Assessing Ideas is Hard

- Doctors take the Hippocratic Oath associated with “Do no harm,” yet David Wootton writes
  For 2,400 years patients have believed that doctors were doing them good; for 2,300 years they were wrong

- For centuries, an illness was thought to be a toxin
  Opening a vein and letting the sickness run out was the best solution – bloodletting

- A British medical text recommended bloodletting for:
  acne, asthma, cancer, cholera, coma, convulsions, diabetes, epilepsy, gangrene, gout, herpes, indigestion, insanity, jaundice, leprosy, ophthalmia, plague, pneumonia, scurvy, smallpox, stroke, tetanus, tuberculosis, and for some one hundred other diseases
  Physicians often reported the simultaneous use of fifty or more leeches on a given patient
  Through the 1830s the French imported about forty million leeches a year for medical purposes
Assessing Ideas is Hard (2)

President George Washington had a sore throat
- Doctors extracted 82 ounces of blood over 10 hours (35% of his total blood), causing anemia and hypotension.
- He died that night

Pierre Louis did an experiment in 1836
- One of the first randomized controlled experiments (clinical trials).
  He treated people with pneumonia either with
  - early, aggressive bloodletting, or
  - less aggressive measures
- At the end of the experiment, Dr. Louis counted the bodies; they were stacked higher over by the bloodletting sink

Most software changes are believed to be positive to the user experience, but are often flat or negative!

Once you objectively evaluate changes, you’re often humbled
Agenda

- Controlled experiments and observational studies
- Examples: you’re the decision maker
- Running experiments at scale and best practices
- The cultural challenge

Two key messages to remember
- It is hard to assess the value of ideas.
  Get the data by experimenting because data trumps intuition
- Make sure the org agrees what you are optimizing

Controlled Experiments in One Slide

Concept is trivial
- Randomly split traffic between two (or more) versions
  - A (Control)
  - B (Treatment)
- Collect metrics of interest
- Analyze

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)
Typical Discovery

- With data mining, we find patterns, but most are correlative, providing hypotheses for possible causes.
- Here is one a real example of two highly correlated variables.

Correlations are not Necessarily Causal

- Real Data for the city of Oldenburg, Germany:
  - X-axis: stork population
  - Y-axis: human population
- What your mother told you about babies and storks when you were three is not correct, despite the strong correlational "evidence."
- Killing the storks won’t solve population growth problems.

Ornithologische Monatsberichte 1936;44(2)
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 a 30” monitor because you bought a DisplayPort cable
- Buy Atonement movie DVD because you bought a Maglite flashlight
- Buy Organic Virgin Olive Oil because you bought Toilet Paper

What about Observational Studies?

Why not compare with historical data?
Here’s an example of Kindle Sales over time. You changed the site, and there was an amazing spike
External Events can Dwarf Your Changes

In this example of an A/B test, you’d be better off with version A.
In controlled experiments, both versions are impacted the same way by external events.

Example: Correlations due to Common Cause

Article in Nature: leaving the night-light on in the nursery made the child myopic later in life.
CNN publishes summary of article.
A year later, an important observation was made:
- Myopic parents are more likely to leave the light on, and
- Myopic parents are more likely to have myopic children
A common factor was identified.
Once you control for the parents’ myopia, the relationship in the original study disappears.
Even the “Best” Observational Studies are Wrong

“[Ioannidis] evaluated the reliability of forty-nine influential studies (each cited more than 1,000 times) published in major journals …

- 90 percent of large randomized experiments produced results that stood up to replication, as compared to only
- 20 percent of nonrandomized studies.”

