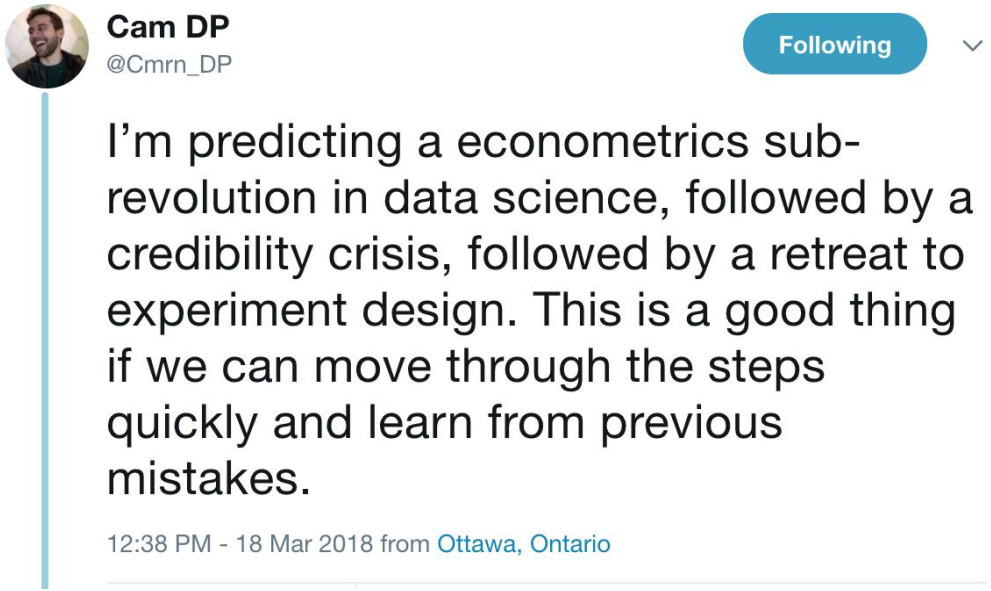
WHAT'S THE SCIENCE IN DATA SCIENCE?

PyData Meetup
Ann Arbor, Michigan
July 19, 2018

Skipper Seabold, Director of Data Science R&D, Product Lead
Civis Analytics
@jseabold

All great ideas come while you're focused on something else. All motivation to act comes from Twitter.

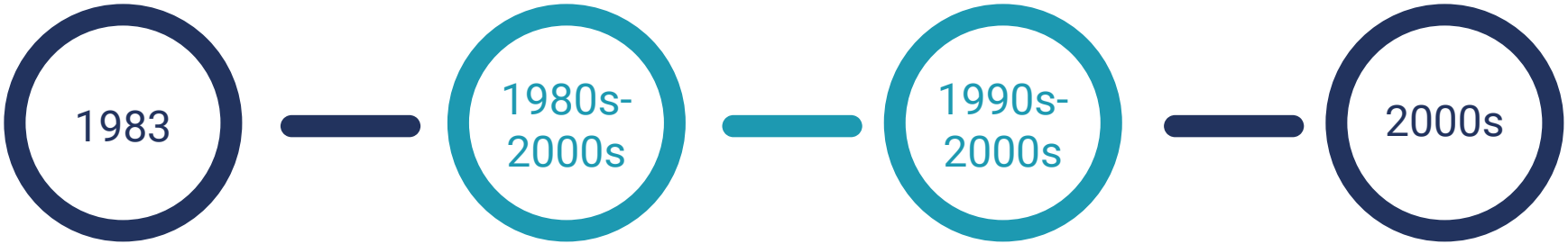


Four decades of econom(etr)ic history

Four decades of econom(etr)ic history

In 120 seconds.

A focus on research design takes the con out of econometrics



"Let's Take the Con out of Econometrics"

Criticism of the current state of econometric research. Calls for focus on threats to validity (sensitivity analysis) of non-randomized control trials.

Quasi-Experimental Design

Lends credibility, especially to micro, through "a clear-eyed focus on research design … [such as] it would command in a real experiment."

Randomized Control Trials

In parallel, increasing use of randomized control trials in economics, especially in development micro.

Worries about External Validity

"Counterrevolutionary" worries that the focus on experimental and quasi-experimental research design comes at the expense of external validity.

These changes lead to an increased relevance in policy decisions

"Good designs have a beneficial side effect: they typically lend themselves to a simple explanation of empirical methods and a straightforward presentations of results."

What does this have to do with data science?

First, some context.

