MIS Quarterly

RESEARCH ARTICLE

# A PRESCRIPTIVE ANALYTICS FRAMEWORK FOR OPTIMAL POLICY DEPLOYMENT USING HETEROGENEOUS TREATMENT EFFECTS[1]

**Edward McFowland III, Sandeep Gangarapu, and Ravi Bapna**
Carlson School of Management, University of Minnesota, Minneapolis, MN, U.S.A.
{emcfowland@hbs.edu} {ganga020@umn.edu} {rbapna@umn.edu}

**Tianshu Sun**
Marshall School of Business, University of Southern California
Los Angeles, CA, U.S.A. {tianshus@marshall.usc.edu}

*We define a prescriptive analytics framework that addresses the needs of a constrained decision-maker facing, ex ante, unknown costs and benefits of multiple policy levers. The framework is general in nature and can be deployed in any utility-maximizing context, public or private. It relies on randomized field experiments for causal inference, machine learning for estimating heterogeneous treatment effects, and on the optimization of an integer linear program for converting predictions into decisions. The net result is the discovery of individual-level targeting of policy interventions to maximize overall utility under a budget constraint. The framework is set in the context of the four pillars of analytics and is especially valuable for companies that already have an existing practice of running A/B tests. The key contribution of this work is to develop and operationalize a framework to exploit both within- and between-treatment arm heterogeneity in the utility response function in order to derive benefits from future (optimized) prescriptions. We demonstrate the value of this framework as compared to benchmark practices—i.e., the use of the average treatment effect, uplift modeling, as well as an extension to contextual bandits—in two different settings. Unlike these standard approaches, our framework is able to recognize, adapt to, and exploit the (potential) presence of different subpopulations that experience varying costs and benefits within a treatment arm while also exhibiting differential costs and benefits across treatment arms. As a result, we find a targeting strategy that produces an order of magnitude improvement in expected total utility for the case where significant within- and between-treatment arm heterogeneity exists.*

**Keywords:** Prescriptive analytics, heterogeneous treatment effects, optimization, observed utility rank condition (OUR), between-treatment heterogeneity

## Introduction

We address the general problem of a budget-constrained decision maker facing, ex ante, unknown costs and benefits from multiple policy levers that can potentially be deployed to optimize an organizational goal. We define and deploy a prescriptive analytics framework as one that folds together the use of (1) randomized field experiments for causal inference around the estimation of ex ante unknown costs and benefits; (2) machine learning to identify heterogeneity in treatment effects and thereby go beyond inference around average treatment effects; and (3) constrained optimization to optimally decide which subpopulation of individuals to treat with which policy levers, maximizing profit in the presence of organizational and individual level constraints. While each of the above three folds—causal inference, supervised machine learning, and constrained optimization—are academic disciplines on their own, the distinct contribution of this paper

[1] Indranil Bardhan was the accepting senior editor for this paper. Balaji Padmanabhan served as the associate editor.

is to stitch specific aspects of them together into a decision-making framework. In doing so we advance the thinking around each of the individual disciplines as well. By using supervised machine learning on data generated by experiments, originally deployed to derive causal inference, we are able to discover patterns of within- and between-treatment arm heterogeneity in the individual and group response functions. This allows our work to go beyond using the average treatment effect as the decision lever from randomized experiments. In estimating unbiased individual level costs and benefits of multiple levers we parameterize the inputs, or "the data," that feeds into the optimization problem. This is generally not the purview of the optimization literature, which assumes that the "data" exists, and focuses on deriving, for example, efficient algorithms to solve computationally complex (often NP-hard) problems.

Consider, as a motivating example, a marketing manager deciding between three different types of calls to action (CTA) of initiators of a referral campaign. Jung et al. (2020) experiment with an altruistic, equitable, and egoistic framing of the call to action, to activate potential senders of a referral. The target population in this situation is the set of past customers, for whom firms typically collect a vast amount of demographic and behavioral data. Further, promotions such as referral programs are incentive-laden word-of-mouth mechanisms, and like all promotions, have underlying costs and benefits associated with them. In the context investigated by Jung et al. (2020), each sender and recipient of a referral is eligible to receive a free product (valued at $25) with free shipping included. Thus, depending on how many referrals a person sends out, how many of these referrals lead to recipient purchases, and the varying dollar amounts of these purchases, the firm can realize many different values of cost and benefit. Furthermore, different value patterns can emerge from different subpopulations of individuals as a consequence of different treatment allocations. Consider the following scenario:

1. User A (targeted using an equitable CTA) invites a friend who buys a $200 item and receives a $25 gift. User A also receives a $25 gift for initiating the referral. The resulting utility to the firm is $150

2. User B (targeted using a selfish CTA) invites a friend who buys a $20 item and receives a $25 gift. User B also buys a $20 item and receives a $25 gift. The resulting utility to the firm is -$10

3. User C (targeted using an altruistic CTA) invites 8 friends, each of whom buys a $35 item and receives a $25 gift. User C also buys a $35 item and receives a $25 gift. The resulting utility to the firm is $90.

The challenge is that ex ante the marketer does not know whether any of these calls to action have a causal impact on profitability, let alone know whether the impact of the three different arms may be heterogeneous, as a function of existing customer "data"—e.g., characteristics and behaviors. Further, given organizational constraints (e.g., budget) or those at the individual level, the problem of allocating individuals to treatments in order to maximize utility is an NP-hard problem (Martelo and Toth 1990). As a motivating example of the generality of our framework, consider an individual-level constraint that restricts the availability of certain treatments to certain individuals based on past interactions with the company. For instance, if there are existing users who have already been targeted with a particular call to action multiple times, the company may not want to overwhelm them with the same call to action yet again.

We demonstrate that challenges of this type can be addressed using a prescriptive analytics approach built atop the key pillars of analytics (Figure 1): causal inference, machine learning, and optimization to enable a budget-constrained decision maker to optimally target interventions with ex ante unknown costs and benefits. A variety of organizations already use some variant of the causal inference paradigm (called "A/B testing" in the industry) to determine which ads to show, what features to deploy, or what type of incentives to provide in order to motivate users to perform an action (Kohavi and Thomke 2017). In the first stage of our approach, the intent is to layer on to this existing practice the ability to use machine learning methods to make inferences beyond the average treatment effects. More specifically, we showcase how to use the data from A/B testing to develop and validate a robust heterogeneous treatment effect (HTE) procedure that can provide estimates of cost and benefit for each individual in each treatment condition. The next stage of our approach utilizes the "randomly optimal" targeting that occurs by chance from random assignment as a means for validating the existence of sufficiently exploitable between-condition heterogeneity. As a result of random assignment, our HTE step gives an unbiased (selection-free) ordering of treatments for each individual. Therefore, we can rely on random assignment to (by chance) produce a targeting of individuals at varying degrees of optimality—e.g., some individuals will experience their optimal assignment. This variation in treatment optimality allows us to examine whether subjects who were randomly placed in their (predicted) optimal condition exhibited higher utility than those that were not. We call this the observed utility rank condition (OUR), a metric that serves as a necessary condition for progressing to the final (optimization) stage of our approach. In the final stage, we exploit the within- and between-treatment heterogeneity in cost and benefit terms for the development of optimal prescriptions, i.e., targeting treatments to individuals while respecting organizational constraints. Although formal definitions are laid

out in the sections that follow, in essence, our process adapts to the existence of heterogeneity across individuals (or subpopulations) for a given treatment and requires the existence of heterogeneity between treatment conditions within individuals (or subpopulations). This makes it worthwhile for the optimization to exploit such heterogeneity and match treatments to the most suitable sets of individuals. If this heterogeneity between treatment conditions does not exist, there is no value in progressing to the prescription phase. When this heterogeneity does exist, we validate this final stage by comparing the utility generated by our prescriptions (allocations) against the current practices of assigning all individuals to the condition that has the highest average treatment effect, as well as uplift modeling (Rzepakowski and Jaroszewicz 2010).

We demonstrate the value of our proposed three-stage process using two real-world settings. First, we consider a public policy setting used to motivate blood donations, and find progressive value in our three-stage framework. In particular, we find that (1) our HTE model has a low out-of-sample mean absolute error, indicating that it is able to capture the data-generating process of the underlying treatment arms; (2) there exists clear observed ordering of the OUR metric, indicating enough heterogeneity across treatment conditions to be exploited; and (3) there exists significant gains in utility from the optimal allocation as compared to ATE and uplift modeling. In contrast, in the context of a call-to-action experiment for referral marketing, while the analysis passes Stage 1 (low MAE), it fails Stage 2. This implies that even though the model captures the underlying data-generation process of each of the treatment arms, there is not sufficient evidence for the existence of heterogeneity between conditions. Furthermore, the absence of between-condition heterogeneity indicates a lack of value to be gained in Stage 3. Thus, our practical contribution stems from providing an overall process that can be applied to any experimental setting of multiple treatment conditions, with ex ante unknown costs and benefits. Step 2, in particular, provides a pragmatic tollgate, that prevents organizations from undertaking a targeting strategy that lacks sufficient evidence for generating increased utility. Such a strategy may lead to prescriptions that are laden with costs (e.g., the cost of chosen suboptimal treatment, opportunity costs, etc.) without sufficiently counterbalancing benefit.

Overall, we contribute by developing and operationalizing a prescriptive analytics framework that combines randomized field experiments for causal inference; machine learning to exploit heterogeneity and advance this inference beyond the average treatment effects; and constrained optimization to optimally decide which subjects to treat with which policy levers, to maximize profit in the presence of organizational and individual level constraints. We believe that this prescriptive analytics framework encourages companies to take an integrated view of the four pillars of analytics (Figure 1). In particular, for sake of completeness, our framework also depends on using managerial judgment and exploratory analytics to generate treatment conditions, where there is the potential for an effect. Our paper is also the first to highlight, and develop a metric (i.e., the OUR metric) around the importance of between-treatment arm heterogeneity of cost and benefit treatment effects.

## Background and Related Literature

Our work is adjacent to the emerging literature around the use of machine learning for heterogeneous treatment effects (Wager and Athey 2018; Athey and Imbens 2016; McFowland III et al. 2018). However, as in Imai and Strauss (2011), we focus on extending these ideas to the design of optimal policy decisions. Moreover, Imai and Strauss's work (2011) (which proposes a targeting approach in the context of "get-out-the-vote" campaigns) and our work argue that if policy makers solely rely on average treatment effects, they fail to exploit potentially valuable sources of treatment heterogeneity. However, while the overall objective in (Imai and Strauss 2011) is similar to ours, we differ in the generality of the optimization approach to determine treatment assignment. In particular, Imai and Strauss (2011) adopt a Bayesian optimization approach, requiring knowledge of a suitable prior that is then subsequently revised with experimental data to determine the posterior distributions of various treatments' effects on the probability of voting. Furthermore, the focus is maximizing the treatment effect on benefit (measured as a probability) subject to a cost constraint. It is not immediately obvious how to optimize general utility (the difference in benefit and cost) where cost is not measured as probabilities (but instead as currency) and costs are unknown ex ante, both of which occur in our setting.

Additionally, our work relates to but is significantly different from the uplift modeling literature (Rzepakowski and Jaroszewicz 2010) and the contextual bandits literature (Auer 2002; Langford and Zhang 2008; Li et al. 2010) in machine learning and the design of experiments. The main commonality lies in the end objective of personalizing decisions, such as marketing interventions, based on consumer characteristics and past behaviors. The key idea in uplift modeling is to go beyond building a model to simply predict who is likely to respond to an intervention, but to target those that will respond *because* they received the intervention. Intuitively, the goal is to develop a model whose estimation of treatment effects and subsequent predictions of who should be targeted will result in better responses relative to control group subjects that are randomly targeted.

