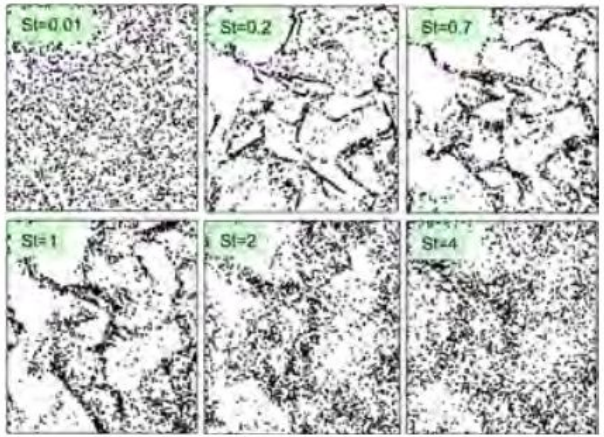
kaggle™

The Home of Data Science (for Science)?

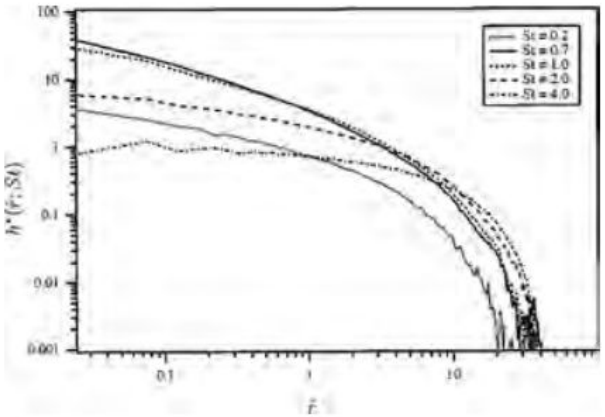
Walter Reade, Ph.D.
inversion@google.com

About Me

- Chemical Engineer (PhD at Penn State)
- Consumer Products
- Data Science / Machine Learning
- Kaggle



