

Somebody: “Check out my kernel! It’s doing well on the public LB and I think it’ll still be high ranking after the shakeup.”

You: “..... What on earth are you talking about?”

kaggle™



Kaggle

- Home of data science
 - Code + data + forums + competitions + courses + you!
- Made-up word
- Rhymes with “gaggle” (a group of geese)



Gaggle of geese ^^



This is Kerneler.

Kernels

- Hosted coding environment
- Runs in the browser (no downloading)
- Write, run & share data science code

Types:

- **Notebooks:** hosted Jupyter notebooks (text + code)
- **Scripts:** Raw Python, R or R Markdown files (code only)



This is Kerneler.

More Kernels Stuff

These are two cells from a notebook kernel

```
# this is a code cell where I can write code  
import pandas as pd
```

And this is a Markdown cell where I can write & format text.

Fancy, right?

- **Cell:** in notebooks, cells are editable blocks of text or code
- **Docker:** A docker is basically a whole file system. We have two (one in Python, one in R) with everything you need to run code written on Kaggle locally. You can download them from the Kaggle GitHub.
- **Commit:** Committing code runs it from top to bottom & saves a static version you can refer back to. (Like a save file in a video game.)

Competitions

A 🌈 rainbow 🌈 of competitions types:

- **Featured**: The most competitive competition with the biggest prizes. Often hosted by corporations looking to improve their current approach.
- **Analytics**: Not supervised machine learning. Many of these are designed to help nonprofits with tricky problems they're facing.
- **Research**: More theoretical problems based on active areas of research (academic & industrial).
- **Getting Started**: Designed to be your first or second competition.
- **Playground**: Especially interesting problems we think you might enjoy trying to solve or fun seasonal competitions.

Competitions

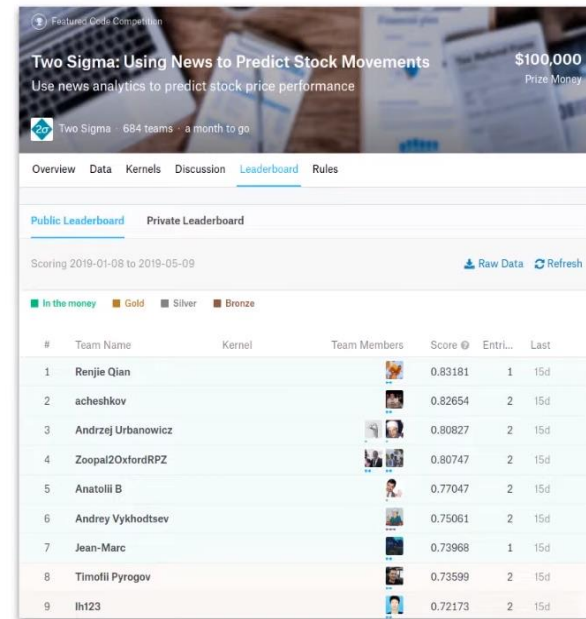
A 🌈 rainbow 🌈 of competitions types:

- **Featured:** The most competitive competition with the biggest prizes. Often hosted by companies to showcase their current approach. These are the most challenging problems we are currently facing. Areas of research we are interested in.
- **Only featured (and some research) competitions award points.** Points are used to determine overall user rankings.
- **Getting Started:** Designed to be your first or second competition.
- **Playground:** Especially interesting problems we think you might enjoy trying to solve or fun seasonal competitions.

Competitions

- **InClass:** Lets you host your own Kaggle competition. (They don't have to be for an actual class, that's just why it was started.)
- **Leaderboard:** The place where all the scores for entries to a specific competition are shown.
 - **Public:** This is the one that you see during the competition.
 - **Private:** You only see this after the competition ends. Usually scored on a different part of the data.
- **Shakeup:** Changes in ranking when the private leaderboard is shown at the end of a competition. (Usually people move down the ranks due to overfitting.)

