# Data Science

## Deriving Knowledge from Data at Scale

Dec 7, 2015



# Lecture 10 Agenda

- Opening discussion 30
- Controlled Experiments 60
- Gold Sets 20
- Course Project Reflection 30
- Course Evaluation online

# Controlled Experiments

*Listen to your customers, not to the HiPPO (Highest Paid Person's Opinion)*

Forbes / Leadership

APR 15, 2013 @ 2:07 PM 41,718 VIEWS

# What Happens When a 'HiPPO' Runs Your Company?

**Chris DeRose and Noel Tichy**
CONTRIBUTOR

*We write about leadership judgment – the good, the bad and the ugly*

FOLLOW ON FORBES (18)

FULL BIO >

Opinions expressed by Forbes Contributors are their own.

Ron Johnson exited JC Penney JCP +0.00% humbled, with a list of accomplishments that he certainly didn't intend to leave as his legacy:

- The company burned through nearly $1 billion in 17 months, taking its cash balance from $1.8 billion to $930 million
- Revenue fell by 25 percent in 2012 for a loss of nearly $1 billion
- The company's market capitalization fell by nearly 50 percent on Johnson's watch

How does a leader initially hailed as the second-coming of Steve Jobs, ready to reinvent a 111-year-old staid retailer, crash and burn in such dramatic fashion? If JC Penney's board members had studied Amazon more closely, they might have discovered the answer. Leaders like Johnson have a special name at Amazon: they're called HiPPOs, which stands for "highest paid person's opinion." HiPPOs are leaders who are so self-assured that they need neither other's ideas nor data to affirm the correctness of their instinctual beliefs. Relying on their experience and smarts, they are quick to shoot down contradictory positions and dismissive of underling's input.


**Source:** http://www.forbes.com/sites/derosetichy/2013/04/15/what-happens-when-a-hippo-runs-your-company

# Introduction

- Three key lessons to share.
  We'll first do a Cambridge version of them.
- You're responsible for operations of the traffic lights in Cambridge, and there's one intersection where there are always problems with the light
- Do you
  - A - Send an engineer to find the root cause, which is estimated to take a few days
  - B – Put up a sign with a phone number to call when the lights don't work

# Traffic Lights

- Key lesson in this talk: it depends on your OEC, the Overall Evaluation Criterion, or what you're optimizing for
- Good OEC: Uptime (% of time light is properly operating)
- The twist: what's the time horizon?
  Are you optimizing for A. uptime this week, or B. for the next year?
- In Cambridge, it's B – put up a sign
- A retailer like Amazon should optimize for long-term customer lifetime value, not short term revenue.

  Picking a good OEC is key

# Local Example 2: Cruise

- You're planning a conference that takes place in June
- There's a dinner cruise planned
- Do you run the conference in
  - A: Seattle, WA
  - B: Cambridge, MA
- Both have good waterfronts and boats

# Cruise 2

- What should you ask first?
- What's the OEC, the Overall Evaluation Criterion
- The OEC can have multiple terms, but let's focus on rain. We want to minimize the probability that it will rain during the cruise
- What do you do?
- Lesson #2: GET THE DATA
- Cambridge gets over double the rainfall in June

US Geography / US Weather / Massachusetts Weather / Cambridge

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Avg. High | 35° | 37° | 45° | 55° | 66° | 76° | 81° | 78° | 72° | 62° | 52° | 40° |
| Avg. Low | 21° | 24° | 31° | 40° | 48° | 58° | 65° | 64° | 56° | 46° | 38° | 26° |
| Mean | 28° | 30° | 38° | 48° | 58° | 68° | 74° | 72° | 65° | 55° | 45° | 34° |
| Avg. Precip. | 3.6 in | 3.6 in | 3.7 in | 3.6 in | 3.3 in | 3.1 in | 2.8 in | 3.3 in | 3.1 in | 3.3 in | 4.3 in | 4.0 in |

US Geography / US Weather / Washington Weather / Seattle

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Avg. High | 45° | 48° | 52° | 57° | 64° | 68° | 75° | 75° | 68° | 58° | 50° | 45° |
| Avg. Low | 35° | 37° | 38° | 41° | 46° | 51° | 55° | 55° | 51° | 45° | 40° | 35° |
| Mean | 40° | 44° | 46° | 48° | 55° | 61° | 65° | 66° | 61° | 54° | 45° | 41° |
| Avg. Precip. | 5.4 in | 4.0 in | 3.5 in | 2.3 in | 1.7 in | 1.5 in | 0.8 in | 1.1 in | 1.9 in | 3.3 in | 5.8 in | 5.9 in |

