

Practical Statistics – part I

‘Basics Concepts’

W. Verkerke (NIKHEF)

What do we want to know?

- Physics questions we have...
 - Does the (SM) Higgs boson exist?
 - What is its production cross-section?
 - What is its boson mass?

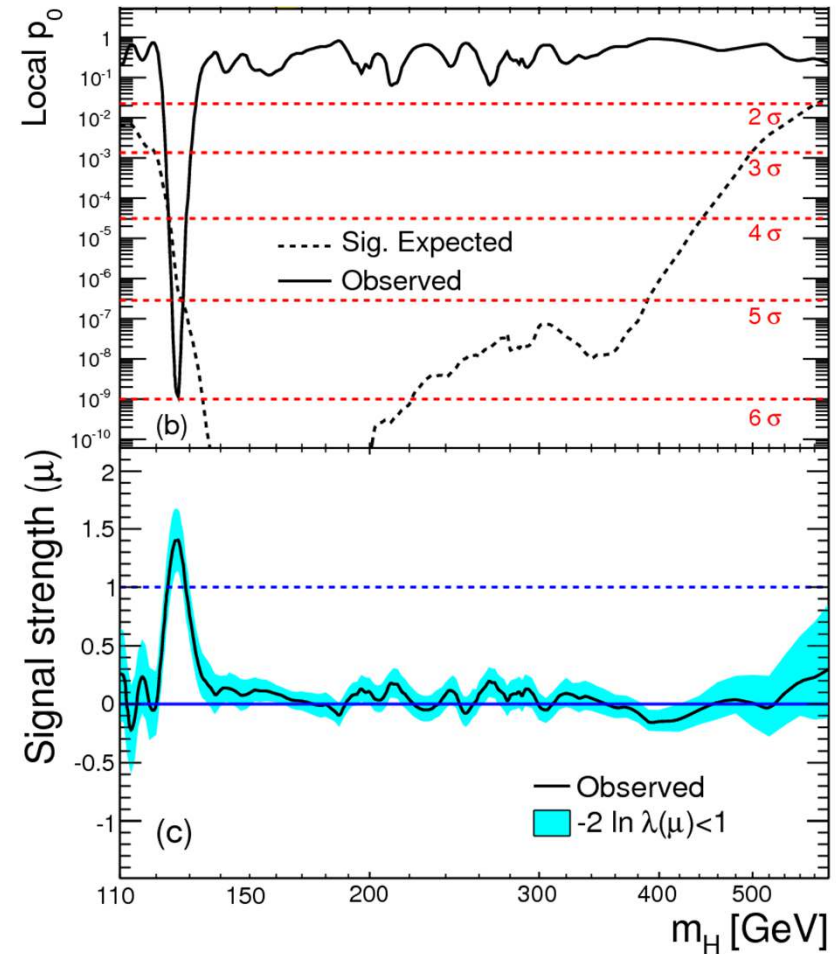


- Statistical tests construct probabilistic statements:
 $p(\text{theo}|\text{data})$, or $p(\text{data}|\text{theo})$
 - Hypothesis testing (discovery)
 - (Confidence) intervalsMeasurements & uncertainties



- Result: *Decision* based on tests

“As a layman I would now say: I think we have it”

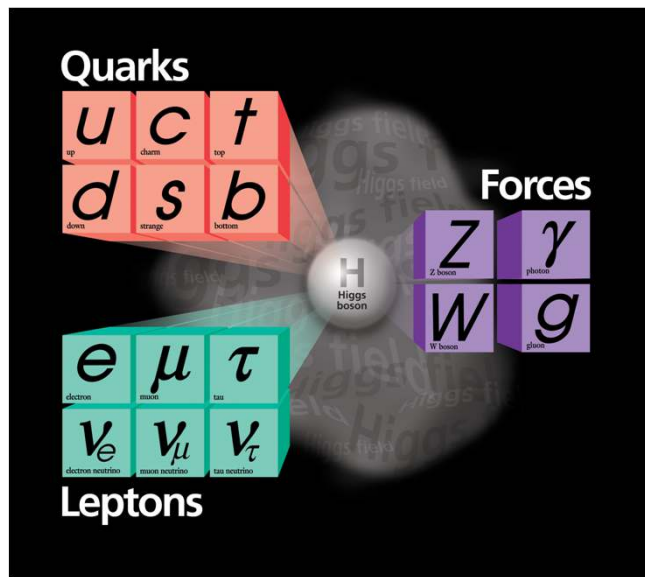


Wo

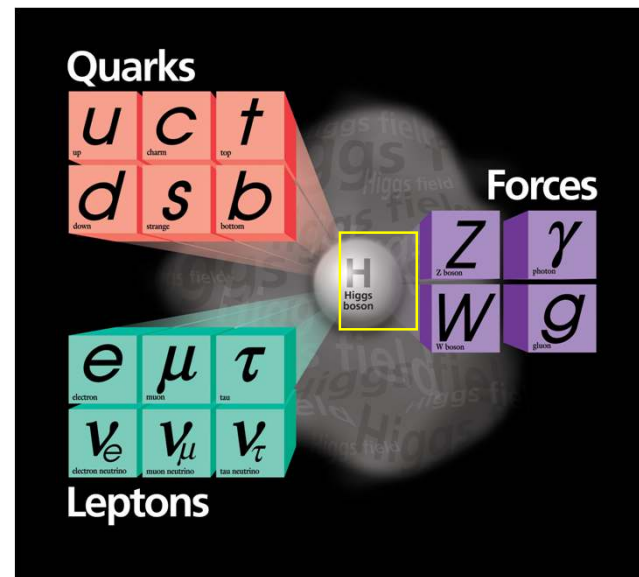
How do we do this?

