Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO

Ronny Kohavi
General Manager, Experimentation Platform, Microsoft

Joint work with Randy Henne and Dan Sommerfield

ronnyk@microsoft.com
http://exp-platform.com

Overview

- **Motivating Examples**
- **OEC – Overall Evaluation Criterion**
- **Controlled Experiments**
- **Limitations**
- **Lessons**
- **Q&A**

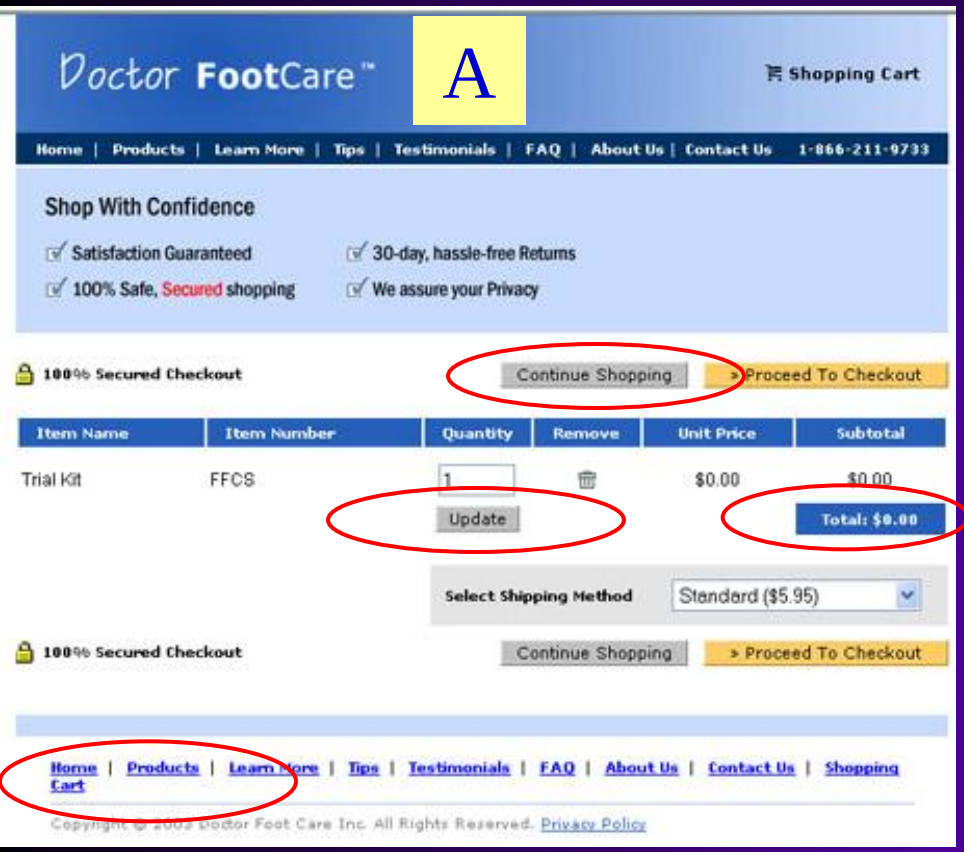
Amazon Shopping Cart Recs

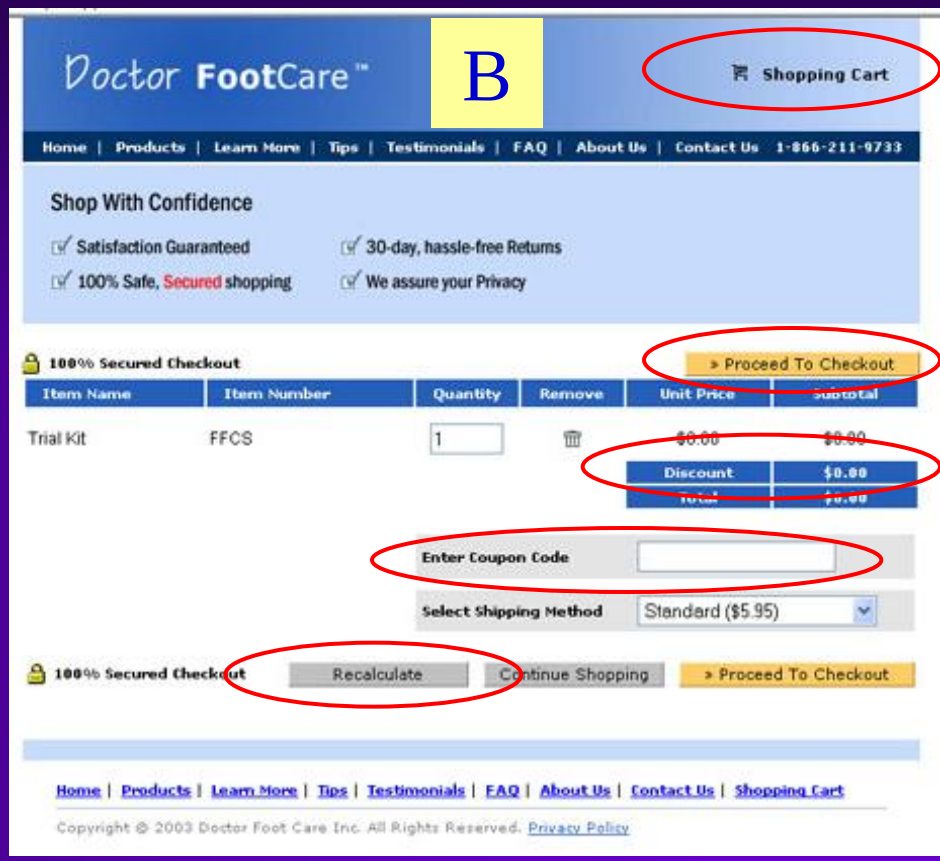
- **Add an item to your shopping cart at a website**
  - Most sites show the cart
- **At Amazon, Greg Linden had the idea of showing recommendations based on cart items**
- **Evaluation**
  - Pro: cross-sell more items (increase average basket size)
  - Con: distract people from checking out (reduce conversion)
- **HiPPO (Highest Paid Person's Opinion) was: stop the project**
- **Simple experiment was run, wildly successful**

