

Emetrics 2007, Washington DC

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

Ronny Kohavi, GM of Experimentation Platform, Microsoft

Based on KDD 2007 paper and IEEE Computer paper with members of ExP team.
Papers available at <http://exp-platform.com>



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**



Overview

- **Controlled Experiments in one slide**
- **Lots of motivating examples**
 - All real and statistically significant
 - Some (but not all) ran with our Experimentation Platform
- **OEC – Overall Evaluation Criterion**
 - It's about the culture, not the technology
- **Controlled Experiments: deeper dive**
 - Advantages & Limitations
 - Lessons
- **Microsoft's Experimentation Platform**

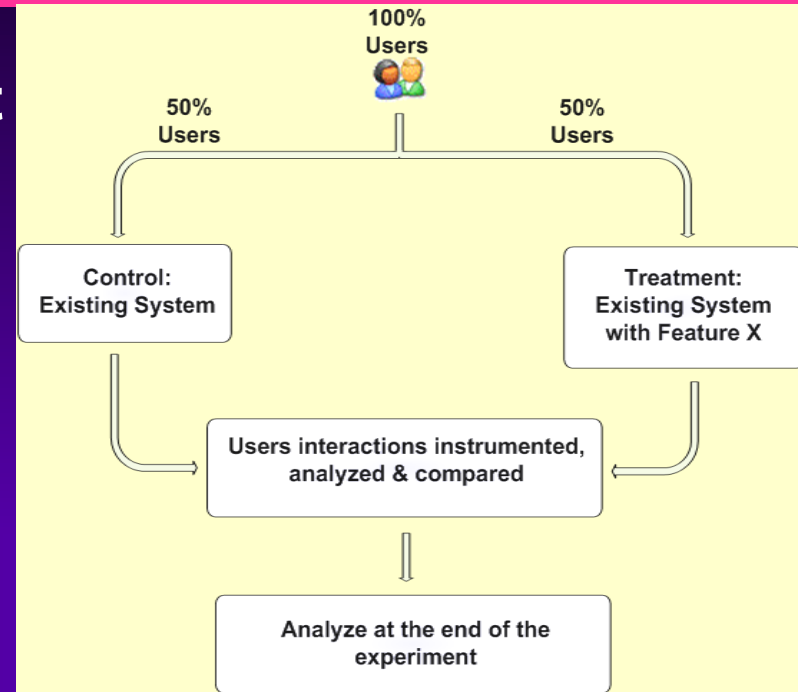
Controlled Experiments

- **Multiple names to same concept**

- A/B tests or Control/Treatment
- Randomized Experimental Design
- Controlled experiments
- Split testing
- Parallel flights

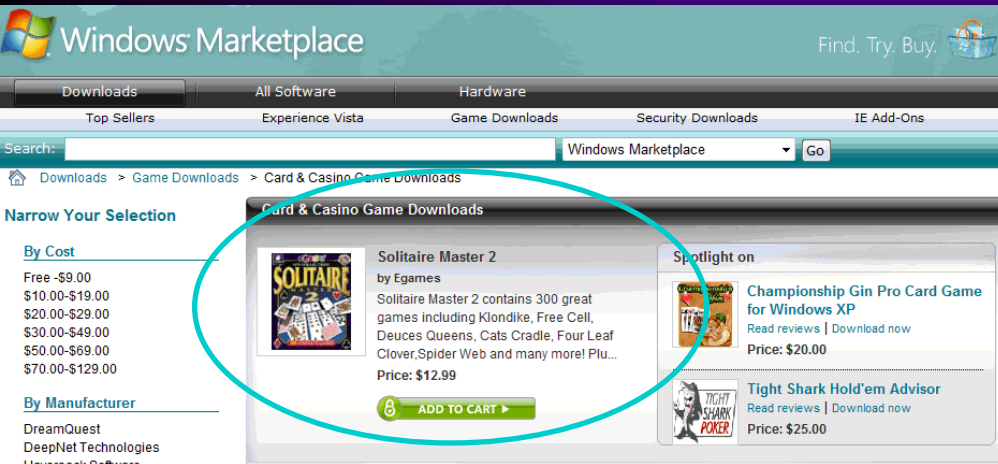
- **Concept is trivial**

- Randomly split traffic between two versions
 - A/Control: usually current live version
 - B/Treatment: new idea (or multiple)
- Collect metrics of interest, analyze (statistical tests, data mining)



