A/B Testing and Beyond

Designed Experiments for Data Scientists

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Continuing Education Certificate offered by:

UNIVERSITY OF
SAN FRANCISCO

DATA INSTITUTE

UNIVERSITY OF SAN FRANCISCO

Administrative Details

Dates:

- Wednesdays 6:00-9:00pm
- September 6 – October 18, 2017

Location:

- 101 Howard Street, San Francisco, CA

Lecture Style:

- 2 hour lecture + discussion
- 1 hour in-class exercise

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Administrative Details

After Lecture Work:

- Assigned exercises and readings

Communication

- Slack: datainstituteusf.slack.com #abtesting

Course Materials

- Slack/ Github

Week 1

Wednesday September 6th, 2017

Outline

- Preface
  - What this course is
  - What this course isn't
- Introduction
  - Notation and Nomenclature
  - Experiments vs. Observational Studies
  - QPDAC
  - Fundamental Design Principles
  - Exercise

But first, introductions

Introductions

- Name
- Employer
- Role
- Previous experience with statistics
- Previous experience with A/B testing
- What you want to get out of the course

PREFACE

What this course is about

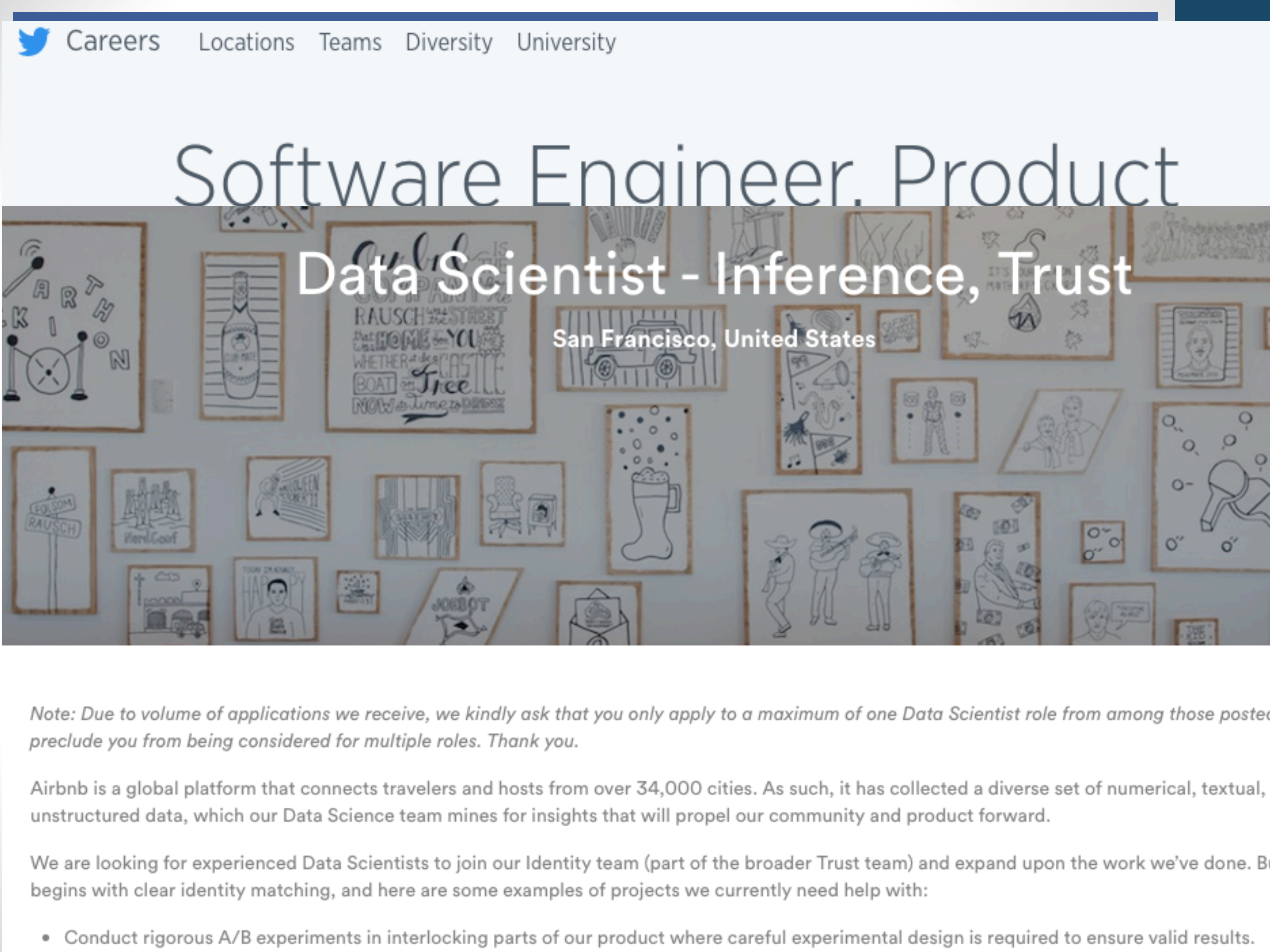
- It is an exciting time to be a data professional – there’s so much data and those with the skills to manage it and draw insight from it are highly valued
- When people think of data science, they tend to think of machine learning
- While machine learning is fascinating and highly useful, it’s uses are limited

