w241: Experiments and Causality

Final Thoughts on Experiments and Causality

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-08-20

Observation vs. Experiment

Observation

- Data analysis allows for making decisions
- Decisions involve counterfactuals
  - Existing in the state of the world where one has done X or Y
  - Eg. should women receive hormone replacement therapy (HRT) or not?
  - Eg. Should prices be raised or not?
- Observational data: Compare units with different X values

Experimentation

- Experiments involve **interventions**
  - Rather than only observing, as an analyst you get involved in giving treatments

- Randomization
- Focus on selection process
  - How do units get different X values?
  - How did units get into the groups?

- Quite often, units have different X values because of pre-existing differences
  - People and firms make choices for a reason
  - Typically implausible to believe X is assigned haphazardly, especially if it's reasonable to think X affects Y
- In experiments, X values are determined by randomization, guaranteeing subjects' Y values would be otherwise be similar if there were no treatment effect
- If we are wrong, it can be proven

- Field experiments allow us to infer **causal relationships** in the real world
  - Study real-world conditions as closely as possible

Prediction vs. Inference

Prediction vs. Causal Inference

- In previous years, there have been huge advances in predictive accuracy of statistical models
- Sometimes only need to predict Y:

Some examples:

Question	Decision	Experiment Needed?
How many shoes will I sell next month?	How many shoes should I stock?	No
How many website visits will I get?	Which web hosting plan to buy?	No
Are men or women more likely to buy my product?	Who to market to?	**Yes**

- Subtle difference: person most likely to do something won't necessarily be most likely to respond

Mistaking Prediction for Causal

Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

Overview

- Specifically branded searches
- Seemingly strong evidence that Google search clicks have great return on investment (ROI)
    - People who click often buy
    - Very strong correlation between number of sales and number of clicks
    - Statisticians didn't want to decrease variable that seemed to predict sales well

Mistaking Prediction for Causal Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

