Modern Design of Experiments for Computational Advertising

NATHANIEL T. STEVENS

STATISTICAL METHODS FOR COMPUTATIONAL ADVERTISING
BANFF INTERNATIONAL RESEARCH STATION
OCTOBER 5, 2021

UNIVERSITY OF WATERLOO

Outline

- What are Online Controlled Experiments (OCE)?
- Open OCE Problems
- Opportunities for Academia
- Summary

What are OCEs?

What are OCEs?

- The Scientific Method is based on skepticism and empiricism.
- Experimentation is key to the Scientific Method, and is necessary for understanding the world around us.
- Historically, experiments have been used in fields such agriculture, manufacturing, physical sciences, social sciences, and medicine.
- Recently, the utility of designed experiments has been recognized within internet and technology companies, where online controlled experiments are a means to optimize products, customer customer experience, and revenue.

What are OCEs?

- In an environment of economic Darwinism, experimentation is key if businesses want to remain competitive.[1]

- The "Big Five" tech organizations (Google, Amazon, Facebook, Apple, and Microsoft) are each running 10,000+ experiments per year engaging millions of users. [2,3]
  - LinkedIn reportedly runs 400+ simultaneous experiments per day.[4]

- 1000's of companies use tools such as Optimizely, Google Optimize, Mixpanel, VWO, AB Tasty, and Split.io to run experiments.
  - Optimizely has around 500 employees and is reportedly worth $500M+.[5]

What are OCEs?

So what exactly *is* an OCE and how does it work?

In a classic A/B test, two groups of experimental units (usually people) are randomized to one of two treatments (usually different versions of a product), and the data collected in each treatment provide information about which product version is superior.

What are OCEs?

What kinds of things are companies experimenting with?

- User acquisition funnels
- User engagement mechanics
- User retention mechanics
- Email promotions and headlines
- Website layout
- Esthetic features
- Checkout experience
- Freemium conversion
- Branding
- Ad Campaigns
- Call to action language
- ML algorithms

For some real-life examples, checkout the “Leaks” on GoodUI:

https://goodui.org/leaks/

What are OCEs?

Concrete Examples:

Ryan Reynolds' face loses in an A/B test: https://youtu.be/OW_OId8aaM4

0.74%

1.61%

What are OCEs?

Concrete Examples:

- Amazon experiments with purchase reassurances
- Airbnb experiments with next available date feature
- The New York Times experiments with article headlines
- Lyft experiments with the hardware and software on their eBikes
- eHow experiments with ad placement
- Lyft worries about interference in experiments on their ridesharing network

What are OCEs?

Concrete Examples:

- Obama's 2008 campaign increased donations by $60M using a factorial experiment[6]

https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/

What are OCEs?

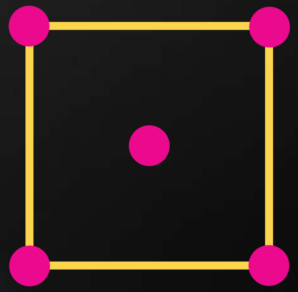
Concrete Examples:

https://de2.surveyengine.com/clients/web_demos_4_0/surveys/facebook_demo_with_data/dss/facebook_ad.html

What are OCEs?

