A/B Testing, Theory, practice and pitfalls

- Srimugunthan ,
Data scientist at Verizon AI&D

Brief about me..

- My name is Srimugunthan
- Data Scientist @ Verizon AI&D
- 12+ years of experience with 8 years of experience in Big data analytics
- Interest in Experimentation and causal inference, Recommender systems, reinforcement learning

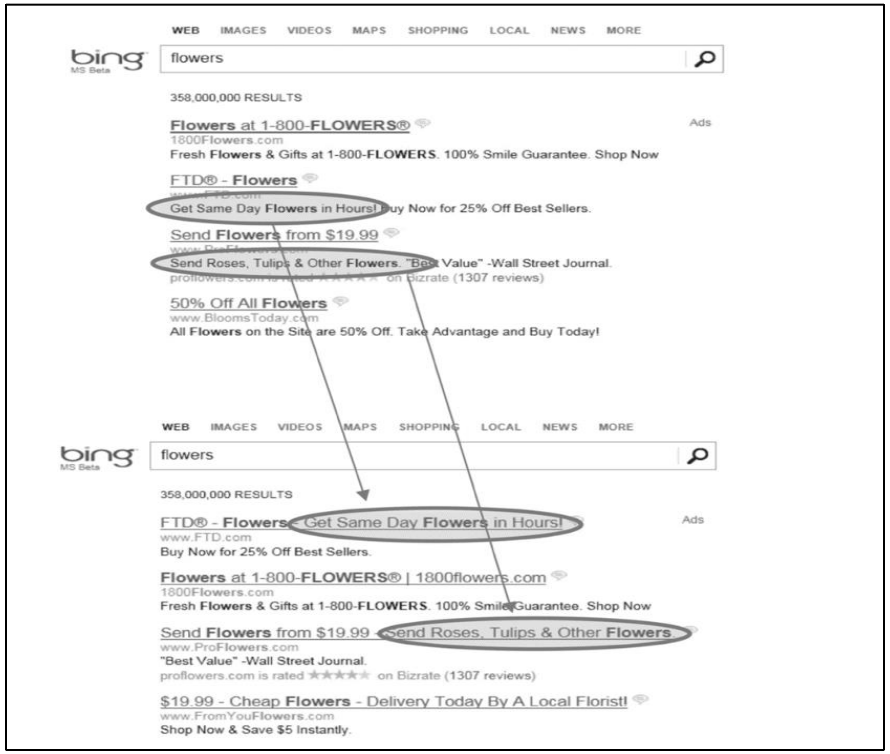
Agenda

- Introduction to Design of experiments
- Steps to A/B Test Design
- Demo and Example of A/B testing
- A/B testing Pitfalls

Design of experiments – Introduction

Motivation: Why experiments?

A simple UI experiment in Microsoft Bing that resulted in 100M$/Y



- Randomized control trials originally prevalent with research is now used for Marketing campaigns. Conversion optimization, Startups, drug trials etc
- Google finds 10% of these experiments have positive results
- “Given a 10 percent chance of a 100 times payoff you should take that bet every time” – from Amazon SEC filing

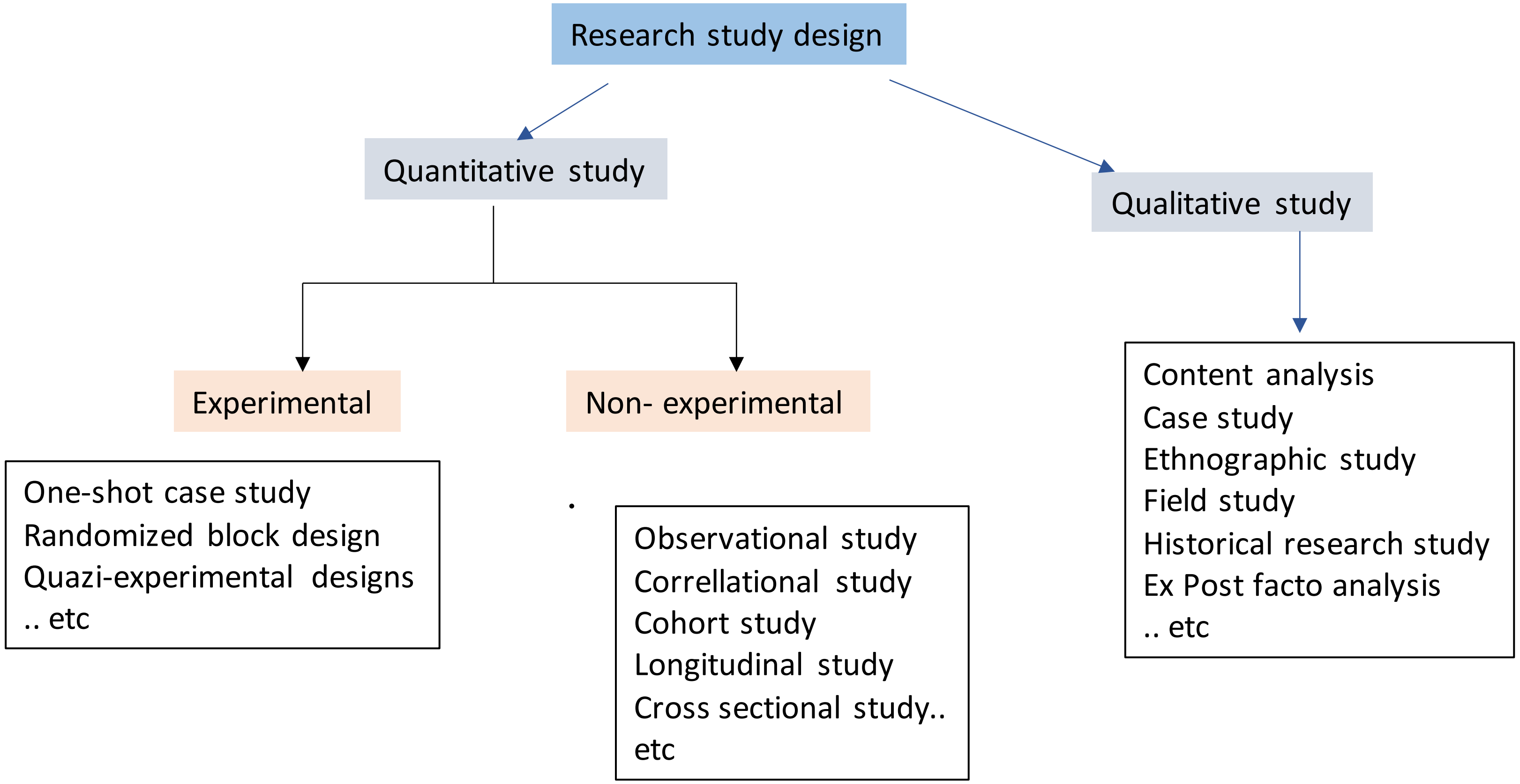
- 10000 experiment rule of scaled experimentation Amazon, Booking.com, Facebook, and Google—each conduct more than 10,000 online controlled experiments annually

