Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes

Alex Deng
Airbnb
Seattle, WA, USA
alex.deng@airbnb.com

Michelle Du
Airbnb
San Francisco, CA, USA
michelle.du@airbnb.com

Anna Matlin
Airbnb
San Francisco, CA, USA
anna.matlin@airbnb.com

Qing Zhang
Airbnb
San Francisco, CA, USA
qing.zhang@airbnb.com

ABSTRACT

Improving statistical power is a common challenge for online experimentation platforms so that more hypotheses can be tested and lower effect sizes can be detected. To increase the power without increasing the sample size, it is necessary to consider the variance of experimental outcome metrics. Variance reduction was previously applied to online experimentation based on the idea of using pre-experiment covariate data to account for noise in the final metrics. Since this method relies on correlations between pre-experiment covariates and experiment outcomes, its effectiveness can be limited when testing features for specific product surfaces. We were also motivated by the challenge of attributing sparse, delayed binary outcomes to individual user-product interactions. We present two novel methods for variance reduction that rely exclusively on *in-experiment* data. The first method is a framework for a model-based leading indicator metric which continually estimates progress toward a delayed binary outcome. The second method is a counterfactual treatment exposure index that quantifies the amount that a user is impacted by the treatment. We applied these methods to past experiments and found that both can achieve variance reduction of 50% or more compared to the delayed outcome metric. The substantial reduction in variance afforded by the two methods presented in this paper has enabled Airbnb's experimentation platform to become more agile and innovative.

CCS CONCEPTS

• **Mathematics of computing → Probabilistic inference problems**; • **Applied computing → E-commerce infrastructure**.

KEYWORDS

A/B testing, experimentation, online evaluation, variance reduction, causal surrogate, counterfactual, recommender system

ACM Reference Format:
Alex Deng, Michelle Du, Anna Matlin, and Qing Zhang. 2023. Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3580305.3599928

1 INTRODUCTION

Online experimentation has been widely adopted within the technology industry for more than a decade [21, 28, 29]. Applications include web search [15, 27, 39], social networks [12, 46, 47], video streaming platforms [7, 18, 45], e-commerce platforms [48] and marketplaces [23, 41]. An online experiment in the simplest one-treatment form is typically referred to as an *A/B test.* In this setting, online traffic is randomly sampled to create two groups of subjects, a control group and a treatment group. The control group receives the existing version of the product or service, while the treatment group receives an altered version, which can also be called an *intervention.* Randomization is the most straightforward and reliable identification strategy to establish a causal link [22, 35] between the treatment intervention and observed outcomes because confounding variables cannot distort results for either group. A systematic difference between the control and treatment group outcomes must be *caused* by the intervention.

With a causal identification strategy in place, the next step is to investigate whether a difference between the two groups exists for outcome metrics of interest. There is a vast literature around this two-sample problem, encompassing frequentist null hypothesis testing, Bayesian methods, sequential testing, and heterogeneous treatment effect estimation (for a recent survey, see [29]). We observe that all of these statistical methods depend on the notion of variance.

High variance poses a serious challenge for online experimentation platforms. For a given level of statistical power, the variance is proportional to the sample size required (more detail in Section 2). Many innovations cannot be tested concurrently, or *parallelized,* without interaction. For example, a user can only experience one user interface and one backend recommendation engine at a time. As a result, running online experiments with high-variance target metrics requires more resources, either in the form of time or user traffic allocation. Simply extending the experiment duration or increasing the traffic introduces an opportunity cost. Yet, maintaining high statistical power is critical to ensure trustworthiness of both the launch decision and effect estimates [26], as low power can lead to both sign error and exaggeration of treatment effect (TypeS and M errors) [19, 33]. Experimentation platforms can often mitigate this problem with commonly used variance reduction techniques such as CUPED [10], constructing estimators for the metric of interest which have lower variance than the original metric.

In addition to the problem of high variance, three notable challenges motivated our work in this paper: delayed outcomes, sparse signals, and effect dilution. The problem of delayed outcomes is well-studied in the context of experiments with long-term outcomes that are not observable in a short experiment period (e.g. effect of

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599928

job training on employment) [2]. In the online experimentation setting, some outcomes are considered to be delayed even if they can occur during the experiment period because they are compared to other metric outcomes that occur earlier. For example, at Airbnb, a booking is considered a delayed outcome because it occurs after a journey of engagements such as searches and listing-views. Sparse signals, meanwhile, are an obstacle for understanding user behavior. Booking patterns on Airbnb illustrate this problem well: because few users reach the final stage of booking, it is unclear which interactions motivated them to book because many users with similar search behavior do not complete a booking. Lastly, in the context of search ranking experiments, we observed that a very small proportion of users exposed to the treatment actually respond to the change, indicating effect dilution. However, it is not possible to guess which users will respond based on pre-experiment covariates.

Alone, any of these challenges can slow down otherwise agile experimentation platforms. When they are combined in one setting, they can become a bottleneck for innovation. We propose two new directions for variance reduction, both of which achieve variance reduction of more than 50% compared to the bookings metric in past experiments. It is worth noting that the first of these approaches also enabled continuous attribution of delayed outcomes to prior user-product interactions.

We make two novel contributions to the online experimentation and measurement science literature:

(1) *Model-based leading indicator with continuous attribution*: We propose a framework for a model-based value function, inspired by reinforcement learning, to continuously track progress towards a delayed outcome and incrementally attribute the final binary reward to individual user-product interactions. This approach yields two notable results. First, it achieves variance reduction of 50% to 85% compared to the baseline bookings metric while tracking the experiment effect sizes with high fidelity. Second, it allows us to create a new metric representing the utility gained towards a purchasing decision based on fine-grain engagements such as listing views. The new *utility metric* can be flexibly aggregated to various levels of granularity, introducing a range of analysis units and opportunities for user understanding.

(2) *In-experiment counterfactual treatment exposure index*: We propose a continuous index that quantifies the amount that a subject is impacted by the treatment based on data from the experiment period to tackle the dilution problem. After subjects are ordered based on their position in the index, it is possible to estimate the treatment effect for the top-k percentile of users. This approach is connected to the the idea of experiment triggering because it excludes subjects with low treatment effect, but it is different because the restriction is based on in-experiment signals. We report 60% or more variance reduction from this novel technique when focusing on subjects with high values in the exposure index. We also show that when in-experiment signals are used for filtering, the standard sample variance formula can underestimate the variance, and appropriate corrections are necessary.

