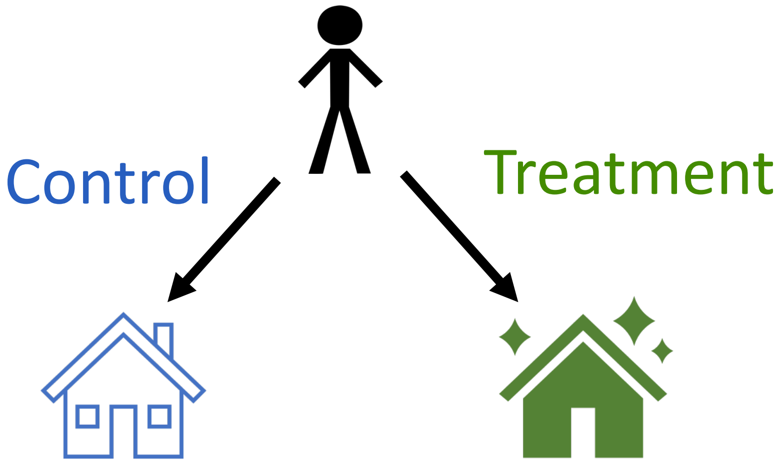
Marketplace Experimentation: Interference, Inference, and Decisions

Hannah Li (MIT)

Simons Workshop on Quantifying Uncertainty

Joint work with Ramesh Johari and Gabriel Weintraub (Stanford)

Decision-making in online marketplaces



"If we show higher quality photos, do the number of bookings increase?"

- Experimentation ("A/B tests")
- Goal: estimate *Global Treatment Effect*

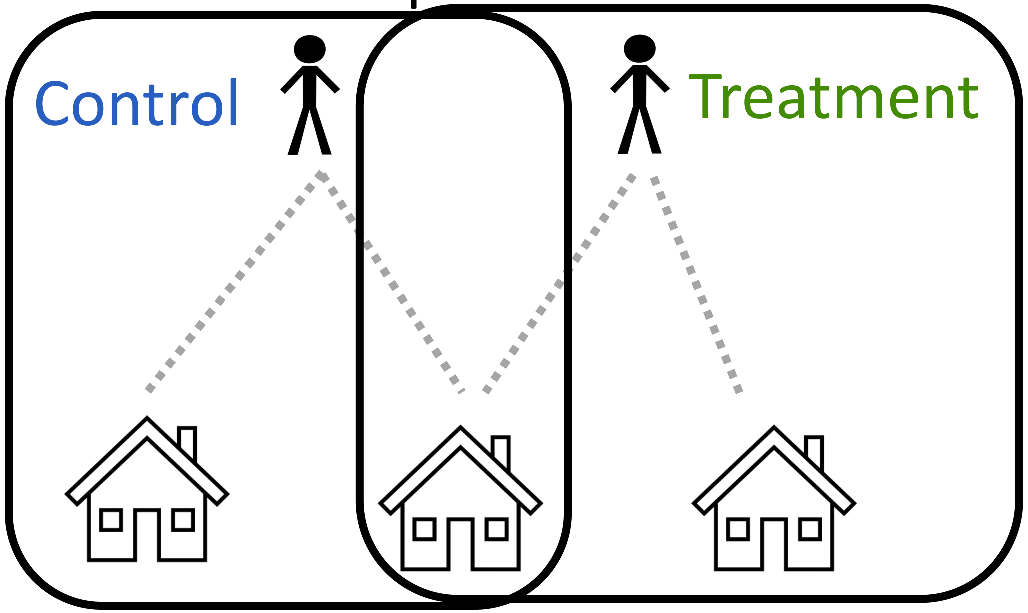
GTE = Bookings in global treatment
− bookings in global control

- Give intervention to some (treatment) and not others (control)
- Large platforms run > 10,000 per year

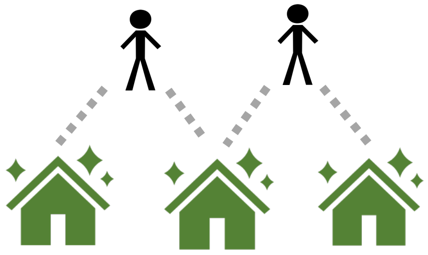
But estimates of GTE in marketplaces often **biased** due to interference!

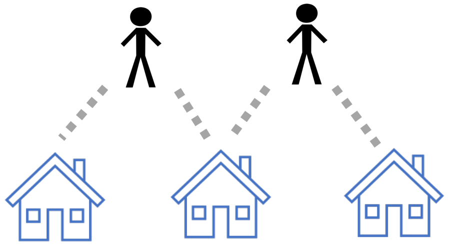
Competition ⟹ Interference ⟹ Bias

Customer-side experiment



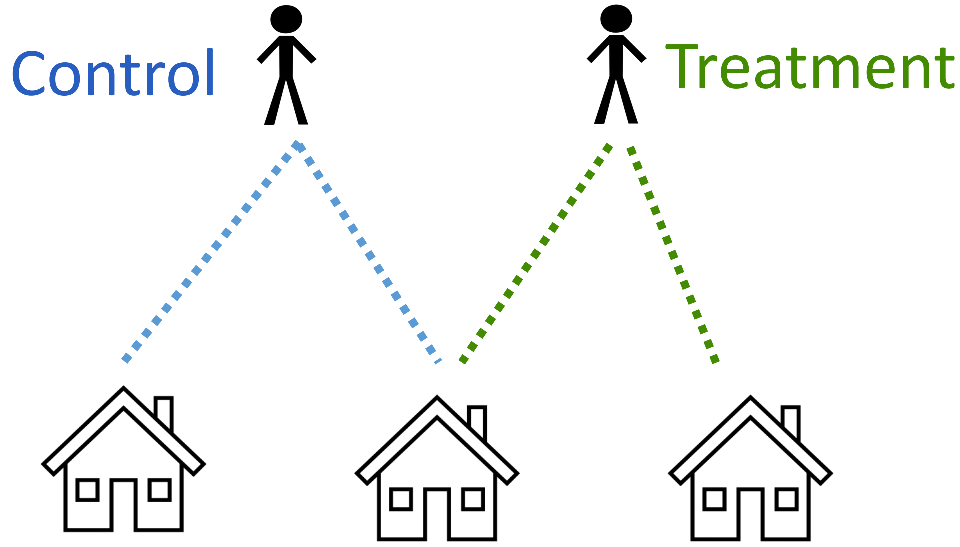
Global Treatment Effect (GTE) = Global Treatment − Global Control





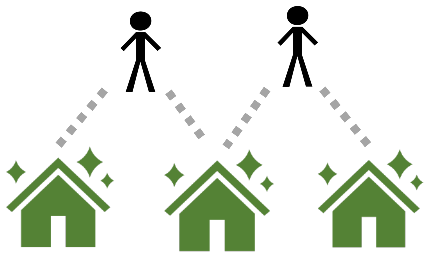
Competition ⟹ Interference ⟹ Bias

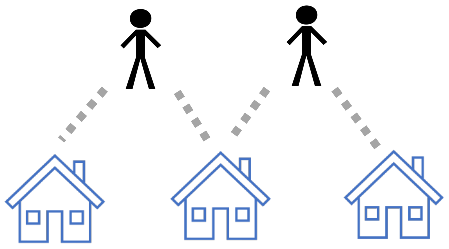
Customer-side experiment



- Suppose feature makes treatment customer more likely to book than control
- Treatment customer books listing
- Reduces supply for control customer
- This instance: overestimate GTE