Marketplace: Solitaire v Poker

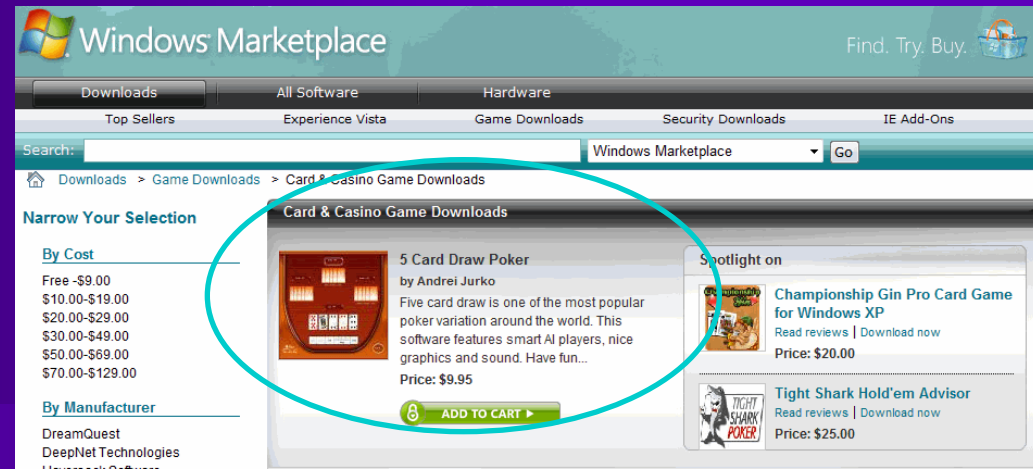
This experiment ran in Windows Marketplace / Game Downloads
Which image has the higher clickthrough? By how much?



A: Solitaire game
in hero position

B: Poker game
in hero position

A is 61% better



Office Online Feedback

A

Please let us know if this content was helpful.

Rate this content:

☆☆☆☆☆

Tell us why you rated the content this way (optional):

Remaining characters: 650

B

How helpful was this information?

Click a star.

Not helpful ☆☆☆☆☆ Very helpful

Click to rate: 3 out of 5 stars

↓

How helpful was this information?

Click a star.

Not helpful ☆☆☆☆☆ Very helpful

Why did you rate the information this way?

