KDD2015

Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 years

Slides at http://bit.ly/KDD2015Kohavi, @RonnyK

Ron Kohavi, Distinguished Engineer, General Manager,
Analysis and Experimentation, Microsoft

Joint work with many members of the A&E/ExP platform team

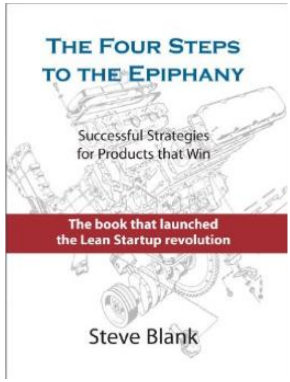
Agenda

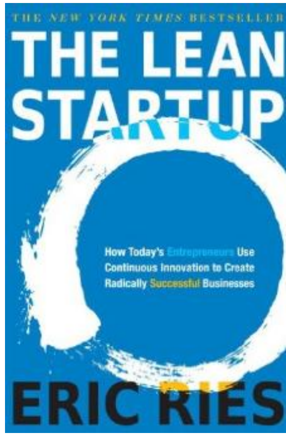
- Introduction to controlled experiments
- Four real examples: you're the decision maker
  Examples chosen to share lessons
- Lessons and pitfalls
- Cultural evolution towards a data-driven org

Motivation: Product Development

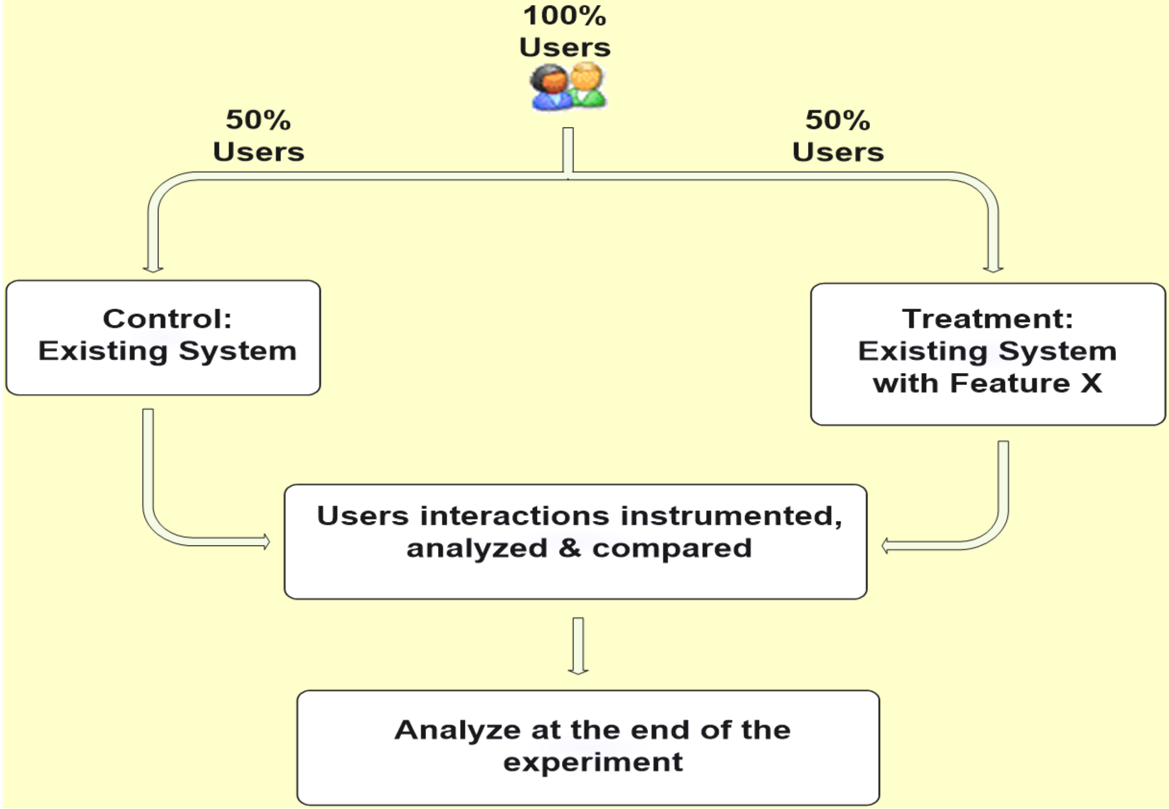
It doesn't matter how beautiful your theory is,
it doesn't matter how smart you are.
If it doesn't agree with experiment[s], it's wrong
-- Richard Feynman

- Classical software development: spec->dev->test->release
- Customer-driven Development: Build->Measure->Learn (continuous deployment cycles)
  - Described in Steve Blank's *The Four Steps to the Epiphany (2005)*
  - Popularized by Eric Ries' *The Lean Startup (2011)*
  - Build a Minimum Viable Product (MVP), or feature, cheaply
  - Evaluate it with real users in a **controlled experiment (e.g., A/B test)**
  - Iterate (or pivot) based on learnings
- Why use Customer-driven Development?
  Because we are poor at assessing the value of our ideas
  (more about this later in the talk)
- Why I love controlled experiments
  In many data mining scenarios, interesting discoveries are made and promptly ignored.
  In customer-driven development, the mining of data from the controlled experiments and insight generation is part of the critical path to the product release


THE FOUR STEPS TO THE EPIPHANY
Successful Strategies for Products that Win
The book that launched the Lean Startup revolution
Steve Blank


THE NEW YORK TIMES BESTSELLER
THE LEAN STARTUP
How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses
ERIC RIES

A/B/n Tests in One Slide


100% Users
50% Users
50% Users
Control: Existing System
Treatment: Existing System with Feature X
Users interactions instrumented, analyzed & compared
Analyze at the end of the experiment

- Concept is trivial
  - Randomly split traffic between two (or more) versions
    - A (Control)
    - B (Treatment)
  - Collect metrics of interest
  - Analyze
- Sample of real users
  - Not WEIRD (Western, Educated, Industrialized, Rich, and Democratic) like many academic research samples
- A/B test is the simplest controlled experiment
  - A/B/n refers to multiple treatments (often used and encouraged: try control + two or three treatments)
  - MVT refers to multivariable designs (rarely used by our teams)
- 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)

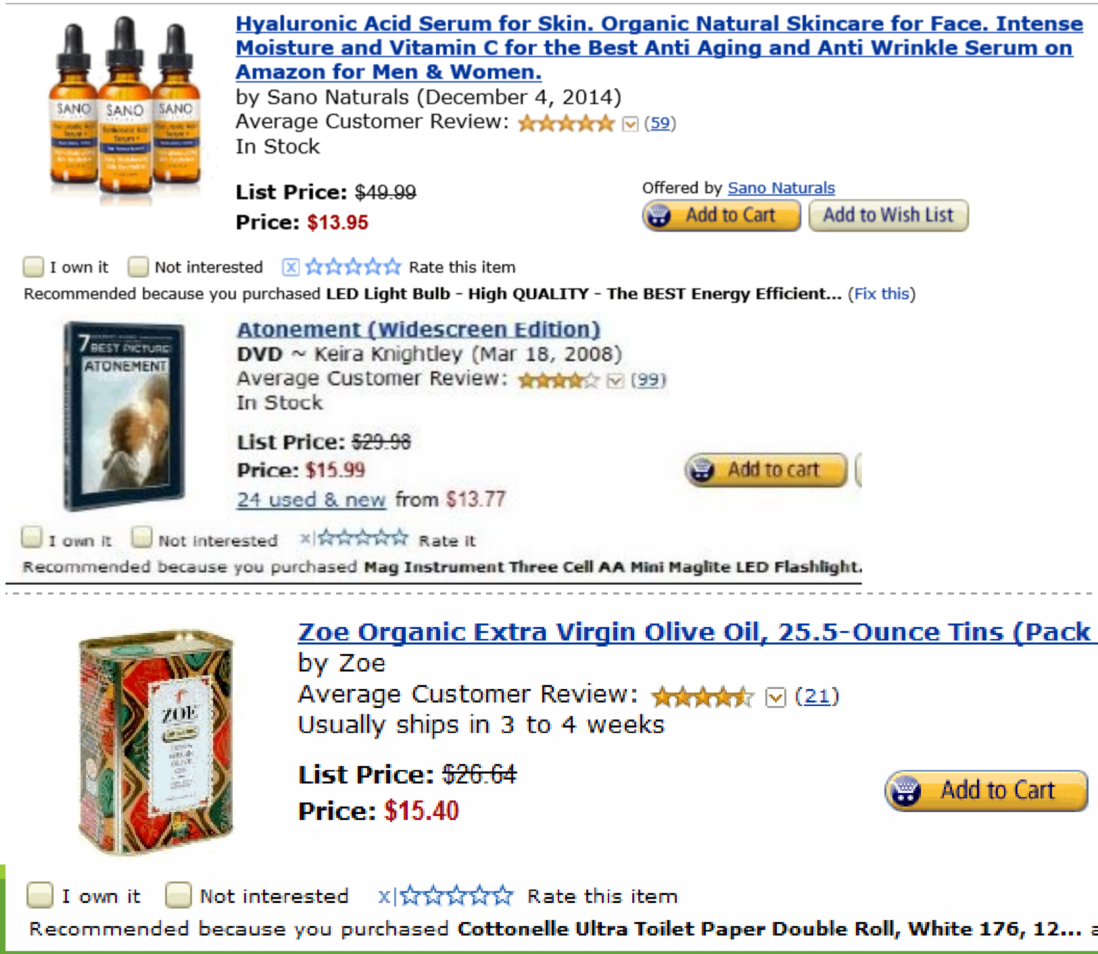
Personalized Correlated Recommendations

- 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 anti aging serum because
  you bought an LED light bulb
  (Maybe the wrinkles show?)

