

# Kaggle Teams!




A quantitative and qualitative review  
of team performance  
How they form, perform, and work

# What's the goal?

- strategizing to become a Master?
- is it better to join a team or go it alone?
- who, and what goes in to a winning team?
- how to get invited to a team?
- how to create a team?
- what are the best tools, stacks?
- how to achieve top results?

# Me - paulperry <https://www.kaggle.com/paulperry>




**paulperry**


Boston, MA, United States  
Joined 5 years ago · last seen in the past day

[K](#) [T](#) [in](#) <http://www.paulperry.net/>


Following 5


  
**Competitions Expert**


[Home](#) [Competitions \(18\)](#) [Kernels \(13\)](#) [Discussion \(72\)](#) [Datasets \(0\)](#) ... [Edit Profile](#)




**Competitions Expert** 


Current Rank	Highest Rank
<b>1902</b>	<b>138</b>
of 82,715	

  
1


  
1


  
5


<a href="#">Liberty Mutual Group - Fire...</a>  · 4 years ago · Top 1%	<b>5<sup>th</sup></b> of 634
<a href="#">Africa Soil Property Predicti...</a>  · 4 years ago · Top 3%	<b>27<sup>th</sup></b> of 1233
<a href="#">Western Australia Rental Pr...</a>  · 3 years ago · Top 23%	<b>13<sup>th</sup></b> of 59


**Kernels Novice** 


**Unranked**

  
0


  
0


  
1


<a href="#">technical_13 does NOT help...</a>  · a year ago	<b>11</b> votes
<a href="#">test_kernel</a> a year ago	<b>0</b> votes
<a href="#">technical 16 - friend of time</a> a year ago	<b>0</b> votes




**Discussion Novice** 

**Unranked**


  
0

  
2

  
19

<a href="#">Data Scientist Hero</a>  · 2 years ago	<b>8</b> votes
<a href="#">Silence xgboost</a>  · 3 years ago	<b>7</b> votes
<a href="#">Can't unzip data</a>  · 3 years ago	<b>3</b> votes


# Sergey Yurgenson <https://www.kaggle.com/ccccat>



**Sergey Yurgenson**

Boston, United States  
Joined 8 years ago · last seen 3 days ago  
<http://www.datarobot.com/>

Followers 146




**Competitions  
Grandmaster**


[Home](#) [Competitions \(59\)](#) [Kernels \(7\)](#) [Discussion \(344\)](#) [Followers \(146\)](#) [Contact User](#) [Unfollow User](#)


