Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments

Xiaolin Shi
Snap Inc.
Santa Monica, CA
xiaolin@snap.com

Pavel Dmitriev
Outreach
Seattle, WA
pavel.dmitriev@outreach.io

Somit Gupta
Microsoft
Redmond, WA
somit.gupta@microsoft.com

Xin Fu
Facebook
Menlo Park, CA
xinfacebook@fb.com

ABSTRACT

A/B Testing is the gold standard to estimate the causal relationship between a change in a product and its impact on key outcome measures. It is widely used in the industry to test changes ranging from simple copy change or UI change to more complex changes like using machine learning models to personalize user experience. The key aspect of A/B testing is evaluation of experiment results. Designing the right set of metrics - correct outcome measures, data quality indicators, guardrails that prevent harm to business, and a comprehensive set of supporting metrics to understand the "why" behind the key movements is the #1 challenge practitioners face when trying to scale their experimentation program [18, 22]. On the technical side, improving sensitivity of experiment metrics is a hard problem and an active research area, with large practical implications as more and more small and medium size businesses are trying to adopt A/B testing and suffer from insufficient power. In this tutorial we will discuss challenges, best practices, and pitfalls in evaluating experiment results, focusing on both lessons learned and practical guidelines as well as open research questions.

CCS CONCEPTS

• General and reference~Surveys and overviews
• General and reference~Metrics
• General and reference~Evaluation
• General and reference~Experimentation
• Information systems~Data analytics
• Social and professional topics~Industry statistics
• Applied computing~Business intelligence
• Mathematics of computing~Probability and statistics

KEYWORDS

Controlled experiments; A/B testing; online metrics; user experience evaluation

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332297

ACM Reference format:
Xiaolin Shi, Pavel Dmitriev, Somit Gupta and Xin Fu. 2019. Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments. In *Proceedings of KDD'19, Aug 4, 2019, Anchorage, Alaska, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332297

1. Target audience and prerequisites for the tutorial

The tutorial does not assume any prior knowledge of A/B testing and open to all researchers and practitioners in the field of data mining, statistics and data analysis, program managers, and business leaders interested in learning how to design effective metrics, get the most value from A/B testing, and make better data-driven decisions. In the later parts of the tutorial, after introducing the necessary pre-requisites, several advanced topics will be covered that will be of interest to current researchers and practitioners of A/B testing. Basic knowledge of probability and statistics is a pre-requisite.

2. Tutors

Xiaolin Shi is a Manager of Applied Research and Data Science at Snap Inc., where she leads a team of scientists with expertise in data mining, machine learning, statistics, and economics. She has over ten years of academic and industrial experience in data science and big data, focusing on online experimentation and metrics, data mining, computational social science, and social network analysis. Xiaolin has published at top tier data mining and data science conferences such as KDD, WWW, WSDM, SIGIR, CIKM, and was the recipient of ACM Douglas Engelbart Best Paper Award (2008). Prior to Snap Inc., Xiaolin was at Stanford University, Microsoft, and Yahoo! Research. Xiaolin received her Ph.D. from the University of Michigan.

Pavel Dmitriev is a Vice President of Data Science at Outreach, where he works on enabling data driven decision making in Sales through machine learning and experimentation. He was previously a Principal Data Scientist with Microsoft's Analysis and Experimentation team, and a Researcher at Yahoo! Labs. Pavel has been working in the field of web mining, machine learning, and experimentation for over 15 years. He published a number of papers at top Data Mining conferences including KDD, WWW, CIKM, ICDM, BigData. He taught tutorials at KDD 2017, SIGIR 2017, and WWW 2010 conferences, gave a keynote

at SEAA 2018 conference, and was an invited lecturer at Russian Summer School on Information Retrieval in 2007 and 2009. Pavel received a Ph.D. degree in Computer Science from Cornell University in 2008, and a B.S. degree in Applied Mathematics from Moscow State University in 2002.

Somit Gupta is a Senior Data Scientist for Microsoft's Analysis and Experimentation team. He helps MSN, edge browser, Office and Windows to innovate faster with trustworthy experimentation. He helps the team run experiments at scale: defining the OEC for the product, design & monitoring of experiments, and interpretation of results to make a ship decision. He co-lectured a tutorial on A/B Testing at KDD 2017. Previously he was a product manager for Windows and Skype. Somit received a master's degree in Computer Algebra from University of Waterloo in 2011. Prior to that he received a bachelor's degree in Computer Engineering from National Institute of Technology, Surathkal, India.

Xin Fu is a Director of Data Science at Facebook where he leads a group of data scientists and engineers to measure and optimize performance, reliability and efficiency of Facebook's global infrastructure. Before Facebook, he was a Senior Director of Data Science at LinkedIn on consumer products, and prior to that, a data scientist at Google and Microsoft. Xin holds a PhD in Human-Computer Interaction from the University of North Carolina at Chapel Hill. He is a frequent speaker at data science and analytics conferences including teaching a data-driven product innovation tutorial at KDD in 2015.

3. Related Tutorials

As far as we know, the proposed tutorial in its current or close to its current form has not been presented to a large audience before. However, some parts of the tutorial are based on or draw from past tutorials, research talks, and lectures, and we describe them below.

Some topics about running A/B tests covered in this tutorial were also covered briefly in the tutorials by A. Deng, et al. [1,7,21] at KDD 2017 and SIGIR 2017, which focused on theory and practical lessons of running online experiments for mobile and online products at scale, and helping industrial teams and companies leading to better data-driven decisions. Two tutors of this tutorial were also part of the KDD 2017 and SIGIR 2017 tutorial. In part, this tutorial is a response to many questions we got from the participants of the 2017 tutorials about experiment evaluation and metric design which were only covered briefly in a 20-minute section. Another tutorial by Roman Budylin, et al. [4] at WWW 2018 covered online metrics in one section briefly. Our proposed tutorial will expand on this topic extensively.

This proposed tutorial will go in depth and include material from the most recent and state-of-the-art work about online experimentation and evaluation of experiments by the authors, such as [1-35]. Very positive response and requests for a more in-depth tutorial from those who attended the our technical talks over the past several years is one of the key motivations for the authors to submit this tutorial proposal. Although many of the state-of-the-art methodologies discussed in this tutorial involves advanced statistical techniques, the materials we are going to present have been adapted to not require advanced statistical knowledge as a prerequisite.