- 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.)



Hyaluronic Acid Serum for Skin. Organic Natural Skincare for Face. Intense Moisture and Vitamin C for the Best Anti Aging and Anti Wrinkle Serum on Amazon for Men & Women.
by Sano Naturals (December 4, 2014)
Average Customer Review: ★★★★★ (59)
In Stock

List Price: ~~$49.99~~
Price: $13.95

Offered by Sano Naturals

□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **LED Light Bulb - High QUALITY - The BEST Energy Efficient...** (Fix this)

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 x|☆☆☆☆☆ 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 x|☆☆☆☆☆ 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

The First Medical Controlled Experiment

- The earliest controlled experiment was a test for vegetarianism, suggested in the Old Testament's Book of Daniel

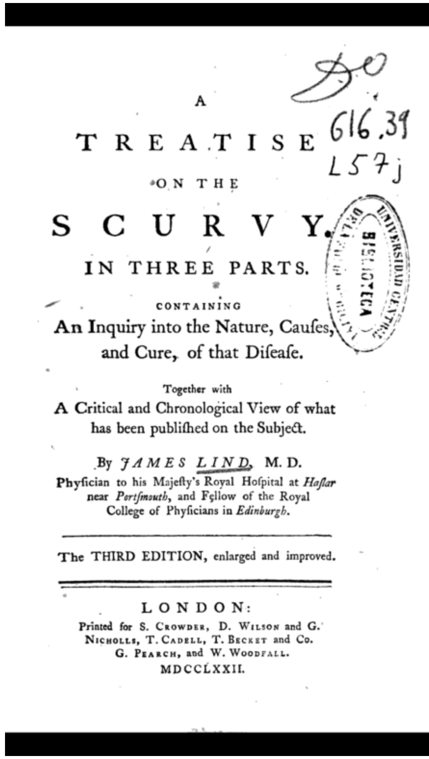
  > *Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see*

- First controlled experiment / randomized trial for medical purposes
  - Scurvy is a disease that results from vitamin C deficiency
  - It killed over 100,000 people in the 16th-18th centuries, mostly sailors
  - Lord Anson’s circumnavigation voyage from 1740 to 1744 started with 1,800 sailors and only about 200 returned; most died from scurvy
  - 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

- Amazing scientific triumph, right? Wrong

The First Medical Controlled Experiment

- Like most stories, the discovery is highly exaggerated
  - The experiment was done on 12 sailors split into 6 pairs
  - Each pair got a different treatment: cider, elixir vitriol, vinegar, sea-water, nutmeg
  - Two sailors were given two oranges and one lemon per day and recovered
  - Lind didn't understand the reason and tried treating Scurvy with concentrated lemon juice called "rob."
    The lemon juice was concentrated by heating it, which destroyed the vitamin C.
  - Working at Haslar hospital, he attended to 300-400 scurvy patients a day for 5 years
  - In his 559 pages massive book A Treatise on the Scurvy, there are two pages about this experiment. Everything else is about other treatments, from Peruvian bark to bloodletting to rubbing the belly with warm olive oil



A

TREATISE

ON THE

SCURVY.

IN THREE PARTS.

CONTAINING

An Inquiry into the Nature, Cauſes,
and Cure, of that Diſeaſe.

Together with
A Critical and Chronological View of what
has been publiſhed on the Subject.

By *JAMES* *LIND*, M. D.
Phyſician to his Majeſty's Royal Hoſpital at *Haſlar*
near *Portſmouth*, and Fellow of the Royal
College of Phyſicians in *Edinburgh*.

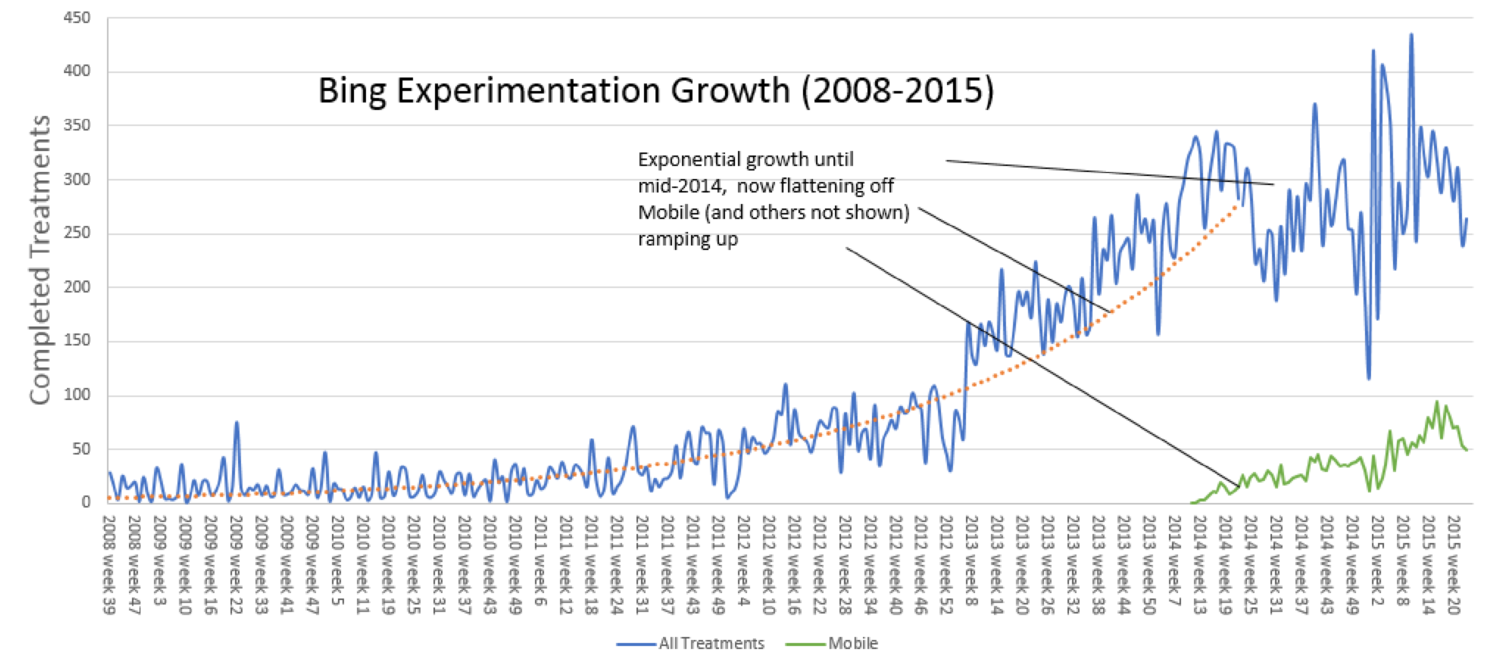
The THIRD EDITION, enlarged and improved.

LONDON:
Printed for S. Crowder, D. Wilson and G.
Nicholls, T. Cadell, T. Becket and Co.
G. Pearch, and W. Woodfall.
MDCCLXXII.

Lesson: Even when you have a winner, the reasons are often not understood.
Controlled experiments tell you which variant won, not why.

