wayfair

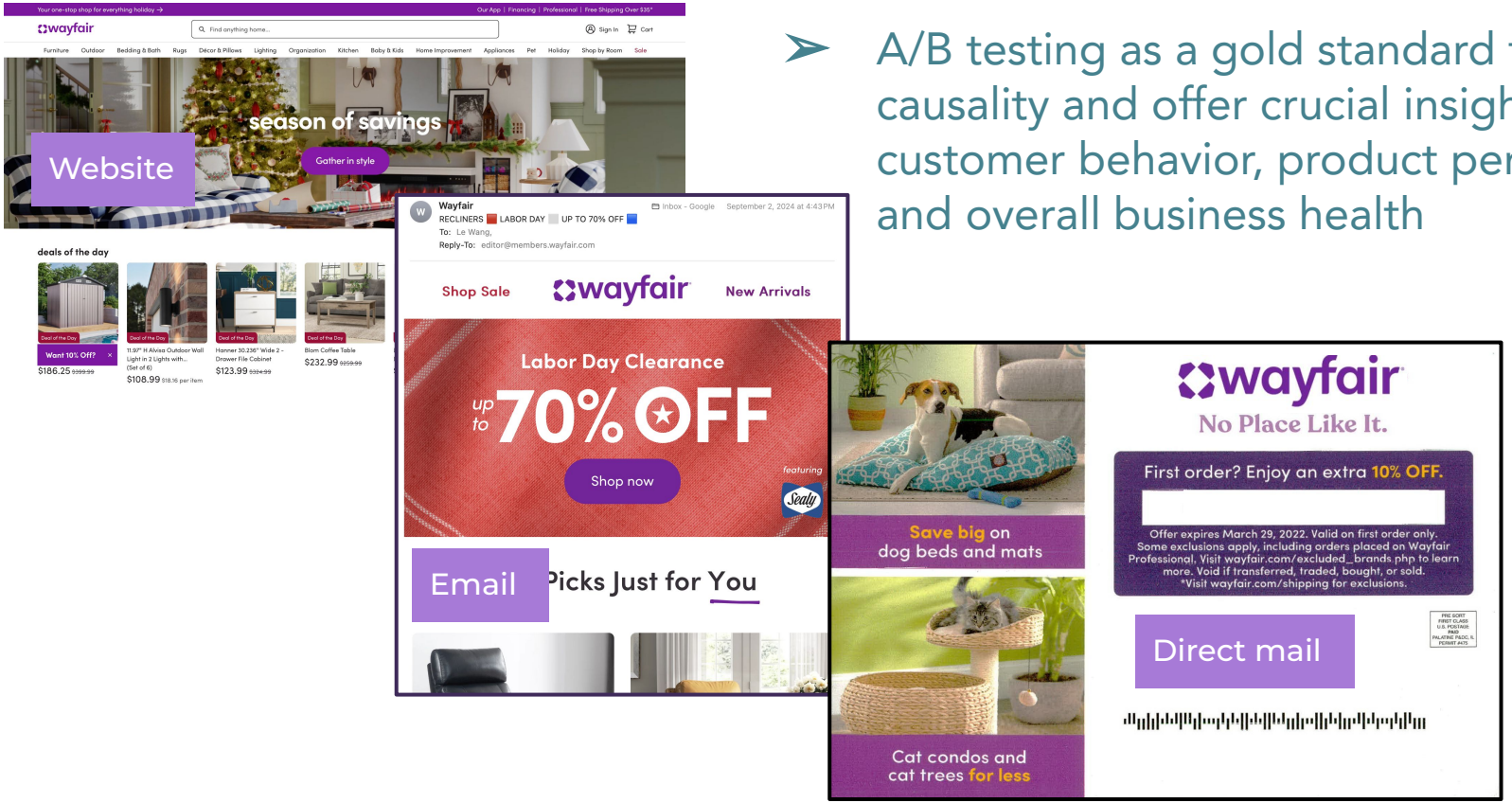
Metrics Evaluation Framework for Controlled Experiments

Le Wang

Oct 29, 2024

NABE TEC | 2024

A/B testing is important at Wayfair



- A/B testing as a gold standard to reveal causality and offer crucial insights into customer behavior, product performance and overall business health

Which metrics to use in controlled experiments?

- Selecting high-quality metrics that capture real business objectives is often challenging
- At Wayfair, we developed a metrics evaluation process to systematically develop, assess and select metrics that align with customer experiences and business objectives
  - Key question it answers:
    - What criteria can be used to evaluate metrics quantitatively?
    - How can we compare and choose primary metrics for experiments?
    - What practical guidelines can help in selecting and refining metrics over time?

