

Introduction to Advertising Measurement

[Kenneth C. Wilbur](#)

Professor of Marketing and Analytics

University of California, San Diego

Revised August 2026 | License: CC BY 4.0 | We use javascript to track readership.

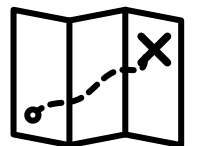
Our main goal is to help advertisers make better decisions. We welcome reuse with attribution. Please share widely.

Contents

- Domain Knowledge
- Causality
- Fundamental Problem of Causal Inference
- Advertising Measurement, Defined
- Correlational Advertising Measurement
- How much should we rely on Correlational Ad Measurement?
- Causal Advertising Measurement
- Who experiments and why?
- Integrating Methods
- Wrapping Up

This Deck's Value Proposition

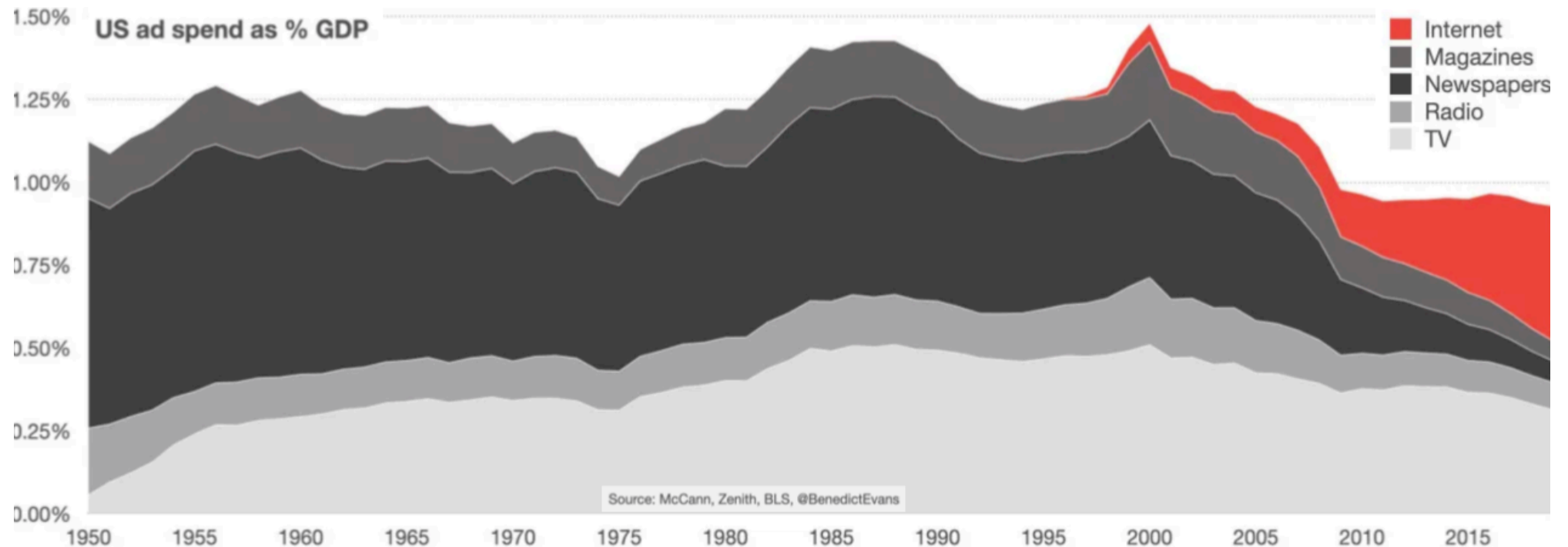
- One day soon, someone may ask, or tell, you how well the ads worked
 - Advertising measurement is a hard, important, interesting problem without any general solution
- After reading, you will be able to:
 - Identify incentive conflicts in the advertising supply chain
 - Describe pros & cons of major advertising measurement methods
 - Understand how multiple methods may be productively combined to optimize spending allocations under a budget
- This knowledge can help you balance maximizing advertising profits with returning profits to shareholders



Domain Knowledge

- Context required to interpret advertising measurement

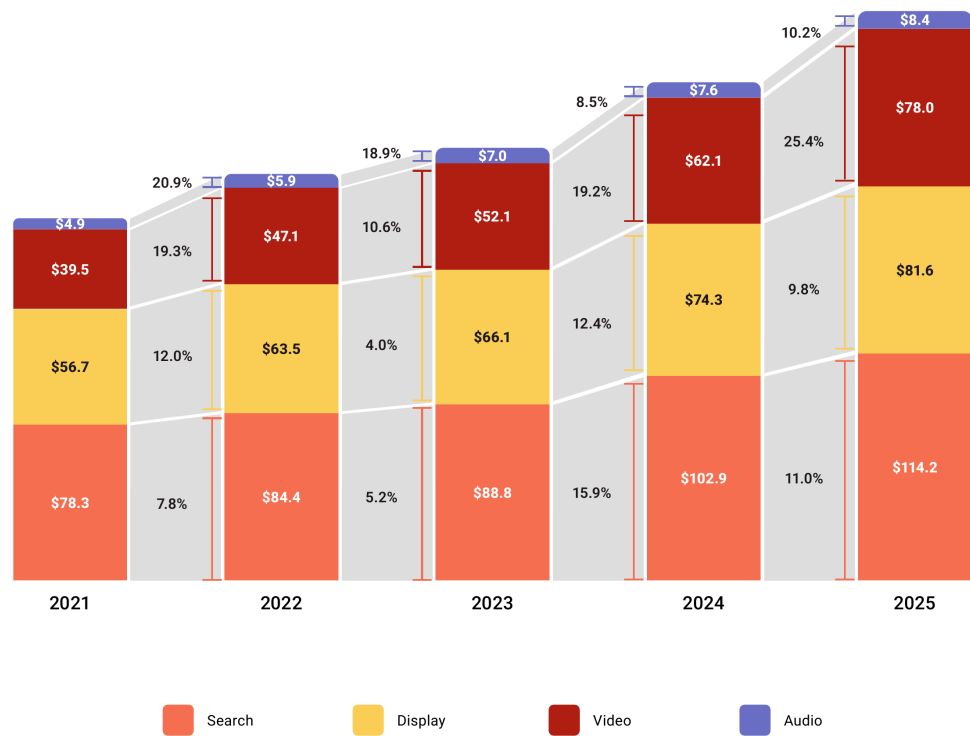
Advertising sales revenue, 1950-2019



For 70 years, between \$0.95-1.50 of every \$100 spent in America bought an ad. These figures report advertising sales revenue to publishers (i.e., entities that attract and sell consumer attention) and exclude supply chain fees (e.g., ad agencies), which are considerable. What else do you see?

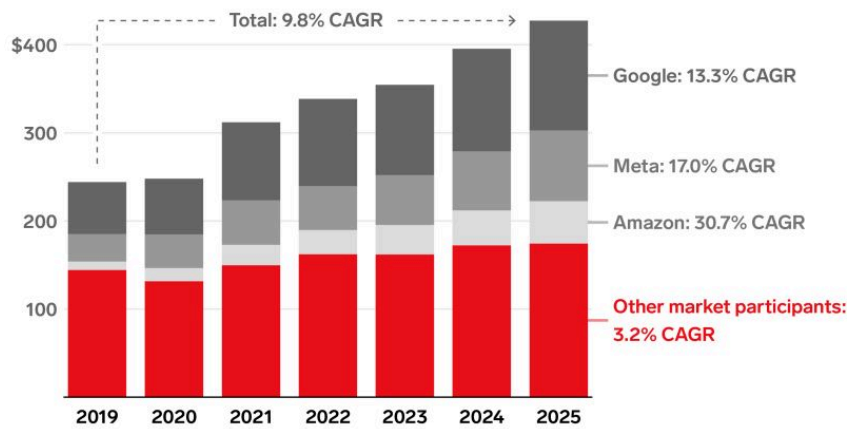
Evans (2020)

More Recent Online Ad Revenues



US Ad Spend Outside of Amazon, Google, and Meta Grew at a 3% CAGR From 2019 to 2025

US total ad revenues in billions, by company, 2019-2025



Note: Google figures represent gross ad revenues before companies pay traffic acquisition costs (TAC) to partner sites, and include YouTube; Meta and Amazon figures represent net ad revenues; Meta includes Facebook, Instagram, and Threads ad revenues, and excludes advertising on WhatsApp; total ad spending includes digital (desktop/laptop, mobile, and other internet-connected devices), directories, magazines, newspapers, out-of-home, radio, and TV

Source: EMARKETER Forecast, Dec 2025

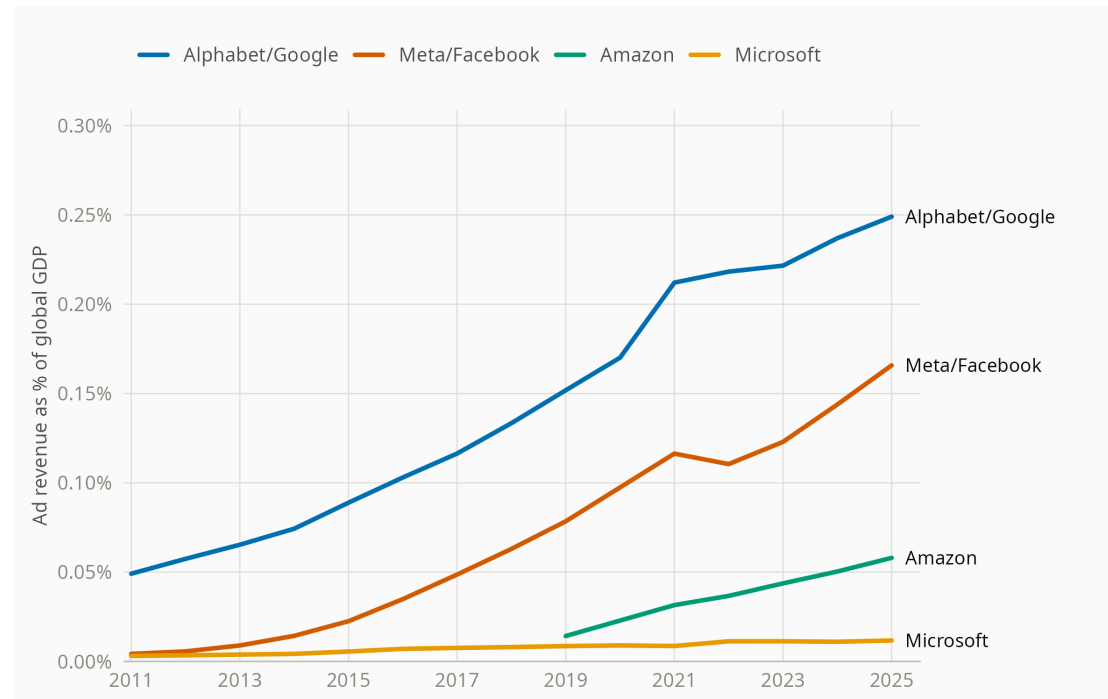
361437

EMARKETER

These data expand the Internet bar. The Interactive Advertising Bureau (IAB) collects ad sales revenue by format from online ad sellers and supply chain firms. Total spending grew from \$179.4B in 2021 to \$282.2B in 2025, 0.92% of gross domestic product (GDP). What else do you see?

Left: IAB (2026). Right: EMARKETER (Dec 2025), via Hanzlik (2026)

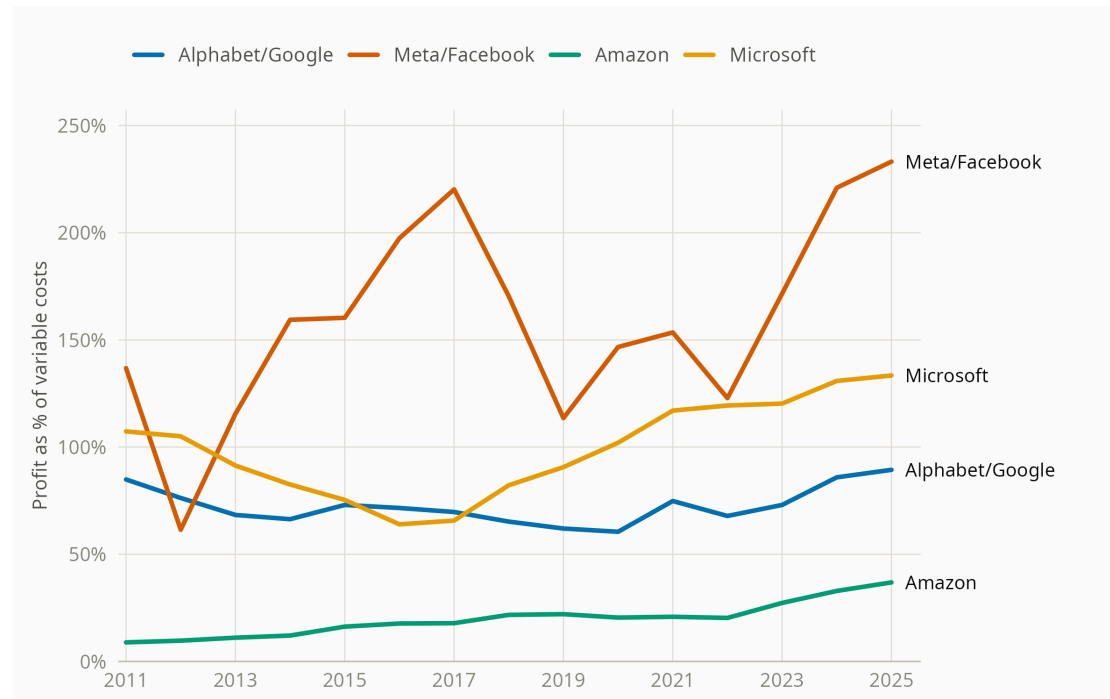
Ad Sales Revenue as % of Global GDP, 2011-2025



Online advertising sales are increasingly dominated by a few large firms; the top 3 sold \$559.5B in ads globally in 2025. We used to call them “the duopoly,” but now we call them “the triopoly.”

The numerator is advertising revenue only, not total company revenue. Each firm reports it as a separate line in its 10-K revenue note — Alphabet’s “Google advertising” (Search & other, YouTube ads, Google Network), Meta’s “Advertising,” Amazon’s “Advertising services,” and Microsoft’s “Search and news advertising.” In 2025 those lines were 73%, 98%, 10% and 5% of each firm’s total revenue, so the chart is not a ranking of company size. The denominator is global nominal GDP. Amazon’s line starts at its first disclosure (2019); Microsoft is plotted at June fiscal years; its FY2014–15 points are inferred from growth rates disclosed in those years’ 10-Ks, which reported no levels.

Large Ad Seller Profit Percentage, 2011-2025



Online advertising sales is a remarkably high-margin line of business, in part due to limited marginal costs, high efficiencies, and supply-side concentration. Note, these percentages are across all lines of business. What might these profit margins indicate to ad buyers?

Profit percentage = $(\text{revenue} - \text{variable costs}) \div \text{variable costs}$, where variable costs are cost of goods sold plus selling, marketing, and general and administrative expense. R&D is excluded. Computed across all lines of business, not advertising alone.

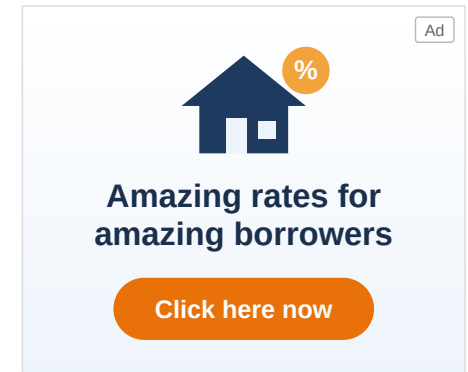
Data: company 10-K filings. Chart concept: [Marto and Le \(2024\)](#)

Toy economics of advertising

- Suppose we pay \$10 to buy 1,000 digital ad Opportunities to See (“OTS”). Suppose 3 people click, 1 person buys.
- Ad profit > 0 if transaction margin > \$10
 - But we bought ads for 999 people who didn’t buy
- Or, ad profit > 0 if customer lifetime value (CLV) > \$10
 - Long-term mentality justifies increased ad budget
- Or, ad profit > 0 if CLV > \$10 *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, conversion rate, margin ... and how we formulate our objective function
 - Exception: Search ads may convert at 1-10+%, but incrementality questions are even bigger

Right ad, right person, right context, right time?

- Imagine you're selling mortgages. Mortgage lenders offer numerous loans at distinct price points. Yet 78% of consumers say they only apply to a single lender/broker for a quote ([FHFA 2024](#)), and most borrowers actively seek a loan for only a few days or less
- To advertise profitably, you may need to find people who
 - Can qualify for a loan
 - Actively want to buy a new home
 - Are thinking about the finance process
 - Have not signed a loan yet
- Predicting which consumers to reach is necessary but insufficient. You also have to identify the brief window of time when an ad might shift each borrower's behavior, and reach them in a context where they might act on your message



Even a perfectly efficient and omniscient advertising industry might struggle to learn how to optimize advertising delivery. What behavioral or contextual signals might indicate mortgage loan receptivity? How much more cost-effective would these targeting signals make the ads?

AI Trades Online Ads

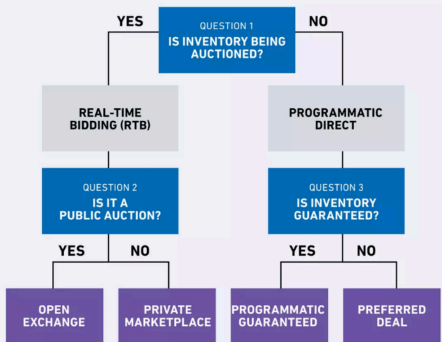
Programmatic: media or ad buying that uses technology to automate and optimize, in real time, the ad buying process. This ultimately serves targeted and relevant experiences to consumers across channels. On the back end, algorithms filter ad impressions derived from consumer behavioral data, which allows advertisers to define budget, goal, and attribution and optimize for reduced risk while increasing ROI.

iab. Glossary of Terminology

iab.com/insights/glossary-of-terminology/#index-17

UNDERSTANDING PROGRAMMATIC ADVERTISING IN THREE EASY STEPS

By asking three simple questions, you can identify the four main approaches to programmatic advertising and better understand the different transaction types used.



