

Dec 6, 2016

# Online Controlled Experiments: Introduction, Pitfalls, and Scaling

Slides at <http://bit.ly/TPKohavi2016>, @RonnyK

Ronny Kohavi, Distinguished Engineer, General Manager,  
Analysis and Experimentation, Microsoft

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



# Agenda

- Introduction and motivation
- Four real examples: you're the decision maker  
Examples chosen to share lessons
- Pitfalls
- Scaling

# A/B Tests in One Slide

## ➤ Concept is trivial

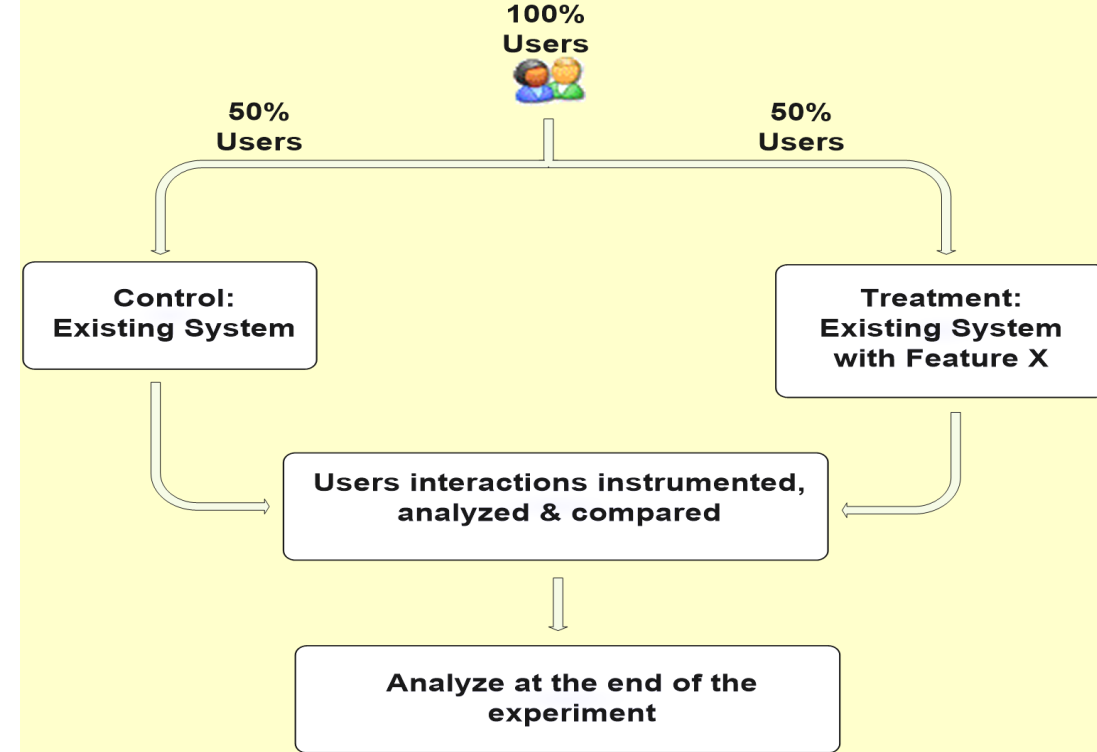
- Randomly split traffic between two (or more) versions
  - A (Control, typically existing system)
  - B (Treatment)
- Collect metrics of interest
- Analyze

## ➤ A/B test is the simplest controlled experiment

- A/B/n refers to multiple treatments (often used and encouraged: try control + two or three treatments)
- MVT refers to multivariable designs (rarely used by our teams)
- Equivalent names: [Bucket tests \(Yahoo!\)](#), Flights (Microsoft), 1% Tests (Google), Field experiments (medicine, Facebook), randomized clinical trials (RCTs, medicine)

## ➤ Must run statistical tests to confirm differences are not due to chance

## ➤ Best scientific way to prove **causality**, i.e., the changes in metrics are caused by changes introduced in the treatment(s)



# Advantage of Controlled Experiments

- Controlled experiments test for **causal** relationships, not simply correlations
- When the variants run concurrently, only two things could explain a change in metrics:
  1. The “feature(s)” (A vs. B)
  2. Random chance

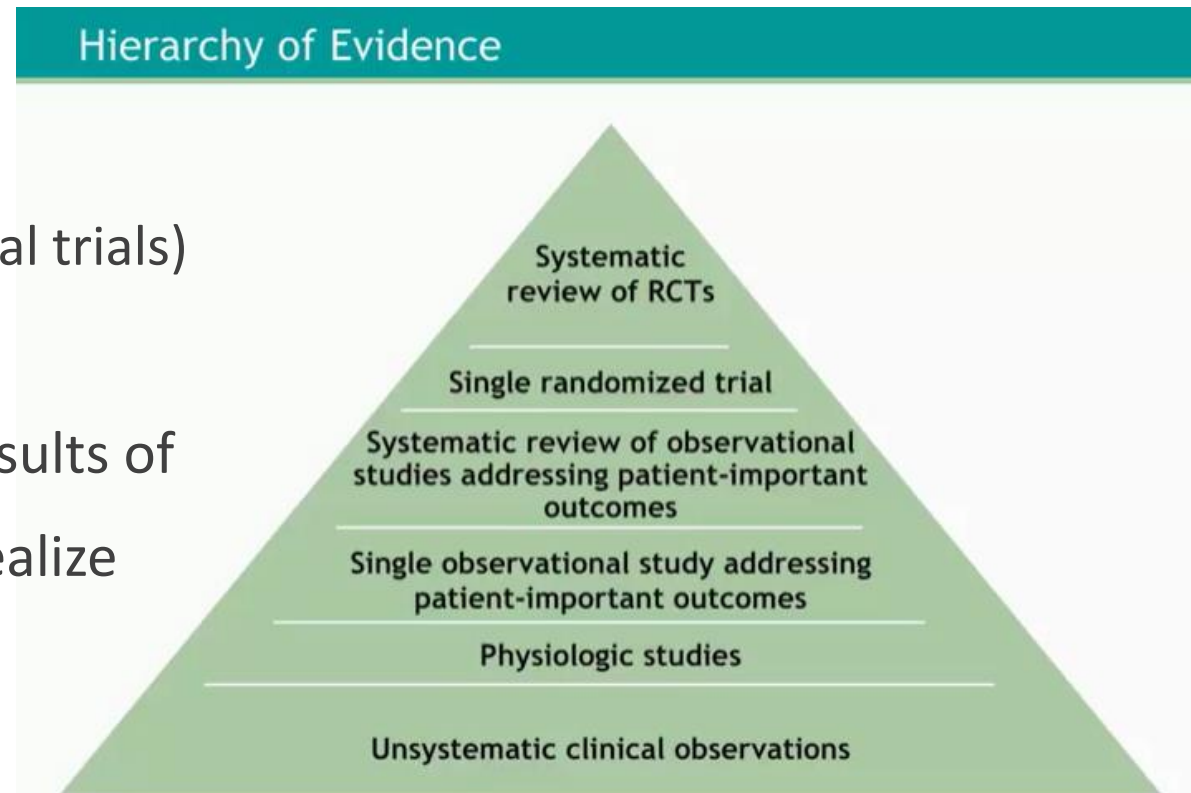
Everything else happening affects both the variants

For #2, we conduct statistical tests for significance
- The gold standard in science and the only way to prove efficacy of drugs in FDA drug tests
- Controlled experiments are not the panacea for everything.