What this course is about

- With experiments we can identify and quantify cause-and-effect relationships
- An experiment is a planned investigation in which the influence of one or more variables on an outcome of interest is quantified
- Here data collection is purposeful as opposed to transactional
- It facilitates causal inference

What this course is about

- Experimentation is key to the *Scientific Method* and our understanding of the world around us
- Historically experimentation has been associated with natural sciences and medicine but recently experimentation has proven useful in business and marketing settings
- Designed experiments are beginning to be thought of as a foundational tenet of data science



Software Engineer, Product

Data Scientist - Inference, Trust

San Francisco, United States

Note: Due to volume of applications we receive, we kindly ask that you only apply to a maximum of one Data Scientist role from among those postec preclude you from being considered for multiple roles. Thank you.

Airbnb is a global platform that connects travelers and hosts from over 34,000 cities. As such, it has collected a diverse set of numerical, textual, unstructured data, which our Data Science team mines for insights that will propel our community and product forward.

We are looking for experienced Data Scientists to join our Identity team (part of the broader Trust team) and expand upon the work we've done. B begins with clear identity matching, and here are some examples of projects we currently need help with:

- Conduct rigorous A/B experiments in interlocking parts of our product where careful experimental design is required to ensure valid results.

What this course is about

- This course exposes you to the value experimentation and provides a thorough treatment of available methods and best practices
  - A/B/n testing in which two or more variants are compared
  - Multivariate experiments such as factorial and fractional-factorial designs
  - Optimization techniques such as multi-armed bandit experiments and response surface methodology

