Aug 12, 2013, KDD

Online Controlled Experiments at Large Scale

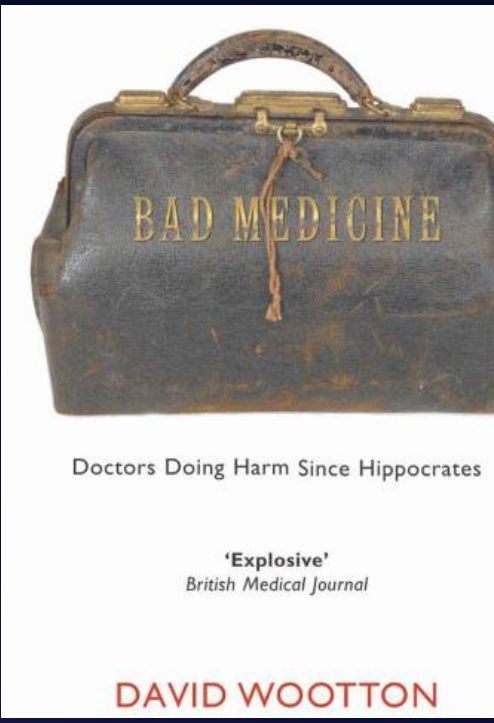
Ron Kohavi

Joint work with Alex Deng, Brian Frasca, Toby Walker, Ya Xu, Nils Pohlmann

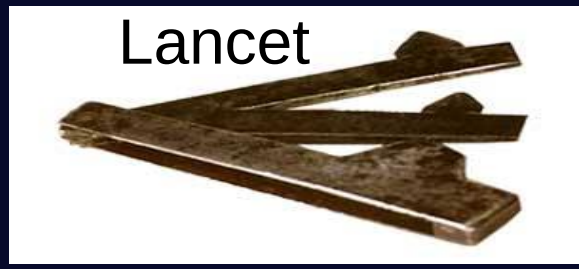
Paper at http://bit.ly/ExPScale

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)



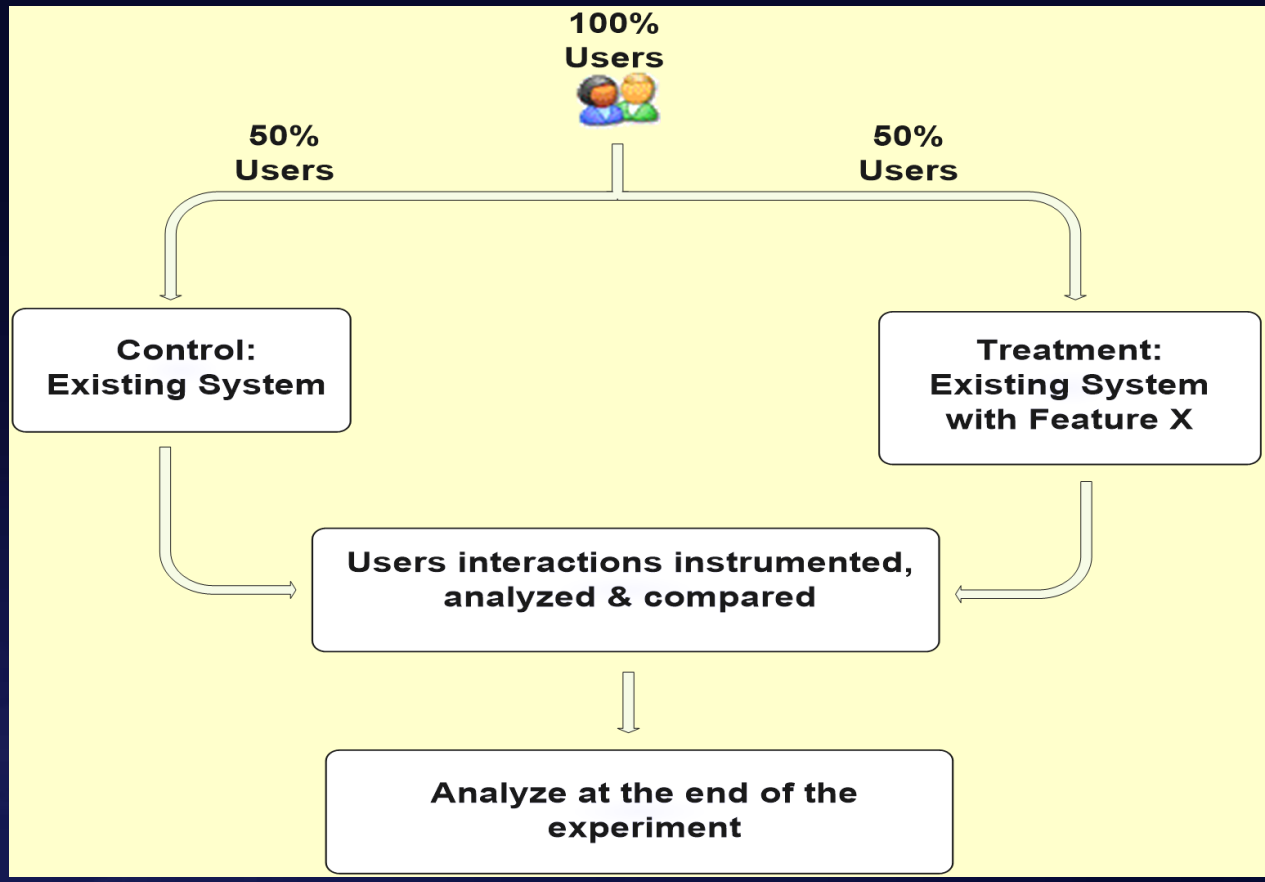
- President George Washington had a sore throat
  - Doctors extracted 82 ounces of blood over 10 hours (35% of his total blood), causing anemia and hypotension.
  - He died that night
- Bloodletting calms patients, but does not help most diseases
- Pierre Louis did an experiment in 1836
  - One of the 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

