## Link to slides, citations and further reading

OR
(QR code and
Bit.ly point to
same link)

https://bit.ly/4ibwML5

# Who gets to "do" responsible AI?

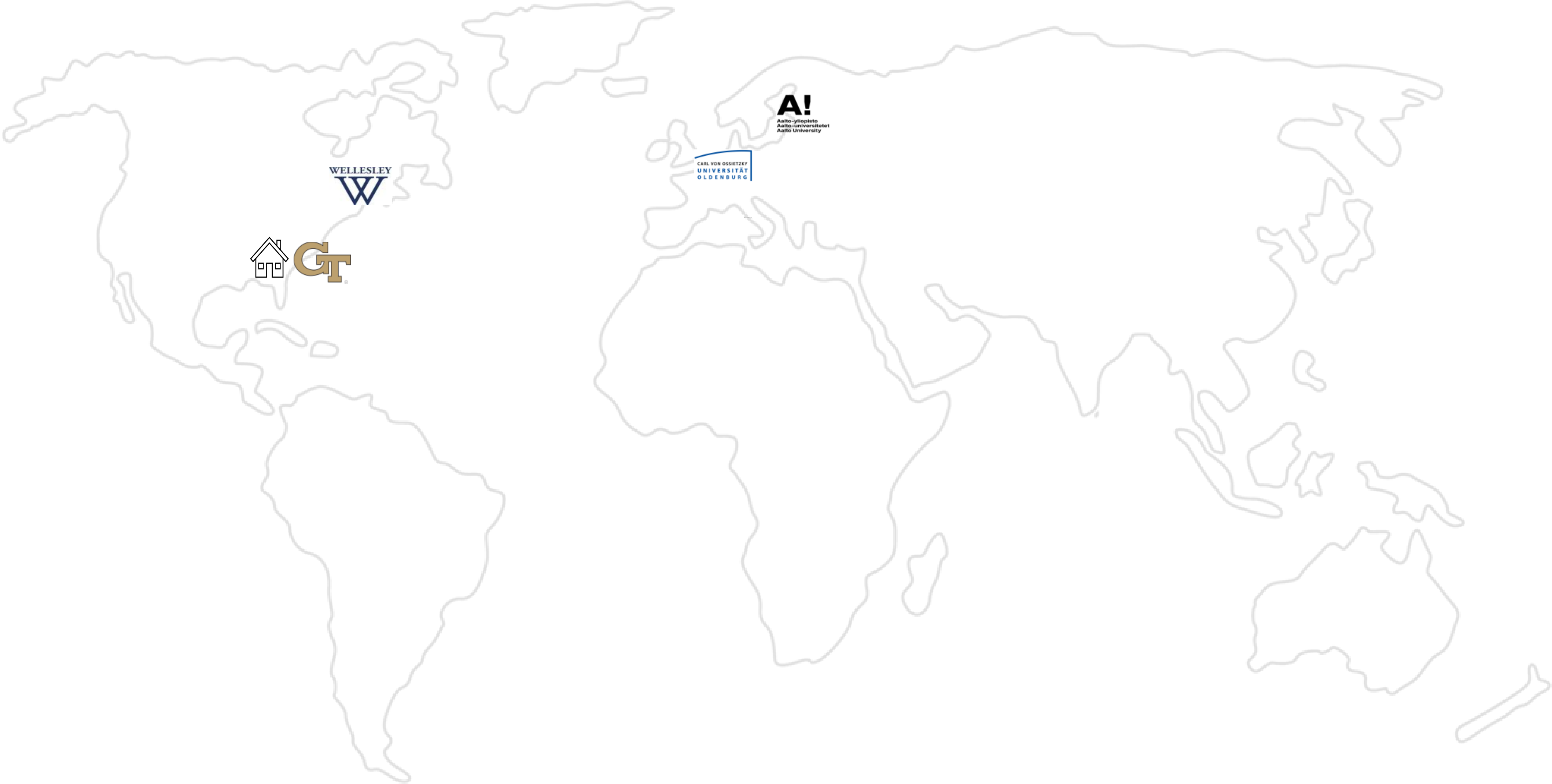

Annabel Rothschild

She/her

arothschild@gatech.edu

www.annabelrothschild.com

## How'd I get here?



System accuracy: numerical or social?

Equality and equity in AI system performance.

(Borenstein & Howard, 2021; Hagendorff, 2020; Jobin et al., 2019)

# Status Quo for Development of "unsafe" (or, unethical) AI systems

Obtain training dataset(s) and use them to develop an AI system

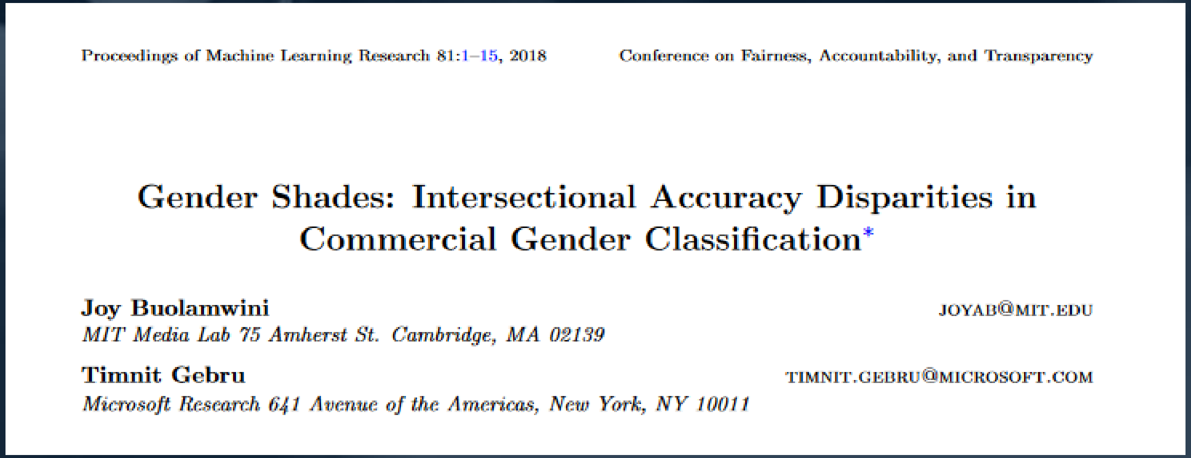
Check generative or predictive performance

It works great! Go to market

And then, an auditing body or individual notices something…

Academic and/or media attention follows

Product developer(s) try to fix the fault



Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

## Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

**Joy Buolamwini** JOYAB@MIT.EDU
*MIT Media Lab 75 Amherst St. Cambridge, MA 02139*

**Timnit Gebru** TIMNIT.GEBRU@MICROSOFT.COM
*Microsoft Research 641 Avenue of the Americas, New York, NY 10011*

## The Harms of AI Systems are not Abstract

### Wrongfully accused by an algorithm

June 24, 2020 at 12:47 pm | *Updated June 24, 2020 at 6:46 pm*

1 of 4 | Robert Julian-Borchak Williams, who was arrested based on a faulty facial recognition match, at home with his wife, Melissa, and their daughters in Farmington Hills, Mich., June 20. (Sylvia Jarrus / The New York Times) **Less**

Two sides of this problem:

1. Our AI systems are faulty along cultural or social axes, because "the dataset wasn't good enough" or "the dataset didn't have enough diversity" .
2. We accept the decisions or predictions of these systems as incontestable truth.

# We had an *unsafe* dataset.

# We need *safe* datasets.

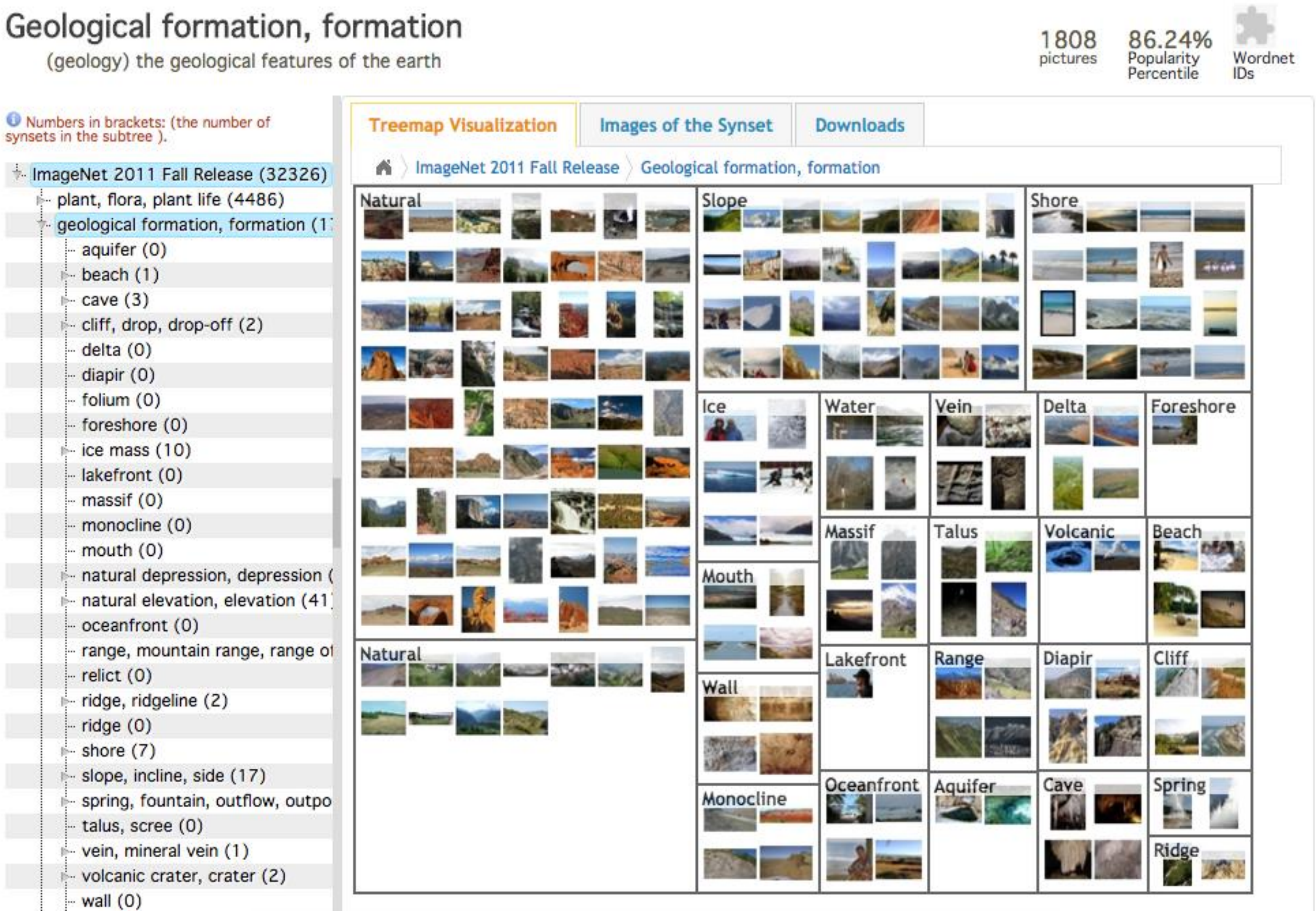
~~1. Who I am, and why I work in computing ethics~~

2. How AI training datasets get developed—ImageNet example
3. Demonstrating the potential of data annotators' insights and experiences
4. *Documenting the Status Quo: Understanding how AI dataset requesters understand and engage with data workers*
5. *Datum Fieldnotes: documenting data workers' insights to support safe dataset use*
6. Future work: the costs of data annotation

# How datasets to train or refine AI systems get developed

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

# What is ImageNet?




- Started in 2006; periodic updates
- 14 million photos of objects with labels—utility for object-recognition (CV) systems
- Labels sourced from Princeton's *WordNet*
- Scrape Goolge Images SERP for term, have Turkers verify

# Why talk about ImageNet?

| TITLE | CITED BY | YEAR |
|---|---|---|
| Imagenet: A large-scale hierarchical image database<br>J Deng, W Dong, R Socher, LJ Li, K Li, L Fei-Fei<br>Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on ... | 76727 | 2009 |

Extremely impactful for the development of the field of computer vision (CV); key benchmarking dataset
(Raji et al. 2021.)