Data science exists to drive better business outcomes

Obtain, Scrub, Explore, Model, and iNterpret (OSEMN)

Mason and Wiggins, 2010

The “ability to [create] prototype-level versions of ... the steps needed to derive new insights or build data products”

Analyzing the Analyzers, 2013

Using multidisciplinary methods to understand and have a measurable impact on a business process or product

Me, Today

Multidisciplinary teams use the (Data) Scientific Method to measure impacts

Question	Start with a **product or business question**. E.g., how do our marketing campaigns perform? What's driving employee attrition?
Hypothesize	Research, if necessary, and write a **falsifiable hypothesis**. E.g., the ROI on our marketing campaigns is greater than break-even.
Research Design	Design a **strategy** that allows you to test your hypothesis, noting all threats to validity.
Analyze	Analyze all data and evidence. **Test for threats to validity**.
Communicate or Productize	**Communicate** results in a way that stakeholders will understand or engineering can use.

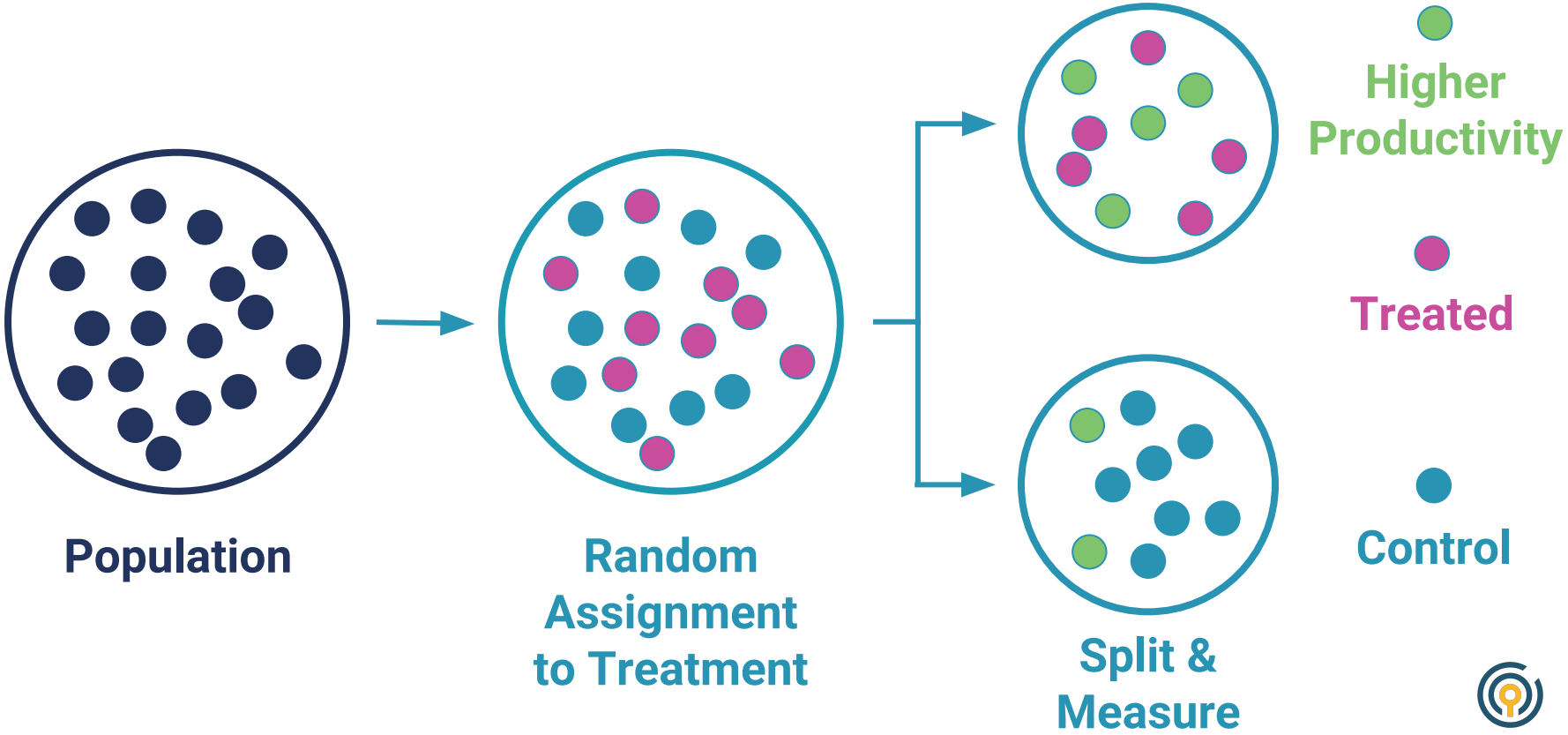
The current and coming credibility crisis in data science

Question	Objectives are often **unclear** and **success is** left **undefined**.
Hypothesize	Research, if necessary, and write a **falsifiable hypothesis**. E.g., the ROI on our marketing campaigns is greater than break-even.
Research Design	**Black-box(-ish) predictive models** are often the focus. Threats to validity are an **afterthought** or brushed aside.
Analyze	Analyze all data and evidence. **Test for threats to validity**.
Communicate or Productize	Decision-makers **don't understand the value** of data science projects.

