# An Investigation of Experiences Engaging the Margins in Data-Centric Innovation

**Gabriella Thompson,**[1] **Ebtesam Al Haque,**[2] **Paulette Blanc,**[3]
**Meme Styles,**[3] **Denae Ford,**[4] **Angela D. R. Smith,**[1] **Brittany Johnson**[2]
[1] University of Texas at Austin
gabriella.thompson@utexas.edu, angela.smith@ischool.utexas.edu
[2] George Mason University
ehaque4@gmu.edu, johnsonb@gmu.edu
[3] Measure
paulette@wemeasure.org, meme@wemeasure.org
[4] Microsoft Research
denae@microsoft.com

**Abstract**

Data-centric technologies provide exciting opportunities, but recent research has shown how lack of representation in datasets, often as a result of systemic inequities and socioeconomic disparities, can produce inequitable outcomes that can exclude or harm certain demographics. In this paper, we discuss preliminary insights from an ongoing effort aimed at better understanding barriers to equitable data-centric innovation. We report findings from a survey of 261 technologists and researchers who use data in their work regarding their experiences seeking adequate, representative datasets. Our findings suggest that age and identity play a significant role in the seeking and selection of representative datasets, warranting further investigation into these aspects of data-centric research and development.

## Introduction & Background

Data is at the center of modern research and development, often providing generalized insights into people and phenomena (Wu et al. 2013). While there is potential in the power of data to offer personalized benefits to broader society, this power is diminished when there are gaps in the data being used. Data-centric innovations in technology continue to demonstrate these gaps through their failure to equally support all users (Miller 2020). Several studies on algorithmic fairness suggest representation bias in training data is a contributing factor to the high error rates for users with historically marginalized identities (Buolamwini and Gebru 2018; Markl 2022; Baack 2024; Chen, Johansson, and Sontag 2018; Lin et al. 2020; Asudeh, Jin, and Jagadish 2019). Attempts to diversify datasets often promote inclusive design methods, such as expanding the possibilities for self-identification on demographic forms (Bivens and Haimson 2016; Slade et al. 2021), or propose novel methods for diverse data collection (Stasaski, Yang, and Hearst 2020; Lin et al. 2020), but there is a gap in scholarship on the experiences of seeking diverse data from the technologists' perspective. Uncovering the common successes and limitations they encounter may illuminate the barriers to equitable data-centric research and development. In this paper, we report on preliminary findings from a survey of technologists regarding their experiences engaging in data-centric work.



## Methodology

To understand our respondents' experiences seeking data, we developed three research questions that guided our survey design and analysis: *What are the factors that impact technologists' decision to use a dataset?* (**RQ1**), *What are the challenges and barriers to finding diverse or representative datasets?* (**RQ2**), and *What methods do technologists use to find relevant and representative datasets?* (**RQ3**).

**Survey Design & Dissemination.** We developed our survey as part of a larger effort to understand the barriers of engaging marginalized groups in data-centric computing research and development. The survey is designed for two audiences: technology users who identify as Black, Indigenous, or People of Color (BIPOC), and technologists and researchers who use data in their work. Our survey consists of four sections: *Contributing Data*, *Seeking Data*, *Collecting Data - Research*, and *Demographics*. In this paper, we focus on the technologists and researchers who completed the Seeking Data section. The questions in the Seeking Data section centered on methods used to search for data, how they make decisions about the data they will use, and their experiences finding adequate data. We also included an attention question for all participants to ensure they were giving the survey due consideration. Our full survey is publicly available for reuse and replication. [1] We administered our survey using Qualtrics. [2] To recruit respondents, we advertised on social media (e.g., X, formerly known as Twitter, LinkedIn) and in our professional networks. From these efforts, we acquired over 900 survey responses. Because we administered the survey online (Griffin et al. 2021), we proceeded to clean the data of any invalid responses.

**Data Preparation & Cleaning.** To ensure validity of our data, we first removed incomplete responses and those

[1] https://inspired-gmu.github.io/engaging-margins/#goals
[2] https://www.qualtrics.com/

that took less than 3 minutes to finish. We then filtered out duplicate responses (based on email) and irrelevant open ended responses. To prepare our data for analysis, we combined the age responses into two new categories: *Under 35* (135 responses) and *35 and Over* (124 responses). This decision was informed by insights from the StackOverflow Developer Survey [3], which found that 43% of professional developers are within the ages of 24–35. We also removed respondents out of scope for our analysis.

