Advertising Measurement

UCSD MGTA 451 — Marketing

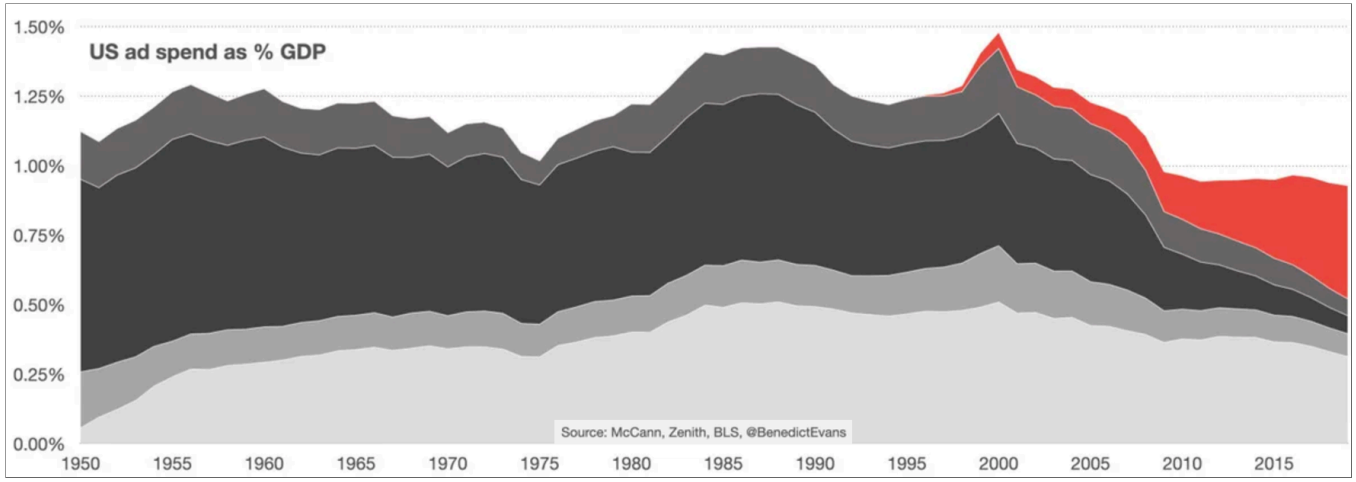
Kenneth C. Wilbur

Agenda

- Advertising Importance
- Causality
- Fundamental Problem of Causal Inference
- Advertising Measurement
- Correlational Advertising Measurement
- Causal Advertising Measurement
- Industry practices
- Marketing Mix Models
- Career considerations

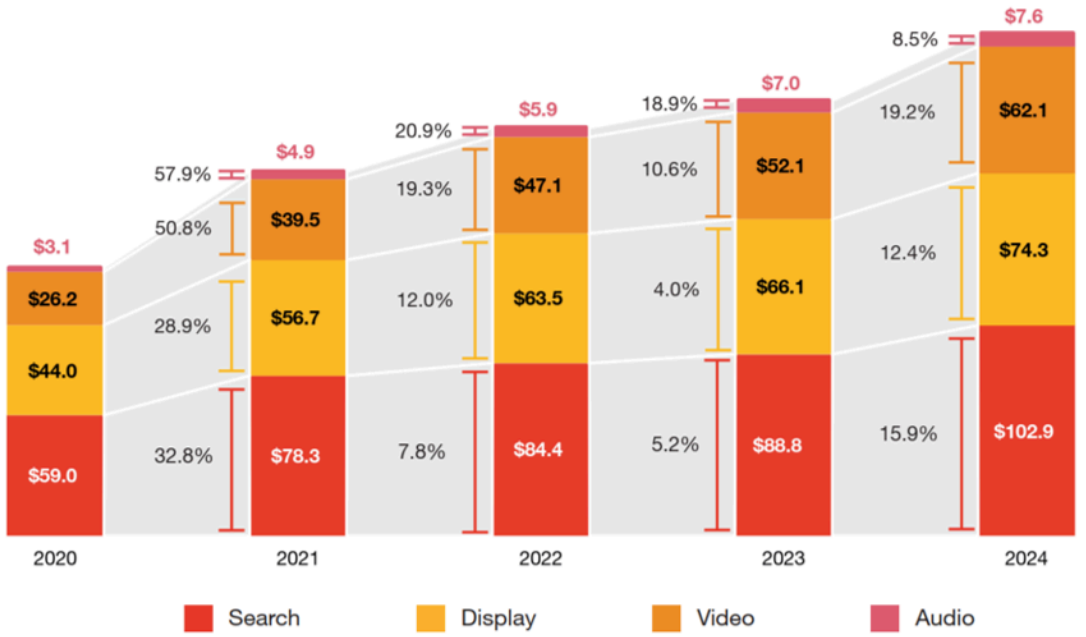
Advertising Importance

Publisher revenue since 1950

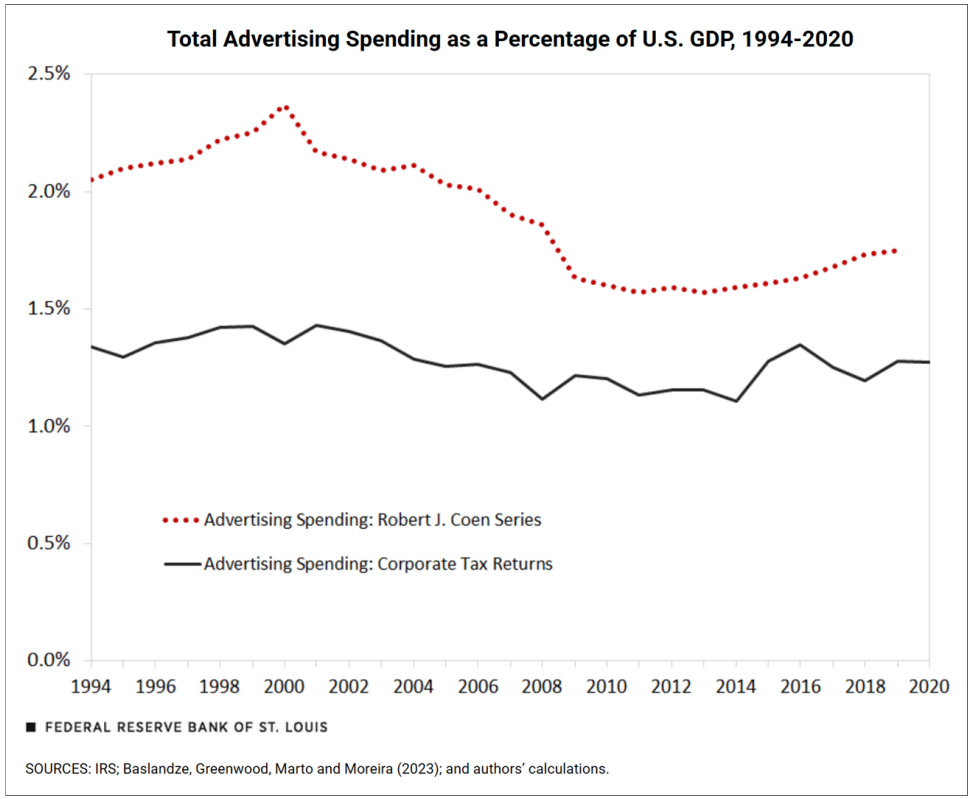


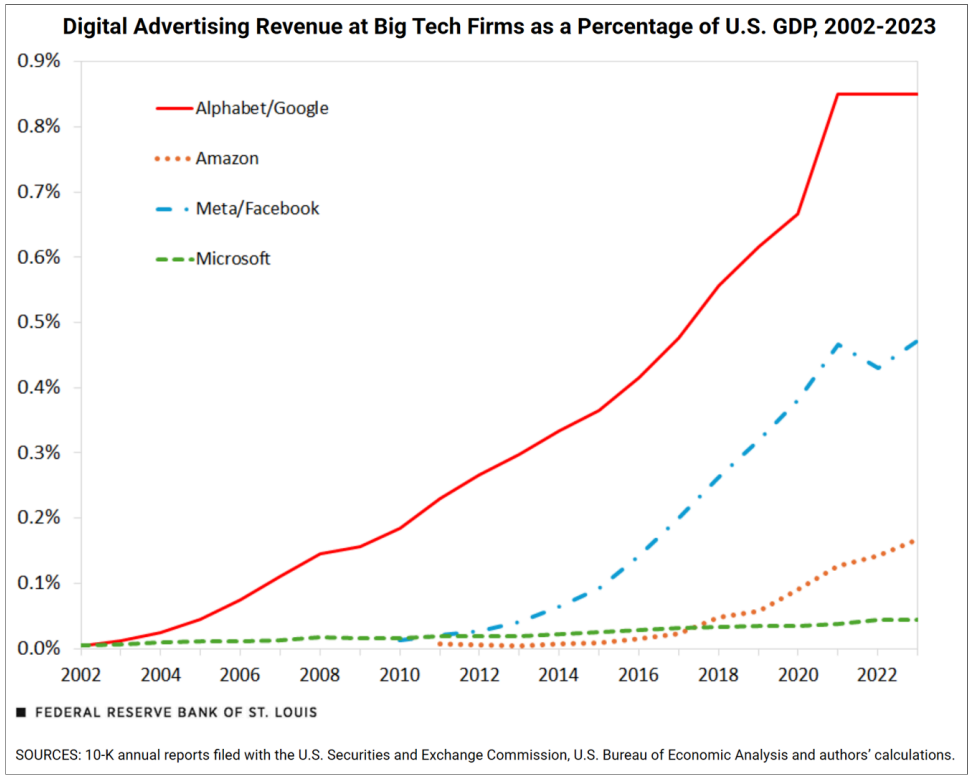
Growth by advertising format

Growth by advertising format (2020-2024) ($ billions)

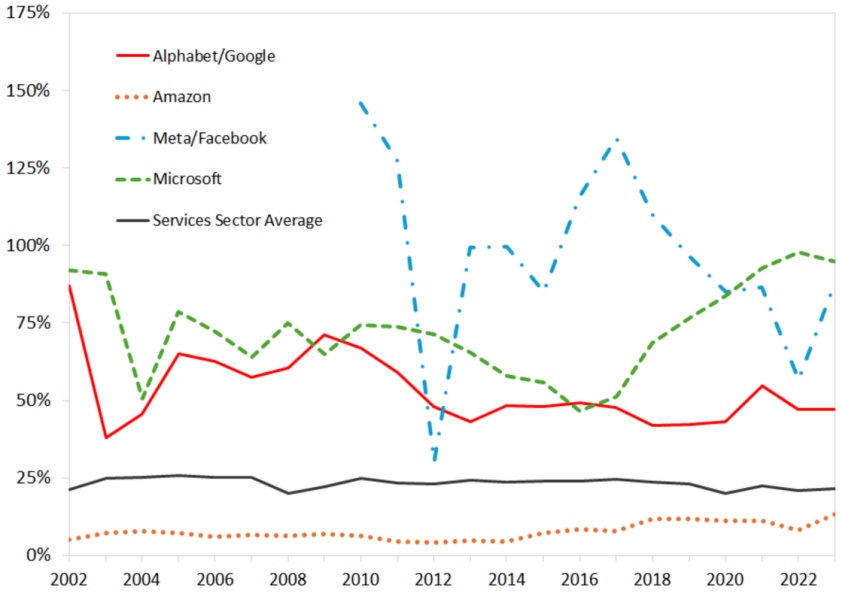


Source: IAB / PwC Internet Ad Revenue Report, FY 2024


Total Advertising Spending as a Percentage of U.S. GDP, 1994-2020
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Advertising Spending: Robert J. Coen Series
Advertising Spending: Corporate Tax Returns
FEDERAL RESERVE BANK OF ST. LOUIS
SOURCES: IRS; Baslandze, Greenwood, Marto and Moreira (2023); and authors' calculations.


