CIKM 2008. Oct 29, 2008

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

Ronny Kohavi, General Manager
Experimentation Platform, Microsoft
ronnyk@microsoft.com

Joint work with multiple people at the Experimentation Platform team.

Amazon Shopping Cart Recs

- **Add an item to your shopping cart at a website**
  - Most sites show the cart

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

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

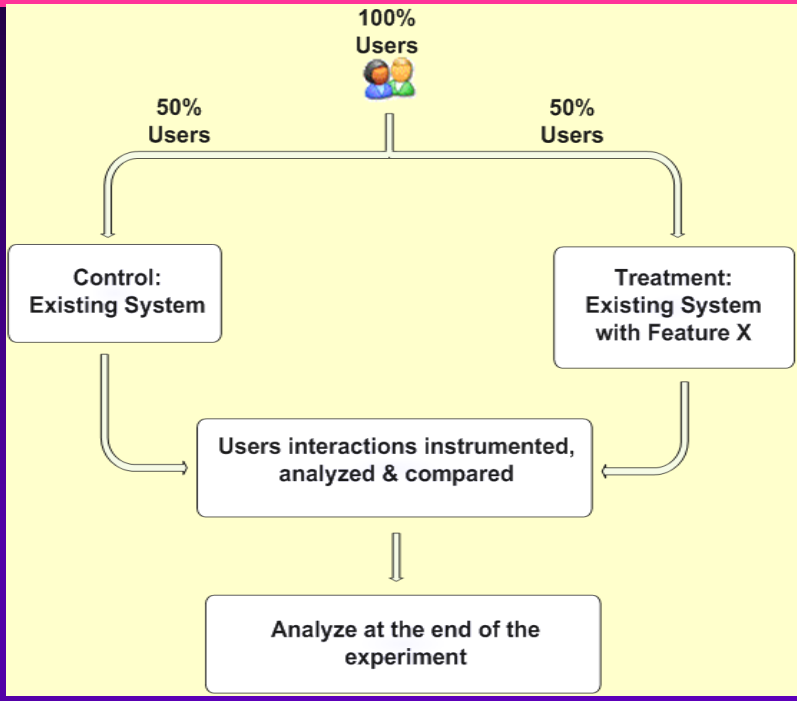
Agenda

- **Controlled Experiments in one slide**
- **Examples: you're the decision maker**
- **Culture, OEC (Overall Evaluation Criterion)**
- **Controlled Experiments: deeper dive**

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

Controlled Experiments

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



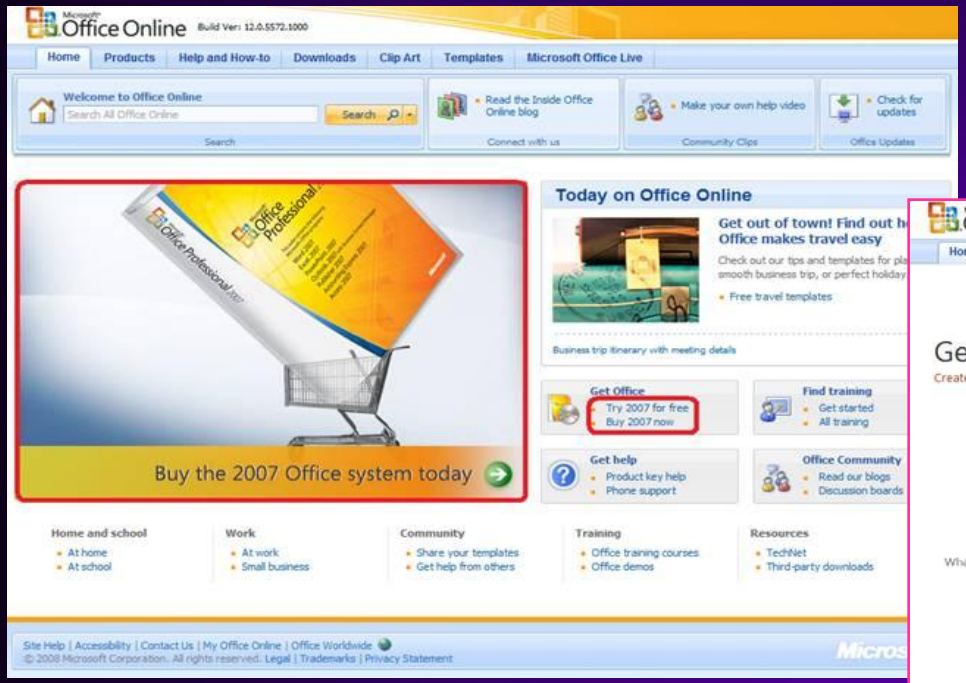
Examples

- **Three experiments that ran with ExP recently**
- **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
  - If you guess randomly
    - 1/3 left standing after first question
    - 1/9 after the second question

