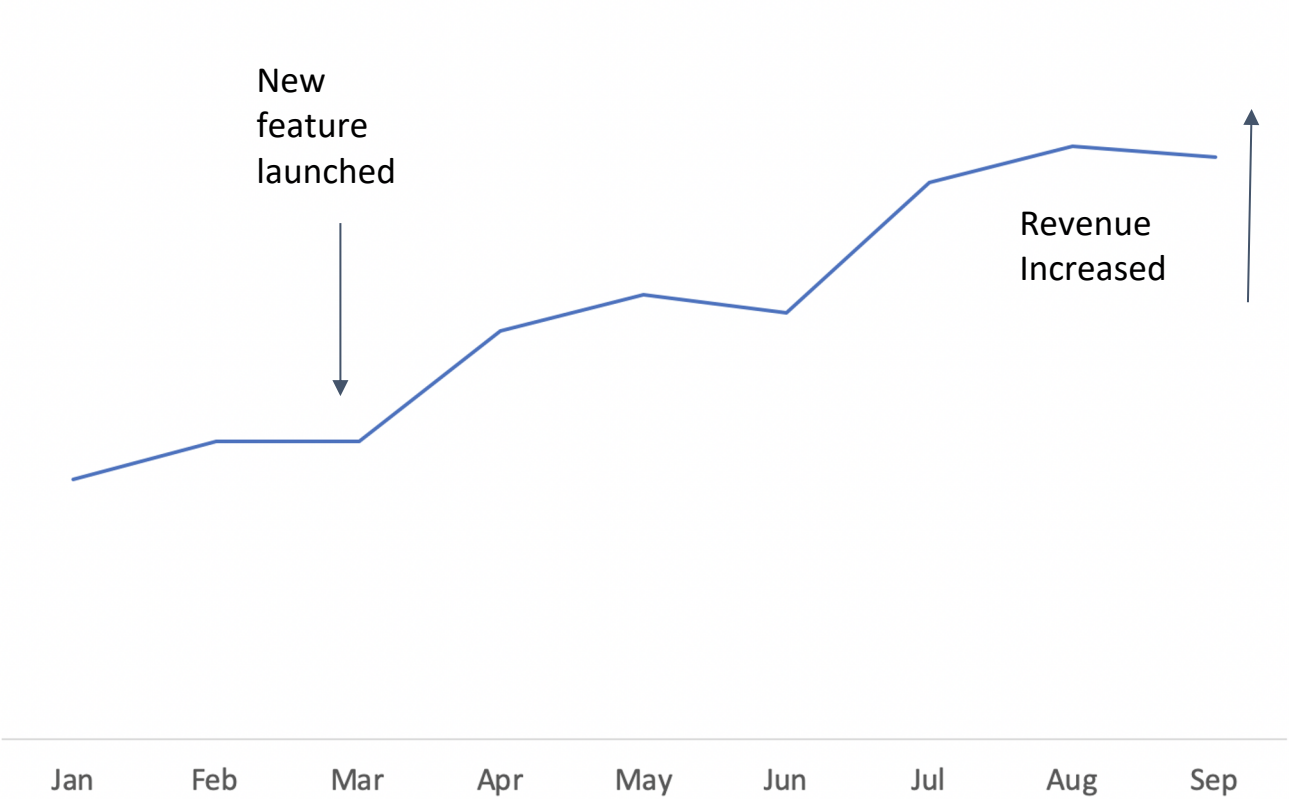
Fundamentals of A/B Testing

Data Application Lab

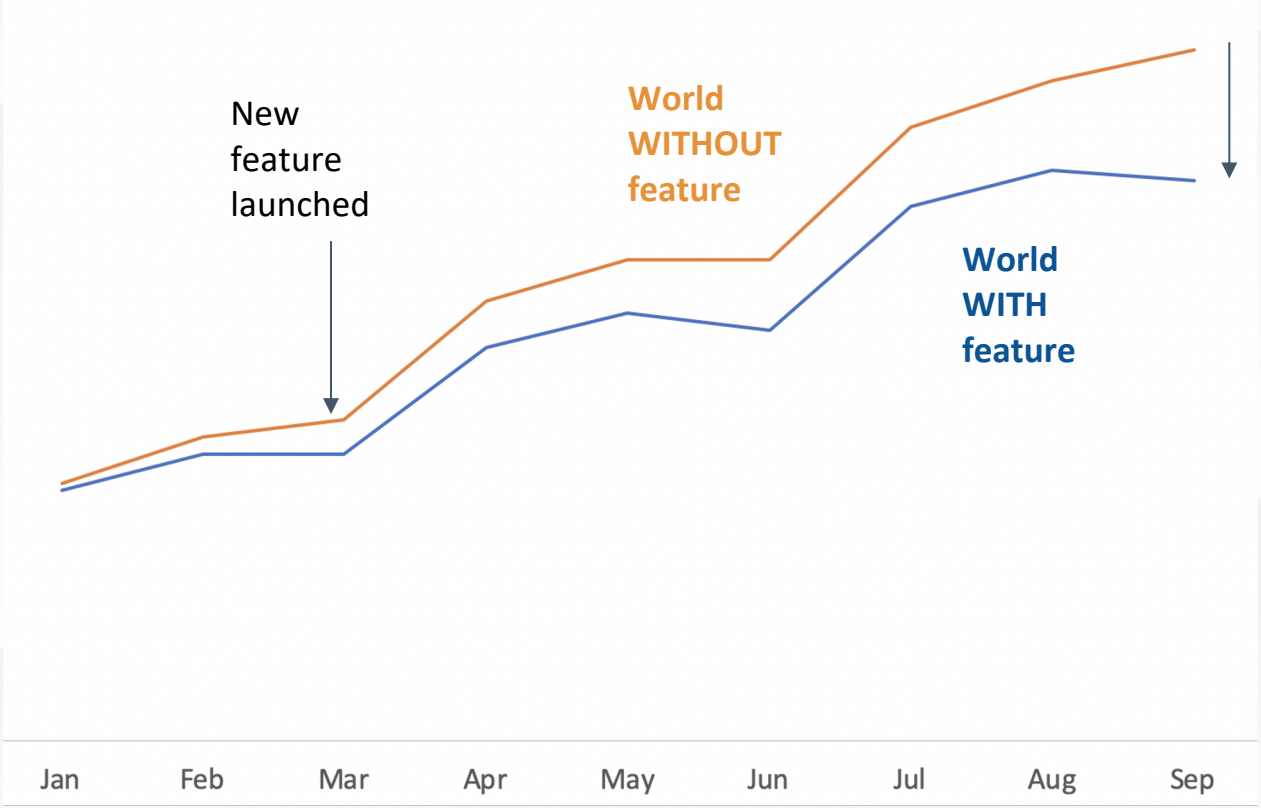
Agenda

- Why and when should you run an A/B test?
- An 8-Step Guide to design your A/B test*
  - Case Study 1: Testing a red vs a green button
  - Case Study 2: Testing a new landing page
  - *: statistical significance level, power, p-value, sample size using R
- How to Analyze A/B Test Results
- Common tools to run A/B tests
- Experimentation Platforms Introduction
- Summary

Why Should You Run A/B Tests?



Why Should You Run A/B Tests?



Why should you run a A/B test?

A/B testing is key to understand what drives your business and make data informed business decisions.

To understand the causal relationship and not simply the correlations.

When to run experiments

- Deciding whether or not to launch a new product or feature
- To quantify the impact of a feature or product
- Compare data with intuition. I.e. to better understand how users respond to certain parts of a product

When not to run experiments

- No clear comparison between control and experimental group
- Emotional changes needs time: Logo/Brand Name
- Response data hard to gather
- Too long to response

An 8-Step Guide to run an Experiment

1. Define your goal and form hypotheses
2. Identify control and treatment groups
3. Identify KPIs to measure
4. Identify what data needs to be collected
5. Determine how small of a difference you would like to detect
6. Determine what fraction of visitors you want to be in the treatment
7. Run a power analysis to decide how much data you need to collect and how long you need to run the test
8. First time trying something new: run an A/A test simultaneously to check for systematic biases

Case Study 1: Testing a red vs a green button

Step 1: Define goals and form hypotheses

Goal: Quantify the impact a different call-to-action button color (from red to green) has on our core metrics.

Hypotheses:

- Compared with a red CTA button, a green CTA button will attract users click more often