Issues discussed in the journal survey paper (<http://bit.ly/expSurvey>)

# Hierarchy of Evidence and Observational Studies

- All claims are not created equally
  - Observational studies are UNcontrolled studies
  - Be very skeptical about unsystematic studies or single observational studies
  - The table on the right is from a Coursera lecture: [Are Randomized Clinical Trials Still the Gold Standard?](#)
  - At the top are the most trustworthy
    - Controlled Experiments (e.g., RCT randomized clinical trials)
    - Even higher: multiple RCTs—replicated results
  - Why show this?  
Our users the experimentation platform trusted results of controlled experiments.  
When seeing observational studies, they did not realize the trustworthiness is MUCH lower
- See <http://bit.ly/refutedCausalClaims>



# Simple Example: Common Cause

- Example Observation (highly stat-sig)

## **Palm size correlates with your life expectancy**

The larger your palm, the less you will live, on average

- Try it out - look at your neighbors and you'll see who is expected to live longer
- But...don't try to bandage your hands, as there is a **common cause**
- Women have smaller palms and live 6 years longer on average
- Obviously you don't believe that palm size is causal, but you may hear this:
  - Users of my feature X churn at a lower rate.  
Unlikely causal. More likely: heavy users use more features and churn less
  - True fact: users that see more errors churn less. Same common cause: heavy users.  
But intentionally showing more errors won't reduce churn

# Hormone Replacement Therapy (HRT)

- In the middle of the Hierarchy of Evidence are systematic observational studies
- Do researchers and doctors make big mistake about life-and-death situations? Yes!
- Large observational studies suggested a reduced risk of Coronary Heart Disease (CHD) among postmenopausal women (e.g., [pubmed 1996](#), [pubmed 2000](#))
- Randomized Control Trial showed the opposite: HRT kills more women!
- Great [Coursera lecture](#) summarizes this fairly complex confounder
  - Time of usage of Hormone Replacement Therapy (HRT)
  - The risk of CHD is highest when you start HRT
- The problem with the observational study?
  - The women who died early on were less likely to get into the study

# Systematic Studies of Observational Studies

- Jim Manzi in the book Uncontrolled summarized papers by Ioannidis showing that 90 percent of large randomized experiments produced results that stood up to replication, as compared to only 20 percent of nonrandomized studies
- Young and Carr looked at 52 claims made in medical observational studies, which were grouped into 12 claims of beneficial treatments (Vitamin E, beta-carotene, Low Fat, Vitamin D, Calcium, etc.)
- These were not random observational studies, but ones that had follow-on controlled experiments (RCTs)
- NONE (zero) of the claims replicated in RCTs, 5 claims were stat-sig in the opposite direction in the RCT
- Their summary  
Any claim coming from an observational study is most likely to be wrong

# The First Medical Controlled Experiment

- The earliest controlled experiment was a test for vegetarianism, suggested in the Old Testament's Book of Daniel

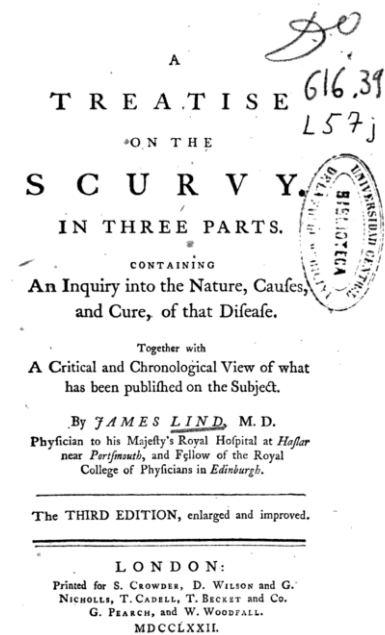
*Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see*

- First controlled experiment / randomized trial for medical purposes
  - Scurvy is a disease that results from vitamin C deficiency
  - It killed over 100,000 people in the 16th-18th centuries, mostly sailors
  - Lord Anson's circumnavigation voyage from 1740 to 1744 started with 1,800 sailors and only about 200 returned; most died from scurvy
  - Dr. James Lind noticed lack of scurvy in Mediterranean ships
  - Gave some sailors limes (treatment), others ate regular diet (control)
  - Experiment was so successful, British sailors are still called limeys
- Amazing scientific triumph, right? Wrong



# The First Medical Controlled Experiment

- Like most stories, the discovery is highly exaggerated
  - The experiment was done on 12 sailors split into 6 pairs
  - Each pair got a different treatment: cider, elixir vitriol, vinegar, sea-water, nutmeg
  - Two sailors were given two oranges and one lemon per day and recovered
  - Lind didn't understand the reason and tried treating Scurvy with concentrated lemon juice called "rob."  
The lemon juice was concentrated by heating it, which destroyed the vitamin C.
  - Working at Haslar hospital, he attended to 300-400 scurvy patients a day for 5 years
  - In his 559 pages massive book [A Treatise on the Scurvy](#), there are two pages about this experiment. Everything else is about other treatments, from Peruvian bark to bloodletting to rubbing the belly with warm olive oil



Lesson: Even when you have a winner, the reasons are often not understood.  
Controlled experiments tell you which variant won, not why.

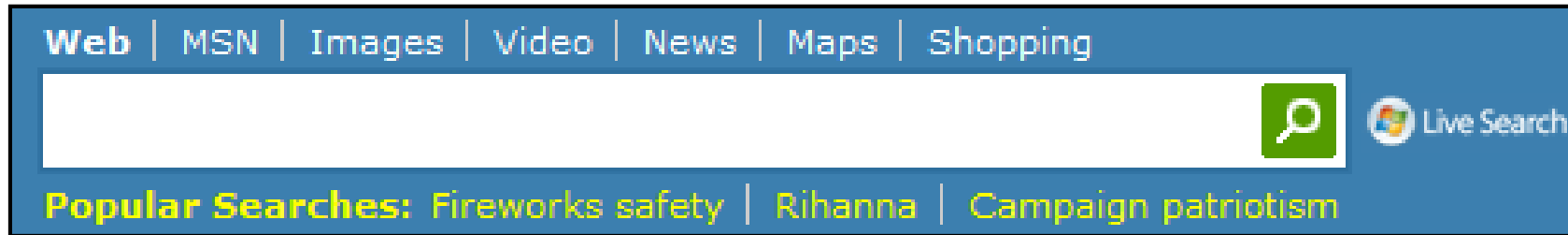
# Real Examples

- Four experiments that ran at Microsoft
- Each provides interesting lessons
- All had enough users for statistical validity
- For each experiment, we provide the OEC, the Overall Evaluation Criterion
  - This is the criterion to determine which variant is the winner
- Game: see how many you get right
  - Everyone please stand up
  - Three choices are:
    - A wins (the difference is statistically significant)
    - A and B are approximately the same (no stat sig diff)
    - B wins
- Since there are 3 choices for each question, random guessing implies  $100\%/3^4 = 1.2\%$  will get all four questions right.  
Let's see how much better than random we can get in this room

# Example 1: MSN Home Page Search Box

➤ OEC: Clickthrough rate for Search box and popular searches

A



B



Differences: A has taller search box (overall size is the same),  
has magnifying glass icon, “popular searches”  
B has big search button, provides popular searches without calling them out

- Raise your left hand if you think A Wins (top)
- Raise your right hand if you think B Wins (bottom)
- Don't raise your hand if they are the about the same

# MSN Home Page Search Box

This slide intentionally left blank

# Example 2: Bing Ads with Site Links

- Should Bing add “site links” to ads, which allow advertisers to offer several destinations on ads?
- OEC: Revenue, ads constraint to same vertical pixels on avg

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!

