# XGBoost

## A Library for Fast and Accurate Gradient Boosting

Tong He

Simon Fraser University

Nov. 7th, 2016

# Machine Learning Competitions

Academic:

- For conferences
- For research topics

Industrial:

- Sales prediction, user behaviour classification, etc
- Mainly on Kaggle.com

# Two Years Ago...

The most popular algorithms are:

- Generalized Linear Models
- Random Forests
- Neural Networks
- Support Vector Machines
- Gradient Boosting Machines

# Higgs Boson Competition

Completed • $13,000 • 1,785 teams

**Higgs Boson Machine Learning Challenge**

Mon 12 May 2014 – Mon 15 Sep 2014 (2 years ago)

Dashboard

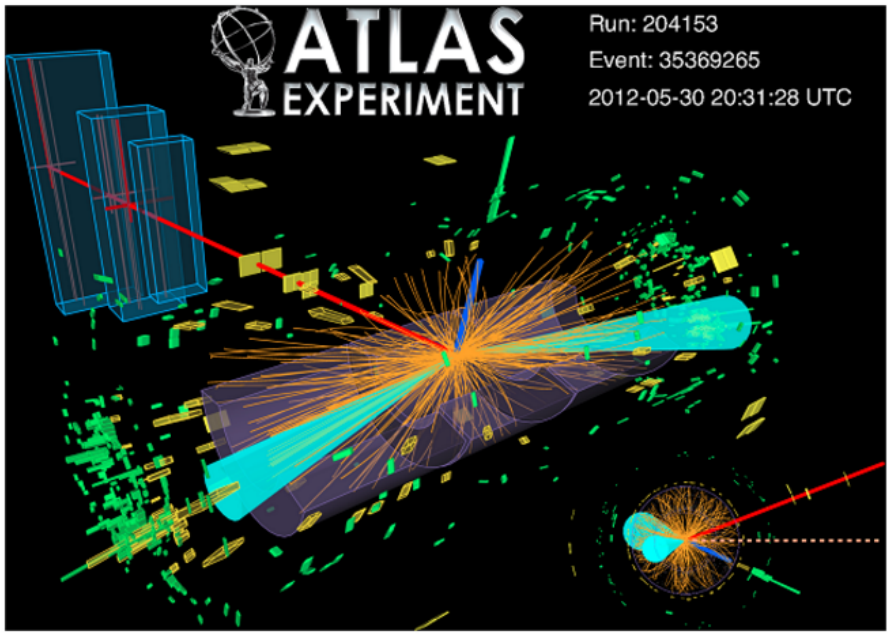
- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - About the Sponsors
  - Timeline
  - Winners
- Forum
- Leaderboard
  - Public
  - Private
- My Team
  - Your model
  - GitHub
- My Submissions

Private Leaderboard

1. Gábor Melis
2. Tim Salimans
3. nhlx5haze

Competition Details » Get the Data » Make a submission

## Use the ATLAS experiment to identify the Higgs boson



# Higgs Boson Competition

An academic competition

- complicated features
- moderate data size
- Missing values
- Imbalanced labels

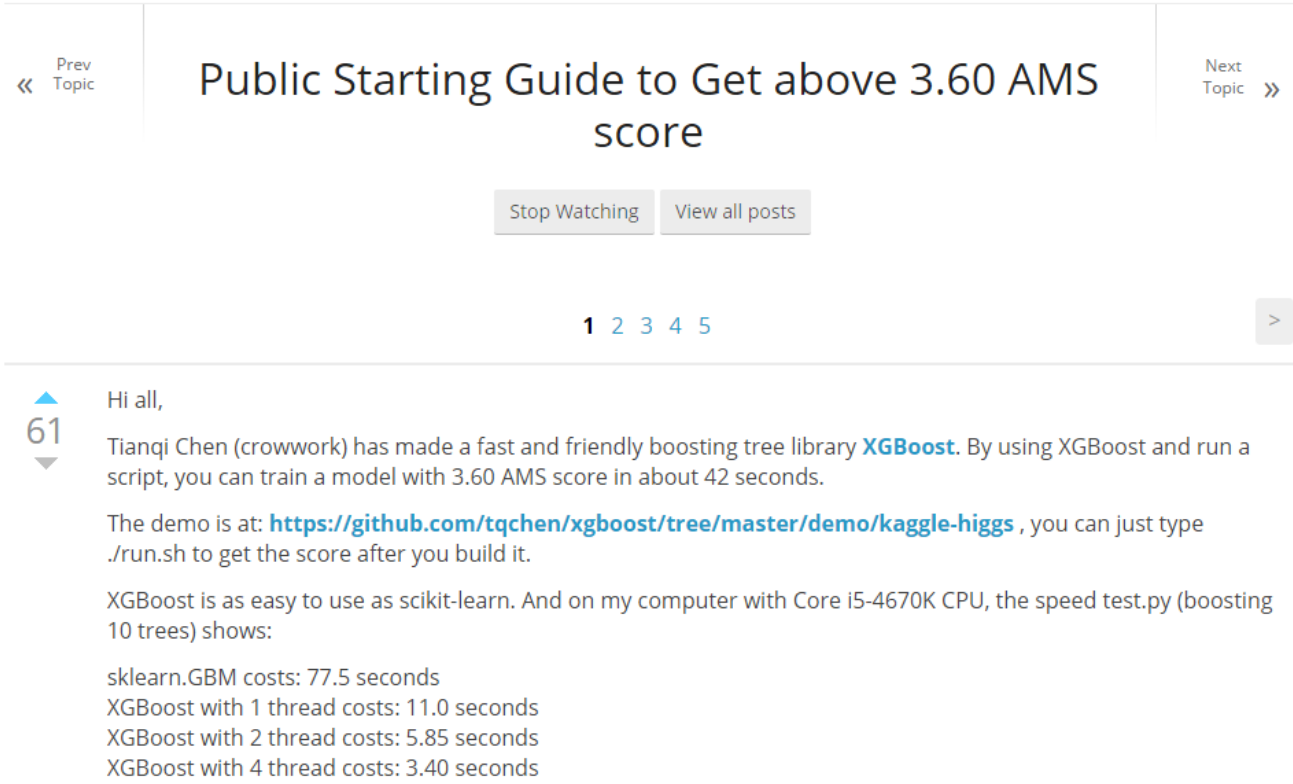
# Beat the Benchmark?

