# Trustworthy Analysis of Online A/B Tests: Pitfalls, Challenges and Solutions

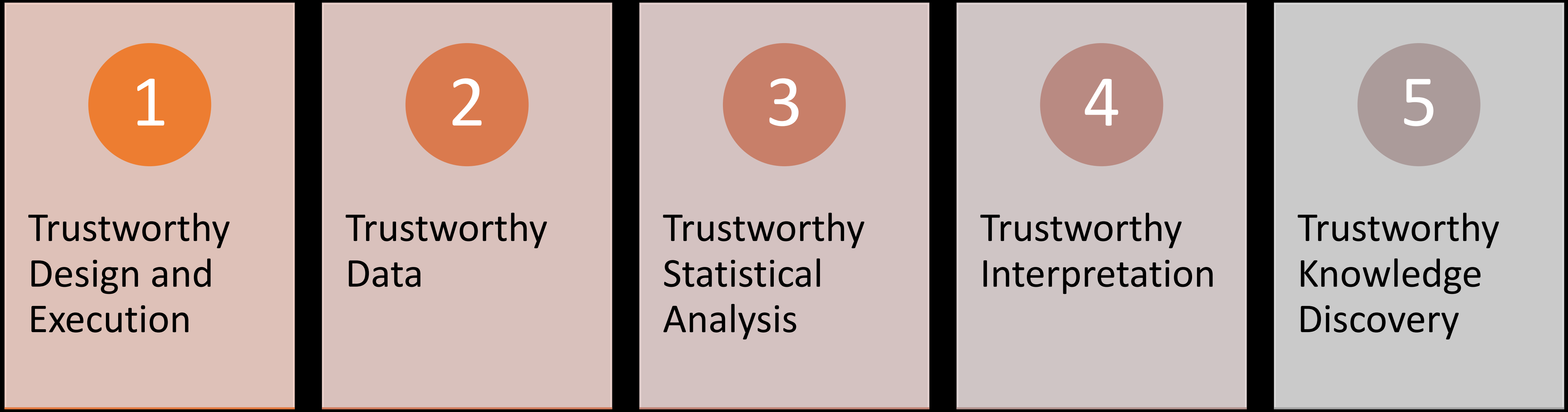
**Alex Deng @ Microsoft 2017**

**Expedia Test Summit**

# About Myself

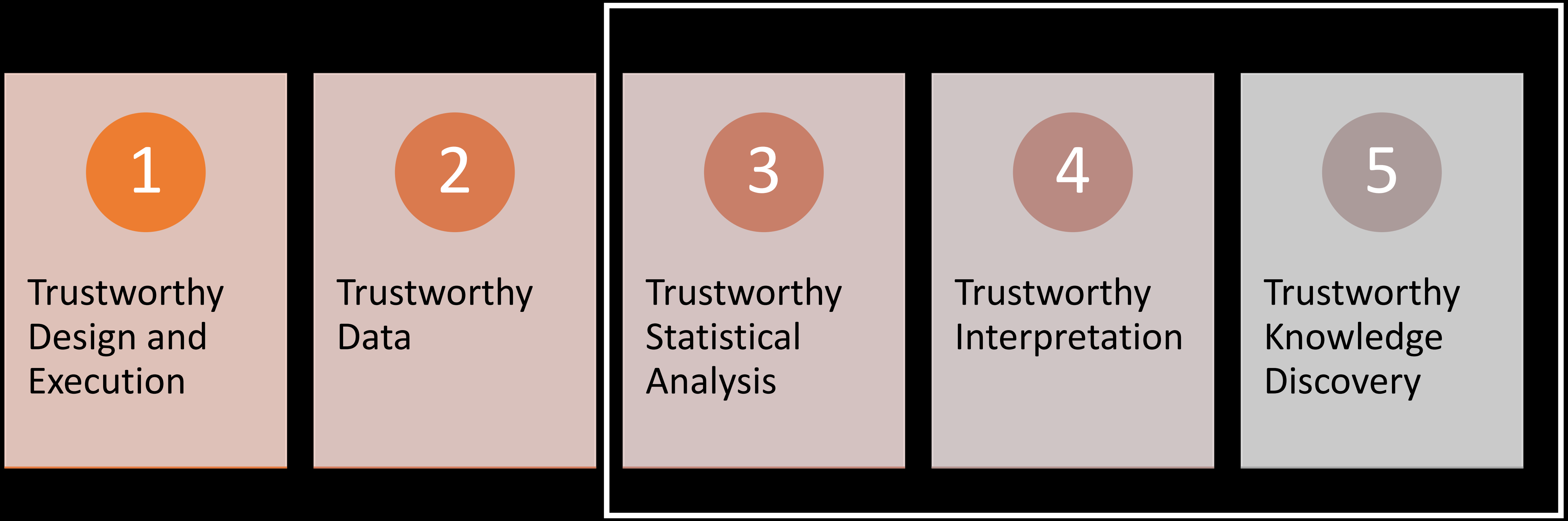
- 7 years in ExP team
- Statistician by training, nowadays better known as Data Scientists
- Publications in KDD, WWW, WSDM
- Interested in statistical problems combined with engineering challenges ☺
- Learn more about my work at alexdeng.github.io

# Trustworthy A/B Tests



1. Trustworthy Design and Execution
2. Trustworthy Data
3. Trustworthy Statistical Analysis
4. Trustworthy Interpretation
5. Trustworthy Knowledge Discovery