# Final Simple Example

- You're responsible for programming the elevators in a bldg

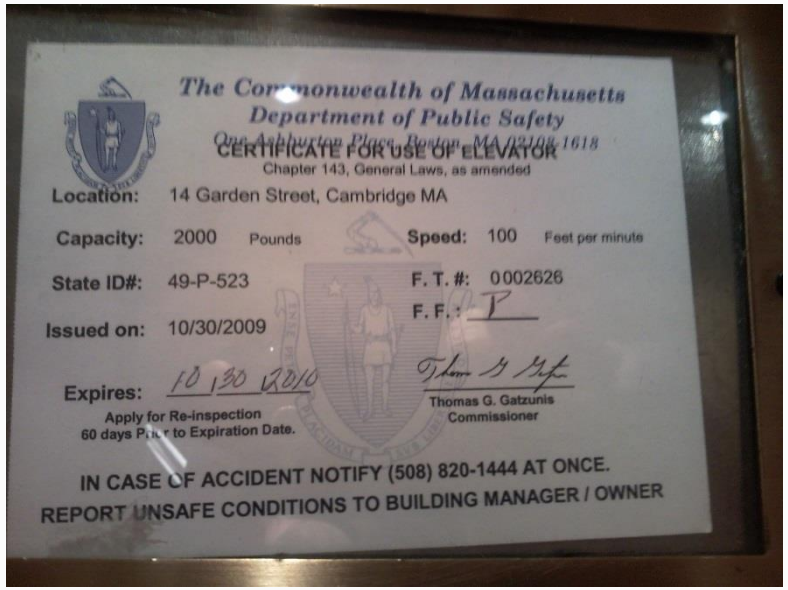
- Scenario:
  - Both elevators are at the Lobby floor, doors closed
  - Someone comes in from the Lobby and hits the "up" button
  - Do you program the elevators to
    - A – Open the left elevator's door
    - B – Open the right elevator's door
    - C – Open both doors

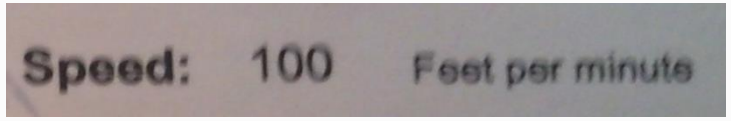
# Building Elevators

- First question you should ask?
- What's the OEC
- Answer: mean time to get customer to their floor
- Now what?
- Lesson #2: Get the data!
- Anyone who works in the building that knows the critical data?

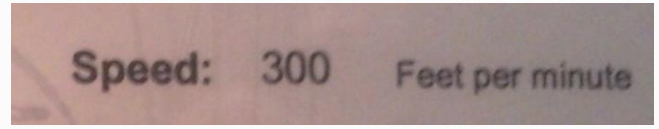
# Sheraton Elevators – Which Door to Open?

Right elevator is 3x faster.  Open that door.
Here's the data:




Left Elevator


Right Elevator

## Lesson #3: Prepare to be humbled

# Three Lessons to Remember from This Lecture

- Lesson #1: Ask what is the OEC?
  (Overall Evaluation Criterion, *recall first lecture…*)
  What are we optimizing for?
- Lesson #2: Get the Data
- Lesson #3: Prepare to be humbled

- In many scenarios getting the data is hard
- On the web, it's easy – we can run controlled experiments
  - *We can get the data, and it trumps our intuition*

