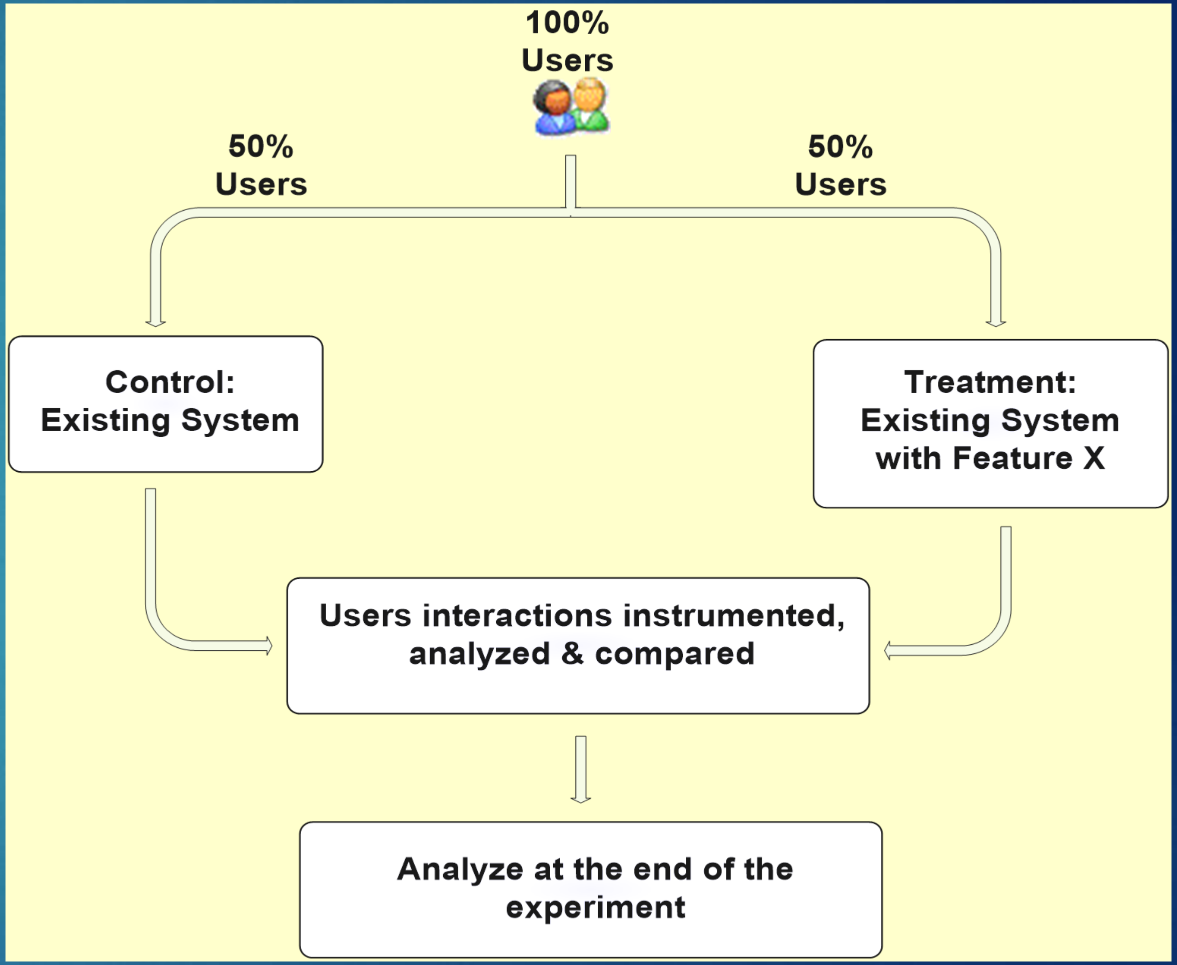
Pitfalls of Long-Term Online Controlled Experiments

Pavel Dmitriev, Brian Frasca, Somit Gupta, Ronny Kohavi, Garnet Vaz

Controlled Experiments aka A/B Tests

- Best scientific way to prove ***causality***, i.e., the changes in metrics are caused by changes introduced in the treatment

- Want to evaluate new ideas fast, so short-term experiments are preferred, as long as can obtain stat sig results

