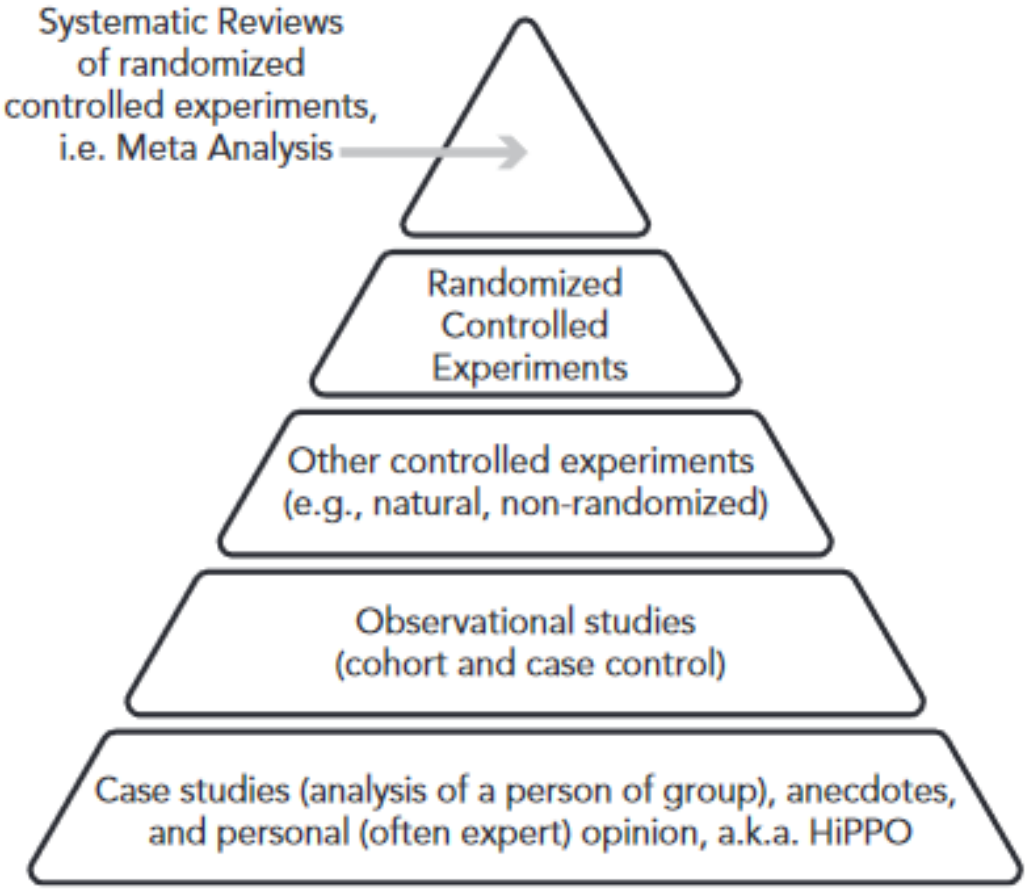
Online Experimentation: Beyond A/B testing