Two critical properties of the metric quality

Sensitivity

- Probability of correctly detecting a treatment effect:

P(Reject null hypothesis & alternative hypothesis is correct)

- Signal to Noise Ratio (SNR):

Mean(metric)/ StandardError(metric)

Trustworthiness

Alignment with key business metrics of interest, such as customer-level revenue

- Test statistics correlation with business north star metric
- Directional alignment score
- Significant alignment score
- Distance based measure

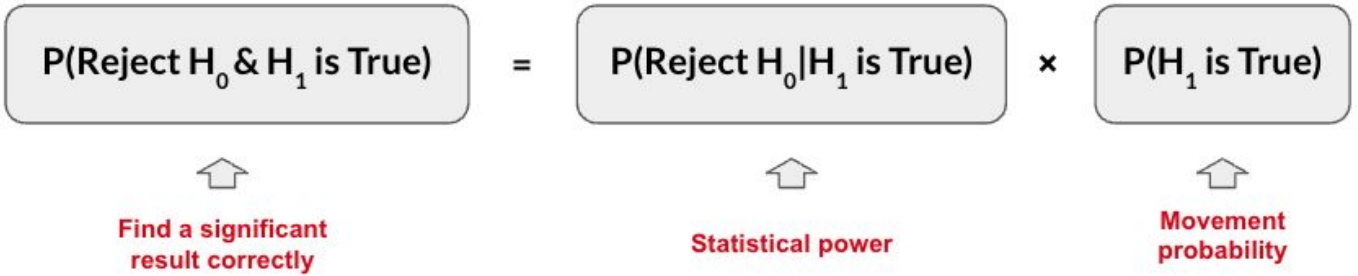
Selecting historical experiments for metrics evaluation

Criteria of selecting historical experiments:

- **Diverse Test Coverage:** The experiments should cover a broad range of test types (e.g., different products, geographical regions, etc.) to enable a comprehensive evaluation.
- **Quality Assurance:** Selected experiments should be free of known issues, such as misconfigurations, which could skew the results.
- **Business Domain Knowledge:** The results of the experiments should align with established business domain knowledge, excluding experiments with exceptionally "extreme" outcomes.


WAIT, THAT'S NOT ENOUGH
I WANT TO GO DEEPER
imgflip.com

Estimate the probability of correctly identifying a significant treatment effect



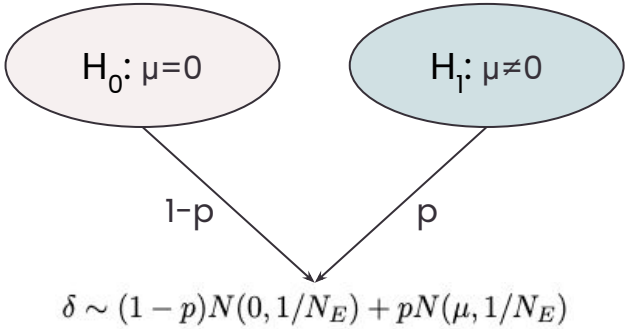
But we don't observe P(H1 is True)....

Bayesian mixture model to the rescue

δ is the scaled observed treatment effect
μ = E(δ) is the scaled true treatment effect
N_E is the effective sample size

Assume a prior distribution for μ under H_1:

$$\pi(\mu) = N(0, V^2)$$



The goal here is to infer the estimates of **p=Pr(H_1 is True)** and **V** based on observed experiment data points.

Sensitivity - simulation

Using a Bayesian mixture model we can estimate both parameter accurately

Number of tests	Truth p	Truth V	Estimated p	Estimated V
20	0.8	0.004	0.7965	0.003961
		0.010	0.8058	0.009910
	0.5	0.004	0.5372	0.003856
		0.010	0.5071	0.009883
	0.2	0.004	0.2659	0.003483
		0.010	0.2140	0.009324
200	0.8	0.004	0.8017	0.004000
		0.010	0.8008	0.009967
	0.5	0.004	0.5003	0.004001
		0.010	0.5007	0.009972
	0.2	0.004	0.2048	0.003960
		0.010	0.1987	0.009956
2000	0.8	0.004	0.7995	0.004000
		0.010	0.7997	0.010002
	0.5	0.004	0.4990	0.004003
		0.010	0.5002	0.010003
	0.2	0.004	0.1996	0.004002
		0.010	0.1995	0.010000

