

# E6893 Big Data Analytics:

## *User Behavior Modeling*

Team Members:

Avery Wu (yw2928)

Tsung-Yi Huang (th2668)

Minghong Zheng (mz2597)



December 15, 2016

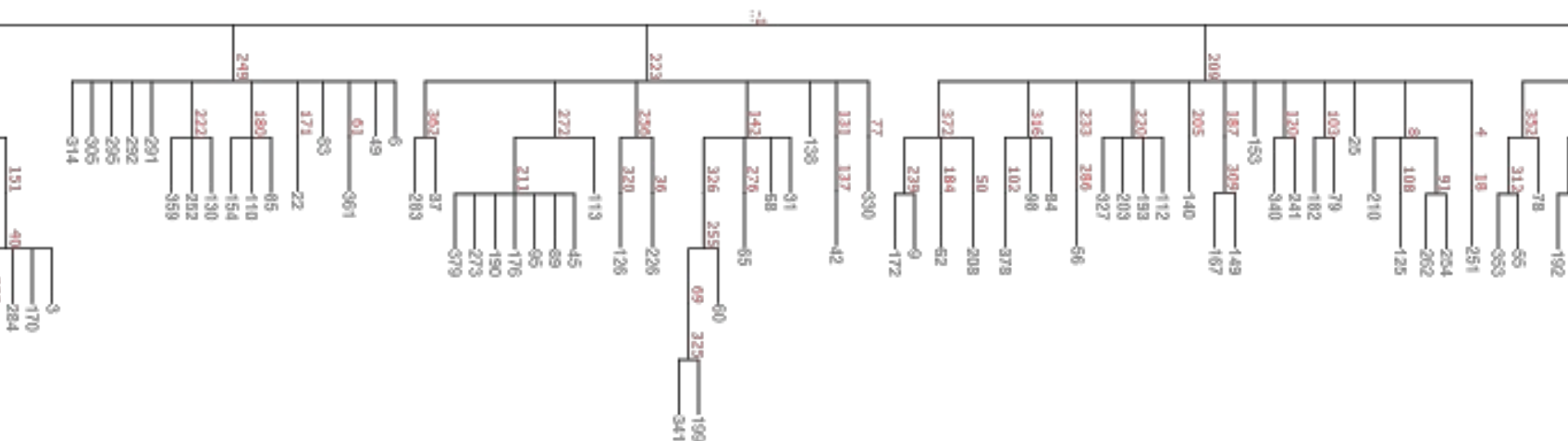
-

- Trying to improve ad click-through conversion rate based on past user activities
- Based on HDFS, HBase, Scikit Learn and Spark
- Our Dataset: *Yahoo Data Targeting User Modeling (4.3 GB)*
- Try out three models with different assumptions
- Build models based on past user activities and test them



## Taxonomy Tree:

- Interest taxonomy used at Yahoo
  - E.g. "Sports/Baseball", "Travel/Europe"
- 380 interest categories in total
- 130 non-leaf nodes and 250 leaves nodes



For each user

- 13346 Features calculated from events
  - E.g. page views, search queries, search result clicks...etc
  - Such as recency and intensity
- 380 Label for each interest category
  - -1 for seeing an ad but didn't click
  - 1 for seeing and clicking the ad
  - 1 and -1 propagate upward till root
  - 1 will overwrite/block -1 along the path

