the Advertising Research Foundation

AUDIENCE xSCIENCE 2024

Foundations of Incrementality

Sophie MacIntyre

Meta

#ARFAxS

Foundations of Incrementality

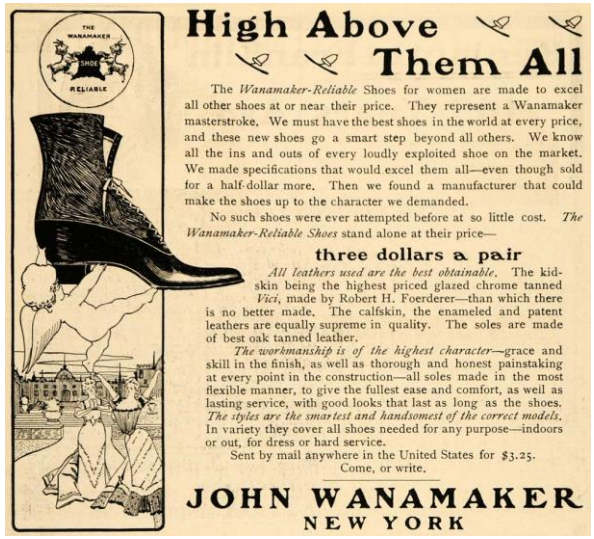
2024

Meta

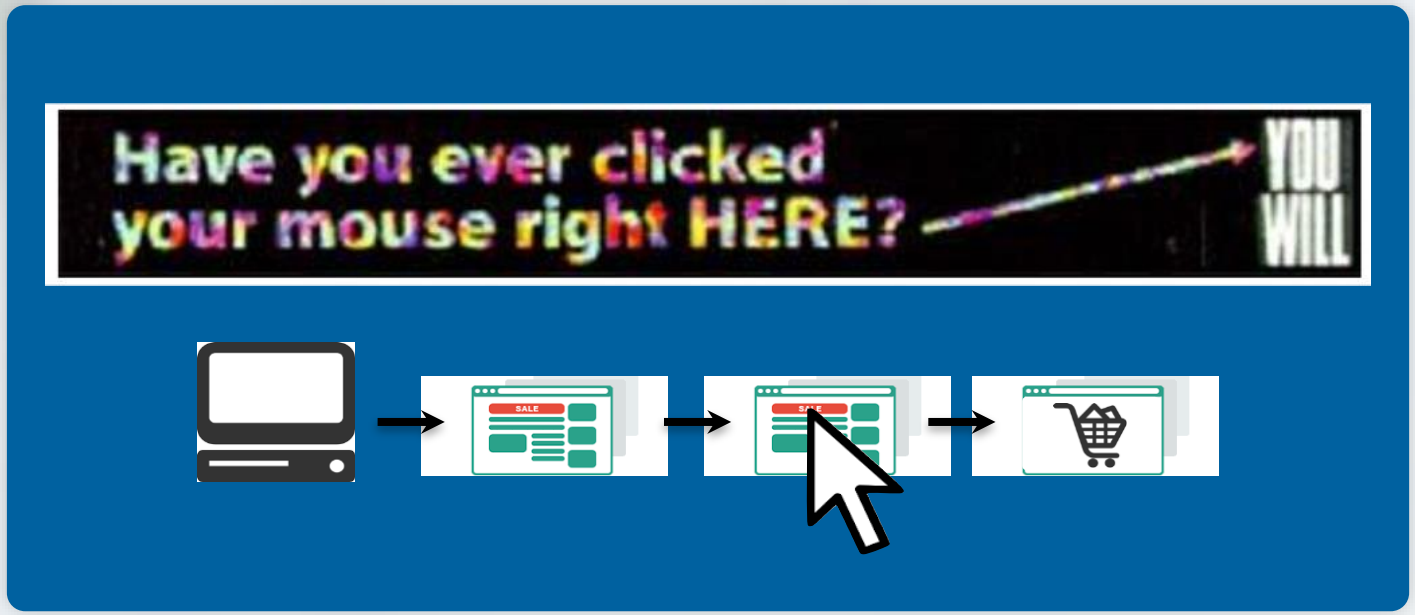
ADVERTISING EFFECTS
HAVE ALWAYS BEEN HARD TO MEASURE

“
Half the money I spend on advertising is wasted; the trouble is I don’t know which half.”

