Seven Rules of Thumb for Web Site Experimenters

First two rules in this talk, rest in PPT appendix
Slides at http://bit.ly/expRulesOfThumb

Ronny Kohavi
Joint work with Alex Deng, Roger Longbotham, Ya Xu

Can We Generalize?

- We have been involved in thousands of experiments
  - Bing and LinkedIn run thousands of experiments per year
  - Experimentation Platform at Microsoft: experiments at over 20 Microsoft properties
  - Roger and Ronny have prior experience from Amazon; Ya Xu at LinkedIn
- Rules of thumb
  - Generalizations from experiments
  - Mostly true, exceptions may be known
  - Similar to financial rule of 72: for interest rate x, 72/x is the time to double the money. Accurate for 4-12% range, where most people are interested in
- Useful for discussions. Will evolve over time, as we understand applicability. We want your feedback!

Data and Process

- All examples are real
- Users randomly sampled, sufficient sample sizes of at least 100K users to millions of users
- Based on statistical significance (p-value < 0.05).
  Surprising result always replicated, and Fisher's Combined Probability Test from the two experiments results in much lower p-values.
- Experiments scrutinized for common pitfalls, so we believe they are trustworthy

Rule #1: Small Changes can have a Big Impact to Key Metrics

- It is easy for small changes to have a big negative impact on key metrics.
  - JavaScript error makes checkout impossible
  - Users on some browser unable to click (this happened to us in a Bing experiment)
  - Servers crashing
- Our focus is on positive differences due to small changes
- We are also not interested in short-term novelty effects.
  - Colleen Szot changed three words to a standard infomercial line.
    Huge increase in the number of people who purchased her product.
  - Instead of the all-too-familiar "Operators are waiting, please call now," it was
    "If operators are busy, please call again."
  - Her show shattered a twenty-year sales record
  - Nice ploy showing the value of "social proof" (must be hot product if everyone is buying), but will have short shelf-life
- We are interested in high sustained ROI