Remaining characters: 650

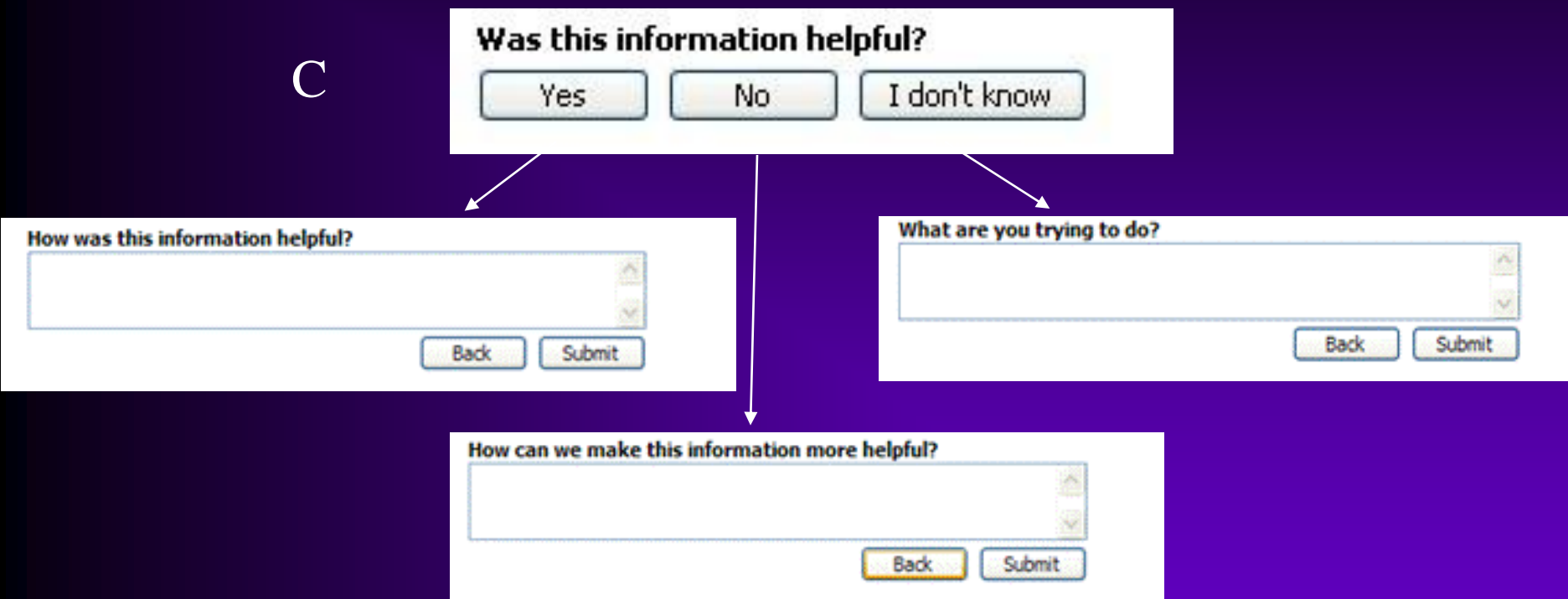
Feedback A puts everything together, whereas feedback B is two-stage: question follows rating.

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



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 !!

MSN US Home Page

Proposal: New Offers module below Shopping

Shopping

- Lancôme: Free deluxe compact w/ purchase
- Special promotions at your favorite stores
- Warm fall fashion styles are here
- Save on top brand digital cameras
- Free shipping on furniture for every room

Advertisements:

 **A smart way to buy a diamond**


- Wal-Mart: Back-to-school
- Our editor picks budget electronics
- Get fit & save money: Sports sale

Control

Shopping


- Lancôme: Free deluxe compact w/ purchase
- Special promotions at your favorite stores
- Warm fall fashion styles are here
- Save on top brand digital cameras
- Free shipping on furniture for every room


Advertisements:


 **A smart way to buy a diamond**

- Wal-Mart: Back-to-school
- Our editor picks budget electronics
- Get fit & save money: Sports sale