**Competitions Grandmaster**

Current Rank  
**131**  
of 82,715

Highest Rank  
**1**

**18**

**22**

**10**


**Predicting a Biological Res...** **1<sup>st</sup>**  
of 699  
🏆 · 6 years ago · Top 1%


**Flu Forecasting** **2<sup>nd</sup>**  
of 50  
🏆 · 4 years ago · Top 4%


**The Big Data Combine Eng...** **2<sup>nd</sup>**  
of 425  
🏆 · 5 years ago · Top 1%

**Kernels Novice**

**Unranked**

**0**

**0**

**1**

**R version of most popular l...** **20**  
votes  
🏆 · 2 years ago


**R version of Santa\_03** **3**  
votes  
a year ago


**R version of most popular l...** **0**  
votes  
2 years ago


**Discussion Expert**

Current Rank  
**218**  
of 58,631

Highest Rank  
**43**

**4**

**13**

**104**

**Data Scientist Hero** **38**  
votes  
🏆 · 2 years ago

**Validation vs LB** **21**  
votes  
🏆 · 2 years ago

**A leak in the data ?** **21**  
votes  
🏆 · 2 years ago

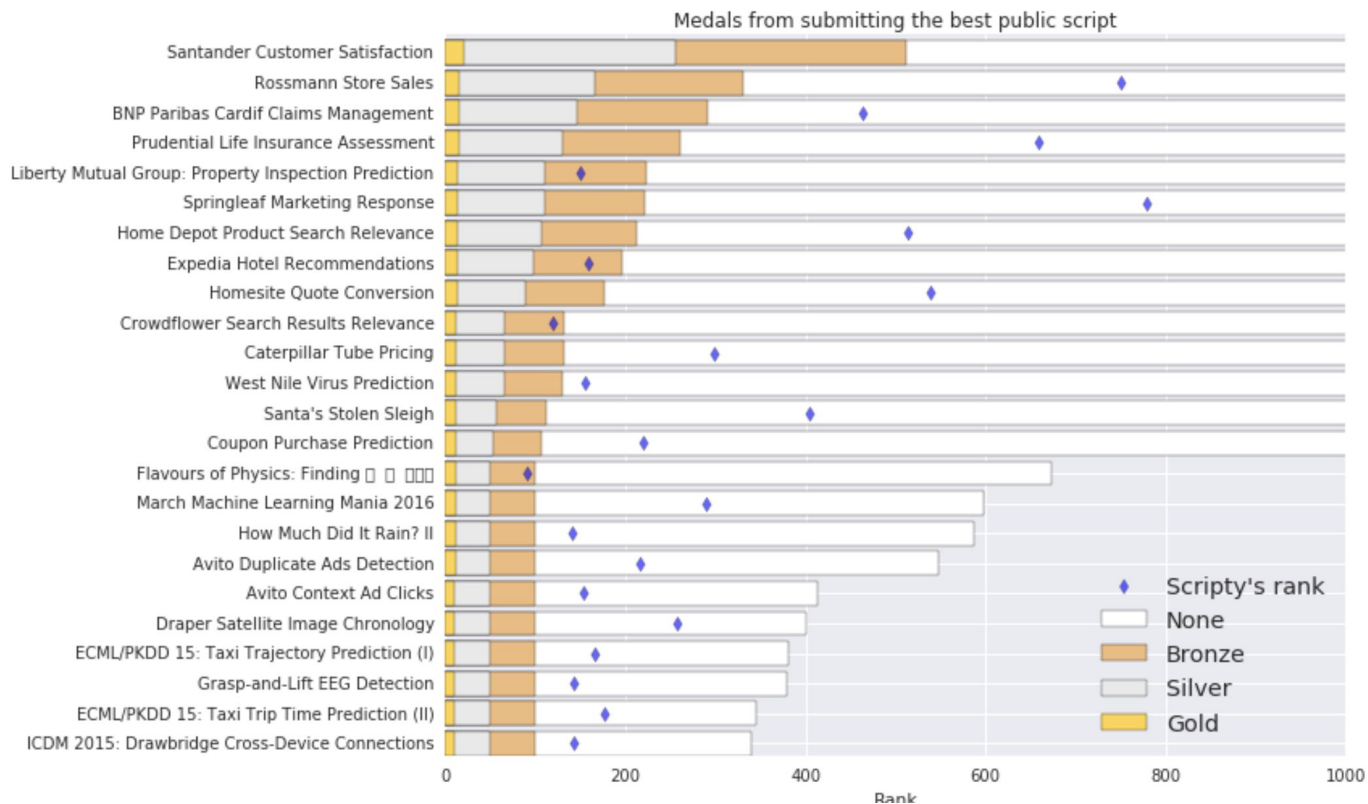
# Data for Kaggle Team analysis

1. Web-scraping Kaggle Leaderboard as of February 17, 2016.
  - a. 200 completed competitions
  - b. 133 competitions with a cash prize that had teams
  - c. Source code for web-scraping at [Jim Thompson's github repo](#).
2. Builds on previous analysis of Kaggle teams [by country](#), [by team structure](#), [by profile](#), and [over time](#).
3. Does not use of the [Meta-Kaggle](#) (Under construction) dataset ... yet.
4. But summarize the EDA from Meta-Kaggle
5. Summary of team related questions to <https://kagglenoobs.slack.com/>
6. Interviews with top Kagglers