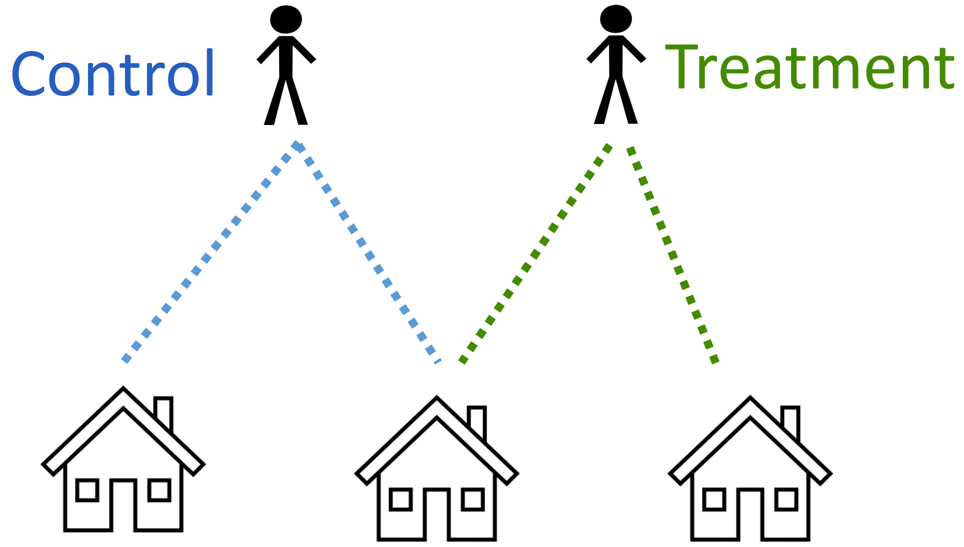
Global Treatment Effect (GTE) = Global Treatment – Global Control





Competition ⟹ Interference ⟹ Bias

Customer-side experiment



- Suppose feature makes treatment customer more likely to book than control
- Treatment customer books listing
- Reduces supply for control customer
- This instance: overestimate GTE

More generally:

- Change a customer's booking prob. ⇒ change supply for other customers
- Change a listing's display ⇒ make other listing relatively more/less attractive

Prior work: Interference and $\widehat{GTE}$-bias

- $\widehat{GTE}$-bias is **30% – 230%** size of GTE. [Blake and Coey '14, Fradkin '19, Holtz et al. '20, Liu et al. '21]
- Methods to reduce $\widehat{GTE}$-bias: Cluster randomization, switchback testing, and TSR. [Holtz '18, Candogan et al. '21, Sneider et al. '19, Glynn et al '20, Bojinov et al. '21, Wager and Xu '19, Ha-Thuc et al. '20, Novak et al. '20, Han et al. '21, Liu et al. '21, Bajari et al. '21, Li et al. '21, Johari et al. '22, Bright et al. '22]
- Size of bias depends on supply and demand imbalance. [Li et al. '21] [Johari et al. '22]

This talk: How do biases affect resulting **decisions**?
Takeaway: Interference creates multiple biases, fixing one bias alone can actually worsen decisions.

Decision-making pipeline

Decision-making pipeline

Given significance level α and launch threshold c:

Run an experiment

Decide whether to launch

$(\alpha = .05)$

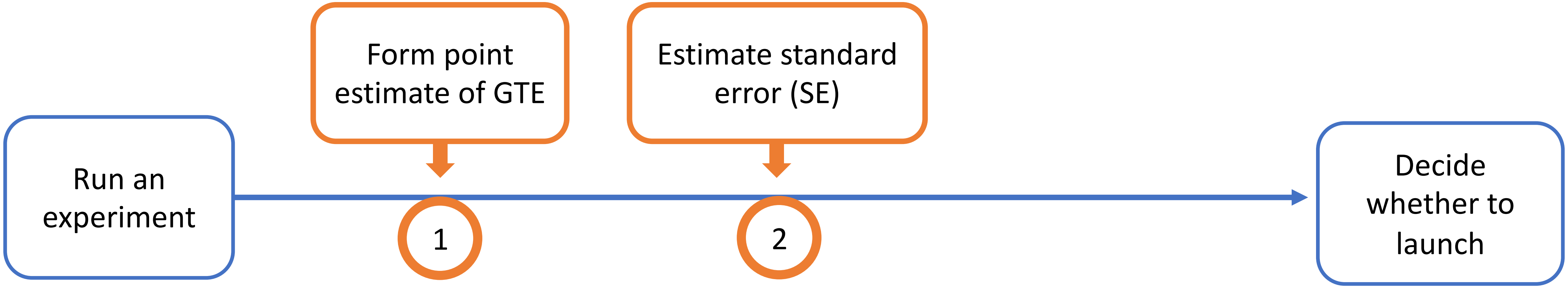
c can represent cost of launching

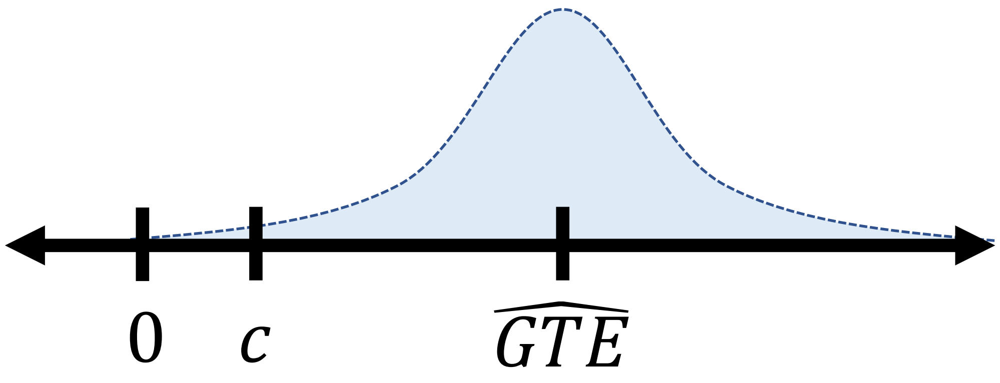
Decision-making pipeline

Given significance level α and launch threshold c:

Decision-making pipeline

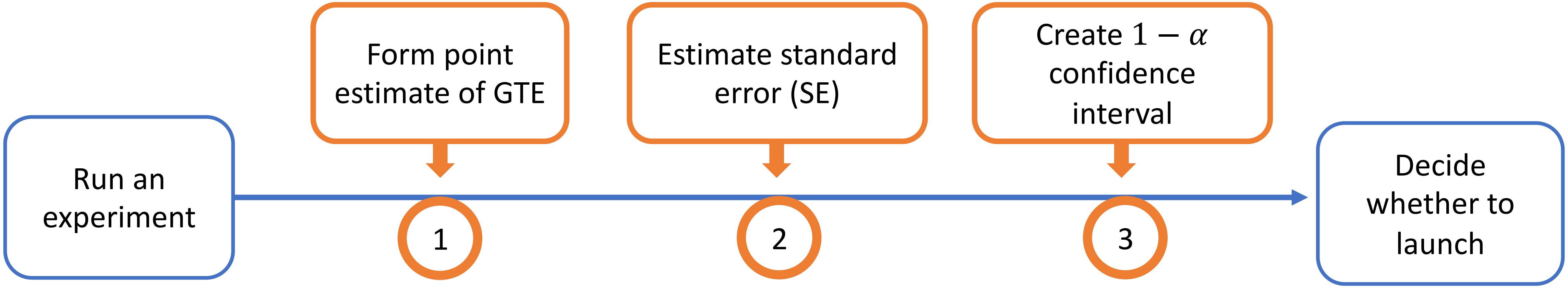
Given significance level α and launch threshold c:

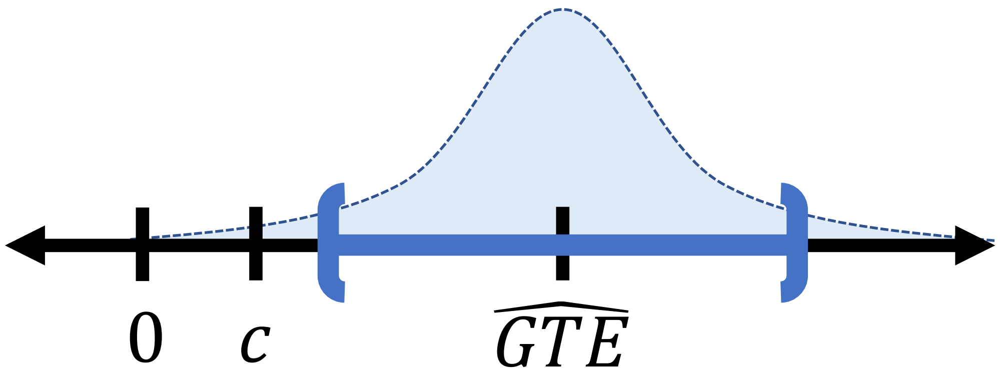




Decision-making pipeline

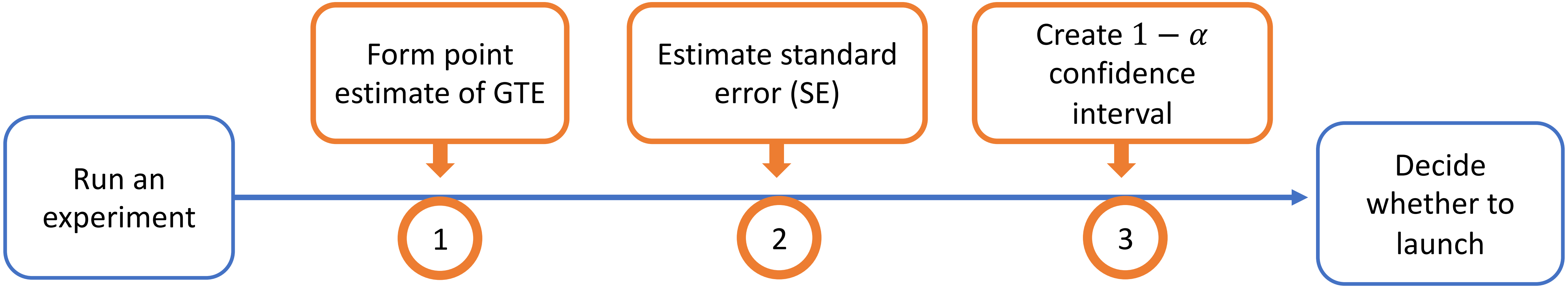
Given significance level α and launch threshold c:



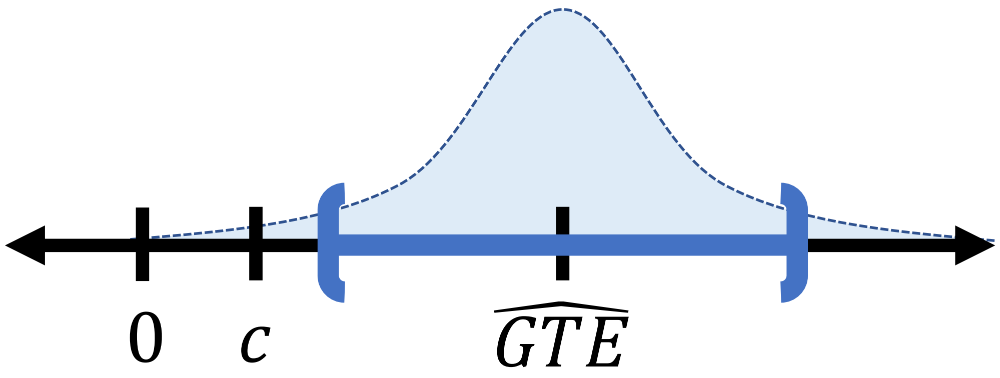


Decision-making pipeline

Given significance level α and launch threshold c:

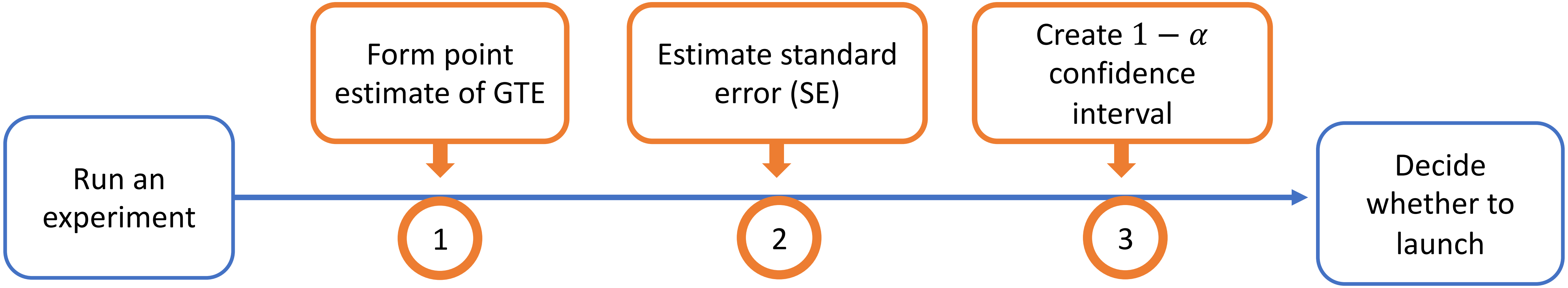


Decision heuristic:
Launch if lower bound of confidence interval $> c$



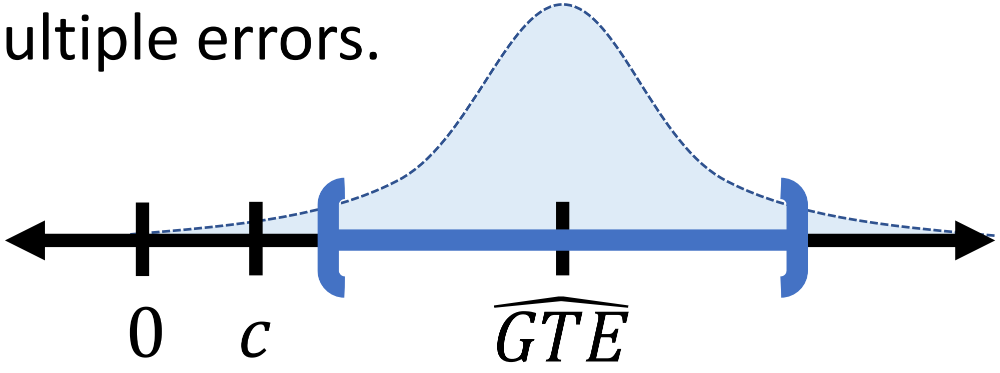
Decision-making pipeline

Given significance level α and launch threshold c:



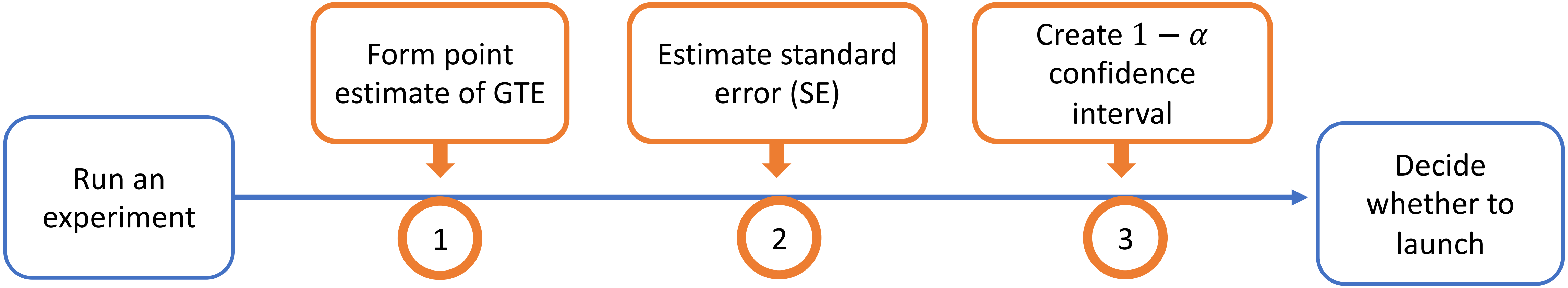
Decision heuristic:
Launch if lower bound of confidence interval $> c$

Interference can create multiple errors.