Digital Advertising Revenue at Big Tech Firms as a Percentage of U.S. GDP, 2002-2023
0.9%
0.8%
0.7%
0.6%
0.5%
0.4%
0.3%
0.2%
0.1%
0.0%
Alphabet/Google
Amazon
Meta/Facebook
Microsoft
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
FEDERAL RESERVE BANK OF ST. LOUIS
SOURCES: 10-K annual reports filed with the U.S. Securities and Exchange Commission, U.S. Bureau of Economic Analysis and authors' calculations.

Profit Percentage among Big Tech Firms, 2002-2023



■ FEDERAL RESERVE BANK OF ST. LOUIS

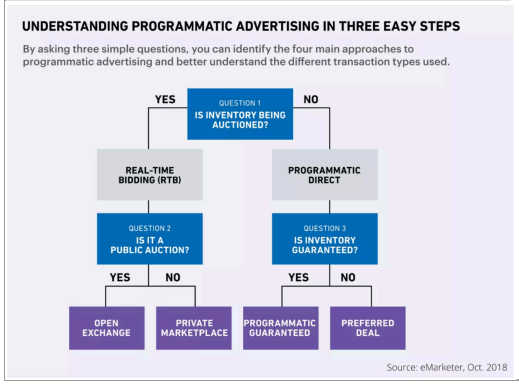
SOURCES: Compustat and authors' calculations.

NOTES: Profit percentage is equal to (revenue minus variable costs)/variable costs, where variable costs correspond to the cost of goods sold and to selling, marketing, general and administrative expenses. The services sector average is the weighted average of profit percentages for all firms in the services sector (NAICS code 42-81), where the weight for each firm is the sum of cost of goods sold and selling, marketing, general and administrative expenses over the total for services firms.

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



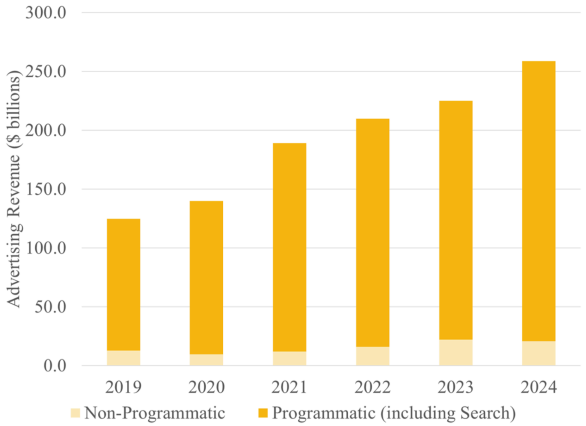
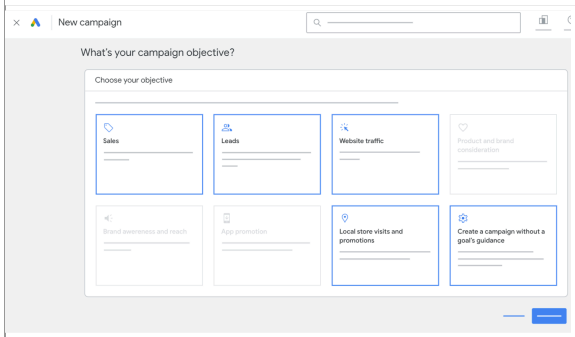


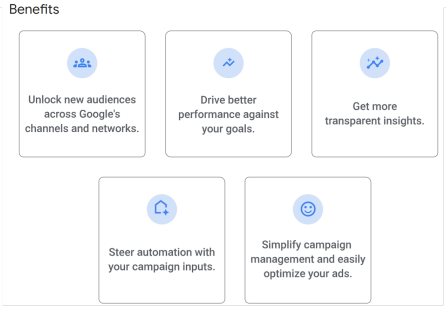
Figure 2. Internet Programmatic Advertising Revenues, 2019-24[48]

[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

Google Ads Help

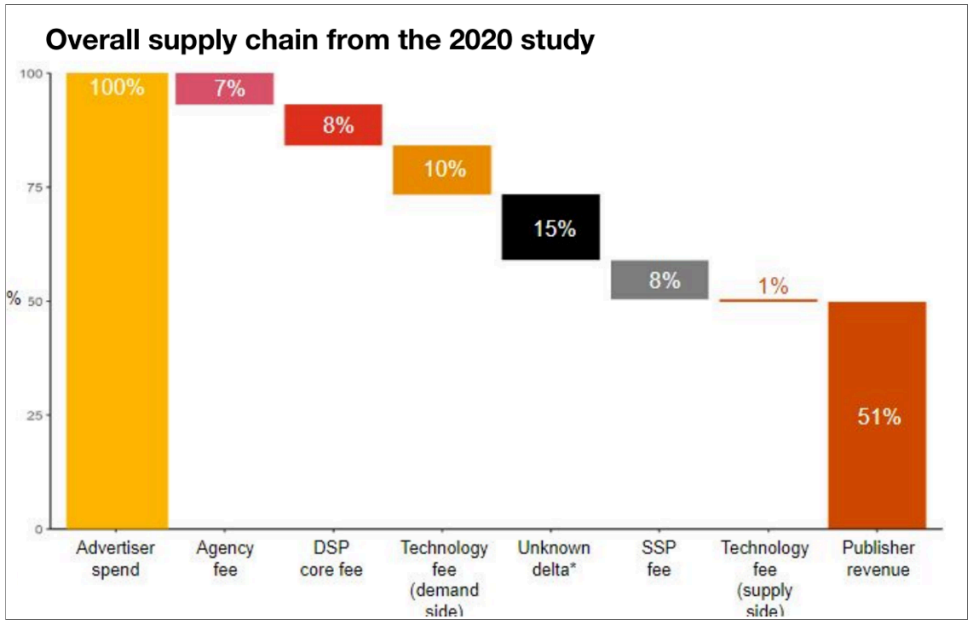
Performance Max is a goal-based campaign type that allows performance advertisers to access all of their Google Ads inventory from a single campaign. It's designed to complement your keyword-based Search campaigns to help you find more converting customers across all of Google's channels like YouTube,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



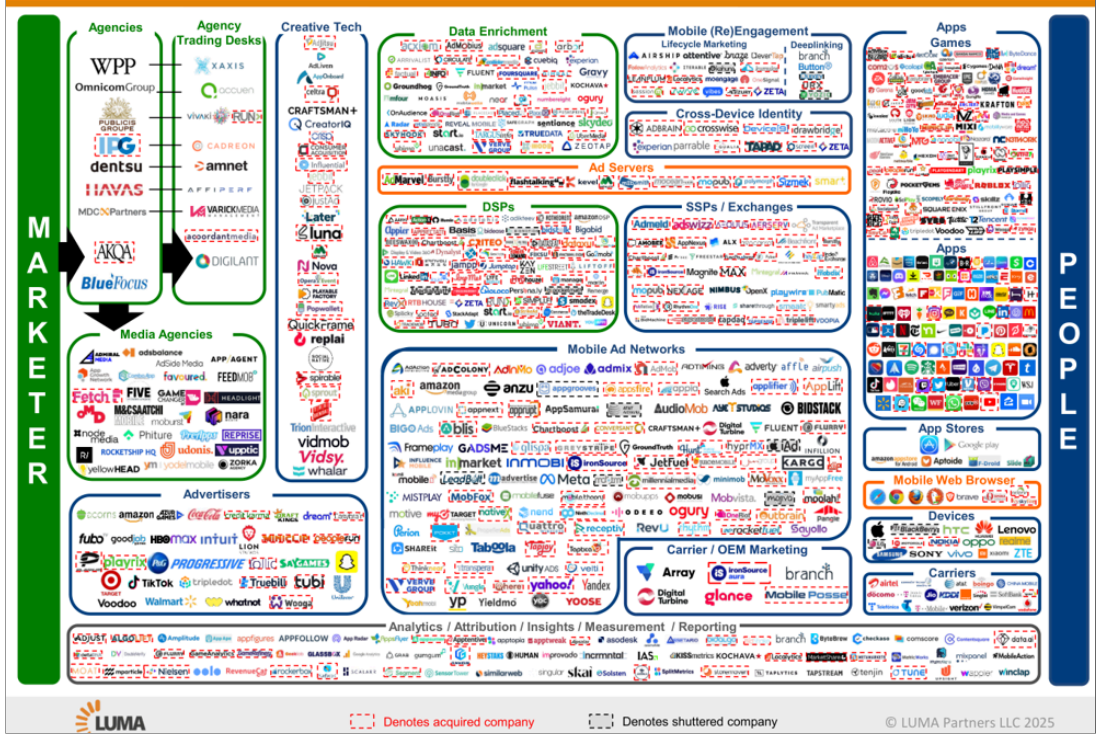


Google Ads & Commerce Blog

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.

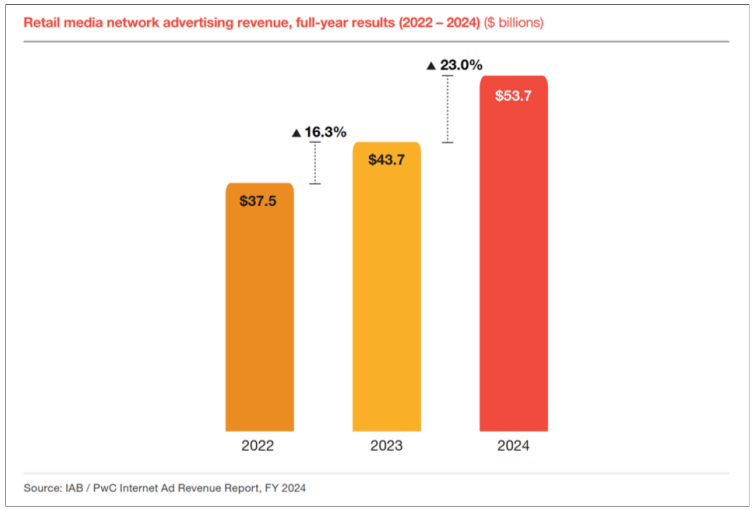

Overall supply chain from the 2020 study
100
75
50
25
0
%
100%
7%
8%
10%
15%
8%
1%
51%
Advertiser spend
Agency fee
DSP core fee
Technology fee (demand side)
Unknown delta*
SSP fee
Technology fee (supply side)
Publisher revenue