2 BACKGROUND

We return to the A/B test from Section 1. When an experiment is completed, two metric outcomes, M_T and M_C, are computed for the two groups, along with the difference $\Delta(M)$:

$$\Delta(M) = M_T - M_C$$

Thanks to the central limit theorem [4], we know $\Delta(M)$ can be approximated by a Normal distribution $Normal(\delta,\sigma^2)$, with δ being the true underlying treatment effect and our interest of inference, and the variance σ^2 can be estimated via:

$$\mathrm{Var}(\Delta(M)) = \mathrm{Var}(M_T) + \mathrm{Var}(M_C)\ .$$

The frequentist 95% confidence interval of δ is

$$\Delta(M) \pm 1.96\sqrt{\mathrm{Var}(\Delta(M))}$$

and the well-known t-statistic is defined as the ratio of the estimator of δ and its standard deviation $\frac{\Delta(M)}{\sqrt{\mathrm{Var}(\Delta(M))}}$. The variance term $\mathrm{Var}(\Delta(M))$ plays a central role in the inference problem, determining the width of the confidence interval and the statistical power of the test. A well-known rule of thumb [42] is that to achieve statistical power of 80%, the total number of samples required can be calculated from sample variance σ^2 and the desired minimum detectable effect size δ:

$$n \approx \frac{16\sigma^2}{\delta^2}\ .$$

Therefore, a reduction in variance σ^2 translates directly into a reduction in the required sample size n which constrains the agility of online experimentation platforms (see 3.2 for more detail about high variance metrics at Airbnb).

A widely adopted method for variance reduction in online experimentation is to leverage pre-experiment covariates. CUPED [10] is a simple, semi-parametric efficiency augmentation method [40] that modifies the original estimator $\Delta(M)$ by augmenting it with another term based on pre-experiment covariates that are correlated with in-experiment outcome metric M but orthogonal to the treatment. That is, define a new family of estimators $\Delta^*(M,\theta)$ as:

$$\Delta^*(M,\theta) = \Delta(M) - \theta\Delta(X)\ ,$$

where $\mathrm{E}(\Delta(X)) = 0$ because treatment cannot impact pre-experiment observations. For any fixed θ, $\Delta^*(M,\theta)$ is still an unbiased estimator for δ just as $\Delta(M)$. Moreover, we can use a θ that minimizes the variance of $\Delta^*(M,\theta)$. [10] showed that the optimal θ resembles a regression coefficient. Despite the resemblance, this efficiency augmentation method does not rely on a *linear model assumption* and can utilize *nonlinear regression* with modern machine learning algorithms, or even in non-randomized settings [11, 20, 32, 36].

The effectiveness of CUPED relies heavily on the correlation between pre-experiment baseline covariates and the in-experiment outcome metric. In a blog post from Microsoft Experimentation Platform [6], it is reported that for one product surface, CUPED substantially reduces the variance by 20% or more for the majority of metrics. However, on another product surface, the majority of metrics see a reduction of less than 5% because pre-experiment covariates are weakly correlated with the outcome metric. Here we tackle a similar situation that CUPED using pre-experiment signals is ineffective in variance reduction over the baseline standard diff-in-means metric.

3 BUSINESS MOTIVATION

We provide an overview of the Airbnb booking flow, which motivated the approaches proposed in Sections 4 and 5. We also review various experimentation challenges introduced by business patterns.

3.1 Airbnb Booking Stages

A typical Airbnb booking journey is visualized in Figure 1.

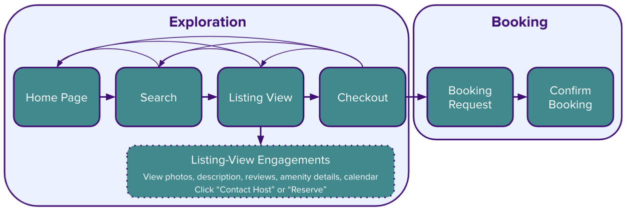


Figure 1: Airbnb booking flow.

Homepage When users arrive at airbnb.com or open the Airbnb App, they are greeted by a homepage experience that is called location-less search. Because users have yet to specify search criteria such as the travel destination, listings are surfaced from a variety of locations. Users looking for inspiration can spend hours clicking around and use the newly introduced *categories* to explore different types of unique Airbnb stays.

Search The majority (about 90%) of visitors will have a specific location in mind and use the search box, which prompts the user to enter key criteria for their upcoming trip: location, calendar dates, and guest count. Location and dates are the most important elements of the query affecting listing retrieval and ranking. Travelers flexible about trip timing can issue a search without specifying calendar dates. The number of guests further excludes listings based on guest capacity. The search results page includes a ranked feed where each listing option is presented as a card containing photos, title and price information. There is also a map view which displays the price at the location of the listing. Most of the time, more than one page of listing cards is returned, and users can continue to review listings from the same search request. Alternatively, users can continue to refine their search by updating their criteria in the search box or by interacting with the map. An average user can issue 10 to 20 requests from the search box, and often even more requests from the map, during a 2 to 3 weeks window of visits.

Listing View Viewing one listing in detail enables users to gain information about their potential stay. The user can reach a listing view from the homepage, search results, saved wishlist, or a direct link (shared via message, 3rd party apps, web search, marketing, etc.)[1] During a listing view, the user can read the full description, house rules, amenities offered, reviews by previous guests, and cancellation policy details. Users can also browse photos, add or modify trip dates, and review payment details such as the nightly price, fees, and tax. They can also reach out to the host directly using the "Contact Host" button.

Booking and Post Booking When a guest decides to book a listing, they click the "Reserve" button on the listing view to go to the checkout stage to complete payment. If the listing allows "Instant Booking", the reservation can be completed immediately. Other listings require the guest to write a message, which the host can review and approve. After the guest makes the booking request, it is typically approved instantly or shortly thereafter. Guests can cancel and re-book, though refunds depend on the listing's cancellation policy.

The process of booking on Airbnb ranges from a few hours to several weeks, with the majority of the planning taking place over a few days, often with gaps. For example, a guest may visit on Tuesday and then return during the weekend to complete the booking. Though the conversion rate from the search to the listing view is relatively high, the conversion rate from the listing view to booking is much lower. The listing view step is central to the booking process because it is immediately upstream of the checkout page. Users with different levels of booking intention can have very different engagement patterns. For instance, most users will look at photos. But reading the full list of amenities, clicking into review details, or checking availability for different calendar dates reveals a stronger booking intention. Finally, clicking on the Reserve button is a signal that the user is interested in continuing the booking process.

3.2 High Variance and Other Challenges

On Airbnb's marketplace, a booking represents a match between a guest and a listing. Helping our guests find high-quality listings is a continuing effort that is measured with a set of key metrics. The *bookings per user* metric is often the target metric or one of the guardrail metrics in an experiment. However, compared to many other user engagement metrics such as listing-views and searches, the bookings metric lacks sensitivity — it requires much more traffic to detect the same effect size. In Table 1, we compare the variances of the bookings, searches, and listing-views metrics, using the variance of bookings as a baseline. Remembering that the sample size required is proportional to the sample variance, the normalized values represent a multiplier for the traffic required to detect the same effect size, compared to bookings.

	Bookings	Bookers	Bookers(CUPED)	Searches	Listing views
Variance of percent lift	1	0.88	0.84	0.43	0.35

Table 1: Comparison of variances from a search ranking experiment, using the bookings metric as the baseline.

