Tencent 腾讯

AB Testing in (Tech) Industry

Shichao Han 韩士超

Weixin Experimentation Platform

Feb 29, 2024

Darwin Wang 王勇

Founder, manager @ Weixin experimentation platform

Engineering, data science

darwinwang@tencent.com

Shichao Han 韩士超

Data scientist @ Weixin experimentation platform

Statistics and computer science

shichaohan@tencent.com

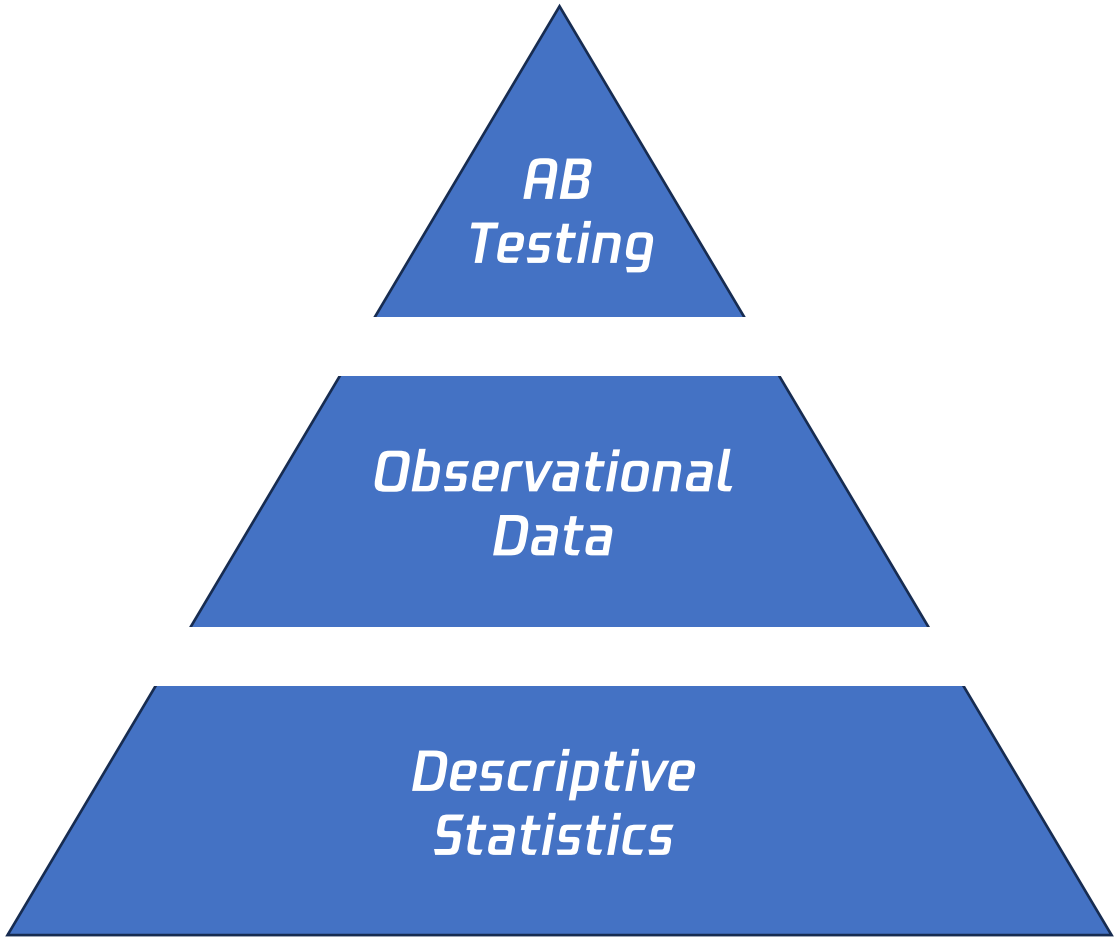
Shichao

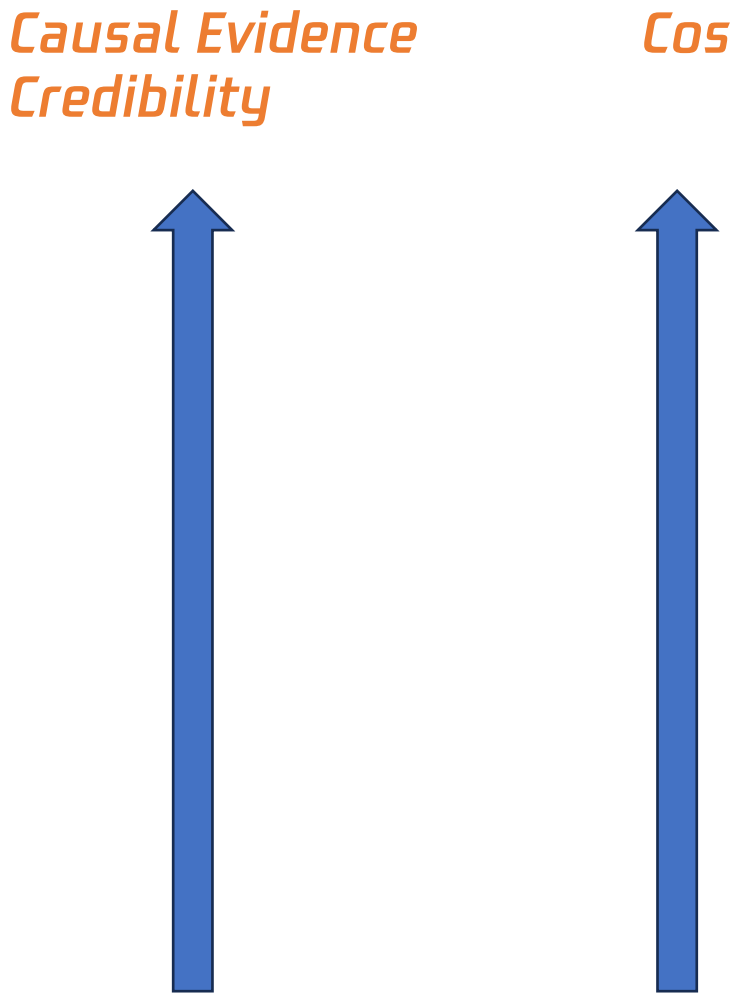
1. Why AB Testing
2. Challenges, solutions, and practices - engineering
3. Challenges, solutions, and practices - design and analysis

1 Why AB Test

1. Why AB Test

(0) Causal inference



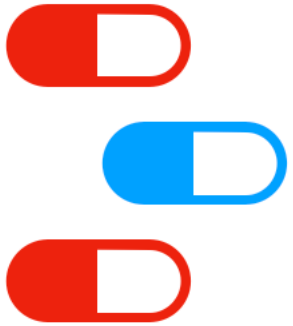


1. Why AB Test

(1) Advantages of online experimentation platforms

We rely on machines for:

- **Randomization:** treatment and control group are highly comparable
- **Implementation:** consistency of treatment consistency of treatment; no defiers
- **Data collection:** we can capture as many pre-treatment covariates as possible; measurement



