Online Advertising Incrementality Testing

Practical Lessons, Paid Search And Emerging Challenges

Joel Barajas, Narayan Bhamidipati, James G. Shanahan

April, 2022

ECIR 2022

Who we are ...

Joel Barajas
Sr Research Scientist
Amazon, Marketing Measurement*,
Sunnyvale, CA, USA

[Linkedin]

Narayan Bhamidipati
Sr. Director, Research
Yahoo! Research,
Sunnyvale, CA, USA

[Linkedin]

James G. Shanahan
Church and Duncan Group Inc
UC Berkeley, CA, USA

[Linkedin]

*Work done while employed at Yahoo! Research.

Tutorial Parts

1. The basics: context and challenges
2. Incrementality Testing: concepts, solutions and literature
3. From concept to production: platform building, challenges, case studies
4. Deployment at Scale: test cycle and case studies
5. Emerging trends: identity challenges, industry trends and solutions

Part 3

From concept to production: platform building, challenges, case studies

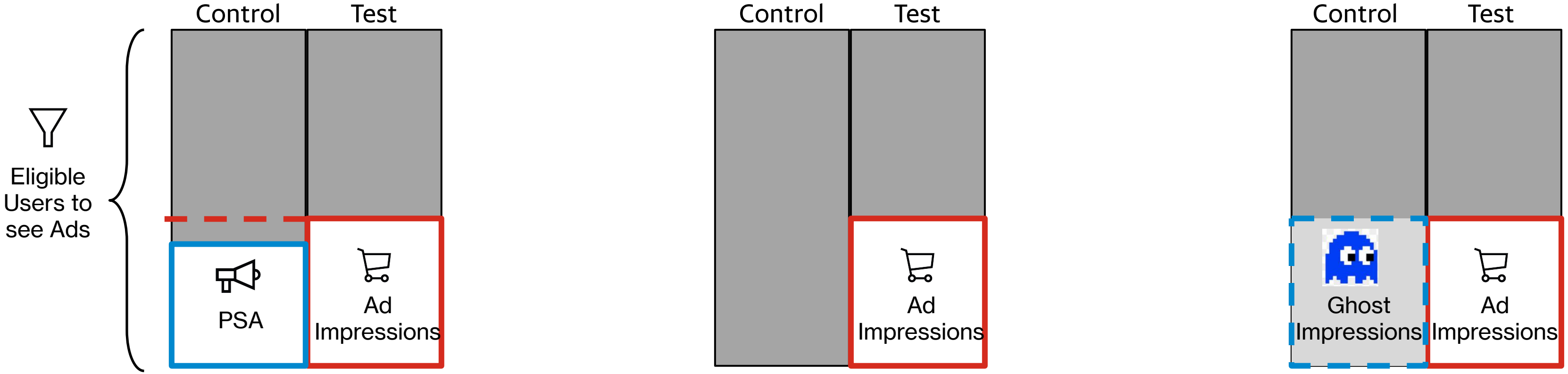
yahoo!

ECIR 2022

Experimentation Typical Designs for incrementality testing

Lewis et al. (2011), Barajas et al. (2016), Johnson et al. (2017), Barajas and Bhamidipati (2021)

Key Challenge: Identify would-be (counterfactual) impressed users



Case 1: Placebo PSA

Case2: Intent to Treat

Case 3: Ghost Ad Approach

Potential misalignment in user groups

Diluted effect design reducing the test power

Impressed and "ghost" impressed users are compared

“In Theory There Is No Difference Between Theory and Practice, While In Practice There Is”

To trust the numbers, a careful experimental design must be executed

yahoo!

