Advertising Measurement

[Kenneth C. Wilbur]

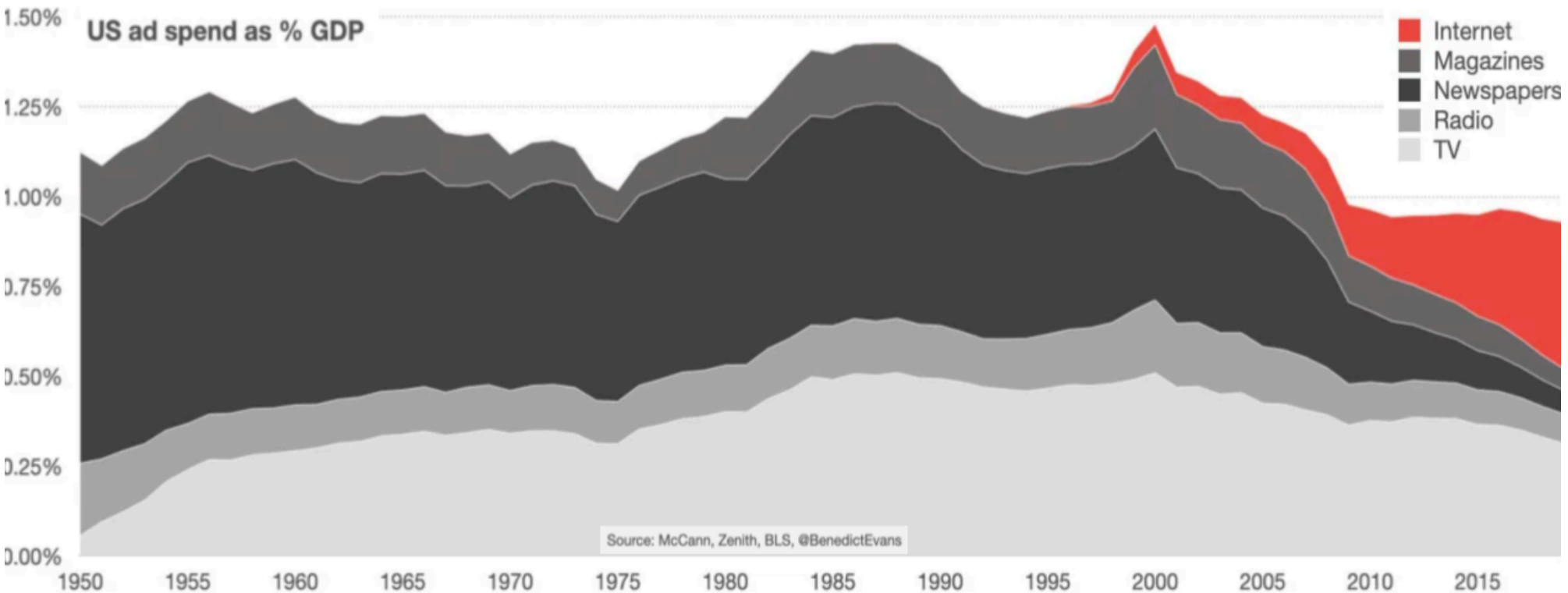
Professor of Marketing and Analytics

University of California, San Diego

Domain Knowledge

- Broad context to motivate and interpret advertising measurement

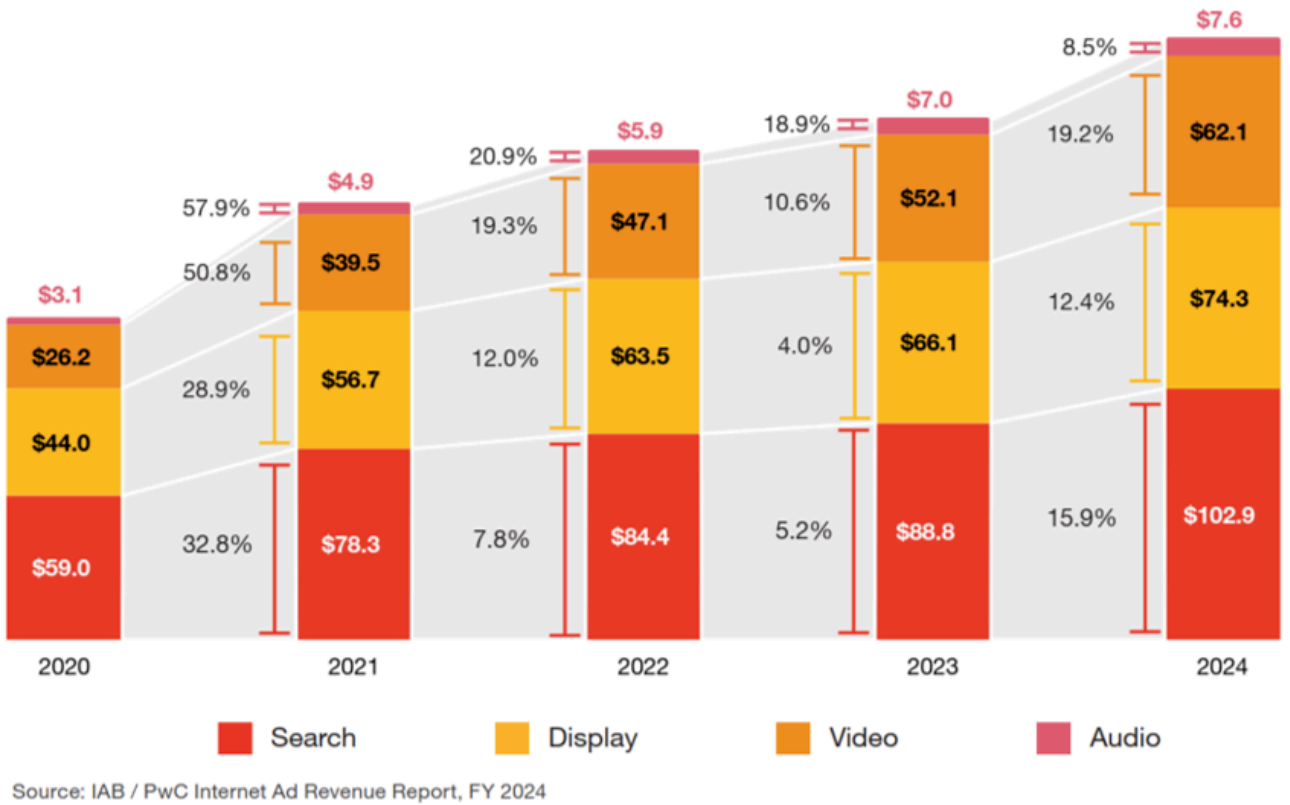
Advertising sales revenue, 1950-2019



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)

Online Ad Revenues, 2020-2024

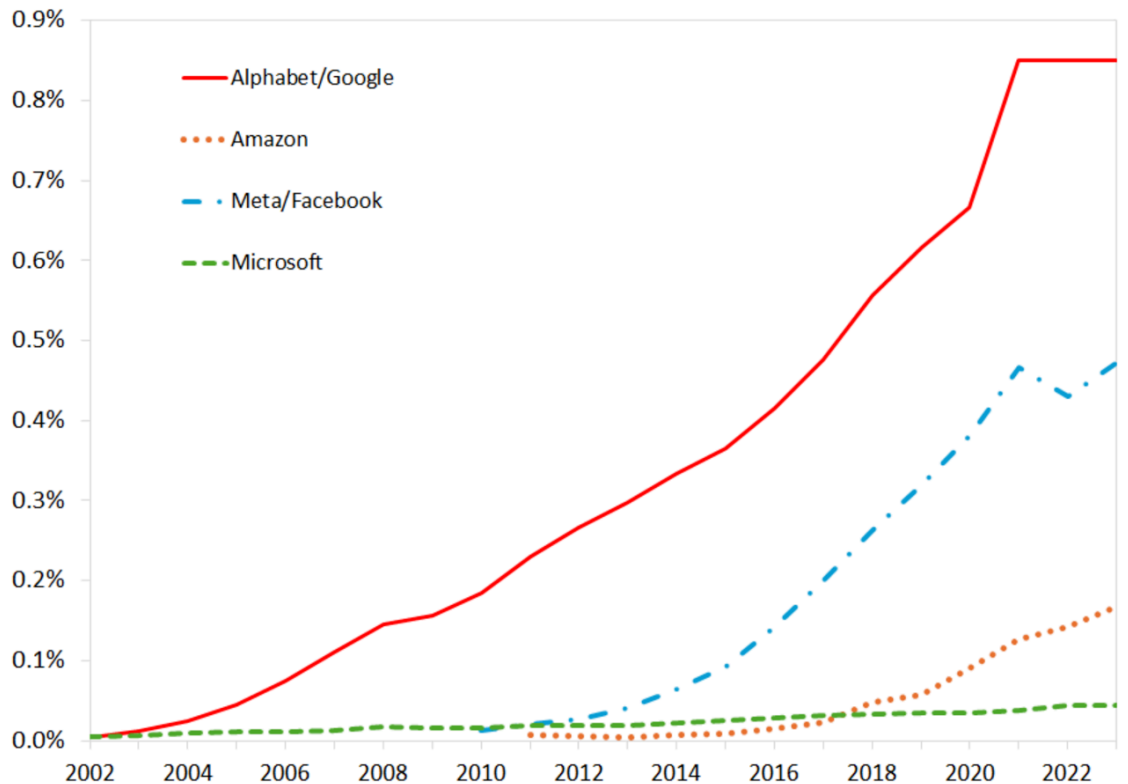


Source: IAB / PwC Internet Ad Revenue Report, FY 2024

These data expand the Internet bar. The Interactive Advertising Bureau collects ad sales revenue by format from online ad sellers and supply chain firms. Total spending grew from $132.3B in 2020 to $246.9B in 2024. What else do you see?

IAB (2025)

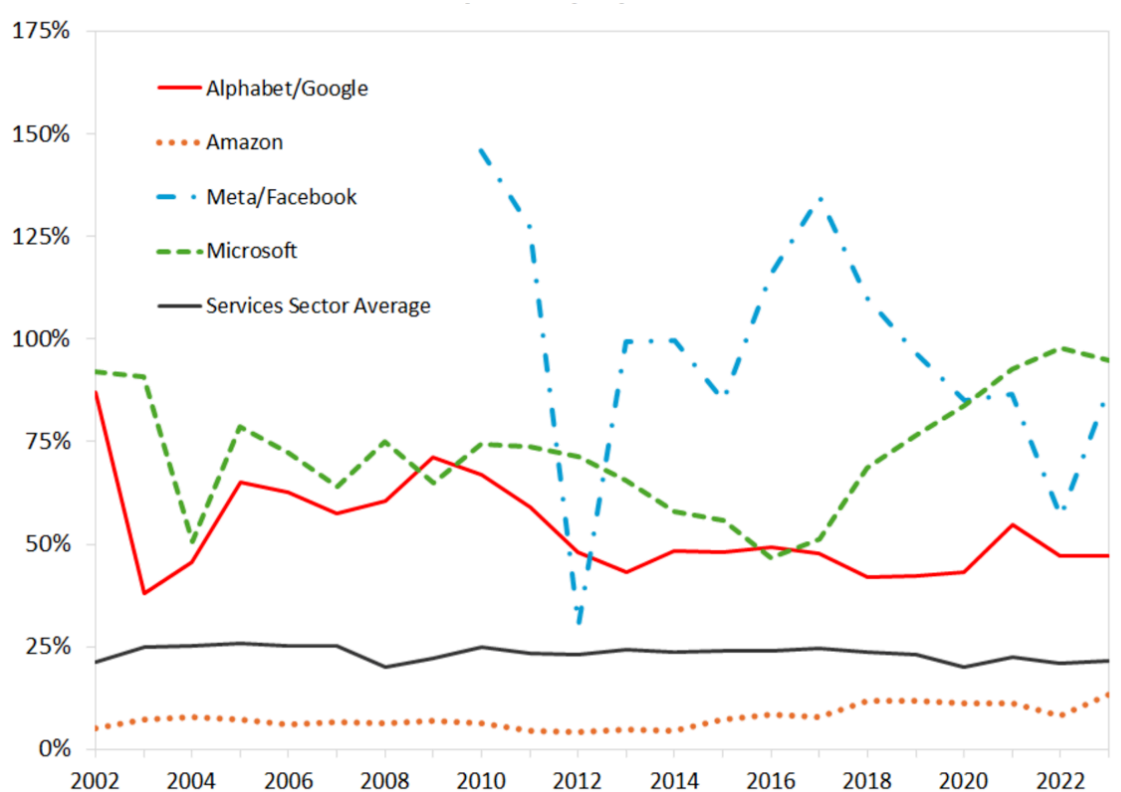
Large Sellers' Ad Revenues as % of US GDP, 2002-23



Online advertising sales are increasingly dominated by a few large firms. We used to call them "the duopoly," but now we call them "the triopoly."

Marto and Le (2024)

Large Ad Sellers' Profit Percentage, 2002-2023



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?

Marto and Le (2024)

Toy economics of advertising

- Suppose we pay $10 to buy 1,000 digital ad 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 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

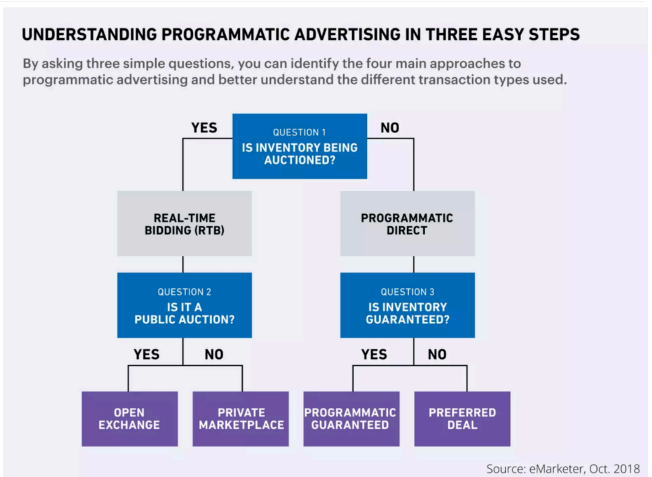
> Even a perfectly efficient and omniscient advertising industry might struggle to learn how to optimize advertising delivery. What behavioral or contextual signals might indicate mortgage loan receptivity? How much more cost-effective would these targeting signals make the ads?