Source: eMarketer, Oct. 2018

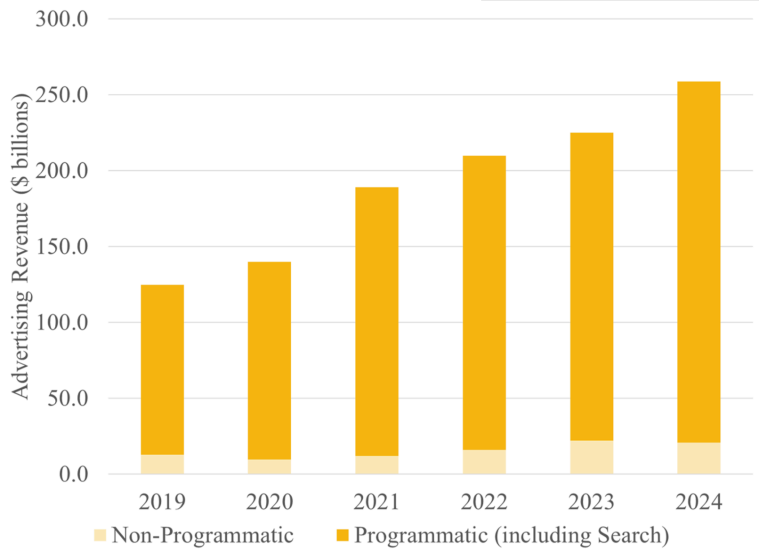
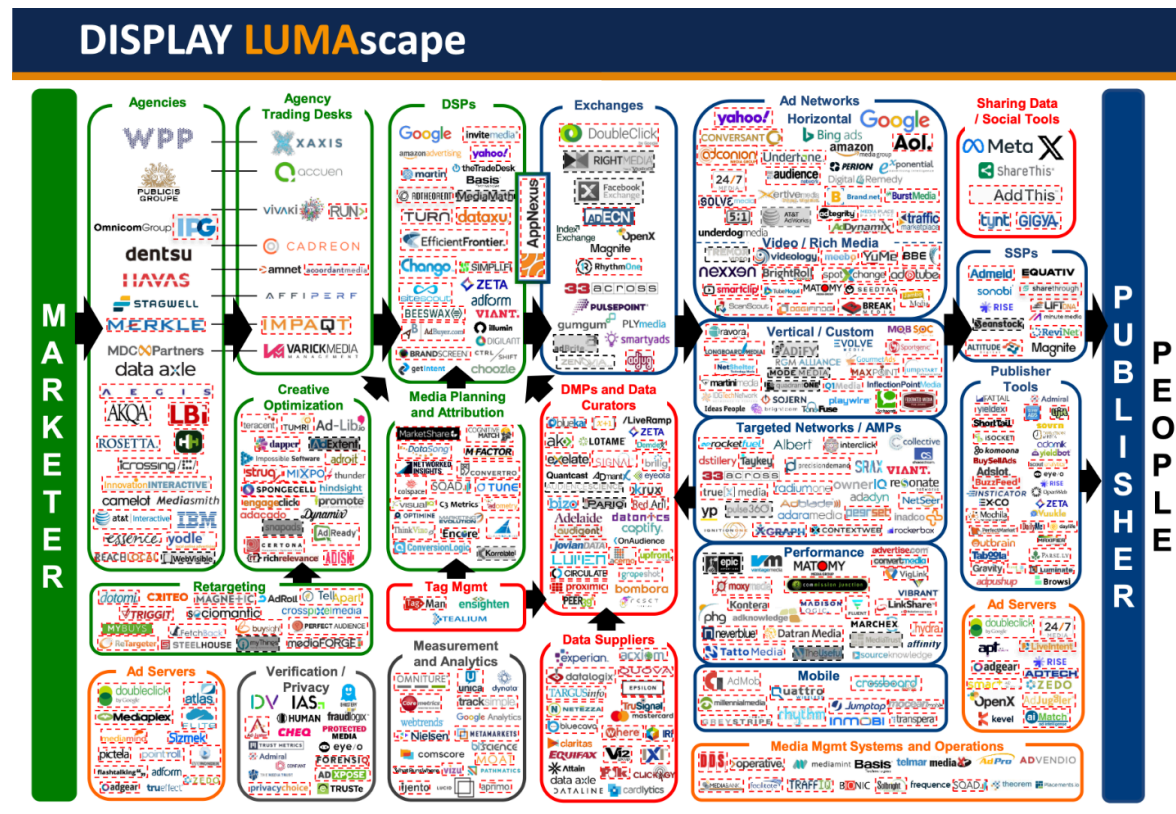


Figure 2. Internet Programmatic Advertising Revenues, 2019-24⁴⁸

⁴⁸ Original analysis and presentation of data reported in IAB/PwC Internet Advertising Revenue Reports between 2019 and 2024. See also Lebow, Sara. "5 recent charts forecasting how ad spend is changing, from retail media to programmatic." *eMarketer*, 16 Feb. 2024, Accessed 21

Advertising was the second industry to automate trading, after finance. ‘Programmatic’ methods are defined by automation and optimization. Over 90% of online advertising revenue flows through Programmatic channels, in which buyers and sellers are both represented by computerized agents. What is being automated and optimized, and for whose benefit?

The Ad Tech Ecosystem

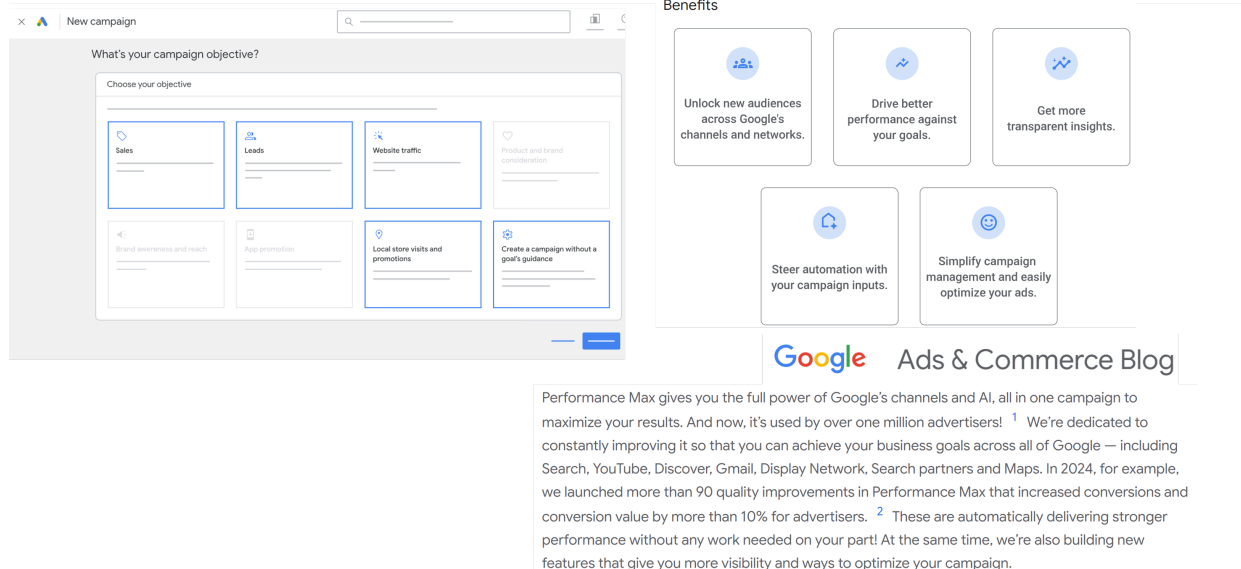


Luma Partners maps ad tech ecosystems. Each logo is a company that intermediates between advertisers and publishers: data/algorithm specialists, representatives, and marketplaces. This map is one among many. Contrast this with “walled gardens,” which vertically integrate ad sales, creation, targeting, delivery & measurement.

Google Performance Max

Google Ads Help

Performance Max is a goal-based campaign type that allows performance advertisers to access all of their [Google Ads inventory](#) from a single campaign. It's designed to complement your keyword-based Search campaigns to help you find more converting customers across all of Google's channels like YouTube, Display, Search, Discover, Gmail, and Maps.

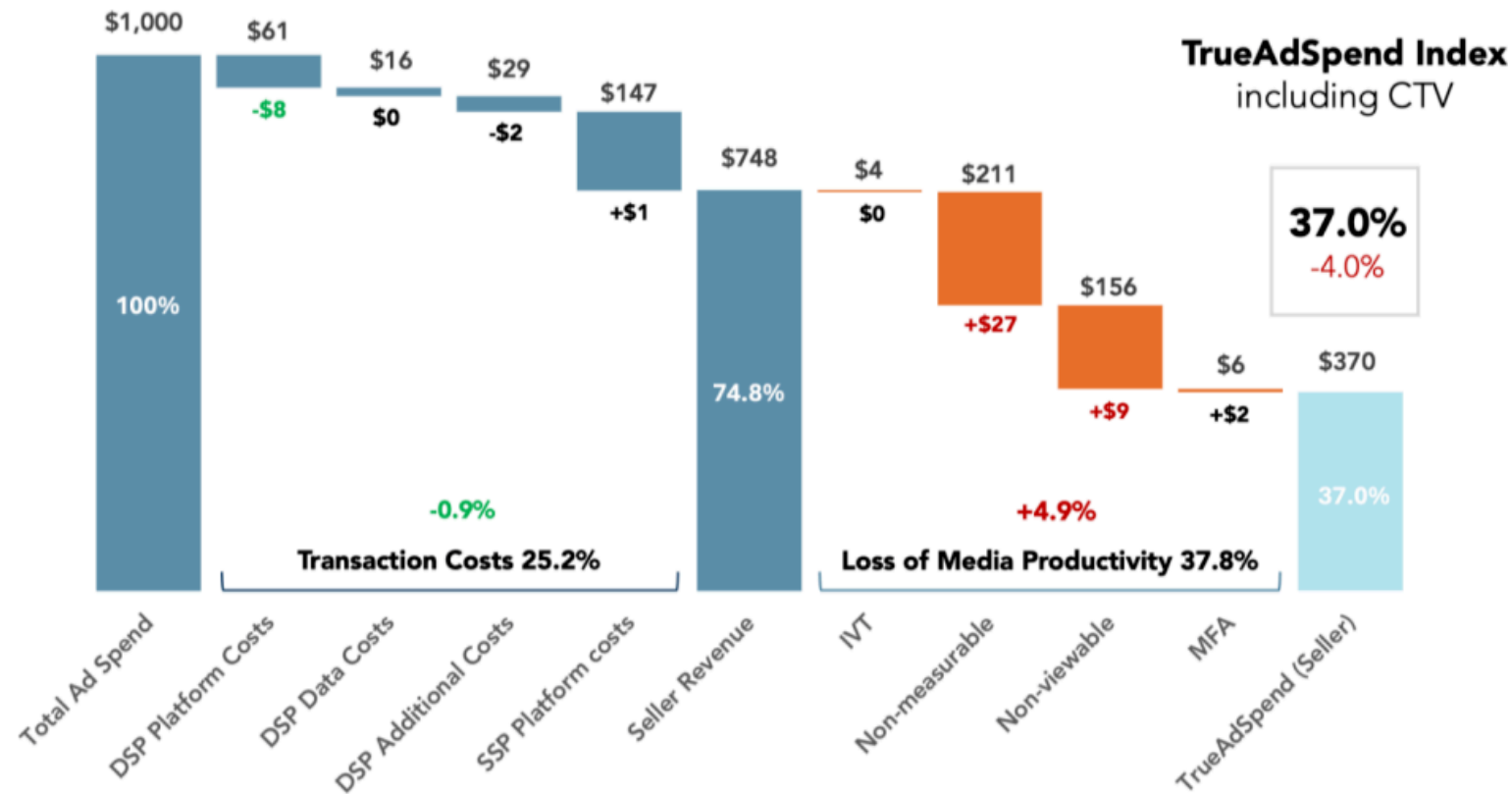


The image shows a screenshot of the Google Ads 'New campaign' setup interface. On the left, a sidebar titled 'What's your campaign objective?' lists various goals: Sales, Leads, Website traffic, Product and brand consideration, Brand awareness and reach, App promotion, Local store visits and promotions, and Create a campaign without a goal's guidance. The 'Website traffic' option is highlighted. On the right, a 'Benefits' section lists five advantages: Unlock new audiences across Google's channels and networks, Drive better performance against your goals, Get more transparent insights, Steer automation with your campaign inputs, and Simplify campaign management and easily optimize your ads. Below this, the 'Google Ads & Commerce Blog' header is visible, followed by a paragraph of text about Performance Max's capabilities and its adoption by over one million advertisers.

Performance Max gives you the full power of Google's channels and AI, all in one campaign to maximize your results. And now, it's used by over one million advertisers! ¹ We're dedicated to constantly improving it so that you can achieve your business goals across all of Google — including Search, YouTube, Discover, Gmail, Display Network, Search partners and Maps. In 2024, for example, we launched more than 90 quality improvements in Performance Max that increased conversions and conversion value by more than 10% for advertisers. ² These are automatically delivering stronger performance without any work needed on your part! At the same time, we're also building new features that give you more visibility and ways to optimize your campaign.

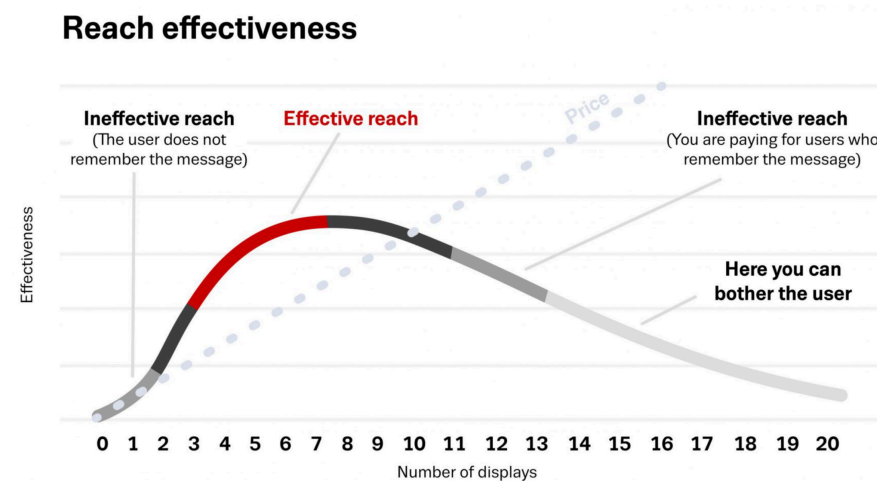
Google Performance Max (Pmax) is the ultimate expression of programmatic advertising. You give Google your goals, your budget, and things it can say in ads. Google decides where, when and how to spend your money, designs your ad, then tells you how well it did. Launched in 2021; over 1 million advertisers served by 2025. Meta's Advantage+ is similar.

The “Ad Tech Tax”

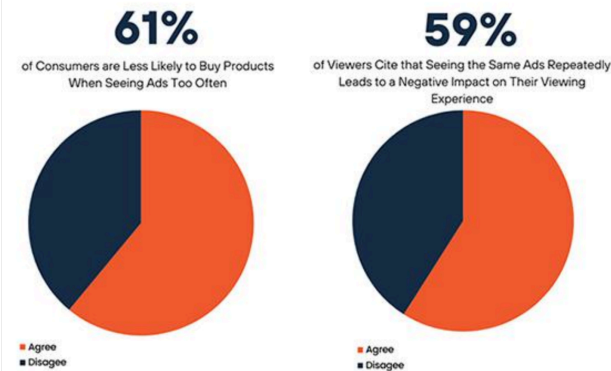


2025Q2 data show that DSP takes 11%, SSP takes 15%, publisher receives 75%, and about half of that is verifiably viewable by human recipients. What are demand-side platforms (DSP), supply-side platforms (SSP), invalid traffic (IVT), Measurable, Viewable, and [made-for-advertising](#) (MFA) sites?

Effective Frequency



How Much Is Too Much?



Source: AD-ID. This survey was conducted online within the United States by The Harris Poll on behalf of AD-ID from September 10-12, 2024 among 2,079 U.S. adults ages 18+, among which 1,921 watch TV or view content on streaming services ("Viewers"). The sampling precision of Harris online polls is measured by using a Bayesian credibal interval. For this study, the total sample data is accurate to within +/- 2.5 percentage points using a 95% confidence level.

Have you ever had an ad “follow you around”?

Age-old advertising theory posits a nonlinear effective frequency curve, which is to say, the marginal effect of an ad on conversion probability depends on how many times the consumer sees the ad.

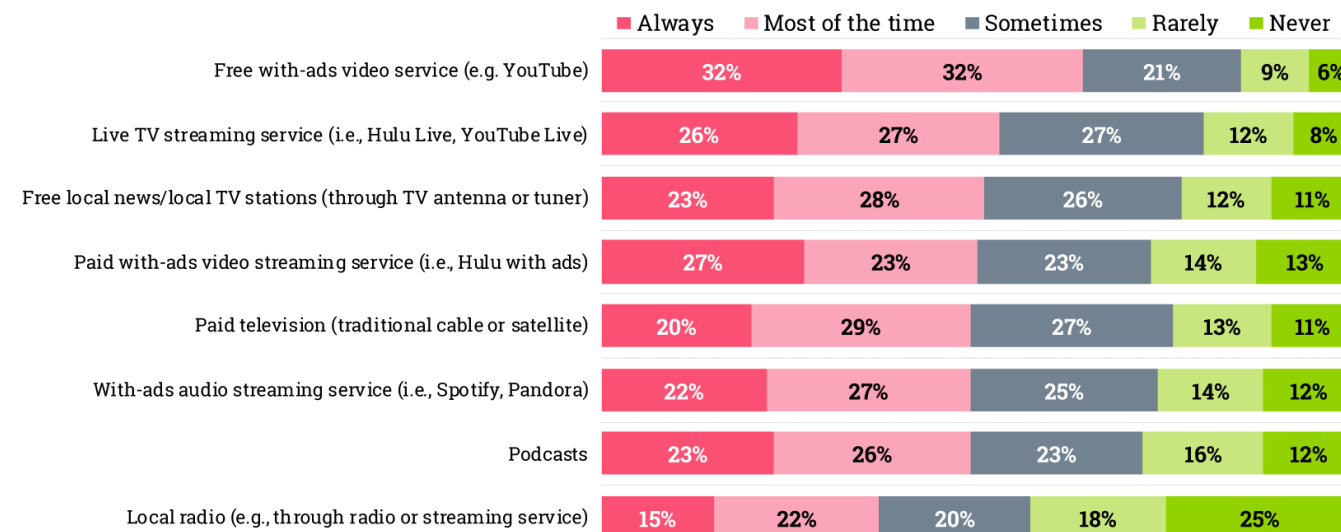
Why is effectiveness convex for exposures 1-3?

Frequency Capping limits ad exposures per individual. Retargeting targets consumers based on past actions (e.g., product detail pageviews, add-to-cart)

Advertising avoidance

US Adults' Stated Ad Avoidance, by Medium

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

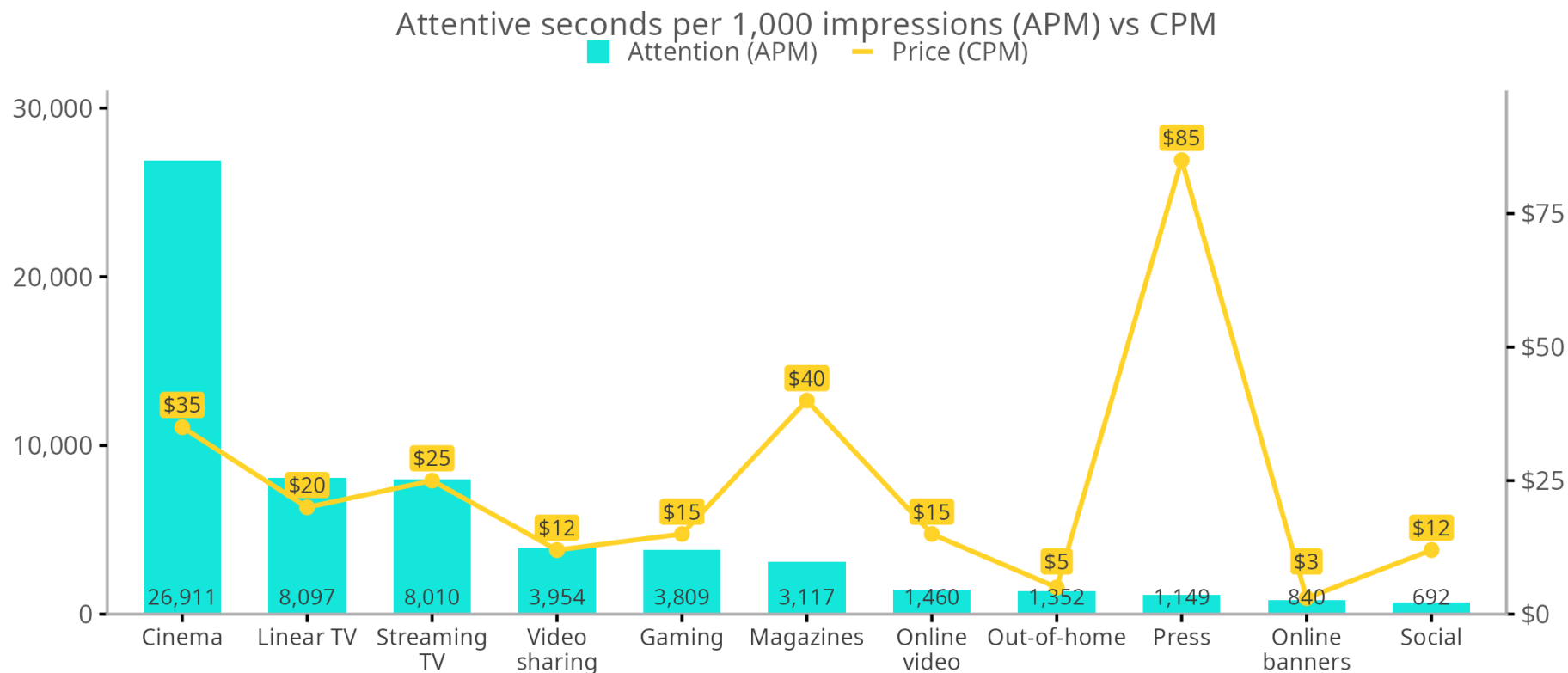


Published on MarketingCharts.com in December 2023 | Data Source: Nielsen

Based on an April survey of 3,000 US adults (18+)

Can an ad work if a consumer avoids it? About half of consumers say they usually or always skip ads. Do you use an ad blocker in your favorite browser? *Soft* ad blockers, like Adblock and Adblock Plus, make money by charging publishers for NOT blocking "Acceptable Ads." *Hard* ad blockers, like uBlock Origin and AdGuard, block all ads and make money with a freemium strategy or they forego revenue.

