



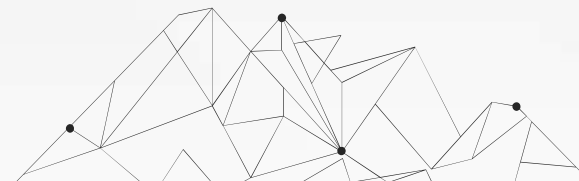
TalkingData 全球算法大赛 盘点

路瑶 数据科学部



What is the...

- **Power of Data Science**
- **Mystery of Data Scientist**
- **Magic of Data Scientist**
- **The Arena of Data Scientist**

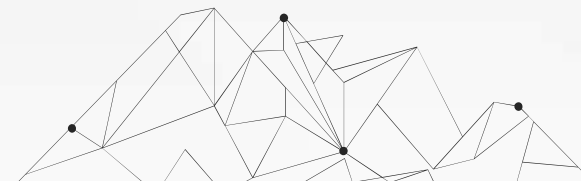




The Arena of Data Scientist

Kaggle: leading platform for crowdsourcing data challenges.

- A community to build the best solution on problems posed by industry, government and academia.
- Over 1,200 data science challenges.
- More than 600,000 registered users over 194 countries from around the world, from a wide variety of educational backgrounds and are often experts in their fields.
- Platform of KDD CUP





Famous competition and scientist of Kaggle

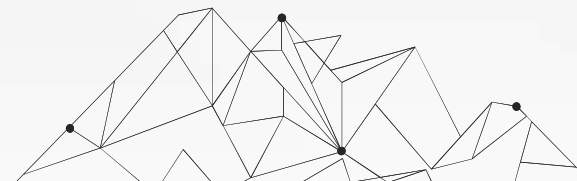
Kaggle Drug Discovery Competition, 2012

The screenshot shows the Merck Molecular Activity Challenge page. At the top left is the Merck logo with the tagline "Be well". To the right, it says "Completed • \$40,000" and "Merck Molecular Activity Challenge". Below this, the dates "Thu 16 Aug 2012 – Tue 16 Oct 2012 (3 years ago)" are listed. A sidebar on the left contains a "Dashboard" section with links to "Home" and "Data", an "Information" section with links to "Description", "Evaluation", "Rules", "Prizes", "Submission Instructions", "Visualization Prospect", and "Winners", a "Forum" section, a "Leaderboard" section with "Public" and "Private" links, and a "Visualization" section. The main content area has the heading "Help develop safe and effective medicines by predicting molecular activity." followed by the text "Help enable the development of safe, effective medicines." and a paragraph explaining the competition's goal: "When [developing new medicines](#) it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. The objective of this competition is to identify the best statistical techniques for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures." It also mentions that the challenge is based on 15 molecular activity data sets and that Merck is hosting a prediction competition.

Geoffrey Hinton



Deep
Learning





TalkingData host the competition in Kaggle for

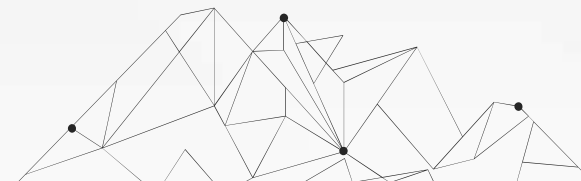
- Provide open data and open platform for
- Smartest scientist and smartest methodology
- To solve most challenging topics

Data Science Summit
 2016 Presented by Turi 

TalkingData Competition was announced in Data Science Summit in San Francisco on July 13th

Acknowledge Turi

- A famous Machine Learning company who created GraphLab.
- Acquired by Apple Inc. in August





TalkingData Mobile User Demographics

Goal of the competition:

- To know your users' profile by
- Learning from their behavior.

Given

- Application usage and trace with time stamps
- Mobile brand and device mode

To optimize

- The estimation of their age and group

Evaluated by

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}),$$

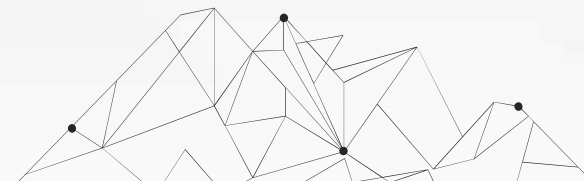
The screenshot shows the TalkingData Mobile User Demographics competition interface. At the top, it states 'Completed • \$25,000 • 1,689 teams' and 'TalkingData Mobile User Demographics' with the dates 'Mon 11 Jul 2016 – Mon 5 Sep 2016 (yesterday)'. The left sidebar contains a 'Dashboard' menu with links to Home, Data, Make a submission, Information (Description, Evaluation, Rules, Prizes, GraphLab Create License, Timeline), Forum, Kernels (New Script, New Notebook), and Leaderboard (Public, Private). The main content area has a header with 'Competition Details', 'Get the Data', and 'Make a submission'. Below this is a section titled 'Get to know millions of mobile device users' with a paragraph about the importance of understanding user behavior and a bar chart showing user demographics. The text mentions that TalkingData is seeking to leverage behavioral data from more than 70% of the 500 million mobile devices active daily in China.



