DIW BERLIN

2091

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2024

# Returns to Data: Evidence from Web Tracking

Hannes Ullrich, Jonas Hannane, Christian Peukert, Luis Aguiar, Tomaso Duso

IMPRESSUM

DIW Berlin, 2024

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
https://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
https://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
https://ideas.repec.org/s/diw/diwwpp.html
https://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

# Returns to Data: Evidence from Web Tracking*

Hannes Ullrich† Jonas Hannane‡ Christian Peukert§

Luis Aguiar¶ Tomaso Duso‖

July 12, 2024

**Abstract**

Tracking online user behavior is essential for targeted advertising and is at the heart of the business model of major online platforms. We analyze tracker-specific web browsing data to show how the prediction quality of consumer profiles varies with data size and scope. We find decreasing returns to the number of observed users and tracked websites. However, prediction quality increases considerably when web browsing data can be combined with demographic data. We show that Google, Facebook, and Amazon, which can combine such data at scale via their digital ecosystems, may thus attenuate the impact of regulatory interventions such as the GDPR. In this light, even with decreasing returns to data small firms can be prevented from catching up with these large incumbents. We document that proposed data-sharing provisions may level the playing field concerning the prediction quality of consumer profiles.

*Keywords:* Prediction quality, Web Tracking, Cookies, Data protection, Competition Policy, Internet Regulation, GDPR.

*JEL codes:* C53, D22, D43, K21, L13, L4.

*We thank Ilia Azizi for excellent research assistance. We are grateful to Johanna Arlinghaus, Amelia Fletcher, Paul Heidhues, Shan Huang, Anna Kerkhof, as well as conference and seminar participants at DIW Berlin, the annual conference of the German Economic Association, the Digital Economy Workshops in Munich and Norwich, and the Berlin School of Economics and DIW Graduate Center Workshops for helpful suggestions. Christian Peukert acknowledges the support from the Swiss National Science Foundation for the project 100013_197807. The order of authors was randomized and archived in the American Economic Association's Random Author Order Archive: https://t.ly/NI6l_.

†DIW Berlin, University of Copenhagen, and CESifo. E-mail: hullrich@diw.de.

‡DIW Berlin and Technische Universität Berlin. E-mail: jhannane@diw.de.

§*Corresponding author.* HEC Lausanne and CESifo. E-mail: christian.peukert@unil.ch.

¶University of Zurich and CESifo. E-mail: luis.aguiar@business.uzh.ch.

‖DIW Berlin, Technische Universität Berlin, CEPR, and CESifo. The views expressed in this article are those of the author and may not in any circumstances be regarded as stating an official position of the German Monopolies Commission. E-mail: tduso@diw.de.

# 1 Introduction

Business models centered around the use of consumer data as a core asset have thrived in digital markets. A prominent example is targeted online advertising, where advertisers often rely on inferred consumer characteristics and preferences to deliver personalized ads (Johnson et al., 2020; Rafieian and Yoganarasimhan, 2021). While consumers value enhanced products and better matches, policymakers increasingly worry that the consolidation of digital platforms and the consequent combination of different data sources might lead to less choice and exploitation in the long run, making consumers worse off (Tucker, 2019; Economides and Lianos, 2021; Chen et al., 2022). Understanding the link between data inputs and the quality of predictions of consumer characteristics is important for marketing strategies as much as for policy discussions on the necessity and design of regulation. Notably, data collection carries a cost in terms of online privacy (Goldfarb and Que, 2023) and may give rise to competition concerns due to potentially unsurmountable barriers to entry created by strong data network effects and externalities (Hagiu and Wright, 2023). Despite the importance of these concerns, empirical evidence on whether the scale and scope of data available to incumbents may represent a barrier to entry for smaller competing firms remains scarce.

Web tracking allows firms to collect information on consumers' browsing behavior to build user profiles based on demographic or interest attributes, which may then be sold to advertising companies (Neumann et al., 2019).[1] Web tracking is enabled by small pieces of code embedded in a website, which send information about the user to a third-party tracking firm when the website is loaded. A prime example is Google Analytics, a free service that provides a dashboard of usage statistics to website owners. After installing Google Analytics, websites effectively share their users' information with Google. The size and scope of a firm's tracking network depend on how many websites choose to use its services.[2] Firms can increase the scale of their data by tracking additional users or the scope of their data by collecting additional information about users they already track.

In this paper, we measure the production function for demographic profile prediction quality using web-browsing data observed by the 52 largest trackers on the Internet. We have access to the full browsing history of over 75,000 US users over 12 months, alongside survey-based demographic information including age, gender, income, and location. Combining these data with domain-level tracking data, we exploit observed and simulated variation in the scale and

[1] Web tracking firms have also been labeled web technology vendors as they often provide services to website publishers such as web analytics, social media sharing, or the placement of advertisements (Peukert et al., 2022; Johnson et al., 2023).

[2] Data may be collected by third-party trackers on websites but also by the (first-party) websites that users visit.

scope of tracker-collected data to estimate the prediction returns to data and study the role of regulatory interventions for profile prediction.

Using a gradient boosting machine learning method, we first test how well consumers' demographics can be predicted based on their own and other users' browsing data. To do so, we define binary outcome variables for distinct demographic categories – such as specific age or income levels – and retrain the prediction model for each outcome conditional on the data available to each firm. This analysis generates a large set of prediction tasks including varying data inputs associated with varying prediction qualities. In addition to using observed variation in the scale and scope of data across trackers, we generate within-tracker variation in the scale and scope of data by drawing random sub-samples of users and websites for each tracker. This allows us to construct counterfactuals of varying data inputs for each tracker. The resulting prediction quality and input data observations provide the basis for a systematic analysis of data-enabled learning curves.

We then explore potential complementary effects on prediction quality when combining individual browsing data with user demographic information. Large firms like Google, Facebook, or Amazon indeed can directly obtain such information, at least partially, for instance by eliciting their users' birth date when creating an account. Such data access might play an important role if data complementarities can help overcome decreasing returns in one type of data (Peukert et al., 2023; Schaefer and Sapi, 2023). Finally, we estimate how prediction quality changes with the introduction of regulatory interventions – such as the European General Data Protection Regulation (GDPR) and the Digital Markets Act (DMA) – and discuss implications for the competitive landscape in the web tracking market.

We find substantial variation in prediction quality across prediction tasks. For some demographic outcomes, web browsing data exhibits only little predictive power. In these instances, we find that larger firms benefit only marginally, if at all, from their access to more extensive browsing data. However, for prediction tasks where web browsing data yields higher prediction quality, larger firms have an advantage over smaller firms with less data. Overall, we find decreasing returns to data in the number of users and tracked websites, both across and within trackers. We also find evidence of data complementarities between web browsing data and demographic information. For Google, Facebook, and Amazon (GFA) – the three largest tracker firms who arguably have more direct access to users' demographic information – combining demographics with browsing data substantially enhances prediction quality. This improvement is equivalent to the gains in prediction quality that can be achieved by relying solely on browsing data from 100,000 domains, and such complementarity may therefore prevent smaller trackers

from catching up with these large incumbents.

Data-sharing provisions, such as those mandated by the DMA, could help level the playing field. We show that giving competing trackers access to demographic data significantly reduces the gap between their prediction quality and that of Google. Finally, our results show that the prediction quality of all trackers decreased after the implementation of the GDPR. This is because websites substantially removed trackers, even in jurisdictions where GDPR does not apply de iure (Peukert et al., 2022). However, large trackers can countervail these negative effects through their ability to combine demographic information with users' browsing data. While the GDPR may have improved consumer privacy due to less tracking, indirectly, it also seems to have increased the competitive position of Google, Amazon, and Facebook in the web-tracking market.

Our study has important policy implications. The consolidation of digital ecosystems by internal and, especially, external growth through acquisitions of smaller competitors to enable the use of rich consumer data raises concerns about reduced competition and potential exploitation. This has been relevant in the past, as the combination and use of different data contributed to raising barriers to entry and strengthening the dominance of large tech firms. Yet, evidence on data complementarities underlying such market dynamics is lacking (Calvano and Polo, 2021). The issue may become even more relevant in the exploding market for generative AI as data is the core input for AI models. Data complementarities are likely to reinforce an already highly concentrated industry. Thus, policymakers must balance the benefits of data accumulation and its positive impact on improving product quality against privacy concerns and the risks associated with market dominance by a few large firms. Effective regulation should aim to ensure data privacy and foster a competitive environment through access to data in which smaller firms can thrive. This requires a deeper understanding of data-driven business models and their impact on market dynamics.

We contribute empirical evidence to a largely theoretical literature that has highlighted the role of data in providing firms with a competitive advantage (Gregory et al., 2021; Hagiu and Wright, 2023). So far, the existing work on the value of consumer data focuses on single online platforms (Bajari et al., 2019; Peukert et al., 2023; Schaefer and Sapi, 2023; Aguiar et al., 2023; Lei et al., 2023; Luca et al., 2023), while we investigate differences across platforms. The three most related papers develop and discuss methods to infer consumer profiles from web tracking data (Trusov et al., 2016), provide evidence that consumer profiles offered by data brokers are of low quality overall (Neumann et al., 2019), and estimate that information from web browsing histories can be much more valuable for personalized pricing than pure demographic information

(Shiller, 2020). However, they are limited to studying a few intermediaries, i.e., search engines, ad networks, and data brokers. Further, they do not focus on analyzing differences among intermediaries of different sizes, or other competitive dynamics, and do not attempt to evaluate policy interventions in the web-tracking market. We add systematic evidence to confirm prior work showing that firms can improve their predictions of consumer profiles with more and broader data, albeit with decreasing returns to scale and scope (Klein et al., 2023; Peukert et al., 2023; Schaefer and Sapi, 2023).

Notably, we show that significant competitive benefits come from non-behavioral data that the largest and vertically integrated firms in the industry can easily collect at scale. Based on our evaluations of the GDPR, which increased the cost of collecting behavioral data and therefore disproportionally hurt small and non-integrated firms, and the DMA's provision mandating large firms to share data, we show that privacy regulation might counteract competition policy goals.

This paper is organized as follows. Section 2 provides the industry and policy background and Section 3 describes the data. Section 4 explains our approach to estimating the returns to data and Section 5 shows the main results. Section 6 shows the *ex ante* and *ex post* evaluations of privacy and data sharing regulations. Section 7 concludes.

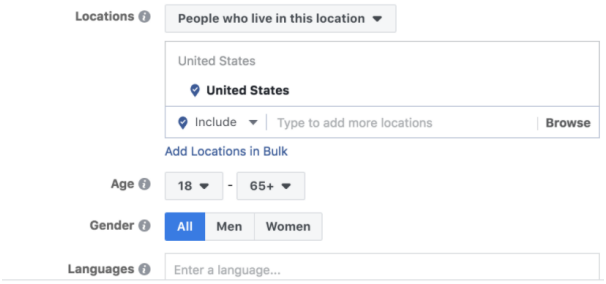
# 2 Industry background and policy discussion

## 2.1 Consumer profiling and web tracking

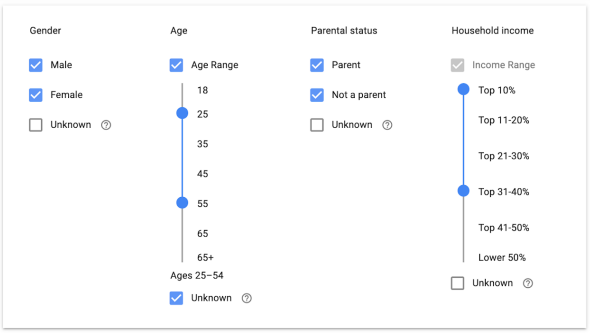
Information about consumer profiles ranges from basic demographics to fine-grained preferences for specific product characteristics and can be valuable to advertisers, both for positive and negative targeting. Positive targeting focuses on consumers who are likely to respond well to specific ads, aiming to increase engagement and conversion rates. In contrast, negative targeting aims to help advertisers avoid unnecessary ad impressions. Online advertising vendors therefore typically offer a variety of targeting options through so-called audience selection tools that let advertisers filter on specific variables (see screenshots from Facebook and Google in Figure 1).

Figure 1: Example: Screenshots of audience selection tools

*Panel A: Facebook*




*Panel B: Google*

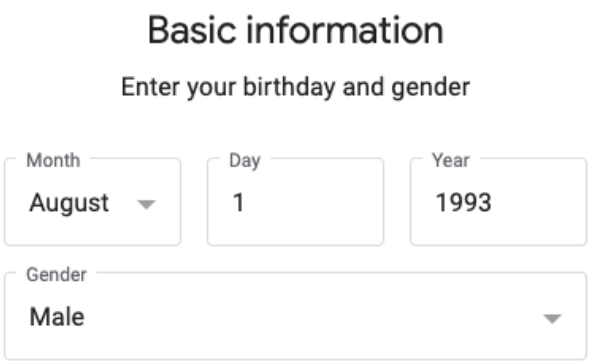



**Source:** https://sproutsocial.com/insights/facebook-ad-targeting/ and https://support.google.com/displayvideo/answer/6071542?hl=en

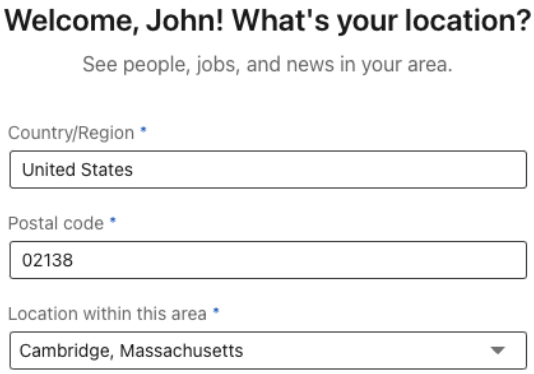
Advertising intermediaries can gather information about Internet users' demographics through several means. As illustrated in Figure 2, online platforms often require users to directly share personal information – including their birthday, gender, or geographic location – when signing up for their services. Often, however, consumer profiles and purchase probabilities are inferred from users' behavioral data (e.g. a user's clickstream; Trusov et al., 2016; De Cnudde et al., 2020; Shiller, 2020).

Figure 2: Example: Screenshots of major platforms' signup procedure

*Panel A: Google*




*Panel B: LinkedIn*




**Note:** Screenshots taken in June 2024.

The machine learning approaches used to predict user profiles are implemented in standard open source software and are computationally relatively inexpensive (Trusov et al., 2016; De Cnudde et al., 2020). This makes consumer profiling scalable to millions of users, and there is likely little room for firms to gain a competitive advantage based solely on software engineering and computational power. The key ingredient appears to be access to data, which has led to the flourishing of an entire industry devoted to online behavioral data collection over the past few decades (Lerner et al., 2016). Field experiments indeed show that the quality of consumer

profiling, i.e. the accuracy of estimates relative to the ground truth, is often not very high and varies substantially across suppliers (Neumann et al., 2019), pointing towards a potential competitive advantage in data.

How does web tracking work? Making use of the modular nature of modern websites and apps, web-tracking technologies pick up the digital traces that users leave behind when accessing the Internet. For example, as a user navigates to *news.com*, the web server sends so-called requests to load resources from third-party domains (e.g., an image hosted on *tracker.com*). Many use cases of third-party requests are related to online advertising. For example, third-party requests can be necessary to deliver content (e.g., advertising creatives), or helpful to identify the same machine or individual across time and websites (e.g., via cookies). The corresponding web server processes the request and delivers the response, but may also store meta information such as the context in which the request originated. For example, information on the website that originated the request and the IP address of the user. Storage can happen both server-side and as cookies on the user's device. The next time the same third-party service is requested from a different website, the same user is identified based on the meta information, enabling cross-website tracking.

## 2.2 Privacy and competition policy issues in digital markets

Web tracking is now ubiquitous, and its massive scale has raised consumer privacy concerns and prompted global regulatory action to curb online data collection and processing. As early as 2002, the EU introduced the e-Privacy Directive (Directive 2002/58/EC) and its amendment "cookie law" (Directive 2009/136/EC), focusing on data protection in electronic communications and cookie consent. The landmark GDPR, implemented in May 2018, replaced the previous directive with a robust, comprehensive framework. GDPR strengthened individuals' rights, increased accountability for data processors, and introduced substantial penalties for non-compliance. Overall, it made cookie-based web tracking more costly, difficult, and transparent.

In contrast, the US has a more fragmented approach with sector-specific and state-level laws. Most recently, the California Consumer Privacy Act (CCPA), which took effect in January 2020, marked a significant step towards comprehensive privacy regulation in the US. The CCPA grants California residents broad rights over their personal data, including the right to know what data is being collected, the right to delete personal data, and the right to opt out of the sale of their data. Following the CCPA, other states have begun to consider or enact similar privacy laws, signaling a potential shift towards more uniform privacy standards across the

country. However, evidence shows that the GDPR effectively already impacts US consumers, as websites with international audiences adopt EU standards globally to comply with the stringent EU regulations (Peukert et al., 2022).

Further, a growing body of literature highlights an important intersection between privacy and competition policy, mostly because compliance costs are often too high for small firms leaving large firms at a competitive advantage (Campbell et al., 2015; Kira et al., 2021; Peukert et al., 2022; Johnson et al., 2023). Hence, while privacy regulations help consumers protect their personal data, in pre-existing concentrated markets they may further stifle competition if data are necessary to compete in markets for personalized products and services. Competition concerns are particularly strong when data are difficult to substitute for new entrants, complementarities between various types of data are strong, or targeting has increasing returns to scale for data inputs (Calvano and Polo, 2021).

The fact that privacy policy cannot be separated from competition policy is now recognized by regulators. With a few large firms dominating many digital markets, potentially stifling innovation and consumer choice, regulators have introduced new competition policy measures that explicitly take data as a source of competitive advantage into account. In a move from an ex-post to an ex-ante approach to competition policy and antitrust, the EU introduced the DMA in 2023. The DMA aims to create more competition by imposing specific obligations on "gatekeepers" – large online platforms with significant market influence. One key aspect of the DMA is the requirement for these gatekeepers to share data with smaller competitors and third parties, fostering a more competitive and dynamic market environment (Article 6(10), DMA). This data-sharing obligation is designed to prevent monopolistic practices, promote interoperability, and ensure that smaller firms have the opportunity to innovate and compete on a level playing field.