Offers

 **Search GM Certified**
With our 117-Point Inspection GM Certified means no worries

 **Online University**
Earn degree from a top school 100% Online. Get Free Info!

 **\$200k Loan, Get Low Rates**
Secure Financing and Increase Cash Flow. Click Here Now!

Treatment

MSN US Home Page Experiment

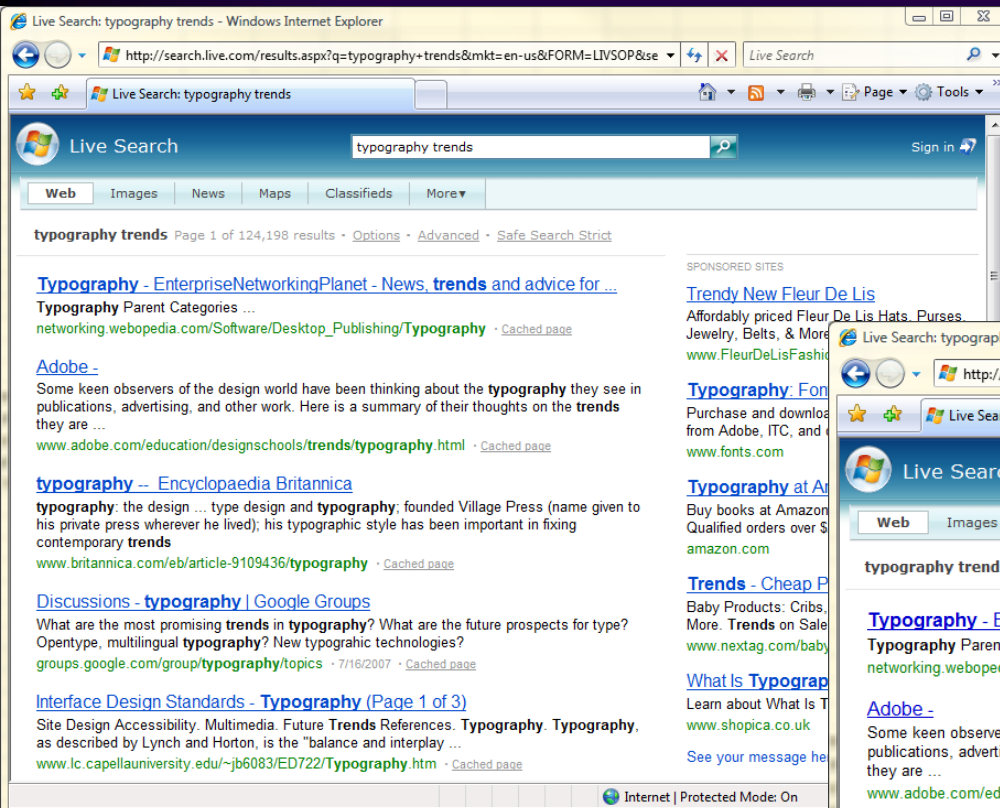
- **Offers module eval**
 - Pro: significant ad revenue
 - Con: do more ads degrade the user experience?
 - How do we trade the two off?
- **Last month, we ran an A/B test for 12 days on 5% of the MSN US home page visitors**

Experiment Results

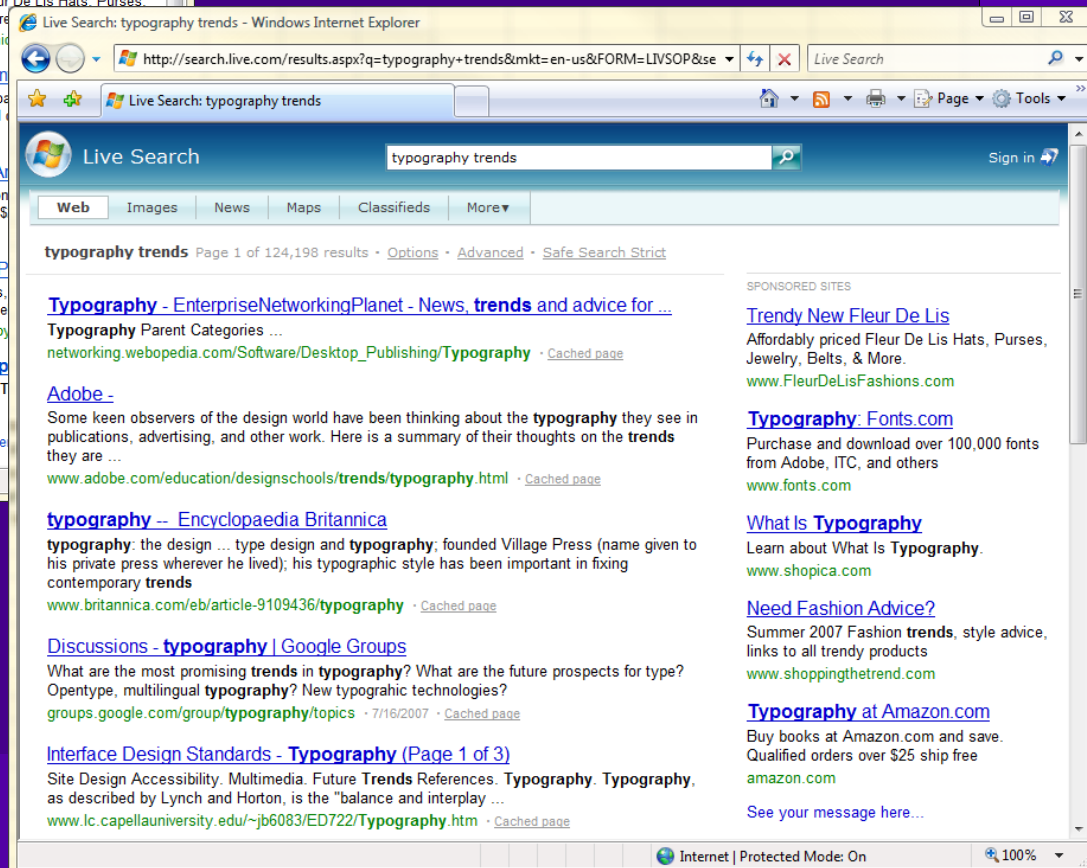
- Clickthrough rate (CTR) decreased 0.49% (p-value <0.0001)
Small change, but highly significant
- Page views per user-day decreased 0.35% (p-value<0.0001)
- Value of click from home page: X cents
Agreeing on this value is the hardest problem
 - Method 1: estimated value of “session” at destination
 - Method 2: what would the SEM cost be to generate “lost” traffic
- Net = Expected Revenue –
direct lost clicks –
lost clicks due to decreased page views

