

Examining User Heterogeneity in Digital Experiments

SRIRAM SOMANCHI, AHMED ABBASI, and KEN KELLEY, University of Notre Dame, USA
DAVID DOBOLYI, University of Colorado, USA
TED TAO YUAN, eBay, USA

Digital experiments are routinely used to test the value of a treatment relative to a status-quo control setting—for instance, a new search relevance algorithm for a website or a new results layout for a mobile app. As digital experiments have become increasingly pervasive in organizations and a wide variety of research areas, their growth has prompted a new set of challenges for experimentation platforms. One challenge is that experiments often focus on the average treatment effect (ATE) without explicitly considering differences across major sub-groups: heterogeneous treatment effect (HTE). This is especially problematic, because ATEs have decreased in many organizations as the more obvious benefits have already been realized. However, questions abound regarding the pervasiveness of user HTEs and how best to detect them. We propose a framework for detecting and analyzing user HTEs in digital experiments. Our framework combines an array of user characteristics with double machine learning. Analysis of 27 real-world experiments spanning 1.76 billion sessions and simulated data demonstrates the effectiveness of our detection method relative to existing techniques. We also find that transaction, demographic, engagement, satisfaction, and lifecycle characteristics exhibit statistically significant HTEs in 10% to 20% of our real-world experiments, underscoring the importance of considering user heterogeneity when analyzing experiment results; otherwise, personalized features and experiences cannot happen, thus reducing effectiveness. In terms of the number of experiments and user sessions, we are not aware of any study that has examined user HTEs at this scale. Our findings have important implications for information retrieval, user modeling, platforms, and digital experience contexts, in which online experiments are often used to evaluate the effectiveness of design artifacts.

CCS Concepts: • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Heterogeneous treatment effects, digital experiments, user heterogeneity, user modeling, double machine learning

ACM Reference format:
Sriram Somanchi, Ahmed Abbasi, Ken Kelley, David Dobolyi, and Ted Tao Yuan. 2023. Examining User Heterogeneity in Digital Experiments. *ACM Trans. Inf. Syst.* 41, 4, Article 100 (March 2023), 34 pages.
https://doi.org/10.1145/3578931

Sriram Somanchi and Ahmed Abbasi contributed equally to this research.

This work was funded in part through U.S. NSF grant IIS-2039915 and an eBay research grant entitled "Machine Learning Methods for Causal Inference in Digital Experimentation Platforms."

Authors' addresses: S. Somanchi, 344 Mendoza College of Business, University of Notre Dame, Notre Dame, IN, USA 46556; email: ssomanch@nd.edu; A. Abbasi, 360 Mendoza College of Business, University of Notre Dame, Notre Dame, IN, USA 46556; email: aabbasi@nd.edu; K. Kelley, 234B Mendoza College of Business, University of Notre Dame, Notre Dame, IN, USA 46556; email: kkelley@nd.edu; D. Dobolyi, 401F KOBL, Leeds School of Business, University of Colorado Boulder, Boulder, CO, USA 80309; email: david.doboloyi@colorado.edu; T. T. Yuan, 2025 Hamilton Ave, San Jose, CA, USA 95125; email: teyuan@ebay.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2023 Copyright held by the owner/author(s).

1046-8188/2023/03-ART100 $15.00
https://doi.org/10.1145/3578931

1 INTRODUCTION

Digital experiments, also commonly referred to as online A/B tests or online controlled experiments, are routinely used by organizations and researchers online to test the value and efficacy of a digital treatment relative to a status-quo control setting [39, 45]. Example digital treatments include a new search relevance algorithm for a website, a new results layout for a mobile app, an adjustment to the font colors used to display sponsored ad copy headlines, an update to the recommendation system/engine, and so on [43, 47]. Digital experiments have become a key mechanism through which data-driven decisions are enacted regarding the design, development, and management of digital artifacts involving user preferences and interaction behaviors.

In recent years, there has been dramatic growth in the number of digital experiments conducted [47, 68]. At the forefront are technology companies such as Microsoft, Google, Amazon, Meta, and Tencent, which each now run well over 10,000 experiments annually [47, 81]. Whereas such experiments generate immense monetary benefit through enhanced user experience and operational efficiency gains, their growth and pervasiveness has also ushered in a new era of challenges [33]. These challenges include *sensitivity* of the experiments and *heterogeneity* of findings. *Sensitivity*, or statistical power, refers to the ability of an experiment to detect differences in key metrics between control and treatment groups when they truly exist [79]. As the number of experiments deployed over time within organizations increases, many of the "low-hanging fruit" outcomes are realized early on, resulting in smaller effect sizes for new experimental treatments [33]. Here, by effect size, we mean "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Reference [41], p. 137), such as lift, the difference in the proportion of individuals making a purchase, or the number of clicks within in a session. Regarding the diminishing size of effects, on average, the 500th improvement to the search relevance machine learning algorithm on an e-commerce platform is not likely to yield the same incremental gain as the first 100 algorithm updates [23, 33]. *Heterogeneity* in experiments, however, refers to the possibility of user sub-groups that might observe a *heterogeneous treatment effect (HTE)*—that is, treatment versus control effects that deviate from the *average treatment effect (ATE)* observed across the entire experiment as a whole [67, 80].

User HTE detection methods have been suggested as a mechanism for alleviating both challenges—directly relevant to the heterogeneity in experiments and implicitly related to sensitivity [67]. For instance, in an experiment evaluating advertising, providing analysts and decision-makers with details about how click-through rates associated with a proposed ad intervention differ for key subgroups might help provide insights that add value beyond insignificant and/or low-effect-size average treatment effects—that is, helping with the sensitivity issue [67]. Relatedly, even when the overall effect sizes are statistically significant, statistically and practically speaking, knowing how the results vary for major sub-populations is important to understand the value proposition in a more holistic manner [33]. Indeed, considering the effect sizes within groups instead of across the entire population provides an opportunity for nuanced and potentially meaningful findings. It is entirely possible, for example, that multiple subgroups have different effects, but they cancel when combined. Thus, what looks like a near-zero ATE may be heterogeneity in effect sizes across the various groups. Despite the importance of using HTE methods to answer questions such as "who drives the changes," there has been limited work examining user heterogeneity in digital experiments [80]. Accordingly, the research objective of this study is to *develop and apply an empirical framework for examining user heterogeneity across a set of digital experiments*. Our three *research questions (RQ)* are as follows:

- **RQ1:** How accurately can we detect user HTEs in digital experiments?
- **RQ2:** In the context of digital experiments, how prevalent are user characteristics in influencing conversion outcomes?

- **RQ3:** (a) What are the user HTE effect sizes and statistical significance in digital experiments? (b) How do these effect sizes vary for different experiment characteristics?

To address these questions, we leverage the user modeling and customer analytics literature to identify a set of user characteristic constructs that have been consistently observed to predict online user behavior outcomes in a variety of contexts, including e-commerce and general digital user experience [11, 24, 42, 52, 57, 77]. These include users' prior session/transaction history, demographics, engagement, satisfaction, channel, choice, messaging, and lifecycle variables. We then adapt these constructs into a framework for examining user HTE in digital experiments. Our framework considers the different session phases encountered in digital settings, the importance of experimental factors such as channels/devices, and key user characteristics that might drive heterogeneity, in conjunction with a double machine learning method called *heterogeneous treatment effect–double machine learning (HTE-DML)*, which is designed to provide an unbiased estimate of user HTE [20, 66]. To address our research questions, we evaluate the framework and HTE-DML on a series of simulated digital experimentation datasets and also use 27 real-world digital experiments from a major e-commerce platform. These real-world experiments, which encompass 1.76 billion user sessions, span various phases of the session journey, including search, merchandising, advertising, viewing items, and checkout. Our simulation is thus based on real-world results and reveals several important findings that have considerable impact for those in the digital experimentation space that have not previously been realized. The benchmark evaluation versus existing HTE detection methods shows that HTE-DML provides better sensitivity (true positive) and reduced false-positive rates for detected user HTEs (RQ1). With respect to the overall impact of user characteristics on conversion outcomes (RQ2), we find that, on average, 20–60% of the user characteristic variables significantly impact session-level conversion outcomes, confirming the importance of such variables in understanding and predicting user behavior. In regards to RQ3, we observe that many user characteristics also have significant (at $\alpha = 0.05$) HTEs in 10% to 20% of experiments. These results hold for various session stages and channels, underscoring the importance of incorporating user HTE assessment as part of the digital experimentation measurement framework.

The main contributions of our work are threefold. First, we propose a framework for holistically assessing user HTE in digital experiments by considering the interplay between session phases, experiment factors, and an array of user characteristics guided by the user modeling and customer analytics literature, in conjunction with powerful machine learning detection methods. Second, we apply our framework to a large set of digital experiments, in partnership with a global online marketplace, and provide empirical insights regarding the extent and magnitude of user HTEs in such online controlled experiments. To the best of our knowledge, no study has examined user HTEs at this scale, in terms of the number of experiments and total user sessions. Third, we use an extensive evaluation, including simulations to evaluate the method and real-world experiments applying the method to estimate the impact, so as to illustrate how and why double machine learning methods are more effective than existing univariate and multivariate techniques commonly used for detecting HTEs. Our work has important implications for research examining digital experiments in large-scale environments, in which we provide a roadmap for how to accurately measure user HTEs and shed light on their potential prevalence. This is especially relevant to researchers working in the information retrieval, user modeling, and digital experience spaces, in which proposed artifacts such as search algorithms, recommender systems, and digital layout designs are routinely evaluated and validated using end user A/B testing [39, 51, 57]. Hence, our three main contributions are the proposed framework/method, empirical insights, and extensive evaluation. Furthermore, our work has important practical implications for applied

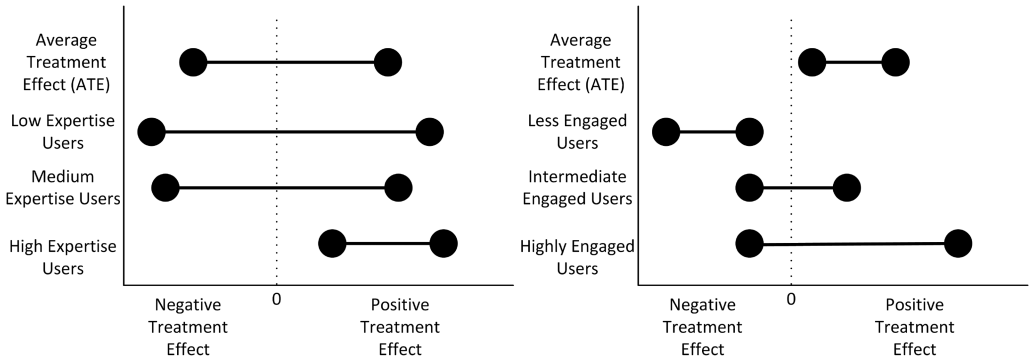


Fig. 1. A motivating example: User HTE in digital experiments.

data science teams, analysts, and managers tasked with designing, running, and analyzing digital experiments in online platforms. Indeed, our work has already benefited our partner.

The remainder of this article is organized as follows. In the ensuing section, we discuss relevant prior work on digital experiments, heterogeneity, user modeling, and customer analytics. In Section 3, we introduce our framework for measuring user HTEs in digital experiments. Section 4 describes our two testbeds, whereas Section 5 presents evaluation results comparing our double machine learning model against existing methods on an array of experiment simulations. In Section 6, we apply the framework to a set of large experiments conducted at a major e-commerce platform and report results related to our research questions. In Section 7, we offer concluding remarks.

2 RELATED WORK

In this section, we review relevant prior work related to online controlled experiments, user modeling and customer analytics, and user HTE detection. We then outline several key areas in need of development, which we contribute to with our methods. But first, we offer a motivating example.

Figure 1 shows two illustrative experiments. The one on the left-hand side is an experiment to test whether clearing the older auto-complete suggestions in a search engine box improves or hurts a downstream outcome, such as the session conversion rate. The first row shows the confidence interval (at $\alpha = 0.05$, i.e., 95% confidence) for the population ATE, that is, the difference in conversion rates for those assigned to the "clear suggestions" setting relative to those in the control group. Although the estimated ATE is slightly positive (i.e., the center of the line is greater than 0), it is not statistically significant, evidenced by the lower confidence interval (left circle) crossing 0. Now, suppose that we are able to measure those that are low, medium, and high expertise users. Here, using the groupings of expertise we can decompose the ATE conditional on each of the groups. In so doing, we are using an HTE approach. The HTE approach, then, suggests that the ATE is not in fact consistent across the three levels of expertise (the three lines below ATE). Because the effectiveness of clearing auto-complete depends on the level of expertise, it is a classic interaction (also known as moderation) effect. From Figure 1, notice that for high expertise users, the HTE is higher than the ATE, and it is statistically significant. In other words, the users with the highest expertise seem to benefit the most from being in the treatment group, here, the clearing of older auto-complete suggestions.

We can now contrast the left side of Figure 1 to the right side. This similarly stylized example shows the impact of display ads appearing on a web page and evaluates how such ads might cannibalize conversion rates. Again, the topmost plot shows the confidence interval for the population ATE, followed by the HTE approach, which examines the effectiveness by conditioning on

engagement. As the figure on the right shows, for less-, intermediate, and highly engaged users, the effectiveness of display ads differ. In this example, the overall conversion ATE is positive and statistically significant, but for the less-engaged users, the HTE has the opposite effect, namely that there is a statistically significantly negative effect.

Although these two examples are hypothetical, they are quite realistic scenarios. In fact, these examples signify a common set of questions asked by experimenters in digital settings [45]. Further, they illustrate the potential value of consideration of user HTE for decision-making, both in terms of uncovering differences in effect size across levels of some other variable(s) or feature(s) and the impact of those on for different levels or sub-groups, as well the amount of variance (i.e., the width of the confidence intervals) [67, 80]. These examples illustrate the three research questions we address in this article. *RQ1* is geared toward identifying the user characteristics that can be included in the consideration set (i.e., the ordinate or y-axis) based on impact on downstream implications such as conversion. *RQ2* is focused on how accurately the confidence interval plots for user HTE (the effect size and significance) can be inferred from large-scale digital experiments. Finally, regarding *RQ3*, whereas Figure 1 is an illustrative example, applying an accurate user HTE detection method on a large set of experiments can help shed light on how many such HTEs are statistically significant in real-world settings. In particular, we later show that DML, state-of-the-art techniques for estimating HTE in digital experiments using two-stages of machine learning, provides advantages over traditional HTE detection methods. Related to this example and our research questions, in the remainder of the section, we discuss relevant prior work on digital experiments, user characteristics, and HTE detection.

2.1 Online Controlled Experiments

Digital experiments, often referred to as online controlled experiments or A/B tests, have been an important measurement tool and evaluation method since the early days of the Internet. In academic research, in situations involving digital platform-oriented artifacts such as search relevance, recommendation engines, result display layouts, conversational agents, and so forth, such controlled user experiments are critical to infer the operational utility of digital artifacts [10, 31, 51]. In industry, such experiments have become the norm for guiding digital platform design decisions. One of the high-profile examples frequently cited comes from the Bing search engine team [45]. One of their digital experiments revealed that for sponsored search ad copy text, darkening the blue and green text while making the black text lighter/grayer dramatically increased click-through-rates, apparently resulting in a $10 million increase in annual revenue at Microsoft [47]. In digital experimentation, to some extent, the paradigm has shifted from convincing experimenters and managers to run such controlled experiments as part of their data-driven decision-making efforts to effectively running "digital experimentation at scale" [46, 53]. In this vein, in recent years, industry and academic research has focused on two major directions.