ECIR 2022

Reviewing the Ad Serving Flow
Barajas et al. (2016), Barajas and Bhamidipati (2021)

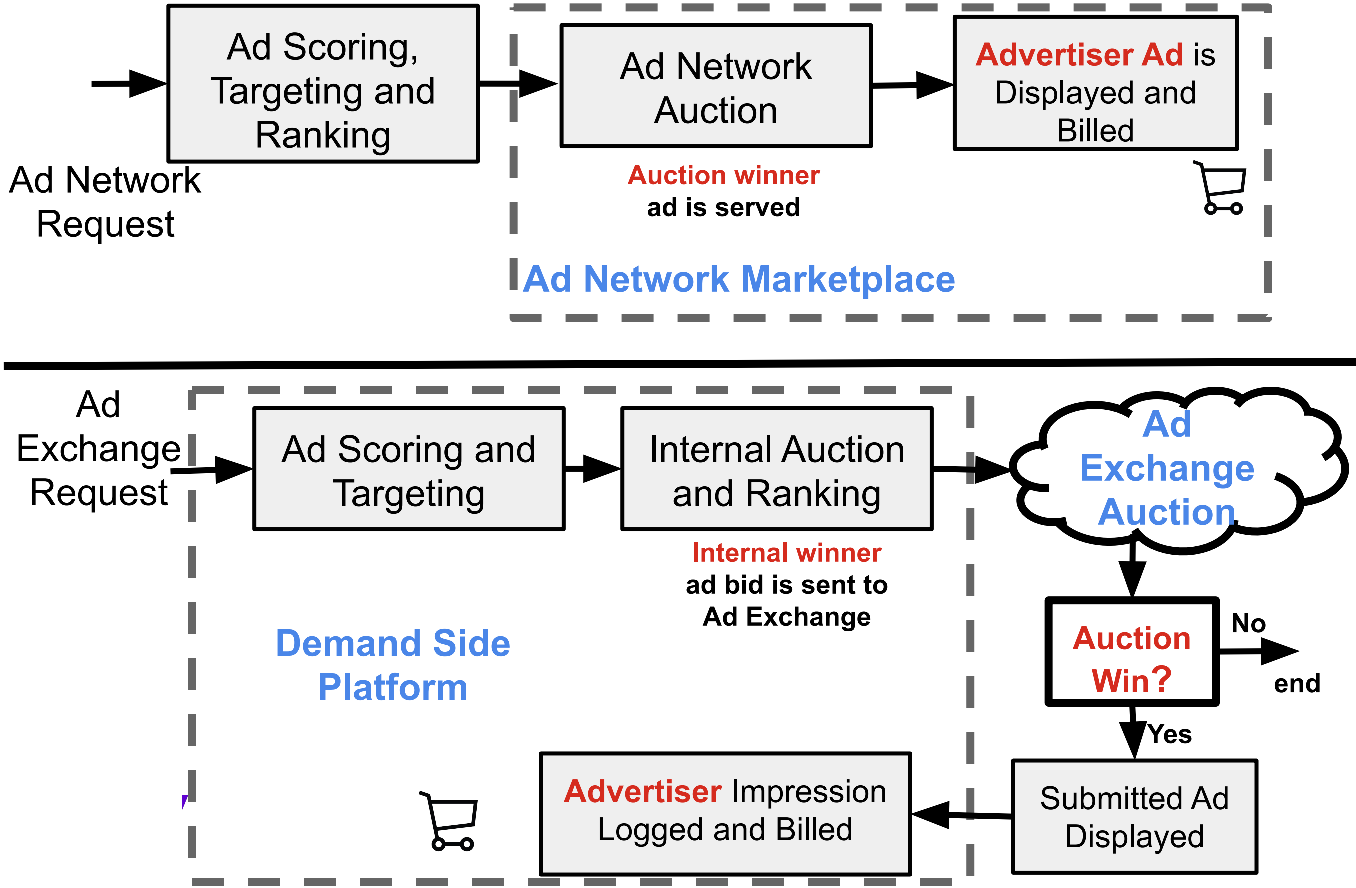
We review the ad serving flow to identify the right experiment intervention

- Typical in experimentation platforms, we need to identify **the experiment eligibility the “exposure” indicators**

