

A/B Testing

Jaime Arguello

INLS 509: Information Retrieval

jarguell@email.unc.edu

- **Credits:** these slides borrow heavily from examples and figures from Ron Kohavi's presentations on A/B testing at Microsoft (available online)

Introduction

- Systems (e.g., search systems) are always trying to improve
- **Basic question:** If a specific change is introduced, will it improve key metrics?
- **Metrics:** measures that are believed to be correlated with the quality of the user experience
- Metrics are often things we want to minimize or maximize
- Examples?

A/B Testing

- Experiments where different populations of users are exposed to different versions of the system for a period of time
- **Control group:** group of users exposed to the “normal” or “baseline” version of the system
- **Experimental group:** group of users exposed to the experimental version of the system
- More often A/B/C/D/E... testing
- Search companies can have about 15 different A/B tests happening at once
- $5^{15} = 30,517,578,125$

The Alternative

- Make the change and measure the same metrics.
- Why is this a bad idea?

The Alternative

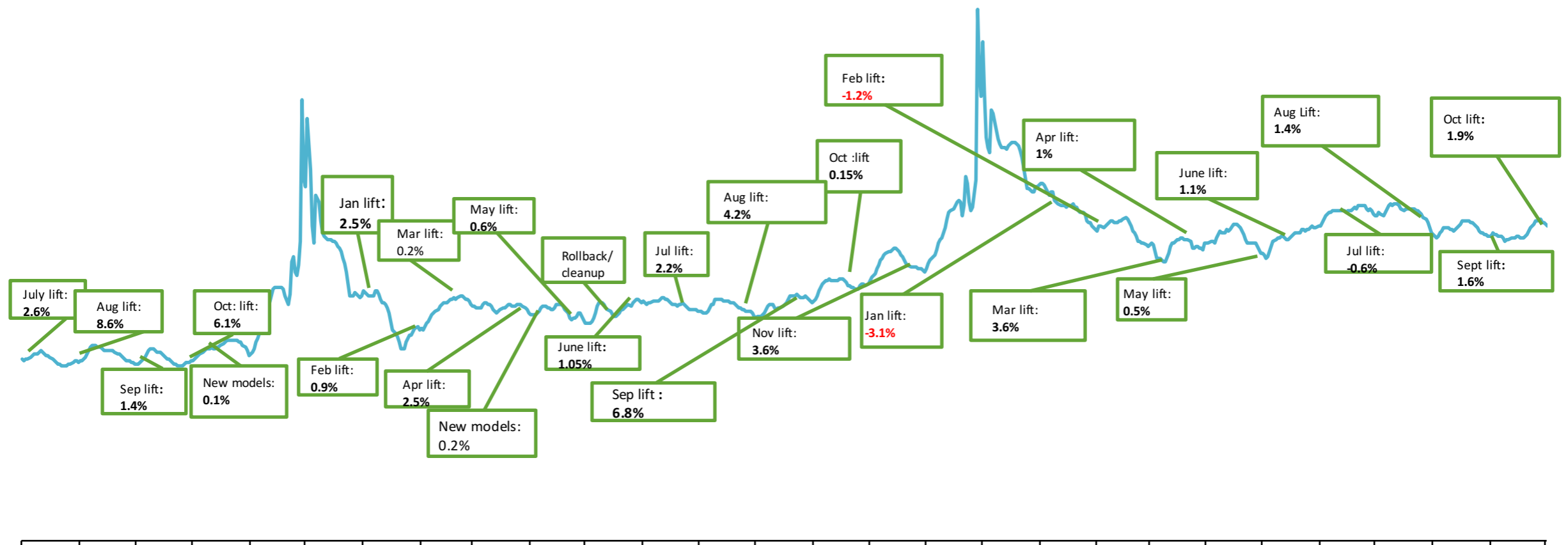
- Make the change and measure the same metrics.
- Why is this a bad idea?
 1. Temporal changes
 2. Good features lead to incremental improvements
 3. It's difficult to assess the value of ideas

Temporal Changes



Source: <http://exp-platform.com/2017abtestingtutorial/>

Temporal Changes + Incremental Improvements



Source: <http://exp-platform.com/2017abtestingtutorial/>

Predicting the value of new features

- 1/3 of ideas improve the intended metric(s)
- 1/3 of ideas have no effect
- 1/3 of ideas degrade the intended metric(s)

Source: <http://exp-platform.com/2017abtestingtutorial/>

Predicting the value of new features

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!

bing MS Beta

flowers

358,000,000 RESULTS

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

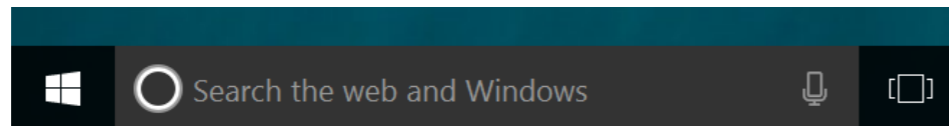
Flowers at 1-800-FLOWERS® | 1800flowers.com 1800Flowers.com
Fresh Flowers & Gifts at 1-800-FLOWERS. 100% Smile Guarantee. Shop Now

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

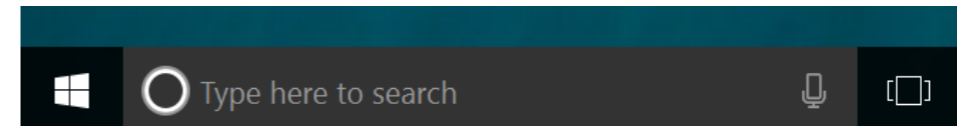
\$19.99 - Cheap Flowers - Delivery Today By A Local Florist! www.FromYouFlowers.com
Shop Now & Save \$5 Instantly.

Source: <http://exp-platform.com/2017abtestingtutorial/>

(1) Predicting the value of new features



A



B

- Overall Evaluation Criterion: no. of searches
- $A > B$, $A < B$, or $A = B$?

