Studying Online Behavior at Scale (SOC 412)

Week 2 Lecture 1

Sherrerd Hall 306

J. Nathan Matias

@natematias

civilservant.io

jmatias@princeton.edu

Department of Psychology
PRINCETON UNIVERSITY

CITP

mit media lab

What we will cover today

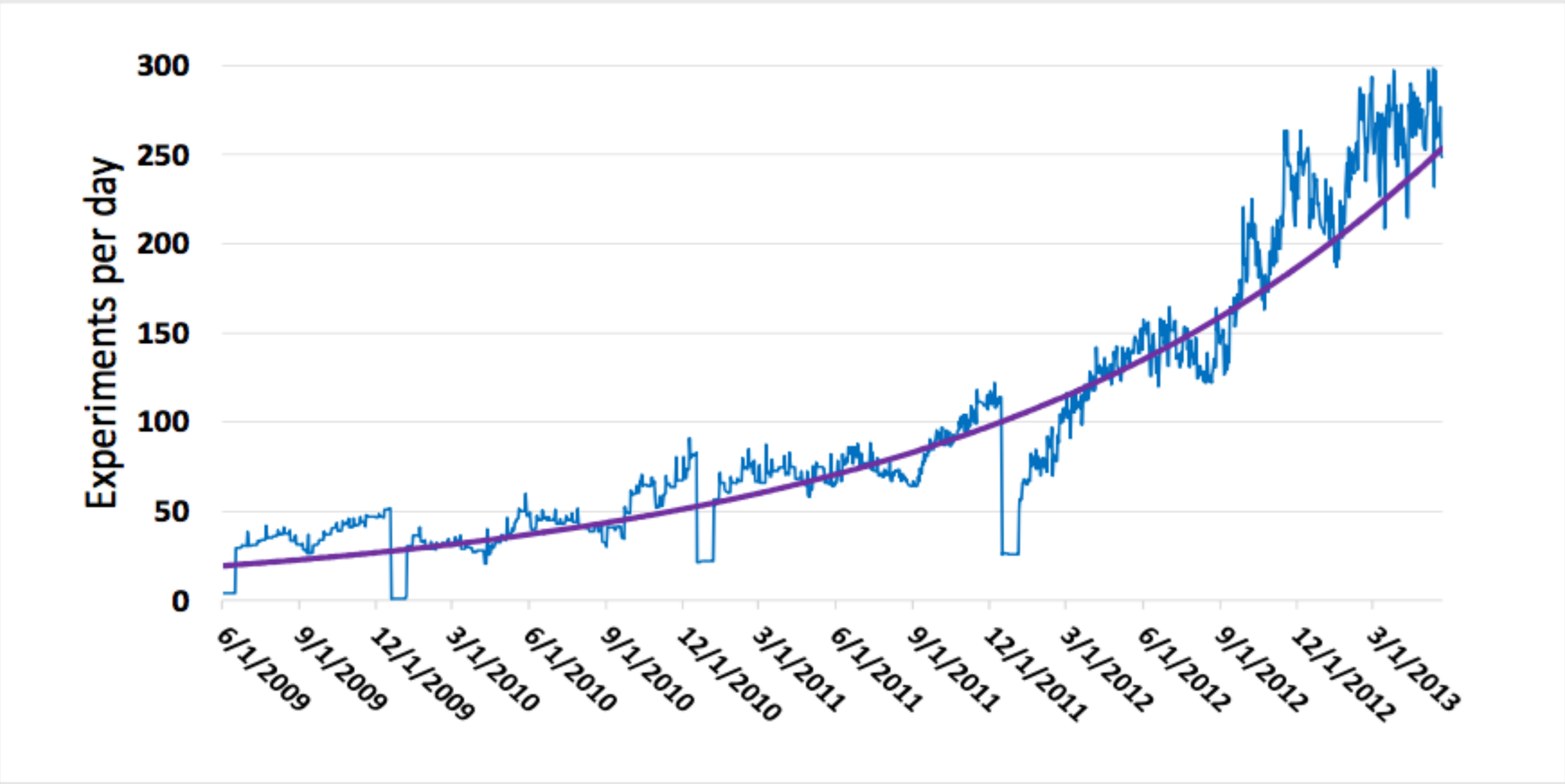
Discuss today’s readings

Discuss upcoming assignments

- The Social Media Color Experiment

- The Cornhole Challenge

A rhythm for readings, discussion, presentations



Experiments Per Day on bing.com

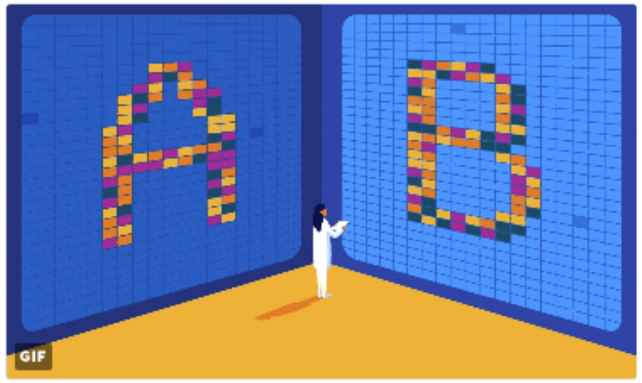
Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). **Online controlled experiments at large scale**. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176). ACM.

bing

MailChimp
@MailChimp

Follow

We looked at nearly 500K A/B testing campaigns to determine how long a test needs to run to give you accurate results. Here's what we found: expi.co/01iJRL



1:00 AM - 9 Feb 2018

19 Retweets 50 Likes

19 50

Jonathan Parks Follow
May 10, 2017 · 8 min read

Scaling Airbnb's Experimentation Platform

Simon Jackson
Data scientist in the Experiment Tool team at Booking.com, Ph.D. in cognitive psychology, R guy.
Jan 22 · 10 min read

How Booking.com increases the power of online experiments with CUPED

Simon Jackson | Data Scientist at Booking.com

Data-supported decisions rule the roost at Booking.com. All product teams are empowered to do controlled experiments (A/B testing) and test any changes they make to the website (Kaufman, Pitchforth, & Vermeer, 2017). Such experiments expose some users to the existing website (base) while others see a new variant, and we statistically test the observed difference.

