



# Trustworthy analyses of online A/B tests

Jiannan Lu

Analysis and Experimentation, Microsoft



# Disclaimer

- This will be a “less technical” talk.
- Outline:
  - Introduction;
  - Overview of A/B testing @MSFT;
  - Three (statistical) stories in experimentation.
- Some details are ignored, but references are provided.

# Introduction

- My [team](#) – making MSFT data-driven, via experimentation:



- [Me](#) – PhD (Harvard, 2015), working as DS/researcher

# Overview of A/B testing

- Most ideas are, frankly, mediocre or terrible.



- Let the (experimental) data speak.

# Metric development

- User engagement/satisfaction vs. revenue<sup>[1]</sup>:

The image shows a Bing search results page for the query 'flowers'. At the top, there are navigation links for WEB, IMAGES, VIDEOS, MAPS, SHOPPING, LOCAL, NEWS, and MORE. The search bar contains the text 'flowers' and a magnifying glass icon. Below the search bar, it says '358,000,000 RESULTS'. The main content area displays several search results, all of which are advertisements. The first ad is for 'Flowers at 1-800-FLOWERS®' from 1800Flowers.com, with the text 'Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now'. The second ad is for 'FTD® - Flowers' from www.FTD.com, with the text 'Get Same Day Flowers in Hours! Buy Now for 25% Off Best Sellers.'. The third ad is for 'Send Flowers from \$19.99' from www.ProFlowers.com, with the text 'Send Roses, Tulips & Other Flowers. "Best Value" -Wall Street Journal. proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)'. The fourth ad is for '50% Off All Flowers' from www.BloomsToday.com, with the text 'All Flowers on the Site are 50% Off. Take Advantage and Buy Today!'. There are two red arrows pointing to the second and third ads. One arrow points from the text 'Gotta make money...' to the 'FTD® - Flowers' ad. The other arrow points from the text 'Don't anger the users...' to the 'FTD® - Flowers' ad.

Gotta make money...

Don't anger the users...

[1] Dmitriev, P. and Wu, X. Measuring metrics. CIKM'16

# Metric development

- Quick back rate (QBR):

	Treatment	Control	Delta	Delta %	P-Value
QuickBack					
QuickBack Rate v2				+0.23%	2e-32
--Web Result				+0.79%	≈0
--Answer				+1.48%	≈0
--Ad				-1.10%	≈0

- The trade-off:

Estd Revenue/UU	-2.17%	≈0
-----------------	--------	----

# Success criterion

- A valuable experiment:
  - a. Confirm great feature;
  - b. Prevent supposedly great but actually meh feature;
  - c. Prevent bad feature;
  - d. Discover supposedly meh but actually great feature.

# Example from Bing

- Longer title for ads<sup>[2]</sup>:

Control – existing display

bing MS Beta

flowers

358,000,000 RESULTS

**Flowers at 1-800-FLOWERS®** 1800Flowers.com  
Fresh **Flowers** & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

**FTD® - Flowers** www.FTD.com  
**Get Same Day Flowers in Hours!** Buy Now for 25% Off Best Sellers.

**Send Flowers from \$19.99** www.ProFlowers.com  
Send Roses, Tulips & Other **Flowers**. "Best Value" -Wall Street Journal.  
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

**50% Off All Flowers** www.BloomsToday.com  
All **Flowers** on the Site are 50% Off. Take Advantage and Buy Today!

Treatment – long titles

bing MS Beta

flowers

358,000,000 RESULTS

**FTD® - Flowers - Get Same Day Flowers in Hours!** www.FTD.com  
Buy Now for 25% Off Best Sellers.

**Flowers at 1-800-FLOWERS® | 1800flowers.com** 1800Flowers.com  
Fresh **Flowers** & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

