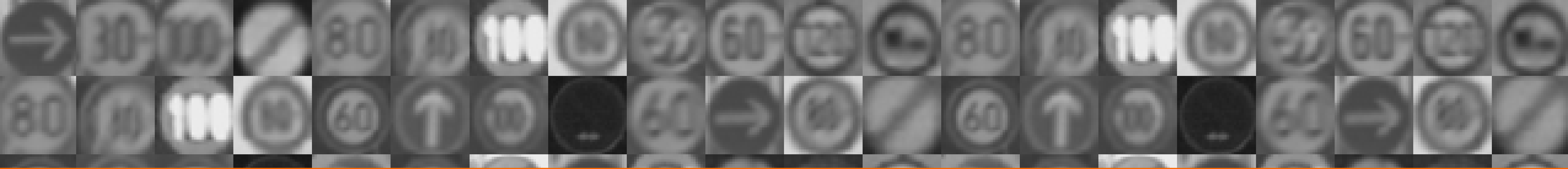
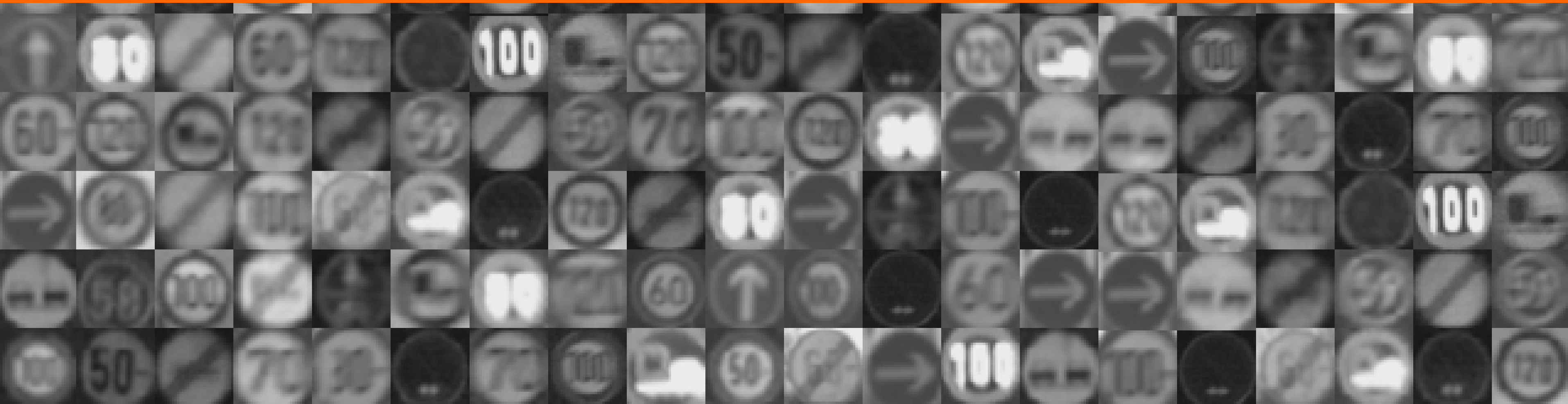


Multivariate Data Analysis and Machine Learning in High Energy Physics

Helge Voss (MPI–K, Heidelberg)

Graduierten-Kolleg , Freiburg, 11.5-15.5, 2009



Outline

- Introduction:
  - the reasons why we need “sophisticated” data analysis algorithms
  - the classification/(regression) problem
  - what is Multivariate Data Analysis and Machine Learning
  - a little bit of statistics
- Classifiers
  - Bayes Optimal Analysis
  - Kernel Methods and Likelihood Estimators
  - Linear Fisher Discriminant
  - Neural Networks
  - Support Vector Machines
  - BoostedDecision Trees
- Will not talk about
  - Unsupervised learning
  - Regression very little only

Literature /Software packages

just a short and biased selection..

Literature:

- T.Hastie, R.Tibshirani, J.Friedman, "*The Elements of Statistical Learning*", Springer 2001
- C.M.Bishop, "*Pattern Recognition and Machine Learning*", Springer 2006

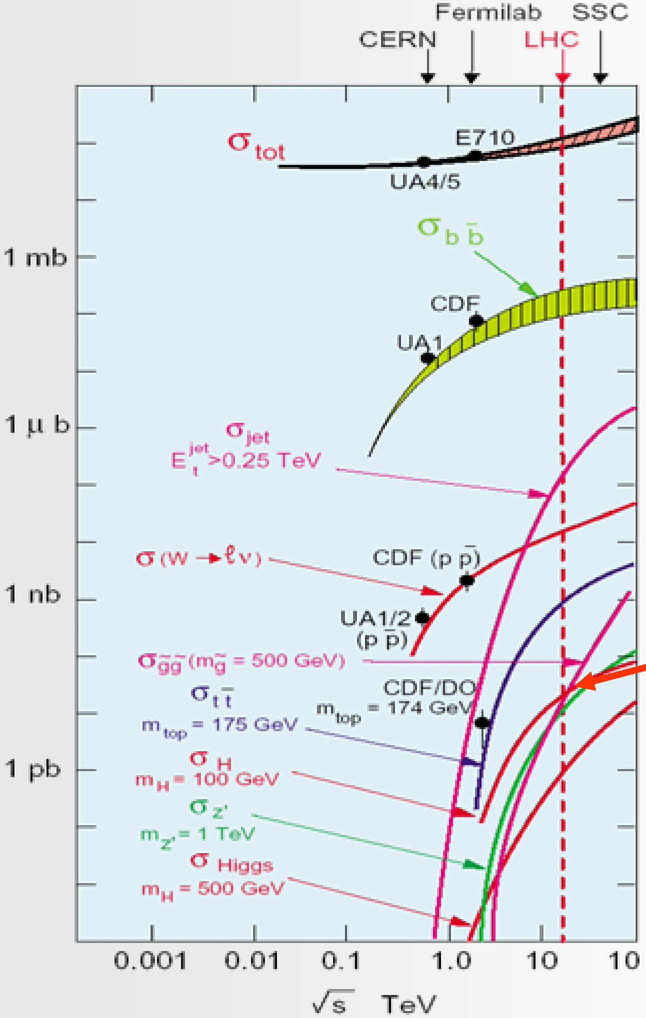
Software packages for Mulitvariate Data Analysis/Classification