- All experimental results start with formulation of a (physics) theory
- Examples of HEP physics models being tested

The Standard Model



The SM without a Higgs boson

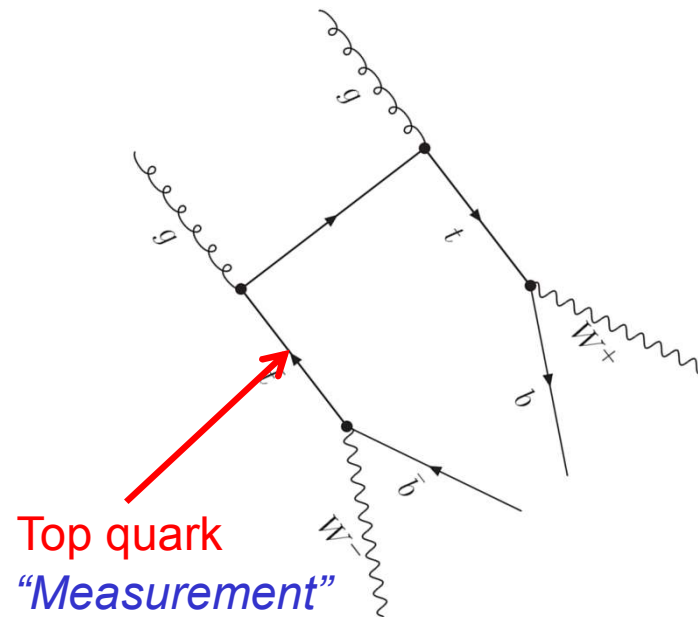
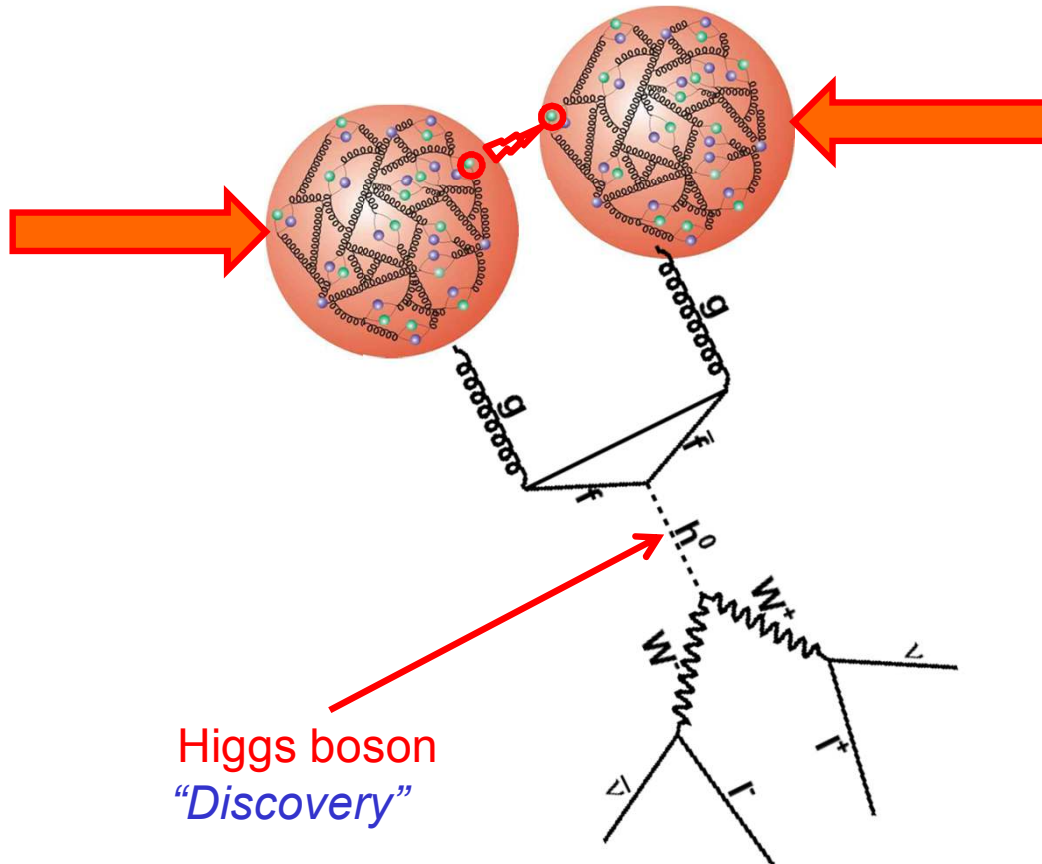


- Next, you design a measurement to be able to *test* model
 - Via chain of physics simulation, showering MC, detector simulation and analysis software, a physics model is reduced to a statistical model

How do we do this?

