

Online Controlled Experiments: Introduction, Learnings, and Humbling Statistics

Ronny Kohavi
Online Services Division, Microsoft



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



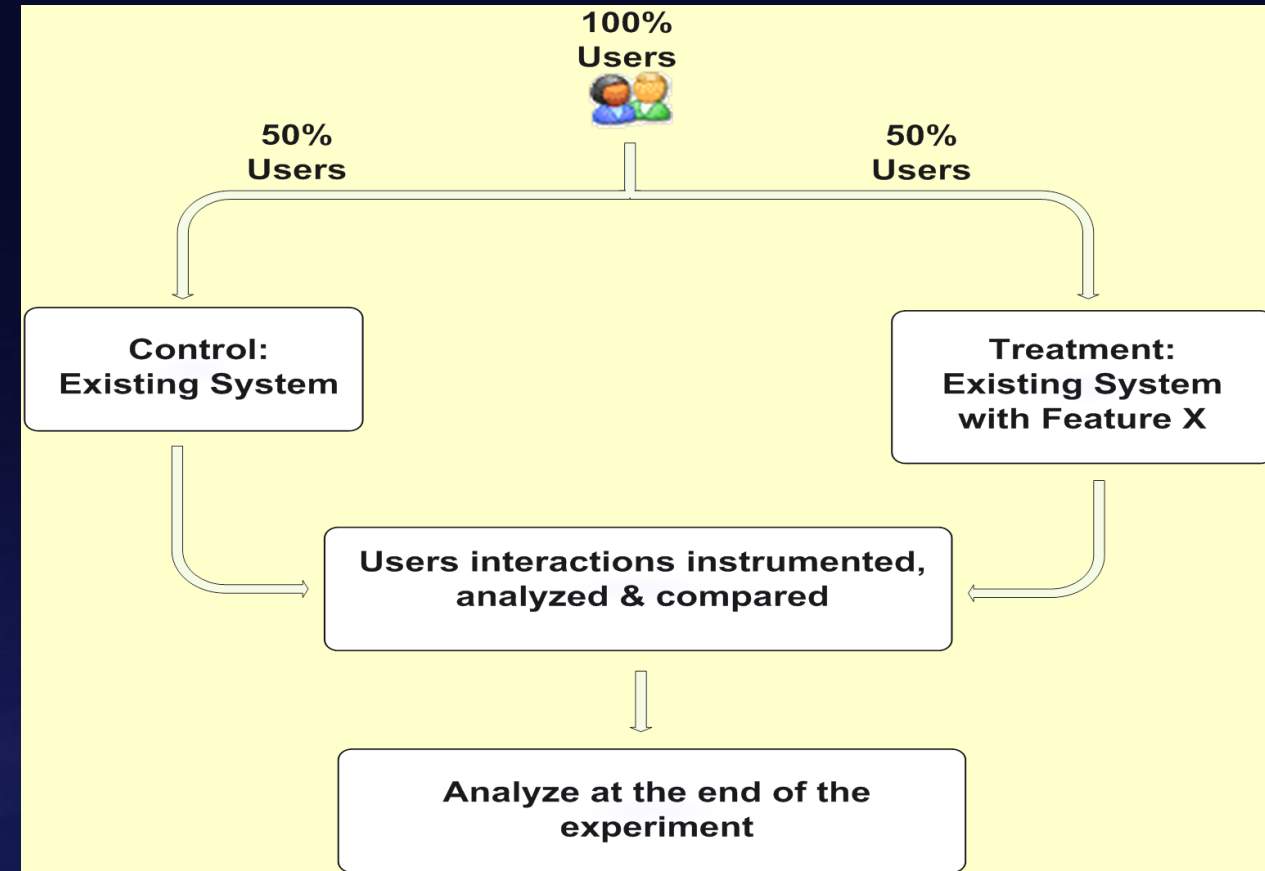
Agenda

- Controlled Experiments
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Running Experiments at scale and best practices
- Recommendation themes