For highly skewed metrics, capping [25] can enable further variance reduction without sacrificing the strength of the treatment effect. Because we focus on experiments that are randomized by guest, the *bookers* metric is the bookings metric capped at 1. Over a corpus of past experiments, we found that the bookers metric reduced variance by 10-15%.

Applying CUPED with a linear formulation to the bookers metric only negligibly reduced variance (about 5%) in search and recommendation experiments at Airbnb. We also applied CUPED in the general form, using a boosting model with a large number of covariates as augmentation. Even in this case, we were only able to achieve ~ 15% variance reduction. There are many business reasons why pre-experiment data may not be helpful for variance reduction. For example, a traveler's booking patterns may change with different occasions and purposes. As a growing business, Airbnb continues to attract new users, who do not have pre-experiment covariate data.

A common approach to identify surrogate metrics is to define a set of candidate metrics and to use a historical experiment corpus to evaluate metrics based on two criteria: sensitivity and alignment [9, 14]. Because the variance of listing-views is lower than that of bookings, and the listing-view stage directly precedes a booking request, it is intuitive to propose a surrogate metric based on listing-view engagement signals or simply the quantity of *dated listing-views* (listing-views with check-in and out dates specified). This is a trial-and-error approach whose success largely depends on the heuristics used to define the candidate metrics. Heuristic-driven metrics often cannot align with the target metric with high fidelity because the causal

[1] A rare usage pattern is to query the Airbnb search box with the listing name and go to the listing page directly.

mechanism can differ by treatment. For example, an increase in dated listing-views is generally considered to be a positive outcome, and often aligns with an increase in the bookings metric. However, there are plenty of experiments in which bookings improved and listing views decreased. We can also argue that reducing listing views while improving booking conversion is a good outcome because users reach the booking stage with less effort. Empirically picking surrogate metrics from existing metrics does not lead to improved understanding of user behavior that can answer deeper questions. In particular, we are interested to identify which moments in the process lead a user to commit to a specific listing and complete a booking.

Below, we explain how additional challenges related to Airbnb's business model motivated new directions for variance reduction.

Delayed outcomes For many travelers on Airbnb, the decision to book a listing can be the culmination of a lengthy search experience. Most trips are large purchase decisions, typically several hundred dollars or more. Achieving confidence in the purchase requires time and effort. Among guests who eventually booked, it is common that they considered more than one listing, yet the motivation for their final selection is not always clear. It is not straightforward to attribute their booking decision to an interaction in the product flow. This is in stark contrast to metrics like listing-views and click-through-rates, where users' preferences and feedback are revealed almost instantly. High cost purchases are intrinsically noisy, as there are various exogenous factors potentially affecting *whether* the guest will book, *when* the guest will make the decision, and *which* listing the guest will book. For example, guests often travel with family or a group of friends, and one guest's booking decision may be influenced by preferences or suggestions unrelated to the guest's search behavior. A guest that is comparing listings on Airbnb to options outside of the platform may suddenly abandon their booking journey. Guests also have varying degrees of urgency. Those who plan months ahead can gather more candidate listings and delay the booking to a later time. As the decision process gets longer, more exogenous noise is added to the booking outcome. Crucially, without clear logic for how to reward the final conversion to previous interactions that the user experienced in the booking process, the bookings metric does not have as much signal density as engagement metrics such as listing-views.

Sparse signals Another challenge is that booking is a sparse signal, for several reasons. Most notably, travel is an infrequent activity for most consumers. As a result, the Airbnb product flow results in a low visitor-to-booker conversion rate compared to the conversion rates of internet products such as search engines, social media, or video streaming platforms. Many users visit Airbnb simply to find inspiration for a future trip idea. Such users often do not plan to make a booking in coming days or weeks. When testing new search and recommendation features, which are high coverage because users can interact with them in early stages of the booking process, it is possible that the majority of user traffic has low intent to book. Though this traffic offers little signal, it introduces a large amount of noise for measuring booking conversion.

Effect dilution We observe that experiment exposure is not the same as impact. For high-exposure features such as search and recommendation, the treatment effect on low conversion rate event such as booking is often concentrated in a fraction of experiment subjects. In our applications, this fraction can be as low as 5 to 10%. If it were possible to guess which subjects would respond to the change using pre-experiment data, we could limit the analysis to those subjects for a more precise effect estimation. [8] pointed out that tight triggering is a special case of CUPED which can properly dilute the effect to estimate average treatment effect for the whole population, achieving variance reduction.

A naive approach to measure treatment exposure is to count the number of exposures and set a threshold. In the setting of search ranking, it is possible to count the number of searches, or searches in which two rankers return different results. However, in practice, the treatment can easily move the number of searches up and down, causing the number of users passing the threshold between treatment and control to be different. This result is common when the index is affected by the treatment and can yield misleading results. For example, if a treatment increases user engagement, more users would pass the threshold in the treatment group, but these users tend to have a lower conversion rate compared to those who have high purchasing intention and would been engaged regardless. The conversion rate when filtered to high-exposure users in this approach can show a negative conversion effect when the true effect is positive. Using in-experiment signal for segmenting and filtering can lead to sample ratio mismatch [17] and render the analyses invalid.

4 MODEL-BASED LEADING INDICATOR WITH CONTINUOUS ATTRIBUTION

We propose a framework backed by the theory of causal surrogacy [2] to construct a model-based value function [38] that continuously assesses the propensity of a delayed conversion event, and attributes the incremental gain of this value function at every step to the source of the action responsible for the improvement of the value function.

Let W_i be the binary treatment assignment of subject i, and Y_i be the delayed outcome (e.g. whether a series of visits and engagements lead to an uncancelled booking in the end). The causal surrogate model posits the existence of a set of observations S that can block (d-separate as in [35]) all the causal pathways from the treatment assignment W to the outcome Y. This leads to the surrogate assumption in the form of conditional independence in [2],

$$W_i \perp Y_i | S_i \, , \tag{1}$$

which entails that the same regression model $\mathrm{E}(Y|S) = \mathrm{E}(Y|S, W = 1) = \mathrm{E}(Y|S, W = 0)$ applies to both the treatment and control group. This also implies that all the causal effects on Y are mediated exclusively through causal effects on S, and pass forward to Y through the functional form of $\mathrm{E}(Y|S)$.

Let $V_i = \mathrm{E}(Y_i|S_i)$, it is straightforward to show that $\Delta(V) = \overline{V}_T - \overline{V}_C$ (average of V over the treatment and control group) is an unbiased estimator for the average treatment effect on Y. For readers familiar with causal graphical model, this surrogate assumption (1) is a special case of the front-door criterion [35] in the randomized treatment setting. We call V_i a surrogate or leading indicator for a delayed outcome Y_i. It is especially critical when Y is a long-term outcome *not observable* for most subjects during the experimentation period, for example, if Y represents a person's future income in 20 years when studying causal effect of education. However, in our application, the outcome is delayed but still mostly observed in the experimentation period. Unlike long-term effect literature, we employ a causal surrogate model for two purposes: variance reduction and continuous attribution.