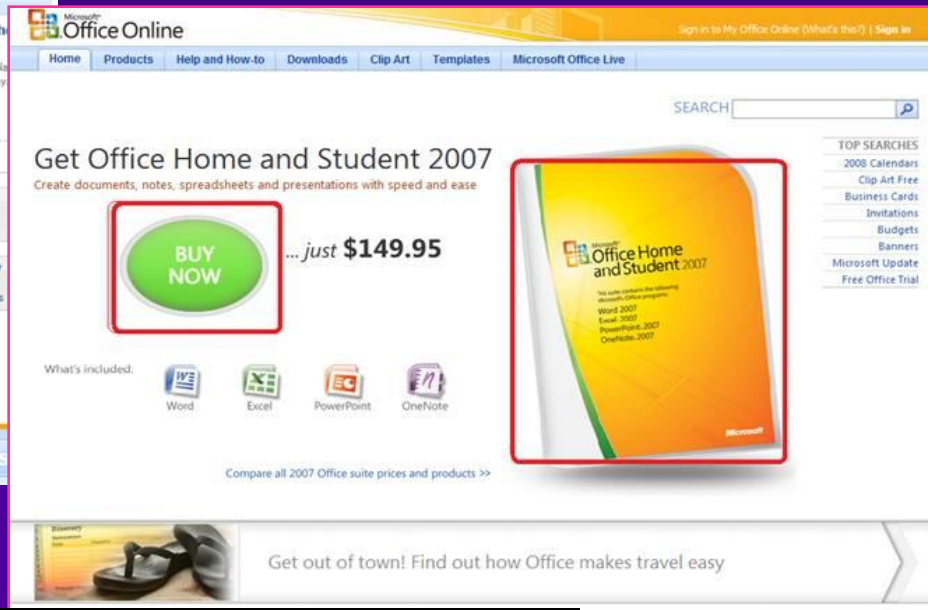
Office Online

Test new design for Office Online homepage

A



OEC: Clicks on revenue generating links (red below)



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

Office Online

- **If you did not raise a hand, please sit down**
- **If you raised your left hand, please sit down**
- **B was 64% worse**

The Office Online team wrote

> *A/B testing is a fundamental and critical Web services… consistent use of A/B testing could save the company millions of dollars*

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

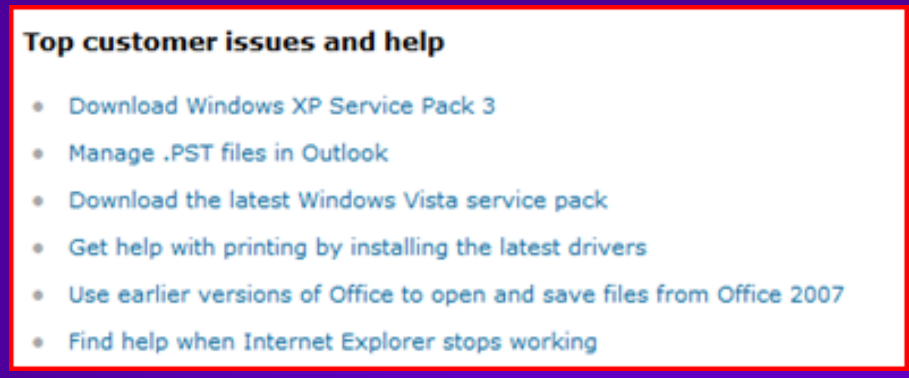
Search Box

- **If you raised any hand, please sit down**
- **Insight**

  Stop debating, it's easier to get the data

Microsoft Support

- **Support.microsoft.com shows "top issues"**
- **OEC = click-through rate**
- **A shows top issues**
- **B filters top issues to OS & Browser used to visit site (useragent)**



Personalization rarely hurts, but does it help?

