accelerator

RESEARCHER PAIN POINTS SURVEY

Introduction

The purpose of this survey, conducted in the summer of 2024, was to gain a comprehensive understanding of the challenges researchers face when accessing and analyzing large datasets related to human behavior in online environments.

As research in the field rapidly evolves, it has become clear that scholars and professionals encounter a variety of barriers that hinder their ability to conduct high-quality research. This initiative sought to gather firsthand insights into these impediments from the global research community, with the aim of identifying key areas where support, resources, or tools could make a significant impact.

To promote broad and diverse participation, the survey was distributed widely through a global network of scholars, professionals, and research groups. Using the Accelerator's network of connections, experts from a variety of fields related to digital information environments and from various countries were reached. This broad dissemination aimed to capture a wide spectrum of experiences, from well-established researchers to early-career professionals.

Ultimately, the goal was to use the results of this survey to inform the development of the Accelerator's infrastructure and tools, ensuring that they are capable of addressing the most pressing challenges currently faced by researchers. Findings can also guide future efforts to improve access to data and analytical resources more broadly.

While we can't be sure how many researchers saw the call, overall, we received 108 responses, primarily from academic researchers (71%) and from the United States (36%). North America (39%) and Europe (38%) are primarily represented, but respondents come from a total of 23 different countries.

Findings suggest the need for better infrastructure and support for digital information systems worldwide. They show that initiatives like the Accelerator could contribute immensely towards overcoming critical barriers to research progress.

A striking finding from our survey was that almost 60% of respondents had abandoned a research project in the past year due to the constraints mentioned in the following sections. Among those who managed to overcome such constraints, 48% cooperated with colleagues with expertise from other institutions, 44% gained researcher access via APIs, and 30% purchased the data using their own research funds – a solution not available to many researchers. The reasons behind project abandonment and how these issues differ across platforms, disciplines, and other factors require further investigation.

More than half of the respondents reported that while working on projects involving large amounts of data on online human behavior, they spend at least 20% of their time on tasks that initiatives like the Accelerator aim to support. These tasks include obtaining and pre-processing/cleaning data, ensuring compliance with IRB/ethics boards and other regulations, and building data infrastructure, all while working on projects involving large-scale online human behavior data. The significant time commitment to these tasks highlights the critical need for efforts that can alleviate some of this burden for researchers.

Additionally, about 75% of our respondents did not apply to any of the major platform (researcher) data access programs (including access to APIs or whitelisted data scraping programs) in the last year.

Constraints – Pain Points

The study asks a series of questions about major constraints and pain points when researchers and professionals try to access and work with large amounts of data on online human behavior. Lack of access comes out as the most important constraint across all respondents.

Figure 1 shows the average Likert value per response category across all respondents. It shows that lack of access is the most important constraint, followed by cost and opportunity cost. The lack of normal datasets and the lack of knowledge about the availability of data are also strong impeding factors for this group of respondents. Figure 2 shows the same constraints per type of respondent (Academia, Civil Society and Other) that follow similar patterns.

As mentioned above, about 75% of our respondents did not apply to any of the main (researcher) data access programs (such as access to APIs or whitelisted data scraping programs) in the past year. According to their responses, this was mostly due to a lack of interest. However, when examining the factors that prevented respondents from applying for APIs (beyond lack of interest, as shown in Figure 3), it appears that reasons vary depending on the platform. Still, the primary reason for not applying for APIs is a lack of awareness of data availability.

