Higgs Uncertainty Challenge (overview and updates)

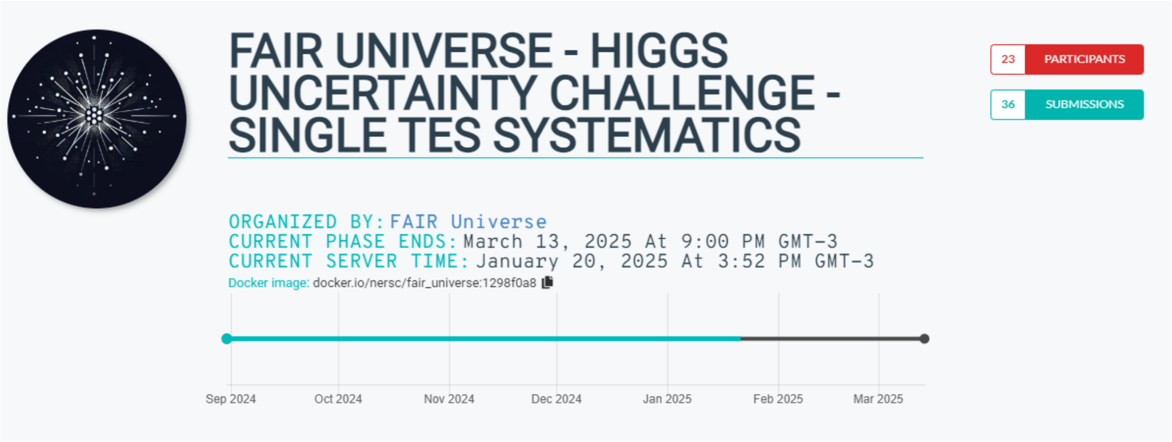
Marcio de Souza

SPRACE

Summary

- ❑ About the Challenge
- ❑ Systematic Biases
- ❑ Dataset Features and Visualization
- ❑ Evaluation Metrics
- ❑ Model Submission
- ❑ Next Steps…

About the Challenge



MAIN GOAL

To develop an estimator for the number of Higgs boson events and its uncertainty in a dataset.

About the Challenge

The dataset consists of ~280 million pp events at 13 TeV created with Pythia 8.2 and Delphes 3.5.0 (~6.5GB)

Process	Number Generated	LHC Events	Label
Higgs	52101127	1015	signal
Z Boson	221724480	1002395	background
Di-Boson	2105415	3783	background
$t\bar{t}$	12073068	44190	background

```
data.get_train_set()['data']
```

✓ 0.0s

	PRI_lep_pt	PRI_lep_eta	PRI_lep_phi	PRI_had_pt	PRI_had_eta	PRI_had_phi	PRI_jet_leading_pt	PRI_jet_leading_eta	PI
0	25.412001	-0.555	1.975	28.997999	0.351	-1.838	54.056000	1.636	
1	39.563000	0.586	3.058	28.274000	0.735	0.084	-25.000000	-25.000	
2	22.320000	0.801	2.521	29.041000	-0.735	-0.864	26.881001	-1.195	
3	46.730999	2.417	-1.388	24.281000	0.999	0.822	48.387001	1.524	
4	24.546000	2.164	1.908	38.452999	0.426	-2.151	-25.000000	-25.000	
...	...	...	...	...	...	...	...	...	
220785	22.063000	-2.229	1.702	52.797001	-2.474	-1.446	-25.000000	-25.000	
220786	37.266998	0.293	-1.102	40.002998	-0.076	2.791	36.099998	-1.481	
220787	20.327999	-0.147	1.123	24.271999	-1.573	-3.082	41.176998	-1.049	
220788	27.358000	0.606	2.957	48.169998	1.732	-0.226	27.389000	1.044	
220789	28.065001	-1.085	-0.326	33.818001	-1.781	3.108	-25.000000	-25.000	

220790 rows × 16 columns

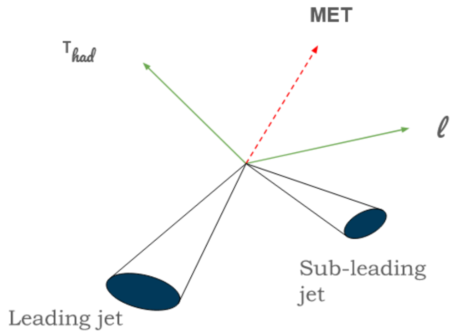


Figure 1: Diagram of the particles in the final state chosen: one lepton, one tau hadron, up to two jets, and the missing transverse momentum vector, see Appendix A for details.

