

# Causality & Advertising

UCSD MGTA 451-Marketing

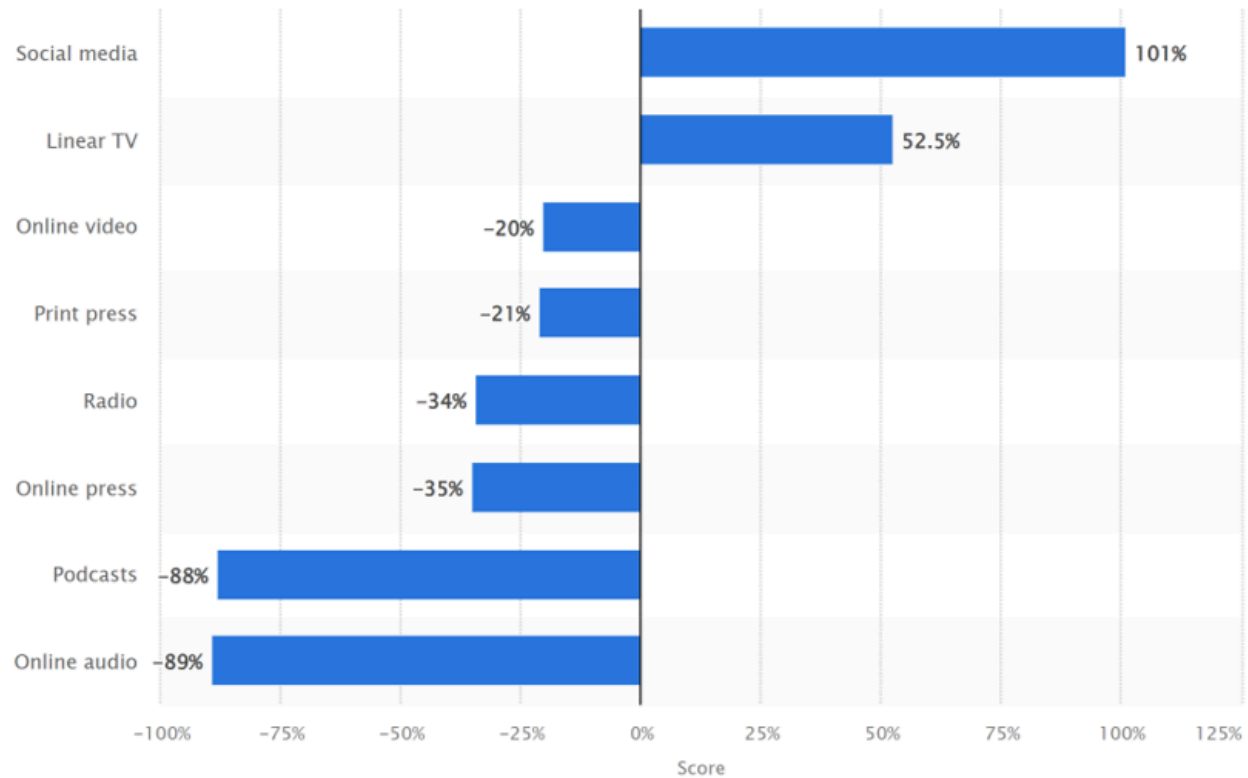
Kenneth C. Wilbur

# Advertising

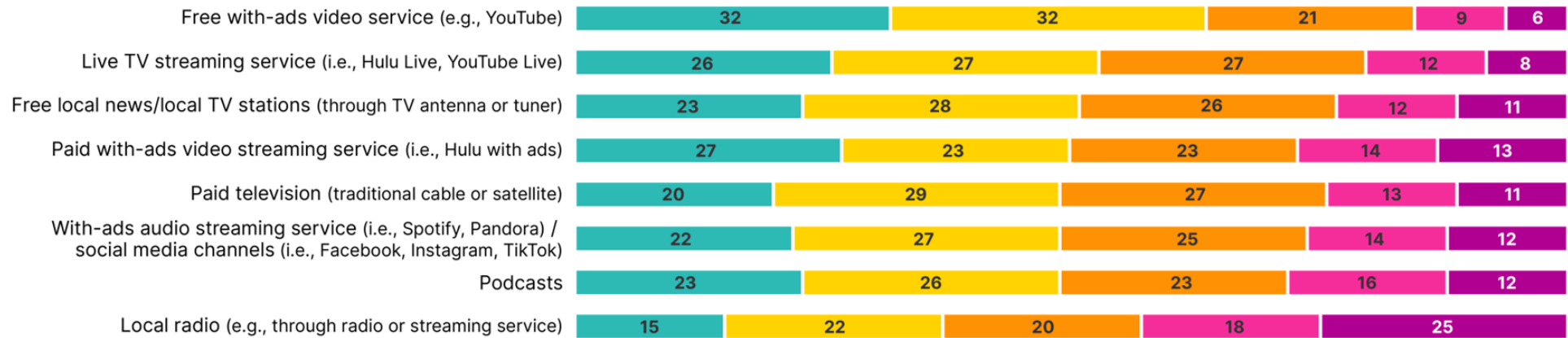
Some introductory and motivating facts

# Difference between advertising spending and time spent with selected media in the United States in 2022

*(index score)*

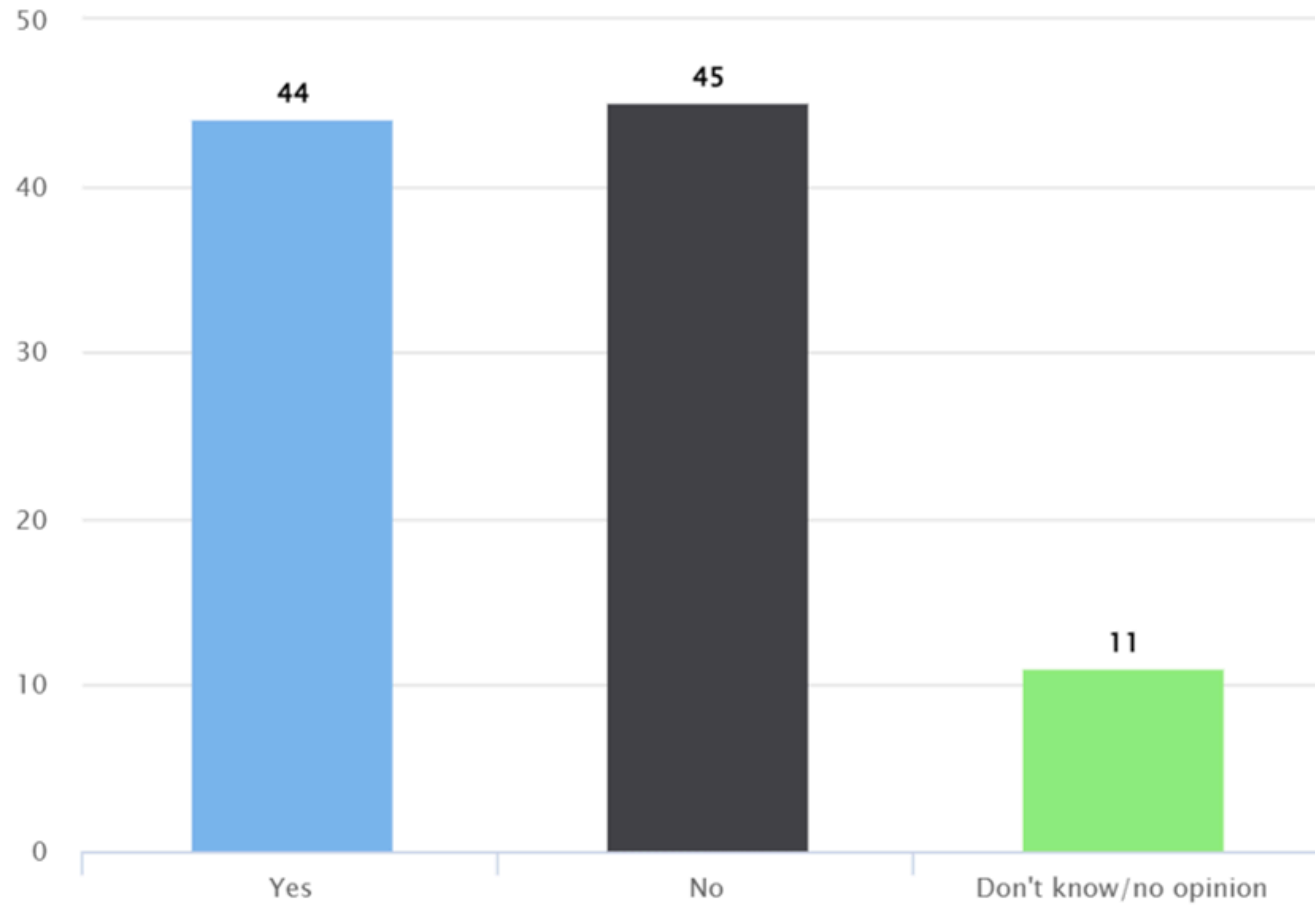


# How often do you intentionally take action to avoid ads on each of the following channels?





## U.S. consumers who purchased products after seeing an internet ad



### 2023 Advertising To Sales Ratios by Industry Sector

Industry Sector	Ad to Sales Ratio %	Ad Growth %	Sales Growth %
Agriculture, Forestry, Fishing	0.40	-17.79	17.64
Mining, Extraction	0.07	13.28	-3.45
Construction	0.35	26.94	0.53
Manufacturing	2.70	10.47	0.69
Transportation, Communications, Utilities	3.60	4.26	5.31
Wholesale Trade	1.27	7.81	0.70
Retail Trade	2.42	11.69	4.09
Finance, Insurance, Real Estate	2.31	-0.05	8.57
Services	4.06	11.27	11.86
<b>All sectors combined</b>	<b>2.83</b>	<b>9.01</b>	<b>4.77</b>

[Advertising Ratios & Budgets](#) is the source for the above data. This detailed report covers over 2,500 companies and 315 industries with fiscal 2023 and 2022 advertising budgets and revenue, 2023 ad-to-sales ratio and ad-to-profit ratio, as well as 2023 annual growth rates in ad spending and sales. Use it to track competition, win new ad agency clients, set and justify ad budgets, sell space and time or plan new media ventures and new products. Includes industry and advertiser ad spending rankings and data on over 350 non-U.S. headquartered companies. Bought by major advertising agencies, media companies, advertisers and libraries. Published May 2024.

