Online Advertising Incrementality Testing

Practical Lessons, Paid Search And Emerging Challenges

Joel Barajas, Narayan Bhamidipati, James G. Shanahan

April, 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

Paid Search Evaluation

Testing Lower Funnel Advertising

ECIR 2022

Demand Captured vs Demand Generation Channels

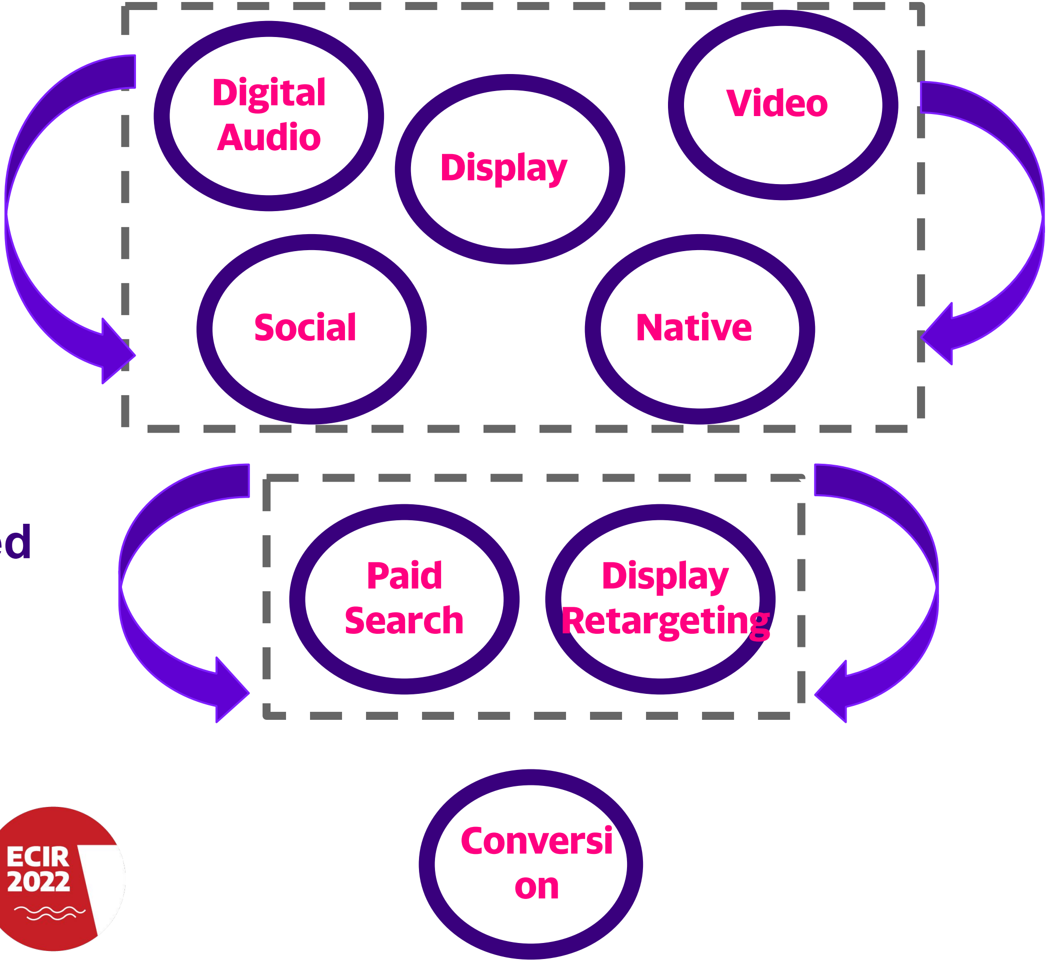
Li and Kannan (2014)

Upper Funnel

Lower Funnel

Advertiser Initiated Ads: Demand Generation

Customer Initiated Ads: Demand Captured

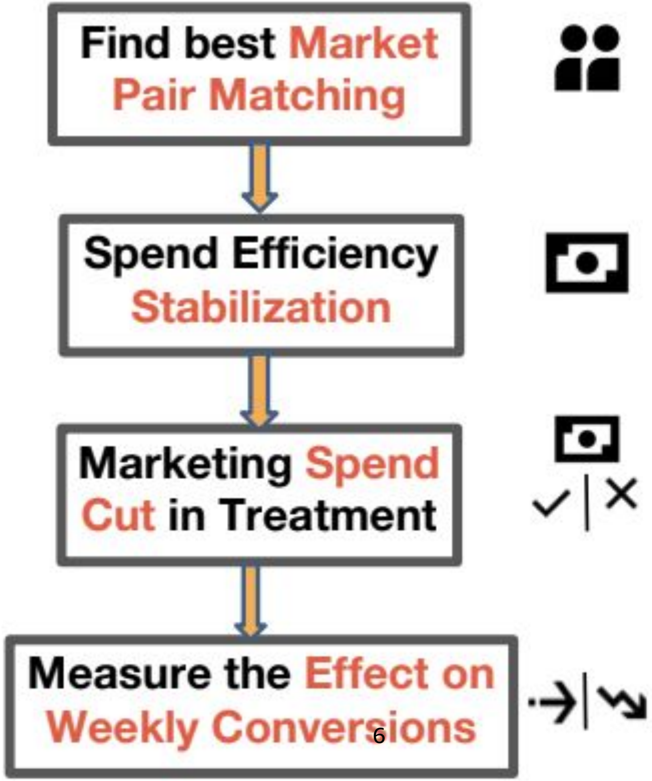


Testing Challenge: No User-level holdout
Blake et al. (2015), Barajas et al. (2020)

Without user level holdout, market pair testing is a viable solution

- Typical incrementality testing for advertisers when the ad network **does not support user-level holdouts**

Same concepts applied to hold out users in paid-search: e.g. display retargeting

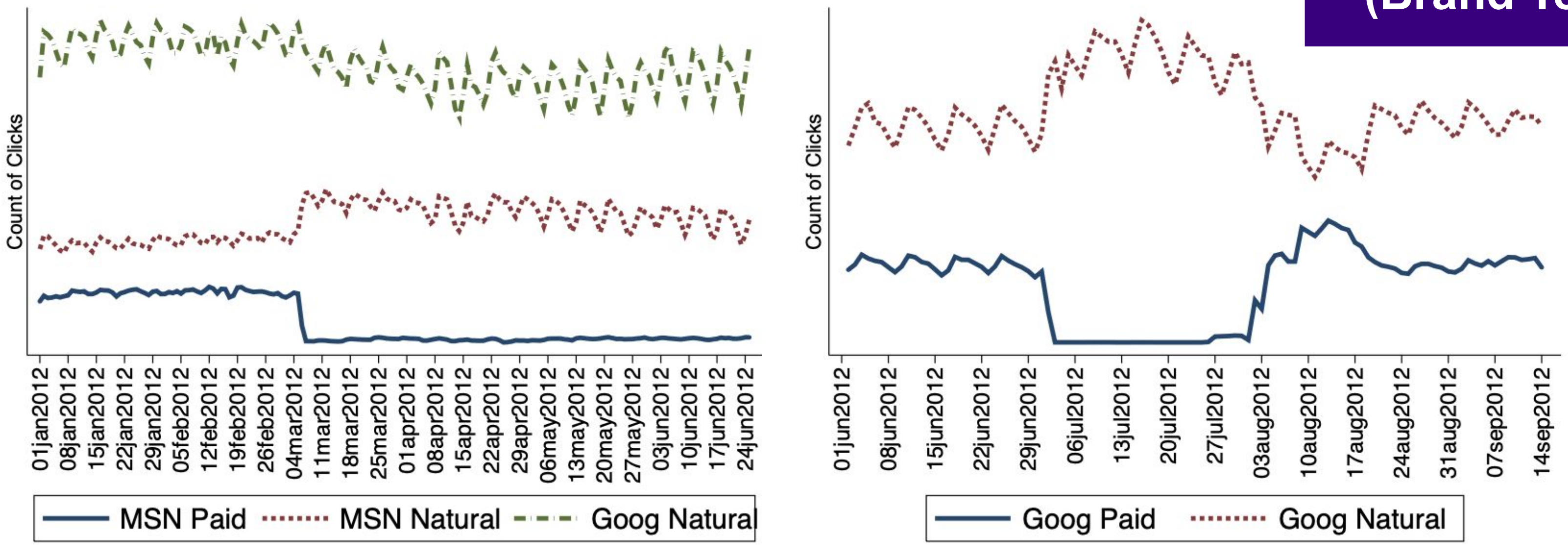


The Effects of Lower Channels: Organic vs Paid Search

Blake et al. (2015)

Cannibalization of Organic Search (Brand Terms)

Figure 2: Brand Keyword Click Substitution



(a) MSN Test

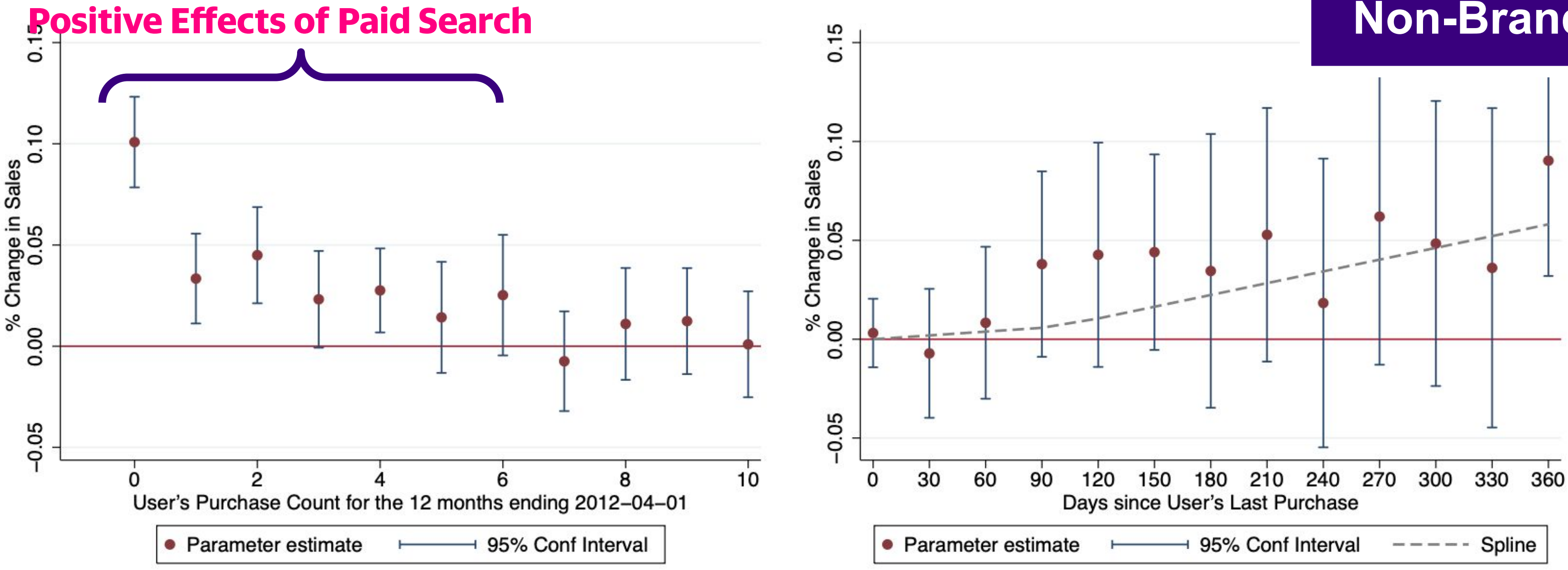
(b) Google Test

Paid Search: Positive effect on new/infrequent users

Blake et al. (2015)

Positive Effect on Least-Active Users: Non-Brand Terms

Figure 4: Paid Search Impact by User Segment



(a) User Frequency

(b) User Recency

yahoo!

Paid Search effect on Yelp Metrics

Dai and Luca (2016)

Dai and Luca (2016)

- **Randomizes Restaurants** (Advertisers) in Yelp
- Assigns **paid search packages to treatment** restaurants only (no ads to control restaurants)
- Positive effects on **upper-funnel metrics**:
  - Page views: 25%
  - Purchase intention metrics (directions to restaurant, browsing, etc): 9% - 18%
  - Number of reviews: 5%

Short-term top-funnel metrics: All effects droop to zero after test

yahoo!

ECIR 2022

Limitations and Caveats

yahoo!

ECIR 2022

Limitations and Caveats

1. Paid search studies suggest the channel **to be effective for short-term effects within the search marketplace** .

2. Ongoing research suggests a more **complex set of effects on marketplace prices** in e-commerce sites (see **Moshary (2021)**)
   a. The study suggests a potentially negative effect on total sales in the platform

3. More studies need to be developed

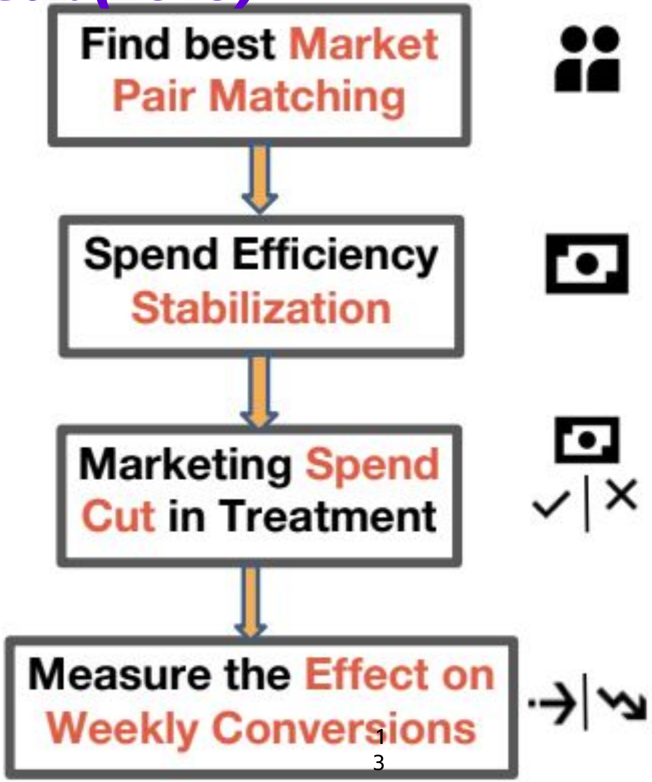
yahoo!

Geo-Testing

Testing with aggregate time series and geo testing units

Controlled Geo-Experiment + Synthetic Control

Barajas et al. (2020), Blake et al. (2015), Abadie et al. (2010)



Without user level holdout, market pair testing is a viable solution

- Typical incrementality testing for advertisers when the ad network **does not support user-level holdouts**

Causal Estimation: Synthetic Control

Barajas et al. (2020)

Bayesian Structural framework with **time series and a regression component** from the control market conversions to predict the treatment conversions (**synthetic control**).

Structural Equation

Time Series trend

Control Market Predictor

$$
\begin{aligned}
y_t^{(treat)} &= \underbrace{F_t\theta_t}_{\text{Time Series trend}} + \underbrace{x_t^{(control)T}\beta}_{\text{Control Market Predictor}} + \epsilon_t, \qquad \epsilon_t \sim N(0,\sigma^2),\\
\theta_t &= G_t\theta_{t-1} + \omega_t, \qquad \omega_t \sim N(0, W),
\end{aligned}
$$

Market Best Match: A/A tests

Barajas et al. (2020)

Best Pair selection given the conversion and estimation method

A/A test estimation: Given Causal Estimation Framework

Filter Best Pairs and Parameters: tightest and interval with zero effect

Estimation Parameters: Markets, Intervention times, Train/follow-up Length

Algorithm 1 Control/Treatment Market Pair Selection

```
Ω: Set of Markets to consider
Φ: Set of placebo intervention times
Δ_t: Time length of historical data
Δ_{t_i}: Time after placebo intervention
for all treatment market: m ∈ Ω do
    for all control market: n ∈ {Ω − m} do
        for all intervention time: d ∈ Φ do
            Fit the synthetic control model of Eq 1:
            Find Θ^s, s = 1, ..., N_s, given {y^(m)_{d−Δ_t:d−1}, x^(n)_{d−Δ_t:d−1}}
            Predict ŷ^{s(m)}_{t_i}, ∀s ∈ {s = 1, ..., N_s} after intervention,
            ∀t_i ∈ {d, ..., d + Δ_{t_i}}
            Estimate Credible Intervals (CI) lift_{cum(d+Δ_{t_i})}, Eq 3
        end for
    end for
    n*, d* ← tightest CI that include lift_{cum(d+Δ_{t_i})} = 0
    Append best control/treatment/time V = {V, (m, n*, d*)}
end for
return V
```

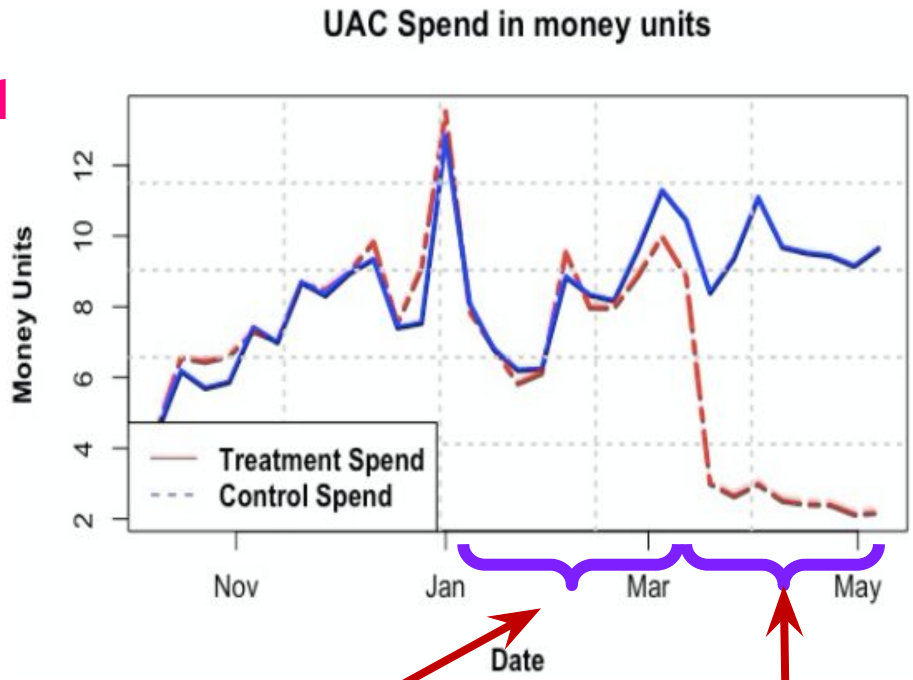
UAC Incrementality: Intervention

Barajas et al. (2020)