Participation of the competition

1689 teams, 1961 players submitted, 24,629 entries, 2729 kernels

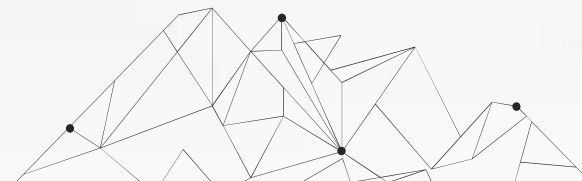
- **Largest** competition hosted by a Chinese company.
- Among the **highest** number of Kernels of all.
- Great proportion of **top** kagglers participated.
- **70+** countries and regions represented by the participant pool.





What is the...

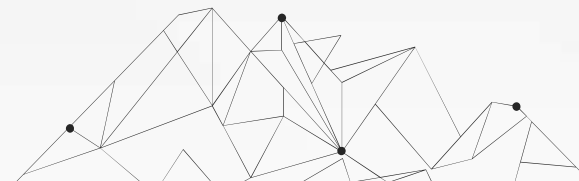
- **Power of Data Science**
- **Mystery of Data Scientist**
- **Magic of Data Scientist**
- **The Arena of Data Scientist**





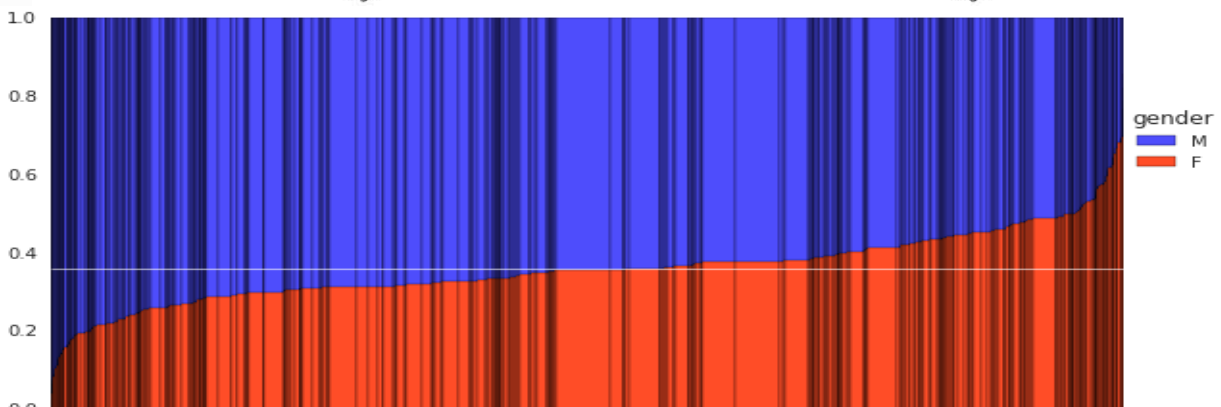
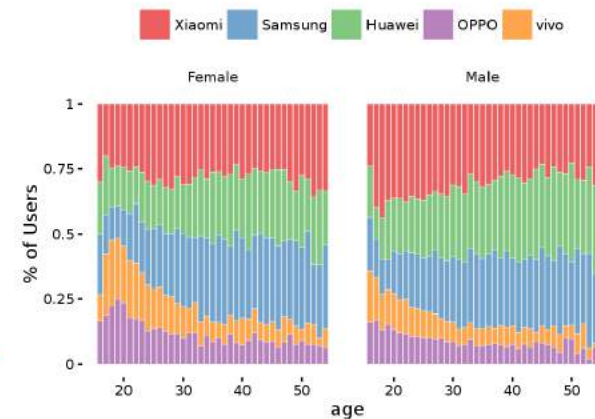
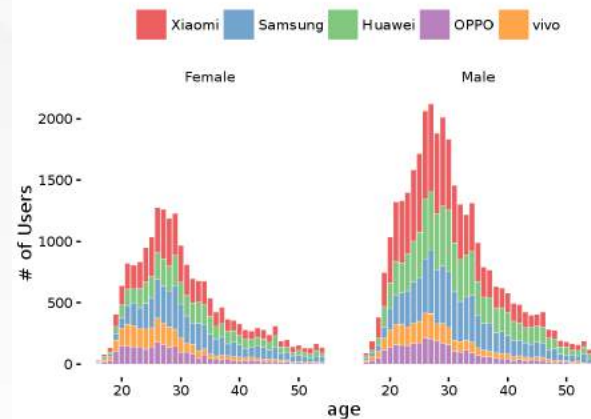
Magic of Data Scientist