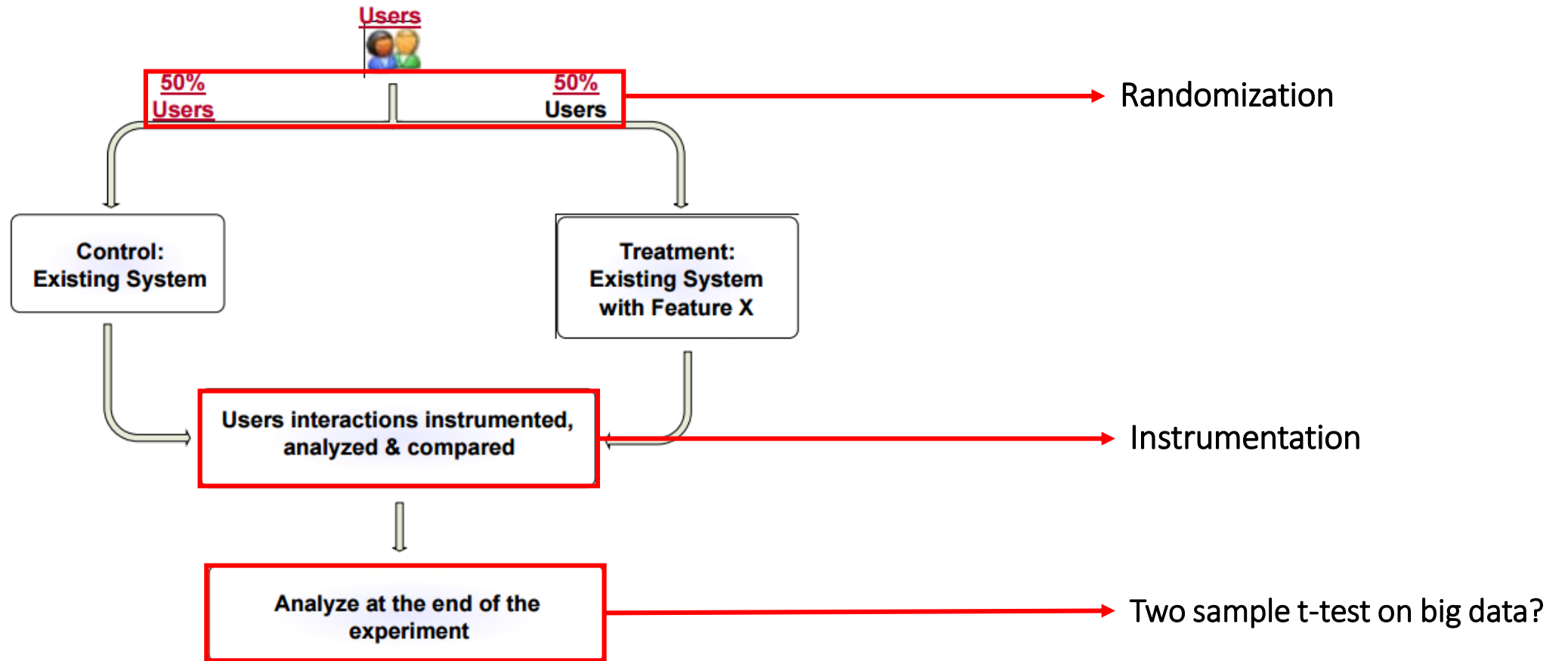
**Send Flowers from \$19.99 - Send Roses, Tulips & Other Flowers.** www.ProFlowers.com  
"Best Value" -Wall Street Journal.  
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

**\$19.99 - Cheap Flowers - Delivery Today By A Local Florist!** www.FromYouFlowers.com  
Shop Now & Save \$5 Instantly.

[2] Deng, A. et al. A/B Testing at Scale: Accelerating Software Innovation. SIGIR'17



# Experimentation pipeline<sup>[3]</sup>



[3] Kohavi, R. and Longbotham, R. Online Controlled Experiments and A/B Tests. Encyclopedia of Machine Learning and Data Mining. ISBN: 978-1-4899-7502-7.

# Story I: Calculating variances

# Background

- In Bing, each user views multiple pages;
- Randomization unit (R):
  1. By user: consistent experience;
  2. By page: larger sample size.
- Analysis unit (A):
  1. Page-level metric (business consideration).

# Notations

- Number of users  $n$ ;
- Randomization probability  $p$ ;
- At user-level ( $i = 1, \dots, n$ ):
  - Treatment assignment:  $W_{ij} = 1$  if treated;
  - Numbers of treated/control/total calls:  $N_{iT}, N_{iC}, N_i$
  - Observed outcomes:  $Y_{ij}^{\text{obs}}$  ( $j = 1, \dots, N_i$ );
  - Sums of treated/control:  $S_{iT}, S_{iC}$
- Estimator:

$$\hat{\tau} = \frac{\sum_{i=1}^n S_{iT}}{\sum_{i=1}^n N_{iT}} - \frac{\sum_{i=1}^n S_{iC}}{\sum_{i=1}^n N_{iC}}$$