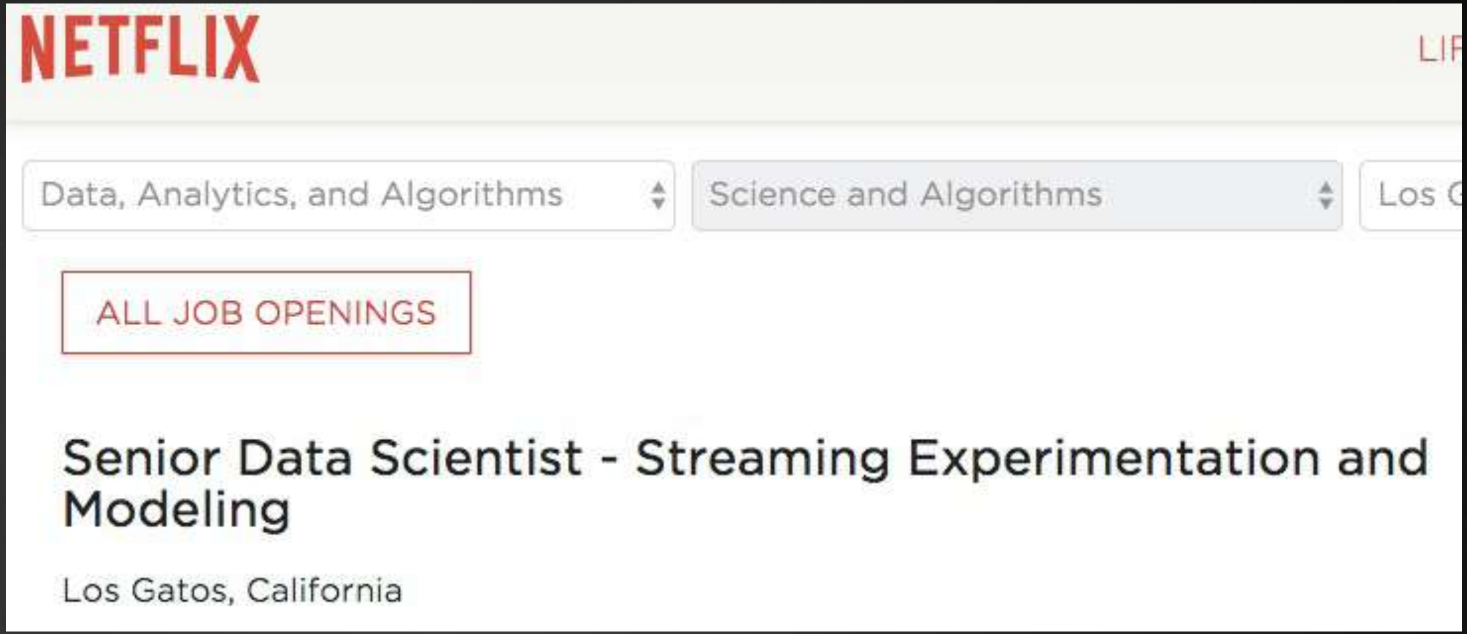
Concrete Examples: lyft

- Lyft was interested in designing a promotional offer to re-engage users that have not booked a ride in while.
- They planned to offer a discount (10% vs. 50%) on the next several (3 vs. 10) rides each user booked.
- They wanted to determine the optimal promotion (i.e., optimal discount amount and discount duration).
- They did so with a 2^2 + center point experiment.[7]



What are OCEs?

Concrete Examples:



- And the experimentation happening here isn't trivial.
- This job ad explicitly called out the need for someone that could:
  - "analyze experimental data with statistical rigor", and
  - "support internal research into new methodologies for experimentation as well as adapt existing methods such as Response Surface Methodology (RSM) to online A/B testing"

What are OCEs?

Concrete Examples:

- Google's infamous 41 shades of blue experiment reportedly increased annual revenue by $200M.[8]
- Bing generated an additional $100M in annual revenue by changing the way the search engine displayed ad headlines.[2]
- Amazon boosted profits by tens of millions per year by moving their credit card offers from the homepage to the checkout page.[2]
- Optimizely helped businesses collectively increase revenue by more than $800M in 2019 alone.[9]

Experimentation Analyst

at Sony Interactive Entertainment PlayStation

San Francisco, CA

Apply Now

PlayStation isn't just the Best Place to Play —it's also the Best Place to Work. We've thrilled gamers since 1994, when we launched the original PlayStation. Today, we're recognized as a global leader in interactive and digital entertainment. The PlayStation brand falls under Sony Interactive Entertainment, a wholly-owned subsidiary of Sony Corporation.

Data Scientist

Turo ★★★★☆ 11 reviews - San Francisco, CA

Apply Now

About You

You are a creative, rigoro
problems. You are comf
ability to creatively apply
business problems into a
to propose practical and

You have a passion for c
enjoy developing models
excited to play a key role

Data Scientist - Product Analytics

THE QUALIFICATIONS

- 2+ years of related work experience in analytics or e-commerce analytics.
- Strong SQL skills

You thrive on ambiguity and enjoy the frequent pivoting that's part of the exploration. Your tear
two pizzas can feed - and team members frequently wear multiple hats

macy's JOBS

sting and Planning team, you will work with a
predictive modeling, media mix models, time
rithms, and A/B testing capabilities, as well as
sets is desirable.

rtners. You need to be able to
ality insights and impactful
usiness growth.

e for you. Working
neering teams, you

nce, Trust

Proc

Apply Now

Ref#: P1217

the use of experimental technique
design. This role is key to help wit

We are building new machine learning p
these Adobe Cloud product lines. This p
matures.