DNS of isotropic turbulence

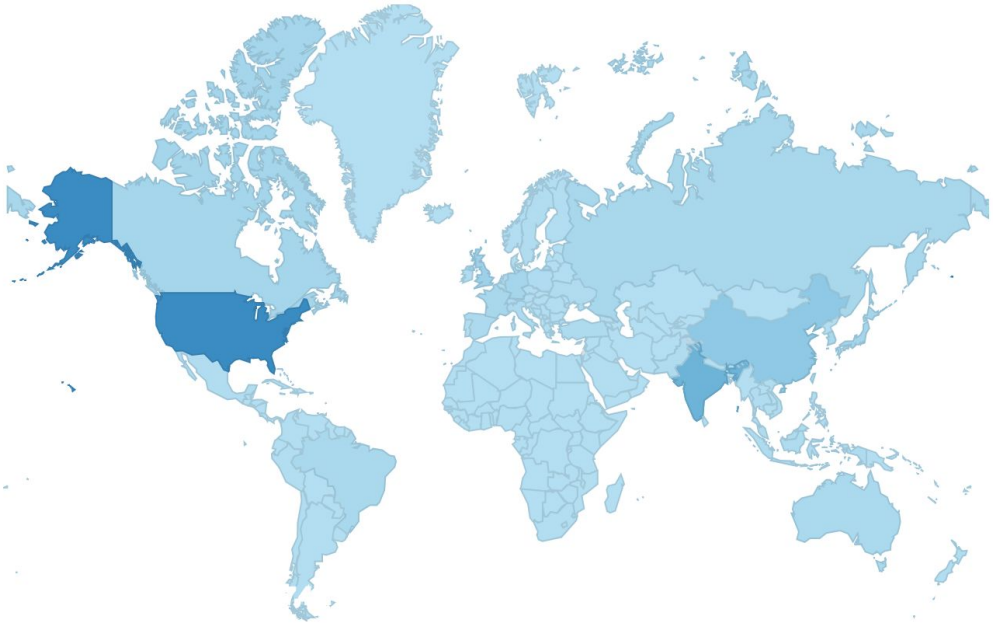


Residual RDF $(g(r)-1)$vs $\widehat{r}$

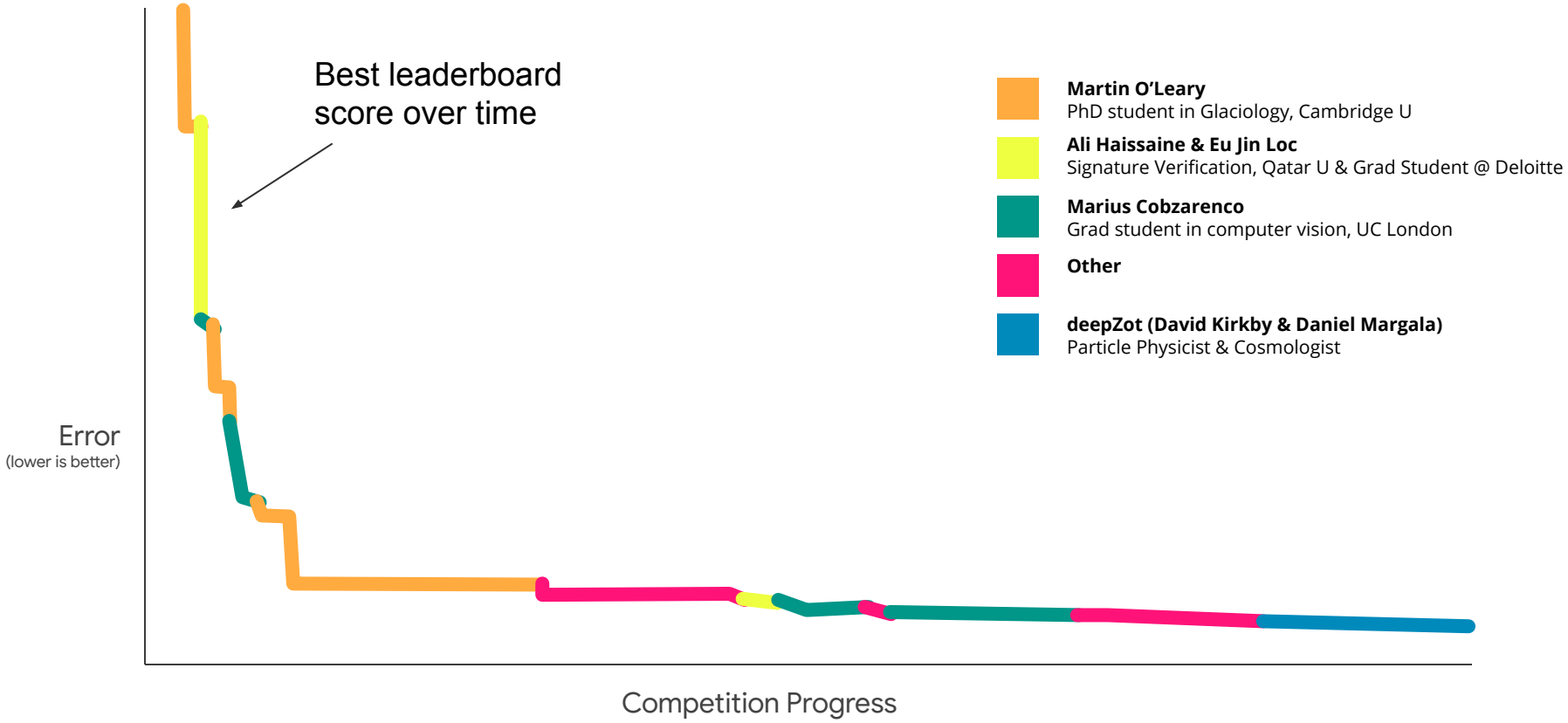
[1]Reade and Collins, Phys. Fluids, Vol. 12, 2000.

Why Kaggle?