# Trustworthy A/B Tests



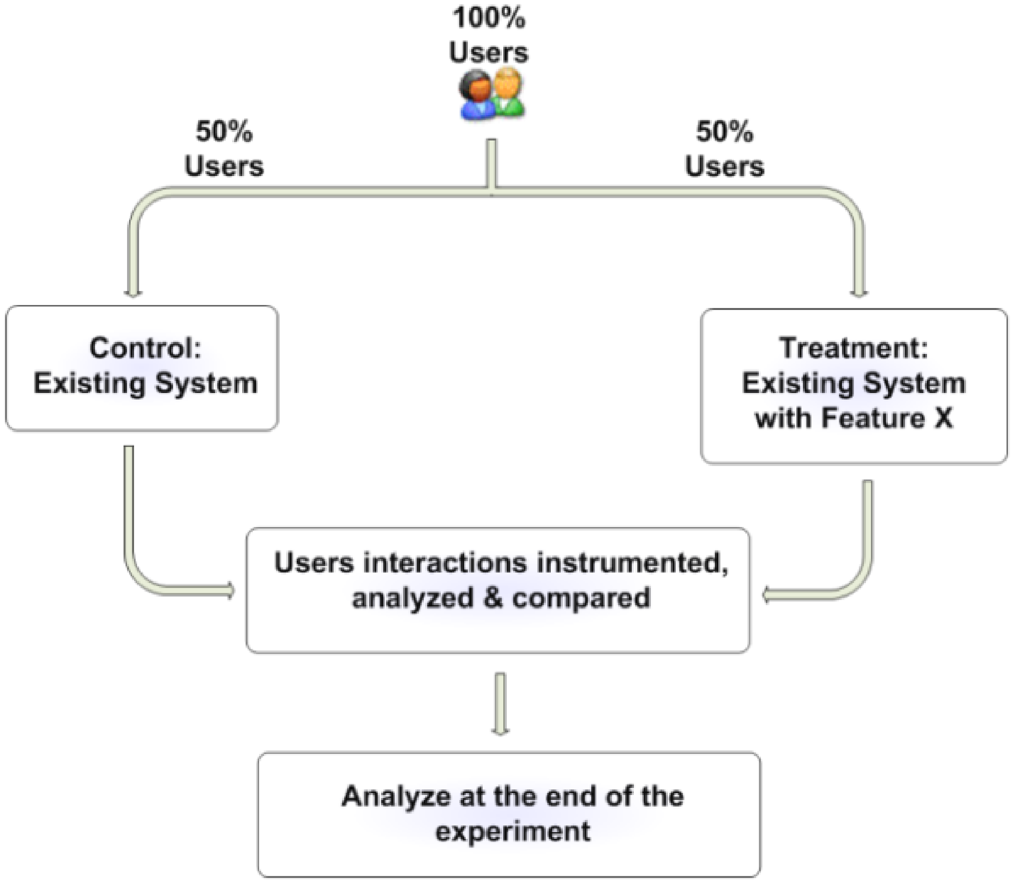
1 Trustworthy Design and Execution

2 Trustworthy Data

3 Trustworthy Statistical Analysis

4 Trustworthy Interpretation

5 Trustworthy Knowledge Discovery


100%
Users
50%
Users
50%
Users
Control:
Existing System
Treatment:
Existing System
with Feature X
Users interactions instrumented, analyzed & compared
Analyze at the end of the experiment

# 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

# Trustworthy Statistical Analysis

Assumptions underneath large scale A/B tests

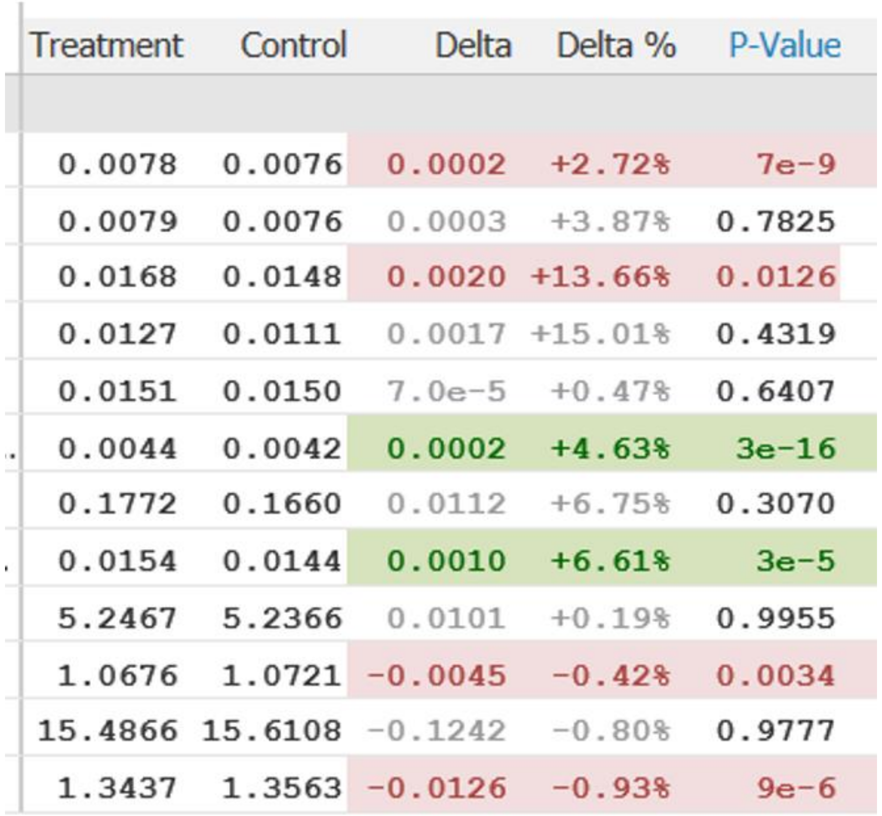
1. Randomization is performed on a fixed unit, e.g. user, page-view, document, game-session
2. Independence (i.i.d.)
3. Normal approximation by central limit theorem

# Beware of two units

- Randomization unit: where randomization is performed
  - Most common: **User based**
  - Also used:  **Page-view based**, **location based**, etc.
- Analysis unit: on which a metric is defined
  - **Click-through-rate: Page-view/impression based**
  - **Revenues/User: User-based**

A VERY common mistake: treat measurements at analysis unit as independent

You get an A/A(Treatment = Control) scorecard like this



| Treatment | Control | Delta | Delta % | P-Value |
|---|---|---|---|---|
| | | | | |
| 0.0078 | 0.0076 | 0.0002 | +2.72% | 7e-9 |
| 0.0079 | 0.0076 | 0.0003 | +3.87% | 0.7825 |
| 0.0168 | 0.0148 | 0.0020 | +13.66% | 0.0126 |
| 0.0127 | 0.0111 | 0.0017 | +15.01% | 0.4319 |
| 0.0151 | 0.0150 | 7.0e-5 | +0.47% | 0.6407 |
| 0.0044 | 0.0042 | 0.0002 | +4.63% | 3e-16 |
| 0.1772 | 0.1660 | 0.0112 | +6.75% | 0.3070 |
| 0.0154 | 0.0144 | 0.0010 | +6.61% | 3e-5 |
| 5.2467 | 5.2366 | 0.0101 | +0.19% | 0.9955 |
| 1.0676 | 1.0721 | -0.0045 | -0.42% | 0.0034 |
| 15.4866 | 15.6108 | -0.1242 | -0.80% | 0.9777 |
| 1.3437 | 1.3563 | -0.0126 | -0.93% | 9e-6 |