Source: <http://exp-platform.com/2017abtestingtutorial/>

(2) Predicting the value of new features

The image shows a Bing search results page for the query "kdd 2015". The search bar at the top contains "kdd 2015" and the Bing logo. Below the search bar, there are navigation tabs for "Web", "Images", "Videos", "Maps", "News", and "Explore". The user's name "Ronny" and a profile picture are visible in the top right corner. The search results are displayed in a list format, with the first result highlighted in blue. The results include:

- 1 KDD 2015, 10-13 August 2015, Sydney**
www.kdd.org/kdd2015
KDD 2015 is a premier conference that brings together researchers and practitioners from data mining, knowledge discovery, data analytics, and big data.
You've visited this page before - See search history
Sub-sections: Research Track, Sponsorship, Tutorials, Kdd-2014, Attending, Organisers.
- 2 KDD 2015 - The 21th ACM SIGKDD International Conference ...**
conference.researchbib.com/view/event/33616
KDD 2015 - The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
- 3 KDD CUP 2015**
https://www.kddcup2015.com
If you have any questions or comments, please send an email to support@kddcup2015.com. Updates: 1) Many people have asked the definition of ...
- 4 KDD 2015 : ACM SIGKDD Conference on Knowledge Discovery ...**
myhuiban.com/conference/136
The Latest Computer Conference and Journal List ... KDD 2015 : ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- 5 KDD 2015 -ACM SIGKDD International Conference on ...**
www.ourglocal.com/wikicfp/?conid=37&year=2015
KDD 2015 - ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Send this CFP to us by mail: cfp@ourglocal.org. Introduction: SIGKDD aims to ...
- 6 KDD-2015 Call for Papers, Workshop proposals - KDnuggets**
www.kdnuggets.com/2015/01/kdd-2015-call-papers.html
ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD) 2015 will be held in Sydney, Australia during August 10-13, 2015. KDD invites submissions of ...
- 7 KDD 2015 | 21st ACM SIGKDD Conference on Knowledge ...**
eventegg.com/kdd-2015
KDD 2015, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Australia, Sydney, 10 - 13 August 2015
- 8 KDD 2015 : 21th ACM SIGKDD Conference on Knowledge ...**
www.wikicfp.com/cfp/servlet/event.showcfp?eventid=40581
KDD 2015 : 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Home. Login; Register; Account; Logout; Categories CFPs. Post a CFP; Conf ...

At the bottom of the search results, there is a pagination bar with numbers 1 through 5 and a right arrow icon.

KDD 2015
We invite submission of papers describing innovative research on all aspects of knowledge discovery and data mining, ranging from theoretical foundations to novel models and algorithms for data mining problems in science, business, medicine, and engineering. Visionary papers on new and emerging topics are also welcome, as are appl...
wikicfp.com
Dates: Aug 10 - 13, 2015
Location: Sydney
Subjects: Data mining - Database - Knowledge extraction
Website: KDD 2015
Submissions due: Feb 20, 2015

People also search for
ICDM 2015 (Nov 14, 2015)
CIKM 2015 (Oct 19, 2015)
ICML 2015 (Jul 06, 2015)
AAAI 2016 (Feb 12, 2016)
WWW 2015 (May 20, 2015)
See more

Data from: Wikicfp.com
Feedback

Related searches
KDD 2014
KDD 2016
WSDM 2015
PAKDD 2015
ICDM 2015
KDD Sydney
SIGIR 2015
KDD Cup 2015

Source: <http://exp-platform.com/2017abtestingtutorial/>

(2) Predicting the value of new features

10 search results

A

8 search results

B

- Overall Evaluation Criterion: clickthrough rate 1st SERP per query
- $A > B$, $A < B$, or $A = B$?

Source: <http://exp-platform.com/2017abtestingtutorial/>

(3) Predicting the value of new features

Esurance® Auto Insurance - You Could Save 28% with Esurance. Ads
www.esurance.com/California
Get Your Free Online Quote Today!

Esurance® Auto Insurance - You Could Save 28% with Esurance. Ads
www.esurance.com/California
Get Your Free Online Quote Today!
[Get a Quote](#) · [Find Discounts](#) · [An Allstate Company](#) · [Compare Rates](#)

A

B

- Overall Evaluation Criterion: revenue
- 4 A ads for every 3 B ads
- $A > B$, $A < B$, or $A = B$?