It was difficult to beat the benchmark:

| | | | | | |
|---|---|---|---|---|---|
| 899 | ↓123 | dtrx | 3.41081 | 45 | Thu, 28 Aug 2014 13:28:45 (-38.2d) |
| 900 | ↑10 | tudu | 3.40918 | 1 | Mon, 25 Aug 2014 08:18:53 |
| 901 | ↓73 | gyong | 3.40909 | 5 | Thu, 14 Aug 2014 22:12:23 |
| 902 | ↓47 | stay crunchy | 3.40638 | 4 | Wed, 30 Jul 2014 21:40:45 (-2.1d) |
| | | **MultiBoost** | 3.40488 | | |
| 903 | ↓102 | sweezyjeezy | 3.40191 | 7 | Mon, 01 Sep 2014 16:41:47 (-4.3d) |
| 904 | ↑8 | asfandyar | 3.40015 | 4 | Tue, 03 Jun 2014 12:35:51 |
| 905 | ↓96 | hussam.ashab | 3.39927 | 31 | Sat, 16 Aug 2014 17:16:38 (-2.8d) |
| 906 | ↑16 | I prefer steak | 3.39907 | 7 | Tue, 20 May 2014 18:23:10 |
| 907 | ↓10 | Nikolas A. | 3.39881 | 14 | Mon, 15 Sep 2014 23:52:15 |

# Debut

Then, someone posted in the forum...



« Prev Topic

Public Starting Guide to Get above 3.60 AMS score

Next Topic »

Stop Watching | View all posts

1 2 3 4 5 >

61

Hi all,

Tianqi Chen (crowwork) has made a fast and friendly boosting tree library **XGBoost**. By using XGBoost and run a script, you can train a model with 3.60 AMS score in about 42 seconds.

The demo is at: **https://github.com/tqchen/xgboost/tree/master/demo/kaggle-higgs** , you can just type ./run.sh to get the score after you build it.

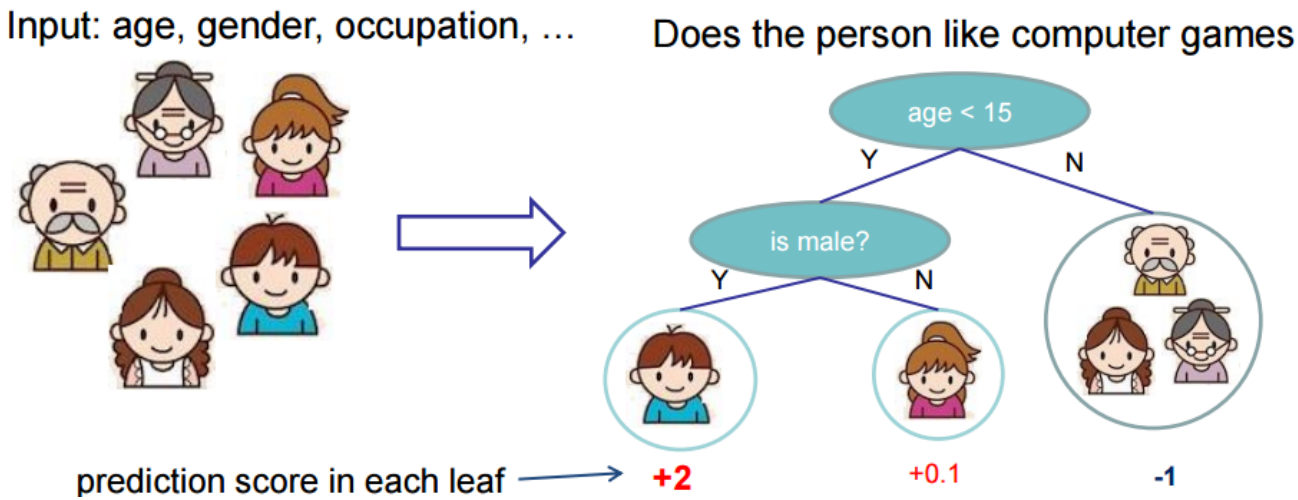
XGBoost is as easy to use as scikit-learn. And on my computer with Core i5-4670K CPU, the speed test.py (boosting 10 trees) shows:

sklearn.GBM costs: 77.5 seconds
XGBoost with 1 thread costs: 11.0 seconds
XGBoost with 2 thread costs: 5.85 seconds
XGBoost with 4 thread costs: 3.40 seconds