Decision-making pipeline

Given significance level α and launch threshold c:



Decision heuristic:
Launch if lower bound of confidence interval $> c$

Interference can create multiple errors.

Prior work: Focuses on **(1)**

This work: Studies **(2)**, **(3)**, and impact on decisions

This Work

Use a dynamic market model to study:

1. What biases arise in $\widehat{SE}$ estimates?
2. When/how do biases in $\widehat{GTE}$ and $\widehat{SE}$ ests. affect decision-making?

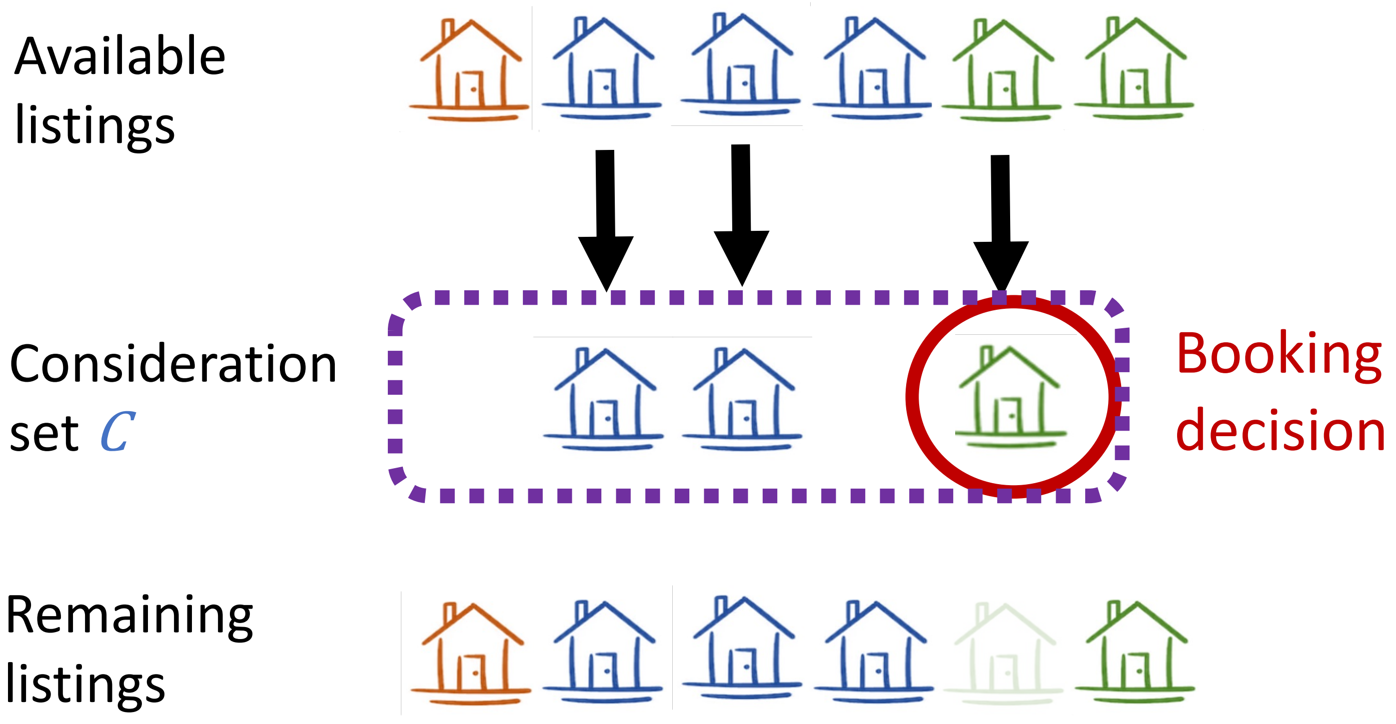
Takeaways:

- In a large class of interventions ("positive interventions"), $\widehat{GTE}$ and $\widehat{SE}$ -bias lead platform to launch too often.
- Two types of biases interact; fixing only one can lead to worse decisions.
- Provide a method to reduce $\widehat{SE}$-bias *and* improve decisions

CTMC model of two-sided markets

[Johari, Li, Liskovich, Weintraub '22]

Customers have type $\gamma \in \Gamma$. Type γ customers arrive at rate Λ_γ (Poisson).



1. Consideration set. Includes each listing l in consideration set w.p. $\alpha_\gamma(\theta_l)$ (independent across listings).

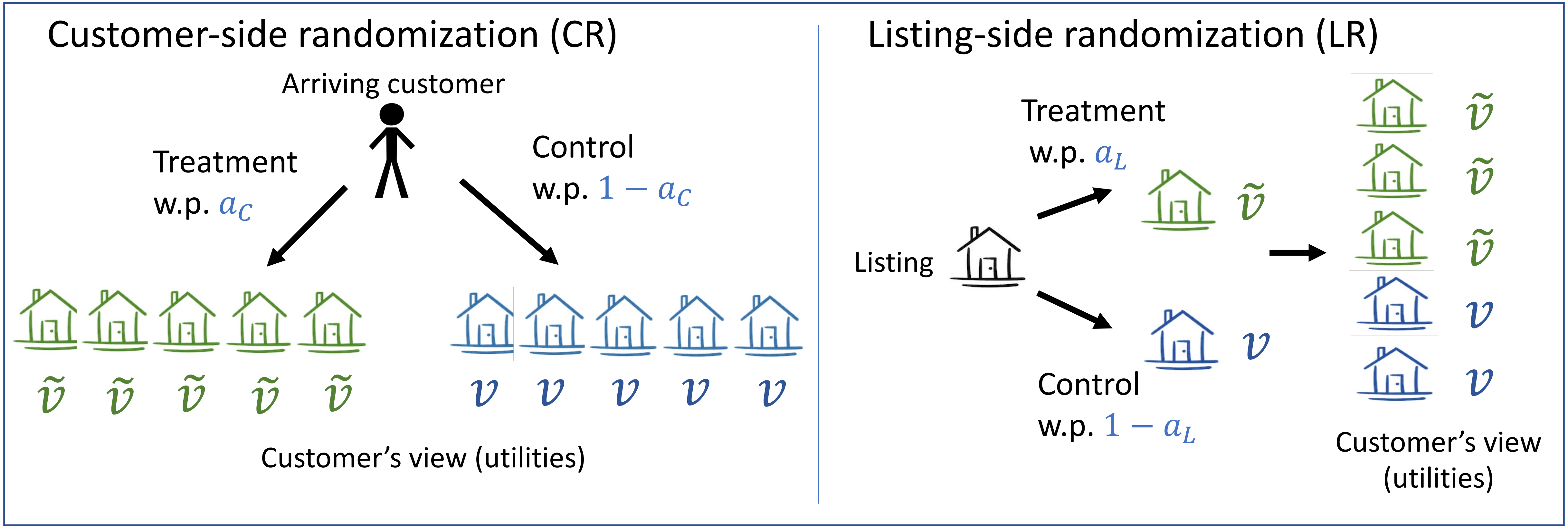
2. Choice. Chooses from consideration set according to multinomial logit model.

$$\mathrm{P}_\gamma(\text{choose } l) = \frac{v_\gamma(\theta_l)}{E_\gamma + \sum_{l' \in C} v_\gamma(\theta_{l'})}$$

Booked listing of type θ becomes unavailable for an exponential time with parameter $\tau(\theta)$.

