# Ab testing

* Testing new features/ UI…-> two dff groups, one with current features while one with new feature to test for the different response
* Help with the same product with new features, but not helpful for choosing different products, not helpful with new experiments, cannot tell you if missing
* Need to have clear control and clear metrics

## Cannot use ab testing:

* Long time range (e.g referral)
* Emotional choose (e.g change logo)
* Specific product cannot tell answers in general (e.g if shopping site complete)
* Should separate people to one group and another (e.g adding premium service)

# Rates vs Probability

* Rate: measure the usability of the site (how often to specific button)
* Probability: measure the total impact (how often will move to another state)

# Binomial distribution

* Two types of outcomes
  + Success, failure
* Independent events
* Identical distribution
  + P same for all

## Difference statistically significant and practically significant

Diagram, schematic, box and whisker chart

Description automatically generated

# Policy and Ethics for Experiments

## IRB Principles:

1. Risk

Whether the risk exceed the ‘minimal risk’

Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. The harm considered encompasses physical, psychological and emotional, social, and economic concerns. If the risk exceeds minimal risk, then informed consent is required.

1. Benefit
2. Choice

Whether participate really has the choice for alternatives

1. Privacy

Data sensitivity

Graphical user interface, text

Description automatically generated

# Choosing and Characterizing Metrics

## Define

* + Define the criteria to be measured, possibly in different metrics, and then combine them into one
  + Each funnel is a metrics
    - Sanity checking metrics -> use multiple metrics
    - Evaluation metrics -> OEC (overall evaluation criteria)
    - Combination metrics (overall evaluation metric)
  + Difficult metrics
    - No access to data, too long to collect
    - Can use other technologies (external data) as proxy

Diagram

Description automatically generated

## Build institution

### Segmenting and Filtering data

### Categories of summary metrics

* + - Sums and counts
      * E.g # users who visited the page
    - Distributional metrics (mean, median, 25%...)
      * Mean age of users who complete the course
      * Median latency of page load
    - Probability and rates
      * Probability has 0 or 1 outcome in each case
      * Rate has 0 or more
    - Ratios
      * Can be any value

### Sensitivity and robustness

* + - Running experiments or using the experiments already have
      * A versus A experiments (control A vs another control A)-> check if its too sensitive
        + Just compare people who saw the same thing to each other
        + Application

Sanity check (compare results to what’s your expected)

Estimate variance and calculate confidence

Directly estimate CI

* + - * Check previous experiments
    - Retrospective analysis of logs

## Characterize

### Variability

* + - Empirical Variability
      * Sometime the analytical estimate of the variance ended up being an underestimates
      * Bootstrapping
        + Random sampled and repeat to get any experiments (run x times in one experiment using bs is the same as doing x times experiments)
      * CI
        + Analysis CI -> need to assume data distrubition and independency

For each experiment, calculate the dff betn each other

Calculate the sdv. of each dff (assume normally distributed)

Calculate the CI

* + - * + Empirical CI

For each experiment, calculate the dff bewn each other

Order the diff

Choose the second min and max values as bound

# Designing an Experiment

### Choose ‘subject’

#### Unit of diversion -> how we define what an individual subject is in an experiment

* + - Event
      * No consistent experiences
    - User ID -> may change due to dff accounts
      * Stackable, unchanging
      * Personal identifiablebroeser
    - Anonymous ID (cookie) -> single browser and singer device
      * Changes when you switch browser/device
      * Users can clear cookies
    - Others
      * Device ID
        + Only available for mobile
        + Tied specific device
        + Unchangeable by user
      * IP Address
        + Changes when location changes

Text, whiteboard

Description automatically generated

#### Consistency

* + - * User visible change -> cookies, user ID
      * User unnoticeable -> event
      * Ip is not usually consistent

A picture containing text

Description automatically generated

#### Ethical

Text, letter

Description automatically generated

#### Variability

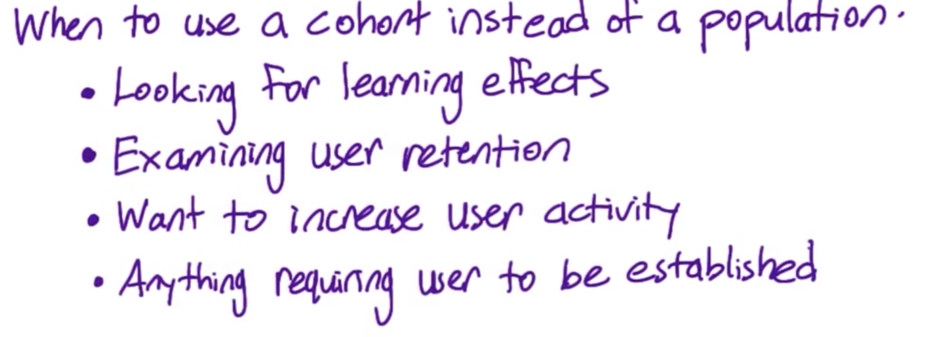
* + - * Unit of analysis
        + The denominator of the metrics
        + Whwhenanalytiven unit of analysis = unit of diversion -> the analytic variance tends to be lower and more likely to match the empirical variance

Text, letter, whiteboard

Description automatically generated

## Choose ‘population’

* + Globally? Us?
  + Intra-user Experiment
    - Observe user behavior in time range
    - In an interleaved ranking experiment, suppose you have two ranking algorithms, X and Y. Algorithm X would show results X1, X2, … XN in that order, and algorithm Y would show Y1, Y2, … YN. An interleaved experiment would show some interleaving of those results, for example, X1, Y1, X2, Y2, … with duplicate results removed. One way to measure this would be by comparing the click-through-rate or -probability of the results from the two algorithms.
  + Target population
  + Cohort
    - Defining entering class
    - For user stability

