

FYS9555 Final project

«Higgs boson production mechanism classification from ATLAS Open Data»

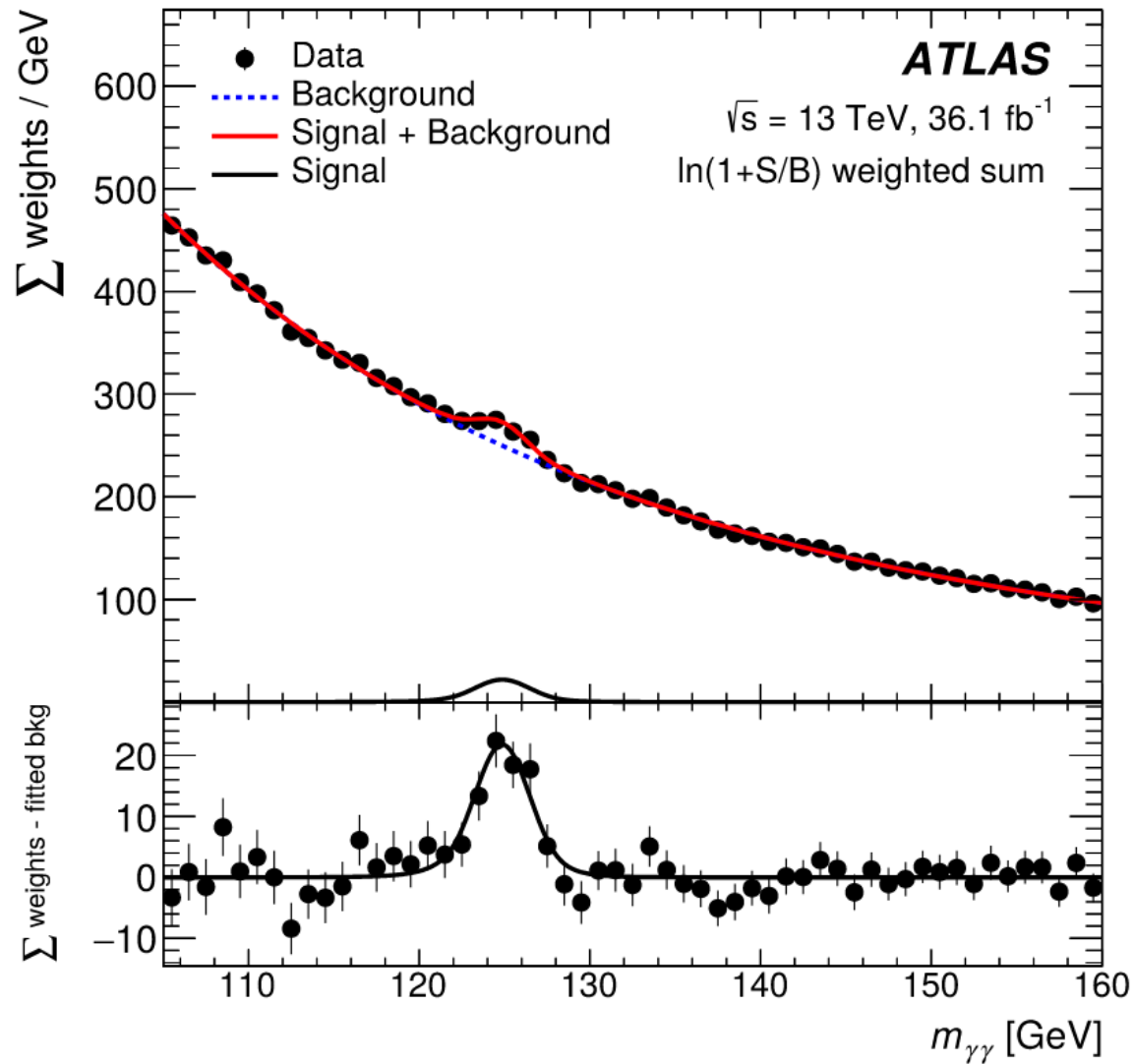
10 June 2020

Victor Ananyev

Outline

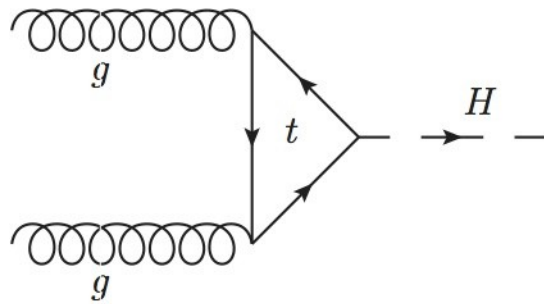
- Di-photon process
- Data acquisition
- What's inside?
- Dealing with jagged data in ML
- Fighting imbalanced data sets
- Fitting baseline ML models

Di-photon process

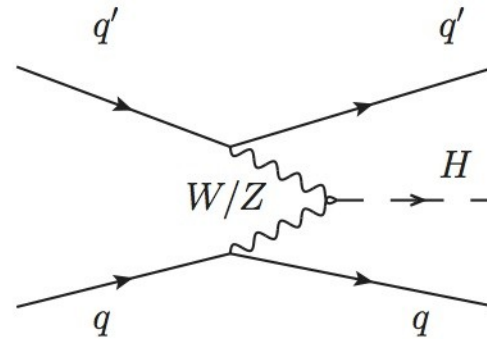


(borrowed from <https://arxiv.org/abs/1806.00242>)

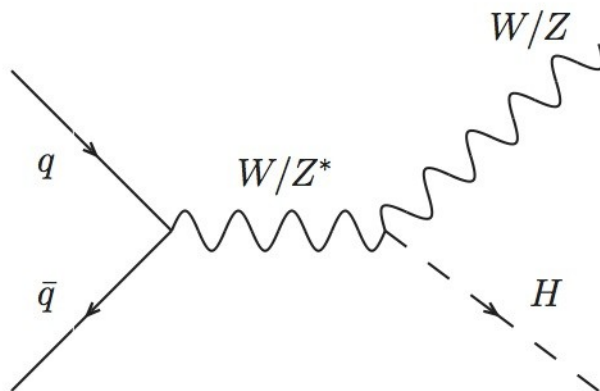
Major production mechanisms



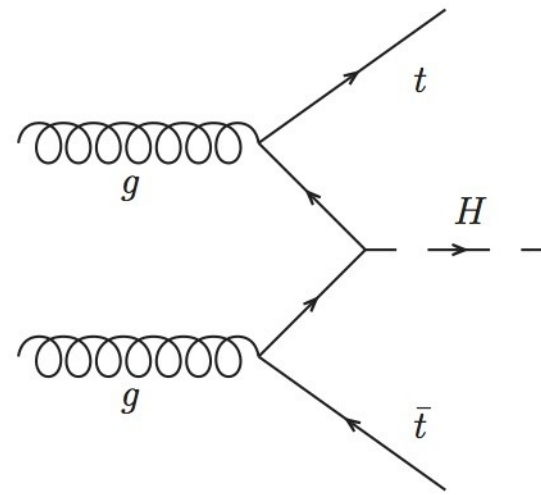
a)



b)



c)



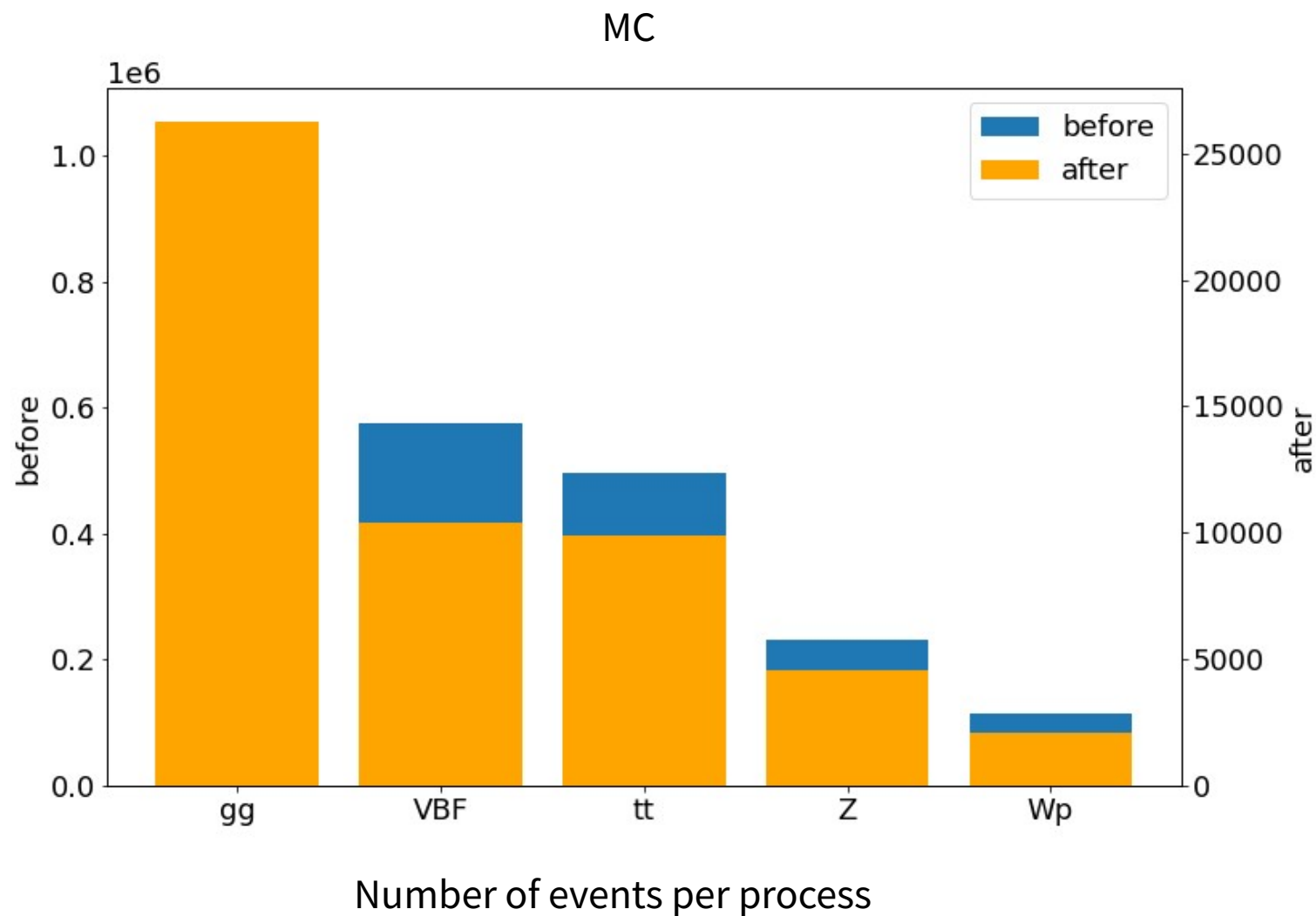
d)

(borrowed from <https://grab-group.ethz.ch/research/higgs-boson.html>)

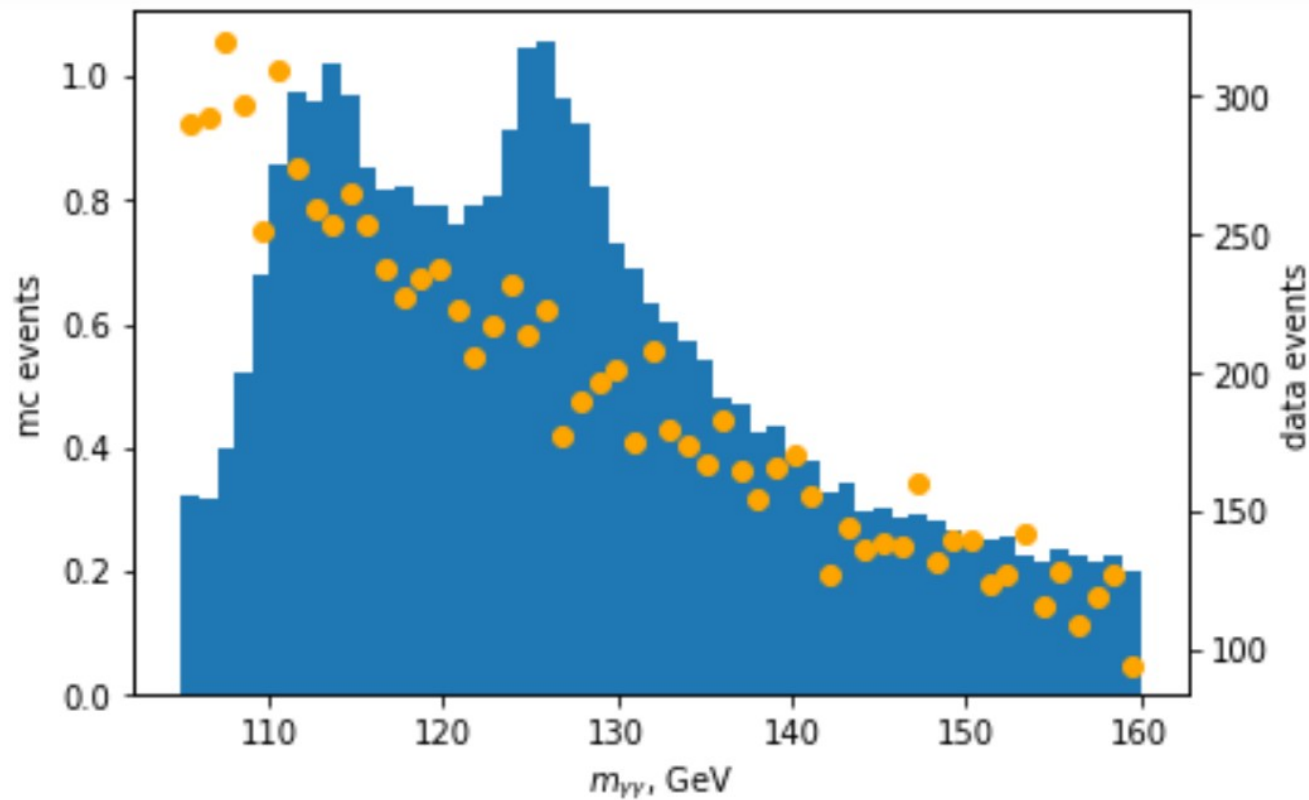
Event selection for $H\gamma\gamma$ peak

- Diphoton trigger is satisfied;
- Exactly two photons with $E_T > 35$ and 25 GeV, respectively;
- Leading and subleading photon candidates are respectively required to have $E_T/m_{\gamma\gamma} > 0.35$ and 0.25;
- Diphoton invariant mass $m_{\gamma\gamma}$ between 105 GeV and 160 GeV.