Example: Font Colors

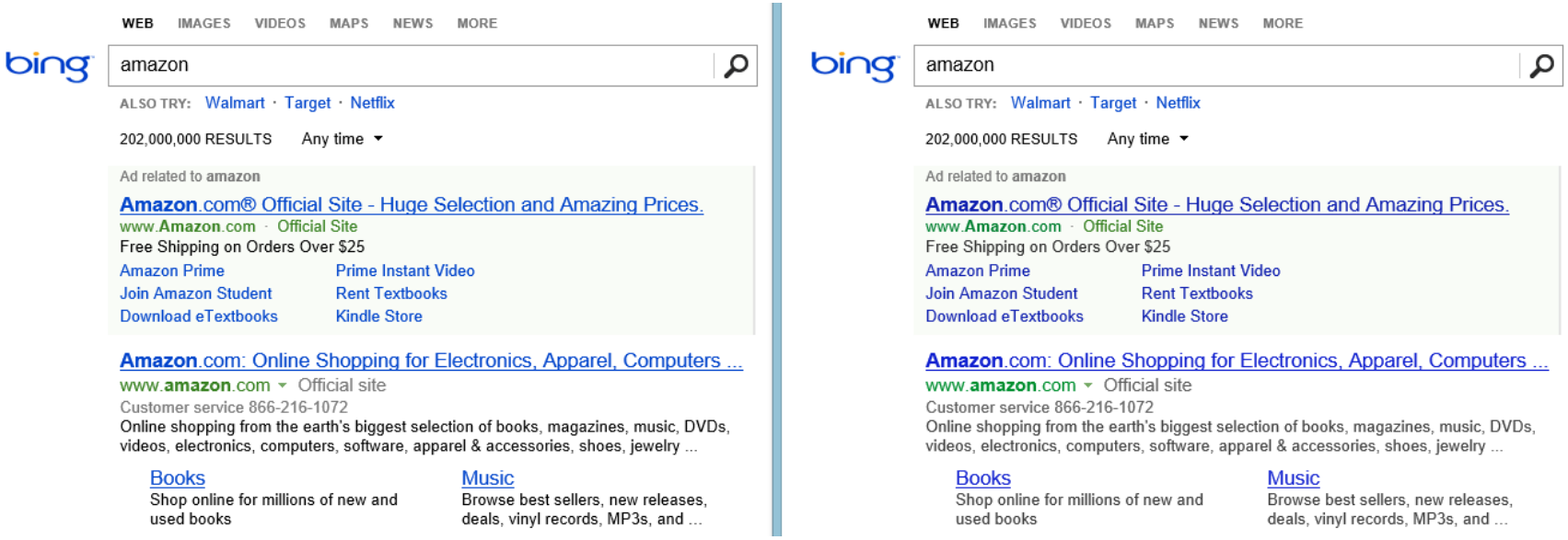


Figure 1: Font color experiment. Can you tell the difference?

- Hard to even tell the difference

Example: Font Colors

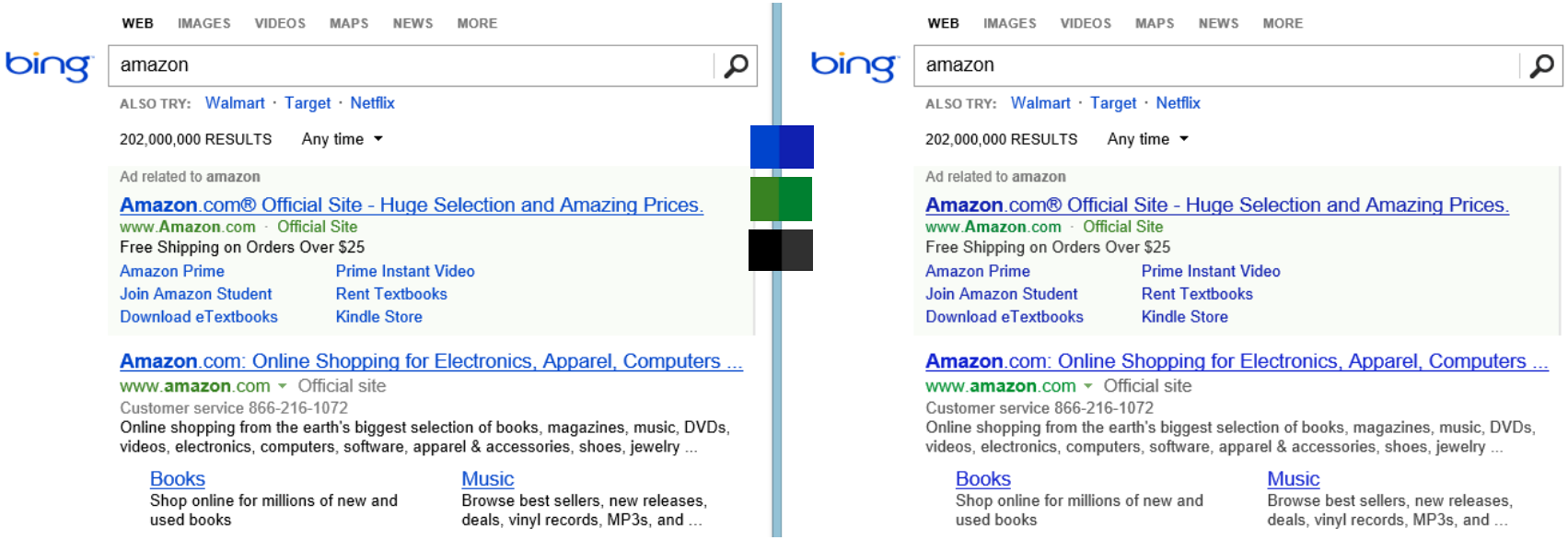


Figure 1: Font color experiment. Can you tell the difference?

- Hard to even tell the difference
- Change is trivial: a few numbers change in the CSS
- Sessions success rate improved, time-to-success improved, +$10M annually

Example: Right Offer at the Right Time

- Amazon in 2004 auto-optimized home page slots
- Amazon's credit-card offer was winning the top slot
  - Surprising because it had very low clickthrough-rate
  - Highly profitable, so expected value was high
- Moved offer to shopping cart (clear intent to purchase)

- This simple change was worth tens of millions of dollars in profit annually.