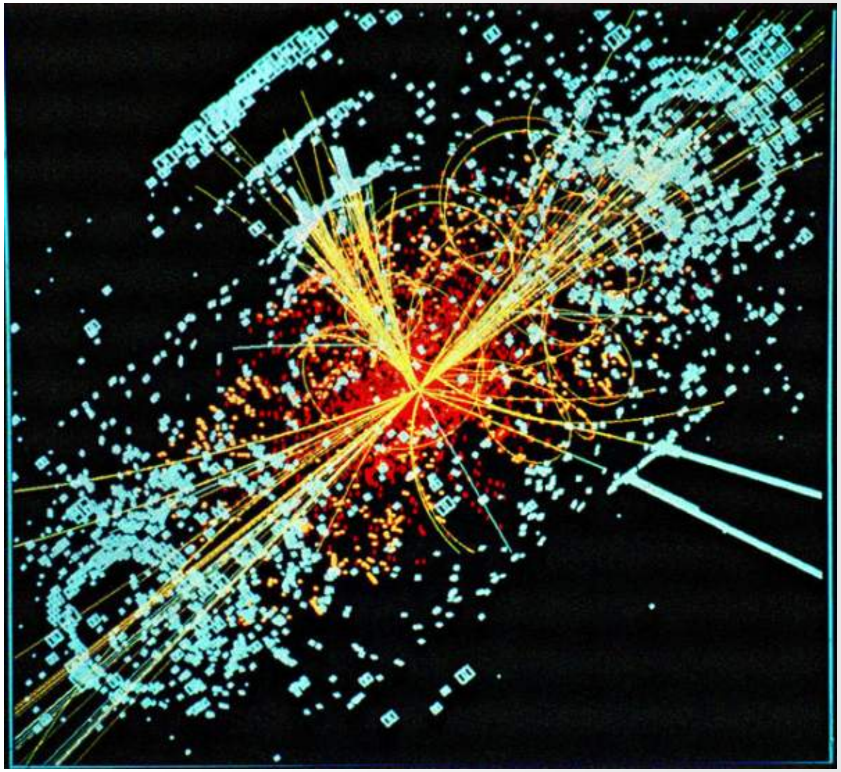
- individual classifier software
  - e.g. "JETNET" C.Peterson, T. Rognvaldsson, L.Loennblad
- attempts to provide "all inclusive" packages
  - StatPatternRecognition: I.Narsky, *arXiv: physics/0507143*
    - *http://www.hep.caltech.edu/~narsky/spr.html*
  - TMVA: Höcker,Speckmayer,Stelzer,Tegenfelt,Voss,Voss, *arXiv: physics/0703039*
    - *http:// tmva.sf.net or every ROOT distribution (not necessarily the latest TMVA version though ☹)*
  - WEKA: http://www.cs.waikato.ac.nz/ml/weka/

Conferences: PHYSTAT, ACAT,…

HEP Experiments: Simulated Higgs Event in CMS

- That's how a "typical" higgs event looks like: (underlying ~23 'minimum bias' events)