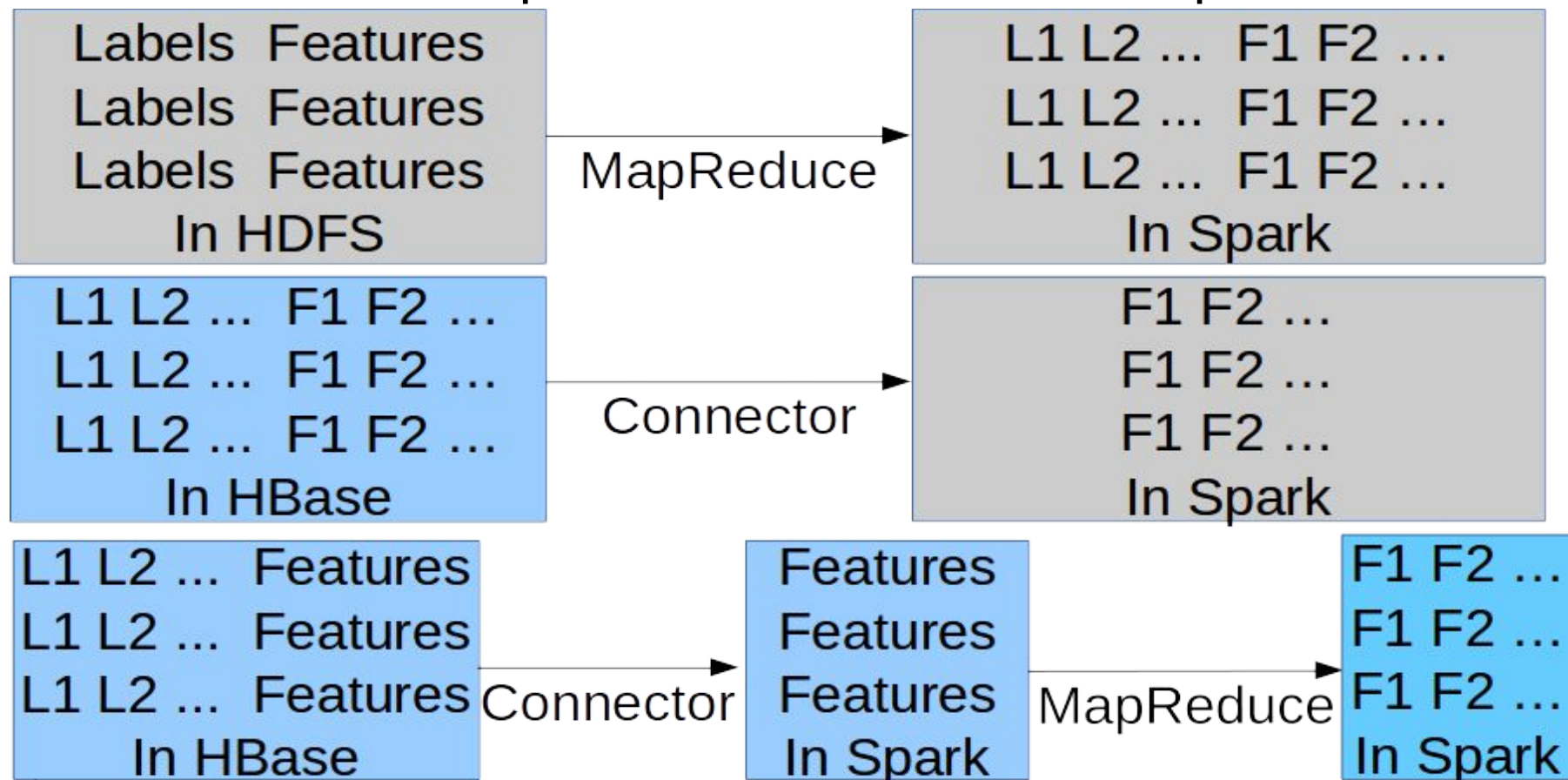
```
83:0.84294 967:68.63747 1106:5 1133:0.86006 1237:0.49984 1527:0.58704 1535:6 12966:0.41712
<tab> 32:-1 45:-1 51:-1 57:-1 198:-1 209:-1 211:-1 223:1 263:-1 268:-1 272:1 279:1 280:-1 290:-1
298:-1
313:6:0.50999 10837:5.33449 10886:16.8626 10911:0.57517 10945:0.41295 10967:47 <tab> 10:-1
17:-1 236:1 245:-1 248:-1 253:-1 270:-1 279:1 281:1 293:-1 316:-1 336:-1 350:1 370:-1 372:-1 373:-1
380:-1
```



# Challenges and Struggles

## Data Preparation:

- Anonymous Data
- Sparse feature
- Load and select sparse feature in HBase & Spark



- PCA with  $K = 100$  to reduce dimension of data
- One-to-All linear SVM for every leaf
- For branch nodes generate prediction using result of leaves
- Error Rate: 27.31%
- Error Rate of leaves: 36.85%
- Error Rate of top level node: 22.94%



- MultinomialNB - naive Bayes algorithm for multinomial data
- Smoothed maximum likelihood estimates the probability of a user with label  $y = +1$  having label  $x = +1$ :