Source: <http://exp-platform.com/2017abtestingtutorial/>

Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

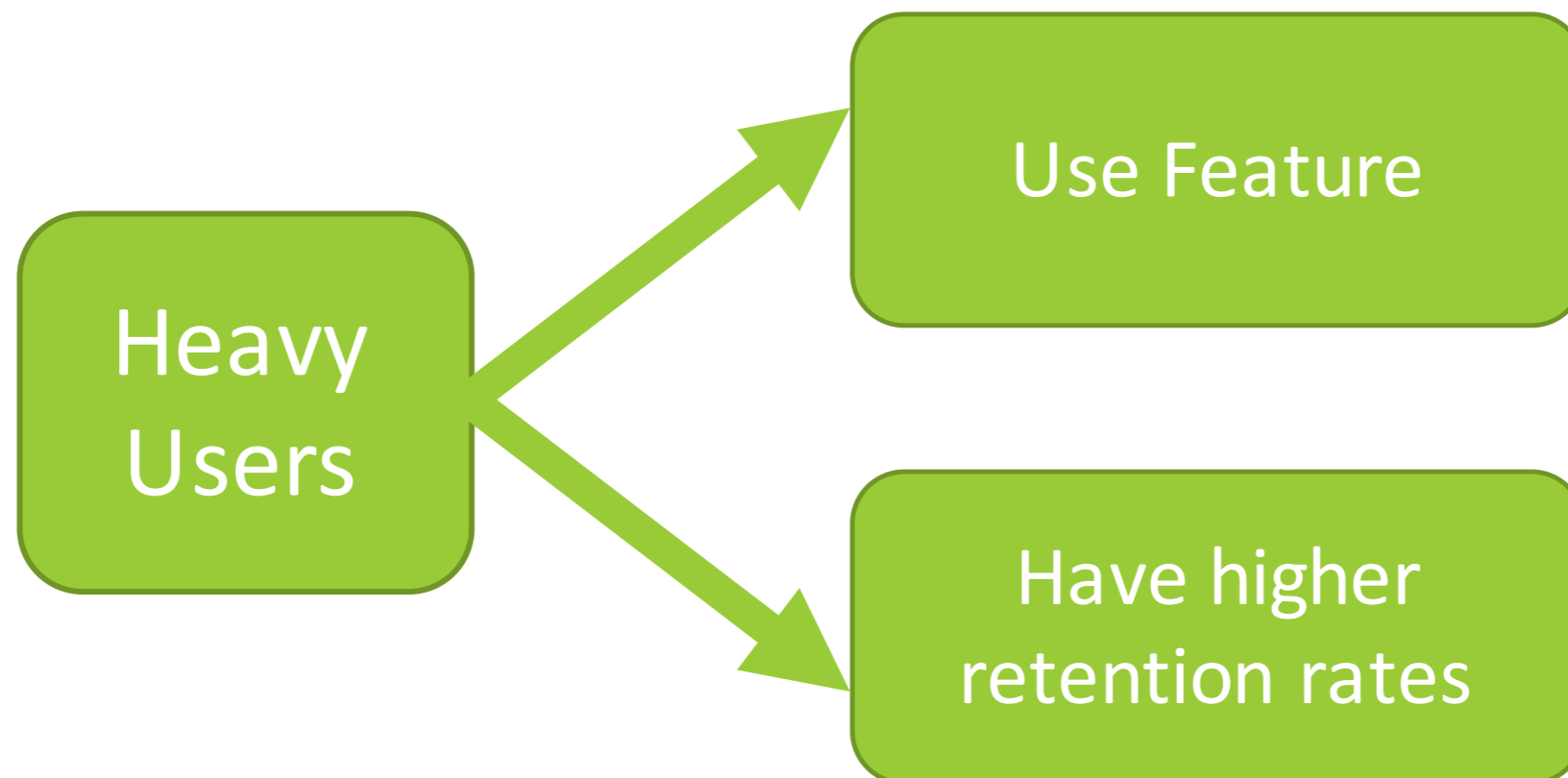
Correlation does not Imply Causation

- Umbrellas cause rain
- People with smaller hands live longer
- A new feature (e.g., a new advanced search tool) increases retention rate

Source: <http://exp-platform.com/2017abtestingtutorial/>

Correlation does not imply Causation

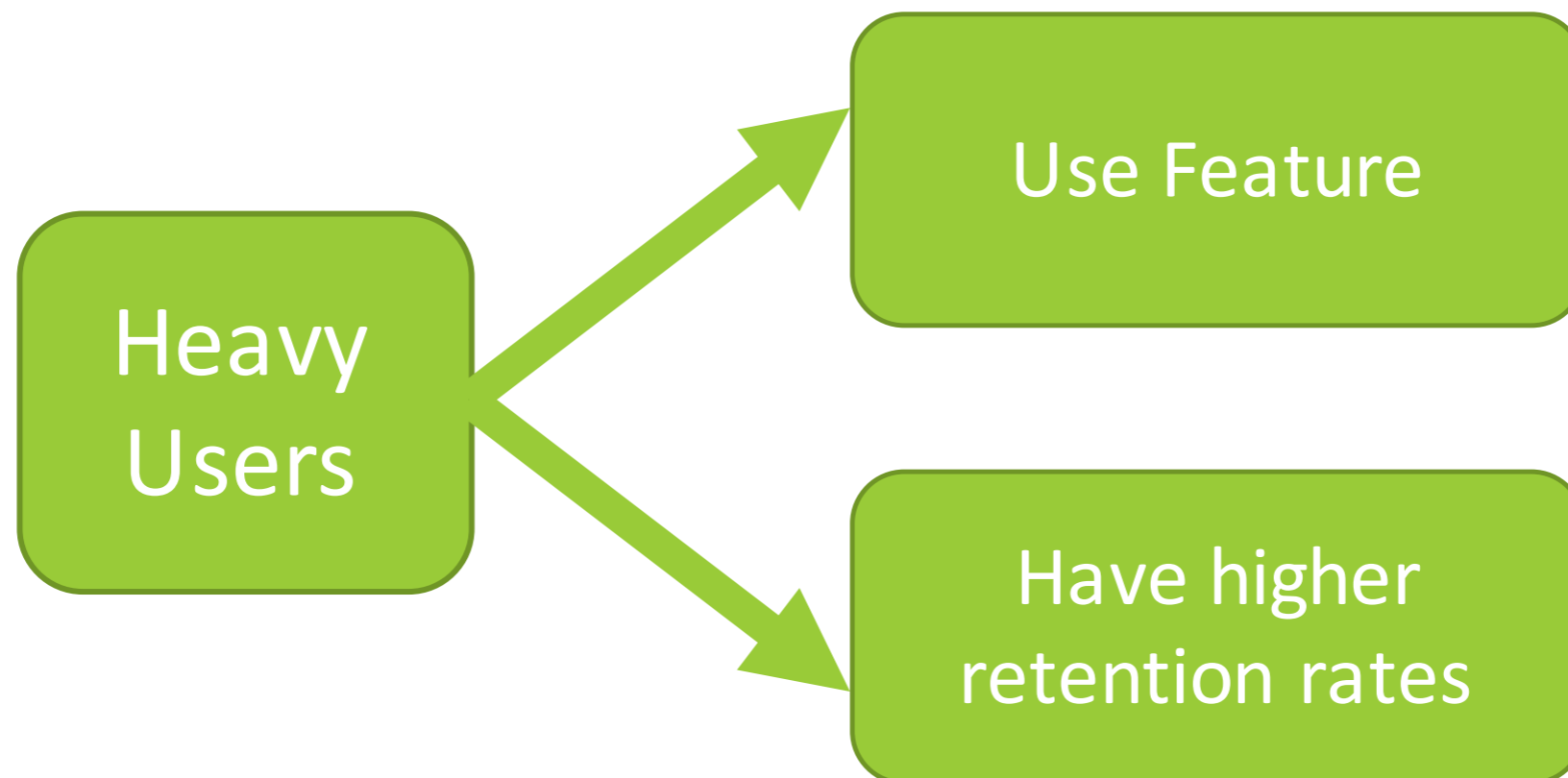
- Particularly important for understanding the impact of system features that are used more by certain types of users than others