MOBILE LUMAscape



Denotes acquired company

Denotes shuttered company



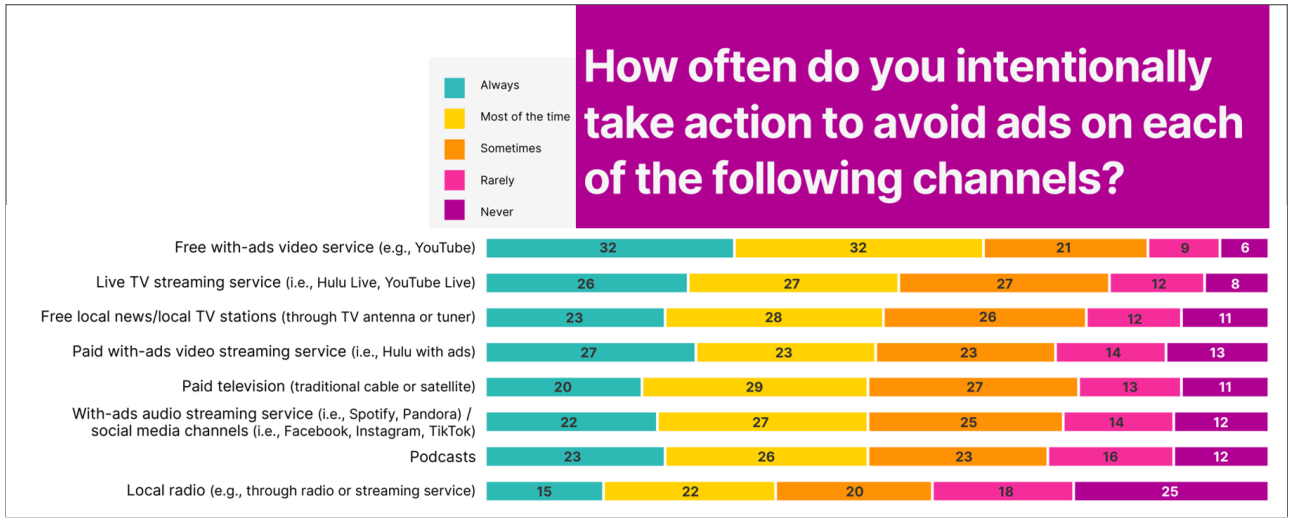
- Retail media networks now
    - provide data for ad targeting
    - sell sponsored product search listings
    - partner with publishers to sell display & video ads

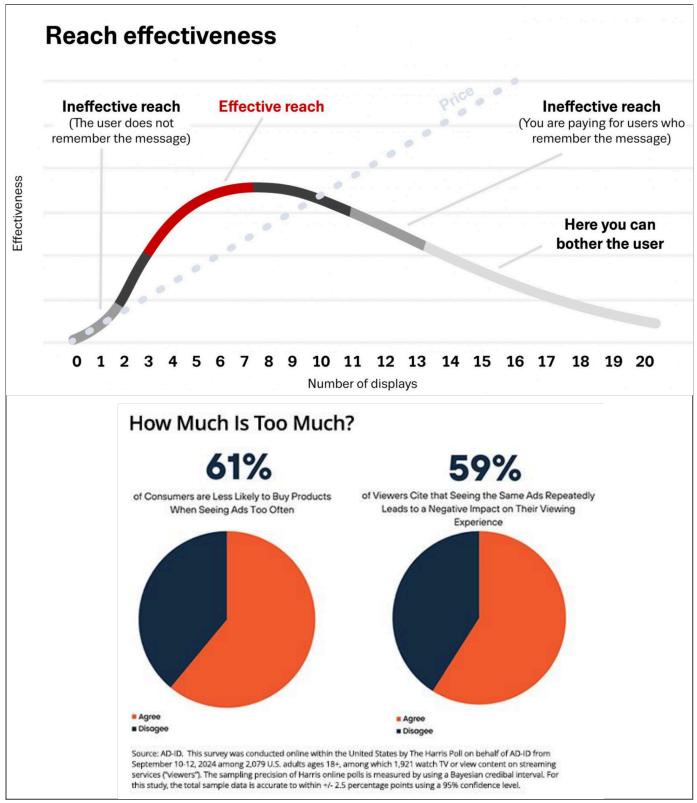
- Brand safety: Protect brands against negative impacts on consumer opinion associated with specific types of content [1]

```
- E.g., proximate to military conflict, obscenity, drugs, ...
- Ad verification vendors monitor ad placements
- Keyword blacklists and whitelists determine contextual ad bids
- Yet fraudulent ad sales may place brand ads in non-safe contexts
```

- Brand suitability: Help a brand avoid brand-inappropriate content, and identify brand-aligned content [1]

```
- Gray areas abound, e.g. cannabis-related content
- Brand safety may over-reach and defund controversial topics
```

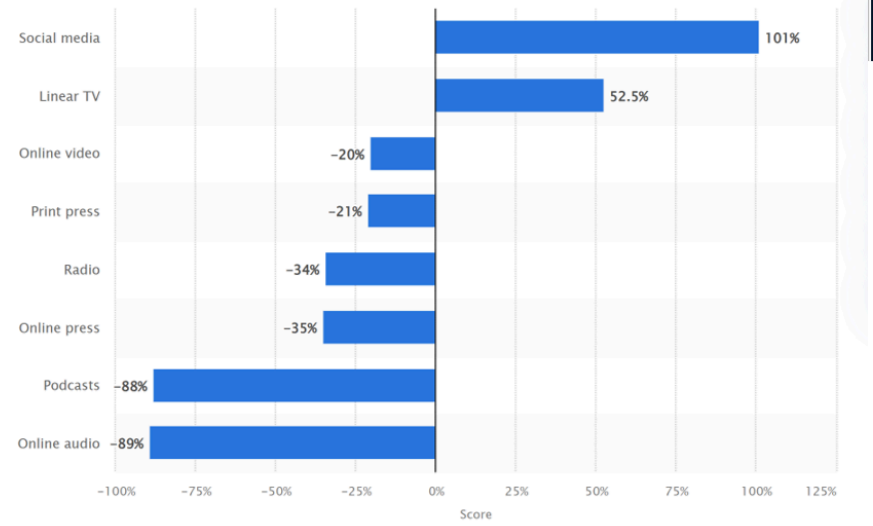

How often do you intentionally take action to avoid ads on each of the following channels?
Always
Most of the time
Sometimes
Rarely
Never
Free with-ads video service (e.g., YouTube)
32
32
21
9
6
Live TV streaming service (i.e., Hulu Live, YouTube Live)
26
27
27
12
8
Free local news/local TV stations (through TV antenna or tuner)
23
28
26
12
11
Paid with-ads video streaming service (i.e., Hulu with ads)
27
23
23
14
13
Paid television (traditional cable or satellite)
20
29
27
13
11
With-ads audio streaming service (i.e., Spotify, Pandora) / social media channels (i.e., Facebook, Instagram, TikTok)
22
27
25
14
12
Podcasts
23
26
23
16
12
Local radio (e.g., through radio or streaming service)
15
22
20
18
25


Reach effectiveness
Ineffective reach
(The user does not remember the message)
Effective reach
Price
Ineffective reach
(You are paying for users who remember the message)
Here you can bother the user
Effectiveness
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of displays

How Much Is Too Much?
61%
of Consumers are Less Likely to Buy Products When Seeing Ads Too Often
59%
of Viewers Cite that Seeing the Same Ads Repeatedly Leads to a Negative Impact on Their Viewing Experience
Agree
Disagee
Agree
Disagee
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.

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

(index score)



statista

2023 Advertising To Sales Ratios by Industry Sector

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

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

Advertising Sales Ratios - SAI Books

- Across sectors, typical ad-sales ratio is 2.83\%
- Typical public company net margin: 8-10%
- So modal firm could increase EBITDA 28-35% by dropping ads: (8+2.83)/8=1.35
- Or could it? What would happen to top-line revenue and cost efficiencies?

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

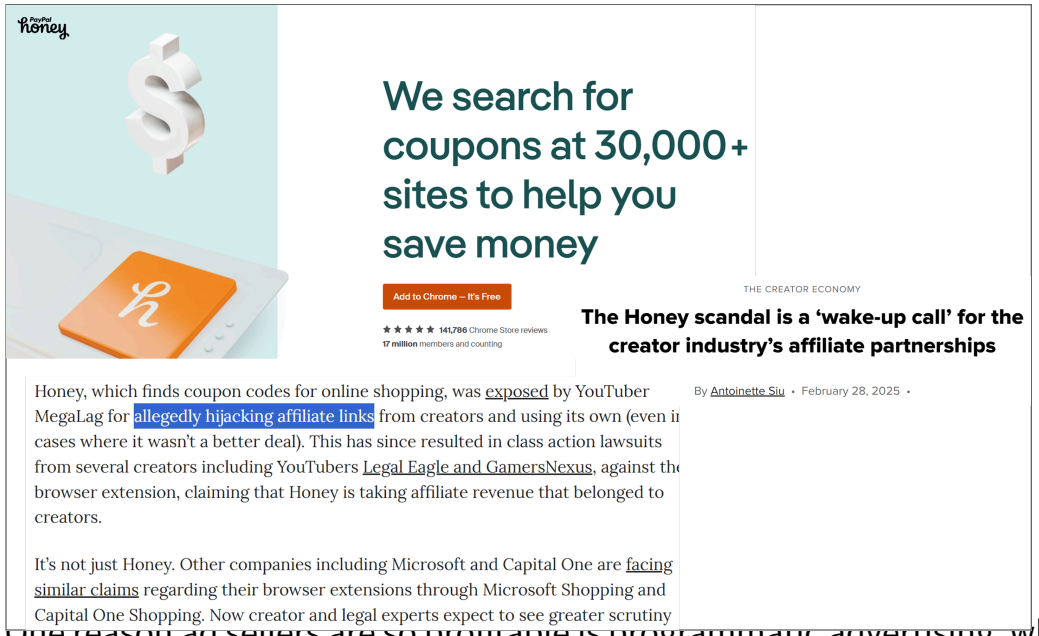
```
- Ad profit depends on ad cost, conversion rate, margin ... and how we formulate our objective function
- Exception: Search ads are atypical, you may see conversion rates like 1-5%, but incrementality questions are even more important
```

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

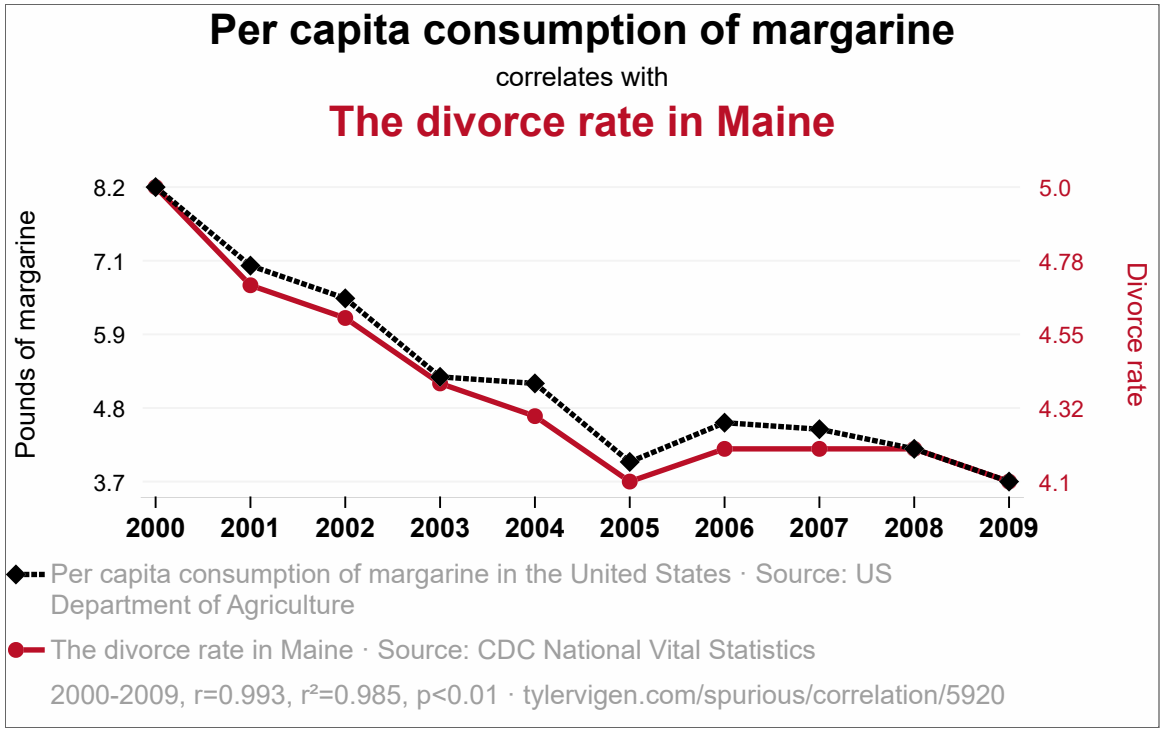
1. **Causal Inference**: Incrementality is rooted in causal inference, aiming to isolate the effect of a specific action from other factors.
2. **Control Groups**: A key methodology involves using control groups to measure what would have happened in the absence of the intervention.
3. **Lift**: Incrementality 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 tr drive results.



