# Lesson 4 Designing an Experiment

## Outline

* Choose ‘subject’
* Choose ‘population’
* Size
* Duration

## Choose “subject” – unit of diversion

We can use event-based assignment for choosing either the control or experiment group for each pageview. But for a user visible change, we want to basically assign people as opposed to events. But how to determine who a person is?

* Each id? But each person could have multiple logins, such as both consumer and corporate accounts.
* Anonymous identifier like cookie, which is tied to a single browser or a single device? But users can use different devices or browsers.

So, in any case, if we’re trying to assign a person, we will use some imperfect proxy.

**Pros and cons for each unit of diversion**

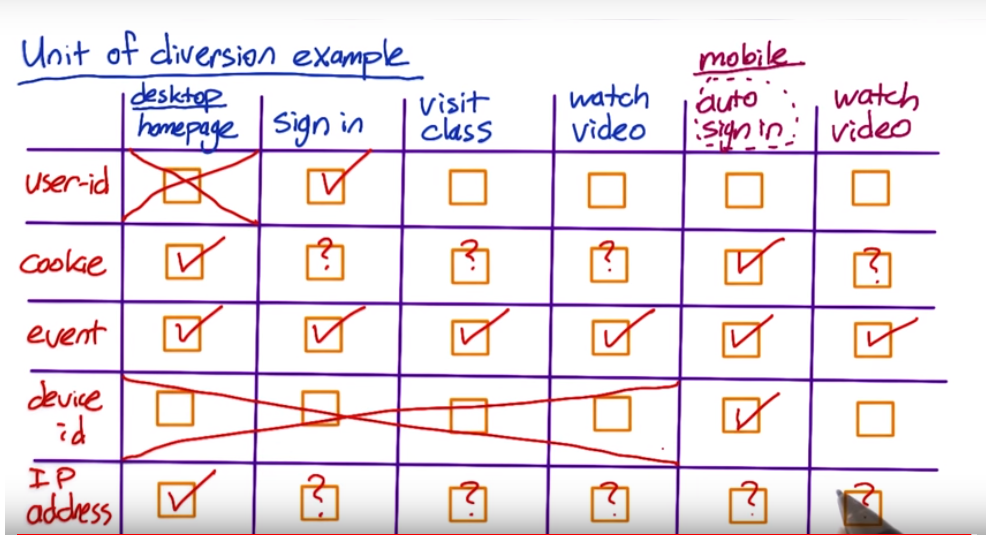
Commonly used:

* User id, such as login email address
  + Stable, unchanging
  + Personally identifiable
  + Requires ethical consideration
* Anonymous id, such as cookie
  + Changes when we switch browser or device
  + Users can clear cookies
* Event, such as pageviews or queries
  + No consistent experience
  + Changes when users refresh pageviews
  + Use only for non-user-visible changes

Less common:

* Device id
  + Only available for mobile
  + Tied to specific device
  + Unchangeable by user
* IP address
  + Changes when location changes

Check if subject can be switched between control and experiment groups.

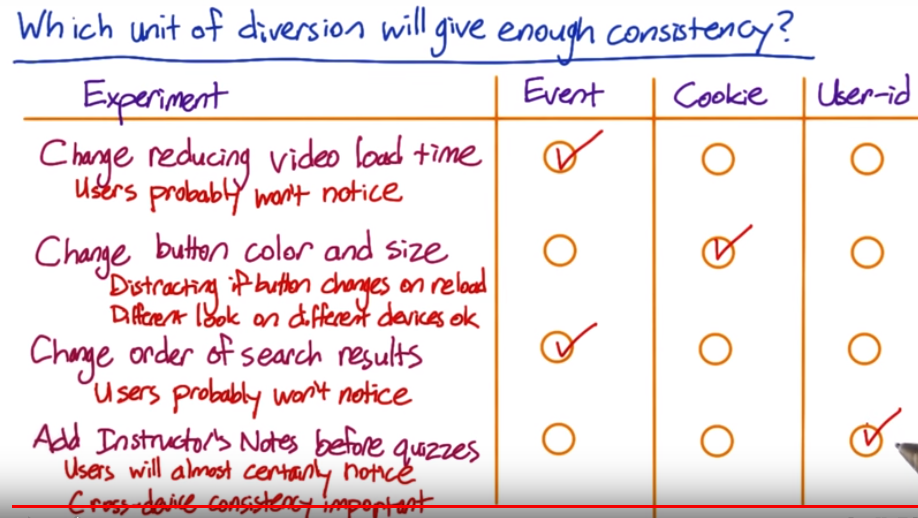


? means maybe

## 3 main considerations for diversion

* Consistency (event/cookie/user id?)
* Ethical considerations (especially for user id, requiring user consent form)
* Variability (unit of analysis vs. unit of diversion)

## Consistency of diversion



## Ethical considerations

Depends on whether there is new information being collected, experiments might require additional ethical review. For example,

