Incrementality Bidding & Attribution

Randall Lewis
Director of Economics

Jeffrey Wong
Senior Research Scientist

NETFLIX

Abstract: The causal effect of showing an ad to a potential customer versus not, commonly referred to as “incrementality,” is the fundamental question of advertising effectiveness. In digital advertising three major puzzle pieces are central to rigorously quantifying advertising incrementality: ad buying/bidding/pricing, attribution, and experimentation. Building on the foundations of machine learning and causal econometrics, we propose a methodology that unifies these three concepts into a computationally viable model of both bidding and attribution which spans the randomization, training, cross validation, scoring, and conversion attribution of advertising’s causal effects. Implementation of this approach is likely to secure a significant improvement in the return on investment of advertising.

Contributors:

- **Management & Ad Operations:** Gagan Hasteer, Kelly Uphoff, Steve McBride, Michael Pow, James Ouska
- **Ad Tech Engineering:** Duo Wang, Kai Hu, Raghu Srinivasan, Devesh Parekh, Stephen Walz
- **Science & Algorithms:** Jeffrey Wong, Vijay Bharadwaj, Benoit Rostykus, Tony Jebara, Dave Hubbard

Introduction to Incrementality

Incrementality
Bidding & Attribution

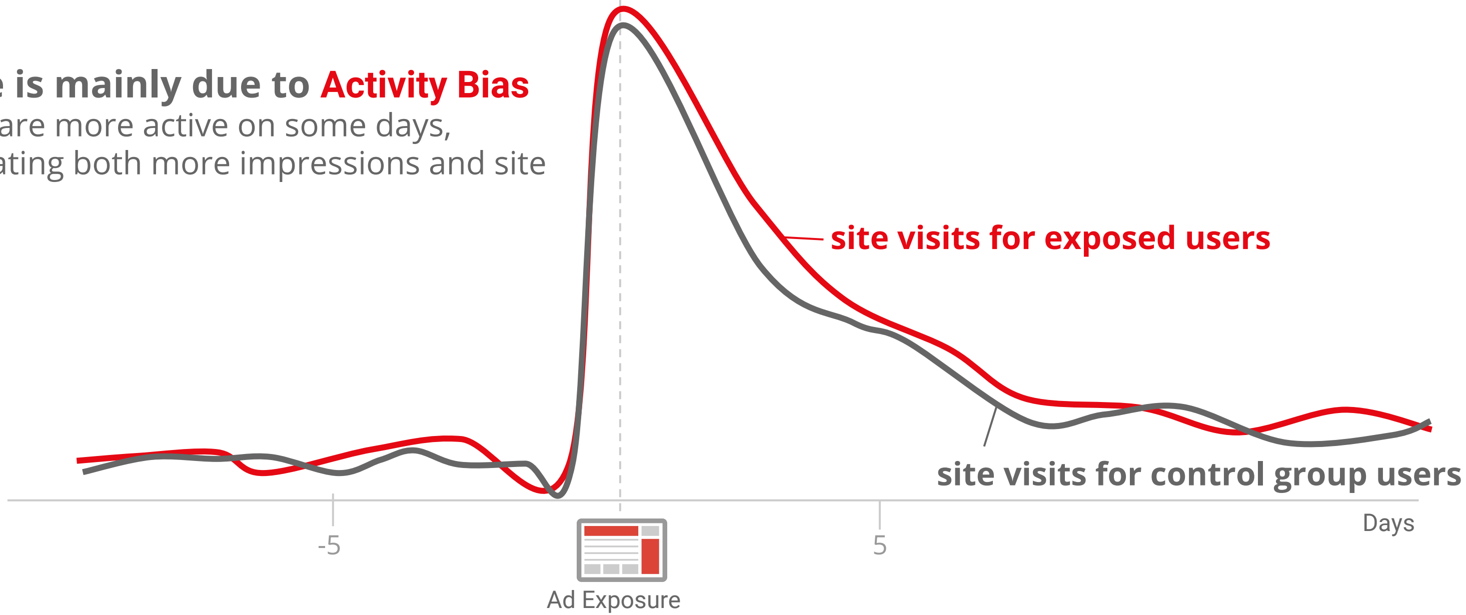
Why Incrementality Matters: Examples of Ad Effectiveness Failures & Challenges

NETFLIX

Advertisers struggle to measure ad effectiveness

Does this Yahoo display campaign drive site visits?

Spike is mainly due to Activity Bias
Users are more active on some days, generating both more impressions and site visits.



Lewis, R.; Rao, J.;, Reiley, D. (2011), “Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising”, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2080235. This chart is a stylized representation of their results.

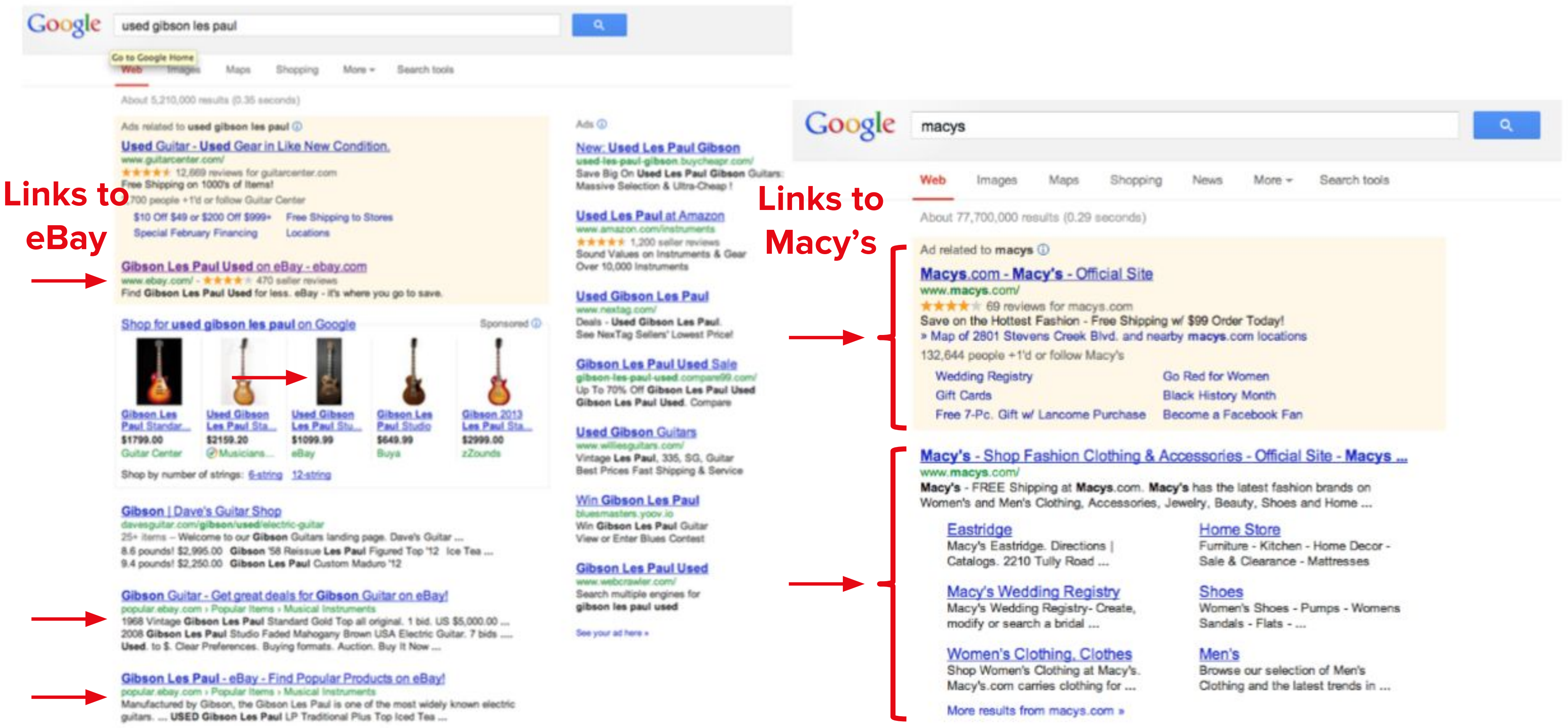
Is Correlation = Causation?

Correlation:
"Measuring the online sales impact of an online [search] ad is straightforward: We determine who has viewed the ad, then **compare online purchases made by those who have and those who have not seen [the ad].**"

--Harvard Business Review article by Magid Abraham, comScore

eBay Ad Tests

Figure 1: Google Ad Examples

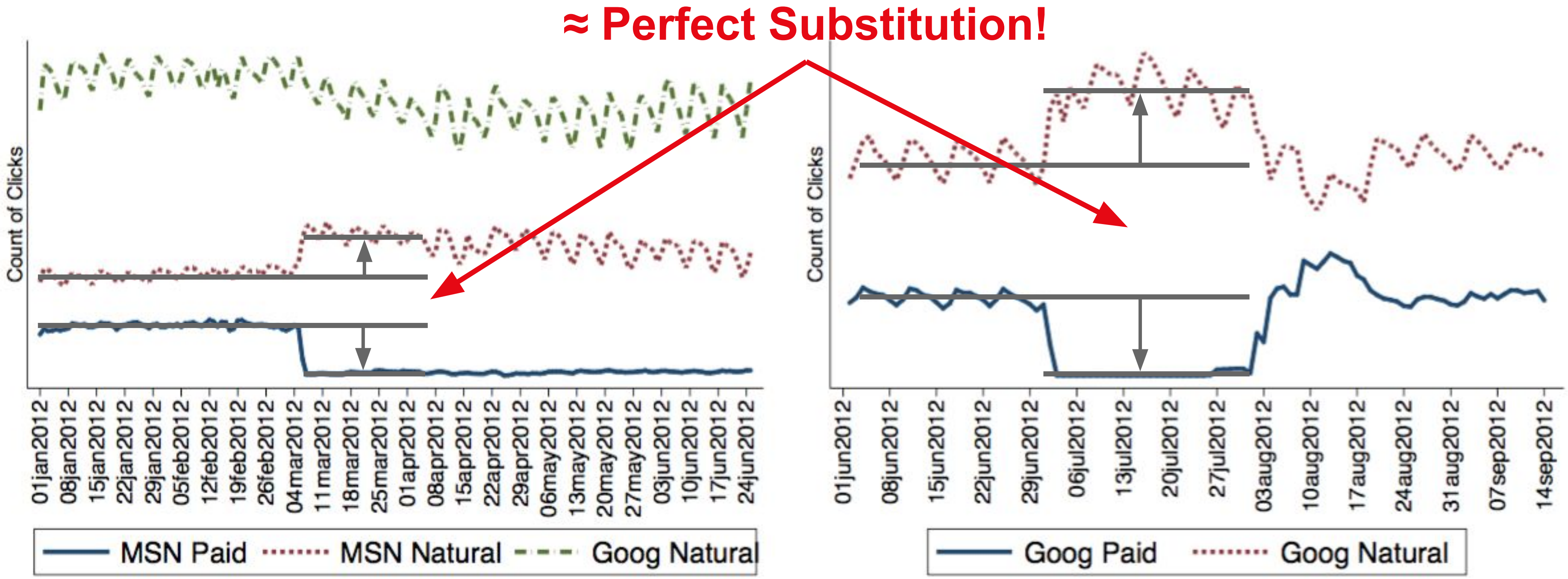


(a) Used Gibson Les Paul

(b) Macys

eBay Ad Tests

Figure 2: Brand Keyword Click Substitution



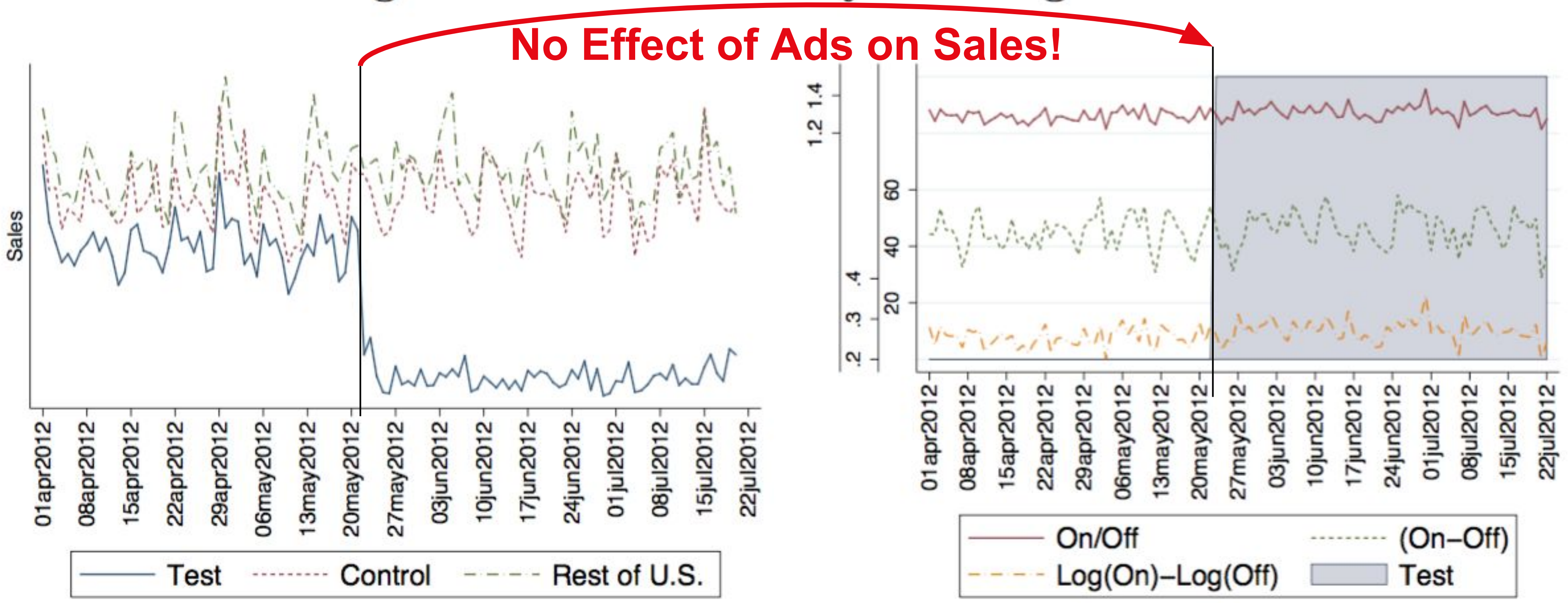
(a) MSN Test

(b) Google Test

Note: MSN and Google click traffic is shown for two events where paid search was suspended (Left) and suspended and resumed (Right).

eBay Ad Tests

Figure 3: Non-Brand Keyword Region Test



(a) Attributed Sales by Region

(b) Differences in Total Sales

Correlation is NOT Causation!

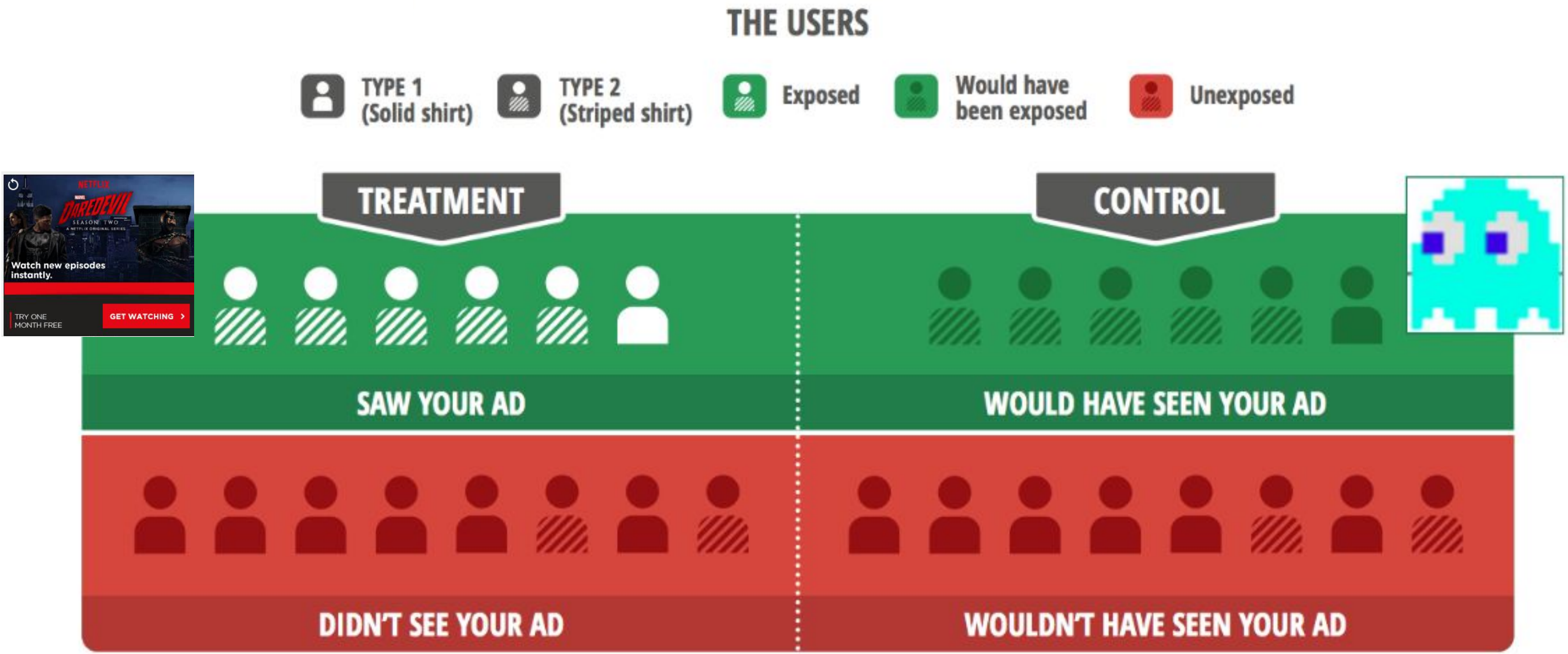
eBay Search Ad Effectiveness

- Blake, Nosko, and Tadelis (2014) "Consumer Heterogeneity in Paid Search Effectiveness," *Econometrica*.
- Compare standard industry practice with natural and controlled field experiments.
- Find **>$50M/year** spent on branded and unbranded search ads **yielded little impact on sales**.
- However, “Consumer Heterogeneity” provides opportunities for eBay to improve the performance of their search advertising expenditures.

Defining Incrementality

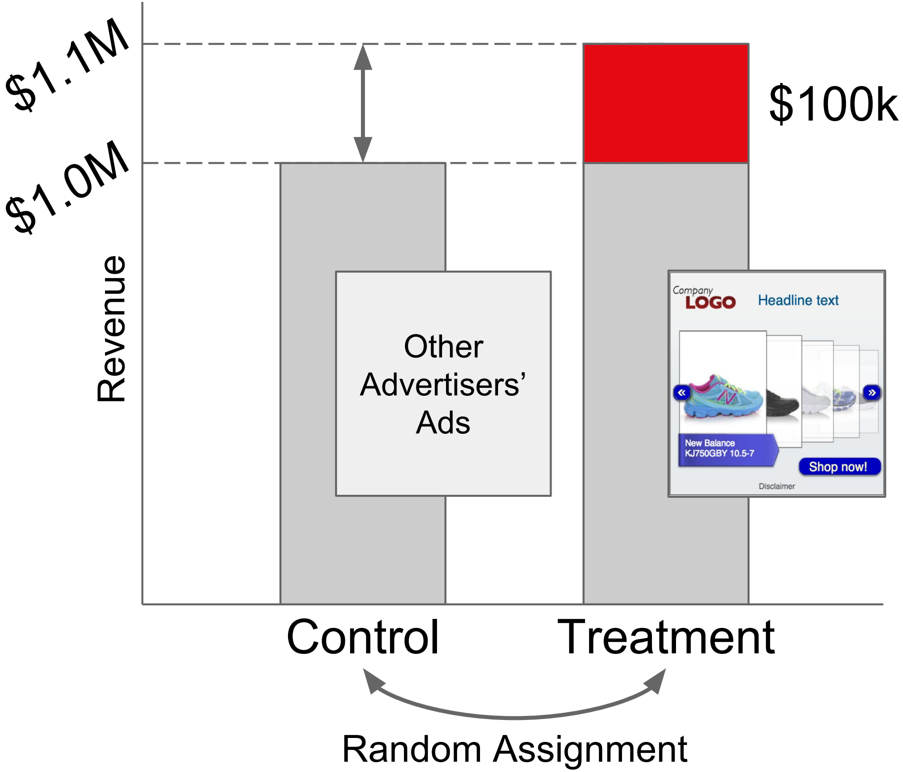
NETFLIX

Ghost Ads: Who Would Have Seen My Ad?



Johnson, G.; Lewis, R.; Rao, J.;, Nubbemeyer, E. (2017), “Ghost Ads: Improving the Economics of Measuring Ad Effectiveness,” *Journal of Marketing Research. Draft available at SSRN:* https://papers.ssrn.com/abstract=2620078.

Incrementality: The Causal Effect of an Ad



Example from "Ghost Ads": Sporting goods retailer who ran an experiment:

- Retargeting
- 570k users
- 2 weeks
- 9 million impressions
- Ad spend: $30,500
- Avg. CPM = **$3.40**

Incrementality: The difference in the outcome because the ad was shown; the causal effect of the ad.

Per impression:
$100k/9M=$0.011 ⇒ **$11 RPM**

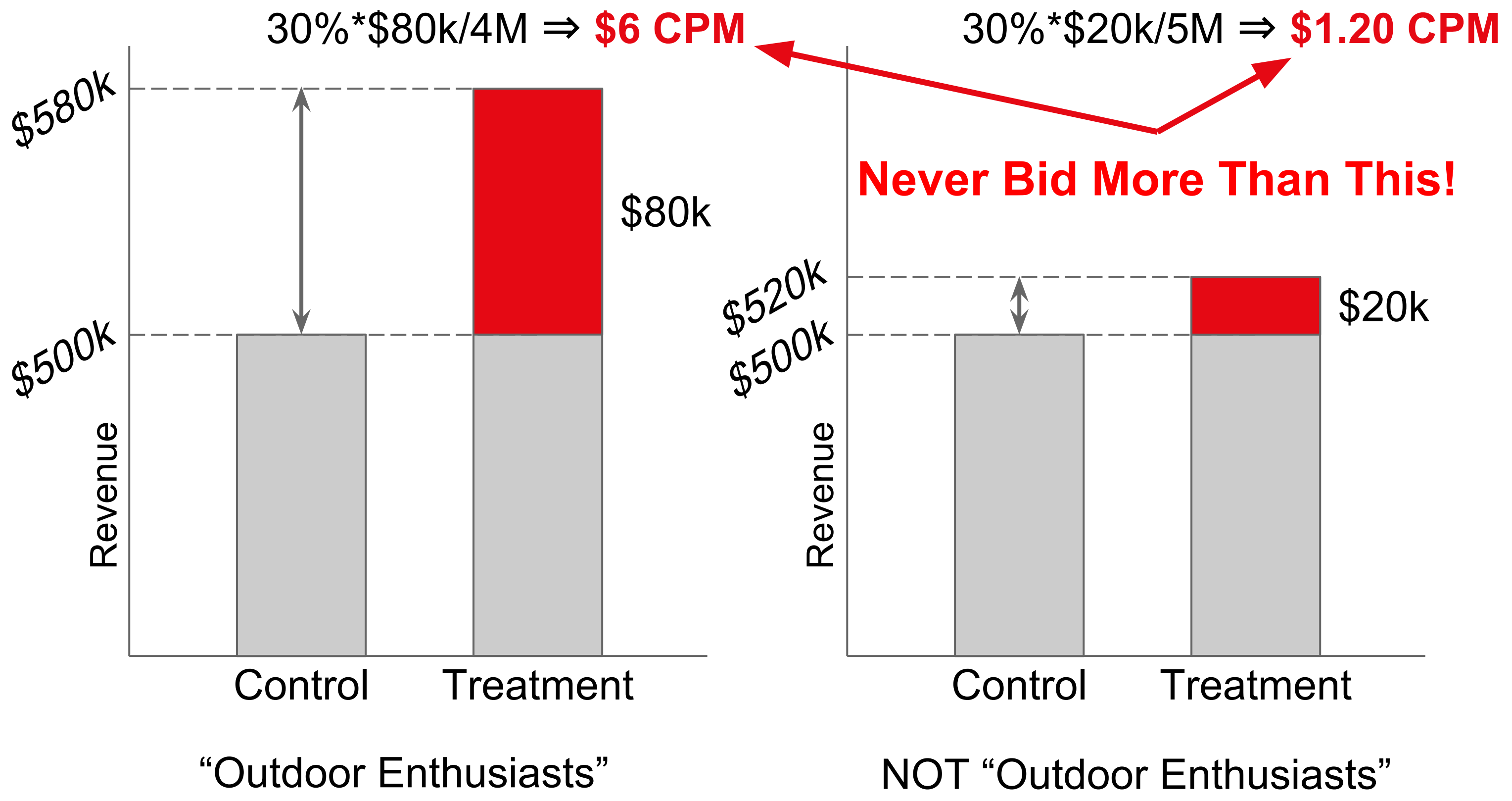
Johnson, Garrett A. and Lewis, Randall A. and Nubbemeyer, Elmar I, Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness (January 12, 2017). Simon Business School Working Paper No. FR 15-21. Available at SSRN: https://ssrn.com/abstract=2620078

Optimizing Incrementality

Introduction to
Incrementality

NETFLIX

Optimizing Incrementality via Attribution

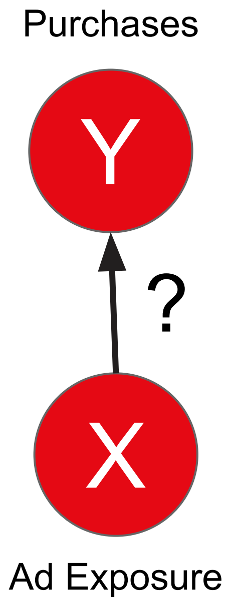


Stylized example not from "Ghost Ads" paper.