About the Challenge

The primary metric is the **signal strength** (μ)

$$N_{tot} = \mu\gamma + \beta = N_{higgs} + N_{bkg}$$

$$\mu = \frac{N_{tot} - N_{bkg}}{\gamma}$$

where γ is the rate predicted by the Standard Model.

Participants need to generate a 68.27% Confidence Interval for μ , incorporating **six different systematics uncertainties** that can alter the dataset.

About the Challenge

The background processes have a rate β, meaning the number of observed events in a given period of time is expected to follow a Poisson distribution

$$Poisson(\beta) = \frac{\lambda^{\beta} e^{-\lambda}}{\beta!}$$

The standard approach used in LHC analyses is to construct a 1D feature, make a histogram, and then estimate μ (and its uncertainty) using **maximum likelihood estimation**.

The **likelihood is a product of bin-by-bin Poisson probabilities** where the expected counts are determined from simulations.

Systematic Biases

Systematic Biases

Variable	Mean	Sigma	Range
α_{tes}	1.	0.01	[0.9, 1.1]
α_{jes}	1.	0.01	[0.9, 1.1]
$\alpha_{\mathrm{soft\_met}}$	0.	3.	[0., +∞]
$\alpha_{\mathrm{ttbar\_scale}}$	1.	0.25	[0., +∞]
$\alpha_{\mathrm{diboson\_scale}}$	1.	0.025	[0., +∞]
$\alpha_{\mathrm{bkg\_scale}}$	1.	0.01	[0., +∞]

Table 2: List of six systematic bias Nuisance Parameter defined in the challenge, with the mean and sigma of their Gaussian distribution and their range. The corresponding α is set to the Mean value whenever a systematic bias is switched off. "No systematics" means all α are set to their Mean value.

α_{tes} is meant to describe the fact that the detector is not calibrated correctly for the measurement of the hadron momentum, meaning when the detector reports a momentum P_{had} it really is :

$$P_{\mathrm{had}}^{\mathrm{biased}} = \alpha_{\mathrm{tes}} P_{\mathrm{had}}$$

Systematic Biases

And similarly, for the jets momentum (when they are defined)

$$P^{\text{biased}}_{\text{jet\_leading}} = \alpha_{\text{jes}} P_{\text{jet\_leading}}$$

$$P^{\text{biased}}_{\text{jet\_subleading}} = \alpha_{\text{jes}} P_{\text{jet\_subleading}}$$

Both have an influence on P_{MET} as:

$$P^{biased}_{\text{MET}} = P_{\text{MET}} + (1 - \alpha_{\text{tes}}) P_{\text{had}} + (1 - \alpha_{\text{jes}}) P_{\text{leading jet}} + (1 - \alpha_{\text{jes}}) P_{\text{subleading jet}}$$

Systematic Biases

As for $\alpha_{\text{soft\_met}}$, it expresses an additional noise source in the measurement of the missing ET vector, which is not present in the simulation.

$$P_{\text{MET}}^{biased} = P_{\text{MET}} + \begin{pmatrix} Gauss(0, \alpha_{\text{soft\_met}}) \\ Gauss(0, \alpha_{\text{soft\_met}}) \end{pmatrix}$$

DERived features are also impacted if they depend on these PRImary features.

Event selection - Thresholds

Higher thresholds are applied **after the calculation** of the biased PRImary parameters so that the thresholds to be observed on `PRI_had_pt`, `PRI_jet_leading_pt` `PRI_jet_subleading_pt` are independent of α_{tes} and α_{jes}.