# 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

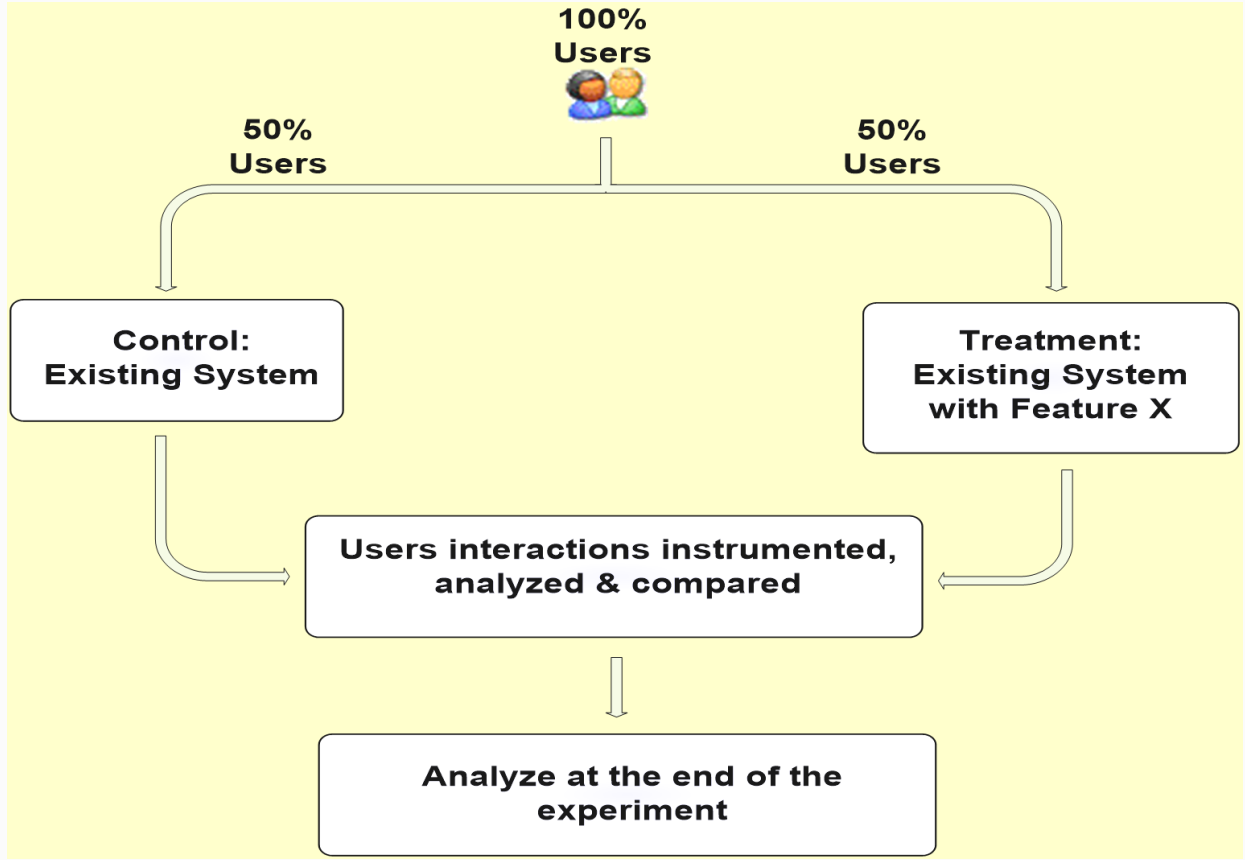

From Greg Linden's Blog: http://glinden.blogspot.com/2006/04/early-amazon-shopping-cart.html

# Agenda for Remainder of HiPPO Lecture

- Controlled Experiments in one slide
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding (Erik Brynjolfsson has identified 4 stages: measurement, experiment, share, replicate; this model starts earlier and ends later – see his TED talk)
- Quick overview of pros/cons of controlled experiments

# 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)

# Examples

- Three experiments that ran at Microsoft
- All had enough users for statistical validity
- Game: see how many you get right
  - As long as you guess correctly, you stay in the game; one incorrect, out!
  - Three choices are:
    - A wins  (the difference is statistically significant)
    - A and B are approximately the same (no stat sig diff)
    - B wins
  - If you guess randomly
    - 1/3 left in the game after first question
    - 1/9 after the second question

# MSN Real Estate

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


A


B

- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if you think they're about the same

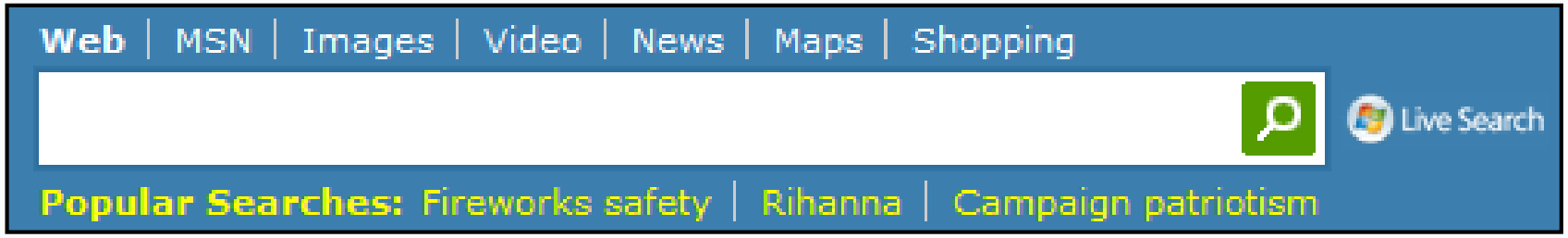
# MSN Real Estate

- If you did not raise a hand, you are OUT
- If you raised your left hand, you are OUT
- 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

# 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 right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if they are the about the same