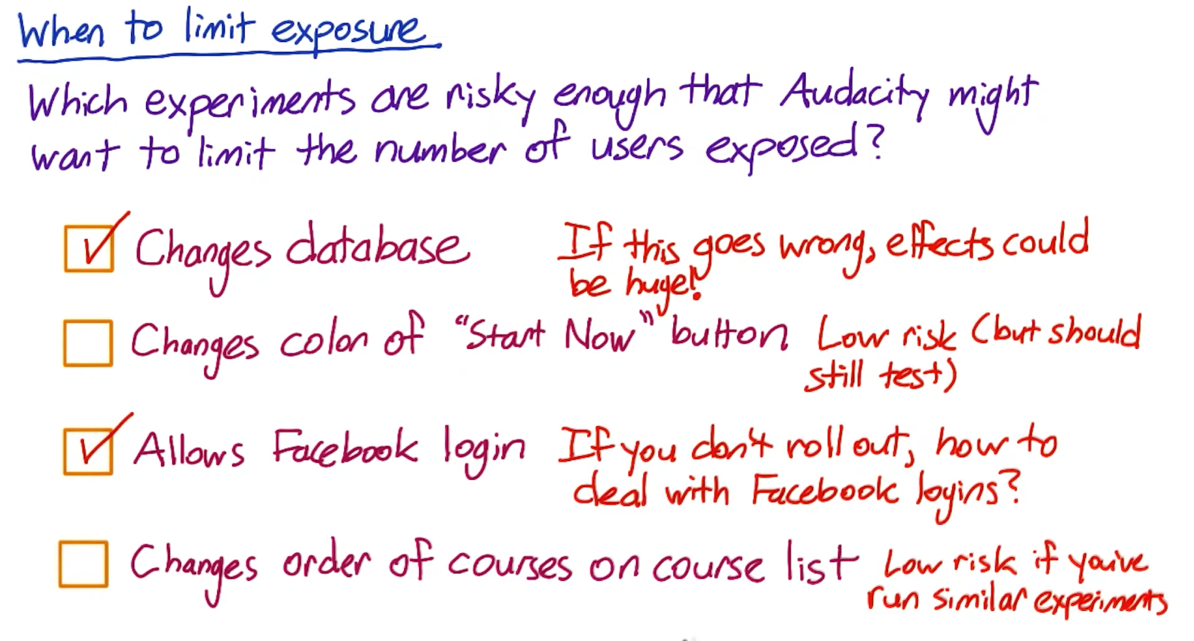


## Size

* + Iterative process
  + Reduce the size
    - Increase dmin, α， β
    - Make unit of diversion and unit of analysis the same
    - Target experiment to specific traffic

## Duration

* + Duration
  + When
  + What fraction of traffic will you send to experiment
    - Cannot be 1 -> safe, maybe other issues affect the results
    - Work on multiple task, each occupy small fraction to make them comparable in the same time ranges



Learning effects

* When you want to measure user learning or effectively whether a user is adapting to a change or not
* Types
  + Change aversion
  + Novelty effect
* Keep in mind
  + Choosing the unit of diversion correctly
  + Dosage – how often for the change
    - Choosing cohort: how long they the change, how many times they see the change
  + Risk and duration
    - Duration: may need a period for the changes (don’t put all users in)
    - Risk: uncertain about the effect (maybe worse)
    - So -> it’s better to run for a small groups for a long period of time
  + Pre-period, post-period

# Analyzing Results

## Sanity Check

1. Choosing invariant metrics
   * Population Sizing metrics
     + Check the experiment population and control are comparable
   * Actual invariants
     + Check if they change or not
2. Checking
   * Technically issue
   * Retrospective analysis
   * Trying to use pre- and post- periods (time-phased inspection)
     + Data capture (especially for new experiment）
     + Experiment setup
     + Rare: Infrastructure, experiment system

## Single Metric

* + If there’s statistically significant of the experiments
  + The magnitude and direction of the change
  + Simpson’s Paradox
    - Sometimes overall results are different to subset groups
    - Recommend: dig deeper
  + OEC: Help to better make recommendation based on the results, create a balance between two metrics

## Multiple Metrics

* + Multiple comparisons -> adjust significance level
  + Automatically detection of differences
    - E.g run multiple time of EDA to check metrics
  + Problem: probability of any false positive increases as you increased number of metrics
    - Solution: using higher CI for individual metrics
      * Assume independence
        + Alpha\_overall = 1 – (1- alpha\_individual)^n
      * Bonferroni Correlation
        + Advantages:

Simple

No assumptions

Conservative - guarantee to get the overall alpha at least as small as you specified

* + - * + Alpha\_individual = alpha\_overall / n
        + Limit:

if tracking metrics are correlated, they maybe move at the same time

too conservatives

* + Conclusion
    - Do you understand the change
    - Do you want to launch the change

## Gotchas: changes over time

* + Ramp up when launching a change
    - The effect maybe flattens out as you ramp up the change, even the initial effect was statistically significant (the results are not repeatable)
      * Seasonality effect / event-driven impact
        + Holdbacks: launch a change to everyone except for a small group of holdbacks (no change for them)
      * Novelty effect / change aversion
        + Cohort analysis
      * Advertisers (budget changes)
  + Using pre- post- period to detect how the users adapt the changes over time
* Conclusion
  + Check
    - if the experiment set up properly
    - variance
    - if experiment metrics looking same
  + Business decision
    - Fine tuning if only works on small group (30%) of users
  + Overall business analysis
    - Engineering cost of maintain the change
    - Customer support or sales issues?
    - Opportunity cost versus potential benefit from the change or not launching
  + Incidental impact
    - Run dff experiments to see the results compared to the first experiment before launching