Doing online experiments the proper way: true effect or due to chance?

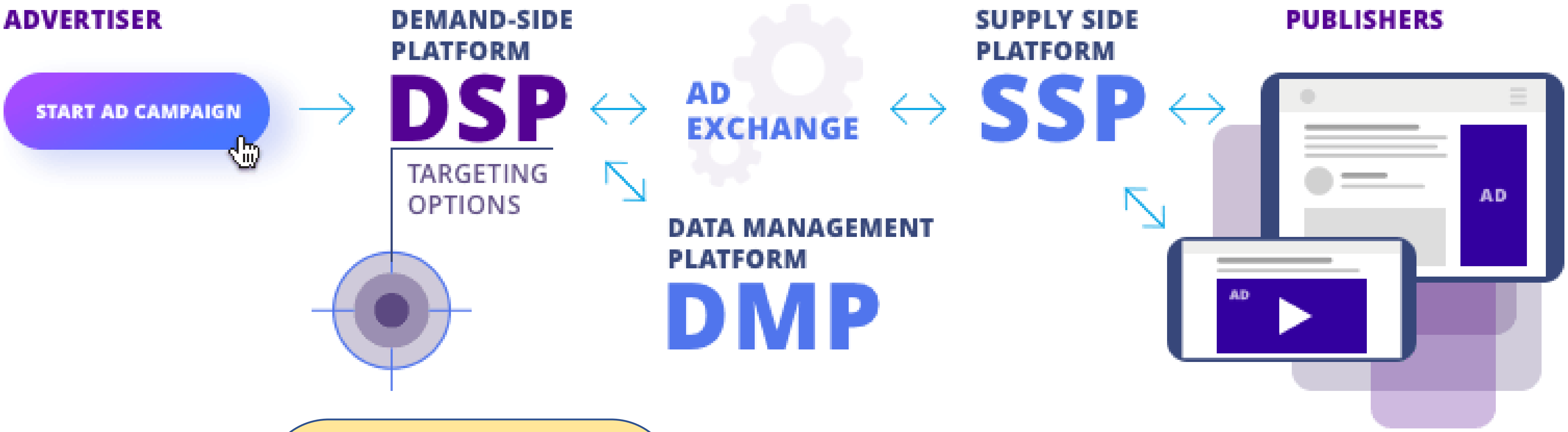
Uyen Nguyen

Data Scientist, Knorex

06/06/2019

Knorex - Precision performance marketing

Demand side platform (DSP)



- Targeting
- Retargeting
- Audience profiling
- Brand safety
- Dynamic creative optimization
- Real time bidding

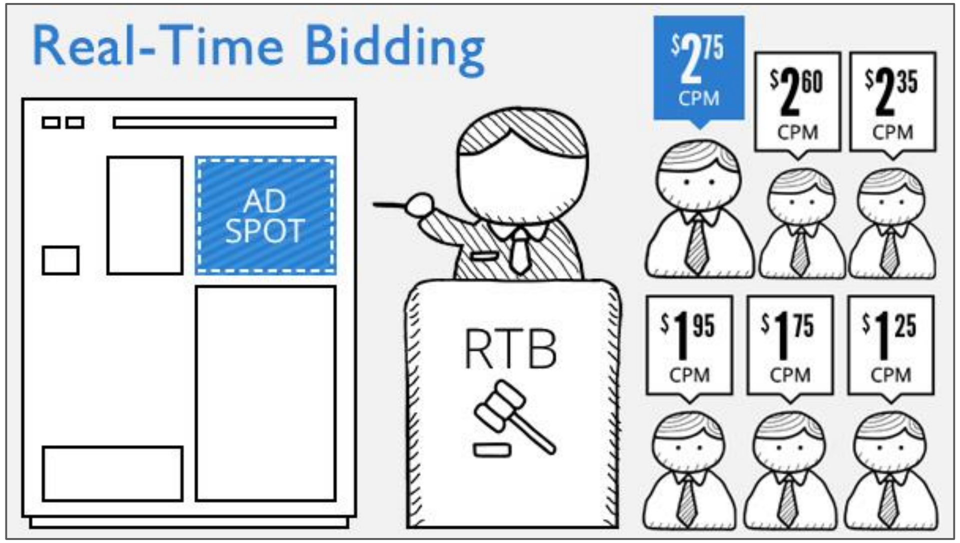
(Source: Quora)

Data Science at Knorex

Bidding strategies to optimize for:

- cost per click (CPC)
- cost per acquisition (CPA)
- cost per installation (CPI)

