

Kaggle

**Crowdsourcing Machine
Learning to Solve Today's
Greatest Data Problems**

kaggle™

Kaggle is the world's largest machine learning community



Overview

Kaggle is a **data modeling** and **data analysis** competition platform.

Businesses and researchers can **publish data** here, and statisticians and data mining experts can **compete** on the platform to produce the **best models**.

Kaggle specializes in the industry of **supervised ML**



Over **1.2MM** members











4MM+ Uploaded Solutions



Nearly **300** competitions



Over **4,500** open datasets

| | | |
|--|--|-----------------------------------|
|  | Titanic: Machine Learning from Disaster Start here! Predict survival on the Titanic and get familiar with ML basics <i>Getting Started</i> · 2 years to go · tabular, binary classification | 8,464 teams |
|  |  Web Traffic Time Series Forecasting Forecast future traffic to Wikipedia pages <i>Research</i> · a month to go · time series, internet, tabular, forecasting | \$25,000 1,095 teams |
|  |  Passenger Screening Algorithm Challenge Improve the accuracy of the Department of Homeland Security's threat recognition algorithms <i>Featured</i> · 2 months to go · terrorism, image, object detection | \$1,500,000 316 teams |
|  | Zillow Prize: Zillow's Home Value Prediction (Zestimate) Can you improve the algorithm that changed the world of real estate? <i>Featured</i> · 3 months to go · housing, real estate | \$1,200,000 3,727 teams |
|  | Cdiscount's Image Classification Challenge Categorize e-commerce photos <i>Featured</i> · 2 months to go · multiclass classification | \$35,000 220 teams |
|  | Porto Seguro's Safe Driver Prediction Predict if a driver will file an insurance claim next year. <i>Featured</i> · 2 months to go · tabular, binary classification | \$25,000 1,797 teams |

What types of problems does Kaggle help solve?

Sales/Marketing

- **Categorizing e-commerce products by image**
- Maximize sales and minimize returns
- Improving search term relevance
- Detect duplicate ads
- Predict if context ads will earn a user's click
- Subscriber churn

Finance/Insurance

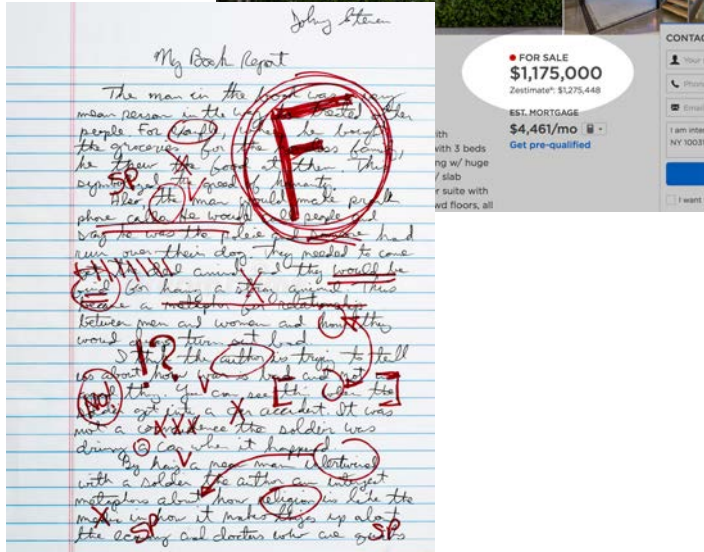
- Uncover predictive value in financial markets
- Pair financial products with potential customer
- Predict if a driver will file an insurance claim next year
- **Spot distracted drivers using computer vision**

Manufacturing

- **Cut the automobile manufacturing time spent on the test bench**
- Reduce manufacturing failures
- **Identify the boundaries of a car in an image**



What types of problems does Kaggle help solve?