- Raise your right hand if you think B Wins by over 30%
- Raise your left hand if you think B Wins by under 30%
- Don't raise your hand if you think they're about the same

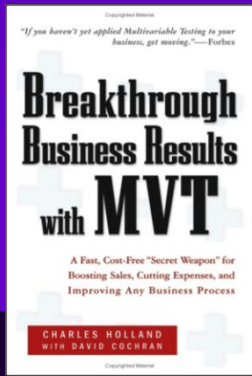
Microsoft Support

- If you did not raise a hand, please sit down
- If you raised your left hand, please sit down
- B was >50% better

Personalization helps more than people think!

Hard to Assess the Value of Ideas: Data Trumps Intuition

- **It is humbling to see how often we are wrong**
  - Experts are often wrong in many domains
  - Doctors did bloodletting for centuries until the 1836 when Pierre Louis ran a controlled experiment (randomized clinical trial)
- **At Amazon, more than half of the experiments failed to show improvement**
  - Every new feature was built because *someone* thought it was a great idea worth implementing (and convinces others)
- **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…



Agenda

- **Controlled Experiments in one slide**
- **Examples: you're the decision maker**
- **Culture, OEC (Overall Evaluation Criterion)**
- **Controlled Experiments: deeper dive**

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

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?"

Experimentation Culture

- **Learn from flat/negative results**
  - Even if an idea failed to improve the OEC, the org **learned** something. Failing fast is good
  - "*If you're not prepared to be wrong, you'll never come up with anything original*" – Sir Ken Robinson (TED 2006)
  - Deploy the positive experiments: only **their** sum really matters
- **To innovate, 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

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

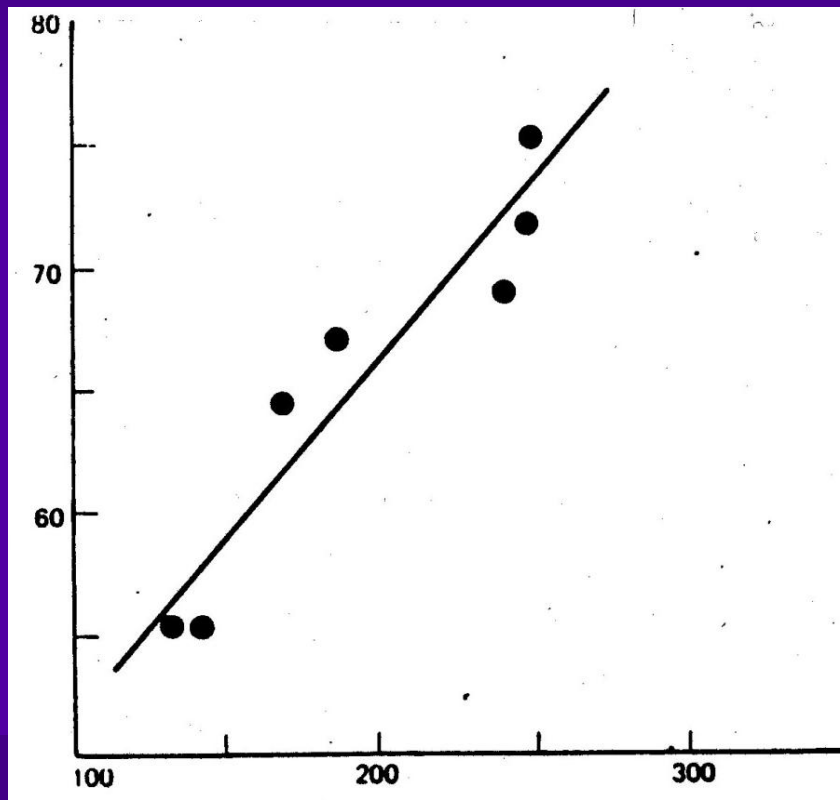
Agenda

- **Controlled Experiments in one slide**
- **Examples: you're the decision maker**
- **Culture, OEC (Overall Evaluation Criterion)**
- **Controlled Experiments: deeper dive**

