Personalized Treatment Selection using Causal Heterogeneity

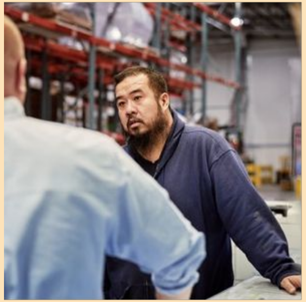
Ye Tu, Kinjal Basu, Cyrus DiCiccio, Romil Bansal, Preetam Nandy, Padmini Jaikumar, Shaunak Chatterjee

The Web Conference, 2021

Overview

Introduction






Heterogeneity of treatment effect

Randomized experimentation (**A/B testing**) is widely used in the internet industry to measure the metric impact obtained by different **treatment variants**.

- e.g., different models, parameter value choices, and UI components.

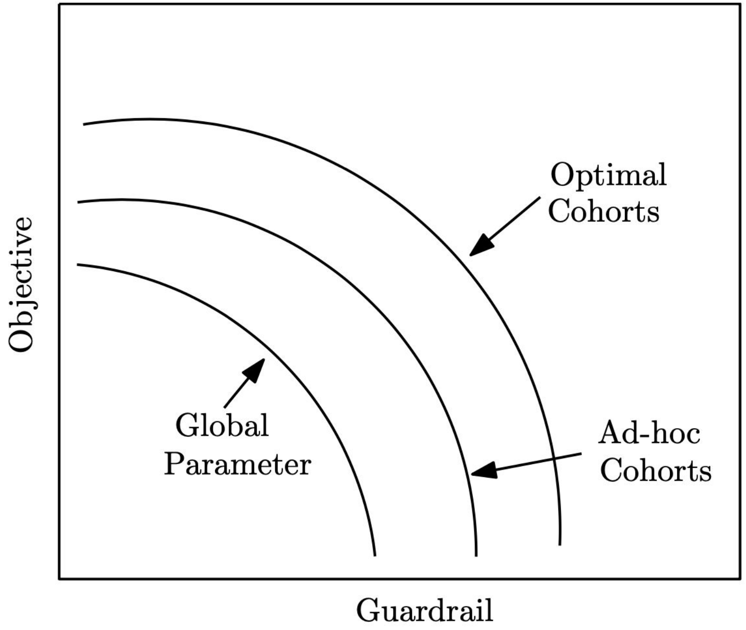
The effect of a given treatment can be **heterogeneous** across experimental units.

Personalized treatment selection

Global allocation: identify the treatment variant that performs the best in the entire population and ramp that variant to everyone.

A **personalized approach** for treatment selection can greatly improve upon the usual global selection strategy.

- **Choosing these cohorts wisely** is one of our main focus areas.



Major Contributions

We develop a **general framework** for selecting optimal treatment variants for members by estimating **heterogeneous causal effects** and solving an **optimization problem**.

- We discuss ways to identify which among the proposed techniques should be chosen for a given application.
- We introduce a novel **merging tree algorithm** to handle **multiple** treatments and metrics of interests.
- We adopt a **multiple cooperative stochastic approximation** to solve multi-objective optimization while considering the variances in estimations.
- We do extensive **simulations** to show the benefit of using our framework.

We describe the **infrastructure** required to put such a system in production.

We show results on a **real-world application** that has resulted in significant metric wins.

Major Contributions

A **general framework** for selecting optimal treatment variants for members by estimating **heterogeneous causal effects** and solving an **optimization problem**.

- **Framework of solutions**: With guidance on which one to pick and when
- **Technical novelty**
  - Merging tree algorithm
  - Multiple cooperative stochastic approximation
- **Real-world application**
  - Building the serving infrastructure
  - Strong, positive results from a large scale industrial application

Problem Set-up

Notations and Objective

Symbol	Meaning
J	Total number of treatment variants or choices.
K	Total number of guardrail metrics
C_i	i-th cohort (the smallest cohort would be a individual member) for $i = 1, \ldots, n$.
$\mathbf{U}_k$	Vectorized version of $U_{i,j}^k$, which is the causal effect in metric k by variant j in cohort C_i.
$\boldsymbol{\mu}_k$	Mean of $\mathbf{U}_k$
$\boldsymbol{\sigma}_k^2\mathbf{I}$	Variance of $\mathbf{U}_k$
$\mathbf{x}$	The assignment vector.

