What We're Getting Wrong about Running Experiments



Eric Metelka
Head of Product, Eppo

LENNY & FRIENDS SUMMIT





WHY EXPERIMENT?

THE FASTEST GROWING COMPANIES ARE HIGHLY EXPERIMENTAL

airbnb
cameo
_zapier
coinbase
Uber
ClickUp
twitch
perplexity
STITCH FIX
GOLD BELLY
descript
DRAFTKINGS
LinkedIn
miro
zalando
Delivery Hero



WINNING COMPANIES ADAPT IN TIMES OF CHANGE BY BEING EXPERIMENTAL

IT'S NOT JUST A PROCESS, IT'S A CULTURE



AGENDA

1. We’re not using rigorous science
2. We’re not running enough experiments
3. We’re not making big enough changes
4. One more thing...





ISSUE 1: WE'RE NOT USING RIGOROUS SCIENCE

IF WE CUT CORNERS ON THE SCIENCE, WE DIG OURSELVES INTO A HOLE WHERE OUR RESULTS WON'T BE TRUSTED

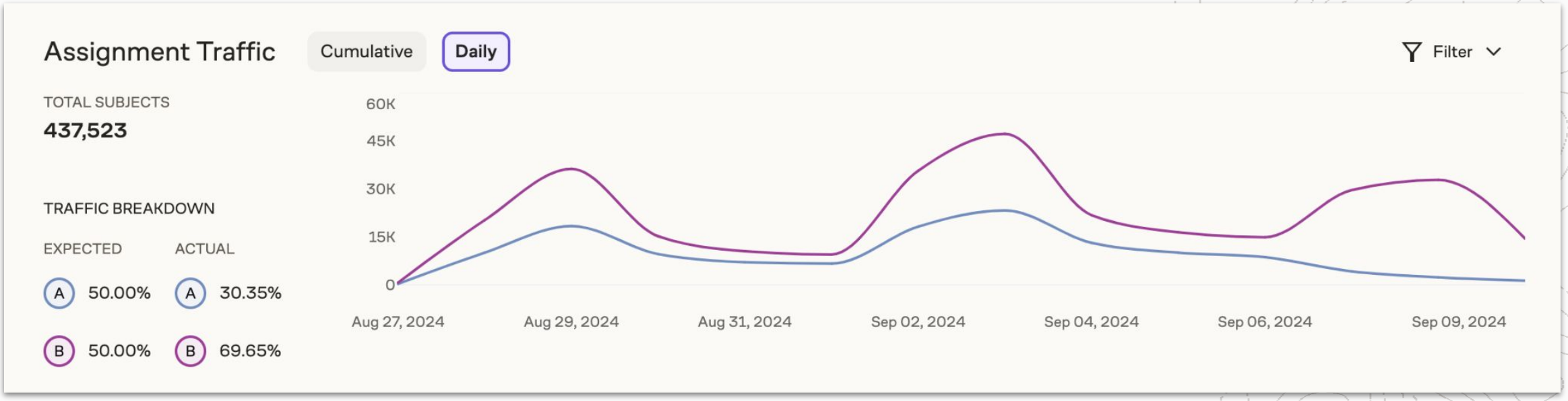


WHAT DO WE NEED TO RUN RIGOROUS EXPERIMENTS?

Siloed data that lives in a proprietary tool	Trustworthy data that matches what we see in other tools
Imbalanced traffic discovered late	Balanced traffic discovered early
Drawing wrong conclusions from early results	Making conclusions from powered experiments