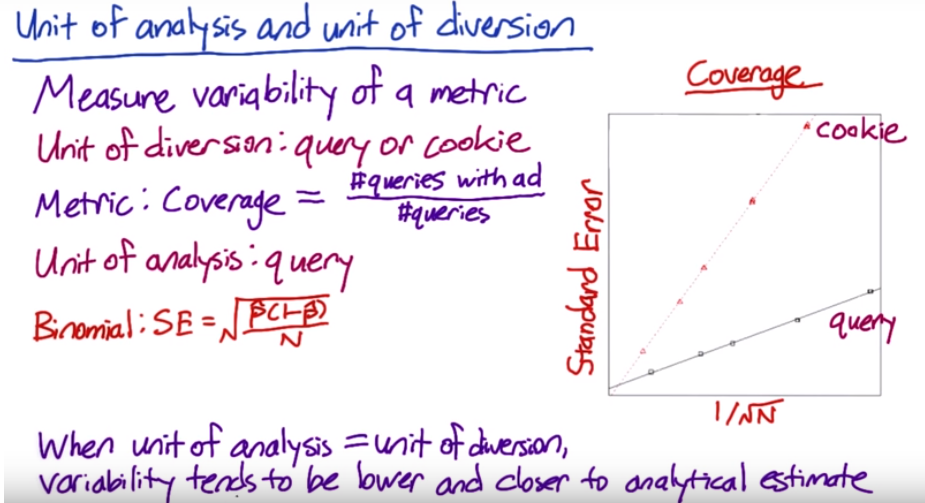
* Newsletter prompt after starting course (user id diversion)
  + No new information being collected
  + Fine if original data collection was approved
* Newsletter prompt on course overview (cookie diversion)
  + Depends: are email address stored by cookie?
  + If yes, cookies become non-anonymous
  + Potentially impacts other data collection
* Changes course overview page (cookie diversion)
  + Not a problem, and probably already being done.

## Unit of analysis vs. unit of diversion

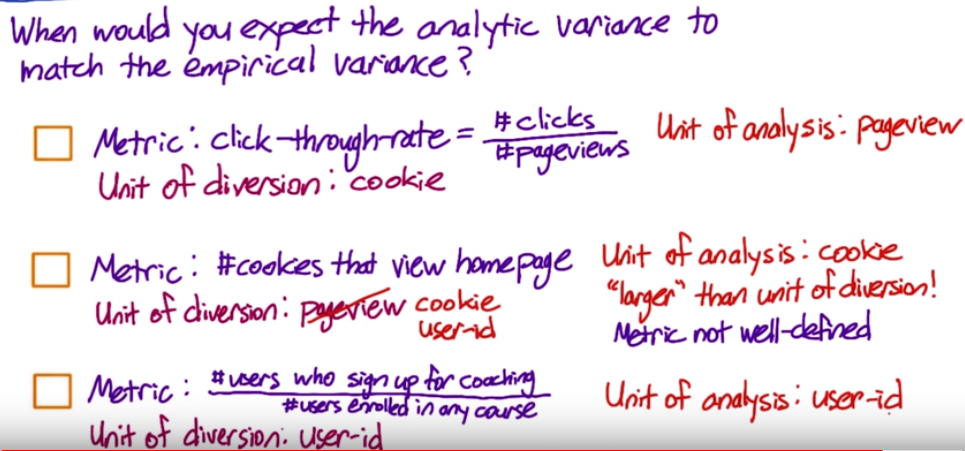
Sometimes the empirically computed variability is much higher than the analytically computed variability. This happens because the unit of analysis is different from the unit of diversion. Assumptions for analytically computed variability are:

* Underlying distribution
* Data collected are independent

Unit of analysis is whatever the unit of our metric is. For CTR, which is the ratio of clicks to page views, then page views will be our unit of analysis. If we also use pageviews (event) as unit of diversion, then the analytically computed variability is likely to be very close to the empirically computed variability. But if we choose user ids or cookies as unit of diversion, our collected data will be correlated to each other (basically we are diverting groups of events and the independent assumption is violated), the variability of the same metric CLR is going to be much higher. Therefore, we should move to empirically computed variability given our unit of diversion.



In reality, we need to make sure our unit of diversion to be at least as big as our unit of analysis, otherwise same unit could be assigned to both control and experiment groups.



## Inter- vs. intra-user experiments

**Intra-user experiments**: we expose the same user to the new feature being on and off over time and analyze how they behave in different time windows. But this has some pitfalls, for example, we have to be really careful about a comparable time window with consideration of holidays/seasonality/weekend. Another downside is that with complicated features, which people learn to use the particular feature in the first two weeks and then when we turn it off, users might get confused, which will result different behaviors.

**Interleaved experiments**: applied to scenarios with ranked order lists, such as search ranking or preferences. Same users will be expose to the A and B side at the same time.

**Inter-user experiments**: we expose different people on the A side and B side. Define cohort is essential for this type of experiment.