The first area of emphasis for digital experimentation at scale is the introduction of orthogonal testing planes, which allow experimentation platforms to assign users/customers to multiple experiments at the same time in a statistically independent way [68, 81]. By deviating from the classic factorial design approach, these testing plane techniques allow organizations to concurrently launch hundreds, or even thousands, of new experiments every month [33, 45, 47, 68].

The second area of emphasis for digital experimentation at scale is a by-product of the increased size, scale, and pervasiveness of digital experiments. It relates to challenges and opportunities for analyzing the results of digital experiments to derive deeper and more insightful patterns and takeaways [33]. Examples include the sensitivity of experimental results, estimating long-term effects, determining the best overall evaluation criterion metrics, and HTE [33, 43]. In this article, we focus primarily on the latter, namely, detection of HTEs in digital experiments. However, our results also

have important implications for the sensitivity of experiments and the overall value proposition for researchers and managers conducting such online controlled A/B tests. Here, sensitivity refers to the situation in which many treatments might cause a very small change (or ATE) relative to the status-quo or business-as-usual control setting [79]. HTE, however, refers to the situation in which, even when the ATE is large or small or statistically significant or not, the practical and statistical significance might differ in magnitude or directionality for important sub-groups [67, 80]. Our work contributes to the nascent research on detecting user HTEs in digital experimentation settings. To better understand the core set of user characteristics, digital session stages, and online user behaviors that might be relevant for user HTE in digital experiments, in the ensuing sub-section, we turn to the literature on user modeling and customer analytics.

2.2 User Modeling and Customer Analytics

The user modeling literature has focused on the interplay between user characteristics and digital outcomes in various contexts including mobile apps, search, recommender systems, and e-commerce. The major thrusts most related to our work here on HTE have focused on using user characteristics to predict digital behavior or vice versa (i.e., using behavior traces to infer user characteristics). Some studies have attempted to develop holistic user profiles as features for predicting ideal recommendations or personalization strategies. In their personal digital assistant system, Reference [57] found that calibrating recommendations using demographics, interests/preferences, affects such as positive sentiment, and user personality traits resulted in higher perceived accuracy, usage, ease of use, and familiarity. User expertise has also been found to impact assessments of online search relevance rankings [63, 71]. Furthermore, latent and observed user characteristics (including skill and competency) have been shown to improve item recommendation predictions [74, 75]. Some user modeling studies have also represented user characteristics as user or personal embeddings for downstream prediction of outcomes such as search intent, personalized search, social-aware recommendation, text psychometrics, or link prediction [4, 5, 76, 77, 82]. Another direction has been the use of user messaging to infer future online behavior [77, 83]. In some user modeling studies, digital behavior traces were used to predict user characteristics. For instance, Reference [24] predicted gender and age based on users' mobile communication patterns. Similarly, Reference [52] inferred user expertise in the context of question-answering communities based on user's search query terms.

Another related area of research is in the customer analytics space. These studies explore the interplay between user characteristics and the three major stages of the customer lifecycle: acquisition, retention, and expansion [42]. Much of this stream of work can be traced back to seminal research on the use of transaction patterns to predict future customer activity [64]. Several subsequent studies have used demographic and transaction-based user characteristics to predict customer churn and lifetime value [11, 60]. Engagement, which provides a continued signal of a user's interest in a given digital product or service [69], has also been used as a cue for future digital behavior [19, 30]. Another important construct for understanding the digital life of persons is user satisfaction—this can manifest in the form of sentiment, ratings, reviews, and other explicit feedback mechanisms [1, 27]. Unsurprisingly, increased user satisfaction can lead to greater future positive behavioral outcomes [55]. Similarly to the signaling effect of engagement and satisfaction, choice and variety preferences manifesting in historical transactions/logs can indicate current trust and future interest levels [42, 54]. Other important predictors of user behavior include channels and messaging. The channels used to engage with digital products, including the Web, mobile Web, and mobile apps, have been shown to predict future digital consumption patterns [34, 37]. In the same vein, the messaging sent to, and received from, users has important discriminatory potential for forecasting subsequent behaviors and/or predicting user profiles [24, 28, 29]. Collectively, the

user modeling and customer analytics literature has noted the importance of a core set of user constructs—demographics, transactions, engagement, satisfaction, choice, channels, messaging—and how these characteristics can inform and predict key digital user behaviors [42]. In Section 3, we use this body of literature to inspire the design of our proposed digital experimentation framework for modeling user HTE.

2.3 User HTE in Digital Experiments

The user modeling and customer analytics methods used to understand and predict user behavior in online settings have often included feature-based and deep learning-based machine learning methods that use user characteristics as predictors to forecast digital behavior. These techniques are not directly applicable for inferring user HTEs, because the latter necessitates considering differences in how users react in treatment versus control settings [22, 59, 67]. One common approach for detecting user HTEs is the naive two-sample *t-test* method [80]. For example, suppose we run an experiment in two countries (A and B) and want to infer if there is user HTE, meaning that the conditional treatment effect for users in country A or B differs from the overall average. We could use a *t*-test to get the HTE for those in country A and B versus the total experiment population [80]. Because this method is prone to high false-positive rates due to multiplicity (i.e., spurious observed statistically significant user HTE effects), an alternative is to perform a *Bonferroni* correction. However, the Bonferroni correction is prone to high false-negative rates with too many comparisons [80]. Accordingly, other pairwise comparison methods have been proposed that rely on the *false discovery rate (FDR)* techniques commonly used in bioinformatics, because such methods have higher detection power [44, 65]. For instance, *FDR-BH* extends the FDR technique by constructing a design matrix, running a linear regression to get *p* values and coefficient estimates for all subgroups, and using the Benjamini-Hochberg procedure to control false-positive rates [80]. There are other alternatives to BH procedure for controlling FDR that have been proposed in the context of digital experimentation. First is the *FDR-Knockoff* procedure that can be used to identify heterogeneous factors in digital experiments [80]. FDR-Knockoff is based on a "Knockoff" procedure proposed in Reference [13], where knockoff for the design matrix (where columns represent characteristics and rows represents users) are created to select the user characteristics that have heterogeneous effects. FDR-BH and FDR-Knockoff have been found to work well on simulated data and in real-world digital experiments on the Snapchat platform [80]. Second is the *FDR-BY* procedure based on *Benjamini-Yekutieli (BY)* method to control for FDR, that works well when there is are dependencies between user characteristics in experiments [12]. Third, *FDR-dBH* is a dependence adjusted Benjamini-Hochberg procedure where a separate *p*-value cutoff is constructed for each coefficient that is adaptive to the correlation matrix of input characteristics [26]. Finally, *HTE-I*^{3} is a recent method [25] proposed to identify individual treatment effects while controlling for false discovery. These individual treatments can then be processed further to identify factors for HTEs.

While FDR-based methods can be extended from pairwise comparisons to a small group of user HTE covariates evaluated against treatment versus control outcomes, they are not as adept at parsimoniously modeling a large set of covariates, control variables, and also considering non-linear interaction effects between user characteristics and digital outcomes. In recent years, machine learning methods have been proposed to address this gap. For instance, Reference [67] used *classification and regression trees (CART)* to model HTEs in online retail experiments at eBay. Also, *matching plus classification and regression trees (mCART)* was proposed to identify heterogeneous treatment effects while preventing false discovery in clinical trials [61]. Similarly, causal decision trees have been used to explore HTE in contexts where an individual threshold needs to be reached to trigger a treatment [70] or to explore the impact of software patches in multiplayer online games

[36]. An advantage of such methods is that they allow usage of non-linear machine learning methods while providing provably valid statistical inference [8, 73]. A set of methods closely related to causal trees/forests is DML [20]. DML methods can estimate HTE in high-dimensional data contexts, in which non-parametric and non-linear methods are necessary [20, 66]. A strength of DML methods, which use a two-stage classification approach, is that they can use any underlying *machine learning (ML)* methods in the first stage to predict the treatment and outcome scores for a given user, thereby allowing flexibility [48, 66]. Related classes of ML methods for causal inference include doubly robust machine learning and meta-learning [9, 48, 66].

In our proposed framework, we use DML due to the following advantages, and empirically demonstrate its effectiveness versus other comepting methods such as FDR-BH, FDR-Knockoff, FDR-BY, FDR-dBH, Bonferroni, and naive t-tests on simulated and real-world data.

- DML has attractive model agnostic properties in the first stage to capture arbitrary relationships between the user characteristics and treatment or outcome. We can then use a more interpretable linear model in the second stage to derive the HTEs.
- DML can be used to extract the HTEs and ATEs simultaneously from the model. This allows for digital platforms to not only show the effectiveness of an experiment based on ATE but also provide user characteristics that were most influential in identifying these treatment effects.
- The cross-fitting technique [20] employed in these methods can help improve the confidence bounds on the HTEs.

2.4 Research Gaps

Based on our review of prior work on digital experiments, user modeling and customer analytics, and HTE detection methods, we tackle the following research gaps that are important to fill to better model and understand HTE in the context of digital experimentation and beyond. These gaps are closely aligned with the three research questions explicated in the introduction section and filling these gaps provides an important contribution to the literature discussed previously:

- *Paucity of User HTE Work in Digital Experimentation Settings* - As noted in the Introduction and Related Work sections, several recent studies have emphasized the importance of identifying user HTEs [33, 44, 67, 80]. Existing work has shown the potential for highly focused covariates such as country or the recency or frequency of user encounters [67, 80]. However, it remains unclear how to consider the array of user characteristics, experiment factors, and session journey considerations in a unified, cogent manner. These questions are especially hard to answer in situations in which there are many covariates and little theoretic directions on how or why various covariates may lead to user HTE. If there were few covariates or strong theory, then these covariates could be planned to be tested from the outset. Yet, we are working in a space that is often atheoretic and with many options (i.e., features) that can be included.
- *Limited Benchmarking of Existing HTE Methods* - Whereas some studies have examined different t-test and FDR methods [80], and causal tree setups applied to offline/survey datasets [70], there has been limited evaluation of the true-positive and false-positive rates for machine learning methods such as DML on simulated and real-world datasets. Such evaluation is important to shed light on the pros and cons of methods with varying detection rates for future research and practice.
- *No Prior Large-scale Assessment of Prevalence of User Characteristics in Digital Experiments* - How often such user characteristics affect digital outcomes in general, and treatment effects in particular, has been underexplored. Most prior studies have been illustrative—showing re-

sults on one or two experiments [67, 80] or other digital contexts such as gaming [36]. Other studies have leveraged two to three survey/simulation datasets spanning 10 to 50 thousand instances [70]. Whereas such studies have been essential in paving the way for future user HTE research, there remains a need for larger-scale examination in online contexts where digital artifact-oriented treatments are examined on tens of millions of user-sessions (i.e., large-scale online controlled experiments).

In the following section, we discuss our proposed user HTE framework for digital experiment settings. This framework is intended to offer a holistic mechanism for examining user HTE in online experiments, though we believe that our framework can extend to other contexts as well. We then use a series of simulations to benchmark the DML model employed in our framework against existing methods and proceed to apply the framework to a large set of 27 real-world experiments spanning over 1.76 billion user sessions, we are not aware of any study that has examined user HTEs at this scale.

3 PROPOSED USER HTE FRAMEWORK

User modeling and customer analytics research has recently focused on coupling a set of user characteristics with feature-based and deep learning ML models to predict digital user behaviors related to stages of the user/customer lifecycle such as acquisition, retention, and expansion [42]. This general trend is depicted in the top half of Figure 2 and embodied in various prior studies (see Table 1). For instance, engagement, satisfaction, transaction, and demographics coupled with support vector machines or Bayesian networks have been used to predict various stages of the user decision funnel in contexts such as auction fraud, e-commerce conversion, and video-on-demand service churn [2, 19]. Customers' first purchase transaction data have been modeled using Logistic regressions to predict partial churn [54]. Similarly, deep learning models such as variational autoencoders and graph neural networks, coupled with transactional accounts of user skills or purchases, have been used to predict employees' course recommendation conversions [74] or consumer product recommendations [75].

The bottom part of Figure 2 shows our adaptation to the digital experimentation context. Most notably, the framework underscores the importance of session phases, and bifurcates between the broad set of user characteristics encompassed in user modeling and customer analytics settings and how they manifest in digital contexts—with some still appearing as user characteristics while others become experiment treatment or context factors. Furthermore, as noted in Section 2, the classification problem (and accompanying ML models) also differ from the traditional supervised ML binary or multi-class classification setup, because it needs to account for the control versus treatment group assignment [8].

Simply put, the five categories of user characteristics depicted in the bottom half of Figure 2—engagement, satisfaction, lifecycle, transaction, and demographic—are the groups of variables we measure for user heterogeneity using DML. The experimental factors and session phases are incorporated as analysis dimensions—meaning we look at the overall results as well as the results across channels/devices and key session phases. In the following two sub-sections, we describe the user characteristics and experiment factors in digital experiments, as well as the DML methods used to estimate heterogeneity.

3.1 User Characteristics and Experiment Factors in Digital Experiments

When thinking about how user characteristics and experiment factors impact user outcomes and treatment heterogeneity, a user's session phase offers crucial context information. Prior work has underscored the importance of user pathways or trajectories as critical signals or indicators of

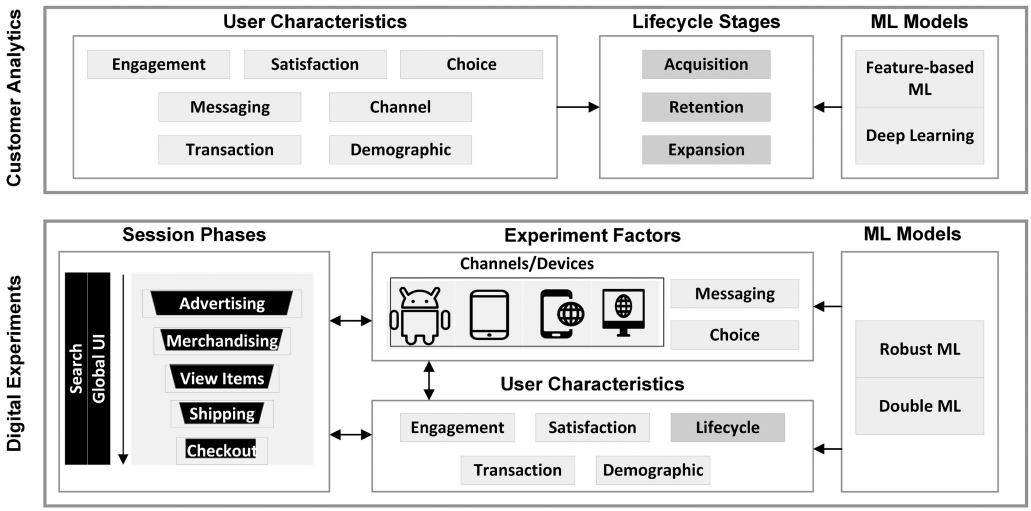


Fig. 2. A framework for user HTE detection and analysis in digital experimentation.

where users are relative to their digital objectives [71]. In the customer analytics and user modeling literature, such customer/user journeys become essential contextual considerations in the digital life of persons [50], which can influence current outcomes and predict future behaviors. Unsurprisingly, in e-commerce settings, session phases such as searching, browsing, viewing items, and delving deeper down "the funnel" have implications for the effect of user behavior [2] on outcomes. We depict such example phases in the bottom left corner of Figure 2. As we later show, the stage of the session at which an experiment treatment occurs—such as a search engine revamp, a merchandising layout tweak, or an ad delivery refinement—can impact which user characteristics are most important.