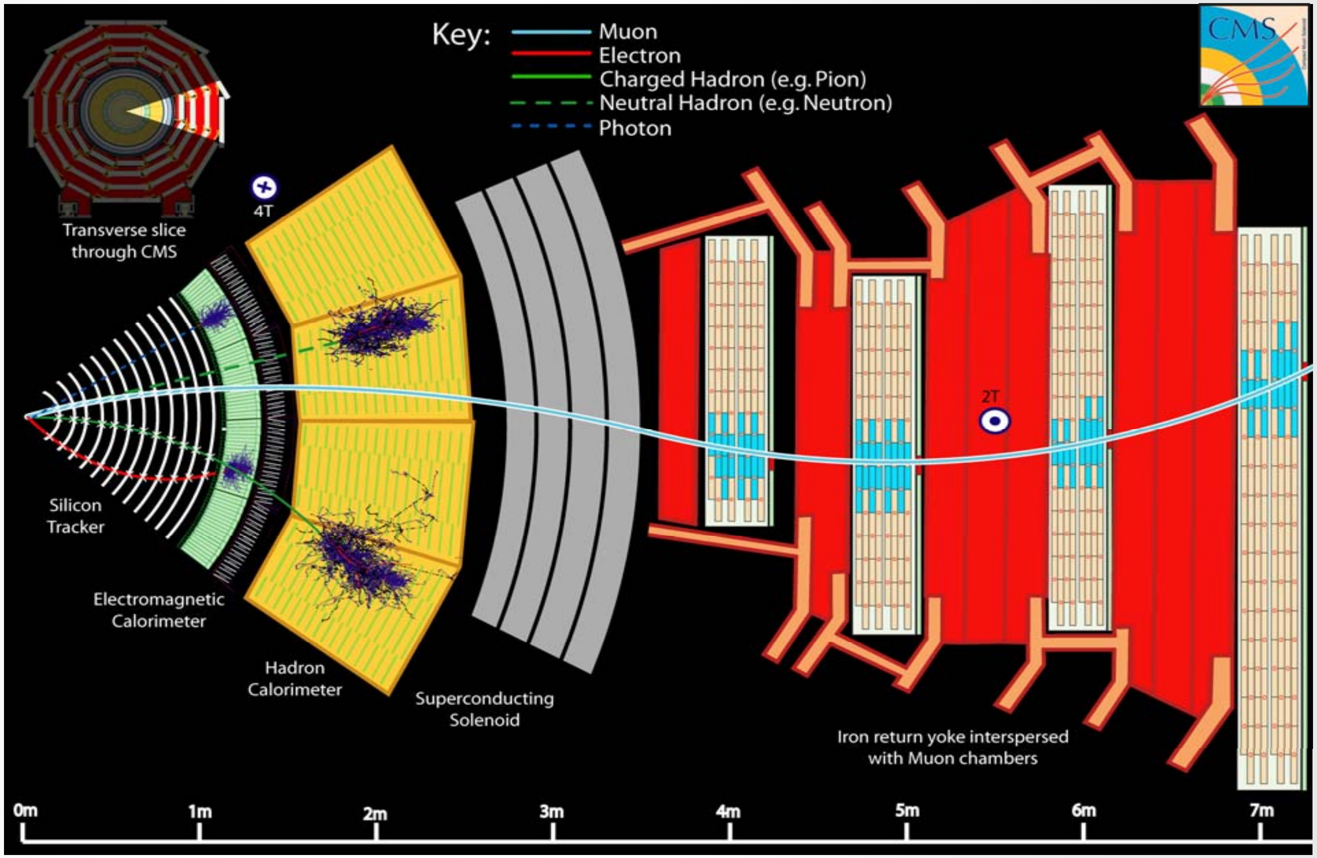




- And not only this: These event happen only in a tiny fraction of the collisions O(10^{-11})

HEP Experiments: Event Signatures in the Detector

- And while a the needle in the hey-stick would be already in one piece:
  - these particles need to be reconstructed from decay products
  - decay products need to be reconstructed from detector signatures
  - etc..



Event Classification in High Energy Phyisics (HEP)

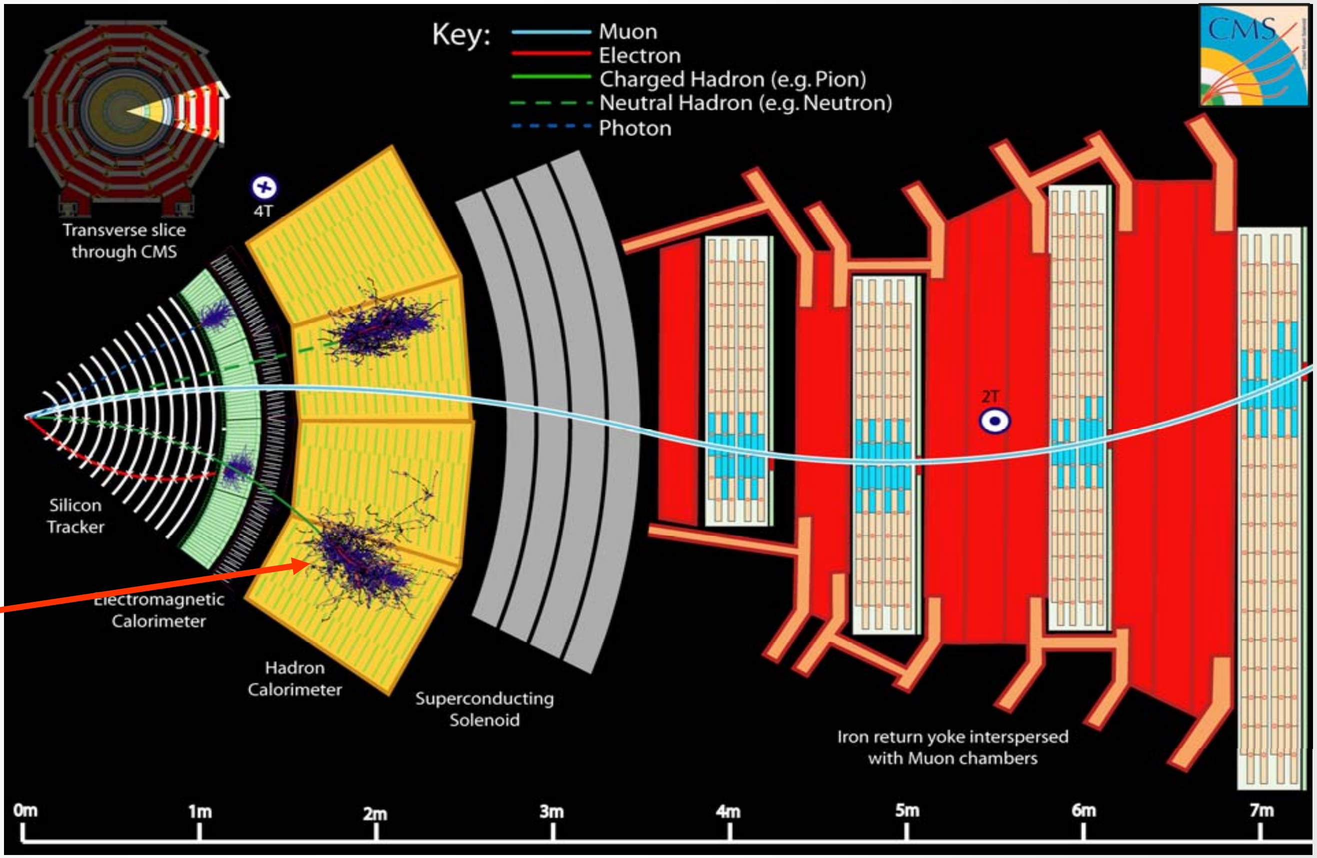
- Most HEP analyses require discrimination of signal from background:
  - event level (Higgs searches, SUSY searches, Top-mass measurement …)
  - cone level (Tau. vs. quark-jet reconstruction, …)
  - track level (particle identification, …)
  - secondary vertex finding (b-tagging)
  - flavour tagging
  - etc.
- Input information from multiple variables from various sources
  - kinematic variables (masses, momenta, decay angles, …)
  - event properties (jet/lepton multiplicity, sum of charges, …)
  - event shape (sphericity, Fox-Wolfram moments, …)
  - detector response (silicon hits, *dE*/*dx*, Cherenkov angle, shower profiles, muon hits, …)
  - etc.
- Traditionally few powerful input variables were combined;

→ new methods allow to use up to 100 and more variables w/o loss of classification power

e.g. MiniBooNE; NIMA 543(2005)577 or D0 single top; Phys.Rev.D78,012005(2008)

HEP Experiments: Event Signatures in the Detector

- And while a the needle in the hey-stick would be already in one piece:
  - these particles need to be reconstructed from decay products
  - decay products need to be reconstructed from detector signatures
  - etc..