Running an experiment

We focus on two common types of marketplace experiments.



$$\widehat{GTE}^{CR} = \frac{\#\, Treatment\ Bookings}{a_C\, T} - \frac{\#\, Control\ Bookings}{(1 - a_C)T}$$

Quantities of interest

Study Markov chain behavior in counterfactual worlds

(global treatment, global control, experiment)

- Estimand: Global Treatment Effect GTE – evaluated in steady state

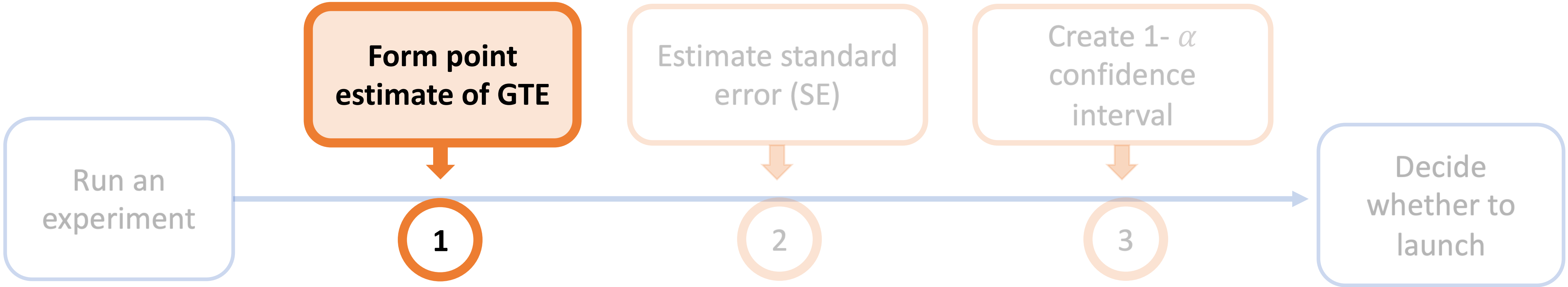
$$GTE = \underbrace{Q^{GT}}_{\text{Global Treatment rate of booking}} - \underbrace{Q^{GC}}_{\text{Global Control rate of booking}}$$

- Estimator (calculated from experiment booking rates):

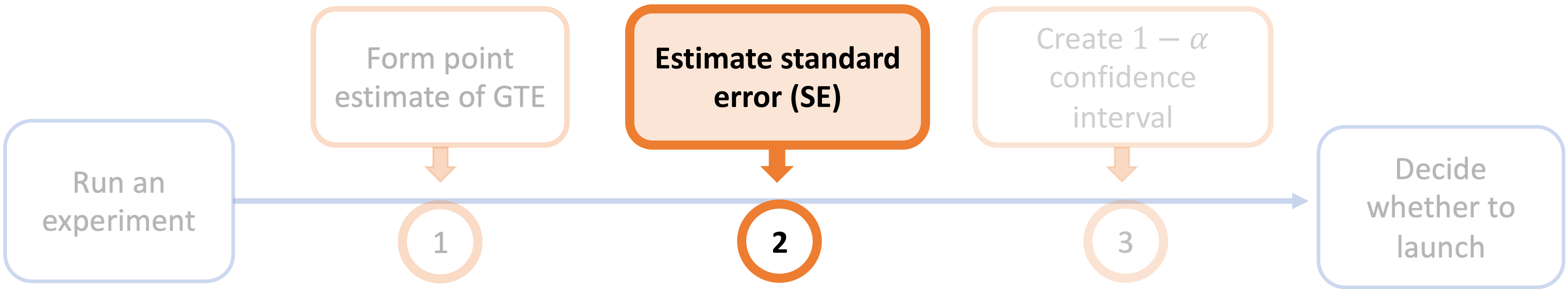
$$\widehat{GTE} = \frac{\#\,Treatment\ Bookings}{a_C\,T} - \frac{\#\,Control\ Bookings}{(1-a_C)T}$$

- Standard Error $\quad SE = \left(Var\left(\widehat{GTE}\right)\right)^{1/2}$

This talk: Focus on a class of interventions that increases utilities, denoted “positive” interventions



Theorem (informal) [JLLW ‘21]:
For a positive intervention, CR and LR *overestimate* the magnitude of the GTE.



Run an experiment
Form point estimate of GTE
1
Estimate standard error (SE)
2
Create $1 - \alpha$ confidence interval
3
Decide whether to launch

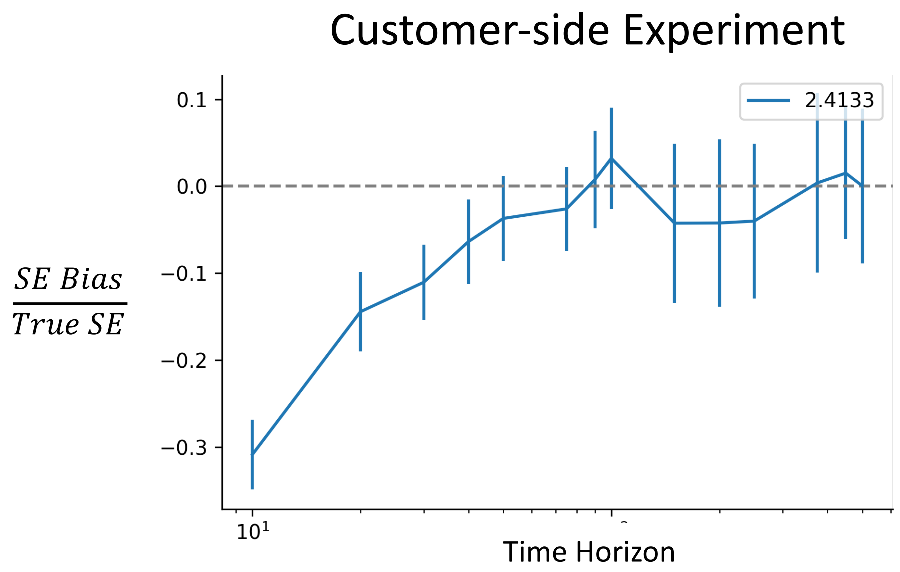
(2) Inference and SE estimation

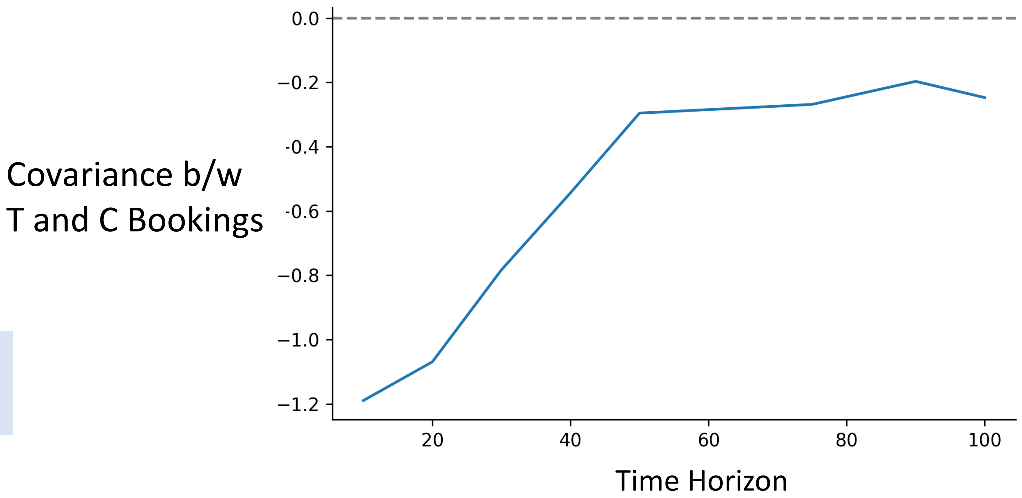
- Estimate $SE = \left(Var\left(\widehat{GTE}\right)\right)^{1/2}$
- "Naive" $\widehat{SE}$ estimate: **Assume individuals are independent**
- Leads to biased estimates of SE