**Pervasive Label Errors in Test Sets**
**Destabilize Machine Learning Benchmarks**

**Curtis G. Northcutt*** ChipBrain, MIT, Cleanlab
**Anish Athalye** MIT, Cleanlab
**Jonas Mueller** AWS

Expensive to build; sticks around
(Northcutt et al., 2021)


One of the first major AI (CV) datasets to be assembled with crowdworkers (Amazon Mechanical Turkers)
(Tsipras et al., 2020)

# The problem(s) with ImageNet

ImageNet given label:
**tailed frog**

Cleanlab guessed: **European green lizard**

MTurk consensus: **European green lizard**

ID: 00032415

ImageNet given label:
**alligator lizard**

Cleanlab guessed: **European green lizard**

MTurk consensus: **European green lizard**

ID: 00033036

ImageNet given label:
**alligator lizard**

Cleanlab guessed: **desert grassland whiptail lizard**

MTurk consensus: **desert grassland whiptail lizard**

ID: 00018979

ImageNet given label:
**alligator lizard**

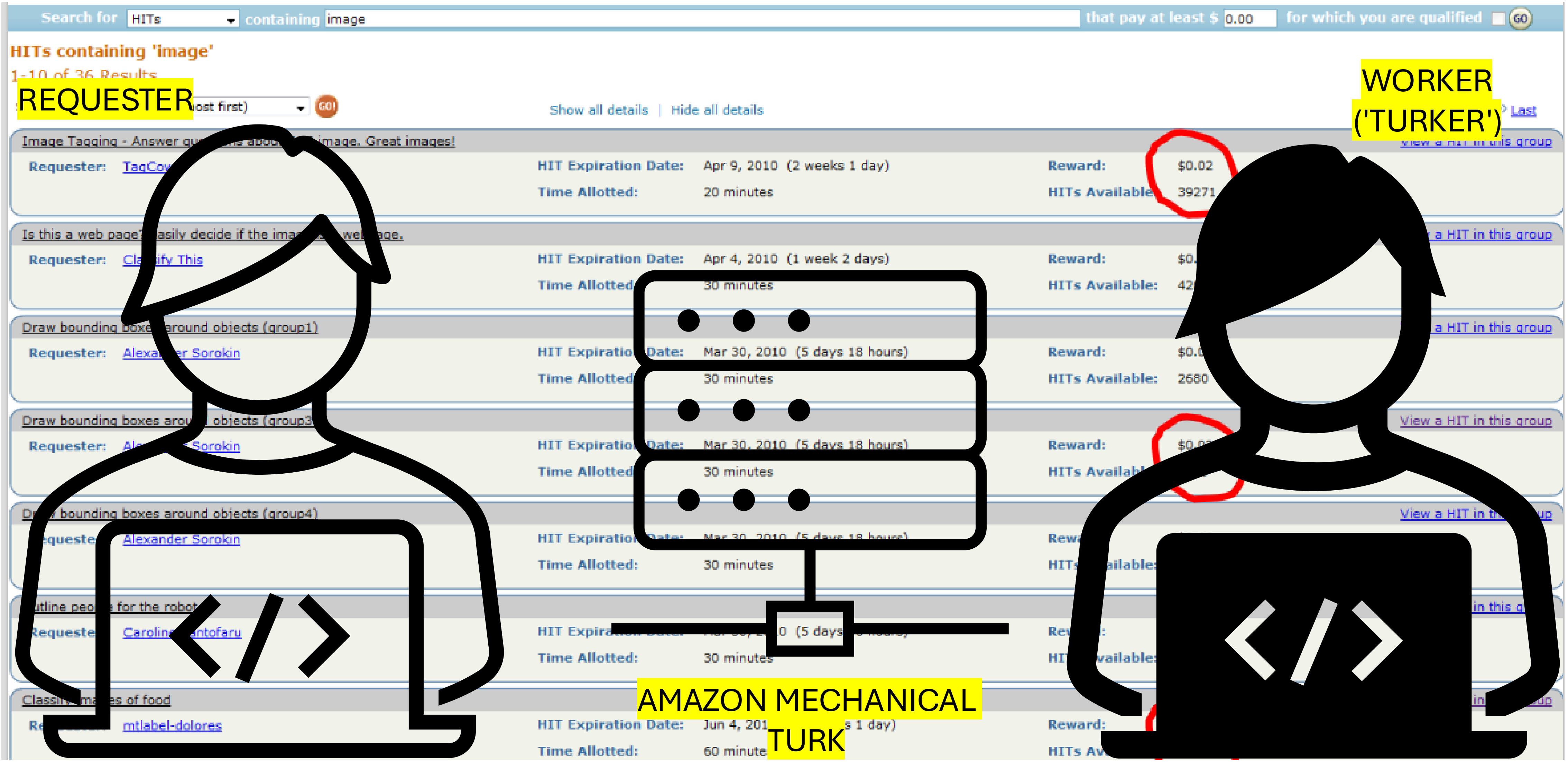
Cleanlab guessed: **desert grassland whiptail lizard**

MTurk consensus: **desert grassland whiptail lizard**

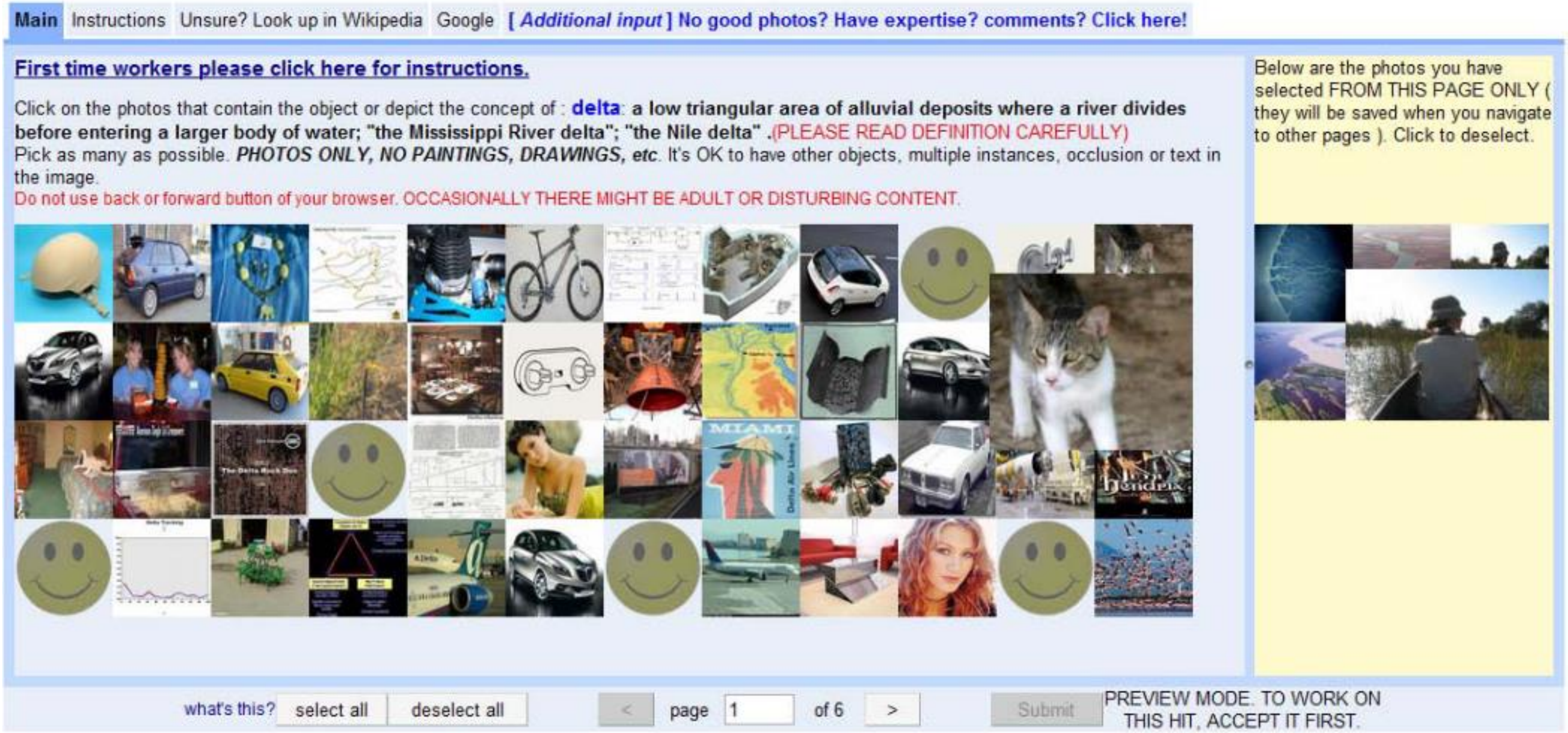
ID: 00028626

(Vasudevan et al., 2022); See for yourself: https://labelerrors.com/

# What is the role of crowdsourcing?



# What workers see when completing an *ImageNet* AMT task ('HIT')...



1. Why weren't data workers asked for their impressions, concerns, and reflections?
2. Going forward, once we have impressions, concerns, and reflections from data workers, how do we record and archive this paradata?

*This is where my work fits in!*

# My work in the context of R-AI more generally

**Collecting, cleaning, organizing, and standardizing data needed to train AI system**

Pre-processing

*Margaret Mitchell*
*Ex-Google, Chief Ethicist at Hugging Face*

In-Processing

**Learning algorithm that comprises AI system is modified**

*Kate Crawford*
*Microsoft*

*Deb Raji*
*Mozilla*

**Predictions generated by AI system are modified**

Post-processing

*Arvind Narayanan & Sayash Kapoor*
*Princeton*

*Alondra Nelson*
*White House OSTP*

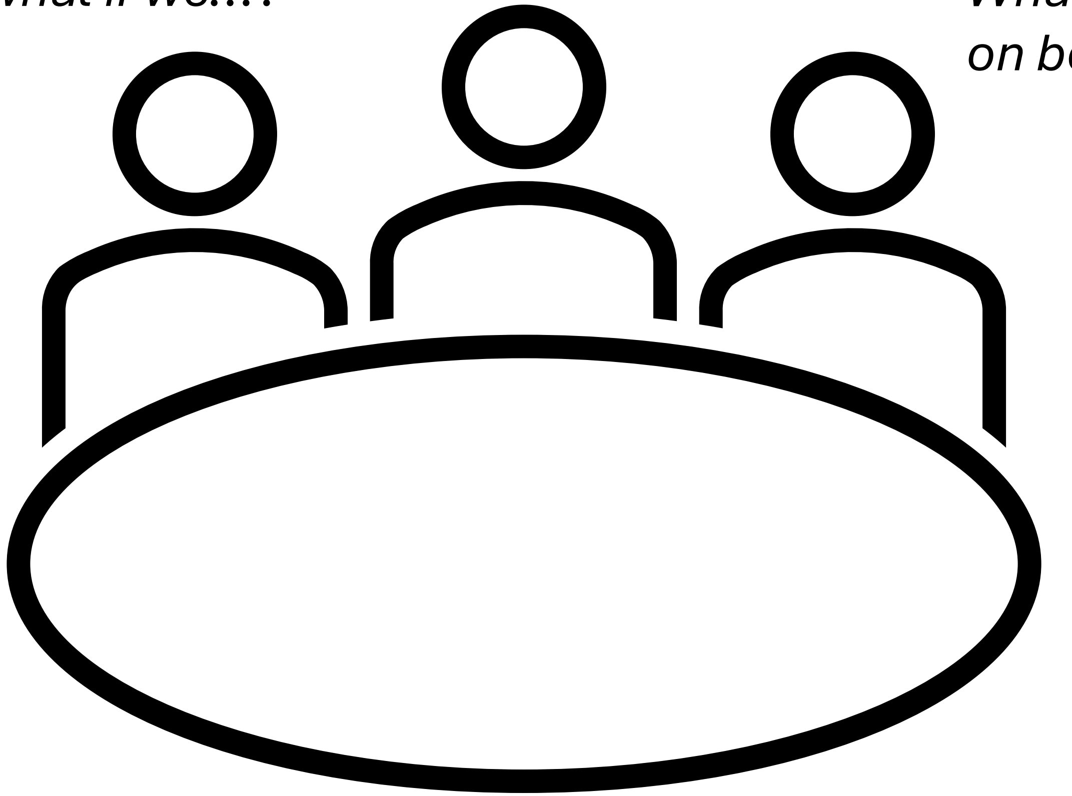
(D'Alessandro et al., 2017)

## DataWorks: *the conference table model*

## An ode to the conference table

*Wait, what if we...?*

*What is the data we work on being used for?*



*What did you mean with this direction?*

*This is close but not quite there—can you readjust...?*

## Critical data literacy: developing dataset auditing skills

- **10 week** Critical Data Literacy curriculum at DataWorks
- **2 modules**: non-computational introducing to AI, data preparation skills
- **90 minute sessions:** mix of short lectures, hands on activities, worksheets, and creative activities

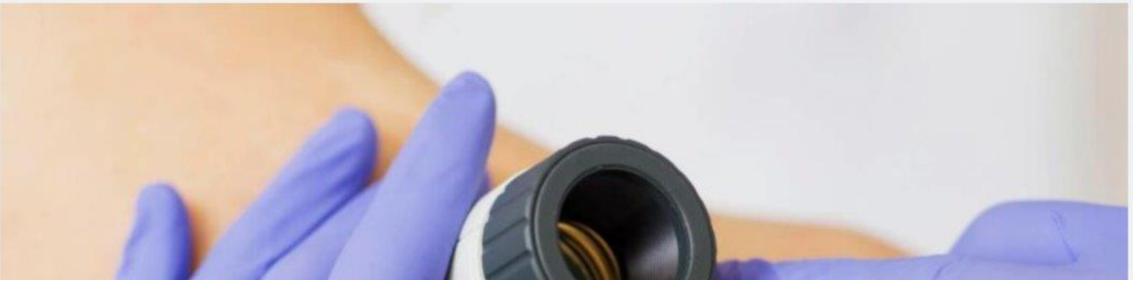
A medical company thinks they can diagnose different kinds of bug bites without patients having to see a doctor in person. Instead, patients will send photos of their bug bites to an algorithm that will tell them what bit them.