Example: Anti-Malware

- Ads are a lucrative business, and “freeware” installed by users often contains malware that pollutes pages with ads
- The red areas are showing the actual experience for Bing’s SERP
- Experiment blocked changes to the DOM
- Results improved Sessions/user, Session Success Rate, Time to Success. Page Load Time improved by hundreds of milliseconds for the triggered pages

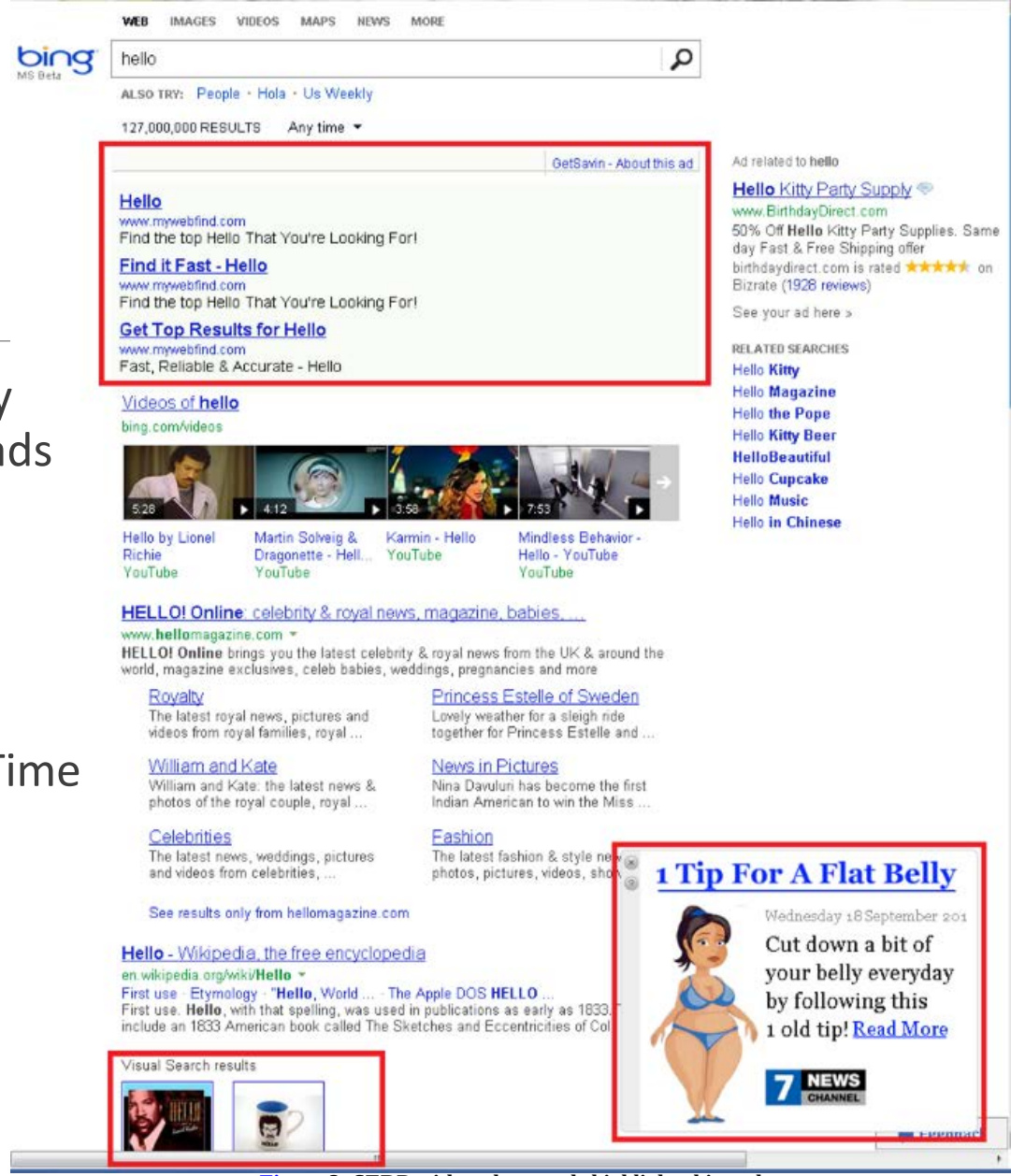


Figure 2: SERP with malware ads highlighted in red

Risks

- Focusing on breakthroughs is tough, as they are rare, maybe 1 in 500 experiments at Bing
- Avoiding Incrementalism: an organization should test small changes that potentially have high ROI, but also take some big bets for the Big Hairy Audacious Goals (from Built to Last book).

  Jack Welch in You're Getting Innovation All Wrong (6/2014)

  innovation is a series of little steps that, cumulatively, lead up to a big deal that changes the game

Rule #2: Changes Rarely have a Big Positive Impact to Key Metrics

- 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 10%
  (but we have had two such experiments).
  This is especially true for well-optimized sites
- Important to highlight
  - Rule applies to key organizational metrics, not some feature metric.
    Think Sessions/user, time-to-success
  - We are looking at diluted effects. A 10% improvement to a 1% segment has an overall impact of approximately 0.1%
- Two sources of false positives that appear like breakthroughs
  - Expected from the Statistics. With p-value of 0.05, hundreds of false positives are expected when one runs 5,000 experiments per year.
  - Those that are due to a bad design, data anomalies, or bugs, such as instrumentation errors

Bayes Rule Applied to Experiments

- Standard hypothesis testing gives us the wrong conditional probabilities P(D|H) not P(H|D)
- Define
  - α is the statistical significant level = 0.05
  - β is the type-II error level = 0.2 (80% power)
  - π is the probability that the alternative hypothesis is true, i.e., the experiment is moving metrics
  - TP is True Positive, and SS is a Stat-sig result, then we have Bayes Rule:

$$P(TP|SS) = P(SS|TP) * \frac{P(TP)}{P(SS)} = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- If we have a prior probability of success of $\pi = 1/3$, which is what we reported is the average across multiple experiments at Microsoft, then the posterior probability for a true positive result given a statistically significant experiment is 89%.
