A/B Testing at Scale: Accelerating Software Innovation

Alex Deng
Microsoft Corporation
One Microsoft Way, Redmond WA
98052, USA
alexdeng@microsoft.com

Pavel Dmitriev
Microsoft Corporation
One Microsoft Way, Redmond WA
98052, USA
padmitri@microsoft.com

Somit Gupta
Microsoft Corporation
One Microsoft Way, Redmond WA
98052, USA
somit.gupta@microsoft.com

Ron Kohavi
Microsoft Corporation
One Microsoft Way, Redmond WA
98052, USA
ronnyk@microsoft.com

Paul Raff
Microsoft Corporation
One Microsoft Way, Redmond WA
98052, USA
paraff@microsoft.com

Lukas Vermeer
Booking.com
Herengracht 597, 1017 CE Amsterdam
Netherlands
lukas.vermeer@booking.com

ABSTRACT

TheInternet provides developers of connected software, including web sites, applications, and devices, an unprecedented opportunity to accelerate innovation by evaluating 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. While the theory of a controlled experiment is simple, and dates back to Sir Ronald A. Fisher's experiments at the Rothamsted Agricultural Experimental Station in England in the 1920s, the deployment and evaluation of online controlled experiments at scale (100's of concurrently running experiments) across variety of web sites, mobile apps, and desktop applications presents many pitfalls and new research challenges.

In this tutorial we will give an introduction to A/B testing, share key lessons learned from scaling experimentation at Bing to thousands of experiments per year, present real examples, and outline promising directions for future work. The tutorial will go beyond applications of A/B testing in information retrieval and will also discuss on practical and research challenges arising in experimentation on web sites and mobile and desktop apps.

Our goal in this tutorial is to teach attendees how to scale experimentation for their teams, products, and companies, leading to better data-driven decisions. We also want to inspire more academic research in the relatively new and rapidly evolving field of online controlled experimentation.

1 PRESENTERS

Alex Deng is a Principal Data Scientist Manager on the Microsoft Analysis and Experimentation Team. He and his team work on methodological improvements of the experimentation platform as well as related engineering challenges. His works in this area are published in conference proceedings like KDD, WWW, WSDM and other statistical journals. He co-lectured a tutorial on A/B Testing at JSM 2015. Alex received a Ph.D. degree in Statistics from Stanford University in 2010 and a B.S degree in Mathematics from Zhejiang university in 2006.

Pavel Dmitriev is a Principal Data Scientist with Microsoft's Analysis and Experimentation team. He was previously a Researcher at Yahoo! Labs. Pavel has been working in the field of web mining, search, and experimentation for close to 15 years. He published a number of papers at top Data Mining conferences including KDD, WWW, CIKM, ICDM, BigData. He was an invited lecturer at Russian Summer School on Information Retrieval in 2007 and 2009, taught tutorials at WWW 2010 and SIGIR 2011, and was an invited speaker at University of Pittsburgh Big Data colloquium in 2016. 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 Data Scientist for Microsoft's Analysis and Experimentation team. He helps MSN and edge browser content teams 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. 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.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5022-8/17/08…$15.00
http://dx.doi.org/10.1145/3077136.3082060

Ronny Kohavi is a Microsoft Distinguished Engineer and the General Manager for Microsoft's Analysis and Experimentation team at Microsoft's Artificial Intelligence and Research group. He was previously Partner Architect at Bing, part of the Online Services Division at Microsoft. He joined Microsoft in 2005 and founded the Experimentation Platform team in 2006. He was previously the director of data mining and personalization at Amazon.com, and the Vice President of Business Intelligence at Blue Martini Software, which went public in 2000, and later acquired by Red Prairie. Prior to joining Blue Martini, Kohavi managed MineSet project, Silicon Graphics' award-winning product for data mining and visualization. He joined Silicon Graphics after getting a Ph.D. in Machine Learning from Stanford University, where he led the MLC++ project, the Machine Learning library in C++ used in MineSet and at Blue Martini Software. Kohavi received his BA from the Technion, Israel. His papers have over 31,000 citations and three of his papers are in the top 1,000 most-cited papers in Computer Science.

Paul Raff is a Principal Data Scientist Lead in Microsoft's Analysis and Experimentation team. Previously, he was a Supply Chain Researcher at Amazon. Paul and his team work to enable scalable experimentation in varied teams around Microsoft, including Windows 10, Office Online, Exchange Online, and Cortana. Additionally, his team focuses on experiment quality, ensuring that all experiments are operating as intended an in a way that allows for the appropriate conclusions to be made. Paul received a Ph.D. degree in Mathematics from Rutgers University in 2009, and prior to that received degrees in Mathematics and Computer Science from Carnegie Mellon University.

