Applied Statistics

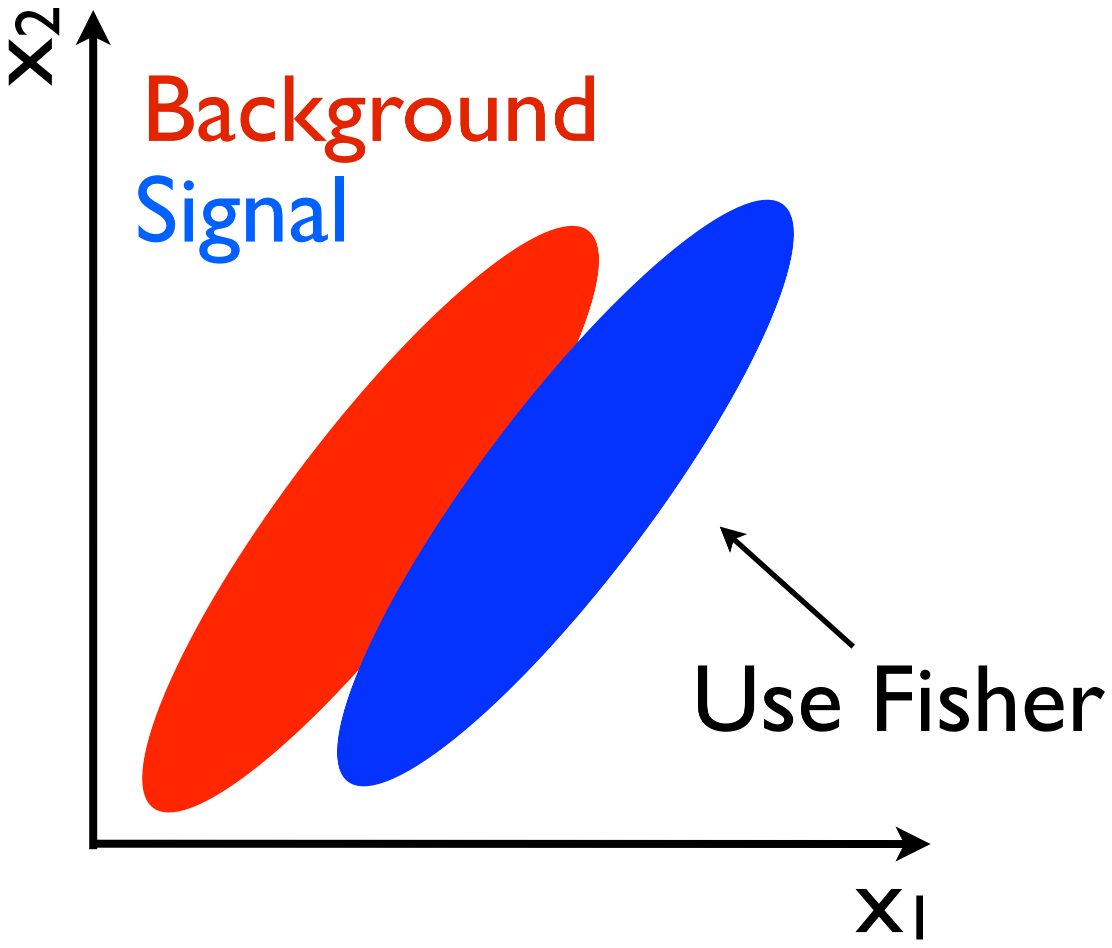
Week 6 - Multivariate Analysis

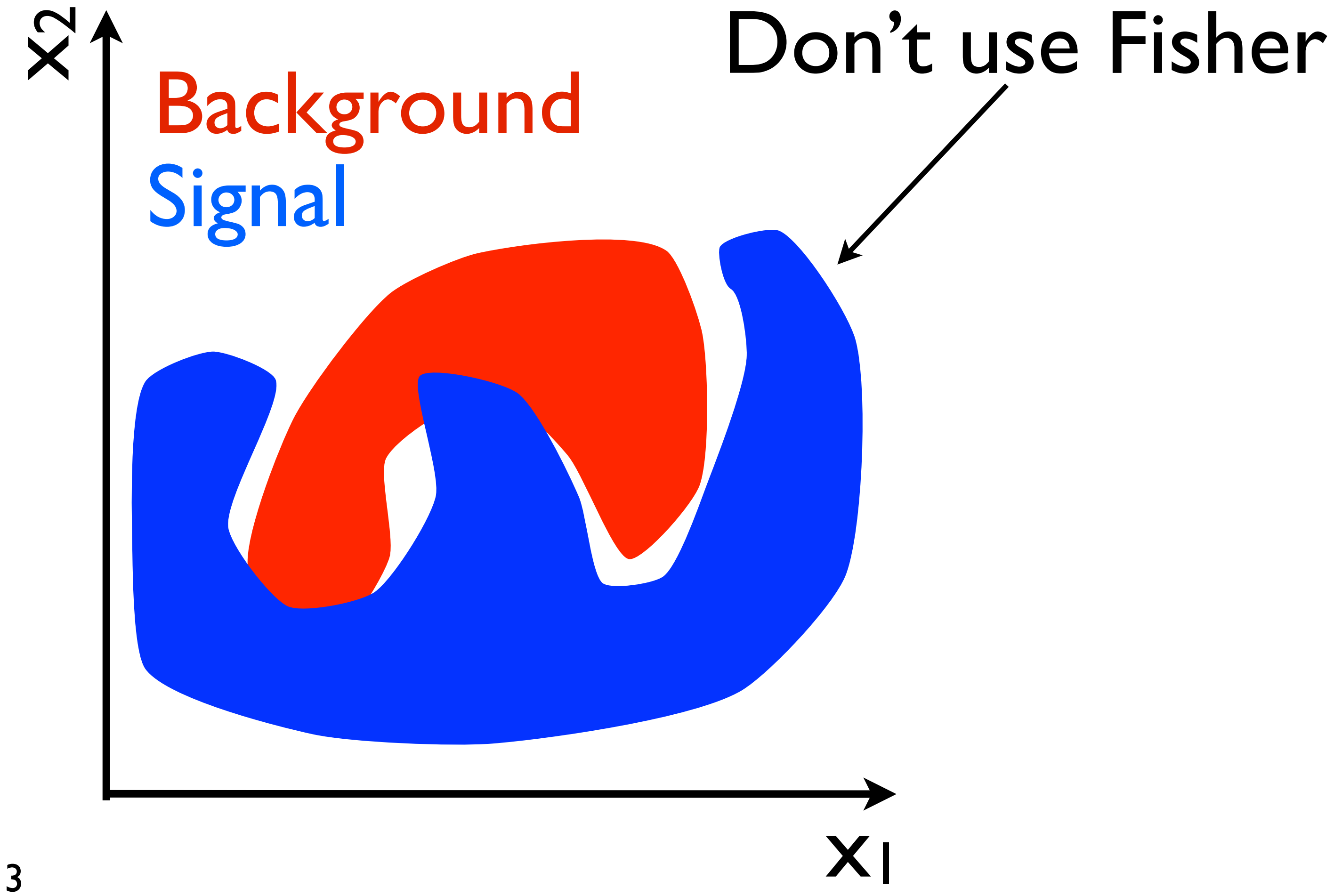
This week

- Monday:
  - Intro to Multivariate analysis - Fisher discriminant/Iris data
- Tuesday:
  - Working on project two
- Friday:
  - A peek into more involved Multivariate techniques / machine learning

Multivariate Analysis

- Monday we saw how to construct the Fisher Discriminant
- Very useful for a lot of applications, but limited in its uses
  - Although you often get the idea in physics, the world is not always linear.





