

From RCT's to A/B: Choosing Lean Methods for Evaluation

Kathryn Vasilaky

GroundTruthData.com,

Columbia University

Ralph Lin

GroundTruthData.com

Overview

- Randomized Controlled Trial(RCT) versus Agile Development (3)
 - What is a RCT? What is Agile? (2 minutes)
 - Motivating Average Treatment Effect (ATE)
- Getting at ATE (5 minutes)
 - Strong Ignorability
- Who can afford a RCT? (2 minutes)
- Going Agile, Mobile, SMS, IVR (5 minutes)
 - Rapid, low cost agile principles
 - What can we measure?
- Breakout Brainstorm (10 minutes)
 - Scenario to brainstorm
 - Design an agile approach via SMS or IVR
- One minute pitches on poster board (5 minutes)

Randomized Controlled Trial?

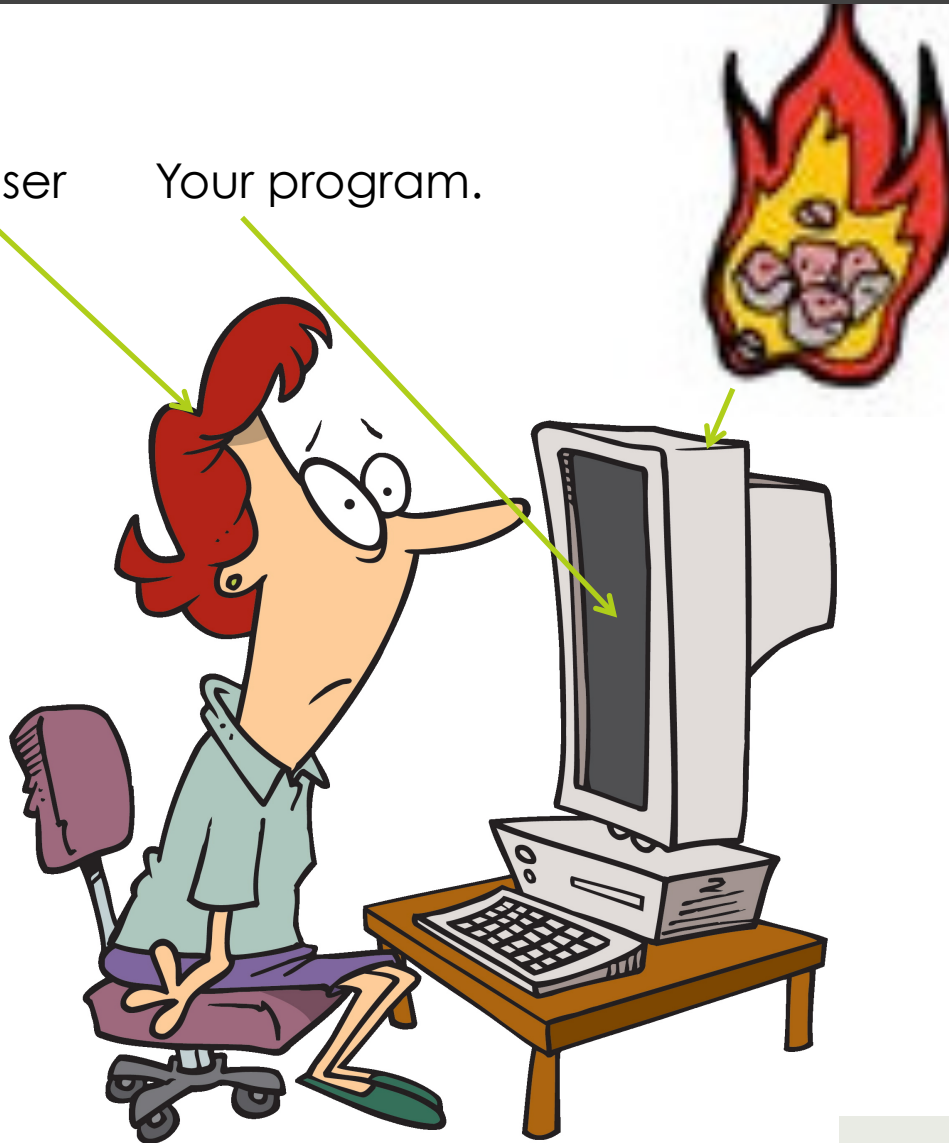
- Randomized Controlled Trial (RCT) is a method of studying the impact of an intervention or program;
 - It's thorough and randomized about one question for one population (external validity)
 - Often takes some time to conduct
 - Treatment is randomized
- Agile Development incorporates iterative software development methodologies.
 - Incorporates monitoring and evaluation, sometimes randomized (A/B Testing)
 - Is cheap
 - You fail fast