Selection efficiency ~ 1%

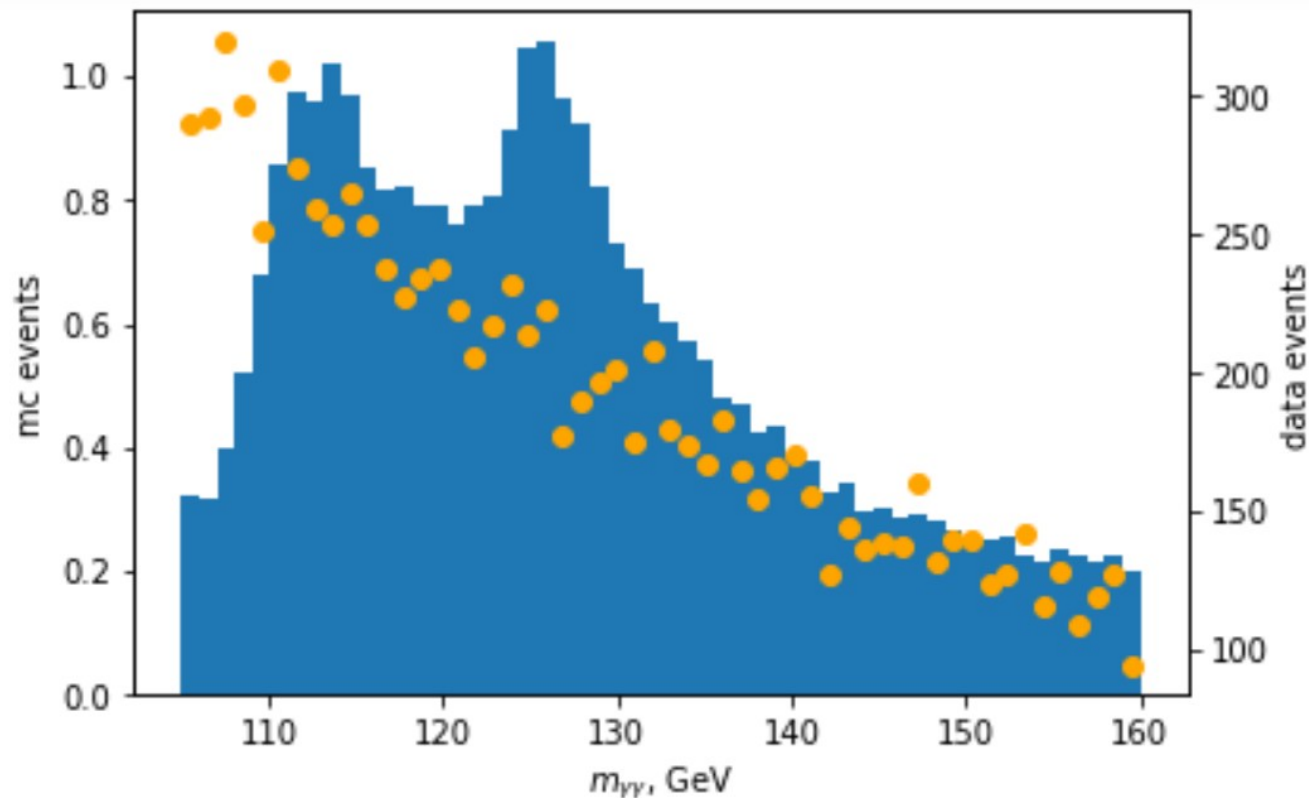


Higgs mass from ATLAS Open Data



8M events in 7 minutes!

Acquire data efficiently



8M events in 7 minutes **with Python!**

Data pipeline



1TB Root files storage

Data pipeline



1TB Root files storage



Processing power and data traffic



Data pipeline



1TB Root files storage



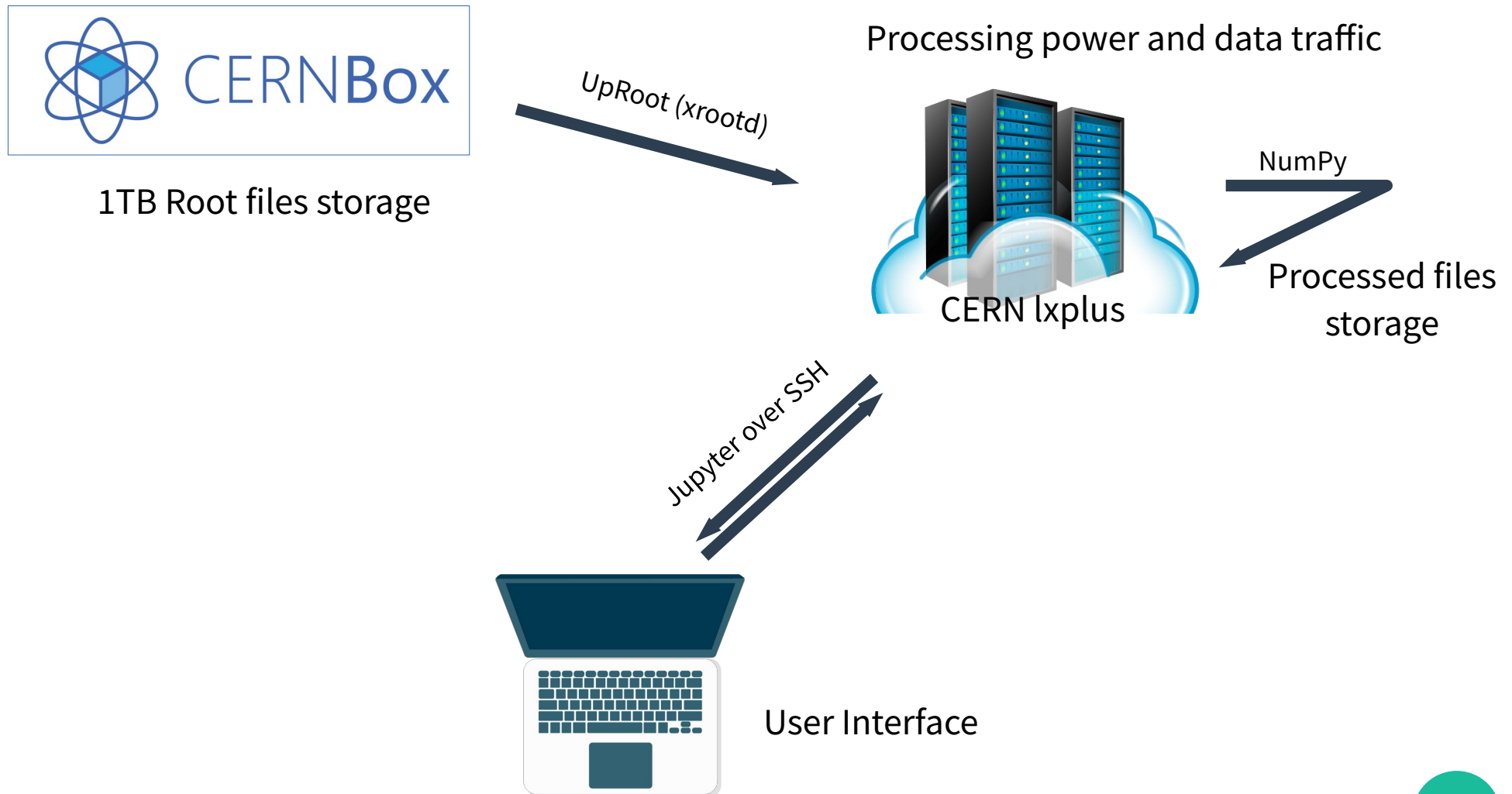
Processing power and data traffic



NumPy

Processed files
storage

Data pipeline



Data pipeline

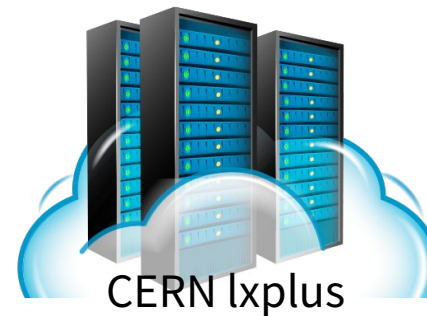
Processing power and data traffic

1TB Root files storage

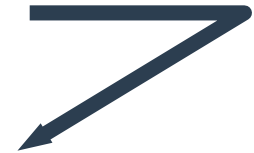


2GB – 100 GB

UpRoot (xrootd)



NumPy



Batches 200MB
~ 100 000 events

Store processed
~ 2MB
~1000 events/batch

What's inside?

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features
1 number per event

- | | | | | |
|-----------------------|----------------------------|--------------------|--------------------|--------------------------|
| • photon_n | • lep_n | • tau_n | • jet_n | • largeRjet_n |
| • photon_truthMatched | • lep_truthMatched | • tau_pt | • jet_pt | • largeRjet_pt |
| • photon_trigMatched | • lep_trigMatched | • tau_eta | • jet_eta | • largeRjet_eta |
| • photon_pt | • lep_pt | • tau_phi | • jet_phi | • largeRjet_phi |
| • photon_eta | • lep_eta | • tau_E | • jet_E | • largeRjet_E |
| • photon_phi | • lep_phi | • tau_isTightID | • jet_jvt | • largeRjet_m |
| • photon_E | • lep_E | • tau_truthMatched | • jet_trueflav | • largeRjet_truthMatched |
| • photon_isTightID | • lep_z0 | • tau_trigMatched | • jet_truthMatched | • largeRjet_D2 |
| • photon_ptcone30 | • lep_charge | • tau_nTracks | • jet_MV2c10 | • largeRjet_tau32 |
| • photon_etcone20 | • lep_type | • tau_BDTid | • jet_pt_syst | • largeRjet_pt_syst |
| • photon_convType | • lep_isTightID | • tau_pt_syst | | |
| • photon_pt_syst | • lep_ptcone30 | • tau_charge | | |
| | • lep_etcone20 | | | |
| | • lep_trackd0pvunbiased | | | |
| | • lep_tracksigd0pvunbiased | | | |
| | • lep_pt_syst | | | |

Micro features
*Jagged. List of numbers
per event*

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...]

MASK

1000 events

800

**Photon
Features**

- | | | | | |
|-----------------------|----------------------------|--------------------|--------------------|--------------------------|
| • photon_n | • lep_n | • tau_n | • jet_n | • largeRjet_n |
| • photon_truthMatched | • lep_truthMatched | • tau_pt | • jet_pt | • largeRjet_pt |
| • photon_trigMatched | • lep_trigMatched | • tau_eta | • jet_eta | • largeRjet_eta |
| • photon_pt | • lep_pt | • tau_phi | • jet_phi | • largeRjet_phi |
| • photon_eta | • lep_eta | • tau_E | • jet_E | • largeRjet_E |
| • photon_phi | • lep_phi | • tau_isTightID | • jet_jvt | • largeRjet_m |
| • photon_E | • lep_E | • tau_truthMatched | • jet_trueflav | • largeRjet_truthMatched |
| • photon_isTightID | • lep_z0 | • tau_trigMatched | • jet_truthMatched | • largeRjet_D2 |
| • photon_cone30 | • lep_charge | • tau_nTracks | • jet_MV2c10 | • largeRjet_tau32 |
| • photon_cone20 | • lep_type | • tau_BDTid | • jet_pt_syst | • largeRjet_pt_syst |
| • photon_Type | • lep_isTightID | • tau_pt_syst | | |
| • photon_syst | • lep_ptcone30 | • tau_charge | | |
| | • lep_etcone20 | | | |
| | • lep_trackd0pvunbiased | | | |
| | • lep_tracksigd0pvunbiased | | | |
| | • lep_pt_syst | | | |

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...]

MASK

MASK

1000 events

800

600

**Photon
Features**

**Lepton
Features**

unbiased
pvunbiased

- jet_n
- jet_pt
- jet_eta
- jet_phi
- jet_E
- jet_jvt
- jet_trueflav
- jet_truthMatched
- jet_MV2c10
- jet_pt_syst
- largeRjet_n
- largeRjet_pt
- largeRjet_eta
- largeRjet_phi
- largeRjet_E
- largeRjet_m
- largeRjet_truthMatched
- largeRjet_D2
- largeRjet_tau32
- largeRjet_pt_syst

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

$$[0, 0, 1, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 0 \dots]$$

MASK

MASK

MASK

1000 events

250 events

800

600

300

Photon Features

Lepton Features

Tau Features

Micro features

Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

$[0, 0, 1, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 0 \dots]$

MASK

MASK

MASK

1000 events

250 events

800

600

250

Photon Features

Lepton Features

Tau Feature

Micro features

MASK

Mask bubbling for feature extraction

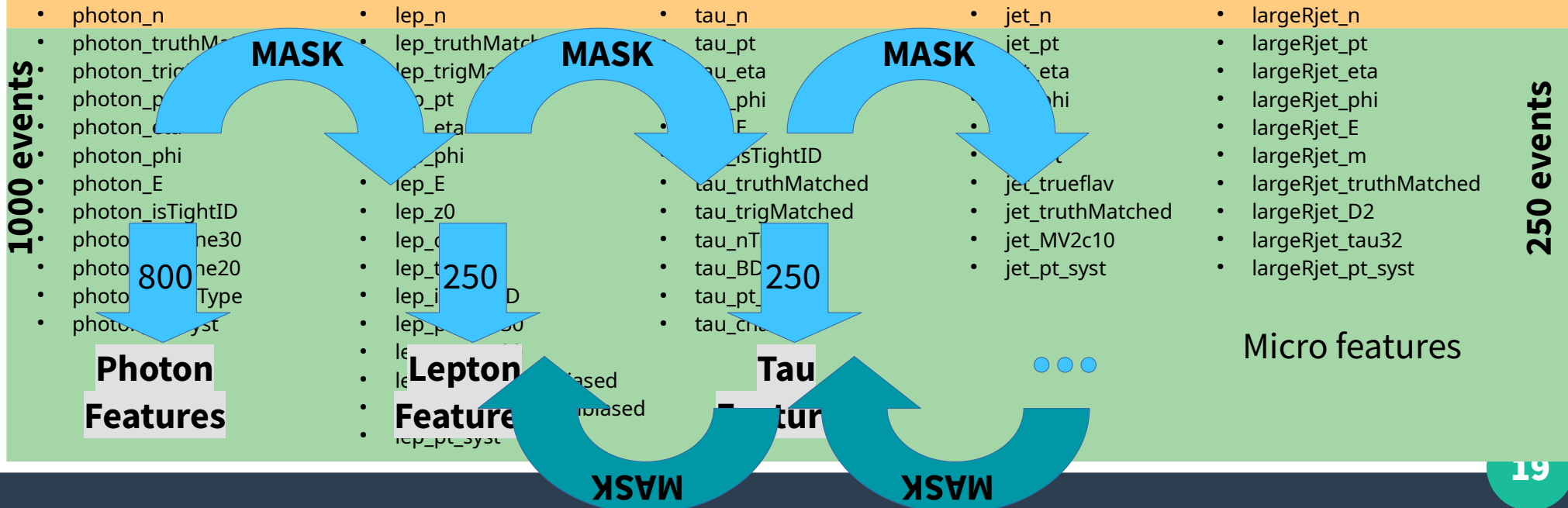
- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

[0, 0, 1, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 1 ...] > [0, 0, 0, 1, 0, 0 ...]