Let k = 0 denote the main success metric (objective). We wish to maximize the objective keeping the constraint metrics at a threshold. Formally, we wish to get the optimal x* by solving the following:

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & \mathbf{x}^T\mathbf{U}_0 \\ \text{subject to} \quad & \mathbf{x}^T\mathbf{U}_k \leq c_k \qquad \text{for } k = 1, \ldots, K. \\ & \sum_j x_{i,j} = 1 \ \ \forall i, \qquad 0 \leq \mathbf{x} \leq 1 \end{aligned}$$

Problem Breakdown

(1) Identify member cohorts $C_1, . . . , C_n$ using data from randomized experiments, and then estimate the cohort-level causal effects U_k.

- At a **member-level** set-up, where **each member represents a cohort**, we directly estimate the individual level causal effects.

(2) Optimally allocate treatment variants x* to each member cohort by solving the optimization problem.

Problem Breakdown

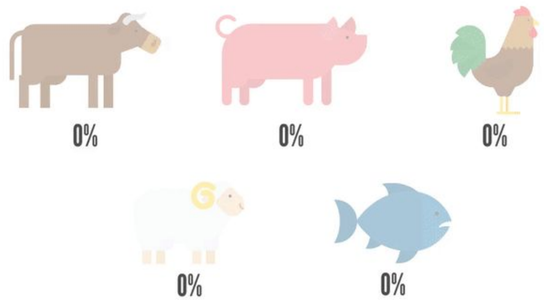
1. Identify member cohorts $C_1, \ldots, C_n$ using data from randomized experiments to estimate the causal effect U_k for each cohort

2. Optimally allocate treatment variants x* to each member cohort by solving the optimization problem.

Methodology

Framework Breakdown

We first begin with how we can estimate **heterogeneous causal effects** at either cohort or member level.



We then describe how we solve the **optimization** problem to select optimal treatment variants for each member.

Heterogeneous Effects Estimations

In this paper, we follow the potential outcomes framework from Rubin (1974) [21] and consider the following assumptions:

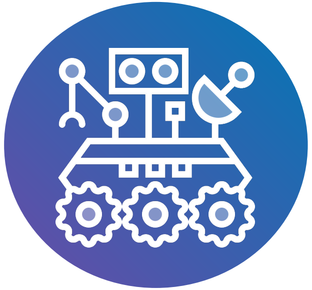
- **Stable Under the Treatment Value Assumption** (SUTVA) [21], which states that the response of the treatment unit only depends on the allocated treatment to that unit and not on the treatment given to other units.
- **Strongly Ignorable Treatment Assignment** [20], which combines the assumption of unconfoundedness and overlap. We refer to [20] for the details.

Cohort-Level Heterogeneity

We use the recursive partitioning technique from Athey and Imbens [1] to identify the heterogeneous cohorts.

Regression Tree (CART)

- Predict Y
- Splitting Objective: MSE(Y)



Causal Tree

- Estimate treatment effect τ E(Y1)-E(Y0)
- Splitting Objective: MSE(τ) + Variance regularizer

Multiple treatments and metrics

Causal tree can only handle one objective metric and a binary treatment definition at a time.

- One option could be merging the $J\ (K + 1)$ tree models into one single cohort assignment.
- Simply merging all the trees would fragment the cohorts into very small subsets with **extremely noisy estimations**.
- We avoid this unwanted noise by carefully exploiting the within cohort homogeneity of the treatment effect by **Algorithm 1**.

Merging Trees - Algorithm 1

We sequentially merge the cohort sets $S_{j,k} = \{ C_1^{j,k}, \ldots, C_n^{j,k} \}$ to obtain the following set of **mutually exclusive** and **exhaustive** cohorts

$$\mathcal{S}_{out} = \left\{ \cap_{j=1}^{J} \cap_{k=0}^{K} C^{j,k} \neq \emptyset \mid C^{j,k} \in \mathcal{S}_{j,k} \right\}.$$

For each treatment j and each metric k, we **retain** the estimated treatment effect and its variance from the **original cohort**. Since each $S_{j,k}$ is exhaustive, this provides estimates of treatment effect and its variance for all sub-partitions.

