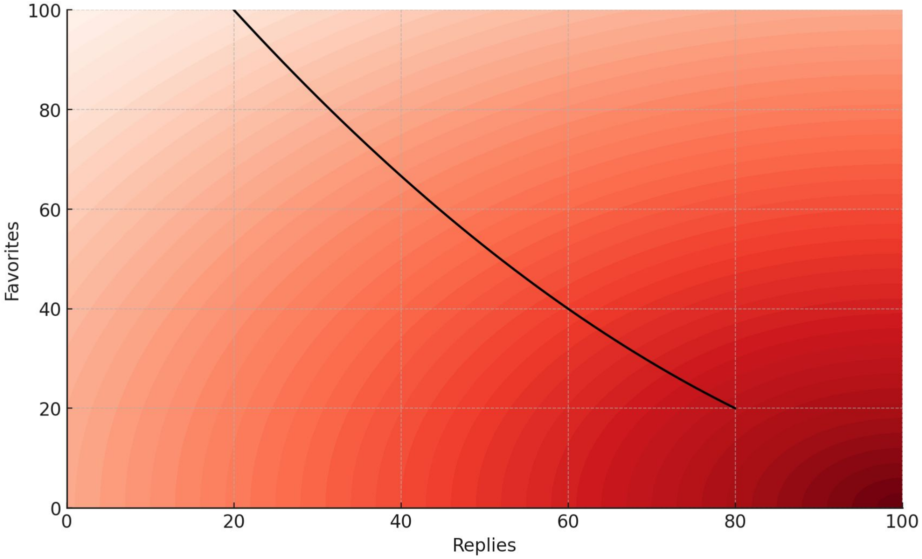
Some Quick Thoughts On Experimentation And Optimization

From Manual Decisions to Automated Strategy



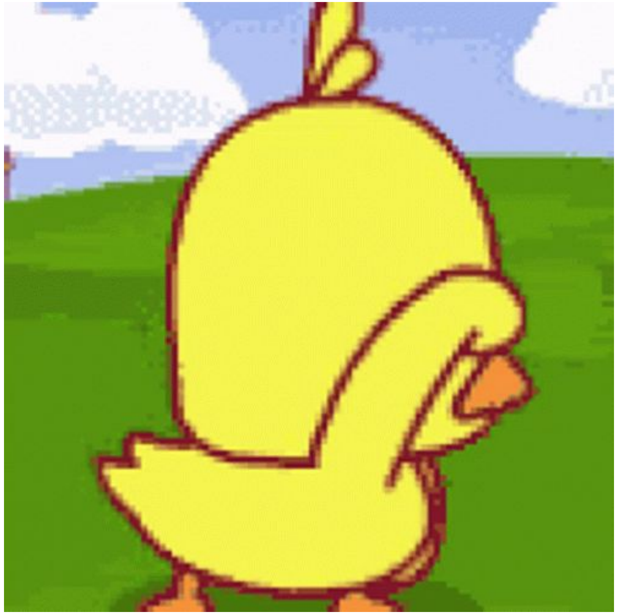
Brent Cohn, March 2024

Acknowledgements

- Thank you to the Experimentation, Data Science and ML teams at Twitter and Stripe that I worked with on the experiments discussed in this talk.

- In particular, thank you to: Anthony Fu, Marika Inhoff, Iuliana Pascu, Tom Cunningham, Niall O'Hara, Ming Chen, Matt DeLand, Ross Kravitz and many others.

- All opinions are my own!



The logo of Duck Duck Goose, Twitter's Experimentation Platform

Gif Source https://tenor.com/view/hi-ho-duck-duckling-looking-gif-14488085

About Me and Slides

- Slides at: https://tinyurl.com/brentdatacouncilslides

- Brent is a data scientist at Stripe where he manages teams focused on optimizing payment costs, fraud, and success rates. Before Stripe, Brent spent 6 years at Twitter where he managed a team focused on improving Twitter's experimentation platform, DDG. Brent is interested in democratizing the application of causal inference to software development, and has built a variety of tools to make it easier and more intuitive for developers to draw causal conclusions from their work.

- Contact: brentjoseph@gmail.com; https://www.linkedin.com/in/brentcohn/

The Problem: Decisions & Strategy

- ‘A/B testing’ implies a comparison between two variants. The conventional view of experimentation in industry has **focused on decision-making**. We run experiments to make faster, more accurate, and on average better, decisions.

- I’ve been a part of many debates around p values; MDEs; variance reduction; and peaking, but (in certain environments) I don’t think the juice was worth the squeeze.

- In modern companies with large amounts data ***running experiments to make decisions is often wasteful***, and we should think more about ***optimizing strategies instead***.

Image Source: A/B Testing and the Benefits of an Experimentation Culture

Defining The Terms

- **What is a Strategy?** Richard Rumelt's Good Strategy, Bad Strategy defines a strategy as having three parts:
    - **Diagnosis**: a theory describing the nature of the challenge.
    - **Guiding policies**: the approaches you'll apply to grapple with the challenge.
    - **Coherent actions**: a set of specific actions directed by guiding policy to address challenge.

- **What is a Decision?** A specific choice made by an organization or individual, aimed at achieving a defined goal or outcome.
    - In our context, the decision is launching an experiment to production and is often made based on a set of metrics that act as shipping criteria to determine if an experiment should be launched to production.