The Unofficial Google Data Science Blog

HOME ABOUT THIS BLOG

Designing A/B tests in a collaboration network

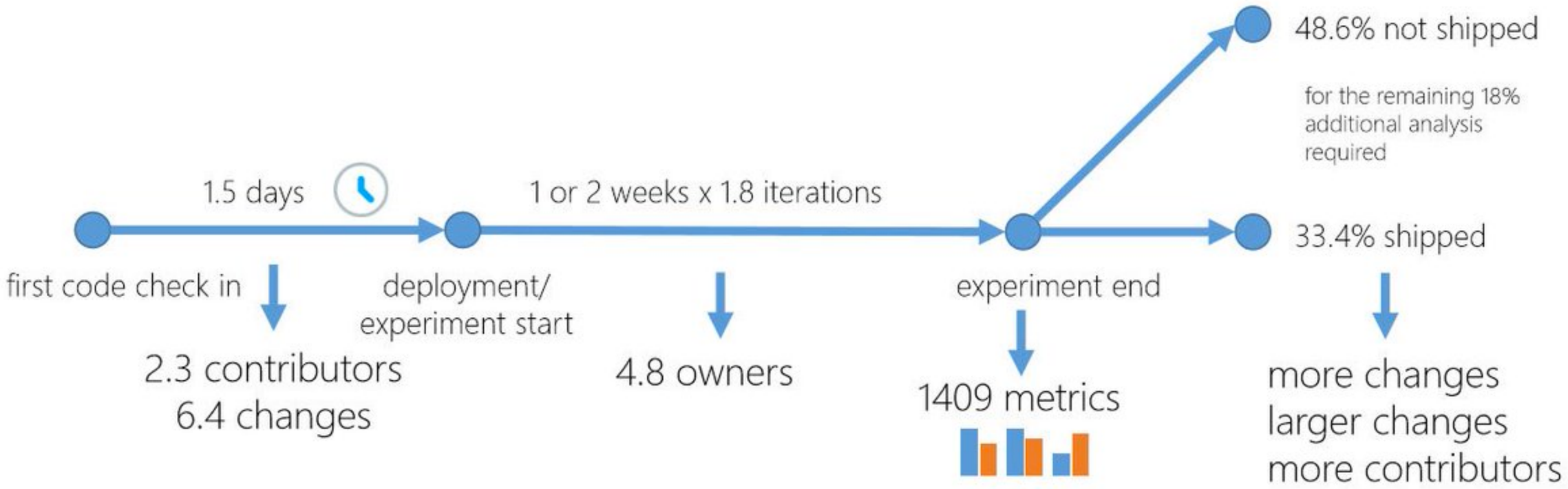
January 16, 2018

BY SANGHO YOON

In this article, we discuss an approach to the design of experiments in a network. In particular, we describe a method to prevent potential contamination (or inconsistent treatment exposure) of samples due to network effects. We present data from Google Cloud Platform (GCP) as an example of how we use A/B testing when users are connected. Our methodology can be extended to other areas where the network is observed and when avoiding contamination is of primary concern in experiment design. We first describe the unique challenges in designing experiments on developers working on GCP. We then use simulation to show how proper selection of the randomization unit can avoid estimation bias. This simulation is based on the actual user network of GCP.

Experimentation on networks

A/B testing is a standard method of measuring the effect of changes by randomizing samples into different treatment groups. Randomization is essential to A/B testing because it removes selection bias as well as the potential for confounding factors in assessing treatment effects.



Kevic, K., Murphy, B., Williams, L., & Beckmann, J. (2017, May). **Characterizing experimentation in continuous deployment: a case study on bing.** In Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Practice Track (pp. 123-132). IEEE Press.

bing

	Category/ Phase	Crawl	Walk	Run	Fly
Technical Evolution	**Technical focus of product dev. Activities**	**(1) Logging of signals** **(2) Work on data quality issues** **(3) Manual analysis of experiments** Transitioning from the debugging logs to a format that can be used for data-driven development.	**(1) Setting-up a reliable pipeline** **(2) Creation of simple metrics** Combining signals with analysis units. Four types of metrics are created: debug metrics (largest group), success metrics, guardrail metrics and data quality metrics.	**(1) Learning experiments** **(2) Comprehensive metrics** Creation of comprehensive set of metrics using the knowledge from the learning experiments.	**(1) Standardized process for metric design and evaluation, and OEC improvement**
	Experimentation platform complexity	**No experimentation platform** An initial experiment can be coded manually (ad-hoc).	**Platform is required** 3rd party platform can be used or internally developed. The following two features are required: • **Power Analysis** • **Pre-Experiment A/A testing**	**New platform features** The experimentation platform should be extended with the following features: • **Alerting** • **Control of carry-over effect** • **Experiment iteration support**	**Advanced platform features** The following features are needed: • **Interaction control and detection** • **Near real-time detection and automatic shutdown of harmful experiments** • **Institutional memory**
