Randomized Experiments and Ad Measurements

Instructor: Davide Proserpio

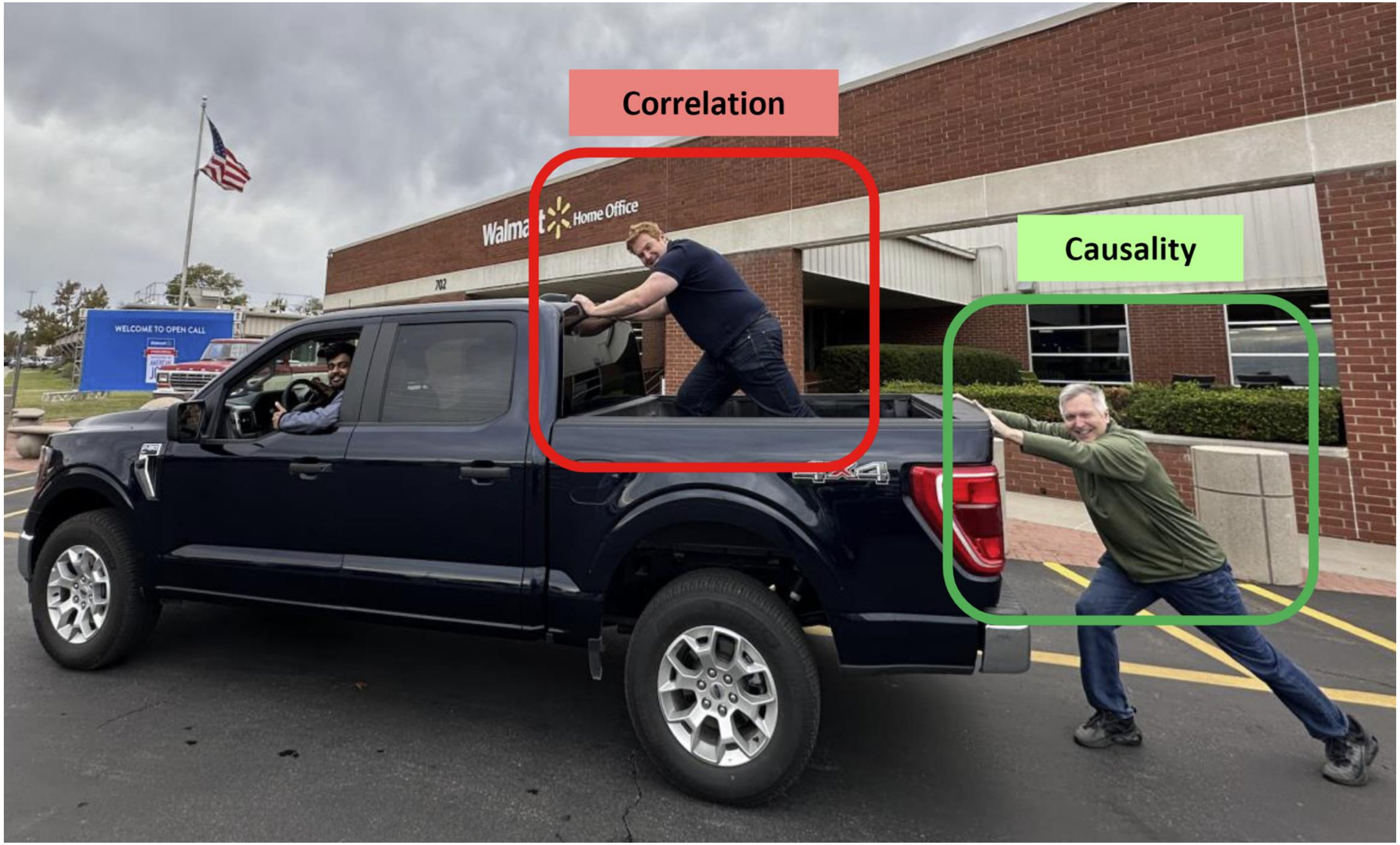
Why We Run Experiments

- Marketers constantly ask: *“Did my action **cause** a change?”*
- The problem: correlation ≠ causation.