# Search Box

- If  you raised any hand, you are out

- Insight

    Stop debating, it's easier to get the data

# MSN US Home Page: Search Box

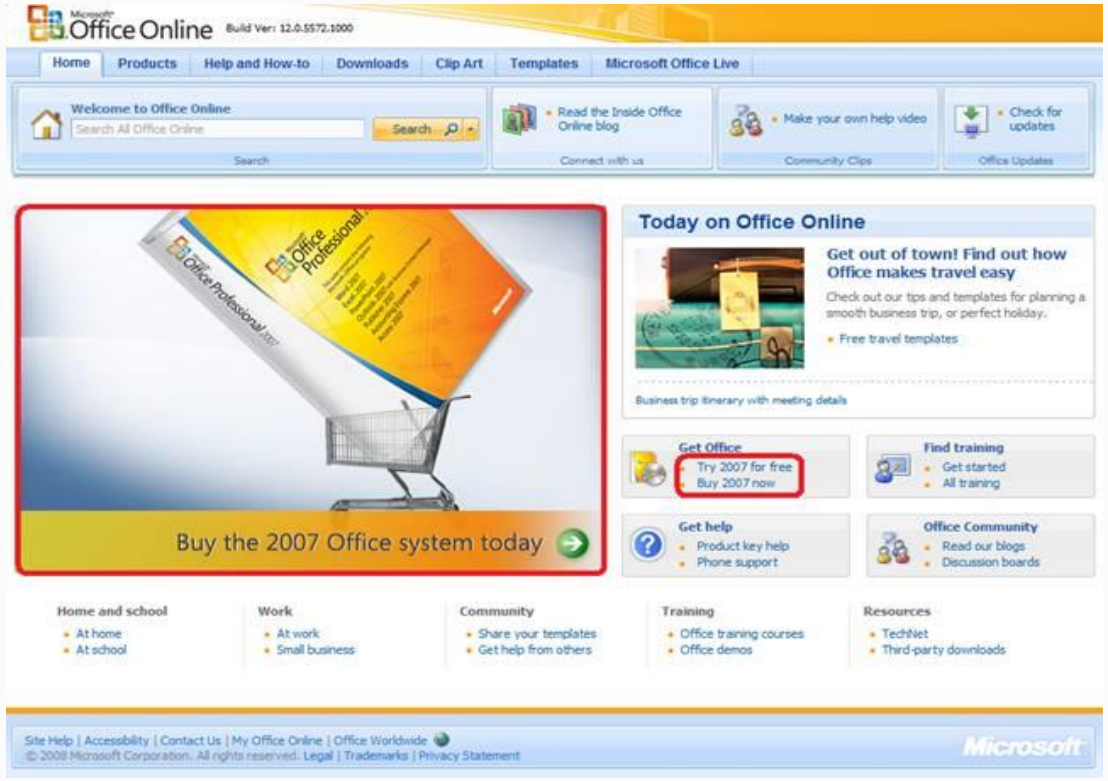
- A later test showed that changing the magnifying glass to an actionable word (search, go, explore) was highly beneficial.
- This:


is better than

# Office Online

## OEC: Clicks on revenue generating links (red below)

A




B


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

# Office Online

- If  you did not raise a hand, you are out
- If you raised your left hand, you are out
- B was 64% worse

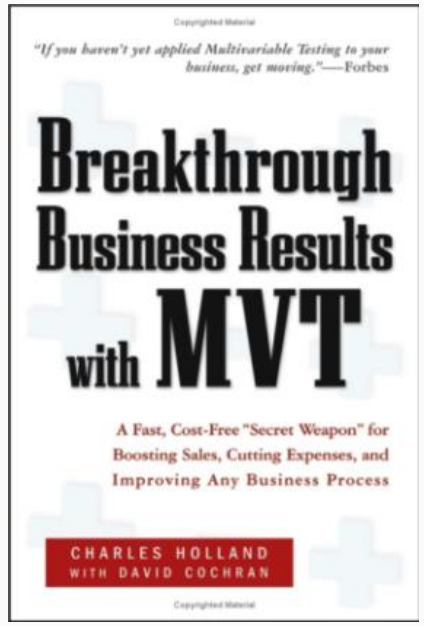
- What % of the audience is still in the game?
- Humbling!
- Remember lessons #2 and #3: get the data and prepare to be humbled

# Twyman's Law

*Any statistic that appears interesting is almost certainly a mistake*

- 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 jobs = Astronaut
- The previous Office example assumes click maps to revenue. Seemed reasonable, but when the results look so extreme, find the flaw (conversion rate is not the same; see why?)

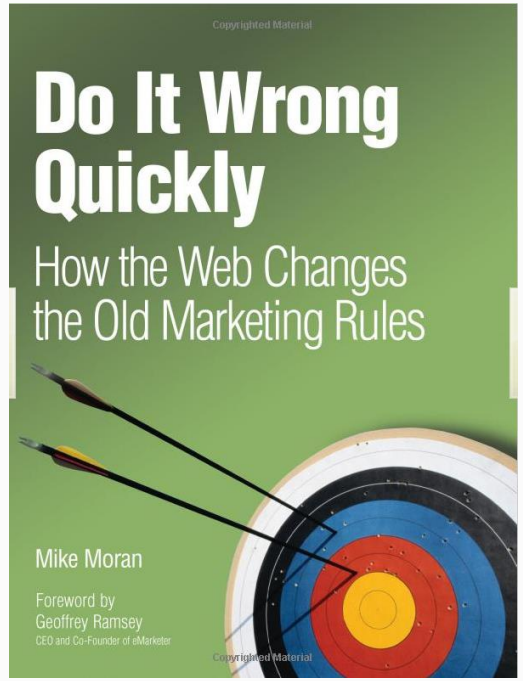
# Hard to Assess the Value of Ideas:
# Data Trumps Intuition




- 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…
- At Amazon, half of the experiments failed to show improvement
- Based on experiments with ExP at Microsoft
  - 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: 2/3rd of ideas do not improve the metric(s) they were designed to improve. 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



# The OEC

If you remember one thing from this lecture, remember this point

- OEC = Overall Evaluation Criterion
  - Agree early on what you are optimizing
  - Getting agreement on the OEC in the organization 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 why the OEC changed and raise new hypotheses