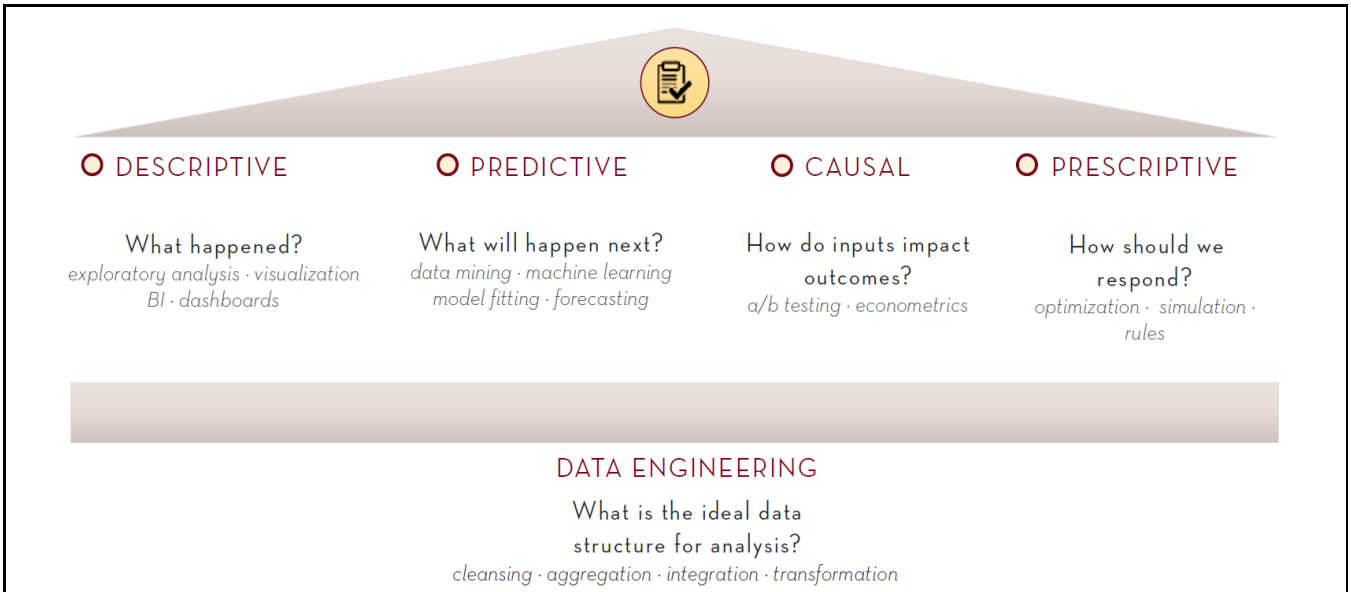



**Figure 1. Four Pillars of Analytics**

The practice of targeting customers who will respond even in the absence of the intervention actually results in unnecessary costs. If we extend this setting to multiple treatments, with sequential allocation of subjects to treatments considering their demographic and behavioral data, we enter the setting of the contextual bandit problem. Therein, the key issue is dealing with the exploration-exploitation balance: trading off the focus on increasing learning with the focus on earning. Exploration is used to estimate individual rewards based on known contextual data, while exploitation simultaneously attempts to maximize cumulative returns in a sequential decision-making framework. Multi-armed bandit approaches are gradually being adopted in more conventional management decision contexts, such as the pricing of multiple products (Misra et al. 2019) and the acquisition of customers using display advertising (Schwartz et al. 2017). Our framework borrows ideas from both these streams of literature and is consistent with what Bertsimas and Kallus (2020) describe as moving from predictions to decisions. A key aspect of this move is the incorporation of organizational- and individual-level constraints; for example, budgets and individual-level ineligibilities, respectively. Further, to demonstrate the value of our proposed approach, we do not require the additional complexity of a sequential decision-making policy optimization formulation, which is characteristic of the known algorithms for the contextual bandit problem (Auer 2002; Li et al. 2010). In our work, we decouple the joint optimization of exploration-exploitation by treating the learning phase—which results from utilizing heterogeneous treatment effect procedures to process and extract information from randomized experiments—as an input into the optimal targeting phase. From a practical point of view, our approach requires less supporting infrastructure and is much easier to deploy than bandit approaches, improving its accessibility to less technically sophisticated organizations. Moreover, our approach can provide immediate value for organizations that have already developed some causal analytics capabilities, and potentially have data from past experimentation contained in their administrative records.

It is important to note that the popularity of multi-armed-bandit-based approaches (Chu et al. 2011; Ding et al. 2013; (Joulani et al. 2013) has risen with the popularity of A/B testing on digital platforms in what some call an online experimentation model (think content testing on a web page, or showing an ad on the Facebook wall, and observing subsequent engagement). We highlight thi=s, because the use of multi-armed bandits is (essentially) predicated on the ability to update the policy of treatment delivery throughout the course of the experimentation process. While such conditions may be satisfied in certain contexts where the customer may immediately respond to the treatment (e.g. click outcome in sponsored ads or website design application) or the response is observed with a delay (e.g. the decision to purchase a product might be observed hours or days after the assignment of treatment), there are many contexts where outcomes are not able to be observed instantaneously and the treatment effects cannot be measured prior to the treatment of subsequent subjects: often, outcomes for all subjects are observed long after the treatment. Take, for example, a blood donation experiment that serves as one of our case studies later

in this paper. Here, treatments can be given weeks (or even earlier) before outcomes are observed. Moreover, outcomes are not observed in sequences, rather they are collected at once on the day of the blood donation. Therefore, all treatments are given before any outcomes can be observed, defeating the purpose and benefit of multi-armed bandit approaches. These are similar scenarios faced in a collage.com experiment (the second case study we discuss in detail later in this paper) where purchases are made over a one-month period. Moreover, almost all interventions designed to move the long-term outcome (e.g., projected customer lifetime value) face certain constraints when using multi-armed bandit methods. While we understand that there are advances made in Multi-armed Bandit research to incorporate delayed feedback (Vernade et al. 2018), we believe that our approach is still useful for those applications and could complement the multi-armed bandit approach to improve firms' practice of randomization experiments. In general, we envision firms developing a cyclical organizational discipline of hypotheses generation using exploratory analytics, judgment, and intuition; the adoption of the scientific method, to validate and test these hypotheses using causal analytics; the use of predictive modeling, to go beyond ATE exploiting heterogeneity; and the use of constrained optimization, for eventual decision making.

## Prescriptive Analytics Framework

We begin by providing notation for establishing the setting of our prescriptive analytics framework. Let $\mathcal{N}$ be a sample of $n$ independent and identically distributed units from the population of interest $\mathcal{P}$, such that sample units are indexed by $i \in \{1, \dots, n\}$, and for each unit we observe characteristics $X_i \in \mathcal{X}$. Let the decision maker have a collection of $J$ treatments $\mathcal{T} = \{T_0, \dots, T_J\}$. Following the potential outcomes framework, we posit for each unit $i$ the existence of potential ex ante unknown outcomes $(B_i(j), C_i(j), U_i(j) \in \mathbb{R})$ which are the respective benefits, costs, and utility $(U_i(j) = B_i(j) - C_i(j))$ that would be realized by assigning unit $i$ to treatment $j$. Let the decision maker's budget be $M$. Let $A_{ij} \in \{0,1\}$ be the decision variable, such that $A_{ij} = 1$ assigns unit $i$ to receive treatment $T_j$ (note that $A_{ij}$ is turned "on" for only one $j$ for each $i$). We elaborate on each of the key aspects of our framework in the following sections. We start by highlighting the growing value of causal analytics in organizations.

### *In Vivo Experimentation*

Two assumptions (preconditions) that are implicit to our framework are that firms can use exploratory analytics to discover new policy levers, and that firms can adopt the randomized control trial to causally estimate the expected effects of different policy levers on benefits and costs. The first assumption, by its exploratory nature, is difficult to explicate precisely or conduct mechanically. Furthermore, in this work, we utilize preexisting large-scale randomized experiments for conducting the empirical analysis. Therefore, given the ex post focus of our work, we do not cover how to develop an initial set of treatment conditions, conduct a randomized experiment with these conditions, explore the experimental results to discover new conditions with potentially larger effects, or repeat this process as would an organization. However, we do highlight that recent advances, including subset-scanning-based approaches for subpopulation discovery (McFowland III et al. 2018; Somanchi et al. 2018), provide promising methods for organizations to engage in such exploratory hypothesis generation. Subset scanning is a concept that originates in anomalous pattern detection (Neill 2012; McFowland III et al. 2013; Neill et al. 2013; Speakman, McFowland III, and Neill 2015), which has been adapted for heterogeneous treatment effects to identify the subpopulations with the most statistically significant treatment effects in randomized experiments (McFowland III et al. 2018) and field studies with multiple treatments (Somanchi et al. 2018). These methods focus solely on identifying subpopulations with sufficient evidence of a treatment effect, which can be characterized as generating treatment hypotheses that are supported by the data. While the idea of the data-generating hypothesis is already prevalent in fields of (bio)informatics (Dopazo and Aloy 2006; Biesecker 2013), there has been a recent call for research in the IS community (Agarwal and Dhar 2014) to take advantage of the power of big data and machine learning to identify phenomena of interest and use the results to investigate causal relationships by exploiting econometric techniques.

The second assumption is consistent with the emerging stream of IS literature that pinpoints the benefits of in vivo large-scale randomized field experiments, with respect to identifying nuanced mechanisms of interest in complex online systems such as social networks and online dating markets (Aral and Walker 2011; Bapna and Umyarov 2015). It is also consistent with the growing recognition of "A/B testing" as a valuable organizational capability to develop. Kohavi and Thomke (2017) describes not only how digital native organizations—e.g., Amazon, Booking.com, Facebook, and Google—each conduct tens of thousands of experiments annually but also how traditional companies—e.g., Walmart, Hertz, and Singapore Airlines—are beginning to rely on experimentation, albeit at a smaller scale. We treat this practice as a baseline capability for organizations and policy makers and develop a framework for increasing the value obtained from such experimentation by using machine learning and optimization to advance current practice.

### Heterogeneous Treatment Effects

With the results of a randomized experiment as input, our framework begins by relying on machine learning to discern heterogeneous treatment effects (HTE). Most approaches for HTE in the literature are built atop the potential outcomes framework, with (as good as) random treatment assignment, enabling causal inference. More precisely, they begin with observing $\mathcal{N}_1$, a sample of $n$ independent and identically distributed units from a population of interest $\mathcal{P}$. The units are indexed by $i \in \{1, \ldots, n\}$, and for each unit there is a binary assignment indicator $W_i \in \{0,1\}$, where $W_i = 0$ indicates assignment to the control group, while $W_i = 1$ indicates assignment to the treatment group. Therefore, there exist two potential outcomes for each unit ($Y(0), Y(1) \in \mathbb{R}$), although only one is realized in practice. Additionally, each unit is described by $X_i \in \mathcal{X}$, a $d$-dimensional vector of covariates, whose support is the set $\mathcal{X}$.

In particular, there is interest in a causal population estimand $\tau$ that is a function of the potential outcome distributions and covariates, and measures the treatment effect

$$\begin{aligned} \tau &= \tau(F_{Y(1)}, F_{Y(0)}, X) \\ &= Div(F_{Y(1)|X}, F_{Y(0)|X}) \end{aligned}$$

where $Div: (F, F') \mapsto \mathbb{R}$ is a general measure of divergence between cumulative distribution functions (CDF) $F$ and $F'$. The most common estimand of interest is the average treatment effect (ATE),

$$\begin{aligned} \tau_{\text{ATE}} &= \int y \, dF_{Y(1)}(y) - \int y \, dF_{Y(0)}(y) \\ &= \mathbb{E}[Y(1) - Y(0)], \end{aligned}$$

which computes the expected difference between the potential outcomes across the population. With the rising interest in HTE, the conditional average treatment effect (CATE)

$$\begin{aligned} \tau_{\text{CATE}}(x) &= \int y \, dF_{Y(1)|X}(y|x) - \int y \, dF_{Y(0)|X}(y|x) \\ &= \mathbb{E}[Y(1) - Y(0)|X = x], \end{aligned} \quad (1)$$

has also garnered much attention, as it considers how the potential outcome distributions vary for each covariate profile. However, the literature does offer alternative estimands (Grimmer et al. 2017), including more general measures of divergence between distribution functions (McFowland III et al. 2018; Somanchi et al. 2018).

The CATE is the typical estimand of interest in the pursuit of HTE using Machine Learning algorithms. More specifically, algorithms attempt to estimate

$$\tilde{\tau}_{CATE}(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \quad (2)$$

However, a challenge arises because $\tilde{\tau}_{\text{CATE}}$ is implicitly a function of both the observed and unobserved potential outcome values. At most one of the potential outcomes is observed for each unit in the sample:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1; \end{cases}$$

therefore, machine learning algorithms cannot be trained directly to estimate $\tilde{\tau}_{\text{CATE}}$. Therefore, the literature commonly makes the additional assumption of unconfoundedness (Rosenbaum and Rubin 1983)

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \mid X_i \; \forall i, \quad (3)$$

because if we let $\gamma(x) = \mathbb{E}[W_i|X_i = x]$—the propensity of receiving treatment for a subject with covariate profile $x$ - (3) implies

$$\mathbb{E}\left[Y_i^{obs}\left(\frac{W_i}{\gamma(x)} - \frac{1 - W_i}{1 - \gamma(x)}\right)|X_i = x\right] = \tilde{\tau}_{CATE}(x). \quad (4)$$

However, in our particular framework, we assume that the data is drawn from a randomized experiment, and therefore can conclude

$$Y_i(0), Y_i(1), X_i \perp\!\!\!\perp W_i \; \forall i. \quad (5)$$

It is clear that Equation (5) is a stronger assumption that implies Equation (3) and therefore Equation (4); more specifically, Equation (4) becomes

$$\mathbb{E}[Y_i|X_i = x, W_i = 1] - \mathbb{E}[Y_i|X_i = x, W_i = 0] = \tilde{\tau}_{CATE}(x). \quad (6)$$

Now that $\tilde{\tau}_{\text{CATE}}$ can be expressed as a function of observed quantities, it can now be estimated empirically from data by

$$\hat{\tau}_{CATE}(x) = \frac{1}{|r(x,1)|}\sum_{i \in r(x,1)} Y_i - \frac{1}{|r(x,0)|}\sum_{i \in r(x,0)} Y_i \quad (7)$$

where $r(x, w) = \{i: X_i = x, W_i = w\}$. We know that theoretically $\hat{\tau}_{\text{CATE}}(x)$ possess good properties in terms of estimating $\tilde{\tau}_{\text{CATE}}(x)$ and subsequently $\tau_{\text{CATE}}(x)$, e.g., $\hat{\tau}_{\text{CATE}}(x)$ is an unbiased and consistent estimator. In practice, this empirical estimator is often substituted for an algorithm from the machine learning literature which has also been demonstrated to exhibit desirable properties theoretically (usually given regularity conditions) and empirically; for example, in this work we use the random forest estimation procedure. Finally, we note that although the HTE literature commonly considers one treatment group, and therefore

defines the estimand and estimators of CATE as parameters by covariate profile $x$, this definition can be extended to account for multiple treatment groups. Essentially, in our previous definitions $W_i$ is a binary indicator, which can be extended to $W_i \in \{0,1,\dots,J\}$ with Equations (1), (2), (7) being respectively redefined as

$$\begin{aligned} \tau_{CATE}(x,j) &= \mathbb{E}[Y(j) - Y(0) | X = x], \\ \tilde{\tau}_{CATE}(x,j) &= \mathbb{E}[Y_i(j) - Y_i(0) | X_i = x], \\ \hat{\tau}_{CATE}(x,j) &= \frac{1}{|r(x,j)|} \sum_{i \in r(x,j)} Y_i - \frac{1}{|r(x,0)|} \sum_{i \in r(x,0)} Y_i . \end{aligned} \quad (8)$$

We describe how organizations can implement and operationalize HTE below.