- How do I estimate best, the energy this particle had originally?

Regression In High Energy Physics

- Most HEP event reconstruction require some sort of “function estimate”, of which we do not have an analytical expression:
  - energy deposit in a the calorimeter: shower profile → calibration
  - entry location of the particle in the calorimeter
  - distance between two overlapping photons in the calorimeter
  - … you can certainly think of more..

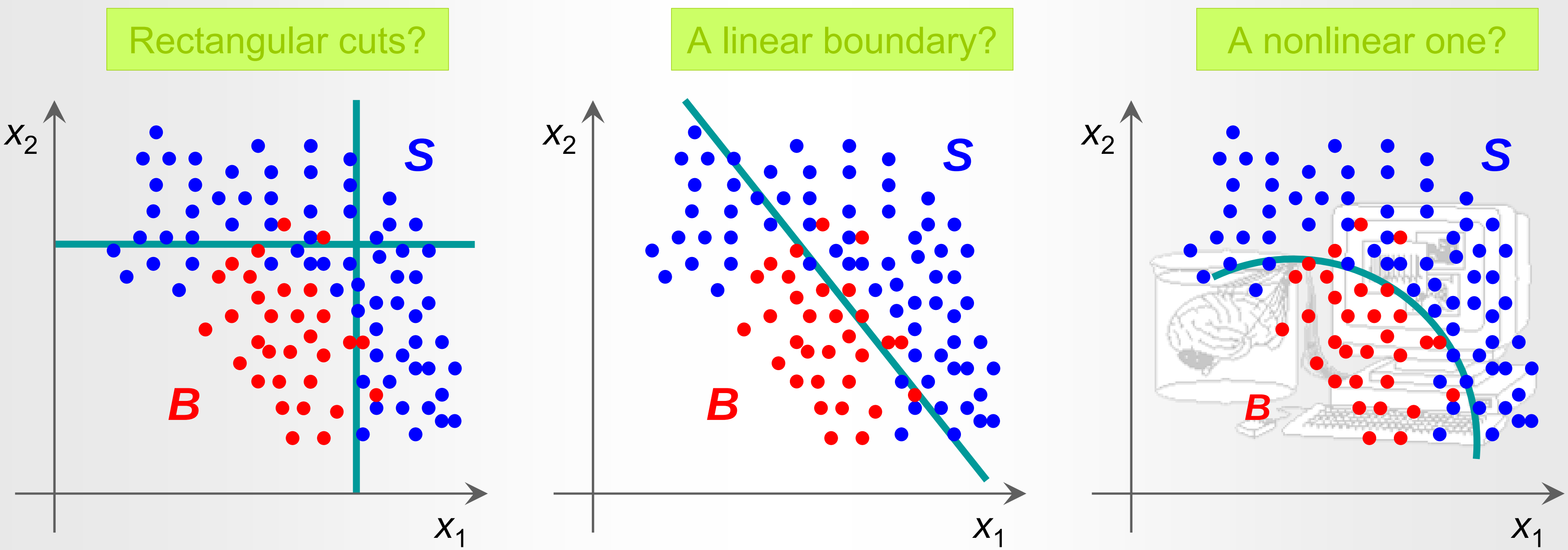
- Maybe you could even imagine some final physics parameter that needs to be fitted to your event data and you would rather use a non analytic fit function from Monte Carlo events than an analytic parametrisation of the theory ???:
  - reweighing method used in the LEP W-mass analysis
  - …

somewhat wild guessing, might not be reasonable to do..

→ Maybe your application can be successfully learned by a machine using Monte Carlo events ??

Event Classification

- Suppose data sample of two types of events: with class labels *Signal* and *Background* (will restrict here to two class cases. Many classifiers can in principle be extended to several classes, otherwise, analyses can be staged)
  - how to set the decision boundary to select events of type *S* ?
  - we have discriminating variables x_1, x_2, …