- Variance of metrics are hugely underestimated (by a factor of 2 or more)

Root cause:

- Analysis units might not be independent of each other
- Metrics are typically defined as an average over the analysis unit
- A naïve engineer will just use sample variance formula from Wikipedia or from his/her favorite package

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$$

- Sample variance formula make the assumption that average was taken over i.i.d. observations. When there exists positive correlation, this variance will underestimate

# Solution: Delta Method

- Suppose we assume users are independent, and our metric is Click-through-rate:

$$\frac{\#Clicks}{\#Page-Views}$$

- Instead of treating it as average of clicks over page-views, treat it as

$$\frac{Clicks/User}{PageViews/User}$$

- Becomes a ratio of two metrics, both are average of i.i.d. observations (since we assume users are independent)
- Delta Method: The ratio metric also converge to a normal distribution and formula for the variance exists

# Independence

- What justifies i.i.d. assumption?
- Are users i.i.d.?
- Are locations i.i.d.?
- Are organizations i.i.d.?

# A short quiz (2min)

There is a large urn full of balls with numbers between 1 to 10

Each time pick one ball, observe the number and then put it back to the urn

Observe a series of numbers. Are these observations independent?

Hint: Independence means knowing previous observations won’t help you predict the next observation

# A short quiz (2min)

Raise your LEFT hand if your answer is YES

Raise your RIGHT hand if your answer is NO

# Answer: Both are correct

Not Independent: Knowing previous numbers help us understand the distribution of numbers in the urn, thus help better predicting the next number

Independent: If we assume the distribution information is public information, e.g. uniform between 1 to 10, then observations are i.i.d. from this distribution

# Answer: Both are correct

Not Independent: Knowing previous numbers help us understand the distribution of numbers in the urn, thus help better predicting the next number

Independent: If we assume the distribution information is public information, e.g. uniform between 1 to 10, then observations are i.i.d. from this distribution

Independence is not justified by theory, but by *choice!*

# Independence and External Validity

Users(or any randomization units) always share some common environment

If we see this common environment as fixed, then we can assume users i.i.d.

If we expect this environment to also be changing, then not

External Validity: whether result can be generalized outside of the context of the experiment

# Randomization Unit Principle

(WSDM2017) RUP: Randomization unit can typically be treated as independent

- Search Ads experiment randomize on page-views -> pageviews i.i.d.
- Xbox game randomize on game-session -> game-sessions i.i.d.
- Skype randomize on call -> calls i.i.d.

Of course randomization unit has to be chosen appropriate to avoid jarring experience switching

# Complex Randomization

01

## Client Experimentation

Mobile app and desktop app need to be working with or without network connection

Randomized assignments are pushed to client every hour

Clients only receive new assignment when connected

Clients apply changes at the next refresh window, e.g. app open or wake from background

02

## Social Sharing

Experiment a new way of sharing. Randomized by sharers and treatment got new sharing

Interested in conversion rate

Can share with multiple people

From a receiver perspective, you receive both treatment and control sharing messages depends on who share with you

# Complex Randomization

01

Client Experime

Mobile app and des
with or without netv

Randomized assignr
hour

Clients only receive ne
connected

Clients apply changes at the next refresh window, e.g. app open or wake from background

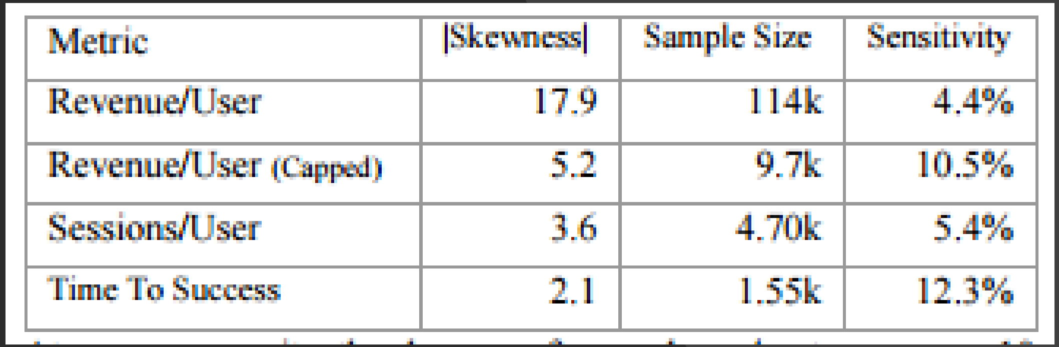
[WSDM2017]
General Variance Formula
that is generally applicable

ring. Randomized by
w sharing

le

e, you receive both treatment and control sharing messages depends on who share with you