https://hbr.org/2017/09/the-surprising-power-of-online-experiments

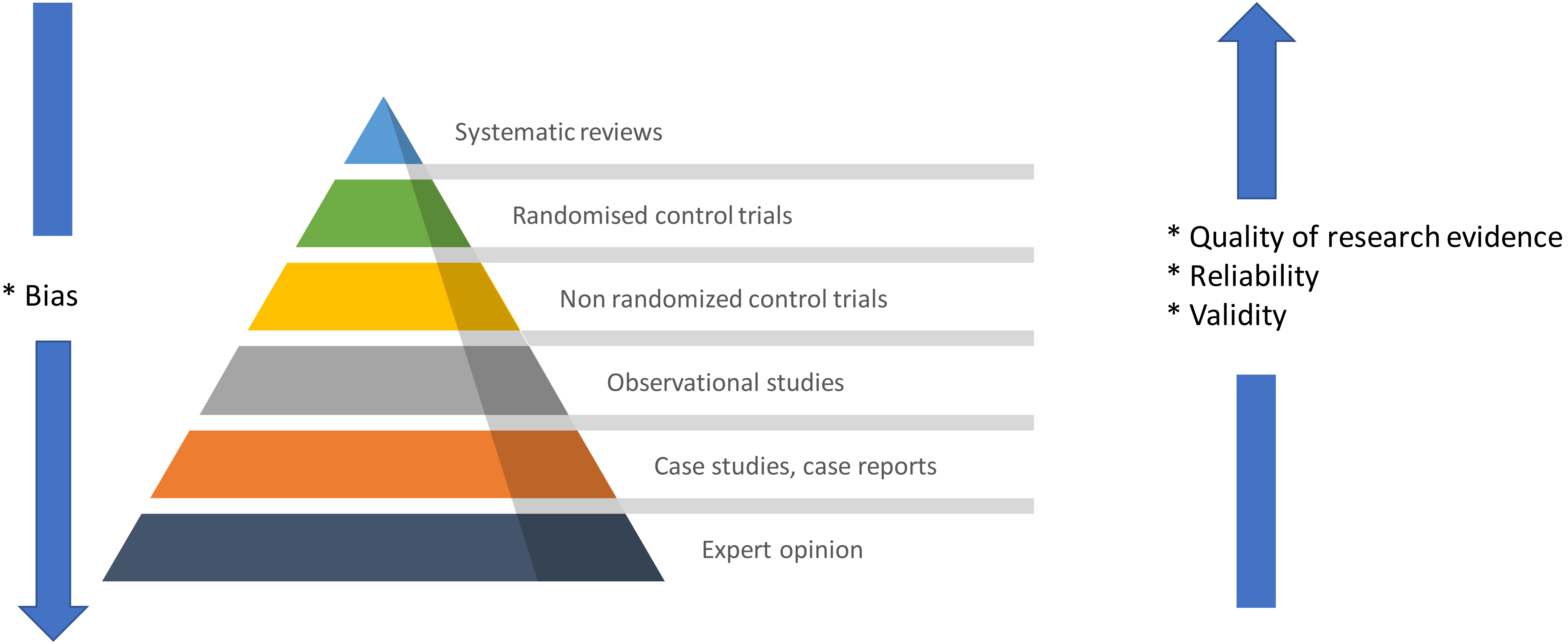
Book, “Trustworthy online controlled experiments” –Rohnny kohavi

Research Study design

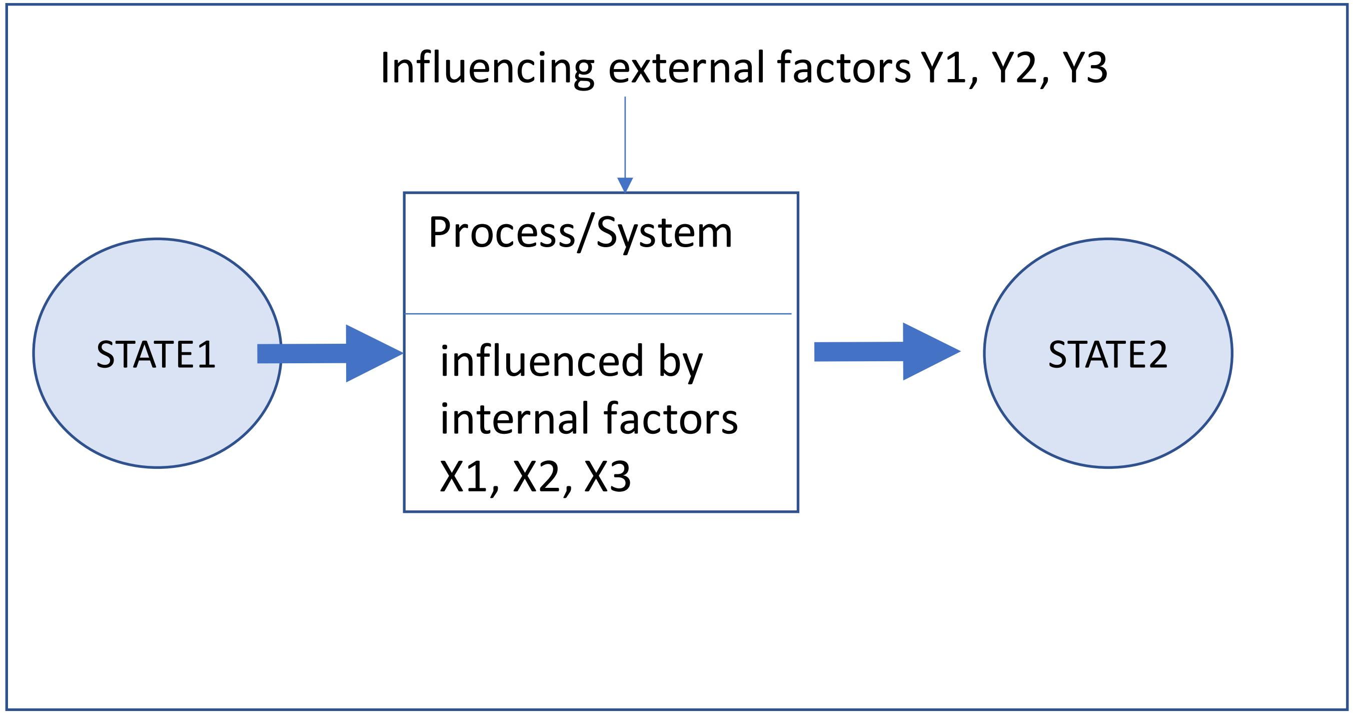
- how researcher goes about finding answer to the research question



Evidence pyramid



Experiment and causal inference



Causal inference: State2 is caused from state1 because of factors X1, X2..

Example: Increase in agricultural yield is caused by the use of nitrogen fertilizer

- What is an experiment?
  - the process of examining the truth of a hypothesis, relating to a research question.
- Origins of Experiment design:
  - Pioneered by Fischer in 1920s when studying agricultural yield
- **Experimental design**
  - Error of an experiment can be reduced by careful design
  - Good experimental design Increases reliability of the experiment
- **Co-variate: Confounding/extraneous variables**
- **Internal validity**: ability to strictly control for confounding variables and avoid selection bias.
- **External validity**: Applicability of inference to population and time period different from the experiment

Experimental design : Principles

Key question: "Did the treatment exert a causal effect?"

