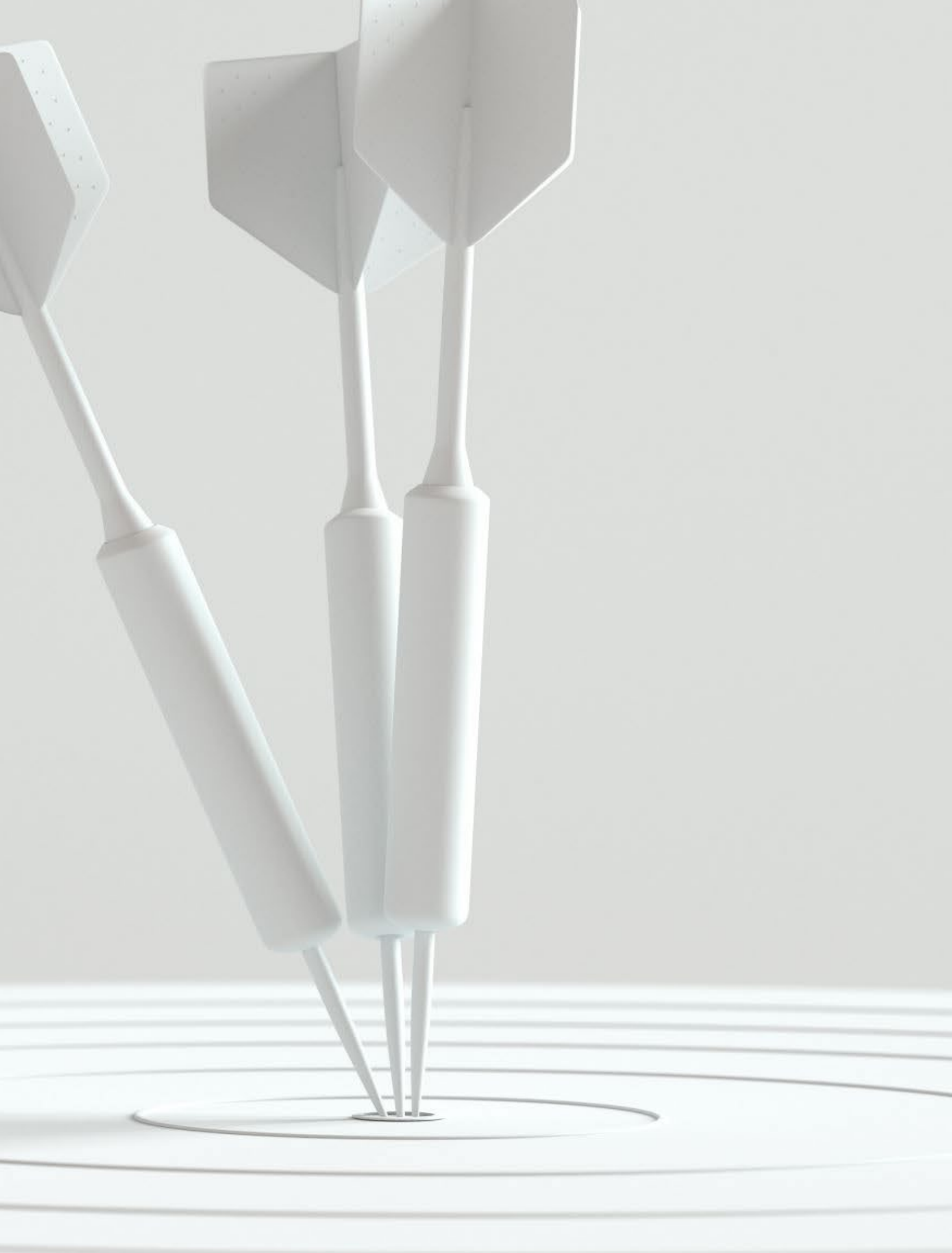
Designing Complex Experiments: Some Recent Developments

SUSAN ATHEY AND GUIDO IMBENS

STANFORD UNIVERSITY & NBER

1. What are the **goals** and **context** for the use of experimental data and results?
2. What are **challenges** in achieving goals?
3. How can we **design experiments** to better achieve goals?



Overview

- Inspiration from Tech
- Working backwards from post-experiment
- Challenges
- Design strategies
- Staggered rollout experiments
- Adaptive experiments
- Interference

Experiments in tech firms

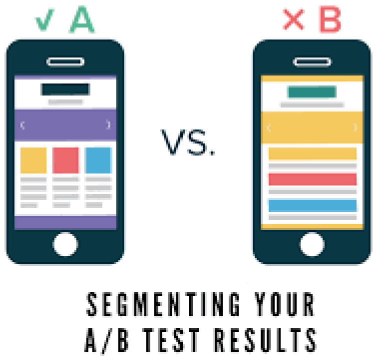
- Widespread adoption & research
- Integral to innovation, business ops
- Many open methodological ?s
- Short term, partial eqm focus

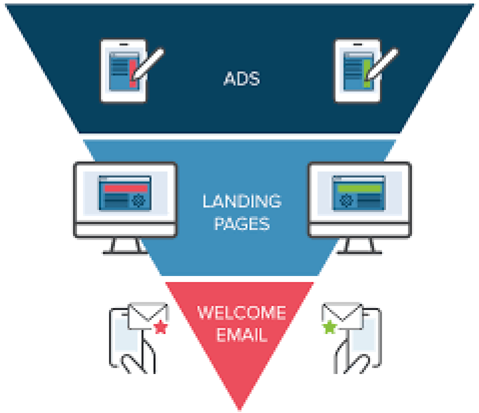
Tech changing economics and experiments

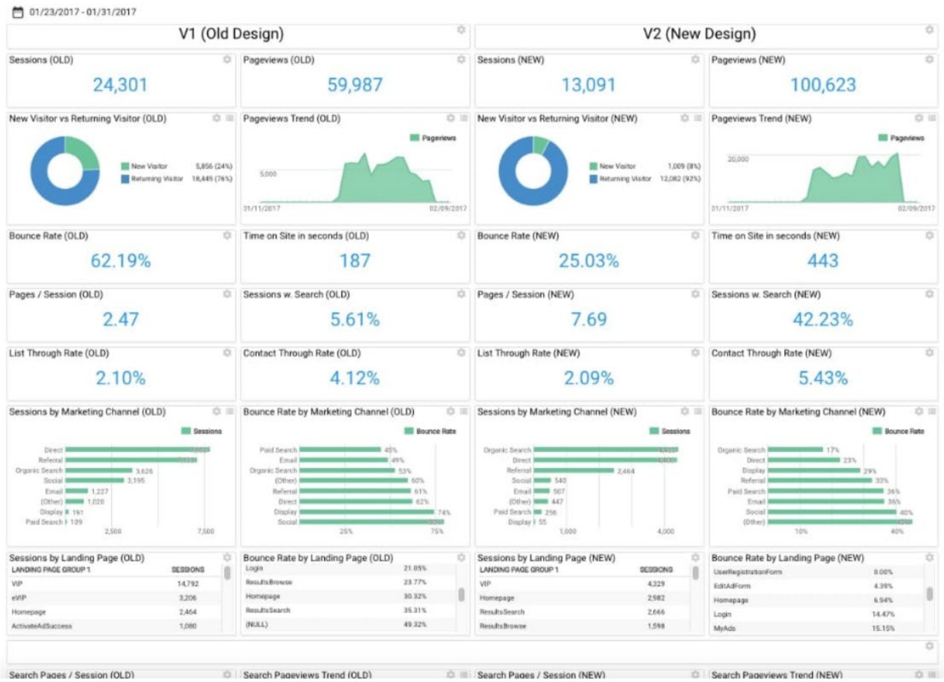
- Digitization: business, gov't, society
- Economist as foundational innovator: idea generation, architect, product designer
- Economist as incremental innovator: embedded in the build, refine, & optimize cycle

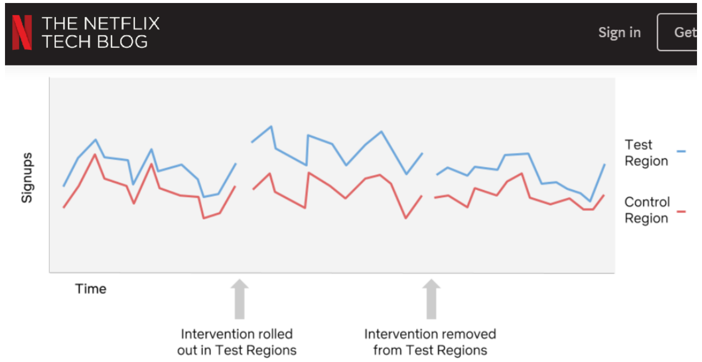
Economic frameworks changing tech and experiments

- Economics of outcomes, eqm, impact
- Theory of data-driven decision-making for orgs/policy-makers
- Applied econometrics in analysis
- Resource allocation problem for scarce experimental units
- Optimize design to achieve objectives









Reimagining Experimentation Analysis at Netflix

Netflix Technology Blog Follow
Sep 10 · 8 min read ★

Toby Mao, Sri Sri Perangur, Colin McFarland

Another day, another custom script to analyze an A/B test. Maybe you've done this before and have an old script lying around. If it's new, it's probably going to take some time to set up, right? Not at Netflix.

ABlaze

METRICS as of Sep 3, 2019	1 - Control	2 - Encode 1	3 - Encode 2
Play Delay (Fabricated Data)			

Above: Screenshots from online blogs about A/B testing

Foundational Innovation

Theory of Impact

- Goals & mechanisms
- Institutional context
- Economic, behavioral, social theories
- Dynamics, equilibrium, spillovers
- Informed by related obs. studies & experiments
- Proposed outcomes based on economic frameworks

Scope of Intervention

- Regulation or market shaping
- Firm, org., or locality
- Service provider vs. consumer in marketplaces

Incremental Innovation

Analysis of Historical Obs/Exp Data

- Off-policy (counterfactual) evaluation
- Heterogeneous treatment effects (HTE) of prior policies
- Combine w/ “foundation models” and/or external data

Refine Interventions

- Informed by theory and analysis
- Policies/algos: targeted treatment assignment, rec systems
- Pilot experiment, poss. recruited subjects

Experimental Design

- Select and validate outcome measures
- Formulate hypotheses and goals
- Design: unit, timing, measures, size, analysis plan
- Adv. experiments: adaptive, staggered rollout, dynamic treatments

Analysis & Decisions

- Revisit outcome measure properties
- “On-policy” evaluation, HTE
- Estimate optimal targeted policies
- Generalizable & tactical insights
- Deployment & decisions
- More experiments?

Foundational Innovation

Theory of Impact

- Goals & mechanisms
- Institutional context
- Economic, behavioral, social theories
- Dynamics, equilibrium, spillovers
- Informed by related obs. studies & experiments
- Proposed outcomes based on economic frameworks

Scope of Intervention

- Regulation or market shaping
- Firm, org., or locality
- Service provider vs. consumer in marketplaces

Incremental Innovation

Analysis of Historical Obs/Exp Data

- Off-policy (counterfactual) evaluation
- Heterogeneous treatment effects (HTE) of prior policies
- Combine w/ “foundation models” and/or external data

Refine Interventions

- Informed by theory and analysis
- Policies/algos: targeted treatment assignment, rec systems
- Pilot experiment, poss. recruited subjects

Experimental Design

- Select and validate outcome measures
- Formulate hypotheses and goals
- Design: unit, timing, measures, size, analysis plan
- Adv. experiments: adaptive, staggered rollout, dynamic treatments

Analysis & Decisions

- Revisit outcome measure properties
- “On-policy” evaluation, HTE
- Estimate optimal targeted policies
- Generalizable & tactical insights
- Deployment & decisions
- More experiments?

Potential Goals

- **Demonstration** of impact for further development
  - Is there anything in this category of interventions that does anything for anyone?
- **Insight** for future foundational innovation
  - ATE, HTE
  - Which arms hurt, help, or neither
  - Which outcomes are affected, and identify tradeoffs in outcomes
- **Deployment** decisions (possibly targeted)
  - What is best on average?
  - For whom should we deploy, given resource constraints?
- Welfare of **those in the experiment** vs. use of learning afterwards

Examples of Challenges & Tradeoffs in Meeting Goals

Internal and external validity for relevant hypotheses

What to pre-specify vs. post-hoc

- **Pre-specified** vs. comprehensive w/ **multiple hypothesis testing** vs. **data-driven hypothesis generation** & sample splitting
- Outcome selection, transformation and modeling

Multiple arms, multiple subgroups

- **Lack of overlap** of collected data with post-experiment policy evaluation & optimization leads to **high variance** for policy evaluation and optimization
- **Recommendation system**: Algorithms that prioritize items for each individual. Two levels of "treatment," algorithm and item. Overlap especially challenging.
- Number of arms to test
  - Eggs in baskets: better precision on fewer arms vs. diversified portfolio of arms

Generic versus tailored intervention design

- Find something that works ok for most people *vs.*
- Interventions that work well for some and poorly for others (amenable to finding HTE & targeted policies)

Exploration vs. Exploitation – experimental subject outcomes

Tradeoffs in outcome selection/collection/modeling (cost, response rate, timing)

Design process to evaluate tradeoffs & optimize may use pilots, semi-synthetic simulations, scenario planning

Working Backwards: How Will Analyst Evaluate Policies?

On-policy evaluation:

Compare outcomes across treatment arms

Common challenges

Low signal-to-noise for key outcomes, e.g. fat tails

How to transform or combine outcomes

Selective attrition/non-response (e.g. Lee Bounds)

Interference

Adaptively collected data

Approaches to reduce variance & refine outcomes

- Change the question/outcome
  - Redefine functional form for outcome
  - Combine outcomes into a surrogate index (Athey, Chetty, Imbens, Kang, 2019)
  - Study combination outcomes (e.g. product of two binary outcomes, for ex. Agrawal, Athey, Kanodia, Palikot (2023)) or conditional outcomes
- Model outcomes/adjust for covariates or lagged outcomes
  - Predictive model from historical cross-section or experimental data
    - E.g. study $Y_i - \mu_0(X_i)$ or $Y_i/\mu_0(X_i)$
  - Panel data methods
    - Attentive to staggered rollout issues, recent econometrics literature
    - TWFE, Synthetic Control, Matrix Completion, SDID
- Model/restrict treatment effects
  - How they vary with covariates/predicted baseline (e.g. additive vs. multiplicative)
  - How they shift distribution of outcomes

$$\tau(u) = F_1^{-1}(u) - F_0^{-1}(u) = h(F_0^{-1}(u), \theta) - F_0^{-1}(u).$$

    - Model *h* as, e.g. linear or multiplicative, without functional form of *F*
    - Athey et al (JRSS-B 2023) propose semi-parametric efficient estimation approach, demonstrate benefits with fat tails (see R:parTreat)
- Issues for pre-analysis plans
  - How to spell out plans for model selection, e.g. cross-validated predictive model in control group
  - How many variations of outcomes to pre-specify

Working Backwards: Staggered Rollout Designs

Here: focus on Xiong, Athey, Bayati & Imbens (Mgmt Science, 2023)

- **Planned Analysis:** Estimate ATE using matrix completion w/ staggered rollout design, e.g. Athey, Bayati, Doudchenko, Imbens, Khosravi (Mgmt Science, 2021)
- **Optimization:** Minimize var of ATE

Synthetic control design

- Selects units for (simultaneous) treatment, anticipating synthetic control estimation
- Doudchenko et al. 2021a,b, Abadie and Zhao 2021

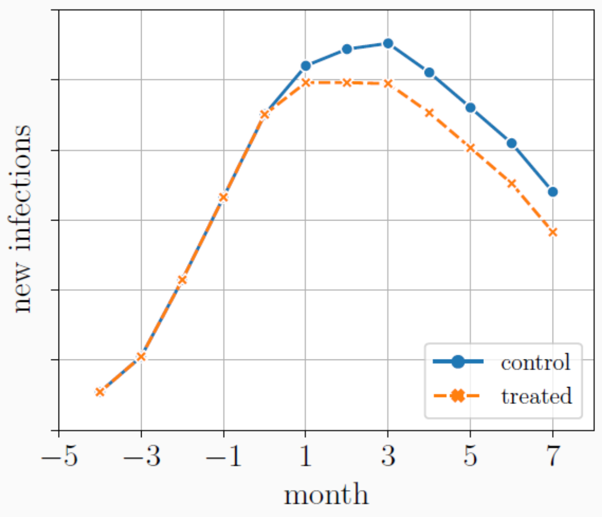
Stepped wedge designs (clinical trials)

- Hussey and Hughes 2007, Hemming et al. 2015, Li, Turner, and Preisser 2018
- No time-varying carryover effects

Estimation of carryover effects

- Minimax temporal experimental design (Basse, Ding, and Toulis 2019)
- Switchback design (Bojinov, Simchi-Levi, and Zhao 2020)

- Estimating treatment effects in panel data with **staggered rollouts**
  - Units $i \in \{1, \cdots, N\}$ observed in time periods $s \in \{1, \cdots, T\}$
  - Design: Treatment assignment $Z_{is} \in \{0, 1\}$
  - Potential outcomes: $Y_{is}(z_{i,s-\ell}, \cdots, z_{is})$ may depend on the history of treatment to date, with known ℓ periods of history that matter
  - Observed outcomes: $Y_{is} = Y_{is}(Z_{i,s-\ell}, \cdots, Z_{is})$
- Staggered rollout designs commonly encountered in observational data:
  - Products/promotions released in different regions at different times
  - State regulations adopted over time



Cumulative effect of treatment for j periods with $\ell = 2$ and $\tau_0, \tau_1, \tau_2 < 0$

Note: another example of multi-dimensional outcomes

Working Backward: Staggered Rollout Experiments

Multiple outcomes aggregated into weighted average for purposes of experiment optimization

Question: How should analyst **design** a staggered rollout experiment?

- How fast should rollout occur?
- How does rollout depend on hypothesized maximum duration of carryover effects?
- How can historical data be used to optimize design?
- Can an **adaptive design**, where analyst updates speed of rollout and termination based on data collected during experiment, improve performance?

Formal objective: Propose experimental designs that optimize the precision of post-experiment estimates of treatment effects

Focus on environment with: Irreversible treatment adoption pattern ($Z_{is} \leq Z_{i,s+1}$)

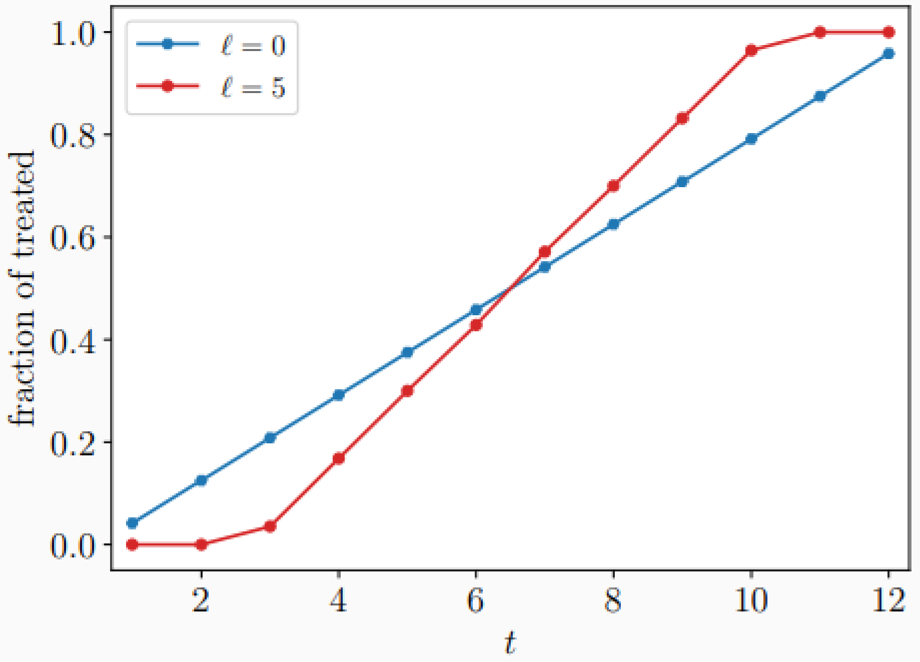
$$
\begin{array}{c|ccccc|}
 & \multicolumn{5}{c}{\text{Time}} \\
\text{SF} & 1 & 1 & 1 & \cdots & \cdots \\
\text{BOS} & 0 & 1 & 1 & \cdots & \cdots \\
\text{ATL} & 0 & 0 & 1 & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & & \\
\vdots & \vdots & \vdots & \vdots & &
\end{array}
$$

0 denotes control and 1 denotes treated

Characterizing the Solution to the (non-adaptive) Experimental Design Optimization Problem

Non-adaptive experiments: N and T are set, and treatment decisions are made, pre-experiment

- Assume after experiment will use GLS to estimate instantaneous and lagged treatment effects from nonstationary observed outcomes
- Analytical optimality conditions for the designs that maximize linearly combined precisions of estimated instantaneous and lagged effects
- Propose an algorithm to choose a treatment design based on the optimality conditions. The design has two features
  - $\Rightarrow$ Fraction of treated units per period takes an S-shaped curve: Treatment rollouts slowly at the beginning and end, and quickly in the middle
    - Bigger ℓ leads to more pronounced S
  - $\Rightarrow$ This rollout pattern is imposed for each stratum of units with the same observed and estimated latent covariate values



Adaptive Experimental Design for Staggered Rollouts

Goal: Most precisely estimate average treatment effects (i.e., increase $\mathrm{Prec}(\hat{\tau}_0; Z)$) with valid inference, while using the least sample size

Two adaptive decisions:

- Stop the experiment early if the desired precision is achieved (i.e., max duration is $T_{\max}$, and duration $\tilde{T} \in [T_{\max}]$ is a random variable)
- Speed of treatment rollout for the next time period is determined after each period's outcomes are collected

Designing the Adaptive Experiment

Design choices

1. Treatment design (rollout speed)
   - Adaptively choose as we gather more information about σ^2 during the experiment
2. Termination rule

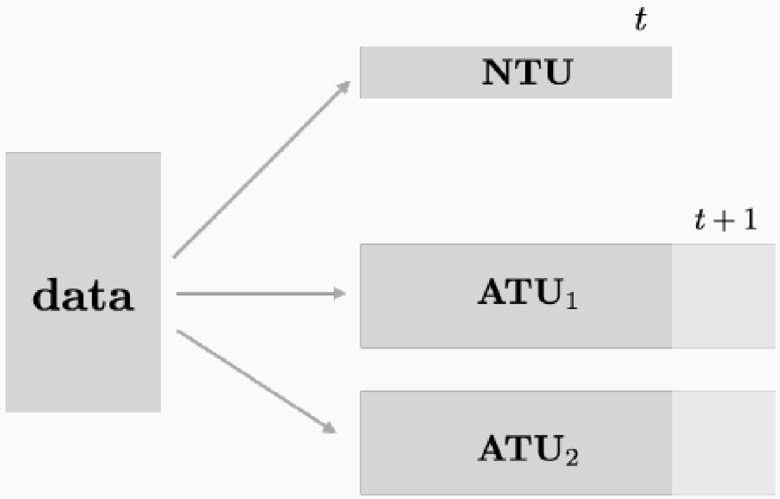
Design in order to **enable efficient estimation** and valid inference for treatment effect **after** experiment