Predicting chances of click/conversion



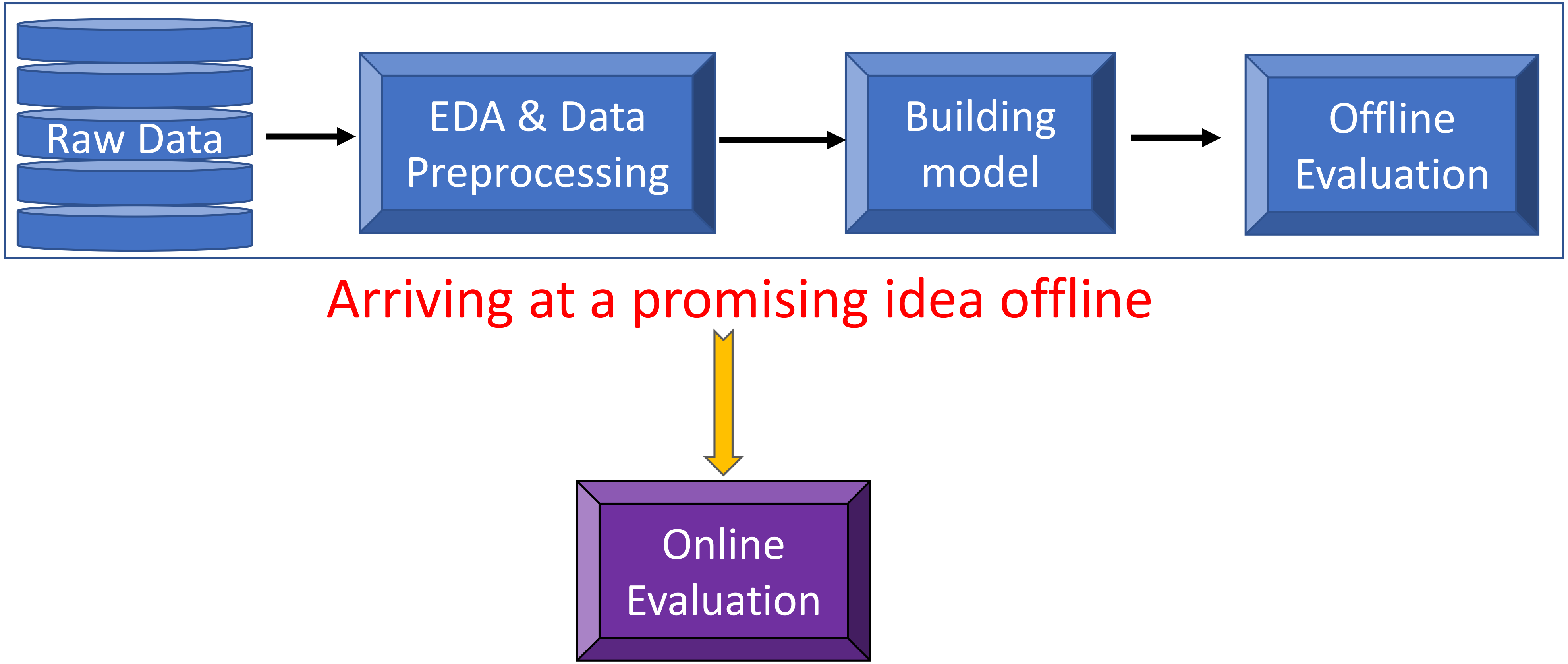
Source: www.sitescout.com

Audience look-alike modeling

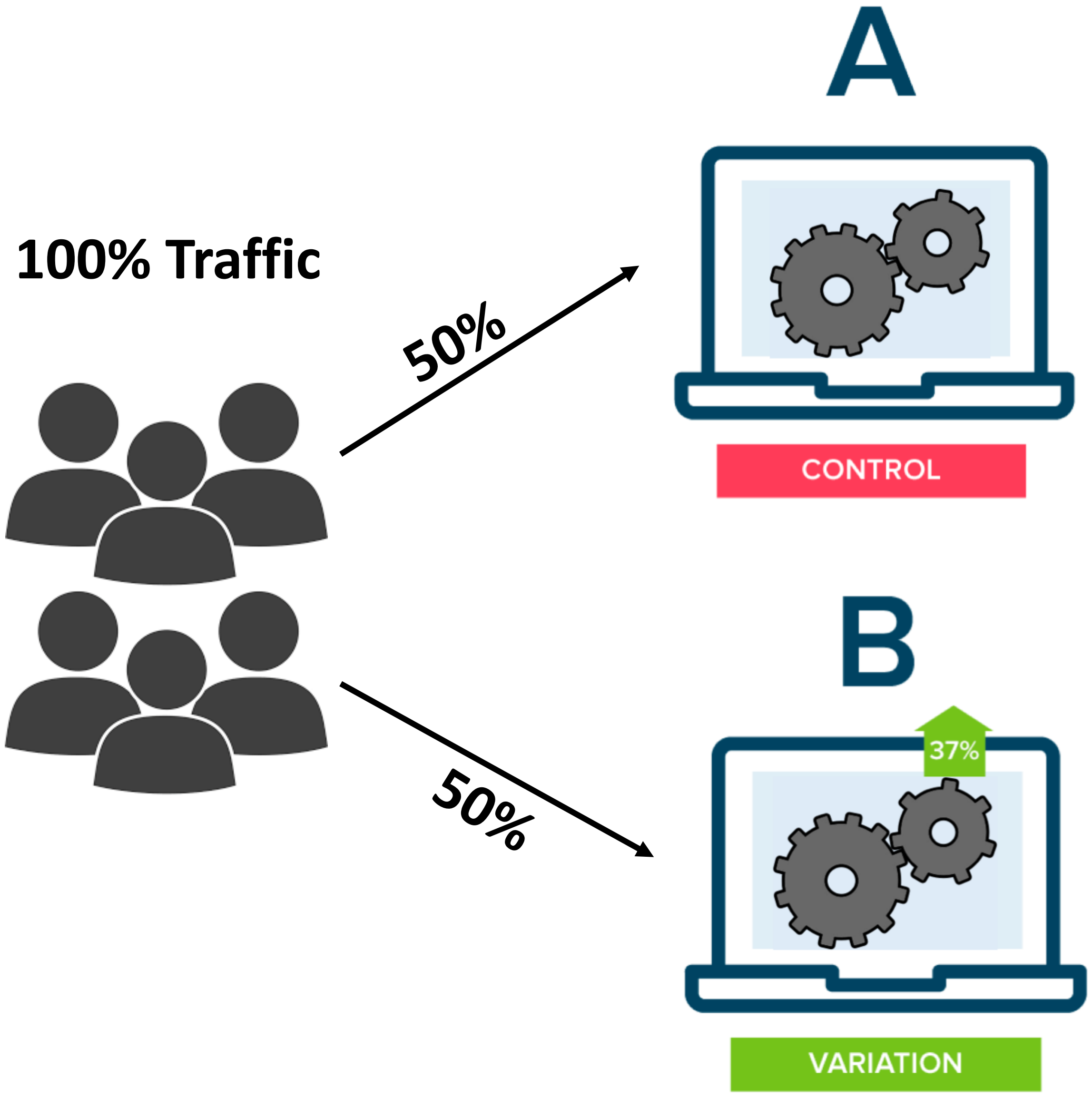
Monitoring performance, strategic troubleshooting