Net was negative, so the offers module did not launch

Typography Experiment Color Contrast on MSN Live Search



← A: Softer colors



B: High contrast →



B: Queries/User up 0.9%
Ad clicks/user up 3.1%

Performance Impact on Search

- **Performance matters a lot**
- **Experiment slowed search results page by 1 second**
 - Queries/User declined 1.0%
 - Ad Clicks/User declined 1.5%
- **Slowed page by 2 seconds**
 - Queries/User declined 2.5%
 - Ad Clicks/User declined 4.4%

The OEC

- **If you remember one thing from this talk, remember this point**
- **OEC = Overall Evaluation Criterion**
 - Agree early on what you are optimizing
 - Experiments with clear objectives are the most useful
 - Suggestion: optimize for customer lifetime value, not immediate short-term revenue
 - Criterion could be weighted sum of factors, such as
 - Time on site (per time period, say week or month)
 - Visit frequency
 - Report many other metrics for diagnostics, i.e., to understand the why the OEC changed and raise new hypotheses

OEC Thought Experiment

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



- 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**
- **Example:**
 - Sites show too many irrelevant ads
 - Groups are afraid to experiment because the new idea might be worse
[but it's a very short term experiment, and if the new idea is good, it's there for the long term]

The Cultural Challenge

It is difficult to get a man to understand something when his salary depends upon his not understanding it.

-- Upton Sinclair

- **Getting orgs to adopt controlled experiments as a key developmental methodology, is hard**
 - Some believe it threatens their job as decision makers
 - At Microsoft, program managers select the next set of features to develop. Proposing several alternatives and admitting you don't know which is best is hard
 - Editors and designers get paid to select a great design
 - Failures of ideas may hurt image and professional standing. It's easier to declare success when the feature launches

Experimentation: the Value

- **Data Trumps Intuition**

- It is humbling to see how often we are wrong at predicting the magnitude of improvement in experiments (most are flat, meaning no statistically significant improvement)
- Every new feature is built because *someone* thinks it is a great idea worth implementing (and convinces others)

- **Encourage Experimentation**

- Learn from flat/negative results. Even if an idea failed to improve the OEC, the org **learned** something
- Deploy the positive experiments: only **their** sum really matters
- To innovate, experiment often. As Thomas Edison said:
To have a great idea, have a lot of them