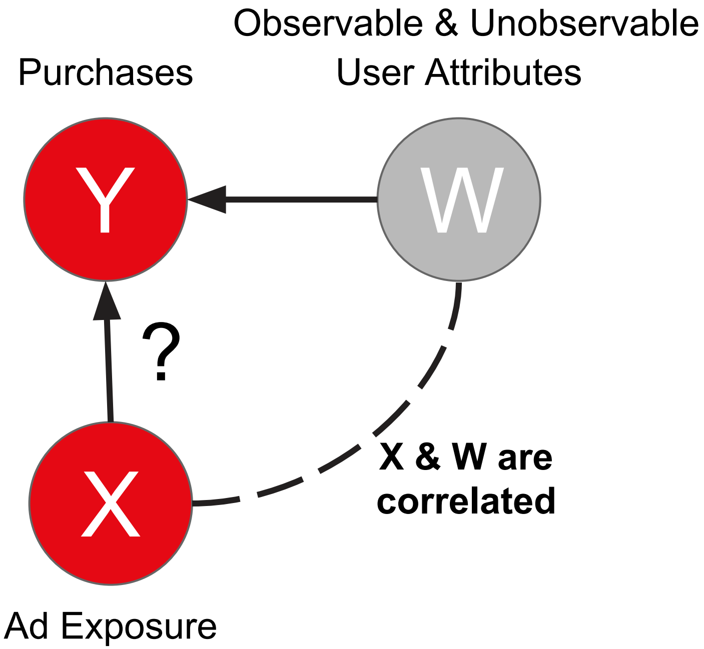
Estimating Incrementality

NETFLIX

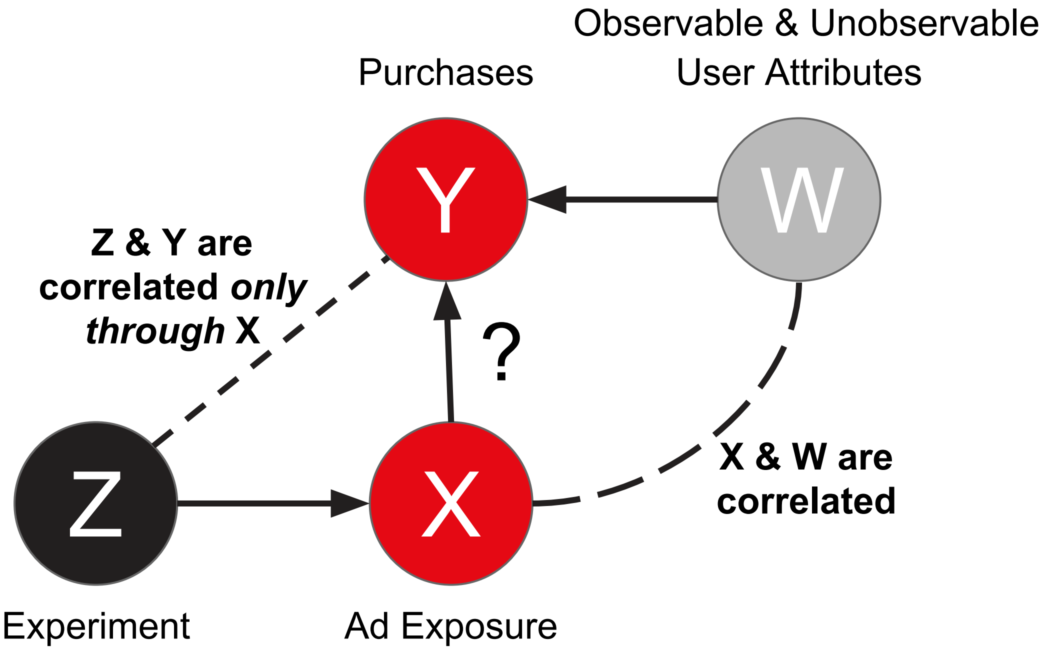
Understanding Causal Estimation



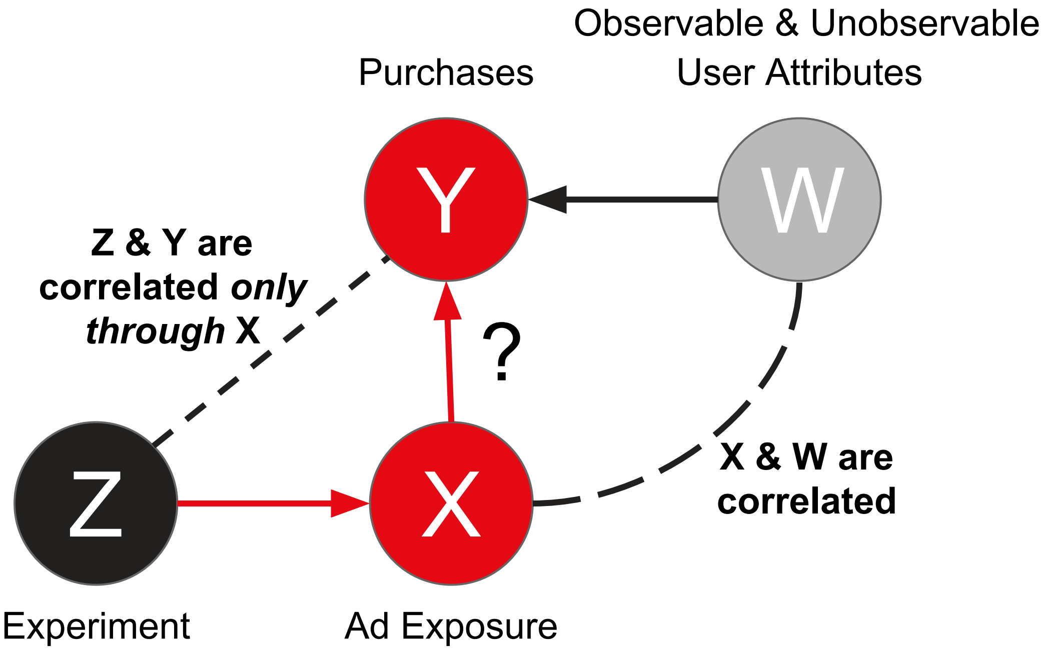
Understanding Causal Estimation



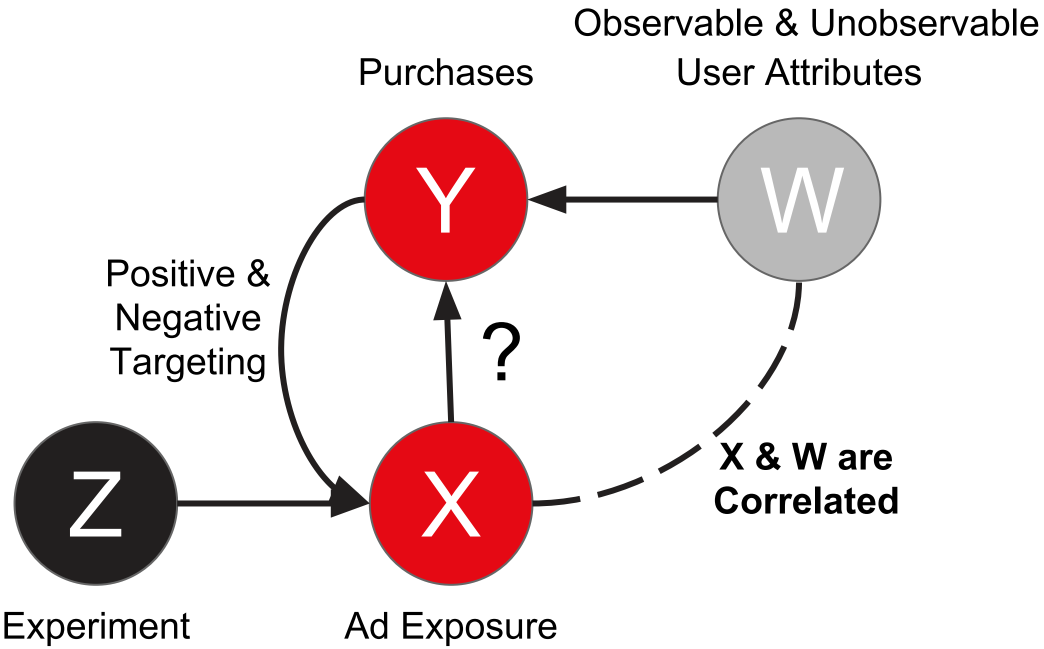
Understanding Causal Estimation



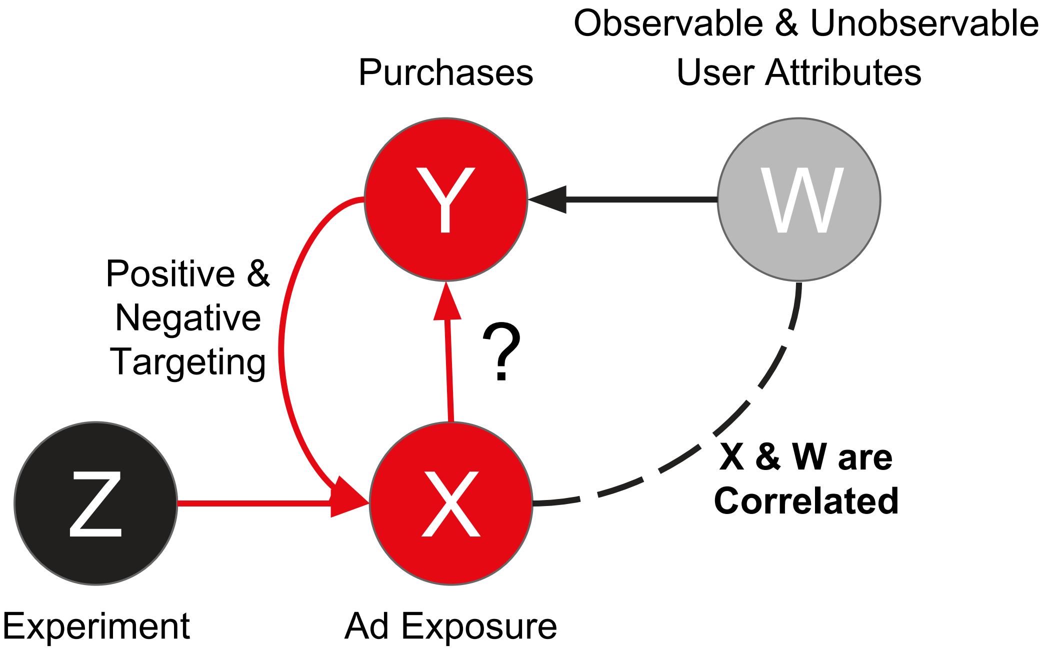
Understanding Causal Estimation



Understanding Causal Estimation

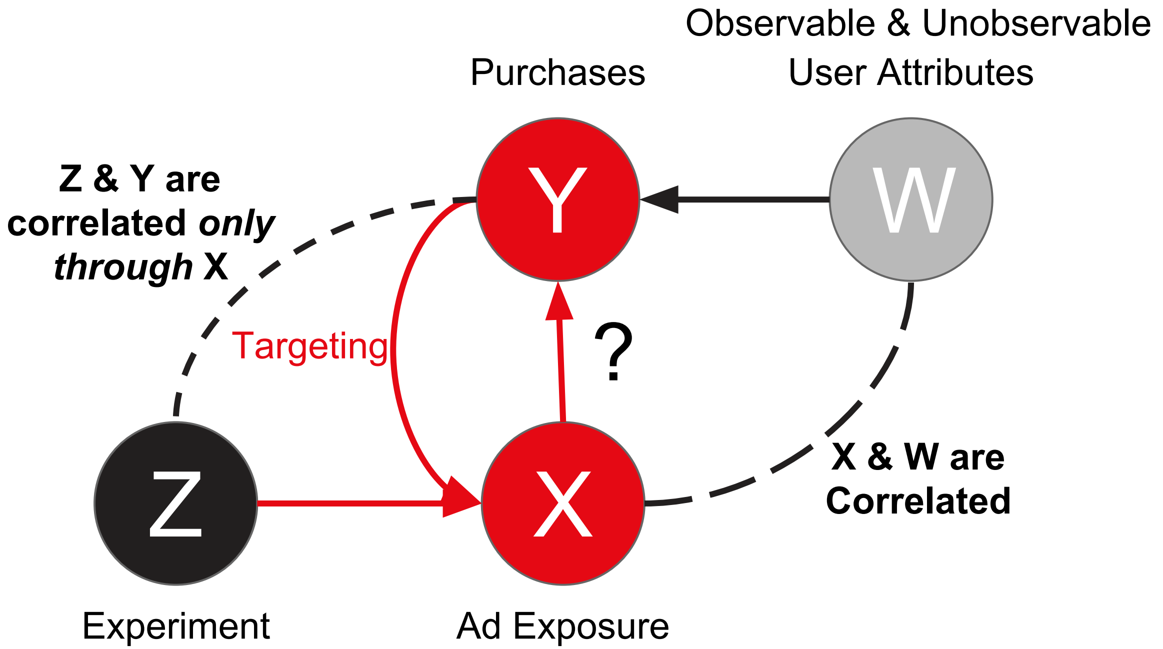


Understanding Causal Estimation



NETFLIX

Understanding Causal Estimation



A Simple Incrementality Model: Heterogeneous Treatment Effects

- **Simple Incrementality Model:** Effect of ads on purchases.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\Delta y_i = E[y_i | show\ ad] - E[y_i | don't\ show\ ad] = \beta$$

- **Heterogeneous Treatment Effects:** Differential effects from different types of ads.

$$y_i = \alpha(W) + \sum_k \beta_k x_{ik} + \varepsilon_i$$

$$\Delta y_{ij} = \sum_k \beta_k x_{ijk}$$

E.g., different weights for "Outdoor Enthusiasts," "country=USA," "ad_size=300x250," etc.

$$x_{ik} = \sum_{j \in impressions} x_{ijk};\ x_{ijk} \equiv 1(impression\ j\ has\ characteristic\ k)$$

Instrumental Variables (IV): Estimating a Causal Model

- **"Second Stage":** Causal effect of ads on purchases.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- **"First Stage":** Causal effect of randomized experiment on ad exposure.

$$x_i = \pi_0 + \pi_1 z_i + \nu_i$$

- **2-Stage Least Squares (2SLS):** Efficient causal estimation of an IV model.

$$Z'(Y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta}_{2SLS} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}\left(X'Z(Z'Z)^{-1}Z'Y\right)$$

Instrumental Variables (IV): Heterogeneous Treatment Effects

- **"Second Stage":** Heterogeneous causal effect of ads on purchases.

$$y_i = \alpha(W) + \sum_k \beta_k x_{ik} + \varepsilon_i$$

- **"First Stage":** Causal effect of randomized experiment on ad exposure.

$$x_{ik} = \pi_{0,k}(W) + \sum_{k'} \pi_{k'} z_{ik'} + \nu_{ik}$$

- **2-Stage Least Squares (2SLS):** Efficient causal estimation of an IV model.

$$Z'(Y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta}_{2SLS} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}\left(X'Z(Z'Z)^{-1}Z'Y\right)$$

Advanced Incrementality for Industry

Incrementality
Bidding & Attribution

Challenges to Incrementality

- >10 billion auctions per day
- >1 billion users per month
- Inexpensive ad impressions
- Sparse conversions
- Continuous stream of data
- Low signal to noise
- High dimensionality
- Correlation != Causation?
- Opportunity cost of experimentation
- Advertiser awareness & demand

Solutions for Incrementality

- $\boxed{\text{Data Volume} \Rightarrow \text{Downsampling}}$
- Causal panel data econometrics
- High impression volume ⇒ Modeling
- Sparse conversions ⇒ “Small Data”
- $\boxed{\text{Continuous-Time Modeling}}$
- Ghost Ads/Bids
- Scalable Sparse IV
- $\boxed{\text{Hausman Causal Correction}}$
- Thompson Sampling, Bayesian Bootstrap
- Critical mass of advertiser demand

Modeling Incrementality in Continuous Time with Ad Stock

Advanced Incrementality
for Industry

NETFLIX

Examples of Ad Effectiveness

Lewis & Reiley 2013:

- Super Bowl 2012 Commercials
- Post-Commercial Search Spikes

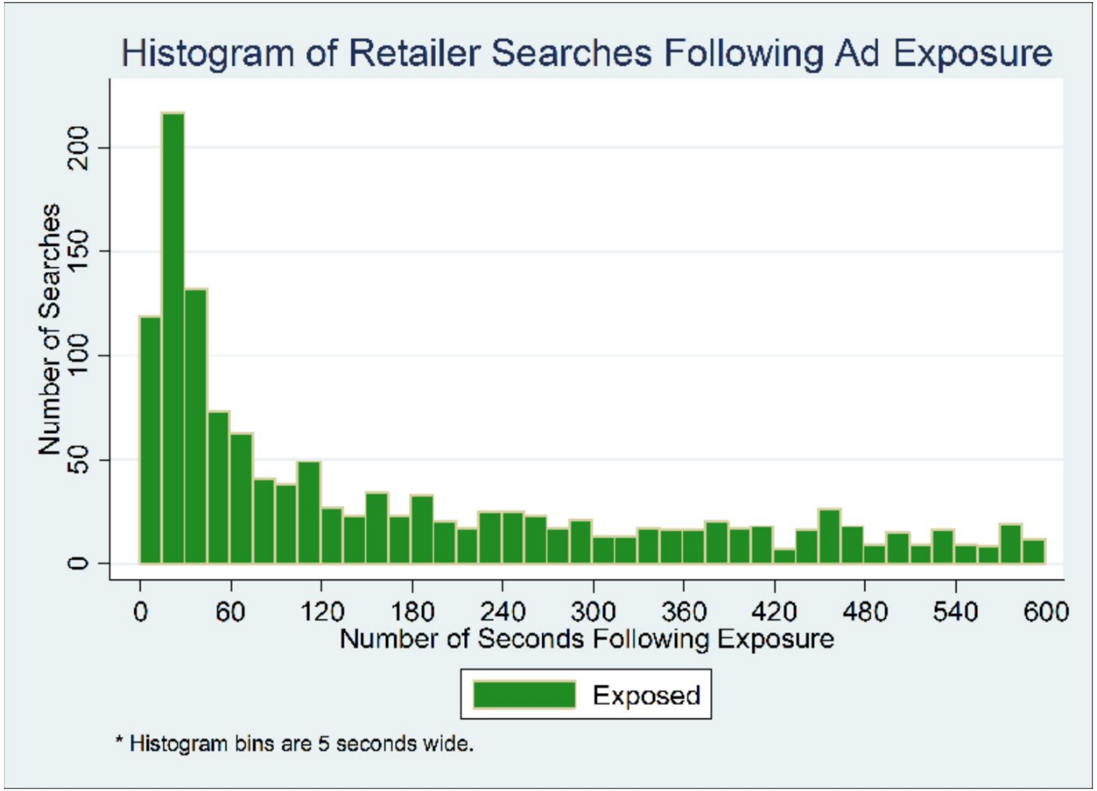
Lewis, Rao, & Reiley 2012:

- Online display ads
- Post-impression search spikes (baseline & lift)

Key Insight:

- Ad effects vary with time
- Modeling can improve statistical power

Examples of Ad Effectiveness



Lewis & Reiley 2013:

- Super Bowl 2012 Commercials
- Post-Commercial Search Spikes

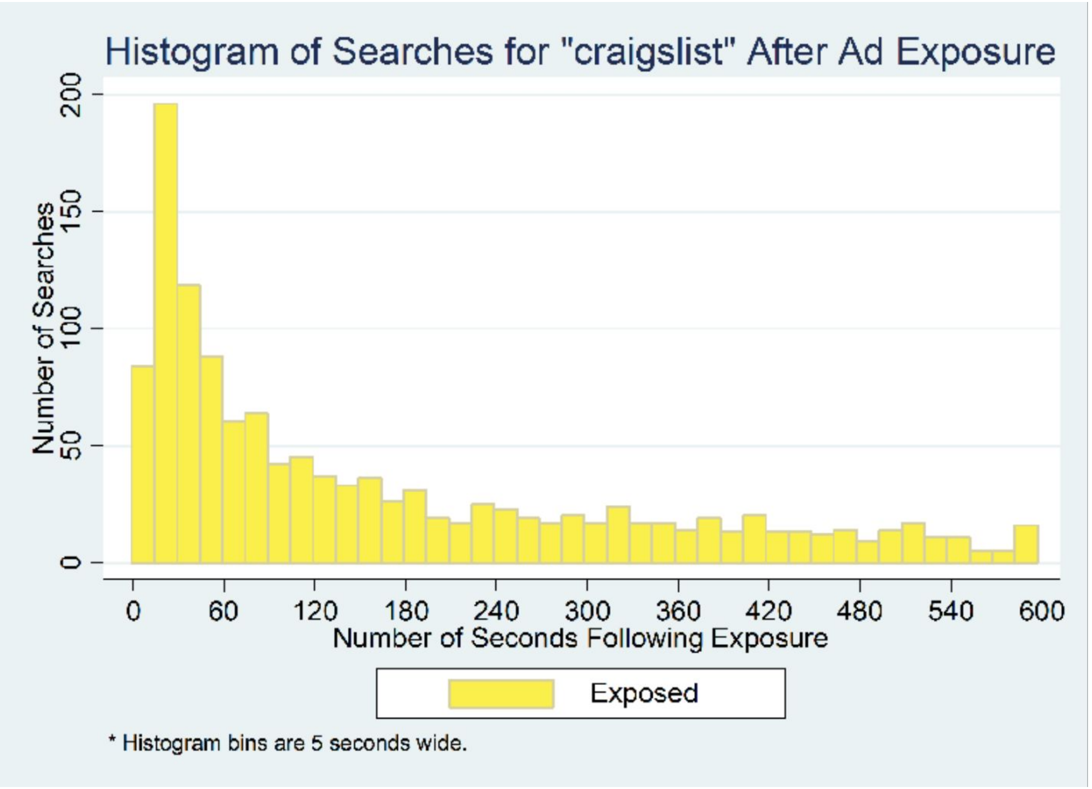
Lewis, Rao, & Reiley 2012:

- Online display ads
- Post-impression search spikes (baseline & lift)

Key Insight:

- Ad effects vary with time
- Modeling can improve statistical power

Examples of Ad Effectiveness



Lewis & Reiley 2013:

- Super Bowl 2012 Commercials
- Post-Commercial Search Spikes

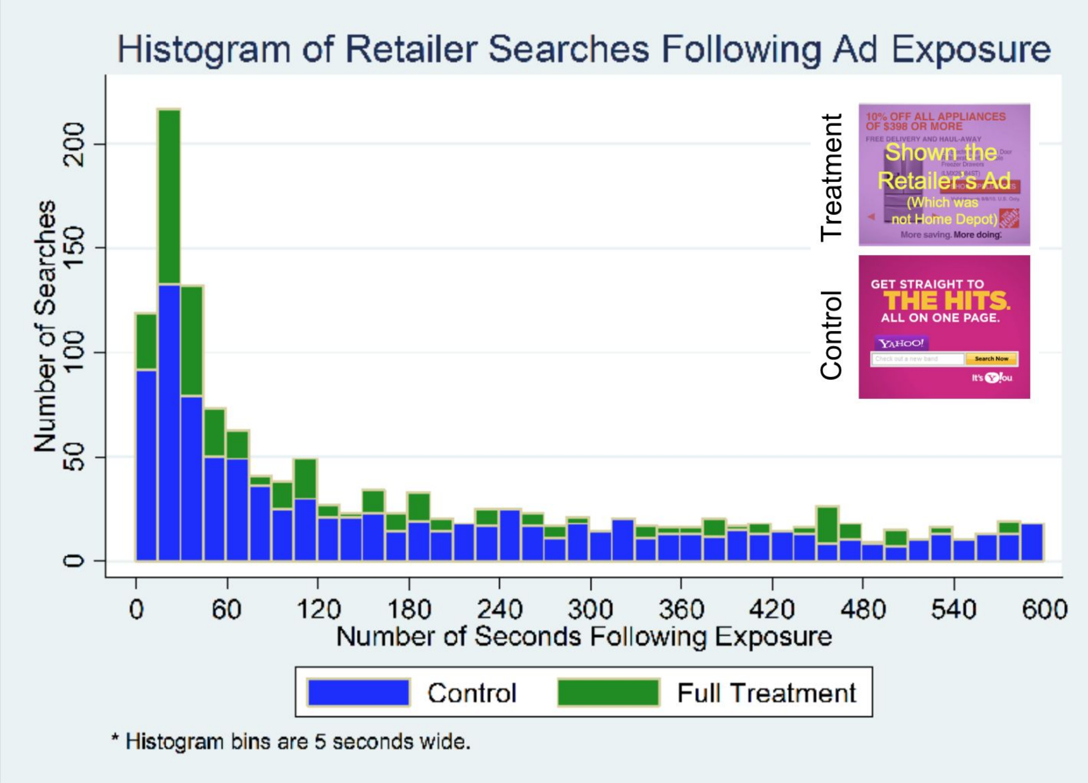
Lewis, Rao, & Reiley 2012:

- Online display ads
- Post-impression search spikes (baseline & lift)

Key Insight:

- Ad effects vary with time
- Modeling can improve statistical power

Examples of Ad Effectiveness



Lewis & Reiley 2013:

- Super Bowl 2012 Commercials
- Post-Commercial Search Spikes

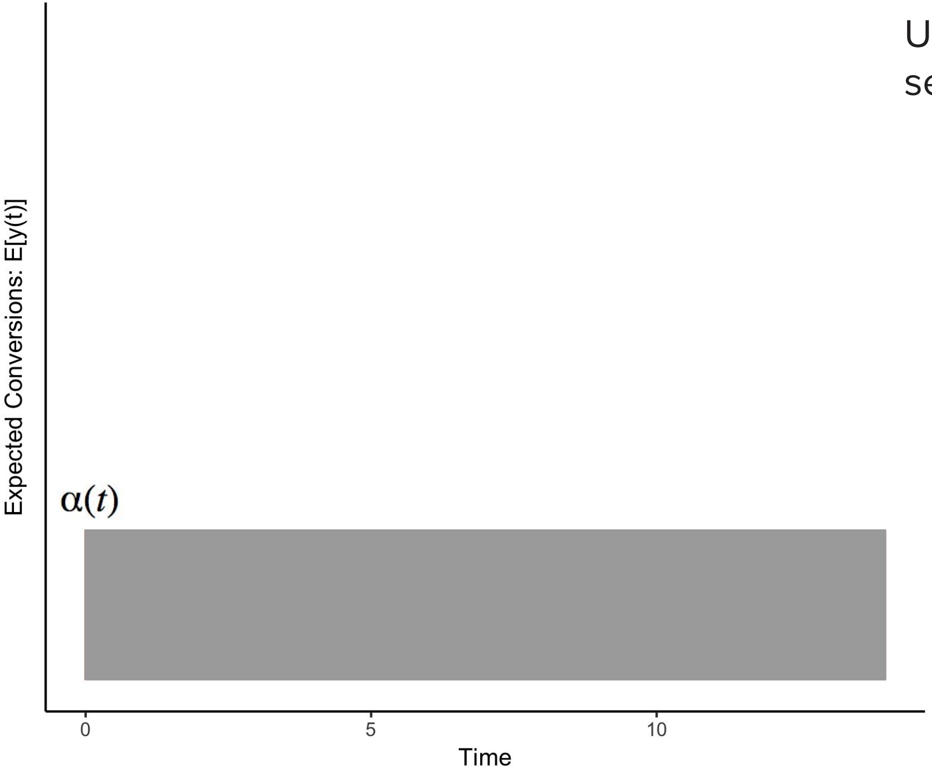
Lewis, Rao, & Reiley 2012:

- Online display ads
- Post-impression search spikes (baseline & lift)

Key Insight:

- Ad effects vary with time
- Modeling can improve statistical power

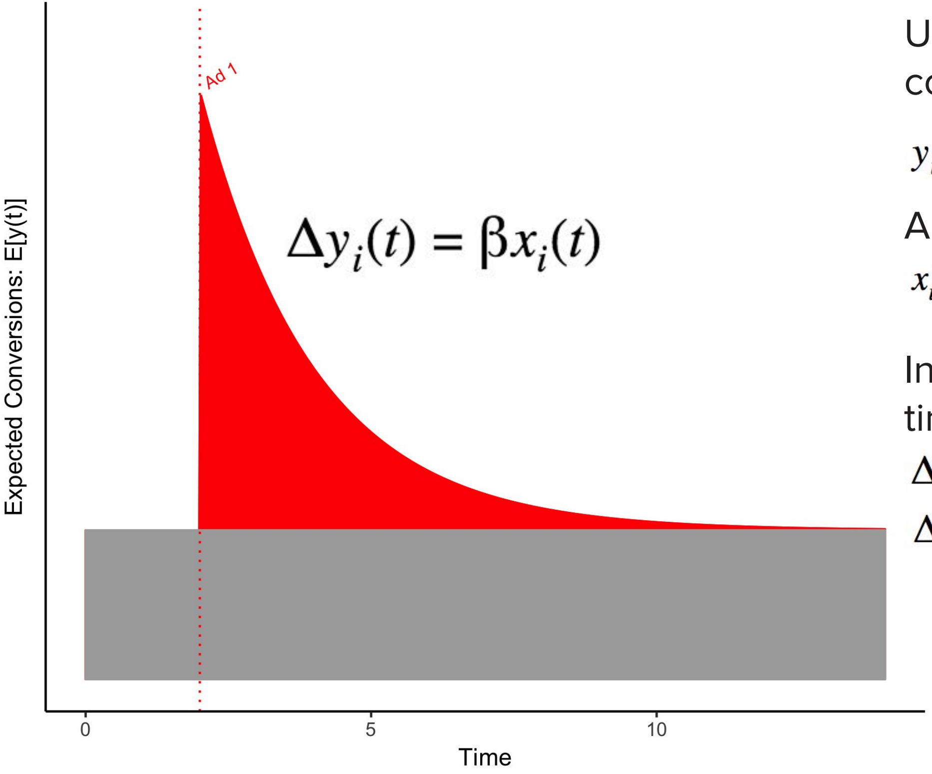
Baseline Conversion Rate



User *i* may convert without seeing an ad.

$$y_i(t) = \alpha(t) + \varepsilon_i(t)$$

Incrementality: 1 Ad



User *i* is more likely to convert after seeing an ad.

$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

Ad stock varies over time.

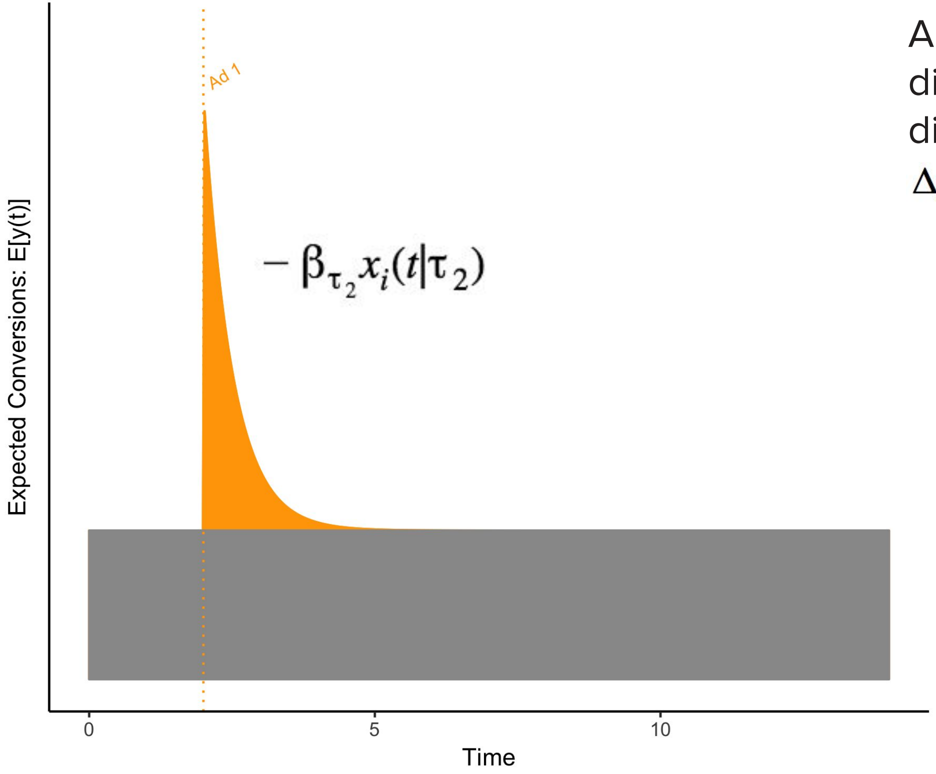
$$x_i(t) = f(t - t_{ad}|\tau) = \frac{1}{\tau} e^{-(t-t_{ad})/\tau}$$

Incrementality varies over time.

$$\Delta y_i(t) = E[y_i(t)|show\ ads] - E[y_i(t)|don't]$$

$$\Delta y_i(t) = \alpha(t) + \beta x_i(t) - \alpha(t) = \beta x_i(t)$$

Incrementality: 1 Ad, 2 Kernels

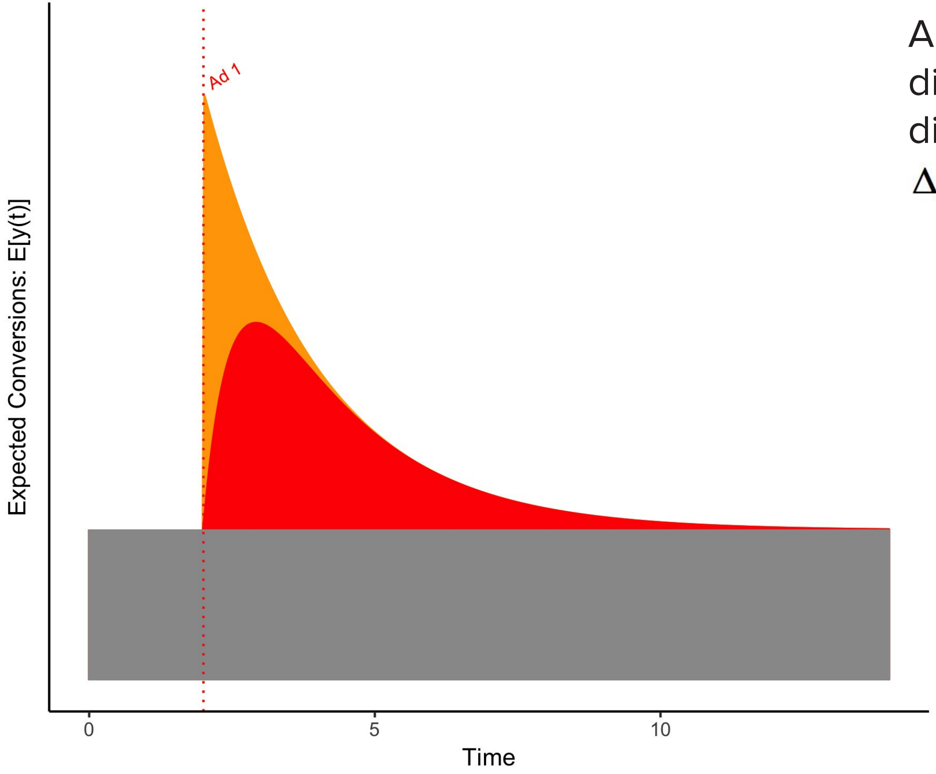


Ad stock can take on different shapes by using different kernels.

$$\Delta y_i(t) = \beta_{\tau_1} x_i(t|\tau_1) + \beta_{\tau_2} x_i(t|\tau_2)$$

NETFLIX

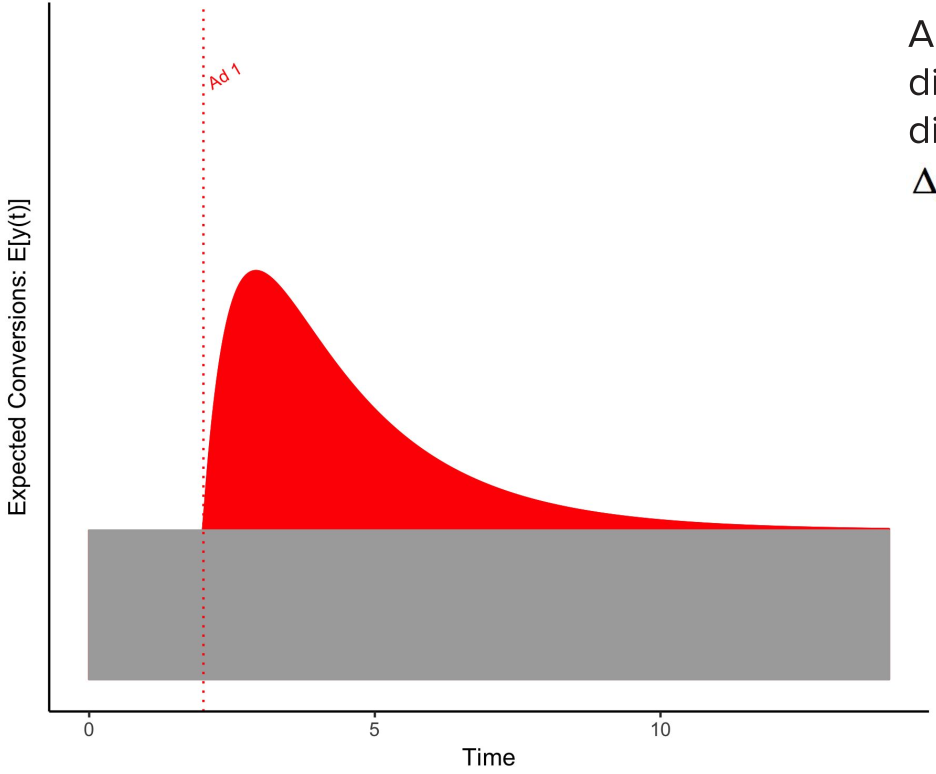
Incrementality: 1 Ad, 2 Kernels



Ad stock can take on different shapes by using different kernels.

$$\Delta y_i(t) = \beta_{\tau_1} x_i(t|\tau_1) + \beta_{\tau_2} x_i(t|\tau_2)$$

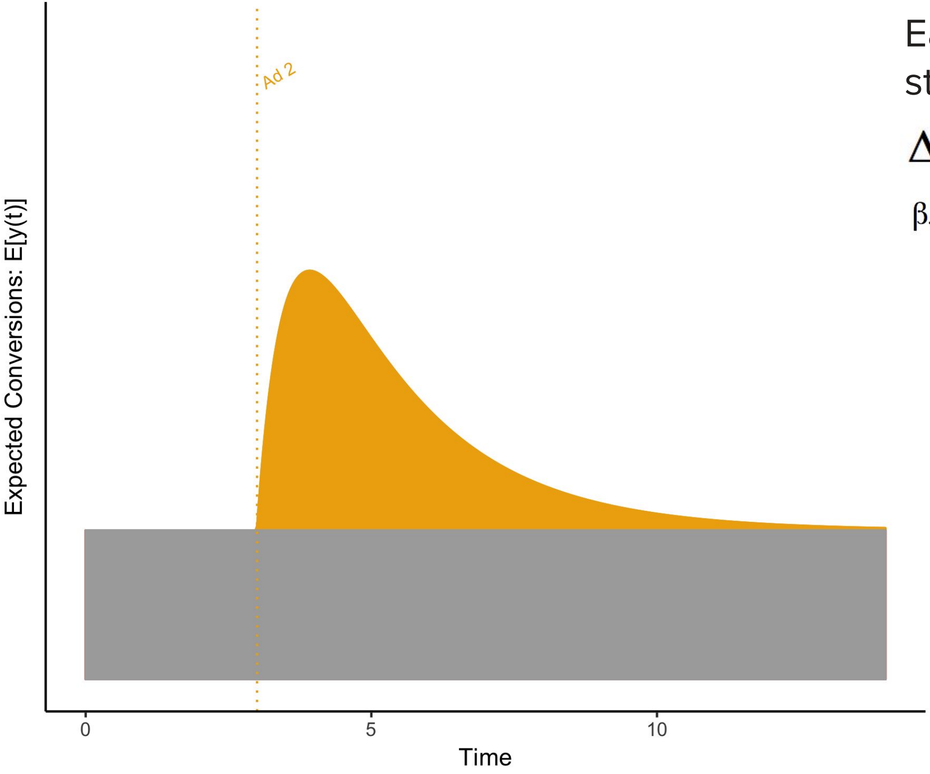
Incrementality: 1 Ad, 2 Kernels



Ad stock can take on different shapes by using different kernels.

$$\Delta y_i(t) = \beta_{\tau_1} x_i(t|\tau_1) + \beta_{\tau_2} x_i(t|\tau_2)$$

Incrementality: 2nd Ad

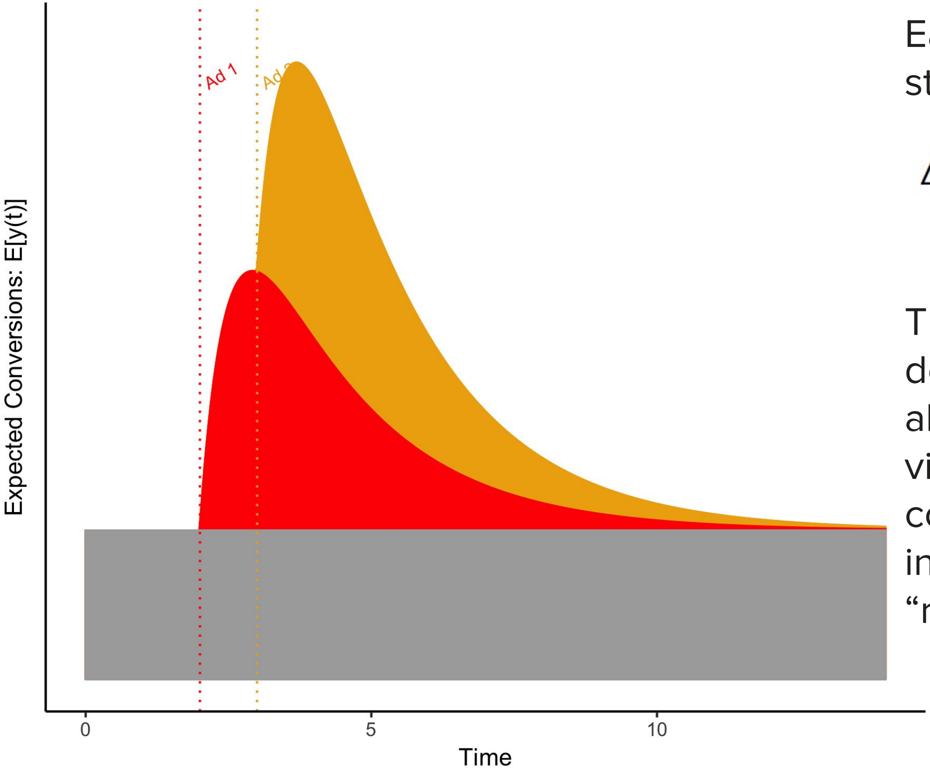


Each ad contributes to ad stock and incrementality.

$$\Delta y_{i,j=2}(t) = \beta x_{i,j=2}(t)$$

$$\beta x_{i,j=2}(t) = \beta_{\tau_1} x_{i,j=2}(t|\tau_1) + \beta_{\tau_1} x_{i,j=2}(t|\tau_2)$$

Incrementality: 2 Ads

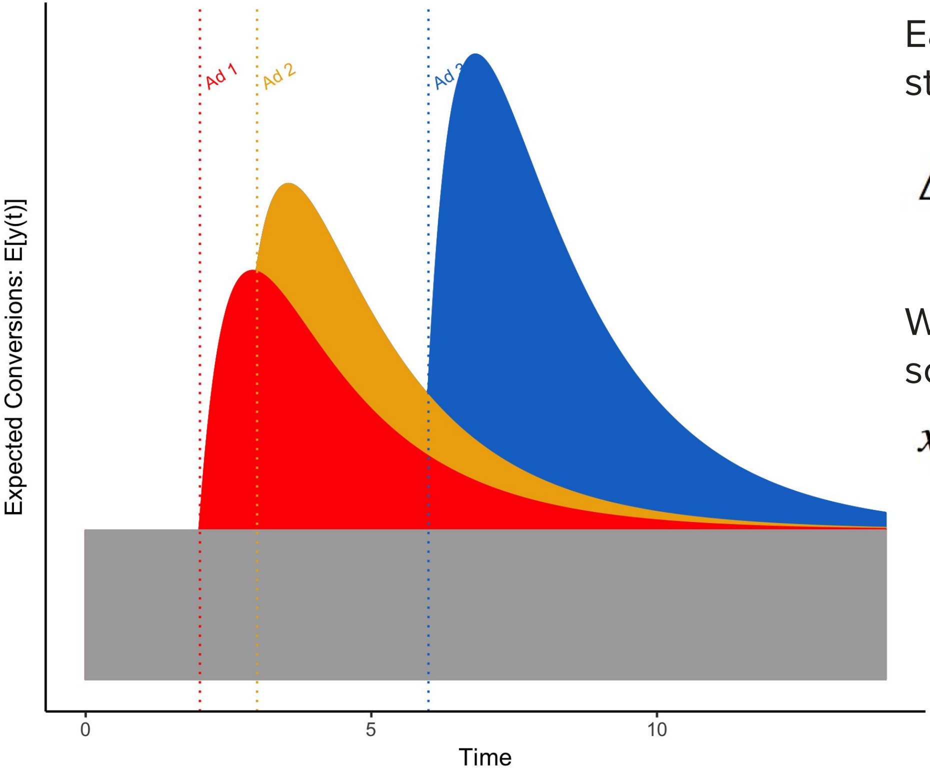


Each ad contributes to ad stock and incrementality.

$$\Delta y_i(t) = \sum_{j=1}^{2} \sum_{\tau} \beta_\tau x_{ij}(t|\tau)$$

The second ad's effect can depend on the presence or absence of the first ad (e.g., via weights w_j or more complex nonlinear interactions---see "retargeting features").