- A fraction of these additional clicks will complete the transaction, increasing revenue
- The change in behavior to be most pronounced on mobile (vs desktop, tablet)

Null hypothesis:

- Green button will cause no difference on Click Through Rate or other user behavior.

Step 2: Identify treatment & control

Step 3: Identify KPIs to measure

- **Revenue**
- **Purchase Rate**: Purchase per visitor
- **Click Through Rate**: Clicks per visit or Clicks per visitor

Other behavioral metrics:

- Bounce rate
- Time on site
- Return visits
- Engagement with other parts of the website
- Referrals
- etc

Step 4: Identify data to be collected

- User ID (if user is logged in)
- Cookie ID
- What platform the visitor is on
- Page loads
- Experiment assignment
- If the user sees a button, and which color
- Clicks on the button, and which color
- Data for other metrics (i.e. engagement behavior)

Step 5: Determine how small of a difference needs to detected

Current button CTR: 3%

Successful experiment: 3.3% or more

Order Now

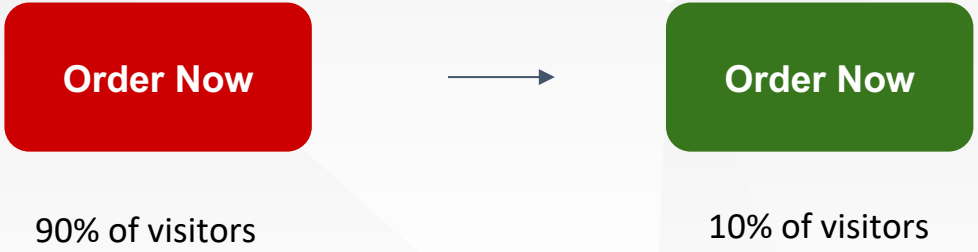
This means that we would need to collect enough data to measure a 10% increase.

Hypothesis Test

My null hypothesis (无效假设) H_0: $d_0 = 0$
there is no difference between the *two treatments(*$d_r - d_g = 0$, rewrite as $d_0 = 0$)

My alternative hypothesis (备择假设) then is: $d_0 \neq 0$
If I know that d_g is for sure better than or worse than d_r then I can choose one-sided test, if not, I will choose two-sided test.

Step 6: Determine what fraction of visitors you want to be in the treatment



50/50 splits are the easiest and take the least amount of time to run, but you might not always want to do that.

Step 7: Calculate how much data to collect and how long to run the experiment

Run a **power analysis** to decide how many data samples to collect depending on your tolerance for:

- Minimum detectable level, aka minimum measurable difference (10% in our case)
- False negative (Type II error)
- False positive (Type I error)