	Experimentation pervasiveness	**Generating management support** Experimenting with e.g. design options for which it's not a priori clear which one is better. To generate management support to move to the next stage.	**Experiment on individual feature level** Broadening the types of experiments run on a limited set of features (design to performance, from performance to infrastructure experiments)	**Expanding to (1) more features and (2) other products** Experiment on most new features and most products.	**Experiment with every minor change to portfolio** Experiment with any change on all products in the portfolio. Even to e.g. small bug fixes on feature level.
Organizational Evolution	**Engineering team self-sufficiency**	**Limited understanding** External Data Scientist knowledge is needed in order to set-up, execute and analyse a controlled experiment.	**Creation and set-up of experiments** Creating the experiment (instrumentation, A/A testing, assigning traffic) is managed by the local Experiment Owners. Data scientists responsible for the platform supervise Experiment Owners and correct errors.	**Creation and execution of experiments** Includes monitoring for bad experiments, making ramp-up and shut-down decisions, designing and deploying experiment-specific metrics.	**Creation, execution and analyses of experiments** Scorecards showing the experiment results are intuitive for interpretation and conclusion making.
	Experimentation team organization	**Standalone** Fully centralized data science team. In product teams, however, no or very little data science skills. The standalone team needs to train the local product teams on experimentation. We introduce the role of Experiment Owner (EO).	**Embedded** Data science team that implemented the platform supports different product teams and their Experiment Owners. Product teams do not have their own data scientists that would analyse experiments independently.	**Partnership** Product teams hire their own data scientists that create a strong unity with business. Learning between the teams is limited to their communication.	**Partnership** Small data science teams in each of the product teams. Learnings from experiments are shared automatically across organization via the institutional memory features.
Business Evolution	**Overall Evaluation Criteria (OEC)**	OEC is **defined** for the first set of experiments with a few key signals that will help ground expectations and evaluation of the experiment results.	**OEC evolves** from a few key signals to a structured set of metrics consisting of Success, Guardrail and Data Quality metrics. Debug metrics are not a part of OEC.	OEC is **tailored** with the findings from the learning experiments. Single metric as a weighted combination of others is desired.	OEC is **stable**, only periodic changes allowed (e.g. 1 per year). It is also used for setting the performance goals for teams within the organization.