One reason ad sellers are so profitable is programmatic advertising, which uses technology to *automate* and *optimize* [illegible] Viewability

Causality

Examples, fallacies and motivations


Per capita consumption of margarine
correlates with
The divorce rate in Maine
Pounds of margarine
8.2
7.1
5.9
4.8
3.7
5.0
4.78
4.55
4.32
4.1
Divorce rate
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Per capita consumption of margarine in the United States · Source: US Department of Agriculture
The divorce rate in Maine · Source: CDC National Vital Statistics
2000-2009, r=0.993, r²=0.985, p<0.01 · tylervigen.com/spurious/correlation/5920

- Suppose 10 outcomes, 1000 predictors, N=100,000 obs

  - Outcomes might include visits, sales, reviews, ...
  - Predictors might include ads, customer attributes & behaviors, device/session attributes, ...
  - You calculate 10k bivariate correlation coefficients

- Suppose everything is noise, no true relationships

  - The distribution of the 10,000 correlation coefficients would be Normal, tightly centered around zero
  - A 2-sided test of {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.

Classic misleading correlations

- “Lucky socks” and sports wins

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

- Commuters carrying umbrellas and rain

```
- Forward-looking behavior
```

- Kids receiving tutoring and grades

```
- Reverse causality / selection bias
```

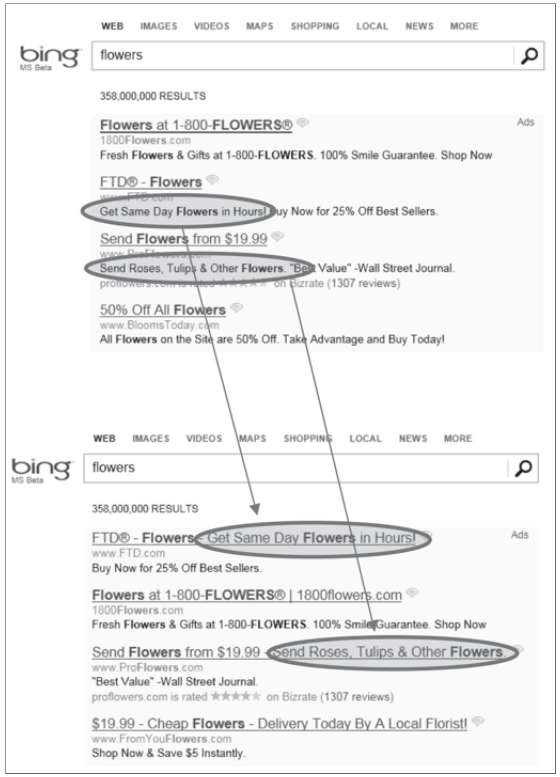
- Ice cream sales and drowning deaths

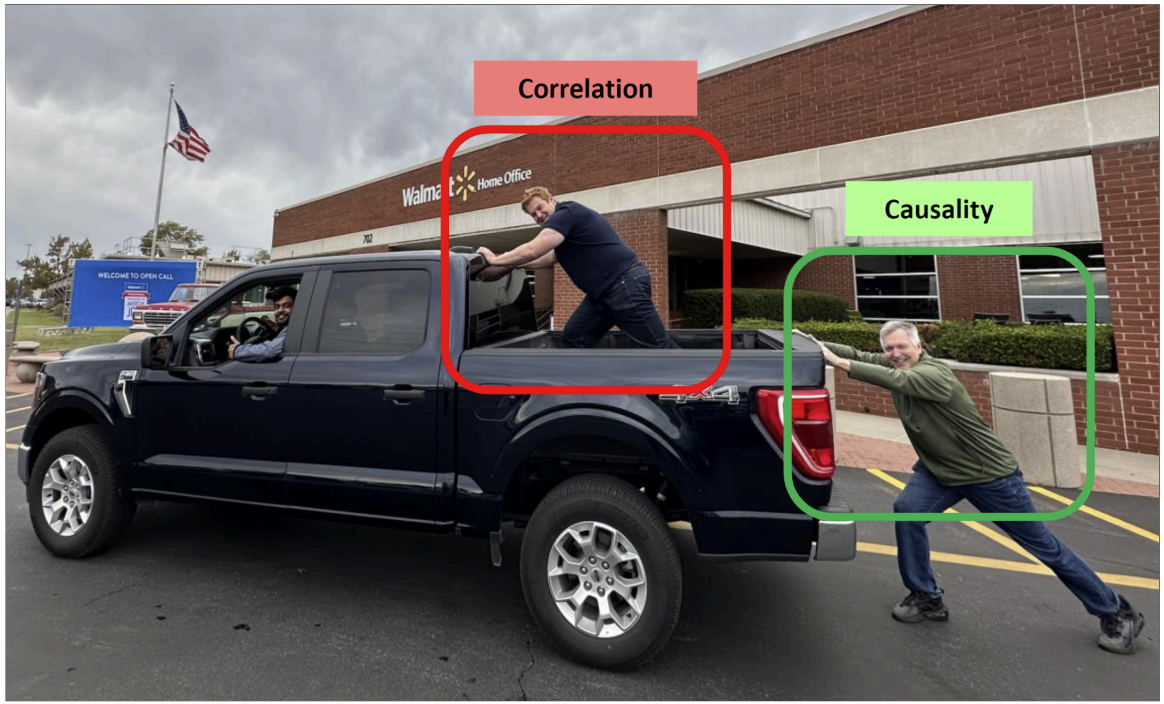
```
- Unobserved confounds
```

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

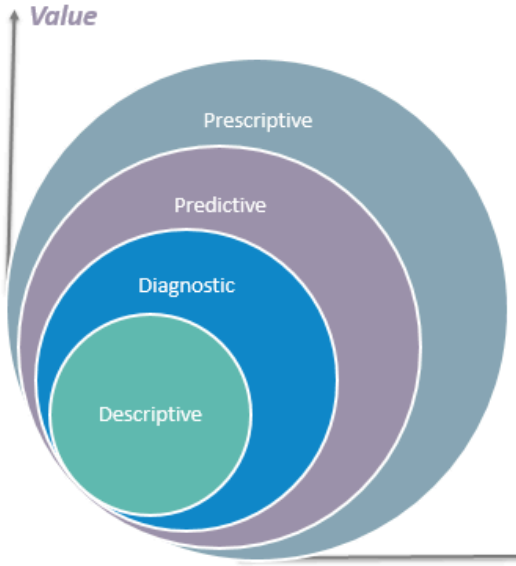
```
- Correlation-based beliefs are hard to disprove and therefore sticky
- Correlations that reinforce logical theories are especially sticky
- Correlation-based beliefs may or may not reflect causal relationships
```

“Revenue too high alert”




Correlation
Causality
Home Office
WELCOME TO OPEN CALL

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?***

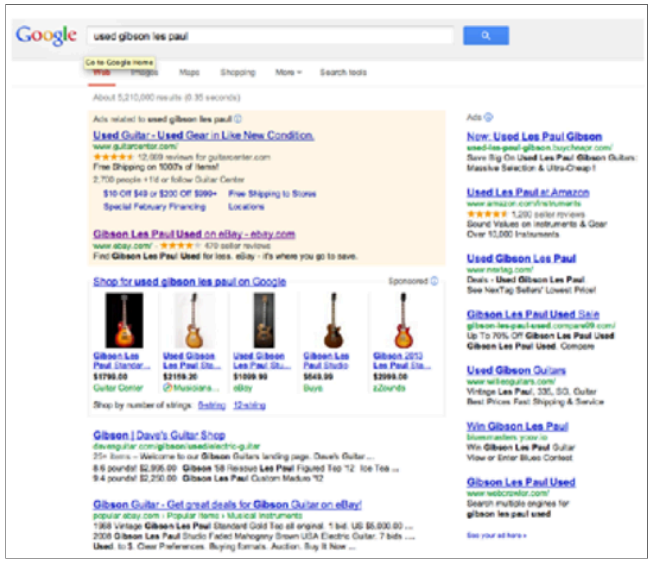
- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

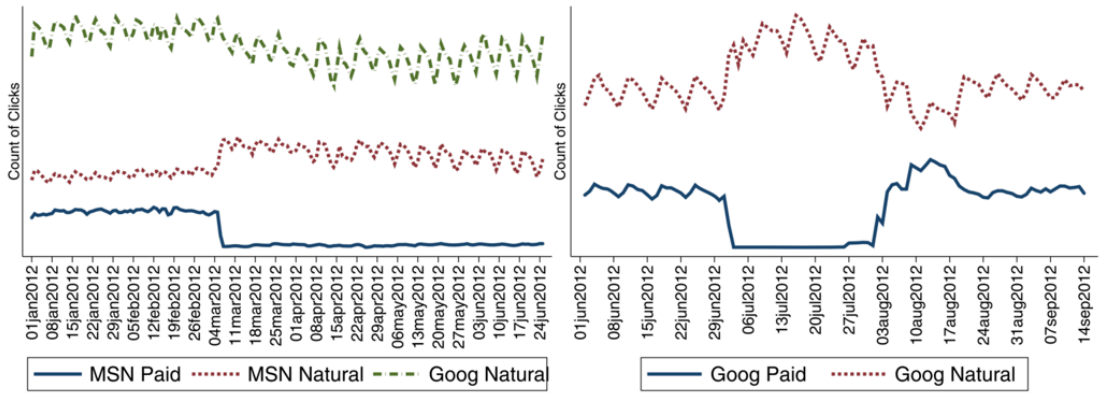
Prescriptive: ***What do I need to do?***

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

Principa
www.principa.co.za

In 2015 economists working at eBay reported a series of geo experiments testing how shutting off paid search ads affected search clicks, sales and attributed sales



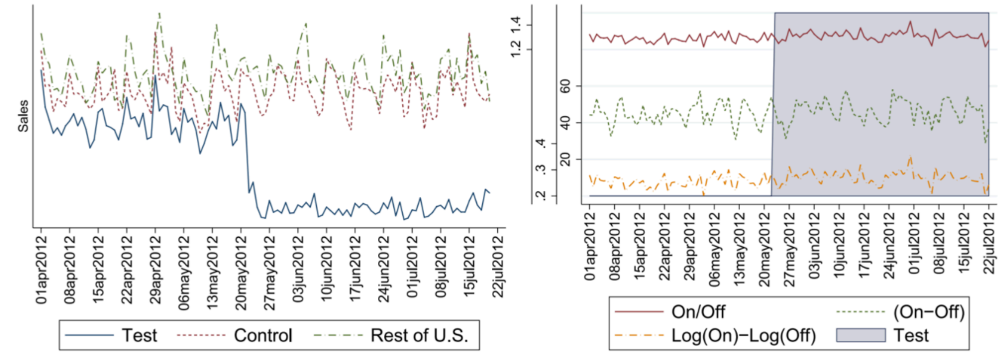


(a) MSN Test

(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.