Incrementality: 3 Ads



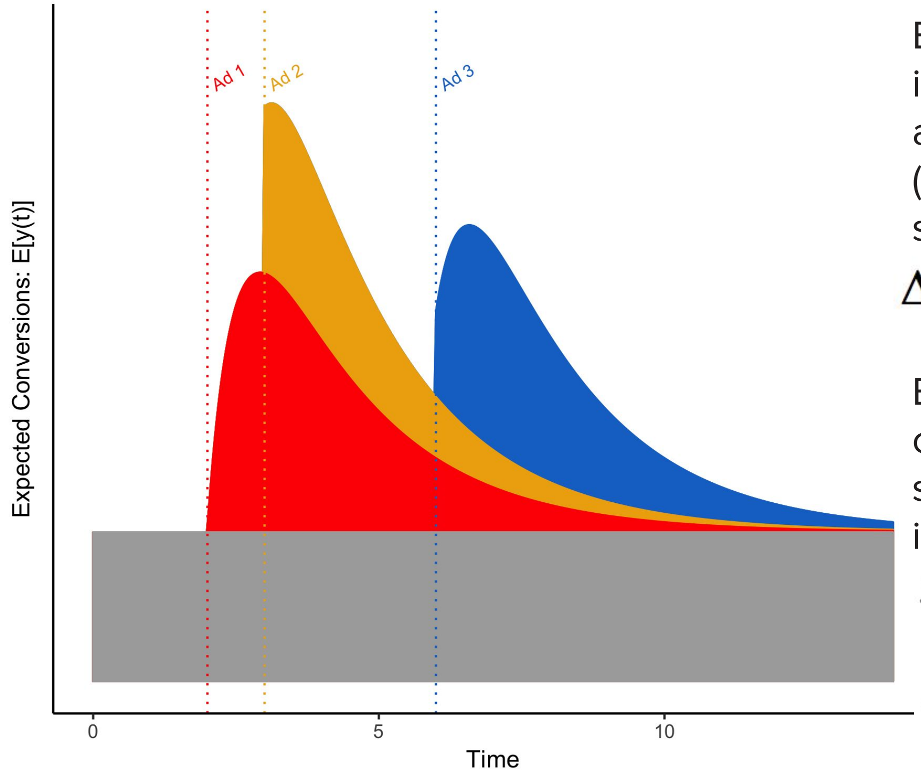
Each ad contributes to ad stock and incrementality.

$$\Delta y_i(t) = \sum_j^J \sum_\tau \beta_\tau x_{ij}(t|\tau)$$

We might want to give some ads more weight.

$$x_{ij}(t|\tau) = w_j \cdot f(t - t_j|\tau)$$

Incrementality: 3 Ads (Different Types)



Each ad contributes to incrementality differently, according to its attributes (e.g., placement, creative, size).

$$\Delta y_i(t) = \sum_j^J \sum_k^K \sum_\tau \beta_{k,\tau} x_{ijk}(t|\tau)$$

Each attribute, *k*, contributes to its own ad stock feature contributes to incrementality.

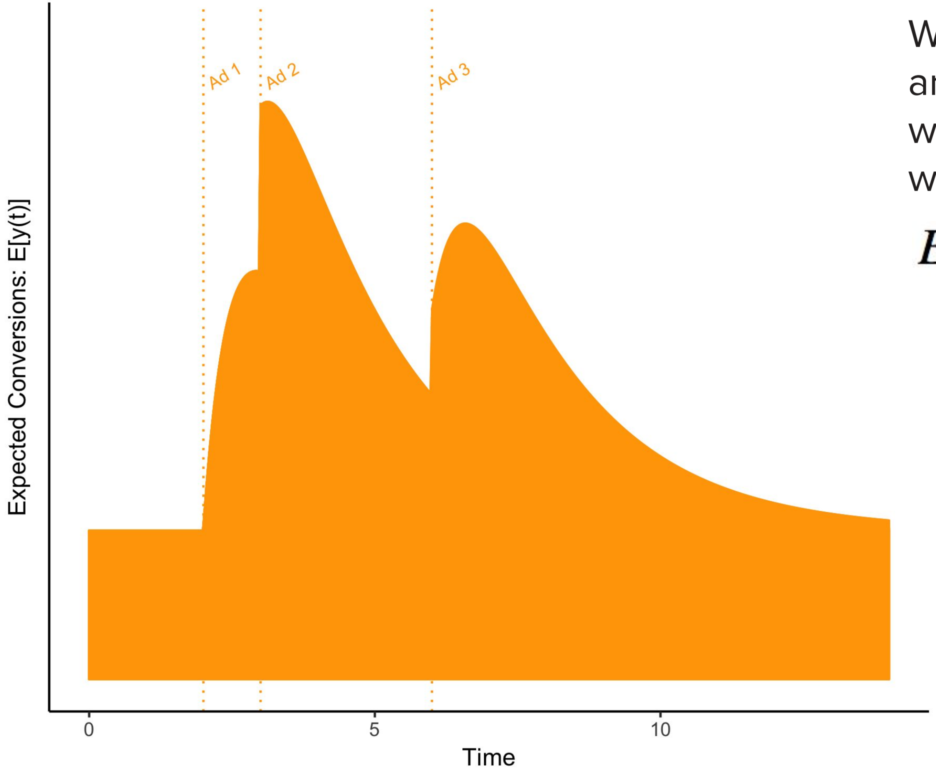
$$x_{ijk}(t|\tau) = w_{ijk} \cdot f(t - t_j|\tau)$$

$$\Delta y_i(t) = \sum_k \sum_\tau \beta_{k,\tau} x_{ik}(t|\tau)$$

$$x_{ik}(t|\tau) = \sum_j x_{ijk}(t|\tau)$$

NETFLIX

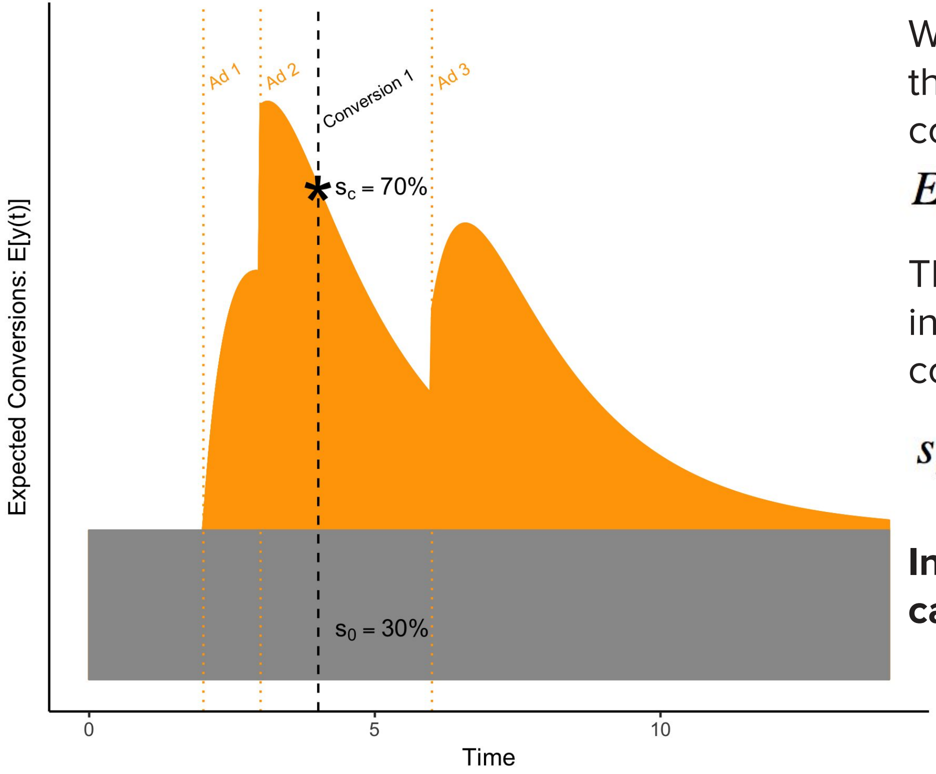
Observed Data: 3 Ads



We observe when the ads are shown and can model when, on average, users will convert.

$$E[y_i(t)] = \alpha(t) + \Delta y_i(t)$$

Observed Data: 3 Ads + 1 Conversion



We evaluate the model at the time t_c when user i converts.

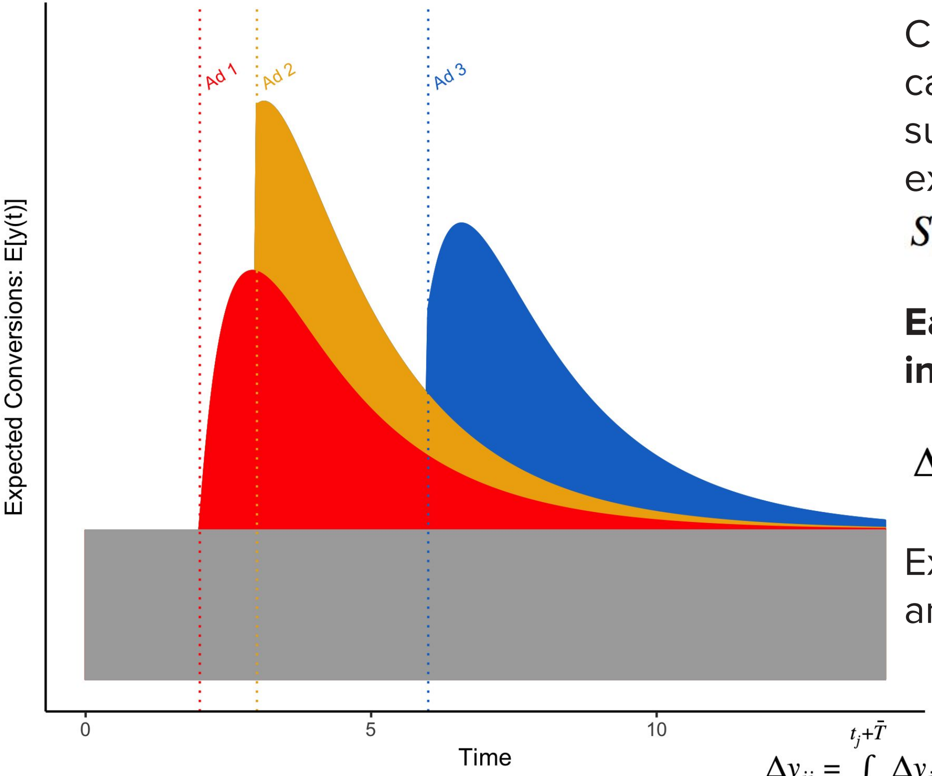
$$E[y_i(t_c)] = \alpha(t_c) + \Delta y_i(t_c)$$

The model identifies the incrementality share of the conversion.

$$s_{ic} = \frac{E[\Delta y_i(t_c)]}{E[y_i(t_c)]} = \frac{\beta x_i(t_c)}{\alpha(t_c)+\beta x_i(t_c)}$$

Incrementality share is causal attribution.

Incrementality Model ⇒ Bids



Campaign incrementality can be obtained by summing up each ad's expected incrementality.

$$S_{campaign} = \sum_{i \in users} \sum_{j \in ads} \Delta y_{ij}$$

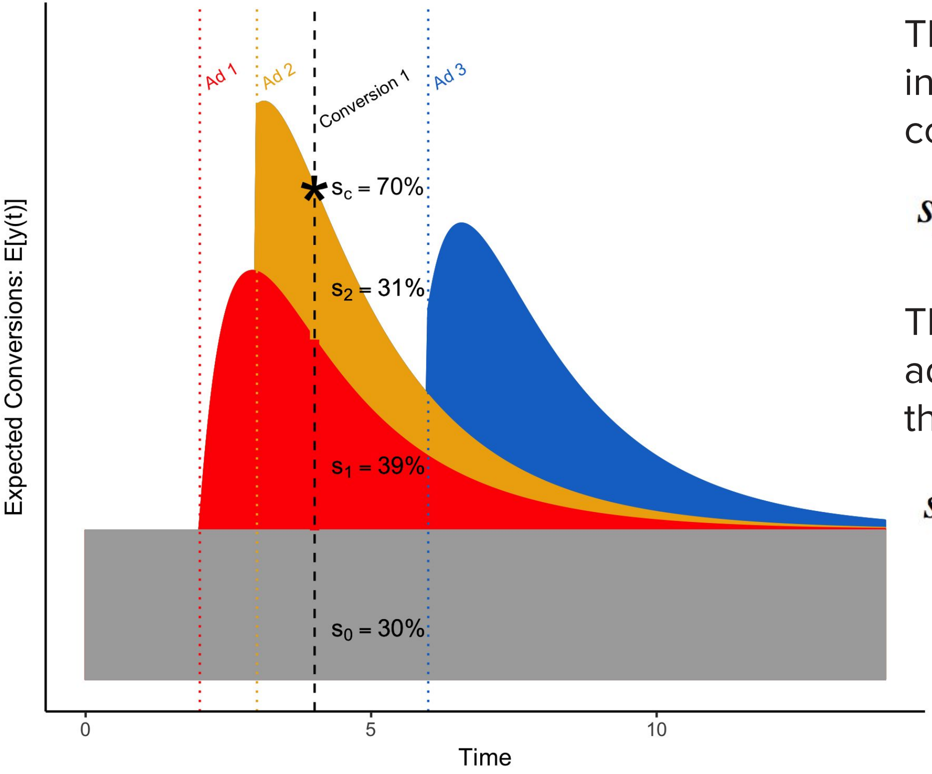
Each ad's expected incrementality is its area.

$$\Delta y_{ij} = \int_{t_j}^{t_j+\bar{T}} \Delta y_{ij}(t)dt = \sum_{k}^{K} \beta_k \cdot w_{ijk}$$

Expected incrementality is an input to bidding.

$$\Delta y_{ij} = \int_{t_j}^{t_j+\bar{T}} \Delta y_{ij}(t)dt = \int_{t_j}^{t_j+\bar{T}} \beta x_{ij}(t)dt = \beta w_{ij} \int_{0}^{\bar{T}} f(t - t_j|\theta)dt = \beta \cdot w_{ij}$$

Incrementality Share: 2 Ads



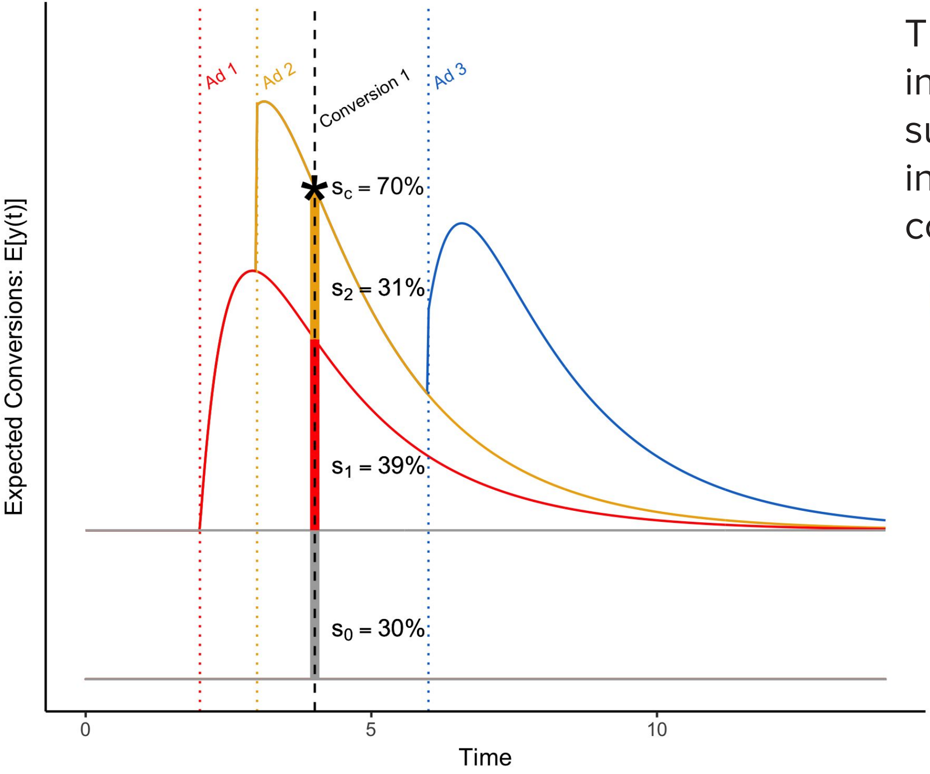
The model identifies the incrementality share of the conversion.

$$s_{ic} = \frac{E[\Delta y_i(t_c)]}{E[y_i(t_c)]} = \frac{\beta x_i(t_c)}{\alpha(t_c)+\beta x_i(t_c)}$$

The model identifies each ad's incrementality share of the conversion.

$$s_{ijc} = \frac{E[\Delta y_{ij}(t_c)]}{E[y_i(t_c)]} = \frac{\beta x_{ij}(t_c)}{\alpha(t_c)+\beta x_i(t_c)}$$

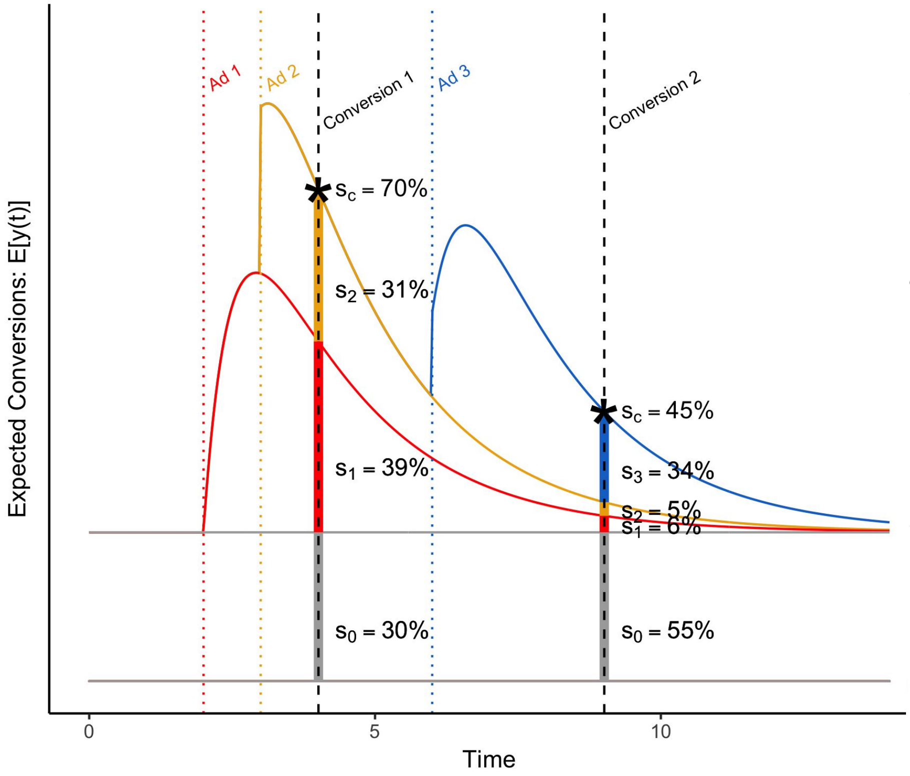
Incrementality Share: 2 Ads



The conversion's incrementality share is the sum of each ad's incrementality share of the conversion.

$$s_{ic} = \sum_{j \in ads} s_{ijc}$$

Incrementality Share: 2 Conversions



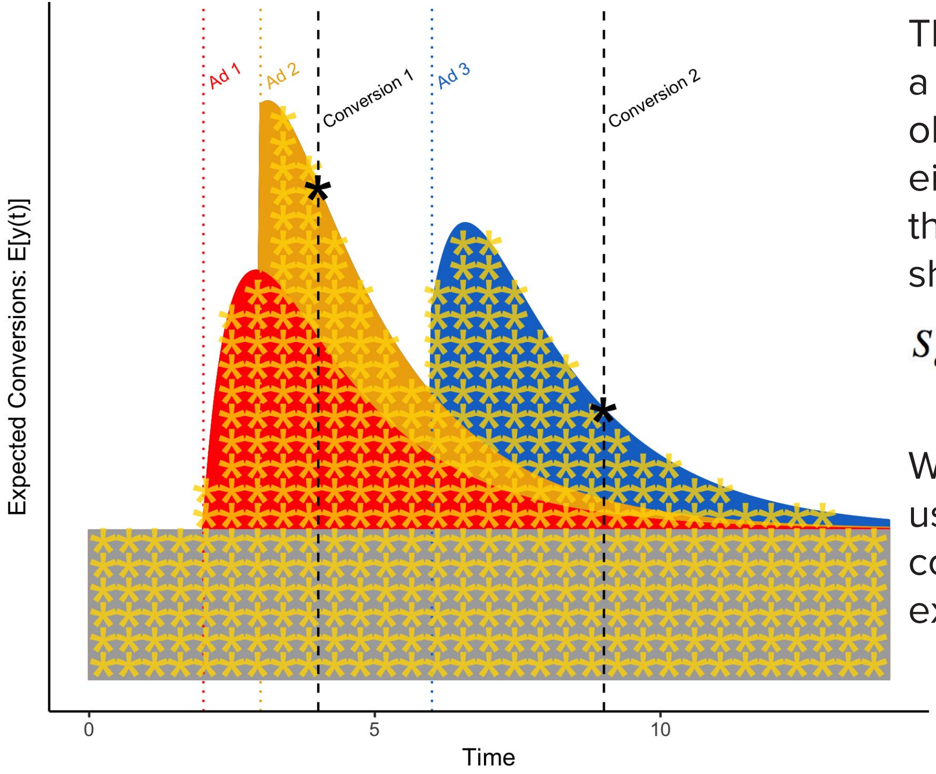
Each ad's incrementality share is the sum of its contribution to each conversion. **This is each impression's total causal attribution credit.**

$$s_{ij} = \sum_{c \in conversions} s_{ijc}$$

Each user's incrementality share is the sum over its conversions' or ads' incrementality shares.

$$s_i = \sum_c s_{ic} = \sum_{j \in ads} \sum_c s_{ijc} = \sum_{j \in ads} s_{ij}$$

Conversions ⇒ Incrementality Model

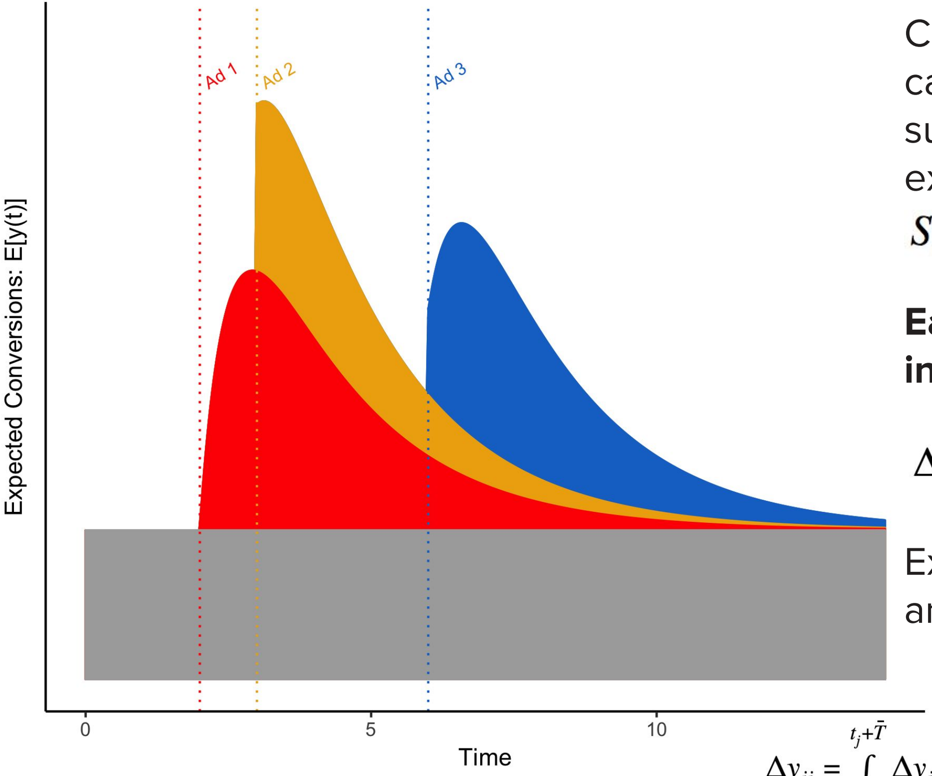


The total incrementality of a campaign can be obtained by summing up either the conversions' or the ads' incrementality shares.

$$S_{campaign} = \sum_{i \in users} \sum_{c} s_{ic} = \sum_{i \in users} \sum_{j \in ads} s_{ij}$$

We estimate the model using many users' conversions and ad exposures.