Attention & Ad Cost by Medium

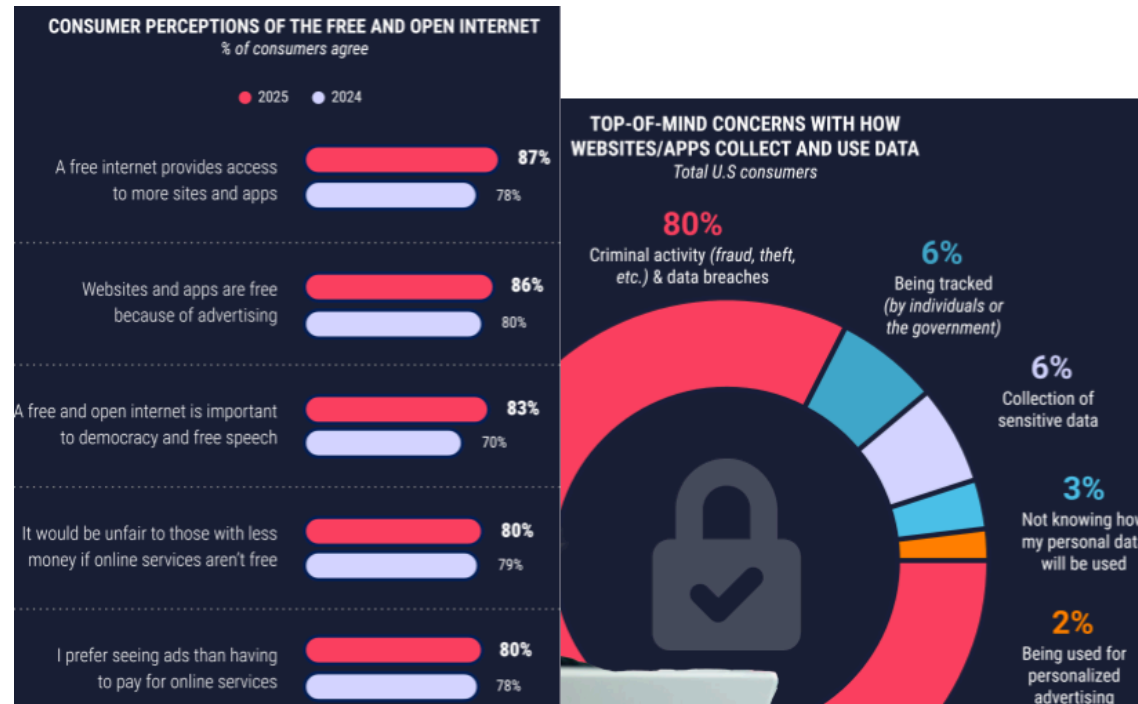


Advertising media vary in average attention attracted (i.e. eyes-on-screen) and advertising price. How does average attentiveness correlate with ad price across media? Why?



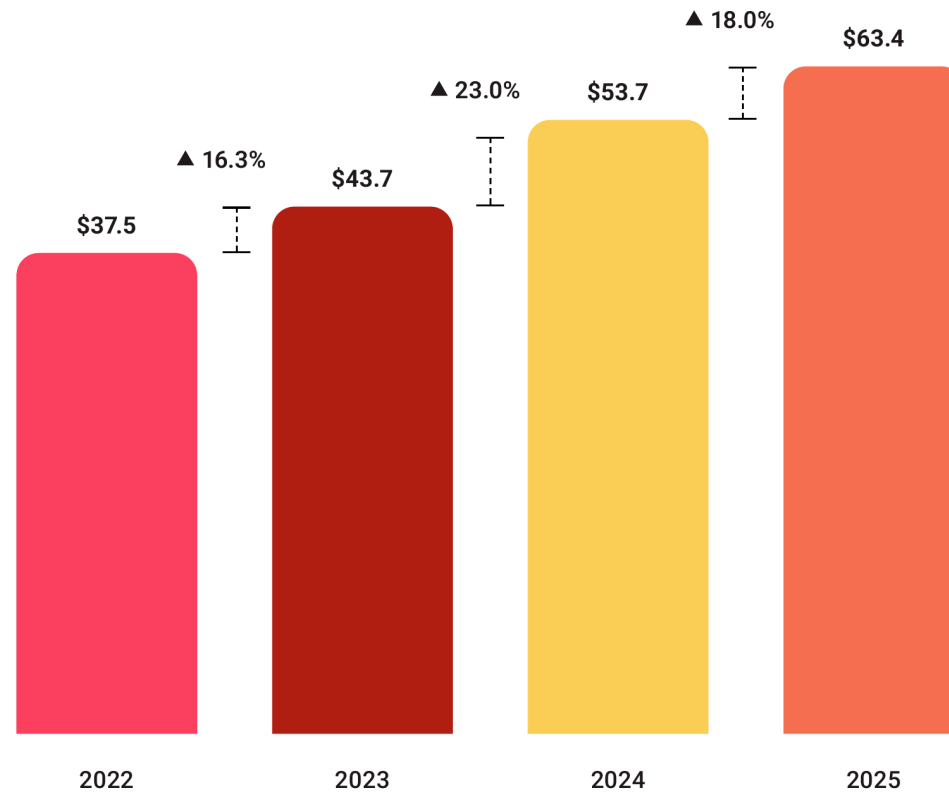
Attention: [Lumen Research](#) 2025 channel norms | Price: [Adspposure \(2026\)](#); cinema [AdQuick](#); gaming [PubScale](#) | Audio formats excluded

Privacy Perceptions



Consumers don't love ad interruptions, but most understand that advertising is the "attentional price" that subsidizes media access, which would otherwise cost more; most consumers prefer to pay with attention rather than money. Personalized advertising ranks low on most consumers' data privacy concerns. Empirical studies usually show that personalized ads generate more conversions because they are more relevant.

Commerce Media



Commerce media (e.g., Amazon, Walmart) provide data for ad targeting, sell sponsored product search listings, sell ads on behalf of publishers, and measure advertising conversions. They grew quickly by cannibalizing trade promotions budgets.



We search for
coupons at 30,000+
sites to help you
save money

Add to Chrome - it's free

THE CREATOR ECONOMY
The Honey scandal is a 'wake-up call'
creator industry's affiliate partners

Honey, which finds coupon codes for online shopping, was exposed by YouTuber Megalag for allegedly hijacking affiliate links from creators and using its own (even in cases where it wasn't a better deal). This has since resulted in class action lawsuits from several creators including YouTubers Legal Eagle and GamersNexus, against the browser extension, claiming that Honey is taking affiliate revenue that belonged to creators.

It's not just Honey. Other companies including Microsoft and Capital One are facing similar claims regarding their browser extensions through Microsoft Shopping and Capital One Shopping. Now creator and legal experts expect to see greater scrutiny

By Adam Osato on February 28, 2025

Cloudflare CEO: Bots have overtaken human traffic online

THE MEDIA COPILOT

For the first time, bots account for more web traffic than humans, according to Cloudflare data.

LAWFARE

Topics ▾ Podcasts & Multimedia ▾ Projects ▾

In an eye-opening investigation, Reuters has revealed that Meta had projected its 2024 advertisements for scams and banned goods would bring in about \$16 billion, or 10 percent of its total revenue.

The report is based on a cache of documents reviewed by Reuters.

In one of those documents, Meta's safety staff estimated that the company's platforms were "involved" in a third of all successful scams in the U.S. That's a stunning figure. But we do wonder how much of that involvement is simply WhatsApp being used to talk to victims. If advertisements weren't the bait that lured victims, it hardly seems fair to blame Meta for running an end-to-end encrypted messaging app.

The company doesn't get such an easy pass elsewhere, though. Other documents revealed that Meta bans advertisers only if its automated systems are 95 percent certain that an account is committing fraud. If the account doesn't meet that threshold, but Meta still believes it is likely a scammer, the company instead charges higher advertisement rates as a "penalty." According to Reuters, the idea here is to discourage suspicious advertisers from buying ads. But in our view it's just as likely to encourage

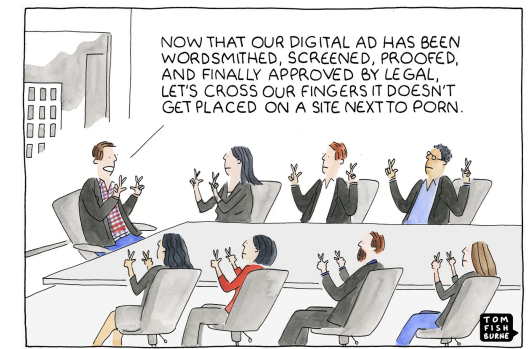
According to Cloudflare's Radar data, bots represented roughly 57% of all HTTP requests as of late April 2026, with humans accounting for the remaining 43%. Bot traffic has held between 53% and 60% in the weeks since. Prince said the actual crossover occurred in the last few months, though the data is messy enough that pinning down an exact date is difficult.

Advertising value delivered is hard to detect, but advertising remains lightly regulated, so fraud is a first-order issue. 3 main types: Fraudulent ads delivered to consumers; fraudulent audiences increasing brand payments to publishers/supply chain (e.g., MFA, made-for-advertising, AKA made-for-arbitrage, sites); and supply-chain participants stealing from each other. Verification firms (e.g., Integral Ad Science) try to detect problems but high-profile failures have occurred. Career professionals believe fraud funds organized crime and hostile governments.

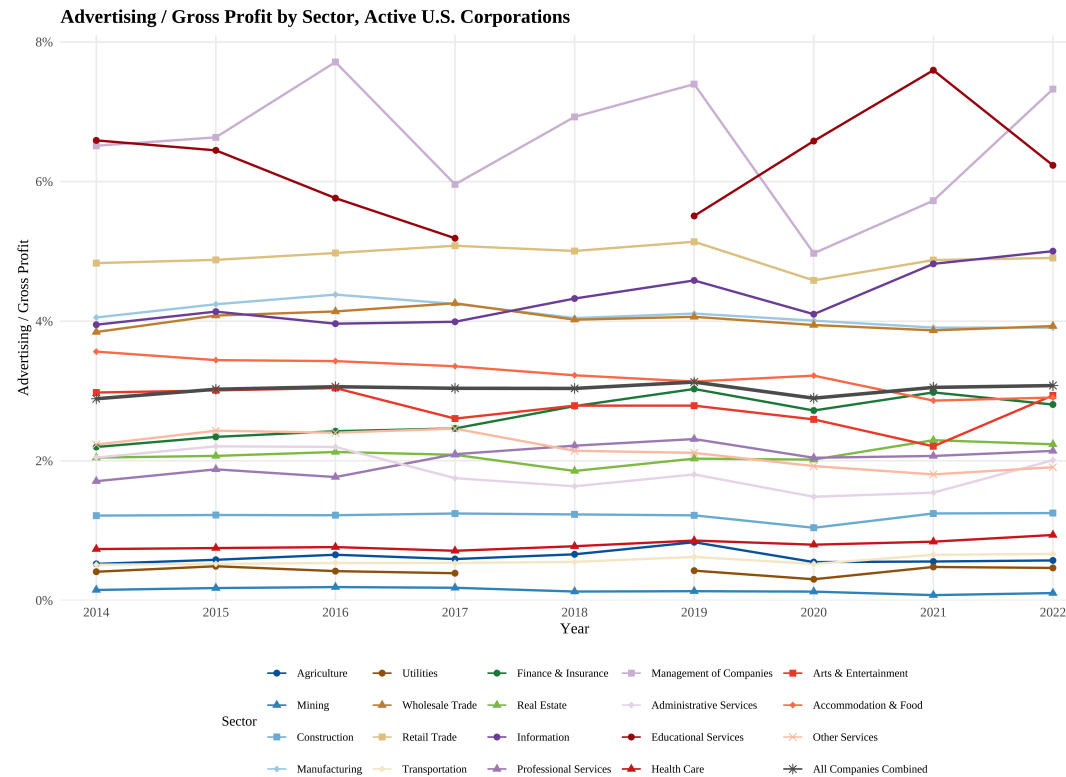
Brand Safety and Suitability

- **Brand safety:** Protect brands against negative impacts on consumer opinion from ads appearing near specific types of content
 - E.g., proximate to military conflict, obscenity, drugs, hate speech
 - Keyword blacklists and whitelists determine contextual ad bids; blacklists **demonetized >50% of Reuters.com news articles**
 - Ad platforms have seen “advertiser boycotts” demanding content moderation improvements; monetization is **expanding recently**
- **Brand suitability:** Identifies brand-aligned content to improve ad delivery

IAB (2020)



How much do companies spend on Advertising?



Source: IRS Statistics of Income, Table 5.1
Ratio: Advertising / Gross Profit
Page 18 of 88

The average US corporation spends about 3.1% of gross profit on advertising deductions. Gross profit usually ranges from 3-5x net profit, so the modal firm could increase net profit by 10.3-18.5% by setting ads to zero (i.e. $100/(100-9.3)$ to $100/(100-15.5)$). Or could it? What would happen to top line revenue and cost efficiencies? [2 missing data points were withheld by source for confidentiality purposes]

What is Incrementality?

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.

Incrementality is the difference between advertising-generated conversions and the conversions that would have occurred anyway without the campaign.

The word incrementality is only used in marketing. However, it is increasingly misused.

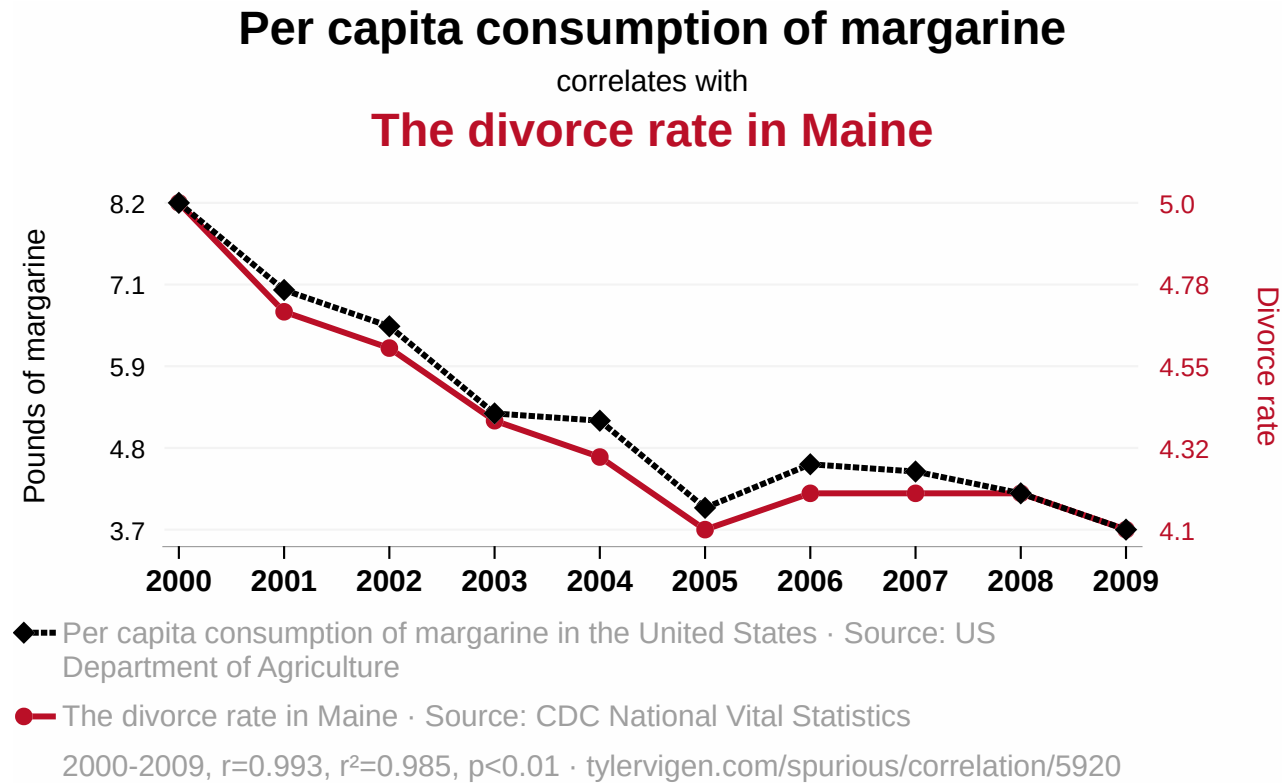
Causality

Examples, fallacies and motivations

“Participants drew causal inferences from non-experimental vignettes as often as they did from experimental vignettes, and more frequently for causal statements and directions of association that fit with intuitive notions than for those that did not.”

Bleske-Rechek et al. (2015)

Correlation $\neq$ Causation



This chart shows a near-perfect correlation between margarine consumption and divorce rates—but does margarine cause divorce?

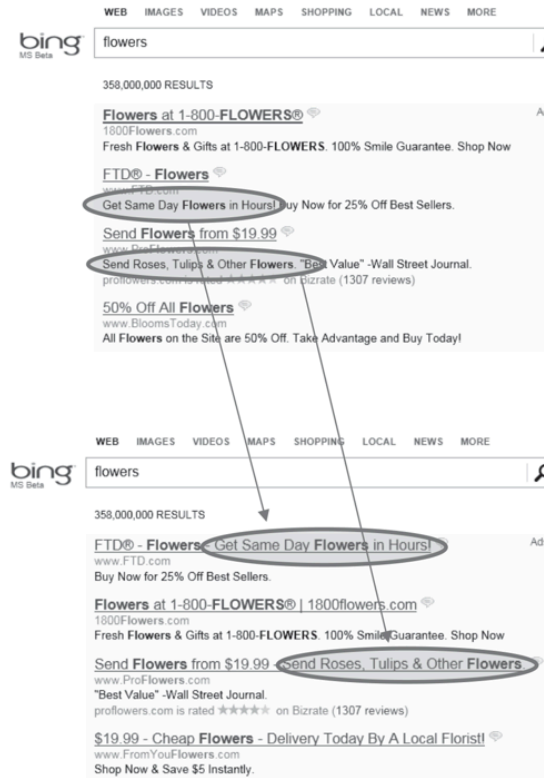
How Many False Positives?