(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. Hmmm
- These results led to changes in eBay's ad measurement
```

- This R Script simulates this scenario / Causal ML can help predictive models, but predictive correlations may suffice
[Source](#), [hbr writeup](#). [Central Control (2023)](#) offers ad geo test designs & code

Fundamental Problem of Causal Inference

Causal Inference

- Suppose we have a binary “treatment” or “policy” variable T_i that we can “assign” to person i

  - Examples: Advertise, Serve a design, Recommend a product

- Suppose person i could have a binary potential “response” or “outcome” variable $Y_i(T_i)$

  - Examples: Visit site, Click product, Add to Cart, Purchase, Rate, Review
  - Looks like the marketing funnel model we saw previously
  - "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 due to an ad

Why care?

- 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)
```

- Profits may decrease if we misallocate T_i

Fundamental Problem of Causal Inference

- We can only observe **either** $Y_i(T_i = 1)$ **or** $Y_i(T_i = 0)$, but not both, for each person i

    - The case we don't observe is called the "counterfactual"

- This is a missing-data problem that we cannot resolve. We only have one reality

    - A major reason we build models is to compensate for missing data

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 $\frac{\partial Y_i}{\partial T_i}$

    - Much easier than 1 or 2
    - But T is only randomly assigned when we run an experiment, so what exactly are we doing here?
    - Are we paying our agencies to distribute our ads randomly?

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

    - Lots of advertisers do this

How much does causality matter?

- How hard should we work?

- 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

Rubin Causal Model

Advertising measurement

Measurement of what?

- Paid media distinct from owned media & earned media

Measurement on what?

- Performance advertising: Campaigns designed to stimulate short-run measurable response. Could be any funnel stage, including awareness, information, consideration, price knowledge, visitation and sales. We should measure all funnel effects

- Brand advertising: Campaigns designed to stimulate long-run response. Measurable in multiple ways, but measurement will usually be incomplete

- *Advertising measurement* quantifies ad delivery, exposure and outcomes to improve advertising efforts
- 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, so must *directly* inform firm choices

    - We have to know how measurements will inform next steps, else measurement is wasting money

What do we measure?

Most often, we measure Return on Advertising Spend (ROAS)

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

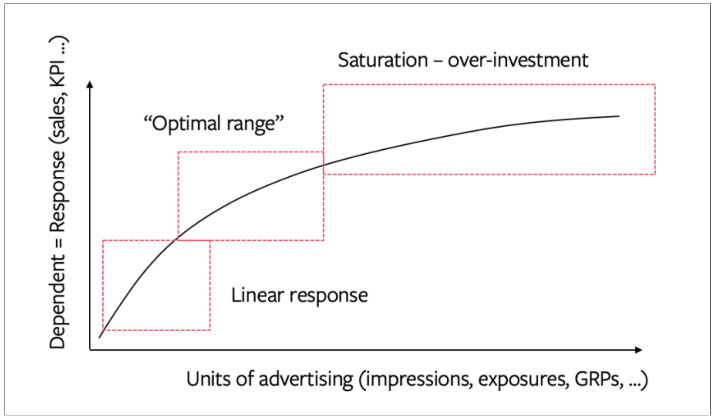
Increasingly, we report incremental ROAS (iROAS) if we have causal identification

```
- ROAS != iROAS bc attribution is usually correlational
```

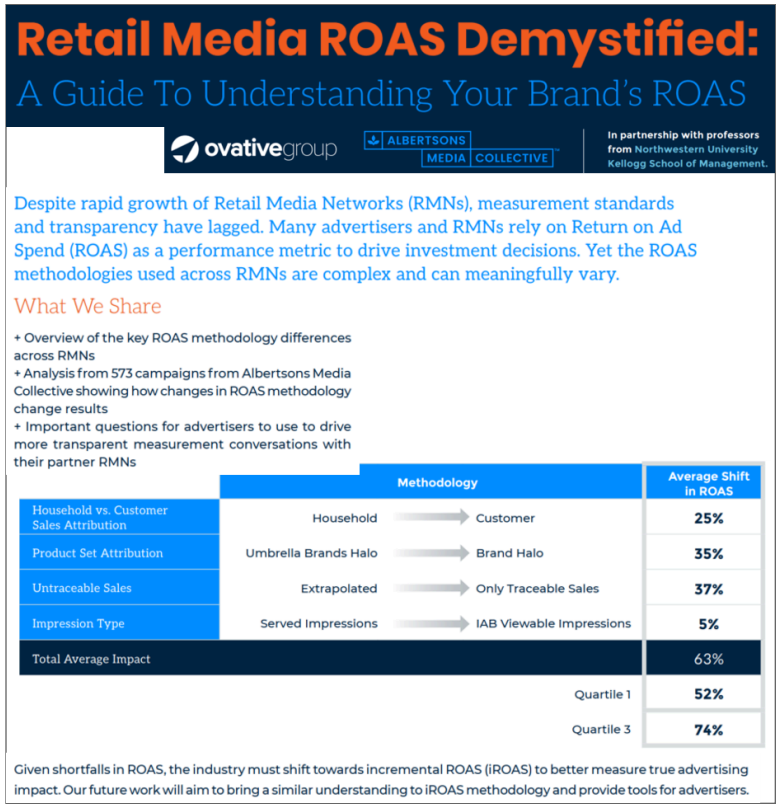
We also should measure delivery and funnel-wide KPIs, e.g. brand metrics, visits, add-to-cart, sales, revenue, …

```
- 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 that ROAS != mROAS
We use ROAS for overall evaluation, and mROAS for budget reallocation
We don't necessarily want to maximize ROAS or mROAS (why not?)
The curve above can be S-shaped, but it's hard to prove empirically
```


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.

source

Correlational advertising measurement

1. Lift Statistics

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

Usually reported as $\frac{Prob.\{Conversion|Ad\}}{Prob.\{Conversion|NoAd\}}$ or $\frac{Prob.\{Conversion|Ad\}-Prob.\{Conversion|NoAd\}}{Prob.\{Conversion|NoAd\}}$

Lift > 1 interpreted as ads are working

```
This is purely correlational, as it ignores all targeting efforts

Lift can be incremental if ads are allocated randomly
```

2. Regression

Get historical data on Y_i and T_i and run a regression

```
Could be across individuals, places, time, or combinations
Most people use OLS or MMM, but Google's CausalImpact R package is also popular
```

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

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

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

3. Multi-Touch Attribution (MTA)

Get individual-level data on every touchpoint for every purchaser

- Includes earned media (PR, reviews, organic social), owned media (website, content marketing, email) & paid media (<--ads; also, paid influencer & affiliate)
- Often sourced from third parties

Choose a rule to attribute purchases to touchpoints

- Single-touch rules: Last-touch, first-touch
- Multi-touch rules: Fractional credit, Shapley

Historically, Last-touch was popular

MTA algorithm searches for touchpoint parameters that best-fit the conversion data given the rule

- Credit then informs future budget allocations
- MTA is designed to maximize attributions
- MTA assumes advertising is the *sole* driver of conversions

MTA is mostly dead due to privacy and platform reporting changes

- Governments, platforms, browsers, OS have all restricted MTA input data for privacy
- Some advertisers' MTA lives on due to inertia, despite signal loss
- Large platforms offer MTA results within the platform

Strongest args for 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
	- Then they get 1-5 calls per 1k viewers, lasting up to ~30 minutes
	- What are some digital analogues to this?
```

However, this argument gets pushed too far

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

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 effect

```
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
```

Many, many firms basically do this

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

Problem 2 with corr(ad,sales)

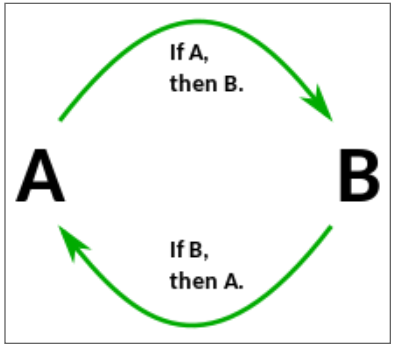
- How do most advertisers set ad budgets? Historically, the top 2 ways were:

1. Percentage of sales method, e.g. 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?



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?
- How can advertisers evaluate brand safety, targeting, context?
```

- Have ad platforms ever left ad budget unspent?

```
- Would you, if you were them?
- If not, why not? What does that imply about incrementality?
```

- To balance platform power, you have 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

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.

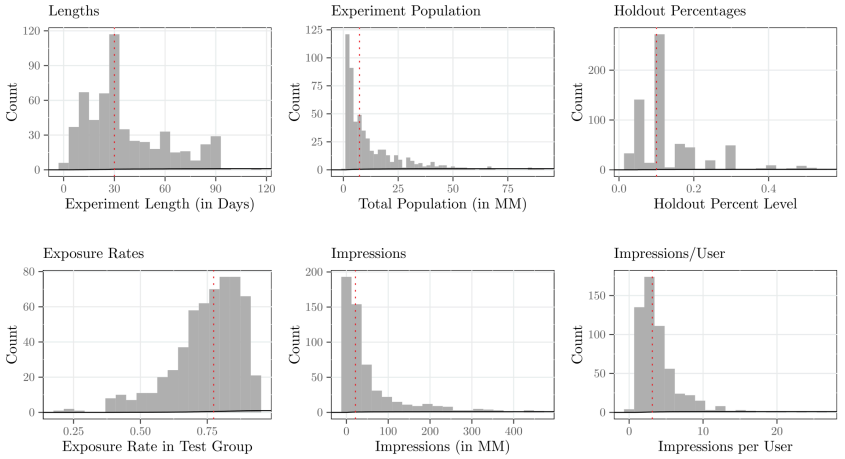
3.1. Experiment Selection

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

Figure 2. ATTs Across All RCTs

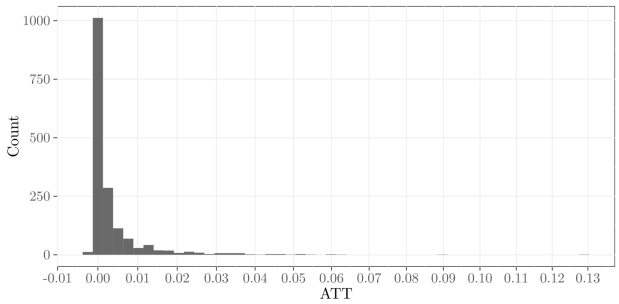
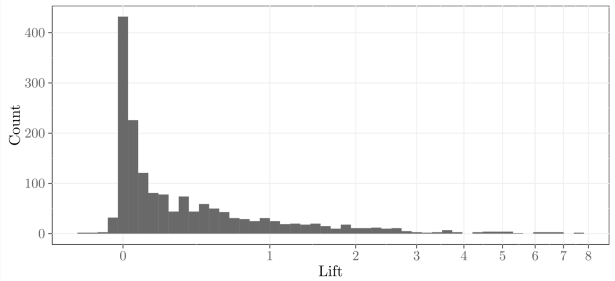


Figure 3. Lifts Across All RCTs

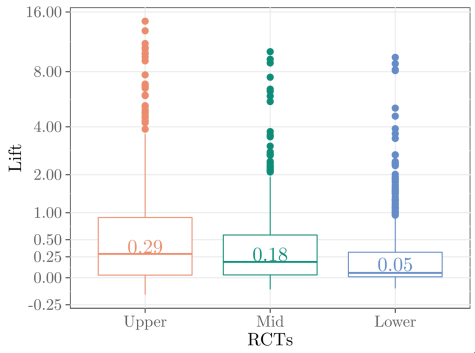


Note. Excludes the top 1% of lifts.

