



# Online Controlled Experimentation at Scale

May 14, 2019

Kaska Adotey, Ph.D.

Data Scientist

Analysis & Experimentation Team ([exp-platform.com](http://exp-platform.com))

Questions: [kaadotey@microsoft.com](mailto:kaadotey@microsoft.com)



Joint work with many members of the A&E/ExP platform team

# The Life of an Idea (2012)

On Bing.com, move ad text to the title line to make it longer

## Control – Existing UI

bing MS Beta

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE

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 Ad Titles

bing MS Beta

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE

flowers

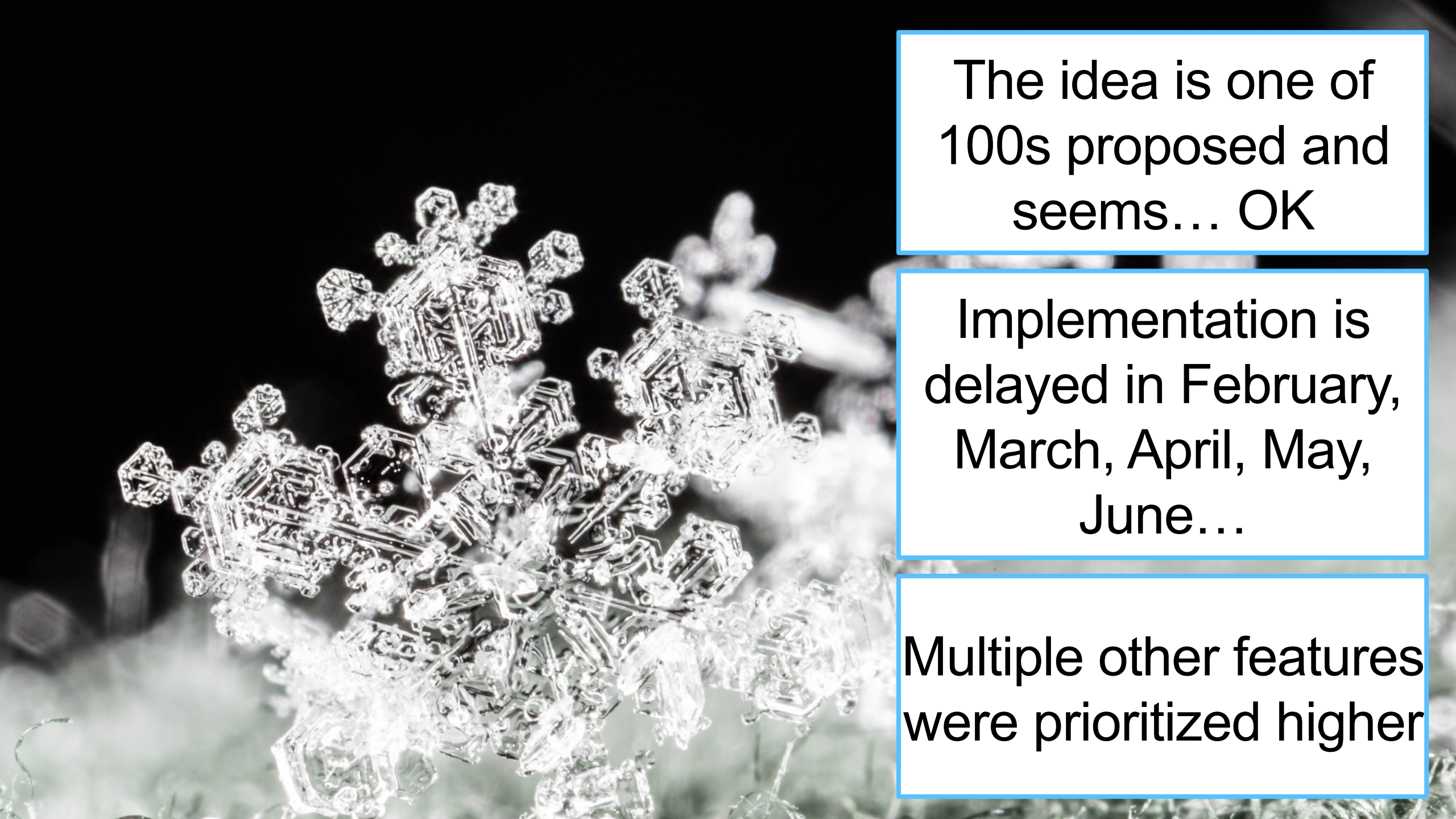
358,000,000 RESULTS

**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.



The idea is one of  
100s proposed and  
seems... OK

Implementation is  
delayed in February,  
March, April, May,  
June...