Algorithm 1 Merging Trees

Input: L cohorts sets: $\{\{C_i^\ell\}_{i=1}^{n_\ell} \mid \ell = 1, \ldots, L\}$ and corresponding treatment effects and variances $\{\{(U(C), \sigma^2(C)) \mid C \in \{C_i^\ell\}_{i=1}^{n_\ell}\} \mid \ell = 1, \ldots, L\}$

Output: $\mathcal{S}_{out}$ and $\mathcal{T}_{out}$

```
Set S_out = {C_i^1}_{i=1}^{n_1} and T_out = {(U_1(C), σ_1^2(C)) | C ∈ S_out}
for ℓ = 2, ..., L do
    for A ∈ S_out do
        for B ∈ {C_i^ℓ}_{i=1}^{n_ℓ} do
            C = A ∩ B
            if C ≠ ∅ then
                S_out = S_out ∪ {C}
                T_out = T_out ∪ {(U_m(C), σ_m^2(C)) | m = 1, ..., ℓ},
```

where

$$U_m(C),\ \sigma_m^2(C) = \begin{cases} U_m(A),\ \sigma_m^2(A) & \text{for } m \leq l-1 \\ U_m(B),\ \sigma_m^2(B) & \text{for } m = \ell \end{cases}$$

```
            end if
        end for
    end for
    S_out = S_out \ {A}
    T_out = S_out \ {(U_m(A), σ_m^2(A)) | m = 1, ..., ℓ − 1}
end for
```

Member-level Heterogeneity

To estimate the heterogeneous causal effects at a **member level**, some of the options include:

(a) **Causal Forest**: The Causal Forest Algorithm [30] is an extension of the Causal Tree which was inspired by Random Forest Algorithm [5] and use ensemble learning to incorporate results from multiple tree models.

b) **Two-Model Approach**: This is a baseline method (commonly applied in uplift modeling domain) that models the causal effect at a member level through the difference of the predicted response in the treatment and control models [24].

Optimization Solution

Stochastic Optimization: the problem is stochastic since both the objective function and the constraints are not deterministic but are coming from a particular distribution (e.g., Gaussian).

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & f(\mathbf{x}) = \mathbb{E}(\mathbf{x}^T \mathbf{U}_0) \\ \text{subject to} \quad & g_k(\mathbf{x}) := \mathbb{E}(\mathbf{x}^T \mathbf{U}_k - c_k) \leq 0, \qquad k = 1, \dots, K. \\ & \sum_j x_{ij} = 1 \ \forall\, i, \qquad 0 \leq \mathbf{x} \leq 1 \end{aligned}$$

Deterministic Optimization: If using sample average approximation (SAA)[13], we replace the stochastic objective and constraints via their empirical sample expectation.

$$\begin{aligned} \underset{\mathbf{x}}{\text{Maximize}} \quad & f(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\mu}}_0 \\ \text{subject to} \quad & g_k(\mathbf{x}) := \mathbf{x}^T \hat{\boldsymbol{\mu}}_k - c_k \leq 0, \qquad k = 1, \dots, K. \\ & \sum_j x_{ij} = 1 \ \forall\, i, \qquad 0 \leq \mathbf{x} \leq 1. \end{aligned}$$