- Suppose you observe 10 customer outcomes, 1,000 predictors, $N=100,000$ obs
 - Outcomes might include visits, sales, reviews, ...
 - Predictors might include ads, customer attributes & behaviors, device/session attributes, ...
 - Suppose you calculate 10k bivariate correlation coefficients
- Suppose everything is noise, no true relationships
 - 10k correlation coefficients would be distributed 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." Anything more requires assumptions. No causal ordering or reason for co-movement can be inferred from a correlation alone.

Classic misleading correlations

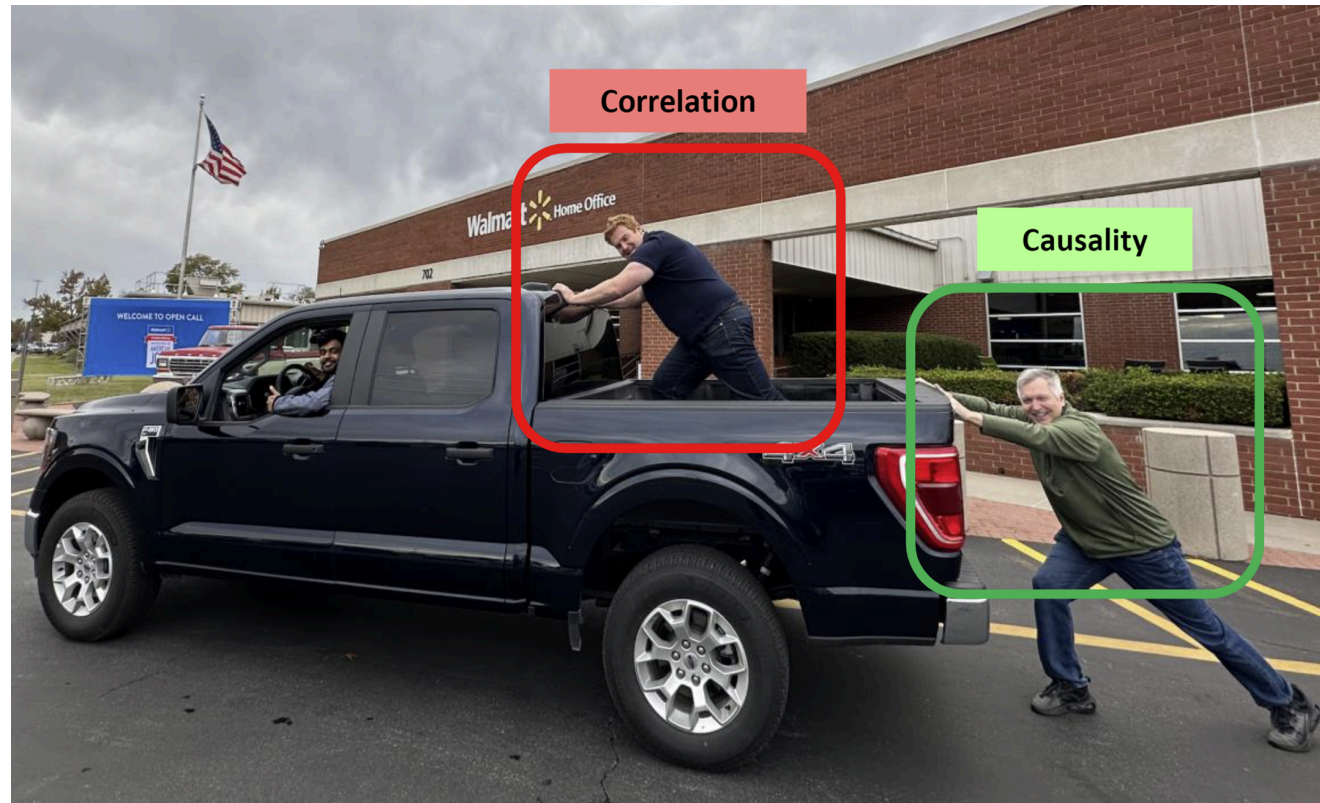
- “Lucky socks” and sports wins
 - Post hoc fallacy (precedence indicates causality AKA superstition)
- Commuters carrying umbrellas followed by rain later in the day
 - Forward-looking behavior: Rooster noises do not cause sunrise
- Kids receiving tutoring often get lower grades
 - Reverse causality / selection bias
- Ice cream sales and drowning deaths co-occur
 - Unobserved confounds (hot temperatures drive both)
- Correlations are measurable & usually predictive, but easy to misinterpret
 - Correlation-based beliefs are hard to disprove and therefore persist
 - Correlations that reinforce logical theories especially persist
 - Correlation-based beliefs may or may not reflect causal relationships

“Revenue too high alert”



This A/B test triggered a “revenue too high” alert at Microsoft Bing in 2012. The treatment improved horizontal space usage, enlarged a selling argument, and expanded the clickable surface in search ads. It increased revenue 12%—over \$100 million per year—without harming user experience metrics.

Correlation vs Causation

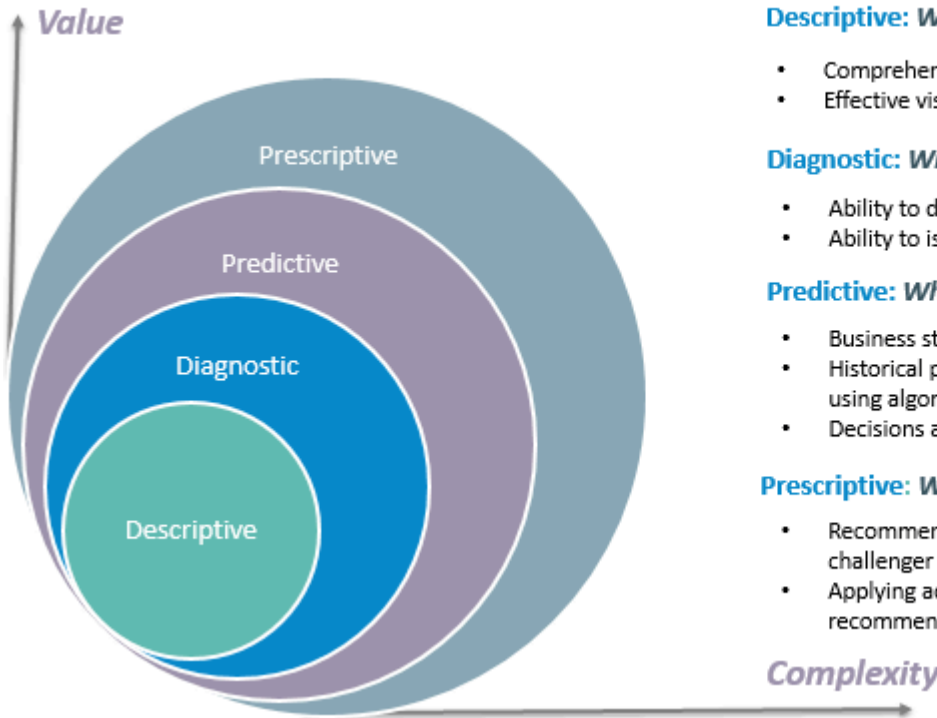


The Correlation guy is silly but he's not harmless. He's weighing down the truck. And there is an opportunity cost: he could be helping to push the truck instead.

List (2023)

Four Types of Analytics

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

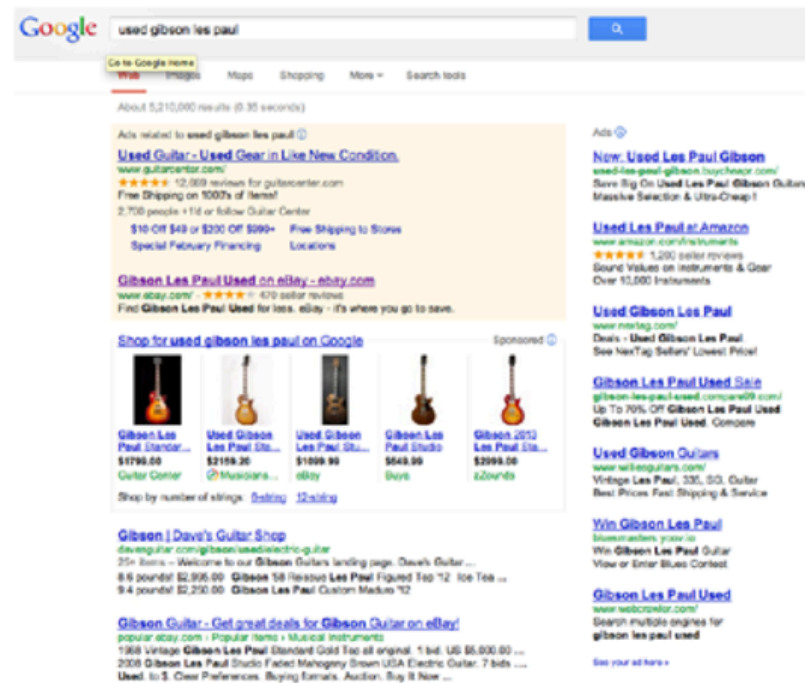
Prescriptive: *What do I need to do?*

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

Correlations are descriptive analytics (“facts”). Causality matters most for diagnostic and prescriptive analytics. The great power of data analytics is cutting through the noise to isolate the effect of a single variable on outcomes of interest, apart from competing and simultaneous causes. Causality can help build predictive models, but predictive correlations often suffice.

eBay Search Ad Experiments

In 2015 economists working at eBay published a series of geo experiments testing how shutting off paid search ads affected search clicks, sales and attributed sales in a random sample of US cities.



eBay Results: Click Substitution

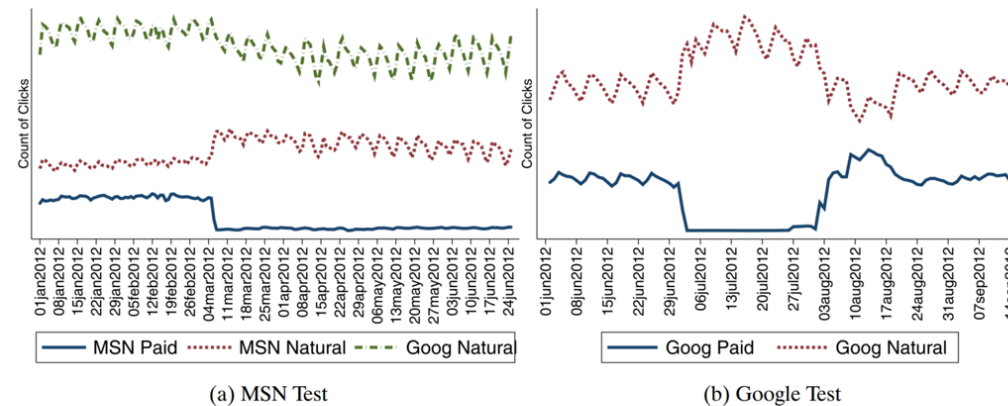


FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for ‘ebay’ terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

In summary, the evidence strongly supports the intuitive notion that for brand keywords, natural search is close to a perfect substitute for paid search, making brand keyword SEM ineffective for short-term sales. After all, the users who type the brand keyword in the search query intend to reach the company’s website, and most likely will execute on their intent regardless of the appearance of a paid search ad.

When eBay turned off paid search ads, clicks on paid branded keywords went to zero—but clicks on organic branded keywords fully replaced them.

eBay Results: Attribution vs Reality

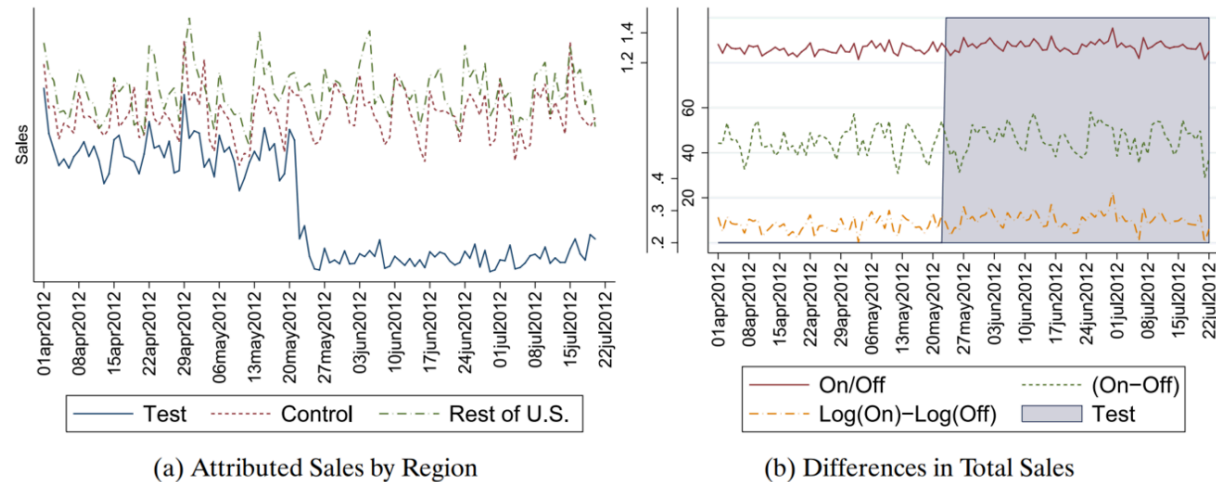


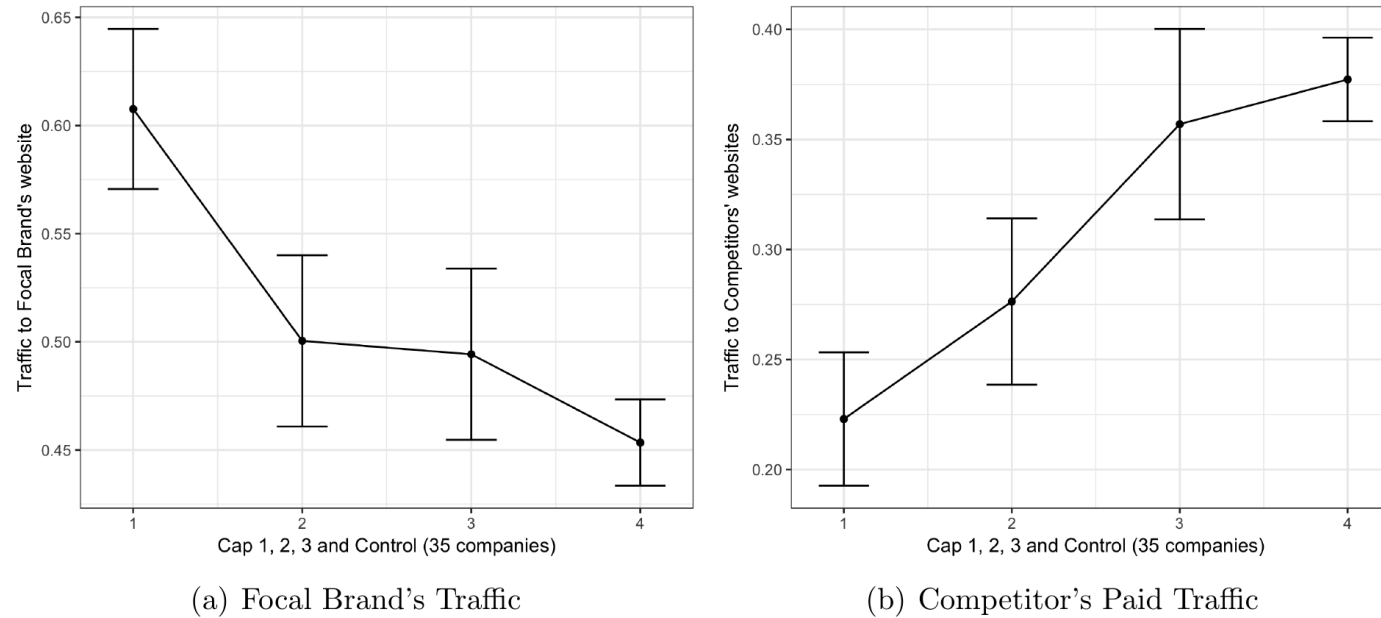
FIGURE 3.—Non-brand keyword region test. Panel (a) plots total purchases by users who clicked on an ad prior to purchase, which drops when the test commences in the test areas. Panel (b) plots three different measures of the difference between test and control regions before and after the test. The y-axis is shown for the ratio, the log difference, and in differences in thousands of dollars per day, per DMA.

Attributed sales fell, but actual sales didn't. (Why?) These results led to changes in eBay ad measurement and Google algorithms. This story became famous for the pitfalls of correlational advertising measurement.

Fisman (HBR 2013): Did eBay Just Prove That Paid Search Ads Don't Work? | Central Control (2025)

Did the eBay result generalize to other companies?

Figure 8: Effect of second, third and fourth competing firms advertising in the top paid position on the page for companies with high CTR.



Point estimates are computed for each brand using the frequency estimator defined in Section 3.4. Results are averaged across brands. Error bars give $\pm$ two standard errors.

A later paper estimated similar effects in Bing search ads. They found that, when competing brands buy ads on a focal firm's branded keywords, sponsored search advertising defends traffic that would not otherwise get to the organic result link. The effects were pretty big. The eBay result did not generalize to companies whose competitors bought their own-branded keyword ads.

Did Other Firms Learn from eBay?

Firms' Reactions to Public Information on Business Practices: The Case of Search Advertising*

Justin M. Rao
HomeAway, Inc.

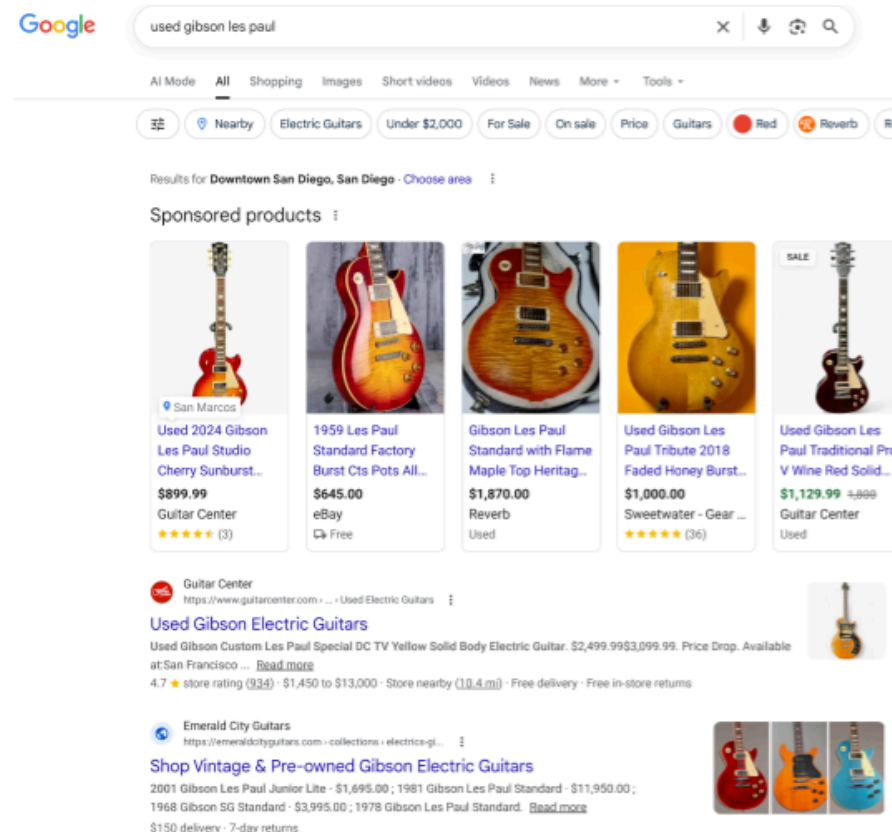
Andrey Simonov
Columbia University

February 24, 2018

We use five years of bidding data to examine the reaction of advertisers to widely disseminated press on the lack of effectiveness of brand search advertising (queries that contain the firm's name) found in a large experiment run by eBay (Blake, Nosko and Tadelis, 2015). We estimate that 11% of firms that did not face competing ads on their brand keywords, matching the case of eBay, discontinued the practice of brand search advertising. In contrast, firms did not react to the information pertaining to the high value and ease of running experiments—we observe no change in the experiment-like variation in advertising levels. Further, while 72% of firms had sharp changes in advertising suitable for estimating causal effects, we find no correlation between firm-level advertising effects and the propensity to advertise in the future. We discuss how a principal-agent problem within the firm would lead to these learning dynamics.