Figure 1. Constraints Across All Respondents

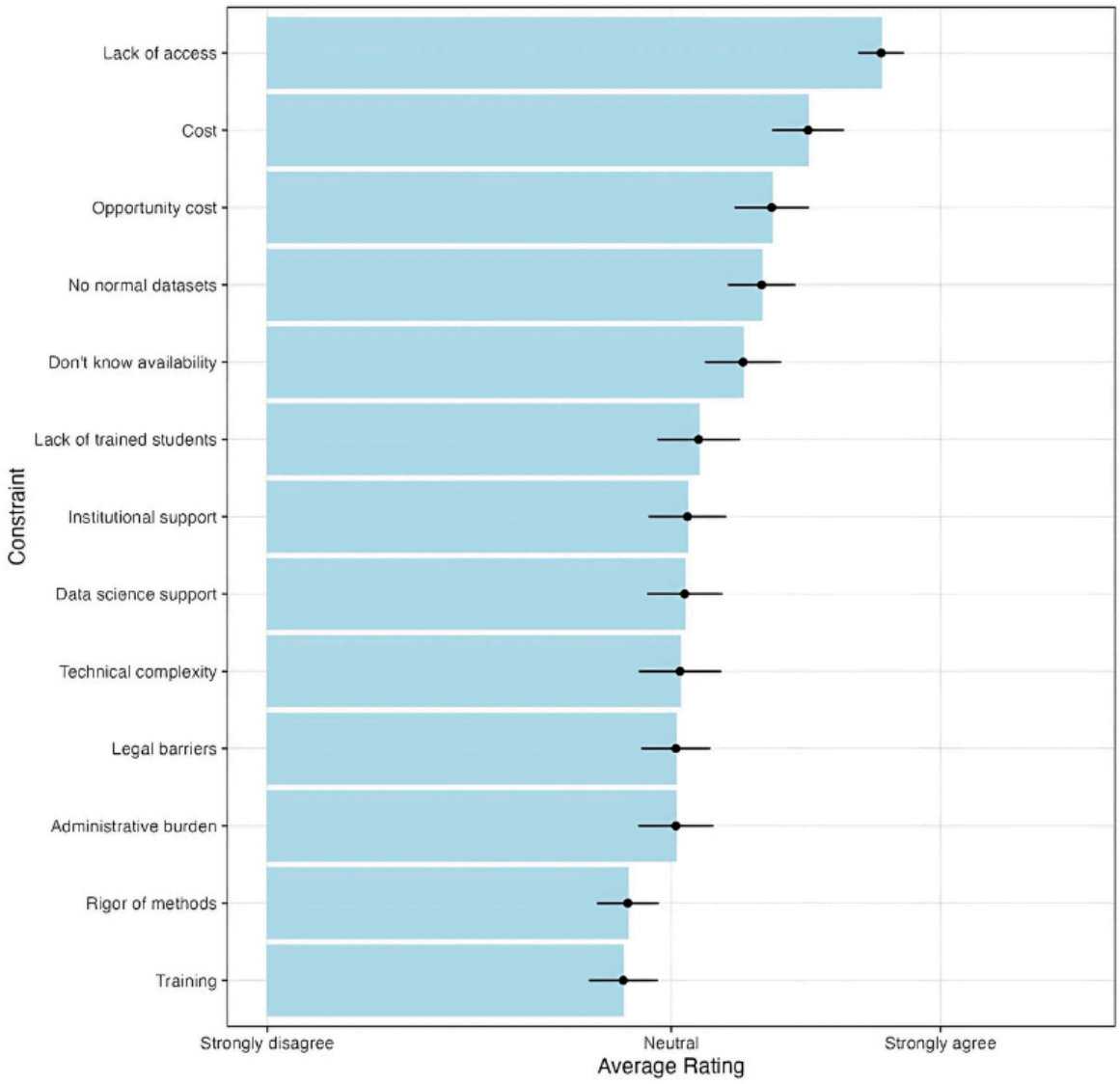


Figure 2. Constraints