Advertising Sales Ratios - SAI Books

- Typical net margin: 8-10% (see [Damodaran](#))
  - So modal firm could increase EBITDA 28-35% by dropping ads:  
 $(8+2.83)/8=1.35$
  - Or could it? What would happen to revenue?

# Toy economics of advertising

- Suppose we pay \$20 to buy 1,000 digital ad OTS. Suppose 3 people click, 1 person buys.
- Ad profit > 0 if transaction margin > \$20
  - But we bought ads for 999 people who didn't buy
- Or, ad profit > 0 if CLV > \$20
  - Long-term mentality justifies increased ad budget
- Or, ad profit > 0 if CLV > \$20 *and* if the customer would not have purchased otherwise
  - This is "incrementality"
  - But how would we know if they would have purchased otherwise?
- Ad effects are subtle—typically, 99.5-99.9% *don't* convert—but ad profit can still be robust
  - Ad profit depends on ad cost, conversions, margin, objective formulation

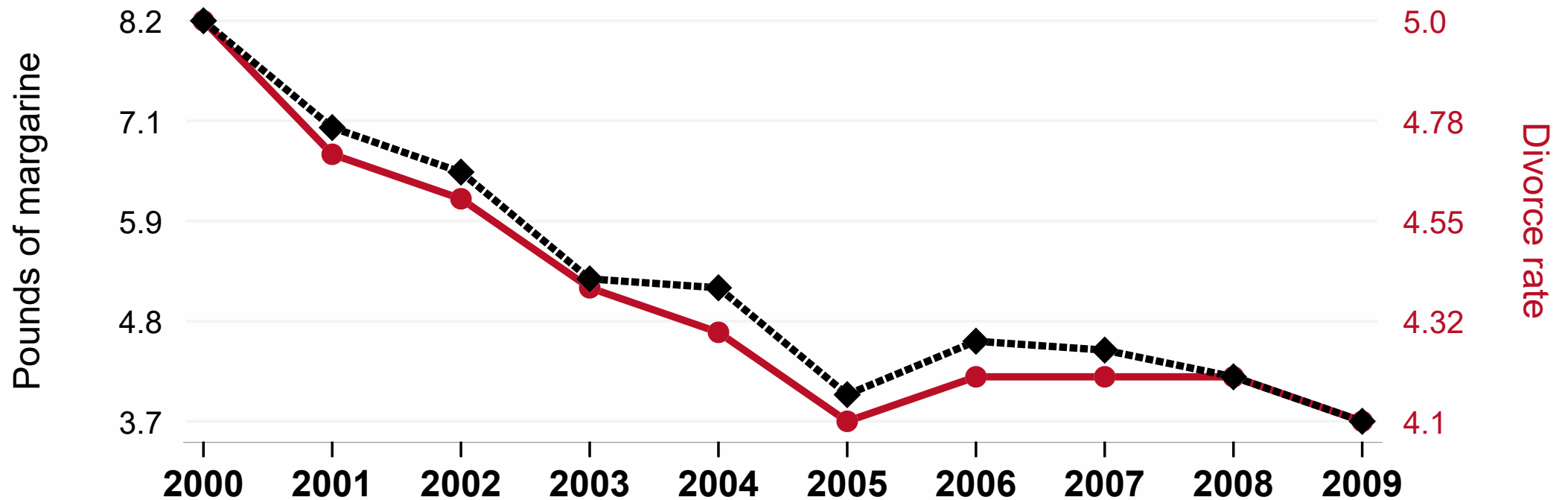
# Causality

Examples, fallacies and motivations

# Per capita consumption of margarine

correlates with

## The divorce rate in Maine



◆ Per capita consumption of margarine in the United States · Source: US Department of Agriculture

● The divorce rate in Maine · Source: CDC National Vital Statistics

2000-2009,  $r=0.993$ ,  $r^2=0.985$ ,  $p<0.01$  · [tylervigen.com/spurious/correlation/5920](http://tylervigen.com/spurious/correlation/5920)

- Suppose 10 outcomes, 1000 predictors,  $N=100,000$  obs
  - Outcomes might include visits, sales, reviews, ...
  - Predictors might include customer attributes, session attributes, ...
- Suppose everything is noise, no true relationships
  - The distribution of the 10,000 correlation coefficients would be Normal, tightly centered around zero
  - A 2-sided test of  $\{\text{corr} == 0\}$  would reject at 95% if  $|r| > .0062$
- We should expect 500 false positives
  - What is a 'false positive' exactly?
- In general, what can we learn from a significant correlation?
  - "These two variables likely move together." Nothing more.

# Classic misleading correlations

- “Lucky socks” and sports wins

- Post hoc fallacy [1] (precedence indicates causality AKA superstition)

- Commuters carrying umbrellas and rain

- Forward-looking behavior

- Kids receiving tutoring and grades

- Reverse causality / selection bias

- Ice cream sales and drowning deaths

- Confounding variables

- Correlations are measurable & usually predictive, but hard to interpret causally

- Correlation-based beliefs are hard to disprove and therefore sticky
  - Correlations that reinforce logical theories are especially sticky
  - Correlation-based beliefs may or may not reflect causal relationships

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE



flowers

358,000,000 RESULTS

**Flowers at 1-800-FLOWERS®** Ads

1800Flowers.com  
Fresh **Flowers & Gifts** at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

**FTD® - Flowers**

www.FTD.com  
**Get Same Day Flowers in Hours!** Buy Now for 25% Off Best Sellers.

**Send Flowers from \$19.99**

www.ProFlowers.com  
**Send Roses, Tulips & Other Flowers. "Best Value"** -Wall Street Journal.  
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

**50% Off All Flowers**

www.BloomsToday.com  
All **Flowers** on the Site are 50% Off. Take Advantage and Buy Today!

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE



flowers

358,000,000 RESULTS

**FTD® - Flowers** Ads

www.FTD.com  
**Get Same Day Flowers in Hours!**  
Buy Now for 25% Off Best Sellers.

**Flowers at 1-800-FLOWERS® | 1800flowers.com**

1800Flowers.com  
Fresh **Flowers & Gifts** at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

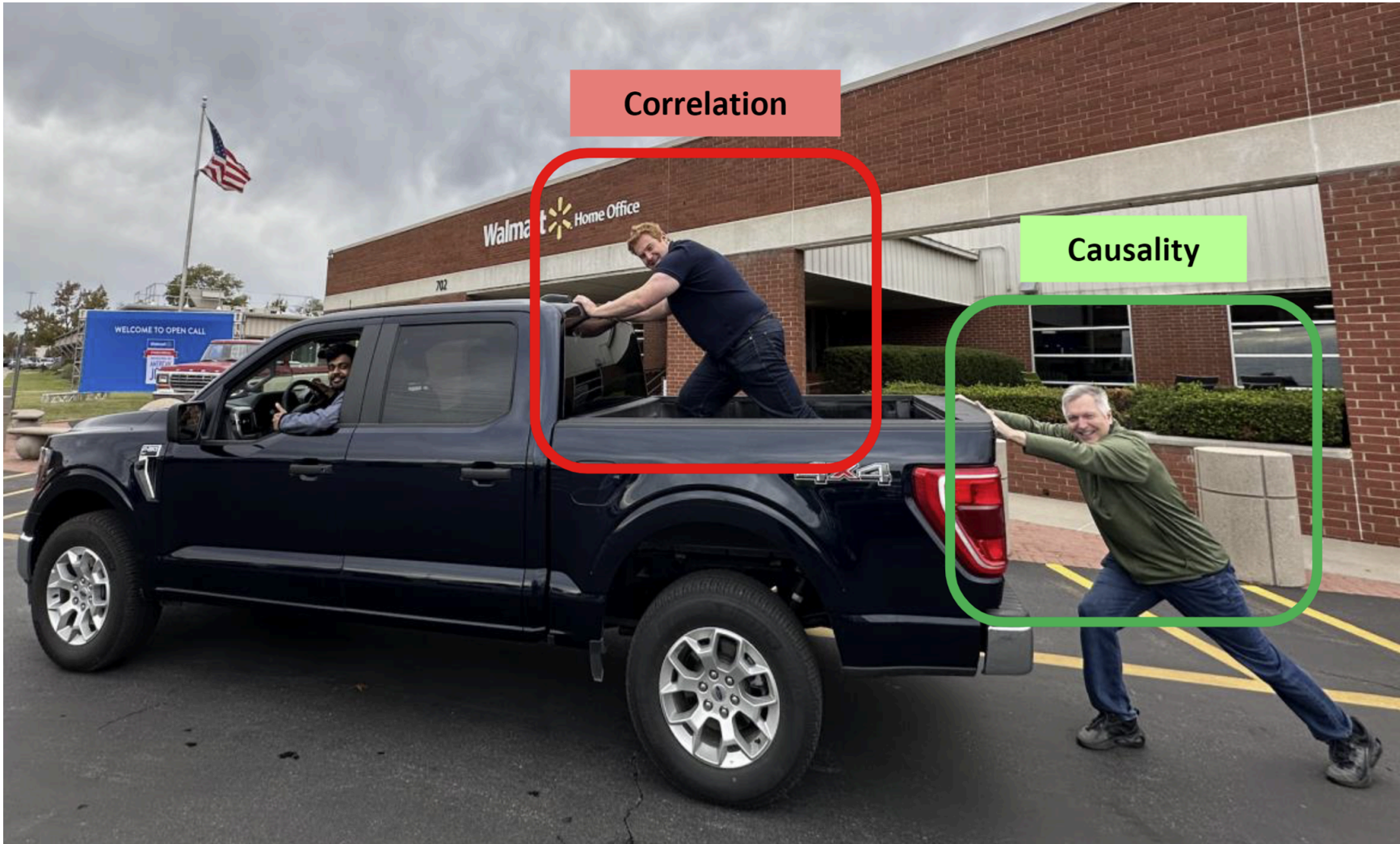
**Send Flowers from \$19.99** **Send Roses, Tulips & Other Flowers**

www.ProFlowers.com  
"Best Value" -Wall Street Journal.  
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

**\$19.99 - Cheap Flowers - Delivery Today By A Local Florist!**

www.FromYouFlowers.com  
Shop Now & Save \$5 Instantly.