So, data science is about running experiments?

Well, kind of.

The randomized control trial is the gold standard of scientific discovery



But sometimes a randomized control trial is a hard internal sell (yes, these are often strawman arguments)

A/B Testing for Outbound Sales	Experiment to understand Digital Ad Effectiveness	Intervention Program for Customer Retention
No one wants to be "B."	**Running a control ad is too expensive.**	**The results will be difficult to interpret.**
"That's taking food out of my mouth."	"I have to pay what? Just to buy space for some PSA?"	"We'll never really know whether what we tried was the real reason people stayed."

And even if we could, sometimes experiments can go wrong

This is where the social science comes in handy.

This is where the social science comes in handy.

And I’m not just justifying my life choices.

This is where the social science comes in handy.

And I’m not just justifying my life choices. This is mostly true.

In the absence of an experiment, we might try to measure an intervention by running a regression

We have a bunch of data on an intervention, say it's a job training program, and we want to know how well it worked.

We ran the program, and about 65% of people went through the training. We know it's not an RCT, so we want to run a regression to control for all possible sources of non-random assignment.

$$y = X\beta + \gamma D + \epsilon$$

where y is a measure of productivity, X contains information about all employees, and D is a dummy variable that says whether or not someone participated in the program and will give us the effect of the training on productivity.

Here are some things that could go wrong with a regression approach

Omitted Variable Bias (Confounding)	Endogeneity Bias (Simultaneity)	Sample Selection Bias
Did we omit variables that could plausibly explain our outcome of interest?	**Are we sure that we understand the direction of causation?**	**Is our sample truly representative of our population of interest?**
People may pursue training or further education on their own, or one office may have changed or improved the curriculum.	Low productivity offices may have been targeted for training programs.	We didn't track who attended trainings but ran a survey after. Only 40% responded and 65% of those attended a training.

Omitted variable bias: a brief math stat digression

Consider the linear model

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2$$

Suppose we've omitted a variable q with an additive effect where the true model is

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \gamma q$$

Since we don't observe q, we estimate the model

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \upsilon$$

where

$$\upsilon \equiv \gamma q + u$$

If q is correlated with *any* X, we can't estimate the βs. Consider the regression

$$q = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + r$$

Plugging this into the equation that we're trying to estimate gives

$$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)x_1 + (\beta_2 + \gamma\delta_2)x_2 + \gamma r + \upsilon$$

Research design strategies from the social sciences

Avoiding threats to validity

Use instrumental variables for overcoming simultaneity or omitted variable biases

Avoid problems like “reverse causation,” $Y \rightarrow X$, or common unobserved confounders, $C \rightarrow X \wedge C \rightarrow Y$, by replacing X with “instruments” that are correlated with X but not caused by Y or that affect Y but *only* through X.

An example from the Fulton fish market: stormy weather as an instrument

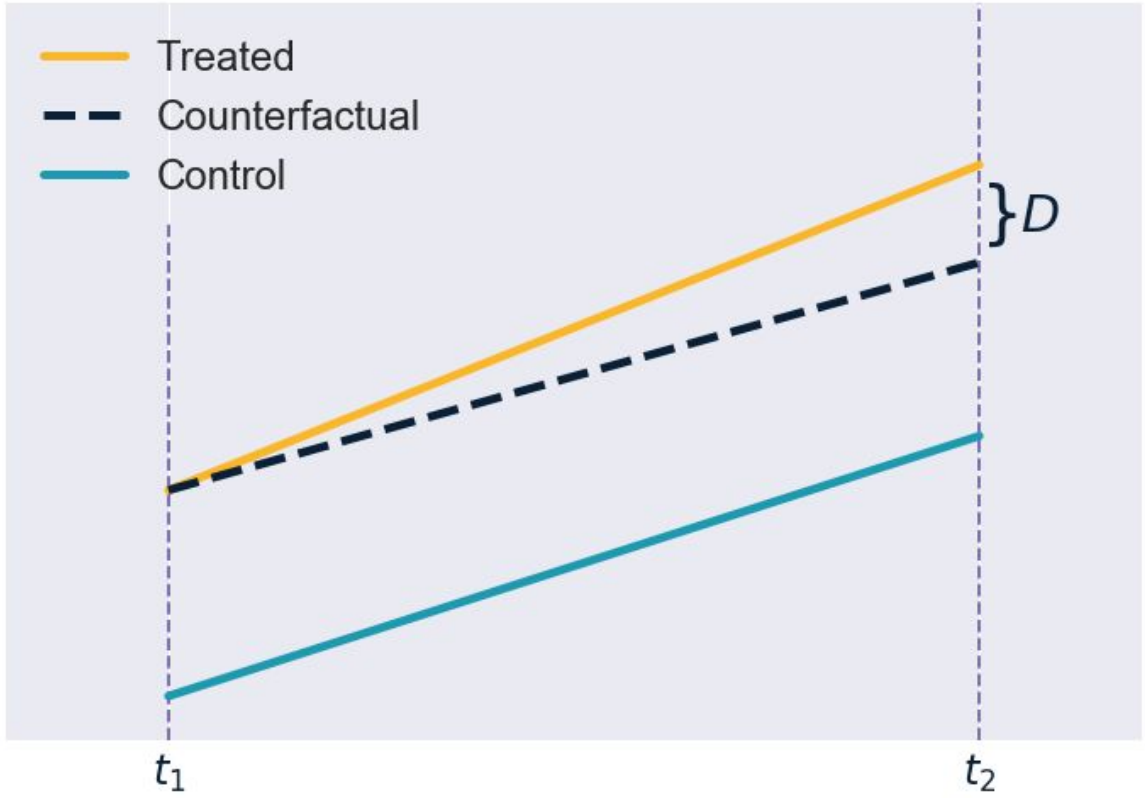
Use matching to overcome non-random treatment assignment