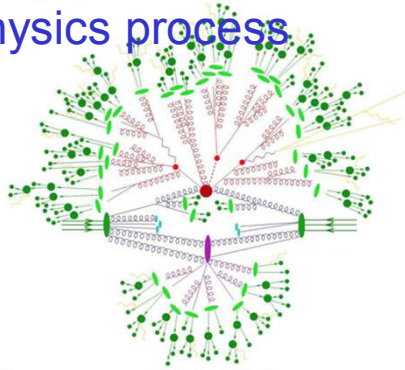


- General infrastructure collider, detector, computing
- Design analysis strategy to
 - observe existence of a particular process,
 - measure property of a produced particle

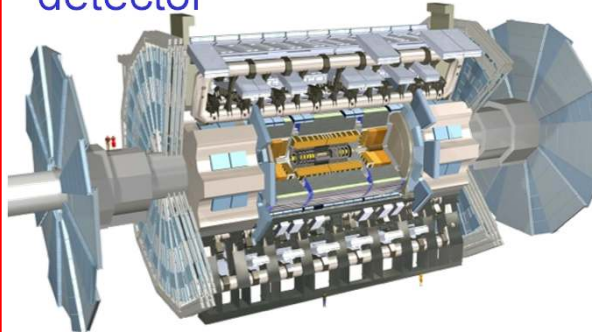


An overview of HEP data analysis procedures

Simulation of 'soft physics' physics process



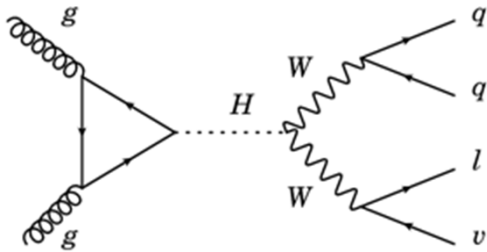
Simulation of ATLAS detector



LHC data

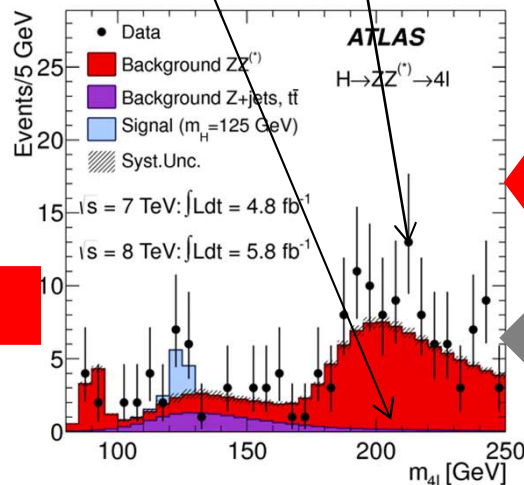


Simulation of high-energy physics process



$P(m_{4l} | \text{SM}[m_H])$

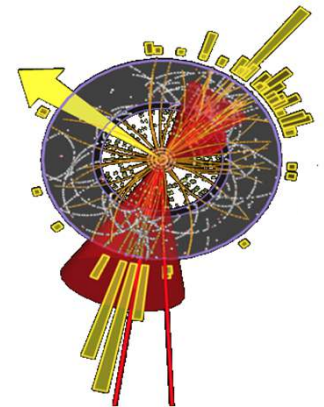
Observed m_{4l}



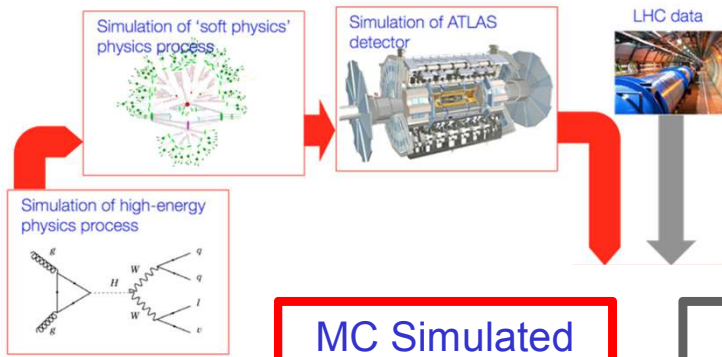