Multivariate Analysis

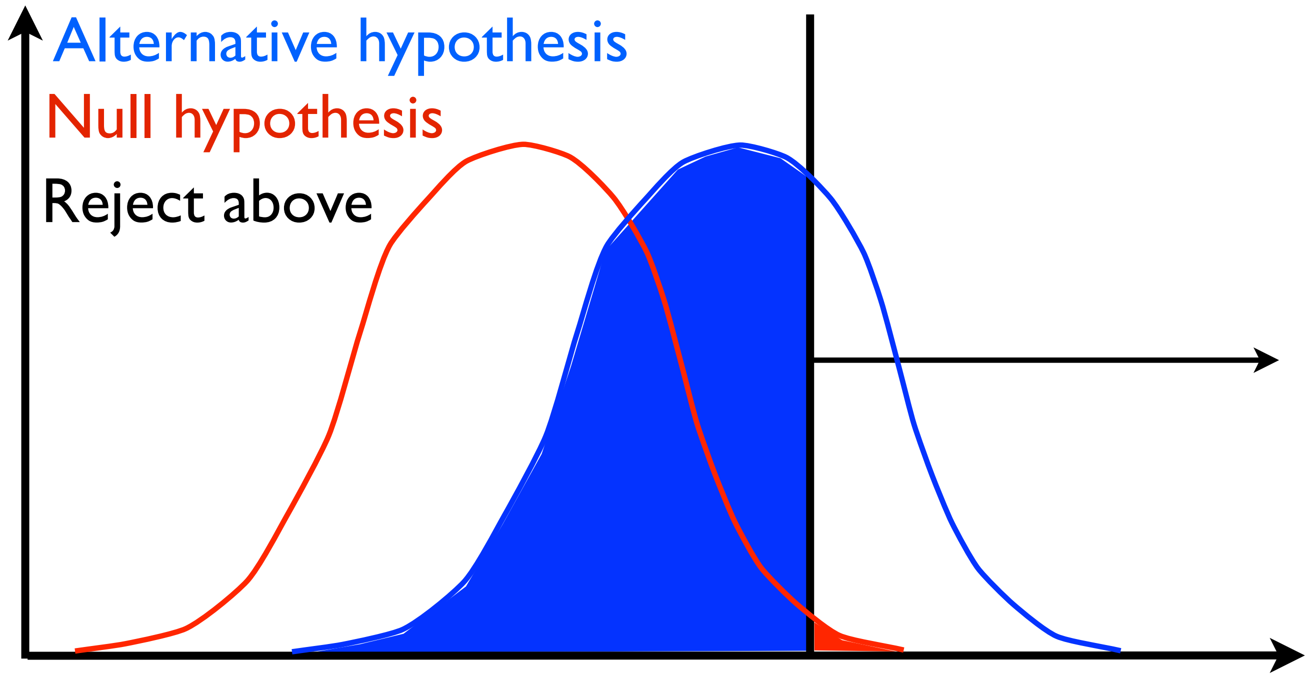
- Today: Introduce a couple of other methods
- Impossible to go into great detail
  - After today: Be able to recognize problems where MVA is applicable
- The Boosted Decision Tree
- Neural Networks
- First introduce a couple of generally useful tools

Separation - ROC*

- Last time: Measured separation with:

$$J(a) = \frac{(\tau_{Null} - \tau_{Alt})^2}{\Sigma^2_{Null} + \Sigma^2_{Alt}}$$

- Easy to understand but also to construct examples where this value is non-sensical
- Idea: Plot Null efficiency vs Alt rejection:
  - I.e. 1-α vs 1-β
- Will contain all information about all possible cuts you can make on distributions

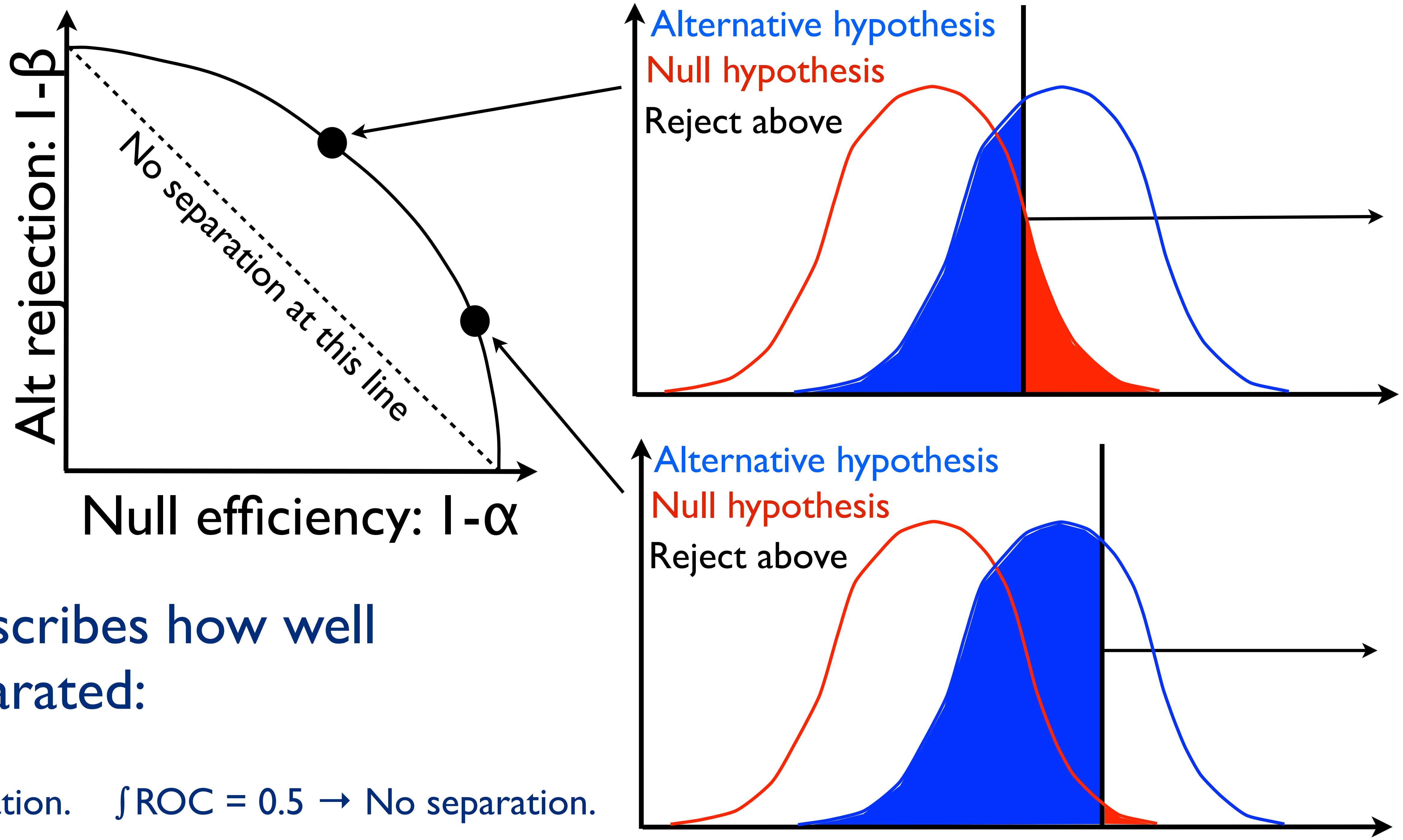


Reality

		Null is	Null is
Statistical decision	Do not reject Null	1-α Correct	β Type II error
	Reject Null	α Type I error	1-β Correct

*Receiver operating characteristic, but nobody calls it that...