*DERMATOLOGY BY AI —*

**Google launches a new medical app—outside the United States**

The dermatology AI app won approval for use in the EU but not with the FDA.

TOM SIMONITE, WIRED.COM - 6/27/2021, 6:30 AM



Lara Schenck, Dana Priest, Gabe Dubose, Zajerria Godfrey, Annabel Rothschild, Ben Rydal Shapiro, and Betsy DiSalvo. 2025. “A Window into Data Apprenticeship: Developing an Integrated Work-Training Curriculum for Novice Adults”. In SIGCSE TS 2025 (ACM Special Interest Group on Computer Science Education).

## Numerous problems quickly arose:

- There is no singular AAVE — locations matter and express differences
- The data was sexually suggestive and racist

## The fundamental premise is suspect:

- “Who really wants to be able to identify Black speakers?”
- “It seems like this could be misused in a way that will hurt my community.”

* Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American Language Bias in Natural Language Generation. https://doi.org/10.48550/arXiv.2305.14291

Carl DiSalvo, Annabel Rothschild, Lara L. Schenck, Ben Shapiro, and Betsy DiSalvo. 2024. “When Workers Want to Say No: A View into Critical Consciousness and Workplace Democracy in Data Work”. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 156 (April 2024)

# Data Use Agreement

Please copy the following agreement below: "I will not use the VALUE dataset for malicious purposes including (but not limited to): deception, impersonation, mockery, discrimination, hate speech, targeted harassment and cultural appropriation. In my use of this resource, I will respect the dignity and privacy of all people." *

Your answer

A reminder that this is human subject research. By clicking "Yes" below, you understand that this dataset is based on a small sample of individuals, and it does not represent all individuals in a culture. *

- YES
- NO

This resource contains synthetic data from transformations based on features of English dialects. This synthetic data was designed to stress-test current NLP systems. It may not fully or accurately represent the natural usage patterns of native speakers. *

- I understand
- I do not understand

This resource is available for non-commercial research purpose only. *

- I understand
- I do not understand

# Data workers **are highly capable** of serving as dataset auditors, when requesters and workers are collaborators.

We also know from citizen science model that this is possible in platform based work.*

Now, how do we make collaboration happen at scale?

*Ashley Boone, Annabel Rothschild, Xander Koo, Grace Pfohl, Alyssa Sheehan, Betsy DiSalvo, Christopher Le Dantec, and Carl DiSalvo. 2024. "Reimagining Meaningful Data Work through Citizen Science". Proc. ACM Hum.-Comput. Interact. January 2024.

How do we build collaborative relationships between data work requesters and workers?

*Pulling chairs up to the conference table*

**Research questions:**

- (RQ1) how do requesters perceive the identity of platform workers?
- (RQ2) what are their views on the workers' motivations and work methods?

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. "The Problems with Proxies: Making Data Work Visible through Requester Practices". AIES (Conference on Artificial Intelligence, Ethics, and Society).

**Final project for Qualitative Methods for HCI (GT MS-HCI course)**

- Students want "real world" projects for their portfolios
- Accrue authentic experience of human factors work (e.g., recruiting)

Two years to conduct this work: starting from course project design in August 22, to publishing work in October 24.

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. “The Problems with Proxies: Making Data Work Visible through Requester Practices”. AIES (Conference on Artificial Intelligence, Ethics, and Society).

# Design Interview Protocol <--> Figure Out Who to Talk to

Semi-structured Interview Protocol

Introduction

1. Introduce the study, the IRB protocol, and yourself.
2. Ask if participant has questions.
3. Ask them to verbally consent.

Ask participants to introduce themselves.
- In general terms, without naming organizations, what do you do? Do you consider that research or product development?

Have participant tell us about their background as a requestor on digital pieceworking platforms (e.g., Amazon Mechanical Turk, Clickworker, Microworker).
- How did you find out about these platforms?
- If from an advisor (if interviewee was a PhD or MS student), how do you think your advisor learned about it?
- Do you ever look at new platforms? If so, how do you think about those new platforms (review them for your purposes)?
- What impact does crowdsourcing, crowd work, or human intelligence have on your career?
- How does it contribute to your productivity or research agenda? (E.g., are there things you are uniquely able to do because of crowdsourcing?)
- Do you anticipate publishing findings from the work you do with pieceworkers?

How did you start requesting on those platforms? E.g., for academic or professional role / duties, for personal project, etc.
- How did you learn to be a requestor? (e.g., Googled it, knew from PhD work, watched YouTube videos, word of mouth, ...)
- What kinds of tasks do you post as a requestor?
- How would you describe the genre(s) of work you post? (E.g., image labeling, 'chat with a robot', quick NLP translation)

Can you walk me through a recent HIT that you posted (or had your students post)?
- What made you choose this platform for this task?
- How did you conceptualize the nature of this task? Why did you need (or choose) to have human annotators / labelers / conversation partners, etc.?
- How did you decide to structure your HIT? What design choices -- e.g., having workers leave AMT proper and move to a Qualtrics survey setup -- did you make and why?

How do you design tasks?
- Do you trial your HITs? Does your lab, workplace (or other professional setting) have any procedures for doing so? (E.g., creator checks for all required questions being answered, or has friends check that wording makes sense, etc.)
- How do you select for workers? Do you have any common defaults (e.g., only accepting 'Master' Turkers)?
- Who do you think works on your task? How do you estimate the time on tasks?
- How do you determine payment? Do you have a mental model for what's an appropriate pay scale, or do you alter it task-by-task?
- How do you use the data that is used to select workers (e.g., criteria info -- like a test to make sure if people are 'qualified')?

Industry and academic professionals who have:

**(1)** Used a crowdsourcing platform (e.g., Amazon MTurk) to source data work, AND:

**(2)** that data work was used to train or refine an AI system.

Potential participants were identified by the students through:

- Relevant online forums
- Own professional networks
- Course staff's professional networks

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. “The Problems with Proxies: Making Data Work Visible through Requester Practices”. AIES (Conference on Artificial Intelligence, Ethics, and Society).

# Memoing --> Thematic Analysis --> Codebook and Final Analysis

Associate interaction

- Contradiction: they want the average American, but they filter works by perceived quality tags (G15P3)
  - Platform tries to qualify workers (eg Master Turker) but rise of requester-imposed qualifications suggests system does not work
- Based in US
  - Getting around GDPR
  - Can circumvent with VPN
- Technical fluency
  - Want to interact with / engage the average American, but posting on a site that requires a fair level of digital fluency and also access
- Associate feedback
  - None of them talk about worker as an equal counterpart; research-participant or requester-worker, where the requester self-describes as being in position of power – demonstrated by lack of feedback requester or solicited
  - No respect for worker – feedback / comment box would indicate respect, at least equal work or fellow worker
- G14P1 – usertesting.com
  - Good result because of modality that platform offers (facilitates think aloud sessions)
- G11P4 – fair payment
  - If people miss the gold standard, they will still pay but they will throw out their data
- Assumptions made, and use them to justify their behaviors, such as picking demographic profiles that are convenient for them (representative of entire United States)
  - Developer developing things – use high end computer, ppl using their systems are using low end devices and things don't render (ICT4D).
    - No one things about how their tasks are being experienced and in what technical setting
  - Requesters think associates are doing it just for pocket money, however people are using this as a primary income source
  - That a native English speaker indicates some kind of standard English (example of a proxy used for a certain level of English) – what does it mean to be a native speaker?
- How do they qualify / justify qualification criteria

| Code in Dedoose | Description of Code | Examples | No |
|---|---|---|---|
| Requestor Background | Quotes that reference who the requestor is and how they approach being a requestor | | |
| > RequestProfessionalContext | Participant describes where they worked AT TIME of requesting | Broken down into two child codes: | |
| >> Research Setting | Work(ed/s) in a research setting | 1 - "I work at a research institute"<br>2 - "I am a research scientist" | |
| >> Commercial Setting | Work(ed/s) in a commercial setting | 1 - "I work on the {FizzBuzz} Model, which we sell to customers looking for a {FizzBuzz} tool"<br>2 - "I was a Senior Engineer at {Large Tech Company}" | |
| > CurrentEmploymentSetting | Participant describes where they NOW work | 1 - "Presently I am an associate professor"<br>2 - "I am currently a software engineer" | |
| > LearningPlatform | When participant mentions how they learned to use the digital pieceworking platform and/or familiarize themselves with the platform | 1 - "I googled 'how do you post tasks on MTurk'"<br>2 - "My adviser sent me a document our lab compiled on how to post tasks"<br>3 - "Appen onboards you, so I worked with my contact there"<br>4 - "My friend {redacted} who had requested before on AMT showed me how to navigate the interface" | |
| > CareerContributions | When the the participant describes how use of (a/the) digital pieceworking platform(s) contributed to their career | 1 - "If I didn't use MTurk, there's no way I could be the reach I need to make an NLP model based on common English variants — I wouldn't be able to do this work otherwise"<br>2 - "I use the datasets I get from MTurk for publications that I need to continue my career"<br>3 - "I actually got my current job because they needed someone who knew how to request on AMT"<br>4 - "Yeah so to train our {FizzBuzz} model [main product participant works on] we needed 1,000 annotated images of cars"<br>5 - "Right, to reach a pool of 20 young adults who had raised a younger sibling, to deploy our survey, we could only do that through AMT" | |
| > Funding | Participant describes where the funding for the tasks they post on the platform(s) comes from | 1 - "Our lab is mostly funded by the NSF"<br>2 - "It's part of our organizational research budget"<br>3 - "Actually our University banned working on AMT in Spring 2020 so we couldn't get reimbursed for work we deployed on the platform" | No |
| PlatformChoiceRationale | Participant mentions using a specific digital pieceworking platform | | |
| > AssociateAccess | Being able to reach outside of the requestor's own network to access a more general or "average" associates | 1 - "I don't know that many people who want to draw bounding boxes"<br>2 - "I wanted members of the 'general public'" | |
| > NumberAssociates | Size of the associate population available to complete their tasks | 1 - "To get the model right, we needed to get at least a million variations"<br>2 - "I needed lots of examples to get a robust response" | |
| > SpecificPopulation | Being able to access a specific population or demographic | 1 - "I don't know where else to find video game players"<br>2 - "I needed to find Swedish speakers"<br>3 - "I didn't know where else to reach Arabic speakers" | |
| > AssociatesReputation | Reputation of associate labor and quality of work as a platform attribute | 1 - "Prolific workers just have a better reputation generally since they have to share their demographics beforehand"<br>2 - "I think AMT workers are mostly bots so I don't really trust them" | |
| > Turnaround | How quickly their work (in the form of the tasks posted) are completed | 1 - "I had a really quick turn around"<br>2 – "We needed just-in-time submissions" | |
| > Cost | Financial specifics of a given platform | 1 - "It's the cheapest platform"<br>2 - "There's no minimum task number to post"<br>3 - "Prolific is generally not the cheapest, but we find the survey data is much more reliable from Prolific participants so the price is worth it" | |
| > CommonPractice | Choosing to use a specific platform because it is a common or accepted practice within their professional or academic field | 1 - "Everyone in psychology now uses MTurk"<br>2 - "Lots of other papers use it, so we don't have to explain it in methodology"<br>3 - "I mean everyone in my lab uses AMT so I just used it too" | |