Medical

- Identify which cancer treatment will be most effective
- Improve detection of lung cancer and heart disease
- Identify nerve structures in ultrasound images of the neck
- Predict the effect of Genetic Variants to enable Personalized Medicine
- Predict seizures in long-term human intracranial EEG recordings

Environmental

- Use satellite data to track the human footprint in the Amazon
- Detect and classify species of fish
- Identify endangered right whales in aerial photographs
- Predict hourly rainfall using data from polarimetric radars
- Predict physical and chemical properties of soil

Other

- Predict Donors Choose funding requests that deserve an A+
- Identify similar question/answer pairs in online forums
- Grade written essays
- Predict what songs a user will listen to next
- Estimating property values in the real estate market



How Do Kaggle Competitions Work?

1

Sign up for, and download data from the Kaggle Competition page

2

Develop a model based on a sample set of the full data

3

Models are evaluated by submitting answers to a reserved set of data, and are provided with a score

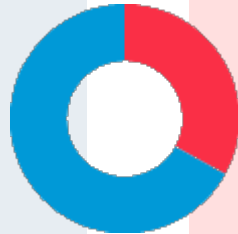
4

Leaderboard maintains competitor success while competition stays active on the platform

2

| Age | Income | Default |
|-----|----------|---------|
| 58 | \$95,824 | True |
| 73 | \$20,708 | False |
| 59 | \$82,152 | False |
| 66 | \$25,334 | True |

Training Data



3

| Age | Income | Default |
|-----|----------|---------|
| 73 | \$53,445 | ??? |
| 61 | \$36,679 | ??? |
| 47 | \$90,422 | ??? |
| 44 | \$79,040 | ??? |

Test Data

Kaggle provides all the functionality to make running a competition easy

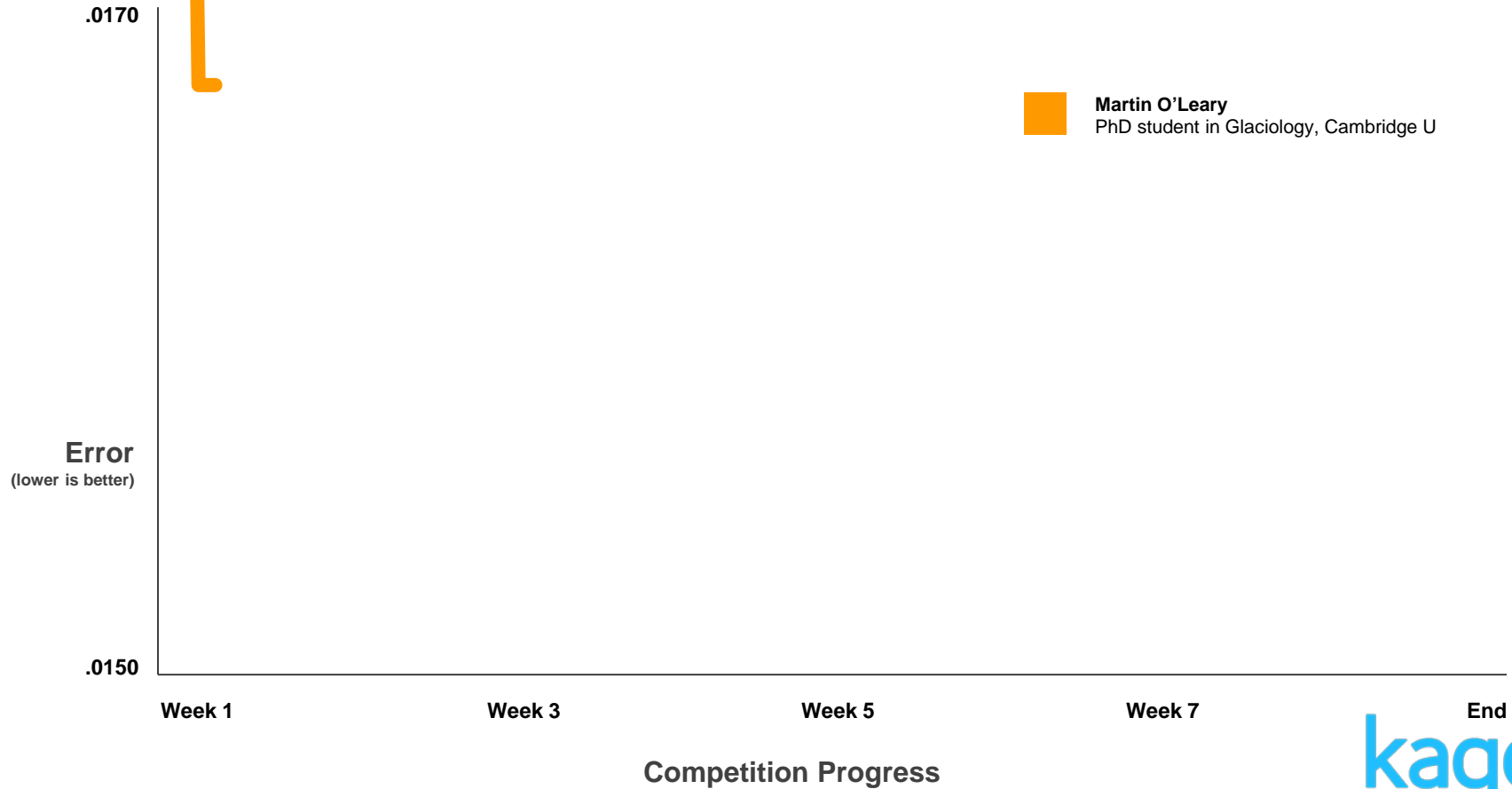
Overview Data Kernels Discussion Leaderboard Rules Host Admin [Join Competition](#)