Why do a Randomized Controlled Trial?

Your user

Your program.

Or external events.



What's causing a change
In the user behavior?

- Your user?
- Your program?
- Or external events?

Simple Rules of Thumb

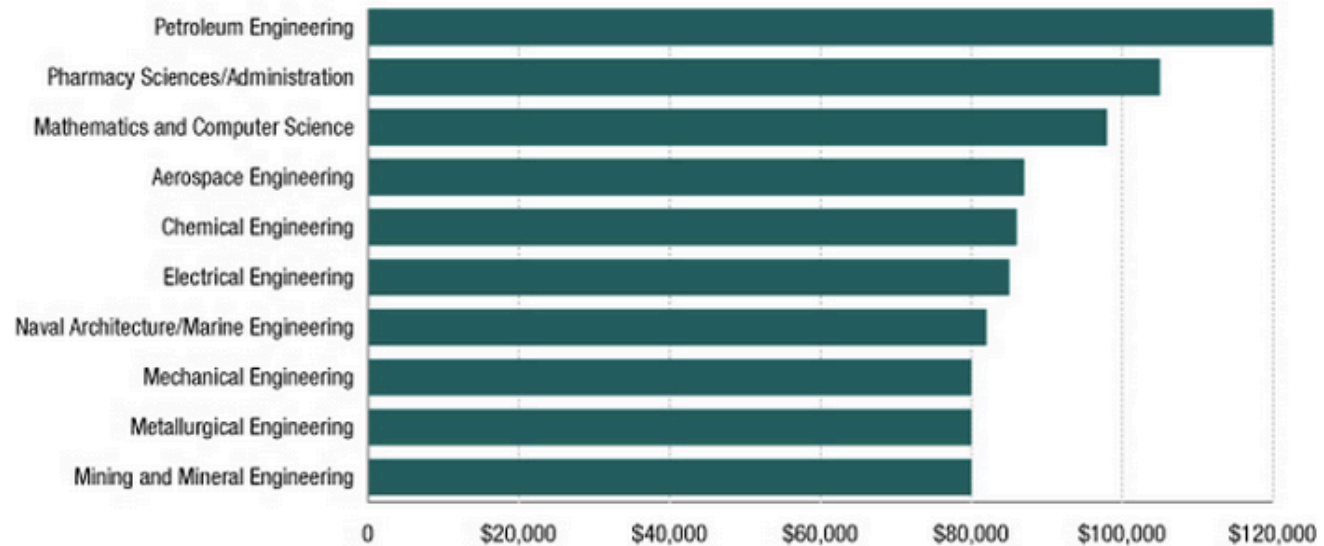
- You need a **control group** (C), and a treatment (T) group
- Only change **one thing** at time in your treatment groups
 - Answer your question-What needs to change?
- Calculate your sample size for T & C (**power**) *before* you begin the experiment
- Individuals should be **randomly** assigned