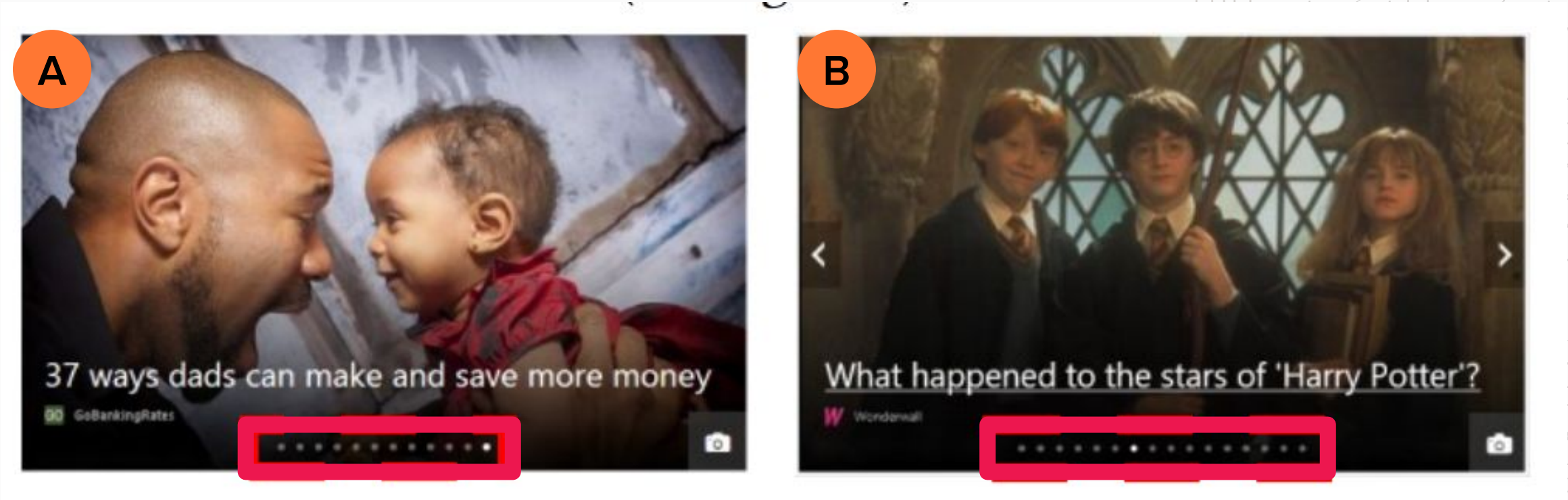


TRAFFIC IMBALANCES ARE THE MOST PERVASIVE REASON RESULTS AREN'T TRUSTWORTHY...AND THEY WASTE TIME



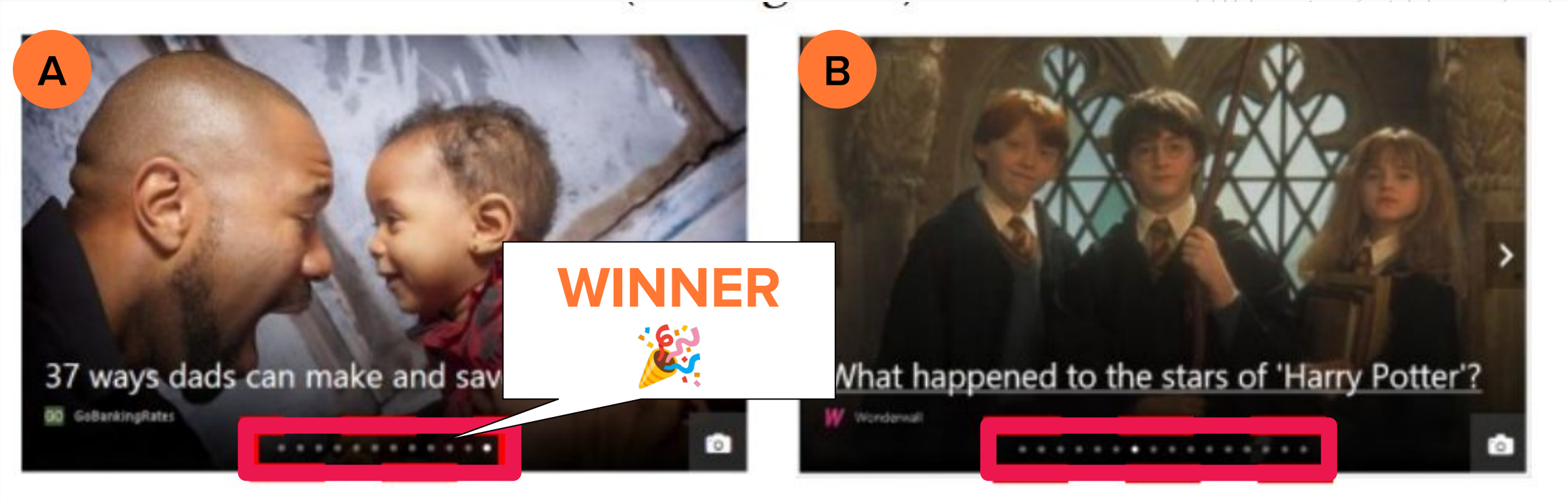


MSN RAN A TEST ON THEIR IMAGE CAROUSEL



Source: Diagnosing Sample Ratio Mismatch in A/B Testing by Aleksander Fabijan et al.

MSN RAN A TEST ON THEIR IMAGE CAROUSEL



Source: Diagnosing Sample Ratio Mismatch in A/B Testing by Aleksander Fabijan et al.

THE USERS EXPOSED TO VERSION B ENGAGED WITH IT ENOUGH TO CONFUSE A BOT DETECTION ALGORITHM WHICH THEN FILTERED THEM OUT OF THE ANALYSIS



PMS LOVE TO LOOK AT EXPERIMENTS

This is called the peeking problem



SIMPLE TESTS ASSUME NO ONE WILL PEEK AT RESULTS DURING THE EXPERIMENT RUNTIME

Certified Core User metrics

Metrics	A control	B	Lift	-4% -2% 0% 2% 4%
AOV (Certified)	36.58	36.75	0.47%	
Total Revenue (Certified)	36.75	37.03	0.78%	
Purchase Conversion Rate (Certified)	0.99	0.99	0.11%	



TO FIX THIS, WE WANT TO RUN SEQUENTIAL TESTS

Test Version	Visitors	Conversions	Conversion Rate	% Lift
Control	82	3	3.66%	
Variant	75	12	16.00%	337%



TO FIX THIS, WE WANT TO RUN SEQUENTIAL TESTS

Test Version	Visitors	Conversions	Conversion Rate	% Lift
Control	82	3	3.66%	
Variant	75	12	16.00%	337%



Relative minimum detectable effect (MDE) per metric over time

Add...

Experiment Metrics

Metric	1 week 932 subjects	2 weeks 2.09K subjects	4 weeks 4.43K subjects	6 weeks 6.96K subjects	8 weeks 9.87K subjects	10 weeks 13.3K subjects
Total revenue	67.23%	26.83%	14.84%	9.20%	6.37%	4.98%
Total purchase count	64.30%	26.05%	14.40%	8.84%	6.07%	4.76%

Calculate by experiment run time

Based on your previous data, running this experiment for 1 week will enroll approximately **932** per variation. With this sample size, you'll likely be able to detect a **67.23%** change in Total revenue.



UNDERSTAND WHAT METHODOLOGY YOU'RE USING AND PEEK ACCORDINGLY





ISSUE 2: WE'RE NOT RUNNING ENOUGH EXPERIMENTS

“What you really want to do company-wide is **maximize the number of experiments you can do** per given unit of time. Since the outcomes of all these things are uncertain, if you can figure out how to conduct an experiment, you can make more bets. **So the key, really, is reducing the cost of the experiments**.”

– Jeff Bezos



ONLY A MINORITY OF EXPERIMENTS ARE SUCCESSFUL

Company/Source	Success rate
Microsoft	33%
Bing	15%
Booking.com, Google Ads, Netflix	10%
Airbnb Search	8%

Source: A/B Testing Intuition Busters © Kohavi, Deng, Vermeer 2022



WE CAN FORECAST THE NUMBER OF EXPERIMENTS YOU NEED TO RUN - IT'S MATH!

$$\text{Number of experiments} = \frac{\text{Relative impact goal}}{\text{Success rate} \times \text{Avg effect size} \times \text{Shrinkage}}$$



WE CAN FORECAST THE NUMBER OF EXPERIMENTS YOU NEED TO RUN - IT'S MATH!