Kaggle is the world's largest data science community



Competitions extract all the signal from a dataset

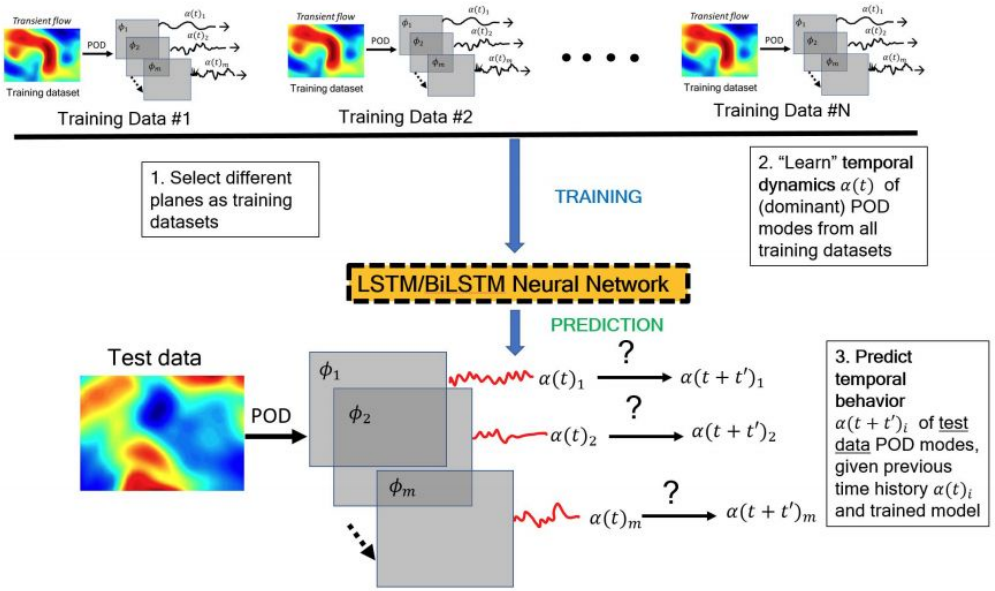


Science on Kaggle?

ML for Science - Exciting vs "Mundane"

Mundane - "We fit our 200 observation dataset with a RandomForest"

Exciting - "We use Proper Orthogonal Decomposition and Galerkin Projection for dimensionality reduction of 3D turbulent flow and modeled temporal dynamics using LSTM NNs."

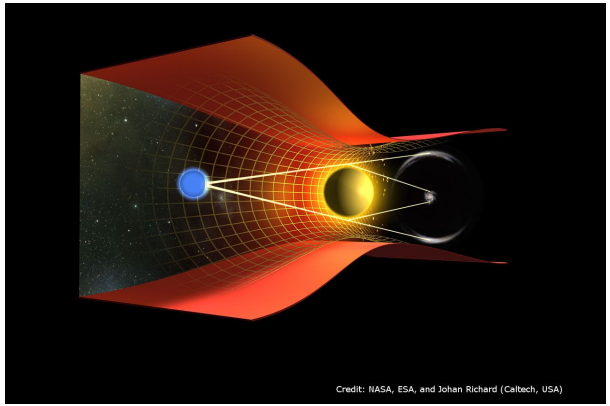


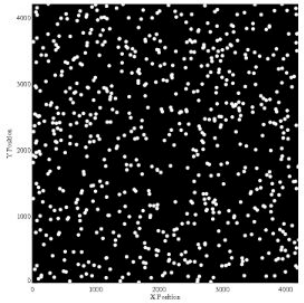
https://arxiv.org/pdf/1804.09269.pdf

Dark Worlds - Gravitational Lensing

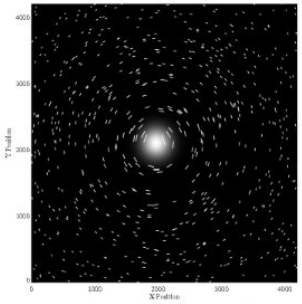
Winton Capital

Predicting dark matter
Top solutions used Bayesian methods

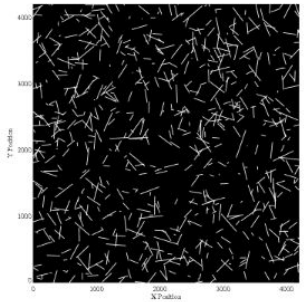




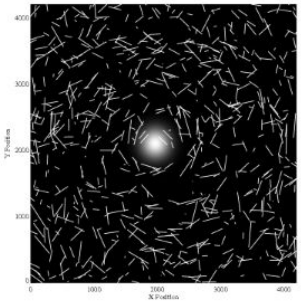
A. Distant circular galaxies (or dots in this case) are randomly distributed in the sky. Each galaxy has an (x,y) coordinate corresponding to the position in the sky from 0:4200



B. By placing a Dark Matter halo in the middle of the sky between us and the background galaxies, they are altered such that they become elliptical. The lines show the orientation and size of the major axis of the galaxy.



C. However unfortunately galaxies are NOT circular and infact they are inherently elliptical. This property is random, however since the Universe has no preferred ellipticity this averages out to zero in the case of no other influence.

