eMetrics SUMMIT

San Francisco

May 17, 2017

Trustworthy A/B Tests: Pitfalls in Online Controlled Experiments

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

@RonnyK

Joint work with many members of the ExP platform team

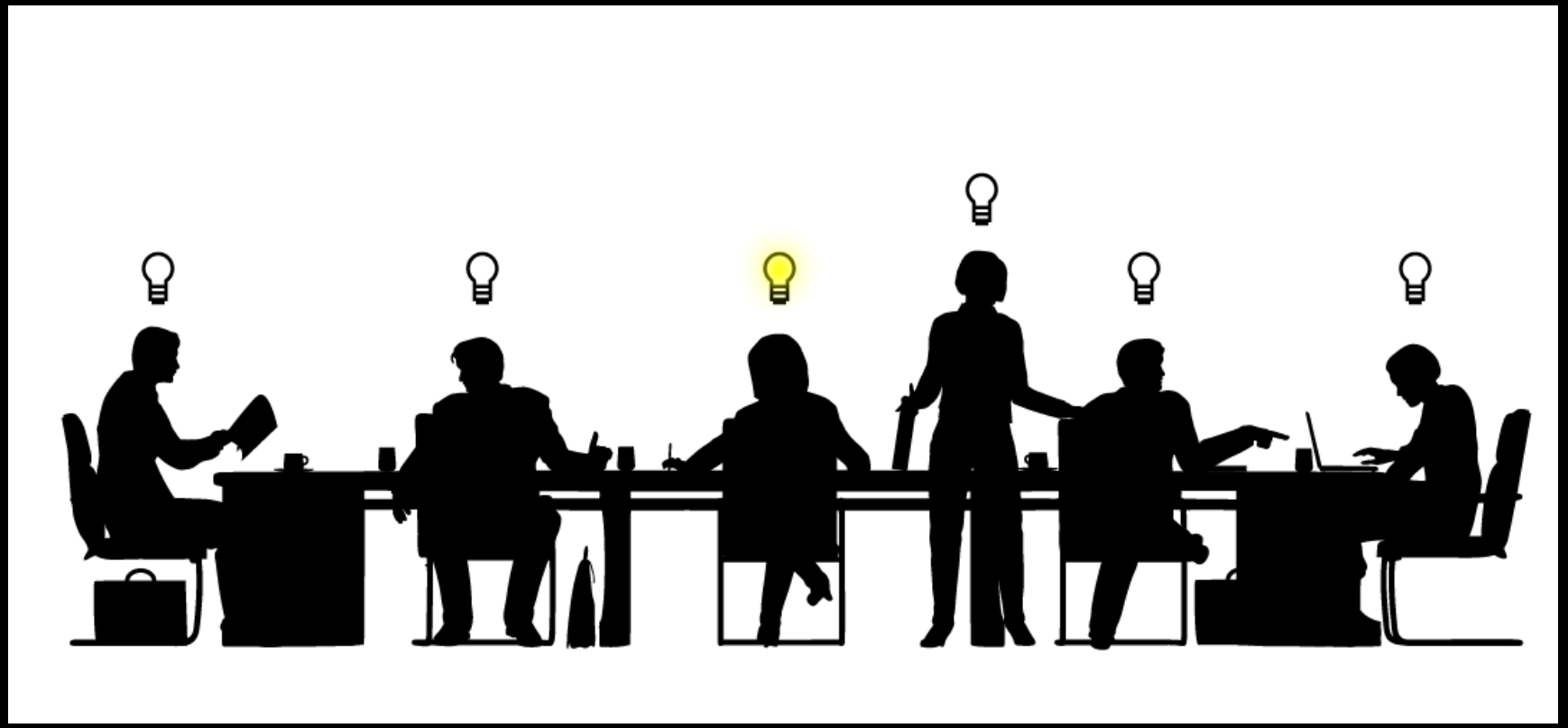
The Life of a Great Idea

- An idea was proposed in early 2012 to change the way ad titles were displayed on Bing.
  It's one of hundreds of ideas proposed
- Implementation was delayed in: Feb March April May June
- Multiple features were stack-ranked as more valuable
- It wasn't clear if it was going to be done by the end of the year
- An engineer thought: this is trivial to implement.
  He implemented the idea in a few days, and started a controlled experiment (A/B test) to evaluate the idea
- An alert fired that something is wrong with revenue, as Bing was making too much money.
  Such alerts have been very useful to detect bugs (such as logging revenue twice)
- But there was no bug. The idea increased Bing's revenue by 12% without hurting guardrail metrics
- Hundreds of engineers work on Bing Ads and increase revenue by 1.5% in a good month.
  This was worth the equivalent of over 100 person years of work
- 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 was badly rated and delayed for months!

Agenda

- Introduction and motivation
- The experimentation platform
- Valuable experiments
- Lessons and pitfalls

We have lots of ideas to make our products better…



But not all ideas are good.
It’s humbling, but most ideas are actually bad

It's Hard to Assess the Value of Ideas

- 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

- Our observations 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

- Similar observations made by many others
  - "Google ran approximately 12,000 randomized experiments in 2009, with [only] about 10 percent of these leading to business changes" was stated in *Uncontrolled* by Jim Manzi
  - "80% of the time you/we are wrong about what a customer wants" was stated in *Experimentation and Testing: A Primer* by Avinash Kaushik, author of *Web Analytics* and *Web Analytics 2.0*
  - QualPro tested 150,000 ideas over 22 years and founder Charles Holland stated, "75 percent of important business decisions and business improvement ideas either have no impact on performance or actually hurt performance…" in *Breakthrough Business Results With MVT*
  - At Amazon, half of the experiments failed to show improvement

How can we tell which ideas are good?