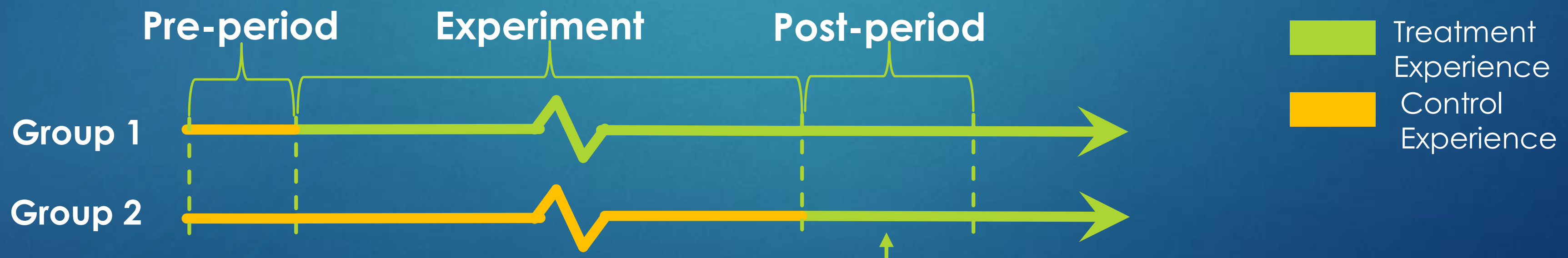

100% Users
50% Users
50% Users
Control: Existing System
Treatment: Existing System with Feature X
Users interactions instrumented, analyzed & compared
Analyze at the end of the experiment

Are results obtained in short-term experiments correctly predict long-term impact?

- The common assumption is that they are: most experiments described in literature are short-term
- But we also believe that sometimes they aren’t
  - A ranking bug in an experiment resulted in very poor search results
  - Degraded (algorithmic) search results cause users to search more to complete their task, and ads appear more relevant
  - Distinct queries went up over 10%, and revenue went up over 30%
  - This won’t last!

Are results obtained in short-term experiments correctly predict long-term impact?

- Two approaches
  1. Design an OEC (Overall Evaluation Criteria) based on metrics that are predictive of long-term value (KDD2012 paper)
  2. Run long term experiments, build a model for learning effect (change in user behavior over the course of experiment). Use model to predict long-term impact based on results from the first few weeks of the experiment (Google's KDD2015 paper = FLT paper)



Measure the difference in this period

Pitfall 1: Cookie stability

- Many sites, including Bing and Google, identify users based on cookies.
- The user is randomized into control/treatment based on their cookie. If user cleans/looses a cookie, a new one is generated resulting in a new experiment assignment.
- Studies estimate 25%-33% users change their cookie in a month.
- While some users clear cookie intentionally, others lose their cookies unintentionally due to e.g. browser bugs. We call phenomenon of unintentional cookie loss ***Cookie Clobbering***.

Pitfall 1: Cookie stability (cont.)

- Cookie churn makes it hard to run long-term experiments correctly
  - Treatment effect spills over into control
  - Results are invalid if treatment impact cookie churn rate
  - Few cookies survive till the end (more on this later)

- Mitigations:
  - Make sure there’s no stat sig difference in cookie churn rate between control and treatment
  - Cookie backup/restore to mitigate cookie clobbering.
    - Store the cookie in two domains associated with the site, e.g. bing.com and www.bing.com
    - If cookie is missing in one domain, restore it from the other
    - Experiments showed significant reduction in cookie clobbering

Pitfall 2: Survivorship Bias

- During WWII British wanted to add armor to bombers. Originally, the military wanted to add armor to the most damaged areas. Statistician Abraham Wald pointed out that instead the armor should be placed where there was no damage, since the bombers hit in those places… never made it back!
- Ignoring survivorship bias may lead to opposite conclusions!
- Measuring survivorship bias in Search experiments: Sessions/User or Tasks/User metrics
  - Degradation indicates increased abandonment, resulting in survivorship bias during the post-period

Pitfall 2: Survivorship Bias (cont.)

- We ran a series of experiments varying the number of Ads users see. Similar to the experiments described in FLT paper
  - FLT paper: detected no impact on Tasks/User (no survivorship bias)
  - We: detected stat sig impact on Sessions/User, persisting in the post-period
- **Cannot trust any measurements in the post-period when survivorship bias is present!**
  - Cannot use FLT paper's methodology
- Mitigation:
  - Some metrics (Clicks/User, Queries/User) may be adjusted by imputing 0's for users who did not show up in the post-period.
  - Ratio metrics such as Ad CTR = #Clicks/#Queries are hard to adjust.

