Robustness and interpretability of machine learning methods applied to LHC data

Steffen Mæland

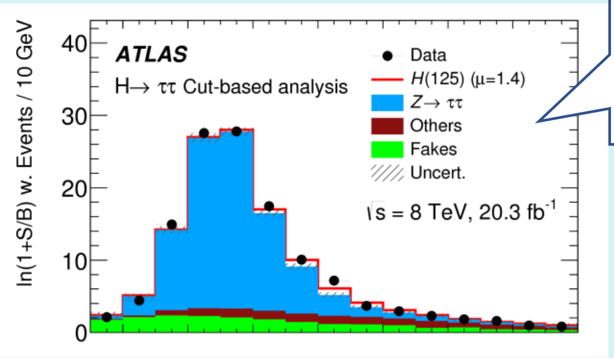
Western Norway University of Applied Sciences

Høgskulen på Vestlandet

NORCC

ML use case: Event classification

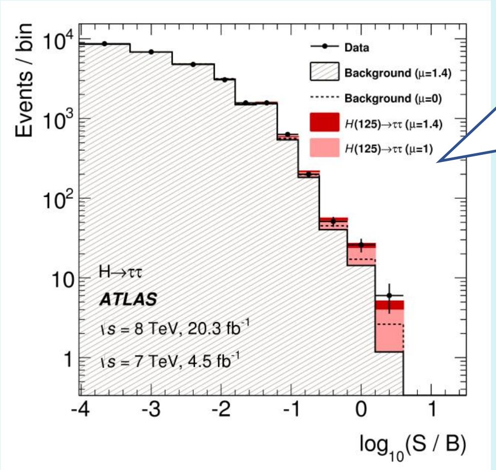
Event ⟶ Observables (*p*, *E*, …) ⟶



⟶ Fit ⟶ Result

Carefully constructed observable

Event ⟶ Observables (*p*, *E*, …) ⟶



⟶ Fit ⟶ Potentially better result

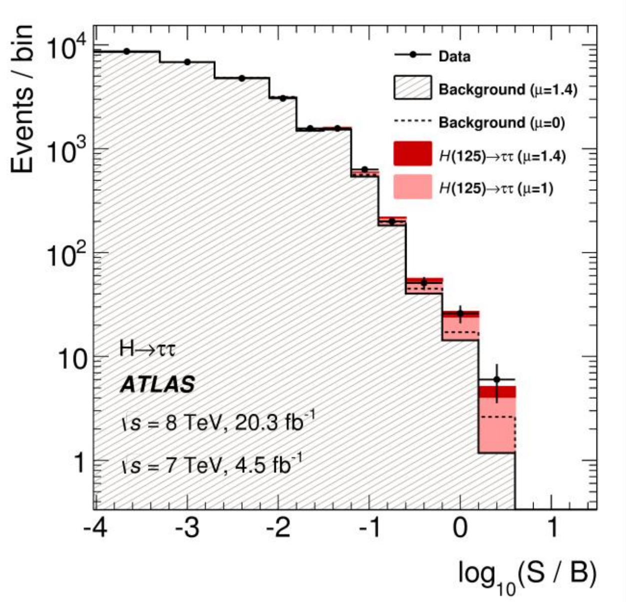
Machine learning

The ATLAS collaboration., Aad, G., Abbott, B. *et al.* Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector. *J. High Energ. Phys.* **2015**, 117 (2015).

ML use case: Event classification

Since the transformation

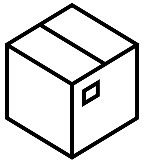
Observables (*p, E, …*) ⟶



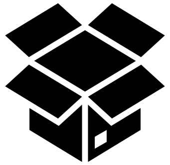
is (highly) non-trivial, verification is typically not straight-forward

Terms

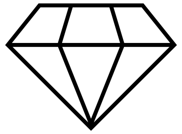
- *Interpretability:*
  Passive characteristic of a model – to what extent it is understandable by humans



- *Explainability:*
  Active characteristic, involving methods that clarify a model's decision process or internal function



- *Robustness:*
  To what extent a model's prediction is affected by perturbations in the input data