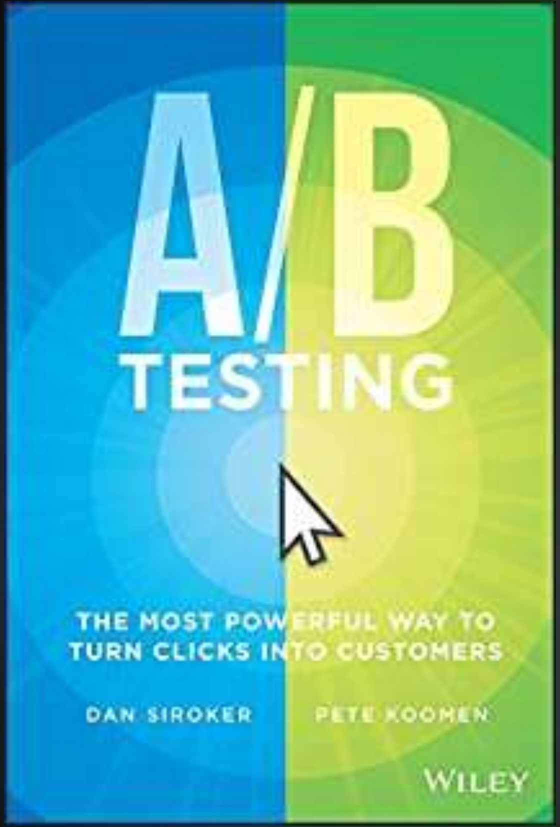
y apply to a maximum of one Data Scientist role from among those posted

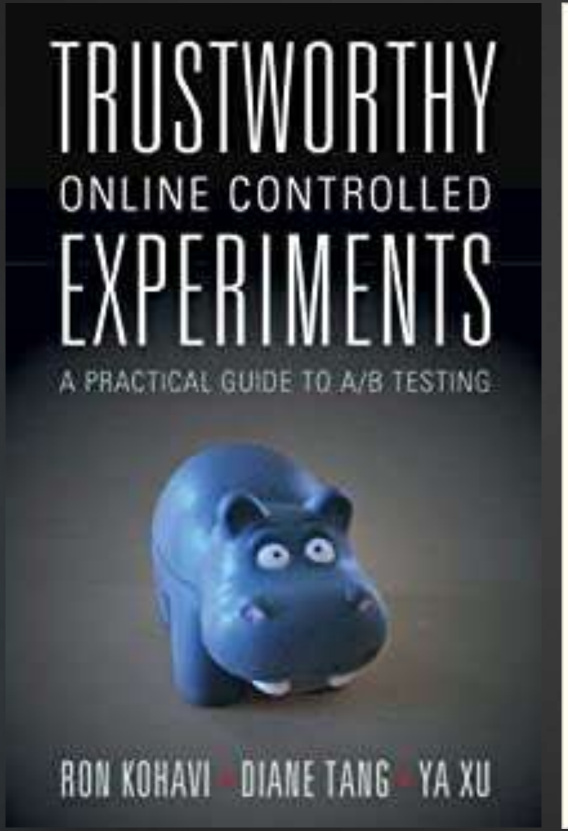
4,000 cities. As such, it has collected a diverse set of numerical, textual,
unstructured data, which our Data Science team mines for insights that will propel our community and product forward.