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



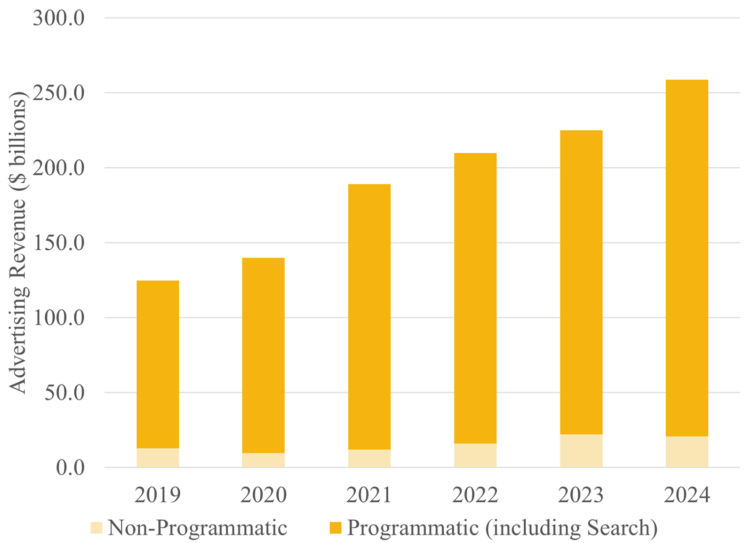


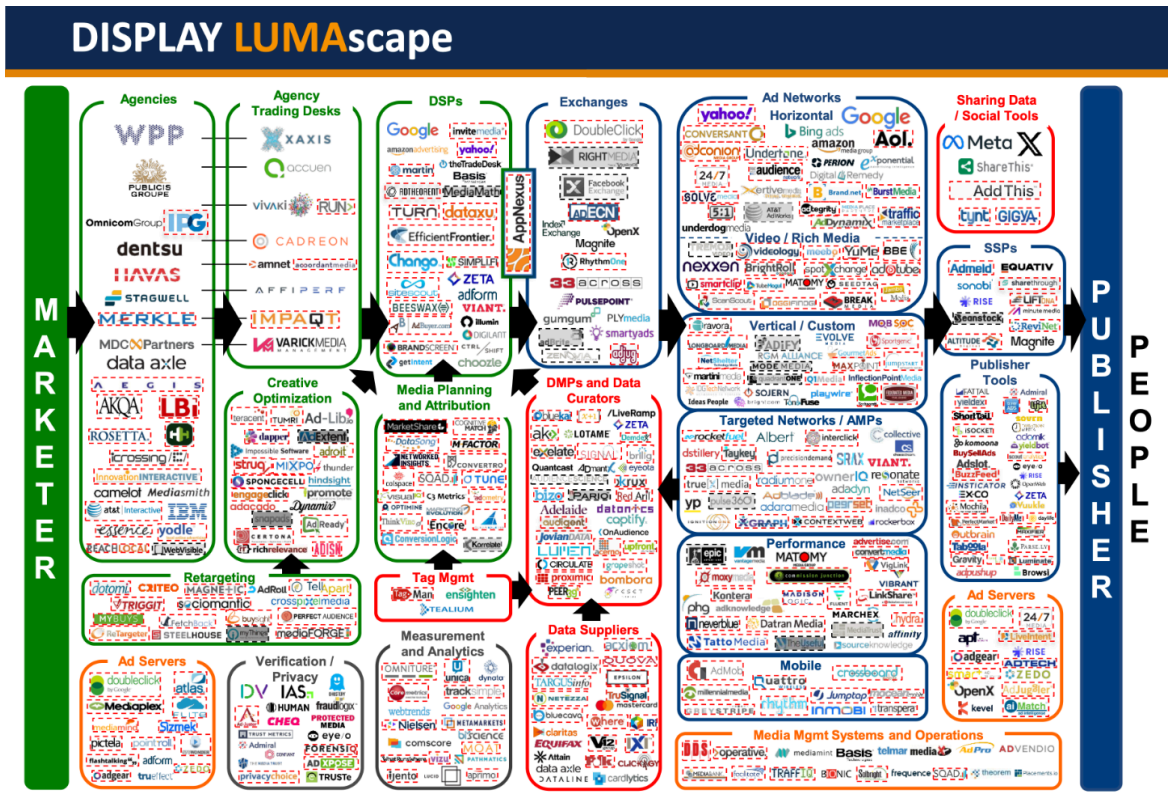
Figure 2. Internet Programmatic Advertising Revenues, 2019-24[48]

[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?

IAB Glossary of Terms (requires free registration)

The Ad Tech Ecosystem



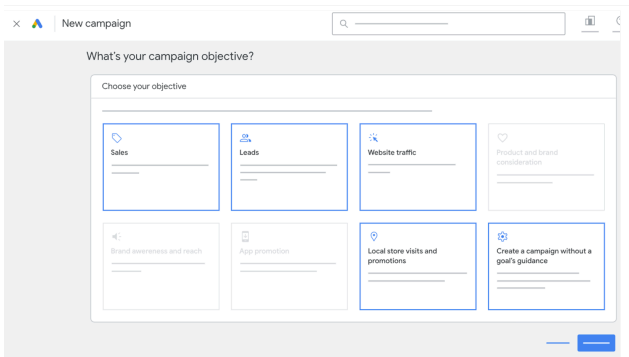
Luma Partners maps ad tech ecosystems. Each logo is a company that intermediates between advertisers and publishers: data/algorithm specialists, representatives, and marketplaces. This map is one among many.

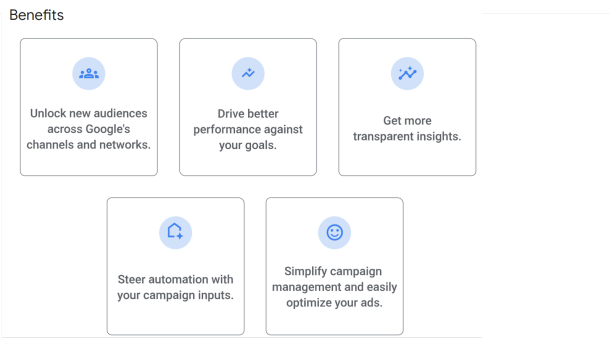
Luma Partners

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.



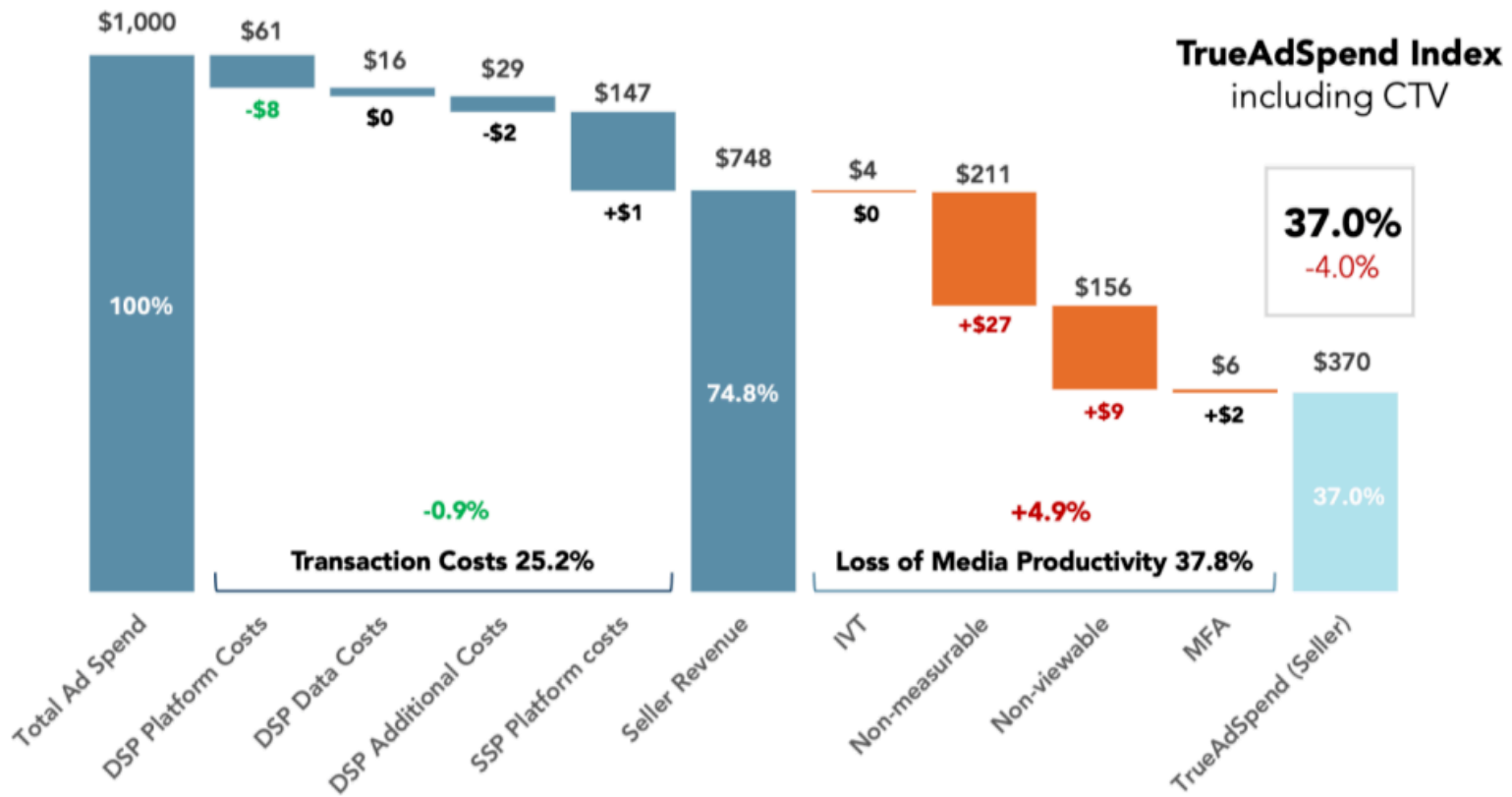


Google Ads & Commerce Blog

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 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.

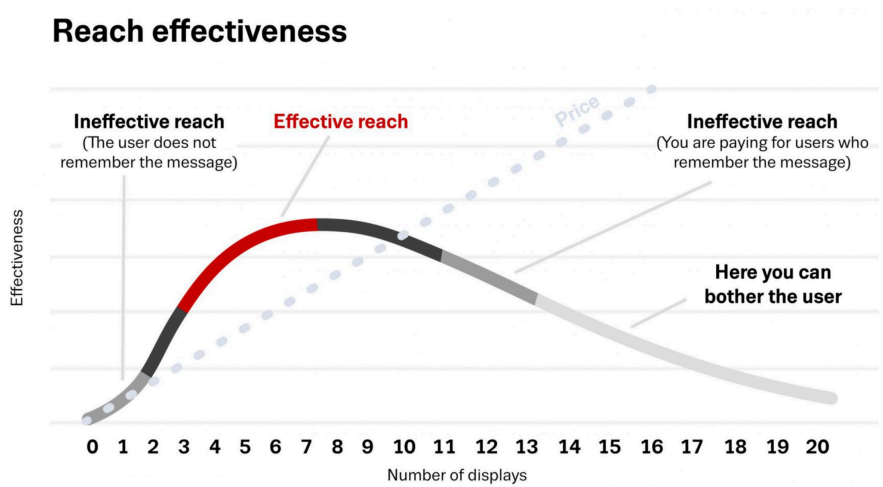
The “Ad Tech Tax”

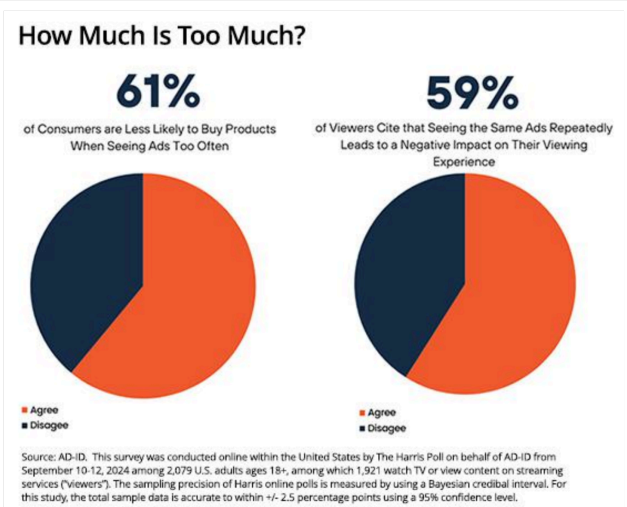


2025Q2 data show that DSP takes 11%, SSP takes 15%, publisher receives 75%, and about half of that is verifiably viewable by human recipients. What are DSP, SSP, IVT, Measurable, Viewable, MFA?

ANA/TAG (2025)

Effective Frequency





Have you ever had an ad "follow you around"?

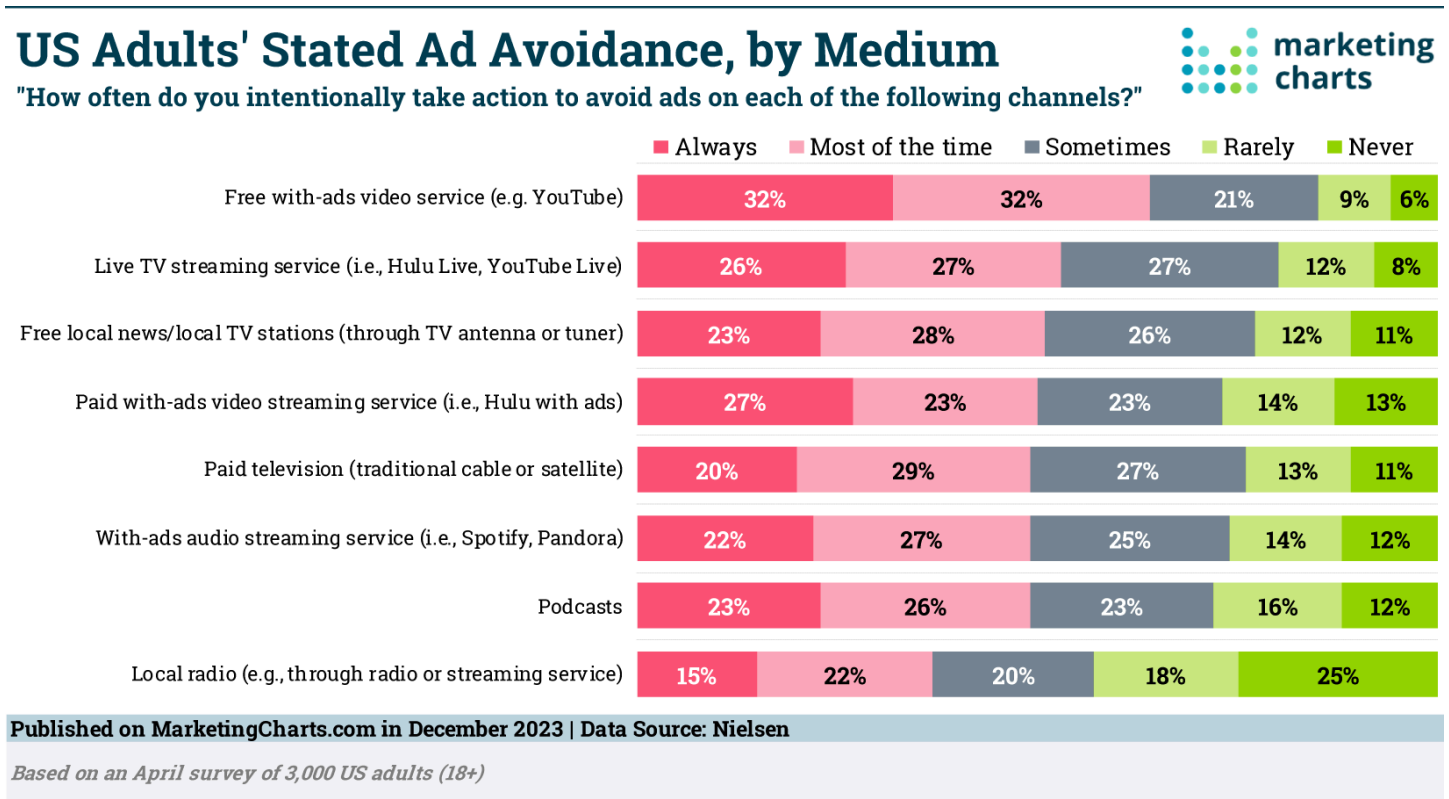
Age-old advertising theory posits a nonlinear effective frequency curve, which is to say, the marginal effect of an ad on conversion probability depends on how many times the consumer sees the ad.