Separation - ROC*

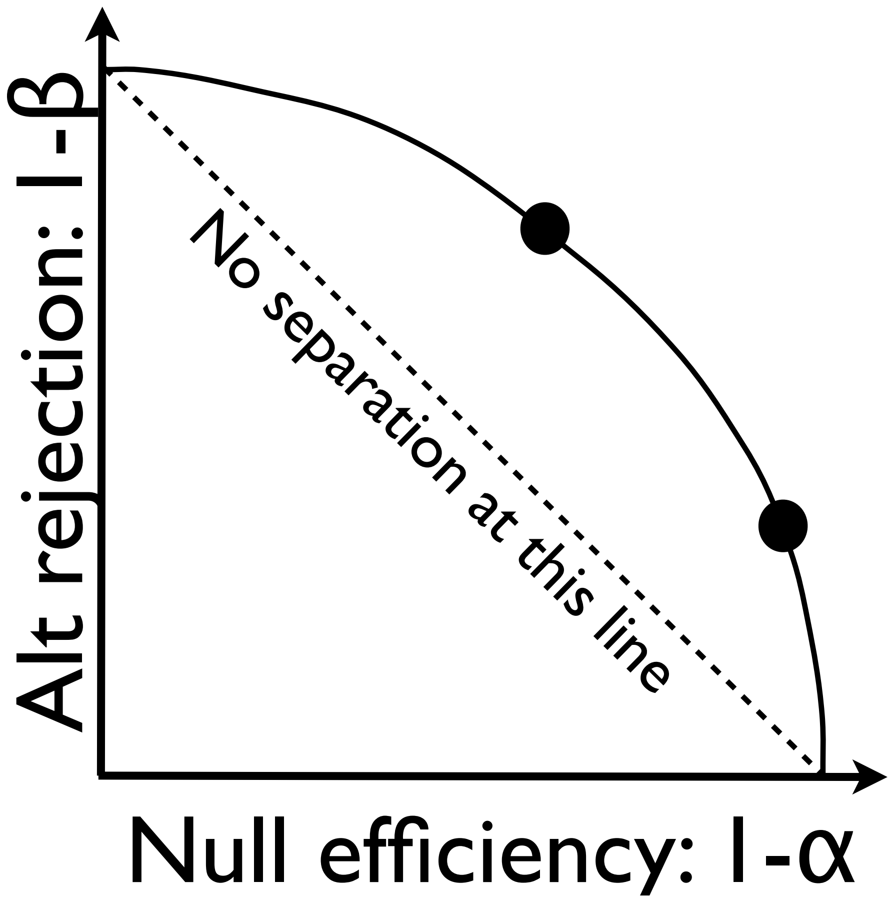


- Integral of curves describes how well distributions are separated:
  - ∫ROC =1.0 → Perfect separation. ∫ROC = 0.5 → No separation.

*Receiver operating characteristic, but nobody calls it that...

Separation - ROC*

- Given "left ↔ right" separation, will always give you a sensible answer.
- Curve tells you directly what a cut on a parameter will do to both Alt and Null distributions.
- Be warned: Here curve is defined as 1-α vs 1-β but there does not exist consensus in literature.
  - Often used: 1-α vs β or α vs 1-β, but information is the same.
  - Personally it makes sense to me that: ∫ROC =1.0 → Perfect separation.
- Use this!



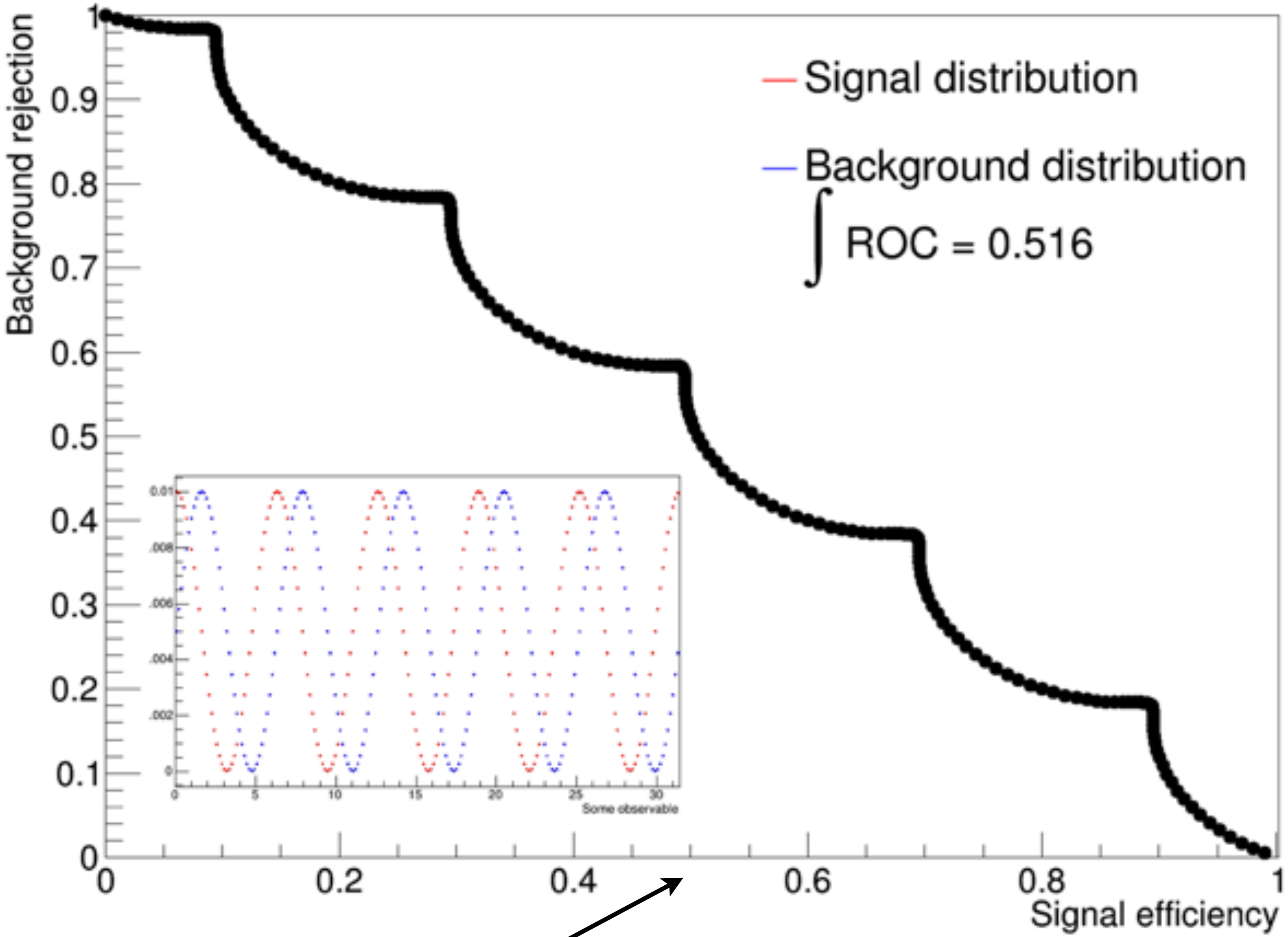
*Receiver operating characteristic, but nobody calls it that...

Separation - What to do if not "left ↔ right" separated

- Probably only case where you should not use ROC curve.

- Often useful to integrate differences in distributions and normalize:

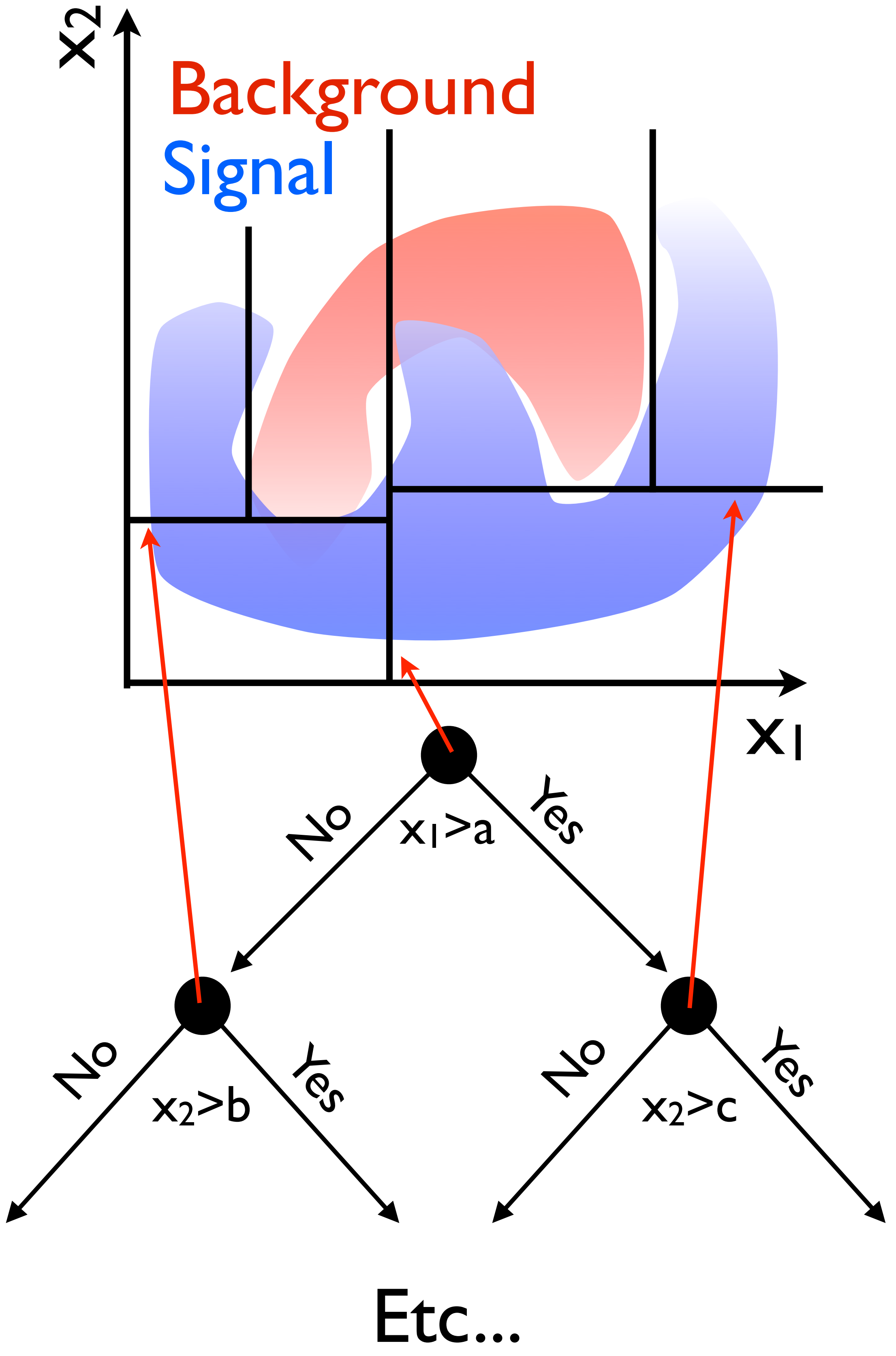
$$\frac{1}{2}\int dx \frac{(f(x|Null) - f(x|Alt))^2}{f(x|Null) + f(x|Alt)}$$



From todays exercise

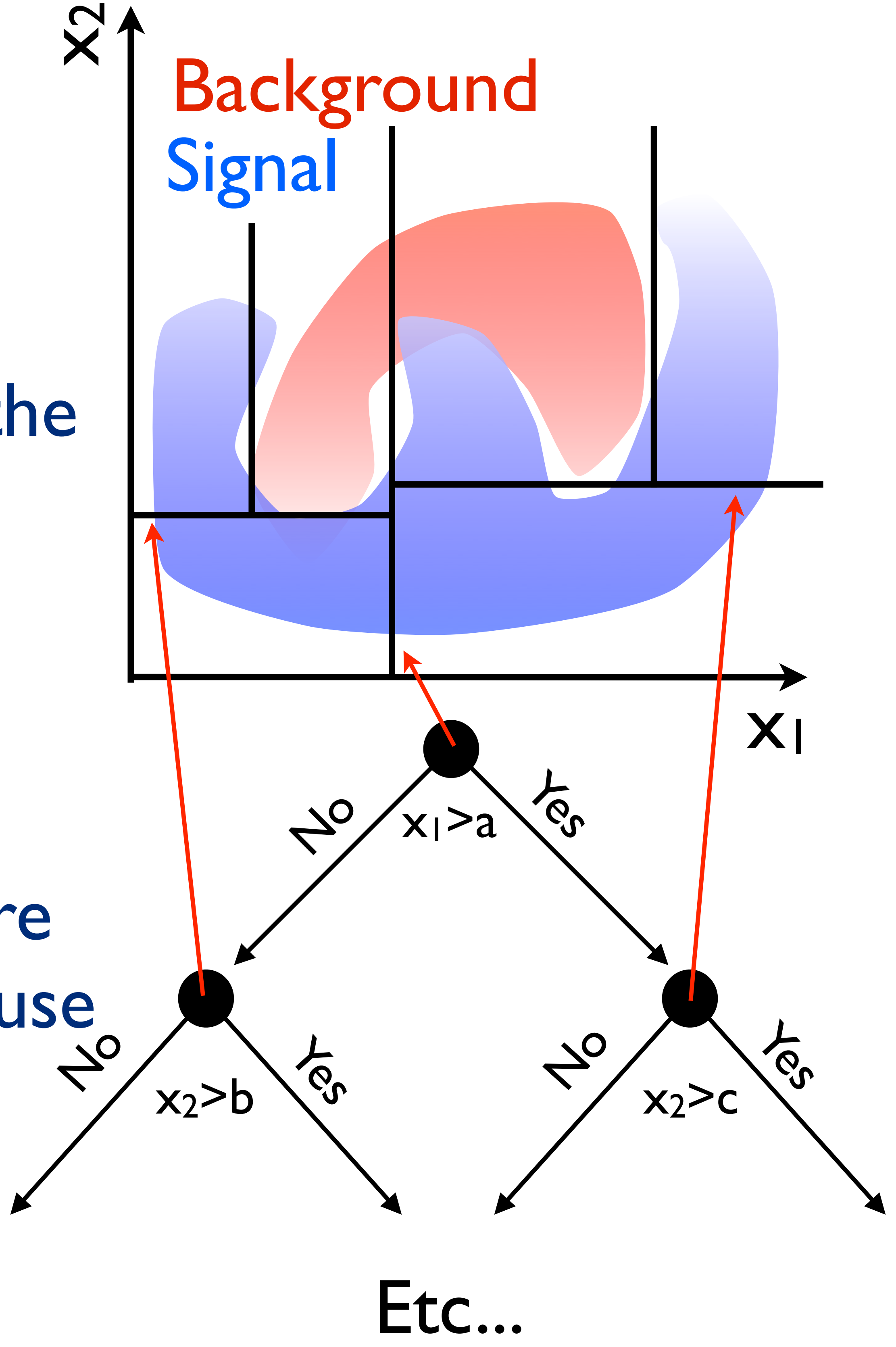
MVA - The Boosted Decision Tree

- Useful in 95% of cases if you have enough statistics
- Fully able to describe non-linear correlations
- Requires high statistics for training
- A decision tree divides the parameter space into squares, each with a probability of being signal or background.
- Create as complex trees as you have statistics for...