Sample leaderboard for a competition



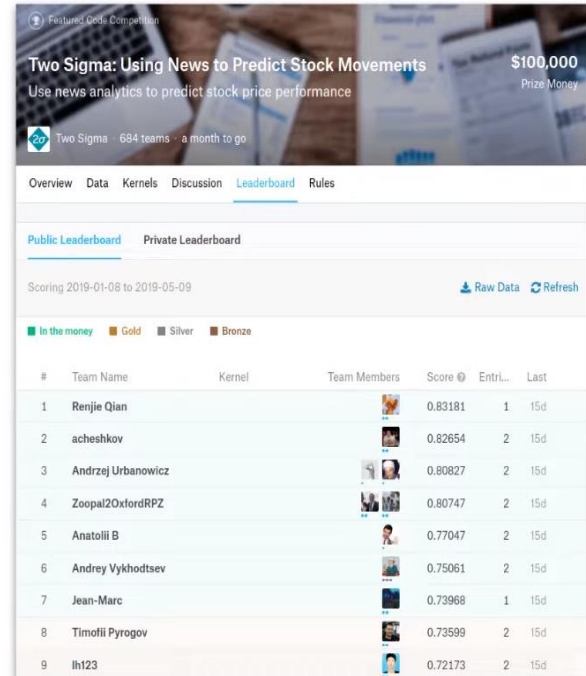
The screenshot shows the 'Two Sigma: Using News to Predict Stock Movements' competition on Kaggle. The title bar indicates a prize of \$100,000 and that 684 teams are participating. The 'Leaderboard' tab is selected, showing the 'Public Leaderboard' for the period 2019-01-08 to 2019-05-09. The table lists the top 9 teams, their scores, and the number of entries. The top team, Renjie Qian, has a score of 0.83181. The table also includes a legend for 'In the money' (green), Gold (yellow), Silver (grey), and Bronze (brown) medals.

#	Team Name	Kernel	Team Members	Score @	Entri...	Last
1	Renjie Qian			0.83181	1	15d
2	acheshkov			0.82654	2	15d
3	Andrzej Urbanowicz			0.80827	2	15d
4	Zoopal2OxfordRPZ			0.80747	2	15d
5	Anatolii B			0.77047	2	15d
6	Andrey Vykhodtsev			0.75061	2	15d
7	Jean-Marc			0.73968	1	15d
8	Timofii Pyrogov			0.73599	2	15d
9	lht23			0.72173	2	15d

Competitions

“In the money” means that people at that rank will win money. (If they still have that rank on the private leaderboard after the shakeup.)

Sample leaderboard for a competition

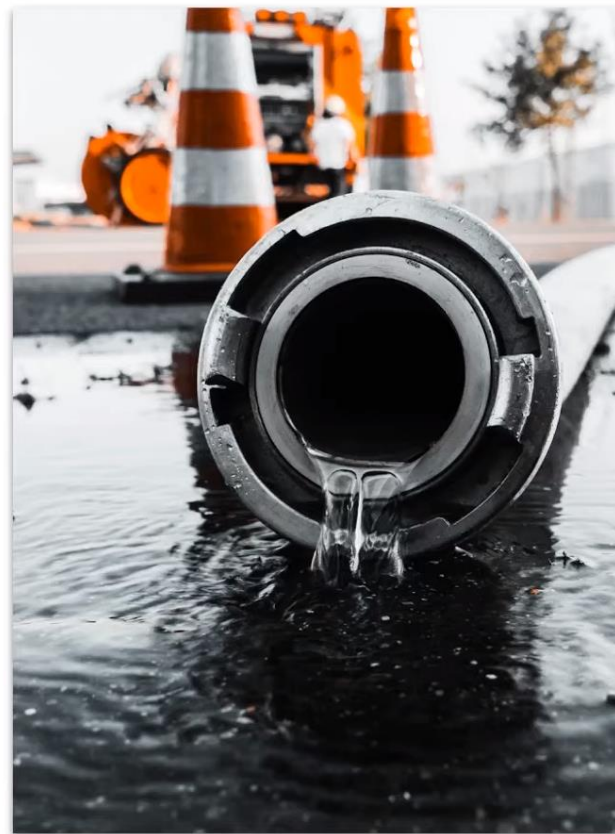


The screenshot shows the 'Two Sigma: Using News to Predict Stock Movements' competition on Kaggle. The title bar indicates a prize money of \$100,000. The competition is currently in the 'In the money' phase, with 684 teams and a month to go. The leaderboard is divided into 'Public Leaderboard' and 'Private Leaderboard' tabs. The 'Public Leaderboard' is selected, showing a table of team rankings. The table includes columns for rank, team name, kernel, team members, score, entries, and last update. The top 9 teams are listed, with the first team, Renjie Qian, having a score of 0.83181. The table also includes a legend for 'In the money', Gold, Silver, and Bronze medals.

