

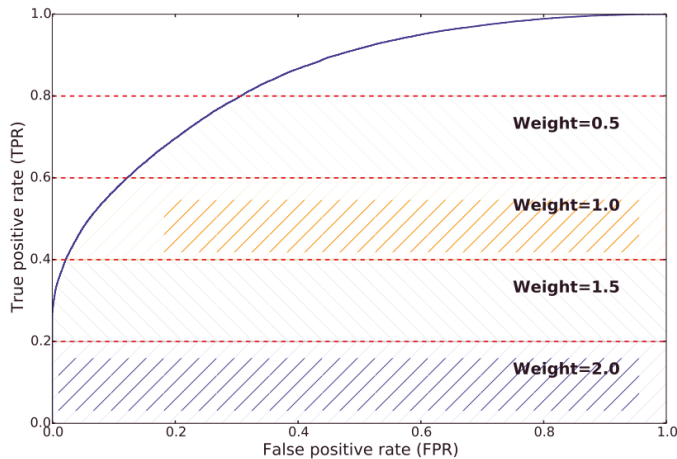
# Fisher information metrics for binary classifier evaluation and training

*Event selection for HEP precision measurements*

Andrea Valassi  
(CERN IT-DI-LCG)

QCHS XIII, Maynooth 2018  
Session H - Statistical Methods for Physics Analysis in the XXI Century

# Why and when I got interested in this topic



T. Blake et al., *Flavours of Physics: the machine learning challenge for the search of  $\tau \rightarrow \mu\mu\mu$  decays at LHCb* (2015, unpublished). [https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb\\_description\\_official.pdf](https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb_description_official.pdf) (accessed 15 January 2018)

The 2015 LHCb Kaggle ML Challenge:

- Develop an event selection in a search for  $\tau \rightarrow \mu\mu\mu$

ML binary classifier problem

- *Evaluation: the highest weighted AUC is the winner*

Figure 3: Weights assigned to the different segments of the ROC curve for the purpose of submission evaluation. The  $x$  axis is the False Positive Rate (FPR), while the  $y$  axis is True Positive Rate (TPR).

- First time I saw an *Area Under the Roc Curve (AUC)*
- My reaction:
  - What is the AUC? Which other scientific domains use it and why?
  - *Is the AUC relevant in HEP? Can we develop HEP-specific metrics?*

# Overview – the scope of this talk (1)

- *Different domains and/or problems → Need different metrics*
  - HEP and other domains require different metrics
  - Different problems within HEP also require different metrics

## ML playground



ML in HEP, David Rousseau, CHEP 2018, Sofia

39

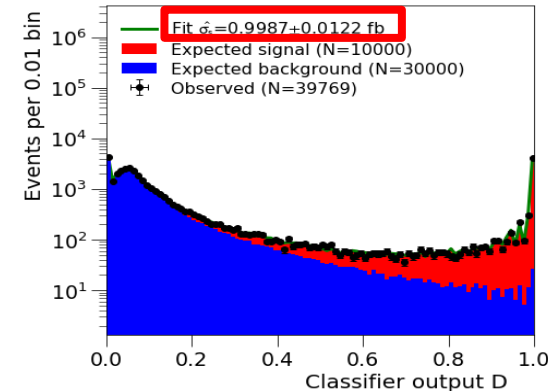
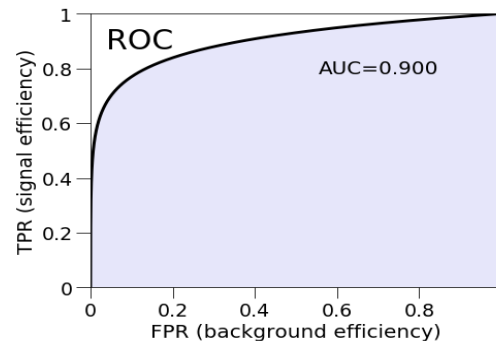
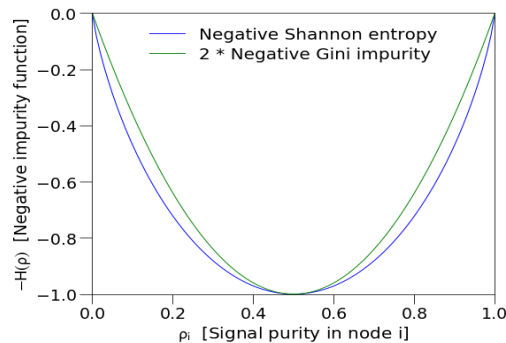
- This talk: one specific HEP example, *event selection to minimize statistical error  $\Delta\theta$  in an analysis for the point estimation of  $\theta$* 
  - I will not discuss: tracking, systematic errors, trigger, searches...

# Overview – the scope of this talk (2)

- *Different domains and/or problems → Need different metrics*
  - Always keep your final goal in mind
- This talk: one specific HEP example, *event selection to minimize statistical error  $\Delta\theta$  in an analysis for the point estimation of  $\theta$*
- *Whenever you take a decision, base it on the minimization of  $\Delta\theta$* 
  - Metrics for physics precision → final goal: minimize  $\Delta\theta$
  - Metrics for binary classifier evaluation → (is the AUC relevant?)
  - Metrics for binary classifier training → (are standard ML metrics relevant?)

# Training, Evaluation, Physics: one metric to bind them all?

Example: event selection using a Decision Tree for a parameter fit



## TRAINING

- (either) **Gini impurity**  
*Economics: inequality*  
*Ecology: diversity*
- (or) **Shannon information**  
*Information theory: entropy*

## EVALUATION

- **ROC Curve** (Receiver Operating Characteristic)  
*Signal detection: radar detection*  
*Psychophysics: sensory detection*
- **AUC** (Area Under the ROC Curve)  
*Radiology, Medicine: diagnostic accuracy*

## PHYSICS

- **Precision**  
*Parameter estimation: measurement error  $\Delta\theta$*

*Proposal: use metrics based on Fisher Information in all three steps*  
*(Fisher Information about  $\theta$   $\sim$  is  $I_\theta = 1/(\Delta\theta)^2$  – maximize  $I_\theta$  to minimize  $\Delta\theta$ )*

# Binary classifier evaluation – reminder

## Discrete classifiers: the confusion matrix

Binary decision:  
signal or background

$$PPV = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP} = 1 - FPR$$

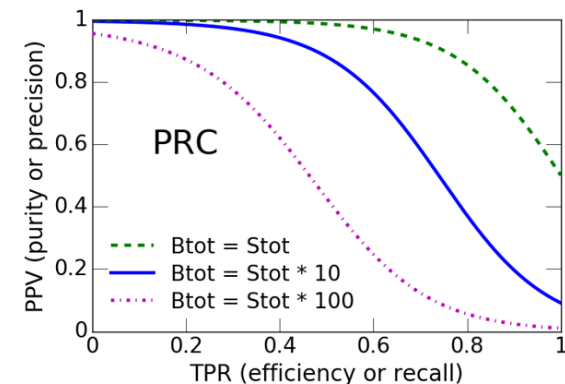
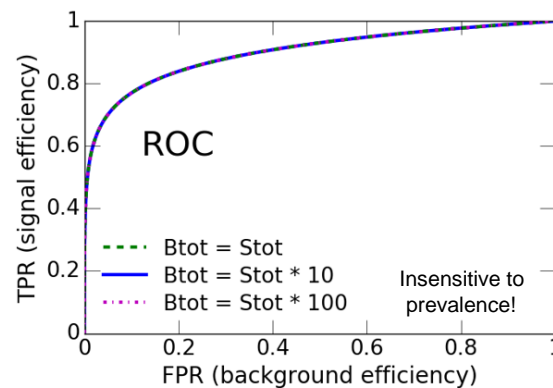
$$\text{Prevalence } \pi_s = \frac{S_{tot}}{S_{tot} + B_{tot}}$$

|   | <u>true class</u> : Positives<br>(HEP: <b>signal Stot</b> )        | <u>true class</u> : Negatives<br>(HEP: <b>background Btot</b> ) |
|---|--|---|
| <u>classified as</u> : positives<br>(HEP: <b>selected</b> ) | <b>True Positives (TP)</b><br>(HEP: selected signal <b>Ssel</b> )  | <b>False Positives (FP)</b><br>(HEP: selected bkg <b>Bsel</b> ) |
| <u>classified as</u> : negatives<br>(HEP: <b>rejected</b> ) | <b>False Negatives (FN)</b><br>(HEP: rejected signal <b>Srej</b> ) | <b>True Negatives (TN)</b><br>(HEP: rejected bkg <b>Brej</b> )  |

## Scoring classifiers: ROC and PRC curves

Continuous output:  
probability to be signal

Vary the binary decision  
by varying the cut  
on the scoring classifier



# Binary classifier evaluation in other domains

**Medical Diagnostics (MD)** → e.g. diagnostic accuracy for cancer

- Symmetric: all patients important, both truly ill (TP) and truly healthy (TN)
- Traditional  $\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$  was too sensitive to prevalence: moved to ROC
  - But now ROC is questioned as too insensitive to prevalence (imbalanced data)
- ROC-based analysis (because ROC insensitive to prevalence)
  - AUC interpretation: probability that diagnosis gives greater suspicion to a randomly chosen sick subject than to a randomly chosen healthy subject

**Information Retrieval (IR)** → e.g. find pages in Google search

- Asymmetric: distinction between relevant and non-relevant documents
- PRC-based evaluation: precision and recall (= purity and efficiency in HEP)
  - Single metric: e.g. Mean Average Precision ~ area under PRC (AUCPR)

Oversimplification: 
$$\boxed{\text{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s} \text{ (MD)} \quad \text{vs.} \quad \text{(IR)} \quad \boxed{\text{AUCPR} = \int_0^1 \rho d\epsilon_s}$$

# Binary classifiers: domain-specific challenges

- Questions valid for all domains, but with different answers:

- **Qualitative imbalance?**

- Are the two classes equally relevant?

In this talk I will focus on these three questions for signal/background discrimination in HEP

- **Quantitative imbalance?**

- Is the prevalence of one class much higher?

- **Prevalence known? Time invariance?**

- Is relative prevalence known in advance? Does it vary over time?

- **Dimensionality? Scale invariance?**

- Are all 4 elements of the confusion matrix needed?
  - Is the problem invariant under changes of some of these elements?

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002

- **Ranking? Binning?**

- Is the scoring classifier used to rank or partition the selected instances?

- **Instance weights?**

- Are all instances in a class equally important? Are instance counts enough?