- Use as many observations as possible

Propose the Precision-Guided Adaptive Experiment (PGAE) algorithm

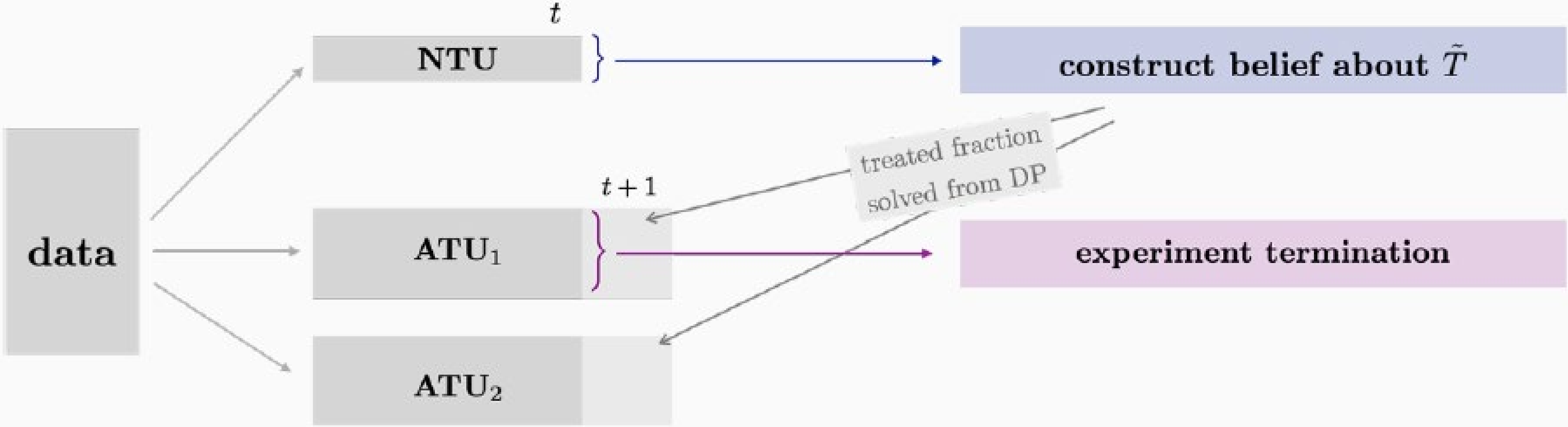
- Simultaneously achieves more efficient rollout and stopping, with efficient post-experiment estimation
- Uses **sample splitting** and **dynamic programming**

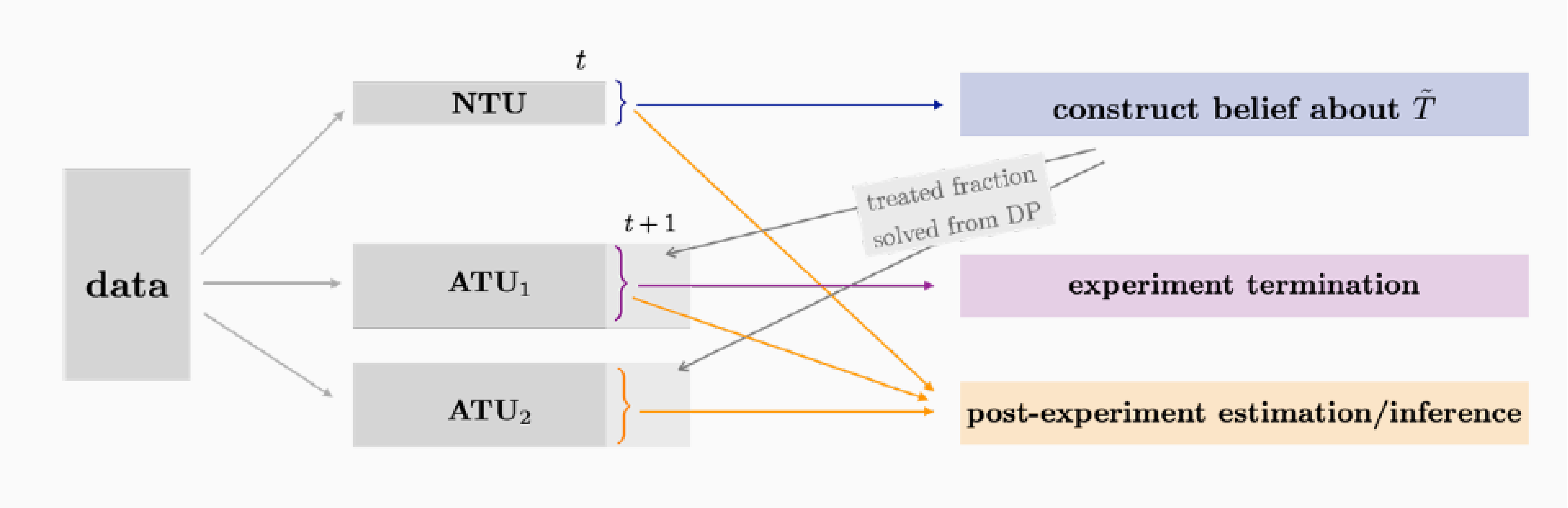
- NTU: Treatment design set pre-experiment (a small set)
    - Set as $\omega_{bm,s} = (2s-1)/(2T_{\max})$ (optimal solution for $T_{\max}$)
- ATU: Treatment design chosen adaptively



At time t, optimize ω_{t+1} for ATU_1 and ATU_2 through dynamic programming (DP)

- In the DP, no intermediate cost and terminal cost is the precision at termination, i.e., $\mathrm{Prec}(\hat{\tau}_0; Z_{:,1:\tilde{T}}) = (N\tilde{T}/\sigma^2) \cdot g_{\tau}(\boldsymbol{\omega}, \tilde{T})$
- Solve ω_{t+1} from DP based on the belief about $\tilde{T}$





- The adaptivity of the design, with the termination time depending on early values of the outcomes, comes at no cost in the estimation of τ_0
  - Compare with a series of experiments with the same distribution of termination times, the average variance of $\hat{\tau}_{\text{all},\tilde{T}}$ is the same
- Adaptive treatment decisions improve the estimation precision of τ_0

Semi-synthetic application: Adaptive staggered rollouts

Imaginary experiment: city-level vaccine campaign to fight influenza

Data: month-city observations on influenza aggregated from MarketScan insurance data; DGP based on data from October to April from 2007-2017

- Artificially assumes flu season lasts longer when analyzing longer potential experiment lengths

Results: Adaptive design lowers estimation error by 20% at lower experiment cost

- Leads to **substantial early stopping**
  - When max possible # months is greater than 7, stop at less than half the max # months.
- Adaptive rollout speeds up as algorithm predicts an early finish

Note: This exact method has not been implemented in practice to my knowledge.

- Industry just beginning to move from ad-hoc midstream decisions, simulation-based planning or heuristics
- Illustrates the ideas of experimental design as a formal optimization problem.

Working Backwards: Takeaways

Experimental Design as an Optimization Problem

Adaptivity improves performance but must be carefully designed to avoid costs

Non-adaptive approach

- Challenge:
  - Power; heterogeneous units & time shocks
- Analysis at the end:
  - ATE using post-experiment outcome modeling of latent time and unit effects
- Design choices:
  - (Stratified) rollout of treatment, length of experiment
- Optimization:
  - Algorithm for rollout design, characterization of solution (manually compare lengths of exp.)

Adaptive approach:

- Design choices:
  - Stopping time (since data-driven, becomes stochastic)
  - Rollout of treatment
- Optimization:
  - Structure algorithm so that data can be re-used for estimating treatment effects; adaptive based on estimates of variance, NOT estimates of outcomes
  - Algorithm optimizes rollout & stopping based on current beliefs about variance

Adaptivity tradeoff:

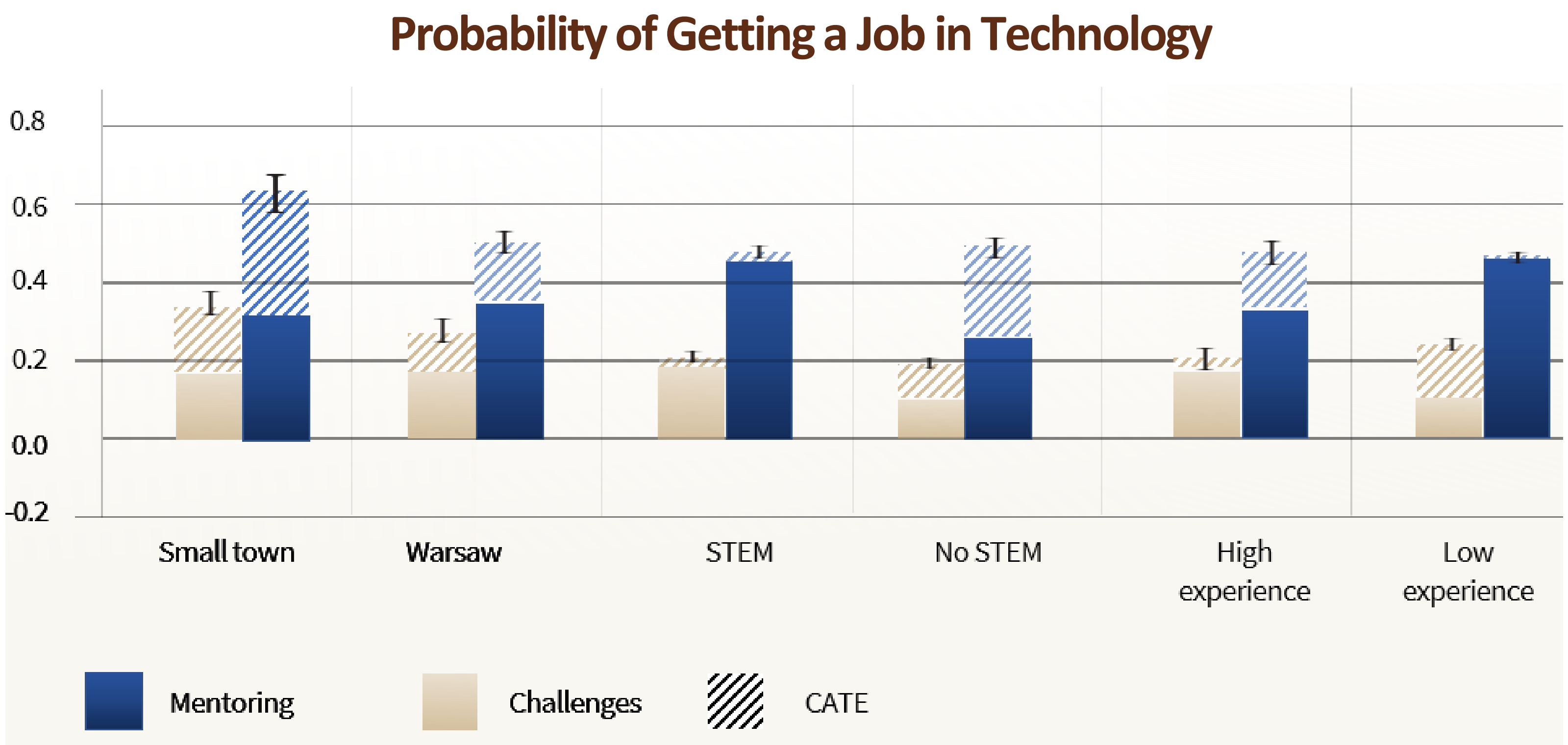
- Optimizing DURING the exp. ***potentially*** sacrifices ability to analyze AFTER
  - Xiong et al shows careful sample splitting and design can ameliorate tradeoff- **no eff. loss!**
  - Xiong et al is adaptive based on learned variance, not learned treatment effect
- Outcomes in early periods correlated w/ assignment in later periods
- Later, we discuss statistical issues with analyzing adaptively collected data, see e.g.:
  - Andrews, Kitagawa, McCloskey (2019), Hadad, Hirschberg, Zhan, Wager & Athey (PNAS, 2021) & Zhan, Ren, Athey & Zhou (KDD, 2021, Mgmt Sci 2023), Deshpande, Mackey, Syrgkanis, Taddy (2019), Howard, Ramdas, McAuliffe, Sekhon (2021), etc.

Goal: Estimate & Deploy Targeted Treatment Assignment Policy

Heterogeneous Treatment Effects

1. Pre-specified or hypothesis driven
2. Comprehensive with MHT corrections
   - See e.g. List, Shaikh, Xu 2017
3. Data-driven hypothesis generation
   - E.g. Causal trees (Athey & Imbens, 2016), causal forests (Wager & Athey, 2018)

Can also consider het. in outcomes (e.g. Ludwig, Mullainathan & Spiess 2017)



Athey & Palikot (2023) created & implemented Challenges program in collaboration with DareIT to help women transition sectors into IT in Poland

2 programs, 2 distinct randomized experiments, 2 control group baselines. Whiskers show standard errors for CATEs.

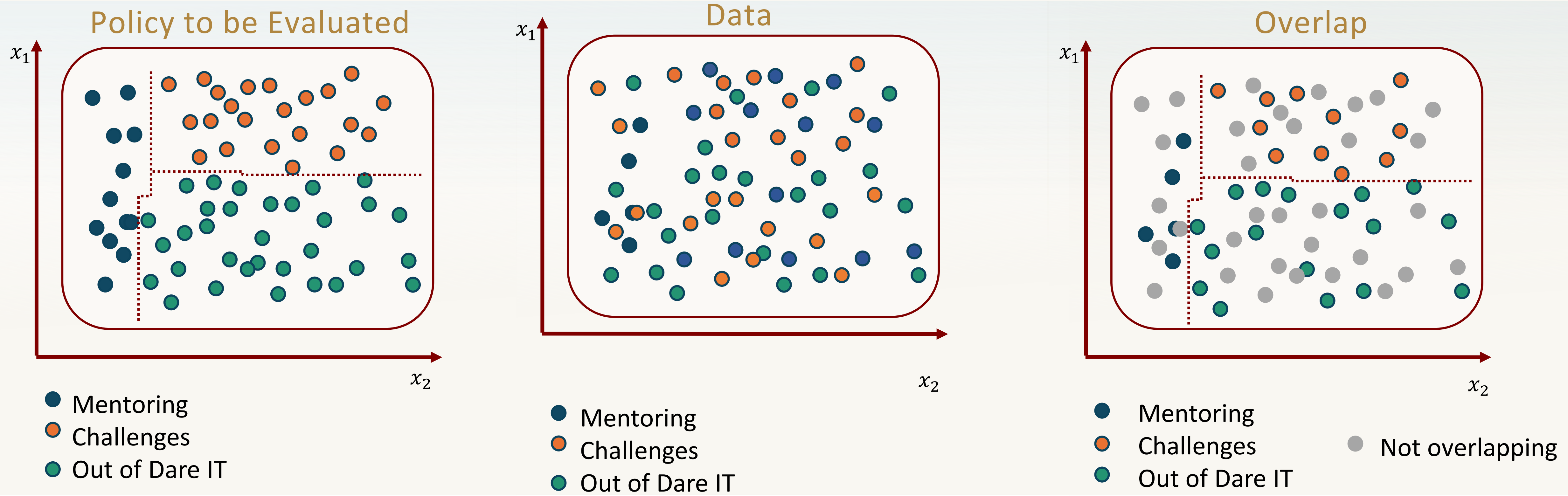
Off-policy evaluation

1. Policy assigns an arm based on covariates and overall capacity $\pi: \mathcal{X} \times \mathcal{Q} \rightarrow \mathcal{A}$

   Potential outcome: $Y(a)$, Expected value of policy: $V(\pi) = \mathbb{E}_X[Y(\pi(X,q))]$

2. Off-Policy Estimators

   $\hat{V}(\pi)$ can be estimated using sample means for overlapping observations (simple RCT)
   Our application blends 2 RCTs, samples; outcome modeling, propensity weighting, or AIPW (w/cross-fitting)



Estimating and Evaluating Treatment Assignment Prioritization Rules

Estimate optimal policy

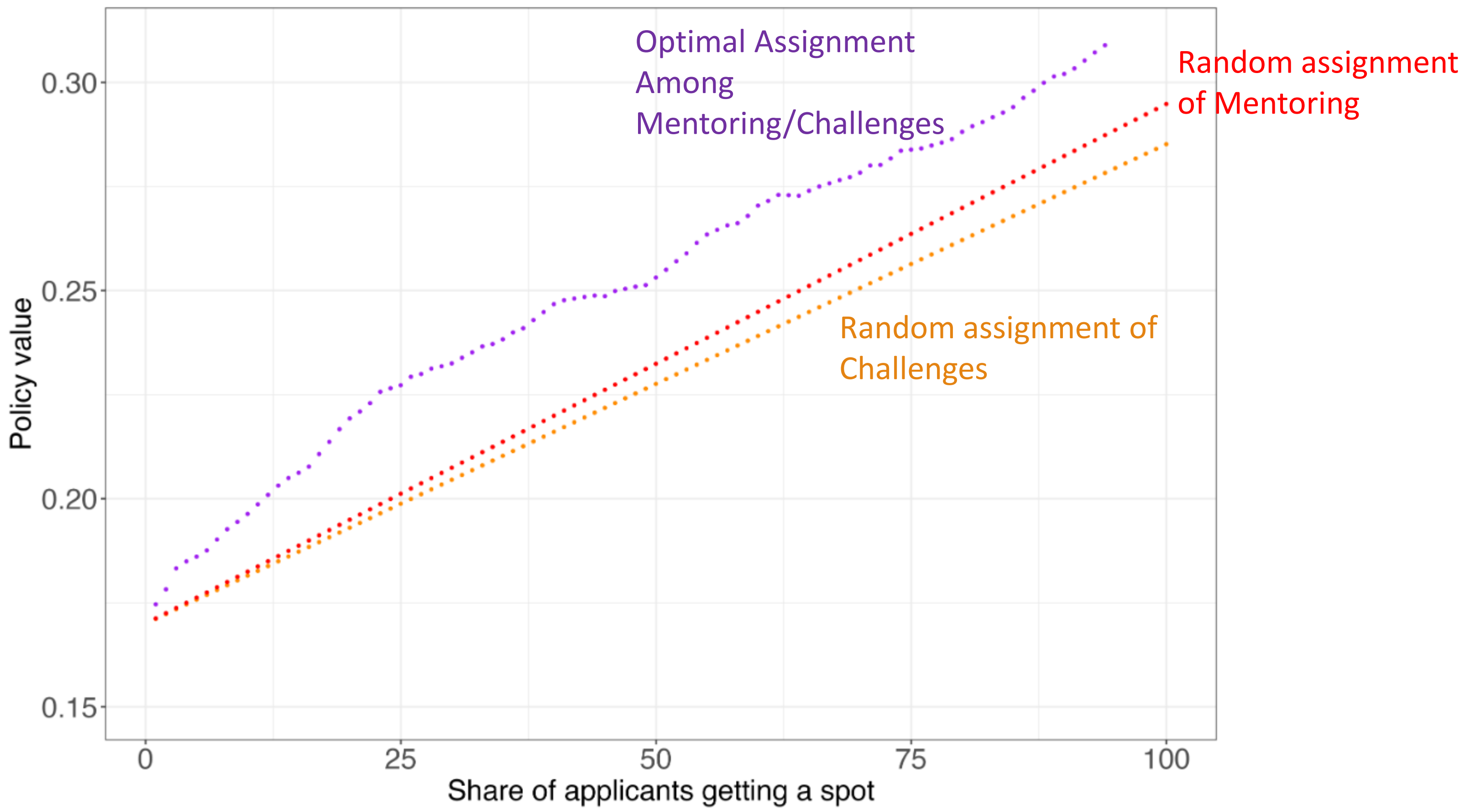
- For each program a and cov. x, estimate $\hat{\tau}_a(x)$
- Optimization algorithm:
  - Prioritize the program and indiv characteristics that are most effective given capacity

Evaluate using test set

For more on methods, see also:

- Sverdrup, Wu, Athey & Wager (2023) & software in R:grf
- Yadlowsky, Fleming, Shah, Brunskill, and Wager (2021)

The value of targeting as a function of program capacity (*q*)



Off Policy Evaluation & Estimation: Important Points from Design Perspective

Overlap in historical data critical for variance of estimates:

$$Var(\widehat{\mathbb{E}}_{X_i}[Y_i(\pi(X_i))]) = \frac{\sigma^2}{N}\frac{1}{Pr(A_i=\pi(X_i))}$$

- Historical deterministic targeted policy -> lack overlap

Policy estimation $\hat{\pi}$: by evaluating many policies (Athey & Wager, 2021; Zhou, Athey & Wager, 2023; grf, policyTree in R)

- Quality of the policy estimate is worse if you optimize over a larger/more complex policy set Π
- Is proportional to the **largest** variance in set of considered policies:

$$\sup \pi \in \Pi \; Var(\widehat{\mathbb{E}}_{X_i}[Y_i(\pi(X_i))])$$

Working backward:

- Adaptive design/iterative exp assigns treatments based on covariates to **manage overlap** for (uncertain) future
- Policy learning during & after may **restrict policy class**, e.g. drop arms or use tree policies (Athey et al, 2022)
- Theory: Krishnamurthy, Zhan, Athey, Brunskill, 2023; Krishnamurthy, Propp, Athey, 2023;

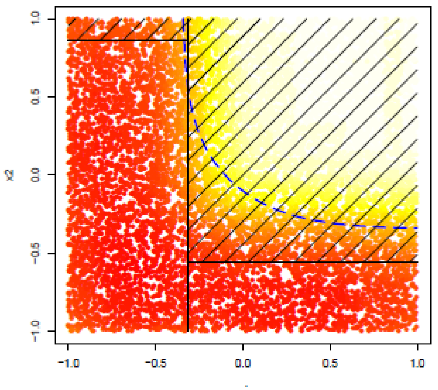
Allowable Policies

Nonparametric, e.g. causal forest (Wager & Athey 2018; Athey, Tibshirani & Wager 2019), based on $\hat{\tau}_i(X_i) > 0$

- Theoretically higher value
- Overfitting, hard to describe, non-monotone
- See eg Manski (2004), Hirano & Porter (2009), Stoye (2009), Kitagawa & Tetenov (2018)

Tree policy

- Athey & Wager, 2021; policyTree
- May do better in practice (regularize)
- Easy to describe, track segments



Tradeoffs between multiple outcomes

Targeting to improve one outcome might improve or hurt a different outcome

- May be tradeoff at individual level
- May be that you treat different people to maximize different outcomes

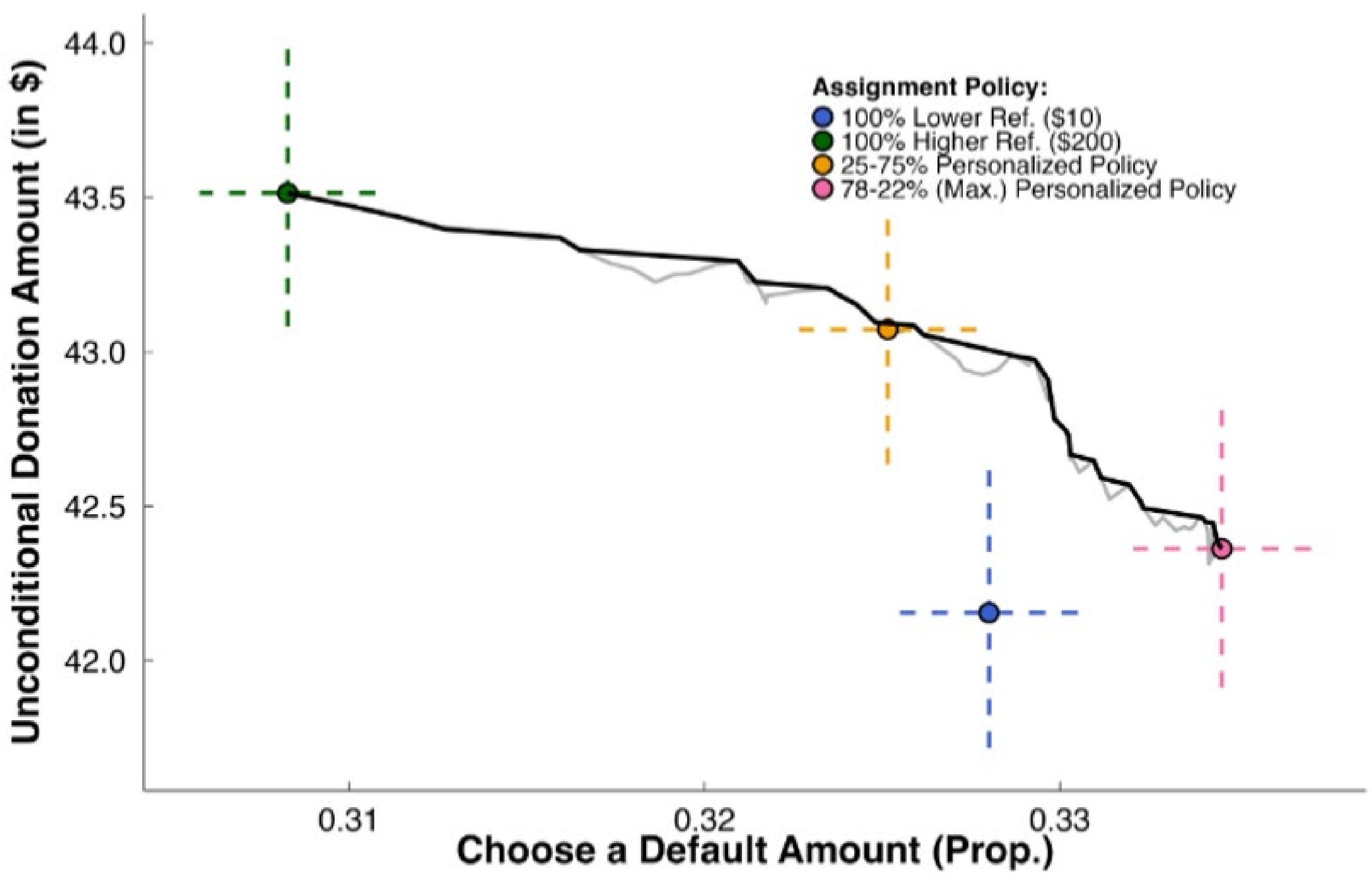
One option: maximize weighted sum of outcomes

- Constrain avg. of each outcome? For subgroups?
- Upweight vulnerable subgroups?