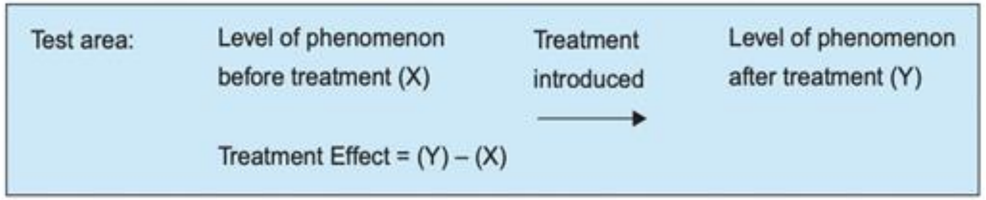
Three principles of experiment design

1. Principle of Replication
Treatment is applied to many experimental units

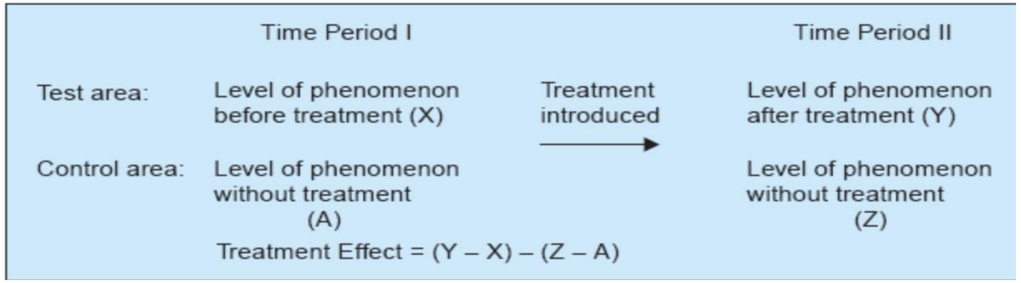
2. Principle of Randomization
the allocation of treatment to experimental units at random to avoid any bias in the experiment . Errors are independently distributed

3. Principle of local control or blocking
arrangement of experimental units into groups (blocks) consisting of units that are similar to one another

Different possible experimental designs



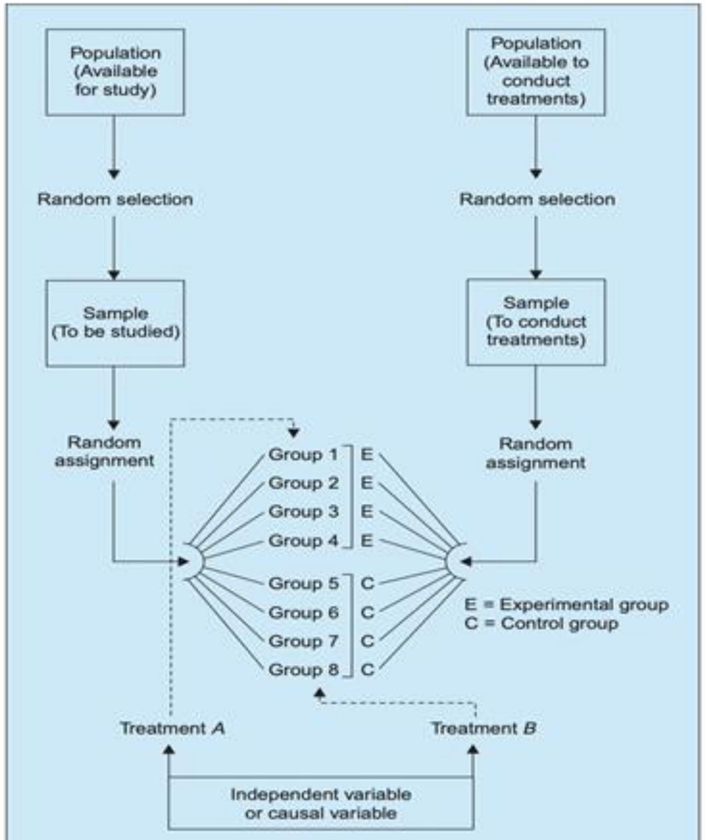
Before and After without control



Before and After with control

	Very low I.Q.	Low I.Q.	Average I.Q.	High I.Q.	Very high I.Q.
	Student A	Student B	Student C	Student D	Student E
Form 1	82	67	57	71	73
Form 2	90	68	54	70	81
Form 3	86	73	51	69	84
Form 4	93	77	60	65	71

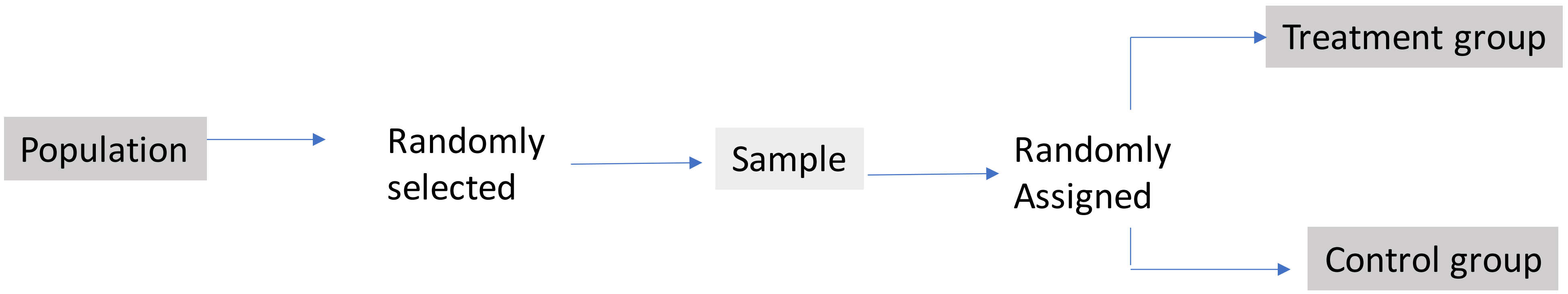
Randomized block design



Random replicated design

https://sites.google.com/site/experimentaldesignandanaly/what-is-experiment/important-experimental-designs

Randomized control design



- A/B testing popularity due to simplicity
- Control group: **minimizing the effect of all variables except the impact of the variable in the treatment**.
- Stable Unit Treatment Value Assumption, SUTVA. **It states that the treatment and control units don't interact with each other; otherwise, the interference leads to biased estimates**
- High internal validity: Achieves co-variate balance in average
- Sometimes Extended with blocking