- 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

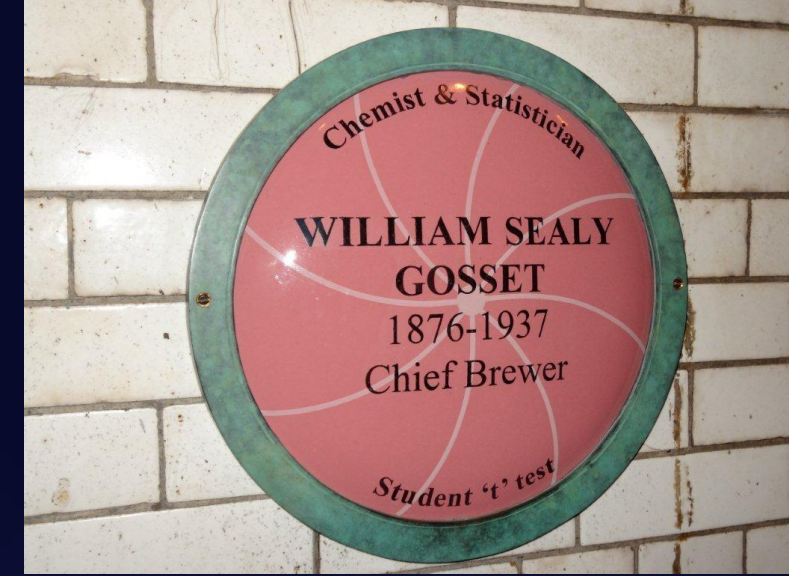
- Concept is trivial
 - Randomly split traffic between two (or more) versions
 - A (Control)
 - B (Treatment)
 - Collect metrics of interest
 - Analyze



- 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)

The Dublin Connection

- It is an honor to present this talk at Dublin, 10 minutes from where the "t-test" was introduced by William Gosset in 1908 (picture I took this week)
- Gosset worked at Guinness, which prohibited its employees from publishing papers. He published the papers under the pseudonym "Student"
- The student t-test is commonly used in determining statistical significance in controlled experiments
- The 2nd floor of the Guinness Storehouse is dedicated to "ads." Imagine if instead of Student t-test, it was Guinness t-test 😊



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 Blackberry because you bought a microSD card
- Buy Atonement movie DVD because you bought a Maglite flashlight
(must be a dark movie)
- Buy Organic Virgin Olive Oil because you bought Toilet Paper.
(If there is causality here, it's probably in the other direction.)



BlackBerry 8300 Skin, Black
by BlackBerry (Jun 7, 2007)
Average Customer Review: ★★★★★ (21)
Not in stock; order now and we'll deliver when available

List Price: \$9.99
Price: \$6.10
12 used & new from \$0.74

I own it Not interested Rate it

Recommended because you purchased **Kingston 2GB microSD Memory Card, Retail Package** (Fix



Atonement (Widescreen Edition)
DVD ~ Keira Knightley (Mar 18, 2008)
Average Customer Review: ★★★★★ (99)
In Stock

List Price: \$29.98
Price: \$15.99
24 used & new from \$13.77

I own it Not interested Rate it

Recommended because you purchased **Mag Instrument Three Cell AA Mini Maglite LED Flashlight**



Zoe Organic Extra Virgin Olive Oil, 25.5-Ounce Tins (Pack
by Zoe
Average Customer Review: ★★★★★ (21)
Usually ships in 3 to 4 weeks

List Price: \$26.64
Price: \$15.40

I own it Not interested Rate this item

Recommended because you purchased **Cottonelle Ultra Toilet Paper Double Roll, White 176, 12...**

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](#)

Examples

- Three experiments that ran at Microsoft
- 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)
 - B wins

MSN Real Estate

- “Find a house” widget variations
- Overall Evaluation Criterion(OEC): Revenue to Microsoft generated every time a user clicks search/find button