Makes up for the lack of a properly randomized control group by “matching” treated people to untreated people who look almost exactly like the treated group to create a “pseudo-control” group. Then look at the average treatment effect across groups.

Use difference-in-differences for repeated measures to account for underlying trends

Used when you don’t have an RCT, but you observe one or more groups over two or more periods of time, and there is an intervention between time periods.

A simple example of difference-in-differences



Regression discontinuity exploits thresholds that determine treatment

Exploits technical, natural, or randomly occurring cutoffs in treatment assignment to measure the treatment effect. Those observations close to but on opposite sides of the cutoff should be alike in all but whether they are treated or not.

A simple example of a regression discontinuity design

Survey experiments can be a great additional tool

Somewhere in between a “true experiment” and a quasi-experiment, a survey experiment gives **slightly changed survey instruments** to **randomly selected** survey participants. They provide an interesting way to test many hypotheses and are heavily and increasingly used in the social and data sciences, particularly in political science.

Some types of survey experiments

Survey Experiment	Description
Scenario-based	Change aspects of a hypothetical scenario. When combined with a conjoint design, can be used, for example, to understand how users view different aspects of a potential product without asking every variation of everyone.
Priming & Framing	Change the context or wording of questions. This can be used, for example, to understand how some message or creative content influences a survey taker vs. a control group.
List	Change whether a sensitive item is included in a list of response options. May be able to avoid types of biases that can arise in other approaches. Can be used to measure brand favorability after some bad publicity, for example.

By using these methods, you discuss particular threats to validity and whether identifying assumptions are met

Method	Identifying Assumption
Instrumental Variables	The instrument *z* affects the outcome *y only* through the variable *x*, not directly (Exclusion Restriction), & the instrument *z* is correlated with the endogenous variable *x* (Relevance Restriction).
Matching	Assignment to treatment is random conditional on observed characteristics (Selection on observables; Unconfoundedness). You have covariate balance between the treatment and control groups. (Overlap).
Regression Discontinuity	Look for clumping at the discontinuity. People may adjust their behavior, if they understand the selection mechanism.

There have also been many interesting advances in machine learning aimed at causal inference

- Bayesian Additive Regression Trees (BART)
- Post-Selection Inference (Taylor, Barber, Imai & Ratkovic)
- SuperLearner
- Interpretable Modeling (LIME, SHAP)
- Causal Trees & Forests (Imbens & Athey)
- G-Estimation
- Double Machine Learning (Chernozhukov, Hansen)

A few examples of research design-based thinking at work

A non-random walk through some (mostly) recent papers.

A few examples of research design-based thinking at work

A non-random walk through some (mostly) recent papers. Pro tip: start or join a journal club at work.

A retreat to experiments

"Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising"

by Randall A. Lewis, Justin M. Rao, and David H. Reiley

Observational experiment methods often overestimate the causal impact of advertising due to “activity bias”

01	**Effects on Searches**	○ 250m impressions on Yahoo! FP w/ 5% control ○ 5.4% increase in search traffic via experiment ○ Matching (1198%), Regression (800%), Diff-in-diff (100%)
02	**Effects on Consumption of Yahoo! Pages**	○ Survey experiment on Amazon Mechanical Turk ○ Treatment (800) and control (800) of ad on Y! activity ○ Control and treatment had the same effects
03	**Competitive Effects**	○ 200M impressions for Major Firm on Y! w/ 10% control ○ Tracked sign-ups on a Competitor Firm ○ Both saw more sign-ups for both ads! (And Y! usage)

The limits of measurement

"The Unfavorable Economics of Measuring the Returns to Advertising"

by Randall A. Lewis and Justin M. Rao

It is extremely difficult *but not impossible* to measure the effect of an advertising campaign

Small Effect Sizes

Campaigns have a low spend per-person, and the ROI on a break-even campaign is small.