From Greg Linden's Blog: http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html

Checkout Page

The *conversion rate* is the percentage of visits to the website that include a purchase



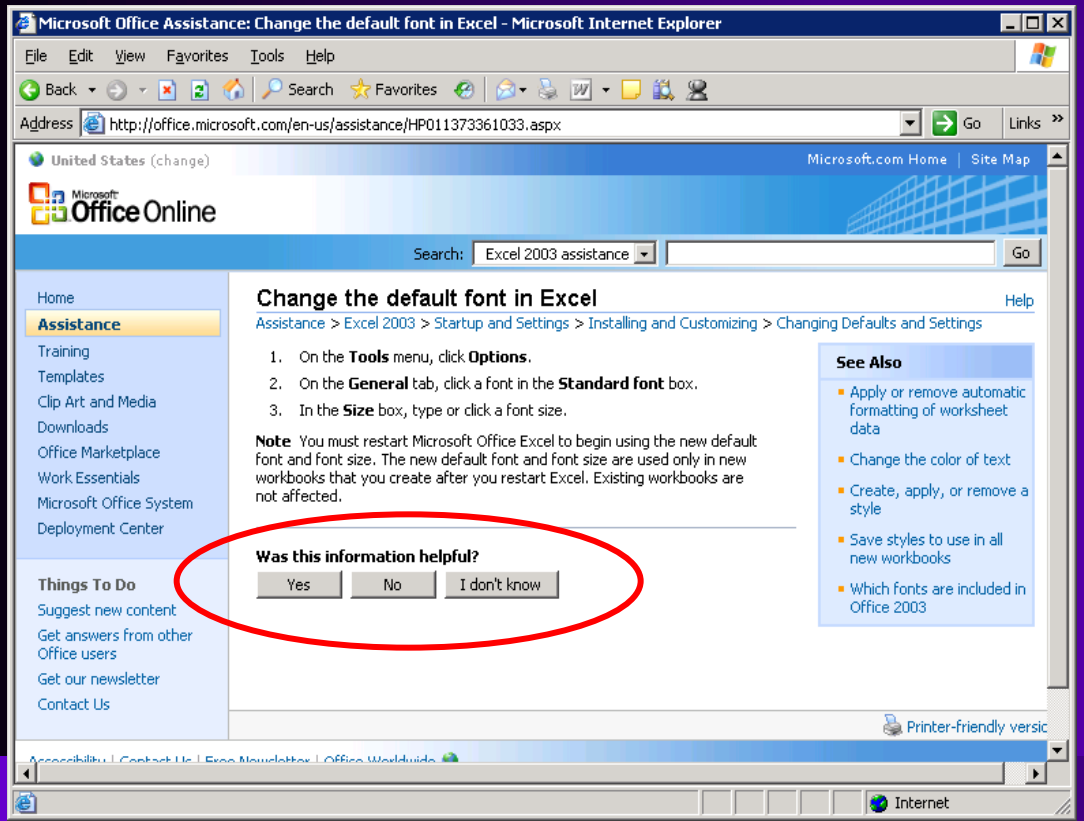


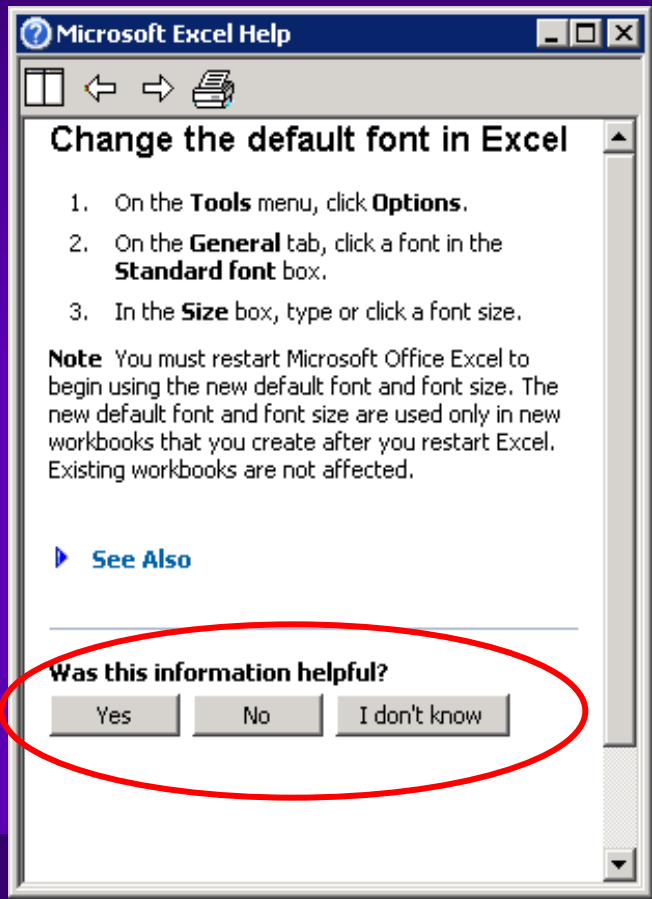
Which version has a higher conversion rate? Why?