# Normal Approximation

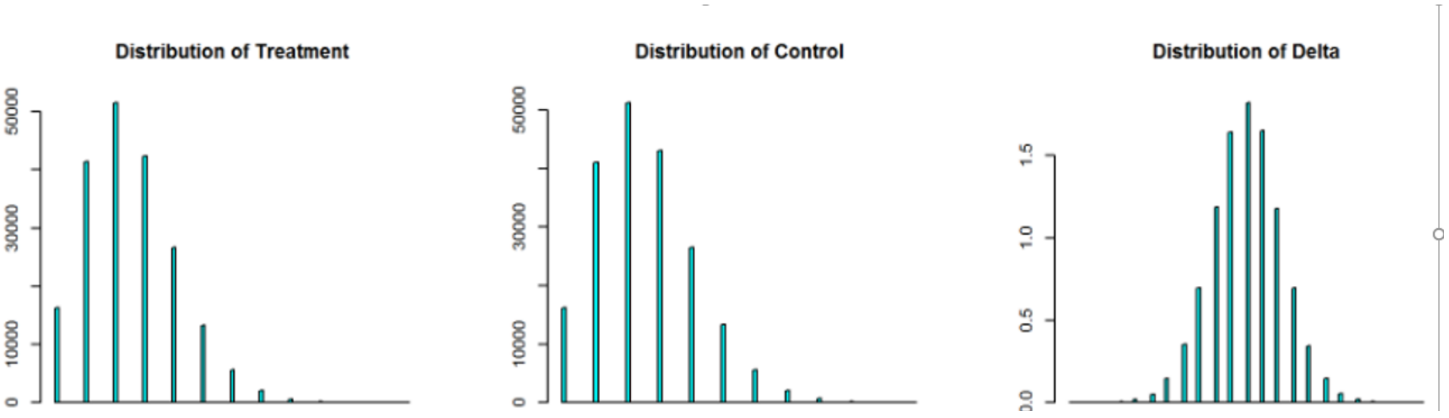


| Metric | \|Skewness\| | Sample Size | Sensitivity |
|---|---:|---:|---:|
| Revenue/User | 17.9 | 114k | 4.4% |
| Revenue/User (Capped) | 5.2 | 9.7k | 10.5% |
| Sessions/User | 3.6 | 4.70k | 5.4% |
| Time To Success | 2.1 | 1.55k | 12.3% |

- Central limit theorem requires “large enough sample”
- How large is enough: rule of thumb 355× $skewness^2$
- Some metrics like Revenue/UU have large skewness (long tail, 0 inflated)
- Another common type of large skewness metric: Rare Event Rate, e.g. event with very small conversion rate

Detail at http://bit.ly/expRulesOfThumb

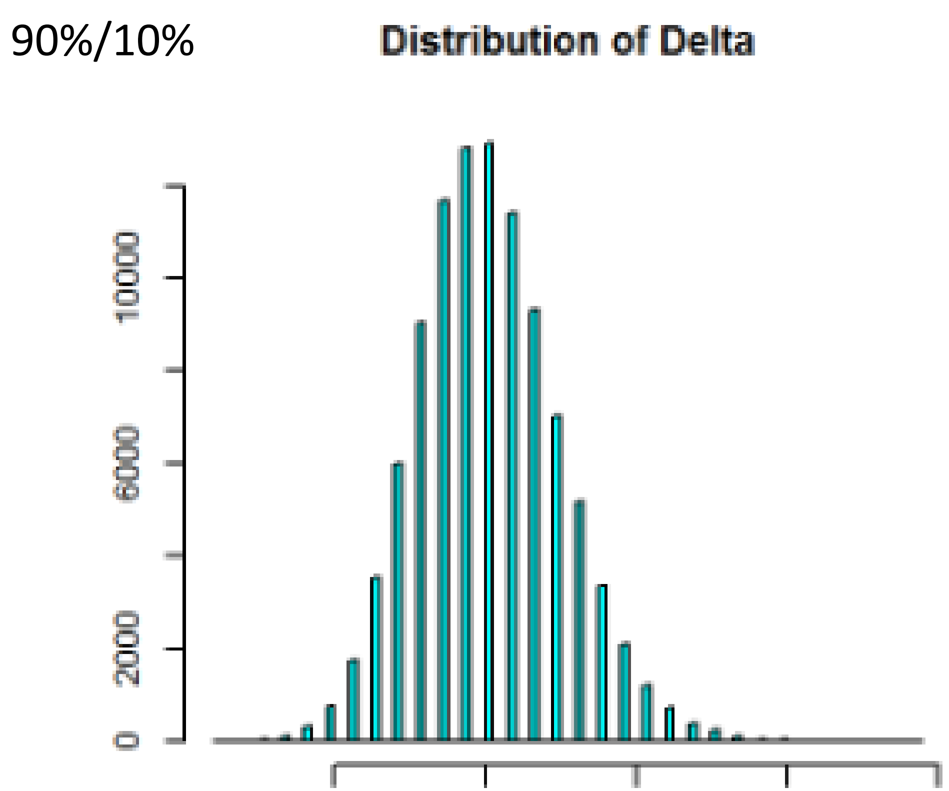
# Solution: Balanced Design

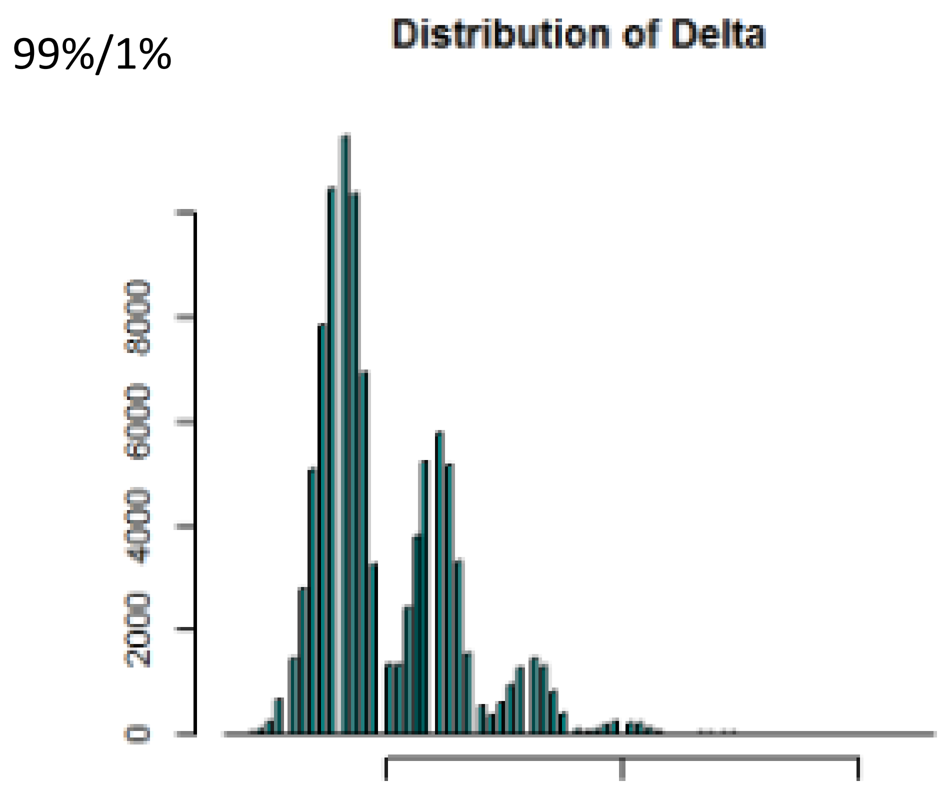


- When treatment and control have same traffic size, normal approximation of the **Δ** of treatment and control metrics kicks in way faster than unbalanced design