Task design -- ALL .XLSX Saved to Drive

File Edit View Insert Format Data Tools Help

Menus 100% $ % .0 .00 123 Calibri 11

H5 False

| | A | B | C |
|---|---|---|---|
| 1 | Media Title | Excerpt Copy | Codes Applied Combined |
| 2 | G6P4.docx | And I am guessing if two years back you found resources good enough to help you navigate it, pretty sure in two years they must have more helpful resources out there. That's nice. Yeah. So, what kind of tasks did you post as a requester? Can you describe it a little?Myeonghan: Oh. So, I deployed an application using Microsoft's Azure Cloud platform, and so the task was basically go open the link and follow the instructions of the app, there were I guess around three visualization tasks. Yeah. | Progress_process, Task_design |
| 3 | G6P4.docx | No, I first tried the task. And also, I asked some people around me to do the task. So, I was thinking it would take around 30 minutes to do the task and then it really took about an hour for others. So, I, I even changed my experiment design from within subject to between subjects. | Task_timing, Testing, Task_design |
| 4 | G6P4.docx | So, do you have like, maybe like you or someone you work with, have any specific procedure on how they specifically design the trials to make it more efficient and user friendly?Myeonghan: Well, I think I thoroughly followed some basic principles of user centered design. So even though it was like not really designing an interface but looking at how people, how people interact and interpret with that, the color of the graphs and all. Mm hmm. And then based on that, I made some changes in the graph for better experience.Upasana: So, when you design these trials, do you have a specific crowd in mind, or you design it so that any layman can understand it and do it? Myeonghan: Oh, yeah. Upasana: So, for anyone? Myeonghan: Yeah. Oh. So. But while design the trial I included a few tests which didn't actually count towards the experiment result which were pretty straightforward. Also, I made it in such a way that they could attempt the activity several times. | Testing, Progress_process, Task_design |
| 5 | G6P4.docx | Okay, so what do you the think about the payment on these platforms? Do you think there is an appropriate pay scale?Myeonghan: I consider the minimum wage in the US. So, I, I was at Purdue University, Indiana for my Master's.Upasana: Oh, really!Myeonghan: Yeah, so, I'm not sure if minimum wage varies by, like, different states or not, but I am not sure about that. But, you know, in a paper I, I read while preparing for this study, they said they consider minimum wage in the US and they pay the amount relative to the time spent. Where, suppose if the minimum wage was $10 for an hour and you spent like 30 minutes, so they paid them $5 instead. | Payment_calibration, Task_design |
| 6 | G6P4.docx | Oh, no, no. Number of participants was in the plan. Okay. Yeah, because I ran an apriori power analysis which helped me decide the proper number of participants. Okay. It was around 200. And then I planned to recruit 200 crowd-workers. Okay. Yeah. Maybe because of some bugs in the system, I couldn't get data from some of them. Then I needed to recruit some additional crowd-workers. But yeah, everything was in the plan. | Progress_process, Task_design |
| 7 | G6P4.docx | Well. So, to make sure participants finish all the tasks, I requested them to show me some quote that they can see on the last page of the task. Oh, okay. Yeah. So, if they make it to the end, then they can see it. And if they tell me the correct quote, then I approved their submission. | Associate_validation, Task_design |
| 8 | | o. I think the Prolific provided a feature to filter participants. So basically, many people just sort of come and go, like, they just come in and just leave. Leave without finishing the task, but only, only the participants who provided the quote for me, were considered valid workers. And also, I added a few attention checking test, which were very straightforward and easy, but to see if they just like mindlessly click anything, anything, and then they might, they might they are likely to give the wrong answers to those | |

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. "The Problems with Proxies: Making Data Work Visible through Requester Practices". AIES (Conference on Artificial Intelligence, Ethics, and Society).

## What We Found: Who is requesting these tasks?

- Mix of *commercial* (14) and *research* (38) participants
- Learning habits from *co-workers and collaborators*, *search engine links*, *platform documentation*, and *YouTube*
- Tasks include *data annotation & classification*, *data collection*, *tool or system feedback*
- Platforms ranging from *Amazon MTurk* to *Tokola* and *iMerit*

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. “The Problems with Proxies: Making Data Work Visible through Requester Practices”. AIES (Conference on Artificial Intelligence, Ethics, and Society).

# What We Found: Proxies to Establish Worker Identity and Abilities

- Workers as "the general public," but simultaneously highly curated by requesters
- Good data vs bad actors
- Testing proxies, rather than actual attributes; quantification of lived experience and skills
  - Worker identity: *English fluency*, *age*, *location*
  - Aptitude: pre hoc (*approval rating, prior task completion*), post hoc (*keyboard interaction, answer pattern, attention check, gut reaction, coherence as trust*)

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. "The Problems with Proxies: Making Data Work Visible through Requester Practices". AIES (Conference on Artificial Intelligence, Ethics, and Society).

# What We Found: Proxies vs. Reality

- FA(C)T(E) notions of reliability and validity (Jacobs & Wallach, 2023)
- *Construct reliability* (reliable, replicable measurement; e.g., 10 vs 10,000 prior tasks)
- *Construct validity* (measurements grounded in construct, encompass all relevant aspects; e.g., worker actually skilled at the task)

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. “The Problems with Proxies: Making Data Work Visible through Requester Practices”. AIES (Conference on Artificial Intelligence, Ethics, and Society).

# Why Do We Care About Proxies?

- Proxies demonstrate lack genuine trust and collaboration, which characterize the "conference table" model of data work.
- Without trust and collaboration, data workers as auditors is impossible…
- But there's hope! Requesters want to learn how to do this better, and we can build the tools, systems, and processes to help them do it!
- Short term future work (students take note!): pro-social task design for requesters.

Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo. 2024. “The Problems with Proxies: Making Data Work Visible through Requester Practices”. AIES (Conference on Artificial Intelligence, Ethics, and Society).

# Archiving and Documenting Data Workers' Auditing Work

*Making a record of conference table conversations*

# Motivation: Paradata is Critical to Algorithmic Accountability

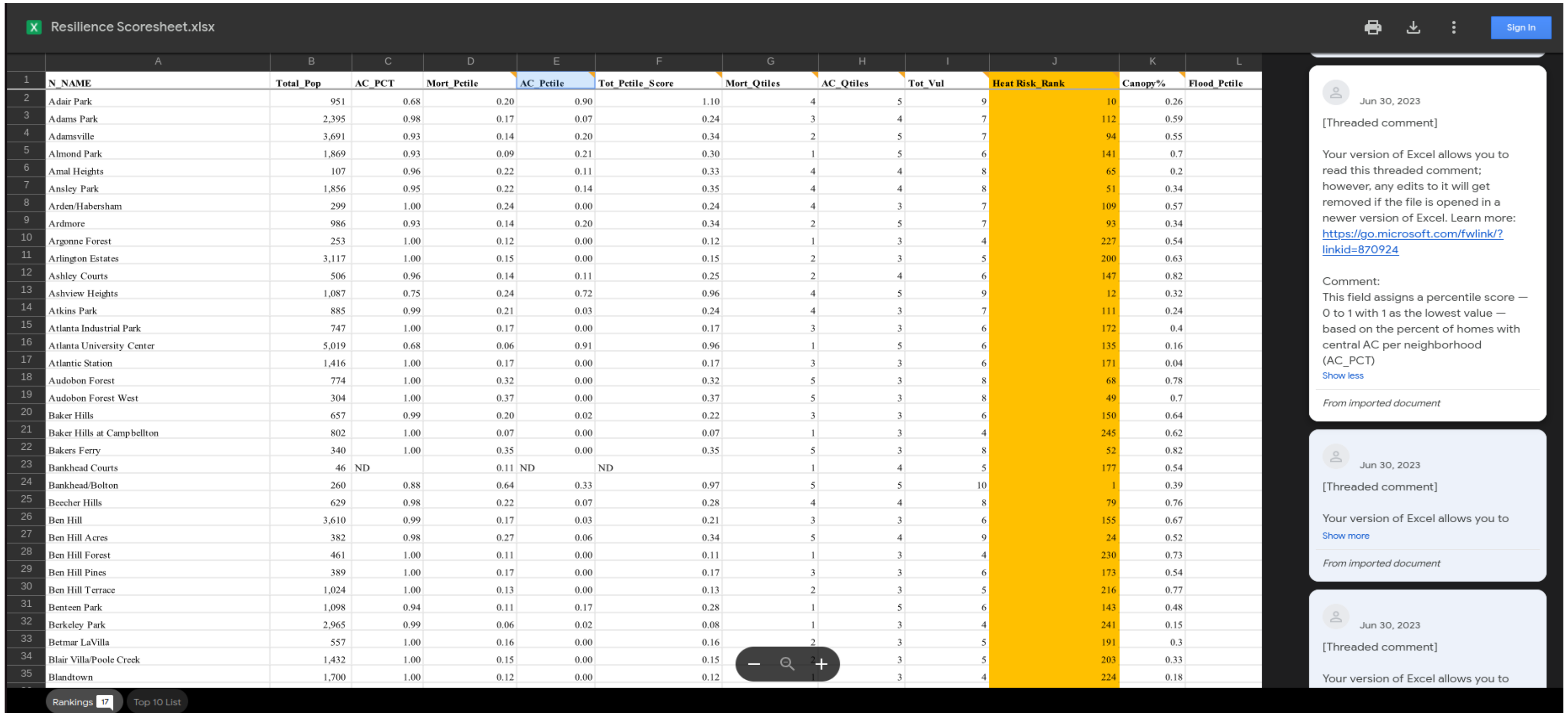
That the AI community should "**use paradata to create unified reporting models** that enhance the explainability of algorithms and algorithmic systems" and that" packaged data in the form of descriptions and documentation are contextualized understandings of work practices and processes."[1]

In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, **we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on**.[2]

1. Ciaran B. Trace and James A. Hodges. 2024. The Role of Paradata in Algorithmic Accountability. In Perspectives on Paradata: Research and Practice of Documenting Process Knowledge, Isto Huvila, Lisa Andersson and Olle Sköld (eds.). Springer International Publishing, Cham, 197–213. https://doi.org/10.1007/978-3-031-53946-6_11
2. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM 64, 12: 86–92. https://doi.org/10.1145/3458723

# Datum Fieldnotes Inspiration

Civic and non-profit (as non tech industry) data workers have novel ways of documenting and contextualizing the datasets they work with.*

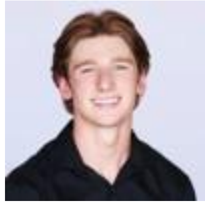



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | N_NAME | Total_Pop | AC_PCT | Mort_Pctile | AC_Pctile | Tot_Pctile_Score | Mort_Qtiles | AC_Qtiles | Tot_Vul | Heat Risk_Rank | Canopy% | Flood_Pctile |
| 2 | Adair Park | 951 | 0.68 | 0.20 | 0.90 | 1.10 | 4 | 5 | 9 | 10 | 0.26 | |
| 3 | Adams Park | 2,395 | 0.98 | 0.17 | 0.07 | 0.24 | 3 | 4 | 7 | 112 | 0.59 | |
| 4 | Adamsville | 3,691 | 0.93 | 0.14 | 0.20 | 0.34 | 2 | 5 | 7 | 94 | 0.55 | |
| 5 | Almond Park | 1,869 | 0.93 | 0.09 | 0.21 | 0.30 | 1 | 5 | 6 | 141 | 0.7 | |
| 6 | Amal Heights | 107 | 0.96 | 0.22 | 0.11 | 0.33 | 4 | 4 | 8 | 65 | 0.2 | |
| 7 | Ansley Park | 1,856 | 0.95 | 0.22 | 0.14 | 0.35 | 4 | 4 | 8 | 51 | 0.34 | |
| 8 | Arden/Habersham | 299 | 1.00 | 0.24 | 0.00 | 0.24 | 4 | 3 | 7 | 109 | 0.57 | |
| 9 | Ardmore | 986 | 0.93 | 0.14 | 0.20 | 0.34 | 2 | 5 | 7 | 93 | 0.34 | |
| 10 | Argonne Forest | 253 | 1.00 | 0.12 | 0.00 | 0.12 | 1 | 3 | 4 | 227 | 0.54 | |
| 11 | Arlington Estates | 3,117 | 1.00 | 0.15 | 0.00 | 0.15 | 2 | 3 | 5 | 200 | 0.63 | |
| 12 | Ashley Courts | 506 | 0.96 | 0.14 | 0.11 | 0.25 | 2 | 4 | 6 | 147 | 0.82 | |
| 13 | Ashview Heights | 1,087 | 0.75 | 0.24 | 0.72 | 0.96 | 4 | 5 | 9 | 12 | 0.32 | |
| 14 | Atkins Park | 885 | 0.99 | 0.21 | 0.03 | 0.24 | 4 | 3 | 7 | 111 | 0.24 | |
| 15 | Atlanta Industrial Park | 747 | 1.00 | 0.17 | 0.00 | 0.17 | 3 | 3 | 6 | 172 | 0.4 | |
| 16 | Atlanta University Center | 5,019 | 0.68 | 0.06 | 0.91 | 0.96 | 1 | 5 | 6 | 135 | 0.16 | |
| 17 | Atlantic Station | 1,416 | 1.00 | 0.17 | 0.00 | 0.17 | 3 | 3 | 6 | 171 | 0.04 | |
| 18 | Audobon Forest | 774 | 1.00 | 0.32 | 0.00 | 0.32 | 5 | 3 | 8 | 68 | 0.78 | |
| 19 | Audobon Forest West | 304 | 1.00 | 0.37 | 0.00 | 0.37 | 5 | 3 | 8 | 49 | 0.7 | |
| 20 | Baker Hills | 657 | 0.99 | 0.20 | 0.02 | 0.22 | 3 | 3 | 6 | 150 | 0.64 | |
| 21 | Baker Hills at Campbellton | 802 | 1.00 | 0.07 | 0.00 | 0.07 | 1 | 3 | 4 | 245 | 0.62 | |
| 22 | Bakers Ferry | 340 | 1.00 | 0.35 | 0.00 | 0.35 | 5 | 3 | 8 | 52 | 0.82 | |
| 23 | Bankhead Courts | 46 | ND | 0.11 | ND | ND | 1 | 4 | 5 | 177 | 0.54 | |
| 24 | Bankhead/Bolton | 260 | 0.88 | 0.64 | 0.33 | 0.97 | 5 | 5 | 10 | 1 | 0.39 | |
| 25 | Beecher Hills | 629 | 0.98 | 0.22 | 0.07 | 0.28 | 4 | 4 | 8 | 79 | 0.76 | |
| 26 | Ben Hill | 3,610 | 0.99 | 0.17 | 0.03 | 0.21 | 3 | 3 | 6 | 155 | 0.67 | |
| 27 | Ben Hill Acres | 382 | 0.98 | 0.27 | 0.06 | 0.34 | 5 | 4 | 9 | 24 | 0.52 | |
| 28 | Ben Hill Forest | 461 | 1.00 | 0.11 | 0.00 | 0.11 | 1 | 3 | 4 | 230 | 0.73 | |
| 29 | Ben Hill Pines | 389 | 1.00 | 0.17 | 0.00 | 0.17 | 3 | 3 | 6 | 173 | 0.54 | |
| 30 | Ben Hill Terrace | 1,024 | 1.00 | 0.13 | 0.00 | 0.13 | 2 | 3 | 5 | 216 | 0.77 | |
| 31 | Benteen Park | 1,098 | 0.94 | 0.11 | 0.17 | 0.28 | 1 | 5 | 6 | 143 | 0.48 | |
| 32 | Berkeley Park | 2,965 | 0.99 | 0.06 | 0.02 | 0.08 | 1 | 3 | 4 | 241 | 0.15 | |
| 33 | Betmar LaVilla | 557 | 1.00 | 0.16 | 0.00 | 0.16 | 2 | 3 | 5 | 191 | 0.3 | |
| 34 | Blair Villa/Poole Creek | 1,432 | 1.00 | 0.15 | 0.00 | 0.15 | 2 | 3 | 5 | 203 | 0.33 | |
| 35 | Blandtown | 1,700 | 1.00 | 0.12 | 0.00 | 0.12 | 1 | 3 | 4 | 224 | 0.18 | |