What this course is about

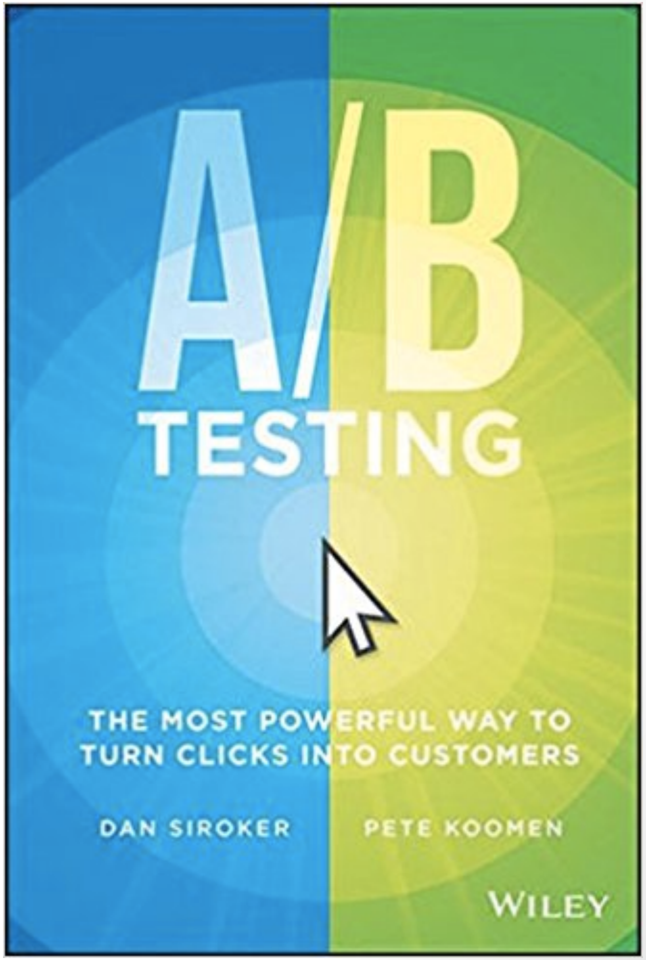
- We emphasize the statistical principles and practical considerations that underlie effective experimentation
- We learn to carefully navigate the choices and nuances associated with the design of an experiment
- We discuss relevant hypothesis tests, power analyses, sample size calculations and analysis methods necessary to draw conclusions and make impactful statements about the question of interest

What this course isn’t about

- Third party experimentation platforms such as:
  - Optimizely
  - Google Analytics
  - Wasabi
  - Apptimize
  - VWO
- The engineering skills required to build your own experimentation platform

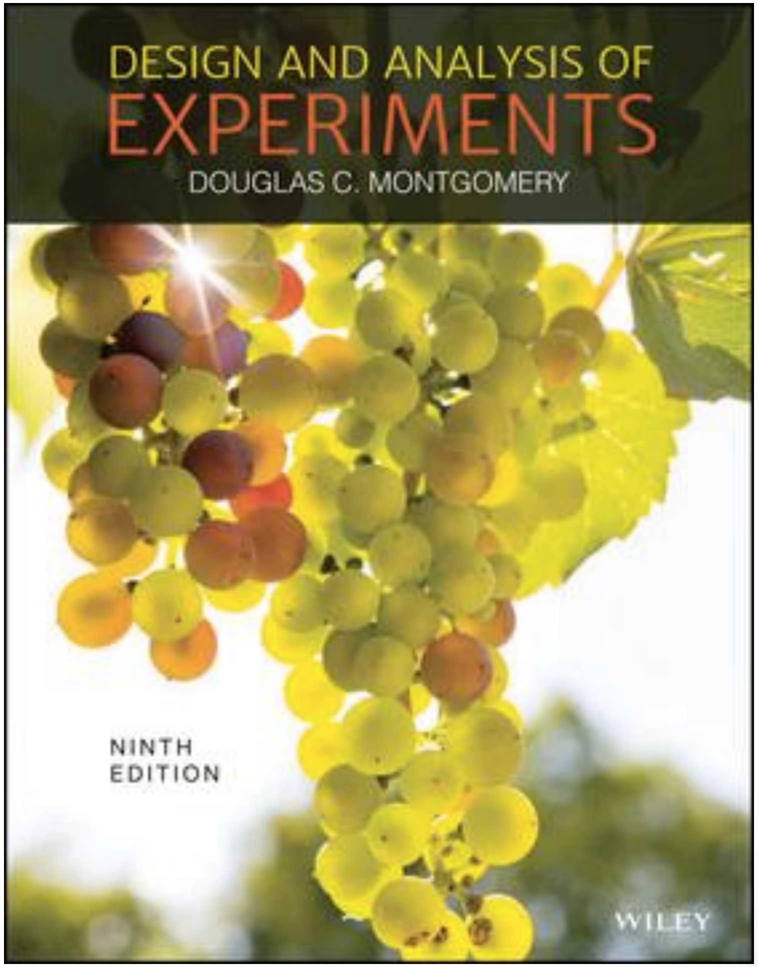
Helpful Resources

- *"A/B Testing: The Most Powerful Way to Turn Clicks into Customers"* by Dan Siroker and Pete Kooman