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)

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

Example: 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 B

- 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)

- 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

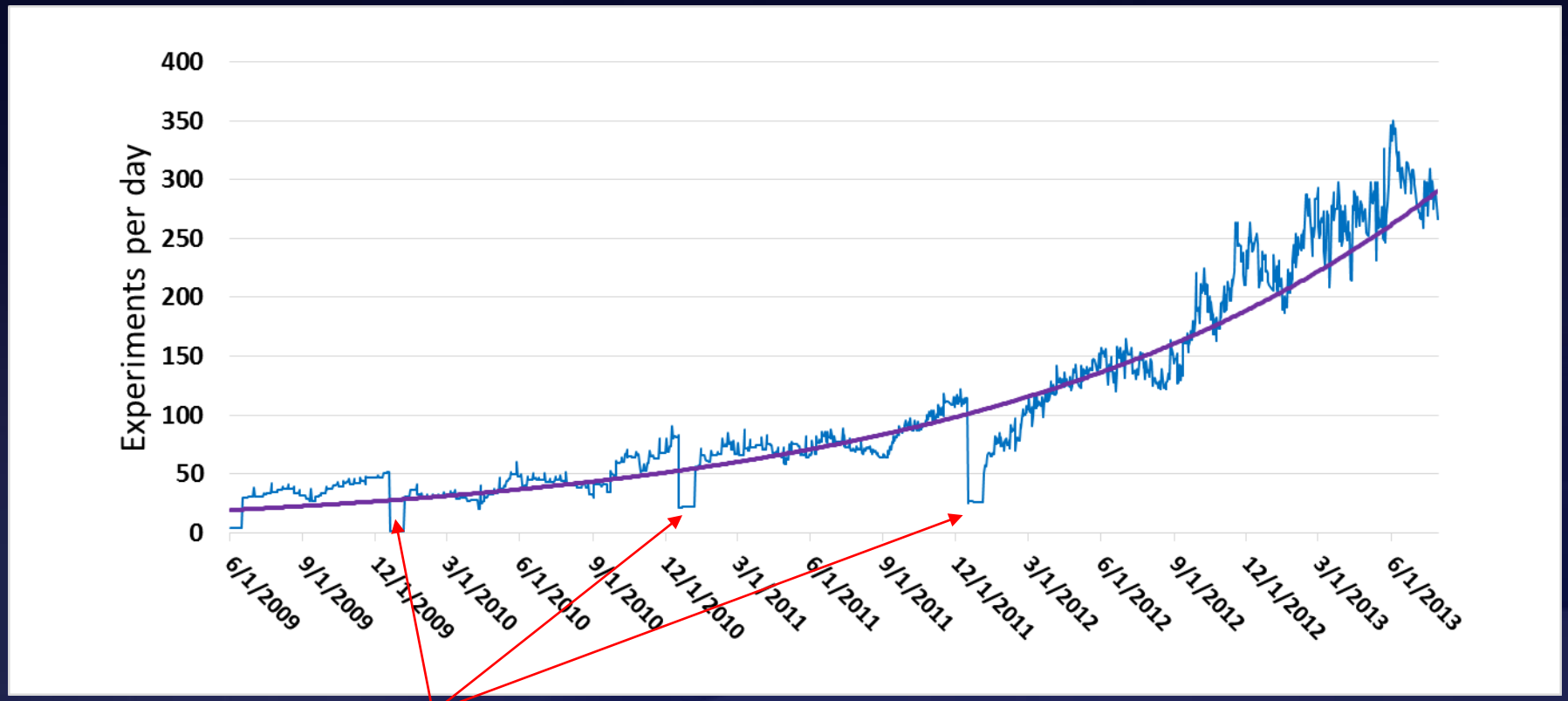
Bing Ads Example

- If you raised your left hand, you were wrong
- If you did not raise a hand, you were wrong
- Site links generate incremental revenue on the order of tens of millions of dollars annually for Bing