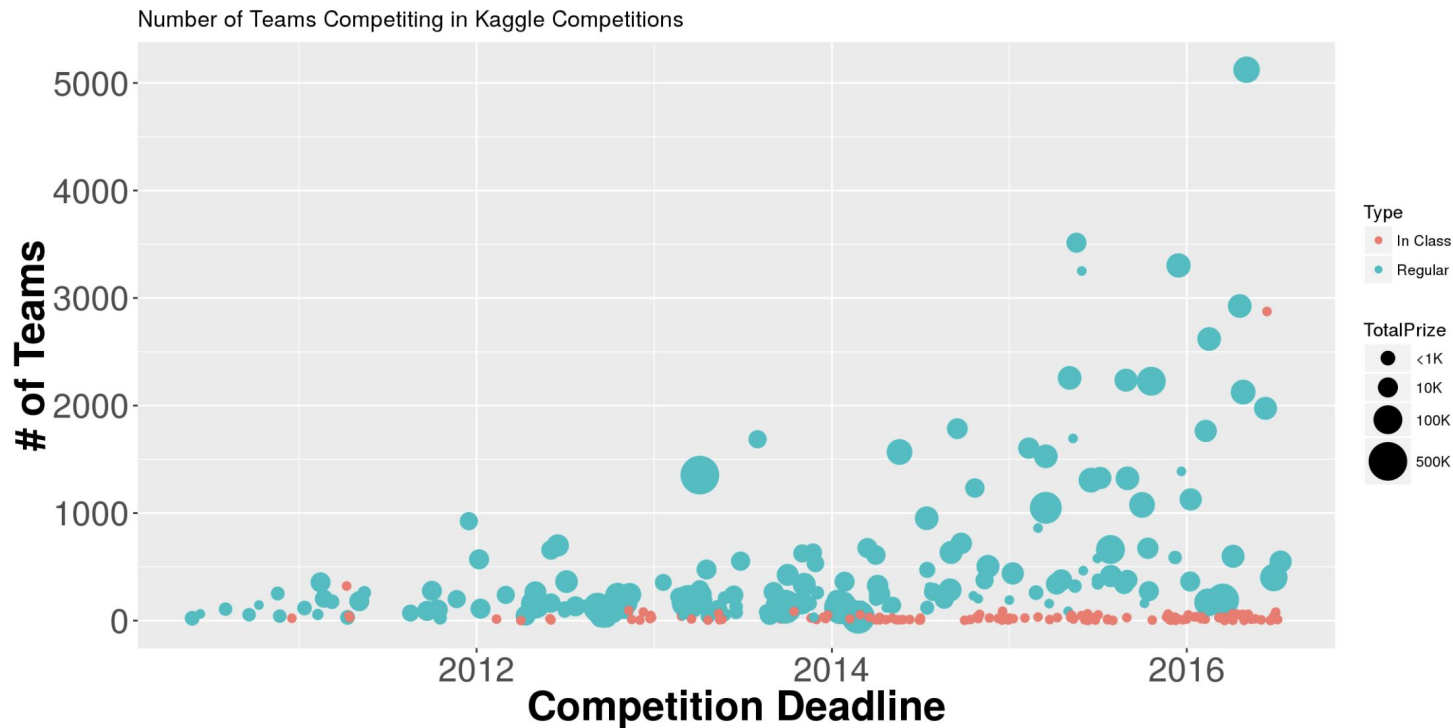
# Strategizing to become a Master

- **Don't chase the leaderboard:** You are at risk of overfitting.
- **Don't chase rank** Don't be a lazy kaggler and chase rank by submitting other people's scripts.
- **Read the Forums:** seems obvious but every little nugget counts.
- **Share some scripts:** scripts do provide great opportunities for code sharing and learning, but beat those scripts
- [How long does it take to place first?](#) About a year.

# Scripty McScriptface the Lazy Kaggler



# Kagglers are teaming up more



Credit [Jeong-Yoon Lee](#)



# Why Team up?

There is a lot to learn from the top data science practitioners.