Left Figure:

- Both treatment and control metrics are clearly not normal
- Delta already close to normal


90%/10%


99%/1%

# Without Balanced Design

## Common Pitfalls

- Continuous monitoring of p-value
- Take p-value as the probability that the null hypothesis is true
- Fail to adjust for multiple comparison/testing

# Trustworthy Interpretation

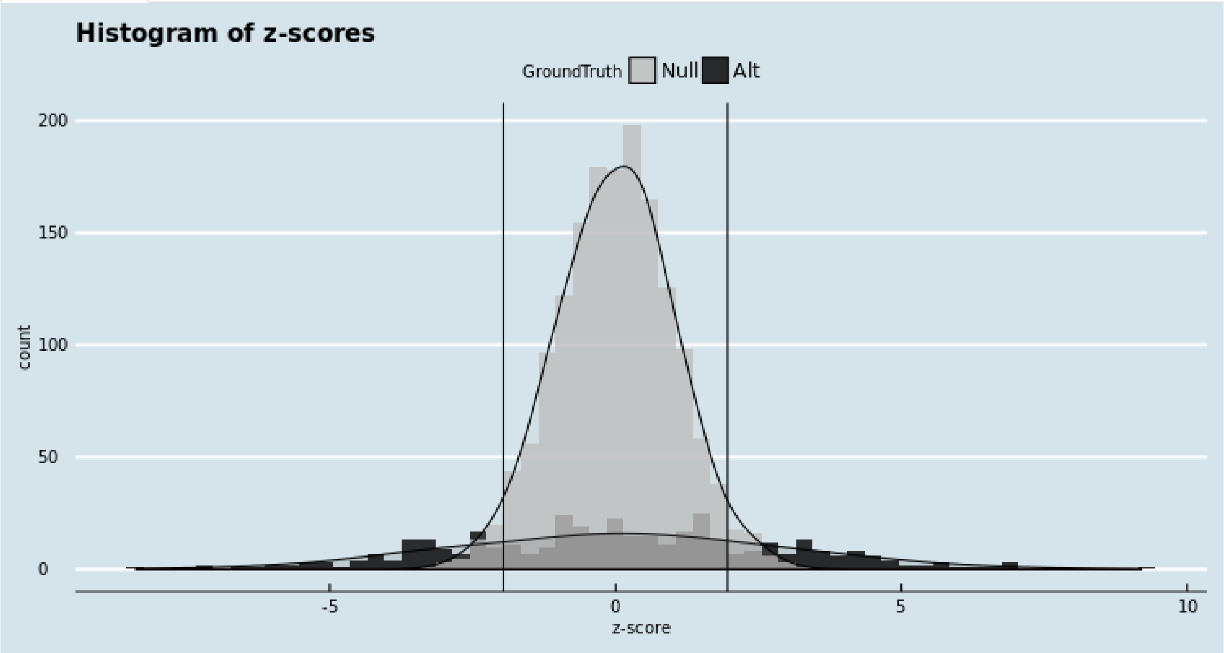
# Misconception of P-value

1. P-value is the probability of the null hypothesis being true
2. Studies with the same p-value provide same evidence against null
3. P-value 0.05 means that if you reject the null hypothesis, the probability of a false positive is only 5%

More from Steven Goodman's *"A Dirty Dozen: Twelve P-Value Misconceptions"*

This is NOT a problem of ***experimenter***, but a problem of the ***platform***

# Using Rich Historical Experiment Data



- P-value assumes the null is true and then computes the False Positive Rate
- But we don’t know P(Null)

Solution: use historical experiments data to estimate P(Null)

# Objective Bayesian Hypothesis Testing

- With P(Null) and P(Alternative), we can truly compute

$$P(\text{Alternative}|\text{Data})$$

- Note 1 – P(Alternative|Data) = P(Null|Data) is also known as the False Discover Rate (FDR)
- FDR allows continuous monitoring: can stop the experiments once FDR is below a threshold
- FDR also adjusts for (most) multiple testing. We can compute FDR for a scorecard of thousands of metrics and make decision

# Challenges

- Are past experiments a good source to learn for new experiments?
- Ignored other rich information:
  - Type of experiment
  - Type of treatment
  - Team’s confidence in the treatment
- Ongoing work

# Trustworthy Knowledge Discovery

Traditional experimentation:

- A fixed set of questions
- Focus heavily on a good design to enable answering all the relevant questions

Modern experimentation with big data:

- A set of primary questions, but more after seeing the results
- Focus on iteration. The goal of one experiment is to figure out what to do next and to test in the next iteration

Hypothesis Testing -> Knowledge Discovery/Machine Learning

# Insight Mining (or chasing noises?)

01

Finding unexpected interactions with other experiments

02

Look at time series of metric difference to reason about novelty effect, trend, weekend effect, etc