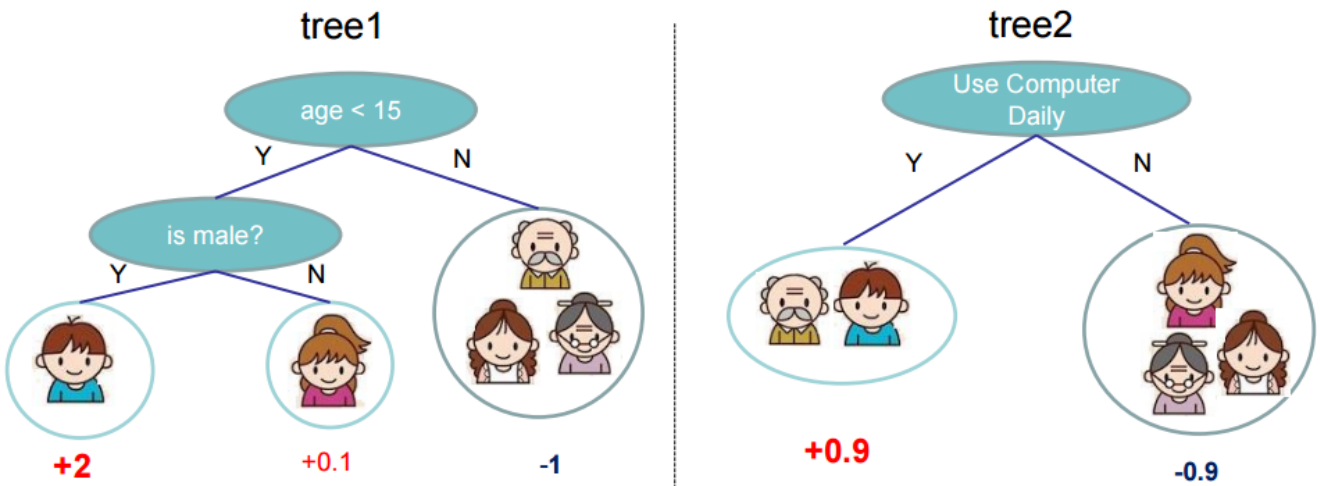
# Debut


Higgs challenge
Higgs Boson Machine Learning Challenge
3 months to go
Monday, May 12, 2014
$13,000 • 182 teams
Monday, September 15, 2014
Dashboard
Leaderboard - Higgs Boson Machine Learning Challenge
This leaderboard is calculated on approximately 18% of the test data.
The final results will be based on the other 82%, so the final standings may be different.
See someone using multiple accounts?
Let us know.
# Δ1d Team Name Score Entries Last Submission UTC
1 - Triskelion & Abhishek * 3.67962 45 Mon, 19 May 2014 21:37:03 (-45.3h)
2 new Dominado * 3.65823 5 Mon, 19 May 2014 20:05:03 (-0.6h)
3 ↓1 Terry Chen * 3.63731 9 Mon, 19 May 2014 19:07:34 (-22.9h)
4 ↓1 ivanhoe 3.63520 9 Tue, 20 May 2014 05:34:35 (-24h)
5 ↑14 TomHall 3.63084 22 Tue, 20 May 2014 07:58:59
Your Best Entry
Top Ten!
You improved on your best score by 0.03081.
You just moved up 17 positions on the leaderboard.

# Tree Model


Input: age, gender, occupation, …
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