# Evaluation: (main) specificities of HEP

1. Qualitative asymmetry: signal interesting, background irrelevant
  - Like Information Retrieval: use purity and efficiency (precision and recall)
    - *True Negatives and the AUC are irrelevant in HEP event selection*
    - *ROC alone is not enough, also need prevalence to interpret it*
2. Distribution fits: several disjoint bins, not just a global selection
  - Analyze *local signal efficiency and purity in each bin*, not just global ones
  - Frequent special case: **fits involving distributions of the scoring classifier**
3. Signal events not all equal: they may have different sensitivities
  - Example: only events close to a mass peak are sensitive to the mass

Illustrated in the following by three examples ( $1 = \text{FIP1}$ ,  $1+2 = \text{FIP2}$ ,  $1+2+3 = \text{FIP3}$ )

- Counting experiments (FIP1) vs. distribution fits (FIP2, FIP3)
- Total cross-section (FIP1, FIP2) vs. generic parameter fit (FIP3)

# Evaluation: Fisher Information Part (FIP)

- Evaluation of an event selection from its effect on the error  $\Delta\hat{\theta}$ 
  - *Compare to “ideal” case where there is no background*
- FIP: fraction of “ideal” FI that is retained by the real classifier
  - Range in  $[0,1]$  → 0 if no signal, 1 if select all signal and no background
  - *Qualitatively relevant: higher is better* → maximize FIP to minimize  $\Delta\hat{\theta}$
  - *Numerically meaningful: related to  $\Delta\hat{\theta}$*
- For a binned fit of  $\theta$  from a (1-D or multi-D) histogram:
  - Consider only statistical errors → sum information from the different bins

$$\text{FIP} = \frac{\mathcal{I}_{\theta}^{(\text{real classifier})}}{\mathcal{I}_{\theta}^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^m \epsilon_i \rho_i \times \frac{1}{S_i} \left( \frac{\partial S_i}{\partial \theta} \right)^2}{\sum_{i=1}^m \frac{1}{S_i} \left( \frac{\partial S_i}{\partial \theta} \right)^2}$$

Remember from the previous slide:

1. Qualitative asymmetry: use  $\epsilon$  and  $\rho$  (as in IR)
2. Distribution fit: need local  $\epsilon_i$  and  $\rho_i$  in each bin
3. Signal events not all equal: need sensitivity  $\frac{1}{S_i} \frac{\partial S_i}{\partial \theta}$

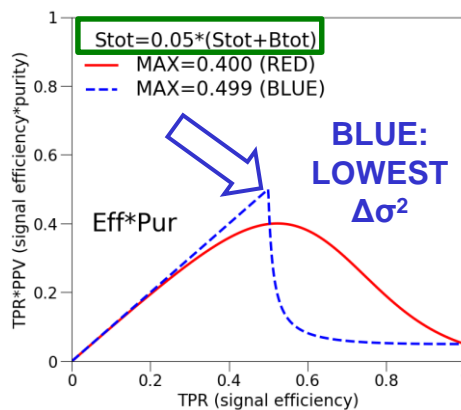
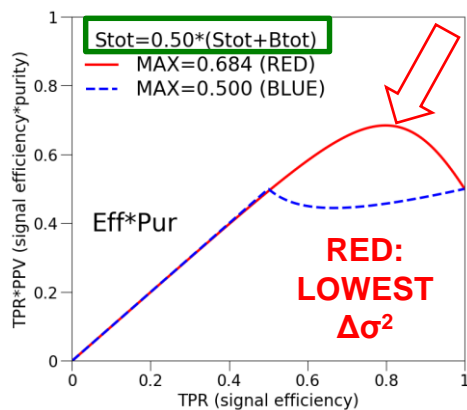
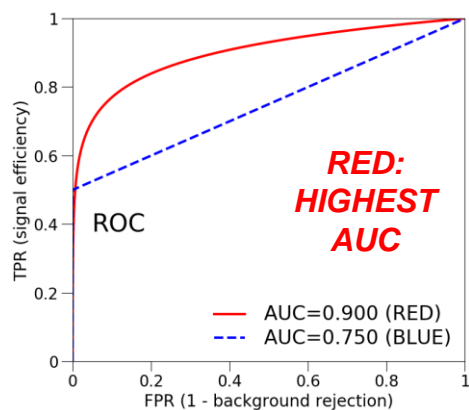
# [FIP1] Cross-section in counting experiment

- Counting experiment: measure a single number  $N_{\text{meas}}$ 
  - Well-known since decades: **maximize  $\epsilon_s * \rho$**  to minimize statistical errors
- FIP special case:
  - Counting experiment (1 bin)  $\rightarrow$  *global* signal efficiency and purity
  - **Cross-section fit  $\theta = \sigma_s \rightarrow$  all events have equal sensitivity  $\frac{1}{S_i} \frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$**

$$\text{FIP} = \frac{\mathcal{I}_{\theta}^{(\text{real classifier})}}{\mathcal{I}_{\theta}^{(\text{ideal classifier})}} = \frac{\sum_{i=1}^m \epsilon_i \rho_i \times \frac{1}{S_i} \left( \frac{\partial S_i}{\partial \theta} \right)^2}{\sum_{i=1}^m \frac{1}{S_i} \left( \frac{\partial S_i}{\partial \theta} \right)^2} \rightarrow \boxed{\text{FIP1} = \epsilon_s * \rho}$$

# Examples of issues in AUCs – *crossing ROCs*

- Cross-section measurement by counting experiment
  - Maximize  $FIP1 = \epsilon_s * \rho \rightarrow$  Minimize the statistical error  $\Delta\sigma^2$
- Compare two classifiers: red (AUC=0.90) and blue (AUC=0.75)
  - The red and blue ROCs cross (otherwise the choice would be obvious!)
- Choice of classifier achieving minimum  $\Delta\sigma^2$  *depends on*  $S_{tot}/B_{tot}$ 
  - *Signal prevalence 50%*: choose classifier with higher AUC (red)
  - *Signal prevalence 5%*: choose classifier with lower AUC (blue)
  - **AUC is irrelevant** – and **ROC is only useful if you also know prevalence**



|                        | FIP1 | AUC |
|------------------------|------|-----|
| Range in [0,1]         | YES  | YES |
| Higher is better       | YES  | NO  |
| Numerically meaningful | YES  | NO  |

# Optimal partitioning in distribution fits

- Does information  $I_\theta$  increase if I split a bin into two ( $n \rightarrow n_L + n_R$ )?
  - Information gain is  $\Delta I_\theta = \left( \rho_L \frac{1}{s_L} \frac{\partial s_L}{\partial \theta} - \rho_R \frac{1}{s_R} \frac{\partial s_R}{\partial \theta} \right)^2 * \frac{n_L n_R}{n_L + n_R}$
- Partition events using optimal binning variables ( $\rightarrow$  two examples)
  - For cross-sections  $\left( \frac{1}{s_i} \frac{\partial s_i}{\partial \sigma_s} = \frac{1}{\sigma_s} \right)$  : *separate bins with different  $\rho_i$*  ( $\rightarrow$  “FIP2”)
  - For a generic parameter  $\theta$  : *separate bins with different  $\rho_i \frac{1}{s_i} \frac{\partial s_i}{\partial \theta}$*  ( $\rightarrow$  “FIP3”)
- Practical ML consequences (focus on cross-section example):
  - Use the scoring classifier (i.e.  $\sim \rho$  !) to partition events, not to reject them
  - Train the scoring classifier to maximize the total Fisher information of the histogram binning, i.e. train it to maximize its partitioning power
    - Use Fisher Information as a node splitting criterion for decision tree training
    - *Use the decision tree more as a regression tree than as a classification tree*

# [FIP2] cross-section fit on the 1-D scoring classifier distribution – evaluation

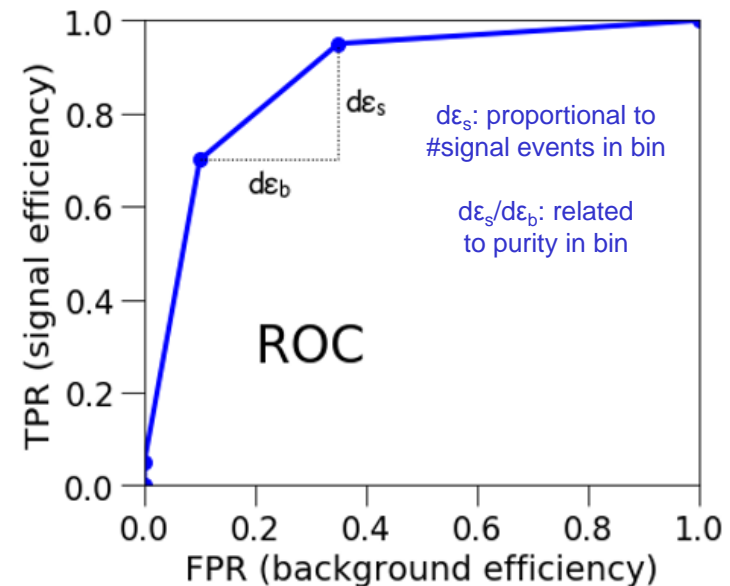
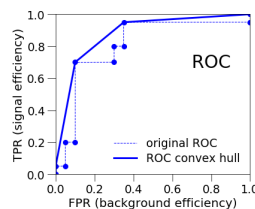
- FIP special case

- Cross-section: constant  $\frac{1}{s_i} \frac{\partial s_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$
- Fit on all events:  $\epsilon_i=1$  in all bins
- Fit scoring classifier: use ROC and prevalence to determine purity  $p_i$ 
  - Region of constant ROC slope is a region of constant signal purity

→ 
$$\text{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s} \frac{d\epsilon_b}{d\epsilon_s}}$$

Compare FIP2 to AUC: 
$$\text{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$$

- \*Technicality: convert ROC to *convex hull*
- ensure decreasing slope, i.e. decreasing purity
  - avoid staircase effect that artificially inflates FIP2 (bins of 100% purity: only signal or only background)

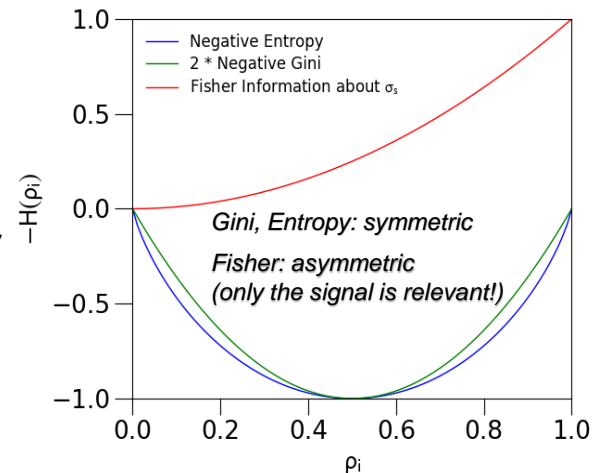


# [FIP2] cross-section fit on the 1-D scoring classifier distribution – training

- Is there a gain if I split a node into two ( $n \rightarrow n_L + n_R$ )?
  - Same question as in optimal partitioning: do I gain by splitting a bin?
- Gain depends on “impurity” function  $H(\rho)$ :  $\Delta = -n_L H(\rho_L) - n_R H(\rho_R) + n H(\rho)$ 
  - two standard choices: Shannon information (entropy) and Gini impurity
  - *I suggest a third option: Fisher information  $I_{\sigma_s}$  about the cross-section  $\sigma_s$*
- Surprise: different functions, but *Gini and Fisher gains are equal!*

$$\Delta_{\text{Fisher}} = \frac{(s_L n_R - s_R n_L)^2}{n_L n_R (n_L + n_R)} = \frac{\Delta_{\text{Gini}}}{2}$$