## Optimal Prescriptions with Heterogeneous Treatment Effects

The final component of our framework utilizes the information gathered from estimating heterogeneous treatment effects in order to inform prescriptions for future subjects. As in the case of estimating HTE, we begin with observing $\mathcal{N}_2$, a sample of $n$ independent and identically distributed units from the population of interest $\mathcal{P}$. Again, the sample units are indexed by $i \in \{1,\dots,n\}$, and for each unit we observe $X_i \in \mathcal{X}$. Note that the sample observed here is different than the sample observed and utilized for estimating HTE ($\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$), although $\mathcal{P}$ and $\mathcal{X}$ must be the same. Furthermore, the decision maker has a collection of $J$ treatments $\mathcal{T} = \{T_0,\dots,T_J\}$, and must decide the value of each binary indicator $A_{ij} \in \{0,1\}$, where $A_{ij} = 1$ assigns unit $i$ to receive treatment $T_j$. Again, following the potential outcomes framework, we posit for each unit $i$ the existence of potential outcomes $(B_i(j), C_i(j), U_i(j) \in \mathbb{R})$ which are the respective benefits, costs, and utility $(U_i(j) = B_i(j) - C_i(j))$ that would be realized by assigning unit $i$ to treatment $T_j$. Therefore, if the decision maker has a budget $M$, her objective is to

$$\begin{aligned} \textit{maximize:} \quad & \sum_{i=1}^{n} \sum_{j=0}^{J} U_i(j) A_{ij} \\ \textit{subject to:} \quad & \sum_{i=1}^{n} \sum_{j=0}^{J} C_i(j) A_{ij} \le M \\ & \sum_{j=1}^{J} A_{ij} \le 1, \forall i \\ & A_{ij} \in \{0,1\}, \qquad \forall i,j. \end{aligned} \quad (9)$$

We label this optimal model in Equation (9) as HTE-OPT, note that its solution is the result of an integer linear program (ILP), and recognize that it is purely theoretical as it is based on unknown (potential outcome) functions. Furthermore, these functions are not directly estimable from data as they rely on all of the potential outcomes, of which at most one can be observed. However, recall the discussion above that we can use machine learning methods to obtain unbiased estimates of potential outcomes if we have data that follows a randomized experiment, as this implies Equation (5). Therefore, we propose substituting the unknown potential outcomes in Equation (9) for the following estimators

$$\begin{aligned} \hat{B}_i(j) &= \hat{\tau}^{B}_{CATE}(X_i,j) + \hat{B}_i(0) | X_i \\ \hat{C}_i(j) &= \hat{\tau}^{C}_{CATE}(X_i,j) + \hat{C}_i(0) | X_i, \end{aligned} \quad (10)$$

which can be computed as in Equation (8). We propose computing the estimators by explicitly modeling the heterogeneity, as the treatment effect is the facet that creates the variability of most interest. Moreover, it has been demonstrated that the variables responsible for the treatment effect heterogeneity are not only qualitatively different from those that capture the response surface for the baseline condition, but they often have relatively weak predictive power (Imai and Ratkovic 2013; Imai and Strauss 2011). Therefore, attempting to estimate the entire function jointly will likely obscure components of the heterogeneity, especially when the true $\tau_{\text{CATE}}$ diverges from the assumptions of the chosen machine learning approach. Even beyond the particular learning approach, there are various forms of modeling for the $\hat{\tau}_{\text{CATE}}$ estimators (e.g., the meta-learners in Künzel et al. 2019), which exhibit different finite sample properties depending on the complexity of the underlying treatment effect, further demonstrating the importance of properly capturing the uniqueness of this heterogeneity (variation). Therefore, our final proposed ILP optimization, which serves as an estimate of the-OPT, is labeltheHTE-EST and is defined as follows:

$$\begin{aligned} \textit{maximize:} \quad & \sum_{i=1}^{n} \sum_{j=0}^{J} \hat{U}_i(j) A_{ij} \\ \textit{subject to:} \quad & \sum_{i=1}^{n} \sum_{j=0}^{J} \hat{C}_i(j) A_{ij} \le M \\ & \sum_{j=1}^{J} A_{ij} \le 1, \forall i \\ & A_{ij} \in \{0,1\}, \qquad \forall i,j. \end{aligned} \quad (11)$$

In order to solve this ILP, we use IBM CPLEX optimizer,[2] an industrial level high-performance mathematical programming solver with state-of-the-art run times.[3]

Having provided the theoretical and mathematical foundations of our framework, we now provide a practical operationalization for the determination of the heterogeneous treatment effects. This is an important aspect of our contribution where we leverage and develop key metrics (serving as tollgates) that determine whether the data generating process within a treatment arm can be properly modeled, and if there exists sufficiently exploitable heterogeneity across treatments. We need both conditions to be satisfied to provide value beyond average treatment effects or uplift modeling.

## Operationalization of Heterogeneous Treatment Effects

Above, we outlined how the four pillars of analytics can be unified to produce a general three-stage prescriptive analytics framework. In this section, we dive deeper into the facet of the framework that is built atop heterogeneous treatment effects. Specifically, we outline how one may exploit HTE given a randomized control trial with a control group and multiple treatment arms.

The building and evaluation of the HTE come from the ability to split the data into training and test datasets. The split percentage can be set by the decision maker. The training set is used to demonstrate the process an organization would take in their attempt to learn the costs and benefits of each intervention. The test set is used to demonstrate how the organization would implement our validation measures—i.e., computing out-of-sample model accuracy metrics and measuring the existence of exploitable heterogeneity across treatment conditions. Furthermore, the test set can also be used later to evaluate the expected utility the organization could obtain on future subjects (from the same population) by following prescriptions derived from the learning. We describe our key steps formally using pseudocode where necessary.

### Model Fitting

The first objective of our process for operationalizing HTE is to fit a statistical model to the training data set, allowing us to develop an understanding of the treatment effects for each intervention. For each treatment intervention, we compare the subjects who experience this treatment to those who experience control, in order to formulate an estimate of the treatment effect. For the ATE approach, this reduces to simply computing the average difference in the outcomes of interest (e.g., firm benefit and firm cost) between the treatment and control groups. However, the prescriptive component of our framework, as well as uplift modeling, is built atop individual-level treatment effects. Therefore, to support its prescriptions, we must build models that will provide individual-level treatment effect estimates. For each outcome of interest, using fivefold cross-validation, we build a model to capture its relationship with provided covariates, under each condition. The collection of these models form an estimate $\widehat{m}(x, j, o)$ of the function which maps from covariate profile $x \in \mathcal{X}$, treatment condition $T_j \in \mathcal{T}$, and outcome of interest $o \in \mathcal{O} = \{\text{cost, benefit}\}$ to the expected realization of the outcome. We note that any class of statistical learning algorithm can be used derive $\widehat{m}$; we select random forest for demonstrative purposes. Algorithm 1 (below) describes the process of building machine learning models that represent the cost and benefit outcomes of the control group and the treatments groups.

### Causal Inference and Treatment Effects

There are two ways to estimate outcomes in each treatment condition. The first approach is to model the baseline outcomes and treatment effect jointly—i.e. $\hat{B}_i(j) = \widehat{m}(x_i, j, B)$—whereas the second approach, which we select in this work, explicitly separates (from the baseline outcome) and estimates the treatment effect heterogeneity—$\hat{B}_i(j) = \hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$—as shown in Equation (10). Above, we provide commentary from the literature explaining why the latter approach can be preferable for accurately modeling heterogeneity, irrespective of the chosen estimation algorithm (Imai and Ratkovic 2013; Imai and Strauss 2011). Moreover, it has been argued from a practical and empirical perspective that utilizing single tree methods—e.g., causal tree or transformed outcome tree (Athey and Imbens 2016)—to capture treatment effects can lead to estimations that are partly fit on the error terms, which eventually find treatment effects even when they are not present (Berry et al. 2016). Therefore, (Berry et al. 2016) proposes a two-stage process (consistent with the $\hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$ approach to estimation), which is demonstrated to have better performance than causal tree (Athey and Imbens 2016).

[2] https://www.ibm.com/analytics/cplex-optimizer

[3] Empirically, we find that this solver can optimize for over a million users in under five minutes and for ten million users in eight hours.

**input** : training data ($\mathcal{N}_1$), treatment condition set ($\mathcal{T}$)
/* $\mathcal{N}_1$ is composed by benefit outcomes ($\vec{B}$), cost outcome ($\vec{C}$), covariates ($\boldsymbol{X}$), treatment condition ($\vec{W}$), and set of Outcomes $o \in \mathcal{O} = \{\text{cost, benefit}\}$ *
**for** $T_j \; in \; \mathcal{T}$ **do**
  $D_j \leftarrow \{(B_i, C_i, X_i) | W_i = j\}$; // separate training data by treatment conditio
  **for** $o \; in \; \mathcal{O}$ **do**
    Use $D_j$ to learn $\hat{m}(x, j, o)$–via cross validation, optimizing MAE, for hyper parameter tuning–which estimates $m : (x, j, o) \mapsto \mathbb{R}$ the true potential outcomes function;
  **end**
**end**
**for** $T_j \; in \; \mathcal{T} \setminus \{T_0\}$ **do**
  **for** $o \; in \; \mathcal{O}$ **do**
    **for** $i \; in \; 1 \ldots |\mathcal{N}_1|$ **do**
      Compute $\hat{\tau}_i(j, o) = \hat{m}(x_i, j, o) - \hat{m}(x_i, 0, o)$, an estimate for the individual-level treatment effect in the outcome of interest;
    **end**
    Use $\{\hat{\tau}_i(j, o)\}_{i=1 \ldots |\mathcal{N}_1|}$ to learn $\hat{\tau}_{\text{CATE}}(x, j, o)$–via cross validation, optimizing MAE, for hyper parameter tuning–which estimates $\tau_{\text{CATE}}(x, j, o)$;
  **end**
**end**
**output:** the functions $\hat{m}(x, j, o)$ and $\hat{\tau}_{\text{CATE}}(x, j, o)$

**Algorithm 1. HTE Algorithm**

For our empirical analysis, we follow the suggestions from the prior literature and therefore decided to model treatment effect heterogeneity separately from the baseline outcome ($\hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$). We utilize an approach similar to that described in (Berry et al. 2016) to model the individual level treatment effects.[4] Assuming that the model accuracy metric computed on the test data indicates that the models ($\widehat{m}$) are well estimated, we therefore can trust their predictions. Moreover, armed with $\widehat{m}$, we can (in a sense) overcome the "the fundamental problem of causal inference"—i.e., a subject is observed in at most one condition—because $\widehat{m}$ provides estimates for each subject's outcomes across all conditions. Furthermore, because of randomization, these estimates are unbiased. Therefore, for each subject in the training data, we predict their outcome under each condition and obtain subsequent estimates of individual level treatment effects: $\hat{\tau}_i(j, o) = \widehat{m}(x_i, j, o) - \widehat{m}(x_i, 0, o) \; \forall i, j, o$. Note that $\hat{\tau}_i(j, o)$ is an unbiased and consistent estimate of $\tau_i(j, o)$, the true, unobservable, individual-level treatment effect for subject $i$, given treatment condition $T_j$ and outcome $o$. Finally, using 5-fold cross validation, from the $\hat{\tau}_i(j, o)$ we learn a model to capture the heterogeneity in individual-level treatment effects, for each outcome of interest and treatment condition. We note that any class of statistical learning algorithm can be used in this context as well; we select decision trees for demonstrative purposes. The collection of these models form Equation (8): an estimate $\hat{\tau}_{\text{CATE}}(x, j, o)$ of the function which maps from covariate profile $x$ and treatment condition $T_j$ to the expected treatment effect for outcome $o$. This process is in the same category of well-documented approaches in machine learning, where a strong—but potentially difficult to interpret—learner (e.g., random forest) is utilized to best capture the underlying complex function, and then a more interpretable—but potentially weaker—learner (e.g., decision tree) is utilized to summarize the complex learner's decision function (Domingos 1997). Evaluation of the predictive modeling exercise described above is outlined in Algorithm 1. The ultimate goal is to ensure that our models are rich enough to capture the (within-heterogeneity of the) data generation process for all treatment conditions.

### *Model Accuracy*

In order to evaluate the accuracy of our chosen estimated functions (i.e., $\widehat{m}$ and $\hat{\tau}_{\text{CATE}}$) we turn to the test data subjects ($\mathcal{N}_2$). More specifically, we measure *mean absolute error* (MAE) which is a standard metric in model fitting to measure

[4] For the sake of completeness, in the Comparing HTE Learners section, we also conducted our analysis using causal forest (Wager and Athey 2018), which is an ensemble with a causal trees base learner. Therefore, we use it to represent models that estimate the baseline outcomes and treatment effect jointly.

the similarity between the estimated and observed quantities of interests, e.g., cost, benefits, and therefore utilities. This is computed across all subjects, comparing the observed quantity of interest to the estimate of this quantity, given the observed treatment condition for the subject. In our context, values of mean absolute error (on the test data) demonstrate to what degree the model is able to capture the true data-generating process of the underlying treatment arms. Furthermore, a small value of MAE is desirable, as it implies that the models are able to capture what true utility a subject would experience under any condition. It is still an open question what a sufficiently small value of MAE is on a given dataset to signify a sufficiently high quality of fit. We treat this threshold value as an external parameter, leaving it to the researcher and the domain expert to determine whether the MAE signals a sufficient quality of fit for their purposes. It is possible that researchers will be unable to find a modeling technique that can produce a fit of sufficient quality. If so, this provides evidence that continuing with the subsequent steps of our process is likely to not be fruitful; there is little practical value in attempting to perform inference (and eventually generate prescriptions) with highly error-prone estimations.