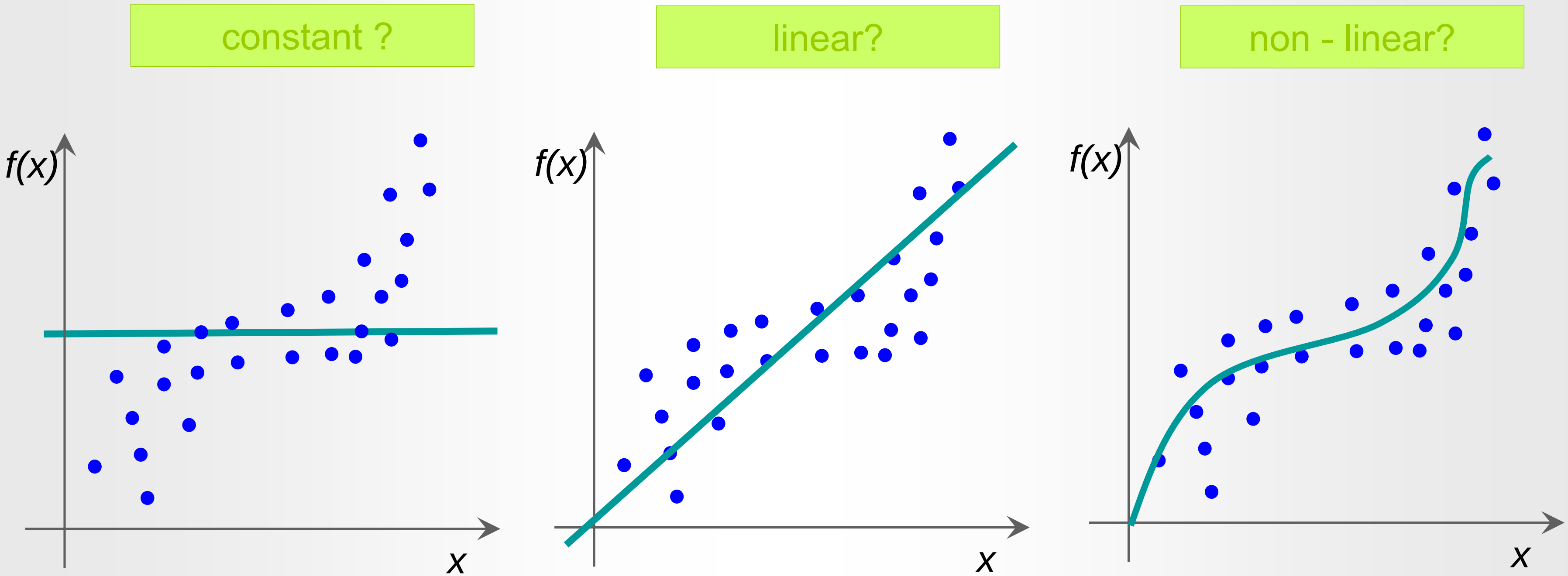


- How can we decide?

- Once decided on a class of boundaries, how to find the “optimal” one?

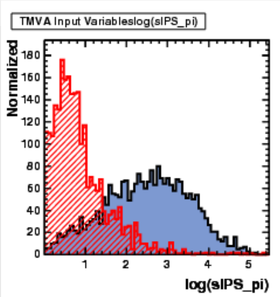
Regression

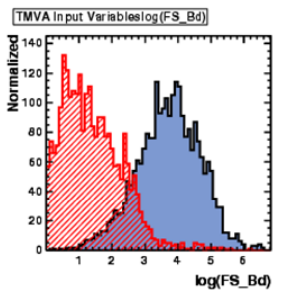
- how to estimate a “functional behaviour” from a given set of ‘known measurements” ?

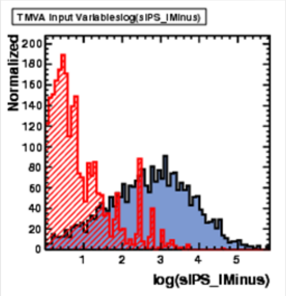


- seems trivial??
- maybe… the human eye and brain behind have very good pattern recognition capabilities!
- but what if you have more variables?

Event Classification







$\mathbb{R}^D$

"feature space"

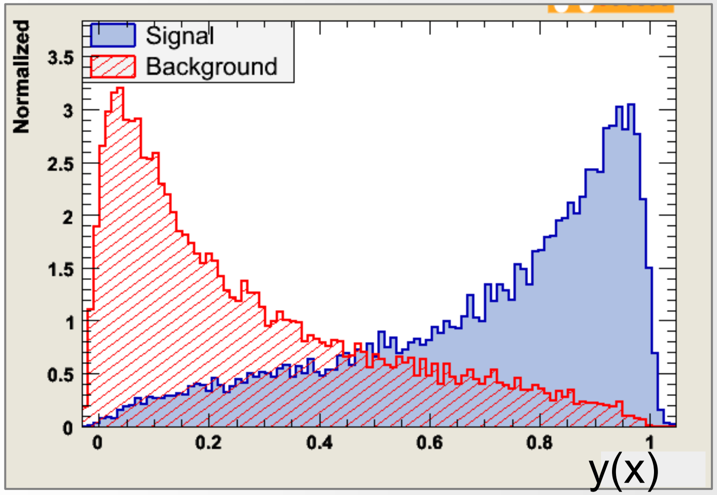
- Each event, if Signal or Background, has "D" measured variables.
- Find a mapping from D-dimensional input/observable/"feature" space to one dimensional output
  → class lables

y(x): $R^n \rightarrow R$:

$\longrightarrow \mathbb{R}$

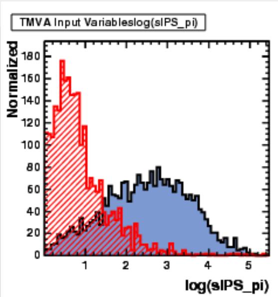
most general form
$y = y(\mathbf{x});\ \mathbf{x} \in \mathbb{R}^D$
$\mathbf{x}=\{x_1,\ldots,x_D\}$: input variables

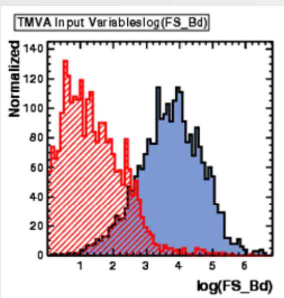
- If one histogramms the resulting y(x) values:

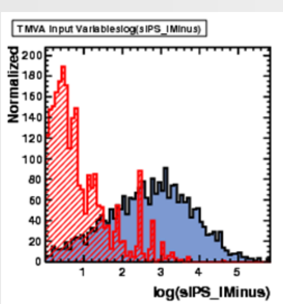


- Who sees how this would look like for the regression problem?

Event Classification







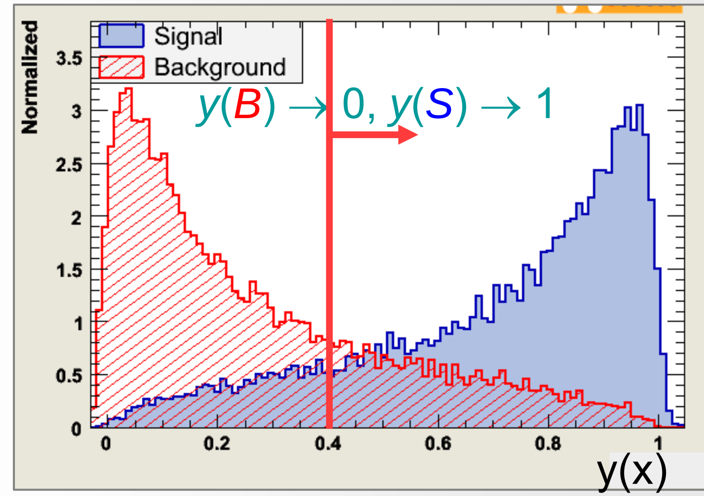
$\mathbb{R}^D$

"feature space"

- Each event, if Signal or Background, has "D" measured variables.
- Find a mapping from D-dimensional input/observable/"feature" space to one dimensional output → class lables

y(x): $R^n \rightarrow R$:

$\longrightarrow \mathbb{R}$



- y(x): "test statistic" in D-dimensional space of input variables
- distributions of y(x): $PDF_S(y)$ and $PDF_B(y)$
- used to set the selection cut!