Treatment mean  $\bar{Y}_T^{\text{obs}}$

Control mean  $\bar{Y}_C^{\text{obs}}$

# Variance formulas

- R = page – standard method:

$$\widehat{\text{Var}}_S(\widehat{\tau}) = \lambda_T^2 + \lambda_C^2,$$

where

$$\lambda_T^2 = \frac{1}{(\sum_{i=1}^n N_{iT})^2} \sum_{i=1}^n \sum_{j: W_{ij}=1} (Y_{ij}^{\text{obs}} - \bar{Y}_T^{\text{obs}})^2.$$

$$\lambda_C^2 = \frac{1}{(\sum_{i=1}^n N_{iC})^2} \sum_{i=1}^n \sum_{j: W_{ij}=0} (Y_{ij}^{\text{obs}} - \bar{Y}_C^{\text{obs}})^2.$$

- R = user – Delta method:

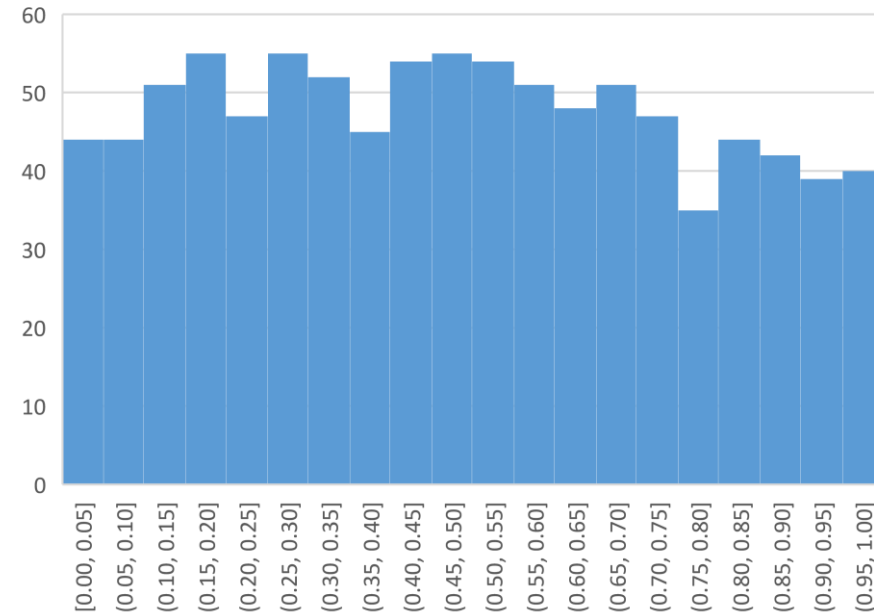
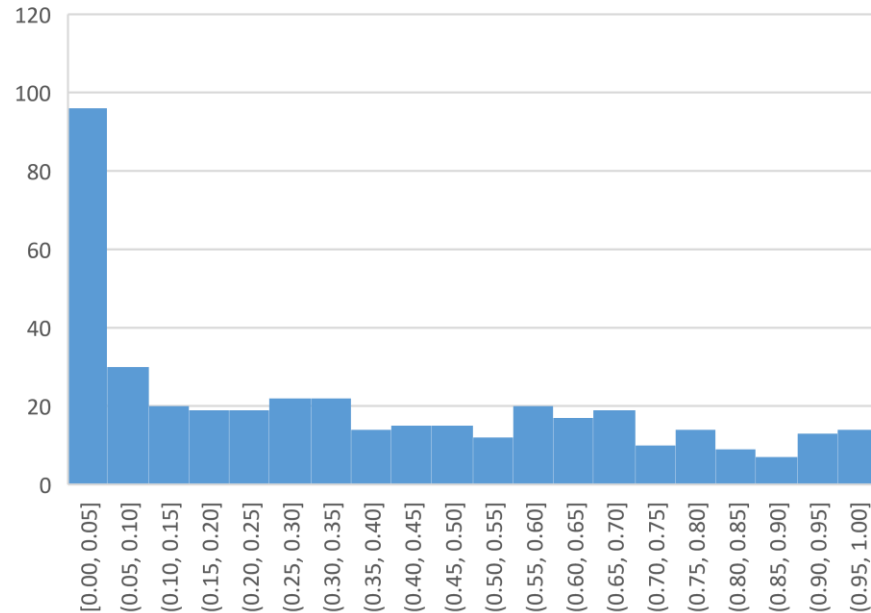
$$\widehat{\text{Var}}_D(\widehat{\tau}) = n^{-1}(\xi_T^2 + \xi_C^2),$$

where

$$\xi_T^2 = \frac{1}{(\widehat{\text{E}}N_{iT})^2} \widehat{\text{Var}}(S_{iT}) + \frac{(\widehat{\text{E}}S_{iT})^2}{(\widehat{\text{E}}N_{iT})^4} \widehat{\text{Var}}(N_{iT}) - 2 \frac{\widehat{\text{E}}S_{iT}}{(\widehat{\text{E}}N_{iT})^3} \widehat{\text{Cov}}(S_{iT}, N_{iT})$$