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



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,

$$e(X_i) \equiv \Pr(W_i = 1 \mid X_i = x). \tag{11}$$

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

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

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

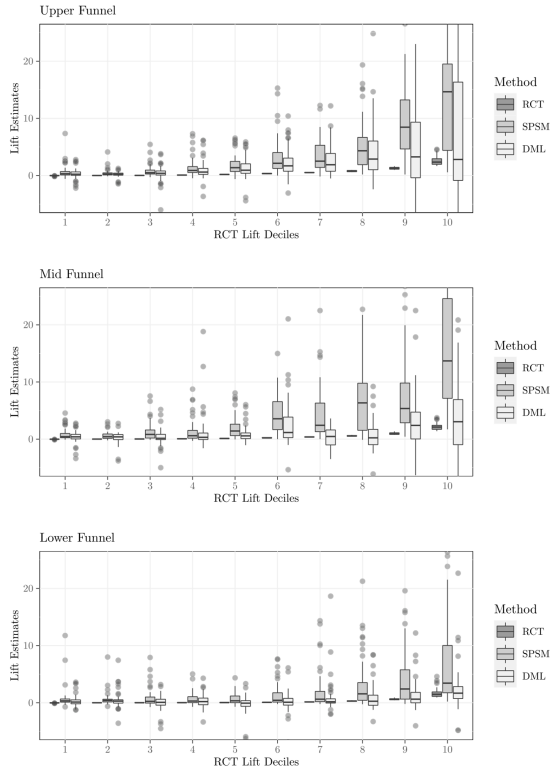
In standard propensity score matching, we find the one (or more) unexposed users with the closest propensity score to each exposed user to estimate the treatment effect. Since this is very computationally burdensome, instead, we stratify on the propensity score: After estimating the propensity score,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

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
    - Downside of lost sales may exceed downside of foregone profits
    - However, premise implies deeper problems, i.e. need to diversify marketing efforts
    - Plus, we can run experiments without setting ads to zero, e.g. weight tests

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

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

Why are some teams OK with corr(ad,sales)?

3. CFO and CMO negotiate ad budget

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

4. Few rigorous analytics cultures or ex-post checks

- In some cultures, ex-post checks may threaten bonuses, turf; may get personal

5. Estimating causal effects of ads is not always easy

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

Why are some teams OK with corr(ad,sales)?

6. 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
- Sometimes platforms respond to marketing clients' demand for good numbers
- "Nobody ever got fired for buying [famous platform brand here]"

7. Historically, agencies usually estimated RoAS

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

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

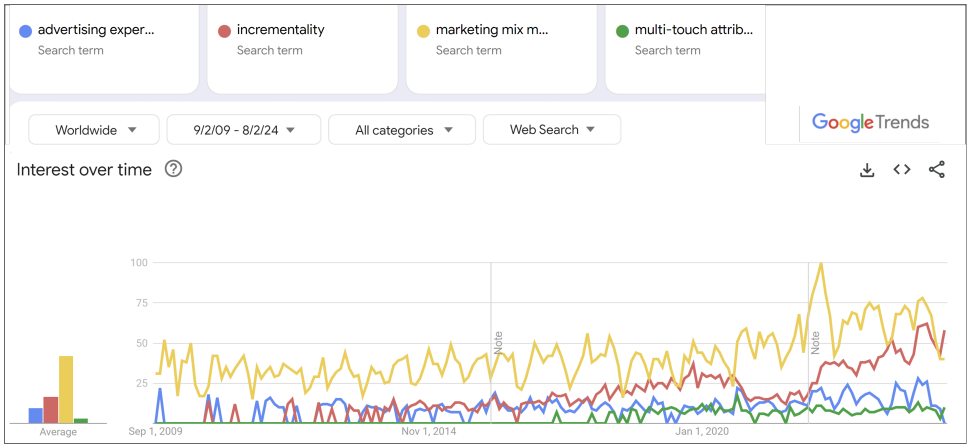
Call it vintage chic, because MMM is back.

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.



- I believe 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
- But sometimes correlational measurement is misdescribed as causal
- What happens when ad measurement is moved from CMO reporting to CFO?

Source, pg. 76-101 **Amazon Ads MTA** embeds incrementality

Causal advertising measurement

- *Causal* means that we rule out alternate explanations

Ad Experiments: Common Designs

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

    - 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

3. Randomize budget across times & places (“Geo tests”)
4. Randomize bids and/or consumer targeting criteria
5. Randomize budget over platforms, publishers, behavioral targets, contexts

    - Experimental design describes how we create data to enable treatment/control comparisons. Experimental data are amenable to any number of models or statistical analyses.
    - Causal identification is a property of the data, not the model

Experimental necessary conditions

1. Stable Unit Treatment Value Assumption (SUTVA)

- Treatments do not vary across units within a treatment group
- One unit's treatment does not change other units' potential outcomes, i.e. treatments in one group do not affect outcomes in another group
- May be violated when treated units interact on a platform
- E.g., successful ad campaign could deplete inventory, leading to periods of product nonavailability, thereby changing other treated consumers' outcomes
- Violations called "interference"; remedies usually start with cluster randomization

2. Observability

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

3. Compliance

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

4. Statistical Independence

- Random assignment of treatments to units

Muy importante

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. A:A test can fail for numerous reasons

    - 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
RoAS at 1.5 with c.i. [-1.1, 4.1]. What actions would follow each?
```

Platform experiments advisory

- Some advertising platforms offer on-platform experiments
- Most of them randomize ads across platform & consumers; not just consumers

```
- Randomization occurs before ad targeting & distribution algorithms
- You're "treating" the algorithms that determine ad distribution, not randomly selecting consumers holding everything else constant
```

- Prominent exception: **Ghost ads**, an ingenious system to randomly withhold ads from auctions in order to estimate causal ad effects

Productive experiments …

- serve customer interests

```
- Working against customers drives customers away
```

- live within theoretical frameworks

```
- Science requires hypotheses if we want to learn from tests
- Theoretical frameworks offer mechanisms and solutions
```

- 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 SR & LR metrics, e.g. trial & repurchase
- Classic example: Pop-ups seeking email subscriptions
```

- acknowledge possible interactions between variables

```
- E.g. price advertising effects will always depend on the price
```

Quasi-experiments Vocab

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

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

```
- A strategy to compare apples with apples, not apples with oranges
```

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

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

```
- This is widely, widely misunderstood
```

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

$\mathrm{avg\{Y\_i(T\_i=1)\text{-}Y\_i(T\_i=0)\}}$

Usually you want both. Models reduce uncertainty by controlling for covariates and enable counterfactual predictions

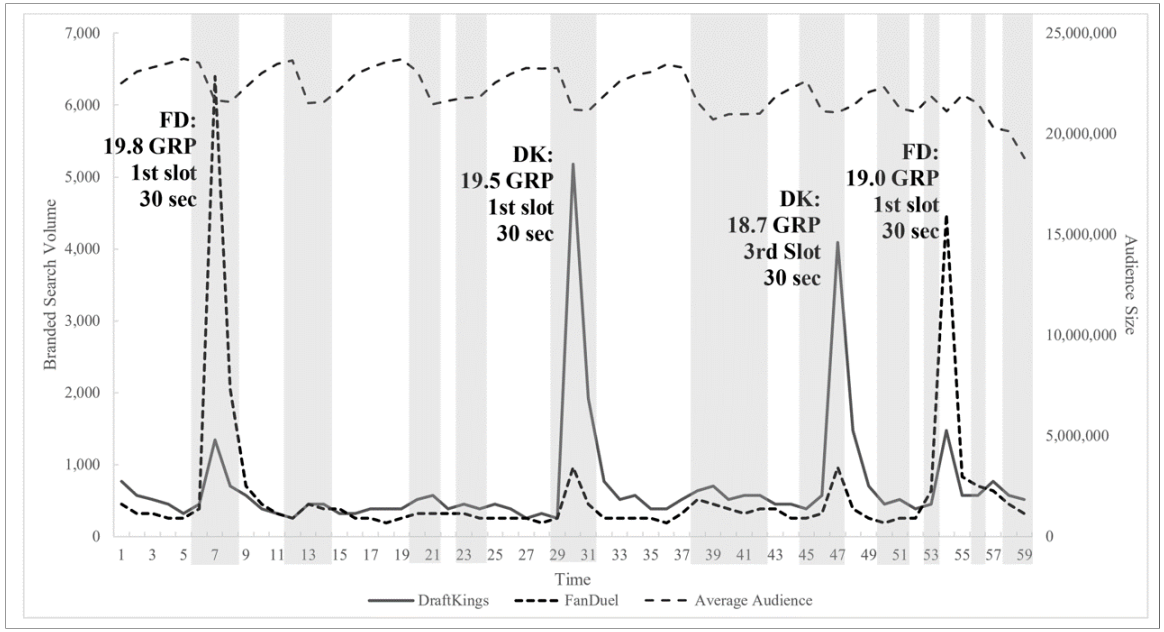
Ad/sales: Quasi-experiments

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

Possibilities:

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

DFS TV ad effects on Google Search



Ad/sales: Quasi-experiments (2)

Or, construct a “quasi-control group”

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

Helpful identification strategies: Difference in differences, Synthetic control, Regression discontinuity, Matching, Instrumental variables
In each case, we try to predict our missing counterfactual data, then estimate the causal effect as observed outcomes minus predicted outcomes
E.g., **Shapiro et al. (2021)** used the county-border approach to identify excessive TV advertising

Experiments vs. Quasi-experiments

- Experimentalists and quasi-experimentalists differ in beliefs, cultures & training

```
- not unlike bayesians v 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 difficult to impossible to validate. We can evaluate treatment & control similarity on observables, but not on unobservables
```