+

Noisy Effects

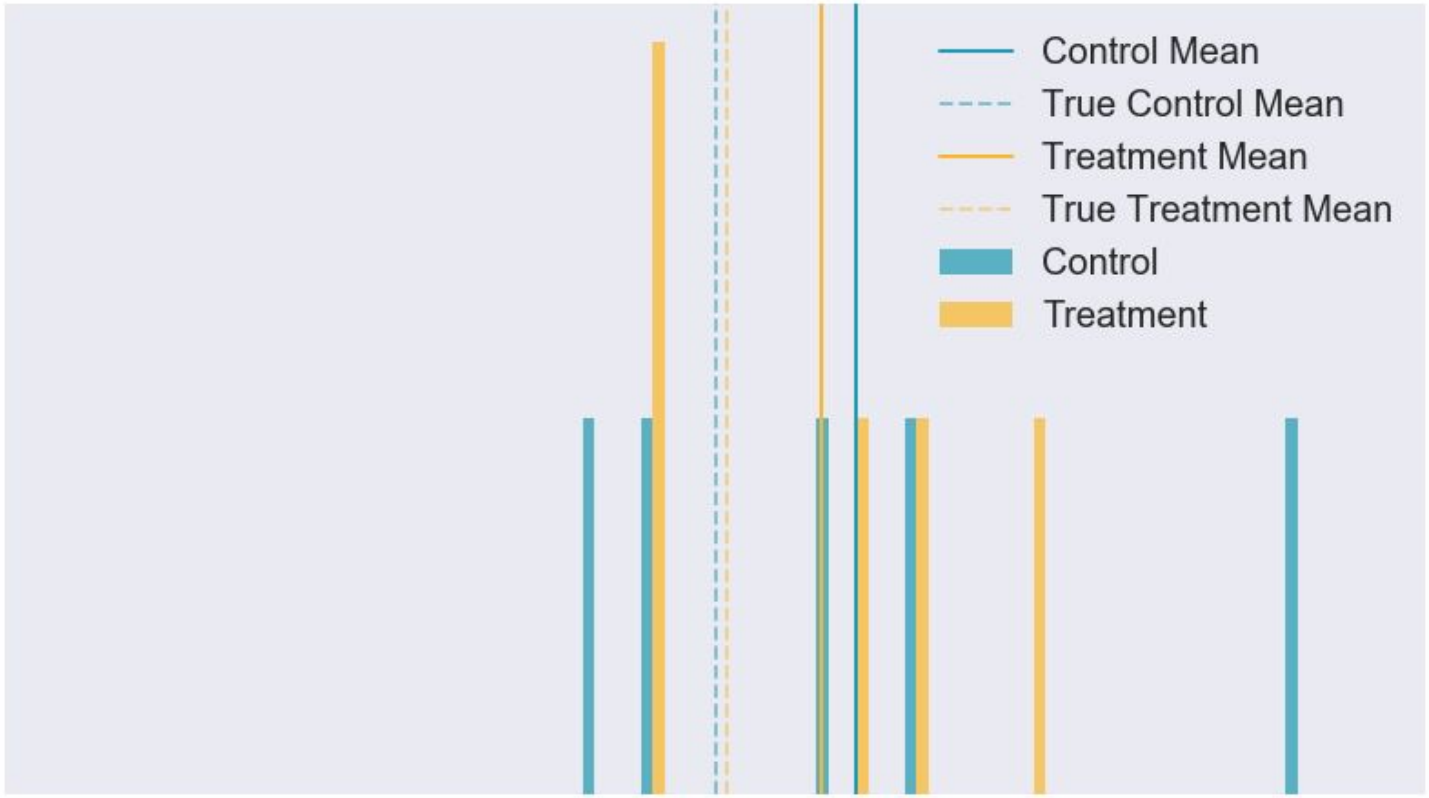
The standard deviation of sales vs. the average is around 10:1 across many industries.