# Results

- P-values for A/A experiments:

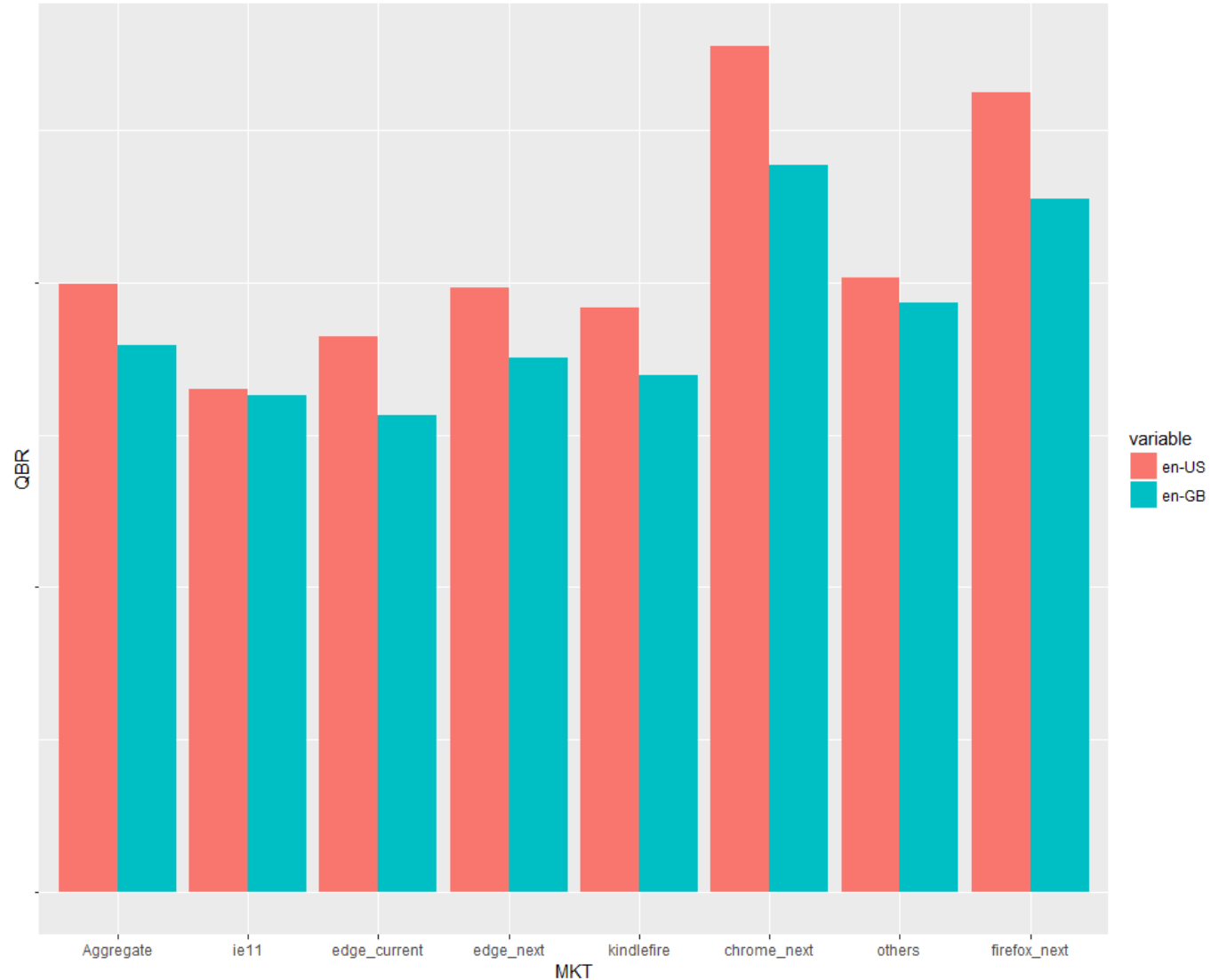


- Actually, there is more – See [Deng, Lu and Litz \(WSDM'17\)](#).