- We want to have the data to reliably discard users whom we are certain they **do not have any effect from the treatment**
    - Adding users without effects **decreases the statistical power** and precision of the randomized design

Typical Ad Serving Flows

Ad Network Traffic



Treat/control checkpoint determines what is blind to the experiment

RTB traffic for DSP ad serving engines

Reviewing the Ad Serving Flow

Barajas et al. (2016), Barajas and Bhamidipati (2021)

The approach of the literature practices within the serving flow

- **PSA Testing**: Randomize users a priori, say via segments, **set up two campaigns equally**
  - **Pros:** segments are easily built, **ad exposure precision,** not major eng effort
  - **Cons:** cost of serving PSAs, **not double blind** and prone to selection bias

- **ITT Testing**: Randomize users at ad request time, say via hashing ids, **and holdout users from that advertiser**
  - **Pros:** **Blind to what is behind** the holdout in the flow
  - **Cons:** Need to include all the users after the holdout point, **even when some never see the ad**, modest eng effort

yahoo!

ECIR 2022

Reviewing the Ad Serving Flow
Barajas et al. (2016), Barajas and Bhamidipati (2021)

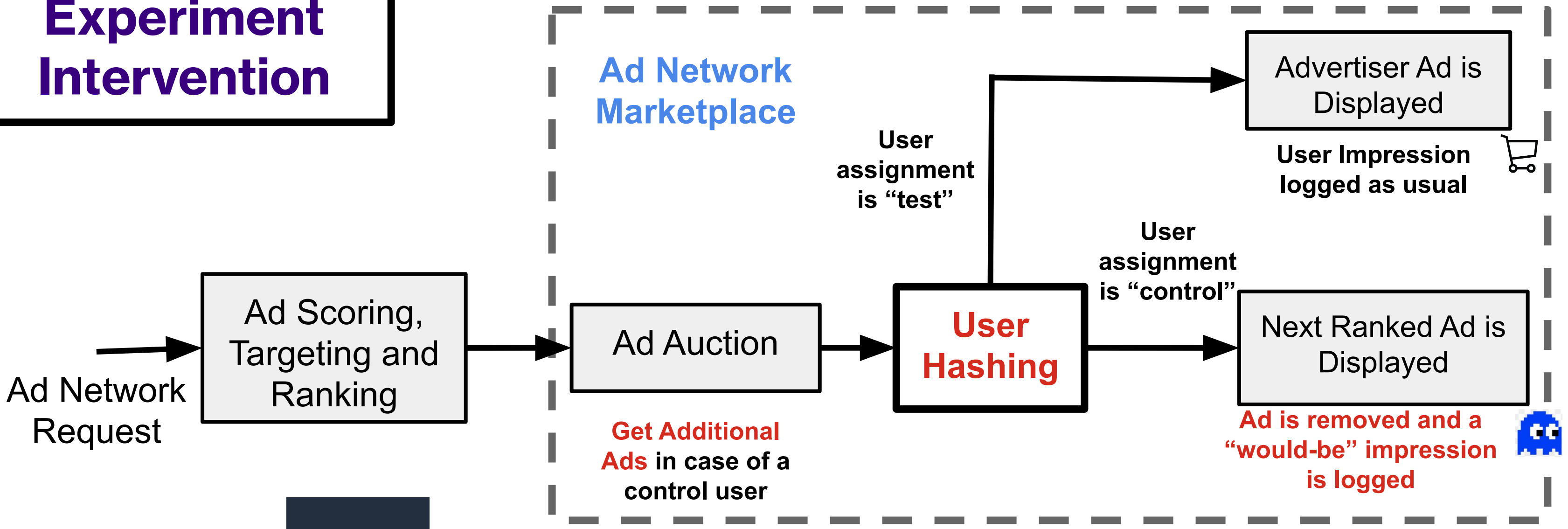
The approach of the literature practices within the serving flow