$$Var(T - C)$$
$$= Var(T) + Var(C) - Cov(T, C)$$

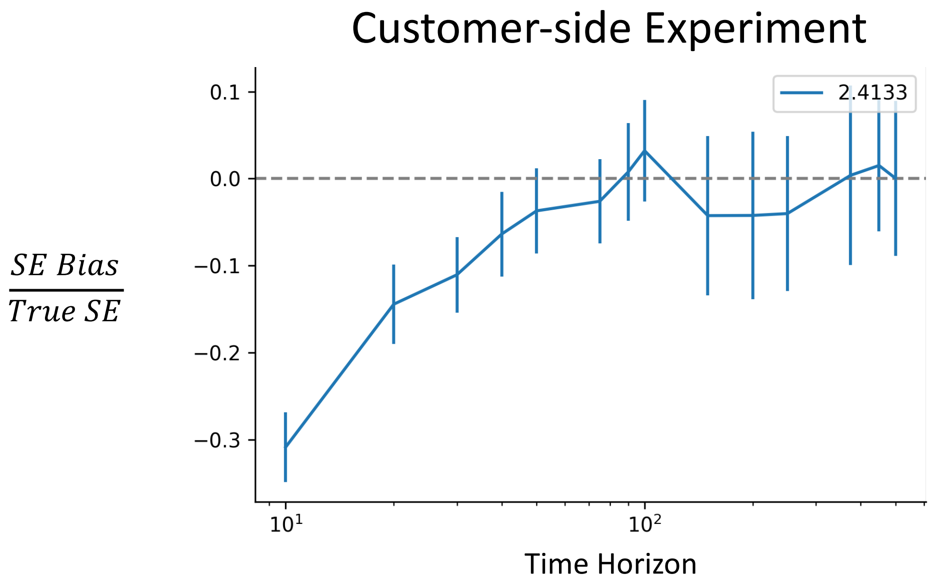
- Ignores correlation between individual outcomes

Competition $\Longrightarrow$ Interference $\Longrightarrow$ Bias





Reducing $\widehat{SE}$ bias
Method 1: Longer experiments



Theorem (informal).

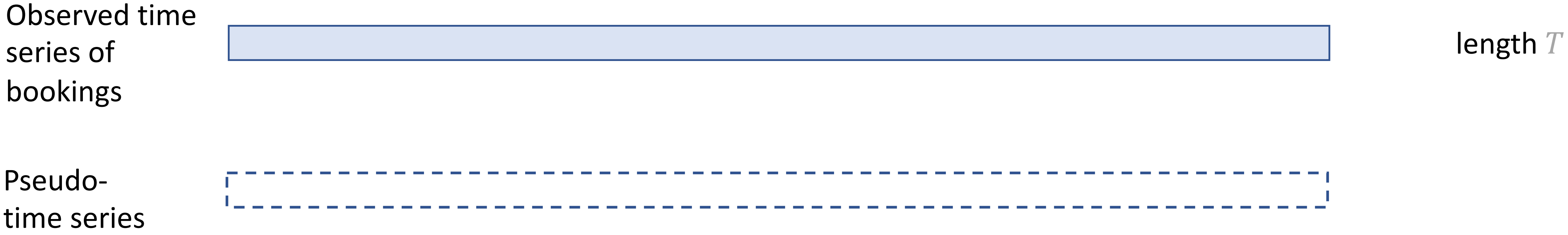
For a customer-side experiment, the bias of the “naive” $\widehat{SE}$ estimate approaches 0 as $T \to \infty$.

Proof idea.

System is a regenerative process.

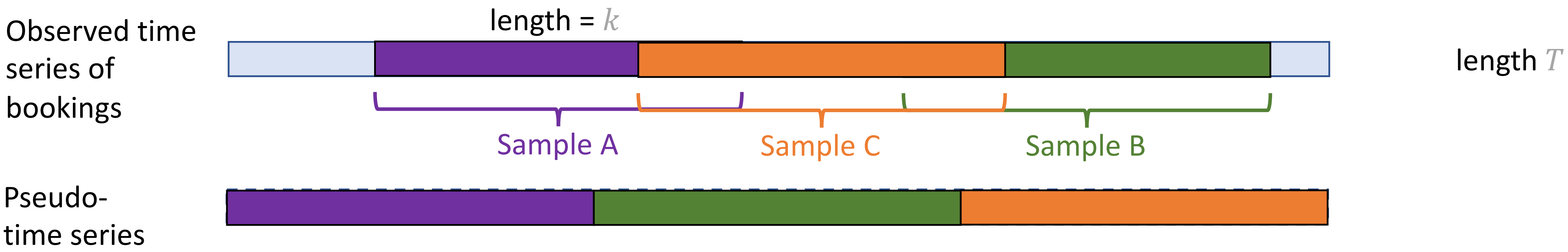
Reducing $\widehat{SE}$ bias
Method 2: Block bootstrap

- Standard bootstrap: resample individuals
- Block bootstrap [Hardle et al. '03]
  - Resample "blocks" from observed time series, create "pseudo-time series"



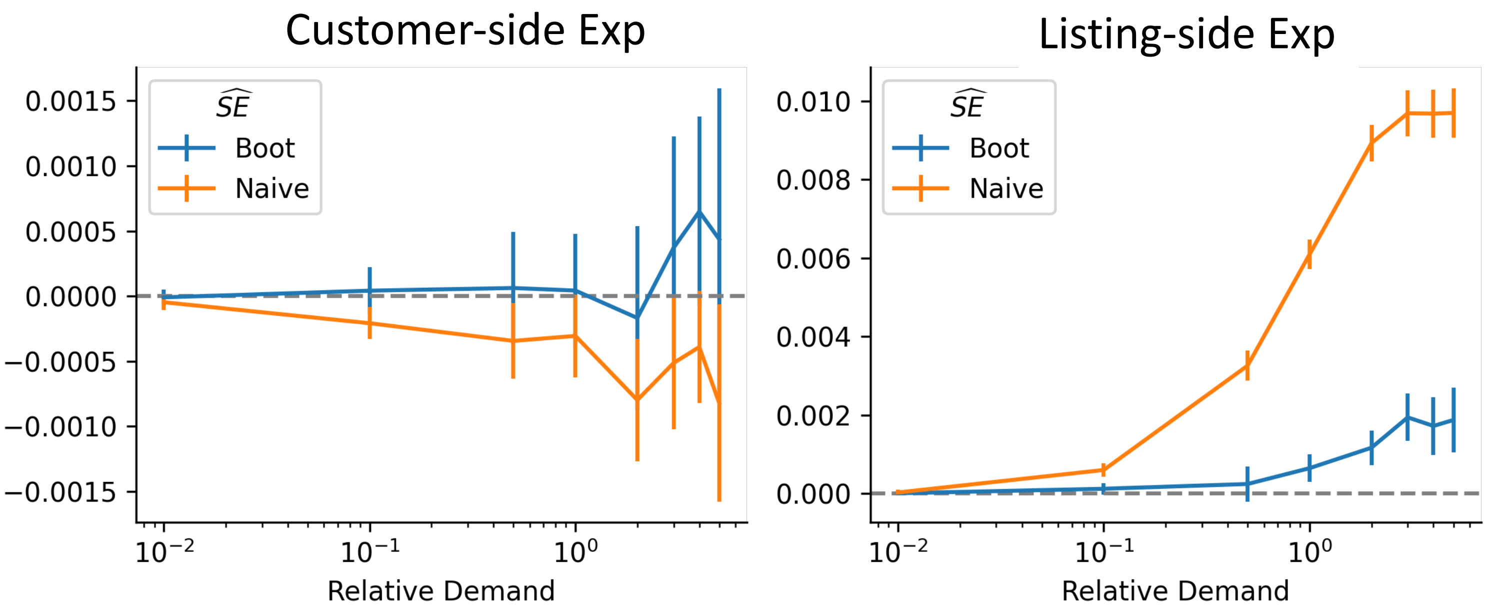
Reducing $\widehat{SE}$ bias
Method 2: Block bootstrap