Mask bubbling for feature extraction

- XSection
- SumWeights

Simulation wide

- runNumber
- eventNumber
- channelNumber
- mcWeight
- scaleFactor_*
- trig*
- met_et
- met_et_syst
- met_phi
- ditau_m

Macro features

$$[0, 0, 1, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 1 \dots] > [0, 0, 0, 1, 0, 0 \dots]$$

-
- The diagram illustrates the relationship between input, intermediate, and output variables. It consists of three columns of variables. The first column (input) has 10 variables. The second column (intermediate) has 10 variables. The third column (output) has 10 variables. Three blue curved arrows labeled "MASK" connect the input variables to the intermediate variables. Three blue straight arrows labeled "250" connect the intermediate variables to the output variables.
- | Input Variables | Intermediate Variables | Output Variables |
|---------------------|------------------------|------------------------|
| photon_n | lep_n | largeRjet_n |
| photon_truthMatched | lep_truthMatched | largeRjet_pt |
| photon_trigMatched | lep_trigMatched | largeRjet_eta |
| photon_pt | lep_pt | largeRjet_phi |
| photon_eta | lep_eta | largeRjet_E |
| photon_phi | lep_phi | largeRjet_m |
| photon_E | lep_E | largeRjet_truthMatched |
| photon_isTightID | lep_z0 | largeRjet_D2 |
| photon_tau30 | lep_tau30 | largeRjet_tau32 |
| photon_tau20 | lep_tau20 | largeRjet_pt_syst |
| photon_Type | lep_tauType | |
| photon_syst | lep_tau_syst | |

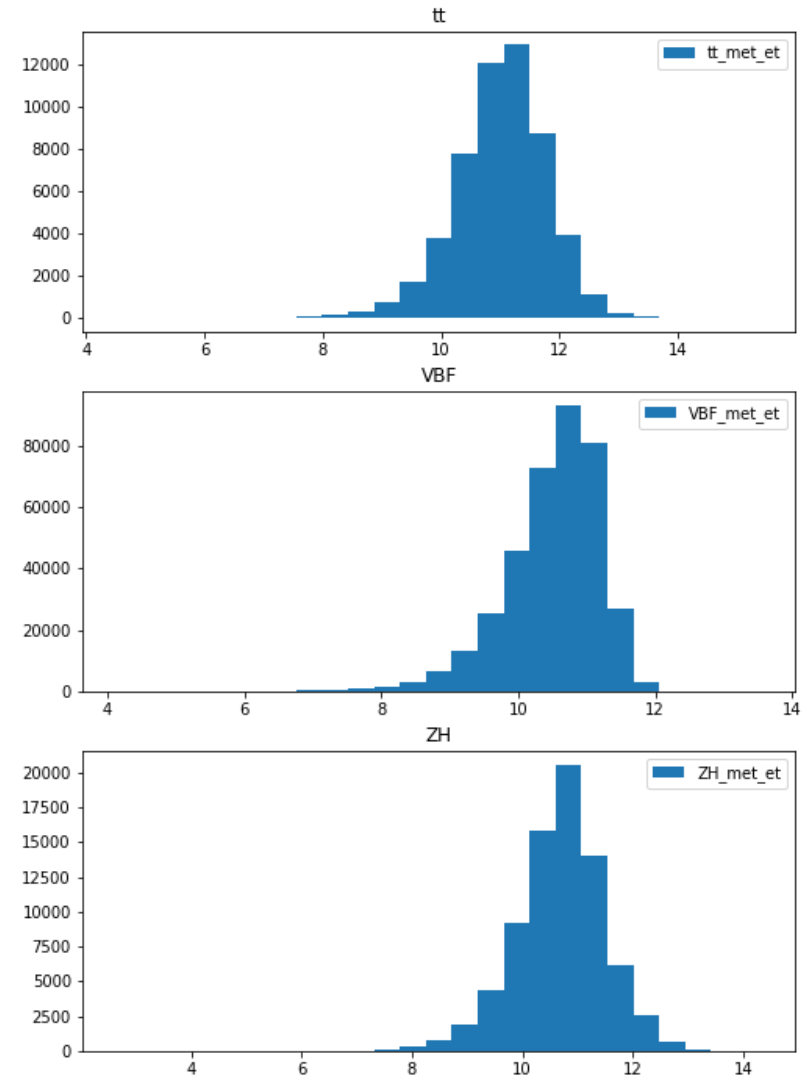
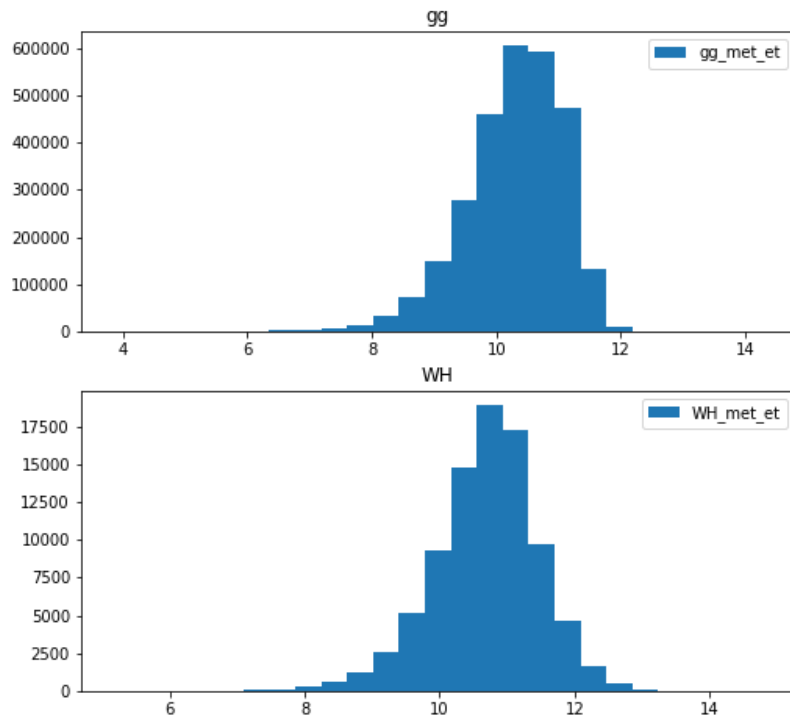
Micro features

1000 events

250 events

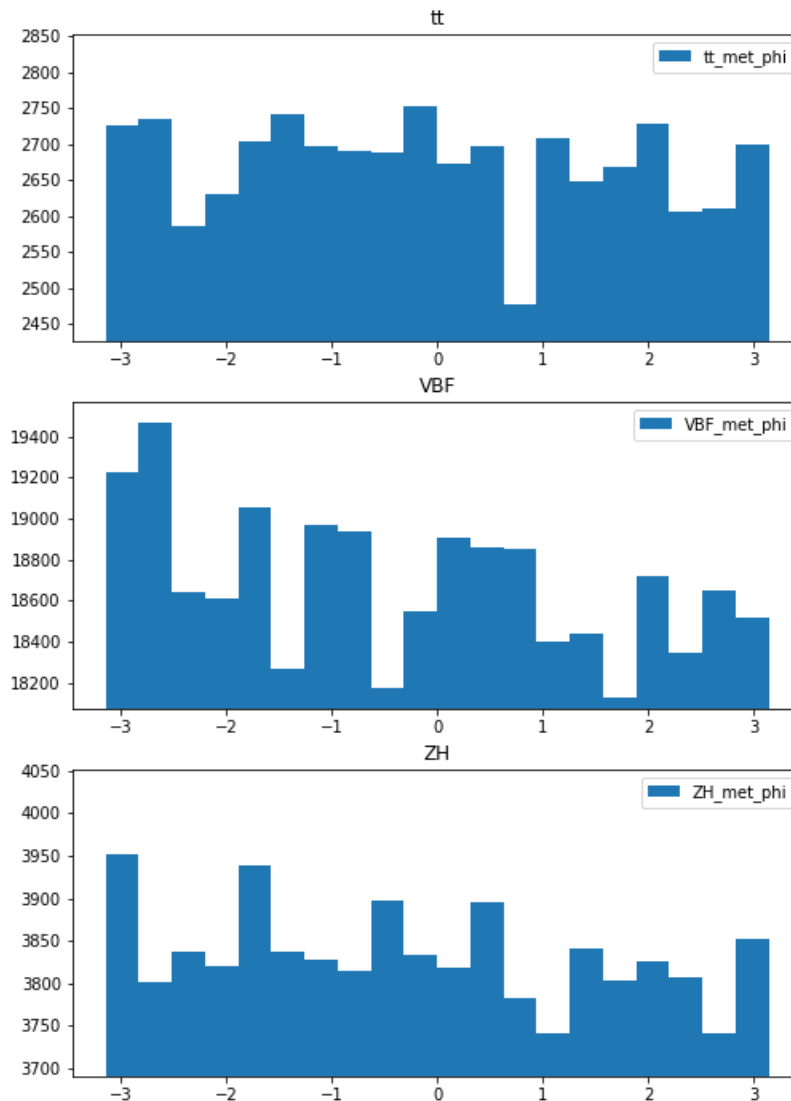
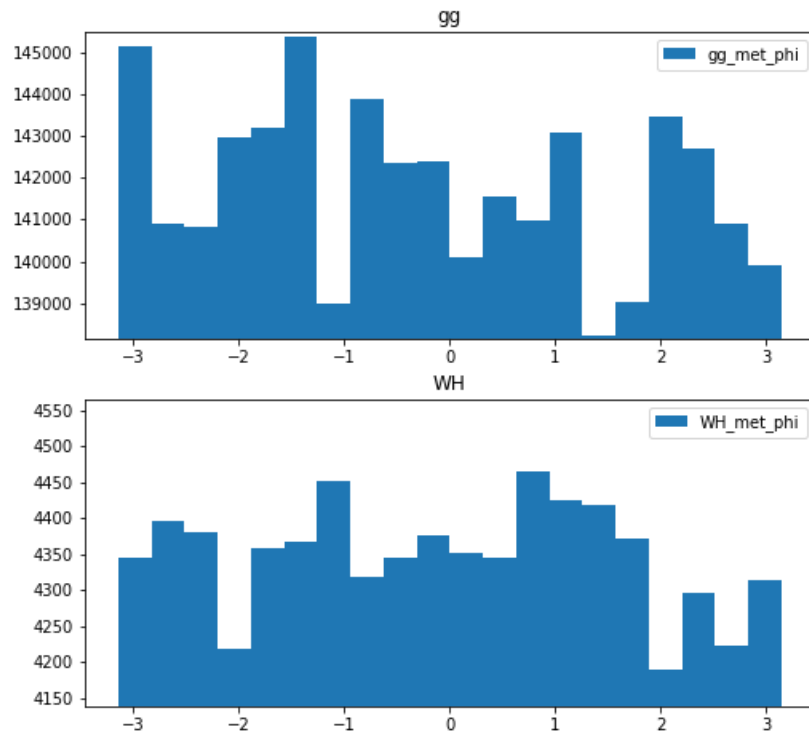
Missing energy E_t (log x scale)

Within 100 000 events batch



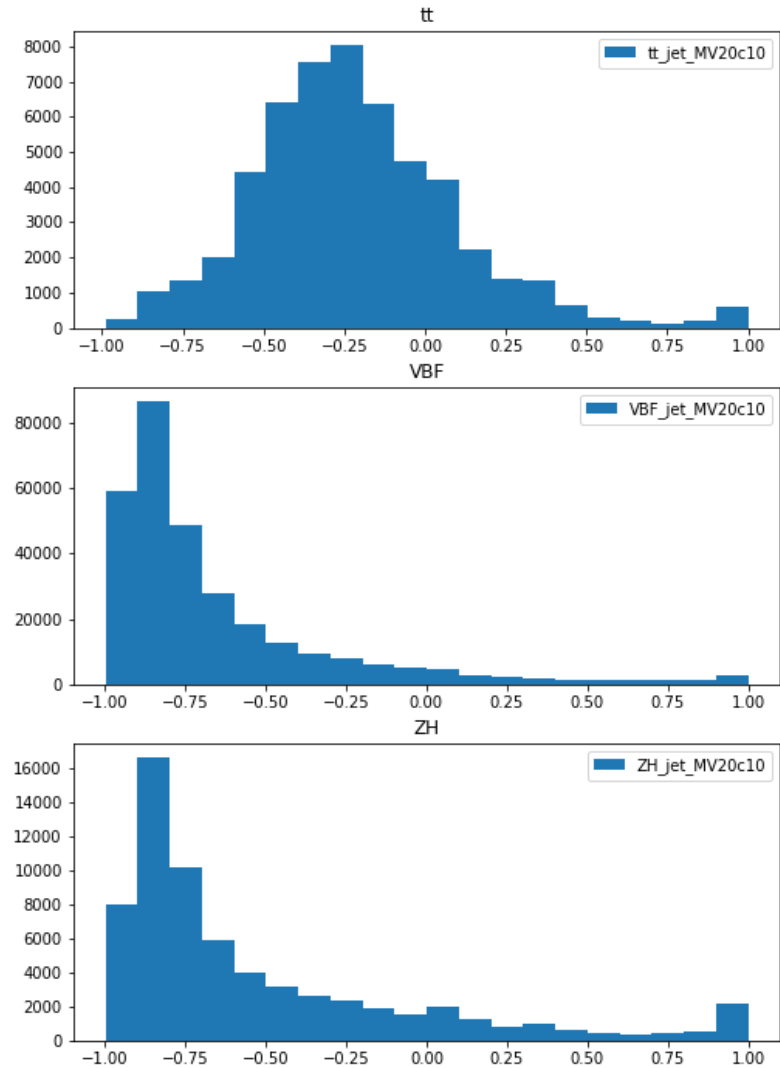
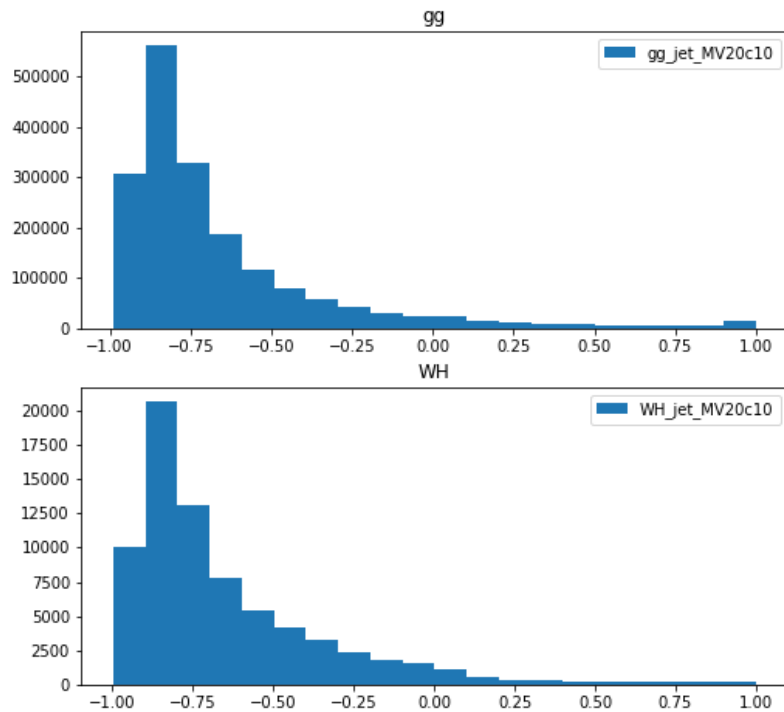
Missing energy Phi