Stochastic Approximation

Algorithm 2 Multiple Cooperative Stochastic Approximation

Input : Initial $\mathbf{x}_1 \in \mathcal{X}$, Tolerances $\{\eta_k\}_t, \{\gamma\}_t$, Iterations N
for $t = 1, \ldots, N$ **do**
  Estimate $\hat{G}_{k,t}$ for all $k \in 1, \ldots, K$ using (4).
  if $\hat{G}_{j,t} \leq \eta_{j,t}$ for all j **then**
    Set $h_t = F'(\mathbf{x}_t, \mathrm{U}_{0,t})$
  else
    Randomly select k^* from $\{k : \hat{G}_{k,t} > \eta_{k,t}\}$
    Set $h_t = G'_{k^*}(\mathbf{x}_t, \mathrm{U}_{k^*,t})$
  end if
  Compute $\mathbf{x}_{t+1} = P_{\mathbf{x}_t}(\gamma_t h_t)$
end for
Define $\mathcal{B} = \left\{1 \leq t \leq N : \hat{G}_{k,t} \leq \eta_{k,t} \; \forall k \in \{1, \ldots, K\}\right\}$
return $\hat{\mathbf{x}} := \frac{\sum_{t \in \mathcal{B}} \mathbf{x}_t \gamma_t}{\sum_{t \in \mathcal{B}} \gamma_t}$

Multiple Cooperative Stochastic Approximation [3] is an iterative algorithm which runs for N steps. At each step t it starts by estimating the constraint function.

$$\hat{G}_{k,t} = \frac{1}{L} \sum_{\ell=1}^{L} G_k(\mathbf{x}_t, \mathrm{U}_{k,\ell}).$$

- if all the estimated constraints are less than a threshold, the algorithm chooses the gradient to be the gradient of the **objective**.
- Otherwise, from the set of violated constraints, it chooses a constraint **at random** and use the gradient of **that constraint**.

Overall Algorithm

Algorithm 4 : Optimal Treatment Selection

1: Run Randomized Experiment to collect data across various treatment variants and metrics.
2: Generate a cohort-level or member-level causal effect for the different parameters using the technique in Section 3.1.
3: Solve the optimization problem (stochastic or deterministic) as given in 3.2.
4: Return bias corrected assignment $\hat{x}$ by following Algorithm 3.

* Algorithm 3: it use bootstrap to improve bias and variance estimations: Section 3.3, Bias and Variance of Optimal Assignment Estimates.

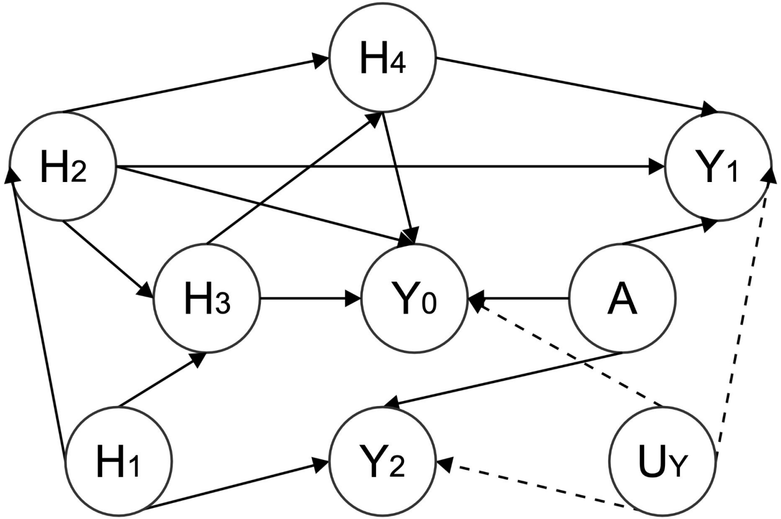
Results

Simulation Analysis

We leverage simcausal R package [23] to generate simulation datasets under self-defined causal **Directed Acyclic Graphs** (DAG).

- *Aj* as the treatment variables
- *Yk* are the metrics (or response variables)
- *Uy* as a latent variable impacts *Yk*
- *Hm* as the heterogeneous variables

We simulate heterogeneity by introducing **interaction terms** between *Aj* and *Hm* on *Yk*.



Evaluation of Simulation

We consider the normalized mean of **individualized treatment effect** (ITE) for metric k at optimal x* as

$$\tau(\mathbf{x}^*)_k = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3}(Y_{j,i,k} - Y_{0,i,k})\, z^*_{i,j}}{\mu_{0,k}},$$

($Yj,i,k - Y0,i,k$) represents the individualized treatment effect. We normalize the ITE by the control group mean μ to make results comparable across different simulated datasets.

Comparing all variants

(1) $HT.ST$: A heuristic cohort-level solution paired with stochastic optimization.