A

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](#)

B

- Pro adding: richer ads, users better informed where they land
- Cons: Constraint means on average 4 “A” ads vs. 3 “B” ads  
Variant B is 5msc slower (compute + higher page weight)

- Raise your left hand if you think A Wins (left)
- Raise your right hand if you think B Wins (right)
- Don't raise your hand if they are the about the same

# Bing Ads with Site Links

This slide intentionally left blank

# Example 3: SERP Truncation

- SERP is a Search Engine Result Page (shown on the right)
- OEC: Clickthrough Rate on 1<sup>st</sup> SERP per query (ignore issues with click/back, page 2, etc.)
- Version A: show 10 algorithmic results
- Version B: show 8 algorithmic results by removing the last two results
- All else same: task pane, ads, related searches, etc.
- Version B is slightly faster (fewer results means less HTML, but server-side computed same set)

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 being "KDD 2015, 10-13 August 2015, Sydney". The results are numbered 1 through 8. On the right side of the page, there is a sidebar with a "KDD 2015" section, which includes a description of the conference, dates, location, subjects, website, and submission deadline. Below this, there is a "People also search for" section with links to related conferences like ICDM 2015, CIKM 2015, ICML 2015, AAAI 2016, and WWW 2015. At the bottom of the page, there is a pagination control showing "1 2 3 4 5" and a right arrow button.

- Raise your left hand if you think A Wins (10 results)
- Raise your right hand if you think B Wins (8 results)
- Don't raise your hand if they are the about the same

# SERP Truncation

This slide intentionally left blank

# Example 4: Underlining Links

➤ Does underlining increase or decrease clickthrough-rate?

This screenshot shows a Bing search result for 'amazon'. The search bar contains 'amazon' and the results page shows 219,000,000 results. The top result is 'Amazon.com® Official Site - Amazon', which is an advertisement. Below the ad, there are several underlined links: 'Amazon Prime', 'Amazon Echo', 'Fire HDX Tablet', 'Amazon Gift Cards', 'Prime Instant Video', and 'Amazon Fire TV'. The main result for 'Amazon.com' includes the company logo, address (440 Terry Ave N, Seattle, WA 98109), founding date (Jul 06, 1994), and founder (Jeff Bezos). It also displays the stock price (427.63) and a recent post about National Wine Day. At the bottom, there are 'People also search for' results including eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

This screenshot shows a Bing search result for 'amazon', identical to the first one but with non-underlined links. The search bar contains 'amazon' and the results page shows 219,000,000 results. The top result is 'Amazon.com® Official Site - Amazon', which is an advertisement. Below the ad, there are several non-underlined links: 'Amazon Prime', 'Amazon Echo', 'Fire HDX Tablet', 'Amazon Gift Cards', 'Prime Instant Video', and 'Amazon Fire TV'. The main result for 'Amazon.com' includes the company logo, address (440 Terry Ave N, Seattle, WA 98109), founding date (Jul 06, 1994), and founder (Jeff Bezos). It also displays the stock price (427.63) and a recent post about National Wine Day. At the bottom, there are 'People also search for' results including eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

# Example 4: Underlining Links

- Does underlining increase or decrease clickthrough-rate?
- OEC: Clickthrough Rate on 1<sup>st</sup> SERP per query

This screenshot shows the first page of search results for 'amazon' on Bing. The search bar at the top contains 'amazon'. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'More'. The search results are displayed in a grid. The first result is 'Amazon.com® Official Site - Amazon', which is underlined. Other results include 'Amazon.com - Official Site', 'Books', 'Kindle eBooks', 'Shop All Departments', 'Prime Instant Video', 'Movies & TV', 'Amazon Local', 'Electronics', 'Shopping', 'Full Store Directory', 'Outlet', and 'News about Amazon'. The 'Amazon.com' result shows a stock price of 427.63, down 4.00 (-0.93%) from the previous close. The 'Recent post' section shows a link to 'What color is in your glass for National Wine Day?'. The 'People also search for' section includes links to eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

A

This screenshot shows the first page of search results for 'amazon' on Bing, identical to screenshot A, but without the underlines. The search bar at the top contains 'amazon'. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'More'. The search results are displayed in a grid. The first result is 'Amazon.com® Official Site - Amazon', which is not underlined. Other results include 'Amazon.com - Official Site', 'Books', 'Kindle eBooks', 'Shop All Departments', 'Prime Instant Video', 'Movies & TV', 'Amazon Local', 'Electronics', 'Shopping', 'Full Store Directory', 'Outlet', and 'News about Amazon'. The 'Amazon.com' result shows a stock price of 427.63, down 4.00 (-0.93%) from the previous close. The 'Recent post' section shows a link to 'What color is in your glass for National Wine Day?'. The 'People also search for' section includes links to eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

B

- Raise your left hand if you think A Wins (left, with underlines)
- Raise your right hand if you think B Wins (right, without underlines)
- Don't raise your hand if they are the about the same

# Underlines

This slide intentionally left blank

# Agenda

- Introduction and motivation
- Four real examples: you're the decision maker  
Examples chosen to share lessons
- Pitfalls
- Scaling

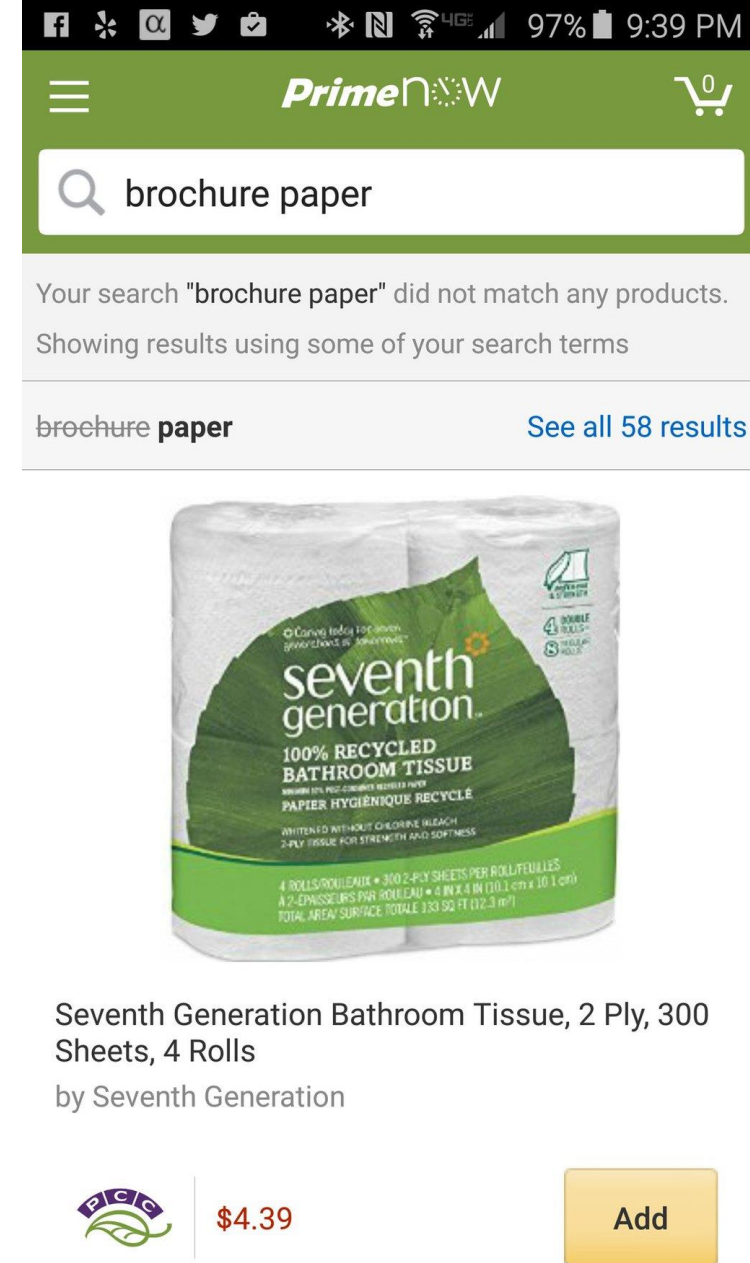
# Pitfall 1: Failing to agree on a good Overall Evaluation Criterion (OEC)

- The biggest issues with teams that start to experiment is making sure they
  - Agree what they are optimizing for
  - Agree on measurable short-term metrics that predict the long-term value (and hard to game)
- Microsoft support example with time on site
- Bing example
  - Bing optimizes for long-term query share (% of queries in market) and long-term revenue. Short term it's easy to make money by showing more ads, but we know it increases abandonment. Revenue is a constraint optimization problem: given an agreed avg pixels/query, optimize revenue
  - Queries/user may seem like a good metric, but degrading results will cause users to search more. Sessions/user is a much better metric (see <http://bit.ly/expPuzzling>). Bing modifies its OEC every year as our understanding improves. We still don't have a good way to measure "instant answers," where users don't click

# Bad OEC Example

- Your data scientists makes an observation: 2% of queries end up with “No results.”
- Manager: must reduce. Assigns a team to minimize “no results” metric
- Metric improves, but results for query **brochure paper** are crap (or in this case, paper to clean crap)
- Sometimes it *\*is\** better to show “No Results.” This is a good example of gaming the OEC.