4.1 Variance Reduction for Delayed Outcomes

The main mathematical property we exploit is the following corollary from the law of total variance:

$$\mathrm{Var}(V)=\mathrm{Var}\{E(Y|S)\}\leq\mathrm{Var}(Y)\ . \tag{2}$$

Because the regression model can smooth out a portion of Y's variance not explained by S, we expect significant variance reduction by replacing a highly noisy Y with a regression prediction of it. This also highlights the fundamental difference between our use of the causal surrogate model and a typical ML model for prediction.

(1) The objective of this model is to smooth out noise that is not part of the causal mechanism. We emphasize that optimizing for prediction accuracy is not the primary goal. In fact, when the outcome is delayed but still observed, one can include the outcome Y itself into the predictors S. This will trivially satisfy the surrogate assumption, but then V is equivalent to Y and no variance reduction is achieved.

(2) The ideal choice of S is the *smallest* set of predictors that satisfies (1). This is because $\mathrm{Var}\{\mathrm{E}(Y|S)\}\leq\mathrm{Var}\{\mathrm{E}(Y|S')\}$ when $S\subset S'$. The guiding principle of selecting such a set S is to picture the causal mechanism as a causal graph [35] and find a small set of nodes between the source of treatment intervention and the outcome.

We illustrate the selection process of surrogate predictors S in Figure 2. In this graph, W represents the treatment intervention and Y is the target outcome. C affects outcome Y but is not impacted by treatment intervention W. X and U represent possible predictors for the surrogate model. In this case, we exclude C from the set of possible predictors. Ignoring the dotted arrow $U\longrightarrow Y$ for a moment, there are two causal pathways from W to Y, both of which go through S. We see that S completely blocks all the causal pathways from W to Y, and S alone will satisfy the causal surrogate assumption (1). If the dotted arrow is real and there does exist a direct causal effect from U to Y that is not mediated through S, then adding U to the predictors is necessary. Otherwise, the partial effect $W\longrightarrow U\longrightarrow Y$ will not be captured by the surrogate model.

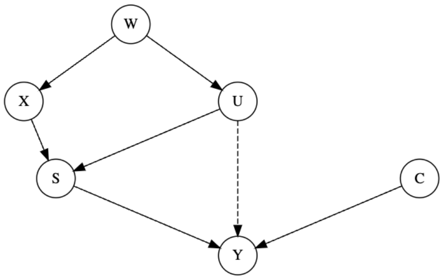


Figure 2: If the dotted line $U\rightarrow Y$ does not exist, S is sufficient and the optimal causal surrogate predictor. If $U\rightarrow Y$ exists, then U needs to be added to the predictors otherwise the effect of $W\rightarrow U\rightarrow Y$ will not be captured by the causal surrogate model.

For Airbnb search and recommendation experiments, treatments change the ranked listing feed, which then affects the set of listings that users click and view. Listing quality, affordability and other aspects of listings further affect users' level of engagement and booking intention. When a user is close to making a booking, they may examine the listing page more closely, going through past reviews and even contacting the host for specific questions. They then click the Reserve button to review payment details. Some will make a booking request, while others may choose to mark the listing as a booking candidate and continue to search for better options. Since the search step is upstream of listing views, all causal effects from the search page will be reflected by user activities during the listing view. Moreover, the checkout flow is downstream of the listing view, and the Reserve button on the listing view is the only entry-point to the checkout flow. This means that adding checkout page signals to our set of predictors does not help with the surrogate assumption. On the contrary, adding checkout page signals such as payment confirmation interactions undermines the variance reduction goal because the prediction problem becomes too simple. Based on these reasons, the main surrogate model features we use are various forms of user engagement and visits during the listing view step. We also included listing attributes such as location, nightly price, review ratings as well as total views, search result appearances and bookings of a listing for a past period (e.g. 90 days). The last category of features is trip attributes, which include the trip dates (if provided) and the number of guests. Note that even though treatments on ranking do not directly affect listing attributes, availability, or capacity, the user's decision to book is jointly affected by the information gathered from listing view and listing attributes. For example, for holidays and popular destinations inventories will run out and users have higher urgency to book. Table 2 lists main features from the three categories.

We construct the causal surrogate at each (user, listing) pair. For training, we collect logs from the preceding 14 days. For each (user, listing) pair, we label the pair using the observed booking outcome. We train a lightGBM model with dropout [24, 43] for binary classification followed by a logistic regression based calibration step [34]. The model is trained every week and used to score the following week. As mentioned above, prediction accuracy is not the primary goal and our main evaluation of our model-based surrogate is via empirical meta-analysis in Section 4.3.

Category	Top Features
Listing View Engagements	Number of Views from various platforms (web, mobile, app). Interaction with photos, descriptions, amenity details, guest reviews Reserve and Contact Host button clicks. Interaction with calendar availability
Listing Attributes	Past 90 days total listing views, search result appearances, booking requests. Location and nightly price. Cancellation policy. Instant bookable or must request to book.
Trip Attributes	Check-in and checkout dates. Lead time (how far from today to check-in). Number of guests.

Table 2: List of main features chosen for causal surrogate model of search ranking treatment on booking conversion. Listing view engagement features are aggregated and updated for each (user, listing) pair when users revisit a previously viewed listing or visit a new listing.

4.2 Continuous Attribution of Delayed Outcomes to Individual User-Product Interactions

Our second goal is to attribute a sparse and delayed outcome to fine-grained individual user-product interactions. We assume that we have identified the surrogate predictors S_t at any time step t for every user, and computed its causal surrogate $V_t=V(S_t)$. We omit user index i in notation for simplicity. Here the "time" t represents an event causing the vector S to update, i.e. when a user interaction with the product updates the surrogate vector to the new state S_t from a previous state S_{t-1}. With an updated causal surrogate $V_t=V(S_t)$, we attribute the difference $R_t:=V_t-V_{t-1}$ ($V_0=0$) as a pseudo-reward to the user-product interaction making that state update at t.

For our application to Airbnb booking behavior, the state vector S_t is defined by a set of listing view engagement features as well as listing and trip attributes in Table 2. As in training, we define a state vector for each (user, listing) pair. As users are browsing listings, each listing view represents a time step t and a state change for a user-listing pair. If it is the user's first time viewing the listing, a user-listing pair is initialized. We look back 14 days to retrieve the cumulative states up to $t-1$ and t and use the most recently (weekly) trained model to score V_{t-1} and V_t. The difference $V_t - V_{t-1}$ is attributed to the listing view at step t. The result is a listing-view level attribution which we call *listing-view utility*.

We remark that the attribution aspect is an even more fundamental improvement than the variance reduction, with numerous applications in online measurement and beyond. For Airbnb's case, not all listing views and user activities contribute equally towards a booking decision. If we can associate different amounts of reward to different listing-view engagements, we can then further attribute that reward to specific *actions*. We can also aggregate it to the level of various analysis units such as host, listing, guest, and session. Continuously attributed reward without delay is a prerequisite for adaptive experimentation [30].

