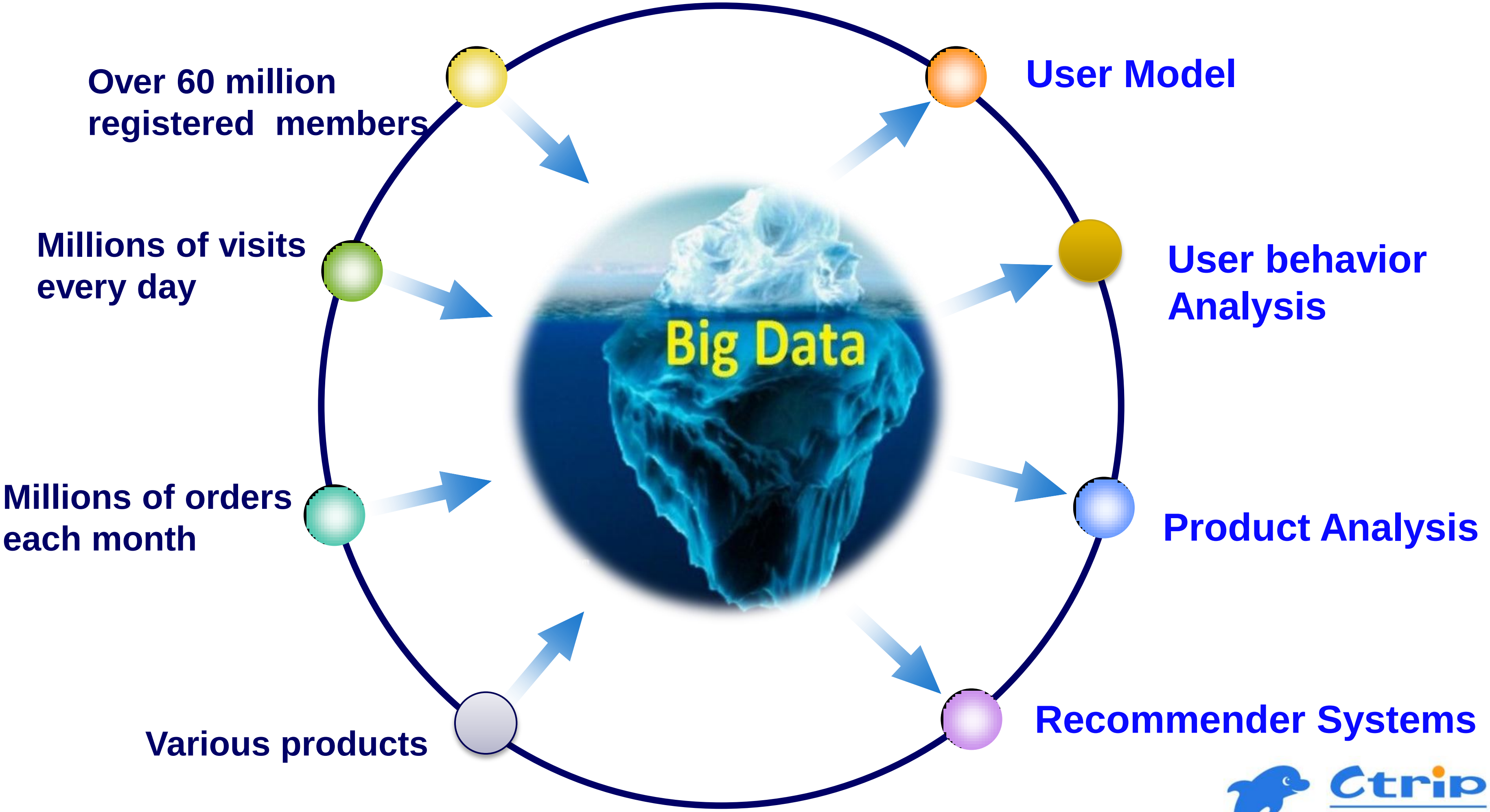
R在旅游行业中的应用

甘华来
商业智能部

Big Data in Ctrip



Over 60 million registered members

Millions of visits every day

Millions of orders each month

Various products

User Model

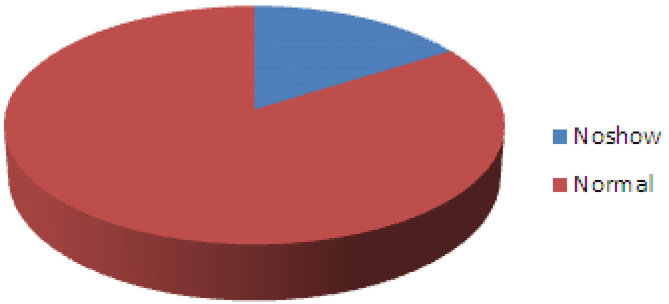
User behavior Analysis

Product Analysis

Recommender Systems