(2) $CT.ST$: Cohort-level estimations using Causal Tree model paired with stochastic optimization.

(3) $CF.DT$: Member-level estimations using the Causal Forest model [30] paired with deterministic optimization.

(4) $TM.DT$: Member-level estimations using a "Two-Model" approach (i.e., build two Random Forest [5] models) paired with deterministic optimization.

(5) *Global*: A best global allocation as baseline.

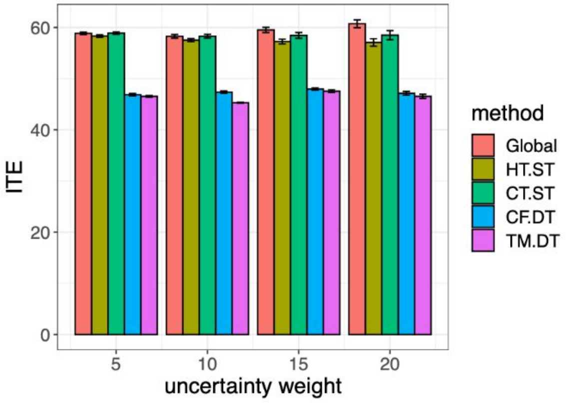
Analysis Results - Exist a global best

First scenario: Aligning the effect on the objective with that of the constraint metrics.

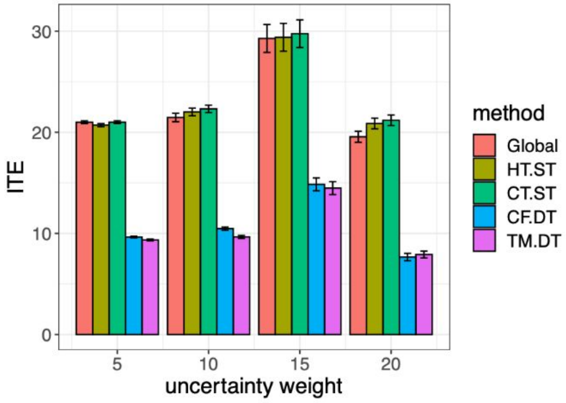
Benefit of the stochastic optimization:

- the cohort-level solutions paired with stochastic optimization ($HT.ST$ and $CT.ST$) perform almost at parity with the oracle global best solution $Global$.
- However, the member-level estimations paired with deterministic optimization ($CF.DT$ and $TM.DT$) show worse performance due to the high variance.

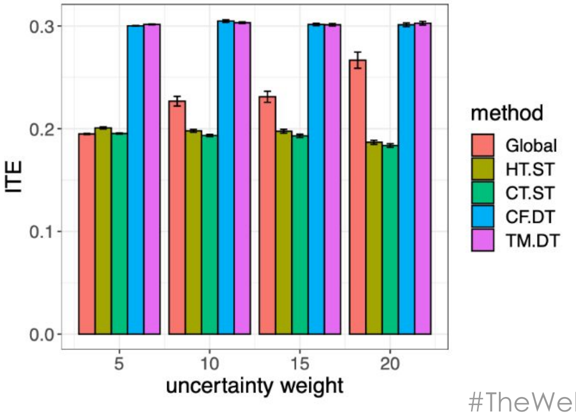
(a) Evaluation on the objective metric Y_0 if there exists a global best policy



(b) Evaluation on the constraint metric Y_1 if there exists a global best policy



(c) Evaluation on the constraint metric Y_2 if there exists a global best policy



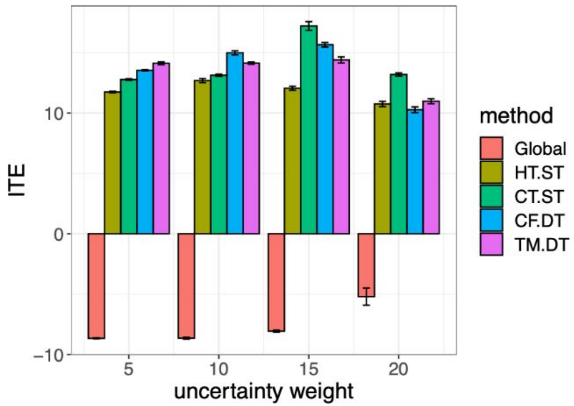
Analysis Results - No global best

Second scenario: the objective metrics move possibly in the opposite direction to some constraint metrics for some treatment.