4.3 Model-Based Surrogate Metrics for Delayed Outcomes

With the causal surrogate model and continuously attributed listing-view utilities, we can now define new surrogate metrics for the bookings metric.

The user-level surrogate metric is straightforward to define — we simply sum the listing-view utilities from the listing-view level a user has been attributed across "time" t, obtaining a user-level value $U(i) = \sum_t R_t(i)$ for each user i. This demonstrates the flexibility of fine-grained continuous attribution and the ease of aggregation to other analysis units. A side note about continuous attribution is that the notion of pre-experiment adjustment from CUPED is implied automatically. If a user already started their trip planning prior to the start of an experiment, we sum only the incremental utilities gained within the experiment window.

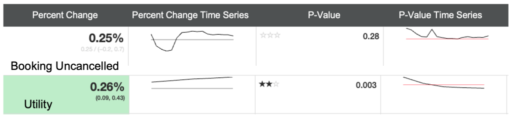
Unlike the bookings metric, which takes integer values, user-level listing-view utility takes a continuous value. Since most users only make 1 booking, we cap the user level utility metric at 1 by default. Under further investigation, we found the distribution of user level utility to be highly skewed: only about 25% of users has utility more than 0.01. We created several versions of the listing-view utility metric, capped at 0.3, 0.2 and 0.1. Note that capping here is not for outlier removal but for extra sensitivity gain when the treatment affects the whole distribution of user-level utility in the same direction. The continuous metric value of user-level utility also opens future development to explore distributional comparison beyond comparing two means [1].

4.4 Results

Variance reduction We report the variance of percent lift for the bookings metric, listing-view metrics, and the newly created listing-view utility metrics in Table 3. The variance of the listing-view utility metric is 57% lower than that of the bookings metric. We see that capping leads to further improvement: listing-view utility metric capped at 0.3, 0.2 and 0.1 achieved 76%, 79% and 84% variance reduction, respectively, compared to bookings. Though the listing-viewers metric is more sensitive than listing-view utility, we will show later that it lacks strong alignment with the bookings metric.

	Bookings	Bookers	Listing Views	Listing Viewers
Variance of percent lift	1	0.88	0.35	0.06
	Utility	Utility Capped 0.3	Utility Capped 0.2	Utility Capped 0.1
Variance of percent lift	0.43	0.24	0.21	0.16

Table 3: Variance of percent lift normalized based on the variance of the bookings metric. The new utility metric (0.43) reduced variance by more than 50% compared to bookings (1), and capping at 0.3, 0.2 and 0.1 further significantly reduced the variance by 75% to 85%.



Percent Change	Percent Change Time Series	P-Value	P-Value Time Series
0.25% 0.25 / (−0.2, 0.7) Booking Uncancelled		☆☆☆ 0.28	
0.26% (0.09, 0.43) Utility		★★☆ 0.003	

Figure 3: A screenshot from a real experiment report showing time series of percent change and p-value. Top (uncancelled bookings) vs. bottom (listing-view utility). Reduced variance resulted in a much more stable percent change time series reaching stat. sig. conclusion earlier.

How big of a difference can 50% (or more) variance reduction make in real-world experimentation? The results of one experiment in Airbnb's experimentation reporting framework are displayed in Figure 3. We see the time series of percent change and p-values changing over the duration of the experiment for the uncancelled bookings metric and the new listing-view utility metric. Not only does the new metric reach a much smaller and more conclusive p-value, but the point estimates also fluctuate much less as a result of its lower variance.

Given Eq (2), the variance reduction and sensitivity gain do not come as much of a surprise. A more important question is whether the causal surrogate assumption holds, and whether the utility metric is truly capturing all or most of the causal effect on booking. We backtested the utility metrics on 141 search ranking experiments from 2021 to 2022 Q2. Naively, one might think that we can just compare or compute the correlation between the two metrics. However, because the bookings metric has much larger variance than utility, the variance of the booking metric accounts for the majority of the noise and can easily dilute the correlation. We restricted the comparison to a subset of 32 experiments where the bookings metric had a statistically significant effect, indicated by a p-value below 0.05 and a sample size of at least 4 million users. The p-value distribution of the total 141 experiments are shown in Figure 4.

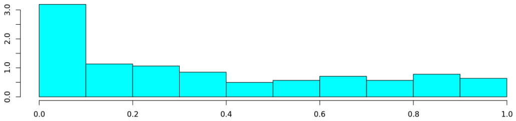


Figure 4: Uncancelled booking p-value distribution from 141 experiments. 32 of them have p-value below 0.05 and sample size at least 4 million users.

In Figure 5, we compare the alignment of surrogate metric candidates with uncancelled bookings over the corpus of 32 selected experiments in a series of plots. Each point represents a single experiment, and the size of the point represents the sample size of the experiment. We compare the effect size estimates for the surrogate metric on the x-axis against the effect size estimates of uncancelled bookings on the y-axis. A diagonal line is included as a reference for perfect alignment between the surrogate metric and the uncancelled bookings metric.

Prior to the utility metric, *dated listing-viewers* (users who viewed a listing with check-in and out dates specified) was the best surrogate metric from a previous meta-analysis. When comparing to dated listing-viewers, the utility metric significantly improves directional agreement. There are zero experiments with disagreement

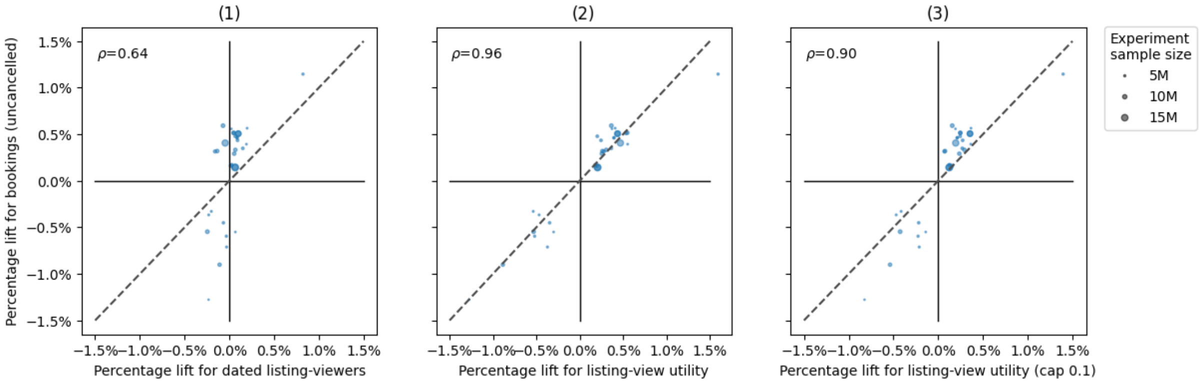


Figure 5: Surrogate metric alignment with the uncancelled bookings metric in a corpus of 32 experiments. Plot (1) shows the alignment of uncancelled bookings with dated listing-viewers, (2) shows alignment with listing-view utility, and (3) shows alignment with the listing-view utility capped at 0.1. Each point represents a single experiment, and the size of the point represents the sample size of the experiment. The Pearson correlation coefficient ρ is also displayed for each plot.

between listing-view utility and uncancelled bookings, whereas listing-viewers has 5/32 cases of disagreement.