Hotel Reservations

- Millions of orders each month, some of them are noshow orders

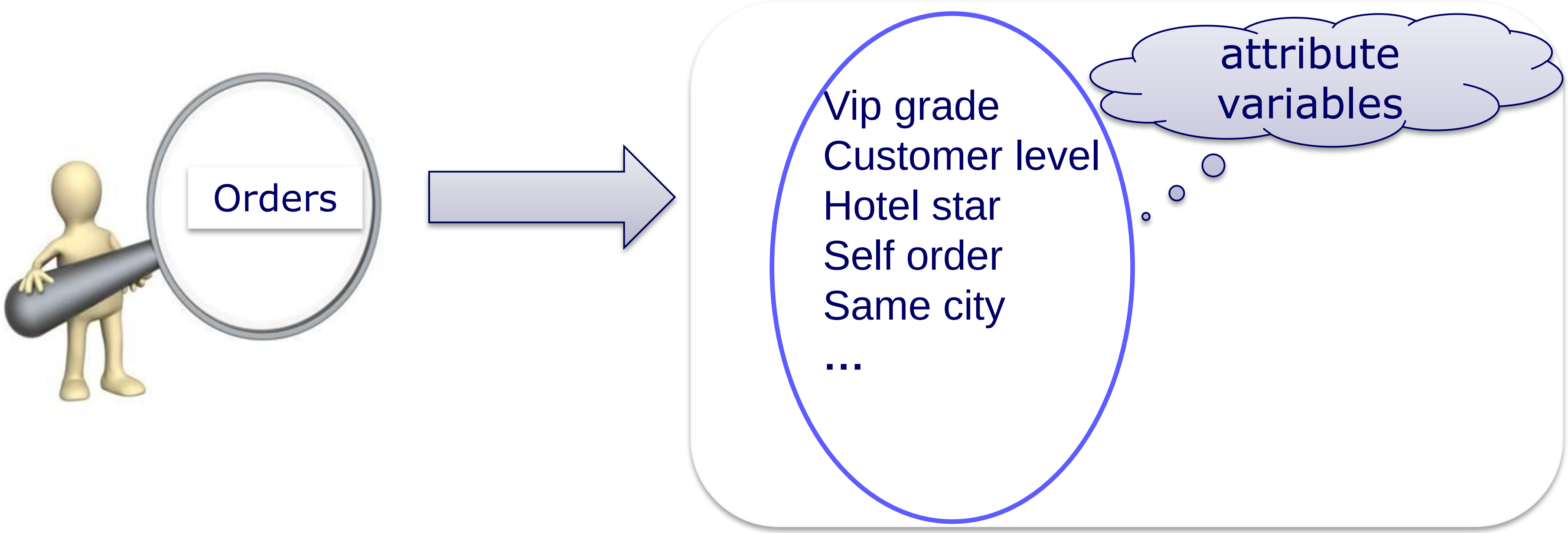


- Noshow orders need human review (hundreds of staffs do this work)

Can predict noshow orders before the manual review???

This is a very interesting case of Machine Learning.