→

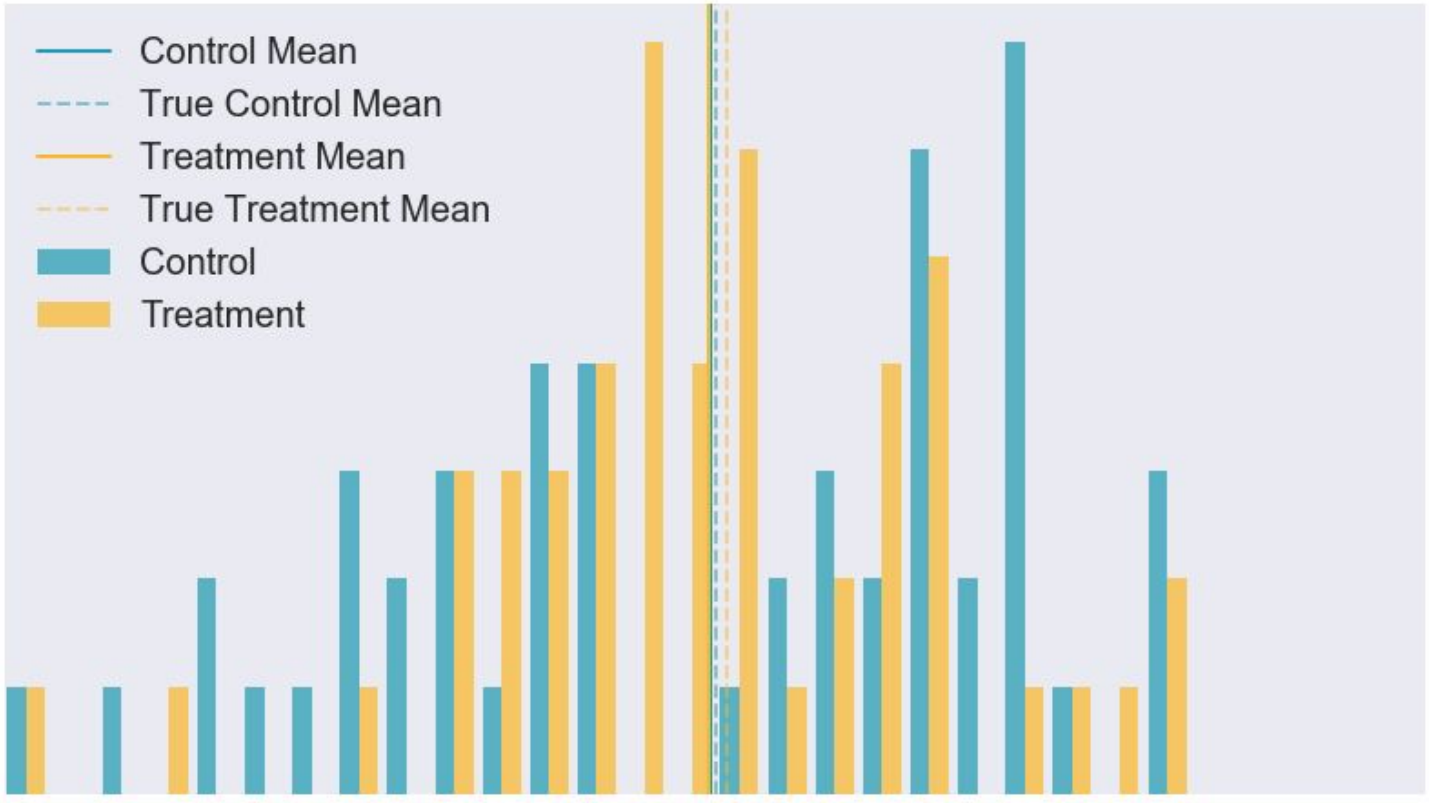
Experiments & Observational Methods

Designing an experiment to measure this effect is hard. Observational methods look attractive but overestimate effects due to targeted advertising.