Testing a new idea in production – Online experimentation

- Ad creative
- Bidding strategy
- Machine learning model

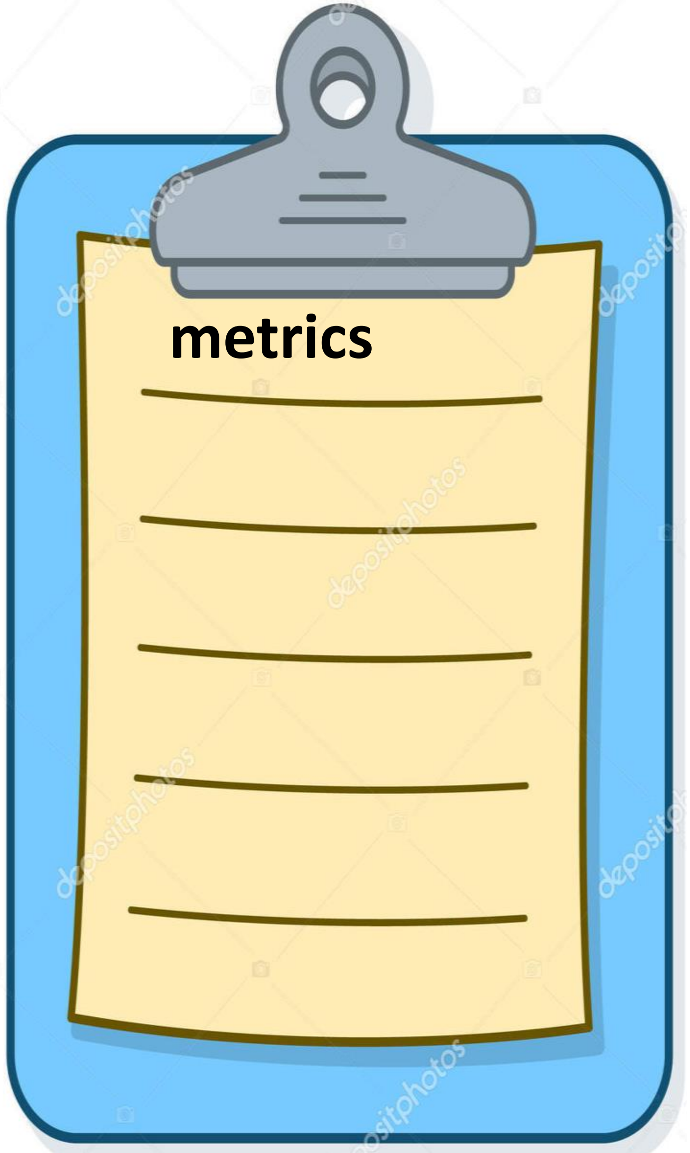


A/B testing – The conventional method



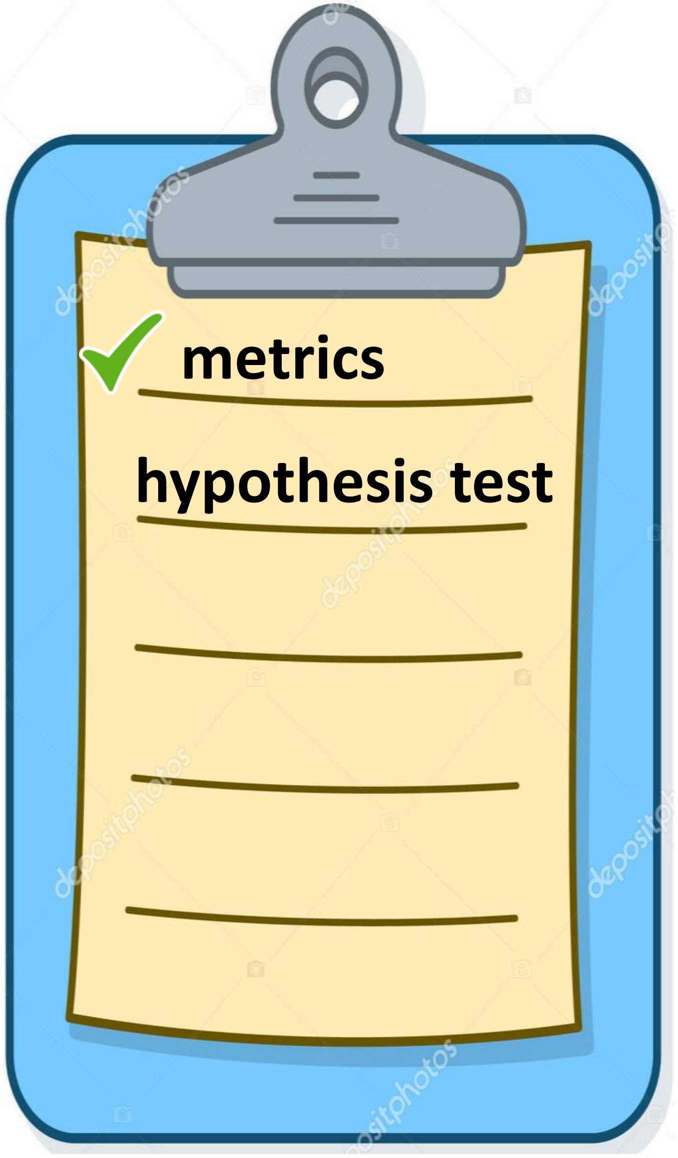
- straight-forward, easy to implement
- 100 years ago: Sir Ronald Fisher – tested different fertilizers on crop yields
- Google ran the first test in 2000. In 2011: they ran > 7000 A/B tests

A/B testing – Choosing relevant metrics



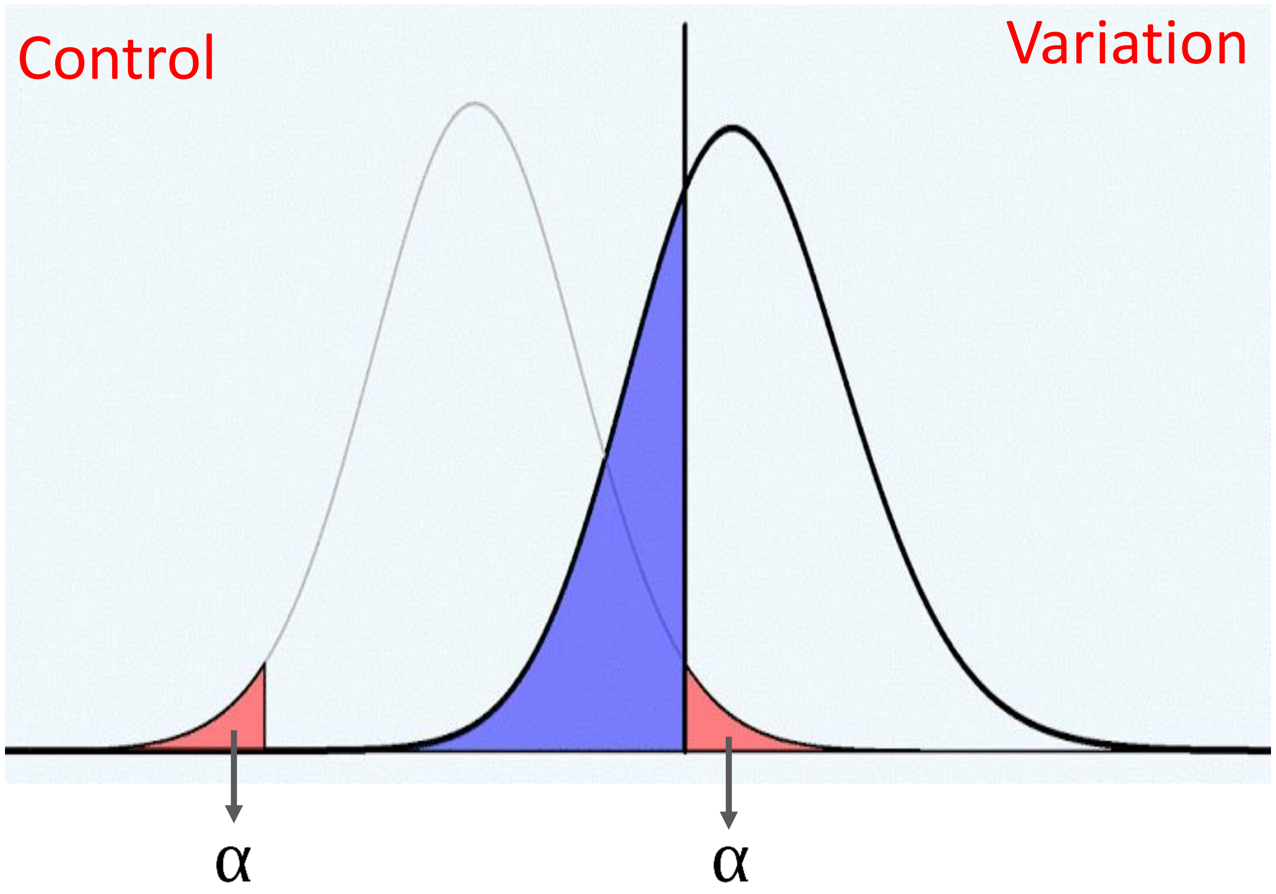
- E.g, CPC, CPA, CTR
- Choosing the right metrics is critical
- Avoiding confounding variables

A/B testing – Is test result statistically significant?