**Respondents.** From our cleaned dataset, we found most technologists work in industry, either in a technical (196) or research role (27), followed by academia (27). We had a handful of respondents from other occupations, such as healthcare (4) and childcare (5). Of the 261 respondents in our sample, 135 are between the ages of 18–34 years old, while 124 are between 35-84. The majority of respondents in the 35–84 age group are under 64 years old (94%). The majority of our respondents (249) identified as a Person of Color (POC). A handful of respondents seek data for technology development alone (37), but the majority in our sample seek data for research (118) or for both.

**Data Analysis.** To answer **RQ1** and **RQ2**, we focused on three independent variables (Data Use purposes, Age, and POC Declaration) across all of our statistical tests. To determine the factors that impact the decision to use a dataset (**RQ1**), we identified five factors (Cost, Diversity, Trust in the Source, and Amount of Data in the Source) and compared responses to relevant survey questions against the independent variables to determine association. We identified the challenges and barriers to finding diverse data (**RQ2**) by analyzing their responses to the questions regarding how often they do not find adequate data, how easily they find trustworthy resources for data, and factors that impact their ability to find diverse datasets. Lastly, we analyzed their methods for finding relevant and representative datasets (**RQ3**) by comparing responses to questions regarding their ability to find adequate data and trustworthy data to their methods for finding data (where they start their search and where they have the most success).

We used Python and the Pandas package (pandas development team 2024) for the majority of our data analysis. We conducted Chi-square tests of independence to determine association for comparisons that involved two categorical variables. Building on methods from prior work (Sharpe 2015), we conducted post-hoc testing for the Chi-Square tests with contingency tables larger than 2x2 to determine which relationships were driving the significance. We then used the Researchpy package (Bryant 2018) and information from Peter Statistics (Statistics n.d.) to calculate Cramer's V for each significant Chi-Square test to determine the strength of the association.

For the tests that did not satisfy the requirements of the Chi-Square test, we conducted a two-tailed Fisher's Exact Test using the Scipy package (Virtanen et al. 2020) to determine significance. If the contingency table was larger than 2x2, we utilized the Fisher-Freeman-Halton Exact Test. Lastly, we performed a Mann-Whitney U test to compare against ordinal dependent variables. For tests that compared more than two ordinal dependent variables, we used a Kruskal-Wallis test to determine if there was a significant difference in means.

[3]https://survey.stackoverflow.co/2023/

## Findings

In this section, we describe the findings from our survey relative to our three research questions.

### Dataset Decision Factors (RQ1)

We identified cost, diversity, dataset structure, trust, and size of the dataset as influential factors in the decision to use a dataset. We compared each factor to three independent variables: Data Use Purposes, Age Group, and POC. Here, we discuss the significant results from the statistical tests we conducted.

**Cost.** We found a strong correlation between cost and data use purposes ($p = 0.0087$). Our post-hoc testing revealed a significant relationship between collecting data for technology development and not paying for data ($p = 0.015$). We affirmed the strength of this association by calculating the Cramer's V, which indicated that there was a small but significant association ($V = 0.1906$) between the two variables. We did not find any significant relationships from the remainder of our tests.

**Diversity.** Our analyses using the Mann-Whitney U Test identified an association between the importance of data diversity and age group ($U = 10969$, $p = 1.2651e-6$). We assigned a rank for the four answer choices in the relevant question (Not important at all = 1; Not very important = 2, Kind of important = 3, and Very important = 4). We conducted a one-sided Mann-Whitney U test to infer the direction of the significance and found a significant p-value for the 'Under 35' age group ($p = 6.325e-7$). The mean for the 'Under 35' age group was 3.759, and the mean for the '35 and over' was 3.403. These results indicate that respondents under 35 may be more likely to rate diversity as more important to their work. We did not find any significant results from the remainder of our tests.