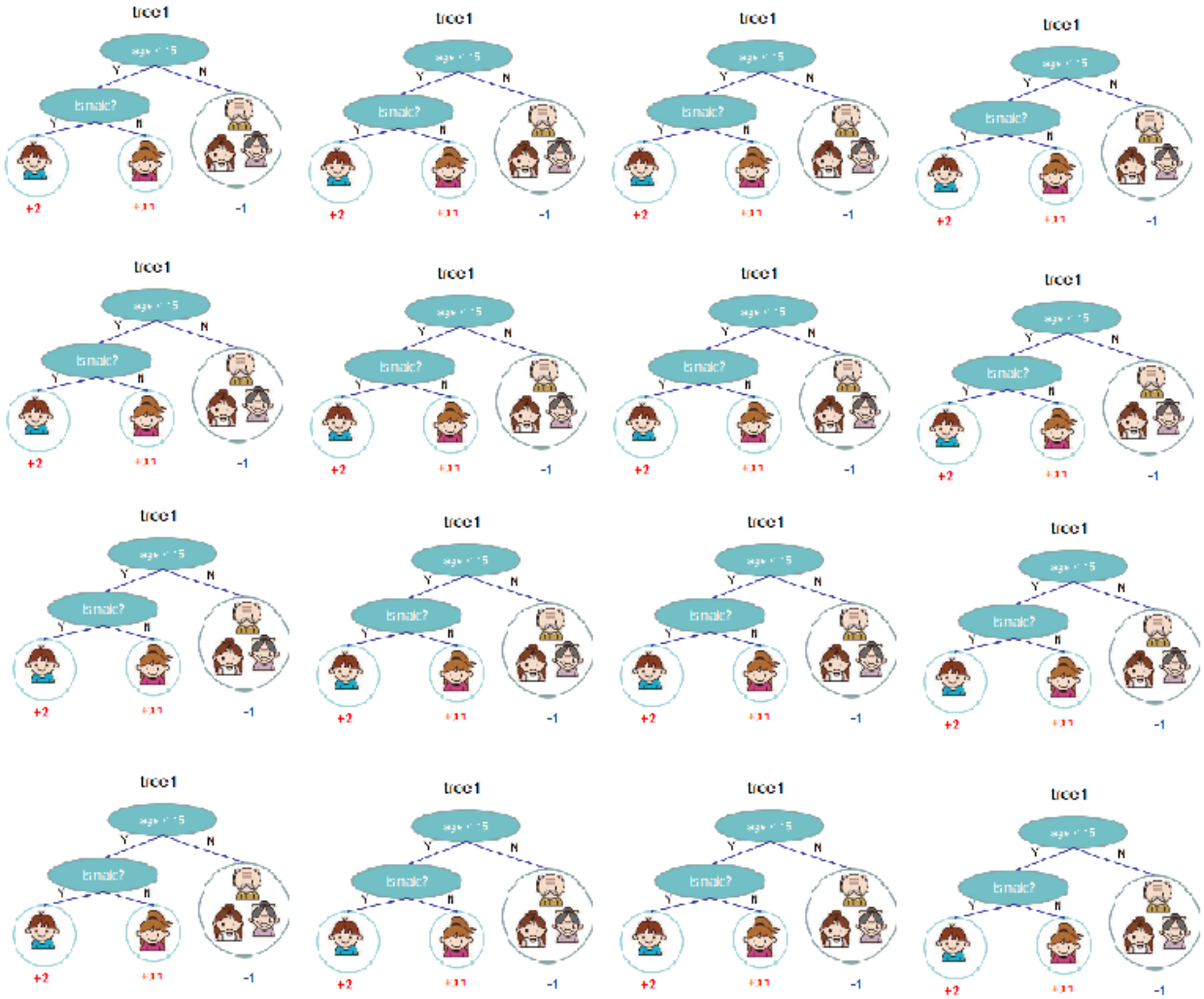
# Tree Ensemble




f( ) = 2 + 0.9= 2.9 f( )= -1 + 0.9= -0.1

Prediction of is sum of scores predicted by each of the tree

# Tree Ensemble

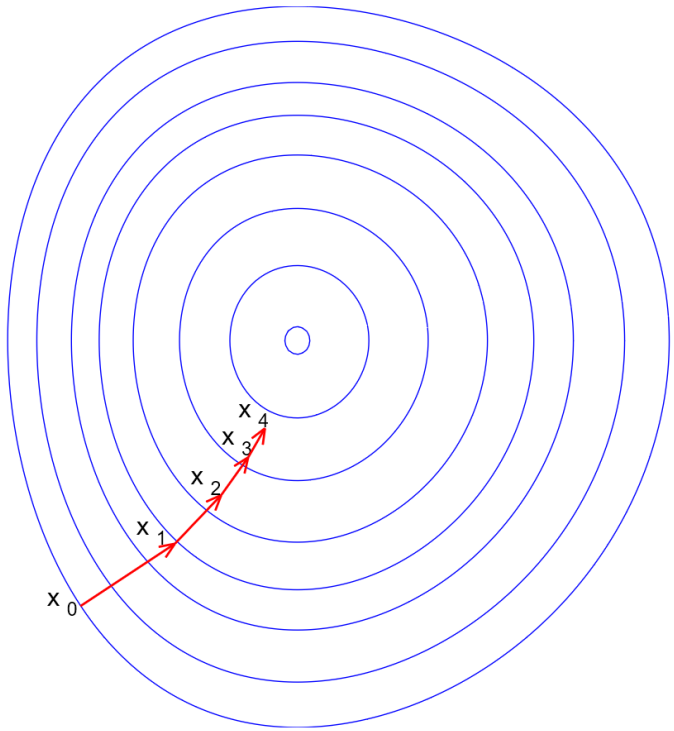


# Tree Ensemble

- Parallel: Random Forests
- Iterative: Boosting

# Gradient Boosting

- Calculate the gradient of the current ensemble
- Add a new tree: take a step along the gradient



# eXtreme Gradient Boosting

- $L_1$ and $L_2$ regularization
- Using both first and second order gradient
- Prune on a full binary tree

# Accuracy

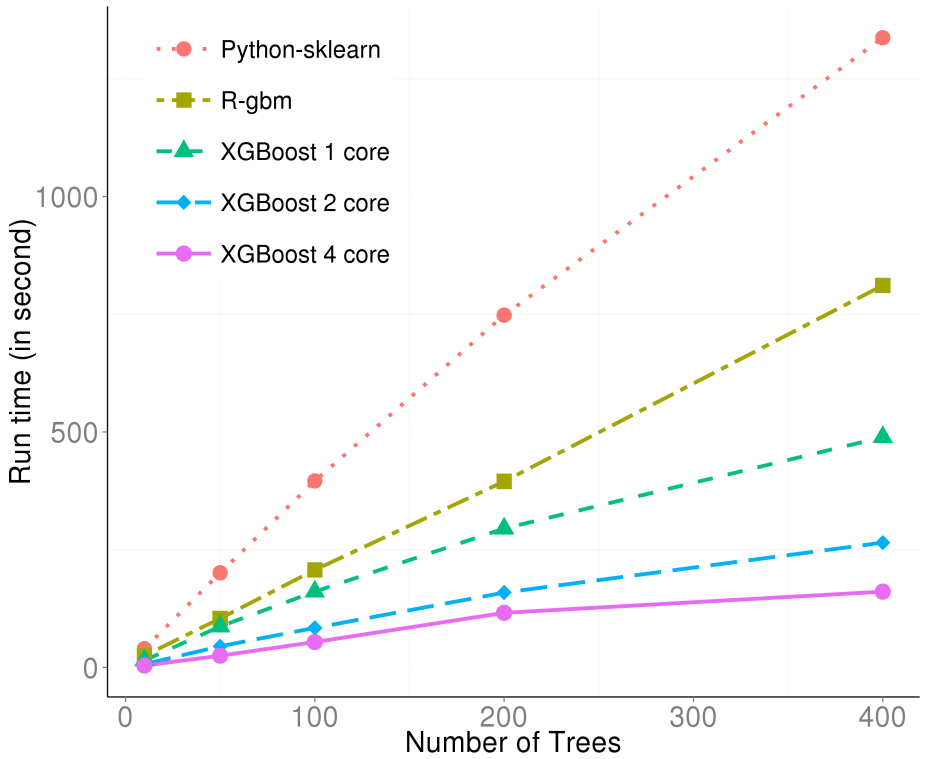
Tested on the Higgs Boson Machine Learning Competition:

| | Original Data | Physical Features |
|---|---|---|
| R-gbm | 3.38356 | 3.38422 |
| python-sklearn | 3.55236 | 3.56985 |
| *XGBoost* | 3.64655 | 3.65860 |
| *XGBoost*(tuned) | 3.71142 | 3.72370 |

Table: 5-fold crossvalidation result, in Approximate Median Significance

# Time Efficiency

- Multi-threading by OpenMP
- Internal data structure in C++
- Pre-sort the features



# Results

Through the entire competition, xgboost

- outperforms most tree-based algorithms.
- was popular among the participants.

People started to try XGBoost!

