When Web Personalization Misleads Bucket Testing

Ariyam Das Harish Ranganath

UEO Workshop, November 1, 2013

Outline

Online experiments for optimizing user engagement

Can personalization impact online experiments ?

Empirical analysis

Mitigating impact of personalization

Future work

Driving User Engagement

Two aspects of a web site broadly influence engagement of its users -

a) content provided on the site.

b) features of the site and layout of its individual pages.

Online Experiments

The Problem

Which layout maximizes CTR for this *personalized* module ?

Big ambitions for much-hyped new currency

Ambitions for new currency | Secret of 'Price is Right' | PS4 launch lineup | 'Deadliest Catch' scares | Ships that make you shiver

VS

GOP blocks Obama nominee for appeals court

GOP blocks Obama nominee

1990s NBA stars today

A-listers' hilarious spoof

Gosselins stopped speaking

The Problem

A higher CTR in one bucket can indicate -

Users liked the layout better.

or

Personalization worked better with users in that bucket.

Relevant content, strongly aligned to user interests, was available and served.

A Simple Experiment

Without any feature or layout changes, the visitors to a site were split into *control* and *test* buckets.

Both groups saw the same version.

Empirical Analysis

Non-Personalized Module	Personalized Module
Percentage deviation of CTR between buckets consistently **~ 1%**	Percentage deviation of CTR between buckets consistently **> 8%**

Empirical Analysis

Users are assigned numeric scores for their interest categories.

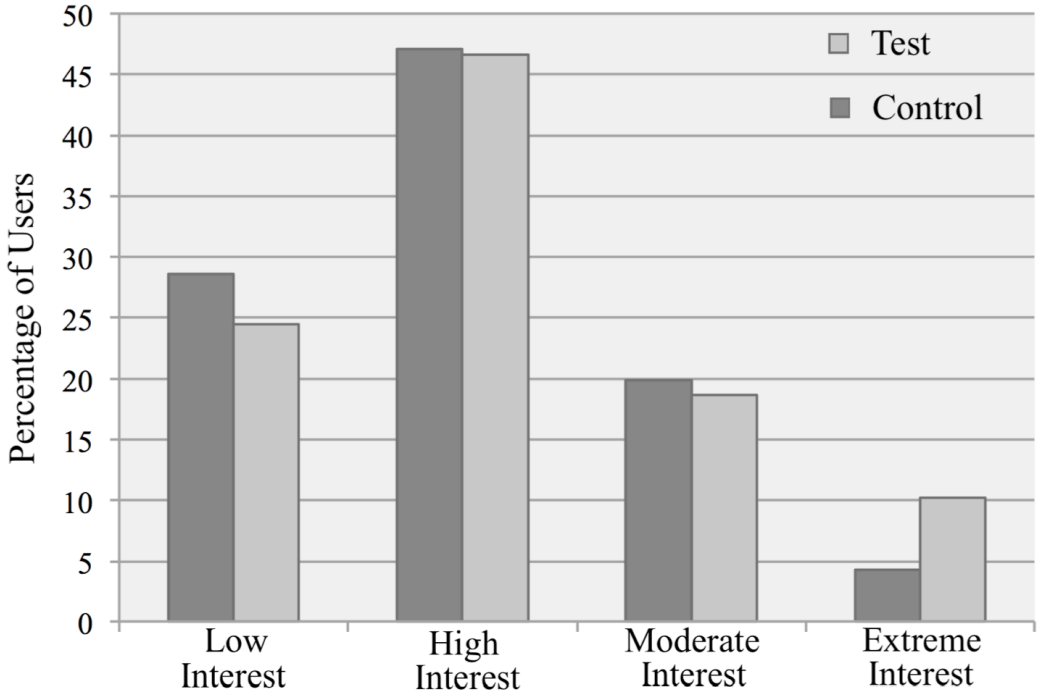
Score in an interest category ← f(user activities)

Given an interest category, we classified users into different interest groups according to their degree of interest based on their scores.