While we use the standard MAE metric for this purpose, there is nuance (detailed in Algorithm 2 below) around how the outcome predictions incorporate the treatment effects for the individuals in their treatment arms. After computing MAE, we can use the user-defined cutoff (or parametric threshold) and proceed further if the MAE values are sufficiently small.

## Observed Utility Rank Condition (OUR) ■

Following from above, we were able to perform causal inference across the entire response surface, modeling the treatment effect across the space of covariate profiles. We now aim to evaluate whether, for individuals, there exists heterogeneity in the effect across treatments. Exploiting such heterogeneity in the effects of the treatments is critical for future prescriptions. Again, we turn to the test data subjects, and propose a new metric to compute: *observed utility rank* (OUR). More specifically, for each subject, we compute and rank (within subject) the expected utility under each of the treatment conditions $T_j$ $(j = 1, \dots, J)$. We then partition the test data into $J$ groups ($P_j$), each capturing subjects who received their (estimated) $j^{th}$ best treatment condition, as measured by individual utility. The value of this measure is built on the fact that the test data is also a random sample from the population of interest, devoid of selection bias, and that the models have captured the data generating process of the treatment conditions well. Given these conditions, the individual level estimation and rank of utility indicate the degree to which there are differences between the various treatment conditions for a given individual. Furthermore, the observed average utility of each partition $P_j$ is an unbiased estimate of the expected utility from placing a subject in their $j^{th}$ best condition. Therefore, if there exists exploitable heterogeneity in the effect across treatments, and the ability of the models to capture it with sufficient accuracy, we expect that the utility of partitions $P_j$ will be monotonically decreasing with $j$, and the differences between the partitions will be significantly different. In that case, we conclude that the OUR metric is satisfied. If this does not occur—i.e., there appears to be no significant improvement in utility (on average) by assigning a subject to an individually preferable condition—this likely indicates one or both of two possibilities. First, the accuracy of models that capture the outcomes of interest is poor. Second, there is not sufficient heterogeneity to exploit for prescriptions. In this case, the practitioner should strive to build models that best capture the outcome distribution. If the model with the best MAE still fails to satisfy the OUR metric, it implies one or both of the conditions have likely occurred and progress should not continue.

Essentially, OUR captures the relative strength of heterogeneity present in the data and the model's ability to capture it. If the heterogeneity is small but the models are accurate enough to capture it, OUR is satisfied. If the heterogeneity is sufficiently large, even if the models are less accurate, OUR is still satisfied because the models can still provide a proper ordering of the treatment.

This process of measuring OUR is described in Algorithm 3 (below). We proceed to empirically validate our framework using two different real-world decision-making scenarios, one successful in that it succeeds at each step and the other that fails the OUR stage. We consider it informative to demonstrate and discuss both scenarios, as we believe there is value in learning from experiments that both succeed and fail to be exploitable by our framework.

In the Discussion and Conclusion section, we acknowledge that there are other procedures for model validation like the one mentioned in Hitsch and Misra (2018), in which a follow-up experiment is conducted to validate the performance of existing models.

### *Empirical Analysis*

In this section, we empirically demonstrate the utility of our proposed prescriptive analytics framework for determining which subjects to target with which interventions, in order to optimize organizational goals.

**input** : test data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), potential outcomes function ($\hat{m}$), treatment effects function ($\hat{\tau}_{\text{CATE}}$)

**for** $i \, in \, 1 \ldots |\mathcal{N}_2|$ **do**
  $\hat{B}_i(0) \leftarrow \hat{m}(x_i, 0, b)$, $\hat{C}_i(0) \leftarrow \hat{m}(x_i, 0, c)$;
  **for** $T_j \, in \, \mathcal{T} \setminus \{T_0\}$ **do**
    $\hat{B}_i(j) \leftarrow \hat{B}_i(0) + \hat{\tau}_{\text{CATE}}(x_i, j, b)$, $\hat{C}_i(j) \leftarrow \hat{C}_i(0) + \hat{\tau}_{\text{CATE}}(x_i, j, c)$;
  **end**
**end**

$$\text{MAE}_b = \frac{\sum_{i=1}^{|\mathcal{N}_2|} |B_i(W_i) - \hat{B}_i(W_i)|}{|\mathcal{N}_2|} \qquad \text{MAE}_c = \frac{\sum_{i=1}^{|\mathcal{N}_2|} |C_i(W_i) - \hat{C}_i(W_i)|}{|\mathcal{N}_2|}$$

/* MAE captures the error in the observed condition ($W_i$), across all $i$ */

**output:** $\hat{B}, \hat{C}, \text{MAE}_b, \text{MAE}_c$

**Algorithm 2. Validation of HTE Models (and within-heterogeneity)**

**input** : test data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), benefit values ($\hat{B}$), cost values ($\hat{C}$), treatment condition ($W_i$)

$\text{P} \leftarrow (\{\emptyset\}_1, \{\emptyset\}_2, \ldots, \{\emptyset\}_{|\mathcal{T}|})$ ; // vector of $|\mathcal{T}|$-many empty set placeholders
**for** $i \, in \, 1 \ldots |\mathcal{N}_2|$ **do**
  $R_i \leftarrow |\mathcal{T}|$ ; // default to the lowest observed rank
  **for** $T_j \, in \, \mathcal{T}$ **do**
    $\hat{U}_i(j) \leftarrow \hat{B}_i(j) - \hat{C}_i(j)$ ; // compute estimate of utility
    $R_i \leftarrow R_i - \mathbb{1}\{\hat{U}_i(W_i) \geq \hat{U}_i(j)\}$ ; // update rank of observed condition
  **end**
  $P_{R_i} \leftarrow P_{R_i} \cup \{i\}$ ; // add $i$ to index $R_i$ of P
**end**
**for** $j \, in \, 1 \ldots |\mathcal{T}|$ **do**
  $\text{OUR}_j \leftarrow \frac{1}{|P_j|} \sum_{i \in P_j} \hat{U}_i(j)$ ; // utility if observed condition is rank $j$
**end**
**output:** OUR

**Algorithm 3. Measuring the Observed Utility Rank (OUR)**

We use data from two different randomized field experiments—one from public policy in the context of stimulating blood donations, and one from the marketing domain in the context of referral marketing—to provide a sense of the generality of our framework. These experiments serve as case studies of our frameworks' ability to identify the existence of exploitable heterogeneity and demonstrate its performance in real-world decision-making.

The first experiment was conducted in collaboration with a major blood bank in China and investigates the impact of different mobile messaging interventions in motivating blood donations (Sun et al. 2019), whereas the second experiment was conducted on the online platform Collage.com and investigates the impact of various call-to-actions on activating referrals and subsequent purchases (Jung et al. 2020). In both experiments, the organizations are budget-constrained in their attempt to maximize utility, namely social welfare and profit, respectively. Furthermore, each organizational decision maker is facing ex ante unknown costs and benefits from the multiple policies they can deploy. Therefore, these experimental settings present realistic case studies in which we can explore and evaluate the ultimate effectiveness of prescriptive strategies.

**Table 1. Example Set of Benefit, Cost, and Utility Matrices Where the Rows Are Future (or Test Data) Subjects and the Columns Correspond to Four Different Treatment Conditions**

| | Benefit matrix | | | | Cost matrix | | | | Utility matrix | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Subject** | B0 | B1 | B2 | B3 | C0 | C1 | C2 | C3 | U0 | U1 | U2 | U3 |
| **1** | 26 | 120 | 1 | 100 | 3 | 92 | 0 | 90 | 23 | 28 | 1 | 10 |
| **2** | 16 | 24 | 0 | 15 | 10 | 12 | 0 | 20 | 6 | 12 | 0 | -5 |
| **3** | 5 | 45 | 15 | 16 | 5 | 25 | 11 | 1 | 0 | 20 | 4 | 15 |
| **4** | 0 | 15 | 100 | 55 | 9 | 5 | 50 | 10 | -9 | 10 | 50 | 45 |

To empirically validate our framework, we benchmarked its HTE-EST procedure from Equation (11) against two other prescriptive approaches within the causal inference paradigm that currently dominates practice. These are the average treatment effect (ATE), which assigns all future subjects to the policy that has the highest estimated population average treatment effect (increase in utility), and uplift modeling (UM), which assigns each future unit to the condition estimated to provide the largest individual increase in utility. We show below that, under the exploitable conditions of within- and between-treatment heterogeneity of costs and benefit, the prescriptions provided by traditional approaches fall short. They are either too nonspecific or too myopic and therefore fail to capture attainable utility. In contrast, our analytics framework amalgamates randomized experiments, causal inference, machine learning, and optimization, in order to overcome these limitations.

### *Comparing Prescriptions*

In order to compare the prescriptions generated by HTE-EST—built atop integer linear programming (ILP)—to the traditional approaches—i.e., ATE, UM—we again turn back to our test data, in conjunction with the $\hat{\tau}_{\text{CATE}}$ and $\hat{m}$. More specifically, we construct matrices—where rows are the test data subjects and columns are treatment conditions—of expected benefit and cost (see Table 1). We note that although we utilize estimates of the expected value in our matrices, the error in these estimates is not taken into account. Therefore, in the robustness checks, we compare the prescriptive results under various procedures for propagating the error. Each of the prescriptive procedures are given these matrices and subsequently provide a set of subject assignments, given organizational (budget) constraints. We then compare the amount of utility that each of the methods' prescriptions can obtain subject to the constraints. For reference, in our empirical evaluations on the datasets described below, we depict the mean and variance of the utility each method obtains across thirty random partitions of training and test data.

It is also important to note that since each prescriptive technique is provided with the same estimates of individual cost and benefit, their comparative performance is still informative concerning their relative ability, even if the true values are not perfectly estimated. Notwithstanding, we expect our framework's HTE-EST to yield better prescriptions given a set of matrices. Let us take Table 1 as an illustrative example, with a budget constraint of 50. We can see that ATE would prescribe putting each subject into Condition 1, as it has the highest average utility, leading to 42 in utility (84 in benefit) at a cost 42; ATE would be unable to service Subject 1 because its cost for Treatment 1 would exceed the budget constraint. UM would provide additional flexibility, as compared to ATE, by selecting the best individual condition for each subject. Therefore, it would prescribe conditions $(1,1,1,2)$ for the subjects respectively, leading to 50 in utility (100 in benefit) at a cost 50, by serving only Subject 4. HTE-EST, would be the most flexible, as it can assign any condition to any subject in an attempt to maximize total utility. Therefore, it would prescribe conditions $(0,1,1,3)$ for the subjects respectively, leading to 100 in utility (150 in benefit) at a cost 50. Essentially, ATE's prescriptions would be based on which column has the highest average utility, while UM's prescriptions would be based on which column, for each row separately, has the highest utility.

From our simple example, it becomes evident that ATE and UM are suboptimal, with ATE being overly general and UM being overly myopic. These methods fail to consider the cost required to achieve the utilities that drive their prescriptions. HTE-EST, however, offers prescriptions that can select any treatment for any subject, taking into consideration the budget constraint and cost associated with the utility being gained. As a result, HTE-EST is able to avoid ATE's overly general prescriptions—with individual prescriptions—and UM's overly myopic prescriptions—by realizing that Subject 1 and 2's individually high utility conditions come at too high a cost. Unlike UM, HTE-EST can utilize the budget saved by placing Subject 4 and Subject 1 into an individually suboptimal condition to obtain additional benefit from being able to serve other subjects, leading to more globally optimal prescriptions.

**Table 2. Mean Absolute Error (MAE) of the Random Forests Models that Estimate the True Data Generating Process of the Treatment Conditions in the Blood Donation Experiment across Subjects**

| Outcome | MAE | St. dev |
|---|---|---|
| Benefit | 8.22 | 0.07 |
| Cost | 0.10 | 0.01 |
| Utility | 7.26 | 0.07 |

**Table 3. *P*-Values for Pairwise Comparison of Average Utility Actually Experienced by the Blood Bank for Different Ranks of Treatment Allocations**

| Rank of treatment allocation | First | Second |
|---|---|---|
| Second | 0.0025 | - |
| Third | < 2e-16 | 8e-11 |

**Table 4. Comparison between the Allocation of Subjects under Their Actual (Random) Assignment and the Allocation under the (Estimated) Individually Optimal Assignment and ATE**

| Allocation | Control | Treatment 1 | Treatment 2 |
|---|---|---|---|
| Actual (random) | 2484 | 3888 | 8022 |
| Individually optimal | 4285 | 2553 | 7556 |
| Average treatment effect | 0 | 0 | 14394 |

### *Blood Donation Case Study*

The first field experiment we use to illustrate our framework for optimal utilization of heterogeneous treatment effects is concerned with motivating blood donation. We use the experiment to illustrate how our methods may be extended to nonprofit applications and accommodate considerations of policy makers on the nonmonetary utility of individuals. The research context, experiment design, and summary statistics are detailed in Sections 4 and 5 of Sun et al. (2019).