Lidija Turkovic 09.03.2021

Hierarchy of evidence



NETFLIX

Microsoft

Google

Booking.com

ebay

amazon

airbnb

Spotify

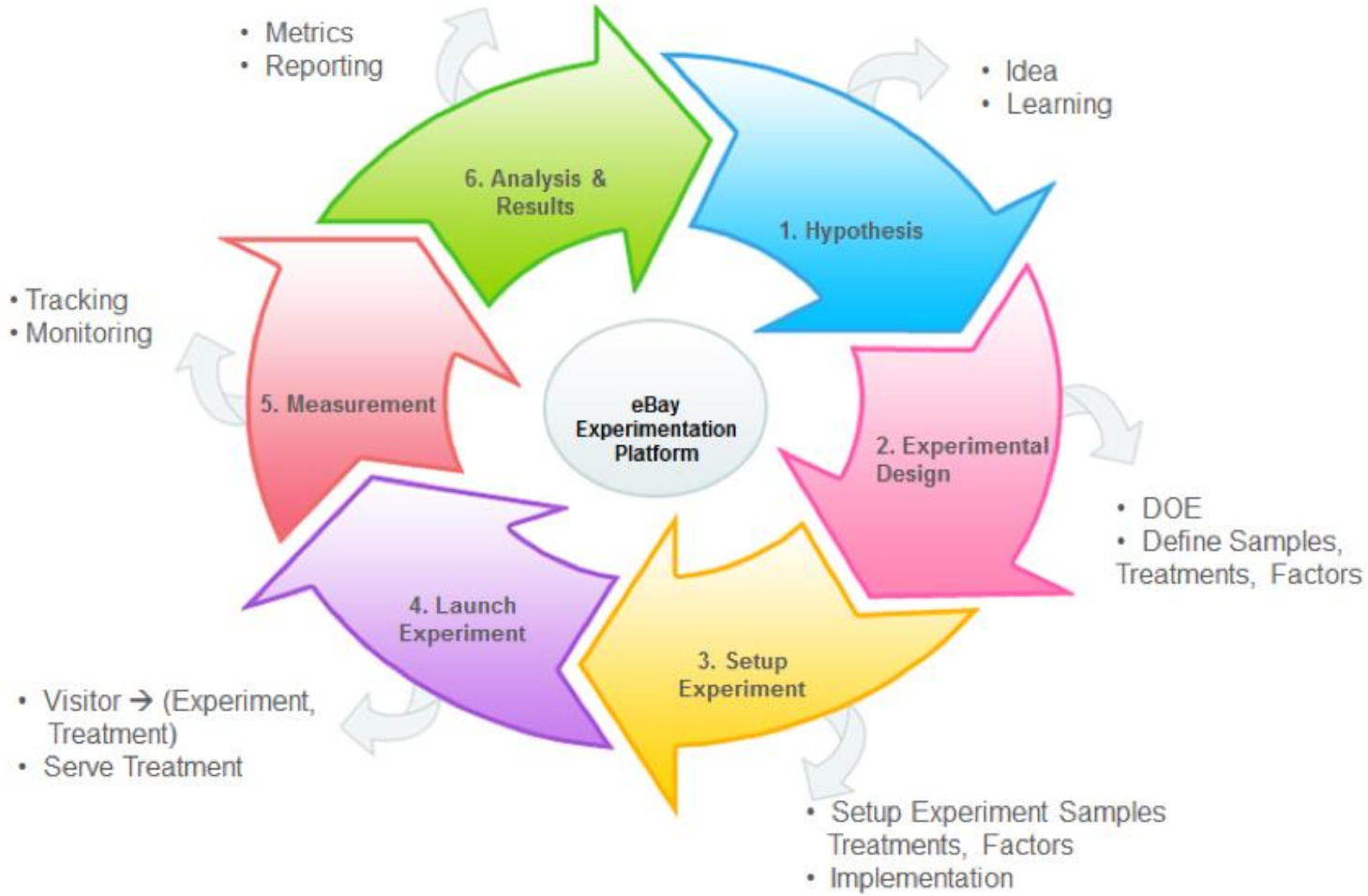
Bing

lyft

Uber

BMJ. 2014; 348: g3725

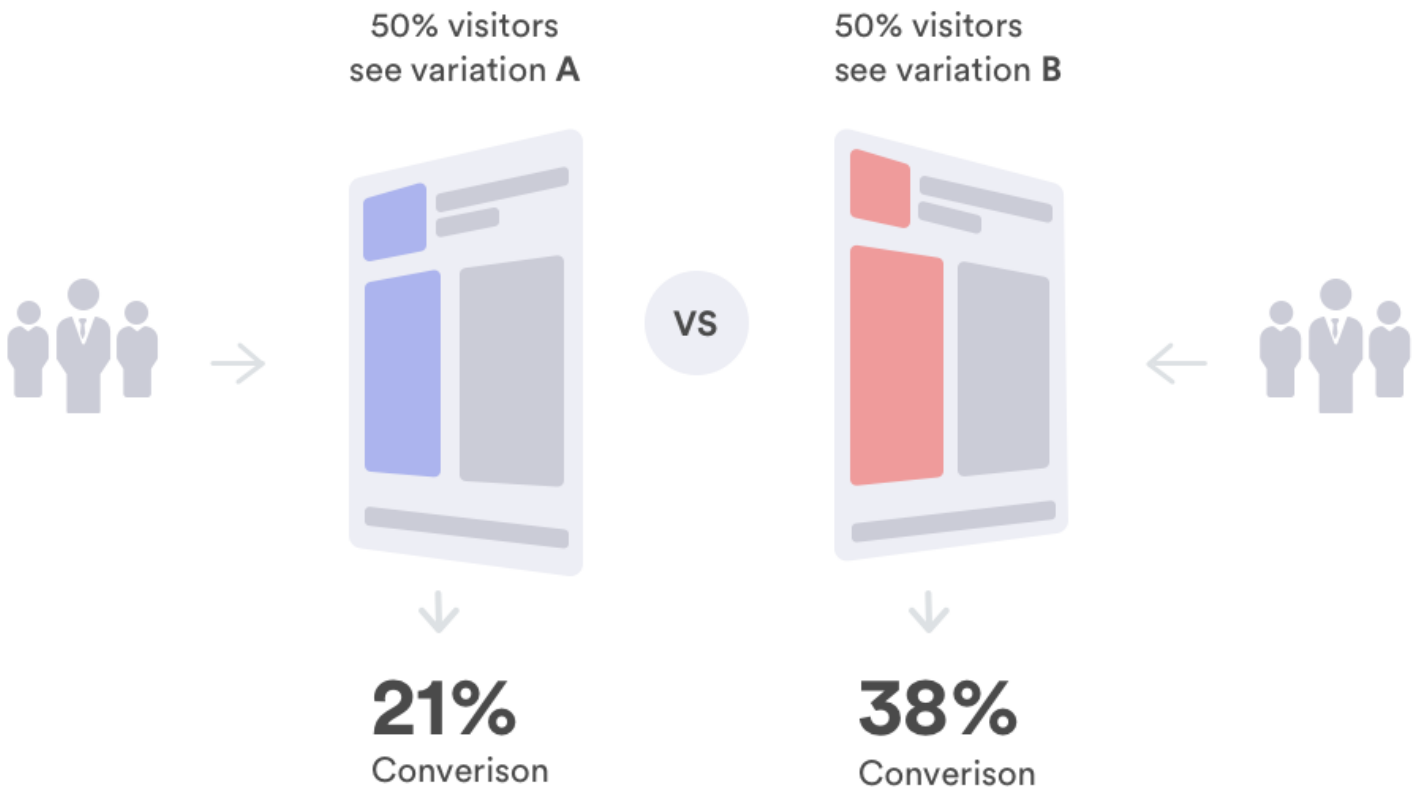
Experiment lifecycle



Common complexities

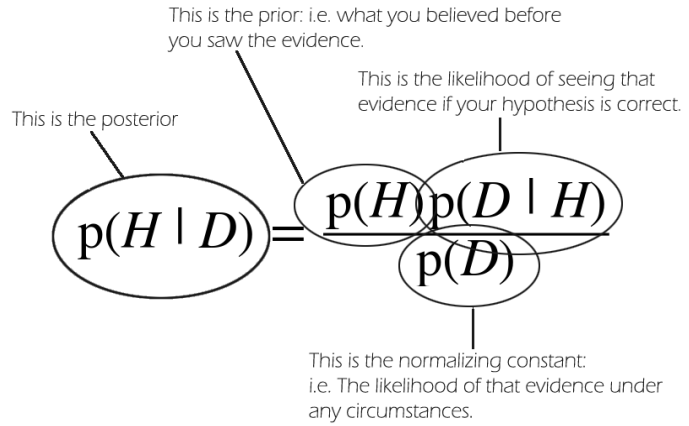
1.	2.	3.	4.	5.
• Experiments take too long to complete • No interim results available	• User dependence on past search and recommendation sessions	• Suspected cannibalisation between experiments	• Suspected interactions between experiments	• Tricky distribution of the primary metric

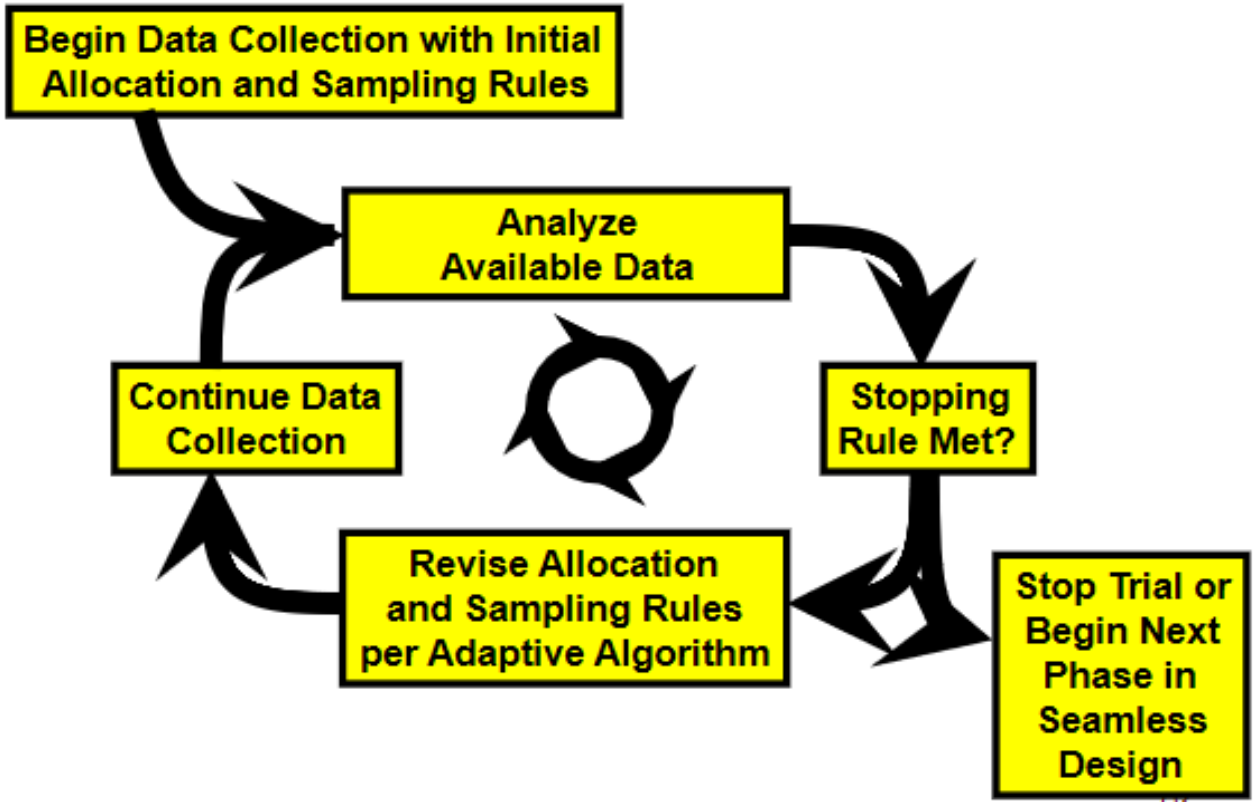
A/B Testing



- Predetermined sample size and length of the experiment
- Have to wait until experiment is over to see results