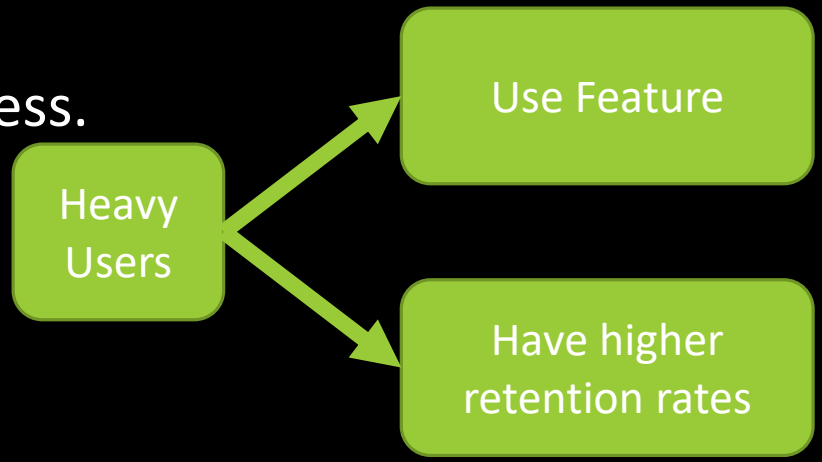
It’s very hard (often impossible) to tell ahead of time which ideas will good. Hence, we want to try many ideas and measure their impact to determine which ones are good.

How can we reliably measure the impact of our changes?

Let’s first examine some commonly used but **flawed** techniques…

Your Feature Reduces Churn!

- You observe the churn rates for users using/not-using your feature:
  - 25% of new users that do NOT use your feature churn (stop using product 30 days later)
  - 10% of new users that use your feature churn
- [Wrong] Conclusion: your feature reduces churn and thus critical for retention
- Flaw: Relationship between the feature and retention is correlational and not causal
- The feature may improve or degrade retention: the data above is insufficient for any causal conclusion
- Example: Users who see error messages in Office 365 churn less. This does NOT mean we should show more error messages. They are just heavier users of Office 365
- See Best Refuted Causal Claims from Observations Studies for great examples of this common analysis flaw



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

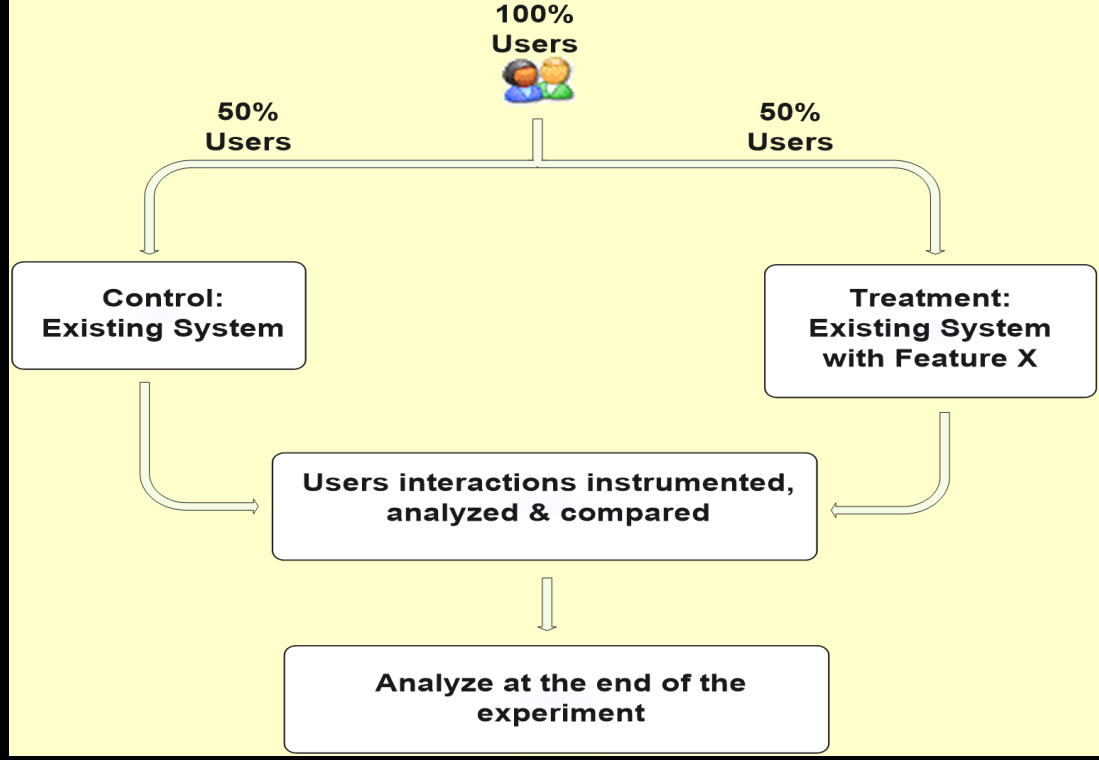
- Try it out - look at your neighbors and you'll see who is expected to live longer
- But…don't try to bandage your hands, as there is a **common cause**
- Women have smaller palms and live 6 years longer on average

Measure user behavior before/after ship

Flaw: This approach misses time related factors such as external events, weekends, holidays, seasonality, etc.