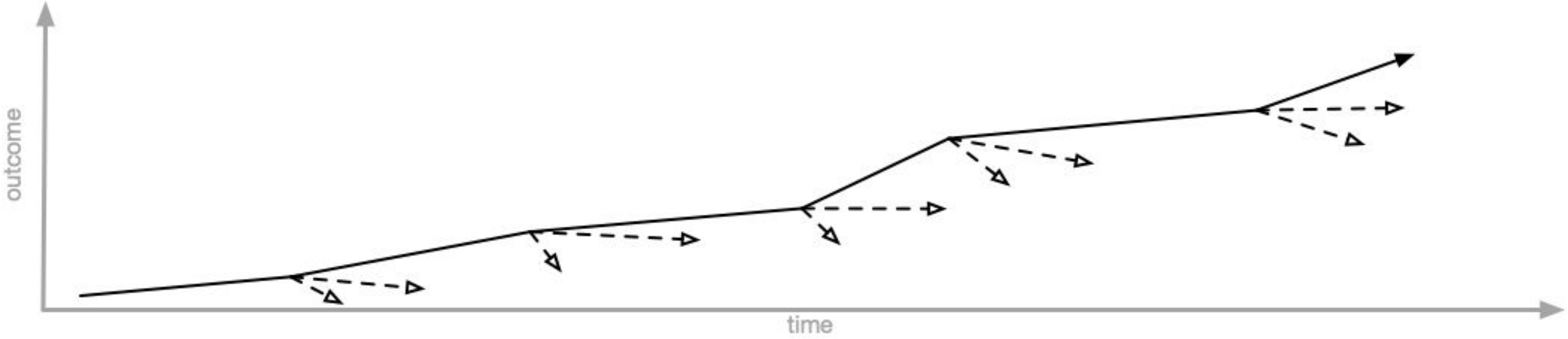


Image Source: Magnitudes of exploration.

Applying The Terms

- The three parts of a Strategy map to experiments and, more broadly, optimization problems in business fairly clearly.
  - **Diagnosis:** How your business's goal connects to your space. For example, maximize paid conversion by showing compelling content in an advertise.
  - **Policies:** The things you can intervene on in your space to achieve your goal. In this example, text copy, multimedia presence, and page layout
  - **Actions:** The specific things you can do to each of your policies.

- **Meanwhile determining Good Shipping Criteria Is A Hard Problem:** it takes a long time through one off experiments to enough evidence to generate data driven shipping criteria, and it's hard for decision-makers to weigh multiple factors when shipping criteria involve many trade-offs.

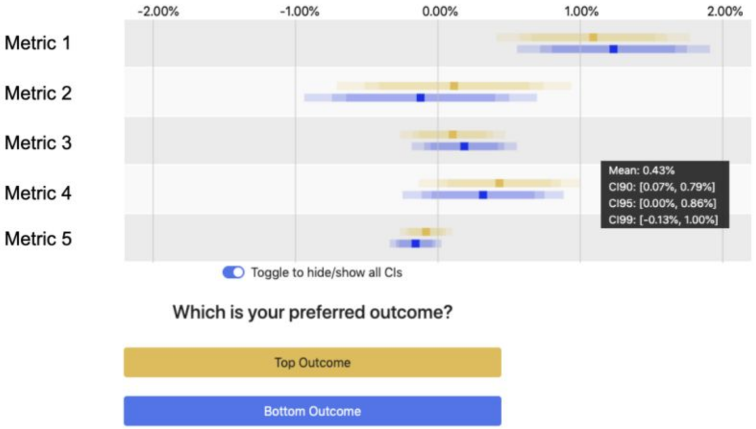


Image Source: Preference Exploration for Efficient Bayesian Optimization with Multiple Outcomes

A Problem Example: Video Size

- In 2016, Twitter pivoted to focus more on video. As part of this pivot, the company decided to emphasize videos (and other rich media) more in the timelines.
  - This meant we wanted to make videos bigger, and leadership wanted to know the effects of this decision.

- I was on the experimentation team at the time and designed and ran a really great experiment. We (with very low variance!) very much learned the effects of this decision. It increased engagement with videos and decreased engagement with text and met the OEC we'd set out for the experiment.

- We shipped it. A year later, I was asked to work on this again.

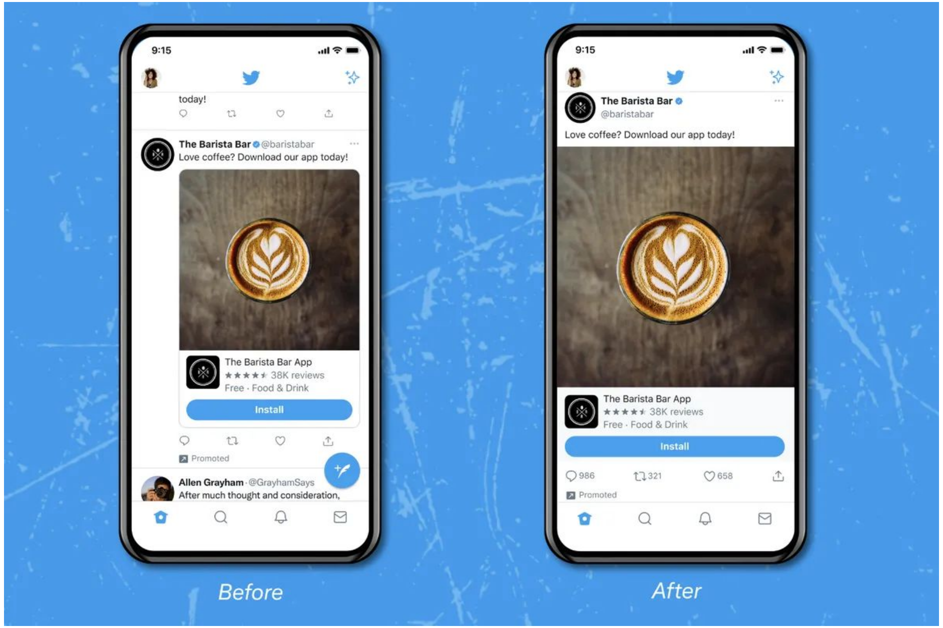


Image Source: Twitter tests a new timeline with edge-to-edge picture and video

A Fun Example: Video Size

- The second time I was asked to help design a “make video bigger experiment”, I started to think more about the strategy underlying the experiment rather than the specific decision.

- The strategy was to increase engagement with video with some acceptable trade off for lost engagement on non-video materials.The key relationship we needed to estimate to implement this strategy was the relationship between video size, video engagement, and non-video engagement.

- This relationship would be hard to estimate by comparing two sizes every years! Maybe we weren’t trying the best size each time?