Figure 5. The "Experimentation Evolution Model".

Fabijan, A., Dmitriev, P., Olsson, H. H., & Bosch, J. (2017, May). **The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale.** In Proceedings of the 39th International Conference on Software Engineering (pp. 770-780). IEEE Press.

bing

The Obligation To Experiment

J. Nathan Matias

Dec 12, 2016 · 13 min read

Tech companies should test the effects of their products on our safety and civil liberties. We should also test them ourselves.

Upcoming Assignments

Why I am asking you to participate in an experiment:

If you're going to be asking other people to participate in experiments, you need to at least be willing to think about research ethics in light of your own experience and the people you know.

Upcoming Assignments

- Facebook Color Experiment
  - https://github.com/natematias/SOC412/tree/master/2-facebook-color
- Cornhole Experiment
  - https://github.com/natematias/SOC412/tree/master/2-cornhole-challenge

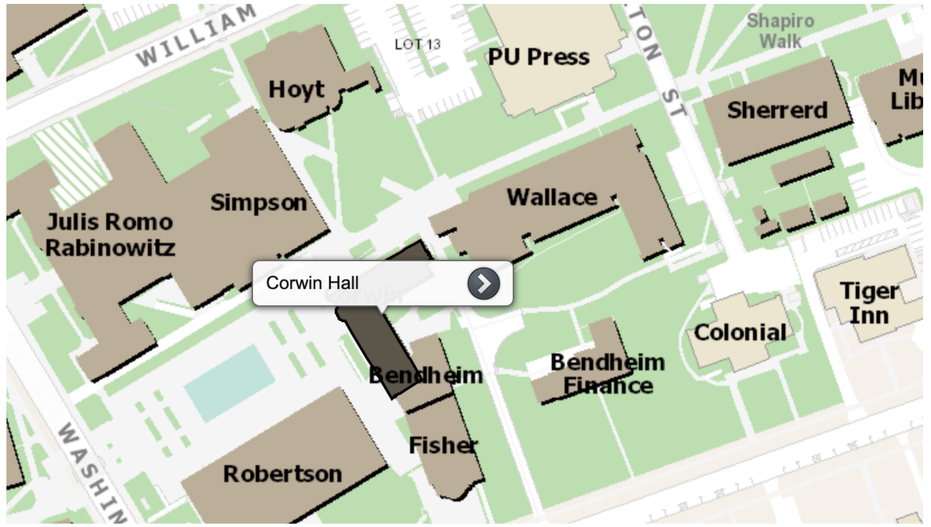
Weekly Rhythm

- **Post to Slack with one observation for the upcoming discussion**
- **Post at least one response to someone else's observation**
- **Submit assignments** by Friday at 5pm
- **Office Hours** Tues/Thurs 11-12

https://meetme.so/natematias-soc412

Precepts

- Times
  - **Wednesdays:**
  - **12:30pm**
  - **1:30pm**
- Location: **Corwin Hall 023**
- Sign up today (in class preferably)



Leading Group Discussions

- Provide a summary of the material (5-8 minutes)
- Pose some questions
- Support the conversation (you can meet with me in advance)

You should expect to lead a session (spreadsheet forthcoming)

References to Know