--- Jim Manzi, Uncontrolled

- We run t-tests at 95% confidence, so 90% replication is reasonable for randomized controlled experiments
- It’s the 20% for uncontrolled experiments that’s shocking, and these are the “best of the best” studies

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 (“Student’s t-test”)
- 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
Examples

- Three experiments that ran at Microsoft
- Each helps share interesting lessons
- All had enough users for statistical validity
- 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, < 2% delta)
    - B wins

MSN Home Page Search Box

OEC: Clickthrough rate for Search box and popular searches

A

Popular Searches: Fireworks safety | Rihanna | Campaign patriotism

B

Differences: A has taller search box (overall size is the same), has magnifying glass icon, “popular searches”
B has big search button

• Raise your left hand if you think A Wins
• Raise your right hand if you think B Wins
• Don’t raise your hand if they are the about the same
Search Box

- Insights
  - Stop debating, it's easier to get the data
  - Most people are overly confident that their idea will work. How confident were you?
  - Reality: most ideas fail to deliver (statistics in later slides)
  - To get insights try OFAT: One Factor At a Time. Don’t tweak too many things at once.
    (But be careful not to fall into Incrementalism)

MSN US Home Page: Search Box

- A later test showed that changing the magnifying glass to an actionable word (search, go, explore) was highly beneficial.
- This:
  - is better than

In line with Steve Krug’s great book: Don’t Make Me Think
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

A
- Pro: 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)

B

• Raise your Left hand if you think A Wins
• Raise your Right hand if you think B Wins
• Don’t raise your hand if you think they’re about the same

Bing Ads

<answer>

The above change was costly to implement.
We made two small changes to Bing, which took days to develop, each increased ad revenue by about $100 million annually.

(One was delayed by 6 months because it was not prioritized high, a prioritization mistake that cost $50M)
Office Online

OEC: Clicks on revenue generating links (red below)

A

B

• Raise your left hand if you think A Wins
• Raise your right hand if you think B Wins
• Don’t raise your hand if they are the about the same

What % of the audience is still standing? Humbling!

Remember: random guesses = $1/3^3 = 1/27$
Twyman’s Law

Any figure that looks interesting or different is usually wrong

- If something is “amazing,” find the flaw!
- Examples
  - If you have a mandatory birth date field and people think it’s unnecessary, you’ll find lots of 11/11/11 or 01/01/01
  - If you have an optional drop down, do not default to the first alphabetical entry, or you’ll have lots of: jobs = Astronaut
  - For most web sites, traffic will spike between 1-2AM November 3, 2013, relative to the same hour a week prior. Why?
- The previous Office example assumes click maps to revenue. Seemed reasonable, but when the results look so extreme, find the flaw

Hard to Assess the Value of Ideas: Data Trumps Intuition

- 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
- We joke that our job is to tell clients that their new baby is ugly
- In Uncontrolled, Jim Manzi writes
  Google ran …randomized experiments… with [only] about 10 percent of these leading to business changes.
- In an Experimentation and Testing Primer by Avinash Kaushik, authors of Web Analytics: An Hour a Day, he wrote
  80% of the time you/we are wrong about what a customer wants
Hard to Assess the Value of Ideas: Data Trumps Intuition

- QualPro tested 150,000 ideas over 22 years
  - 75 percent of important business decisions and business improvement ideas either have no impact on performance or actually hurt performance...
- 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
- Our intuition is poor: 60-90% of ideas do not improve the metric(s) they were designed to improve (domain dependent). Humbling!

Key Lessons

- Avoid the temptation to try and build optimal features through extensive planning without early testing of ideas
- Experiment often
  - To have a great idea, have a lot of them -- Thomas Edison
  - If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster -- Mike Moran, Do it Wrong Quickly
- Try radical ideas. You may be surprised
  - Doubly true if it’s cheap to implement (e.g., shopping cart recommendations)
  - If you’re not prepared to be wrong, you’ll never come up with anything original – Sir Ken Robinson, TED 2006 (#1 TED talk)
The OEC

- If you remember one thing from this talk, remember this point
- OEC = Overall Evaluation Criterion
  - Agree early on what you are optimizing
  - Getting agreement on the OEC in the org is a huge step forward
  - 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 for Search