In addition to strong agreement, the listing-view utility metric effect size estimates are also closer to those of uncancelled bookings. We display the Pearson correlation coefficients for each plot. Both listing-view utility metrics achieved more than 0.9 correlation while the previous best surrogate, dated listing-viewers, scored only 0.64. Because listing-view utility is calibrated and capped to 1 by default, we expect to see points close to $y = x$ if the surrogate model reflects the underlying causal mechanism. The larger the sample size, the closer the bookings metric is to its ground truth. We see that the largest points are closer to the diagonal line than the others. In the third subplot of 5, we still observe good alignment even when capped at 0.1 (87.5% percentile). There is some hint that capped utility effect size estimates are smaller than booking effect size estimates. Because we are selecting experiments in which the bookings metric was statistically significant, we do expect that the selection process will cause the booking metric effect size estimates to have a systematic upward bias. We do observe this bias in the the third plot. At the same time, capping too aggressively can lead us to ignore treatment effects on the upper percentiles of the utility distribution. In practice, we found that capped utilities can be great surrogate metrics for the utility metric itself.

Attribution to search level At Airbnb, online experiments in the search ranking domain are typically implemented with a randomization unit of a logged-in user or a visitor to ensure that we can have a comprehensive understanding of the treatment effect to the whole guest booking journey. Randomization at the search request level can improve the statistical power of metrics at the cost of observability for metrics aggregated above the search request level because it is no longer possible to capture interactions spanning multiple searches. Nevertheless, search-level randomization and its variations are commonly used in exploratory experiments. One important variation of search level randomization for ranking is interleaving [37].

When the objective is to understand booking behavior, a key challenge remains: how do we define a search-level metric that can indicate user-level booking impact? Before the introduction of the listing-view utility metric, several approaches to attribution were proposed, and they all focused on the booked listing. If a user didn't book in the end, then all of the engagement signals from their exploration process remained untapped. Among users who booked, they often searched and clicked on the same listing multiple times before booking. Distributing the reward of the completed booking to preceding listing views and searches was a challenging problem.

By comparison, the listing-view utility metric offers a data-driven way to attribute a booking to each listing view and the corresponding referral searches. Reward is distributed according to how much information the user gathered on the page toward their booking propensity. Importantly, this also applies to *almost-bookings* – those listings that the user seriously considered, and compared against, but didn't book in the end. Thus, it is more *efficient* to gather signals from the vast majority of unbooked listings, and also more *targeted* because reward is assigned to the searches and search results that lead to the greatest elevation of booking intention.

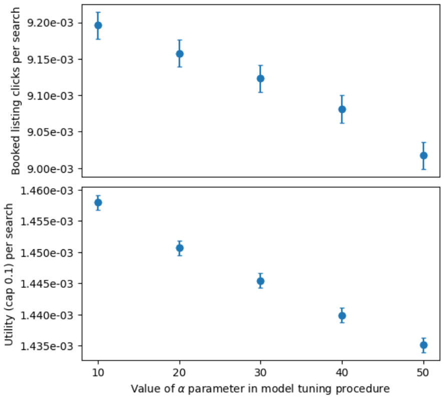


Figure 6: One-dimensional parameter tuning results for the α parameter at the search request level. Point estimates and confidence intervals of booked listing-views per search (top) and the new listing-view utilities per search (bottom) are shown for each value of the model parameter α.

We demonstrate two applications of search-level attribution. The first is a hyper-parameter tuning procedure that we performed to identify a good trade-off between bookings and customer support cost. We separated online traffic into five different arms based on possible values of a model parameter *alpha* that controls the weight between the two objectives, randomizing at the search request level. We then estimated the impact on bookings and listing-view utility. At the search request level, some preliminary attribution method for the bookings metric was needed to enable comparison with listing-view utility. The *booked listing-views* metric is a count of listing-views occurring after a search request containing a listing that was later booked.

We found that the variance of the percent change of the listing-view utility metric capped at 0.1 was 85% lower than the variance of booked listing-views per search. Results are displayed in Figure 6. We report point estimates and confidence intervals for booked listing-views per search (top) and the new listing-view utility per search capped at 0.1 (bottom). We note that the ranges of the y-axes are substantially different because booked listing-views is a less sophisticated metric and includes many equally weighted interactions. The two graphs share a decreasing trend, but the booked listing-views metric has a much wider confidence interval than its utility counterpart. The confidence intervals on the top plot often cover adjacent parameter's point estimates, while the utility metric's confidence intervals are clearly separated and close to disjoint. We also applied a similar enhancement to interleaving metrics, achieving similar variance reduction and improved alignment with follow-up experiments.

Another promising application of the model-based surrogate metric is to better understand patterns of repeated engagement. An example of this is when a user views the same Airbnb listing multiple times before making a booking decision. Figure 7 shows three cohorts: users with 6 listing-views (left), 12 listing-views (middle) and 20 listing-views (right) prior to a booking request. We plot the 75th percentile of listing-view utility for each listing-view, from the first interaction to the last one before booking request. (The utility distribution is highly skewed. A similar trend can be shown with the average or the median, but the 75th percentile manifests the uptick of utility better.) We found that listing-views close to booking request unsurprisingly have high utility because users actively inspect and verify many details about a listing before making a booking decision. Contrary to common beliefs about first impressions, the results show that first listing view often does not yield high utility for those users, and the first uptick of utility happens at 3rd to 4th views of the same listing. Further investigation reveals that many early listing-views are characterized by photo scrolling behavior. These interactions are less meaningful engagements than looking at reviews, amenities, and descriptions. This discovery has direct impact on how we value *retargeting* — upranking results that a user has repeatedly clicked will lead to further page-views and may improve conversion up to a point. For the 20 listing-view cohort, the utility metric has a decreasing trend after 5 views, and the booking decision takes place after a gap of low-utility views. This pattern suggests that such intermediate low-utility views may not provide incremental value to the user, and can be better replaced with diversified results.

5 IN-EXPERIMENT COUNTERFACTUAL TREATMENT EXPOSURE INDEX

We share another novel application of search-level utility attribution. Constructing a continuous user-level index that quantifies the amount of treatment exposure enables us to focus on the subset of users that were most impacted by the treatment. The idea is motivated by tight triggering as an efficient variance reduction method [8, 27], but with an important twist: the definition of the exposure index is built from signals in-experiment, and are expected to be affected by treatment itself. This also echoes the theme of this paper, breaking away from traditional variance reduction thinking.

To avoid sample ratio mismatch, we propose to segment by percentile of a continuous exposure index per treatment group. We order subjects by an exposure index and compute percentile in each group separately. Then we select the top-k percentile from each group and compare them. By design, as long as there is no sample ratio mismatch for the original dataset and there are no ties for the index values, with both groups selecting the same percentage of subjects, the resulting subset won't have sample ratio mismatch.