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

Typical Discovery

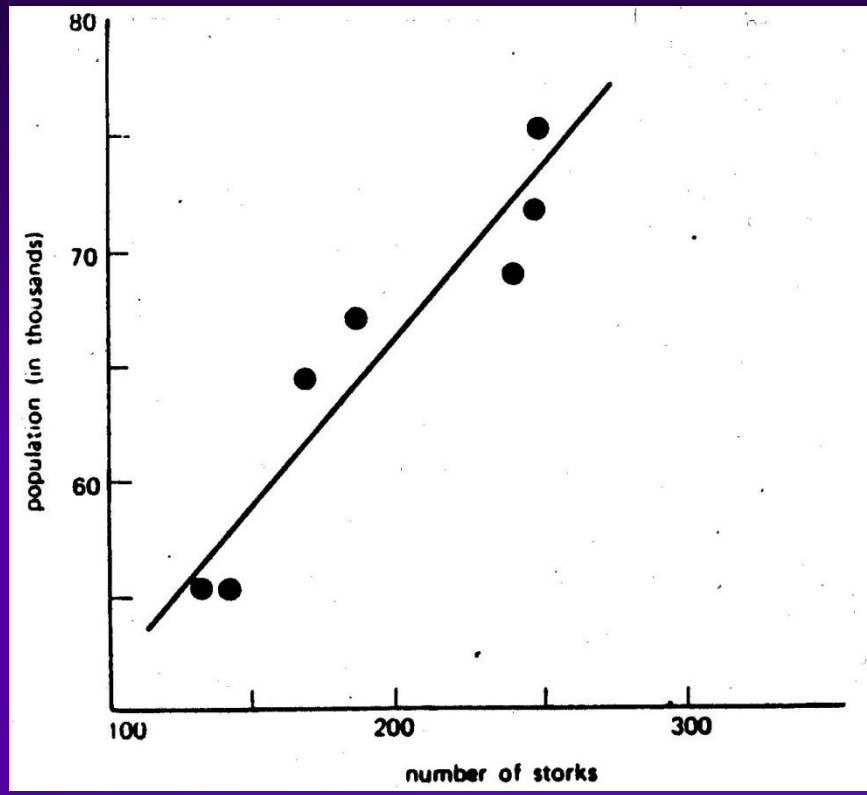
- **With data mining, we find patterns, but most are correlational**
- **Here is one a real example of two highly correlated variables**



Correlations are not Necessarily Causal

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

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



- Example 2:
  True statement (but not well known):
  Palm size correlates with your life expectancy

  The larger your palm, the less you will live, on average.

  Try it out - look at your neighbors and you'll see who is expected to live longer.

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

Ornitholigische Monatsberichte 1936;44(2)

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

- **Org has to agree on OEC (Overall Evaluation Criterion).**
  This is hard, but it provides a clear direction and alignment

- **Quantitative metrics, not always explanations of "why"**
  - A treatment may lose because page-load time is slower.
    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/newness effect**
  - Changing navigation in a website may degrade the customer experience (temporarily), even if the new navigation is better
  - Evaluation may need to focus on new users, or run for a long period
- **Multiple experiments**
  - Even though the methodology shields an experiment from other changes, statistical variance increases making it harder to get significant results. There can also be strong interactions (rarer than most people think)
- **Consistency/contamination**
  - On the web, assignment is usually cookie-based, but people may use multiple computers, erase cookies, etc. Typically a small issue
- **Launch events / media announcements sometimes preclude controlled experiments**
  - The journalists need to be shown the "new" version

Experimentation Platform Team

Mission: accelerate software innovation through trustworthy experimentation

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

Summary

The less data, the stronger the opinions

1. **It is hard to assess the value of ideas**
   - Listen to your customers
   - Get the data by experimenting because data trumps intuition
   - Examples are humbling. More at http://exp-platform.com/cikm.aspx
2. **Replace the HiPPO with an OEC**
   - Make sure the org agrees **what** you are optimizing (long term lifetime value)
3. **Compute the statistics carefully**
   - Power, 95% confidence, ramp-up
   - Stats/details described at http://exp-platform.com/hippo_long.aspx
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

http://exp-platform.com

Accelerating software Innovation through trustworthy experimentation

Extra Slides

MSN UK Hotmail experiment

Hotmail module on home page

MSN UK Hotmail experiment

A: When user clicks on email

hotmail opens in same window

B: Open hotmail in separate window

Trigger: only users that click in the module are in experiment (no diff otherwise)

OEC: clicks on home page (after trigger)

Penalty for users annoyed with new widow (opinionlab feedback)

- 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

UK Hotmail

- If you didn't raise a hand, please sit down
- If you raised your right hand, please sit down
- For those in the experiment, clicks on MSN HP increased +8.9%
- <0.001% of users in B wrote negative feedback about the new window

Data Trumps Intuition

- The experiment report was sent by the BI/CI team to all multiple teams across the world
- Someone who saw the report wrote

  This report came along at a really good time and was VERY useful.