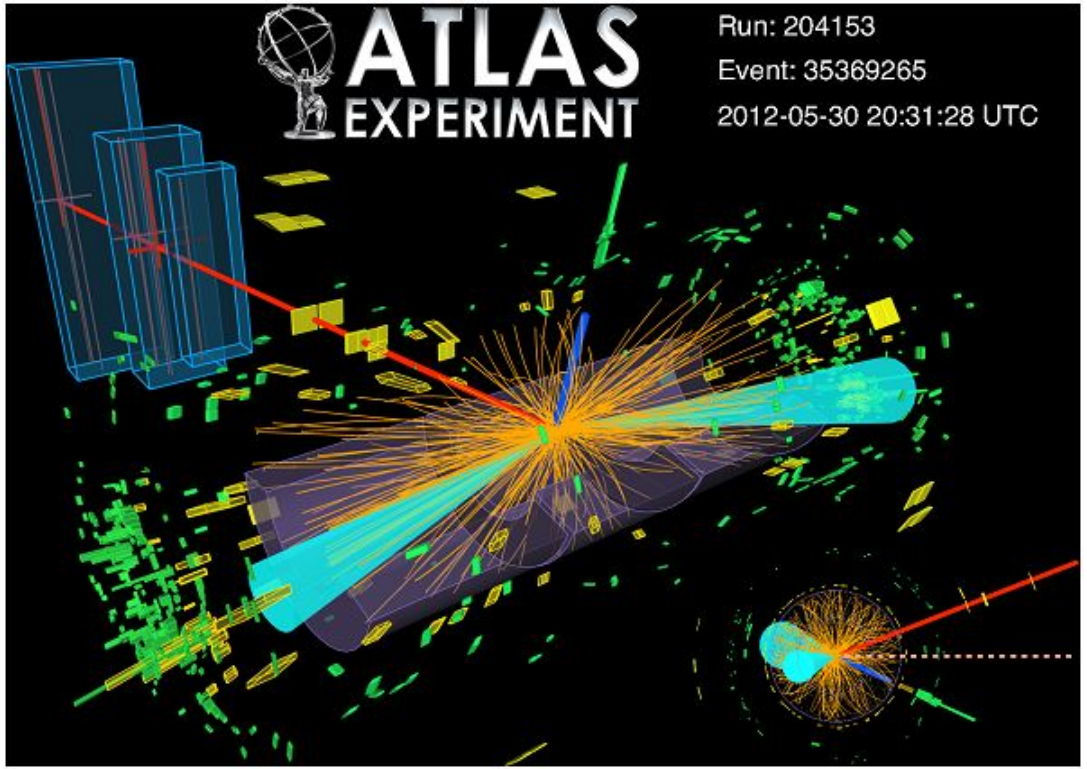


D. Therefore if we placed a Dark Matter halo into a field of randomly elliptical galaxies we would get a field that does not average out to zero. If we can use the fact that Dark Matter makes the pattern seen in B, we should be able to detect the position of the central halo.

Higgs Boson Machine Learning Challenge

CERN

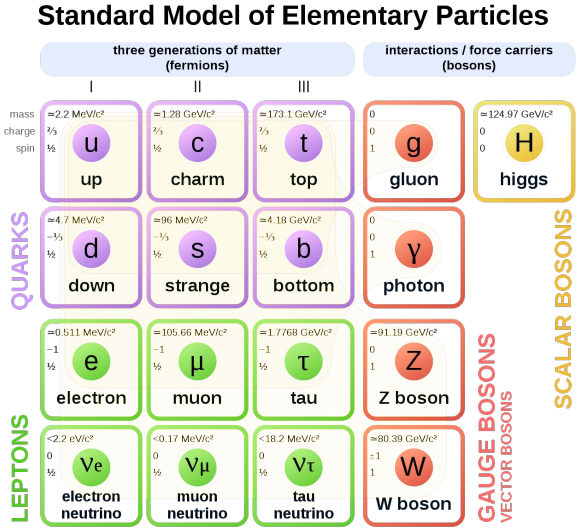
Use the ATLAS experiment to identify the Higgs boson, predicting tau tau decay of a Higgs boson vs background