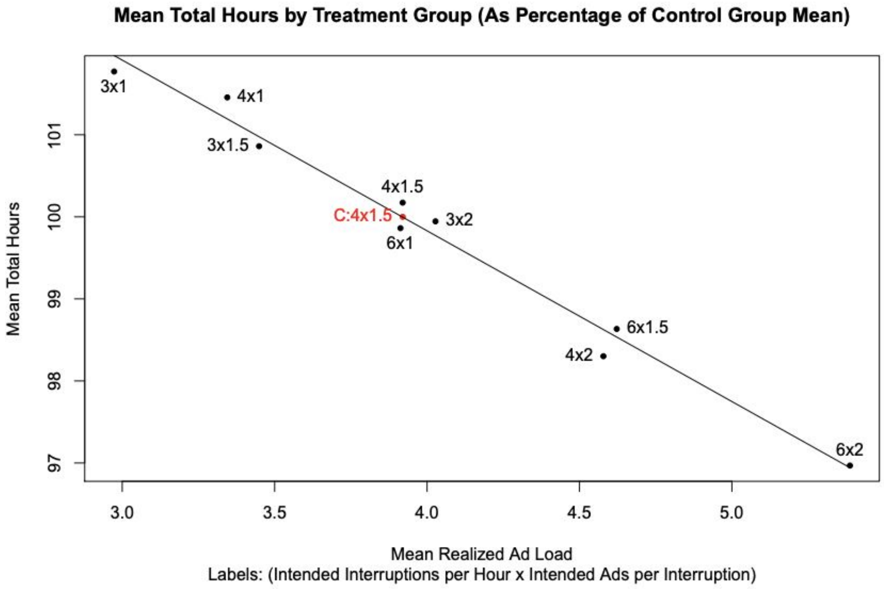


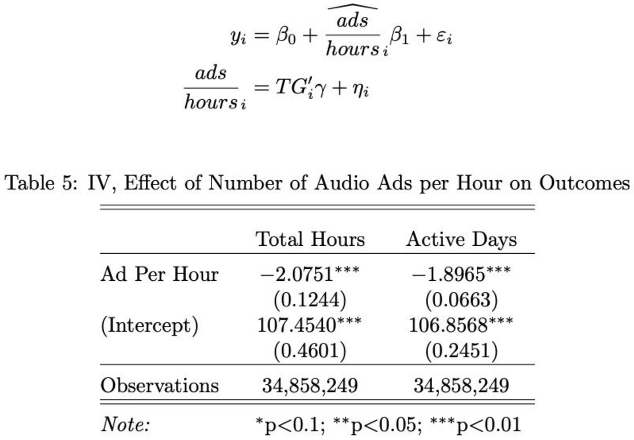
Image Source: Measuring Consumer Sensitivity to Audio Advertising:A Field Experiment on Pandora Inte
Radio

A Fun Example: Video Size

- How could we estimate this relationship? By this point we’re in 2018 and I read this wonderful paper by some folks at Pandora, Measuring Consumer Sensitivity to Audio Advertising:A Field Experiment on Pandora Internet Radio

- They wanted to estimate the effect of ads on usage at Pandora. So they tested a bunch of variants and fit a regression to estimate the dose response curve.

- They found a surprisingly linear relationship, but this approach was fairly straightforward to translate to Twitter and **helped us generate a “formula” for decision makers to trade off video real estate with text real estate at Twitter**.



$$y_i = \beta_0 + \widehat{\frac{ads}{hours}}_i \beta_1 + \varepsilon_i$$

$$\frac{ads}{hours}_i = TG_i'\gamma + \eta_i$$

Table 5: IV, Effect of Number of Audio Ads per Hour on Outcomes

	Total Hours	Active Days
Ad Per Hour	−2.0751***	−1.8965***
	(0.1244)	(0.0663)
(Intercept)	107.4540***	106.8568***
	(0.4601)	(0.2451)
Observations	34,858,249	34,858,249
Note:	*p<0.1; **p<0.05; ***p<0.01	

Image Source: Measuring Consumer Sensitivity to Audio Advertising:A Field Experiment on Pandora Inte
Radio

Looping Back To The Big Idea

- **Diagnosis**: “More consumption of videos is good for social media platforms because it increases revenue through pre-roll advertisements and drives retention. We want to increase video consumption without hurting other engagement too much”

- **Guiding policies**: Video size.

- **Coherent actions**: Varying video size at the user, device level.

- We have gone from an experiment to an optimization problem! In this circumstance we only generated a few points and drew a line through them, but this was still more effective than generating an additional point per year and picking the higher value.

Image Source: ChatGPT

From Decisions To Strategy: Video Size

- Focusing on directly estimating the strategic question, “what is the trade off between video size and engagement”, was more efficient than testing a single decision (“what is the effect of a new design with bigger videos”).

- Using the simple model we estimated an executive could say “I want to increase the engagement on videos by 50%” and this would be fairly straightforward to turn into a product launch rather than “guess and checking” video sizes.

- In my experience almost all decision making A/B tests have an underlying strategic question that can be estimated – often through experimentation!

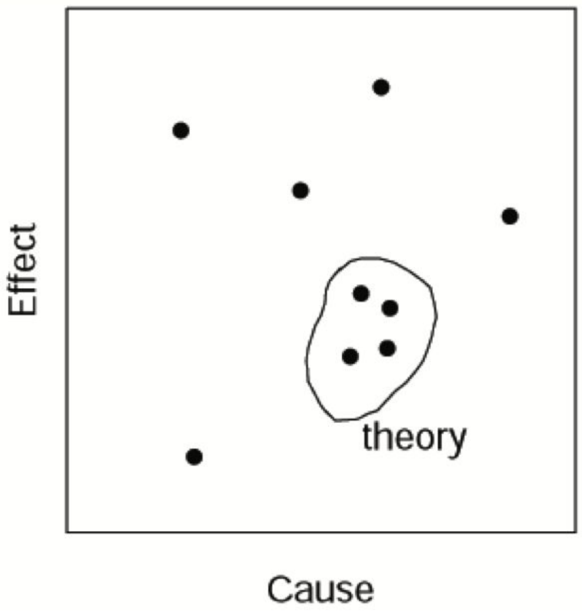
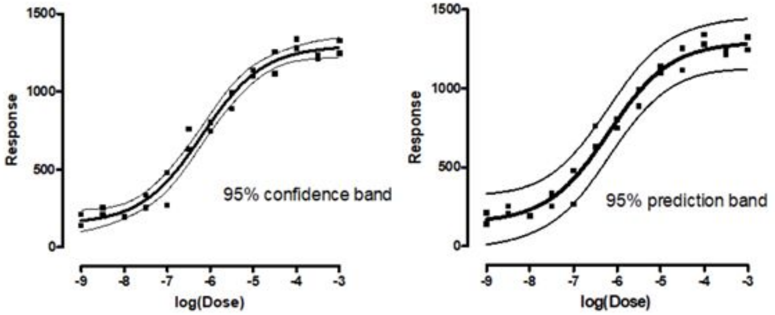


Figure 31. Cause-effect space. Each point represents a cause-effect linkage – a case where x causes y. A theory indicates a region of the space with an above-average density of points.