Table 1 shows the user characteristics and experiment factors incorporated in our framework. In regards to user characteristics, following prior studies [11, 42, 50, 60], we include foundational constructs such as transactions and demographics. *Transaction* characteristics based on prior transaction history have been shown to be highly predictive of various future behaviors [54, 75, 78]. *Demographics* have also been found to have a major effect on digital user behaviors [2, 4, 50]. *Engagement* related to the frequency, recency, and depth of visitation and browsing across digital channels may signify important user heterogeneity characteristics [19, 30, 69]. As noted, *Satisfaction* manifesting in the form of feedback received and/or offered, can affect digital user behaviors [42, 55]. And, finally, whereas *Lifecycle* states such as acquisition, retention, and expansion have been studied extensively in the customer analytics and user modeling literature as a final predicted outcome or dependent variable of interest [42], in the digital experimentation context, such characteristics may have important implications for heterogeneity.

With respect to experiment factors (bottom part of Table 1), *Messaging*, *Choice*, and *Channels/Devices* may occur as facets of the experiment treatment or target audience (i.e., experiment design) [39, 45]. It is important to note that the exact variables employed could vary from context to context. The ones appearing in Table 1 are merely illustrative of constructs of interest and are the ones used in this study—they are not intended to be exhaustive. Indeed, there are many ways to think about and use psychological and behavioral variables in the digital space ready to be used and extend behavioral-based research [3]. In the following section, we show how these user characteristic variables are coupled with DML methods for estimating HTE.

Table 1. User Characteristics and Experiment Factors Incorporated in the HTE Detection Framework

Constructs		Description	Variables	Select User Modeling Work
User Heterogeneity Characteristics	Transaction	Transaction characteristics are based on prior transaction history and represent one of the foundational user/customer constructs.	Buyer type, user status, user designation, credit	[54, 75, 78]
	Demographic	Demographics are another foundational user construct for understanding behavior and heterogeneity.	Gender, country	[2, 4, 50]
	Engagement	In digital contexts, the frequency and recency of visit and browsing on web, mobile, app, email, and other channels.	Disengaged, decreasing, stable, and increasing levels of engagement	[19, 30, 69]
	Satisfaction	Feedback scores, ratings, reviews provided (and received) can be an important indicator of future behaviors and outcomes.	Feedback score	[42, 55]
	Lifecycle	In traditional customer analytics and user modeling, lifecycle stages relate to acquisition, retention, and expansion.	Beginner, stagnant, advanced, expert buyer, expert buyer-seller	[42]
Experiment Factors	Channel	The types of digital channels used to engage and interact with users, including web, mobile web, mobile apps, email, etc.	Web, mobile web, apps, email	[34, 37]
	Choice and Messaging	These constructs may manifest in the design of the digital experiment control/treatment settings.	Control/treatment characteristics	[24, 28, 29]
	Session Stage	The stage of the session related to the treatment.	Advertising, merchandising, search, views, checkout	[50, 71]

3.2 Estimating HTE with DML Methods

We adopt a novel technique to investigate user HTE using DML methods [9, 20, 58]. DML methods are state-of-the-art techniques that can be used to estimate conditional average treatment effects in randomized experiments. These models have several benefits over traditional "single-stage/model" regression-based techniques. First, DML methods are advantageous in situations in which the effects of control variables on the treatment and the outcome cannot be satisfactorily modeled by parametric functions [20]. Second, the cross-fitting techniques [20] employed by these methods can help improve the estimation of the effects (the finite population convergence rates are faster). Finally, and most importantly, they help identify heterogeneous treatment effects on observed characteristics.

Our proposed HTE-DML approach is shown conceptually in Figure 3 and includes user characteristics as covariates to examine user HTE for a given experiment with outcome Y and a session's control versus treatment group affiliation denoted by $T \in \{0, 1\}$. The user characteristics include the five construct categories we discussed earlier that we explore the user HTE. These user characteristics (X) encompass all the observable transactional, demographic, engagement, satisfaction, and lifecycle variables previously mentioned in Table 1. Furthermore, we control for session characteristics from orthogonal testing planes. Orthogonal testing planes allow digital experimentation

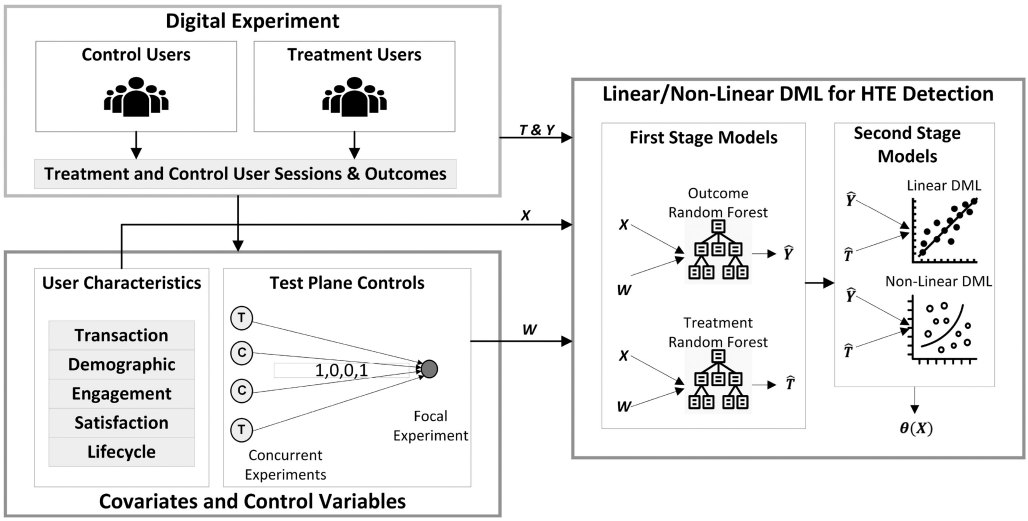


Fig. 3. Proposed HTE-DML method for identifying user HTEs in digital experiments.

platforms to conduct multiple experiments at the same time while minimizing interference from other experiments [68, 81]. Therefore, to minimize any residual bias in our estimated user HTE of a given focal experiment, we control for other experimental treatments that the user session has experienced. These session characteristics are signified by W, which consists of elements that can be continuous or discrete, in Figure 3. In the following sub-section, we elaborate further on the double machine learning mechanism including how HTE-DML controls for the orthogonal test plane.

DML methods first build two predictive models using classic machine learning models to (1) predict the outcome from a set of control variables and (2) predict the treatment from the control variables. These predictive models built in the first stage are then used in the final stage model to estimate the heterogeneous treatment effect. There are two kinds of control variables that can be used in the DML methods. As noted, the first set of variables (X) are the ones that need further investigation for heterogeneity of the treatment effect (e.g., user characteristics), and the second set of variables (W) are other covariates (e.g., session characteristics) that are further used as controls. More formally, DML models assume the following structural equations on data generation:

$$\begin{aligned} Y &= \theta(X) \cdot T + g(X, W) + \epsilon_1, \\ T &= f(X, W) + \epsilon_2, \end{aligned} \tag{1}$$

where T is the treatment indicator, Y is the outcome of interest (e.g., conversion), and $\theta(X)$ is the *conditional average treatment effect (CATE)*. Further, it is assumed that $E[\epsilon_1|X, W] = 0$, $E[\epsilon_2|X, W] = 0$, and $E[\epsilon_1 \cdot \epsilon_2|X, W] = 0$ to make valid inference. However, there are no further assumptions on the functional form (e.g., a linear function) for the functions g and f in the above structural equation models, reducing the possibility of model misspecification and bias [20]. These functions are estimated using machine learning methods, making them attractive for capturing arbitrary non-linear relationships. Therefore, DML models build two predictive machine learning models in the first stage: (1) predicting the outcome Y from the variables X, W and (2) predicting the treatment T from variables X, W. The residuals from these two predictive models in the first stage feed into the final stage to estimate CATE $\hat{\theta}(X)$. Specifically, in this study, we use Linear

DML models where the final stage is a linear model that performs well when the variables used for heterogeneity identification are in a low dimensional space. Furthermore, the Linear DML models offer more interpretable coefficients (and corresponding confidence intervals) for each variable for which we are interested in investigating heterogeneity.

The structural equations in (1) can be rewritten as follows, which helps us estimate CATE:

$$Y - E[Y|X, W] = \theta(X) \cdot (T - E[T|X, W]) + \epsilon. \tag{2}$$

Here we can learn the conditional expectations $E[Y|X, W]$ and $E[T|X, W]$ non-parametrically using machine learning techniques. The decomposition in Equation (2) was originally proposed in Reference [62] and, for completeness, we include here a short derivation to show how Equation (2) comes from Equation (1). First, we can write $E[Y|X, W] = \theta(X) \cdot E[T|X, W] + g(X, W)$ as we assume $E[\epsilon_1|X, W] = 0$. Therefore, we have the following:

$$\begin{aligned} Y - E[Y|X, W] &= \theta(X) \cdot T + g(X, W) + \epsilon - E[Y|X, W] \\ &= \theta(X) \cdot T + g(X, W) + \epsilon - \theta(X) \cdot E[T|X, W] - g(X, W) \\ &= \theta(X) \left(T - E[T|X, W]\right) + \epsilon. \end{aligned}$$

Once we have estimated the conditional expectations, we can find the residuals, which are given by the following:

$$\begin{aligned} \tilde{Y} &= Y - E[Y|X, W] \\ \tilde{T} &= T - E[T|X, W]. \end{aligned} \tag{3}$$

In this study we use random forest-based [18] estimation for identifying $E[Y|X, W]$ and $E[T|X, W]$. That is,

$$E[Y|X = x, W = w] = \hat{Y} = \frac{1}{B^Y} \sum_{b^Y=1}^{B^Y} E[Y|(x, w) \in L_{b^Y}(x, w)],$$

$$E[T|X = x, W = w] = \hat{T} = \frac{1}{B^T} \sum_{b^T=1}^{B^T} E[T|(x, w) \in L_{b^T}(x, w)],$$

where B^Y and B^T correspond to the number of trees in the forest for outcome and treatment, respectively. Further, $E[Y|(x, w) \in L_{b^T}(x, w)]$ and $E[T|(x, w) \in L_{b^T}(x, w)]$ capture the expected value of the outcome and treatment, respectively, in the corresponding leaf nodes $L_{b^Y}(x, w)$ and $L_{b^T}(x, w)$ for the trees b^Y and b^T where x and w fall into.

In the final stage, we estimate CATE $\theta(X)$ [58] using the following model:

$$\tilde{Y} = \theta(X) \cdot \tilde{T} + \epsilon. \tag{4}$$

The estimator for $\theta(X = x)$ from the above equation can then help us identify the treatment effect for a given value of $X = x$. More specifically, the Linear DML methods help us estimate the coefficients for each observed user characteristic.

Once we have estimated CATE using the DML model, we adopt the following procedure to estimate the HTE. Importantly, it is helpful to emphasize that the HTEs we compute is analogous to finding the interaction effect of treatment with each of the observed user variables. First, we create binary (or dummy) variables for all the discrete non binary variables. Then for each of the binary variable X_u, we estimate the CATEs as follows:

$$\begin{aligned} \theta(X_u = 1) &= E[Y^1|X_u = 1] - E[Y^0|X_u = 1], \\ \theta(X_u = 0) &= E[Y^1|X_u = 0] - E[Y^0|X_u = 0], \end{aligned} \tag{5}$$

where Y^1 and Y^0 are the potential outcomes for treated and control user sessions, respectively. Note that we use shorthand notation $X_u = x$ to represent the set of covariate profiles X, where just the variable X_u is set to x. Similarly, $\theta(X_u = x)$ represents the treatment effect among all user sessions with the variable X_u is x. Once we have CATEs from Equation (4), then we estimate the treatment effect with respect to the variable X_u as

$$\beta_{ut} = \theta(X_u = 1) - \theta(X_u = 0). \tag{6}$$

The procedure for estimating the HTE for continuous variable X_u follows a similar procedure and is given by

$$\beta_{ut} = \frac{\partial}{\partial x}\theta(X_u = x). \tag{7}$$

This procedure for the continuous variables is analogous to estimating the marginal effects in econometric analysis [14]. Note that the DML procedure also provides uncertainty estimates of the treatment effects, $\theta(X_u = x)$, that can be used to calculate the uncertainty estimates for β_{ut}.

In our linear DML formulation, the coefficient β_u estimates the main effect of the variable X_u, and the coefficient β_t estimates the average treatment effect. We can, therefore, rewrite Equation (5) for a binary variable X_u as

$$\begin{aligned} \theta(X_u = 1) &= (\beta_0 + \beta_t + \beta_u + \beta_{ut}) - (\beta_0 + \beta_u) = \beta_t + \beta_{ut} \\ \theta(X_u = 0) &= (\beta_0 + \beta_t) - \beta_0 = \beta_t, \end{aligned} \tag{8}$$

which helps understand the derivation for the treatment effect in Equation (6).

As discussed, the orthogonal testing planes are used to run multiple experiments on the platform simultaneously, without interfering with each other [68]. However, there may be practical limitations to the multiple test plane setup [33]. Hence, controlling for other experimental treatments could help identify the effect of the focal experimental treatment on the outcomes and reduce bias in our HTE estimations. Note that here when describing the orthogonal test plane, we refer to the focal treatment (previously T in Equation (4)) associated with each focal experiment f as T_f to distinguish from other treatments a user might experience in a given session. Therefore, in our framework, W includes binary treatment indicators, $T_1, \ldots, T_K$, for the top K experiments that co-occur with the focal treatment. Specifically, the top K experimental treatments are identified using the following procedure. Let $G_f = (V, E)$ be a network over the set of all simultaneously launched experiments along with the focal experiment T_f. Let $numSessions(T_f, T_j)$ be the number of user sessions that experience treatment condition in both experiments T_f and T_j. The nodes of the network are the set of experiments, $V = \{T_f, T_1, \ldots, T_J\}$, the edges are defined if there are user sessions that experience treatment condition in both the experiments, $E = \{(T_f, T_j) | T_j \in \{T_1, \ldots, T_J\}\ and\ numSessions(T_f, T_j) > 0\}$. Finally, the edge weights are defined as the total number of user sessions that experience treatment conditions in both experiments that the edge connects, $W(T_f, T_j) = numSessions(T_f, T_j)$. Therefore, our top K co-occurring experiments are defined as $TopK = argmax_{V_k \subseteq V, |V_k|=k} \sum_{T_j \in V_k} W(T_f, T_j)$.