Signal-to-noise affects tradeoff

- May be much better for some outcomes than others, often better for short-term/simple engagement outcomes like clicks

Figure 10: Trade-Off between Unconditional Donation and Choosing a Default



Experiment randomizing the buttons for charitable donations for hundreds of thousands of PayPal donors, removing intermediate $75 button and replacing with either $10 or $200
Athey, Koutout, and Nath, 2024 WP

Iterative Experimentation to Develop Targeted Treatment Assignment Policies

Iterative experimentation and policy estimation

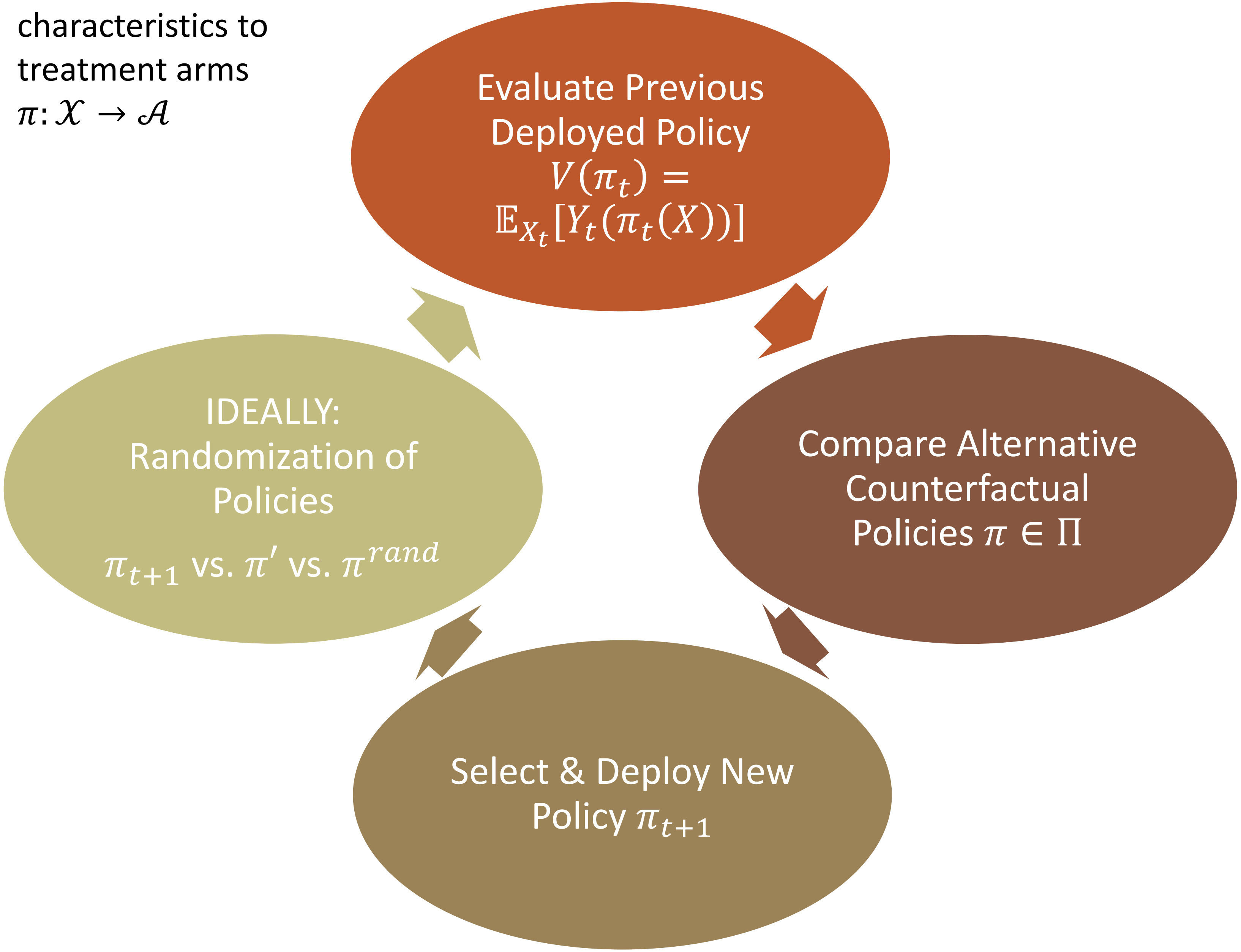
Collect data via randomization to do off-policy evaluation of alternative targeted policies

- See e.g. Hitsch et al., 2024; Simester et al., 2020a; Yoganarasimhan et al., 2023

Deploy and test performance

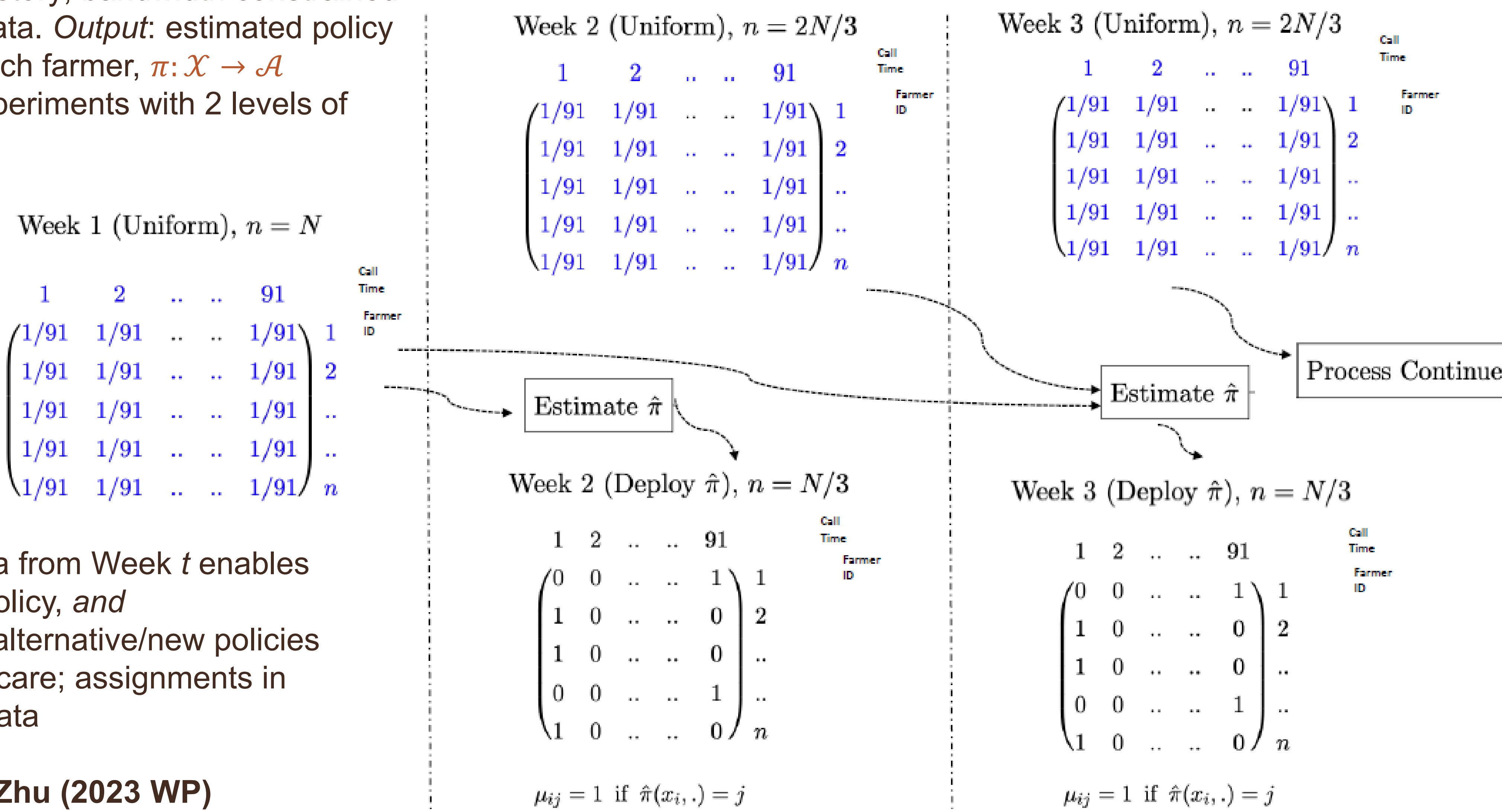
- Hitsch et al. (2024)
- Simester et al., 2020a
- Yang et al., 2023

Recall: policies map characteristics to treatment arms $\pi: \mathcal{X} \to \mathcal{A}$



Iterative Experimentation: Sequence of On-Policy Evaluation & New Policy Estimation

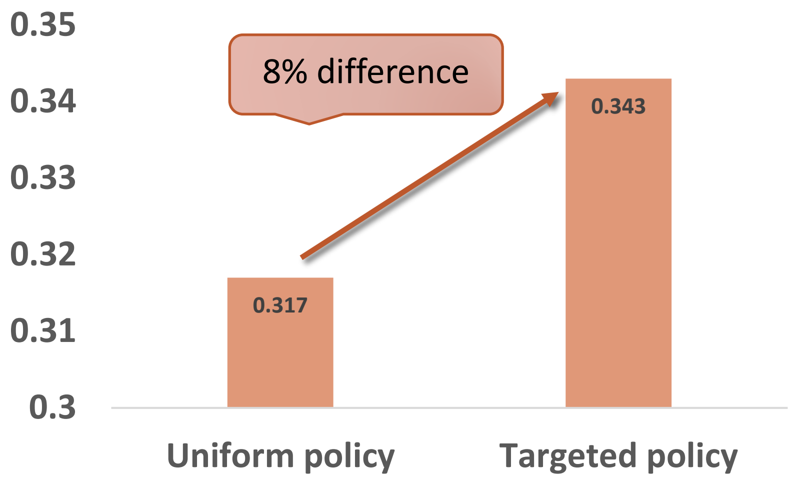
Goal: Deploy algorithm assigning call times to farmers based on engagement history; bandwidth constrained
Algorithm: *Input*: past data. *Output*: estimated policy assigning call times to each farmer, $\pi: \mathcal{X} \rightarrow \mathcal{A}$
Design: Sequence of experiments with 2 levels of randomization.

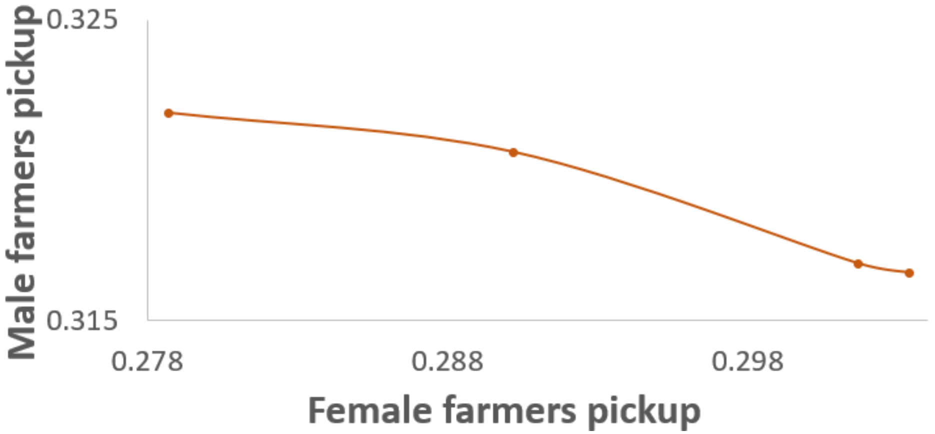


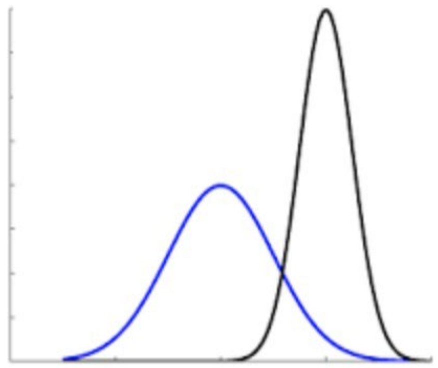
Benefits of design: Data from Week *t* enables evaluation of deployed policy, *and* estimation/evaluation of alternative/new policies
Caution: Pool data with care; assignments in week *t* depend on past data

Athey, Cole, Nath, and Zhu (2023 WP)

Impact of Personalization in Call Times







Value of Targeting

- Estimate targeted policy under capacity constraints (new methods)
- 8% gain in engagement.
- Potential to reach 26,000-33,000 additional farmers with educational content.

Tradeoffs between Outcomes

- Scarce bandwidth per hour
- Female farmers lower average engagement. Can re-optimize giving them greater weight
- Can improve engagement from women by 9% if we reduce men's engagement by 1.7%

Shocks/external validity

- On-policy estimates worse than off-policy predictions
- Show: Pref./Tech. Shocks.
- Distribution Shifts.
- Weight more recent data for better perf. (tradeoff w/ variance)

Adaptive Experiments: Bandits & Contextual Bandits

Bandits: Goals

1. Regret
2. Learn good policy
3. Hypothesis testing/ precise estimation

Key Tradeoffs

Exploration vs. Exploitation

Exploitation targets optimal policy but risks low overlap with it

Overlap w/ optimal policy vs. overlap with all policies to be evaluated

Low "Regret" – DURING Experiment

- Dropping harmful arms (e.g. medical)
- Expected outcomes of subjects DURING experiment

Policy Learning - AFTER Experiment

- "Policy Learning" (Kasy & Sautmann) or "Simple Regret" (ML)
- Very little theory for *contextual* bandit (Qin and Russo, 2022, Krishnamurthy, Zhan, Athey & Brunskill, NeurIPS 2023)
- When does it help? Pure RCT puts lower bound on overlap

Tight standard errors, specific hypothesis tests at END

- Best arm/policy vs. control?
- Policy learning and pure RCT keep exploring forever—always some benefit to more learning
- Real world policy learning *often doesn't converge in time*
- Policy makers are going to pick one choice and want to know how good it is, compare to baseline, do budgeting/planning etc.
- To optimize need to be aware what happens after experiment ends
- Adaptivity requires special treatment for hypothesis at end

Tradeoffs: bandits can be designed to manage

Goal: Adaptive pilot to inform email design for nudge

Want to find the best arms at the end

See Rosenberg et al 2021 tutorial: https://gsbdbi.github.io/ml_for_behavioral_science/index.html

Practitioner's guide: https://www.gsb.stanford.edu/faculty-research/publications/practitioners-guide-designing-adaptive-experiments

From: EnforcementWarning@finance.nyc.gov **C = control**

Subject: Take Action Now to Avoid Booting or Towing

NYC Department of Finance

Dear Customer,

This is a courtesy notice to let you know that at least one vehicle associated with your email address has accrued nearly $350 in unpaid parking or camera summons judgment debt. A vehicle with more than $350 in judgment debt may be booted or towed to recover the outstanding amount due - and if the owner of these plates owns other vehicles, those vehicles may also be subject to booting or towing.

We urge you to visit www.nyc.gov/citypay to pay or dispute the out- standing summonses associated with the vehicles or vehicles shown below. Please note that you may have other vehicles at risk of booting. You can search all of the plates you own at www.nyc.gov/citypay.

For more information about disputing a ticket, including who is eligible to request a hearing, you may visit www.nyc.gov/disputeticket.

This is the only email you will receive on this matter. If you have questions, you may contact us at 311. This mailbox is unattended. To contact the NYC Department of Finance, visit www.nyc.gov/contactfinance. Contacting us without payment will not prevent your vehicle from being booted. Please visit a DOF business center if you need immediate assistance or to establish a payment plan if you qualify.

Sincerely,

The New York City Department of Finance

Plate	State	Type
ABC6789	NY	PAS

From: NYCparking@finance.nyc.gov **A**

Subject: Your vehicle is at risk of being booted

Dear NYC Driver,

A [vehicle type] with plate [plate #] associated with this email has unpaid parking fines. **If you or the vehicle owner do not pay or dispute by [DATE]*, this vehicle will be booted or towed.**

- **To pay tickets, click Pay my parking ticket and use this violation number: [oldest summons #]**
- If you think this is a mistake and want to dispute a ticket, click Dispute a ticket
- If you are interested in a payment plan, click Learn about payment options

Need some help? We answer questions at Contact Finance or you can come by in person to one of our Business Centers.

Remember, you or the vehicle owner must take action by **[DATE]*** or this vehicle and any other associated vehicle(s) will be at risk of being booted or towed.

Thank you in advance,

NYC Department of Finance

P.S. This email address was previously used to pay a parking ticket for the vehicle(s) listed below. All vehicles registered to the owner are at risk of being booted or towed.

*Booting or towing may occur before listed date

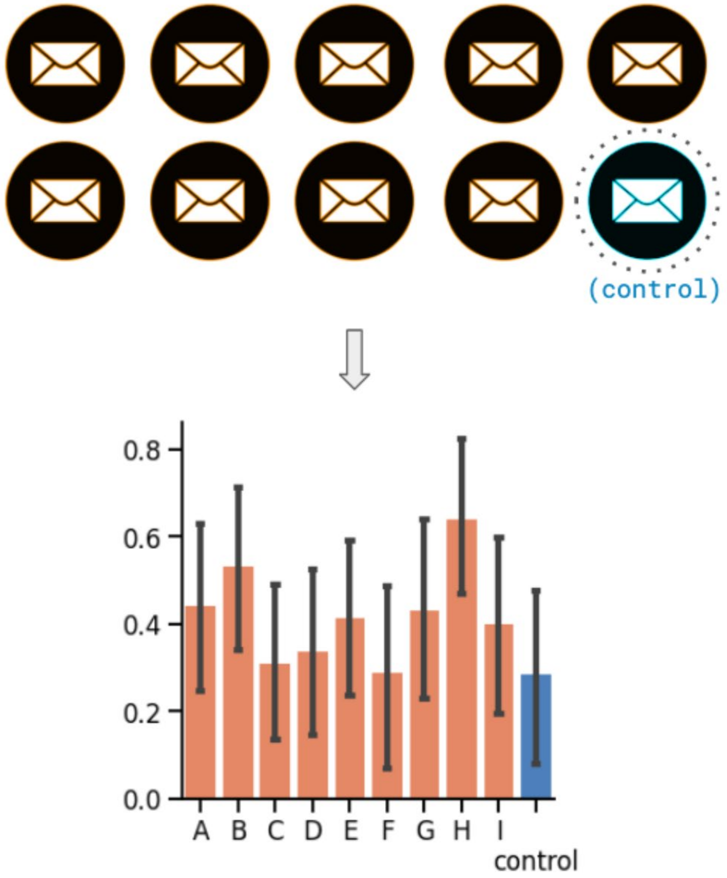
Plate	State	Type
CJK 876	NY	PAS

NYC Department of Finance

Prototype B:
No personalization in 1st sentence

Experiments in practice: Email nudges
Challenges with many arms

Suboptimal design: Ten-armed RCT
- Requires large sample size.
- Some arms may be bad enough that precise estimates are not useful.
- Some arms may be good but similar, irrelevant which is chosen.

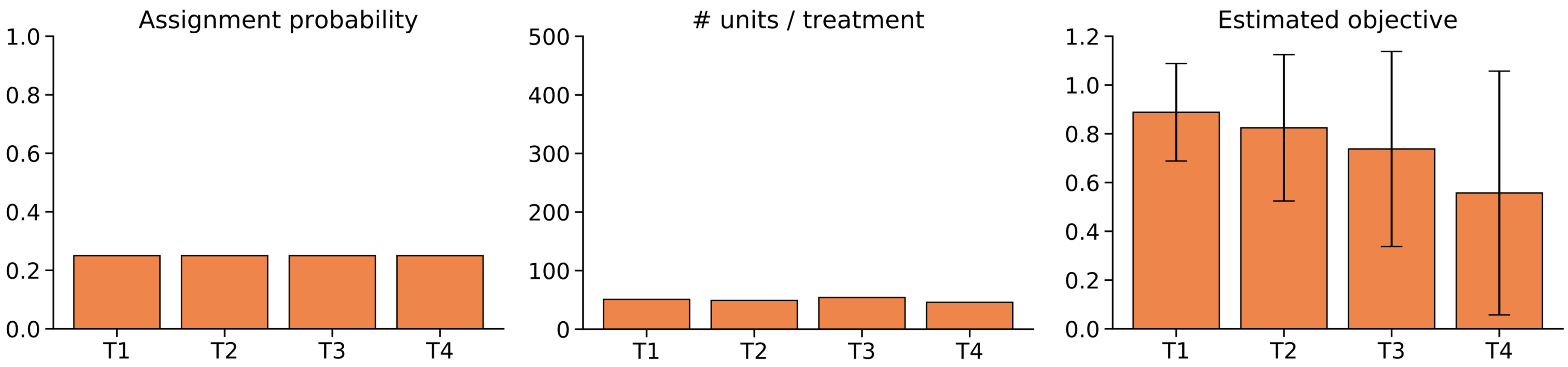


Adaptive experimentation

Learning by design

Multi-armed bandits

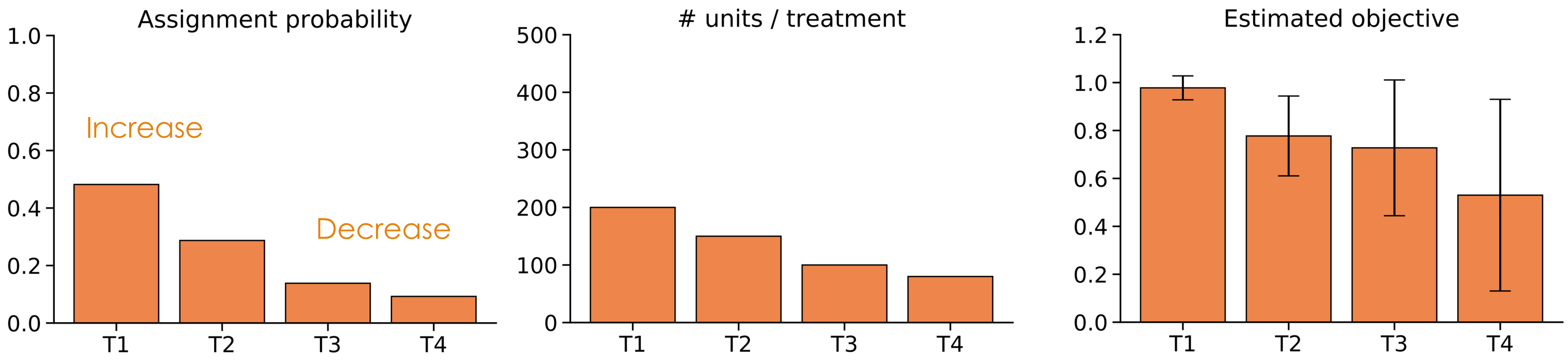
An example of adaptive design



Step 1: At the *beginning* of the experiment, assign treatments uniformly at random

Adaptive experimentation
Learning by design

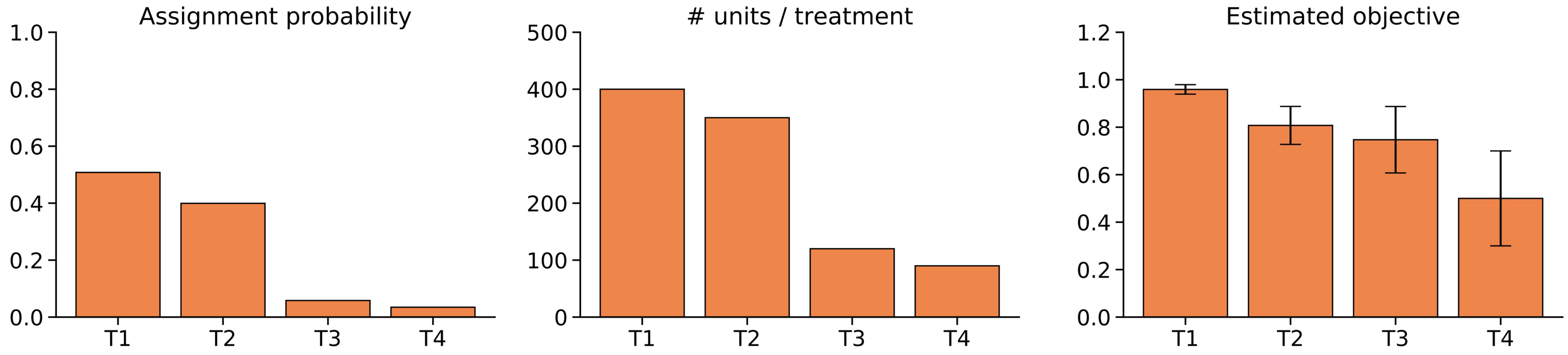
Multi-armed bandits
An example of adaptive design



Step 2: Once some data has been collected, increase the probability of assignment to more promising arms.