Experimentation at Scale

- I’ve been fortunate to work at an organization that values being data-driven
- We finish about ~300 experiment treatments at Bing every week.
  (Since most experiments run for a week or two, there are a similar number of concurrent treatments running. These are “real” useful treatments, not 3x10x10 MVT = 300)
- See Google’s KDD 2010 paper on Overlapping Experiment Infrastructure and
  Our KDD 2013 paper on challenges of scaling experimentation: http://bit.ly/ExPScale
- Each variant is exposed to between 100K and millions of users, sometimes tens of millions
- 90% of eligible users are in experiments (10% are a global holdout changed once a year)
- There is no single Bing. Since a user is exposed to 15 concurrent experiments, they get one of 5^15 = 30 billion variants
  (debugging takes a new meaning).
- Until 2014, the system was limiting usage as it scaled.
  Now the limits come from engineers’ ability to code new ideas



Exponential growth until mid-2014, now flattening off
Mobile (and others not shown) ramping up

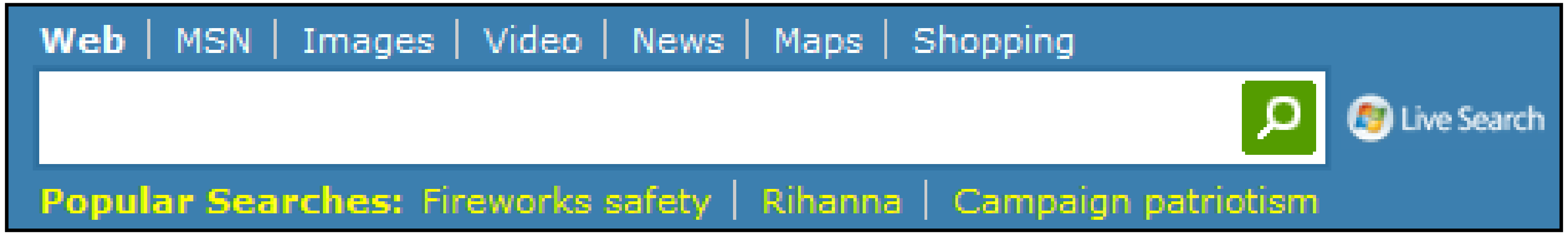
Real Examples

- Four experiments that ran at Microsoft
- Each provides interesting lessons
- All had enough users for statistical validity
- For each experiment, we provide the OEC, the Overall Evaluation Criterion
  - This is the criterion to determine which variant is the winner
- 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
- Since there are 3 choices for each question, random guessing implies
  100%/3^4 = 1.2% will get all four questions right.
  Let’s see how much better than random we can get in this room

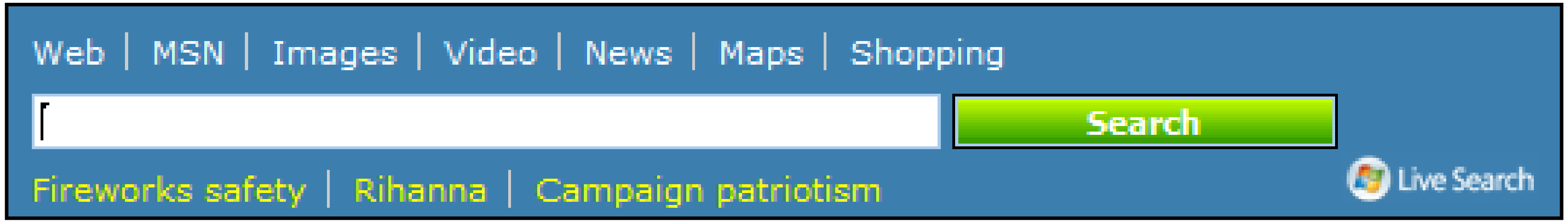
Example 1: MSN Home Page Search Box

- OEC: Clickthrough rate for Search box and popular searches

A


Web | MSN | Images | Video | News | Maps | Shopping
Live Search
Popular Searches: Fireworks safety | Rihanna | Campaign patriotism

B


Web | MSN | Images | Video | News | Maps | Shopping
Search
Fireworks safety | Rihanna | Campaign patriotism
Live Search

Differences: A has taller search box (overall size is the same),
has magnifying glass icon, “popular searches”
B has big search button, provides popular searches without calling them out

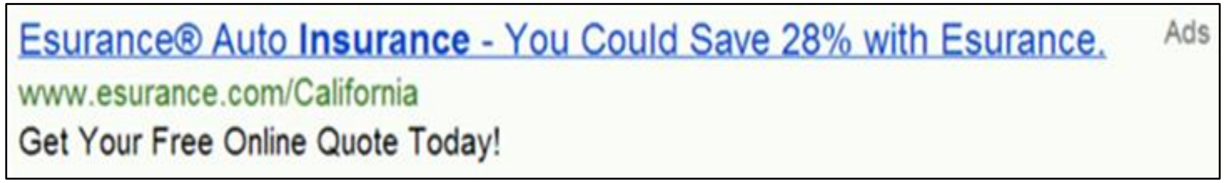
- Raise your left hand if you think A Wins (top)
- Raise your right hand if you think B Wins (bottom)
- Don’t raise your hand if they are the about the same

MSN Home Page Search Box

[You can’t cheat by looking for the answers here]

Example 2: 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



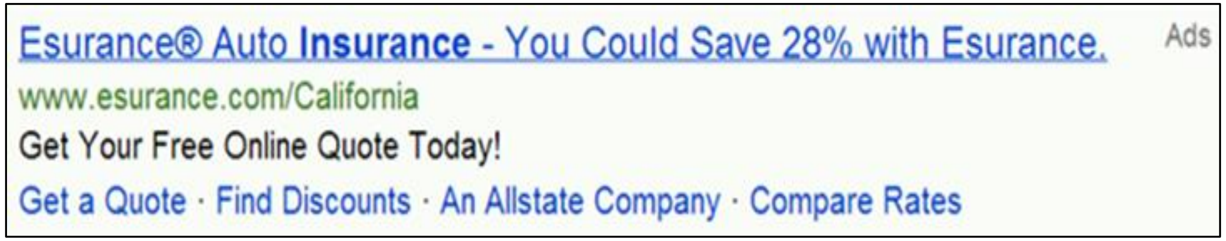
Esurance® Auto **Insurance** - You Could Save 28% with Esurance.

Ads

www.esurance.com/California

Get Your Free Online Quote Today!

A



Esurance® Auto **Insurance** - You Could Save 28% with Esurance.

Ads

www.esurance.com/California

Get Your Free Online Quote Today!

Get a Quote · Find Discounts · An Allstate Company · Compare Rates

B

- Pro adding: 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)

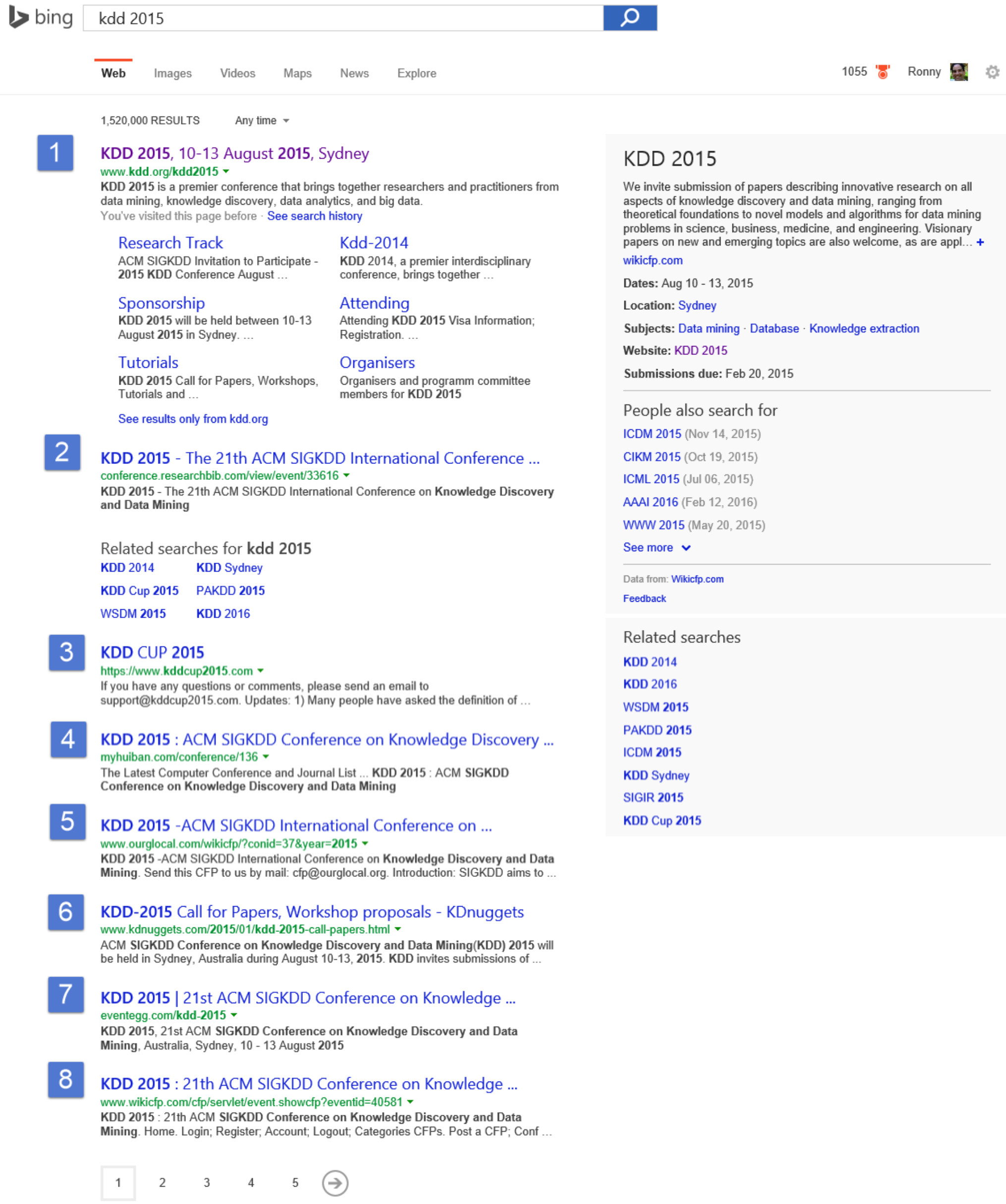
- Raise your left hand if you think A Wins (left)
- Raise your right hand if you think B Wins (right)
- Don’t raise your hand if they are the about the same