Figure 4 shows an illustration of how the test plane control features W are derived. In the left side of the figure, the set of focal experiments are depicted by the dashed circle (i.e., dark gray nodes). For a given focal experiment T_f (denoted by the small dashed red circle), we collect the top K co-occurring experiments—the set $\{T_1, \ldots, T_J\}$ depicted as nodes with edges to the focal experiment, and inside the red solid circle. Ultimately, we can represent each user in the given T_f through their treatment co-occurrence vector W that is analogous to a bi-partite graph such as the one depicted in the right-hand side of Figure 4. To the best of our knowledge, this is one of the first studies to control for test-plane-based session heterogeneity that might manifest in

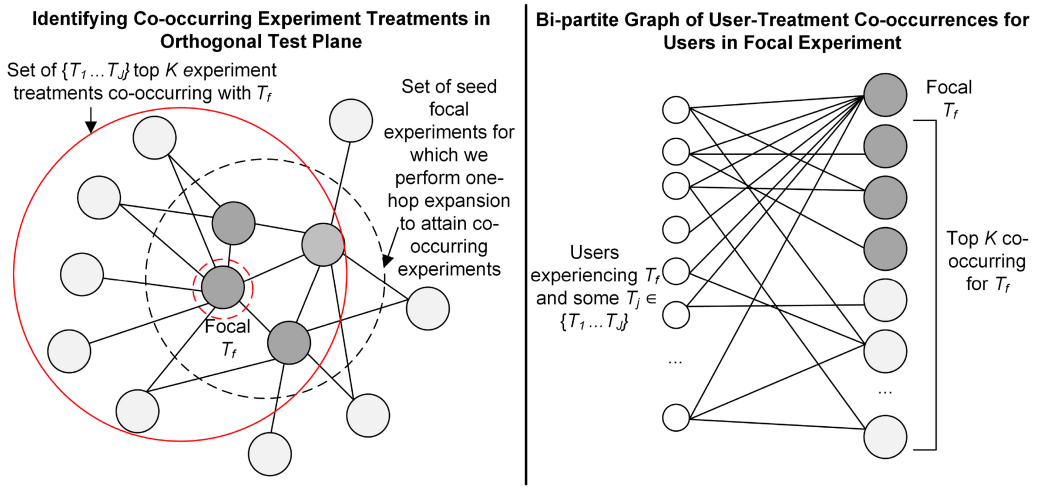


Fig. 4. How orthogonal testing plane is converted to test plane control variables.

digital experimentation settings [33]—an important consideration for isolating the effect of user characteristics on digital outcomes.

3.2.1 Linear DML vs. Nonlinear DML Models. As shown on the right side of Figure 3, the second stage of the DML model can be linear or nonlinear. In the above formulation, we chose a linear DML formulation that has multiple advantages in our setting. First, it allows us to estimate the coefficient β_u, the main effect of the variable X_u, from the model. This is important for us to understand the effect of each user characteristic on the outcome before we understand its interaction effect with the treatment. In other words, we need these main effects to answer RQ2, and linear models are better suited for deriving them. Second, this formulation is easy to interpret rather than using a nonlinear model, which are typically black-box models. This is important for RQ2 and RQ3, since our goal is to shed light on the extent of user heterogeneity in digital experiments. Finally, linear DML also allows us to get uncertainty estimates via asymptotic normality-based theoretical arguments [21] that allow for tighter confidence bounds. However, our framework can be extended to use a nonlinear DML model, such as Causal Forest DML, where the second stage of the model is a Causal Forest [9]. We can still use Equation (6) to derive the interaction effects from Causal Forest DML models. We empirically compare linear and Causal Forest DML models and show that they detect similar interaction patterns (see Section 5.2.1 for more details). Note that both Linear DML and Causal Forest DML employ the same first-stage models that capture the nonlinear relationship between user characteristics and the treatment, as well as user characteristics and the outcome. Hence, as we later show, on our testbeds, both linear and non-linear DML second-stage models perform comparably on our test bed in terms of detection rates.

4 TESTBED OVERVIEW

While real-world data are invaluable for examining the prevalence of user HTE in digital experiments, it does not provide a gold standard for assessing our ability to accurately detect significant HTEs with our methodological framework. Accordingly, consistent with prior work [67, 80], we used a combination of simulated data and real-world data to evaluate HTE-DML (see Table 2). Our real-world data encompassed 27 large-scale digital experiments conducted over a 4-month period at a major two-sided marketplace headquartered in the United States and having a global buyer and

Table 2. Digital Experimentation Testbed Overview

Testbed Type	Testbed Characteristic	Description
Real-world Experiment Data	Number of Experiments	27
	Session Stages	Merchandising (14), Search (7), Advertising (2), View Item (2), Checkout (1), GlobalHeader (1)
	Average Number of Sessions	65.4 million (max = 314 million)
	Total Number of Sessions	1.76 billion
	Average Number of Users	6.8 million (max = 35.9 million)
Simulated Experiment Data	Number of Datasets	6×20 null, 6×20 alternate
	Number of Sessions Per Dataset	300 million
	Number of User Variables Per Dataset	20 discrete, 10 continuous

seller base. These 27 experiments spanned six different session stages, including merchandising, search, advertising, item viewing, checkout, and the global web page header (with merchandising and search being the most common). On average, each experiment encompassed about 65.4 million user sessions and 6.8 million users. Collectively, the real-world dataset spanned 1.76 billion total user sessions. For each user session, as noted in Section 3, we observed an outcome Y (whether or not the session ended in a conversion), whether or not the user was in the treatment or control group T, and the user X and session heterogeneity characteristics W.

Our second dataset was a collection of simulated digital experiments [80]. In total, we incorporated 240 settings with each setting spanning 300 million user sessions and 30 simulated user variables (namely, 20 discrete and 10 continuous). Most notably, we generated two types of gold standards, one under a null condition and one in which is an alternate setting in which the null hypothesis is false under different scenarios. In the null setting, all user variables had the same population values across groups (i.e., the null hypothesis was true). Conversely, in the alternate setting, certain variables were set to low, medium, or high levels effect sizes. Further details regarding the simulated dataset generation process appear in the ensuing section.

In the following two sections, we describe how we used these two datasets to evaluate HTE-DML and infer the extent of user HTE in real-world digital experiments as follows:

- *Evaluating HTE Detection on Simulated Data* - In Sections 5.1 and 5.2, we compare HTE-DML to existing baseline and benchmark detection methods on the various null and alternate simulated experiment datasets. *This evaluation relates to our first research question that asks how accurately user HTEs can be detected in digital experiments.*
- *Evaluating HTE Detection on Actual Experiment Data* - Although it is impossible to report true-/false-positive and -negative rates on actual experiments, as ground truth is unknown, in Section 5.3 we examine the consistency of the HTE effect sizes and p values on the 27 real-world experiments as an indicator of detection accuracy/quality. *This evaluation also relates to our first research question related to detection accuracies for user HTE identification methods in digital experiments.*
- *Analysis of User HTE Prevalence* - After demonstrating the efficacy of HTE-DML in Sections 5 and 6, we report the prevalence of user HTEs in our 27 real-world digital experiments. *This analysis relates to our second and third research questions focused on empirical insights into how user characteristics impact outcomes and HTE.*

5 EVALUATION: HTE DETECTION

In this section, we empirically demonstrate the performance of our HTE-DML framework for detecting user heterogeneity. We first evaluate HTE-DML and commonly used comparison methods such as FDR-BH, t-test, and Bonferroni t-test on simulated data (Sections 5.1 and 5.2). The purpose of such evaluation on simulated data is to make sure we can assess our FDR and power to detect heterogeneity in a controlled environment where we are aware of the ground truth. Additionally, we compare the performance of HTE-DML to the aforementioned baseline and benchmark techniques on the actual real-world digital experiments (Section 5.3).

5.1 Experiment Simulation Setup

Our simulation aims to replicate conditions similar to digital experimentation platforms, where our framework could be used to discover user heterogeneity, and to compare our results with other common techniques. We used the parameters from our partner digital experimentation platform and injected specific subpopulations with a treatment effect that we intend to discover. Our simulations generated user sessions with a treatment indicator, user and session characteristics, and an outcome (e.g., conversion) as described below. For each user session, we generated a set of user characteristics $X_1, \ldots, X_U$, both discrete and continuous, that describe the user. We made sure our characteristics are generated as closely related to the customer analytics framework as possible to replicate the structure and challenges faced by the business managers on an e-commerce platform. Each user session is associated with a treatment indicator T_f for the focal experiment and an outcome Y. Finally, we generated session characteristics to emulate multiple *other* experiments that the user experiences in a given session. Specifically, every session also includes $T_1, \ldots, T_K$ treatment indicators for the set of K *other* experiments that the user experiences in a given session. We then generated two sets of experimental data: (1) where there is no HTE introduced based on user characteristics (the null setting) and (2) where HTEs based on user characteristics are injected in our data generation process (the alternate alternate setting). The former experimental simulated data helps us identify the false discovery rate, that is, the possibility of finding HTEs when the null hypothesis that there are no HTEs is true. The latter data setting simulates when the alternative hypothesis is true and allows us evaluate the power to detect the induced heterogeneity.

The process we follow to generate simulated user sessions begins with creating a user base with a set of characteristics $X_1, \ldots, X_U$. We then generated user sessions with a possibility of a user being in multiple sessions. As observed on our e-commerce platform, we followed Pareto distribution for the number of sessions each user has in an experiment. Each user session generated is randomly assigned to a focal treatment indicator (T_f) with probability p_{ft}. We further generated other treatment indicators $T_1, \ldots, T_K$ such that the total number of treatment indicators for each session, $\sum_{k=1}^{K} T_k$, follows a Pareto distribution. Pareto distribution for the number of sessions per user and the number of treatments indicators per session is known to model human behavior [7]. Finally, we used the following generative model for the outcome Y for each session based on user and session characteristics:

$$Y = f(T_f, X_1, \ldots, X_U, T_1, \ldots, T_K), \tag{9}$$

where the functional form of f is both linear and nonlinear in our experiments. We generate two sets of data where the outcomes for each record follow a null hypothesis of no HTEs or the outcome for reach record follows the alternative hypothesis with varying degrees of heterogeneity induced.

5.1.1 Our DML-based Approach and Comparison Methods. We evaluated the performance of our DML-based HTE-DML technique on the simulated data. In our first stage, we used random forest-based classifiers to predict outcome Y and the treatment T. We employed cross-validation

techniques and optimized the area under the ROC curve to learn the random forest parameters. We then used the final stage model to estimate the CATE, followed by the HTE procedure in Section 3.2 to identify significant interaction of user characteristics and the treatment.

We compared our results to techniques commonly used by prior researchers and practitioners to identify HTE, as discussed in Section 2.3. The naive approach in detecting heterogeneity, based on a discrete variable, is to perform a two independent-samples t-test. For instance, if there is a binary variable X_u, then we can perform two independent-sample t-test for outcome sets $\{Y^1 - Y^0 | X_u = 1\}$ and $\{Y^1 - Y^0 | X_u = 0\}$, where Y^1 and Y^0 are the outcomes for treated and control, respectively. We refer to this approach as *Naive t-test* in our experiments. It can be easily shown that this naive t-test approach could lead to spurious HTEs due to multiple testing problems [80]. One simple alternative is to perform a Bonferroni correction to mitigate family-wise error rate, which we refer to as *Bonferroni t-test*. Though this correction can reduce false positives, it is very conservative as the number of tests increases, leading to low statistical power. Next, we include multiple methods that are based on the controlling false discovery rate. First, recent work has improved the false discovery rate in detecting HTEs for online controlled experiments [80]. More specifically, the authors proposed a Benjamini-Hochberg procedure [15] to control for FDR, which we refer to as FDR-BH approach in our experiments. The authors also proposed a knockoff procedure [13] to further discover heterogeneous factors, which we refer to as FDR-Knockoff approach in our experiments. Second, further extensions of the FDR-based methods are proposed in the literature for identifying treatment effects in large-scale experiments [12]. Specifically, the authors proposed a BY method [16] for controlling FDR, which we refer to as FDR-BY approach in our experiments. The authors also proposed a dependence-adjusted BH procedure [26], which we refer to as the FDR-dBH approach.

We also compare our results to a recent procedure for identifying individual treatment effects controlling for FDR [25]. This procedure is designed to identify individual users that have positive and non-positive treatment effects. Once these individual users are identified by the procedure, we then post-process these users to find characteristics that are different from *all* the users. Specifically, we conduct a t-test for each user characteristic between the users identified by the procedure and all the users and flag those characteristics that are significantly different. We refer to this procedure as HTE-I^3 in our experiments. Though this procedure is not designed to detect HTEs in data, we included it to show efficacy of directly detecting HTEs rather than post-process other known methods in literature. Finally, we use mCART proposed in Reference [61], a tree-based procedure, for detecting HTEs. Once HTE tree is constructed by the procedure, we post-process that tree to identify all the characteristics that appear in the non-leaf nodes of the tree, which we refer to as HTE-mCART in our experiments. The idea is that any characteristic that appears in a non-leaf node is important in defining the subpopulations that exhibit HTEs, identified by the mCART procedure.

5.2 Simulation Results

In the prior section, we provided a general experimental simulation setup for evaluating the discovery of user HTE in digital experiments. In this section, we provide specific details of our data generation process and provide simulation results to compare the performance of HTE-DML with other techniques. Our main goal is to evaluate the false discovery rate when there are no HTEs and power of detecting HTEs. To provide interpretable results and cleanly induce HTEs, we generated outcome (Y) using the following model:

$$logit(Y) = \beta_0 + \beta_f * T_f + \boldsymbol{\beta_X} * \boldsymbol{X} + \boldsymbol{\beta_{Xf}} * \boldsymbol{X} * T_f + \sum_{k=1}^{K} \beta_k * T_k + \epsilon.$$

Table 3. Comparison of Percentage Significant HTEs Identified under the Null Hypothesis Where There Is No HTE and under the Alternate Hypothesis

Method	Percentage of Significant HTEs Identified (Std. Err)											
	Null Hypothesis						Alternate Hypothesis					
	Linear model			Nonlinear model			Linear model			Nonlinear model		
	Large β	Medium β	Small β	Large β	Medium β	Small β	Large β	Medium β	Small β	Large β	Medium β	Small β
HTE-DML	**9 (3.5)**	**11 (3.6)**	**9 (3.7)**	**9 (1.8)**	**10 (2.7)**	**7 (1.6)**	**93 (1.3)**	**86 (1.9)**	**86 (1.9)**	**81 (1.9)**	**83 (1.5)**	**83 (1.4)**
Naive t-test	30 (3.3)	36 (3.0)	32 (3.3)	31 (2.2)	32 (2.9)	30 (2.5)	73 (0.7)	71 (1.1)	71 (1.3)	69 (1.3)	71 (1.1)	70 (1.2)
Bonferroni t-test	25 (3.7)	30 (3.5)	25 (3.7)	16 (2.0)	20 (2.8)	17 (2.3)	62 (0.9)	58 (1.3)	60 (1.6)	57 (1.7)	60 (1.3)	58 (1.5)
FDR-BH	25 (3.9)	31 (3.5)	25 (4)	18 (2.1)	21 (3)	19 (2.3)	72 (0.8)	69 (1.1)	69 (1.6)	67 (1.4)	70 (1.4)	68 (1.4)
FDR-BY	24 (3.8)	30 (3.4)	24 (3.8)	16 (2)	20 (2.9)	17 (2.2)	72 (1)	67 (1.4)	69 (1.8)	65 (1.7)	70 (1.4)	68 (1.5)
FDR-dBH	22 (3.3)	28 (3.1)	25 (3.6)	20 (2.5)	18 (2.1)	18 (2.3)	61 (0.9)	58 (1.1)	60 (0.8)	54 (3.4)	57 (1.7)	57 (1.6)
FDR-Knockoff	20 (3.8)	20 (3.8)	17 (3.7)	**6 (2.3)**	**6 (2.1)**	**14 (2.9)**	66 (5.6)	54 (4.5)	25 (3.8)	41 (4.4)	**70 (4.9)**	44 (4.6)
HTE-I^3	**6 (0.7)**	**8 (0.8)**	**7 (0.8)**	14 (0.4)	14 (0.4)	15 (0.2)	14 (0.3)	12 (1.0)	13 (0.9)	14 (0.5)	15 (0.02)	14 (0.6)
HTE-mCART	19 (0.7)	19 (0.8)	16 (0.8)	**6 (0.7)**	**5 (0.8)**	13 (0.8)	59.2 (0.3)	49 (1.0)	23 (0.9)	37 (0.4)	63 (0.4)	39 (0.5)

We bold the cells that the lowest (or the highest) percentage detected in null (or alternate) hypothesis setting and the ones that are statistically indistinguishable from the lowest (or the highest).