Adaptive experimentation
Learning by design

Multi-armed bandits
An example of adaptive design



Step k: Repeat this procedure in batches, increasing probabilities of assignment as we become more certain about which treatments are good.

Adaptive experimentation
Learning by design

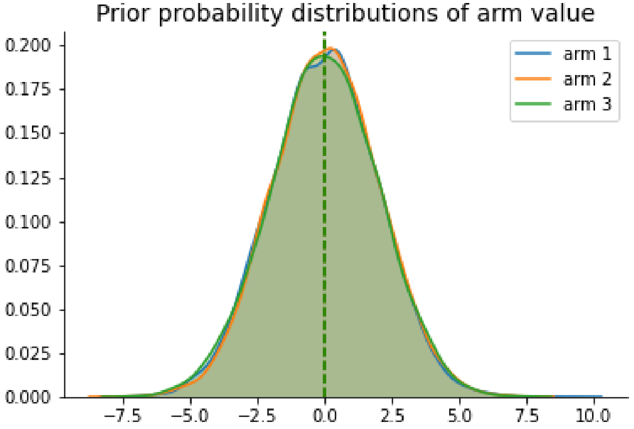
Thompson sampling
A Bayesian multi-armed bandit algorithm

1. Start with a **prior** distribution on arm values.

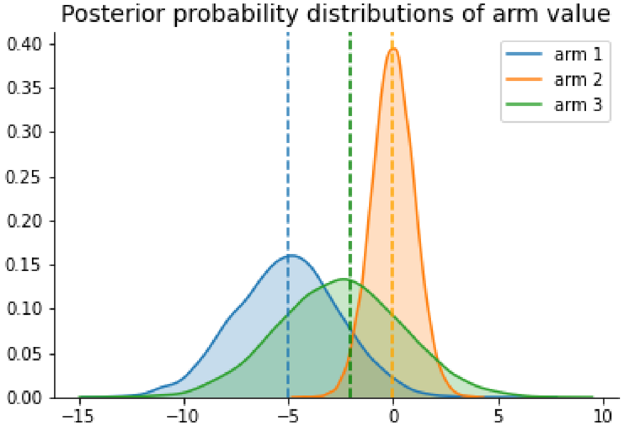
2. Collect first batch of data by assigning treatments uniformly at random.

3. Observe outcomes and **update the posterior** distribution.

4. Next batch, assign treatments according to their **posterior probability of being optimal**. (Repeat)



P(arm 1 is optimal) = ⅓
P(arm 2 is optimal) = ⅓
P(arm 3 is optimal) = ⅓

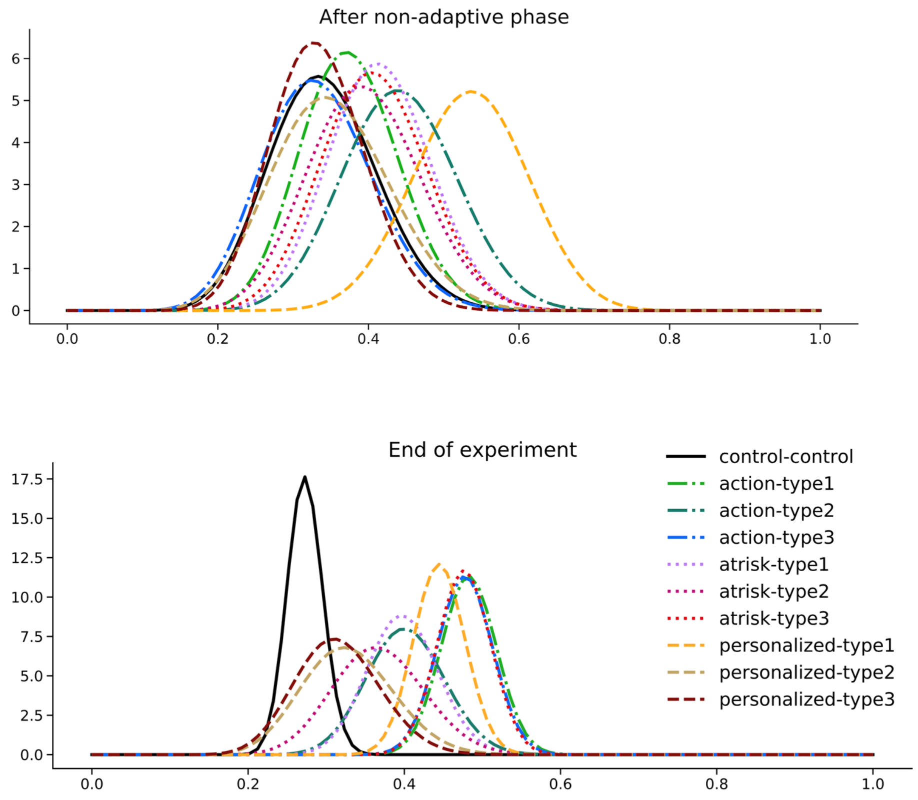


P(arm 1 is optimal | Data) = 0.05
P(arm 2 is optimal | Data) = 0.70
P(arm 3 is optimal | Data) = 0.25

The **Thompson Sampling** heuristic dictates these assignment probabilities.

Has good properties balancing exploration & exploitation.

Adaptive exerimentation
Pilot experiment result for email nudge

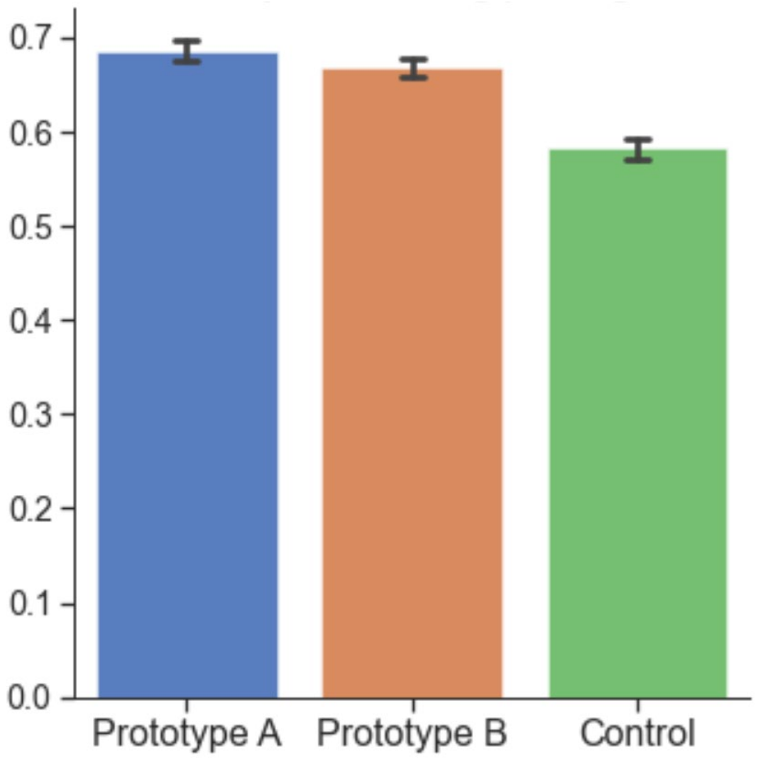


- Data was collected via an adaptive experiment, where more units assigned to the treatments that were doing better
    - Modified Thompson sampling algorithm (ensuring a minimum # of obs allocated to control).
    - Better: “Exploration sampling” to target policy learning. Here little difference (didn’t converge).
- Experiment allowed us to learn that
    - Control is indeed suboptimal; three prototypes “in the lead”.

[Left] Snapshots of posterior distributions of the probability that a participant will engage with each email prototype after initial phase of experiment and at the end of the experiment.

Experiments in practice: Email nudges for NY Dept of Finance

Results from main experiment



Pilot experiment informed selection of two prototypes

(note: due to organizational constraints, these ended up being slightly different from pilot winners).

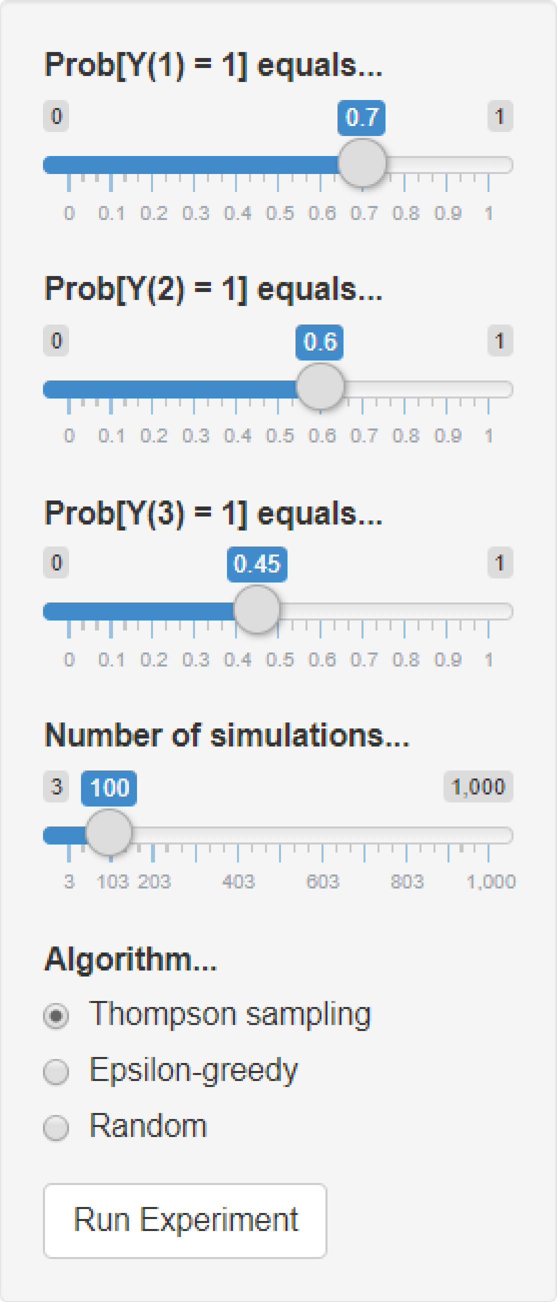
Main experiment was an RCT with ~22k actual NYC drivers.

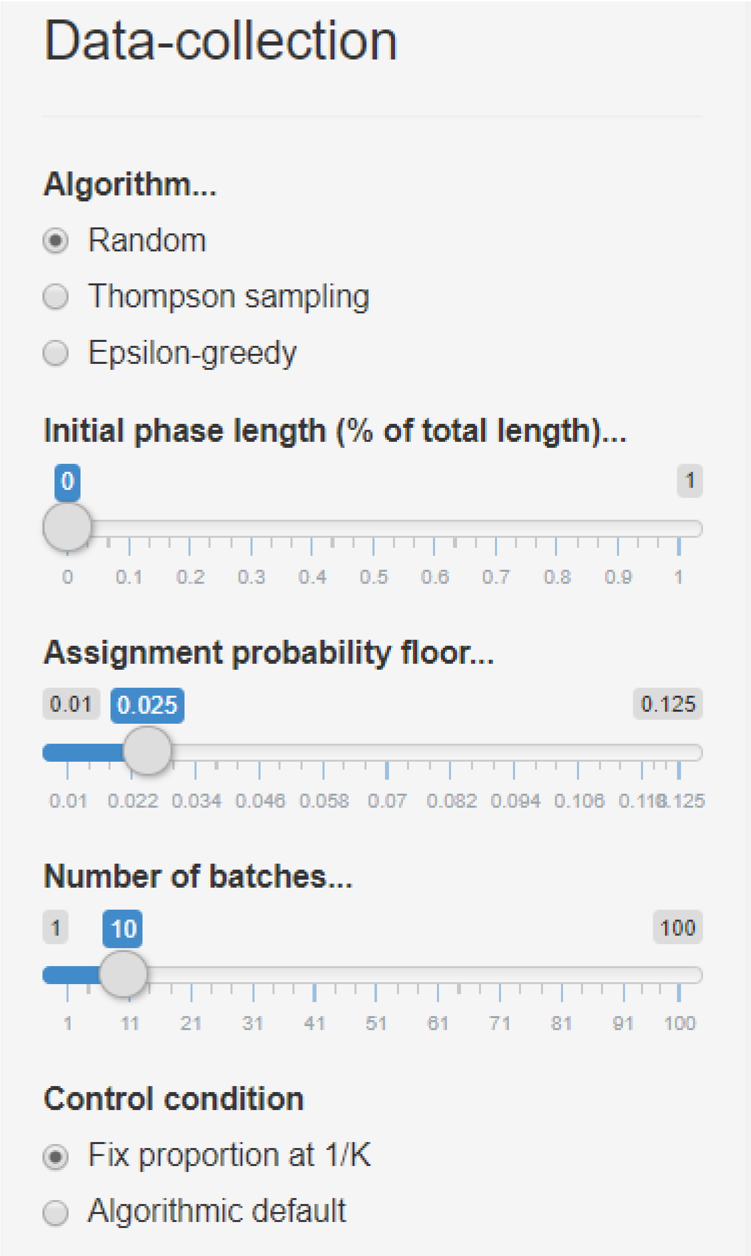
[Left] Main experiment results, showing average outcome (defined as indicator of payment or dispute within two weeks of receiving the email).

Check out our shiny app!

You can try different arm means, and different algorithms

https://gclab.shinyapps.io/bernoulli-bandit/





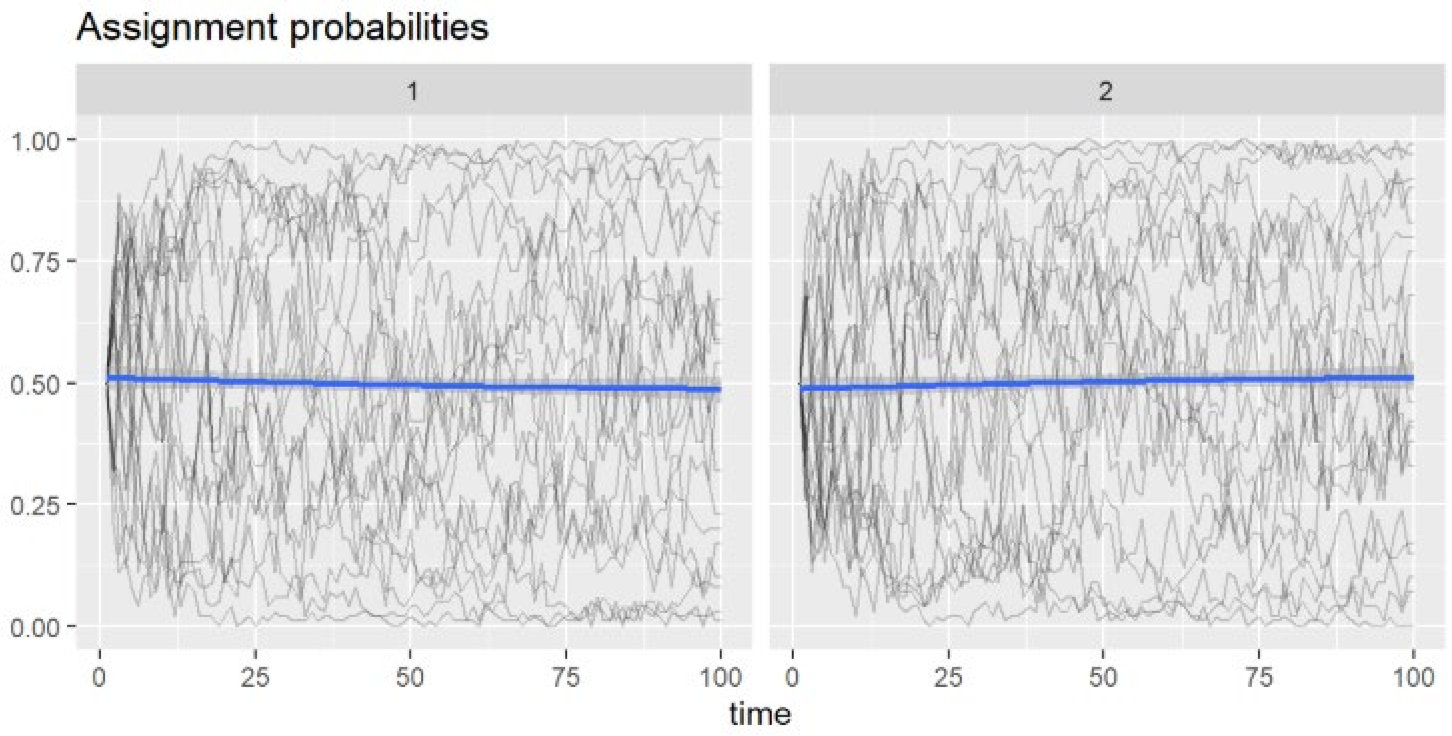
Instability is often a problem

Bandit may not have converged before you run out of money/experimental units

Heuristics for stopping can be used if you have sufficient budget (but creates an additional form of adaptivity to address in analysis)

Recall earlier discussion of adaptivity after experiment, e.g. Hadad, Hirschberg, Zhan, Wager & Athey, PNAS 2021 & references therein

Case of equally good arms



Testing Hypotheses After Bandit

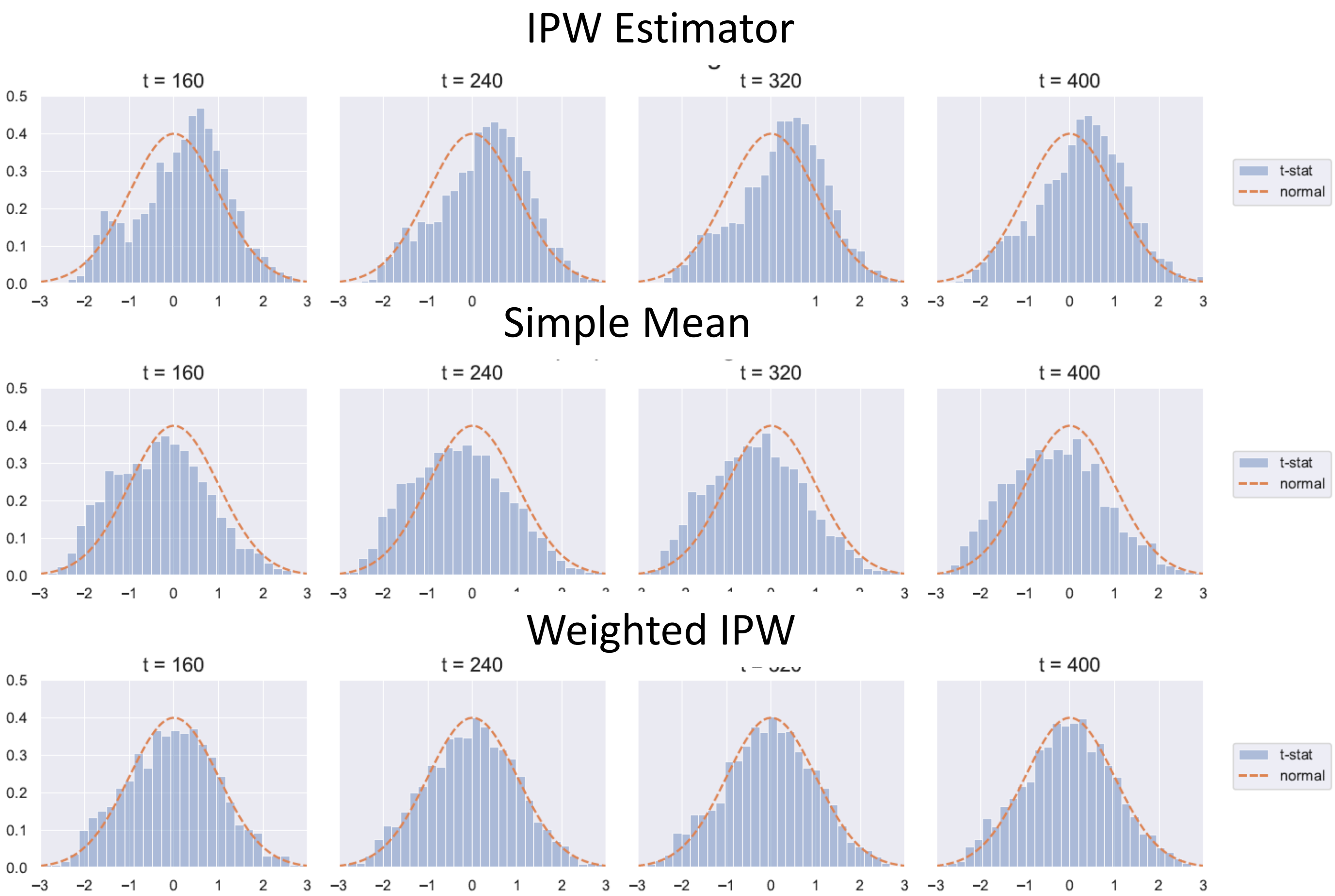
- Simple mean is biased estimator of true arm mean, and estimator is often multi-modal
  - If an arm does badly initially, receives lower assignment probability later, upweighting initial bad outcomes—low mean, low sample size go together
- Weighting by inverse of assignment probability (IPW) restores "equal weighting of each batch," eliminates bias
- IPW is not asymptotically normal (variance & mean pf estimates are still related)

Solution: adaptively weight data to stabilize variance, retain consistency, restore normality (Hadad et al, 2021 PNAS)

Introduce *evaluation weights* $h_t(w)$.

Adaptively-Weighted Augmented Inverse Propensity Score Estimator

$$\widehat{Q}_T^{AW}(w) := \sum_{t=1}^{T} \frac{h_t(w)}{\sum_{s=1}^{T} h_s(w)} \left\{ \frac{\mathbb{I}\{W_t = w\}}{e_t(w)} Y_t + \left(1 - \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}\right) \hat{\mu}_t \right\}$$



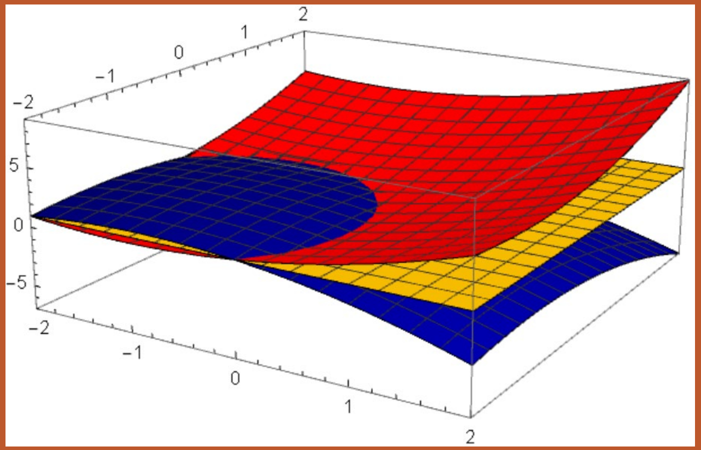
See also Andrews, Kitagawa, McCloskey (2019), Zhan, Ren, Athey & Zhou (KDD, 2021, Mgmt Sci 2023), Deshpande, Mackey, Syrgkanis, Taddy (2019), Howard, Ramdas, McAuliffe, Sekhon (2021), etc.