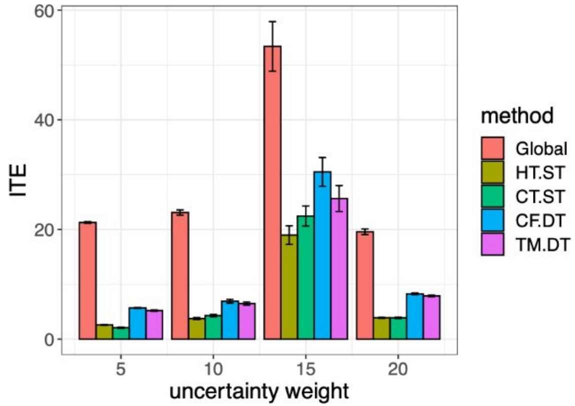
Benefit of heterogeneity estimation and personalization:

- All the proposed approaches perform better than the *Global* solution.
- With low noise levels, member-level solutions (*CF .DT* and *TM.DT*) perform better than the cohort-level solution (*HT.ST*, *CT.ST*). Along with an increase in the noise level, *CT.ST* quickly starts to catch up and can outperform the member-level solutions.

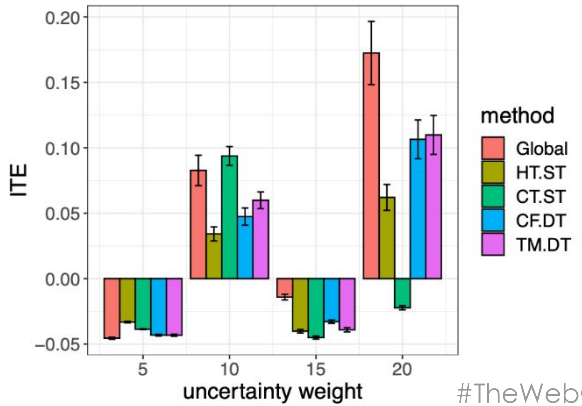
(d) Evaluation on the objective metric Y_0 if personalization can benefit the system



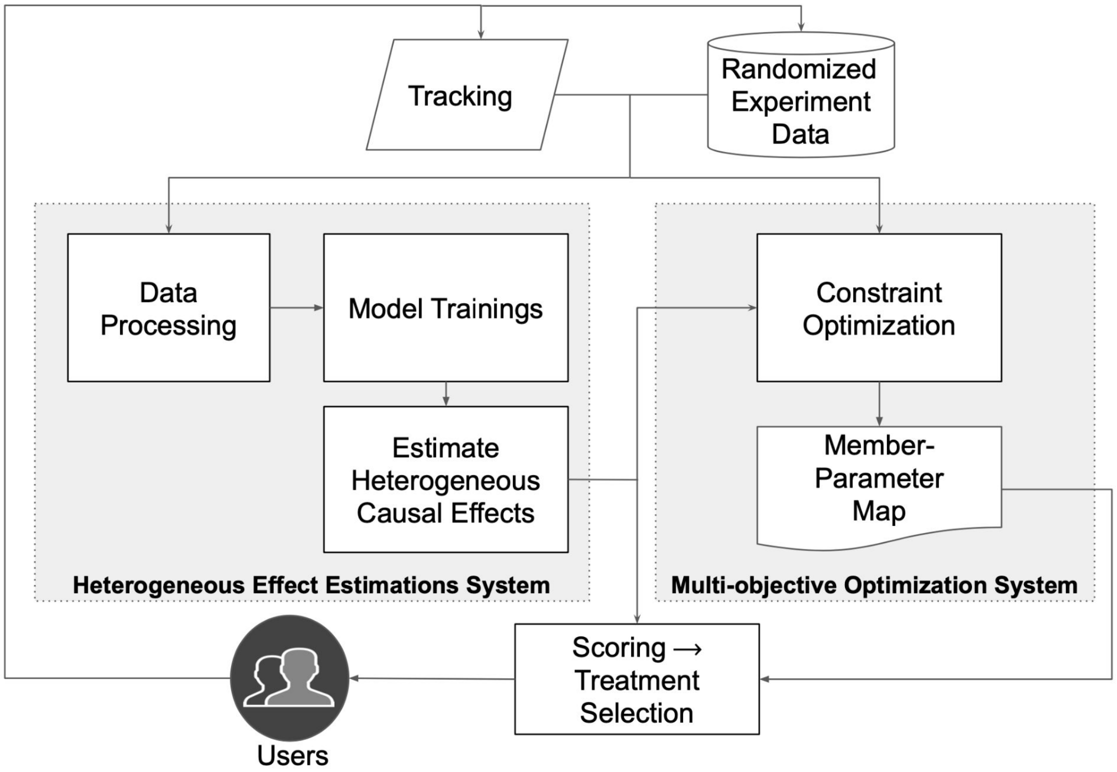
(e) Evaluation on the constraint metric Y_1 if personalization can benefit the system