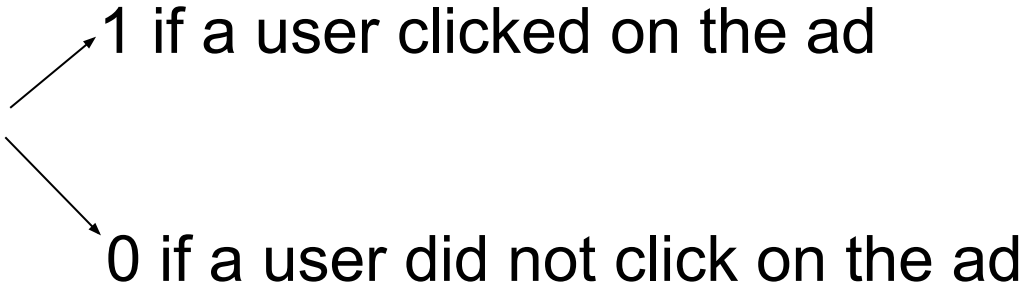
$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

- Laplace smoothing ( $\alpha = 1$ ) prevents zero probabilities
- Cross validation: 5-fold on training data
- Average error rate on testing data: 0.253
  - Significantly better than random guessing

## Input:

- Label ID(e.g.interest areas)
- User's average feature value (e.g. to measure users' engagement)

**Output: Binary Variable**



```
graph LR; A[Output: Binary Variable] --> B[1 if a user clicked on the ad]; A --> C[0 if a user did not click on the ad]
```

**Assumption:** only look at the value of lowest level that each user can reach; therefore, the labels should be independent

# Logistic Regression

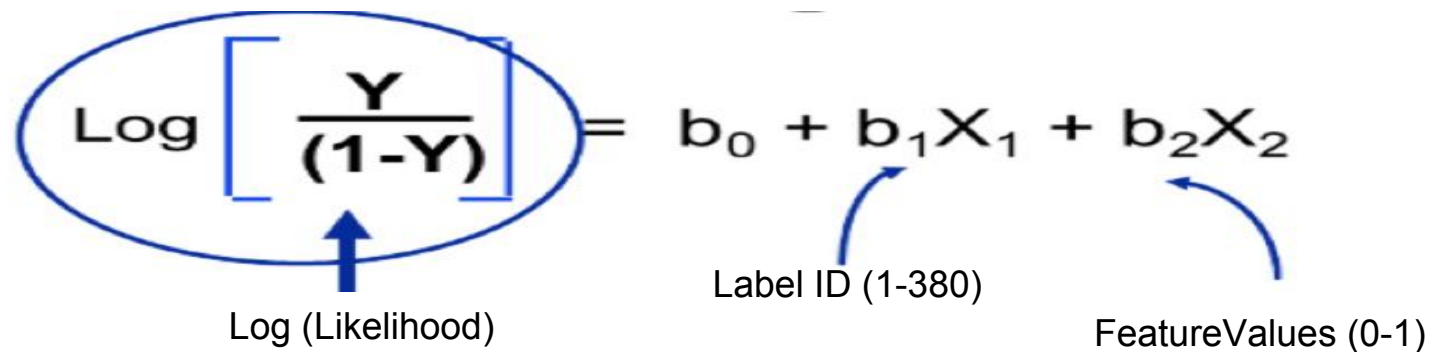
**Purpose:** given a user's interest area and feature values, predict the chance that the user will click on the ad related to this interest.

**Algorithm:**

$$\text{Log} \left[ \frac{Y}{(1-Y)} \right] = b_0 + b_1 X_1 + b_2 X_2$$

↑
Label ID (1-380)
FeatureValues (0-1)

Log (Likelihood)



**Average Error rate on testing data: 13.3%**

- The attitude of a user towards a specific interest category (represented by labels) can be predicted using other categories with high precision.
- Use the logistic regression model to improve the click-through conversion rate for advertising by targeting users who more potentially click on an ad.

# E6893 Big Data Analytics:

## *User Behavior Modeling*

Team Members:

Avery Wu (yw2928)

Tsung-Yi Huang (th2668)

Minghong Zheng (mz2597)



December 15, 2016