Example from Bryan Eisenberg's article on clickz.com

Office Online

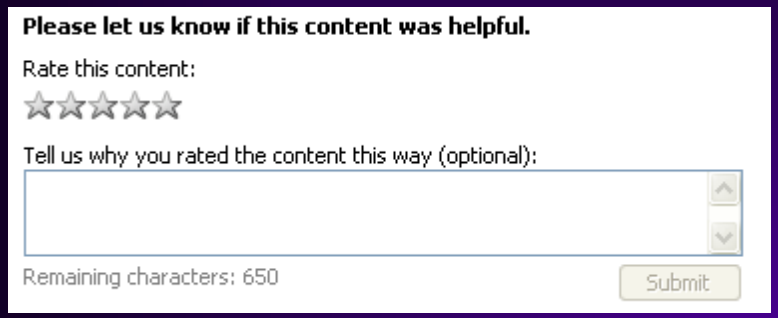
- **Small UI changes can make a big difference**
- **Example from Microsoft Help**
- **When reading help (from product or web), you have an option to give feedback**



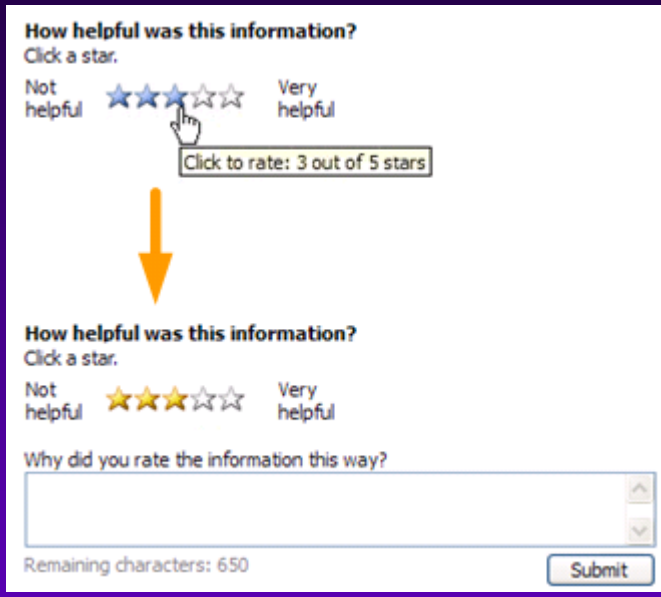


Office Online Feedback

A



B



Feedback A puts everything together, whereas feedback B is two-stage: question follows rating.

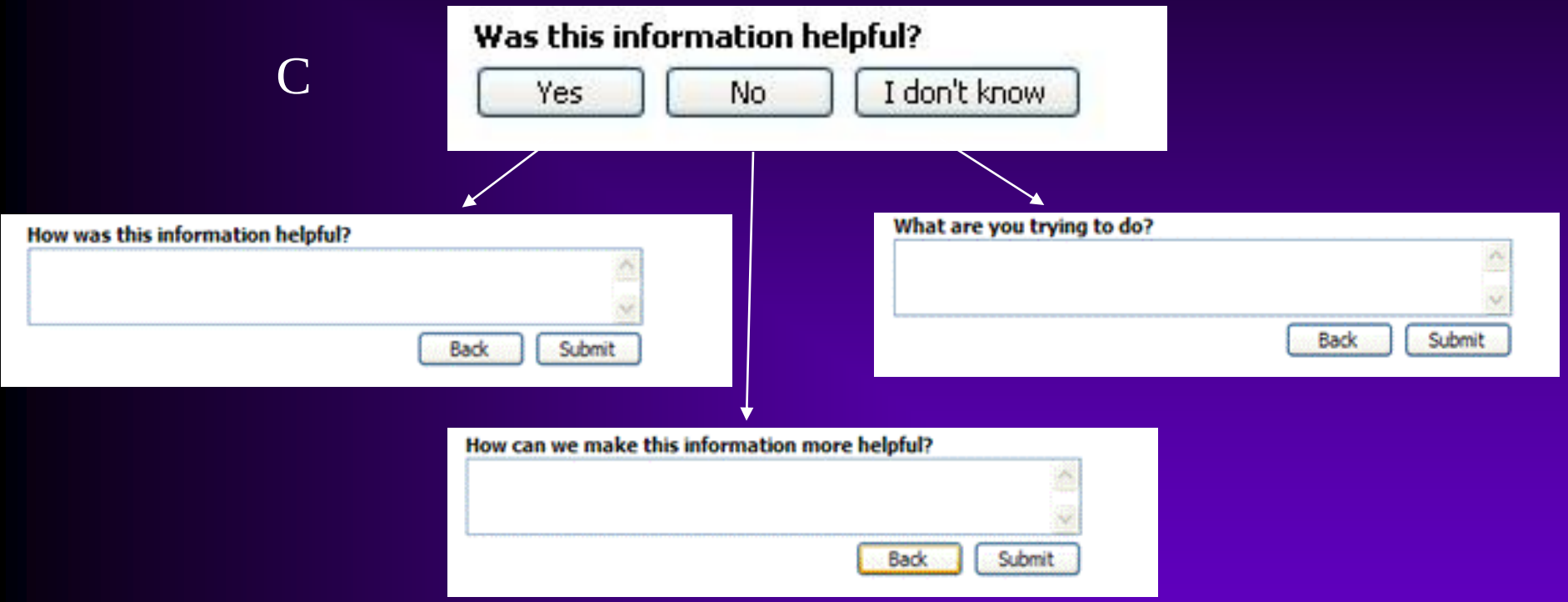
Feedback A just has 5 stars, whereas B annotates the stars with "Not helpful" to "Very helpful" and makes them lighter

Which one has a higher response rate? By how much?

B gets more than double the response rate!

Another Feedback Variant

C



Call this variant C. Like B, also two stage.

Which one has a higher response rate, B or C?

C outperforms B by a factor of 3.5 !!