Bing Ads with Site Links

[You can’t cheat by looking for the answers here]

Example 3:
SERP Truncation

- SERP is a Search Engine Result Page (shown on the right for the query KDD 2015)
- OEC: Clickthrough Rate on 1st SERP per query (ignore issues with click/back, page 2, etc.)
- Version A: show 10 algorithmic results
- Version B: show 8 algorithmic results by removing the last two results
- All else same: task pane, ads, related searches, etc.



1,520,000 RESULTS Any time

1 **KDD 2015, 10-13 August 2015, Sydney**
www.kdd.org/kdd2015
KDD 2015 is a premier conference that brings together researchers and practitioners from data mining, knowledge discovery, data analytics, and big data.
You've visited this page before · See search history

Research Track
ACM SIGKDD Invitation to Participate - 2015 KDD Conference August …

Kdd-2014
KDD 2014, a premier interdisciplinary conference, brings together …

Sponsorship
KDD 2015 will be held between 10-13 August 2015 in Sydney. …

Attending
Attending KDD 2015 Visa Information; Registration. …

Tutorials
KDD 2015 Call for Papers, Workshops, Tutorials and …

Organisers
Organisers and programm committee members for KDD 2015

See results only from kdd.org

2 **KDD 2015 - The 21th ACM SIGKDD International Conference ...**
conference.researchbib.com/view/event/33616
KDD 2015 - The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Related searches for kdd 2015
KDD 2014 · KDD Sydney · KDD Cup 2015 · PAKDD 2015 · WSDM 2015 · KDD 2016

3 **KDD CUP 2015**
https://www.kddcup2015.com
If you have any questions or comments, please send an email to support@kddcup2015.com. Updates: 1) Many people have asked the definition of …

4 **KDD 2015 : ACM SIGKDD Conference on Knowledge Discovery ...**
myhuiban.com/conference/136
The Latest Computer Conference and Journal List ... KDD 2015 : ACM SIGKDD Conference on Knowledge Discovery and Data Mining

5 **KDD 2015 -ACM SIGKDD International Conference on ...**
www.ourglocal.com/wikicfp/?conid=37&year=2015
KDD 2015 -ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Send this CFP to us by mail: cfp@ourglocal.org. Introduction: SIGKDD aims to ...

6 **KDD-2015 Call for Papers, Workshop proposals - KDnuggets**
www.kdnuggets.com/2015/01/kdd-2015-call-papers.html
ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD) 2015 will be held in Sydney, Australia during August 10-13, 2015. KDD invites submissions of ...

7 **KDD 2015 | 21st ACM SIGKDD Conference on Knowledge ...**
eventegg.com/kdd-2015
KDD 2015, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Australia, Sydney, 10 - 13 August 2015

8 **KDD 2015 : 21th ACM SIGKDD Conference on Knowledge ...**
www.wikicfp.com/cfp/servlet/event.showcfp?eventid=40581
KDD 2015 : 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Home. Login; Register; Account; Logout; Categories CFPs. Post a CFP; Conf ...

1 2 3 4 5

KDD 2015

We invite submission of papers describing innovative research on all aspects of knowledge discovery and data mining, ranging from theoretical foundations to novel models and algorithms for data mining problems in science, business, medicine, and engineering. Visionary papers on new and emerging topics are also welcome, as are appl… +
wikicfp.com

Dates: Aug 10 - 13, 2015
Location: Sydney
Subjects: Data mining · Database · Knowledge extraction
Website: KDD 2015
Submissions due: Feb 20, 2015

People also search for
ICDM 2015 (Nov 14, 2015)
CIKM 2015 (Oct 19, 2015)
ICML 2015 (Jul 06, 2015)
AAAI 2016 (Feb 12, 2016)
WWW 2015 (May 20, 2015)
See more

Data from: Wikicfp.com
Feedback

Related searches
KDD 2014
KDD 2016
WSDM 2015
PAKDD 2015
ICDM 2015
KDD Sydney
SIGIR 2015
KDD Cup 2015

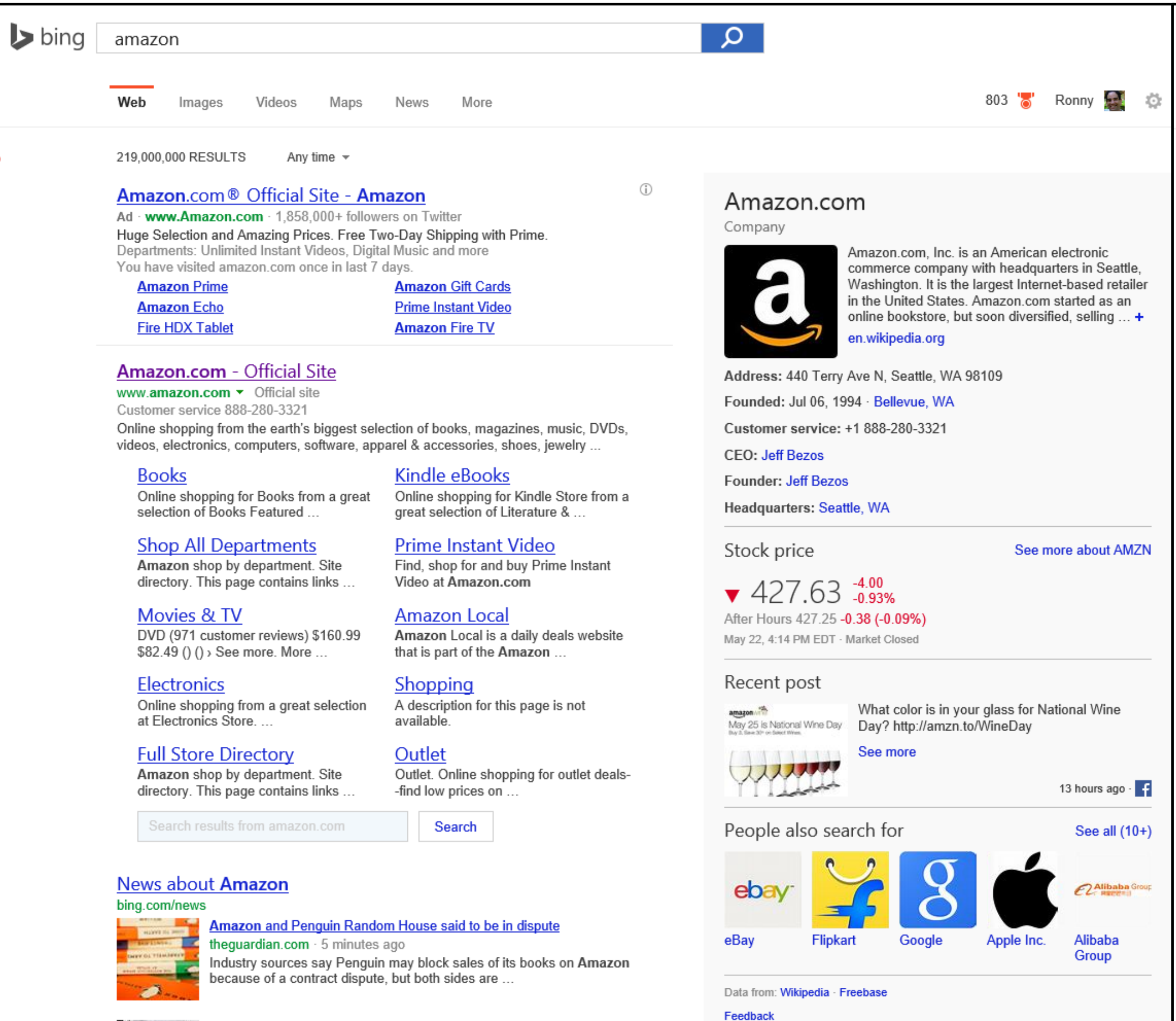
- Raise your left hand if you think A Wins (10 results)
- Raise your right hand if you think B Wins (8 results)
- Don't raise your hand if they are the about the same