Getting at the ATE

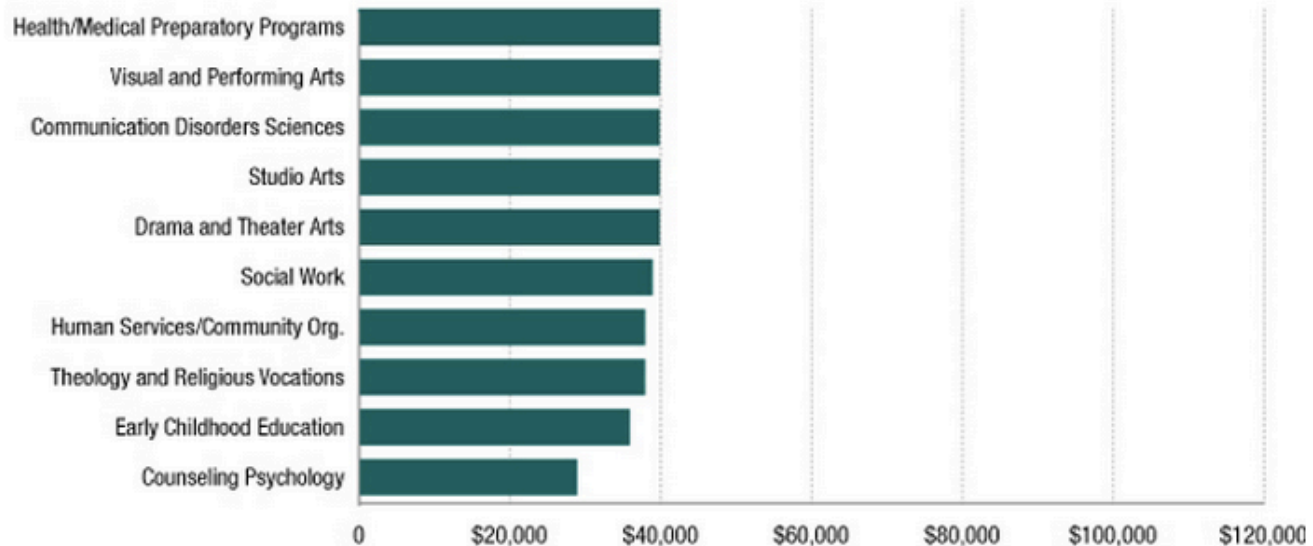
What's the effect of your college degree on your future earnings?

Effect of your major on your future earnings?

Majors With The Highest Earnings



Majors With The Lowest Earnings



Effect Size?

- Looks like if you go into petroleum engineering then you'll earn 100 K more than with a psychology major.

$$E(\text{Earnings} \mid \text{Engineering}) - E(\text{Earnings} \mid \text{Performing Arts}) = 80K$$

Probably Not

Individuals who go into engineering (E),



Probably Not

Are probably different from those who go into the performing arts (PA).



Rubin Causal Model

Difference in difference

$E(\text{Earnings} \mid \text{PE, After College}) - E(\text{Earnings} \mid \text{PE, Before College})$

–

$E(\text{Earnings} \mid \text{PA, After College}) - E(\text{Earnings} \mid \text{PA, Before College})$

Rubin Causal Model

Difference in difference

$E(\text{Earnings} \mid \text{EE, After College}) - E(\text{Earnings} \mid \text{EE, Before College})$

–

$E(\text{Earnings} \mid \text{PA, After College}) - E(\text{Earnings} \mid \text{PA, Before College})$

Why?

- What does the difference in difference accomplish?

Why?

A counterfactual is a “what if”:

- ▣ We hadn't received the program?
- ▣ We hadn't seen a policy change?
- ▣ We hadn't received aid?

Very broadly, it's what would have occurred in the absence of an event.