Our user variables X contain both discrete and continuous variables. More specifically, we generated three dummy variables that take five unique values, five independent binary variables, and 10 continuous variables. Therefore, the user vector X is of length 30. We chose these specific values, as they align well with our real data from the aforementioned e-commerce platform. Furthermore, this procedure to generate the outcomes also aligns with the established simulation framework in the literature [35]. We then assigned the focal treatment indicator T_f for each session with probability $p_{ft} = 0.5$. We further set the number of concurrent experiments that a user session can be part of as 20 (i.e., $K = 20$). For the non-linear models, we generated the outcomes using the above equation, but we introduced a random set of 40 two-way interactions among user variables and treatment indicators $(T_1, \ldots, T_K)$, except for the focal treatment T_f. Furthermore, we generated multiple datasets by selecting each of the β values from a mixture of two normal distributions with mean at μ_β and $-\mu_\beta$, respectively. We chose a bimodal distribution to make sure we represent the positive as well as negative effects that can be identified by our models. Also, note that the $abs(\mu_\beta)$ indicates the signal strength, specifically, the mean for the interaction coefficients, $\boldsymbol{\mu}_{\boldsymbol{\beta}_{Xf}}$, indicate the strength of the HTE introduced in our outcome model. In our setting, we chose three levels of mean for the β coefficients, which we regard as large ($abs(\mu_\beta) = 1$), medium ($abs(\mu_\beta) = 0.5$), and small ($abs(\mu_\beta) = 0.1$). Finally, we generated 20 simulated datasets for each experimental setting as explained in Section 4. The results for both experimental settings along with the comparison methods are show in Table 3.

In our first simulation data setting where there were no HTEs (i.e., the null setting), we set $\boldsymbol{\beta}_{Xf} = \mathbf{0}$, that is, in this setting there were no interactions between the user characteristics and the focal experiment's treatment indicator. Figure 5 visualizes the results from the analysis, where the Y axis shows the average percentage of significant interactions (at $\alpha = 0.05$) identified by HTE-DML and each of the comparison methods. A coefficient is significant if the confidence interval does not include 0. We can see that HTE-DML identified a significantly lower percentage of coefficients as significant compared to other methods. This trend is true for both linear and non-linear simulation models for the outcome variable. Furthermore, in most cases, the confidence interval for the percentage of significant coefficients identified (error bars) includes 5%, which shows that we are within the false-positive rate expected at an $\alpha = 0.05$. That is, the HTE-DML method does have a bounded false discovery rate compared to other methods. One reason for lower false discovery rate by the HTE-DML method is because it detects the HTE for a given user characteristic while controlling for other user and session characteristics. With respect to the comparison methods, as expected, the naive t-test had the highest false discovery rate followed by the Bonferroni t-test.

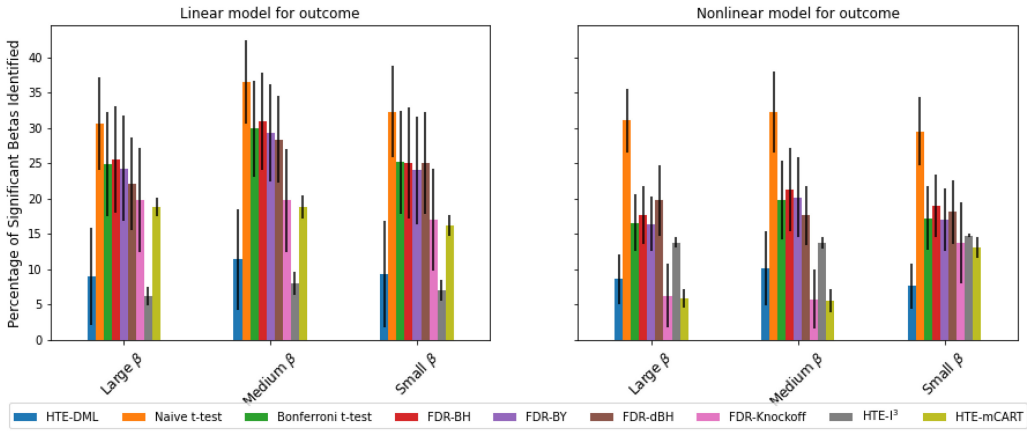


Fig. 5. Comparison of percentage significant HTEs identified under the null hypothesis where there is no HTE.

Among FDR-based methods, FDR-Knockoff seems to outperform most of the methods, though it still garnered markedly higher false discoveries than HTE-DML. The HTE-mCART method has similar number of false discoveries as the FDR-based methods. Finally, HTE-I^3 does have lower false discoveries than HTE-DML in the linear model, they are not significantly different, and from Table 3 we can see that in the alternate hypothesis setting, HTE-I^3 has very low power to detect HTEs. The low false discovery rate associated with HTE-DML is important for inferential validity and to minimize the effort expended by analysts and decision-makers on irrelevant user subgroups.

In our second simulation data, we introduced significant HTE interactions (i.e., the alternate setting). More specifically, we set $\boldsymbol{\beta}_{Xf}$ to large, medium, and small values as explained earlier. We introduced a treatment interaction with all the discrete variables. The average percentage of significant interactions identified by each method is shown in Figure 6. Figure 6 shows that the average percentage of statistically significant interactions identified by the HTE-DML method was itself statistically significantly higher than other comparison methods. For instance, HTE-DML identified over 80% of the interactions in all experiment settings. It yielded detection rates that were 10 to 20 percentage points higher than FDR-based methods on all six settings, and 20 to 30 points better than HTE-mCART, t-test, and Bonferroni t-test. Finally, HTE-I^3 has a very low power to detect HTEs in this alternate setting, which is not surprising, as the method is not designed for that purpose. That is, HTE-I^3 is designed to identify users who have positive or non-positive treatment effects may not help to identify factors for HTE. Overall, these results highlight the HTE detection capabilities of HTE-DML relative to comparison methods. Although the average performance of all methods drops in the non-linear setting (right chart in Figure 6), HTE-DML still yields a higher true-positive rate in terms of percentage of significant interactions identified. Collectively, the results from the null and alternate setting experiments underscore the efficacy of HTE-DML for HTE detection. From a downstream implications perspective, this ability to better detect HTEs could allow decision-makers to siphon the wheat from the chaff to accurately find subgroups with statistically significant HTE, thus allowing managers to customize specific features evaluated in an experiment. Overall, our simulation analyses presented in this section supports the notion that HTE-DML has high accuracy in detecting user HTEs in digital experiments (RQ1).

5.2.1 Simulation Results Comparing Linear DML and Causal Forest DML. We use Linear DML in most of our analysis because of its advantages, such as computing the main effects, interpretability,

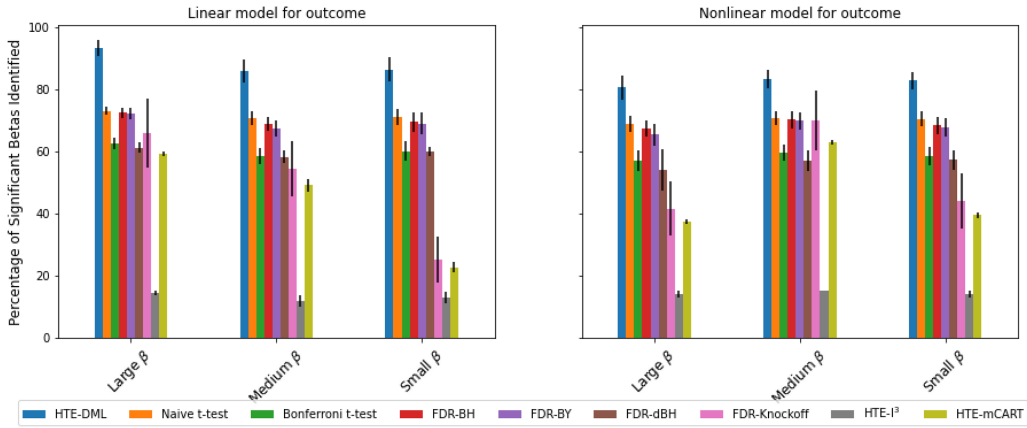


Fig. 6. Comparison of percentage statistically significant HTEs identified under the alternate hypothesis.

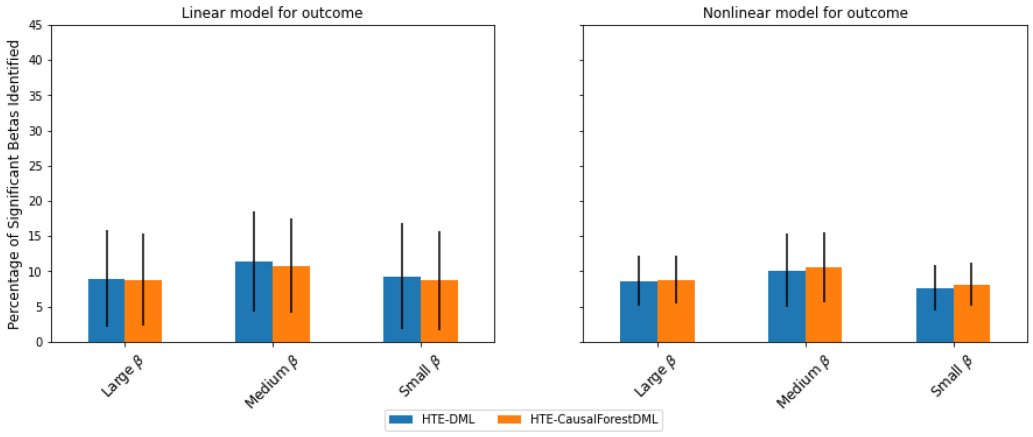
and tighter confidence bounds, as explained in Section 3.2.1. In this section, we compare the detection results of the Linear DML model with Causal Forest DML model, that uses a nonlinear (causal forest) model in the second stage. We use the same simulation test bed with varying strength of HTEs controlled by the size of the coefficient (β). In our first simulation data, where there were no HTEs (i.e., the null setting), we can see from Figure 7(a) that Linear DML and Causal Forest DML have very similar false discovery rates in all six settings. Not surprisingly, we can see that Causal Forest DML has slightly better performance in nonlinear outcome model, though they are not significantly different than Linear DML. Finally, Figure 7(b) compares the detection power of Linear DML and Causal Forest DML in simulated data, where there were significant HTEs introduced (i.e., alternate setting). We can again see that both methods have no significant differences in all six settings. This further validates our choice of using Linear DML model, which has multiple advantages without losing the capability of detecting the HTEs.

5.3 Comparison of HTEs Identified on Real-world Digital Experiments

Whereas the prior section focused on evaluating HTE-DML and comparison methods on simulated data, we also compared the HTEs identified by each of the methods on all 27 actual digital experiments observed in the e-commerce platform. In the absence of a gold standard for true-positives/-negative user HTEs, we used the variance in effect size, directionality, and statistical significance for our user variables as a proxy of detection quality. The key intuition being that because the t-test, Bonferroni t-test, and FDR-BH methods are more prone to false positives and false negatives, their user HTE detection profiles would more closely resemble those of a random model (i.e., one with a detection AUC closer to 0.5) [17]. Guided by this intuition, we organized the data such that for each method (i.e., HTE-DML, naive t-test, Bonferroni t-test, HTE-I^3, and FDR-based methods),[1] each experiment was a row vector with columns representing the p values and the direction of significance identified for each user characteristic (i.e., the user HTE significance and effect direction for that experiment). The direction of significance was set to 1 if the HTE was positive and statistically significant, -1 if the HTE was negative and significant, and 0 if the HTE was not significant. Based on our aforementioned intuition, if the HTE estimates detected are accurate, then the effect directions and statistical significant findings would likely exhibit certain patterns

[1]In this analysis, we have not included the results for HTE-mCART as our post-processing step only identifies characteristics that appear in a non-leaf node of the tree. Therefore, we cannot quantify the interaction coefficient in this case.

(a) Under the Null Hypothesis setting where there is no HTE



(b) Under the Alternate Hypothesis setting

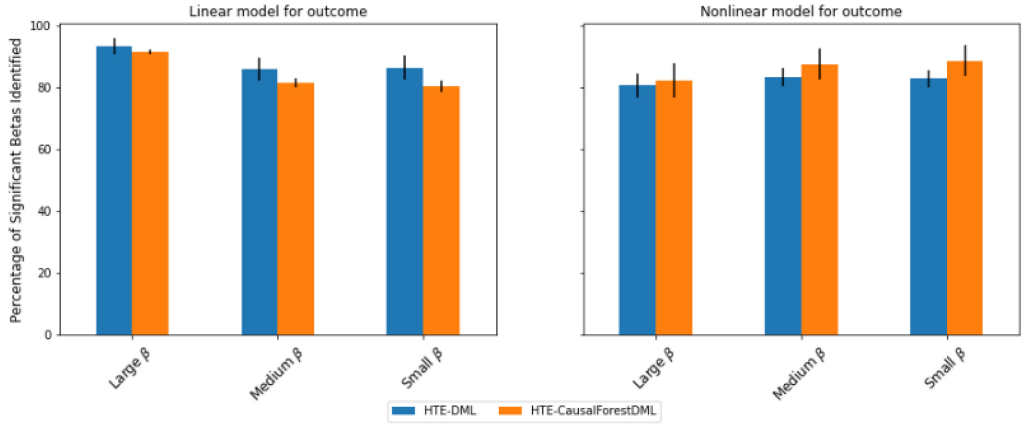


Fig. 7. Comparison of linear DML and causal forest DML in terms of percentage of statistically significant HTEs identified.

more consistently observed across experiments. Conversely, methods with lower accuracy—and hence greater randomness—would have look more sporadic across experiments.

