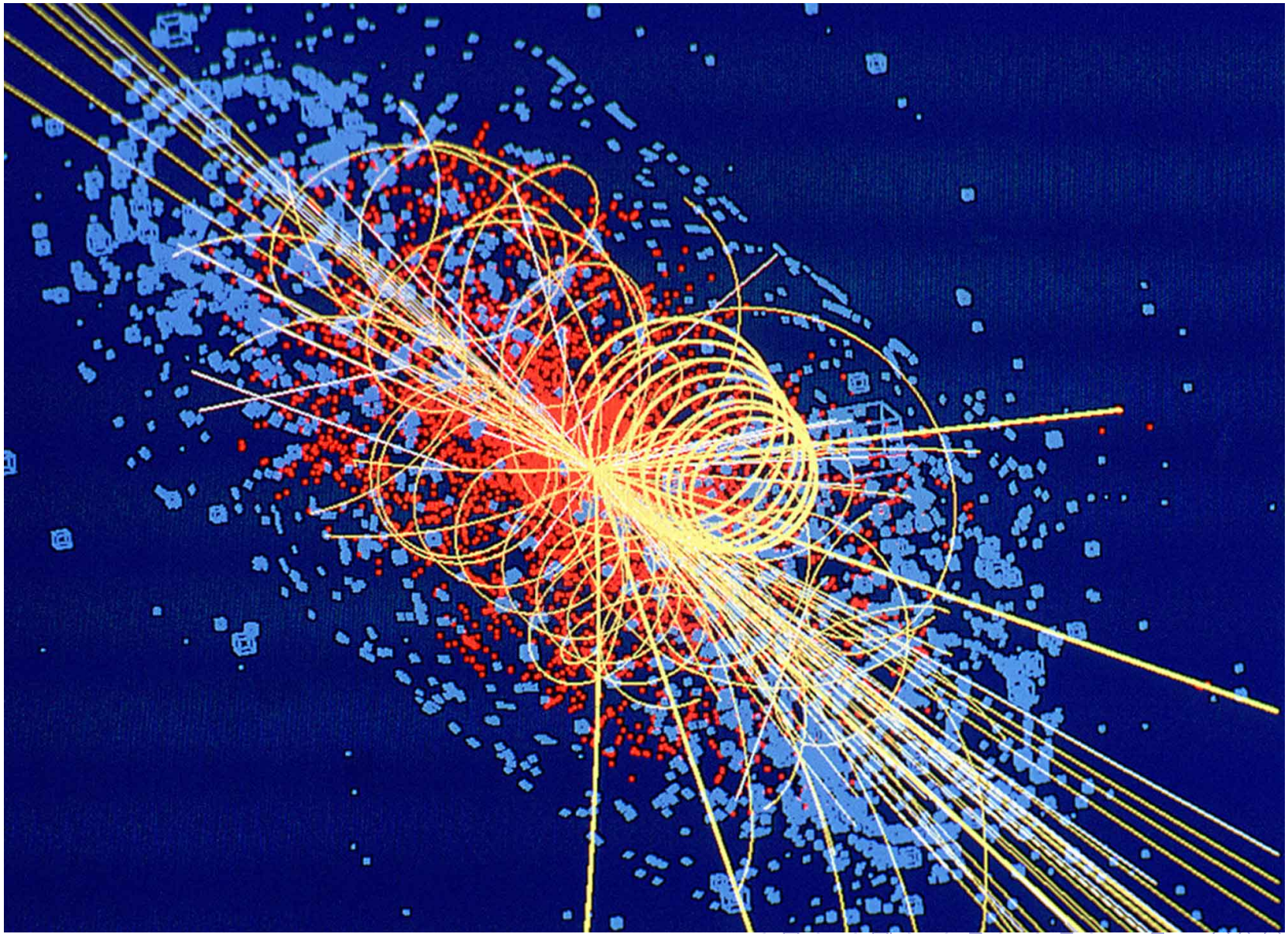
From Description to Ensemble

Robert Castellano, Yannick Kimmel, Wanda Wang, Ho Fai Wong



First section title

Models

Our models

- Random forest

Our models

- Random forest
- Gbm

Our models

- Random forest
- Gbm
- Xgboost

Random forest model

- Tuning parameters
  - mtry: Number of splits per tree

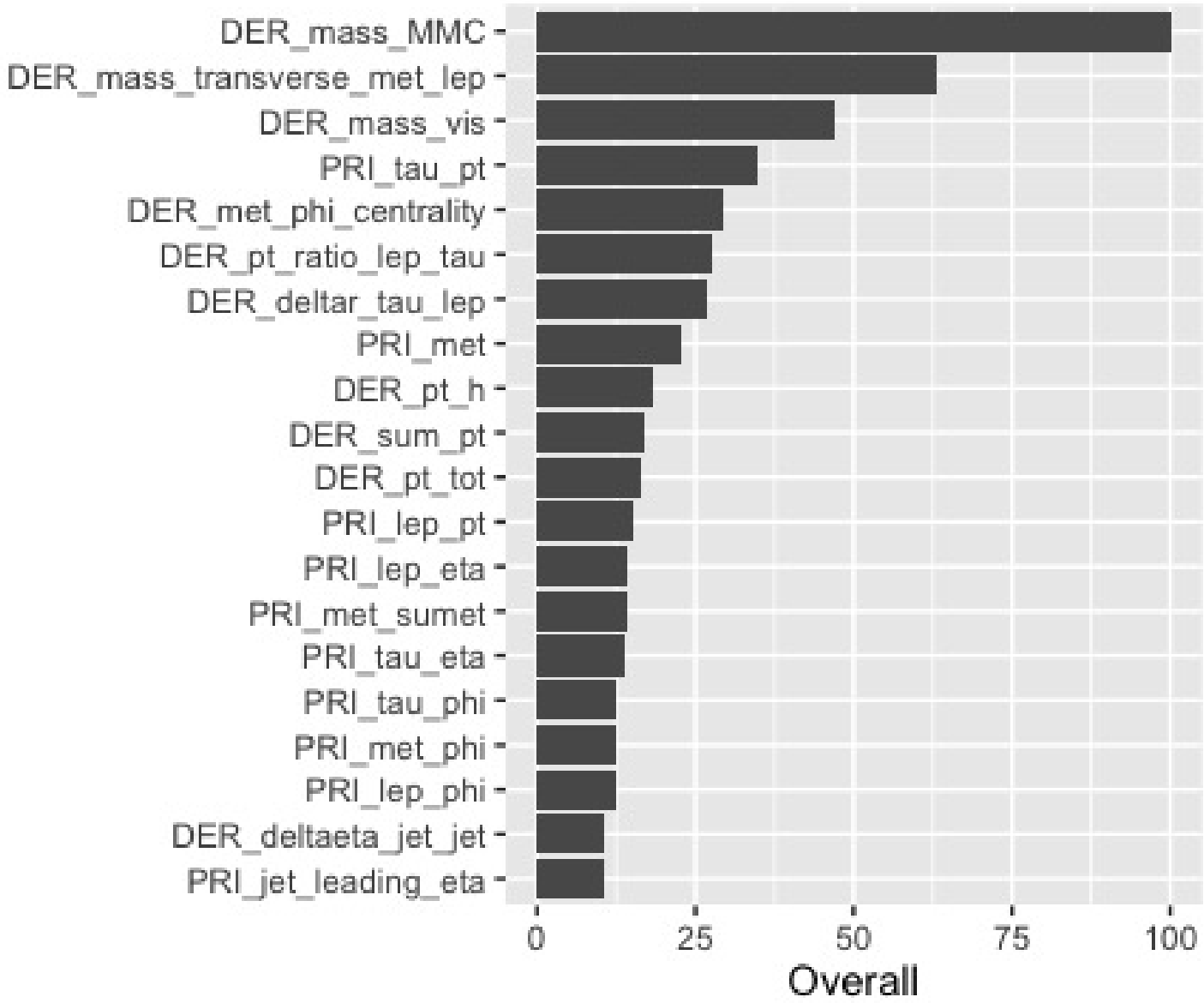
Random forest model

- Tuning parameters
  - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
  - mtry $= 5$

Random forest model

- Tuning parameters
  - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
  - mtry $= 5$
- AUC on training data $= .9071$
- Kaggle rank $= 1311$
- AMS $= 2.57949$

Random forest variable importance



About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters
- Not completely greedy in tree creation

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters
- Not completely greedy in tree creation
- Generally faster and performs better than gbm.

Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate

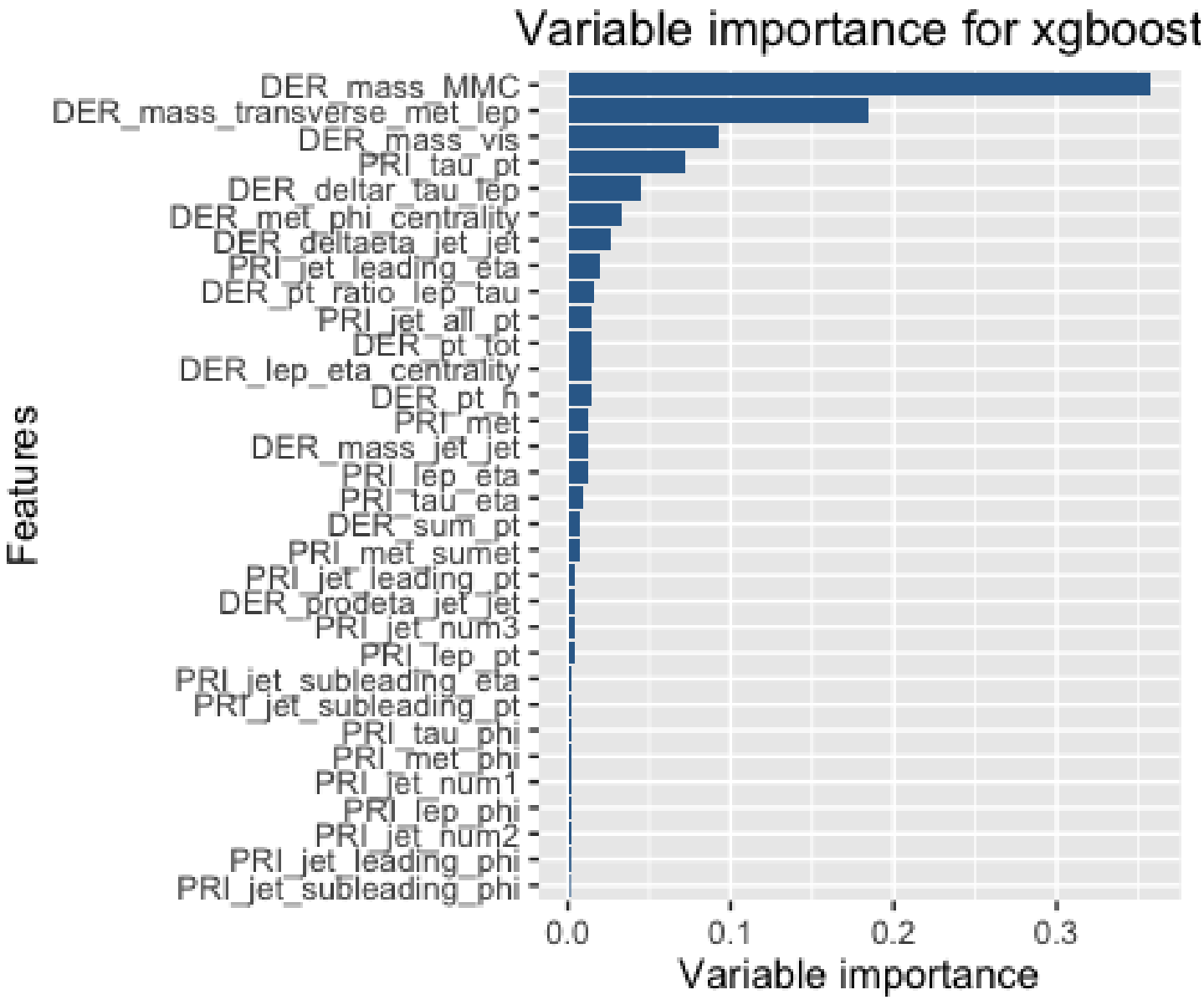
Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate
- Performed 5-fold CV to tune parameters.
  - nrounds = 200
  - max_depth = 5
  - colsample_bytree = .85
  - eta = .2

Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate
- Performed 5-fold CV to tune parameters.
  - nrounds = 200
  - max_depth = 5
  - colsample_bytree = .85
  - eta = .2
- AUC on training data = .9254
- Kaggle rank = 1340
- AMS = 2.49958

Xgboost variable importance



Ensemble

- Combined three models by majority vote

Ensemble

- Combined three models by majority vote
- Kaggle rank $= 1309$

Ensemble

- Combined three models by majority vote
- Kaggle rank $= 1309$
- AMS $= 2.58510$

Room for improvement

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physcics: e.g. CAKE variable

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physcics: e.g. CAKE variable
  - Better understand the physics of additional models

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physcics: e.g. CAKE variable
  - Better understand the physics of additional models
- Log transforms

Models

- More models

Models

- More models
- More sophisticated emsemble

Models

- More models
- More sophisticated emsemble
- Run different random seeds for the same model