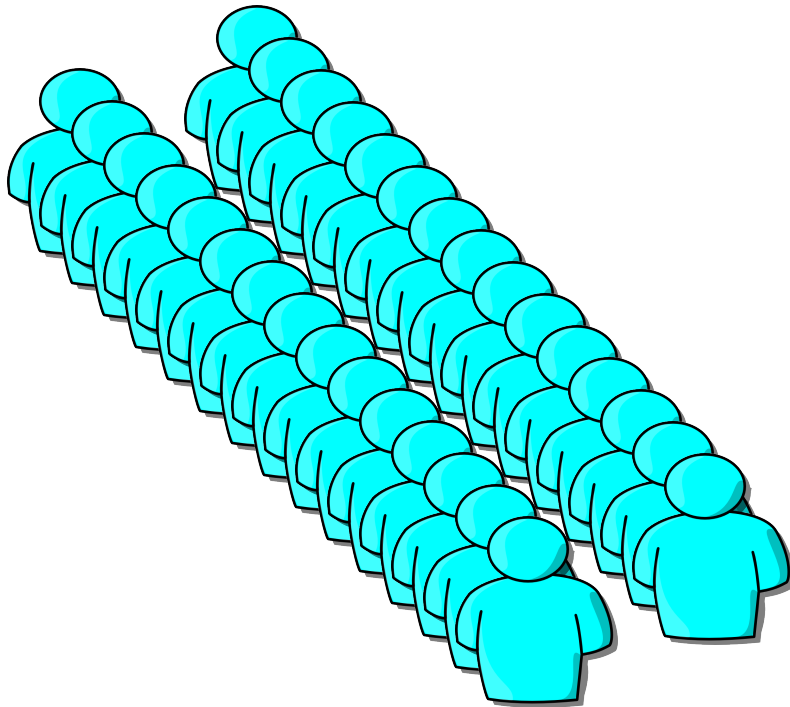
Would John Travolta have earned as much as an engineer as a performing artist.

Is it that simple?

- Not quite.
- A true counterfactual means that there is absolutely no correlation between individuals' characteristics who receive a treatment and who do not.
- Hence the randomization.

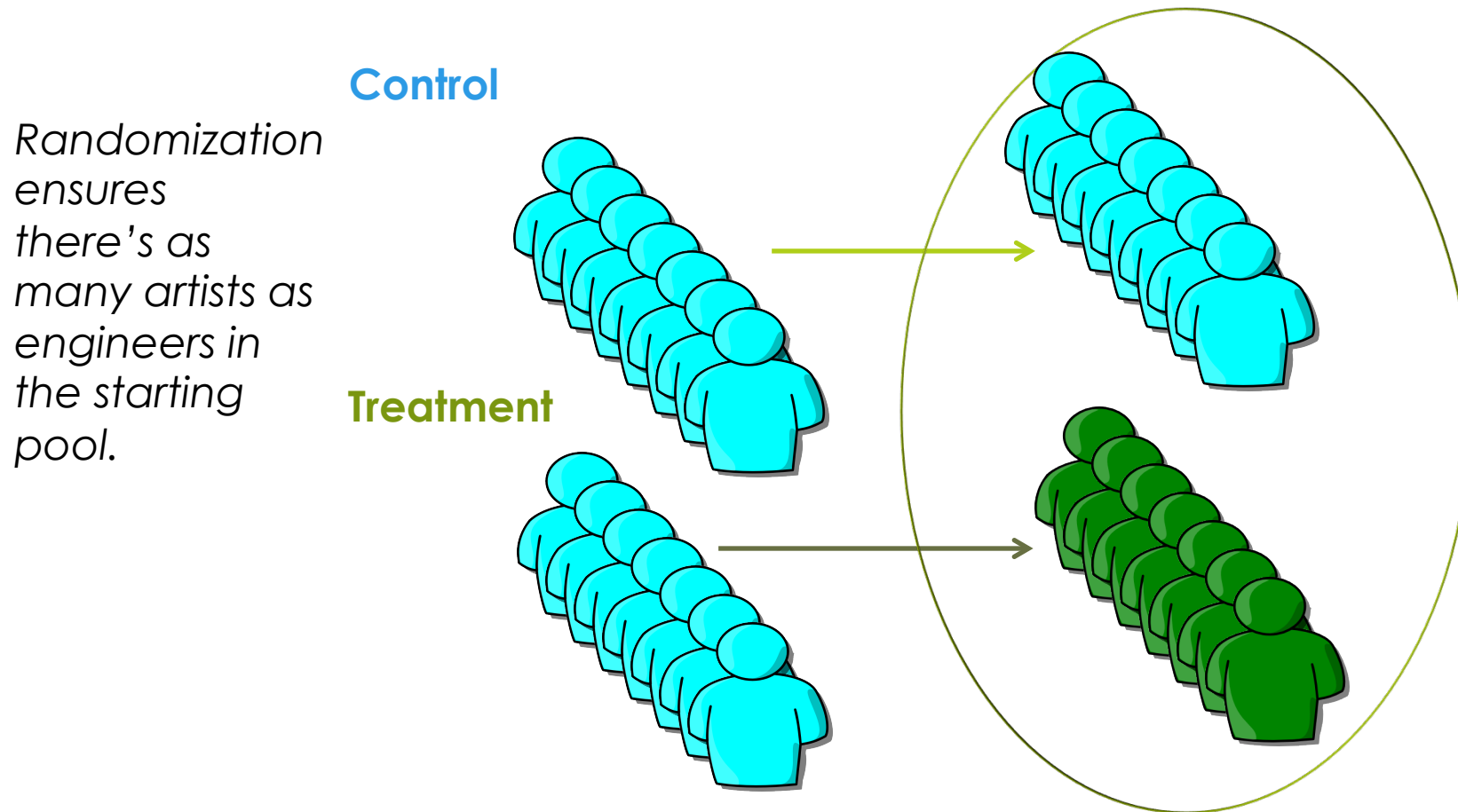
How?