Real example from my Amazon Prime now search 3/26/2016  
<https://twitter.com/ronnyk/status/713949552823263234>



The screenshot shows the Amazon Prime Now mobile app interface. At the top, there are social media icons and a battery level of 97% at 9:39 PM. The Prime Now logo is in the top right. A search bar contains the text "brochure paper". Below the search bar, a message states: "Your search 'brochure paper' did not match any products. Showing results using some of your search terms". Underneath, the search term "brochure paper" is displayed, followed by a link to "See all 58 results". The main product shown is a roll of Seventh Generation Bathroom Tissue, 2 Ply, 300 Sheets, 4 Rolls. The product packaging is white with a green leaf design and text that reads: "seventh generation 100% RECYCLED BATHROOM TISSUE". Below the product image, the text reads: "Seventh Generation Bathroom Tissue, 2 Ply, 300 Sheets, 4 Rolls by Seventh Generation". The price is listed as \$4.39, and there is an "Add" button.

# Pitfall 2: Misinterpreting P-values

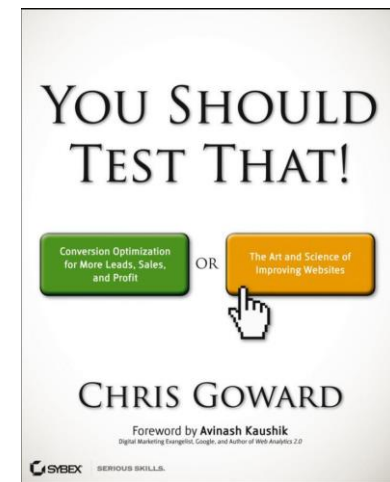
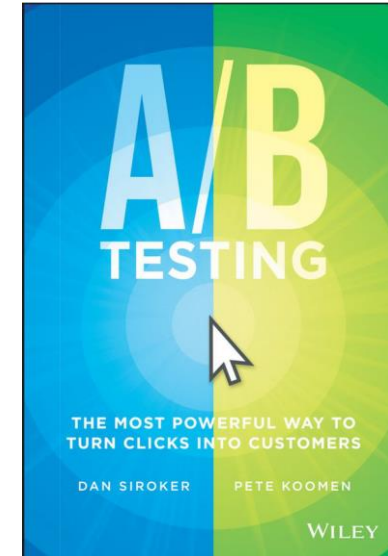
- NHST = Null Hypothesis Statistical Testing, the “standard” model commonly used
- P-value  $\leq 0.05$  is the “standard” for rejecting the Null hypothesis
- P-value is often mis-interpreted.

Here are some incorrect statements from Steve Goodman’s A Dirty Dozen

1. If  $P = .05$ , the null hypothesis has only a 5% chance of being true
  2. A non-significant difference (e.g.,  $P > .05$ ) means there is no difference between groups
  3.  $P = .05$  means that we have observed data that would occur only 5% of the time under the null hypothesis
  4.  $P = .05$  means that if you reject the null hyp, the probability of a type I error (false positive) is only 5%
- The problem is that p-value gives us  $\text{Prob}(X \geq x \mid H_0)$ , whereas what we want is  $\text{Prob}(H_0 \mid X = x)$

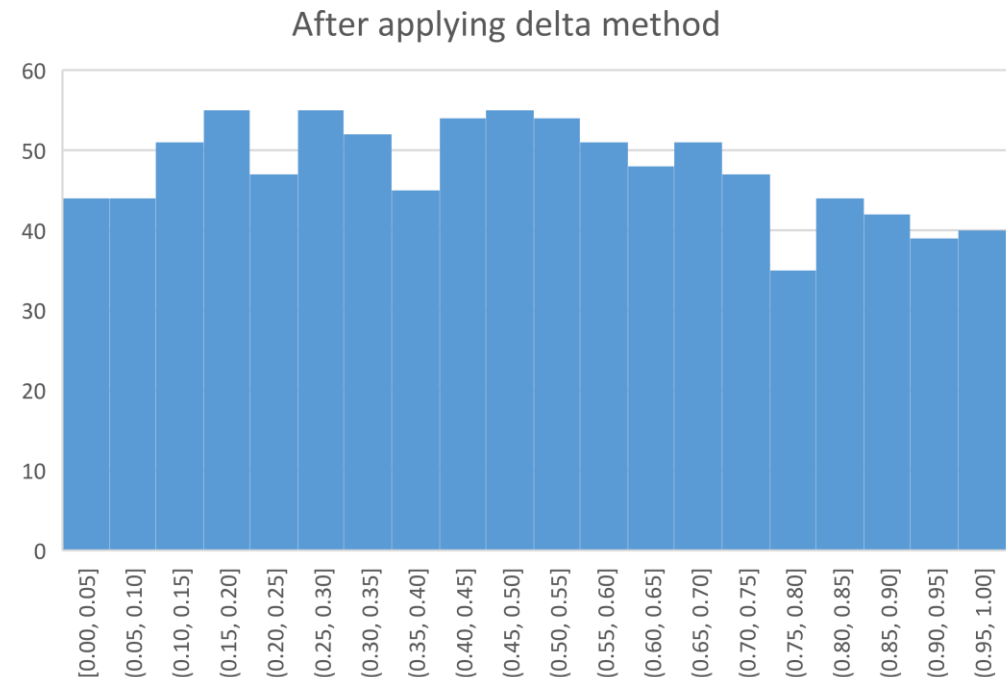
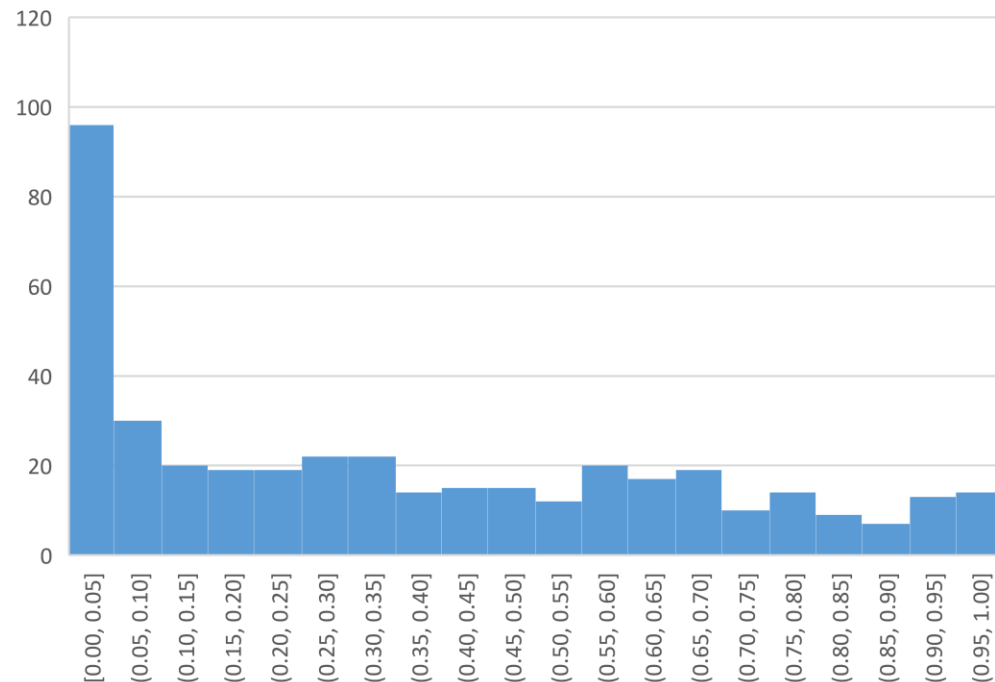
# Pitfall 3: Failing to Validate the Experimentation System