**Trust.** This Chi-square test indicated a significant relationship between respondents who felt trust contributed the most to their decision to use a dataset and their age ($p = 1.4136e-05$) We compared the expected and observed counts of the Chi-square test and found that respondents over 35 relied more on trust as a factor in their decision than older technologists. The strength of this association was small but significant according to the Cramer's V ($V = 0.2767$). We also found a significant relationship between identifying as a POC and considering trust as a factor in using a dataset. Due to the small sample size of non-POC respondents, we conducted a Fisher's Exact test ($p = 0.0207, OR = 4.909$). The odds ratio indicates that the odds of considering trust in the decision to use a source for

POC was 4.9 times higher than that of non-POC. We failed to find a significant relationship for the rest of our tests.

**Amount of data.** We identified an association between the amount of data in a dataset contributing to the decision to use it and age group. A Chi-square test of independence revealed a significant relationship between the two factors ($p = 0.0104$). Upon comparison of the expected and observed counts of the Chi-square test, we found that respondents under the age of 35 were more likely to consider amount of data in the decision to use a dataset. The strength of this association was small but significant ($V = 0.1662$). No other tests from this factor were significant.

### Experiences Finding Diverse Datasets (RQ2)

To find the challenges and barriers that affect respondents' ability to find diverse data, we asked questions regarding factors they believe impact their ability to find diverse datasets, the frequency with which they are able to find adequate data, and their experiences finding appropriate data. We compared the responses to the independent variables: data use goals, age, and whether they identify as a POC. In this section, we will report the significant relationships from our tests.

**Factors that impact ability to find diverse sources.** We identified a significant relationship between Q55 and age group. We conducted a Chi-square test of independence for the 3x3 contingency table and identified a significant relationship between factors selected and age ($p = 0.8297$). We performed post-hoc testing to ensure the validity of results by calculating the adjusted residuals and correcting the significance level using a Bonferroni corrected alpha. From this analysis, we found that two factors have significant associations with age: **resources** (e.g, money, data sources) and **tooling** (e.g., language support). A comparison of expected and observed counts found that respondents under the age of 35 felt resources ($p = 7.78e-07$, adjusted residual = 5.27) impacted their ability to find diverse datasets more than tooling ($p = 2.32e-5$, adjusted residual = $-4.61$), whereas those over the age of 35 felt tooling impacted their ability more. Cramer's V indicated a medium strength association ($V = 0.3324$). Analysis of the other independent variables did not reveal any significant associations.

**Difficulty finding trustworthy sources.** We conducted a Mann-Whitney U Test to determine any associations between our independent variables and the ability to find adequate, representative datasets. We identified a significant association between difficulty and identifying as a POC ($U = 478.0$, $p = 0.0005$). This significant relationship encouraged us to evaluate the direction through a one-sided Mann-Whitney U test. From this test, we determined that POC respondents found it more difficult to find trustworthy sources than non-POC respondents ($p = 0.0002$). The mean difficulty score (on a Likert scale 1–5, Extremely Easy to Extremely Difficult) for POC was 3.1325, while the mean for non-POC was 2.0. We did not find significant relationships from the remainder of our tests.

### Finding relevant and representative datasets (RQ3)

To better understand differences in experiences finding relevant and representative datasets, we compared the ability to find adequate and trustworthy sources for data with the first and most successful methods respondents use to find them. In this section, we will elaborate on the significant results from our tests of association.

**Difficulty finding adequate data.** Our analyses indicated a relationship between frequency not finding adequate data and beginning with a general web search ($U = 5563$,$p = 0.0061$). Given the relationship, we performed a one-sided Mann-Whitney U Test to infer the direction. Our results indicated that using a web search first was associated with less difficulty finding adequate data ($p = 0.003$). The mean difficulty for beginning with a web search was 1.608, while the mean difficulty for respondents who did not begin with a web search was 1.813. The rest of our comparisons did not find any significant relationships.

## Discussion

Our findings thus far provide valuable insights for advancing our efforts and others interested in the role of technologists and researchers in equitable data-centric innovation.

**The Role of Expertise in Data Seeking.** Expertise plays a significant role in technologists' processes and strategies (LaToza et al. 2020). Prior research indicates significant relationships between expertise in computing and age, emphasizing the heightened expertise found in older adults (Arning and Ziefle 2008). Our findings suggest a relationship between age and the methods and considerations involved in seeking and using data for innovation, including the willingness to explore smaller datasets which studies have shown may be the case for datasets centered on historically marginalized groups (Philip, Schuler-Brown, and Way 2013; Warren et al. 2022). We will use these insights to delve deeper into the role of expertise in data seeking behaviors, which can lead to broader, actionable insights for improving practice.