Find Your Dream Home or Apartment

City, State or ZIP

Existing homes New construction
 Foreclosures Rentals

Search listings ▶

A



Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State ▼
or
Enter Zip

Find homes ▶

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

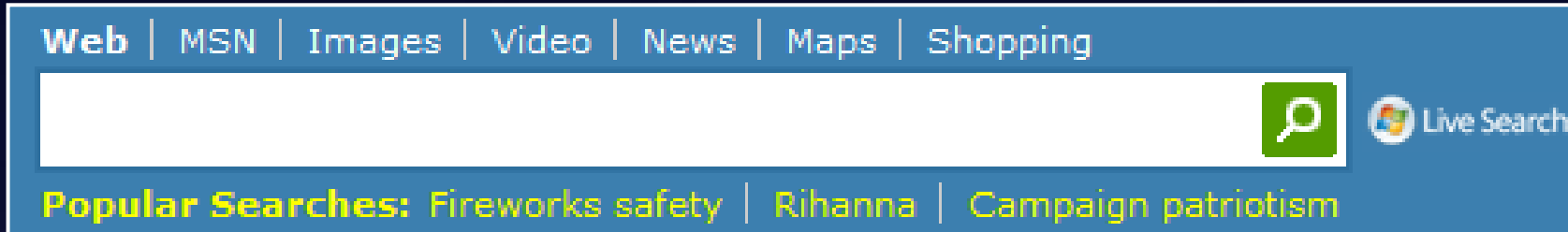
MSN Real Estate

- 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

MSN Home Page Search Box

OEC: Clickthrough rate for Search box and popular searches

A



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

- Insight

Stop debating, it's easier to get the data

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

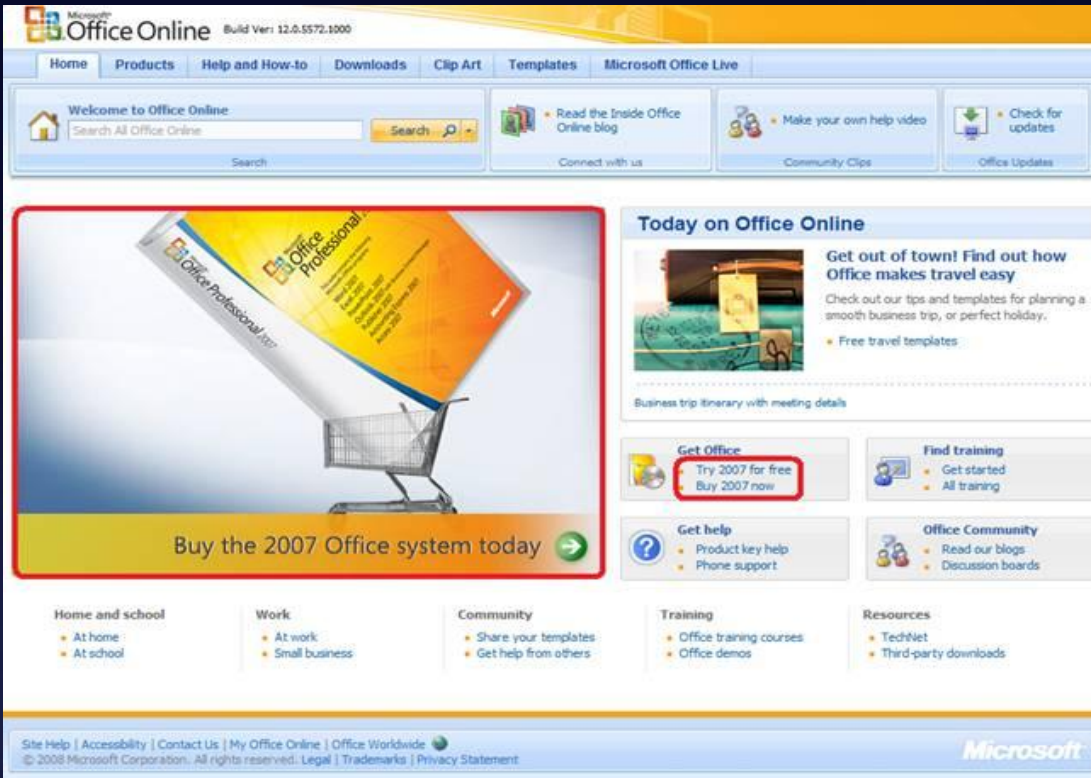


In line with Steve Krug's great book: Don't Make Me Think

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

Office Online

Twyman's Law

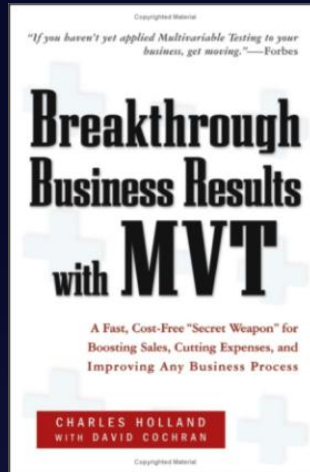
Any figure that looks interesting or different is usually wrong

- 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
 - Traffic to web sites doubled between 1-2AM November 6, 2011 for many sites, 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 the recently published book *Uncontrolled*, Jim Manzi writes
Google ran approximately 12,000 randomized experiments in 2009, 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

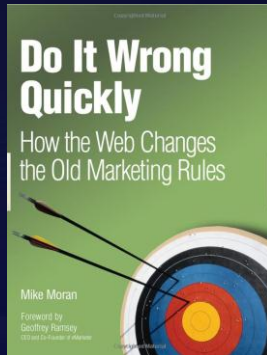
Hard to Assess the Value of Ideas: Data Trumps Intuition



- 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!