Power calculation in R:

```
power.prop.test(n = NULL, p1 = NULL, p2 = NULL, sig.level = 0.05,
                power = NULL,
                alternative = c("two.sided", "one.sided"),
                strict = FALSE, tol = .Machine$double.eps^0.25)

power.prop.test(p1 = 0.03, p2 = 0.033, sig.level = 0.05, power = 0.8)
```

```
     Two-sample comparison of proportions power calculation

              n = 53210.3
             p1 = 0.03
             p2 = 0.033
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power Analysis, continued...

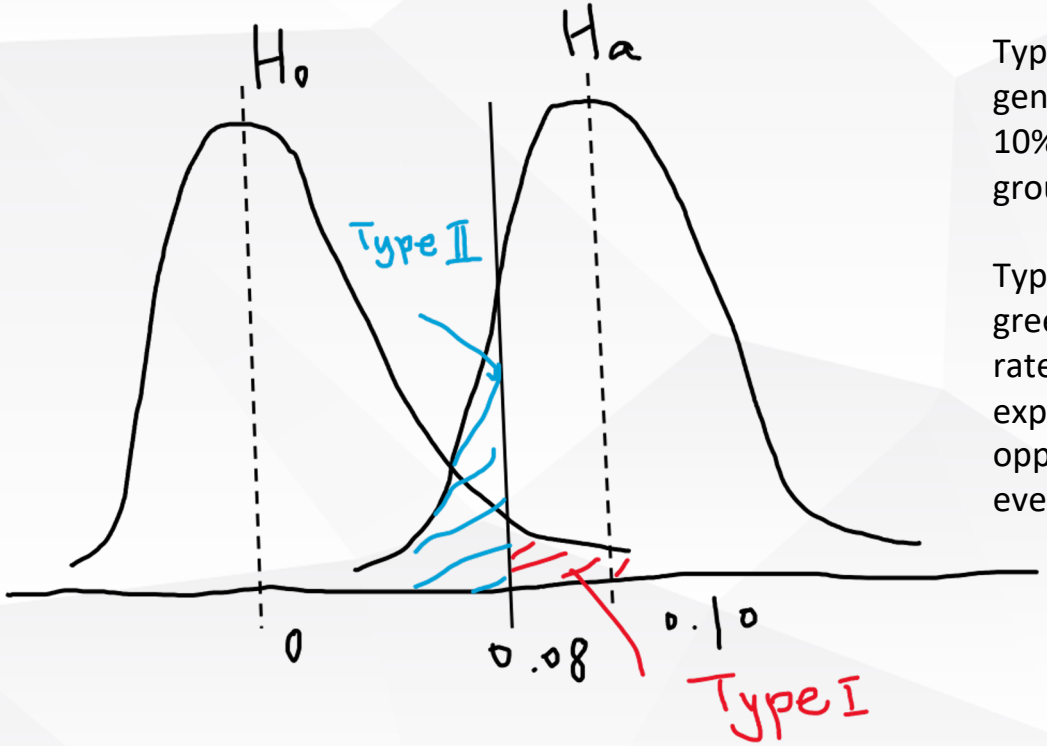
False Positive (Type I error): we see significant result when there isn't one

- Falsely reject the null hypothesis
- Typically, we want a false positive rate of < 5%, $\boldsymbol{\alpha} = 0.05$
- False positive rate is equivalent to the **significant level** of a statistic test

False Negative (Type II error): there is an effect, but we weren't able to measure it

- We should reject the null hypothesis, but we did not.
- Typically, we want a false negative rate of < 20%, $\boldsymbol{\beta} = 0.2$
- The "**power**" of the test is $1 - \boldsymbol{\beta} = 0.8$

Type I and Type II Revisited



Type I: False Positive - we think green button generates a better click through rate by at least 10%, so that we adopted the experimental group, but actually, it does not.

Type II: False Negative - we DIDN'T believe green button generates a better click through rate by 10%, so that we DID NOT adopt the experimental group, so that we lost the opportunity for more clicks (revenue eventually).

Figure out how long to run the test and run the test for AT LEAST that long...

For 50/50 split test:

We need at least 53,210 x 2 = 106,420 users