By Type of Respondent

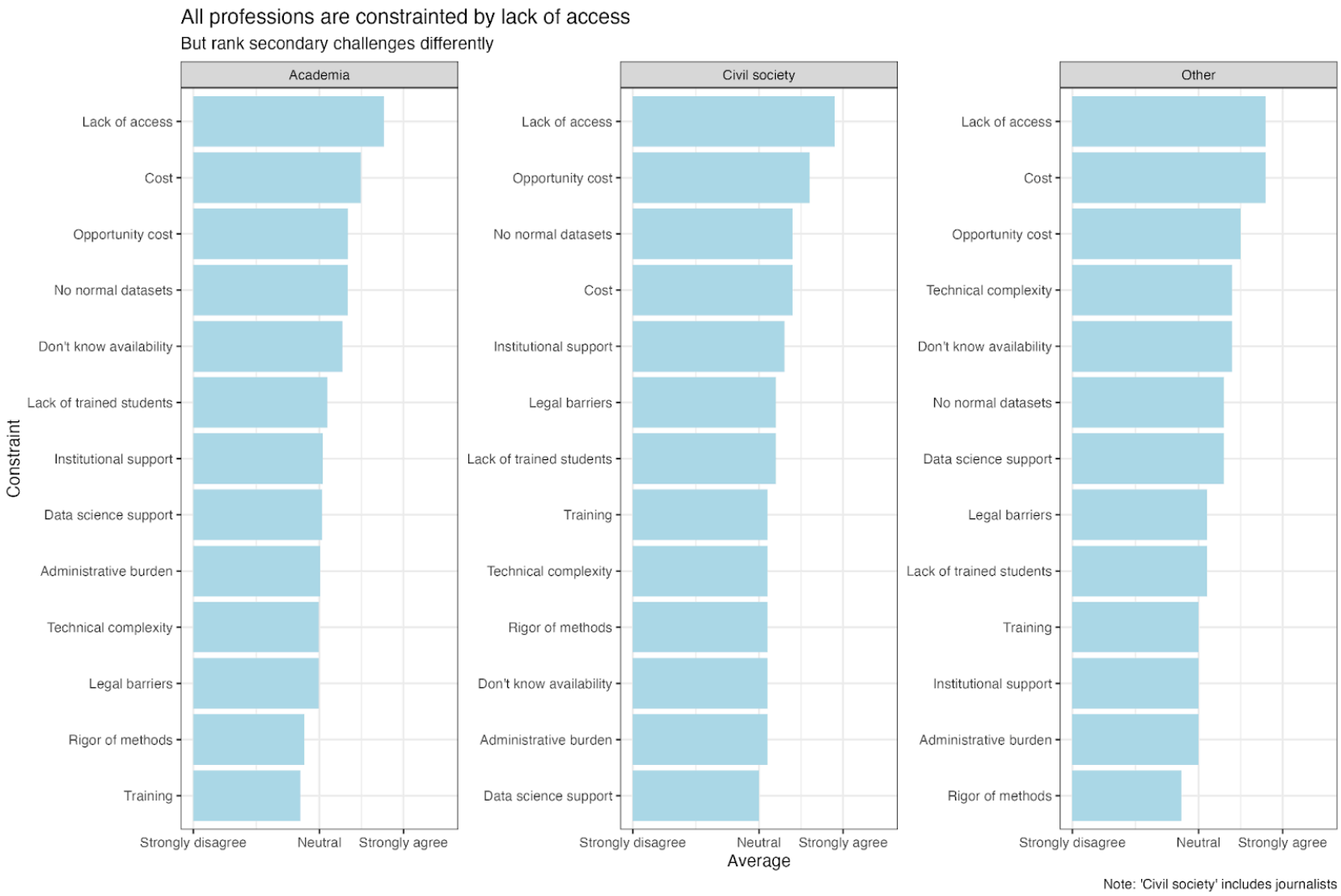
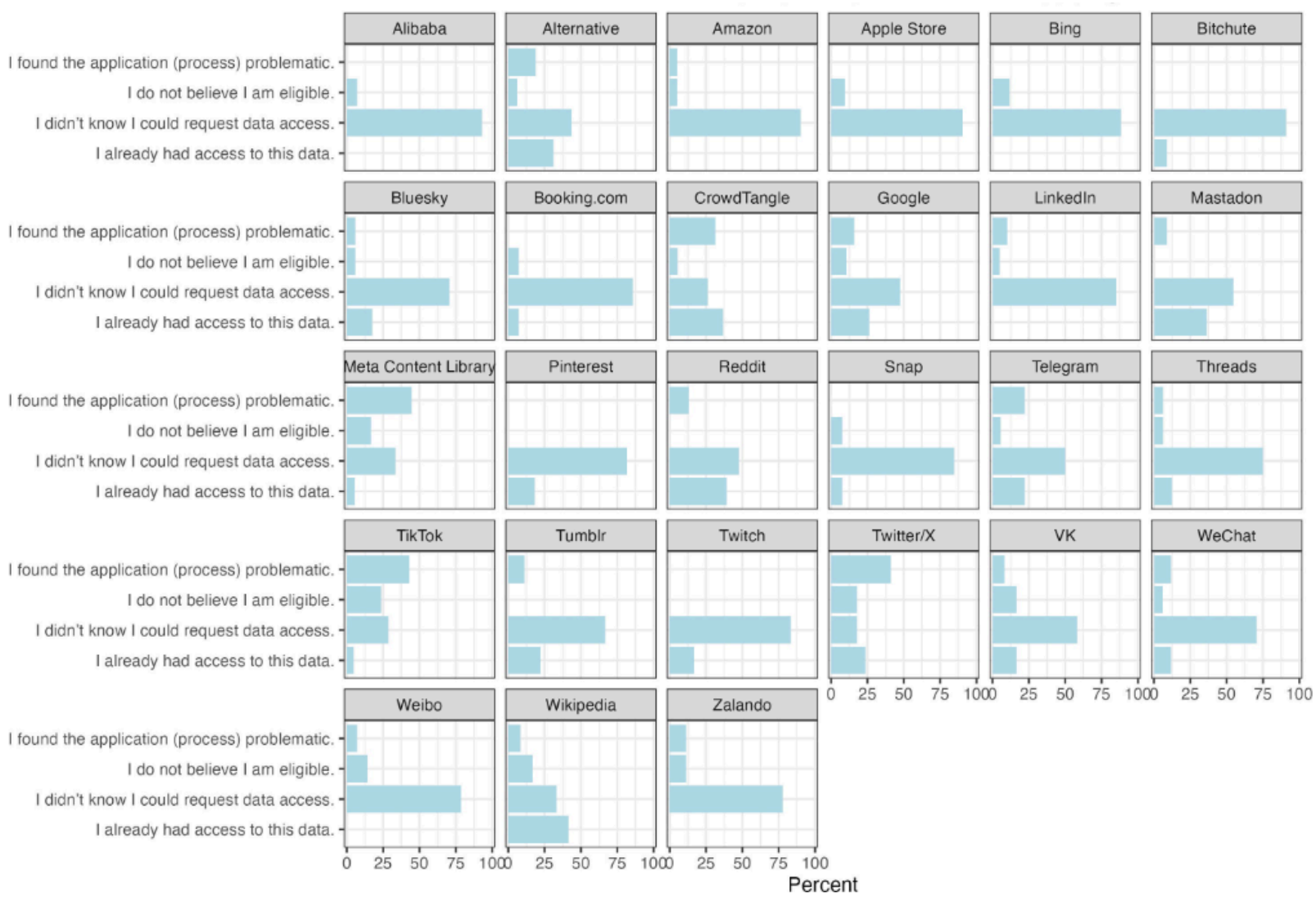