- So, Gini is OK for cross-sections (or searches?)
- *But more intuitive physics interpretation for Fisher*
- No practical gain here, but important principle
  - And proof-of-concept for generic parameter  $\theta$

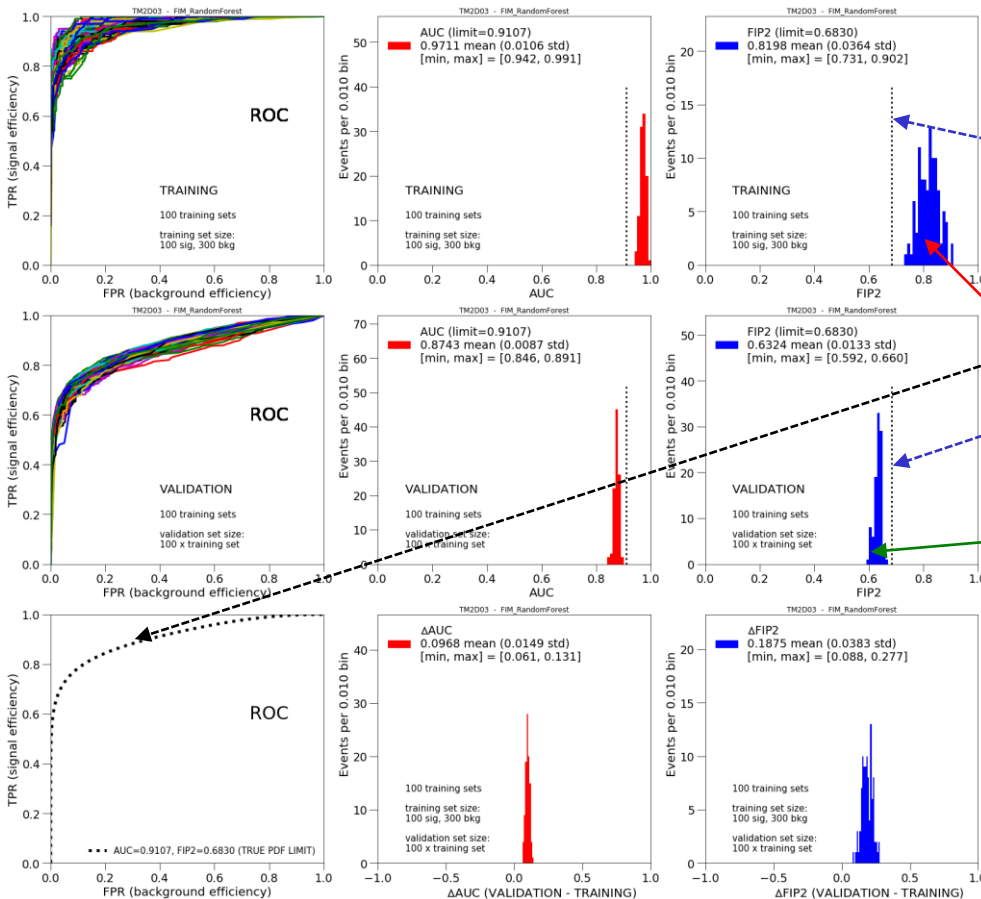


# Limits to knowledge

- FIP2 range is  $[0,1]$  → but it does not mean that 1 is achievable
  - 1 represents the *ideal* case where there is no background
- In some regions of phase space, signal and background events may be undistinguishable based on the available observations
  - There is a **limit ROC** which depends on the signal and background pdf's
  - There is a **limit FIP2** which depends on prevalence and the limit ROC
- Example – toy model, you know the real pdf's and prevalence
  - See next slide about overtraining

# Overtraining

- Using the same metric for training and evaluation also simplifies the interpretation of overtraining



- Example: toy model where you know the real pdf
  - You know the limit ROC
  - You know the limit FIP2
  - You want your validation FIP2 as close as possible to the limit, but it will be **lower**
  - To get there you maximize your training FIP2, but it will be **higher** than the real limit
    - You may trace back every increase to one node split
  - You may study the effects of things like `min_sample_leaf`

# [FIP3] generic parameter fits including the scoring classifier distribution – *work in progress*

- Not a cross-section, e.g. a coupling fit: signal events not all equal
  - [FIP2] Fit for  $\sigma_s \rightarrow$  should partition events into bins with different  $\rho_i$
  - [FIP3] Fit for  $\theta \rightarrow$  **should partition events into bins with different  $\rho_i \frac{1}{s_i} \frac{\partial s_i}{\partial \theta}$** 
    - Closely related to the “optimal observables” technique
- Example: 2-D fit for  $\theta$  of the  $\rho$  and  $\frac{1}{s} \frac{\partial s}{\partial \theta}$  distributions
  - Train a *regression tree* for  $\frac{1}{s} \frac{\partial s}{\partial \theta}$  (on MC weight derivative) using signal alone
  - Train a *regression tree* for  $\rho$  using signal (weighted by  $\frac{1}{s} \frac{\partial s}{\partial \theta}$ ) and background
  - Use Fisher Information about  $\theta$  as the gain function in both cases

*Boundary between classification and regression even more blurred*

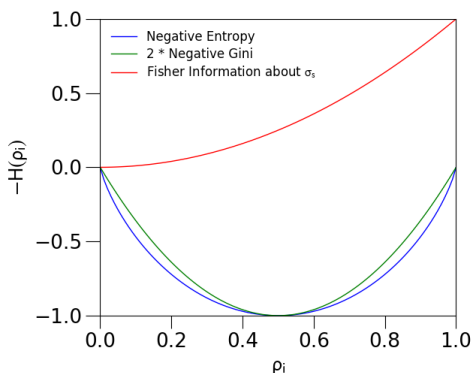
# Software technicalities

- I use Python (SciPy, iminuit, bits of rootpy) on SWAN at CERN
  - Thanks to all involved in these projects!
- **Custom impurity not available in sklearn DecisionTree's**
  - Planned for future sklearn releases (issue [#10251](#) and MR [#10325](#))?
  - I implemented a very simple DecisionTree from scratch, starting from the excellent iCSC [notebooks](#) by Thomas Keck (thanks!)
  - (May try XGBOOST in the future, where custom impurities are available)
- I plan to make the software available when I find the time...

# Conclusions and outlook

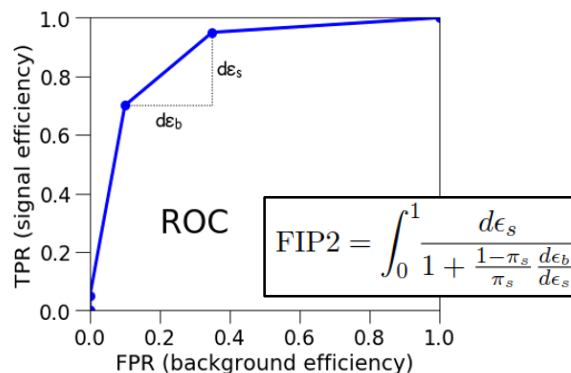
A paper will be on arxiv soon with all details

- Fisher Information: one metric to bind them all**



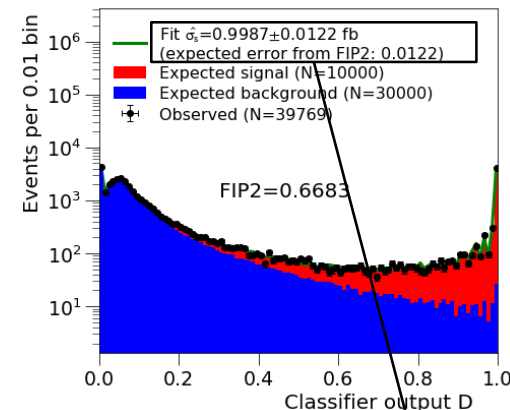
## TRAINING

- Fisher Information  
= measurement error



## EVALUATION

- Fisher Information  
= measurement error



## PHYSICS

- Fisher Information  
= measurement error

- Use scoring classifiers to partition events, not to reject them**
  - The boundary between classification and regression is blurred
- We must and can define our own HEP specific metrics**
  - I described one case, there are others (searches, systematics, tracking...)
  - Focus on signal. Describe distribution fits. Signal events are not all equal.
  - Can we please stop using the AUC now?* 😊

# Backup slides

Including selected slides from my previous IML talks  
in April (<https://indico.cern.ch/event/668017/contributions/2947015>)  
and January (<https://indico.cern.ch/event/679765/contributions/2814562>)

# Backup – statistical error in binned fits

- Data: *observed event counts*  $n_i$  in  $m$  bins of a (multi-D) distribution  $f(x)$ 
  - *expected event counts*  $y_i = f(x_i, \theta) dx$  depend on a parameter  $\theta$  that we want to fit
  - [NB here  $f$  is a differential cross section, it is not normalized to 1 like a pdf]
- Fitting  $\theta$  is like combining the independent measurements in the  $m$  bins
  - expected error on  $n_i$  in bin  $x_i$  is  $\Delta n_i = \sqrt{y_i} = \sqrt{f(x_i, \theta) dx}$
  - expected error on  $f(x_i, \theta)$  in bin  $x_i$  is  $\Delta f = f * \Delta n_i / n_i = \sqrt{f / dx}$
  - expected error on estimated  $\hat{\theta}_i$  in bin  $x_i$  is  $\frac{1}{(\Delta \hat{\theta})^2_{(\text{bin } dx)}} = \left( \frac{\partial f}{\partial \theta} \right)^2 \frac{1}{(\Delta f)^2} = \left( \frac{\partial f}{\partial \theta} \right)^2 \left( \frac{\sqrt{dx}}{\sqrt{f}} \right)^2 = \left( \frac{\partial f}{\partial \theta} \right)^2 \frac{dx}{f}$
  - expected error on estimated  $\hat{\theta}$  by combining the  $m$  bins is  $\left( \frac{1}{\Delta \hat{\theta}} \right)^2 = \int \frac{1}{f} \left( \frac{\partial f}{\partial \theta} \right)^2 dx$
- A bit more formally, joint probability for observing the  $n_i$  is  $P(\mathbf{n}; \theta) = \prod_{i=1}^m \frac{e^{-y_i} y_i^{n_i}}{n_i!}$ 
  - Fisher information on  $\theta$  from the data available is then
 
$$\mathcal{I}_\theta = E \left[ \frac{\partial \log P(\mathbf{n}; \theta)}{\partial \theta} \right]^2 \quad \text{i.e.} \quad \mathcal{I}_\theta = \sum_{i=1}^m \frac{1}{y_i} \left( \frac{\partial y_i}{\partial \theta} \right)^2 = \int \frac{1}{f} \left( \frac{\partial f}{\partial \theta} \right)^2 dx$$
  - The minimum variance achievable (Cramer-Rao lower bound) is  $(\Delta \hat{\theta})^2 = \text{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_\theta}$