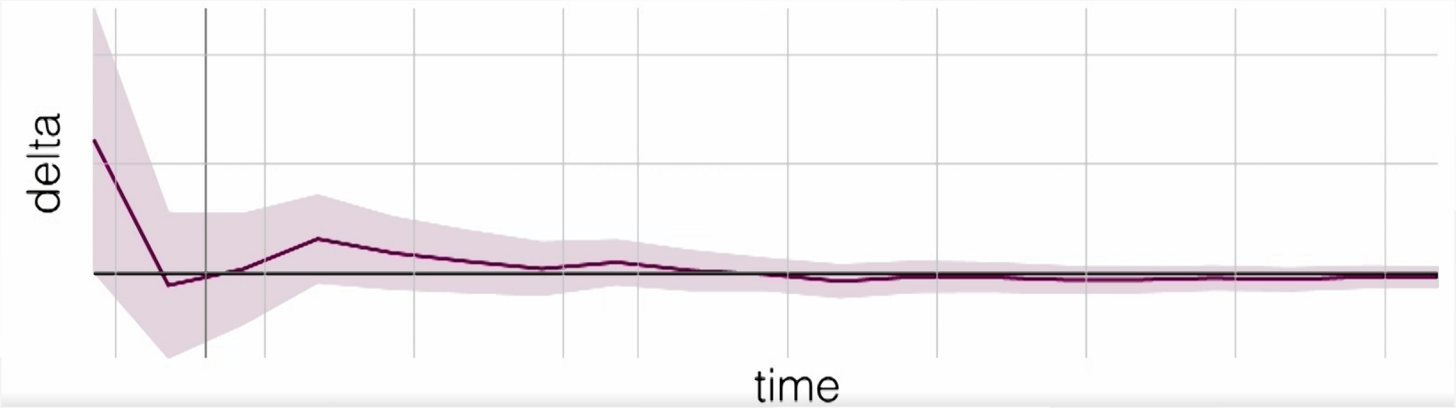
If our traffic is about ~100,000 unique users/day, we need to run the test for at least a day.

Note:

If we were to run a 90/10 test, we would need 53,210 x 10 = 532,100 users, the 90% of 532,100 will go to the 90% of the traffic, so that the duration of test will be around 5 times more than the 50/50 split.

Step 8: If this is your first time, run a dummy test

You should see something like this for your metrics over time:



Quiz

- Which of the followings can be evaluated by A/B testing?
  - A. An online training website to launch premium service
  - B. A retail brand to apply new customer segmentation for email marketing
  - C. A social media site wants to know if changing the color of “Sign up” button helps with conversion
  - D. A e-commerce site wants to know if they are missing some good selling-products
  - E. A career service site wants to know if the new job recommendation algorithm can improve hiring rate
  - F. Video sharing sites wants to test new video recommendation algorithm
  - G. Company wants to change brand name

Case Study 2: Testing a New Landing Page

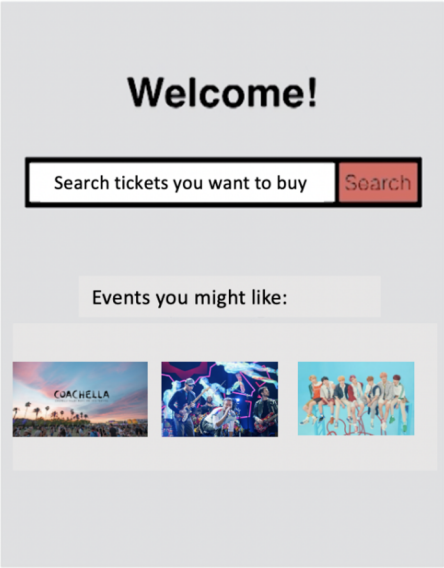
Personalized Recommendation

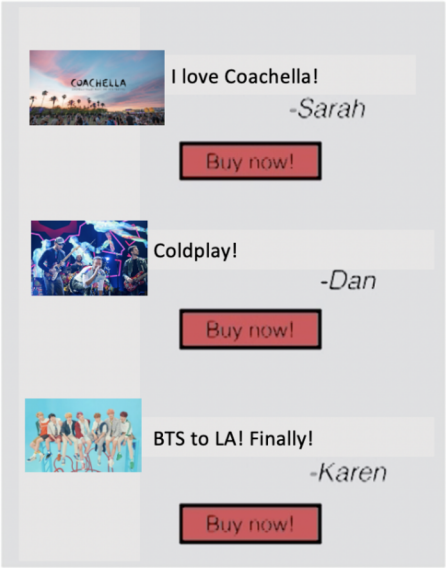
Vanilla version

Home Page

Content Page

vs.





Step 1: Define goals and form hypotheses

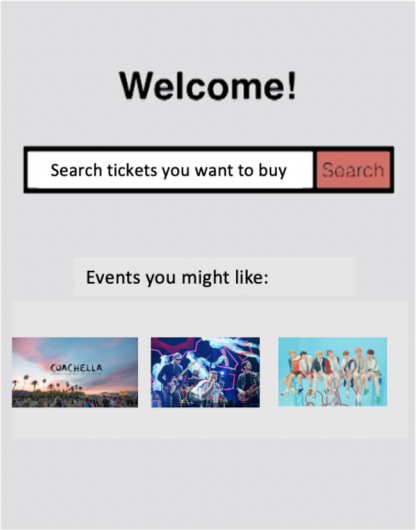
Goal: Identify the impact of the new landing page and recommendation page on conversion and user behavior.

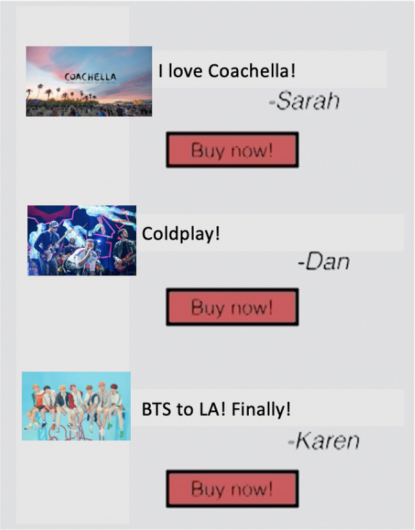
Hypotheses:

- Personalized suggestions on homepage will engage visitors who would have otherwise bounced.
- Relevant information on the next page will inspire visitors to make purchases who would not have otherwise.