- The above change was costly to implement. We made two small changes to Bing, which took days to develop, each increased annual revenues by about $100 million
- (One was delayed by 6 months because it was not prioritized high, a prioritization mistake that cost $50M)

Scaling Experiments at Bing

- We now run over 250 concurrent experiments at Bing

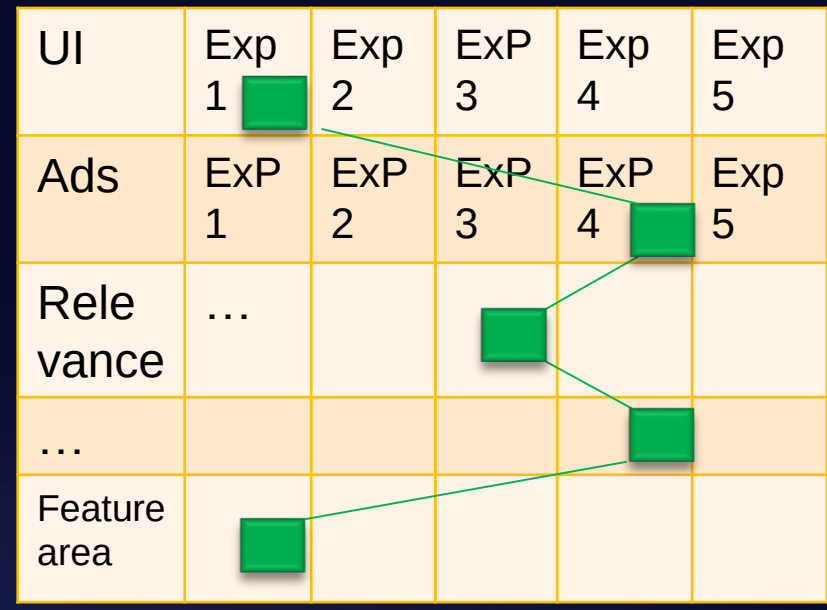


- We used to lockdown for Dec holidays. No more

Running Controlled Experiments at Scale

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 per line ^15 lines)
  - 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 evaluation criteria (KDD 2012 paper), is hard to move, so we're looking for small effects
  - Important experiments run on 10-20% of users



Cultural Lessons

- Ideas Funnel – we have too many ideas
  - Doug Hubbard's EVI: Expected Value of Information
    - Controlled experiments provide nearly perfect information
    - But ideas may be expensive to implement
  - Use cheaper means that have lower fidelity first: sketches, mockups, prototypes, surveys usability studies, tests against historical data
  - Observation: for ideas that are cheap to code, skip everything and just run the experiment
- Test everything
  - Code rewrites and platform changes frequently fail to be "equal"
  - Amazon's Gurupa app server lost 2% of revenue several times

Negative Experiments and Performance

- Should we ever knowingly degrade the customer experience?
- Yes, for a short-term experiment. Learn about tradeoffs
- Example: understand performance tradeoffs
  - Experiment slowed server by 100msec and 250msec
  - Multiple metrics were impacted
  - Simple rule-of-thumb:

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

Alerts, Aborts, Interactions

- With many experiments, we built an alerting system
  - Alert on user/business impact, not just stat-sig.
    Even if page-load-time is different with very low p-value, if the delta is a few msec, let it run (and optimize later)
  - Correct for multiple testing: we test whether to alert/abort multiple times, so the false positive rate will be high without corrections
  - Severe degradations cause automatic shutdowns
- We assume no interactions among different product areas
  - Prevention: we designed our system so that a use falls into one experimentation area (number line)
  - Run all-pairs test for interactions nightly

More in Paper

- Twyman's law: the "best" stories are usually wrong
- Be careful of incrementalism
- Complex MVT designs (multi-variable) are less useful in the online world: lots of simple experiments are better
- Architecture: our experimentation system
- Impact of experimentation system (30msec overhead)

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
   - Prepare to be humbled: data trumps intuition
2. Accelerate innovation through trustworthy experimentation
   - Make controlled experiments easy (cheap) to run
   - Surprising results often lead to useful insights (e.g., puzzling results)
3. Lessons shared
   - Cultural
   - Engineering
   - Trustworthiness