- 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**
- Example: Two very good books on A/B testing get the stats wrong (see Amazon reviews)
- Recommendation:
  - Run A/A tests: if the system is operating correctly, the system should find a stat-sig difference only about 5% of the time
  - Do a Sample-Ratio-Mismatch test. Example
    - Design calls for equal percentages to Control Treatment
    - Real example: Actual is 821,588 vs. 815,482 users, a 50.2% ratio instead of 50.0%
    - Something is wrong! Stop analyzing the result.  
The p-value for such a split is  $1.8e-6$ , so this should be rarer than 1 in 500,000.  
SRMs happens to us every week!



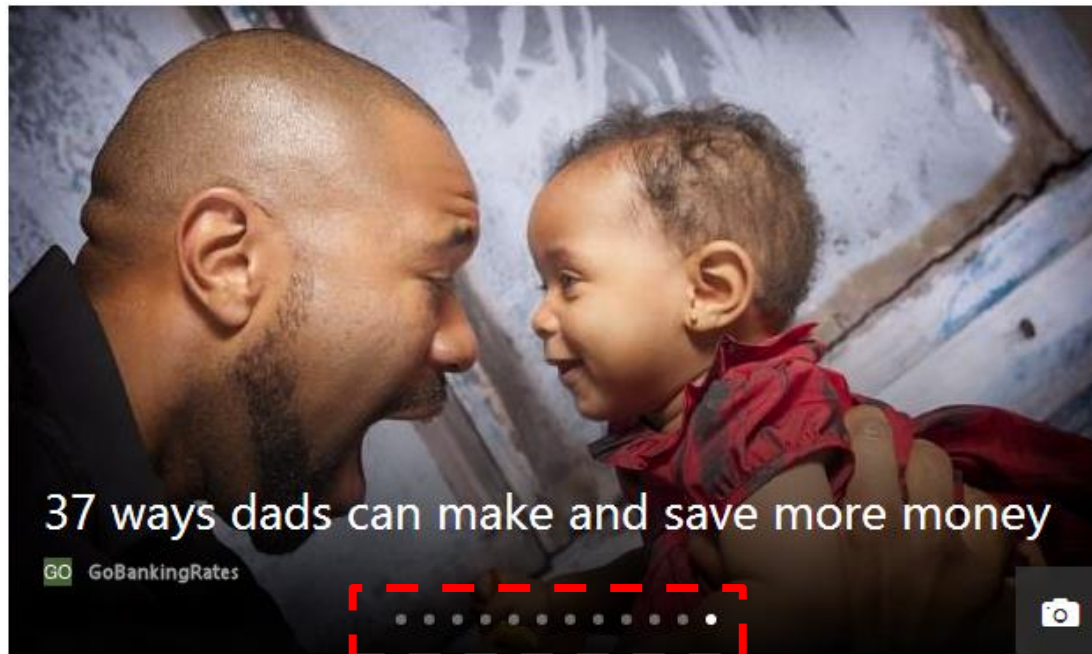
# Example A/A test

- P-value distribution for metrics in A/A tests should be uniform
- When we got this for some Skype metrics, we had to correct things

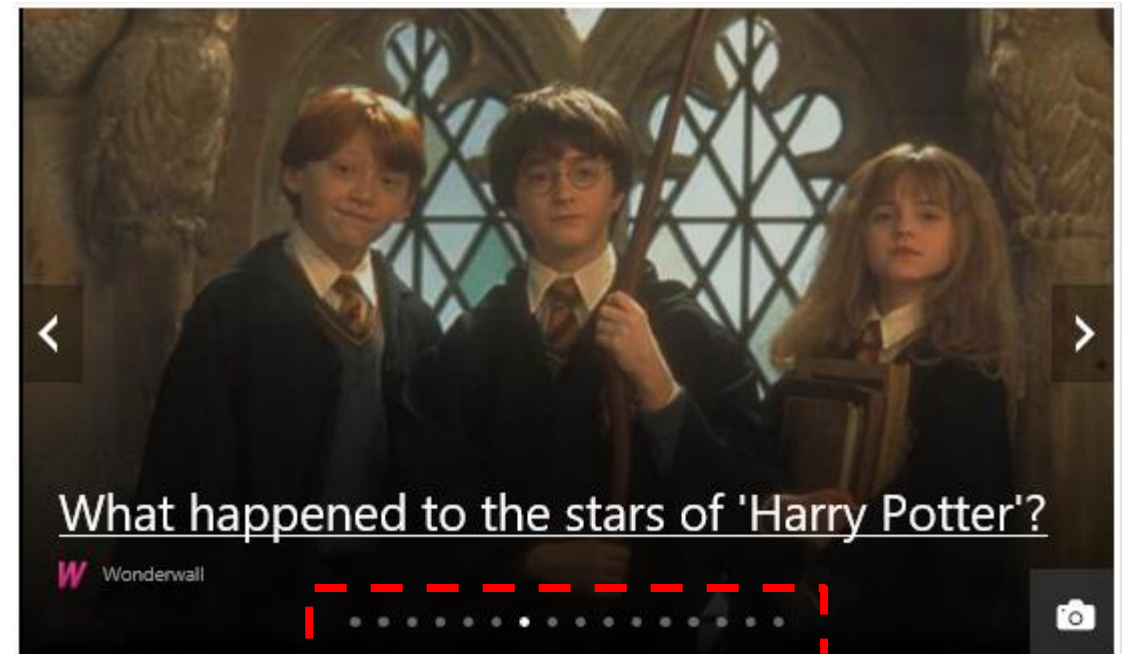


# MSN Experiment: Add More Infopane Slides

The infopane is the “hero” image at MSN, and it auto rotates between slides with manual option



**Control: 12 slides**



**Treatment: 16 slides**

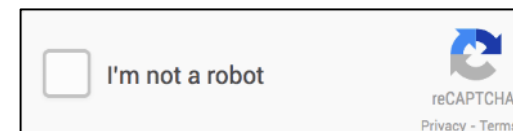
Control was much better than treatment for engagement (clicks, page views)

# MSN Experiment: SRM

- Except... there was a sample-ratio-mismatch with fewer users in treatment (49.8% instead of 50.0%)
- Can anyone think of a reason?
- User engagement increased so much for so many users, that the heaviest users were being classified as bots and removed
- After fixing the issue, the SRM went away, and the treatment was much better