Preselection Cuts

Criteria	Pre-selected cut	Post selection cut
Number of τ_{had}	1	
Number of τ_{lep}	1	
$p_T \tau_{had}$	> 20GeV	> 26GeV
$p_T \tau_{lep}$	> 20GeV	> 20GeV
$p_T leadingjet$	> 20GeV	> 26GeV
$p_T subleadingjet$	> 20GeV	> 26GeV
Charege	Opposite Charges	

⚠ **Note:** The Post selection cuts are the cuts made after systematics is applied.

Dataset Features

Dataset Primary Features

There are **16 features** (columns) prefixed with PRI (for PRImitives) are “raw” quantities about the bunch collision as measured by the detector, essentially parameters of the momenta of particles. Those are:

```
PRI_had_pt
PRI_had_eta
PRI_had_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
```

```
PRI_jet_num
PRI_jet_leading_pt
PRI_jet_leading_eta
PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt
```

Dataset Derived Features

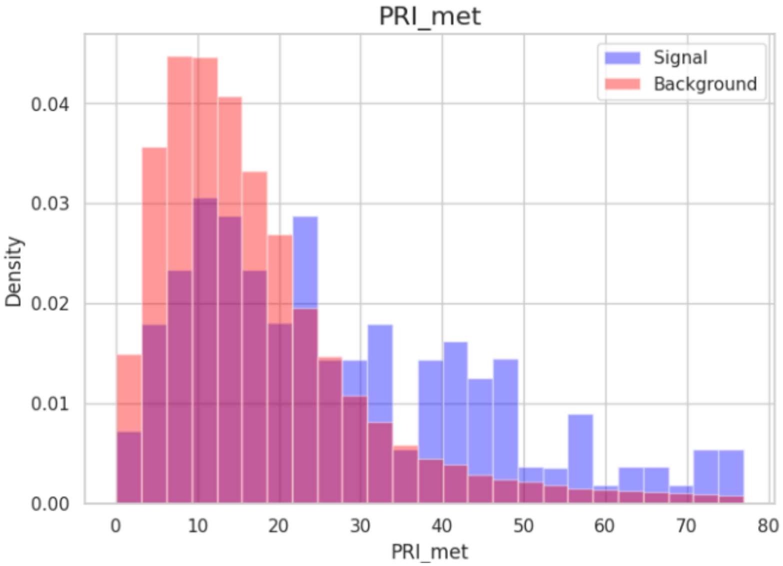
There are also **12 variables** (also columns) prefixed with DER (for DERived), that are quantities computed from the primitive features on the fly from PRImary features (including possible systematics shifts)

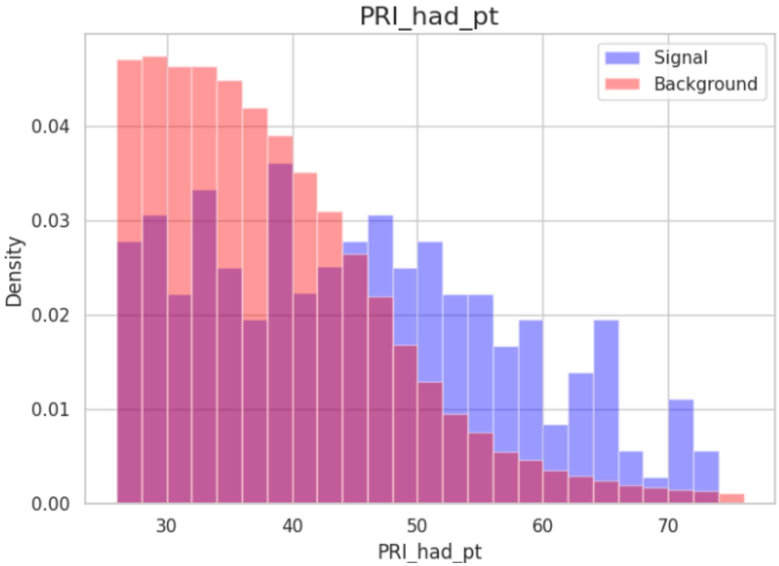
```
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_had_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality
```

These quantities **were selected by the physicists of ATLAS**, either to select regions of interest or as **features for the Boosted Decision Trees** used in this analysis in order to enhance signal Higgs boson events separation from background events.