# Agenda

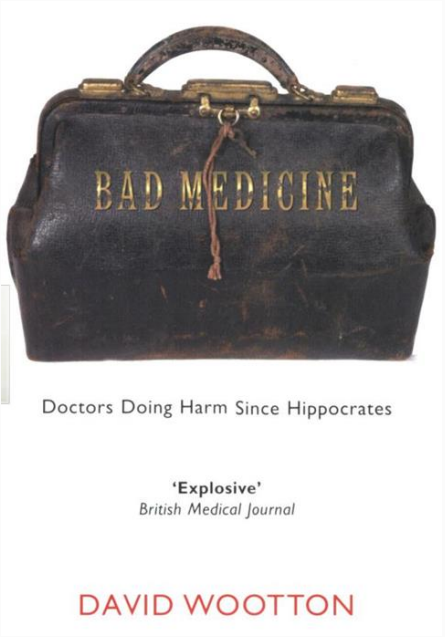
- Controlled Experiments in one slide
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Quick overview of pros/cons of controlled experiments

# 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
  - In a technology company, 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
  - Often 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

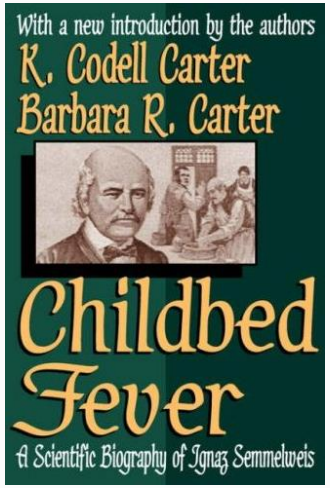



- The org goes through stages in its cultural evolution
- Stage 1: we know what to do and we're sure of it
  - True story from 1849
  - John Snow claimed that Cholera was caused by polluted water
  - A landlord dismissed his tenants' complaints that their water stank
    - Even when Cholera was frequent 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 tried 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! *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

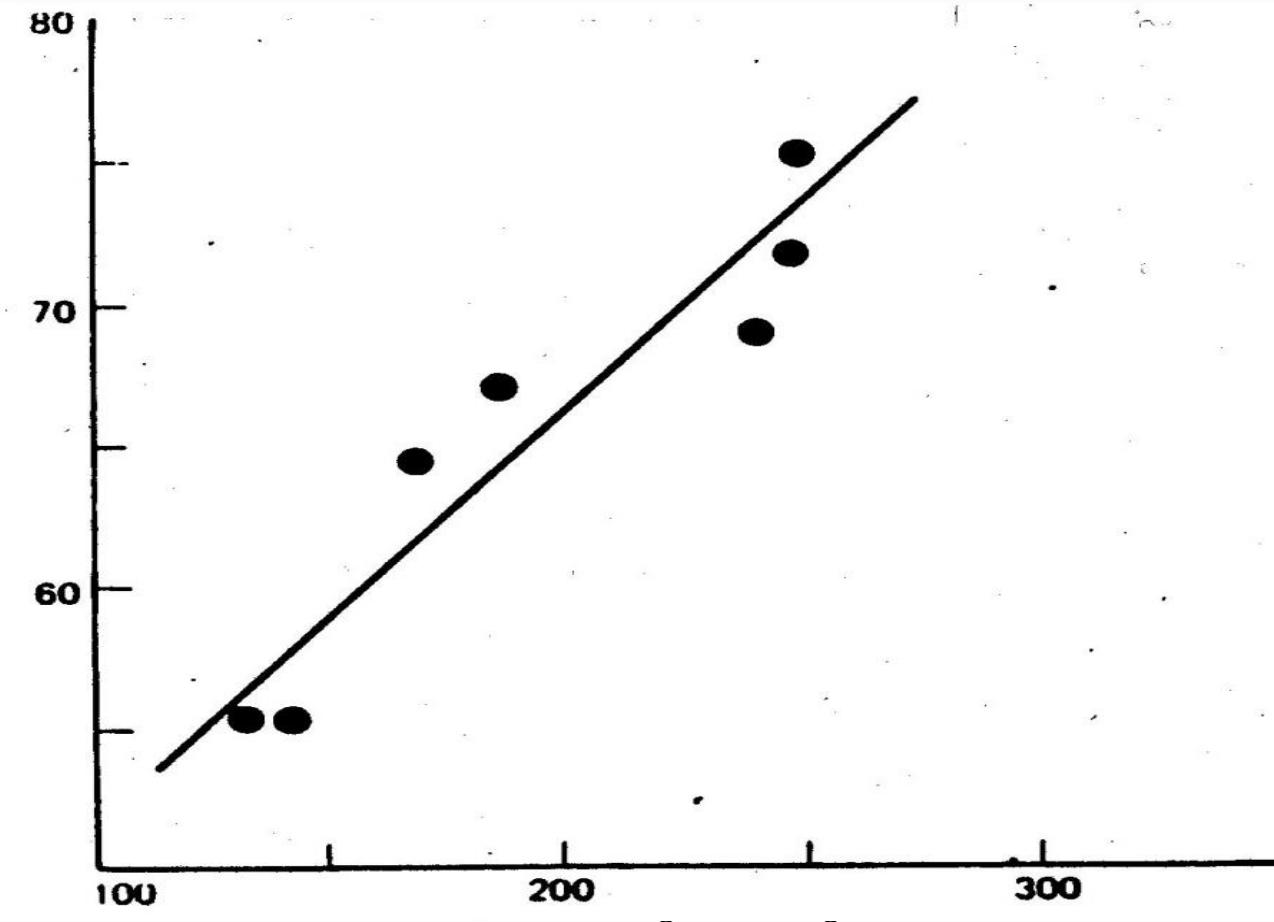

- In many areas we're in the 1800s in terms of our understanding, so controlled experiments can help
  - First in doing the right thing, even if we don't understand the fundamentals
  - Then developing the underlying fundamental theories

