Challenges in A/B Testing

ALEX DENG @ MICROSOFT 2015

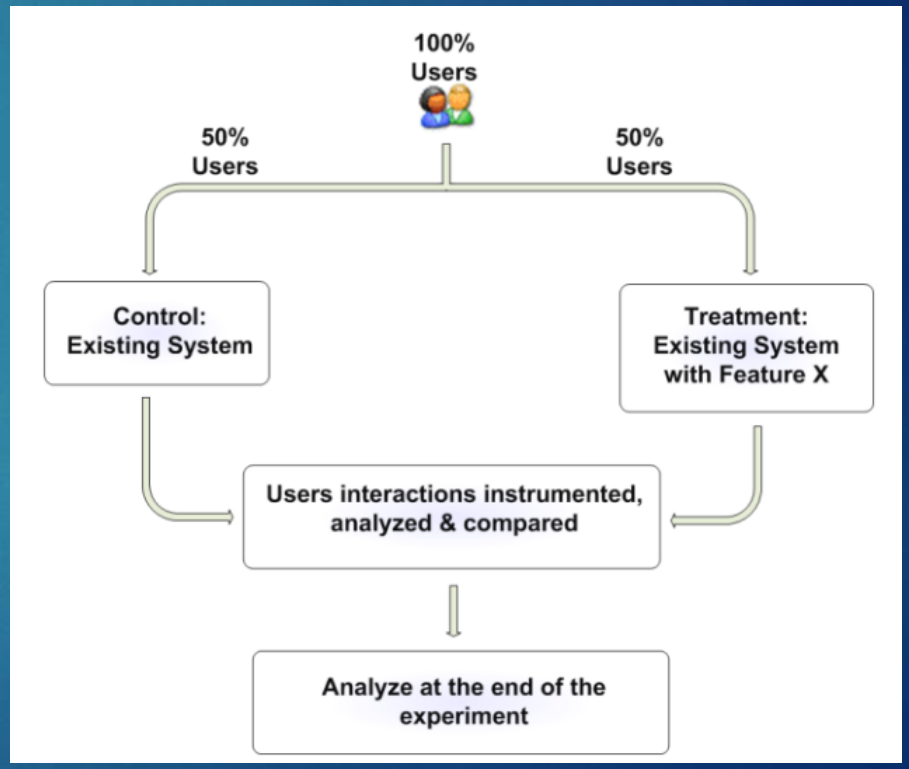
AMAZON TECH TALK

A/B/n Tests in One Slide

- Randomly split traffic between two (or more) versions
  - A (Control)
  - B (Treatment(s))
- Collect data and analyze

- Online Controlled Experiment
- Best scientific way to establish causality
  - Observational data analyses hard and error prone

- …… But there are still many challenges



Five Challenging Problems

- Ronny Kohavi summarized 5 challenging problems in his recent follow-up talk after KDD
- Five challenges (paraphrase)
  1. Metric sensitivity
  2. Problems with NHST and p-value
  3. Beyond Population Average Treatment Effect
  4. Novelty Effects and Experiment Duration
  5. Leakage/Violations of SUTVA(stable unit treatment value assumption)

This Talk

- ExP's mission: "***accelerating*** *innovation through* ***trustworthy*** *analysis and experimentation*"
- Tension between Agility and Quality, closely related to Challenge #1 and #2
- I will spend most of time on #1 and #2 and share some ongoing work for #3. #4 and #5 are something I haven't had thought yet
- Many works are published and can be found on www.exp-platform.com and on my website
- I will stay high level for things involving ongoing and unpublished work
- Time permitted, I want to talk about my opinion of several popular competitions of A/B Testing: multi-armed bandit, Bayesian A/B Testing, MOE

Challenge 1: Metric Sensitivity

- P(Detect a Movement) = P(Detect a Movement | Movement) (1)
  × P(Movement) (2)
- Statistical Power mainly concerns (1)
- P(Movement) can be more fundamental. If your ideas don't work, don't expect ***trustworthy*** analysis to save you
- If your OEC don't move, understand which part is the bottleneck (how?)