# Winning Solutions

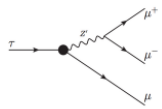
There are a number of machine learning competition winners using xgboost as a part of their solutions

CrowdFlower

Completed • $20,000 • 1,326 teams

**Crowdflower Search Results Relevance**

Mon 11 May 2015 – Mon 6 Jul 2015 (11 months ago)



Completed • $15,000 • 673 teams

**Flavours of Physics: Finding τ → μμμ**

Mon 20 Jul 2015 – Mon 12 Oct 2015 (8 months ago)

Liberty Mutual INSURANCE

Completed • $25,000 • 2,236 teams

**Liberty Mutual Group: Property Inspection Prediction**

Mon 6 Jul 2015 – Fri 28 Aug 2015 (10 months ago)

Dato

Completed • $10,000 • 274 teams

**Truly Native?**

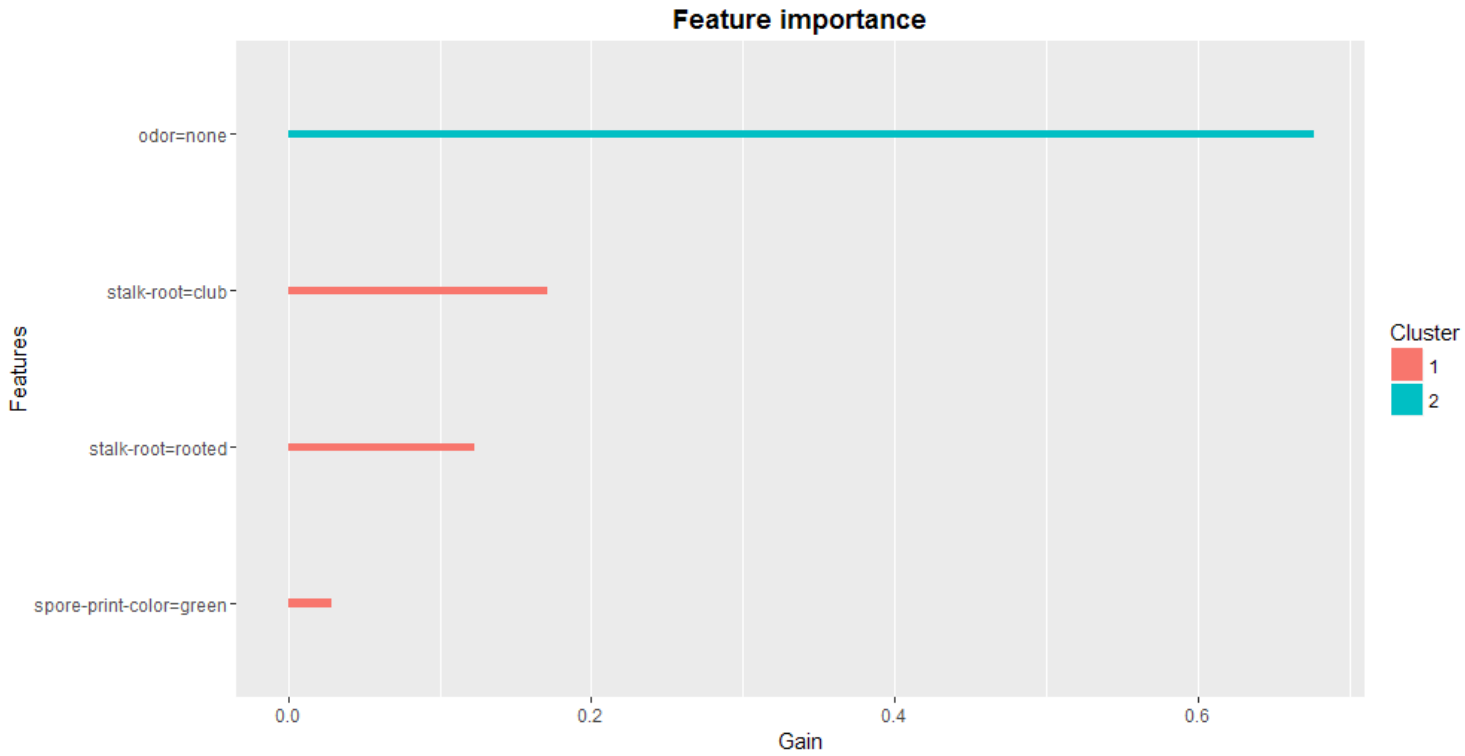
Thu 6 Aug 2015 – Wed 14 Oct 2015 (8 months ago)