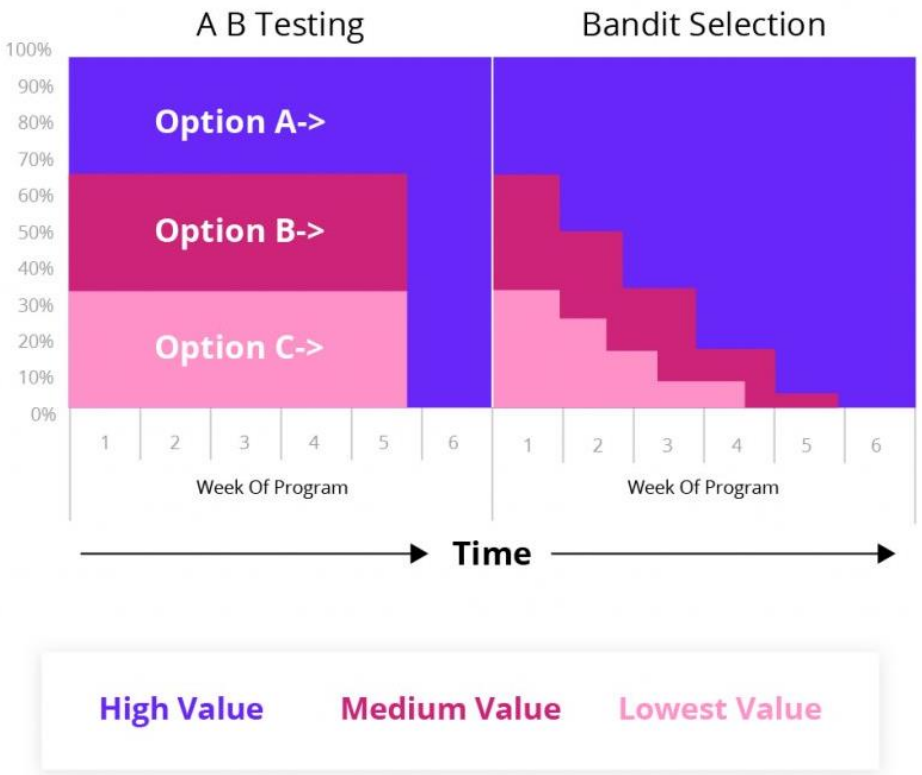
Bayesian Adaptive Design





JAMA 2006;296:1955-1957.

Bayesian Bandits



Exploration/Exploitation trade off

A/B test

- Current site: 4% conversion rate (CR)
- New site: unknown conversation rate
- Target effect size: 1% increase in conversion rate

- H_0: Current site CR = New site CR
- H_1: New site CR – Current site CR ≥1

- Significance level: 5%
- Power: 80%
- Sample size needed: $n_1 = 11165$, $n_2 = 11165$

- Visitors to the site: 100 per day
- Length of the experiment: 223 days

Bayesian Bandits

- Current site: 4% conversion rate
- Visitors to the site: 100 per day

- Non-informative prior: assign visitors 50%/50% on Day 1
- Interim analysis: Daily
- Stopping rule: 95% probability that new site has increased CR by 1%

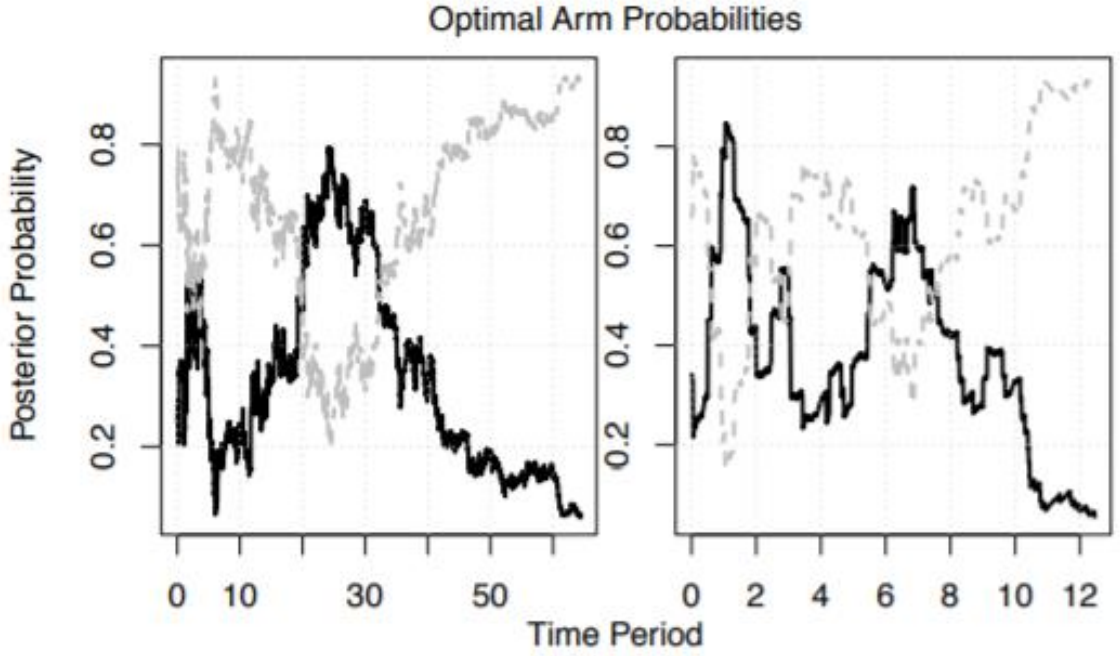
- Sample size: unknown
- Length of experiment: unknown

- Day 2 interim: current site appears superior to new site, new posterior probabilities 30%/70% - weights for Day 3 etc.

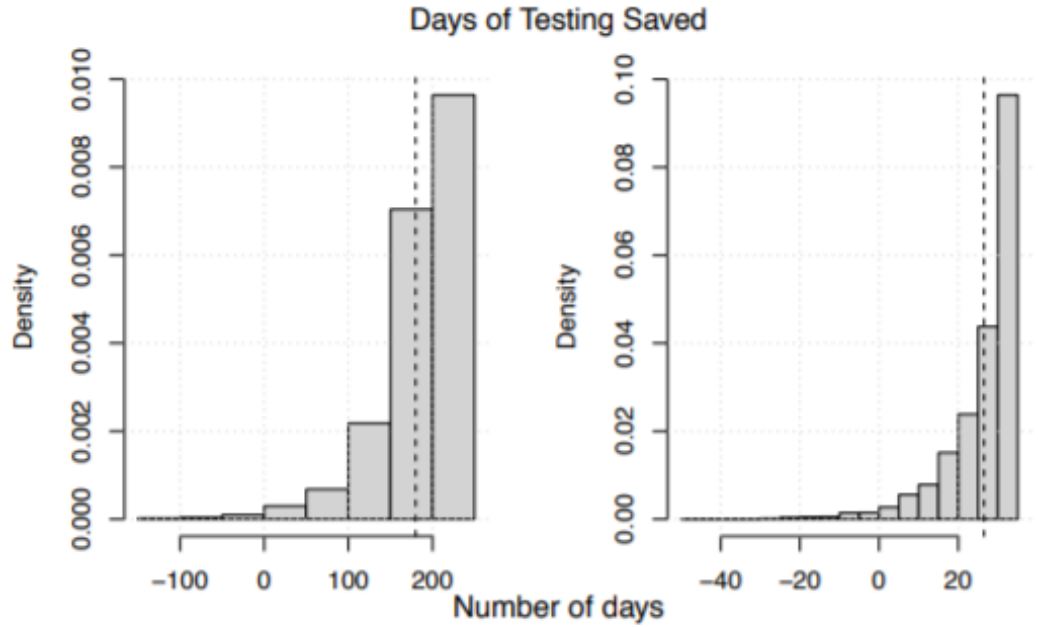
International Review of Economics and Business 2018, 21(1):95-104, Google Analytics 2012