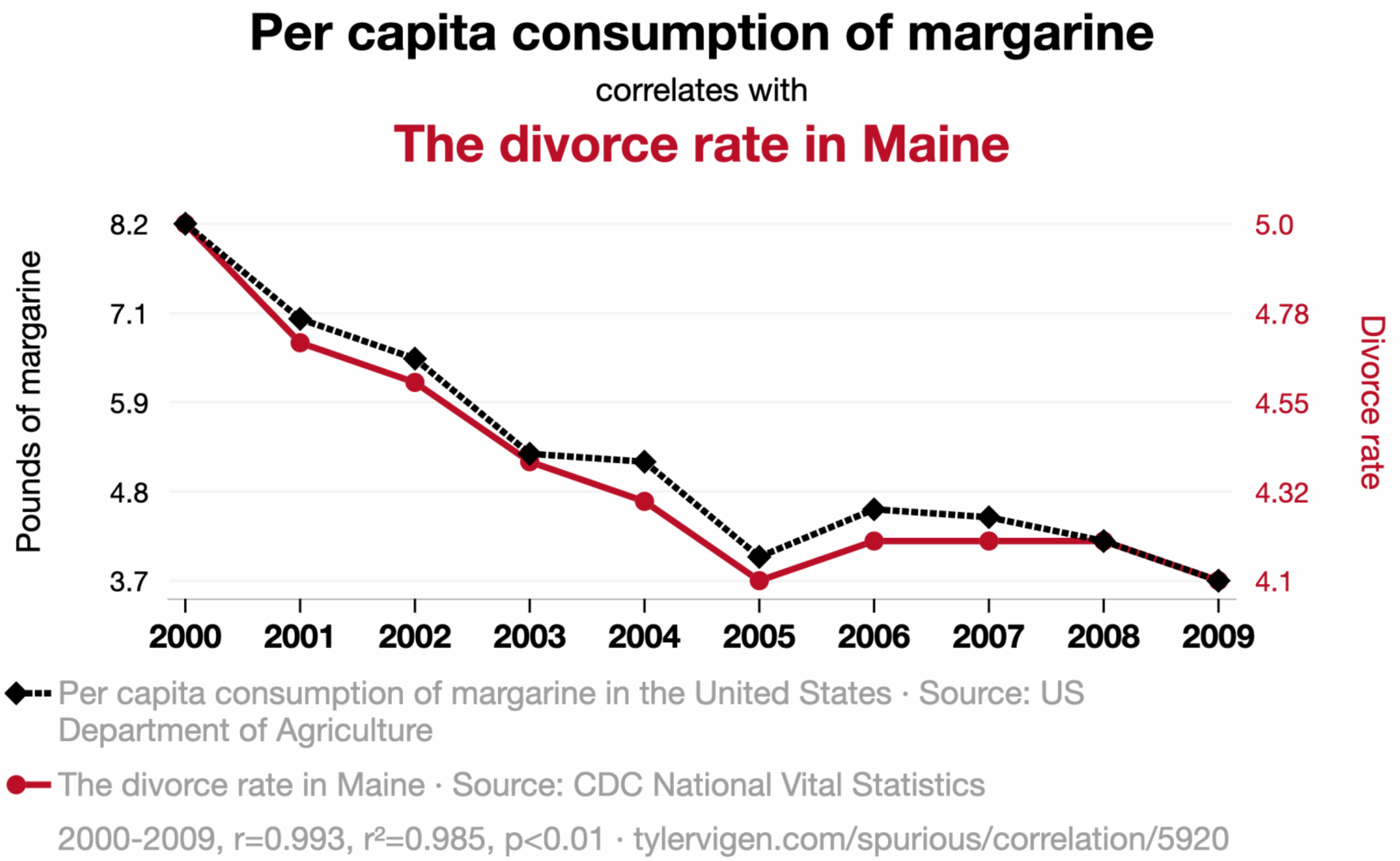
Why We Run Experiments

- Marketers constantly ask: *“Did my action* ***cause*** *a change?”*
- The problem: correlation ≠ causation.
- In general, what can we learn from a significant correlation?
  - “These two variables likely move together." Anything more requires assumptions.