Incrementality Model ⇒ Bids



Campaign incrementality can be obtained by summing up each ad's expected incrementality.

$$S_{campaign} = \sum_{i \in users} \sum_{j \in ads} \Delta y_{ij}$$

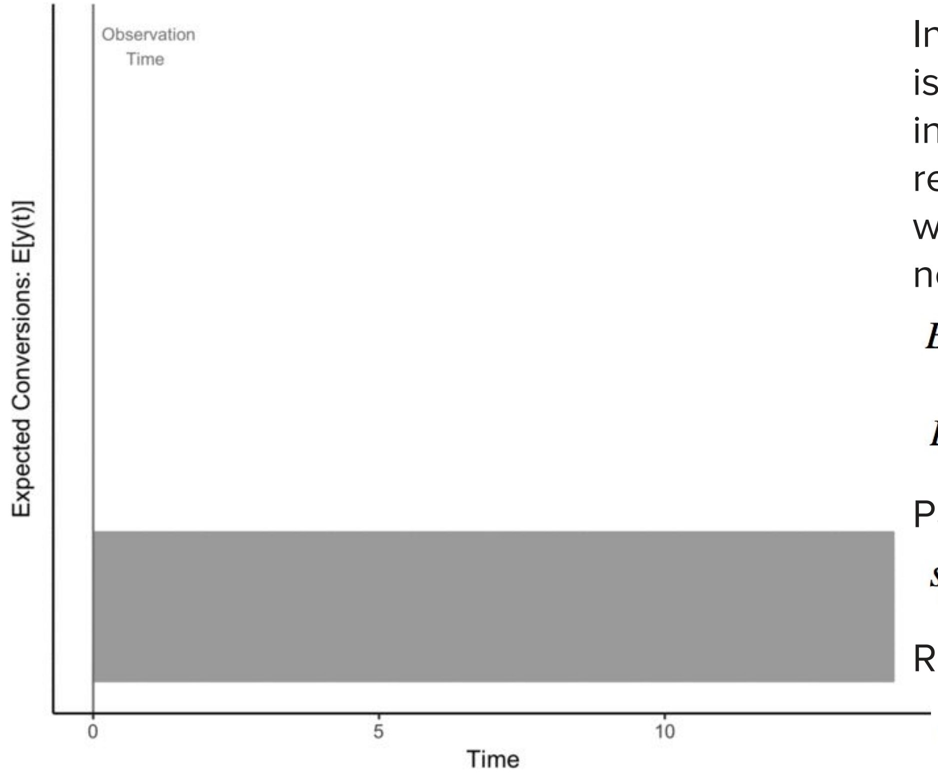
Each ad's expected incrementality is its area.

$$\Delta y_{ij} = \int_{t_j}^{t_j+\bar{T}} \Delta y_{ij}(t)dt = \sum_{k}^{K} \beta_k \cdot w_{ijk}$$

Expected incrementality is an input to bidding.

$$\Delta y_{ij} = \int_{t_j}^{t_j+\bar{T}} \Delta y_{ij}(t)dt = \int_{t_j}^{t_j+\bar{T}} \beta x_{ij}(t)dt = \beta w_{ij} \int_{0}^{\bar{T}} f(t - t_j|\theta)dt = \beta \cdot w_{ij}$$

Incrementality Through Time



Incrementality through time is the sum of observed incrementality shares and residual incrementality whose conversions have not yet been observed.

$$E[\Delta y|t] = \sum_{i \in users} \sum_{j \in ads} E[\Delta y_{ij}|t]$$

$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

Partial Incrementality Share

$$s_{ij}(t) = \sum_{c} 1(t_j < t_c < t) \cdot s_{ijc}$$

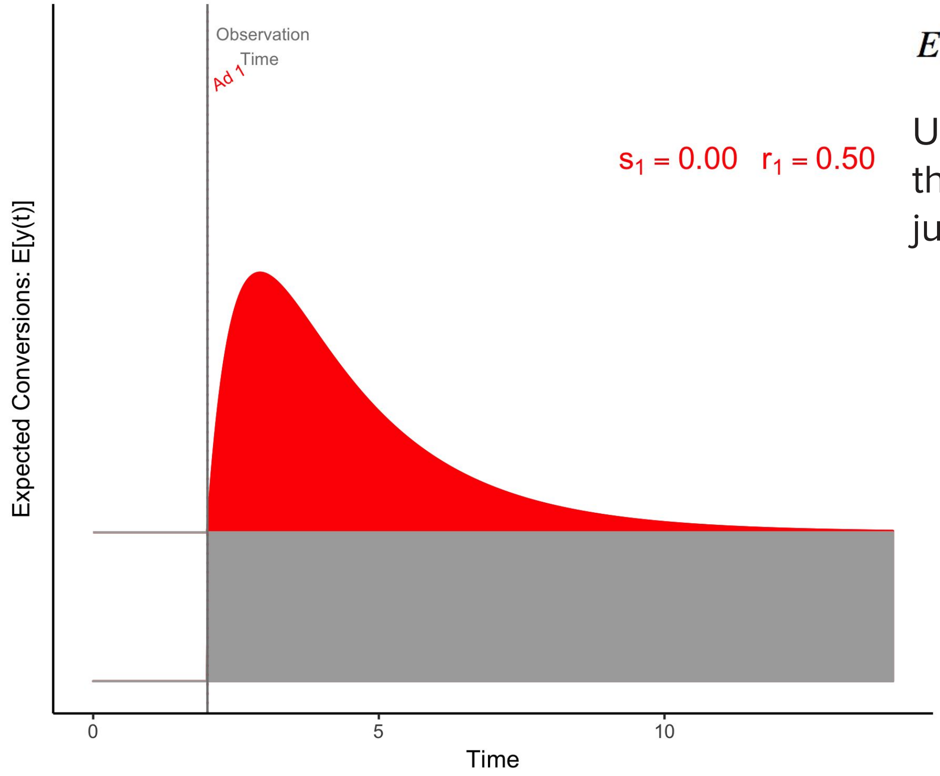
Residual Incrementality

$$r_{ij}(t) = \beta \left(1 - F(t - t_j|\theta)\right)$$

$$r_{ij}(t) = \int_{t \geq t_j}^{\bar{T}} \Delta y_{ij}(t)dt = \beta \int_{t \geq t_j}^{\bar{T}} f(t - t_j|\theta)dt = \beta \left(1 - F(t - t_j|\theta)\right) \qquad r_{ij}(t) = \sum_{k}^{K} \sum_{\tau} \beta_{k,\tau} A_{jk} w_j \left(1 - F(t - t_j|\tau)\right)$$

NETFLIX

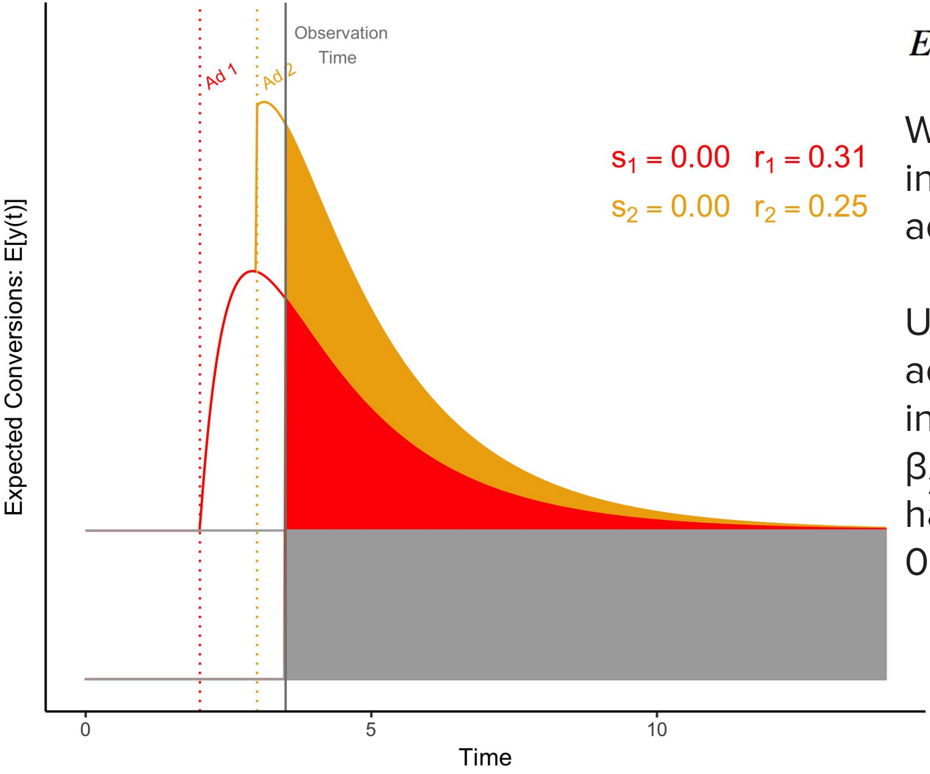
Incrementality Through Time



$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

Upon winning our first ad, the residual incrementality jumps to β_1=0.5.

Incrementality Through Time

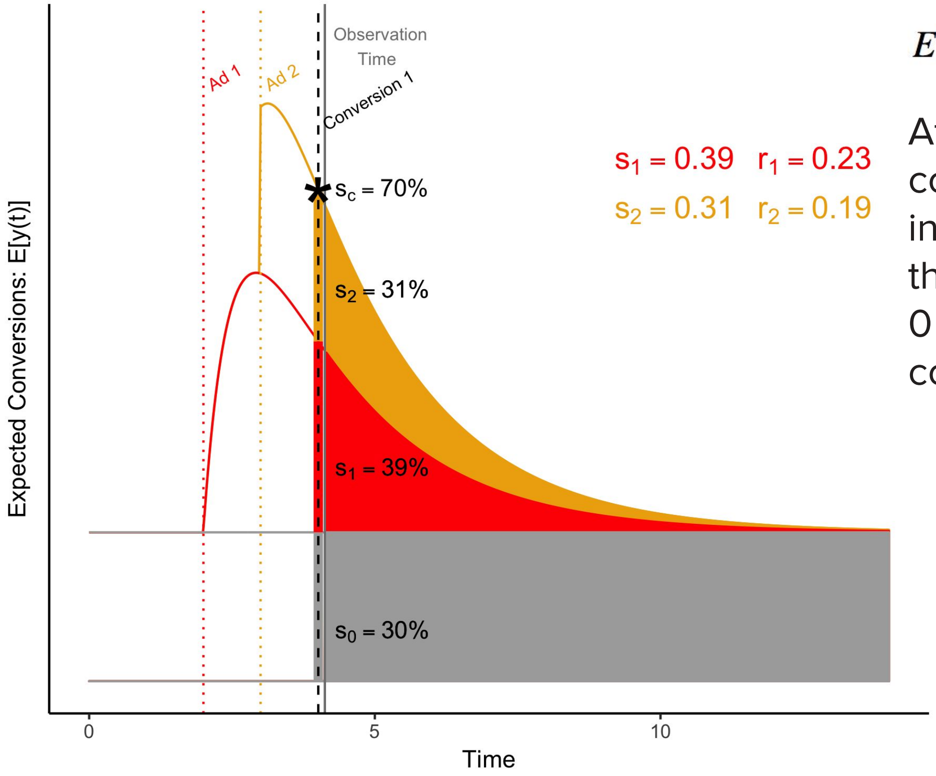


$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

With time, the residual incrementality of our first ad declines to $0.6\beta_1$.

Upon winning our second ad, its residual incrementality increased to β_2=0.3 but at this point time has already decreased to $0.8\beta_2$.

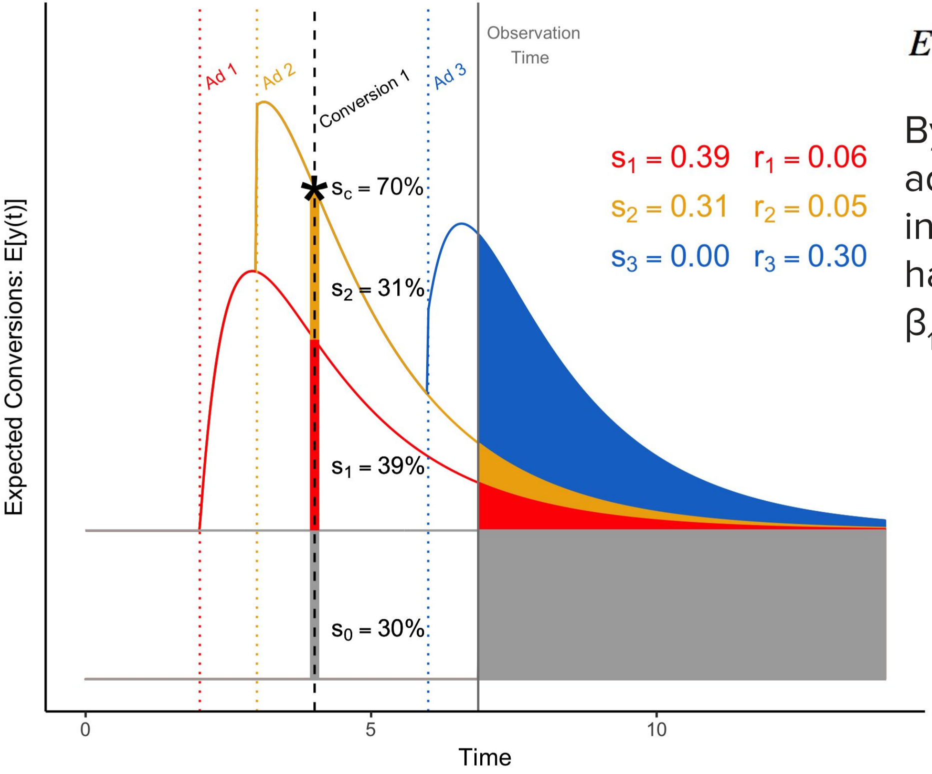
Incrementality Through Time



$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

At t=4, we observe our first conversion, and the incrementality shares of the two ads increase from 0 to 0.39 and 0.31 conversions.

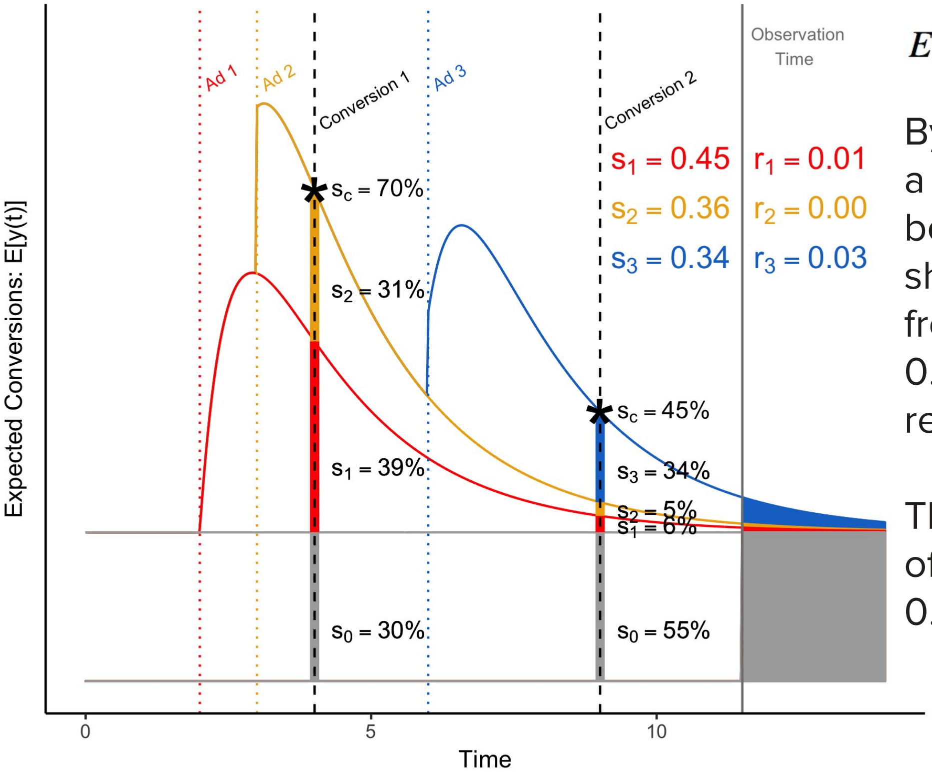
Incrementality Through Time



$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

By t=7, we have won a third ad, and the residual incrementality of all three has further declined to 0.1 β_1, 0.1β_2, 0.7β_3.

Incrementality Through Time



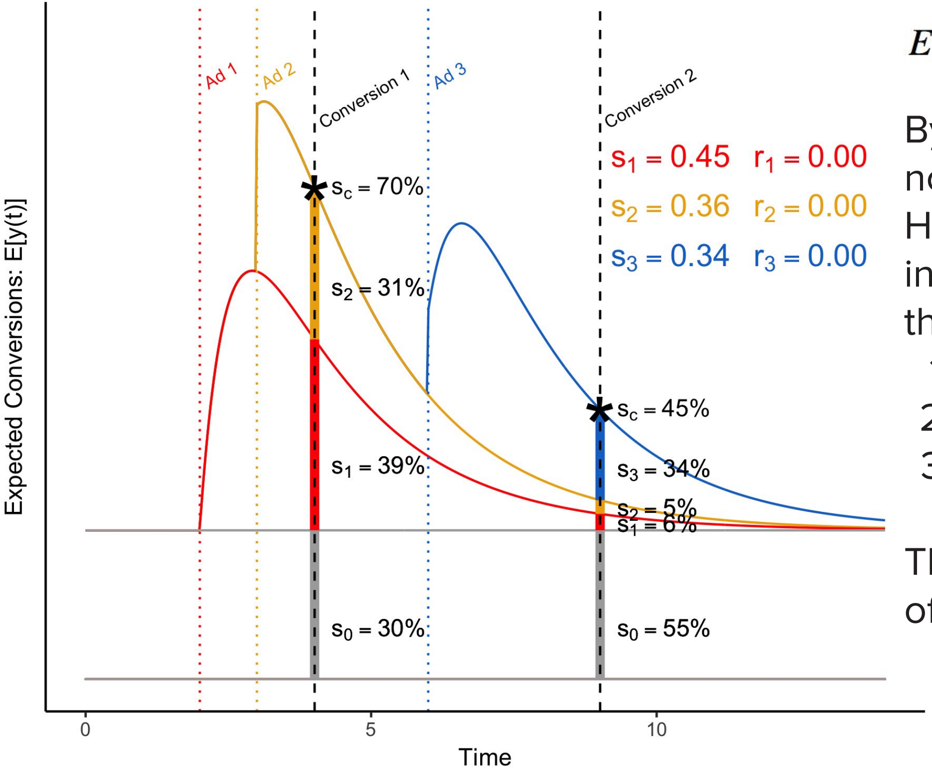
$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

By t=11, we have observed a second conversion, boosting the incrementality shares of the three ads from 0.39, 0.31, and 0 to 0.45, 0.36, and 0.34, respectively.

The residual incrementality of all three has declined to $0.01\beta_1$, $0.01\beta_2$, $0.1\beta_3$.

NETFLIX

Incrementality Through Time



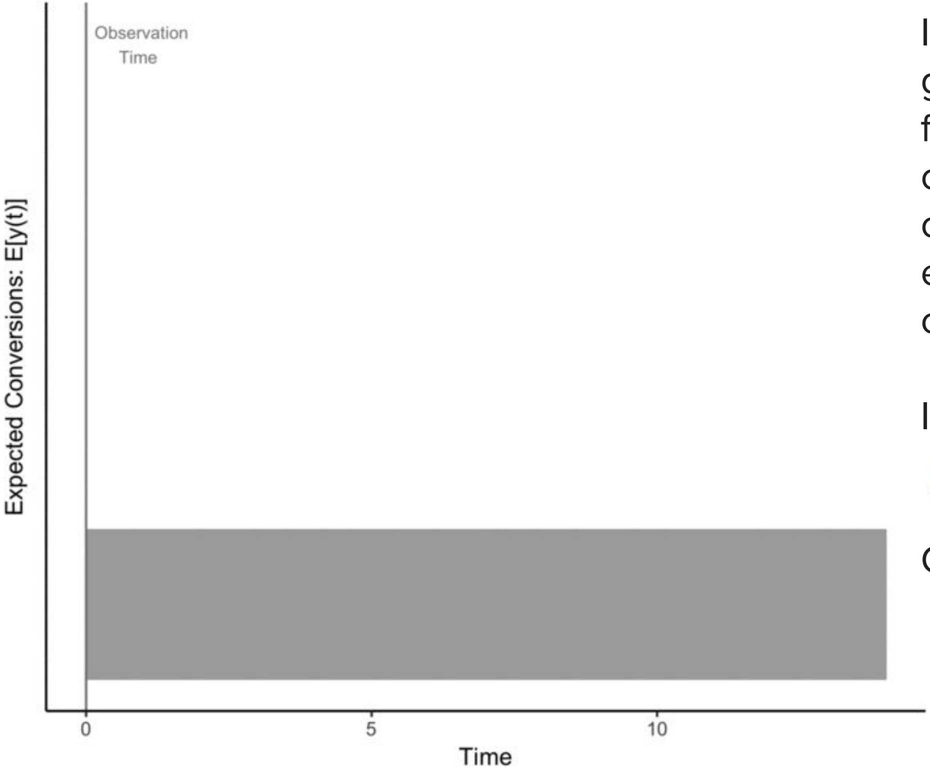
$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

By t>15, we have observed no additional conversions. Hence, the finalized incrementality shares of the three ads are:

1. 0.39 + 0.06 = 0.45
2. 0.31 + 0.05 = 0.36
3. 0 + 0.34 = 0.34

The residual incrementality of all three ads is now 0.

“Black Box” Incrementality Model Training



Incrementality through time gives us attribution scores for both impressions and conversions. These scores can inform a “black box” of expected and realized campaign performance.

