

A/B Testing at Scale

Slides at http://www.exp-platform.com/Pages/talks.aspx

Pavel Dmitriev, Principal Data Scientist Analysis and Experimentation, Microsoft

Joint work with members of the A&E/ExP team Some slides borrowed from talks by Ron Kohavi



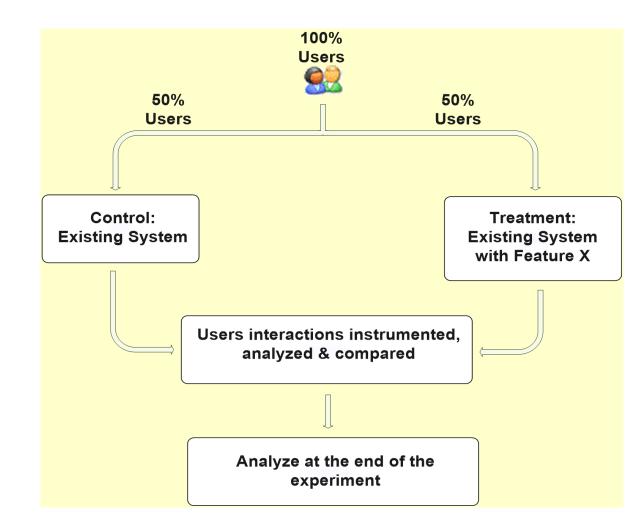
>Introduction to controlled experiments

> Four real examples: you're the decision maker!

➢ Five Challenges

A/B/n Tests aka Controlled Experiments

- A/B test is the simplest controlled experiment
 - A/B/n refers to multiple treatments
 - MVT refers to multivariable designs (rarely used at Microsoft)
- 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



Brief History

The Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.

lisecoury Results Pellers, Pelline et Michanalcal Budinis, CALRO, blowshy of Alabakie Fueng Anousini, Dalind Samo, Nicolan Jacoberg di Sciences and Perlip Honorary Miniter, Anetina Andrary of Arist and sizecour Norder Methods of the Neural Rayd Academy of Estimate, and the Royal Dashk Anademy of Belanam and Leaten: Masher of the Postline Anadorey, Nather of the Greenes Anademy of Mashera, and Anadres, Nather of the Greenes Anademy of London; (Langhlan)) formacity Galam Perlanes; Uthership of London; and Arthur Ballow Thomers at Genetics.

The <u>earliest reference to a controlled experiment</u> was a test for benefits of vegetarianism, suggested in the Old Testament's Book of Daniel

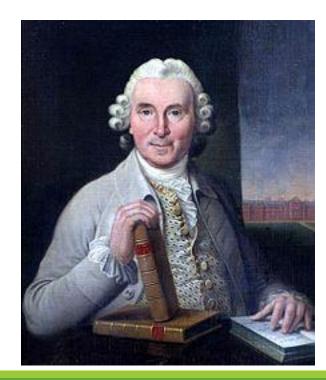
Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see

First controlled experiment / randomized trial for medical purposes: Dr. James Lind, 1747

Scurvy is a disease that results from vitamin C deficiency
It killed over 100,000 people in the 16th-18th centuries, mostly sailors
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

Theory of controlled experiments was formalized by Sir Ronald A. Fisher in 1920's

HAFNER PRESS A DIVISION OF MACMILLAN PUBLISHING CO., JNC. New York COLLIER MACMILLAN PUBLISHERS Loadon



Motivation for A/B testing: Evolving Product Development Process

Classical software development: Spec->Dev->Test->Release

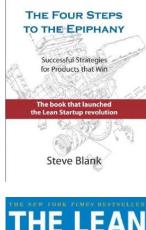
Customer-driven development: Build->Measure->Learn (continuous deployment cycles)

ODescribed in Steve Blank's *The Four Steps to the Epiphany (2005)*

• Popularized by Eric Ries' *The Lean Startup (2011)*

• Measure and Learn parts is where A/B testing comes in!

