A/B Testing at Scale: Accelerating Software Innovation

Alex Deng
Analysis & Experimentation Team
Microsoft

Somit Gupta
Analysis & Experimentation Team
Microsoft

Pawel Janowski
Analysis & Experimentation Team
Microsoft

Ronny Kohavi
Analysis & Experimentation Team
Microsoft

Jeff Omhover
Analysis & Experimentation Team
Microsoft

Abstract

The Internet and the general digitalization of products and operations provides an unprecedented opportunity to accelerate innovation while applying a rigorous and trustworthy methodology for supporting key product decisions. Developers of connected software, including web sites, applications, and devices, can now evaluate ideas quickly and accurately using controlled experiments, also known as A/B tests. From front-end user-interface changes to backend algorithms, from search engines (e.g., Google, Bing, Yahoo!) to retailers (e.g., Amazon, eBay, Etsy) to social networking services (e.g., Facebook, LinkedIn, Twitter) to travel services (e.g., Expedia, Airbnb, Booking.com) to many startups, online controlled experiments are now utilized to make data-driven decisions at a wide range of companies. The theory of a controlled experiment is simple, but for the practitioner the deployment and evaluation of online controlled experiments at scale (100's of concurrently running experiments) across a variety of web sites, mobile apps, and desktop applications presents many pitfalls and new research challenges. In this tutorial, we will introduce the overall A/B testing methodology, walkthrough use cases using real examples, and then focus on practical and research challenges in scaling experimentation. We will share key lessons learned from scaling experimentation at Microsoft to thousands of experiments per year and outline promising directions for future work.

ACM Reference format:

Alex Deng, Somit Gupta, Pawel Janowski, Ronny Kohavi and Jeff Omhover. 2019. A/B Testing at Scale: Accelerating Software Innovation. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320093

WWW '19, May 13–17, 2019, San Francisco, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320093

1 Previous editions

The tutorial in its current or close to its current form has been presented to a large audience before at SIGIR 2017 [1] , KDD 2017, and STRATA 2018 [2]. Some topics covered in the tutorial were discussed in Ronny Kohavi's keynote talk at the KDD 2015 conference [3]. While the keynote covered a range of topics in a brief fashion, the tutorial will go in depth and will also include material from works such as [1] [8] [12] [7] [9] [4]. Requests for a more in-depth tutorial from those who attended the keynote and other conference talks the authors have given over the last two years is one of the key motivations for us to submit this tutorial proposal. Parts of the tutorial are based on the "Introduction to Experimentation" training course Analysis and Experimentation team conducts internally at Microsoft on a monthly basis. Some of the advanced statistical techniques discussed in this tutorial were also presented in a tutorial by Deng et. al at JSM 2015 conference. However, the material has been adapted to not require advanced statistical knowledge as a prerequisite. A tutorial on experimentation on the web was given by Ronny Kohavi at el at KDD 2009 [6]. Since that time, both the theory of online A/B testing and its use in practice have evolved greatly, and the overlap of our current tutorial proposal with that one is not large. Slides and videos from some of the past talks given by authors can be viewed at http://exp-platform.com/talks/

References

[1] P. D. S. G. R. K. P. R. a. L. V. Alex Deng, "A/B Testing at Scale: Accelerating Software Innovation," in *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*, Shinjuku, Tokyo, Japan , 2017.

[2] R. K. A. D. P. R. Somit Gupta, "A/B Testing at Scale Tutorial," March 2018. [Online]. Available: https://exp-platform.com/2018StrataABtutorial/.

[3] R. Kohavi, "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years," Conference on Knowledge Discovery and Data Mining (KDD).

[4] A. Fabijan, P. Dmitriev, H. Holmstrom and J. Bosch, "The Evolution of Continuous Experimentation in Software Product Development," in *International Conference on Software Engineering (ICSE)*, 2017.

[5] A. Deng and X. Shi, "Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[6] R. Kohavi, R. Longbotham and J. Quarto-vonTivadar, "Planning, Running, and Analyzing Controlled Experiments on the Web," in *tutorial at Conference on Knowledge Discovery and Data Mining*, 2009.

[7] R. Kohavi, "Pitfalls in Online Controlled Experiments," in *MIT COnference on Digital Experimentation (CODE)*, 2016.

[8] Z. Zhao, M. Chen, D. Matheson and M. Stone, "Online Experimentation Diagnosis and Troubleshooting Beyond AA Validation," in *Conference on Data Science and Advanced Analytics*, 2016.

[9] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker and Y. Xu, "Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

[10] A. Deng, Y. Xu, R. Kohavi and T. Walker, "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data," in *Conference on Web Search and Data Mining (WSDM)*, 2013.

[11] P. Dmitriev and X. Wu, "Measuring Metrics," in *Conference on Information and Knowledge Management (CIKM)*, 2016.

[12] W. Machmouchi and G. Buscher, "Principles for the Design of Online A/B Metrics," in *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2016.

[13] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu and N. Pohlmann, "Online Controlled Experiments at Large Scale," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[14] A. Deng, "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments," in *World Wide Web Conference (WWW)*, 2015.

[15] A. Deng, J. Lu and S. Chen, "Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing," in *Conference on Data Science and Advanced Analytics*, 2016.

[16] A. Deng, P. Zhang, S. Chen, D. Kim and J. Lu, "Concise Summarization of Heterogeneous Treatment Effect Using Total Variation Regularized Regression," in *In submission*, 2017.

[17] R. Kohavi, "SPEED MATTERS: A SLOWDOWN EXPERIMENT," [Online]. Available: https://1drv.ms/w/s!AuRxCGEOCRKGkbtoabZo0OcKoo6xhw.