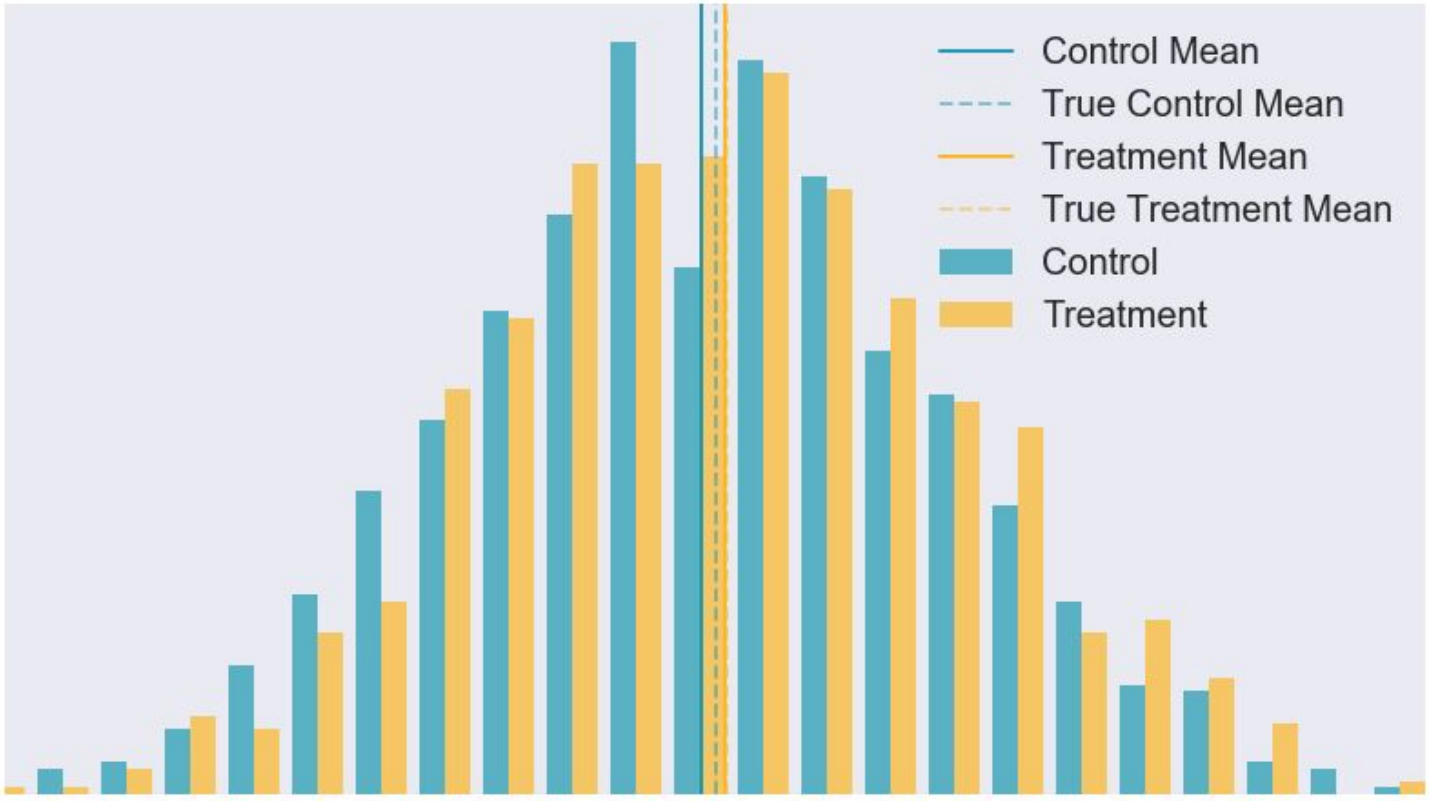
Measuring effects (in a sampling framework) (n = 10)



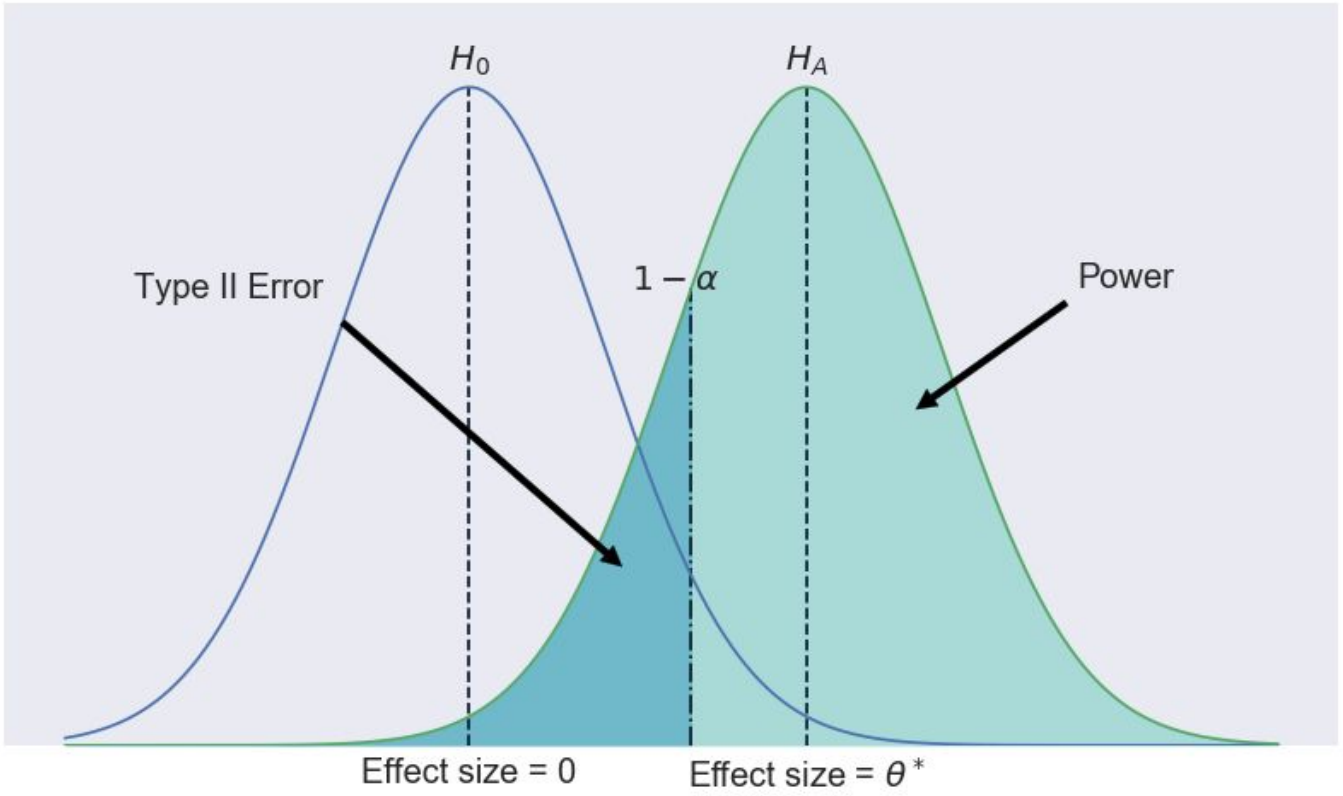
Measuring effects (in a sampling framework) (n = 50)