Within 100 000 events batch



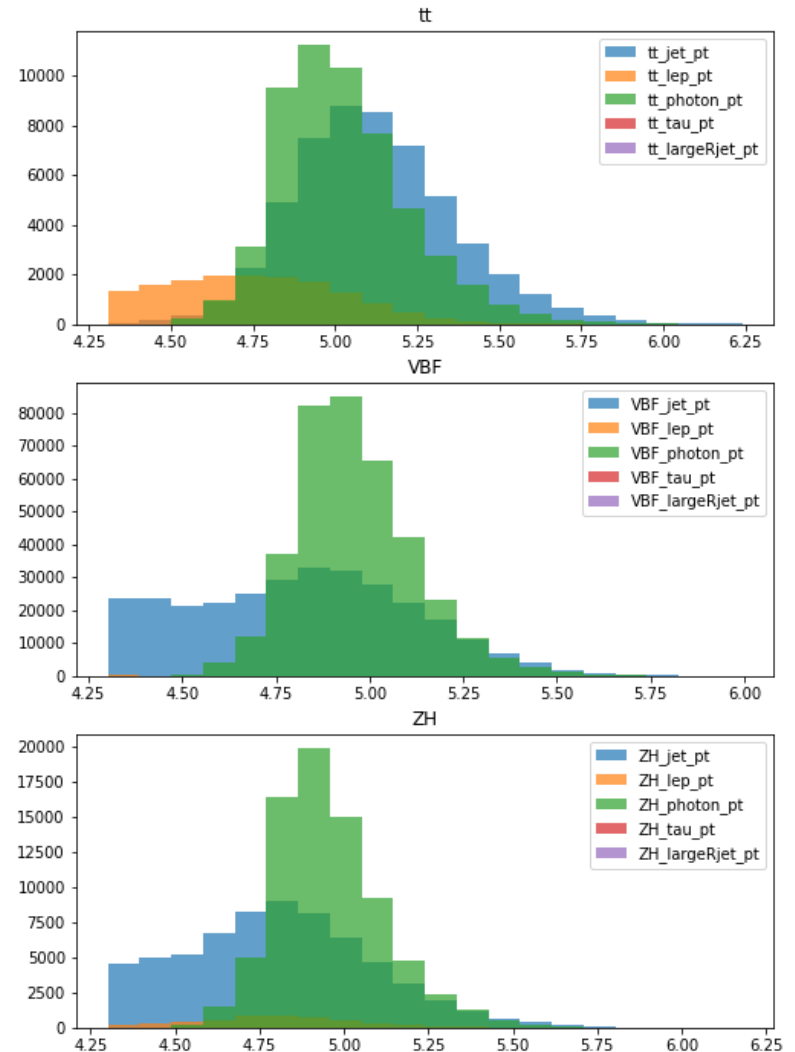
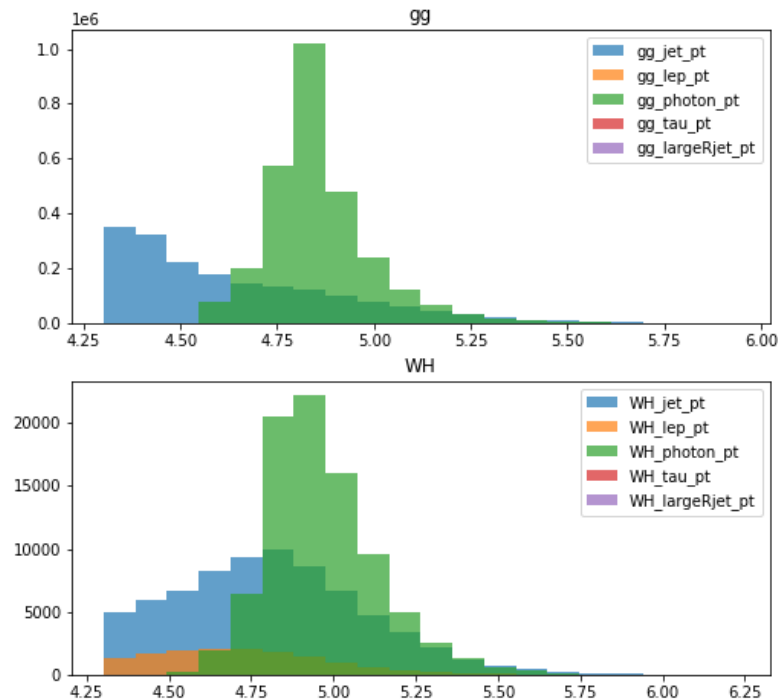
Jet b-tagging score

Within 100 000 events batch



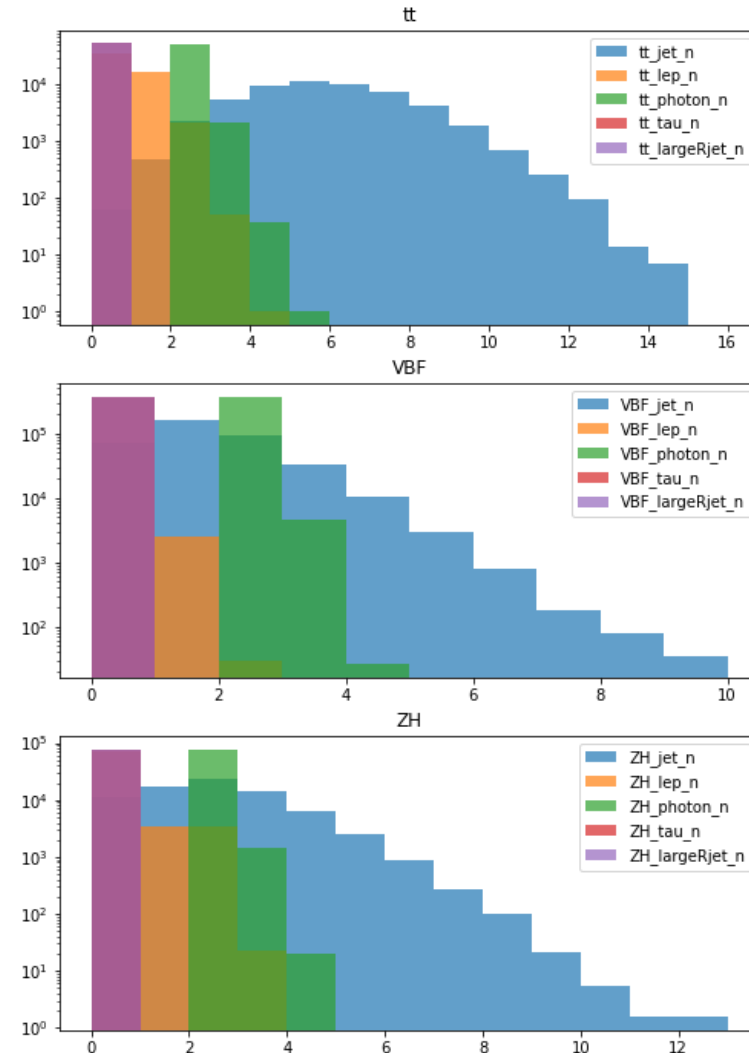
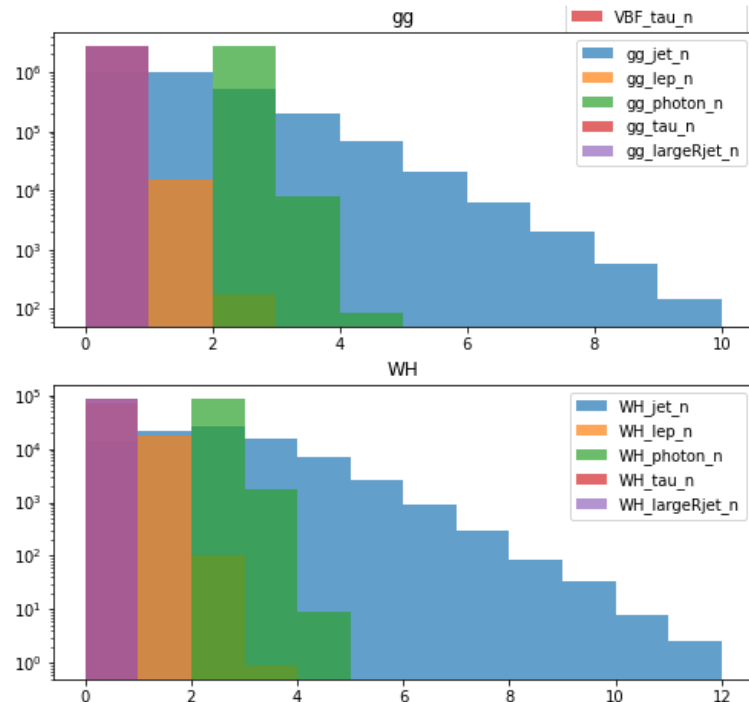
Max pT per particle type (log x scale)

Within 100 000 events batch

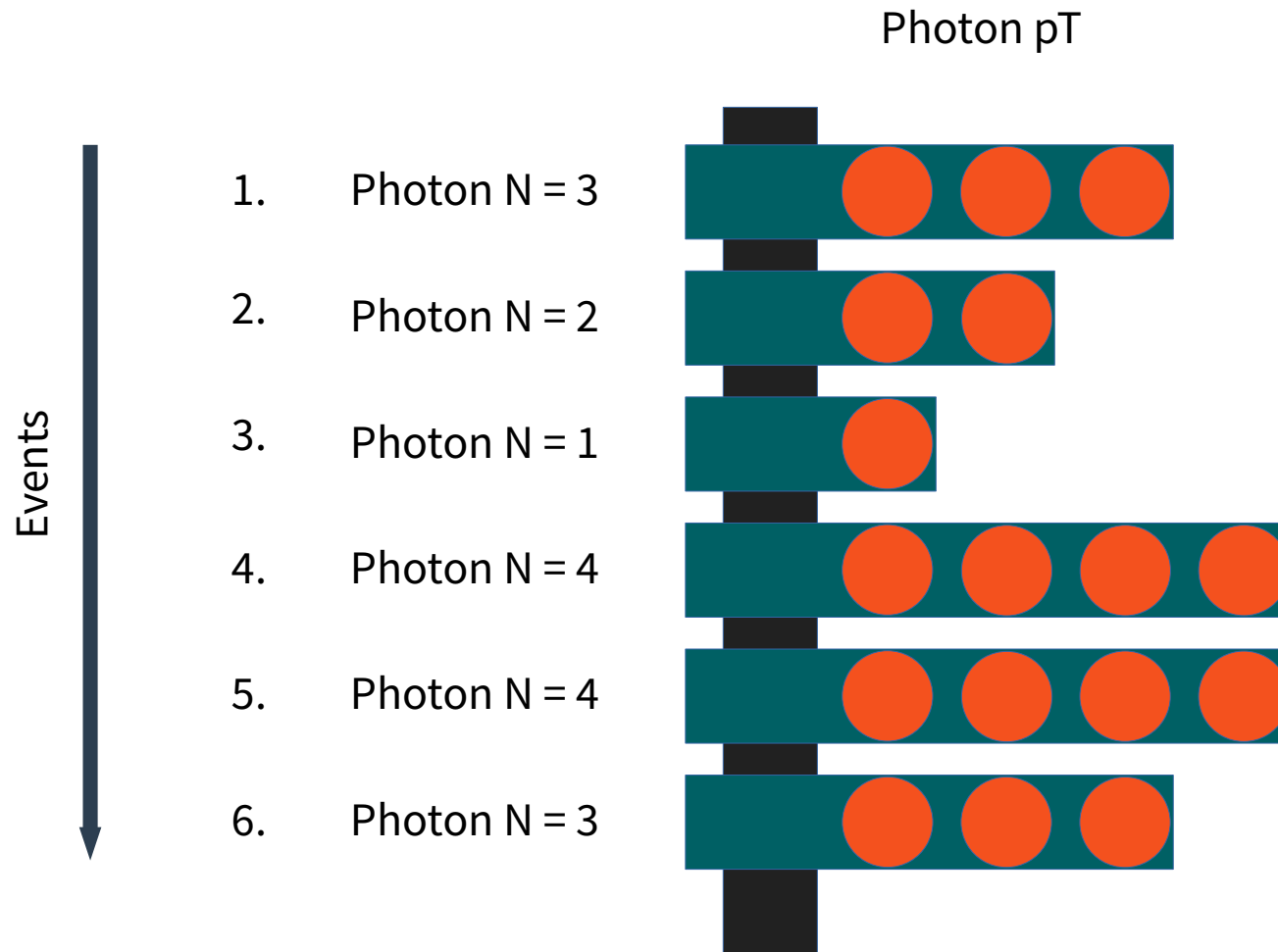


Number of particles (log y scale)