$\text{prob}(\text{data} | \text{SM})$

Analysis Event selection

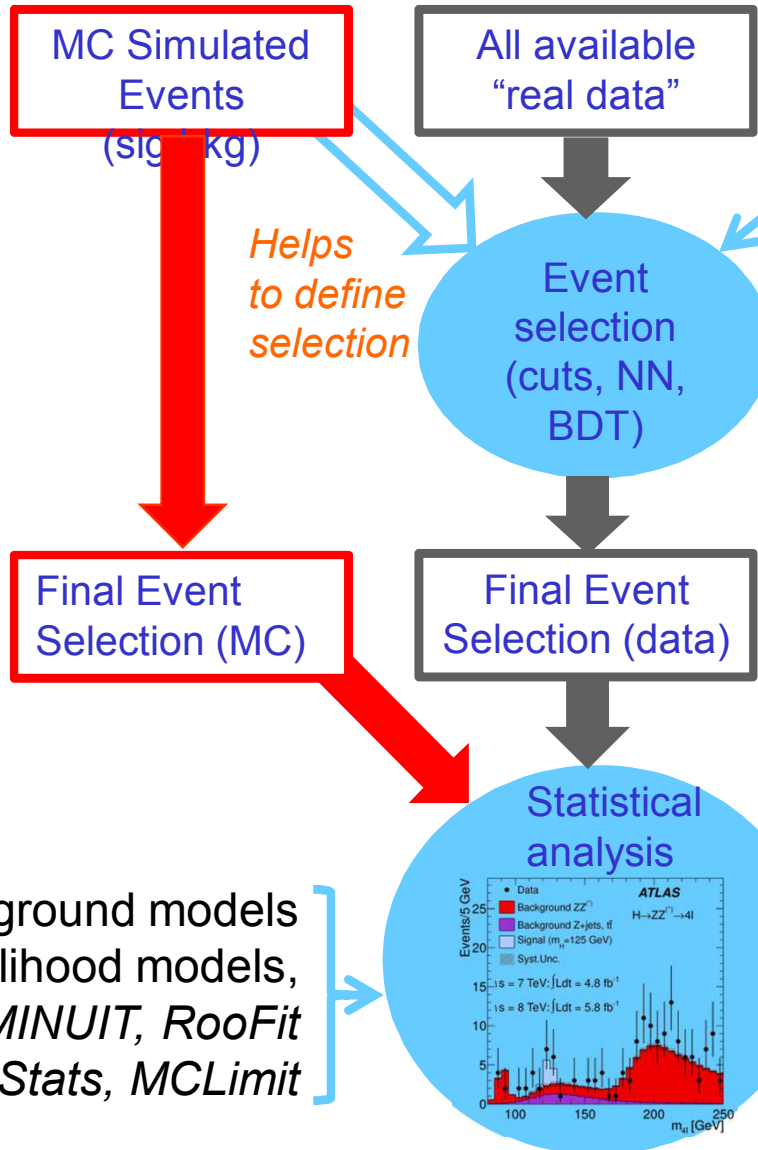
Reconstruction of ATLAS detector



An overview of HEP data analysis procedures



HEP workflow: data analysis in practice



N-tuples
Cut-flows,
Multi-variate analysis (NN, BDT)
ROOT, TMVA, NeuroBayes

Helps
to define
selection

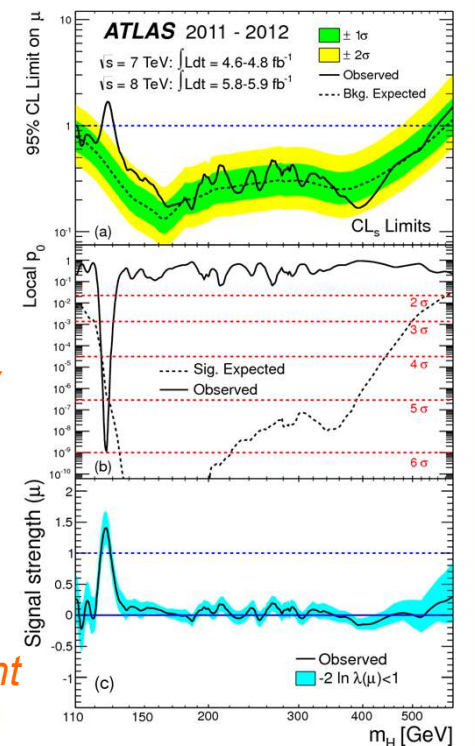
Signal, background models
Likelihood models,
MINUIT, RooFit
RooStats, MCLimit

Final Result

Limit

Discovery

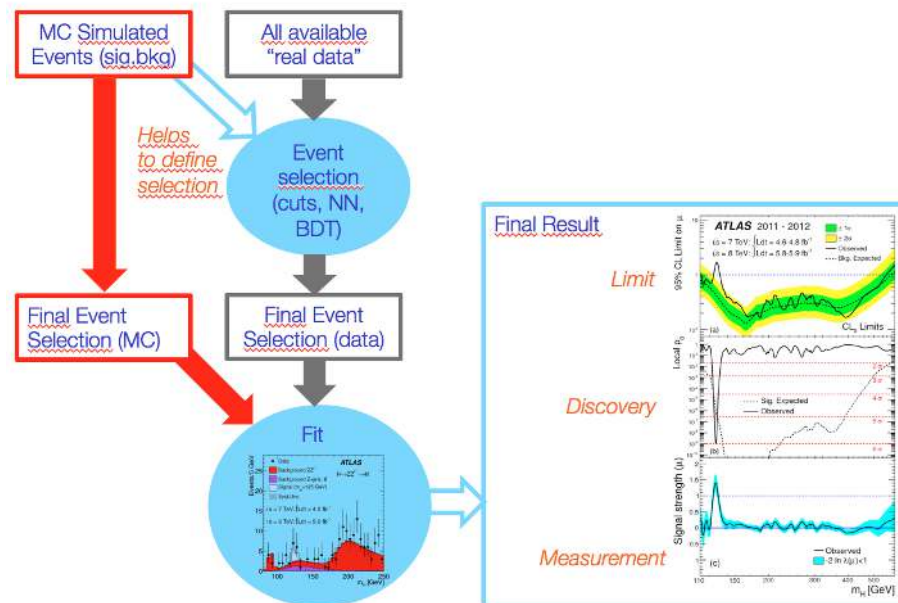
Measurement



From physics theory to statistical model

- HEP “Data Analysis” is for large part the reduction of a physics theory to a statistical model