Correlation ≠ causation



Correlation ≠ causation

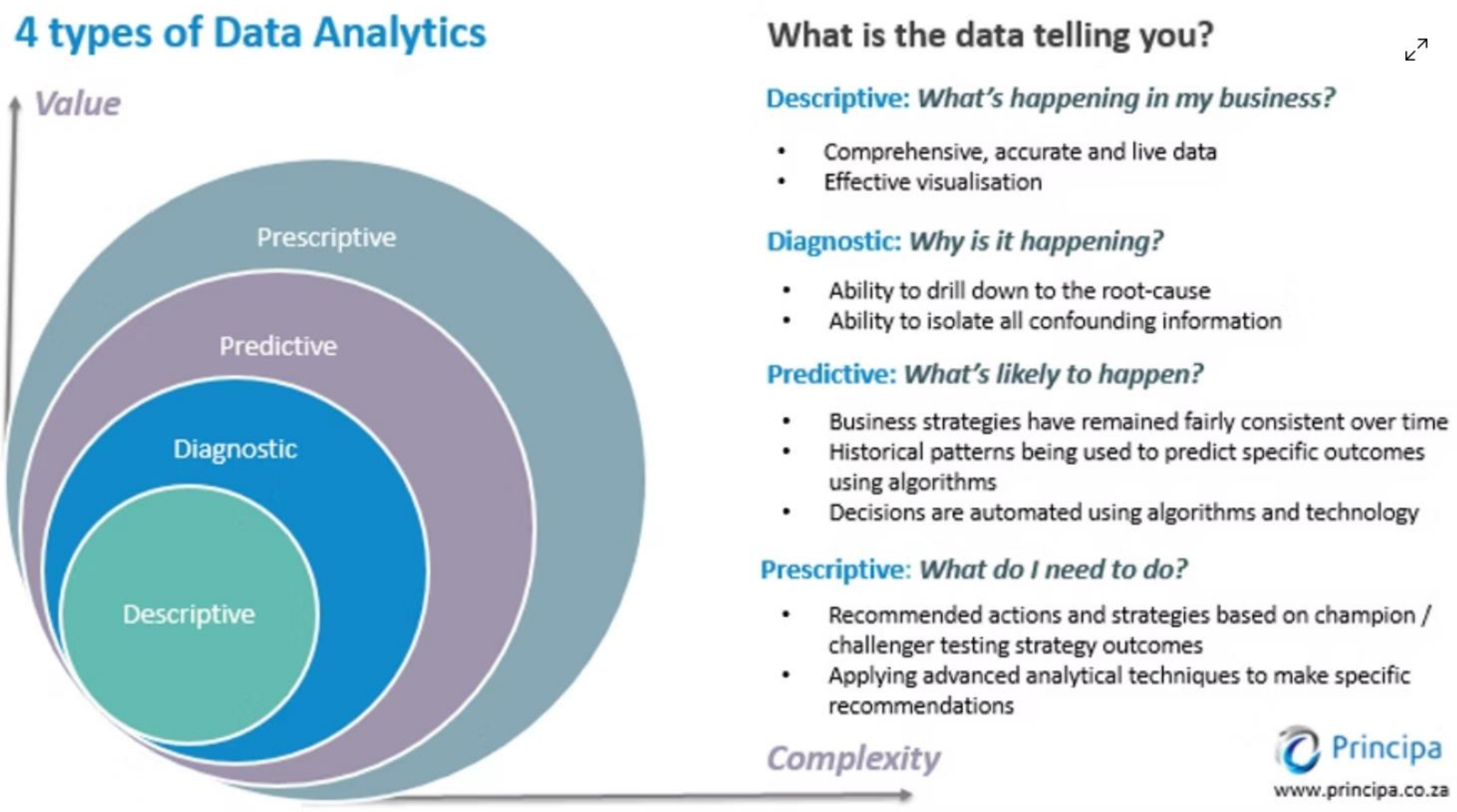


Correlation ≠ causation

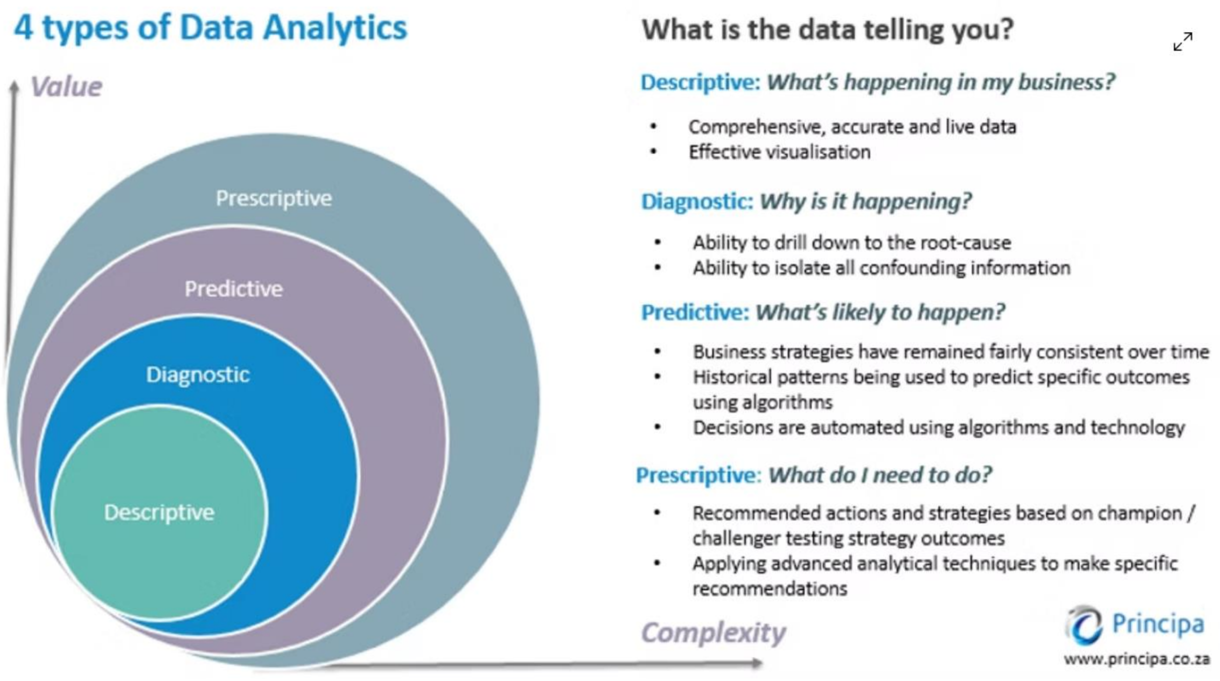
Classic misleading correlations

- 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

Why causality matters?



Why causality matters?



- Correlations are **descriptive** analytics ("facts")
- Causality matters most for **diagnostic** and **prescriptive** analytics
- Causality can help build predictive models, but correlations suffice most of the time for **predictions**

Why we need experiments: The Counterfactual Problem

- For each unit, we observe only one outcome:
  - what happened with the treatment, or without, not both.
  - 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 significant reason we build models is to compensate for missing data.
- Randomization creates comparable groups → approximates missing world.

The Logic of an Experiment

- Compare **treated** vs **control** units.
- Everything else held constant through **randomization**.
- Any average difference = **causal effect**.