- **KDD 2012** paper (*)
- Search engines (Bing, Google) are evaluated on query share (distinct queries) and revenue as long-term goals

Puzzle

- A ranking bug in an experiment resulted in very poor search results
- Distinct queries went up over 10%, and revenue went up over 30%
- What metrics should be in the OEC for a search engine?
- Degraded (algorithmic) search results cause users to search more to complete their task, and ads appear more relevant

(*) KDD 2012 paper with Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, Ya Xu
Puzzle Explained

Analyzing queries per month, we have

$$ \frac{\text{Queries}}{\text{Month}} = \frac{\text{Queries}}{\text{Session}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Users}}{\text{Month}} $$

where a session begins with a query and ends with 30-minutes of inactivity. (Ideally, we would look at tasks, not sessions).

Key observation: we want users to find answers and complete tasks quickly, so queries/session should be smaller

In a controlled experiment, the variants get (approximately) the same number of users by design, so the last term is about equal

The OEC should therefore include the middle term: sessions/user

Agenda

- Controlled Experiments
- Examples: you’re the decision maker
- Running Experiments at scale and best practices
- The cultural challenge
Scaling Experiments at Bing

- We now run over 250 concurrent experiments at Bing

We used to lockdown for Dec holidays. No more

Running Controlled Experiments at Scale (1)

Numbers below are approximate to give sense of scale

- In a visit, you’re in about 15 experiments
- There is no single Bing. There are 30B variants (5^15)
- 90% of users are in experiments. 10% are kept as holdout

Sensitivity: we need to detect small effects
- 0.1% change in the revenue/user metric > $1M/year
- Not uncommon to see unintended revenue impact of +/-1% (> $10M)
- Sessions/UU, a key component of our OEC, is hard to move, so we’re looking for small effects
- Important experiments run on 10-20% of users
Running Controlled Experiments at Scale (2)

Challenges
- QA. You can’t QA all combinations, of course. What are the equivalence classes?
- For UI change, no need to QA combinations of relevance exps
- Alarming on anomalies is critical: notify experiment owners that there’s a big delta on metric M (100 metrics) for browser B
- Interactions (optimistic experimentation): everyone experiments. Run statistical tests for pairwise interactions, and notify owners.
- Carryover effects: reuse of “bucket of users” from one experiment to the next is problematic

Important Lesson: Performance

- Bing server time is under one second at the 95th percentile
- Is it worth improving?
- We ran slowdown experiments to see the impact: we introduce an artificial server delay
- Performance matters a LOT. Here’s the summary:

> 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
Lesson: Small Changes can have High ROI

- We made small changes to font colors in August 2013
- Can you see? Can you figure out which is better?

Lesson: Small Changes (2)

- The change was from the left version to the right version
- Users were more successful in their tasks (SSR)
- Users completed tasks faster (time-to-success)
- We made more money (over $10M annually)
- Companies set standard company color/fonts without appreciating the impact it can have
Best Practice: A/A Test

- Run A/A tests – simple, but highly effective
  - Run an experiment where the Treatment and Control variants are coded identically and validate the following:
    1. Are users split according to the planned percentages?
    2. Is the data collected matching the system of record?
    3. Are the results showing non-significant results 95% of the time?

This is a powerful technique for finding problems
- Generating some numbers is easy
- Getting correct numbers you trust is much harder!