$$13.8 = \frac{5\%}{29\% \times 2.5\% \times .5}$$

Number of experiments

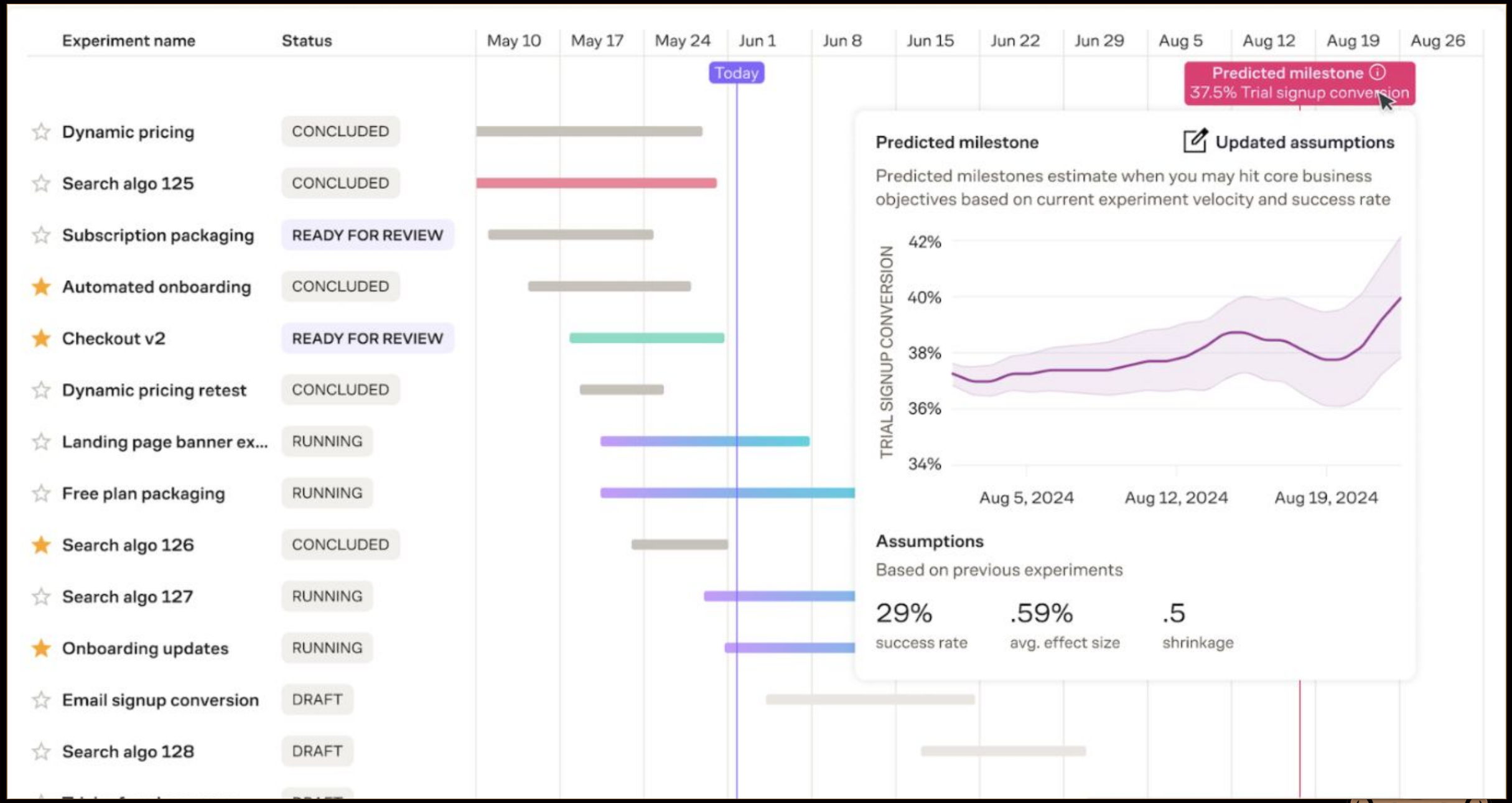
Relative impact goal

Success rate

Avg effect size

Shrinkage



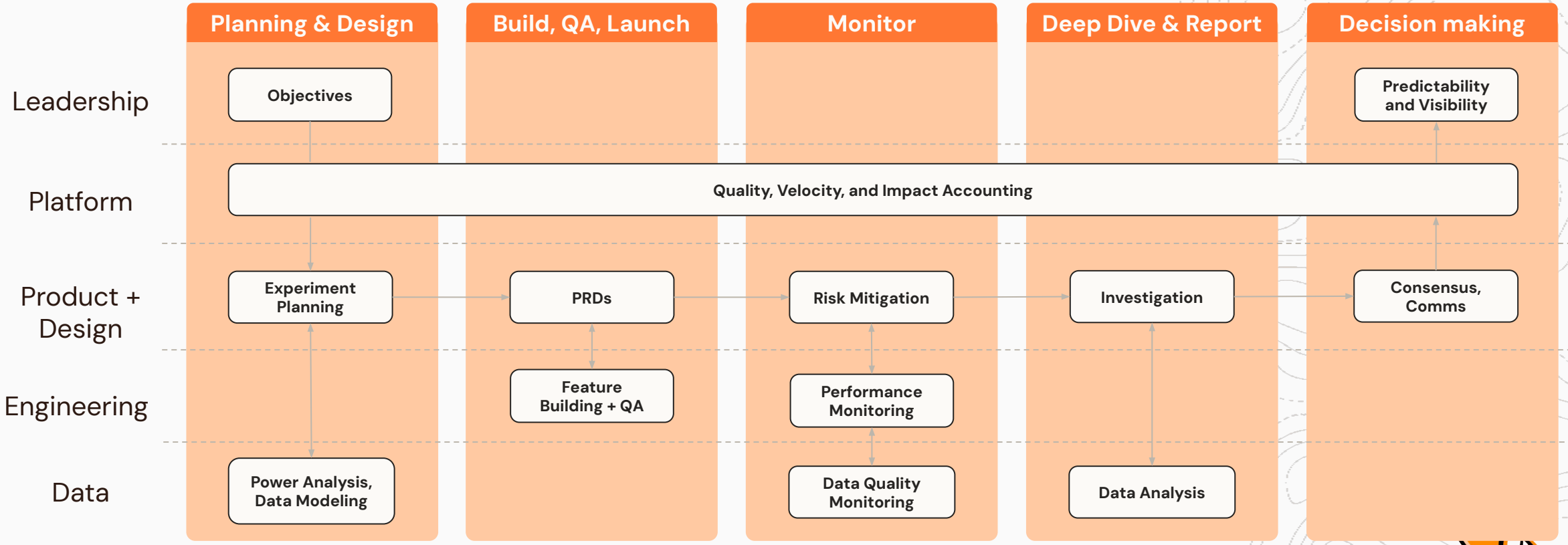

Experiment name
Status
May 10
May 17
May 24
Jun 1
Jun 8
Jun 15
Jun 22
Jun 29
Aug 5
Aug 12
Aug 19
Aug 26
Today
Predicted milestone
37.5% Trial signup conversion
Dynamic pricing
CONCLUDED
Search algo 125
CONCLUDED
Subscription packaging
READY FOR REVIEW
Automated onboarding
CONCLUDED
Checkout v2
READY FOR REVIEW
Dynamic pricing retest
CONCLUDED
Landing page banner ex...
RUNNING
Free plan packaging
RUNNING
Search algo 126
CONCLUDED
Search algo 127
RUNNING
Onboarding updates
RUNNING
Email signup conversion
DRAFT
Search algo 128
DRAFT
Predicted milestone
Updated assumptions
Predicted milestones estimate when you may hit core business objectives based on current experiment velocity and success rate
TRIAL SIGNUP CONVERSION
42%
40%
38%
36%
34%
Aug 5, 2024
Aug 12, 2024
Aug 19, 2024
Assumptions
Based on previous experiments
29%
success rate
.59%
avg. effect size
.5
shrinkage

“The most useful thing Booking.com did was figure out ways to allow non-experts to run experiments and understand statistical thinking.”

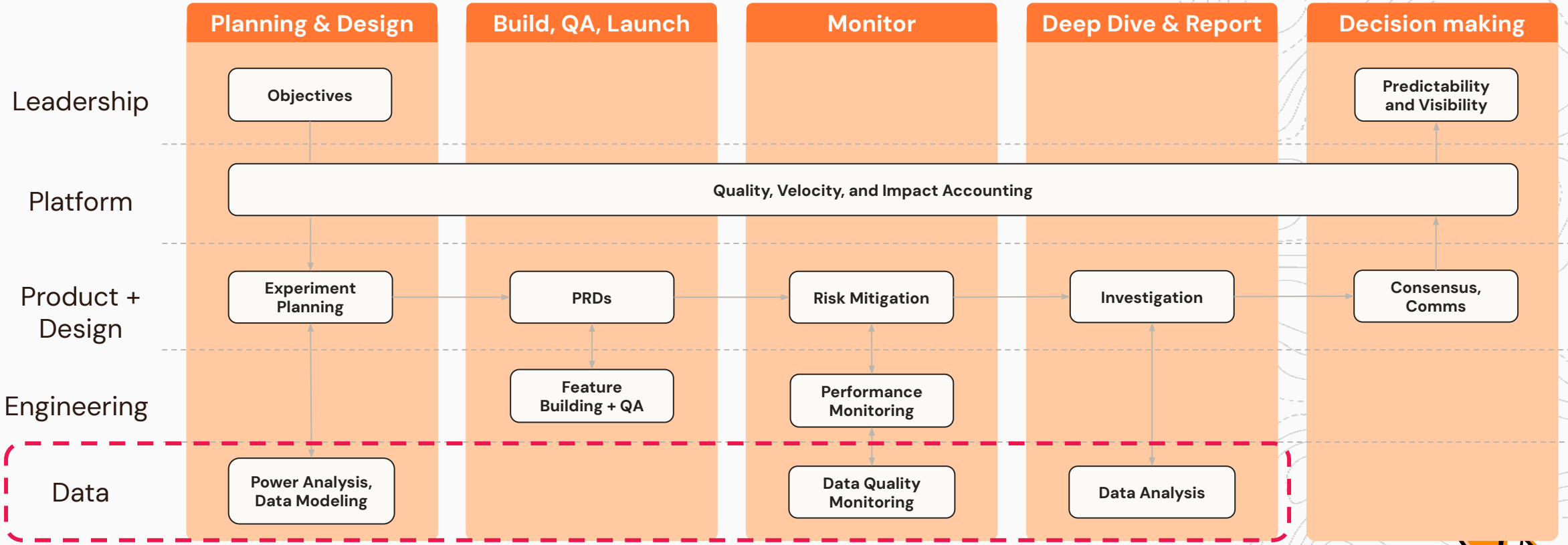
– Lukas Vermeer, Director of Experimentation at Booking.com