- Standard bootstrap: resample individuals
- Block bootstrap [Hardle et al. '03]
  - Resample "blocks" from observed time series, create "pseudo-time series"



- From each bootstrap run b (pseudo-time series): calculate $\widehat{GTE}_b$
- Repeat B times, calculate std. dev. across $\widehat{GTE}_b$ estimates $\rightarrow \widehat{SE}^{boot}$

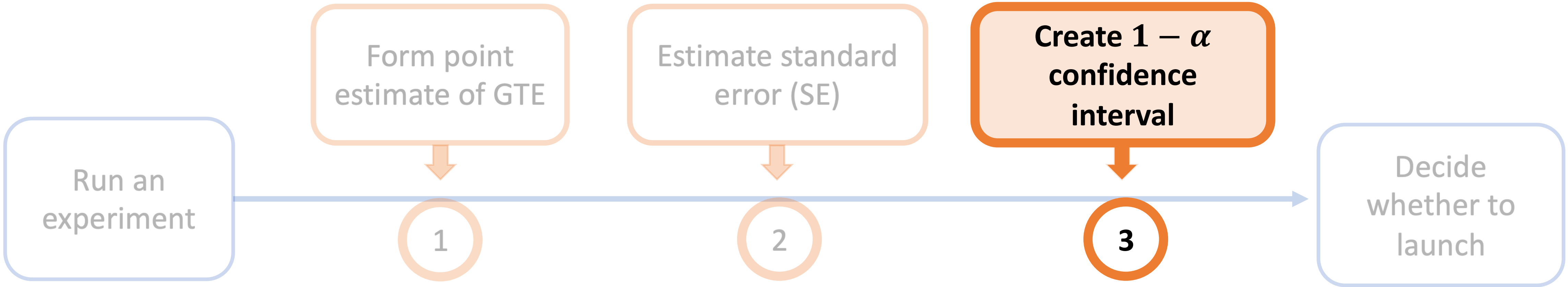
Reducing SE bias
Method 2: Block bootstrap



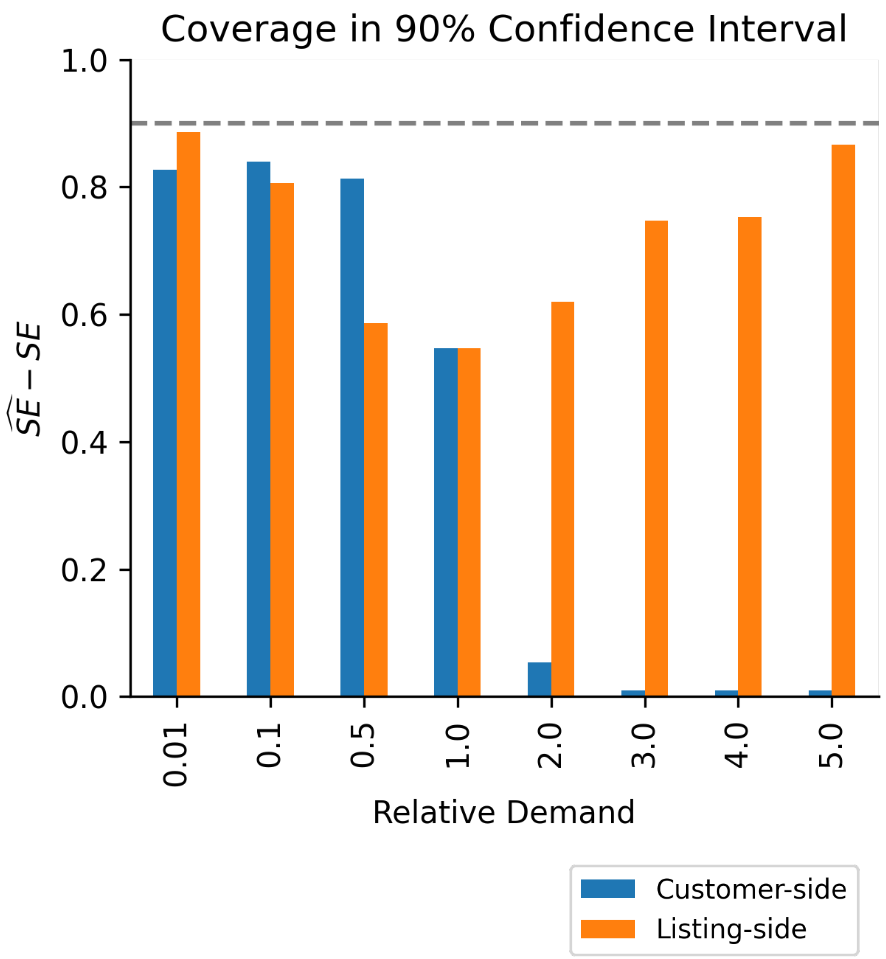
Takeaway: Bootstrapping can mitigate biases.

Caveat: Need to tune block length.

Decision-making pipeline



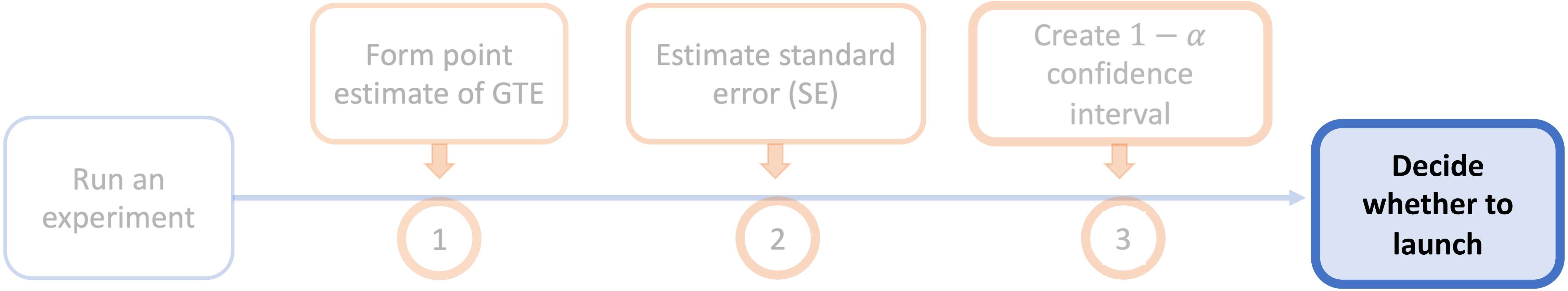
(3) Coverage of confidence intervals



- $\widehat{GTE}$-bias shifts confidence intervals
- $\widehat{SE}$-bias changes width of intervals
- Interactions between $\widehat{GTE}$-bias and $\widehat{SE}$-bias determine coverage