- However, if the probability of success is one in 500, then the posterior probability drops to 3.1%.

Corollary: Following Tail Lights

- Following taillights is easier than innovating in isolation
- Features introduced by statistical-savvy companies that we see out there have a higher chance of having positive impact for us
  - If our success rate on ideas at Bing is about 10-20%, in line with other search engines, the success rate of features that the competition has tested and shipped is higher.
  - The converse is also true: other search engines tend to test and ship positive changes that Bing introduces.

Twyman's Law

- **Twyman**: *Any figure that looks interesting or different is usually wrong!*
- Sessions per User in most of Bing's experiments is close to zero (hard to improve). Assume it is Normal(0, $0.25\%^2$) based on thousands of experiments.
- If an experiment shows +2.0% improvement to Sessions/user, we will call out Twyman, pointing out that 2.0% is "extremely interesting" but also eight standard-deviations from the mean, and thus has a probability of 1e-15
- Twyman's law is regularly applied to proofs that $P = NP$.
  - No modern editor will celebrate a submission
  - Instead, they will send it to a reviewer to find the bug, attaching a template that says "with regards to your proof that $P = NP$, the first major error is on page x."

Examples of Twyman's Law

- Office ran an experiment that redesigned their page, which was pitching try or buy.
  They saw a decline of 56% in clicks.
  Reason? The new variant listed the price, so it sent more qualified users to the pipeline
- JavaScript added to Bing's page, expected to slow things down a bit.
  Instead of slightly worse metrics, clicks-per-user improved significantly.
  Reason? Click fidelity improved because the web beacon had more time to reach our servers
- Multiple groups, such as the Bing home page, reported great improvements to clicks per user in late 2013.
  Reason? The deployment of Bing's edge improved click fidelity.
- E-mail campaign added link to order at an e-commerce site; future conversions improved 10%.
  Reason: triggering condition counted users in Control/Treatment who clicked through
- MSN massively improved search transfers to Bing.
  Reason: auto-suggest clicks initiated two searches at Bing (one always aborted).
- Which Test Won claimed that sending e-mails at 9AM PST is better than 1PM PST for users in that time zone (July 16, 2014)
  - They claimed the lift was 4,090%. It doesn't pass the sniff test a mile away.