Bayesian Bandits example continued

Simulation 1 Simulation 2

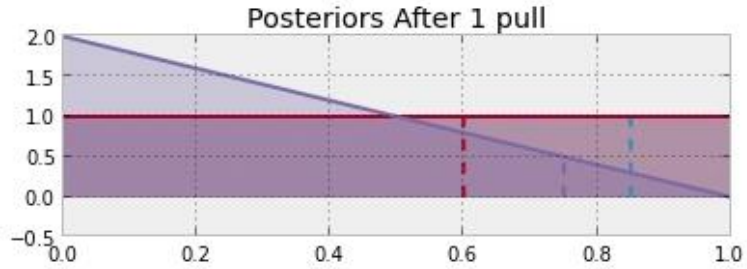


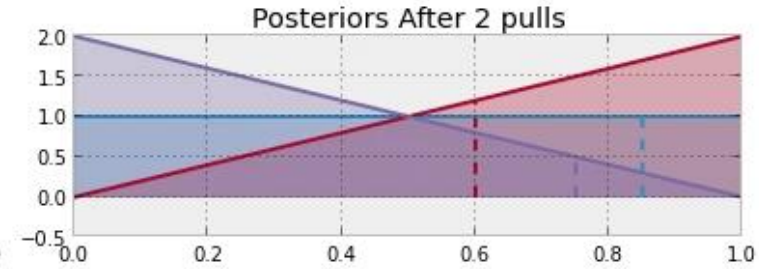
Effect size = 1% Effect size = 3%

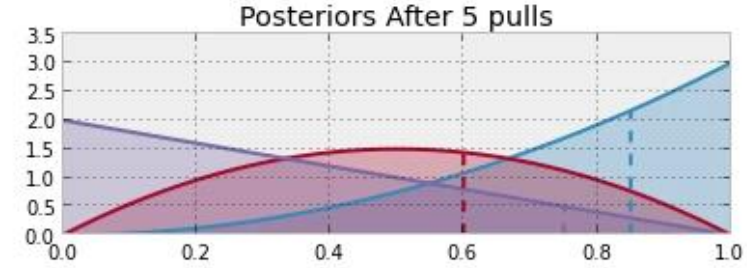


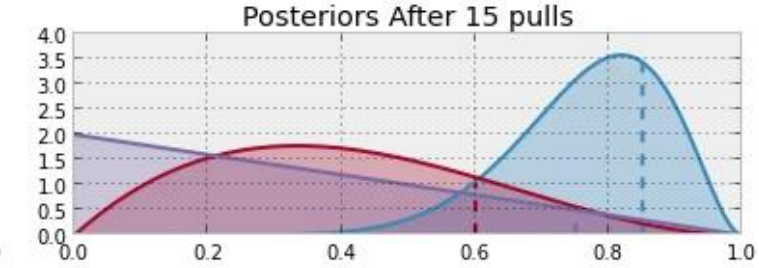
Experiment was 157 days shorter using Bayesian Bandits design (66 days) compared to A/B testing (223 days).

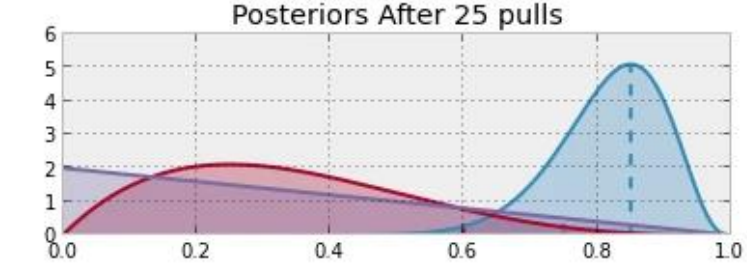
International Review of Economics and Business 2018, 21(1):95-104, Google Analytics 2012