Anatomy of a Randomized Experiment

Element	**Description**
Population	Users, customers, products
Treatment	Message, ad, feature, policy
Randomization	50/50 split or stratified design
Outcome	Click, purchase, satisfaction, etc.

Randomization in Practice

- Unit of randomization matters:
    - user, session, region, campaign, etc.
- Always check unit **balance**: are treated/control groups similar?
- If not, the experiment may be biased.

A/B Testing: Experiments in Product and Marketing

- **A/B test** = simplest form of an RCT.
- Test two versions (A and B) differing in *one* element.
- Measure difference in outcomes → decide which performs better.
- Used for web design, pricing, ad creatives, and recommendations.
- Substack example

Estimate treatment effect

- Observed differences can arise by chance.
- Use **hypothesis testing** to judge if effect is real.
- Report confidence intervals or p-values.
- Small samples → noisy results.

Hypothesis testing (a reminder)

- **Setup**
  - **Null hypothesis (H_0):** The ad / change has *no effect*
    $$H_0: \mu_T = \mu_C$$
  - **Alternative (H_1):** The treatment *changes* the outcome
    $$H_1: \mu_T \neq \mu_C$$
- **Logic**
  - Compute the **difference in means** between groups.
  - Estimate its **sampling uncertainty** (standard error).
  - Compare to what random chance would produce (p-value).
  - If the difference is unlikely under H_0 → **reject H_0**.

Common Mistakes in A/B Testing

- Peeking early (stopping when results look good)
- Running time too short → low statistical power
- Multiple tests → false positives
- Spillovers between users
- Focusing on statistical significance, not business value

Ads Measurements

Ads Measurements

- **Ad measurement** refers to the set of methods used to **quantify the effect of advertising exposure on desired outcomes** — such as awareness, clicks, conversions, or sales — across channels, audiences, and time.
- Advertising measurement is hard because ad effects depend on ad content, context, timing, targeting, current market conditions, past advertising & past outcomes
- 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?

- Often, Return on Ad Spend (ROAS) or Incremental ROAS (iROAS)):

  ROAS = (Revenue **attributed** to ads – Ad Spend) / Ad Spend

- ROAS != iROAS because attribution is usually correlational

Attribution vs. Incrementality

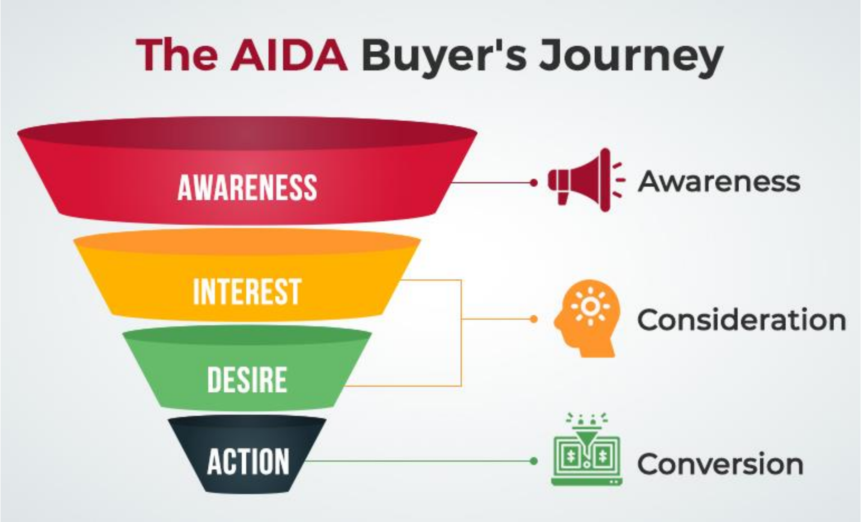
Concept	Main Question	What It Measures	Data Basis	Typical Output	Core Limitation
Attribution	*“Which channel, ad, or touchpoint gets credit for the conversion?”*	Allocation of credit among exposures	Observational data (click paths, impressions)	% contribution per channel (e.g., search 40%, display 30%, social 30%)	Correlational: cannot tell if ads *caused* the conversion
Incrementality	*“How much of the observed behavior would have happened **without** the ad?”*	True causal lift — incremental effect	Experimental or quasi-experimental data (holdout tests, ghost ads, geo experiments)	Lift %, incremental conversions, iROAS	Requires randomization or strong identification design