JoAnn.com Sewing Machines

- **Several promotions were tried to increase sales of sewing machines**
- **The winner: "buy two, get 10% off" was initially ranked as *least* likely to be useful. After all, who needs two sewing machines.**
- **Martin Westreich, CFO, said: "We initially thought, why waste a week's worth of sales on this promotion?"**
- **But the sewing community has small clubs and many times one person (e.g., grandma) called another to buy together**

http://www.cfo.com/article.cfm/5193417/1/c_2984283

Data Trumps Intuition

Our intuition is poor, especially on novel ideas

- The less data, the stronger the opinions
- Get the data through experimentation

Define Your OEC

Optimize for the long term, not just clickthroughs

- The sewing machine ad did not win on clickthrough, but it won on sales because they sold many pairs
- Example long-term metrics
  - Time on site (per time period, say week or month)
  - Visit frequency
- Phrased differently: optimize for customer lifetime value
- We use the term OEC, or Overall Evaluation Criterion, to denote the long-term metric you really care about
- Continue to evaluate many metrics to understand the specifics and for understanding why the OEC changed

OEC Thought Experiment

- **Tiger Woods comes to you for advice on how to spend his time: improving golf, or improving ad revenue**
- **Short term, he could improve his ad revenue by focusing on ad revenue (Nike smile)**

- **But to optimize lifetime financial value (and immortality as a great golf player), he needs to focus on the game**

OEC Thought Experiment (II)

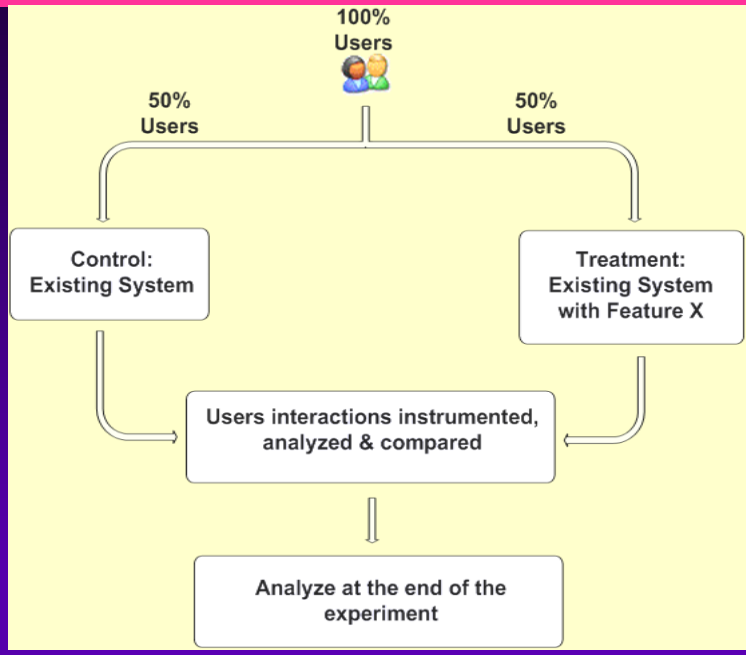
- **While the example seems obvious, organizations commonly make the mistake of focusing on the short term**
- **Groups are afraid to experiment because the new idea might be worse**
  [but it's very short term, and if the new idea is good, it's there for the long term]
- **This is the toughest cultural problems we see: getting clear alignment on the "goal."**

Lesson: Drill Down

- **The OEC determines whether to launch the new treatment**
- **If the experiment is "flat" or negative, drill down**
  - Look at many metrics
  - Slice and dice by segments (e.g., browser, country)

Controlled Experiments

- **Multiple names to the same concept**
  - Parallel flights (at MSN)
  - A/B tests or Control/Treatment
  - Randomized Experimental Design
  - Controlled experiments
  - Split testing
- **Concept is trivial**
  - Randomly split traffic between two versions
    - Control: usually current live version
    - Treatment: new idea (or multiple)
  - Collect metrics of interest, analyze (statistical tests, data mining)

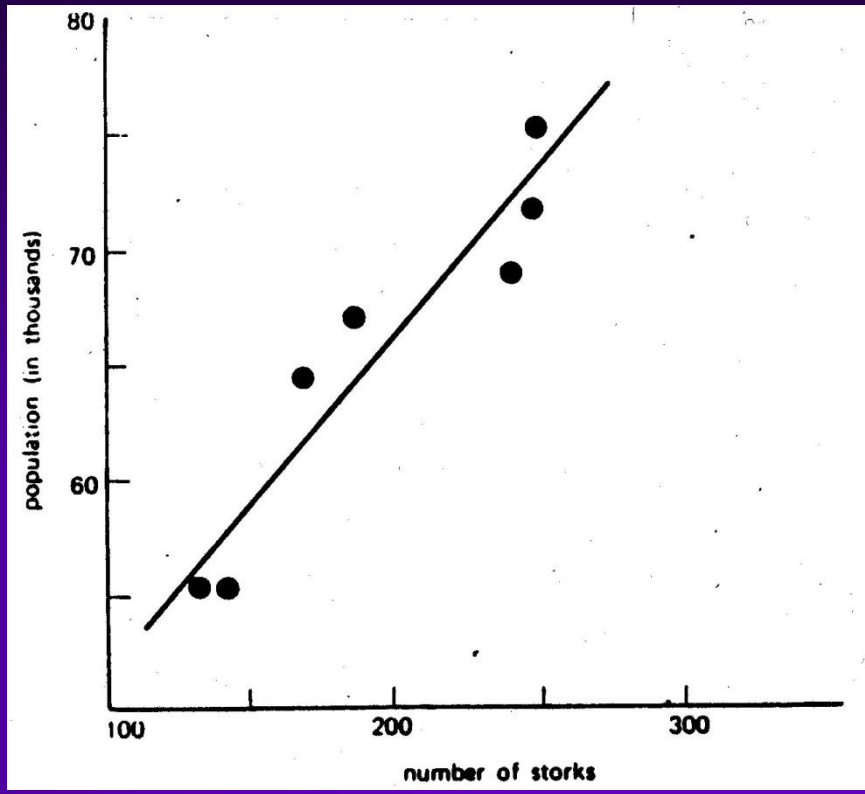


Advantages of Controlled Experiments

- **Controlled experiments test for causal relationships, not simply correlations (example next slide)**
- **They insulate external factors**
  - History/seasonality impact both A and B in the same way
- **They are the standard in FDA drug tests**
- **They have problems that must be recognized (discussed in a few slides)**

Correlations are not Necessarily Causal

- A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year, 1930-1936.
- Excellent correlation, but one should not conclude that storks bring babies



- Example 2:
  True statement (but not well known):
  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.