The Seven Rules of Thumb

- Rule #1: Small Changes can have a Big Impact to Key Metrics
- Rule #2: Changes Rarely have a Big Positive Impact to Key Metrics
- Rule #3: Your Mileage WILL Vary: most amazing stories that you see out in the wild will not replicate for you
- Rule #4: Speed Matters a LOT: At Bing, an engineer that improves server performance by 10msec (that's 1/30 of the speed that our eyes blink) more than pays for his fully-loaded annual costs
- Rule #5: Reducing Abandonment is Hard, Shifting Clicks is Easy
- Rule #6: Avoid Complex Designs: Iterate: multi-variable tests are good for one-shot offline tests. In the online world, it is better to run many simple experiments
- Rule #7: Have Enough Users: Statistic books say the Central limit theorem implies converges to a normal distribution around n≥30 users. Depends on the metric of interest. Typically need thousands

Slides with all seven rules at http://bit.ly/expRulesOfThumb

Appendix - the Rest of the Rules

Rule #3: Your Mileage WILL Vary

- Many documented examples of successes using controlled experiments
  - Anne Holland's "Which Test Won?" site (http://whichtestwon.com), has hundreds of case studies of A/B tests, and a new case is added about every week.
- These are great idea generators, but there are several problems
  - Quality varies. Some reports are not statistically significant at 5% level.
  - Domain may differ. Red button beat green button is unlikely to generalize
  - Novelty Effects. Experiments may not have run long enough
  - Misinterpretation of results

Historical Example: Lack of medical knowledge about Vitamin C

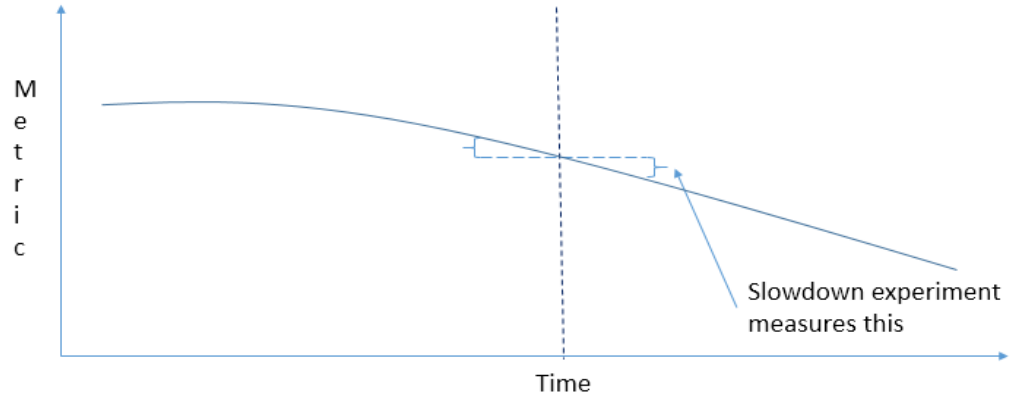
- Scurvy is a disease that results from vitamin C deficiency.
- It killed over 100,000 people in the 16th-18th centuries, mostly sailors who went out for long-distance voyages and stayed at sea longer than perishable fruits could be stored
- In 1747, Dr. James Lind noticed lack of scurvy in Mediterranean ships and gave some sailors oranges and lemons (Treatment), and others ate regular diet (Control).
- The experiment was very successful, but Dr. Lind did not understand the reason.
- At the Royal Naval Hospital in England, he treated scurvy patients with concentrated lemon juice called “rob.” He concentrated the lemon juice by heating it, thus destroying the vitamin C.
- He lost faith in the remedy and became increasingly reliant on bloodletting.
- In 1793, a formal trial was done and lemon juice became part of the daily rations throughout the navy; Scurvy was quickly eliminated and British sailors are called Limeys to this day.