Remove Bots for Analysis

- Bots are lucrative business, but they skew the statistics
- At Bing, >50% of traffic comes from bots
Best Practice: Ramp-up

- 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
- Abort the experiment if treatment is significantly worse on key metrics

Best Practice: Large User Samples

- 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)
- Higher user samples increase sensitivity, which helps confidence (lower p-values for same effect size)
- Fastest way to achieve that exposure is to run equal-probability variants (e.g., 50/50% for A/B)
- Exception: biggest sites in the world. On the Bing, we run experiments on 10-20% of users instead of 50/50%
- Small sites? You want larger effects, so you need less users, but still run 50/50%
Agenda

- Controlled Experiments
- Examples: you’re the decision maker
- Running Experiments at scale and best practices
- The cultural challenge

The Cultural Challenge

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

-- Upton Sinclair

Why people/orgs avoid controlled experiments

- 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
- We’ve heard: “we know what to do. It’s in our DNA,” and “why don’t we just do the right thing?”
Cultural Stage 1: Hubris

- The org goes through stages in its cultural evolution
- Stage 1: we know what to do and we’re sure of it
  - True story from 1849
    - John Snow claimed that Cholera was caused by polluted water
    - A landlord dismissed his tenants’ complaints that their water stank
    - Even when Cholera was frequent among the tenants
    - One day he drank a glass of his tenants’ water to show there was nothing wrong with it
  - He died three days later
  - That’s hubris. Even if we’re sure of our ideas, evaluate them
  - Controlled experiments are a powerful tool to evaluate ideas

Cultural Stage 2: Insight through Measurement and Control

- Semmelweis worked at Vienna’s General Hospital, an important teaching/research hospital, in the 1830s-40s
- In 19th-century Europe, childbed fever killed more than a million women
- **Measurement:** the mortality rate for women giving birth was
  - 15% in his ward, staffed by doctors and students
  - 2% in the ward at the hospital, attended by midwives
Cultural Stage 2: Insight through Measurement and Control

- He tries to control all differences
  - Birthing positions, ventilation, diet, even the way laundry was done
- He was away for 4 months and death rate fell significantly when he was away. Could it be related to him?
- Insight:
  - Doctors were performing autopsies each morning on cadavers
  - Conjecture: particles (called germs today) were being transmitted to healthy patients on the hands of the physicians
- He experiments with cleansing agents
  - Chlorine and lime was effective: death rate fell from 18% to 1%

Cultural Stage 3: Semmelweis Reflex

- Success? No! Disbelief. Where/what are these particles?
  - Semmelweis was dropped from his post at the hospital
  - He goes to Hungary and reduced mortality rate in obstetrics to 0.85%
  - His student published a paper about the success. The editor wrote
    We believe that this chlorine-washing theory has long outlived its usefulness… It is time we are no longer to be deceived by this theory
- In 1865, he suffered a nervous breakdown and was beaten at a mental hospital, where he died
- **Semmelweis Reflex** is a reflex-like rejection of new knowledge because it contradicts entrenched norms, beliefs or paradigms
- Only in 1800s? No! A 2005 study: inadequate hand washing is one of the prime contributors to the 2 million health-care-associated infections and 90,000 related deaths annually in the United States
Cultural Stage 4: Fundamental Understanding

- In 1879, Louis Pasteur showed the presence of Streptococcus in the blood of women with child fever
- 2008, 143 years after he died, there is a 50 Euro coin commemorating Semmelweis

True Story – Scurvy and Vitamin C

- Without fundamental understanding, you make mistakes
- Scurvy is a disease that results from vitamin C deficiency
- It killed over 100,000 people in the 16th-18th centuries, mostly sailors
- First known controlled experiment in 1747
  - 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
- But Lind didn’t 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
Summary: Evolve the Culture

- In many areas we’re in the 1800s in terms of our understanding, so controlled experiments can help
  - First in doing the right thing, even if we don’t understand the fundamentals
  - Then developing the underlying fundamental theories

---

Summary

The less data, the stronger the opinions

1. Empower the HiPPO with data-driven decisions
   - HiPPO = Highest Paid-Person in Org, or Highest Paid-Person’s Opinion
   - Hippos kill more humans than any other (non-human) mammal (really)
   - OEC: make sure the org agrees what you are optimizing (long term lifetime value)

2. It is hard to assess the value of ideas
   - Listen to your customers – Get the data
   - Prepare to be humbled: data trumps intuition