Figure 3. Reasons for Not Applying to APIs

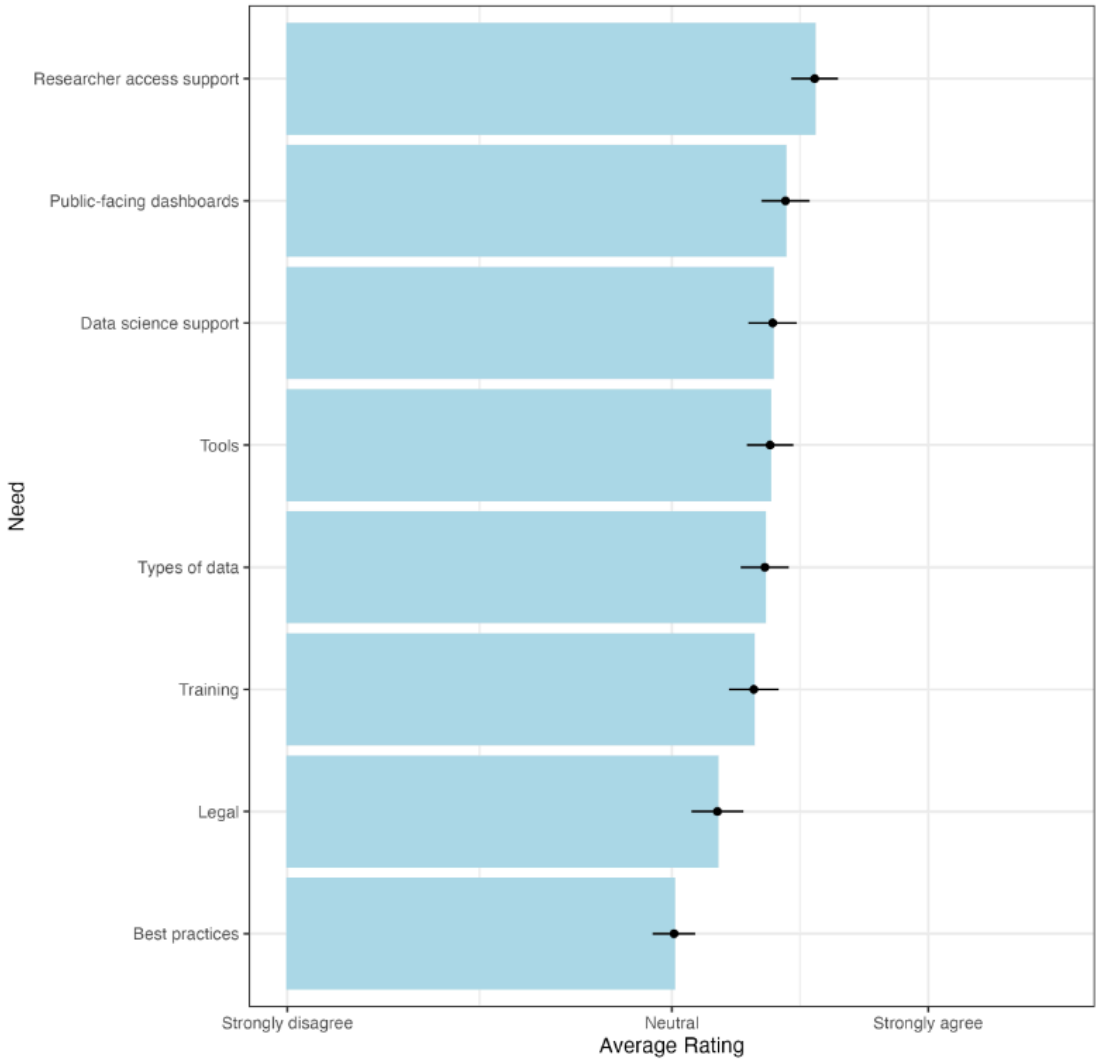


Needs – Types of Data and Tools

When asked which area would benefit most from additional support to enhance their research, respondents identified access to research resources as the top priority for improvement. This includes support with overcoming difficulties related to the lack of compliance with the Digital Services Act (DSA) on the side of the platforms, as well as navigating the legal procedures involved in gaining access to platform data. DSA compliance enforcement and navigation were also recurring themes in some open-ended questions, where respondents shared their experiences. One respondent noted, "sometimes asking for data from platforms is like talking to a wall," highlighting the frustration with data access challenges.