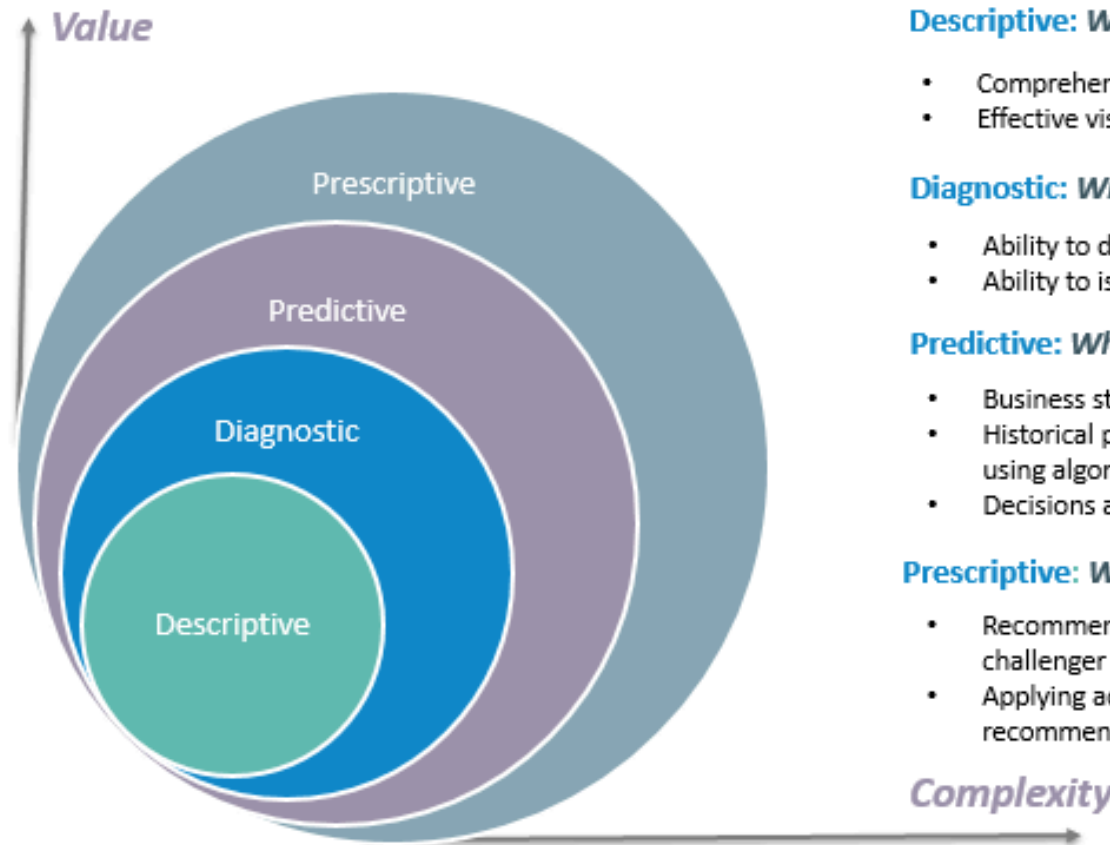
**Correlation**

**Causality**

# Agenda

- Causality
- Experiments, quasi-exp & corr, applied to ads
- Why are correlations used so often?
- Ad/sales modeling frameworks

## 4 types of Data Analytics



### What is the data telling you?

#### **Descriptive:** *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

#### **Diagnostic:** *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

#### **Predictive:** *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

#### **Prescriptive:** *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

# Causal Inference

- Suppose we have a binary “treatment” or “policy” variable  $T_i$  that we can “assign” to person  $i$ 
  - Examples: Advertise, Serve a design, Recommend
  - "Treatment" terminology came from medical literature
- Suppose person  $i$  could have a binary potential “response” or “outcome” variable  $Y_i(T_i)$ 
  - Examples: Visit site, Click product, Add to Cart, Purchase, Rate, Review
  - Looks like the marketing funnel model we saw previously
- Important:  $Y_i$  may depend fully, partially, or not at all on  $T_i$ , and the dependence may be different for different people
  - Person 1 may buy due to an ad; person 2 may stop due to an ad

# Why care?

- We want to maximize profits  $\pi_i(Y_i(T_i), T_i)$
- Suppose  $Y_i = 1$  contributes to revenue; then  $\frac{d\pi_i}{dY_i} > 0$
- Suppose  $T_i = 1$  is costly; then  $\frac{d\pi_i}{dT_i} = \frac{\partial\pi_i}{\partial Y_i} \frac{\partial Y_i}{\partial T_i} + \frac{\partial\pi_i}{\partial T_i}$
- We have to know  $\frac{\partial Y_i}{\partial T_i}$  to optimize  $T_i$  assignments
  - Called the "treatment effect" (TE)
- Profits may decrease if we misallocate  $T_i$

# Fundamental Problem of Causal Inference

- We can only observe **either**  $Y_i(T_i = 1)$  **or**  $Y_i(T_i = 0)$ , but not both, for each person  $i$ 
  - The case we don't observe is called the "counterfactual"
- This is a missing-data problem that we cannot resolve. We only have one reality
  - Models can only compensate for missing data by assumption

# So what can we do?

1. Experiment. Randomize  $T_i$  and estimate  $\frac{\partial Y_i}{\partial T_i}$  as  $\text{avg } Y_i(T_i = 1) - Y_i(T_i = 0)$

- Called the "Average Treatment Effect"
- Creates new data; costs time, money, attention; deceptively difficult to design and then act on

2. Use assumptions & data to estimate a “quasi-experimental” average treatment effect using archival data

- Requires expertise, time, attention; difficult to validate; not always possible

3. Use correlations: Assume past treatments were assigned randomly, use past data to estimate  $\frac{\partial Y_i}{\partial T_i}$

- Easier than 1 or 2; but T is only randomly assigned when we run an experiment, so what exactly are we doing here?

4. Fuhgeddaboutit, go with the vibes, do what we feel

# How much does causality matter?

- How hard should we work?

- Organizational returns or costs of getting it right?
- Data thickness: How likely can we get a good estimate?
- How does empirical approach fit with organizational analytics culture? Will we act on what we learn?
- Individual: promotion, bonus, reputation, career; Will credit be stolen or blame be shared?
- Accountability: Will ex-post attributions verify findings? Will results threaten or complement rival teams/execs?

- Analytics culture starts at the top



# Ad/sales example: Experiment

## 1. Randomly assign ads to customer groups on a platform; measure sales in each group

- Often called "incrementality" in ad/sales context
- Pros: AB testing is easy to understand, easy to implement, easy to validate
- Cons: Can we trust the platform's "black box"? Will we get the data and all available insights? Could platform knowledge affect future ad costs?

## 2. Randomize over messages within a campaign

## 3. Randomize over times, places, consumer segments

## 4. Randomize over budgets and bids

## 5. Randomize over platforms, publishers, behavioral targets, etc., to compare RoAS across options

RoAS = Return on Ad Spend. RoAS defined as  $\text{Sales} / \text{AdSpend}$  or  $(\text{Sales} - \text{AdSpend}) / \text{AdSpend}$

# Experimental necessary conditions

## 1. Stable Unit Treatment Value Assumption (SUTVA)

- Treatments do not vary across units within a treatment group
- One unit's treatment does not change other units' potential outcomes, i.e. treatments in one group do not affect outcomes in another group
- Often violated when treated units interact on a platform
- Violations called "interference"; remedies usually start with cluster randomization

## 2. Observability

- Non-attrition, i.e. unit outcomes remain observable

## 3. Compliance

- Treatments assigned are treatments received
- We have partial remedies when noncompliance is directly observed

## 4. Statistical Independence

- Random assignment of treatments to units

# 2. Ad/sales example: Experiment

## Key issues for any experimental design:

- Always run A:A test first. Validate the infrastructure before trusting a result
- Can we agree on the opportunity cost of the experiment? "Priors"
- How will we act on the (uncertain) findings? Have to decide before we design. We don't want "science fair projects"
- Simple example: Suppose we estimate RoAS at 1.5 with c.i. [1.45, 1.55]. Or, suppose we estimate RoAS at 1.5 with c.i. [-1.1, 4.1]. How will we act?