The new site (B) always does worse than the original (A)

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: Flights (Microsoft), 1% Tests (Google), Bucket tests (Yahoo!), 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

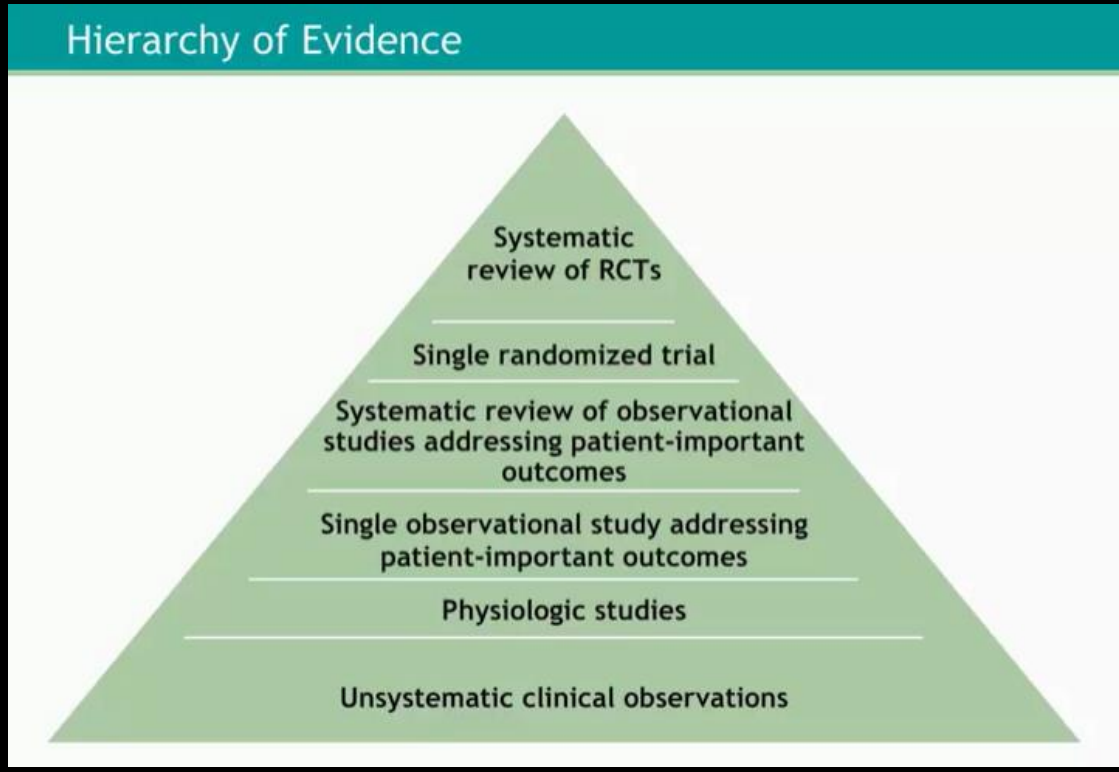
- Controlled experiments test for causal relationships, not simply correlations
- When the variants run concurrently, only two things could explain a change in metrics:
  1. The "feature(s)" (A vs. B)
  2. Random chance

  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
- See http://bit.ly/refutedCausalClaims

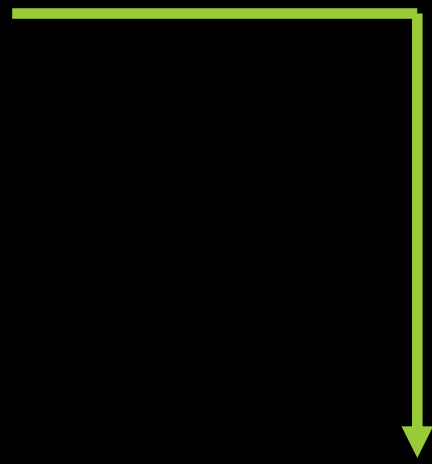


Systematic Studies of Observational Studies

- 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

Agenda

- Introduction and motivation
- The experimentation platform
- Valuable experiments
- Lessons and pitfalls



Why I can talk about controlled experiments and pitfalls.
I'm not here to sell you anything. Our platform is internal.
The 5%/80%/15% rule for machine learning and experimentation.
    We worked hard to reduce that 80% (and thus the 15%)

About the Team

- Analysis and Experimentation team at Microsoft:
  - Mission: Accelerate innovation through trustworthy analysis and experimentation. Empower the HiPPO (Highest Paid Person's Opinion) with data
  - About 80 people
    - 40 developers: build the Experimentation Platform and Analysis Tools
    - 30 data scientists
    - 8 Program Managers
    - 2 overhead (me, admin).