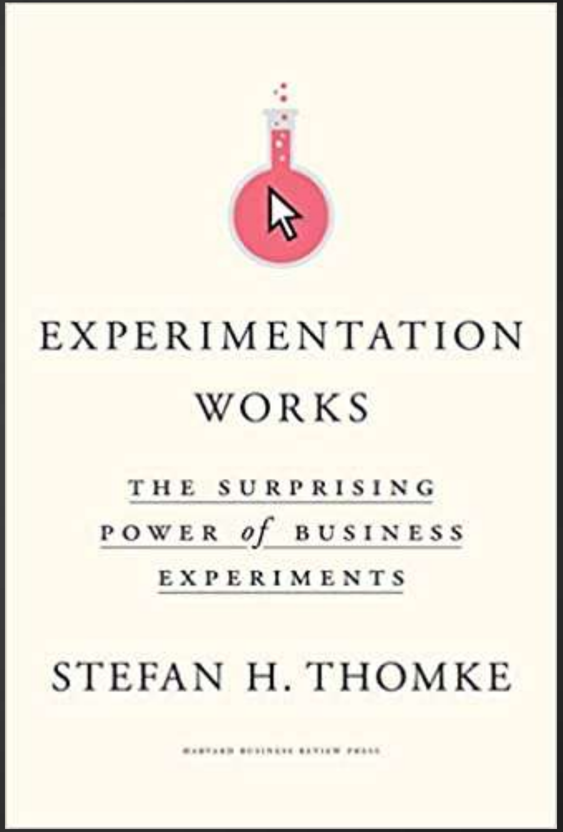
We are looking for experienced Data Scientists to join our Identity team (part of the broader Trust team) and expand upon the work we've done. B
begins with clear identity matching, and here are some examples of projects we currently need help with:

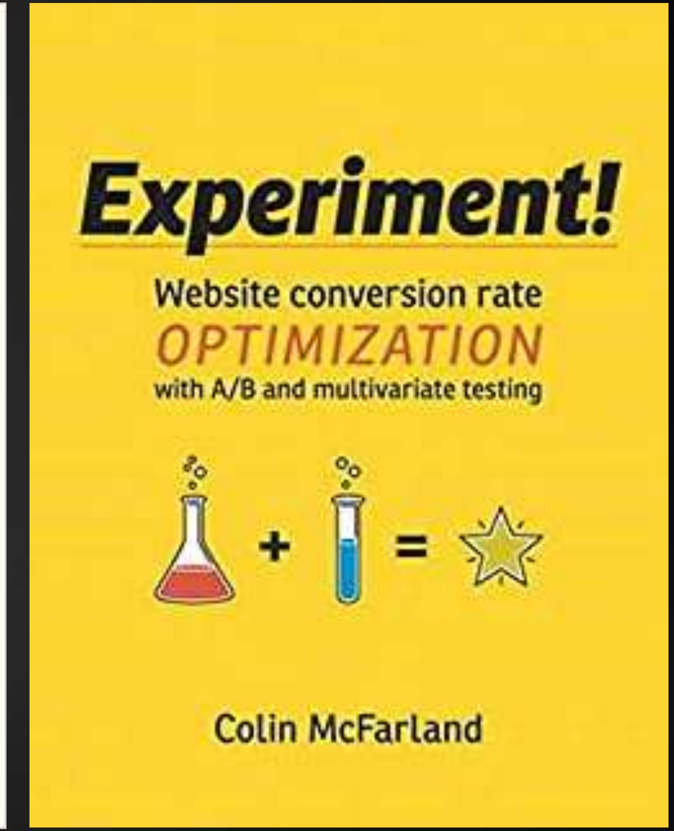
- Conduct rigorous A/B experiments in interlocking parts of our product where careful experimental design is required to ensure valid results.

What are OCEs?









Open OCE Problems

2021 Conference on Digital Experimentation @ MIT (CODE@MIT)

Join us virtually November 4-5 for the 8th annual Conference on Digital Experimentation @ MIT (CODE@MIT).

Registration is now open! Early bird pricing closes on October 4, 2021.

Please note: if you are submitting work to be considered to present and are selected, you will receive a gratis code to register free of charge for CODE 2021 along with your acceptance letter on September 30, 2021.

November 04 - 05, 2021

9:00 am - 6:00 pm EDT

Register Now

CODE @MIT

Open OCE Problems

Appears in the June 2019 issue of SIGKDD Explorations Volume 21, Issue 1
https://bit.ly/OCESummit1

Top Challenges from the first Practical Online Controlled Experiments Summit

Somit Gupta (Microsoft)[1], Ronny Kohavi (Microsoft)[2], Diane Tang (Google)[3], Ya Xu (LinkedIn)[4], Reid Andersen (Airbnb), Eytan Bakshy (Facebook), Niall Cardin (Google), Sumitha Chandran (Lyft), Nanyu Chen (LinkedIn), Dominic Coey (Facebook), Mike Curtis (Google), Alex Deng (Microsoft), Weitao Duan (LinkedIn), Peter Forbes (Netflix), Brian Frasca (Microsoft), Tommy Guy (Microsoft), Guido W. Imbens (Stanford), Guillaume Saint Jacques (LinkedIn), Pranav Kantawala (Google), Ilya Katsev (Yandex), Moshe Katzwer (Uber), Mikael Konutgan (Facebook), Elena Kunakova (Yandex), Minyong Lee (Airbnb), MJ Lee (Lyft), Joseph Liu (Twitter), James McQueen (Amazon), Amir Najmi (Google), Brent Smith (Amazon), Vivek Trehan (Uber), Lukas Vermeer (Booking.com), Toby Walker (Microsoft), Jeffrey Wong (Netflix), Igor Yashkov (Yandex)