Multiple other features  
were prioritized higher

An engineer thought  
“this is trivial to  
implement.”

They did so in a few  
days and then started  
a controlled  
experiment (A/B test).



# What happened?



- An alert fired to say something was wrong with revenue (it was too high).
- This is useful in case you do something like log revenue twice.



- But in this case, it was not a logging bug.



- The change increased Bing's revenue by 12% (over \$120M at the time) without hurting any guardrail metrics.



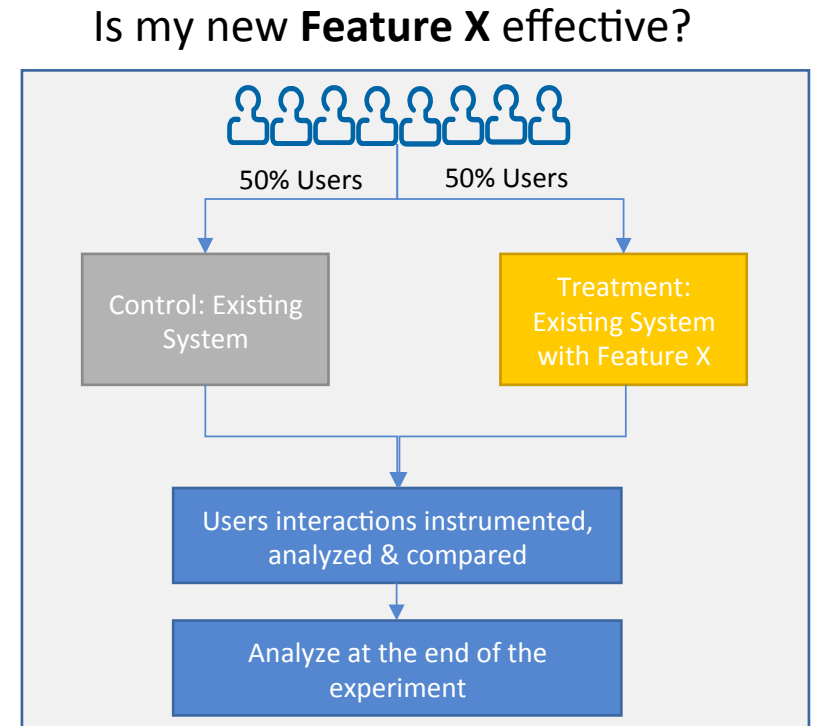
- **We are terrible at assessing the value of ideas.**
- Few ideas generate over \$100M in incremental revenue (as this idea), but the **best revenue-generating** idea in Bing's history to that point was badly rated and delayed for months!

# Agenda

- **Experimentation at scale:** How to manage an experimentation lifecycle for over 24,000 experiments/year
- **Three examples of real A/B tests:** You are the decision maker
- **Four important lessons:** Take home wisdom for online experiments at scale.

# Online Controlled Experiments (A/B/n Tests)

- Simple concept:
  1. Randomly split traffic between two (or more) versions
    - Control: Existing System
    - Treatment(s): Feature(s) being tested
  2. Collect metrics of interest
  3. Analyze
- Must run statistical tests to confirm differences are not due to chance
- Sample of real users
  - Not **WEIRD** (Western, Educated, Industrialized, Rich, and Democratic) like many academic research samples
- Scientific way to prove **causality**, i.e., the changes in metrics are caused by changes introduced in the treatment(s)
- Used by Microsoft, Google, Facebook, Netflix, LinkedIn, and many others



Our Mission:















# Accelerate innovation through trustworthy analysis and experimentation

- Currently serving multiple Microsoft organizations

## TRUSTWORTHINESS

- Empower the HiPPO (Highest Paid Person’s Opinion) with data



 Bing	 Exchange	 msn
 Cortana	 OneNote	 Store
 Windows	 Skype	 Photos
 Office	 Teams	 Visual Studio
	 Edg e	 XBOX