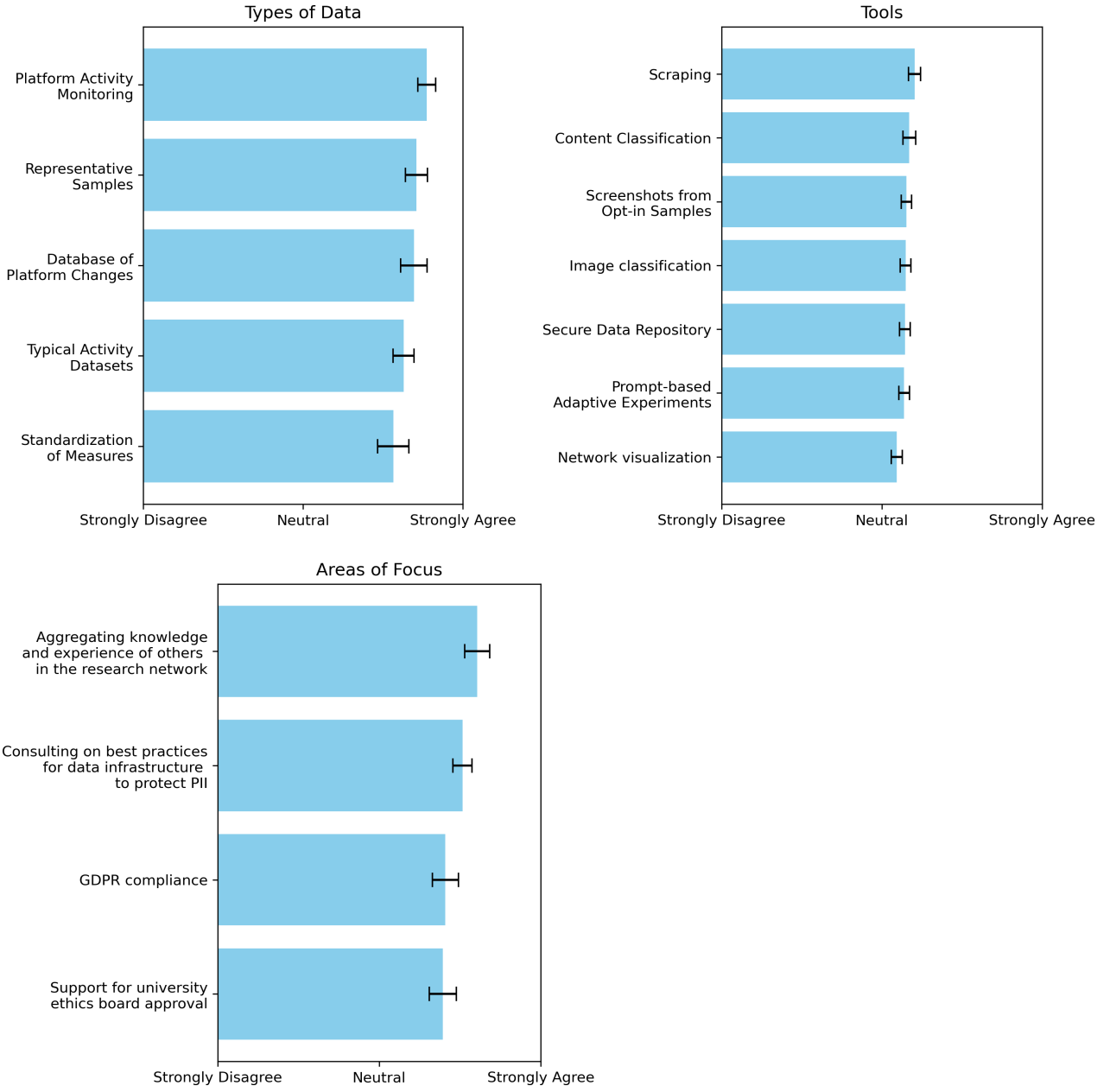
Additionally, there is a noted need for better APIs and tools to streamline data access, with respondents mentioning that existing APIs are often inadequate or overly complicated to use. Respondents also pointed out data request difficulties, with one respondent stating that "data requests by researchers from the global south return more errors compared to requests by the same researchers sent from American or EU universities," underlining the disparities in access based on geographical location. Finally, public-facing dashboards are another area where respondents expressed the need for more support.

Figure 4. Research Support Needs Across All Respondents



Additionally, when respondents are asked about what they would like to see in initiatives like the Accelerator (Figure 5), although they do not show clear variation across the options offered, some important insights emerge. Platform activity monitoring and representative samples seem to be priorities in terms of types of data that should be offered while scraping and content classification come out on top of the list of tools ideally covered. Finally, knowledge and experience aggregation seem to be an important area of focus for respondents.

Figure 5. Priorities for Shared Research Resources



Needs - Interest in Platforms

When looking at respondents' interest in specific platforms (Figure 6), X/Twitter, TikTok, Google, and Meta remain strong and important platforms when it comes to working with large datasets on online human behavior. More specifically, academic researchers are slightly more interested than the rest of the respondents on TikTok, Twitter/X, and Google as Figure 7 shows. Finally, when we look at how platform preferences are distributed by respondents' region (Figure 8), it turns out there is a disproportionate European interest in Telegram, while US respondents are more interested in Reddit, LinkedIn, Weibo, and Wikipedia.

Figure 6. Platform Preferences Across All Respondents

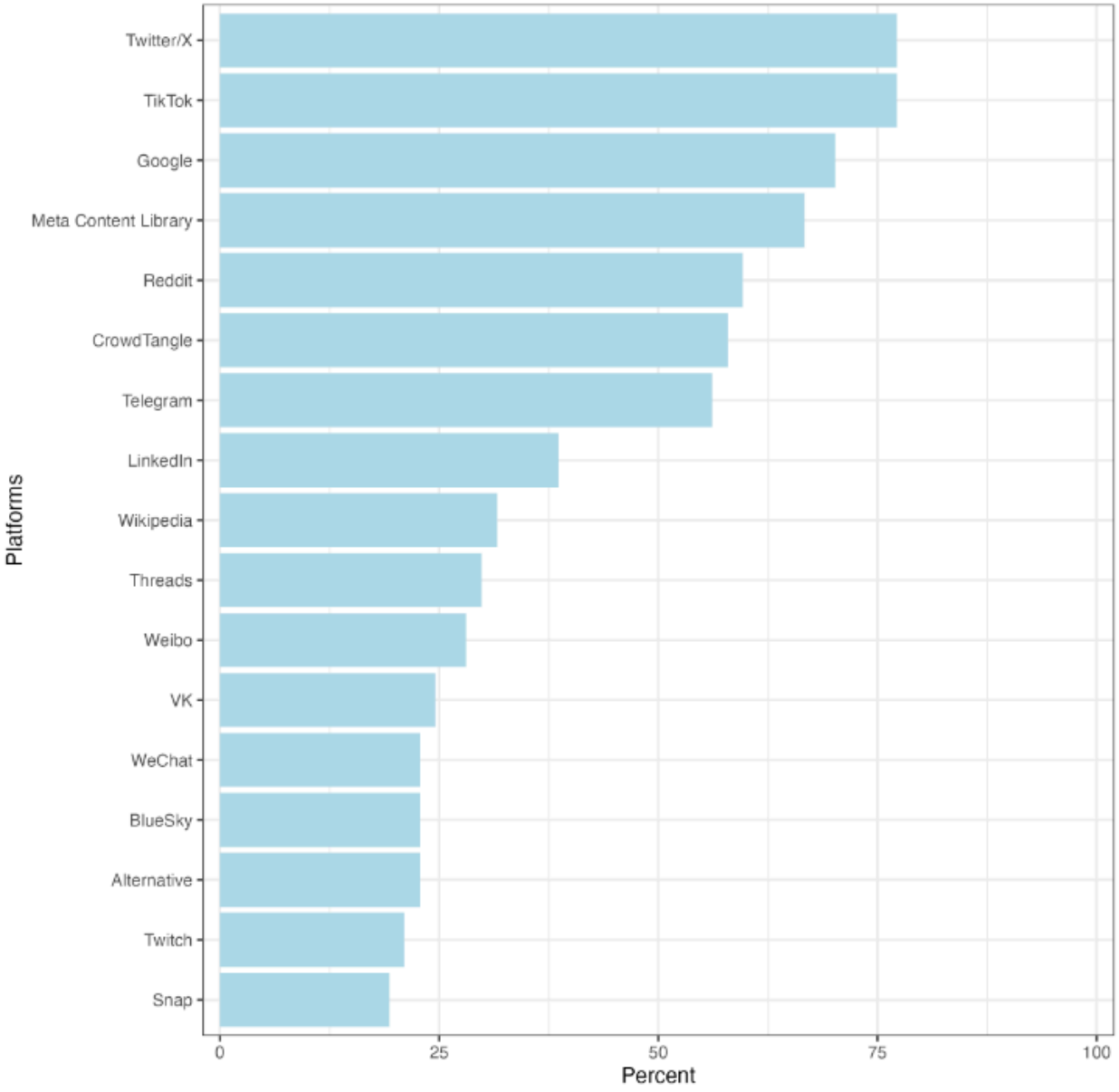


Figure 7. Platform Preferences for Academic Respondents

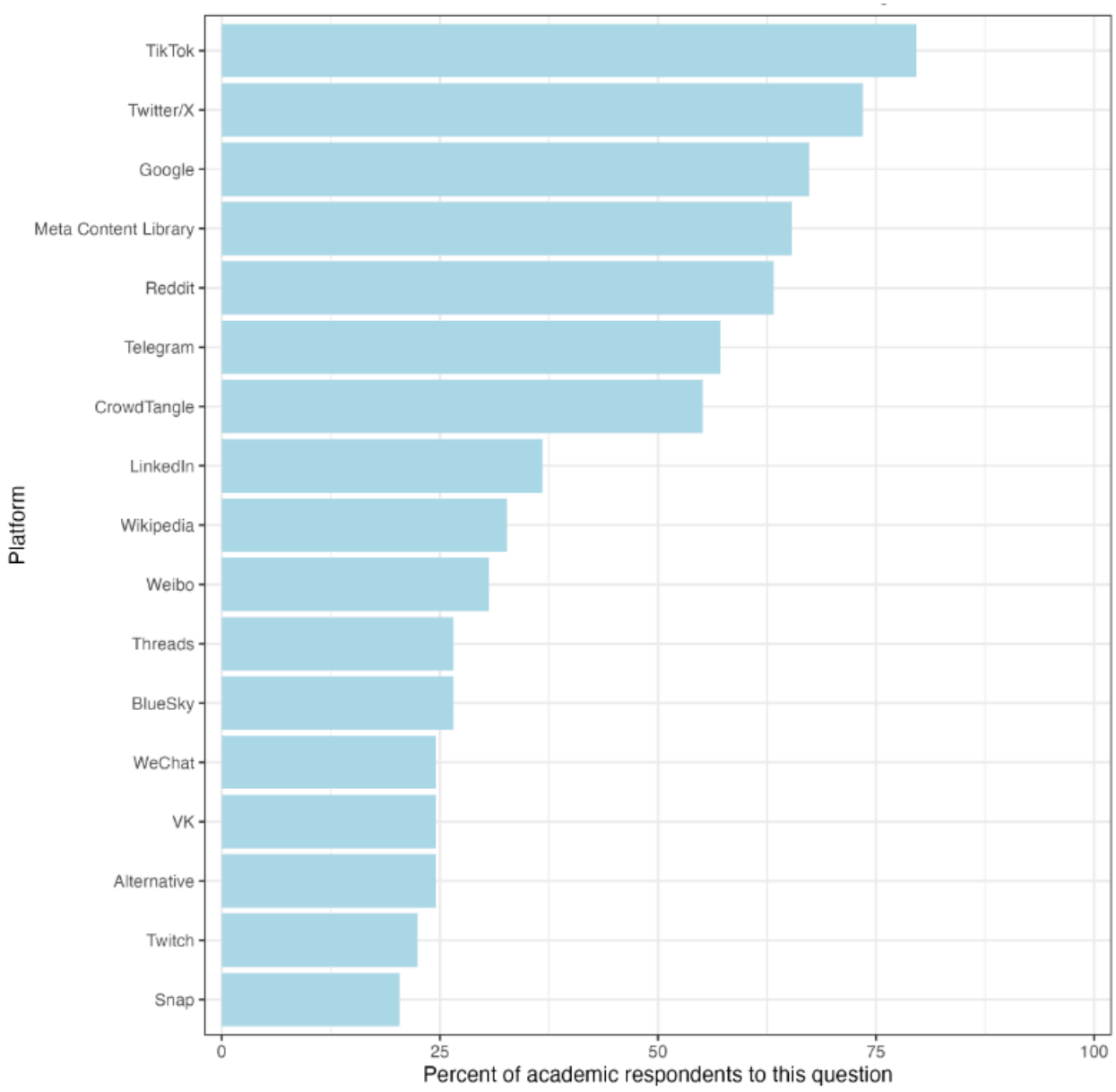
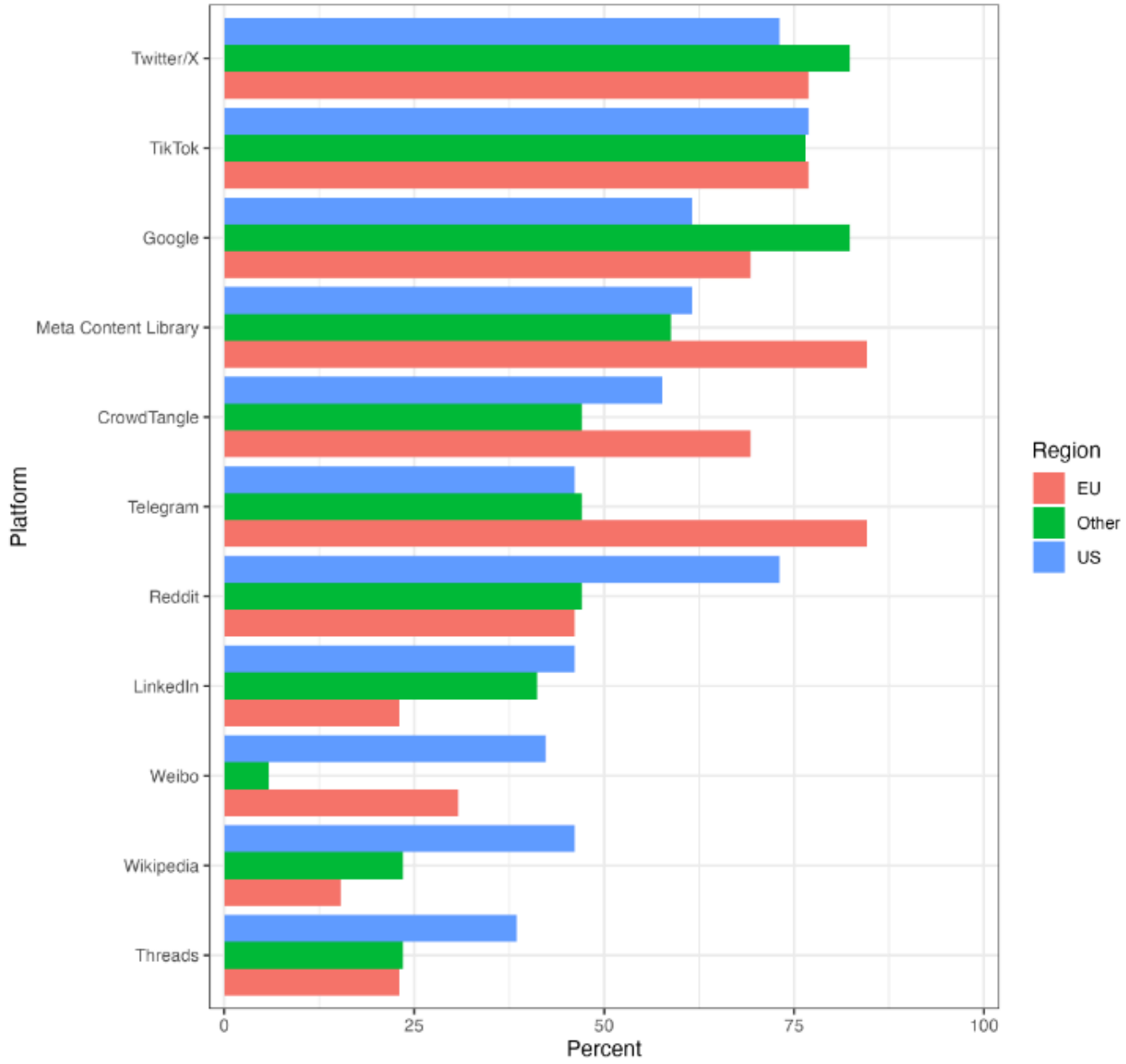


Figure 8. Platform Preferences by Region



Conclusions

The study highlights several key findings about the constraints and pain points faced by researchers and professionals when accessing large datasets on online human behavior. Lack of access emerges as the most significant barrier, followed closely by cost and opportunity costs. A substantial portion of respondents (75%) did not apply for major data access programs, largely due to a lack of awareness of data availability. Furthermore, platform preferences indicate a strong interest in X/Twitter, TikTok, Google, and Meta, with regional differences showing European interest in Telegram and US interest in platforms like Reddit, LinkedIn, Weibo, and Wikipedia. Respondents also emphasized the need for better APIs, tools, and legal support, especially concerning compliance with the Digital Services Act (DSA).

Respondents highlighted challenges in accessing large datasets on online human behavior, even when available, often requiring institutional support and better tools like APIs. Platform-specific restrictions, such as changes to Twitter's academic API, complicate access, while compliance with legal frameworks like the DSA is another obstacle. Respondents called for improved institutional and community support, including collaboration on data donation efforts. Technical approaches such as web scraping were seen as valuable, and many advocated for focusing on aggregate data to ease access and address privacy concerns.

Overall, there was an appreciation for ongoing efforts to improve data accessibility. Interestingly, respondents seem to not be using commercial tools. Around 83.5% of them have not used any commercial tool for getting access to online human behavior data for their research in the past year. Researchers might not be aware of the existence of such tools, or the cost might be prohibitively high. Unfortunately, we did not include such questions in the survey, and further investigation is needed. Coupled with the previous findings, this points to the need for further exploration of the issue.

Initiatives like the Accelerator should focus on tackling these problems. Clear paths have been highlighted with this survey. Building accessible data collections on online human behavior that are aggregable, multi-platform, standardized, and representative is crucial in the eyes of researchers. Furthermore, summarizing access options would solve a large portion of the ongoing pain points for most researchers. This would also include guides to accessing commercial tools readily available. Finally, it would be crucial to include guides around compliance and IRB processes that would be globally applicable and cover various legal settings (US, EU, etc.).