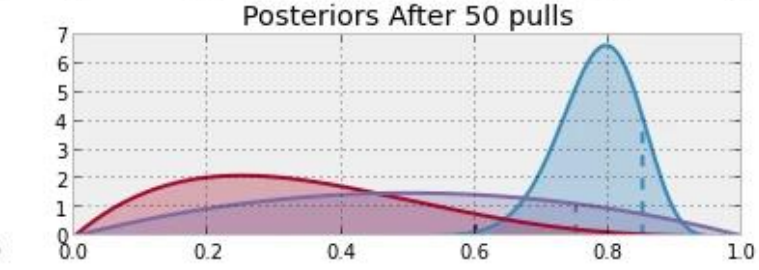

Posteriors After 1 pull

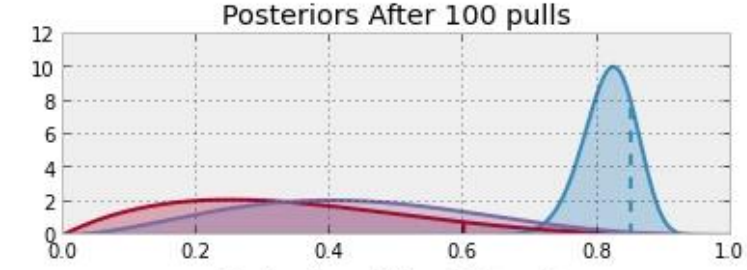

Posteriors After 2 pulls

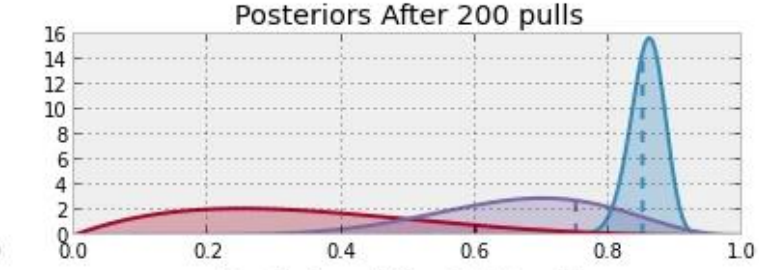

Posteriors After 5 pulls

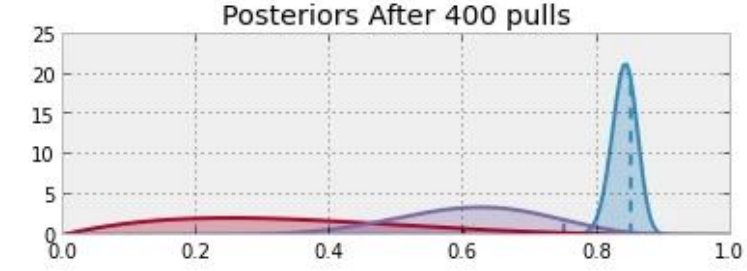

Posteriors After 15 pulls

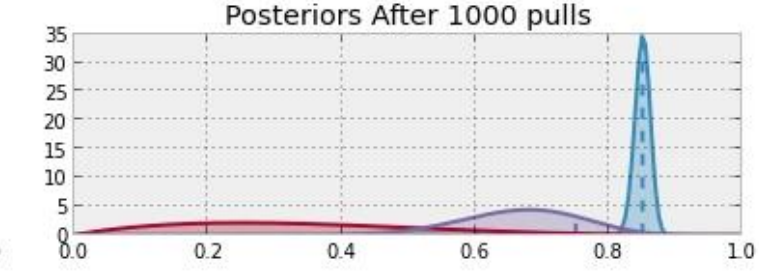

Posteriors After 25 pulls


Posteriors After 50 pulls


Posteriors After 100 pulls


Posteriors After 200 pulls


Posteriors After 400 pulls

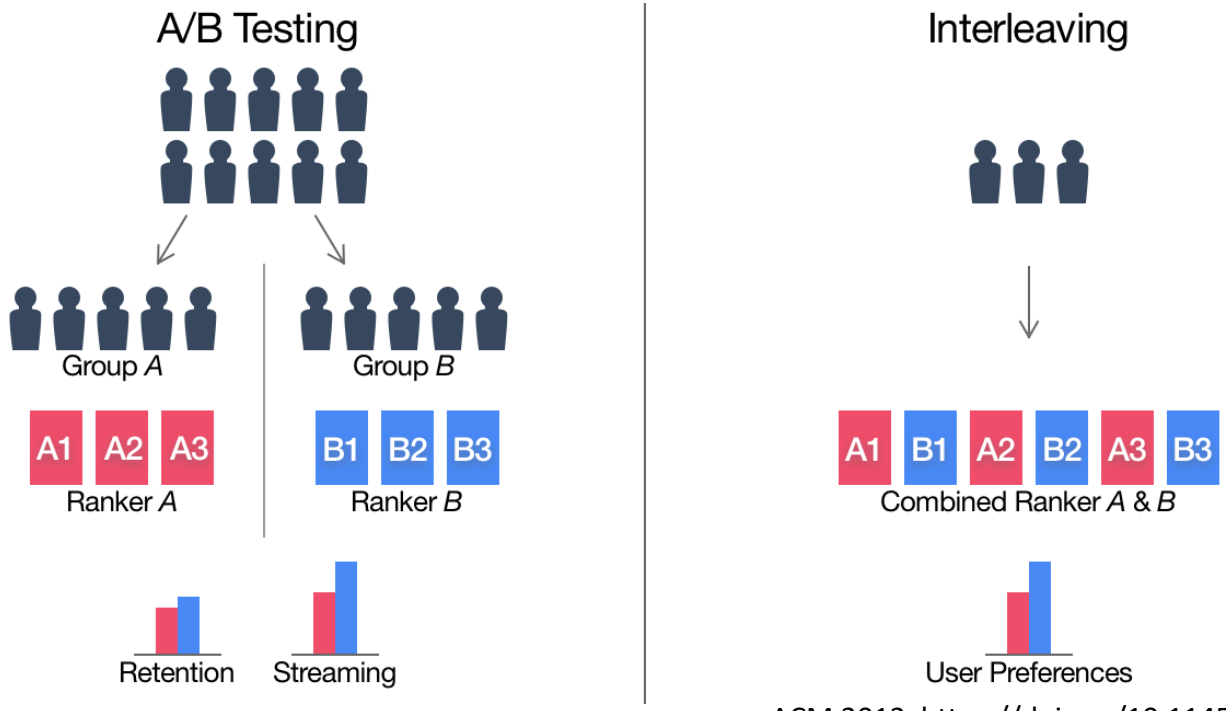

Posteriors After 1000 pulls

Limitations

- Needs careful thinking when setting up:
  - Stopping rules
  - Frequency of interim analysis
- Assumes that effect doesn’t vary over time
  - For long experiments, may need reset the experiment periodically
- Needs understanding of math behind convergence algorithms before implemented
- Alternative: Contextual Bandits (https://arxiv.org/pdf/2002.00467.pdf)

Past search/recommendations sessions dependency

- User as randomization unit
- Clinical trials: measure some outcome metric pre and post "treatment" on the same person
- Interleaving: repeated measures design



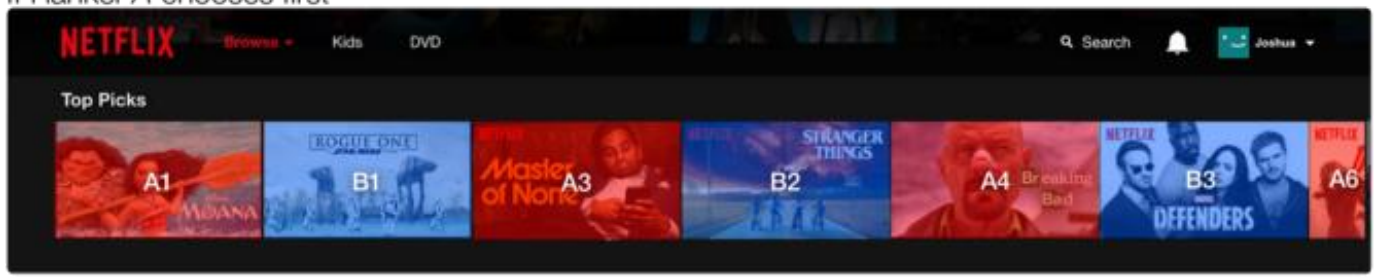
ACM 2012 https://doi.org/10.1145/2094072.2094078
https://netflixtechblog.com/tagged/ab-testing

Interleaving

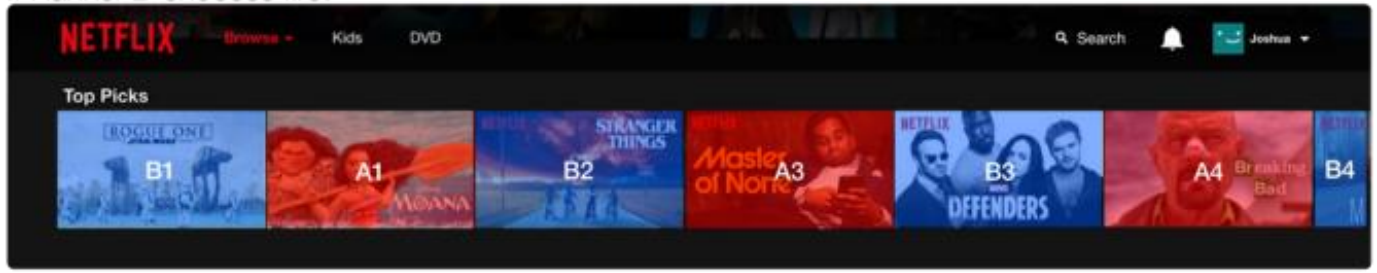
Ranker *A*

Ranker *B*

If Ranker *A* chooses first



If Ranker *B* chooses first



ACM 2012 https://doi.org/10.1145/2094072.2094078
https://netflixtechblog.com/tagged/ab-testing

Limitations

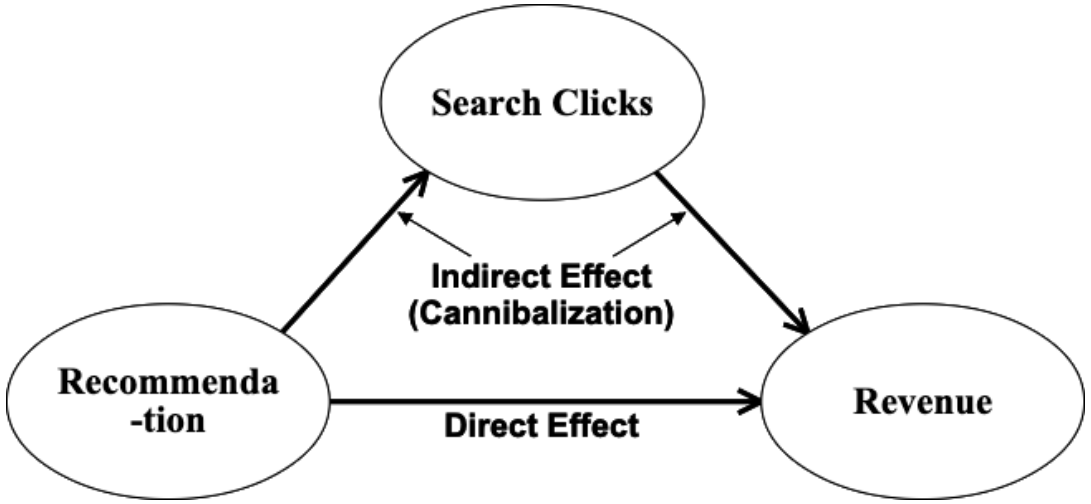
- Implementing an interleaving framework can be fairly complex
- Lack of scalability due to difficulty in automating and building consistency checks
- It is a relative measure of user preferences
- Doesn’t allow us to directly measure some important metrics such as user retention

Cannibalisation between products

Causal inference framework

	% Change = Effect/Mean of Control
Recommendation Clicks	+28%***
Search Clicks	-1%***
Conversion	+0.2%
GMS	-0.3%

GMS = Gross Merchandise Sale
* Level of statistical significance

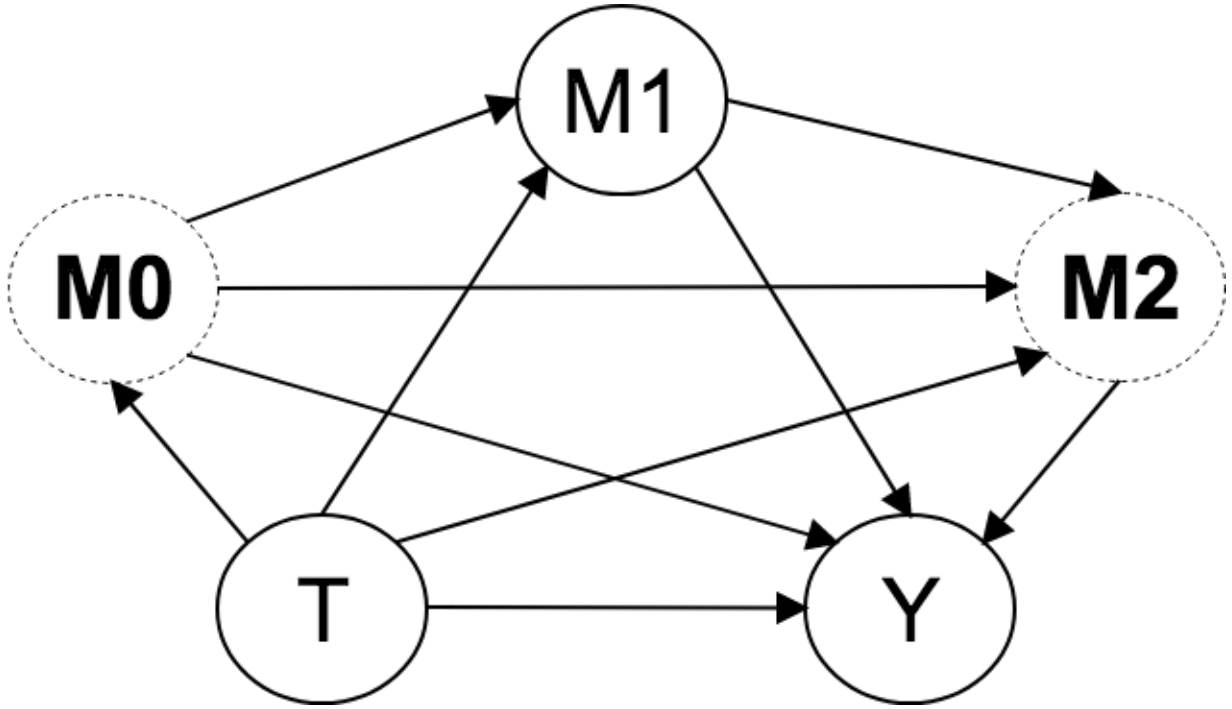


Directed Acyclic Graph (DAG)