Example: Speed vs. More Results

- In a Web 2.0 talk by Marissa Mayer, then at Google, she described an experiment where Google increased the number of search results from ten to thirty.
  - Traffic and revenue dropped by 20%.
  - Her explanation? The page took half a second more to generate.
- Performance is a critical factor, but this is an order of magnitude too large
- Here are three reasons:
  - We ran slowdown experiments at Bing, isolating just the performance factor.
    500msec impacts revenue about 3% not 20%, and clickthrough-rate declines 0.50%, not 20%.
  - Jake Brutlag from Google blogged about an experiment showing that slowing down the search results page by 100 to 400 milliseconds reduced searches per user by 0.2% to 0.6%
  - We ran an experiment where we showed 20 results instead of 10. We were able to nullify the revenue loss by adding another mainline ad (which slowed the page a bit more).
    We believe the ratio of ads to algorithmic results plays a more important role than performance.

Rule #4: Speed Matters a LOT

- Using a slowdown experiment, we can assess the impact of performance on a given metric
- If the linear approximation is reasonable (see graph), then the slowdown ≈ speedup
- Example metric: every 100msec speedup improves revenue by 0.6%.
  - *An engineer that improves server performance by 10msec (that's 1/30 of the speed that our eyes blink) more than pays for his fully-loaded annual costs*.
  - Every millisecond counts.
- But the above was server slowdown



Which Time?

- Slowing down Bing's right pane by 250msec did not have any significant difference on key metrics
- Amazon renders in 2.0 above the fold, but windows.onload fires at over 5 seconds. It's important to optimize what users care about: perceived performance
- Multiple metrics have been proposed
  - AFT: Above the fold time.
    Suffers from videos playing, and unimportant pixels showing late.
  - Speed index, which generalizes AFT by averaging visible elements
  - Page Phase Time and User-Ready Time
- At Bing, we believe Time-To-Success is a key metric, where success is defined as clicking on a result and not returning back in less than 30 seconds

Rule #5: Reducing Abandonment is Hard, Shifting Clicks is Easy

- A key metric we look at is abandonment rate (1 –PCR on Bing scorecards)
- Many experiments shift traffic, but rarely does abandonment rate decline
- **Example**: **Related Searches in right column.** Some related searches were removed from the right column (task pane). Clicks shifted to other areas of the page, but abandonment rate did not change statistically significantly (p-value 0.64).
- **Example**: **Related Searches below bottom ads.**
  - Related searches, which usually "float" with mainline results, were pinned to the bottom.
  - Clickthrough-rate on these related searches declined 17%
  - Abandonment rate did not change statistically significantly (p-value 0.71)
- **Example: Truncation.** We truncated SERPs with Deep-Link cards (Dcards) to 4 and abandonment rate did not change statistically significantly for these pages (p-value 0.92).

Rule #5: Reducing Abandonment is Hard, Shifting Clicks is Easy (2)

- **Example: page extension**. After users come back to the SERP using a back button, we extended the page to show 20 results and a bug caused related searches not to show up
  - Revenue declined 1.8%
  - Pages were slower by 30msec
  - Pagination was reduced by 18% (good)
  - Abandonment rate did not change significantly (p-value 0.93).
  - This change was not released
- **Example: Ad background color.** The ad background color was changed, causing a 12% decline in revenue (an annual loss of over $150M if this change were made). Users shifted their clicks from ads to other areas of the page, but abandonment rate did not change statistically significantly (p-value 0.83)

Rule #5: Reducing Abandonment is Hard, Shifting Clicks is Easy (3)

- When are we able to reduce abandonment?
  - The Anti-malware flight made a big difference
  - Significant relevance improvements reduce abandonment
- Why is this a useful rule of thumb?
  - Many projects add a module or widget, and it gets clicked
  - Teams usually claim goodness for users because they are clicking on the new area
  - But if the module is just cannibalizing other areas, as in the examples shown, then it is only useful if these clicks are “better” on some axis
- Phrased differently: local improvements are easy;
  Global improvements are much harder.