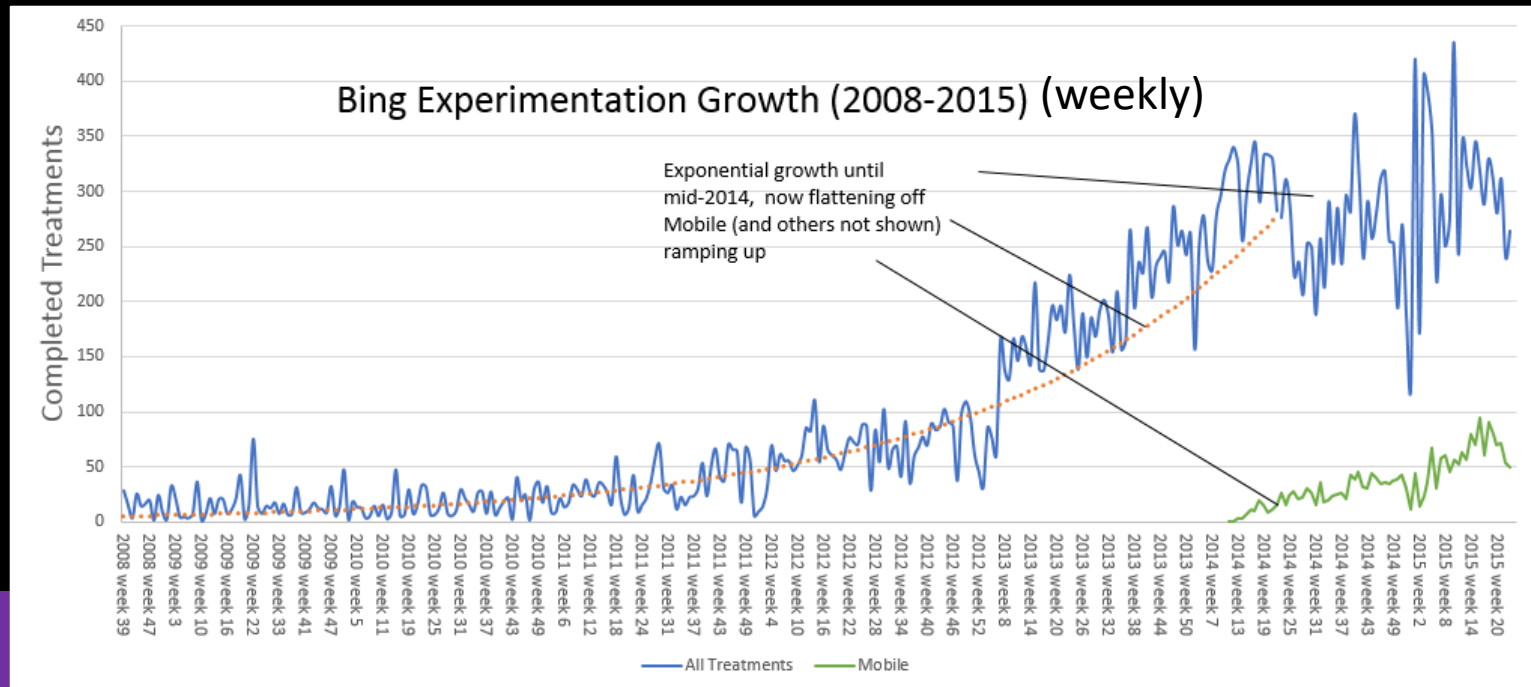
    Team includes people who worked at Amazon, Facebook, Google, LinkedIn

The Experimentation Platform

- Experimentation Platform provides full experiment-lifecycle management
  - Experimenter sets up experiment (several design templates) and hits “Start”
  - Pre-experiment “gates” have to pass (specific to team, such as perf test, basic correctness)
  - System finds a good split to control/treatment (“seedfinder”).
    Tries hundreds of splits, evaluates them on last week of data, picks the best
  - System initiates experiment at low percentage and/or at a single Data Center.
    Computes near-real-time “cheap” metric and aborts in 15 minutes if there is a problem
  - System wait for several hours and computes more metrics.
    If guardrails are crossed, auto shut down; otherwise, ramps-up to desired percentage (e.g., 20% of users for each variant)
  - After a day, system computes many more metrics (e.g., thousand+) and sends e-mail alerts about interesting movements (e.g., time-to-success on browser-X is down D%)

Experimentation at Scale

- We run over 1,000 experiment treatments per month on Bing, and over 100 per month across Office (client and online), OneNote, Xbox, Cortana, Skype, and Exchange.
  (These are "real" useful treatments, not 3x10x10 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



Agenda

- Introduction and motivation
- The experimentation platform
- Valuable experiments
- Lessons and pitfalls

What is the value of an 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

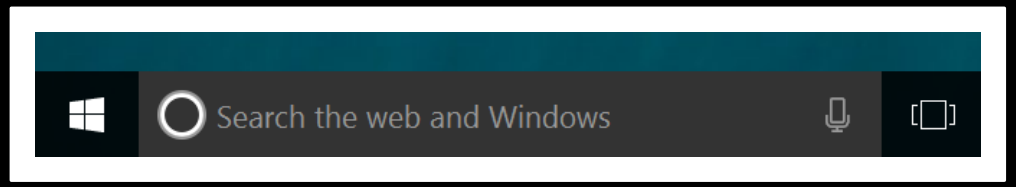
Real Examples

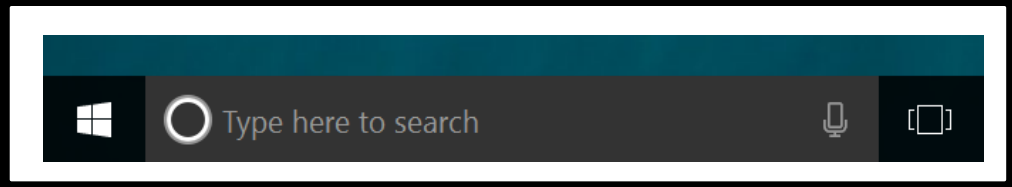
- Three experiments that ran at Microsoft
- Each provides interesting lessons
- All had enough users for statistical validity
- For each experiment, we provide the OEC, the Overall Evaluation Criterion
  - This is the criterion to determine which variant is the winner
- Let's see how many you get right
  - Everyone please stand up
  - You will be given three options and you will answer by raising you left hand, right hand, or leave both hand down (details per example)
  - If you get it wrong, please sit down
- Since there are 3 choices for each question, random guessing implies 100%/3^3 =~ 4% will get all three questions right.
  Let's see how much better than random we can get in this room

Windows Search Box

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

Here are two variants:





OEC (Overall Evaluation Criterion): user engagement--more searches (and thus Bing revenue)