Stress HiPPO

The less data, the stronger the opinions

- To help the cultural shift, we created the stress-HiPPO
- Whenever you feel stressed that a decision is made without data, squeeze the Stress-HiPPO



- You can pick one up after the talk

Overview

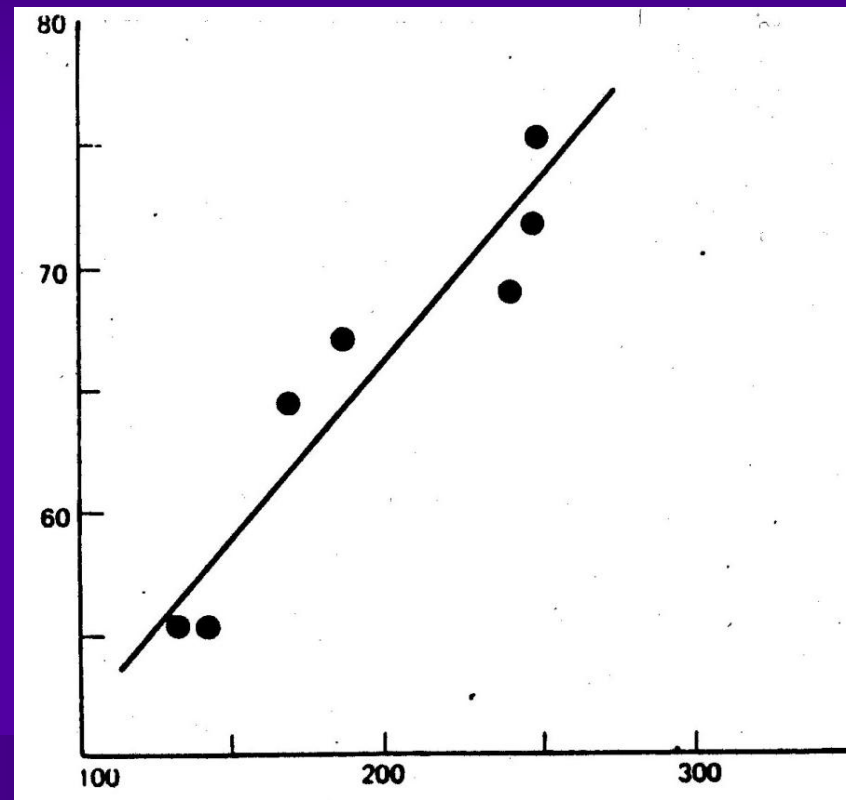
- **Controlled Experiments in one slide**
- **Lots of motivating examples**
 - All real and statistically significant
 - Some (but not all) ran with our Experimentation Platform
- **OEC – Overall Evaluation Criterion**
 - It's about the culture, not the technology
- **Controlled Experiments: deeper dive**
 - Advantages & Limitations
 - Lessons
- **Microsoft's Experimentation Platform**

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**

Typical Discovery

- With data mining, we find patterns, but most are correlational
- Here is one a real example of two highly correlated variables