— John Wanamaker (1838 - 1922) Department Store Merchant



Digital media was supposed to make it easier…



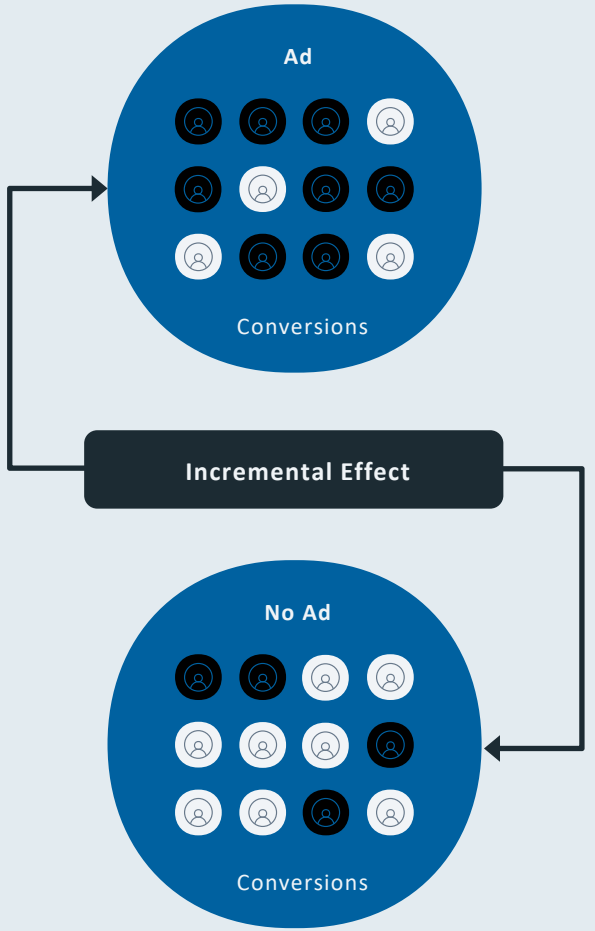
01 Why Incrementality matters

What we really want to measure…

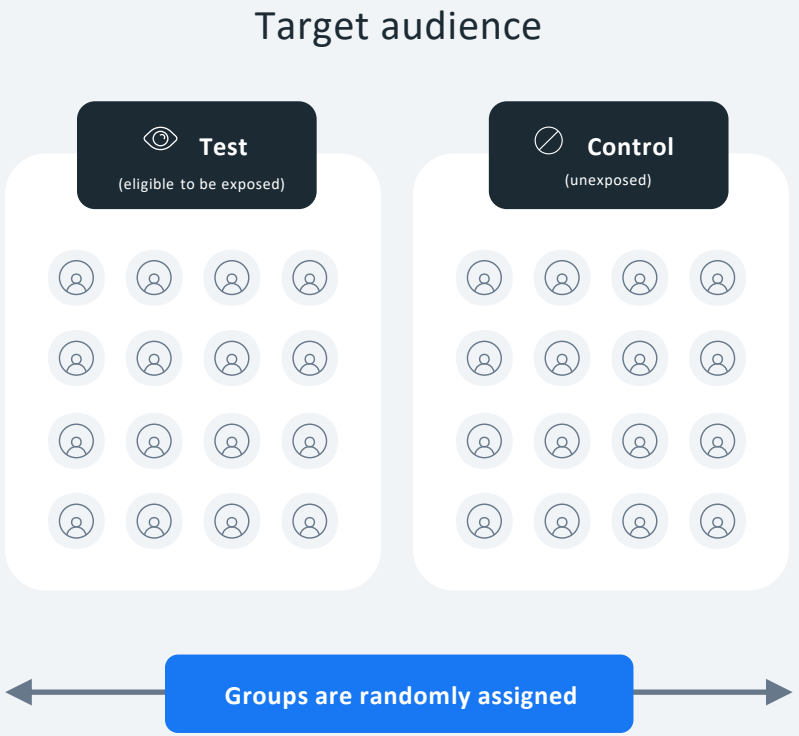
How would a consumer behave in two alternative worlds that are identical, except for one difference:

1. In one world they see an ad
2. In the other world they do not see an ad

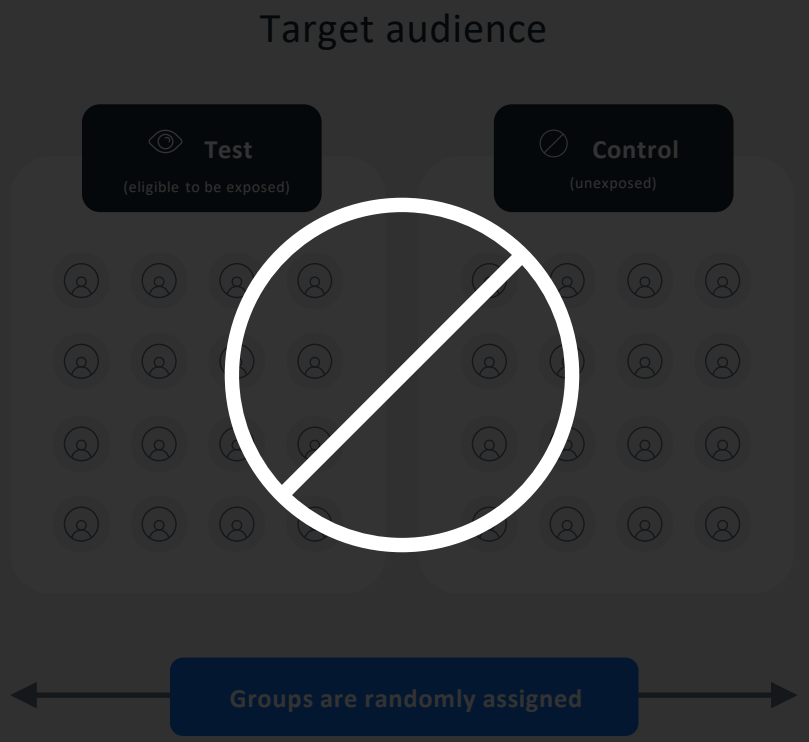
The fundamental problem in casual measurement:
No person can see and not see an ad at the same time



The “gold standard” for this is to run a Randomized Control Trial (RCT)



But what if you can’t run an RCT?



Imagine we don't have an RCT…

You could use a simple naïve approach and directly compare Exposed and Unexposed users

PROBLEM

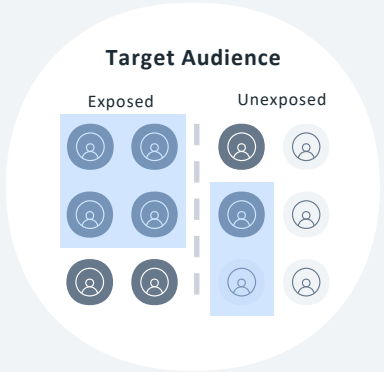
Exposed and Unexposed users aren’t “comparable” and are different for specific reasons.

Imagine we don't have an RCT…

USING "TRADITIONAL" CAUSAL INFERENCE METHODS…

Find unexposed users who look "similar" to exposed users based on observable characteristics.

The more characteristics we observe, the better!



If an advertiser had not run an RCT, how close to the RCT ad effect could they get using a non-experimental method?

1,673

RCTs from Facebook's Conversion Lift[1] platform with outcomes measured using conversion pixels

5,000+

user-level characteristics to aid model adjustment

OUTCOMES

- **601** upper funnel (e.g., viewing a web page)
- **597** mid funnel (e.g., adding a product to a cart)
- **475** lower funnel (e.g., purchasing)

Selected to be representative of RCTS run between 11/1/19 and 3/1/20 with 1M+ de-identified users in the US (~7M users)

OBSERVATIONAL METHODS COMPARED:

- **SPSM** → stratified propensity score matching
- **DML** → double/debiased machine learning

1. https://www.facebook.com/business/m/one-sheeters/conversion-lift

We use a significant number of user-level features and different observational models and compare non-experimental results to RCTS