Image Source: Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, weight

From Decisions To Strategy: Ideas

- **Draw A Line!** Add a bunch of variants to an experiment to explore effects across a continuum and plot a dose response curve.



From Past Decisions To Future Strategy!

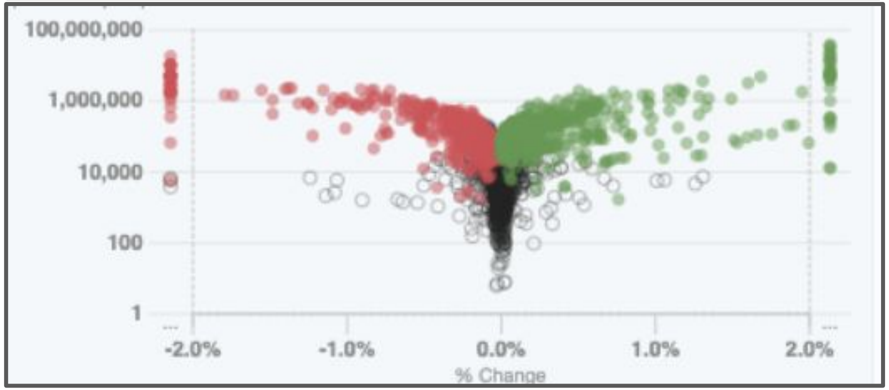
- By 2020, Twitter ***had run more than 10,000 experiments*** on this platform.

- These past experiments constitute a vast corpus of historical information (results, design, etc.) that could potentially be used to better estimate future strategies.

- However, the information from these experiments was disorganized, hard to digest, and **all too often forgotten**.

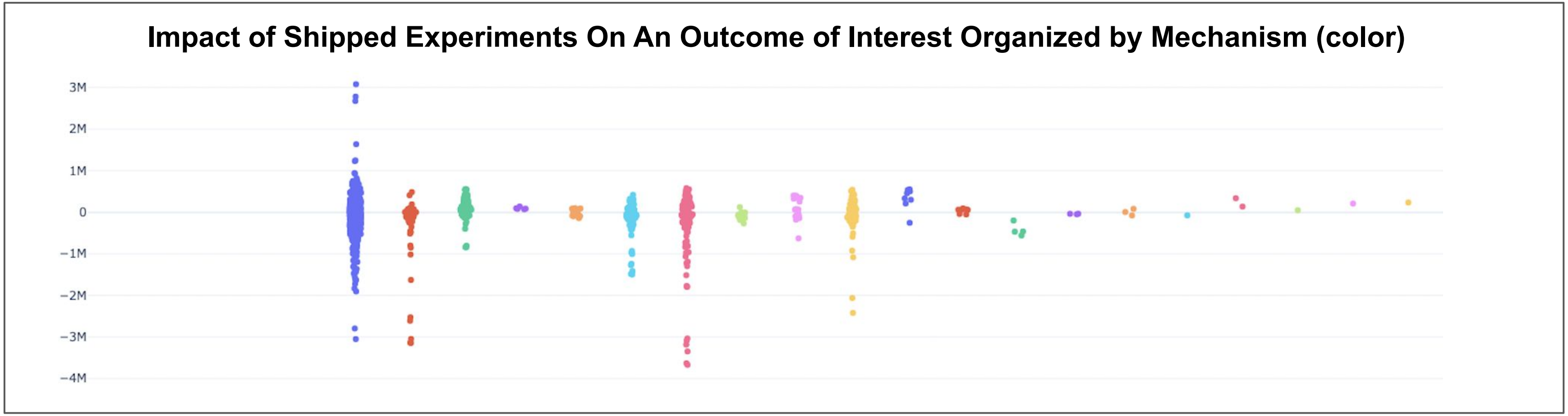
- So we built some tools to make the connection between past decisions and future strategy more obvious.

Relevant Past Experiment Results by Impact and Significance



Institutional Memory Tooling

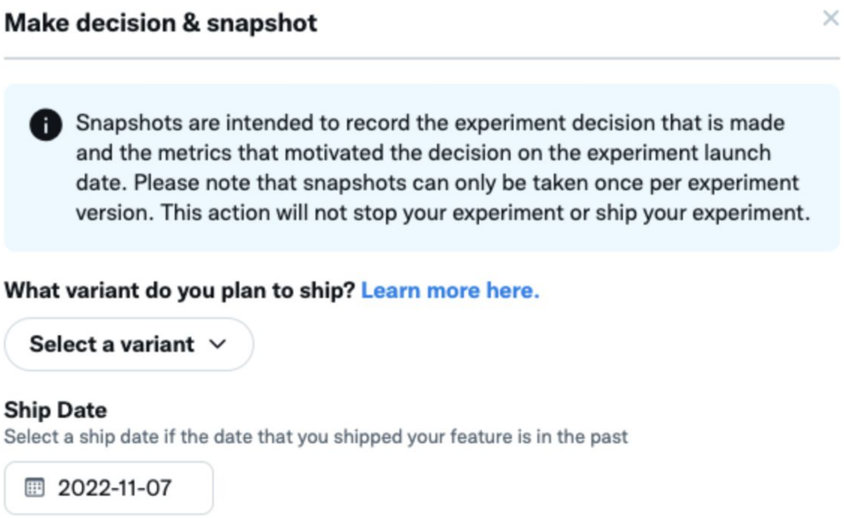
- Tooling to support institutional memory at Twitter:
  - **Tracking and organization of meta-data**. Backfill as much as we can to create structured, queryable datasets.
  - **Build Theory:** Create theories based on a large amount of experiment results.
  - **Design Better Experiments:** Use past experiments for power calculations.
  - **More Comprehensive Evaluation:** Track and map the impact of individual shipped experiments to an overall product's success.