and much more

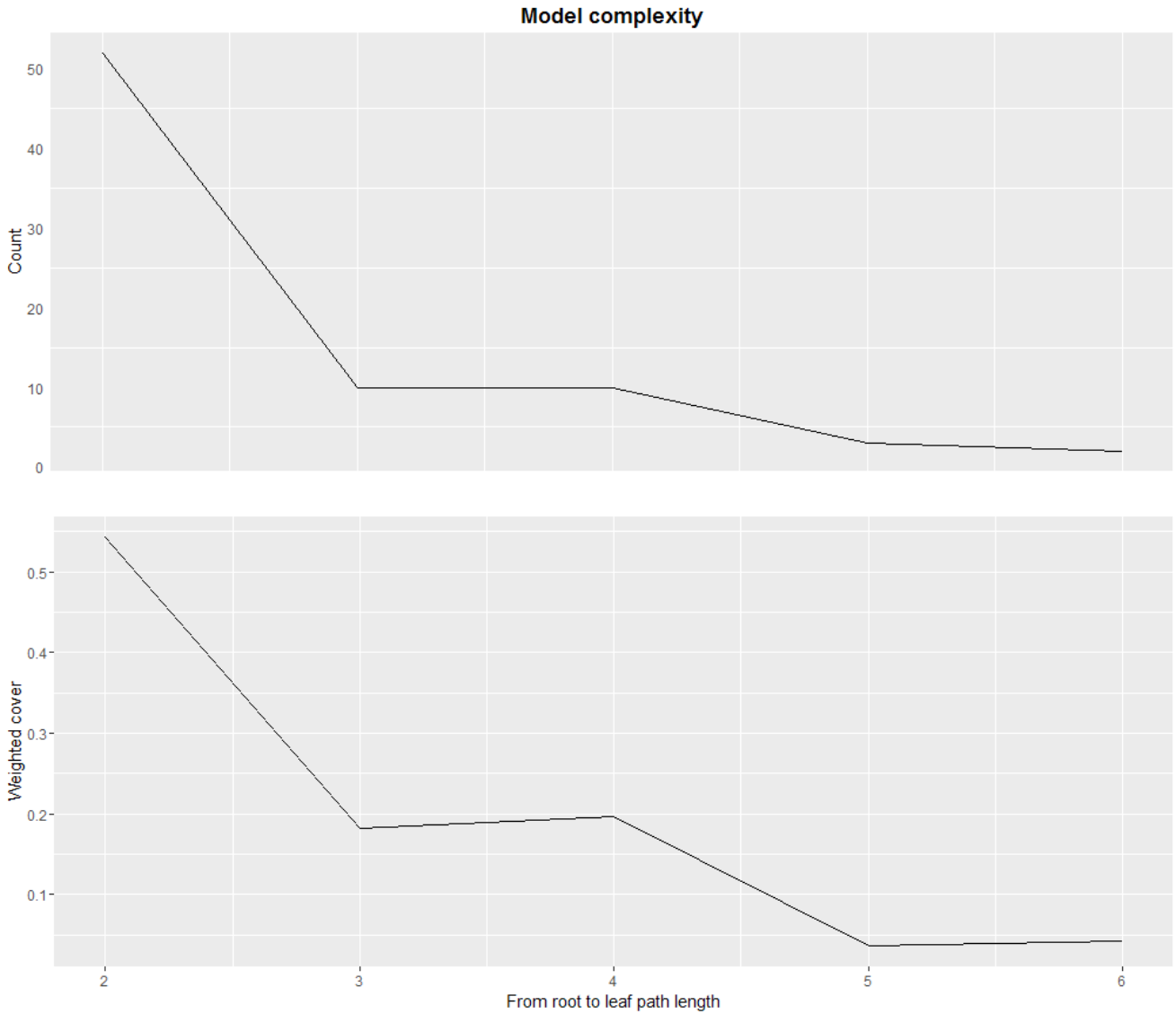
# Visualization

- Feature importance
- Number of leaves per level
  - inspired by Aysen Tatarinov
- Merge multiple tree plots into one
  - inspired by Sean Welleck