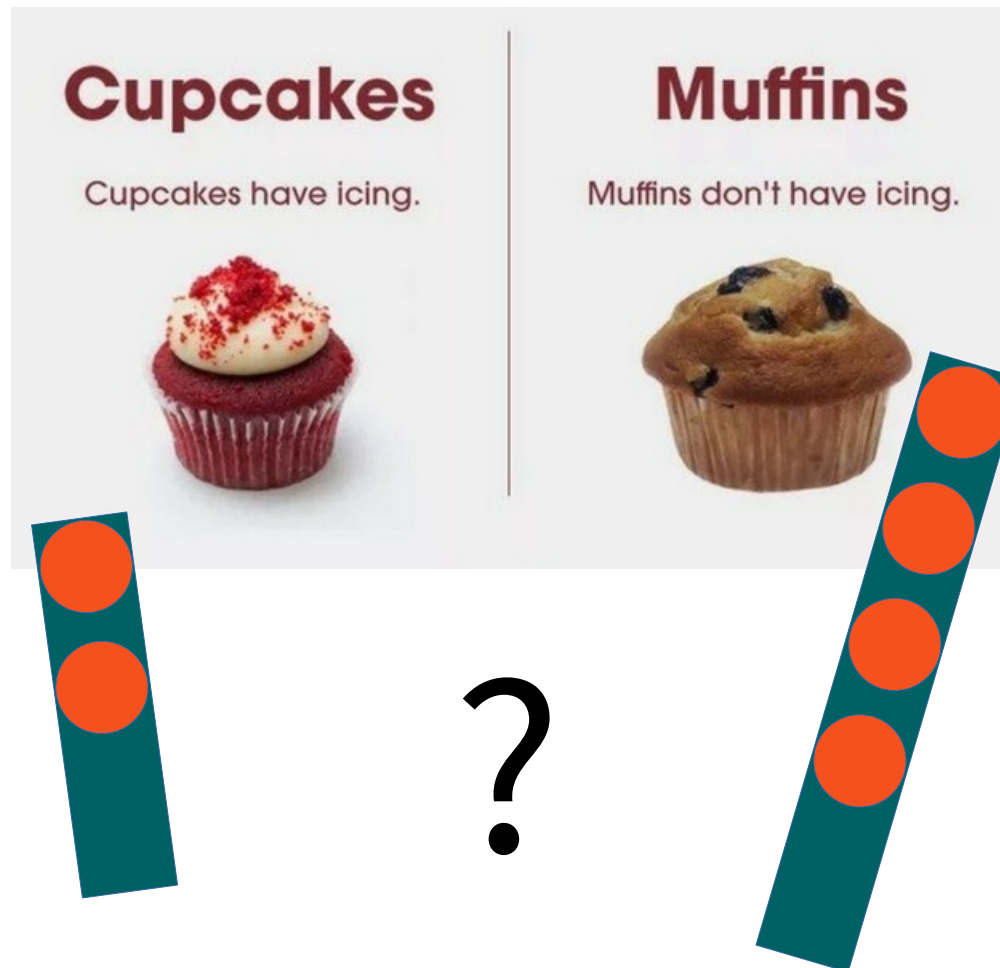
Within 100 000 events batch



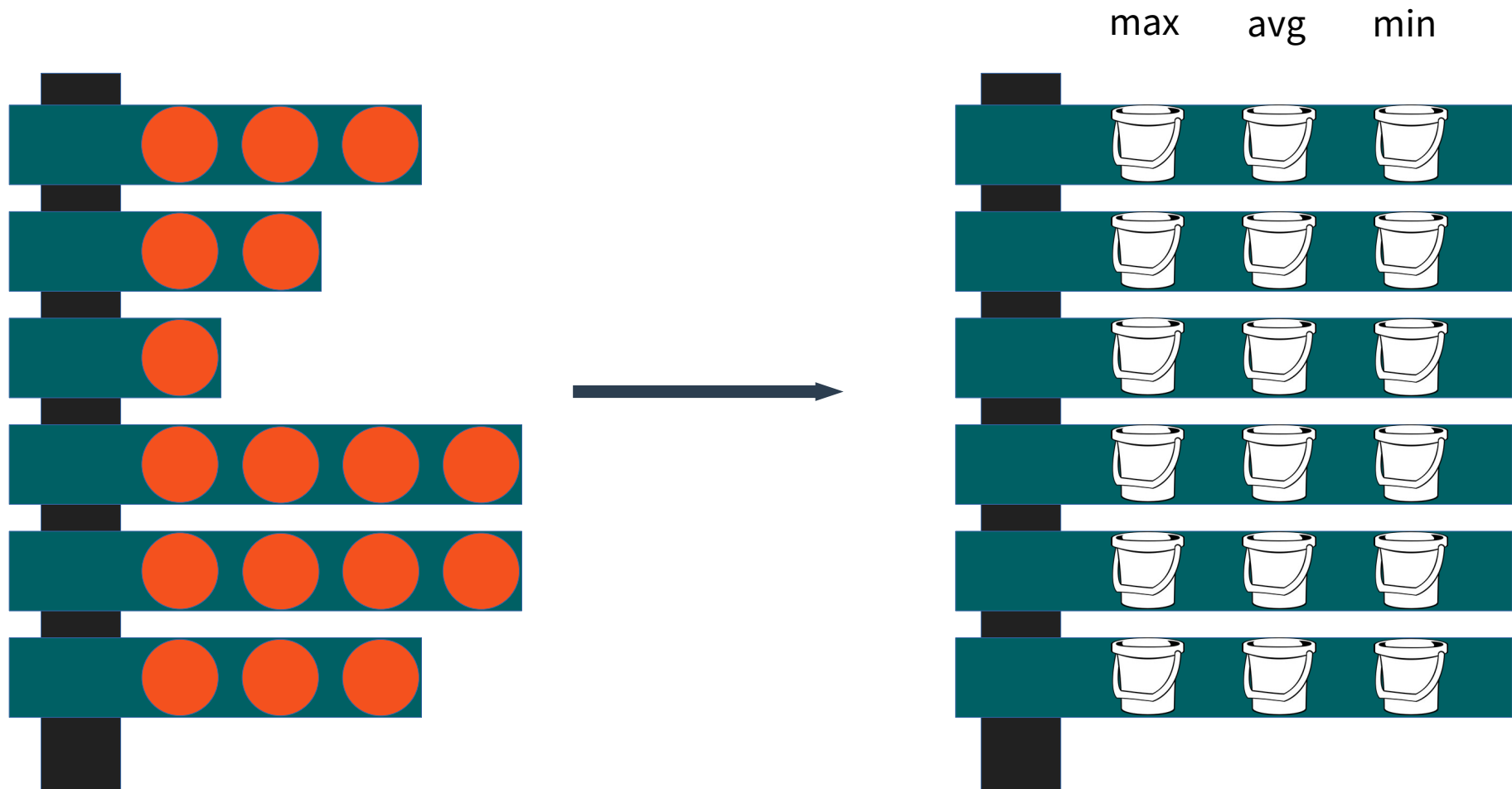
Jagged data. Structure of micro features



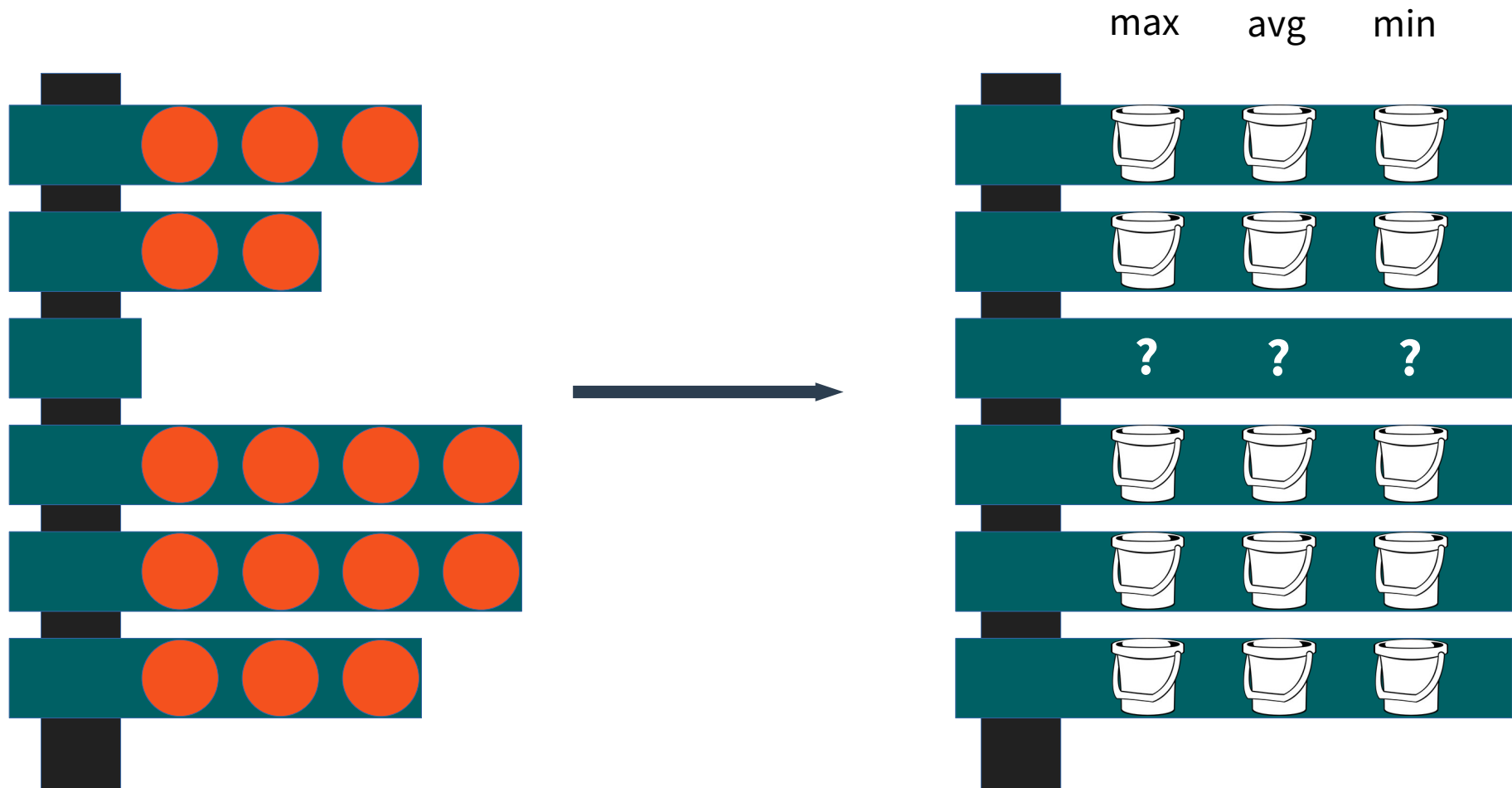
How similar are them?



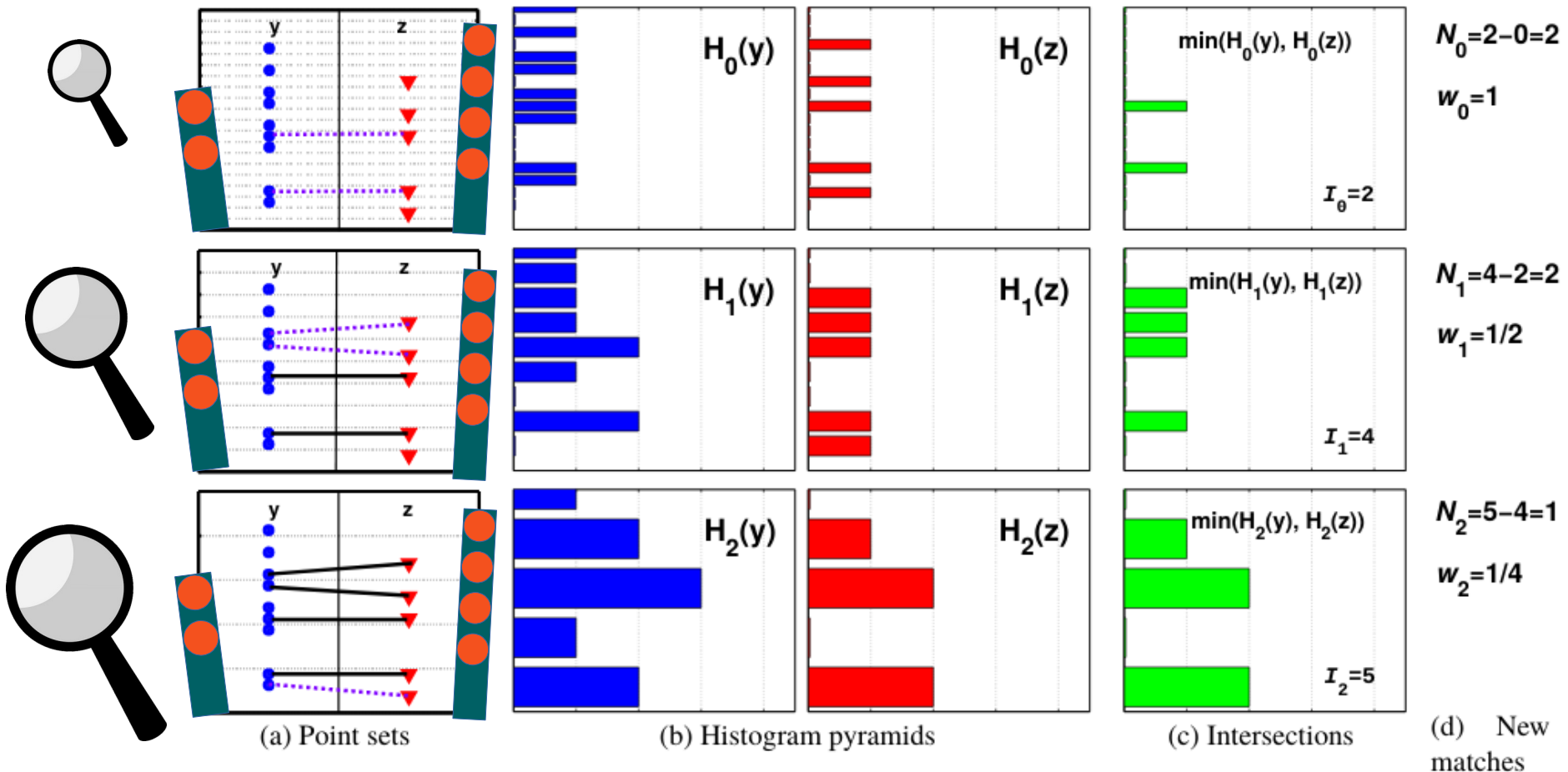
Option: Aggregation



Option: Aggregation



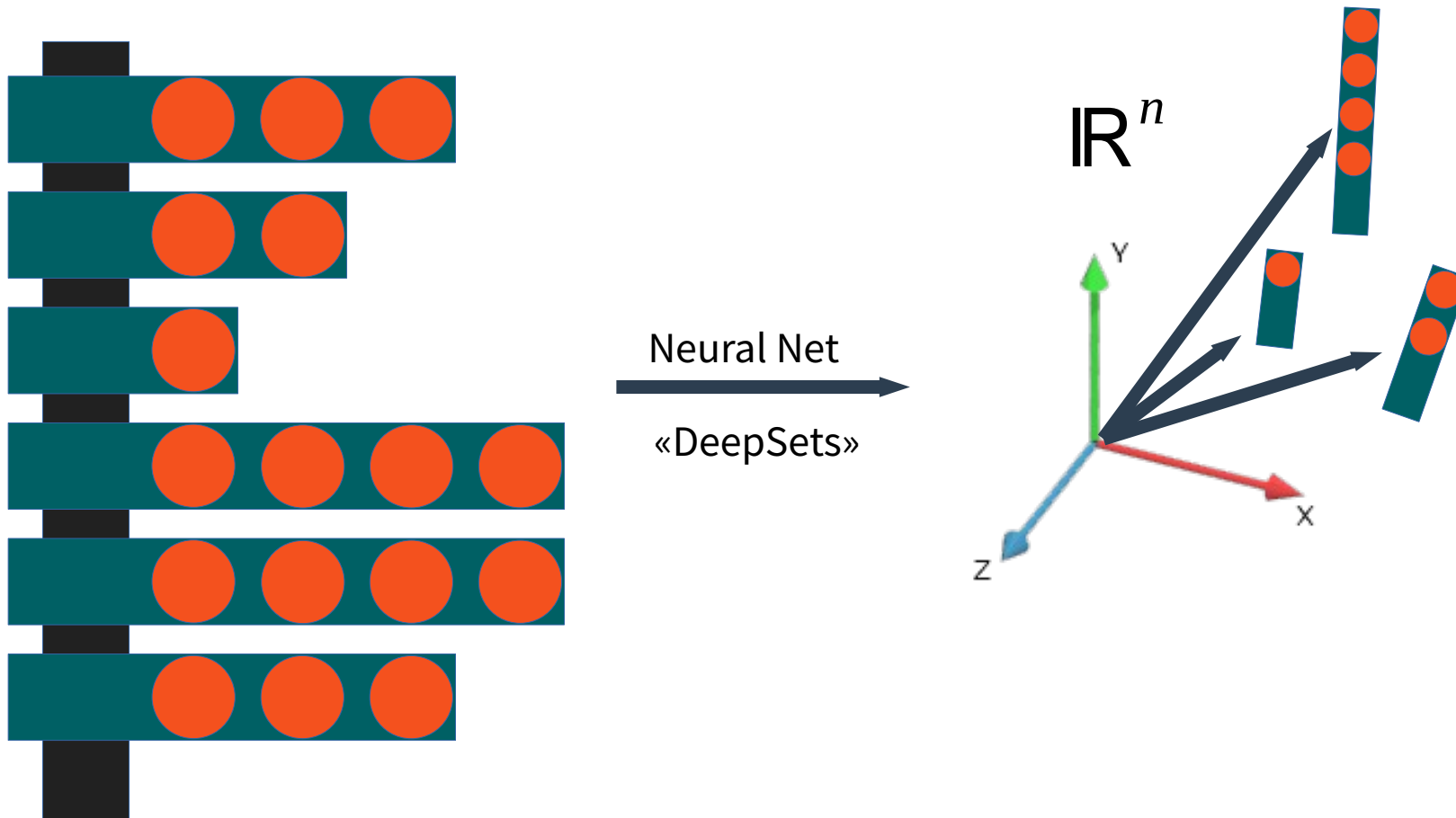
Option: Set Kernels (Pyramid match)



Option: Set Kernels

- + Operation in non-linearly transformed feature space at no cost (Kernel trick)
 - Interpretable. Not a black box.
- Compute and store $N \times N$ matrix of distances (N – number of training instances)
 - N^2 memory consumption
 - Incremental learning is not easily achievable. 1 extra training point triggers N calls to kernel.