User-level features

1. Prior campaign outcomes
2. Estimated action rates
3. Dense features
4. Sparse features

Observational models

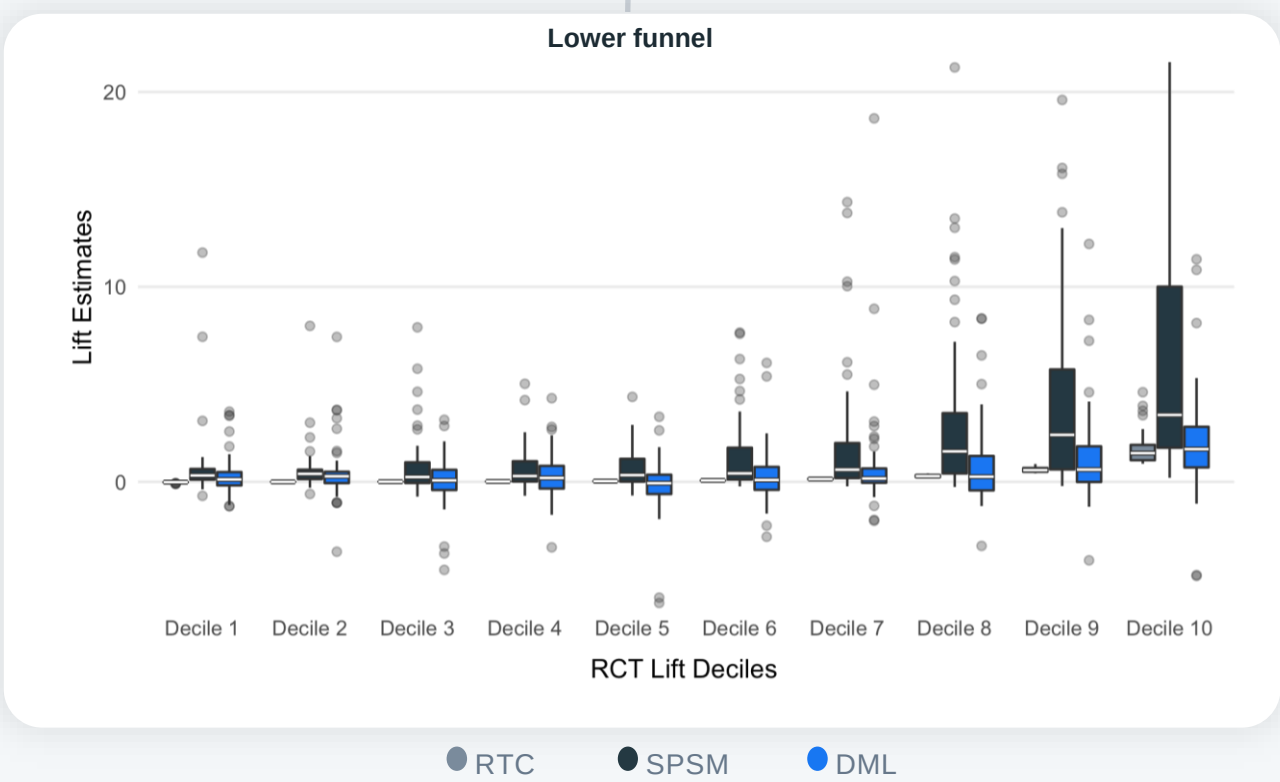
- Stratified Propensity Score Model (SPSM) (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015)
- Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018)

RCT vs. Machine Learning-based Causal Models

Compare RCT Lifts to those from SPSM (Stratified Propensity Score Matching) and DML (Double/Debiased Machine Learning)[1]

- RCT and DML estimates are statistically different (α=0.05) in 75% of experiments

FUNNEL LEVEL	RTC	SPSM	DML
Upper	29%	173%	83%
Mid	18%	176%	58%
Lower	5%	64%	24%



[1]Chernozhukov et al. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," Econometrics Journal, 21:1-68.

Conclusions

Given the data available, DML generally fails to measure the true effect of advertising accurately

To improve on this, ad platforms would probably need to log data at an extremely granular level (e.g., bid-user)