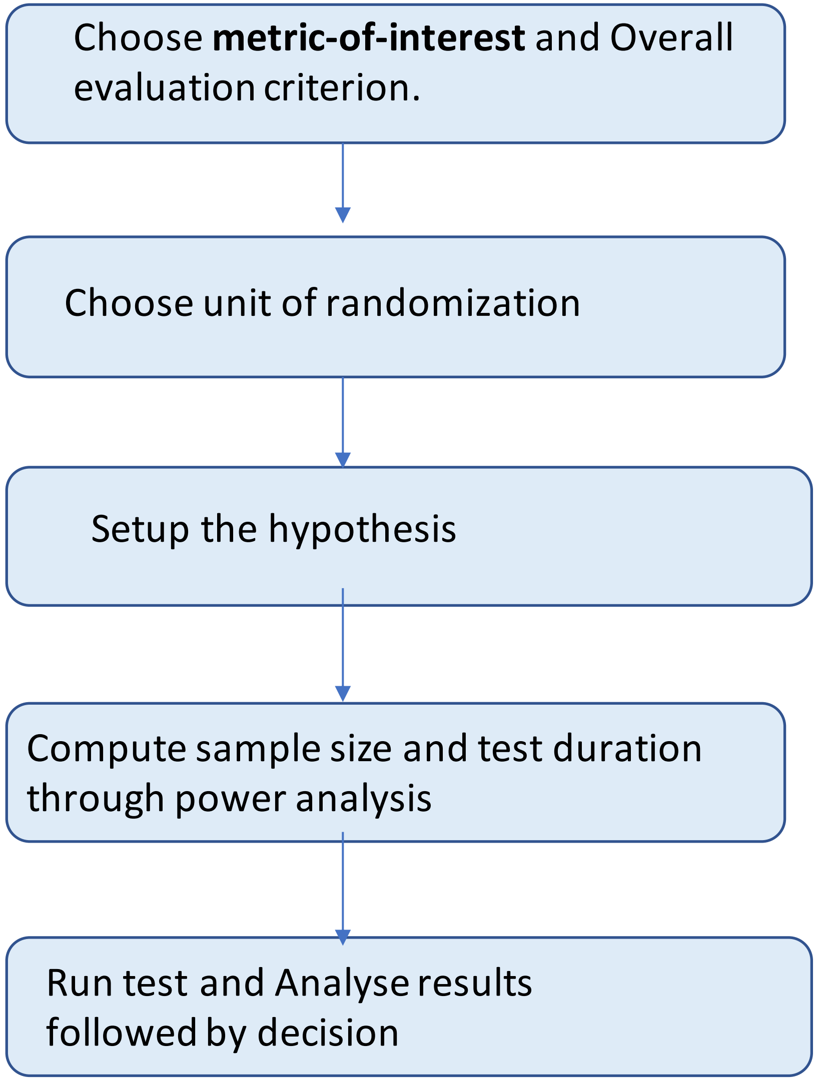
https://towardsdatascience.com/randomization-blocking-and-re-randomization-f0e3ab4d79ca

Steps to Designing A/B test

Designing A/B Test

Experiment design Questions:

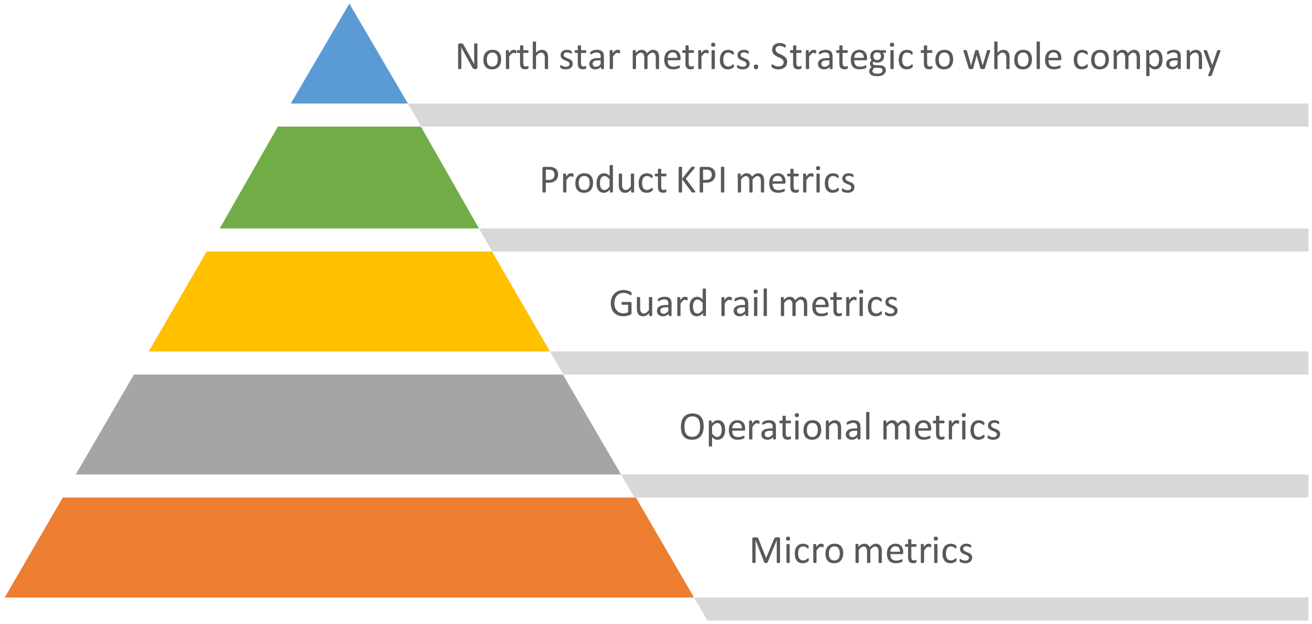
- What are the metrics you want to optimize? What is the overall evaluation criterion?
- What is your hypothesis?
- How safe and ethical is the experiment?
- What is the population you want to target against?
- What is the unit of randomization ?
- How may variants you might need?
- Are there any interference between control and treatment group?
- How large will the experiment?
- How long do we have to run the experiment?
- How do you split the traffic? Does this experiment need to share traffic with other experiments?
- How are you accounting for novelty effect of a change?
- When do you run the experiment?



Step1: Overall evaluation criterion

- **What are the success metrics?**
  - Different types of metrics
    - Engagement metrics
    - Guardrail metrics
    - Invariant metrics

- Combine metrics to Overall Evaluation criterion



Step2: choose randomization unit

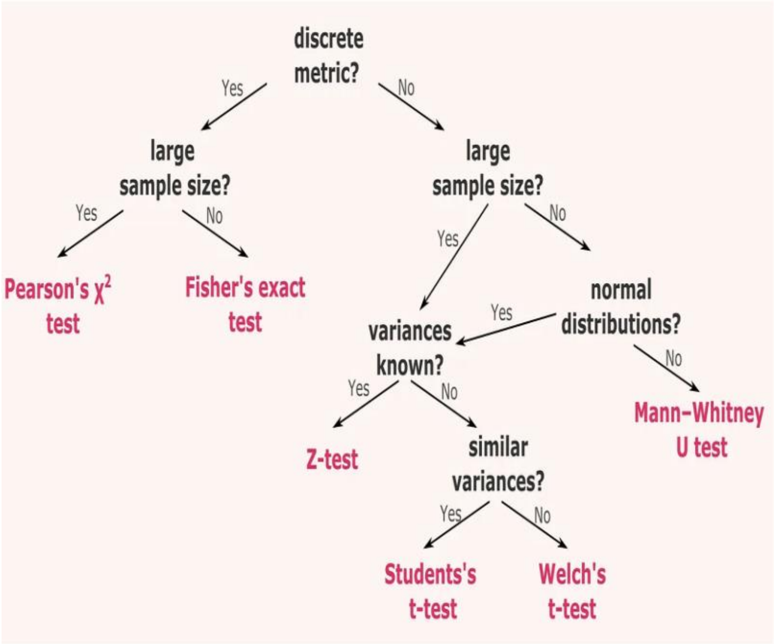
- **Choice of randomization unit**
  - Page level
  - Session level
  - User level