ABSTRACT

Online controlled experiments (OCEs), also known as A/B tests, have become ubiquitous in evaluating the impact of changes made to software products and services. While the concept of online controlled experiments is simple, there are many practical challenges in running OCEs at scale and encourage further academic and industrial exploration. To understand the top practical challenges in running OCEs at scale, representatives with experience in large-scale experimentation from thirteen different organizations (Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yandex, and Stanford University) were invited to the first Practical Online Controlled Experiments Summit. All thirteen organizations sent representatives. Together these organizations tested more than one hundred thousand experiment treatments last year. Thirty-four experts from these organizations participated in the summit in Sunnyvale, CA, USA on December 13-14, 2018.

While there are papers from individual organizations on some of the challenges and pitfalls in running OCEs at scale, this is the first paper to provide the top challenges faced across the industry for running OCEs at scale and some common solutions.

1. INTRODUCTION

The Internet provides developers of connected software, including web sites, applications, and devices, an unprecedented opportunity to accelerate innovation by evaluating ideas quickly and accurately using OCEs. At companies that run OCEs at scale, the tests have very low marginal cost and can run with thousands to millions of users. As a result, OCEs are quite ubiquitous in the technology

1.1 First Practical Online Controlled Experiments Summit, 2018

To understand the top practical challenges in running OCEs at scale, representatives with experience in large-scale experimentation from thirteen different organizations (Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yandex, and Stanford University) were invited to the first Practical Online Controlled Experiments Summit. All thirteen organizations sent representatives. Together these organizations tested more than one hundred thousand experiment treatments last year. Thirty-four experts from these organizations participated in the summit in Sunnyvale, CA, USA on December 13-14, 2018. The summit was chaired by Ronny Kohavi (Microsoft), Diane Tang (Google), and Ya Xu (LinkedIn). During the summit, each company presented an overview of experimentation operations and the top three challenges they faced. Before the summit, participants completed a survey of topics they would like to discuss. Based on the popular topics, there were nine breakout sessions detailing these issues. Breakout sessions occurred over two days. Each participant could participate in at least two breakout sessions. Each breakout group presented a summary of their session to all summit participants and further discussed topics with them. This paper highlights top challenges in the field of OCEs and common solutions based on discussions leading up to the summit, during the summit, and afterwards.

1.2 Online Controlled Experiments

Online Controlled Experiments, A/B tests or simply experiments, are widely used by data-driven companies to evaluate the impact of

Open OCE Problems

Problems Highlighted in the Summit [10]:

1. Estimation of long-term effects
2. Estimation of heterogeneous treatment effects
3. Experimentation in the presence of network interference
4. Interacting experiments

Additional Problems:

5. Sequential experimentation
6. Non-identifiable experimental units
7. Post-selection inference
8. Causal inference via observational studies & Ethics

Open OCE Problems

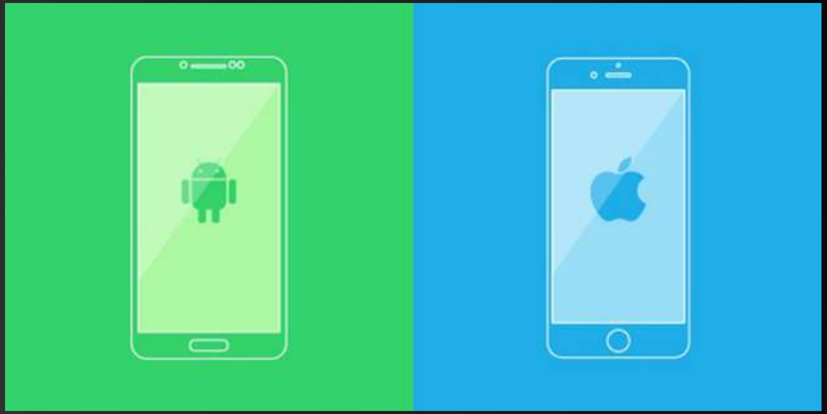
1. Estimation of long-term effects

- OCEs typically run for 2 weeks – how then can we estimate longer term treatment effects?
- Estimating long term effects is important to protect oneself from primacy and newness effects[11].
  - Primacy: When a change that proves to be better over time temporarily degrades performance to begin with
  - Newness: When a change that proves to be poor in the long run looks great initially
- Simply running the experiment longer is not typically a viable option.