Dataset Visualization

On the challenge main page, they offer a "starting kit" on `GitHub` with some helper functions to visualize and manipulate some features of the dataset, as well a simplified model to start a dummy submission.

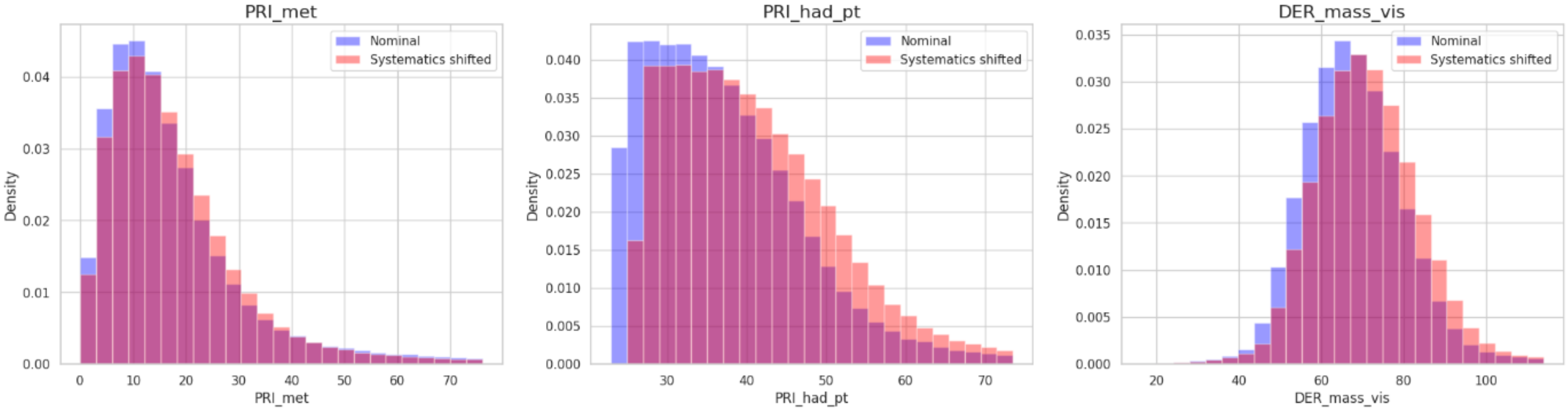




Dataset Visualization

```python
syst_train_data = systematics(data_vis, tes = 1.1, jes=0.9,  bkg_scale=50, ttbar_scale=5, dopostprocess=True)
train_visualize.histogram_syst(syst_train_data['data'],syst_train_data['weights'])
```

✓ 3.6s



Evaluation Metrics

Evaluation Metrics

Average Interval Width over N *pseudo-experiments*

$$w = \frac{1}{N_{test}} \sum_{i=1}^{N} \left| \mu_{84,i} - \mu_{16,i} \right| ,$$

The second component $\boldsymbol{c}$ quantifies the **frequency** with which the true value of $\boldsymbol{\mu_{true}}$ **falls within the 68.27% Confidence Interval (CI)**

$$c = \frac{1}{N_{test}} \sum_{i=1}^{N} 1 \text{ if } \mu_{\text{true},i} \in [\mu_{16,i} - \mu_{84,i}] .$$