Heuristic Attribution Models

- **Last-touch**: credit to last channel before conversion.
- **First-touch**: credit to first interaction.
- **Linear / Time-decay**: spread credit across touchpoints.
- Easy, but **not causal** → can double-count effects.

How can we measure ads incrementality?

- RCTs can aid ad measurements
- They are very useful to measure the **incremental (causal) impact** of ads on outcomes.
- Note that it is much easier to measure outcomes for campaigns designed to stimulate **short-run** responses (e.g., sales) rather than **long-run**



Incrementality Testing in Ads

Run an **ad holdout experiment**:

- Randomly suppress ads for a control group.
- Compare outcomes → estimate incremental lift.

The “Ghost Ads” Design

- Serve ads as if everyone participated, but only some see them.
- Randomization integrated into ad auction logic.
- Ensures fair control → avoids targeting bias.
- Used by Amazon, Meta, and Google.

Geo-Split Test

- Randomize ad exposure across markets (cities/regions).
- Measure aggregate lift.
- Useful for TV, brand, or offline campaigns.
- Requires large samples and market comparability.

Some firms may be ok with Cor(Ad Spend, Sales), why?

Some firms may be ok with Cor(Ad Spend, Sales), why?

1. Some firms assume that correlations indicate the direction of causal results
   - The guy in the truck bed is pushing forward, right?
   - Biased estimates might lead to unbiased decisions (keyword: "might")
   - But direction is only part of the picture; what about effect size?
2. 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

Some firms may be ok with Cor(Ad Spend, Sales), why?

3. 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]"
4. Historically, agencies usually estimated RoAS
   - Agency compensation usually relies on spending, not incremental sales
   - Advertising attribution is all about maximizing credit to ads

Marketing Mix Models (MMM)

- **Marketing Mix Models** are **statistical models** that estimate how **sales or revenue** respond to **different marketing inputs** over time.
  - However, they often report correlation
- They help answer:
  - “How much does each channel — TV, search, display, email, etc. — contribute to performance?”
- Because correlations are still important and very much used to inform decision, MMMs continue to be very popular

Marketing Mix Models (MMM)

- Google Meridian
- Facebook Robyn

MMM: Core idea

- Regress **sales** on **spend and other drivers**:

$$\text{Sales}_t = \beta_0 + \beta_1 TV_t + \beta_2 Search_t + \beta_3 Display_t + \cdots + \varepsilon_t$$

- The estimated β's capture the **average (correlational) effect** of each channel.
- **Typical Inputs**
  - Weekly or monthly data
  - Media spend (TV, search, social, radio, etc.)
  - Control variables (seasonality, price, promotions, holidays)
  - Sometimes lagged or decayed ad effects (adstock): the carryover effect of advertising how the *impact* of an ad persists over time even after spending stops.

What It’s Used For

- Estimate **ROI per channel**
- Support **budget allocation** decisions
- Complement experiments when user-level data are unavailable
- Capture **long-term brand effects**

MMM: Limitations

- Correlation bias — no randomization
- Requires strong modeling assumptions
- Sensitive to multicollinearity among channels
- Slower feedback — often quarterly or annual

Practical Takeaways

- Randomization = best path to causality.
- Attribution ≠ incrementality — don’t confuse credit with causation.
- Experiments should inform budget allocation, not replace it.
- Cor(Ad Spend, Sales) are still pretty popular and so are Marketing Mix Models