Mediator = potential mechanism by which an independent variable can produce changes on a dependent variable.

Statistics Applications 2019; arXiv:1906.09757v1, Nature 2018 487–493

Cannibalisation between products



		% Change = Effect/Mean of Control	
Cannibalization in Gain	**Causal Mediation**	**Conversion**	**GMS**
The Original Gain from recommendation	GADE(0) (Direct Component)	0.5%*	0.2%
The Loss Through Search	GACME(1) (Indirect Component)	-0.3***	-0.4%***
The Observed Gain	ATE (Total Effect)	0.2%	-0.3%

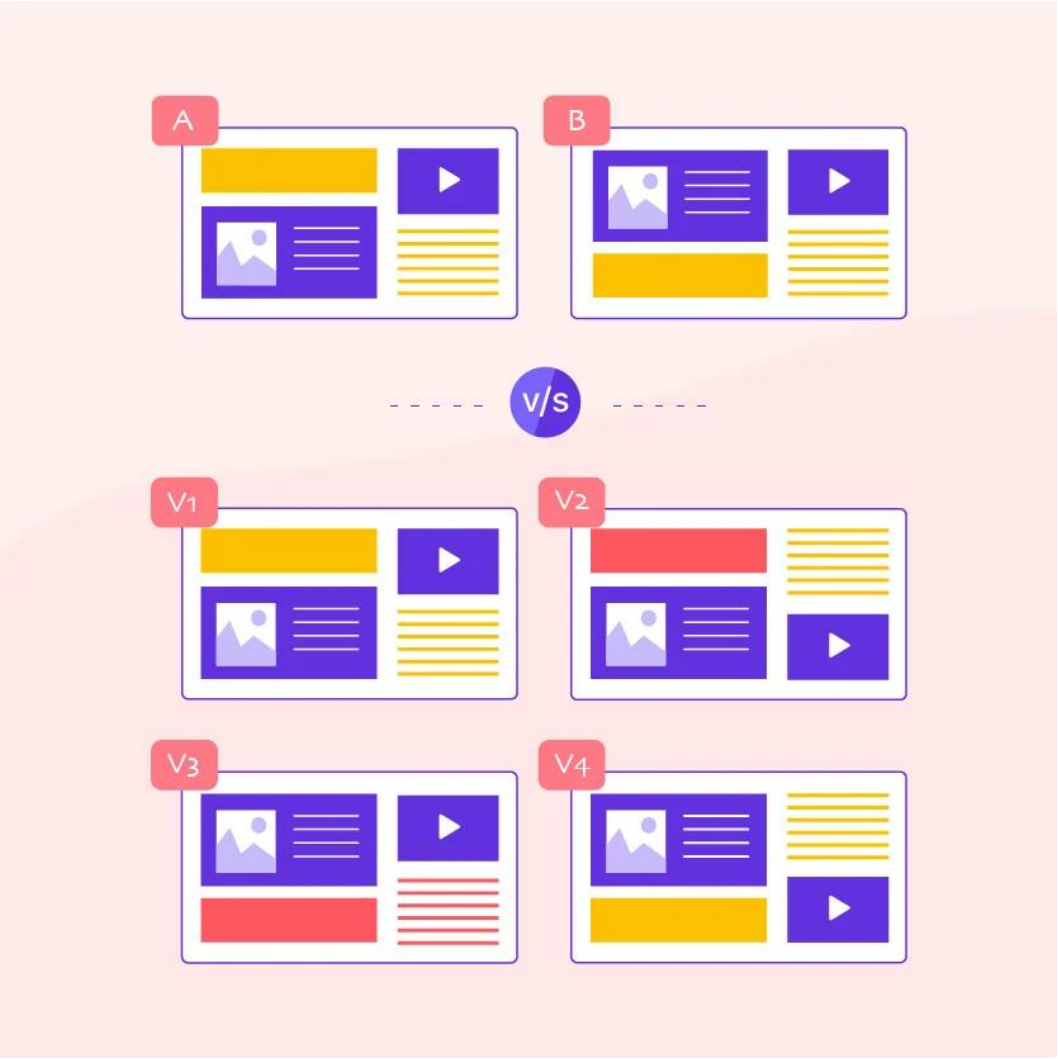
GADE = Generalized Average Direct Effect
GACE = Generalized Average Causal Effect
* Level of statistical significance

Statistics Applications 2019; arXiv:1906.09757v1

Limitations

- Causal diagrams do not specify size of the associations and remain qualitative in nature
- Causal diagrams are only as good as the background information used to create them
- Ability to detect systematic error only (not random error)

Interactions between experiments



If suspected, can test every possible combination using multivariate testing (MVT)

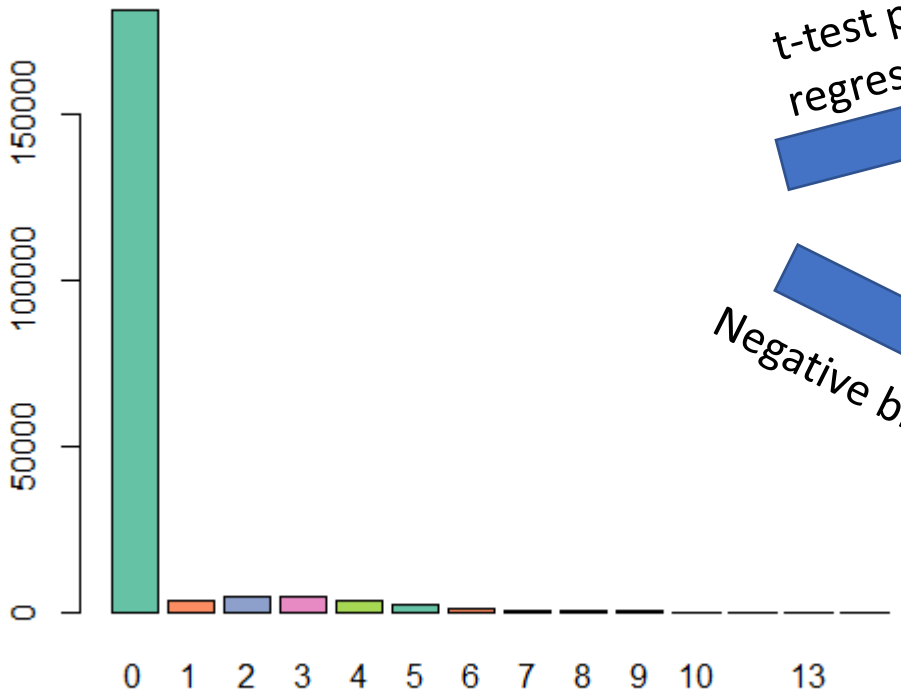
Limitations

- Interactions are often not known
- MVT is very time consuming, especially as number of variations grow
- Schedule either mutually exclusive experiments on the same traffic or experiments with weak interactions only via an experimentation platform

Highly skewed, zero inflated primary metric

Can we improve precision of the estimates (i.e. lower standard error) at analysis stage by using a better fitting model and therefore can decrease the sample size needed?