In collaboration with a centralized blood bank in a major city in China, the researchers conducted a large, randomized field experiment to test the effectiveness of different interventions in motivating blood donations from subjects and their friends. Specifically, 80,000 eligible subjects were chosen from the pool of past donors to the blood bank and randomly assigned into several treatment conditions. The first condition is a control group with 14,000 subjects. For the remaining treatment groups, the researchers send a mobile message and vary its content across groups. The message content explores two treatments to overcome the hurdle of blood donation. The first message condition is a behavioral intervention (with 22,000 subjects) in the form of a message that reminds a potential donor to come to donate or to donate together with friend(s). The second treatment (with 44,000 subjects) informs the potential donor they will receive an economic reward for donation.[5] The details of the mobile messages for the test groups using the behavioral or economic interventions, the choice of sample, the time horizon, the variable collected, as well as randomization check, the summary statistics, and the main results for the experiment can be found in Sun et al. (2019), Sections 4 and 5. This original study of the data is concerned with the ATE, i.e., identifying the treatment condition that would motivate the most donations from the subjects. For the purpose of our study, we directly use this field experiment data and further augment it with rich archival data, including demographics (age, gender, education, occupation, marriage status, resident status, and health indicators) and donation history (across 10 years). We follow the practice of the blood bank to define the benefits and costs associated with each donation. Specifically, the benefit is determined by the volume of each donation (1, 1.5, or 2 units * 220RMB/unit). Similarly, the cost is associated with the reward given out to each donor (30RMB gift for 1 unit, 40RMB for 1.5 units, and 50RMB for 2 units). An additional cost of 1RMB is also incurred to send each mobile message, which is therefore not applicable to the control treatment group. We focus on whether and how the blood bank may leverage the heterogeneity in the treatment effect and design optimal policy at an individual level, as well as compare the performance across different prescriptive policies. The setting features a low response rate in the experiment (about 1%) and the potential for rich heterogeneity across subjects given the large number of covariates available. Such a setting mimics the characteristics of many nonprofit and for-profit practical applications, such as charitable donation solicitation and digital advertising.

[5] While only one condition contained information of the economic reward, all eventual donors, regardless of treatment condition, receive the same economic reward for donation.

Table 2 shows the mean absolute error (MAE) from the application of Algorithms 1 and then 2 of our framework for each treatment condition and outcome of interest. As desired, we observe that the errors from our models are quite small on average and tend to vary closely around their small error values. Figure 2 depicts the observed utility rank (OUR) from Algorithm 3 for our models. The x-axis captures the (estimated) rank of the treatment into which a subject is randomly allocated and the y-axis captures the average utility actually experienced by the blood bank. As desired, we observe that there is a significant negative and monotonic relationship between the (estimated) quality of the treatment a subject receives and the expected utility achieved by the blood bank. We conducted pairwise *t*-tests with Bonferroni adjustment between different ranks of treatment allocation and find that they are significantly different. We report the *p*-values in Table 3.

Given that our models appear to accurately capture the expected utility of each treatment for an individual and the OUR metrics signal the existence of exploitable heterogeneity between treatments, we proceed to construct matrices of our outcomes of interest (as in Table 1). These matrices may assist in a deeper exploration of the experiment and help evaluate the effectiveness of the prescriptive techniques. For example, Table 4 presents the number of subjects for whom each condition is their individually optimal allocation, based on the estimated utility, and compares the results with the actually observed random allocation and ATE. When there are no constraints, the optimal policy is simply to put each subject into their individual best conditions—i.e., diverging from the treatment allocations defined in Row 2 of Table 4 will lead to suboptimal utility. Hence, although UM and HTE-EST will provide different prescriptions when facing an organizational constraint, when constraint-free, both will output the same (individually optimal) set of prescriptions. However, recall that ATE allocates all subjects to the treatment condition that is optimal at an aggregate level, which, in this experiment, is Treatment Condition 2, corresponding to the monetary incentives. Furthermore, Table 4 shows that for a large number of subjects, the control group is ideal; the implications are that the prescriptions from ATE lead to an individually suboptimal allocation for many subjects, and the total utility obtained from ATE's prescriptions will be strictly less than HTE-EST (and UM). Therefore, the organization should prefer the allocation suggested by our framework, even with an infinite budget, because ATE will place many subjects in the more costly monetary incentive condition incurring large and unnecessary costs. Such costs originate from two sources: moving some subjects from the control to the costly treatment group without enough gain, and moving some subjects from a treatment group to the control group (or the other treatment group) without properly motivating the subject.

Interestingly, our framework proves even more valuable when the decision maker has certain organizational constraints, such as a budget constraint.[6] In Figures 3 and 4, respectively, we compare two outcomes—the total expected utility realized and number of users served—by each prescriptive targeting strategy for different levels of a budget constraint. With respect to expected utility, our framework's HTE-EST performs significantly better than the other strategies across the range of budget constraints, followed by UM and then ATE. This behavior is rather intuitive, as UM is a greedy approach that simply allocates as many subjects as possible to their individual optimal condition. As a consequence, it first serves subjects with high individual utility; therefore, when the budget it small, it will be exhausted after a few subjects are treated (Figure 4). Conversely, HTE-EST may elect to place subjects in less optimal individual conditions to preserve its precious budget in order to obtain additional total utility. As a result, we find that prescriptions from HTE-EST result in up to 240% and 340% higher utility than UM and ATE, respectively. With a sufficiently large budget, the constraint is no longer relevant, and as described above, HTE-EST and UM would obtain the same (individually optimal) set of prescriptions. However, even with a large budget, HTE-EST still leads to a 19% increase in utility over ATE, suggesting that a great deal of utility is sacrificed by following the simplified guideline from ATE (Row 3, instead of Row 2 of Table 4).

HTE-EST's ability to strategically place a (specific) set of subjects in a secondary condition enables it to capture a significantly larger number of subjects that can obtain additional utility from the treatment. Moreover, HTE-EST is also able to identify subjects for whom the control condition is optimal—i.e., no treatment will affect their ultimate decision to (not) donate. Figure 4 shows that these two facets together make HTE-EST fairly conservative with (and effective at) how many users to target with treatments. Such selectivity can be beneficial, as we know that following a prescriptive strategy like ATE and indiscriminately targeting users (with uninteresting treatments) can have negative consequences (Ghose et al., 2017).

In summary, our framework presents an ideal prescriptive strategy by combining machine learning, causal inference, and optimization to exploit heterogeneity and parsimoniously identify precisely the right subjects to target in order to maximize the number of blood donations received, at any level of budget constraint.

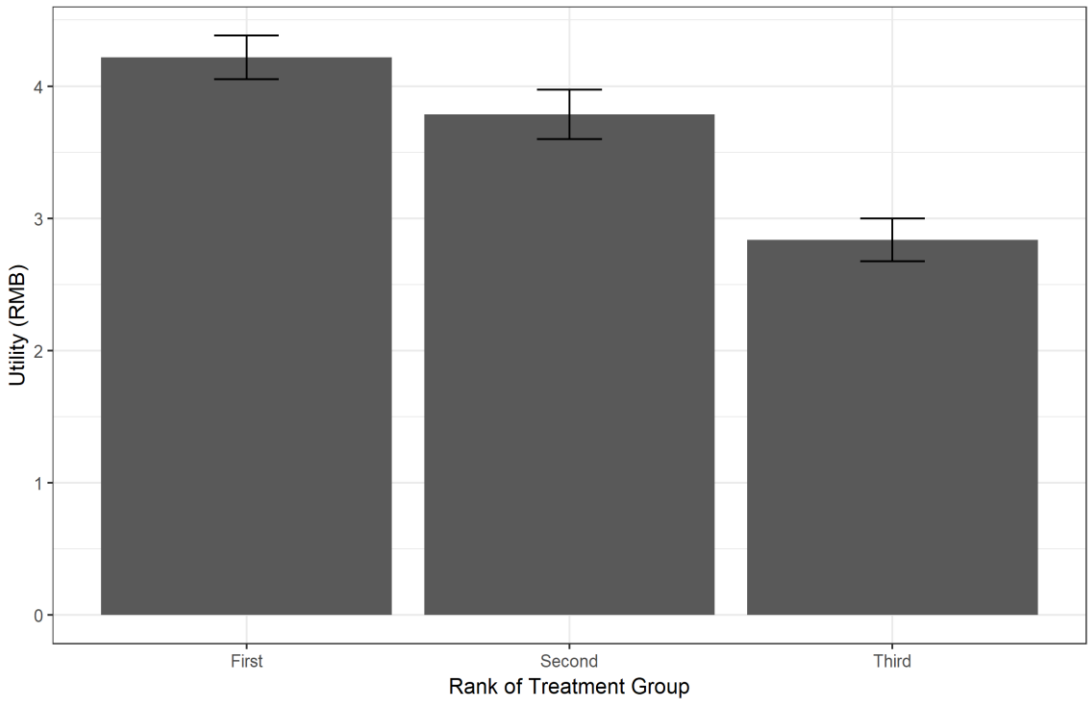
[6] In our context, blood banks often face a strict budget constraint because of a rigid financial planning process at the beginning of each quarter.

**Table 5. Mean Absolute Error (MAE) of the Random Forest Models That Estimate the True Data Generating Process of the Treatment Conditions, in the Referral Marketing Experiment, across Subjects**

| Outcome | MAE | St. dev. |
|---|---|---|
| Benefit | 7.72 | 0.006 |
| Cost | 7.81 | 0.023 |
| Utility | 0.38 | 0.044 |

**Table 6. Accuracy Metrics (MAE, MSE, ME) of the Models that Estimate the True Data Generating Process of the Treatment Conditions by Two Different Learners across Subjects**

| | MAE | | MSE | | ME | |
|---|---|---|---|---|---|---|
| **Outcome** | **Causal forest** | **Berry 2S** | **Causal forest** | **Berry 2S** | **Causal forest** | **Berry 2S** |
| **Benefit** | 7.77 | 8.22 | 1689.65 | 1687.14 | 0.43 | -0.12 |
| **Cost** | 0.92 | 0.97 | 23.34 | 23.28 | 0.05 | -0.01 |
| **Utility** | 6.86 | 7.26 | 1317.16 | 1315.55 | 0.39 | -0.11 |




**Note:** Subjects generate significantly higher utility on average when placed in more optimal conditions as estimated by our models

**Figure 2. Observed Utility Ranking (OUR) Graphically, Demonstrating the Existence of Heterogeneity across Treatment Conditions for Individuals (or Subpopulations).**

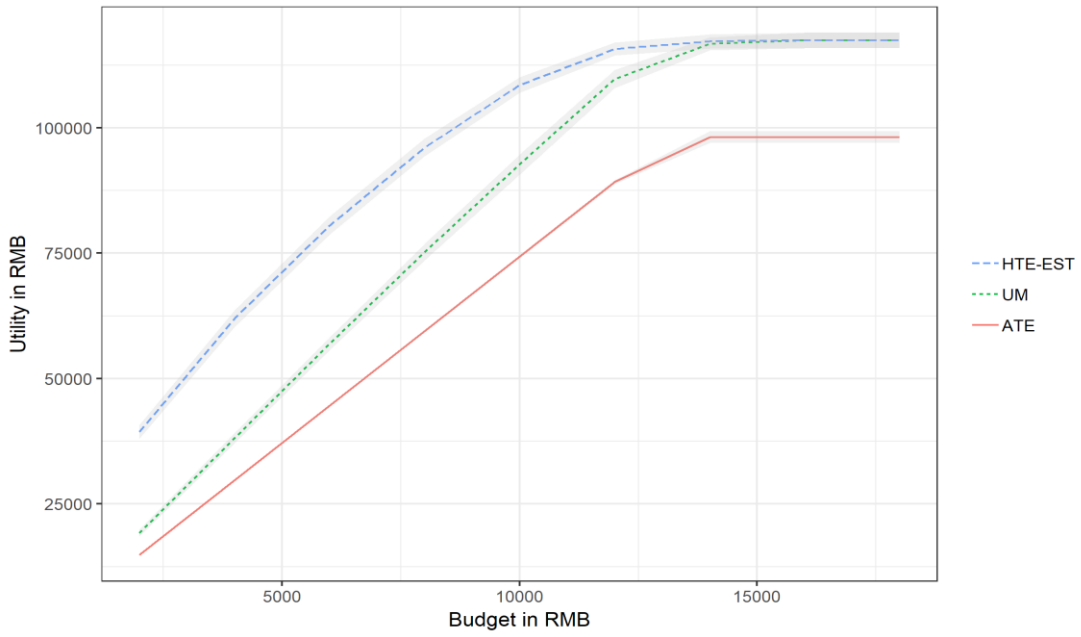



**Figure 3. Expected Total Utility Generated by Each Prescriptive Method over a Range of Budget Constraints**

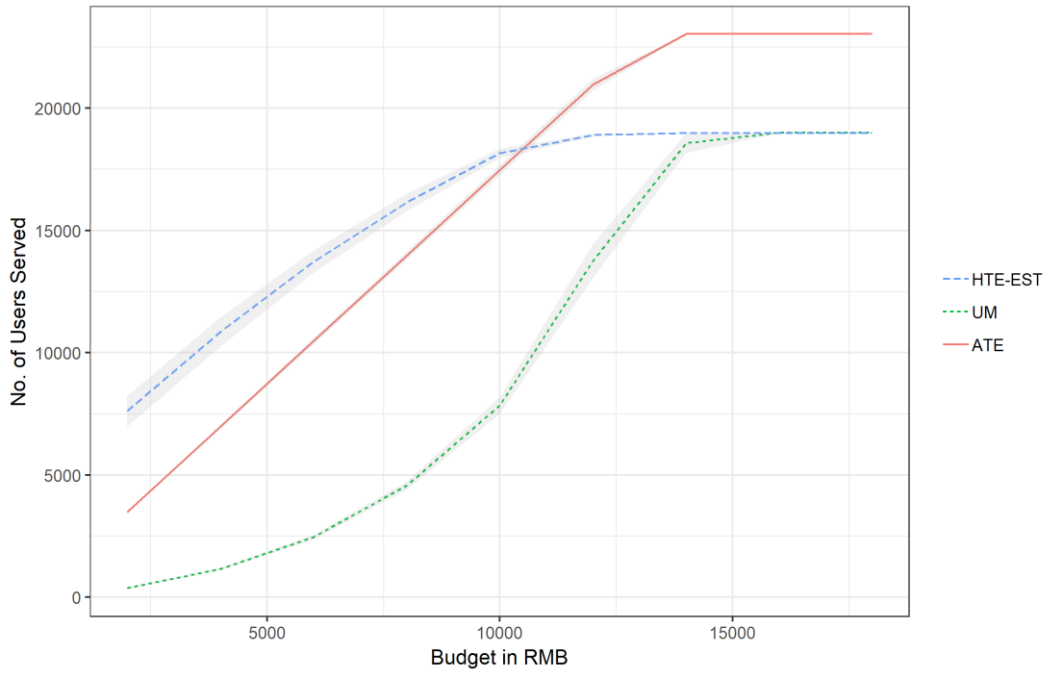



**Figure 4. Expected Number of Subjects Targeted with a Treatment, by Each Prescriptive Method, over a Range of Budget Constraints**