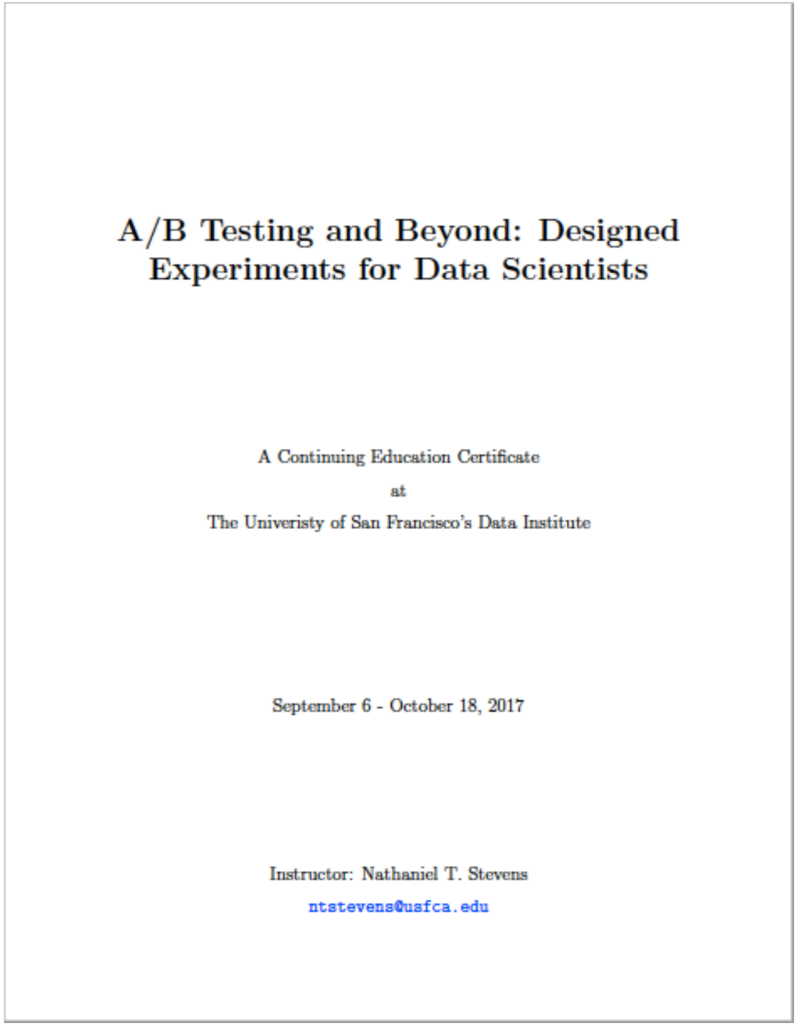
Helpful Resources

- "*Design and Analysis of Experiments*" by Douglas C. Montgomery



Helpful Resources

- My notes!



A/B Testing and Beyond: Designed Experiments for Data Scientists

A Continuing Education Certificate

at

The Univeristy of San Francisco's Data Institute

September 6 - October 18, 2017

Instructor: Nathaniel T. Stevens

ntstevens@usfca.edu

INTRODUCTION

Notation and Nomenclature

Example 1: Button Message

- Interest lies in comparing different messages on a call-to-action button to find which message maximizes the click through rate (CTR).

Example 2: Webpage Design

- Interest lies in the comparison of different webpage designs to decide which one minimizes that page's bounce rate.