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)

Seznam (2023)

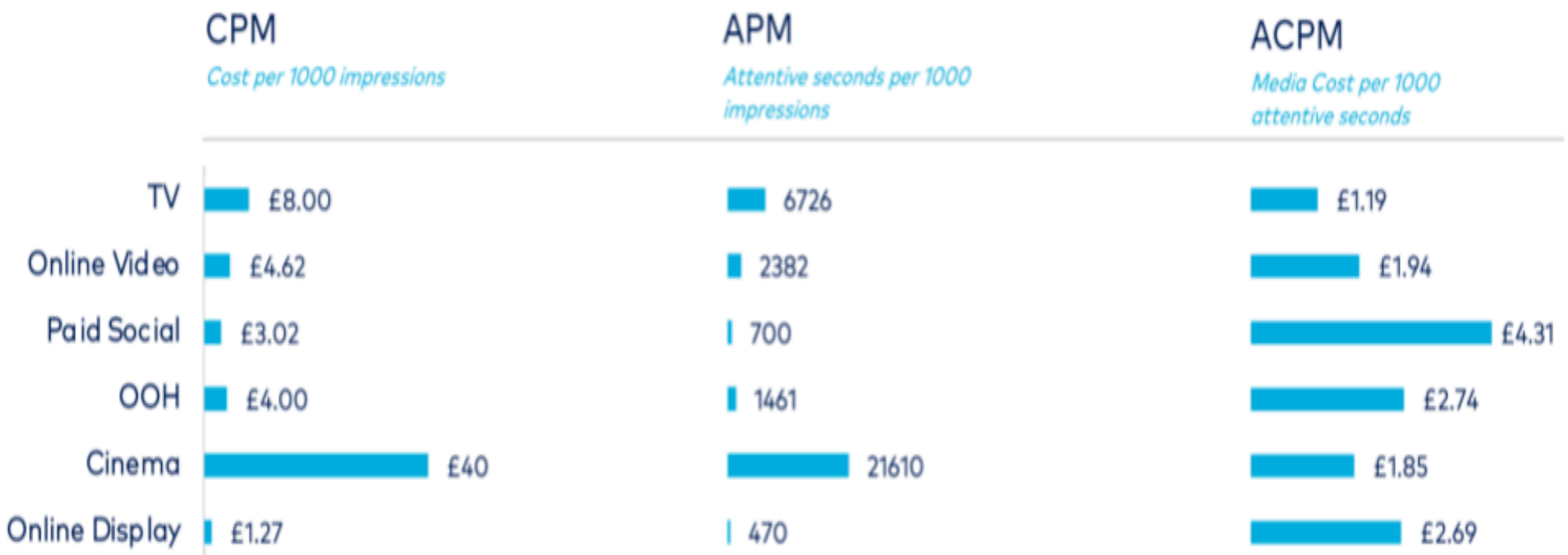
Advertising avoidance



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? Ad load and ad nuisance are the “attentional prices” that subsidize our media: without ads, we would pay more for content.

Nielsen (2023)

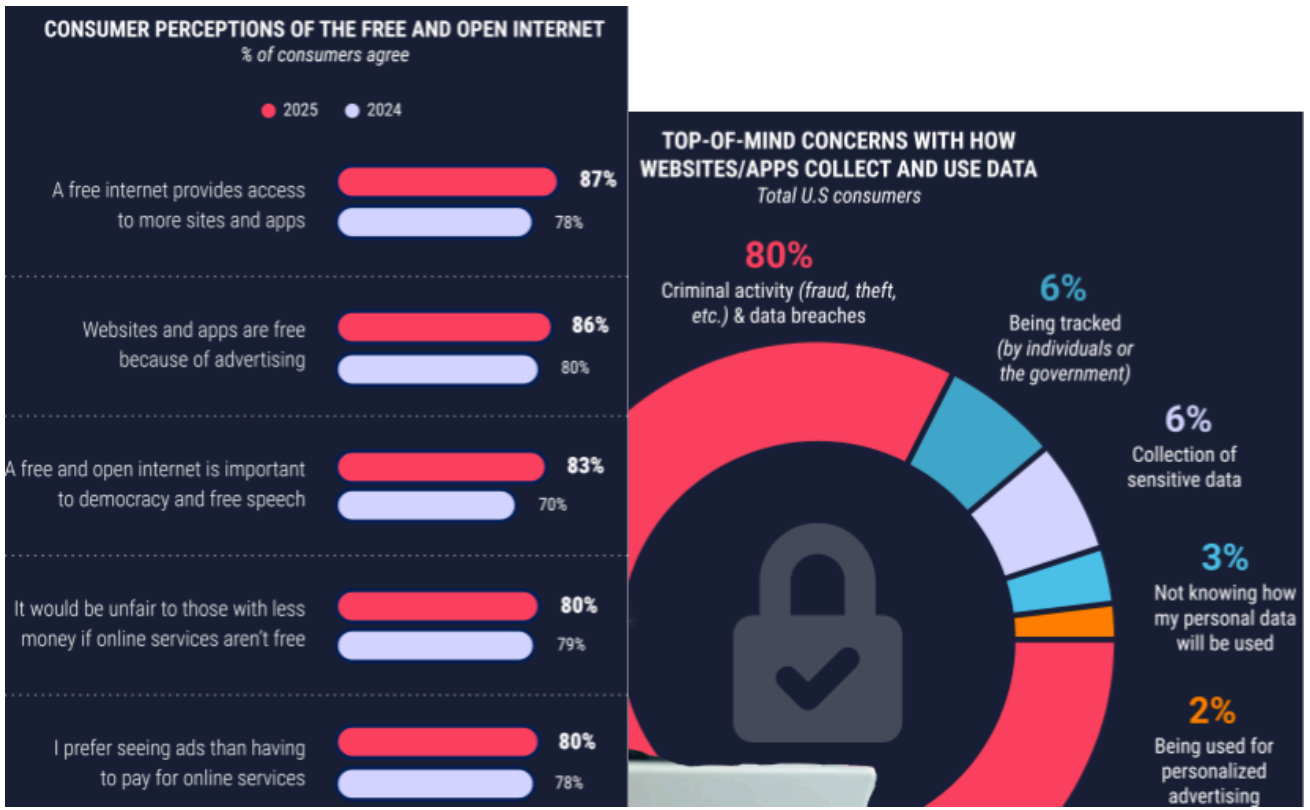
Consumer Attention by Medium



Advertising media vary in average attention attracted (i.e. eyes-on-screen) and advertising price. These data reflect attention measurements and prices by medium. What do you see?

Ebiquity & Lumen (2024)

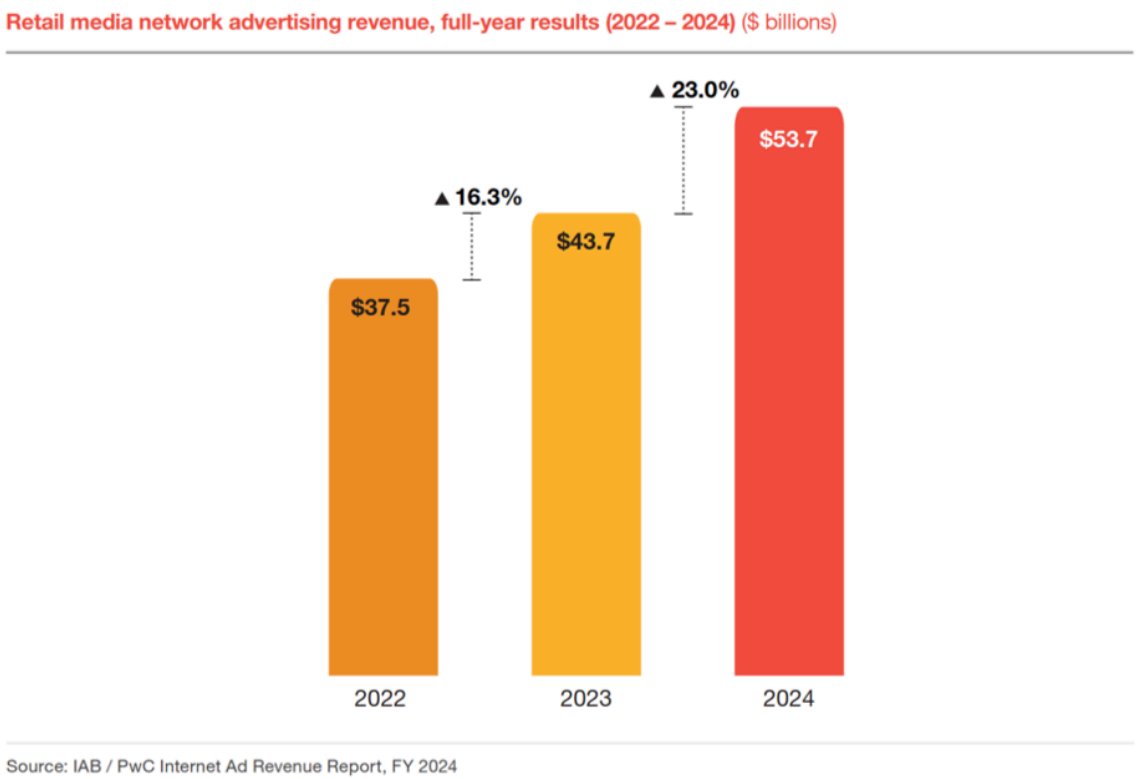
Privacy Perceptions



Consumers don't especially love advertising but they mostly understand that advertising subsidizes media access, and most 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.

IAB (2025) | Browser Fingerprinting (2025) | Yeo et al. (2025)

Retail Media Networks



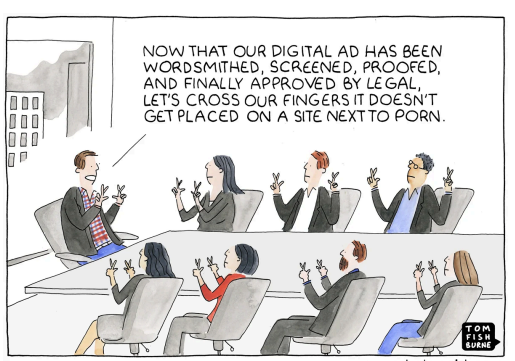
Retail media networks (e.g., Amazon, Walmart) provide data for ad targeting, sell sponsored product search listings, sell ads on behalf of publishers, and measure advertising conversions. RMNs are growing quickly as they cannibalize older trade promotions budgets.

Gabel et al. (2024)

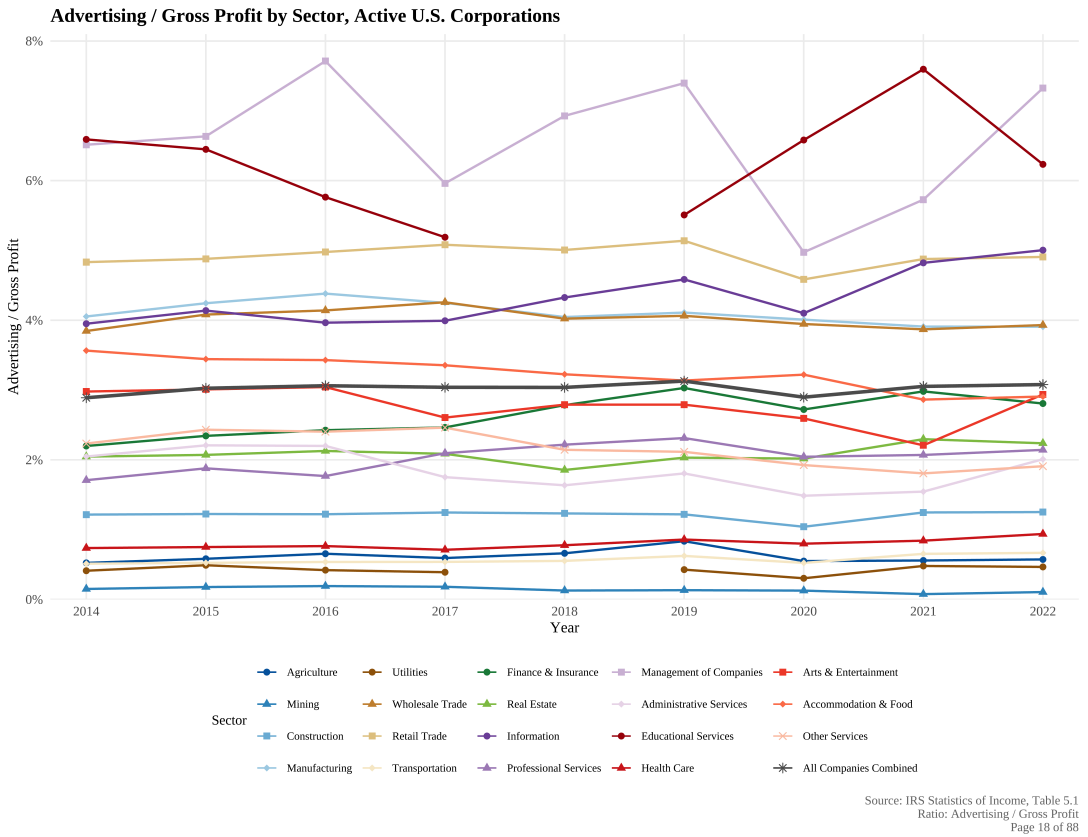
Brand Safety and Suitability

- **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, yet fraudulent ad sales may place brand ads in non-safe contexts
  - → 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)



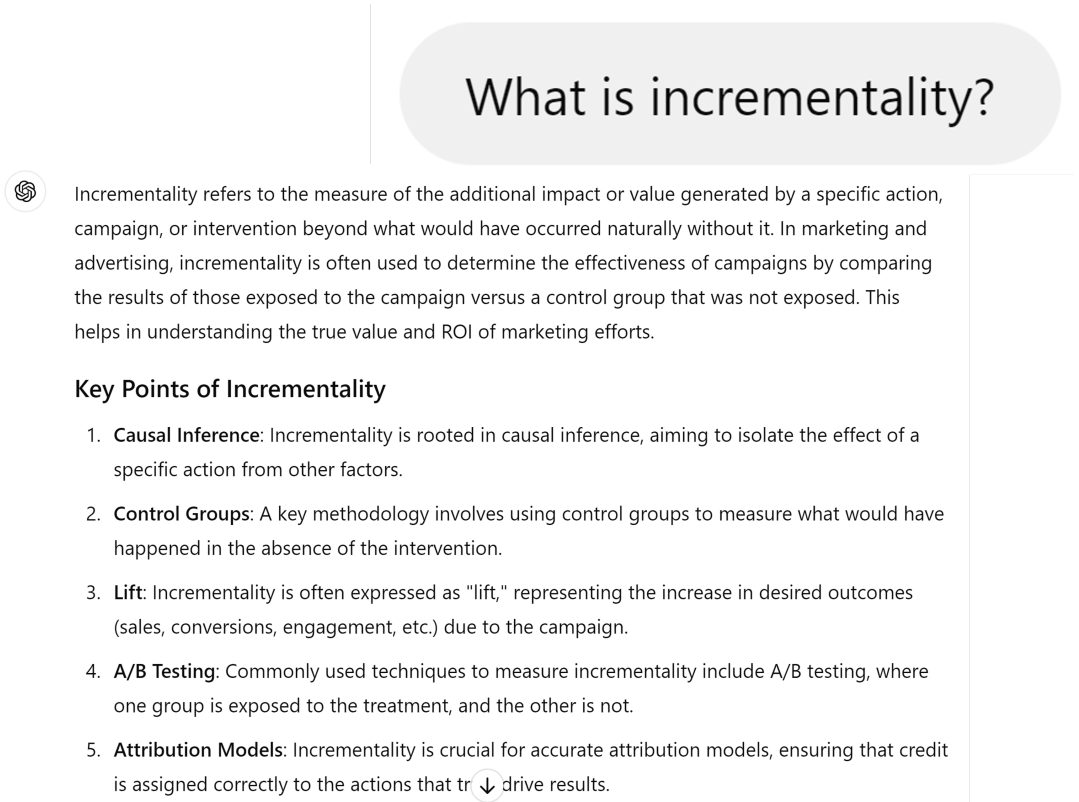
How much do companies spend on Advertising?