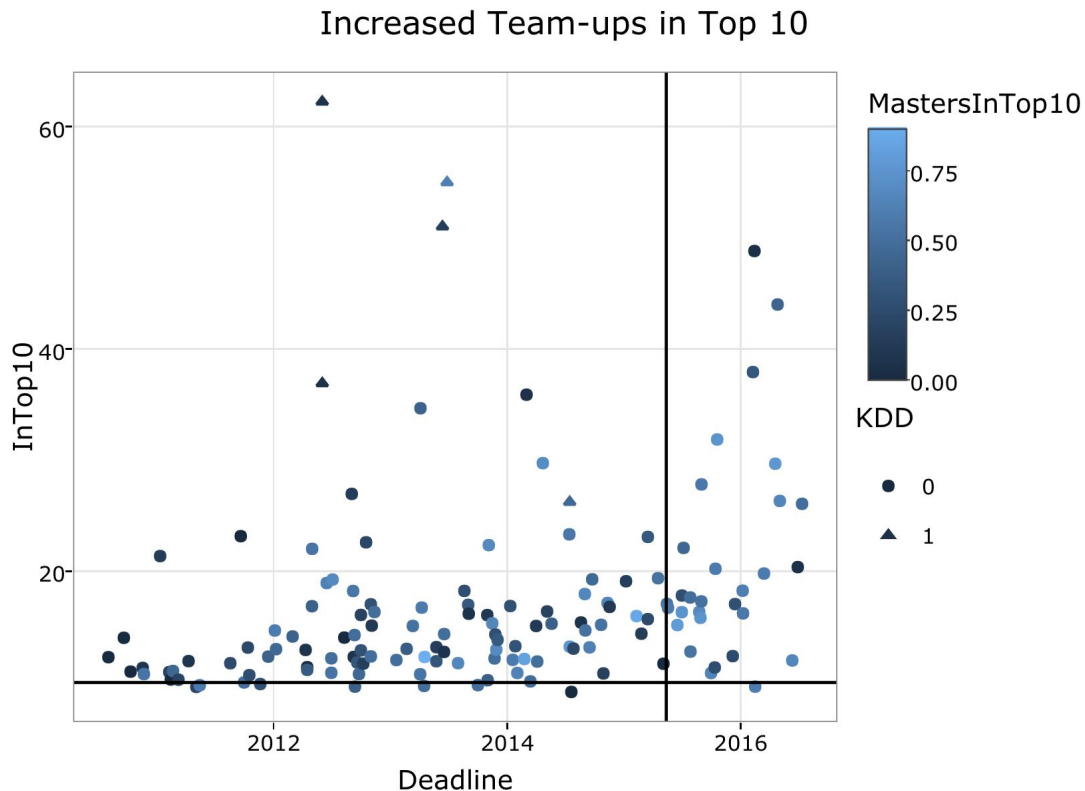
- They have dedicated a lot of time to develop their workflows
- have great intuition of what methods are likely to work
- and are efficient in their use of time.

Teams are winning more!

# And teams are winning more

In [May 13, 2015](#), Kaggle updated the ranking system to penalize teams less than before.