- Randomisation unit same or coarser than unit required for metrics
  - User level randomisation for metric like revenue per user

Step3: Setup the hypothesis

- **Set the null and alternate hypothesis**
  - Null hypothesis: change has no effect
  - Alternate hypothesis: change has an effect
- Decide on parameters
  - Minimum detectable effect
  - Statistical significance level

- Pick a statistical test . The choice of test depends on
  - Probability density function of Metric-of-choice
  - Sample size
  - Nature of the hypothesis



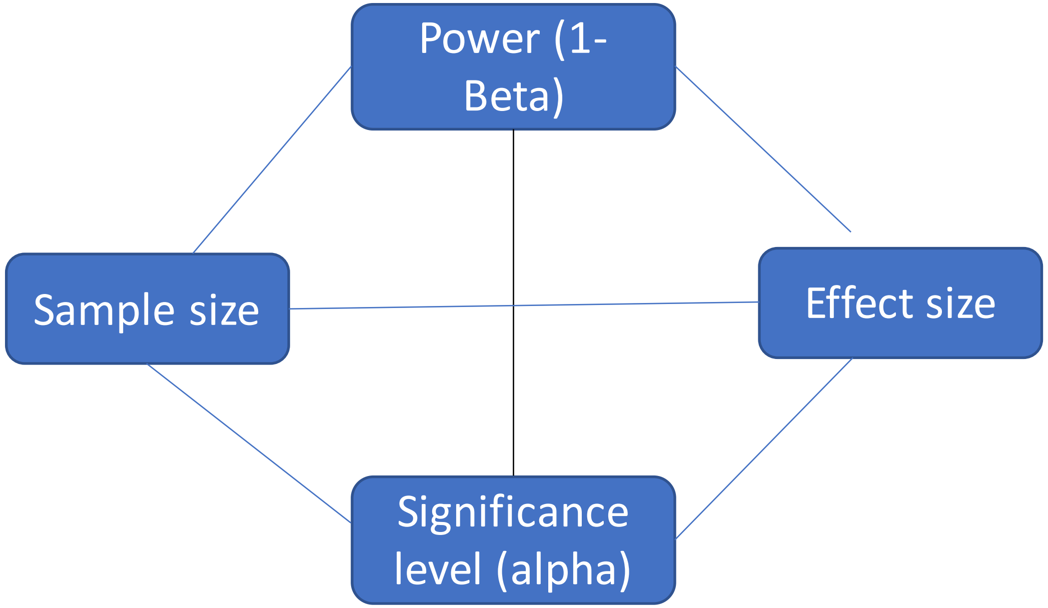
https://towardsdatascience.com/a-b-testing-a-complete-guide-to-statistical-testing-e3f1db140499

https://towardsdatascience.com/simple-and-complet-guide-to-a-b-testing-c34154d0ce5a

Step4: Power analysis

Power analysis: determine the sample size to achieve a certain statistical power

- Type1 error: concluding significant difference between treatment and control when actually there is no difference
- Type2 error: concluding no significant difference when actually there is
- Statistical power, or the power of a hypothesis test is the probability that the test correctly rejects the null hypothesis. The higher the statistical power for a given experiment, the lower the probability of making a Type II (false negative) error.
- **Low Statistical Power**: Large risk of committing Type II errors, e.g. a false negative.
- **High Statistical Power**: Small risk of committing Type II errors.

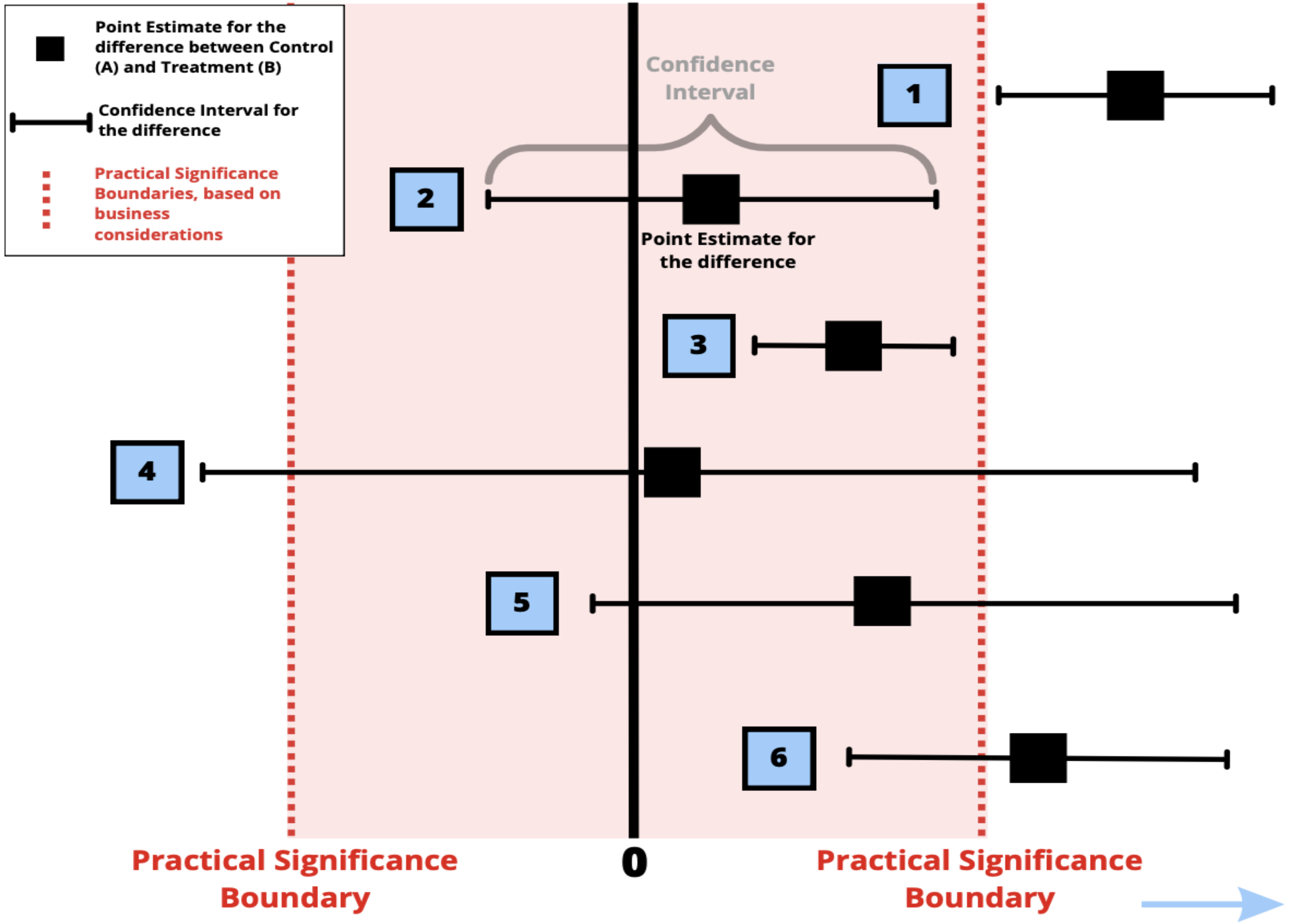


Sample Size. The number of observations in the sample.

Significance level Alpha value (α) - . The significance level used in the statistical test, e.g. alpha. Often set to 5% or 0.05.