Public Leaderboard Private Leaderboard

The private leaderboard is calculated with approximately 51% of the test data. [Refresh](#)

| # | Δpub | Team Name | Kernel | Team Members | Score 📉 | Entries | Last |
|---|------|--------------------|--------|--------------|---------|---------|------|
| 1 | — | GMV | | | 0.04875 | 13 | 14d |
| 2 | — | Terry | | | 0.06505 | 80 | 14d |
| 3 | — | Not hotdog | | | 0.06989 | 45 | 14d |
| 4 | — | UncleCat | | | 0.07028 | 70 | 14d |
| 5 | — | DLUT_VLG | | | 0.07041 | 54 | 14d |
| 6 | — | Deepinsight | | | 0.07488 | 103 | 14d |
| 7 | — | JobHunting | | | 0.07501 | 45 | 14d |

Mapping Dark Matter





Home • The Administration • Office of Science and Technology Policy

Search WhiteHouse.gov

Search



Office of Science and Technology Policy

“The world’s brightest physicists have been working for decades on solving one of the great unifying problems of our universe”

Competition Shines

Posted by Jason Rhodes on June 27, 2011 at 04:32 PM EDT

The world’s brightest physicists have been working for decades on solving one of the great unifying problems of our universe. It is a problem that explores our place in the cosmos and, as was the case with Newton’s law of gravitation and Einstein’s theory of relativity, the nature of the Universe if solved. Recently, top experts from an unlikely place

On May 23, a consortium of the very best from the American Physical Society posted the problem on the data-mining website Kaggle and Challenge.gov for all the world to weigh in. In less than a week, Martin O’Leary, a PhD student in glaciology, crafted an algorithm that outperformed the state-of-the-art algorithms most commonly used in astronomy for mapping dark matter.

Chalk another one up for the power of crowdsourcing, and this Administration’s commitment to using prizes and challenges to find solutions to some of our most pressing problems—here on Earth as well as in the furthest reaches of space!