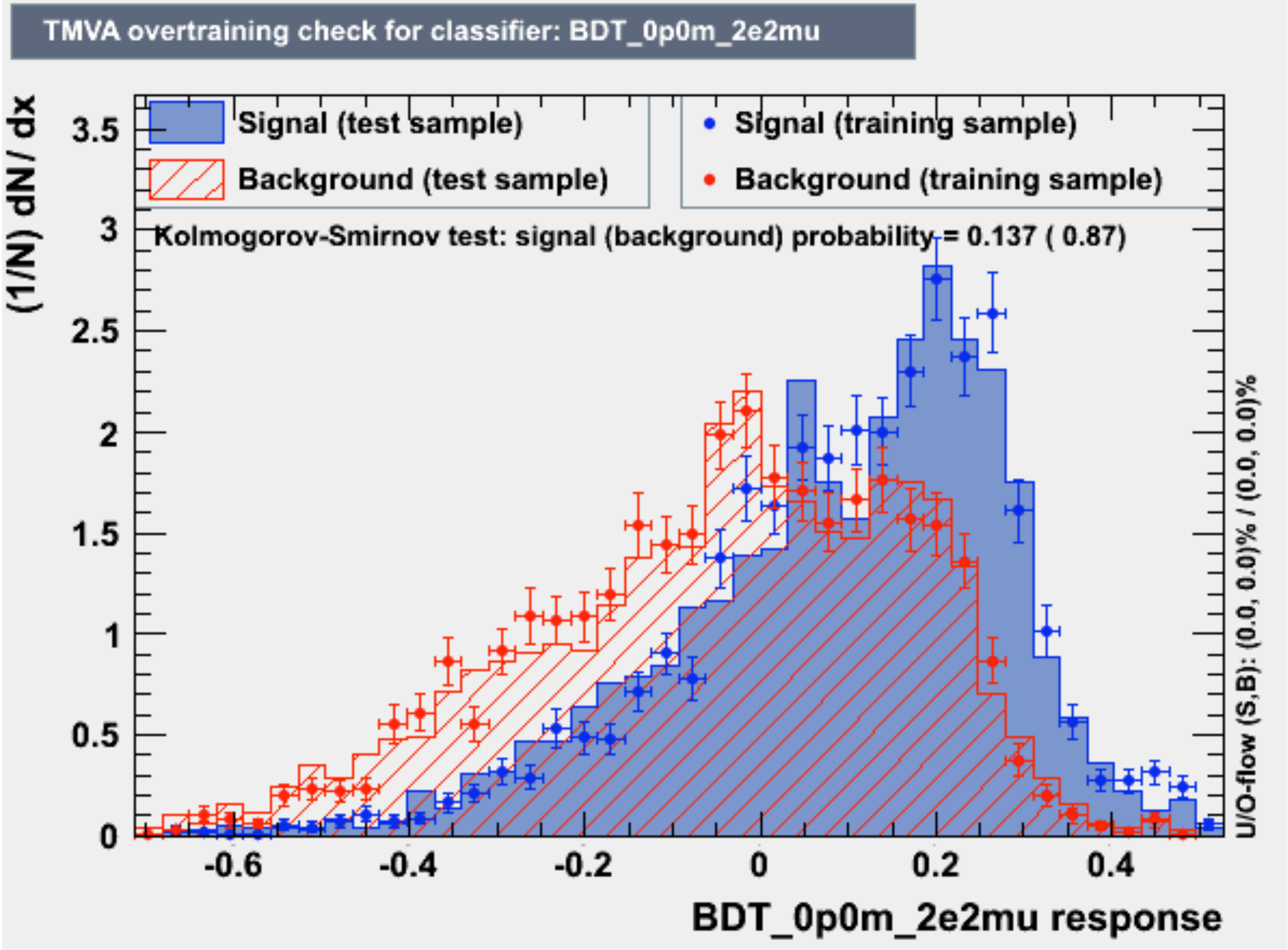
MVA - The Boosted Decision Tree

- Algorithm starts by scanning through one parameter
- At the point of maximal purity the a cut is made and the remaining parameter space is examined
- Could be repeated until each small box only contains signal or background.
- If you found such a region, would you think it was more likely that it was due to a statistical limitation, or because you have 100% pure signal?



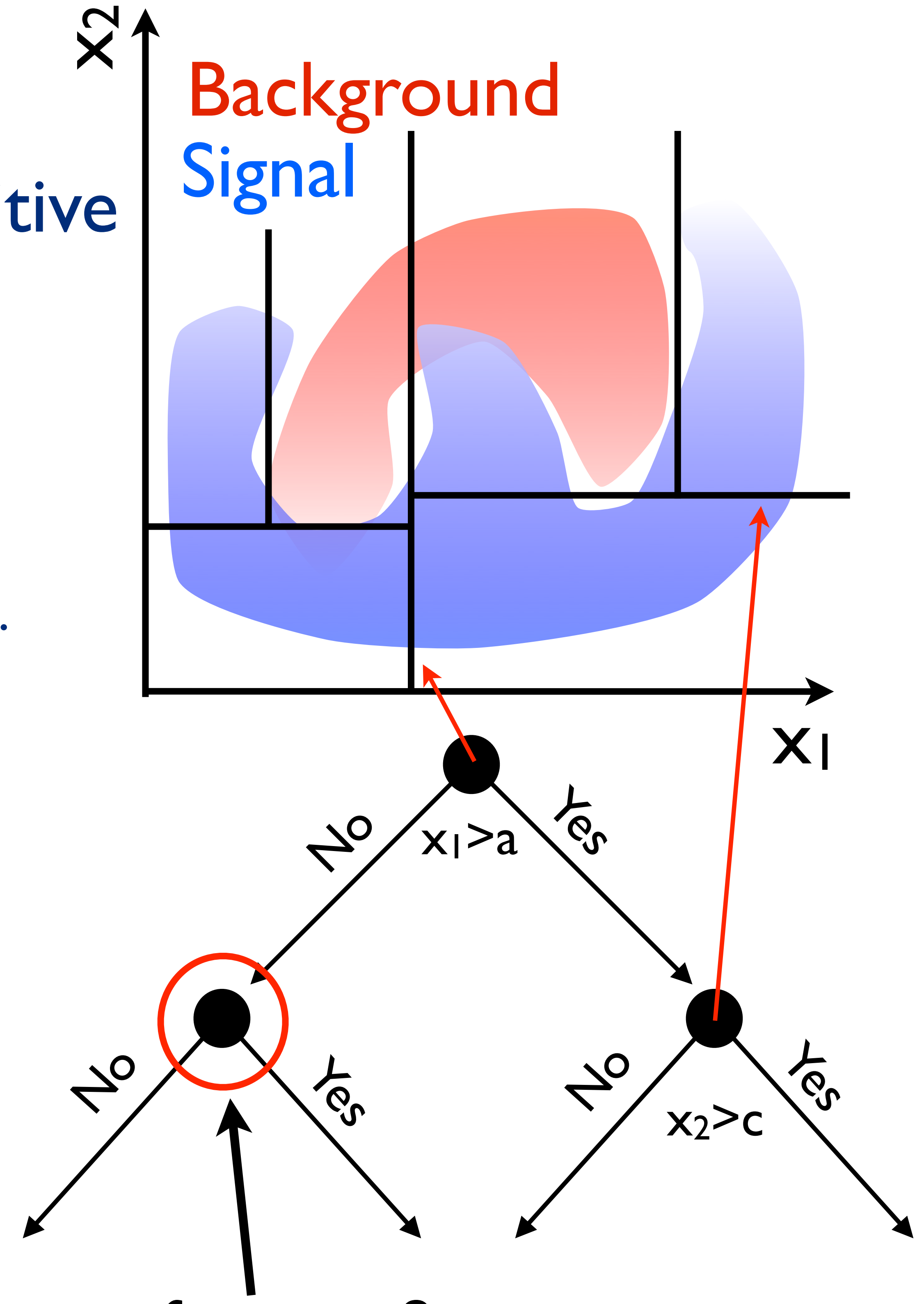
MVA - Overtraining

- General concept within multivariate analysis:
- How do you know if your classifier is picking up on fluctuations?
- Divide your sample into two: A training and a testing sample.
  - For Fischer, calculate means and covariance matrices on training sample. Apply on test sample
- Compare resulting distributions, with e.g. a Kolmogorov Smirnov test:



MVA - The Boosted Decision Tree

- The BDT requires a lot of statistics because it is sensitive to overtraining.
- Overtraining is reduced by ‘pruning’ the tree:
  - If the statistical fluctuations of one of the boxes is too large, remove the node.
  - Is anything lost by removing this node?
- Overtraining can also be reduced by limiting its complexity: Number of nodes and number of trees.
- Remove insignificant parameters!



Remove if insignificant...