Evaluation Metrics

$$c \in [0.6827 - 2\sigma_{68}, 0.6827 + 2\sigma_{68}] : f(c) = 1$$

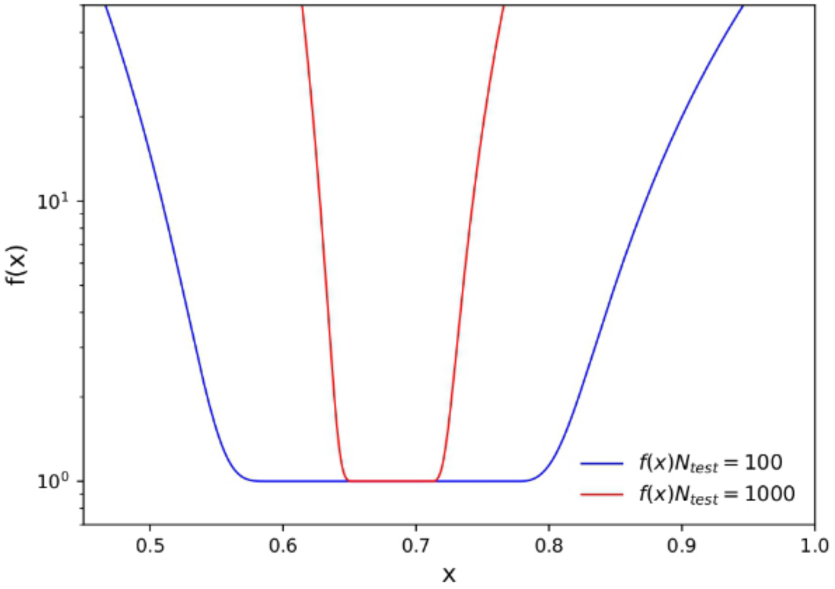
$$c < 0.6827 - 2\sigma_{68} : f(c) = 1 + \left|\frac{c - (0.6827 - 2\sigma_{68})}{\sigma_{68}}\right|^4$$

$$c > 0.6827 + 2\sigma_{68} : f(c) = 1 + \left|\frac{c - (0.6827 + 2\sigma_{68})}{\sigma_{68}}\right|^3$$

If the confidence interval accurately represents the 68.27% quantile, the true value of μ should lie within this interval in 68.27% of the pseudo-experiments.

Consequently, they employ a penalising function f that penalises models that deviate from this 68.27% reference

They opted for an **asymmetric penalty function** because, in HEP, *overestimating uncertainty is deemed more acceptable than underestimating it.*



Evaluation Metrics

The final Coverage Score used to rank participants is calculated as follows:

$$\text{score} = -\ln((w+\epsilon)f(c)),$$

$\boldsymbol{w}$ represents the average width of the Confidence Interval, $\boldsymbol{c}$ is the coverage parameter, and $\boldsymbol{\epsilon = 10^{-2}}$ is a regularization term to guard against submissions that report unrealistically narrow CIs.

Using the − ln() function ensures that the score variations remain within a modest range.

Model Submission

Model Submission

The Model Class

We import a class named `Model` from the submission file (`model.py`). This `Model` class has the following methods:

- `init`: receives train set and systematics class as input
- `fit`: can be used for training
- `predict`: receives one test set and outputs a dictionary with the following keys
  - `mu_hat` : predicted mu $\hat{\mu}$
  - `delta_mu_hat`: $\Delta\hat{\mu}$ bound for μ
  - `p16`: 16th percentile
  - `p84`: 84th percentile