The posted problem had to do with how scientists can go about mapping “dark matter.” Our Universe, it turns out,



YOUR FEDERAL TAXPAYER RECEIPT

Your Federal Taxpayer Receipt
UNDERSTAND HOW AND HOW YOUR TAX DOLLARS ARE BEING SPENT

Calculate Your Receipt • Refund the Numbers

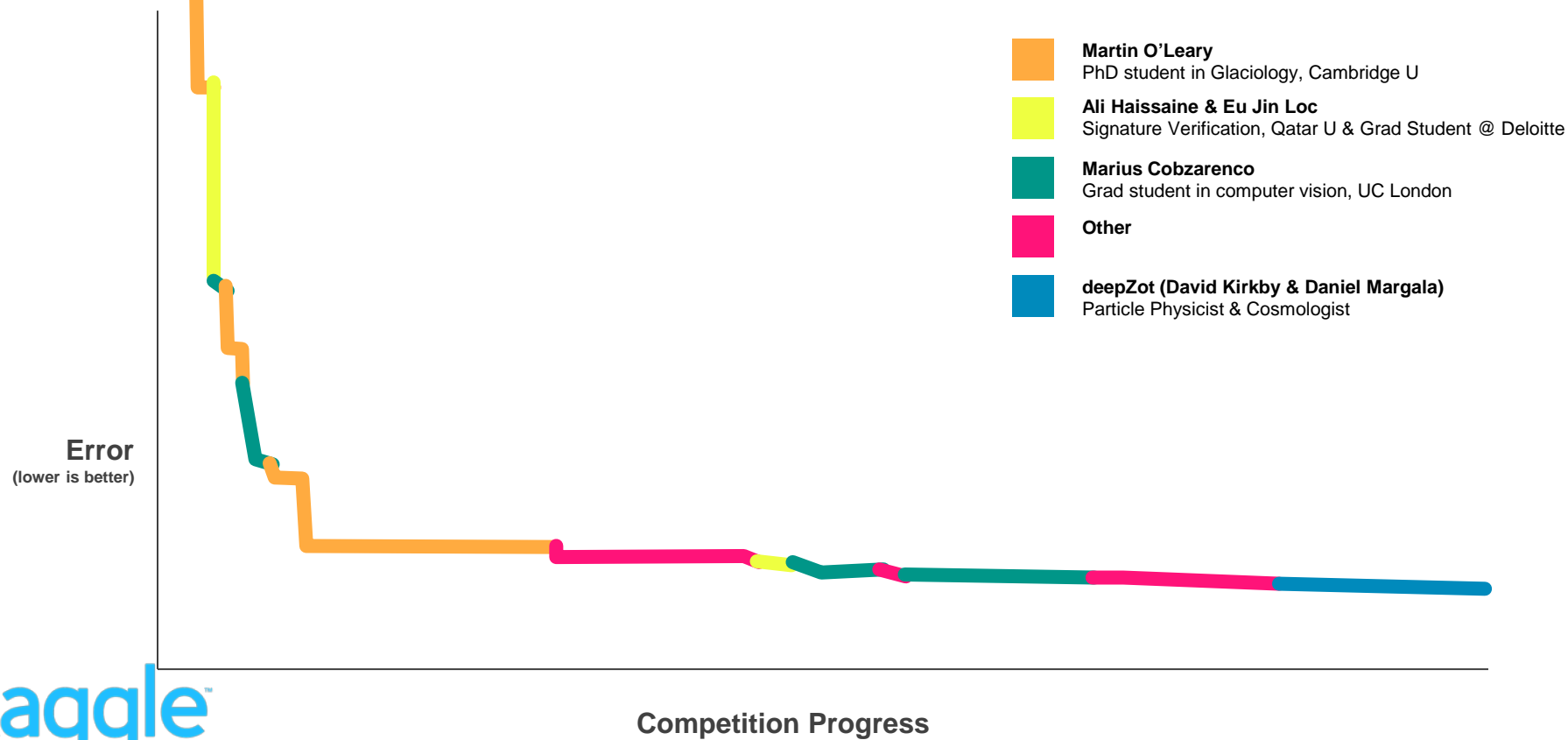
Enter your 2010 adjusted gross income for federal income tax

Standard Deduction: \$12,000
 Exemption: 1
 Taxable Income: \$38,000
 Estimated Refund: \$1,200

Don't have your identification ready? [Click on Home icon](#)