Parametric Decomposed Sensitivity for different metrics

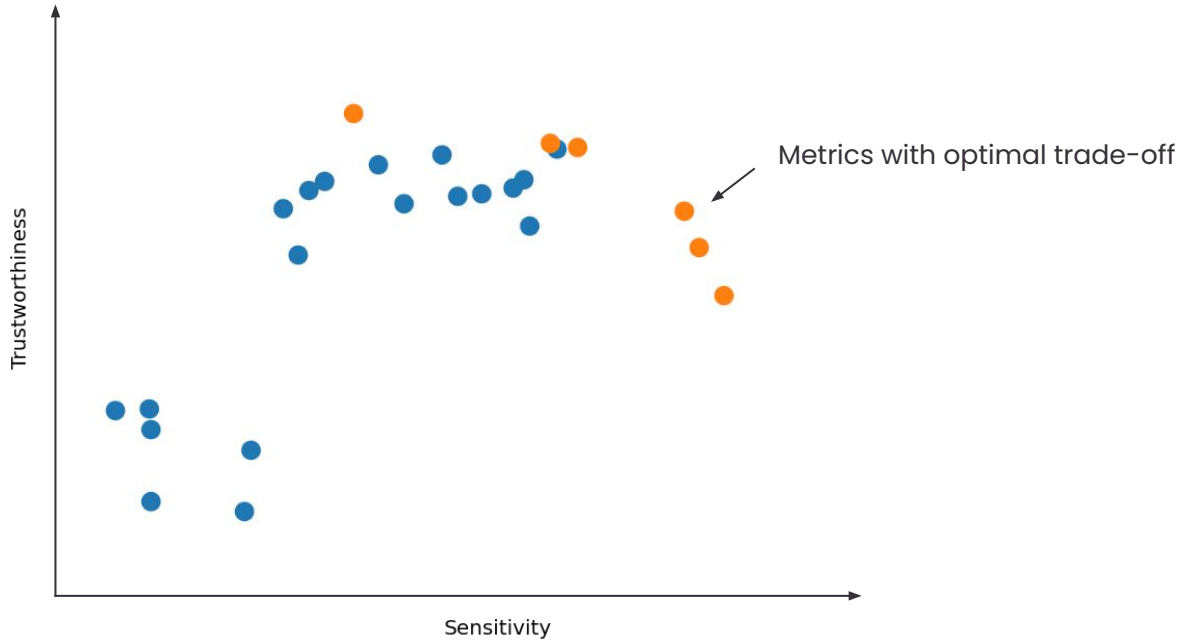
(numbers rescaled for demo purpose)

Metrics	Movement Probability	Estimated Power	Estimated Correct Rejection Probability
A	0.300537	0.759001	0.228108
B	0.416782	0.433084	0.180502
C	0.639996	0.859820	0.550281
D	0.868146	0.825113	0.716318

Only the power is not enough - movement probability (p) is also important

Sensitivity - Trustworthiness Tradeoff

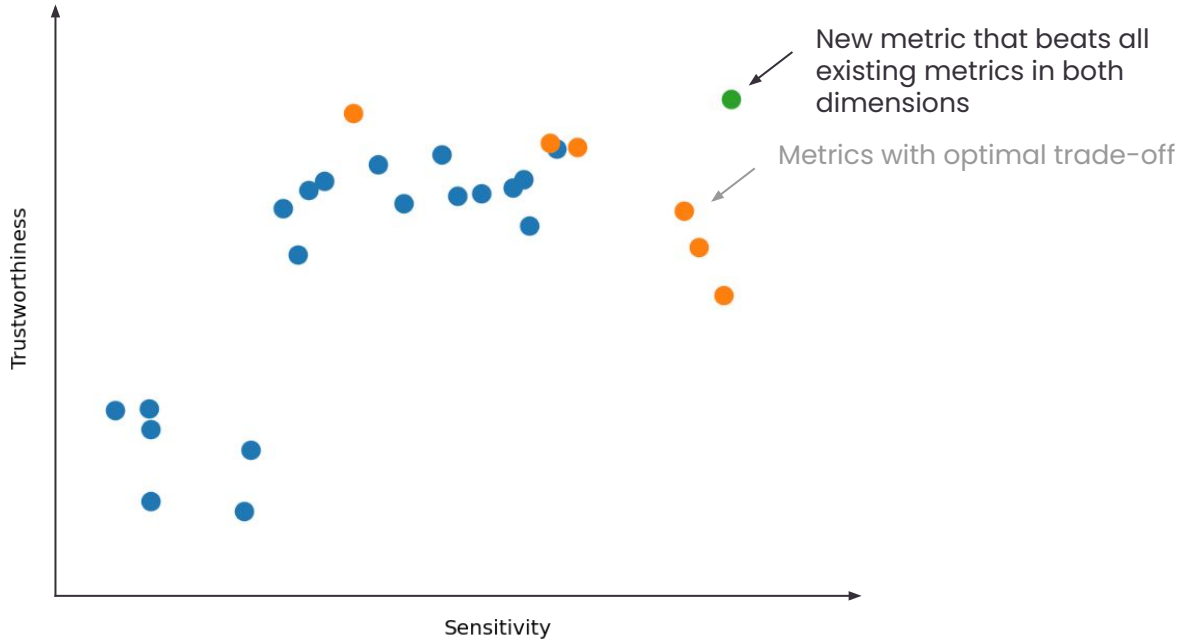
Leverage Pareto optimization when evaluating many metrics



- We found a range of metrics beyond simple engagement metrics which offer near optimal trade-offs in sensitivity vs trustworthiness.
- We found a substantial fraction of commonly used metrics were pareto dominated by top metrics.

Sensitivity - Trustworthiness Tradeoff

Leverage Pareto optimization when evaluating many metrics



We can create new metrics based on existing high quality metrics to achieve better sensitivity-trustworthiness trade-off

Conclusions

- Metrics evaluation framework matters
- Define measures for Sensitivity and Trustworthiness
- Identify high quality metrics and use them to drive test decisions towards business north star

Measurement Science team at Wayfair

- Ziwei Liao
- Enes Dilber
- Wiley Jennings
- Egemen Pamukcu
- Jerry Chen
- Le Wang
- Faith Shin
- Yusheng Jiao
- Joy Sun
- Patrick Phelps