Given **treatment/control pairs and the estimation method**, we execute the experiment

1. Cost-per-attributed-signup (CPA) **stabilization** both groups

2. **Suspend** spend for treatment market



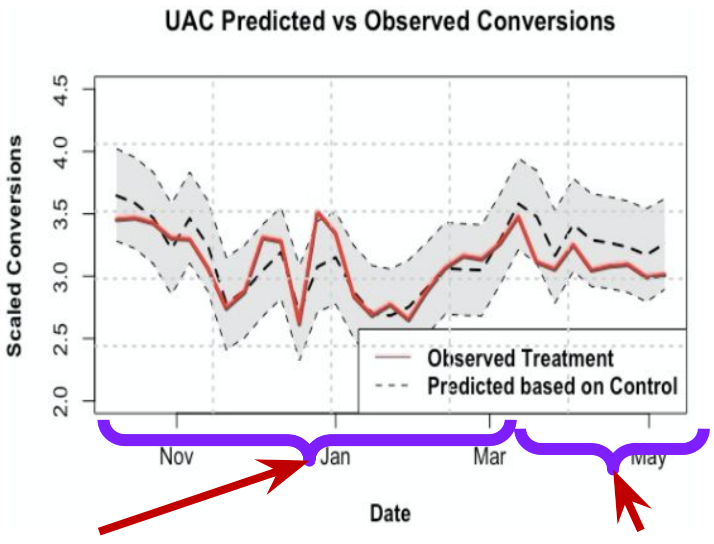
Stabilization, same CPA in both markets 01/15 - 03/12

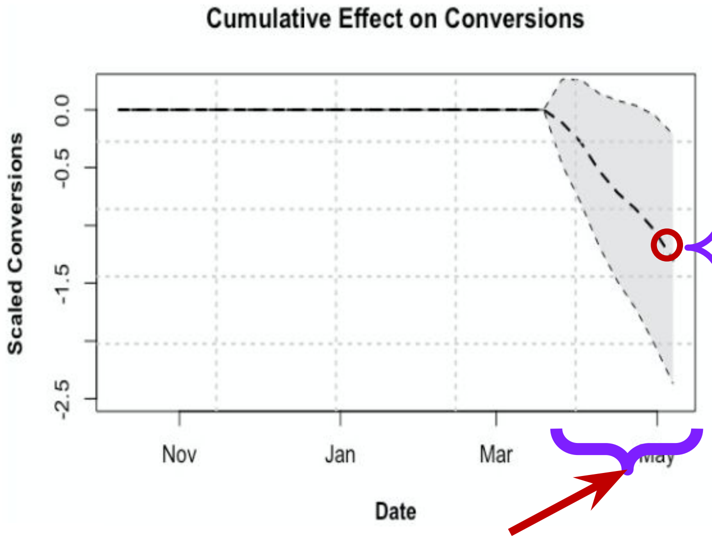
Actual Intervention: Treatment Spend cut 03/19 - 05/13

UAC Incrementality: Effect on Weekly Conversions

Barajas et al. (2020)

Consistently lower predictive (synthetic control) treatment conversions than observed





Conversion Lift:
-6.57%

Spend Lift:
-72.33%

Cost per Incremental Scaled conversion:
39.30 money units

Training period:
09/09 - 03/12

Intervention period.
Predictive vs observed
03/12 - 03/15

Cumulative effect and 95% credible intervals

Limitations and Caveats

Limitations and Caveats

1. Comparisons between aggregate market conversions **require large intervention effects (spend)** since we are **unable to identify users not exposed to the ads** leading to less precision.

2. Rigorously designed experiments provide **valuable data to build channel cost curves of incremental conversions** and to calibrate Media Mix Models for optimal spend allocation

3. **Testing during holidays is noisy and problematic**, which is a big limitation compared to user holdout testing

Thank you!!!

Feedback welcome.

joelbz__AT__amazon.com