TO ACHIEVE A GOOD RUN RATE, WE NEED TO REMOVE BOTTLENECKS



REMOVING BOTTLENECKS AROUND DATA SCIENCE TENDS TO BE HIGHEST IMPACT



Create Experiment Analysis

1. Analysis Set Up
2. Decision Metrics
3. Recommended Decision

2. Select the metrics that will factor into your decision criteria

These are the metrics that influence the end roll out vs roll back decision when the experiment is ready for review.

* Primary Metric

This is the metric that measures the success of the change being tested.

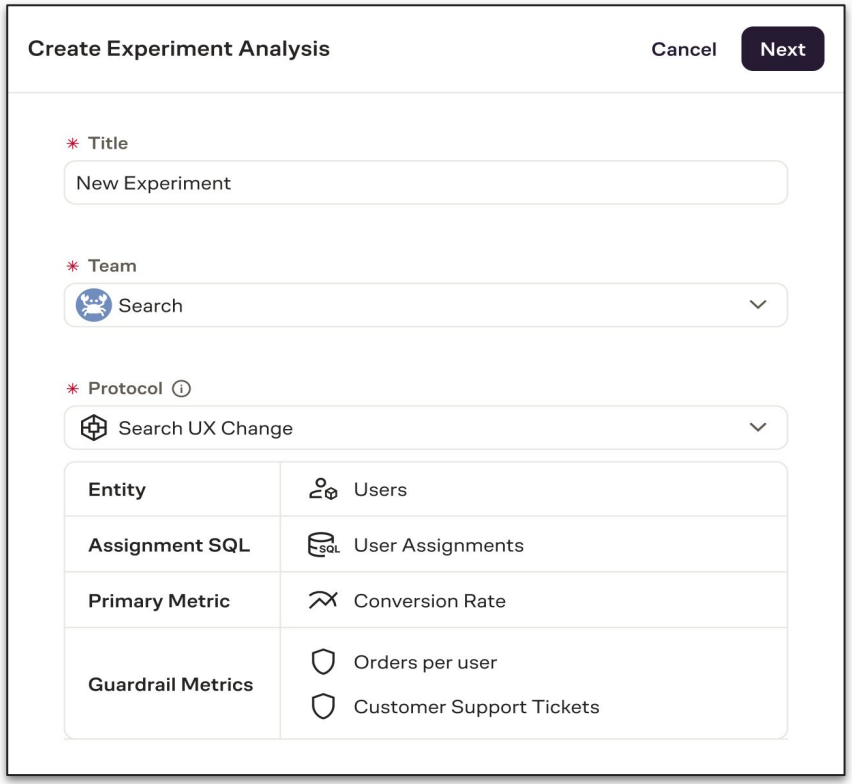
Add a Primary Metric

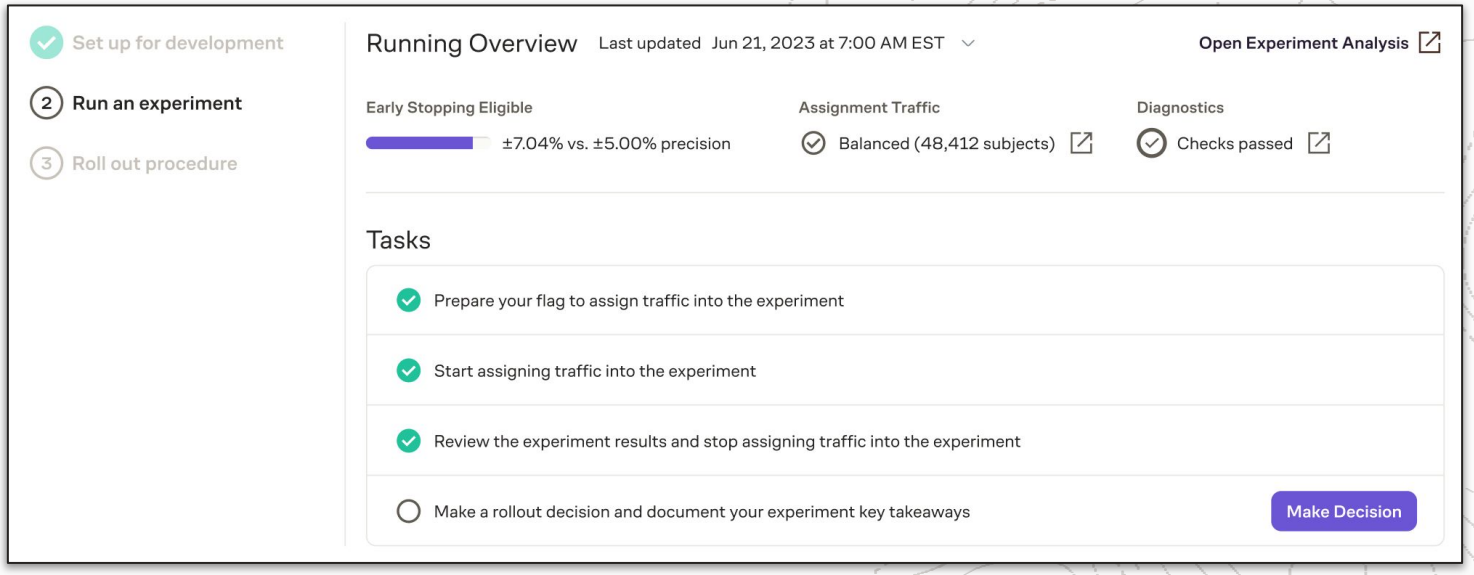
Guardrail Metrics

These metrics check that the change doesn't unintentionally cause negative side effects.