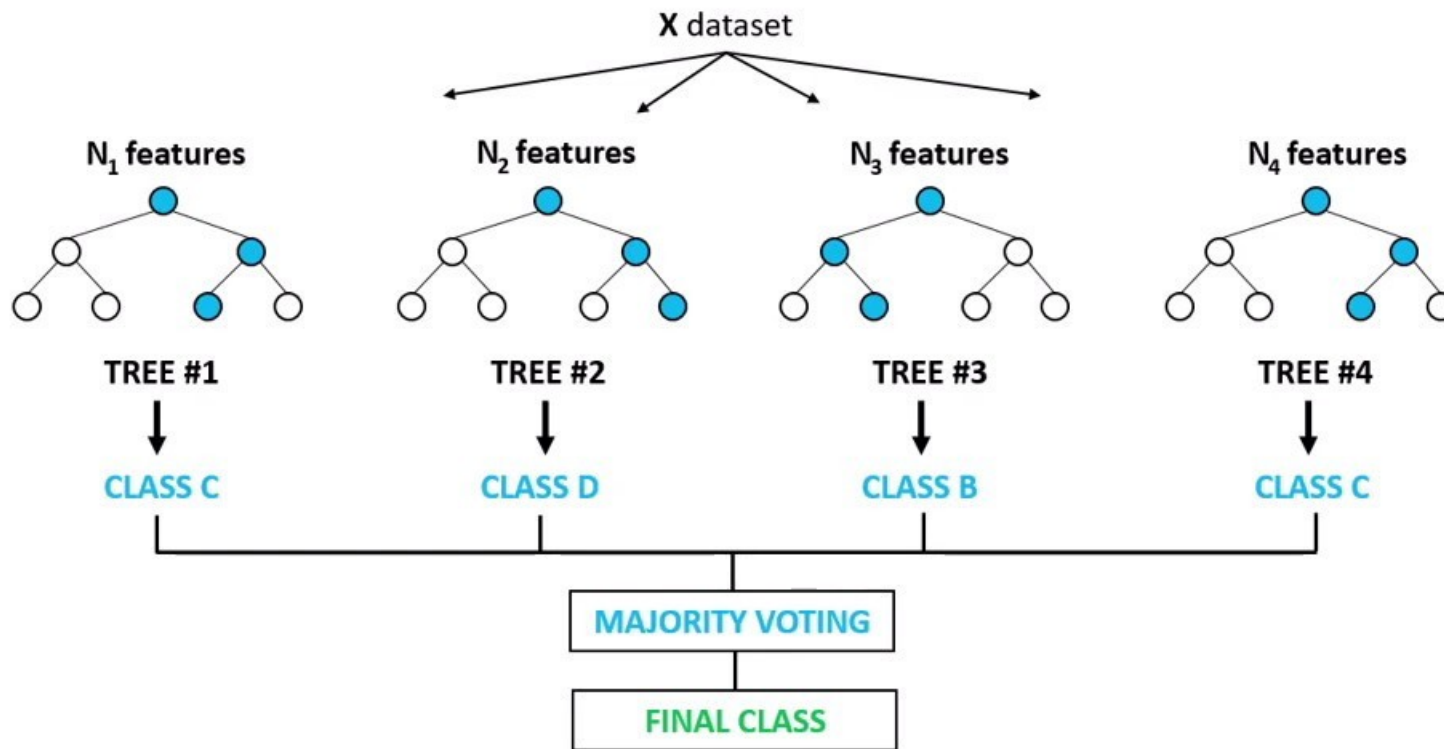
Option: Embedding with NN «DeepSets»



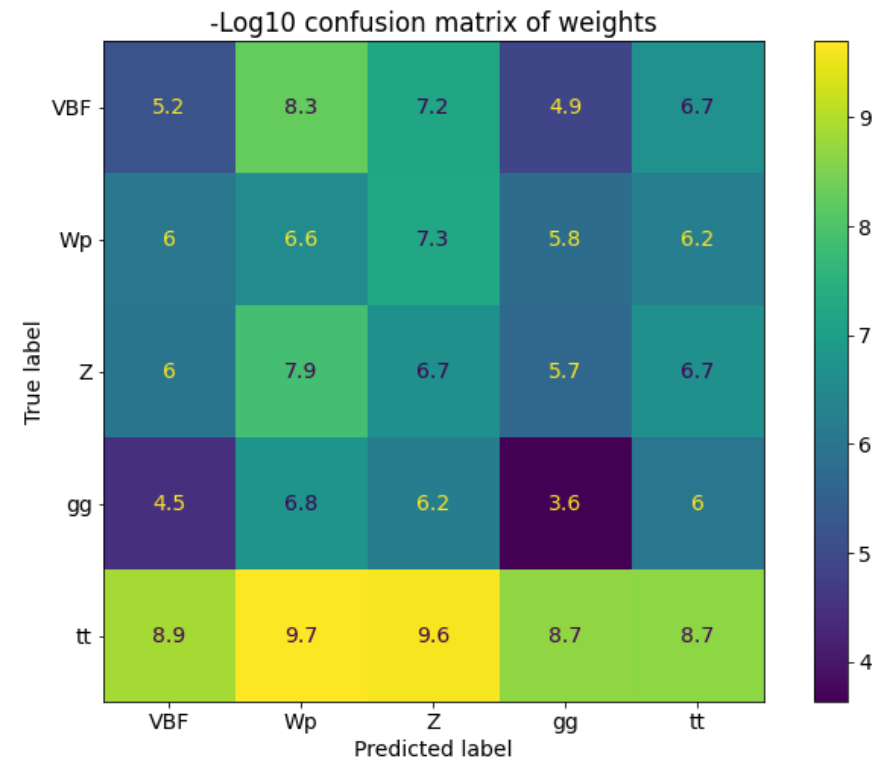
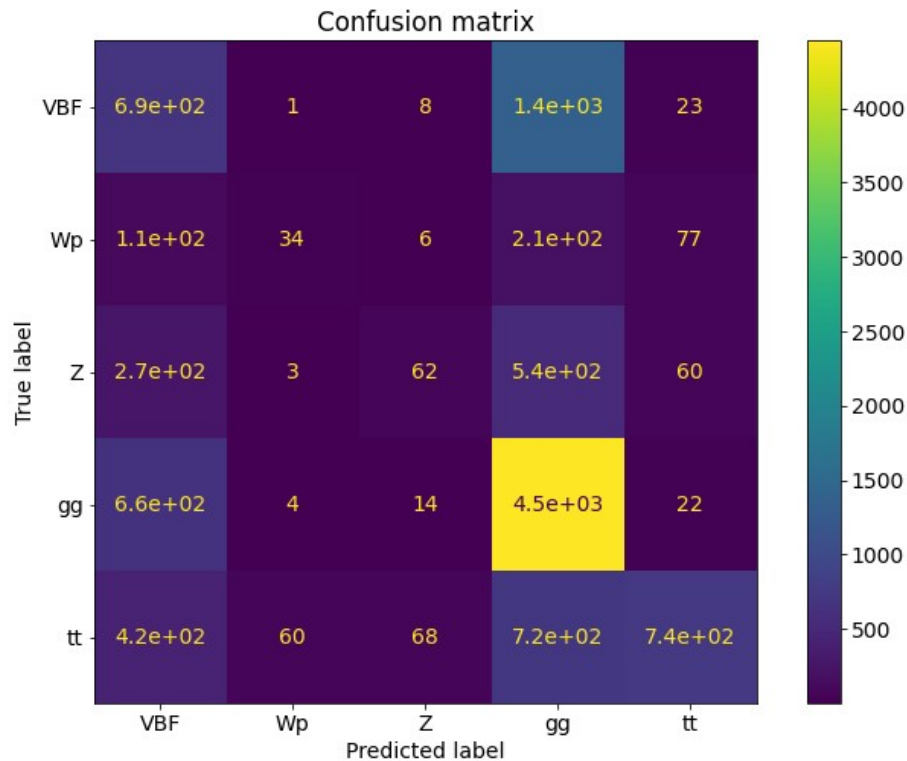
Option: Embedding with NN «DeepSets»

- + Order independence. Treat set of particles as a *set*
 - Get feature embedding model in addition to the classifier
- Neural networks are black box
 - Training requires significant amount of data