(f) Evaluation on the constraint metric Y_2 if personalization can benefit the system



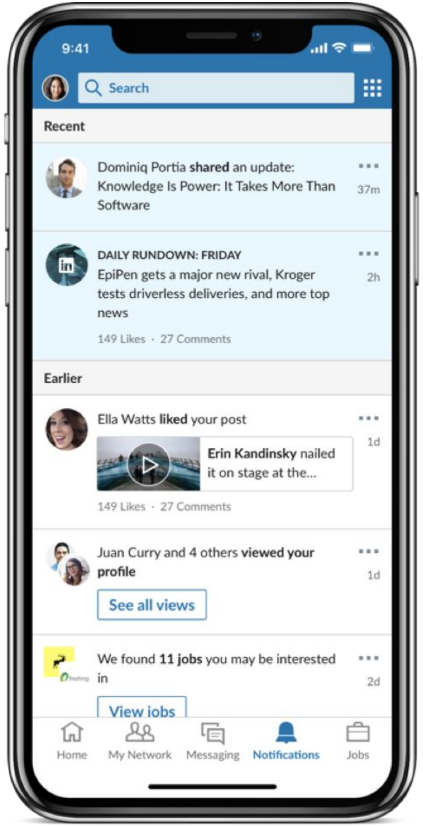
System Architecture



The general engineering architecture consists of two major components:

- One for **heterogeneous causal effect** estimations
- The other for the **optimization** module.

Application in Notification System



- Notifications are an important driver for **member visits** and **engagement**.
- Sending more notifications can increase visits, but it also has **negative consequences** (reduction in click-through rate) and increase in notifications disables.
- The system initially had a **fixed cap parameter** which was the **same for all** members.
- Our goal with introducing personalized volume caps is to **maximize visits** to LinkedIn with **constraints** on click-through rate and Notification disables metrics.

Notification System Results

- We implemented the cohort-level solution $CT.ST$.
- Heuristic Cap A and B are based a cohort definition where members are grouped into four segments according to their visit frequency.
- Personalized cap treatment showed significant positive impact on Sessions, while the impact on the constraint metrics remained within acceptable bounds. It also outperforms the both heuristic solutions.

Metrics	Descriptions
Sessions (Objective)	Number of visits to the LinkedIn site/app
Notification Sends	Volume of notifications sent to members
Notification CTR	Click through rate on notifications
Total Disables	Number of total disables on notifications

Table 3: Metrics of Interest for Personalized Capping

Metrics	ATE % Personalized Cap	ATE % Heuristic Cap A	ATE % Heuristic Cap B
Sessions	**+1.39%**	+1.31%	+0.54%
Notification Sends	**+1.64%**	+6.62%	+3.07%
Notification CTR	−1.24%	−1.73%	-1.18%
Total Disables	Neutral	+9.23%	Neutral

Table 4: Notification Cap Experiment Results

Discussions

Future work

A few non-trivial, but likely **impactful extensions** for future consideration include:

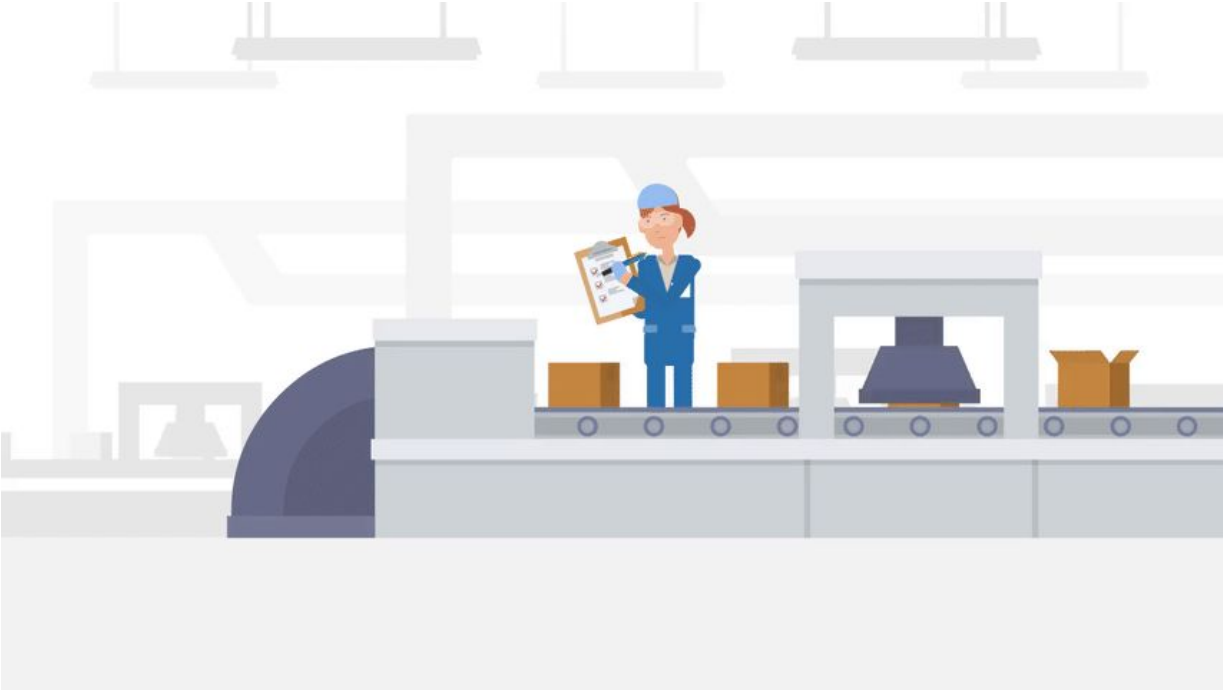
(1) Designing a more **cost-efficient data collection framework** or leveraging observational data to achieve the same performance would be beneficial.

(2) Users can potentially move in and out of cohorts. Extending this framework to incorporate the **dynamic nature of cohorts** could be an interesting research topic.

(3) Future work on generating one single optimal cohort definition based on effects from **multiple treatments with various metrics of interests** could further improve the method.

Reproducibility

We share example scripts for conduct **simulation analysis** in examining the proposed methods and stochastic optimization algorithms in the following Github link: https://github.com/tuye0305/prophet.



Reference

[1] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 27 (2016), 7353–7360.
[3] Kinjal Basu and Preetam Nandy. 2019. Optimal convergence for stochastic optimization with multiple expectation constraints. arXiv preprint arXiv:1906.03401 (2019).
[5] Leo Breiman. 2001. Random Forests. Machine Learning 45, 1 (2001), 5–32.
[23] Oleg Sofrygin, Mark J. van der Laan, and Romain Neugebauer. 2017. simcausal R Package: Conducting Transparent and Reproducible Simulation Studies of Causal Effect Estimation with Complex Longitudinal Data. Journal of Statistical Software 81, 2 (2017), 1–47.
[24] Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. 2015. Ensemble methods for uplift modeling. Data Mining and Knowledge Discovery 29, 6 (2015), 1531–1559.
[30] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. J. Amer. Statist. Assoc. 113, 523 (2018), 1228–1242.

Thank you

LinkedIn
#TheWebConf