Open OCE Problems

2. Estimation of heterogeneous treatment effects

- Treatment effects are rarely the same across all user segments
  - Market/ region
  - User activity level
  - Device type
  - Temporal windows



- Estimating these heterogenous treatment effects can be a challenge
  - Signal-to-noise is small
  - Multiple testing problem
  - Correlation vs. causation

Open OCE Problems

3. Experimentation in the presence of network interference

- How do you design and analyze an experiment when the Stable Unit Treatment Value Assumption (SUTVA) is violated?
- In this case, the treatment effect estimator is biased if the network effect is not appropriately accounted for.
- Google[12], Facebook[13], and LinkedIn[14] have ideas, but additional research is warranted.

Open OCE Problems

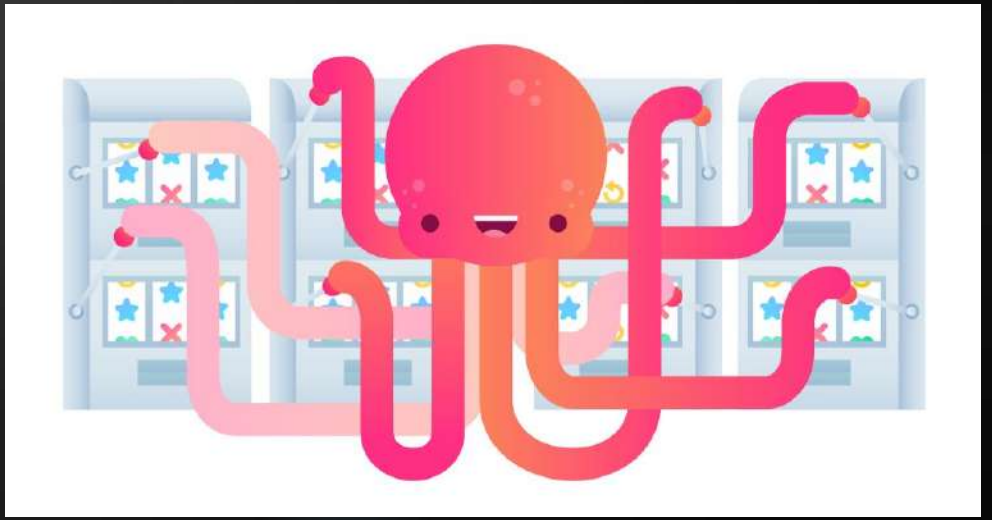
4. Interacting experiments

- Experimentally mature organizations are running 100s of experiments at the same time, sometimes independently trying to move the same metric.
- How do you know if the lift observed in your experiment is due to your treatment and not one in another team's experiment?
- Factorial designs are acknowledged as an obvious (albeit complicated) solution to this problem, but most practitioners seem to implement practical solutions that aim to limit the exposure of units to multiple experiments[4,15].

Open OCE Problems

5. Sequential experimentation

- Companies with large user bases have the capacity to engage millions of users in a single experiment with near real time data collection.
  - The volume and velocity of this experimental data – *when properly handled* – can facilitate expedited decision making by way of sequential experimentation.
- Methods like always valid p-values[16] and multi-armed bandit experiments[17,18] have become popular sequential alternatives to traditional experiments in which sample sizes are static and predetermined.



Open OCE Problems

6. Non-identifiable experimental units

- When experimental units are users, and randomization is cookie-based, cookie churn and private browsing can lead to individuals entering the experiment (and generating data) multiple times.
  - The same person could be in a single treatment more than once.
  - The same person could be in different treatments simultaneously.
  - Both of the above might happen.

- Different users might also use the same device, thereby contaminating the data.