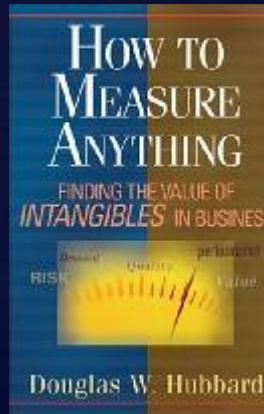
Key Lessons

- 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

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

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

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!

OFFER 7 8 9 10

Ron, you now have 60 minutes before these deals expire!
Order now for special Gold Box™ savings, or hold the deal you prefer to see another.

Le Creuset 9-1/2-Inch Nonstick Omelet Pan, Cherry Red [Kitchen & Housewares]

Pirates of the Caribbean - The Curse of the Black Pearl [DVD]

List Price: ~~\$98.00~~
Our Regular Price: ~~\$49.99~~
Gold Box Coupon: **\$7.50**

List Price: ~~\$29.99~~
Our Regular Price: ~~\$19.49~~
Gold Box Coupon: **\$1.95**

Special Gold Box Price: \$42.49
You Save: **\$47.51**

Special Gold Box Price: \$17.54
You Save: **\$12.45**

Availability: Usually ships within 24 hours

Availability: Usually ships within 24 hours

[See full product details](#) Avg. Customer Rating: ★★★★★

[See full product details](#) Avg. Customer Rating: ★★★★★

Buy now! **Hold this offer**

Buy now! **Hold this offer**

Agenda

- Controlled Experiments
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Running Experiments at scale and best practices
- Recommendation themes



The Cultural Challenge

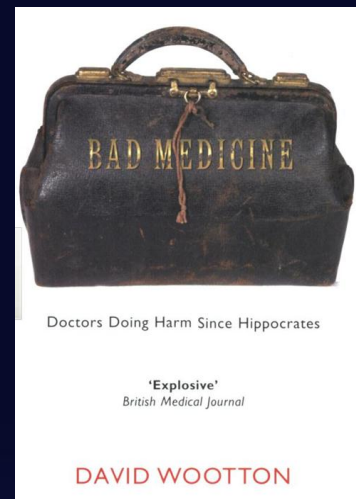
It is difficult to get a man to understand something when his salary depends upon his not understanding it.

-- Upton Sinclair

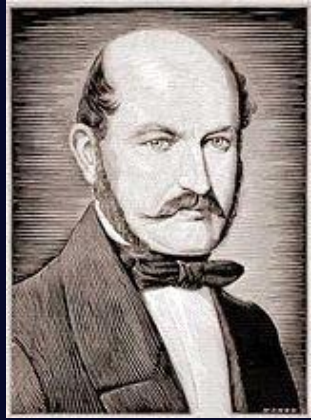
- Why people/orgs avoid controlled experiments
 - Some believe it threatens their job as decision makers
 - At Microsoft, program managers select the next set of features to develop. Proposing several alternatives and admitting you don't know which is best is hard
 - Editors and designers get paid to 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

With a new introduction by the authors

R. Codell Carter
Barbara R. Carter



**Childbed
Fever**

A Scientific Biography of Ignaz Semmelweis

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



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

Agenda

- Controlled Experiments
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Running Experiments at scale and best practices
- Recommendation themes



Running Controlled Experiments at Scale (1)

Numbers below are approximate to give sense of scale

- At Bing, we now run over 50 “useful” concurrent experiments
 - In a visit, you’re in about 10 experiments
 - There is no single Bing. There are 10M variants (5^{10})
- 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

| | | | | | |
|--------------|-------|-------|-------|-------|-------|
| UI | Exp 1 | Exp 2 | Exp 3 | Exp 4 | Exp 5 |
| Ads | Exp 1 | Exp 2 | Exp 3 | Exp 4 | Exp 5 |
| Relevance | ... | | | | |
| ... | | | | | |
| Feature area | | | | | |