Tracking and Organization

- General Principles
  - **Structured data over unstructured data:** capture as much information as possible in structured, queryable formats.
  - **Add no friction**: ensure collecting additional metadata adds minimal work to individual experimenters.
  - **Flexibility:** allow experimenters to add new tags and labels to better classify experiments since they are the experts (e.g. onboarding modals types).
  - **End to end data collection:** ensure that experiment decisions and rationales are automatically captured, don't just focus on set up.

UI that Records the State of the Experiment at Ship



Make decision & snapshot

Snapshots are intended to record the experiment decision that is made and the metrics that motivated the decision on the experiment launch date. Please note that snapshots can only be taken once per experiment version. This action will not stop your experiment or ship your experiment.

What variant do you plan to ship? Learn more here.

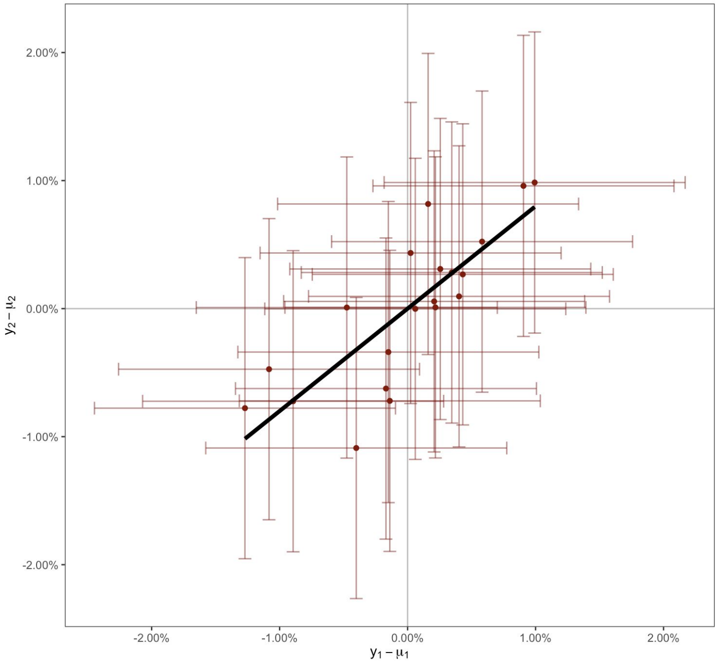
Ship Date

Select a ship date if the date that you shipped your feature is in the past

2022-11-07

From Tooling To Strategy

- By analyzing joint distribution of results across experiments we learned about causal relationships between mechanisms.
  - Examples:
    - Push recommendations and email recommendations are substitutes.
    - Push recommendations and in-app algorithmic selection of content are complements.
  - Implications:
    - Better external validity of experiment impacts.
    - Higher impact product direction.



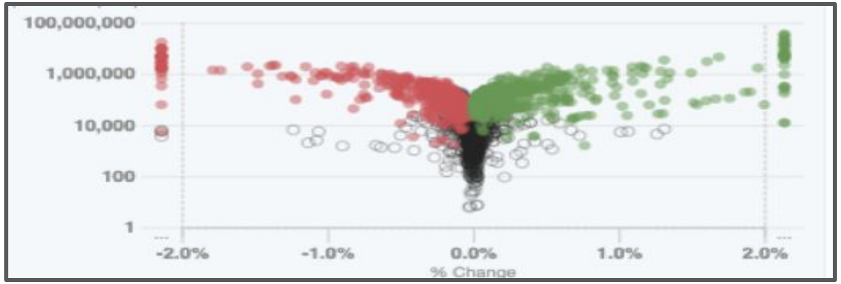
Source: Experiment Interpretation and Extrapolation

From Decisions To Strategy: Ideas

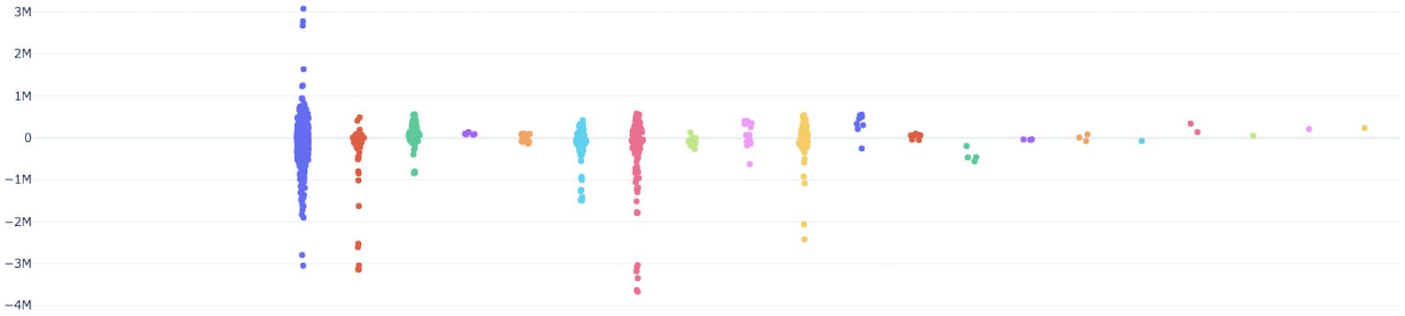
- **Draw A Line!** Add a bunch of variants to an experiment to explore effects across a continuum and plot a dose response curve.

- **Make A Scatter Plot!** Collect and Analyze Structured Metadata from Prior Experiments.

Relevant Past Experiment Results by Impact and Significance



Impact of Shipped Experiments On An Outcome of Interest Organized by Mechanism (color)



Another Fun Example: Reply Guys!