MVA - Boosting

- The individual tree in the collection is called a 'weak classifier'.
- No reason why you don't want more trees.
  - Get a more stable result with many simple trees (i.e. few nodes) than a few complex.
  - But, if you train several trees, why don't you just end up with the same one 100 times?
- Boosting: Enhancing performance of weak classifiers.
  - Identify which events are difficult to classify
  - Give difficult event a higher weight and repeat training on next tree

Boost weight

$$\alpha = \frac{1 - \text{err}}{\text{err}}$$

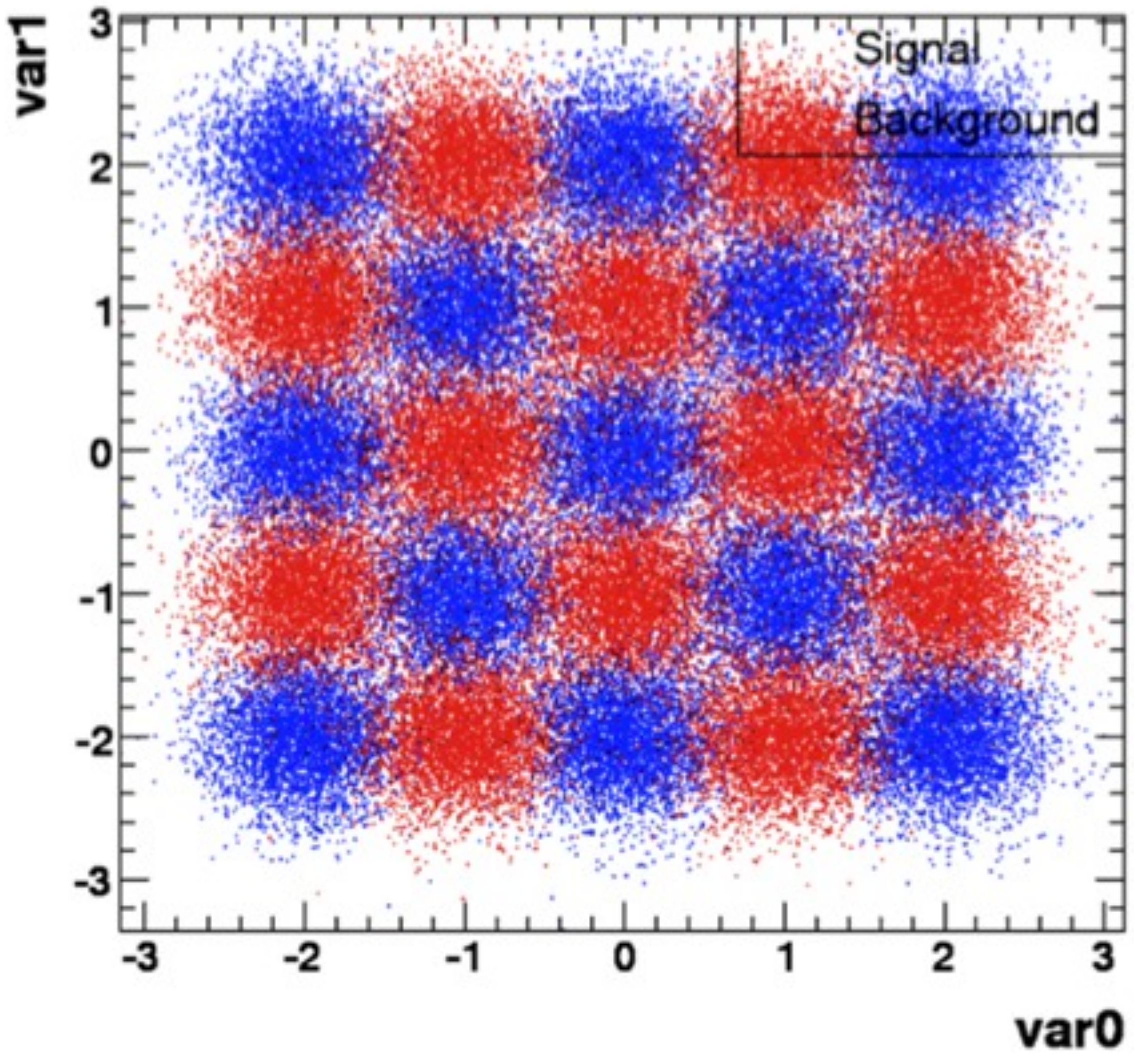
Final classifier

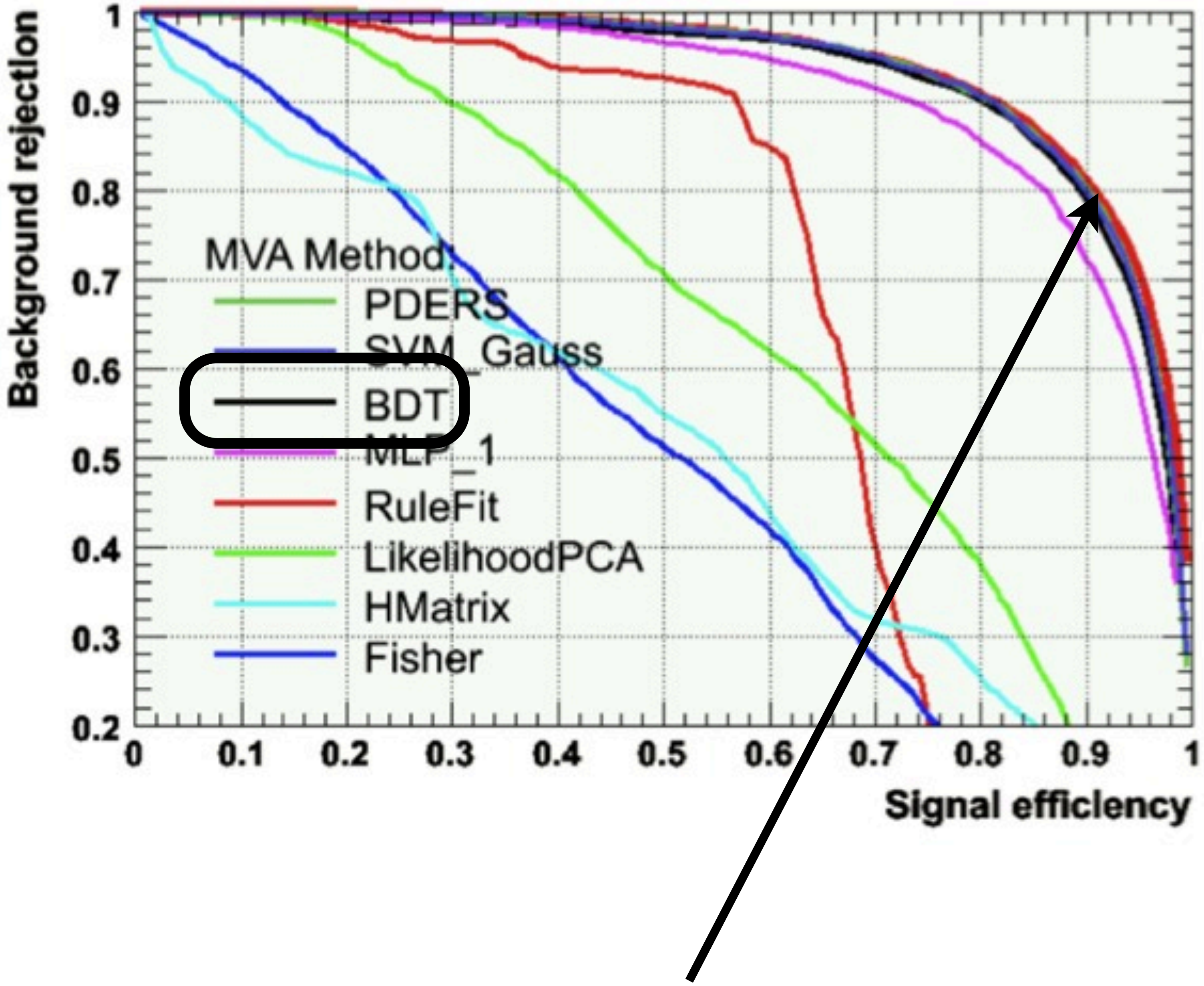
$$y_{\text{Boost}}(\mathbf{x}) = \frac{1}{N_{\text{collection}}} \cdot \sum_{i}^{N_{\text{collection}}} \ln(\alpha_i) \cdot h_i(\mathbf{x})$$

Parameters in event N

Individual tree

MVA - The Boosted Decision Tree





Theoretically
optimal separation

MVA - The Neural Network

- A special case of the Artificial Neural Network: Perceptron.
- Generalize the Fisher test statistics: $t(x) = \sum a_i x_i$