# Quasi-experiments Vocab

**Model:** Mathematical relationship between variables that simplifies reality, eg  $y=xb+e$

**Identification strategy:** Set of assumptions that isolate a causal effect  $\frac{\partial Y_i}{\partial T_i}$  from other factors that may influence  $Y_i$

- A system to compare apples with apples, not apples with oranges

We say we “identify” the causal effect if we have an identification strategy that reliably distinguishes  $\frac{\partial Y_i}{\partial T_i}$  from possibly correlated unobserved factors that also influence  $Y_i$

If you estimate a model without an identification strategy, you should interpret the results as correlational

- This is widely, widely misunderstood

You can have an identification strategy without a model, e.g.

$$\text{avg } Y_i(T_i = 1) - Y_i(T_i = 0)$$

Usually you want both. Models help with quantifying uncertainty and estimating treatment effects by controlling for relevant observables

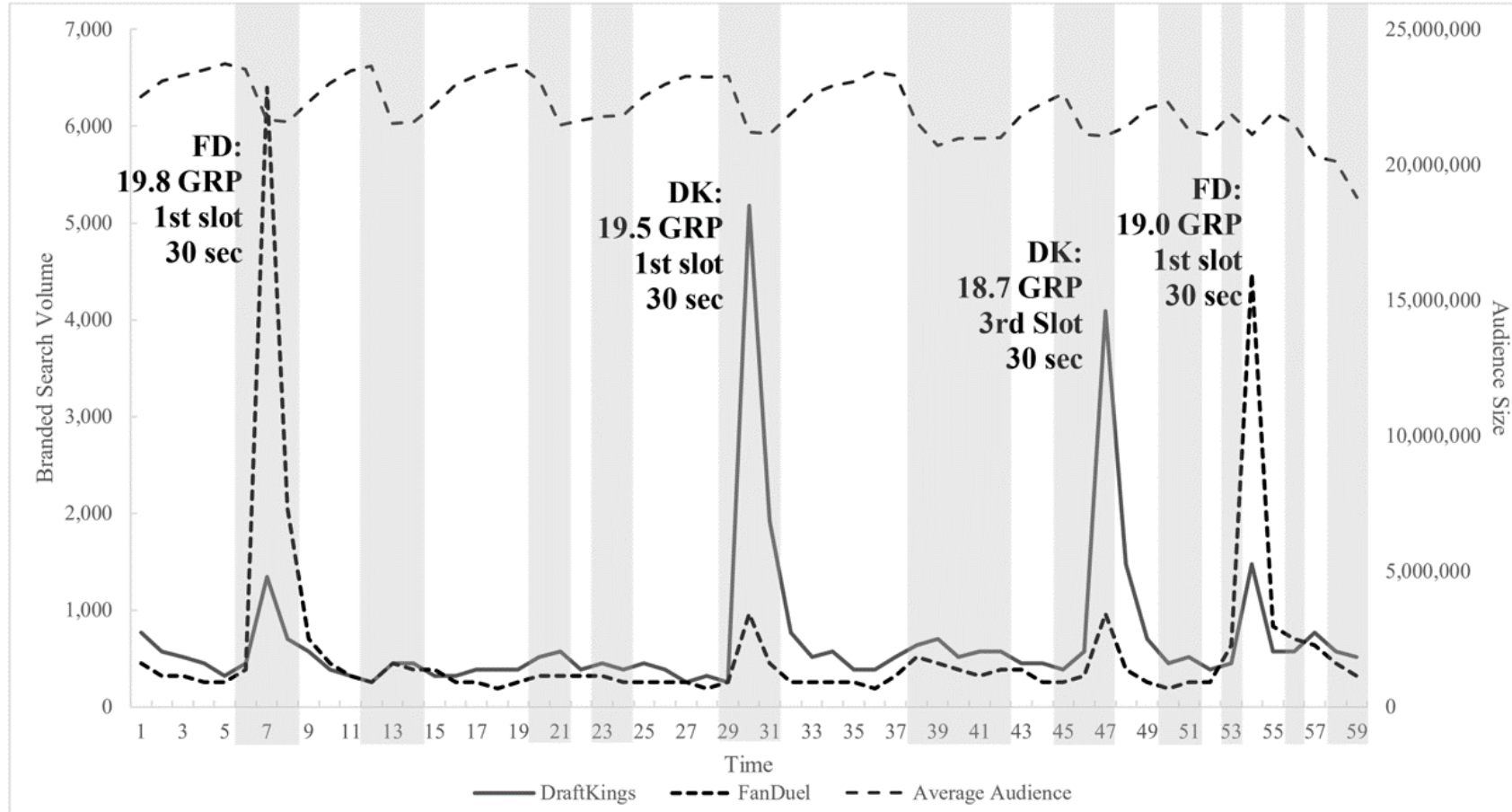
## 2. Ad/sales: Quasi-experiments

Goal: Find a “natural experiment” in which  $T_i$  is “as if” randomly assigned, to identify  $\frac{\partial T_i}{\partial Y_i}$

### Possibilities:

- Firm starts, stops or pulses advertising without changing other variables, especially when staggered across times or geos
- Competitor starts, stops or pulses advertising
- Discontinuous changes in ad copy
- Exogenous changes in ad prices, availability or targeting (e.g., biannual elections)
- Exogenous changes in addressable market, website visitors, or other factors

# DFS TV ad effects on Google Search



# Ad/sales: Quasi-experiments (2)

Or, construct a “quasi-control group”

- Customers or markets with similar demand trends where the firm never advertised
- Competitors or complementors with similar demand trends that don't advertise

Helpful identification strategies: Difference in differences, Synthetic control, Regression discontinuity, Matching, Instrumental variables

In each case, we try to predict our missing counterfactual data, then estimate the causal effect as observed outcomes minus predicted outcomes

# 3. Ad/sales example: Correlational

Just get historical data on  $Y_i$  and  $T_i$  and run a regression

Most people use OLS, but Google's CausalImpact R package is also popular

The implicit assumption is that past ads were allocated randomly, i.e. correlation==causality

"Better to be vaguely right than precisely wrong"  
But are we the guy in the truck bed?

In truth, past ads were only random if we ran an experiment



# Strongest args for $\text{corr}(\text{ad}, \text{sales})$

$\text{Corr}(\text{ad}, \text{sales})$  should contain signal

- If ads cause sales, then  $\text{corr}(\text{ad}, \text{sales}) > 0$  (probably) (we assume)

Some products/channels just don't sell without ads

- E.g., Direct response TV ads for telephone response
- Career professionals say advertised phone #s get 0 calls without TV ads, so we know the counterfactual
- Then they get 1-5 calls per 1k viewers, lasting up to ~30 minutes
- What are some digital analogues to this?

However, this argument gets pushed too far

- For example, when search advertisers disregard organic link clicks when calculating search ad click profits
- Notice the converse:  $\text{corr}(\text{ad}, \text{sales}) > 0$  does not imply a causal effect of ads on sales

# Problem 1 with $\text{corr}(\text{ad}, \text{sales})$

Advertisers try to optimize ad campaign decisions

E.g. surfboards in coastal cities, not landlocked cities

If ad optimization increases ad response, then  $\text{corr}(\text{ad}, \text{sales})$  will confound actual ad effect with ad optimization effect

More ads in san diego, more surfboard sales in san diego

$\text{Corr}(\text{ad}, \text{sales})$  usually overestimates the causal effect, encourages overadvertising

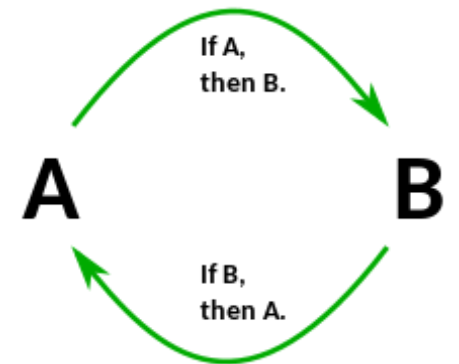
Many, many firms basically do this

It's ironic when firms that don't run experiments assume that past ads were randomized

# Problem 2 with $\text{corr}(\text{ad}, \text{sales})$

- How do most advertisers set ad budgets? Top 2 ways:
  1. Percentage of sales method, e.g. 3% or 6%
  2. Competitive parity
  3. ...others...

Do you see the problem here?



# Problem 3 with corr(ad,sales)

- Leaves marketers powerless vs ~~big~~ colossal ad platforms
- Google and Meta withhold data and obfuscate algorithms
  - How many ad placements are incremental?
  - How many ad placements target likely converters?
  - How can advertisers react to adversarial ad pricing?
  - How can advertisers evaluate brand safety, targeting, context?
- Have ad platforms ever left ad budget unspent?
  - Would you, if you were them?
  - If not, why not? What does that imply about incrementality?
- To balance platform power, know your ad profits, vote with your feet

# U.S. v Google (2024, search case)

UNITED STATES DISTRICT COURT  
FOR THE DISTRICT OF COLUMBIA

UNITED STATES OF AMERICA et al.,

Plaintiffs,

v.

GOOGLE LLC,

Defendant.

Case No. 20-cv-3010 (APM)

263. When it made pricing changes, Google took care to avoid blowback from advertisers. For instance, records show that Google had concerns about the impact of transparency on their efforts to increase prices. See UPX507 at .015 (“Worry that if we tell advertisers they will be impacted, they will attempt to game us and convince us to abandon the experiment. . . . But, if influence our decision at all.”); UPX519 at .003 (“A sudden step function might create adverse reaction.”).

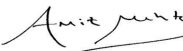
264. Google therefore endeavored to raise prices incrementally, so that advertisers would view price increases as within the ordinary price fluctuations, or “noise,” generated by the auctions. See, e.g., UPX507 at .023 (describing a 10% CPC increase as “safe” because it is “within usual WoW noise”); UPX519 at .003 (acknowledging that advertisers would notice a 15% price increase, but “this change is to [be] put in perspective with CPC noise,” that is, “50% of advertisers seeing 10%+ WoW CPC changes”); *id.* (comment stating that 15% is “probably an acceptable level of change (from a perception point of view) because these are magnitudes of fluctuations they are used to see[ing]”).

265. With respect to format pricing, one Google document states: “A progressive ramp up leaves time to internalize prices and adjust bids appropriately[.]” UPX519 at .003; UPX509 at 870 (stating that “[i]ncremental launches and monitoring should help us manage” the risk that price increases would lead advertisers to “lower[] their bids or modify[] other settings . . . to get back to a given ROI, leading to less revenue for Google than the initial impact hinted to”). Similarly, in 2020, Google raised prices on navigational queries using multiple knobs and recognized that it was “[o]bviously a very large change that we don’t intend to roll out at once,” instead planning a “[s]low 18 months rollout” to “[l]eave[] time for advertiser[s] to respond rationally[.]” UPX503 at 034; *id.* at 038 (“A slow roll ensures we don’t shock the system, gives time for advertisers to respond and us to monitor changes and stop early if needed.”); see also, e.g., UPX505 at 312 (prior to implementing squashing, concluding that “[a]dvertisers should perceive AdWords as a consistent system, and not be subject to constant large impacts due to Google changes,” in part to “improve[] advertiser stickiness”); UPX506 at .018 (Momiji slide deck: “Unlikely that advertisers will notice by themselves and respond. However, a bad press cycle could put us in jeopardy.”).