Data Features



Orders

Vip grade
Customer level
Hotel star
Self order
Same city
…

attribute variables

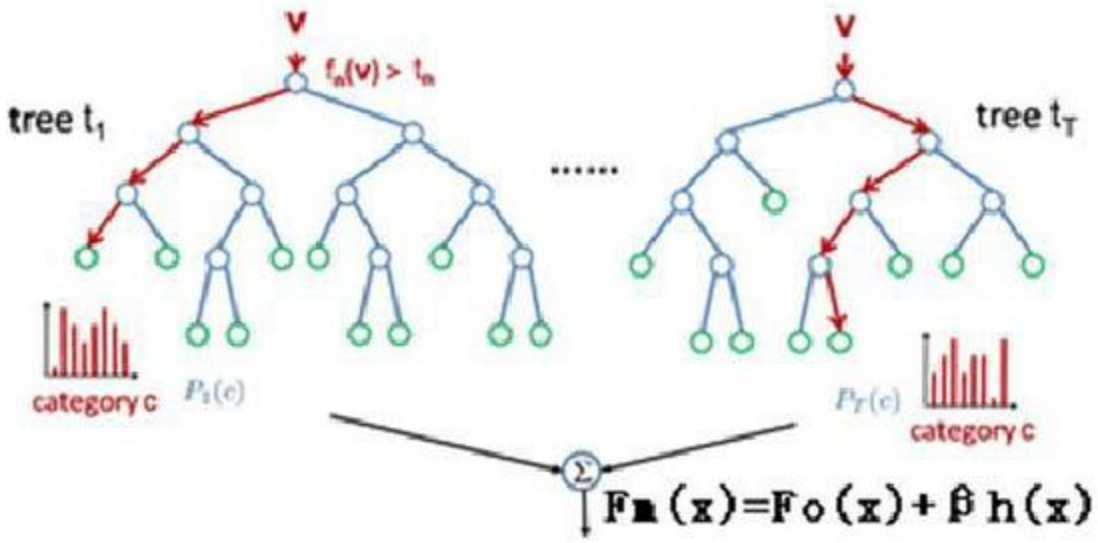
GBM(Gradient Boosted Models) :

- •One of the most widely used learning algorithms in machine learning today
- •It is adaptable, easy to interpret, and produces highly accurate models
- •Successful applied in Yahoo/eBay/Amazon/Linkedin…

GBM

GBM: Gradient Boosted Models, was invented by Jerome H. Friedman in 1999.

GBDT (Gradient Boosted Decision Trees); **GBRT** (Gradient Boosted Regression Trees)
MART (Multiple Additive Regression Trees); **TreeNet/Treelink**



GBM Package

```
> library(gbm)
> model <- gbm(formula = formula(data), shrinkage=0.01, bag.fraction = 0.5
+                   distribution='bernoulli',cv.folds=5,n.trees=3000,
+                   interaction.depth=3,verbose=F)
```

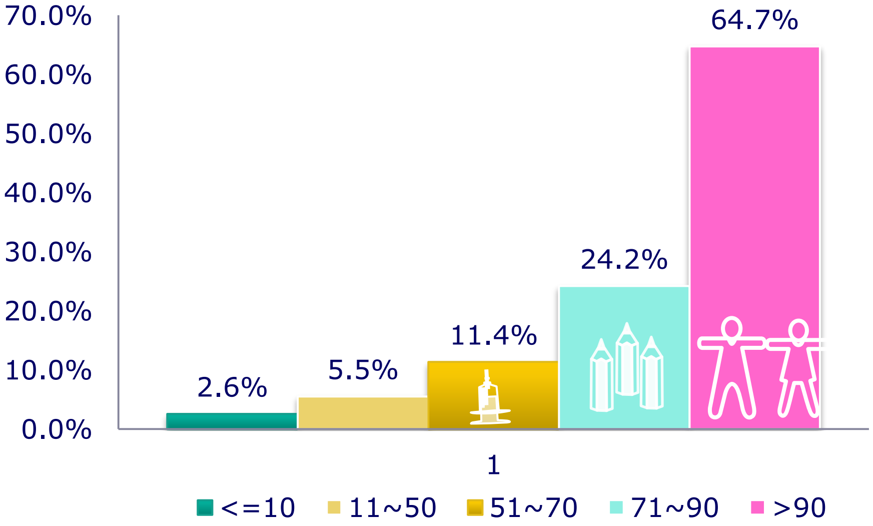
Arguments

shrinkage：the learning rate or step-size reduction
distribution：the form of loss function
cv.folds：number of cross-validation folds to perform
n.trees：the total number of trees to fit
bag.fraction：subsampling fraction, 0.5 is probably best
interaction.depth：the maximum depth of variable interactions