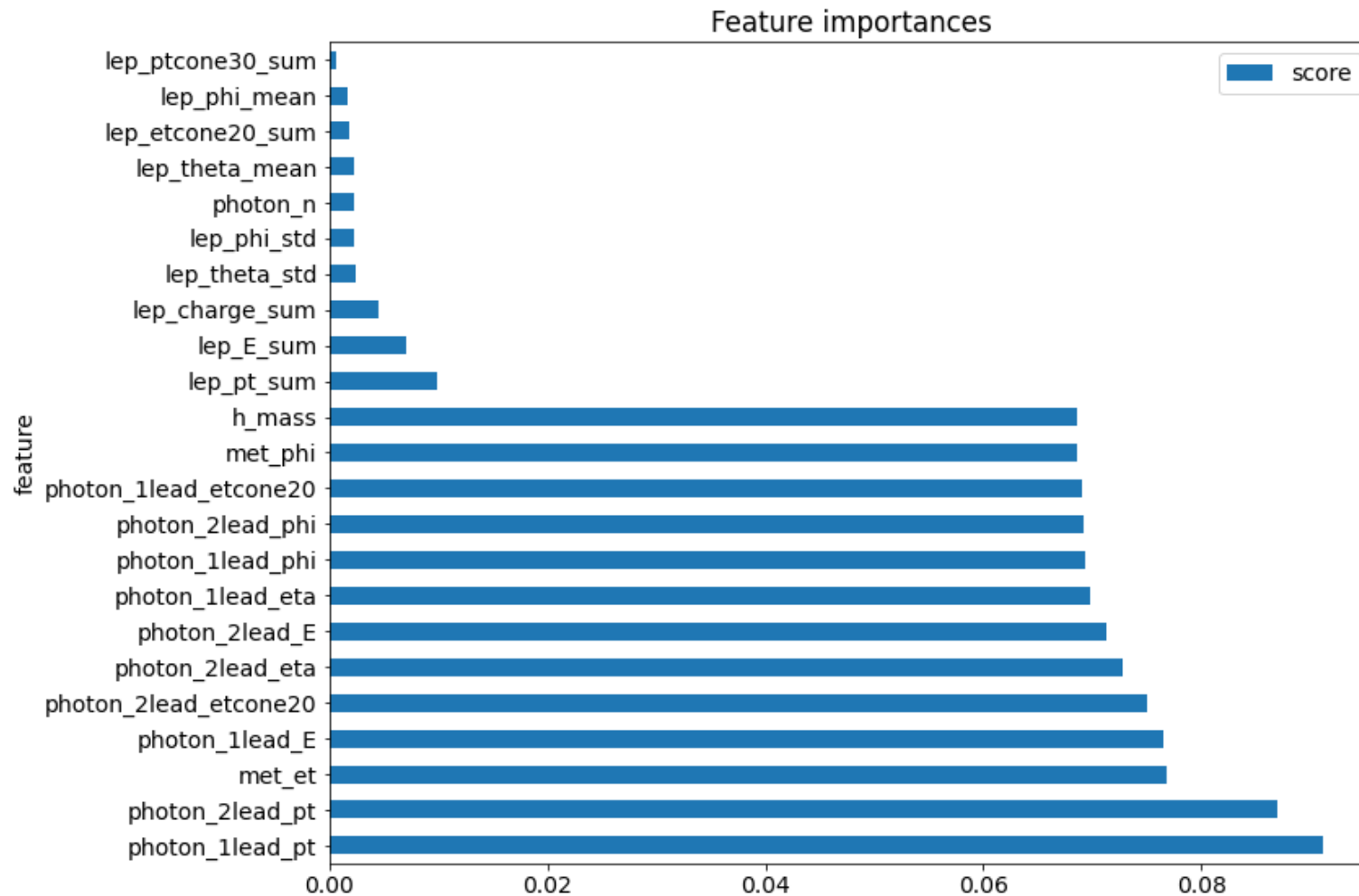
Baseline classifier: Random forest



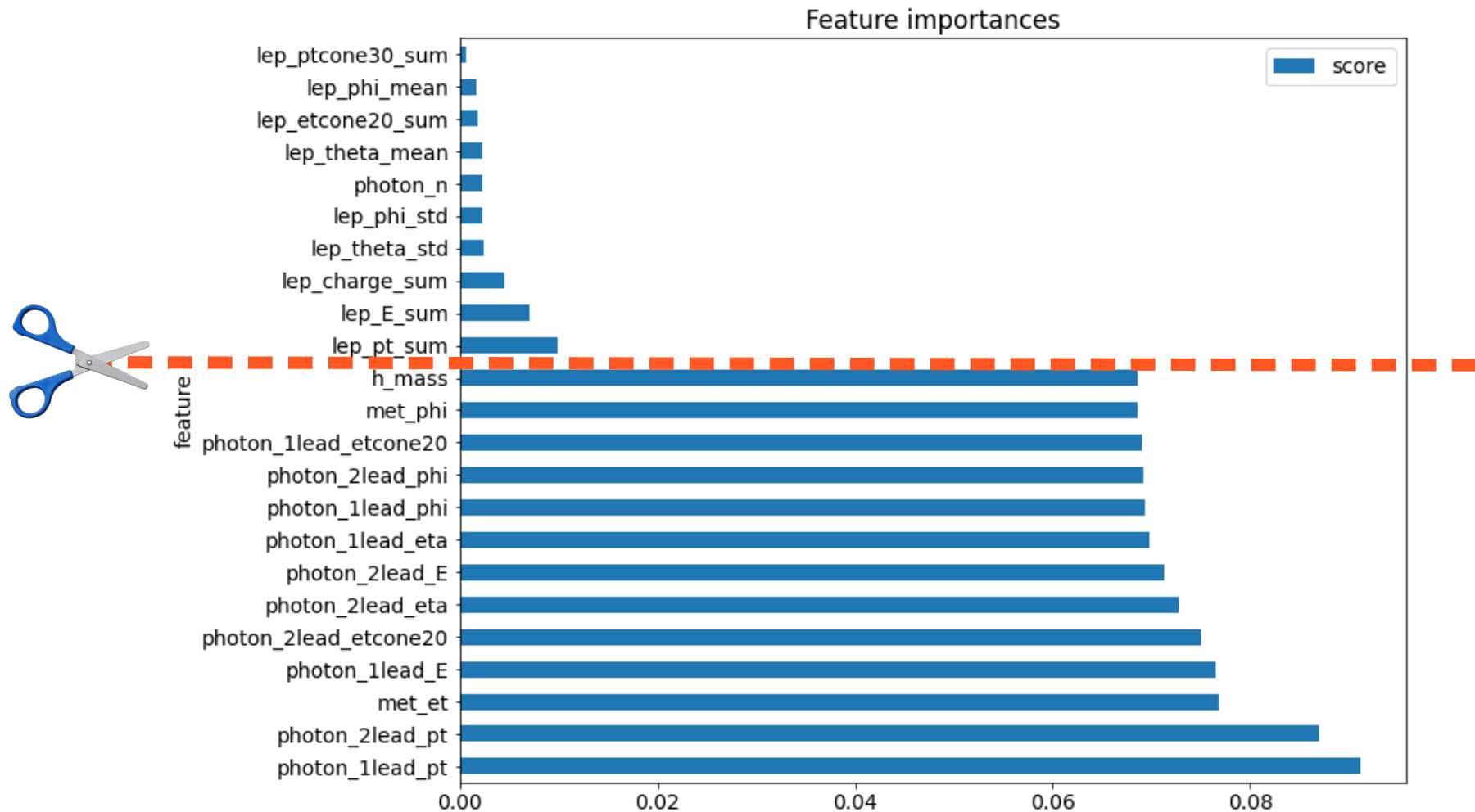
Baseline classifier: Random forest



Baseline classifier: Random forest



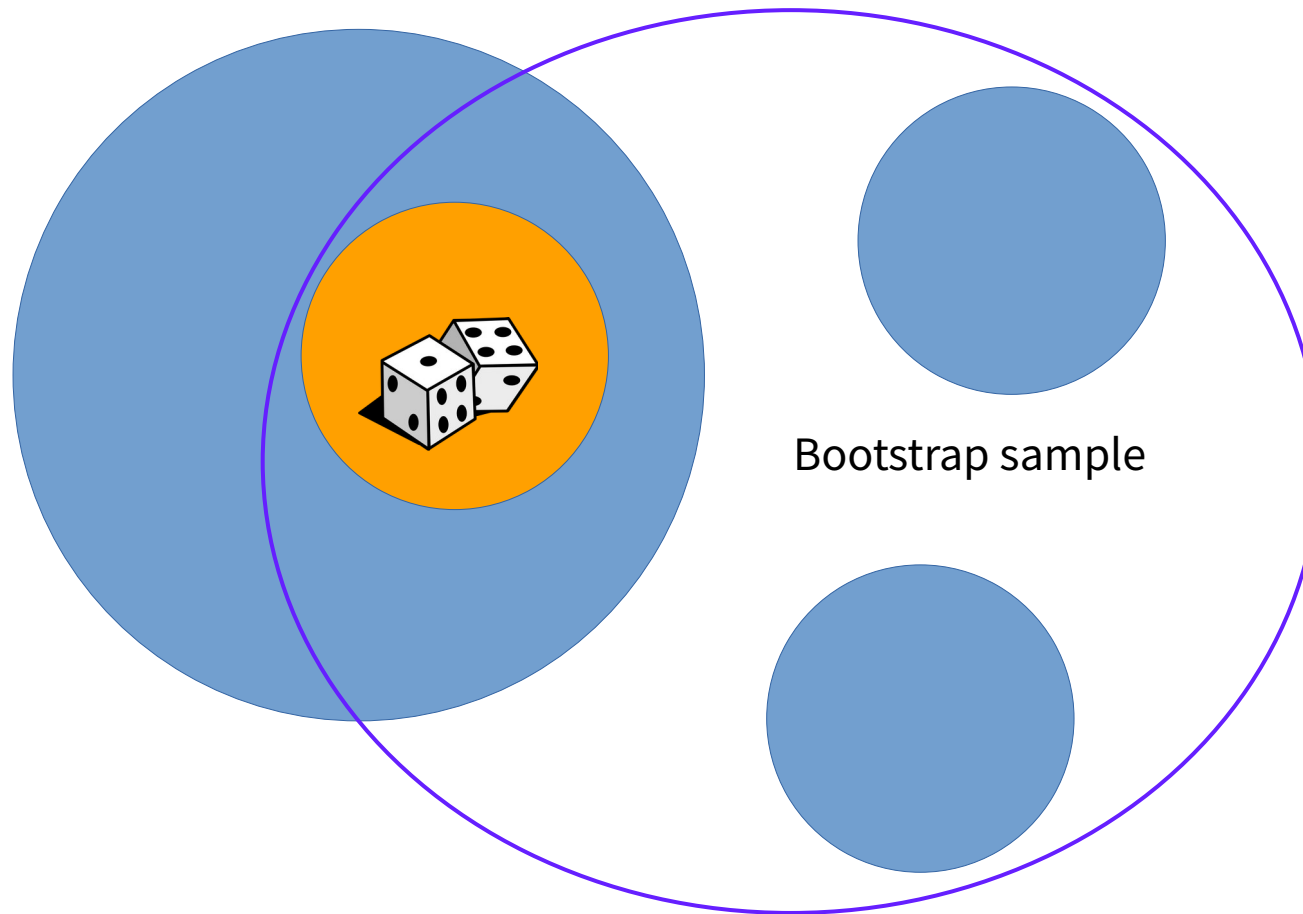
Baseline classifier: Random forest



Imbalanced data

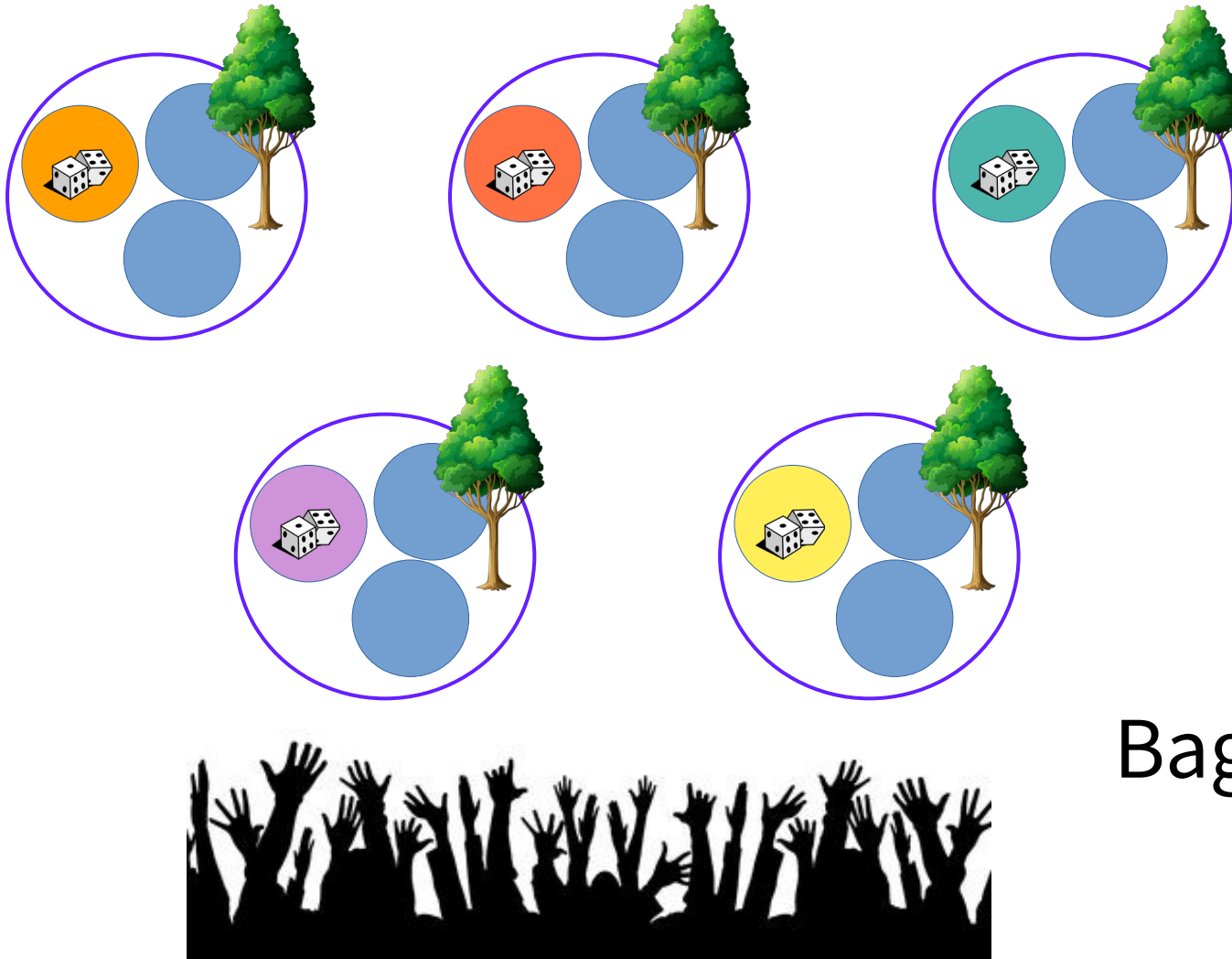
	event_num	weight_sums	weight_frac	weight_per_event
VBF	10369	9.612055e-05	0.064293	9.269992e-09
Wp	2112	1.702021e-05	0.011384	8.058813e-09
Z	4571	1.697019e-05	0.011351	3.712578e-09
gg	26258	1.364896e-03	0.912952	5.198020e-08
tt	9861	2.851057e-08	0.000019	2.891245e-12

Fighting imbalanced data with bagging



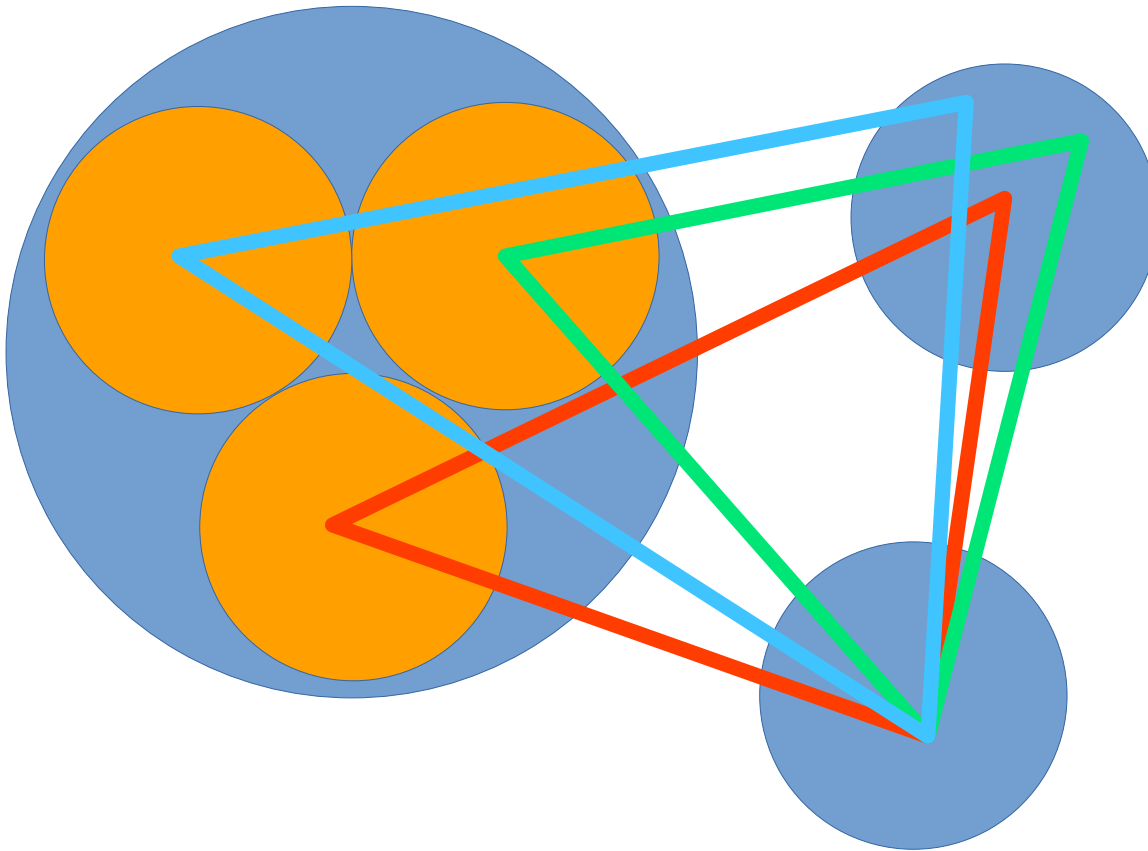
Undersampling

Fighting imbalanced data with bagging



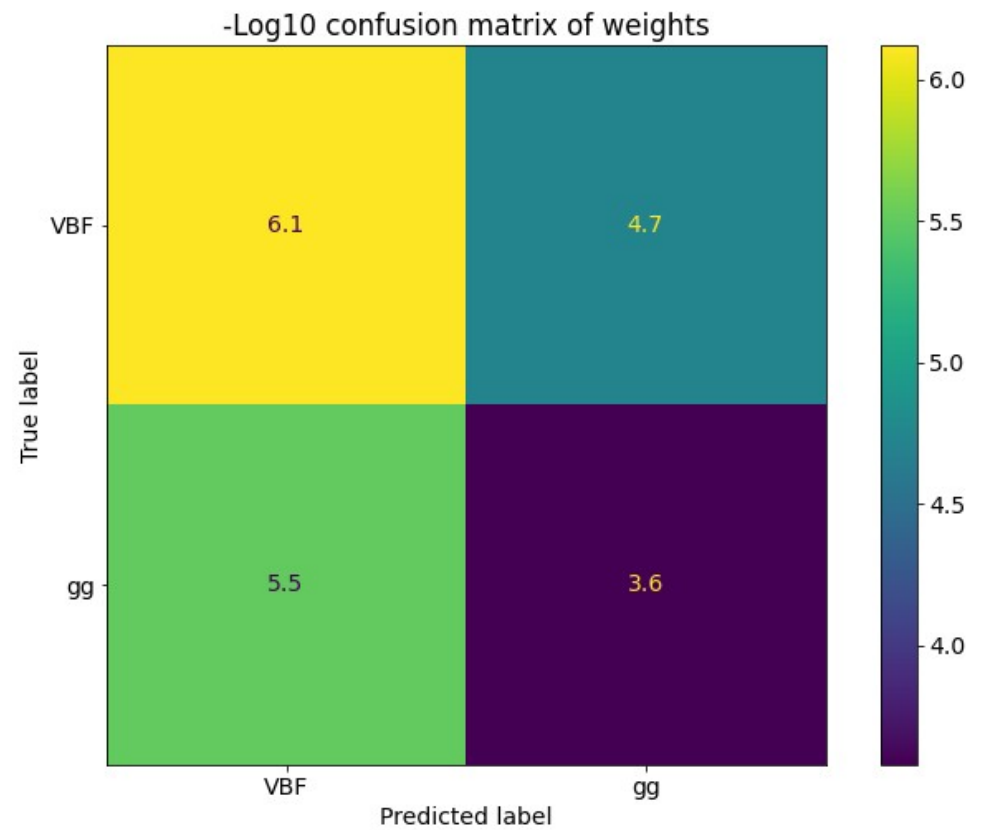
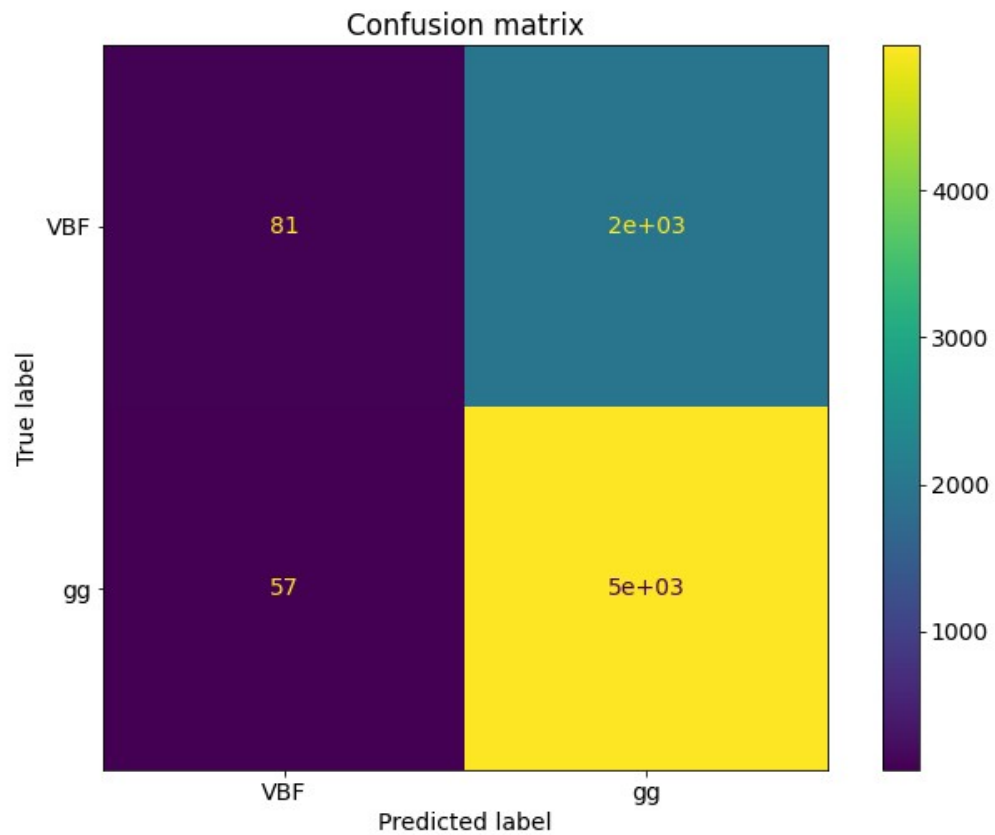
Bagging