Notation and Nomenclature

- Response Variable: the variable we are primarily interested in, denoted by y
- The question of interest will be defined in terms of this variable
- Common choices for response variables are KPIs such as
  - Conversion rate
  - Revenue per visitor
  - Session duration
  - Bounce rate

Notation and Nomenclature

Example 1: Button Message

- Response variable = click through rate
- Want to maximize this

Example 2: Webpage Design

- Response variable = bounce rate
- Want to minimize this

Notation and Nomenclature

- Explanatory Variables: are variables that we expect may influence the response variable
- We tend to think of these variables as having secondary importance relative to the response
- In the context of experiments we call these variables factors and denoted them by x
- The different values that a factor takes on in an experiment are referred to as levels

Notation and Nomenclature

Example 1: Button Message

- Factor = button message
- Levels = {“Submit”, “Go”, “Let’s Go!”}

Example 2: Webpage Design

- Factor = webpage design
- Levels = {photo, carousel}

Notation and Nomenclature

- Experimental conditions are defined by unique combinations of the levels of one or more factors
- Experimental units are then assigned to each condition and their response values are observed
- These response values are compared across conditions in order to find an optimal one

Experiments vs. Observ. Studies

- An experiment is composed of a series of conditions defined by purposeful changes to one or more factors
- The goal is to evaluate the change in response elicited by a change in the factors
- In order to truly understand this relationship it would be ideal to observe how a given set of units responds to each of the conditions

Experiments vs. Observ. Studies

Unit	Condition 1	Condition 2
1	y_{11}	y_{21}
2	y_{12}	y_{22}
3	y_{13}	y_{23}
$\vdots$	$\vdots$	$\vdots$
n	y_{1n}	y_{2n}

Experiments vs. Observ. Studies

- However, a given set of units can't be exposed to *just one* experimental condition
- Their unobserved response in the conditions to which they were not assigned represents a counterfactual
- Because counterfactuals are unobservable we instead assign different groups of units to the different conditions, but we try to make these groups as homogenous as possible

Experiments vs. Observ. Studies

Unit	Condition 1	Condition 2
1	y_{11}	×
2	y_{12}	×
3	y_{13}	×
⋮	⋮	⋮
n_1	y_{1n}	×
1	×	y_{21}
2	×	y_{22}
3	×	y_{23}
⋮	⋮	⋮
n_2	×	y_{2n}

Experiments vs. Observ. Studies

- When this is the case it is reasonable to believe that the only difference between the units in each condition is the fact that they are in different conditions

- Thus, if there is a marked difference in the response between the conditions, then this difference can be attributed to the conditions themselves and we conclude that the observed difference in response values was caused by the condition they were in

Experiments vs. Observ. Studies

- Different from an experiment, an observational study has no measure of control in the data collection process
- Data are recorded passively and any relationship between the response and factors is observed organically
- Unfortunately, with data arising in this manner, it is difficult to perform causal inference

Experiments vs. Observ. Studies

Experiments: Advantages

- Causal inference is straight forward

Experiments: Disadvantages

- May be risky or costly
- May be unethical

Experiments vs. Observ. Studies

Observational Studies: Advantages

- Less risk inherent
- Ethical

Observational Studies: Disadvantages

- Causal inference is more difficult

QPDAC

- QPDAC is a general framework for planning and executing a data-driven investigation and is a particularly useful approach to designing and analyzing experiments
- QPDAC is an acronym that stands for
  - Question
  - Plan
  - Data
  - Analysis
  - Conclusion

QPDAC

QUESTION: Develop a clear statement of the question that needs to be answered.