# Agenda

- Controlled Experiments in one slide
- Examples: you're the decision maker
- Cultural evolution: hubris, insight through measurement, Semmelweis reflex, fundamental understanding
- Quick overview of pros/cons of controlled experiments

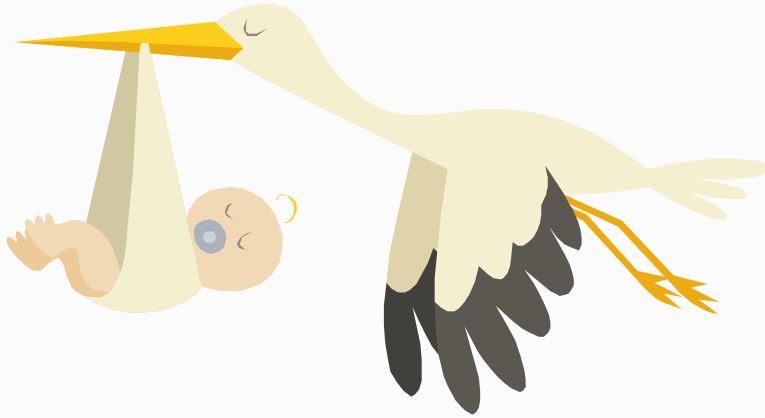
# Typical Discovery

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

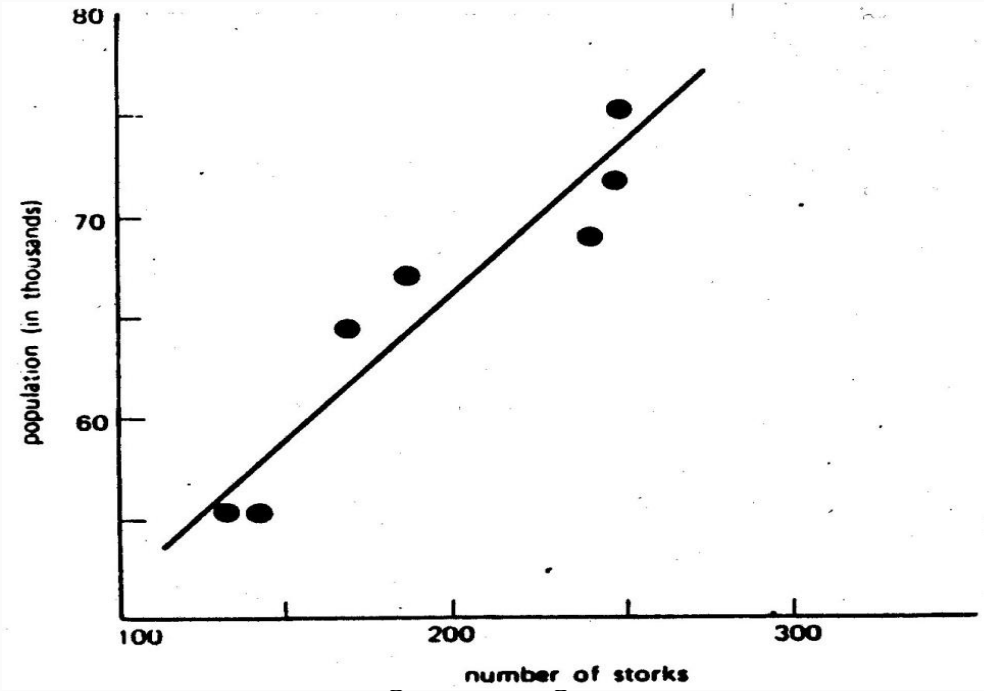


# Correlations are not Necessarily Causal

- 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 still not right, despite the strong correlational "evidence"




Ornitholigische Monatsberichte 1936;44(2)

# Correlation: 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

But…don't try to bandage your hands

Women have smaller palms and live 6 years longer on average

# Advantages of Controlled Experiments

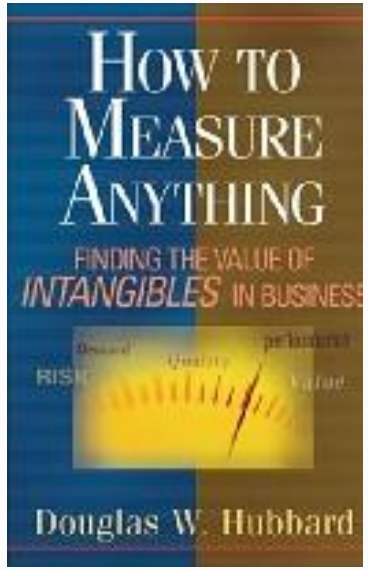
- Controlled experiments test for causal relationships, not simply correlations
- When the variants run concurrently, only two things could explain a change in metrics:
    1. The "feature(s)" (A vs. B)
    2. Random chance