# Optimal partitioning – information inflow

- Information about  $\theta$  in a binned fit  $\rightarrow \mathcal{I}_\theta = \sum_{i=1}^m \frac{1}{y_i} \left( \frac{\partial y_i}{\partial \theta} \right)^2$
- Can I reduce  $\Delta \hat{\theta}$  by splitting bin  $y_i$  into two bins?  $y_i = w_i + z_i$ 
  - Is the “information inflow” positive?  $\frac{1}{w_i} \left( \frac{\partial w_i}{\partial \theta} \right)^2 + \frac{1}{z_i} \left( \frac{\partial z_i}{\partial \theta} \right)^2 - \frac{1}{w_i + z_i} \left( \frac{\partial (w_i + z_i)}{\partial \theta} \right)^2 = \frac{(w_i \frac{\partial z_i}{\partial \theta} - z_i \frac{\partial w_i}{\partial \theta})^2}{w_i z_i (w_i + z_i)} \geq 0$
  - information increases (error  $\Delta \hat{\theta}$  decreases) if  $\frac{1}{w_i} \frac{\partial w_i}{\partial \theta} \neq \frac{1}{z_i} \frac{\partial z_i}{\partial \theta}$
- In the presence of background:  $\frac{1}{y_i} \frac{\partial y_i}{\partial \theta} = \rho_i \frac{1}{S_i} \frac{\partial S_i}{\partial \theta}$ 
  - information increases if  $\rho_w \frac{1}{s_w} \frac{\partial s_w}{\partial \theta} \neq \rho_z \frac{1}{s_z} \frac{\partial s_z}{\partial \theta}$
  - therefore: **try to partition the data into bins of different  $\rho_i \frac{1}{s_i} \frac{\partial s_i}{\partial \theta}$** 
    - for cross-section measurements,  $\frac{1}{S_i} \frac{\partial S_i}{\partial \sigma_s} = \frac{1}{\sigma_s}$  : *split into bins of different  $\rho_i$*
- Two important practical consequences:
  - *1. use scoring classifiers to partition the data, not to reject events*
  - *2. information can be used also for training classifiers like decision trees*

# Limited scope of this talk

- Different problems *also within HEP* require different metrics
- In this talk, I will focus on one specific problem:
  - Optimize event selection to minimize statistical errors in point estimation
- Three specific examples (I will focus on the second one)
  - [FIP1] *Total cross-section measurement in a counting experiment*
  - [FIP2] *Total cross-section measurement by distribution fit*
  - [FIP3] *Generic model parameter fit (e.g. mass/coupling) by distribution fit*
    - Even more specific: FIP2 and FIP3 use fits of the scoring classifier distribution

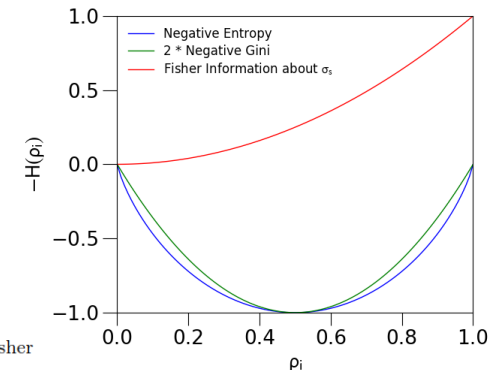
# FIP2 for training decision trees

- Decision Tree → **partition training set into nodes of different  $\rho_i$** 
  - The best split  $(n,s)=(n_L,s_L)+(n_R,s_R)$  maximizes  $\Delta = -n_L H(\rho_L) - n_R H(\rho_R) + n H(\rho)$

- Current metrics are Gini and entropy: add Fisher information!

- negative Gini impurity →  $-n_i H(\rho_i) = n_i \times [-2\rho_i(1 - \rho_i)]$
- Shannon information →  $-n_i H(\rho_i) = n_i \times [\rho_i \log_2 \rho_i + (1 - \rho_i) \log_2(1 - \rho_i)]$
- Fisher information on  $\sigma_s$  →  $-n_i H(\rho_i) = n_i \times [\rho_i^2]$

- Functions look different, but** (modulo a constant factor)...
- ... information gain is the same for Fisher and Gini!**



$$\Delta_{\text{Fisher}} = \frac{s_L^2}{n_L} + \frac{s_R^2}{n_R} - \frac{(s_L + s_R)^2}{n_L + n_R} = \frac{(s_L n_R - s_R n_L)^2}{n_L n_R (n_L + n_R)}$$

$$\frac{\Delta_{\text{Gini}}}{2} = -s_L \left(1 - \frac{s_L}{n_L}\right) - s_R \left(1 - \frac{s_R}{n_R}\right) + (s_L + s_R) \left(1 - \frac{s_L + s_R}{n_L + n_R}\right) = \Delta_{\text{Fisher}}$$

- But interpretation is clearer for Fisher: reduce the error on the fit
  - And this is a proof-of-concept for FIP3: split *into nodes of different  $\rho_i$*   $\frac{1}{s_i} \frac{\partial s_i}{\partial \theta}$

*Technicality: user-defined criteria for DecisionTree's will only be available in future sklearn releases  
 → I implemented a DecisionTree from scratch, reusing the excellent iCSC [notebooks](#) by Thomas Keck (thanks!)*

# FIP2 from the ROC (+prevalence) or from the PRC

- From the previous slide: 
$$\text{FIP2} = \frac{\sum_{i=1}^m \rho_i s_i}{\sum_{i=1}^m s_i}$$

FIP2: integrals on ROC and PRC, more relevant to HEP than AUC or AUCPR! (well-defined meaning for distribution fits)

- FIP2 from the ROC (+prevalence  $\pi_s = \frac{S_{\text{tot}}}{S_{\text{tot}} + B_{\text{tot}}}$ ):

$$\begin{aligned} S_{\text{sel}} &= S_{\text{tot}} \epsilon_s & \Rightarrow & s_i = dS_{\text{sel}} = S_{\text{tot}} d\epsilon_s \\ B_{\text{sel}} &= B_{\text{tot}} \epsilon_b & \Rightarrow & b_i = dB_{\text{sel}} = B_{\text{tot}} d\epsilon_b \end{aligned} \quad \Rightarrow \quad \rho_i = \frac{1}{1 + \frac{B_{\text{tot}}}{S_{\text{tot}}} \frac{d\epsilon_b}{d\epsilon_s}} \quad \Rightarrow \quad \text{FIP2} = \int_0^1 \frac{d\epsilon_s}{1 + \frac{1-\pi_s}{\pi_s} \frac{d\epsilon_b}{d\epsilon_s}}$$

Compare FIP2(ROC) to AUC

$$\text{AUC} = \int_0^1 \epsilon_s d\epsilon_b = 1 - \int_0^1 \epsilon_b d\epsilon_s$$

- FIP2 from the PRC:

$$\begin{aligned} S_{\text{sel}} &= S_{\text{tot}} \epsilon_s & \Rightarrow & s_i = dS_{\text{sel}} = S_{\text{tot}} d\epsilon_s \\ B_{\text{sel}} &= S_{\text{sel}} \left( \frac{1}{\rho} - 1 \right) & \Rightarrow & b_i = dB_{\text{sel}} = S_{\text{tot}} \left[ d\epsilon_s \left( \frac{1}{\rho} - 1 \right) - \epsilon_s \frac{d\rho}{\rho^2} \right] \end{aligned} \quad \Rightarrow \quad \rho_i = \frac{\rho}{1 - \frac{\epsilon_s}{\rho} \frac{d\rho}{d\epsilon_s}} \quad \Rightarrow \quad \text{FIP2} = \int_0^1 \frac{\rho d\epsilon_s}{1 - \frac{\epsilon_s}{\rho} \frac{d\rho}{d\epsilon_s}}$$

Compare FIP2(PRC) to AUCPR

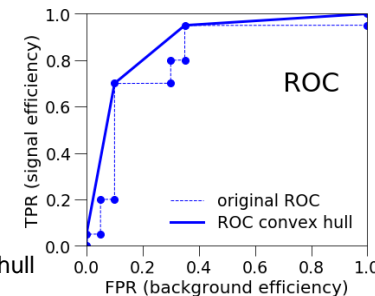
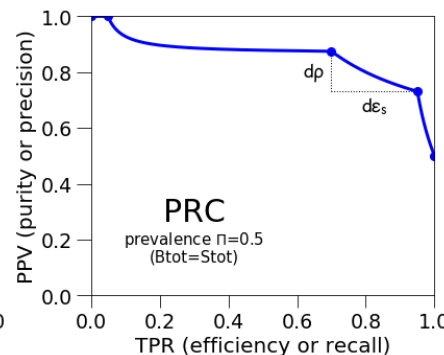
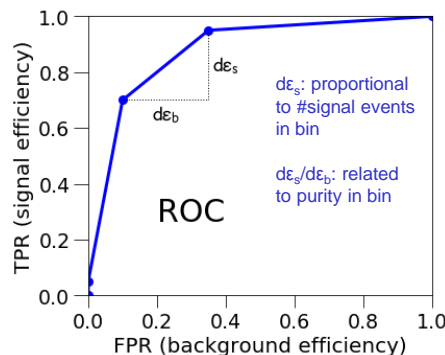
$$\text{AUCPR} = \int_0^1 \rho d\epsilon_s$$