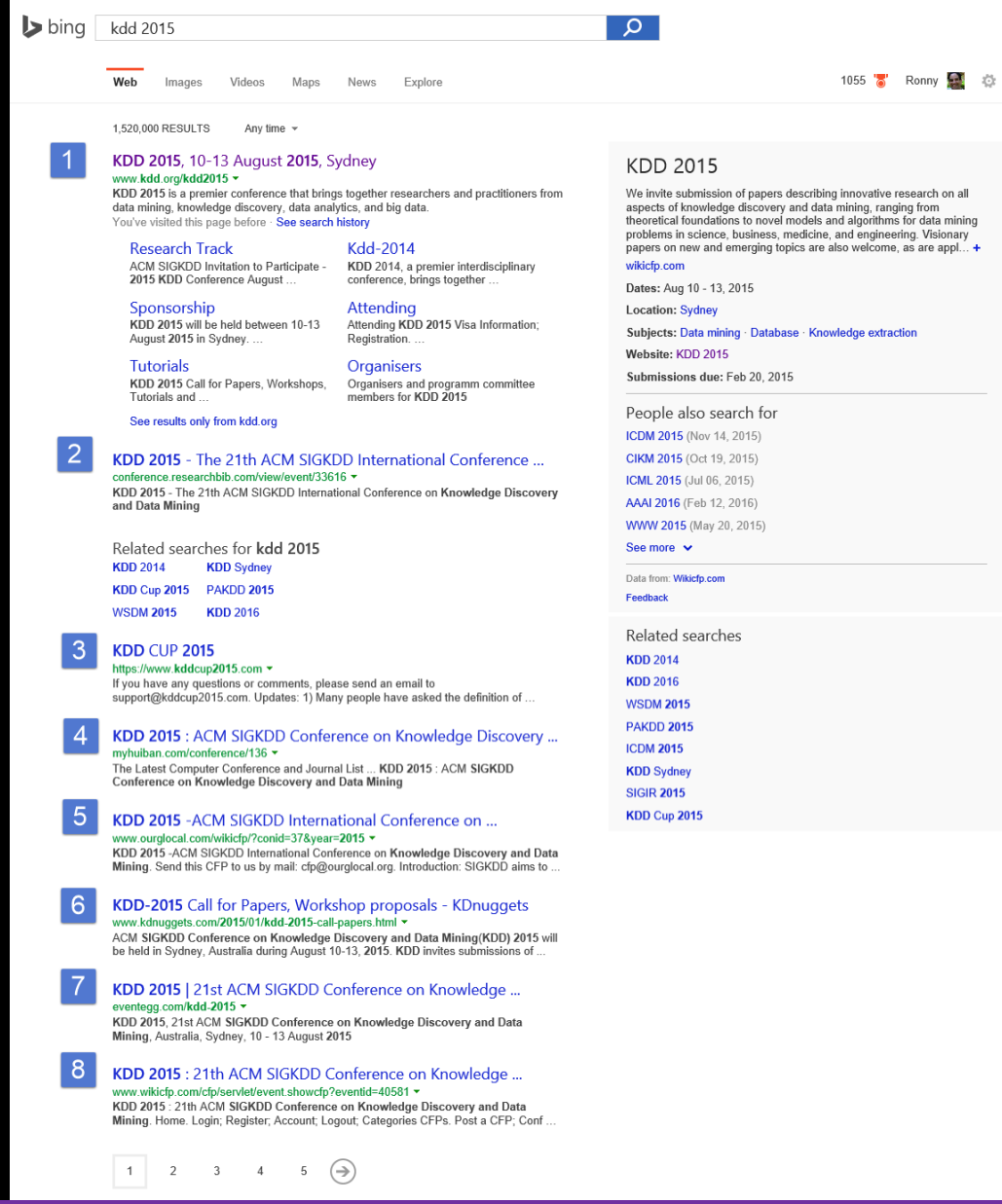
- Raise your left hand if you think the Left version wins (stat-sig)
- Raise your right hand if you think the Right version wins (stat-sig)
- Don’t raise your hand if they are the about the same

Windows Search Box (cont)

- This slide intentionally missing content

Example 2: SERP Truncation

- SERP is a Search Engine Result Page (shown on the right)
- OEC: Clickthrough Rate on 1st 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)



- 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

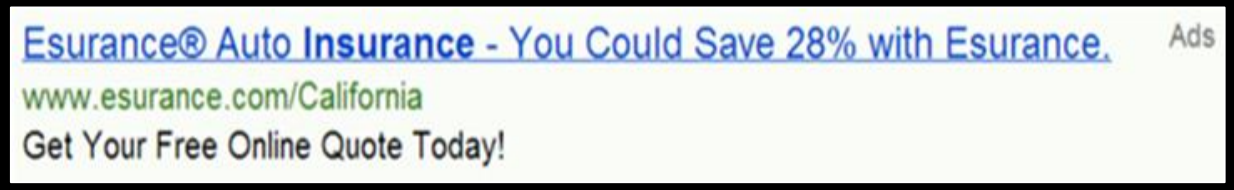
SERP Truncation

- This slide intentionally missing content
- We 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 3: 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



- 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 Left version wins
- Raise your right hand if you think Right version wins
- Don’t raise your hand if they are the about the same