Source: <http://exp-platform.com/2017abtestingtutorial/>

Correlation does not imply Causation

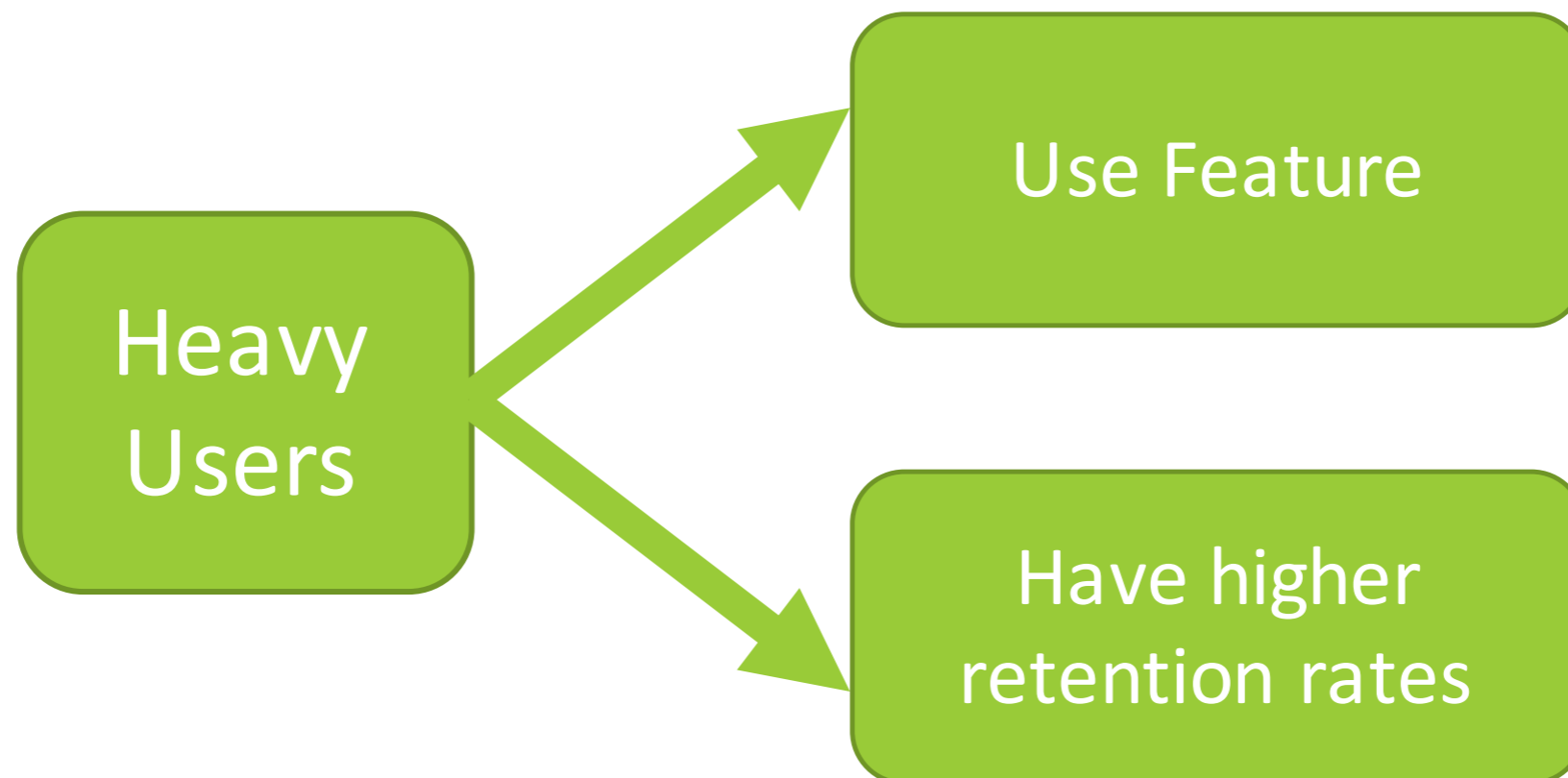
- What are features used more by heavy users?



Source: <http://exp-platform.com/2017abtestingtutorial/>

Correlation does not imply Causation

- What are features used more by new users?



Source: <http://exp-platform.com/2017abtestingtutorial/>

Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

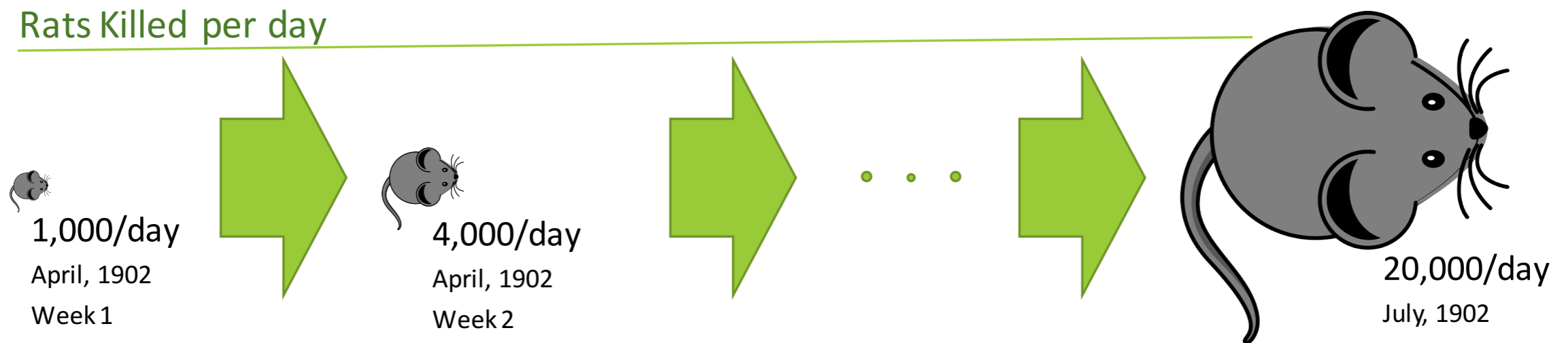
Short-term vs. Long-term Metrics

- An increase in ad clicks suggests an increase in revenue
- Showing lots of ads (often) hurts the user experience and decreases retention (i.e., long-term ad-click revenue)