Impression Scores:

$$E[\Delta y_{ij}|t] \equiv s_{ij}(t) + r_{ij}(t)$$

Conversion Scores:

$$E[s_{ic}|t] \equiv s_{ic} + r_{ic}(t)$$

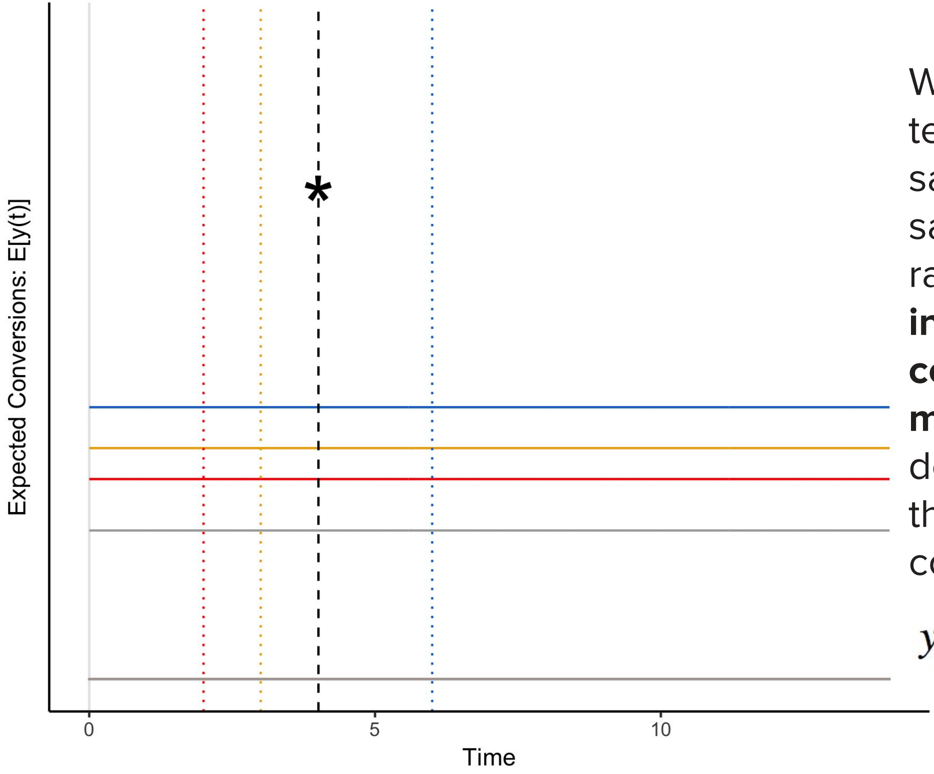
$$r_{ic}(t) = \sum_{j \in ads} s_{ijc} \cdot \frac{r_{ij}(t)}{\Delta y_{ij} - r_{ij}(t)}$$

Estimating an Incrementality Model in Continuous Time

Advanced Incrementality
for Industry

NETFLIX

Continuous-Time Panel Data

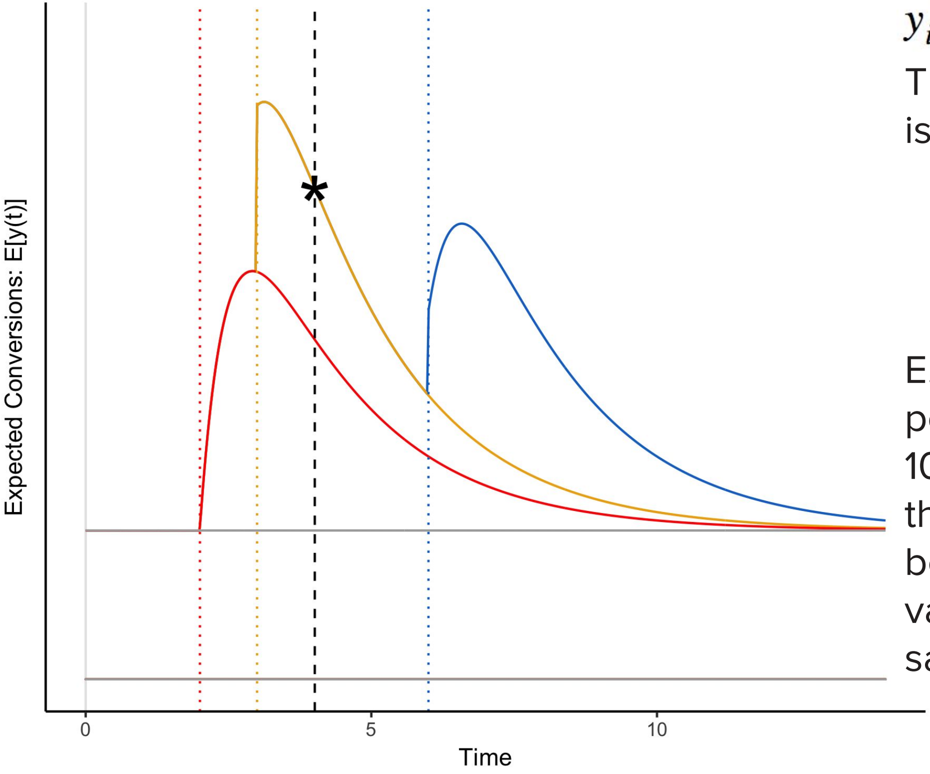


$$y_{it} = x_{it}\beta + \varepsilon_{it}$$

We need to increase the temporal precision of our samples. However, sampling by millisecond rather than by hour **increases computational costs by a factor of 3.6 million**. Fortunately, we can downsample in time. Given this, we go straight to continuous time sampling:

$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

Continuous-Time Panel Data



$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

The cost of downsampling is in terms of variance:

$$Var(\hat{\beta}) \propto 1 + \frac{1}{C}$$

$$C = \frac{\#\{Y=0\}}{\#\{Y=1\}} >> 1$$

E.g., if we have 1,000 positives (Y=1) and sample 10,000 negatives (Y=0), then C=10. Hence, we will be within 10% of the variance of using an infinite sample of negatives.

Continuous-Time Panel Data

$$\mu = \frac{\#\{Y=1\}}{N \cdot T} = \frac{N^{+}}{(N^{-}+N^{+}) \cdot T} = E[Y]\ per\ unit\ of\ time$$

$$\hat{\beta} = (\Sigma_{-}X'X + \Sigma_{+}X'X)^{-1}(\Sigma_{+}X'Y + \Sigma_{-}X'0)$$

$$\hat{\beta} = (\Sigma_{-}X'X + \Sigma_{+}X'X)^{-1}(\Sigma_{+}X'Y)$$

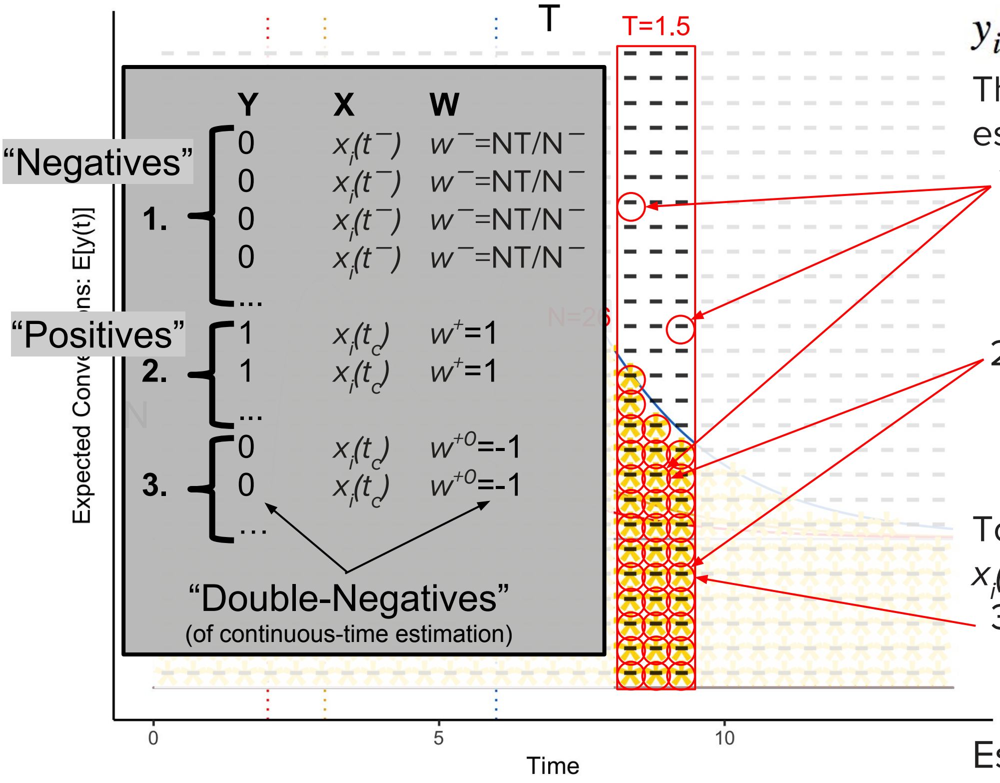
"Double-Negative"

$$\boxed{\hat{\beta} = \left(\frac{NT}{N^{-}}\Sigma_{-}X'X + \Sigma_{+}X'X + (-1)\Sigma_{+}X'X\right)^{-1}(\Sigma_{+}X'Y + (-1)\Sigma_{+}X'0)}$$

$$\hat{\beta} = \left(\frac{NT}{N^{-}}\Sigma_{-}X'X + \Sigma_{+}X'X + (-1)\Sigma_{+}X'X\right)^{-1}(\Sigma_{+}X'Y)$$

$$\boxed{\hat{\beta} = \left(\frac{NT}{N^{-}}\Sigma_{-}X'X\right)^{-1}(\Sigma_{+}X'Y) \approx E[X'X]^{-1}E[X'Y] \approx \beta}$$

Continuous-Time Panel Data



$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

The continuous-time estimator is simple:

1. $Y_i(t^-)$=0 *"Negatives"*: Uniformly sample N^- observations over *i*, *t* to obtain $x_i(t^-)$. $\boxed{w^- = NT/N^-}$
2. $Y_i(t_c)$=1 *"Positives"*: Sample all, obtain $x_i(t_c)$. $\boxed{w^+ = 1.}$

Infeasible!

Measure=39

34 obs.

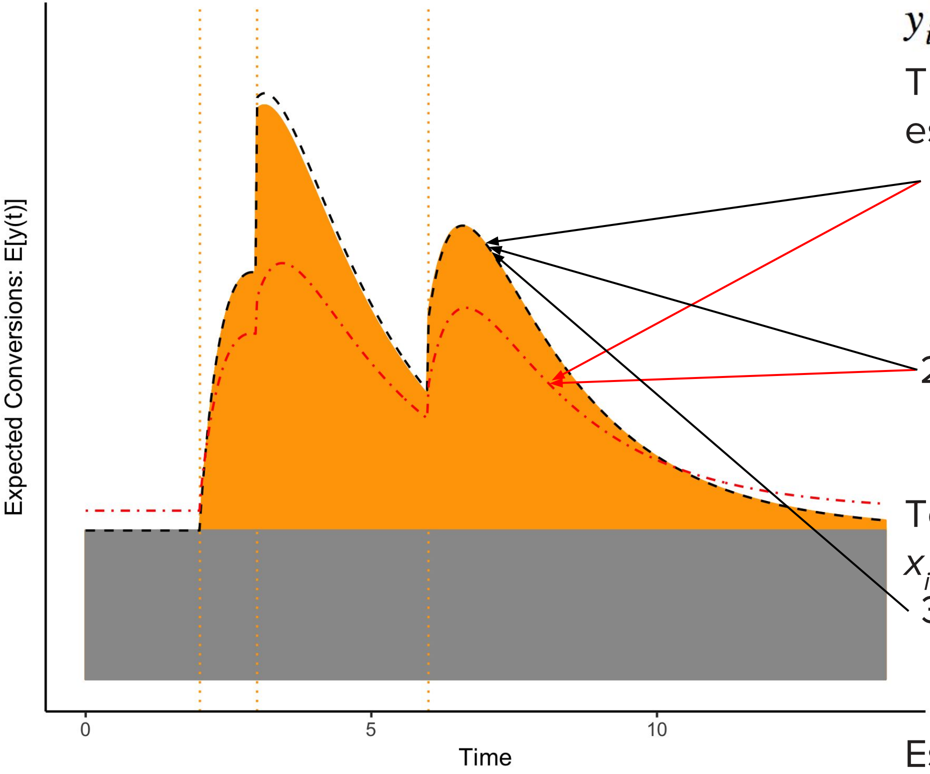
To offset double-sampling $x_i(t_c)$ in $Y_i(t)$=0 and $Y_i(t_c)$=1:

3. $Y_i(t_c)$=0: obtain $x_i(t_c)$. $\boxed{w^{+0} = -1.}$

-34 obs.

Estimate on 1, 2, and 3.

Continuous-Time Panel Data



$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

The continuous-time estimator is simple:

1. $Y_i(t)$=0: Sample N^- observations uniformly over *i*, *t* to obtain $x_i(t)$. $\boxed{w^- = NT/N^-}$. **"Negatives"**
2. $Y_i(t_c)$=1: Sample all, obtain $x_i(t_c)$. $\boxed{w^+ = 1}$. **"Positives"**

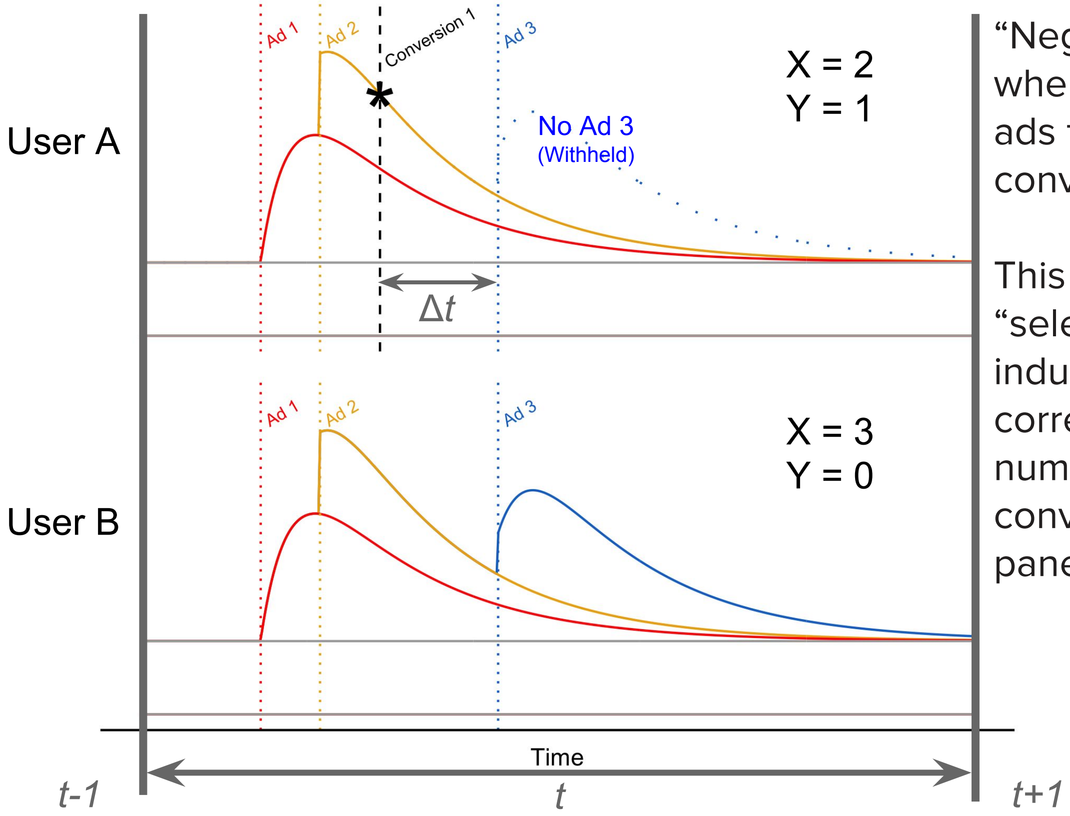
To offset double-sampling $x_i(t_c)$ in $Y_i(t)$=0 and $Y_i(t_c)$=1:

3. $Y_i(t_c)$=0: obtain $x_i(t_c)$. $\boxed{w^{+0} = -1}$. **"Double-Negatives"**

Estimate on 1, 2, and 3.

NETFLIX

The Worst Endogeneity: Negative Targeting



"Negative Targeting" is when the server withholds ads from users who have converted recently.

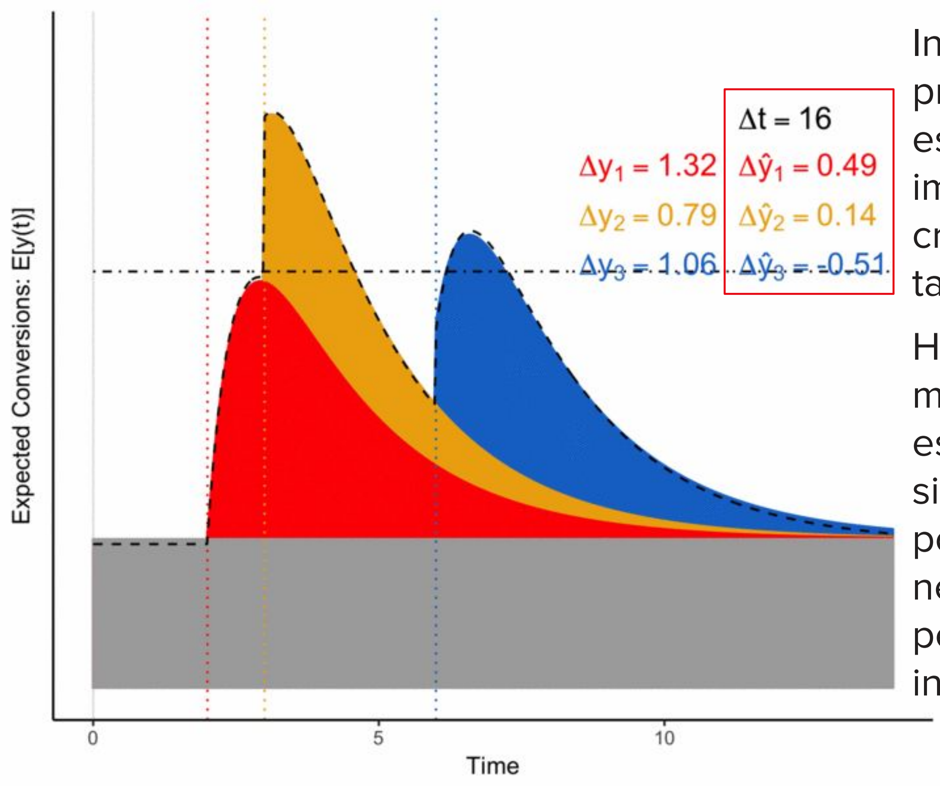
This contemporaneous "selection on the outcome" induces a negative correlation between the number of ads and conversions in a simple panel regression:

$$Y_{it} = \beta_0 + X_{it}\beta_1$$

$$\boxed{\beta_1 = -1 < 0\ ???}$$

Biased model!

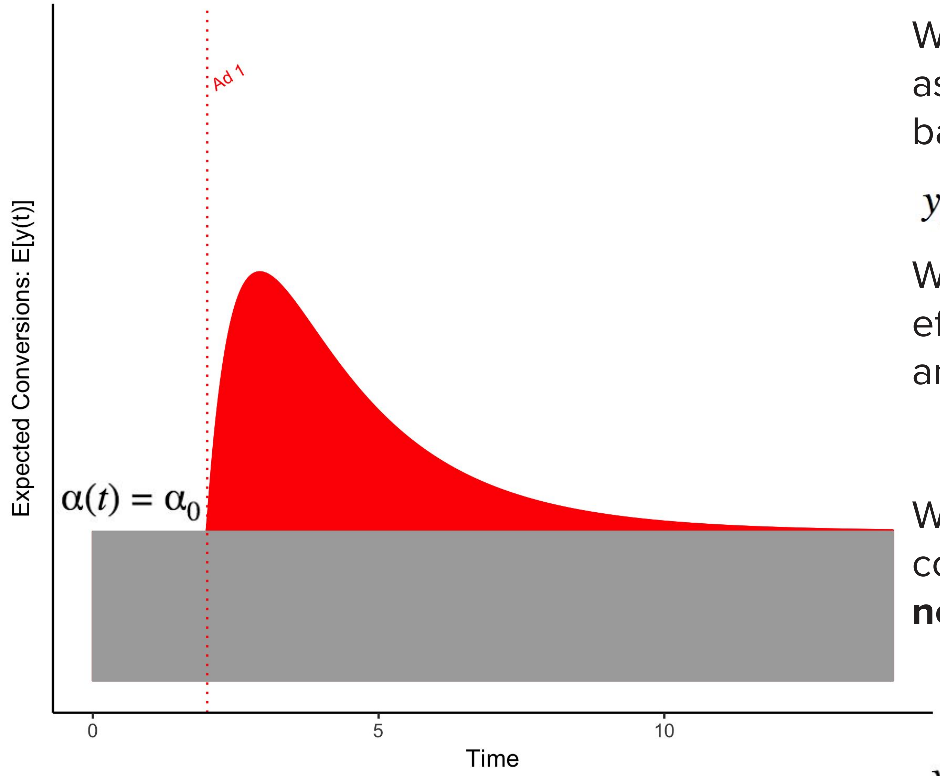
The Worst Endogeneity: Negative Targeting



Increasing the temporal precision of our panel estimates reduces the impact of the endogeneity created by negative targeting.

Here, we see that our model's incrementality estimates go from being significantly biased, some positively and others negatively, to being perfectly calibrated with increased precision.

The Worst Endogeneity: Simultaneous Treatment



We revisit our first assumption of a constant baseline.

$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

We generalized the causal effects to the exact time and attributes of the ad.

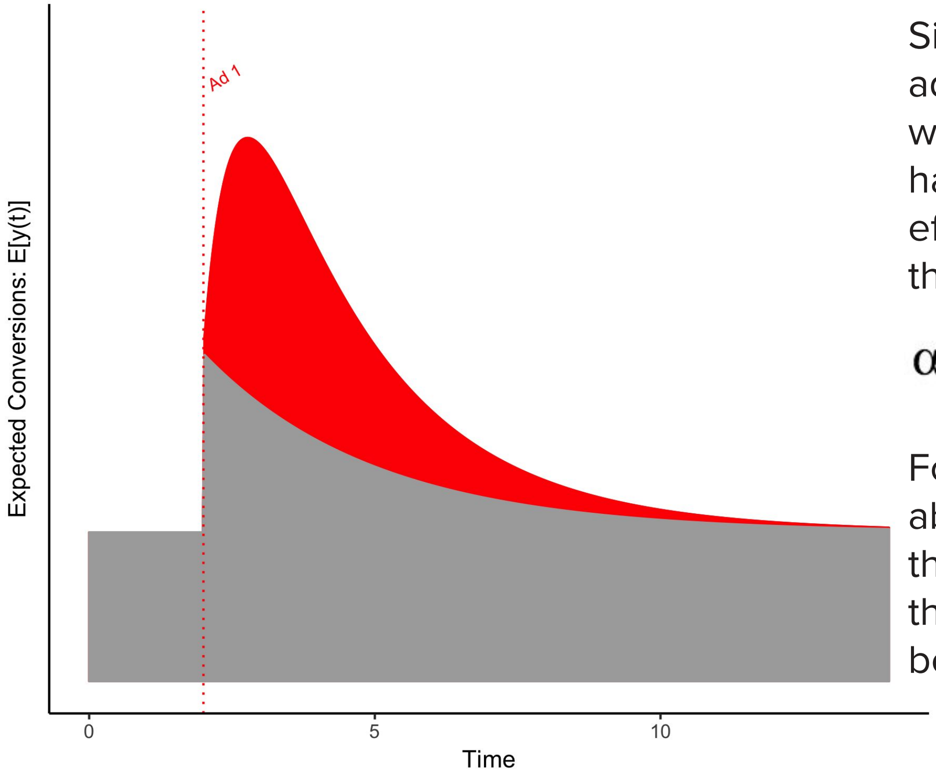
$$\Delta y_i(t) = \beta x_i(t)$$

We now consider the consequences of a **non-constant baseline**.

$$\alpha_0 \Longrightarrow \alpha_i(t)$$

$$y_i(t) = \alpha_i(t) + \Delta y_i(t) + \varepsilon_i(t)$$

The Worst Endogeneity: Simultaneous Treatment

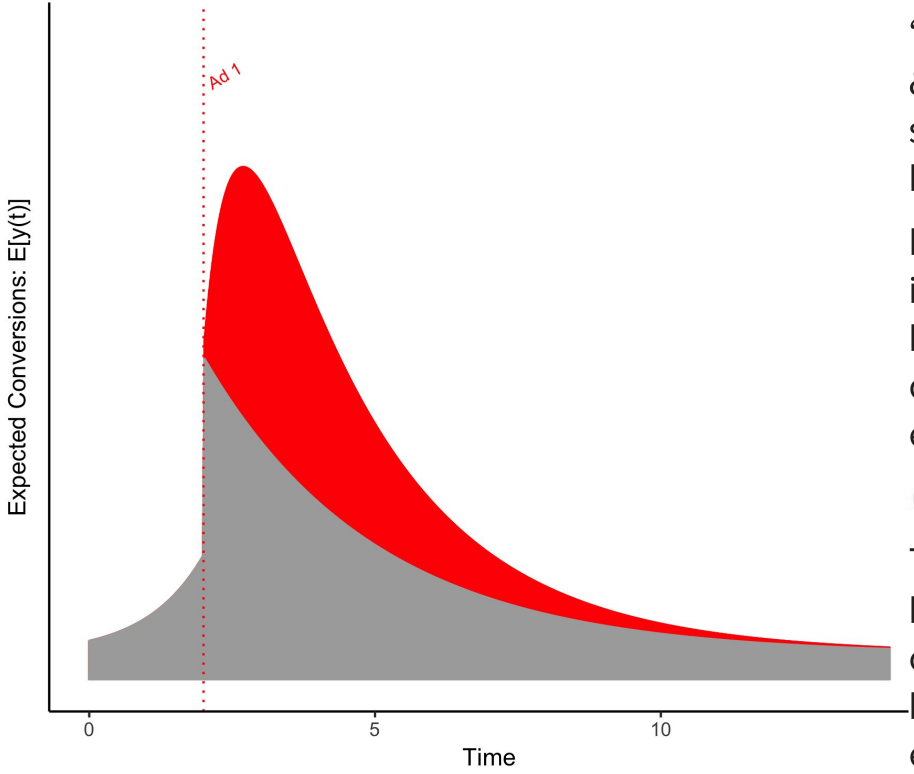


Simultaneous treatment in advertising results from the websites that the user visits having a direct or indirect effect on the likelihood of the outcome.

$$\alpha_i(t) = \alpha_i(t \mid page\ views_i)$$

For example, if a website is about a TV show or movie that is available on Netflix, the webpage content might boost conversions.