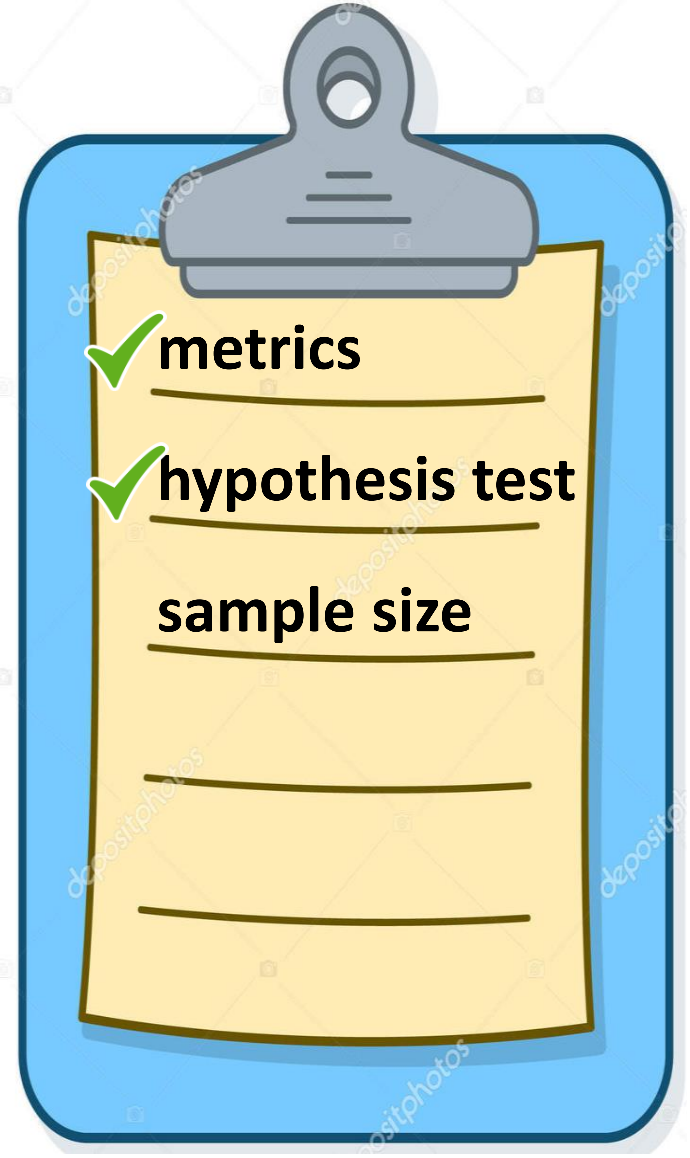


Null hypothesis (H_o): A = B
Alternative hypothesis (H_a): A # B

α: significant level
p-value < α: reject H_o



A/B testing – How long to run the test?