Biomedical

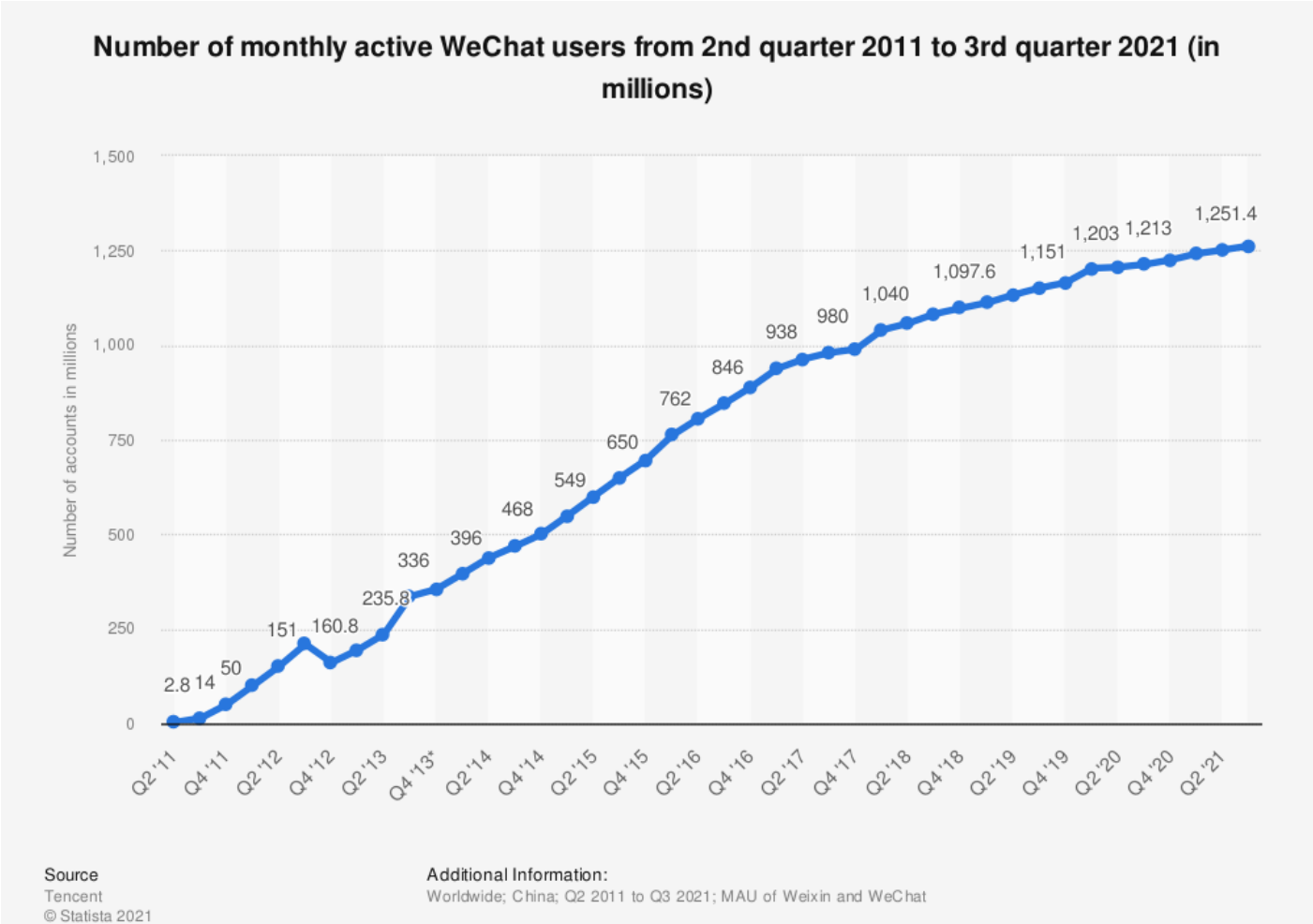
Weixin

Survey, poll

1. Why AB Test

(2) "SMALL" increment, "HUGE" return

- Weixin: since 2011
- A/B Testing at Weixin starts later
- Growth of MAU slows down
- Very large population
- Before: fewer users => acquisition
- Now: more users => increment
- The outcomes of interest usually have small effect size (Cohen's d <= 0.5)



Graph from: https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts/

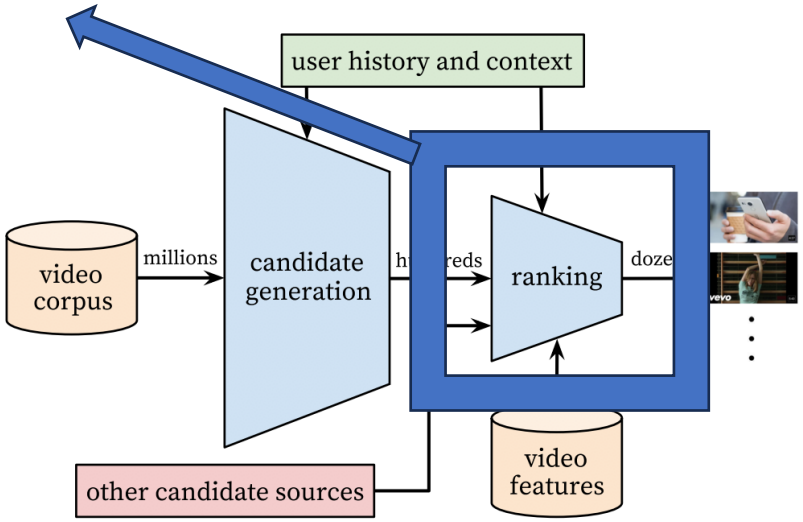
1. Why AB Test Engineers

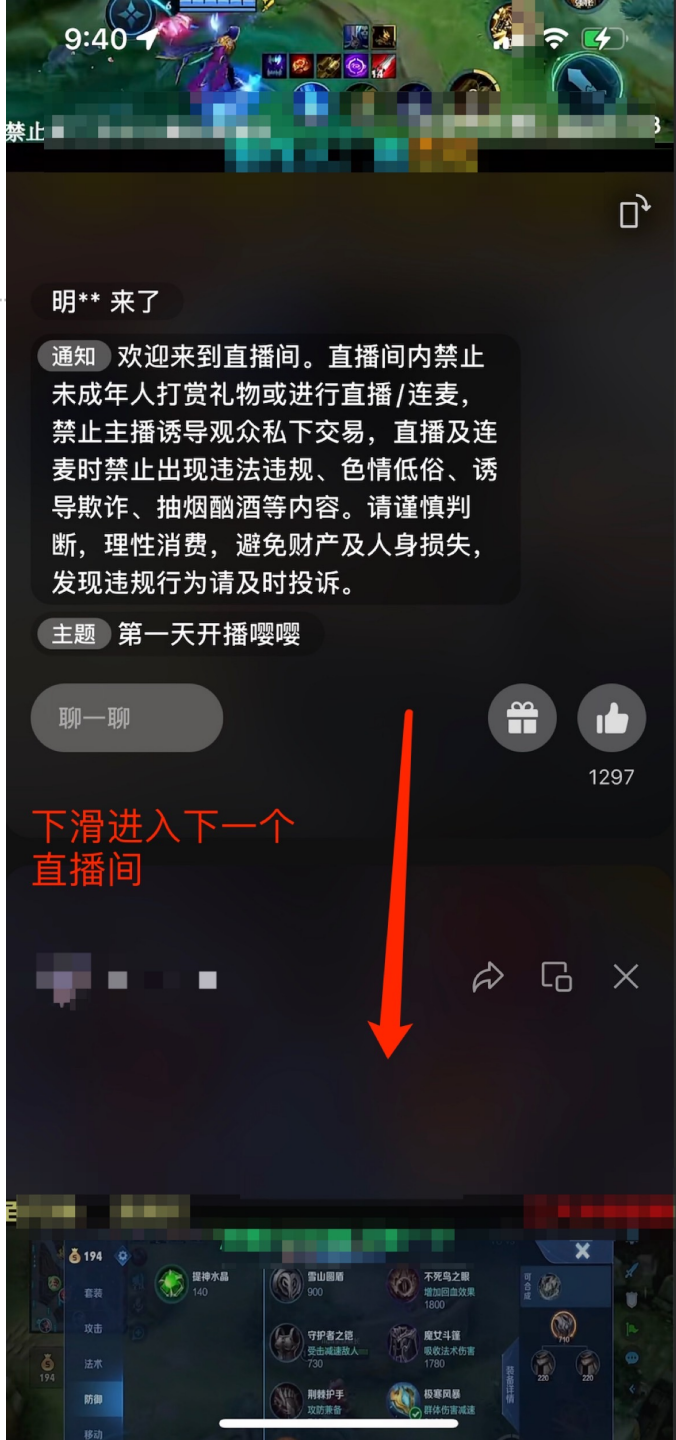
Step I: Develop and optimize algorithm (classification, prediction ...)

Step II: Historical data -> L2 loss, NDCG loss, Accuracy, Precision, Recall ...

Step III: Online experimentation -> Business decision metrics

Step IV: Report to the manager





1. Why AB Test Engineers

Step I: Develop and optimize algorithm (classification, prediction …)

Step II: Historical data -> L2 loss, NDCG loss, Accuracy, Precision, Recall …

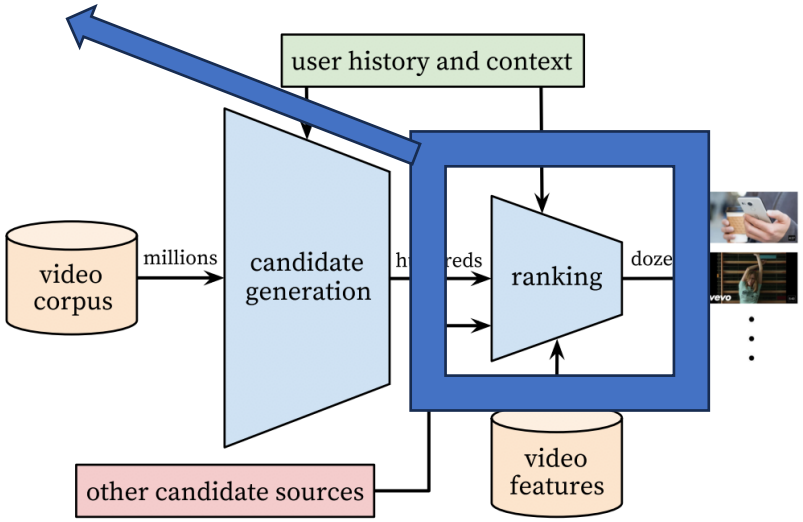
Step III: Online experimentation -> Business decision metrics