More details in Appendix to this Presentation

Adaptive Experiments and Targeted Treatments

System interacts with its environment, taking actions or assigning treatments

Outcomes for different arms depend on contexts

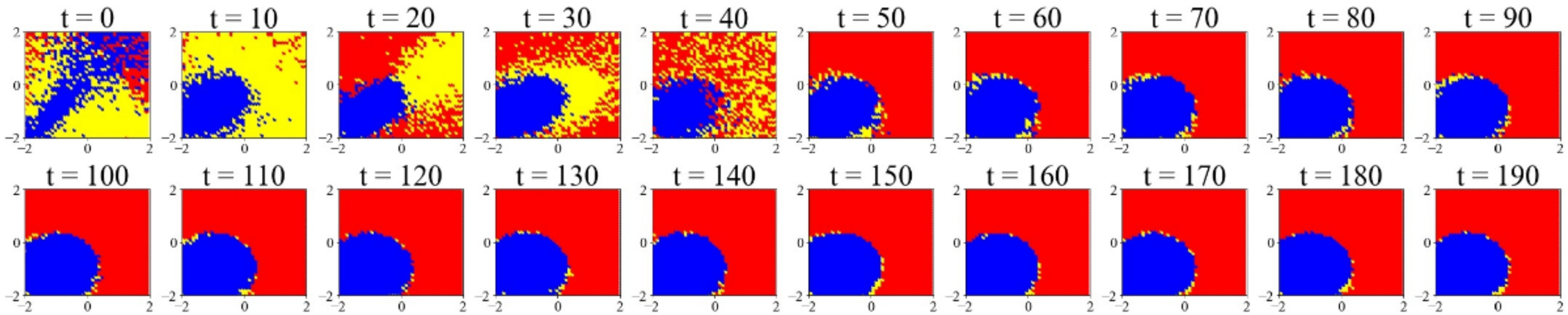


Contextual bandits:

- Learn a targeted treatment assignment policy mapping from individual characteristics to treatments

$$\pi: \mathcal{X} \rightarrow \mathcal{A}$$

- Consider batches of subjects
- **Outcome modeling** approach: After each batch, estimate a model mapping characteristics to (counterfactual) outcomes for each treatment $\hat{f}_k(x, a)$
- Then apply bandit heuristics as each *x* in next batch arrives



Real-World Applications of Contextual Bandits

Contextual Bandits in a Survey Experiment on Charitable Giving: Within-Experiment Outcomes versus Policy Learning

Athey et al 2022

Design a “contextual bandit” - an **adaptive experiment** with multiple arms

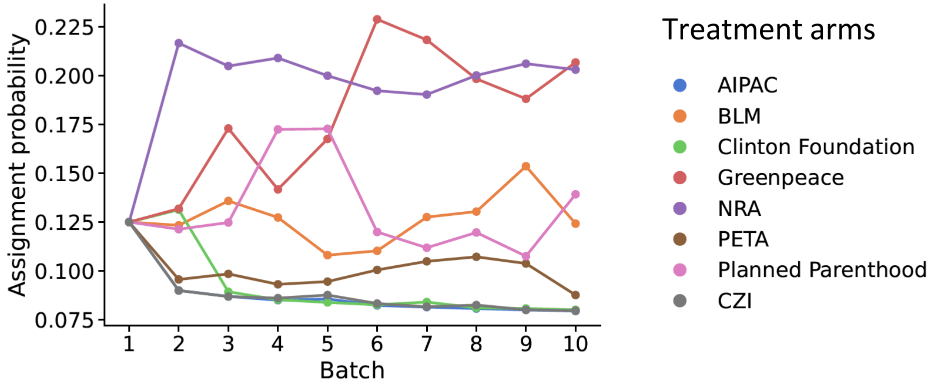
Tension arises between

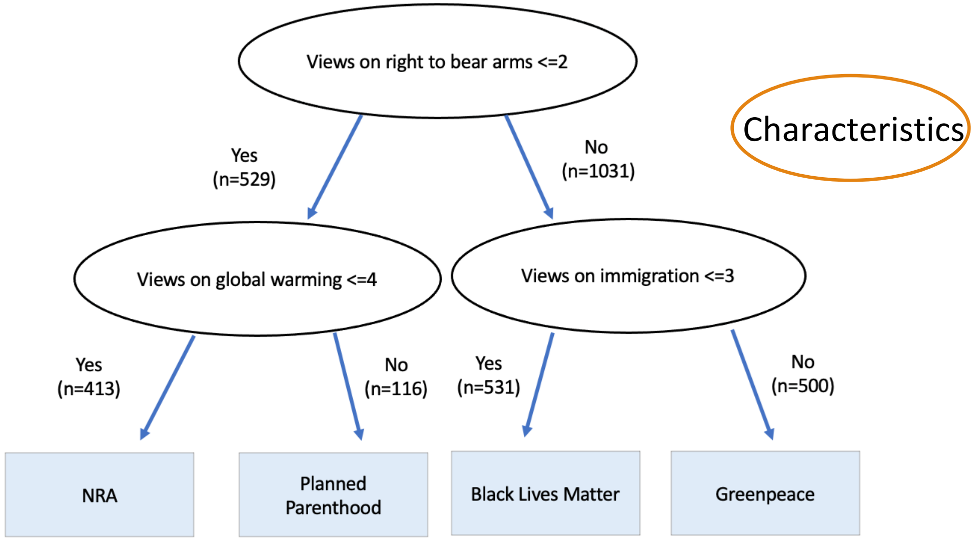
- **Cumulative regret** (**within-experiment** outcomes), and
- Finding best policy to use AFTER experiment (**“policy learning”**)

Propose a heuristic algorithm that **balances the two goals**.

Implement in charitable giving field experiment.

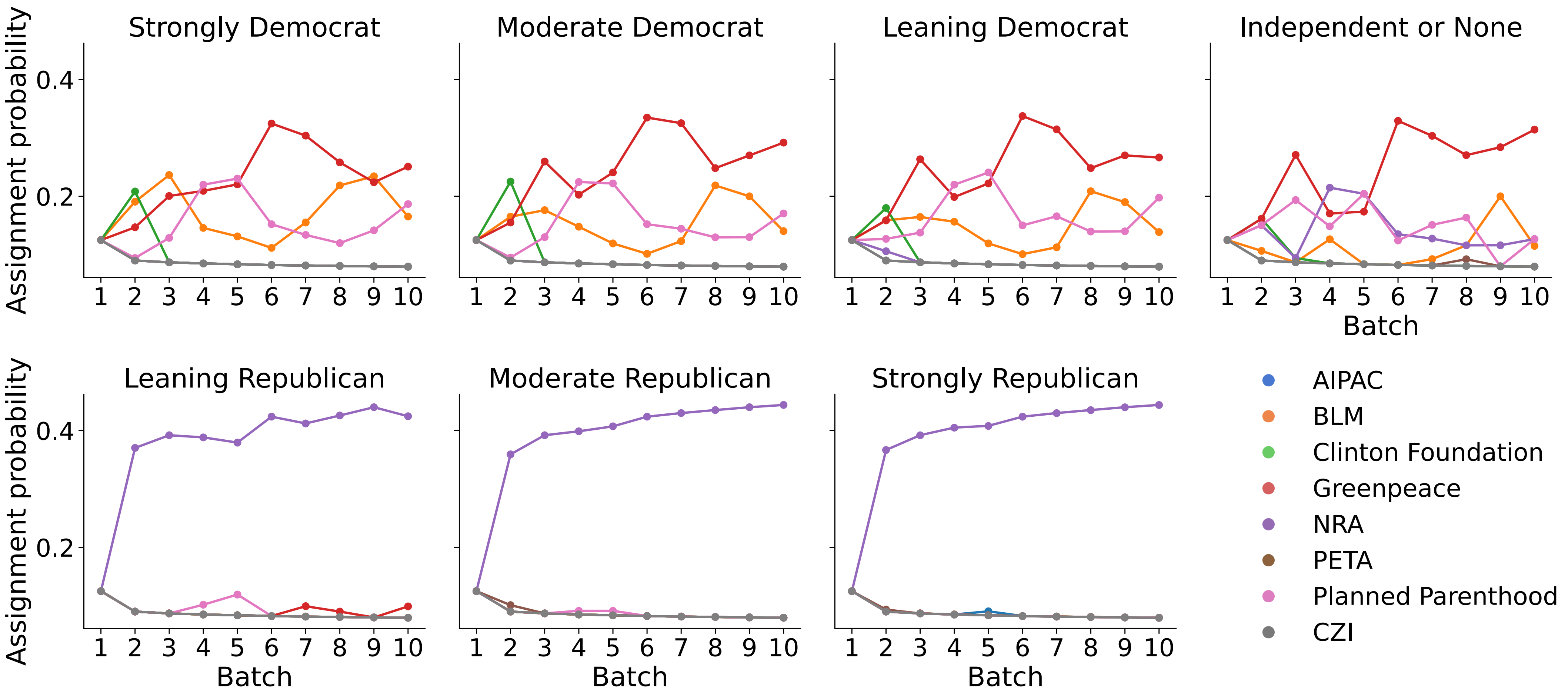
Compare with other existing contextual bandit algorithms using semi-synthetic data based on our experimental data.



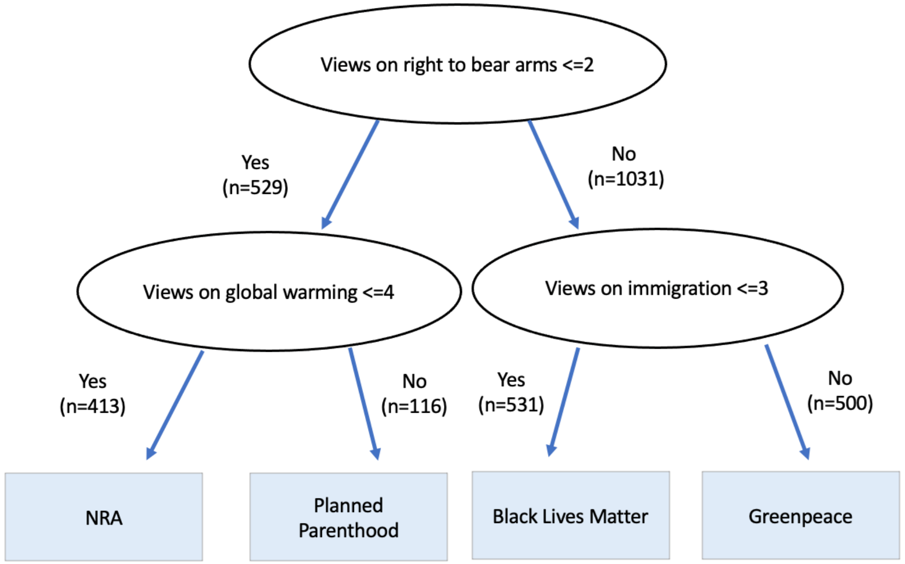


Treatment arms

Contextual bandit over time



Targeted policy vs. best non-targeted policy



	Value	Std.err	Diff	Std.err	p-value
Best non-targeted policy (Greenpeace)	4.687	0.208			
Targeted policy	5.653	0.216	0.966	0.300	0.001

Please drag the slider to indicate your estimate, with ***-10 being extremely dissatisfied***, and ***10 being extremely satisfied***.

-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

Views on immigration: The US government needs to get tougher on immigration
Views on global warming: The US government should do more to prevent global warming
Views on right to bear arms: The right to bear arms should be limited

1- Strongly disagree, 2 - Somewhat disagree, 3 - Neither agree nor disagree, 4 - Somewhat agree, 5 - Strongly agree

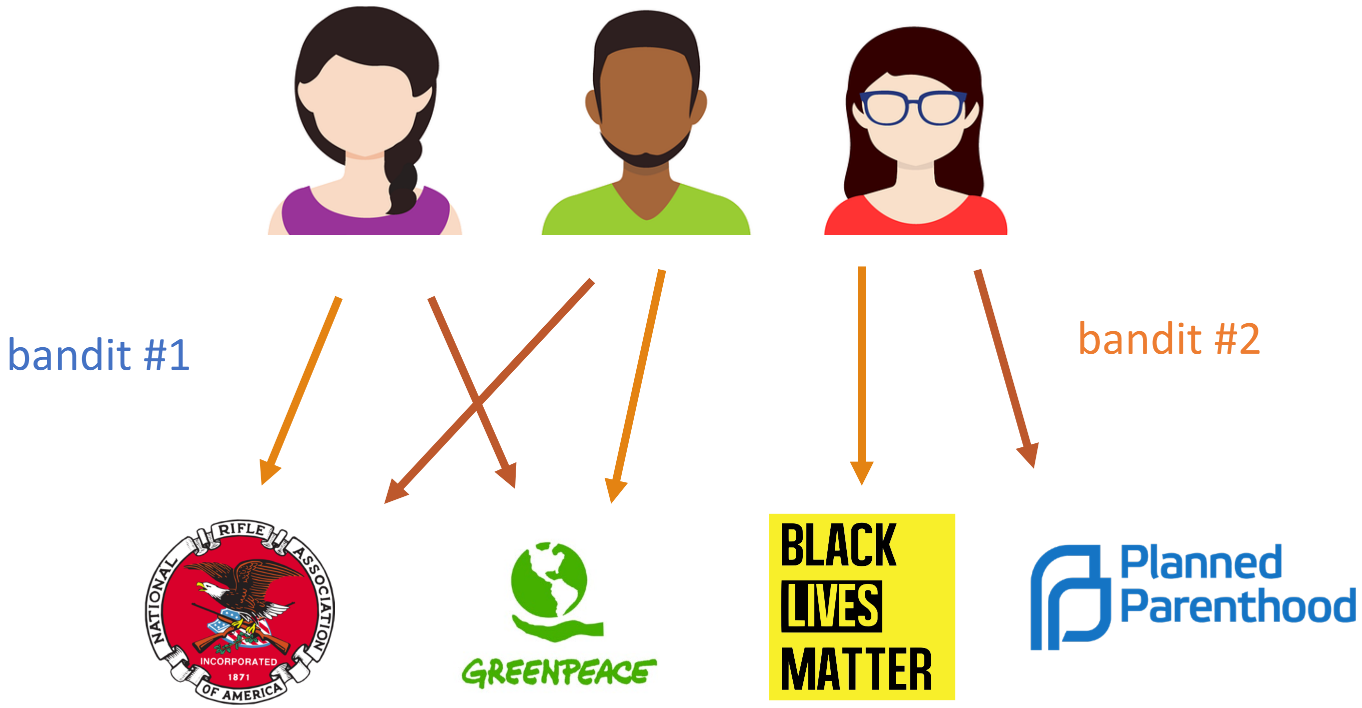
Simulations based on semi-synthetic data

Contextual bandits algorithms guide data collection ⇒

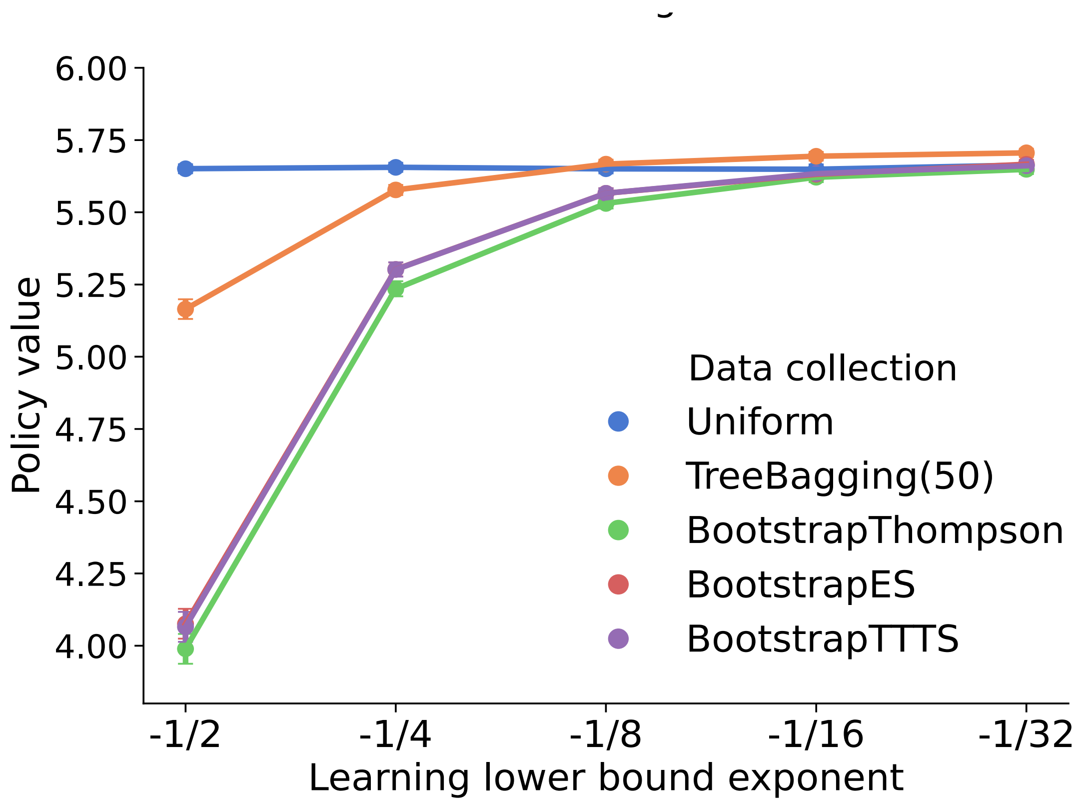
- Not straightforward to reanalyze historically collected data to compare algorithms
- **For a given x**, a **different algorithm** would assign a **different treatment** than what was observed

Running many parallel experiments to compare algorithms can be costly ⇒

- Rely on simulations based on semi-synthetic data
- See Athey et al 2021 for more realistic simulations based on GANs
- Here we change complexity of treatment effect heterogeneity across simulation designs



Uniform Randomization does better than CB algorithms at **Policy Learning** without aggressive lower bounds



Compare performance of different algorithms in semi-synthetic experiments

Plot show average value of learned policy with varying tuning parameters

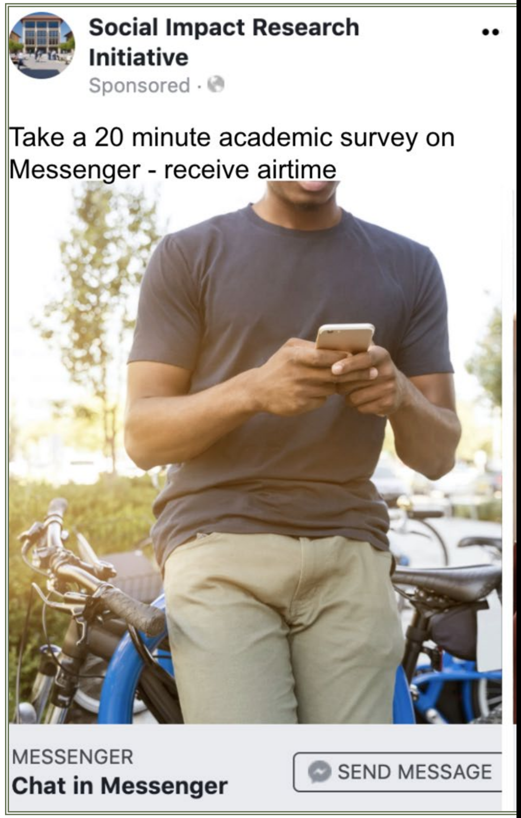
Takes significant hyperparameter tuning to SLOW DOWN exploitation in order for variants of TS to beat uniform randomization.

[1] Athey, Byambadalai, Hadad, Krishnamurthy, Leung, Williams (2022)

Misinformation Interventions Research Design

Offer-Westort, Rosenzweig, & Athey (2023 Nature Hum Behavior)

- **Goals:** First, to narrow treatments, and second, to estimate and evaluate a targeted treatment assignment policy
- **Design:** Contextual Adaptive Experiment to **narrow down treatments** and **learn a targeted treatment assignment policy**, with an evaluation phase to gather more data to precisely estimate the benefits to the policy (and test null of no benefits to targeting)
- **Respondents**: Facebook users in Kenya and Nigeria (WHO priority)
- **Treatments**: Interventions to combat the spread of COVID misinformation
- **Outcomes**: Sharing intentions and behaviors (aggregated)



Treatments

Respondent-Level Treatment: Pledge

Do you want to help keep your family, friends, and community safe from COVID-19?

Yes

Did you know that false information about ways to prevent or cure COVID-19 threaten the health and wellbeing of everyone around us?

Yes

Are you committed to keeping your family, friends, and community safe from COVID-19 **misinformation?**

Yes

Great! Take our pledge by posting this pledge to your timeline now.

Post to timeline

Not now

Headline-Level Treatment: Real Info

Osadare Olumide
March 28 ·

SIMPLE SOLUTION TO CORONA VIRUS REVEALED.

God is wonderful, as deadly as Corona Virus is, it has been confirmed and tested that Palm oil can stop the spread of the virus.

Experts have advised, we drink two spoonful of palm oil every morning to avoid the continuous spread of the virus.

Plz help share and spread this important message urgently to your families, friends and love ones.

Your message can save millions of lives.

WORLD HEALTH ORGANIZATION (WHO)

World Health Organization

According to the WHO, there is currently **no proven** cure for COVID-19.

Palm oil is simple solution to Corona
Facebook user

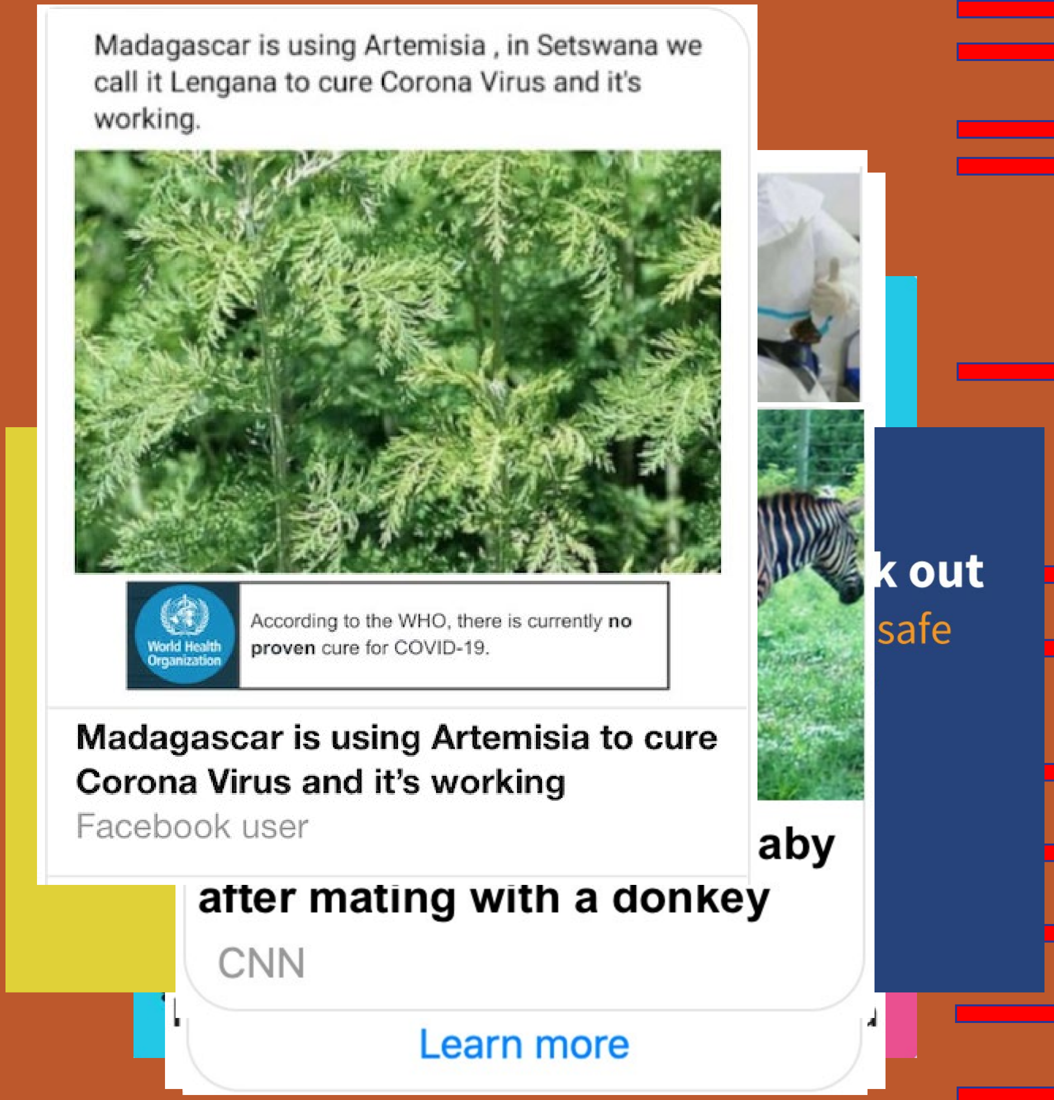
Learn more

Would you like to share this post on your timeline?