Add Metrics

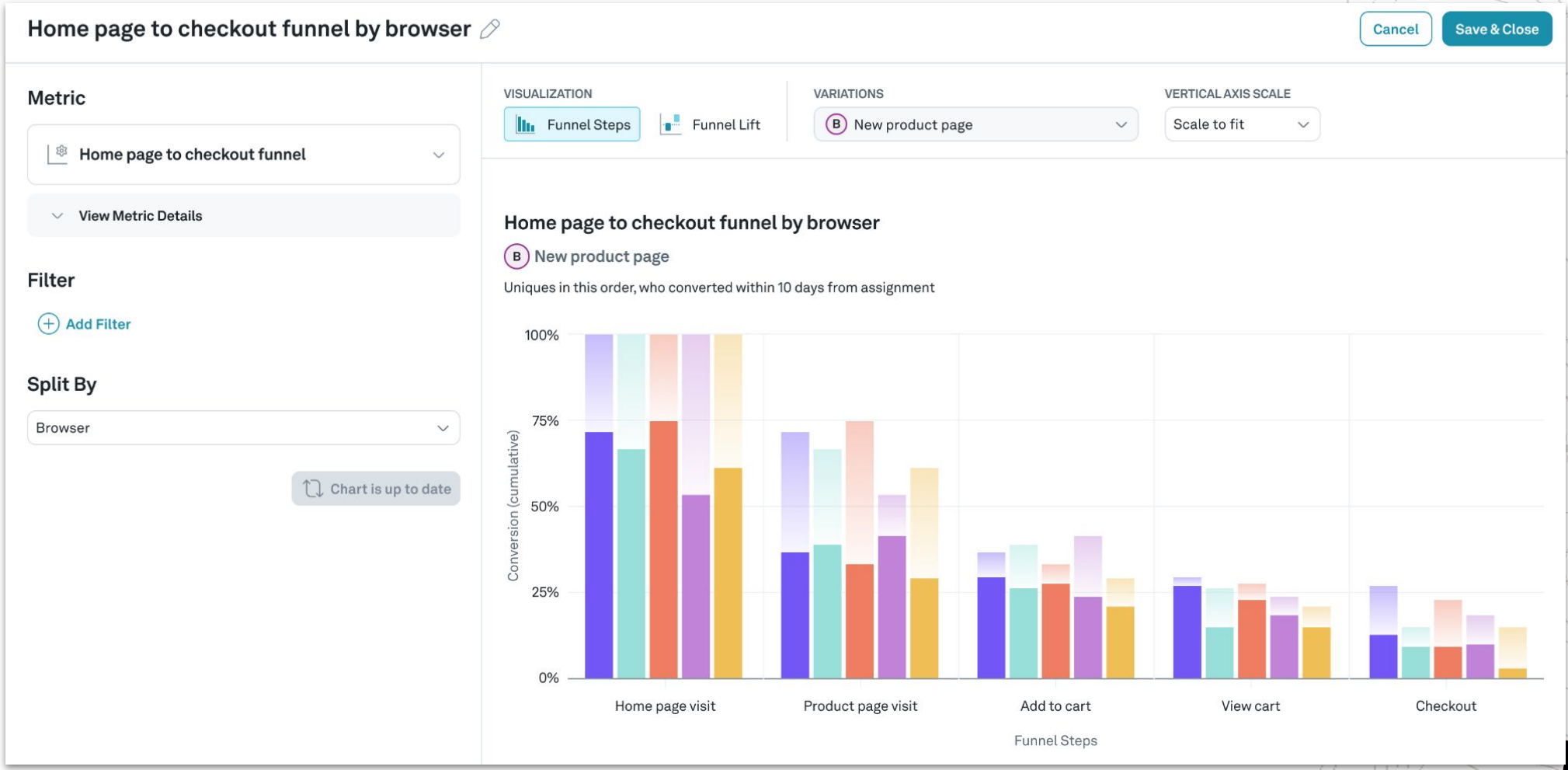
INPUTS: CREATE REPEATABLE DEFAULTS







OUTPUTS: SELF-SERVICE ANALYSIS TOOLS





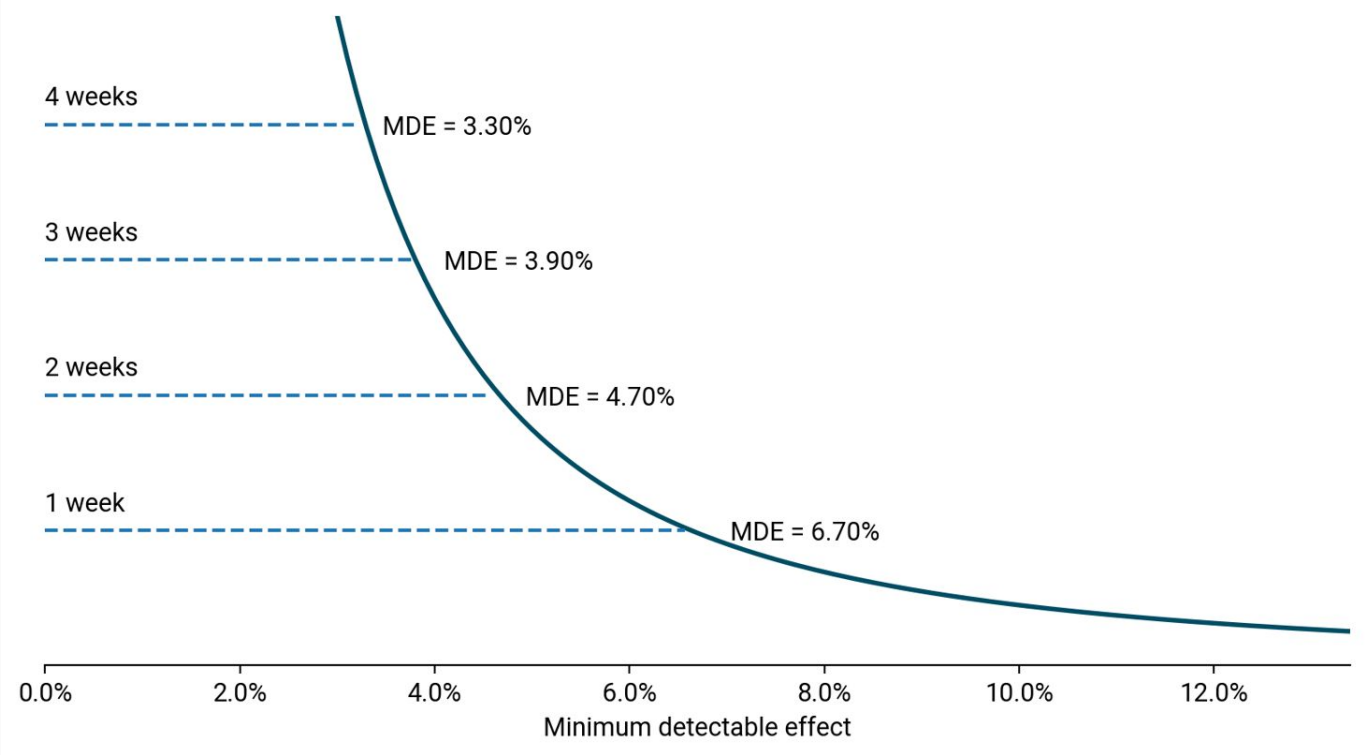
ISSUE 3: WE'RE NOT EXPERIMENTING WITH BIG ENOUGH CHANGES

WE NEED TO BE TESTING ALL THE THINGS!