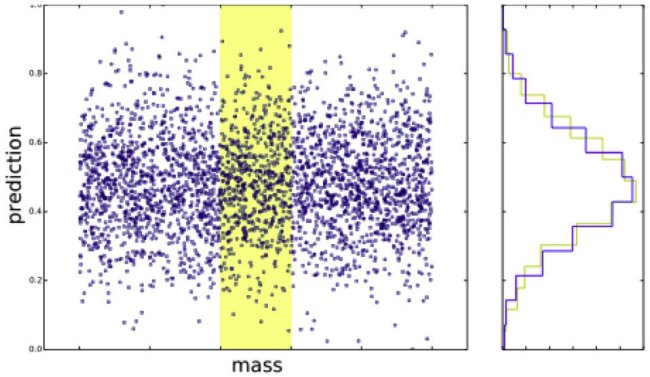


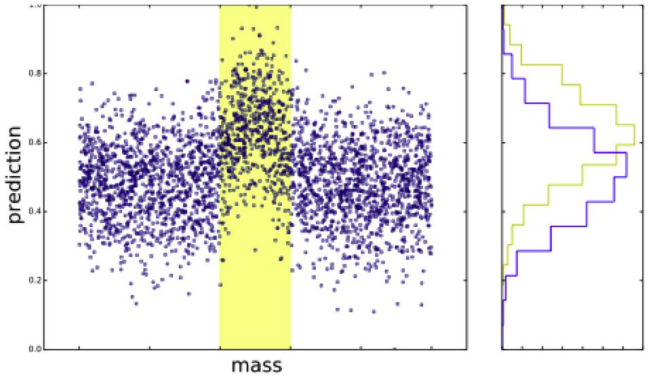
Finding τ ⟶ μμμ

CERN

Given a list of collision events and their properties. predict whether a τ ⟶ 3μ decay happened in the collision.