A second follow-up study estimated how Bing advertisers changed their advertising policies after the eBay study was publicized. It found that advertisers largely either (a) maintained the status quo, or (b) stopped advertising entirely. However, advertisers did not start running more experiments. (Why not?)

eBay Case Study: Key Takeaways



eBay taught us that correlational advertising measurement is questionable, and that firms should use experiments to measure causal advertising effects. However, most companies were not ready for that message yet. This 2026 screenshot shows that eBay lost its organic search engine results page (SERP) real estate and started buying Google search ads again. As it should do when ads are profitable.

“Grading Your Own Homework”



Why didn't most advertisers get the right message from eBay? A likely culprit: Textbook principal/agent problems. Today, more marketers have internal agencies, better data, and better capacities to run experiments. It may help if advertising measurement team reports to CFO.

2024 Ad Measurement Trends

The Ad Measurement Trends That Reshaped Online Advertising This Year



By James Hercher



MONDAY, DECEMBER 30TH, 2024 - 12:55 AM

SHAR

2024 was a year of hectic change for ad measurement.

Third-party cookies may have been given a reprieve by Chrome, perhaps even an indefinite lifeline. But, still, user-level data is running dry, to the point that last-click and multitouch attribution have lost their edge entirely.



Stop Setting Money On Fire

Incrementality measurement

Media buyers were consumed by “curation” mania this year. But for ad measurement, 2024 was the breakout year of “incrementality,” a hard term to define.

“I really need a better answer to this question,” Olivia Kory, head of strategy at the incrementality measurement startup Haus in an [AdExchanger Talks](#) podcast this month, when asked what actually is incrementality measurement.

She sums it up as **a marketing measurement model that is geared toward establishing causation, rather than correlation.**

Incrementality measurement achieves this through the sophisticated use of holdout groups and geo-testing. One way to benchmark Instagram’s incremental contribution, say, or that of a large DOOH campaign, is to run that campaign in some markets while not serving those ads at all in other, similar markets.

Mix modeling

Call it vintage chic, because MMM is back.

Not so long ago, data-driven marketers would have scoffed at the idea of a reversion to MMM. It’s an old-school method of campaign measurement built for TV, radio and print, and which takes *months* to establish results.

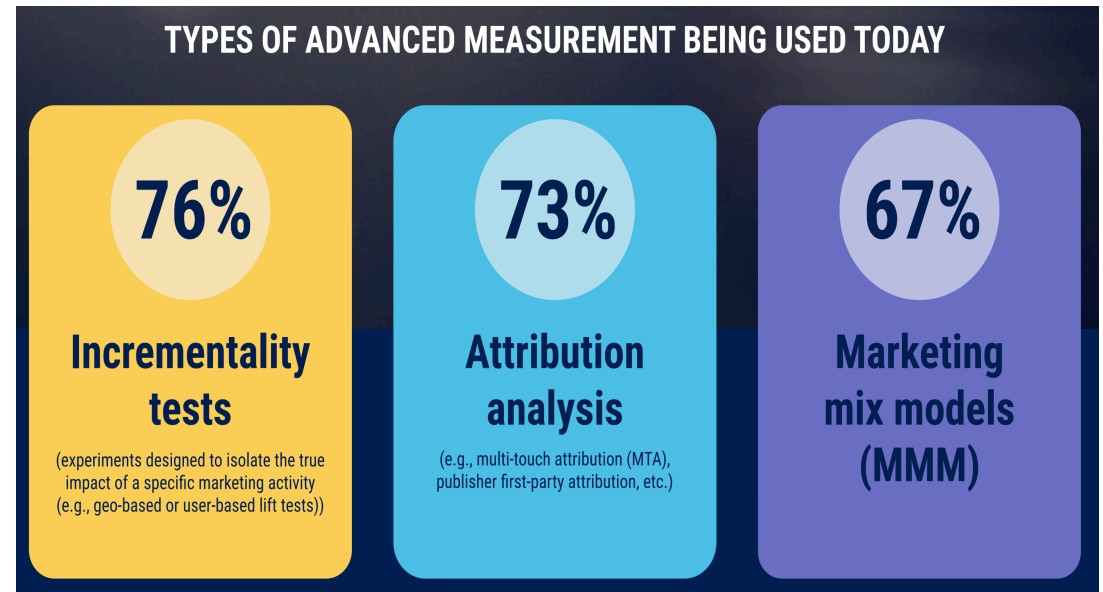
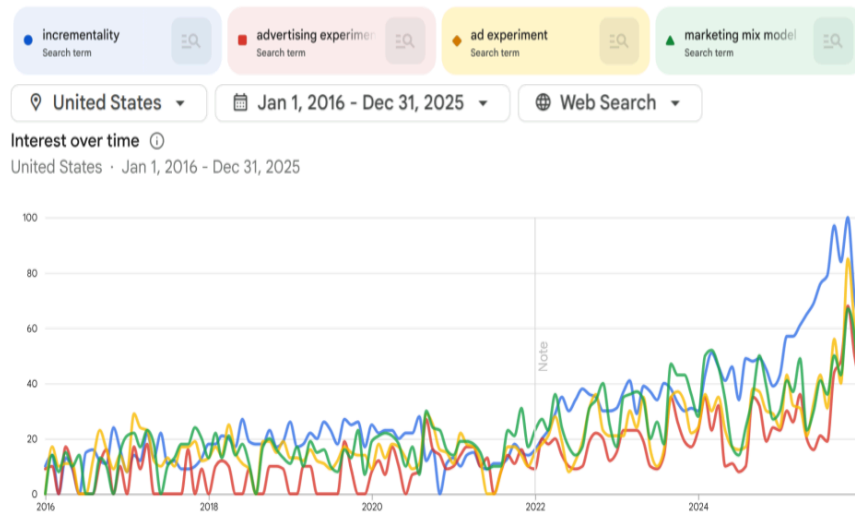
But **with user-level data running dry and walled gardens hoovering up all the ad demand, MMM becomes a feasible way to attribute platforms as a whole, without having visibility into the platform itself.**

The platforms have heard the MMM requests, and answered in 2024.

Google this year launched Meridian, its open-source MMM service. Meta already had one, called Robyn.

Incrementality & marketing mix models (MMM) were trends #1 & #2; the only other trend was e-commerce metric proliferation.

Ad Experiments in Google Trends & Buy-side Surveys



We're a few years into a generational shift. Smaller, independent ad agencies are making the most noise about incrementality. However, the correlation between advertising and sales ($\text{corr}(\text{ad}, \text{sales})$) is not going away. Union(correlations, experiments) should exceed either alone.

Google Trends | IAB (2026)

Fundamental Problem of Causal Inference

- Why causal effects are estimable but not directly observable

Causal Inference Framework

- Suppose we have a binary “treatment” or “policy” variable T_i that we can “assign” to person i
 - Examples: Send an ad, Serve a webpage, Recommend a product
- Suppose person i could have a binary potential “response” or “outcome” variable $Y_i(T_i)$
 - Marketing funnel examples: Visit site, open app, search products, enter email, add to cart, purchase
 - “Treatment” terminology came from medical literature; Y could be patient outcome
- Important: Y_i may depend fully, partially, or not at all on T_i , and the relationship may differ across people
 - Person 1 may buy due to an ad; person 2 may stop buying due to an ad

Why Care About Causal Effects?

- We want to maximize profits $\Pi = \sum_i \pi_i(Y_i(T_i), T_i)$, where π_i is person i 's profit contribution
- Suppose $Y_i = 1$ contributes to revenue; then $\frac{\partial \pi_i}{\partial Y_i} > 0$
- Suppose $T_i = 1$ has a known cost, so $\frac{\partial \pi_i}{\partial T_i} < 0$
- Effect of $T_i = 1$ on π_i is $\frac{d\pi_i}{dT_i} = \frac{\partial \pi_i}{\partial Y_i} \frac{\partial Y_i}{\partial T_i} + \frac{\partial \pi_i}{\partial T_i}$
- We have to know $\frac{\partial Y_i}{\partial T_i}$ to optimize T_i assignments
 - Called the “treatment effect” (TE); can be approximated by $Y_i(T_i=1) - Y_i(T_i=0)$
- Profits may decrease if we misallocate T_i
 - E.g., buy ads targeting people with inefficiently low response rates

The Fundamental Problem

- 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”
 - Causality is a missing-data problem that we cannot fully resolve. We only have one reality
 - ⇒ We can build models to help compensate for missing counterfactuals

The Fundamental Problem of Causal Inference: We cannot directly observe counterfactual outcomes. Therefore, we cannot directly compare $Y_i(T_i = 1)$ to $Y_i(T_i = 0)$ to measure the treatment effect on person i .

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, effort; 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, effort; 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_i is only randomly assigned when we run an experiment, so what exactly are we doing here?
 - Are we paying our demand-side platforms (DSPs) to distribute our ads randomly?
4. **Fuhgeddaboutit,** do not measure
 - Some advertisers do this
 - Measurement is costly; may be a net negative when not possible to do well

Sophisticated companies usually combine 1, 2 and 3

How Much Does Causality Matter?

- Are organizational incentives aligned with profits?
- Data thickness: How likely can we get a good estimate?
- Organizational analytics culture: Will we act on what we learn? Or are we just trying to get a number to negotiate the budget?
- 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. The value of causal measurement, and danger of correlational measurement, depends on whether the organization will act on what it learns.

Advertising Measurement, Defined

- What we measure, challenges, classic eBay measurement case

Measurement of What?

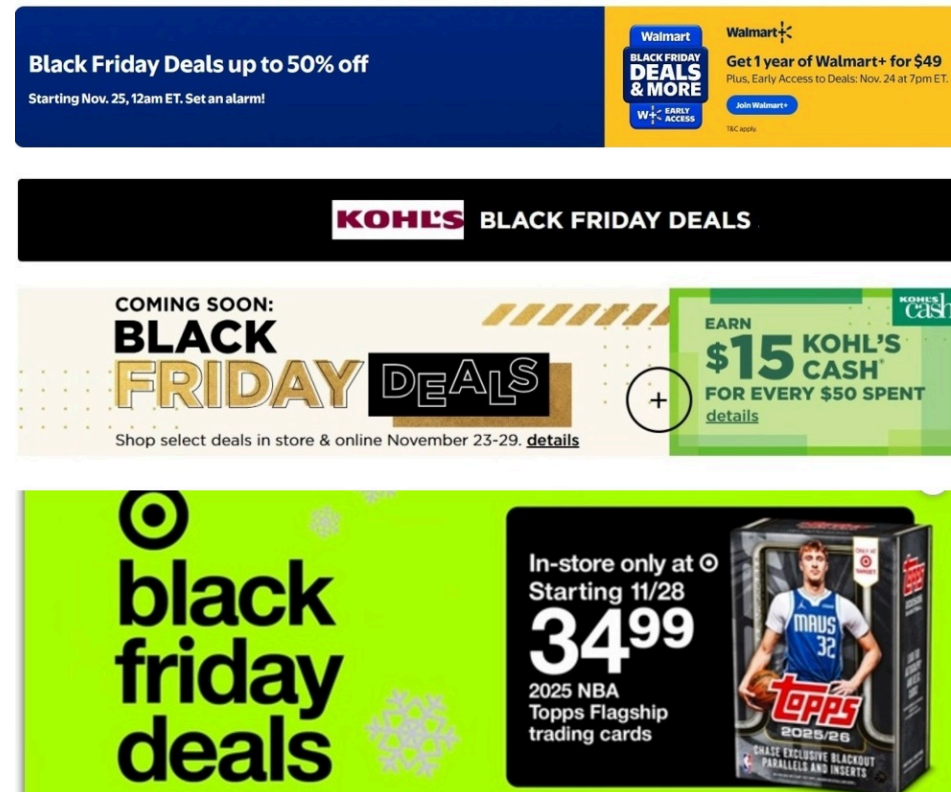
advertising noun

ad·ver·tis·ing ('ad-vər-, tī-zɪŋ ◀▶)

[Synonyms of advertising >](#)

- 1 : the action of calling something to the attention of the public especially by paid announcements
- 2 : **ADVERTISEMENTS**
| a magazine full of *advertising*
- 3 : the business of preparing **advertisements** for publication or broadcast
| looking for a job in *advertising*

Many people use 'advertising' to refer to any commercial speech. In marketing, 'advertising' refers to paid media, as distinct from owned media (e.g., organic social, website, emails, direct mail) & earned media (e.g., reviews, news stories). Paid media implies that a 'publisher' generated the advertising opportunity by attracting consumer attention; sells the ad; and may constrain the advertiser's message, to maintain its own relationship with the consumer.



- **Brand advertising:** Campaigns designed to generate awareness, create associations, change attitudes and stimulate long-run response. Believed to increase 'mental availability' to increase choice probability in the consumer's next choice occasion. Testable prior to launch
- **Performance advertising:** Campaigns designed to stimulate short-run measurable response, especially visitation and/or sales. Highly amenable to experimentation

Most large advertisers run both brand and performance campaigns; many believe they work better together. But, brand and performance compete for budget (both internal teams, and external agencies), and sometimes denigrate each other. Some people argue the distinction is artificial, we should test ads on both long-run & short-run metrics.

Ad Measurement

- *Advertising measurement* quantifies ad delivery, exposure and outcomes to improve advertising efforts
 - We focus on sales/conversions, as they matter most
 - Brand campaign measurement usually focuses on cost and intermediate metrics such as attention and brand attitudes. General principles include independence and transparency in measurement; these must be checked, as they cannot be assumed
- Advertising measurement is hard because ad effects depend on ad content, context, timing, targeting, current market conditions, ad prices, past advertising & past outcomes—all of which change
 - Shooting at a moving target
- Advertising measurement is expensive: must *directly* inform future choices

Bruner (2025)



What Do We Measure?

Often, Return on Advertising Spend (ROAS)

$$\frac{\text{Revenue Attributed to Ads}}{\text{Ad Spending}} \text{ or } \frac{\text{Revenue Attributed to Ads} - \text{Ad Spending}}{\text{Ad Spending}}$$

Increasingly, we report incremental ROAS (iROAS) if we have causal identification, i.e. we isolated causal ad effects

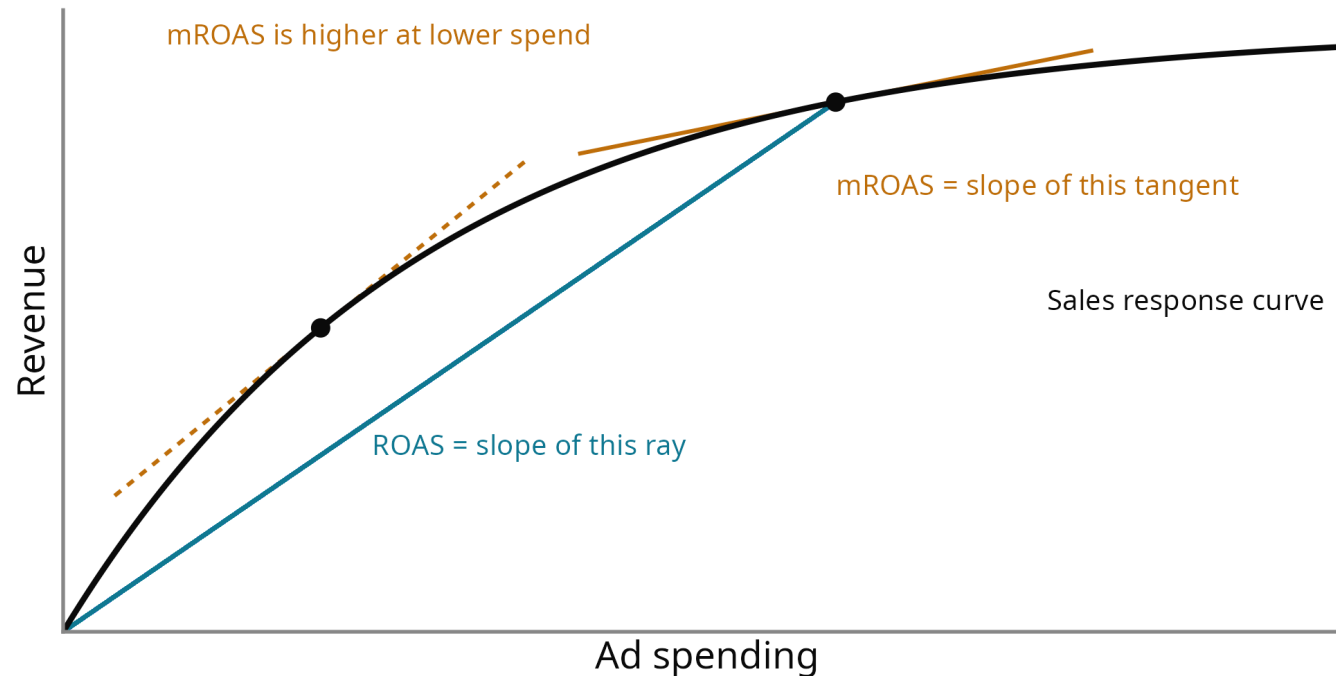
- ROAS $\neq$ iROAS because attribution is usually correlational

We also should measure delivery and funnel-wide key performance indicators (KPIs), e.g. brand metrics, visits, add-to-cart, sales, revenue, ...

- We usually get economies of scope in measurement

Brand Lift Tests measure ad effects on brand attitude surveys, but these are often underpowered. CLV/CAC (customer acquisition cost) is increasingly common within subscription businesses

Diminishing Returns



In theory, we buy the best ad opportunities first, so increasing spend should lower marginal returns (“saturation”). Marginal ROAS (mROAS) is the tangent to the curve. Nonlinearity means $ROAS \neq mROAS$. We use ROAS for overall evaluation, and mROAS for budget reallocation. The common adage to “max your ROI” usually leaves money on the table. (Why?) An optimal budget allocation equalizes mROAS across channels. (Why?) What level of ad spending actually maximizes mROAS?