### Referral Marketing Case Study

The second field experiment we use to illustrate our method on optimal utilization of heterogeneous treatment effects is testing the optimal framing (altruistic, selfish, equity) that a firm can use to motivate product referrals. We use the example to illustrate how a profit-maximizing firm may benefit from our method. The research context and experiment design are detailed in Jung et al. (2020). Specifically, the randomized field experiment targets existing customers of a large U.S.-based online platform called Collage.com. On this platform, users can design a collage by uploading photos and customizing the layout with the proprietary software tools. Once a user creates the layout, the user can purchase various types of customized printed products, such as blankets, photo books, canvases, etc. A large number of customers purchase a variety of products from the platform every day (with $22 million in revenue for 2015). The treatments in the experiment manipulate the framing, i.e., solely focus on varying the words and emphasizing certain phrases of the call-to-action while keeping all other aspects of the incentive and messaging constant across groups. The experiment offers both the sender of the referral and the recipient a free product voucher that comes with free shipping and has no expiration date ($25 worth in listing value), which should appeal to many types of users.

The randomized field experiment design allows clean identification of the causal effect of the framing of the calls to action on customers' referral decisions: whether they share, to what extent they share, as well as on their induced referral outcomes, as measured by the number of successful referrals. Specifically, in the experiment, the researchers randomly assigned 100,000 customers who have made purchases on the platform in the past into four test groups (10,000 in control and 30,000 in each of the three treatment groups) and emailed each group with different calls to action. The data on customers' referral behaviors and outcomes were collected within a five-week window following the experiment. Based on extensive discussions with the CEO and the marketing team of the partner company, we arrived at the following benefits and costs associated with each referral and voucher redemption. Specifically, the benefits of a referral come from three parts: the impression value to the recipient(s) at $0.01/referral, the user acquisition value if the recipient registered with Collage.com at $3/registration, and the transaction value if the recipient(s) actually made purchases, calculated based on the actual transaction amount. Conversely, the cost of a referral is associated with the redemption of the free product voucher by the sender, recipient(s), or both. Each redemption incurs an average cost of $6. It is critical to note that relatively few customers who sent out or received a referral redeemed the free voucher. This serves as the source of heterogeneity on the cost side, with the benefit-side heterogeneity resulting from the differential response to the promotion.

For the purpose of our study, we directly use the field experiment data and augment it with rich archival data, including product characteristics, individual characteristics, customers' past purchases and their net promoter scores (NPS).[7] The data from the large randomized experiments and the archival data allows us to identify the causal effect of different calls to action and explore the heterogeneity underlying the treatment effect. As mentioned earlier, the key challenge is that ex ante the marketer does not know whether any of these calls to actions have a causal impact on profitability, let alone know whether the impact of the three different arms may be heterogeneous as a function of existing customer "data"—i.e., characteristics and behaviors.

We begin with Table 5, which shows the MAE of our random forest models for each treatment condition and outcome of interest. As desired, we observe across all combinations that the errors from our models are quite small on average and tend to vary closely around their small error values. Furthermore, Figure 5 depicts the observed utility rank (OUR) for our models, where the x-axis captures the (estimated) rank of the treatment a subject was randomly allocated into and the y-axis captures the average utility actually experienced by subjects. In this case, we fail to observe a negative or monotonic relationship between the (estimated) quality of the treatment a subject received and the expected utility achieved by the subject. More specifically, there is no observable significant difference in the expected utility across the treatment conditions. Given that Table 5 demonstrates a rather small amount of prediction error in our models, we conclude that there is not a sufficient amount of individual-level heterogeneity across treatment conditions present to exploit for optimization. As a result, we see no benefit in continuing to the optimization stage. We do speculate that the lack of between-treatment heterogeneity may be driven by three factors: (1) the drastic difference between the magnitude for the benefit (typically on the order of cents) and that of cost (on the order of dollars)[8]; (2) the inherent challenge in predicting the benefit measure, as revealed by the large variance in the summary statistics associated with the total spend and total discount measures given in the Appendix; and possibly (3) the lack of useful and relevant individual-level features.

[7] The net promoter score is an index used to measure loyalty and customer satisfaction, as the willingness to recommend a company's products or services to others.

[8] To no avail, we attempted a variety of different machine learning models as well as up and down sampling to determine whether the lack of heterogeneity across conditions was a consequence of implicit modeling assumptions or imbalance.

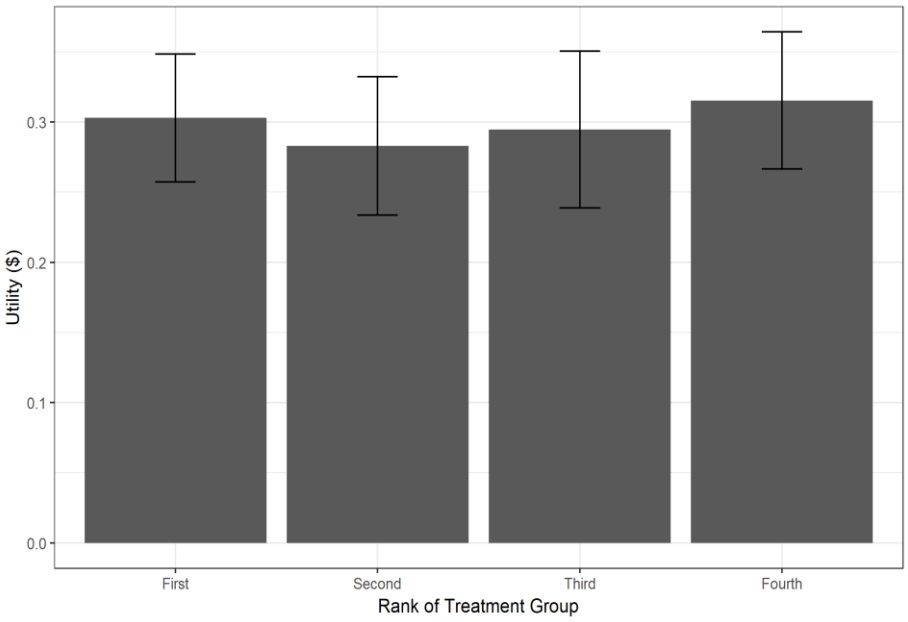



**Figure 5. Mean Absolute Error (MAE) of the Random Forest Models That Estimate the True Data-Generating Process of the Treatment Conditions in the Referral Marketing Experiment across Subjects**

We note that the lack of significant individual heterogeneity across treatments provides no commentary on the existence of a (heterogeneous) treatment effect across the whole sample. Our three-stage prescriptive analytic framework allows us to identify the potential challenges at an early stage in the analytic cycle and save the efforts and expense on further optimization. It could be useful to the organization in terms of future data collection efforts on customer features, which could be added to the framework and the OUR condition to check for between-treatment heterogeneity.

## Robustness Checks

### *Error Simulation Study*

In the results above, we use our models' estimates of cost and benefit to construct the matrices necessary to carry out the prescription generation. This procedure assumes that these estimates are correct and although we know that the estimates from our models are consistent (under appropriate regularity conditions), they (as all estimates) have a degree of error. Therefore, we conducted a simulation analysis of the prescriptions generated from our process (Figure 6), in the attempt to (more properly) propagate the error into our prescriptions. Specifically, we took our final tree models, recognizing that when providing the estimate for cost and benefit for a subject, it is common to use the mean of the tree's leaf the subject falls into. We now consider two additional ways to model the values in the leaves of our trees, which are used to generate these estimates in our matrix of benefits and costs: "Gauss," where a Gaussian distribution is fit to the (training) data points that reside in a tree's leaves, and "Emp," where we utilize the empirical distribution of the (training) data points that reside in the leaves of the tree. Therefore, when we have a subject for whom we want to estimate benefit and cost values, we locate their leaf in the tree and draw a value from the Gaussian distribution (Gauss) or empirical distribution (Emp) at this leaf. This allows the degree of uncertainty that exists to propagate into our matrices and eventual optimization.[9]

Figure 6 shows the results for each of the three prescriptive analytics allocation approaches (ATE, HTE-EST, and UM) for the two error propagation approaches (Emp, Gauss), along with our original mean-based approach for reference. We calculate confidence intervals by splitting the entire dataset into training and test using a different seed ten times.

[9] There are other ways of introducing errors through the global perturbation of nodes. We believe the qualitative pattern of the results is likely similar to our simulation below.

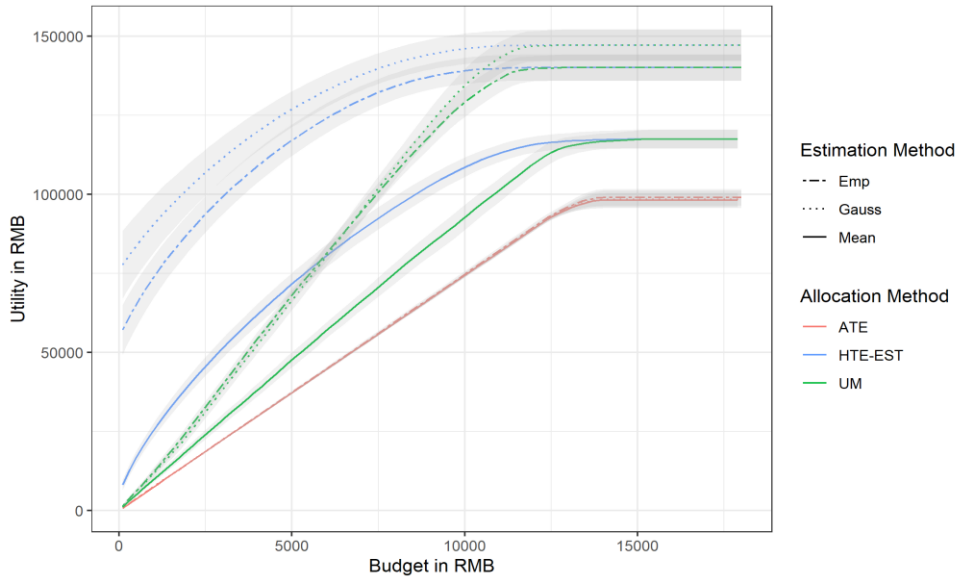



**Figure 6. Expected Total Utility Generated by Each Allocation Method for Different Estimation Procedures over a Range of Budget Constraints**

From this graph, we can make three observations about the final utility derived. First, regardless of which error-propagation procedure we use, the ordering of the prescriptive methods remains constant: HTE-EST, UM, ATE. Intuitively, if we consider the set of potential allocations under each of the methods, ATE is a subset of UM, which is itself a subset of HTE-EST; therefore, HTE-EST should always dominate in utility, followed by UM, and ATE.

Second, within a given allocation approach, the ordering of propagation approaches remains constant: Gauss, Emp, Mean. Intuitively, by using the mean, as we did originally, when the prescriptive algorithms attempt to optimize allocations, they are forced to consider all customers who fall into the same leaf as providing the same cost and benefit. Essentially, units in the same leaf are exchangeable deterministically from an estimated utility point of view, as their estimated values of interest have no sampling variation. Therefore, even though (technically) the optimization of allocations occurs at the unit level, the values used to determine these allocations are solely determined at the leaf level. When we introduce estimation error into these values, we now introduce sampling variation into the utility measures at the unit level; therefore, units are now only exchangeable stochastically. This increased flexibility allows the allocation methods to optimize even more, ensuring that the units that happen to render large(r) amounts of utility are prioritized, even if others in their leaves render small amounts of utility. Conversely, those that render small(er) amounts of utility can better be separated out and placed in the (low-cost) control condition, even if others in their leaves render high amounts of utility. Using the (parametric) Gaussian distribution at each leaf, as opposed to the less restrictive (nonparametric) empirical distribution, tempers this flexibility a bit, but the logic, and thus the observed ordering, still carries over.

Finally, the confidence intervals around the utility estimates are larger when a form of error propagation is used. Intuitively, though the estimated mean utility for these error propagating regimes is higher because of the flexibility in optimization, there is more uncertainty around these values, given the additional variation introduced at the unit level. We also note that this process is data driven and replicable by any practitioner wanting to get a sense of how different the allocations and utilities would be under any of the allocation and error propagation regimes. Our recommendation for practitioners is to conduct an analysis similar to ours, comparing the results from the various allocation methods, following the conservative mean prediction approach for their planning, and recognizing the potential for further upside.