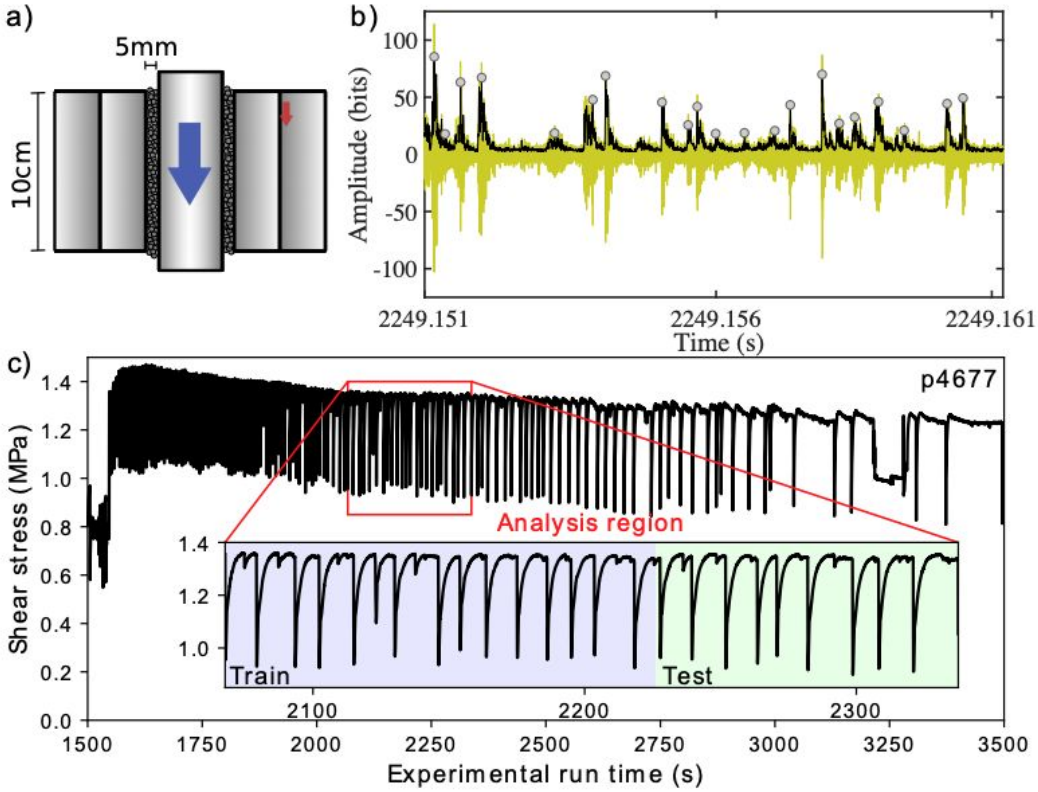




Earthquake Prediction

LANL / Penn State / Purdue

Model laboratory earthquake data



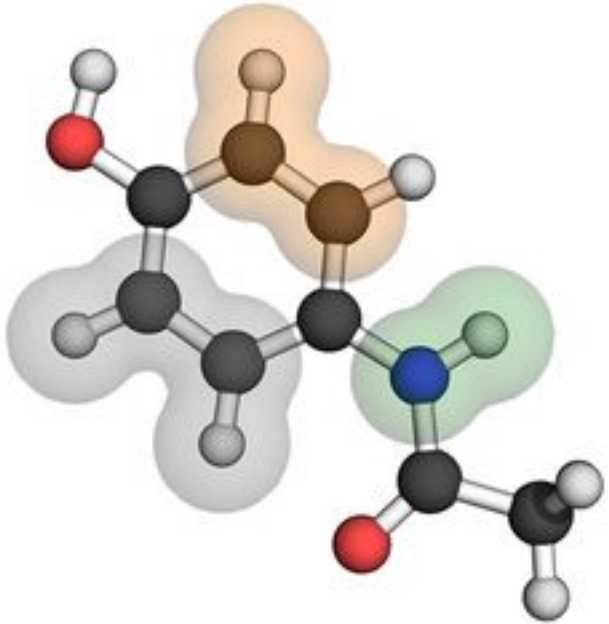
https://arxiv.org/pdf/1810.11539.pdf

Predicting Molecular Properties

CHemistry and Mathematics in Phase Space (CHAMPS)

University of Bristol, Cardiff University, Imperial College, and the University of Leeds

Predict atomic scalar coupling



Challenges with Science on Kaggle

Challenges

- Reliance on Host (for issues with Leakage, etc.)
- Public Data
  - Earthquake
- "Defeating the Purpose" Data
  - CHAMPS
    - Can't anonymize molecules
- Compute Constraints
  - Speed
  - (Memory, etc.)

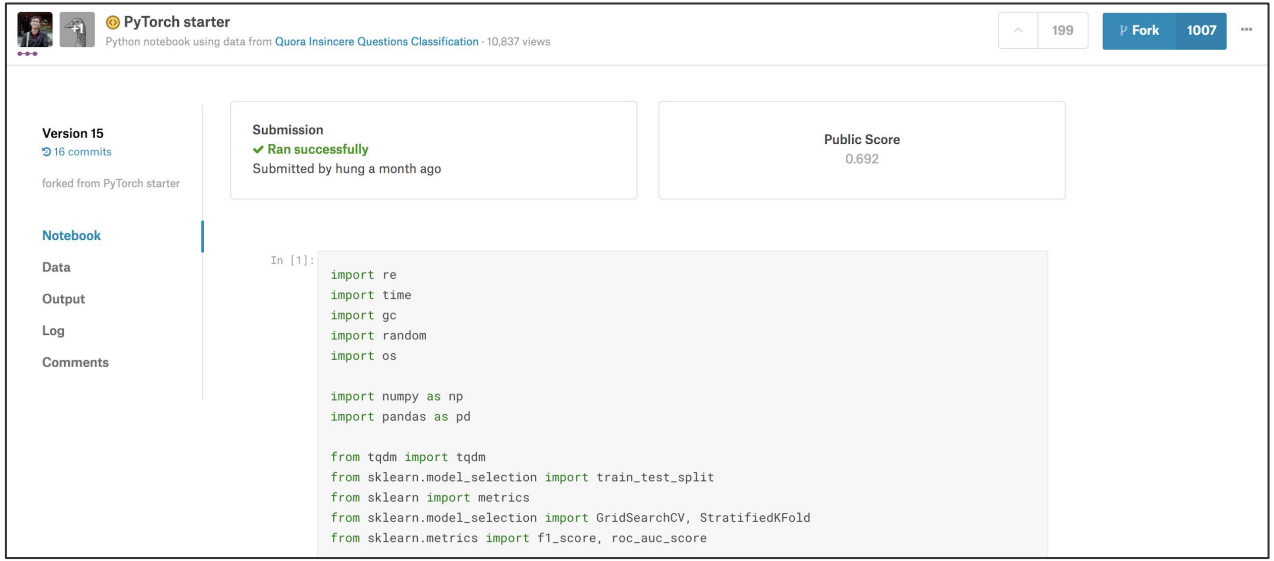
What is the Future of Kaggle?