## Team of 110+ people

- ~60 developers
- ~40 data scientists
- ~10 program managers



# Three Real Examples

1. Bing Search Engine Results Page Truncation
2. Windows Search Box
3. Killer Instinct Initial Character

# About the Real Examples

Three real **experiments that ran at Microsoft**

All had **enough users for statistical validity**

For each, I provide the **Overall Evaluation Criterion (OEC)**

**The challenge:** You predict which variant will do best (by the OEC).

- Everyone please stand up.
- You will be given three options and answer by raising **left hand**, **right hand**, or **neither hand**.
- If you're wrong, please sit down.
- Random guessing implies  $100\%/(3^3) = \sim 4\%$  will get all three questions right.
- **Let's see if the room can beat random guessing!**

# Example 1: Bing SERP Truncation

SERP = Search Engine Result Page

**Version A:** show 10 algorithmic search results

**Version B:** show 8 algorithmic search results by removing the last 2 results

**OEC:** Clickthrough Rate on 1<sup>st</sup> SERP per query

**The challenge:**

**Raise your left hand** if you think **A Wins**

**Raise your right hand** if you think **B Wins**

**Don't raise your hand** if you think they are **about the same**

The screenshot 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 search results are displayed in a list format, with the first result highlighted in blue. The results are as follows:

- 1** **KDD 2015, 10-13 August 2015, Sydney**  
[www.kdd.org/kdd2015](http://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](#)  
**Research Track**  
ACM SIGKDD Invitation to Participate - 2015 KDD Conference August ...  
**Sponsorship**  
KDD 2015 will be held between 10-13 August 2015 in Sydney. ...  
**Tutorials**  
KDD 2015 Call for Papers, Workshops, Tutorials and ...  
**Kdd-2014**  
KDD 2014, a premier interdisciplinary conference, brings together ...  
**Attending**  
Attending KDD 2015 Visa Information; Registration. ...  
**Organisers**  
Organisers and program committee members for KDD 2015  
[See results only from kdd.org](#)
- 2** **KDD 2015 - The 21th ACM SIGKDD International Conference ...**  
[conference.researchbib.com/view/event/33616](http://conference.researchbib.com/view/event/33616)  
KDD 2015 - The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining  
**Related searches for kdd 2015**  
[KDD 2014](#) [KDD Sydney](#)  
[KDD Cup 2015](#) [PAKDD 2015](#)  
[WSDM 2015](#) [KDD 2016](#)
- 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 ...**  
[myhuban.com/conference/136](http://myhuban.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](http://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](http://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 ...**  
[eventlegg.com/kdd-2015](http://eventlegg.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](http://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 showing "1 2 3 4 5" with a right arrow icon.

On the right side of the page, there is a sidebar for the first result, "KDD 2015". It contains the following information:

- 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](#)

# Bing SERP Truncation Results

If you raised any hand, please sit down

While there are obviously exceptions, most of the time users click at the same rate.

*In this case, with over 3M users in each variant, we could not detect a stat-sig delta.*

*Users simply shifted the clicks from the last two algorithmic results to other elements of the page.*

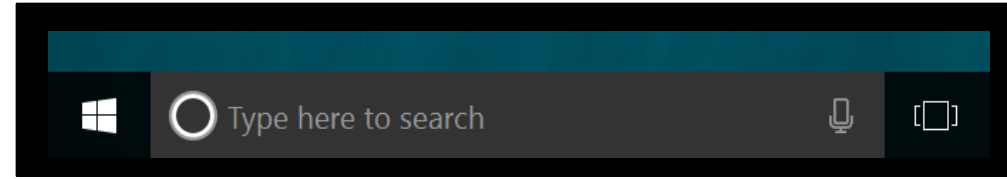
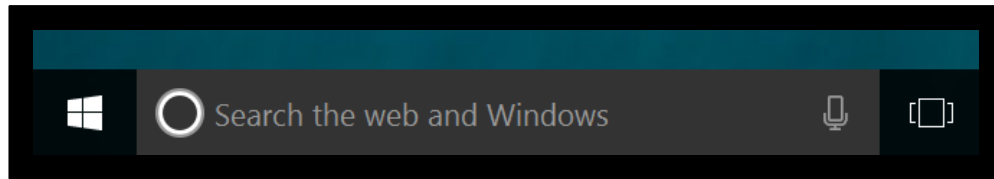
A&E wrote a paper with several rules of thumb (<http://bit.ly/expRulesOfThumb>)