Launch the Receipt

Competitions are very powerful for extracting all the signal from a dataset



We have worked with around 50 Global 1000 companies

Oil & Gas

Shell

Apache


Top 5 E&P


Top 20 E&P

Consumer Internet

amazon

 Expedia

facebook

 Microsoft

yelp

Finance


Allstate

 MasterCard

Top Credit Card Issuer

Global Bank

 Liberty Mutual

FICO

Consumer Marketing

CLOROX

Neiman Marcus

TESCO

\$50b+ Beverage Co.
ABInBev

Walmart*

Healthcare & Pharma

Boehringer Ingelheim

Genentech

HERITAGE PROVIDER NETWORK

 MERCK

Pfizer

Industrial

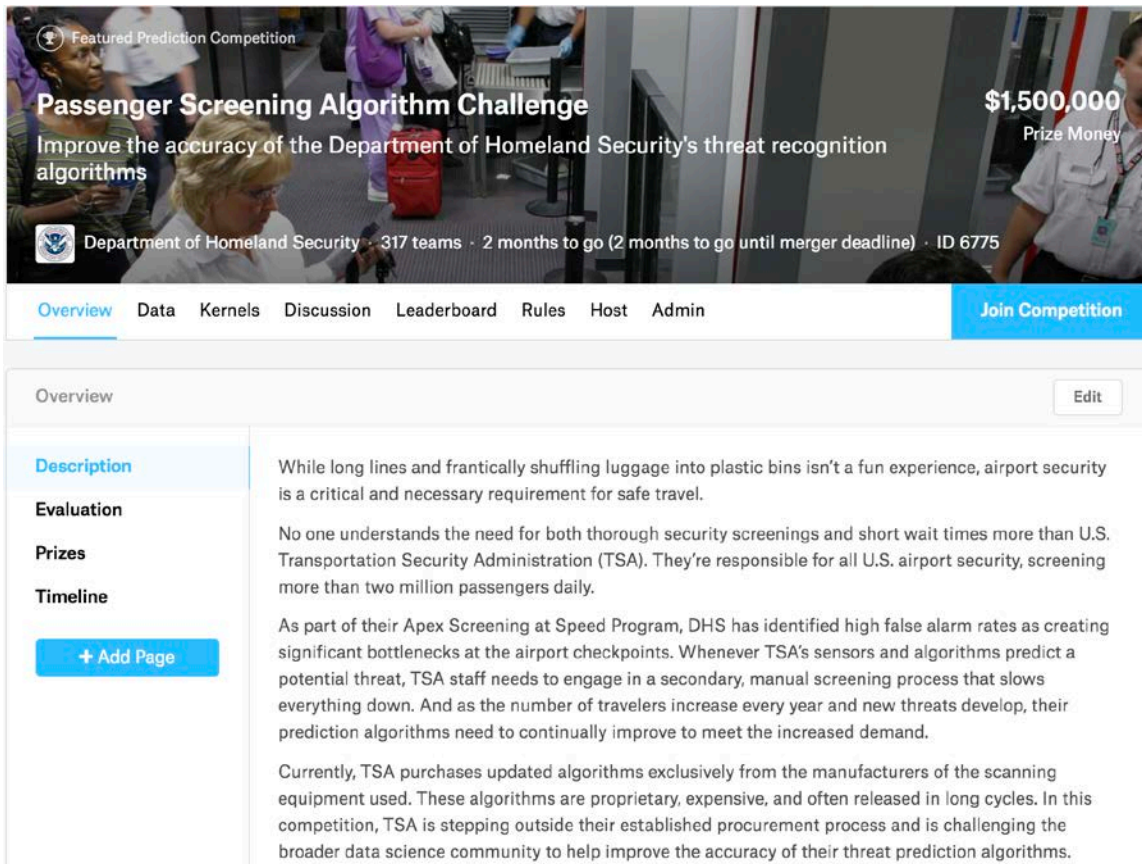
Ford

GE

kaggle

The DHS/TSA Passenger Screening Challenge has over 300 active teams

- Participants are challenged to perform detection on millimeter wave AIT scan using **representative objects**.
- Over **1,500** members have been approved to work on the data.
- Evaluates models using **logarithmic loss** to determine the likelihood/confidence level of a threat existing in one of many zones on the body.
- Entry Deadline: 12/4/17, Phase 1 concludes 12/15/17.
- **\$1.5 Million** offered to **Top 8** winners