effect size = $\left|\frac{\mu_a - \mu_o}{\sigma}\right|$

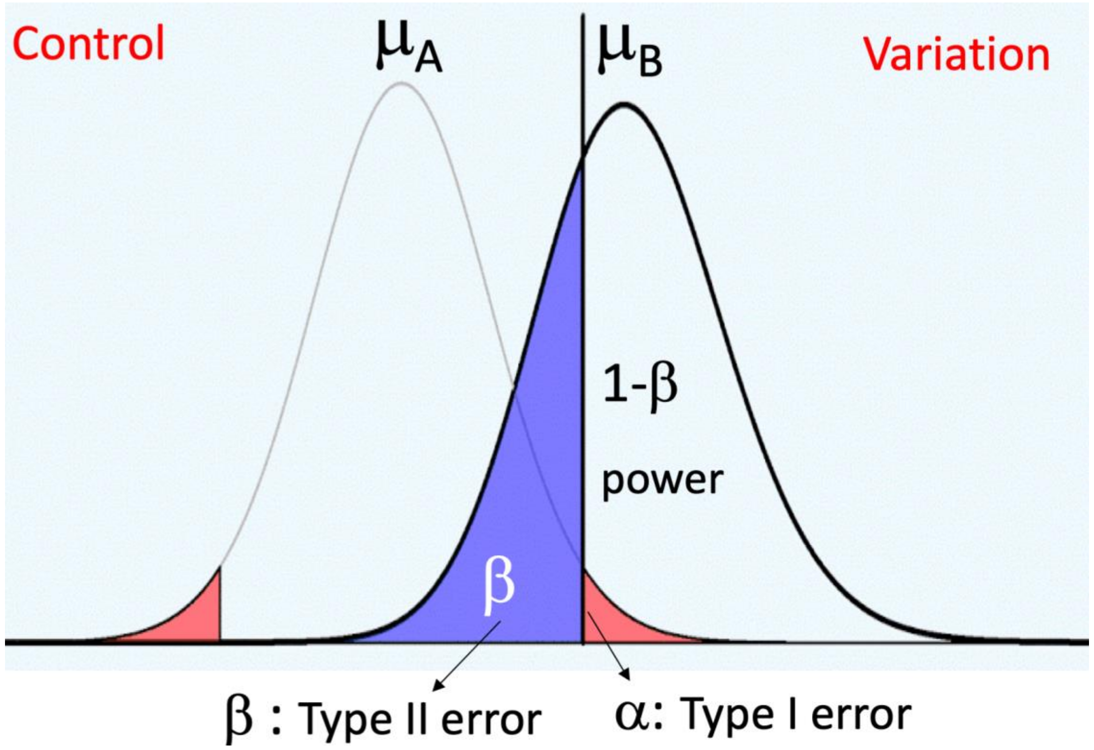
significant level = α

power = 1- β

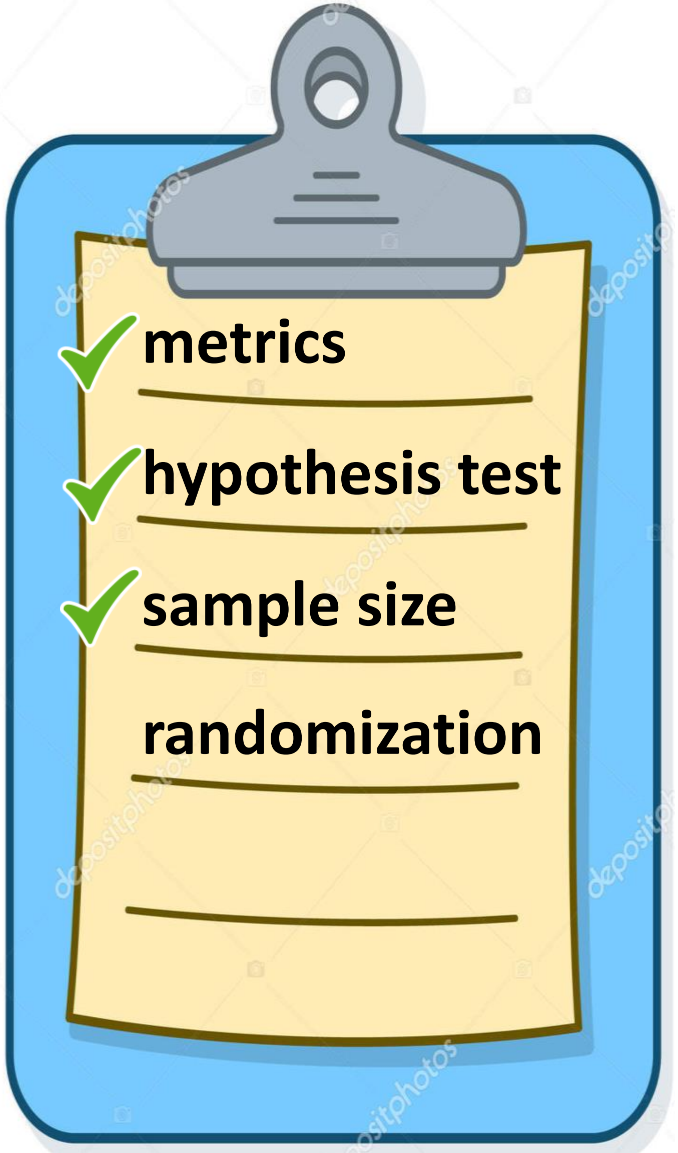
Power analysis → Sample size = n

↓

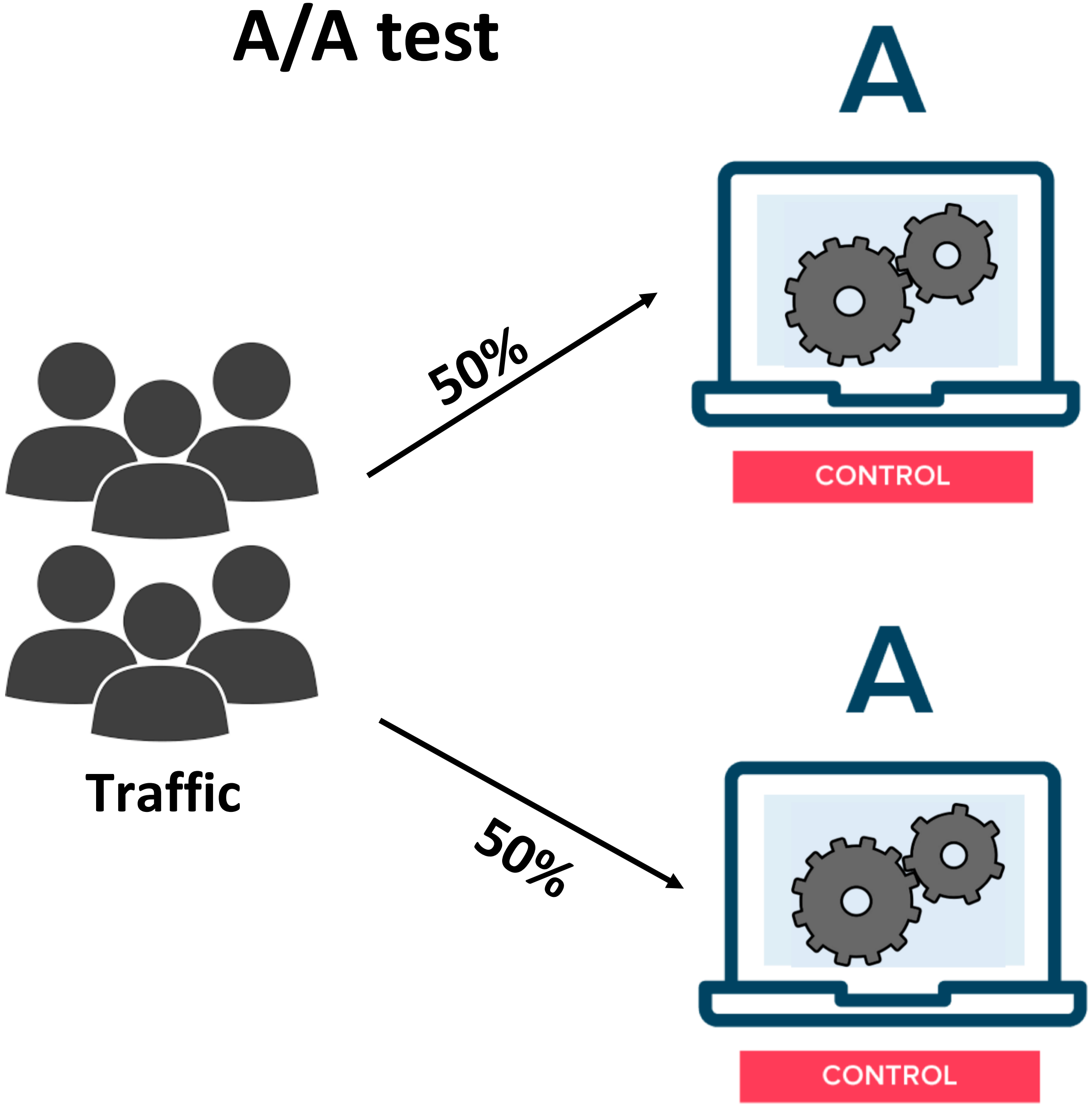
Duration of test



A/B testing – Is randomization guaranteed?