Empirical Analysis

Sports article in personalized module had highest CTR in test bucket.

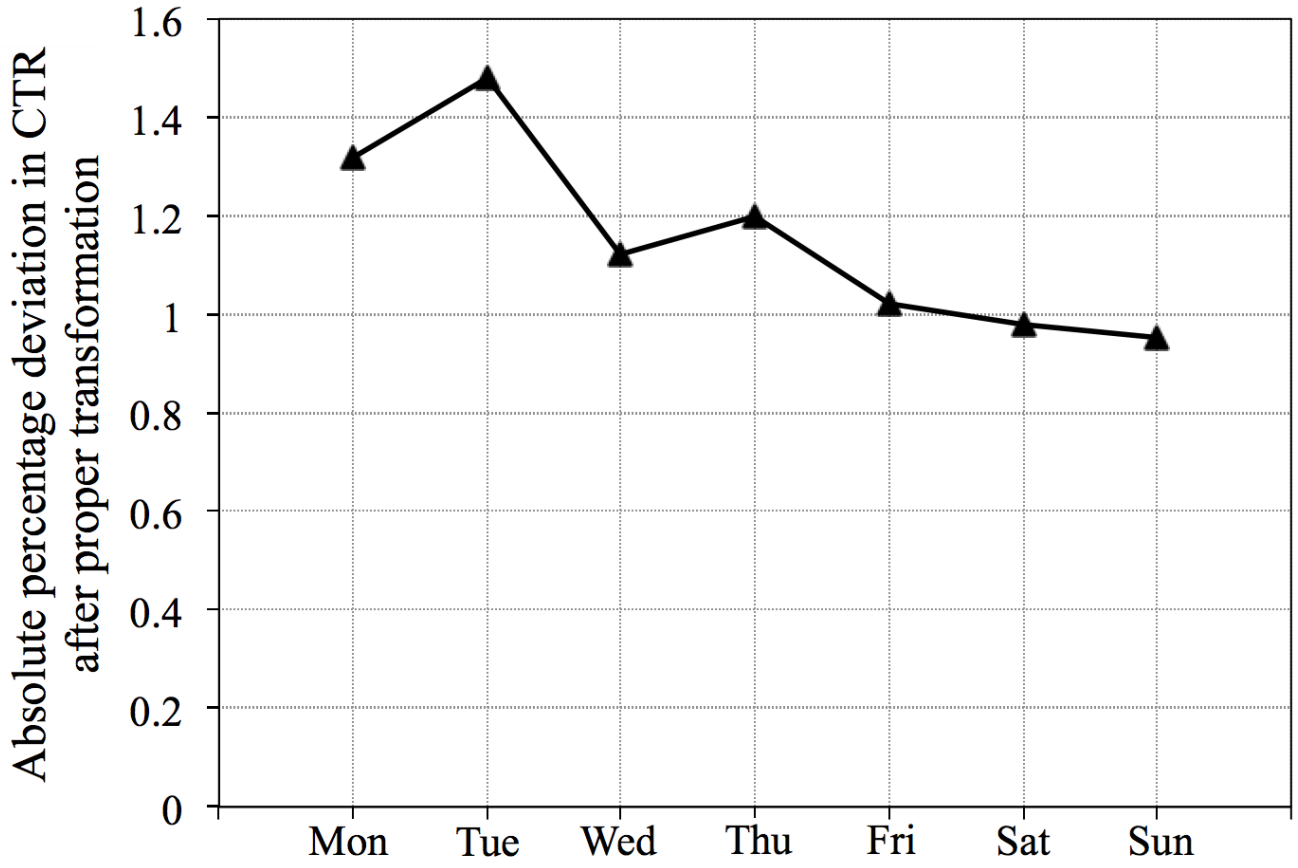
Same article received much lower CTR in control bucket.