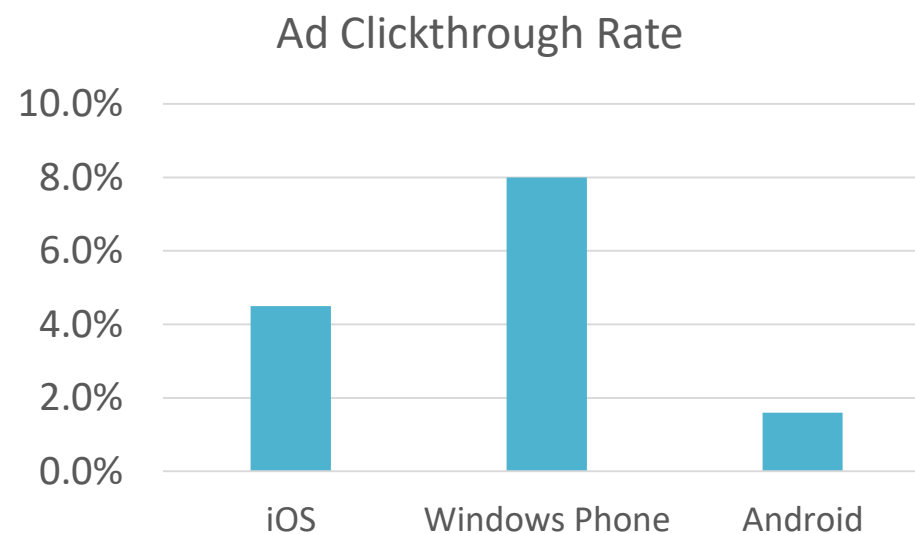
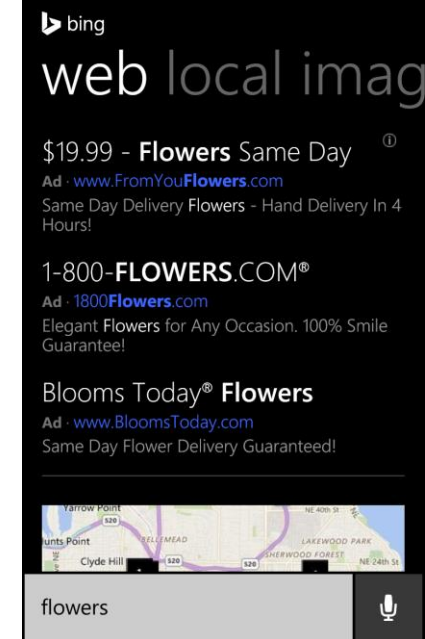
# Pitfall 4: Failing to Validate Data Quality

- Outliers create significant skew: enough to cause a false stat-sig result
- Example:
  - An experiment treatment with 100,000 users on Amazon, where 2% convert with an average of \$30. Total revenue =  $100,000 * 2% * \$30 = \$60,000$ . A lift of 2% is \$1,200
  - Sometimes (rarely) a “user” purchases double the lift amount, or around \$2,500. That single user who falls into Control or Treatment is enough to significantly skew the result.
  - The discovery: libraries purchase books irregularly and order a lot each time
  - Solution: cap the attribute value of single users to the 99<sup>th</sup> percentile of the distribution
- Example:
  - Bots at Bing sometimes issue many queries
  - Over 50% of Bing traffic is currently identified as non-human (bot) traffic!



# Instrumentation is Critical

- Ad clickthrough rate(\*) for Bing mobile appeared very different
- All sorts of hypotheses developed about how Windows Phone users are better, then iOS, then Android
- The delta is almost all due to instrumentation artifacts for clicks
  - iOS used redirects (slow but reliable)
  - Android used asynchronous click tracking (fast but lossy)
  - Windows Phone used asynchronous click tracking. Bug caused swipe actions to register as clicks



(\*) y-values scaled by constant factor

# The Best Data Scientists are Skeptics

- The most common bias is to accept good results and investigate bad results. When something is too good to be true, remember Twyman

## Twyman's law

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

<http://bit.ly/twymanLaw>

# Pitfall 5: Failing to Keep it Simple

- With Offline experiments, many experiments are “one shot” so a whole science developed of how to make efficient use of a single large/complex experiment. For example, it’s very common offline to vary many factors at the same time (Multivariable experiments, fractional-factorial designs, etc)
- In Software, new features tend to have bugs
  - Assume a new feature has 10% probability of having an egregious issue
  - If you combine seven such features in an experiment, then the probability of failure is  $1-(0.9^7)=52\%$  so would have to abort half the experiments
- Examples
  - LinkedIn unified search attempted multiple changes that had to be isolated when things failed. See “Rule #6: Avoid Complex Designs: Iterate” at <http://bit.ly/expRulesOfThumb>
  - Multiple large changes at Bing failed because too many new things were attempted

# The Risks of Simple

➤ There are two major concerns we have heard about why “simple” is bad

1. Leads to incrementalism.

We disagree. Few treatments does not equate to avoiding radical designs.

Instead of doing an MVT with 6 factors, each with 3 values each thus creating  $3^6 = 729$  combinations, have a good designer come up with four radical designs.

2. Interactions.

A. We do run multivariable designs when we suspect strong interactions, but they are small (e.g., 3x3). Our observation is that interactions are relatively rare: hill-climbing single factors works well.

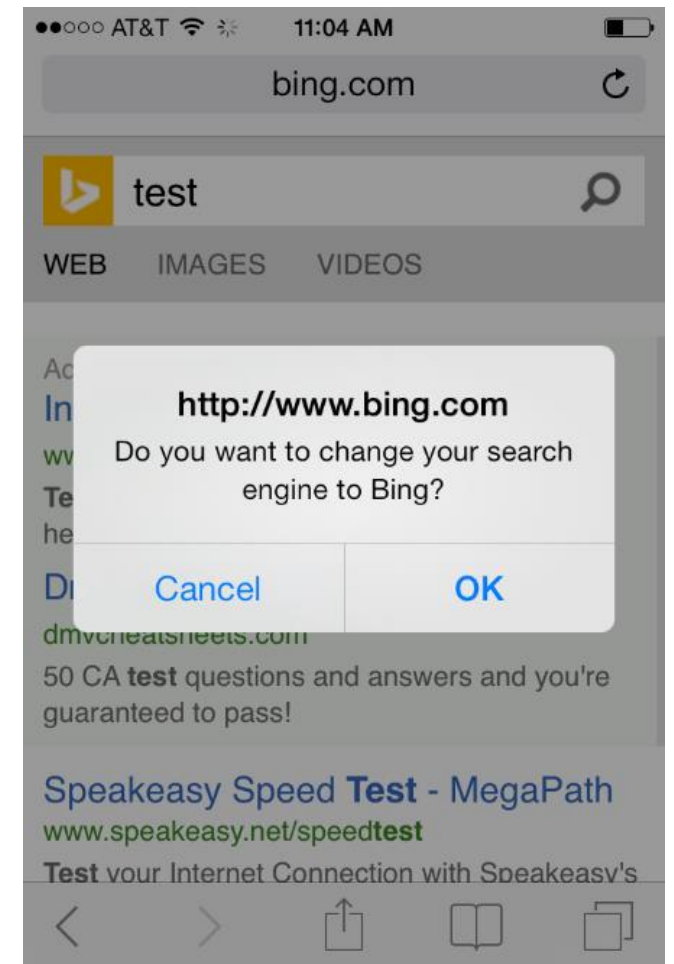
B. Because our experiments run concurrently, we are in effect running a full-factorial design. Every night we look at all pairs of experiments to look for interactions. With hundreds of experiments running “optimistically” (assuming no interaction), we find about 0-2 interactions per week and these are usually due to bugs, not because the features really interact

# Pitfall 6: Failing to Look at Segments

- It is easy to run experiments and look at the average treatment effect for “ship” or “no-ship” decision
- Our most interesting insights have come from realizing that the treatment effect is different for different segments (heterogeneous treatment effects)
- Examples
  - We had a bug in treatment that caused Internet Explorer 7 to hang. The number of users of IE7 is small, but the impact was so negative, it was enough to make the average treatment effect negative. The feature was positive for the other browsers
  - Date is a critical variable to look at. Weekend vs. Weekdays may be different. More common: a big delta day-over-day may indicate an interacting experiment or misconfiguration
- We automatically highlight segments of interest to our experimenters