Model Submission

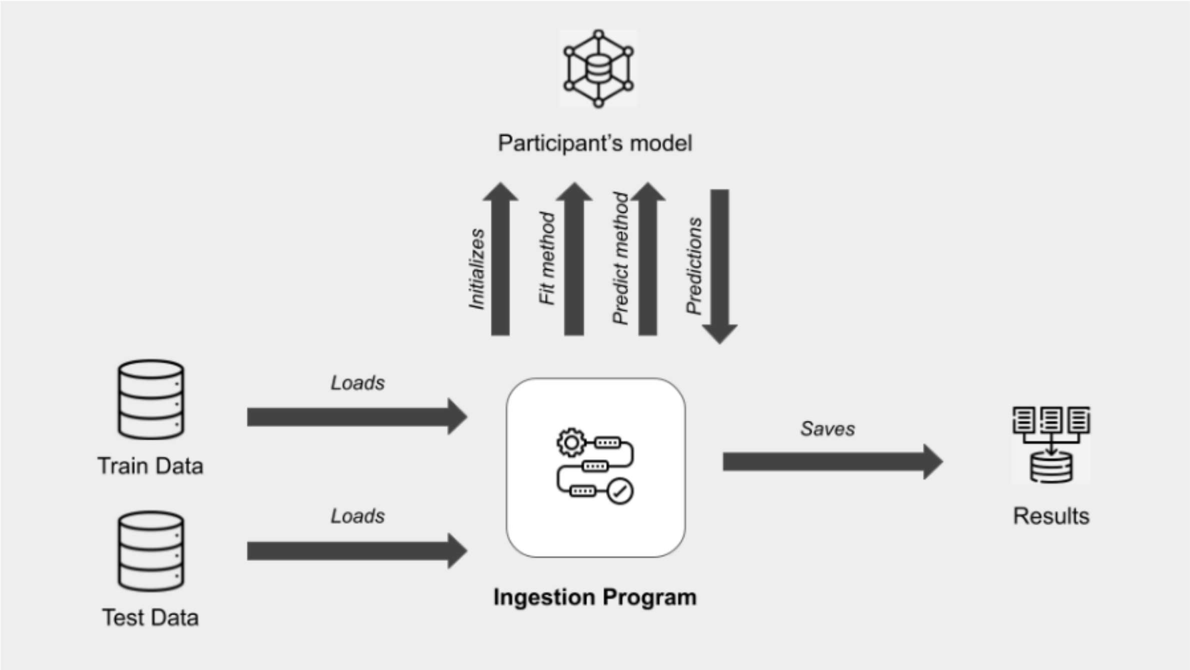


Figure 5: Ingestion Program execution flow

Model Submission

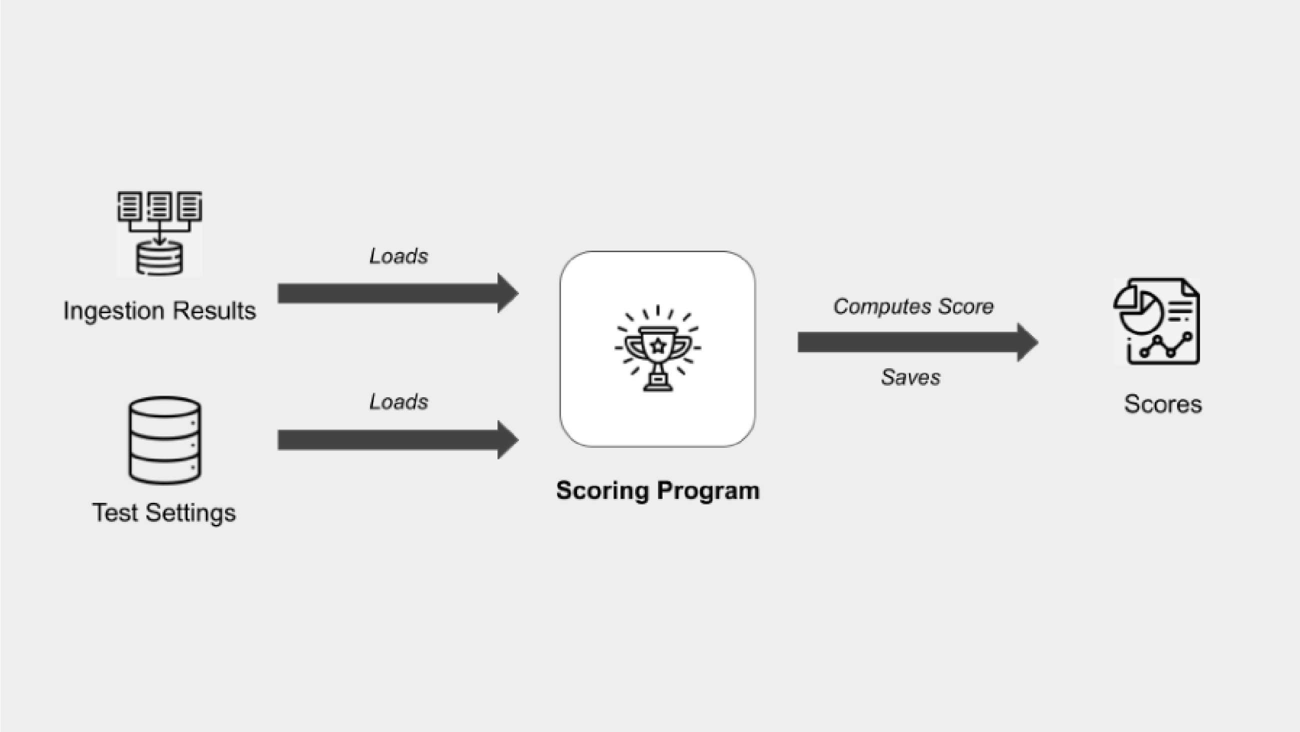
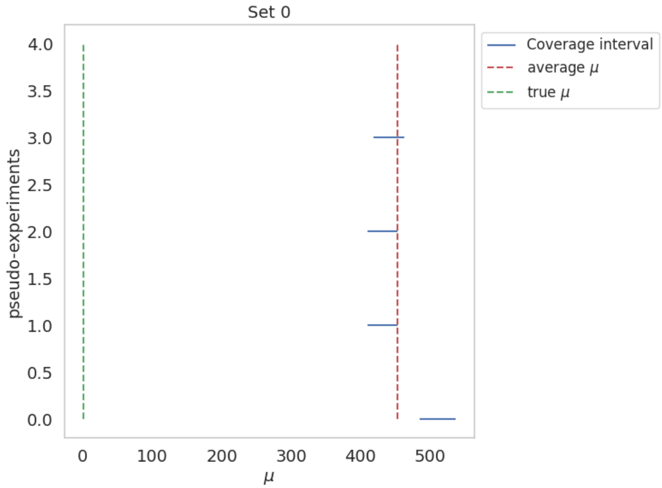
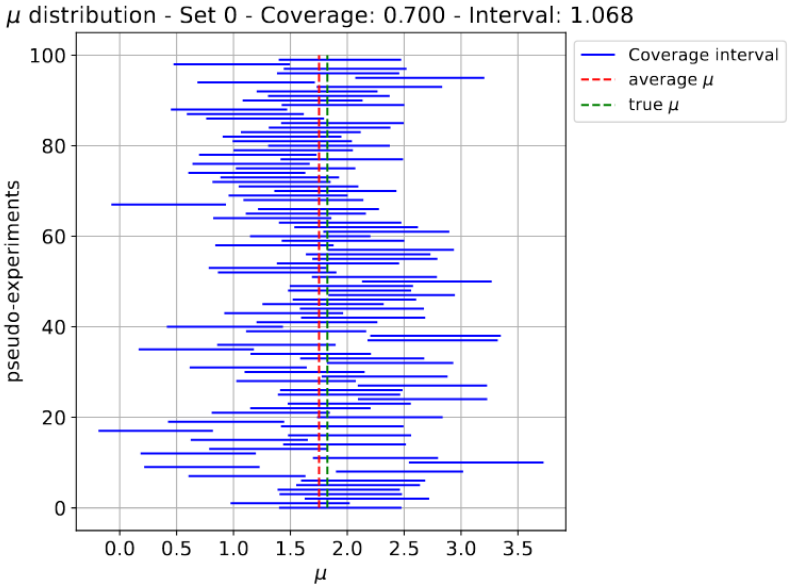


Figure 6: Scoring Program execution flow

Model submission



They offer this…



But we'll need something like this!

Next steps…

Next steps…

- ✓ Understand the challenge goal DONE
- ✓ Read through the fair-universe documentation and examples DONE
- ✓ Reproduce the dummy model locally DONE
- ✓ Gain access to the Codabench mainboard (thanks to Thiago) DONE
- ✓ Learn how to use the GPU available at the NCC DONE
- ✓ Learn about the problem statistics () UNDERWAY
- ✓ Search for hints in the previous Kaggle challenge UNDERWAY

TO DO:

- ☐ Understand how to modify the Model Class to a different architecture STARTED
- ☐ Prepare a model using the simplified TES challenge PLANNED
- ☐ Submit our own model to de competition (deadline March 13) PLANNED

Extra

In a machine learning context, the task resembles a transduction problem with distribution shift: it requires constructing a μ interval estimator from labelled training data and biased unlabelled test data. One possibility is to train a classifier to distinguish Higgs boson from the background, with robustness against bias achieved possibly through data augmentation (or adversarial approach, or black box optimisation or any other novel approach) via the provided script.

?