3. Compute the statistics carefully
   - Getting a number is easy. Getting a number you should trust is harder

4. Experiment often to accelerate innovation
   - Triple your experiment rate and you triple your success (and failure) rate. Fail fast & often in order to succeed
Resources and Q&A

- [http://exp-platform.com](http://exp-platform.com) has papers, talks including
  - Controlled Experiments on the Web: Survey and Practical Guide (Data Mining and Knowledge Discovery journal)
  - Online experiments at Microsoft (Third Workshop on Data Mining Case Studies and Practice Prize)
  - Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained (KDD 2012)
  - Online Controlled Experiments at Large Scale (KDD 2013)

Appendix
Failures are Not Mistakes

- One page article by Stefan Thomke, May 2006
- Mistakes refer to the wrong actions that result from poor judgments or inattention; they should be avoided because they produce little new or useful information.
- A poorly planned or badly conducted experiment that results in ambiguous data, forcing researchers to repeat the experiment, is a mistake.
- Another common mistake is repeating a prior failure or learning nothing from the experience.

Failures are Not Mistakes

Story about Tom Watson Sr., IBM's founder

- A promising young executive involved in a risky new venture, managed to lose more than $10 million while trying to make the venture work
- When the nervous man was called into Watson's office, he offered to accept the logical consequence of losing the company such a large amount of money: "I guess you want my resignation, Mr. Watson."
- Much to his surprise, Watson countered: 'You can't be serious! We've just spent 10 million [dollars] educating you.'
Randomization

Good randomization is critical.
It’s unbelievable what mistakes developers 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)

Issues with Controlled Experiments (1 of 2)

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

Scope: Experimentation is not applicable everywhere

- Necessary ingredients for experimentation to be useful detailed in paper at http://exp-platform.com/expMicrosoft.aspx
- Sweet spot: websites and services that practice agile development

Org has to agree on OEC (Overall Evaluation Criterion)

- This is hard, but it provides a clear direction and alignment
- Some people claim their goals are “soft” or “intangible” and cannot be quantified. Give them Hubbard’s How to Measure Anything

Quantitative metrics, not always explanations of “why”

- A treatment may lose because page-load time is slower.
  At Amazon, we slowed pages by 100-250msec and lost 1% of revenue
- A treatment may have JavaScript that fails on certain browsers, causing users to abandon
Issues with Controlled Experiments (2 of 2)

- **Primacy/novelty 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

- **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

---

Dangers of Fragmentation: False Positives

- Richard Peto’s classical article showed that Aspirin helped patients after a heart attack
- The Lancet journal required analysis by subgroup as a condition for publication: which patients have the greatest benefit (males vs. females, age groups, etc)
- He refused to fragment the data; the journal insisted
- He finally agreed, but analyzed by astrological signs. Libra and Gemini don’t respond to Aspirin as well
  
OEC Example: Amazon Goldbox

- From eMetrics 2003 talk: Front Line Analytics at Amazon.com (PDF)
- Goldbox was a cross-sell and awareness raising tool
- We allowed customers to buy items at an additional discount
- We got a lot of suggestions on how to improve our goldbox offers to make them more personalized, but...

It’s by design. We discounted items to encourage purchases in new categories!
Different OEC!

Which Recommendations?