WE LIKELY DON'T HAVE THE SAMPLE SIZE (OR TIME!) TO DETECT SMALL CHANGES

It takes 4 times longer to detect an effort twice as small

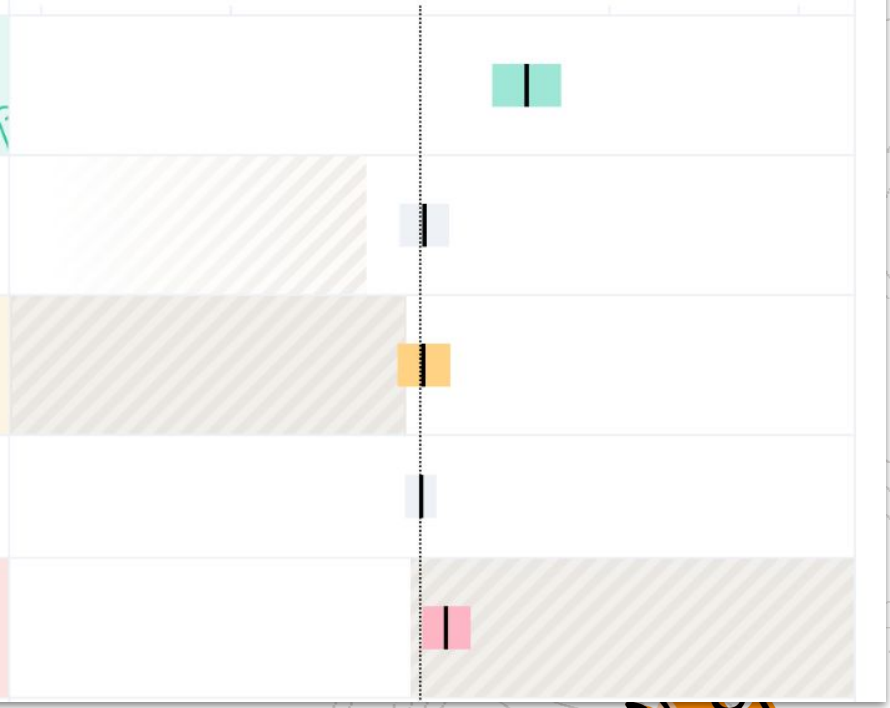




GUARDRAIL THRESHOLDS WARN ABOUT DOWNSIDE RISK

Experiment metrics

Metrics		A control	B	Lift
Total revenue	PRIMARY	137.07	175.92	28.34%
Conversion rate	GUARDRAIL	20.52%	20.80%	1.37%
Total purchase count	GUARDRAIL	134.98	136.53	1.15%
Total visits		6.95	6.99	0.46%
Total support tickets	GUARDRAIL	0.29	0.32	7.20%





“The thing Netflix focuses on is automated, continuous testing. Once we know a **product area is sensitive** and we can move metrics there, [we test] as rapidly and as much as possible.”

– Martin Tingley, Head of Experimentation Platform at Netflix



WE NEED TO KEEP A LOG OF EXPERIMENTS RUN AND THEIR OUTCOMES

Experiment Reports

Sort: Popularity

Filter by keyword | All Owners | All Primary Metrics | Ready for Review, Concluded | My Starred

New Pricing

152 views | Share

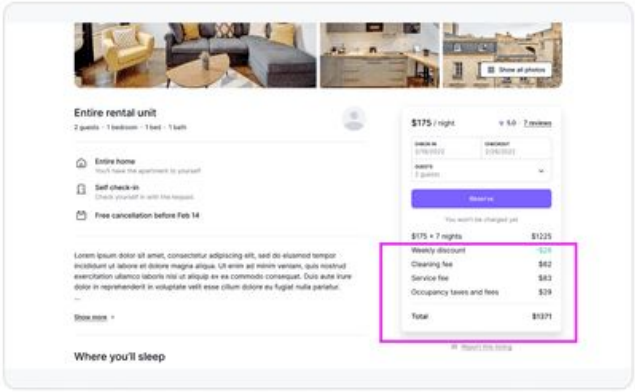
DATES Sep 02, 2023 - Sep 16, 2023

OWNER Josh Hanson

DECISION (B) Show total fees on product page

PRIMARY METRIC Purchase Conversion within 1 Day – 11.69%

KEY TAKEAWAYS The New Pricing experiment yielded significant results, revealing an impressive 11.69% increase in the primary metric of 1-day purchase conversion. This outcome strongly supports the hypothesis that displaying total fe...