Case 1: Statistical Power

- Increase traffic/"power up", limited to capacity
- Run longer, won't always work (Ronny's Sessions/UU example)
- Variance Reduction (CUPED/Doubly Robust Estimation)
  - improve your statistical test, same metric but pure power increase
- Transformation and capping of highly-skewed metrics
  - changed metric definition a little bit
  - Interpret with caution

Case 2: P(Movement) is low

- A perfectly designed metric is not actionable if you can not move it
- What should I do?
  - Re-engineer your metric. You need different OEC at different stages of your product. DAU(daily active users) easy to move for a new product, but harder for matured sites
  - Session Success Rate and Time To First Success moves a lot more than Sessions/UU
  - Define a new surrogate metric as a function of metrics with higher P(Movement) and reasonably statistical power. Calibrate the function form so that the surrogate metric aligns with your OEC
    - Linear combination is easy to work with
    - Optimization problem: maximize metric sensitivity given constraint of alignment

Decompose Metric Sensitivity

- p = P(Movement)
- Observe $\Delta_i \sim N(\mu_i, \sigma_i^2)$, where σ_i assumed to be known
- $\mu = 0$ if no movement
- $\mu_i \sim F$ if $\mu_i \neq 0$ (movement)
- Problem: given a dataset of historical observations Δ_i, how can we estimate p and distribution F
- For a parametric F (e.g. normal or doubly exponential), EM algorithm can help to fit p and F (Deng 2015, Johnstone and Silverman 2004)
- For general nonparametric F, active research area

Bing Results

Metric	P(H0)	P(H1)
X1	97.63%	2.37%
X2	99.80%	0.20%
X3	90.60%	9.40%
X4	98.77%	1.23%
X5	78.55%	21.45%
X6	97.41%	2.59%
X7	97.75%	2.25%
X8	35.50%	64.50%
X9	85.73%	14.27%
X10	98.35%	1.65%
X11	89.25%	10.75%
X12	81.02%	18.98%
X13	73.79%	26.21%
X14	65.57%	34.43%
X15	71.18%	28.82%
X16	66.74%	33.26%
X17	68.12%	31.88%

Device	Metric	PFlat
Mobile	X	66.07%
Desktop	X	81.02%
Mobile	X(Capped)	61.85%
Desktop	X(Capped)	75.19%

- User Engagement Metrics harder to move, e.g. active days per user, visits per user
- Revenue easier to move than engagement
- Signals on a module or part of page much easier to move than whole page
- Capping metrics for highly skewed distribution increased sensitivity (KDD 2013, Online Controlled Experiments at Large Scale) by increasing power
- Variance Reduction method helps (CUPED, WSDM 2013) by increasing power
- Different devices, product areas have different priors

Challenge 2: NHST and p-value

- NHST (Null Hypothesis Statistical Test): assume null hypothesis(no movement) as ground truth, try gathering enough evidence to reject this assumption
- P-value quantifies the strength of your evidence. Loosely speaking p-value = P(Data | H0)
- NHST is the de-facto standard in most scientific research today, including A/B testing
- But it was born in early 20th century. We need new methodology for the Internet era.

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research[1].

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia, tweeted:

- Many published research findings found not reproducible.
  - Notable/Surprising results even more so

  "The fluctuating female vote: Politics, religion, and the ovulatory cycle"
  - Many results with small p-value fails Twyman's law
  - Many cheerful results we observed won't survive confirmation run
- P-value hack
  - Multiple Testing: Team keep making minor changes to a feature and keep running A/B testings until they get small p-value (5% chance to get p-value<0.05!)
  - Optional stopping/continuous monitoring: stop the test once the p-value is "statistically significant"

Problems of NHST

- Null and Alternative is asymmetric.
  - Test only try to reject null, and gather evidence against the null
  - Even with infinite data, will never accept the null with 100% confidence
- Multiple testing: because of the asymmetry, multiple testing can only favor the alternative
- Optional Stopping/Early stopping
- "Genuine" Prior information not used
  - Researchers motivated to publish counter-intuitive results, which are more often not reproducible
  - Twyman's law: any piece of data that looks interesting or unusual is probably wrong

Objective for A/B Testing

- Feature owner: I want 0% Type-I error and 100% statistical power, i.e. test me whether my feature is good or bad correctly every time
  - Mission impossible under uncertainty. Intrinsic trade-off
- Organization:
  - There are always mistakes as long as noise. 0% Type-I error -> 0% statistical power
  - In the **long run**, we want **majority** of features shipped are good for our users
- Long run could be a month in a company like Microsoft, where different teams are using A/B Testing
- "Majority" should be quantified and controlled, is 51% enough?
- That's the beauty of A/B testing **at scale**, we benefit from law of large number