  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

# 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
  - 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, they 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

# Best Practice: A/A Test

- Run A/A tests
  - 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 bugs and other integration issues before teams try to make data-driven decisions
  - 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 OEC or other key metrics (e.g., time to generate page)

# Best Practice: 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)
- Rare exception: biggest sites in the world.

  On the MSN US home page, we sample 10% of traffic
- If you perceive risk, don't start an experiment at 50/50% from the beginning: Ramp-up over a short period

# 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

# Lecture 10

## Out of Class Reading

Appears in KDD 2007.


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

Ron Kohavi
Microsoft
One Microsoft Way
Redmond, WA 98052
ronnyk@microsoft.com

Randal M. Henne
Microsoft
One Microsoft Way
Redmond, WA 98052
rhenne@microsoft.com

Dan Sommerfield
Microsoft
One Microsoft Way
Redmond, WA 98052
dans@microsoft.com

**ABSTRACT**

The web provides an unprecedented opportunity to evaluate ideas quickly using controlled experiments, also called randomized experiments (single-factor or factorial designs), A/B tests (and their generalizations), split tests, Control/Treatment tests, and parallel flights. Controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior. We provide a practical guide to conducting online experiments, where end-users can help guide the development of features. Our experience indicates that significant learning and return-on-investment (ROI) are seen when development teams listen to their customers, not to the Highest Paid Person's Opinion (HiPPO). We

**1. INTRODUCTION**

*One accurate measurement is worth more than a thousand expert opinions*
— Admiral Grace Hopper

In the 1700s, a British ship's captain observed the lack of scurvy among sailors serving on the naval ships of Mediterranean countries, where citrus fruit was part of their rations. He then gave half his crew limes (the Treatment group) while the other half (the Control group) continued with their regular diet. Despite much grumbling among the crew in the Treatment group, the experiment was a success, showing that consuming limes

*Eight (8) page conference paper*

Data Min Knowl Disc (2009) 18:140–181
DOI 10.1007/s10618-008-0114-1

**Controlled experiments on the web: survey and practical guide**

Ron Kohavi · Roger Longbotham · Dan Sommerfield · Randal M. Henne

Received: 14 February 2008 / Accepted: 30 June 2008 / Published online: 30 July 2008
Springer Science+Business Media, LLC 2008

**Abstract** The web provides an unprecedented opportunity to evaluate ideas quickly using controlled experiments, also called randomized experiments, A/B tests (and their generalizations), split tests, Control/Treatment tests, MultiVariable Tests (MVT) and parallel flights. Controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior. We provide a practical guide to conducting online experiments, where end-users can help guide the development of features. Our experience indicates that significant learning and return-on-investment (ROI) are seen when development teams listen to their customers, not to the Highest Paid Person's Opinion (HiPPO). We provide several examples of controlled experiments with surprising results. We review the important ingredients of running controlled experiments, and discuss their limita-

*40 page journal version…*

# *Short Break…*

# *Gold Sets*

# Gold Sets

Documented subsets of production data, specifically **curated** to *completely specify a problem* and *measure progress* on its solutions. Ideally, the gold set is **paired with a metric**, **target SLAs**, and a **score board** of existing algorithms. Advantages of having gold sets include:

- Reproducibility / Verifiability
- Low startup cost
- Merit based culture (meritocracy)
- Documented goals & achievements
- Automatic testing
- Detection and tracking of environment changes
- Transparency
- High innovation speed

# *This isn't easy…*

- *Building high quality gold sets is a challenge.*
- *It is time consuming.*
- *It requires making difficult and long lasting choices, and the rewards are delayed…*

# Gold Set Data Principles

If the system being built is expected to generalize to new data, be predictable, and be easy to improve on, it is advantageous to ***enforce a few principles***. Some of these are borrowed from machine learning.

1. **Distribution parity**: The training set and test set are drawn from same I.I.D. distribution (I.I.D. = independently identically distributed).
2. **Testing blindness**: No information from the test set can inform a training decision. *Blind sets further enforce this principle*.
3. **Production parity**: The distribution of the input and output of the production component and the gold set used to train it are the same.
4. **Single metric**: For each model, only one metric is optimized. Training and testing metric are the same.
5. **Reproducibility**: The gold sets are "pickled" and "sufficient". All information is contained (no external links or pointers) and frozen. Yardstick results should be easy to reproduce.
6. **Experimentation velocity**: Testing a solution on gold set is virtually free. This decouples algorithms from production. Analyze the errors, add features, train, repeat often.
7. **Data is gold**: The data and SLA define the problem. The model and implementation are ephemeral.