The Worst Endogeneity: Activity Bias



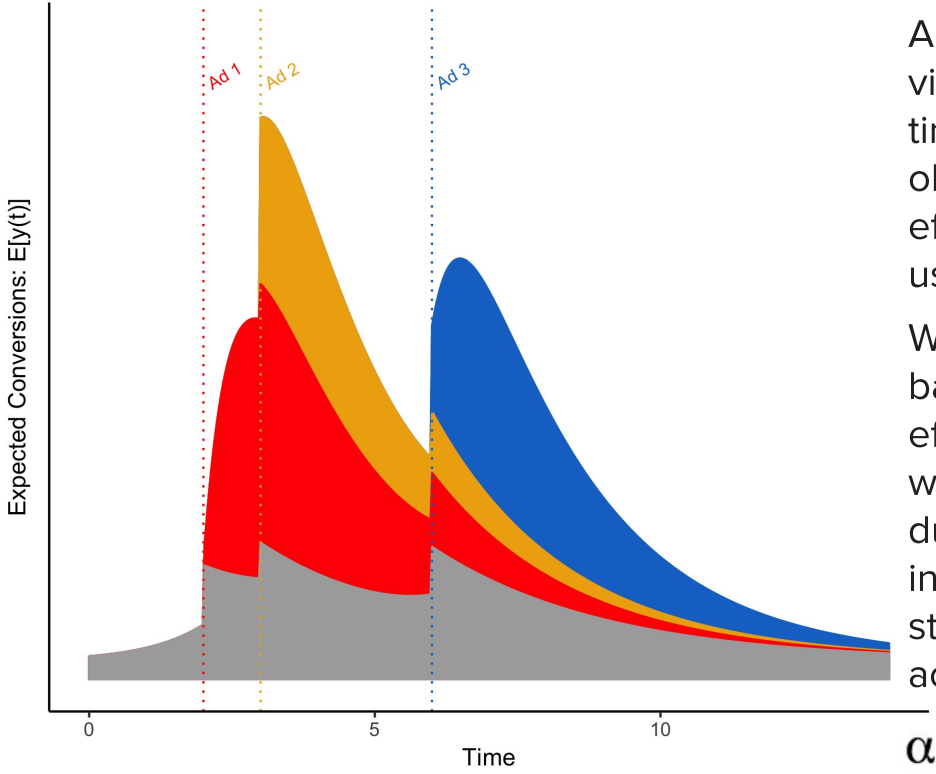
"Activity bias" (Lewis, Rao, & Reiley 2011) is another source of non-constant baselines.

Experiments show spikes in conversion activity both before and after other online events, absent ad exposure (e.g., placebos).

$$\alpha_i(t) = \alpha_i(t | page\ views_i)$$

These contemporaneous, but not causal, spikes are called "activity bias" because they bias causal estimators on panel data.

The Worst Endogeneity: Activity Bias



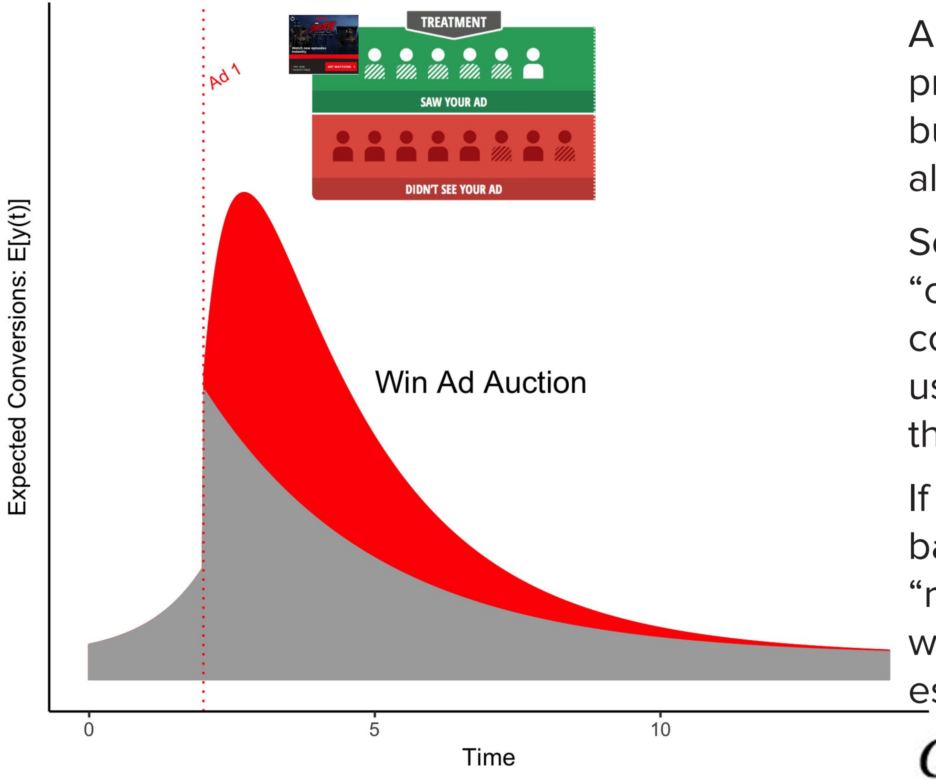
Activity bias, when visualized in continuous time, illustrates how hard obtaining causal treatment effect estimates can be using observational data.

While “controlling for baseline activity” can be effective in some settings, we are pessimistic for ads due to the selection bias introduced by a continuous stream of endogenous user activity.

$$\alpha_i(t) = \alpha_i(t \mid page\ views_i)$$

NETFLIX

The Worst Endogeneity: Random Non-Compliance?



Advertising auctions provide many chances to buy ads. But we do not always win.

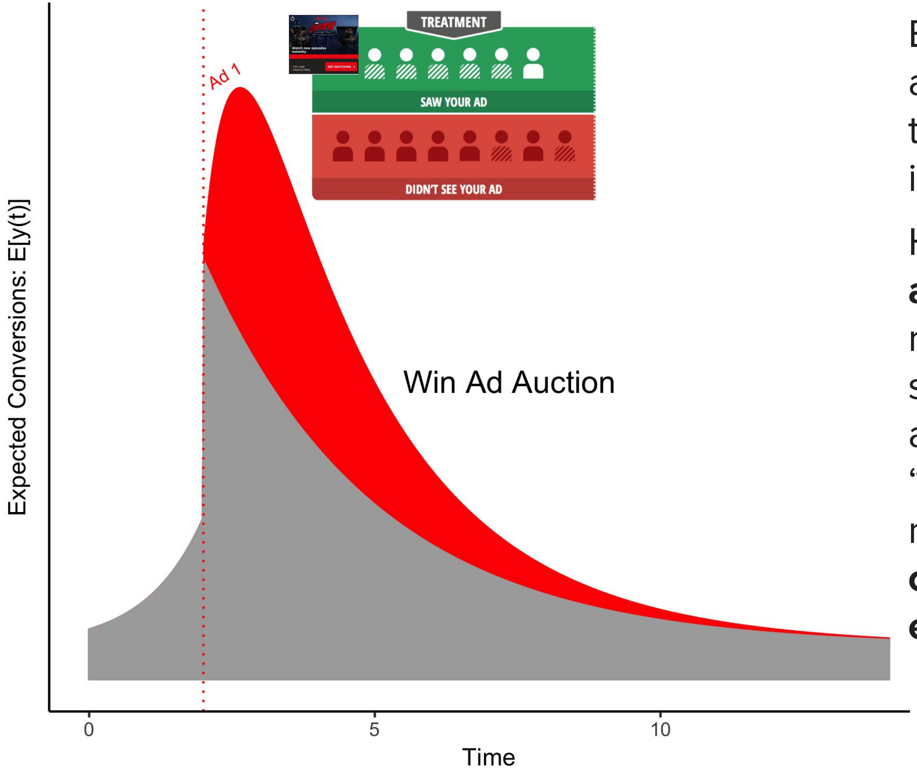
So, perhaps, we could “control” for activity bias by comparing purchases of users who see the ads to those who do not.

If winning the auction is basically random, implying “random non-compliance,” we can interpret our estimates as causal.

$$Cov(x_i(t), \varepsilon_i(t)) = 0$$

Source: Johnson, Lewis, and Nubbemeyer (2016), Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

The Worst Endogeneity: Non-Random Non-Compliance



But advertising auctions are ranking mechanisms that pool private information across bidders.

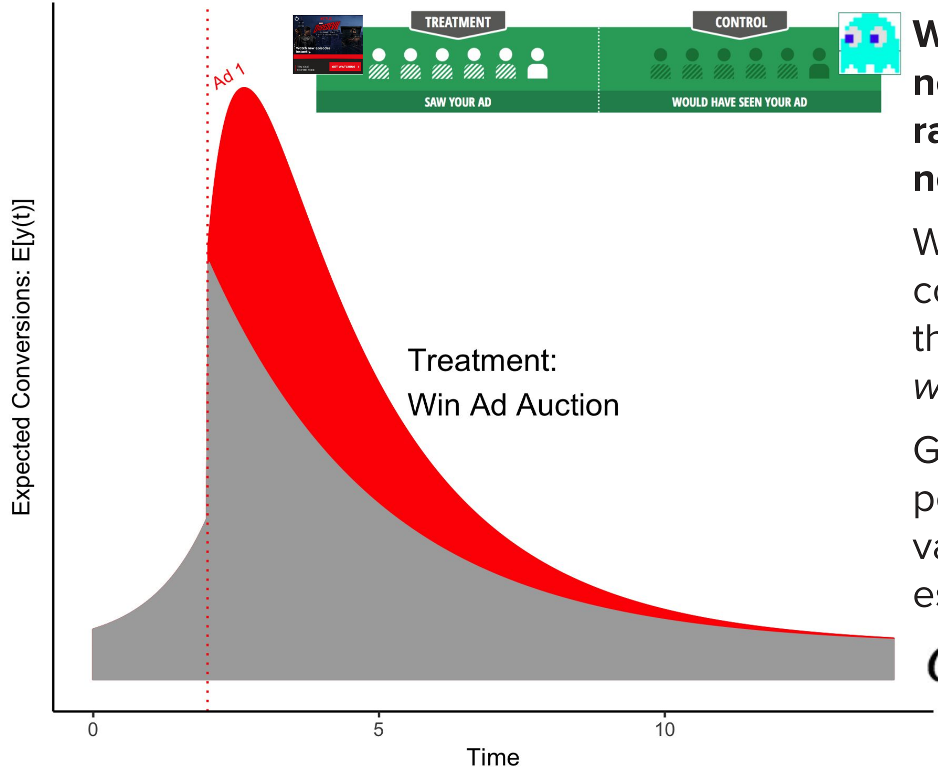
Hence, **winning the auction is not random**, but rather correlated with user socioeconomics, behavior, and ad quality. Due to this "***non***-random non-compliance," **we cannot interpret our estimates as causal.**

$$Cov(x_i(t), \varepsilon_i(t)) \neq 0$$

Source: Johnson, Lewis, and Nubbemeyer (2016), Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

NETFLIX

The Worst Endogeneity Solved: Instrumental Variables (IV)



We can engineer random non-compliance by randomizing whether or not we show an ad.

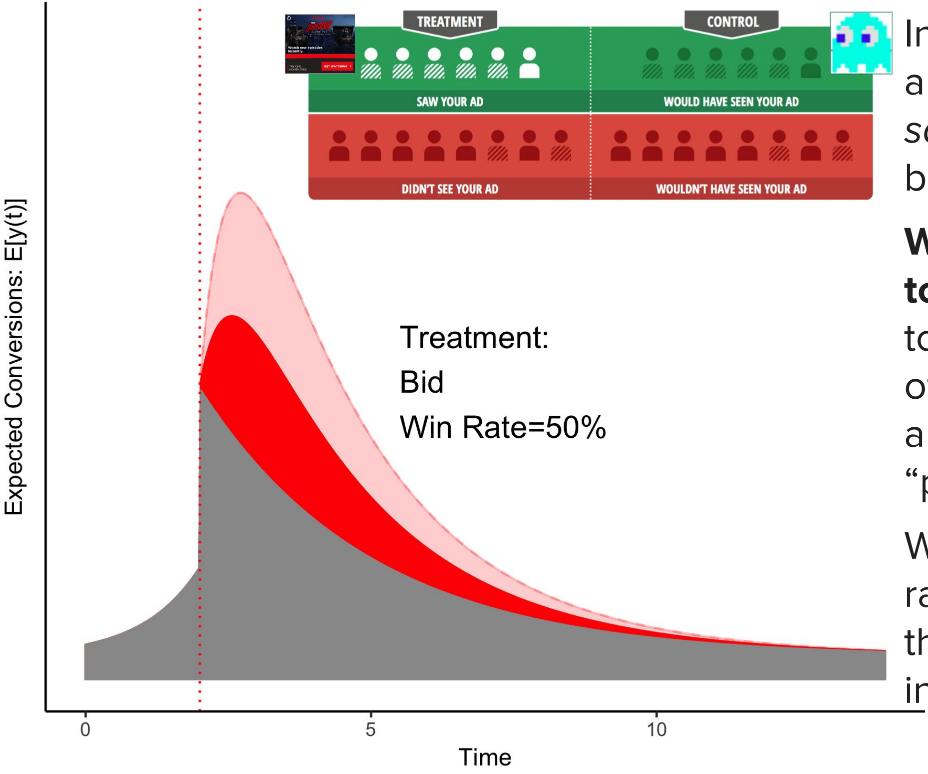
With "ghost ads" we compare users who saw the ad with those who *would have* seen the ad.

Ghost ads are the most powerful instrumental variable and ensure our estimates are causal.

$$Cov(z_i(t), \varepsilon_i(t)) = 0$$

Source: Johnson, Lewis, and Nubbemeyer (2016), Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

The Worst Endogeneity Solved: Ghost Bids → Predicted Ghost Ads



In ad auctions, we both win and lose, even with the *same* bid due to other bidders' behavior.

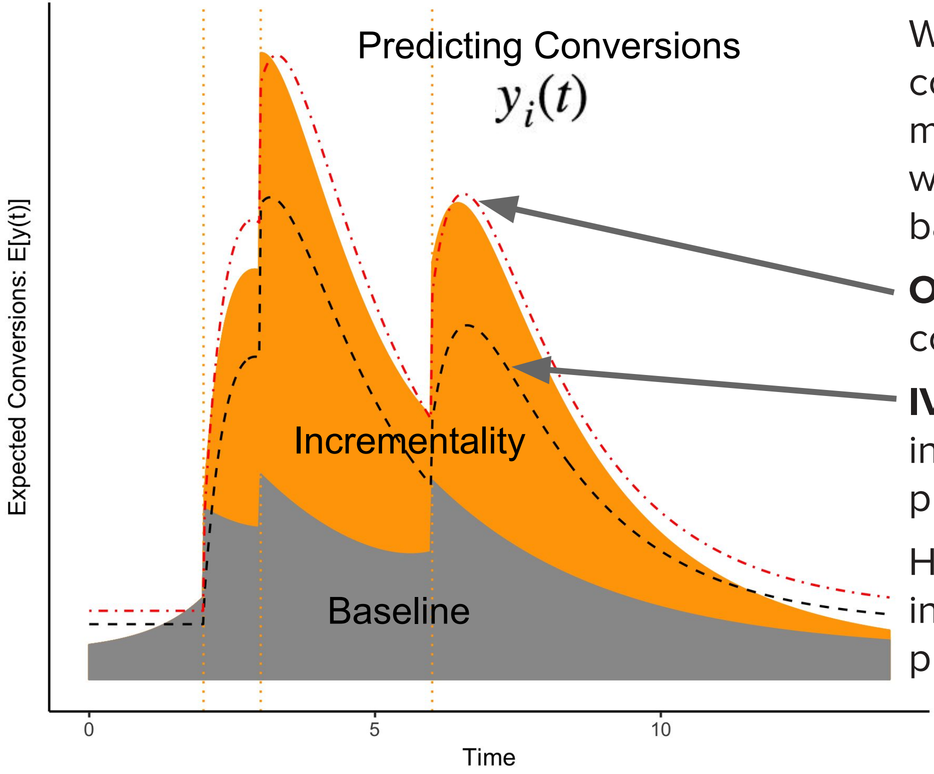
We record the bid we want to submit as a "ghost bid" to simulate the probability of winning that type of auction at our bid, yielding "predicted ghost ads."

When interacted with the randomization, these are the most powerful feasible instrumental variables.

Source: Johnson, Lewis, and Nubbemeyer (2016), Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

NETFLIX

The Worst Endogeneity Solved: Instrumental Variables with Ghost Bids



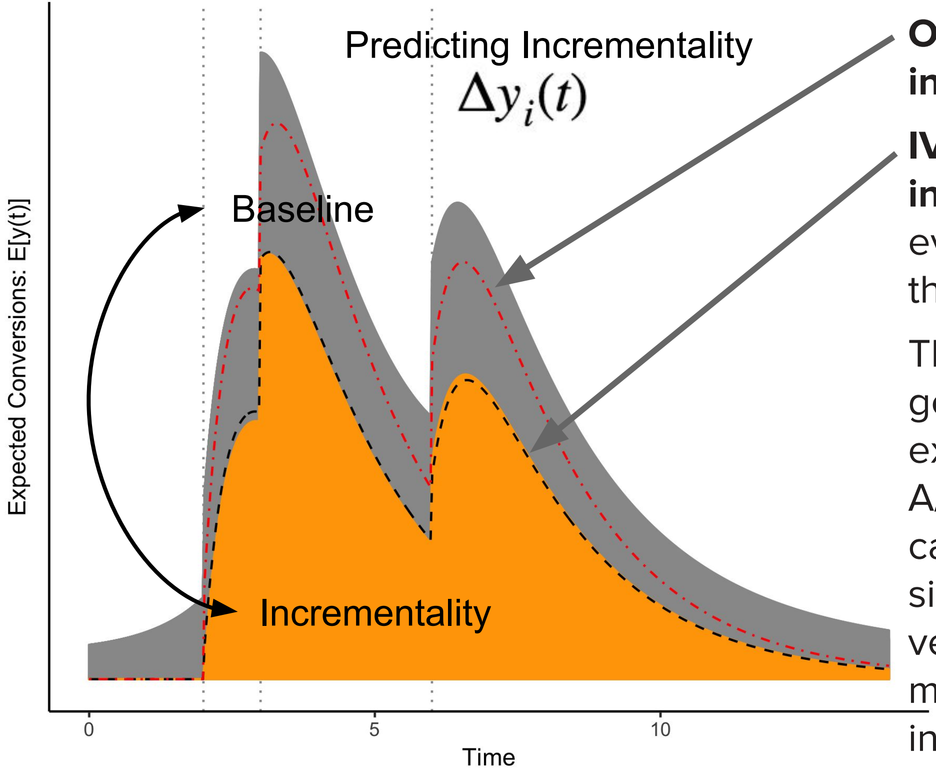
We can estimate the continuous time regression model with OLS or IV without fully modeling the baseline.

OLS can give good conversion predictions.

IV is expected to give inferior conversion predictions.

However, predicting incrementality is not predicting conversions.

The Worst Endogeneity Solved: Instrumental Variables with Ghost Bids



OLS estimates of incrementality are biased.

IV estimates the correct incrementality effects, even without fully modeling the baseline.

This works because IV generalizes randomized experiments. For example, A/B testing is a special case of IV, estimating a simple causal effect of A versus B, without fully modeling all factors that influence the outcome.