Step 2: Identify treatment & control

vs.

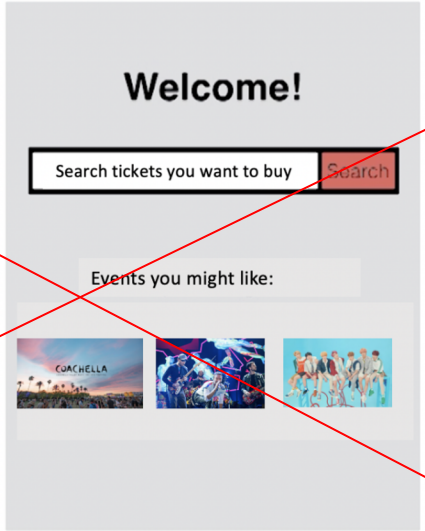


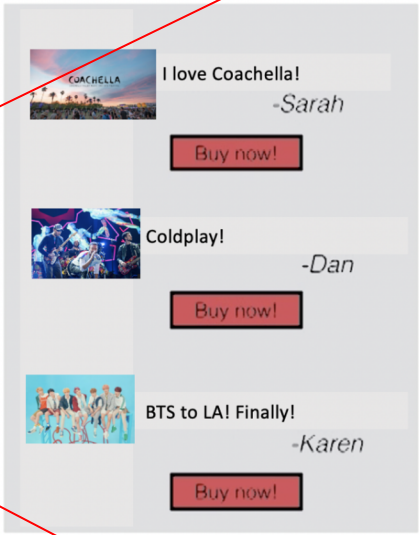


Control:
Visitors who see old landing page

Treatment:
Visitors who see new landing page and new content page

vs.



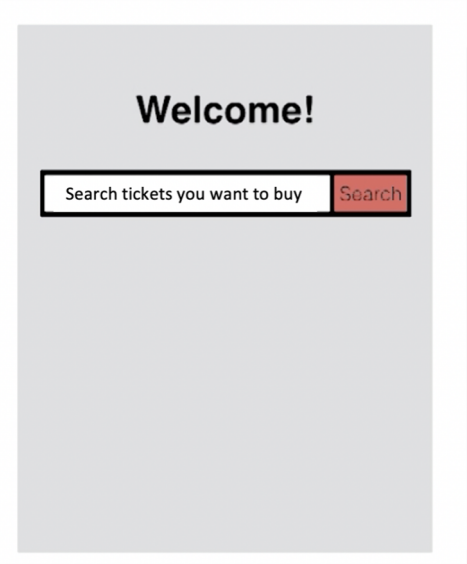


Control:
Visitors who see old landing page

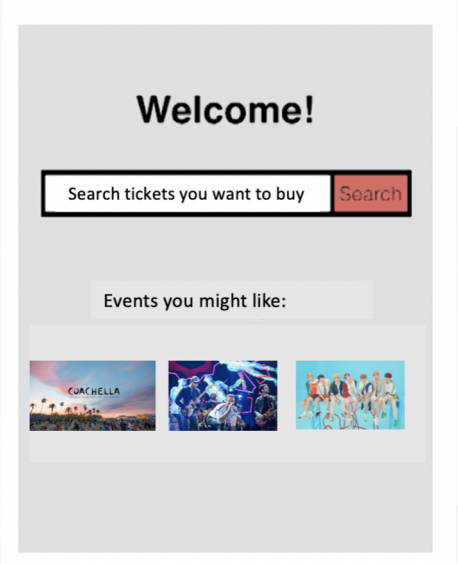
Treatment:
Visitors who see new landing page and new content page

Tip: Only one thing should change between treatment and control

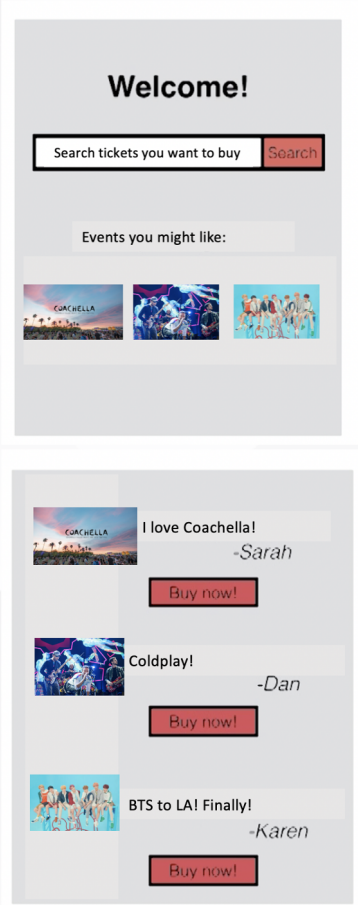
Control:
No personalization



Treatment 1:
Personalized home page



Treatment 2:
Personalized home + content page



Step 3: Identify KPIs to measure

- Between Control & Treatment 1: **Bounce Rate** on home page
- Between Treatment 1 & Treatment 2: **Transactions**
- Keep track of other metrics too for clues:
  - Click through rates
  - Searches
  - Time on site
  - etc

Step 4: Identify data to be collected

- User ID (if user is logged in)
- Cookie ID
- What platform the visitor is on
- Page loads
- Experiment assignment
- If the user clicks on a recommendation and which one
- Any engagement on the content page

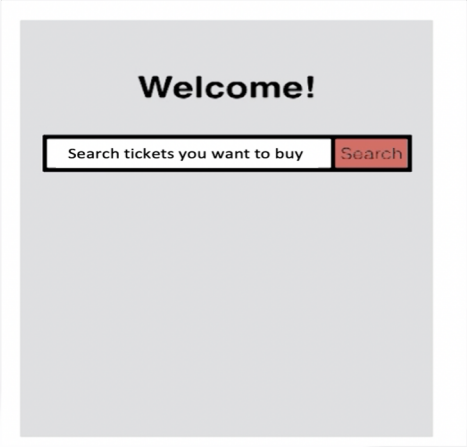
Step 5: Determine how small of a difference needs to detected

Current transactions/visitor: 0.05 (5%)

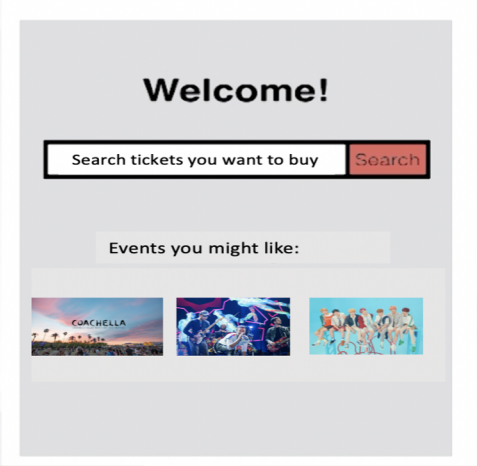
Successful experiment: 0.051 (5.1%)

Step 6: Determine what fraction of visitors you want to be in the treatment

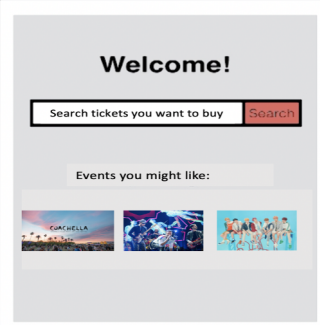
Control: (33%)

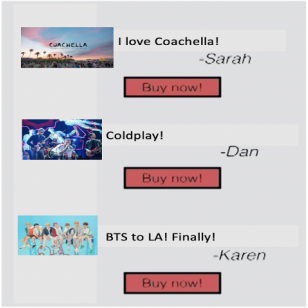


Treatment 1: (33%)



Treatment 2: (33%)





Step 7: Calculate how much data to collect and how long to run the experiment