- **Interpret the raw data**
- **Execute feature engineering**
- **Fine tune individual and ensemble models**
- **Avoid overfitting**
- **Fetch the best tools**



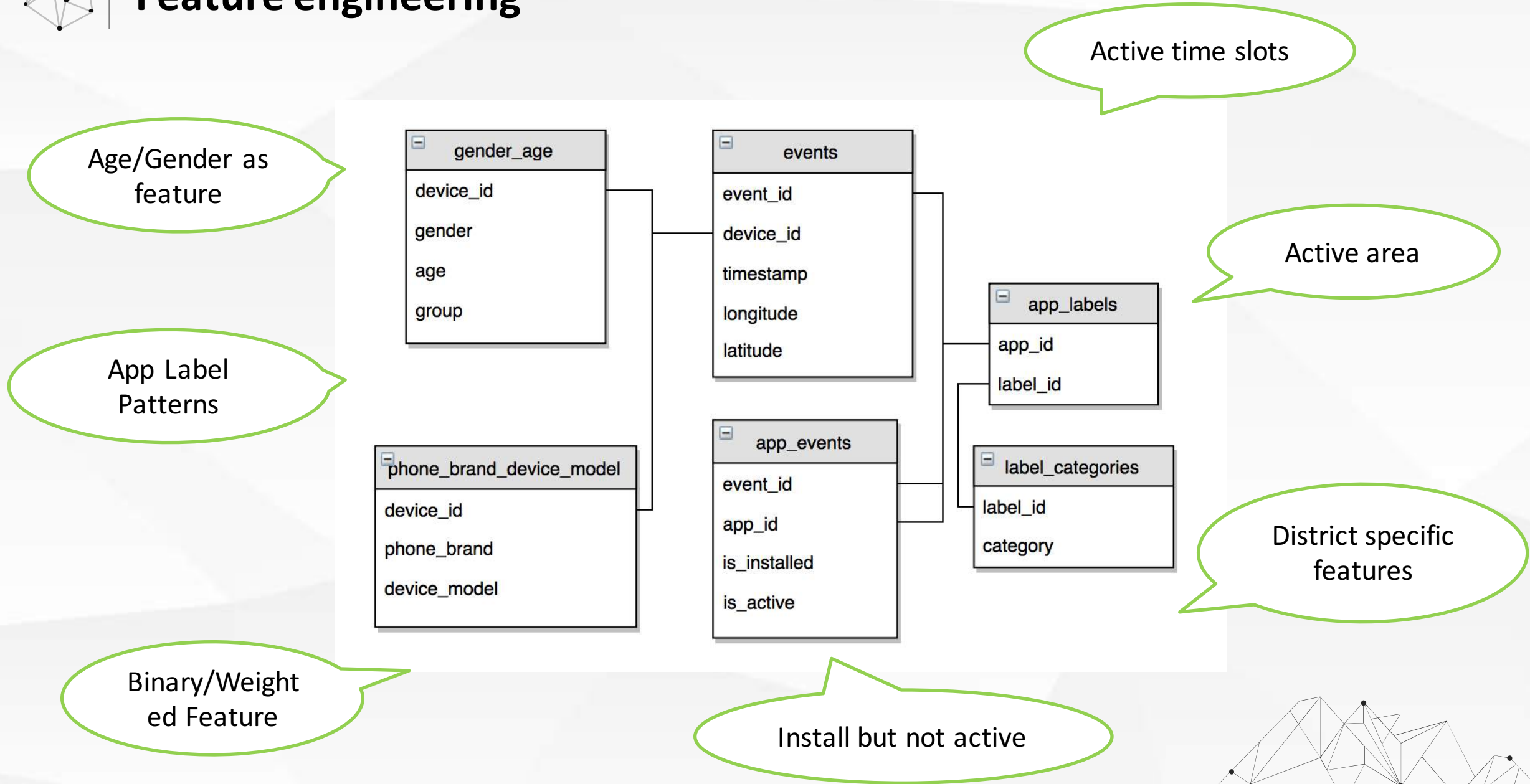


Interpret your data





Feature engineering





Fine tune individual and ensemble models

Random Forest

GBDT

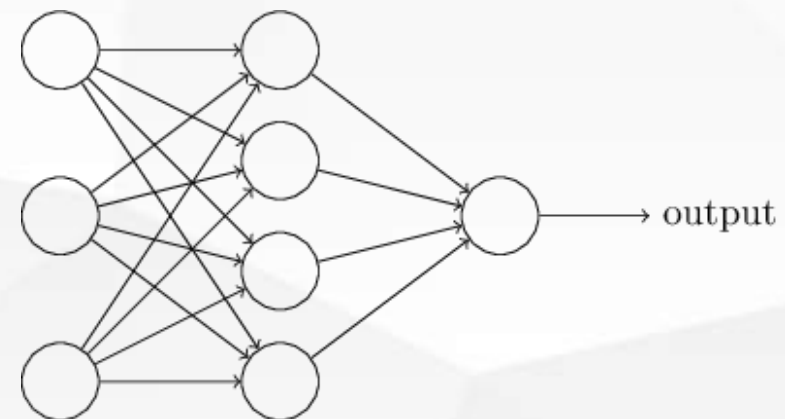
SVM

Logistic
Regression

Neural Network

Adaboost

Stack





Avoid Overfitting

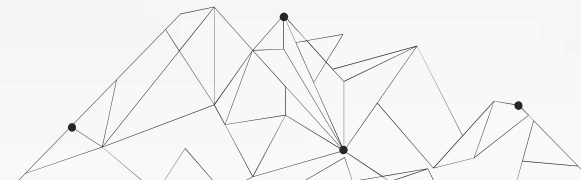
Train Set

Test Set

Always do
Cross Validation

Public
Board

Private Board





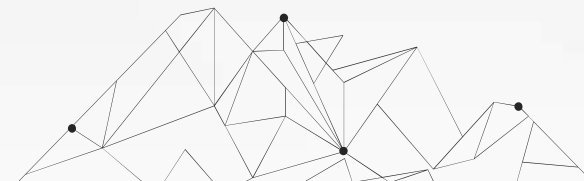
Fetch the best tools

Keras

- highly modular neural networks library, written in Python
- capable of running on top of either Tensorflow or Theano.
- allows for easy and fast prototyping
- runs seamlessly on CPU and GPU.

XGboost

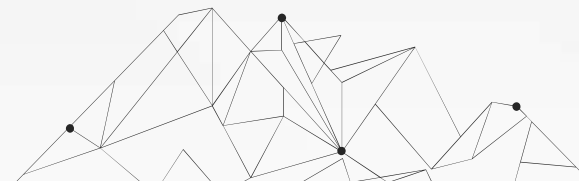