Physics Theory: Standard Model with 125 GeV Higgs boson

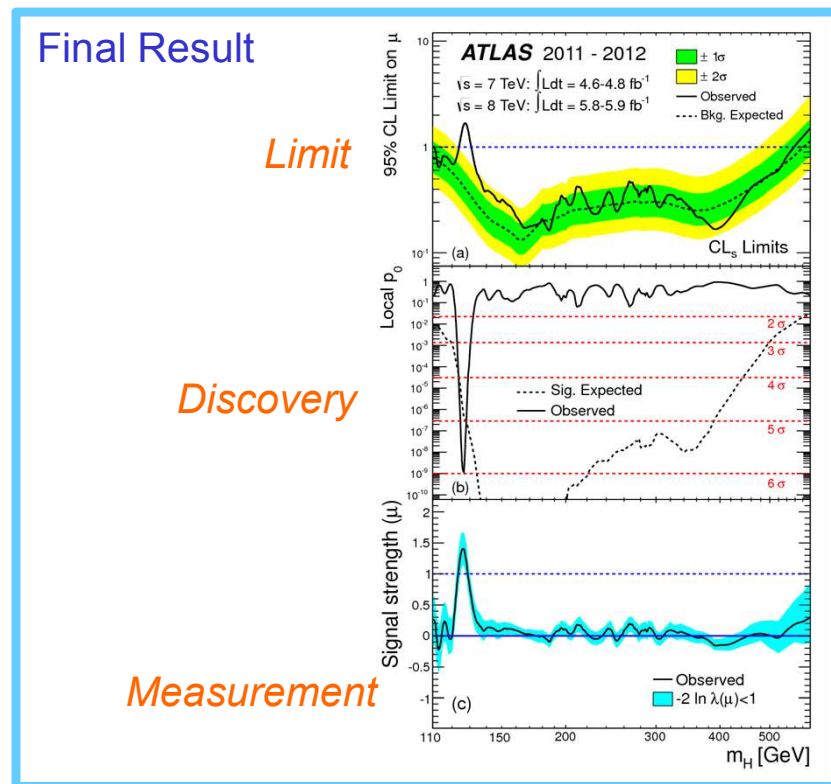


Statistical Model: *Given a measurement x (e.g. an event count)
what is the probability to observe each possible value of x ,
under the hypothesis that the physics theory is true.*

Once you have a statistical model, all physics knowledge has been abstracted into the model, and further steps in statistical inference are ‘procedural’ (no physics knowledge is required in principle)

From statistical model to a result

- The next step of the analysis is to confront your model with the data, and summarize the result in a probabilistic statement of some form



‘Confidence/Credible Interval’

$$\sigma/\sigma_{\text{SM}} (H \rightarrow ZZ) |_{m_H=150} < 0.3 \text{ @ 95\% C.L.}$$

‘p-value’

“Probability to observed this signal or more extreme, under the hypothesis of background-only is 1×10^{-9} ”

‘Measurement with variance estimate’

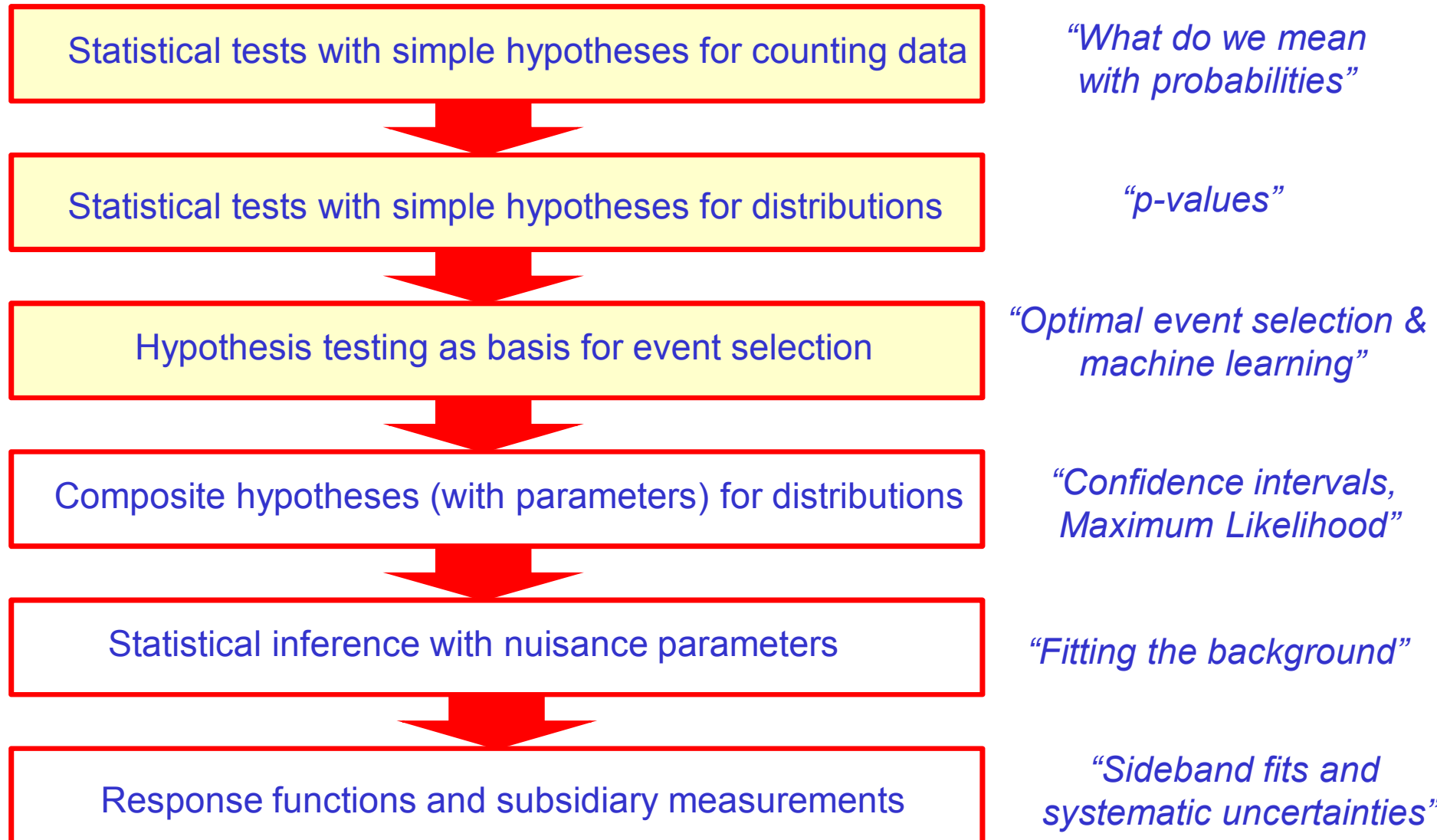
$$\sigma/\sigma_{\text{SM}} (H \rightarrow ZZ) |_{m_H=126} = 1.4 \pm 0.3$$