The average US corporation spends about 3.1% of gross margin on advertising deductions. Gross margin ranges from 3-5x net margin, so the modal firm could increase net income 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]

Wilbur (2026, p. 18) | Saibooks

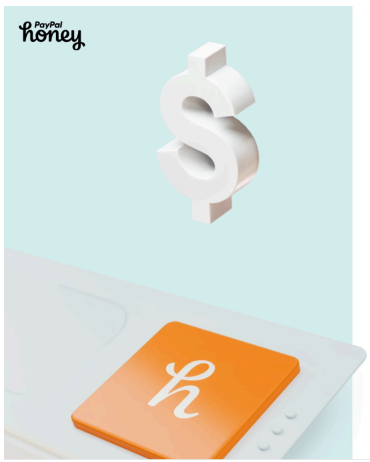
What is Incrementality?



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

The word incrementality is only used in marketing.

Honey: A Case Study in Attribution Fraud



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

Add to Chrome – It's Free

★★★★★ 141,786 Chrome Store reviews
17 million members and counting

THE CREATOR ECONOMY

The Honey scandal is a 'wake-up call' for the creator industry's affiliate partnerships

By Antoinette Siu • February 28, 2025 •

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 ir 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

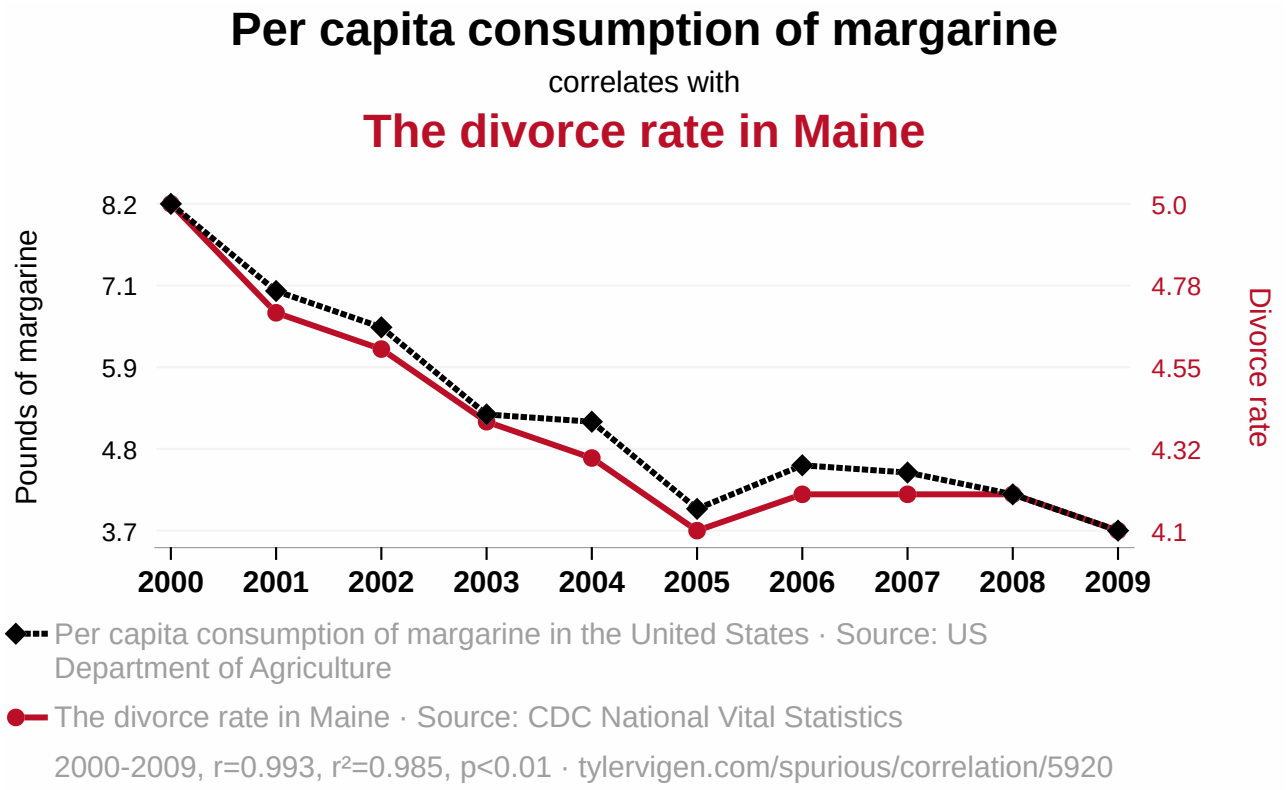
Honey, a PayPal-owned browser extension with 17 million members, was found to replace influencers' affiliate codes with its own, stealing affiliate marketing fees from partners. This is one example of "cookie stuffing" attribution fraud, and helps to illustrate why ad buyers don't always trust ad sellers. Incrementality estimates verify marketing value delivered without relying on seller statements.

MegaLag (2024)

Causality

Examples, fallacies and motivations

Correlation ≠ 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 {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.

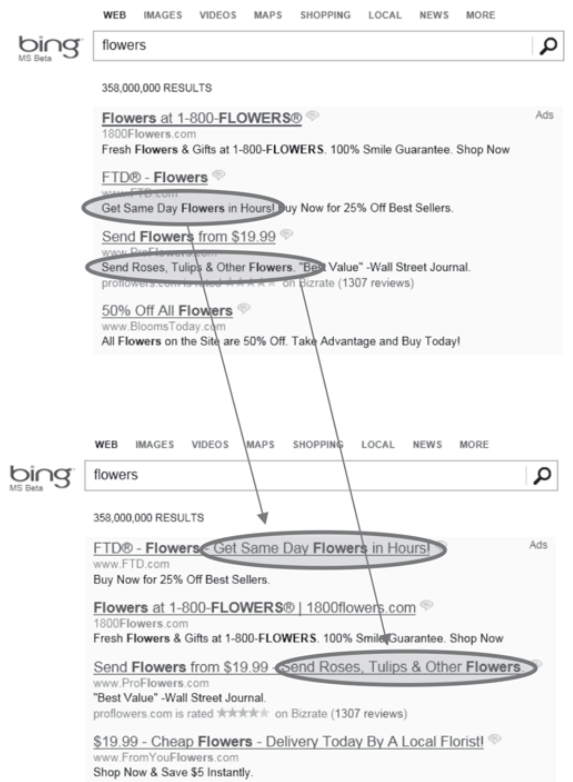
R Script simulating this scenario

Classic misleading correlations

- “Lucky socks” and sports wins
  - → Post hoc fallacy (precedence indicates causality AKA superstition)
- Commuters carrying umbrellas and rain
  - → Forward-looking behavior
- Kids receiving tutoring and grades
  - → Reverse causality / selection bias
- Ice cream sales and drowning deaths
  - → Unobserved confounds
- Correlations are measurable & usually predictive, but hard to interpret causally
  - → Correlation-based beliefs are hard to disprove and therefore sticky
  - → Correlations that reinforce logical theories are especially sticky
  - → Correlation-based beliefs may or may not reflect causal relationships

Wikipedia

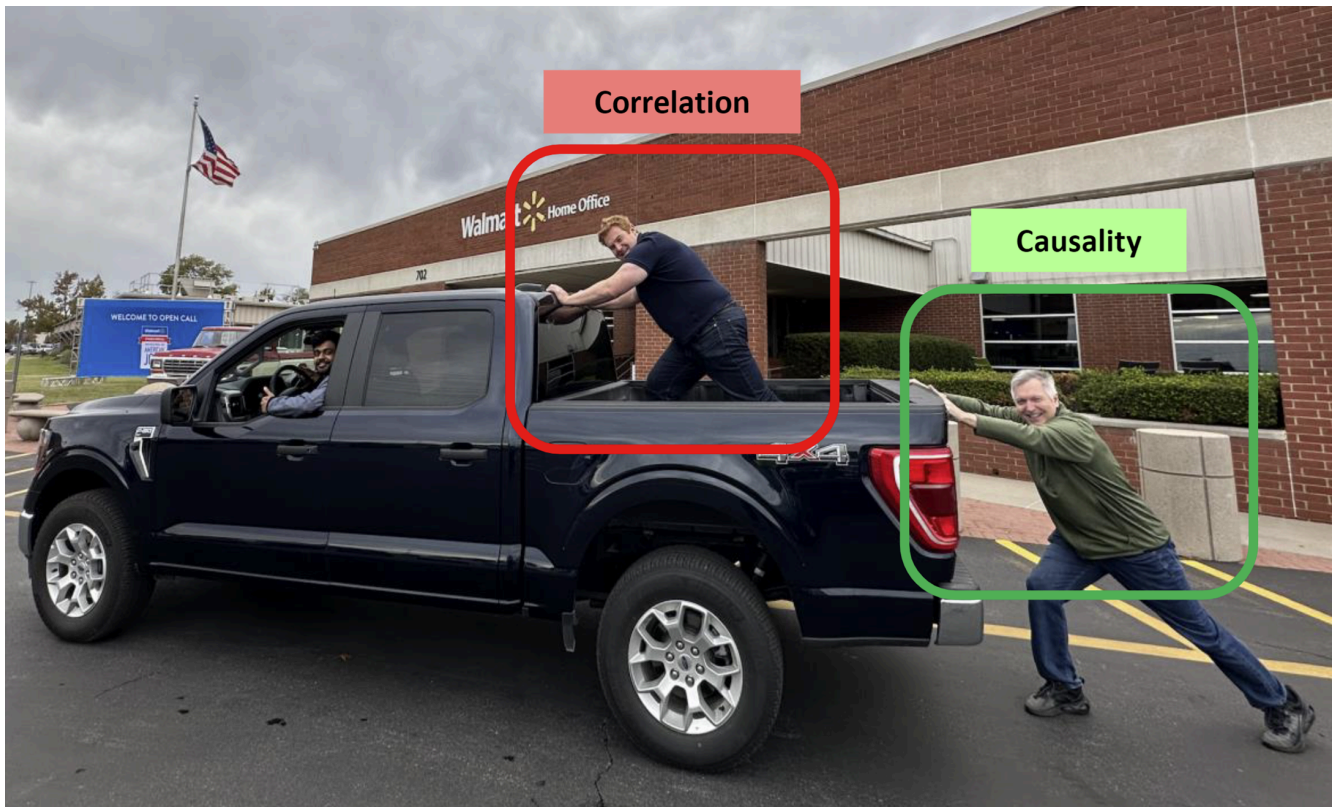
“Revenue too high alert”



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

Kohavi & Thomke (2017)

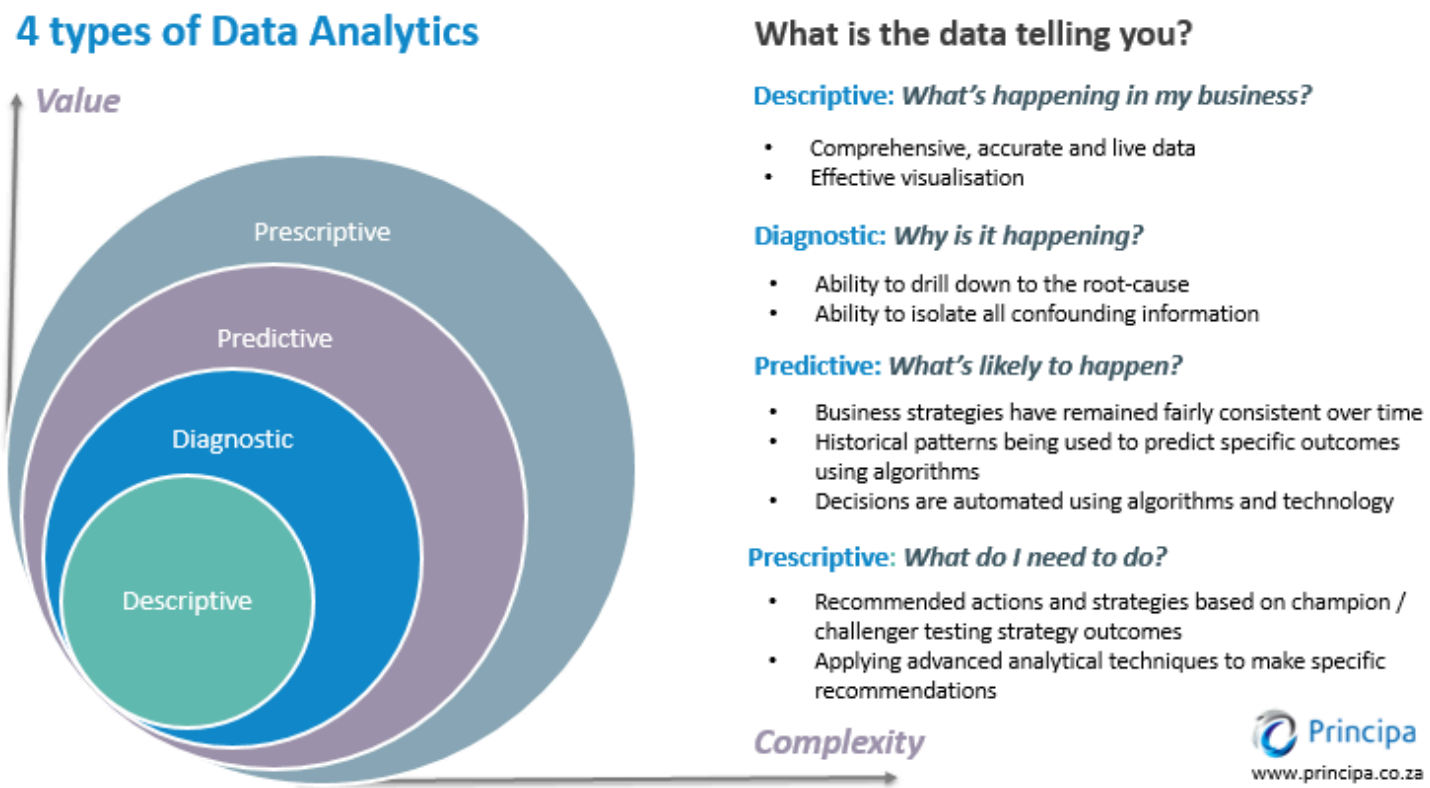
Correlation vs Causation



> 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)

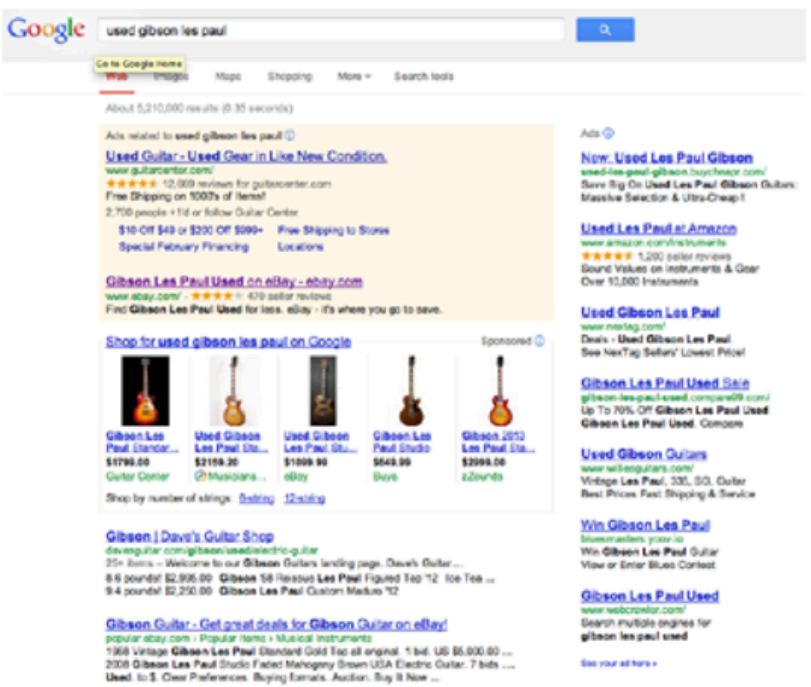
Four Types of Analytics



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 may suffice.