- In 2019, Twitter made a strategic focus to increase conversation on the platform.

- I worked with some phenomenal ranking engineers to hit our KRs by increasing ranking weights!

- Our goal was to increase replies without hurting engagement. We tried a few guess and check weights for the respective models and

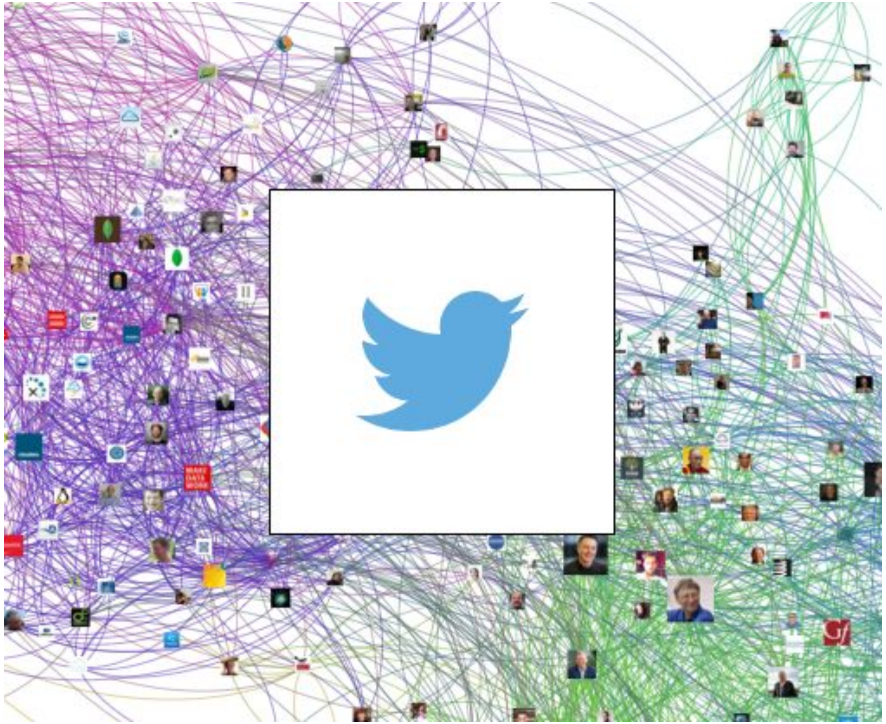


Image Source: How to Visualize Your Twitter Network

Another Fun Example: Reply Guys!

- We did a series of experiments to determine the effects from upweighting tweets with high p(reply) relative to p(favorite).
- All sorts of neat network effects implications since upweighting some tweets downweights others. This causes other content to be displaced and users to receive fewer favorites.
- We did some HPO and “drew an ellipse” in a series of production experiments.
- However! we found that tweets that were high p(reply) and low p(favorite) were likely to receive many reports!
- We needed to optimize among many axes…

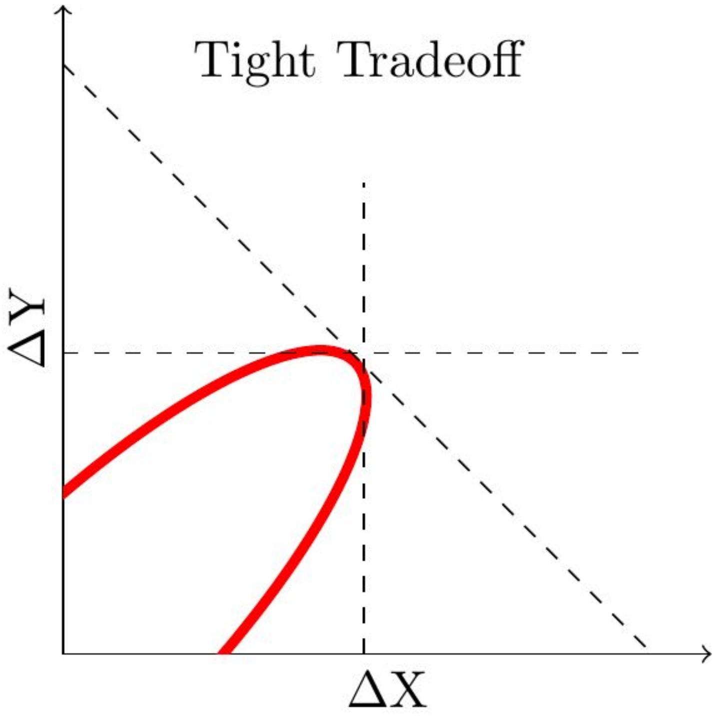
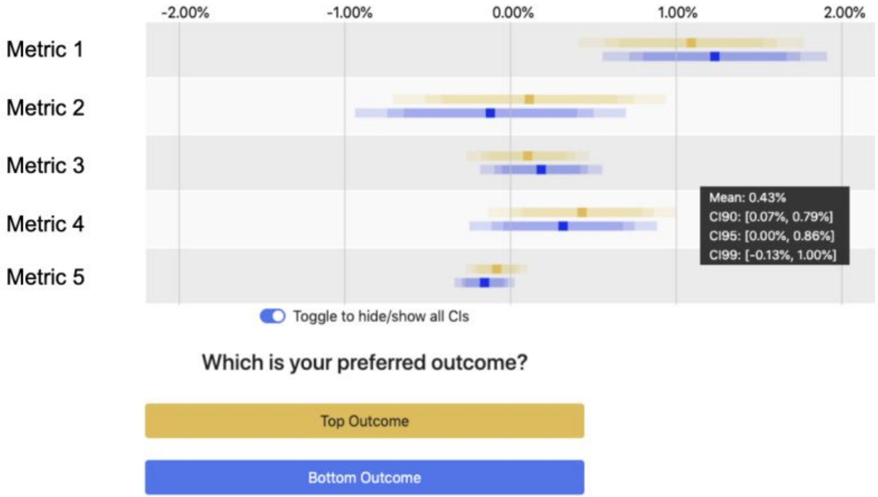


Image Source: Thinking About Tradeoffs? Draw an Ellipse

Another Fun Example: Reply Guys!

- Turns out it’s really hard for a human to pick an optimal point across many different metrics with probabilistic in production!