```
> power.prop.test(p1 = 0.05, p2 = 0.051, sig.level = 0.05, power = 0.8)

     Two-sample comparison of proportions power calculation

              n = 752702.5
             p1 = 0.05
             p2 = 0.051
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

752,702 x 3 = 2.25M users

What if we relaxed significance level to 10%?

```
> power.prop.test(p1 = 0.05, p2 = 0.051, sig.level = 0.1, power = 0.8)

     Two-sample comparison of proportions power calculation

              n = 592903.1
             p1 = 0.05
             p2 = 0.051
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

592,903 x 3 = 1.78M users

Figure out how long to run the test and run the test for AT LEAST that long...

For **α** = 0.05

We need at least 2.25M users

If our traffic is ~ 100,000 unique users/day, we need to run the test for 22.5 days.

For **α** = 0.1

We need at least 1.78M users

If our traffic is ~ 100,000 unique users/day, we need to run the test for 18 days.

Advanced: Multi-armed bandit approach

Problem: We want to find the best variant, so we are willing to let the experiment run for longer. But we also want to maximize revenue during this experiment period.

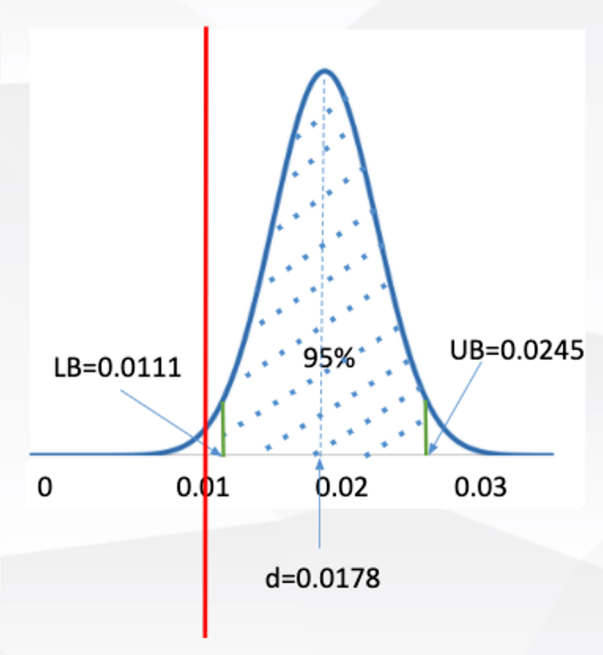
Solution: Adjust fraction of (new) users in treatment/control according to which group seems to be doing better.

Analyze Results

Group	Control (red button)	Treatment (green button)
Unique Visitors	16,242	16,173
Number of Clicks	1,641	1,921
Click Through Rate	10.1%	11.9%

Is Green Button actually better?

Steps to Analyze an A/B Test Results:



Analysis

- Pa = 1641/16242 = 0.1010
- Pb = 1921/16173 = 0.1188
- d = 0.1188 – 0.1010 = 0.0178
- Ppool = (1641+1921)/(16242+16173) =0.1099
- SE = $\sqrt{Ppool * (1 - Ppool) * (\frac{1}{Na} + \frac{1}{Nb})}$ = 0.0034
- m = z*SE = 1.96 * 0.0034 = 0.0067
- Confidence Interval= (d-m, d+m) = (0.0111, 0.0245)

Conclusion

- Lower Bound CI > 0
  => Pass Statistical Significance
- Lower Bound CI > Practical Significance (0.0111 > 0.01)
  => Pass Practical Significance

Note that this analysis is based on the assumption that the distribution is a normal distribution but not a binomial distribution. There is another way to calculate this using binomial distribution which we can cover in future lectures.

Some Popular A/B Testing Tools

Google™ Analytics

- Free
- Allows for easy tracking of business targets
- Raw data – interpretation and use are on you

Optimizely

- Cost-efficient premium option for entry-level users
- Easy, intuitive interface
- Allows for easy multivariate testing
- Mobile-focused testing options
- Less control over analysis

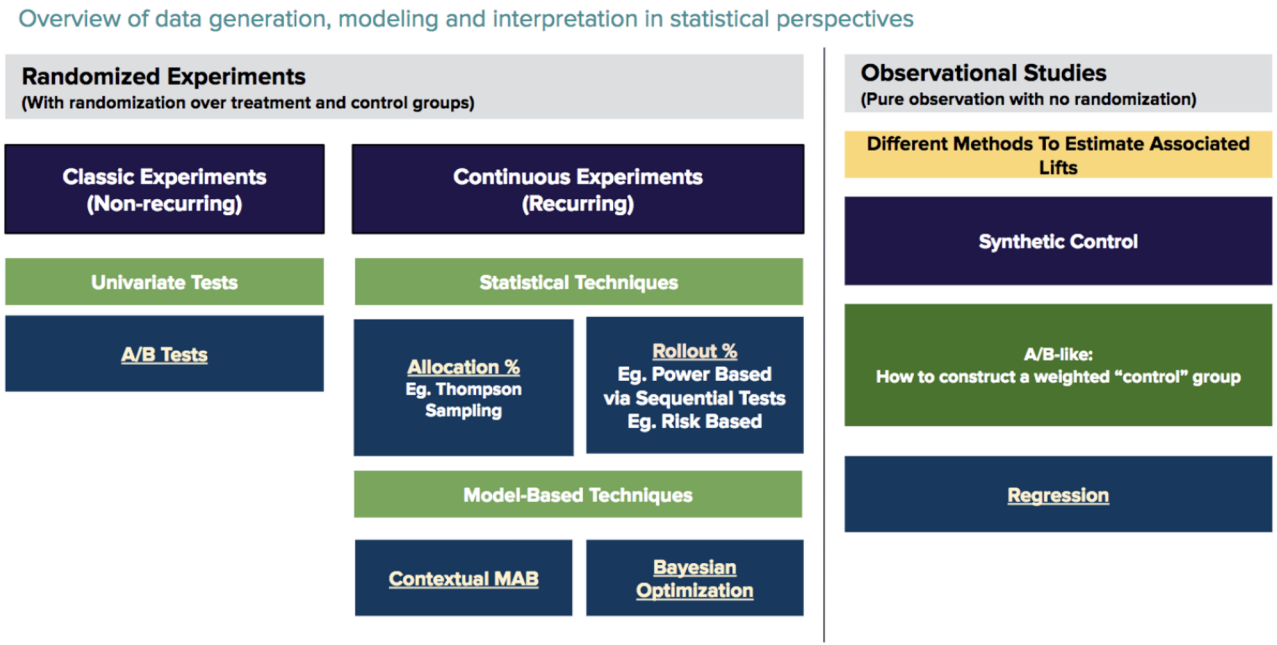
unbounce

- Focuses exclusively on landing pages
- Provides a bunch of templates optimized for conversion
- Great for businesses starting from scratch

Visual Website Optimizer

- Similar to Optimizely, a little more expensive with more features

Uber Experimentation Platform