Effect. The difference in OECs for the variants, i.e. the mean of the Treatment minus the mean of the Control. How big of a difference we expect there to be between the Target metric

Step5: Go/No-go Decision



1 LAUNCH

2 Change has no impact. Iterate or abandon

3 No practical significance. No launch

4 No statistical significance. Test again

5 Repeat test with greater power

6 Repeat test with greater power, but can launch

Compare Confidence Interval and compare its lower bound to the MDE .
if the lower bound of CI is larger than the MDE (delta), then you can state that you have a practical significance. For example, if the CI = [5%, 7.5%] and the MDE = 3% then you can conclude to have a practical significance since 5% > 3%.

Demo and example of A/B test

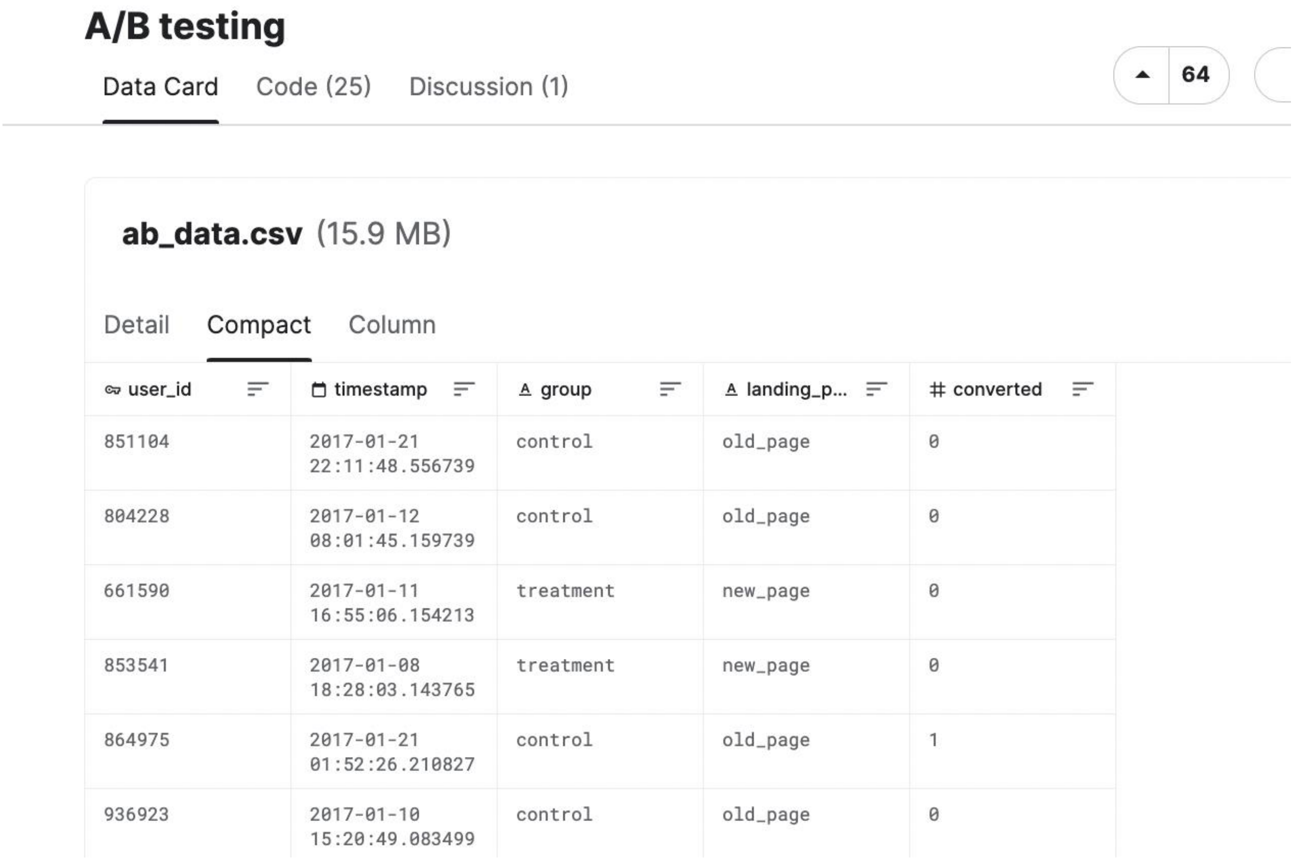
A/B test example: conversion on landing page

Assume a medium-sized ***online e-commerce business****.*

A new version of the landing page is available which promises higher conversion rate

The ***current conversion rate*** *is about* ***13%*** *on average throughout the year*

an ***increase of 2% is targeted***



A/B testing

Data Card Code (25) Discussion (1) 64

ab_data.csv (15.9 MB)

Detail Compact Column

user_id	timestamp	group	landing_p...	converted
851104	2017-01-21 22:11:48.556739	control	old_page	0
804228	2017-01-12 08:01:45.159739	control	old_page	0
661590	2017-01-11 16:55:06.154213	treatment	new_page	0
853541	2017-01-08 18:28:03.143765	treatment	new_page	0
864975	2017-01-21 01:52:26.210827	control	old_page	1
936923	2017-01-10 15:20:49.083499	control	old_page	0

https://www.kaggle.com/datasets/zhangluyuan/ab-testing
https://towardsdatascience.com/ab-testing-with-python-e5964dd66143

Demo

A/B test pitfalls

A/B testing Limitations and Pitfalls

PITFALLS

1. A/A Testing Ignored
2. Peeking: Continuous Monitoring of the experiment in fixed sample size testing
3. Random Split on the entire user base is not truly random on power users
4. Some users exposed to both treatment /control
5. Contamination / Spillover effect b/w control and treatment
6. Multiple Testing Problem
7. Simpson's paradox - A/B test on a subset population shows a different trend post full release
8. Novelty and Primacy Effect Ignored
9. Survivorship bias and Selection bias

LIMITATIONS

Not possible to conduct an A/B test if

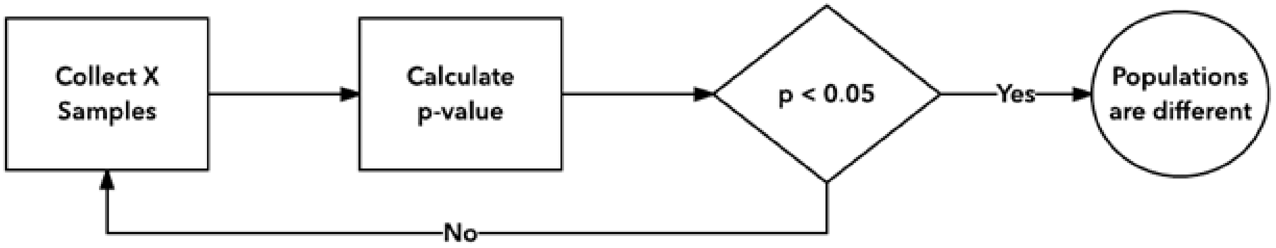
- It is not possible to randomize
- Not enough time to run the experiment
- Not enough samples
- No control

A/B testing pitfalls: https://www.linkedin.com/pulse/ab-testing-netflix-uber-pinterest-google-linkedin-spotify-uppal/

Pitfall1: Early stopping/Peeking

- ▷ Continuously monitor p-value until significance threshold

- ▷ Solution: commit to the experiment parameters calculated before test.