Lukas Vermeer combines industry experience in online experimentation with an academic background in computing science and machine learning. He explains science using historical narratives and teach statistics through storytelling. At Booking.com, the world's leading accommodation website, Lukas is responsible for the internal tooling and training that helps product development validate improvements to the customer experience through experimentation. Booking.com has been running online experiments for over a decade and has well over a thousand concurrent experiments running at any given time.

2 OBJECTIVES

In this tutorial we will give an introduction to A/B testing, share key lessons learned from scaling experimentation at Bing to thousands of experiments per year, present real examples, and outline promising directions for future work. The tutorial will go beyond applications of A/B testing in information retrieval and will also discuss on practical and research challenges arising in experimentation on web sites and mobile and desktop apps.

Our goal in this tutorial is to teach attendees how to scale experimentation for their teams, products, and companies, leading to better data-driven decisions. We also want to inspire more academic research in the relatively new and rapidly evolving field of online controlled experimentation.

3 RELEVANCE TO INFORMATION RETRIEVAL COMMUNITY AND RELATED TUTORIALS AND TALKS

Controlled experiments are more and more commonly used for evaluating performance of information retrieval systems, including web search, product search, and e-mail search, in addition to offline measures such as precision, recall, NDCG. Industry leaders such as Bing, Google, Amazon are running thousands of experiments per year, striving to build a culture where any change, even a bug fix, is evaluated via an experiment. Therefore, we believe the topic of online experimentation is relevant to anyone working in information retrieval domain. The tutorial will discuss many real example experiments from the information retrieval domain, dealing with issues such as handling bots, developing Overall Evaluation Criteria (OEC) for a search engine, detecting interaction between search experiments, etc.

We also believe the information retrieval community will benefit from learning about more general topics in A/B testing: the currently active research areas, such as heterogeneous treatment effect and alternatives to null hypothesis testing, open research questions arising from violations of basic assumptions of A/B experiment design in practice, as well as the challenges arising when running experiments in mobile and desktop apps.

Some topics covered in the tutorial were discussed in Ronny Kohavi's keynote talk at the KDD 2015 conference [1]. While the keynote covered a range of topics in a brief fashion, the tutorial will go in depth and will also include material from recently published works such as [2] [3] [4] [5] [6] [7]. 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.

A tutorial on experimentation on the web was given by Ronny Kohavi at el at KDD 2009 [8]. 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://www.exp-platform.com.

3 FORMAT AND DETAILED SCHEDULE

1. Introduction to A/B testing (40 min)
    a. What is A/B testing
    b. Brief history
    c. Why use A/B testing
    d. Examples

2. Evolution of experimentation in product development [1] (10 min)
    a. Stages a company or product go through as they ramp up experimentation

   b. Technical and cultural challenges for each stage
   c. Ethical Considerations

3. Design of Experiments (before: Statistical Foundations) (20 min)
   a. Null-hypothesis testing, confidence intervals, p-values
   b. Central Limit Theorem
   c. Power analysis

4. Challenges in scaling A/B testing: from a handful of experiments/year to multiple experiments/day

   a. Ensuring Trustworthiness [2] [3] [4] (35 min)
      i. Importance and example issues
         1. Data Quality (e.g. click reliability)
         2. Impact of Bots
         3. Carry-over effects
         4. Random imbalance
         5. Data Loss
         6. Simpson's paradox
      ii. Solutions
         1. A/A tests
         2. Sample Ratio Mismatch tests
         3. Dealing with carry-over effects and random imbalance
         4. Twyman's law
      iii. Violations of Assumptions in Practice
         1. Unstable user identifiers (cookie churn and clobbering, multiple devices)
         2. Leaks due to shared resources
         3. Network interactions resulting in spill-over effects

   b. Designing Metrics (25 min)
      i. Metrics taxonomy, importance of good Overall Evaluation Criteria (OEC), search engine OEC example [5]
      ii. Designing a good OEC [6] [7] [8] [9]
         1. Triggering & counterfactual logging
         2. Variance reduction
         3. Metric transformations & surrogates, principles and pitfalls in metric design
         4. Metric evaluation based on experiment corpus

   c. Protecting Users [10] (15 min)
      i. Start small then ramp up
      ii. Near-real-time detection and shut down of bad experiments
      iii. Interaction prevention via isolation groups
      iv. Interaction detection

   d. Active Research Areas and Recent Developments (30 min)
      i. Issues with Null Hypothesis Testing, Bayesian alternative [11] [12]
      ii. Heterogeneous Treatment Effects [13]

5. Summary and Open Challenges (5 min)

4 SUPPORT MATERIALS

Slides from the tutorial will be made available to the attendees.

REFERENCES

[1] R. Kohavi, "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years," in *Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.