## Target population

If we can identify what population will be affected by the experiments, we might want to target our experiment to that traffic and only run our experiment on the affected traffic. But filtering the traffic can change the variability of our metric too.

## Population vs. Cohort

Cohort is defined as people who enter the experiment at the same time in both groups, which limit our experiment to a subset of the population.

When to use a cohort instead of a population:

* Looking for learning effects
* Examining user retention
* Want to increase user activity
* Anything requiring user to be established

## Sizing

The size of experiment depends on practical significance, statistical significance and the sensitivity (from lesson 1). But the size of experiment also depends on variability of our metric.

For example, if we are running an experiment on changing order of courses on course list.

Metric: click-through-rate

Unit-of-diversion: cookie

for 1000 pageviews

Result: need 300,000 pageviews per group!

The following strategies could reduce the number of pageviews:

* Increase ,
* Change unit of diversion to pageview
  + Make unit of diversion same as unit of analysis, which reduces the variability of the metric and be closer to the analytical estimate.
  + But will less consistent experience be okay?
  + If , -> only 34,000 pageviews per group
* Target experiment to specific traffic
  + If English is the only targeted area, the non-english traffic will dilute the results of the experiment, which will require more page views
  + Even though the size of English pageviews might not decrease, we can run other experiments on non-english traffic. So, it’s still worthy doing.
  + This strategy can also impact choice of practical significance boundary.
  + If SE = 0.00188 and d\_min = 0.015 -> only 12,000 pageviews per group
* Change metric to cookie-based click-through-probability
  + Doesn’t make significant difference
  + If there is a difference, variability would probably go down

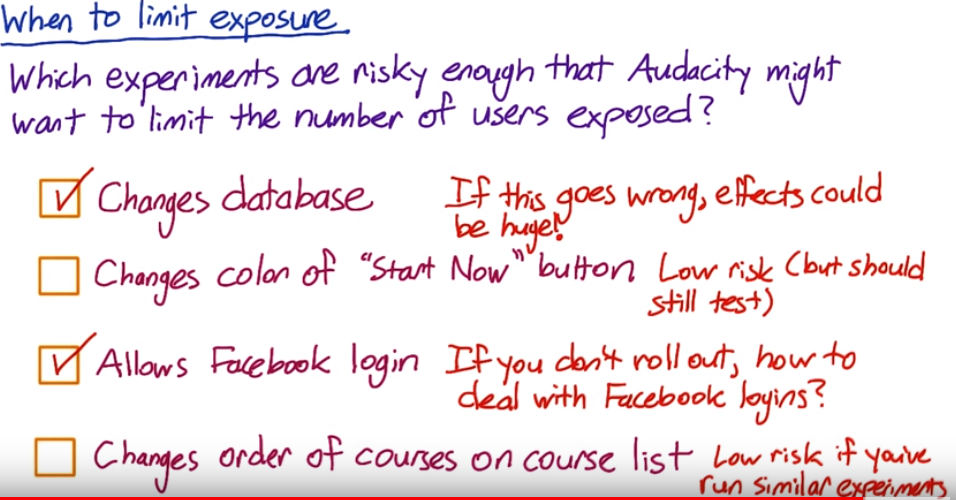
## Duration

After we decide the size of our experiment, we need to find out:

* What’s the duration of the experiment?
* When to run the experiment? (holidays, students back to school…)?
* What fraction of traffic we will send through the experiment? (how many percent of daily traffic should we use for the experiment) – exposure

The reasons for not using all traffic for the experiment are:

* Safety. For example, if we have a new UI feature, and we’re not sure either how well it functions in all browsers or how our users are going to react. Therefore, we should keep our site mostly the same, and only expose a few people to it until we feel more comfortable with it.
* Press. For a new feature that we are not sure if we are going to keep it, we wouldn’t want all people to see it or blogging about.
* Other things that can impact the variability of our metric, such as holidays, weekday/weekend. We might prefer to run a small percentage across multiple days to get a sense for how the differences are varied by time/holidays.



## Learning effects

Definition: whether a user is adapting to a change or not. There are two types of learning effects:

* Change aversion: I don’t like changes
* Novelty effect: try everything

Overtime, users are going to plateau to a very different behavior. It takes time for user to adapt the change and often we don’t have the luxury of taking that much time to make a decision.

Cohort analysis can help us to understand learning effects. We will choose a cohort in both the experiment and control groups based on either how long users are being exposed to the changes or how many times they’ve seen it.

Pre-period A/A test and post-period A/A test is another way to study leaning effects. Both tests will be used in a way that’s specific to our experiment and control (such as same target population, same duration, same size…):

* First we run a pre-period on the exact same population as A/B test but with exact same version. The difference we measure is due to system variability or user variability.
* After A/B test, run post-period, which is another A/A test. If there is additional difference discovered, we can attribute those differences to user leaning that happened in the experiment period.

## Summary

* Pay attention to the relationship between the metric and the unit of diversion.
* Building intuition during the process.