Rule of Thumb: Reducing abandonment (1-clickthrough-rate) is hard.  
Shifting clicks is easy

# Example 2: Windows Search Box

The Search box is the lower left part of the taskbar for most of the ~500M machines running Windows 10

Here are the two variants:



**Version A:** search box says “Search the web and Windows”

**Version B:** search box says “Type here to search”

**OEC:** User engagement, i.e. more searches (and thus more Bing revenue)

**The challenge:**

**Raise your left hand** if you think **A Wins**

**Raise your right hand** if you think **B Wins**

**Don't raise your hand** if you think they are **about the same**

# Windows Search Box Results

If you didn't raise a hand, please sit down

If you raised your left hand, please sit down

If you raised your right hand, you are good! (or you've looked closely at the lower left of your Windows 10 UI sometime since 2017)

We actually tested 4 variants here (listed in order of performance):

1. "Type here to search" (**WINNER**)
2. "What can I help you find?"
3. "Ask me anything" (Control – the design that shipped with Windows 10)
4. "Search the web and Windows" (the one shown in prior slide)

*This change was worth several million dollars/year*

## Example 3: Xbox – Killer Instinct

Fighting video game that runs on the Xbox One console

**Freemium Model:** can play one character for free (Jago) but must pay to play others

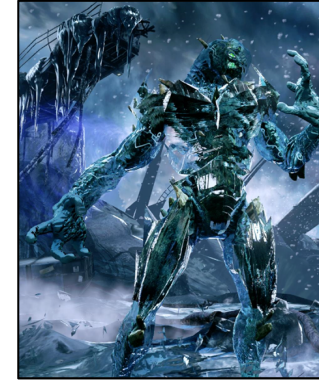
Team hopes to **increase revenue** by getting players to purchase additional characters



# Example 3: Xbox – Killer Instinct

Should Killer Instinct use Jago or Glacius as initial free character?

Here are the two variants:



**Version A:** Jago as initial free character

**Version B:** Glacius as initial free character

**OEC:** Revenue

**The challenge:**

**Raise your left hand** if you think **A Wins**

**Raise your right hand** if you think **B Wins**

**Don't raise your hand** if you think they are **about the same**



# Killer Instinct

If you didn't raise a hand, please sit down

If you raised your left hand, please sit down

If you raised your right hand, you are right!

Revenue increased, but here engagement (time in game) decreased

*Takeaway: Stop debating or voting on what to ship → **get the data.***

# Four Important Lessons

1. Agree on a Good Overall Evaluation Criteria (OEC)
2. Most Ideas Fail
3. Small Changes can have Big Impacts
4. Validate the Experimentation System

# Lesson #1: Agree on a good

## OEC

- Overall Evaluation Criteria → OEC
- Getting agreement on the OEC in an org is a huge step forward
- OEC should:
  - Predict long-term value
  - Be hard to game
- Criterion could be function of various factors, such as:

*Conversion/action, time to action, visit frequency*

See also:

<http://exp-platform.com/advanced-topics-in-online-experiments/>



# Example of a Bad OEC

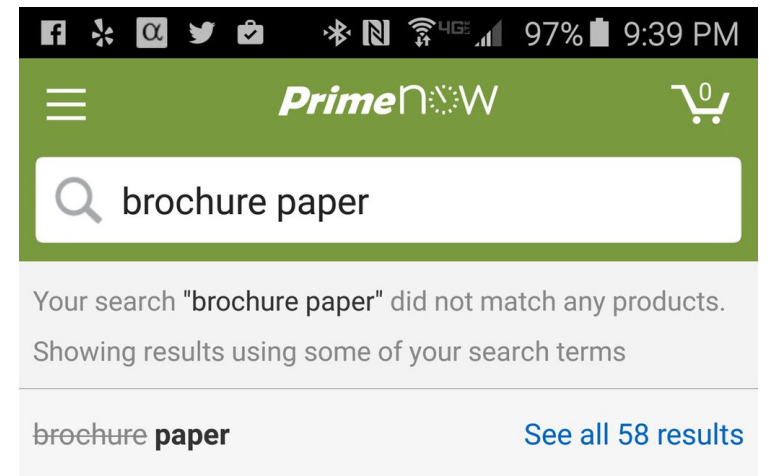
Your data science team makes an observation:

*2% of queries end up with “No results.”*

Team decides to minimize the “count of no results” metric

*... But this optimization can have unintended consequences*

This is a real example from Amazon Prime Now search (via [@RonnyK](#)).