To visualize the consistency or sporadic nature of the experiment vectors generated by each HTE detection method, we used the *t-distributed stochastic neighbor embedding (t-SNE)* plots [72] to visualize each experiment point for a given method in a series of two-dimensional plots.[2] In other words, we ran *t*-SNE separately within the five matrices of experiment vectors (e.g., the HTE-DML matrix, the FDR matrix, etc.). The *t*-SNE algorithm embodies the same intuition as a non-linear principal component analysis in that it can accurately and efficiently represent the most essential variance in a two- or three-dimensional space [72]. *t*-SNE plots are very useful for our setting as they can help reduce the number of dimensions such that similar points (consistent HTEs across experiments) are highly likely to be shown as nearby points in the lower dimensional space while

[2]The HTE coefficients identified by all the FDR-based methods, i.e., FDR-BH, FDR-BY, FDR-dBH, FDR-BH-Knockoff, are the same. Therefore, we only show one visualization for all these methods together.

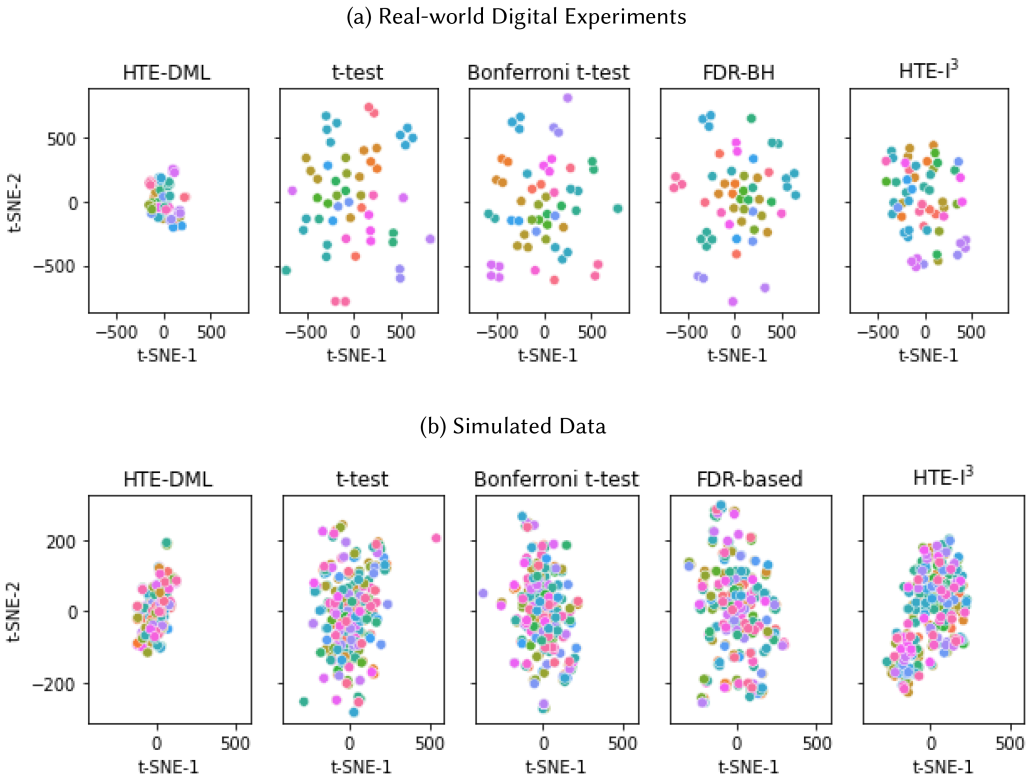


Fig. 8. t-SNE-based visualization of p value and significance of user heterogeneity identified by each method. Each input row vector was an experiment with the p value and direction of significance $(-1, 0, 1)$ of the HTE identified for each of the user characteristics in that experiment. The fourth panel in both figures represents all FDR-based methods: FDR-BH, FDR-BY, FDR-dBH, and FDR-BH-Knockoff.

dissimilar points (inconsistent HTEs) are highly likely to be shown as distant points. We depict the first two dimensions of t-SNE as X axis and Y axis, respectively. The resulting two-dimensional scatter plots appear in Figures 8 and 9.

Each point in Figure 8(a) represents one of the 27 experiments in our testbed. To add additional robustness to our observed results, for each experiment, we randomly resampled 10 million user sessions (note that the average number of session peer experiment was over 65 million) and reran the respective HTE detection methods. Hence, the total number of experiment points in each plot are greater than 27—further, points of the same color represent a given 10 million session sampling for that respective experiment. Looking at the results appearing in Figure 8(b), we observe a couple of important findings. First, as suspected, we observe that the points representing HTEs identified by HTE-DML are very close to each other compared to the other methods. With respect to the comparison methods, pursuant with our simulation results in the prior section, the FDR-based plots appear relatively closer/tighter than the Bonferroni and naive t-test methods. HTE-I^3 method also has relatively tighter points; however, these could be because of relatively low power to detect HTEs (i.e., most HTEs are not significant) as we have observed in our simulation analysis. Second, within each individual plot, we see that the points for multiple sampled sessions from the same experiments appear very close to each other, showing that all four methods also have consistency across their respective 10 million session samples. This also offers a type of validation

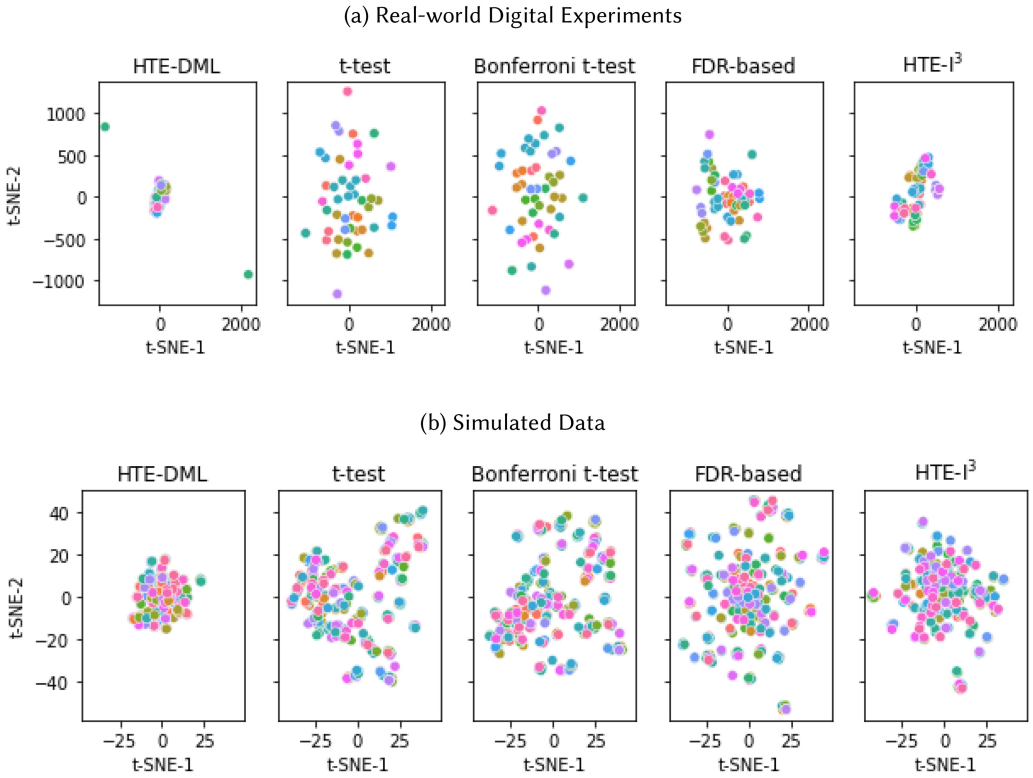


Fig. 9. t-SNE-based visualization of upper and lower effect size bounds of the user HTEs identified by each method. Each input row vector was an experiment's identified HTE upper and lower effect size confidence interval tuples for each of the user characteristics in that experiment. The fourth panel in both figures represents all FDR-based methods: FDR-BH, FDR-BY, FDR-dBH, and FDR-BH-Knockoff.

of our intuition for using the t-SNE plots to illustrate performance in a real-world context devoid of gold-standard labels. To further validate t-SNE-based validation, we performed a similar visualization of the results obtained from the simulated data where we introduced significant HTE interactions (i.e., the alternate setting), shown in Figure 8(b). We can observe similar patterns in the visualization based on the simulated data, where the HTEs identified by HTE-DML are very close to each other compared to other methods. Overall, the results further reinforce the notion that HTE-DML's detection capability seems more accurate and robust on simulated and actual large-scale digital experiments.

We also performed a similar analysis using the HTE size estimates for each of the four methods on various user characteristics across the 27 experiments. Similarly to prior analysis, we created an input dataset for each detection method in which each row was an experiment. However, the columns now represented the upper and lower bounds of the 95% confidence intervals of the user HTEs for a given method. We again visualized this multidimensional HTE effect size data using t-SNE plots. The visuals appear in Figure 9(a). Once again we observe that data points in HTE-DML plot are very close to each other (relative to comparison methods), suggesting consistency in the effect sizes identified for user HTE. Again, we visualize the results from simulated data from alternate setting in Figure 9(b), which exhibits similar patterns. Collectively, these results further underscore the credibility and efficacy of the user HTEs identified by the HTE-DML method.

Researchers and practitioners could incorporate user HTE insights guided by HTE-DML, with greater confidence, into the design of personalized or sub-group level platform features. Having evaluated the HTe detection capability, in the ensuing section, we delve deep into the user characteristics that drive HTEs in digital experiments and offer empirical insights regarding the pervasiveness of user HTEs.

6 ANALYSIS: PREVALENCE OF HTES IN DIGITAL EXPERIMENTS

Having shown the suitability of HTE-DML for detecting user HTEs, in this section we examine the prevalence of user HTEs in digital experiments. In Section 6.1, we report on how frequently user characteristics significantly impact overall session conversion outcomes (RQ2). Section 6.2 reports results for how often user characteristics interact with the treatment; that is, user HTE (RQ3).

6.1 Impact of User Characteristics on Conversion Outcomes

As noted, the user modeling and customer analytics literature has already shown that user constructs such as prior transactions, demographics, engagement, satisfaction, and lifecycle stages can all have a profound impact on digital user behavior related to conversions, churn, and other outcomes [2, 42, 50, 56]. In our HTE-DML method formulations depicted in Section 3.2, this equates to examining the significance of the effect of β_u on Y. It stands to reason that just as these constructs have been found to predict behavior in prior non-experimentation studies, such user constructs should also inform conversion outcomes in digital experiment settings. However, how predictive they are is an important pre-cursor to user HTE analysis, because one would expect the number of β_{ut} user HTEs that impact Y to be lower than the number of main effects β_u.

Figure 10 shows the impact of user characteristics on our binary conversion outcome. The top chart depicts the overall effect across all 27 digital experiments, whereas the two charts in the bottom row show the effects within the subset of experiments related to the merchandising and search session stages. In each chart, the x-axis depicts each of the 18 user variables previously presented in Table 1. The y-axis shows what percentage of the experiments that variable is significantly impacting the conversion outcome. Bar chart colors correspond to the five construct categories with which a respective user variable is associated. Looking at the overall results, we see that most user variables significantly impact session conversions in 20% to 40% of the experiments. This is especially true for the transaction, demographic, and satisfaction variables. Within engagement, as expected, the decreasing, stable, and increasing engagement levels are most likely to be significant. The category most significant overall are the lifecycle variables, with users that are advanced, expert buyers, or expert buyers and sellers significantly affecting conversion behavior. It is important to note that we intentionally omitted the effect sizes and directionality, because the purpose was simply to show how pervasive these user characteristics are.

Looking at the bottom part of Figure 10, we see that the percentage of experiments where user variables are significant does vary based on the session stage where the experiment treatment occurs. For instance, the aforementioned lifecycle variables are most pronounced on the search experiments—where the advanced and expert buyers and sellers have a significant impact on conversion in over 80% of experiments. In the search experiments, many of the other user variables are significant around 20% of the time. Conversely, in the merchandising experiments, many of the transaction, demographic, engagement, and lifecycle variables are significant in nearly 40% of the experiments. Most notably, user engagement variables such as decreasing, stable, or increasing levels of engagement become more crucial in such experiments.

To delve deeper into the impact of user characteristics on session conversions, we examined the significance of user variables within different experiment channels/devices—namely, mobile web, web, Android app, and iOS app. The results appear in Figure 11. It is important to note that whereas

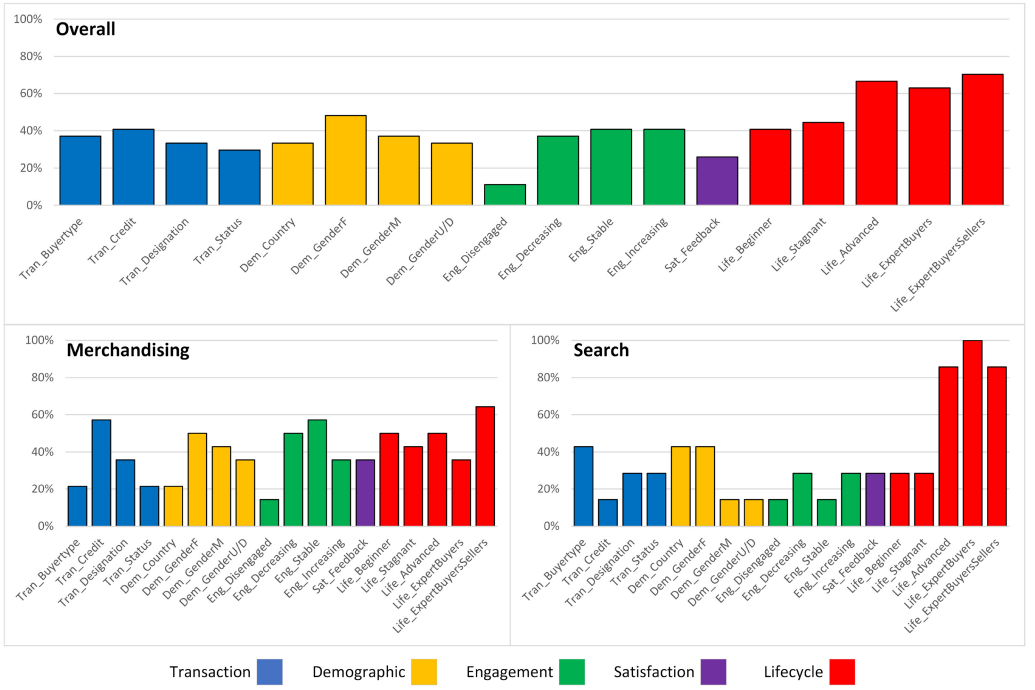


Fig. 10. Impact of user characteristics on conversion outcomes: Overall and for key session stages.

experiment session stages are mutually exclusive, experiment channels may overlap, meaning a given digital experiment may focus on all app users (Android and iOS). Looking at the four charts, we observe a few interesting patterns. First, the lifecycle variables seem most significant in predicting session outcomes in experiments that include mobile web and web sessions and are least influential for experiments encompassing iOS app-based sessions. Second, satisfaction has a more pronounced affect in experiments involving mobile web and Android user sessions. Third, transaction, demographic, and engagement variables also vary in terms of the proportion of experiments in which they are significant. Collectively, in regards to RQ2, these results underscore the importance of the five categories of user constructs for modeling user session behavior and are consistent with results from prior non-experiment digital modeling research. Further, the results lend credence to our framework's focus on the interplay between user characteristics, session stages, and channels/devices in digital experiments. In the ensuing section, we show how pervasive these characteristics are for user HTE in digital experiments.