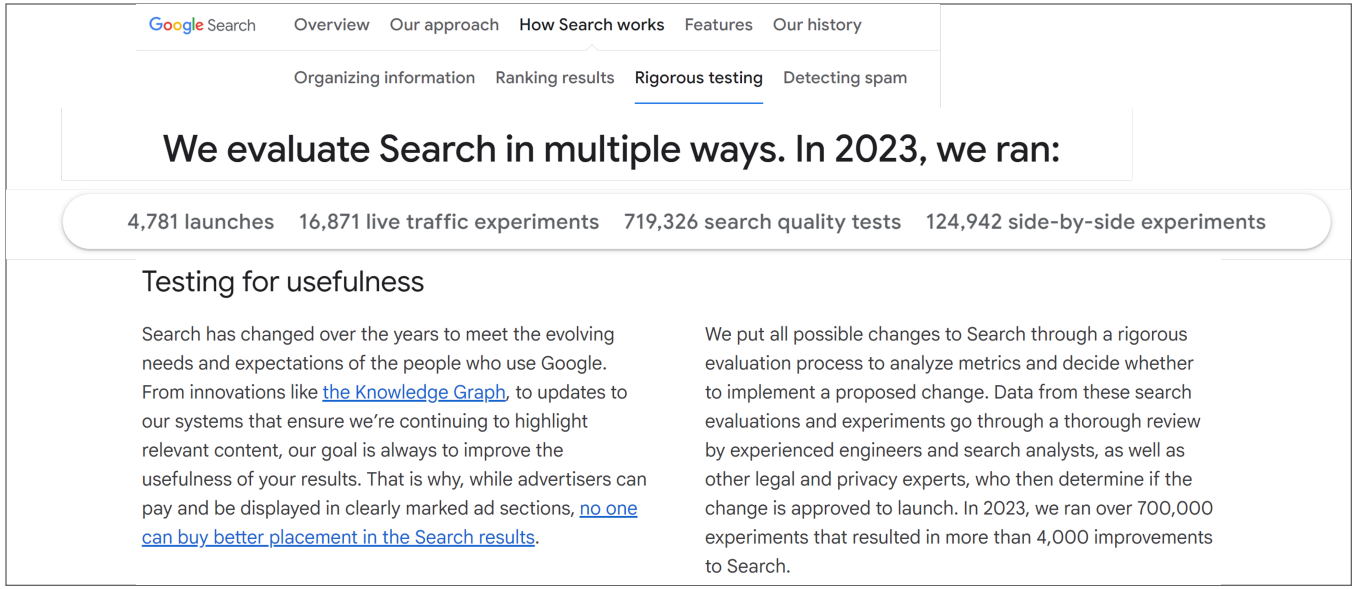
- Experiments & quasi-experiments should be “yes-if-valid,” not “either-or”

- See **List (2025)** for much more

Source: [illegible] & Schwartz (2025), [illegible]hausen et al.

Industry practices

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

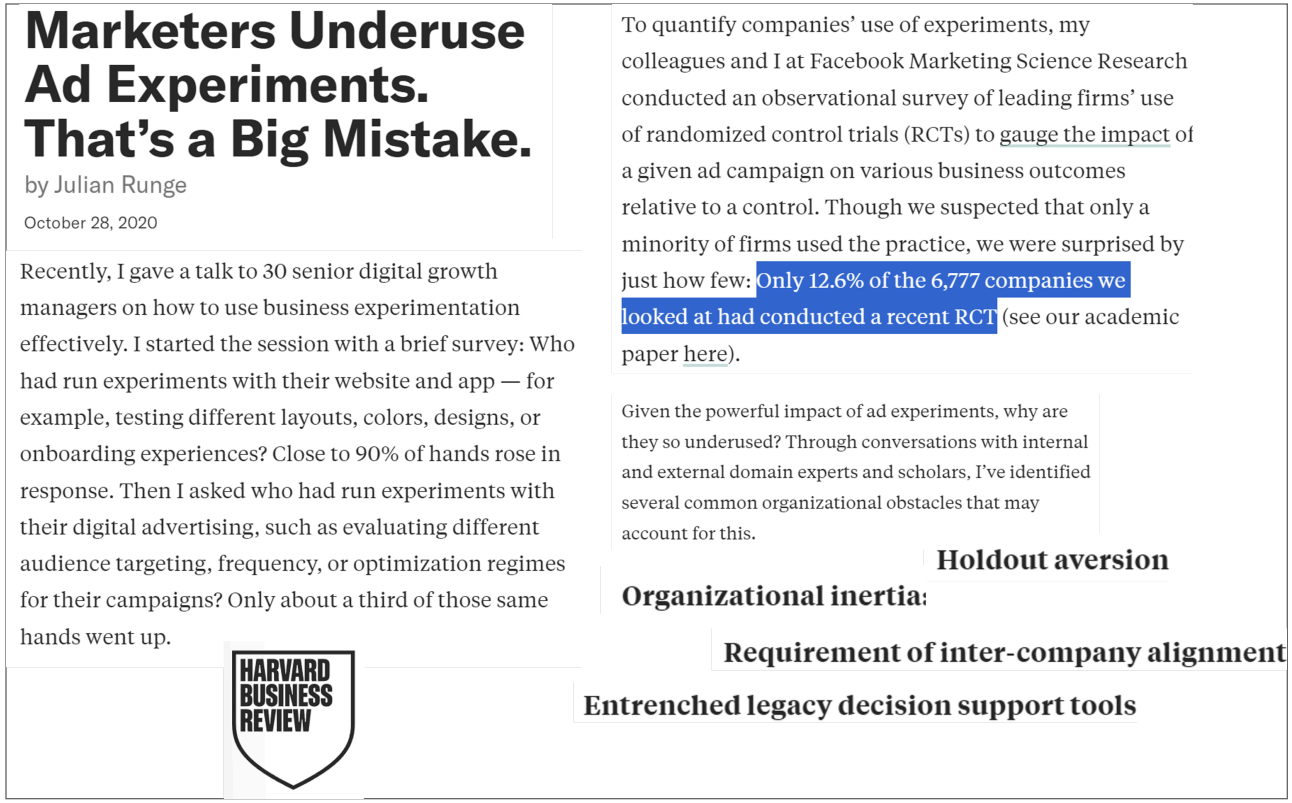
Testing for usefulness

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.

“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

Advertising experiment effectiveness

- Authors describe how companies' experimental practices relate to last-click conversion/ad$ metrics, and incremental conversions/ad$ in Meta experiments

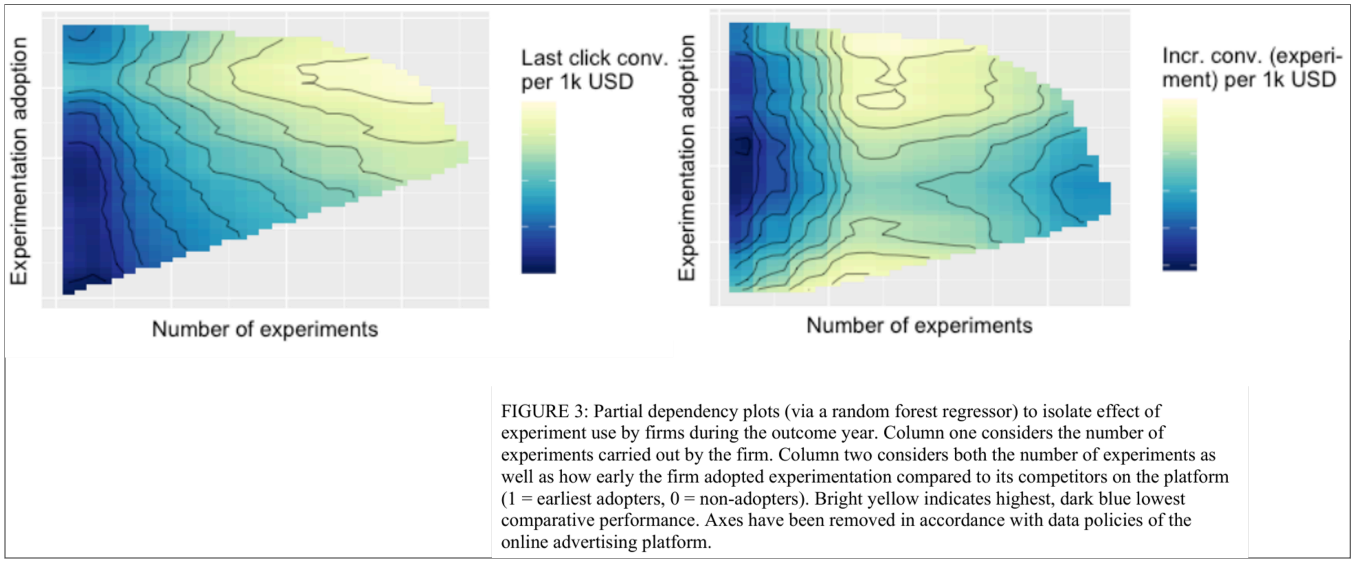


FIGURE 3: Partial dependency plots (via a random forest regressor) to isolate effect of experiment use by firms during the outcome year. Column one considers the number of experiments carried out by the firm. Column two considers both the number of experiments as well as how early the firm adopted experimentation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

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

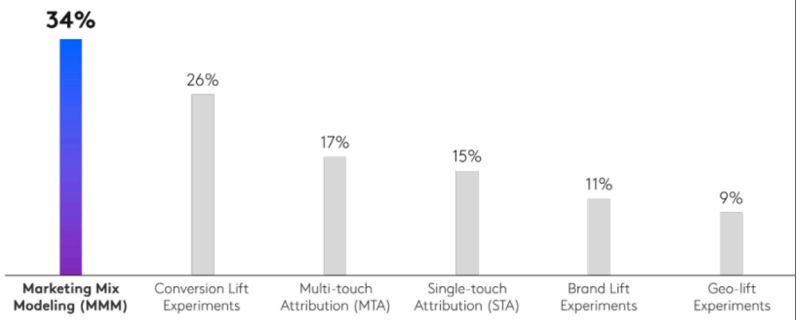
Megan DeMello

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

#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.



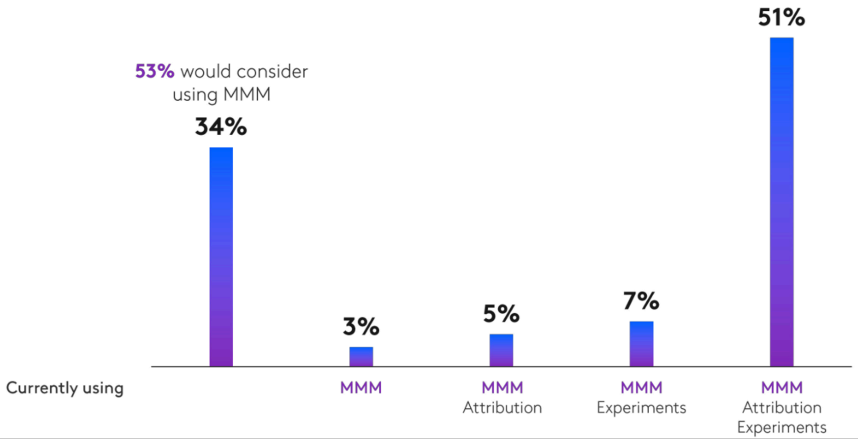
Measurement associations among total	MMM
Holistic understanding of media performance	**59%**
Data driven approach	**55%**
Used to make strategic decisions	**51%**
Worth the investment	**47%**
Backed by senior stakeholders	**46%**

#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.

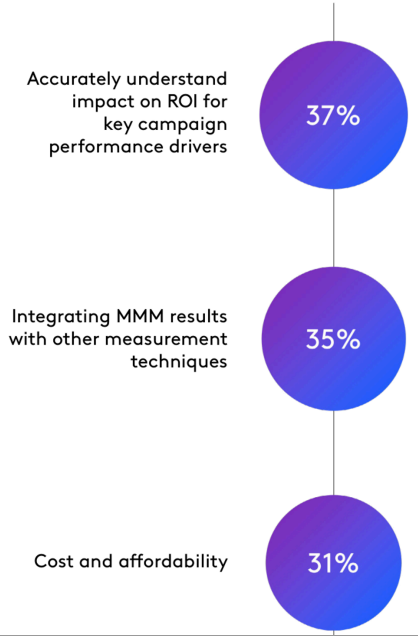
This increases to 4.2 for the largest advertisers.