- Easier calculation and interpretation from ROC (+prevalence) than from PRC

– *region of constant ROC slope\* = region of constant signal purity*

– decreasing ROC slope = decreasing purity

- technicality (my Python code): convert ROC to convex hull\*\* first

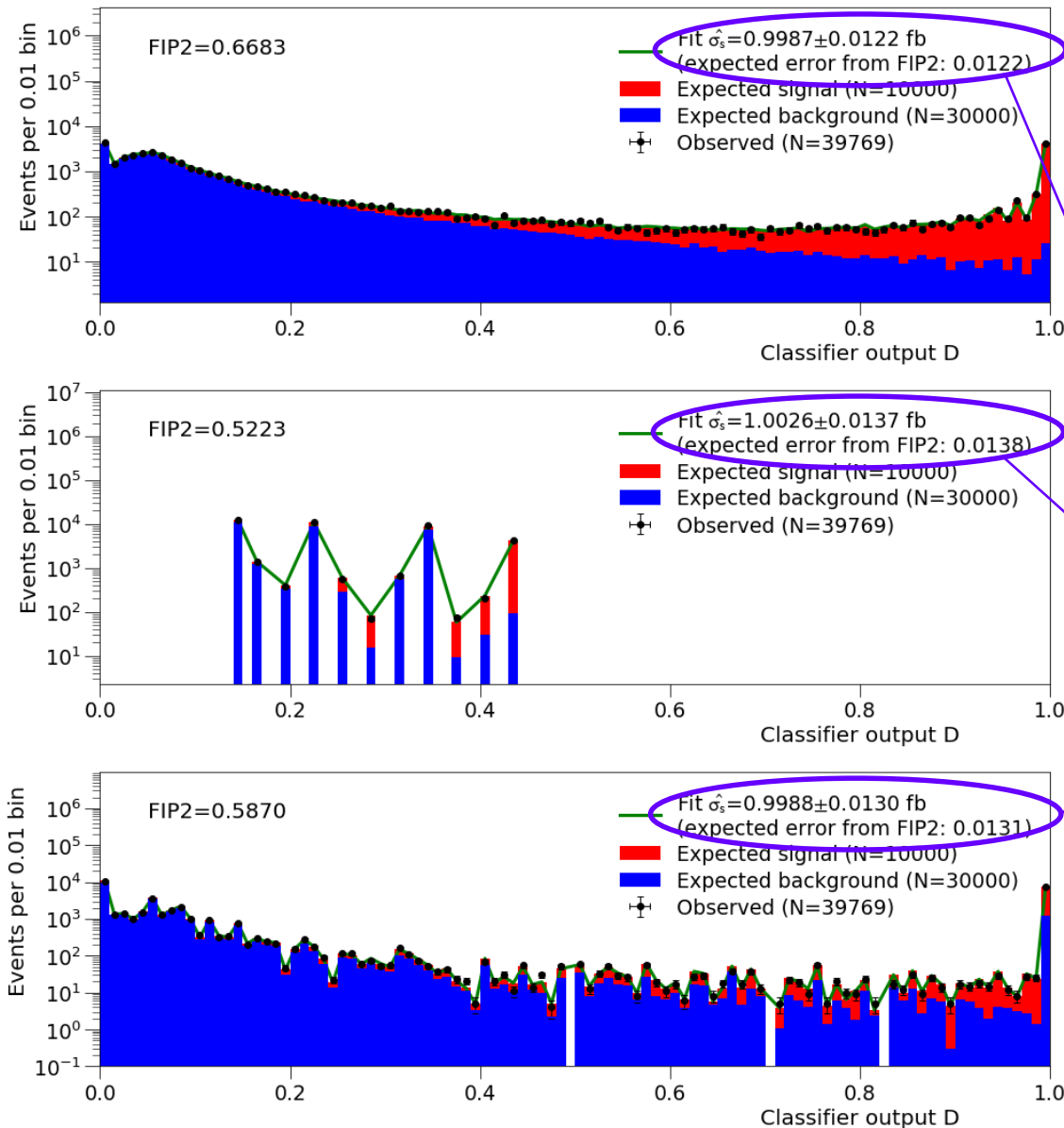


\*\*Convert ROC to convex hull

- ensure decreasing slope
- avoid staircase effect that would artificially inflate FIP2 (bins of 100% purity: only signal or only background)

\*ROC slopes are also discussed in medical literature in relation to diagnostic likelihood ratios [Choi 1998], but their use does not seem to be widespread(?)

# Sanity check



- Three random forests (on the same 2-D pdf)
  - reasonable
  - undertrained
  - overtrained
- Fit  $\sigma_s$  from the distribution of the classifier output
  - Errors consistent with FIP2

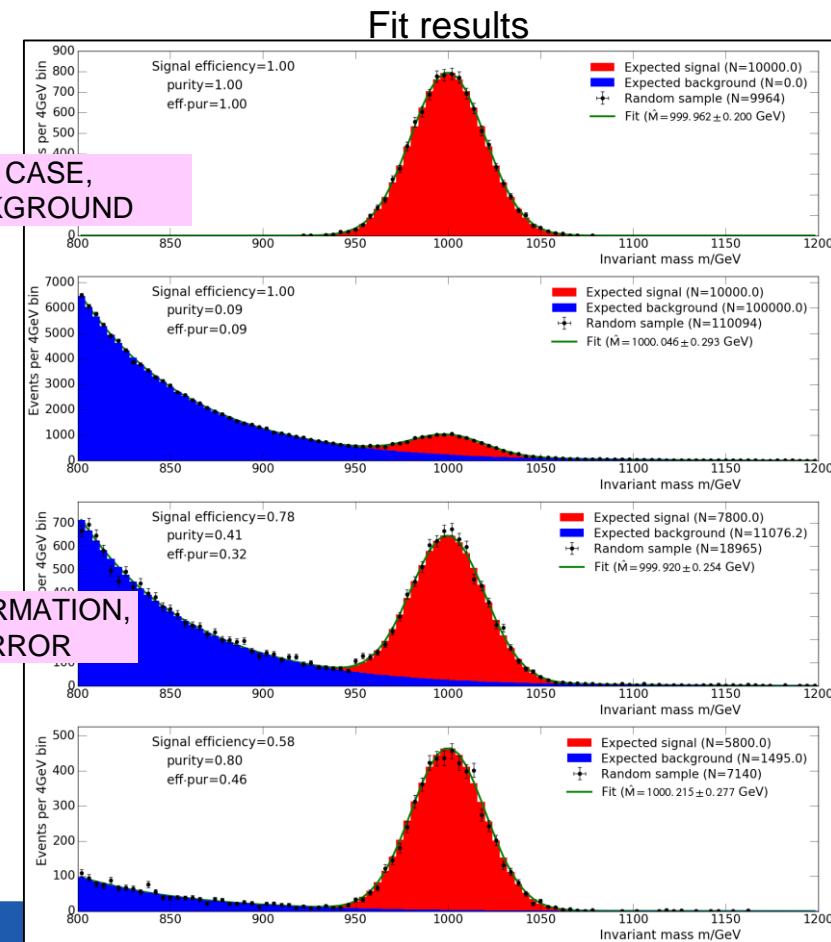
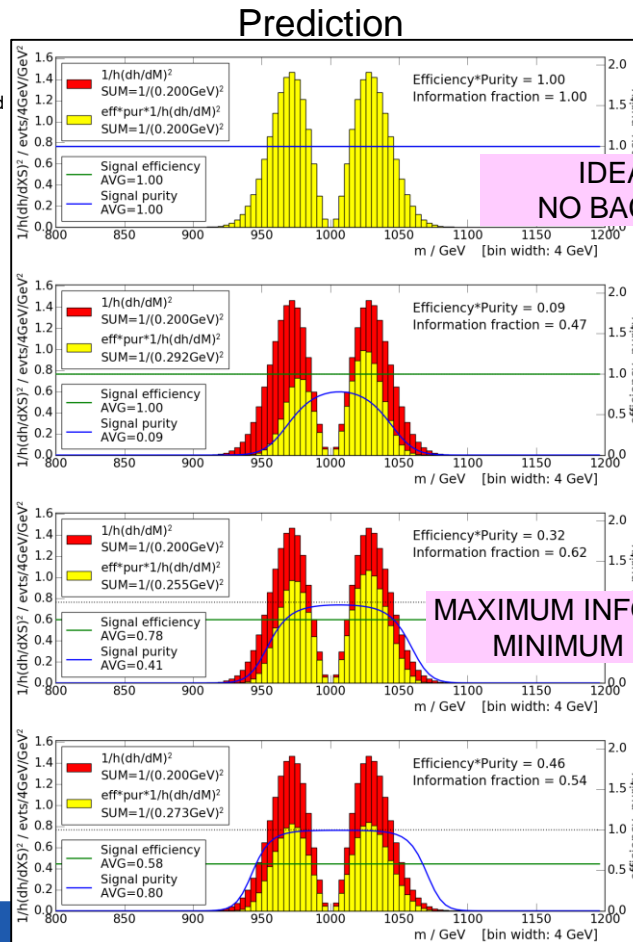
$$(\Delta\hat{\theta}^{(\text{real classifier})})^2 = \frac{1}{\text{FIP}} (\Delta\hat{\theta}^{(\text{ideal classifier})})^2$$

*My development environment: SciPy ecosystem, iminuit and bits of rootpy, on SWAN at CERN.  
Thanks to all involved in these projects!*

# M by 1D fit to m – visual interpretation

- Information after cuts:  $\sum_i \frac{1}{s_i} \left( \frac{\partial s_i}{\partial M} \right)^2 * \epsilon_i * \rho_i \rightarrow$  show the 3 terms in each bin  $i$ 
  - fit = combine N different measurements in N bins  $\rightarrow$  local  $\epsilon_i, \rho_i$  relevant!
  - important thing is: maximise purity, efficiency in bins with highest sensitivity!

Ideal case - yellow histogram (after cuts) coincides with and covers red histogram (ideal)



IDEAL CASE,  
NO BACKGROUND

MAXIMUM INFORMATION,  
MINIMUM ERROR

Red histogram:  
information per bin,  
ideal case  $\frac{1}{s_i} \left( \frac{\partial s_i}{\partial M} \right)^2$

Blue line: local  
purity in the bin,  $\rho_i$

Green line: local  
efficiency in the bin,  $\epsilon_i$

Yellow histogram:  
information per bin,  
after cuts  $\epsilon_i * \rho_i * \frac{1}{s_i} \left( \frac{\partial s_i}{\partial M} \right)^2$



# Event selection in HEP searches

- Statistical error in searches by counting experiment → “significance”
  - several metrics → but optimization always involves  $\epsilon_s$ ,  $\rho$  alone → TN irrelevant

$$Z_0 = \frac{S_{\text{sel}}}{\sqrt{S_{\text{sel}} + B_{\text{sel}}}} \implies (Z_0)^2 = S_{\text{tot}} \epsilon_s \rho$$

$Z_0$  – Not recommended? (confuses search with measuring  $\sigma_s$  once signal established)

C. Adam-Bourdarios et al., *The Higgs Machine Learning Challenge*, Proc. NIPS 2014 Workshop on High-Energy Physics and Machine Learning (HEPML2014), Montreal, Canada, PMLR 42 (2015) 19. <http://proceedings.mlr.press/v42/cowa14.html>

$Z_2$  – Most appropriate? (also used as “AMS2” in Higgs ML challenge)

$$Z_2 = \sqrt{2 \left( (S_{\text{sel}} + B_{\text{sel}}) \log\left(1 + \frac{S_{\text{sel}}}{B_{\text{sel}}}\right) - S_{\text{sel}} \right)}$$

$\implies$

$$(Z_2)^2 = 2S_{\text{tot}} \epsilon_s \left( \frac{1}{\rho} \log\left(\frac{1}{1-\rho}\right) - 1 \right) = S_{\text{tot}} \epsilon_s \rho \left( 1 + \frac{2}{3} \rho + \mathcal{O}(\rho^2) \right)$$

$$Z_3 = \frac{S_{\text{sel}}}{\sqrt{B_{\text{sel}}}} \implies (Z_3)^2 = S_{\text{tot}} \epsilon_s \frac{\rho}{1-\rho} = S_{\text{tot}} \epsilon_s \rho (1 + \rho + \mathcal{O}(\rho^2))$$

$Z_3$  (“AMS3” in Higgs ML) – Most widely used, but strictly valid only as an approximation of  $Z_2$  as an expansion in  $S_{\text{sel}}/B_{\text{sel}} \ll 1$ ?