P(H0 | Data), not P(Data | H0)

- If we only ship a feature if P(H1 | Data) > x% (P(H0 | Data) < 1-x%), then we know x% of the ship decisions are correct
- P(H1 | Data) is the Bayesian posterior belief of the alternative hypothesis, it is closely related to the concept of FDR(False discovery Rate)
- Many people misunderstood p-value as P(H0 | Data), and therefore treat 1-pvalue as "confidence" of a correct ship decision
  - A few commercial A/B testing tools use "confidence" instead of p-value

Bayesian Two Sample Hypothesis Testing

1. H0 and H1, with prior odds

$$PriorOdds = \frac{P(H1)}{P(H0)}$$

2. Given observations, likelihood ratio

$$LR = \frac{P(Data|H1)}{P(Data|H0)}$$

3. Bayes Rule

$$\frac{P(H1|Data)}{P(H0|Data)} = PriorOdds \times LR = \frac{P(H1)}{P(H0)} \times \frac{P(Data|H1)}{P(Data|H0)}$$

Frequentist NHST vs Bayesian: Two Trial Systems

- Frequentist:
  - One group of jury, with **presumption of innocence**, reckoning evidence of being guilty
- Bayesian:
  - Two groups of jury, one reckon the evidence of being guilty, the other reckon the evidence of being innocent
  - Judge make final decision based on decisions of both jury, together with prior belief
- Benefit of two jury system
  - Symmetry
  - Principled, not opportunistic anymore. Think multiple testing, both two groups of jury will share the same multiple testing dividend and the judge can still make a balanced call

Bayesian Advantages

- Solves many(not all) multiple testing issues
- Supports optional stopping/early stopping
- Useful Prior information
  - Prior should be learned objectively, not set subjectively! (P(Movement) in #1!)
- More intuitive result
- Accepting the Null: ship based on no harm
- Meta Analysis: combine results from different studies
  - Very useful if you run same experiment multiple times

P-Assessment

- Empirical Bayesian Framework allows us to estimate posterior of negative, positive, and flat movement
- We call it P-Assessment: [PNeg, PFlat, PPos]
- Use PNeg and PFlat for feature shipping
- Use PFlat for shipping with no harm

Delta	Delta %	Conf. Interval for Delta %	P-Assessment
-0.0019	-0.59%	(-1.05%,-0.13%)	96.0% loss, 4.0% flat
-0.0009	-0.27%	(-1.23%,+0.68%)	10.5% loss, 0.5% win, 89.0% flat
-0.0031	-0.93%	(-1.89%,+0.02%)	81.9% loss, 18.1% flat

Challenge 3: Beyond Population Average Treatment Effect

- When we say "treatment effect" most cases we refer to Population Average Treatment Effect (ATE or PATE)
- We know treatment effect differs from person to person
  - A feature might not be popular in some markets -> improvement
  - A feature might be broken on one browser -> bug
- There could be many micro-structure in subpopulations, where treatment effect varies, or even flip sign!
- Heterogeneous Treatment Effect (HTE): Hot topic in Economics/Policy Evaluation, Personalized Treatment/Drug, etc.

- ExP creates different segments and provide segmented scorecard
  - Date: daily metric and effect time series
  - Browsers
  - Markets
  - …
- Challenge
  - we need to find HTE for people, not expecting them to look for it
  - Segment only provides first order HTE, what about higher order effect such as Browser X on weekends

Machine Learning Framework

- Recall $\tau = Y(1) - Y(0)$ (difference of potential outcomes/counterfactual)
- Given covariates X, we want to learn $\tau(X) = E(\tau|X)$, i.e. regression of τ on X. **WLOG, we assume τ is a function of X, i.e. don't distinguish τ and $E(\tau|X)$**
- If we observe (τ, X), this would be a typical supervised learning task. Find a predictor $\hat{\tau}(X)$ according to optimization criteria
- Theoretical loss: $E((\tau - \hat{\tau})^2)$
- Empirical loss function: $\frac{1}{\mathrm{N}} \times \sum(\tau_i - \hat{\tau}(X_i))^2$
- But we don't observe τ!