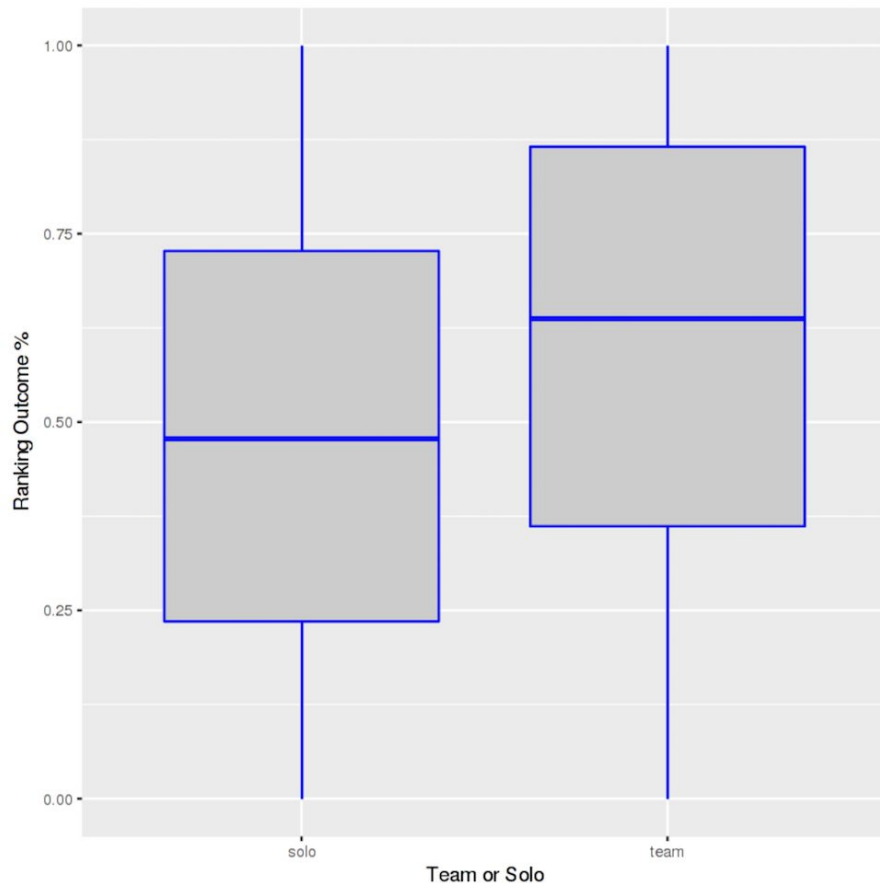
Recent competitions feature big and rich data sets conducive to blending has meant increased team-ups in the top of the leaderboard.



# Are you better off solo?

solo wins are rare

and only after a lot of submissions



# When do you team up?

- “strategic” teaming: somewhere after 1/3 of the competition.
  - plan to work together, learn, cooperate
  - may select teammates based on complementary strengths.
  - Assess if potential teammate actively participates in the competition and will not be a ballast
  - Also, by then you know if you want to dedicate time to competition yourself
- “tactical” teaming: Just before teaming deadline
  - Not much time to cooperate, but lets blend our models.
  - Makes sense for people close to each other on leaderboard especially near some threshold (teams 4 and 5 hoping to jump to second place after teaming-up) or near gold medal threshold
- “Grandmaster problem” - needs individual gold medal
  - Easy to do during recruiting competitions when teams are not allowed and scripts are limited.

# Who do you team up with?

- user ranking in current competition
- user ranking in previous competition
- someone with a different solution than yours:

*"I usually ask the potential member about the solution and I try to figure out if its relatively different from my current solution. One thing that makes you win is diversity of solutions. Remember that blending 2 solution with low correlation is much better that blend 0.999 correlated solutions." [Titericz](#) (aka giba1) [ref](#)*

- availability for the competition
- user computer resources
- someone who is active in the forums and who shares kernels
- Someone with some experience, good work ethic, and positive attitude

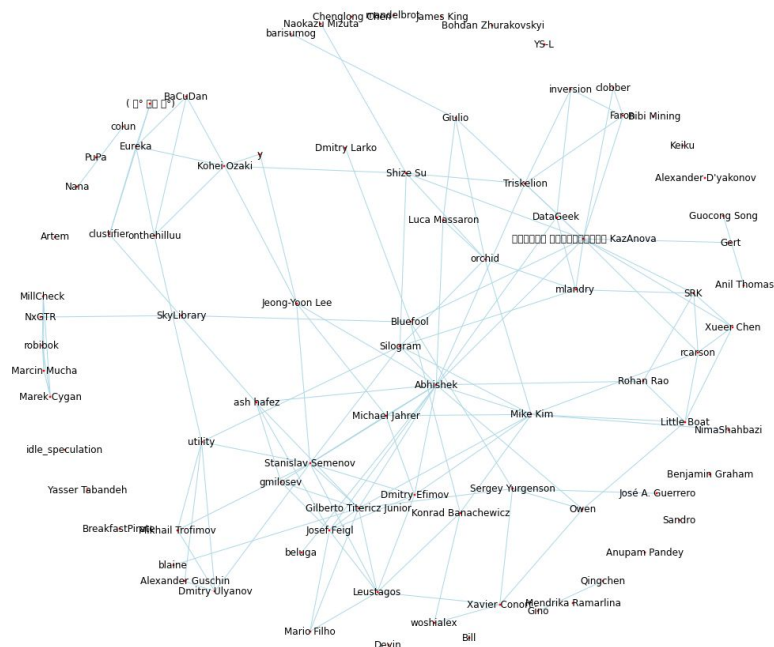
*"There has been exceptions, but in general would like someone that has played at least a few kaggle competitions . This is mainly to ensure that he/she understands the rules well, will not attempt cheating or do something funny by mistake . I like a good work ethic. Good communications with positive attitude. There is no other requirement. In general I don't see a collaboration as a business thing. I am much more happy if after a potential merge we can be friends and long-lasting teammates with the group." [Kazanova](#)*