$$\frac{S_{\text{sel}}}{B_{\text{sel}}} = \frac{\rho}{1-\rho} = \rho (1 + \rho + \mathcal{O}(\rho^2))$$

Expansion in  $\rho \ll 1$ ? – use the expression for  $Z_2$  if anything

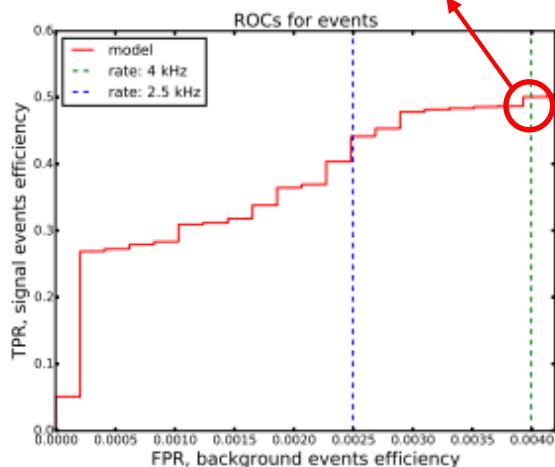
G. Punzi, *Sensitivity of searches for new signals and its optimization*, Proc. PhyStat2003, Stanford, USA (2003). [arXiv:physics/0308063v2](https://arxiv.org/abs/physics/0308063v2) [physics.data-an]  
G. Cowan, E. Gross, *Discovery significance with statistical uncertainty in the background estimate*, ATLAS Statistics Forum (2008, unpublished). <http://www.pp.rhul.ac.uk/~cowan/stat/notes/SigCalcNote.pdf> (accessed 15 January 2018)

R. D. Cousins, J. T. Linnemann, J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, Nucl. Instr. Meth. Phys. Res. A 595 (2008) 480. doi:10.1016/j.nima.2008.07.086  
G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C 71 (2011) 15. doi:10.1140/epjc/s10052-011-1554-0

- Several other interesting open questions → **beyond the scope of this talk**
  - optimization of systematics? → e.g. see AMS1 in Higgs ML challenge
  - predict significance in a binned fit? → integral over  $Z^2$  (=sum of log likelihoods)?

Maximise  $\epsilon_s$  at 4 kHz

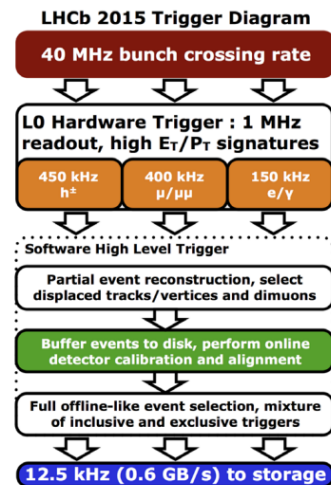
# Trigger



T. Likhomanenko et al., *LHCb Topological Trigger Reoptimization*, Proc. CHEP 2015, J. Phys. Conf. Series 664 (2015) 082025. doi:10.1088/1742-6596/664/8/082025

**Figure 2.** Trigger events ROC curve. An output rate of 2.5 kHz corresponds to an FPR of 0.25%, 4 kHz — 0.4%. Thus to find the signal efficiency for a 2.5 kHz output rate, we take 0.25% background efficiency and find the point on the ROC curve that corresponds to this FPR.

IIUC, 4kHz is  $\epsilon_b$  (FPR) = 0.4% of 1 MHz L0 hw rate

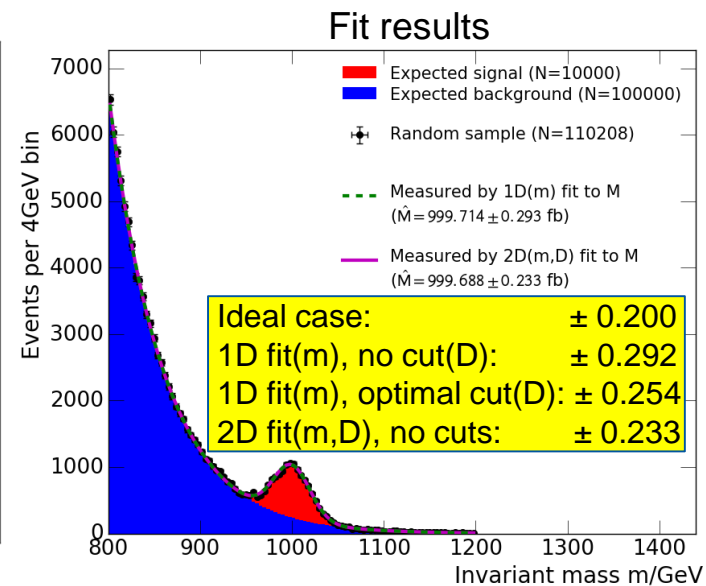
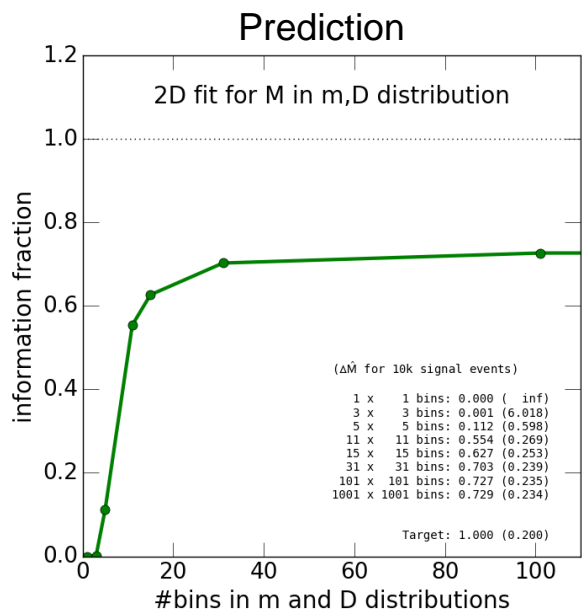
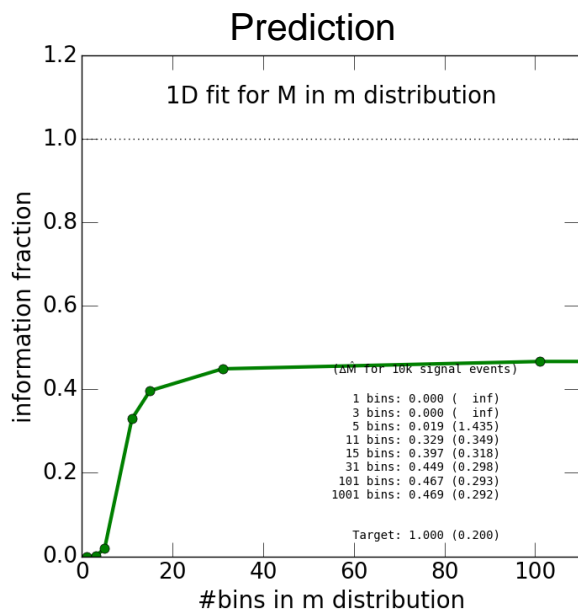


F. Dordei, *LHCb detector and trigger performance in Run II*, Proc. 5th Int. Conf. on New Frontiers in Physics (IC-NFP 2016), EPJ Web of Conferences 164, 01016 (2017). doi:10.1051/epjconf/201716401016

- Different meaning of absolute numbers in the confusion matrix
  - Trigger → events per unit time i.e. trigger rates
  - (Physics analyses → total event sample sizes i.e. total integrated luminosities)
- Binary classifier optimisation goal: maximise  $\epsilon_s$  for a given  $B_{sel}$  per unit time
  - i.e. maximise  $TP/(TP+FN)$  for a given  $FP \rightarrow TN$  irrelevant
- Relevant plot →  $\epsilon_s$  vs.  $B_{sel}$  per unit time (i.e.  $TPR$  vs  $FP$ )
  - ROC curve ( $TPR$  vs.  $FPR$ ) confusing – **AUC irrelevant**
  - e.g. maximise  $\epsilon_s$  for 4 kHz trigger rate, whether L0 rate is 1 MHz or 2MHz

# M by 2D fit – use classifier to partition, not to cut

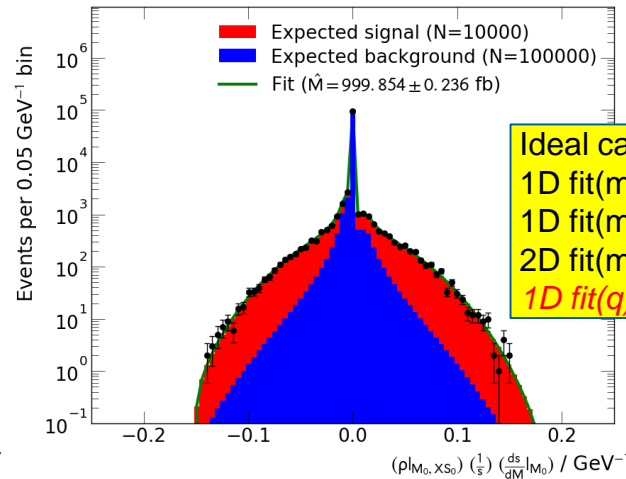
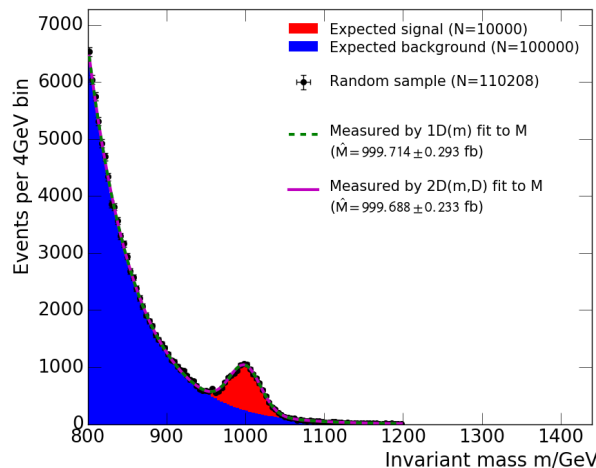
- Showed a fit for M on m, after a cut on D → can also fit in 2-D with no cuts  
– again, use the scoring classifier D to partition data, not to reject events
- Why is binning so important, especially using a discriminating variable?  
– next slide...