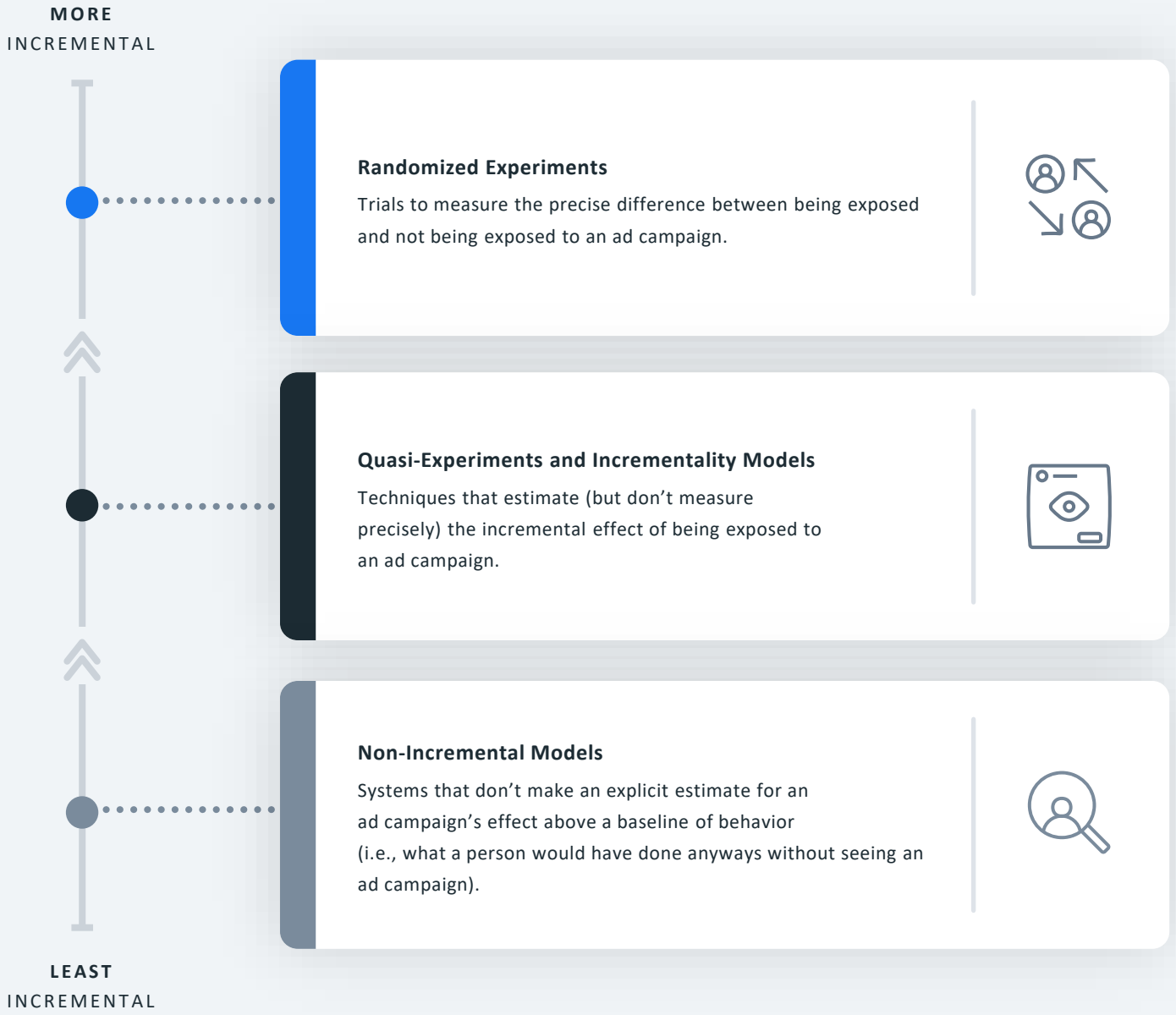
DML does relatively better for prospecting campaigns and those with low baseline conversion rates—but still not accurate