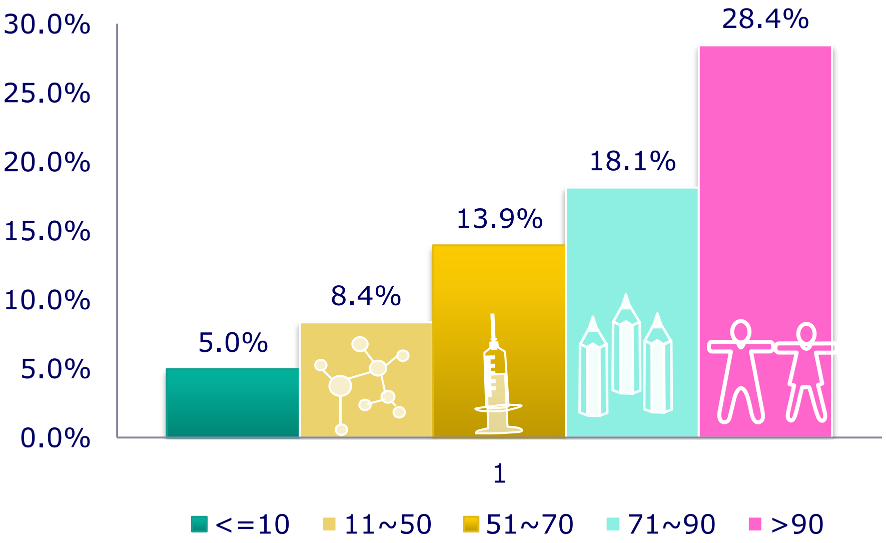
```
print(pretty.gbm.tree(gbm1,1))   ##  compactly print the first tree
```

GBM vs. Regression

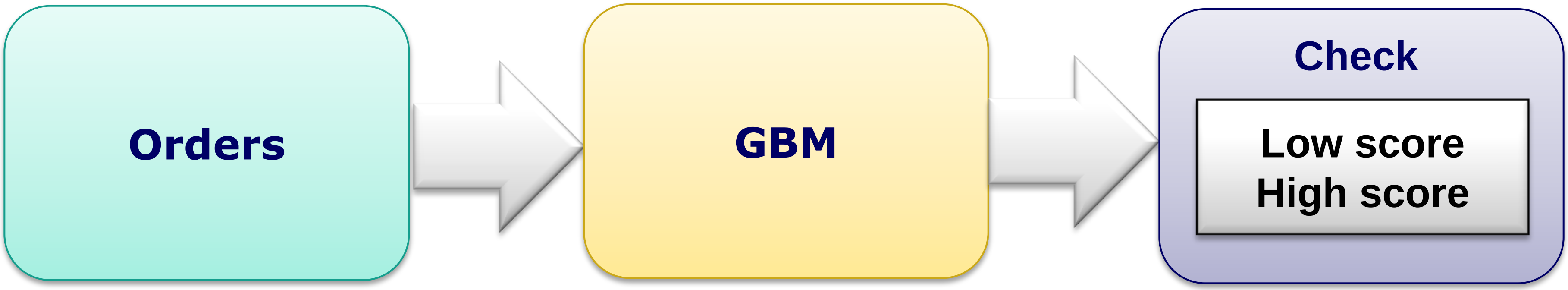
1. GBM



2. Logistic Regression



Ctrip 携程



Orders

GBM

Check

Low score
High score

Thank you!