Bing Ads with Site Links

- This slide intentionally missing content

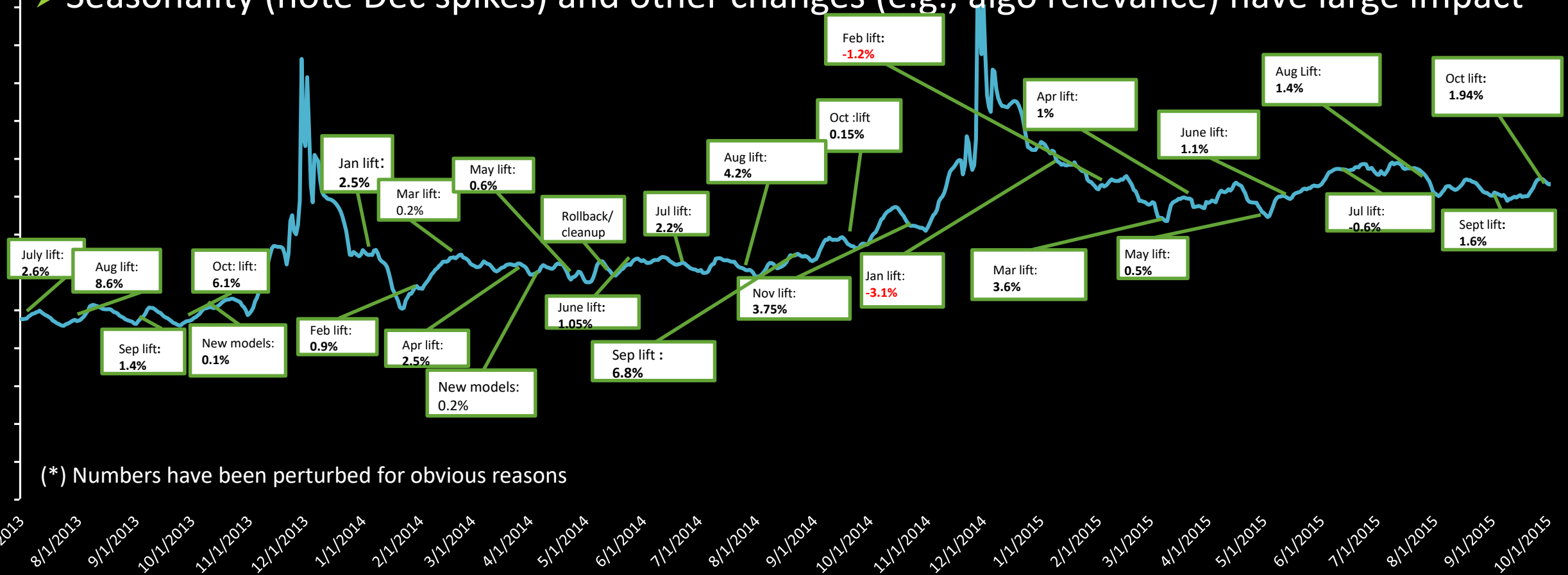
- Stop debating – get the data

Breakthroughs are Rare

- In talks like this, we focus on breakthrough examples
- As Al Pacino says in the movie Any Given Sunday, winning is done inch by inch
- Most progress is made by small continuous improvements: 0.1%-1% after a lot of work. Rare are the experiments that improve overall revenue by 2% or more
- It is very hard to improve key organizational metrics (not vanity metrics)
  - For Bing, Sessions/User is a key metric, but improving it is extremely rare (1 in 5000 experiments). Most "successes" are false positives and do not replicate
- The next slide shows the inch-by-inch progress of Bing ads

Bing Ads Revenue per Search(*) – Inch by Inch

- Emarketer estimates Bing revenue grew 55% from 2014 to 2016 (not official statement)
- About every month a "package" is shipped, the result of many experiments
- Improvements are typically small (sometimes lower revenue, impacted by space budget)
- Seasonality (note Dec spikes) and other changes (e.g., algo relevance) have large impact



(*) Numbers have been perturbed for obvious reasons

Agenda

- Introduction and motivation
- The experimentation platform
- Valuable experiments
- Lessons and pitfalls

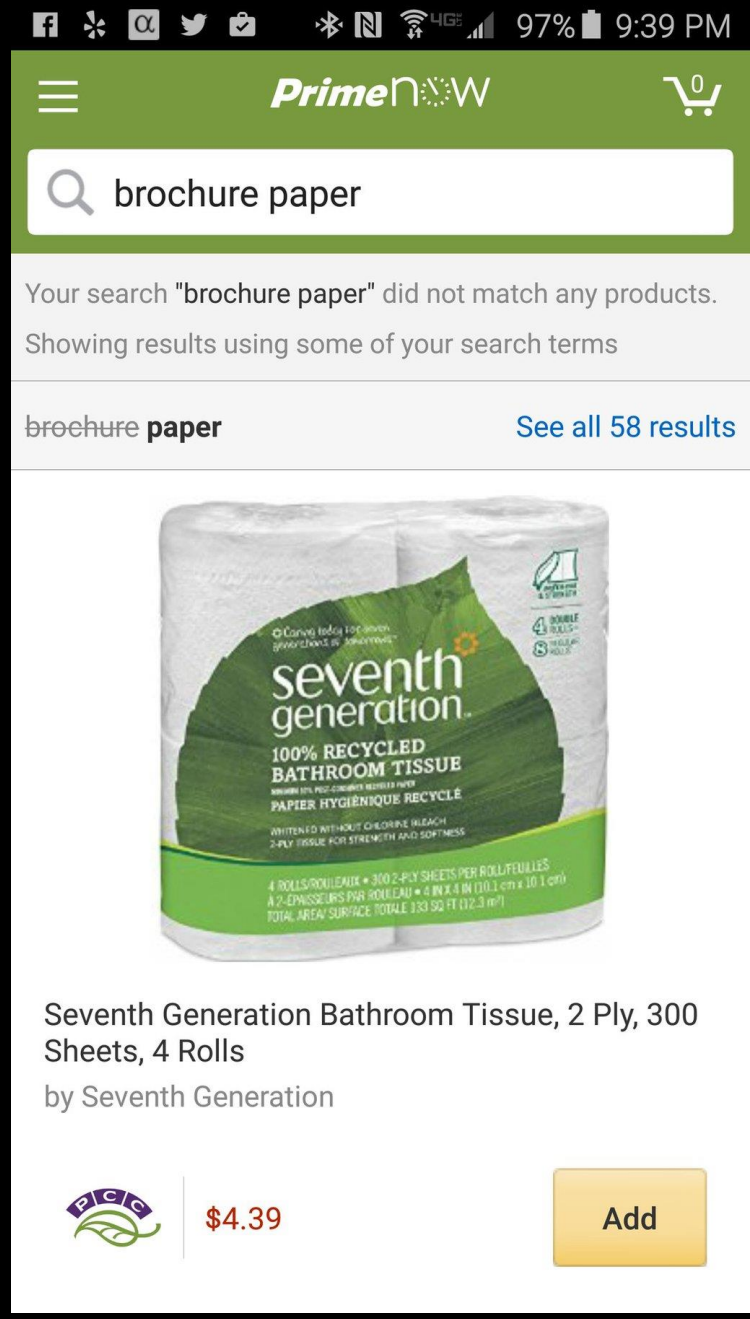
Pitfall 1: Failing to agree on a good Overall Evaluation Criterion (OEC)