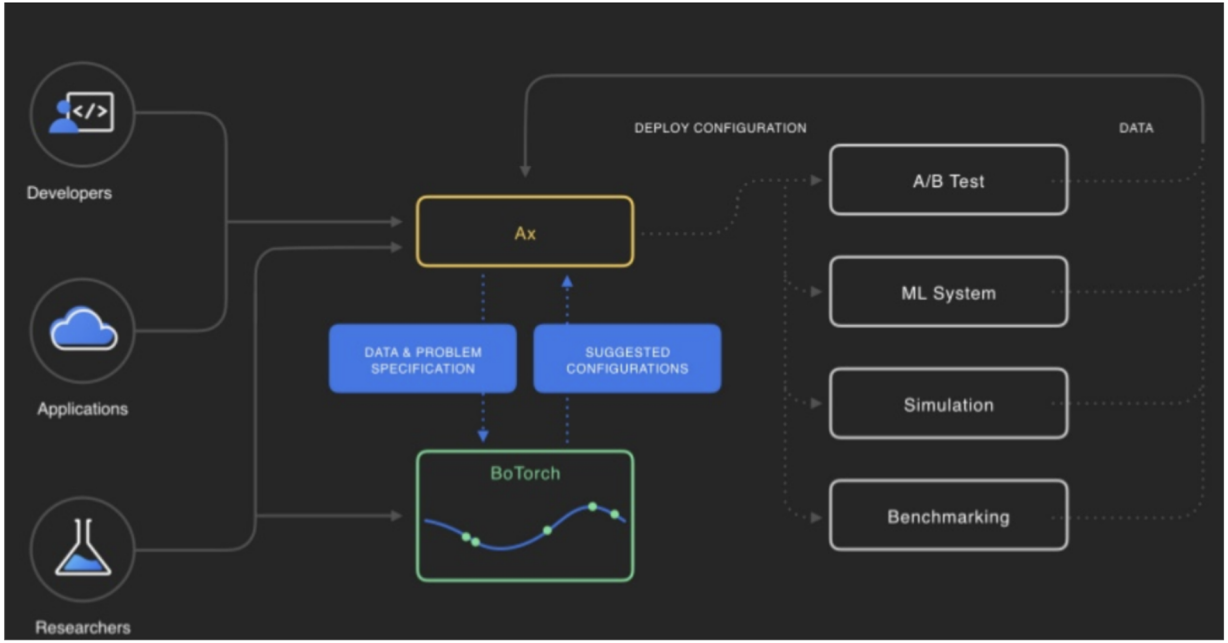


Reference:
[Under the Hood of Uber's Experimentation Platform]

Highlights

- One-size-fits-most methodologies: define decision metrics that can be applied to use cases across the company. Standardized. Reusable.
- Sequential Testing: traditional A/B testing methods inflate Type I error by repeatedly subsampling, so that added another layer as a checkpoint for system outage or other degradations.

Facebook Experimentation Platform



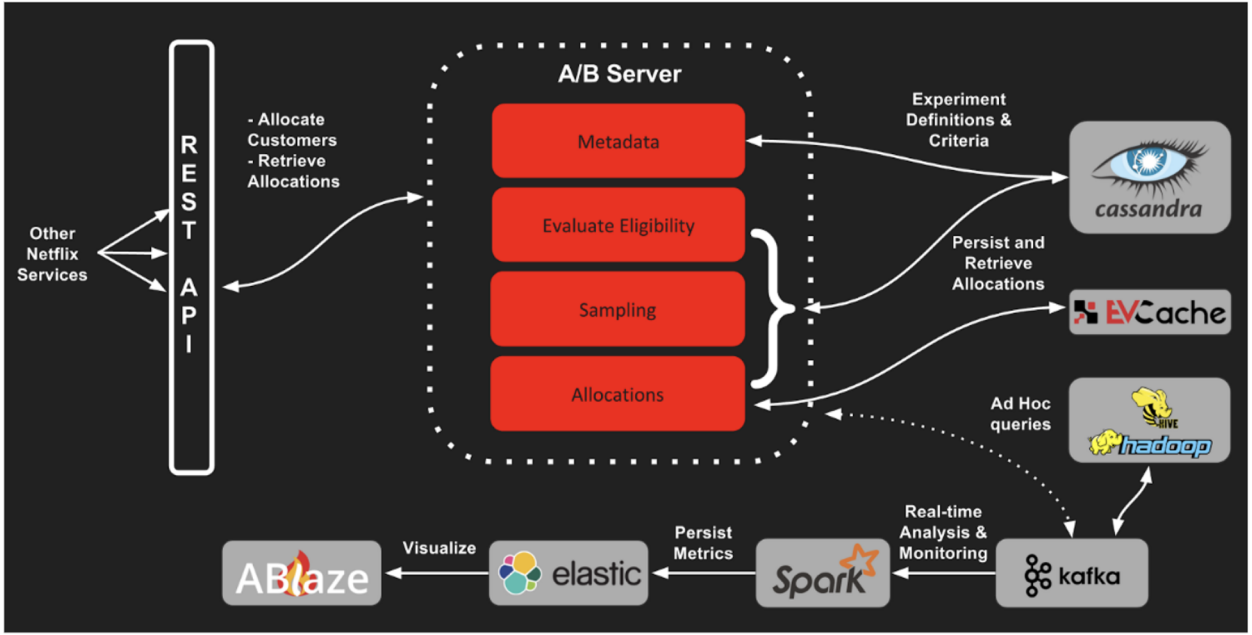
Reference: [Facebook Talk at Netflix ML Platform Meet-up Sep 2019]

Highlights

- They augment online experiments with an offline simulator and apply multi-task Bayesian optimization to tune live machine learning systems, including their recommendation and ranking systems.
- Offline simulators provide a higher-throughout mechanism of generating the estimates while they are not be possible to accurately simulate the interactions between human and an ML system given its nature.
- The solution the FB Adaptive Experimentation team proposes: Bayesian Optimization with the aid of a simulator is a special case of multi-fidelity optimization, in which there is a target function to be optimized along with a collection of cheaper approximations of varying fidelity.

Reference: [Bayesian Optimization for Policy Search via Online-Offline Experimentation]

Netflix Experimentation Platform



Reference: [It's All A/Bout Testing: The Netflix Experimentation Platform]

Highlights

- “Cell”: each member belongs exclusively to one cell within a given experiment, with one of the cells always designated the control group cell.
- Allocation: how members get allocated to a given test
  - Batch: allows analysts to populate tests using custom queries.
  - Real-time: allows analysts to configure rules which are evaluated as the user interacts with Netflix.
- Stratified Sampling: Netflix would like to have each cell contains miliar proportions of members from each country, device type etc.
- Caching: they persist allocation details in Cassandra and invalidate the A/B caches associated with this member.
- Ad-hoc Analysis: publish feed to Hive tables through Kafka data pipeline.
- Near real-time display of A/B test results: use Spark streaming to ingest and transform data from Kafka streams before persisting them in ElasticSearch.

Airbnb Experimentation Platform

Metric (Filter metrics...)	Global Coverage	control Mean	treatment Mean	Percent Change	P-Value	MDE
Metric 1 target metric	97.14%	0.05 430k/9.29M	0.05 429k/9.29M	0% neutral	0.51	0.73%
Metric 2 target metric		0.6 258k/430k	0.6 258k/429k	▾ 0.05% ± 0.35	0.80	2.4%
Metric 3 core metric	96.54%	0.17 1.58M/9.29M	0.17 1.59M/9.29M	0.35% ± 0.85	0.42	1.2%
Metric 4 core metric	95.28%	0.06 577k/9.29M	0.06 580k/9.29M	0.41% ± 1.4	0.58	2.1%
Metric 5 core metric		0.33 140k/430k	0.33 140k/429k	▾ 0.32% ± 0.84	0.32	4.3%
Metric 6 core metric	96.33%	2.5 23M/9.29M	2.5 23M/9.29M	▾ 0.00% ± 0.75	1.00	1.1%