Why use Customer-driven Development? Because we are poor at assessing the value of our ideas





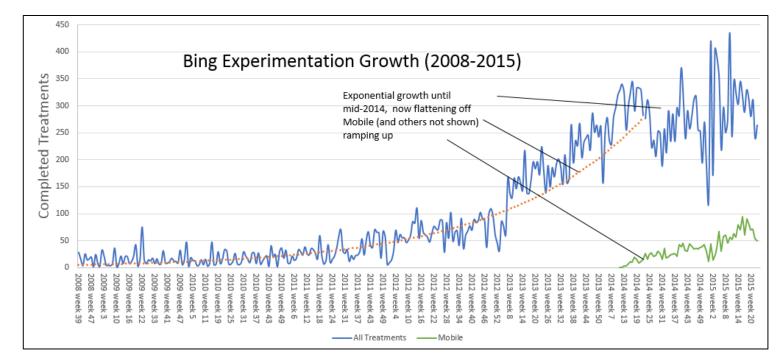
Experimentation at Microsoft



Bing Example

>~300-500 experiments are running concurrently at any given point

- > 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)
- Until 2014, the system was limiting usage as it scaled. Now the limits come from engineers' ability to code new ideas
- There is no single Bing. Since a user is exposed to 30+ concurrent experiments, they get one of 2^30 = over 1 billion variants.



Play Time!

>Four real experiments that ran at Microsoft

>All had enough users for statistical validity

> For each experiment, I tell you the OEC (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:

OA wins (the difference is statistically significant)

○A and B are approximately the same (no stat sig difference)

 $\circ B$ 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



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

Slide intentionally left blank

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. A	ds	Esurance® Auto Insurance - You Could Save 28% with Esurance.	Ads
www.esurance.com/California		www.esurance.com/California	
Get Your Free Online Quote Today!		Get Your Free Online Quote Today!	
		Get a Quote · Find Discounts · An Allstate Company · Compare Rates	

Α

В

> 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

Slide intentionally left blank

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	1,520,000 RESULTS Any time - KDD 2015, 10-13 August 2015, Svdn	OV.	KDD 2015
· •	www.kdd.org/kdd2015 ~		KDD 2015
		and big data. ry dd-2014	We invite submission of papers describing innovative research to aspects of knowledge discovery and data mining, ranging from theoretical foundations to novel models and algorithms for data problems in science, business, medicine, and engineering. Visio papers on new and emerging topics are also velcome, as are ap
	ACM SIGKDD Invitation to Participate - KI 2015 KDD Conference August co	DD 2014, a premier interdisciplinary nference, brings together	wikicfp.com
	Sponsorship A	ttending	Dates: Aug 10 - 13, 2015 Location: Sydney
	KDD 2015 will be held between 10-13 At	tending KDD 2015 Visa Information;	Subjects: Data mining · Database · Knowledge extraction
	August 2015 in Sydney Re	gistration	Website: KDD 2015
	KDD 2015 Call for Papers, Workshops, Or	rganisers ganisers and programm committee	Submissions due: Feb 20, 2015
		embers for KDD 2015	People also search for
_	See results only from kdd.org		ICDM 2015 (Nov 14, 2015)
2	KDD 2015 - The 21th ACM SIGKDD	International Conference	CIKM 2015 (Oct 19, 2015)
	conference.researchbib.com/view/event/33616 -		ICML 2015 (Jul 06, 2015)
	KDD 2015 - The 21th ACM SIGKDD International and Data Mining	Conference on Knowledge Discovery	AAAI 2016 (Feb 12, 2016)
	5		WWW 2015 (May 20, 2015)
	Related searches for kdd 2015 KDD 2014 KDD Sydney		See more 🐱
	KDD Cup 2015 PAKDD 2015		Data from: Wikicfp.com
	WSDM 2015 KDD 2016		Feedback
_			Related searches
3	KDD CUP 2015		KDD 2014
_	https://www.kddcup2015.com - If you have any questions or comments, please se	end an email to	KDD 2016
	support@kddcup2015.com. Updates: 1) Many pe	ople have asked the definition of	WSDM 2015
4	KDD 2015 : ACM SIGKDD Conference	a an Knowledge Discovery	PAKDD 2015
4	myhuiban.com/conference/136 -	e on knowledge Discovery	ICDM 2015
	The Latest Computer Conference and Journal Lis Conference on Knowledge Discovery and Data	t KDD 2015 : ACM SIGKDD	KDD Sydney
	control of thomage blace by and bad		SIGIR 2015
5	KDD 2015 - ACM SIGKDD Internation		KDD Cup 2015
_	www.ourglocal.com/wikicfp/?conid=37&year=201 KDD 2015 -ACM SIGKDD International Conferen Mining. Send this CFP to us by mail: cfp@ourglo	ce on Knowledge Discovery and Data	
6	KDD-2015 Call for Papers, Workshop www.kdnuggets.com/2015/01/kdd-2015-call-pape	ers.html 🔻	
_	ACM SIGKDD Conference on Knowledge Disc be held in Sydney, Australia during August 10-13,	overy and Data Mining(KDD) 2015 will 2015. KDD invites submissions of	
7	KDD 2015 21st ACM SIGKDD Confe eventegg.com/kdd-2015 -	erence on Knowledge	
_	KDD 2015, 21st ACM SIGKDD Conference on K Mining, Australia, Sydney, 10 - 13 August 2015	nowledge Discovery and Data	
8	KDD 2015 : 21th ACM SIGKDD Confe www.wikicfp.com/cfp/servlet/event.showcfp?even		
	KDD 2015 : 21th ACM SIGKDD Conference on I Mining. Home. Login; Register, Account, Logout;	Knowledge Discovery and Data	