User Interest Group	Absolute Percentage Deviation of CTR in Test Bucket over Control Bucket
Low	1.12%
High	0.96%
Moderate	1.15%
Extreme	0.78%
Overall	10.59%

Metric Normalization

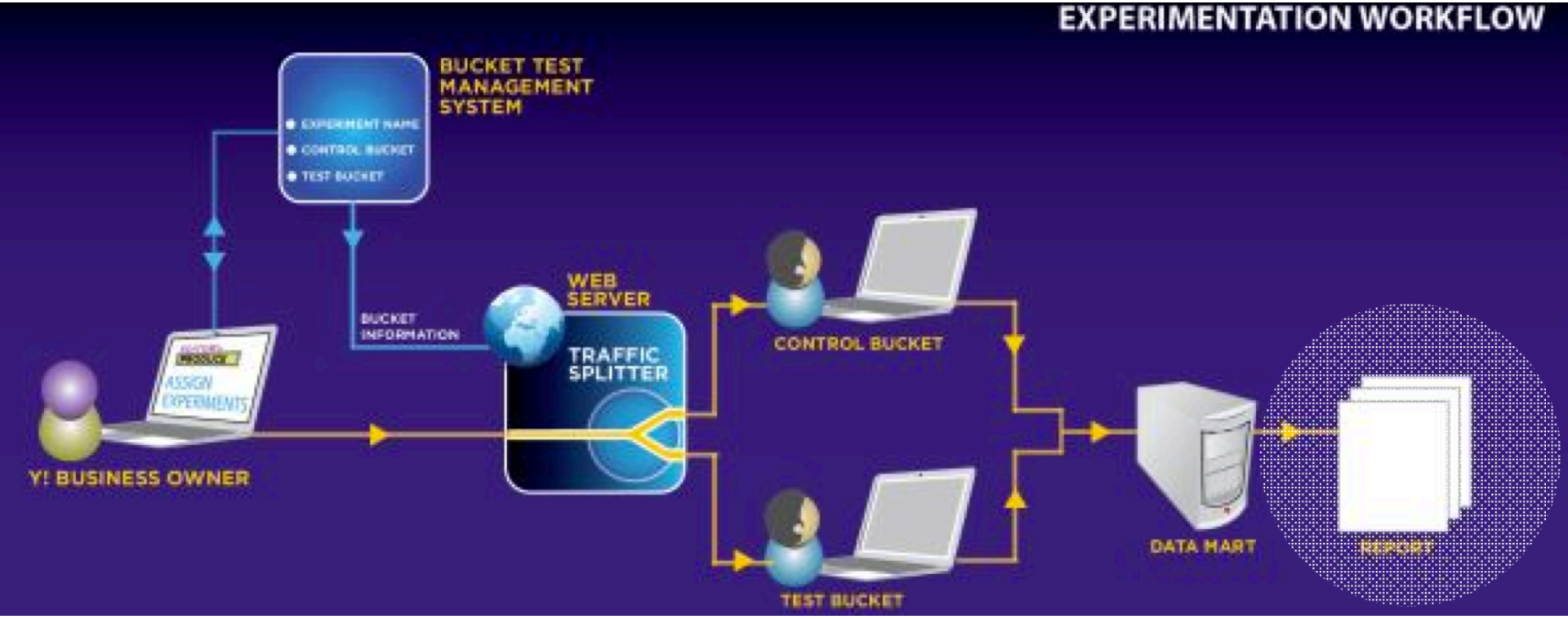
Metrics can be normalized to avoid misinterpreting bucket testing results.



Future Work

Build intelligence into traffic splitter to partition users effectively.

Page loading time for test and control groups *cannot to be compromised*.



Summary

Bucket testing used to forecast impact of product changes.

Personalization can adversely impact bucket testing.

Normalizing metrics, based on user interest scores, mitigates the impact.

Enhance traffic splitter to perform effective bucketization at the onset.

Acknowledgments

Shrikant Srinivas Naidu, *Yahoo Labs, Bangalore*

Rashmi Mohan, *Yahoo Labs, Bangalore*