How Code Competitions Work

Build models in Kaggle Kernels

Collaborate with teammates

Submit code, instead of .CSV predictions



Kernels-only Competitions

True holdout set

- Evaluated on a withheld, private test set, enforcing true model generalization.
- Allows proprietary or sensitive information to be tested, without public release.

Reproducible

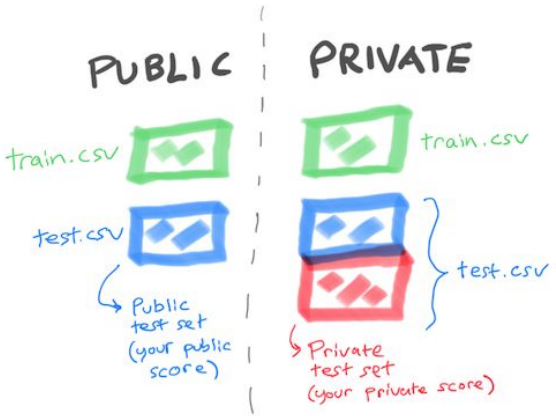
- Models run directly in Kernels with the click of a button or via a docker image for ease of solution transfer.

Equitable & Contained

- Runtime, memory, GPU, and data constraints minimize excessively large ensembles.
- Compute constraints level the playing field for participants.

Important Considerations

- Supports Python and R only.
- Handles medium-to-large datasets, but may not be the best choice for the largest problems.



How It Works

Submission Deadline

Training / Model-Building (in Kernels!)

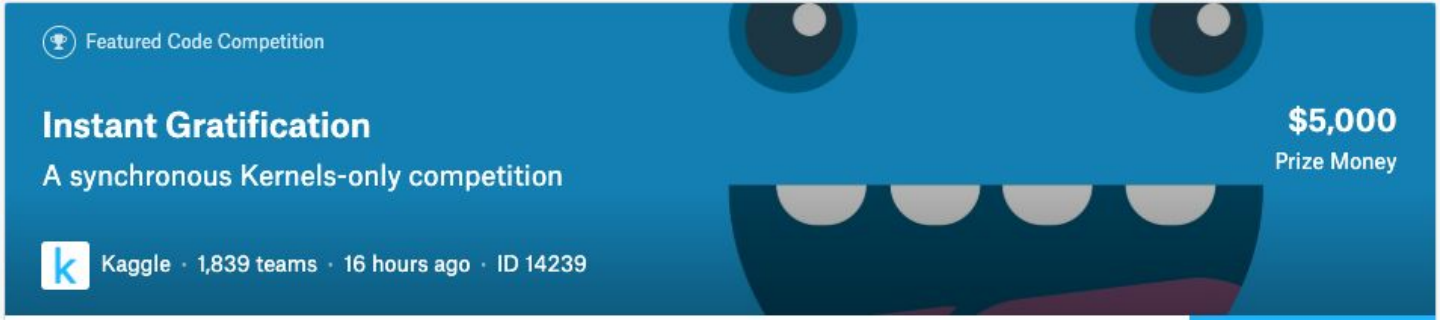
Kernel Output Submission

Public Test Set Feedback

NEW!!

Held-out Private Test Set Bulk Run

Final Leaderboard!



Overview Edit

+ Add Page

Welcome to Instant (well, *almost*) Gratification!

In 2015, Kaggle introduced Kernels as a resource to competition participants. It was a controversial decision to add a code-sharing tool to a competitive coding space. We thought it was important to make Kaggle more than a place where competitions are solved behind closed digital doors. Since then, Kernels has grown from its infancy--essentially a blinking cursor in a docker container--into its teenage years. We now have more compute, longer runtimes, better datasets, GPUs, and an improved interface.

We have iterated and tested several Kernels-only (KO) competition formats with a true holdout test set, in particular deploying them when we would have otherwise substituted a two-stage competition. However, the experience of submitting to a Kernels-only competition has typically been asynchronous and imperfect; participants wait many days after a competition has concluded for their selected Kernels to be rerun on the holdout test dataset, the leaderboard updated, and the winners announced. This flow causes heartbreak to participants whose Kernels fail on the unseen test set, leaving them with no way to correct tiny errors that spoil months of hard work.

Questions