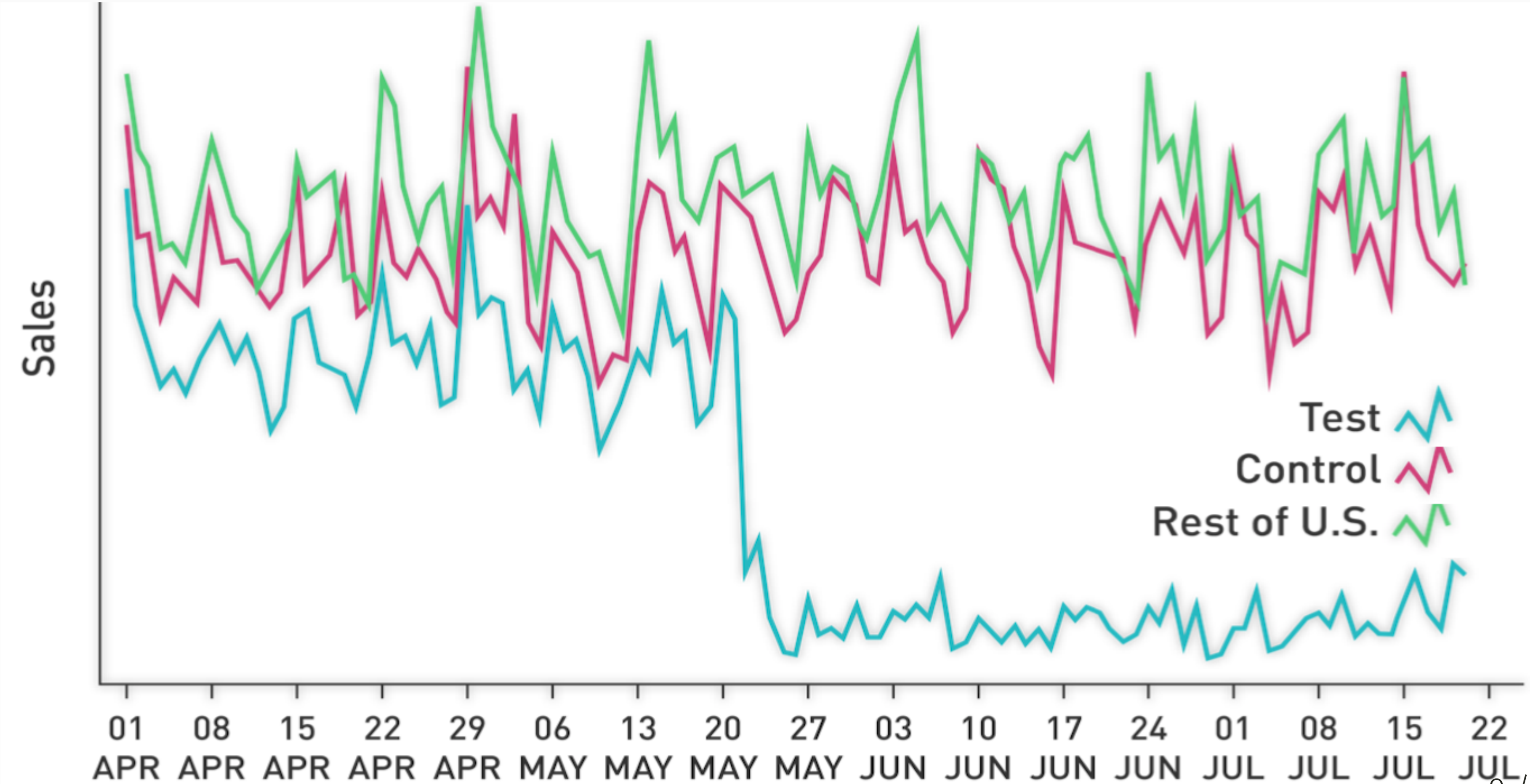
Experiment

- If ads weren't shown, would people searching "eBay" end up on eBay.com anyway?

Mistaking Prediction for Causal

Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"



Mistaking Prediction for Causal Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