- The last step, usually not in a (first) paper, that you, or your collaboration, *decides* if your theory is valid



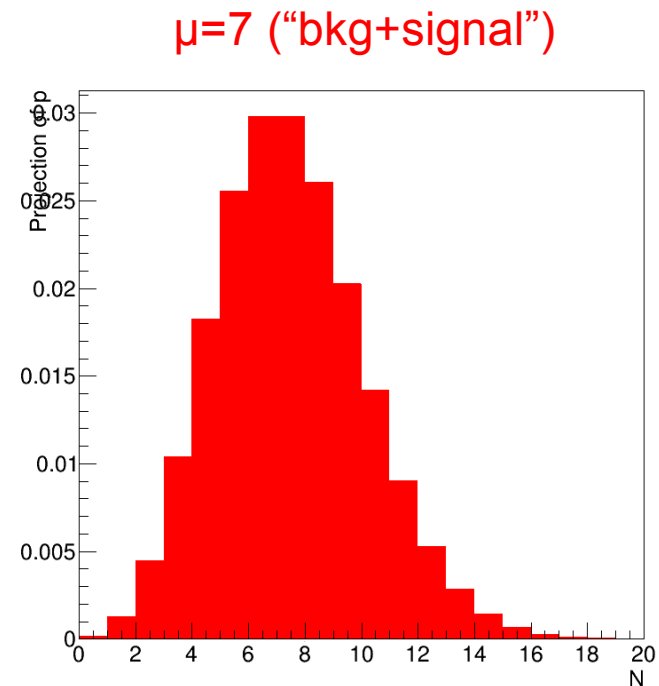
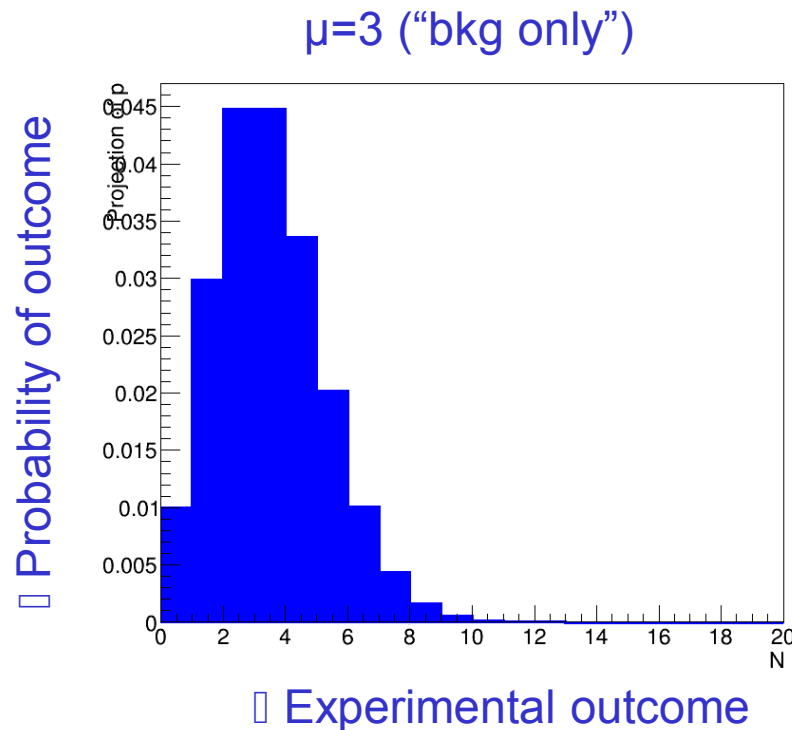
Roadmap for this course