→ efficiency and purity

y(x): { >cut: signal; =cut: decision boundary; <cut: background }

- y(x)=const: surface defining the decision boundary.
- overlap of $PDF_S(y)$ and $PDF_B(y)$ → separation power , purity

MVA and Machine Learning

- The previous slide was basically the idea of "Multivariate Analysis" (MVA)
  - rem: What about "standard cuts" ?

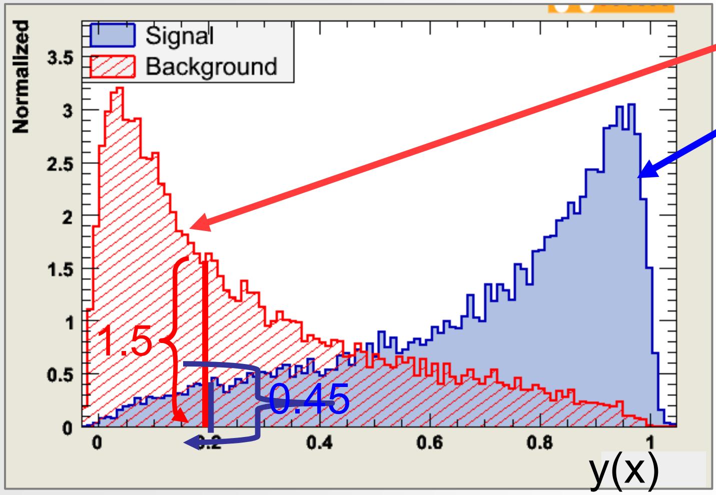
- Finding $y(x) : R^n \rightarrow R$
  - given a certain type of model class y(x)
  - in an automatic way using "known" or "previously solved" events
    - i.e. learn from known "patterns"
  - such that the resulting y(x) has good generalization properties when applied to "unknown" events

→ that is what the "machine" is supposed to be doing: supervised machine learning

- Of course… there's no magic, we still need to:
  - choose the discriminating variables
  - choose the class of models (linear, non-linear, flexible or less flexible)
  - tune the "learning parameters" → bias vs. variance trade off
  - check generalization properties
  - consider trade off between statistical and systematic uncertainties

Event Classification

y(x): $R^n \rightarrow R$: the mapping from the "feature space" (observables) to one output variable



$PDF_B(y)$. $PDF_S(y)$: are the normalised distribution of y=y(x) for background and signal events (i.e. the "function" that describes the shape of the distribution)

with y=y(x) one can also say $PDF_B(y(x))$, $PDF_S(y(x))$: :

Probability densities for background and signal

now let's assume we have an unknown event from the example above for which y(x) = 0.2

$\rightarrow PDF_B(y(x)) = 1.5$ and $PDF_S(y(x)) = 0.45$

let f_S and f_B be the fraction of signal and background events in the sample, then:

$$\frac{f_S PDF_S(y)}{f_S PDF_S(y) + f_B PDF_B(y)} = P(C = S \mid y)$$

is the probability of an event with measured **x**=$\{x_1,\ldots,x_D\}$ that gives y(x) to be of type signal

Quick digression: What is Probability

A measure of how likely it is that some event will occur; a number expressing the ratio of favorable cases to the whole number of cases (wordnet.princeton.edu/perl/webwn)

- Frequentist probability:
  - $$P(E_{\text{vent}}) = \lim_{n \to \infty} \frac{\#\text{outcome is } E_{\text{vent}}}{\text{n} - \text{"trials"}}$$
- Bayesian probability:
  - $P(E_{\text{vent}}) = \text{degree of belief that } E_{\text{vent}} \text{ is going to happen}$
  - Fraction of possible worlds in which E_{vent} is going to happen….

People like to spend hours on philosophy over these statements… I don't

- Axioms of probability: Kolmogorov (1933)
  - $P(A) \geq 0$
  - $\int P(A)dA = 1$
  - if $A \cap B = 0$ (i.e disjoint events) then $P(A \cup B) = P(A) + P(B)$

→ given those we can define: **conditional probability:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Quick digression: Probability

- conditional probability: $P(\mathrm{A}|\mathrm{B}) = \dfrac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{B})}$

$\rightarrow P(\mathrm{A}|\mathrm{B}) * P(\mathrm{B}) = P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{B} \cap \mathrm{A}) = P(\mathrm{B}|\mathrm{A}) * P(\mathrm{A})$

$\rightarrow$ Bayes Theorem: $P(\mathrm{A}|\mathrm{B}) = \dfrac{P(\mathrm{B}|\mathrm{A}) * P(\mathrm{A})}{P(\mathrm{B})}$

B.t.w. Nobody argues about the validity of “Bayes Theorem”.
Discussions start only if one uses it to:

turn frequentist like statements about the

“Probability of the observed data given a certain model” $P(\text{data} \mid \text{model})$

into something that reads like

“Probability of a certain model being correct” $P(\text{model} \mid \text{data})$

Event Classification

$P(Class=C|\mathbf{x})$ (or simply $P(C|x)$) : probability that the event class is of C, given the measured observables $\mathbf{x}=\{x_1,\ldots,x_D\} \rightarrow y(x)$

probability density distribution according to the measurements x and the given mapping function

prior probability to observe an event of "class C" i.e. relative abundance of "signal" versus "background"

$$P(Class = C\,|\,y) = \frac{P(y\,|\,C)\cdot P(C)}{P(y)}$$

posterior probability

overall probability density to observe the actual measurement y(x). i.e. $P(y) = \sum_{Classes} P(y\,|\,Class)P(Class)$

Bayes Optimal Classification

$$P(Class = C\,|\,y) = \frac{P(y\,|\,C)P(C)}{P(y)}$$

$\mathbf{x}=\{x_1,\ldots,x_D\}$: measured observables

$y = y(x)$

+

minimum error in misclassification if C chosen such that is has maximum P(C|**y**)

i.e. to select S(ignal) over B(ackground), place decision on

$$\frac{P(S\,|\,y)}{P(B\,|\,y)} = \frac{P(y\,|\,S)}{P(y\,|\,B)} \cdot \frac{P(S)}{P(B)} > c$$

the constant "c" determines efficiency and purity

or any monotonic function of P(S|y)/P(B|y)

Likelihood ratio
as discriminating function y(x)

prior odds of choosing a signal event
(relative probability of signal vs. bkg)

$$\boxed{\text{odds} = \frac{p}{1-p}}$$

Any decision involves a certain risk ….