SERP Truncation

[You can’t cheat by looking for the answers here]

Example 4: Underlining Links

- Does underlining increase or decrease clickthrough-rate?



bing amazon

Web Images Videos Maps News More 803 Ronny

219,000,000 RESULTS Any time

Amazon.com® Official Site - Amazon
Ad · www.Amazon.com · 1,858,000+ followers on Twitter
Huge Selection and Amazing Prices. Free Two-Day Shipping with Prime.
Departments: Unlimited Instant Videos, Digital Music and more
You have visited amazon.com once in last 7 days.

Amazon Prime
Amazon Echo
Fire HDX Tablet
Amazon Gift Cards
Prime Instant Video
Amazon Fire TV

Amazon.com - Official Site
www.amazon.com · Official site
Customer service 888-280-3321
Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry ...

Books
Online shopping for Books from a great selection of Books Featured ...

Kindle eBooks
Online shopping for Kindle Store from a great selection of Literature & ...

Shop All Departments
Amazon shop by department. Site directory. This page contains links ...

Prime Instant Video
Find, shop for and buy Prime Instant Video at Amazon.com

Movies & TV
DVD (971 customer reviews) $160.99 $82.49 () () › See more. More ...

Amazon Local
Amazon Local is a daily deals website that is part of the Amazon ...

Electronics
Online shopping from a great selection at Electronics Store. ...

Shopping
A description for this page is not available.

Full Store Directory
Amazon shop by department. Site directory. This page contains links ...

Outlet
Outlet. Online shopping for outlet deals--find low prices on ...

Search results from amazon.com Search

News about Amazon
bing.com/news
Amazon and Penguin Random House said to be in dispute
theguardian.com · 5 minutes ago
Industry sources say Penguin may block sales of its books on Amazon because of a contract dispute, but both sides are ...

Amazon.com
Company
Amazon.com, Inc. is an American electronic commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailer in the United States. Amazon.com started as an online bookstore, but soon diversified, selling ... +
en.wikipedia.org

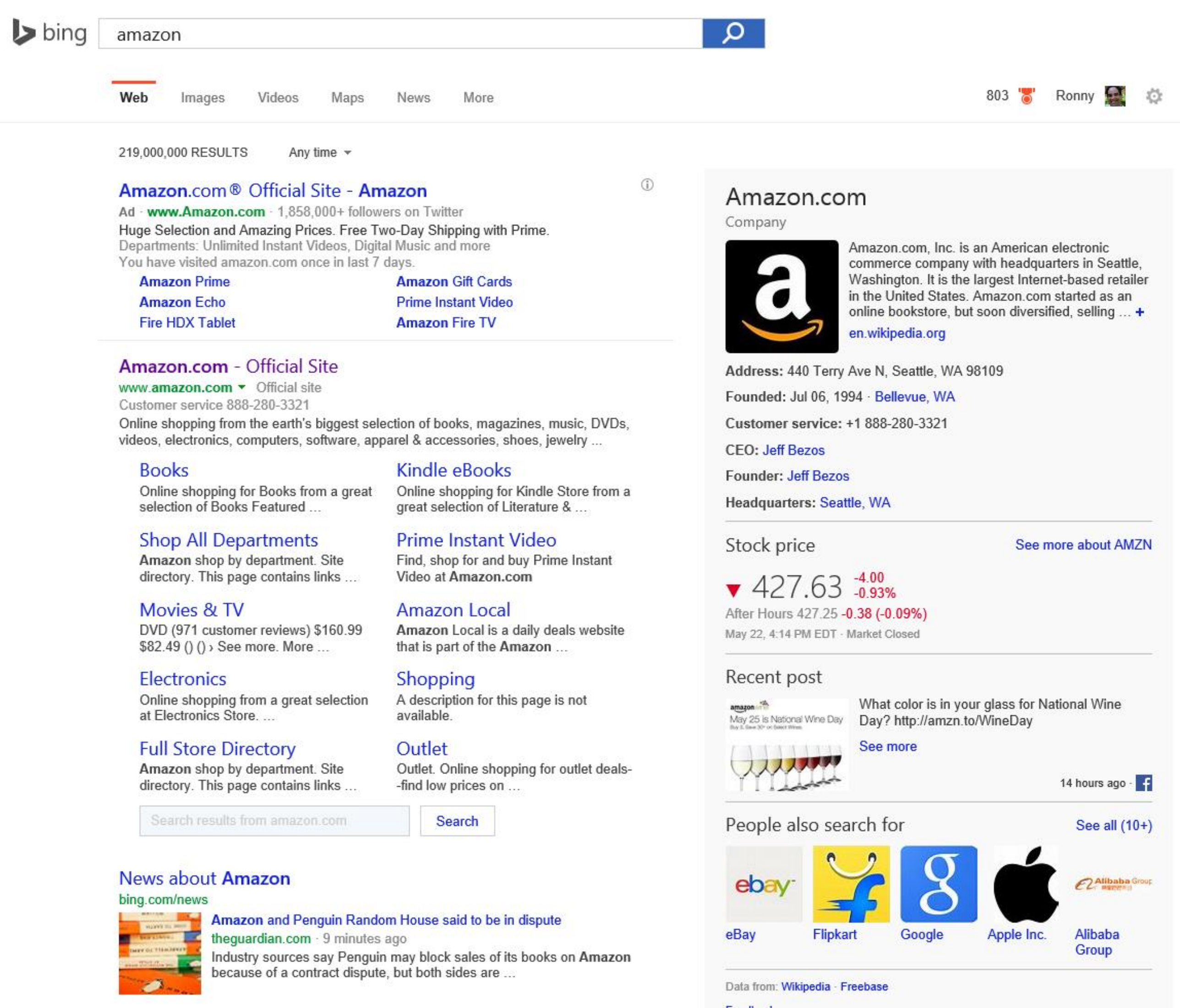
Address: 440 Terry Ave N, Seattle, WA 98109
Founded: Jul 06, 1994 · Bellevue, WA
Customer service: +1 888-280-3321
CEO: Jeff Bezos
Founder: Jeff Bezos
Headquarters: Seattle, WA

Stock price See more about AMZN
▼ 427.63 -4.00 -0.93%
After Hours 427.25 -0.38 (-0.09%)
May 22, 4:14 PM EDT · Market Closed

Recent post
What color is in your glass for National Wine Day? http://amzn.to/WineDay
See more
13 hours ago · f

People also search for See all (10+)
eBay Flipkart Google Apple Inc. Alibaba Group

Data from: Wikipedia · Freebase
Feedback



bing amazon

Web Images Videos Maps News More 803 Ronny

219,000,000 RESULTS Any time

Amazon.com® Official Site - Amazon
Ad · www.Amazon.com · 1,858,000+ followers on Twitter
Huge Selection and Amazing Prices. Free Two-Day Shipping with Prime.
Departments: Unlimited Instant Videos, Digital Music and more
You have visited amazon.com once in last 7 days.

Amazon Prime
Amazon Echo
Fire HDX Tablet
Amazon Gift Cards
Prime Instant Video
Amazon Fire TV

Amazon.com - Official Site
www.amazon.com · Official site
Customer service 888-280-3321
Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry ...

Books
Online shopping for Books from a great selection of Books Featured ...

Kindle eBooks
Online shopping for Kindle Store from a great selection of Literature & ...

Shop All Departments
Amazon shop by department. Site directory. This page contains links ...

Prime Instant Video
Find, shop for and buy Prime Instant Video at Amazon.com

Movies & TV
DVD (971 customer reviews) $160.99 $82.49 () () › See more. More ...

Amazon Local
Amazon Local is a daily deals website that is part of the Amazon ...

Electronics
Online shopping from a great selection at Electronics Store. ...

Shopping
A description for this page is not available.

Full Store Directory
Amazon shop by department. Site directory. This page contains links ...

Outlet
Outlet. Online shopping for outlet deals--find low prices on ...