- **Ghost Ads Testing**: Randomize users at ad request time, say via hashing ids, hold out users from **the closest point to the ad exposure, and log the events**
  - **Pros:** All active users who reached the holdout point are in the analysis, with the **benefits of ITT and with the same precision of PSA testing**
  - **Cons:** **largest engineering effort**

Holdout Design: Ad Targeting and Auction blind

Barajas and Bhamidipati (2021)

Post-auction Experiment Intervention



Reviewing the Ad Serving Flow
Barajas and Bhamidipati (2021)

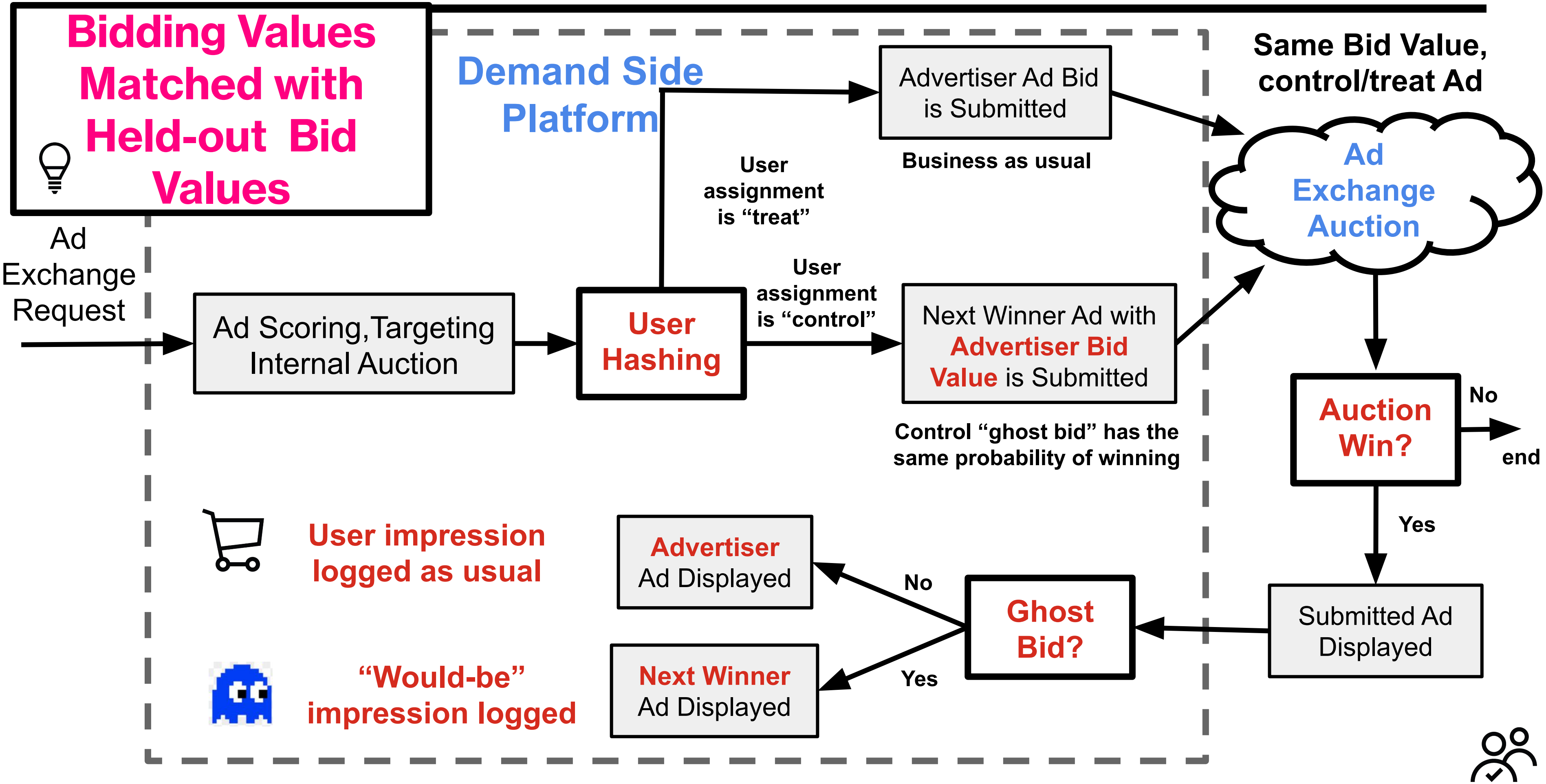
Ad Network Traffic when the ad auction is controlled