**Understanding Identity as a Factor in Data Seeking.** Prior studies have provided insights into the role of identity and positionality in computing research and innovation (Schwarz and Watson 2005; Scheuerman and Brubaker 2024; Secules et al. 2021). We found that the role of identity in innovation may extend to the intentionality behind and challenges with finding diverse datasets. Our findings suggest racial disparities in data trust and accessibility, where POC reported greater difficulty finding trustworthy data sources. This points back to persistent systemic inequalities and underscores the importance of addressing issues of representation and bias in data collection and use. In our efforts to better understand experiences and strategies involved in the usage of representative datasets, we will ensure that we are engaging with BIPOC technologists in research and development to better understand how we can more effectively support the use of representative datasets in practice.

## References


Arning, K.; and Ziefle, M. 2008. Development and validation of a computer expertise questionnaire for older adults. *Behaviour & Information Technology*, 27(1): 89–93.

Asudeh, A.; Jin, Z.; and Jagadish, H. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 554–565. IEEE.

Baack, S. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2199–2208. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Bivens, R.; and Haimson, O. L. 2016. Baking gender into social media design: How platforms shape categories for users and advertisers. *Social Media+ Society*, 2(4): 2056305116672486.

Bryant, C. 2018. Researchpy.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Griffin, M.; Martino, R. J.; LoSchiavo, C.; Comer-Carruthers, C.; Krause, K. D.; Stults, C. B.; and Halkitis, P. N. 2021. Ensuring survey research data integrity in the era of internet bots. *Quality & quantity*, 1–12.

LaToza, T. D.; Arab, M.; Loksa, D.; and Ko, A. J. 2020. Explicit programming strategies. *Empirical Software Engineering*, 25(4): 2416–2449.

Lin, Y.; Guan, Y.; Asudeh, A.; and Jagadish, H. 2020. Identifying insufficient data coverage in databases with multiple relations. *Proceedings of the VLDB Endowment*, 13(11).

Markl, N. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 521–534. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Miller, K. 2020. A matter of perspective: Discrimination, bias, and inequality in ai. In *Legal regulations, implications, and issues surrounding digital data*, 182–202. IGI Global.

pandas development team, T. 2024. pandas-dev/pandas: Pandas.

Philip, T. M.; Schuler-Brown, S.; and Way, W. 2013. A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning*, 18: 103–120.

Scheuerman, M. K.; and Brubaker, J. R. 2024. Products of positionality: How tech workers shape identity concepts in computer vision. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Schwarz, G. M.; and Watson, B. M. 2005. The influence of perceptions of social identity on information technology-enabled change. *Group & Organization Management*, 30(3): 289–318.

Secules, S.; McCall, C.; Mejia, J. A.; Beebe, C.; Masters, A. S.; L. Sánchez-Peña, M.; and Svyantek, M. 2021. Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education*, 110(1): 19–43.

Sharpe, D. 2015. Your chi-square test is statistically significant: now what?. *Practical assessment, research & evaluation*, 20(8): n8.

Slade, T.; Gross, D. P.; Niwa, L.; McKillop, A. B.; and Guptill, C. 2021. Sex and gender demographic questions: improving methodological quality, inclusivity, and ethical administration. *International Journal of Social Research Methodology*, 24(6): 727–738.

Stasaski, K.; Yang, G. H.; and Hearst, M. A. 2020. More Diverse Dialogue Datasets via Diversity-Informed Data Collection. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4958–4968. Online: Association for Computational Linguistics.

Statistics, P. n.d. Nominal vs. Nominal - Part 3b: Post-hoc test — peterstatistics.com.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.

Warren, C. M.; Brown, E.; Wang, J.; and Matsui, E. C. 2022. Increasing representation of historically marginalized populations in allergy, asthma, and immunologic research studies: challenges and opportunities. *The Journal of Allergy and Clinical Immunology: In Practice*, 10(4): 929–935.

Wu, X.; Zhu, X.; Wu, G.-Q.; and Ding, W. 2013. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1): 97–107.


## Acknowledgments


This work is supported by the National Science Foundation (NSF) under grant #2224674 and #2224675.