Story II:  
Finding heterogeneity

# Example

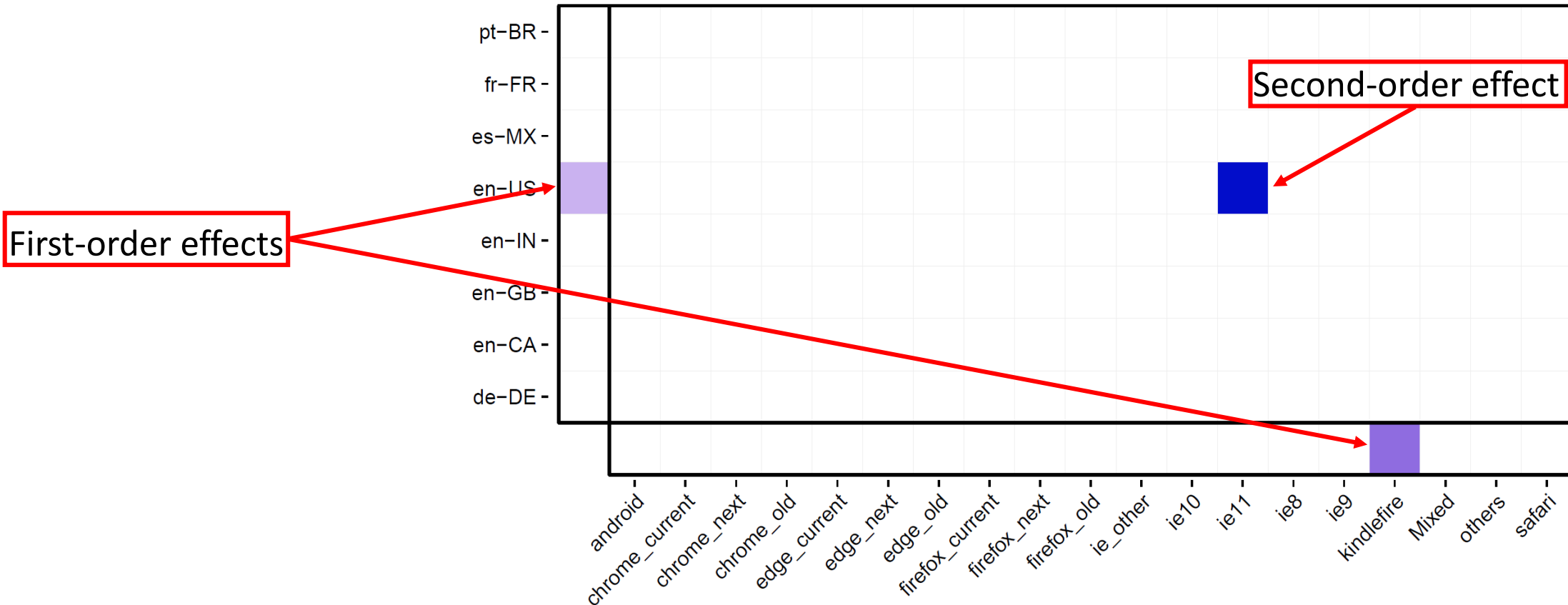
- User behaviors vary across different segments.
- “Personalized” treatment seems necessary.





# Results

- A Lasso-type solution<sup>[4]</sup>:

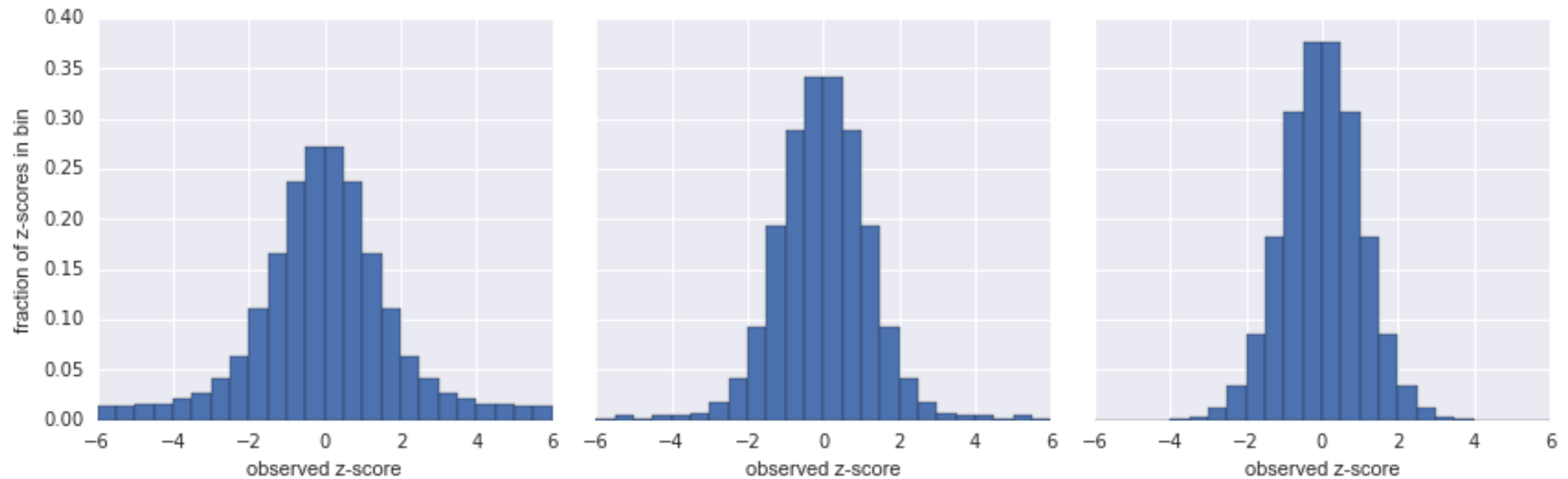


[4] Deng, A. et al. (2016) Concise summarization of heterogeneous treatment effect using total variation regularized regression. ArXiv:1610.03917

Story #3:  
(Try to) be Bayesian

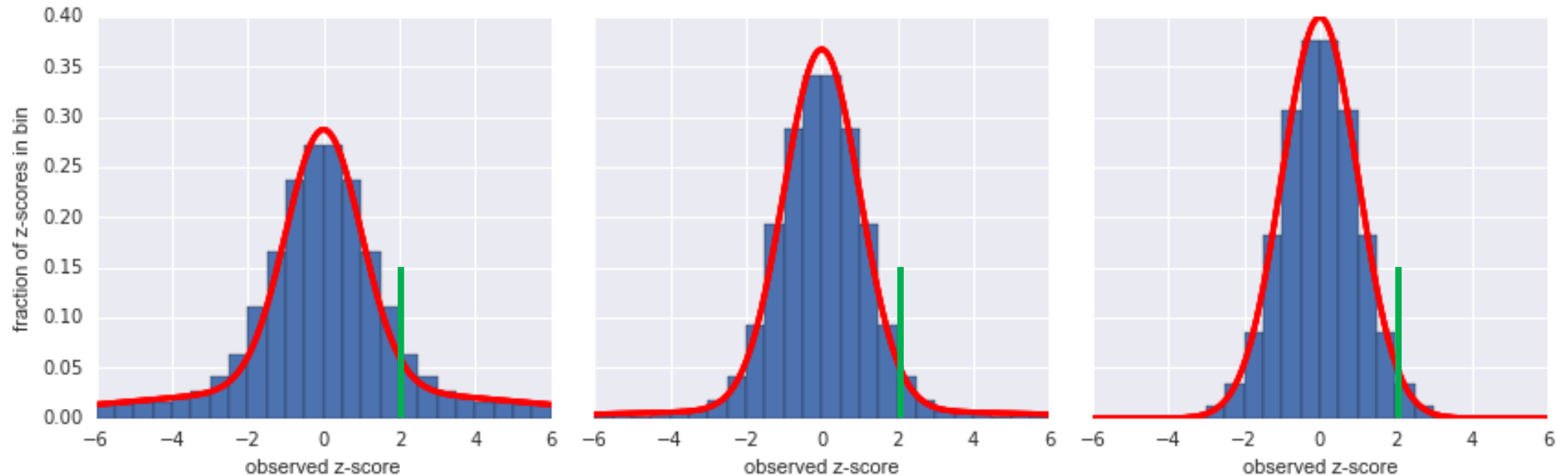
# Background

- NOT all metrics are created equal:
  1. Prior information can impact the interpretation of new results;
  2. We want **p-move**:  $\Pr(\text{true move} | \text{data})$ .



# Results

- Classic two-group model<sup>[5]</sup>:



p-move = 47.6%

p-move = 13.1%

p-move = 1.37%

- Small p-value might NOT indicate real movements.

[5] Efron, B. Microarrays, empirical Bayes and the two-group model. Statistical Science.

# Concluding remarks

- Experimentation is at the front-line of technology innovation;
- Trustworthiness is the foundation of experimentation;
- Principled statistical thinking is critical in the age of “Big Data” and “Machine Learning.”