Open OCE Problems

7. Post-selection inference

- Treatment effects observed in an experiment are rarely replicated when the experiment ends and the winning treatment is rolled out.
  - Berman et al.[19] estimate that on the order of 20-30% of tests result in false discoveries owing largely to mis-attributed true-null effects.
- This may be due in part to post-selection bias.
  - The problem is that if only statistically significant treatment effects are estimated, these will be biased upward.
  - In a one-sided Z-test, it can be shown (where δ is the true lift) that

$$\mathrm{E}[\bar{X}-\bar{Y} \mid \bar{X}-\bar{Y} \geq w] = \delta + \sigma \frac{\phi\left(\frac{w-\delta}{\sigma}\right)}{1-\Phi\left(\frac{w-\delta}{\sigma}\right)} > \delta$$

- Bias adjustment in this context is an important/ active research area[20].

Open OCE Problems

8. Causal inference via observational studies & Ethics

- Controlled experiments are not always ethical or even possible.
- This suggests that the causal inference literature has relevance.
  - For example, Mozilla recently used propensity score matching in an observational study to determine whether Firefox users that installed an ad blocker were more engaged with the browser[21].
- This raises the broader issue of ethics and fairness in online controlled experiments.
  - These are typically *human subjects trials* but they do not undergo the same scrutiny as in academia. Should there be better regulation in place?

Opportunities for Academia

Opportunities for Academia

Research

- The aforementioned open problems serve as exciting research opportunities.
- But useful innovation requires a synergistic bridge between academia and industry.
- The proprietary nature of most OCEs means that the problem, data, and methods are not often readily available to academics.
- This leads to a disconnect between academia and industry, and ultimately Type III Errors, where academics don’t fully appreciate the context or understand the problems.

Opportunities for Academia

Education

- This modern application area can "breath new life" into otherwise "stodgy" DOE courses.

- The non-triviality of the practical application of – and research in – online controlled experiments indicates the need for dedicated OCE courses in data science degree programs.

Summary

Summary

- Online controlled experiments are an exciting modern playground for design of experiment (DOE) researchers and practitioners
- Many novel practical challenges arise in this area that require innovative statistical solutions
- Industrial statisticians who are well-versed in the DOE literature should be involved
- Useful contributions from academia to industry will happen only with meaningful collaboration

References

1. Thomke S. (2020). *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Review Press.
2. Kohavi R. and Thomke S. (2017). The Surprising Power of Online Experiments. *Harvard Business Review*.
3. Kovahi R., Tang D. and Xu Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.
4. Xu, Y., Chen, N., Fernandez, A., Sinno, O., & Bhasin, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2227-2236).
5. Crunchbase: https://www.crunchbase.com/organization/optimizely#section-overview
6. Siroker D. and Koomen P. (2013). *A/B testing: The most powerful way to turn clicks into customers*. Wiley.
7. Xing, George, Lyft Head of Analytics (2016). Driving Growth Through Analytics https://www.meetup.com/USF-Seminar-Series-in-Data-Science/events/233118005/
8. Why Google has 200M reasons to put engineers over designers: https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers
9. Jay Larson, Optimizely CEO (2019). Opticon19 Keynote: https://blog.optimizely.com/2019/07/10/the-opticon19-agenda-is-here/

References

10. Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., ... & Yashkov, I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20-35.
11. McFarland C. (2013). *Experiment! Website conversion rate optimization with A/B and multivariate testing*. New Riders.
12. Yoon, S. (2018). Designing A/B tests in a collaboration network.
13. Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1).
14. Saint-Jacques, G., Varshney, M., Simpson, J., & Xu, Y. (2019). Using ego-clusters to measure network effects at LinkedIn. *arXiv preprint arXiv:1903.08755*.
15. Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 17-26).
16. Johari, R., Pekelis, L., & Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*.
17. Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, *26*(6), 639-658.

References

18. Scott, S. L. (2015). Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, *31*(1), 37-45.
19. Berman, R., & Van den Bulte, C. (2020). False Discovery in A/B Testing. *Preprint*.
20. Deng, A., Li, Y., Lu, J., & Ramamurthy, V. (2019). On Post-selection Inference in A/B Testing. *arXiv preprint arXiv:1910.03788*.
21. Miroglio, B., Zeber, D., Kaye, J., & Weiss, R. (2018). The effect of ad blocking on user engagement with the web. In *Proceedings of the 2018 world wide web conference* (pp. 813-821).