[Reminder: don't share it now, you will be able to share it at the end of the survey]

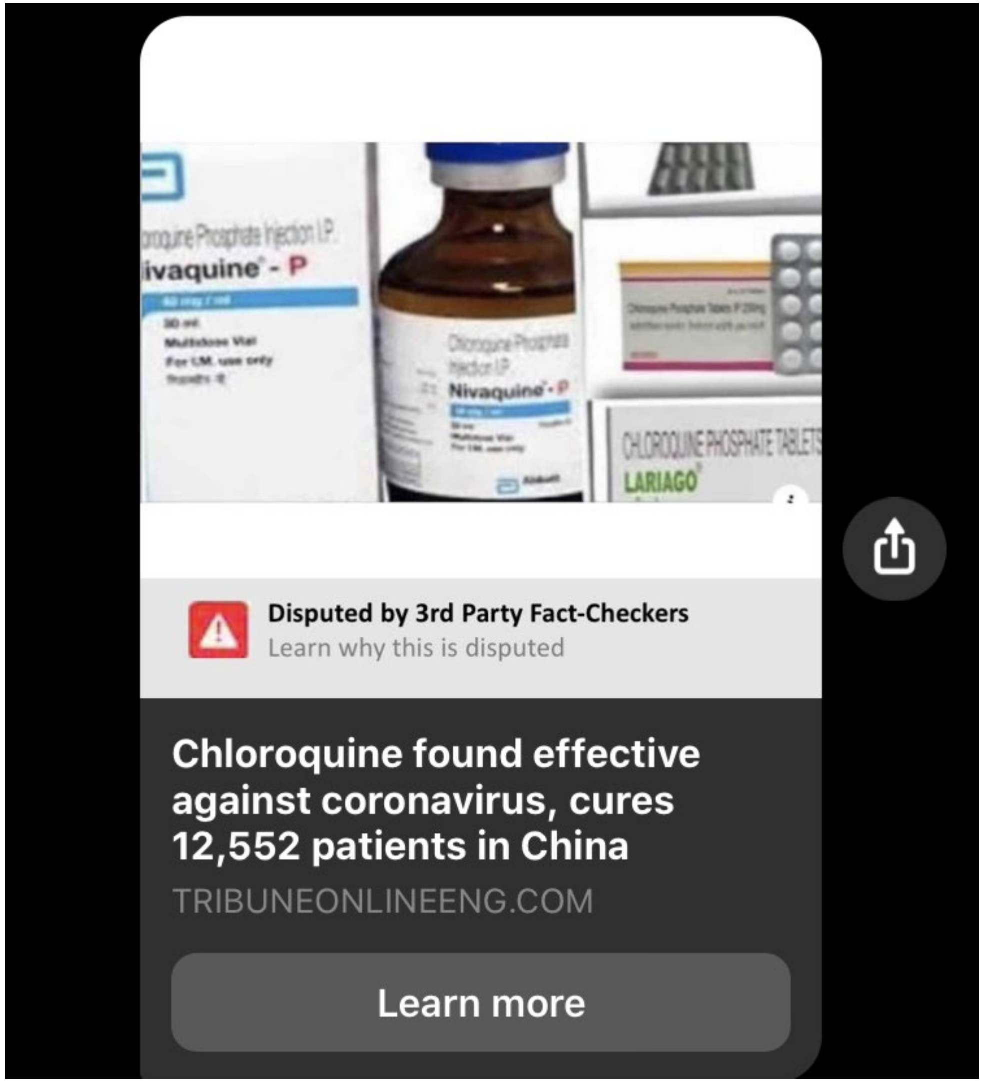
Yes No

Treatments



Shorthand Name	Treatment Level	Treatment
1. Facebook tips	Respondent	Facebook's "Tips to Spot False News"
2. AfricaCheck tips	Respondent	`Africacheck.org`'s guide: "How to vet information during a pandemic"
3. Video training	Respondent	BBC Video training
4. Emotion suppression	Respondent	Prompt: "As you view and read the headlines, if you have any feelings, please try your best not to let those feelings show. Read all of the headlines carefully, but try to behave so that someone watching you would not know that you are feeling anything at all" (Gross, 1998).
5. Pledge	Respondent	Prompt: Respondents will be asked if they want to keep their family and friends safe from COVID-19, if they knew COVID-19 misinformation can be dangerous, and if they're willing to take either a *private* or *public* pledge to help identify and call out COVID-19 misinformation online
6. Accuracy nudge	Respondent	Placebo headline: "To the best of your knowledge, is this headline accurate?" (Pennycook et al., 2020, 2019).
7. Deliberation nudge	Respondent	Placebo headline: "In a few words, please say *why* you would like to share or why you would not like to share this headline." [open text response]
8. Related articles	Headline	Facebook-style related stories: below story, show one other story which corrects a false news story
9. Factcheck	Headline	Fact checking flag from third party (e.g., Facebook, AFP, AfricaCheck, etc)
10. More information	Headline	Provides a link to "Get the facts about COVID-19" as per Twitter flags
11. Real information	Headline	Provides a *true* statement: "According to the WHO, there is currently **no proven** cure for COVID-19.
12. Control	N/A	Control condition

Response measurement



Outcomes:

- *Would you like to share this post on your timeline?*
- *Would you like to send this post to a friend on Messenger?*

M_i = Sum of *misinformation* outcomes

T_i = Sum of *true information* outcomes

Response function (weighted sum):

$$Y_i = -M_i + 0.5T_i$$

Review of experiment design

Data collection:

- 4.5k in adaptive learning (Feb/March 21)
- 12.1 k in evaluation split (July 21)
    - 1,451 simple balanced random assignment
    - 10,681 on-policy targeted assignment

Analysis:

- Response function: *Weighted sum* of sharing intentions,
  M = misinformation, T = True stimuli:
  $Y_i = -M_i + 0.5T_i$

Evaluation arms:

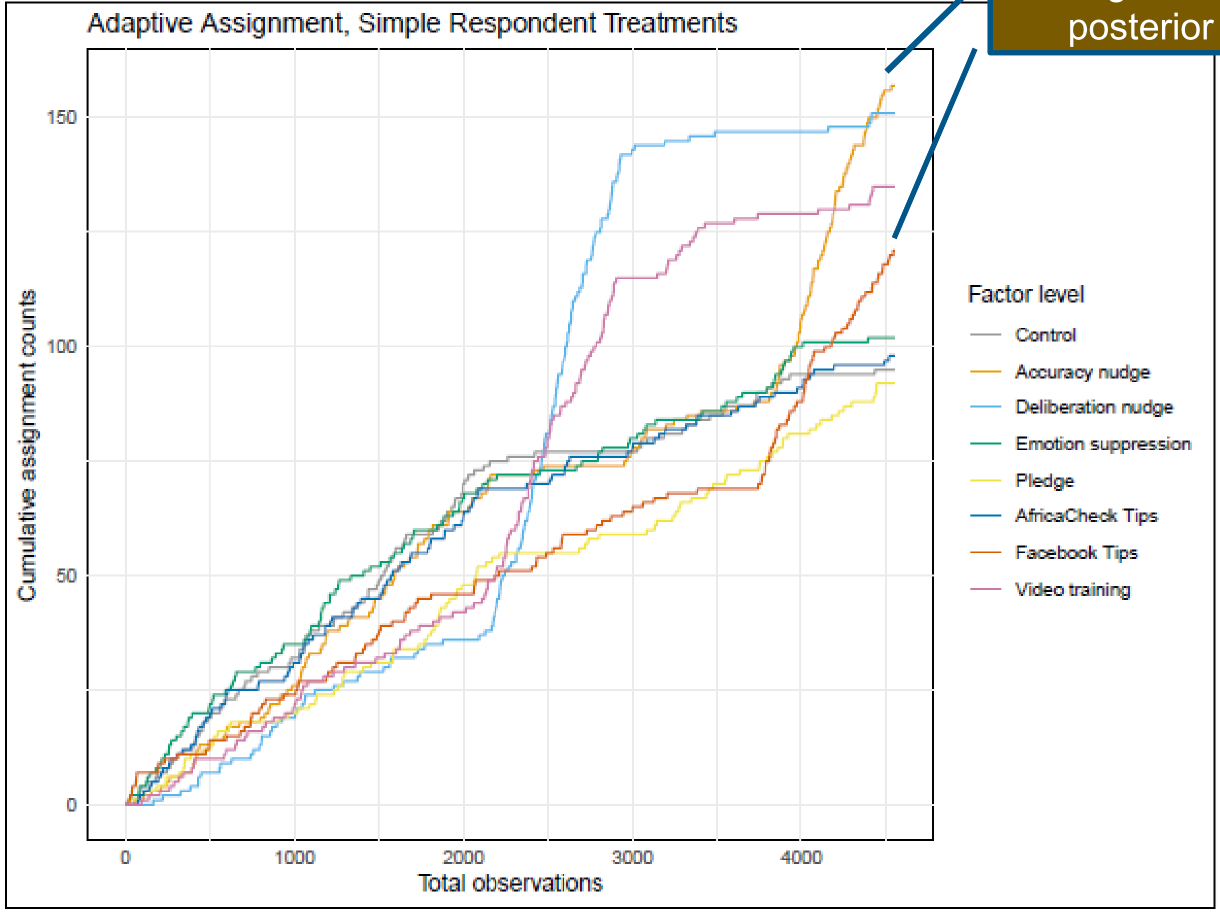
- Pure control
- Headline only:
    - Factcheck
    - Related Articles
- Respondent only:
    - Accuracy
    - Facebook Tips
    - Optimal contextual (accuracy/Facebook tips/video/emotion 83/15/1/1)

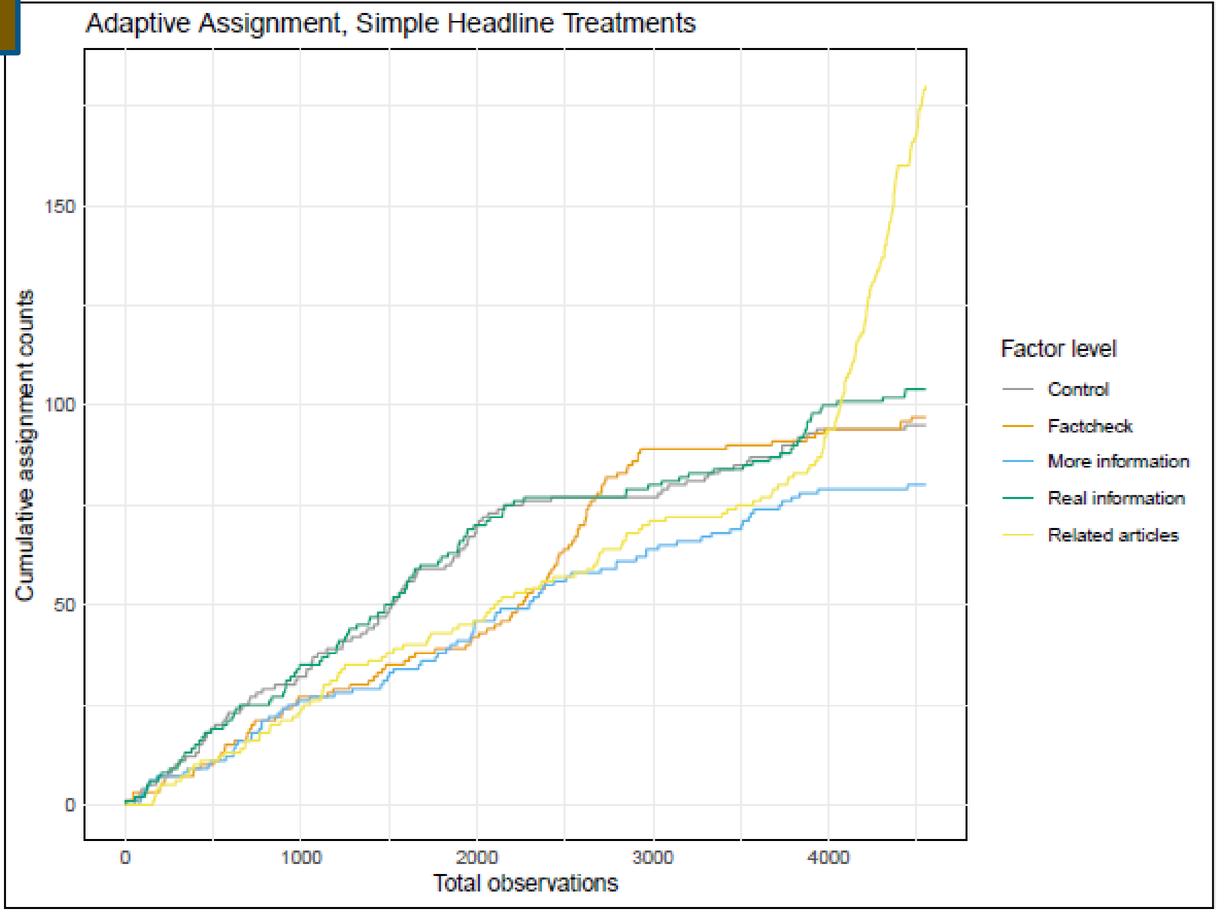
Design Discussion

- Adaptive phase/contextual bandit
  - Ideally, collect more data about treatment arms that are more effective for each type of subject
  - Challenge: instability can *increase* variance when estimating optimal policies
  - Post-hoc analysis (informed by econometric theory): estimate a policy where only options are the best-performing arms. That policy appears to perform better when evaluated in test data.
- Evaluation phase
  - Gather purely randomized data across smaller number of arms
    - Enables off-policy evaluation of variety of targeted policies, with sufficient overlap between data collection and policy evaluated
  - Gather additional data on learned targeted policy
    - Enables hypothesis testing after the experiment,

Assignment Probabilities in Adaptive Phase

Steep slope = high assignment/ posterior





Frequent problem with moderate size experiments:
convergence is not achieved, environment is changing

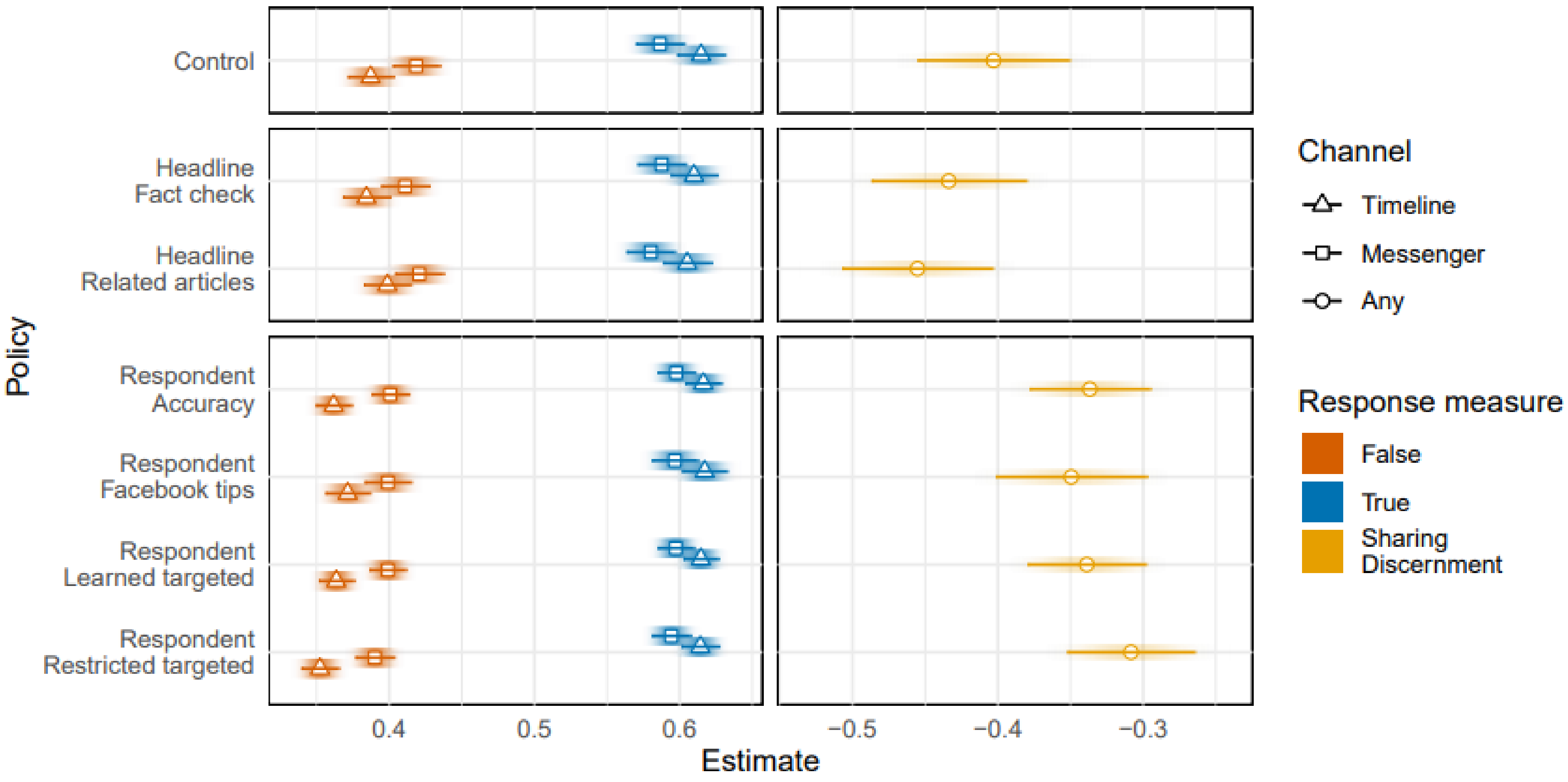
Outcomes in Evaluation Phase & Mechanisms

Learned targeted policy is estimate of best policy at the end of adaptive phase (and deployed in evaluation phase)

Restricted targeted policy is targeted policy restricted to the top two individual arms (also uses adaptive phase data). Since greater signal on the false sharing outcome, it is optimized on the latter outcome.

The Restricted Targeted Policy has greater discernment than the control, TE=0.029, s.e. = 0.013

It also achieves a decrease in false intentions sharing relative to control of −3.3 pp (s.e. = 1.0)



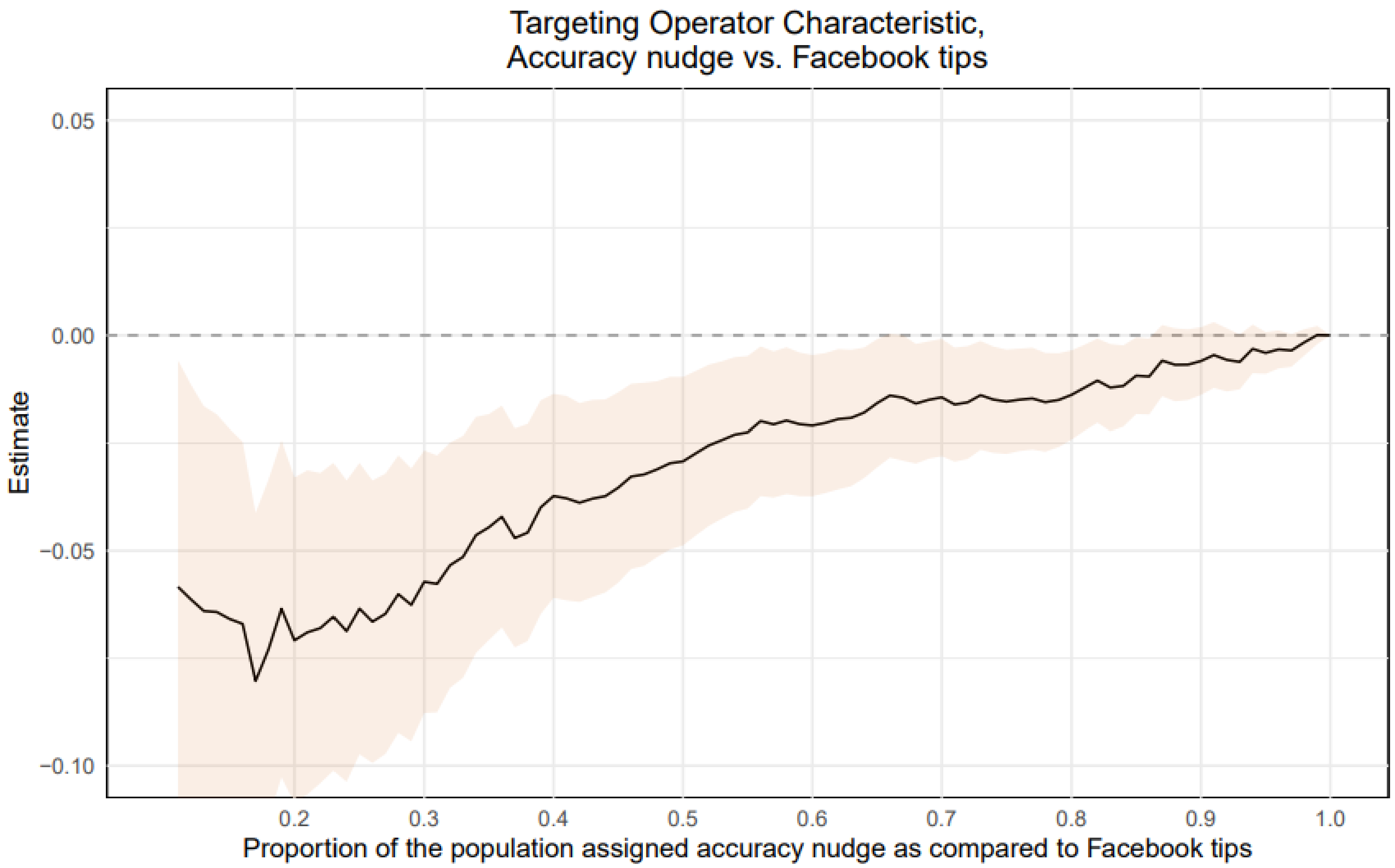
Subjects appear to discern true vs. false posts

Restricted Contextual Policy: Average Characteristics of Users Assigned to Each Arm

Covariate	Optimal policy == accuracy (n = 8,309)	Optimal policy == FB tips (n = 2,222)	Difference
Digital literacy index	14.050 (0.043)	12.570 (0.094)	−1.486 (0.103) Z = −14.427, p = <0.001 [−1.688, −1.284]
Age	27.310 (0.085)	29.070 (0.168)	1.757 (0.188) Z = 9.346, p = <0.001 [1.389, 2.125]
Supports governing party	0.293 (0.005)	0.331 (0.010)	0.038 (0.011) Z = 3.455, p = <0.001 [0.016, 0.060]
Male	0.539 (0.005)	0.512 (0.011)	−0.027 (0.012) Z = −2.250, p = 0.024 [−0.051, −0.003]
Scientific knowledge index	1.369 (0.007)	1.376 (0.014)	0.007 (0.016) Z = 0.438, p = 0.662 [−0.024, 0.038]

Standard deviation on normalized distribution: −0.2 −0.1 0.0 0.1 0.2

Outcomes in Evaluation Phase



Plot shows differences in average outcomes between two scenarios for prioritizing accuracy nudge assignment:

X% randomly chosen vs.