Single layered perceptron

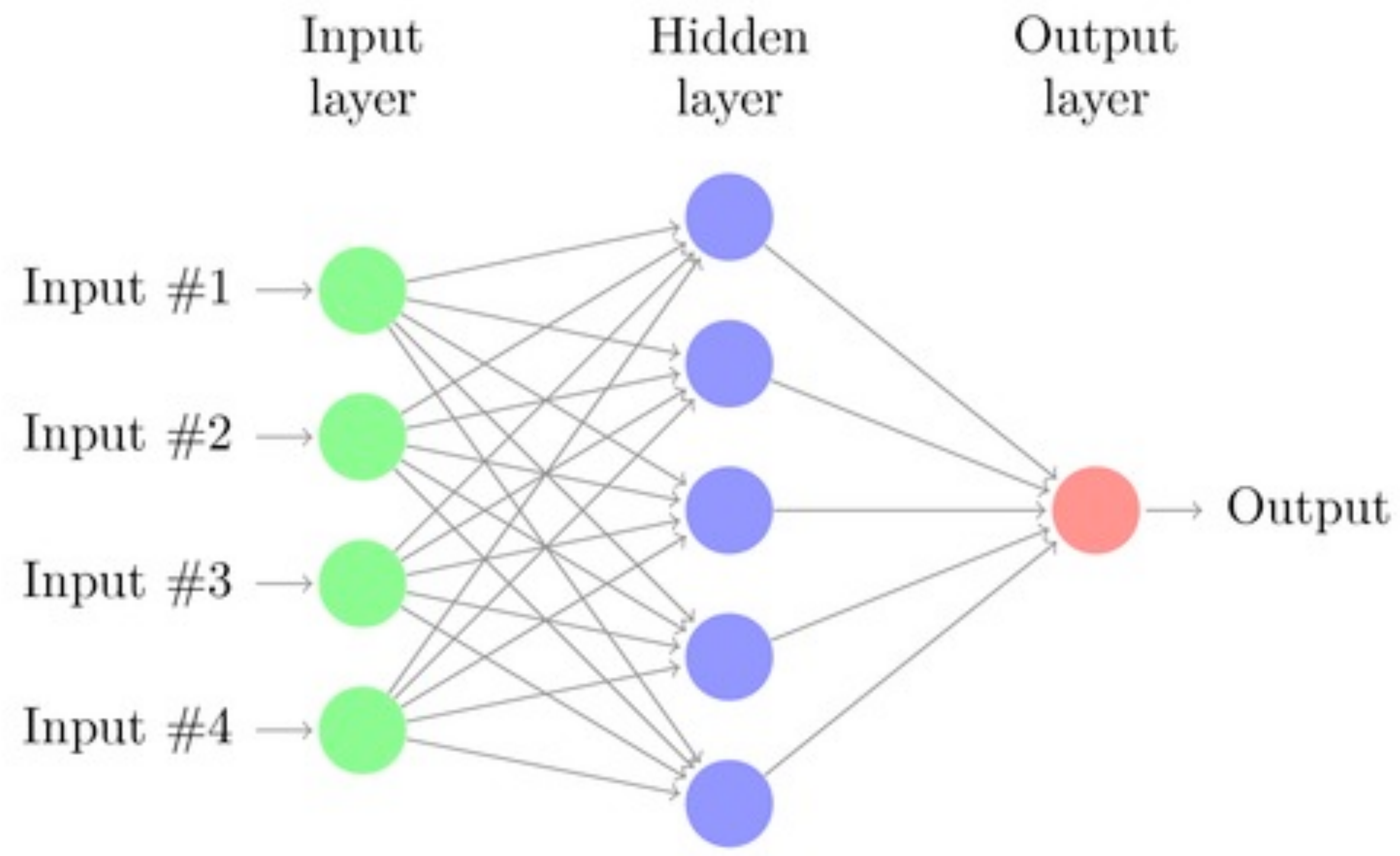
$$t(x) = s\left(a_0 + \sum a_i x_i\right)$$

Generalize to multiple layers

$$t(x) = s\left(a_i + \sum a_i h_i(x)\right)$$

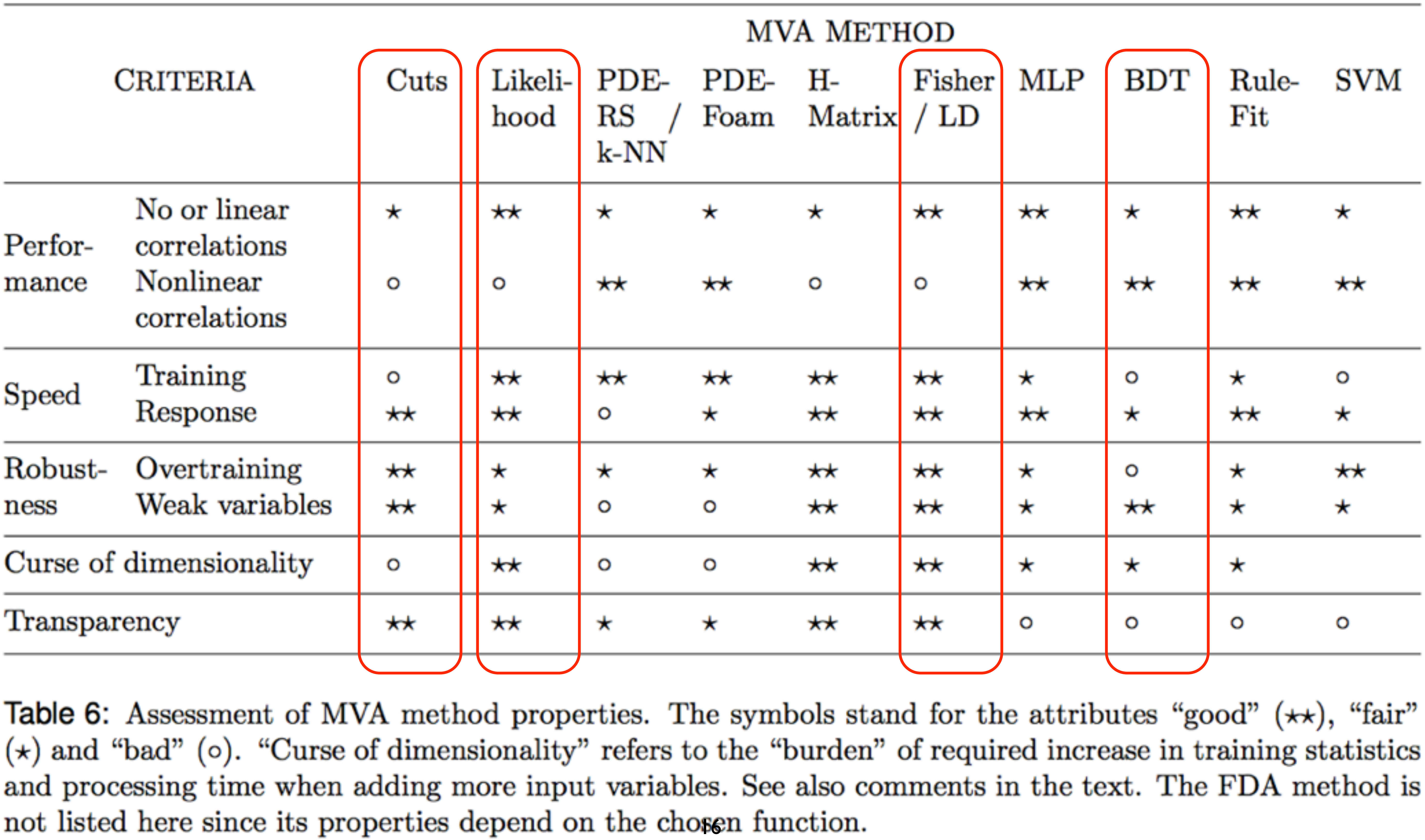
$$h_i(x) = s\left(w_{i0} + \sum w_{ij} x_j\right)$$

Activation function s(...) can be anything monotone. Normally ‘s’ shaped...