Finding correlated items is easy.
Deciding what, how, and when to present to the user is hard
-- Francisco Martin’s RecSys 2009 keynote

Amazon is well known for Bought X -> Bought Y

- Don’t tweak the algo to compute P(Y|X), apply it differently!
  - We tried Viewed X -> Viewed Y
  - Then Viewed X -> Bought Y

Useful, as it warns you if users who are viewing the product you’re viewing end up buying something else!
- Then Searched X -> Bought Y. This was a home run (next slides)
Amazon Behavior-Based Search (BBS) - Motivation

Searches for “24” are underspecified, yet most users want the TV program

Without BBS’s Search X->Bought Y, you get random stuff:
- 24 count Crayola
- 2.4Ghz USB adapter
- Dress for 24-months-old girls

- The screen shot was generated Sept 2012 by adding “-foo” to the query. Since nobody searches for that, the BBS algorithm doesn’t kick in

Built Searched X -> Bought Y (Behavior based Search)

- Ran controlled experiment with MVP (Minimum Viable Product):
  - Very thin UI integration (search team was busy)
  - Strong correlations shown at top of page, pushing search results down
  - Simple de-duping of results

- Result : +3% increase to revenue(*), i.e., 100s of millions of dollars!

- More [here](#)

(*) Based on UW iEdge Seminar talk by Amazon, 4/2006
Right Offer at the Right Time

- 2003 eMetrics: Front Line Analytics at Amazon.com (PDF): Amazon’s home page was auto-optimizing: offers in slots were evaluated based on real-time experiments
- Credit-card offer was winning the top slot, which seemed wrong since it had very low clickthrough-rate
- The reason: very profitable (high expected value)
- My team moved it from the home page to the shopping cart (purchase intent) with simple math UI
- Highly successful, both for Amazon and for users: right offer at the right time
- You now see this on other sites (e.g., disney)

Education is Hard: Interrupting Tasks

- In 2003, Amazon was well known as a book seller
- Wanted to educate users that it sells other things
- Added trivia questions, such as
  - How many pots and pans are available on Amazon.com?
    - a. Zero: Amazon only sells only books
    - b. Two
    - c. Over 100
  - If you “guessed” correctly (usually the highest number), Amazon added a nickel to your account
- Not shown now because of long-term controlled experiment
- Most “education” campaigns with pop-ups/eye-grabbing-UI/videos annoy users and are useless when properly evaluated
Explanations Help

- This was stated many times at RecSys 2012
- Telling users *why* they’re getting a recommendation is useful

- Amazon: people who bought X bought Y explains why you’re getting a recommendation for Y
- Amazon e-mails: as someone who bought from Author X, …
- Netflix: more like X, Watch it Again, …

- Allow users to “fix” the reason (e.g., don’t use X for recommendations)

Amazon Shopping Cart Recommendations

- 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, and the rest is history

Controlling in Observational Studies

One could try to control for confounding factors, something called \textit{causal sufficiency}, but this is very hard to do.

Example (summary)

- Study on teen sex concluded that youngsters who lose their virginity earlier than their peers are more likely to become juvenile delinquents.
- Controlled for over 10 variables, including: Gender, race, receipt of public assistance, parental education, family structure, previous substance use and depression, importance of religion, school GPA, relative pubertal status, virginity pledge status.
- Study was accepted and published.
- A PhD student then used the same data, but found 534 same-sex twins, so it controlled for genetics. The results were reversed!
- Methodology is stronger, so new result was published (in the same journal).
- Causal sufficiency is a strong assumption, easy to get wrong.

Observational Studies are Likely Wrong

- Stanley Young and Alan Karr \textit{compared} published results from hypotheses shown to be significant using observational studies with randomized clinical trials, considered more reliable.
- Their conclusion
  
  \textit{Any claim coming from an observational study is most likely to be wrong.}
“Find a house” widget variations

Overall Evaluation Criterion (OEC): Revenue to Microsoft generated every time a user clicks search/find button

- Raise your Left hand if you think A Wins
- Raise your Right hand if you think B Wins
- Don’t raise your hand if you think they’re about the same

If you did not raise a hand, please sit down.
Most people are highly opinionated, which is the opposite of what happens in practice: many experiments are “flat.”

If you raised your right hand, please sit down

A was 8.5% better

Since this is the #1 monetization, it effectively raised revenues significantly

Actual experiment had 6 variants.
If you’re going to experiment, try more variants, especially if they’re easy to implement