New Search Algorithm

123 views | Share

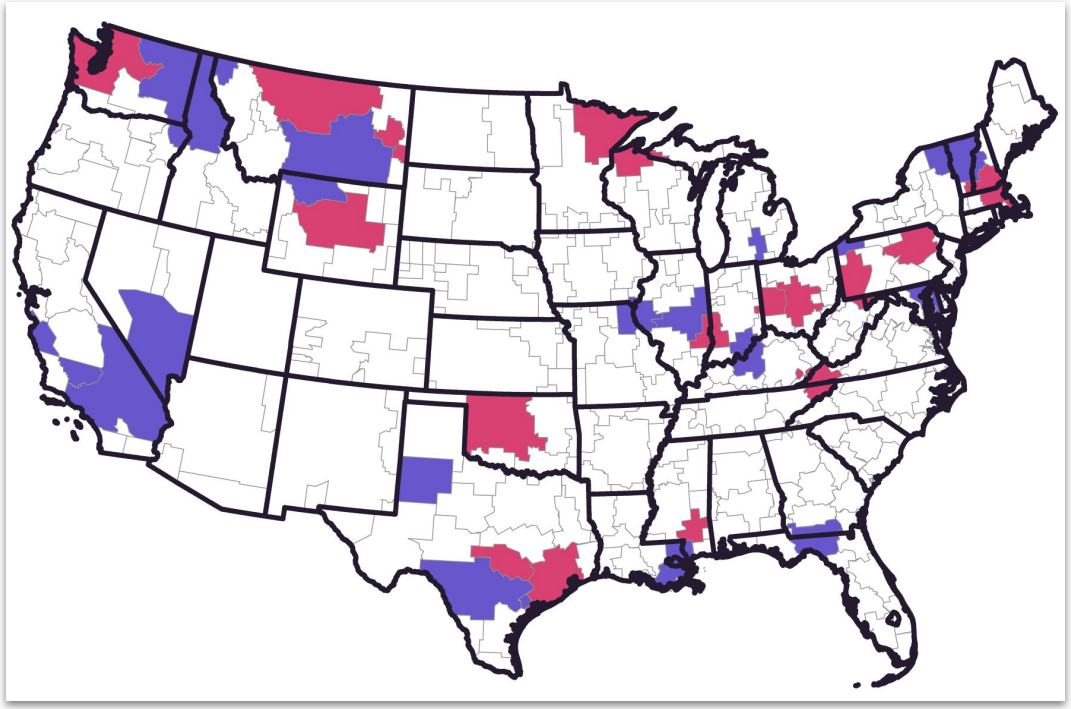
DATES Sep 20, 2023 - Sep 28, 2023

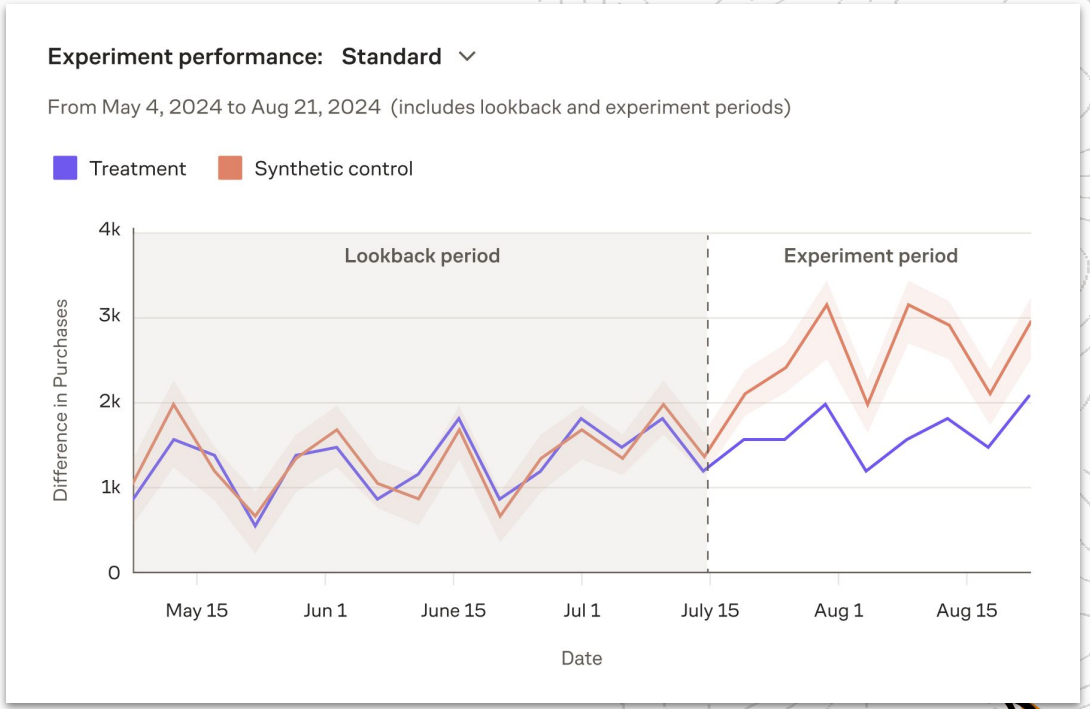
OWNER Summer Tan

DECISION (B) Price conscious algo



WE NEED TO TEST ON AREAS WHERE YOU ARE SPENDING THE MOST MONEY



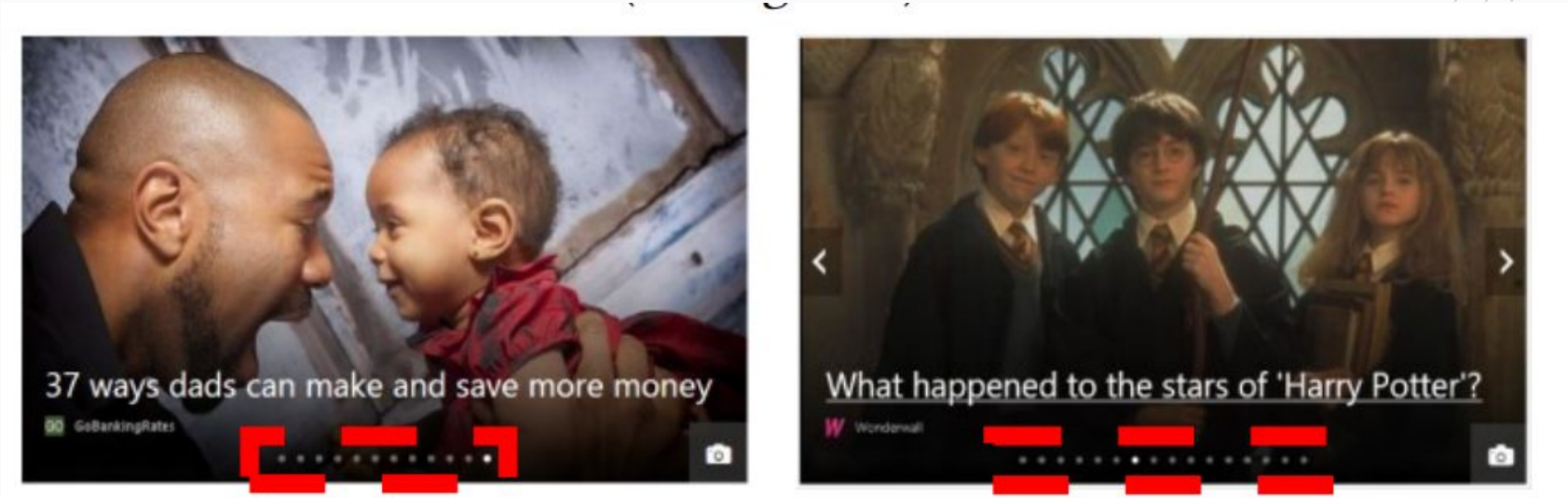






CONCLUSION

WE'RE USING RIGOROUS SCIENCE



Check assignment balance

Ensure experiments are powered

WE'RE RUNNING ENOUGH EXPERIMENTS

Create repeatable defaults

Self-service analysis tools

Company/Source	Success rate
Microsoft	33%
Bing	15%
Booking.com, Google Ads, Netflix	10%
Airbnb Search	8%

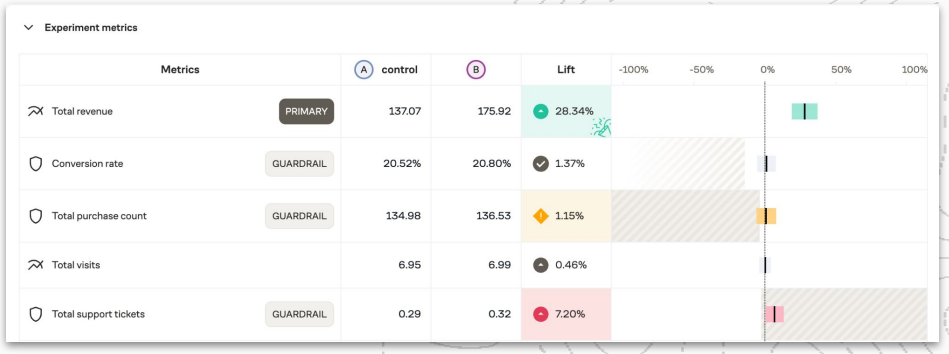


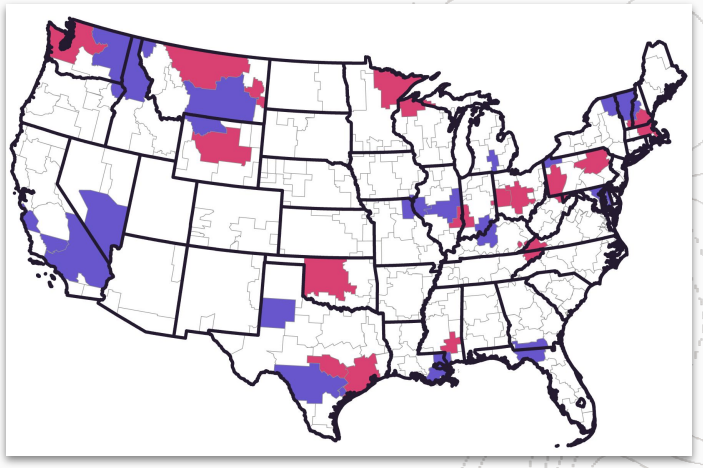
WE'RE EXPERIMENTING WITH BIG ENOUGH CHANGES