Modified/Transformed Outcome

- $T = \pm 1$, 1 for Treatment and -1 for control
- Define $Y^* = 2YT$
- Observation: $E(Y^*|X) = \tau(X)$!
- Loss function

$$E\left(\left(Y^* - \hat{\tau}(X)\right)^2\right) = E\left((Y^* - \tau)^2\right) + E\left(\left(\tau - \hat{\tau}(X)\right)^2\right) + E\left((Y^* - \tau)\left(\tau - \hat{\tau}(X)\right)\right)$$

Note that

- $E\left((Y^* - \tau)\left(\tau - \hat{\tau}(X)\right)\right) = 0$ (first condition on X, $E(Y^* - \tau|X) = 0$)
- $E\left((Y^* - \tau)^2\right)$ is constant, i.e. does not depend on $\hat{\tau}$

Minimize $E\left(\left(Y^* - \hat{\tau}(X)\right)^2\right)$ is equivalent to minimize $E\left(\left(\tau - \hat{\tau}(X)\right)^2\right)$!

Empirical loss: $\frac{1}{N}\sum\left(Y_i^* - \hat{\tau}(X_i)\right)^2$

- Now we can employ different machine learning algorithms
- Different algorithms cover different spaces of $\hat{\tau}$
- **Athey and Imbens**: use regression tree and cross-validation for tree pruning
  - No problem with nested segments/multiple covariates X
  - Good for categorical covariate as well as continuous covariate
- Athey and Imbens also changed loss function slightly for training(in sample) and testing(out of sample) and demonstrate improvement
  - They called their winning method Causal Tree (CT)

Issues and Limitations of Tree

- Works only on absolute delta, not percent delta
  - This is due to the function search space of the tree algorithm. Need some custom modifications for percent delta
- Overfit
  - Cross validation only controls the tree size, not the structure
  - Tree structure very unstable/high variance
- Each split reduces sample size, and make subsequent signal weaker
- Could be hard to interpret
  - Tree splitting are hard to follow
  - You need to summarize from all leaves, and number of leaves could be large