Why?
Women have smaller palms and live 6 years longer on average

Issues with Controlled Experiments (1 of 2)

If you don't know where you are going, any road will take you there
—Lewis Carroll

- **Org has to agree on OEC (Overall Evaluation Criterion).**
  This is hard, but it provides a clear direction and alignment
- **Quantitative metrics, not always explanations of "why"**
  - A treatment may lose because page-load time is slower.
    Example: Google surveys indicated users want more results per page. They increased it to 30 and traffic dropped by 20%.
    Reason: page generation time went up from 0.4 to 0.9 seconds
  - A treatment may have JavaScript that fails on certain browsers, causing users to abandon

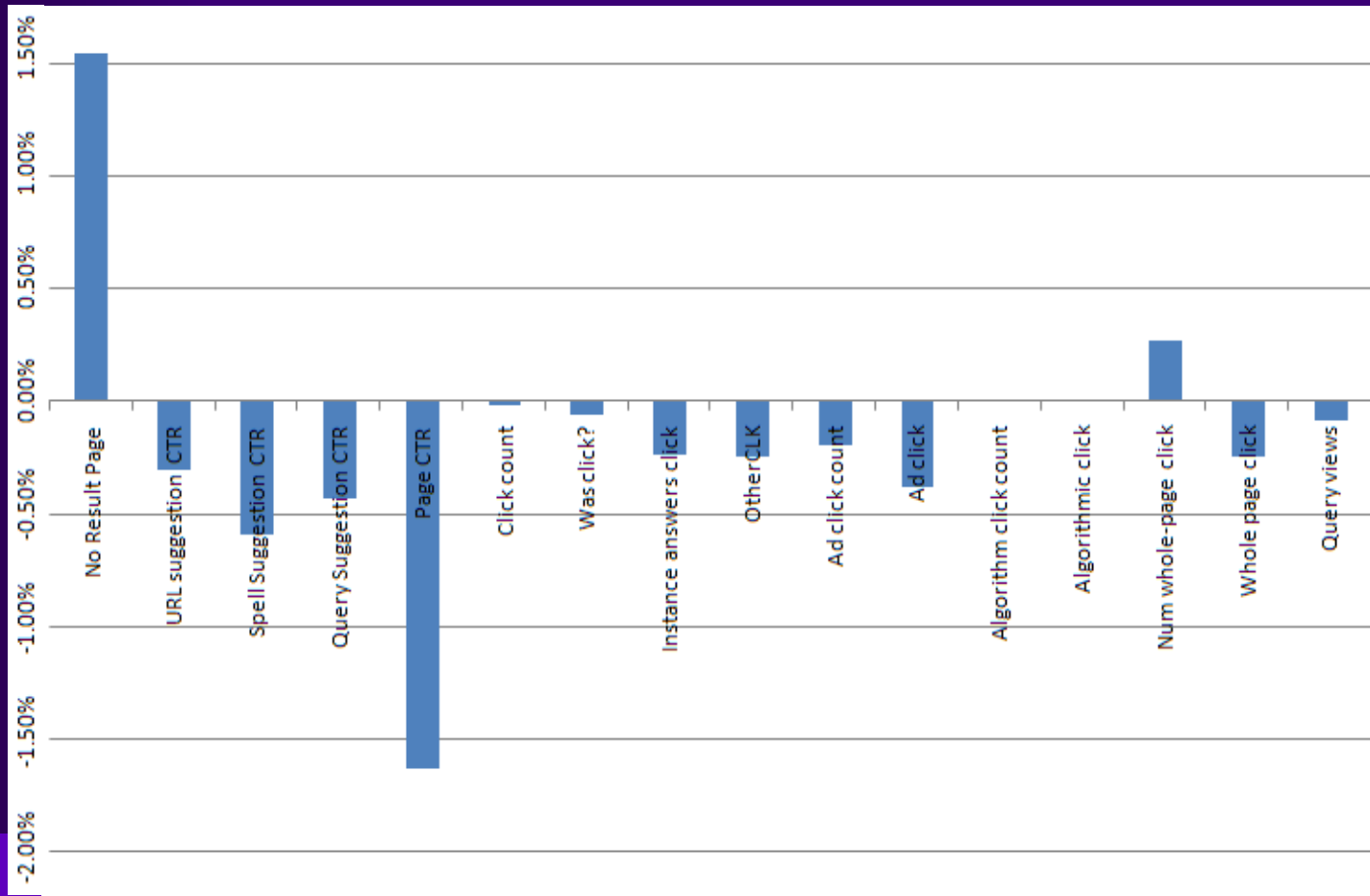
Issues with Controlled Experiments (2 of 2)

- **Primacy effect**
  - Changing navigation in a website may degrade the customer experience (temporarily), even if the new navigation is better
  - Evaluation may need to focus on new users, or run for a long period
- **Multiple experiments**
  - Even though the methodology shields an experiment from other changes, statistical variance increases making it harder to get significant results. There can also be strong interactions (rarer than most people think)
- **Consistency/contamination**
  - On the web, assignment is usually cookie-based, but people may use multiple computers, erase cookies, etc. Typically a small issue
- **Launch events / media announcements sometimes preclude controlled experiments**
  - The journalists need to be shown the "new" version

Typical Experiment

- Here is an A/B test measuring 16 metrics in search
- It has one problem. Guesses?