  I argued this point to my team (open Live services in new window from HP) just some days ago.
  They all turned me down.

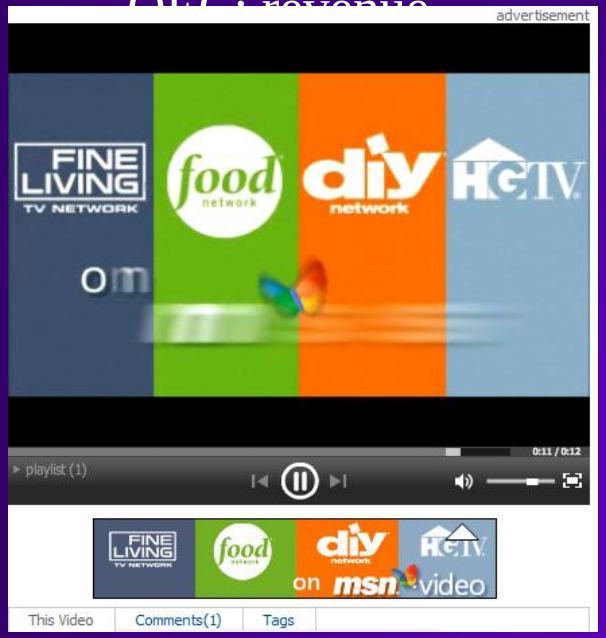
  Funny, now they have all changed their minds.

MSN Entertainment and Video Services (EVS)

Determine whether showing the first ad **after** the first video rather than before it would increase user engagement and loyalty without sacrificing ad revenue

A: Show ad then video

OEC: revenue



B: Show Video then ad

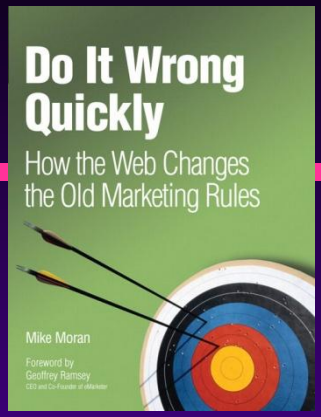
- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if about the same

MSN EVS

- **If you did not raise a hand, please sit down**
- **If you raised your left hand, please sit down**
- **Ad starts =revenue (OEC) for B down 56%**
  - Content starts per session up 8.5%
  - Repeat users up 2%
- **EVS wrote**

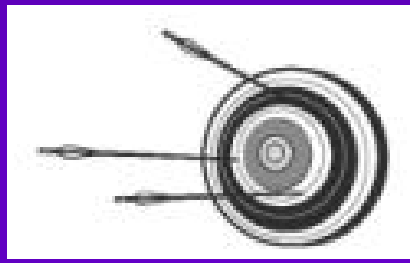
There is a preponderance of opinion driven design…
The results of the experiment were in some respect counterintuitive.
They completely changed our feature prioritization. It dispelled long held assumptions about video advertising. Very, very useful.

Do It Wrong Quickly



- We work on “the plan,” which is reviewed and approved by execs, then we execute flawlessly (or do we?)
- We’re looking to hit the arrow in the center—the bulls-eye

- But what if we the game is to score the most points, i.e., the sum of arrow scores.
- Shooting three arrows may be much more effective



Bloodletting (1 of 2)

- For many years, the prevailing conception of illness was that the sick were contaminated by some toxin
- Opening a vein and letting the sickness run out – bloodletting.
- One 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

Bloodletting (2 of 2)

- **President George Washington had a sore throat and 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 that is now recognized as one of the first clinical trials, or randomized controlled experiment. 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.**

Lesson: Compute Statistical Significance and run A/A Tests

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

Run Experiments at 50/50%

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

Ramp-up and Auto-Abort

- **Ramp-up**
  - Start an experiment at 0.1%
  - Do some simple analyses to make sure no egregious problems can be detected
  - Ramp-up to a larger percentage, and repeat until 50%
- **Big differences are easy to detect because the min sample size is quadratic in the effect we want to detect**
  - Detecting 10% difference requires a small sample and serious problems can be detected during ramp-up
  - Detecting 0.1% requires a population 100^2 = 10,000 times bigger
- **Automatically abort the experiment if treatment is significantly worse on OEC or other key metrics (e.g., time to generate page)**

Randomization

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