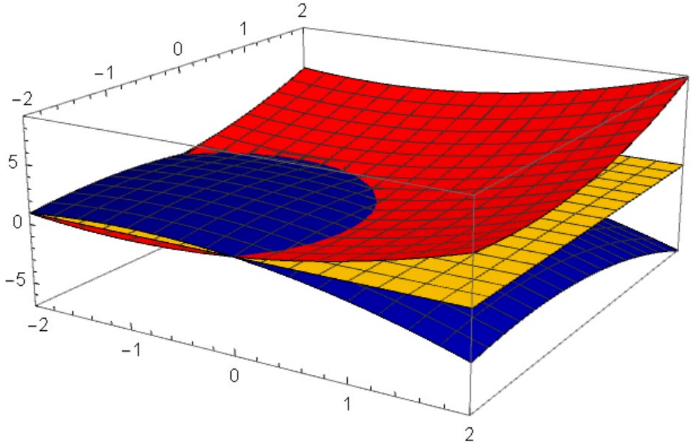
X% prioritized by treatment effect

False sharing outcome (.44 baseline average)

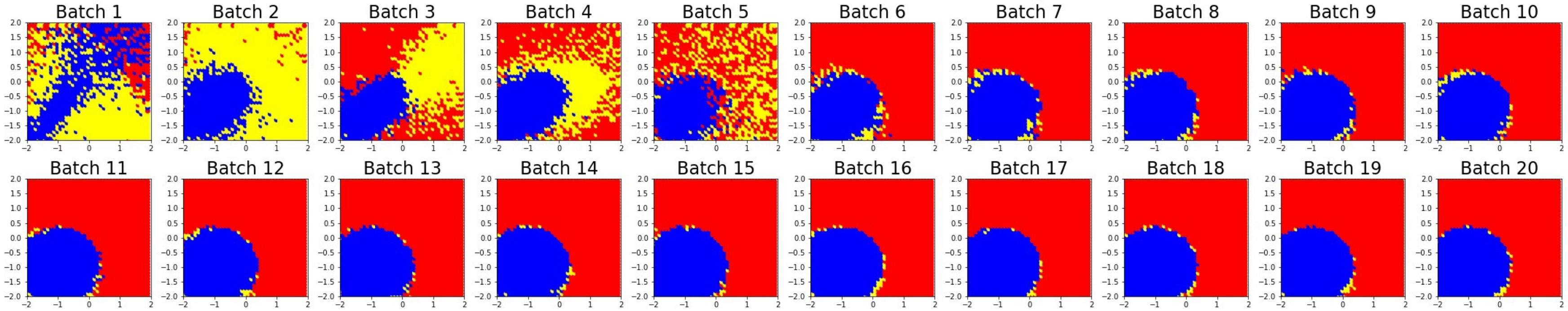
Priority from causal forest treatment effects (grf) Athey, Tibshirani & Wager (2019)

TOC (grf) from Yadlowsky, Fleming, Shah, Brunskill, and Wager (2021)

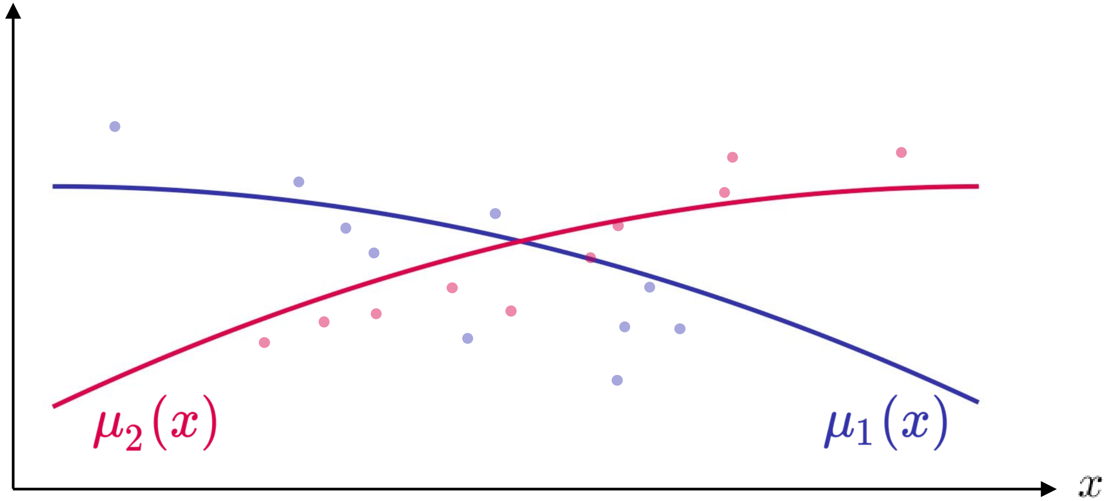
Mean outcomes $\mu_a(x)$ for different arms depend on contexts

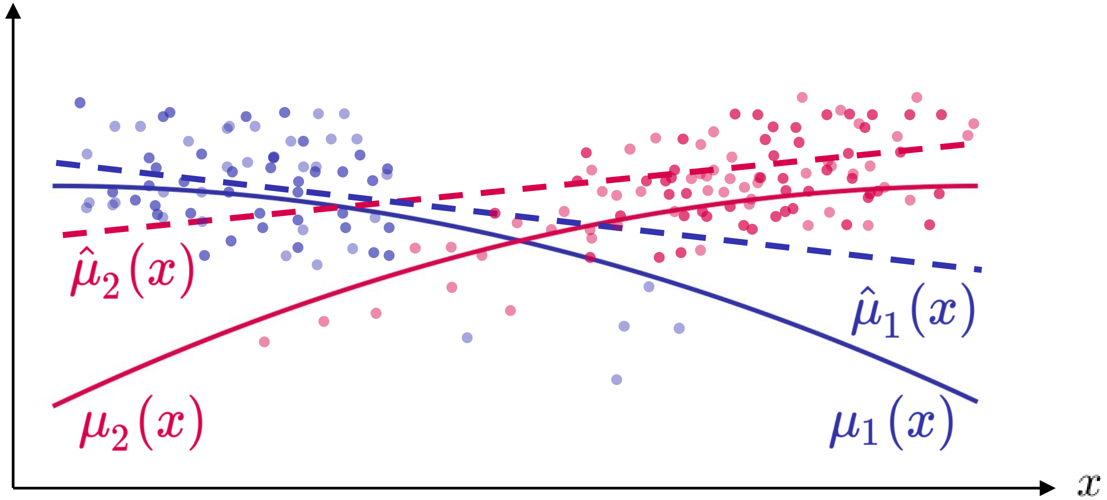


Doubly robust contextual bandit learns the optimal treatment assignment policy: Dimakopoulou, Zhou, Athey & Imbens 2019



Estimation at each batch plagued by adaptivity of assignment process. Weighting creates variance due to **lack of overlap** as assignment probabilities get more concentrated





Contextual bandits algorithm issues – policy learning and regret

Recall problems encountered:

- Selecting functional form complexity
- Uncertainty quantification needed to do data-driven experimental design (guide exploration)
- Instability & poor overlap with optimal policies implies hard to beat RCT for policy learning

Goal tradeoff:

- Post-experiment policy learning vs. in-experiment outcomes

Solutions make heavy use of tools from semi-parametric approaches for causal inf w/ unconfoundedness

- See eg Carranza, Krishnamurthy & Athey (AISTATS 2023)
- But also very different issues around uncertainty quantification and bandit heuristics: need **algorithm** to choose **how to explore arms**

Challenge: **Functional forms** for outcomes

- Need func. form to extrapolate to make decisions for new context
- Early, func. Form too complex for data size and later, too simple for reality
- Solution: **data-driven model selection** using cross-validation, together with **specification test** to show that uncertainty quantification based on outcome model is reliable to use for **exploration rate**
  - Func. Form increases in complexity as more data collected
  - Specification test compares policy evaluation using (known) propensity weighting and outcome modeling approach (Krishnamurthy, Athey & Brunskill, arxiv 2024)
  - Result (Krishnamurthy, Propp, & Athey AISTATS 2024): can use resulting uncertainty to guide exploration rate, "costless model selection"

Challenge: **Controlling overlap** for policy learning

- Algo with tuning parameter that traces frontier for regret vs. policy learning
- New algorithms that directly target risk of excluding potentially optimal arms
- Krishnamurthy, Athey & Brunskill (NeurIPS 2023)

$$Var(\widehat{\mathbb{E}}_{X_i}[Y_i(\pi(X_i))]) = \frac{\sigma^2}{N}\frac{1}{Pr(A_i=\pi(X_i))}$$

Takeaways

Economics toolkit applied for high-level design of intervention, experiment structure, and outcomes

Experimentation cycle is part of operation of digital systems

Detailed design choices solve an **optimization problem**

- **Work backwards**
- Goal: what will you do with the results & how?
  - E.g. Insight, deploy once, deploy as part of experimentation cycle
- What **outcomes** are you measuring to achieve goal?
- Solve **optimization problem** for experiment design
  - Formally or informally, theoretically or with simulations
  - One-time/in advance, series of iterative experiments, or fully adaptive
- Full adaptive issues
  - Fully adaptive is “AI replacement” for experimental design who is solving a causal inference problem with unconf. at each step
  - Needed a lot of refinement to automate solutions to common problems
  - Getting closer to working reliably, but still cumbersome to implement

Appendices

Details on Hypothesis Testing With Bandit (Adaptively Collected) Data

Adaptive experimentation
Learning by design

Caveats
Statistics under adaptivity

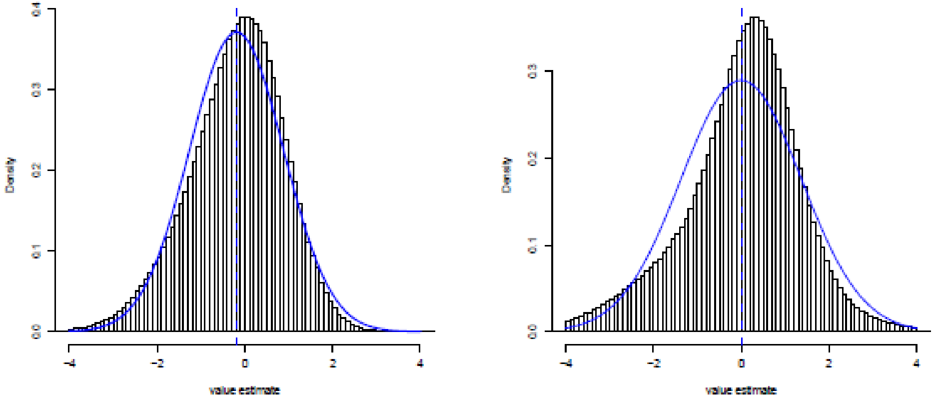
In an adaptive experiment, collected data are **not independent**.

Usual methods for inference will often give the wrong answer. More sophisticated methods are needed.

Hadad, Hirshberg, Zhan, Wager, Athey (2021, PNAS)

This is an area of active research:

- Luedtke and van der Laan (2016)
- Deshpande, Mackey, Syrgkanis, Taddy (2017)
- Howard, Ramdas, McAuliffe, Sekhon (2019ab)
- Zhang, Janson, Murphy (2020)



Sample average: biased and not Normal

IPW: unbiased, but higher variance and not Normal

Example: distribution of the sample mean after an adaptive experiment. Estimates are biased and do not have a normal distribution (usual t-tests are not valid!).

Let's consider a simple experiment.

- There are two arms.
- Both arms have the same expected value of zero.
- Data is collected **adaptive**, in two batches:

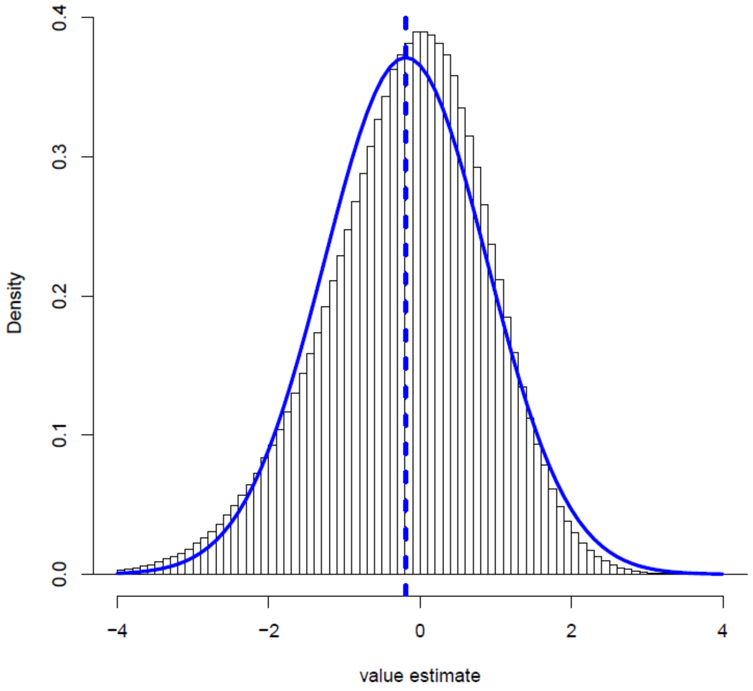
Periods t = 1 to T/2	Period t = T/2	Periods t = T/2 + 1 to T
Assign each arm with 50% probability	Estimate sample means	Assign arm with highest estimated mean 90% of the time

Goals:

- Estimate the **value** of the arm $Q(w) := E[Y(w)]$.
- Produce a valid confidence interval for it.

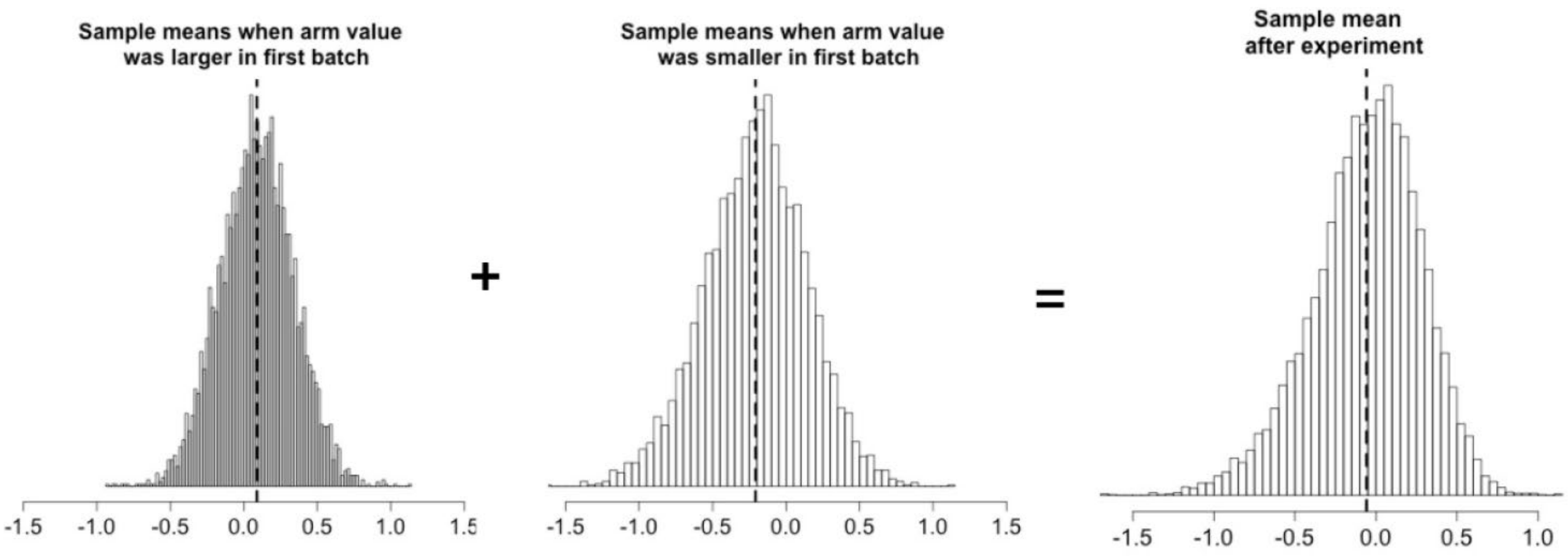
Idea 1: Use the sample mean:

$$\widehat{Q}_{w,T}^{AVG} := \frac{1}{n_{w,T}} \sum_{\{t: W_t = w\}} Y_t$$



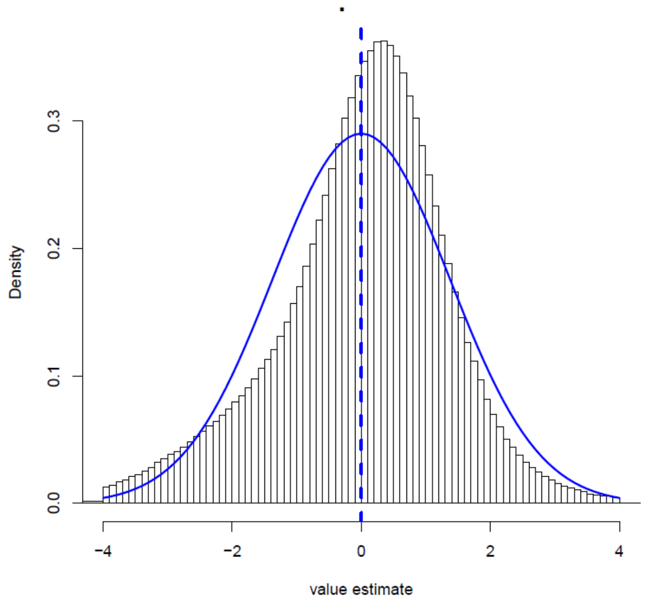
The sample mean is **biased** and **not Normal**!

- Why is the sample mean **biased**?
- When we get a low estimate of the average after the first batch (by chance!), we collect less data in future batches, which has the effect of overweighting the batches with (spuriously) lower outcomes.
- Why is it not **normal**?
- Because it's a mixture of two distributions:



Idea 2: Use the inverse-propensity weighted (IPW) mean:

$$\widehat{Q}^{IPW}_{w,T} := \frac{1}{T}\sum_{t=1}^{T} \frac{Y_t 1\{W_t = w\}}{e_t(w)} \qquad e_t(w) \text{ assignment prob.}$$



Interpretation with batches:

- First, take average outcome for an arm within each batch.
- Then, equally weight the batches.

The IPW mean is **unbiased**, but it's still **not Normal**!
It can also exhibit **high-variance** due to small propensity weights.

- What's the problem with IPW averages?

$$\widehat{Q}^{IPW}_{w,T} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\Gamma}^{IPW}_t(w) \quad \text{with} \quad \widehat{\Gamma}^{IPW}_t(w) \quad = \frac{1(\{W_t = w\})}{e_t(w)} Y_i$$

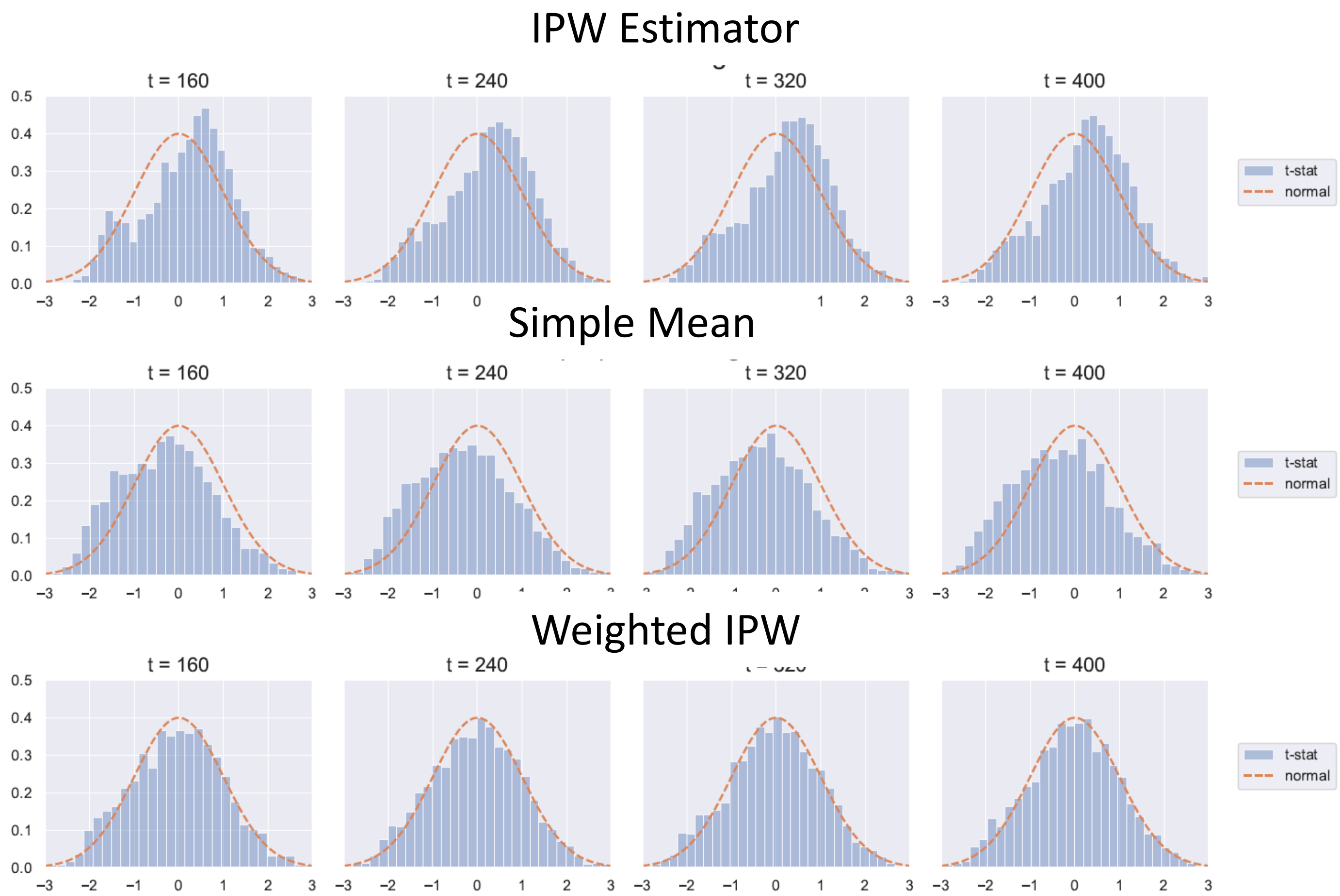
- It's an average of unbiased terms: $\widehat{\Gamma}^{IPW}_t(w) = E[Y_t(w)]$.
- But in a bandit experiment, $e_t(w) \to 0$ for bad arms.
- For bad arms, variance is small at the beginning of the experiment, but very large at the end.
- In spite of that, we're giving them the same weight $1/T$!
- Idea: average using **non-uniform weights**.

Let's see how to apply this insight next.

Testing Hypotheses After Bandit

- Simple mean is **biased** estimator of true arm mean (initial bad outcomes upweighted)
- **Weighting** by inverse of assignment probability restores "equal weighting of each batch," eliminates bias
- IPW is not asymptotically normal

Solution: adaptively weight data to stabilize variance, retain consistency, restore normality (Hadad et al, PNAS 2021)



Introduce *evaluation weights* $h_t(w)$.