Pitfall 3: Selection bias

1936 Roosevelt-Landon presidential election

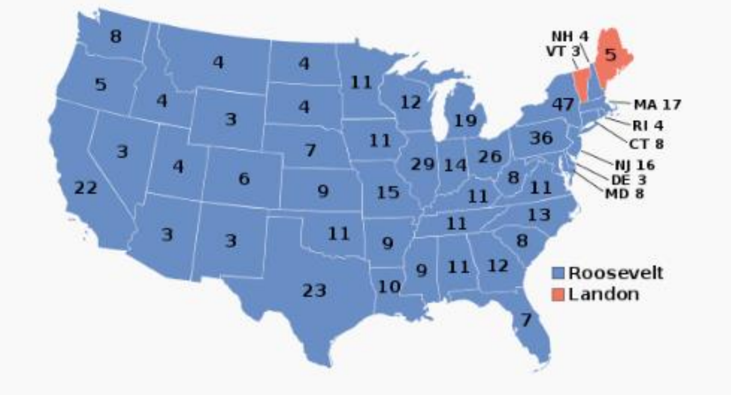
The 1936, The Literary Digest poll showed that the Republican candidate, Governor Alfred Landon of Kansas, was likely to be the overwhelming winner.

Roosevelt won that election in landslide by winning all but 8 electoral votes.

Literary magazine has polled ten million individuals (of whom about 2.4 million responded), it had a large selection bias:

- It surveyed its own readers first, then two other readily available lists: that of registered automobile owners and that of telephone users
- All wealthier than the average American at the time.

That same year, George Gallup, an advertising executive who had begun a scientific poll, predicted that Roosevelt would win the election, based on a quota sample of 50,000 people.



Presidential election results map. Blue denotes those won by Roosevelt/Garner, red denotes states won by Landon/Knox. Numbers indicate the number of electoral votes allotted to each state.

https://en.wikipedia.org/wiki/United_States_presidential_election,_1936

Pitfall 3: Selection bias

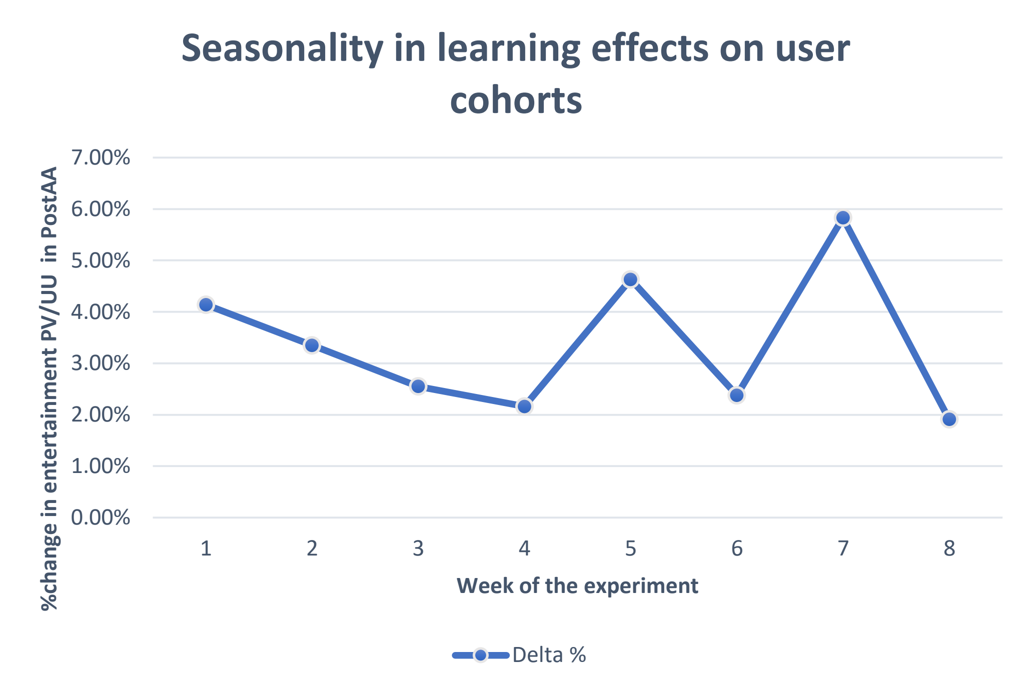
- Even in cases where there is no survivorship bias, FLT analysis can suffer from selection bias. User cohorts from the pre-period that survived in the post-period can have very different overall characteristics than the cohort of all users in pre-period.
- After 3 months, only 23% of the cookies from the pre experiment survived in the post experiment period. The characteristics of the survived population are very different from the overall population.
- **Mitigation:** We believe that the best estimate of the long –term impact on all users is provided by measuring the changes at the end of the experiment period.