Source: <http://exp-platform.com/2017abtestingtutorial/>

Using the wrong metric

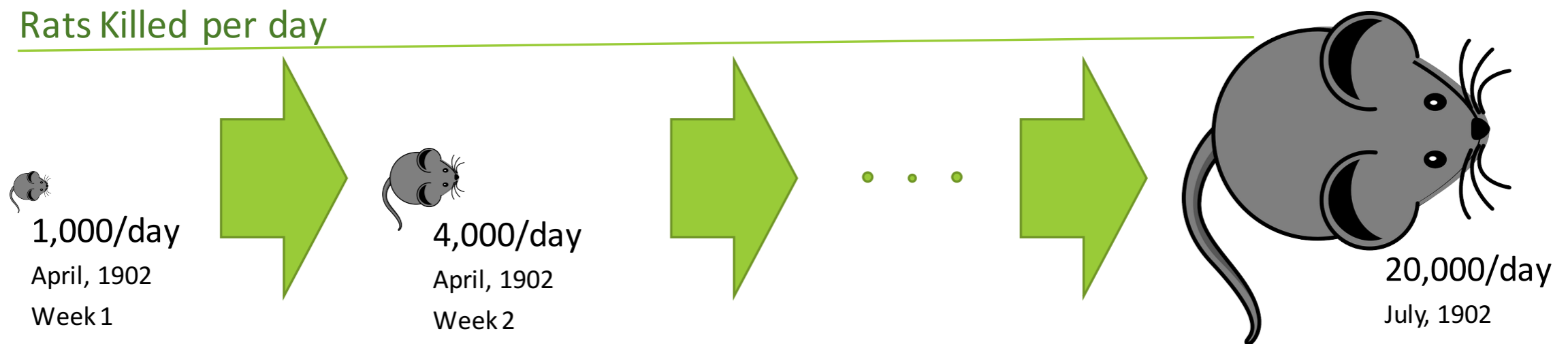
- Hanoi's French Quarter rat problem in 1902



Source: <http://exp-platform.com/2017abtestingtutorial/>

Using the wrong metric

- Hanoi's French Quarter rat problem in 1902



- What you do not measure, does not improve.
- **Goodhart's law:** "when a measure becomes a target, it ceases to be a good measure"

Source: <http://exp-platform.com/2017abtestingtutorial/>

Unexpected Effects on Important Metrics

- **Example:** a hyperlink on the SERP was changed to open on a new browser tab.
- It increased avg. SERP load time by 8.32%
- Why?

```
<a href="https://www.thesitewizard.com/" target="_blank">thesitewizard.com</a>
```

Source: <http://exp-platform.com/2017abtestingtutorial/>

Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Making Untested Claims

- **Question:** What is the effect of SERP load-time on ad-click revenue?
- Artificially increase SERP load-time and measure decrease in ad-click revenue
- Make the claim that decreasing the SERP load-time will have a comparable increase in ad-click revenue
- What's wrong with this?

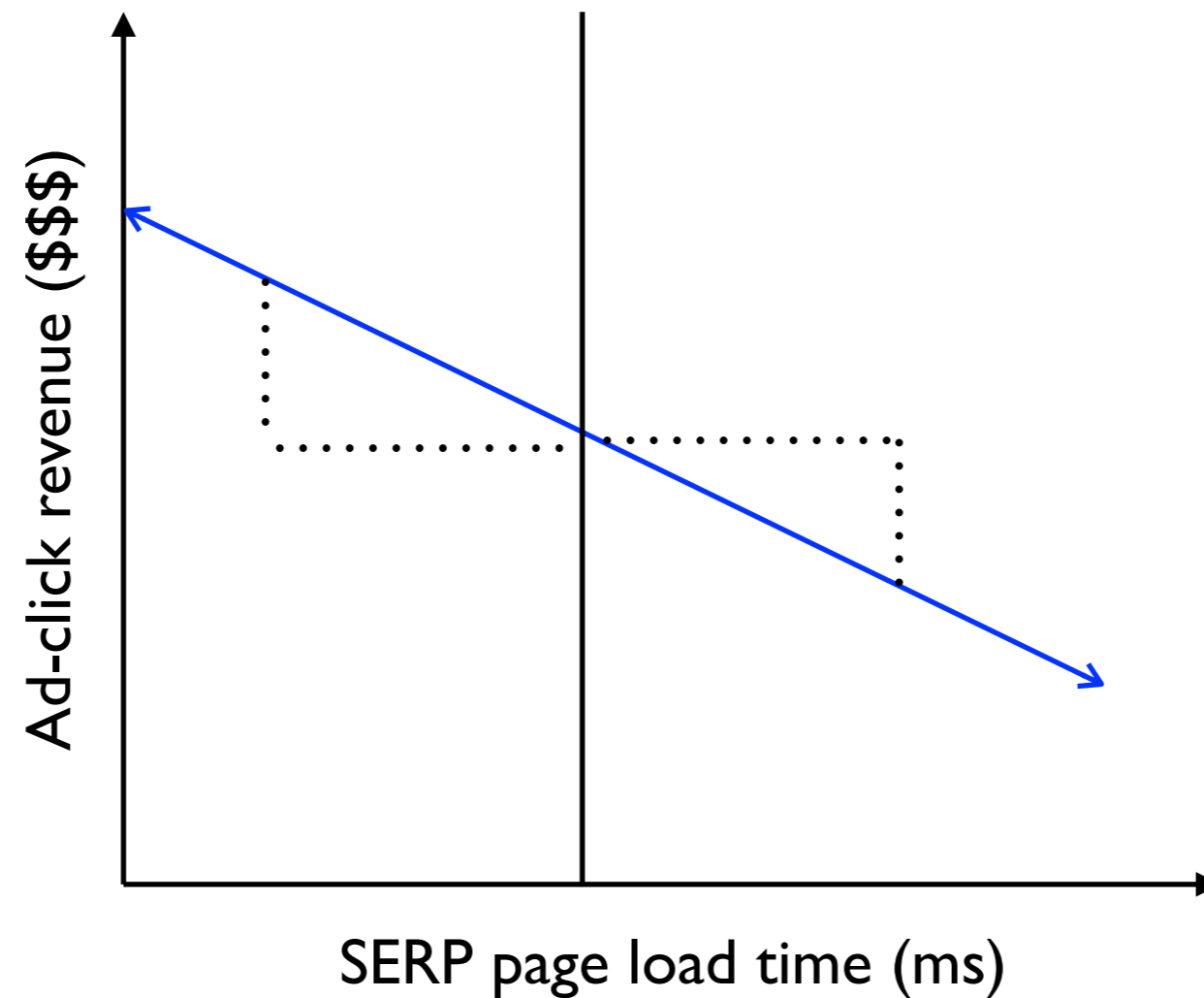
Source: <http://exp-platform.com/2017abtestingtutorial/>

Making Untested Claims

- **Question:** What is the effect of SERP load-time on ad-click revenue?
- Artificially increase SERP load-time and measure decrease in ad-click revenue
- Make the claim that decreasing the SERP load-time will have a comparable increase in ad-click revenue
- What's wrong with this?
- Assumes (bi-directional) linear relationship