- Start with basics, gradually build up to complexity of



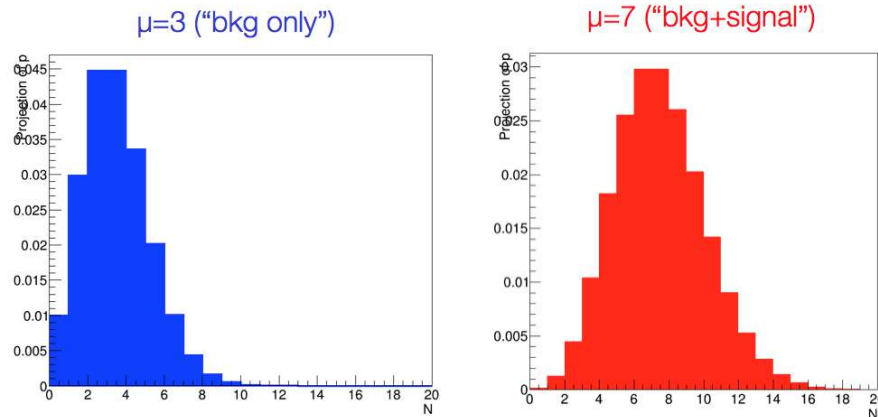
The statistical world

- Central concept in statistics is the ‘probability model’
- *A probability model assigns a probability to each possible experimental outcome.*
- Example: a HEP counting experiment $P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$
 - Count number of ‘events’ in a fixed time interval \square Poisson distribution
 - Given the *expected event count*, the probability model is fully specified



Probabilities vs conditional probabilities

- Note that probability models strictly give *conditional* probabilities (with the condition being that the underlying hypothesis is true)



Definition:
 $P(\text{data}|\text{hypo})$ is called
the likelihood

$$P(N) \rightarrow P(N | H_{bkg}) \quad P(N) \rightarrow P(N | H_{sig+bkg})$$

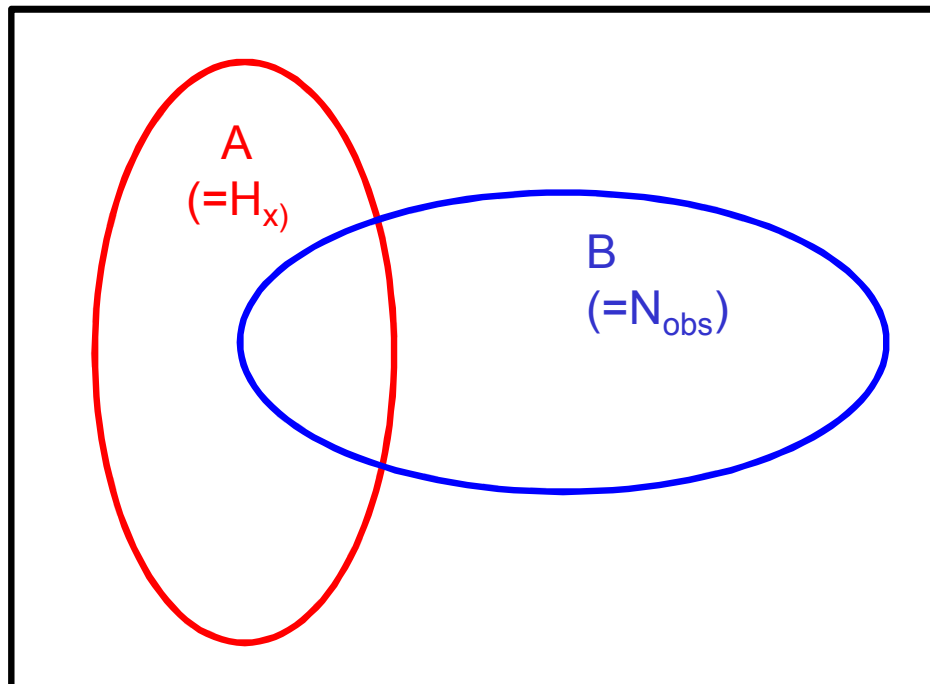
- Suppose we measure $N=7$ then can calculate

$$L(N=7|H_{bkg})=2.2\% \quad L(N=7|H_{sig+bkg})=14.9\%$$

- Data is more likely under sig+bkg hypothesis than bkg-only hypo*
- Is this what we want to know? Or do we want to know $L(H|N=7)$?

Inverting the conditionality on probabilities

- Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$
- No!
- Image the 'whole space' and two subsets A and B