Albertsons: ROAS Varies with Measurement Choices

+ Analysis from 573 campaigns from Albertsons Media Collective showing how changes in ROAS methodology change results
+ Important questions for advertisers to use to drive more transparent measurement conversations with their partner RMNs

	Methodology		Average Shift in ROAS
Household vs. Customer Sales Attribution	Household	→ Customer	25%
Product Set Attribution	Umbrella Brands Halo	→ Brand Halo	35%
Untraceable Sales	Extrapolated	→ Only Traceable Sales	37%
Impression Type	Served Impressions	→ IAB Viewable Impressions	5%
Total Average Impact			63%
Quartile 1			52%
Quartile 3			74%

Given shortfalls in ROAS, the industry must shift towards incremental ROAS (iROAS) to better measure true advertising impact. Our future work will aim to bring a similar understanding to iROAS methodology and provide tools for advertisers.

Albertsons media group reported a meta-analysis of campaigns showing that correlational ROAS results strongly depend on intermediate measurement choices. In a follow-up study, the same authors stress-tested incrementality (iROAS) measurement and found that methodology choices alone produced 6.5x average variation in iROAS within the same campaign, with 83% of campaigns able to flip sign. What does this imply about hiring black-box vendors vs. doing your own measuring?

Correlational Advertising Measurement

- Lift statistics, multi-touch attribution, marketing mix models

Correlational Ad Measurement

- Correlational advertising measurement is defined by the absence of an identification strategy to isolate causal advertising effects from confounding drivers of sales
 - Equivalently, by the assumption (usually implicit) that past ads were distributed randomly
 - Or, by the belief that we should maximize sales attributed to advertising
- Most common approach historically, but many marketers have adopted incrementality over the past two decades

Corr(ad,sales) is defined by the data, not by the methods, but we will review some of the most common methods used within this paradigm

1. Lift Statistics

Compare conversion rates between people exposed to ads and people not exposed to ads

$$\frac{Prob.\{Conv.|Ad\}}{Prob.\{Conv.|NoAd\}} \quad \text{or} \quad \% \text{ Lift: } \frac{Prob.\{Conv.|Ad\} - Prob.\{Conv.|NoAd\}}{Prob.\{Conv.|NoAd\}}$$

- E.g., if ad-exposed users convert at 0.6% and non-exposed at 0.4%,
Lift Ratio = 1.5, % Lift = 50%

The name 'Lift' implies a causal ad effect, but lift statistics can only be incremental when the data contain a treatment/control analogue. Otherwise lift stats encompass all differences between ad-exposed and non-ad-exposed consumer groups, including ad targeting, context, timing, recent behaviors and platform usage, as well as ad effects. Lift stats are easy to compute and communicate, but often misunderstood as causal.

2. Multi-Touch Attribution (MTA)

- Get individual-level data on every touchpoint for every purchaser
 - Should include earned media, owned media & paid media (ads, paid influencer & affiliate)
- Choose an attribution rule (First-touch, last-touch, fractional, Shapley)
- MTA algorithm searches for touchpoint parameters that best-predict conversions given the rule
 - Credit then informs future budget allocations across touchpoints
 - MTA assumes touchpoints *solely* drive conversions & disregards nonpurchasers
- Advertiser-side MTA arose from the open web display market, linking tracking cookies to sales. Has challenges integrating walled gardens due to privacy rules and platform reporting limitations. Some advertiser MTAs live on, but some are zombies. Large platform-side MTA will remain viable and efficient, though limited to data within each walled garden; can advertisers trust/verify?

Amazon Ads MTA combines experiments, machine learning and shopping signals.

3. Marketing Mix Models (MMM)

- The “marketing mix” consists of the 4 P’s. A “marketing mix model” (MMM) typically uses marketing mix variables to explain sales
 - Idea goes back to the 1950s; e.g., suppose we increase price & ads at the same time; what happens to sales?
 - When possible, MMM should include competitor variables also
- A “media mix model” (mMM) relates sales to ads/marketing communications (marcom) channels
 - MMM and mMM share many attributes and techniques
- MMM goal is to evaluate past marketing ROAS by channel and inform future budgeting decisions

Pioneering works: [Magee \(1953\)](#), [Weinberg \(1956\)](#), [Vidale & Wolfe \(1957\)](#), [Little \(1972\)](#)

MMM Components

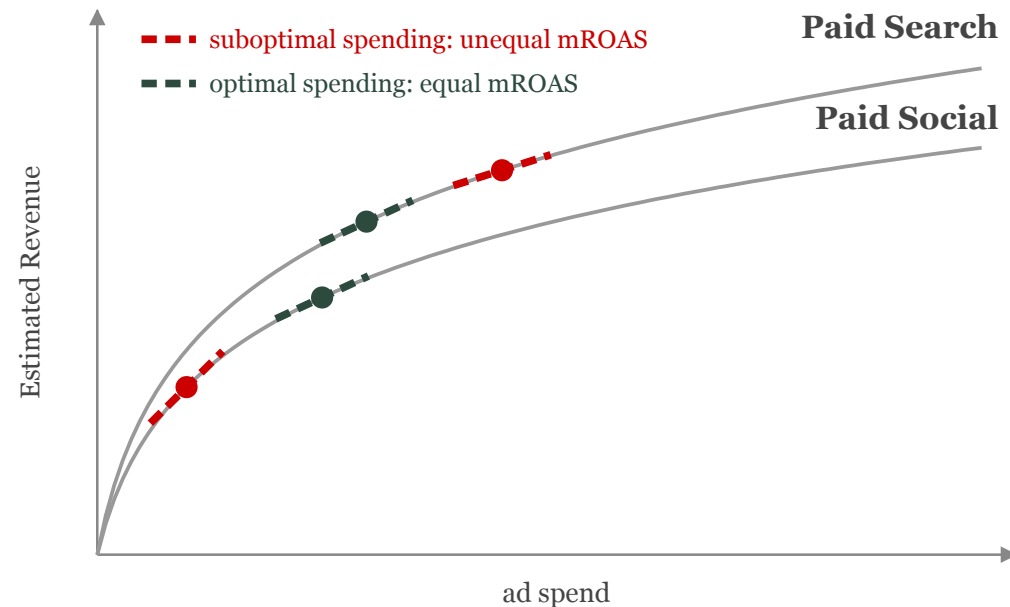
- We usually estimate MMM with 1-5 years of weekly or monthly data, sometimes across a panel of geographic markets
 - Aggregate data are privacy-compliant & do not usually require platform participation
- Main predictors are ad spend or exposures in each channel, assuming diminishing marginal returns (“saturation”) and possibly long-lasting ad effects (“carryover”); outcome is usually sales units, volume or revenue
- MMM often controls for (a) time trends, (b) seasonality, (c) macroeconomic factors, (d) observable demand shifters, (e) competitor marketing
- Outputs include ad elasticities, correlational mROAS estimates, counterfactual budget reallocations

In theory, MMM could be highly granular, such as hours and census blocks. However, compromises are required to balance data measurement intervals, refresh rates, accuracy and integrability.

[Johnson et al. \(2017\)](#) meta-analyzed 432 display ad experiments, finding carryover could be positive, zero or even negative

Using mROAS to Reallocate Spending

- MMM fits a sales response curve per channel; mROAS is the slope at current spend, i.e., revenue from the next ad dollar
- ROAS maximization requires harmonizing mROAS across channels
 - Estimation error makes this an uncertain exercise



- Theory says: Max profits by spending until $\text{avg. cont.} * \text{mROAS} = 1$ in all channels, but practice is more complicated than that

MMM Considerations

- MMM requires sufficient independent variation in each ad spend predictor, else coefficient estimates will be imprecise due to collinearity
 - Deliberate variation in ad spending, e.g., inversely-correlated pulses across channels, time & geos, helps MMM separate each channel's contribution
- “Model uncertainty”: Results can be strongly sensitive to modeling choices, so we usually evaluate multiple models to gauge sensitivity
- MMM results are correlational without experiments or quasi-experimental identification
 - Correlations can be unstable; Bayesian estimation can help regularize
 - MMM results can be calibrated using causal measurements more later
- Advertising media can interact, for example when TV ads generate branded search queries due to media multitasking, which then lead to more search ad clicks

Google [Meridian](#) offers Bayesian estimation, hierarchical geo-level modeling, reach & frequency data, experiment-informed ROI priors and a budget-reallocation optimizer. Other open-source frameworks: [BayesianMMM](#), [mmm_stan](#), [PyMC-Marketing](#), Meta [Robyn](#); data generator: [siMMMulator](#)

How much should we rely on Correlational Ad Measurement?

- Pros & Cons, Research, Explaining Persistence

Steel-manning $\text{Corr}(\text{ad}, \text{sales})$

- $\text{Corr}(\text{ad}, \text{sales})$ 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 marketing
 - E.g., Direct response TV ads for 1-800 phone numbers get 0 calls without TV ads, so we know the counterfactual (what is it?)
 - New Shopify stores offering copycat products usually sell nothing without marketing
- 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; maybe ads got shown to the most loyal customers

Famous books present descriptive evidence about how brands have grown, then extrapolate to prescriptive “laws” about how marketers should act. Parsimonious advice can be appealing, but has been called pseudo-science

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

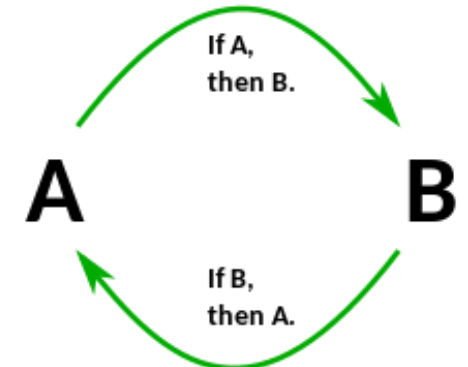
- Advertisers try to optimize ad campaign decisions
 - E.g. ads for surfboards in coastal cities, not landlocked cities
- If ad optimization increases ad response, then $\text{corr}(\text{ad}, \text{sales})$ will confound actual ad effect with ad optimization effort
 - More ads in San Diego, more surfboard sales in San Diego. But would we have 0 sales in SD without ads? If no, then we need a counterfactual prediction
 - $\text{Corr}(\text{ad}, \text{sales})$ usually overestimates the causal effect, encourages overadvertising at the expense of profits

Google's Chief Economist [explains](#) in greater detail.

Problem 2 with Corr(ad,sales)

- How do most marketers set ad budgets? Top 2 ways historically:
 1. Percentage of sales method, e.g. 1%, 3% or 6%
 - Ads:sales ratios are often measured for benchmarking
 2. Competitive parity
 3. ...others...
- Do you see the problem here?

This problem is called simultaneity ([Bass 1969](#)).



Problem 3 with $\text{Corr}(\text{ad}, \text{sales})$

- Leaves marketers powerless vs ~~big~~ colossal ad platforms
- Platforms withhold data and obfuscate algorithms
 - How many ad placements are incremental?
 - How many ad placements target likely converters?
 - How can advertisers respond to adversarial ad pricing?
- Have ad platforms ever left ad budget unspent?
 - Would you, if you were them?
 - If not, why not? What does that imply about incrementality?
- The only way to balance platforms' pricing power is to 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.	)	Case No. 20-cv-3010 (APM)
GOOGLE LLC,	)	
Defendant.	)	

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.”).

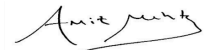
264. Google therefore endeavored to raise prices incrementally, so that advertisers would view price increases as within the ordinary price fluctuations, or “noise,” generated by the auctions. See, e.g., UPX507 at .023 (describing a 10% CPC increase as “safe” because it is “within usual WoW noise”); 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

This was written by a federal judge who heard mountains of evidence on both sides. Judge Mehta describes Google’s efforts to hide price increases from advertisers, based on internal documents.

U.S. v. Google (2024)

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.

Kellogg faculty and Meta data science collaborated to analyze Meta’s large trove of advertising experiments. Their main research question: Can we estimate causal advertising effects on sales by applying machine learning models to advertising treatment data alone? I.e., can we recover true causal estimates without non-advertising control condition data?

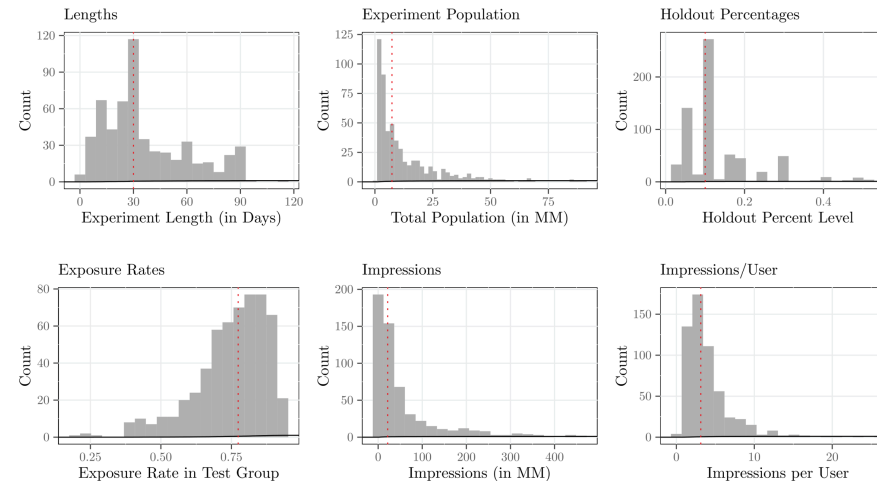
3.1. Experiment Selection

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.¹³ For each experiment, we selected all outcomes with at least 5,000 conversions in the test group.¹⁴

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	N	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	N	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

The setting was auspicious. Machine learning methods work best when applied to thick data with numerous predictors, as is the case in Facebook data. Additionally, Facebook served most ads from content servers to facilitate consistent measurement and reduce ad-blocking.

Figure 2. ATTs Across All RCTs

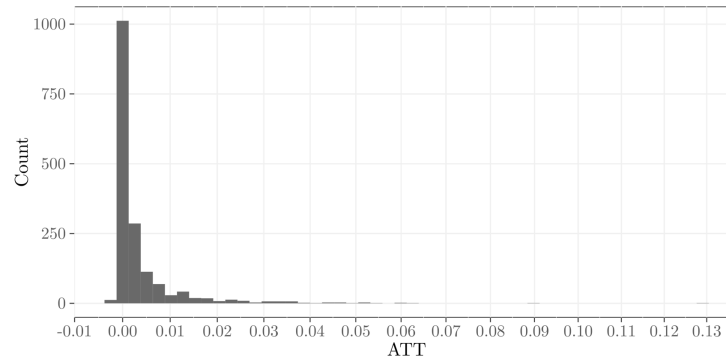
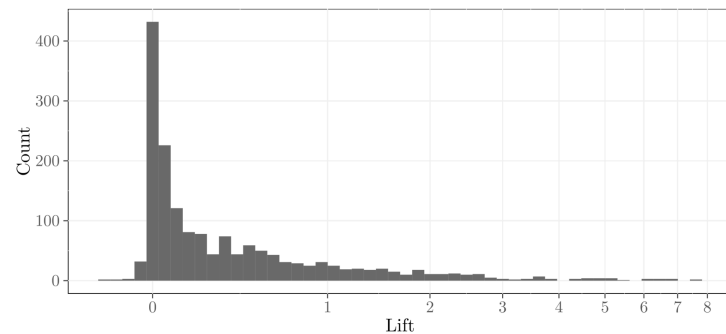


Figure 3. Lifts Across All RCTs

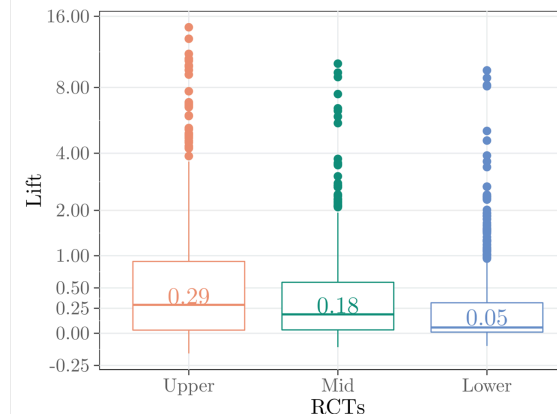


Note. Excludes the top 1% of lifts.

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,

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

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



Most ad experiments show causal ad effects on conversions of 0-0.25%, with median lift ratios of 0.05-0.29. Ads had clearer effects on upper-funnel actions (e.g., shopping) than on lower-funnel actions (e.g., purchase), as price or other factors can discourage purchases

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.

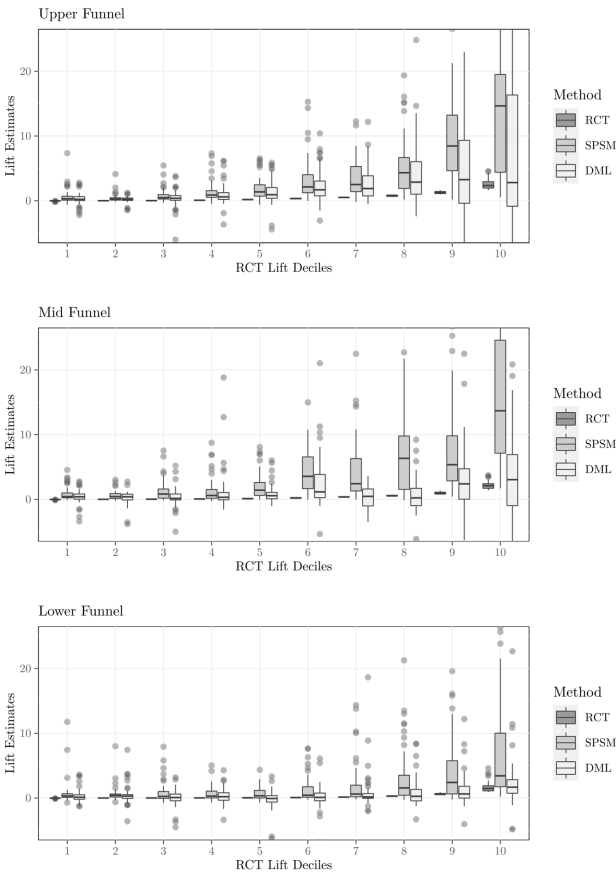
In standard propensity score matching, we find the one (or more) unexposed users with the closest propensity score to each exposed user to estimate the treatment effect. Since this is very computationally burdensome, instead, we stratify on the propensity score: After estimating the propensity score, $\hat{e}(X_i)$, 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.

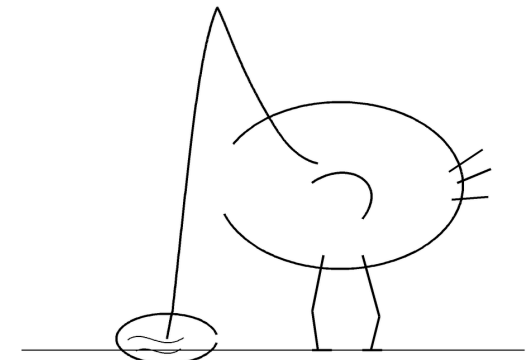


Note. Figures excludes the top 1% lifts for each position in the purchase funnel.

Both Machine Learning frameworks tested failed to recover true incremental ad effects. The correlational advertising effects were mostly overestimated, but not always. This offers strong empirical evidence that models alone cannot substitute for causal identification strategies. Causality is a “data problem,” not a “modeling problem.”

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, without other marketing channels, this may be good logic
 - However, premise implies deeper problems, i.e. need to diversify marketing efforts and find cheaper sources of sales
 - Plus, we can run experiments without setting ads to zero, e.g. test 50% vs. 150%
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 (key word: “might”)
 - But direction is only part of the picture; what about effect size?
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
 - Should ad measurement team report to CFO, CMO, or both?



Why Are Some Teams OK with Corr(ad,sales)?