Take a random sample of individuals from the population.



How?

Expose them to a stimuli at random, and compare the average outcomes.



Rubin Causal Model

Difference in difference

$$\begin{aligned} & E(Y \mid \text{Treat}, \text{time}=1) - E(Y \mid \text{Treat}, \text{time}=0) \\ & \quad - \\ & E(Y \mid \text{Control}, \text{time}=1) - E(Y \mid \text{Control}, \text{time}=0) \end{aligned}$$

So what does randomization give us?

Strong Ignorability & SUTVA

Compare treated individuals to similar individuals who were not treated.

$$E(Y | \text{Treat}, \text{time}=1) - \cancel{E(Y | \text{Treat}, \text{time}=0)}$$

–

$$E(Y | \text{Control}, \text{time}=1) - \cancel{E(Y | \text{Control}, \text{time}=0)}$$

Strong ignorability: $E(Y | \text{Treatment}, \text{time}=0) = E(Y | \text{Control}, \text{time}=0)$

So that comparing $E(\text{Treatment}, \text{time} = 1) - E(\text{Control}, \text{time} = 1)$ (or treatment on the treated), will actually give us the average treatment effect, if matching is good enough (as having a counterfactual).

Experiments in Regression

- DID in regression form:

$$Y = a + b * \text{Treatment} + c * \text{Time} + \textcolor{red}{d} * \text{Treatment} * \text{Time} + \text{error}$$

Treatment=1 if treated

- b is the effect of random treatment assignment
- c is your time effect
- **d** is your treatment impact

Unbiased Estimates

- We want $b=0$
- Any time the $\text{Corr}(\text{Treatment}, \text{error})=0$, o.t. estimate of **d** will be biased.

$$Y = a + b \cdot \text{Treatment} + c \cdot \text{time} + d \cdot \text{Treatment} \cdot \text{time} + \text{error}$$



Experiments

In Reality:

RCTS will still difference out the starting points, because randomization may not go as planned, *and* you might also add in controls.

$$E(Y \mid \text{Treat, time}=1 \mid X) - E(Y \mid \text{Treat, time}=0 \mid X)$$

–

$$E(Y \mid \text{Control, time}=1 \mid X) - E(Y \mid \text{Control, time}=0 \mid X)$$

In Reality:

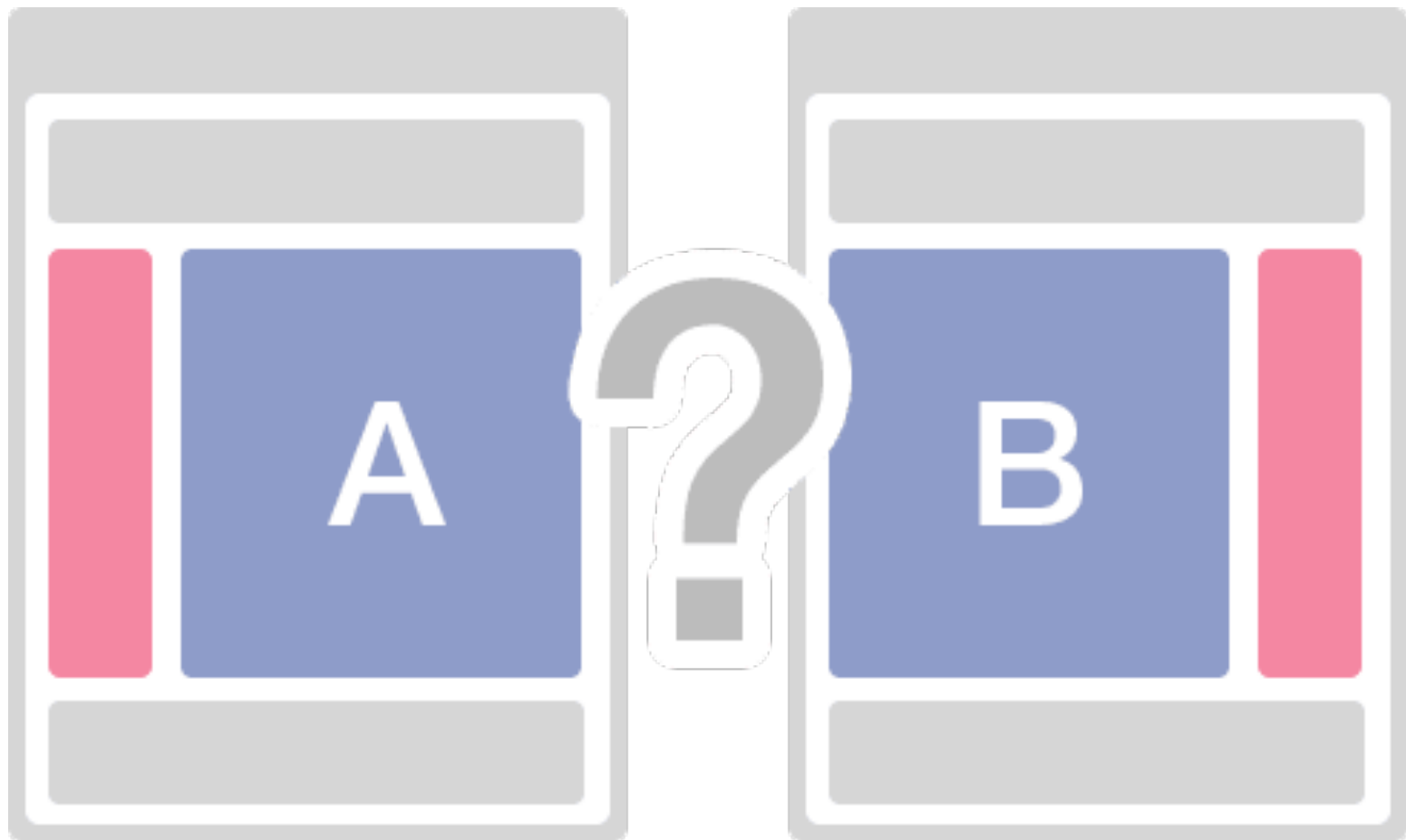
- Lab experiments and **A/B testing** tend to just look at:

$$E(Y | \text{Treat}, \text{time}=1) - E(Y | \text{Treat}, \text{time}=0)$$

$$E(Y | \text{Control}, \text{time}=1) - E(Y | \text{Control}, \text{time}=0)$$

- Because the environment is controlled and randomization is easy: e.g. a change in UI of an app.
- Often possible to run A/B tests *more quickly* for *that* reason.

A/B Testing or Split Testing



What if we can't randomize?

Without randomization

We are still trying to find suitable comparison groups (matching), and we'll need more controls (X), and much bigger data to chop up what's driving outcomes.

$$E(Y \mid \text{Treat, time}=1 \mid X) - E(Y \mid \text{Treat, time}=0 \mid X)$$

—

$$E(Y \mid \text{Control, time}=1 \mid X) - E(Y \mid \text{Control, time}=0 \mid X)$$

Questions?

Who can afford this?

- RCT's will cost around \$100 per respondent (all costs considered).
- NGOs that have “succeeded” or scaled but need validation and improvements.
- Observational studies cost much less, but rarely is the data rich enough for matching and controls to establish anything close to causality.