But this is costly and experimental solutions are already available

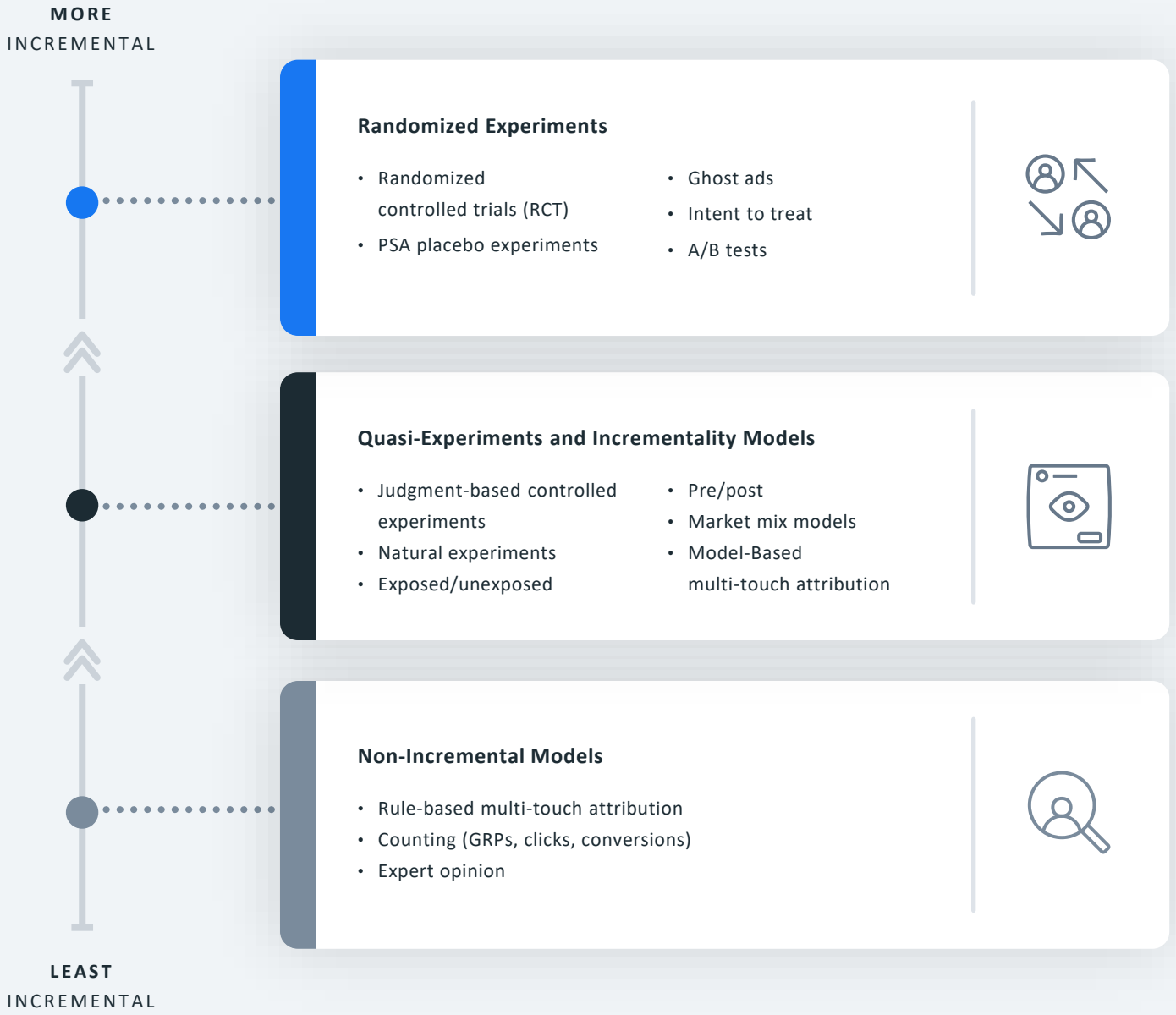
02 The Ladder & Benchmarking

THE LADDER & BENCHMARKING

You can think of incrementality as a ladder of options that get closer to measuring true business value as you climb



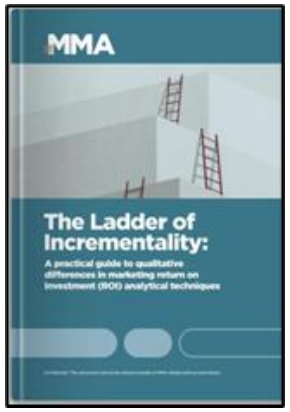
Many different techniques fall into each rung



Developed a series of 3 papers that lay out how to think about Incrementality and the challenges

The What and Why of Incrementality

- Introduces the concept of incrementality and explains why adopting this approach can improve marketing programs



The Ladder of Incrementality

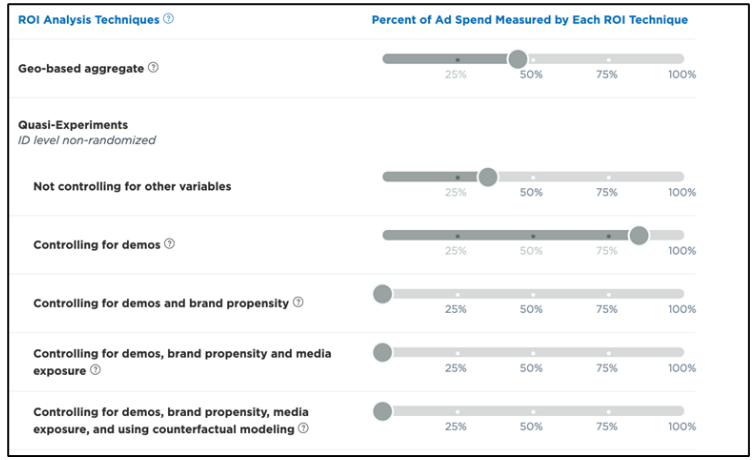
- Describes and orders different measurement techniques by how rigorous and accurate they are

Climbing the Ladder of Incrementality

- Provides actionable recommendations to improve the accuracy of measurement

. www.mmaglobal.com/incrementality

The benchmarking tool helps to assess measurement by channel



THREE INPUTS ARE REQUIRED

1. How much money did your organization spend in a year on advertising, marketing and promotions on major paid advertising channels?

2. For each channel, what proportion is measured with the various analysis techniques and methods (e.g., counting methods, rule-based MTA, MMM, pre/post, etc.)