4. Managing analytics well requires skill and discipline

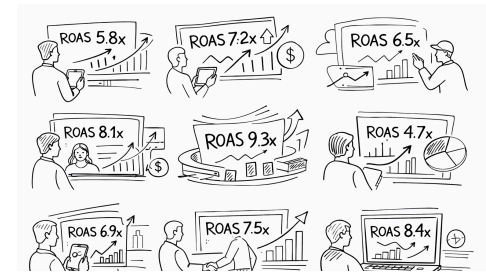
- Managers must integrate correlational and causal analyses when making decisions
- Analysts must have causal inference skillsets
- Organization must tolerate failure in search of data-driven incremental improvements
- How many shops go back and check their ex-ante predictions? What are the internal incentives for accuracy?

5. Platforms often provide correlational ad/sales estimates

- Which are larger, correlational or experimental ad effect estimates?
- Which one might many client marketers prefer?
- “Nobody ever got fired for buying [___].” [Amazon, Google, Meta]

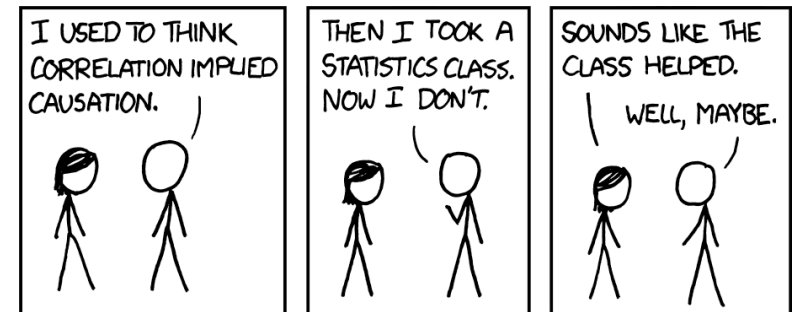
6. Historically, agencies usually estimated ROAS

- Agency compensation usually relies on ad spending, not incremental sales; principal/agent problems are common
- These days, more marketers have in-house agencies, and split work



Causal Advertising Measurement

- Experimental designs, necessary conditions, quasi-experiments



Causal Ad Measurement

- Causal advertising measurement is defined by the presence of an identification strategy to isolate causal advertising effects from confounding drivers of sales
 - Valid experiments create identification by design
 - We can analyze as-if-random variation, AKA quasi-experiments
 - We often use models to predict credible counterfactuals, but models vary in the plausibility of their identification strategies

In science, *Causal* means we isolate the treatment effect from known confounds *and* from unknown confounds. Often misinterpreted as evidence consistent with a hypothesis, a much lower bar which is prone to motivated reasoning

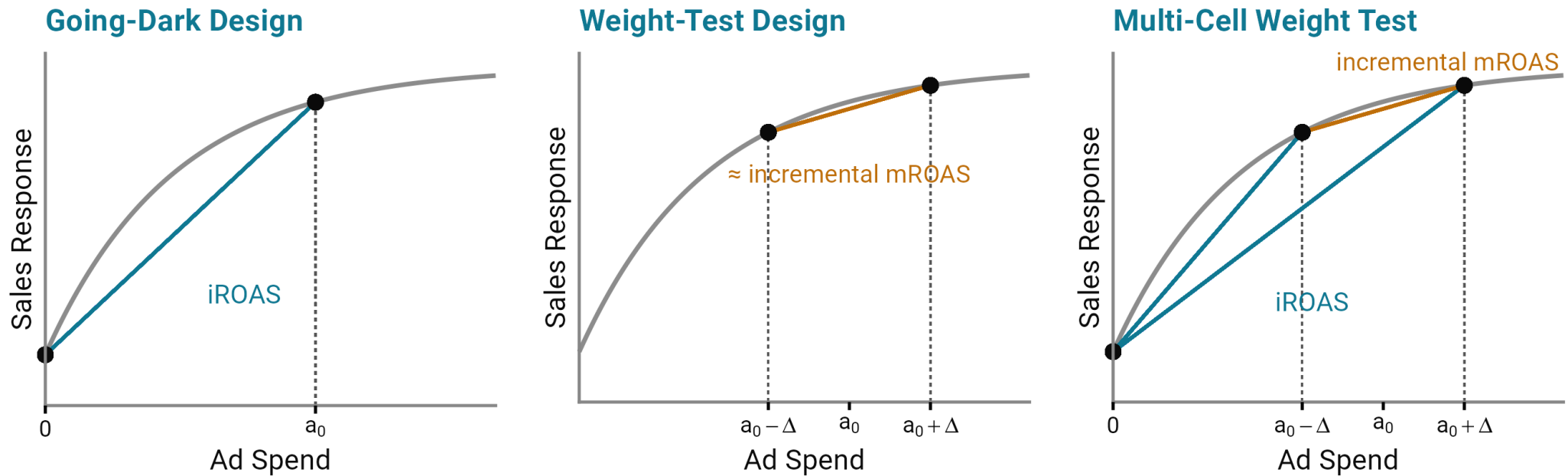
Popular ad experiments

1. Going-dark design. Randomly withhold ads. Estimates iROAS
2. Weight test: Randomize spending between high-spend and low-spend treatments. Estimates incremental mROAS
3. Randomize messages within a campaign
 - Nominally free; Mine competitor messages in [ad libraries](#) for ideas
 - Advertising pros often believe that creative messages are first-order drivers of ad effects (e.g., [Circana 2023](#), [Kantar 2024](#), [Magna+Yahoo 2025](#))
4. Randomize budget across platforms, publishers, behavioral targets, contexts

Geo-experiments randomize treatments across places (e.g., cities) and time periods. Consumer-level experiments usually randomize treatments across devices, accounts or cookies. Geo-Experiments have been around forever, but often return inconclusive results. Consumer-level randomization became widespread with the internet, but it too may yield inconclusive results. [Lewis and Rao \(2015\)](#) taught us that the challenge in both cases involves low response rates and noisy environments

[Recast \(2026\)](#) uses simulation to evaluate popular geo-experiment packages, including its own

Different Designs Estimate Different Things



Going-dark design estimates iROAS: the causal effect of all observed advertising vs. zero advertising. It quantifies campaign performance overall, but cannot say whether spending level was optimal

Weight-test design estimates (approximate) incremental mROAS: can inform spending decisions, but does not focus on overall profitability

Multi-cell weight tests can estimate both iROAS and incremental mROAS, but spread sample across more treatments

The challenge is to estimate both the height and the curvature of the response curve

Recall that MMM can estimate correlational mROAS; incremental mROAS is the causal counterpart

Waisman and Gordon (2025)

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:
May be 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 regardless of outcome

3. Compliance

- Treatments assigned are treatments received
- Ad blocking can threaten compliance; we usually resolve by estimating Intent-to-Treat effects

4. Statistical Independence

- Random assignment of treatments to units. “Balance tests” help to check
- Ad platforms usually distribute ads nonrandomly, threatening random assignment

List (2024); Kohavi et al. (2020)

Before You Kick Off Your Test...

- Run A/A test before your first A/B test. Validate the infrastructure before you rely on the result
 - Shows the effect size a given test duration is powered to detect
 - Shows whether random assignment is working, as it's sometimes coded incorrectly; [Kelly \(2025\)](#) explains how randomization is easy to mess up, e.g. assignment based on nonrandom factors (user IDs, time) and pseudorandom generator misuse
- Can we agree on the opportunity cost of the experiment? “Priors”
 - Confident execs argue control-condition exposures reduce profits due to foregone advertisements
- 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 iROAS at 1.5 with confidence interval (c.i.) [1.45, 1.55]. Or, suppose we estimate iROAS at 1.5 with c.i. [-1.1, 4.1]. What actions would follow each?

Platform Experiments Advisory

- Meta offers two experimentation frameworks:
 - Lift tests randomize users to ad eligibility against a no-ad holdout—a valid intent-to-treat estimate of ads eligibility
 - A/B tests randomize users across campaign configurations with no holdout. The issue is that an ad delivery algo sits between configuration and delivery, so the configurations effectively “treat” the delivery algorithm and violate the user compliance assumption—a confound called “divergent delivery”
- Some platforms require advertisers meet minimum spend levels to use on-platform experimentation tools. You can roll your own experiments by randomizing ad budget across time and targeting criteria
- **Ghost ads** is an ingenious system to randomly withhold ads from auctions and maximize experiment efficiency

[Haus \(2025\)](#) | [Braun & Schwartz \(2025\)](#) | [Burtch et al. \(2025\)](#)

Productive Experiments...

- Serve customer interests
 - Working against customers drives customers away
- Live within theoretical frameworks
 - We require hypotheses if we want to learn from tests
- Test quantifiable hypotheses
 - Choose test size & statistical power based on hypothesis
- Analyze all relevant customer metrics
 - Test positive & negative metrics, e.g. conversions & bounce rates
 - Test short-run & long-run metrics, e.g. trial & repurchase
- Acknowledge possible interactions between variables
 - E.g. price advertising effects will always depend on the price

Zumsteg (2022), *Unchecked AB Testing Destroys Everything It Touches*

Problems with Ad Experiments

1. Low power, noisy estimates -> limited decision guidance

- Sales volatility often swamps true ad effects ([Lewis & Rao 2015](#))
- [Feit & Berman \(2019\)](#) choose test size to maximize expected profit

2. Low external validity / generalizability, transportability

- A finding depends heavily on its context, and may change dramatically when the context changes ([Shapiro, Hitsch & Tuchman 2021](#))

3. No true control

- If an experiment changes ads, and ads change demand, then demand may change optimal recommendations, prices, reviews, inventory, competitor actions or other market signals, which then affect both treatment and control cells ([Sahni 2016](#), [Blake & Coey 2014](#), [Du et al. 2019](#))
- **Identity fragmentation** means a user may be in both cells on different devices ([Lin & Misra 2022](#))
- In going-dark designs, carryover effects from pre-experimental campaigns can bias control-cell sales upwards

[Johnson \(2023\)](#) reviews best practices for digital advertising experiments.

Problems with Ad Experiments

4. Silent failure of other conditions

- Ad fraud, ad blocking, non-visibility & non-attention: All cause noncompliance & distort analyses ([Johnson 2023](#))
- Divergent delivery ([Braun & Schwartz 2025](#))

5. Costs & biases: Large experiments are too costly to run continuously

- Opportunity cost of nontargeted consumers ([Ghost ads](#) as potential remedy)
- Time delay to decision relevance
- Experiments may bias us toward short-term effects of performance ads on certain platforms
- Platforms often intervene to maximize total ad-linked conversions, rather than maximizing the incremental effect of ads on conversions ([Haus 2025](#), [Braun & Schwartz 2025](#))

Quasi-experiments Vocabulary

- Model: Mathematical relationship between variables that simplifies reality, e.g. $y = x\beta + \varepsilon$
 - May be estimated using either experimental data, non-experimental data, or combination
- 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
 - Compare apples with apples, not apples with oranges
 - May be baked into a model, or may stand alone, e.g. A/B test comparison of means
- Popular quasi-experimental techniques: Difference-in-differences, regression discontinuity, instrumental variables, synthetic control, matching. Each technique predicts what counterfactual would have occurred without treatment
- We “identify” the causal effect when our identification strategy reliably distinguishes $\frac{\partial Y_i}{\partial T_i}$ from possibly correlated unobserved factors
- If you estimate a model without a valid identification strategy, results are correlational AKA descriptive
 - Common misconception: Identification is a property of a model. Correction: It’s a logic of comparison, whether embedded within a model or not

Sant’Anna (2026) discusses difference-in-differences theory and code.

Diff-in-Diffs Helped Identify Cholera Cause

In the 1850s, an English doctor named John Snow suspected that cholera spread via food and drink, rather than the popular theory of airborne transmission. Snow realized a natural experiment offered identification.

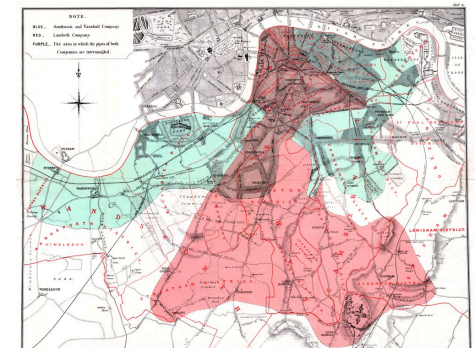
Some London neighborhoods were served by multiple water companies. One company, Lambeth, moved its intake pipes higher up the Thames to obtain cleaner water, whereas its competitor Southwark and Vauxhall (S&V) kept its downstream intake.

Snow went door to door to count customers who subscribed to each water company. He also matched those households' records against the city's mortality records to calculate cholera death rates by water company over time. In 1849, death rates were 85 per 100k Lambeth customers and 135 per 100k S&V customers. In 1854, after the water intake change, death rates were 19 per 100k Lambeth customers, and 147 per 100k S&V customers.

Assuming household cholera risk factors were unrelated to water company choice, then the change in S&V death rates estimated the counterfactual change for Lambeth, showing that cleaner water meaningfully reduced cholera deaths. This discovery came before the germ theory of disease (1860s) and modern experimental methods. (What are the two diffs?)

This exemplifies an identification strategy without a complicated mathematical model

Cunningham (2026)



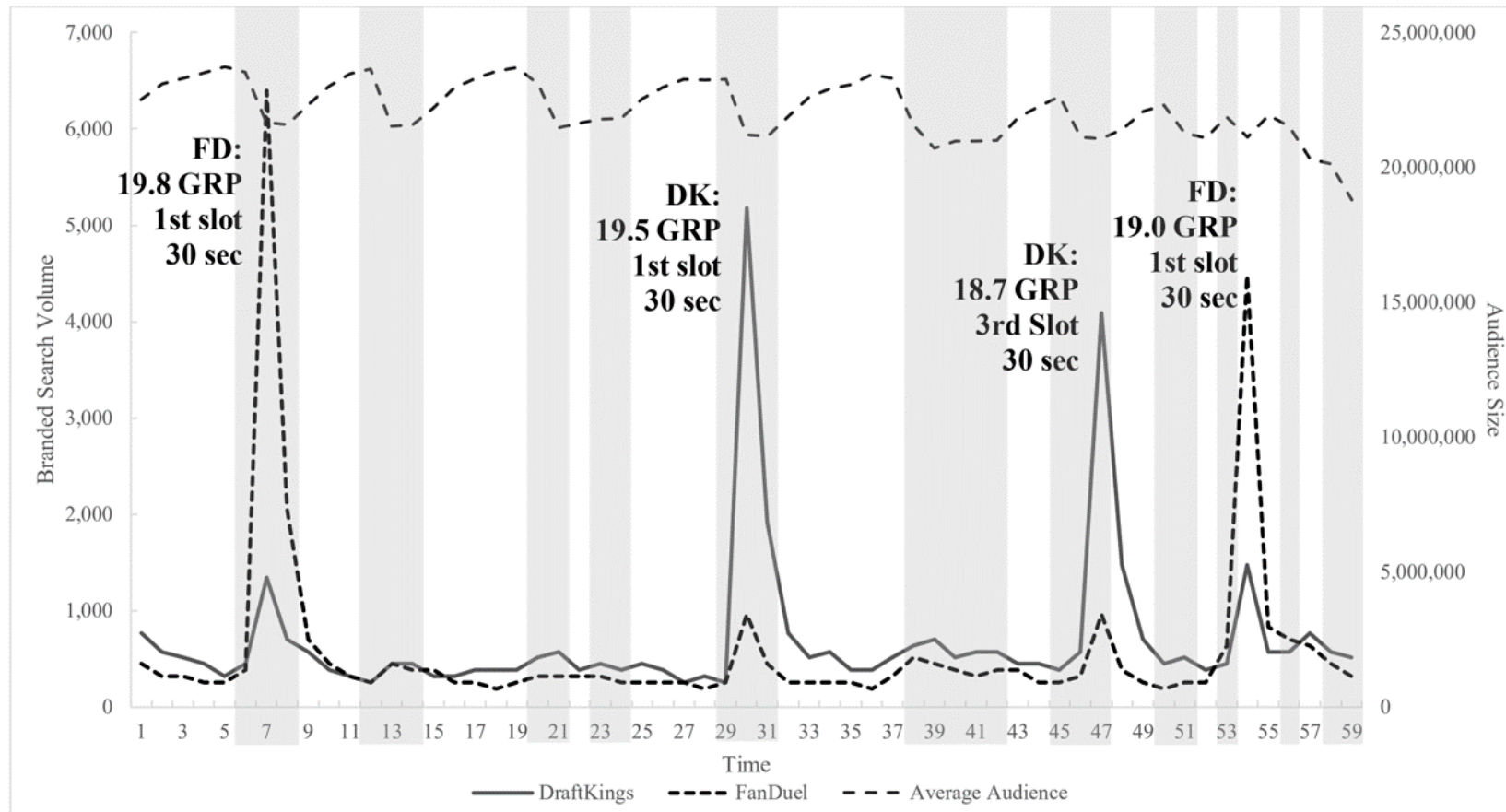
Ad/Sales: Quasi-experiments

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

Possibilities:

- Firm starts, stops or pulses advertising without changing other variables
- Competitor starts, stops or pulses advertising
- Discontinuous changes in ad copy
- Exogenous changes in ad prices, availability or targeting (e.g., elections increase ad prices)
- Exogenous changes in addressable market, web traffic or other factors

TV Ad Effects on Google Search



I made this graph showing DraftKings and FanDuel branded keyword search volume from 9:01-9:59pm ET during the 2015 National Football League (NFL) season opener. TV ads increased search volume by 15-25x, with positive competitive spillovers, and effects that returned to baseline within 5 minutes. Commercial minutes are shaded, showing the spikes came from ads, not just commercial breaks.

Ad/Sales: Quasi-experiments (2)

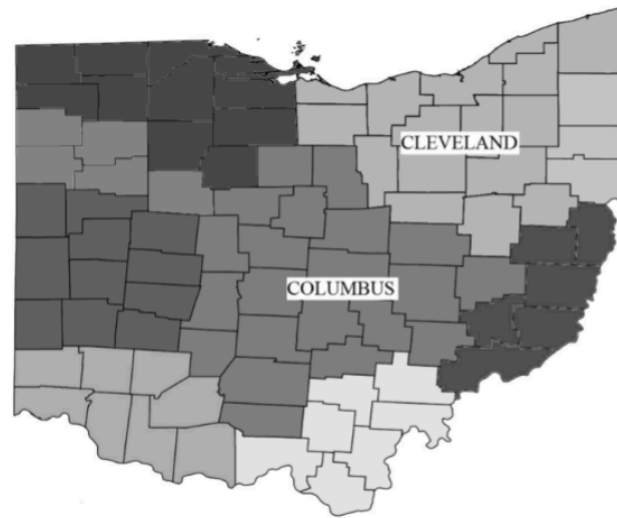


FIG. 5.—Ohio and its DMAs

[Shapiro et al. \(2021\)](#) used a county-border approach to identify how local TV advertising affected packaged goods sales. Media market boundaries reflect broadcast signal footprints, not consumer markets, so consumers on either side of a geographic market boundary are very similar. Therefore, border-county sales predict what in-market sales would have been without advertising.

See also [Shapiro \(2018\)](#)

Experiments vs. Quasi-Experiments

- Experimentalists and quasi-experimentalists differ in beliefs, cultures & training, not unlike Bayesians vs. Frequentists
 - Like B&F, E & Q-E are more similar than different, which is why the debates can get fierce
- Generally speaking, quasi-experiments:
 - Always depend on untestable assumptions (as do experiments)
 - Are bigger, faster & cheaper than experiments when valid
 - Will lead us astray when not valid
 - Are easy to apply even when invalid
 - Range from challenging-to-validate to impossible-to-validate
- E & Q-E should be “yes-and-when-valid,” not “either-or”

Who experiments and why?