Search results from amazon.com Search

News about Amazon
bing.com/news
Amazon and Penguin Random House said to be in dispute
theguardian.com · 9 minutes ago
Industry sources say Penguin may block sales of its books on Amazon because of a contract dispute, but both sides are ...

Amazon.com
Company
Amazon.com, Inc. is an American electronic commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailer in the United States. Amazon.com started as an online bookstore, but soon diversified, selling ... +
en.wikipedia.org

Address: 440 Terry Ave N, Seattle, WA 98109
Founded: Jul 06, 1994 · Bellevue, WA
Customer service: +1 888-280-3321
CEO: Jeff Bezos
Founder: Jeff Bezos
Headquarters: Seattle, WA

Stock price See more about AMZN
▼ 427.63 -4.00 -0.93%
After Hours 427.25 -0.38 (-0.09%)
May 22, 4:14 PM EDT · Market Closed

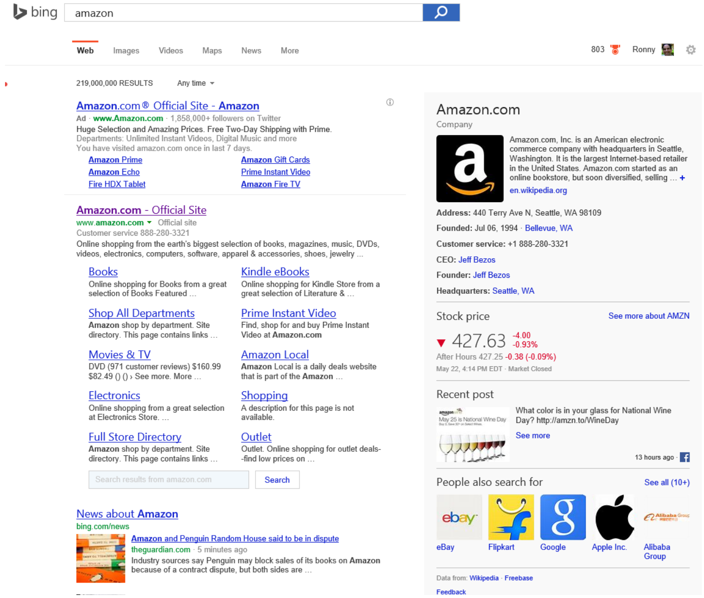
Recent post
What color is in your glass for National Wine Day? http://amzn.to/WineDay
See more
14 hours ago · f

People also search for See all (10+)
eBay Flipkart Google Apple Inc. Alibaba Group

Data from: Wikipedia · Freebase

Example 4: Underlining Links

- Does underlining increase or decrease clickthrough-rate?
- OEC: Clickthrough Rate on 1st SERP per query



bing amazon

Web Images Videos Maps News More

803 Ronny

219,000,000 RESULTS Any time

Amazon.com® Official Site - Amazon
Ad · www.Amazon.com · 1,858,000+ followers on Twitter
Huge Selection and Amazing Prices. Free Two-Day Shipping with Prime.
Departments: Unlimited Instant Videos, Digital Music and more
You have visited amazon.com once in last 7 days.

Amazon Prime · Amazon Gift Cards · Amazon Echo · Prime Instant Video · Fire HDX Tablet · Amazon Fire TV

Amazon.com - Official Site
www.amazon.com · Official site
Customer service 888-280-3321
Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry ...

Books
Online shopping for Books from a great selection of Books Featured ...

Kindle eBooks
Online shopping for Kindle Store from a great selection of Literature & ...

Shop All Departments
Amazon shop by department. Site directory. This page contains links ...

Prime Instant Video
Find, shop for and buy Prime Instant Video at Amazon.com

Movies & TV
DVD (971 customer reviews) $160.99 $82.49 () () › See more. More ...

Amazon Local
Amazon Local is a daily deals website that is part of the Amazon ...

Electronics
Online shopping from a great selection at Electronics Store. ...

Shopping
A description for this page is not available.

Full Store Directory
Amazon shop by department. Site directory. This page contains links ...

Outlet
Outlet. Online shopping for outlet deals--find low prices on ...

Search results from amazon.com Search

News about Amazon
bing.com/news
Amazon and Penguin Random House said to be in dispute
theguardian.com · 5 minutes ago
Industry sources say Penguin may block sales of its books on Amazon because of a contract dispute, but both sides are ...

Amazon.com
Company
Amazon.com, Inc. is an American electronic commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailer in the United States. Amazon.com started as an online bookstore, but soon diversified, selling ... +
en.wikipedia.org

Address: 440 Terry Ave N, Seattle, WA 98109
Founded: Jul 06, 1994 · Bellevue, WA
Customer service: +1 888-280-3321
CEO: Jeff Bezos
Founder: Jeff Bezos
Headquarters: Seattle, WA

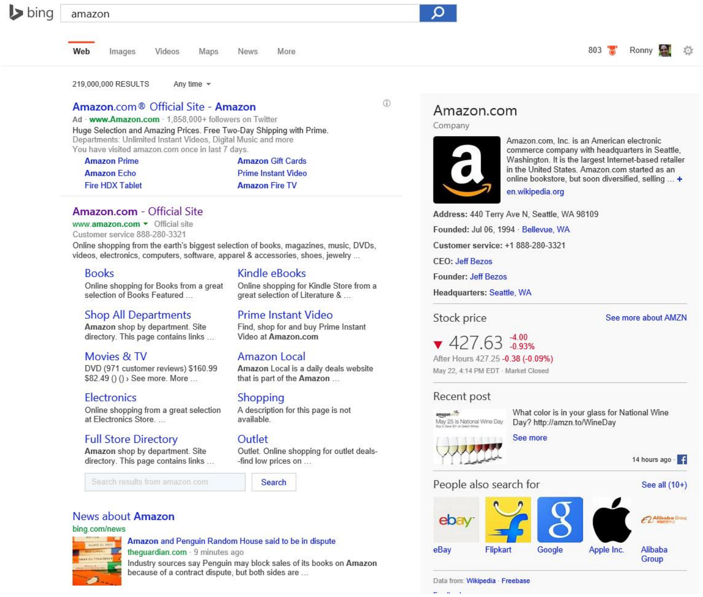
Stock price See more about AMZN
▼ 427.63 -4.00 -0.93%
After Hours 427.25 -0.38 (-0.09%)
May 22, 4:14 PM EDT · Market Closed

Recent post
What color is in your glass for National Wine Day? http://amzn.to/WineDay
See more
13 hours ago

People also search for See all (10+)
eBay · Flipkart · Google · Apple Inc. · Alibaba Group

Data from: Wikipedia · Freebase
Feedback

A



bing amazon

Web Images Videos Maps News More

803 Ronny

219,000,000 RESULTS Any time

Amazon.com® Official Site - Amazon
Ad · www.Amazon.com · 1,858,000+ followers on Twitter
Huge Selection and Amazing Prices. Free Two-Day Shipping with Prime.
Departments: Unlimited Instant Videos, Digital Music and more
You have visited amazon.com once in last 7 days.

Amazon Prime · Amazon Gift Cards · Amazon Echo · Prime Instant Video · Fire HDX Tablet · Amazon Fire TV

Amazon.com - Official Site
www.amazon.com · Official site
Customer service 888-280-3321
Online shopping from the earth's biggest selection of books, magazines, music, DVDs, videos, electronics, computers, software, apparel & accessories, shoes, jewelry ...

Books
Online shopping for Books from a great selection of Books Featured ...

Kindle eBooks
Online shopping for Kindle Store from a great selection of Literature & ...

Shop All Departments
Amazon shop by department. Site directory. This page contains links ...

Prime Instant Video
Find, shop for and buy Prime Instant Video at Amazon.com

Movies & TV
DVD (971 customer reviews) $160.99 $82.49 () () › See more. More ...

Amazon Local
Amazon Local is a daily deals website that is part of the Amazon ...

Electronics
Online shopping from a great selection at Electronics Store. ...

Shopping
A description for this page is not available.

Full Store Directory
Amazon shop by department. Site directory. This page contains links ...

Outlet
Outlet. Online shopping for outlet deals--find low prices on ...

Search results from amazon.com Search

News about Amazon
bing.com/news
Amazon and Penguin Random House said to be in dispute
theguardian.com · 9 minutes ago
Industry sources say Penguin may block sales of its books on Amazon because of a contract dispute, but both sides are ...