Explainability approaches

Assume we have an ML model f with a bunch of parameters $\boldsymbol{\theta}$, taking in data $\boldsymbol{x}$ and returning predictions y:

$$y = f(\boldsymbol{x}, \theta)$$

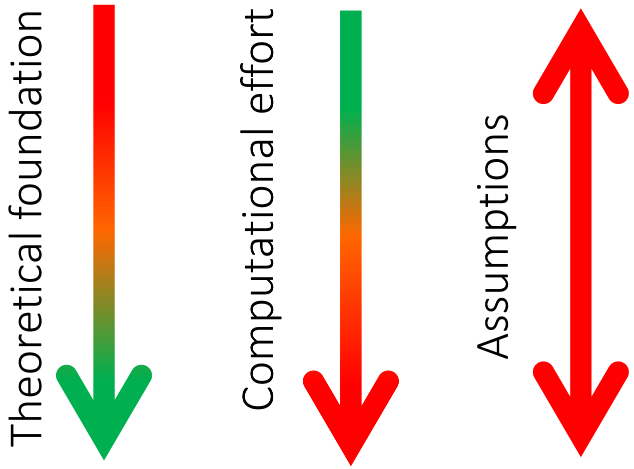
Three options for explaining y:

- Replace f by an interpretable model: *Surrogate model explanations*
- Vary $\boldsymbol{x}$ and observe the effect: *Extrinsic explanations*
- Study θ (for some given $\boldsymbol{x}$): *Intrinsic explanations*

Extrinsic explanations

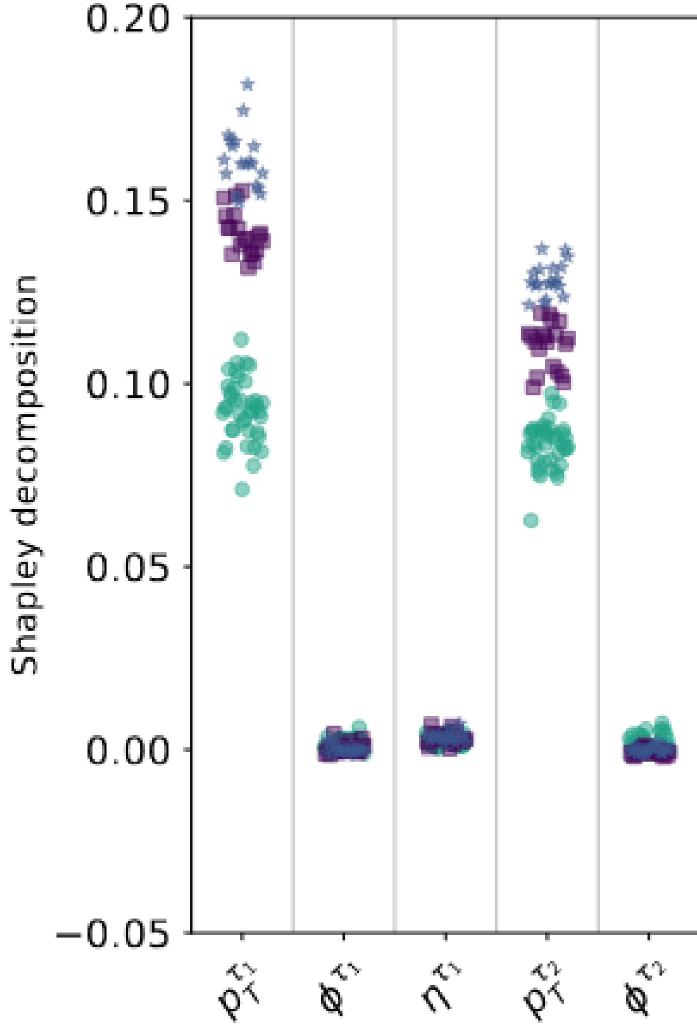
Model agnostic – study y for different $\boldsymbol{x}$

→ Feature importance explanation
(either on average or for single events)



- Randomise feature values
  https://arxiv.org/abs/1801.01489
- SHAP
  https://arxiv.org/abs/1705.07874
- Shapley values
  https://arxiv.org/abs/1705.07874