Rule #6: Avoid Complex Designs: Iterate

- Simple designs are better for the online world, where we see a lot of pitfalls. Complex designs can hide bugs
- **Example: LinkedIn unified search**
  - Single search for people, job, companies
  - Massive effort that touched many areas of the web site and impacted multiple features
  - Experiment was negative on many key metrics
  - Had to bring back one feature at a time to realize that certain features (removed from final launch), not the unified search, were responsible for bringing down clicks and revenue.
  - After restoring these features, unified search was shown to be positive to the user
- **Example: LinkedIn Contacts**
  - Created complicated eligibility requirements for being exposed to new feature
  - After long investigations, it turned out that one rule effectively mapped to: if you have been exposed once, you are no longer in the experiment…

Rule #6: Avoid Complex Designs: Iterate (2)

- With offline experiments, where experiments are expensive relative to the design and analysis, it makes sense to make maximum use of the users (experimental units).
- Online, we can run hundreds of concurrent experiments
- We usually find it best to run simple uni-variable (e.g., A/B/C/D variant of a feature) or bi-variable designs
- Also useful: align with agile software methodologies and building minimum viable products (MVPs)
  - Experiment with features right after they're built
  - Don't wait for all five features for an MVT to be done
- Deploy new code quickly and use experiments to ramp-up the treatment(s), providing a form of exposure control: start with small 1% treatments
- Careful not to repeat a deployment like the one Knight Capital did, which in Aug 2012 caused a $440 million loss and erased 75% of Knight's equity value

Rule #7: Have Enough Users

- How many users?
  - Central limit theorem says our metrics, which are typically averages, converge to a normal distribution
  - How fast? Statistical books say n≥30
  - But we are looking at tails of distributions for confidence intervals
  - In the past we said “thousands” of users are needed
- There are really two factors
  1. Minimum sample size can be computed given the metric’s variance and sensitivity (the amount of change one wants to detect), assuming it’s normally distributed
  2. When does the metric become normally distributed?

  We look at factor #2

Rule #7: Have Enough Users (2)

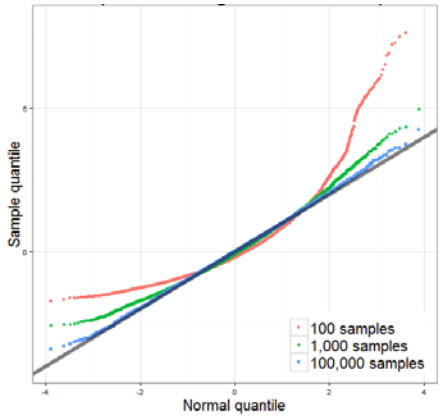
- Given the skewness of the metric, defined as $s = \frac{E[X-E(X)]^3}{[Var(X)]^{3/2}}$

  a lower bound on the number of samples is $355 \times s^2$.

- Here are some metrics and a QQ-norm plot for sample sizes and revenue/user

Metric	*skewness squared*	Min Sample Size	Sensitivity: % change detectible at 80% power
Revenue/User	322.4	114k	4.4%
Revenue/User(Truncated)	27.4	9.7k	10.5%
Sessions/User	13.2	4.70k	5.4%
TimeToSuccess	4.4	1.55k	12.3%
TimeToSuccess (Truncated)	0.15	0.05k	27.9%

Revenue per user is our most skewed metric.
Truncation helps a lot



QQ-norm plot for averages of different sample sizes showing convergence to Normal when skewness is about 18

Rule #1 bonus example: Opening Links in New Tabs

- Aug 2008, MSN UK opened Hotmail link in new tab.
  +8.9% increase in clicks/user for triggered users (those who clicked on the Hotmail link)
- June 2010, replicated experiment in the US.
  Similar results
- Apr 2011, MSN US experimented with opening search results in a new tab.
  +5% increase to clicks/user (overall).
  - One of the best features MSN ever shipped to improve user engagement
  - Trivial change = high ROI
- Philosophical debates about this for Bing.com home page.
  - Experiment showed +8% increase in searches from home page
- Multiple experiments ran at all major search engines doing this for SERP (e.g, for ads)