Over 1M users in each variant



Lesson: Compute Statistical Significance, Run A/A Tests, and Compute Power

- **A=B, i.e., no difference in treatment.**
  This was an A/A test
- **A very common mistake is to make conclusions based on random variations**
- **Compute 95% confidence intervals on the metrics to determine if the difference is due to chance or whether it is statistically significant**
- **Continuously run A/A tests in parallel with other A/B tests**
- **Do power calculations to determine how long you need to run an experiment (minimum sample size)**

Run Experiments at 50/50%

- **Novice experimenters run 1% experiments**
- **To detect an effect, you need to expose a certain number of users to the treatment (based on power calculations)**
- **Fastest way to achieve that exposure is to run equal-probability variants (e.g., 50/50% for A/B)**
- **But don’t start an experiment at 50/50% from the beginning: that’s too much risk. Ramp-up over a short period**

Ramp-up and Auto-Abort

- **Ramp-up**
  - Start an experiment at 0.1%
  - Do some simple analyses to make sure no egregious problems can be detected
  - Ramp-up to a larger percentage, and repeat until 50%
- **Big differences are easy to detect because the min sample size is quadratic in the effect we want to detect**
  - Detecting 10% difference requires a small sample and serious problems can be detected during ramp-up
  - Detecting 0.1% is extremely hard, so you might want 50% for two weeks
- **Automatically abort the experiment if treatment is significantly worse on OEC or other key metrics (e.g., time to generate page)**

Randomization

- **Good randomization is critical.**
  It's unbelievable what mistakes devs will make in favor of efficiency
- **Properties of user assignment**
  - Consistent assignment. User should see the same variant on successive visits
  - Independent assignment. Assignment to one experiment should have no effect on assignment to others (e.g., Eric Peterson's code in his book gets this wrong)
  - Monotonic ramp-up. As experiments are ramped-up to larger percentages, users who were exposed to treatments must stay in those treatments (population from control shifts)

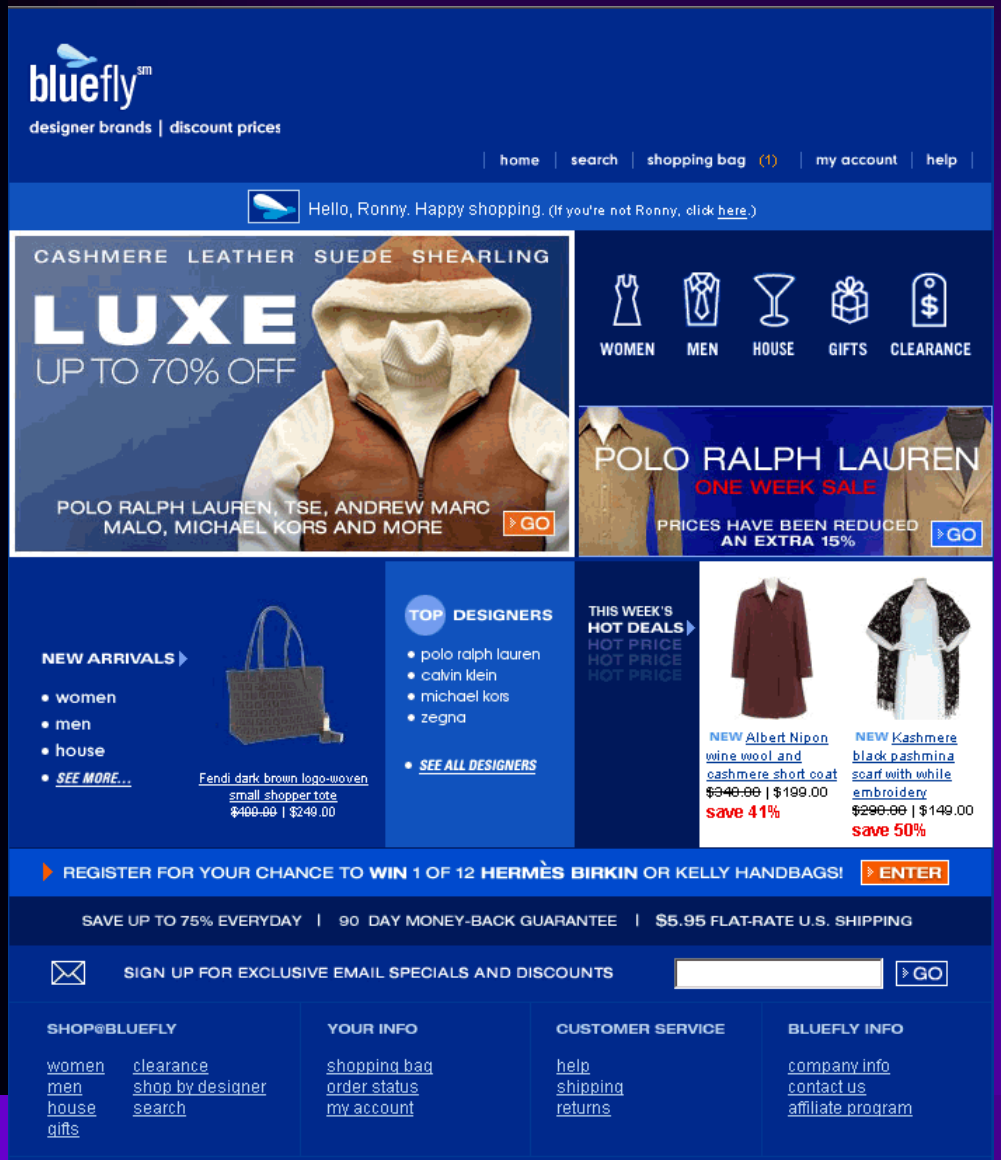
A Real Technical Lesson: Computing Confidence Intervals