- Question statement should address some hypothesis that you'd like to prove or disprove with the experiment
- This statement should be clear, concise and quantifiable
- Everyone involved in the experiment should be aware of the question of interest and hence the goal of the experiment

QPDAC

PLAN: Address all relevant pre-experimental questions and design the experiment

- Choose your response variable and factors
- Distinguish design factors from allowed-to-vary and nuisance factors
- Choose the levels of the design factors and define the experimental conditions
- Define the experimental units and decide how they will be assigned to the conditions and how many will be assigned to each condition

QPDAC

DATA: Collect the data according to the plan

- It is extremely important that this step be done correctly
- If the data quality is compromised, the resulting analysis may be invalid in which case any conclusions drawn will be irrelevant.
- It’s good to check, using an A/A test, that randomization is working properly

QPDAC

ANALYSIS: Analyze the data in order to draw objective conclusions about the question

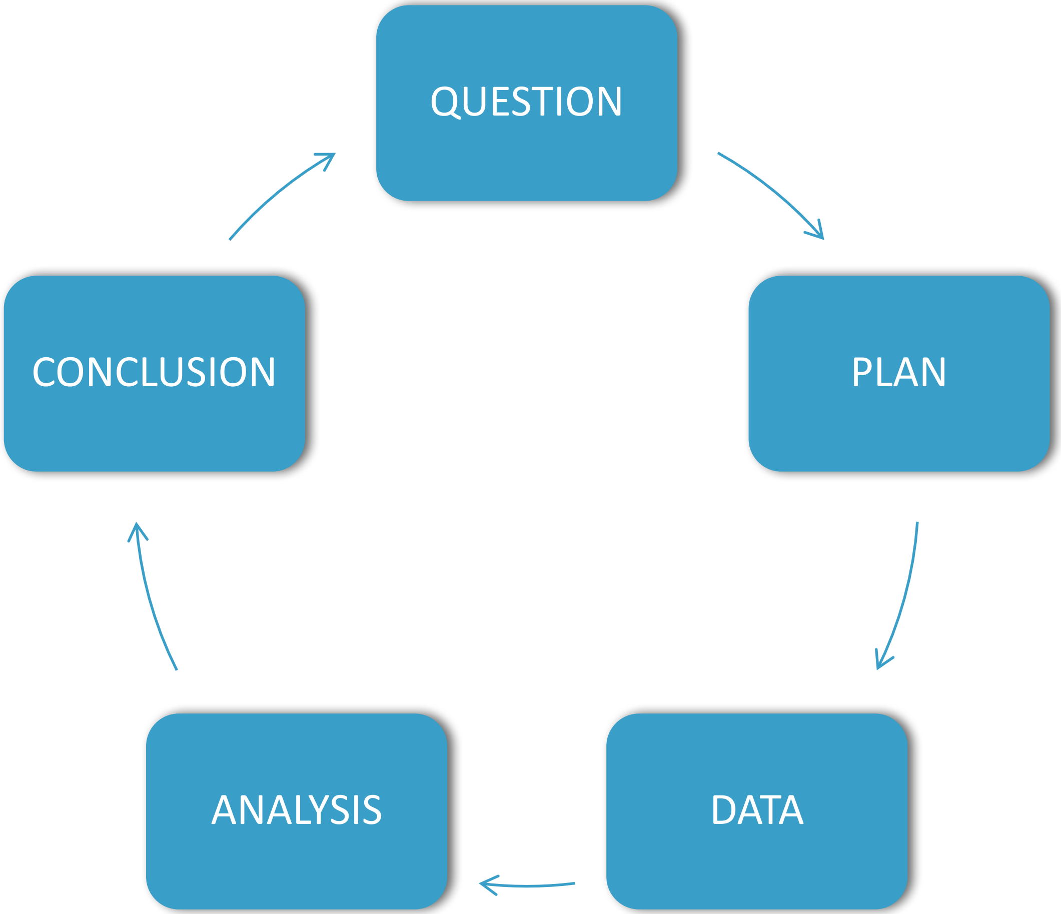
- If the data was collected correctly and the experiment was well defined, this stage should be straightforward
- This typically involves estimating parameters, fitting models, and carrying out statistical tests of hypotheses

QPDAC

CONCLUSION: Draw conclusions based on your analysis and communicate what you have learned about the question

- Conclusions should be clearly communicated to all parties involved in - or impacted by - the experiment
- Doing so will help to foster a culture within your organization that highly values experimentation

QPDAC



Design Principles

Randomization refers both to the manner in which experimental units are selected for inclusion in the experiment and the manner in which they are assigned to experimental conditions.