- Hold out execution is placed ***after the auction and before pricing***
    - It requires ***additional ads*** to be run to the ad serving flow
    - A separate data feed is needed to log ***these ghost events***
    - **Blind design** to the ad serving, as treatment administrator
    - It supports ***any targeting policy*** and/or regular targeting adjustments
    - Support for **long-term testing** as there is no constraint to targeting or user selection a priori
    - **It eliminates any ad targeting or auction bias**

yahoo!

ECIR 2022

Holdout Design: Programmatic Ad Exchanges Traffic



Reviewing the Ad Serving Flow
Barajas and Bhamidipati (2021)

RTB Traffic when the ad auction is executed by a third party

- Hold out execution is placed ***after the internal auction*** along with ***bid price matching***
  - It requires ***matching the bid price*** of the held out ad to the ad sent to the exchange in the control group
  - If the alternative ad sent to the exchange ***wins the action a ghost impression is logged*** for the advertiser running the test
  - It excludes users who never win any exchange auction when **typical winning impression rates are less than 15%**

Experimentation Units

The Role of the Identity Graph

yahoo!

ECIR 2022

Typical Randomization Units

A "user" is often a fragmented definition which varies by vendors

- **Cookie ids**: Cookies are the **core user representation** in web advertising
  - **Pros:** most traffic will have **assigned cookies**
  - **Cons:** **subject to block and deletion**, **limited to a browser**, and a light representation of a real person (studies suggests around 7 cookies per email).

- **Device ids**: Sticky ids to a mobile device which (sometimes) allow **a view of conversions across ad vendors**
  - **Pros:** They are sticky and **updated with low frequency** (with OS updates)
  - **Cons:** Only work for **in-app ads**, and subject to increasing privacy challenges

Typical Randomization Units

- **Email ids**: Relatively sticky and **easier to keep on a experiment group**
    - **Pros:** allows **cross-vendor and cross-device view** in the conversion joins
    - **Cons:** users must provide the **email address to convert AND to see the ad**, which adds friction and potential biases in both sides

- **Logged-in ids**: Robust id and typically used in product experimentation, but **NOT necessarily compatible** with the advertiser conversions
    - **Pros:** Cross-device user hold out and one of the **most stable ids** for a user
    - **Cons:** Need to create **shared ids with the advertiser** to do response link to the experimental group

yahoo!

ECIR 2022

Typical Randomization Units

- **Household ids**: Supports **spillovers among ids in a given household** but it is often probabilistic id
  - **Pros:** segments are easily built, **ad exposure precision**, not major eng effort
  - **Cons:** relies on estimated clustering groups, ie **probabilistic links**, which make the group less stable

- **Identity Graph ids**: Support a **combination of all ids** available to the graph
  - **Pros:** More **robust and stickier** than all other isolated ids
  - **Cons:** Need to handle **user spillovers** within the graph id escalation and expansion

Randomization Units: Takeaways
Kohavi et al. (2020)

Conditional on engineering trade-offs, always prioritize use of the most stable ids

- Identity graph based randomization units provide a balanced approach between **weaker ids and practical trade-offs**
  - Expect **spillovers** and have a mechanism ready to handle them
  - Validate with **A/A tests and run regular audits** in case of technology changes
  - Testing in the **open web** inevitable leads to a combination of ids
  - The statistical power is highly dependent on **the power of the identity graph**

ECIR 2022

User Conversion Joins

The Role of Last-touch Attribution

The Role of Last-touch Attribution?

None, nichts, aucun, ninguno, nessuno, ללא, 没有

yahoo!

ECIR 2022

The Role of Last-Touch Attribution: None

Attribution was created to “attribute” conversion value to a specific ad impression without a control group

- A control group of users allows us to observed the ***counterfactual user response and organic conversion rate***
    - Since everything we observe in **the treatment group is compared with the control group**, the concept of attribution becomes irrelevant

- The attribution rules often introduce **biases in the user response** even within the experimental setup
    - **Example:** **Video ads are rarely favored by last-touch attribution** compared to display or paid search

Effect of one or more ad impressions

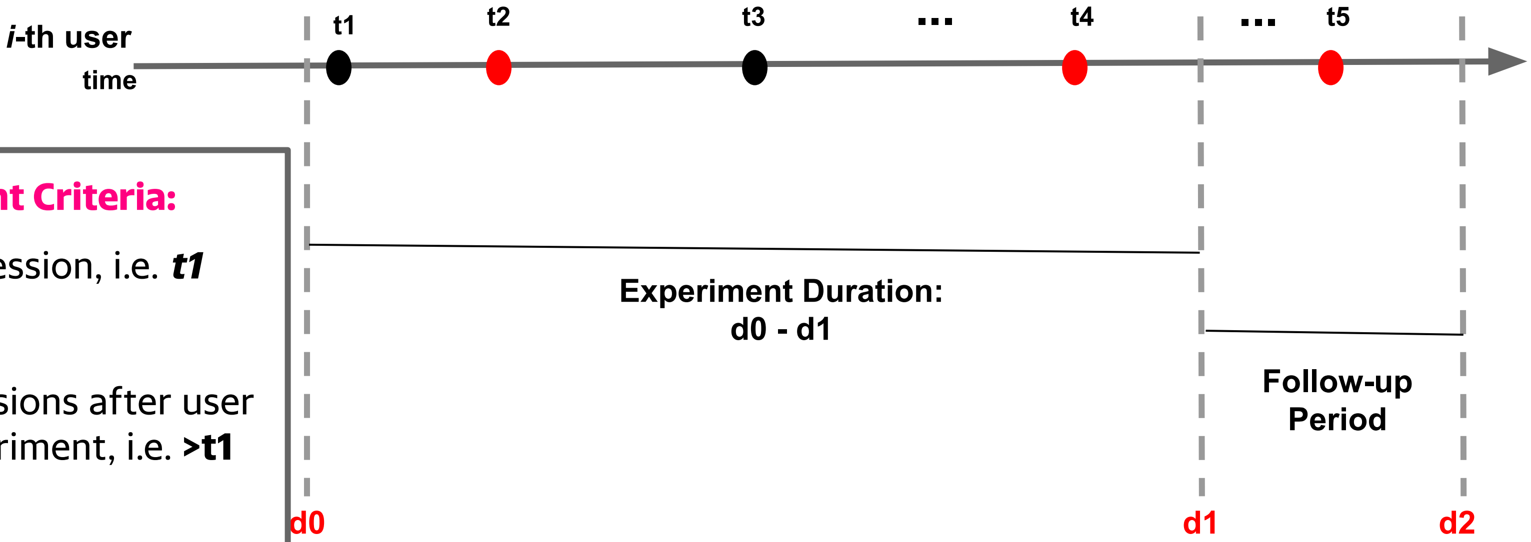
✓

User Timeline of Events

● User Qualifying Event (impression)

● User Response (conversion)

Last-touch *attribution* is just a user response join given the control group average user response



User Experiment Criteria:

First impression, i.e. ***t1***

User Response:

All conversions after user joins experiment, i.e. **>t1**

Metric Definitions

We define the following metrics (Marketing Effects):

1. **Converter Rate Lift (%):** Average Effect on **user converter probability** (binary indicator) over control converter probability

2. **User Conversions Lift (%):** Average Effect on **number of user conversions** (conversions per user) over control user conversions probability

3. **Cost Per Incremental Converter:** Aggregate marketing spend over average number of **incremental converters**.

4. **Cost Per Incremental Conversions:** Aggregate marketing spend over average number of **incremental conversions**.

Causal Estimation and Metrics

Average Treatment Effect (ATE) and Lift:

$Z_i = 0$ for the control and $Z_i = 1$ for the test group.

$$ATE = E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0] \qquad CR\ lift = ATE / E[Y_i | Z_i = 0]$$

Leveraging Central Limit Theorem:

$$ATE \sim N(\bar{Y}_1 - \bar{Y}_0, \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0})$$

$N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2

Metric: Y = Converters

$$n_Z = \#\ of\ users\ in\ group\ Z$$

$$\bar{Y}_Z = \frac{\#\ converters\ in\ group\ Z}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$S_Z^2 = \bar{Y}_Z * (1 - \bar{Y}_Z)$$

Metric: Y = Conversions

$$\bar{Y}_Z = \frac{\#\ conversions\ in\ group\ Z}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$\bar{Y^2}_Z = \frac{(\#\ conversions\ in\ group\ Z)^2}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$S_Z^2 = \bar{Y^2}_Z - (\bar{Y}_Z)^2$$

Cost per Incremental Converter

$$CPiA = \frac{marketing\ spend\ (\$)}{ATE \times (\#\ of\ users\ in\ test\ group)}$$

Incremental Return on Ad Spend

$$iROAS = \frac{ATE \times (\#\ of\ users\ in\ test\ group)}{marketing\ spend\ (\$)}$$

!

Potential Outcomes Causa Model:

Randomized Units must be aligned

Ignorable Treatment Assignment to Features:

No stratification or blocking necessary in the estimation

yahoo!

Example: COVID 19 vaccines effectiveness in clinical trials Pfizer vs J&J effective rates (source)

Among U.S. adults without immunocompromising conditions, vaccine effectiveness against COVID-19 hospitalization during March 11–August 15, 2021, was higher for the Moderna vaccine (93%) than the Pfizer-BioNTech vaccine (88%) and the Janssen vaccine (71%).

That is lift!!
They are not statistically different!!

yahoo!

The Role of Last-Touch Attribution: None

The test answers the incrementality value of *all* impressions in aggregate delivered to users

- The test ***can not*** answer:
  - The **ad impression** that caused a conversion
  - The **interactions among ads** (*halo effect*), eg prospecting and retargeting
  - The ideal **frequency cap**
  - The effect of ads on **the time to convert**, which is a censored data problem

yahoo! ECIR 2022

The Role of Last-Touch Attribution: None

The test answers the incrementality value of *all* impressions in aggregate delivered to users

- The test ***can*** answer:
  - The **aggregate channel effect (lift)** during a period of time including holidays
  - The channel efficiency (CPIA or iROAS) which is **comparable to other channels**
  - The interactions between **multiple conversions in the funnel**
  - The **aggregate user ad frequency** to achieve a minimum detectable lift
  - The best **look-back conversion join window** from a set of values

Limitations and Caveats

Limitations and Caveats

1. **No effect is the "Null Hypothesis"**: Limitations to measure the effect leading to no effect **does not imply** the effect does not present

2. **Measurement relies heavily on user groups:** Reliable user holdout depends on being able to **consistently identify users** (via their ids). Spillovers between groups lead to diluted effects and consequently to **value under-estimation**

3. **Cookie deletion:** Cookie deletion **is not an observable event**. Effects of these events on measurement lead to **value under-estimation** because we can not filter deleted cookie-based users of the analysis

4. **Other running campaigns** (could even be on the same platform) can dilute the effects if they are served to the holdout/control population.

With a testing framework it comes the testing cycle and planning

We'll review this cycle and marketing use cases in the next part of the tutorial...