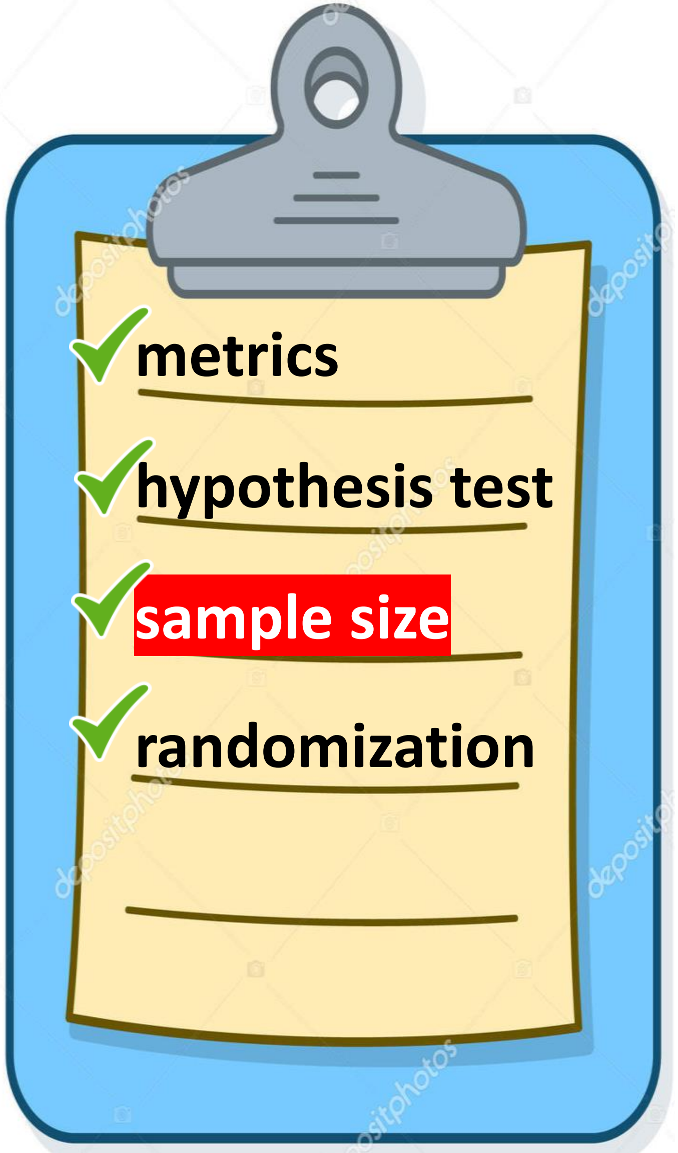


A/A test

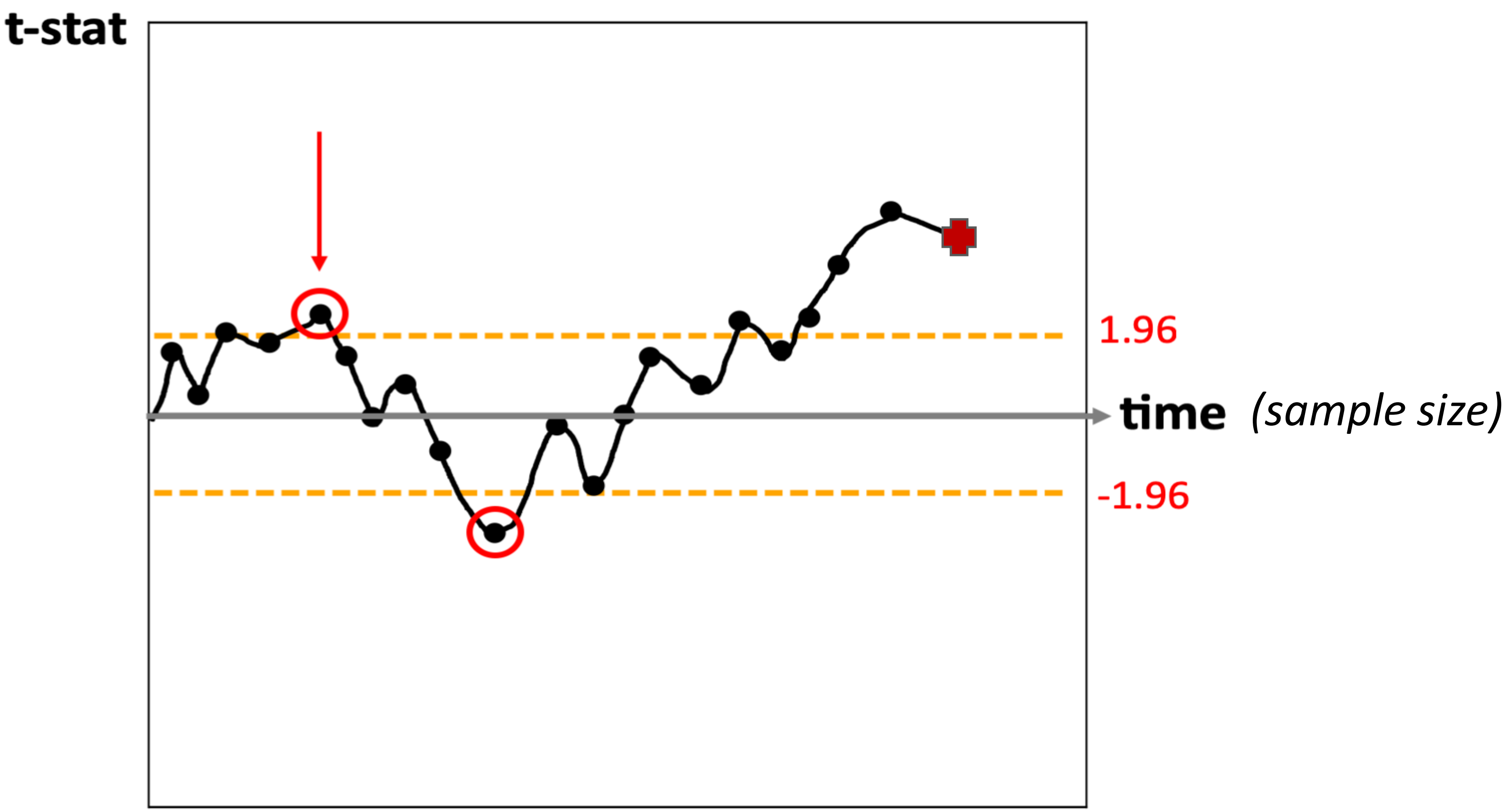


Peeking at A/B Tests and early stopping – Why it matters?

KNOREX



- Impatient experimenters: peeking at the test and early stopping

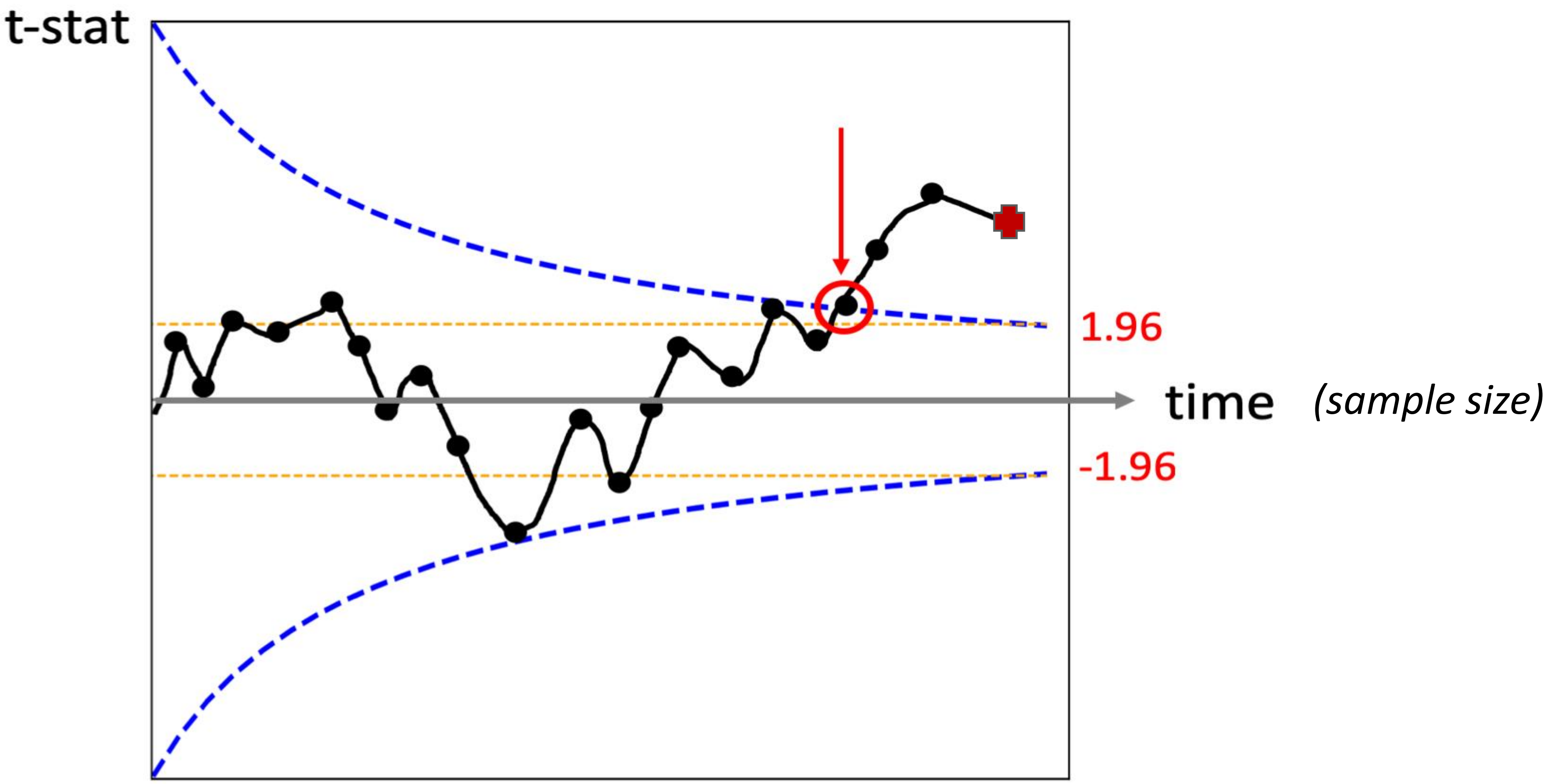


- But time is money!

MOBILEDAY

Peeking at A/B Tests and early stopping – What to do about it?

Sequential A/B testing (frequentist): Sequential probability ratio test



Peeking at A/B Tests and early stopping – What to do about it?