Adaptively-Weighted Augmented Inverse Propensity Score Estimator

$$\widehat{Q}_T^{AW}(w) := \sum_{t=1}^{T} \frac{h_t(w)}{\sum_{s=1}^{T} h_s(w)} \left\{ \frac{\mathbb{I}\{W_t = w\}}{e_t(w)} Y_t + \left(1 - \frac{\mathbb{I}\{W_t = w\}}{e_t(w)}\right) \hat{\mu}_t \right\}$$

- We started with the IPW estimator:

$$\widehat{Q}^{IPW}_{w,T} = \sum_{t=1}^{T} \frac{1}{T} \widehat{\Gamma}^{IPW}_t(w) \quad \text{with} \quad \widehat{\Gamma}^{IPW}_t(w) = \frac{1(\{W_t = w\})}{e_t(w)} Y_i$$

- Identified major flaw: variance depends on (time-varying, adaptive) $1/e_t(w)$, but averaged with (static) $1/T$.
- Proposed the **adaptively-weighted estimator**

$$\widehat{Q}^{h}_{w,T} = \sum_{t=1}^{T} \frac{h_t(w)}{\sum_{t=1}^{T} h_t(w)} \widehat{\Gamma}^{AIPW}_t(w) \quad \text{with}$$

$$\widehat{\Gamma}^{AIPW}_t(w) = \frac{1(\{W_t = w\})}{e_t(w)} Y_i + \left(1 - \frac{1(\{W_t = w\})}{e_t(w)}\right) \hat{m}_t(w)$$

- Next question: which weights h_t should we use?

Central limit theorem for adaptively-weighted estimates

Evaluation weights h_t and assignment mechanism e_t satisfy:

A1. Infinite sampling

$$\left(\sum_{t=1}^{T} h_t\right)^2 \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2}{e_t}\right] \xrightarrow[T\to\infty]{p} \infty.$$

A2. Variance convergence

$$\sum_{t=1}^{T} \frac{h_t^2}{e_t} \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2}{e_t}\right] \xrightarrow[T\to\infty]{L_p} 1.$$

A3. Bounded moments

$$\sum_{t=1}^{T} \frac{h_t^{2+\delta}}{e_t^{1+\delta}} \Big/ \mathbb{E}\left[\sum_{t=1}^{T} \frac{h_t^2}{e_t}\right]^{1+\delta/2} \xrightarrow[T\to\infty]{p} 0.$$

Central limit theorem for adaptively-weighted estimates

Theorem. (Intuitive version) Suppose assumptions A1-A3 are satisfied, and suppose that either $\hat{m}_t$ is consistent or e_t has an almost-sure limit. Then,

$$\widehat{Q}_T^h(w) \xrightarrow[T\to\infty]{p} Q(w) \qquad \text{[Consistency]}$$

$$\frac{\widehat{Q}_T^h(w) - Q(w)}{\widehat{V}_T(w)^{\frac{1}{2}}} \xrightarrow[T\to\infty]{d} \mathcal{N}(0,1) \qquad \text{[Asymptotic Normality]}$$

Notes:

- $\widehat{V}_T(w)$ is a simple estimate of the variance (see paper).
- Our general theory also covers many other estimands, including the difference in value $E[Y(w) - Y(w')]$.

Weights that satisfy this recursive condition allow for our CLT:

$$\frac{h_t^2}{e_t} = \left(1 - \sum_{s=1}^{t-1} \frac{h_s^2}{e_s}\right) \lambda_t \qquad \lambda_t \in [0, 1], \lambda_T = 1$$

- Observation t has variance proportional to h_t^2/e_t.
- Recursive construction forces $\sum_{t=1}^{T} h_t^2/e_t = 1$, so that the variance convergence condition (A2) is satisfied (prevents issues with bandits in the no-signal case).
- Allocation rate λ_t governs the weight observation t receives.
- We can do more than just satisfying CLT conditions.
- Make λ_t adaptive – **get lower variance**!

Susan Athey’s Public Resources

ONLY INCLUDES MY OWN PAPERS, NOT A COMPREHENSIVE BIBLIOGRAPHY

Resources: Tutorials & User Guides

MACHINE LEARNING & CAUSAL INFERENCE TEACHING MATERIAL & TUTORIALS

Videos and slides for ML & Causal Inference:

2021 YouTube Playlist https://bit.ly/MLCIplaylist
2018 AEA 2-day Course https://bit.ly/MLCI2018

Bookdown tutorials can be downloaded and run on public or private data. Covers prediction and cross-validation, ATE, HTE, policy estimation, causal panel data https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/

Public report on Paypal Giving Experiments: https://www.gsb.stanford.edu/sites/default/files/publication/pdfs/report-2021-mar-paypal-giving-experiments_2.pdf

Computational Applications to Behavioral Science report with ideas42 (gentle introduction to machine learning): https://gsbdbi.github.io/ml_for_behavioral_science/index.html

RESOURCES FOR EXPERIMENT PLANNING AND TEACHING

Shiny app for simulating and teaching about bandits: https://www.gsb.stanford.edu/faculty-research/labs-initiatives/sil/research/bandit-experiment-application

Practitioner's Guide for Designing Adaptive Experiments: https://www.gsb.stanford.edu/sites/default/files/publication/pdfs/academic-publication-desiging-adaptive-experiments-2021-mar.pdf

Repository with many datasets from economic field experiments: https://github.com/gsbDBI/ExperimentData

Statistical Software Packages

AVERAGE TREATMENT EFFECTS: CROSS-SECTIONAL AND PANEL DATA

Average treatment effects

- BalanceHD: Residual balancing algorithms for average treatment effects or policy evaluation under unconfoundedness
- Grf: Implementation of an AIPW method for ATE using causal forests for nuisance parameters
- DS-WGAN: Generative adversarial networks to design data-driven simulations to compare methods
- ParTreat: Software for estimating treatment effects in experiments where the outcome distributions may have “fat tails”
- Tutorial for average treatment effects, with R code: https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/ate-i-binary-treatment.html

Panel data

- Torch Choice: A Library for flexible, fast discrete choice modeling designed for both estimation and prediction
- MCPanel: Matrix completion for causal panel data models simulations for benchmarking causal estimators
- Synthdid in R and in Stata: Synthetic difference-in-differences (software paper here https://dx.doi.org/10.2139/ssrn.4346540)
- Tutorial for causal panel data methods with R code: https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/causal-panel-data.html

HETEROGENEOUS TREATMENT EFFECTS AND POLICY EVALUATION/ESTIMATION

Grf package + tutorials https://grf-labs.github.io/grf/

- Honest random forests
- Causal forest under unconfoundedness or with IV
- Causal forest with many outcomes/treatments
- Causal survival forest
- Quantify/test for HTE
- Average treatment effects with unconfoundedness, including with covariate shift
- Qini curves
- Sufficient representations of categorical variables

Tutorial for HTE, with R code: https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html

PolicyTree for estimating tree-based policies: https://grf-labs.github.io/policytree/

CausalTree: Heterogeneous treatment effects with causal trees

Methods: HTE & Policy Evaluation

HETEROGENEOUS TREATMENT EFFECTS

Athey, Susan, and Guido Imbens. “Recursive Partitioning for Heterogeneous Causal Effects.” *Proceedings of the National Academy of Sciences* 113, no. 27 (2016): 7353-7360. https://doi.org/10.1073/pnas.1510489113

Athey, Susan, Julie Tibshirani, and Stefan Wager. “Generalized Random Forests.” *Annals of Statistics* 47, no. 2 (2019): 1148-1178. arXiv:1610.01271

Friedberg, Rina, Julie Tibshirani, **Susan Athey**, and Stefan Wager. “Local Linear Forests.” *Journal of Computational and Graphical Statistics* (2020): 1-15. arXiv:1807.11408

Wager, Stefan, and **Susan Athey**. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.” *Journal of the American Statistical Association* 113, no. 523 (2018): 1228-1242. arXiv:1510.04342

COUNTERFACTUAL EVALUATION OF POLICIES AND METHODS FOR AVERAGE EFFECTS

Athey, Susan, Peter J. Bickel, Aiyou Chen, Guido Imbens, and Michael Pollmann, “Semi-Parametric Estimation of Treatment Effects in Randomized Experiments,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, (2023). https://doi.org/10.1093/jrsssb/qkad072

Athey, Susan, and Guido Imbens. “A Measure of Robustness to Misspecification.” *American Economic Review* 105, no. 5 (2015): 476-80. https://doi.org/10.1257/aer.p20151020

Athey, Susan, Guido Imbens, Jonas Metzger, and Evan Munro. “Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations.” *Journal of Econometrics* (2021). arXiv:1909.02210

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. “Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges.” *American Economic Review* 107, no. 5 (2017): 278-81. arXiv:1702.01250

Athey, Susan, Guido Imbens, and Stefan Wager. “Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions.” *Journal of the Royal Statistical Society-Series B,* 80(4), (2018): 597-623 arXiv:1604.07125 (formerly titled “Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing.”)

Powell, Michael, Allison Koenecke, James Brian Byrd, Akihiko Nishimura, Maximilian F. Konig, Ruoxuan Xiong, Sadiqa Mahmood, Vera Mucaj, Chetan Bettegowda, Liam Rose, Suzanne Tamang, Adam Sacarry, Brian Caffo, **Susan Athey**, Elizabeth A. Stuart, and Joshua T. Vogelstein. “Ten Rules for Conducting Retrospective Pharmacoepidemiological Analyses: Example COVID-19 Study,” *Frontiers in Pharmacology*, 12 (2021). https://doi.org/10.3389/fphar.2021.700776

Methods: Targeted Treatment Assignment Policies

POLICY ESTIMATION/LEARNING

Athey, Susan, and Stefan Wager. “Policy Learning with Observational Data.” *Econometrica* 89, no. 1 (2021): 133-161. arXiv:1702.02896 (formerly titled “Efficient Policy Learning.”)

Sverdrup, Erik, Han Wu, **Susan Athey**, and Stefan Wager. “Qini Curves for Multi-Armed Treatment Rules.” *arXiv preprint* (2023). arXiv:2306.11979

Sverdrup, Erik, Ayush Kanodia, Zhengyuan Zhou, **Susan Athey**, and Stefan Wager. “Policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization over Trees.” *Journal of Open Source Software* 5, no. 50 (2020): 2232. https://doi.org/10.21105/joss.02232

Zhou, Zhengyuan, **Susan Athey**, and Stefan Wager. “Offline Multi-Action Policy Learning: Generalization and Optimization.” *Operations Research* 71, no. 1 (2023): 148-183. https://doi.org/10.1287/opre.2022.2271

APPLICATIONS OF HTE/POLICY ESTIMATION

Agrawal, Keshav,**Susan Athey**, Ayush Kanodia, and Emil Palikot. "Personalized Recommendations in EdTech: Evidence from a Randomized Controlled Trial." *arXiv preprint* (2022). arXiv:2208.13940

Athey, Susan, David Blei, Robert Donnelly, Francisco Ruiz, and Tobias Schmidt. “Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time using Mobile Location Data.” In *AEA Papers and Proceedings*, vol. 108, (2018): 64-67. arXiv:1801.07826

Athey, Susan, Shawn Allen Cole, Shanjukta Nath, and S. Jessica Zhu. "Targeting, Personalization, and Engagement in an Agricultural Advisory Service." *Available at SSRN 4536641* (2023). https://dx.doi.org/10.2139/ssrn.4536641

Athey, Susan, Lisa K. Simon, Oskar N. Skans, Johan Vikstrom, and Yaroslav Yakymovych. “The Heterogeneous Earnings Impact of Job Loss Across Workers, Establishments, and Markets.” *arXiv preprint* (2023). arXiv:2307.06684

Athey, Susan, and Stefan Wager. “Estimating Treatment Effects with Causal Forests: An Application.” *Observational Studies* (2019). arXiv:1902.07409

Inoue, Kosuke, **Susan Athey**, and Yusuke Tsugawa. “Machine-learning-based High-benefit Approach versus Conventional High-risk Approach in Blood Pressure Management.” *International Journal of Epidemiology* (2023). https://doi.org/10.1093/ije/dyad037

Methods: Adaptive Experiments & Bandits

HYPOTHESIS TESTING & POLICY EVALUATION/ESTIMATION WITH ADAPTIVE DATA

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and **Susan Athey**. "Confidence Intervals for Policy Evaluation in Adaptive Experiments." *Proceedings of the National Academy of Sciences* 118, no. 15 (2021). arXiv:1911.02768

HYPOTHESIS TESTING & POLICY EVALUATION/ESTIMATION WITH ADAPTIVE DATA: CONTEXTUAL

Zhan, Ruohan, Zhimei Ren, **Susan Athey**, and Zhengyuan Zhou. "Policy Learning with Adaptively Collected Data." *Management Science* (2023). https://doi.org/10.1287/mnsc.2023.4921

Zhan, Ruohan, Vitor Hadad, David Hirschberg, and **Susan Athey**. "Off-Policy Evaluation via Adaptive Weighting with Data from Contextual Bandits." *SIGKDD (Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining)* (2021): 2125-2135. https://doi.org/10.1145/3447548.3467456

Methods: Adaptive Experiments with Targeting (“Contextual Bandits”)

CONTEXTUAL BANDIT ALGORITHMS TO LEARN TARGETED TREATMENT ASSIGNMENTS

Carranza, Aldo Gael, Sanath Kumar Krishnamurthy, and **Susan Athey**. “Flexible and Efficient Contextual Bandits with Heterogeneous Treatment Effect Oracle.” in *26th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR* Volume 206, (2023): 7190-7212. https://proceedings.mlr.press/v206/carranza23a/carranza23a.pdf

Dimakopoulou, Maria, Zhengyuan Zhou, **Susan Athey**, and Guido Imbens. “Balanced Linear Contextual Bandits.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, (2019): 3445-3453. https://doi.org/10.1609/aaai.v33i01.33013445

Dimakopoulou, Maria, Zhengyuan Zhou, **Susan Athey**, and Guido Imbens. “Estimation Considerations in Contextual Bandits.” *arXiv preprint* (2017). arXiv:1711.07077

Krishnamurthy, Sanath Kumar, and **Susan Athey**. “Towards Costless Model Selection in Contextual Bandits: A Bias-Variance Perspective.” *forthcoming in the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)* 2024. arXiv:2106.06483v3

Krishnamurthy, Sanath Kumar, Vitor Hadad, and **Susan Athey**. “Adapting to Misspecification in Contextual Bandits with Offline Regression Oracles.” *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR* (2021): 139:5805-5814 arXiv:2102.13240

Krishnamurthy, Sanath Kumar, Vitor Hadad, and **Susan Athey**. “Tractable Contextual Bandits Beyond Realizability.” In *24th International Conference on Artificial Intelligence and Statistics (AISTATS)* 2021, *PMLR* Volume 130, (2021). arXiv:2010.13013

Krishnamurthy, Sanath Kumar, Ruohan Zhan, **Susan Athey**, and Emma Brunskill. “Proportional Response: Contextual Bandits for Simple and Cumulative Regret Minimization.” *Neural Information Processing Systems (NeurIPS)*, (2023). arXiv:2307.02108

Methods: Contextual Bandit Applications

ADAPTIVE EXPERIMENT/CONTEXTUAL BANDIT APPLICATIONS

Athey, Susan, Undral Byambadalai, Vitor Hadad, Sanath Kumar Krishnamurthy, Weiwen Leung, Joseph Jay Williams. "Contextual Bandits in a Survey Experiment on Charitable Giving: Within-Experiment Outcomes versus Policy Learning." *arXiv preprint arXiv:2211:12004* (2022). arXiv:2211.12004

Offer-Westort, Molly, Leah R. Rosenzweig, and **Susan Athey**. "Battling the Coronavirus Infodemic' Among Social Media Users in Africa." *Nature Human Behavior,* 8, pages 823–834, 2024. https://www.nature.com/articles/s41562-023-01810-7

Methods: Causal Panel Data Analysis, Experimental Design for Staggered Rollouts

ANALYSIS OF TREATMENT EFFECTS WITH PANEL DATA

Arkhangelsky, Dmitry, **Susan Athey**, David A. Hirshberg, Guido Imbens, and Stefan Wager, "Synthetic Difference in Differences." *American Economic Review* 11 (12), (2021): 4088-4118. arXiv:1813.09970

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* (2021): 1-41. arXiv:1710.10251

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu. "Ensemble Methods for Causal Effects in Panel Data Settings." In *AEA Papers and Proceedings*, vol. 109 (2019): 65-70. arXiv:1903.10079

Athey, Susan, and Guido Imbens. "Design-based Analysis in Difference-in-differences Settings with Staggered Adoption." *Journal of Econometrics* (2021). arXiv:1808.05293

Donnelly, Rob, Francisco R. Ruiz, David Blei, and **Susan Athey**. "Counterfactual Inference for Consumer Choice Across Many Product Categories." *Quantitative Marketing and Economics* (2021): 1-39. arXiv:1906.02635

Ruiz, Francisco JR, **Susan Athey**, and David M. Blei. "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements." *Annals of Applied Statistics* 14, no. 1 (2020): 1-27. arXiv:1711.03560

DESIGN OF STAGGERED ROLLOUT EXPERIMENTS

Xiong, Ruoxuan, **Susan Athey**, Mohsen Bayati, and Guido Imbens. "Optimal Experimental Design for Staggered Rollouts." *Management Science* (2023). arXiv:1911.03764

Methods: Federated Causal Inference

ESTIMATING CAUSAL EFFECTS WITHOUT COMBINING DATA (FEDERATED CAUSAL INFERENCE)

Carranza, Aldo Gael, and **Susan Athey**. "Federated Offline Policy Learning with Heterogeneous Observational Data." *arXiv preprint* (2023). arXiv:2305.12407

Xiong, Ruoxuan, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T. Vogelstein, and **Susan Athey**. "Federated Causal Inference in Heterogeneous Observational Data." *Statistics in Medicine*. 2023; 42(24): 4418–4439. https://doi.org/10.1002/sim.9868

Methods: Prediction and Causal Inference

PREDICTION VS. CAUSAL INFERENCE

Athey, Susan. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355, no. 6324 (2017): 483-485. https://doi.org/10.1126/science.aal4321

STABLE PREDICTION

Athey, Susan, and Peng Cui, “Stable Learning Establishes Some Common Ground Between Causal Inference and Machine Learning.” *Nature Machine Intelligence* 4, (2022):110–115. https://doi.org/10.1038/s42256-022-00445-z

Kuang, Kun, Peng Cui, **Susan Athey**, Ruoxuan Xiong, and Bo Li. “Stable Prediction Across Unknown Environments.” In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (2018): 1617-1626. https://doi.org/10.1145/3219819.3220082

Kuang, Kun, Ruoxuan Xiong, Peng Cui, **Susan Athey**, and Bo Li. “Stable Prediction with Model Misspecification and Agnostic Distribution Shift.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, (2020): 4485-4492. https://doi.org/10.1609/aaai.v34i04.5876

Methods: Combining Experimental and Observational Data

COMBINING EXPERIMENTAL AND OBSERVATIONAL DATA: SURROGATE METHODS

Athey, Susan, Raj Chetty, and Guido Imbens. "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes." (2020). *arXiv preprint* arXiv:2006.09676

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. "The Surrogate Index: Combining Short-term Proxies to Estimate Long-term Treatment Effects More Rapidly and Precisely." No. w26463. *National Bureau of Economic Research*, 2019. https://www.nber.org/papers/w26463

SURROGATE APPLICATIONS

Athey, Susan, Juan Camilo Castillo, and Bharat Chandar. "Service Quality in the Gig Economy: Empirical Evidence about Driving Quality at Uber." *Available at SSRN 3499781* (2019). https://dx.doi.org/10.2139/ssrn.3499781

Athey, Susan, and Scott Stern. "The Impact of Information Technology on Emergency Health Care Outcomes." *The Rand Journal of Economics* 33, no. 3 (2002): 399-432. https://doi.org/10.3386/w7887

Methods: Treatment Effects in Networks

ANALYSIS OF EXPERIMENTS IN NETWORKS

Athey, Susan, Dean Eckles, and Guido Imbens. "Exact p-values for Network Interference." *Journal of the American Statistical Association* 113, no. 521 (2018): 230-240. https://doi.org/10.1080/01621459.2016.1241178

Referenced Methods Topics and Papers: Artificial Intelligence & Foundation Models

CAUSAL INFERENCE & CONFOUNDING WITH FEATURE EXTRACTION FROM TEXT, IMAGES

Zeng, Jiaming, Michael F. Gensheimer, Daniel L. Rubin, **Susan Athey** & Ross D. Shachter. “Uncovering Interpretable Potential Confounders in Electronic Medical Records.” *Nature Communications* 13, (2022). https://doi.org/10.1038/s41467-022-28546-8

AI FOUNDATION MODELS

Liu, Li-Ping, Francisco JR Ruiz, **Susan Athey**, and David M. Blei. “Context Selection for Embedding Models.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4819-4828. 2017. http://papers.nips.cc/paper/7067-context-selection-for-embedding-models

Rudolph, Maja, Francisco Ruiz, **Susan Athey**, and David Blei. “Structured Embedding Models for Grouped Data.” *Neural Information Processing Systems (NeurIPS), 250-260* (2017). arXiv:1709.10367

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, **Susan Athey**, and David Blei. "CAREER: Transfer Learning for Economic Prediction of Labor Data." *Transactions on Machine Learning (TMLR),* 2024. arXiv:2202.08370

Applications: Implementation of Digital Interventions for Social Impact

Agrawal, Keshav,**Susan Athey**, Ayush Kanodia, and Emil Palikot. "Personalized Recommendations in EdTech: Evidence from a Randomized Controlled Trial." *arXiv preprint* (2022). arXiv:2208.13940

Agrawal, Keshav, **Susan Athey**, Ayush Kanodia, and Emil Palikot. "Digital Interventions and Habit Formation in Educational Technology." *arXiv preprint* (2023). arXiv:2310.10850

Athey, Susan, Katy Bergstrom, Vitor Hadad, Julian C. Jamison, Berk Özler, Luca Parisotto, and Julius Dohbit Sama. "Can Personalized Digital Counseling Improve Consumer Search for Modern Contraceptive Methods?" *Science Advances* (2023). https://doi.org/10.1126/sciadv.adg4420

Athey, Susan, Juan Camilo Castillo, and Bharat Chandar. "Service Quality in the Gig Economy: Empirical Evidence about Driving Quality at Uber." *Available at SSRN 3499781* (2019). https://dx.doi.org/10.2139/ssrn.3499781

Athey, Susan, Shawn Allen Cole, Shanjukta Nath, and S. Jessica Zhu. "Targeting, Personalization, and Engagement in an Agricultural Advisory Service." *Available at SSRN 4536641* (2023). https://dx.doi.org/10.2139/ssrn.4536641

Athey, Susan, Matias Cersosimo, Kristine Koutout, and Zelin Li. "Emotion-versus Reasoning-based Drivers of Misinformation Sharing: A Field Experiment Using Text Message Courses in Kenya." *Available at SSRN 4489759 (2023).* https://dx.doi.org/10.2139/ssrn.4489759

Athey, Susan, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. "Digital Public Health Interventions at Scale: The Impact of Social Media Advertising on Beliefs and Outcomes Related to COVID Vaccines." *Proceedings of the National Academy of Sciences* 120, no. 5 (2023): e2208110120. https://doi.org/10.1073/pnas.2208110120

Athey, Susan, Dean Karlan, Emil Palikot, and Yuan Yuan. "Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces." No. w30633. *National Bureau of Economic Research* (2022). https://doi.org/10.3386/w30633

Athey, Susan, Niall Keleher, and Jann Spiess. "Machine Learning Who to Nudge: Causal vs. Predictive Targeting in a Field Experiment on Student Financial Aid Renewal." *arXiv preprint* (2023). arXiv:2310.08672

Athey, Susan, and Emil Palikot. "Effective and Scalable Programs to Facilitate Labor Market Transitions for Women in Technology." *arXiv preprint* (2022). arXiv:2211.09968

Athey, Susan, and Emil Palikot. "The Value of Non-traditional Credentials in the Labor Market." Unpublished manuscript (2024).

Offer-Westort, Molly, Leah R. Rosenzweig, and **Susan Athey**. "Battling the Coronavirus Infodemic' Among Social Media Users in Africa." *Nature Human Behavior,* 8, pages 823–834, 2024. https://www.nature.com/articles/s41562-023-01810-7

Additional Applications

META-ANALYSIS OF DIGITAL INTERVENTIONS

Athey, Susan, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. “Digital Public Health Interventions at Scale: The Impact of Social Media Advertising on Beliefs and Outcomes Related to COVID Vaccines.” *Proceedings of the National Academy of Sciences* 120, no. 5 (2023): e2208110120. https://doi.org/10.1073/pnas.2208110120