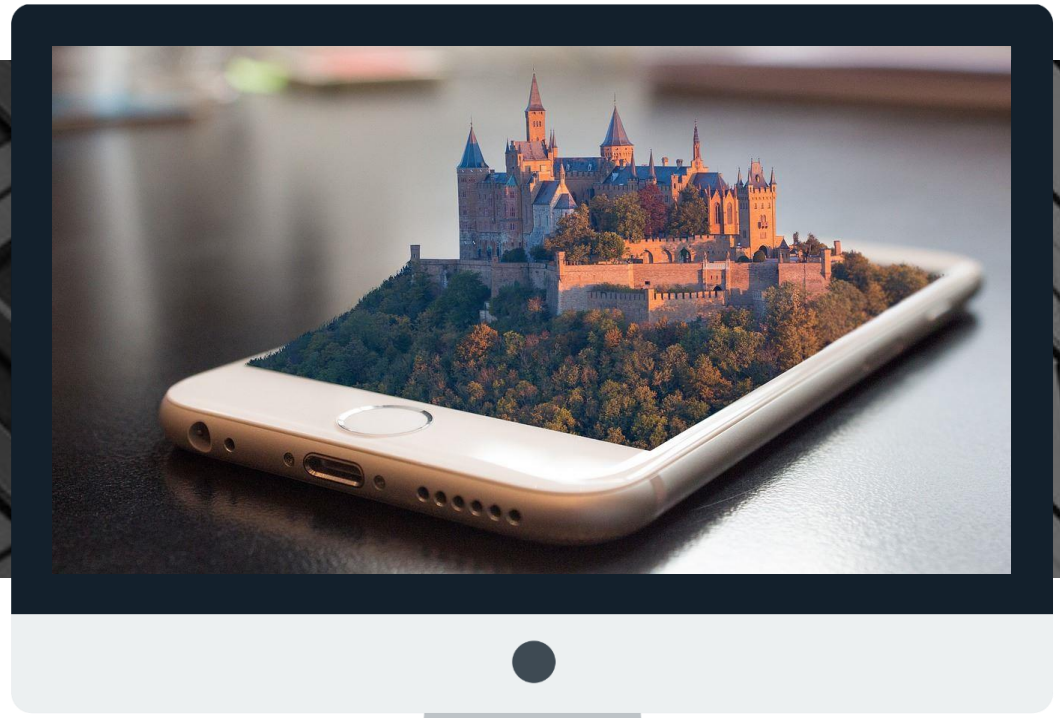
Kaggle

TalkingData Mobile User Demographics



CHAO WANG

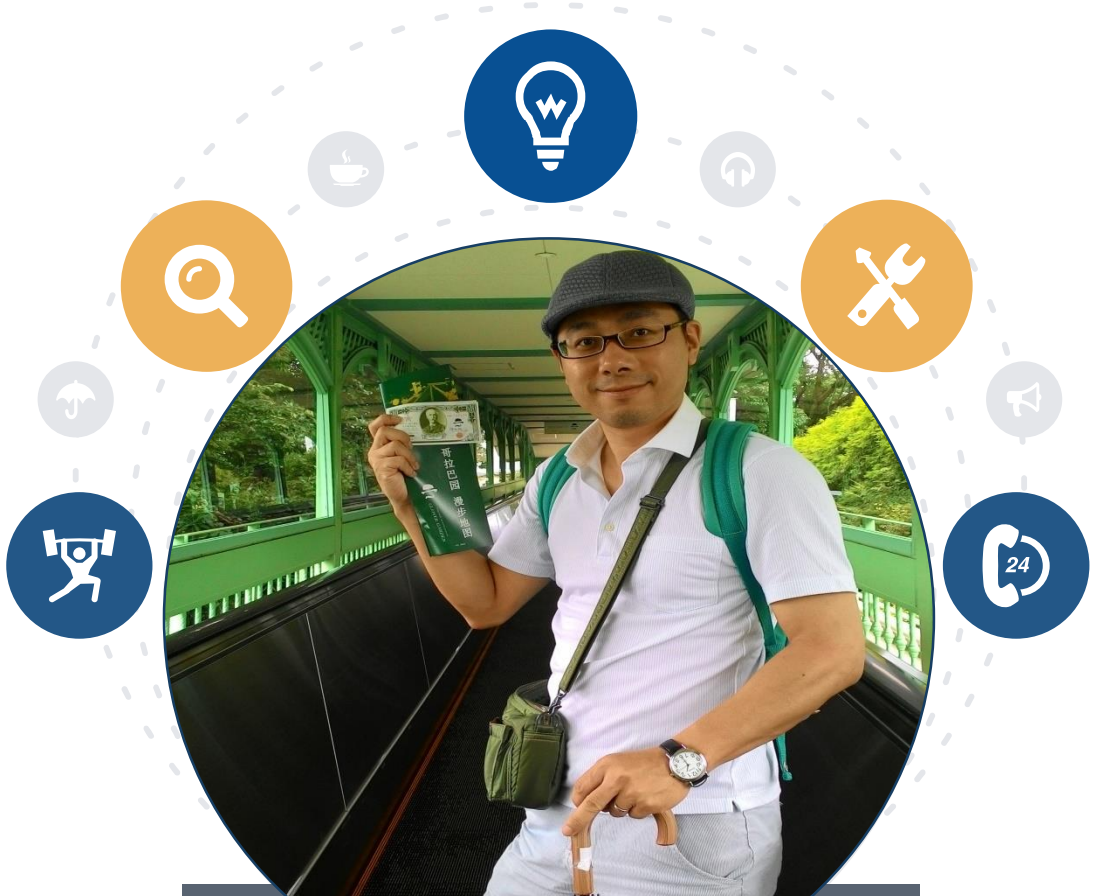
Self Introduction

Experience

- worked 3 years for information industry
- AIGO competition : Masterpiece award

Skill

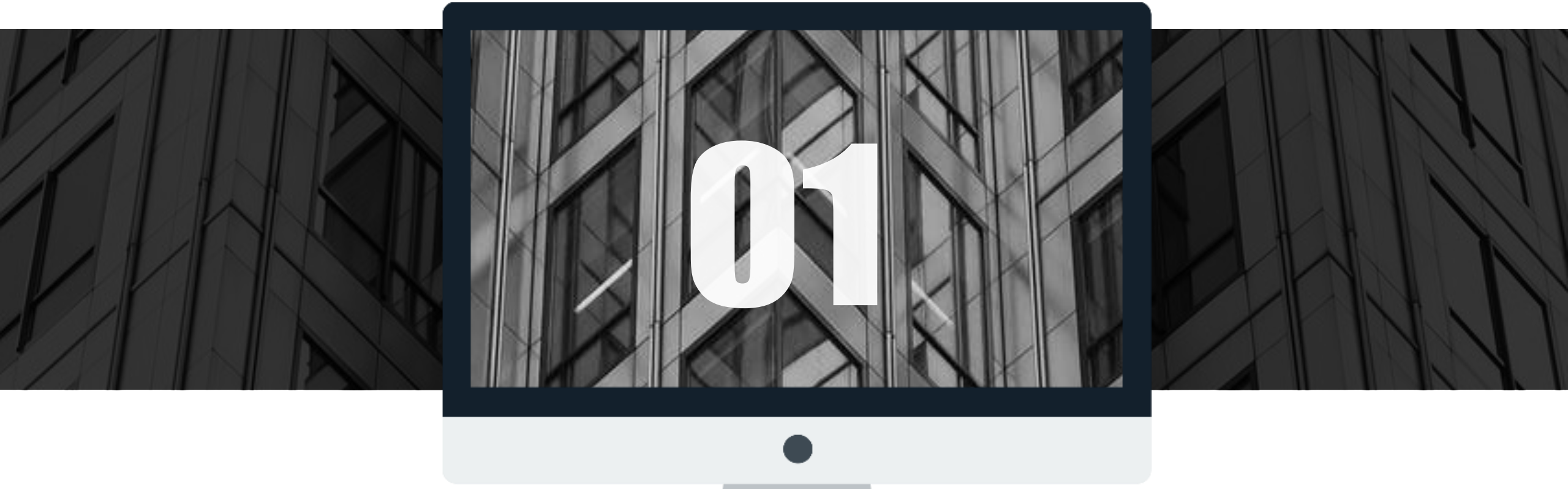
- Python
- MySQL
- ETL
- Java



Education

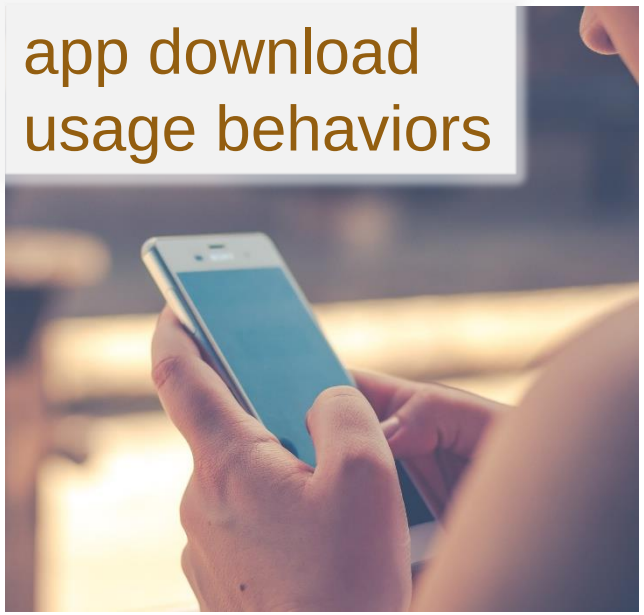
- Master
- Shanghai Normal University

Project Introduction



Project Objective

Predict the demographics of a user (gender and age) based on their app download and usage behaviors

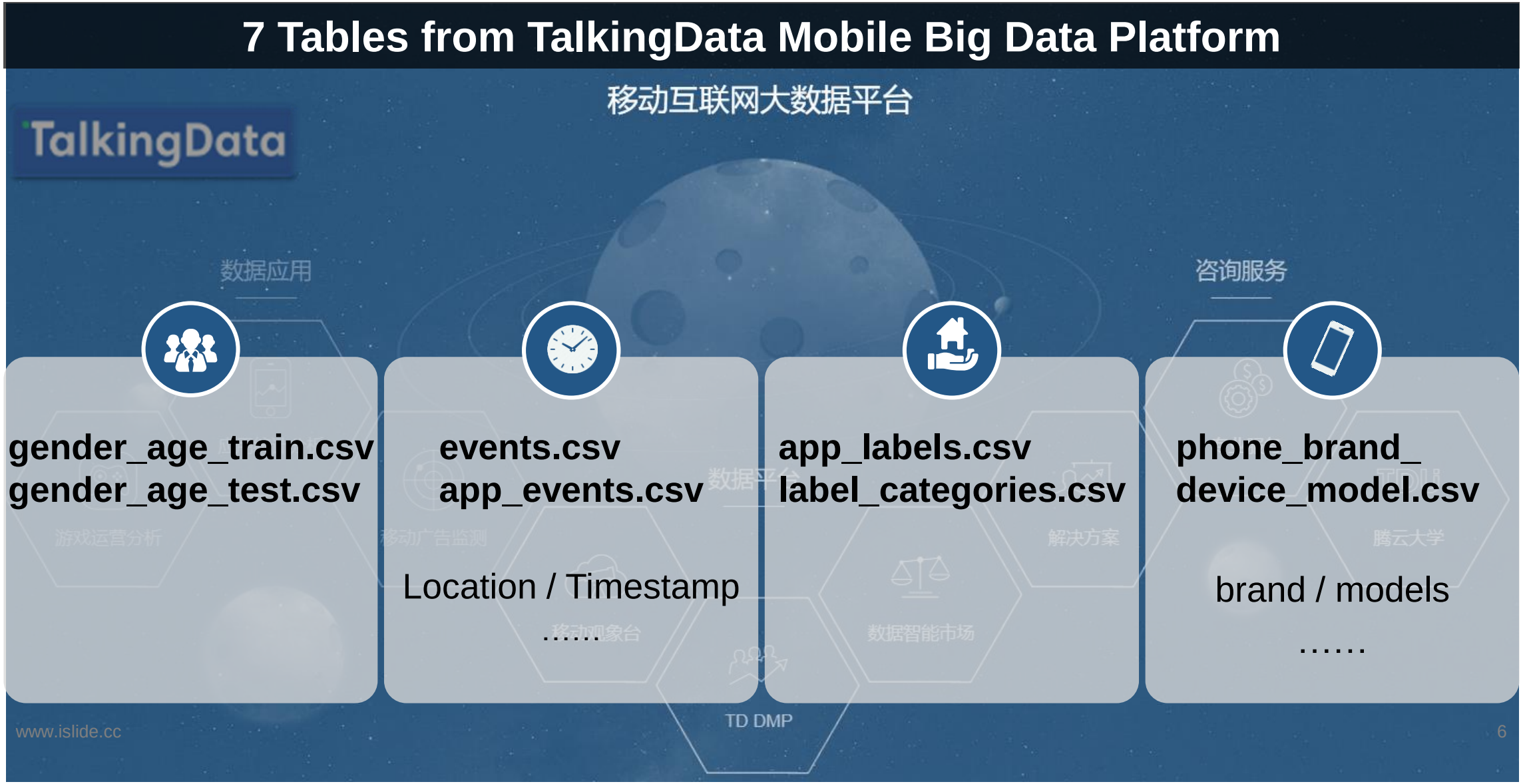


Input

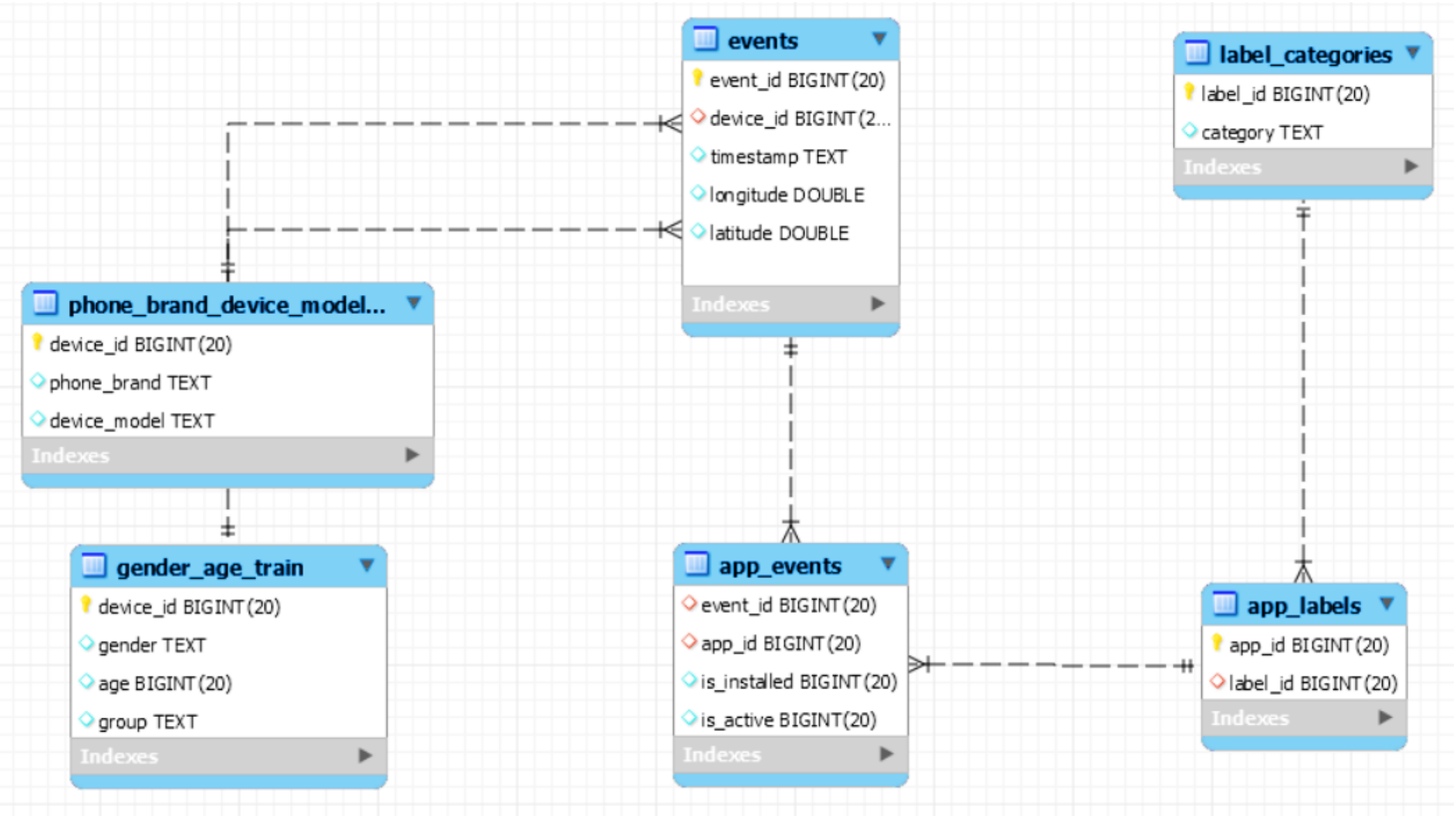
predict

Output

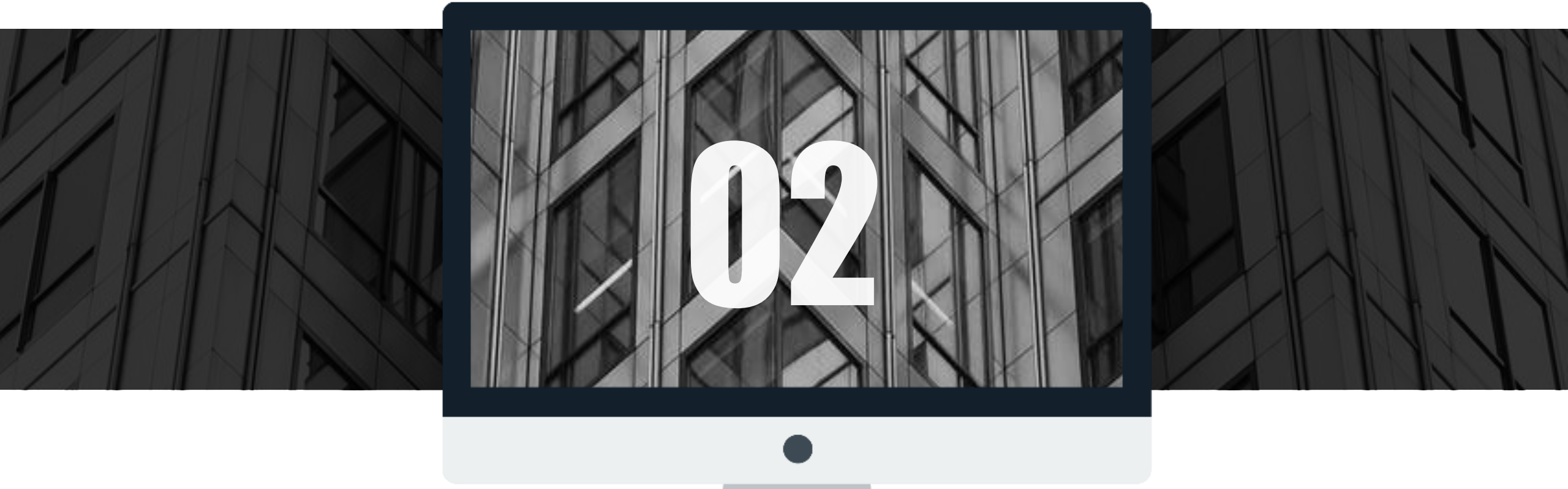
Collected data



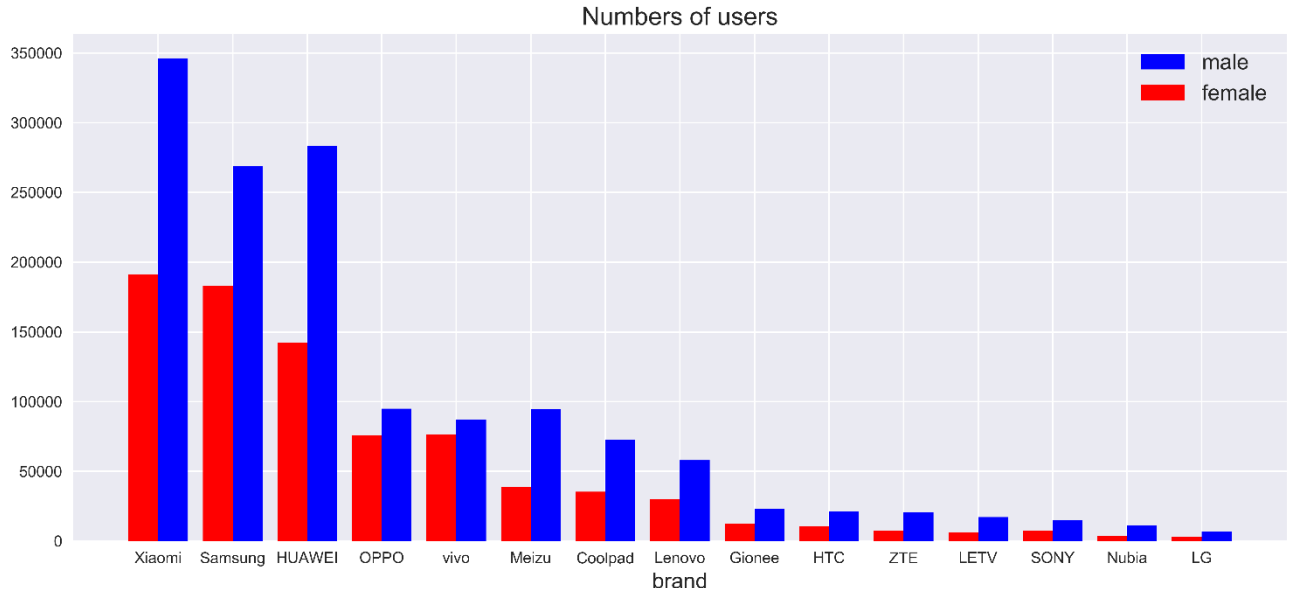
Entity-Relationship Model

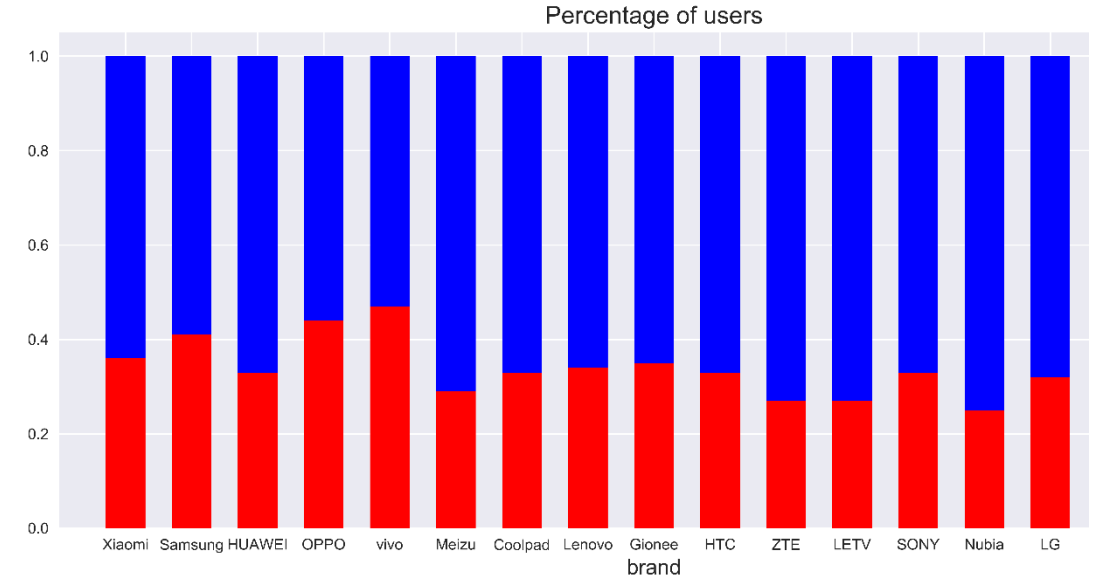


Exploratory Data Analysis



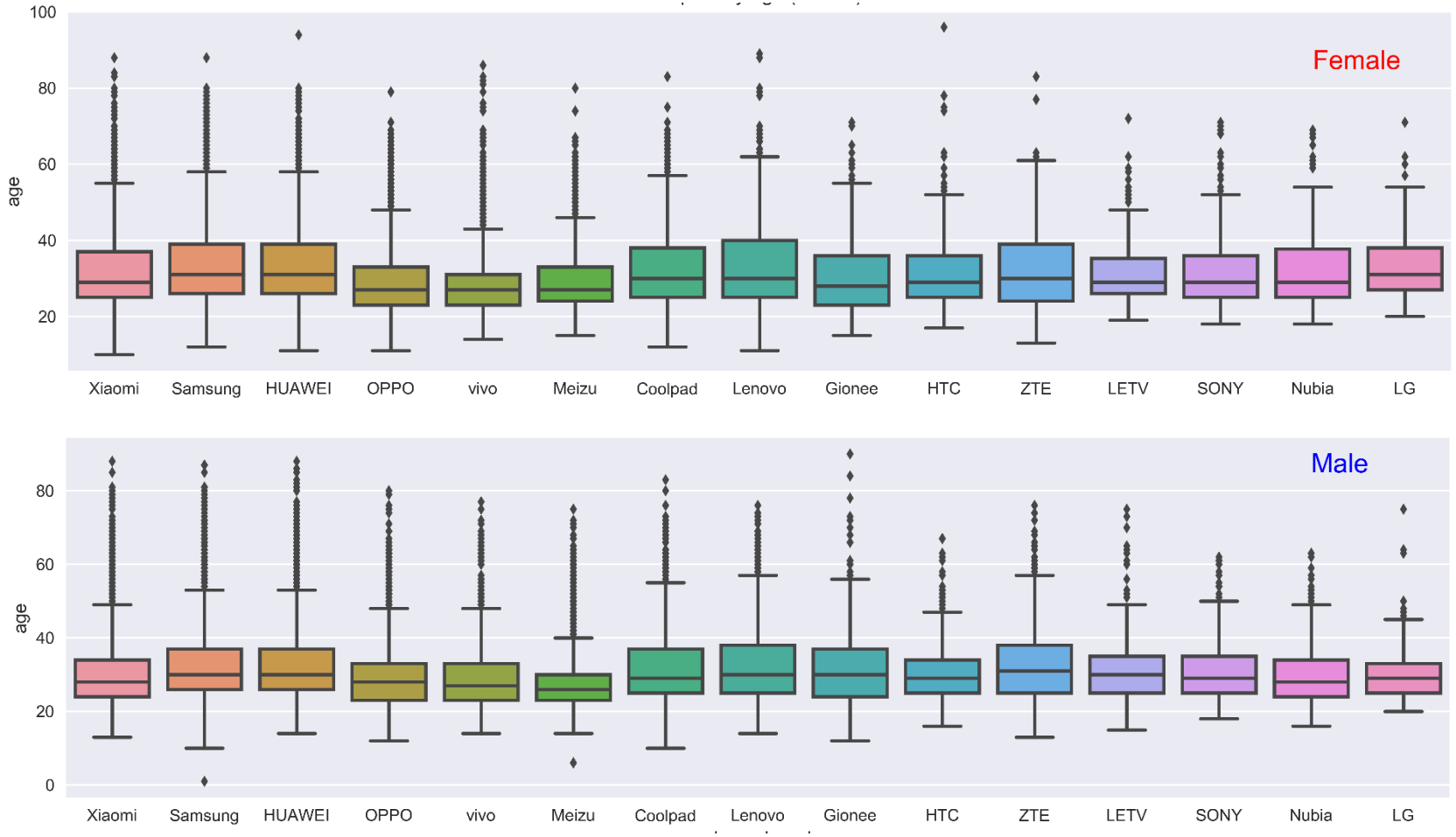
Top 15 by gender





- XIAOMI is the king here
- HUAWEI and MEIZU preferred by male
- OPPO and vivo are preferred by female

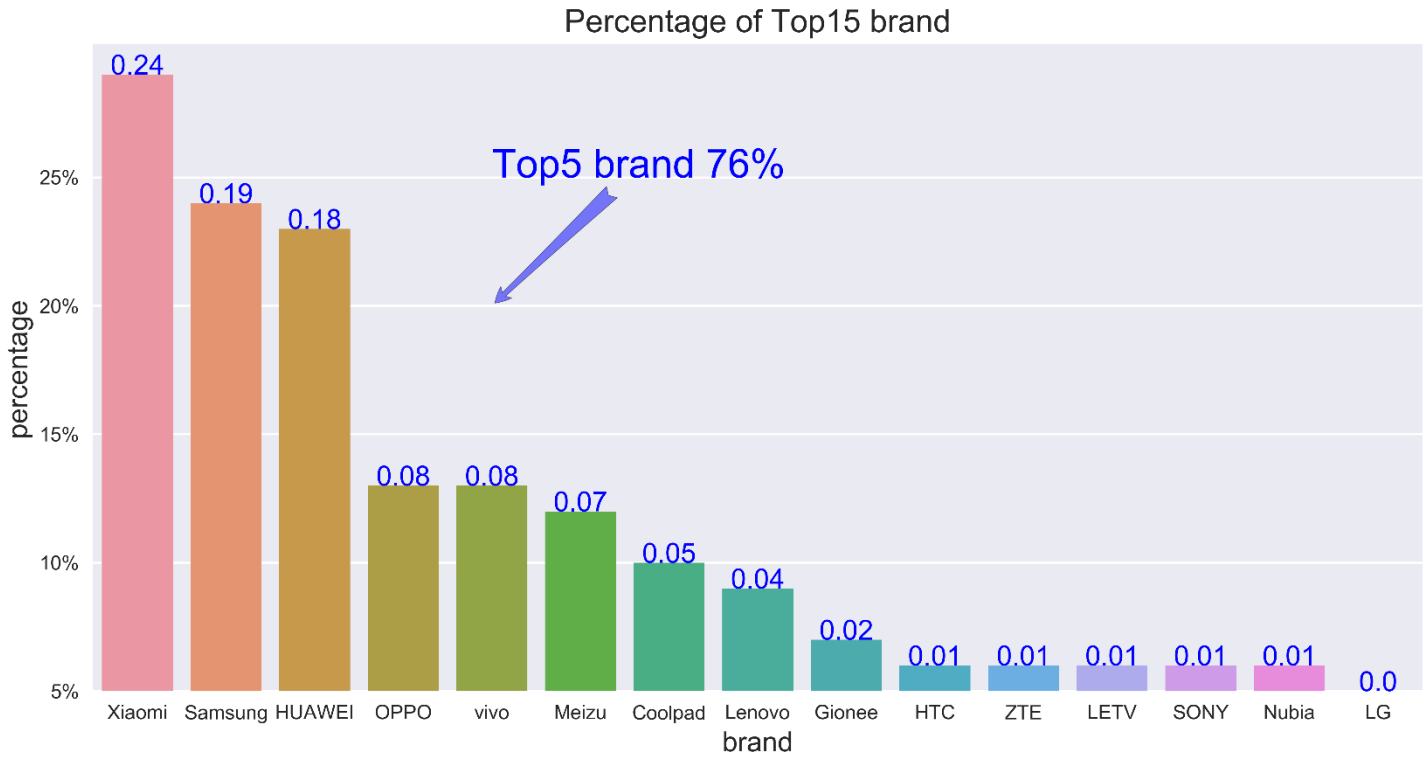
Top 15 by Age

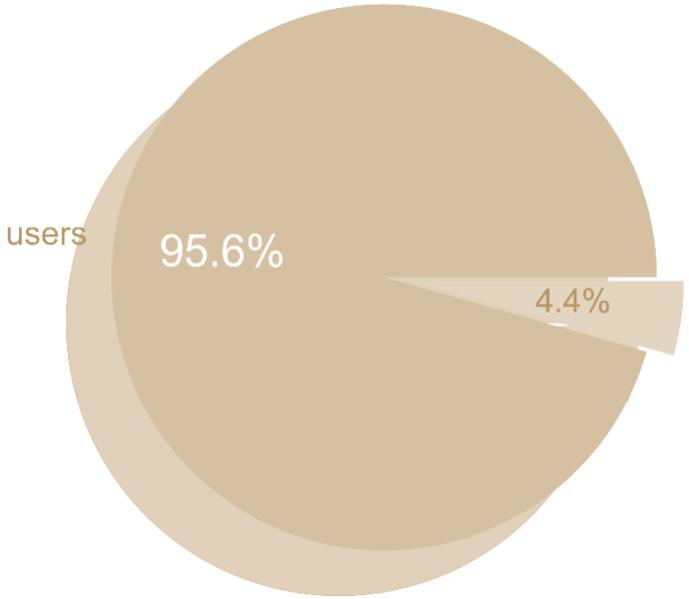


- User age is mainly in the range of 20 ~ 40
- Similar between top 15 brands

User Percentage

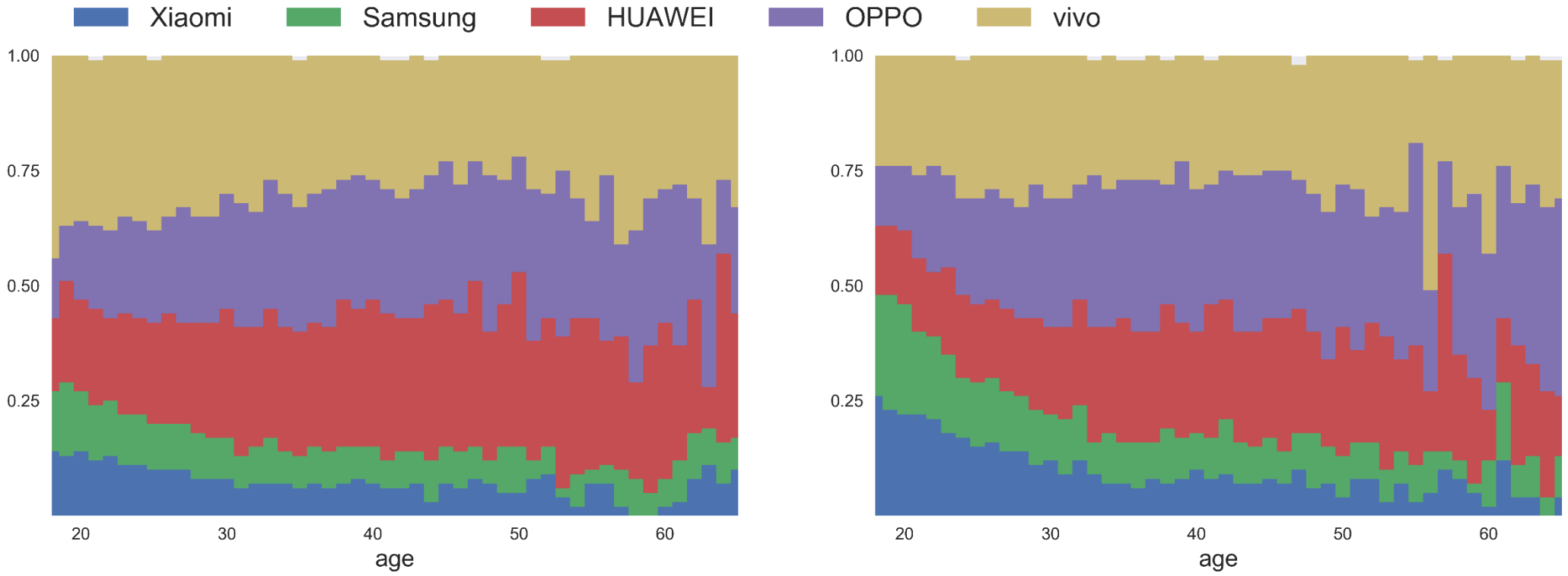
- Top 5 brands : XIAOMI、Samsung、HUAWEI、vivo、OPPO
- User percentage of Top 5 brands :
  accounting for more than 76% of the total population

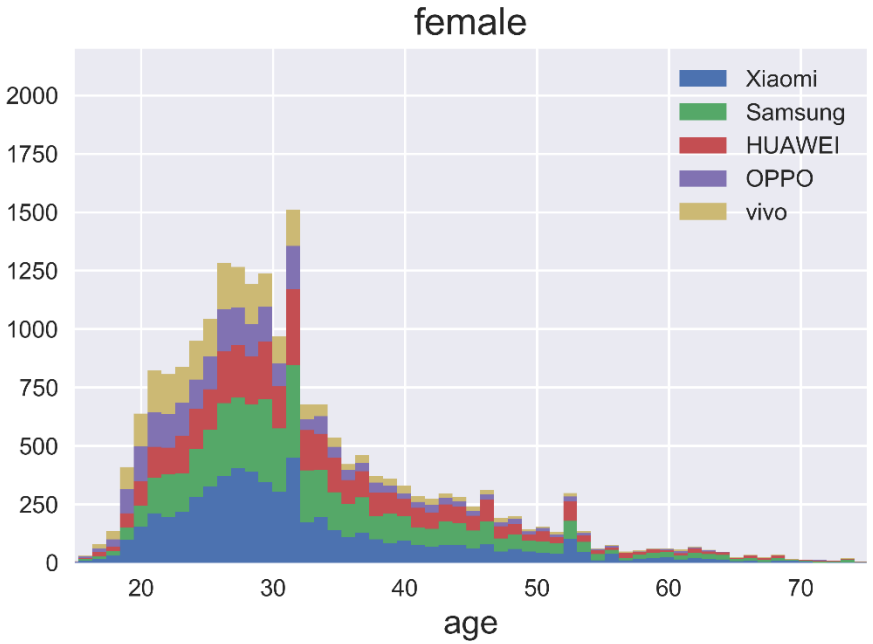




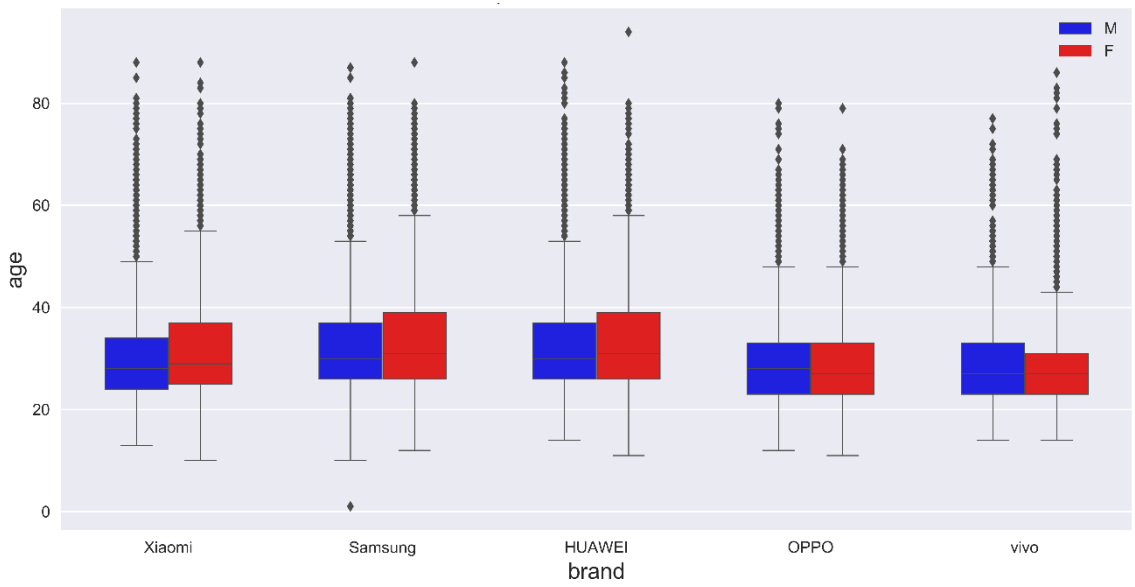
Top 5 Distribution

- Young generation prefer OPPO and vivo (may be more affordable)
- Senior people prefer HUAWEI and Samsung
- Xiaomi is the dominant




female
Xiaomi
Samsung
HUAWEI
OPPO
vivo
2000
1750
1500
1250
1000
750
500
250
0
20
30
40
50
60
70
age

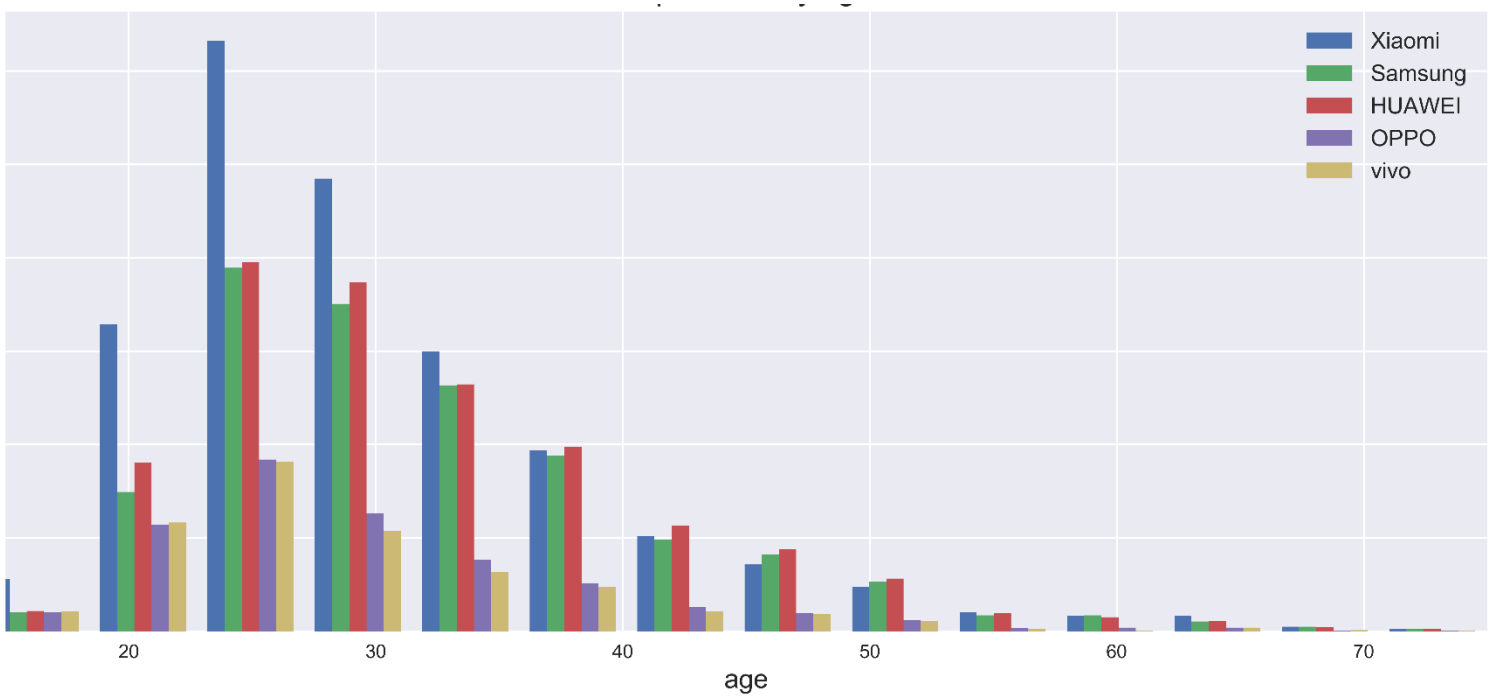

male
Xiaomi
Samsung
HUAWEI
OPPO
vivo
2000
1750
1500
1250
1000
750
500
250
0
20
30
40
50
60
70
age


M
F
80
60
40
20
0
age
Xiaomi
Samsung
HUAWEI
OPPO
vivo
brand

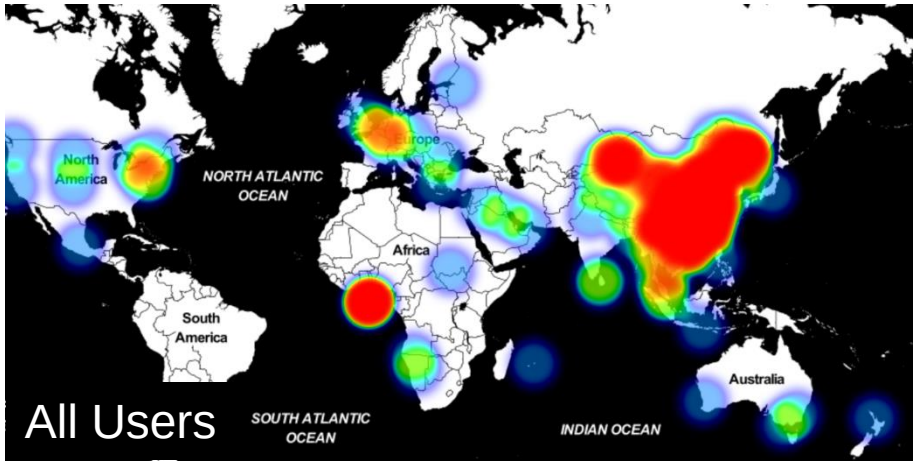
Top 5 count by age

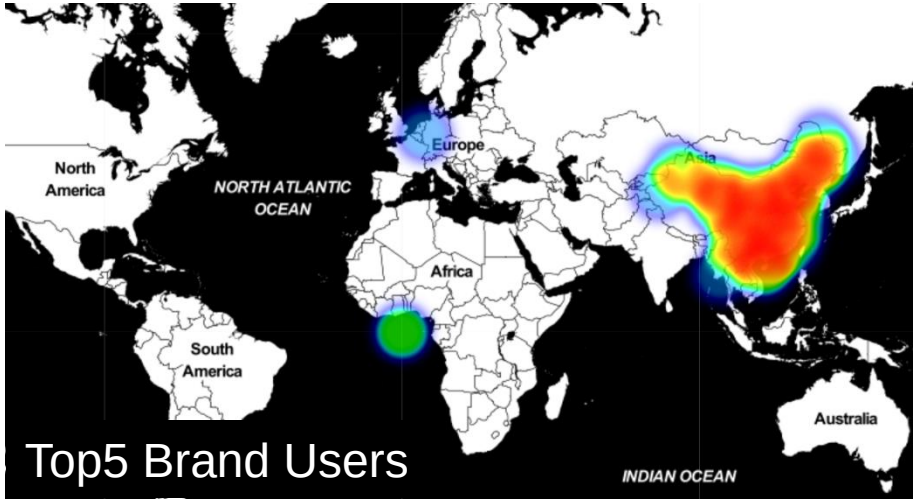
XIAOMI

- particularly popular among young people between the ages of 10 and 20
- with the improvement of user’s age, its dominant position is becoming less and less obvious

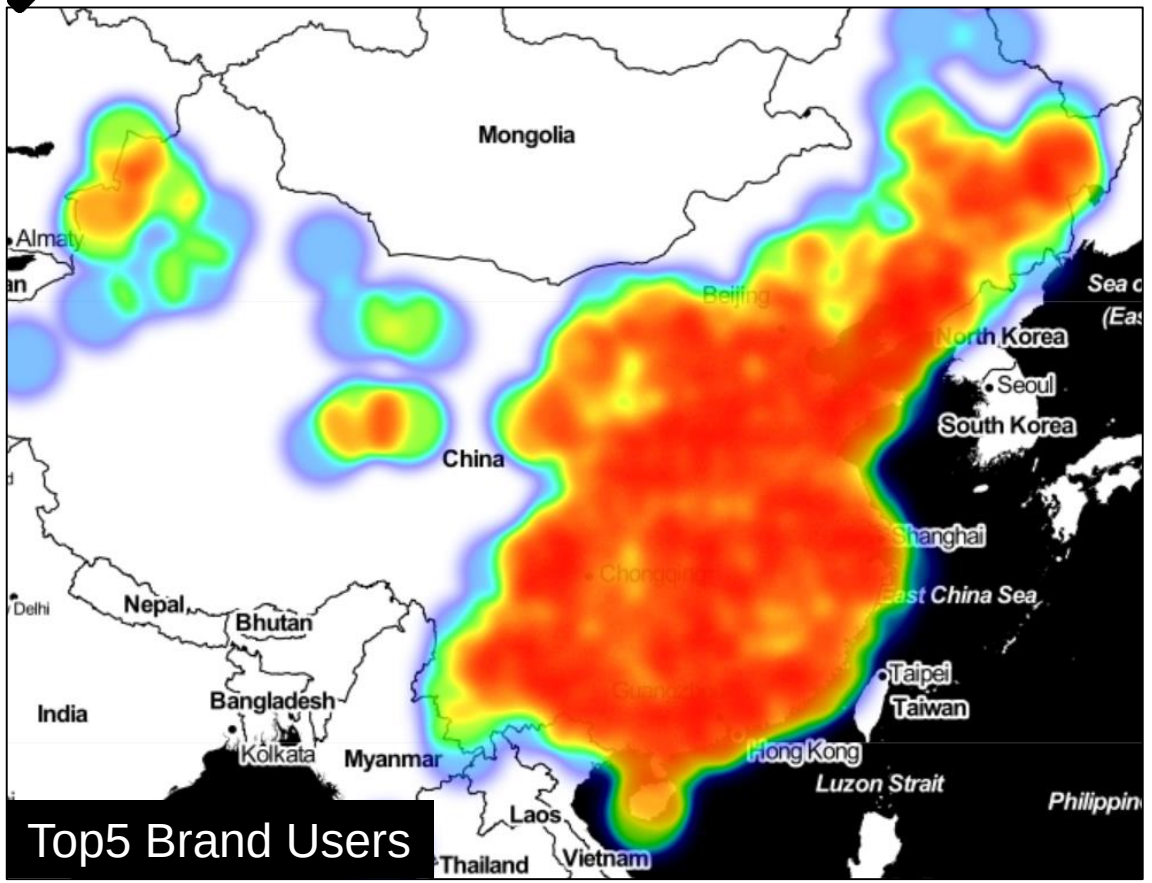


Users Distribution HeatMap （Folium）

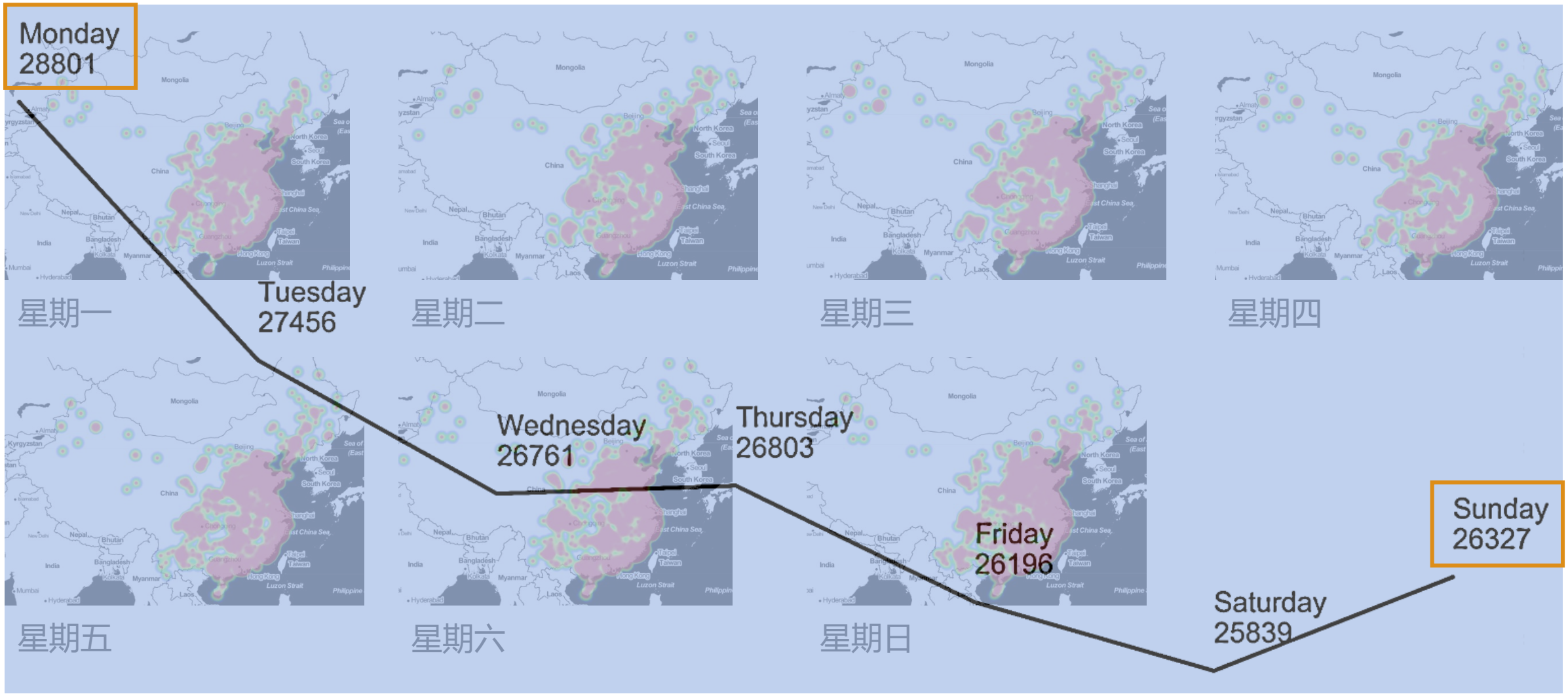




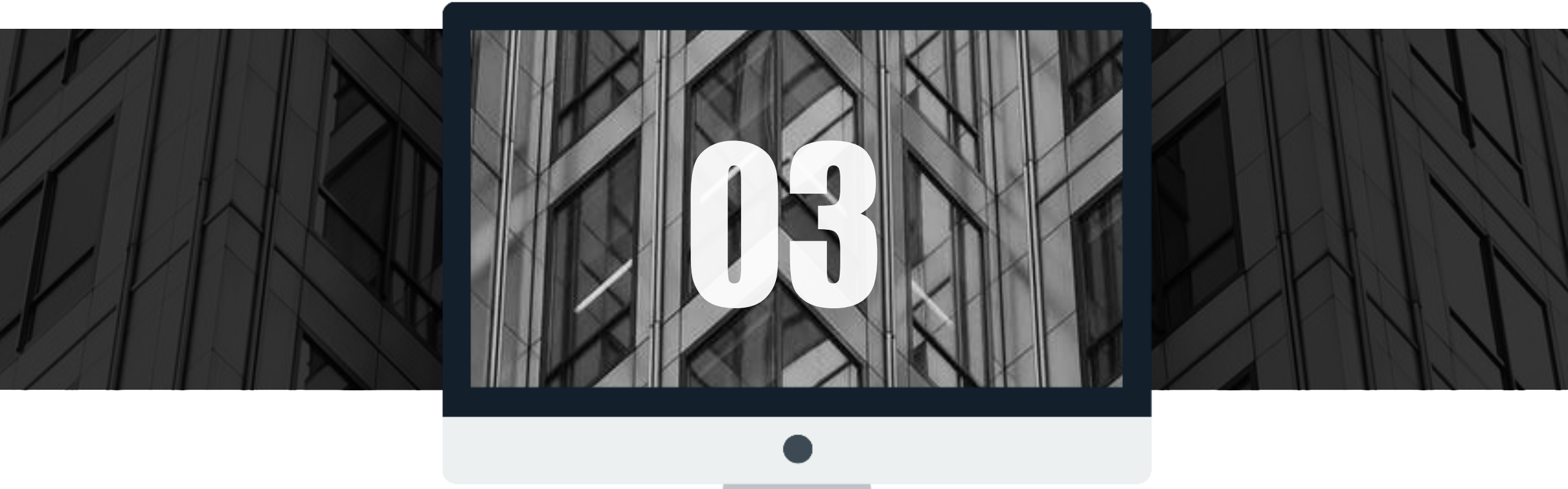
Coastal areas and Sichuan of China



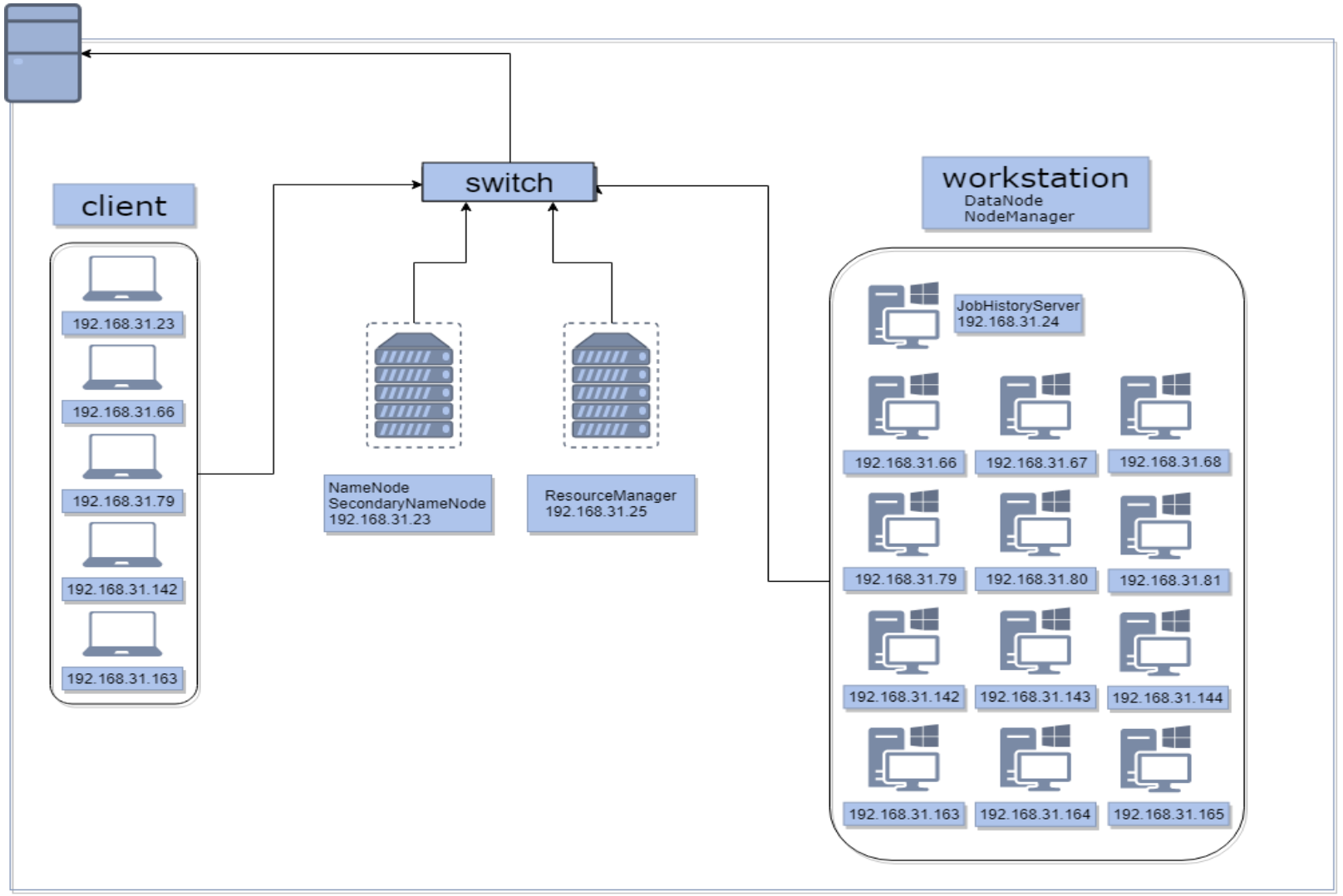
HeatMap with Time Series



Data Preprocessing



Hadoop Cluster



- Memory：78GB
- Cores：26
- Worker:13
- NameNode:1
- ResourceManager:1
- JobHistoryServer:1
- Passwordless login

Table 1 : Device_id vs Brand and Model

Duplicated data : 529
Missing data : 0
One hot encoding

Shape : 186,716 rows × 1,731 columns
(device_id) (brand and model)

Device_id	HTC	XIAOMI	HUAWEI	Samsung	LG
1	0	1	0	0	0
2	0	0	1	0	0
3	1	0	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

Table 2 : Device_id vs App_id

Duplicated data : 0
Missing data : 0
One hot encoding

Shape : 60,669 rows × 19,237 columns
(device_id) (app_id)

Device_id	APP_1	APP_2	APP_3	APP_4	APP_5
1	0	0	1	1	0
2	1	0	1	0	0
3	0	1	0	0	0
4	0	0	0	0	1
5	1	0	0	1	0

Table 3 : Device_id vs App categories

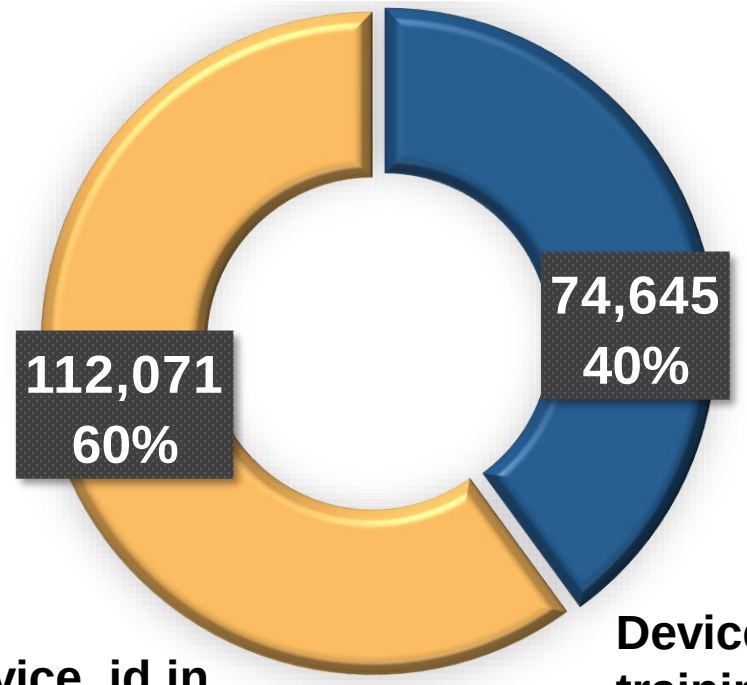
Shape : 60,669 rows × 473 columns
(device_id) (app categories)

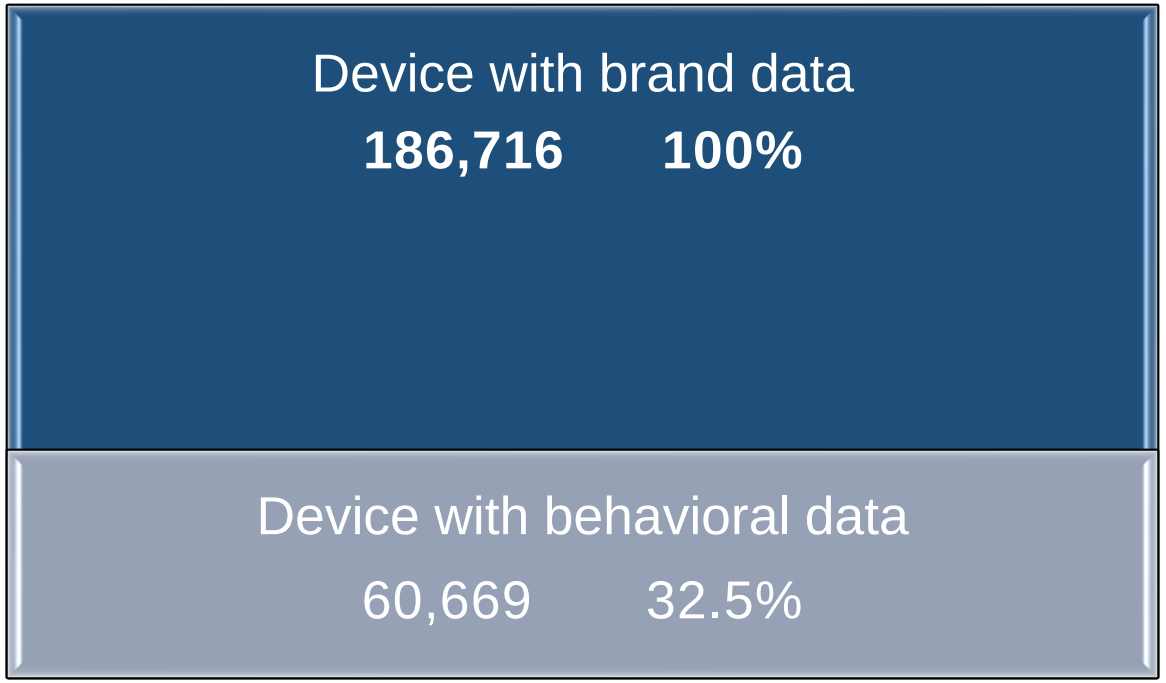
Duplicated data : 491
Missing data : 0
One hot encoding

Device_id	reputation	vitality	kingdom game	Japanese comic	RPG
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

Primary key : Device_id

Numbers of Device_id : 186,716





Join Table

- gender_age_train.csv
- Table 1
- Table 2
- Table 3

→

Training data

- Sparse matrix
- 74,645 rows × 21,441 columns

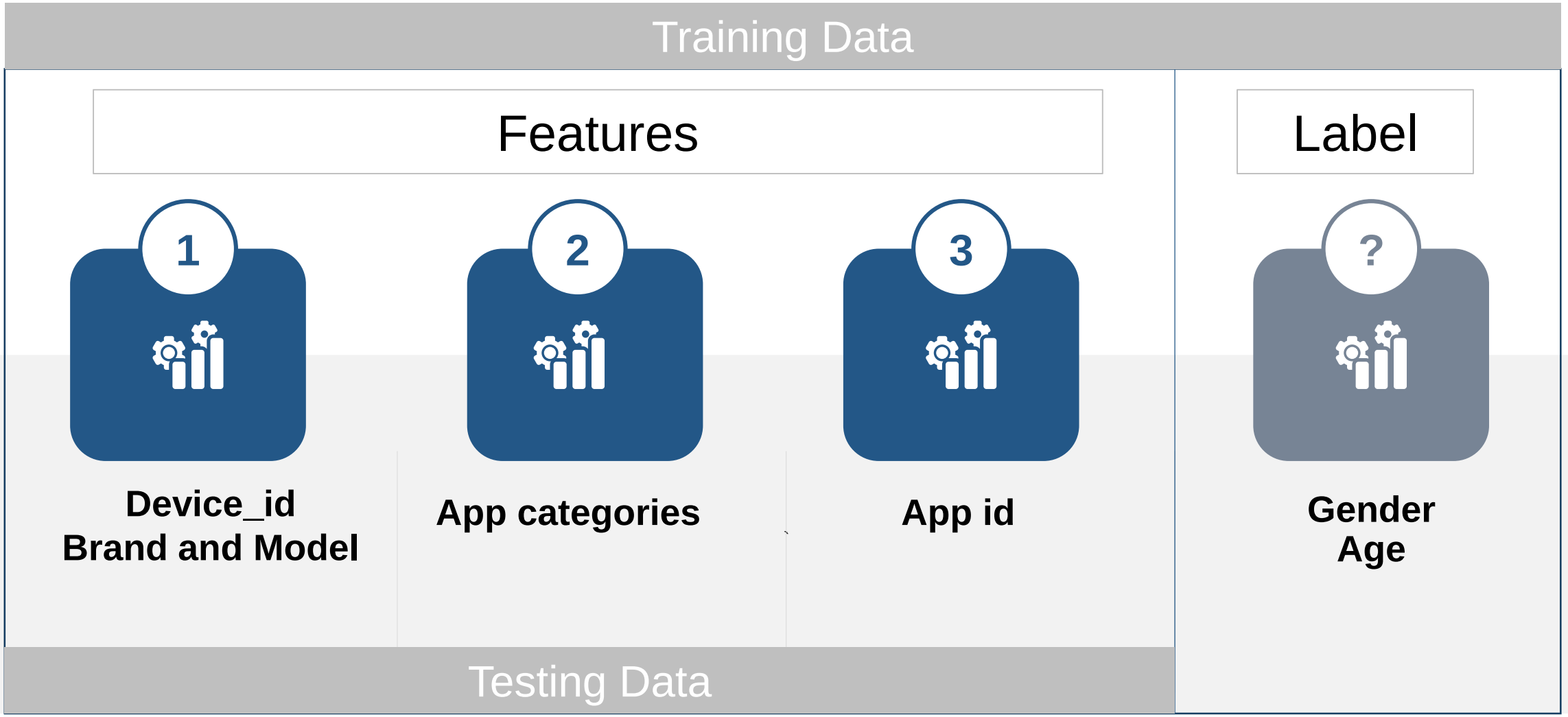
- gender_age_test.csv
- Table 1
- Table 2
- Table 3

→

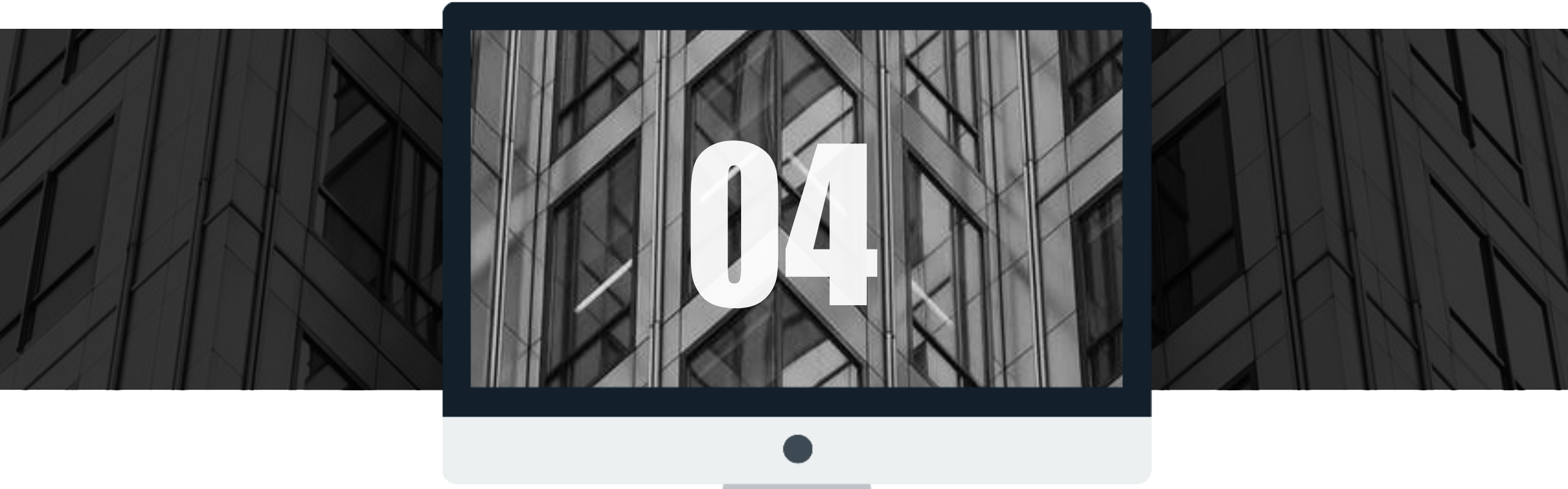
Testing data

- Sparse matrix
- 112,071 rows × 21,441 columns

Tables


Training Data
Features
Label
1
2
3
?
Device_id
Brand and Model
App categories
App id
Gender
Age
Testing Data

Model



Kaggle Submission

Device_id	F23-	F24-26	F27-28	F29-32	F33-42	F43+	M22-	M23-26	M27-28	M29-31	M32-38	M39+
1	0.05%	0.13%	0.20%	0.72%	2.52%	5.20%	0.36%	2.25%	4.02%	10.04%	23.72%	50.50%
2	0.77%	2.15%	2.48%	5.35%	8.14%	6.82%	1.56%	8.35%	9.47%	14.23%	23.19%	17.45%
3	2.73%	4.70%	4.83%	10.71%	16.52%	10.97%	1.46%	4.15%	4.58%	8.14%	15.06%	16.16%
4	0.72%	1.14%	1.09%	1.89%	4.52%	8.26%	6.59%	12.87%	8.48%	12.10%	16.59%	25.76%
5	3.79%	4.96%	4.10%	6.18%	6.90%	5.51%	8.25%	13.56%	9.06%	11.94%	14.16%	11.56%

Device_id	Gender / Age
1	M39+
2	M32-38
3	F33-42
4	M39+
5	M32-38

Target field

Kaggle Rank:

334 / 1689 （19.8%）

Build Models

Multiclass classification

Neural Network

- 2 Hidden Layers
- Validation Accuracy: 8.3% → 22.4%
- Kaggle Rank: 334 / 1689 (19.8%)

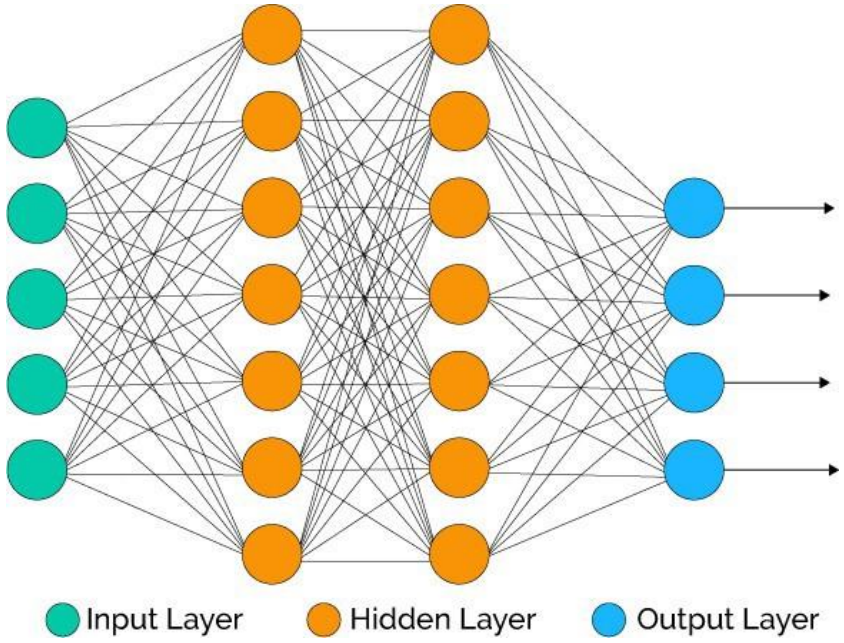
XGBoost

- 500 Trees
- Validation Accuracy: 8.3% → 19.2%
- Kaggle Rank: 810 / 1689 (48.0%)

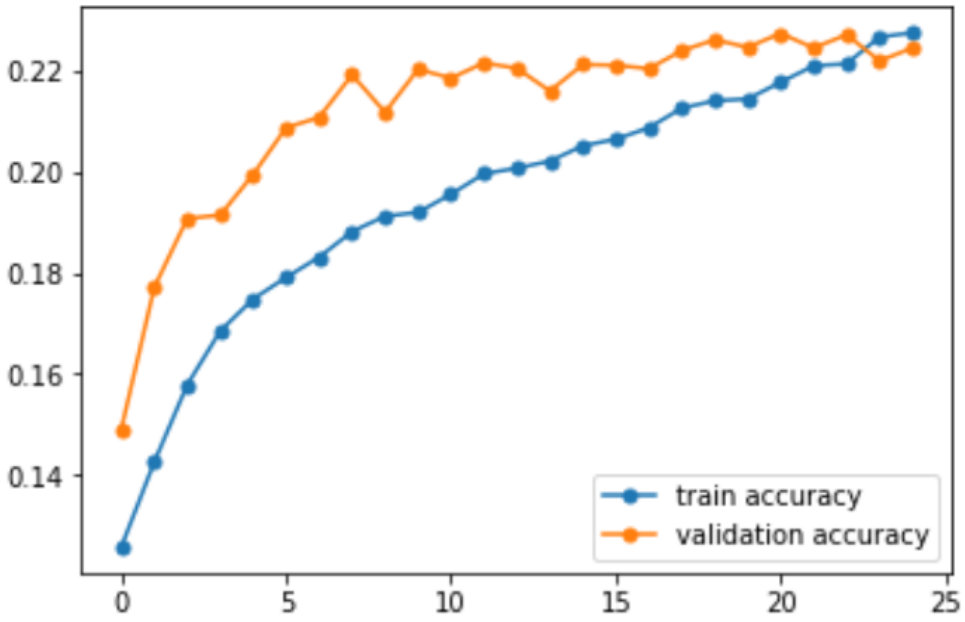
Random Forest

- 500 Trees
- Validation Accuracy: 8.3% → 20.8%
- Kaggle Rank: 774 / 1689 (45.8%)

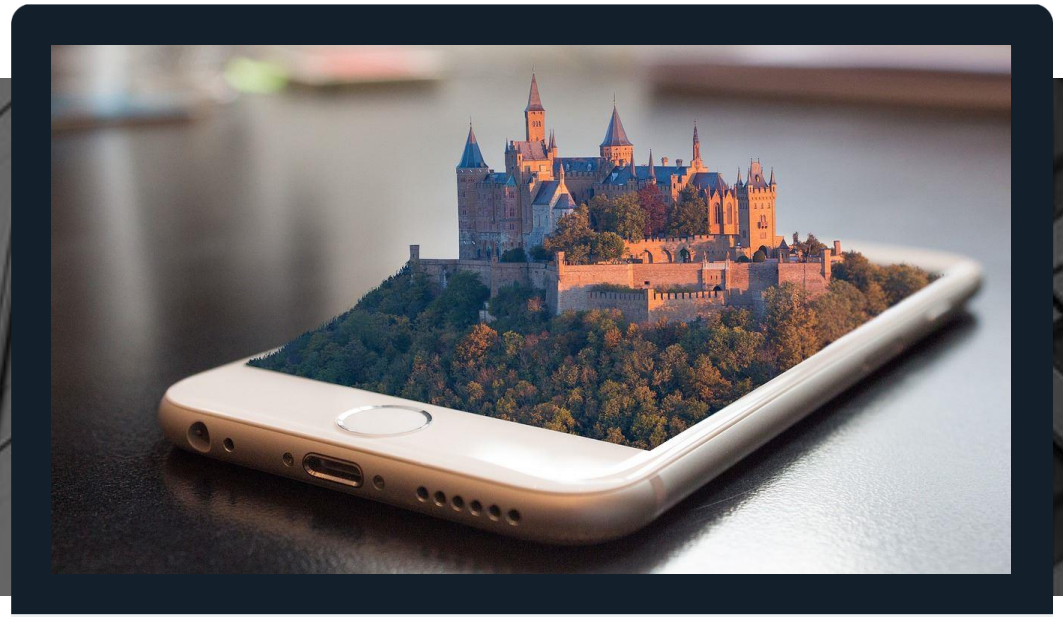
Neural Network



Framework



Training



Thanks

Get to know millions of mobile device users

CHAO WANG