but sharing datasets—an essential part of contextualization—is tricky, particularly when it comes to annotations.

*Annabel Rothschild, Amanda Meng, Carl DiSalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. ”Interrogating Data Work as a Community of Practice”. Proceedings of the ACM on Human-Computer Interaction 6, Article 307 (November 2022) (2022), 29.

# Datum Fieldnotes




Google Sheets add on

Automates micro-documentation (per datum)

Supports collaboration and local expertise

Paper in preparation: William Eickman*, Mukhlisa Nematova*, Annabel Rothschild, Carl DiSalvo, and Betsy DiSalvo. "Datum Fieldnotes: Automating and Archiving Data Workers' Contextualization Practices and Insights". CHI LBW, 2025.

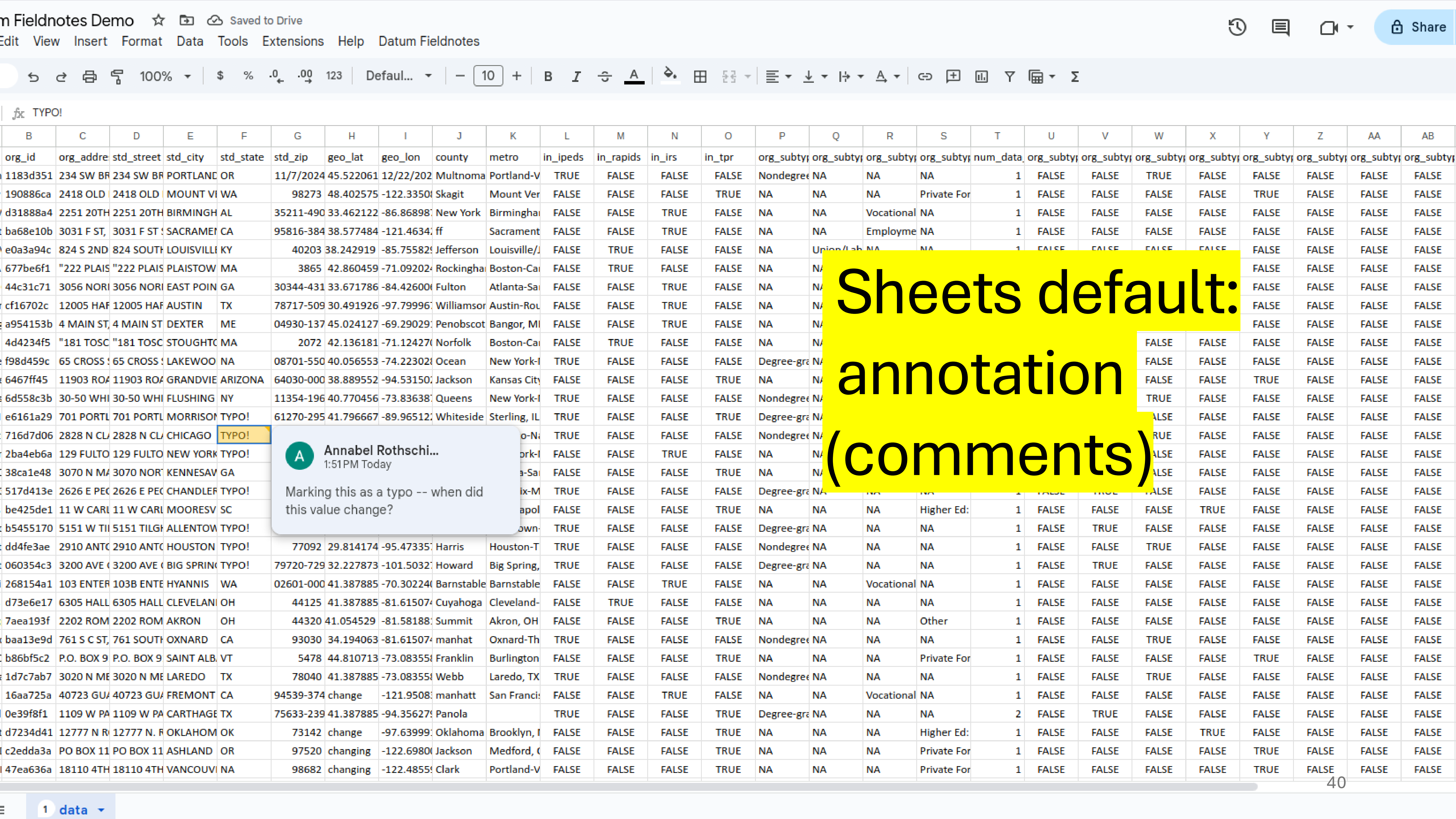
**(RQ1)** Can we offload the labor of dataset micro-documentation from data workers?

**(RQ2)** How do data workers perform dataset contextualization practices in situ (in sheet)?

*"Infrastructural inversion" – bringing background work to the forefront so that you can reflect on the social angles of production and the notion of quantification (p. 34) (Bowker & Star, 2008)*

## Design Criteria

- **Needs to be cheap or free: Google Sheets**
  - Civic and non-profits have high staff turnover (Nault, 2020)
  - Build on existing features of accessibility, adaptability, and replicability (Shapiro & Oystrick, 2018)
  - Google's existing privacy infrastructure and cloud-hosting (Harmon et al., 2017)
  - No or little new organizational learning (Benjamin et al., 2018)
  - Ability to share and transfer between organizations (Voida et al., 2011; Erete et al., 2016)
- **Log, and in CSV format:**
  - Missing data is an ethical issue for civic and non-profit groups (Nault et al., 2020)
  - Operability between different organizations (Davies & Frank, 2013)
  - Losing control of dataset (Darian et al., 2023)
  - Assistive automation (Shapiro & Oystrick, 2018)


m Fieldnotes Demo
Saved to Drive
Edit View Insert Format Data Tools Extensions Help Datum Fieldnotes
Share
TYPO!
org_id org_addre std_street std_city std_state std_zip geo_lat geo_lon county metro in_ipeds in_rapids in_irs in_tpr org_subtyp org_subtyp org_subtyp org_subtyp num_data org_subtyp org_subtyp org_subtyp org_subtyp org_subtyp org_subtyp org_subtyp org_subtyp
Annabel Rothschi...
1:51 PM Today
Marking this as a typo -- when did this value change?
Sheets default: annotation (comments)
data