- 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
https://twitter.com/ronnyk/status/713949552823263234



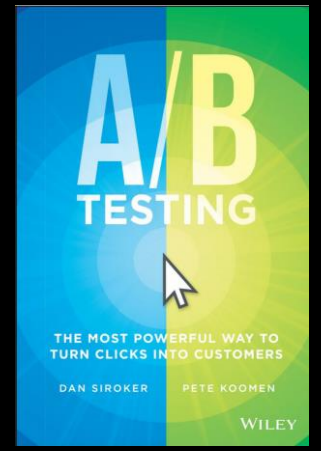
Pitfall 2: Misinterpreting P-values

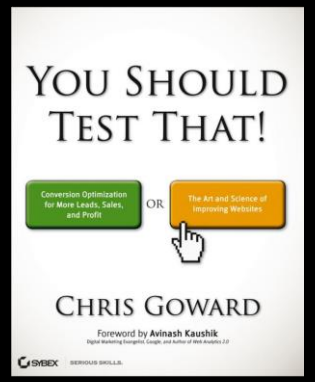
- NHST = Null Hypothesis Statistical Testing, the “standard” model commonly used
- P-value <= 0.05 is the “standard” for rejecting the Null hypothesis
- P-value is often mis-interpreted.
  Here are some incorrect statements from Steve Goodman’s A Dirty Dozen
  1. If P = .05, the null hypothesis has only 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 Prob (X >= x | H_0), whereas what we want is Prob (H_0 | X = x)

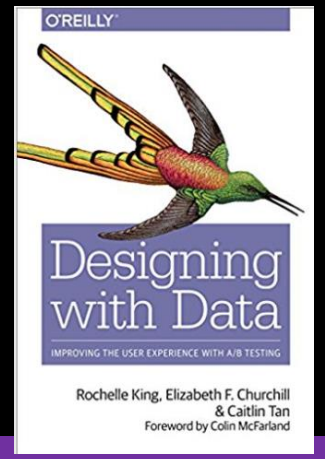
Pitfall 3: Failing to Validate the Experimentation System

- 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). The recent book on Designing with Data also gets its wrong
- 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!