# Optimal partitioning – optimal variables

- The previous slide implies that  $q = \rho \frac{1}{s} \frac{\partial s}{\partial \theta}$  is an optimal variable to fit for  $\theta$ 
  - proof of concept  $\rightarrow$  1-D fit of  $q$  has the same precision on  $M$  as 2-D fit of  $(m, D)$
  - closely related to the “optimal observables” technique

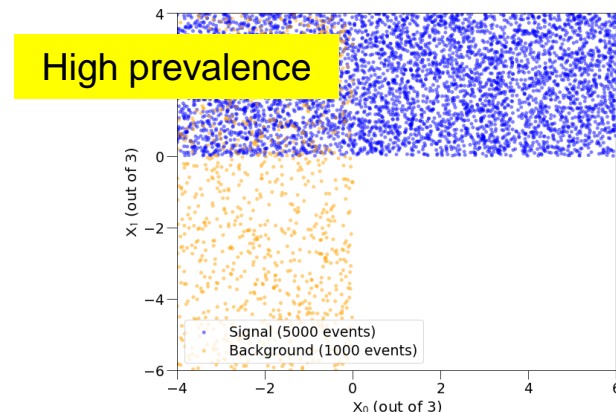
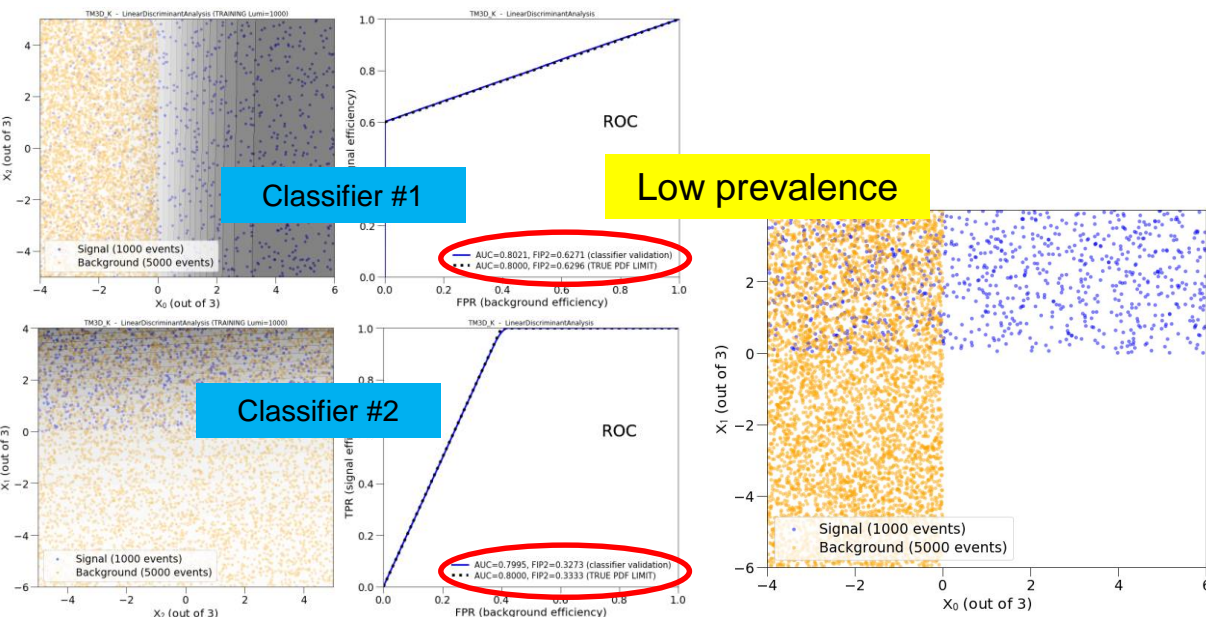
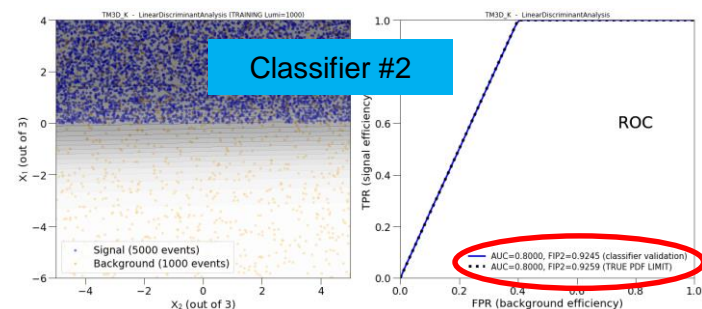
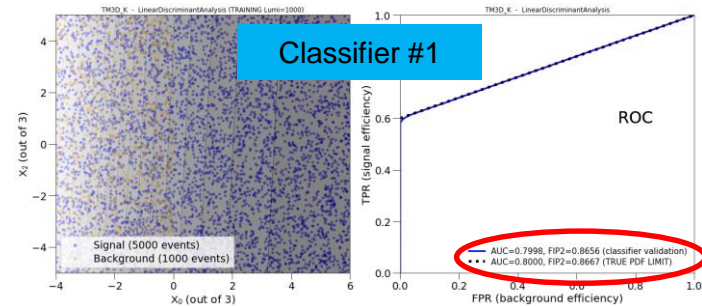
M. Davier, L. Duflot, F. LeDiberder, A. Rougé, *The optimal method for the measurement of tau polarization*, Phys. Lett. B 306 (1993) 411. doi:10.1016/0370-2693(93)90101-M  
M. Diehl, O. Nachtmann, *Optimal observables for the measurement of three-gauge-boson couplings in  $e^+e^- \rightarrow W^+W^-$* , Z. Phys. C 62 (1994) 397. doi:10.1007/BF01555899  
O. Nachtmann, F. Nagel, *Optimal observables and phase-space ambiguities*, Eur. Phys. J. C40 (2005) 497. doi:10.1140/epjc/s2005-02153-9



|                            |             |
|----------------------------|-------------|
| Ideal case:                | $\pm 0.200$ |
| 1D fit(m), no cut(D):      | $\pm 0.292$ |
| 1D fit(m), optimal cut(D): | $\pm 0.254$ |
| 2D fit(m,D), no cuts:      | $\pm 0.233$ |
| 1D fit(q):                 | $\pm 0.236$ |

- In practice: train one ML variable to reproduce  $\frac{1}{s} \frac{\partial s}{\partial \theta}$ ?
  - not needed for cross-sections or searches (this is constant)

- Prepared a model just to show that AUC is misleading
  - pdf with two useful features and a third random one
  - two classifiers, each trained only one useful feature
  - two prevalence scenarios:  $S/B=5$  and  $S/B=1/5$
- Same AUC (0.80) in all four cases
  - it is well known that AUC is insensitive to prevalence
  - ROC curves of the two classifiers cross*
- Low prevalence: FIP2 favors classifier #1 ( $0.63 > 0.33$ )
- High prevalence: FIP2 favors classifier #2 ( $0.87 < 0.93$ )
- Do not choose the best classifier based on AUC**
  - not for a cross-section fit on the classifier output, nor in general!



**FIP2 vs AUC**

# Understanding domain-specific challenges

- Many domain-specific details → but also general cross-domain questions:
  - **1. Qualitative imbalance?**
    - Are the two classes equally relevant?
  - **2. Quantitative imbalance?**
    - Is the prevalence of one class much higher?
  - **3. Prevalence known? Time invariance?**
    - Is relative prevalence known in advance? Does it vary over time?
  - **4. Dimensionality? Scale invariance?**
    - Are all 4 elements of the confusion matrix needed?
    - Is the problem invariant under changes of some of these elements?
  - **5. Ranking? Binning?**
    - Are all selected instances equally useful? Are they partitioned into subgroups?
- Point out properties of MED and IR, attempt a systematic analysis of HEP

M. Sokolova, G. Lapalme, *A Systematic Analysis of Performance Measures for Classification Tasks*, Information Processing and Management 45 (2009) 427.  
[doi:10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002)

# Medical diagnostics (1)

## and ML research

H. Sox, S. Stern, D. Owens, H. L. Abrams, *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions*, The National Academies Press (1989). doi:10.17226/1432

X. H. Zhou, D. K. McClish, N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine* (Wiley, 2002). doi:10.1002/9780470317082

**- Medical Diagnostics (MED)**  
does Mr. A. have cancer?

- Binary classifier optimisation goal: maximise “diagnostic accuracy”
  - patient / physician / society have different goals → many possible definitions
- Most popular metric: “accuracy”, or “probability of correct test result”:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \pi_s \times \text{TPR} + (1 - \pi_s) \times \text{TNR}$$

|  |  |
|--|--|
| TP (correctly diagnosed as ill)          | FP (truly healthy, but diagnosed as ill) |
| FN (truly ill, but diagnosed as healthy) | TN (correctly diagnosed as healthy)      |

- Symmetric → all patients important, both truly ill (TP) and truly healthy (TN)
- Also “by far the most commonly used metric” in ML research in the 1990s

F. J. Provost, T. Fawcett, *Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions*, Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, USA (1997). <https://aaai.org/Library/KDD/1997/kdd97-007.php>

L. B. Lusted, *Signal Detectability and Medical Decision-Making*, Science 171 (1971) 1217 doi:10.1126/science.171.3977.1217

J. A. Swets, *Measuring the accuracy of diagnostic systems*, Science 240 (1988) 1285. doi:10.1126/science.3287615

- Since the ‘90s → shift from ACC to ROC in the MED and ML fields
  - TPR (sensitivity) and TNR (specificity) studied separately
    - solves ACC limitations (imbalanced or unknown prevalence – rare diseases, epidemics)
  - Evaluation often AUC-based → two perceived advantages *for MED and ML fields*
    - **AUC interpretation: “probability that test result of randomly chosen sick subject indicates greater suspicion than that of randomly chosen healthy subject”**
    - ROC comparison without prior  $D_{\text{thr}}$  choice (prevalence-dependent  $D_{\text{thr}}$  choice)

F. J. Provost, T. Fawcett, R. Kohavi, *The Case against Accuracy Estimation for Comparing Induction Algorithms*, Proc. 15th Int. Conf. on Machine Learning (ICML '98), Madison, USA (1998). <https://www.researchgate.net/publication/2373067>

A. P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition 30 (1997) 1145. doi:10.1016/S0031-3203(96)00142-2

J. A. Hanley, B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology 143 (1982) 29. doi:10.1148/radiology.143.1.7063747



# Medical diagnostics (2)

## and ML research

- ROC and AUC metrics → currently widely used in the MED and ML fields
  - Remember: moved because *ROC better than ACC with imbalanced data sets*
- Limitation: evidence that *ROC not so good for highly imbalanced data sets*
  - may provide an overly optimistic view of performance
  - PRC may provide a more informative assessment of performance in this case
    - PRC-based reanalysis of some data sets in life sciences has been performed
- Very active area of research → other options proposed (CROC, cost models)
  - Take-away message: *ROC and AUC not always the appropriate solutions*

J. Davis, M. Goadrich, *The relationship between Precision-Recall and ROC curves*, Proc. 23rd Int. Conf. on Machine Learning (ICML '06), Pittsburgh, USA (2006). [doi:10.1145/1143844.1143874](https://doi.org/10.1145/1143844.1143874)

C. Drummond, R. C. Holte, *Explicitly representing expected cost: an alternative to ROC representation*, Proc. 6th Int. Conf. on Knowledge Discovery and Data Mining (KDD-00), Boston, USA (2000). [doi:10.1145/347090.347126](https://doi.org/10.1145/347090.347126)

D. J. Hand, *Measuring classifier performance: a coherent alternative to the area under the ROC curve*, Mach Learn (2009) 77: 103. [doi:10.1007/s10994-009-5119-5](https://doi.org/10.1007/s10994-009-5119-5)