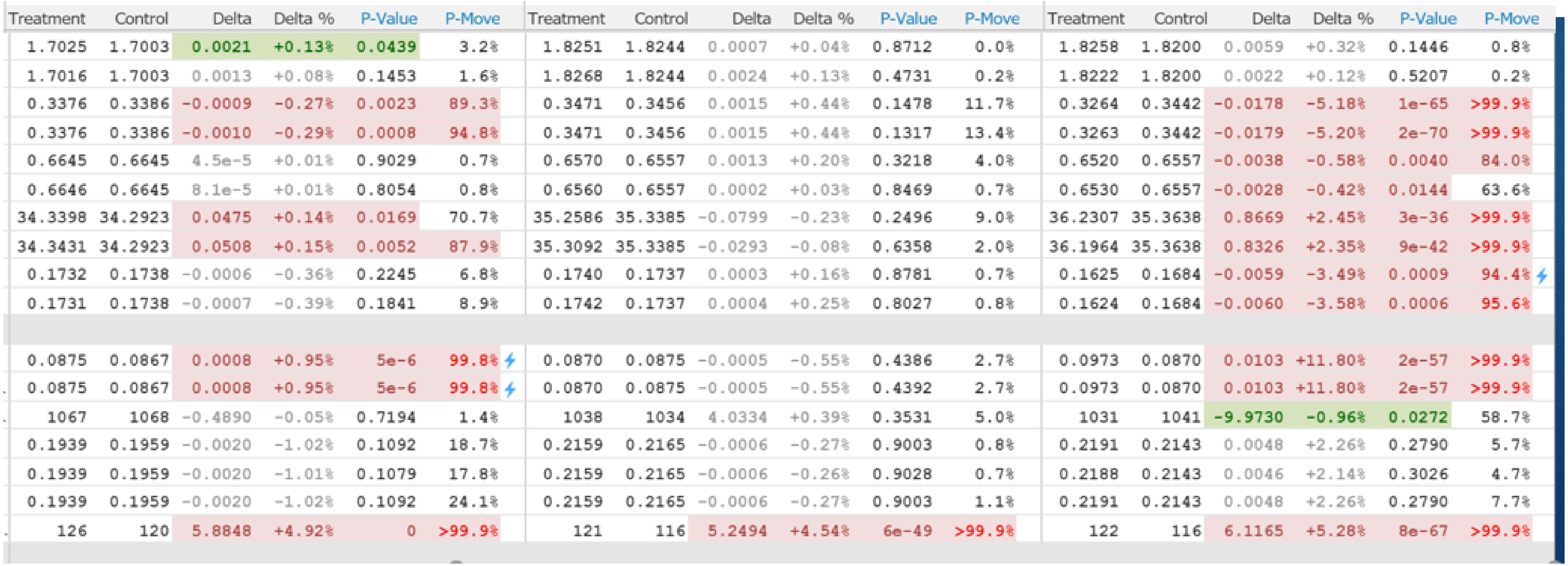
03

Slice and dice using different dimensions and attributes to finally find a subpopulation that the treatment performs well/badly

# Key Ideas

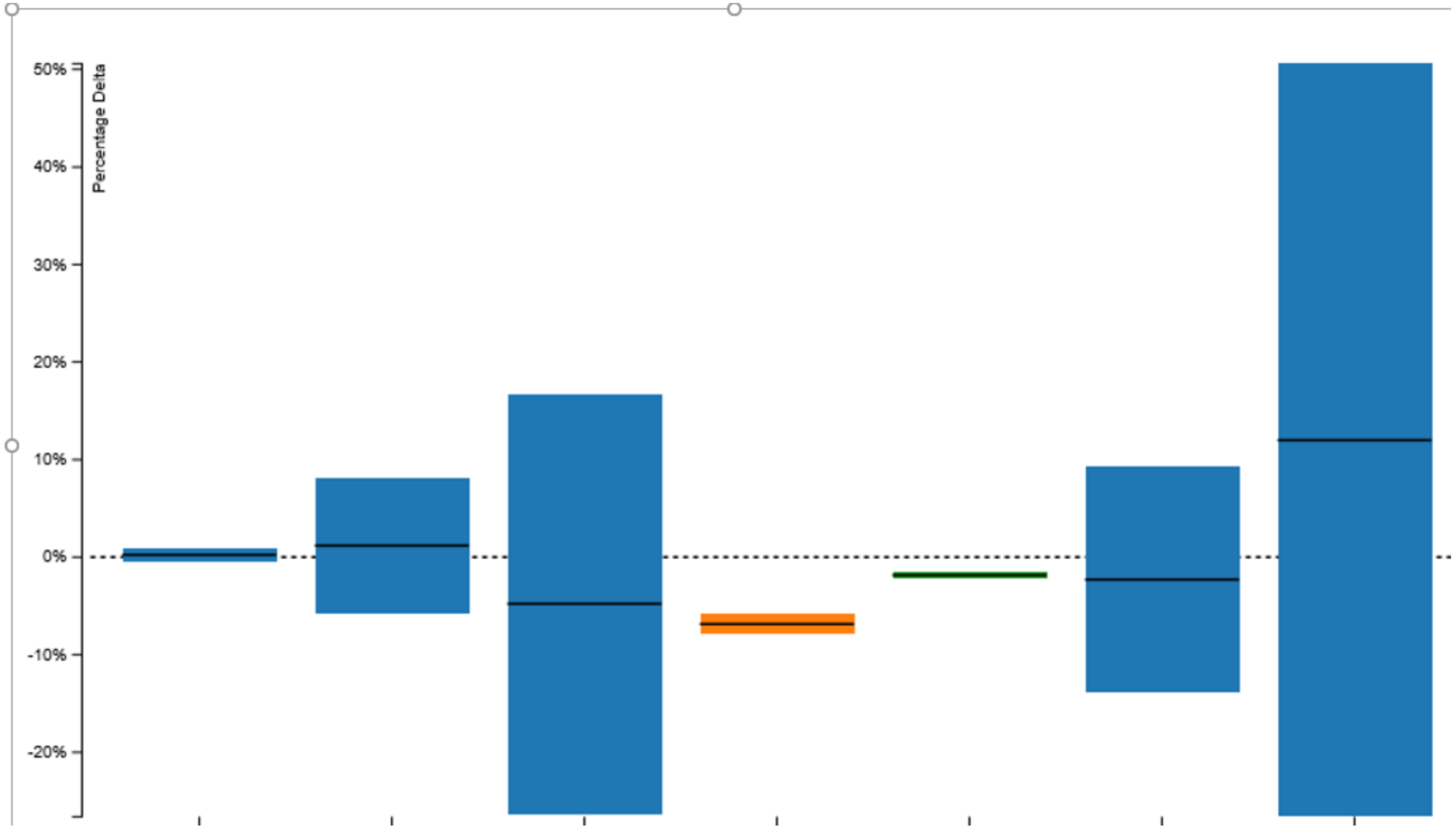
- When signal to noise ratio is low, be skeptical of any ***selected*** results
- Principle of Sparsity/Occam Razor: interesting insights are sparse
  - Truly sparse
  - Not sparse but we are only interested in top few items due to resource constraint
- Research Area: **H**eterogeneous **T**reatment **E**ffect, Personalized Treatment