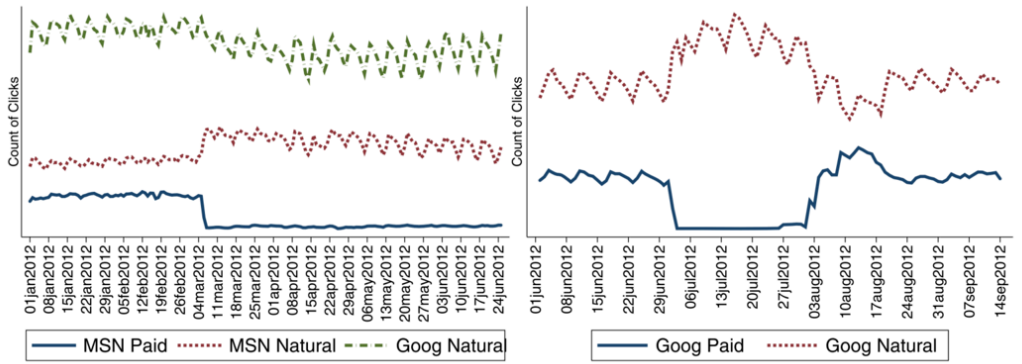
eBay Search Ad Experiments

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.



Blake, Nosko & Tadelis (2015)

eBay Results: Click Substitution



(a) MSN Test (b) Google Test

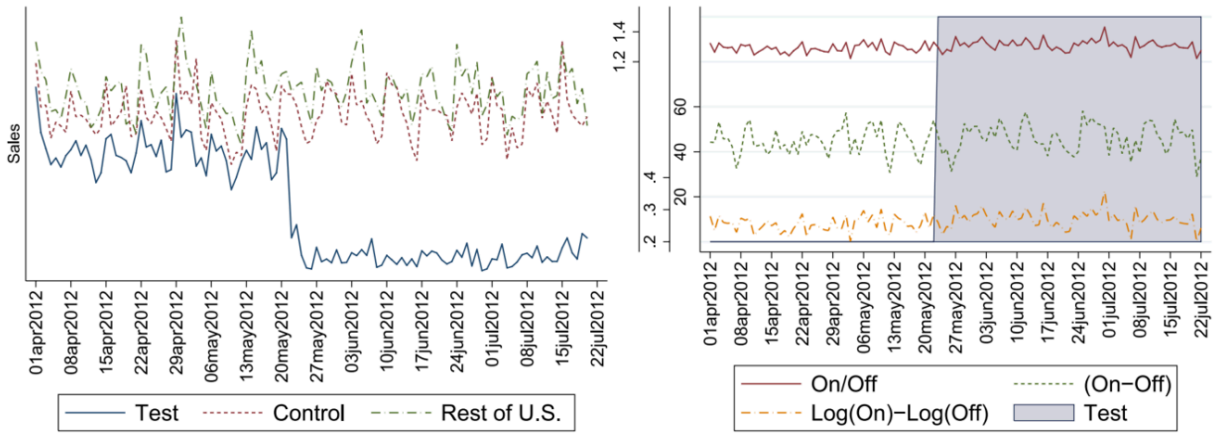
FIGURE 2.—Brand keyword click substitution. MSN and Google click-traffic counts to eBay on searches for 'ebay' terms are shown for two experiments where paid search was suspended (panel (a)) and suspended and resumed (panel (b)).

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.

Blake, Nosko & Tadelis (2015)

eBay Results: Attribution vs Reality



(a) Attributed Sales by Region (b) Differences in Total Sales

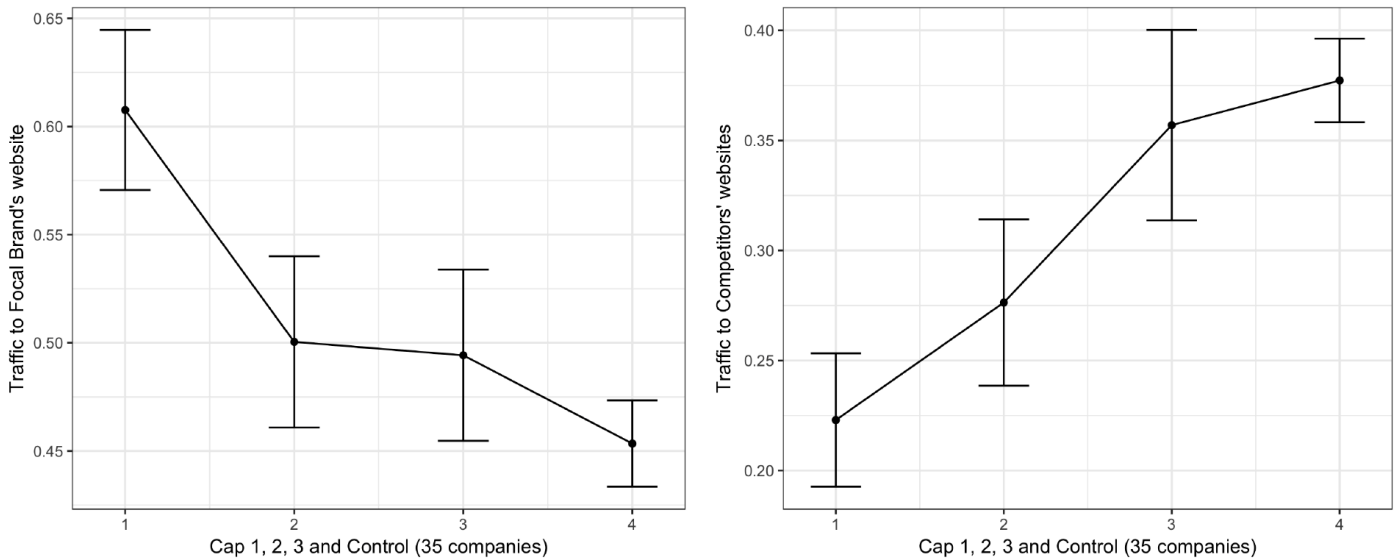
FIGURE 3.—Non-brand keyword region test. Panel (a) plots total purchases by users who clicked on an ad prior to purchase, which drops when the test commences in the test areas. Panel (b) plots three different measures of the difference between test and control regions before and after the test. The *y*-axis 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 2023): 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.



(a) Focal Brand's Traffic (b) Competitor's Paid Traffic

Point estimates are computed for each brand using the frequency estimator defined in Section 3.4. Results are averaged across brands. Error bars give +/- 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.

Simonov, Nosko & Rao (2017)

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

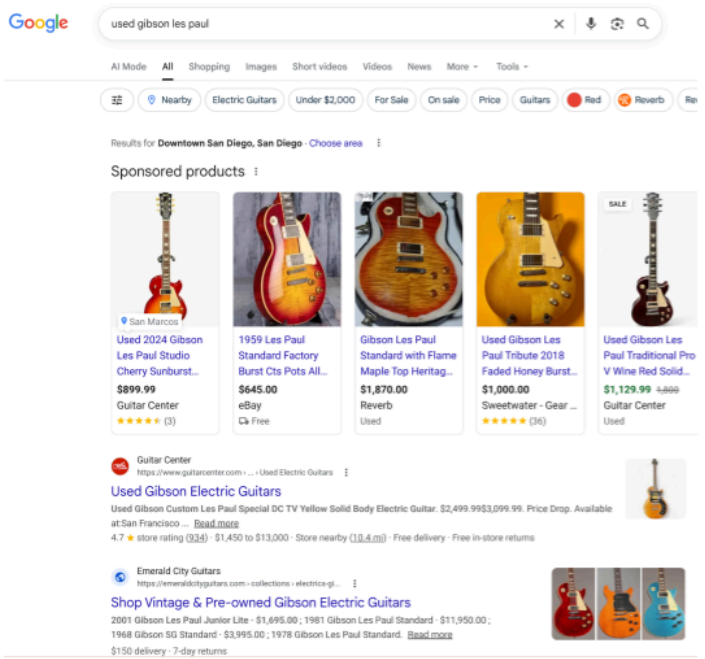
February 24, 2018

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?)

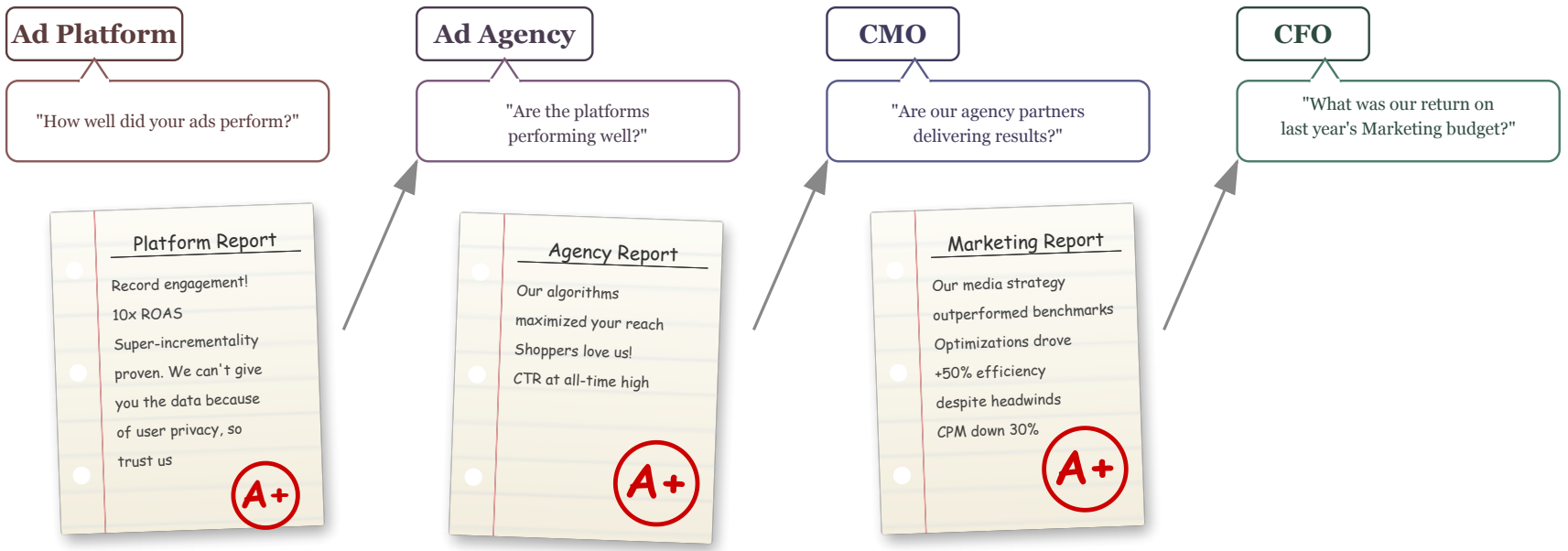
Simonov & Rao (2018)

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 SERP real estate and started advertising again. That's exactly what 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

adexchanger

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.

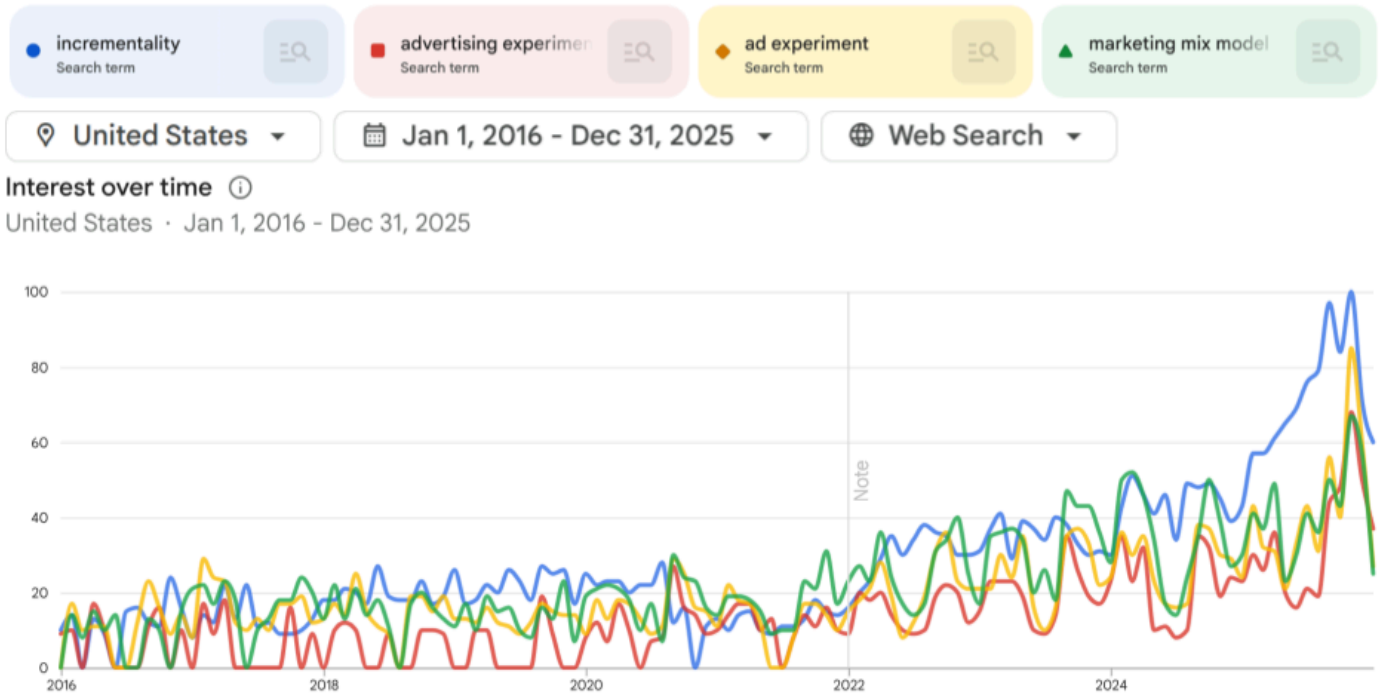
The platforms have heard the MMM requests, and answered in 2024.