Production-Ready Causal Machine Learning

Advanced Incrementality
for Industry

NETFLIX

Requirements: Production-Ready Causal Machine Learning

- **Causal**: Its predictions are not dependent on the distribution of the training data remaining stable. E.g., offline training ⇒ online performance. $E[\hat{\beta}|X] = \beta$
- **Predictive**: Its predictions are as precise as possible out of sample. E.g., "regularization" tuning via a valid, automatic, and feasible cross-validation procedure. $min_{\lambda} \sum_{i \in CV} \left(Y_i - \hat{Y}_i\left(\hat{\beta}(\lambda)\right)\right)^2$
- **Scalable**: The model can be estimated with a large number of sparse features. I.e., no matrix inverses, use of importance sampling to utilize informative gradients. $\hat{\beta} = argmin_{\beta} \ L(\beta|x_{ik}); \ k >> 10{,}000$
- **Efficient**: Minimum variance estimator within its class. $min_{\hat{\beta} \in B} \ Var(\hat{\beta})$

$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

Optimal Instrumental Variables: Causal & Efficient

- **Causal**: Instrumental Variables estimation.

$$g(Z)'\varepsilon(\hat{\beta}_{IV}) = 0 \Rightarrow \hat{\beta}_{IV} = (g(Z)'X)^{-1}(g(Z)'Y)$$

$$plim_{N\to\infty}\hat{\beta}_{IV} = \beta$$

- **Efficient**: Minimum variance nonlinear basis functions & regularization.

$$Var(\hat{\beta}_{IV}) \propto \sigma^2(X'g(Z)g(Z)'X)^{-1}$$

$$max_{g(\cdot)}\ g(Z)'X$$

Hausman Test: Causal *or* Predictive

$$\hat{\varepsilon}(\beta) = Y - X\beta$$

- **Causal (Consistent)**: Instrumental variables, e.g., 2-Stage Least Squares (2SLS).

$$\hat{\beta}_{IV} \; solves \; Z'\hat{\varepsilon}(\beta) = 0$$

- **Predictive (Efficient)**: Ordinary Least Squares (OLS).

$$\hat{\beta}_{OLS} \; solves \; X'\hat{\varepsilon}(\beta) = 0$$

- **Hausman Test (Frequentist): Consistent *or* Efficient**

Looks like L^2-penalization!

$$H = \frac{(\hat{\beta}_{IV} - \hat{\beta}_{OLS})^2}{Var(\hat{\beta}_{IV}) - Var(\hat{\beta}_{OLS})} \sim \chi^2(1)$$

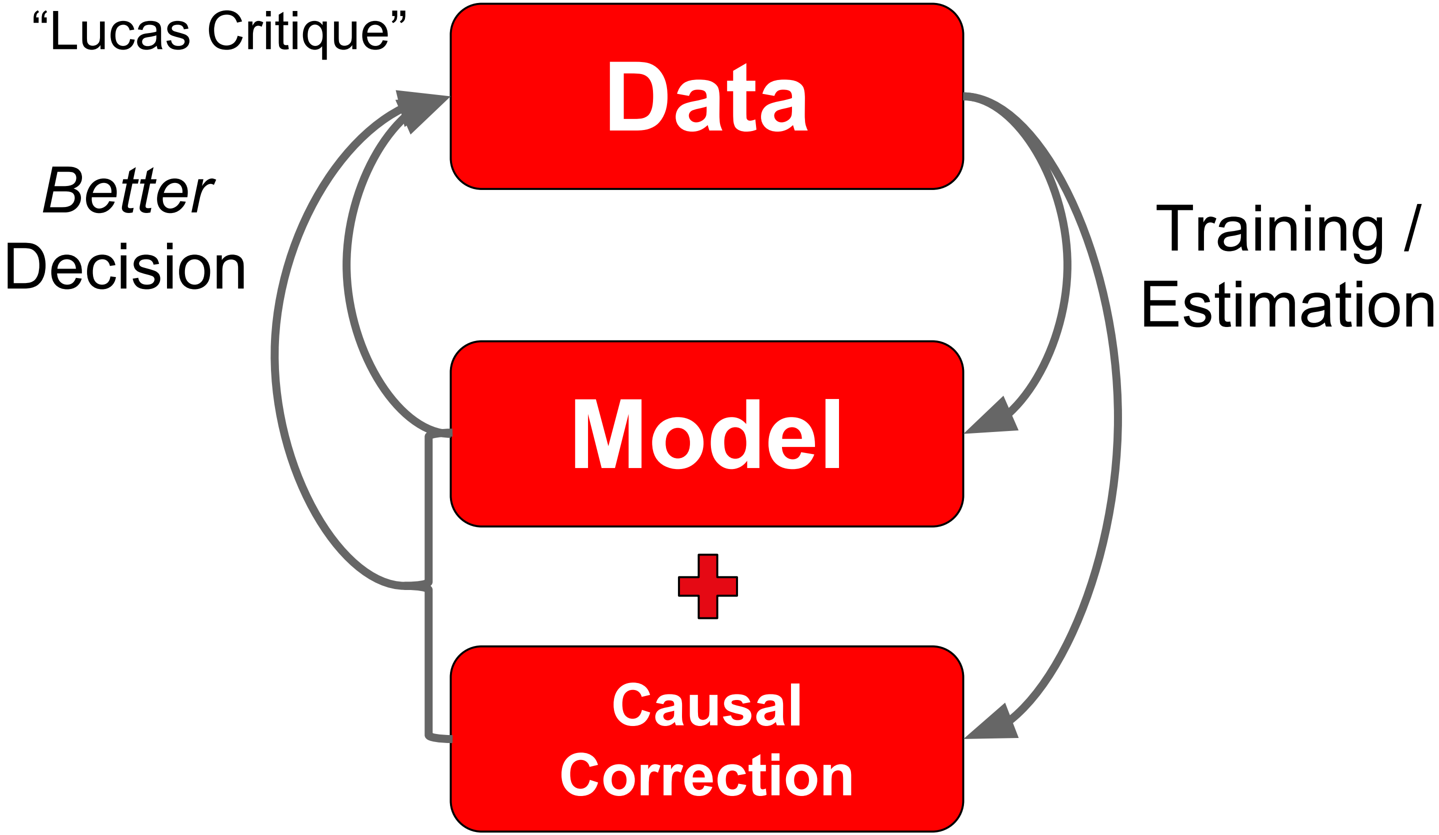
Hausman Penalization: Causally Consistent *and* Predictive

- **Causal**: Optimal IV.
- **Predictive**: Hausman Penalization to OLS/Ridge Regression (or other “best-in-class” predictive estimator).

$$\hat{\beta}_{Hausman} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)'Z\tilde{\Omega}^{-1}Z'\hat{\varepsilon}(\beta) + \boxed{\lambda_{Hausman}||\beta - \hat{\beta}_{Ridge}||^2}$$

$$\hat{\beta}_{Ridge} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)'\hat{\varepsilon}(\beta) + \lambda_{Ridge}||\beta||^2$$

$$\tilde{\Omega}^{-1} \approx Var(\varepsilon'Z)^{-1}$$


"Lucas Critique"
Data
Better Decision
Training / Estimation
Model
Causal Correction

Simple Hausman Penalization: Control Fn. via Ridge Regression

- **Control Function Approach to 2SLS :**

1. Estimate OLS of X on Z to obtain $\hat{\nu} = X - \hat{X} = X - Z\hat{\pi}_{OLS}$.
2. Estimate OLS of Y on $X, \hat{\nu}$ to obtain $\hat{\beta}_{2SLS}, \hat{\beta}_{\hat{\nu}}$.
3. Test $\hat{\beta}_{\nu} = 0$ for the Hausman test .

Key Observation!

$$\boxed{\hat{\beta}_{\hat{\nu}} = \hat{\beta}_{OLS} - \hat{\beta}_{2SLS}}$$

NETFLIX

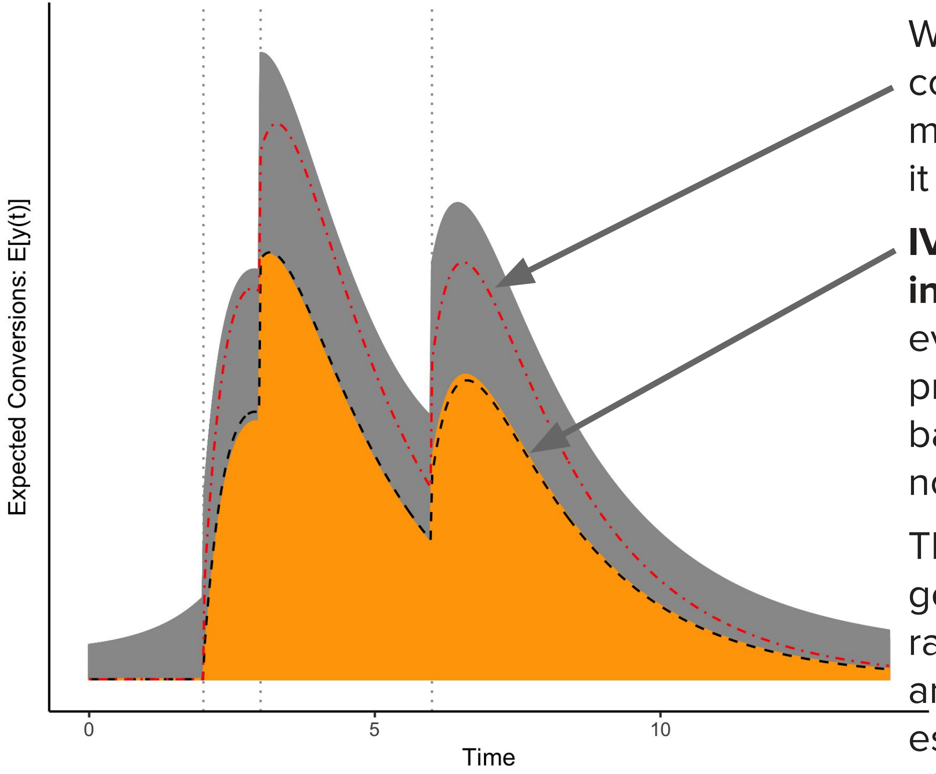
Simple Hausman Penalization: Control Fn. via Ridge Regression

- **Control Function Approach to 2SLS (Hausman):**

  1. Estimate OLS of X on Z to obtain $\hat{v} = X - \hat{X} = X - Z\hat{\pi}_{OLS}$.
  2. Estimate OLS (Ridge) of Y on $X, \hat{v}$ to obtain $\hat{\beta}_{2SLS}, \hat{\beta}_{\hat{v}}$.
  3. Test $\hat{\beta}_v = 0$ (L^2 penalize $\hat{\beta}_v$) for the Hausman test (penalization).

- Cross validation just works! Asks "are X's correlational and causal coefficients different?"
- Obvious generalizations: Elastic Net (2^{nd} stage), ~~ML (1^{st} stage)~~?
- **Not Scalable:** $\hat{v} = X - \hat{X} = X - Z\hat{\pi}_{OLS}$ is dense and $\hat{\pi}_{OLS}$ is O(dim(X)*dim(Z)).

Simple Hausman Penalization: Control Fn. via Ridge Regression

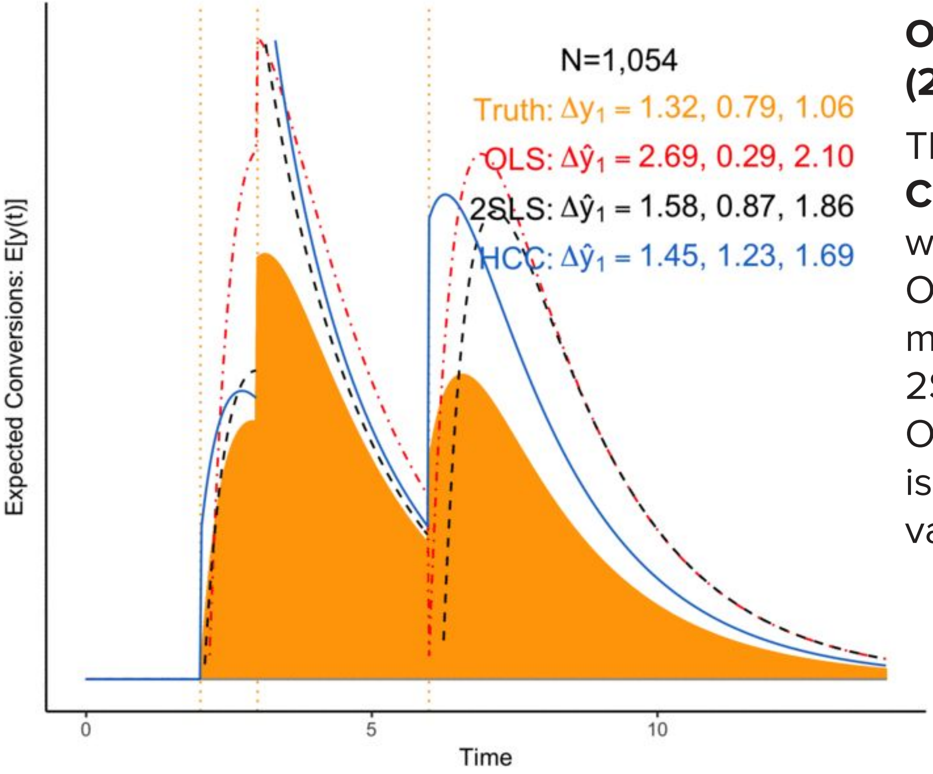


We can estimate the continuous time regression model with **OLS**; however, it will be biased.

IV estimates the correct incrementality effects, even while failing to properly model the baseline---because we did not even attempt to do so.

This strength of IV is the generalization of randomized experiments and A/B testing which estimate a simple causal effect of A versus B.

Simple Hausman Penalization: Control Fn. via Ridge Regression

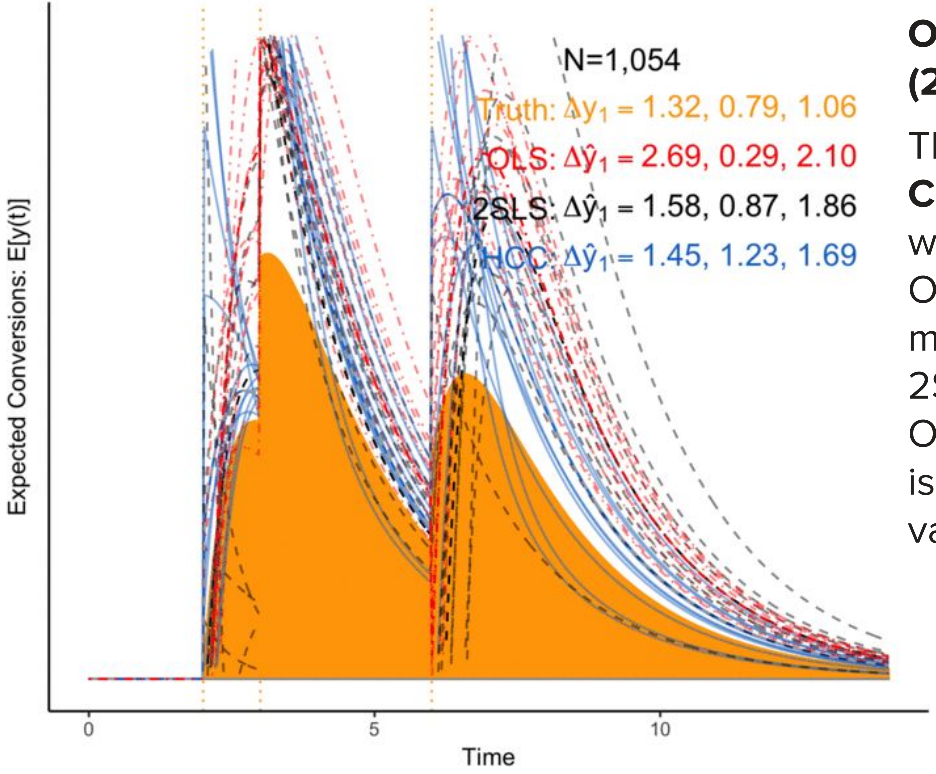


OLS is biased, but **IV (2SLS)** is not.

The **Hausman Causal Correction (HCC)** begins with an estimate close to OLS but then eventually migrates all the way to 2SLS once it is obvious that OLS != 2SLS. Hence, HCC is consistent but reduces variance early on.

NETFLIX

Simple Hausman Penalization: Control Fn. via Ridge Regression



OLS is biased, but **IV (2SLS)** is not.

The **Hausman Causal Correction (HCC)** begins with an estimate close to OLS but then eventually migrates all the way to 2SLS once it is obvious that OLS != 2SLS. Hence, HCC is consistent but reduces variance early on.

Hausman Causal Correction: Hausman Penalization in Practice

- Estimate correlational (e.g., classical machine learning) model. Compute residual.

$$\hat{\varepsilon} = Y - f_{Corr}(X|\beta)$$

- Estimate causal model on residual with penalization on Δβ.

$$\hat{\varepsilon} = f_{Causal}(X, Z|\Delta\beta)$$

- Model is a hybrid model: Initial marginal effect with causal correction.

$$\frac{\Delta y}{\Delta x} = \beta + \Delta\beta \qquad \frac{\Delta y}{\Delta x} = \frac{df_{Corr}}{dx} + \Delta\beta \qquad \frac{\Delta y}{\Delta x} = \frac{df_{Corr}}{dx} + \frac{df_{Causal}}{dx}$$

Linear | Quasi-Linear | Nonlinear

NETFLIX

SGD IV: Scalable

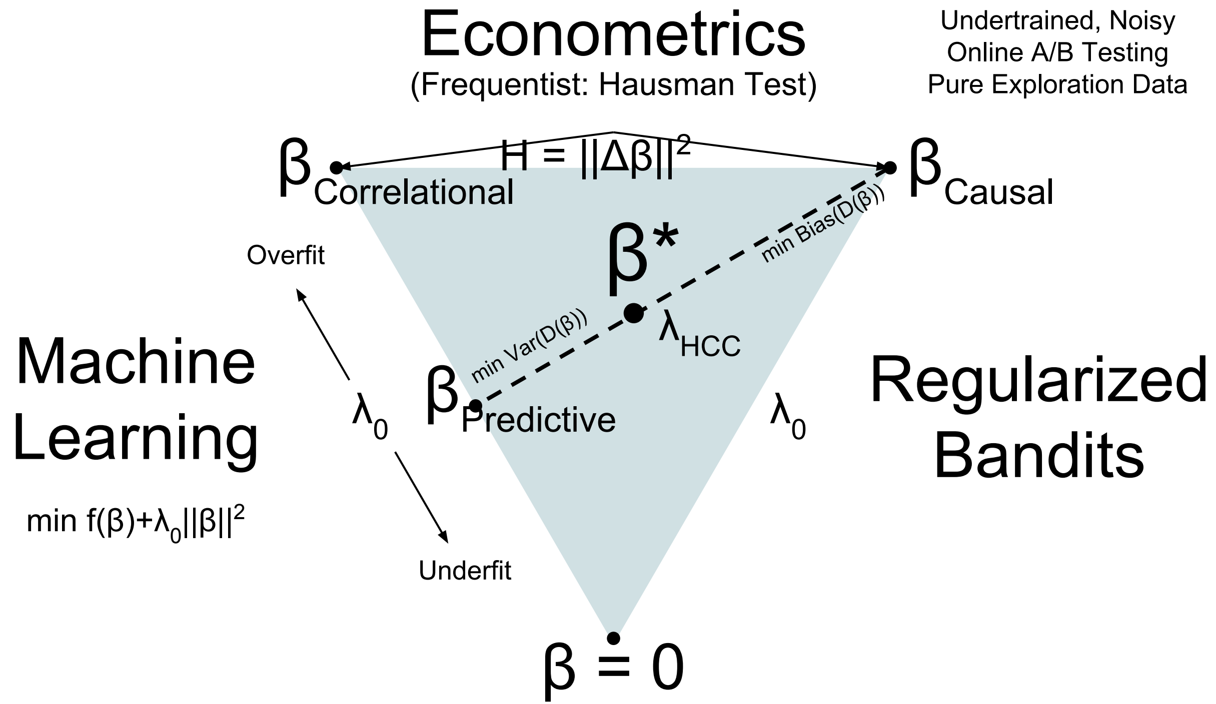
- **Causal**: Optimal IV.
- **Predictive**: Hausman Penalization to OLS/Ridge Regression (or other "best-in-class" predictive estimator).

$$\hat{\beta}_{Hausman} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)'Z\tilde{\Omega}^{-1}Z'\hat{\varepsilon}(\beta) + \lambda_{Hausman}||\beta - \hat{\beta}_{Ridge}||^2$$

$$\hat{\beta}_{Ridge} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)'\hat{\varepsilon}(\beta) + \lambda_{Ridge}||\beta||^2$$

$$\boxed{\tilde{\Omega}^{-1} \approx Var(\varepsilon'Z)^{-1}}$$

Consistent ML: Hausman Penalization



$$\beta^* = \text{argmin } f_{Corr}(\beta)+\lambda_0||\beta||^2+f_{Causal}(\beta,\Delta\beta)+\lambda_{Hausman}||\Delta\beta||^2$$

NETFLIX

Requirements: Production-Ready Causal Machine Learning

- **Causal**: Linear IV. $E[\hat{\beta}|X] = \beta$
- **Predictive**: Hausman Penalization via HCC. $min_{\lambda} \sum_{i \in CV} \left(Y_i - \hat{Y}_i\left(\hat{\beta}(\lambda)\right)\right)^2$
- **Scalable**: Estimation via SGD IV (>5,000 features) or control function approach using PCG (<=5,000 features). $\hat{\beta} = argmin_{\beta} L(\beta|x_{ik});\ k >> 10{,}000$
- **Efficient**: Large Scale Sparse Designer IVs + Feasible Optimal 2-Step GMM. $min_{\hat{\beta} \in B} Var(\hat{\beta})$

$$y_i(t) = \alpha(t) + \beta x_i(t) + \varepsilon_i(t)$$

$$\hat{\beta}_{Hausman} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)' Z \tilde{\Omega}^{-1} Z' \hat{\varepsilon}(\beta) + \lambda_{Hausman} ||\beta - \hat{\beta}_{Ridge}||^2$$

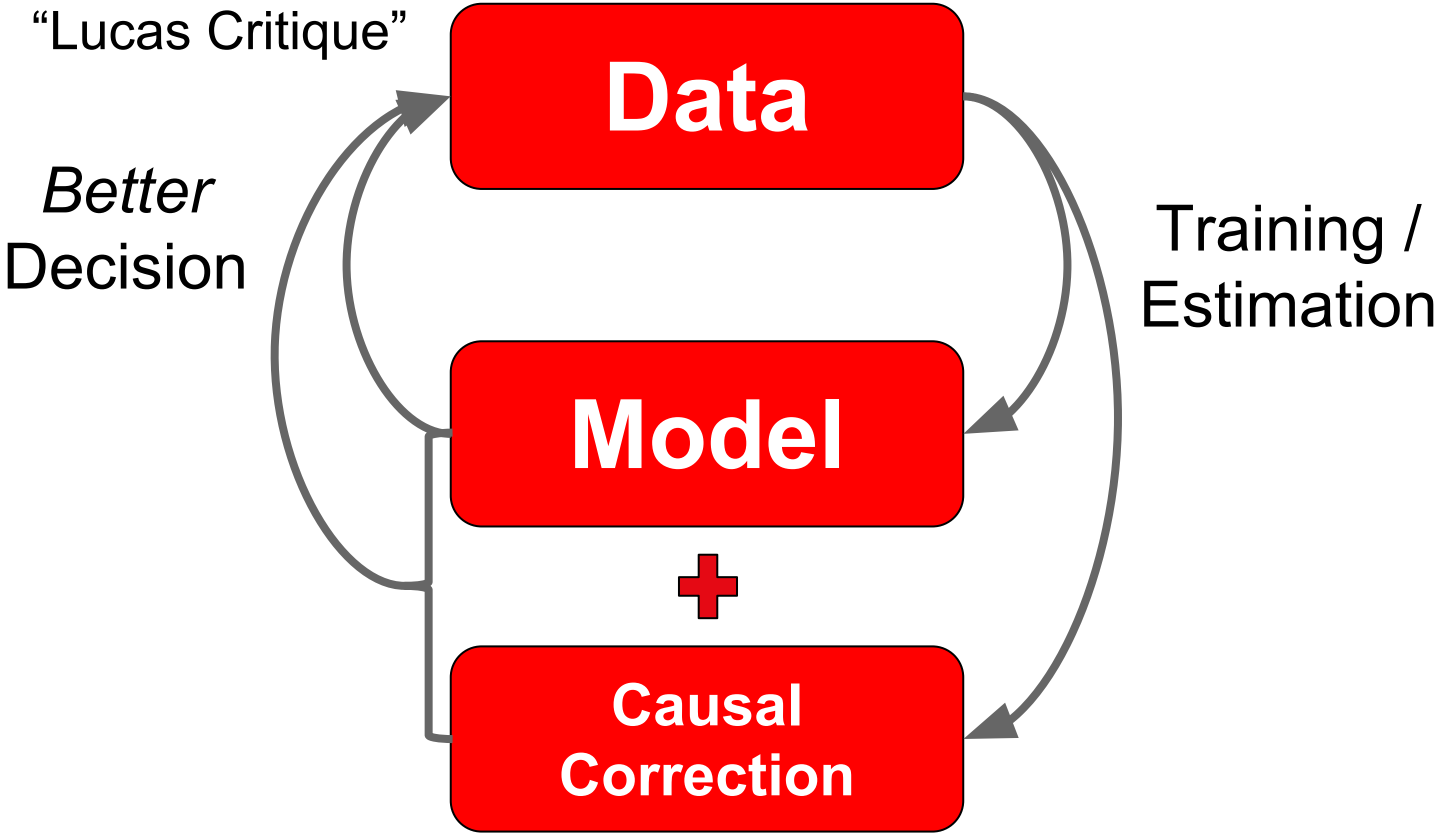
$$\hat{\beta}_{Ridge} \equiv argmin_{\beta}\ \hat{\varepsilon}(\beta)' \hat{\varepsilon}(\beta) + \lambda_{Ridge} ||\beta||^2$$

$$\tilde{\Omega}^{-1} \approx Var(\varepsilon' Z)^{-1}$$

Practical Causal Inference, Exploration, & Cross Validation

Advanced Incrementality
for Industry

NETFLIX


“Lucas Critique”
Data
Better
Decision
Training /
Estimation
Model
Causal
Correction

Thank you.

NETFLIX