A few important remarks. First, this method can be seen as matching subjects between two groups based on an order statistic. The success of this approach depends on how effectively the resulting index aligns with individual treatment effects. Second, it is important that the index is continuous and not discrete so that the percentile cut does not need to resolve ties (at least for the desired high percentile level). Third, a statistical analysis using this data-adaptive segmentation approach must take the percentile cutting process into account. In particular, variance estimations of any metric after the top $k\%$ selection should not rely on the sample variance formula, and would require additional adjustment such as using bootstrap [16]. The amount of variance adjustment depends on the joint distribution of the subject-level exposure index and the metric values. Lastly, [1] proposed to match quantiles of a metric of interest to engineer a new metric more sensitive than the mean or a fixed quantile. We are not matching the quantile of a metric of interest, but rather matching to the quantile of a continuous index engineered to reflect treatment effect exposure.

The treatment exposure index is constructed with the help of the listing-view utility metric. For each search request and each listing view from this search, we compute the *ranker impact-weighted utility* as:

$$u \times \left(1 - \gamma^{\max(|p_T - p_C| - \eta, 0)}\right),$$

where u is the listing-view utility attributed to the search and p_T and p_C are the positions of the listing in the ranked feed in treatment and control. The difference $|p_T - p_C|$ is called the *counterfactual positional difference* because only one position is available during the experiment, and the other one is computed offline. η controls the minimum position difference that will impact the click-through rate, and γ controls the click-through decay and is a value between 0 and 1. When $\eta = 1$, a position difference less than or equal to 1 is assumed to have no impact and the ranker impact weighted utility is 0. The smaller the γ, the more ranking position matters. At the user level, we sum the ranker impact-weighted utility to represent the degree to which a user's booking decision is impacted by the ranking change. A high value of this index requires both a high-utility click from the search results and a large counterfactual positional difference between two rankers.

		Percent lift	Stand Error	t-stat	VR	bootstrap correction factor
full size	Whole population	0.58%	0.20%	2.87	NA	NA
	Top 5% diluted	0.62%	0.12%	5.18	65%	1.08
	Bottom 95% diluted	-0.04%	0.16%	-0.22	NA	1.05
1/5 size	Whole population	0.41%	0.45%	0.91	NA	NA
	Top 5% diluted	0.59%	0.25%	2.26	65%	1.08
	Bottom 95% diluted	-0.18%	0.35%	-0.5	NA	1.05

Table 4: A real experiment using in-experiment treatment exposure index to tease out the top 5% users from the rest. All effects and standard errors were diluted to effect on overall population for comparison. Top three rows are full data and bottom 3 rows we sample only 20% to emphasize the benefit of 65% variance reduction (VR). Bottom 95% results were shown to check assumption that those users were much less affected. Bootstrap correction factor shows how much bootstrap variance estimate deviates from naive sample variance formula.

We present results from a past ranking experiment In Table 4. The treatment showed a percent lift of 0.58% with standard error of 0.2%

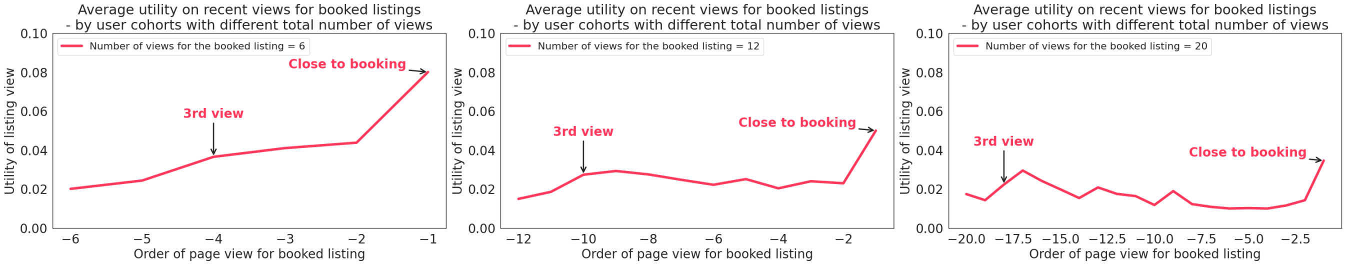


Figure 7: First uptick of utility happens usually around the 3rd and 4th page-views, not the first page-view.

and t-statistics of 2.87, indicating high statistical significance. However, this result required more than 12M users in each group to obtain this level of sensitivity. In addition to the experiment results for the whole population, we includes the results for the same experiment with the counterfactual treatment exposure index, restricting users based on the ranker impact-weighted listing-view utility attribution (setting $\gamma = 0.9$ and $\eta = 1$). Matching the top 5% of users retains almost the same effect with standard error almost halved (we already diluted the effect to overall effect [8]) and the variance was reduced by 65%. The bottom 95% comparisons check the assumption that the treatment effect for this lower exposure tier is close to 0.

What if we only ran this experiment with a fifth of traffic? We further down-sampled the data to simulate this scenario. The same procedure results in the same 65% variance reduction, with t-statistics changed from 0.91 (inconclusive) to 2.26 (statistically significant). The percent lift estimate for the top 5% after dilution (0.59%) is much closer to the results using the full dataset (0.58%). Finally, the bootstrap correction factor shows that the naive sample variance formula underestimates variance by 5% to 8%. It is worth noting this correction factor depends on the correlation between the exposure index and target metrics so it varies case by case: the higher the correlation, the larger the correction factor. An ideal exposure index has low correlation to the target metric (correction factor close to 1), but has high correlation to the individual treatment effect.

6 RELATED WORK

Model-based surrogates and the surrogate assumption were proposed in [2] and can be seen as a special case of the front-door criterion in [35]. [2] also noted the efficiency gain, but their main objective was to study long-term outcomes that are not observable in a short experiment period (e.g. effect of job training on employment). The focus of our method is to exploit the variance reduction, even when the delayed outcome is short-term and can be observed in a normal experiment period. In particular, we use the model prediction even for subjects with an observed outcome Y. We further use the causal surrogate model to approximate the value function [38] for incremental reward attribution. [15] presented a similar work with the variance reduction angle, but focused on predicting a future value of the same metric that is already continuously observed, e.g. sessions-per-user after 2 weeks using 1 week's outcome. Their work shares the same source of variance reduction as ours and causal surrogacy in general — due to smoothing effect of conditional expectation, but without the surrogate modeling and not suitable for delayed outcome. [41] proposed a model-based metric utilizing listing-view engagements, but based on individual listing view instead of tracking (user, listing) pair over time with continuous attribution. In A/B testing literature, most variance reduction work has been focused on exploiting pre-assignment covariates[10, 31, 45].

For delayed outcomes and sparse signals, there exists prior work modeling the delay or jointly modeling delay and binary prediction as training with missing data [5, 44]. These works aim at improving the prediction of a delayed outcome and evaluated their methods based on the accuracy of their forecast. Our work differs because the main objective is not to predict a future metric value but to better estimate the treatment effect, and evaluate our methods based on efficiency gain and empirical alignment with existing goal metrics' effect estimation. Missing data is not a core challenge in our application.