Source: <http://exp-platform.com/2017abtestingtutorial/>

Making Untested Claims

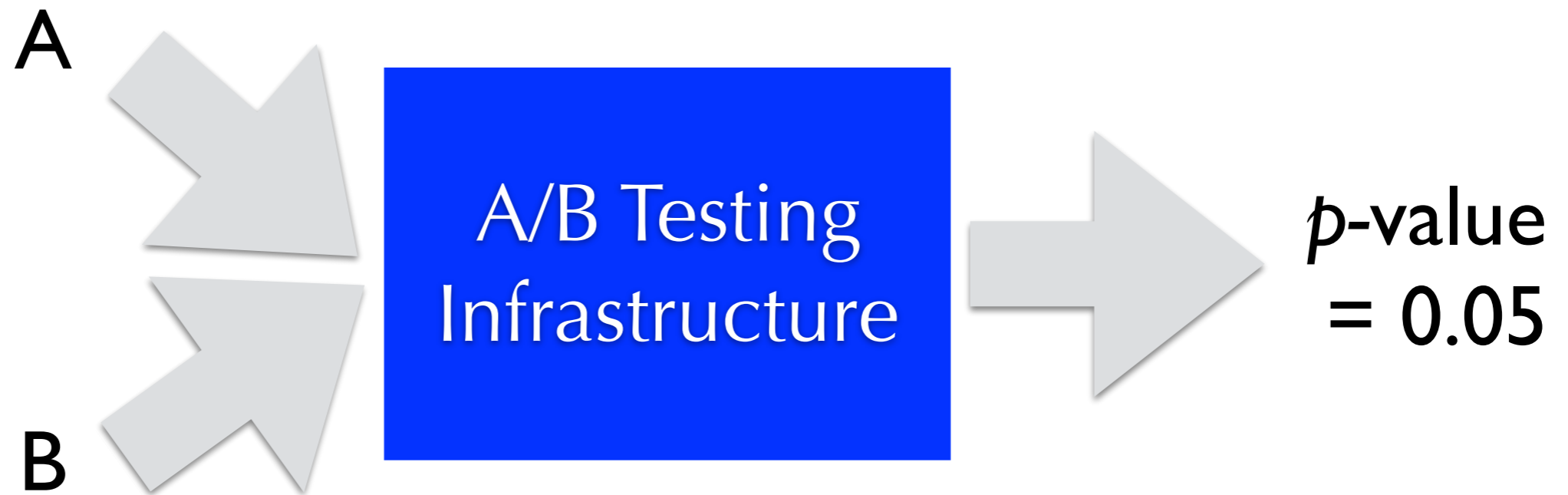


Source: <http://exp-platform.com/2017abtestingtutorial/>

Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Bugs in the Experimental Infrastructure



- User sampling + measurement + statistics
- How can we debug this infrastructure without opening the “black box”?

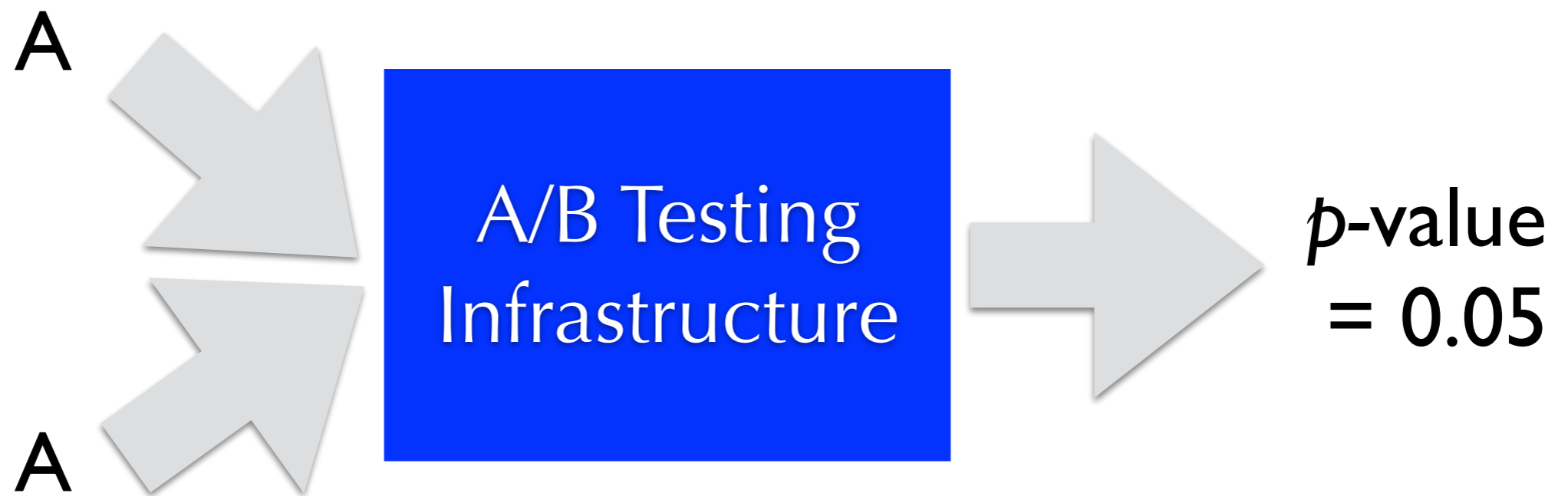
Source: <http://exp-platform.com/2017abtestingtutorial/>

Bugs in the Experimental Infrastructure



- Run lots of A/A tests (no differences between experimental and control conditions)
- How often should we observe a p -value of 0.05 or less?

Sound Statistical Methods



- Even when there is no difference between the two systems, it is still possible to observe a p -value of less than 0.05
- Why?