, 2:01PM

# Sheets default: version history

| B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| org_id | org_addre | std_street | std_city | std_state | std_zip | geo_lat | geo_lon | county | metro | in_ipeds | in_rapids | in_irs | in_tpr | org_subtyp | org_subtyp | org_subtyp | org_subtyp | num_data | org_subtyp | org_subtyp | org_subtyp | org_subtyp |
| 1183d351 | 234 SW BR | 234 SW BR | PORTLAND | OR | 11/7/2024 | 45.522061 | 12/22/202 | Multnoma | Portland-V | TRUE | FALSE | FALSE | FALSE | Nondegree | NA | NA | NA | 1 | FALSE | FALSE | TRUE | FALSE |
| 190886ca | 2418 OLD | 2418 OLD | MOUNT VI | WA | 98273 | 48.402575 | -122.3350 | Skagit | Mount Ver | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Private For | 1 | FALSE | FALSE | FALSE | FALSE |
| d31888a4 | 2251 20TH | 2251 20TH | BIRMINGH | AL | 35211-490 | 33.462122 | -86.86898 | New York | Birminghai | FALSE | FALSE | TRUE | FALSE | NA | NA | Vocational | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| ba68e10b | 3031 F ST, | 3031 F ST | SACRAMEI | CA | 95816-384 | 38.577484 | -121.4634 | ff | Sacrament | FALSE | FALSE | TRUE | FALSE | NA | NA | Employme | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| e0a3a94c | 824 S 2ND | 824 SOUTH | LOUISVILLI | KY | 40203 | 38.242919 | -85.75582 | Jefferson | Louisville/ | FALSE | TRUE | FALSE | FALSE | NA | Union/Lab | NA | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 677be6f1 | "222 PLAIS | "222 PLAIS | PLAISTOW | MA | 3865 | 42.860459 | -71.09202 | Rockingha | Boston-Cai | FALSE | TRUE | FALSE | FALSE | NA | NA | NA | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 44c31c71 | 3056 NORI | 3056 NORI | EAST POIN | GA | 30344-431 | 33.671786 | -84.42600 | Fulton | Atlanta-Sa | FALSE | FALSE | TRUE | FALSE | NA | NA | Employme | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| cf16702c | 12005 HAF | 12005 HAF | AUSTIN | TX | 78717-509 | 30.491926 | -97.79996 | Williamsor | Austin-Rou | FALSE | FALSE | TRUE | FALSE | NA | NA | Vocational | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| a954153b | 4 MAIN ST, | 4 MAIN ST | DEXTER | ME | 04930-137 | 45.024127 | -69.29029 | Penobscot | Bangor, MI | FALSE | FALSE | TRUE | FALSE | NA | NA | Vocational | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 4d4234f5 | "181 TOSC | "181 TOSC | STOUGHTC | MA | 2072 | 42.136181 | -71.1242 | | | | | | | | | | | | | | | |
| f98d459c | 65 CROSS | 65 CROSS | LAKEWOO | NA | 08701-550 | 40.056553 | -74.2230 | | | | | | | | | | | | | | | |
| 6467ff45 | 11903 ROA | 11903 ROA | GRANDVIE | ARIZONA | 64030-000 | 38.889552 | -94.5315 | | | | | | | | | | | | | | | |
| 6d558c3b | 30-50 WHI | 30-50 WHI | FLUSHING | NY | 11354-196 | 40.770456 | -73.8363 | | | | | | | | | | | | | | | |
| e6161a29 | 701 PORTL | 701 PORTL | MORRISOI | NA | 61270-295 | 41.796667 | -89.96512 | Whiteside | Sterling, IL | TRUE | FALSE | FALSE | TRUE | Degree-gra | NA | NA | NA | 2 | FALSE | TRUE | FALSE | FALSE |
| 716d7d06 | 2828 N CL | 2828 N CL | CHICAGO | IL | 60657 | 41.933392 | -87.64554 | Cook | Chicago-Na | TRUE | FALSE | FALSE | FALSE | Nondegree | NA | NA | NA | 1 | FALSE | FALSE | TRUE | FALSE |
| 2ba4eb6a | 129 FULTO | 129 FULTO | NEW YORK | NA | 10038-271 | 40.710295 | -74.00745 | New York | New York-I | FALSE | FALSE | TRUE | FALSE | NA | NA | Employme | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 38ca1e48 | 3070 N MA | 3070 NOR | KENNESAV | GA | 30144 | 34.029759 | -84.62645 | Cobb | Atlanta-Sa | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Other | 1 | FALSE | FALSE | FALSE | FALSE |
| 517d413e | 2626 E PEC | 2626 E PEC | CHANDLEF | NA | 85225-249 | 33.295398 | -111.7963 | KINGS | Phoenix-M | TRUE | FALSE | FALSE | FALSE | Degree-gra | NA | NA | NA | 1 | FALSE | TRUE | FALSE | FALSE |
| be425de1 | 11 W CARL | 11 W CARL | MOORESV | SC | 46158 | 39.616964 | -86.37989 | Morgan | Indianapol | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Higher Ed: | 1 | FALSE | FALSE | FALSE | TRUE |
| b5455170 | 5151 W TI | 5151 TILGH | ALLENTOW | TYPO! | 30516.5 | 40.591817 | -75.56888 | Lehigh | Allentown- | TRUE | FALSE | FALSE | FALSE | Degree-gra | NA | NA | NA | 1 | FALSE | TRUE | FALSE | FALSE |
| dd4fe3ae | 2910 ANTC | 2910 ANTC | HOUSTON | TYPO! | 77092 | 29.814174 | -95.47335 | Harris | Houston-T | TRUE | FALSE | FALSE | FALSE | Nondegree | NA | NA | NA | 1 | FALSE | FALSE | TRUE | FALSE |
| 060354c3 | 3200 AVE ( | 3200 AVE ( | BIG SPRIN | TYPO! | 79720-729 | 32.227873 | -101.5032 | Howard | Big Spring, | TRUE | FALSE | FALSE | FALSE | Degree-gra | NA | NA | NA | 1 | FALSE | TRUE | FALSE | FALSE |
| 268154a1 | 103 ENTEF | 103B ENTE | HYANNIS | WA | 02601-000 | 41.387885 | -70.30224 | Barnstable | Barnstable | FALSE | FALSE | TRUE | FALSE | NA | NA | Vocational | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| d73e6e17 | 6305 HALL | 6305 HALL | CLEVELANI | OH | 44125 | 41.387885 | -81.61507 | Cuyahoga | Cleveland- | FALSE | TRUE | FALSE | FALSE | NA | NA | NA | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 7aea193f | 2202 ROM | 2202 ROM | AKRON | OH | 44320 | 41.054529 | -81.58188 | Summit | Akron, OH | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Other | 1 | FALSE | FALSE | FALSE | FALSE |
| baa13e9d | 761 S C ST, | 761 SOUTH | OXNARD | CA | 93030 | 34.194063 | -81.61507 | manhat | Oxnard-Th | TRUE | FALSE | FALSE | FALSE | Nondegree | NA | NA | NA | 1 | FALSE | FALSE | TRUE | FALSE |
| b86bf5c2 | P.O. BOX 9 | P.O. BOX 9 | SAINT ALB. | VT | 5478 | 44.810713 | -73.08355 | Franklin | Burlington | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Private For | 1 | FALSE | FALSE | FALSE | FALSE |
| 1d7c7ab7 | 3020 N ME | 3020 N ME | LAREDO | TX | 78040 | 41.387885 | -73.08355 | Webb | Laredo, TX | TRUE | FALSE | FALSE | FALSE | Nondegree | NA | NA | NA | 1 | FALSE | FALSE | TRUE | FALSE |
| 16aa725a | 40723 GUA | 40723 GUA | FREMONT | CA | 94539-374 | change | -121.9508 | manhatt | San Franci | FALSE | FALSE | TRUE | FALSE | NA | NA | Vocational | NA | 1 | FALSE | FALSE | FALSE | FALSE |
| 0e39f8f1 | 1109 W PA | 1109 W PA | CARTHAGE | TX | 75633-239 | 41.387885 | -94.35627 | Panola | | TRUE | FALSE | FALSE | TRUE | Degree-gra | NA | NA | NA | 2 | FALSE | TRUE | FALSE | FALSE |
| d7234d41 | 12777 N R | 12777 N. F | OKLAHOM | OK | 73142 | change | -97.63999 | Oklahoma | Brooklyn, I | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Higher Ed: | 1 | FALSE | FALSE | FALSE | TRUE |
| c2edda3a | PO BOX 11 | PO BOX 11 | ASHLAND | OR | 97520 | changing | -122.6980 | Jackson | Medford, ( | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Private For | 1 | FALSE | FALSE | FALSE | FALSE |
| 47ea636a | 18110 4TH | 18110 4TH | VANCOUVI | NA | 98682 | changing | -122.4855 | Clark | Portland-V | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Private For | 1 | FALSE | FALSE | FALSE | FALSE |
| f94b8bb3 | 1215 E ELN | 1215 E. ELI | PHOENIX | AZ | 94539-374 | change | -112.0555 | | Phoenix-M | FALSE | FALSE | FALSE | TRUE | NA | NA | NA | Private For | 1 | FALSE | FALSE | FALSE | FALSE |

Sheet1 | Log3 | in | Log5 | Log4 | Sheet99 | data | Log | Datasheet for Dataset Use and Distribution