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

> Incrementality & MMM were trends #1 & #2; the only other trend was e-commerce metric proliferation.

AdExchanger (2024)

Google Trends: Ad Experiments



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

Google Trends

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

Rubin Causal Model

Why Care About Causal Effects?

- We want to maximize profits $\Pi = \Sigma_i \pi_i(Y_i(T_i), T_i)$
- 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 *Yi*(*Ti* =1) - *Yi*(*Ti* =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 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 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

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?
- 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 depends on whether the organization will act on what it learns.

Advertising Measurement

- What we measure, challenges, classic eBay measurement case

Measurement of What?

advertising noun

ad·ver·tis·ing ˈad-vər-ˌtī-zin

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*

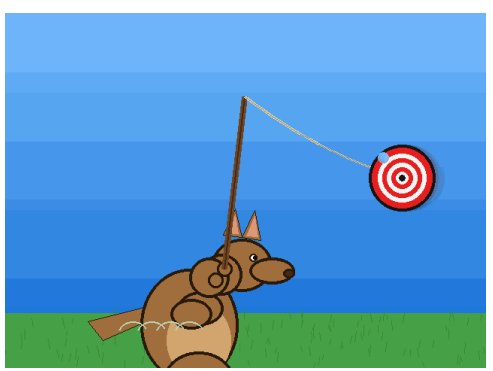
Many people use ‘advertising’ to refer to all commercial speech. In marketing, ‘advertising’ refers to paid media, as distinct from owned media (e.g., organic social, website, emails, direct mail) & earned media (e.g., reviews, news stories). Paid media implies that a ‘publisher’ generated the advertising opportunity by attracting consumer attention; controls the sale; and may constrain the advertiser’s message. Two main ad types:

- Performance advertising: Campaigns designed to stimulate short-run measurable response. Could be any funnel stage, including awareness, consideration, visitation and/or sales.
- Brand advertising: Campaigns designed to stimulate long-run response or change attitudes. Measurable in multiple ways, but measurement will usually be incomplete.

Ad Measurement

- *Advertising measurement* quantifies ad delivery, exposure and outcomes to improve advertising efforts
  - → Our focus here is on outcomes/conversions, as these inform future budget decisions
  - → Delivery and exposure matter most for brand ads. Principles include independence and transparency in measurement; these must be checked, cannot be assumed
- Advertising measurement is hard because ad effects depend on ad content, context, timing, targeting, current market conditions, 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)

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

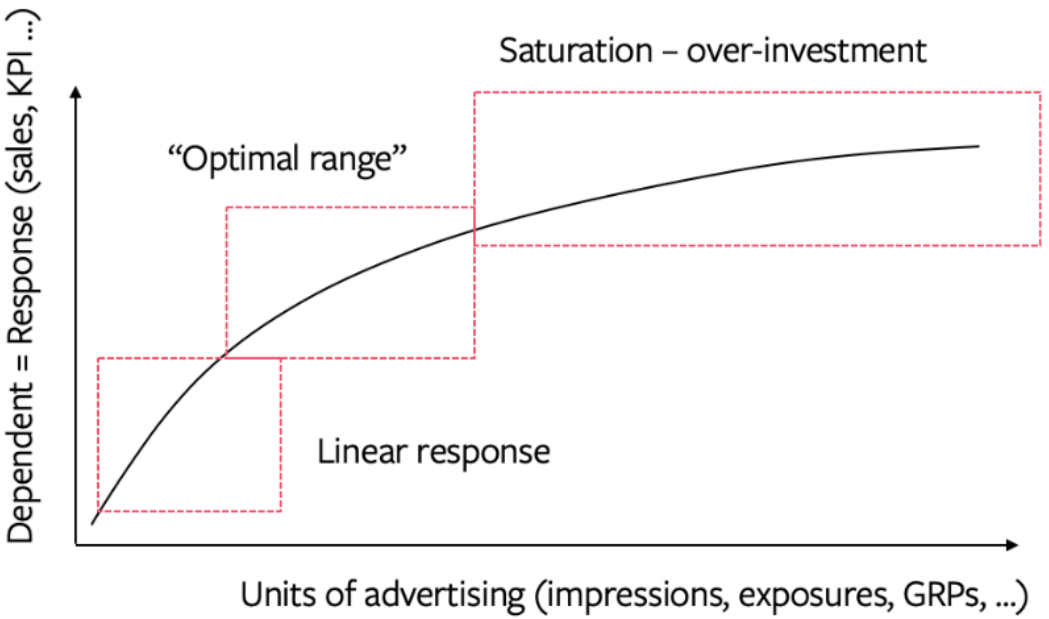
Increasingly, we report incremental ROAS (iROAS) if we have causal identification, i.e. we isolated causal ad effects

- ROAS ≠ iROAS because attribution is usually correlational

We also should measure delivery and funnel-wide KPIs, e.g. brand metrics, visits, add-to-cart, sales, revenue, …

- We usually get economies of scope in measurement

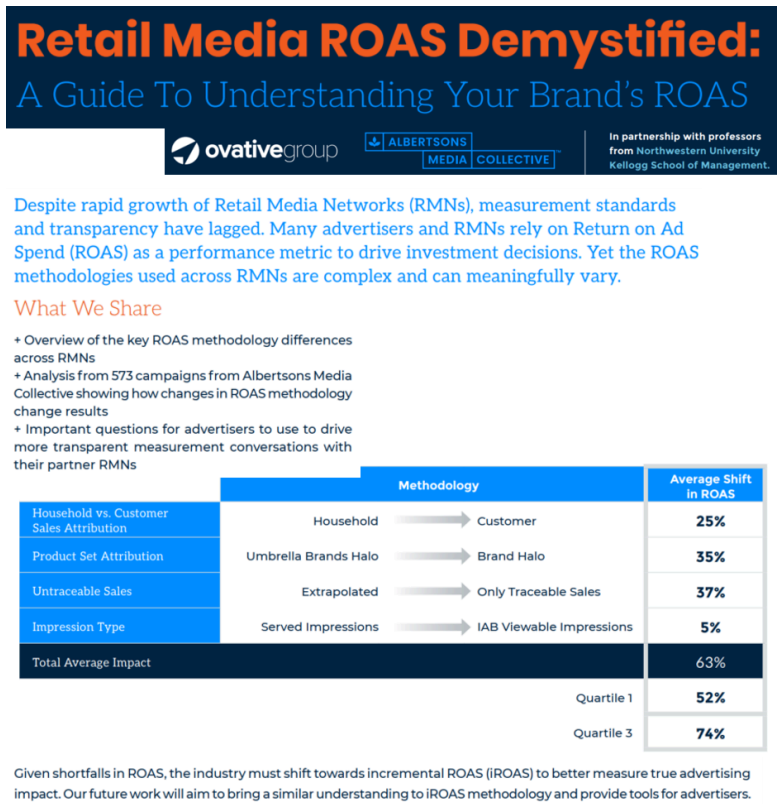
Diminishing Returns



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 ≠ 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?)

Meta (2024) | Google (2024, pgs. 22-25) | Galiza (2025)

Albertsons: ROAS Varies with Measurement Choices



Retail Media ROAS Demystified:

A Guide To Understanding Your Brand's ROAS

ovativegroup | ALBERTSONS MEDIA COLLECTIVE

In partnership with professors from Northwestern University Kellogg School of Management.

Despite rapid growth of Retail Media Networks (RMNs), measurement standards and transparency have lagged. Many advertisers and RMNs rely on Return on Ad Spend (ROAS) as a performance metric to drive investment decisions. Yet the ROAS methodologies used across RMNs are complex and can meaningfully vary.

What We Share

+ Overview of the key ROAS methodology differences across RMNs
+ 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.

Albertsons Media Group (2025)

Correlational Advertising Measurement

- Frameworks, Problems, Facebook study, Why Correlations Persist

Correlational Ad Measurement

- Correlational advertising measurement is defined by the absence of a treatment/control logic 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

Correlational advertising measurement is not defined by an analytical or modeling technique, but we will review 3 common approaches.

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 calculated using experimental data. Otherwise they reflect 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.

Google | Meta

2. Regression, usually controlling for other observables

Get historical data on Y_i and T_i and run a regression

- *i* could index individuals, places, times, or combinations
- Many frameworks exist, including least squares, vector autoregressions, Marketing Mix Models (MMM), bayesian frameworks
  - → Google's CausalImpact R package is popular
  - → We will go deeper on MMM later

3. 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 a rule to attribute purchases to touchpoints
  - → Single-touch rules: Last-touch, first-touch
  - → Multi-touch rules: Fractional credit, Shapley
- MTA algorithm searches for touchpoint parameters that best-fit the conversion data given the rule
  - → Credit then informs future budget allocations across touchpoints
  - → MTA is designed to maximize attributions; MTA often disregards non-purchasers
  - → MTA assumes touchpoints are the *sole* drivers of conversions
- 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.

Steel-manning Corr(ad,sales)

- Corr(ad,sales) should contain signal
  - → If ads cause sales, then corr(ad,sales) > 0 (probably) (we assume)
- Some products/channels just don't sell without ads
  - → E.g., Direct response TV ads for 1-800 phone numbers
  - → Career professionals say advertised phone #s get 0 calls without TV ads, so we know the counterfactual (what is it?)
- 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: corr(ad,sales) > 0 does not imply a causal effect of ads on sales; it could be that the ads got shown to the most loyal customers

Problem 1 with Corr(ad,sales)

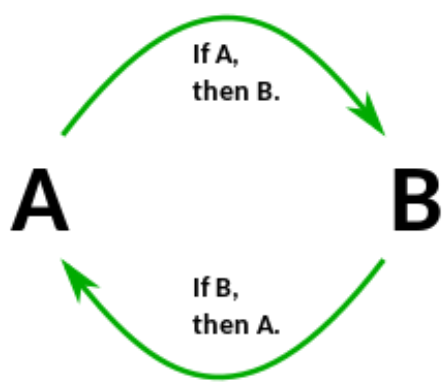
- Advertisers try to optimize ad campaign decisions
  - → E.g. surfboards in coastal cities, not landlocked cities
- If ad optimization increases ad response, then corr(ad,sales) will confound actual ad effect with ad optimization effort
  - → More ads in San Diego, more surfboard sales in San Diego. But would we have 0 sales in SD without ads?
  - → Corr(ad,sales) usually overestimates the causal effect, encourages overadvertising

> Google's Chief Economist explains in greater detail.

Problem 2 with Corr(ad,sales)

- How do most advertisers set ad budgets? Top 2 ways historically:
  1. Percentage of sales method, e.g. 1%, 3% or 6%
     → That's why ads:sales ratios are so 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 Corr(ad,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 react 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 platform 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.").

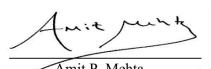
264. Google therefore endeavored to raise prices incrementally, so that advertisers would view price increases as within the ordinary price fluctuations, or "noise," generated by the auctions. *See, e.g.*, UPX507 at .023 (describing a 10% CPC increase as "safe" because it is "within usual WoW noise");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

Gordon, Moakler & Zettelmeyer (2023)

Gordon, Moakler & Zettelmeyer (2023): Data

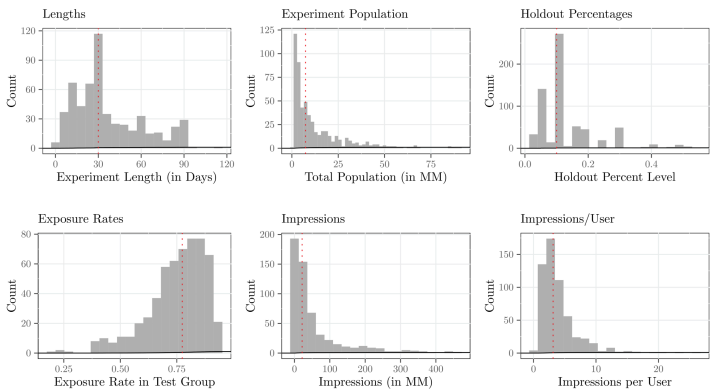
3.1. Experiment Selection

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.

Gordon, Moakler & Zettelmeyer (2023)

Gordon, Moakler & Zettelmeyer (2023): Figures

Figure 2. ATTs Across All RCTs

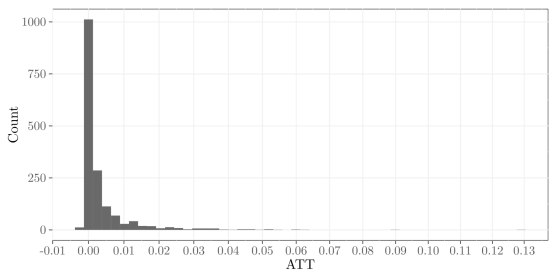
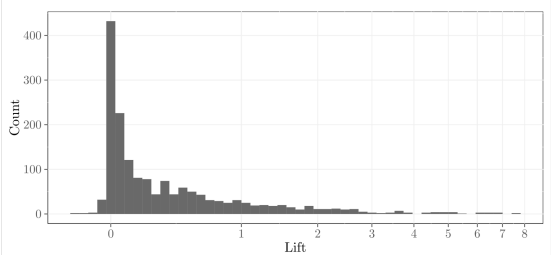


Figure 3. Lifts Across All RCTs

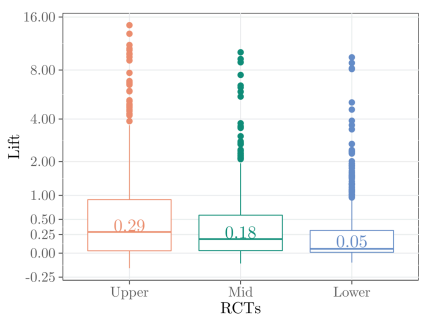


Note. Excludes the top 1% of lifts.

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,

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

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



Most of the ad experiments shows 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); this is common as price or other factors can discourage sales during the shopping process.

Gordon, Moakler & Zettelmeyer (2023)

Gordon, Moakler & Zettelmeyer (2023): Results

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.