Lean Methodology

- Philosophy born out high risk endeavors that has been adopted in and driven by silicon valley startups
 - Emphasizes learning as quickly as possible
 - Reduced uncertainty
- Build-Measure-Learn

Lean Startup – Why?

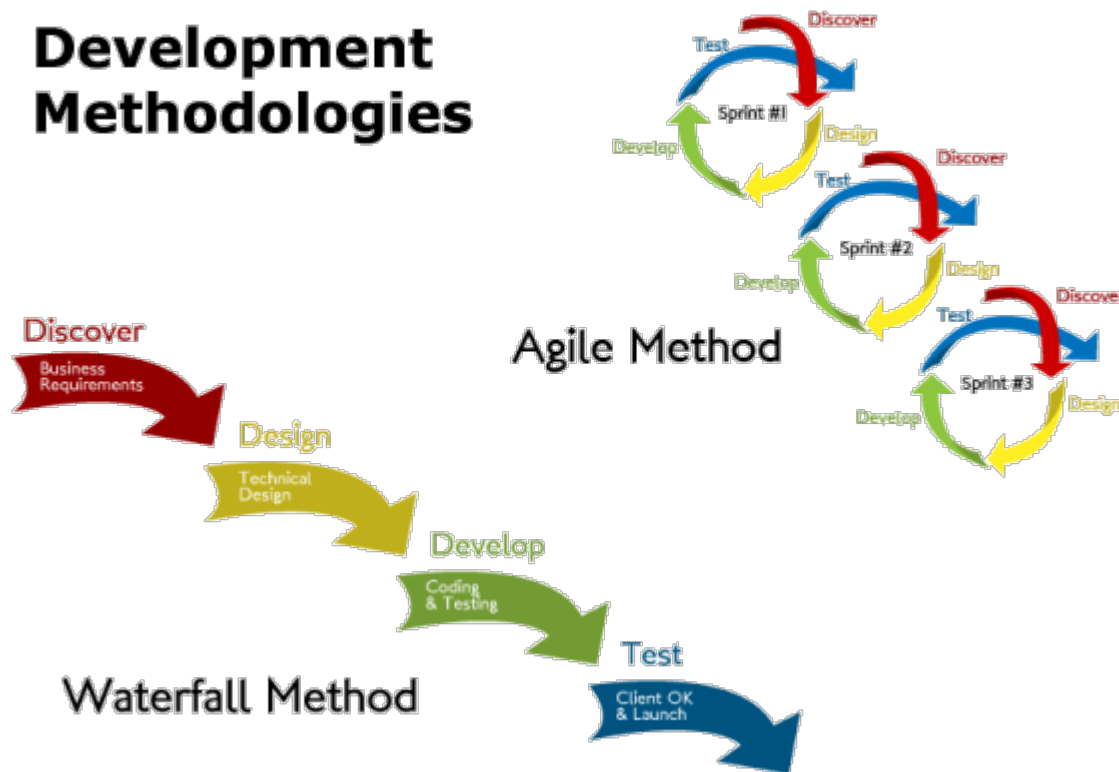
- Startups are not small versions of big companies.
- Master plans do not work well => continuous learning.
- Companies execute business plan, startups look for one.

Lean Startup - Concepts

- Experimentation/rapid iteration > elaborate planning
 - A/B Testing (e.g. UX design)
- Direct feedback > intuition
- Iterative design > design everything first
- Validated learning

Agile Methodologies

Development Methodologies



Lean Startup

- On day 1: list of assumptions
 - Write them down!
- Get out of the building!
- Minimum Viable Product (MVP)
 - E.g. Zappos was a person in a basement mailing back sneakers, before an automated system was ever built out.