[13] used quantile function of a metric to bound individual treatment effect variance. [3] estimated individual treatment effect using pairs matched by quantiles of the metric values.

7 CONCLUSION AND FUTURE WORK

In this work, we constructed a model-based causal surrogate to serve as a value function representing the progress toward a sparse and delayed outcome. Using in-experiment signal for variance reduction is a new direction in the literature of online experimentation. By leveraging user activity as surrogate signal, we demonstrated that the causal surrogate model enabled a variance reduction of 50% to 85% compared to the bookings metric. Furthermore, we can attribute incremental gain to individual user-product interactions as well as the actions that lead to them. This enables us to experiment and evaluate at a more granular level than before, enabling flexible aggregation and improved user understanding. We have also identified opportunities to measure the value of retargeting and personalization.

We presented another novel idea of engineering a treatment exposure index based on listing view attribution weighted by counterfactual ranker impact, and use this index to exclude subjects with close to 0 treatment effect. By selecting the top-k percentile of subjects from treatment and control groups respectively, we achieved more than 60% variance reduction for the bookings metric at Airbnb. Improving understanding of various theoretical aspects of statistical analysis post quantile matching remains an important area for future exploration. The selection of the cutoff k should be determined before the analysis. In practice, making this selection step data-adaptive is useful, but requires post-selection inference adjustment.

ACKNOWLEDGEMENT

We would like to thank Austin Chan and Chun How Tan for providing listing view engagement data, Wangfan Fu and Do-kyum Kim for hyperparameter tuning application, Liwei He, Huiji Gao and Xin Liu for adoption of surrogate metric, Ron Kohavi, Lo-Hua Yuan, Amit Gandhi and four anonymous reviewers for giving feedback.

REFERENCES

[1] Susan Athey, Peter J Bickel, Aiyou Chen, Guido Imbens, and Michael Pollmann. 2021. *Semiparametric Estimation of Treatment Effects in Randomized Experiments.* Technical Report. National Bureau of Economic Research.

[2] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

[3] John Cai and Weinan Wang. 2022. A Systematic Paradigm for Detecting, Surfacing, and Characterizing Heterogeneous Treatment Effects (HTE). *arXiv preprint arXiv:2211.01547* (2022).

[4] George Casella and Roger L Berger. 2002. *Statistical Inference.* Duxbury Press: Pacific Grove, CA.

[5] Olivier Chapelle. 2014. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 1097–1105.

[6] Laura Cosgrove, Jen Townsend, and Jonathan Litz. [n. d.]. Deep Dive Into Variance Reduction. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/deep-dive-into-variance-reduction/.

[7] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems.* 293–296.

[8] Alex Deng and Victor Hu. 2015. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* 349–358.

[9] Alex Deng and Xiaolin Shi. 2016. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[10] 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 WSDM Conference.* 123–132.

[11] Alex Deng, Lo-Hua Yuan, and Alexandre Salama-Manteau. 2021. Variance Reduction for Experiments with One-Sided Triggering using CUPED. *arXiv preprint arXiv:2112.13299* (2021).

[12] Drew Dimmery, Eytan Bakshy, and Jasjeet Sekhon. 2019. Shrinkage Estimators in Online Experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2914–2922.

[13] Peng Ding, Avi Feller, and Luke Miratrix. 2019. Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* 114, 525 (2019), 304–317.

[14] Pavel Dmitriev and Xian Wu. 2016. Measuring metrics. https://archive.org/details/MeasuringMetricsCIKM2016.

[15] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the 24th International Conference on World Wide Web.* 256–266.

[16] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap.* CRC press.

[17] Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 2019. Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, 2156–2164.

[18] Jean Garcia-Gathright, Brian St Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* ACM, 55–64.

[19] Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651.

[20] Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. 2021. Machine Learning for Variance Reduction in Online Experiments. *arXiv preprint arXiv:2106.07263* (2021).

[21] Somit Gupta et al. 2019. Top Challenges from the First Practical Online Controlled Experiments Summit. *SIGKDD Explor. Newsl.* 21, 1 (May 2019), 20–35.

[22] G. W. Imbens and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge University Press.

[23] Raphael Lopez Kaufman, Jegar Pitchforth, and Lukas Vermeer. 2017. Democratizing online controlled experiments at Booking. com. *arXiv preprint arXiv:1710.08217* (2017).

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[25] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *Proceedings of the 20th ACM SIGKDD Conference (KDD '14).* 1857–1866.

[26] Ron Kohavi, Alex Deng, and Lukas Vermeer. 2022. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 3168–3177.

[27] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (2009), 140–181.

[28] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy online controlled experiments: A practical guide to a/b testing.* Cambridge University Press.

[29] Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel Stevens. 2022. Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology. *arXiv preprint arXiv:2212.11366* (2022).

[30] Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. 2019. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis* 14, 2 (2019), 495–519.

[31] Winston Lin. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7, 1 (2013), 295–318.

[32] John A List, Ian Muir, and Gregory K Sun. 2022. *Using Machine Learning for Efficient Flexible Regression Adjustment in Economic Experiments.* Technical Report. National Bureau of Economic Research.

[33] Jiannan Lu, Yixuan Qiu, and Alex Deng. 2019. A note on Type S/M errors in hypothesis testing. *Brit. J. Math. Statist. Psych.* 72, 1 (2019), 1–17.

[34] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Obtaining Calibrated Probabilities from Boosting.. In *UAI*, Vol. 5. 413–20.

[35] Judea Pearl. 2009. *Causality.* Cambridge university press.

[36] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 235–244.

[37] Filip Radlinski and Nick Craswell. 2013. Optimized interleaving for online retrieval evaluation. In *Proceedings of the 6th ACM WSDM Conference.* 245–254.

[38] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[39] 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.

[40] Anastasios Tsiatis. 2006. *Semiparametric Theory and Missing Data.* Springer-Verlag.

[41] Bradley C Turnbull. 2019. Learning Intent to Book Metrics for Airbnb Search. In *The World Wide Web Conference.* ACM, 3265–3271.

[42] Gerald Van Belle. 2011. *Statistical rules of thumb.* John Wiley & Sons.

[43] Rashmi Korlakai Vinayak and Ran Gilad-Bachrach. 2015. Dart: Dropouts meet multiple additive regression trees. In *Artificial Intelligence and Statistics.* PMLR, 489–497.

[44] Zenan Wang, Carlos Carrion, Xiliang Lin, Fuhua Ji, Yongjun Bao, and Weipeng Yan. 2022. Adaptive Experimentation with Delayed Binary Feedback. In *Proceedings of the ACM Web Conference 2022.* 2247–2255.

[45] 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.* ACM, 645–654.

[46] 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.

[47] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2227–2236.

[48] Xuan Yin and Liangjie Hong. 2019. The identification and estimation of direct and indirect effects in A/B tests through causal mediation analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2989–2999.