

LIV.INNO

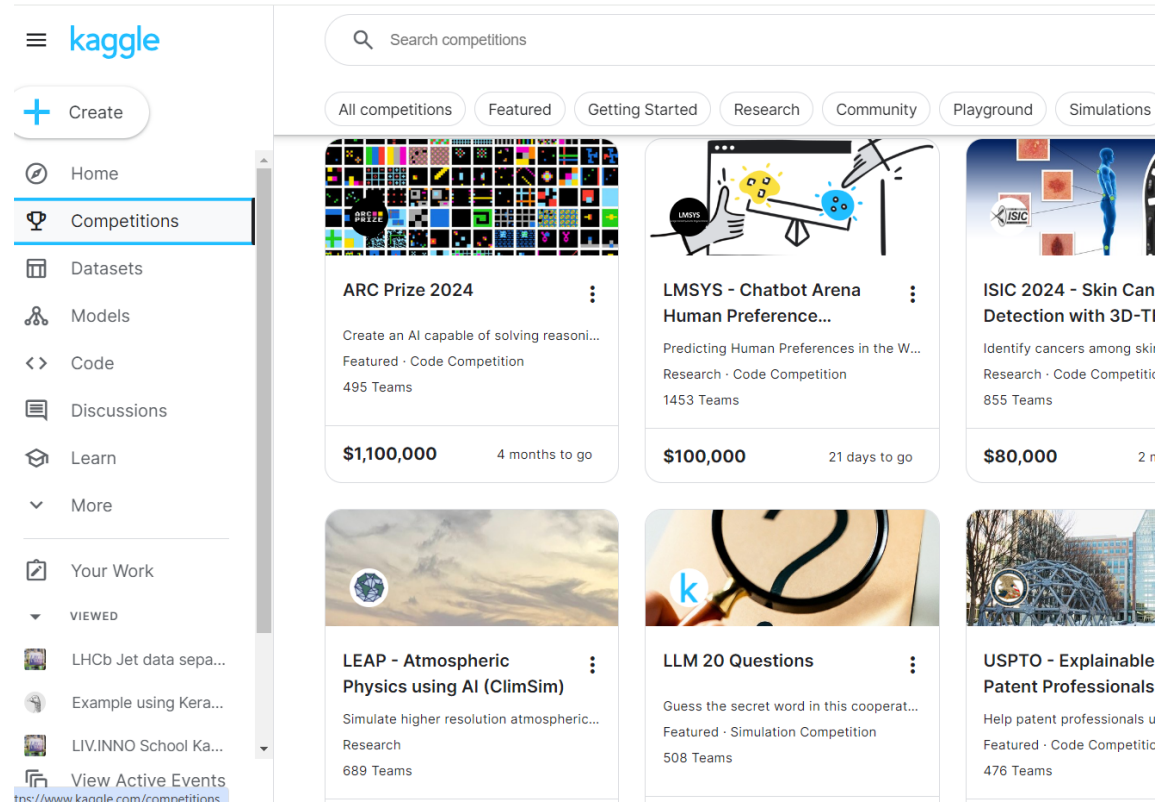
# Kaggle competition

Separating LHCb b-jets from lighter jets



# Kaggle: <https://kaggle.com>

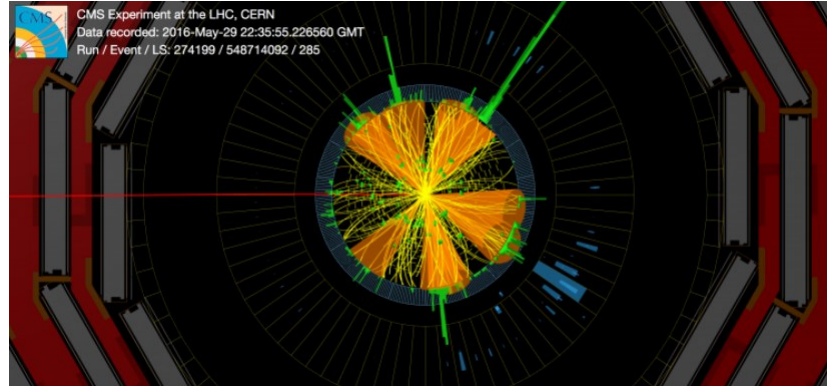
- Kaggle is a website based around competitive category separations
- They host competitions and associated datasets where people are interested in finding patterns
- I set up a private competition for you to try