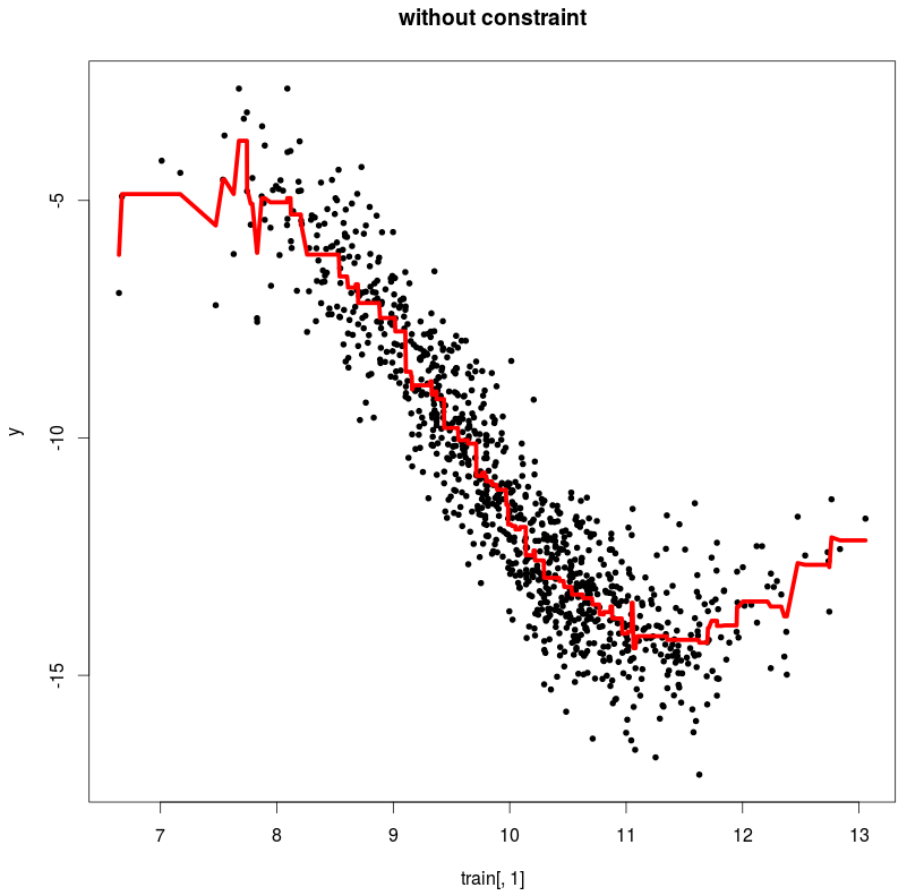
# Visualization

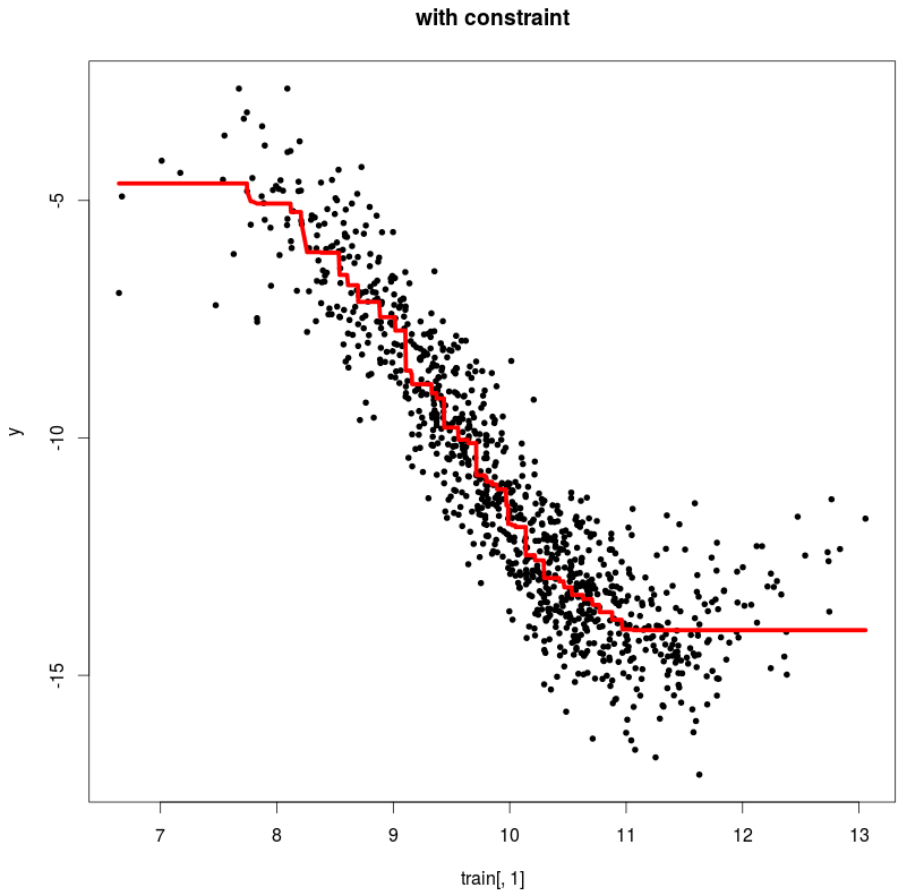


# Visualization



# Monotonic Constraints



# Features

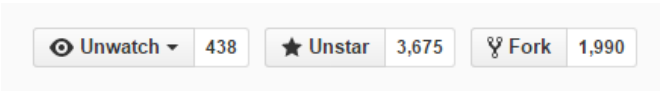
Other useful features:

- Early-stopping in training/cross-validation
- Customized loss function
- Missing values compatible
- Continue training
- Boosted linear model
- Boosted random forest
- ...

# The Project *XGBoost*

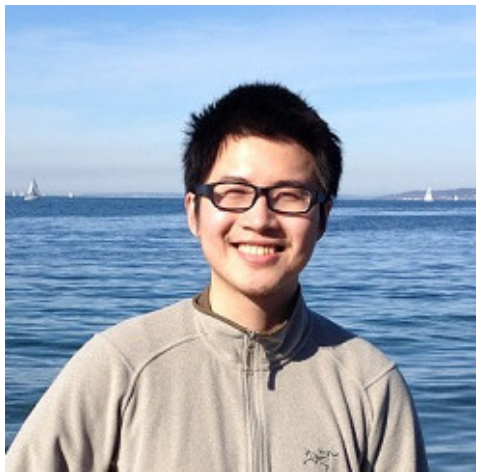
XGBoost: *eXtreme Gradient Boosting*

- on Github: dmlc/xgboost




- Written in C++
- interfaces in *R*, python, Julia, Java
- Widely applied in competitions and industrial activities

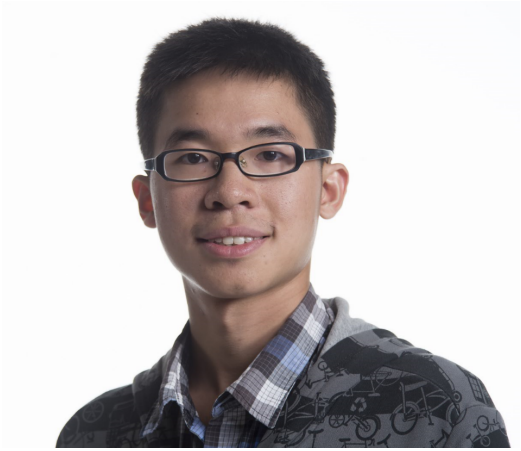
- Project Creator: *Tianqi Chen* from University of Washington



# The R package *xgboost*

- John M. Chambers Statistical Software Award in 2016
- On CRAN, around 8k downloads in the last month