# Concrete examples of data mishaps

- Test set blindness: handwriting team reported 94% word accuracy on US English and 88% word accuracy on UK English. Reshuffled data between train and test for each and retrained both – results were 89% and 88%, respectively. *Suspicion: Some testing data knowledge found its way to the training set.*
- Reproducibility and Data is gold: Two systems have competed for Ad relevance, each claiming to be vastly superior to the other. Only recently, they were both tested on the same data. They performed identically. The good news is that the combined solution (ensemble) performed better. *Suspicion: A lot of energy was wasted in politics.*

# Gold Set Checklist

Building Gold sets is hard work. Many common and avoidable mistakes are made. This suggests having a checklist. Some questions will be trivial to answer or not applicable, some will require work…

1. Metrics: For each gold set, choose one (1) metric. Having two metrics on the same gold set is a problem (you can't optimize both at once).
2. Weighting/Slicing: Not all errors are equal. This should be reflected in the metric, not through sampling manipulation. Having the weighting in the metric has two advantages: 1) it is explicitly documented and reproducible in the form of a metric algorithm, and 2) production, train, and test sets results remain directly comparable (automatic testing).
3. Yardstick(s): Define algorithms and configuration parameters for public yardstick(s). There could be more than one yardstick. A simple yardstick is useful for ramping up. Once one can reproduce/understand the simple yardstick's result, it becomes easier to improve on the latest "production" yardstick. Ideally yardsticks come with downloadable code. The yardsticks provide a set of errors that suggests where innovation should happen.

# Gold Set Checklist

4. Sizes and access: What are the set sizes? Each size corresponds to an innovation velocity and a level of representativeness. A good rule of thumb is 5X size ratios between gold sets drawn from the same distribution. Where should the data live? If on a server, some services are needed for access and simple manipulations. There should always be a size that is downloadable (< 1GB) to a desktop for high velocity innovation.
5. Documentation and format: Create a format/API for the data. Is the data compressed? Provide sample code to load the data. Document the format. Assign someone to be the curator of the gold set.

# Gold Set Checklist

6. Features: What (gold) features go in the gold sets? Features must be pickled for result to be reproducible. Ideally, we would have 2, and possibly 3 types of gold sets.
   a. One set should have the *deployed features (computed from the raw data)*. This provides the production yardstick.
   b. One set should be Raw (e.g. contains all information, possibly through tables). This allows contributors to create features from the raw data to investigate its potential compared to existing features. This set has more information per pattern and a smaller number of patterns.
   c. One set should have an extended number of features. The additional features may be "building blocks", features that are scheduled to be deployed next, or high potential features. Moving some features to a gold set is convenient if multiple people are working on the next generation. Not all features are worth being in a gold set.

7. Feature optimization sets: Does the data require feature optimization? For instance, an IP address, a query, or a listing id may be features. But only the most frequent 10M instances are worth having specific trainable parameters. A pass over the data can identify the top 10M instance. This is a form of feature optimization. Identifying these features does not require labels. If a form of feature optimization is done, a separate data set (disjoint from the training and test set) must be provided.

# Gold Set Checklist

8. Stale rate, optimization, monitoring: How long does the set stay current? In many cases, we hide the fact that the problem is a time series even though the goal is to predict the future and we know that the distribution is changing. We must quantify how much a distribution changes over a fixed period of time. There are several ways to mitigate the changing distribution problem:
   a. Assume the distribution is I.I.D. (I.I.D. = independently identically distributed). Regularly re-compute training sets and Gold sets. Determine the frequency of re-computation, or set in place a system to monitor distribution drifts (monitor KPI changes while the algorithm is kept constant).
   b. Decompose the model along "distribution (fast) tracking parameters" and slow tracking parameters. The fast tracking model may be a simple calibration with very few parameters.
   c. Recast the problem as a time series problem: patterns are (input data from t-T to t-1, prediction at time t). In this space, the patterns are much larger, but the problem is closer to being I.I.D.
9. The gold sets should have information that reveal the stale rate and allows algorithms to differentiate themselves based on how they degrade with time.

# Gold Set Checklist

10. Grouping: Should the patterns be grouped? For example in handwriting, examples are grouped per writer. A set built by shuffling the words is misleading because training and testing would have word examples for the same writer, which makes generalization much easier. If the words are grouped per writers, then a writer is unlikely to appear in both training and test set, which requires the system to generalize to never seen before handwriting (as opposed to never seen before words). Do we have these type of constraints? Should we group per advertiser, campaign, users to generalize across new instances of these entities (as opposed to generalizing to new queries)? ML requires training and testing to be drawn from the same distribution. Problems arise when one partially draw examples from the same entity on both training and testing on a small set of entities. This breaks the IID assumption and makes the generalization on the test set much easier than it actually is.
11. Sampling production data: What strategy is used for sampling? Uniform? Are any of the following filtered out: fraud, bad configurations, duplicates, non-billable, adult, overwrites, etc? Guidance: use the production sameness principle.

# Course Project:

# Data Science

## Deriving Knowledge from Data at Scale

*Have a great holiday season!*

Dec 7th, 2015