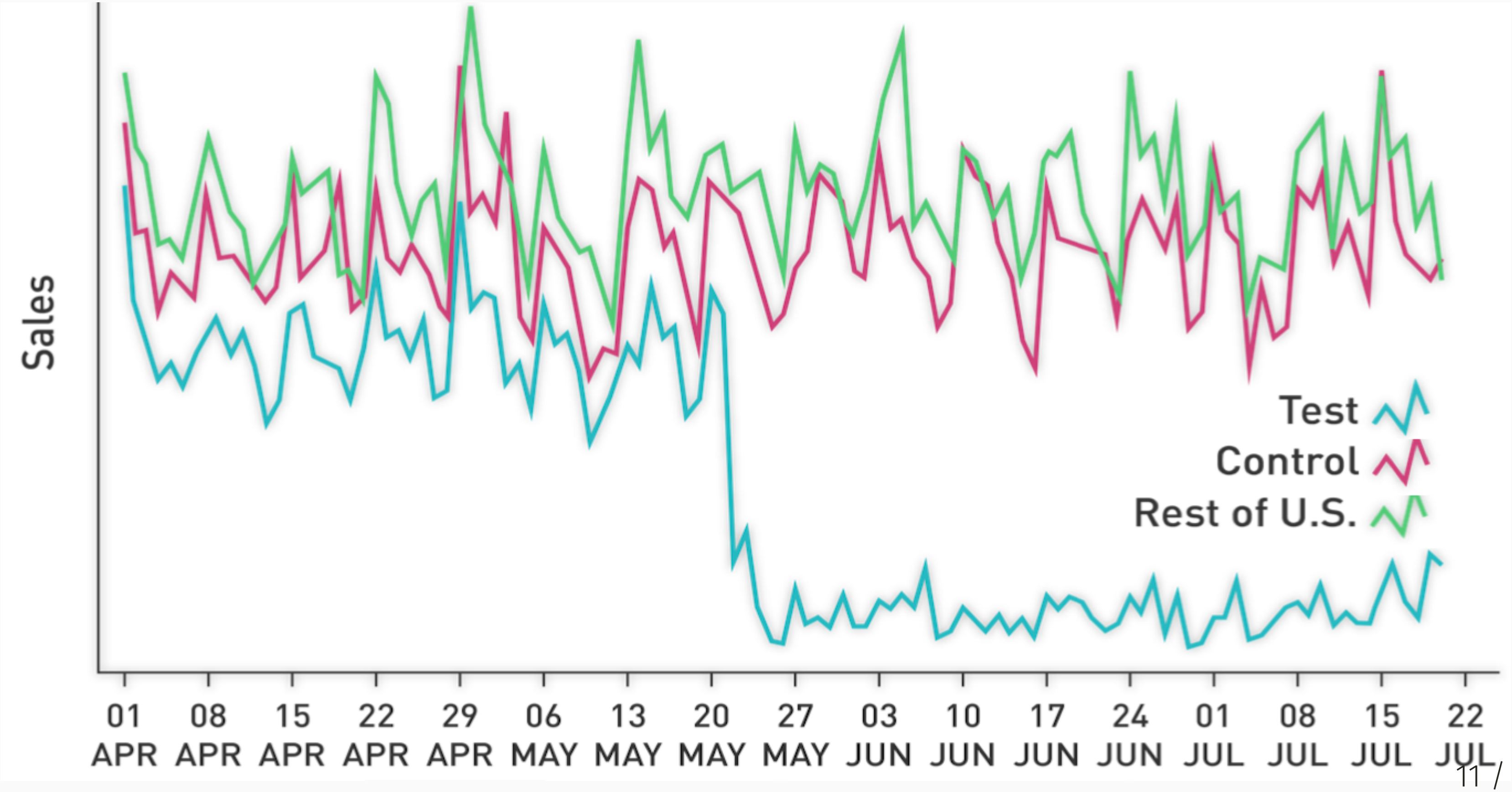
Experiment

- If ads weren't shown, would people searching "eBay" end up on eBay.com anyway?
    - Randomly assigned some regions to get ads while others didn't

Mistaking Prediction for Causal

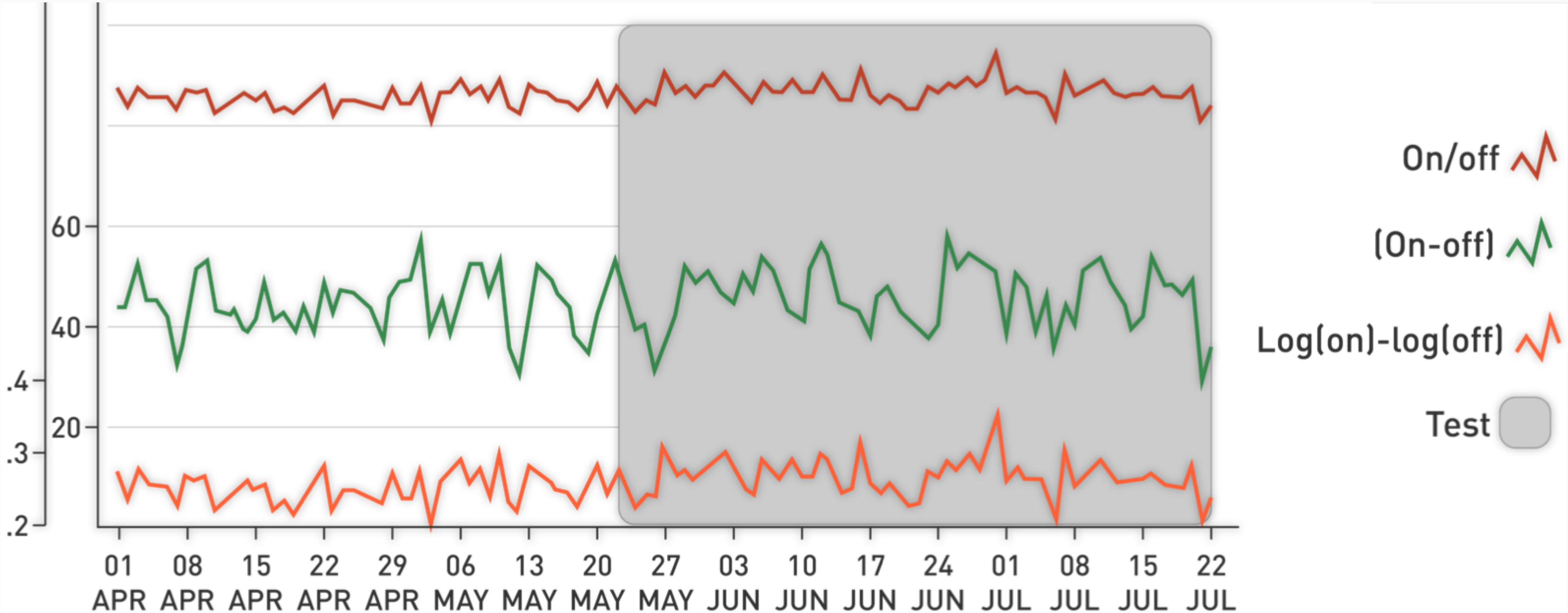
Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"



Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"



Mistaking Prediction for Causal Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

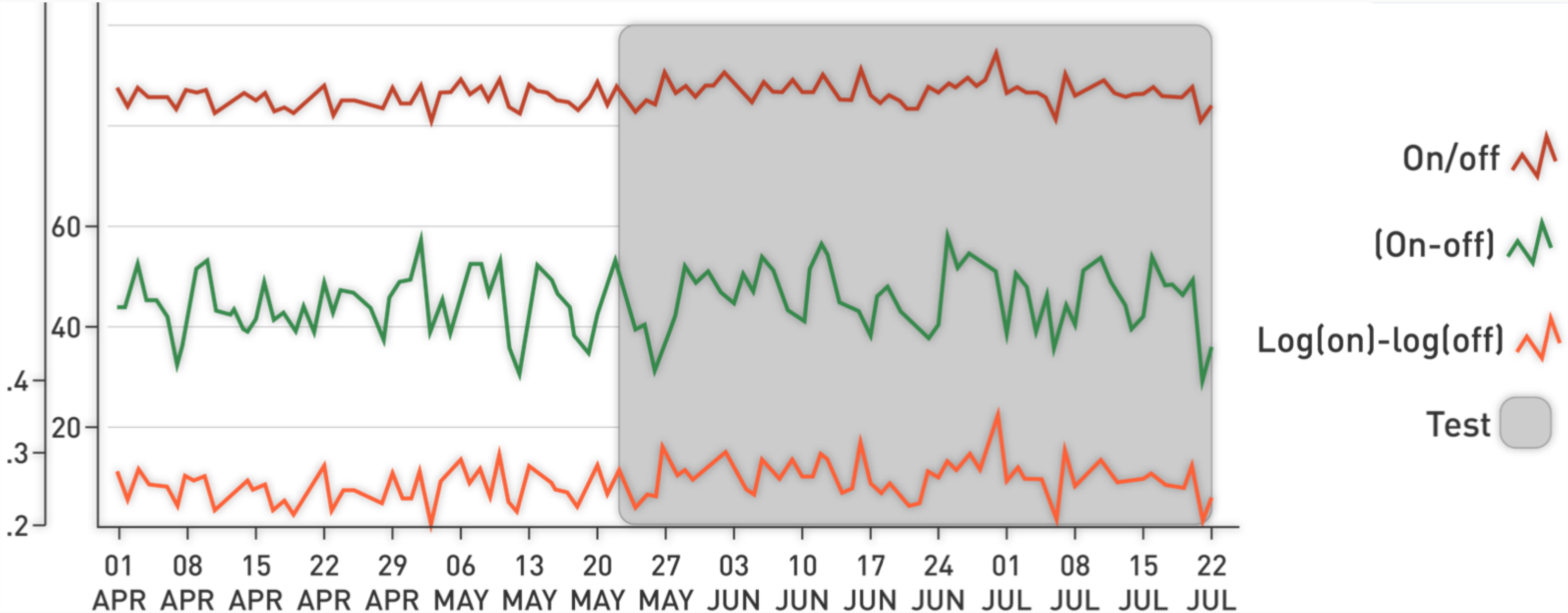
Experiment

- If ads weren't shown, would people searching "eBay" end up on eBay.com anyway?
  - Randomly assigned some regions to get ads while others didn't
  - If attribution model is correct, total sales should go down too

Mistaking Prediction for Causal

Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"



Mistaking Prediction for Causal

Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

Experiment

- If ads weren't shown, would people searching "eBay" end up on eBay.com anyway?
  - Randomly assigned some regions to get ads while others didn't
  - If attribution model is correct, total sales should go down too
  - Observation alone predicts 1 USD spent yields 417.3 USD in revenue (4,173% ROI)

Mistaking Prediction for Causal

Inference

Example: Blake *et al.*: eBay ads on Google searches for "eBay"

Results

- Experiment shows people who clicked ad would have gone to website anyway
- Actual Return on Investment: -63
- 1 USD spent yields 0.37 USD revenue

Experiment showed there wasn't a causal effect

Misuses of Predictive models

Predictions and Decisions

Example: You are a marketer who learns that women are more likely to buy product than men

- Should we advertise more to women?
- Can predict effect of advertising

Misuses of Predictive Scores

- Firms often create predictive model scores
  - Predicts likelihoods
- Predictive models can yield predicted values without clear causal implications

Example: Magazines

- Model for subscription cancellations
  - Percentage chance of cancellation over next few months
- Discount for people likely to cancel
- Problem: Not known if people most likely to leave will be responsive to discounts
- Experiment where random sampling of subscribers received discount
  - Heterogeneous treatment effect
- Only way to be sure it by running intervention

Example: Voting

- Likelihood of voting for Republican candidate
  - Idea: Target "moderates" (40-60) with persuasive appeals
  - 40-60 not moderates, just people we are bad at predicting

Even if person is a moderate, doesn't mean he or she would be receptive to appeals

Predictive models often don't work out in practice

Common Themes

1. Treatment effect different from Y
2. Assumptions can exist without being aware of them

Attempts to Fix Observational Data

Fix Observational Data

Three Techniques

1. **Matching:** Compare units with similar values
2. **Regression adjustment:** Multivariate regression
3. **Propensity scores:** Model likelihood of receiving treatment

The truth of causal epistemology:

There is no free lunch

- All modeling choices will make tradeoffs

Matching

Compare subjects with very similar values of covariates

> - "Among women of the same race, with similar incomes, blood pressure, height, and weight..."
> - "...those who take HRT are less likely to get cancer than those who don't"

- Still don't know if we have all the necessary covariates
- Potential for unobserved heterogeneity still exists in matching analyses
- What are the reasons people who are so similar get different treatments?
- There can be unknowns that don't exist in data set

Regression Adjustment

Imposes a functional form on the link between covariates, treatment, and outcome

- Extremely similar to matching
- **Example:** People who weight more are more likely to take HRT
    - Remove effect of weight by adjusting for it
- Some underlying move as matching
- Covariates don't always have linear relationship between outcome and treatment
- Compare people within similar values of covariates
- Still don't know why some subjects got treatment and some did not

Omitted Variable Bias

Unobserved heterogeneity big problem with experiments

- Incinerator example
    - Researchers had done regression adjustment (ie. "controlled for")

Matching won't always show unobserved differences

- Can't measure everything
- Experimentation allows for unobserved things to be balanced

Fixing Observational Data

Part II

Propensity Scores

- Key challenge in causal inference is potential connection between **likelihood of treatment assignment** and **outcome**
- If units more likely to get a treatment *also* have different Y values for other reasons, comparisons between treatment and control will reflect these non-causal differences

Propensity scores estimate likelihood of receiving treatment directly

- Strategy: Compare units with similar probability of treatment

Example: Propensity Scores

- Overweight rich women and underweight poor women have a similar chance of receiving treatment, so compare those groups with and without HRT to each other
    - Model suggests similar likelihood of receiving treatment
    - If probability of treatment is known, can get unbiased causal effects
    - Problem remains: We don't know all the reasons why people get treated
- Propensity score can be wrong for many units
    - Eg. underweight poor women have 80% chance of treatment
    - Possible some have 99% chance of treatment due to poor health
    - Other group could have 10% chance
    - Therefore, unclear what is true chance of being treated

Propensity score matching is another way to do matching

Have all reasons for treatment been measured?

Common Themes

1. Controlling for observables

- We can control for observable features, but we cannot control for things that we cannot measure
- Might seem tautological, but this point can be challenging to communicate

2. We might ask

- Are there differences between the kind of people who either receive treatment or do not receive treatment?
- Are we able to observe what is different?
- What if we cannot observe what is different?

Shoe Leather

Tremendous Effort Examples

Bertrand and Mullainathan

- Thousands of fake resumes
- Thousands of employer listings

Careful research takes real effort

- Many people don't want to do careful research because its difficult and requires effort
- Many times it doesn't feel "fancy"
- The right kind of data often hard to get

Snow and Cholera

- **Hypothesis:** Disease isn't spread through "miasma"
    - Contended that cholera is a waterborne disease
- Ideal Experiment: Randomly assign houses to water companies
- Natural experiment existed
    - Pipes were laid many years prior in same neighborhoods
    - Arbitrary who has which water company
- Took lots of effort to gather data
- Knocked on doors to determine people's water company and if they had cholera
- The table he produced was very simple

	Number of Houses	**Deaths**	**Deaths per 10,000 Houses**
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Making the Effort

Put onus on those making assumptions

- Why do units get their X values?
- What determines which units get in groups being compared?
- Why believe an artificial setting speaks to the setting that's important?

Some people will say it's impossible to do an experiment that will rigorously answer the questions -- take this as a challege!

- Think carefully about how to conduct an experiment that will answer big question

Worth the work to do careful research

- People will say they can't help
- Then they will be surprised people are cooperating
- Then they will fight the findings

Worth the time and effort

Deception and Privacy

Deception and Privacy

Field experiments affected in particular

- Intervention is occurring
- Affecting real people in real world

Consider ethical implications of choices

Example: Food Poisoning Letters

- Fake letters sent to restaurants, claiming food poisoning
- Testing customer-service responses
- Restaurant employees were fired erroneously
- Professor conducting study got in big trouble
- This really was not an ethical study

Example: Bertrand and Mullainathan

- Measured racial discrimination in job market
- Sought to quantify effects of race during hiring process
- Firms receiving fake resumes were misled and had time wasted
- People were unknowingly participating in study without giving consent

Privacy

- Ethical intuitions still evolving
- Privacy policies make research difficult
- Often want to observe/match data but can't
  - Make case for importance of data desired
  - Find ways to not violate policy such as anonymizing data or randomly assign units in clusters
- Think creatively about how to conduct an experiment consistent with a privacy policy if it can't be changed

Ethics

- Consider costs and benefits of research
- Research ethics are cost/benefit analysis
- Look at subject's point of view
  - Tendency to treat subjects as objects
  - Consider human impact

Ethical principle: Always see your "treatment units" as real people

Ethics

Common argument: Withholding treatment from people in certain situations would be unethical

- Eg. bed nets to protect people from malaria

Two potential responses:

1. Often can't give treatment to everyone anyway
    - Consider alternatives
    - Random assignment and treating everyone possible are not incompatible
2. Consider benefits of research
    - If control group yields good results, it will benefit many more people in the long run

Change the World!

Experiments Are Changing the World

1. Development
2. Politics
3. Conservation
4. Business

Development Economics

See More than Good Intentions by Karlan and Appel, and Poor Economics by Banerjee and Duflo

- How do we increase education?
  - Provide uniforms to girls?
  - Ask teachers to take pictures of themselves?
  - Deworm kids at school?
  - Give cash to families?
- Prior to 2000, most development programs were never really tested
- With limited resources, allocate randomly
  - Can know which pilot programs to expand with additional funds
- Without experiments, no way to know the counterfactual

Politics

- Persuasion and mobilization of voters and volunteers
- How do we register minorities to vote and turn out?
- How do we make sure voters hold elected officials accountable for corruption?
- Which governance structures protect minority rights?
- How can activists affect politicians' behavior?

Questions can start to be answered based on science rather than philosophy

- Experimentation has transformed political world

Conservation

- Typical approach: blandishments to conserve
- Opower sends mail comparing neighbors' power use
  - Had large effect on people's conservation
  - Frequency of mailings?
  - Amount of social judgement?
  - Effect diminished after several months; new mailings needed
  - Optimal number of mailings to preserve "shock value"?

Business

- Employee incentives
- Pricing
- Advertising
- Audit studies for quality control
- Have to admit ignorance in order to justify experiment
  - People often set in their ways
  - Gathering data that proves you wrong can be uncomfortable

Be in the ignorance-reduction business!

- Reduce ignorance through the use of random assignment

Run experiments!

Redeem for:

Advice on designing experiments

david@davidreiley.com
david.broockman@gmail.com
dhughes@berkeley.edu

Expires: never