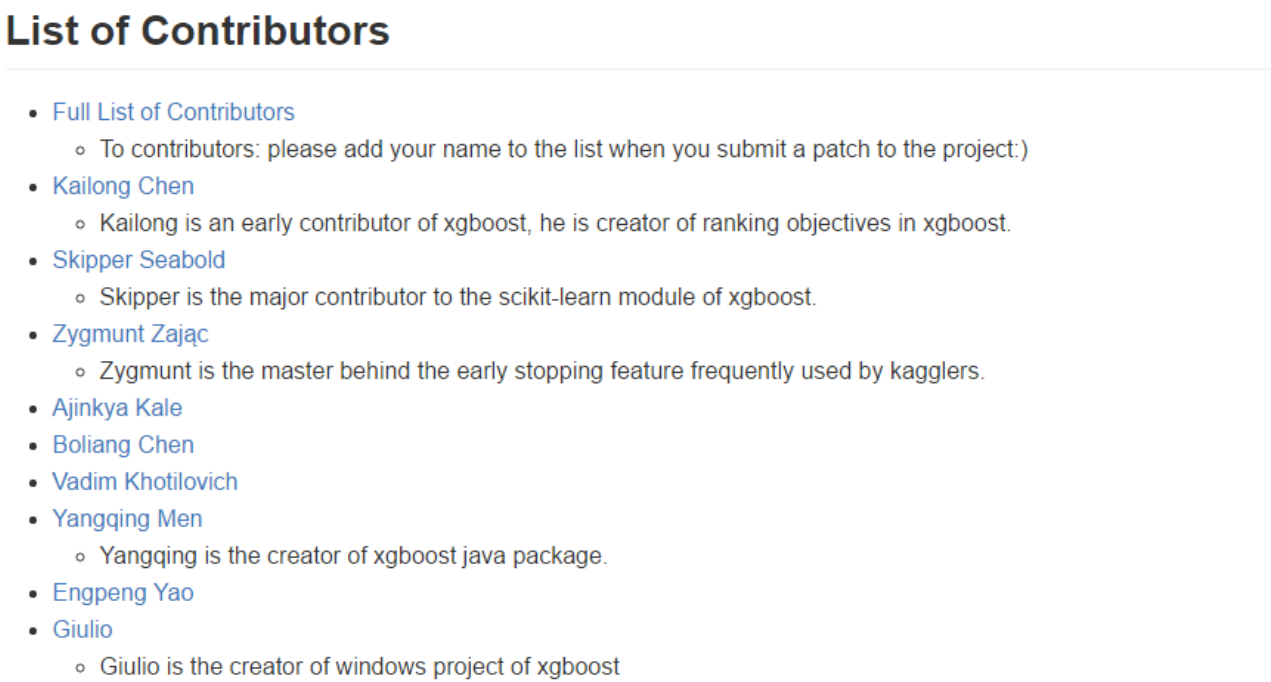

- R-package maintainer: *Tong He* from Simon Fraser University



# Welcome to Contribute

- Michael, Vadim: Contributors of the R package
- A long list of contributors all over the world!

- Pull Request
  - Contribute your code
- Issue
  - Contribute your ideas
  - Ask us a questions
  - Report a bug

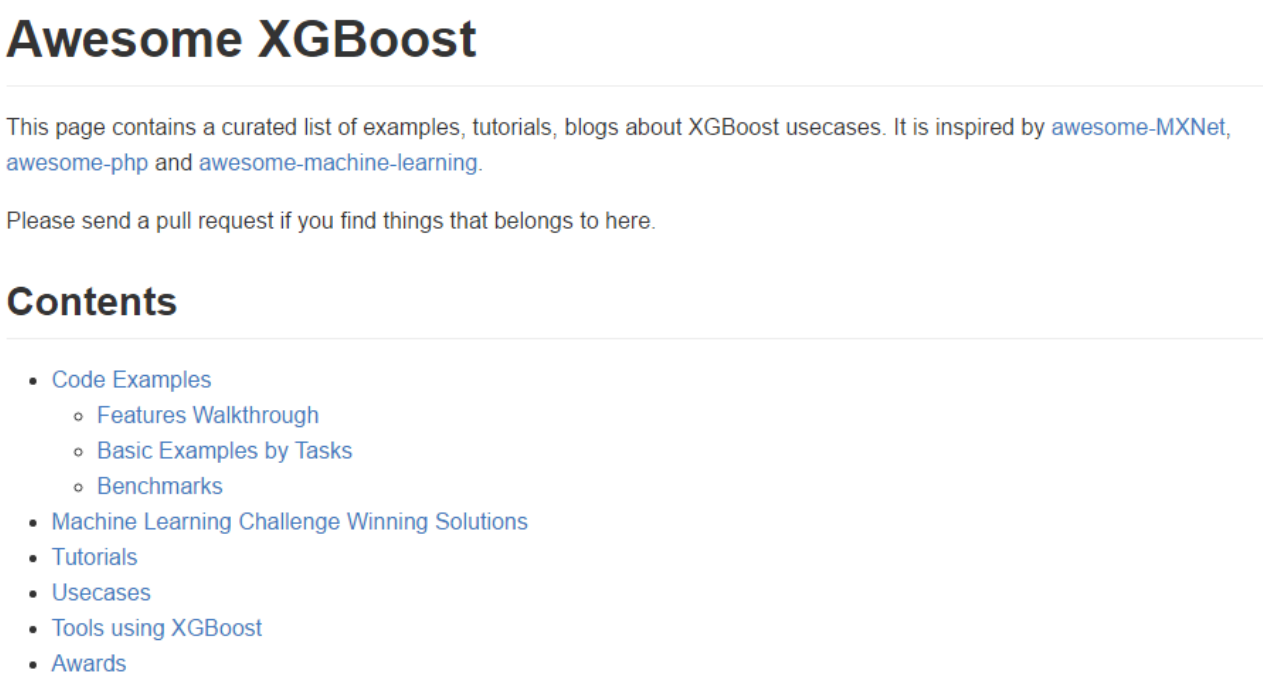


## List of Contributors

- Full List of Contributors
  - To contributors: please add your name to the list when you submit a patch to the project:)
- Kailong Chen
  - Kailong is an early contributor of xgboost, he is creator of ranking objectives in xgboost.
- Skipper Seabold
  - Skipper is the major contributor to the scikit-learn module of xgboost.
- Zygmunt Zając
  - Zygmunt is the master behind the early stopping feature frequently used by kagglers.
- Ajinkya Kale
- Boliang Chen
- Vadim Khotilovich
- Yangqing Men
  - Yangqing is the creator of xgboost java package.
- Engpeng Yao
- Giulio
  - Giulio is the creator of windows project of xgboost

# Resources

*Demo* of XGBoost:



## Awesome XGBoost

This page contains a curated list of examples, tutorials, blogs about XGBoost usecases. It is inspired by awesome-MXNet, awesome-php and awesome-machine-learning.

Please send a pull request if you find things that belongs to here.

### Contents

- Code Examples
  - Features Walkthrough
  - Basic Examples by Tasks
  - Benchmarks
- Machine Learning Challenge Winning Solutions
- Tutorials
- Usecases
- Tools using XGBoost
- Awards

The slides for this talk: github.com/hetong007/Van2016

# References


- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." arXiv preprint arXiv:1603.02754 (2016).
- Chen, Tianqi, and Tong He. "Higgs boson discovery with boosted trees." Cowan et al., editor, JMLR: Workshop and Conference Proceedings. No. 42. 2015.