Linear Models + Lasso

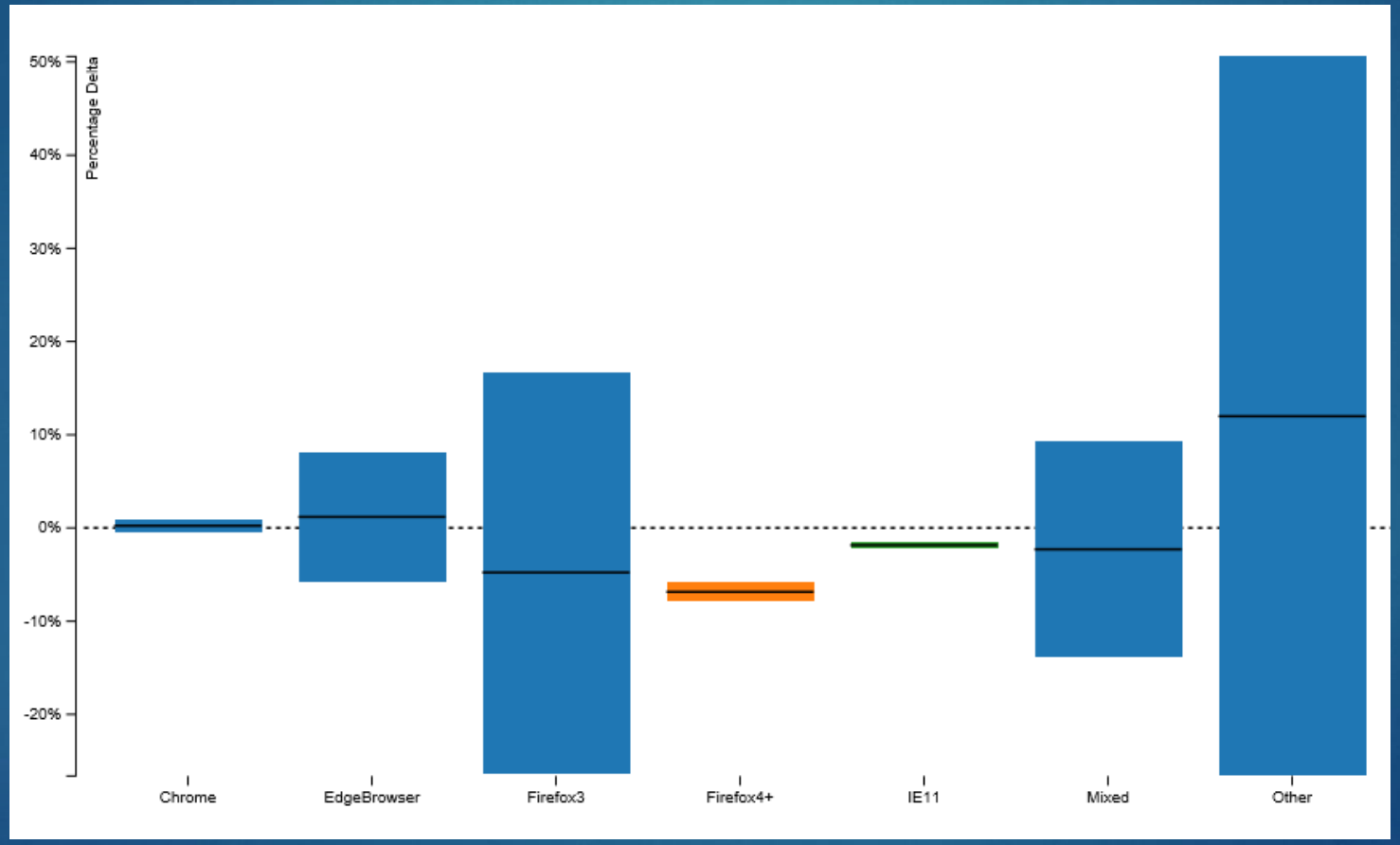
- τ = GlobalEffect + FirstOrder Effect + SecondOrder Effect
- Global Effect is intercept
- First Order Effect is the main effect of each segments: WeekEnd is a% more than WeekDay, etc.
- Second Order Effect is the interaction effect between segments
- Tian, et. al. (with Tibshirani) used this together with Transformed Covariate (Equivalent to Transformed Outcome in our case)
- Pros: Good interpretation. **Parsimonious representation**.
- Cons:
  - Still high False positive (40% when 50 covariates)
  - Lasso on categorical variables need special care, Grouped Lasso still not satisfactory.