- **In many situations we need to compute confidence intervals, which are simply estimated as:** acc_h +- z*stdDev
  - where acc_h is the estimated mean (e.g., clickthrough or accuracy),
  - stdDev is the estimated standard deviation, and
  - z is usually 1.96 for a 95% confidence interval)
- **This fails miserably for small amounts of data**
  - For Example: If you see three coin tosses that are head, the confidence interval for the probability of head would be [1,1]
- **Use a more accurate formula**

$$\frac{2h \cdot \mathtt{acc}_h + z^2 \pm z \cdot \sqrt{4h \cdot \mathtt{acc}_h + z^2 - 4h \cdot \mathtt{acc}_h^2}}{2(h + z^2)}$$

  - It's not used often because it's more complex, but that's what computers are for
  - See Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection" in IJCAI-95

Collect Many Metrics (e.g., Form Errors)



Here is a good example of data collection that we introduced at Blue Martini without knowing apriori whether it will help: form errors

If a web form was filled and a field did not pass validation, we logged the field and value filled

This was the Bluefly home page when they went live

Looking at form errors, we saw thousands of errors every day on this page

Any guesses?

Cleansing

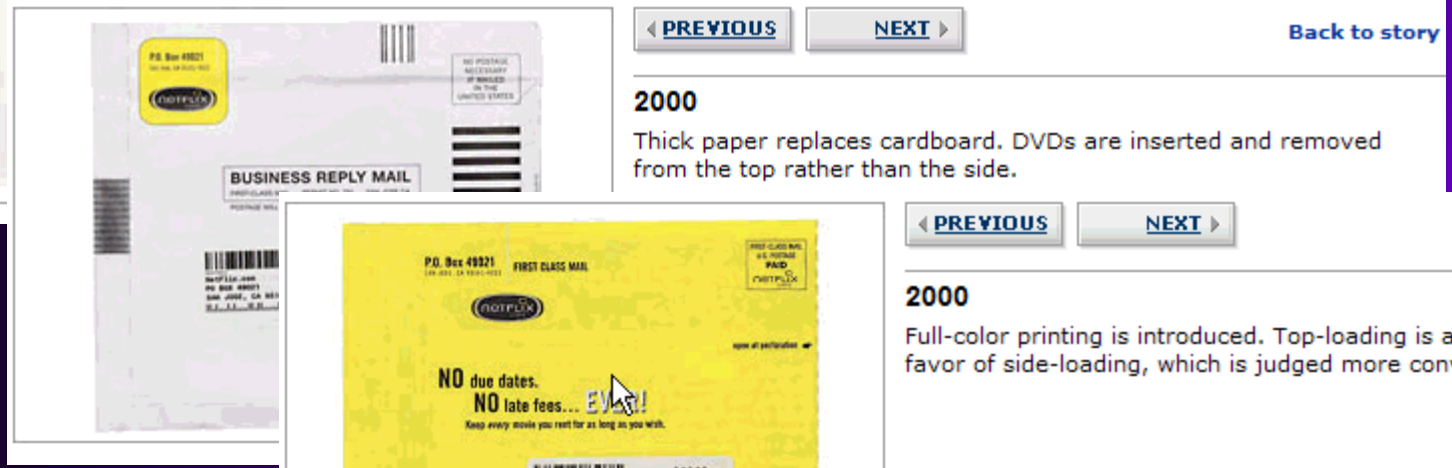
- **Remove test data**
  - QA organizations may be testing live features
  - Performance systems may be generating traffic that adds noise
- **Remove robots/bots/spiders**
  - 5-40% of site e-commerce site traffic is generated by crawlers from search engines and students learning Perl.
    These can significantly skew results or reduce power
- **Do outlier detection and sensitivity analysis**

Cultural Lessons

- **Beware of launching experiments that "do not hurt."**
  - It is possible that the experiments was negative but underpowered
  - To test for "equality" on migrations, make sure to avoid false negatives (type II errors)
- **Weight feature maintenance cost**
  - Statistical significance does not imply new feature is justified against its maintenance costs
- **Drive to a Data-Driven Culture**
  - Test often, run multiple experiments all the time

TIMITI – Try It, Measure It, Tweak It(*)

- **Netflix's envelopes are a great example of a company tweaking things**



(*) TIMITI acronym by Jim Sterne

TIMITI – Try It, Measure It, Tweak It (II)

TIMITI – Try It, Measure It, Tweak It (III)

PREVIOUS NEXT Back to story

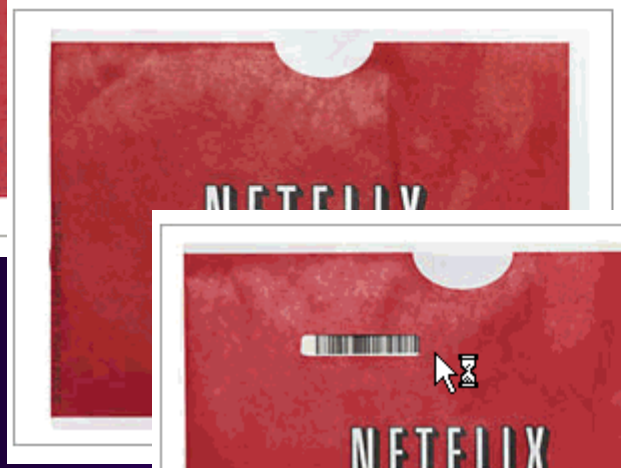
2001

Foam padding is dropped because the benefits don't justify the cost. The company gives top-loading another try.

PREVIOUS NEXT Back to story

2001

Marking a return to side-loading, this mailer is a direct ancestor of the one the company uses today.



PREVIOUS NEXT Back to story

2003

Instead of sealing the entire top and bottom, Netflix introduces a circular sticker, affixed only on the top.

PREVIOUS NEXT Back to story

2004

A window shows the disc bar code. Speculation is that this enables storing discs in mailers prior to shipping.