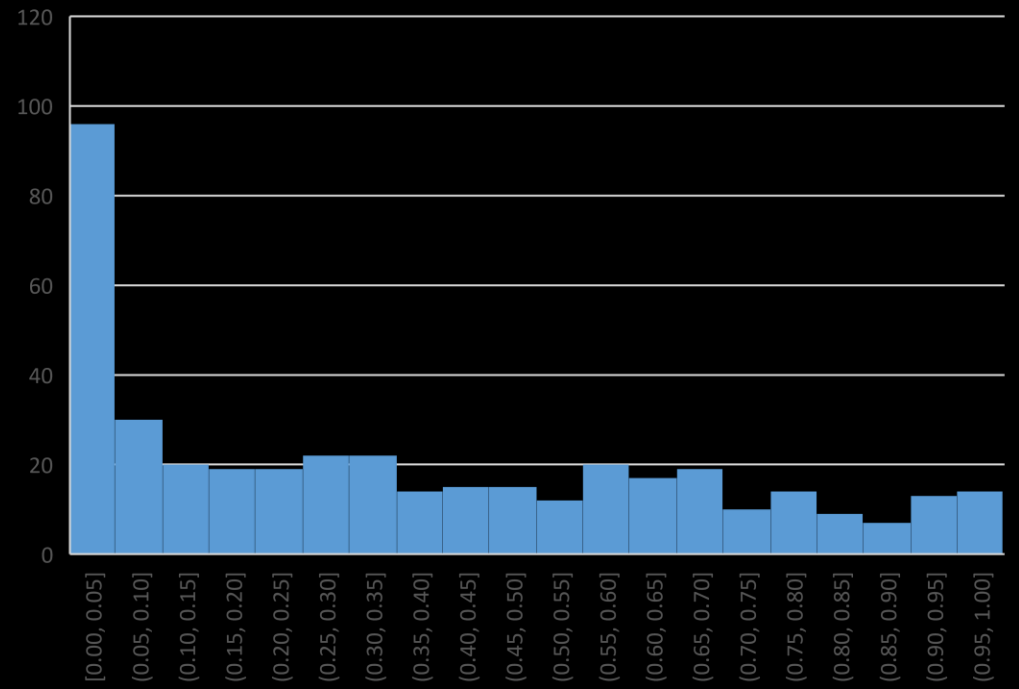


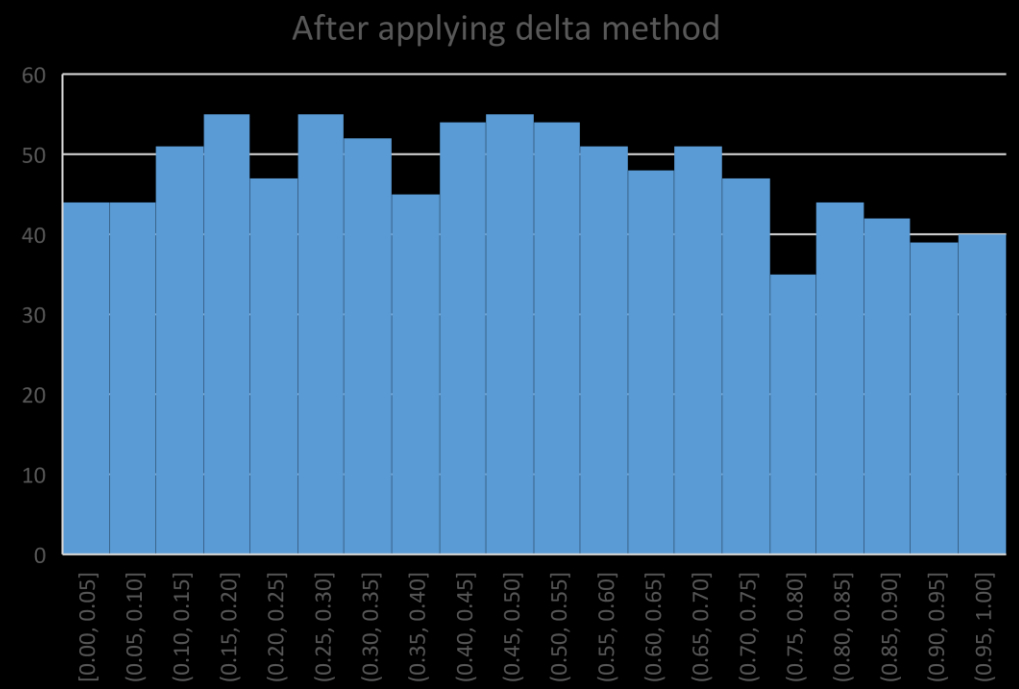




Example A/A test

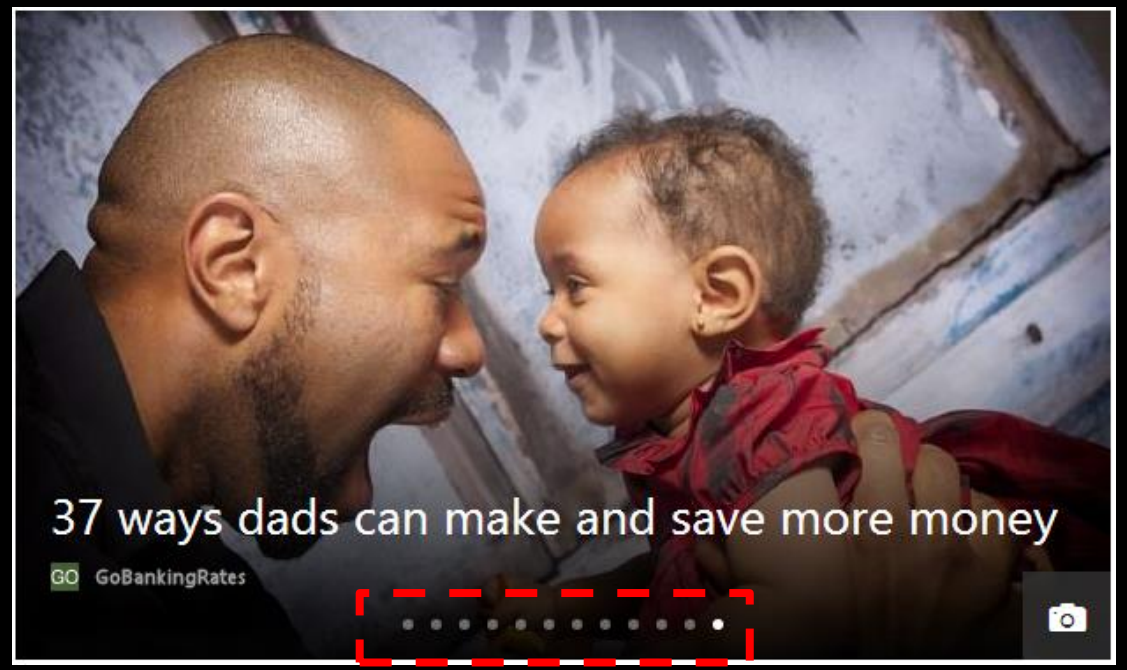
- P-value distribution for metrics in A/A tests should be uniform
- Do 1,000 A/A tests, and check if the distribution is uniform
- When we got this for some Skype metrics, we had to correct things





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)

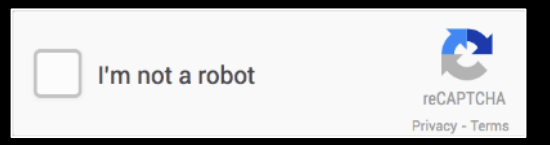
MSN Experiment: SRM

- 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 99th 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!



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

- In the book *Exploring Data: An Introduction to Data Analysis for Social Scientists,* the authors wrote that Twyman's law is

"perhaps the most important single law in the whole of data analysis."

http://bit.ly/twymanLaw

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 HiPPOs outside, and booklets with selected papers for you to take

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
- This talk is at **http://bit.ly/emetrics2017expPitfalls (or see @RonnyK)**