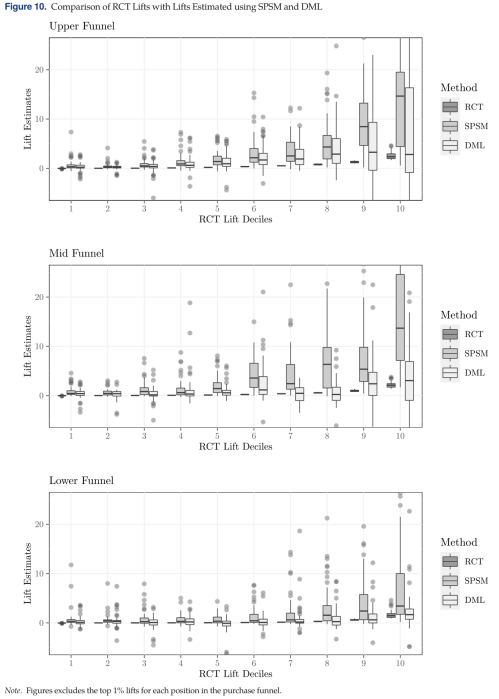
In standard propensity score matching, we find the one (or more) unexposed users with the closest propensity score to each exposed user to estimate the treatment effect. Since this is very computationally burdensome, instead, we stratify on the propensity score: After estimating the propensity score,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



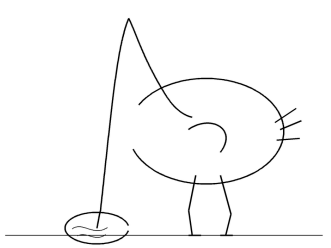
Note. Figures excludes the top 1% lifts for each position in the purchase funnel.

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.”

Gordon, Moakler & Zettelmeyer (2023)

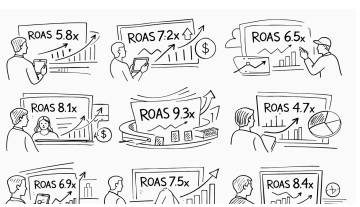
Why Are Some Teams OK with Corr(ad,sales)?

1. Some worry that if ads go to zero → 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 or CMO?



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

4. Managing analytics well requires skill and discipline
    - Managers must understand how to integrate experiment results into decisions
    - Analysts must have causal inference skillsets
    - Organization must tolerate failure in search of data-driven incremental improvements
5. Platforms often provide correlational ad/sales estimates
    - Which is larger, correlational or experimental ad effect estimates?
    - Which one might many client marketers prefer?
    - Platform estimates are typically “black box” without neutral auditors
    - ““Nobody ever got fired for buying IBM.” [Amazon, Google, Meta]
6. Historically, agencies usually estimated ROAS
    - Agency compensation usually relies on 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 a treatment/control logic to isolate causal advertising effects from confounding drivers of sales
  - → We can run experiments to create random variation in advertising treatments
  - → We can analyze naturally occurring random variation, AKA quasi-experiments
- *Causal* means we isolate the treatment effect from confounds
  - → Sometimes misinterpreted as evidence consistent with a hypothesis

A regression may be either causal or correlational depending on the nature of the advertising data variation.

Popular ad experiments

1. Randomly assign ads eligibility / holdout to customer groups
    - Pros: AB testing is easy to understand, rules out alternate explanations
    - Cons: Can we trust the platform's "black box"? Will we get the data and all available insights?
2. Randomize messages within a campaign. Mine competitor messages in ad libraries for ideas
    - Often a great place to start as it's nominally free.
    - Advertising professionals frequently believe that creative messages are first-order drivers of ad effects (e.g., Circana 2023, Kantar 2024, Magna+Yahoo 2025)
3. Randomize bids and/or consumer targeting criteria
4. Randomize budget across platforms, publishers, times, places, behavioral targets, contexts

> Platforms usually tune ad delivery algorithms to maximize post-advertising conversions, not incremental conversions. That is why changing a campaign attribute usually induces nonrandom variation in advertising treatments.

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
3. Compliance
    - Treatments assigned are treatments received
    - Ad blocking can induce noncompliance, which we usually resolve by estimating Intent-to-Treat effects
4. Statistical Independence
    - Random assignment of treatments to units. "Balance tests" help to check
    - When platform algorithms distribute ads nonrandomly, we call it "divergent delivery"

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
- Can we agree on the opportunity cost of the experiment? "Priors"
- How will we act on the (uncertain) findings? Have to decide before we design. We don't want "science fair projects"
  - → Simple example: Suppose we estimate iROAS at 1.5 with 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

- On Meta, Lift Tests are true experiments, whereas “A/B Tests” confusingly do not control for algorithmic user/ad selection (called “divergent delivery”). In Meta’s “A/B Test,” consumers are randomized to treatment eligibility, rather than treatment itself, “treating” the algorithm that determines ad delivery, not the consumers themselves. Burtch et al. (2025) go deeper.
- 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
- Prominent exception: Ghost ads, an ingenious system to randomly withhold ads from auctions and maximize experiment efficiency

Haus (2025) | Braun & Schwartz (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

Quasi-experiments Vocabulary

- Model: Mathematical relationship between variables that simplifies reality, e.g. y = xβ + ε
- Identification strategy: Set of assumptions that isolate a causal effect $\frac{\partial Y_i}{\partial T_i}$ from other factors that may influence Y_i

  → A strategy to compare apples with apples, not apples with oranges
- 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 say we "identify" the causal effect if we have an identification strategy that reliably distinguishes $\frac{\partial Y_i}{\partial T_i}$ from possibly correlated unobserved factors
- If you estimate a model without an identification strategy, you should interpret the results as correlational

  → This is widely misunderstood. We can learn from correlational models, but too many people mistakenly infer causality

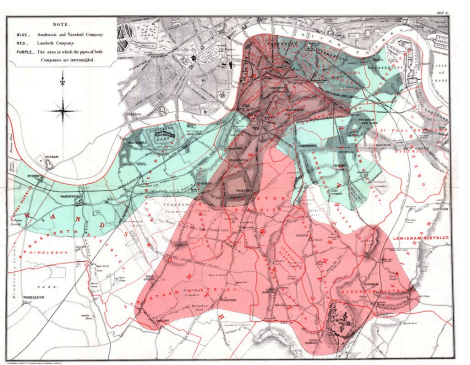
Sant'Anna (2026) maintains a free online resource for difference-in-differences theory and estimation code.

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 would let him test his theory.

> 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 maintained its nearby intake location.
>
> 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 provider and by time. He calculated that cholera death rates in 1849 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.
>
> If household cholera risk factors were unrelated to drivers of water company selection, then the Southwark and Vauxhall cholera death rate in 1854 estimated the Lambeth counterfactual, showing that cleaner water meaningfully reduced cholera death rates. This discovery came before the germ theory of disease in the 1860s or the modern development of experimental methods. (What are the two diffs?)

Cunningham (2026)



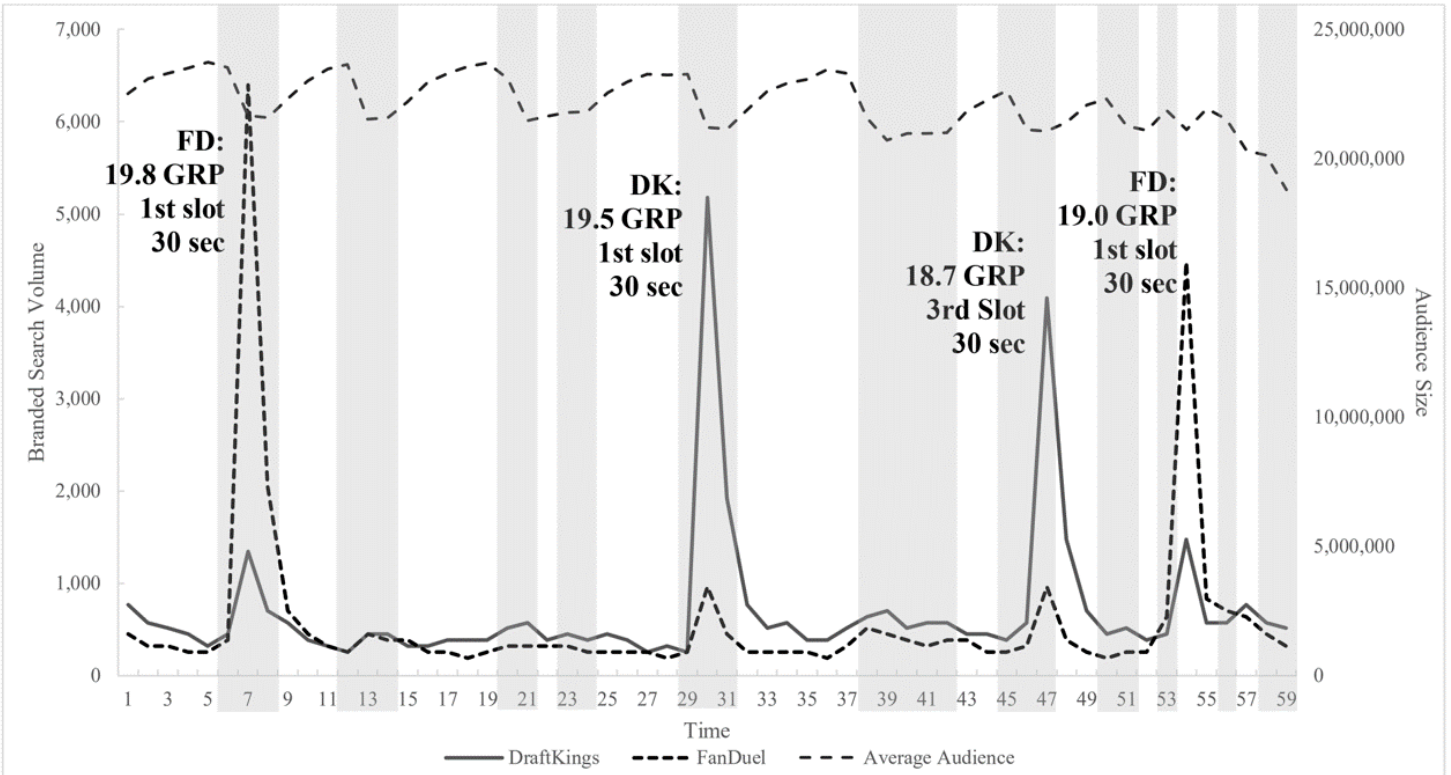
Ad/Sales: Quasi-experiments

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 started, stopped or pulsed 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)
- Exogenous changes in addressable market, web traffic, other factors

DFS TV Ad Effects on Google Search



I made this graph showing DraftKings and FanDuel branded keyword search volume from 9:01-9:59pm E.S.T. during the 2015 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 it was the presence of DFS ads, not just the absence of the game.

Du et al. (2019)

Ad/Sales: Quasi-experiments (2)

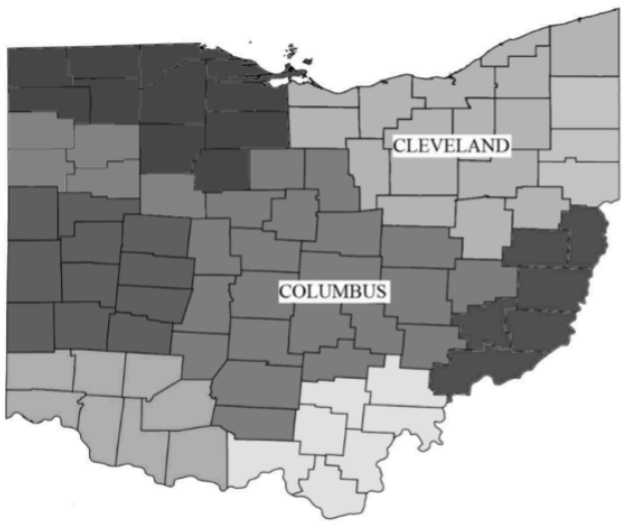


FIG. 5.—Ohio and its DMAs

Shapiro et al. (2021) used a county-border approach to identify how local TV advertising affected package goods sales. The idea is that geographic media market boundaries are drawn based on broadcast signal patterns based on differences from city centers, such that consumers living on either side of the boundary are very similar. Therefore, boundary county sales can predict what in-market counterfactual sales would have been in the absence of advertising.

See also Shapiro (2018)

Experiments vs. Quasi-experiments

- Experimentalists and quasi-experimentalists differ in beliefs, cultures & training, not unlike Bayesians vs. frequentists
- 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 without validity
  - → Range from challenging to impossible to validate
- Experiments & quasi-experiments should be “yes-and-when-valid,” not “either-or”

Marketing Mix Models

- Definition, components, considerations, and open-source tools

Marketing Mix Models (MMM)

- The “marketing mix” consists of the 4 P’s
  - → Product line, length and features; price & promotions; advertising, PR, social media, reviews; retail distribution
- 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/marcom channels or publishers
  - → MMM and mMM share many attributes and techniques
- MMM goal is to evaluate past marketing ROAS by channel and support future budgeting decisions

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

MMM Components

- MMM analyzes aggregate data, usually 3-5 years of weekly or monthly intervals, usually across a panel of geographic markets
  - → Aggregate data are privacy-compliant & often do not require platform participation; helps explain MMM comeback
- Predictors include ad spending/exposures by ad type; outcomes measure sales, volume or revenue
- MMM usually controls for (a) trends, (b) seasonality, (c) macroeconomic factors, (d) known category-specific demand shifters, (e) diminishing marginal returns, (f) possibly long-lasting advertising effects ("carryover")
- Outputs include ad elasticities, ROAS measures, sales predictions based on counterfactual budget reallocations
- MMM parameters can be interpreted causally if and only if adspend data are generated with a randomization strategy built in

MMM Considerations

- Data availability, accuracy, granularity and refresh rate are all critical
- MMM requires sufficient variation in marketing predictors, else it cannot estimate coefficients
- “Model uncertainty”: Results can be strongly sensitive to modeling choices, so we usually evaluate multiple models to gauge sensitivity to alternate assumptions
- MMM results are correlational without experiments or quasi-experimental identification
  - → Correlations can be unstable; Bayesian estimation can help regularize
  - → MMM results can be causal if you induce exogenous variation in ad spending
  - → MMM results can be calibrated using causal measurements, e.g., by using experiments to calibrate Bayesian prior beliefs, or by rewarding models for conformance to external incrementality estimates

MSI White Paper | Sellforte

Open-Source MMM Frameworks

- Meta Robyn (2024). Excellent training course
  - → Cool features: Causal estimate calibration, Set your own objective criteria, Smart multicollinearity handling
- Google Meridian (2025). Excellent self-starter guide
  - → Cool features: Bayesian implementation, Hierarchical geo-level modeling, reach/frequency distinctions
  - → Robyn & Meridian both include budget-reallocation modules

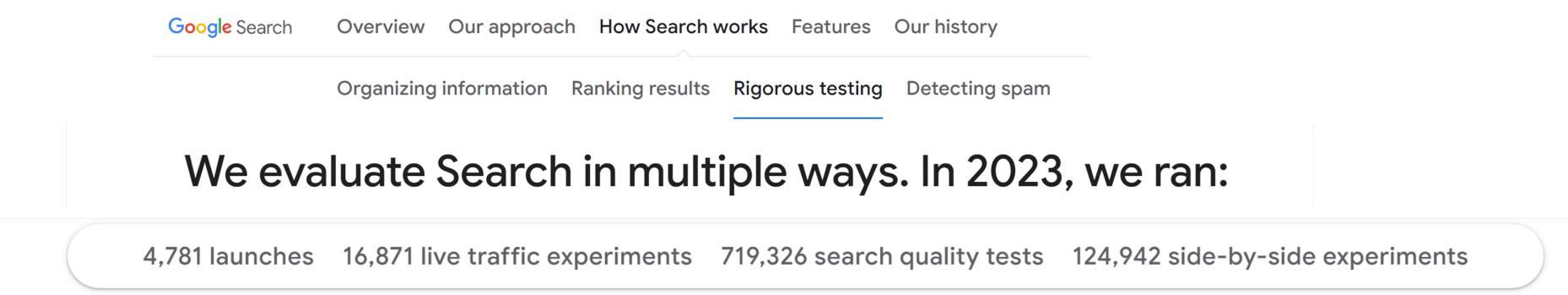
Others: PyMC-Marketing, mmm_stan, BayesianMMM

Also relevant: MMM data simulator

Putting ideas into practice

- Who’s doing what and why

Who Tests the Most?