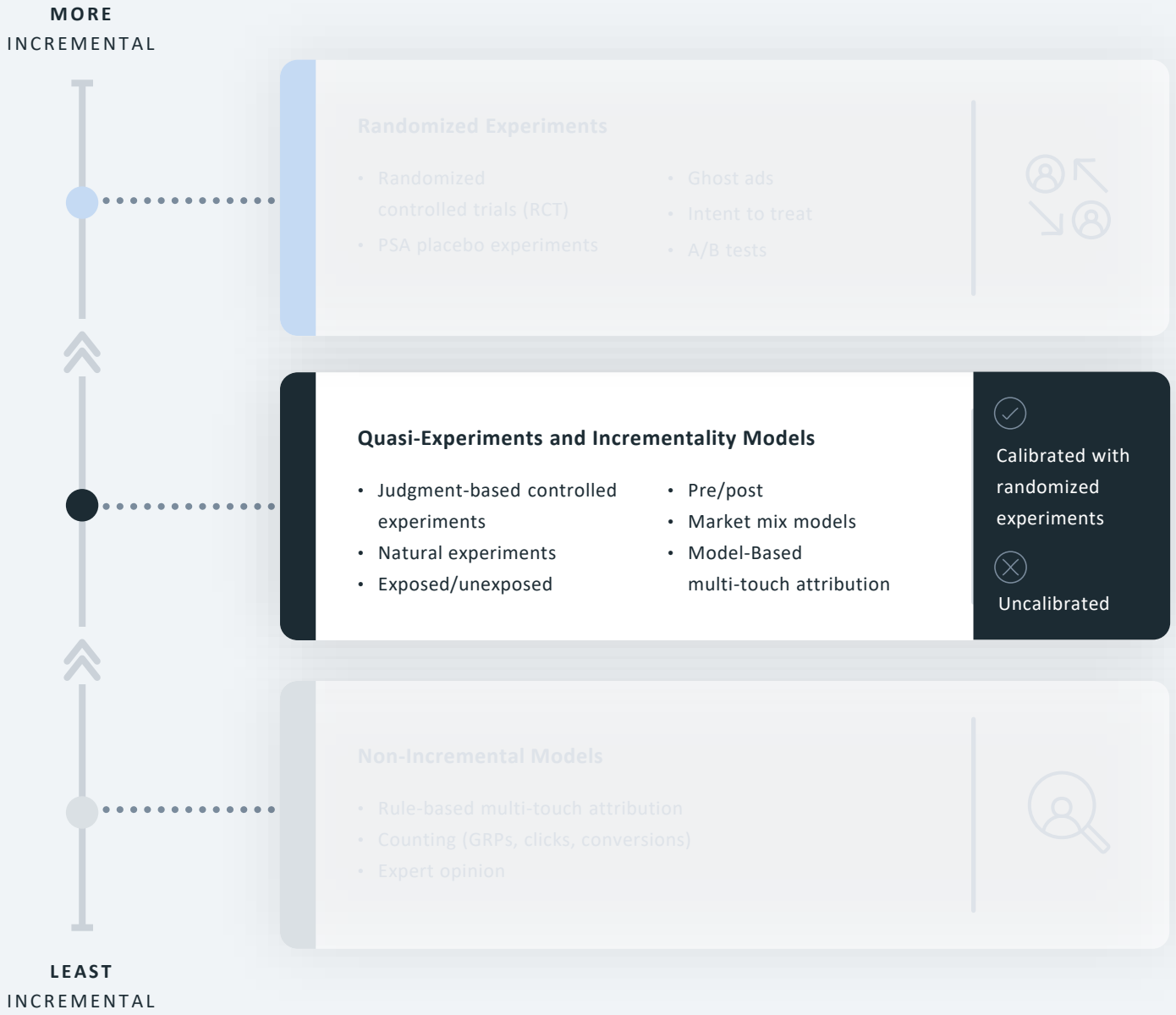
3. How thorough is your organization's process for unifying measurement results across marketing channels into actionable decisions on optimizing marketing spend?

Output:
Report card with a score for how your organization utilizes incrementality-based methods for each channel and across channels

03 Leveraging observational and experimental data together

THE LADDER OF INCREMENTALITY

Techniques in the middle of the ladder can be improved upon through calibration with experiments



By calibrating and moving up the ladder, businesses can better identify ROI

Increasingly, businesses are calibrating MTA or MMM with experiments to evaluate performance.

While not as rigorous as randomized experiments, calibration allows advertisers to advance up the ladder without abandoning the measurement they already use.