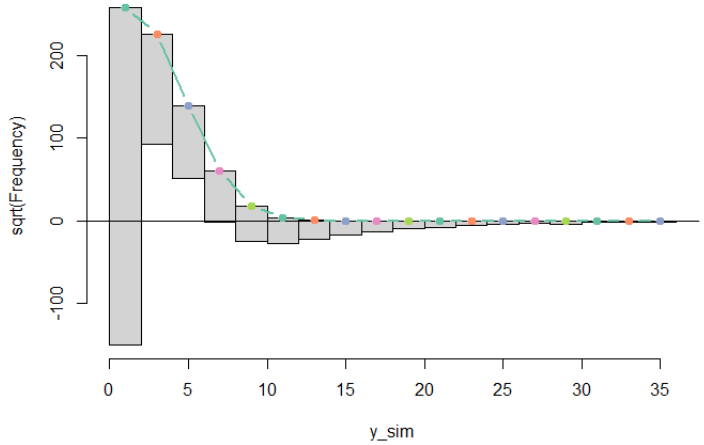
Simulate some zero inflated data n=200000

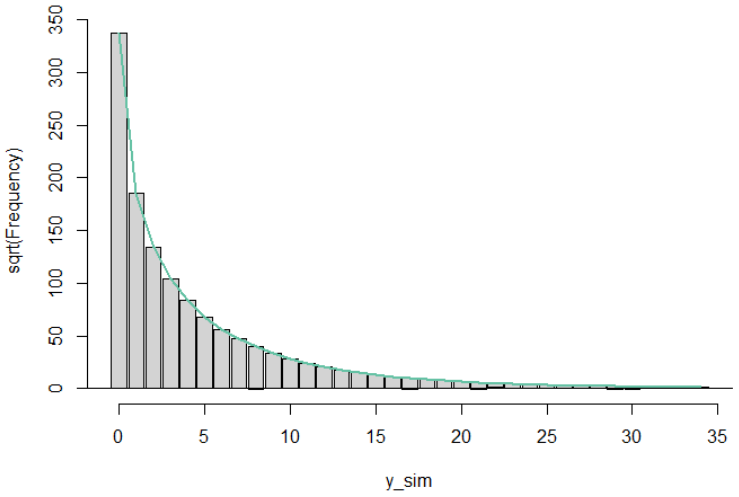


t-test placed into linear regression model

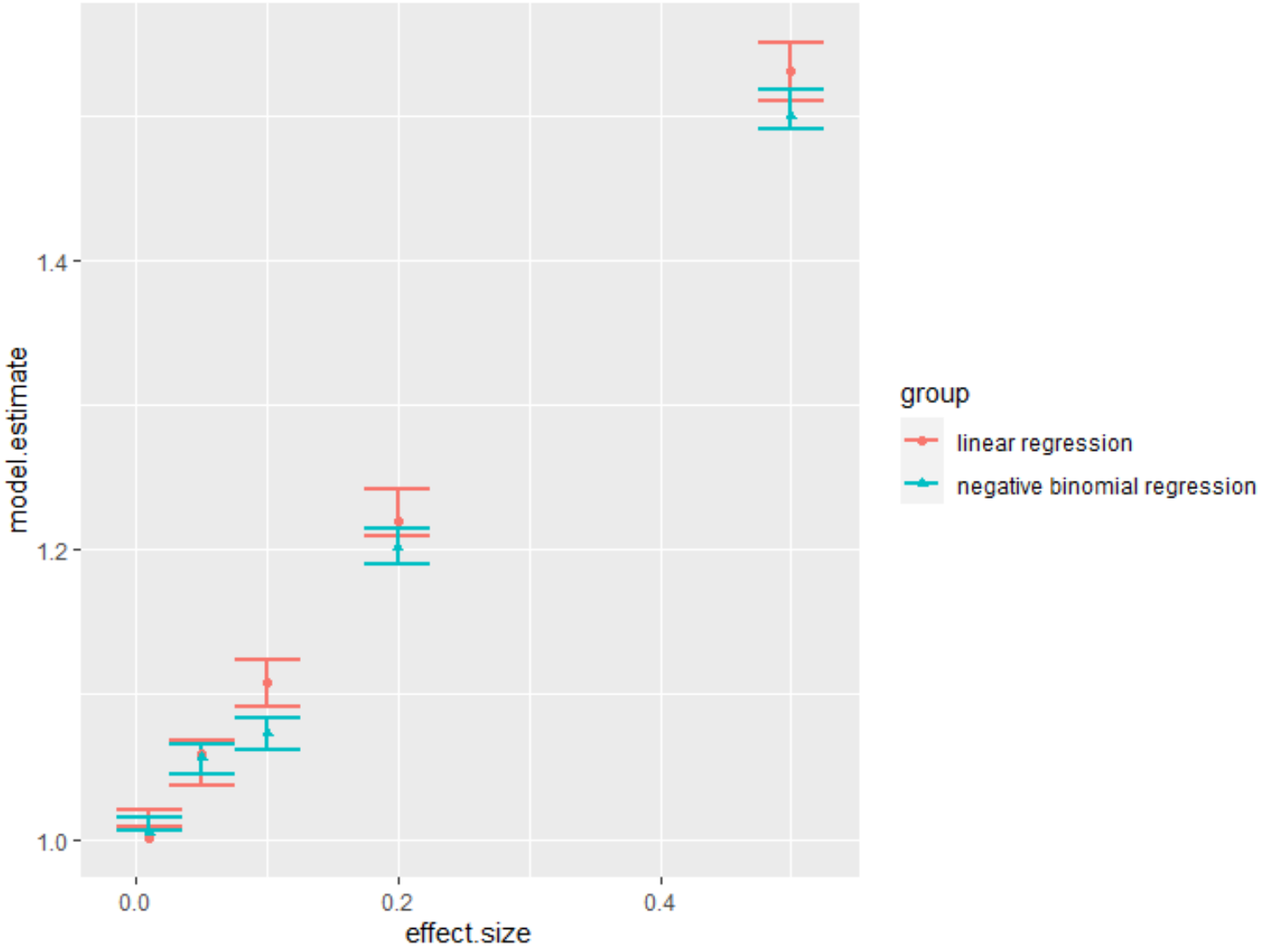
Negative binomial regression

Model Fit





95% confidence intervals for both model estimated coefficients, over 1000 simulations and a range of effect sizes (1% to 50%) with n=200000