Ω

bing kdd 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

Slide intentionally left blank

Example 4: Underlining Links

> Does underlining increase or decrease clickthrough-rate?

Data from: Wikipedia - Freebase

Feedback

Ann.

bing amazon			bing 🎝	amazon		Q	
Web Images Videos N	laps News More	803 🐷 Ronny 🛃	ġ:	Web Images Videos Maps	News More		803 泼 Ronny 🌉 🤻
219,000,000 RESULTS Any time	*			219,000,000 RESULTS Any time 👻			
Amazon.com ® Official Site Ad · www.Amazon.com · 1,858,000 Huge Selection and Amazing Prices. Departments: Unlimited Instant Video You have visited amazon.com once i <u>Amazon Prime</u> <u>Amazon Echo</u> <u>Fire HDX Tablet</u>	⊧ followers on Twitter Free Two-Day Shipping with Prime. s, Digital Music and more	Amazon.com Company Amazon.com, Inc. is an American electronic commerce company with headquarters in Seattle, Washington. It is the largest Internet-based retailed in the United States. Amazon.com started as an online bookstore, but soon diversified, selfing + en.wikipedia.org		Amazon.com ® Official Site - An Ad · www.Amazon.com · 1,858,000+ follow Huge Selection and Amazing Prices. Free T Departments: Unlimited Instant Videos, Digit You have visited amazon.com once in last 7 Amazon Prime Amazon Echo Fire HDX Tablet	vers on Twitter wo-Day Shipping with Prime. tal Music and more	commerce company Washington. It is the in the United States	s an American electronic y with headquarters in Seattle, e largest Internet-based retailer . Amazon.com started as an ut soon diversified, selling +
	est selection of books, magazines, music, DVDs, are, apparel & accessories, shoes, jewelry <u>Kindle eBooks</u> great great selection of Literature &	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		Amazon.com - Official Site www.amazon.com ▼ Official site Customer service 888-280-3321 Online shopping from the earth's biggest sele videos, electronics, computers, software, app Books Online shopping for Books from a great selection of Books Featured		Address: 440 Terry Ave N, Seattle, WA Founded: Jul 06, 1994 · Bellevue, WA Customer service: +1 888-280-3321 CEO: Jeff Bezos Founder: Jeff Bezos Headquarters: Seattle, WA	98109