Which is your preferred outcome?

Image Source: Eytan Bakshy’s X: https://twitter.com/eytan/status/1508430614944440327

Another Fun Example: Reply Guys!

- Turns out it's really hard for a human to pick an optimal point across many different metrics with probabilistic in production!

- So we built a tool to evaluate pareto quality of various outcomes in a hypervolume and picked the pareto optimal points that balanced positive outcomes (replies/favorites) without increasing reports.

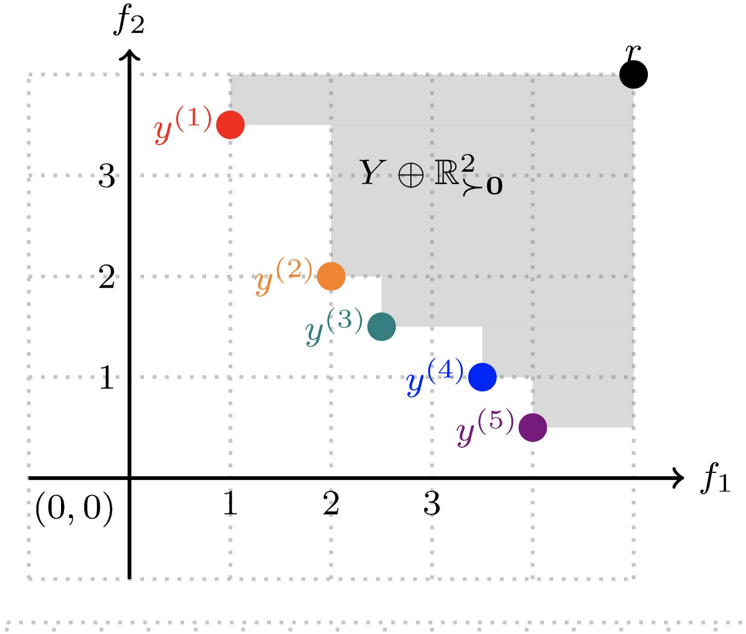


Image Source: Ax documentation: https://ax.dev/versions/0.1.18/tutorials/multiobjective_optimization.html

From Decisions To Strategy: Ideas

- **Draw A Line!** Add a bunch of variants to an experiment to explore effects across a continuum and plot a dose response curve.
- **Look At A Plot!** Collect and Analyze Structured Metadata from Prior Experiments.
- **Write Down Your Objective Function And Map Out A Pareto Frontier** This is helpful when the business faces trade offs in more than two metrics and it's not clear what the obvious optimum between all three would be…. A 3d plot!?

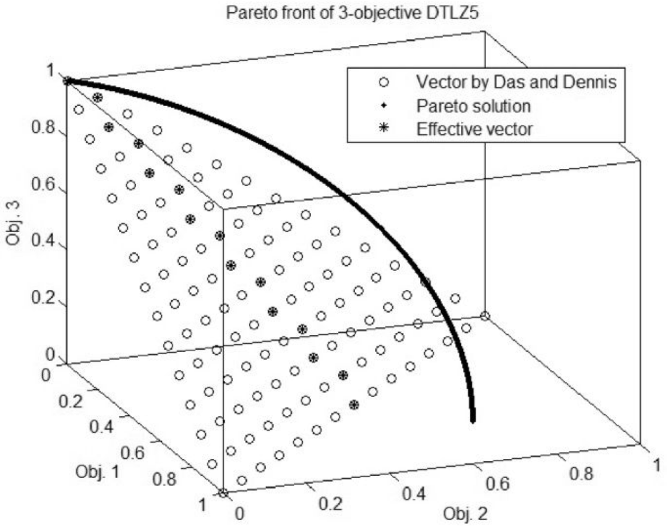


Image Source: Generating multiple reference vectors for a class of many-objective optimization problems with degenerate Pareto fronts | Complex & Intelligent Systems:

A Newer Example: Profit Maximization

- Patio11 has a nice blog post: “The optimal amount of fraud is non-zero” which discusses the tradeoff between willingness to block fraud and a business’s margin.

- We built a model to help each merchant pick the optimal point on the fraud versus conversion curve based on their own information.

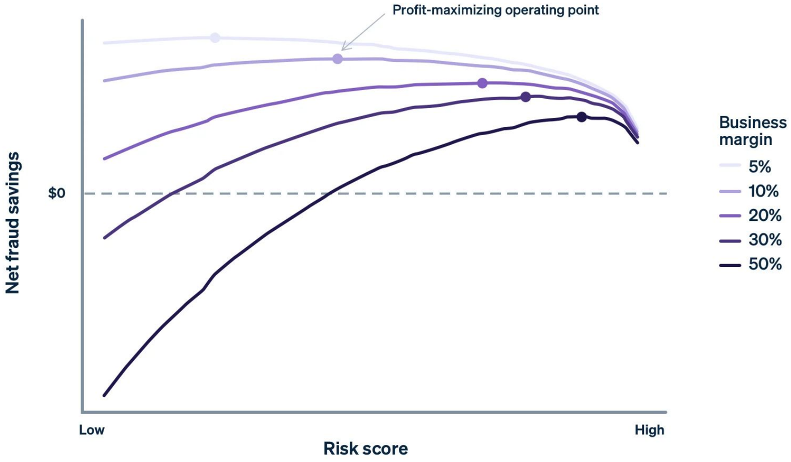


Image Source: https://stripe.com/guides/state-of-online-fraud

A Newer Example: Profit Maximization

- Patio11 has a nice blog post: “The optimal amount of fraud is non-zero” which discusses the tradeoff between willingness to block fraud and a business’s margin.

- This model expanded to include multiple interventions beyond just blocking such as throwing a captcha, 2 factor on phone or browser, card image verification, and sending additional information to the banks for a more accurate response.