## Version history

All versions

Today

November 7, 2:01PM
*Current version*
Annabel Rothschild

November 7, 1:52PM
Annabel Rothschild

November 7, 1:51PM
Annabel Rothschild

November 7, 1:47PM
Annabel Rothschild

November 7, 1:47PM
Annabel Rothschild

November 7, 1:46PM
Annabel Rothschild

November 7, 1:45PM
Annabel Rothschild

November 7, 1:45PM
Annabel Rothschild

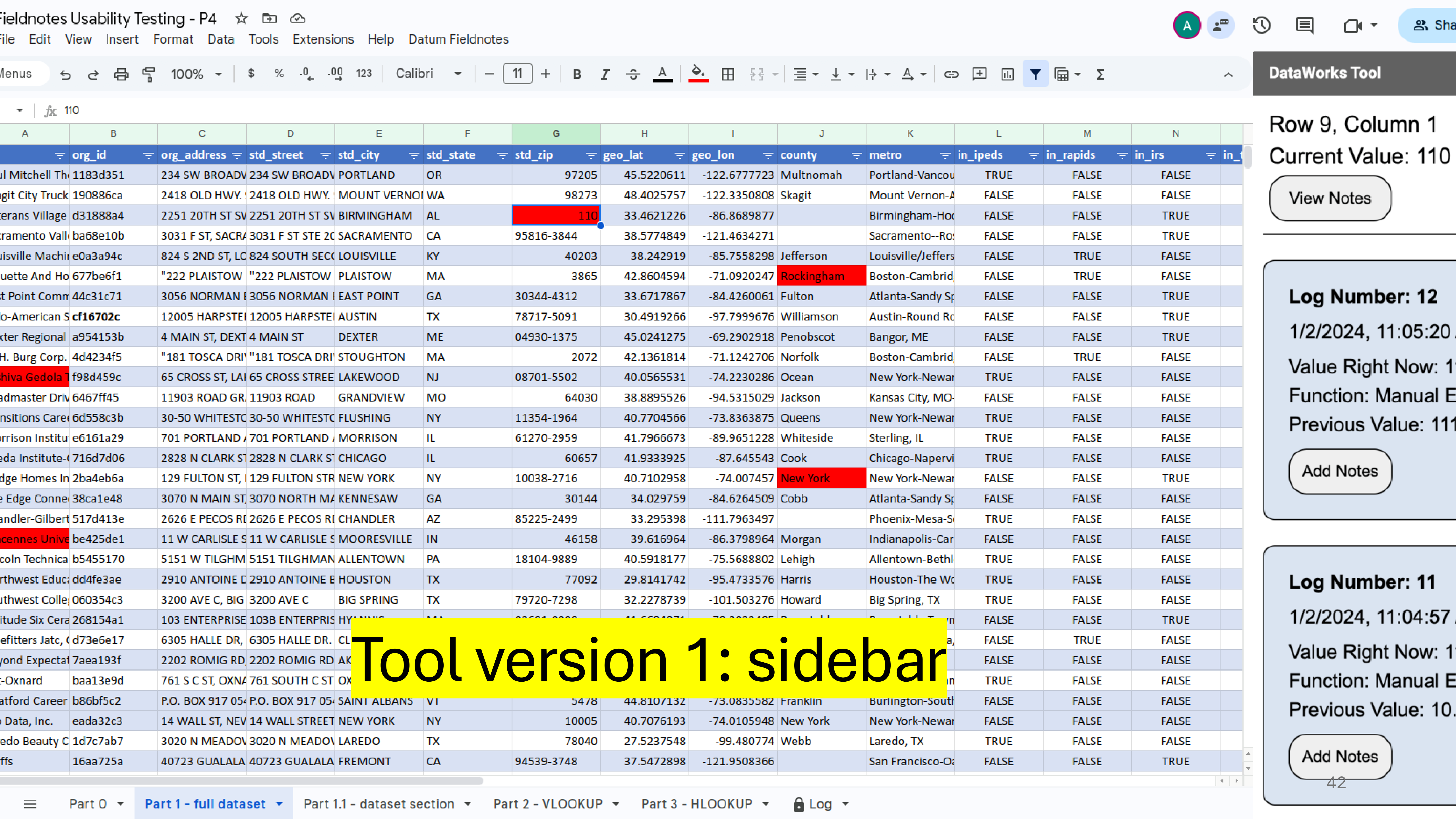
November 7, 1:45PM
Annabel Rothschild

November 7, 1:42PM
Annabel Rothschild

November 7, 1:37PM

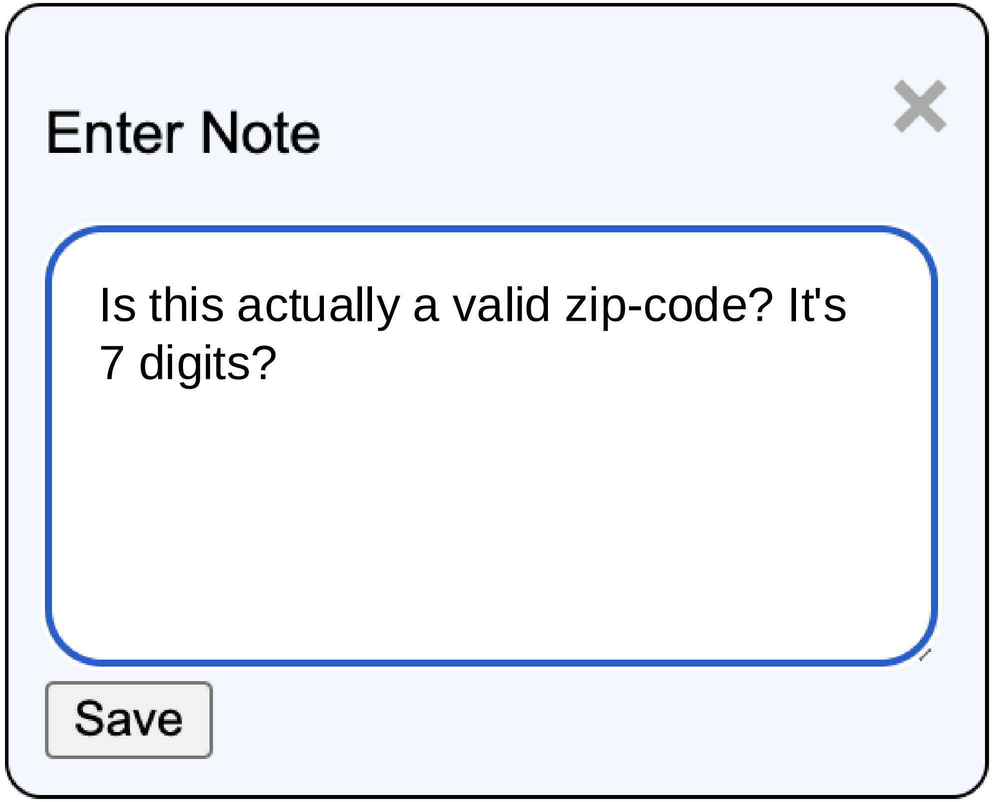
41

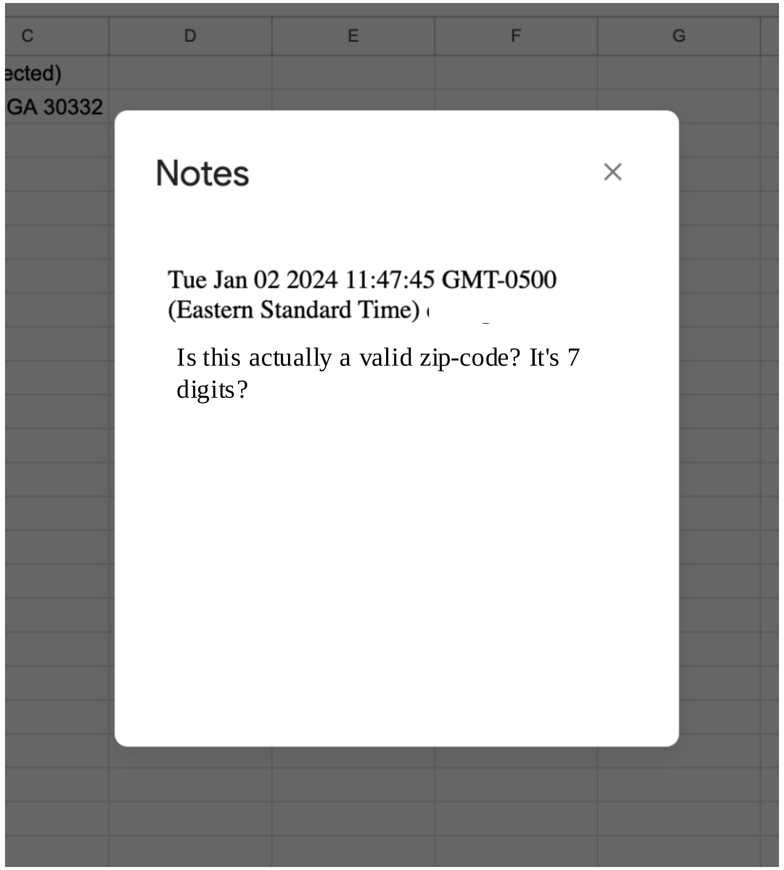
Show changes




| | org_id | org_address | std_street | std_city | std_state | std_zip | geo_lat | geo_lon | county | metro | in_ipeds | in_rapids | in_irs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ul Mitchell Th | 1183d351 | 234 SW BROADV | 234 SW BROADV | PORTLAND | OR | 97205 | 45.5220611 | -122.6777723 | Multnomah | Portland-Vancou | TRUE | FALSE | FALSE |
| agit City Truck | 190886ca | 2418 OLD HWY. | 2418 OLD HWY. | MOUNT VERNOI | WA | 98273 | 48.4025757 | -122.3350808 | Skagit | Mount Vernon-A | FALSE | FALSE | FALSE |
| erans Village | d31888a4 | 2251 20TH ST SV | 2251 20TH ST SV | BIRMINGHAM | AL | 110 | 33.4621226 | -86.8689877 | | Birmingham-Hoo | FALSE | FALSE | TRUE |
| cramento Vall | ba68e10b | 3031 F ST, SACR/ | 3031 F ST STE 2C | SACRAMENTO | CA | 95816-3844 | 38.5774849 | -121.4634271 | | Sacramento--Ro | FALSE | FALSE | TRUE |
| uisville Machi | e0a3a94c | 824 S 2ND ST, LC | 824 SOUTH SECC | LOUISVILLE | KY | 40203 | 38.242919 | -85.7558298 | Jefferson | Louisville/Jeffers | FALSE | TRUE | FALSE |
| uette And Ho | 677be6f1 | "222 PLAISTOW | "222 PLAISTOW | PLAISTOW | MA | 3865 | 42.8604594 | -71.0920247 | Rockingham | Boston-Cambrid | FALSE | TRUE | FALSE |
| st Point Comm | 44c31c71 | 3056 NORMAN I | 3056 NORMAN I | EAST POINT | GA | 30344-4312 | 33.6717867 | -84.4260061 | Fulton | Atlanta-Sandy Sp | FALSE | FALSE | TRUE |
| o-American S | cf16702c | 12005 HARPSTEI | 12005 HARPSTEI | AUSTIN | TX | 78717-5091 | 30.4919266 | -97.7999676 | Williamson | Austin-Round Rc | FALSE | FALSE | TRUE |
| xter Regional | a954153b | 4 MAIN ST, DEXT | 4 MAIN ST | DEXTER | ME | 04930-1375 | 45.0241275 | -69.2902918 | Penobscot | Bangor, ME | FALSE | FALSE | TRUE |
| H. Burg Corp. | 4d4234f5 | "181 TOSCA DRI | "181 TOSCA DRI | STOUGHTON | MA | 2072 | 42.1361814 | -71.1242706 | Norfolk | Boston-Cambrid | FALSE | TRUE | FALSE |
| hiva Gedola T | f98d459c | 65 CROSS ST, LAI | 65 CROSS STREE | LAKEWOOD | NJ | 08701-5502 | 40.0565531 | -74.2230286 | Ocean | New York-Newar | TRUE | FALSE | FALSE |
| admaster Driv | 6467ff45 | 11903 ROAD GR. | 11903 ROAD | GRANDVIEW | MO | 64030 | 38.8895526 | -94.5315029 | Jackson | Kansas City, MO- | FALSE | FALSE | FALSE |
| nsitions Care | 6d558c3b | 30-50 WHITESTC | 30-50 WHITESTC | FLUSHING | NY | 11354-1964 | 40.7704566 | -73.8363875 | Queens | New York-Newar | TRUE | FALSE | FALSE |
| orrison Institu | e6161a29 | 701 PORTLAND / | 701 PORTLAND / | MORRISON | IL | 61270-2959 | 41.7966673 | -89.9651228 | Whiteside | Sterling, IL | TRUE | FALSE | FALSE |
| eda Institute-( | 716d7d06 | 2828 N CLARK ST | 2828 N CLARK ST | CHICAGO | IL | 60657 | 41.9333925 | -87.645543 | Cook | Chicago-Napervi | TRUE | FALSE | FALSE |
| dge Homes In | 2ba4eb6a | 129 FULTON ST, | 129 FULTON STR | NEW YORK | NY | 10038-2716 | 40.7102958 | -74.007457 | New York | New York-Newar | FALSE | FALSE | TRUE |
| e Edge Conne | 38ca1e48 | 3070 N MAIN ST, | 3070 NORTH MA | KENNESAW | GA | 30144 | 34.029759 | -84.6264509 | Cobb | Atlanta-Sandy Sp | FALSE | FALSE | FALSE |
| andler-Gilbert | 517d413e | 2626 E PECOS RI | 2626 E PECOS RI | CHANDLER | AZ | 85225-2499 | 33.295398 | -111.7963497 | | Phoenix-Mesa-S | TRUE | FALSE | FALSE |
| ncennes Unive | be425de1 | 11 W CARLISLE S | 11 W CARLISLE S | MOORESVILLE | IN | 46158 | 39.616964 | -86.3798964 | Morgan | Indianapolis-Car | FALSE | FALSE | FALSE |
| coln Technica | b5455170 | 5151 W TILGHM | 5151 TILGHMAN | ALLENTOWN | PA | 18104-9889 | 40.5918177 | -75.5688802 | Lehigh | Allentown-Bethl | TRUE | FALSE | FALSE |
| rthwest Educ | dd4fe3ae | 2910 ANTOINE D | 2910 ANTOINE D | HOUSTON | TX | 77092 | 29.8141742 | -95.4733576 | Harris | Houston-The Wc | TRUE | FALSE | FALSE |
| uthwest Colle | 060354c3 | 3200 AVE C, BIG | 3200 AVE C | BIG SPRING | TX | 79720-7298 | 32.2278739 | -101.503276 | Howard | Big Spring, TX | TRUE | FALSE | FALSE |
| itude Six Cera | 268154a1 | 103 ENTERPRISE | 103B ENTERPRIS | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | FALSE | FALSE | TRUE |
| efitters Jatc, ( | d73e6e17 | 6305 HALLE DR, | 6305 HALLE DR. | CL | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | FALSE | TRUE | FALSE |
| yond Expecta | 7aea193f | 2202 ROMIG RD | 2202 ROMIG RD | AK | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | FALSE | FALSE | FALSE |
| -Oxnard | baa13e9d | 761 S C ST, OXN/ | 761 SOUTH C ST | OX | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | TRUE | FALSE | FALSE |
| atford Career | b86bf5c2 | P.O. BOX 917 05 | P.O. BOX 917 05 | SAINT ALBANS | VT | 5478 | 44.8107132 | -73.0835582 | Franklin | Burlington-South | FALSE | FALSE | FALSE |
| Data, Inc. | eada32c3 | 14 WALL ST, NEV | 14 WALL STREET | NEW YORK | NY | 10005 | 40.7076193 | -74.0105948 | New York | New York-Newar | FALSE | FALSE | FALSE |
| edo Beauty C | 1d7c7ab7 | 3020 N MEADO\ | 3020 N MEADO\ | LAREDO | TX | 78040 | 27.5237548 | -99.480774 | Webb | Laredo, TX | TRUE | FALSE | FALSE |
| ffs | 16aa725a | 40723 GUALALA | 40723 GUALALA | FREMONT | CA | 94539-3748 | 37.5472898 | -121.9508366 | | San Francisco-O | FALSE | FALSE | TRUE |

# Tool Version 1: Add Notes



| Timestamp | Cell | Sheet | Previous Val | New Value | Formula | User | Notes |
|---|---|---|---|---|---|---|---|
| 11/16/2023 11:00:56 | A13 | Examples | | change | Manual Entry | | |
| 11/16/2023 11:01:09 | A13 | Examples | change | boo | Manual Entry | | Thu Nov 16 2023 11:01:33 G boo |
| 11/16/2023 11:02:25 | A13 | Examples | boo | bool | Manual Entry | | |
| 11/16/2023 11:03:01 | A13 | Examples | bool | hi | Manual Entry | | |
| 11/16/2023 11:03:56 | A13 | Examples | hi | fff | Manual Entry | | |
| 11/16/2023 11:05:13 | A13 | Examples | fff | | Manual Entry | | |
| 11/16/2023 11:06:34 | A13 | Examples | | add | Manual Entry | | |
| 1/2/2024 10:09:43 | A1 | Sheet4 | | hi | Manual Entry | | |
| 1/2/2024 10:10:04 | A1 | Sheet4 | hi | 1 | Manual Entry | | Tue Jan 02 2024 10:10:27 GI We shifted from text to using |
| 1/2/2024 11:04:58 | A9 | Examples | 10 | 111 | Manual Entry | | |
| 1/2/2024 11:05:21 | A9 | Examples | 111 | 110 | Manual Entry | | |

Tool version 1: Log

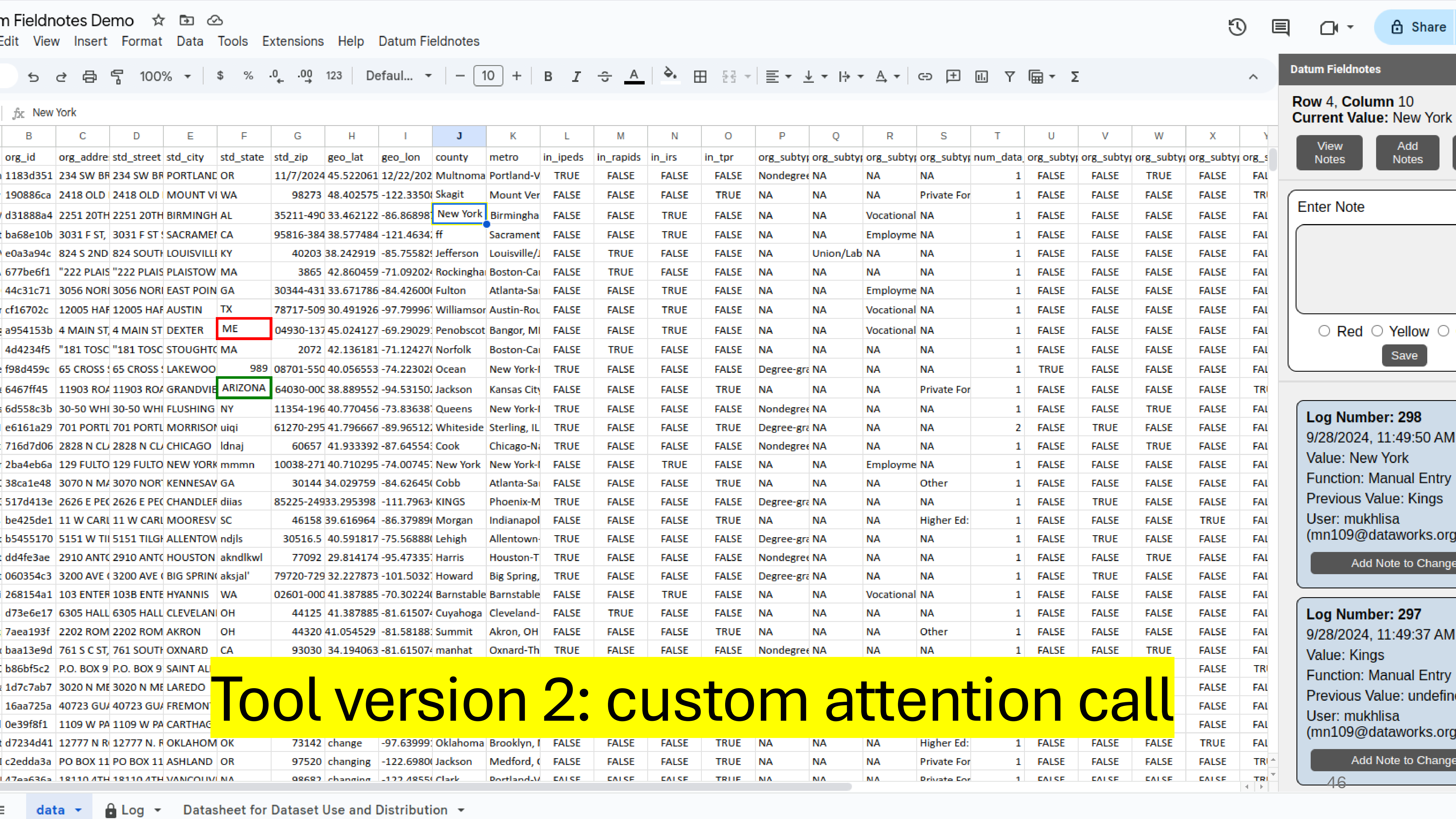
## Usability testing: Follow the Data protocol

**12 interviewees**: 7 Data Fellows, 5 civic and non-profit data workers from around the United States

**Follow the Data**[1] **protocol**: use LLM to generate synthetic dataset, researchers markup as if colleagues, then ask participants to reconstruct a calculation or trace a concern. End with D4D questions.