## How do you find team members?

### Top 100 Users with Most Team Memberships

## Common Collaborators

## Who are the most connected Kagglers?



# The Kaggle Social Graph of Top Players

# How does the team organize their work?

Kaggle team interview questions: What works? What doesn't work?

1. How do you select who you want to team with?
2. How do you communicate? Email, chat, video chat?
3. How do you share data (or features)?
4. Do you share code? If so, how?
5. Do you share code frameworks?
6. How do you decide to divide the work?
7. How do you avoid duplicative work? Or repeated work?
8. How do you avoid team members finding the same stuff?
9. How do you keep track of model performance?
10. How do you decide who gets to submit what on each day?
11. Who picks and how do you pick the final submissions?
12. If you wanted to learn what worked on other Kaggle teams, what questions would you ask?

# Interview with Lucas Eustaquio da Silva (Leustagos) 2/29/2016

1. How do you communicate? Email, chat, video chat? - **chat**
2. How do you share code? - **usually we don't, but when we do, its on dropbox**
3. How do you share data (or features)? - **csv on dropbox**
4. Do you share code frameworks? - **people don't like much to mess with another code**
5. How do you decide to divide the work? - **chat, but we usually have some slightly different approaches. its rare but we can suggest each other based on availability which ideas we can pursue first. Usually telling what i will do prevent others from doing the same**
6. How do you avoid duplicative work? Or repeated work? - **we don't avoid, but its not a big issue. i don't team up from the start so we can have distinct approaches when merging teams**
7. How do you avoid team members finding the same stuff? - **we don't**
8. How do you keep track of model performance? - **each one is responsible for keeping its versioning. i use git, some just duplicate and enumerate files. on each submission we describe which models we used to generate it**
9. How do you decide who gets to submit what on each day? - **common sense. we divide equally the number of submissions, but if someone need more he asks**
10. Who picks and and how do you pick the final submissions? - **the leader picks it. its a consensus. i never had any trouble to do it. with the right reasons its very easy to choose. Of course some times i don't agree and we just go with the majority. just don't be very picky and it will go smoothly. its very rare for me to not pick my best submission. except on some competitions that are too random.**



# A proposed hierarchy of Kaggle team cooperation

First, forget ["team camaraderie"](#). This is your last priority.

1. Agree on how and when to communicate: Slack, Discord, etc
2. Share strategies, approaches, but not ideas
3. Avoid sharing too many ideas to the point every team member is building exactly the same solution
4. Keep track of model performance across the team
5. Share resultsets
6. Agree on a CV strategy (fix fold indexes)
7. Ensemble model results: Usually one team member is responsible to keep the stack ensembling working.
8. Have a way to compare models
9. Share derived features
10. Share code fragments

# What teams could do better

1. Share code frameworks
2. Share cloud infrastructure
3. Share workload (split feature generation)
4. Review each other's work. Point to any obvious errors or omissions.
5. Share a computing instance
6. Share code platform
7. Develop a framework !

# Kaggle Teams: Bottom Line

TL;DR: New masters are still being made.

**Keep learning, climb the leaderboard, and team up.**