- Each intervention works best for specific types of fraud (e.g. captcha for low sophistical bots, 2 factor for stolen cards but not identify)

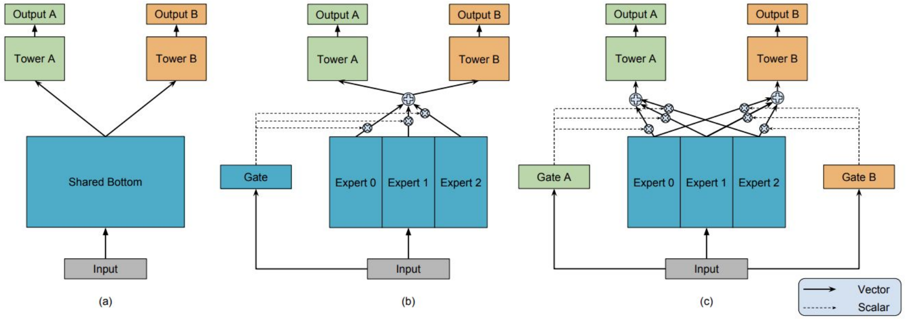


Figure 1: (a) Shared-Bottom model. (b) One-gate MoE model. (c) Multi-gate MoE model.

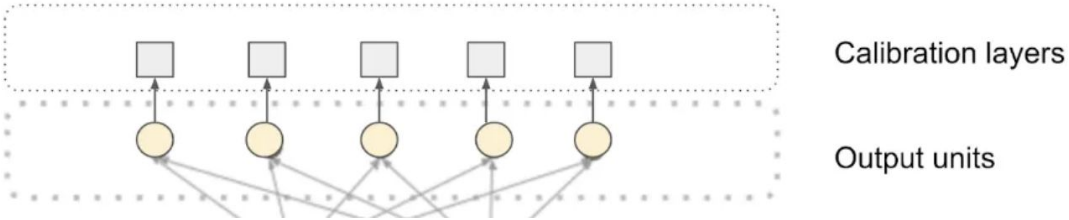


Image source Medium

A Newer Example: Profit Maximization

- Calibrated model output can be used to help merchants maximize profit on a per transaction level.

- We use an online contextual bandit to explore between hundreds of different possible options and predict the best option using a multitask model.

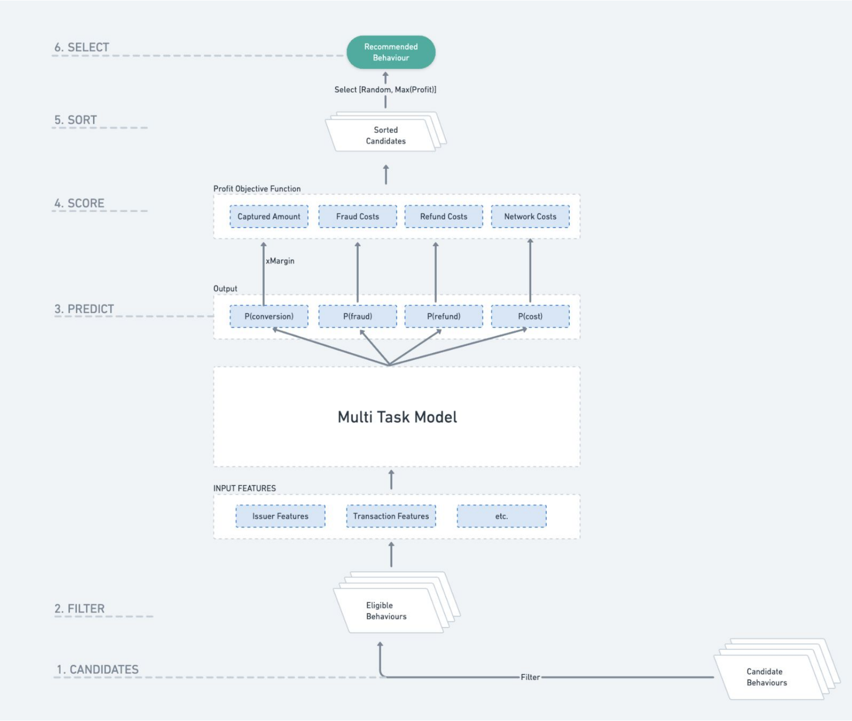


Image Source: Internal System Diagram

A Newer Example: Profit Maximization

- Fun things to discuss!
    - Adding new levers and exploring quickly.
    - Model architecture choices
        - Why multitask MoE?

From Decisions To Strategy: Ideas

- **Draw A Line!** Add a bunch of variants to an experiment to explore effects across a continuum and plot a dose response curve.
- **Look At A Plot!** Collect and Analyze Structured Metadata from Prior Experiments.
- **Write Down Your Objective Function And Map Out A Pareto Frontier** This is helpful when the business faces trade offs in more than two metrics and it's not clear what the obvious optimum between all three would be…. A 3d plot!?
- **Design An Entire Engineering System to Explore Your Multidimensional Space And Find An Optimum!** Hooray - you are the strategy now!

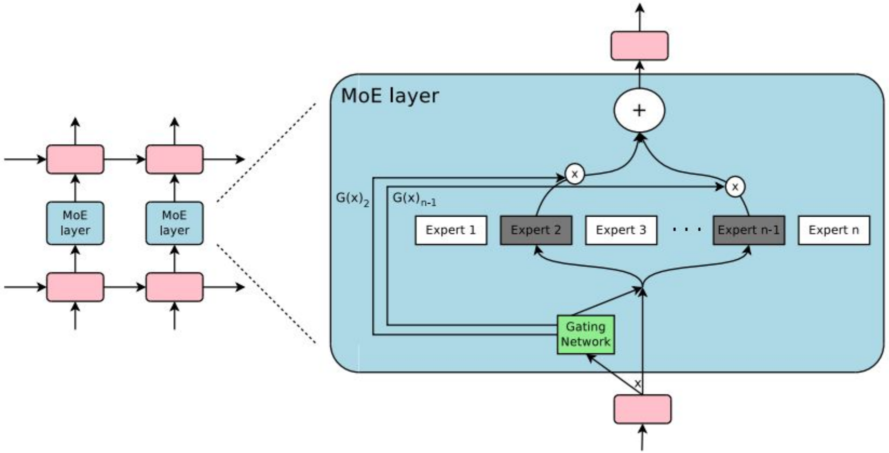


Image Source: Mixture of Experts Explained

Discussion!