266. Google’s incremental pricing approach was successful. In 2018 and 2019, Google conducted ROI Perception Interviews, which raised no red flags about advertisers’ attitudes as to ad spending on Google. See generally DX187; DX119. While advertisers could tell that prices were increasing, they did not understand those changes to be Google’s fault. Google’s studies revealed that advertisers facing CPC changes “dominantly attribute[d] these shifts to themselves, competition[,] and seasonality (85%)—not Google.” UPX1054 at 061; see also UPX737 at 464 (“They often attribute these changes to things in the world or what they’ve done, not just things happening on the backend[.]”).

## CONCLUSION

For the foregoing reasons, the court concludes that Google has violated Section 2 of the Sherman Act by maintaining its monopoly in two product markets in the United States—general search services and general text advertising—through its exclusive distribution agreements. The court thus holds that Google is liable as to Counts I and III of the U.S. Plaintiffs’ Amended Complaint, Am. Compl. ¶¶ 173–179, 187–193. To the extent that Counts I and III of the Plaintiff States’ Complaint are co-extensive with the U.S. Plaintiffs’ Counts I and III, the court finds Google liable. *Colorado Compl.* ¶¶ 212–218, 226–232.

  
Amit P. Mehta  
United States District Court

# Does $\text{Corr}(\text{ad}, \text{sales})$ work?

[Home](#) > [Marketing Science](#) > [Vol. 42, No. 4](#) >

## Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement

Brett R. Gordon , Robert Moakler, Florian Zettelmeyer

Despite their popularity, randomized controlled trials (RCTs) are not always available for the purposes of advertising measurement. Non-experimental data are thus required. However, Facebook and other ad platforms use complex and evolving processes to select ads for users. Therefore, successful non-experimental approaches need to “undo” this selection. **We analyze 663 large-scale experiments at Facebook to investigate whether this is possible with the data typically logged at large ad platforms.** With access to over 5,000 user-level features, these data are richer than what most advertisers or their measurement partners can access. We investigate how accurately two non-experimental methods—double/debiased machine learning (DML) and stratified propensity score matching (SPSM)—can recover the experimental effects. Although DML performs better than SPSM, neither method performs well, even using flexible deep learning models to implement the propensity and outcome models. The median RCT lifts are 29%, 18%, and 5% for the upper, middle, and lower funnel outcomes, respectively. Using DML (SPSM), the median lift by funnel is 83% (173%), 58% (176%), and 24% (64%), respectively, indicating significant relative measurement errors. We further characterize the circumstances under which each method performs comparatively better. Overall, despite having access to large-scale experiments and rich user-level data, we are unable to reliably estimate an ad campaign’s causal effect.

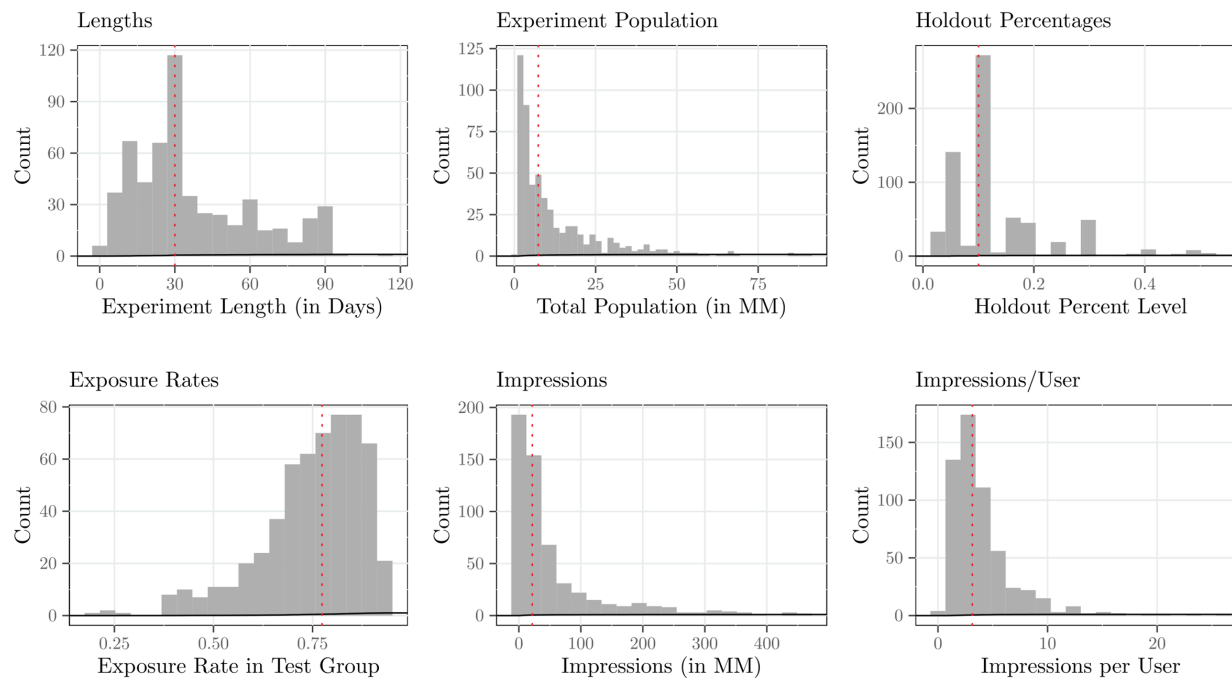
### 3.1. Experiment Selection

The advertising experiments analyzed in this paper were chosen to be representative of large-scale advertising experiments run in the United States on the Facebook ad platform. Ads in these experiments can appear on Facebook, Instagram, or the Facebook Audience Network. These experiments cover a wide range of verticals, targeting choices, campaign objectives, conversion outcomes, sample sizes, and test/control splits. The experiments we analyze are a random subset from the set of experiments started between November 1, 2019, and March 1, 2020, and had at least one million users in the test group.<sup>13</sup> For each experiment, we selected all outcomes with at least 5,000 conversions in the test group.<sup>14</sup>

As Figure 1 shows, experiments vary widely by length, by population size, by the fraction of users in the holdout group, by the rate at which targeted consumers were exposed, and the number of impressions. The median of experiment length is 30 days and includes 7,372,103 users across test and control groups. The median holdout percentage places 90% of users in the test group and 10% in the control group. For those in the test group, the median exposure percentage was 77%, while 23% of users were never exposed. The median of ad impressions per experiment is 22,115,390. Overall, our data set represents approximately 7.9 billion user-experiment observations with 38.4 billion ad impressions.

Most experiments measure several different conversion outcomes, such as purchases, page views, downloads, etc. We treat all such outcomes as binary events, that is, a user either viewed a particular web page or they did not. Industry practitioners classify conversion outcomes by whether they occur earlier or later in a hypothetical purchase funnel. For example, page views occur early in the purchase funnel, adding items to a cart occurs later, and purchase occurs last. Our 663 experiments capture a total of 1,673 conversion events, measuring different conversion outcomes. Henceforth, we will refer to each experiment-conversion event as an “RCT.” We classify RCTs into “Upper Funnel” (601), “Mid Funnel” (475), and “Lower Funnel” (597). As we describe in Section 2.1, outcomes are measured using “pixels,” which advertisers choose to place on their

Figure 1. (Color online) Distribution of Experiment Characteristics



Note. Histogram excludes the top 1% of experiment population size; Dashed line shows median.

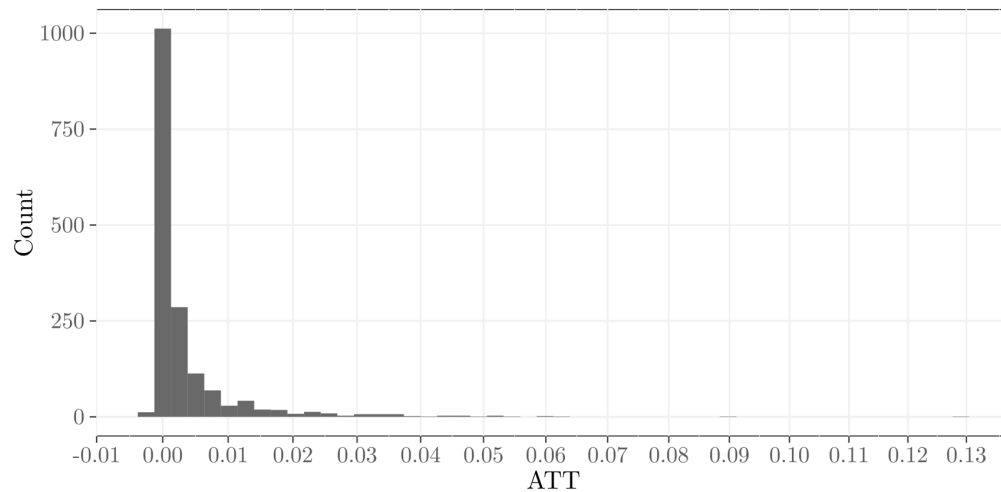
Table 1. Distribution of Conversion Events

Pixel name	Funnel position	<i>N</i>	Percent
view_content	Upper	410	24.5
search	Upper	121	7.2
lead_referral	Upper	70	4.2
add_to_cart	Mid	266	15.9
initiate_checkout	Mid	138	8.2
add_to_wishlist	Mid	34	2
add_payment_info	Mid	21	1.3
tutorial_completion	Mid	16	1
purchase	Lower	409	24.4
app_activate_launch	Lower	97	5.8
complete_registration	Lower	91	5.4