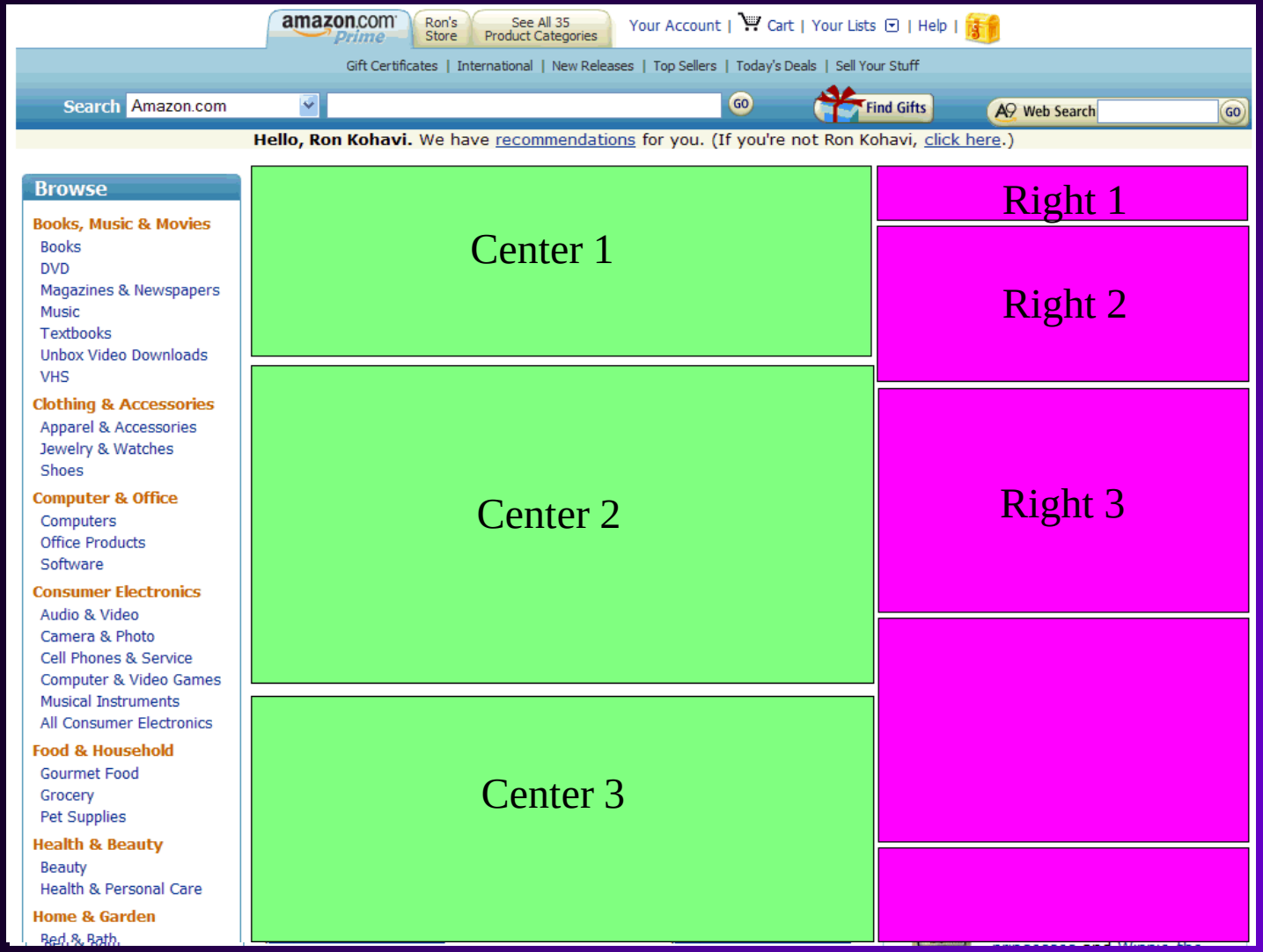
Details in Business 2.0 Apr 21, 2006.
The evolution of the NetFlix envelope

Extensions

- **Integrate controlled experiments into systems so experiments don't require coding.**
  For example, content management systems
- **Near-real-time optimizations**

- **Example of the above two: Amazon**

Amazon Home Page Slots



Amazon Home Page(*)

- **Amazon's home page is prime real-estate**
- **The past: arguments devoid of data**
  - Every category VP wanted top-center
  - Friday meetings about placements for next week were long and loud
  - Decisions based on guesses and clout, not data
- **Now: automation based on real-time A/B tests**
  - Home page is made up of slots
  - Anyone (really anyone) can submit content for any slot
  - Real-time experimentation chooses best content using the OEC
  - People quickly saw the value of their ideas
    - relative to others, and
    - encouraged to try variants to "beat" themselves and others!!

(*) From emetrics 2004 talk by Kohavi and Round (http://www.emetrics.org/summit604/index.html)

Beware of Twyman's Law

Any statistic that appears interesting is almost certainly a mistake

- **Validate "amazing" discoveries in different ways. They are usually the result of a business process**
  - 5% of customers were born on the exact same day (including year)
    - 11/11/11 is the easiest way to satisfy the mandatory birth date field
  - For US and European Web sites, there will be a small sales increase on Nov 4th, 2007
    - Hint: increase in sales between 1-2AM
    - Due to Daylight Saving Time ending, clocks at 2AM are moved back to 1AM, so there is an extra hour in the day

Summary

1. **Listen to customers because our intuition at assessing new ideas is poor**
2. **Replace HiPPOs with an OEC**
3. **Compute the statistics carefully**
4. **Experiment Often**
   Triple your experiment rate and you triple your success (and failure) rate. Fail fast & often in order to succeed
5. **Create a trustworthy system to accelerate innovation**

Experimentation Platform
http://exp-platform.com

Accelerating software innovation through trustworthy experimentation