KNOREX

The slightly better model

Scenario: version B (CTR = 0.142%) is slightly better than version A (CTR = 0.14%), p-value = 0.11 > significant level α (0.05)

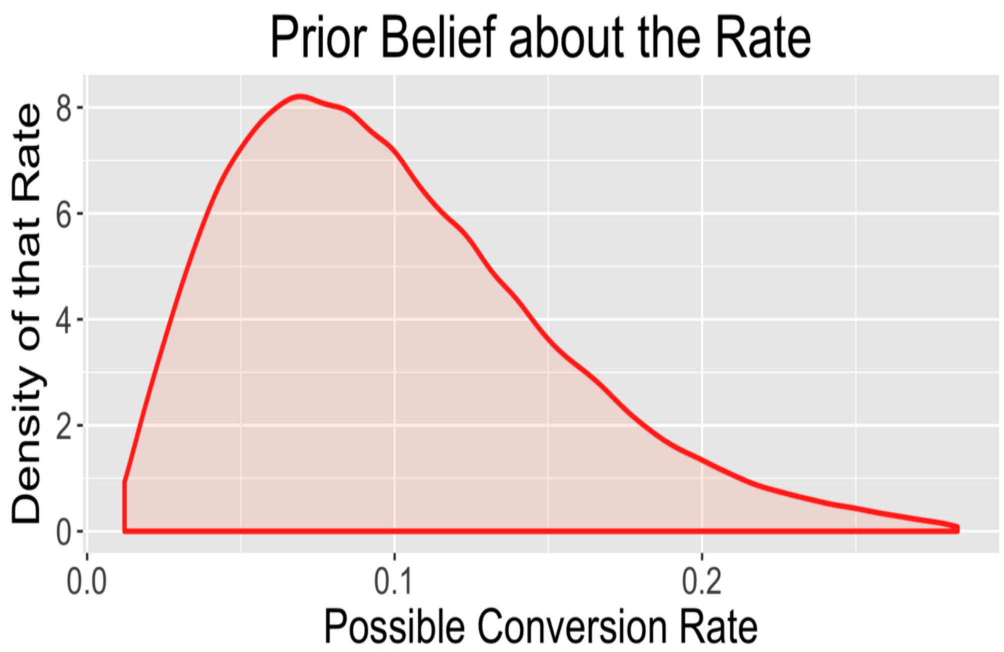
--> Keep A or use B?

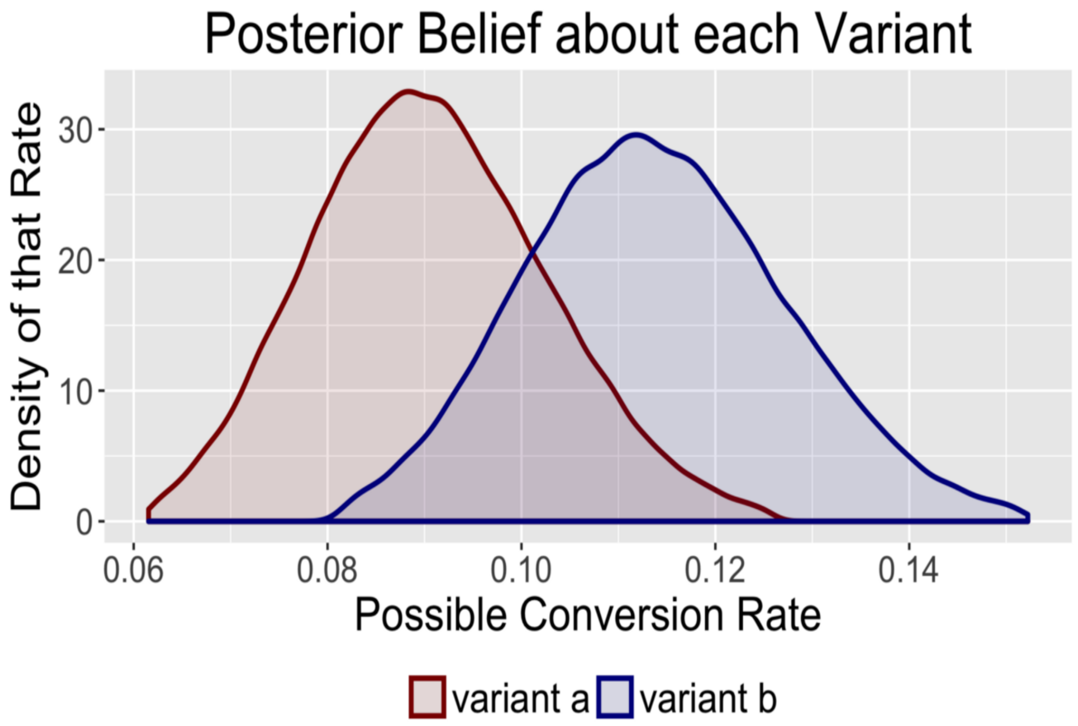
Bayesian A/B testing

Given some prior knowledge, what is the probability that the metric under variant B is larger than the metric under variant A?

Peeking at A/B Tests and early stopping – What to do about it?

Bayesian A/B testing



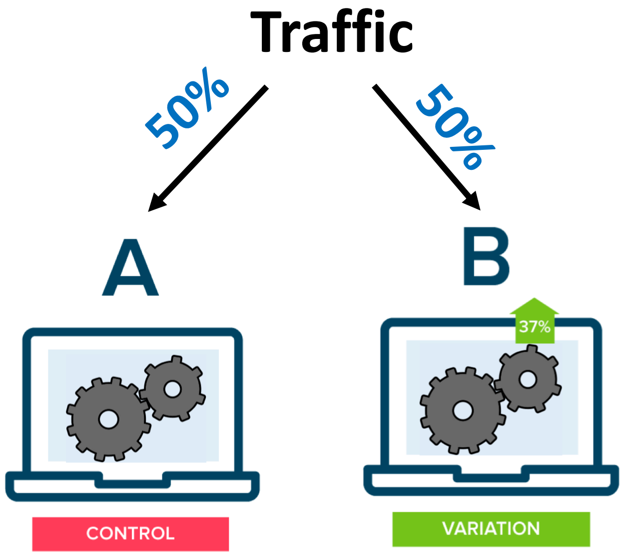


Source: [***Medium***]

- Initiate the experiment and collect data
- Choose a loss threshold **epsilon**
- Periodically, compute expected loss *E[L](A)* or *E[L](B)* if we choose version A or B
- Once either E[L](A) or E[L](B) < **epsilon**: stop the experiment and record the winning variation