The screenshot displays the Kaggle website interface. On the left is a sidebar menu with the following items: 'Create', 'Home', 'Competitions' (highlighted), 'Datasets', 'Models', 'Code', 'Discussions', 'Learn', 'More', 'Your Work', and 'VIEWED' (with sub-items: 'LHCb Jet data sepa...', 'Example using Kera...', 'LIV.INNO School Ka...'). Below the menu is a link to 'View Active Events' with the URL 'https://www.kaggle.com/competitions'. The main content area features a search bar and navigation tabs for 'All competitions', 'Featured', 'Getting Started', 'Research', 'Community', 'Playground', and 'Simulations'. A grid of competition cards is shown, including:

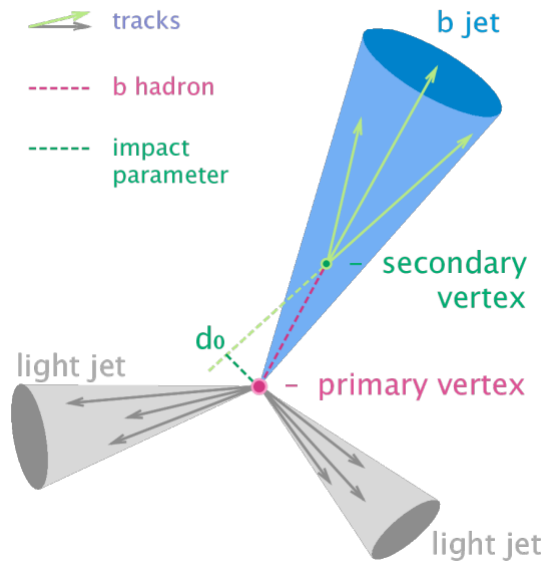
- ARC Prize 2024**: Create an AI capable of solving reason... Featured · Code Competition, 495 Teams, \$1,100,000, 4 months to go.
- LMSYS - Chatbot Arena Human Preference...**: Predicting Human Preferences in the W... Research · Code Competition, 1453 Teams, \$100,000, 21 days to go.
- ISIC 2024 - Skin Cancer Detection with 3D-TI**: Identify cancers among skin... Research · Code Competition, 855 Teams, \$80,000, 21 days to go.
- LEAP - Atmospheric Physics using AI (ClimSim)**: Simulate higher resolution atmospheric... Research, 689 Teams.
- LLM 20 Questions**: Guess the secret word in this cooperat... Featured · Simulation Competition, 508 Teams.
- USPTO - Explainable Patent Professionals**: Help patent professionals u... Featured · Code Competition, 476 Teams.