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.

HARVARD BUSINESS REVIEW

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

Runge (2020) | Runge et al. (2020)

Advertising Experiment Effectiveness

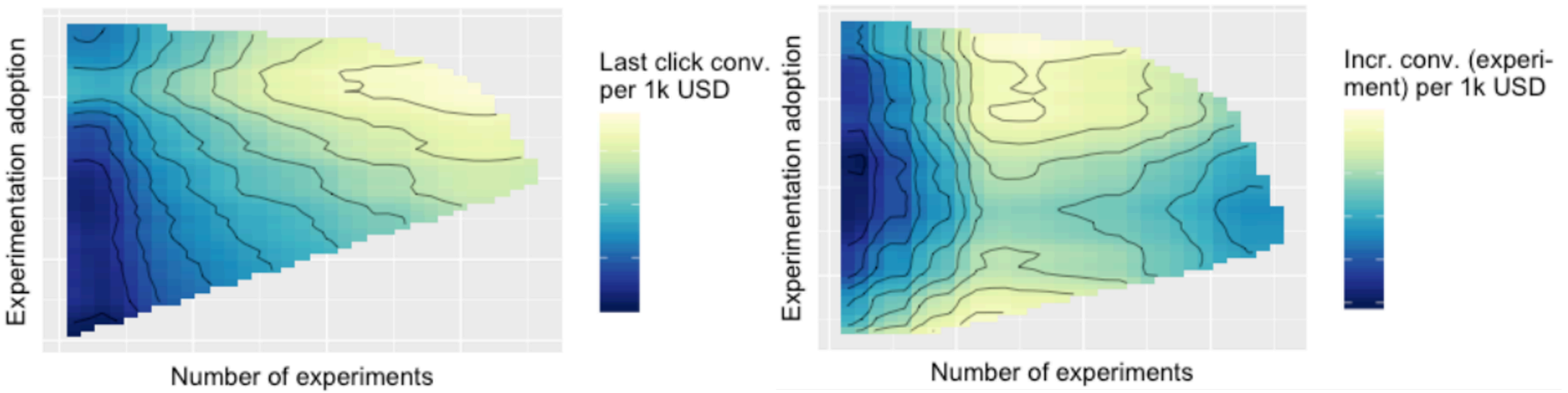


FIGURE 3: Partial dependency plots (via a random forest regressor) to isolate effect of experiment use by firms during the outcome year. Column one considers the number of experiments carried out by the firm. Column two considers both the number of experiments as well as how early the firm adopted experimentation 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.

Runge et al. (2020)

KANTAR

Understanding "Marketing Mix Modeling" (MMM) adoption dynamics in the industry

In an increasingly complex measurement landscape, large advertisers are leaning on Marketing Mix Modeling (MMM) more than ever — not just as a strategic compass, but as a foundational tool in a multi-solution ecosystem. Yet with growing reliance comes growing responsibility: advertisers face fractured results, mounting pressure to align tools, and persistent pain points around cost, clarity, and brand equity impact.

Kantar worked with Meta to conduct a comprehensive global research study on the state of play for measurement practices among large advertisers, with a deeper focus on Marketing Mix Modeling (MMM) adoption dynamics and its role in decision-making. Based on a total of 1,935 interviews conducted with measurement professionals from companies investing over $1 million in digital marketing annually, we uncovered several important findings and implications through this research:

By

Kirsten Hjort

Megan DeMello

Source: May 2025, n=1,935 decision makers across regions, industries and company sizes, investing over $1MM in Digital Advertising

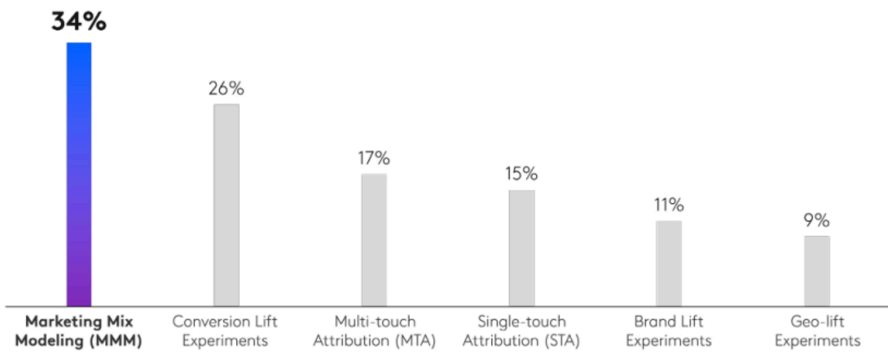
Kantar (2025)

Kantar 2025: Measurement Methods

#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%**

Kantar (2025)

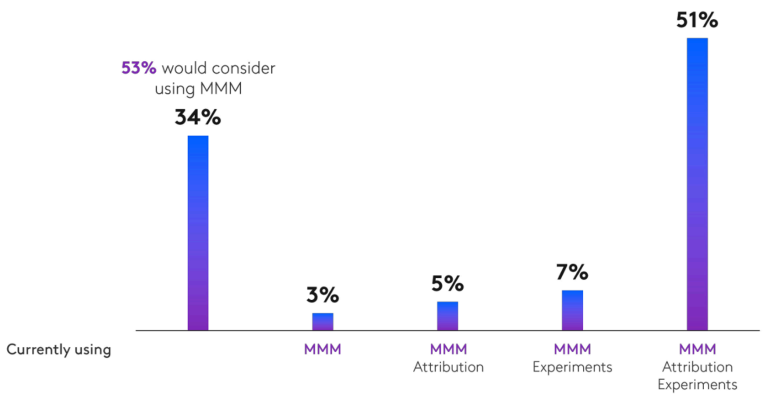
Incrementality Adoption & Barriers

#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.

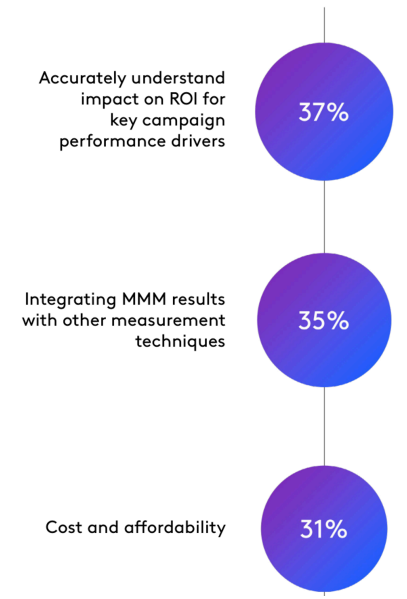
This increases to 4.2 for the largest advertisers.

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



The number one challenge - 37% of MMM users cite the need for greater granularity and actionability in understanding key drivers of campaign performance. This suggests that advertisers are looking for MMM to provide ever more detailed insights into their efforts, as well as actionable recommendations for improvement.

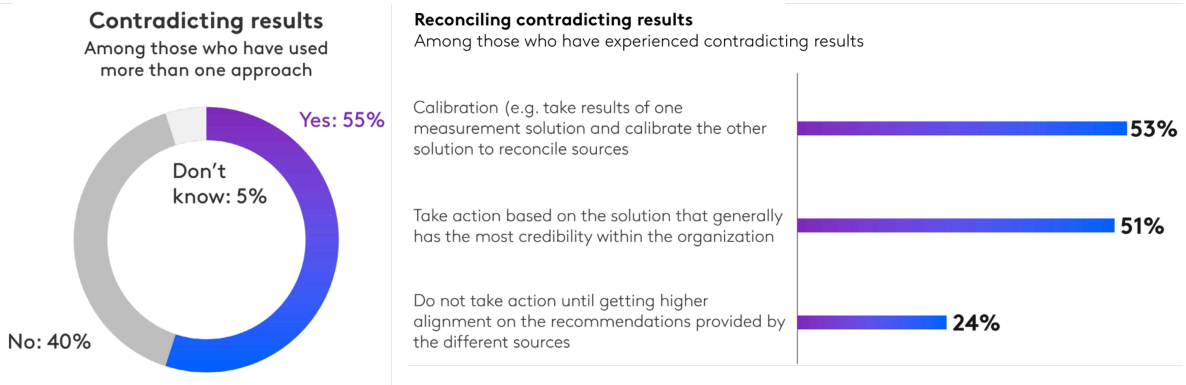
While we see multi-modal stacks as well as validation and calibration as a strong trend, over a third of respondents highlight the challenge of integrating MMM results with other measurement techniques, indicating a need for practical guidance.

Cost and affordability is an issue for users, and also a barrier for non-MMM users who tend to be more cost-conscious with fewer in-house resources compared to MMM users.

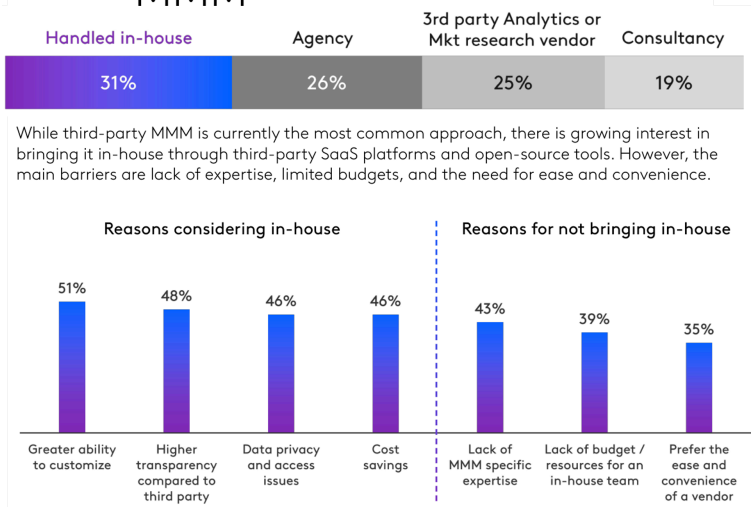
Kantar (2025)

MMM Usage & Priorities

#4 Maximizing accuracy through validation and calibration

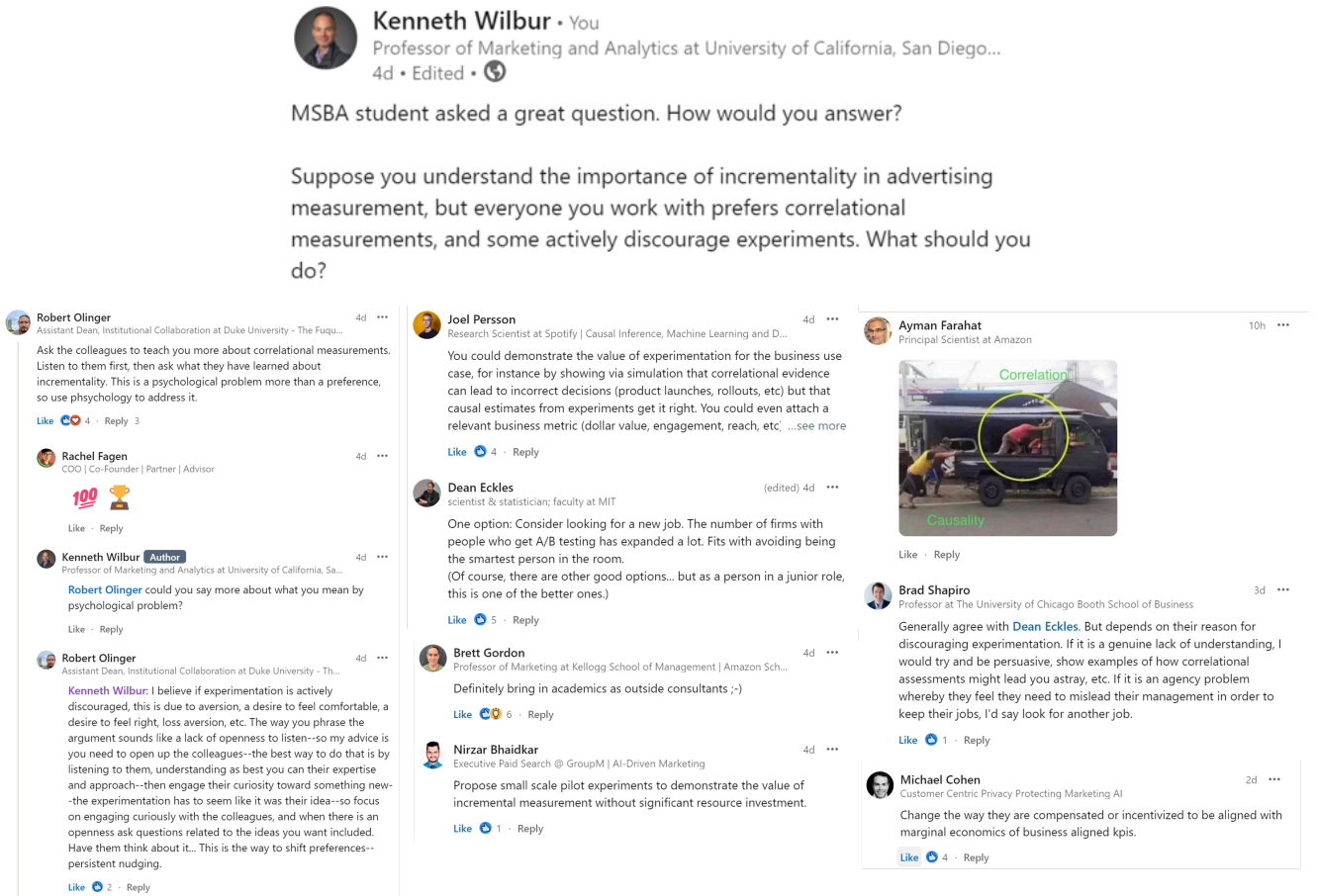


#5 Growing trend of in-housing MMM



Kantar (2025)

What would you do?



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.

Takeaways

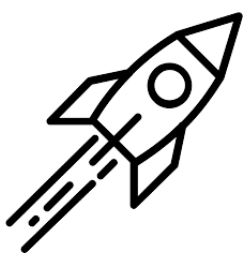
- Fundamental Problem of Causal Inference: We can't observe all data needed to optimize actions. This is a missing-data problem, not a modeling problem.

  → Common remedies: Experiments, Quasi-experiments, Correlations, Triangulate; Ignore

- 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

Going Deeper

- Paparo (2025): Insider's account of programmatic advertising development from 2000-2025
- Johnson (2023): Covers frequent problems in online advertising experiments
- 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
- Athey & Imbens (2024): on designing complex experiments
- Barajas et al. (2021): Online Advertising Incrementality Testing And Experimentation: Industry Practical Lessons



Acknowledgements

- Joel Barajas, Rick Bruner, Peter Daboll, Tom Flanagan, and Prabhath Nanisetty 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