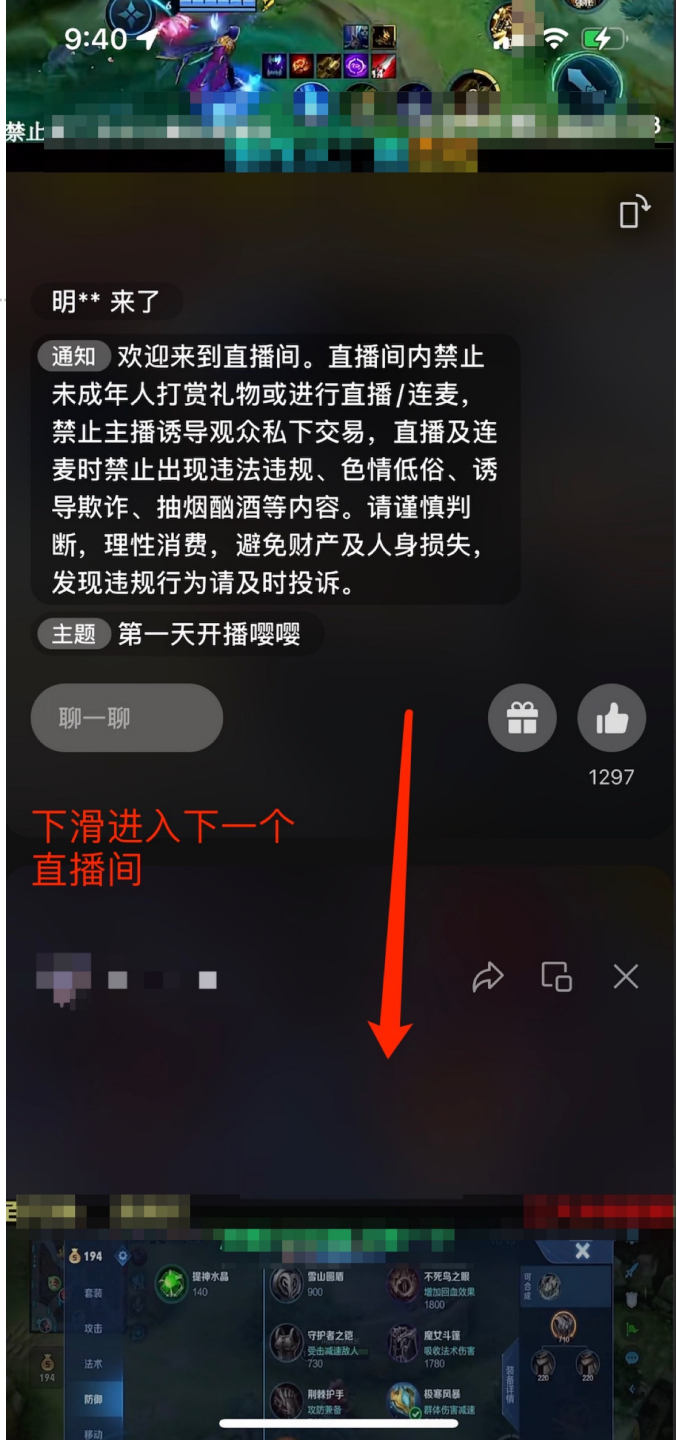
Step IV: Report to the manager

Experimentation Platform

1. Infrastructure – experimentation
2. Techniques – inference

Fast, robust, reliable results





1. Why AB Test

Product managers, data scientists

- Randomized controlled experiments are gold standard for causal inference
- Decisions: whether to launch decisions based on treatments' impact on metrics

1. Why AB Test

Product manager, data scientists

- **Example:** new feature in Weixin’ s PCR test location search tab to help users better find the test centers.
- Hypothesized the change will increase CTR; users don’ t need to search twice within a short time
- Experiment results: (1) # of search queries goes down, but the proportion of users who only searched once within short time goes up => their search need is better satisfied (2) CTR go up
- Decision: Launch the change
- Provide valid causal inference

AB Test Result (Difference in means)

+

Mediation analysis

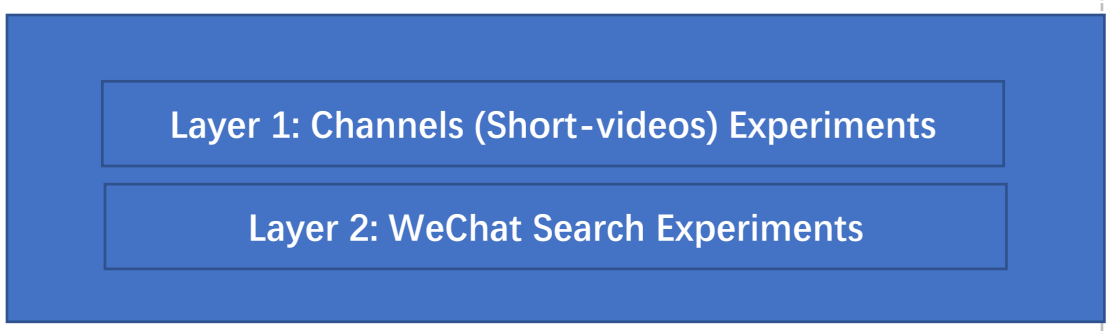
2 Engineering

2. Challenges and solutions - engineering

(1) Efficiency – run experiments in parallel

We can run "orthogonal" experiments at low cost, because we can easily manipulate randomization scheme => each user can be in multiple experiments

- Each user can be in multiple experiments
- The design of "Layers" : each user can only be in 1 experiment within each layer
- If two experiments interact with each other => make sure they are in the same layer



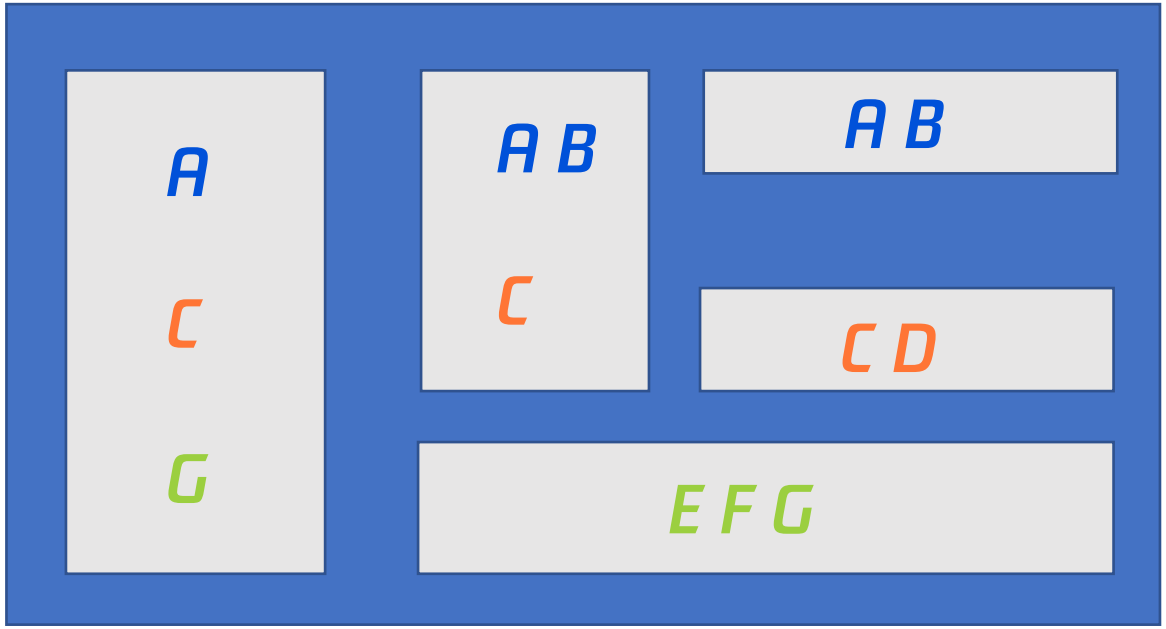
2. Challenges and solutions - engineering

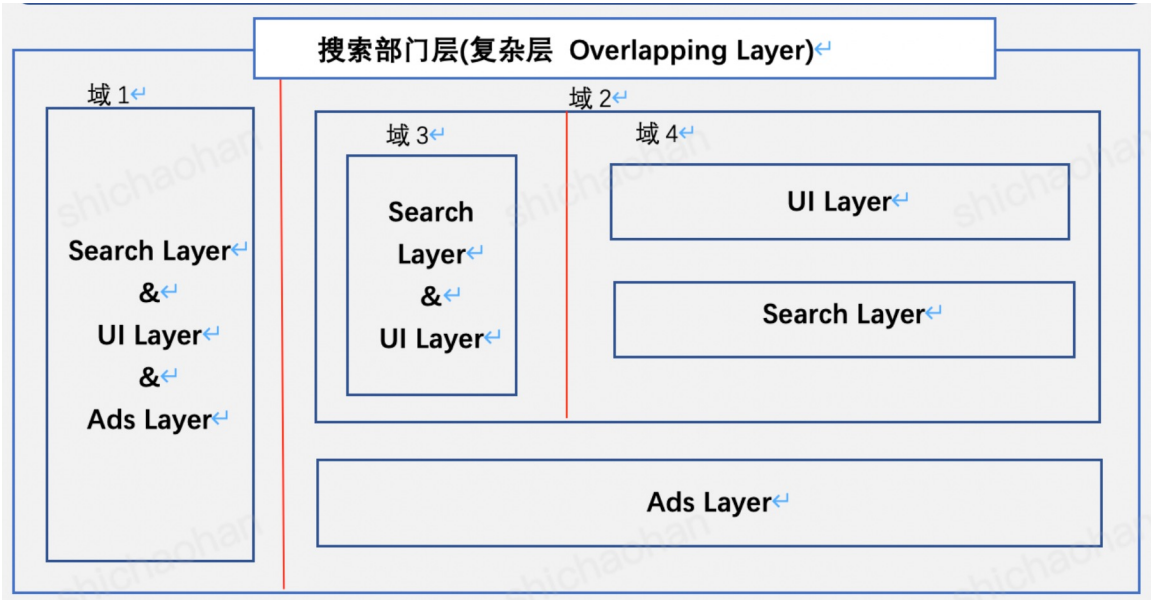
(1) Efficiency – run experiments in parallel

What if my treatments involve changes across different layers?