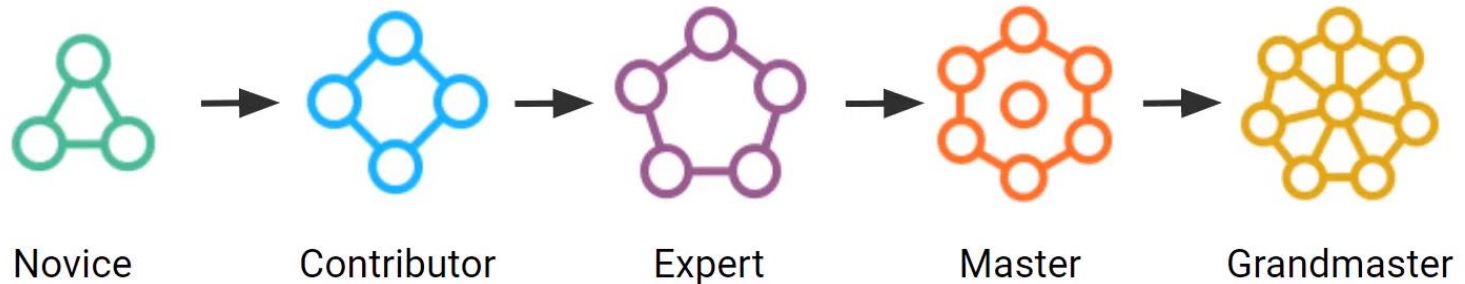
#	Team Name	Kernel	Team Members	Score @	Entri...	Last
1	Renjie Qian			0.83181	1	15d
2	acheshkov			0.82654	2	15d
3	Andrzej Urbanowicz			0.80827	2	15d
4	Zoopal2OxfordRPZ			0.80747	2	15d
5	Anatolii B			0.77047	2	15d
6	Andrey Vykhodtsev			0.75061	2	15d
7	Jean-Marc			0.73968	1	15d
8	Timofii Pyrogov			0.73599	2	15d
9	lht23			0.72173	2	15d

Leakage

- When information about the target variable is accidentally included in the training data
 - All photos of the target class were taken with a specific camera, so you end up actually building a camera type detector
 - You're trying to predict if someone will get an illness, but you don't remove information about whether someone was *treated* for that illness
- Kaggle data scientists work very hard to minimize leakage in competitions

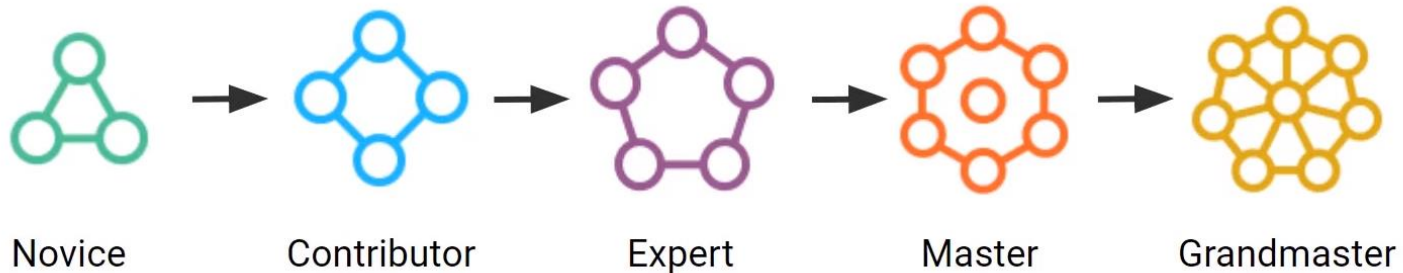


Kaggle Progression System



- Once you reach **Contributor**, your votes count for awarding other people medals!
- Once you reach **Expert**, you get ranked separately in three categories: **discussion**, **kernels** and **competitions**
- **Medals**: These are awarded for having your content upvoted or being near/at the top of the leaderboard for competitions.

Kaggle Progression System



Out of >3 Million Kagglers there are:

- **Competitions:** 145 **Grandmasters**, 1220 **Masters**
- **Kernels:** 11 **Grandmasters**, 49 **Masters**
- **Discussion:** 12 **Grandmasters**, 20 **Masters**

Counts taken June, 2019

It's pretty hard to become a **Master** or **Grandmaster**,
but you can get there with hard work & dedication! 💪💪