Seventh Generation Bathroom Tissue, 2 Ply, 300 Sheets, 4 Rolls

by Seventh Generation



\$4.39

Add

# Lesson #2: Most Ideas Fail

Features are built because teams believe they are useful.

*But most experiments show that features fail to move the metrics they were designed to improve*

Based on experiments at Microsoft ([paper](#))

- 1/3 of ideas were positive ideas and statistically significant
- 1/3 of ideas were flat: no statistically significant difference
- 1/3 of ideas were negative and statistically significant

Experiment often

*“If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster“*

-- Mike Moran, Do it Wrong Quickly

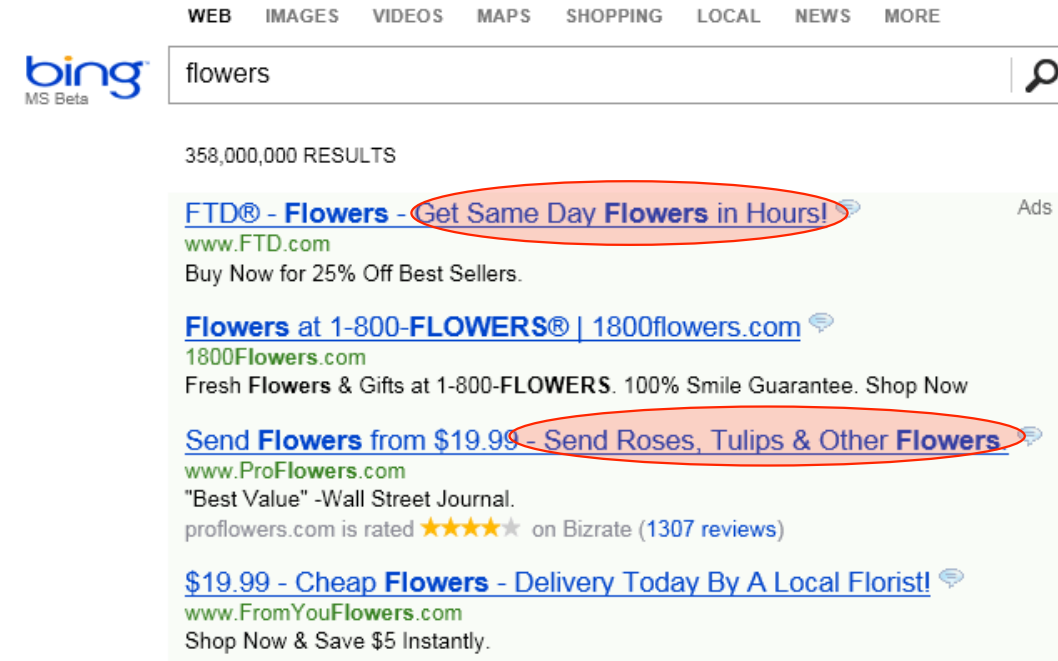
Try radical ideas. You may be surprised

- Doubly true if it's cheap to implement



# Lesson #3: Small Changes can have a Big Impact to Key Metrics

Opening example (Bing Ads) worth over \$120M annually



WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE

bing MS Beta

flowers

358,000,000 RESULTS

**FTD® - Flowers - Get Same Day Flowers in Hours!** Ads  
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

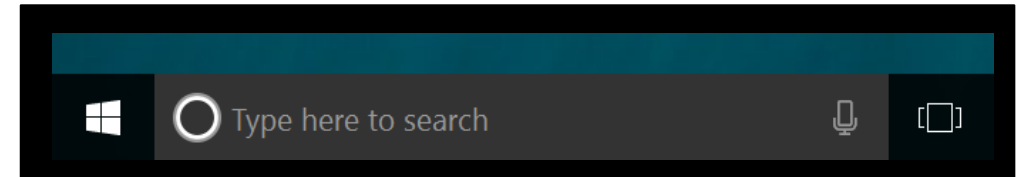
**Send Flowers from \$19.99 - Send Roses, Tulips & Other Flowers.**  
www.ProFlowers.com  
"Best Value" -Wall Street Journal.  
proflowers.com is rated ★★★★★ on Bizrate (1307 reviews)