# Interaction

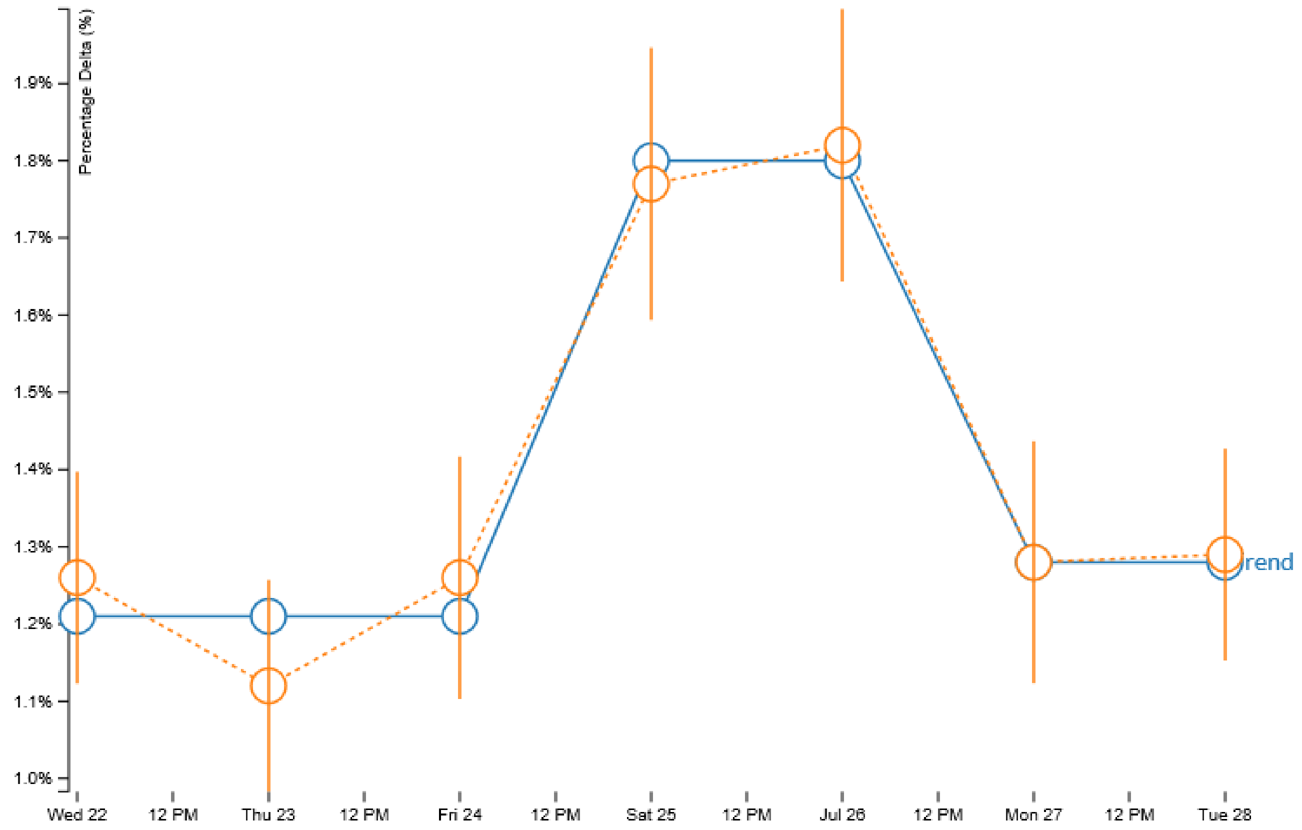


| Overall | | | | | | Without Interaction | | | | | | Interaction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Treatment | Control | Delta | Delta % | P-Value | P-Move | Treatment | Control | Delta | Delta % | P-Value | P-Move | Treatment | Control | Delta | Delta % | P-Value | P-Move |
| 1.7025 | 1.7003 | 0.0021 | +0.13% | 0.0439 | 3.2% | 1.8251 | 1.8244 | 0.0007 | +0.04% | 0.8712 | 0.0% | 1.8258 | 1.8200 | 0.0059 | +0.32% | 0.1446 | 0.8% |
| 1.7016 | 1.7003 | 0.0013 | +0.08% | 0.1453 | 1.6% | 1.8268 | 1.8244 | 0.0024 | +0.13% | 0.4731 | 0.2% | 1.8222 | 1.8200 | 0.0022 | +0.12% | 0.5207 | 0.2% |
| 0.3376 | 0.3386 | -0.0009 | -0.27% | 0.0023 | 89.3% | 0.3471 | 0.3456 | 0.0015 | +0.44% | 0.1478 | 11.7% | 0.3264 | 0.3442 | -0.0178 | -5.18% | 1e-65 | >99.9% |
| 0.3376 | 0.3386 | -0.0010 | -0.29% | 0.0008 | 94.8% | 0.3471 | 0.3456 | 0.0015 | +0.44% | 0.1317 | 13.4% | 0.3263 | 0.3442 | -0.0179 | -5.20% | 2e-70 | >99.9% |
| 0.6645 | 0.6645 | 4.5e-5 | +0.01% | 0.9029 | 0.7% | 0.6570 | 0.6557 | 0.0013 | +0.20% | 0.3218 | 4.0% | 0.6520 | 0.6557 | -0.0038 | -0.58% | 0.0040 | 84.0% |
| 0.6646 | 0.6645 | 8.1e-5 | +0.01% | 0.8054 | 0.8% | 0.6560 | 0.6557 | 0.0002 | +0.03% | 0.8469 | 0.7% | 0.6530 | 0.6557 | -0.0028 | -0.42% | 0.0144 | 63.6% |
| 34.3398 | 34.2923 | 0.0475 | +0.14% | 0.0169 | 70.7% | 35.2586 | 35.3385 | -0.0799 | -0.23% | 0.2496 | 9.0% | 36.2307 | 35.3638 | 0.8669 | +2.45% | 3e-36 | >99.9% |
| 34.3431 | 34.2923 | 0.0508 | +0.15% | 0.0052 | 87.9% | 35.3092 | 35.3385 | -0.0293 | -0.08% | 0.6358 | 2.0% | 36.1964 | 35.3638 | 0.8326 | +2.35% | 9e-42 | >99.9% |
| 0.1732 | 0.1738 | -0.0006 | -0.36% | 0.2245 | 6.8% | 0.1740 | 0.1737 | 0.0003 | +0.16% | 0.8781 | 0.7% | 0.1625 | 0.1684 | -0.0059 | -3.49% | 0.0009 | 94.4% ⚡ |
| 0.1731 | 0.1738 | -0.0007 | -0.39% | 0.1841 | 8.9% | 0.1742 | 0.1737 | 0.0004 | +0.25% | 0.8027 | 0.8% | 0.1624 | 0.1684 | -0.0060 | -3.58% | 0.0006 | 95.6% |
| | | | | | | | | | | | | | | | | | |
| 0.0875 | 0.0867 | 0.0008 | +0.95% | 5e-6 | 99.8% ⚡ | 0.0870 | 0.0875 | -0.0005 | -0.55% | 0.4386 | 2.7% | 0.0973 | 0.0870 | 0.0103 | +11.80% | 2e-57 | >99.9% |
| 0.0875 | 0.0867 | 0.0008 | +0.95% | 5e-6 | 99.8% ⚡ | 0.0870 | 0.0875 | -0.0005 | -0.55% | 0.4392 | 2.7% | 0.0973 | 0.0870 | 0.0103 | +11.80% | 2e-57 | >99.9% |
| 1067 | 1068 | -0.4890 | -0.05% | 0.7194 | 1.4% | 1038 | 1034 | 4.0334 | +0.39% | 0.3531 | 5.0% | 1031 | 1041 | -9.9730 | -0.96% | 0.0272 | 58.7% |
| 0.1939 | 0.1959 | -0.0020 | -1.02% | 0.1092 | 18.7% | 0.2159 | 0.2165 | -0.0006 | -0.27% | 0.9003 | 0.8% | 0.2191 | 0.2143 | 0.0048 | +2.26% | 0.2790 | 5.7% |
| 0.1939 | 0.1959 | -0.0020 | -1.01% | 0.1079 | 17.8% | 0.2159 | 0.2165 | -0.0006 | -0.26% | 0.9028 | 0.7% | 0.2188 | 0.2143 | 0.0046 | +2.14% | 0.3026 | 4.7% |
| 0.1939 | 0.1959 | -0.0020 | -1.02% | 0.1092 | 24.1% | 0.2159 | 0.2165 | -0.0006 | -0.27% | 0.9003 | 1.1% | 0.2191 | 0.2143 | 0.0048 | +2.26% | 0.2790 | 7.7% |
| 126 | 120 | 5.8848 | +4.92% | 0 | >99.9% | 121 | 116 | 5.2494 | +4.54% | 6e-49 | >99.9% | 122 | 116 | 6.1165 | +5.28% | 8e-67 | >99.9% |