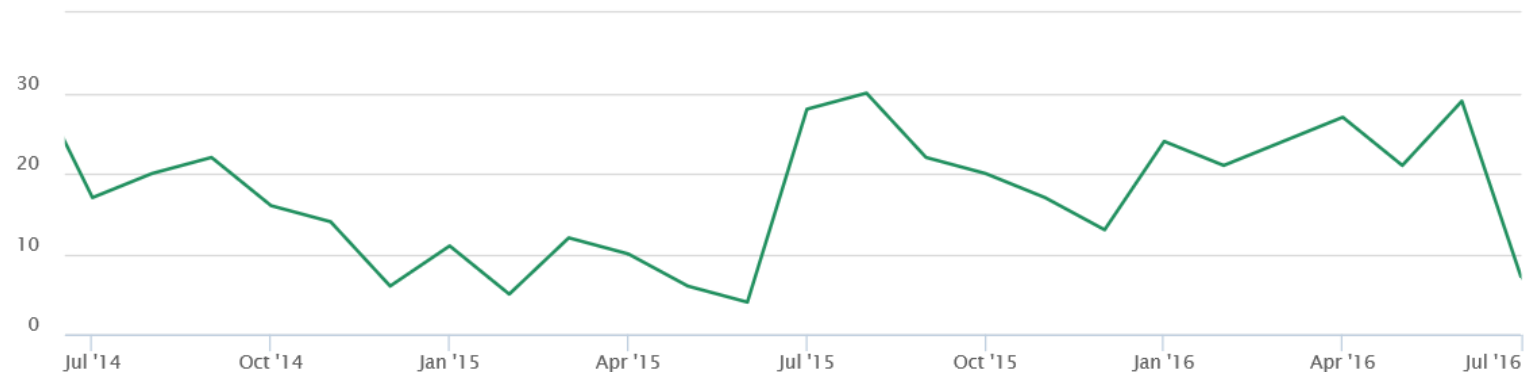
# Example: New Bing Mobile Release

- Major revamp of Bing UX for mobile happened (a few years back)
- Scorecard shows a large drop in Sessions/User, our north star metric. Treatment was losing users relative to control. Terrible!
- Segmenting by browser showed iOS issue
- Segmenting by iOS showed iOS7 issue
- Root cause: missing JavaScript resource that implemented a pop-up asking users if they want to change their search engine to Bing



# Pitfall 7: Failing to Evaluate Early

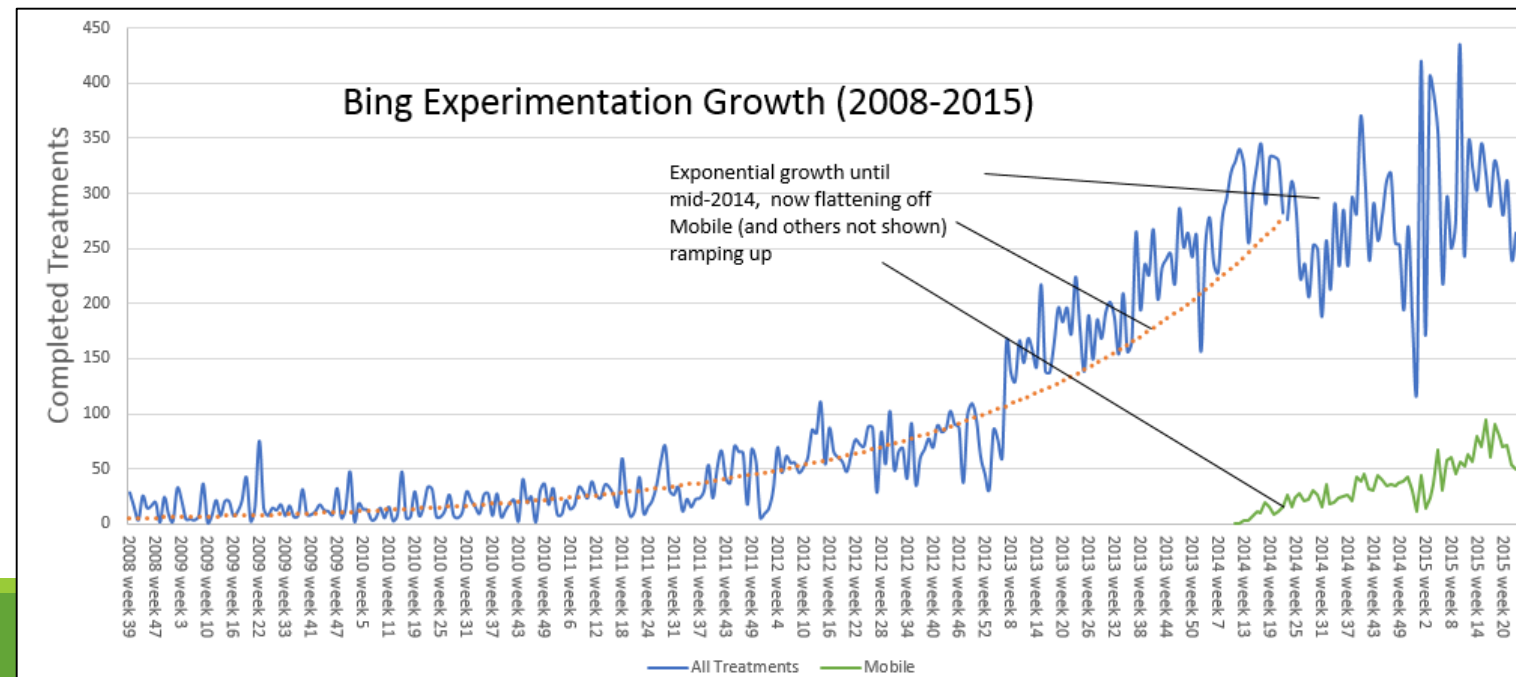
- In theory, the experimenter does the power calculation and that determines the number of users required and hence the duration (we commonly run for 1 week or 2 weeks)
- Early “peeking” is discouraged by the Statistics, as it leads to multiple-hypothesis testing issues
- However, in the software world, early detection of bugs is critical, and the most critical time is at the start of an experiment
- Graph of the number of experiments (per month) that are shutdown by our automated system.  
Critical!



Number of experiments/month that are shutdown due to metric degradation

# Scale

- We finish about ~300 experiment treatments per week, mostly on Bing, MSN, but also on Office (client and online), OneNote, Xbox, Cortana, Skype, and Exchange. (These are “real” useful treatments, not  $3 \times 10 \times 10$  MVT = 300.)
- Typical treatment is exposed to millions of users, sometimes tens of millions.
- There is no single Bing. Since a user is exposed to over 15 concurrent experiments, they get one of  $5^{15} = 30$  billion variants (debugging takes a new meaning)
- Until 2014, the system was limiting usage as it scaled. Now limits come from engineers' ability to code new ideas



# Scaling Experimentation (1 of 2)

- Cultural Challenges (see <http://bit.ly/KDD2015Kohavi>)  
Not discussed here
- What is a valuable experiment?
  - **Absolute value of delta between expected outcome and actual outcome is large**
  - If you thought something is going to win and it wins, you have not learned much
  - If you thought it was going to win and it loses, it's valuable (learning)
  - If you thought it was “meh” and it was a breakthrough, it's HIGHLY valuable
- Managing idea pipeline / roadmap (more at <http://bit.ly/quoraABRoadmap>)
  - The critical limiting resource is feature development cost.  
If an idea is easy to A/B test, stop the debates and just run the test
  - When looking at ROI, the Return in many cases is uncorrelated with the Investment.  
At Bing, highly successful experiments worth over \$100M/year each, took days to develop.  
Conversely, Bing's social integration took over 100 person years to develop and had little value.
  - Develop MVPs (minimum viable products) and reprioritize based on data