- The first level of randomization exists to ensure the sample of units included in the experiment is representative of those that were not
- The second level of randomization exists to balance out the effects of extraneous variables not under study

Design Principles

Replication corresponds to when more than one unit is assigned to each condition, which provides assurance that the observed results are genuine and not just due to chance

- As the number of units in each condition increases we can become increasingly sure of the results we observe
- How much replication is necessary?
- How long does the experiment need to run for?
- Power analyses and sample size calculations answer these questions

Design Principles

Blocking is the mechanism by which nuisance factors are controlled for.

- We want the only source of variation in response values to be the experimental conditions themselves
- So we run the experiment at fixed levels of the nuisance factors, so as not to conflate their influence with the influence of the experimental conditions
- Example: Email promotions

Exercise

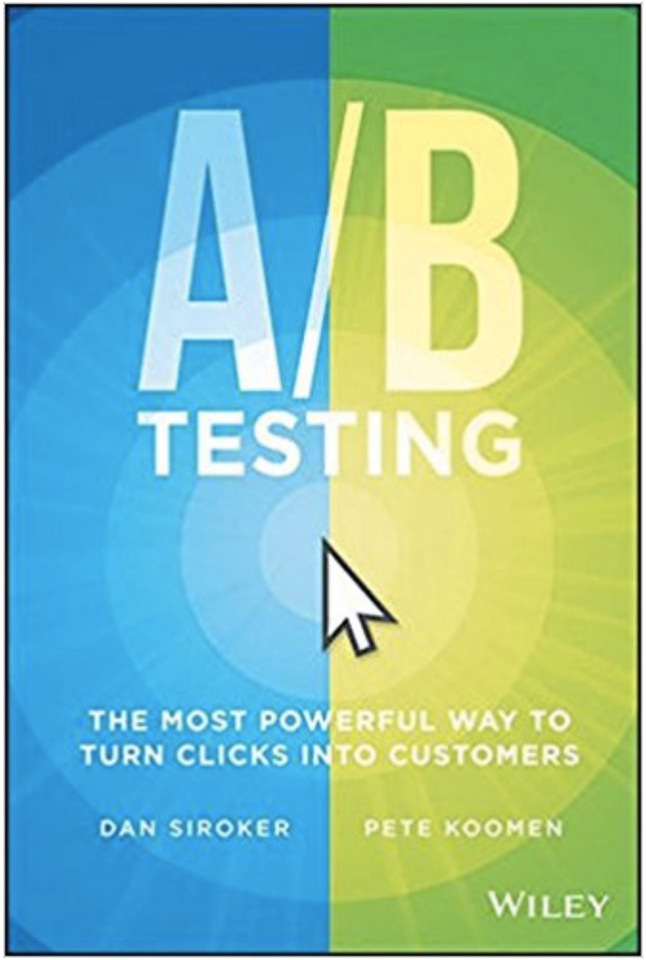
The Instagram Experiment

While sponsored ads serve as a source of revenue for Instagram, they also serve as a source of frustration and annoyance to users. Thus, we would like to run an experiment to gain insight into the interplay between ad revenue, user engagement and factors such as ad frequency, ad type (photo/video), whether the ad's content is targeted or not, etc. Ultimately the goal is to identify a condition that maximizes ad revenue without simultaneously plummeting user engagement below some minimally acceptable threshold.

How would you design such an experiment?

Take Home Task

Read this book!



See you next week!