### *Comparing HTE Learners*

We operationalize heterogeneous treatment effects above to understand the expected effect of placing a subject into a given treatment condition on the outcome of the subject. We use a two-stage procedure proposed by (Berry et al. 2016) ($\hat{B}_i(j) = \hat{\tau}^{B}_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$), which is demonstrated to have

better performance than causal trees (Athey and Imbens 2016), a one-stage, single-tree learning procedure. The argument is that utilizing single tree methods—e.g., causal tree or transformed outcome tree (Athey and Imbens 2016)—to capture treatment effects can lead to estimations that are partly fit on the error terms, which eventually find treatment effects even when they are not present.

We estimate treatment effects using causal forest (Wager and Athey 2018) (an ensemble method, with a causal trees base learner) and predict outcomes by adding control group predictions to treatment effect estimates. Table 6 shows the comparison of accuracy metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Error (ME)—on the out-of-sample test data for causal forest and Berry-2S (the two-stage method proposed by (Berry et al. 2016)). We first note that MAE and MSE for both the learners are comparable, paying attention to $MSE = Bias^2 + Variance$, i.e., mean squared error can be decomposed into the square of bias plus variance. Moreover, we note that ME is essentially an empirical measure of bias; therefore, we see that causal forest has higher bias (ME) than Berry-2S. This shows that causal forest may trade off additional bias for lower variance given its very similar MSE. The presence of this additional bias introduced by casual forest, in addition to its base learner's potential to partly fit on the error term, might result in inaccurate predictions of rank of treatment groups for each subject. Therefore, we perform the OUR analysis for causal forest predictions and see that the condition fails (Figure 7), recalling that the OUR metric evaluates a model performance by obtaining the unbiased estimate of the utility it would provide across treatment conditions. It therefore appears that causal forest's additional bias undermines its ability to properly rank treatment conditions and, following the logic presented in above, we conclude that the causal forest learner fails to capture the exploitable individual heterogeneity between treatment groups. Moreover, we have evidence that this between-heterogeneity exists, given that the two-stage procedure passed the OUR condition and thus the outcome predictions from causal forest are not reliable for further budget-constrained optimization.

## Extension to the Contextual Bandits[10]

In this paper, we focus on leveraging data from existing large-scale randomized field experiments. An organization often accumulates a large number of historical experiments and faces the objective of maximizing the utility in the presence of a budget constraint when designing new policy interventions. With the use of multi-armed bandit algorithms, our framework can be extended to situations where no historical experiment data is available. Multi-armed bandits are sequential experimentation procedures that use a combination of exploration and exploitation strategies with the goal of maximizing overall utility. In these procedures (Agrawal and Goyal 2012), the treatment allocation is determined sequentially every time an outcome or a batch of outcomes are observed (Zhou, Xu, and Blanchet 2019). Contextual bandits (Langford and Zhang 2007) are an extension to the bandit approach where treatment allocations are performed not only based on the values of outcomes but also on the context at hand. For example, a context can be the known characteristics of an individual being allocated. There are parallels between our framework and the contextual bandits approach in that the aim is to improve the overall utility of the allocation process and the usage of context in order to make treatment allocations.

While contextual bandits are continuous learning procedures with exploration and exploitation trade-offs, our framework is a three-stage discrete process that folds together exploration via randomized experiments, learning via heterogeneous treatment effect estimation, and exploitation via optimal allocation of subjects in a step-by-step process. When there is no preexisting experiment data, contextual bandit algorithms like LinUCB (Chu et al. 2011) offer a nice way to jump-start and optimize the utility while exploring the variation. When there is preexisting data from randomized experiments, the approach we have presented may be readily used. In addition, we developed a modified version of the LinUCB algorithm that uses experiment data as an extension of our method.[11] The modified LinUCB also allows us to make a comparison between HTE-EST and LinUCB. The main idea in the modified LinUCB algorithm is to apply ridge regression on the randomized experiment data to calculate initial learning parameters and use them as inputs to the LinUCB algorithm. After this stage, sequential allocations are done based on exploration-exploitation trade-offs instead of pure exploitation done in HTE-EST. The detailed algorithm can be found at the end of the Appendix.

Figure 8 compares the total expected utility realized by HTE-EST and modified LinUCB for different budget constraints. We see that HTE-EST still performs better, compared to the modified LinUCB. We believe that, given the matrices of benefit, cost and utility (as in Figure 4), the expected realized utility generated by the prescriptions of HTE-EST will always be higher than any other prescription method because, for a matrix of nonchanging predicted values, the solution (prescriptions) generated by the ILP optimization used by HTE-EST is theoretically optimal.

[10] We thank one of the reviewers for insightful suggestions, which inspired us to make the extension.

[11] Again, LinUCB does not need any preexisting experiment data and offers an advantage in that scenario.

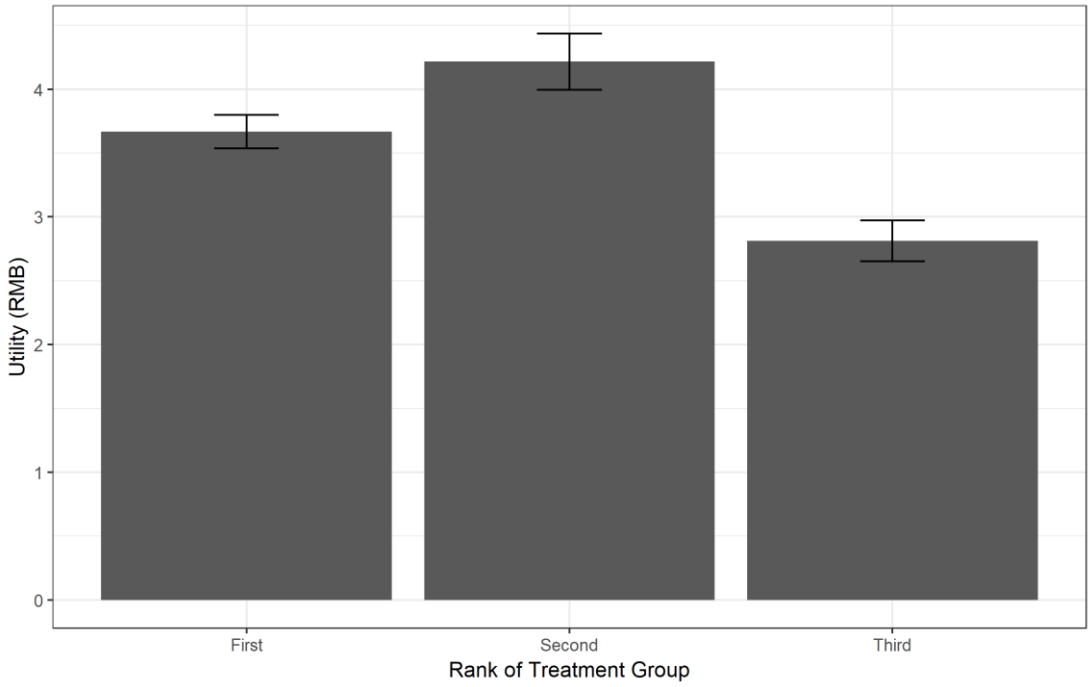



**Figure 7. Observed Utility Ranking (OUR) for Blood Donation Experiment Using Causal Forest Learner**

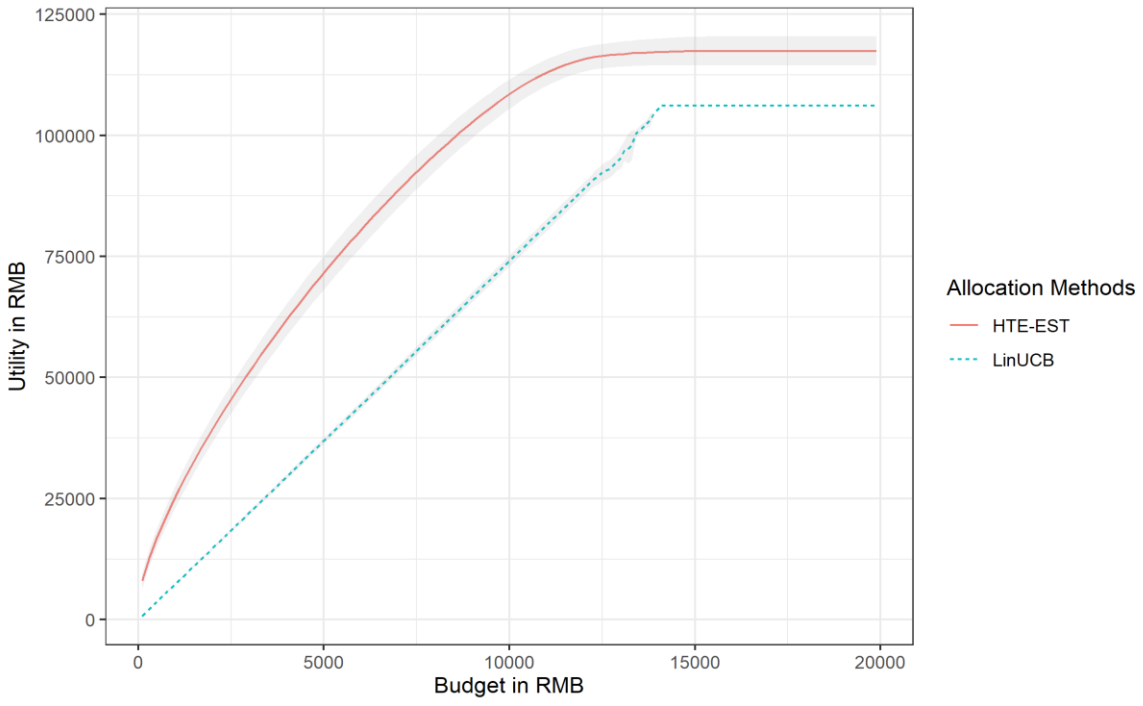



**Figure 8. Expected Total Utility Generated by Modified LinUCB Algorithm and HTE-EST over a Range of Budget Constraints**

However, in a real-world setting, where LinUCB uses realized outcomes instead of predicted outcomes, we believe there is a possibility for realized utility to be higher than that of HTE-EST. Finally, from an implementation perspective, our framework requires less supporting infrastructure, compared to bandit approaches, and is also readily accessible to less sophisticated organizations.

## Discussion and Concluding Remarks

In this study, we address the general problem of a budget-constrained decision maker facing ex ante unknown costs and benefits from multiple policy levers that can potentially be deployed to optimize an organizational goal. We define and deploy a three-stage, prescriptive analytics approach as one that folds together the use of (1) randomized field experiments and causal inference, (2) machine learning to identify heterogeneity in treatment effects, and (3) constrained optimization to optimally decide which subpopulations to treat with which policy levers to maximize profit.

We show that there is potential in combining the four pillars of analytics, as they are different components of a symbiotic process: *randomized experiments* enable the *exploration* of policy outcomes, *causal inference* enables the *estimation* of these policies' impact, *machine learning* enables the *analysis* and *prediction* of the impact variation across subpopulations, and *optimization* enables the *prescription* of a future optimal strategy under constraints.

We use datasets from two large-scale, randomized field experiments—one from public policy in the context of stimulating blood donations, and one from referral marketing—to illustrate the generality of our framework and demonstrate its performance in real-world decision-making, as compared to traditional approaches using the average treatment effect and uplift modeling. The first study on donor recruitment shows that that the prescriptions from our framework may result in up to 340% and 240% increases in overall utility, as compared to the prescriptions provided by ATE and UM, respectively. We show that ATE and UM suffer from being either too nonspecific or too myopic and therefore fail to attain the best utility. Our approach is aware of treatment heterogeneity and the budget constraint and can thus optimally allocate the treatments across individuals. In addition, as demonstrated in the second study on referral marketing, we propose new criteria—observed utility rank (OUR)—to detect heterogeneity in the effect across treatments. Identifying such heterogeneity in the effects of the various treatments is critical for future prescriptions. When between-treatment heterogeneity is unavailable in the data, the organization can save the effort and cost of further (likely unsuccessful) exploitation. Instead, the organization might collect more data to explore, aided by exploratory analytics methods, for other sources of heterogeneity.[12]

Our work is motivated by the observation that a vast majority of F500 companies have a *culture and capabilities deficit* to systematically make sense of the data they already have to create value (Hosanagar and Saxena 2017). Most approaches in the extant literature to solving the proposed decision-making problem of this paper do not combine all the pillars of analytics. While many companies are starting to use predictions from historical (observational) data, such an approach does not address the (often critical) causal questions. Even among the organizations (and policy makers broadly) that subscribe to casual inference, few actually go beyond the average treatment effect, which effectively assumes the population has a homogeneous response to treatment. We believe that our proposed full-spectrum approach to business analytics can make a significant contribution in reducing this deficit.