Shop All Departments Amazon shop by department. Sit directory. This page contains links <u>Movies & TV</u> DVD (971 customer reviews) \$16 \$82.49 () > See more. More	Video at Àmazon.com Amazon Local Amazon Local s a daily deals website that is part of the Amazon	Stock price See more about AMZI ▼ 427.63 -4.00 -0.93% -4.00 -0.93% After Hours 427.25 -0.38 (-0.09%) May 22, 4:14 PM EDT · Market Closed		Shop All Departments Amazon shop by department. Site directory. This page contains links Movies & TV DVD (971 customer reviews) \$160.99 \$82.49 () () > See more. More	Prime Instant Video Find, shop for and buy Prime Instant Video at Amazon.com Amazon Local Amazon Local is a daily deals website that is part of the Amazon	Stock price ▼ 427.63 -4.00 -0.93% After Hours 427.25 -0.38 (-0.09%) May 22, 4:14 PM EDT - Market Closed	See more about AMZN
Electronics Online shopping from a great sele at Electronics Store	Ction A description for this page is not available.	Recent post		Electronics Online shopping from a great selection at Electronics Store	Shopping A description for this page is not available.	Recent post What color is in yo May 25 is National Whe Day Day? http://amzn.	our glass for National Wine to/WineDay
Full Store Directory Amazon shop by department. Sit directory. This page contains links		See more 13 hours ago -		Full Store Directory Amazon shop by department. Site directory. This page contains links	Outlet Outlet. Online shopping for outlet deals- -find low prices on	See more	14 hours ago -
Search results from amazon.com	n Search	People also search for See all (10+		Search results from amazon.com	Search	People also search for	See all (10+)
theguardian.com · 5 r Industry sources say	Random House said to be in dispute ninutes ago Penguin may block sales of its books on Amazon dispute, but both sides are	ebay eBay Flipkart Google Apple Inc. Alibaba Group		theguardian.com · 9 minutes	n may block sales of its books on Amazon	ebay eBay Flipkart Google	Apple Inc. Alibaba Group

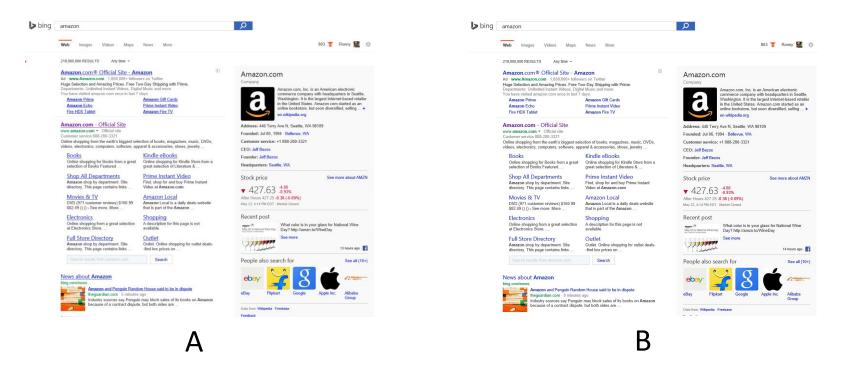
1

Data from: Wikipedia · Freebase

Example 4: Underlining Links

>Does underlining increase or decrease clickthrough-rate?

>OEC: Clickthrough Rate on 1st SERP per query



- 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

Underlining Links

Slide intentionally left blank

Key Lesson: 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

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



Introduction to controlled experiments
 Four real examples: you're the decision maker
 Five Challenges

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

Challenge 1: Trustworthiness

Getting numbers is easy. Getting numbers you can trust is hard. -- Ronny Kohavi

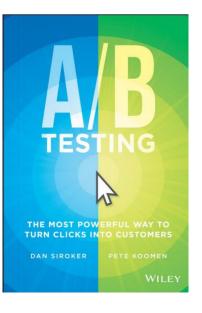
> Two very good books on A/B testing get the stats wrong (see Amazon reviews).