Sound Statistical Methods

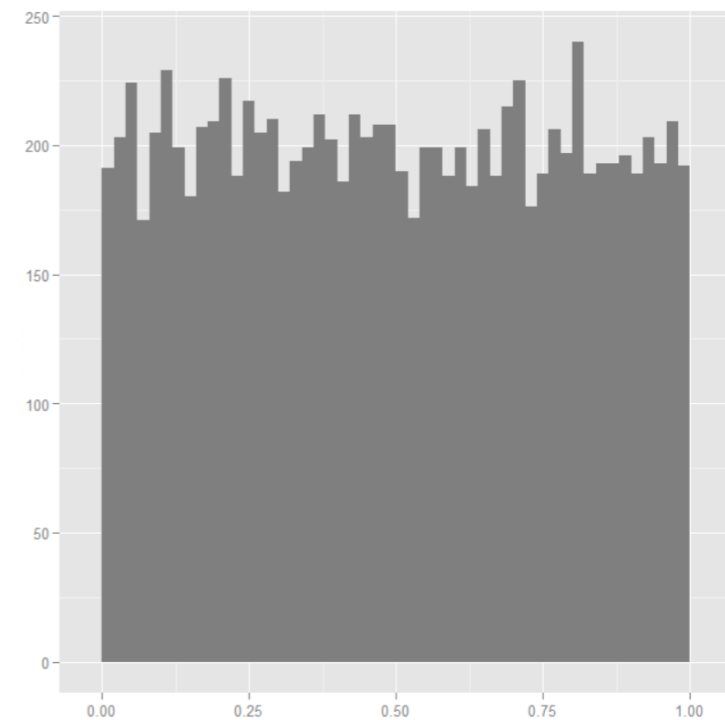
- By definition, the p -value is the probability of the observed difference in means (or a more extreme difference) under the null hypothesis!

Source: <http://exp-platform.com/2017abtestingtutorial/>

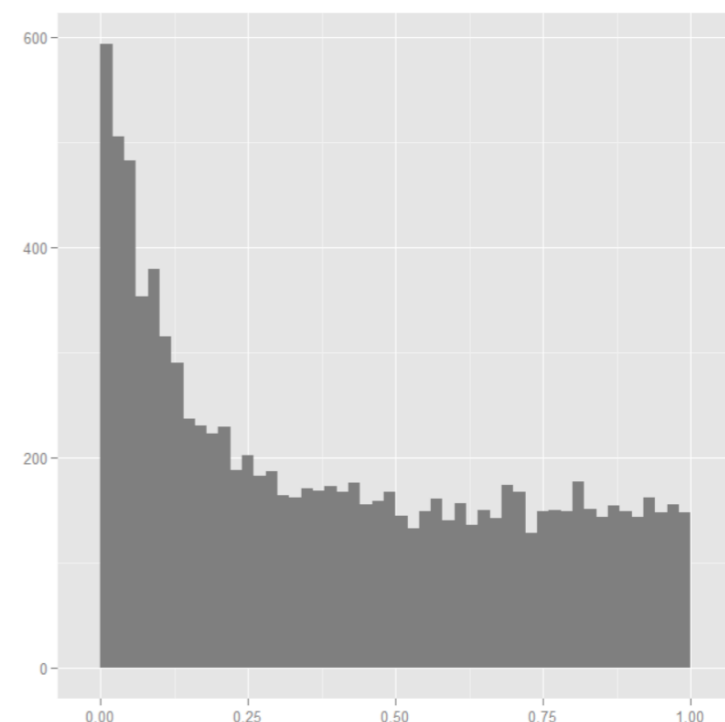
A/A Testing

- Run lots of A/A tests (no differences between experimental and control conditions)
- We should only observe p -values of 0.05 or less about 5% of the time
- The p -value distribution should be uniform rather than skewed to low or high values

Good p -values



Bad p -values



Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Causes of Type I Errors (False Positives)

- Running the same A/B test many times until we observe a significant difference
- Using 100+ metrics and focusing on the ones that are significant
- Running an experiment for as long as it takes to reach significance
- Running an experiment and stopping early because we reached significance

Source: <http://exp-platform.com/2017abtestingtutorial/>

Causes of Type I Errors (False Positives)

- **Bonferroni correction:** multiplying the p -value by the number of comparisons

Source: <http://exp-platform.com/2017abtestingtutorial/>

Hurting the User Experience

- Less manual monitoring of experiments
- Buggy features or bad ideas
- Interactions between concurrent experiments: the whole is less than the sum of its parts

Source: <http://exp-platform.com/2017abtestingtutorial/>

Cautionary Steps: Starting Small

- Starting internally (within the company)
- Starting with only a few users
- Starting with only partial exposure (1/10 queries)

Source: <http://exp-platform.com/2017abtestingtutorial/>

Cautionary Steps: Different types of Metrics

- **Data quality metrics:** ensure that the feature was implemented correctly
- **Overall evaluation criteria:** single metric that measures improvement in user experience (e.g., number of satisfied clicks)
- **Guardrail metrics:** metrics used to shutdown an experiment (e.g., queries with no clicks)
- **Local metrics:** metrics that measure what the user is doing less of (because of the new feature)