# Scaling Experimentation (2 of 2)

- Technical challenges (see more at <http://bit.ly/ExPScale>)
  - Ensure trustworthy results
    - Automatically run tests that detect that something is wrong (e.g., SRM)
    - Warn about unexpected variations over time (e.g., a single day is significantly different)
  - Lower the cost of experimentation
    - Can experimenters define new metrics easily?
    - What about integrating a new source of data, such as end-user feedback?
  - Test for interacting experiments
    - Every night we test for interactions between all pairs of overlapping treatments (must use control false positives through a mechanism like FDR: False Discovery Rate)
  - Detect egregiously bad experiments early
    - We start experiments at low percentage and ramp-up if results look OK
    - We test for big movements for guardrail metrics in near-real-time (15 minutes)
  - Debugging tools to help understand why some metric moved e.g., for Bing, which queries caused a given metric to move

# Challenges (1 of 2)

- Oct 2015 MIT CODE talk on challenges at <https://bit.ly/CODE2015Kohavi>
- OEC: Overall Evaluation Criteria
  - What are good OECs or different domains? Challenge is to define short-term metrics that are predictive of long-term impact (e.g., lifetime value)
- Improving sensitivity /reducing variance (e.g., [CUPED](#))
- Bayesian methods
  - When success is rare (e.g., Sessions/UU improves in only 0.02% of experiments), we have to correct for the classical hypothesis testing, which has 5% false positive rate (with p-value 0.05).  
How do we use historical data to compute the posterior that a metric really moved?
- Deep analyses: what segments improved/degraded?  
Use of machine learning techniques to find these
- Form factors
  - Reasonable understanding of web page design for desktop
  - Weak understanding of small-screen (e.g., mobile), touch interactions, apps

# Challenges (2 of 2)

- Are there long-term impacts that we are not seeing in 1-2 week experiments?  
See Google's paper in KDD 2015
- Aborting experiments
  - With 30-50 experiments starting every day at Bing, some have bugs.  
What are good metrics that have high statistical power to detect issues in near-real-time?
- “Leaks” from experiment variants
  - Social experiments (see multiple papers by Facebook)
  - Consumption of shared resources (e.g., memory/disk by one variant).  
In a well-remembered case, one treatment consumed memory slowly causing the servers to crash, but you see very little in the controlled experiment results
- Also See Ya Xu's [A/B testing challenges in social networks](#)

# The HiPPO

- HiPPO = Highest Paid Person's Opinion
- We made thousands toy HiPPOs and handed them at Microsoft to help change the culture
- Fact: Hippos kill more humans than any other (non-human) mammal
- Listen to the customers and don't let the HiPPO kill good ideas
- There is a box with several hundred HiPPOs outside, and booklets with selected papers



# Summary

*The less data, the stronger the opinions*

- Think about the **OEC**. Make sure the org agrees **what** to optimize
- It is hard to assess the value of ideas
  - Listen to your customers – **Get the data**
  - **Prepare to be humbled**: data trumps intuition
- Compute the statistics carefully
  - Getting numbers is easy. Getting a number you can **trust** is harder
- Experiment often
  - Triple your experiment rate and you triple your success (and failure) rate. Fail fast & often in order to succeed
  - Accelerate innovation by lowering the cost of experimenting
- See <http://exp-platform.com> for papers
- More at Twitter: @RonnyK

# Extra Slides

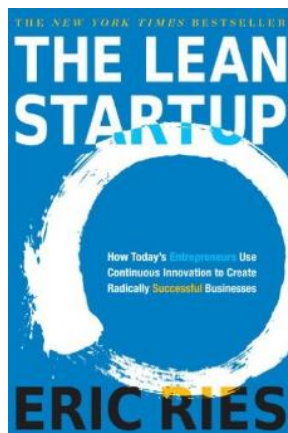
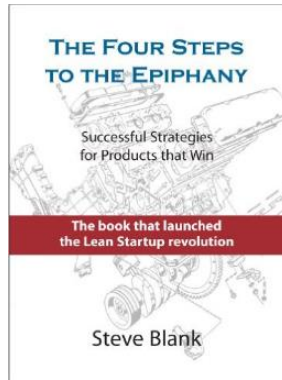
# Motivation: Product Development

*It doesn't matter how beautiful your theory is,  
it doesn't matter how smart you are.*

*If it doesn't agree with experiment[s], it's wrong*

*-- Richard Feynman*

- Classical software development: spec->dev->test->release
- Customer-driven Development: Build->Measure->Learn (continuous deployment cycles)
  - Described in Steve Blank's *The Four Steps to the Epiphany* (2005)
  - Popularized by Eric Ries' *The Lean Startup* (2011)
  - Build a Minimum Viable Product (MVP), or feature, cheaply
  - Evaluate it with real users in a **controlled experiment (e.g., A/B test)**
  - Iterate (or pivot) based on learnings
- Why use Customer-driven Development?  
Because we are poor at assessing the value of our ideas
- Why I love controlled experiments  
In many data mining scenarios, interesting discoveries are made and promptly ignored. In customer-driven development, the mining of data from the controlled experiments and insight generation is part of the critical path to the product release



# There are Never Enough Users

- Assume a metric of interest, say revenue/user
  - Denote the variance of the metric by  $\sigma^2$
  - Denote the sensitivity, i.e., the amount of change we want to detect by  $\Delta$
- From statistical power calculations, the number of users ( $n$ ) required in experiment is proportional to  $\sigma^2 / \Delta^2$
- The problem
  - Many key metrics have high-variance (e.g., Sessions/User, Revenue/user)
  - As the site is optimized more, and as the product grows, we are interested in detecting smaller changes (smaller  $\Delta$ )
- Example: A commerce site runs experiments to detect 2% change to revenue and needs 100K users per variant.  
For Bing US to detect 0.1% (\$2M/year), we need  $20^2 \times 100K = 40M \times 2$  variants = 80M users (Bing US has about 100M users/month)

# Personalized Correlated Recommendations

- Actual personalized recommendations from Amazon.  
(I was director of data mining and personalization at Amazon back in 2003, so I can ridicule my work.)
- Buy anti aging serum because you bought an LED light bulb  
(Maybe the wrinkles show?)
- Buy Atonement movie DVD because you bought a Maglite flashlight  
(must be a dark movie)
- Buy Organic Virgin Olive Oil because you bought Toilet Paper.  
(If there is causality here, it's probably in the other direction.)

**Hyaluronic Acid Serum for Skin. Organic Natural Skincare for Face. Intense Moisture and Vitamin C for the Best Anti Aging and Anti Wrinkle Serum on Amazon for Men & Women.**  
by Sano Naturals (December 4, 2014)  
Average Customer Review: ★★★★★ (59)  
In Stock  
List Price: \$49.99  
Price: \$13.95  
Offered by Sano Naturals  
Add to Cart Add to Wish List

I own it  Not interested  ★★★★★ Rate this item  
Recommended because you purchased LED Light Bulb - High QUALITY - The BEST Energy Efficient... (Fix this)

**Atonement (Widescreen Edition)**  
DVD ~ Keira Knightley (Mar 18, 2008)  
Average Customer Review: ★★★★★ (99)  
In Stock  
List Price: \$29.98  
Price: \$15.99  
24 used & new from \$13.77  
Add to cart

I own it  Not interested  ★★★★★ Rate it  
Recommended because you purchased Mag Instrument Three Cell AA Mini Maglite LED Flashlight.

---

**Zoe Organic Extra Virgin Olive Oil, 25.5-Ounce Tins (Pack)**  
by Zoe  
Average Customer Review: ★★★★★ (21)  
Usually ships in 3 to 4 weeks  
List Price: \$26.64  
Price: \$15.40  
Add to Cart

I own it  Not interested  ★★★★★ Rate this item  
Recommended because you purchased Cottonelle Ultra Toilet Paper Double Roll, White 176, 12... a