In the US, competition policy in digital markets has also been a topic of significant debate, though it has not yet resulted in legislation as comprehensive as the DMA. Recent years have seen increased scrutiny of major tech companies by federal and state regulators, with several high-profile antitrust lawsuits filed against firms like Google, Facebook, and Amazon. These cases focus on practices such as self-preferencing, exclusionary contracts, and acquisition of potential competitors. While data-sharing obligations similar to those in the DMA have not been formally legislated, there are growing calls from lawmakers and advocacy groups for measures that would level the playing field.

Despite these efforts, the US approach remains less prescriptive than the EU's, reflecting a different regulatory philosophy that balances competition enforcement with concerns about

over-regulation. However, the ongoing scrutiny and legislative initiatives indicate a shift towards a more robust competition policy in the US digital markets, mirroring some of the objectives of the EU's DMA. US consumers may already be affected by the DMA because of the extraterritorial reach of EU law. Large tech companies may implement changes globally rather than maintain different standards for different regions, leading to improved data access and increased competition to the benefit of consumers outside of the EU.

## 3 Data

We use individual-level desktop browsing data from a sample of the U.S. population collected by market research firm Comscore, which covers just over 75,000 users.[3] The complete web browsing history of each user is collected by a software tool that runs in their desktop's background. Participants, who receive monetary compensation and other incentives, fill in a survey of basic demographics, such as age, education level, geographic location, and household income. For each user, we observe the domain names of all websites visited in 2018, as well as the time spent on each domain. For the main part of our analysis, in Sections 5 and 6.1, we restrict the browsing data to the 20 weeks after the GDPR was introduced on May 25, 2018.[4] In Section 6.2, we also make use of the browsing records from the 20 weeks before GDPR was introduced to measure the change in trackers' data collection and prediction quality induced by the introduction of this regulation.

We combine the individual-level clickstream data with historical information on websites' connections to third-party services at the website domain level (e.g. digitalecon.org). We obtain historical data on websites' requests from the HTTPArchive, a non-profit project that periodically crawls hundreds of thousands of websites, records the data, and makes it publicly available.[5]

To reduce the complexity of our dataset, we aggregate the information from HTTPArchive over time. If a website makes one request to a third-party service during our observation period, we consider the website to be "tracked" by that specific third-party service throughout our sample period. That is, we make the implicit assumption that a websites' connections

[3]Our paper is not the first to use clickstream data from media measurement and analytics companies such as Comscore or Nielsen (Trusov et al., 2016; Dambra et al., 2022; Aguiar et al., 2023). Prior research has established that these data provide a good representation of the population with Internet access (Aguiar et al., 2018). Even if users signing up for Comscore's service have different privacy preferences than the general population, this is only a concern for our analysis in so far as privacy preferences affect the scope and intensity of web browsing. With these caveats, it remains to note that clickstream data collected with voluntary consent of users is the best available data source for our study.

[4]Specifically, we restrict the observation period to the calendar weeks 21-40 of the year 2018.

[5]These data are available via httparchive.org.

to third-party trackers remain stable over time. Table 1 provides summary statistics of the Comscore data alone and the merged data with HTTPArchive, distinguishing between the pre- and post-GDPR periods. While about 12% of the domains from the Comscore sample appear in the HTTPArchive data, these account for over 77% of all recorded clicks (78% pre-GDPR). More than 99% of the sample users have at least one of their visited domains appear in the HTTPArchive data.

Table 1: Comscore and HTTPArchive data

| Period | Number of users | Number of domains | Number of clicks | Data source | Share of tracked clicks |
|---|---|---|---|---|---|
| Pre-GDPR | 75,407 | 1,275,203 | 278,363,427 | Clickstream | |
| Pre-GDPR | 75,158 | 172,120 | 217,985,811 | Clickstream & Tracking | 78% |
| Post-GDPR | 75,523 | 1,274,237 | 256,614,844 | Clickstream | |
| Post-GDPR | 75,150 | 155,247 | 197,536,318 | Clickstream & Tracking | 77% |

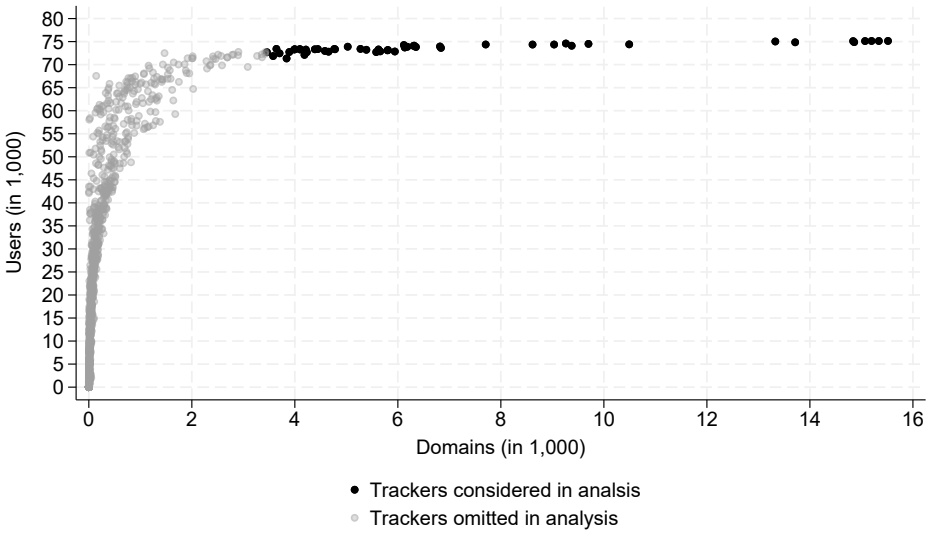
The HTTPArchive data provides the list of all domains to which a website sends communication requests when a user visits that website. We use information from whotracks.me to match these requested domains to trackers and to the firms operating these trackers. For our analysis, we include Google and Amazon, which operate multiple individual trackers, as combined entities and the 50 largest trackers in terms of website clicks. These 50 trackers also include individual trackers operated by Google and Amazon. Figure 3 shows the distribution of tracked users and website domains across trackers, distinguishing between the 52 trackers used in our analysis in black and the excluded trackers in grey. Google, Facebook, and Amazon (GFA) are the largest firms in terms of tracked domains in our sample period, with 155,160, 133,285, and 97,011 tracked domains, respectively. Table 5 in Appendix A provides of a descriptive overview of the trackers in the analysis sample.

Computing the number of clicks per website a user visits over the observation period, we jointly observe, for each user, the intensity of visits per domain, the trackers active on each domain, and the user's characteristics.

## 4 Evaluating consumer profile predictions with varying data

To characterize the impact of varying scale and scope of web-tracking data on the quality of consumer profiling, we estimate how well firm $i$ can predict a specific demographic $s$ of consumer $j$, given the browsing history of all observed consumers ($N_i$) across all tracked domains ($K_i$). We focus on 39 demographic subgroups by age, geographic location (census regions), racial background and country of origin, education levels, household size, household income, and

Figure 3: User - Domains distribution across trackers




whether children live in the household. The classification of users into these subgroups represents 39 binary prediction tasks.

In the data, each observation corresponds to a user. We construct a predictor variable per website domain that takes the value of a user's total number of clicks in the entire sample period. This setup yields a high-dimensional dataset, where the number of predictors largely exceeds the number of observations. For instance, when we perform a prediction task using the full sample including all tracked domains, we have 155,247 predictors (domains) for 75,150 observations (users).

To implement a large number of tracker-specific prediction tasks, at a minimum 52 times 39 tasks, using these high-dimensional data within reasonable time, we use the Light Gradient Boosting Machine (LightGBM) algorithm (Ke et al., 2017). As other gradient boosting algorithms such as XGBoost, LightGBM builds decision trees sequentially. However, LightGBM incorporates two novel techniques, namely gradient-based one-side sampling and exclusive feature bundling, which allow drastic reductions in training time on large datasets.[6]

For each prediction task, we minimize the binary logistic loss function. The target prediction measure is the area under the receiver operation characteristic (ROC) curve (AUC), which

[6]Gradient-based one-side sampling discards data instances with small gradients (i.e. errors) and puts more weight to high gradients during training. Therefore, training is focused on data points that will benefit the model the most. Exclusive feature bundling groups mutually exclusive predictors together, reducing the data's dimensionality. Ke et al. (2017) show that LightGBM can significantly outperform other gradient-boosting algorithms in terms of computational speed and memory consumption.

we evaluate using 5-fold cross-validation.[7] The AUC quantifies the location of the ROC curve, which represents all achievable trade-offs between false positive rates and true positive rates by a given prediction technology. It ranges from 0.5 (random classification) to 1 (perfect classification). The AUC is a robust metric insensitive to imbalanced datasets, where the number of positive and negative examples is not equal. This makes it a suitable measure for our purposes since some demographic groups such as high-income or specific age ranges are imbalanced.

By applying the same algorithm to each tracker and prediction task, we abstract from differences in prediction technologies across trackers. That is, we isolate the effect of information content in the meta-data generated by users' browsing histories on prediction quality. If larger tracker firms such as Google or Facebook could improve their prediction technology relative to smaller trackers, e.g. due to larger computational resources or better algorithms, we would underestimate the predictive returns that can be extracted from meta-data collected on their tracked websites. However, as discussed above, the methods broadly used for profile predictions using structured data such as browsing histories are largely off-the-shelf machine learning methods that are used in research and industry (Trusov et al., 2016; De Cnudde et al., 2020). Thus, we expect the advantage of superior engineering and computational power to be limited in this context, which is in line with empirical results from online search (Klein et al., 2023).[8]

# 5 Results

## 5.1 Tracker size and prediction quality

We first provide a graphical description of how increasing web browsing data translates into prediction quality. Figure 4 depicts the returns to data across trackers for four distinct prediction tasks, where we measure the amount of data by the number of tracked domains on the horizontal axis. The figure reports four subplots ordered by the full-sample mean AUC. The top-left pane shows the results for the task of classifying users into the group "Household Size: 3 people", which has the lowest mean total sample AUC of 0.53. The low value for the total sample AUC shows this is a difficult prediction task for which machine learning predictions are only slightly better than random guessing. The top-right ("Age: 65 and above") and bottom-left ("Have Children") panes show prediction tasks with average mean total sample AUCs of 0.68 and 0.76. The bottom right pane shows the results for the classification task "Census Region: North

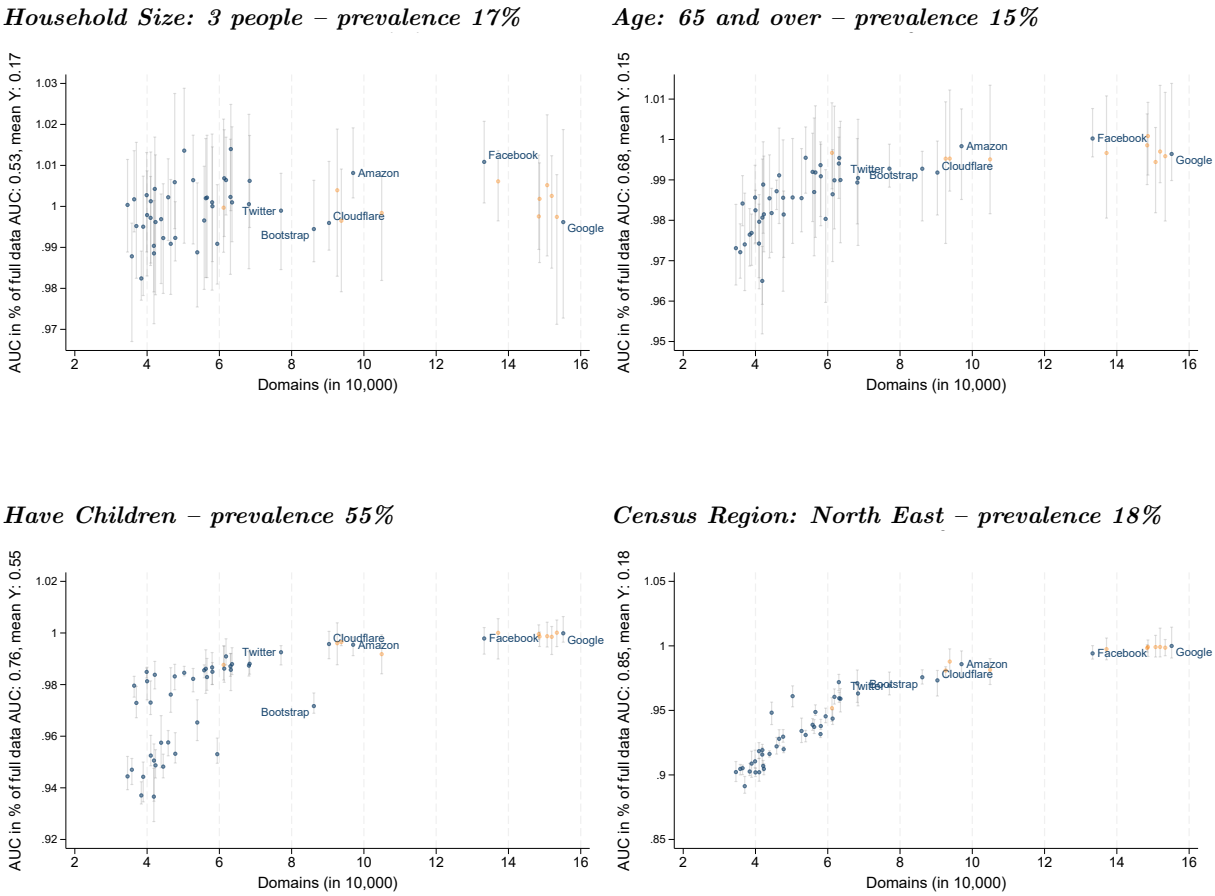
[7] Specifically, we randomly split the available data into five equal-sized folds which we will denote by $f$. We repeat the training and prediction steps five times, using one fold as the test sample and the remaining four as the training sample. By doing so, we obtain five prediction samples for each task and tracker to avoid relying on one test sample in the assessment of prediction quality.

[8] Computational resources may be more important for training large language or image models based on unstructured data but the key to these large models' success has been the vast accumulation of data.

East", which has the highest mean total sample AUC of 0.85 and can thus be considered the simplest prediction task.

For difficult prediction tasks with a low mean AUC in the full sample, for instance "Household Size: 3 people", large trackers such as GFA do not have an advantage. Given that the overall mean total sample AUC for this task is only slightly above 0.5, this result is intuitive: If data are virtually useless to perform a specific prediction, it does not matter how much data a firm has. Prediction quality will always remain low in this case. However, as the overall prediction quality grows, the advantage of large trackers, who observe more websites, becomes discernible. In the bottom panes of Figure 4, large trackers such as GFA achieve better results than all other trackers.

Figure 4: Returns to data in predicting user profiles from web browsing history

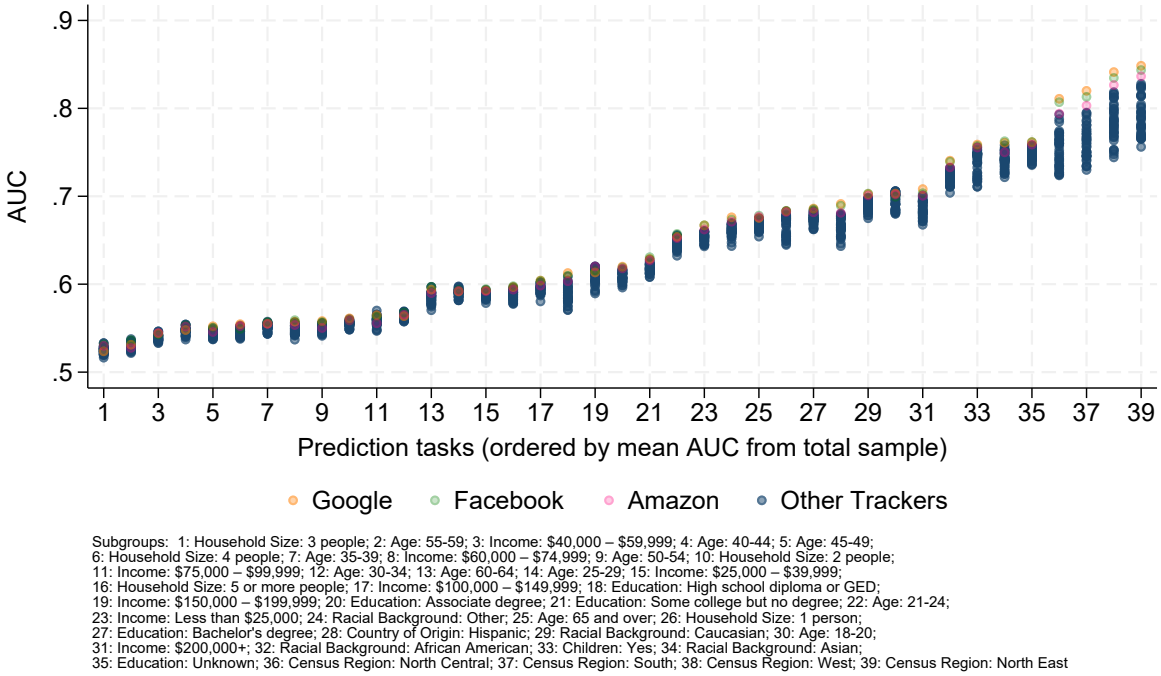



**Note:** The vertical axis denotes performance measured as AUC relative to a hypothetical tracker with access to all tracked users' browsing histories. The horizontal axis shows the number of tracked domains in 10,000s. The largest tracker firms, Google, Facebook, Amazon, Cloudflare, Bootstrap, and Twitter are labeled on the plot. Individual trackers belonging to Google (e.g. Doubleclick or Google Analytics) are colored in light orange. We perform 5-fold cross-validation to assess each tracker's prediction quality. The scatter dots represent the mean relative AUC over these 5 repetitions. The vertical grey lines connect the mean relative AUC of the upper and the lower folds.

Figure 5 shows the variation in prediction quality across trackers for all prediction tasks. On the horizontal axis, all 39 tasks are sorted in increasing order of the mean total sample

AUC. The vertical axis corresponds to the mean AUC achieved by the trackers.[9] Prediction tasks on the left side of the plot, i.e. those with a low mean AUC of the total sample, display little variation in prediction quality. Moving to the right along the horizontal axis as predictions become more accurate, the variation across trackers increases. For the tasks on the very right of the plot, GFA perform markedly better than the smaller firms.

Figure 5: Distribution of AUC across trackers and prediction tasks




**Note:** This figure plots the distribution of prediction quality measured by the AUC across trackers, for each prediction task. Prediction tasks are ordered along the horizontal axis by the mean AUC achieved using the total sample. The sample is restricted to Google, Facebook, Amazon, and the subsequent 36 largest tracker firms. We exclude individual trackers operated by Google (e.g. Doubleclick or Google Analytics) or Amazon (e.g. Amazon Cloudfront or Amazon Web Services) here.

## 5.2 Prediction returns to data

One focus of our analysis is the characterization of returns to data for prediction quality. If returns to data decrease slowly, or not at all, then superior data access can represent a substantial barrier to entry in the web tracking market. We have shown that large firms achieve better prediction results for some demographic prediction tasks. However, the available observational data are insufficient to fully assess the returns to data because we only observe cross-sectional variation across trackers. Large firms such as GFA track almost all domains and all users in the sample period. Even for smaller trackers, the variation in the number of domains and users

[9] We show only the combined trackers for Google and Amazon here to reduce noise. Figure 4 shows that prediction quality for the individual trackers of these firms does not vary much.

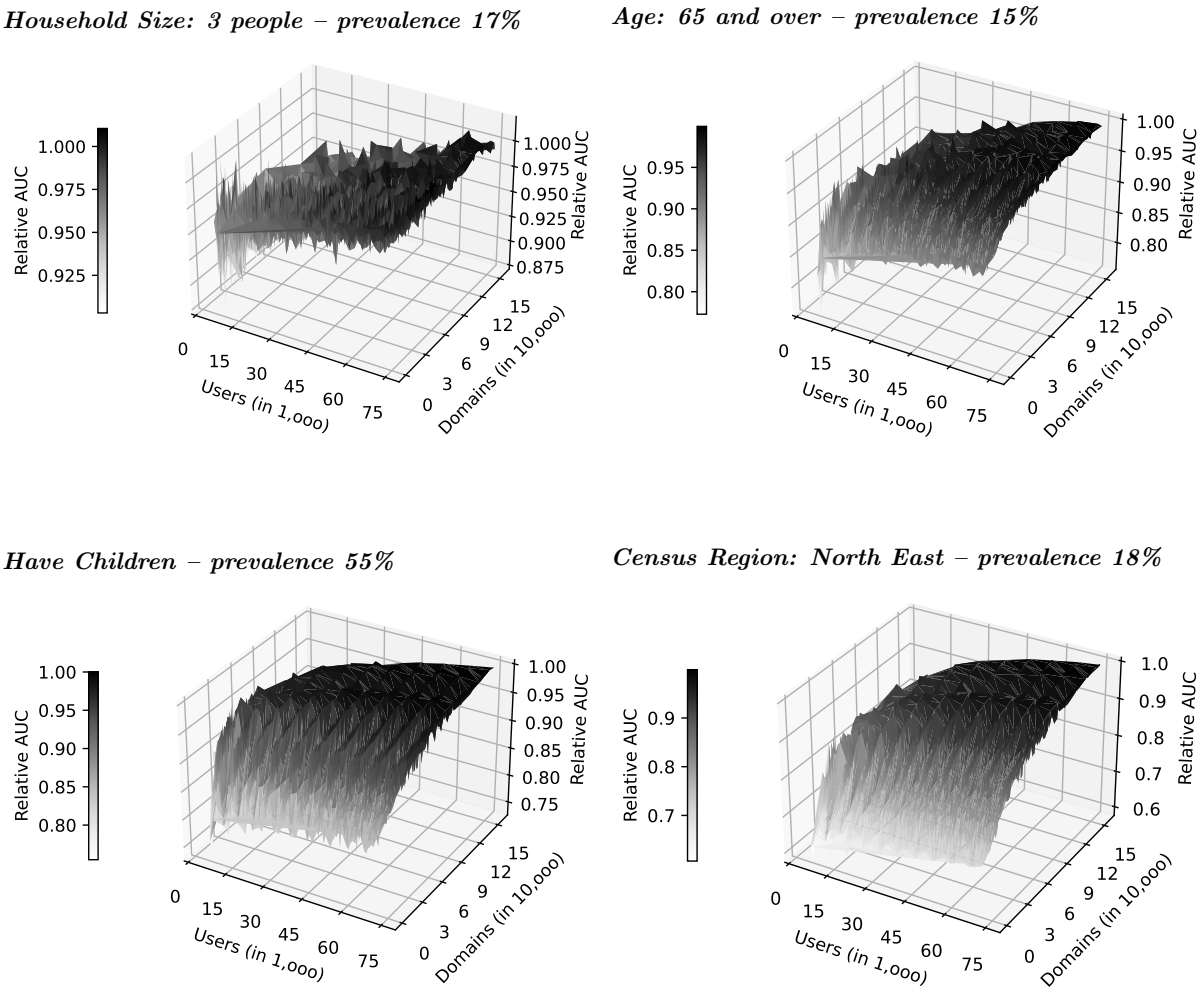
over time is limited. Yet, to identify the relationship between prediction quality and the two data dimensions, we require tracker-specific variation over larger ranges of data inputs.

Therefore, we generate synthetic data that guarantees sufficient variation to assess global returns to data. To generate the data, we vary the scale and scope of data on a 10-by-10 grid for $N$ and $K$. On the grid, we drop fractions of 0%, 10%, 20%, ..., 90% of randomly chosen users and a tracker's observed domains. We create one grid for each tracker and prediction task. Appendix C provides an in-depth description of how we construct the simulated data and a graphical illustration of the grid for all trackers and prediction tasks.

We then apply the approach described in Section 4 to the simulated rather than the observational data. Here, we train the LightGBM classification algorithm for every prediction task, tracker, and grid point. We collect the AUC values, computed relative to the mean AUC using the complete clickstream data, from all five folds of the cross-validated model in each iteration. Collecting prediction qualities for all trackers and prediction tasks at all grid points provides a new data set mapping the varying tracker-specific data into prediction qualities.

Figure 6 shows the prediction qualities over all trackers and grid points for the same four different prediction tasks also displayed in figure 4. Prediction quality is plotted on the vertical axis and the number of tracked users and domains are shown on the two axes spanning the floor. In line with existing studies, we find globally decreasing returns to data (Bajari et al., 2019; Peukert et al., 2023; Yoganarasimhan, 2020; Schaefer and Sapi, 2023). We observe these decreasing returns both across and within trackers.

Figure 6: Returns to data over the entire grid




While we observe significant increases in prediction quality starting at low levels of tracked users and domains, the relative AUC improves only slowly at higher levels, indicating rather strong diminishing returns. This is even more true for the number of users than for the number of domains, especially for tasks that are difficult to predict. Hence, even firms observing only a small share of users' browsing records may be able to predict users' demographic characteristics well. Large trackers such as Google and Doubleclick virtually always attain the same mean AUC as a hypothetical tracker with access to all clickstream data.[10]

## 5.3 Data complementarities: combining clickstream and demographic data

We have established that trackers face decreasing returns to data – both in the number of domains and users tracked – when predicting user characteristics based on users' browsing behavior. Yet, large digital platforms can also collect a variety of additional data on their own

[10] To systematically estimate the returns to data along both dimensions, we run polynomial regressions on our prediction result data; see Appendix F.

users. For example, Meta observes detailed personal information for individuals who actively use their Facebook or Instagram platforms.[11] Likewise, Google observes personal information in their search, maps, or e-mail services, and Amazon observes purchase preferences in its user profiles.[12] Because large platforms directly observe these data across the services they offer, they can combine personal demographic information and off-platform browsing data to infer additional missing personal information.

Motivated by this practice, we investigate the extent to which the large GFA trackers can improve their predictive accuracy by combining web browsing data with additional and potentially complementary data on demographic characteristics. To do so, we use the same algorithm as in Section 5 but expand the data available to Google, Facebook, and Amazon with a subset of each user's demographic information.[13] Specifically, for all prediction tasks of these three trackers, we use all demographic variables as predictors excluding the user characteristic we aim to predict. For instance, for the task of classifying users into the group "Age: 65 and over", we train and test the model based on the web browsing data and all demographic information except the user's age.

Repeating an analysis analogous to Section 5.1, Figure 7 shows the classification results across trackers for all prediction tasks. The results are identical to Figure 5 for all trackers, using the same clickstream data as before, except for GFA. These three platforms experience large jumps in prediction quality across prediction tasks due to the combination of clickstream data with demographic information in their predictive models. We estimate an average increase in prediction quality, measured by the AUC, of 0.065 (10.17%) for these three firms across all tasks.

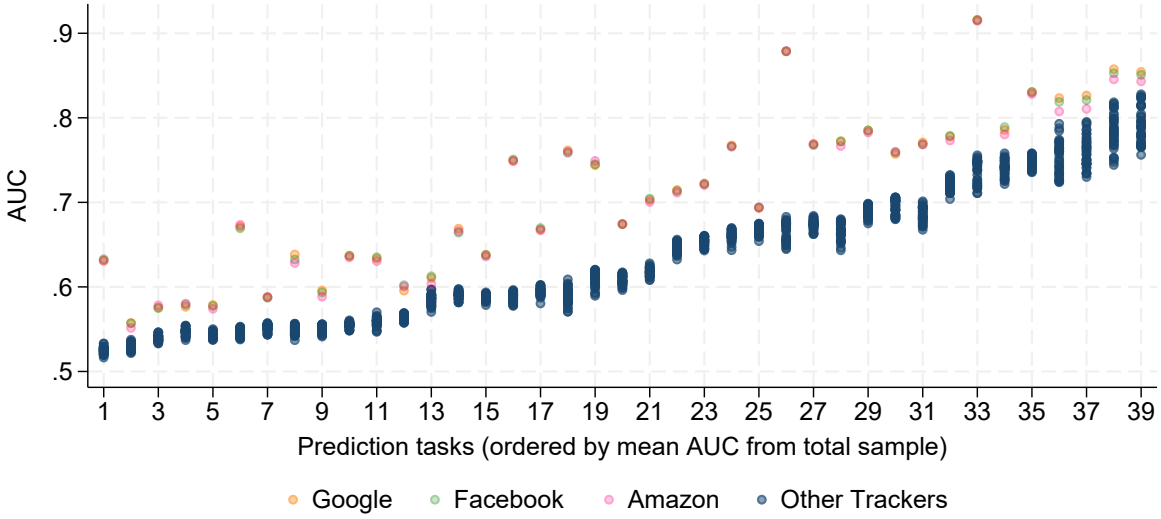
Given our previous finding of diminishing returns to data, the significant increments in prediction quality obtained by adding external demographic data have important implications. Table 7 in Appendix F reports coefficients from linear regressions of prediction quality on the scale and scope of data. The results in column (1) show that trackers can improve their AUC on average by a maximum of 0.066 when increasing the number of tracked domains from 0 to 101,250, which is the maximum number of tracked domains beyond which, ceteris paribus, returns start to decrease. This improvement is equivalent to the jump in prediction quality experienced by GFA when combining external demographic information with their web browsing

[11] For instance, Meta asks users for their date of birth and uses this information for advertising purposes: https://www.meta.com/en-gb/help/quest/articles/accounts/account-settings-and-management/why-we-ask-for-your-birthday/

[12] Cyphers and Gebhart (2019) provide a detailed account of the various sources of information used by big tech firms such as Google, Meta, and Amazon for tracking purposes.

[13] We use the LightGBM algorithm to minimize the binary logistic loss function and evaluate predictions using 5-fold cross-validation.

Figure 7: Distribution of AUC across trackers and prediction tasks with additional data




Subgroups: 1: Household Size: 3 people; 2: Age: 55-59; 3: Income: $40,000 – $59,999; 4: Age: 40-44; 5: Age: 45-49;
6: Household Size: 4 people; 7: Age: 35-39; 8: Income: $60,000 – $74,999; 9: Age: 50-54; 10: Household Size: 2 people;
11: Income: $75,000 – $99,999; 12: Age: 30-34; 13: Age: 60-64; 14: Age: 25-29; 15: Income: $25,000 – $39,999;
16: Household Size: 5 or more people; 17: Income: $100,000 – $149,999; 18: Education: High school diploma or GED;
19: Income: $150,000 – $199,999; 20: Education: Associate degree; 21: Education: Some college but no degree; 22: Age: 21-24;
23: Income: Less than $25,000; 24: Racial Background: Other; 25: Age: 65 and over; 26: Household Size: 1 person;
27: Education: Bachelor's degree; 28: Country of Origin: Hispanic; 29: Racial Background: Caucasian; 30: Age: 18-20;
31: Income: $200,000+; 32: Racial Background: African American; 33: Children: Yes; 34: Racial Background: Asian;
35: Education: Unknown; 36: Census Region: North Central; 37: Census Region: South; 38: Census Region: West; 39: Census Region: North East

**Note:** This figure plots the distribution of prediction quality measured by the AUC across trackers, for each prediction task. We assume that Google, Facebook, and Amazon combine clickstream data with demographic data for their predictions while all other trackers only use clickstream data. Prediction tasks on the horizontal axis are sorted in increasing order of the mean AUC achieved using the total tracking sample. The sample is restricted to Google, Facebook, Amazon, and the subsequent 36 largest tracker firms. We exclude individual trackers operated by Google (e.g. Doubleclick or Google Analytics) or Amazon (e.g. Amazon Cloudfront or Amazon Web Services) here.

data. Thus, even if smaller firms significantly expand their tracking activity by collecting more browsing data on domains and users, decreasing returns will prevent them from achieving the same prediction quality as firms that can combine information from user profiles with clickstream data.

# 6 Evaluating regulatory interventions

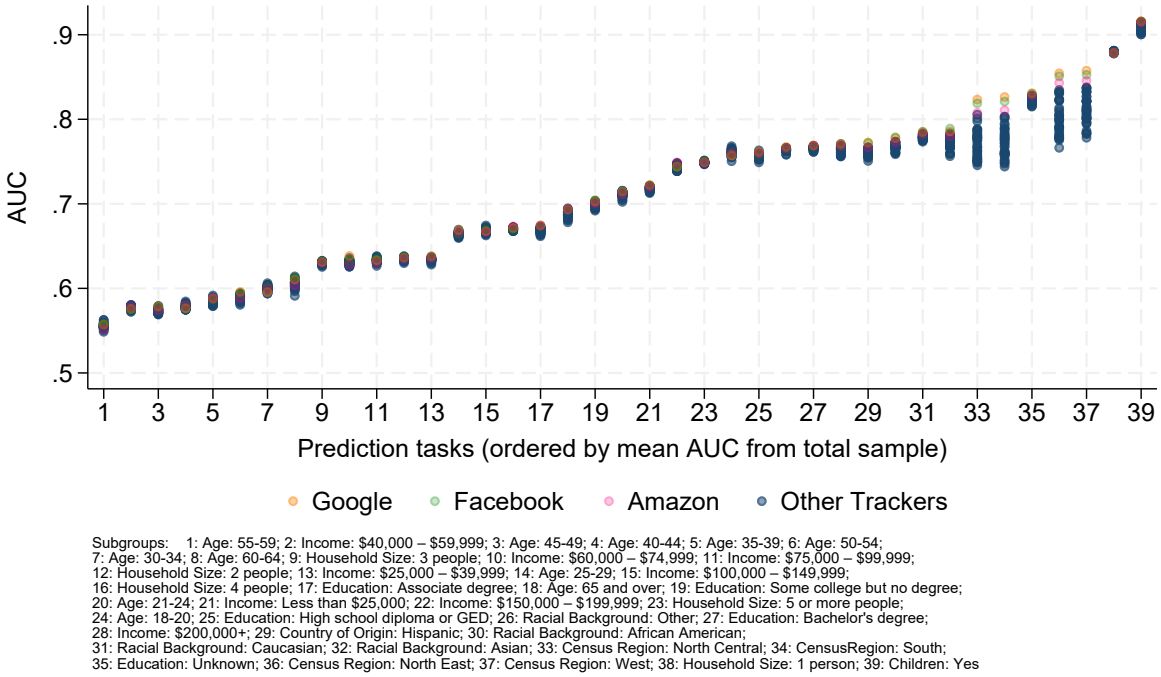
## 6.1 Ex-Ante: Data sharing obligations in the DMA

The DMA is a regulation by the European Union aimed at creating a fairer competitive landscape in digital markets. The DMA targets large online platforms who control access to certain services or markets, and as of September 2023, the EU had identified six gatekeeper companies.[14] The DMA aims to limit their power and ensure a level playing field for smaller firms. Among several obligations, it requires the gatekeepers to share (certain) data that they collect with other competitors, upon request.

[14] Alphabet, Amazon, Apple, ByteDance, Meta, and Microsoft.

In this subsection, we perform a counterfactual analysis in which we assume all trackers obtain access to the demographic information of their tracked users. This scenario could occur if, for instance, Alphabet, identified as a gatekeeper under the DMA, was to be mandated to share part of its user data with tracker firms competing on the market for targeted advertisements. Following the procedure described in Section 5.3, we simulate this scenario by repeating the prediction tasks for all 52 trackers while combining users' demographic information to each tracker's available clickstream data.Figure 8 shows the resulting distribution of prediction qualities across trackers.

Figure 8: Distribution of AUC across trackers and prediction tasks (everyone using additional demographic data)




**Note:** This figure plots the distribution of prediction quality measured by the AUC across trackers, for each prediction task. We assume that every tracker combines clickstream data with demographic data for its predictions. Prediction tasks on the horizontal axis are sorted in increasing order of the mean AUC achieved using the total tracking sample. The sample is restricted to Google, Facebook, Amazon, and the subsequent 36 largest tracker firms. We exclude individual trackers operated by Google (e.g. Doubleclick or Google Analytics) or Amazon (e.g. Amazon Cloudfront or Amazon Web Services) here.

To more precisely quantify how prediction quality improves relative to a baseline in which trackers do not have access to additional user data, we specify a linear regression model of the form:

$$Y_{isfp} = \alpha + \beta_1 demographics_p + \gamma X_{isfp} + \varepsilon_{isfp}, \tag{1}$$

where the outcome variable $Y_{isfp}$ is the AUC for tracker $i$, prediction task $s$, fold $f$, and prediction model $p$, which includes one prediction model with and one without additional demo-

graphic data. The variable of interest, *demographics* is an indicator equal to one for predictions based on clickstream data combined with demographic data and equal to zero for predictions based on clickstream data only. The vector $X$ includes fold, tracker, and prediction task fixed effects.

From an antitrust perspective, an important question regarding data-sharing provisions is whether such policies can reduce the distance between each tracker and the incumbent gatekeepers. To explore how access to demographic data would impact trackers' distance to Google – the largest tracker in our sample – we specify the outcome in equation (1) as $AUC_{isfp}/AUC_{google_{sfp}}$, the ratio between the AUCs of tracker $i$ and Google for prediction task $s$, fold $f$, and prediction model $p$, where Google is assumed to use both clickstream and demographic data for prediction.

The first column in Table 2 shows that combining external data with clickstream data creates sizable improvements in prediction quality across trackers. We estimate that, on average, the AUC increases by .07 (or around 11%) when demographic data is added to the LightGBM prediction models. The second column shows that providing non-GFA trackers with access to demographic data significantly reduces the gap in prediction quality relative to Google, compared to a world where non-GFA trackers only have access to clickstream data. More specifically, competing tracker firms become 9.9 percentage points closer to Google's prediction quality if able to rely on demgraphic data, on average. These results suggest that the DMA's data-sharing requirements can reduce gatekeepers' competitive advantage by improving prediction quality for smaller competitors.

Table 2: Effect of data combination on prediction quality

| VARIABLES | AUC | AUC relative to Google |
|---|---|---|
| Adding Demographic Data | 0.07009*** | 0.09880*** |
| | (0.00765) | (0.00052) |
| Constant | 0.63578*** | 0.89003*** |
| | (0.00382) | (0.00037) |
| Observations | 20,280 | 14,040 |
| R-squared | 0.93 | 0.77 |
| Tracker FE | Yes | Yes |
| Task FE | Yes | Yes |
| Fold FE | Yes | Yes |
| Cluster | Task | Task |

**Note:** The unit of observation is at the tracker-task-fold-prediction model level in all regression models. The dependent variable in column (1) is the in AUC and in column (2) is the AUC relative to Google's. Robust standard errors in parentheses. *** represent p<0.01.

### 6.2 Ex-Post: GDPR-induced changes in tracking and prediction quality

In 2018, the European Commission introduced the GDPR to harmonize data protection law and enforcement throughout the European Union. Under the GDPR, websites are held accountable for privacy breaches, even if third-party technology is used to collect data, because both websites and web technology providers act as so-called "joint controllers" of user data.[15] Joint controllership applies even if the website cannot control what data is collected by a tracker. Furthermore, the GDPR drastically increased possible fines for privacy violations, which can now reach €20 million or 4% of the total worldwide annual turnover, whichever is higher.[16] As a result, after the GDPR became effective on 25 May 2018, websites reduced their compliance risks by reducing the number of third-party web technology providers they use (Peukert et al., 2022).

In this subsection, we assess how prediction quality across web trackers changed after the introduction of the GDPR. For this, we utilize the data from after the introduction of the GDPR and the first 20 calendar weeks of 2018 before the GDPR became effective.[17] Table 1 in Section 3 shows the reduction in tracking activity levels after the introduction of the GDPR. While the number of visited domains remained relatively stable over the pre- and post-GDPR period, the number of tracked domains decreased from 172,120 to 155,247 in our sample, which confirms that the GDPR reduced tracking activity levels in line with previous research. We restrict the sample to the 48 trackers that appear in both the pre- and post-GDPR periods and run the machine learning prediction exercise described in Section 4. We then measure the change in prediction qualities after the introduction of the GDPR using the linear regression model

$$Y_{isfpt} = \alpha + \beta_1 post_t + \gamma X_{isft} + \varepsilon_{isfpt}, \tag{2}$$

where the outcome variable $Y$ is the AUC for tracker $i$, prediction task $s$, fold $f$, and prediction model $p$ in period $t$ defined as either pre- or post-GDPR. The variable of interest, *post* is an indicator equal to one for predictions based on data of the post-GDPR period. The vector $X$ includes tracker, prediction task, and fold fixed effects.

Table 3 shows the results of regressions on the full sample as well as the sub-samples of GFA and non-GFA trackers. In the first column, overall prediction quality decreases slightly by -0.0023 with GDPR. Columns 2 and 3 show the prediction quality of non-GFA trackers decreases significantly less compared to GFA trackers, if all trackers used prediction models based only

[15] See GDPR Art. 26.
[16] See GDPR Art. 83 (5) and (6).
[17] GDPR came into force on 25 May 2018, which lies in the 21st calendar week of 2018.

on clickstream data. However, if GFA can also use demographic features in their prediction models, the decrease is attenuated, as shown in the last two columns. Interestingly, in this case, prediction quality for GFA trackers decreases less than for the non-GFA trackers that do not have access to demographic data. Finally, in a counterfactual scenario in which all trackers have access to demographic variables for their prediction models, as described in Section 6.1, the decrease in prediction quality is the smallest. Hence, access to data that is more difficult to obtain for smaller trackers at the same time appears the most valuable for these small trackers.

Table 3: Prediction quality and the introduction of GDPR

| | Clickstream data | | | Clickstream and demographic data | |
|---|---|---|---|---|---|
| | All Trackers | Non-GFA | GFA | Non-GFA | GFA |
| Post-GDPR | -0.00233*** | -0.00199*** | -0.00309*** | -0.00060*** | -0.00139*** |
| | (0.00014) | (0.00015) | (0.00020) | (0.00012) | (0.00016) |
| Constant | 0.63888*** | 0.63495*** | 0.64752*** | 0.70485*** | 0.71213*** |
| | (0.00010) | (0.00010) | (0.00014) | (0.00008) | (0.00012) |
| Observations | 18,720 | 12,870 | 5,850 | 12,870 | 5,850 |
| R-squared | 0.99 | 0.99 | 0.99 | 0.99 | 0.996 |
| Tracker FE | Yes | Yes | Yes | Yes | Yes |
| Task FE | Yes | Yes | Yes | Yes | Yes |
| Fold FE | Yes | Yes | Yes | Yes | Yes |

**Note:** The unit of observation is on a tracker-task-fold-period level in all regression models. The sample is restricted to the 48 trackers that appear in both the pre- and post-GDPR periods. The dependent variable is the AUC. GFA trackers include Google, Facebook, and Amazon, as well as the individual trackers belonging to Google or Amazon (see Table 5). Robust standard errors in parentheses. *** represent p<0.01.

There can be multiple explanations for why some trackers, especially larger ones such as GFA, are more strongly affected by the introduction of the GDPR than others. To shed light on potential determinants, we assess the association between the change in prediction quality (AUC) pre- and post-GDPR and possible observable factors using a regression analysis at the tracker-task-fold level. The results of this analysis reported in Table 4 show that trackers that trace fewer domains post-GDPR have a greater decrease in prediction quality. The coefficient on the magnitude of a tracker's prediction quality before the GDPR is negative, large, and significant. This is driven by the nature of binary classification, where the AUC bounded from below at 0.5. Hence, for trackers with a high AUC before GDPR prediction quality can drop much more than for those with low AUC values before AUC. Finally, we find that trackers with an EU top-level tracker domain and the number of tracked domains from an EU country, which exposes trackers more to GDPR enforcement, are associated with a stronger decrease in prediction quality after the introduction of GDPR, as one would expect.

Table 4: Potential factors behind changes in prediction quality with GDPR

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Change in Number of Domains | 0.00103*** | | | |
| | (0.00010) | | | |
| Pre-GDPR AUC | | -0.32892*** | | |
| | | (0.01127) | | |
| Number of EU Top-Level tracker Domains | | | -0.00009*** | |
| | | | (0.00002) | |
| Number of tracked EU Domains (1,000) | | | | -0.00048*** |
| | | | | (0.00003) |
| Constant | -0.00032 | 0.20780*** | -0.00225*** | 0.00021 |
| | (0.00023) | (0.00721) | (0.00009) | (0.00021) |
| Observations | 9,360 | 9,360 | 9,360 | 9,360 |
| R-squared | 0.46 | 0.62 | 0.45 | 0.46 |
| Tracker FE | | Yes | | |
| Task FE | Yes | Yes | Yes | Yes |
| Fold FE | Yes | Yes | Yes | Yes |

**Note:** The unit of observation is on a tracker-task-fold level in all regression models. The sample is restricted to the 48 trackers that appear in both the pre- and post-GDPR periods. The dependent variable in all regression models is the change in AUC between the post- and pre-GDPR period. Robust standard errors in parentheses. *** represent $p<0.01$.

# 7 Conclusion

In this paper, we characterize the relationship between the prediction quality of consumer profiling and the data web tracking firms can use in their prediction models. Our results indicate a strong variation in prediction quality across prediction tasks. For some demographic groups, web browsing data exhibits only little predictive power. In these instances, larger firms benefit only marginally, if at all, from accessing richer web browsing data. However, for prediction tasks where web browsing data yields higher prediction quality, larger firms with access to more data have an advantage over smaller firms with less data.

Overall, we find decreasing returns to data on the number of users and tracked websites across and within tracker firms. While this result mitigates competition concerns at first glance, we document that combining web browsing data with additional consumer data, represented here by the example of a subset of demographic information, yields substantial increases in prediction quality. The three largest tracker firms Google, Facebook, and Amazon operate digital ecosystems that enable them to combine web browsing data with further user-level information from multiple sources. We find an average increase in prediction quality, measured by AUC, of 0.065 (10.17%) for these three firms across all subgroups when we enrich their prediction model with demographics. This is comparable to the increase a tracker would obtain, on average, when tracking over 100,000 additional domains. In this light, the decreasing returns

to data may actually have prevented smaller web tracking firms from ever catching up with GFA.

The results of our study have significant implications for policymakers, particularly in the context of data privacy and competition regulation. We analyze the impact of regulatory interventions on the prediction quality and competitive landscape in the web tracking market. Our findings suggest that regulatory measures like the GDPR, while enhancing consumer privacy, may inadvertently consolidate the market power of large tech firms. By reducing the amount of trackable data, the GDPR disproportionately affects smaller firms that cannot compensate for these data losses with complementary demographic data. While GFA incurs more severe losses in prediction quality compared to the other firms when only using web browsing data, this finding is reversed in a scenario where GFA can combine data and the other firms cannot. Our analysis also indicates that data-sharing mandates, such as those included in the DMA, could mitigate the competitive disadvantages faced by smaller firms. By requiring large incumbents to share certain data, these provisions could help foster a more competitive and dynamic market environment. Our simulation results show that access to demographic data can significantly narrow the gap in prediction quality between smaller trackers and large tech firms like Google.

While data-sharing provisions may be a solution to level the playing field in markets that rely on consumer profiles for targeted services, they may come at a cost to consumer privacy. Policymakers need to consider this trade-off when designing privacy and antitrust policies, for example by adopting a privacy-preserving approach to data access. Indeed, the DMA requires gatekeepers to publish general terms and conditions for access on fair, reasonable and non-discriminatory terms, including dispute resolution mechanisms. These terms should protect users' privacy and prevent misuse of personal data.

## References


Aguiar, L., Claussen, J., and Peukert, C. (2018). "Catch me if you can: Effectiveness and consequences of online copyright enforcement." *Information Systems Research*, *29*(3), 656–678. 8

Aguiar, L., Peukert, C., Schäfer, M., and Ullrich, H. (2023). "Off-platform tracking and data externalities." *Working Paper*. 3, 8

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. (2019). "The impact of big data on firm performance: An empirical investigation." *AEA Papers and Proceedings*, *109*, 33–37. 3, 14

Calvano, E., and Polo, M. (2021). "Market power, competition and innovation in digital markets: A survey." *Information Economics and Policy*, *54*, 100853. 3, 7

Campbell, J., Goldfarb, A., and Tucker, C. (2015). "Privacy regulation and market structure." *Journal of Economics & Management Strategy*, *24*(1), 47–73. 7

Chen, Z., Choe, C., Cong, J., and Matsushima, N. (2022). "Data-driven mergers and personalization." *The RAND Journal of Economics*, *53*(1), 3–31. 1

Cyphers, B., and Gebhart, G. (2019). "Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance." Tech. rep., Electronic Frontier Foundation. 16

Dambra, S., Sanchez-Rola, I., Bilge, L., and Balzarotti, D. (2022). "When sally met trackers: Web tracking from the users' perspective." In *31st USENIX Security Symposium (USENIX Security 22)*, 2189–2206. 8

De Cnudde, S., Martens, D., Evgeniou, T., and Provost, F. (2020). "A benchmarking study of classification techniques for behavioral data." *International Journal of Data Science and Analytics*, *9*(2), 131–173. 5

De Cnudde, S., Martens, D., Evgeniou, T., and Provost, F. (2020). "A benchmarking study of classification techniques for behavioral data." *International Journal of Data Science and Analytics*, *9*(2), 131–173. 5, 11

Economides, N., and Lianos, I. (2021). "Restrictions on privacy and exploitation in the digital economy: A market failure perspective." *Journal of Competition Law & Economics*, *17*(4), 765–847. 1

Goldfarb, A., and Que, V. F. (2023). "The economics of digital privacy." *Annual Review of Economics*, *15*, 267–286. 1

Gregory, R. W., Henfridsson, O., Kaganer, E., and Kyriakou, H. (2021). "The role of artificial intelligence and data network effects for creating user value." *Academy of Management Review*, *46*(3), 534–551. 3

Hagiu, A., and Wright, J. (2023). "Data-enabled learning, network effects and competitive advantage." *RAND Journal of Economics*, *forthcoming*. 1, 3

Johnson, G. A., Shriver, S. K., and Du, S. (2020). "Consumer privacy choice in online advertising: Who opts out and at what cost to industry?" *Marketing Science*, *39*(1), 33–51. 1

Johnson, G. A., Shriver, S. K., and Goldberg, S. G. (2023). "Privacy and market concentration: intended and unintended consequences of the gdpr." *Management Science*, forthcoming. 1, 7

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). "Lightgbm: a highly efficient gradient boosting decision tree." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 3149–3157, Red Hook, NY, USA: Curran Associates Inc. 10

Kira, B., Sinha, V., and Srinivasan, S. (2021). "Regulating digital ecosystems: bridging the gap between competition policy and data protection." *Industrial and Corporate Change*, *30*(5), 1337–1360. 7

Klein, T., Kurmangaliyeva, M., Prüfer, J., Prüfer, P., and Park, N. N. (2023). *How Important are User-generated Data for Search Result Quality?: Experimental Evidence*. Centre for Economic Policy Research. 4, 11

Lei, X., Chen, Y., and Sen, A. (2023). "The value of external data for digital platforms: Evidence from a field experiment on search suggestions." *Available at SSRN*. 3

Lerner, A., Kornfeld Simpson, A., Kohno, T., and Roesner, F. (2016). "Internet Jones and the Raiders of the Lost Tracker: An Archaeological Study of Web Tracking from 1996 to 2016." *Proceedings of the 25th USENIX Security Symposium*. 5

Luca, M., Nagaraj, A., and Subramani, G. (2023). "Getting on the map: The impact of online listings on business performance." Tech. rep., National Bureau of Economic Research. 3

Neumann, N., Tucker, C. E., and Whitfield, T. (2019). "How effective is third-party consumer profiling? evidence from field studies." *Marketing Science*, *38*(6), 918–926. 1, 3, 6

Peukert, C., Bechtold, S., Batikas, M., and Kretschmer, T. (2022). "Regulatory spillovers and data governance: Evidence from the gdpr." *Marketing Science*, *41*(4), 746–768. 1, 3, 7, 20

Peukert, C., Sen, A., and Claussen, J. (2023). "The editor and the algorithm: Recommendation technology in online news." *Management Science*, *forthcoming*. 2, 3, 4, 14

Rafieian, O., and Yoganarasimhan, H. (2021). "Targeting and privacy in mobile advertising." *Marketing Science*, *40*(2), 193–218. 1

Schaefer, M., and Sapi, G. (2023). "Complementarities in learning from data: Insights from general search." *Information Economics and Policy*, *65*, 101063. 2, 3, 4, 14

Shiller, B. R. (2020). "Approximating purchase propensities and reservation prices from broad consumer tracking." *International Economic Review*, *61*(2), 847–870. 4, 5

Trusov, M., Ma, L., and Jamal, Z. (2016). "Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting." *Marketing Science*, *35*(3), 405–426. 3, 5, 8, 11

Tucker, C. (2019). "Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility." *Review of Industrial Organization*, *54*(4), 683–694. 1

Yoganarasimhan, H. (2020). "Search personalization using machine learning." *Management Science*, *66*(3), 1045–1070. 14

# A Tracker descriptives

Table 5: Tracker descriptive statistics

| Tracker | Users | Domains | Owned by Google | Owned by Amazon |
|---|---|---|---|---|
| Google* | 75,149 | 155,160 | X | . |
| Google Analytics | 75,138 | 153,387 | X | . |
| Googleapis.Com | 75,144 | 151,971 | X | . |
| Gstatic | 75,121 | 150,713 | X | . |
| Google (Individual Tracker) | 74,891 | 148,602 | X | . |
| Doubleclick | 75,109 | 148,395 | X | . |
| Google Tag Manager | 74,855 | 137,142 | X | . |
| Facebook | 75,032 | 133,285 | . | . |
| Google Syndication | 74,410 | 104,920 | X | . |
| Amazon* | 74,511 | 97,011 | . | X |
| Google Adservices | 74,106 | 93,757 | X | . |
| Youtube | 74,597 | 92,624 | X | . |
| Cloudflare | 74,374 | 90,340 | . | . |
| Bootstrap | 74,363 | 86,153 | . | . |
| Twitter | 74,381 | 77,060 | . | . |
| Amazon Cloudfront | 73,642 | 68,373 | . | X |
| Appnexus | 73,946 | 68,184 | . | . |
| Rubicon | 73,837 | 63,511 | . | . |
| Yahoo | 74,124 | 63,169 | . | . |
| Tradedesk | 74,082 | 63,065 | . | . |
| Adobe Audience Manager | 73,880 | 61,844 | . | . |
| Openx | 73,725 | 61,341 | . | . |
| Google Photos | 74,260 | 61,167 | X | . |
| Pubmatic | 72,853 | 59,420 | . | . |
| Wordpress Stats | 73,175 | 58,077 | . | . |
| Jquery | 73,108 | 57,995 | . | . |
| Amazon Web Services | 72,892 | 56,582 | . | X |
| Addthis | 73,303 | 56,248 | . | . |
| Index Exchange | 72,732 | 55,801 | . | . |
| Yandex | 73,232 | 53,885 | . | . |
| Quantcast | 73,421 | 52,744 | . | . |
| Digicert Trust Seal | 73,895 | 50,274 | . | . |
| Bidswitch | 73,347 | 47,800 | . | . |
| Jsdelivr | 73,400 | 47,651 | . | . |
| Liveramp | 72,809 | 46,547 | . | . |
| Criteo | 72,977 | 45,841 | . | . |
| Advertising.Com | 73,383 | 44,469 | . | . |
| Drawbridge | 73,341 | 43,882 | . | . |
| Bluekai | 72,710 | 42,326 | . | . |
| Mediamath | 73,275 | 42,138 | . | . |
| Aggregate Knowledge | 72,865 | 41,901 | . | . |
| Hotjar | 72,122 | 41,870 | . | . |
| Amazon Associates | 73,295 | 41,022 | . | X |
| New Relic | 73,376 | 40,967 | . | . |
| Fontawesome Com | 73,350 | 39,995 | . | . |
| Scorecard Research Beacon | 73,287 | 39,882 | . | . |
| Media Innovation Group | 72,747 | 38,924 | . | . |
| Lotame | 71,348 | 38,397 | . | . |
| Blogspot Com | 72,466 | 37,031 | . | . |
| Bing Ads | 73,422 | 36,422 | . | . |
| Exelate | 71,871 | 35,786 | . | . |
| Typekit By Adobe | 72,697 | 34,587 | . | . |

**Note:** Trackers marked with a star (*) consist of multiple individual trackers.

# B Subgroup descriptives

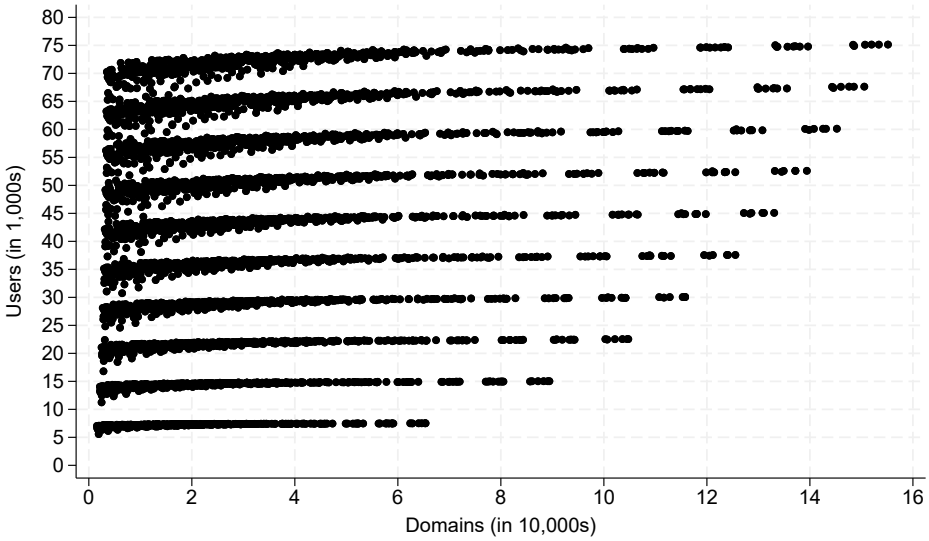
Table 6: Prediction task descriptive statistics

| Prediction task | Mean Y (Mean) | Mean Y (Sd) | AUC (Mean) | AUC (Sd) | AUC (Min) | AUC (Max) | Mean AUC from entire Sample |
|---|---|---|---|---|---|---|---|
| Age: 18–20 | .03 | 0 | .66 | .04 | .42 | .78 | .71 |
| Age: 21–24 | .04 | 0 | .62 | .03 | .45 | .71 | .65 |
| Age: 25–29 | .06 | 0 | .57 | .02 | .45 | .64 | .59 |
| Age: 30–34 | .07 | 0 | .55 | .02 | .43 | .63 | .57 |
| Age: 35–39 | .08 | 0 | .54 | .02 | .43 | .63 | .56 |
| Age: 40–44 | .1 | 0 | .53 | .02 | .43 | .6 | .55 |
| Age: 45–49 | .14 | 0 | .54 | .01 | .45 | .6 | .55 |
| Age: 50–54 | .15 | 0 | .54 | .01 | .46 | .6 | .56 |
| Age: 55–59 | .11 | 0 | .52 | .02 | .42 | .6 | .53 |
| Age: 60–64 | .08 | 0 | .55 | .03 | .41 | .62 | .59 |
| Age: 65 and over | .15 | 0 | .64 | .03 | .5 | .69 | .68 |
| Census Region: North Central | .2 | 0 | .66 | .07 | .45 | .82 | .81 |
| Census Region: North East | .18 | 0 | .7 | .07 | .47 | .86 | .85 |
| Census Region: South | .41 | 0 | .68 | .06 | .5 | .83 | .82 |
| Census Region: West | .21 | 0 | .71 | .07 | .47 | .85 | .84 |
| Children: Yes | .56 | .01 | .71 | .03 | .55 | .77 | .76 |
| Country of Origin: Hispanic | .19 | 0 | .62 | .04 | .46 | .7 | .69 |
| Education: Associate degree | .21 | 0 | .58 | .02 | .46 | .63 | .62 |
| Education: Bachelor's degree | .14 | 0 | .64 | .03 | .48 | .71 | .68 |
| Education: High school diploma or GED | .03 | 0 | .55 | .04 | .36 | .67 | .61 |
| Education: Some college but no degree | .2 | 0 | .59 | .02 | .46 | .64 | .63 |
| Education: Unknown | .4 | .01 | .73 | .02 | .62 | .77 | .76 |
| Household Size: 1 person | .19 | 0 | .64 | .03 | .51 | .7 | .68 |
| Household Size: 2 people | .33 | 0 | .54 | .01 | .47 | .59 | .56 |
| Household Size: 3 people | .17 | 0 | .51 | .01 | .42 | .57 | .53 |
| Household Size: 4 people | .13 | 0 | .53 | .02 | .43 | .59 | .55 |
| Household Size: 5 or more people | .17 | 0 | .56 | .02 | .45 | .61 | .59 |
| Income: $100,000 – $149,999 | .11 | 0 | .57 | .02 | .43 | .62 | .6 |
| Income: $150,000 – $199,999 | .04 | 0 | .56 | .04 | .37 | .68 | .62 |
| Income: $200,000+ | .05 | 0 | .64 | .04 | .45 | .74 | .71 |
| Income: $25,000 – $39,999 | .18 | 0 | .57 | .01 | .49 | .63 | .59 |
| Income: $40,000 – $59,999 | .16 | 0 | .53 | .01 | .45 | .59 | .55 |
| Income: $60,000 – $74,999 | .08 | 0 | .52 | .02 | .38 | .63 | .56 |
| Income: $75,000 – $99,999 | .1 | 0 | .53 | .02 | .41 | .59 | .56 |
| Income: Less than $25,000 | .28 | 0 | .63 | .02 | .52 | .68 | .67 |
| Racial Background: African American | .21 | 0 | .69 | .03 | .56 | .75 | .74 |
| Racial Background: Asian | .06 | 0 | .68 | .05 | .45 | .78 | .76 |
| Racial Background: Caucasian | .56 | 0 | .66 | .02 | .54 | .71 | .7 |
| Racial Background: Other | .18 | 0 | .64 | .02 | .5 | .69 | .68 |
| Total Observations | 1,014,000 | | | | | | |

# C Generating Additional Data using a Grid

To measure the relationship between prediction quality and the two data dimensions, i.e. the number of observed users and the extent of browsing observed per user, we use counterfactual simulations at the tracker-level. For each tracker, given one prediction task, we vary the scale and scope of data on a 10-by-10 grid for $N$ and $K$. Doing so, we respectively drop fractions of 0%, 10%, 20%, ..., 90% of randomly chosen users and domains for each tracker in our sample. Figure 9 displays the variation in the user and domain dimensions obtained through our grid. Compared to the factual distribution of users and domains across trackers shown in figure 3, this simulation creates a much richer dataset in terms of variation in both the user and domain dimensions.

Figure 9: User - Domains distribution across trackers over the entire grid

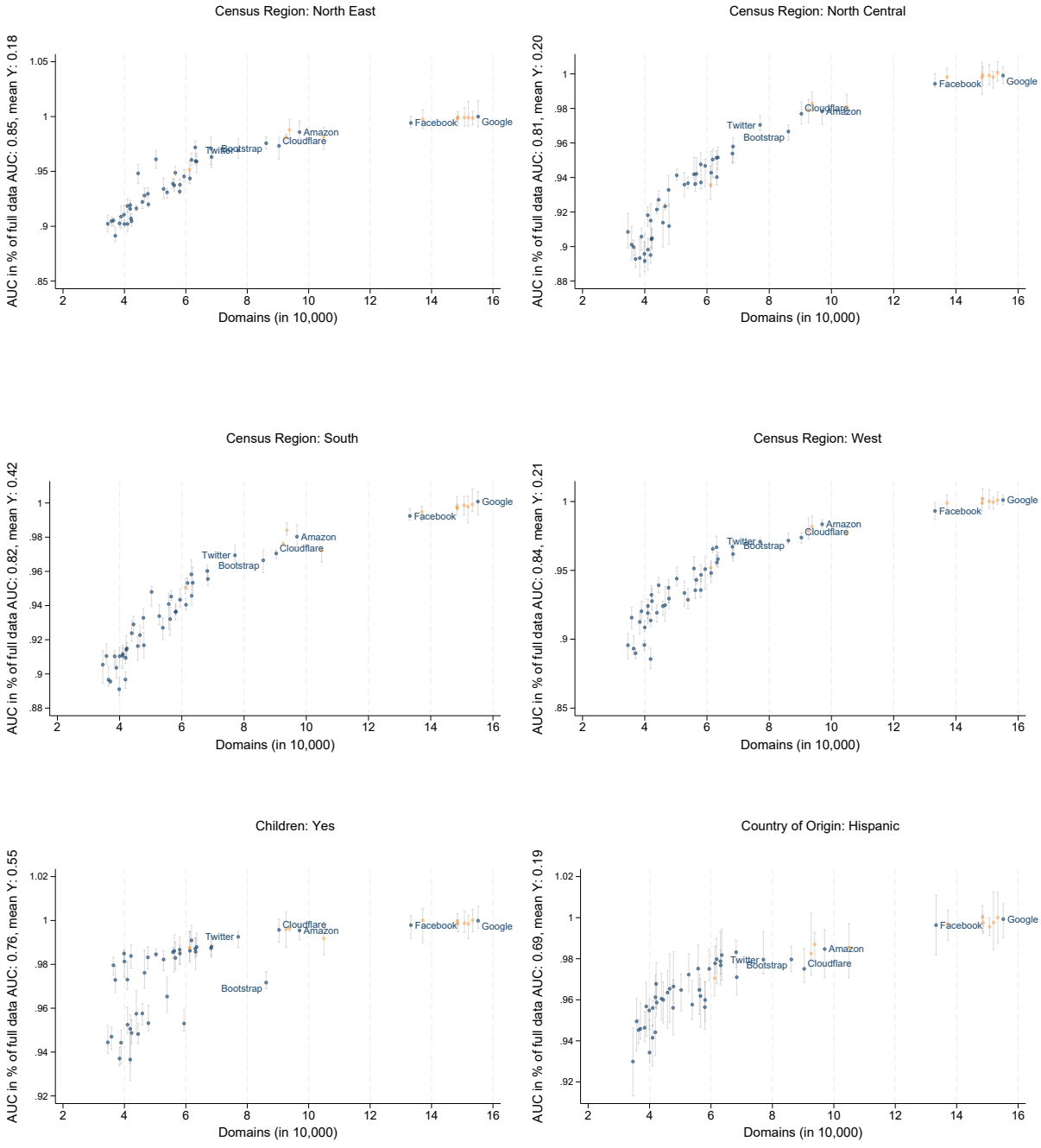


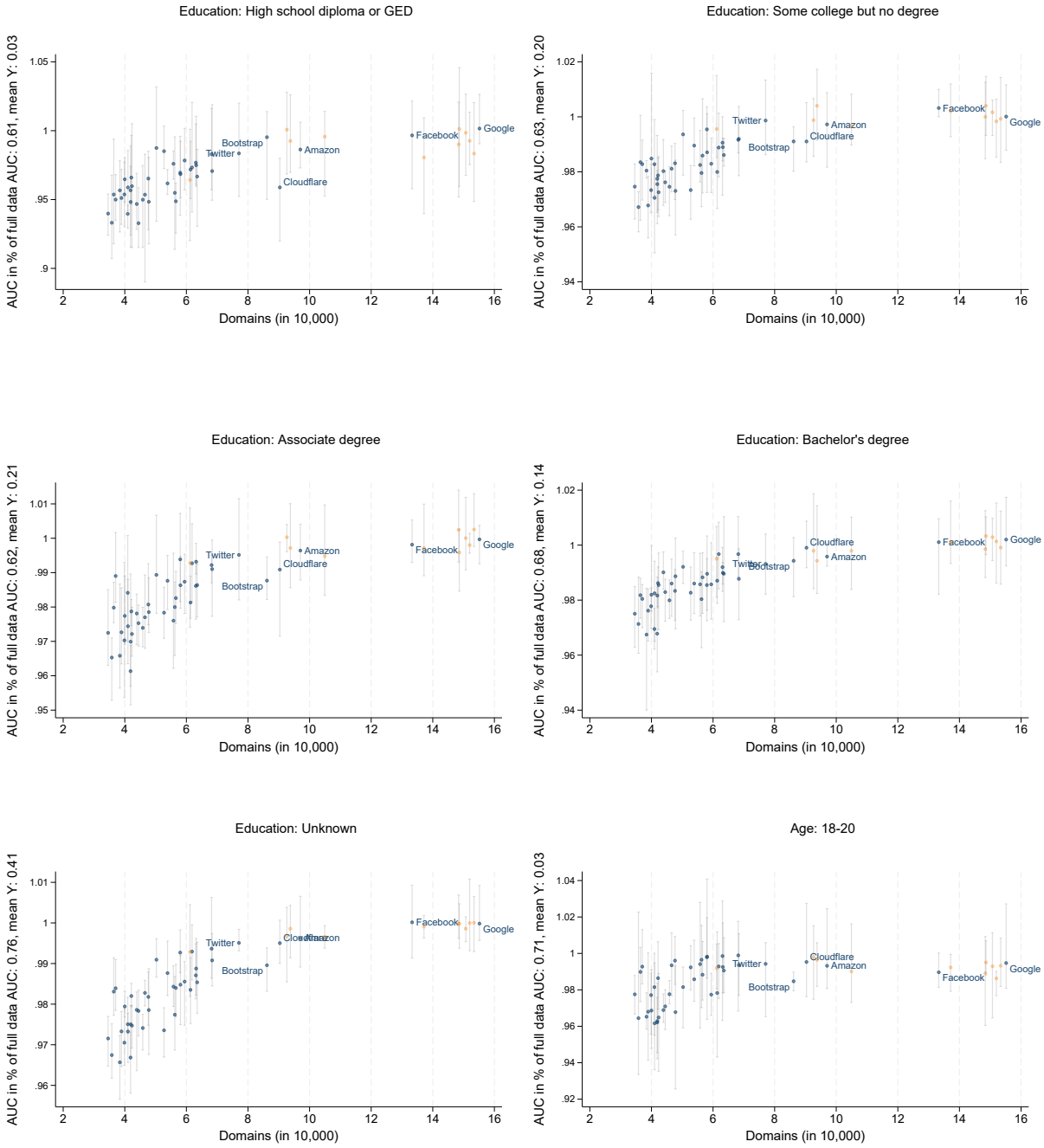

For every prediction task, every tracker and at every gridpoint, we train our LightGBM classification model, as described in the previous subsection. In each iteration, we collect the AUCs from all 5-folds of the cross-validated model. We evaluate these AUCs relative to the mean AUC obtained from the total tracked clickstream data for the respective prediction task. By collecting the achieved prediction quality for all trackers, at all grid points and all prediction tasks, we construct a new data set mapping the varying tracker-specific amount of data into prediction quality.

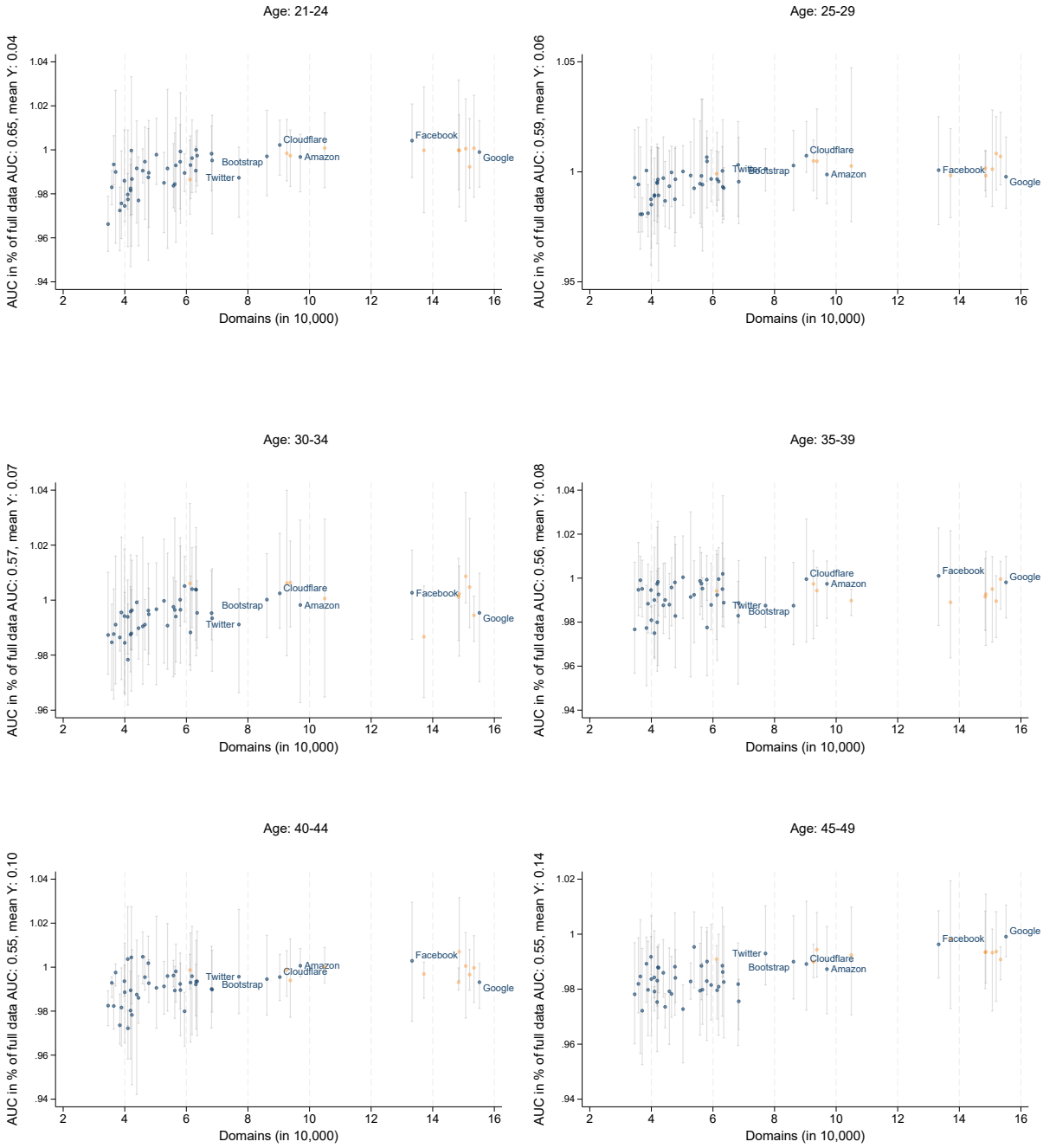
Table 6 reports summary statistics of this data set. The first column "Subgroup" indicates the prediction task, while the second and third columns indicate the prevalence and its standard deviation of that subgroup. Columns 4-7 display summary statistics about the AUC achieved by

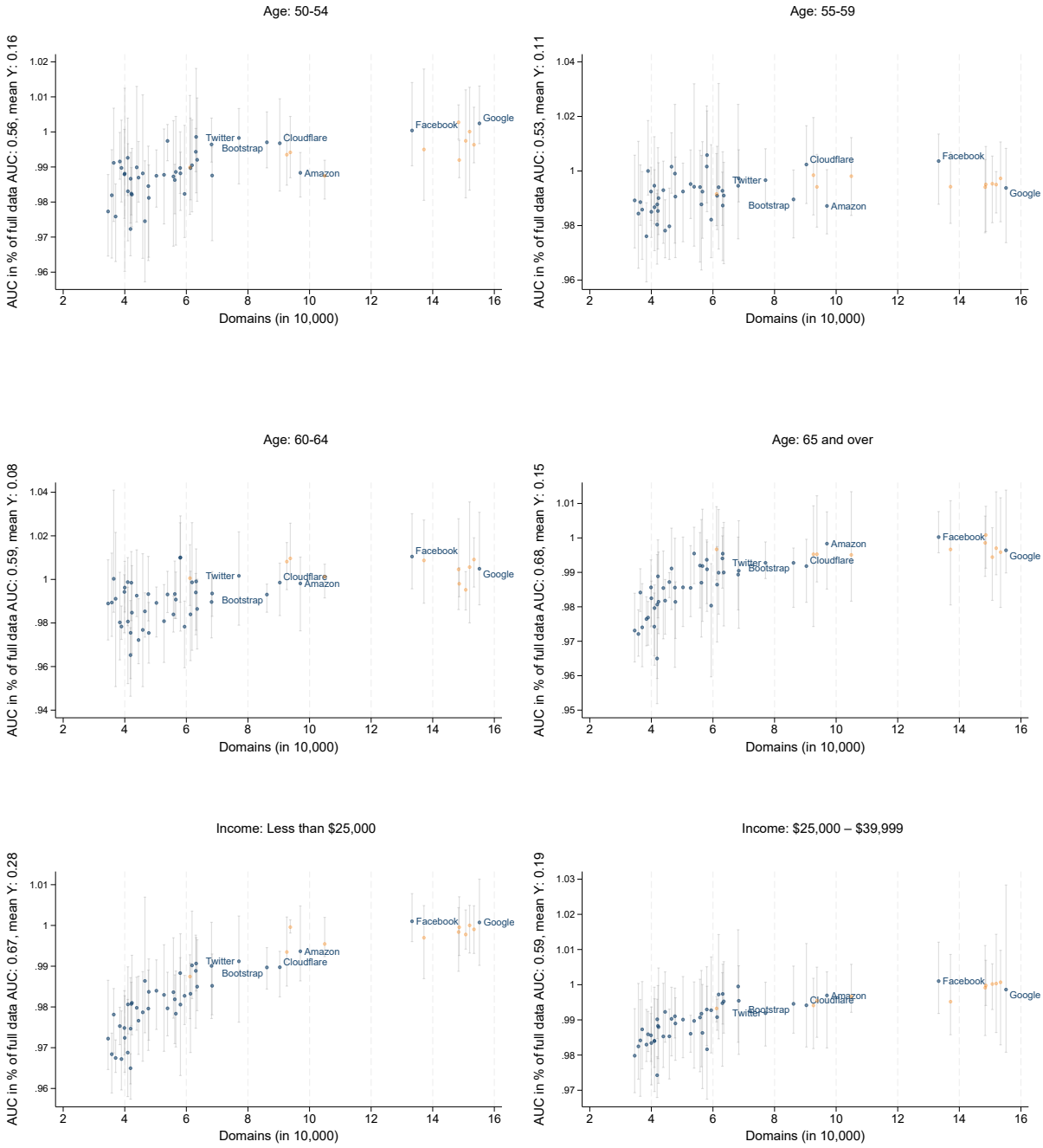
our machine learning algorithm, across trackers, gridpoints and folds. The last column displays the mean AUC obtained when using the entire clickstream data (mean total sample AUC). Note that the maximum AUC reported for a tracker (column 7) can be higher than the mean AUC from the entire sample (column 8), since the we capture the fold with the highest out-of-sample AUCs and compare it to the mean AUC (i.e. the AUC across the 5-folds) of the entire sample.

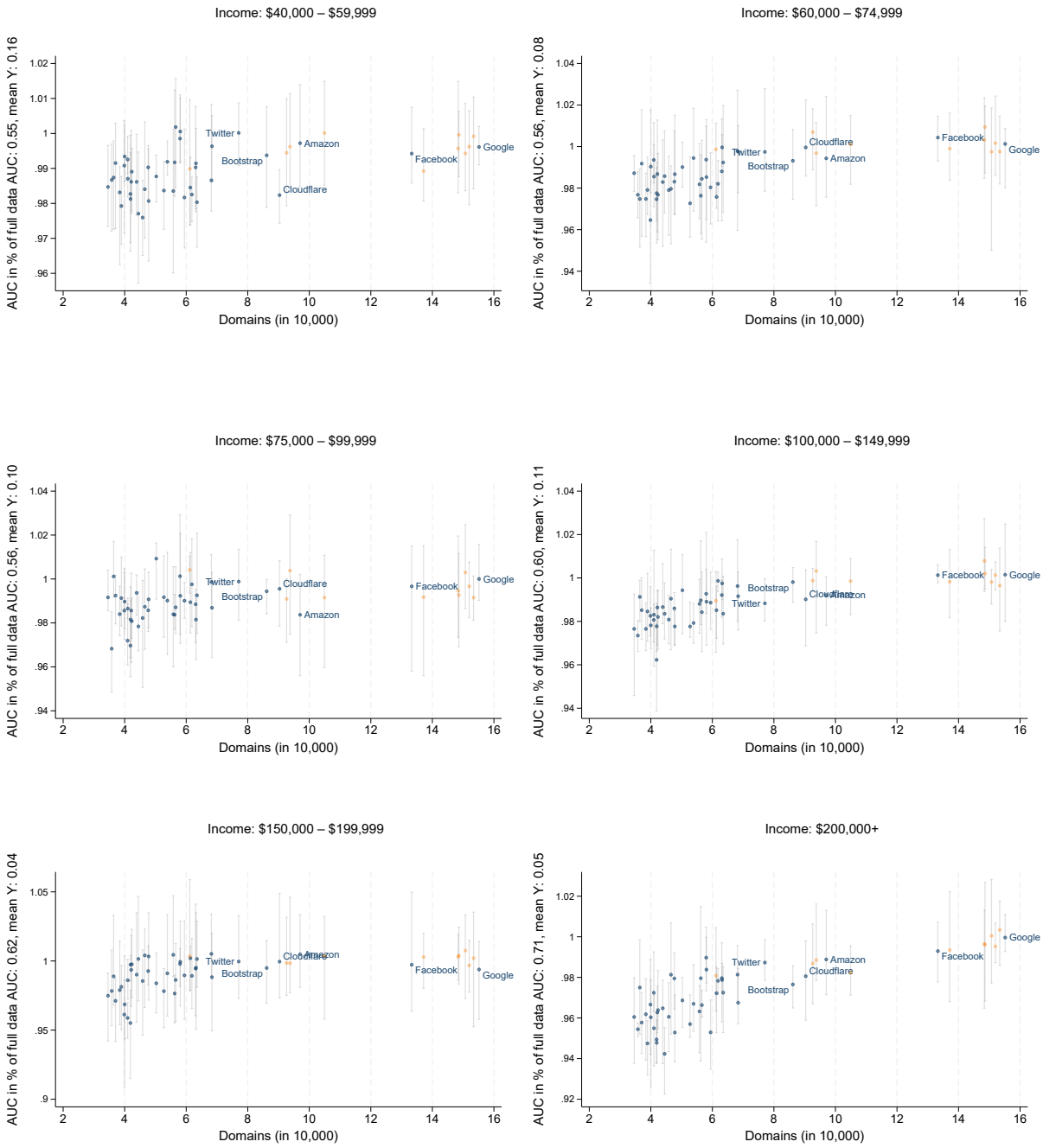
# D Across Trackers Plots

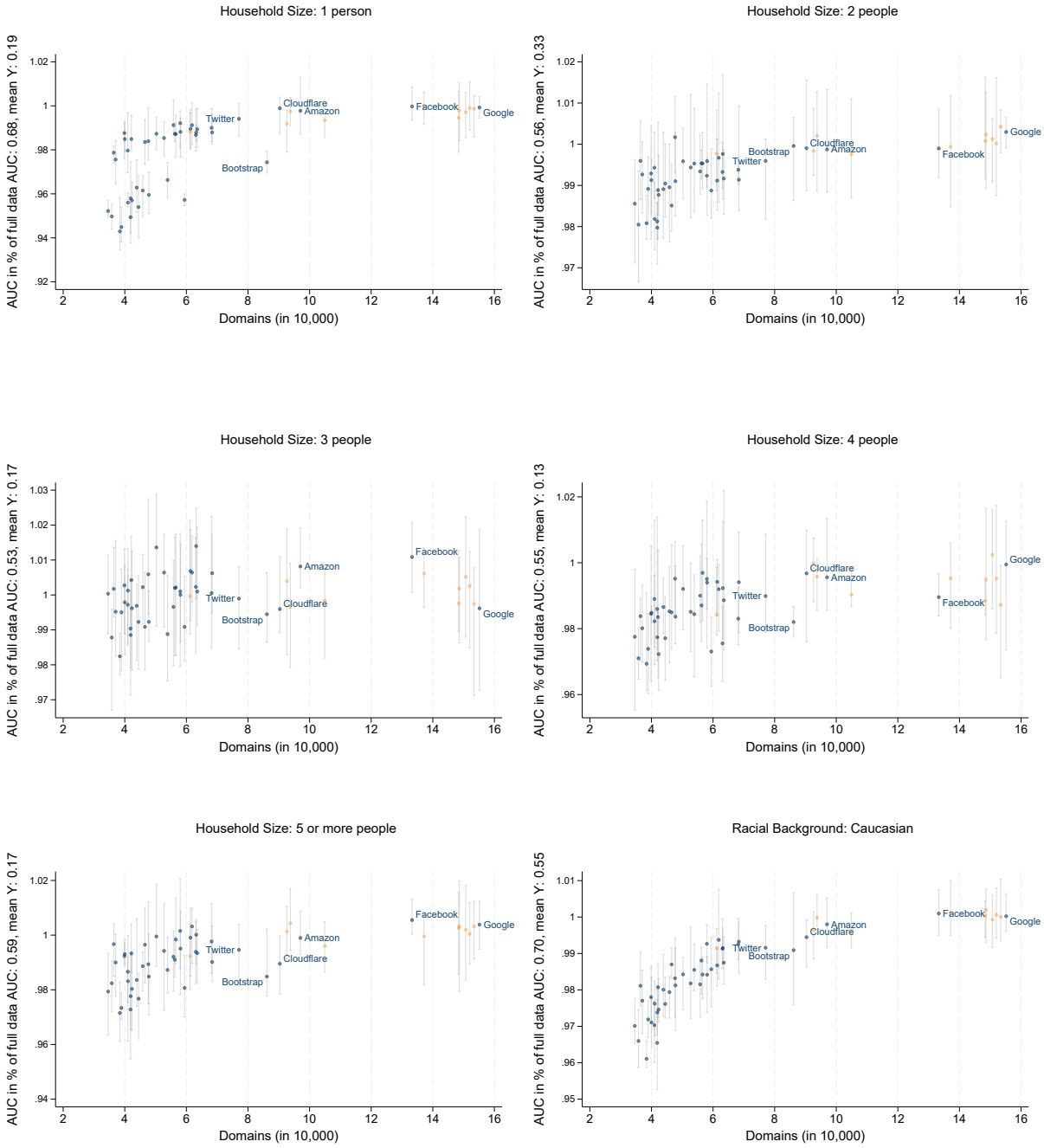


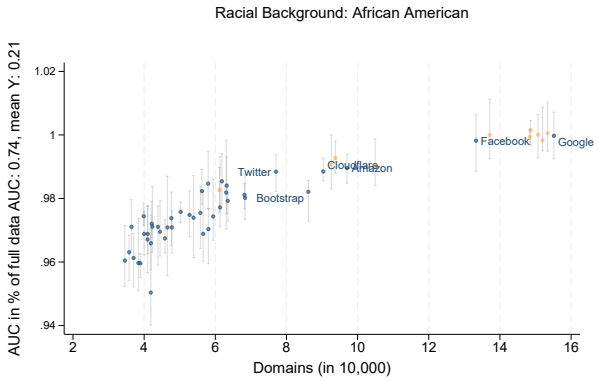

Education: High school diploma or GED
AUC in % of full data AUC: 0.61, mean Y: 0.03
Domains (in 10,000)
Bootstrap
Twitter
Amazon
Cloudflare
Facebook
Google
Education: Some college but no degree
AUC in % of full data AUC: 0.63, mean Y: 0.20
Domains (in 10,000)
Twitter
Bootstrap
Amazon
Cloudflare
Facebook
Google
Education: Associate degree
AUC in % of full data AUC: 0.62, mean Y: 0.21
Domains (in 10,000)
Twitter
Amazon
Cloudflare
Bootstrap
Facebook
Google
Education: Bachelor's degree
AUC in % of full data AUC: 0.68, mean Y: 0.14
Domains (in 10,000)
Cloudflare
Amazon
Twitter
Bootstrap
Facebook
Google
Education: Unknown
AUC in % of full data AUC: 0.76, mean Y: 0.41
Domains (in 10,000)
Twitter
Bootstrap
Facebook
Google
Age: 18-20
AUC in % of full data AUC: 0.71, mean Y: 0.03
Domains (in 10,000)
Twitter
Cloudflare
Amazon
Bootstrap
Facebook
Google

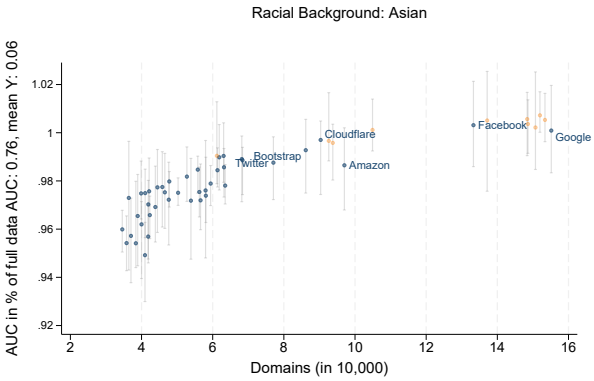

Age: 21-24
AUC in % of full data AUC: 0.65, mean Y: 0.04
Domains (in 10,000)
Cloudflare
Amazon
Bootstrap
Twitter
Facebook
Google
Age: 25-29
AUC in % of full data AUC: 0.59, mean Y: 0.06
Domains (in 10,000)
Cloudflare
Amazon
Twitter
Bootstrap
Facebook
Google
Age: 30-34
AUC in % of full data AUC: 0.57, mean Y: 0.07
Domains (in 10,000)
Cloudflare
Amazon
Bootstrap
Twitter
Facebook
Google
Age: 35-39
AUC in % of full data AUC: 0.56, mean Y: 0.08
Domains (in 10,000)
Cloudflare
Amazon
Twitter
Bootstrap
Facebook
Google
Age: 40-44
AUC in % of full data AUC: 0.55, mean Y: 0.10
Domains (in 10,000)
Amazon
Cloudflare
Twitter
Bootstrap
Facebook
Google
Age: 45-49
AUC in % of full data AUC: 0.55, mean Y: 0.14
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google

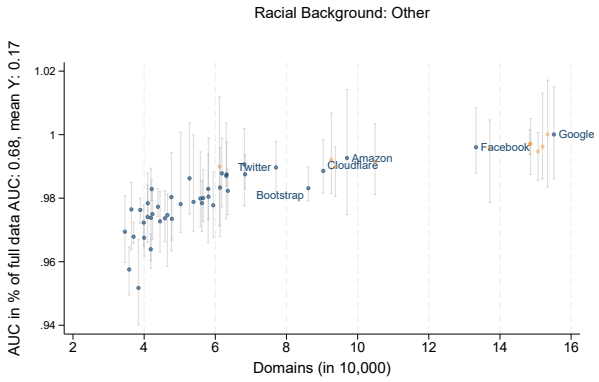

Age: 50-54
AUC in % of full data AUC: 0.56, mean Y: 0.16
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google
Age: 55-59
AUC in % of full data AUC: 0.53, mean Y: 0.11
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google
Age: 60-64
AUC in % of full data AUC: 0.59, mean Y: 0.08
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google
Age: 65 and over
AUC in % of full data AUC: 0.68, mean Y: 0.15
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google
Income: Less than $25,000
AUC in % of full data AUC: 0.67, mean Y: 0.28
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google
Income: $25,000 – $39,999
AUC in % of full data AUC: 0.59, mean Y: 0.19
Domains (in 10,000)
Twitter
Bootstrap
Cloudflare
Amazon
Facebook
Google


Income: $40,000 – $59,999
AUC in % of full data AUC: 0.55, mean Y: 0.16
Domains (in 10,000)
Twitter
Amazon
Bootstrap
Cloudflare
Facebook
Google
Income: $60,000 – $74,999
AUC in % of full data AUC: 0.56, mean Y: 0.08
Domains (in 10,000)
Twitter
Cloudflare
Bootstrap
Amazon
Facebook
Google
Income: $75,000 – $99,999
AUC in % of full data AUC: 0.56, mean Y: 0.10
Domains (in 10,000)
Twitter
Cloudflare
Bootstrap
Amazon
Facebook
Google
Income: $100,000 – $149,999
AUC in % of full data AUC: 0.60, mean Y: 0.11
Domains (in 10,000)
Bootstrap
Twitter
Facebook
Google
Income: $150,000 – $199,999
AUC in % of full data AUC: 0.62, mean Y: 0.04
Domains (in 10,000)
Twitter
Bootstrap
Facebook
Google
Income: $200,000+
AUC in % of full data AUC: 0.71, mean Y: 0.05
Domains (in 10,000)
Twitter
Amazon
Cloudflare
Bootstrap
Facebook
Google

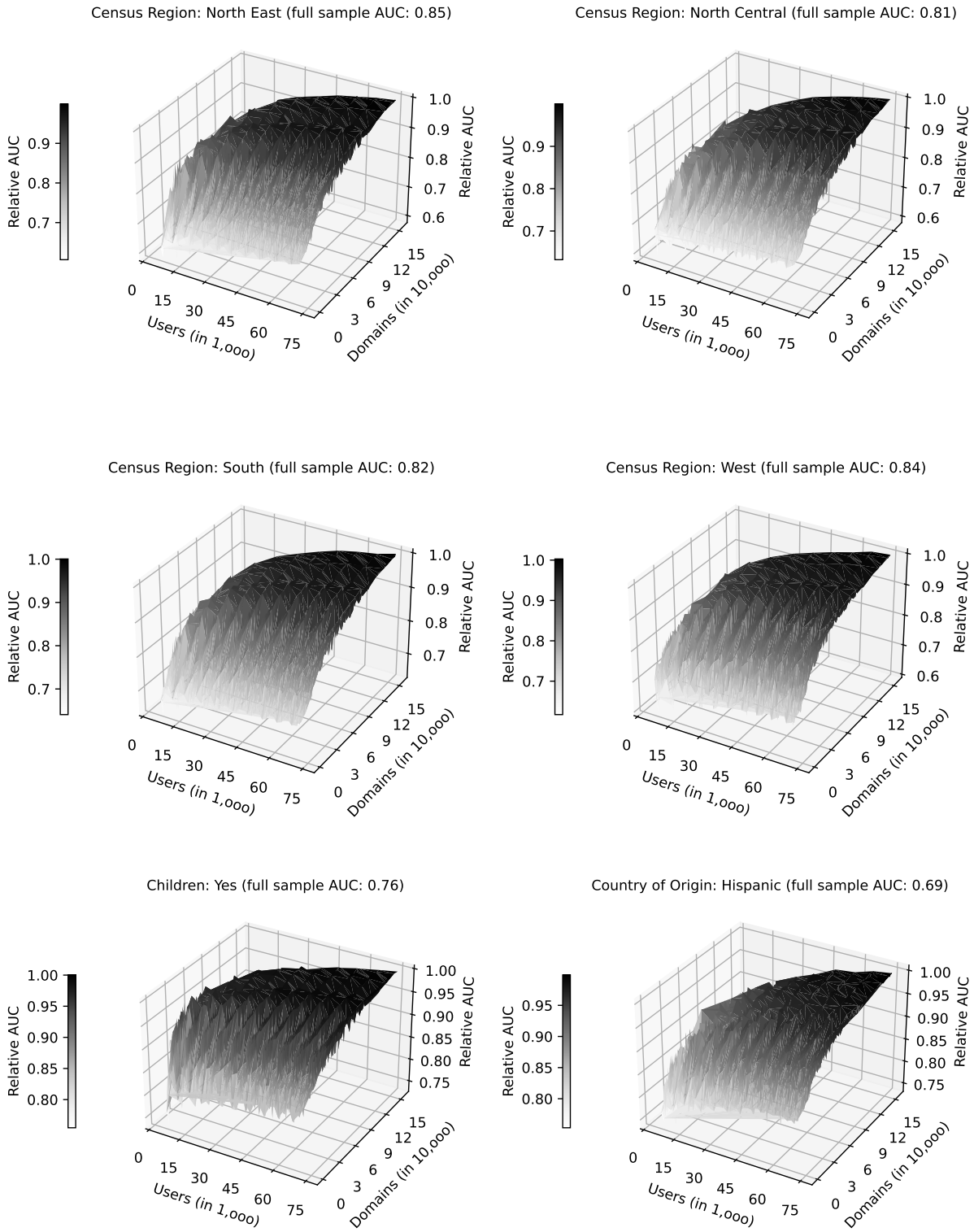

Household Size: 1 person
AUC in % of full data AUC: 0.68, mean Y: 0.19
Domains (in 10,000)
Twitter
Cloudflare
Amazon
Facebook
Google
Bootstrap
Household Size: 2 people
AUC in % of full data AUC: 0.56, mean Y: 0.33
Domains (in 10,000)
Bootstrap
Twitter
Cloudflare
Amazon
Facebook
Google
Household Size: 3 people
AUC in % of full data AUC: 0.53, mean Y: 0.17
Domains (in 10,000)
Facebook
Amazon
Twitter
Cloudflare
Bootstrap
Google
Household Size: 4 people
AUC in % of full data AUC: 0.55, mean Y: 0.13
Domains (in 10,000)
Cloudflare
Amazon
Twitter
Google
Facebook
Bootstrap
Household Size: 5 or more people
AUC in % of full data AUC: 0.59, mean Y: 0.17
Domains (in 10,000)
Facebook
Google
Amazon
Twitter
Cloudflare
Bootstrap
Racial Background: Caucasian
AUC in % of full data AUC: 0.70, mean Y: 0.55
Domains (in 10,000)
Facebook
Google
Amazon
Cloudflare
Twitter
Bootstrap


Racial Background: African American
AUC in % of full data AUC: 0.74, mean Y: 0.21
1.02
1
.98
.96
.94
2
4
6
8
10
12
14
16
Domains (in 10,000)
Facebook
Google
Twitter
Cloudflare
Amazon
Bootstrap

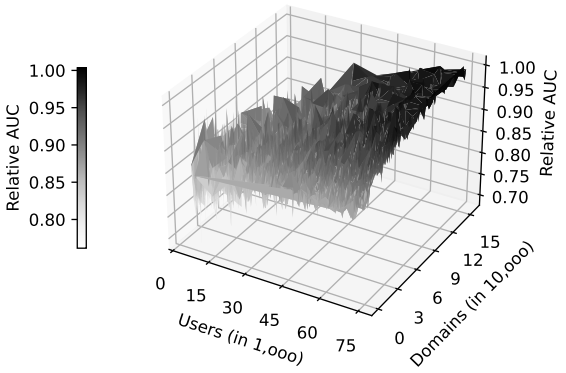

Racial Background: Asian
AUC in % of full data AUC: 0.76, mean Y: 0.06
1.02
1
.98
.96
.94
.92
2
4
6
8
10
12
14
16
Domains (in 10,000)
Facebook
Google
Cloudflare
Bootstrap
Twitter
Amazon


Racial Background: Other
AUC in % of full data AUC: 0.68, mean Y: 0.17
1.02
1
.98
.96
.94
2
4
6
8
10
12
14
16
Domains (in 10,000)
Facebook
Google
Amazon
Twitter
Cloudflare
Bootstrap

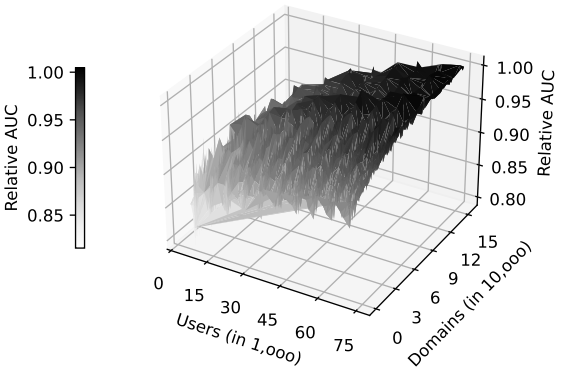
# E 3D Plots


Census Region: North East (full sample AUC: 0.85)
Census Region: North Central (full sample AUC: 0.81)
Census Region: South (full sample AUC: 0.82)
Census Region: West (full sample AUC: 0.84)
Children: Yes (full sample AUC: 0.76)
Country of Origin: Hispanic (full sample AUC: 0.69)
Relative AUC
Users (in 1,000)
Domains (in 10,000)

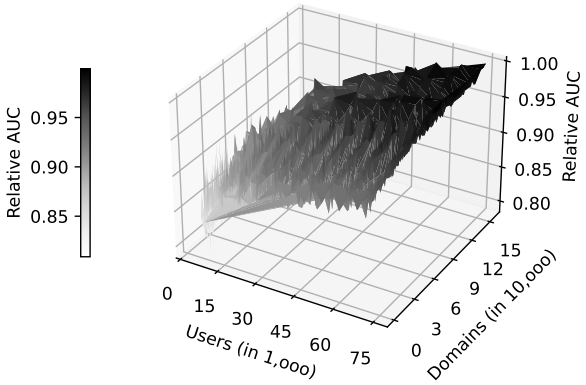
Education: High school diploma or GED (full sample AUC: 0.61)



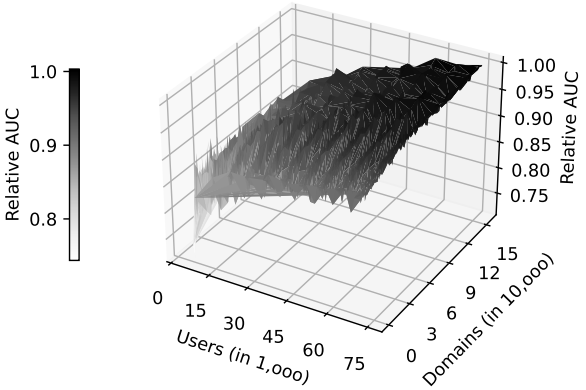

Education: Some college but no degree (full sample AUC: 0.63)



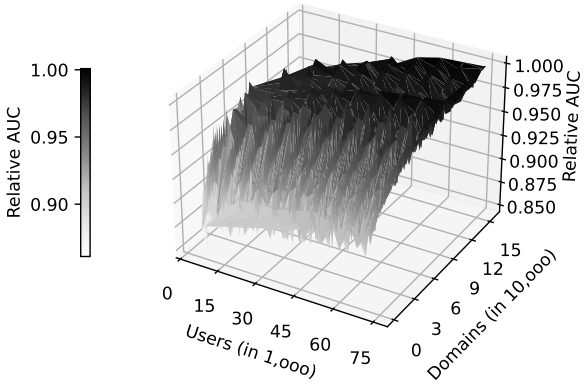

Education: Associate degree (full sample AUC: 0.62)



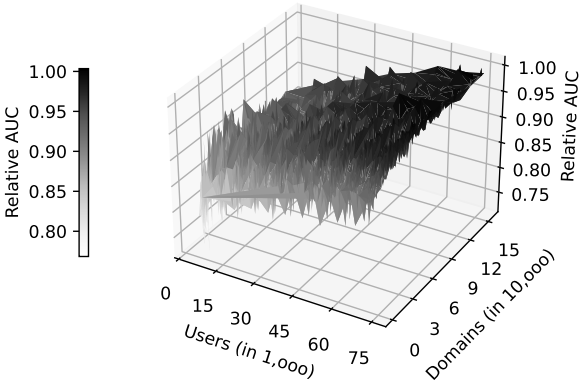

Education: Bachelor's degree (full sample AUC: 0.68)

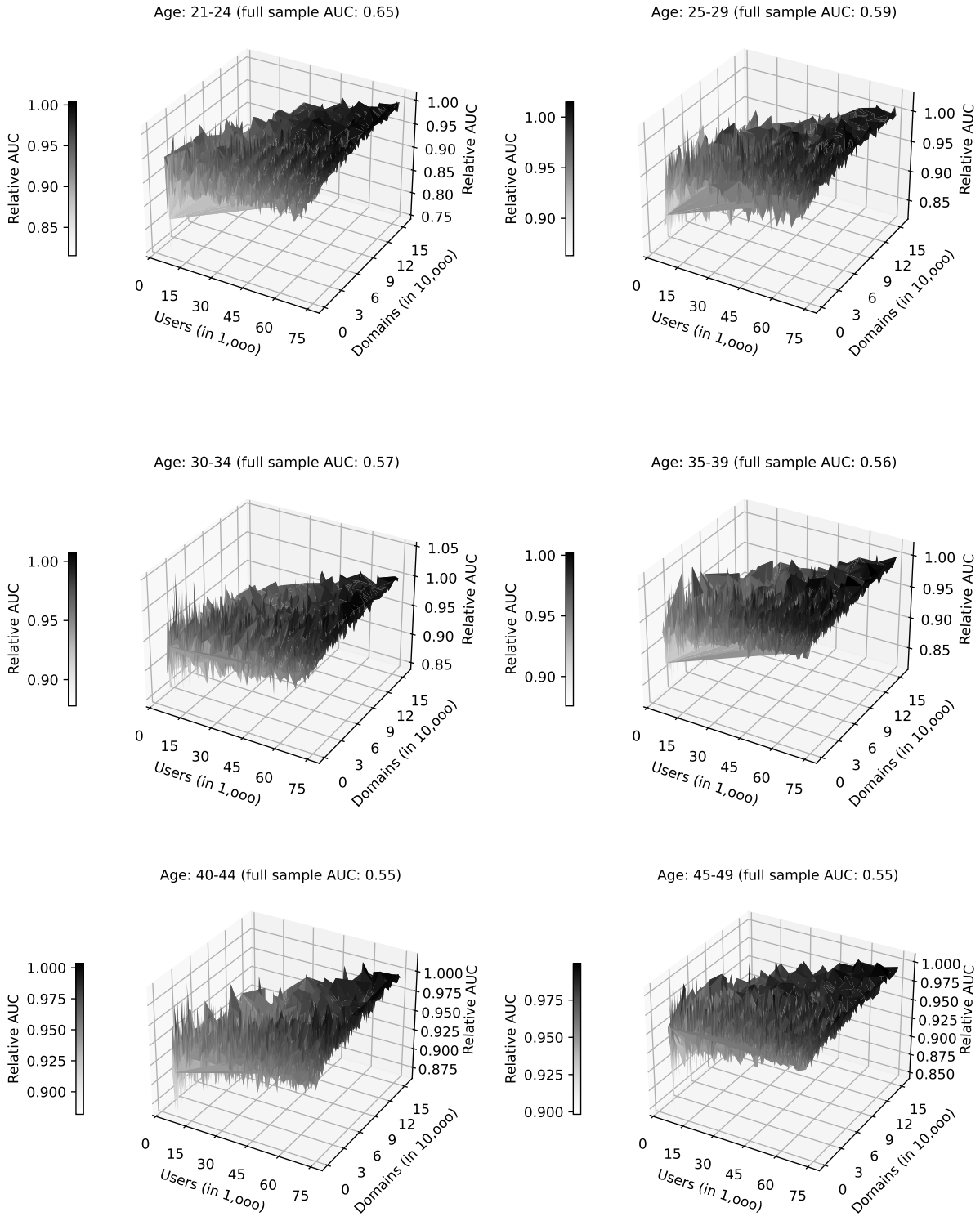


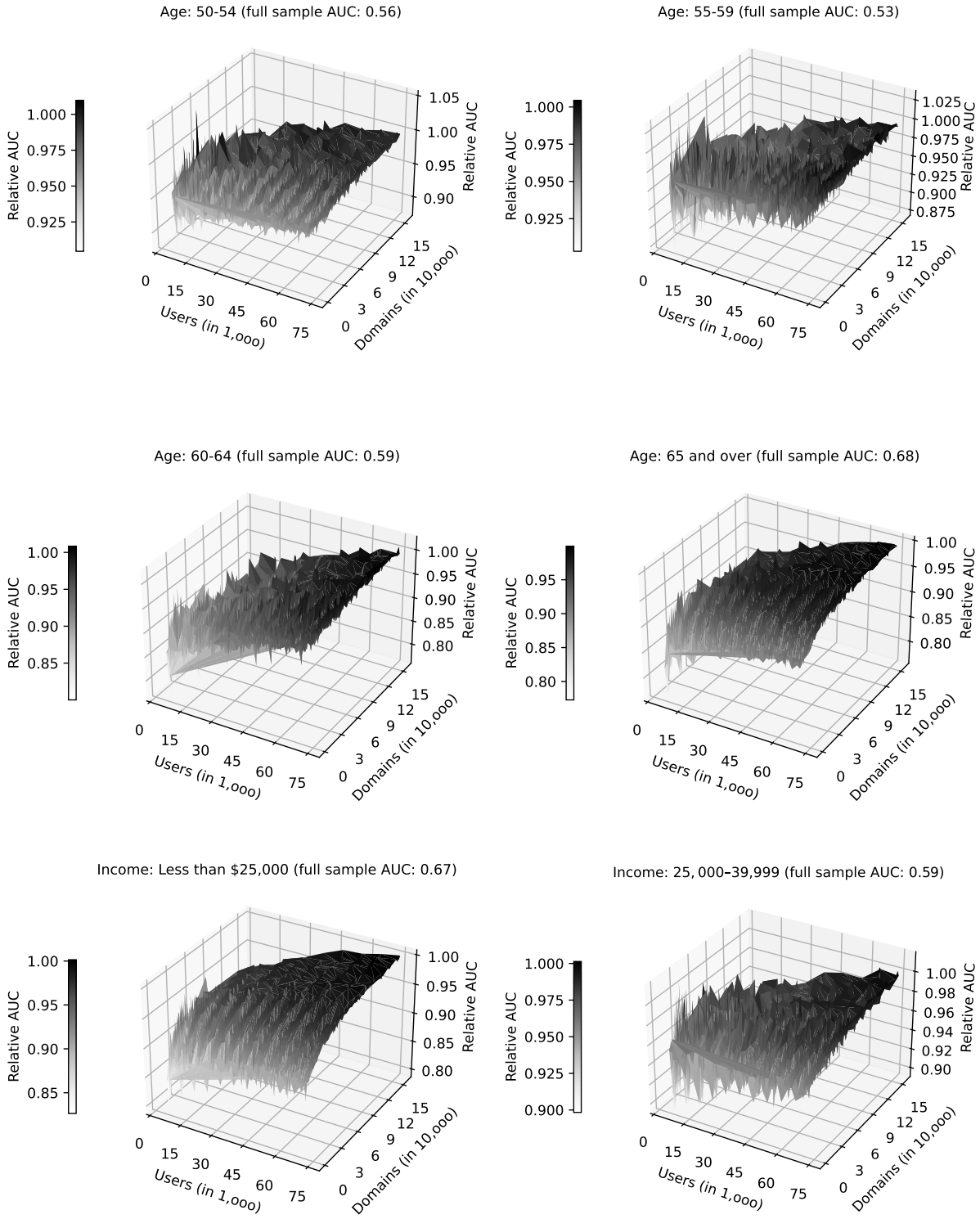

Education: Unknown (full sample AUC: 0.76)

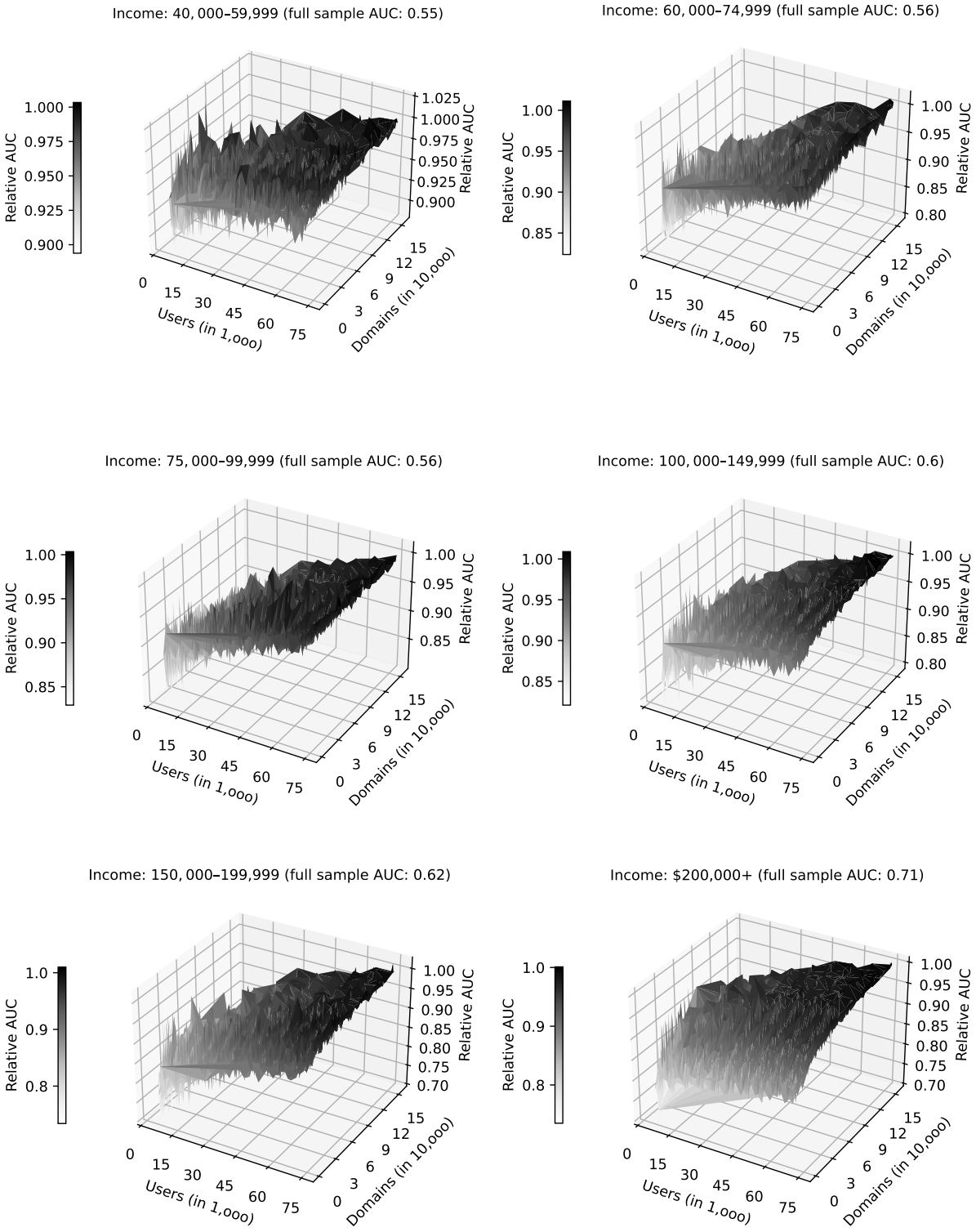


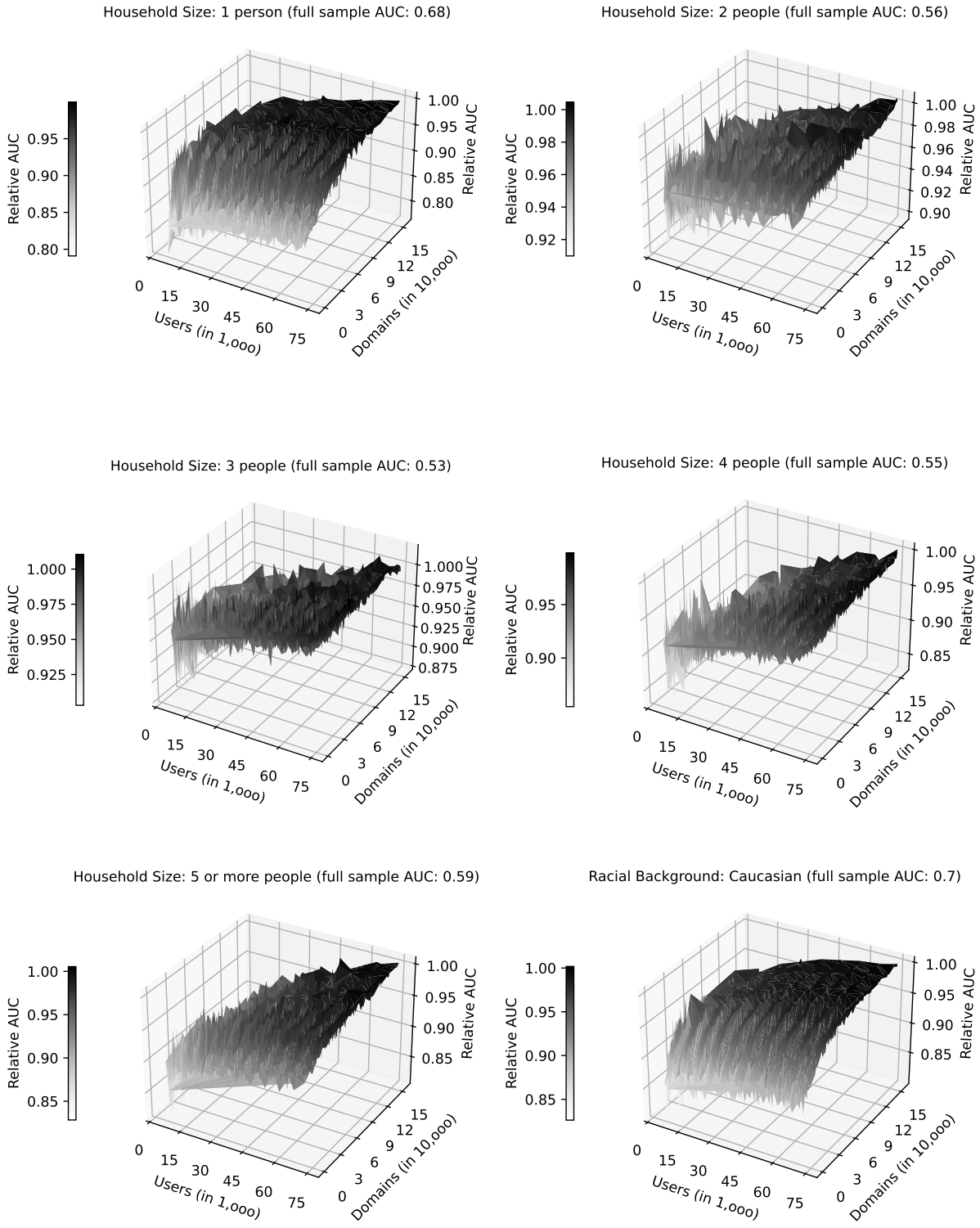

Age: 18-20 (full sample AUC: 0.71)

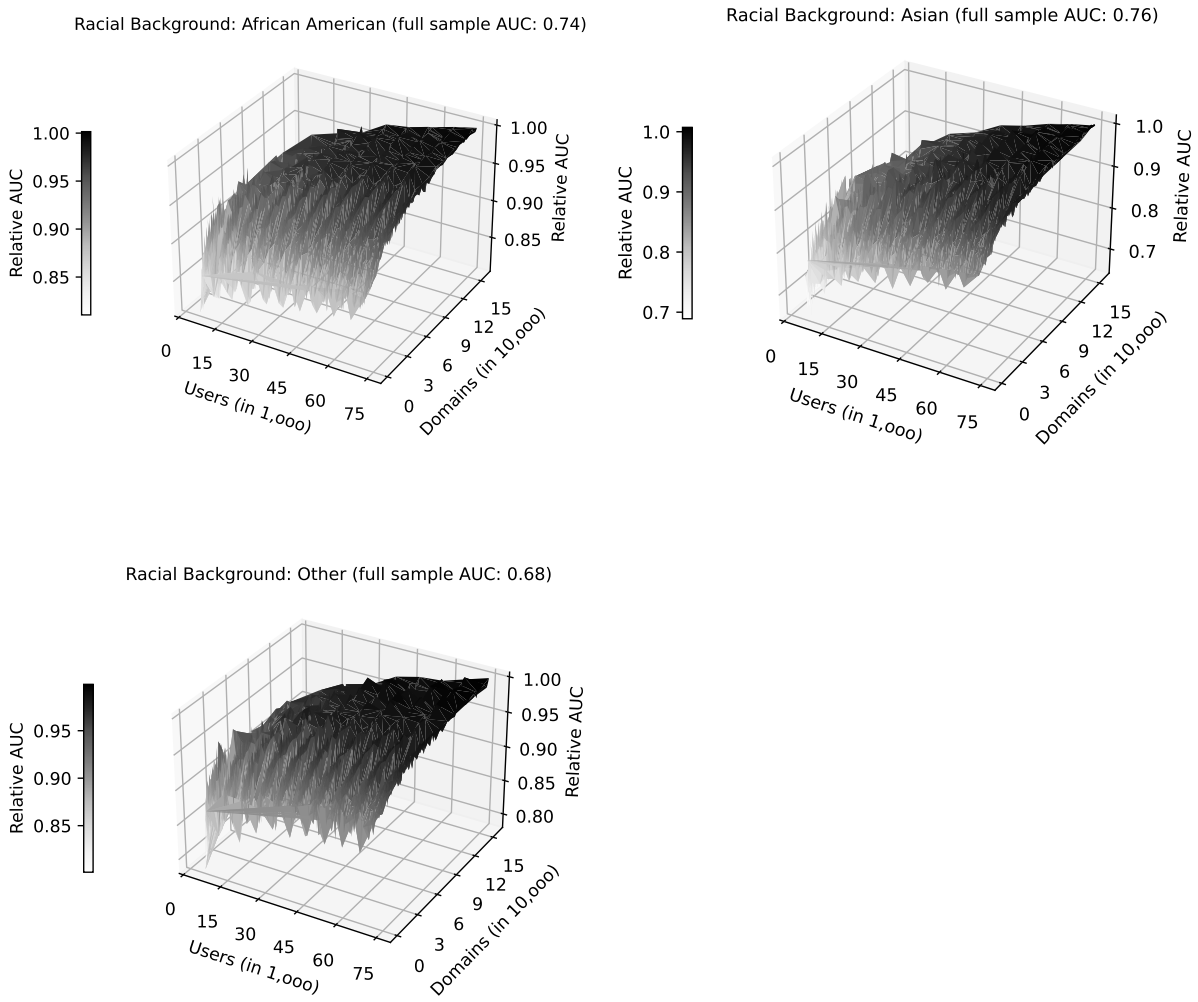



Age: 21-24 (full sample AUC: 0.65)
Age: 25-29 (full sample AUC: 0.59)
Age: 30-34 (full sample AUC: 0.57)
Age: 35-39 (full sample AUC: 0.56)
Age: 40-44 (full sample AUC: 0.55)
Age: 45-49 (full sample AUC: 0.55)
Relative AUC
Users (in 1,000)
Domains (in 10,000)


Age: 50-54 (full sample AUC: 0.56)
Age: 55-59 (full sample AUC: 0.53)
Age: 60-64 (full sample AUC: 0.59)
Age: 65 and over (full sample AUC: 0.68)
Income: Less than $25,000 (full sample AUC: 0.67)
Income: 25, 000–39,999 (full sample AUC: 0.59)
Relative AUC
Users (in 1,000)
Domains (in 10,000)


Income: 40, 000–59,999 (full sample AUC: 0.55)
Income: 60, 000–74,999 (full sample AUC: 0.56)
Income: 75, 000–99,999 (full sample AUC: 0.56)
Income: 100, 000–149,999 (full sample AUC: 0.6)
Income: 150, 000–199,999 (full sample AUC: 0.62)
Income: $200,000+ (full sample AUC: 0.71)
Relative AUC
Users (in 1,000)
Domains (in 10,000)


Household Size: 1 person (full sample AUC: 0.68)
Household Size: 2 people (full sample AUC: 0.56)
Household Size: 3 people (full sample AUC: 0.53)
Household Size: 4 people (full sample AUC: 0.55)
Household Size: 5 or more people (full sample AUC: 0.59)
Racial Background: Caucasian (full sample AUC: 0.7)
Relative AUC
Users (in 1,000)
Domains (in 10,000)




# F Regression Analysis with Simulated Data

To systematically estimate the returns to data along both dimensions, we run a polynomial regression based on our prediction results. Our baseline specification for tracker $i$, prediction task $s$, grid point $g$, and fold $f$ is given by:

$$AUC_{isgf} = \alpha + \beta_1 n_{isgf} + \beta_2 n_{isgf}^2 + \beta_3 k_{isgf} + \beta_4 k_{isgf}^2 + \gamma X_{isgf} + \varepsilon_{isgf}$$

where $n$ denotes the number of observed users and $k$ denotes the number of observed domains in firm $i$'s prediction model. The outcome variable $AUC$ corresponds to the area under the ROC curve, our measure of prediction quality. $X$ captures control variables, such as the mean and squared mean Y,[18] as well as dummy variables for each fold. We estimate this model on the entire sample, i.e. across trackers, and for largest individual tracker firms, exploiting the

[18] Mean Y corresponds to the prevalence of the relevant subgroup in a prediction task.

artifical within-tracker variance in the number of users and domains generated through our grid described in section C.

Table 7: Regression analysis with AUC as outcome variable

| VARIABLES | (1) All Trackers | (2) Google | (3) Facebook | (4) Amazon | (5) Bootstrap | (6) Cloudflare |
|---|---|---|---|---|---|---|
| Users (in 1,000s) | 0.00127*** | 0.00109*** | 0.00122*** | 0.00123*** | 0.00141*** | 0.00130*** |
| | (0.00006) | (0.00006) | (0.00006) | (0.00006) | (0.00006) | (0.00006) |
| Users Squared | -0.00001*** | -0.00001*** | -0.00001*** | -0.00001*** | -0.00001*** | -0.00001*** |
| | (0.00000) | (0.00000) | (0.00000) | (0.00000) | (0.00000) | (0.00000) |
| Domains (in 10,000s) | 0.01296*** | 0.00756*** | 0.00921*** | 0.01232*** | 0.01194*** | 0.01322*** |
| | (0.00026) | (0.00022) | (0.00028) | (0.00035) | (0.00038) | (0.00039) |
| Domains Squared | -0.00064*** | -0.00027*** | -0.00039*** | -0.00071*** | -0.00072*** | -0.00082*** |
| | (0.00002) | (0.00001) | (0.00002) | (0.00004) | (0.00004) | (0.00004) |
| Observations | 1,014,000 | 19,500 | 19,500 | 19,500 | 19,500 | 19,500 |
| R-squared | 0.90815 | 0.93361 | 0.93648 | 0.92837 | 0.92780 | 0.93203 |
| Task FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Tracker FE | Yes | | | | | |
| Fold FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Cluster | Gridpoint-Fold | Gridpoint-Fold | Gridpoint-Fold | Gridpoint-Fold | Gridpoint-Fold | Gridpoint-Fold |

Standard errors in parentheses are clustered at the grid point-fold level
*** p<0.01, ** p<0.05, * p<0.1

Table 7 reports the regression results from our main specification for all trackers pooled together (column 1), ii) Google (column 2), as well as the next 4 largest tracker firms in terms of observed domains (columns 3-6). The marginal returns to data, in the user and domain dimensions are respectively given by the partial derivatives

$$\frac{\partial AUC}{\partial n} = \beta_1 + 2\beta_2 n \text{ , and } \frac{\partial AUC}{\partial k} = \beta_3 + 2\beta_4 k$$

Throughout across and within trackers, we observe decreasing returns to data in both dimensions, as indicated by the positive "Users" and "Domains" coefficients ($\beta_1$ and $\beta_2$) but negative squared terms ($\beta_3$ and $\beta_4$).