Source: <http://exp-platform.com/2017abtestingtutorial/>

Cautionary Steps: Measuring interactions

Exp. 2

Exp. 1

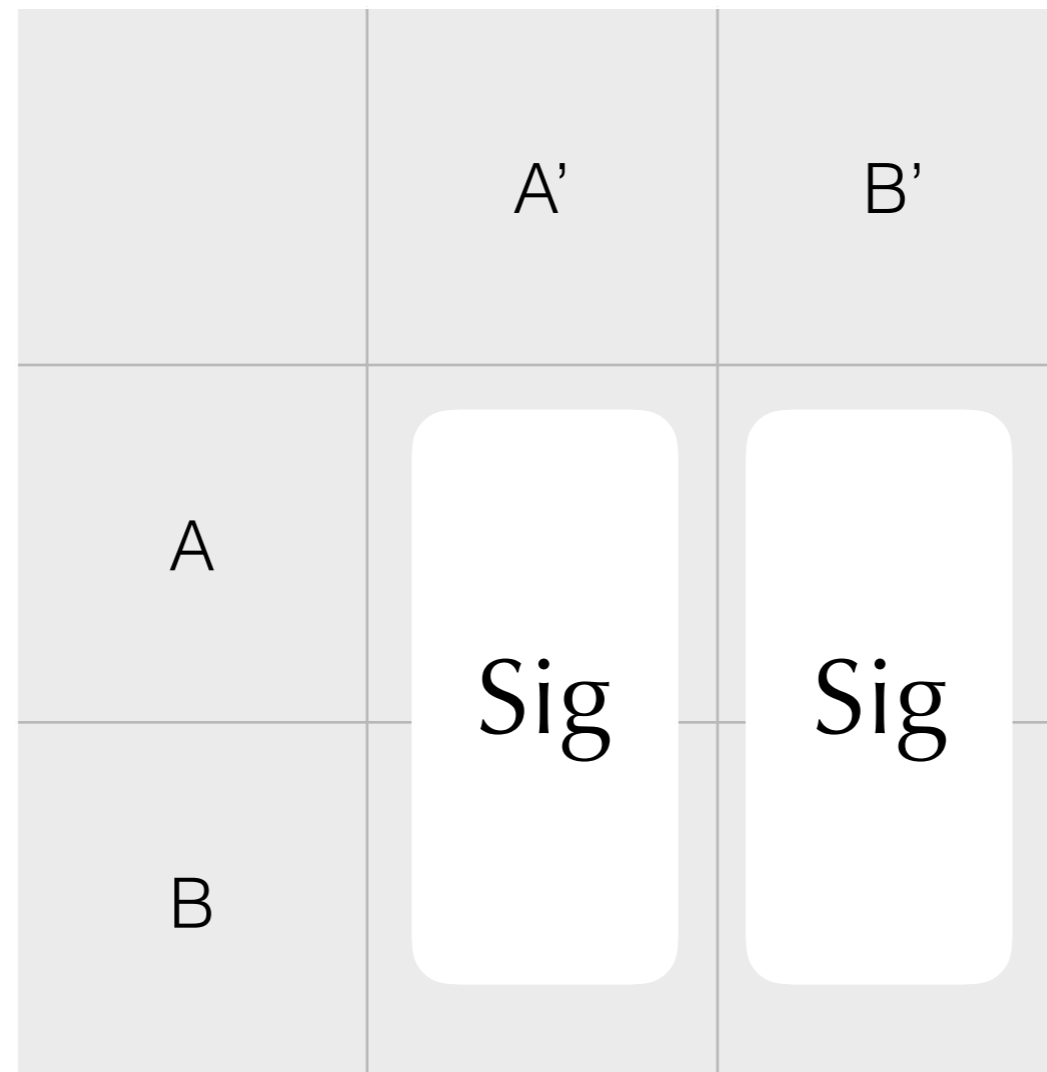
	A'	B'
A	Sig	No Sig
B	Sig	No Sig

Source: <http://exp-platform.com/2017abtestingtutorial/>

Cautionary Steps: Measuring interactions

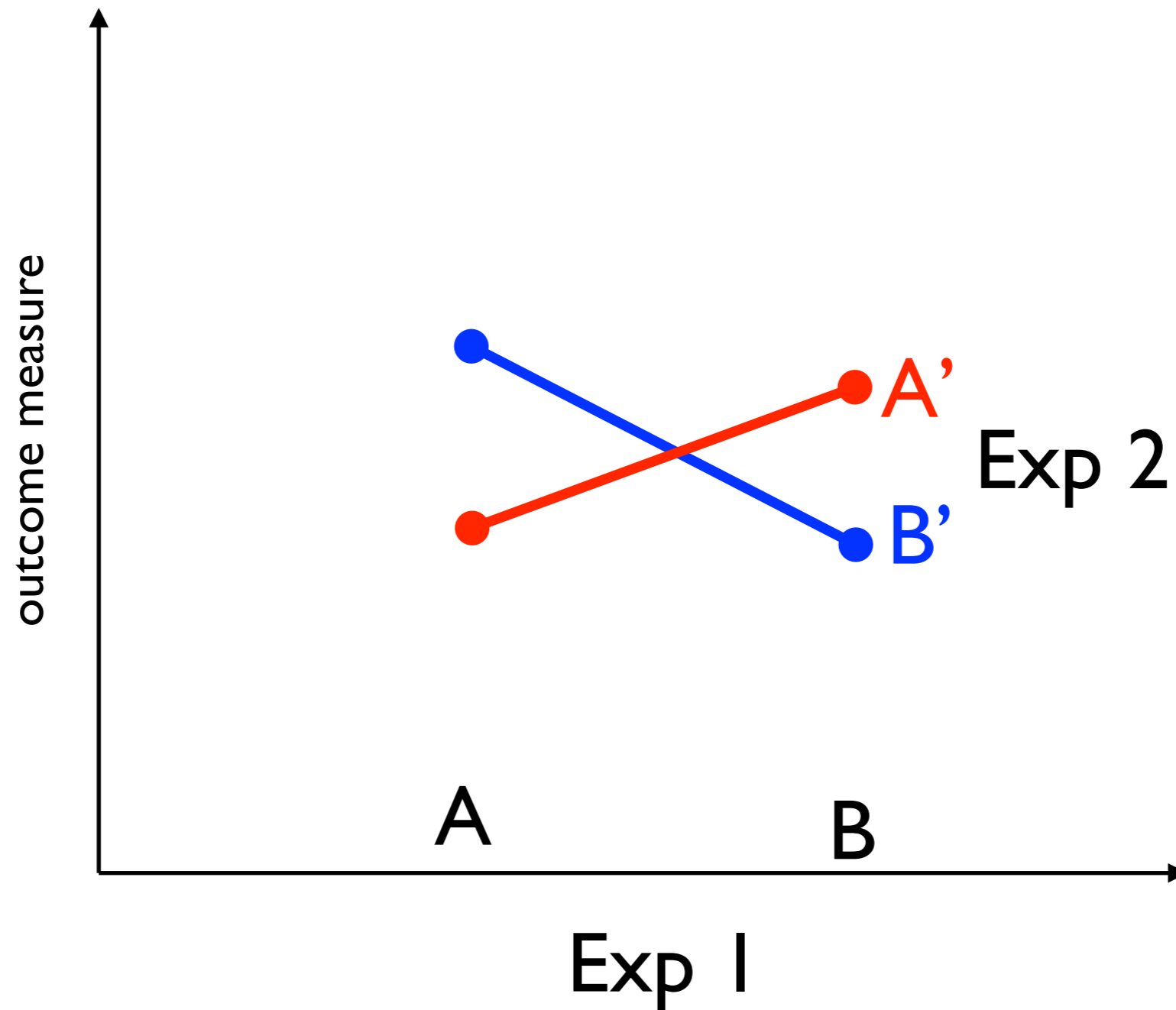
Exp. 2

Exp. 1



Source: <http://exp-platform.com/2017abtestingtutorial/>

Cautionary Steps: Measuring interactions



Source: <http://exp-platform.com/2017abtestingtutorial/>

Ethical Considerations

- System development is influenced by the majority
- Certain communities may be under-represented in the data
- While there is an “average user”, there is also high variance (nobody is close to the average)
- Metrics used in A/B tests are crude measures of “user experience”
- Users may need to experience extreme differences to show (positive or negative) changes in behavior
- A/B tests are done without considering whether the user is in a vulnerable state

Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience