# TalkingData

Kaggle Project

# Problem Description

Fraud risk is everywhere, but for companies that advertise online, click fraud can happen at an overwhelming volume, resulting  in misleading click data and wasted money.

# Problem Description

Talkingdata, China’s largest independent big data service platform, covers over 70% of active mobile devices nationwide. They handle 3 billion clicks per day, of which 90% are potentially fraudulent.

# Problem Description

We need to build an algorithm that predicts whether a user will download an app after clicking a mobile app ad.

# Data Description

We have **training set** 7.5Gb:

total number of instances = 184 903 890, number of features = 7,

number of instances in each class is:

≈0,25% in class 1 and ≈99,75% in class 0

And **test set** 860Mb:

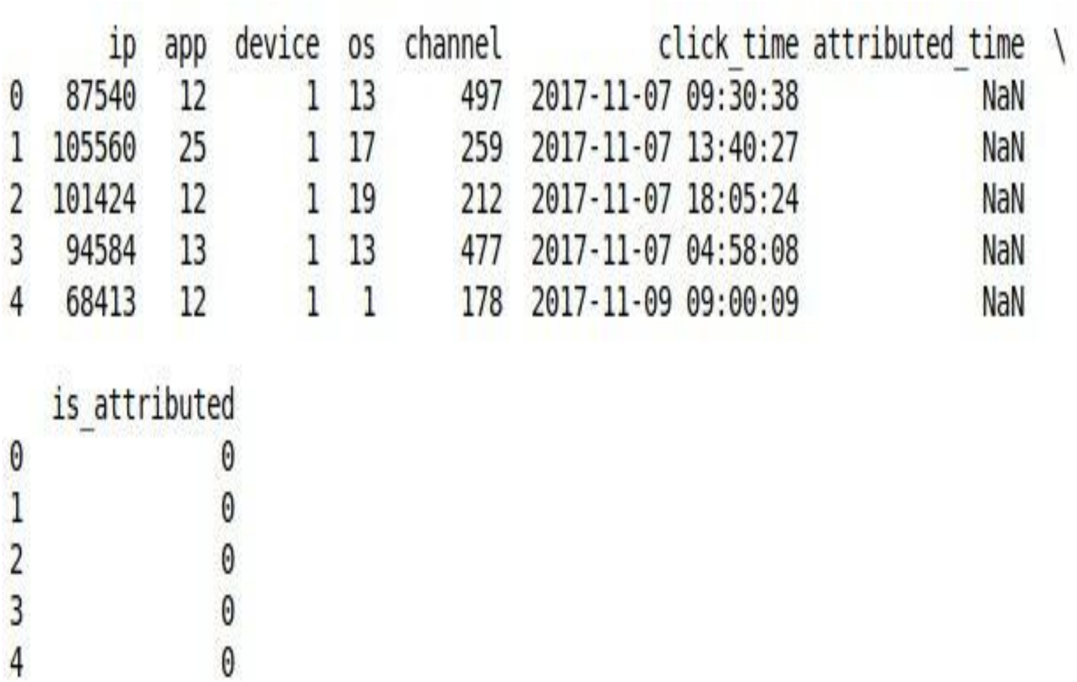
total number of instances = 18 790 469

# Features Description

- **ip**: ip address of click.
- **app**: app id for marketing.
- **device**: device type id of user mobile phone (e.g., iphone 6 plus, iphone 7, huawei mate 7, etc.)
- **os**: os version id of user mobile phone

# Features Description

- **channel**: channel id of mobile ad publisher
- **click_time**: timestamp of click (UTC)
- **attributed_time**: if user download the app for after clicking an ad, this is the time of the app download
- **is_attributed**: the target that is to be predicted, indicating the app was downloaded



```
       ip  app  device  os  channel           click_time attributed_time  \
0   87540   12       1  13      497  2017-11-07 09:30:38             NaN
1  105560   25       1  17      259  2017-11-07 13:40:27             NaN
2  101424   12       1  19      212  2017-11-07 18:05:24             NaN
3   94584   13       1  13      477  2017-11-07 04:58:08             NaN
4   68413   12       1   1      178  2017-11-09 09:00:09             NaN

   is_attributed
0              0
1              0
2              0
3              0
4              0
```

# Features Description

| | ip | app | device | os | channel | click_time |
|---|---|---|---|---|---|---|
| number of different values in each feature | 277396 | 706 | 3475 | 800 | 202 | 259620 |
| range of each feature | min 1<br>max 364778 | min 0<br>max 768 | min 0<br>max 4227 | min 0<br>max 956 | min 0<br>max 500 | min 2017-11-06 14:32:21<br>max 2017-11-09 16:00:00 |

# Problems:

1. Almost all features are encoded, so we cannot extract any additional information

2. Classes in the training set are not balanced

# Possible Solutions:

1. Logistic regression

2. Random forest

3. Neural network

4. Gradient boosting