Continuing one more iteration as p value is slightly close to 0.05
Stopping the experiment when p value less than 0.05

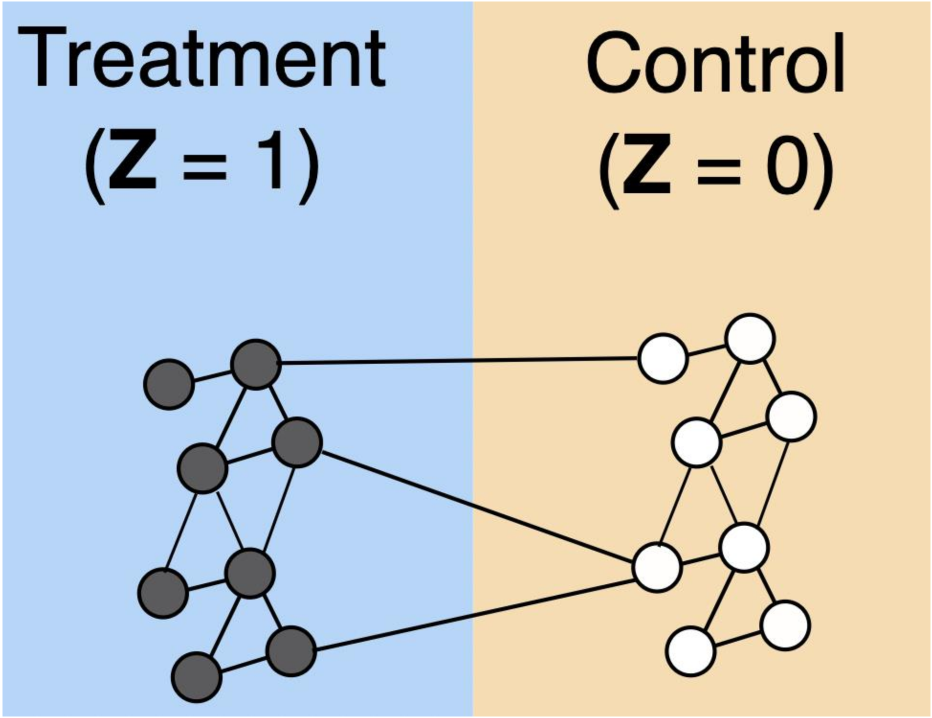
https://www.lucidchart.com/blog/the-fatal-flaw-of-ab-tests-peeking

Pitfall2: Interference/Spillover effect

- **Violation of SUTVA. (Stable Unit Treatment Value Assumption):** one unit's action does not interfere with another one.

- Interference of treatment and control

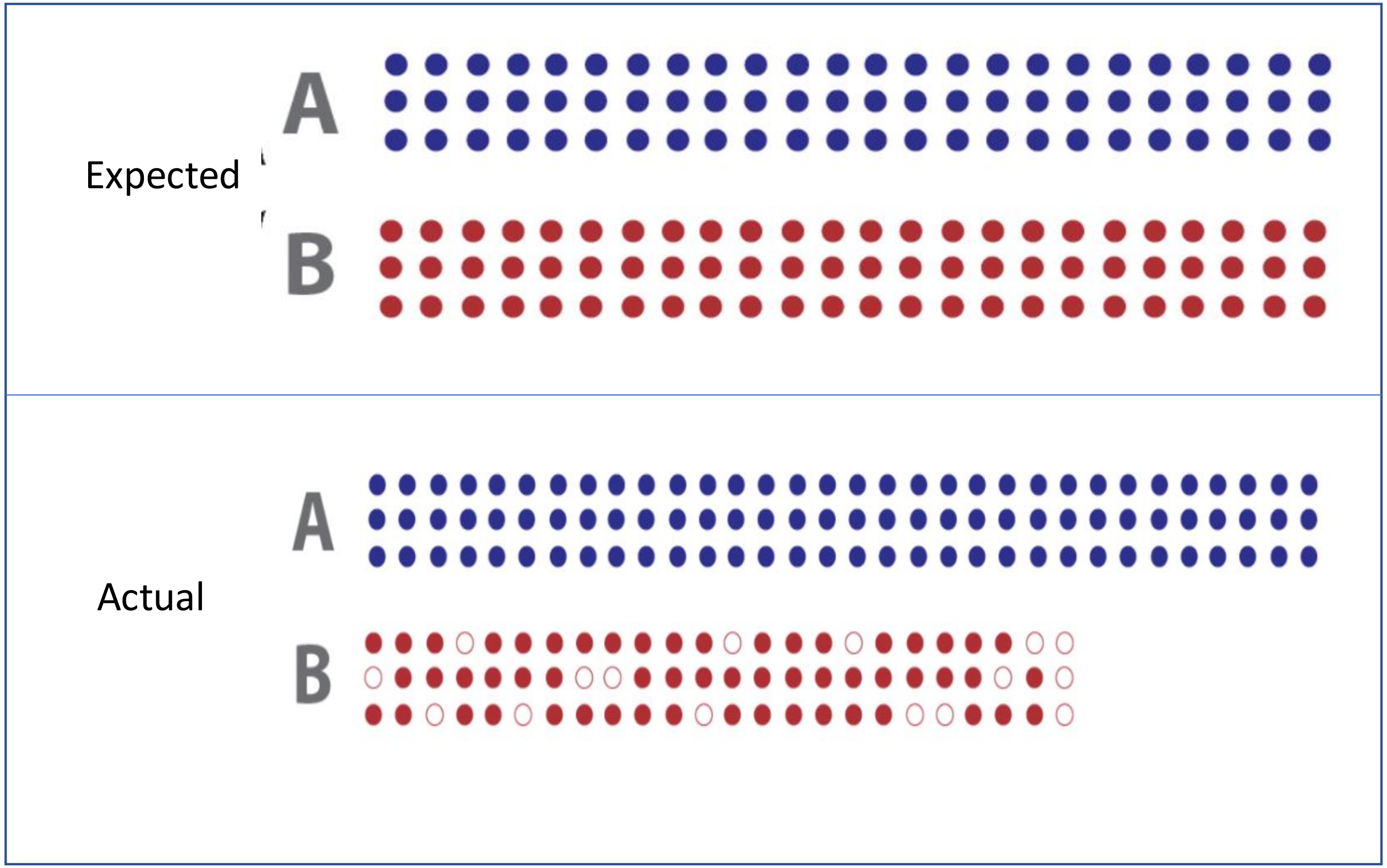
- Example: A new messaging UI with a target metric of Time-spent on chatting



https://jean.pouget-abadie.com/sutva.html
https://medium.com/analytics-vidhya/network-effects-a95c42c10ff2

Pitfall3: Sample ratio mismatch

- Actual traffic allocation != Expected traffic allocation
- Caused due to any reason:
  - Poor randomization
  - Buggy implementation
  - User behavior and time-of-day effect



https://towardsdatascience.com/the-essential-guide-to-sample-ratio-mismatch-for-your-a-b-tests-96a4db81d7a4

Pitfall4: Simpson's paradox

- ***Simpson's Paradox*** *is a statistical phenomenon that a trend appears in the combined data, but disappears or reverses when the data are partitioned into several different groups*
- *Example:*