- Which companies bother with causality? Does it impact results?

Who Tests the Most?

Google Search

Overview

Our approach

How Search works

Features

Our history

Organizing information

Ranking results

Rigorous testing

Detecting spam

We evaluate Search in multiple ways. In 2023, we ran:

4,781 launches 16,871 live traffic experiments 719,326 search quality tests 124,942 side-by-side experiments

Testing for usefulness

Search has changed over the years to meet the evolving needs and expectations of the people who use Google. From innovations like [the Knowledge Graph](#), to updates to our systems that ensure we're continuing to highlight relevant content, our goal is always to improve the usefulness of your results. That is why, while advertisers can pay and be displayed in clearly marked ad sections, [no one can buy better placement in the Search results](#).

We put all possible changes to Search through a rigorous evaluation process to analyze metrics and decide whether to implement a proposed change. Data from these search evaluations and experiments go through a thorough review by experienced engineers and search analysts, as well as other legal and privacy experts, who then determine if the change is approved to launch. In 2023, we ran over 700,000 experiments that resulted in more than 4,000 improvements to Search.

Google

CEO Quotes on Experimentation

“To invent you have to experiment, and if you know in advance that it’s going to work, it’s not an experiment.”
—Bezos, Amazon

“In a culture that prioritizes curiosity over innate brilliance, ‘the learn-it-all does better than the know-it-all.’”
—Nadella, Microsoft

“We ship imperfect products but we have a very tight feedback loop and we learn and we get better.”
—Altman, OpenAI

“You do a lot of experimentation, an A/B test to figure out what you want to do.”
—Chesky, Airbnb

“The only way to get there is through super, super aggressive experimentation.”
—Khosrowshahi, Uber

“Create an A/B testing infrastructure.”
—Huffman, on his top priority as Reddit CEO

Advertising Experiment Frequency

Marketers Underuse Ad Experiments. That's a Big Mistake.

by Julian Runge

October 28, 2020

Recently, I gave a talk to 30 senior digital growth managers on how to use business experimentation effectively. I started the session with a brief survey: Who had run experiments with their website and app — for example, testing different layouts, colors, designs, or onboarding experiences? Close to 90% of hands rose in response. Then I asked who had run experiments with their digital advertising, such as evaluating different audience targeting, frequency, or optimization regimes for their campaigns? Only about a third of those same hands went up.



To quantify companies' use of experiments, my colleagues and I at Facebook Marketing Science Research conducted an observational survey of leading firms' use of randomized control trials (RCTs) to gauge the impact of a given ad campaign on various business outcomes relative to a control. Though we suspected that only a minority of firms used the practice, we were surprised by just how few: **Only 12.6% of the 6,777 companies we looked at had conducted a recent RCT** (see our academic paper [here](#)).

Given the powerful impact of ad experiments, why are they so underused? Through conversations with internal and external domain experts and scholars, I've identified several common organizational obstacles that may account for this.

Holdout aversion

Organizational inertia:

Requirement of inter-company alignment

Entrenched legacy decision support tools

These four obstacles are all management challenges.

Advertising Experiment Effectiveness

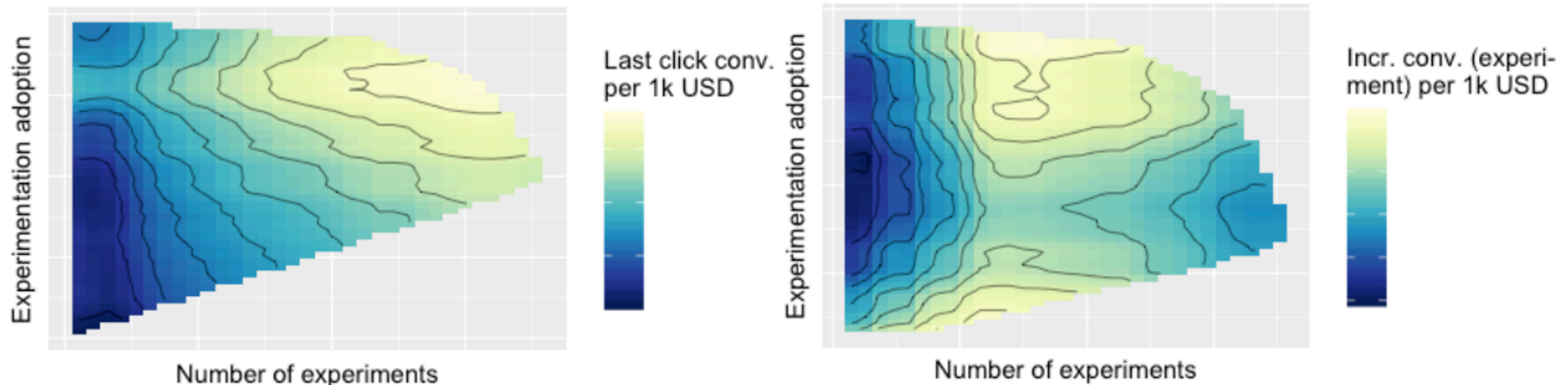


FIGURE 3: Partial dependency plots (via a random forest regressor) to isolate effect of experiment use by firms during the outcome year. Column one considers the number of experiments carried out by the firm. Column two considers both the number of experiments as well as how early the firm adopted experimentation compared to its competitors on the platform (1 = earliest adopters, 0 = non-adopters). Bright yellow indicates highest, dark blue lowest comparative performance. Axes have been removed in accordance with data policies of the online advertising platform.

Companies with deep experimental practices tend to get much better results per ad dollar spent. Ironically, results are correlational; experimentation is not randomly assigned.

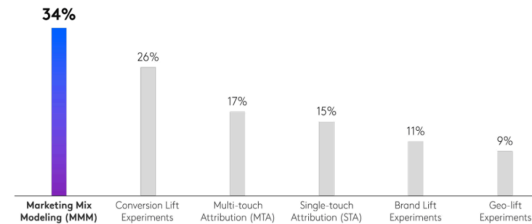
Integrating Methods

- Ways to best combine correlational and causal ad measurement

#1 MMM is often prioritized in advertisers' decision making

MMM is often prioritized in advertisers' decision-making, with 34% of respondents favoring it over all other solutions, followed by conversion lift at 26%.

34% prioritize MMM over all other.



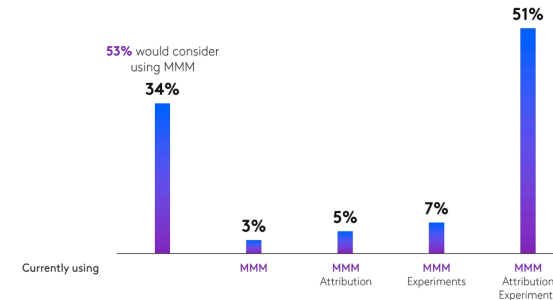
Measurement associations among total	MMM
Holistic understanding of media performance	59%
Data driven approach	55%
Used to make strategic decisions	51%
Worth the investment	47%
Backed by senior stakeholders	46%

#2 Building a suite of truth: Advertisers employ a multimodal measurement stack

Only 3% of advertisers use MMM in isolation, while 51% integrate it with attribution and experiments. Advertisers using a combination of MMM, Multi-touch Attribution/Single-touch Attribution and Experiments are often using these measurement tools together for different purposes. Advertisers typically employ a combination of measurement solutions for decision-making, using an average of 3.8 different solutions.

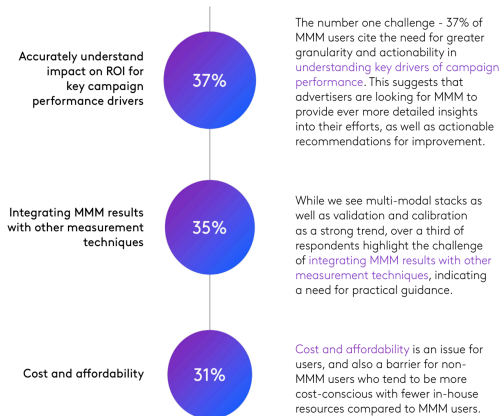
This increases to 4.2 for the largest advertisers.

While 1/3 of businesses surveyed are not using MMM, more than half would consider it.



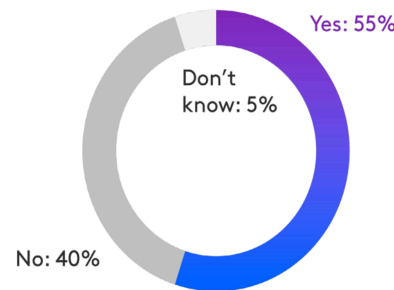
#3 Key opportunities for future MMM improvement

MMM pain points among those who have ever used MMM



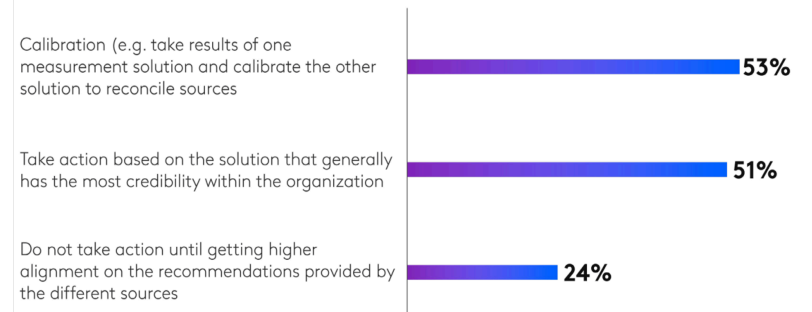
#4 Maximizing accuracy through validation and calibration

Contradicting results
Among those who have used more than one approach



Reconciling contradicting results

Among those who have experienced contradicting results



Kantar (2025) surveyed 1,935 decision makers across regions, industries and company sizes investing over \$1MM in digital advertising

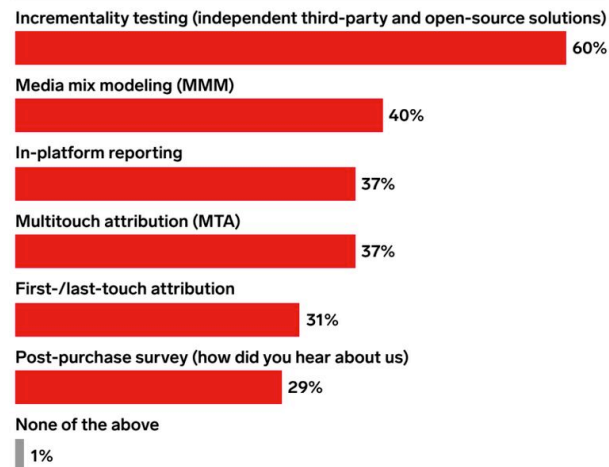
Integrating MMM with Causal Measurement

- Correlational methods are cheap & always-on; Experiments are credible but scarce & costly; Both are noisy; Measures will disagree

No Single Marketing Measurement Approach Commands Full Trust

% of US senior decision-makers, Jan 2026

Q: Which of the following, if any, are your most trusted marketing measurement solutions? Select up to 3.



Source: Haus, "2026 Marketing Decision Confidence Index," March 18, 2026

365580

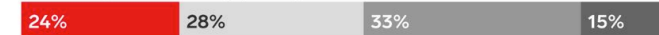
EMARKETER

MMM Experience Lags Behind Other Measurement Approaches

% of marketers worldwide, Sep 2025

Q: Does your marketing team use any of these advanced analytics methods?

Marketing mix modeling (MMM)



Incrementality testing



Attribution modeling



A/B and multivariate testing



4%

■ No experience/use ■ Beginner ■ Intermediate ■ Advanced

Note: n=238; numbers may not add up to 100% due to rounding

Source: Funnel, "The 2026 Marketing Intelligence Report" conducted by Ravn Research, Dec 2, 2025

364954

EMARKETER

Management challenge: What do we do when Attribution says ROAS was 5.9, MMM says ROAS was 2.3 +/- 1.2 and Experiment says iROAS was 0.6 +/- 2.1?

Using Experiments to Validate MMM Reallocations

- A natural first step is to validate correlational ad measurement with experiments
- Dropbox did this with small experiments in Canada, then a month-long full-US blackout on mobile and search, which showed that search attribution was 117% too high
- Mobile-trial starts fell 5%, but earnings guidance increased, citing performance marketing efficiency as a key driver
- More generally, when MMM recommends a change, run a test alongside implementing the change, and compare the resulting test result to the MMM prediction c.i.
 - This tests MMM output's predictive validity



Using Experiments to Calibrate MMM

- Use causal ad measurements to constrain & improve MMM
 - Priors: experiment results set Bayesian channel-ROI priors ([Meridian docs](#); [Google Research 2024](#)). [Kaminsky \(2025\)](#) advises to apply experiment's full confidence interval, but only to MMM observations in which the experiment ran
 - Model Selection: [Robyn](#) scores candidate models on ability to reproduce incrementality estimates, alongside business and statistical criteria ([Runge et al. 2024](#))
 - Experiments can enter the likelihood as observations, grounding each channel's response curve using the experimental result ([PyMC-Marketing](#))

Prior-based and likelihood-based calibration deliver similar accuracy improvements ([Orduz 2024](#))

[See also Recast](#)

Unified Marketing Measurement

- Management considers MMM, Experiments, Attributions, and Brand Lift Tests holistically, seeking to understand the relationships between them, and pros and cons of each, rather than giving primacy to any single measure in all situations ([Andrew 2024](#), [Nair 2026](#))
- When measures overlap, we can compare them directly, to estimate the bias of the less-credible measure. E.g., if an experiment result is half as large as a comparable attribution, then similar attributions might be deflated by 50%. The stability of intermeasure comparisons then becomes important
- [Gordon, Moakler & Zettelmeyer \(2026\)](#) propose Predicted Incrementality by Experimentation (PIE) in which (1) past campaigns with experiments are used to estimate how campaign features predict incremental ad effects, then (2) estimates are used to predict untested campaigns' incrementality. The model generalized well to nontraining campaigns. [PyMC-Marketing](#) includes an open-source implementation

PIE stability should be tested in each instance, not assumed

Iteration

- Use MMM estimates to identify high-value experiments to invest in; which experiments would maximally inform our next MMM?
- Induce some randomness in ad spend by Channel/Time/Geo so that MMM can estimate causal parameters
 - Switchback design could be used to iterate between heavy and light spending levels
 - A fixed proportion of MMM observations could be reserved for random deviations from optimal spending
 - Approaches like these may require platform APIs to set campaign budgets at granular levels. Sophisticated buying requires expertise

All integration approaches can be used simultaneously

Frontier Topics

- Optimal Brand Advertising Measurement: How to maximize loyalty, minimize price response
- Incrementality in Time: Can we distinguish purely incremental purchases from time-accelerated purchases that would have occurred later?
 - Requires counterfactual predictions of *future* events ([Lambrecht, Tucker & Zhang 2024](#))
- Incrementality $\neq$ Optimality: Incrementality asks whether the ads truly worked; optimality asks whether the ads maximized profit
 - How do we learn profit-maximizing combinations of Message, Target, Media, Context, Ad Price, Ad Frequency?
- How best to integrate quasi-experimental analyses
- How to choose experiments to maximize learning value ([Epanomeritakis & Viviano 2026](#))

Optimality should be the next frontier after incrementality is better established and integrated

Frontier: Causal Brand Ad Measurement

- Brand ads are designed to work over a long time horizon, but maintaining a long-term control condition is very expensive and often impossible
 - May violate experimental necessary conditions, e.g., SUTVA
 - Most marketers test brand ads before they run, rely on correlational brand ad measurement, and/or don't causally measure brand ads
- Most promising imho: Test how brand ad spend changes causal performance ad effects
 - Performance ads can be hard to scale due to increasing CAC, and 'fishing out the pond'
 - Using upper-funnel ads to 'warm up' the target predicts a brand × performance interaction
 - Brand ads tend to be beautiful and clear; performance ads often look cheap or tacky; performance ads alone might be a local maximum
 - Some sophisticated companies started as performance-only & later became huge brand advertisers; how do you think they justify it internally?

Ad measurement debates often confuse absence of evidence with evidence of absence. As a performance marketer, it can be uncomfortable to acknowledge that we cannot reasonably estimate causal brand ad effects. But does that mean the effects are zero? Can we know that?



What would you do?

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

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
COO | Co-Founder | Partner | Advisor
4d ...

100 🏆

Like · Reply

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

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

Like · Reply

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

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

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

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

Definitely bring in academics as outside consultants :-)

Like 6 · Reply

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

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

Like 1 · Reply

Ayman Farahat
Principal Scientist at Amazon
10h ...



Like · Reply

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

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
Customer Centric Privacy Protecting Marketing AI
2d ...

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

Like 4 · Reply

Leading a traditional team to adopt incrementality can be a resume headline and interesting challenge, especially if you apply it to solve your hardest challenge. However, it requires leadership support, you usually cannot do it alone. If structural incentives misalign, consider a new role.

Wrapping Up

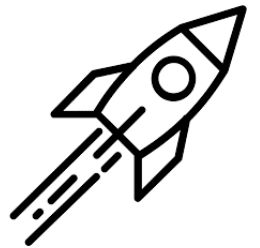
- Takeaways, Further Study, AdMeas the Game, Thanks to

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.
 - Common remedies: Experiments, Quasi-experiments, Correlations, Triangulate; Ignore
- Incrementality-based advertising measurement is a generational shift improving marketing profits, but we still have a long way to go
- Experiments are the gold standard, but are costly and challenging to design, implement and act on
- Ad effects are subtle but that does not imply unprofitable. Measurement is challenging but required to optimize profits

Resources for Further Study

- [Paparo \(2025\)](#): Insider's account of programmatic advertising development from 2000-2025
- Content providers to follow: [Adexchanger](#), [Adweek](#), [Digiday](#), [Marketecture](#)
- [Project Eidos](#): IAB's effort to define admeas principles, standards, and frameworks
- [Gordon et al. \(2020\)](#): Discusses iROAS estimation challenges and remedies
- [Dew et al. \(2024\)](#): Smart discussion of key MMM assumptions
- [Luca & Bazerman \(2020\)](#): Goes deep on digital test-and-learn considerations
- [Barajas et al. \(2021\)](#): Online Advertising Incrementality Testing And Experimentation: Industry Practical Lessons



AdMeas, the Game

AdMeas, the Game
Zippity — playtest build (calibrated placeholder parameters)

Welcome
[Instructions](#)

Enter your gameplay code
Your instructor issued you a one-game code. Enter it to start your year as Zippity's CMO — or to resume a game in progress from any device.

Gameplay code

ZIP-XXXX-XXXX

Continue

This code is fresh — tell us who is playing, then your year begins.

Full name

Email address (UCSD email preferred)

PID (optional)

Start my year

AdMeas, the Game

Apply your understanding in AdMeas the Game. You're Zippity's first CMO, tasked with advertising budgets and measurement, to maximize company profit. Allocate money to two ad types in six channels, informed by attribution, experiments and MMM. Gameplay requires a code—students get it on [Canvas](#), anyone else can get one by messaging [Ken on LinkedIn](#).

<https://admeas-game.kennethcwilbur.workers.dev>

Thanks to

- Joel Barajas (Walmart), Rick Bruner (Central Control), Peter Daboll (DAIVID), Tom Flanagan (Flywire), Mike Follett (Lumen), Carl Mela (Duke), Prabhath Nanisetty (Snowflake), and Koen Pauwels (Northeastern) for helpful comments
- Colleagues and students who helped improve earlier versions
- Benedict Evans for inspiring the assertion-evidence slide format; McDermott & Butts for the quarto theme