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

# Lesson #3: Small Changes can have a Big Impact to Key Metrics

Opening example (Bing Ads) worth over \$120M annually

Windows Search box example: \$5M+



# Lesson #3: Small Changes can have a Big Impact to Key Metrics

Opening example (Bing Ads) worth over \$120M annually

Windows Search box example: \$5M+

Site Links in Ads: \$50M annually

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



# Lesson #3: Small Changes can have a Big Impact to Key Metrics

Opening example (Bing Ads) worth over \$120M annually

Windows Search box example: \$5M+

Site Links in Ads: \$50M annually

Changed text color for fonts in Bing:  
\$10M annually



# Lesson #3: Small Changes can have a Big Impact to Key Metrics

Opening example (Bing Ads) worth over \$120M annually

Windows Search box example: \$5M+

Site Links in Ads: \$50M annually

Changed text color for fonts in Bing: \$10M annually

100msec improvement in Bing server perf: \$18M annually



# Lesson #3: Small Changes can have a Big Impact to Key Metrics

But these are the rare gems amongst tens of thousands of experiments

4,000+

Users/month

2,000+

Experiments/month

250,000+

Scorecards/year



# Lesson #4: Validate the Experimentation System

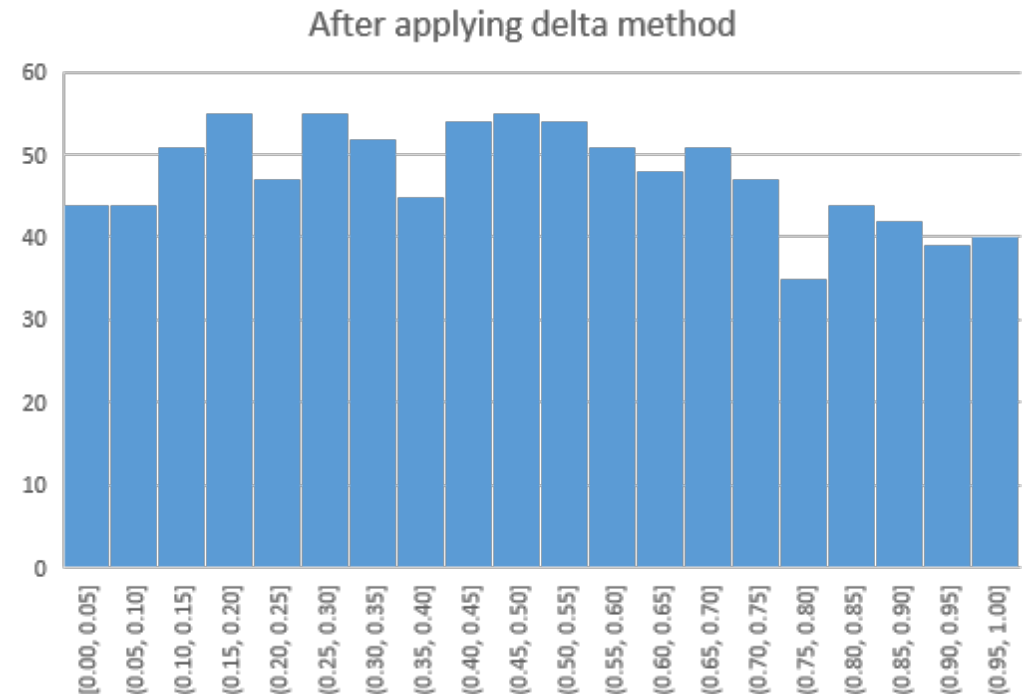
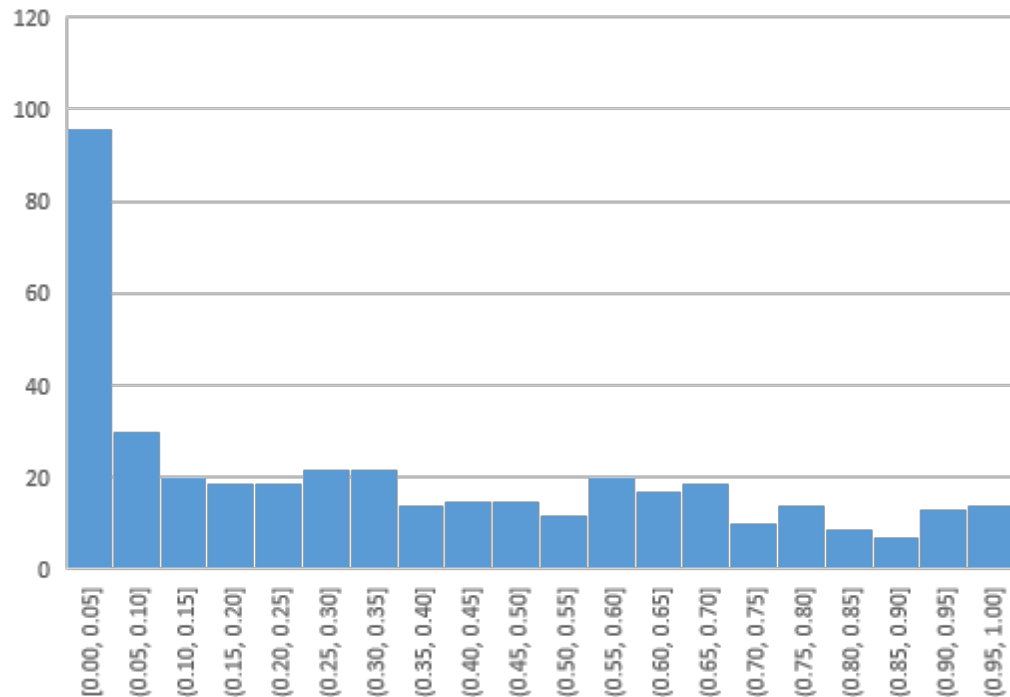
- Software that shows p-values with many digits of precision leads users to trust it, but the statistics or implementation behind it could be buggy
- **Getting numbers is easy; getting numbers you can trust is hard**
- Let's cover 3 validation recommendations
  - Bot Detection
  - A/A Tests
  - Sample Ratio Mismatch (SRM) Checks