Create domains – each domain contains multiple layers

Each parameter can be only belong to one layer





Example of Parameter List for Layers in Domains

2. Challenges and solutions - engineering

(2) Computational cost

$O(n)$ $v.s.$ $O(10000n)$

- 7,000+ experiments running on our platform
  - Number of experimental units per experiment: >> millions
  - Number of outcomes of interests: dozens to hundreds
  - Causal estimators + variance estimates run on **every experiment** for **each treatment group** and **every metric**

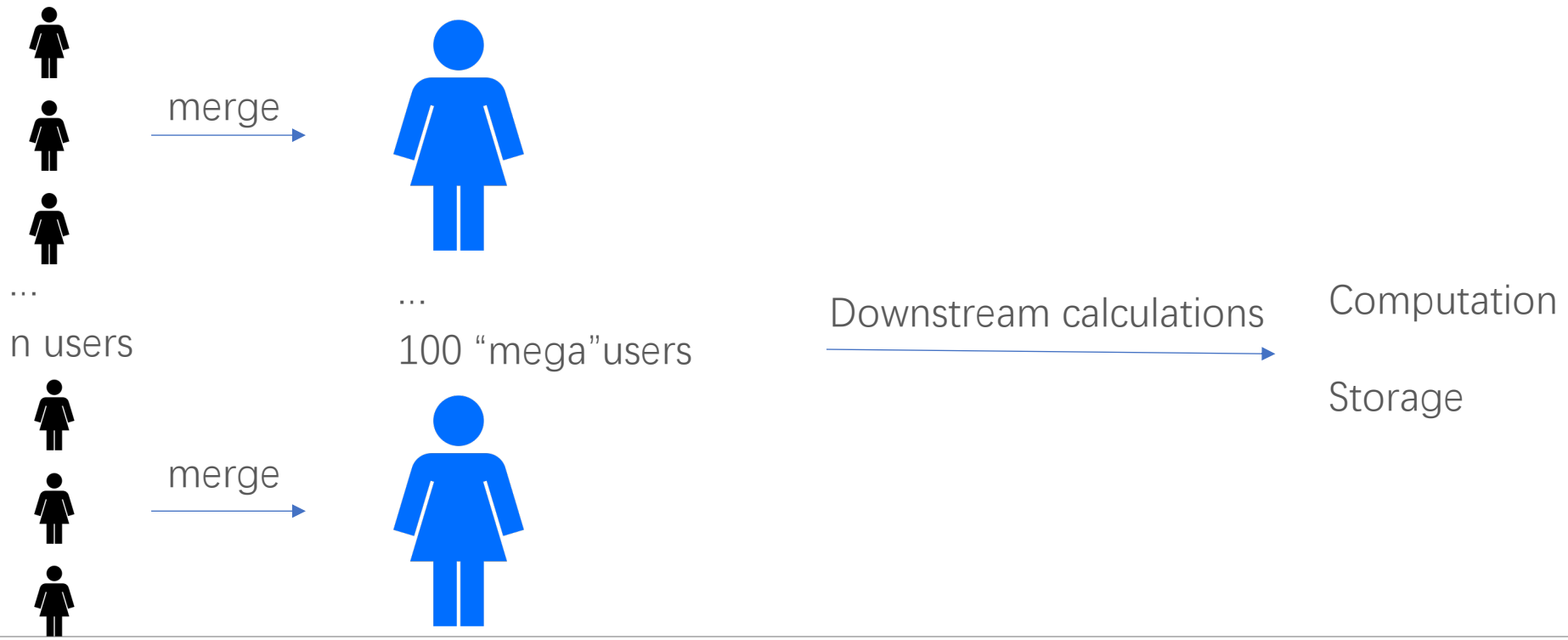
组别	命中人数 (去重累计)	[redacted]	[redacted]	[redacted]	[redacted]
A1,A2	3,016,093	0.01704	0.03508	0.13588	0.08958
B1	3,018,245	-0.697% ±1.874% 0.01692	-0.053% ±1.340% 0.03506	+0.254% ±0.649% 0.13622	+0.134% ±0.679% 0.08970
B2	3,020,680	-2.583% ±1.891% 0.01660	-1.870% ±1.361% 0.03442	-1.630% ±0.662% 0.13366	-1.738% ±0.716% 0.08802
B3	3,021,662	-1.307% ±1.877% 0.01682	-0.470% ±1.323% 0.03491	-0.075% ±0.626% 0.13578	-0.112% ±0.661% 0.08948
B4	3,021,184	-2.362% ±1.871% 0.01664	-1.665% ±1.245% 0.03449	-1.991% ±0.663% 0.13317	-2.197% ±0.696% 0.08761

2. Challenges and solutions - engineering

(2) Computational cost

We put our **n** users randomly into **100** bins

- Rather than **n** rows of data (usually n > millions)
- We only need **100 rows of data for each treatment group**



2. Challenges and solutions - engineering

(2) Computational cost

$O(n)$ $v.s.$ $O(10000n)$

<uin, metric> table has 1,000,000,000 rows

- Data query takes a long time (5 mins)
- (1) Calculate **difference in means estimator + closed form variance estimator** = 1 mins
- (2) Doubly **Robust Estimator + Bootstrap variance estimator** = 2mins x 10000

Doubly Robust Estimator: $\tilde{\mu}_1 - \tilde{\mu}_0$
Variance Estimator: Bootstrap

$$\tilde{\mu}_1^{\mathrm{dr}} = E\left[\frac{Z\{Y-\mu_1(X,\beta_1)\}}{e(X,\alpha)}+\mu_1(X,\beta_1)\right],$$
$$\tilde{\mu}_0^{\mathrm{dr}} = E\left[\frac{(1-Z)\{Y-\mu_0(X,\beta_0)\}}{1-e(X,\alpha)}+\mu_0(X,\beta_0)\right]..$$

2. Challenges and solutions - engineering

(2) Computational cost

$O(n)$ $v.s.$ $O(10000n)$

<uin, metric> table has 1,000,000,000 rows

- Data query takes a long time (5 mins)
- (1) Calculate **difference in means estimator + closed form variance estimator** = 1 mins
- (2) Doubly **Robust Estimator + Bootstrap variance estimator** = 2mins x 10000

2024 Update:
Weixin's **fast causal inference** can compute causal estimators and conduct inference for billions of data **within seconds**

Doubly Robust Estimator: $\tilde{\mu}_1 - \tilde{\mu}_0$
Variance Estimator: Bootstrap

$$\tilde{\mu}_1^{\mathrm{dr}} = E\left[\frac{Z\{Y-\mu_1(X,\beta_1)\}}{e(X,\alpha)}+\mu_1(X,\beta_1)\right],$$

$$\tilde{\mu}_0^{\mathrm{dr}} = E\left[\frac{(1-Z)\{Y-\mu_0(X,\beta_0)\}}{1-e(X,\alpha)}+\mu_0(X,\beta_0)\right]..$$

2. Challenges and solutions - engineering

(2) Computational cost

Algorithm	Compare	Performance (600 million rows , 20 columns , 300 cores)		
		Our Model	Single Machine Model	Distributed Model
Ttest using deltamethod	No package in the industry	0.32s	...	...
OLS	same	3.7s	OOM	1.85min
lasso	same	6.32s	OOM	2.02min
matching	same	1.03s	OOM	3.53min
bootstrap(mean)	same	16.93s	OOM	21.85min
causal tree	same	2.71min	OOM	...
DML(LinearDML/nonParamDML)	same	4.23s	OOM	...

2. Challenges and solutions - engineering

(2) Computational cost

Packages/Tool : Clickhouse/spark/starrocks [distributed computation] + sql/python [Interacting]

Basic operator

- Estimators
- Matrix multiplication
- Gradient descent
- Kfold
- Sampling with/ without replacement
- Tree
- Data cleaning tool
  - Cut bins
  - One hot encoding
  - Polynomial Features

High operator

Model

- OLS
  - WLS
  - ClusteringOLS
  - Panel OLS
- Covariance matrix
- Lasso/Elastic/logistic
- Tree-based model

Simulation

- Bootstrap
- Permutation

Application

- Estimators and Variance
  - Difference in means
  - Deltamethod
  - Post-stratification
  - CUPAC
  - other estimators
- Hypothetical test
  - T test
  - KStest
  - Quantile test
  - Outlier-detection
- Observational studies
  - Meta-learner
  - DML
  - Uplift tree
  - Matching
  - RDD
  - Synthetic control
  - DID
  - IV
  - Mediation Analysis
  - Panel (data) analysis : Fixed effect/ pooled/ feiv/event study

3 Challenges and solutions

3. Design & Inference

(1) External validity problem

- **In-experiment** data reveals a significant positive effect
- **After deployment** the key metrics decrease

3. Design & Inference

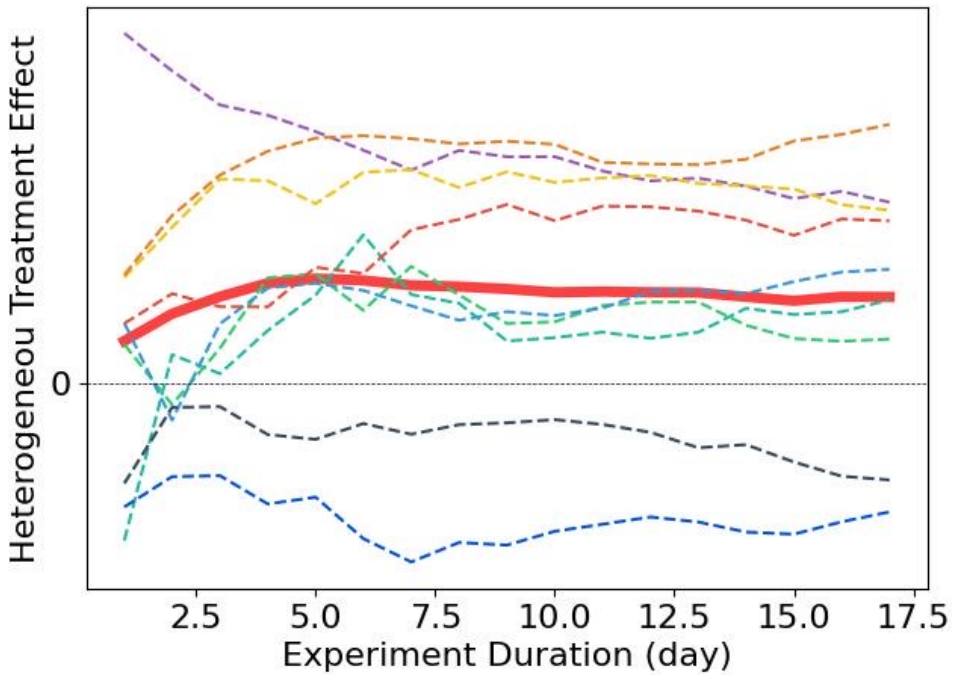
(1) External validity problem

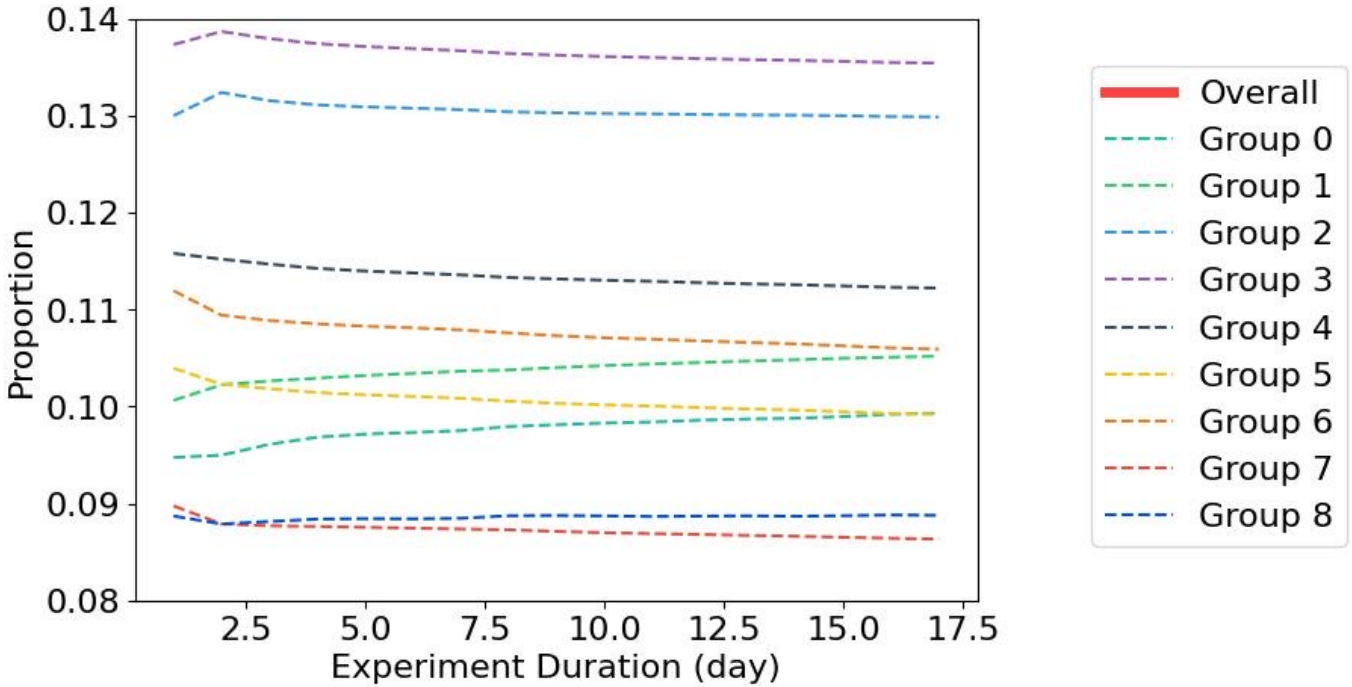
- **In-experiment** data reveals a significant positive effect
- **After deployment** the key metrics decrease

3. Design & Inference

(1) External validity problem

- **In-experiment** data reveals a significant positive effect
- **After deployment** the key metrics decrease

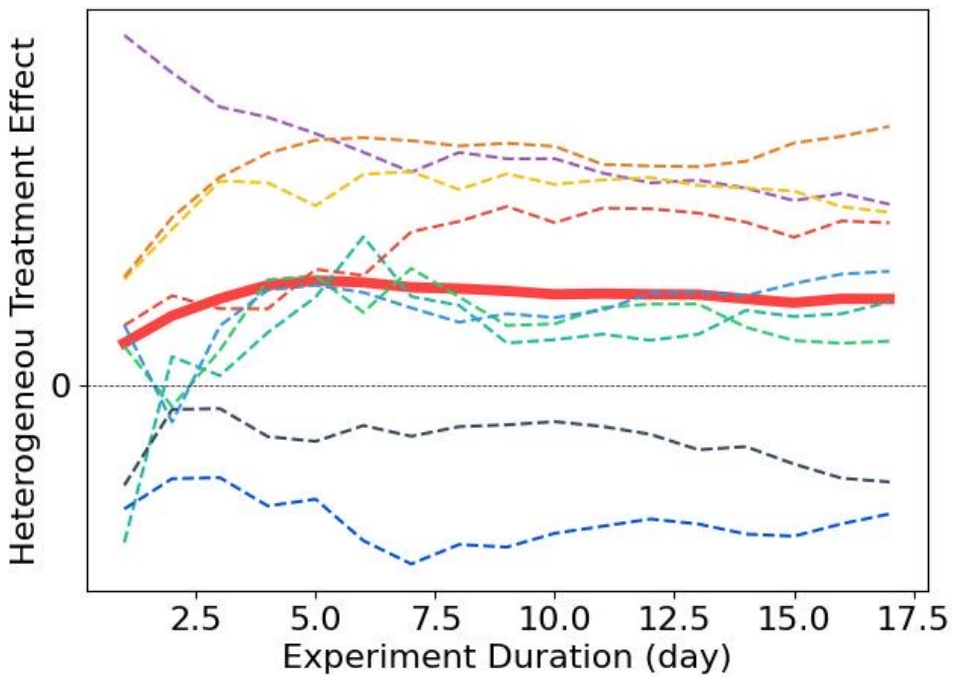


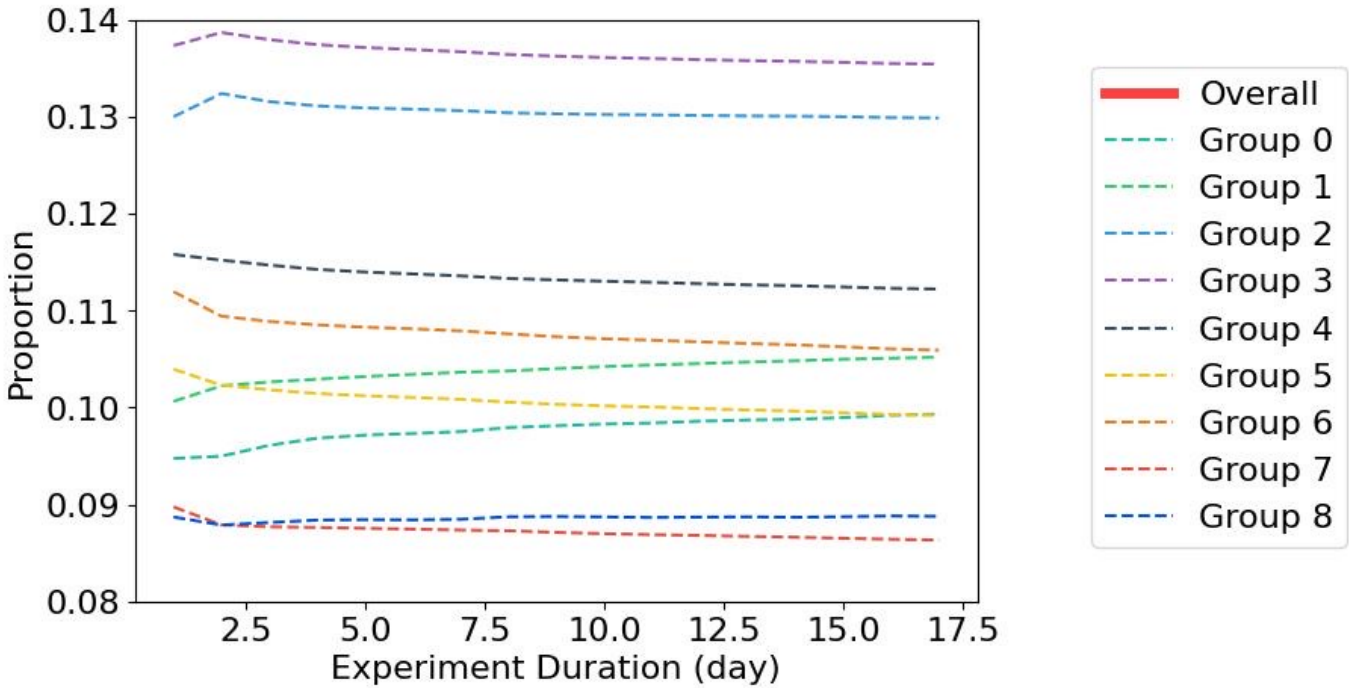


3. Design & Inference

(1) External validity problem

- **Stop wisely:** until when the in-experiment sample can be representative of population
- **Estimator:** adjust for participation probabilities

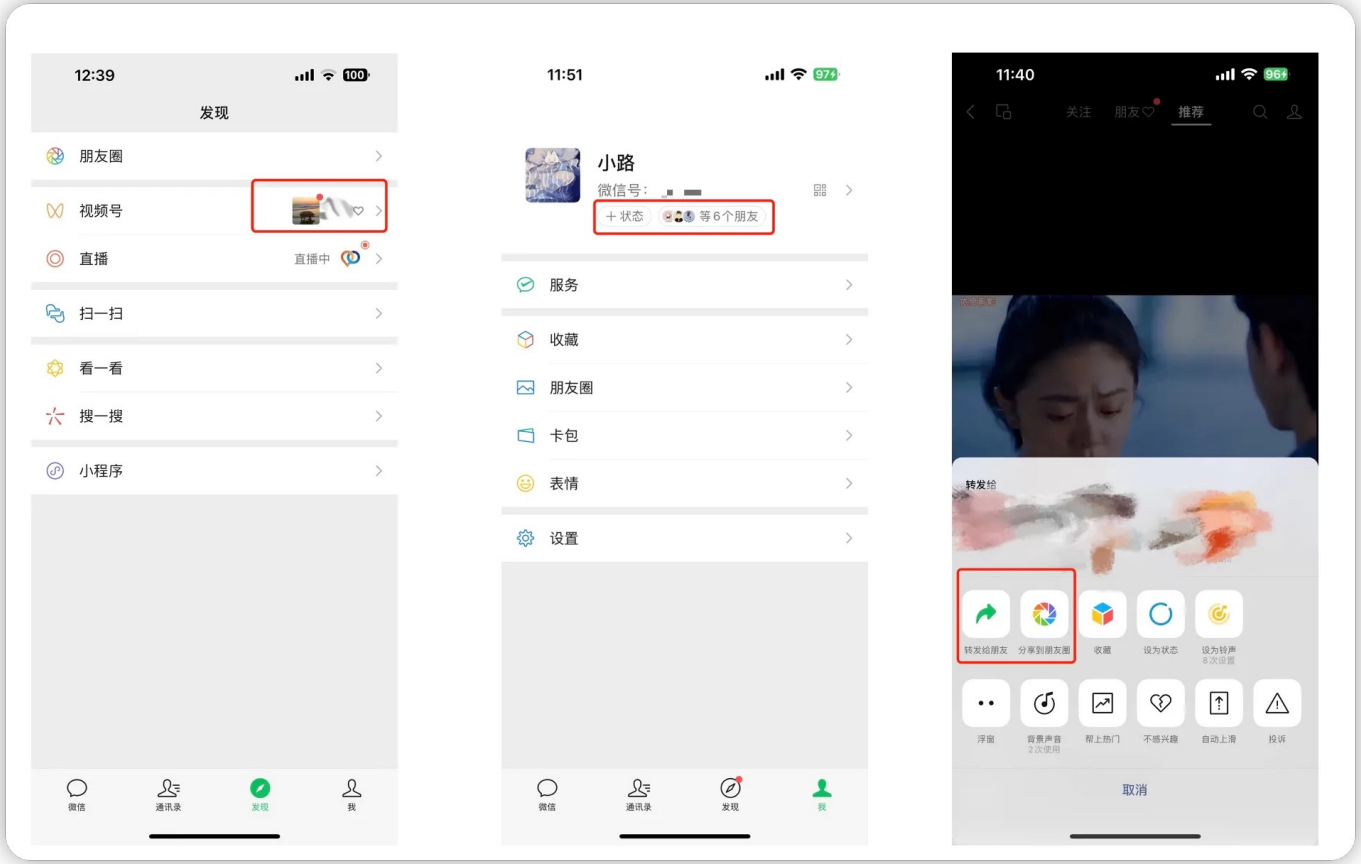




3. Design & Inference

(2) Cluster Randomization to Account for Network Interference

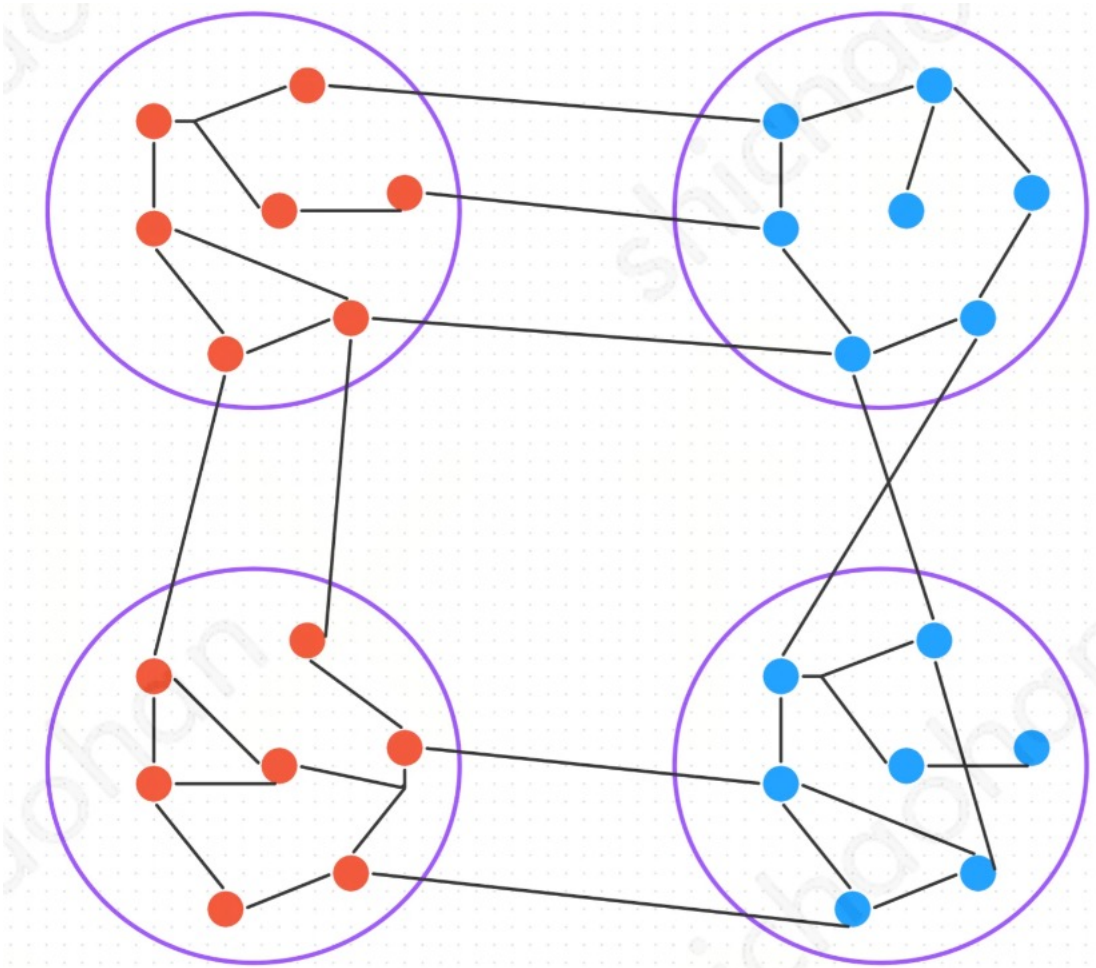
- **Stable Unit Treatment Value Assumption(SUTVA)** is violated
- Example:
  - In WeChat, you can see your friends' sharing; your friends can also share videos to DM
  - Your potential outcomes are affected by your friends'potential outcome
  - Imagine a treatment that will increase the average watching duration of videos-> increase the probability of sharing
  - 0% of your friends get treated v.s. 50% of your friends get treated v.s. 100% of your friends get treated



3. Design & Inference

(2) Cluster Randomization to Account for Network Interference

- **Cluster Randomization**
- **Step 1:** Build a graph
- **Step 2:** Clustering (some ML clustering algorithm)
- **Step 3:** Randomization of clusters
- **Step 4:** Each cluster is a “mega” user -> Inference



3. Design & Inference

(2) Cluster Randomization to Account for Network Interference

- Example
  - Thumbs up: likes, but your friends will not see
  - Heart: likes, and your friends will see you liked it
  - **Treatment**: double tap screen = heart
  - **Control**: double tap screen = thumbs up
  - **Outcomes of interest**: stay duration, number of total likes, clicks, forwards, comments …

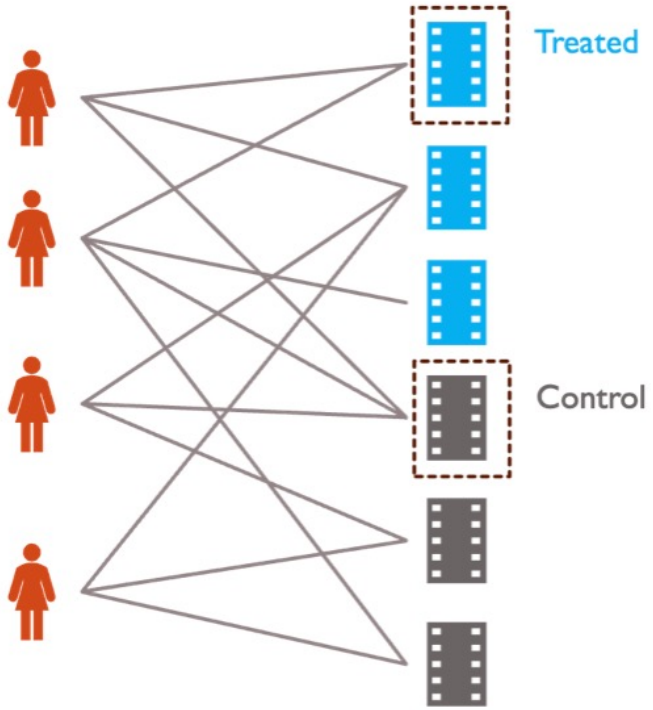
3. Design & Inference

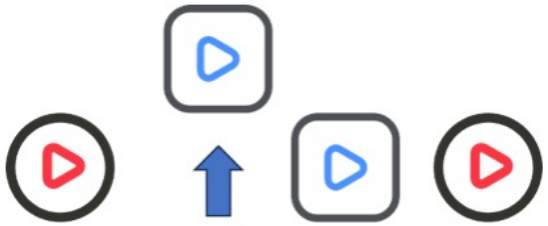
(3) Marketplace Interference

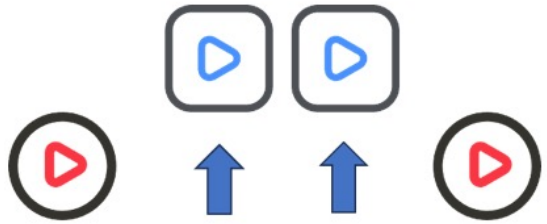
- Content creators and consumers in out platforms
- However, creators **compete** in the platform => **Interference and spillover** => ATE estimator is **biased**
- Example:
  - A treatment that change the exposure probability of new creators in an experiment of 20% of the whole creator population; 10% control and 10% treat
  - New creators in the control group did not receive the treatment => in the experiment, new creators in different groups have different exposure probability
  - If the treatment is applied to everyone in the platform => 100% of the new creators have the same exposure probability
  - **Outcomes of interest**: DAU, number of videos created, CTR …

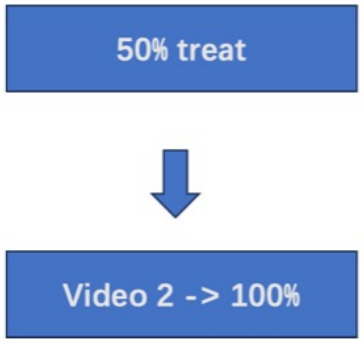
3. Design & Inference

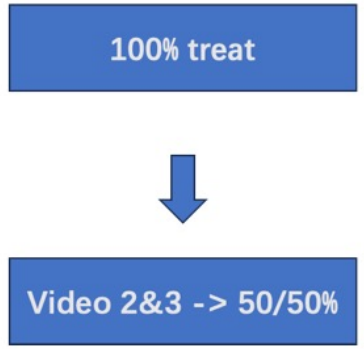
(3) Marketplace Interference











Video-side randomization

- Each user sees both treated and control videos
- Competitions among videos -> Each video's (potential) outcomes depend on other videos' treatment status

3. Design & Inference

(3) Marketplace Interference

Experimental design stage solution:

- Assign consumers and creators to T/C group
- Treated consumers can only see treated creators
- Control consumers can only see control creators

	Control Creators	Treated Creators
Control Consumers	Control Group	
Treated Consumers		Treated Group

3. Design & Inference

(3) Marketplace Interference

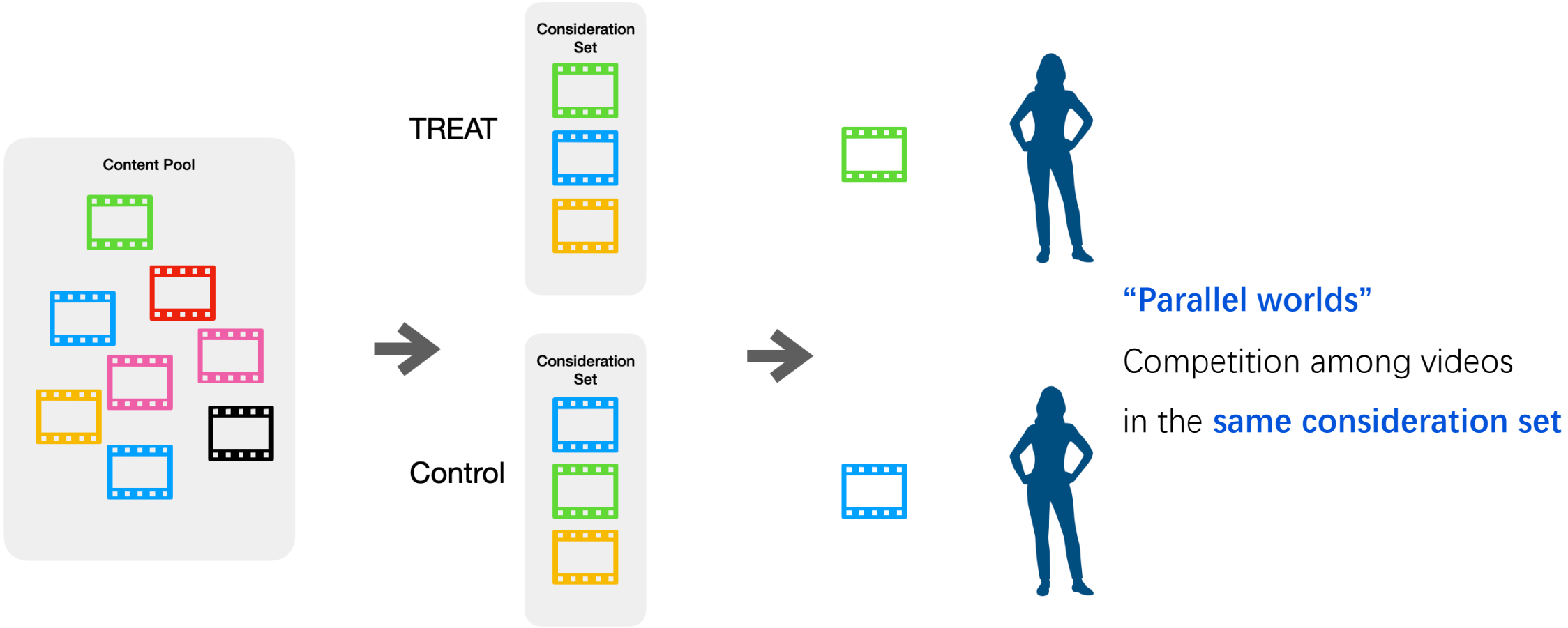
Analysis stage solution:

- Modeling the exposure mapping in a creator-side randomized experiment
- In potential outcome framework=> Potential outcome is determined by "binary" treatment indicator $Y_i(1), Y_i(0)$ **depends on $Z_i \in \{0,1\}$ (binary indicator)**
- Now: $Y_i(Z) = \tau Z_i + \beta e_i,$ where e_i is the proportion of creator i's competitors being in treated group => potential outcome now **depends on $Z \in \{0,1\}^n$ (n-dimensional vector)**

3. Design & Inference

(3) Marketplace Interference

Analysis stage solution - econ modeling



“Parallel worlds”

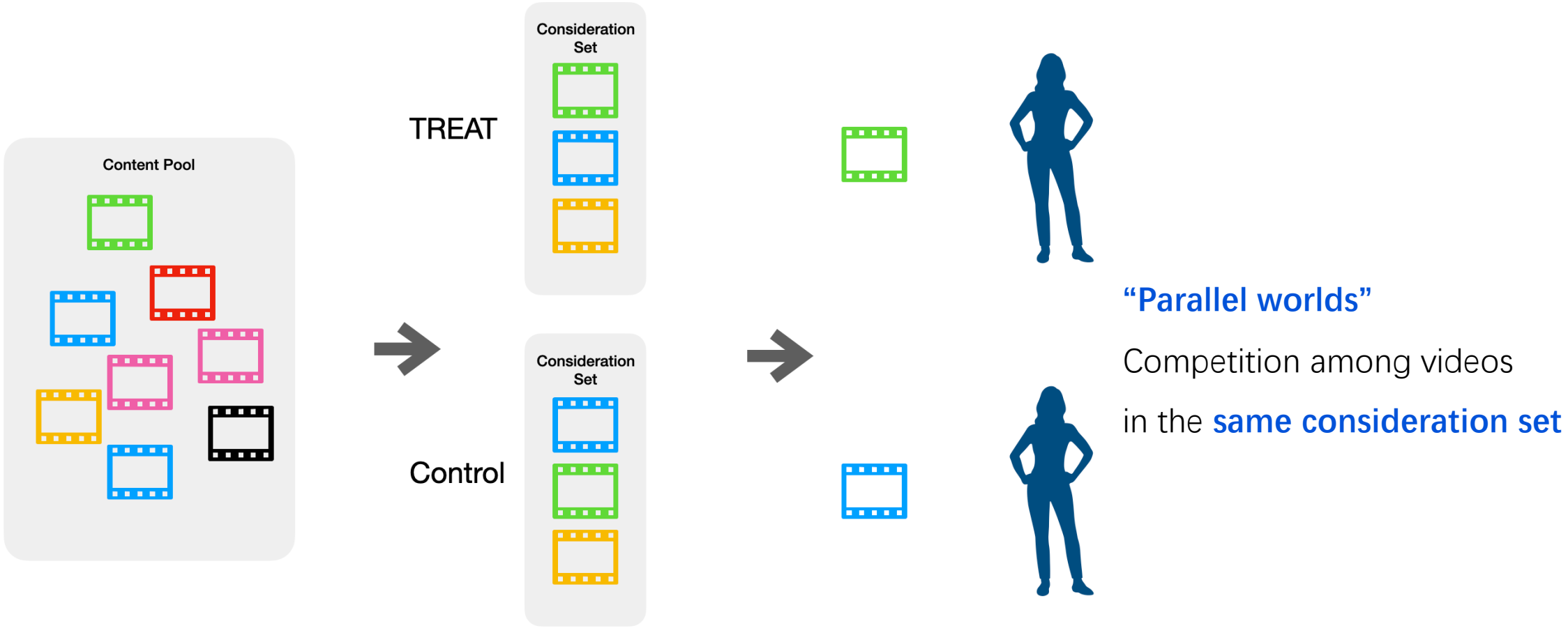
Competition among videos

in the **same consideration set**

3. Design & Inference

(3) Marketplace Interference

Analysis stage solution - econ modeling



“Parallel worlds”