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:
 - Feature areas (rows) get 5 experiment with disjoint users (no worries)
 - Optimistic experimentation: assume experiments between feature areas do no interact.
Run statistical tests for pairwise interactions, and notify owners.
- Carryover effects: reuse of "bucket of users" from one experiment to the next is problematic. Must rehash users (see KDD 2012 paper)

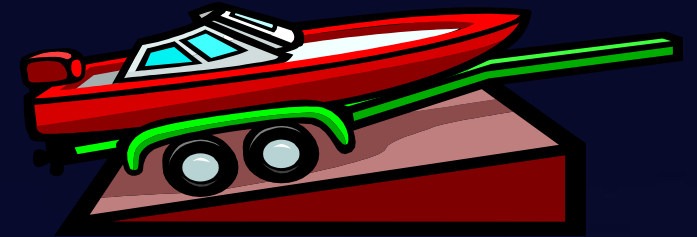
Best Practice: A/A Test

- 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!

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
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Running Experiments at scale and best practices
- Recommendation themes



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

Customers Who Bought This Item Also Bought

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

What Other Items Do Customers Buy After Viewing This Item?

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

"24 -foo"


Showing 1 - 16 of 7,668,656 Results Choose a Dep



Crayola 24 Ct Crayons by Crayola

~~\$4.99~~ **\$2.62** Add-on Item
 Add it to a qualifying order within **13 hours** to get it by Wednesday, Sep 12
 More Buying Choices
\$0.01 new (130 offers)


★★★★★ (22)
 Manufacturer recommended age: 4 - 15 Years
Toys & Games: See all 38,441 items



Medialink - Wireless N USB Adapter - 802.11n - 150Mbps - 2.4ghz - V
2003 / XP 32-Bit and 64-Bit / Vista 32-Bit and 64-Bit / Windows 7 32-B
Compatible by Mediabridge

~~\$65.99~~ **\$19.99** Prime
 Order in the next **13 hours** and get it by Wednesday, Sep 12.
 More Buying Choices
\$19.99 new (7 offers)
\$17.24 used (6 offers)


★★★★★ (1,793)
Electronics: See all 478,331 items



24 Season 1

\$0.00 Prime Instant Video
 \$1.99 to buy episodes \$29.99 (\$1.25 per episode) to buy season
 Watch **instantly** on your PC, Mac, compatible TV or device.

★★★★★ (14)
 Prime members have unlimited
 Amazon Instant Video
 Add to Watchlist
Movies & TV: See all 9,862 items



Youngland Baby-Girls Infant Legging Striped Dress by Youngland

~~\$32.00~~ **\$16.00** Prime
 Subscribe to Clothing E-mails for Discount See Details
Clothing & Accessories: See all 184,966 items

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](#)

"24"

Related Searches: [24 blu ray](#), [24 dvd](#), [24 season 1](#).

Showing 1 - 16 of 6,217,643 Results

24 Season 1 Starring Kiefer Sutherland, Leslie Hope, Sara...
\$0.00 Prime Instant Video
 \$1.99 to buy episodes \$29.99 (\$1.25 per episode) to buy season
 Watch **instantly** on your PC, Mac, compatible TV or device.

24 Season 2 Starring Kiefer Sutherland, Elisha Cuthbert, ...
\$0.00 Prime Instant Video
 \$1.99 to buy episodes \$38.99 (\$1.62 per episode) to buy season
 Watch **instantly** on your PC, Mac, compatible TV or device.

24 Season 3
\$0.00 Prime Instant Video
 \$1.99 to buy episodes \$38.99 (\$1.62 per episode) to buy season
 Watch **instantly** on your PC, Mac, compatible TV or device.

24 Season 4
\$0.00 Prime Instant Video
 \$1.99 to buy episodes \$38.99 (\$1.62 per episode) to buy season
 Watch **instantly** on your PC, Mac, compatible TV or device.

(*) 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., airlines)



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)

Summary

The less data, the stronger the opinions



1. Empower the HiPPO with data-driven decisions

- 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

- 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

Resources and Q&A

- <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)
- This talk at <http://www.exp-platform.com/Pages/2012RecSys.aspx>