	Delta between pre- and post-periods
Number of Users	-77%
Sessions / User	81%
Queries / User	81%
Revenue / User	80%
Ad Clicks / User	74%
Overall Clicks / User	86%
Ad CTR	6%
Overall CTR	5%

Delta on key metrics between the groups in pre-experiment and post-experiment periods, 2-week time period, control experience.

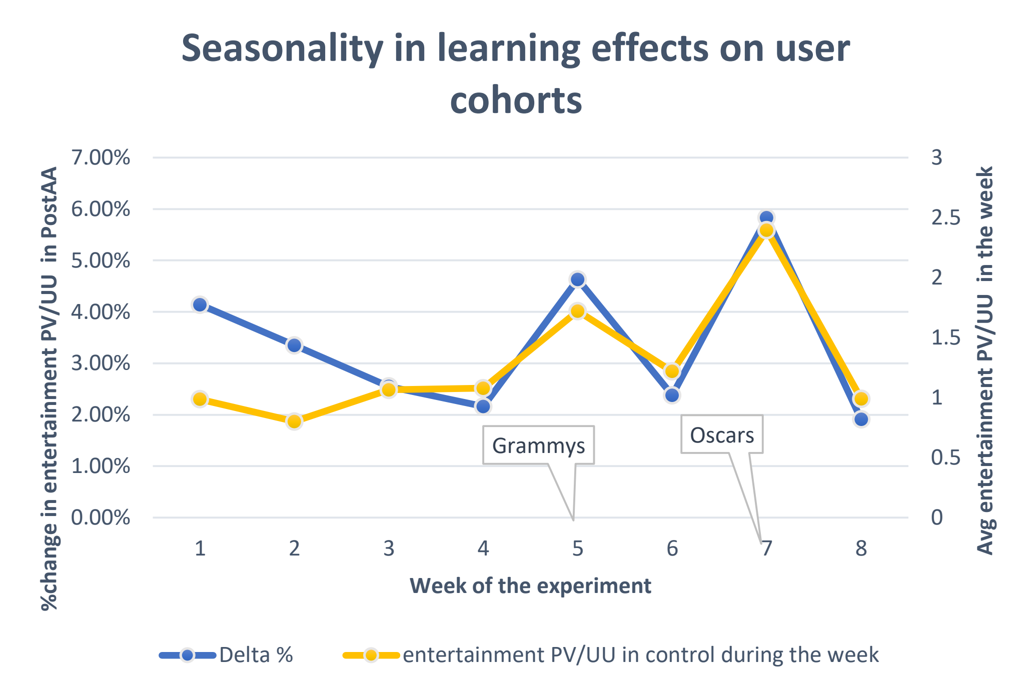
Pitfall 4: Seasonality effects

- Cohort analysis as suggested in FLT paper can also suffer from seasonality. FLT assumes the learning effect increases with the increase in period of exposure
- We ran a 2 month long experiment on MSN homepage where the users in treatment group were shown more entertainment articles than those in the control. We measured the learning effect on user cohorts with different time periods of exposure to the experiment.
- We detected spikes in the learning effect on cohort of users that were first exposed to the experiment in week 5 and 7.



Pitfall 4: Seasonality effects

- Turned out that Grammy's were held in week 5 and Oscars in week 7 of the experiment which might explain spikes in the learning effect.
- In general, user cohorts from different periods can have different propensity of using different features and can have different survival rates in the post-period.
- **Seasonality can impact learning effect**
- **Mitigation:** In such cases, we recommend analyzing all users exposed to the experiment. This analysis will have less seasonality, survivorship and selection bias, and higher sensitivity.



Conclusion

- There are several pitfalls when running a long-term online experiment that may undermine the external validity of the experiment.
  - Cookie stability
  - Survivorship bias
  - Selection bias
  - Seasonality
- A few more pitfalls are mentioned in the paper: http://bit.ly/longTermExP
- There are a lot of open issues in analysis of long-term running experiments. We hope to see more research in this area.

HiPPOs & booklets with selected papers are available at the back!