→ decide to treat an event as "Signal" or "Background"

- Type 1 error:
  classify event as Class C even though it is not
  (accept a hypothesis although it is not tue/false)
  (reject the null-hypothesis although it would have been the correct one)
  → loss of purity (in the selection of signal events)

- Type 2 error:
  fail to identify an event from Class C as such
  (reject a hypothesis although it would have been correct/true)
  (fail to reject the null-hypothesis/accept null hypothesis although it is false)
  → loss of efficiency (in selecting signal events)

Trying to select signal events:
(i.e. try to disprove the null-hypothesis stating it were "only" a background event)

truly is: \ accept as:	signal	back-ground
signal	☺	Type II error
back-ground	Type I error	☺

"A": region of the outcome of the test where you accept the event as signal:

- significance α: rate at which you make a Type I error:
  (= p-value): 1- α : background selection "efficiency"

$$\alpha = \int_{\mathrm{All}-A} P(x\,|\,S)dx$$

should be small

- size β: rate at which you make a Type II error:

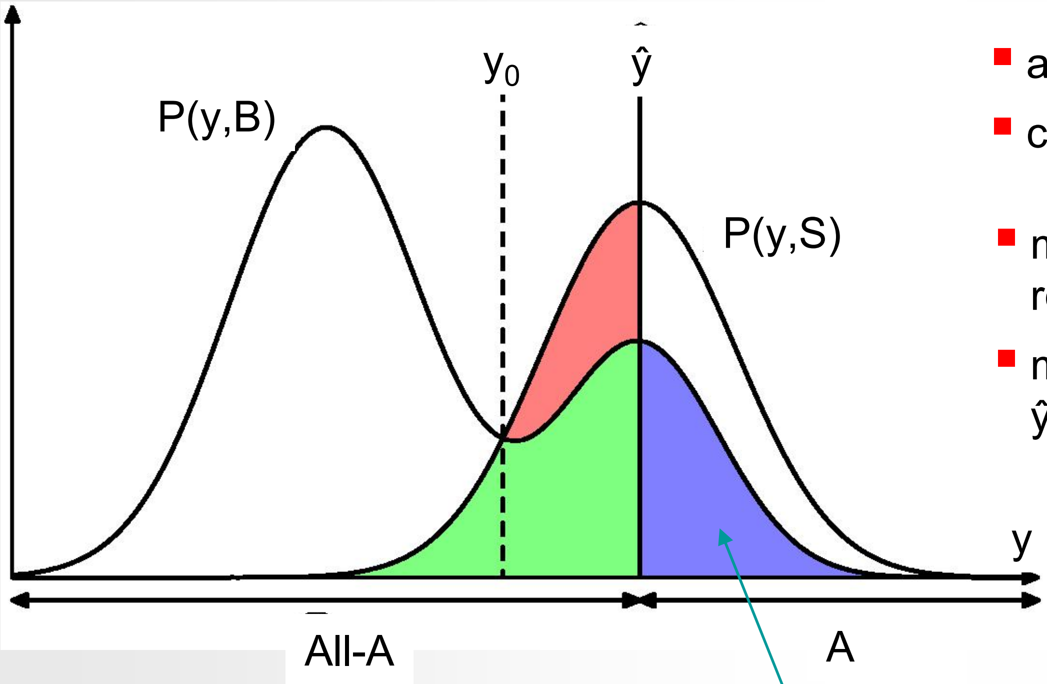
power 1-β: = selection efficiency

$$\beta = \int_{A} P(x\,|\,B)dx$$

should be small

most of the rest of the lecture will be about methods that make as little mistakes as possible ☺

Any decision involves a certain risk ….



Type 1 error for signal selection

- accept as Signal every thing above ŷ
- call everything Background below ŷ

- misclassification occurs in the colored regions:
- minimum number of misclassifications if $\hat{y}=y_0$

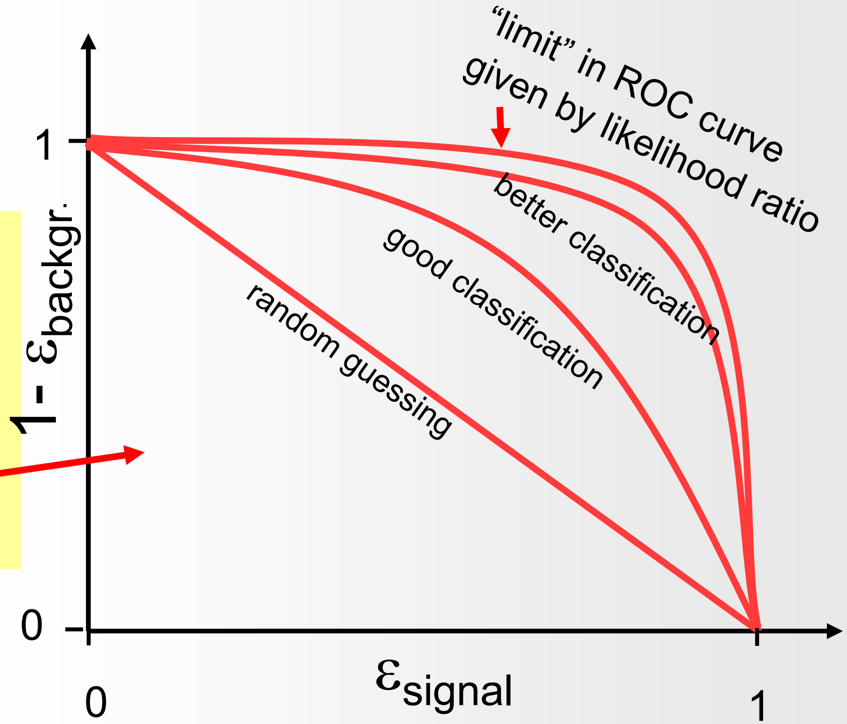
- the longer I look at this picture… P(y|B), P(y|S) are not normalized …. bizarre…

But yes if they would rather represent the “posteriori probabilities” P(B|y), P(S|y) then….