Amazon.com
Company
Amazon.com, Inc. is an American electronic commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailer in the United States. Amazon.com started as an online bookstore, but soon diversified, selling ... +
en.wikipedia.org

Address: 440 Terry Ave N, Seattle, WA 98109
Founded: Jul 06, 1994 · Bellevue, WA
Customer service: +1 888-280-3321
CEO: Jeff Bezos
Founder: Jeff Bezos
Headquarters: Seattle, WA

Stock price See more about AMZN
▼ 427.63 -4.00 -0.93%
After Hours 427.25 -0.38 (-0.09%)
May 22, 4:14 PM EDT · Market Closed

Recent post
What color is in your glass for National Wine Day? http://amzn.to/WineDay
See more
14 hours ago

People also search for See all (10+)
eBay · Flipkart · Google · Apple Inc. · Alibaba Group

Data from: Wikipedia · Freebase

B

- Raise your left hand if you think A Wins (left, with underlines)
- Raise your right hand if you think B Wins (right, without underlines)
- Don’t raise your hand if they are the about the same

Underlines

[You can’t cheat by looking for the answers here]

Agenda

- Introduction to controlled experiments
- Four real examples: you’re the decision maker
- Lessons and pitfalls
- Cultural evolution towards a data-driven org

Hard to Assess the Value of Ideas: Data Trumps Intuition

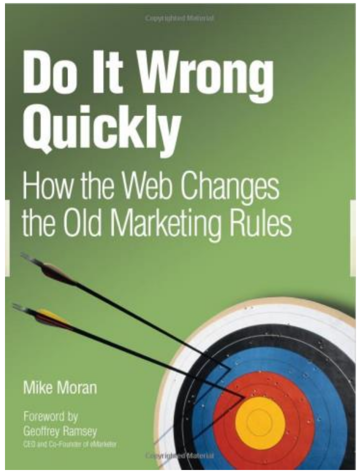
- Features are built because teams believe they are useful.
  But most experiments show that features fail to move the metrics they were designed to improve
- We joke that our job is to tell clients that their new baby is ugly
- Based on experiments at Microsoft (paper)
  - 1/3 of ideas were positive ideas and statistically significant
  - 1/3 of ideas were flat: no statistically significant difference
  - 1/3 of ideas were negative and statistically significant
- At Bing, the success rate is lower
- The low success rate has been documented many times across multiple companies

If you start running controlled experiments, you will be humbled!

Key Lesson Given the Success Rate

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
  - *If you're not prepared to be wrong, you'll never come up with anything original* – Sir Ken Robinson, TED 2006 (#1 TED talk)


Do It Wrong Quickly
How the Web Changes the Old Marketing Rules
Mike Moran
Foreword by Geoffrey Ramsey

Twyman’s Law

Any figure that looks interesting or different is usually wrong

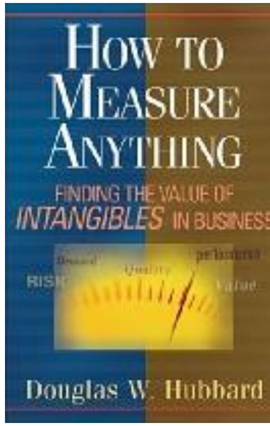
- If something is “amazing,” find the flaw!
- Examples
  - If you have a mandatory birth date field and people think it’s unnecessary, you’ll find lots of 11/11/11 or 01/01/01
  - If you have an optional drop down, do not default to the first alphabetical entry, or you’ll have lots of: jobs = Astronaut
  - Traffic to web sites doubled between 1-2AM November 2, 2014 for many sites, relative to the same hour a week prior. Why?
- If you see a massive improvement to your OEC, call Twyman’s law and find the flaw. Triple check things before you celebrate.

The OEC

- If you remember one thing from this talk, remember this point

Agree early on what you are optimizing

- OEC = Overall Evaluation Criterion
  - *Lean Analytics* call it OMTM: One Metric That Matters (OMTM).
  - Getting agreement on the OEC in the org is a huge step forward
  - Suggestion: optimize for **customer lifetime value**, not short-term revenue. Ex: Amazon e-mail
    Look for success indicators/leading indicators, avoid lagging indicators and vanity metrics.
  - Read Doug Hubbard's *How to Measure Anything*
  - Funnels use Pirate metrics: acquisition, activation, retention, revenue, and referral — AARRR
  - Criterion could be weighted sum of factors, such as
    - Conversion/action, Time to action, Visit frequency
  - Use a few KEY metrics. Beware of the Otis Redding problem (Pfeffer & Sutton)
    "I can't do what ten people tell me to do, so I guess I'll remain the same."
  - Report many other metrics for diagnostics, i.e., to understand the why the OEC changed, and raise new hypotheses to iterate. For example, clickthrough by area of page
  - See KDD 2015 papers by Henning etal. on Focusing on the Long-term: It's Good for Users and Business, and Kirill etal. on Extreme States Distribution Decomposition Method



OEC for Search

- KDD 2012 Paper: *Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained*
- 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
  - Degraded (algorithmic) search results cause users to search more to complete their task, and ads appear more relevant
  - Distinct queries went up over 10%, and revenue went up over 30%
- What metrics should be in the OEC for a search engine?

Puzzle Explained

- Analyzing queries per month, we have

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

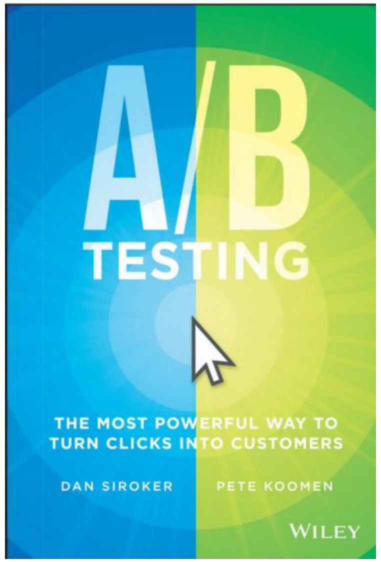
where a session begins with a query and ends with 30-minutes of inactivity. (Ideally, we would look at tasks, not sessions).

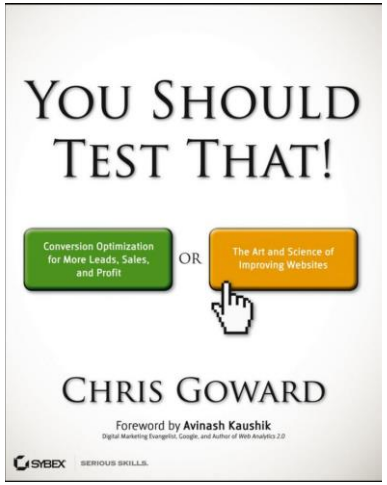
- Key observation: we want users to find answers and complete tasks quickly, so queries/session should be smaller
- In a controlled experiment, the variants get (approximately) the same number of users by design, so the last term is about equal
- The OEC should therefore include the middle term: sessions/user

Get the Stats Right

Getting numbers is easy;
getting numbers you can trust is hard

- Two very good books on A/B testing (A/B Testing from Optimizely founders Dan Siroker and Peter Koomen; and You Should Test That by WiderFunnel's CEO Chris Goward) get the stats wrong (see Amazon reviews).
- Optimizely recently updated their stats in the product to correct for this
- Best techniques to find issues: run A/A tests
  - Like an A/B test, but both variants are exactly the same
  - Are users split according to the planned percentages?
  - Is the data collected matching the system of record?
  - Are the results showing non-significant results 95% of the time?




YOU SHOULD TEST THAT!
Conversion Optimization for More Leads, Sales, and Profit
OR
The Art and Science of Improving Websites
CHRIS GOWARD
Foreword by Avinash Kaushik
SYBEX
SERIOUS SKILLS.