Fighting imbalanced data with bagging

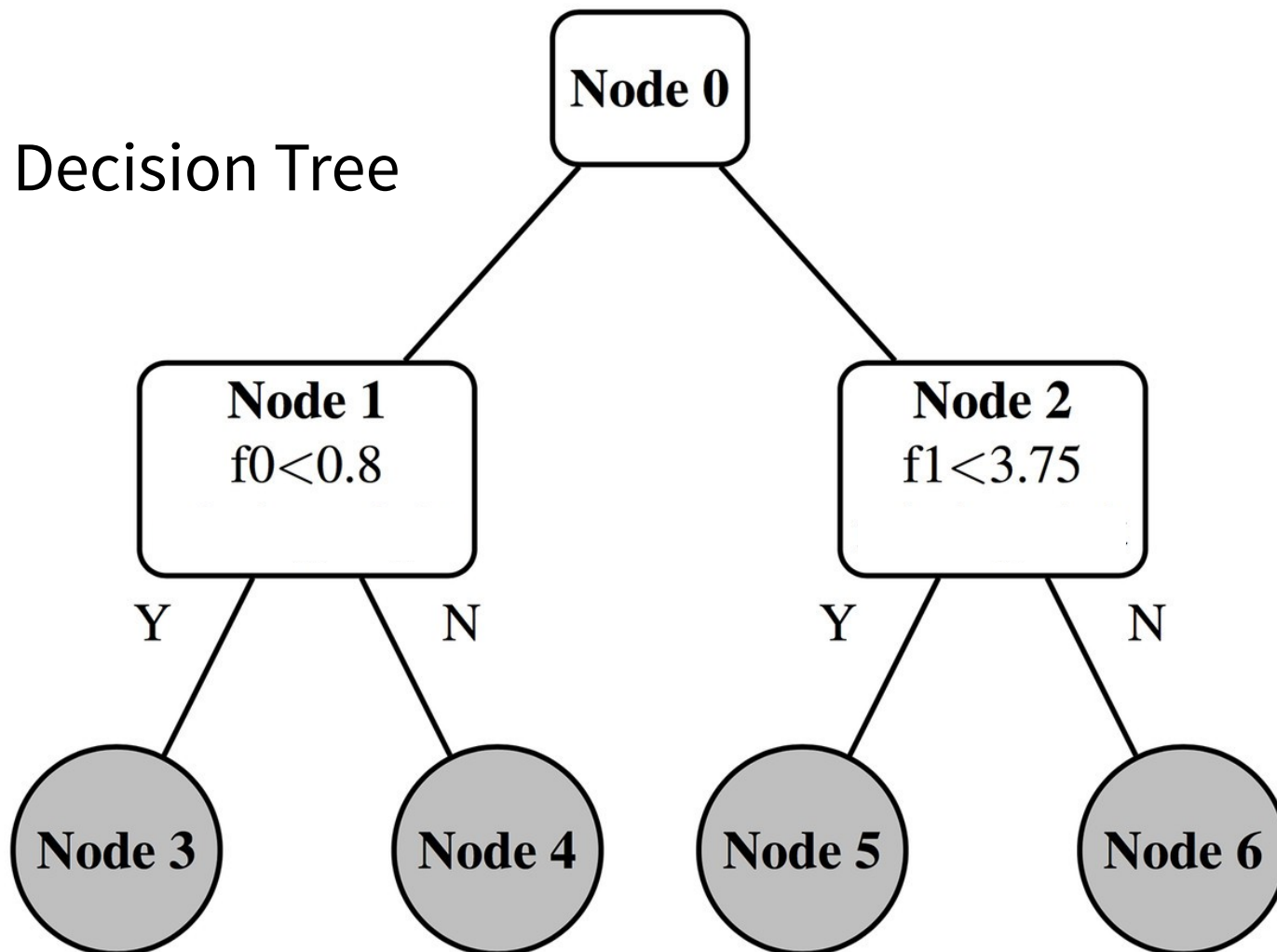


EasyEnsemble

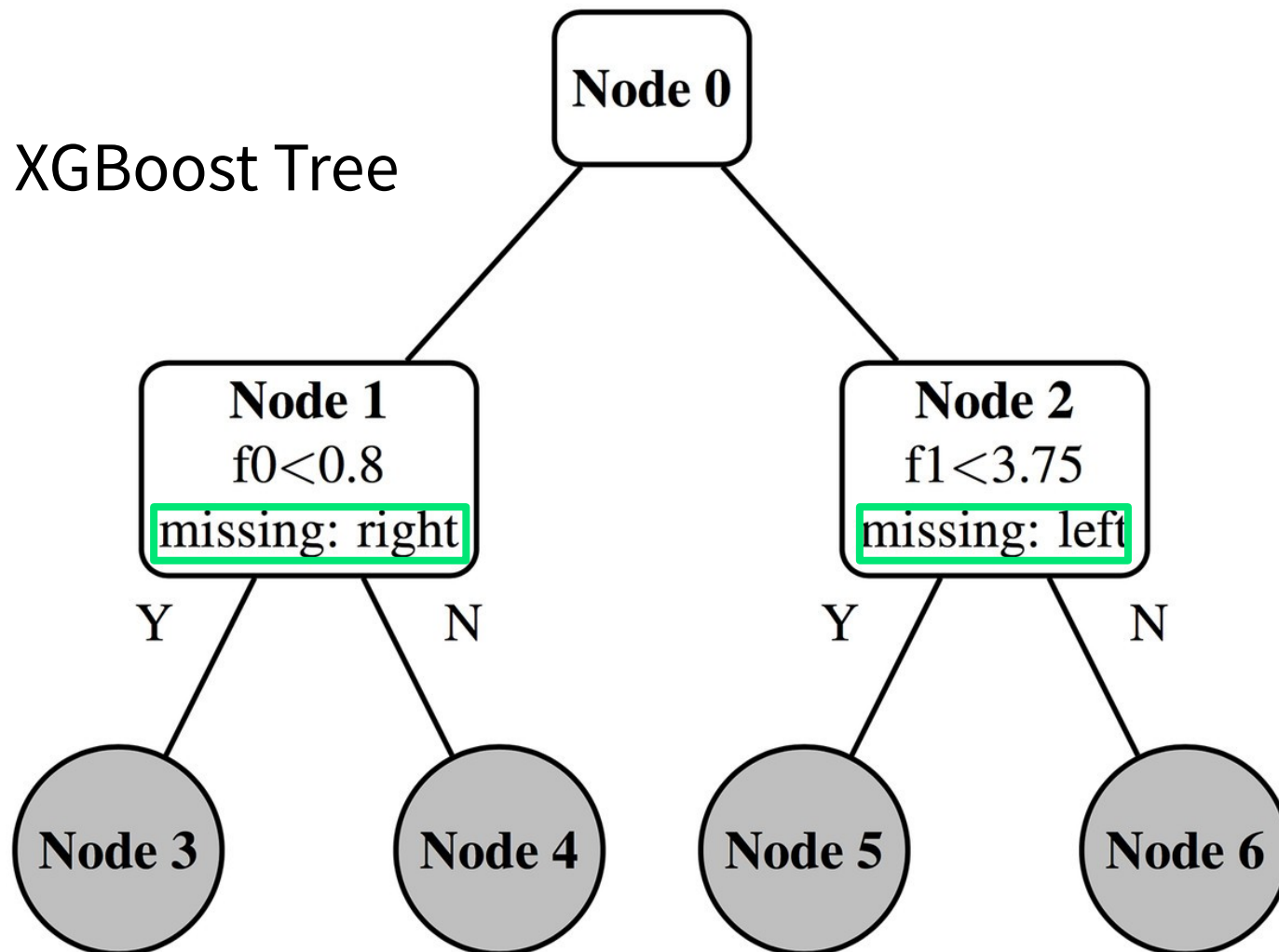
EasyEnsemble model



Alternative to downsampling. XGBoost



Alternative to downsampling. XGBoost



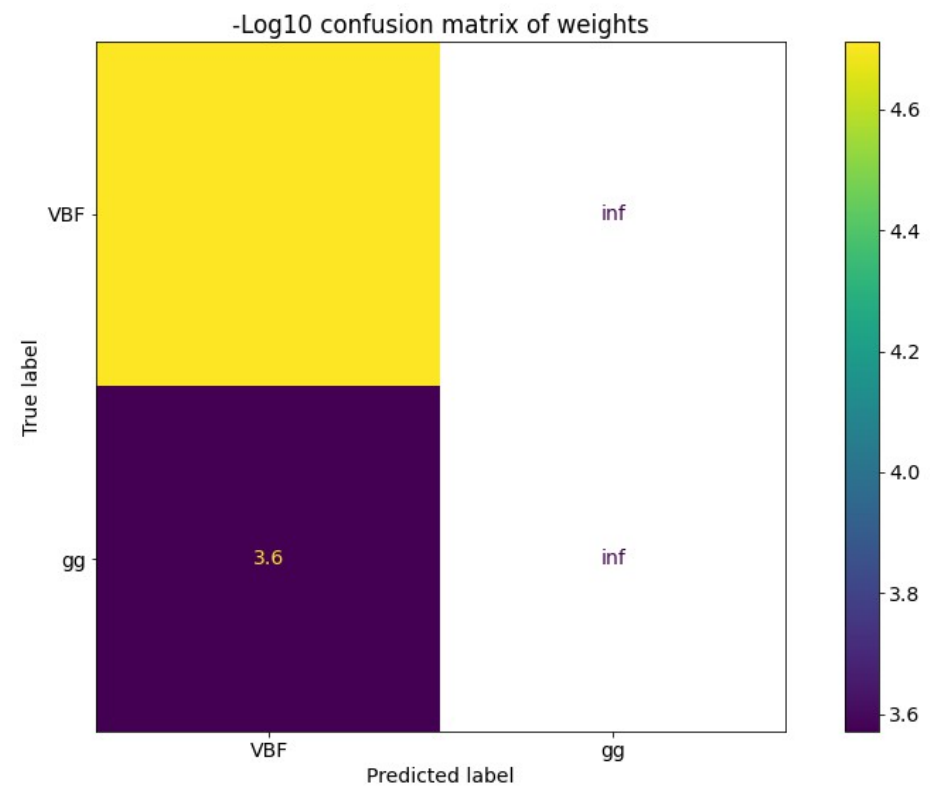
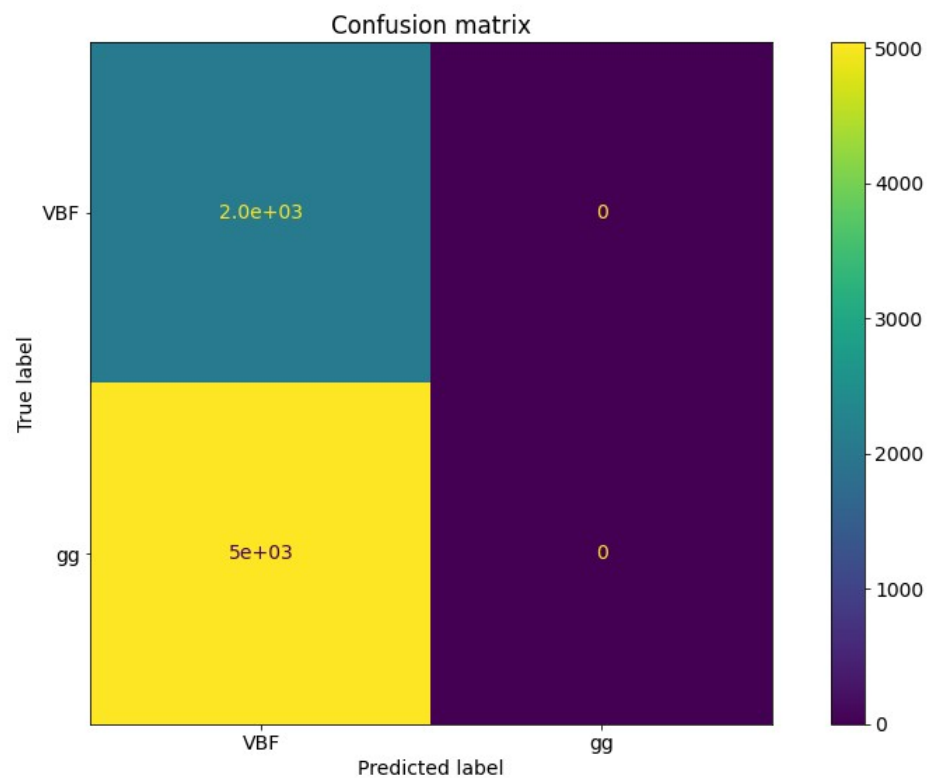
Conclusion

- It is possible to analyze HEP data with Python without loss in efficiency (in comparison to C++)
- HEP datasets are jagged. Possible solutions include:
 - feature aggregation
 - set kernels (Pyramid Match)
 - feature embedding (DeepSets or alternatives)
- HEP datasets are imbalanced. Possible solutions include:
 - downsampling
 - oversampling
 - missing data distribution fit (xgboost)



Backup

XGBoost



XGBoost