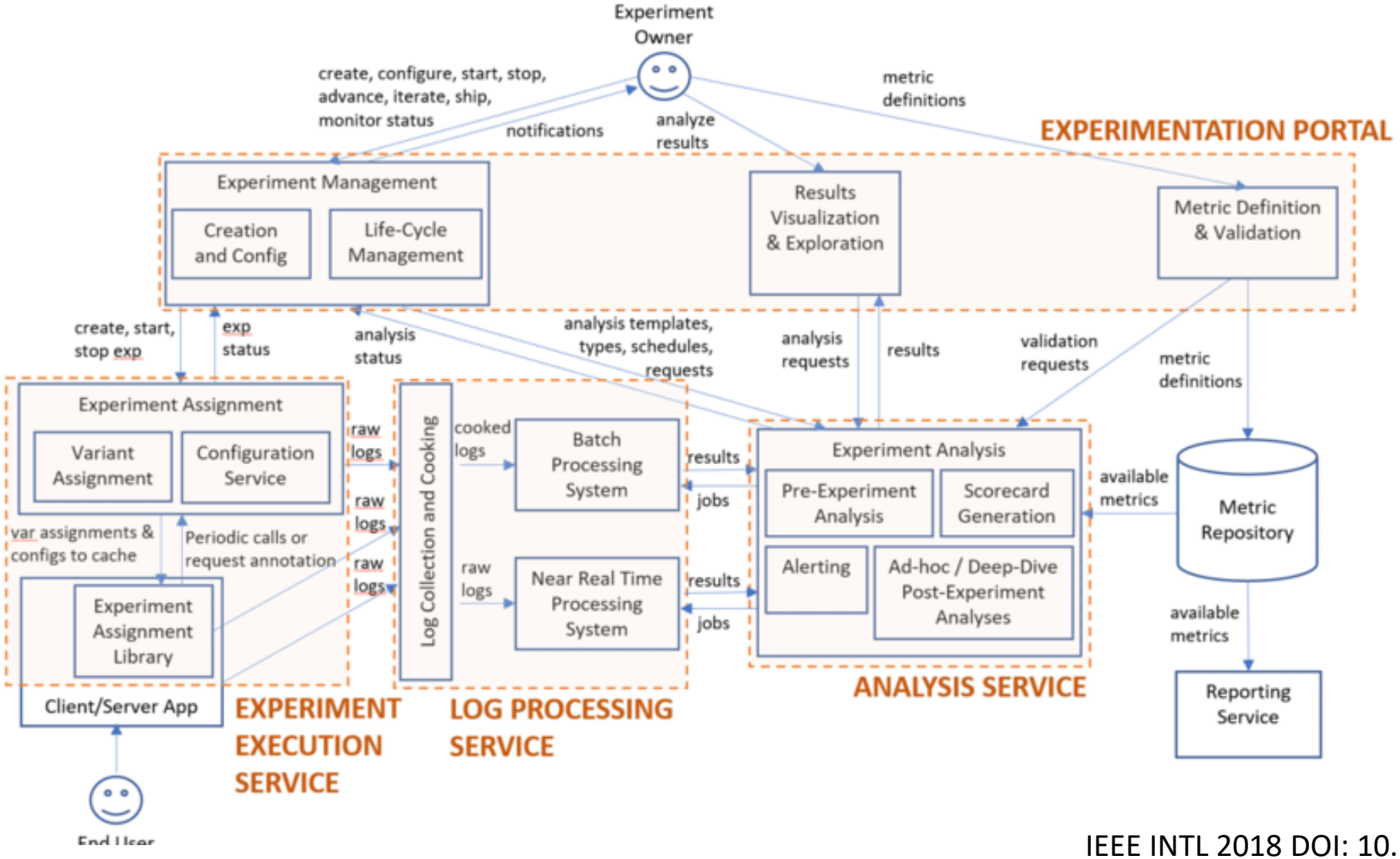


Not much difference….

Limitations

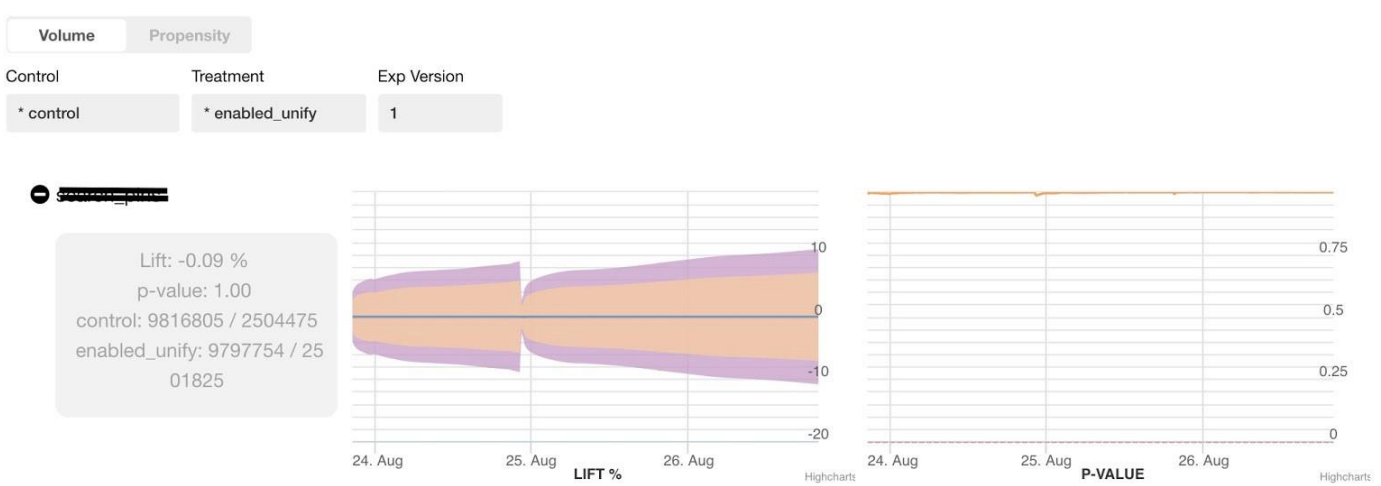
- Simulate different models and scenarios (sample size/effect size)
- Model coefficients may no longer be on linear scale
- Only likely to have modest reduction in sample size
- Alternative: develop a composite metric that has a more desirable distribution

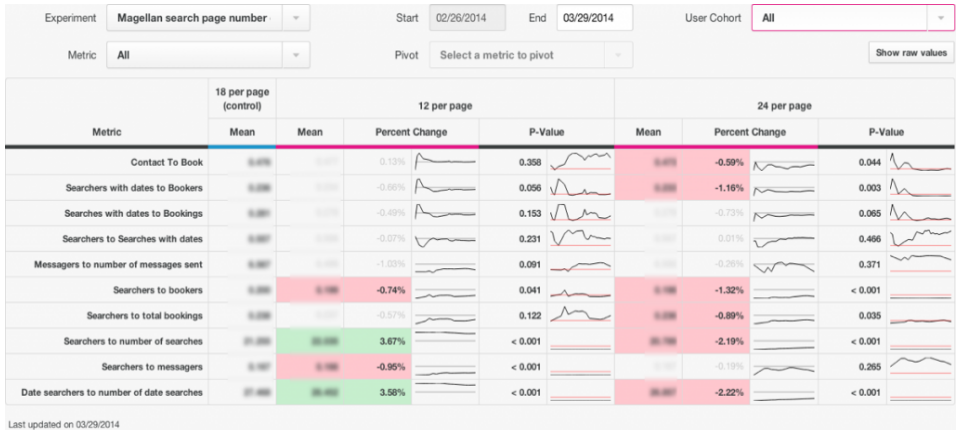
Anatomy of large scale experimentation platform



IEEE INTL 2018 DOI: 10.1109/ICSA.2018.00009

Experimentation platform dashboards





Found 285 Results

Download Displayed Rows | Fullscreen

Date	Control [redacted]	Enabled_unify [redacted]	Control NUM_USERS_ALL_TIME	Enabled_unify NUM_USERS_ALL_TIME
2019-08-26T19:45:00Z	9,816,805	9,797,754	2,504,475	2,501,825
2019-08-26T19:30:00Z	9,742,048	9,723,179	2,490,994	2,488,583
2019-08-26T19:15:00Z	9,668,363	9,649,427	2,477,710	2,475,271
2019-08-26T19:00:00Z	9,596,902	9,578,829	2,464,542	2,462,340
2019-08-26T18:45:00Z	9,524,444	9,507,648	2,451,504	2,449,352
2019-08-26T18:30:00Z	9,453,664	9,436,471	2,438,605	2,436,487
2019-08-26T18:15:00Z	9,383,393	9,365,170	2,426,242	2,423,707

Metric	T	C	Delta (%)	*p*-value
Overall Click Rate	0.9206	0.9219	-0.14%	8e-11
Web Results	0.5743	0.5800	-0.98%	~0
Answers	0.1913	0.1901	+0.63%	5e-24
Image	0.0262	0.0261	+0.38%	0.1112
Video	0.0280	0.0278	+0.72%	0.0004
News	0.0190	0.0190	+0.10%	0.8244
Related Search	0.0211	0.0207	+1.93%	7e-26
Pagination	0.0226	0.0227	-0.44%	0.0114
Other	0.0518	0.0515	+0.58%	0.0048

Summary

Problem	Method(s) suggested
Shorter experiment time	Bayesian Bandits, Interleaving (under specific conditions), Experimentation Platform
Interim results available	Bayesian Bandits, Interleaving (limited), Experimentation Platform
Dependence on past seasons	User as randomization unit, Adjust for baseline, Interleaving,
Potential cannibalisation between experiments	Causal Inference, Experimentation Platform
Potential interactions between experiments	MVT, Experimentation Platform
Distribution of the primary metric	Model based estimates for zero inflated distribution, another metric

One accurate measurement is worth more than a thousand expert opinions.

–Admiral Grace Hopper


QUESTIONS