- Scalable, Portable and Distributed Gradient Boosting Library,
- Runs on single machine, Hadoop, Spark, Flink and DataFlow
- Higher precision, winner in Higgs Boson signal competition in Kaggle, 2014





What is the...

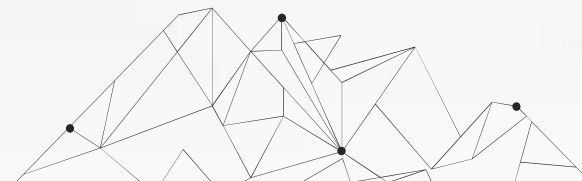
- **Power of Data Science**
- **Mystery of Data Scientist**
- **Magic of Data Scientist**
- **The Arena of Data Scientist**





How are our competitors

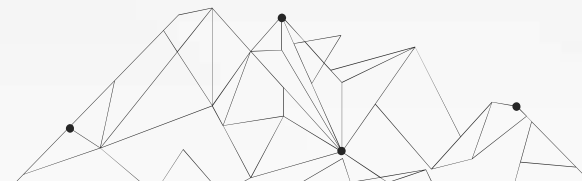
- **Open and helpful**
- **Smart and creative, willing to solve problems in reality**
- **Elegant and rational**
- **Hard working**





What we offered

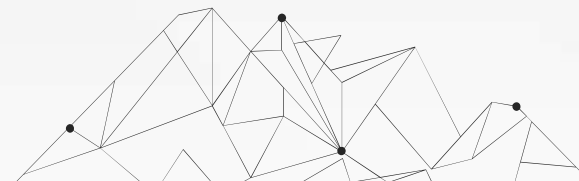
- **Industrial data and valuable business problems**
- **Honor to Data Scientists' professional skills and spirit**





What we achieved from the competition

Mind and Heart of the smartest scientist Over the World





THANKS

聘

Let's rock data together!

hr@tendcloud.com