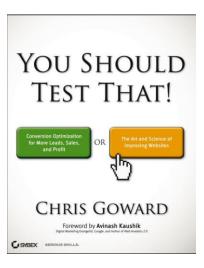
As new experiment designs and statistical techniques get deployed, chance of error increases

○ It took us ~2 years to get implementation of the <u>CUPED</u> variance reduction technique right

- > Metrics are added and modified ~daily, instrumentation changes ~weekly
- > Bots may cause significant skews (over 50% of Bing traffic are bots)
- > Great technique 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?

>Twyman's Law: any figure that looks interesting or different is usually wrong





Challenge 2: Protecting the User

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

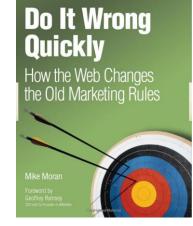
>As more and more ideas get tested, possibility of user harm increases

- O Buggy feature or a bad idea making it to real users
- O Less manual monitoring of experiments
- O Interactions are possible between concurrently running experiments

Need to minimize harm to users!

Requires a combination of approaches

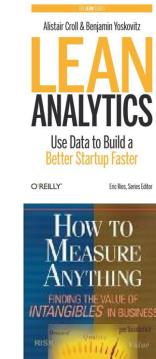
 Automatically detect and shut down bad experiments, fast!
 Start small and then ramp up
 Run with partial exposure (e.g. only on 1 out of 10 queries)
 Run non-overlapping experiments when suspect interactions
 Automatically detect interactions (we run all-pairs test nightly)



Challenge 3: The OEC

It's not enough to do your best; You must know what to do, and then do your best. -- W. Edwards Deming

- OEC = Overall Evaluation Criterion
 OLean Analytics call it OMTM: One Metric That Matters
- Two key properties (paper):
 Alignment with long-term company goals
 - Ability to impact (Sensitivity)
- A single metric or a few KEY metrics. Beware of the Otis Redding problem: "I can't do what ten people tell me to do, so I guess I'll remain the same."
- Designing a good OEC is hard
 - Example: OEC for a search engine





Douglas W. Hubbard

Challenge 3: The OEC (Metric Sensitivity)

Queries

Session

Sessions

User

>OEC for a search engine:

> Problem: almost never moves in our experiments

• Width of the confidence interval is proportional to CV/\sqrt{n} , where $CV = coefficient of variation = <math>\sigma/\mu$.

Queries

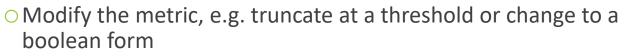
Month

• For many metrics CV is stable as experiment goes on, so confidence interval shrinks ~ $1/\sqrt{n}$.

○ Not the case for Sessions/User

➢Solutions:

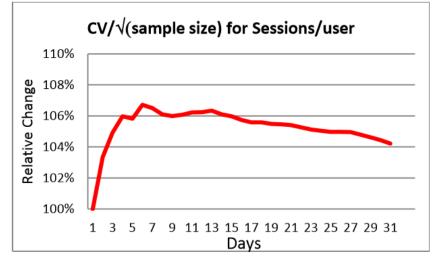
- Run larger experiments (e.g. 50/50)
- <u>Triggering</u>: analyze only users who were exposed to the change
- Variance Reduction: <u>CUPED</u> technique uses delta from pre-experiment period



Users

Month

- Use a more sensitive surrogate metric. E.g. Session Success Rate is predictive of Sessions/User move and is more sensitive
 - Optimization problem: maximize metric sensitivity given constraint of alignment



Challenge 4: Violations of classical assumptions of a controlled experiment

Every theory is correct in its own world, but the problem is that the theory may not make contact with this world. -- W. Edwards Deming

Unstable user identifiers due to e.g. cookie churn and multiple device usage. Leads to the same real user potentially exposed to both treatment and control

OMUID backup/restore (paper) helps but does not completely solve the problem

Leaks due to shared resources

• Cache is a shared resource. If control and treatment are of different size (e.g., control is 90%, treatment is 10%), then control has a big advantage because its elements are cached more, leading to performance improvements

 If treatment leaks memory on the server that servers requests for both control and treatment, performance slows down equally for both variants and degradation is not reflected in the scorecard

>Network interactions resulting in spill-over effects

 Facebook's emotional contagion experiment (<u>paper</u>) suppressed positive posts for users. As a results users started posting fewer positive posts themselves. This impacted their friends from both control and treatment in the same way.