# Anatomy of a proton-proton collision



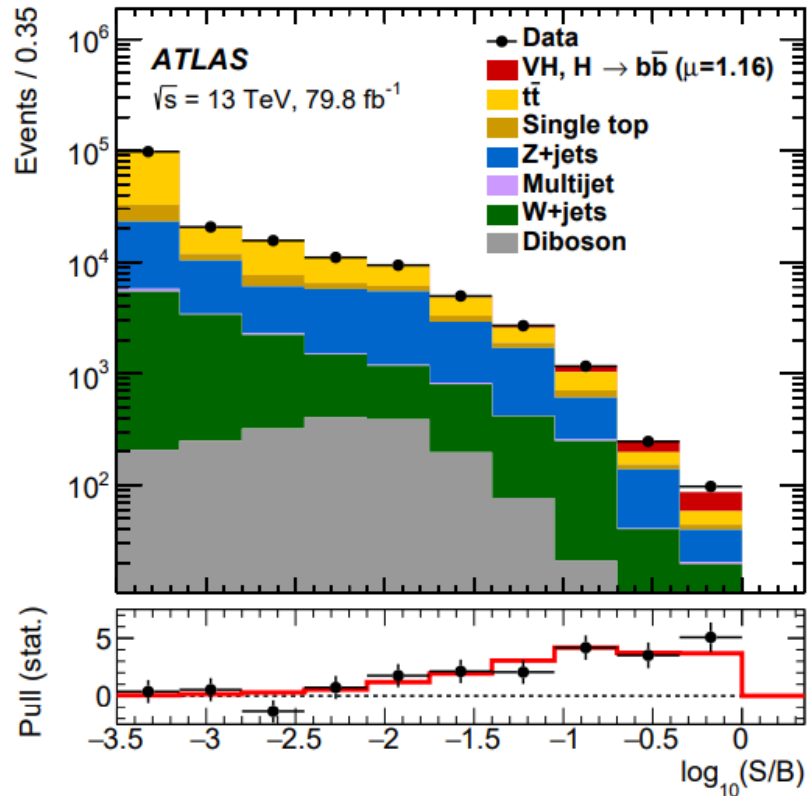
- The collision of protons at high energies at the Large Hadron Collider (LHC) is a messy business...
- For any given measurement, the signal events must be extracted from a huge volume of background events
- A very common signature is a narrow stream of high energy particles, known as a jet
- The study and classification of jets is a core part of the physics program of all experiments

# Quarks, Gluon and Jets



- Quarks and Gluons are fundamental particles that cannot exist in isolation due to the principle of colour confinement
- They subsequently produce “jets” of particles which can be identified and classified
  - The jets arising from the heavy beauty and charm quarks, can be classified as b and c-jets
  - The other jets, which can arise from up, down and strange quarks or gluons, are classified as light jets
- The jets produced by these jets have subtle differences in their substructure that can be used to distinguish them at a particle detector
  - heavy quarks are more massive - larger multiplicity, larger momentum
  - heavy quarks travel a short distance before decaying - can find a “secondary vertex” within the jet

# Why is jet tagging important?



## Example the Higgs Boson

- The most common way a Higgs boson decays is to pairs of b-jets
- However, it was only discovered through this signature in 2018, 6 years after the initial discovery.
- Why? Many processes at the Large Hadron Collider produce jets, both light and heavy
  - The signal is swamped by background events
- Need machine learning techniques to “see” our signal

# Your dataset

- You have simulated datasets of a common process at the LHC, the production of high energy pairs of jets as reconstructed by the LHCb detector, one of the four main detectors on the LHC ring
- The jets have already been pre-selected to contain a secondary vertex, but the abundance of light jets mean that a significant number are still present in any data sample
- Available are observables related to the jet, its constituents, and the properties of the secondary vertex
- Can you separate the  $b$  jets from both  $c$  jets and light jets, to make a measurement at the LHC?

# Competition link

<https://www.kaggle.com/t/2f3ef0f9cd2649fb50edd209c30dac88>

- The link above is needed to enter the competition (see also the email I sent you)
- The data is available from the competition data tab and also a copy is hosted here: <https://hep.ph.liv.ac.uk/~hutchcroft/JetTagging/> (that has the same csv files).
- There is an example (unoptimized) NN solution using Keras under public notebooks, also on the local link above, Kaggle will run your Jupyter notebooks on their hosts <https://www.kaggle.com/code/davidhutchcroft/example-using-keras-submit-solution>
- Do the best separation your team can, 30% of the data is used for a public leaderboard, the other 70% is hidden and used to rank your submissions for the final tally
- You have up to 25 submissions per team, the deadline for submissions is 12:35 today
- Good luck!

# Competition results:

Private Leaderboard