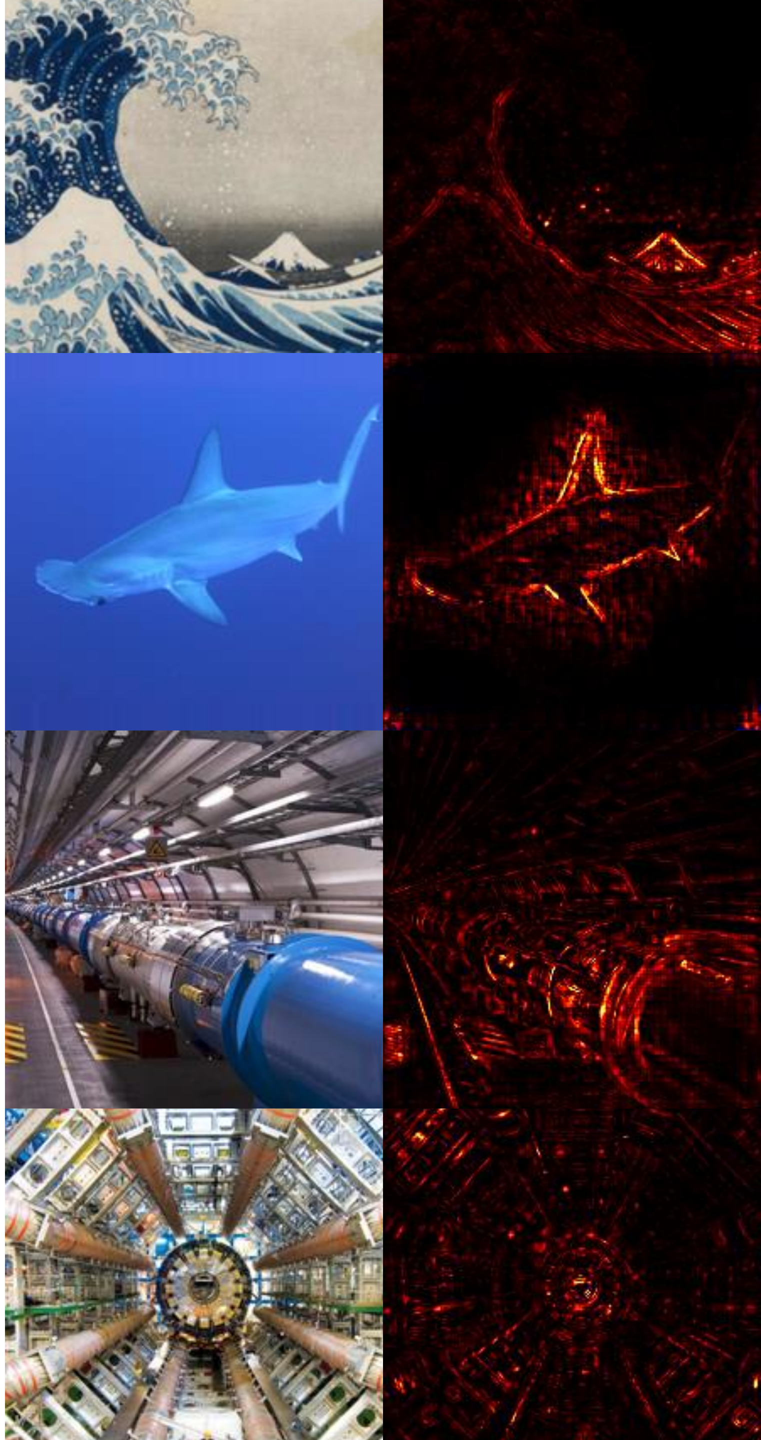
https://arxiv.org/pdf/2108.03125.pdf



Intrinsic explanations

- Model-specific – requires all model parameters
- Use gradients to quantify how a change in input would change the prediction (per event)
- Can be combined with randomisation of feature values

https://lrpserver.hhi.fraunhofer.de/image-classification



Robustness

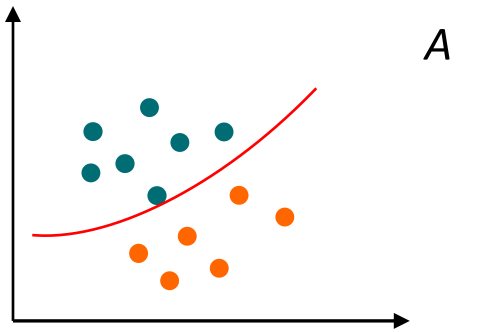
- Again using our ML model f and a test data point $\boldsymbol{x}$, how robust is the prediction $\boldsymbol{y}$ to a perturbation $\boldsymbol{x}'$ in the input?
  i.e. is

$$y = f(\boldsymbol{x}) \quad \text{equal to} \quad y' = f(\boldsymbol{x} + \boldsymbol{x}')?$$