>While some partial mitigations exist, these are largely open problems

Challenge 5: Analysis of Results (NHST)

Everything should be made as simple as possible, but not simpler.

-- Albert Einstein

- > NHST = Null Hypothesis Testing, p-value <= 0.05 is the common threshold for rejecting the null hypothesis
- \geq P-value is often misinterpreted. Here are some incorrect statements (from Steve Goodman's *A Dirty Dozen*):
 - If P = .05, the null hypothesis has only a 5% chance of being true 1.
 - 2. A non-significant difference (e.g., P > .05) means there is no difference between groups
 - 3. P = .05 means that if you reject the null hypothesis, the probability of a type I error (false positive) is only 5%
- > NHST is asymmetric: can only reject the null
- \geq Other problems: multiple testing, early stopping



Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015



A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of Basic and Applied Social Psychology (BASP) announced that the journal would no longer publish papers containing P values because the statistics were too often used to support lowerguality research¹.

 \succ The problem is that (loosely speaking) p-value is P(Data|H₀), but we want P(H₀|Data)

> One approach: Bayesian framework to estimate P(H1|Data), using a prior learned from past experiments (paper)

Challenge 5: Analysis of Results (Heterogeneous Treatment Effect)

>We know treatment effect differs from person to person:

• Feature is not popular in one country but good in others

• The feature does not render correctly in a certain browser

There could be many sub-populations (segments) where treatment effect varies or even flips sign: browser, location, age, gender, etc.

Need to find them automatically: thousands of metrics and hundreds of segments are impossible to examine manually; multiple testing issues

>Machine Learning framework: $\tau = Y(T) - Y(C) = f(X)$

 τ = treatment effect for a user, the difference between potential effects in treatment and control, X = segments. The goal is to learn f, and then use it to identify "interesting"/"different" segments. Note: τ is not observed.

 Active research area, see e.g. <u>paper</u>

➢ Visualization?

Challenge 5: Analysis of Results (Novelty and Learning Effects)

>A common response we hear when someone's experiment fails to move metrics positively is "users just need time to adapt to the new experience"

- Are results observed in a short-term (e.g. 2-week) experiment good predictors of the long-term impact?
- >In Bing we run several long-term experiments and only observed small to no changes
- Google reported experiments manipulating the number and quality of ads having significant learning effect (<u>paper</u>)
- There are many caveats with running and interpreting results of long-term experiments (<u>paper</u>)
 - For example Selection Bias: users who remain till the end of a long-running experiment are very different from the average user.

Summary

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

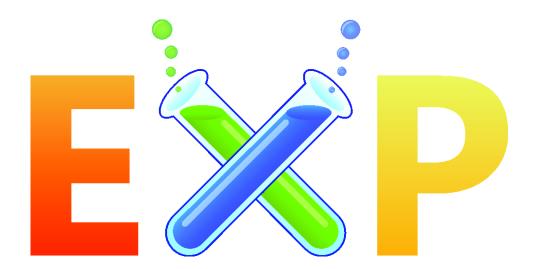
>We are poor at assessing the value of ideas. Run experiment and get the data!

- While the theory of experimentation is well established, scaling experimentation to millions of users, devices, platforms, websites, apps, social networks, etc. presents new challenges.
 - Trustworthiness
 - Protecting the users
 - The OEC
 - Violations of classical assumptions
 - Analysis of results

Exciting research area! Would love to have more academic involvement. No need for access to industry data; easy to setup and run experiments on your own web site, e-mail, social network, etc.



Questions?



http://exp-platform.com