Check for Bots, which can cause significant skews



# Run A/A Tests

- If the system is operating correctly, the system should find a stat-sig difference only about 5% of the time.
- P-value distribution for metrics in A/A tests should be uniform
- Do 1,000 A/A tests, and check if the distribution is uniform for each metric
- We tried this for some Skype metrics, and we had to correct things (delta method)



# Run Sample Ratio Mismatch (SRM) Checks

## Real example

- Control: 821,588 users
- Treatment: 815,482 users
- Ratio: 50.2% (should have been 50%)

Should I be worried?



## Absolutely!

The p-value is  $1.8e-6$ , so the probability of this split (or more extreme) happening by chance is less than 1 in 500,000 (the Null hypothesis is true by design)

# Summary

- Think about the OEC. Agree on what to optimize and measure it.
- Compute the statistics carefully
  - Getting numbers is easy. Getting a number you can **trust** is harder
- Be skeptical and triple check things before you celebrate. See <http://bit.ly/twymanLaw>

*Any figure that looks interesting or different is usually wrong*

- Experiment often:
  - Triple your experiment rate and you triple your success (and failure) rates. Fail fast & often!
  - Accelerate innovation by lowering the cost of experimenting.
- See <http://exp-platform.com> for papers

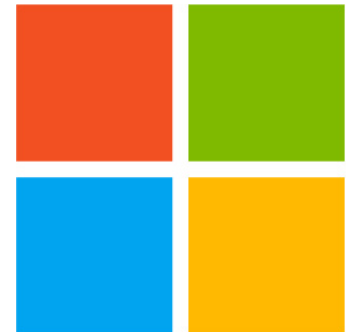


Are we hiring?

Yes!

Email me for details:

[kaadotey@microsoft.com](mailto:kaadotey@microsoft.com)



Microsoft