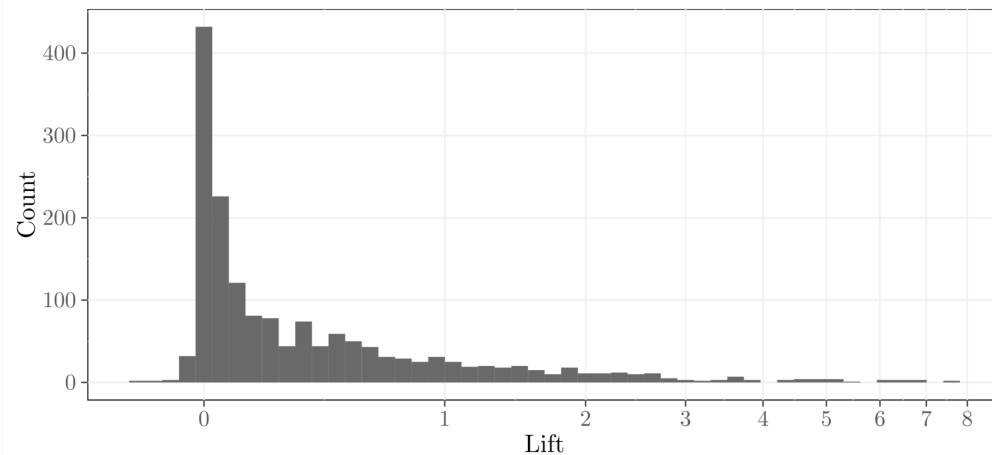
Table 2. Conversion Events by Industry Vertical

Industry vertical	<i>N</i>	Percent
E-commerce	504	30.1
Retail	377	22.5
Financial services/travel	322	19.2
Entertainment/media	145	8.7
Tech/telecom	124	7.4
Consumer packaged goods	105	6.3
Other	96	5.7

**Figure 2.** ATTs Across All RCTs



**Figure 3.** Lifts Across All RCTs

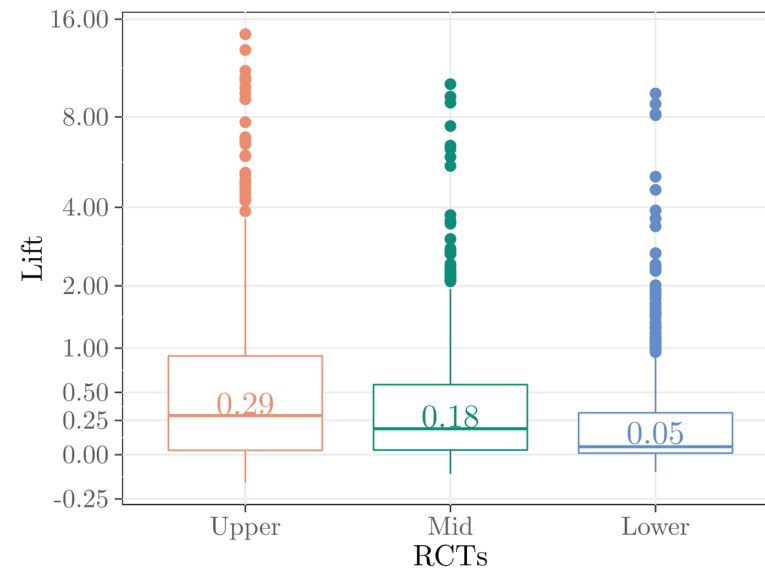


*Note.* Excludes the top 1% of lifts.

However, ATTs are difficult to interpret since they contain no information on whether the ATT is “small” or “large.” Hence, to more easily interpret outcomes across RCTs, we report most results in terms of *lift*, the incremental conversion rate among treated users expressed as a percentage,

$$\begin{aligned} \ell &= \frac{\text{Conversion rate due to ads in the treated group}}{\text{Conversion rate of the treated group if they had not been treated}} \\ &= \frac{\tau}{\mathbb{E}[Y | Z = 1, W = 1]} \end{aligned} \quad (4)$$

**Figure 4.** (Color online) Lifts by Purchase Funnel Position





**5.1.1. Stratified Propensity Score Matching (SPSM).** The first method we use to address the nonrandomness of treatment is propensity score matching (Dehejia and Wahba 2002, Stuart 2010). The propensity score,  $e(X_i)$ , is the conditional probability of treatment given features  $X_i$ ,

$$e(X_i) \equiv \Pr(W_i = 1 \mid X_i = x). \quad (11)$$

Under strong ignorability, Rosenbaum and Rubin (1983) establish that treatment assignment and the potential outcomes are independent, conditional on the propensity score,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid e(X_i). \quad (12)$$

This result shows that the bias from selection can be eliminated by adjusting for the propensity score.

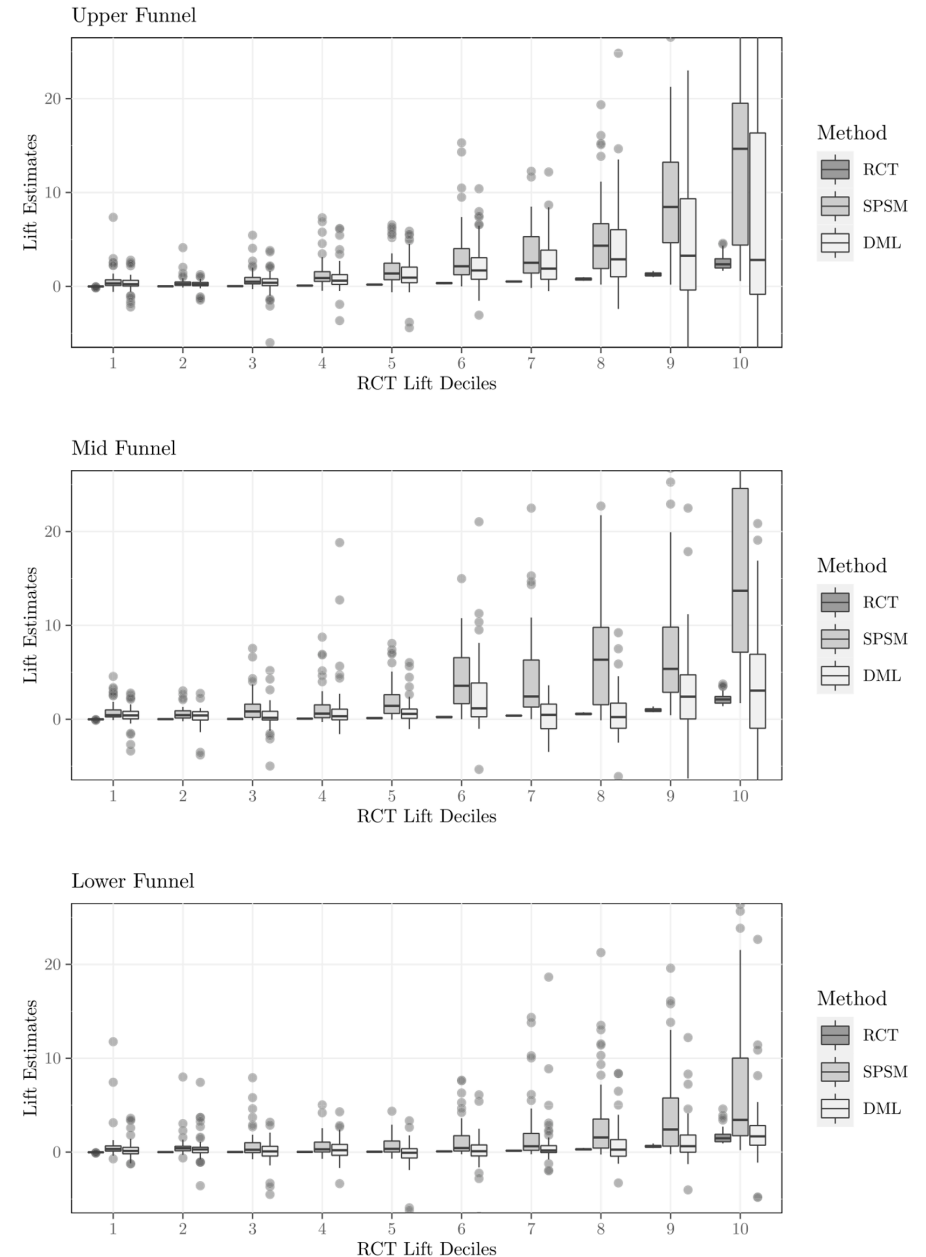
In standard propensity score matching, we find the one (or more) unexposed users with the closest propensity score to each exposed user to estimate the treatment effect. Since this is very computationally burdensome, instead, we stratify on the propensity score: After estimating the propensity score,  $\hat{e}(X_i)$ , we divide the sample into strata such that within each stratum, the estimated propensity scores are approximately constant. This method, known as stratified propensity score matching (SPSM), scales well and achieves good feature balance without an over-reliance on extrapolation (Imbens and Rubin 2015).

**5.1.2. Double/Debiased Machine Learning (DML).** In the past few years, the machine learning community has made vast improvements to predictive modeling procedures with new statistical methods and advances in computational hardware. Given the focus of these models on making accurate predictions, they are trained on data sets for which the true answer is known for a set of records and are then applied to new, unseen data. However, in causal inference settings, where the goal is not simply predictive power and where we will never observe true outcomes for any individual record, a direct application of machine learning methods to estimate causal effects can lead to invalid, biased, results.

In recent years, new work had aimed to combine the advantages of machine learning with the causal inference goals of traditional econometrics. Specifically, new literature has addressed the main reasons why predictive models may struggle with causal inference, namely the bias that arises from regularization and overfitting. The double/debiased machine learning (DML) approach introduced by Chernozhukov et al. (2018) corrects for both of these sources of bias by using orthogonalization to account for the bias introduced by regularization and by implementing cross-fitting to remove bias introduced by overfitting. Double machine learning methods build on common econometric approaches by combining the benefits of cutting-edge machine learning with causal inference methods such as propensity score matching.

The models and data we use surpass what individual advertisers are able to use for ad measurement and represent close to the peak of what third-party measurement partners and large advertising platforms currently employ. Nonetheless, despite the quality of the data available and the flexibility of the models employed, we found these were inadequate to consistently control for the selection effects induced by the advertising platform.

Figure 10. Comparison of RCT Lifts with Lifts Estimated using SPSM and DML



Note. Figures excludes the top 1% lifts for each position in the purchase funnel.

# Do ad experiments work?

## **Experimentation and Performance in Advertising: An Observational Survey of Firm Practices on Facebook<sup>1</sup>**

Julian Runge, Steve Geinitz & Simon Ejdemyr  
*(Facebook, Marketing Science Research)*

---

### **ABSTRACT**

It is widely assumed that firms experiment with their online advertising to identify more profitable approaches to then increase their investment in more profitable advertising, increasing their overall performance. Generalizable evidence on the actual use of such experiment-based learning by firms is sparse. The study herein addresses this shortcoming – detailing the extent to which large advertisers are utilizing experimentation along with evidence on the benefits of doing so. The findings are gleaned from firms’ marketing and experimentation practices on a large online advertising platform and indicate that, while experimentation is utilized by some, adoption is far from perfect. Among the few firms making use of experiments, even fewer invest a significant share of their advertising spend in experimentation. This finding is surprising in light of broadly assumed regular experimentation by firms. Experimenting firms further experience higher concurrent and subsequent performance, suggesting that leading firms indeed successfully use experiment-based learning to improve their advertising policies – and that many firms may fall short of their potential by not (yet) using experiments in advertising.

# Do ad experiments work?

<i>Measurement frame</i>	<i>Prior to outcome year</i>		<i>During outcome year</i>				
	<b>Number of firms</b>	<b>Average years on platform</b>	<b>Avg. adv. spend</b> (as % of e-comm. vertical)	<b>Avg. distinct ad campaigns</b>	<b>Avg. distinct ads</b>	<b>Share of spend on purchase objectives</b>	<b>Share of firms running experiments</b>
E-commerce (sorted by overall adv. spend)	971	3.7	100.0%	10,158	25,281	66.0%	22.2%
Industry 2	684	3.2	72.7%	3,368	9,499	22.1%	6.7%
Industry 3	534	4.1	90.7%	5,352	16,470	34.5%	27.9%
Industry 4	1,070	1.7	42.4%	4,589	15,139	32.0%	8.9%
Industry 5	498	3.5	71.3%	10,433	33,934	33.8%	10.4%
Industry 6	618	3.6	41.5%	3,632	14,986	36.2%	7.6%
Industry 7	313	3.4	79.1%	52,398	83,342	38.5%	16.9%
Industry 8	289	2.9	65.6%	5,137	16,377	40.8%	4.2%
Industry 9	316	4.3	58.4%	3,752	17,542	41.4%	29.4%
Industry 10	128	4.7	83.4%	6,427	19,808	33.3%	26.6%
Industry 11	245	3.3	38.9%	3,368	15,388	45.0%	8.6%
Industry 12	154	3.9	44.0%	5,531	16,462	52.9%	10.4%
Industry 13	418	2.7	10.2%	920	3,638	35.9%	1.4%
Industry 14	291	2.9	13.3%	2,286	5,909	41.8%	2.7%

# Do ad experiments work?

To investigate possible associations between the use of experimentation and performance, we study purchase conversions which closely mirror the economic success of advertising (Jankowski et al. 2016). We propose three measures:

- Purchase conversions, as obtained from last-click attribution (Li et al. 2016), per 1,000 USD of advertising spend;
- Incremental purchase conversions per 1,000 USD (experiments), as obtained from experiments and hence only observed for advertisers who run experiments;
- Incremental purchase conversions per 1,000 USD (DDA), as obtained from a data-driven attribution (DDA) model that provides estimates of incremental conversions for all campaigns, including the ones without holdout conditions, and all advertisers, including those who have not run any experiments.

TABLE 3: Results of linear regression for the three performance outcomes and the two model specifications in the e-commerce vertical; p-values in brackets, \* significant at 10%-, \*\* at 5%-, \*\*\* at 1%-level. For confidentiality reasons only qualitative results are shown for meta variables R-squared of random forest regressor is included to assess increase in explanatory ability when allowing non-linearities and additional model complexity.

Outcome measure / dep. var.	During outcome year (spec. 1)			Prior to outcome year (spec. 2)		
	Last-click conv.	Exper. incr. conv.	DDA incr. conv.	Last-click conv.	Exper. incr. conv.	DDA incr. conv.
Years active on platform	+ ** (.0166)	- (.3973)	+ *** (.0002)	+ ** (.0134)	- (.5615)	+ *** (.0002)
Years managed	+ *** (.0000)	+ * (.0581)	+ *** (.0000)	+ *** (.0000)	+ ** (.0233)	+ *** (.0000)
All-time advertising spend	- (.8528)	+ (.8890)	- (.8681)	- (.8342)	+ (.8815)	- (.8454)
Accounts used	+ (.8797)	- (.8546)	+ (.5600)	+ (.8225)	- (.6669)	+ (.5138)
Account admins	- (.6250)	- (.8804)	- (.5209)	- (.5213)	- (.9707)	- (.4299)
Advertising objectives used	+ (.2967)	+ (.5759)	+ (.3706)	+ (.3160)	+ (.6226)	+ (.3950)
Experimentation adoption	.5289 (.2551)	.1746 (.7551)	.7296 * (.0525)	.3361 (.5323)	-.9384 (.3152)	.5802 (.1831)
Number of experiments	.019 * (.0737)	.0067 (.4308)	.02 ** (.0224)	.033 (.1080)	.0217 (.2015)	.031 * (.0621)
Intercept	+ *** (.0000)	+ *** (.0000)	+ *** (.0003)	+ *** (.0000)	+ *** (.0002)	+ *** (.0003)
R-squared	.0929	.0549	.1348	.0921	.0652	.1329
R-sq. (RF)	.1797	.0687	.1510	.1713	.1181	.1579
N	776	216	776	776	131	776

Ironic note: Results are correlational

# Why are some teams OK with $\text{corr}(\text{ad}, \text{sales})$ ?

## 1. Some worry that if ads go to zero $\rightarrow$ sales go to zero

- For small firms or new products, this may be good logic
- Downside of lost sales may exceed downside of foregone profits
- However, claim may imply a customer satisfaction problem. Happy customers usually share their experiences with others. If you really believe this, try a referral program
- Plus, we can run experiments without setting ads to zero, e.g. weight tests

## 2. Some firms assume that correlations indicate direction of causal results

- The guy in the truck bed is pushing forwards right?
- Biased estimates might lead to unbiased decisions
- But direction is only part of the picture; what about effect size?

# Why are some teams OK with $\text{corr}(\text{ad}, \text{sales})$ ?

## 3. CFO and CMO negotiate ad budget

- CFO asks for proof that ads work
- CMO asks ad agencies, platforms & marketing team for proof
- CMO sends proof to CFO ; We all carry on

## 4. Few rigorous analytics cultures or ex-post checks

- In some cultures, ex-post checks can get personal

## 5. Estimating causal effects of ads can be pretty difficult

- Many firms lack design expertise, discipline, execution skill
- Ad/sales tests may be statistically inconclusive, especially if small
- Tests are often designed without subsequent actions in mind, then fail to inform future decisions ("science fair projects")

# Why are some teams OK with $\text{corr}(\text{ad}, \text{sales})$ ?

## 6. Platforms often provide correlational ad/sales estimates

- Which is larger, correlational or experimental ad effect estimates?
- Which one would most client marketers prefer?
- Platform estimates are typically "black box" without neutral auditors
- Sometimes platforms respond to marketing executive demand for good numbers
- "Nobody ever got fired for buying [famous platform brand here]"

## 7. Historically, agencies usually estimated RoAS

- Agency compensation usually relies on spending, not incremental sales
- Principal/agent problems are common
- Many marketing executives start at ad agencies
- "Advertising attribution" is all about maximizing credit to ads
- These days, more marketers have in-house agencies, and split work
- Should adFX team report to CFO or CMO?