Neyman-Pearson Lemma

$$\text{Likelihood Ratio}: \quad y(x) = \frac{P(x|S)}{P(x|B)}$$

Neyman-Peason:
The Likelihood ratio used as "selection criterium" y(x) gives for each selection efficiency the best possible background rejection.
i.e. it maximises the area under the "Receiver Operation Characteristics" (ROC) curve



- y(x) is the discriminating function given by your estimator (i.e. the likelihood ratio)
- varying y(x)>"cut" moves the working point (efficiency and purity) along the ROC curve
- where to choose your working point? → need to know prior probabilities (abundances)

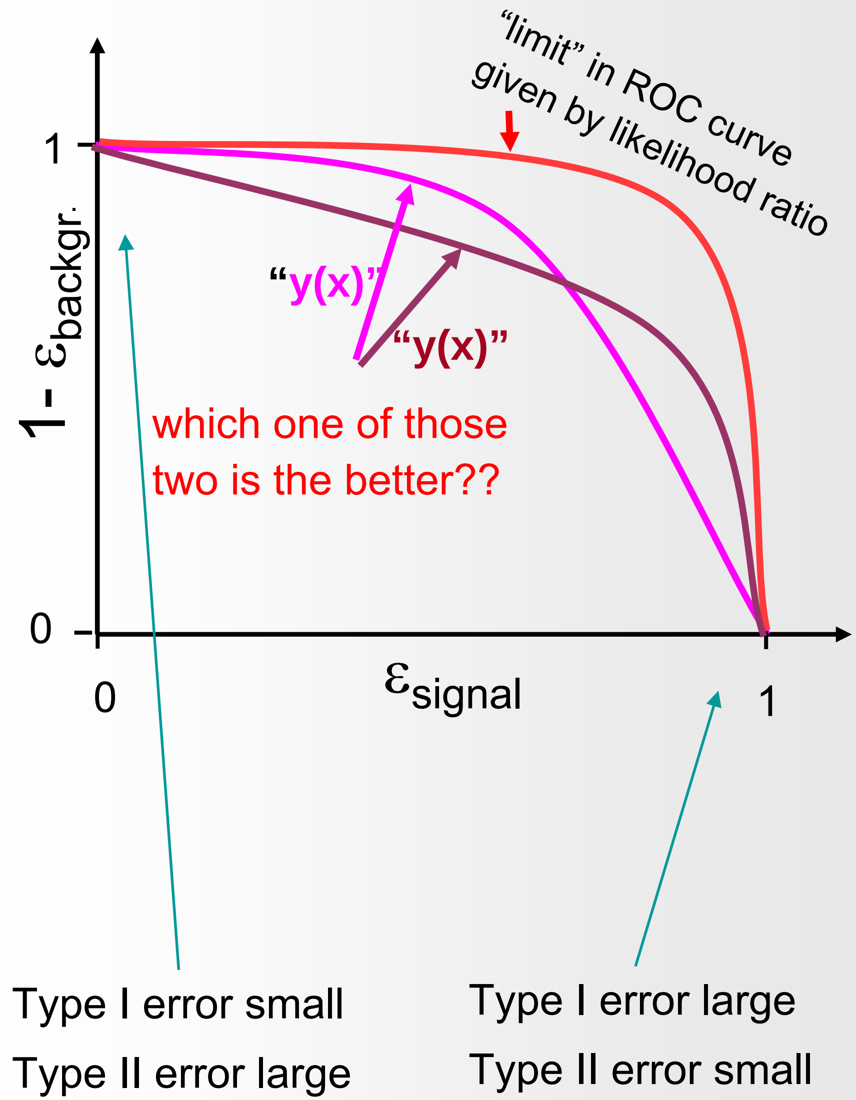
measurement of signal cross section:	maximum of $S/\sqrt{(S+B)}$ or equiv. $\sqrt{(\varepsilon \cdot p)}$
discovery of a signal (typically: S<<B):	maximum of $S/\sqrt{(B)}$
precision measurement:	high purity (p)
trigger selection:	high efficiency (ε)

Efficiency or Purity ?

if discriminating function y(x) = “true likelihood ratio”
→ optimal working point for specific analysis lies somewhere on the ROC curve

y(x) ≠ “true likelihood ratio” differently, point
→ y(x) might be better for a specific working point than **y(x)** and vice versa

- Note: for the determination of your working point (e.g. S/ √(B)) you need the prior S and B probabilities! → number of events/luminosity



Event Classification

- Unfortunately, the true probability densities functions are typically unknown:

→ Neyman-Pearsons lemma doesn't really help us directly

- HEP → Monte Carlo simulation or in general cases: set of known (already classified) "events"

Use these "training" events to:

- try to estimate the functional form of p(x|C): (e.g. the differential cross section folded with the detector influences) from which the likelihood ratio can be obtained
  → e.g. D-dimensional histogram, Kernel densitiy estimators, …

- find a "discrimination function" y(x) and corresponding decision boundary (i.e. hyperplane* in the "feature space": y(x) = const) that optimially separates signal from background
  → e.g. Linear Discriminator, Neural Networks, …

→ supervised (machine) learning

* hyperplane in the strict sense goes through the origin. Here I mean "affine set" to be precise