Experiment on (Almost) All Users

Run Experiments on Large Percentages: 50/50%

- To detect an effect, you need to expose a certain number of users to the treatment (based on statistical power calculations).
  We usually want more users than are available
- Larger user samples increase sensitivity (lower p-values for same effect size) and allows evaluating segments
- 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%.
  At 20%, we could only run 5 disjoint variants, which is why users are in multiple experiments (equivalent to full-factorial designs)

Reduce the Variance of Metrics

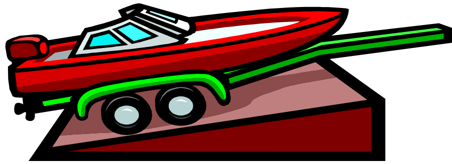
- At Bing, 1% change in the revenue/user in an experiment = tens of millions/year
  If we run an experiment that can only detect 1% change, we might lose tens of millions without realizing it! Before shipping a feature, a sufficiently large experiment must run
- To improve sensitivity of experiments, you can either get more users or reduce the variance of metrics
- Good techniques for reducing the variance of metrics
  - Triggering: analyze only users who were actually exposed to change.
    Important sanity check: the complement users should look like an A/A experiment
  - Use lower-variance metrics (e.g., trim revenue, or look at Boolean metrics like conversion rate vs. revenue; see paper Section 3.2.1)
  - Use pre-experiment period: before the experiment started, there was no difference between the control and treatment. We can use the deltas in the pre-experiment period to reduce the variance.
    Nice trick called CUPED
  - Rejecting randomizations that fail the pre-experiment A/A test (see paper Section 3.5).
    When you start an experiment, you generate a random hash function to distribute users to variants.
    If you look back in time (before the experiment started), does the pre-experiment split look like an A/A?
    In the Mastering ‘Metrics book, they call it “checking for balance.”
    We automated it and try multiple “seeds” to optimize the balance in the randomization

Online is Different than Offline

- The theory of controlled experiments was formalized by Sir Ronald A. Fisher’s in the 1920s, but as the saying goes

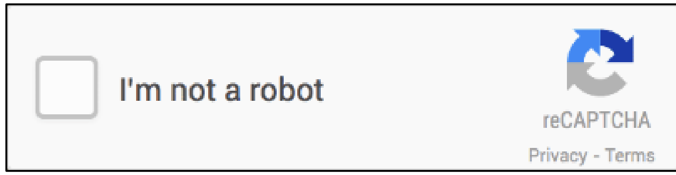
  The difference between theory and practice is larger in practice than the difference between theory and practice in theory

- Key differences from Offline to Online
  - Ramp-up: tens of experiments start every day at Bing, but initial code tends to be buggy.
    - Experiments start small, the system looks for egregious effects, and if there’s a problem, it shuts down the experiment automatically;
    - Conversely, if there are no alerts, it ramps up automatically to a large percentage
  - Massive data logged: tens of terabytes per day, thousands of attributes for every page shown
    - Data quality challenges, big data challenges
    - New metrics created every week (although OEC stays stable)
  - False positives issues and multiple testing (scale paper). Replication is key.
    Credit is assigned by running a reverse/holdback experiment (ship, test in reverse)



Online is Different than Offline (2)

- Key differences from Offline to Online (cont)
  - Click instrumentation is either reliable or fast (but not both; see paper)
  - Bots can cause significant skews.
    At Bing over 50% of traffic is bot generated!



  - Beware of carryover effects: segments of users exposed to a bad experience will carry over to the next experiment.
    Shuffle users all the time (easy for backend experiments; harder in UX)
  - Performance matters a lot! We run “slowdown” experiments.

    A Bing 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

The Cultural Challenge

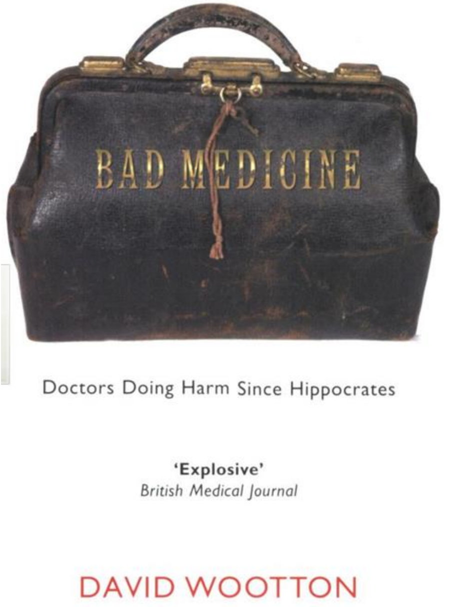
It is difficult to get a man to understand something when his salary depends upon his not understanding it.
-- Upton Sinclair

- 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?"
- The next few slides show a four-step cultural progression towards becoming data-driven

Cultural Stage 1: Hubris

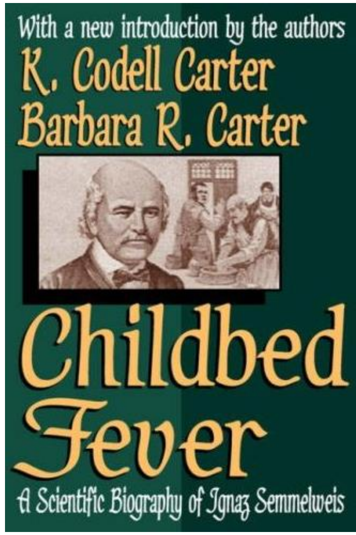
Experimentation is the least arrogant method of gaining knowledge
—Isaac Asimov

- 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, although prevailing theory at the time was that it was caused by miasma: bad air
  - 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



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

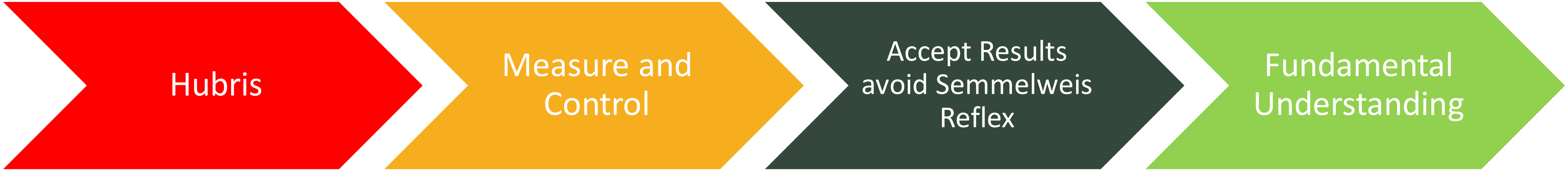
    > *We believe that this chlorine-washing theory has long outlived its usefulness… It is time we are no longer to be deceived by this theory*

- 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, a 50 Euro coin commemorating Semmelweis was issued

Summary: Evolve the Culture



Hubris

Measure and Control

Accept Results avoid Semmelweis Reflex

Fundamental Understanding

- 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

Challenges (1 of 2)

- OEC: Overall Evaluation Criteria
  - What are good OECs or different domains? Challenge is to define short-term metrics that are predictive of long-term impact (e.g., lifetime value)
- Improving sensitivity /reducing variance (e.g., CUPED)
- Bayesian methods
  - When success is rare (e.g., Sessions/UU improves in only 0.02% of experiments), we have to correct for the classical hypothesis testing, which has 5% false positive rate (with p-value 0.05).
    How do we use historical data to compute the posterior that a metric really moved?
- Deep analyses: what segments improved/degraded?
  Use of machine learning techniques to find these
- Form factors
  - Reasonable understanding of web page design for desktop
  - Weak understanding of small-screen (e.g., mobile), touch interactions, apps

Challenges (2 of 2)

- Are there long-term impacts that we are not seeing in 1-2 week experiments?
  Google and Bing act differently here with Google running long-running experiments (e.g., 90-days) and Bing focusing on agility with 1-2 week experiments and iterations
- Aborting experiments
  - With 30-50 experiments starting every day at Bing, some have bugs.
    What are good metrics that have high statistical power to detect issues in near-real-time?
- "Leaks" from experiment variants
  - Social experiments
  - Consumption of shared resources (e.g., memory/disk by one variant).
    In a well-remembered case, one treatment consumed memory slowly causing the servers to crash, but you see very little in the controlled experiment results
- Also See Ya Xu's papers in the main conference and Social Recommender Systems workshop: A/B testing challenges in social networks

The HiPPO

- HiPPO = Highest Paid Person’s Opinion
- We made thousands toy HiPPOs and handed them at Microsoft to help change the culture
- Change the culture at your company
- Fact: Hippos kill more humans than any other (non-human) mammal
- Listen to the customers and don’t let the HiPPO kill good ideas

Summary

The less data, the stronger the opinions

- Think about the OEC. Make sure the org agrees **what** to optimize
- It is hard to assess the value of ideas
  - Listen to your customers – Get the data
  - Prepare to be humbled: data trumps intuition
- Compute the statistics carefully
  - Getting numbers is easy. Getting a number you can trust is harder
- 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
- See http://exp-platform.com for papers
- Visit the Microsoft booth where members of my team will share more and give HiPPOs

Extra Slides

There are Never Enough Users

- Assume a metric of interest, say revenue/user
  - Denote the variance of the metric by σ^2
  - Denote the sensitivity, i.e., the amount of change we want to detect by Δ
- From statistical power calculations, the number of users (n) required in experiment is proportional to σ^2/Δ^2
- The problem
  - Many key metrics have high-variance (e.g., Sessions/User, Revenue/user)
  - As the site is optimized more, and as the product grows, we are interested in detecting smaller changes (smaller Δ)
- Example: A commerce site runs experiments to detect 2% change to revenue and needs 100K users per variant.
  For Bing US to detect 0.1% ($2M/year), we need $20^2 \times 100K$ = 40M × 2 variants = 80M users (Bing US has about 100M users/month)