Future research can further extend our framework in a few ways. First, we have studied two representative real-world decision scenarios, one involving the nonprofit (blood donation) sector and the other involving the for-profit (referral marketing) sector. Future studies may explore decision scenarios in other contexts, such as pricing, new product development and testing, and user-interface design. We expect that the importance of treatment heterogeneity and organizational constraints might differ across contexts and thus influence the improvement that the prescriptive analytics framework can help achieve. Second, in our current study, we treat the data acquisition process as exogenous and given. In other words, we assume an organization may utilize all existing data within the organization in the prescriptive analytic approach. However, with the emergence of a large number of third-party data vendors and data exchange (e.g., TowerData, BlueKai), organizations can increasingly acquire new customer data to improve advertising campaigns and expand the customer base. When acquiring consumer data from external sources, it is important to decide which features of the customer data are of value and should be acquired. Our framework, especially the development of the OUR criteria, may help organizations systematically evaluate the value of external data. Such data acquisition process would also help

[12] Specifically, the OUR condition is not satisfied when the between-heterogeneity is too small to create a statistical difference or when the chosen models are too inaccurate to find the heterogeneity present. The latter is why we recommend searching for and selecting the model that minimizes MAE, using whatever means at the company's disposal. After minimizing MAE with the present data, if either case still exists, the recommendation for the company would be to collect more data and reselect models to reduce the MAE. Both these steps will increase the probability of OUR being satisfied.

organizations better understand whether the lack of treatment heterogeneity (as in our second study) is due to limited customer features available or to the intrinsic (unpredictable) nature of the decision scenario. In the same vein, firms can (and should) take into account potential heterogeneity when thinking of the size of the experiment. A larger experiment with a larger sample may not only boost the power to detect the main effect, but also create variation for the estimation of heterogeneous treatment and further optimization. Future research could take the sample size as an experiment design choice and investigate its implication to the estimation of heterogeneity and optimal policy. Finally, our study validates the performance of the proposed approach using the data from the two large-scale experiments. Future research could directly compare the performance of our suggested optimal policy with other personalized policy using a field test.

We envision that the integrated prescriptive analytic framework can be further enriched, customized, and deployed in a wide range of nonprofit or for-profit, digital native or traditional, and established or emerging organizations. We hope that our study serves as a valuable first step for such future efforts.

## References


Agarwal, R., and Dhar, V. 2014. "Big Data, Data Science, and Analytics: The Opportunity and Challenge for I: Research," *Information Systems Research* (25:3), 443-448.

Agrawal, S., and Goyal, N. 2012. "Analysis of Thompson Sampling for the Multi-Armed Bandit Problem," in *Proceedings of the 25th Conference on Learning Theory*, pp. 39-31.

Aral, S., and Walker, D. 2011. "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management Science* (57:9), pp. 1623-39.

Athey, S., and Imbens. G. 2016. "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences* (113:27), pp. 7353-7360.

Auer, P. 2002. "Using Confidence Bounds for Exploitation-Exploration Trade-Offs," *Journal of Machine Learning Research* (3:Nov), pp. 397-422.

Bapna, R., & Umyarov, A. 2015. "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks," *Management Science* (61:8), pp. 1902-1920.

Berry, G., Franco, A., Peysakovich, A., and Taylor, S. 2016. "Two Stage: A Simple Framework for Finding Cates," presented at the Conference on Digital Experimentation, Cambridge, MA.

Bertsimas, D., & Kallus, N. (2020). "From Predictive to Prescriptive Analytics," *Management Science* (66:3), pp. 1025-1044.

Biesecker, L. G. 2013. "Hypothesis-Generating Research and Predictive Medicine," *Genome Research* (23:7), pp. 1051-1053.

Chu, W., Li, L., Reyzin, L., and Schapire, R.. 2011. "Contextual Bandits with Linear Payoff Functions," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 208-14.

Ding, W., Qin, T., Zhang, X.-D., and Liu, T.-Y. 2013. "Multi-Armed Bandit with Budget Constraint and Variable Costs," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, WA.

Domingos, P. 1997. "Knowledge Acquisition Form Examples Vis Multiple Models," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 98-106.

Dopazo, J., and Aloy, P. 2006. "Discovery and hypothesis generation through bioinformatics," *Genome Biology* (7:2), Article 307.

Ghose, A., Singh, P. V., and Todri, V. 2017. "Got Annoyed? Examining the Advertising Effectiveness and Annoyance Dynamics," in *Proceedings of the International Conference on Information Systems*, Seoul, South Korea.

Grimmer, J., Messing, S., & Westwood, S. J. 2017. "Estimating Heterogeneous Treatment Effects and the Effects Of Heterogeneous Treatments with Ensemble Methods," *Political Analysis* (25:4), pp. 413-434.

Hitsch, G. J., and Misra, S. 2018. "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," Working Paper, (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3111957).

Hosanagar, K., and Saxena, A. 2017. "The Democratization of Machine Learning: What It Means for Tech Innovation," *Knowledge@Wharton* (https://knowledge.wharton.upenn.edu/article/ democratization-ai-means-tech-innovation/).

Imai, K., and Ratkovic, M. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics* (7:1), pp. 443-470.

Imai, K., and Strauss, A. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign," *Political Analysis* (19:1), pp. 1-19.

Joulani, P., Gyorgy, A., and Szepesvári, C. 2013. "Online Learning Under Delayed Feedback," in Proceedings of *International Conference on Machine Learning*, pp. 1453-1461.

Jung, J., Bapna, R., Golden, J. M., & Sun, T. 2020. "Words Matter! Toward a Prosocial Call-to-Action for Online Referral: Evidence from Two Field Experiments," *Information Systems Research* (31:1), pp. 16-36. Kohavi, R., and Thomke, S. 2017. "The Surprising Power of Online Experiments," *Harvard Business Review* (95:5), pp. 74-82.

Künzel, S. R, Sekhon, J. S., Bickel, P. J., and Yu, B. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proceedings of the National Academy of Sciences* (116:10), pp. 4156-4165.

Langford, J., and Zhang, T. 2007. "The Epoch-Greedy Algorithm for Contextual Multi-Armed Bandits," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 817-824.

Langford, J., and Zhang, T. 2008. "The Epoch-Greedy Algorithm for Multi-Armed Bandits with Side Information," in *Advances in Neural Information Processing Systems*, pp. 817-824.

Li, L., Chu, W., Langford, J., and Schapire, R. E. 2010. "A Contextual-Bandit Approach to Personalized News Article

Recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, pp. 661-70.

Martelo, S., and Toth, P. 1990. *Knapsack Problems: Algorithms and Computer Implementations*, S. Martello and P. Toth (eds.), Wiley.

McFowland III, E., Somanchi, S., and Neill, D. B. 2018. "Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection," Working Paper (https://arxiv.org/pdf/1803.09159.pdf).

McFowland III, E., Speakman, S. D., and Neill, D. B. 2013. "Fast Generalized Subset Scan for Anomalous Pattern Detection," *The Journal of Machine Learning Research* (14:1), pp. 1533-1561.

Misra, K., Schwartz, E. M., & Abernethy, J. 2019. "Dynamic Online Pricing with Incomplete Information Using Multiarmed Bandit Experiments," *Marketing Science* (38:2), pp. 226-252.

Neill, D. B. 2012. "Fast Subset Scan for Spatial Pattern Detection," *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* (74:2), pp. 337-360.

Neill, D. B, McFowland III, E., and Zheng, H. 2013. "Fast subset scan for multivariate event detection," *Statistics in Medicine* (32:13), pp. 2185-2208.

Rosenbaum, P. R, and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* (70:1), pp. 41-55

Rzepakowski, P., and Jaroszewicz, S. 2010. "Decision Trees for Uplift Modeling," in *Proceedings of the IEEE 10th International Conference on Data Mining*, pp. 441-450.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. 2017. "Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments," *Marketing Science* (36:4), pp. 500-522.

Somanchi, S., McFowland III, E., and Neill, D. B. 2018. "Discovering Heterogeneous Patterns of Care Using Observational Data: Evidence from Studies of Healthcare," Working Paper.

Speakman, S. D., McFowland III, E., and Neill, D. B. 2015. "Scalable Detection of Anomalous Patterns with Connectivity Constraints," *Journal of Computational and Graphical Statistics* (24:4), pp. 1014-1033.

Sun, T., Gao, G., and Jin, G. Z. 2019. "Mobile Messaging for Offline Group Formation in Prosocial Activities: A Large Field Experiment," *Management Science* (65:6), pp. 2717-36.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. 2018. "Linear Bandits with Stochastic Delayed Feedback," Working Paper (https://arxiv.org/pdf/1807.02089v3.pdf).

Wager, S., and Athey, S. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association* (113:523), pp. 1228-1242.

Zhou, Z., Xu, R., and Blanchet, J. 2019. "Learning in Generalized Linear Contextual Bandits with Stochastic Delays," in *Proceedings of the Advances in Neural Information Processing Systems Conference*, pp. 5197-5208.

# Appendix

## Summary Statistics of the Two Case Studies

**Table A1. This Table Reports the Summary Statistics of Variables and Outcomes in the Referral Marketing Experiment**

| | N | Mean | St. dev. | Min | Max |
|---|---|---|---|---|---|
| **No. of past orders** | 99,881 | 2.68 | 4.46 | 1 | 499 |
| **Total spent (USD)** | 99,881 | 77.87 | 180.13 | 0 | 37,794.07 |
| **Total discount (USD)** | 99,881 | 206.58 | 505.43 | - 0.51 | 102,011.1 |
| **Total refunded (USD)** | 99,881 | 2.65 | 22.91 | 0 | 2,383.78 |
| **Last purchase (days)** | 99,881 | 368.01 | 149.49 | 124 | 559 |
| **NPS** | 99,881 | 1.84 | 3.75 | 0 | 10 |
| **No. of NPS comments** | 99,881 | 0.28 | 0.7 | 0 | 37 |
| **Firm cost (USD)** | 99,881 | 0.02 | 0.73 | 0 | 60 |
| **Firm gain (USD)** | 99,881 | 0.32 | 8.181 | 0 | 781.8 |
| **Firm utility (USD)** | 99,881 | 0.29 | 8.17 | - 47.95 | 781.8 |

**Table A2. This Table Reports the Summary Statistics of Variables and Outcomes in the Blood Donation Experiment.**

| | N | Mean | St. dev. | Min | Max |
|---|---|---|---|---|---|
| **Weight (kgs)** | 57,575 | 65.06 | 11.08 | 45 | 115 |
| **Age** | 57,575 | 28.26 | 8.88 | 18 | 61 |
| **Total past donation (ml)** | 57,575 | 483.75 | 407.64 | 0 | 6,400 |
| **Most recent donation (months)** | 57,575 | 15.04 | 6.75 | 0 | 116 |
| **Num voluntary donation** | 57,575 | 1.02 | 1.38 | 0 | 69 |
| **Num group donation** | 57,575 | 0.35 | 0.68 | 0 | 9 |
| **Num mutual donation** | 57,575 | 0.09 | 0.29 | 0 | 2 |
| **Num plasma donation** | 57,575 | 0.02 | 0.64 | 0 | 68 |
| **Firm cost (RMB)** | 57,575 | 1.32 | 4.91 | 0 | 52 |
| **Firm gain (RMB)** | 57,575 | 4.22 | 41.66 | 0 | 440 |
| **Firm utility (RMB)** | 57,575 | 3.72 | 36.77 | 0 | 389 |

## Pseudo Code for Modified LinUCB Algorithm

**input :** training data ($\mathcal{N}_1$), testing data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), $\alpha \in \mathrm{R}^+$ (Hyper-parameter for LinUCB algorithm), Budget (M)

/* $\mathcal{N}_1$ is composed of benefit outcomes ($\vec{B}$), cost outcome ($\vec{C}$), Utility outcome ($\vec{U} = \vec{B} - \vec{C}$), covariate matrix ($\mathbb{X}$), and treatment condition ($\vec{W}$) */

/* $\mathcal{N}_2$ is composed of covariate matrix ($\mathbb{X}$), predicted outcomes ( $\hat{\vec{B}}, \hat{\vec{C}}, \hat{\vec{U}} = \hat{\vec{B}} - \hat{\vec{C}}$) for each treatment group (Output from Algorithm 2) */

/* We use $i$ to index variables to individual units, and $j$ to index variables related to treatment conditions */

// Training

**for** $T_j \, in \, \mathcal{T}$ **do**

  $\vec{U_j} \leftarrow \{U_i \,|\, W_i = j \,\forall\, i\}$

  $\mathbb{X}_j \leftarrow \{\vec{X_i} \,|\, W_i = j \,\forall\, i\}$ ; // separate training data by treatment condition

**end**

**for** $T_j \, in \, \mathcal{T}$ **do**

  $\mathcal{A}_j \leftarrow \mathbb{X}_j^T \mathbb{X}_j + I$

  $\vec{b_j} \leftarrow \mathbb{X}_j^T \vec{U_j}$

**end**

// Testing (Allocation procedure with budget constraints)

**for** $i \, in \, 1 \ldots |\mathcal{N}_2|$ **do**

  $m_r \leftarrow M$

  **do**

    **for** $T_j \, in \, \mathcal{T}$ **do**

      $\hat{\vec{\theta}}_j \leftarrow \mathcal{A}_j^{-1} \vec{b_j}$

      $\vec{p}_{i,j} \leftarrow \hat{\vec{\theta}}_j^T \vec{X_i} + \alpha \sqrt{\vec{X_i}^T \mathcal{A}_j^{-1} \vec{X_i}}$

    **end**

    Choose treatment $T_j = \arg\max_j p_{i,j}$ with ties broken arbitrarily and observe predicted utility $\hat{U}_i(j)$ and cost $\hat{C}_i(j)$ for that treatment.

    $\mathcal{A}_j \leftarrow \mathcal{A}_j + \vec{X_i} \vec{X_i}^T$

    $\vec{b_j} \leftarrow \vec{b_j} + \hat{U}_i(j) \vec{X_i}$

    $m_r \leftarrow m_r - \hat{C}_i(j)$ ; // Remaining Budget

  **while** $m_r > 0$;

**end**

**output:** Predicted Utility $\hat{\vec{U}}$ and Predicted Cost $\hat{\vec{C}}$

**Algorithm 4. LinUCB Implementation**