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



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.

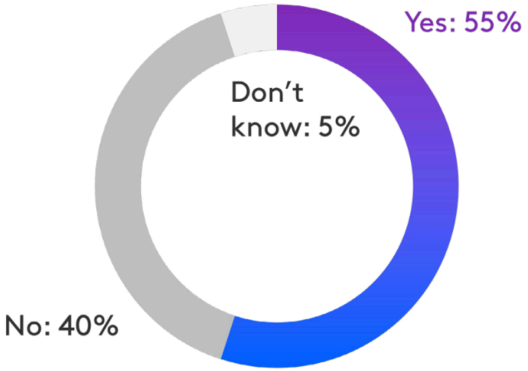
While we see multi-modal stacks as well as validation and calibration as a strong trend, over a third of respondents highlight the challenge of integrating MMM results with other measurement techniques, indicating a need for practical guidance.

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.

#4 Maximizing accuracy through validation and calibration

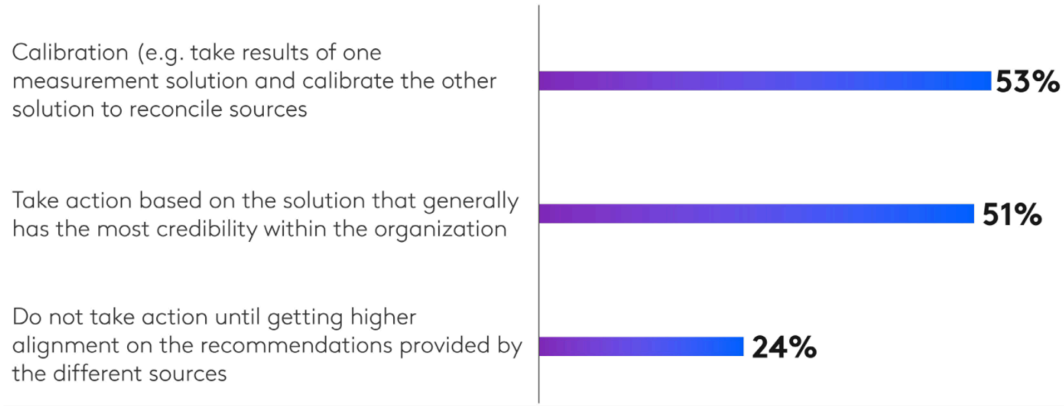
Contradicting results

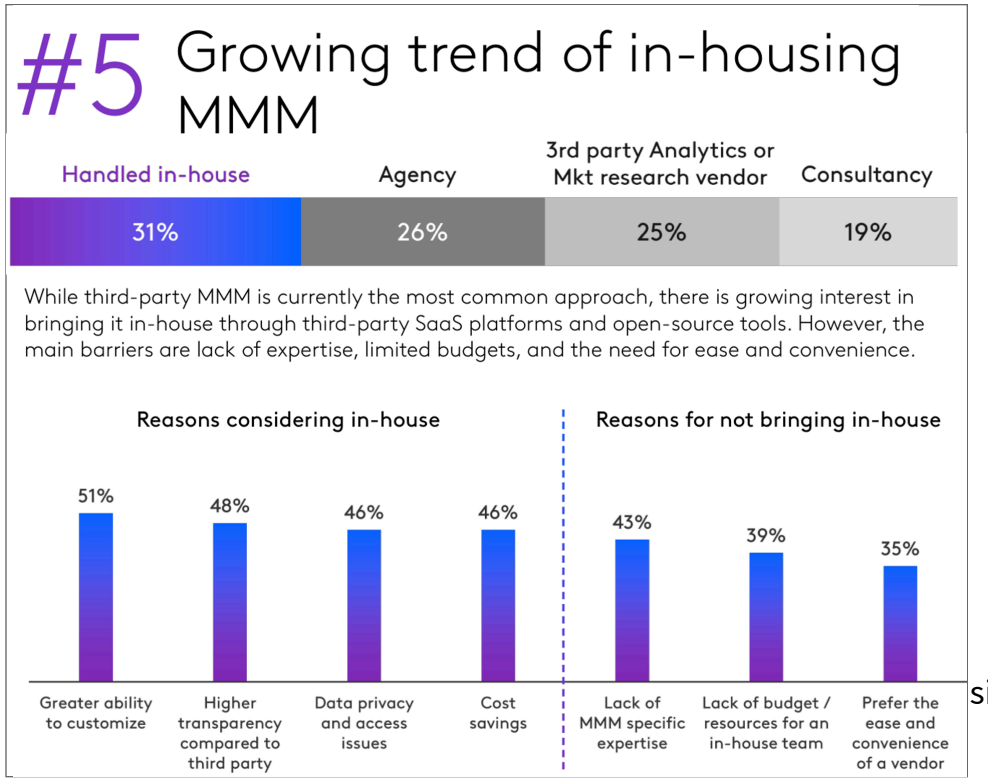
Among those who have used more than one approach



Reconciling contradicting results

Among those who have experienced contradicting results





- siness leaders?
- What does this imply for your career?

Source: Results are correlational, experimentation is not randomly assigned

Marketing Mix Models

Marketing Mix Models

- The “marketing mix” consists of the 4 P’s

  - E.g. product line, length and features; price & promotions; advertising, PR, social media and other marcoms; retail distribution breadth, intensity and quality

- A “marketing mix model” (MMM) relates sales to marketing mix variables

  - Idea goes back to the 1950s
  - E.g., suppose we increase price & ads at the same time
  - Or, suppose ads increased demand, and then inventory-based systems raised prices
  - When possible, MMM should include competitor variables also

- A “media mix model” (mMM) relates sales to ads/marcom channels

  - MMM and mMM share many models and techniques

- MMM goal is to evaluate past marketing efforts, and better inform future efforts

MMM components

MMM analyzes aggregate data, usually 3-5 years of weekly or monthly intervals, usually across a panel of geos (e.g., states, counties, CMSAs)

- 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

- Spending data is nearly always available. Exposures can be better if measured accurately

MMM usually controls for (a) trends, (b) seasonality, (c) macroeconomic factors, (d) known demand shifters, (e) saturation, (f) carryover

- Sometimes accounts for interactions between push/pull channels as well

Analysts use data to select model specifications under these constraints. Estimation can be frequentist, Bayesian, ML-based
Outputs include ad elasticities, ROAS measures, and budget reallocation
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
- 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
```

Open-source 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**, **bayesmmm**, **BayesianMMM**
Also relevant: an **MMM data simulator**

Pioneering works: **Magee (1953)**, **Weinberg (1956)**, **Vidale and Wolfe (1957)**, **Lambin (1972)**, **Little (1972)**

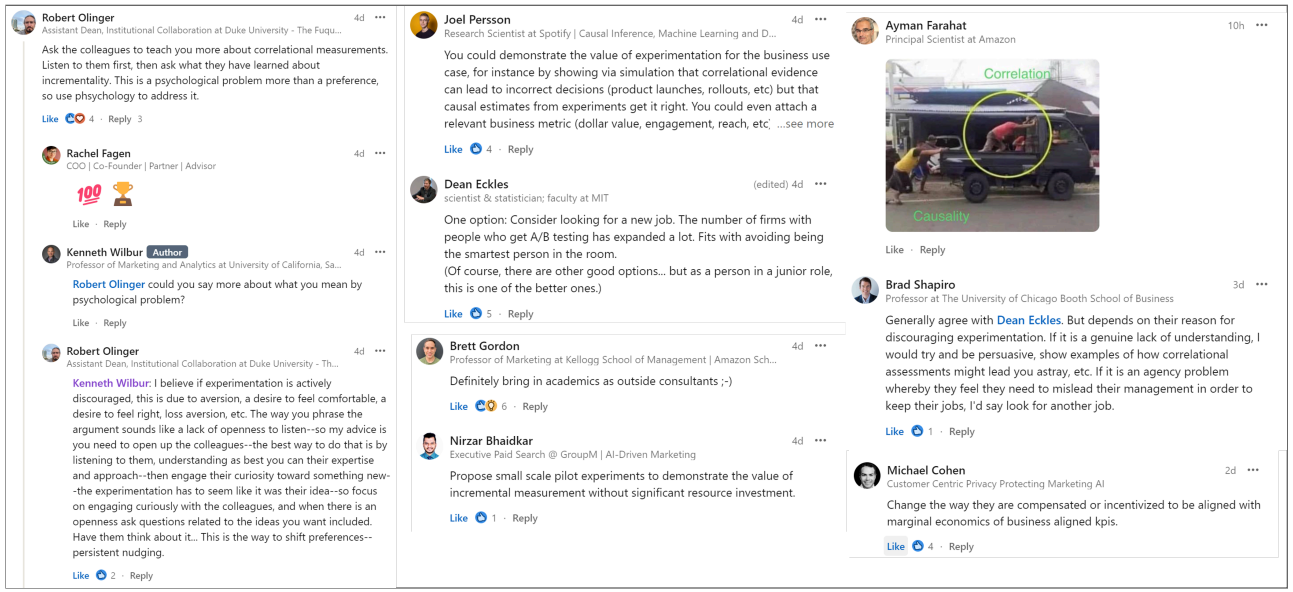
This **MSI White Paper** offers more details on **calibration** here

Career considerations

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

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

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



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

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

Like 4 · Reply 3

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

💯 🏆

Like · Reply

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

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

Like · Reply

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

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

Like 2 · Reply

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

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

Like 4 · Reply

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

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

Like 5 · Reply

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

Definitely bring in academics as outside consultants ;-)

Like 6 · Reply

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

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

Like 1 · Reply

Ayman Farahat 10h
Principal Scientist at Amazon

Like · Reply

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

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

Like 1 · Reply

Michael Cohen 2d
Customer Centric Privacy Protecting Marketing AI

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

Like 4 · Reply

Ken’s take

- Adopting incremental methods is a resume headline & interesting challenge

```
- Team may have a narrow view of experiments or how to act on them
- Understanding that view is the first step toward addressing it
- Reach up the org chart, you will need leadership onboard
```

- Correlational + Incremental > Either alone

```
- What incrementality might be valuable? What's our hardest challenge?
- What quasi-experimental measurement opportunities exist?
- Can we estimate the relationship between incremental and correlational KPIs?
```

- Going-dark design

```
- Turn off ads in (truly) random 5% of places/times; nominally free, though arguably costly if it foregoes some sales
- How do going-dark sales data compare to correlational model's predicted sales?
- Can we improve the model & motivate more informative experiments?
```

- If structural incentives misalign, consider a new role

```
- You can't reform a culture without being in the right position
- Life is short, do something meaningful
```

Takeaways

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

```
- 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

- **Retail Media ROAS Demystified: A Guide to Understanding Your Brand's ROAS** Shows how some common modeling and measurement change correlational ROAS measurements
- **Inferno: A Guide to Field Experiments in Online Display Advertising**: Covers frequent problems in online advertising experiments
- **Inefficiencies in Digital Advertising Markets**: Discusses iRoAS estimation challenges and remedies; also, principal/agent problems, adblocking and ad fraud
- **Your MMM is Broken**: Smart discussion of key MMM assumptions
- **The Power of Experiments**: Goes deep on digital test-and-learn considerations
- **New Developments in Experimental Design and Analysis (2024)** by Athey & Imbens