S. J. Swamidass, C.-A. Azencott, K. Daily, P. Baldi, *A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval*, Bioinformatics 26 (2010) 1348. [doi:10.1093/bioinformatics/btq140](https://doi.org/10.1093/bioinformatics/btq140)

D. Berrar, P. Flach, *Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them)*, Briefings in Bioinformatics 13 (2012) 83. [doi:10.1093/bib/bbr008](https://doi.org/10.1093/bib/bbr008)

H. He, E. A. Garcia, *Learning from Imbalanced Data*, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263. [doi:10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239)

T. Saito, M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*, PLoS One 10 (2015) e0118432. [doi:10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432)

- Qualitative distinction between “relevant” and “non-relevant” documents
  - also a very large quantitative imbalance
- Binary classifier optimisation goal: make users happy in web searches
  - minimise # relevant documents not retrieved → maximise “recall” i.e. efficiency
  - minimise # of irrelevant documents retrieved → maximise “precision” i.e. purity
  - retrieve the more relevant documents first → ranking very important
  - maximise speed of retrieval
- IR-specific metrics to evaluate classifiers based on the PRC (i.e. on  $\epsilon_s$ ,  $\rho$ )
  - unranked evaluation → e.g. F-measures 
$$F_\alpha = \frac{1}{\alpha/\epsilon_s + (1-\alpha)/\rho}$$
    - $\alpha \in [0,1]$  *tradeoff between recall and precision* → equal weight gives  $F1 = \frac{2\epsilon_s\rho}{\epsilon_s + \rho}$
  - ranked evaluation → precision at k documents, mean average precision (MAP), ...
    - MAP approximated by the Area Under the PRC curve (AUCPR)

C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2008).  
<https://nlp.stanford.edu/IR-book>

NB: Many different of meanings of “Information”!  
IR (web documents), HEP (Fisher), Information Theory (Shannon)...

| Domain<br>Property   | Medical diagnostics   | Information retrieval   | HEP event selection  |
|--|---|---|--|
| Qualitative class imbalance  | <b><u>NO.</u></b> Healthy and ill people have “equal rights”.<br><i>TN are relevant.</i>  | <b><u>YES.</u></b> “Non-relevant” documents are a nuisance.<br><i>TN are irrelevant.</i>  | <b><u>YES.</u></b> Background events are a nuisance.<br><i>TN are irrelevant.</i>  |
| Quantitative class imbalance   | <b><u>From small to extreme.</u></b><br>From common flu to very rare disease.   | <b><u>Generally very high.</u></b><br>Only very few documents in a repository are relevant.   | <b><u>Generally extreme.</u></b><br>Signal events are swamped in background events.  |
| Varying or unknown prevalence $\pi$  | <b><u>Varying and unknown.</u></b><br>Epidemics may spread.   | <b><u>Varying and unknown</u></b><br>in general (e.g. WWW).   | <b><u>Constant in time</u></b><br>(quantum cross-sections).<br><b><u>Unknown</u></b> for searches.<br><b><u>Known</u></b> for precision measurements.  |
| Dimensionality and invariances<br><small>M. Sokolova, G. Lapalme, A Systematic Analysis of Performance Measures for Classification Tasks, Information Processing and Management 45 (2009) 427. doi:10.1016/j.ipm.2009.03.002</small> | <b><u>3 ratios <math>\epsilon_s</math>, <math>\epsilon_b</math>, <math>\pi</math> + scale.</u></b><br>New metrics under study because ROC ignores $\pi$ .<br>Costs scale with $N_{tot}$ . | <b><u>2 ratios <math>\epsilon_s</math>, <math>\rho</math> + scale.</u></b><br>$\epsilon_s$ , $\rho$ enough in many cases.<br>Costs and speed scale with $N_{tot}$ .<br>Show only $N_{sel}$ docs in one page.<br><i>TN are irrelevant.</i> | <b><u>2 ratios <math>\epsilon_s</math>, <math>\rho</math> + scale.</u></b><br>$\epsilon_s$ , $\rho$ enough in many cases.<br>Lumi is needed for: trigger, syst. vs stat., searches.<br><i>TN are irrelevant.</i> |
| Different use of selected instances  | <b><u>Binning – NO.</u></b><br><b><u>Ranking – YES?</u></b><br>Treat with higher priority patients who are more likely to be ill?   | <b><u>Binning – NO.</u></b><br><b><u>Ranking – YES.</u></b><br>Precision at k, R-precision, MAP all involve <u>global</u> precision-recall (“top $N_{sel}$ documents retrieved”)  | <b><u>Binning – YES.</u></b><br>Fits to distributions:<br><u>local <math>\epsilon_s</math>, <math>\rho</math> in each bin</u><br>rather than global $\epsilon_s$ , $\rho$ .                                      |

# Different HEP problems → Different metrics

## Binary classifiers for HEP event selection (signal-background discrimination)

|   |                                    |   |  |   |  |
|---|------------------------------------|---|--|---|--|
| <div>Statistical error minimization</div> <div>(or statistical significance maximization)</div> | Cross-section (1-bin counting)     | Only 2 or 3 global/local variables – TN, AUC irrelevant | 2 variables: global $\epsilon_s$ , $\rho$ (given $S_{\text{tot}}$ )  | Maximise $S_{\text{tot}} * \epsilon_s * \rho$ (at any $S_{\text{tot}}$ )  |  |
|   | Searches (1-bin counting )         |   | Simple and CCGV – 2 variables: global $S_{\text{sel}}$ , $B_{\text{sel}}$ (or equivalently $\epsilon_s$ , $\rho$ ) | Maximise $\frac{S_{\text{sel}}}{\sqrt{S_{\text{sel}}+B_{\text{sel}}}}$ (i.e. $\sqrt{S_{\text{tot}}*\epsilon_s*\rho}$ )  |  |
|   |                                    |   |  | Maximise $\sqrt{2((S_{\text{sel}} + B_{\text{sel}}) \log\left(1 + \frac{S_{\text{sel}}}{B_{\text{sel}}}\right) - S_{\text{sel}})}$  |  |
|   |                                    |   | HiggsML – 2 variables: global $S_{\text{sel}}$ , $B_{\text{sel}}$  | Maximise $\sqrt{2((S_{\text{sel}} + B_{\text{sel}} + K) \log\left(1 + \frac{S_{\text{sel}}}{B_{\text{sel}}+K}\right) - S_{\text{sel}})}$  |  |
|   |                                    |   | Punzi – 2 variables: global $\epsilon_s$ , $B_{\text{sel}}$  | Maximise $\frac{\epsilon_s}{A/2 + \sqrt{B_{\text{sel}}}}$   |  |
|   | Cross-section (binned fits)        |   | 2 variables:<br>local $\epsilon_{s,i}$ and $\rho_i$ in each bin<br>(given $s_{\text{tot},i}$ in each bin)          | Maximise $\sum_i s_{\text{tot},i} * \epsilon_{s,i} * \rho_i$<br>Partition in bins of equal $\rho_i$   |  |
|   | Parameter estimation (binned fits) |   |  | Maximise $\sum_i s_{\text{tot},i} * \epsilon_{s,i} * \rho_i * \left(\frac{1}{s_{\text{tot},i}} \frac{\partial s_{\text{tot},i}}{\partial \theta}\right)^2$<br>Partition in bins of equal $\rho_i * \left(\frac{1}{s_{\text{tot},i}} \frac{\partial s_{\text{tot},i}}{\partial \theta}\right)$ |  |
|   | Searches (binned fits)             |   | 3 variables: local $s_{\text{sel}}$ , $s_{\text{tot}}$ , $s_{\text{sel}}$ in each bin (2 counts or ratios enough?) | Maximise a sum? *   |  |
| Statistical + Systematic error minimization   |                                    |   |  | 3 variables: $\epsilon_s$ , $\rho$ , lumi (lumi: tradeoff stat. vs. syst.)  | No universal recipe *<br>(may use local $S_{\text{sel}}$ , $B_{\text{sel}}$ in side band bins) |
| Trigger optimization  |                                    |   |  | 2 variables: global $B_{\text{sel}}$ /time, global $\epsilon_s$   | Maximise $\epsilon_s$ at given trigger rate  |

## Binary classifiers for HEP problems other than event selection

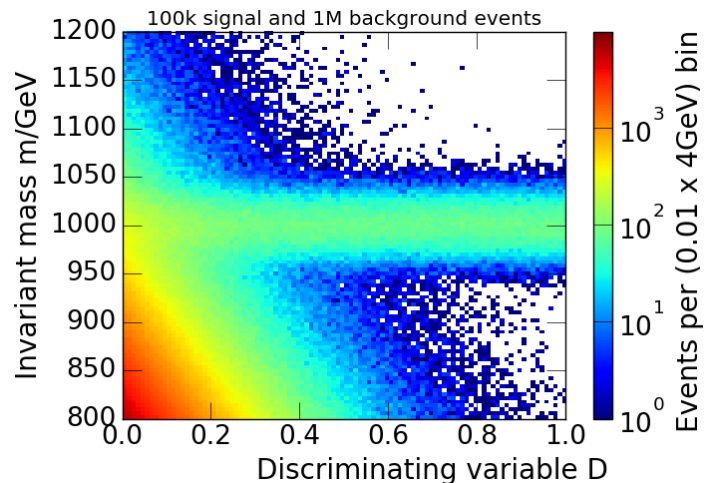
|   |   |                                   |
|---|---|-----------------------------------|
| <b>Tracking and Particle-ID optimizations</b> | All 4 variables? * (NB: TN is relevant) | ROC relevant – is AUC relevant? * |
| <b>Other? *</b>                               | ? *                                     | ? *                               |

\* Many open questions for further research



# Numerical tests with a toy model

- I used a simple toy model to make some numerical tests
  - Verify that my formulas are correct – and also illustrate them graphically
  - Two-dimensional distribution (m,D) → signal Gaussian, background exponential
- Two measurements:
  - total cross-section measurement by counting and 1-D or 2-D fit
  - mass measurement by 1-D or 2-D fits
- Details in the backup slides



*Using scipy / matplotlib / numpy  
and iminuit in Python from SWAN*

# M by 1D fit to m – optimizing the classifier

- Choose operating point  $D_{\text{thr}}$  optimizing information fraction for  $\theta=M$  in m-fit
  - NB: different to operating point maximising  $\varepsilon^*\rho$  (IF for  $\theta=\sigma_s$  in a 1-bin fit)
- To compute IF as sum over bins  $\rightarrow$  need average  $\frac{1}{s} \frac{\partial s}{\partial \theta}$  in each bin
  - proof-of-concept  $\rightarrow$  integrate by toy MC with *event-by-event weight derivatives*
    - in a real MC, could save  $\frac{1}{|\mathcal{M}|^2} \frac{\partial |\mathcal{M}|^2}{\partial \theta}$  for the matrix element squared  $|\mathcal{M}|^2$