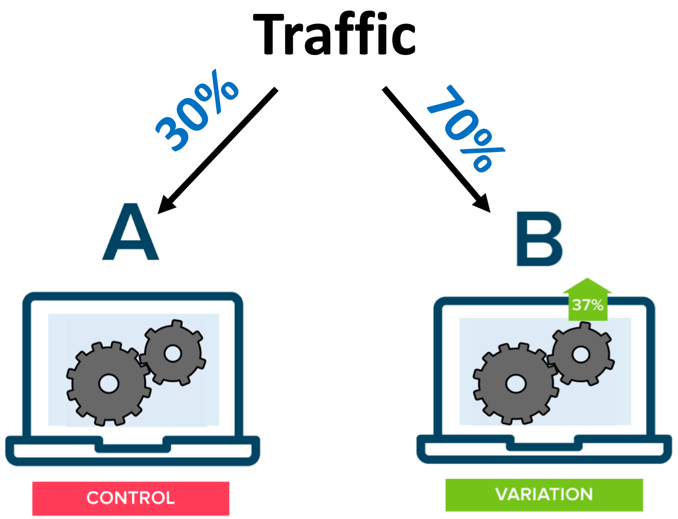
Multi-armed bandit test

t_0

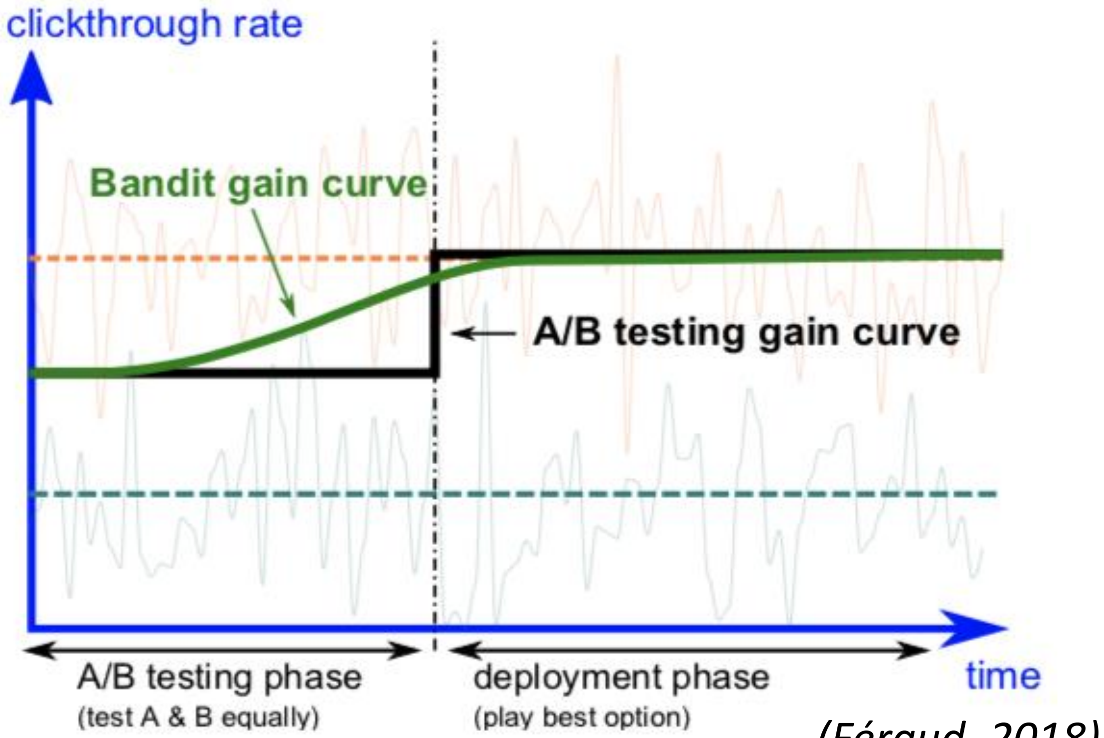


CTR = 0.0015 < CTR = 0.002

t_1



- Earn while you learn



(Féraud, 2018)

- Give the poorer performing versions a second chance
- Automation
- >2 arms
- More tricky to implement

Challenges in online testing in RTB

- Eg., Implement a new click predictor to optimize for cpc → Metric: ctr, cpc
- It might take long time to observe a significant difference in the new model
- Heterogeneity → assumption of ‘iid’ fails
- Testing at ad-group level: data size is even smaller
- Time is money and we don’t want to compromise user’s experience

→ A/B test might not an optimal method

- Multi-armed bandit: the more appropriate approach

Summary

- Online evaluation of new model/new idea is important but challenging, especially in RTB
- A/B testing is the most widely used method, though early stopping without a plan is dangerous → sequential A/B test or Bayesian A/B test
- A/B testing: strict experiments where focus is on statistical significance
- Multi-armed bandit (MAB): continuous optimization where focus is on maintaining optimal KPIs

References

Bayesian A/B test:

https://medium.com/convoy-tech/the-power-of-bayesian-a-b-testing-f859d2219d5

https://cdn2.hubspot.net/hubfs/310840/VWO_SmartStats_technical_whitepaper.pdf

Peeking at A/B test:

https://www.youtube.com/watch?v=AJX4W3MwKzU

https://www.evanmiller.org/how-not-to-run-an-ab-test.html

Sequential A/B test:

https://www.evanmiller.org/sequential-ab-testing.html

https://www.aarondefazio.com/tangentially/?p=83

https://en.wikipedia.org/wiki/Sequential_probability_ratio_test

Backup slides

Peeking at A/B Tests and early stopping – What to do about it?

KNOREX

Sequential A/B testing (frequentist): Sequential probability ratio test

H_o: θ_o
H_a: θ_a

$S_o = 0 \rightarrow S_i = S_{i-1} + \log \Lambda_i$ for i in 1,2,..,t

$\Lambda_I = \frac{L(\theta_o|x)}{L(\theta_a|x)}$: Likelihood ratio

Stopping rule:
$a < S_i < b$
$S_i \geq b$: Accept H_a
$S_i \leq a$: Accept H_o

$a \sim \log\left(\frac{\beta}{1-\alpha}\right)$ $b \sim \log\left(\frac{1-\beta}{\alpha}\right)$

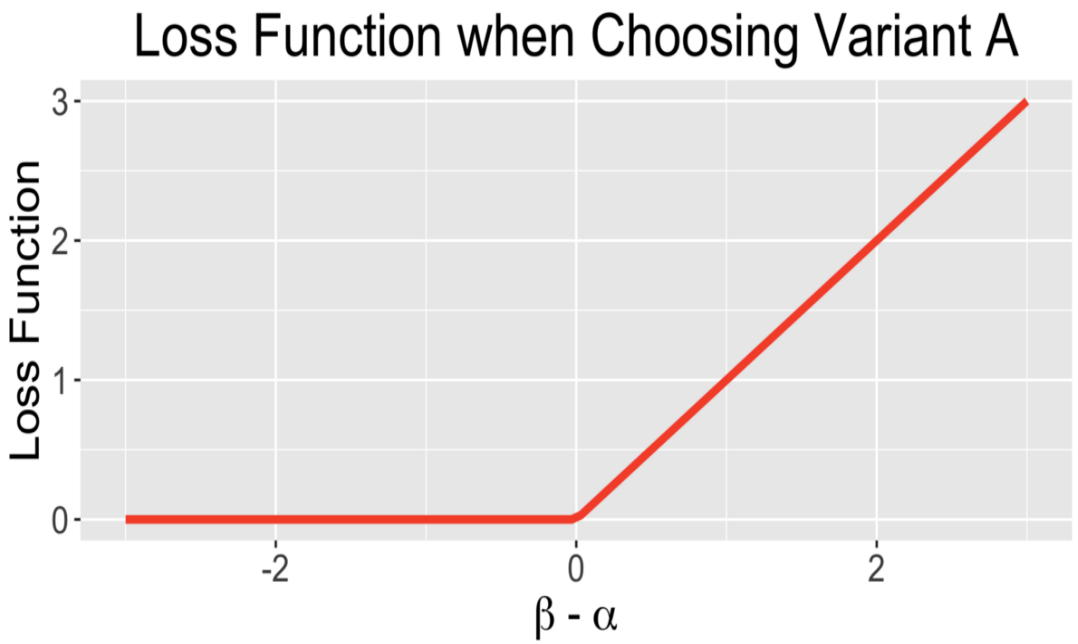
The slightly better model

Scenario: version B (CTR = 0.142%) is slightly better than version A (CTR = 0.14%), p-value = 0.11 > significant level α (0.05)

--> Keep A or use B?

Peeking at A/B Tests and early stopping – What to do about it?

Bayesian A/B testing



$$L(\alpha, \beta, x) = \begin{cases} max(\beta - \alpha, 0) & x = a \\ max(\alpha - \beta, 0) & x = b \end{cases}$$

$$E[L](x) = \int_A \int_B L(\alpha, \beta, x) \underbrace{f(\alpha, \beta)}_{\text{joint posterior density}} \, d\alpha \, d\beta$$

- Initiate the experiment and collect data
- Periodically, compute E[L](A) and E[L](B)
- If either E[L](A) or E[L](B) < **epsilon**: stop the experiment and record the winning variation
- Else: continue the experiment

Source: [Medium]