The image shows a screenshot of the Kaggle competition page for the "Passenger Screening Algorithm Challenge". At the top, it features a banner with a photo of people at an airport security checkpoint. The banner text includes "Featured Prediction Competition", "Passenger Screening Algorithm Challenge", "Improve the accuracy of the Department of Homeland Security's threat recognition algorithms", and "\$1,500,000 Prize Money". Below the banner, it says "Department of Homeland Security · 317 teams · 2 months to go (2 months to go until merger deadline) · ID 6775". The navigation bar includes "Overview", "Data", "Kernels", "Discussion", "Leaderboard", "Rules", "Host", "Admin", and a "Join Competition" button. The "Overview" section is active, showing a "Description" tab. The description text reads: "While long lines and frantically shuffling luggage into plastic bins isn't a fun experience, airport security is a critical and necessary requirement for safe travel. No one understands the need for both thorough security screenings and short wait times more than U.S. Transportation Security Administration (TSA). They're responsible for all U.S. airport security, screening more than two million passengers daily. As part of their Apex Screening at Speed Program, DHS has identified high false alarm rates as creating significant bottlenecks at the airport checkpoints. Whenever TSA's sensors and algorithms predict a potential threat, TSA staff needs to engage in a secondary, manual screening process that slows everything down. And as the number of travelers increase every year and new threats develop, their prediction algorithms need to continually improve to meet the increased demand. Currently, TSA purchases updated algorithms exclusively from the manufacturers of the scanning equipment used. These algorithms are proprietary, expensive, and often released in long cycles. In this competition, TSA is stepping outside their established procurement process and is challenging the broader data science community to help improve the accuracy of their threat prediction algorithms."

Competitions are also used to find top talent



Completed • Jobs • 1,047 teams

Walmart Recruiting: Trip Type Classification

Mon 26 Oct 2015 – Sun 27 Dec 2015 (5 months ago)

| |
|-------------------|
| Dashboard |
| Home |
| Data |
| Make a submission |
| Information |
| Description |
| Evaluation |
| Rules |
| Timeline |
| Forum |
| Leaderboard |
| Public |
| Private |
| My Team |
| My Submissions |



















































Competition Details » [Get the Data](#) » [Make a submission](#)

Use market basket analysis to classify shopping trips

Walmart uses both art and science to continually make progress on their core mission of better understanding and serving their customers. One way Walmart is able to improve customers' shopping experiences is by segmenting their store visits into different trip types.



We have a community of over 1MM data scientists all ranked

| | | | | | | | | |
|----|---|---|-----------------------------------|--------------------|--|--|---|---------|
| 1 |  |  | Gilberto Titericz Junior | joined 5 years ago |  34 |  25 |  19 | 204,153 |
| 2 |  |  | Stanislav Semenov | joined 4 years ago |  27 |  9 |  0 | 179,828 |
| 3 |  |  | Μαριος Μιχαηλιδης KazAnova | joined 4 years ago |  25 |  23 |  21 | 164,603 |
| 4 |  |  | Faron | joined 3 years ago |  14 |  4 |  3 | 139,673 |
| 5 |  |  | Eureka | joined 3 years ago |  15 |  13 |  3 | 126,370 |
| 6 |  |  | raddar | joined 2 years ago |  9 |  6 |  3 | 124,584 |
| 7 |  |  | idle_speculation | joined 4 years ago |  7 |  8 |  6 | 122,333 |
| 8 |  |  | bestfitting | joined a year ago |  5 |  3 |  0 | 113,009 |
| 9 |  |  | utility | joined 3 years ago |  13 |  7 |  3 | 100,769 |
| 10 |  |  | Little Boat | joined 3 years ago |  10 |  15 |  5 | 99,782 |