	Page A		Page B		Page A Conversion Rate	Page B Conversion Rate
	Visits	Conversions	Visits	Conversions		
Aggregate	1,490,000	25,000	510,000	6,230	1.68%	1.22%
Friday (C/T Split: 99%/1%)	990,000	20,000	10,000	230	2.02%	2.30%
Saturday (C/T Split: 50%/50%)	500,000	5,000	500,000	6,000	1.00%	1.20%

Simpson's Paradox due to change in traffic allocation between experiences

https://medium.com/swlh/how-simpsons-paradox-could-impact-a-b-tests-4d00a95b989b

Pitfall5: Current base not reflecting target

- Music sales product (not established).
- Impressions and purchases
- A/b test : how track details are displayed. should it be [Artist Title BPM] or [BPM Artist Title]

Treatment	Impressions	Sales	Conversion Rate	Delta From Control
[Artist Title] (control)	10000	400	4.00±0.38%	-
[Artist Title BPM]	10000	500	5.00±0.43%	+1.00±0.57%
[BPM Artist Title]	10000	600	6.00±0.47%	+2.00±0.60%

EDM users (8,000 impressions):

Treatment	Impressions	Sales	Conversion Rate	Delta From Control
[Artist Title] (control)	8000	320	4.00±0.43%	-
[Artist Title BPM]	8000	440	5.50±0.50%	+1.50±0.66%
[BPM Artist Title]	8000	570	7.12±0.56%	+3.12±0.71%

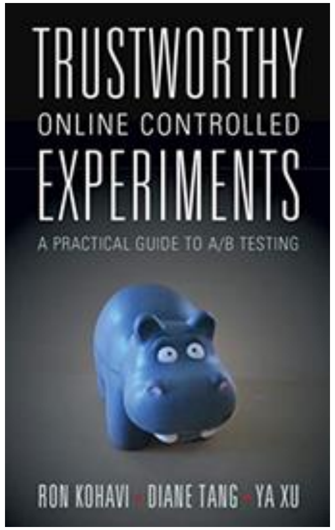
Non-EDM users (2,000 impressions):

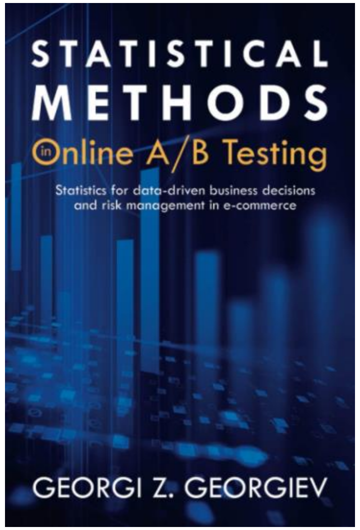
Treatment	Impressions	Sales	Conversion Rate	Delta From Control
[Artist Title] (control)	2000	80	4.00±0.86%	-
[Artist Title BPM]	2000	60	3.00±0.75%	-1.00±1.14%
[BPM Artist Title]	2000	30	1.50±0.53%	-2.50±1.01%

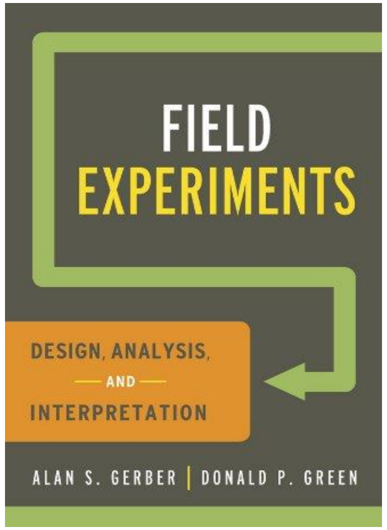
https://www.unofficialgoogledatascience.com/2019/04/misadventures-in-experiments-for-growth.html

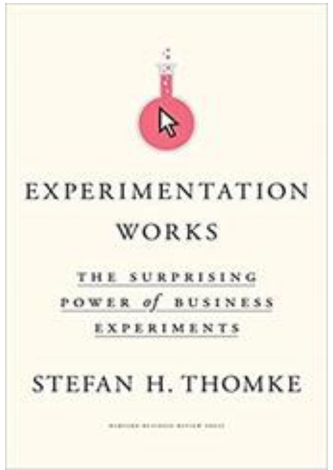
That’s it!.

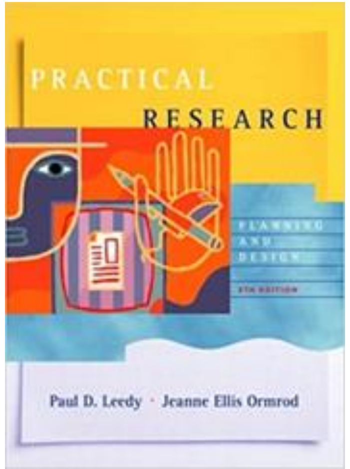
Books

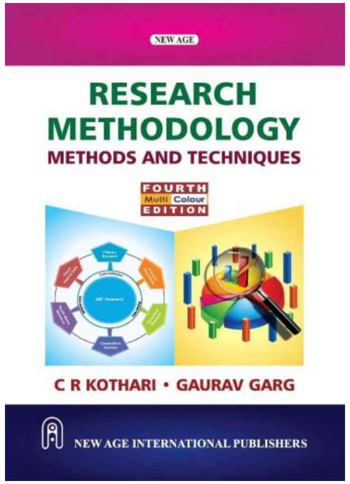

TRUSTWORTHY
ONLINE CONTROLLED
EXPERIMENTS
A PRACTICAL GUIDE TO A/B TESTING
RON KOHAVI • DIANE TANG • YA XU


STATISTICAL
METHODS
in Online A/B Testing
Statistics for data-driven business decisions
and risk management in e-commerce
GEORGI Z. GEORGIEV


FIELD
EXPERIMENTS
DESIGN, ANALYSIS,
AND
INTERPRETATION
ALAN S. GERBER | DONALD P. GREEN


EXPERIMENTATION
WORKS
THE SURPRISING
POWER of BUSINESS
EXPERIMENTS
STEFAN H. THOMKE


PRACTICAL
RESEARCH
Paul D. Leedy · Jeanne Ellis Ormrod


NEW AGE
RESEARCH
METHODOLOGY
METHODS AND TECHNIQUES
FOURTH
Multi Colour
EDITION
C R KOTHARI • GAURAV GARG
NEW AGE INTERNATIONAL PUBLISHERS

Thanks

Give session feedback

https://docs.google.com/forms/d/1Z7n9y6lLMIsqFgC41ev3shAKtxm8ZDZ5otEud8pgiBA/

Linkedin connect

https://www.linkedin.com/in/srimugunthan-dhandapani/

Appendix

Experimental design terminology

~~Co-variate~~
~~effect~~
~~control~~
~~treatment~~
~~extraneous variable~~
~~internal validity and external validity~~
~~confounding~~
~~randomization unit~~
~~interference~~
~~Stable unit Treatment value assumption~~
~~Blocking~~
~~Factor~~

interaction effect
overlapping experiments
Variance
Orthogonality
Experiment error
Variant
sampling variability
Counterfactual
selection bias
Variance reduction