Competition among videos

in the **same consideration set**

3. Design & Inference

(3) Marketplace Interference

Analysis stage solution - econ modeling

Observation (X_i, W_i, Y_i); $X_i := (U_i, V_i), Y_i := (E_i, R_i)$

Estimand $\tau = \mathbb{E}\left[\sum_{k=1}^{K} \frac{e^{s_0^*(U_i,V_{ik})+s_1^*(U_i,V_{ik})}}{\sum_{k'=1}^{K} e^{s_0^*(U_i,V_{ik'})+s_1^*(U_i,V_{ik'})}} \cdot z^*(U_i, V_{ik})\right] - \mathbb{E}\left[\sum_{k=1}^{K} \frac{e^{s_0^*(U_i,V_{ik})}}{\sum_{k'=1}^{K} e^{s_0^*(U_i,V_{ik'})}} \cdot z^*(U_i, V_{ik})\right]$

Global treatment — Global control

Estimator with nuisance $(\hat{s}_0, \hat{s}_1, \hat{z}) = \arg\min \left\{\frac{1}{n}\sum_{i=1}^{n} \ell(X_i, W_i, Y_i; s_0, s_1, z)\right\}$

$\tilde{\tau} = \frac{1}{n}\sum_{i=1}^{n} h(X_i; \hat{s}_1, \hat{s}_0, \hat{z}),$

Global treatment — Global control

$h(X_i; \hat{s}_1, \hat{s}_0, \hat{z}) = \left\{\sum_{k=1}^{K} \frac{e^{\hat{s}_0(U_i,V_{ik})+\hat{s}_1(U_i,V_{ik})}}{\sum_{k'=1}^{K} e^{\hat{s}_0(U_i,V_{ik'})+\hat{s}_1(U_i,V_{ik'})}} \cdot \hat{z}(U_i, V_{ik})\right\} - \left\{\sum_{k=1}^{K} \frac{e^{\hat{s}(U_i,V_{ik})}}{\sum_{k'=1}^{K} e^{\hat{s}(U_i,V_{ik'})}} \cdot \hat{z}(U_i, V_{ik})\right\}$

3. Design & Inference

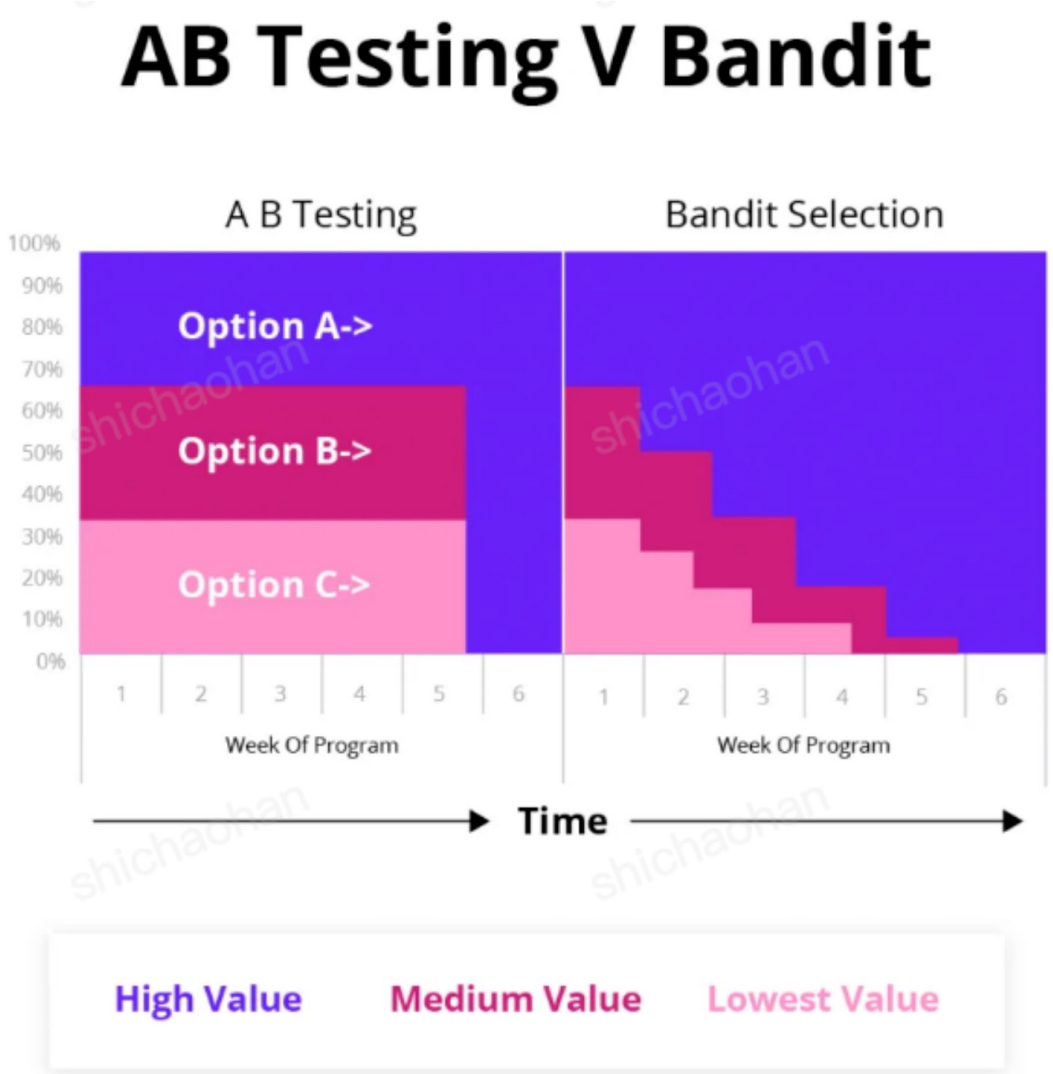
(3) Multi-Armed Bandit for Experimentation

Challenge

- End the experiment faster
- Higher cumulative rewards during the experiment

MAB algorithm and bandit experiment

- Experimental units arrive sequentially
- Rather than uniformly randomly assign users to each arm/treatment, **adaptively change the probability** so that arm/treatment with better outcome gets higher probability



Graph from: optimizely.com

3. Design & Inference

(4) Heterogeneous Treatment Effect(HTE)

- **Heterogeneous treatment effects** on different subgroups of experimental units
- How metrics are different on each subgroup of experimental units
- Covariates example: demographic features – gender, location, device type
- How to identify the subgroups?

3. Design & Inference

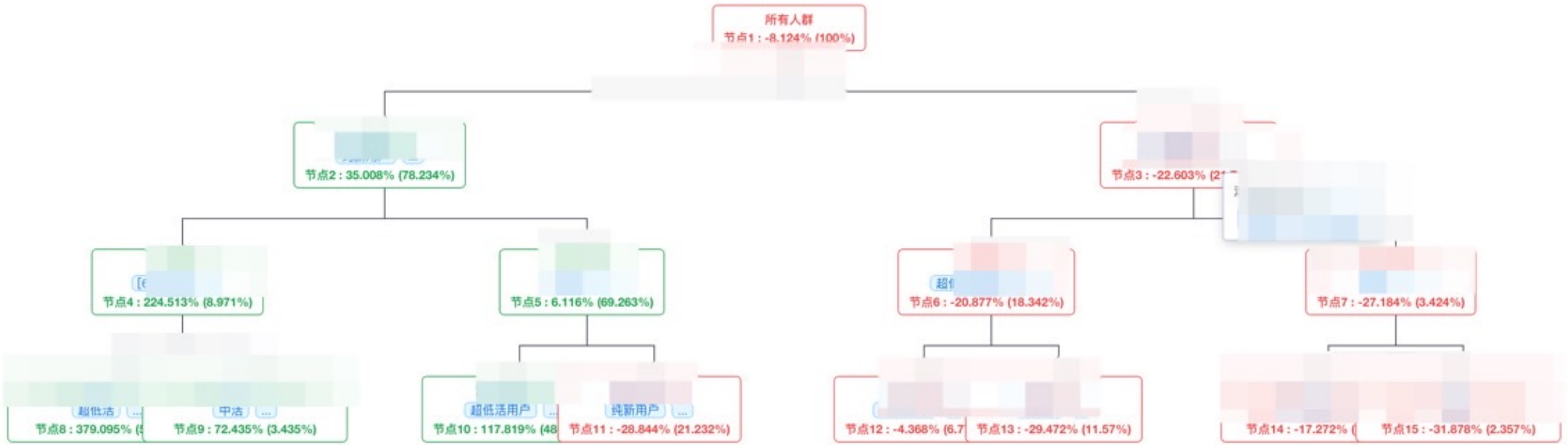
(4) Heterogeneous Treatment Effect(HTE)

- **Heterogeneous treatment effects** on different subgroups of experimental units
- How metrics are different on each subgroup of experimental units
- Covariates example: demographic features – gender, location
- Example: Male users in Shenzhen v.s. Male users in Shanghai v.s. Female users in Shenzhen …
- Challenges

3. Design & Inference

(4) Heterogeneous Treatment Effect(HTE)

- Causal Tree – recursively partitioning
- Automatically give subgroups that differ in average treatment effect



Summary

- AB testings are golden standards for decision making in IT industry
- IT industry has advantages in implementing AB testings in large scale
- Engineers, data scientists and statistics work together to ensure reliable and replicable inference results
- Challenges and opportunities

Thanks

Tencent 腾讯