2 in 3

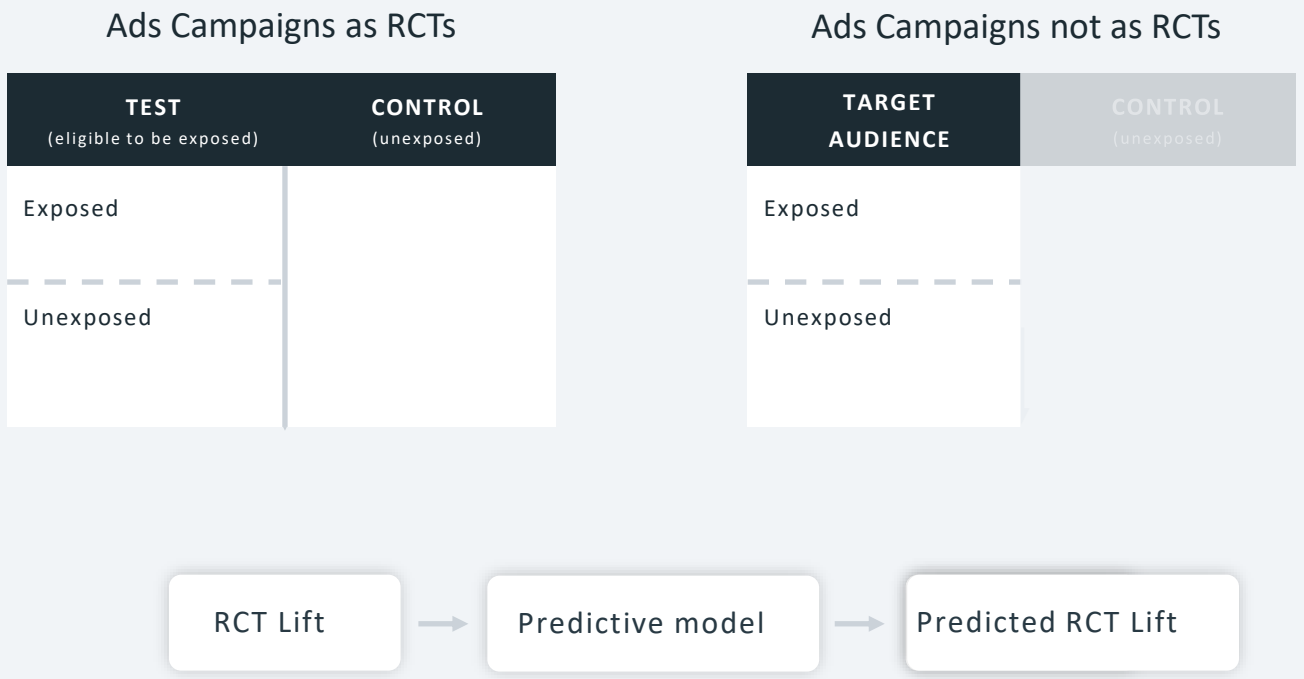
of analyzed MMM studies significantly changed Meta ROI results after calibration

25%

Average variation in ROI results after calibration

Source: “MMM calibration project” by Analytic Edge (Meta commissioned meta-analysis covering 15 MMM studies and 195 conversion lift studies across 9 regions). May 2022

But in practice, we often have RCTs for a subset of advertising campaigns…



Many ad platforms track a variety of counting and attribution metrics

For example, “Last Click” (LC) counts

Start with an outcome (purchase)

Attribution window (7 or 14 days)

“Attribute” purchase to ad that was clicked last in attribution window

Consider

- They don’t act as perfect proxies for RCT Lift
- These types of metrics are available even without an RCT

We shift to using predictive models of incrementality where the unit of observation is an RCT campaign

Question

If we had access to the RCTs in our data,
how well could we predict a new campaign's
RCT Lift that was not run as an RCT?

Approaches to Modeling

1. A simple "calibration" model
What kind of multiplier on the proxy metric would get us as close as possible to the incremental metric?

2. Expand "calibration" and control for observable campaign features
Control for additional campaign features like targeting strategy, industry vertical, prior experiment experience, etc.

RESULTS

The best PIE model yields a % WRMSE of around 70%

PIE does much better than DML

Conclusion

RCTs are considered the gold standard for unbiased measurement, however, we don't always need RCTs

Using traditional non-experimental models like propensity score matching and double machine learning is difficult and leads to large errors

However, if you have some RCTs, incrementality measurement can be achieved with modeling:

Predictive Incrementality by Experimentation (PIE)

- Even simple calibration factors show some promise
- Using PIE estimates for decision making regularly leads to similar experiment-based decisions
- More sophisticated modeling is the subject of further research


Meta