Numerical optimization of weights

MVA - Summary

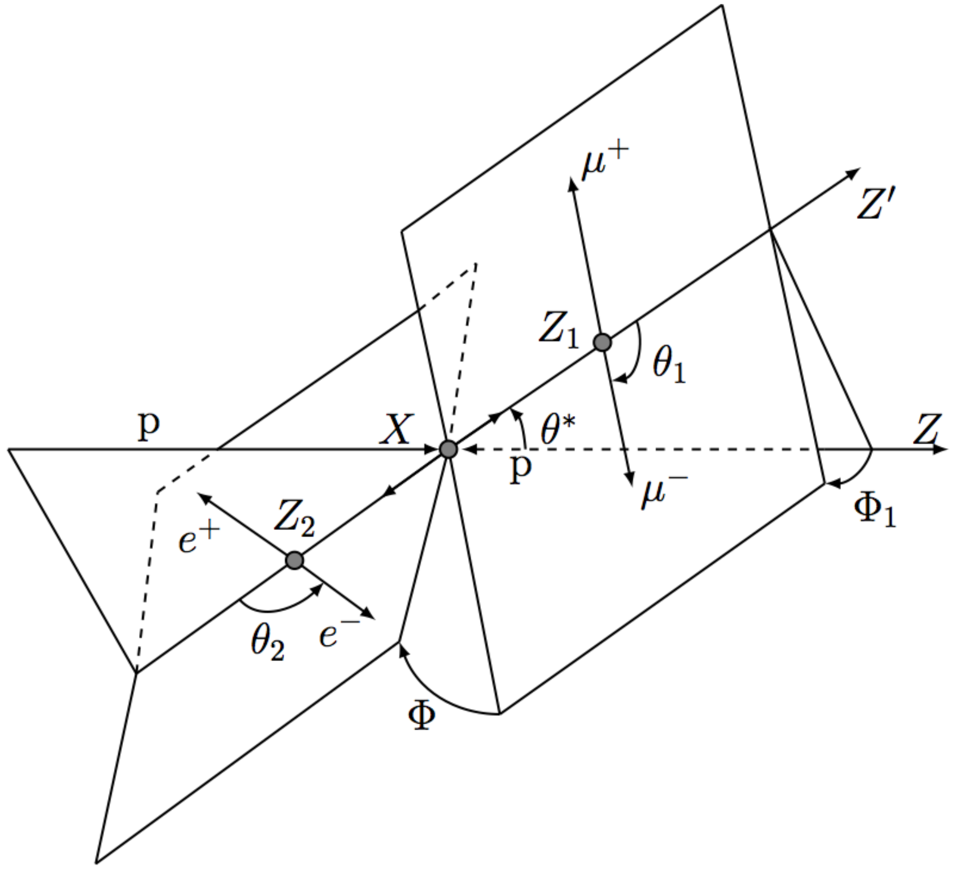


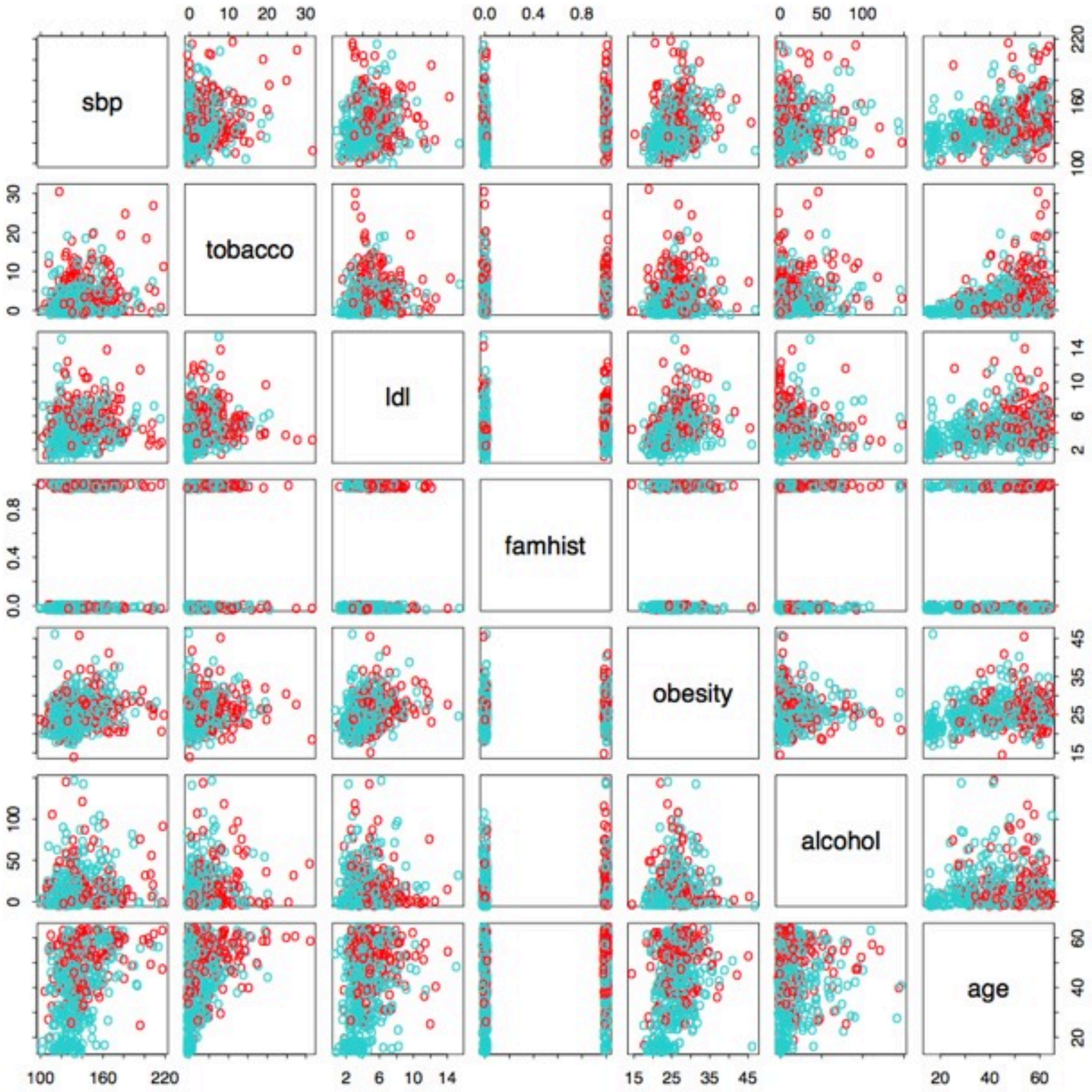
CRITERIA		MVA METHOD									
		Cuts	Likelihood	PDE-RS / k-NN	PDE-Foam	H-Matrix	Fisher / LD	MLP	BDT	Rule-Fit	SVM
Performance	No or linear correlations	⋆	⋆⋆	⋆	⋆	⋆	⋆⋆	⋆⋆	⋆	⋆⋆	⋆
	Nonlinear correlations	∘	∘	⋆⋆	⋆⋆	∘	∘	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Speed	Training	∘	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆	∘	⋆	∘
	Response	⋆⋆	⋆⋆	∘	⋆	⋆⋆	⋆⋆	⋆⋆	⋆	⋆⋆	⋆
Robustness	Overtraining	⋆⋆	⋆	⋆	⋆	⋆⋆	⋆⋆	⋆	∘	⋆	⋆⋆
	Weak variables	⋆⋆	⋆	∘	∘	⋆⋆	⋆⋆	⋆	⋆⋆	⋆	⋆
Curse of dimensionality		∘	⋆⋆	∘	∘	⋆⋆	⋆⋆	⋆	⋆	⋆	
Transparency		⋆⋆	⋆⋆	⋆	⋆	⋆⋆	⋆⋆	∘	∘	∘	∘

Table 6: Assessment of MVA method properties. The symbols stand for the attributes “good” (⋆⋆), “fair” (⋆) and “bad” (∘). “Curse of dimensionality” refers to the “burden” of required increase in training statistics and processing time when adding more input variables. See also comments in the text. The FDA method is not listed here since its properties depend on the chosen function.

Multivariate Analysis

- Todays exercise:
  - Start by calculating the ROC integral between two distributions
  - Play around with the Roots TMVA package:





Some recommended literature

- Glen Cowan: Statistical Data analysis, mentions some of the basics
- http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf
  (yes it is a manual, but it actually contains a lot of valuable explanation)
- http://www-stat.stanford.edu/~tibs/ElemStatLearn/
  (pdf can be found on web page)