Measuring effects (in a sampling framework) (n = 1000)



The power of an experiment



Experimental design for control vs. treatment digital ads

Study Aspect	Description
Industries	Retail Sales (including department stores) & Financial Services
Measured Outcome	New Sales & New Account Sign-Ups
25 Experiments	Ran for 2-135 days (median 14); Cost $10k - $600k+; Measured 100k to 10m impressions

Every single experiment studied did not have enough power to measure a precise 10% return

Question	Measurable?	Result
No Impact vs Positive Return	9 out of 25 fail to reject the null of no impact; 10 out of 25 have enough power to test.	The powered experiments generally reveal a significant, positive effect for advertising.
Profitable Campaign	12 of 25 are severely underpowered; only 3 had sufficient power to conclude they were “wildly profitable” (ROI of 50%).	The median campaign would have to be 9 times larger to have sufficient power.
Greater than 10% Return	Every experiment is underpowered.	The median retail campaign would have to be 61 times larger. For financial services, 1241 times larger (!).

Change the course of industry

"*Courtyard by Marriott*: Designing a Hotel Facility with Consumer-Based Marketing Models"

By Jerry Wind, Paul E. Green, Douglas Shifflet, and Marsha Scarbrough

A well-designed survey experiment resulted in a wildly successful product ideation for Marriott

“Marriott used conjoint analysis to design a new hotel chain [in the early 80s].”

The survey design asked two target segments about 7 facets, considering 50 attributes with 2 to 8 levels each

7 FACETS	Leisure
External Factors	Lounge
Rooms	Services
Food	Security

How successful was this experiment?

As of 1989, Marriott went **from 3 tests cases** in 1983 to 90 hotel in 1987 with more than $200 million in sales. The chain was expected to grow to **300 hotels by 1994** with an expected **$1 billion in sales**.

Captured **within 4%** of predicted market share.

Created **3,000 new jobs** within Marriott with an expected **14,000 new jobs by 1994**.

Affected a **restructuring of all competitors** the midprice-level lodging industry (upgrades, prices, amenities, and followers).

Wrapping Up

A credibility crisis in data science is preventable

Question	Work with stakeholders to **focus on business relevant** and **measurable outcomes** from the start.
Hypothesize	**Be clear** about the causal mechanisms that you will test.
Research Design	**Keep it sophisticatedly simple**. A well thought out research design leads to better communication of results.
Analyze	**Be honest about any threats to validity**. You can test these in the future. This is scientific progress.
Communicate or Productize	**Go the last mile.** This step enables data science to have **decision-making relevance**.

Did I miss some good papers? Come @ me on twitter. Until then, here are some more papers and books on my desk.

- "A comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook"
- "Confounding in Survey Experiments"
- "Using Big Data to Estimate Consumer Surplus: The Case of Uber"
- *Mastering Metrics: The Path from Cause to Effect* [Looks good!]
- "A/B Testing" [Experiments at Microsoft using Bing EXP Platform]
- "Measuring Consumer Sensitivity to Audio Advertising: A Field Experiment on Pandora Internet Radio"
- "Attribution Inference for Digital Advertising using Inhomogeneous Poisson Models"