Working in Progress

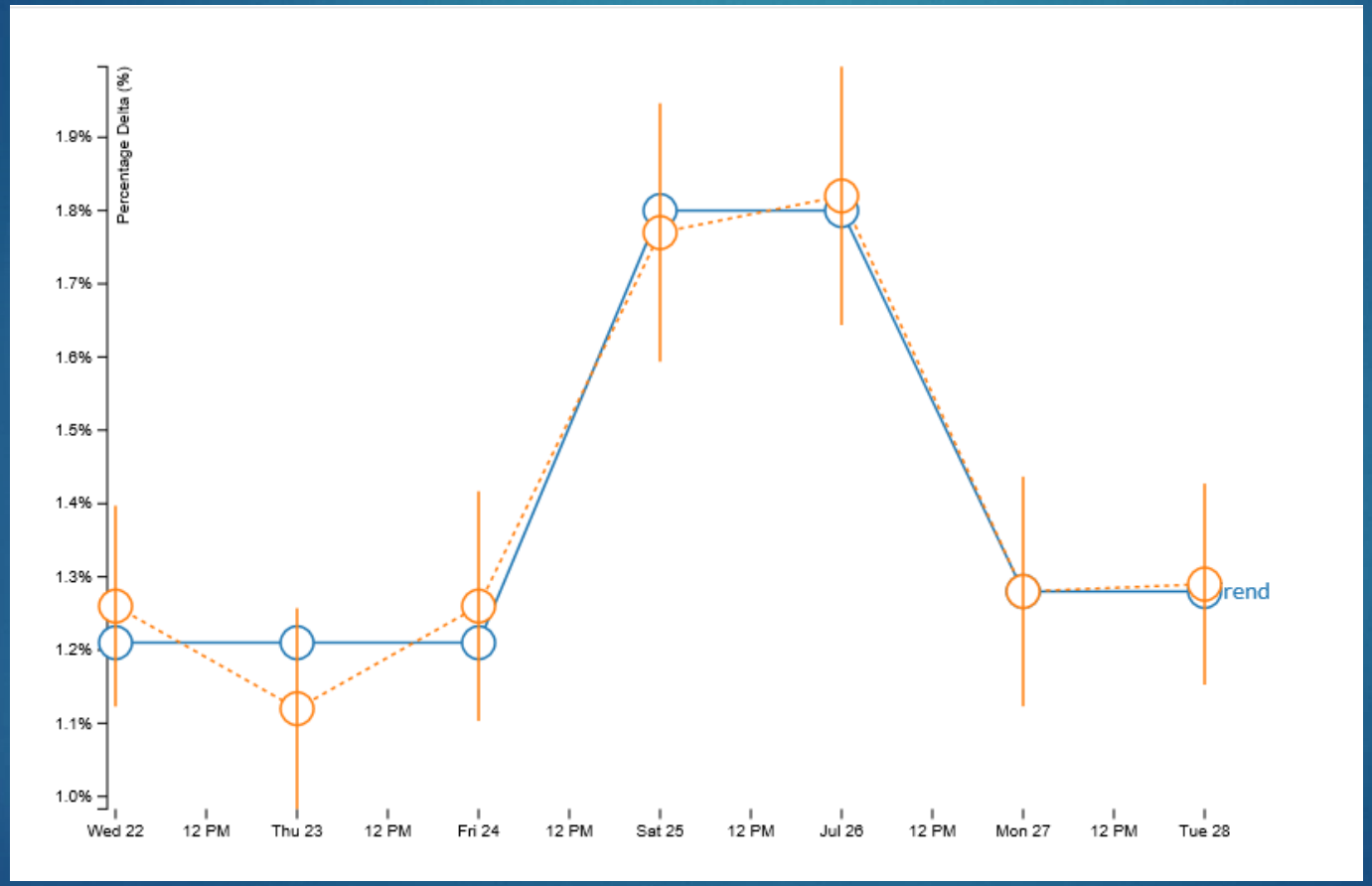
Main ideas:

- Use linear model/multiplicative linear model(percent delta) for good interpretation
- First order effect is like “clustering”
- Step-wise regression
  - Find the covariate with the “highest first order effect”
  - Take residual, and then continue, until no first order effect remains
- Then use Lasso like algorithm to
  - Find second order effects
  - Choose a parsimonious representation of the effects