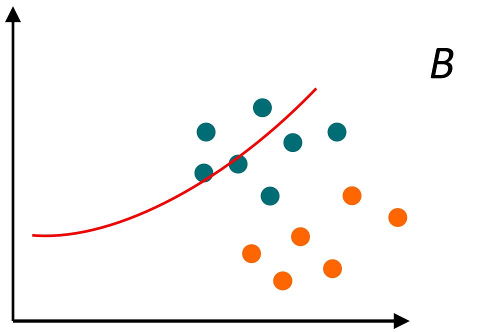
- Types of perturbations:
  - Random noise ($\mathrm{E}(x') = 0$)
  - Distribution shifts ($\mathrm{E}(x') \neq 0$)
  - Adversarial ($\boldsymbol{x}'$ selected so that $y \neq y'$)

Robustness under distribution shifts

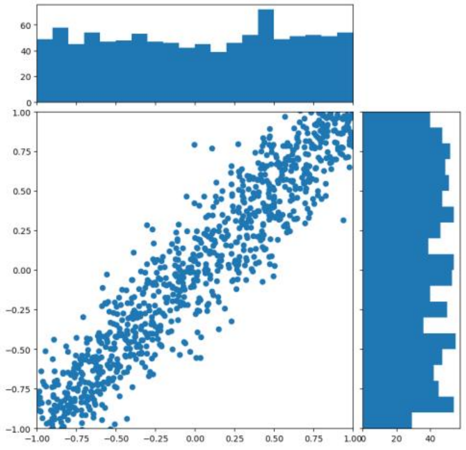
- Under distribution shifts, feature correlations remain but numerical values are consistently shifted
  - ML methods typically not happy about this

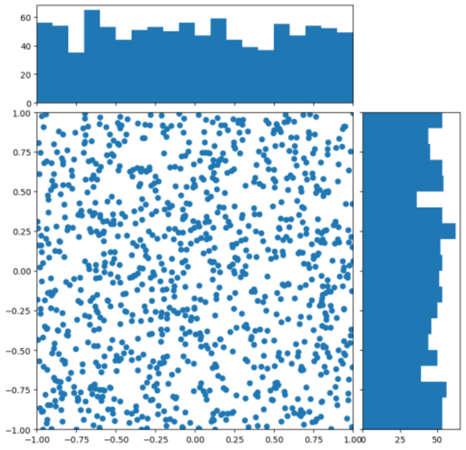


- Mitigated by *domain adaptation*
  - Methods applicable to analysis re-interpretation
    https://arxiv.org/abs/2207.09293



Robustness under random noise





- Can be improved through data augmentation (randomly sampling $\boldsymbol{x}'$ during training)
    - Requires augmentation to be realistic
- Common measures of robustness rely on the same sampling, at different noise levels
    - Estimate is only as good as the sampled values
    - Sampling from the marginal distribution leads to unlikely data points if features are correlated
- Realistic sampling gives
    - Data augmentation ☑
    - Feature relevance estimate ☑
    - Robustness estimate (at statistical level) ☑

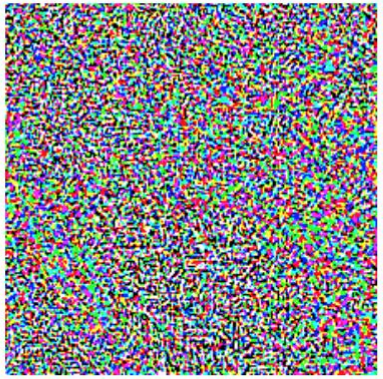
Robustness to adversarial examples

- Prediction accuracy will vary in different regions of feature space
- Adversarial attacks exploit this to find and insert the smallest x' that will change the prediction
- Won't see this in HEP data, ***but*** method is useful for identifying regions of low robustness

$\boldsymbol{x}$

y = «panda»

$\boldsymbol{x}'$



$\boldsymbol{x} + 0.007 \times \boldsymbol{x}'$

y = «gibbon»

Our projects

- Realistic data augmentation for improved robustness

  Framework for

  - Data augmentation (improves also generalisation)
  - Adversarial testing (model diagnostics and verification)
    - Mostly NN specific
  - Develop suitable robustness score for HEP ML