**Major findings**: color coding (as visual indicators), need for localized definitions for D4D[2] terminology, additional security and privacy features.

1. Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2012). Follow the data: How astronomers use and reuse data. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–3. https://doi.org/10.1002/meet.14504901341
2. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM 64, 12: 86–92. https://doi.org/10.1145/3458723




Tool version 2: custom attention call

m Fieldnotes Demo

Edit View Insert Format Data Tools Extensions Help Datum Fieldnotes

Share

100% $ % .0 .00 123 Arial 12

Datum Fieldnotes

Row 1, Column 1
Current Value: Timestamp

View Notes | Add Notes

No session history

fx Datasheet for Dataset Use and Distribution

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| eet for Dataset Use and Distribution | | | | | |
| contents | | | | | |
| a Workers | 2 - Datasheets for Datasets questions | | | | |
| mation | Motivation | | | | |
| er Reflections | Composition | | | | |
| | Collection process | | | | |
| | Preprocessing/cleaning/labeling | | | | |
| | Uses | | | | |
| | Distribution | | | | |
| | Maintenance | | | | |
| ns for Data Workers | | | | | |
| | Basic Information | | | | |
| this dataset created? | | | | | |
| worked on this dataset? | mukhlisa (mn109@data.org), arothschild@gatech.edu, Annabel (annabel@dataworks.hq.com) | | | | |
| ld you contact if you have questions about this dataset? | | | | | |
| data use contract or agreement that someone accessing this dataset must ? Where can potential dataset users access that contract? | | | | | |
| | Data Worker Reflections | | | | |
| any ways you feel this dataset should not be used? | | | | | |
| otice any interesting patterns or surprising entries in this dataset? | | | | | |
| ncounter difficulty working on this dataset? What kinds? | | | | | |
| eets for Datasets Questions | | | | | |
| | Motivation | | | | |
| purpose was the dataset created? Was there a specific task in mind? Was ecific gap that needed to be filled? Please provide a description. | | | | | |
| ed the dataset (e.g., which team, research group) and on behalf of which ., company, institution, organization)? | | | | | |
| ed the creation of the dataset? If there is an associated grant, please provide of the grantor and the grant name and number. | | | | | |

Tool version 2: robust D4D

47 Privacy

data | Log | Datasheet for Dataset Use and Distribution

**Five participants** (Data Fellows)
*How do you talk about data?*

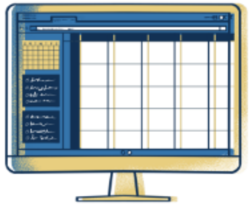
**Findings:**

- "Pre-processing" collapses several distinct data labor types into one
- Dealing with harmful or anxiety-provoking content cannot be made universal

# Additional Suite of Learning Resources

## Datum Fieldnotes

Datum Fieldnotes is an open-source Google Sheets extension that transforms how you work with data. It empowers you to track changes, add context-rich notes, and document your dataset's purpose and usage, all within the familiar Google Sheets environment. By providing a comprehensive view of your data's evolution, Datum Fieldnotes fosters transparency, encourages responsible data practices, and enhances collaboration among team members.



Privacy Policy | Terms of Service | Limited Use Policy | Developer Contact Information

### User Guide

**Tool Setup**
Getting started with Datum Fieldnotes is quick and easy! Follow these simple steps to install the extension and set up your spreadsheet for enhanced data documentation.

**Getting Started**
Now that you have Datum Fieldnotes set up, let's explore how to use its core features to track changes, add notes, and document your dataset effectively.

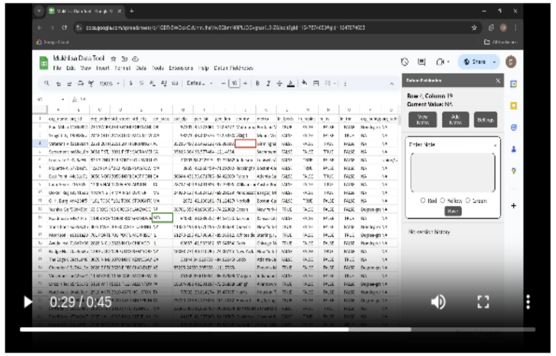
**Exploring the Tool**
Datum Fieldnotes offers advanced features to help you analyze your data's evolution and ensure its quality. Let's delve into these capabilities.

### FAQ

- What is Datum Fieldnotes and what is its purpose?
- Is Datum Fieldnotes a standalone tool or an extension only?
- Can I use Datum Fieldnotes offline?
- Is there a way to export or backup my notes and change log data?
- Does the tool support multi-user collaboration on the same sheet or workbook?
- What is the datasheet for dataset use and distribution?
- Where are the notes and change history data stored? Is it within the

## Exploring/Querying the Log

Aside from the tools sidebar, Datum Fieldnotes Log sheet allows for further data analysis and understanding of your tool. Here are some examples in ways you can use the Log Sheet of the tool to further understand your data.




**Accessing the Log sheet**

Learn how to quickly locate and open the Datum Fieldnotes log sheet, your central hub for tracking and understanding changes in your spreadsheet.

**Filtering the Log sheet**

Discover how to use powerful filters to zero in on specific information within your change log, whether you're looking for edits by a certain user, within a



See more: https://dataworkforce.gatech.edu/datum-fieldnotes/

# Human-AI collaboration around the conference table

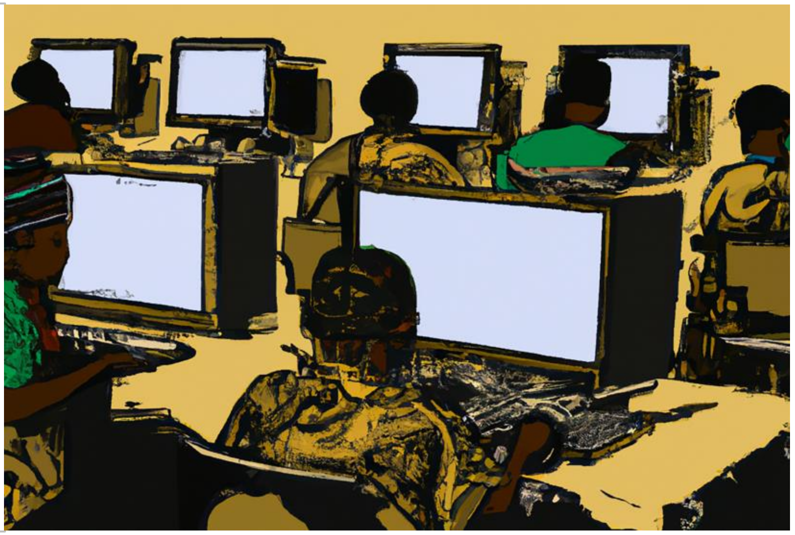
# The Content Moderation Problem

- We can ask data workers about how they performed a task, but we can't do the same for computational (AI) approaches...
- However, human-AI collaboration can offload some of the difficulties of data annotation.

BUSINESS • TECHNOLOGY

## Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



# What are the costs of data annotation across different methods?


- Paying a fair wage
- Opportunities for professional growth
- Financial cost

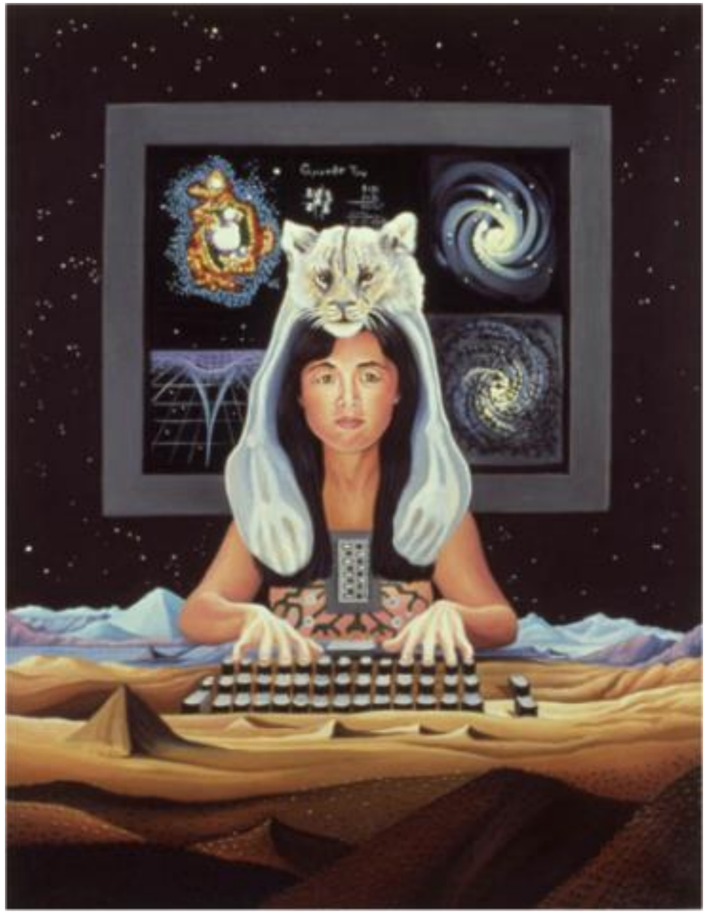
amazon mechanical turk

- Paying a fair wage
- Constructing tasks correctly to ensure feedback mechanisms
- Creating labor conditions that are pro-social

ChatGPT

- Environmental harms
- Auditing for perspective
- How to balance need for human review

Grace Kim, Annabel Rothschild, Carl DiSalvo, and Betsy DiSalvo. 2024. “What’s Your Stake in Sustainability of AI?: An Informed Insider’s Guide”. AIES (Conference on Artificial Intelligence, Ethics, and Society).



To understand the ethics behind a decision, we frequently interview decision-makers and interrogate the surrounding socio-technical environment. But, how can you understand the standpoint of 'someone' from nowhere?

Building safe datasets requires partnering with the people who know and can best contextualize dataset entries in the wider socio-technical world.

**My work:**
(1) Partnering with data workers to channel their observations and feedbacks into dataset audits, resulting in safe datasets.
(2) Building tools and processes for the formalization of dataset audits by data workers.

# The value of safe datasets?

*Priceless!*

**DataWorks Fellows** past and present, especially Dana Priest, Justin Booker, and Christa Davoll.

**Advisors** (Drs. Betsy DiSalvo, Carl DiSalvo), **collaborators** (Lara Schenk, Dr. Ben Rydal Shapiro), **dissertation committee** (Drs. Lauren Klein, Ding Wang, Richmond Wong, Ellen Zegura, Shaowen Bardzell).

**Student collaborators and mentees:** Will Eickman, Grace Kim, Mukhlisa Nematova.

**Financial support from:** NSF, Google, Mellon Foundation, Kapor Foundation Dissertation Fellowship.

*Special thanks to Catherine Wieczorek for design assistance.*

## Link to slides, citations and further reading

OR
(QR code and
Bit.ly point to
same link)

https://bit.ly/4ibwML5