# What is incrementality?

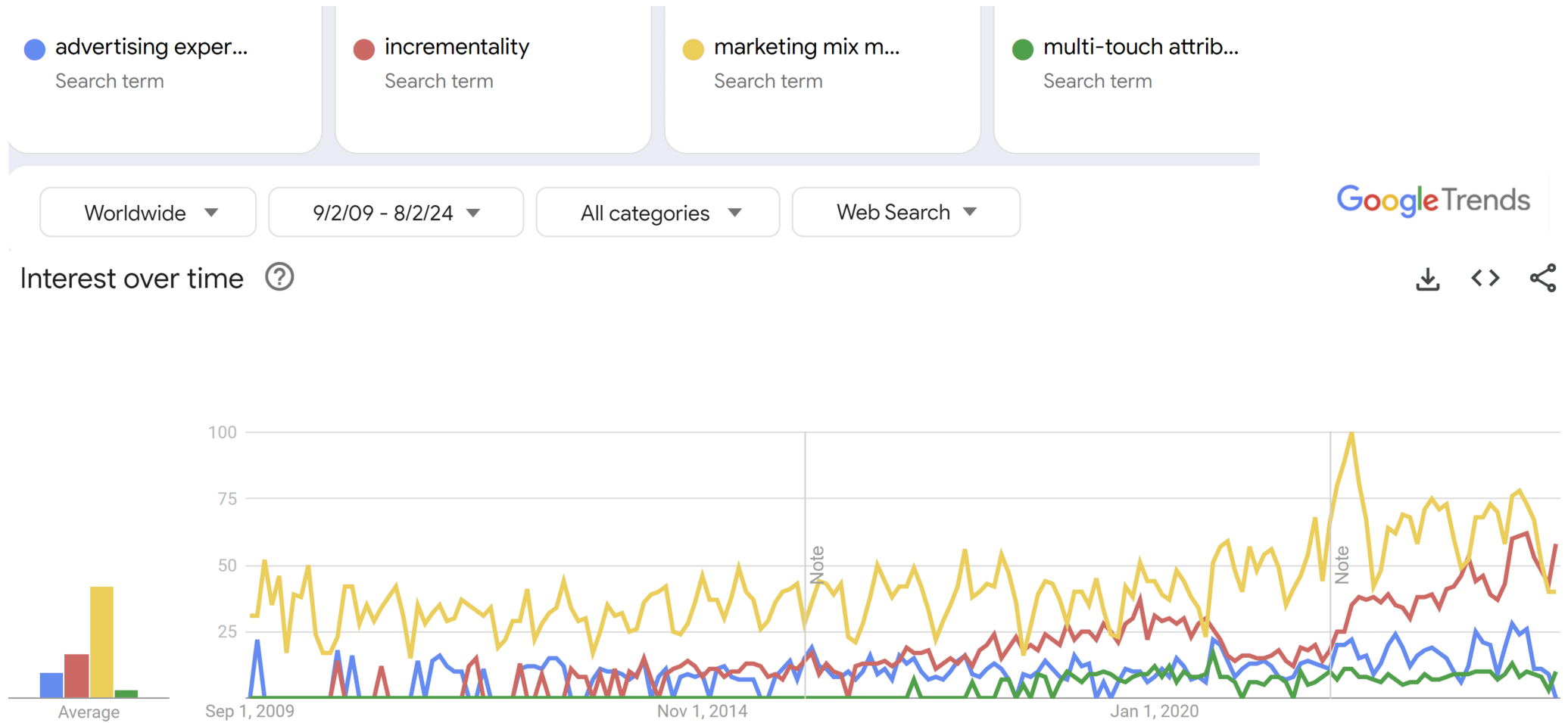


Incrementality refers to the measure of the additional impact or value generated by a specific action, campaign, or intervention beyond what would have occurred naturally without it. In marketing and advertising, incrementality is often used to determine the effectiveness of campaigns by comparing the results of those exposed to the campaign versus a control group that was not exposed. This helps in understanding the true value and ROI of marketing efforts.

## Key Points of Incrementality

1. **Causal Inference:** Incrementality is rooted in causal inference, aiming to isolate the effect of a specific action from other factors.
2. **Control Groups:** A key methodology involves using control groups to measure what would have happened in the absence of the intervention.
3. **Lift:** Incrementality is often expressed as "lift," representing the increase in desired outcomes (sales, conversions, engagement, etc.) due to the campaign.
4. **A/B Testing:** Commonly used techniques to measure incrementality include A/B testing, where one group is exposed to the treatment, and the other is not.
5. **Attribution Models:** Incrementality is crucial for accurate attribution models, ensuring that credit is assigned correctly to the actions that drive results.





- I believe we're a few years into a generational shift
- However,  $\text{corr}(\text{ad}, \text{sales})$  is not going away
- $\text{Union}(\text{correlations}, \text{experiments})$  should exceed either alone

# Marketing Mix Model

- The “marketing mix” consists of quantifiable marketing efforts, such as product line, length and features; price and price promotions; advertising, PR, social media and other communication efforts; retail distribution intensity and quality; etc.
- A “marketing mix model” quantifies the relationship between marketing mix variables and outcomes
  - Idea goes back to the 1950s
  - E.g., suppose we increase price & ads at the same time
  - Or, suppose ads increased demand, and then inventory-based systems raised prices
- A “media mix model” quantifies numerous advertising efforts & relates them to outcomes
  - For example, suppose the brand bought ads from 000s of publishers
  - Confusingly, both abbreviated MMM (or mmm) and often feature similar structures
- MMM goal is to quantify past marketing mix effects, to better inform future efforts

# MMM elements

Typically, MMM uses market/time data

- Outcome: usually sales. Could include more funnel metrics (visits, leads, ...)
- Predictors: Marketing mix factors under our control, plus competitor variables, seasonality, macroeconomic factors, + any other demand shifters

Model structure is usually some type of panel regression, vector autoregression, bayesian model, or machine learning model

- Often includes lags, nonlinear ad effects, interactions between variables
- Regressions typically estimate marginal effects, not average effects
- Nonlinearities built into the model, such as Inc or Dec returns to ad spend, can drive key results

MMM often used to retrospectively evaluate advertising media and copy, advertising interactions, and inform future ad budgets

- MMM coefficient estimation requires sufficient variation in marketing actions

# MMM Considerations

- MMM results are correlational without experiments or quasi-experimental identification strategy
- Data availability, accuracy, granularity and refresh rate are all critical
- MMM requires sufficient variation in predictors, else it cannot estimate coefficients
- “Model uncertainty” : Results can be strongly sensitive to modeling choices
- MMM is gaining traction as digital privacy rules limit user data: E.g. [Google’s Meridian](#) or [Meta’s Robyn](#)
- For much more, see this [MSI White Paper](#) or the [MMM Wikipedia article](#)

# Other Popular Ad/Sales Approaches

Remember, model <> identification strategy

- Lift Tests
- Multi-touch attribution (MTA)
  - Seeks to allocate "credit" for sales across advertising touchpoints
  - Related: First-touch attribution, last-touch attribution
- Cookie-based approaches vs. Google's Privacy Sandbox
- Ghost ads
- Other platform-provided experimentation tools



**Kenneth Wilbur** · You  
 Professor of Marketing and Analytics at University of California, San Diego...  
 4d · Edited · 🌐

MSBA student asked a great question. How would you answer?

Suppose you understand the importance of incrementality in advertising measurement, but everyone you work with prefers correlational measurements, and some actively discourage experiments. What should you do?



**Robert Olinger** 4d ...  
 Assistant Dean, Institutional Collaboration at Duke University - The Fuqu...

Ask the colleagues to teach you more about correlational measurements. Listen to them first, then ask what they have learned about incrementality. This is a psychological problem more than a preference, so use psychology to address it.

Like 🍷 4 · Reply 3



**Rachel Fagen** 4d ...  
 COO | Co-Founder | Partner | Advisor



Like · Reply



**Kenneth Wilbur** **Author** 4d ...  
 Professor of Marketing and Analytics at University of California, Sa...

**Robert Olinger** could you say more about what you mean by psychological problem?

Like · Reply



**Robert Olinger** 4d ...  
 Assistant Dean, Institutional Collaboration at Duke University - Th...

**Kenneth Wilbur:** I believe if experimentation is actively discouraged, this is due to aversion, a desire to feel comfortable, a desire to feel right, loss aversion, etc. The way you phrase the argument sounds like a lack of openness to listen--so my advice is you need to open up the colleagues--the best way to do that is by listening to them, understanding as best you can their expertise and approach--then engage their curiosity toward something new--the experimentation has to seem like it was their idea--so focus on engaging curiously with the colleagues, and when there is an openness ask questions related to the ideas you want included. Have them think about it... This is the way to shift preferences--persistent nudging.

Like 🍷 2 · Reply



**Joel Persson** 4d ...  
 Research Scientist at Spotify | Causal Inference, Machine Learning and D...

You could demonstrate the value of experimentation for the business use case, for instance by showing via simulation that correlational evidence can lead to incorrect decisions (product launches, rollouts, etc) but that causal estimates from experiments get it right. You could even attach a relevant business metric (dollar value, engagement, reach, etc' ...see more

Like 🍷 4 · Reply



**Dean Eckles** (edited) 4d ...  
 scientist & statistician; faculty at MIT

One option: Consider looking for a new job. The number of firms with people who get A/B testing has expanded a lot. Fits with avoiding being the smartest person in the room. (Of course, there are other good options... but as a person in a junior role, this is one of the better ones.)

Like 🍷 5 · Reply



**Brett Gordon** 4d ...  
 Professor of Marketing at Kellogg School of Management | Amazon Sch...

Definitely bring in academics as outside consultants ;-)

Like 🍷 6 · Reply



**Nirzar Bhaidkar** 4d ...  
 Executive Paid Search @ GroupM | AI-Driven Marketing

Propose small scale pilot experiments to demonstrate the value of incremental measurement without significant resource investment.

Like 🍷 1 · Reply



**Ayman Farahat** 10h ...  
 Principal Scientist at Amazon



Like · Reply



**Brad Shapiro** 3d ...  
 Professor at The University of Chicago Booth School of Business

Generally agree with **Dean Eckles**. But depends on their reason for discouraging experimentation. If it is a genuine lack of understanding, I would try and be persuasive, show examples of how correlational assessments might lead you astray, etc. If it is an agency problem whereby they feel they need to mislead their management in order to keep their jobs, I'd say look for another job.

Like 🍷 1 · Reply



**Michael Cohen** 2d ...  
 Customer Centric Privacy Protecting Marketing AI

Change the way they are compensated or incentivized to be aligned with marginal economics of business aligned kpis.

Like 🍷 4 · Reply

# Ken's take

- Adopting incremental methods is a resume headline & interesting challenge
  - Team may have a narrow view of experiments or how to act on them
  - Understanding that view is the first step toward addressing it
- Correlational + Incremental > Either alone
  - What incrementality might be valuable? What's our hardest challenge?
  - What quasi-experimental measurement opportunities exist?
  - Can we estimate the relationship between incremental and correlational KPIs?
- Going-dark design
  - Turn off ads in (truly) random 10% of places/times; nominally free
  - How does going-dark result compare to correlational model's predicted sales?
  - Can we improve the model & motivate more informative experiments?
- If structural incentives misalign, consider a new role
  - It's hard to reform a culture unless you're in the right position
  - Life is short, do something meaningful

# Takeaways

- Fundamental Problem of Causal Inference:  
We can't observe all data needed to optimize actions.  
This is a missing-data problem, not a modeling problem.
  - Experiments, Quasi-experiments, Correlations, Ignore
- Experiments are the gold standard, but are costly and difficult to design, implement and act on
- Ad effects are subtle but that does not imply unprofitable





# Going deeper

- [What is Incrementality? And How Do We Measure it in 2024?](#)
- [Inferno: A Guide to Field Experiments in Online Display Advertising](#): Covers frequent problems in online advertising experiments
- [Inefficiencies in Digital Advertising Markets](#): Discusses digital RoAS estimation challenges and remedies
- [Your MMM is Broken](#): Smart discussion of key MMM assumptions
- [The Power of Experiments](#): Goes deep on digital test-and-learn considerations
- [New Developments in Experimental Design and Analysis \(2024\)](#) by Athey & Imbens
- [Mostly Harmless Econometrics](#): Covers quasi-experimental techniques