Agile Methodologies

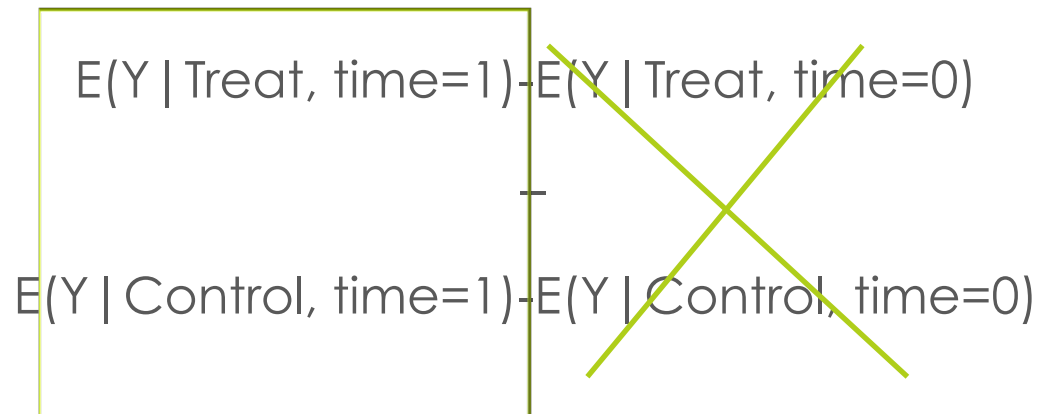
■ SCRUM

- Prioritized wish list with backlog
- Rapid iteration cycles (sprints)
- Daily standups
- ScrumMaster keeps team on task
- Sprint Review

So why not work Agilely always?

Constraints:

- ▣ Agile framework is concerned with testing if enough demand exists to cover costs


$$\begin{array}{l} E(Y \mid \text{Treat, time}=1) - E(Y \mid \text{Treat, time}=0) \\ - \\ E(Y \mid \text{Control, time}=1) - E(Y \mid \text{Control, time}=0) \end{array}$$

- ▣ RCTs are concerning with testing impacts (perhaps even impacts on demand) as long as there is a control group
 - ▣ Randomized control groups are costly
 - ▣ Non-randomized control groups are also costly

Should we just give up?

- Forget about rigor and go for breadth - look at the treatment on the treated across many locations**?
 - i.e. The effect of vitamins on people who choose to take vitamins and remember to take them)
- That's might be ok for software development → the bottom line is profit (but you still want to know *why*)
- For economic development → bottom line is reducing poverty.
 - Most impoverished have the least access to programs, products, and information, and we need a measure of successful reach.

Middle Ground?

- With SMS and IVR data is faster to collect, can make the assignment of an intervention easier to administer

Issues with remote data:

- validation/misreporting/comprehension
- higher attrition
- literacy for SMS
- phone access

Breakout Scenarios

- You are a NGO; you have 50,000 USD to test if a school training program is “effective;” you have 6 months to show if there is an effect.
- The training program trains teachers on best teaching practices. The government of XXX will pay for the program and they’re flexible on where they roll it out first.
- You can get cell phones numbers of the teachers once they are enrolled in the program, but not beforehand.
- Power calculations say you would need to sample about 1,000 schools (2 teachers per school), for a sample of 2,000 teachers.
- You have a list of all the schools and teachers who can be randomly assigned into the study; and then randomly assigned to treatment or control groups.

Breakout Scenario

You have 50,000 USD to spend potentially on:

- Obtaining the control group's contact info
- Calling or texting the teachers
- Sensitizing the teachers regarding the calls
- Verifying the percentage of honest reporting via calls
- And of course, you need to spend time on what questions you will want to ask

Rules for Breakouts

- Choose a ScrumMaster
- Write a list of items that you need: method and frequency of surveying; any on site visits and costs; and technology development
- Order the items
- Design how you will test impact:
 - Can be randomized (e.g. if SMS/IVR and the intervention is information)
 - Can be matching
 - Can be controlling for likely confounders
 - Can be none of the above, but you need to list potential biases below
- Describe a plan and timeline; We're looking for a SCRUM like approach, where you can iterate quickly rather than waiting one year.
- Think about how responses verified, if at all.
- Will you spend your resources on participatory processes and focus group meetings.
- Put a ballpark price tag on your intervention (phone calls, fuel costs, staff)
- Identify the pitfalls
- Bullet your innovations on your board with a 1-minute pitch

End