6.2 Impact of User Characteristics on HTE

Figure 12 shows the user HTEs for our digital experiments. Similarly to the prior section, the top chart depicts the overall effect across all 27 digital experiments whereas the two charts in the bottom row show the effects within the subset of experiments related to the merchandising and search session stages. In each chart, the x-axis depicts individual user variables while the y-axis shows the percentage of experiments in which that variable's HTE is significant. Bar chart colors correspond to the five construct categories with which a respective user variable is associated.

Looking at the overall HTEs (top chart), unsurprisingly, the proportion of significant HTEs is lower than the proportion impacting conversion outcomes. This is partly to be expected, because randomly assigning users to control versus treatment settings lessens the β_{ut} interaction effect.

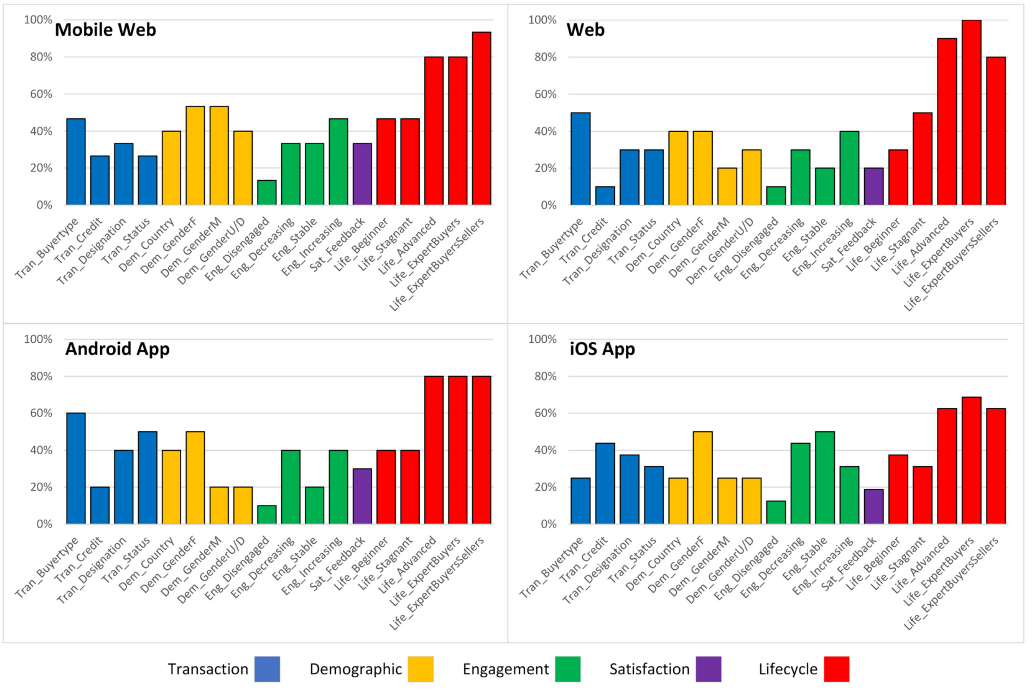


Fig. 11. Impact of user characteristics on conversion outcomes by channels/devices.

However, we do still observe quite a bit of significant user HTEs—in 10% to 30% of experiments. These effects are most pervasive for engagement variables, but we also see demographics and the beginner expertise state being significant in 10% to 20% of experiments. The bottom part of Figure 12 shows the percentage of experiments where user HTEs are significant for merchandising and search session stages. We observe a few notable differences. For instance, the lifecycle variable HTEs are once again are most pronounced on the search experiments but with a key difference. Whereas we noticed that the main effects were most pronounced for the advanced and expert buyers and sellers, for user HTE the only significant lifecycle states are beginner (in over 40% of search experiments) and stagnant users. The stable engagement variable is highly significant for both search and merchandising session stage experiments. Similarly to the main effects of user variables on conversions in merchandising experiments, decreasing and increasing engagement levels are once again more pronounced in statistical significance when considering user HTEs.

We also examined the significance of user HTEs within different experiment channels/devices, as done for the main effects in Section 6.1: mobile web, web, Android app, and iOS app. The results appear in Figure 13. Looking at the four charts, we observe a few interesting patterns. Variables such as beginner lifecycle state and stable engagement are most often significant across all channels. Further, the country demographic variable is only a significant user HTE dimension for mobile web. Additionally, transaction variables also vary in terms of the proportion of experiments in which they are a significant source of user HTE. Collectively, in regards to RQ3, these results lend credence to the notion that the five categories of user constructs are viable dimensions for measuring HTE in digital experiments and, importantly, that user HTEs are prevalent in such online controlled experiments. In fact, across our observed experiments, about 14% of total user variable-experiment HTE tuples were significant. We have no reason to believe that this degree of user

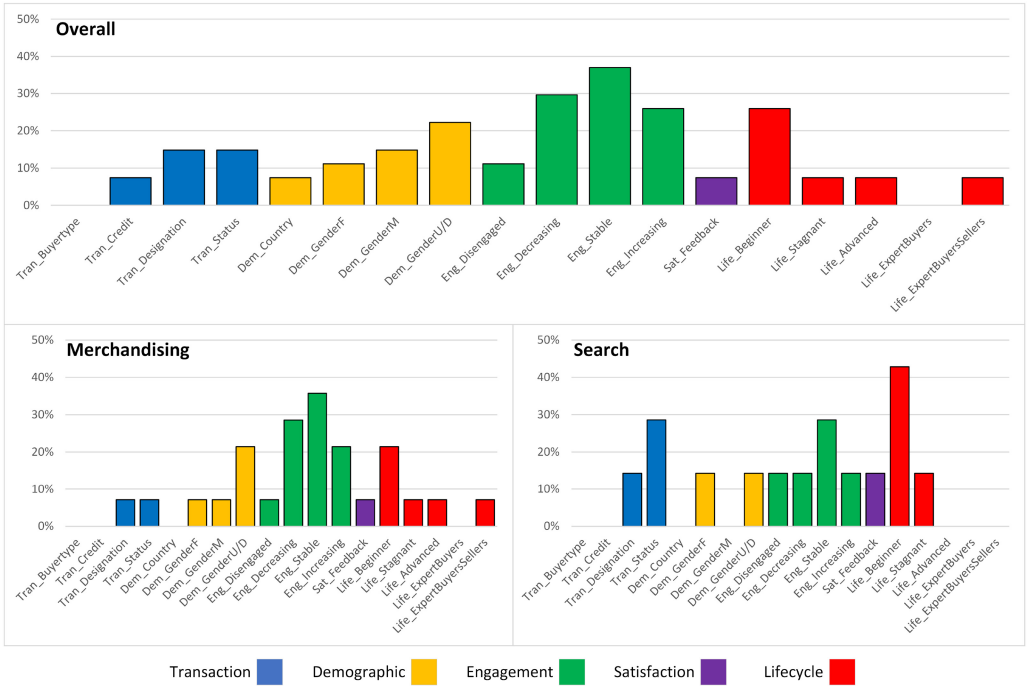


Fig. 12. Impact of user characteristics on HTE: Overall and for key session stages.

heterogeneity is unique to our data and, rather, believe that it is quite common and potentially highly problematic if ignored.

We had previously used a stylized example of two experiments where the user HTEs differed from the overall ATE (Figure 1). To better understand the implications of significant user HTEs for decision-making, we present similar plots from four actual experiments—as done in prior studies [67, 80]. Figure 14 shows results from four experiments related to the following respective session stages: advertising (top left chart), search (top right chart), and merchandising (two bottom charts). In each chart, the vertical solid line indicates the ATE, the vertical dashed lines depict the 95% confidence interval for the ATE. The x-axis shows the relative session-level conversion rate delta/improvement for the treatment versus the control setting. For example, 0.01 indicates that the treatment setting had a conversion rate that was 1% higher than the status-quo control group. The horizontal lines depict the confidence intervals for each user variable labeled on the y-axis. We only included select, statistically significant user HTEs (i.e., ones where the entire line is on a single side, either positive or negative). The line colors match the user characteristic category colors presented in earlier figures in the article (e.g., Figure 13).

All four experiments in Figure 14 have positive ATEs meaning that the treatment groups in these experiments had higher conversion rates than the control groups. The ATE is most pronounced on the advertising experiment, with a value around 1%. To put this effect size into perspective, most e-commerce platforms have session level conversion rates between 5% and 10%. The user HTE results also show several important patterns. First, we see that in each of the charts, there are many user characteristics with significant HTEs—and that often they are significant in the opposite direction from the ATE. In the aforementioned advertising experiment, while increasing engagement (second green line) is close to the ATE, decreasing and stable engagement levels have a strong negative treatment effect (the first and third green lines). Similarly, the expert buyers and

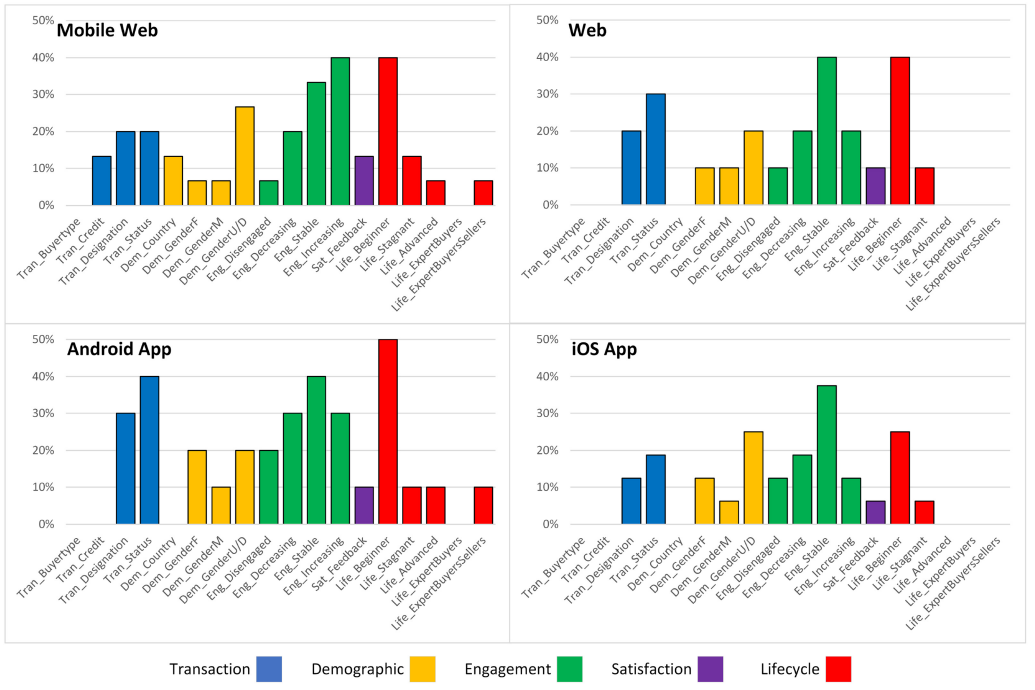


Fig. 13. Impact of user characteristics on HTE by channels/devices.

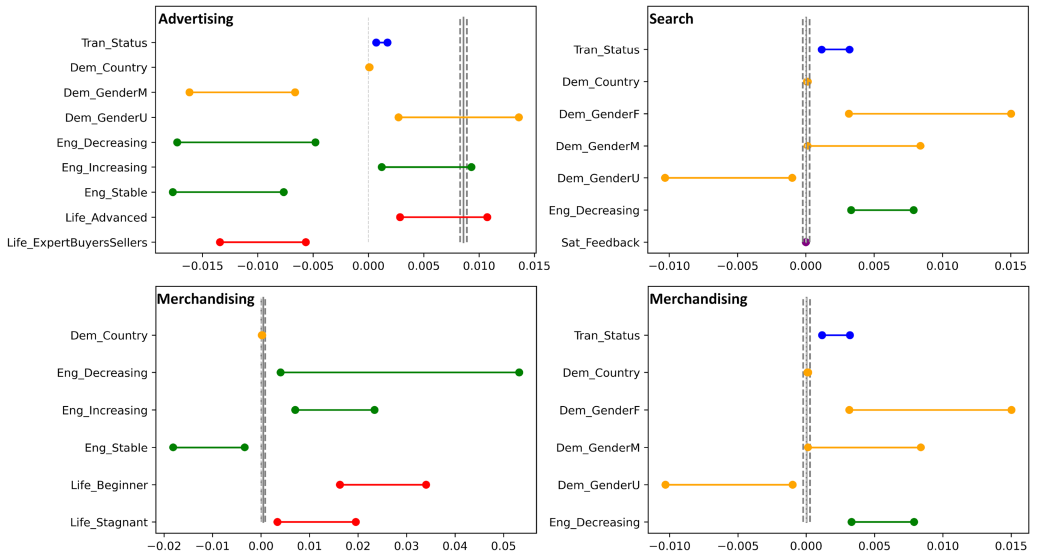


Fig. 14. Examples or user HTE in experiments with significant and insignificant ATEs.

sellers also have a strong negative reaction to the treatment. For these sub-groups, the treatment decreases conversion rates by 1% or more. Second, we observe differing HTE trends within the same category of user characteristics such as engagement or lifecycle states. Third, in addition to identifying complementary and opposing HTEs relative to the ATE, the size of the intervals can

indicate variance or stability of the effect for sub-groups. As noted in Reference [67], because the user HTEs provide insights into the underlying "composition" of the ATE, decision-makers or researchers might wish to avoid acting on small positive ATEs driven by a long tail of positive user HTEs with wide confidence intervals, thus illustrating estimates that are not accurately estimated (e.g., Reference [40]). Overall, as Figure 14 illustrates, user HTE analysis provides richer insights and a value-add when conducting digital experiments that can be expensive and time-consuming. Our proposed user HTE framework, which combines theory-guided and literature-based user characteristics and the HTE-DML method, offers an accurate mechanism for performing such analysis in large-scale online controlled experiments.

7 CONCLUSION

Digital experiments have become increasingly pervasive for assessing the value of a new digital treatment relative to the status quo in various stages of the user session journey including search, recommendation, layout design or merchandising, advertising, and checkout/conversion. However, this growth has prompted two challenges for digital experimentation platforms tasked with overseeing hundreds or thousands of concurrent experiments on a weekly basis. First, the average effect sizes for treatment versus control settings (i.e., the ATE) have decreased as the volume of experiments has increased and low-hanging fruit have been identified (i.e., the experiment sensitivity challenge) such that the incremental value of the insights and actions resulting from subsequent experiments has decreased. Second, experiments often focus on the ATE, even when spanning 50–100+ million user sessions, without considering the varying effects on major user sub-groups, that is, the heterogeneous treatment effect or user HTE.

To tackle these challenges, we propose a novel framework for examining user HTEs in digital experiments. Our framework combines an array of user characteristics with the HTE-DML method. Experiments on a series of simulated datasets and on 27 real-world experiments spanning 1.76 billion sessions demonstrate the effectiveness of HTE-DML relative to existing HTE detection methods. We also find that in digital experimentation contexts, transaction, demographic, engagement, satisfaction, and lifecycle characteristics are all predictive of session conversion outcomes, with most variables being statistically significant in 40% to 60% of experiments. More importantly, these same variables also have statistically significant heterogeneous treatment effects in 10% to 20% of cases. We use examples to illustrate how treatment effects might differ for various user sub-groups, and why being aware of such differences might offer valuable insights. In the same vein as prior studies, our main contributions include the framework/method [5], empirical insights yielded [6, 38, 71], and extensive evaluation of existing HTE methods.