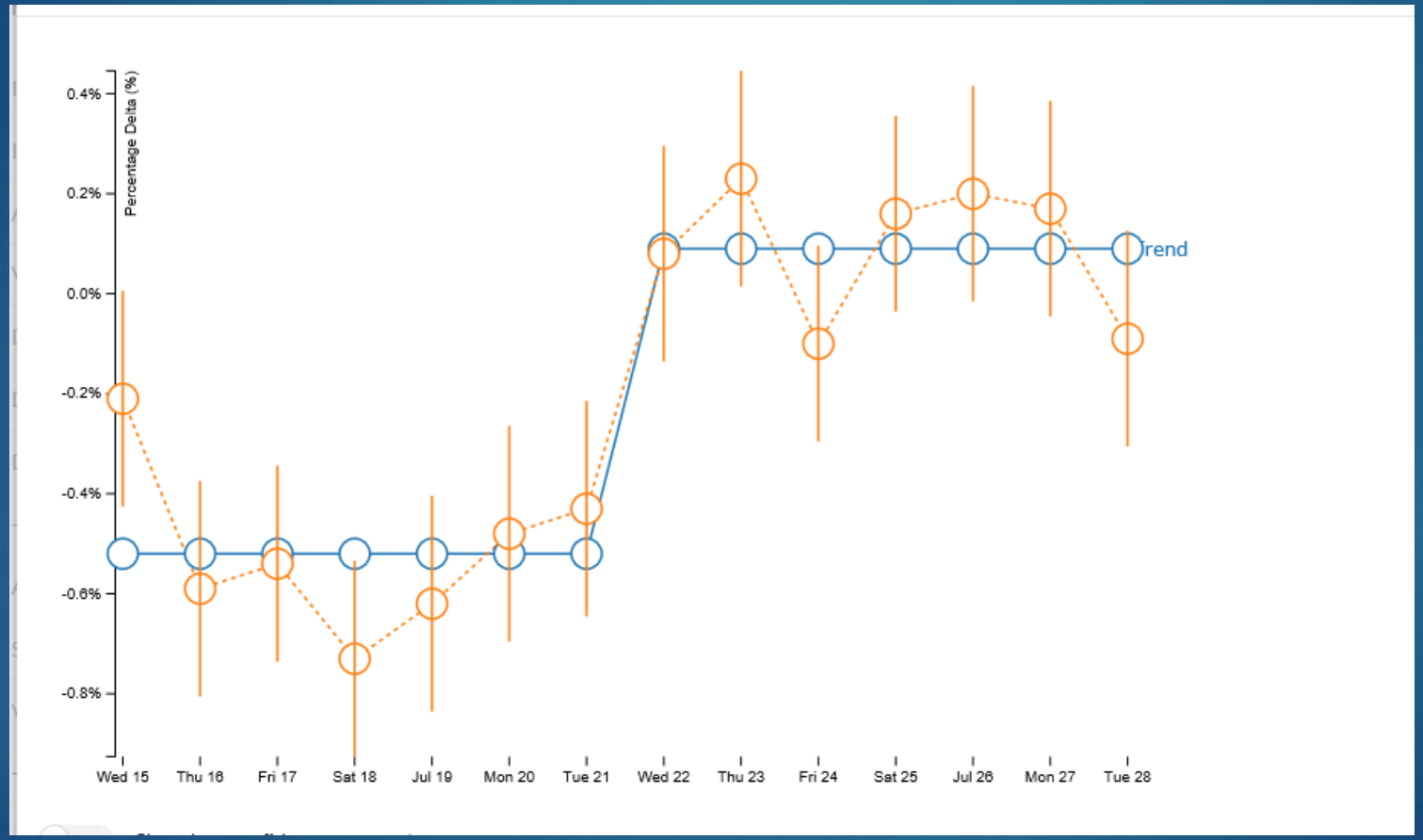
Browser difference



Weekend vs weekday



Shift


Percentage Delta (%)
0.4%
0.2%
0.0%
-0.2%
-0.4%
-0.6%
-0.8%
Wed 15
Thu 16
Fri 17
Sat 18
Jul 19
Mon 20
Tue 21
Wed 22
Thu 23
Fri 24
Sat 25
Jul 26
Mon 27
Tue 28
rend

Competitions

Bayesian A/B Testing

- Not a competition, a complement.
- I don't agree many Bayesian A/B testing procedure you can find online where you just use a uniform or "non-informative" prior. Any prior contains information!
- I believe we are in a **unique** position that we can utilize historical data to use objective prior information, instead of using subjective prior or "non-informative" prior. The subject called Empirical Bayes shines in big data scenario
  - Machine Learning community has been using EB in many problems, where they call it MLE-II
  - It has many nice properties such as "adaptive sparsity", and closely connected to frequentist multiple testing
  - The fact that you learn prior using your data is critical for FDR control

Multi-armed Bandit

- Multi-armed bandit allows you to change traffic splitting dynamically given data.
- Limited application: static effect, independent observations, no carry-over effect, etc.
- Contextual multi-armed bandit/multi-verse experiment is a very interesting active learning idea
  - In some sense many products maps context to result(recommender system, search engine, etc.),
  - You use live traffic feedback as your labeled data for (context, action) pairs
  - Unbiased evaluation of different policy exist with inverse propensity weighting. Idea is with some level of exploration/randomization, your algorithm can gradually learn to perform better
  - In practice, the propensity need to be bounded away from 0 and 1, so randomization tend to be fixed, which makes it very close to A/B testing with fixed traffic split

Metric Optimization Engine(MOE)

- "optimize anything that has tunable parameters"
- I think it is different from A/B Testing with a very specialized problem
- You have one or more tuning parameter and want to find the optimum points
- You don't know the curve, but you can get i.i.d. observations with noises for each given parameter setting
- Naive method:
  - Learn values for different parameters and compare them -> naïve A/B testing with a large number of treatments
  - Multi-armed bandit
- MOE:
  - You can "learn experience from others" by putting a model for the curve. (think regression)
  - MOE at its core is a Bayesian smoothing trick + multi-armed bandit

Question?

- www.exp-platform.com
- alexdeng.github.io