Utilize guardrail metrics to protect from downside risk

Test in areas that are sensitive

Test on areas where you're spending the most money



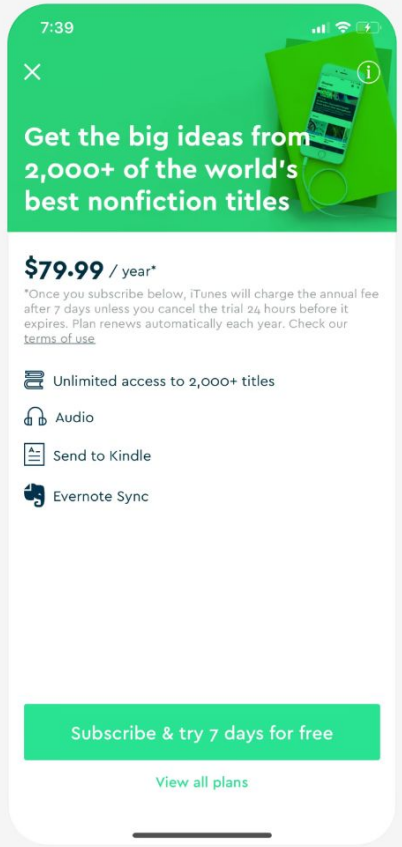




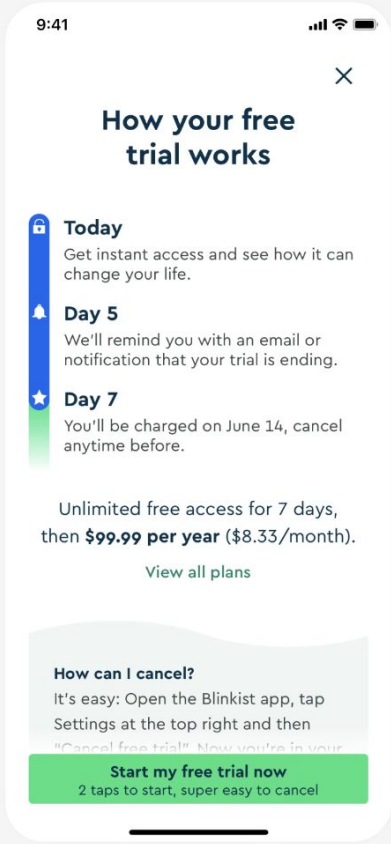
ONE MORE THING...

EDUCATION

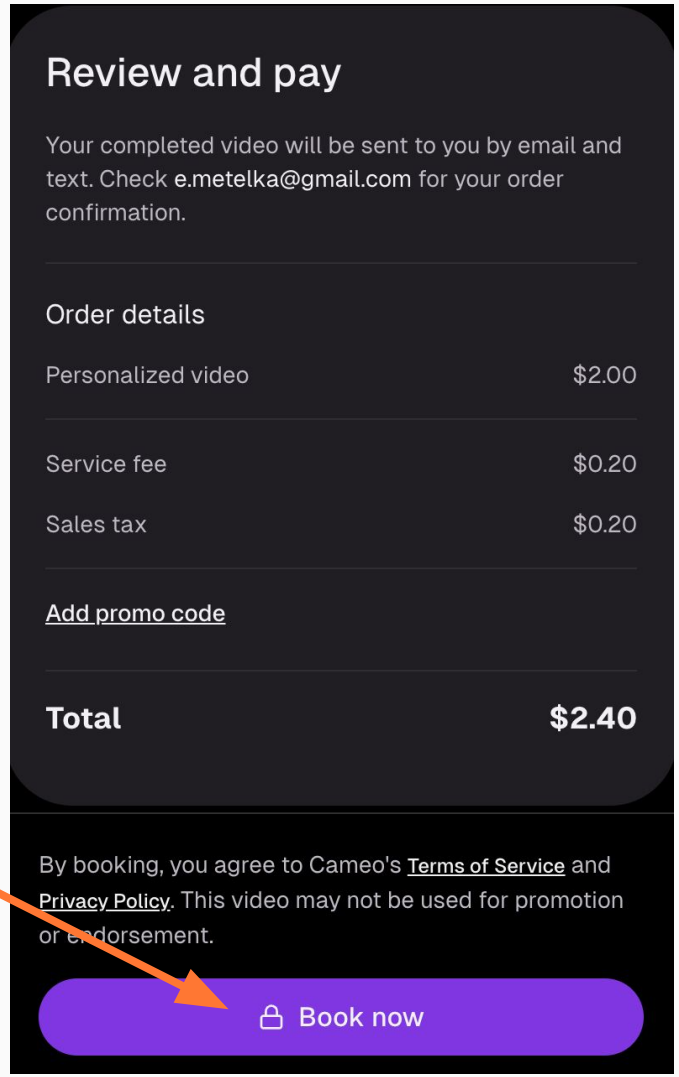
A



B



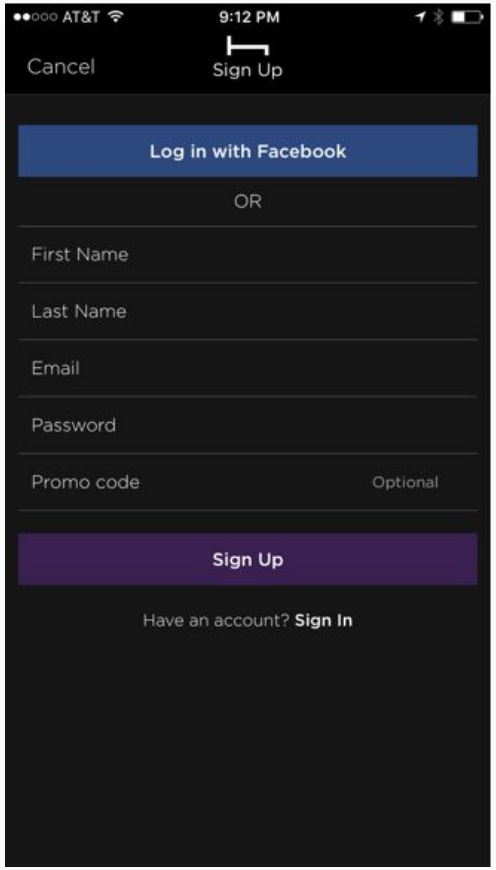
TRUST


Review and pay
Your completed video will be sent to you by email and text. Check e.metelka@gmail.com for your order confirmation.
Order details
Personalized video $2.00
Service fee $0.20
Sales tax $0.20
Add promo code
Total $2.40
By booking, you agree to Cameo's Terms of Service and Privacy Policy. This video may not be used for promotion or endorsement.
Book now

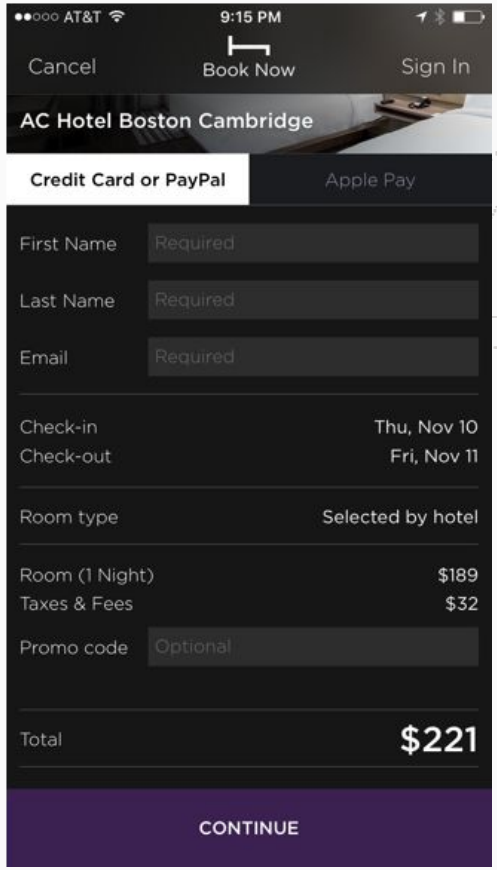


FRICTION REMOVAL

CONTROL
Account Creation



VARIANT
In-line Checkout





THANK YOU