The results of our work have important implications for research and practice involving digital experimentation. Many digital experience and/or information retrieval-related researchers proposing new search engine algorithms, novel recommender system models, enhanced layout or user interface elements, use online controlled experiments or "A/B tests" to test and validate their results. These studies generally report only the ATEs. Examining and reporting user HTE results could offer more granular and actionable insights into how such innovations are perceived or by different subgroups from a larger sample or population. In contexts involving ML-based treatments, user HTE can also shed light on the fairness of digital treatments/policies [32, 49]. Our framework provides concrete guidelines and methods that researchers can employ. With respect to practitioners, our work sheds light on the art of the possible, practical, and valuable regarding user HTEs in large-scale digital experiments. In particular, we show that in real-world digital experiments, user HTEs can be accurately detected, are reasonably prevalent in this context, and may inform decision-making by shedding light on heterogeneous effects that are similar or different from the average effect, and the extent of variance and uncertainty surrounding these effects.

These findings are important for various stakeholder groups, including product managers, analysts, and data scientists engaged in digital design and data-driven decision-making. Our work is not without its limitations. For instance, our user HTE framework requires user characteristics as input (e.g., the transaction, demographic, engagement, satisfaction, and lifecycle variables in this study). In primary data research environments, researchers would need to collect such data from experiment participants as part of the research design—though this is becoming commonplace [71]. In organizational settings, firms may not always have collected user characteristics [33], however with the impetus on data monetization and advanced insights, this, too, is becoming more common [67]. Additionally, controlling for user test planes requires knowledge of co-occurring session-level treatments. Future work may consider proposing new algorithms and/or models for user HTE detection. Nevertheless, we believe our study constitutes an important step toward the call for more research exploring user heterogeneity in online controlled experiments [33, 45].

REFERENCES

[1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.* 26, 3 (2008), 1–34.

[2] Ahmed Abbasi, Raymond Y. K. Lau, and Donald E. Brown. 2015. Predicting behavior. *IEEE Intell. Syst.* 30, 3 (2015), 35–43.

[3] Idris Adjerid and Ken Kelley. 2018. Big data in psychology: A framework for research advancement. *Am. Psychol.* 73, 4 (2018), 899–917.

[4] Faizan Ahmad, Ahmed Abbasi, Brent Kitchens, Donald A. Adjeroh, and Daniel Zeng. 2020. Deep learning for adverse event detection from web search. *IEEE Trans. Knowl. Data Eng.* 34, 6 (2020), 2681–2695.

[5] Faizan Ahmad, Ahmed Abbasi, Jingjing Li, David G. Dobolyi, Richard G. Netemeyer, Gari D. Clifford, and Hsinchun Chen. 2020. A deep learning architecture for psychometric natural language processing. *ACM Trans. Inf. Syst.* 38, 1 (2020), 1–29.

[6] Jaime Arguello and Bogeum Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Trans. Inf. Syst.* 37, 3 (2019), 1–34.

[7] Barry C. Arnold. 2015. *Pareto Distribution*. John Wiley & Sons, Ltd, 1–10.

[8] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A.* 113, 27 (2016), 7353–7360.

[9] Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *Ann. Stat.* 47, 2 (2019), 1148–1178. https://doi.org/10.1214/18-AOS1709

[10] Xiao Bai, Ioannis Arapakis, B. Barla Cambazoglu, and Ana Freire. 2017. Understanding and leveraging the impact of response latency on user behaviour in web search. *ACM Trans. Inf. Syst.* 36, 2 (2017), 1–42.

[11] Michel Ballings and Dirk Van den Poel. 2012. Customer event history for churn prediction: How long is long enough? *Expert Syst. Appl.* 39, 18 (2012), 13517–13522.

[12] Yihan Bao, Shichao Han, and Yong Wang. 2021. Treatment effect detection with controlled FDR under dependence for large-scale experiments. arXiv:2110.07279. Retrieved from https://arxiv.org/abs/2110.07279.

[13] Rina Foygel Barber and Emmanuel J. Candès. 2015. Controlling the false discovery rate via knockoffs. *Ann. Stat.* 43, 5 (2015), 2055–2085.

[14] Tamás Bartus. 2005. Estimation of marginal effects using margeff. *Stata J.* 5, 3 (2005), 309–329. https://doi.org/10.1177/1536867X0500500303

[15] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[16] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 4 (2001), 1165–1188.

[17] J. Martin Bland and Douglas G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 327, 8476 (1986), 307–310.

[18] Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.

[19] Donald E. Brown, Ahmed Abbasi, and Raymond Y. K. Lau. 2015. Predictive analytics: Predictive modeling at the micro level. *IEEE Intell. Syst.* 30, 3 (2015), 6–8.

[20] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *Am. Econ. Rev.* 107, 5 (2017), 261–65.

[21] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.
[22] Scott Cunningham. 2021. *Causal Inference: The mixtape*. Yale University Press.
[23] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. 123–132.
[24] Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, and Yang Yang. 2017. User modeling on demographic attributes in big mobile social networks. *ACM Trans. Inf. Syst.* 35, 4 (2017), 1–33.
[25] Boyan Duan, Larry Wasserman, and Aaditya Ramdas. 2021. Interactive identification of individuals with positive treatment effect while controlling false discoveries. arXiv:2102.10778. Retrieved from https://arxiv.org/abs/2102.10778.
[26] William Fithian and Lihua Lei. 2022. Conditional calibration for false discovery rate control under dependence. *Ann. Stat.* 50, 6 (2022), 3091–3118.
[27] Tianjun Fu, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2012. Sentimental spidering: Leveraging opinion information in focused crawlers. *ACM Trans. Inf. Syst.* 30, 4 (2012), 1–30.
[28] Shen Gao, Xiuying Chen, Li Liu, Dongyan Zhao, and Rui Yan. 2021. Learning to respond with your favorite stickers: A framework of unifying multi-modality and user preference in multi-turn dialog. *ACM Trans. Inf. Syst.* 39, 2 (2021), 1–32.
[29] Juan Carlos Gázquez-Abad, Marie Hélène De Canniére, and Francisco J. Martínez-López. 2011. Dynamics of customer response to promotional and relational direct mailings from an apparel retailer: The moderating role of relationship strength. *J. Retail.* 87, 2 (2011), 166–181.
[30] Priyanga Gunarathne, Huaxia Rui, and Abraham Seidmann. 2017. Whose and what social media complaints have happier resolutions? Evidence from Twitter. *J. Manage. Inf. Syst.* 34, 2 (2017), 314–340.
[31] Asela Gunawardana and Guy Shani. 2015. Evaluating recommender systems. In *Recommender Systems Handbook*. Springer, 265–308.
[32] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1023.
[33] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explor. Newslett.* 21, 1 (2019), 20–35.
[34] Shuguang Han, Zhen Yue, and Daqing He. 2015. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Trans. Inf. Syst.* 33, 4 (2015), 1–34.
[35] Trevor Hastie and Robert Tibshirani. 1987. Non-parametric logistic and proportional odds regression. *J. Roy. Stat. Soc.: Ser. C (Appl. Stat.)* 36, 3 (1987), 260–276.
[36] Yuzi He, Christopher Tran, Julie Jiang, Keith Burghardt, Emilio Ferrara, Elena Zheleva, and Kristina Lerman. 2021. Heterogeneous effects of software patches in a multiplayer online battle arena game. In *Proceedings of the 16th International Conference on the Foundations of Digital Games (FDG'21) 2021*. 1–9.
[37] Lorin M. Hitt and Frances X. Frei. 2002. Do better customers utilize electronic distribution channels? The case of PC banking. *Manage. Sci.* 48, 6 (2002), 732–748.
[38] Melody Y. Ivory and Rodrick Megraw. 2005. Evolution of web site design patterns. *ACM Trans. Inf. Syst.* 23, 4 (2005), 463–497.
[39] Avinash Kaushik. 2009. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*. John Wiley & Sons.
[40] Ken Kelley and Scott E. Maxwell. 2003. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychol. Methods* 8, 3 (2003), 305–321.
[41] Ken Kelley and Kristopher J. Preacher. 2012. On effect size. *Psychol. Methods* 17, 2 (2012), 137–152.
[42] Brent Kitchens, David Dobolyi, Jingjing Li, and Ahmed Abbasi. 2018. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *J. Manage. Inf. Syst.* 35, 2 (2018), 540–574.
[43] Ron Kohavi. 2015. Online controlled experiments: Lessons from running a/b/n tests for 12 years. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1–1.
[44] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1168–1176.
[45] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to a/b Testing*. Cambridge University Press.

[46] Ron Kohavi, Diane Tang, Ya Xu, Lars G. Hemkens, and John P. A. Ioannidis. 2020. Online randomized controlled experiments at scale: Lessons and extensions to medicine. *Trials* 21, 1 (2020), 1–9.
[47] Ron Kohavi and Stefan Thomke. 2017. The surprising power of online experiments. *Harv. Bus. Rev.* 95, 5 (2017), 74–82.
[48] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U.S.A>* 116, 10 (2019), 4156–4165.
[49] John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3598–3609.
[50] Jingjing Li, Ahmed Abbasi, Amar Cheema, and Linda B. Abraham. 2020. Path to purpose? How online customer journeys differ for hedonic versus utilitarian purchases. *J. Market.* 84, 4 (2020), 127–146.
[51] Jingjing Li, Kai Larsen, and Ahmed Abbasi. 2020. TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning. *MIS Quart.* 44, 4 (2020).
[52] Shangsong Liang, Yupeng Luo, and Zaiqiao Meng. 2021. Profiling users for question answering communities via flow-based constrained co-embedding model. *ACM Trans. Inf. Syst.* 40, 2 (2021), 1–38.
[53] Eveliina Lindgren and Jürgen Münch. 2016. Raising the odds of success: The current state of experimentation in product development. *Inf. Softw. Technol.* 77 (2016), 80–91.
[54] Vera L. Miguéis, Dirk Van den Poel, Ana S. Camanho, and João Falcão e Cunha. 2012. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Syst. Appl.* 39, 12 (2012), 11250–11256.
[55] Vikas Mittal and Wagner A. Kamakura. 2001. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *J. Market. Res.* 38, 1 (2001), 131–142.
[56] Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty. 2004. Modeling online browsing and path analysis using clickstream data. *Market. Sci.* 23, 4 (2004), 579–595.
[57] Cataldo Musto, Fedelucio Narducci, Marco Polignano, Marco De Gemmis, Pasquale Lops, and Giovanni Semeraro. 2021. MyrrorBot: A digital assistant based on holistic user models for personalized access to online services. *ACM Trans. Inf. Syst.* 39, 4 (2021), 1–34.
[58] X. Nie and S. Wager. 2020. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (09 2020), 299–319. https://doi.org/10.1093/biomet/asaa076
[59] Judea Pearl. 2009. Causal inference in statistics: An overview. *Stat. Surv.* 3 (2009), 96–146.
[60] Werner J. Reinartz and Vita Kumar. 2003. The impact of customer relationship characteristics on profitable lifetime duration. *J. Market.* 67, 1 (2003), 77–99.
[61] Joseph Rigdon, Michael Baiocchi, and Sanjay Basu. 2018. Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials* 19, 1 (2018), 1–15.
[62] Peter M. Robinson. 1988. Root-N-consistent semiparametric regression. *Econometrica* 56, 4 (1988), 931–954.
[63] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance assessments for web search evaluation: Should we randomise or prioritise the pooled documents? *ACM Trans. Inf. Syst.* 40, 4 (2022), 1–35.
[64] David C. Schmittlein, Donald G. Morrison, and Richard Colombo. 1987. Counting your customers: Who-are they and what will they do next? *Manage. Sci.* 33, 1 (1987), 1–24.
[65] Korbinian Strimmer. 2008. A unified approach to false discovery rate estimation. *BMC Bioinf.* 9, 1 (2008), 1–14.
[66] Vasilis Syrgkanis, Greg Lewis, Miruna Oprescu, Maggie Hei, Keith Battocchi, Eleanor Dillon, Jing Pan, Yifeng Wu, Paul Lo, Huigang Chen, et al. 2021. Causal inference and machine learning in practice with econml and causalml: Industrial use cases at microsoft, tripadvisor, uber. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4072–4073.
[67] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. 2016. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econ. Stat.* 34, 4 (2016), 661–672.
[68] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 17–26.
[69] Yuan Tian, Ke Zhou, and Dan Pelleg. 2021. What and how long: Prediction of mobile app engagement. *ACM Trans. Inf. Syst.* 40, 1 (2021), 1–38.
[70] Christopher Tran and Elena Zheleva. 2019. Learning triggers for heterogeneous treatment effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5183–5190.
[71] Kelsey Urgo and Jaime Arguello. 2022. Understanding the "Pathway" towards a searcher's learning objective. *ACM Trans. Inf. Syst.* 40, 4 (2022), 1–42.
[72] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 11 (2008).
[73] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113, 523 (2018), 1228–1242.

[74] Chao Wang, Hengshu Zhu, Peng Wang, Chen Zhu, Xi Zhang, Enhong Chen, and Hui Xiong. 2021. Personalized and explainable employee training course recommendations: A bayesian variational approach. *ACM Trans. Inf. Syst.* 40, 4 (2021), 1–32.

[75] Hongwei Wang and Jure Leskovec. 2021. Combining graph convolutional neural networks and label propagation. *ACM Trans. Inf. Syst.* 40, 4 (2021), 1–27.

[76] Hao Wang, Defu Lian, Hanghang Tong, Qi Liu, Zhenya Huang, and Enhong Chen. 2021. HyperSoRec: Exploiting hyperbolic user and item representations with multiple aspects for social-aware recommendation. *ACM Trans. Inf. Syst.* 40, 2 (2021), 1–28.

[77] Lili Wang, Chenghan Huang, Ying Lu, Weicheng Ma, Ruibo Liu, and Soroush Vosoughi. 2021. Dynamic structural role node embedding for user modeling in evolving networks. *ACM Trans. Inf. Syst.* 40, 3 (2021), 1–21.

[78] Wei Wang, Jiaying Liu, Tao Tang, Suppawong Tuarob, Feng Xia, Zhiguo Gong, and Irwin King. 2020. Attributed collaboration network embedding for academic relationship mining. *ACM Trans. Web* 15, 1 (2020), 1–20.

[79] Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 645–654.

[80] Yuxiang Xie, Nanyu Chen, and Xiaolin Shi. 2018. False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 876–885.

[81] Tao Xiong, Yong Wang, and Senlie Zheng. 2020. *Orthogonal Traffic Assignment in Online Overlapping A/B Tests.* Technical Report. Tencent EasyChair Whitepaper.

[82] Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2021. Clarifying ambiguous keywords with personal word embeddings for personalized search. *ACM Trans. Inf. Syst.* 40, 3 (2021), 1–29.

[83] Peng Zhang, Baoxi Liu, Tun Lu, Xianghua Ding, Hansu Gu, and Ning Gu. 2022. Jointly predicting future content in multiple social media sites based on multi-task learning. *ACM Trans. Inf. Syst.* 40, 4 (2022), 1–28.

Received 8 April 2022; revised 20 October 2022; accepted 17 December 2022