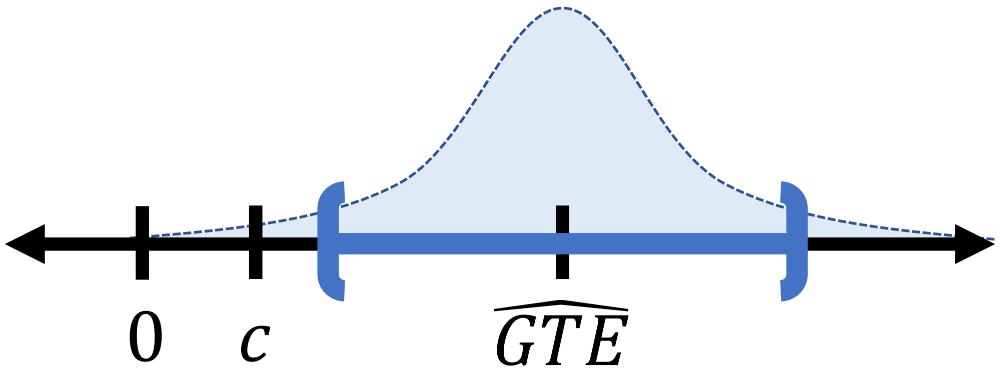
We characterize asymptotic coverage of conf. ints. as a function of $\widehat{GTE}$-bias, $\widehat{SE}$-bias, and SE

Decision-making pipeline



Implications for decision-making

Goal:	Launch if $GTE > c$.
Decision Heuristic:	Launch if lower bound of conf int > c.
Evaluating Decision:	Decision is correct if we launch only when $GTE > c$.



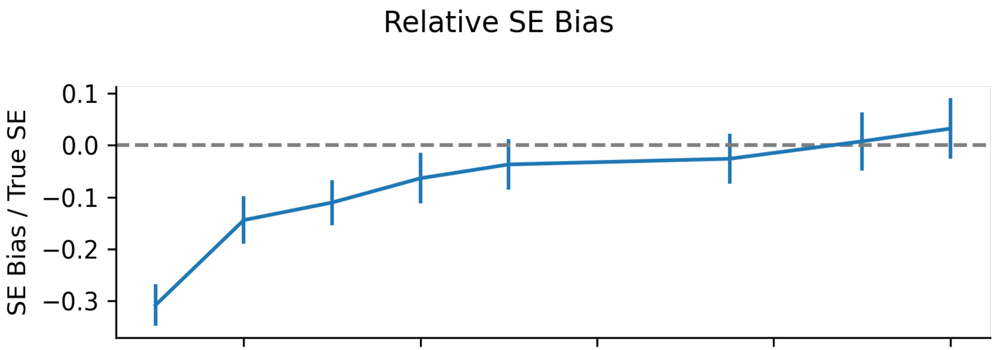
In positive interventions, we see:

1. Overestimation of GTE in CR and LR experiments
2. Underestimation of SE in CR experiments

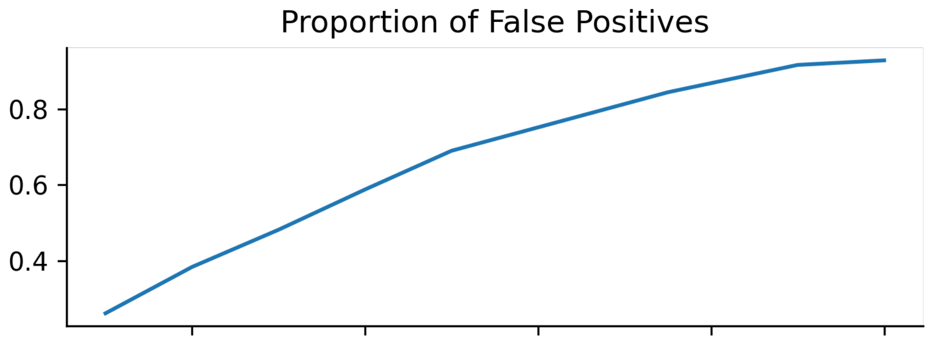
Combination leads to more **false positives** (launch feature when $GTE < c$).

Fixing bias ≠ improving decisions

- Scenario: $GTE < c$
- Any decision to launch is a "false positive"
- Implement **Method 1** for SE-bias reduction: Run longer experiment
- As time horizon increases:
  - Actual SE of CR estimator decreases together with $\widehat{SE}$ bias
  - More confident about our biased $\widehat{GTE}$

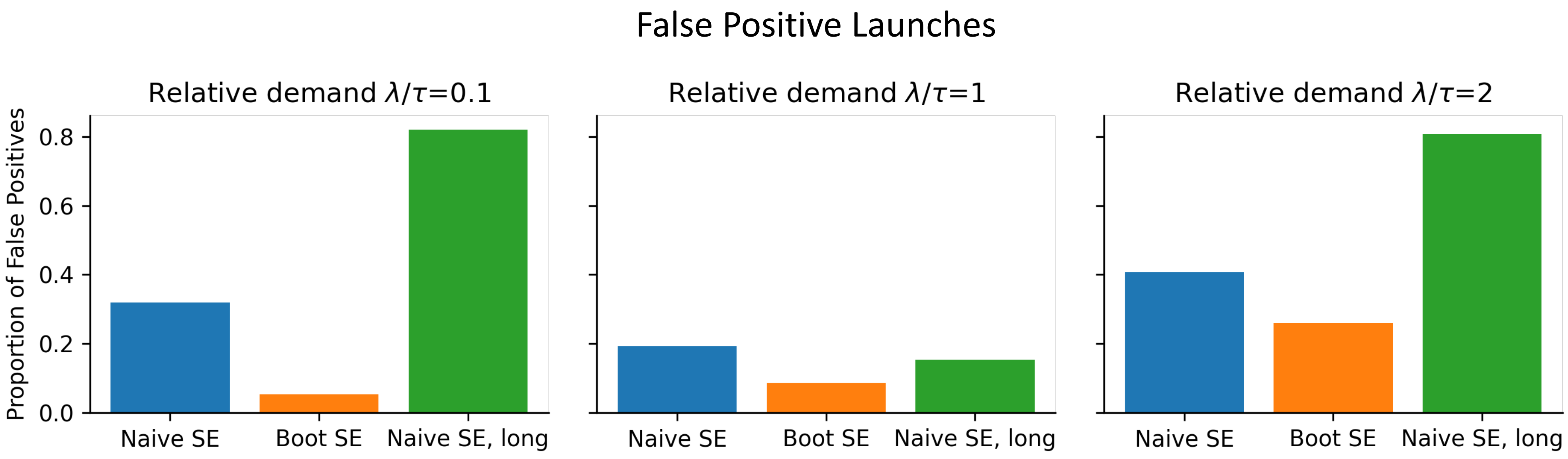


Time Horizon



Alternative: Reduce $\widehat{SE}$ bias with bootstrap

- Scenario: $GTE < c$
- (With appropriate block length) bootstrap method reduces SE bias and reduces false positive launches



Takeaways

- $\widehat{GTE}$ and $\widehat{SE}$-biases interact and cause incorrect decisions
- Propose two methods to reduce $\widehat{SE}$-bias
  1. Increasing time horizon – can worsen decisions
  2. Block bootstrapping – can improve decisions

Open questions

- Combining $\widehat{SE}$-bias reduction with $\widehat{GTE}$-bias reduction
- Increased attention on decisions made from experiments
- Marketplace interactions complicate many statistical methods. How do complications interact with the ways platforms utilize experiments?
  - e.g., simultaneous experiments, ramp-up experiments