Metric 7 core metric		0.32 566k/1.75M	0.32 565k/1.74M	0.62% ± 0.58	★☆☆ 0.04	1.9%
Metric 8 core metric	42.47%	0.00 2.57k/9.29M	0.00 2.55k/9.29M	▾ 0.82% ± 5.5	0.77	7.8%
Metric 9 core metric See tickets , See issue breakdown	77.13%	0.00 23.4k/9.29M See tickets	0.00 23.4k/9.29M See tickets	▾ 0.04% ± 1.3	0.97	2.7%
Metric 10 core metric	85.01%	0.03 319k/9.29M	0.03 320k/9.29M	0% neutral	0.08	0.69%
Metric 11 1 day behind		4.7 1.56M/331k	4.7 1.54M/328k	▾ 0.05% ± 0.03	★★☆ < 0.01	0.24%

Reference: [Scaling Airbnb's Experimentation Platform]

Highlights

- Experiment Reporting Platform
  - Metric Hierarchy: Core Metrics (predefined list of critical business metrics); Target Metrics (specifically for each experiment); Certified metrics
  - Dimensional Cuts: User Attributes, Event Attributes.
  - Pipeline Logic and Entities: to organize data to minimize computational requirements and prevent duplicate processing.
  - Monitoring: plotting p-values over time etc

Summary

- A/B testing is key to understand what drives your business and make data informed business decisions.
- An 8-Step Guide to start your A/B test
- Case Study 1: Testing a red vs a green button
- Case Study 2: Testing a new landing page
- Common tools that run A/B tests

8 Steps Guide to start your A/B test

1. Define goals and form hypotheses
2. Identify control and treatment groups
3. Identify KPIs to measure
4. Identify what data needs to be collected
5. Determine how small of a difference you would like to detect
6. Determine what fraction of visitors you want to be in the treatment
7. Run a power analysis to decide how much data you need to collect and how long you need to run the test
8. First time trying something new: run an A/A test simultaneously to check for systematic biases

Post Class Poll

Multiple Choice: what are the below topics you are also interested related to A/B testing:

a. More in depth knowledge and techniques in statistics (Bayesian Inference, Monte Carlo Simulations etc)
b. Programming: How to program A/B test calculator in Python
c. Advanced Topics in A/B testing: Scale A/B test in enterprise platforms, Multi-Armed Bandit, Bayesian Optimization etc.
d. More about business process and knowledge in A/B testing
e. Interview questions related to A/B testing
f. Hands on practice on A/B testing

Thank you

APPENDIX

$$n = \frac{2(\bar{p})(1-\bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_B - p_A)^2}$$

Z_β : z-score that corresponds to the level of statistical power
$Z_{\alpha/2}$: z-score that corresponds to the level of significance or confidence level

$\bar{p}$: pooled probability or average of p_A and p_B
p_A : success rate of control group
p_B : success rate of test group

References

- *A/B Testing, A Data Science Perspective, An Introduction to Data and Statistics for Improved U/X - Lisa Qian*
- *Under the Hood of Uber's Experimentation Platform - Anirban Deb, Suman Bhattacharya, Jeremy Gu, Tianxia Zhou, Eva Feng, and Mandie Liu*