Correlations are not Necessarily Causal

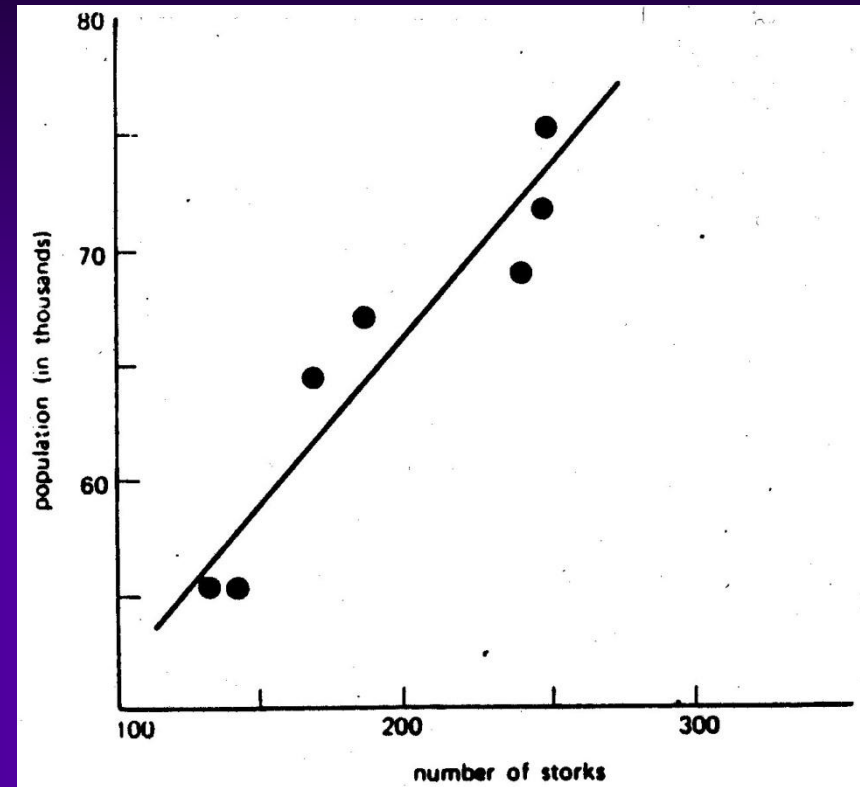
- City of Oldenburg, Germany
- X-axis: stork population
- Y-axis: human population

What your mother told you about babies when you were three is still not right, despite the strong correlational “evidence”

- 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

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)

Lesson: Compute Statistical Significance and run A/A Tests

- **A very common mistake is to declare a winner when the difference could be due to random variations**
- **Always run A/A tests**
(similar to an A/B test, but besides splitting the population, there is no difference)
- **Compute 95% confidence intervals on the metrics to determine if the difference is due to chance or whether it is statistically significant**
- **Increase percentage if you do multiple tests**
(e.g., use 99%)
- **Idea: run an A/A test in concurrent to your A/B test to make sure the overall system doesn't declare it as significant more than 5% of the time (great QA)**

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% requires a population $100^2 = 10,000$ times bigger
- **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)

Controversial Lessons

- **Run concurrent univariate experiments**
 - Vendors make you think that MVTs and Fractional Factorial designs are critical---they are not. The same claim can be made that polynomial models are better than linear models: true in theory, less useful in practice
 - Let teams launch multiple experiments when they are ready, and do the **analysis** to detect and model interactions when relevant (less often than you think)
- **Backend integration (server-side) is a better long-term approach to integrate experimentation than Javascript**
 - Javascript suffers from performance delays, especially when running multiple experiments
 - Javascript is easy to kickoff, but harder to integrate with dynamic systems
 - Hard to experiment with backend algorithms (e.g., recommendations)

Microsoft's Experimentation Platform

Mission: accelerate software innovation through trustworthy experimentation

- **Build the platform**
- **Change the culture towards more data-driven decisions**
- **Have impact across multiple teams at Microsoft , and**
- **Long term: Make platform available externally**

Design Goals

- **Tight integration with other systems (e.g., content management) allowing “codeless experiments”**
- **Accurate results in near real-time**
 - Trust is important
 - Quickly detect and abort poorly performing experiments
 - High-performance data pipeline with built-in data loss detection
- **Minimal risk for experimenting applications**
 - Encourage bold innovations with reduced QA cycles
 - Auto-abort catches bugs in experimental code
 - Client library insulates app from platform bugs
- **Experimentation should be easy**
 - Client library exposes simple interface
 - Web UI enables self-service
 - Service layer enables platform integration

Summary

- 1. Listen to customers because our intuition at assessing new ideas is poor**
- 2. Replace the HiPPO 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 by lowering the cost of running experiments**

<http://exp-platform.com>



**Accelerating software Innovation through
trustworthy experimentation**