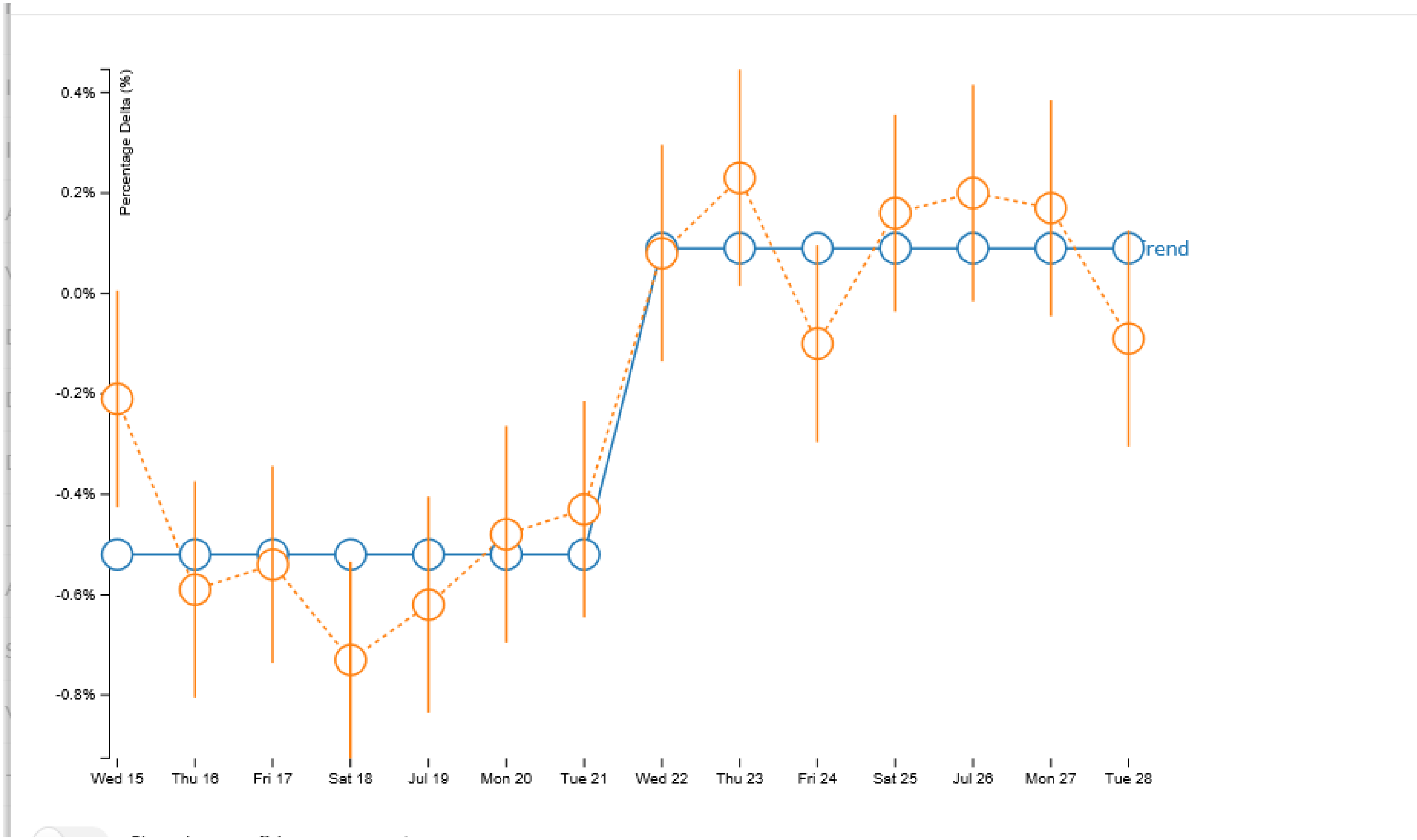
# Browser difference



# Weekend vs weekday



# Shift


Percentage Delta (%)
0.4%
0.2%
0.0%
-0.2%
-0.4%
-0.6%
-0.8%
rend
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

# Question?

More at http://exp-platform.com

Slides available at aka.ms/expedia-summit