$$P(A) = \frac{\text{small blue oval}}{\text{large blue rectangle}}$$

$$P(B) = \frac{\text{small blue oval}}{\text{large blue rectangle}}$$

$$P(A|B) = \frac{\text{tiny blue oval}}{\text{medium blue oval}}$$

$$P(B|A) = \frac{\text{tiny blue oval}}{\text{medium blue oval}}$$

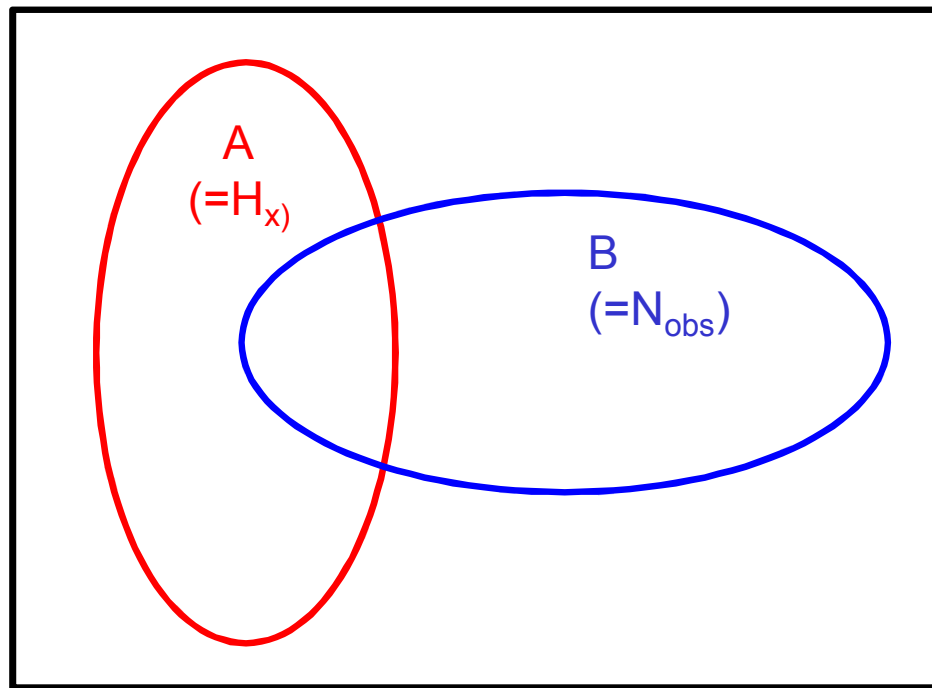
↓

$$P(A|B) \neq P(B|A)$$

↓

$$P(7|H_b) \neq P(H_b|7)$$

Inverting the conditionality on probabilities



$$P(A) = \frac{\text{blue oval}}{\text{blue square}}$$

$$P(B) = \frac{\text{blue oval}}{\text{blue square}}$$

$$P(A|B) = \frac{\text{small blue oval}}{\text{blue oval}}$$

$$P(B|A) = \frac{\text{small blue oval}}{\text{blue oval}}$$



$$P(A|B) \neq P(B|A)$$



but you can deduce
their relation



$$P(A) \times P(B|A) = \frac{\text{blue oval}}{\text{blue square}} \times \frac{\text{small blue oval}}{\text{blue oval}} = \frac{\text{small blue oval}}{\text{blue square}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{blue oval}}{\text{blue square}} \times \frac{\text{small blue oval}}{\text{blue oval}} = \frac{\text{small blue oval}}{\text{blue square}} = P(A \cap B)$$

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

Inverting the conditionality on probabilities

- This conditionality inversion relation is known as Bayes Theorem

$$P(B|A) = P(A|B) \times P(B)/P(A)$$

Essay "Essay Towards Solving a Problem in the Doctrine of Chances" published in Philosophical Transactions of the Royal Society of London in 1764



Thomas Bayes (1702-61)

- And choosing A =data and B =theory

$$P(\text{theo}|\text{data}) = P(\text{data}|\text{theo}) \times P(\text{theo}) / P(\text{data})$$

- Return to original question:*

Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$

- No! \square Need $P(A)$ and $P(B)$ \square Need $P(H_b)$, $P(H_{sb})$ and $P(7)$

Inverting the conditionality on probabilities

- What is $P(\text{data})$?
- It is the probability of the data under *any* hypothesis
 - For Example for two competing hypothesis H_b and H_{sb}

$$P(N) = L(N|H_b)P(H_b) + L(N|H_{sb})P(H_{sb})$$

and generally for N hypotheses

$$P(N) = \sum_i P(N|H_i)P(H_i)$$

- Bayes theorem reformulated using law of total probability

$$P(\text{theo}|\text{data}) = \frac{L(\text{data}|\text{theo}) \times P(\text{theo})}{\sum_i L(\text{data}|\text{theo-i})P(\text{theo-i})}$$

- *Return to original question:* Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you
enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$
No! □ Still need $P(H_b)$ and $P(H_{sb})$

Prior probabilities

- What is the meaning of $P(H_b)$ and $P(H_{sb})$?
 - They are the probability assigned to hypothesis H_b *prior to the experiment*.
- What are the values of $P(H_b)$ and $P(H_{sb})$?
 - Can be result of an earlier measurement
 - Or more generally (e.g. when there are no prior measurement) they quantify *a prior degree of belief* in the hypothesis
- **Example** – suppose prior belief $P(H_{sb})=50\%$ and $P(H_b)=50\%$

$$\begin{aligned} P(H_{sb}|N=7) &= \frac{P(N=7|H_{sb}) \times P(H_{sb})}{[P(N=7|H_{sb})P(H_{sb})+P(N=7|H_b)P(H_b)]} \\ &= \frac{0.149 \times 0.50}{[0.149 \times 0.5 + 0.022 \times 0.5]} = 87\% \end{aligned}$$

- Observation $N=7$ strengthens belief in hypothesis H_{sb}
(and weakens belief in $H_b \rightarrow 13\%$)

Interpreting probabilities

- We have seen

probabilities assigned observed experimental outcomes
(probability to observed 7 events under some hypothesis)

probabilities assigned to hypotheses
(prior probability for hypothesis H_{sb} is 50%)

which are conceptually different.

- How to interpret probabilities – two schools

Bayesian probability = (subjective) degree of belief $\frac{P(\text{theo}|\text{data})}{P(\text{data}|\text{theo})}$

Frequentist probability = fraction of outcomes in $P(\text{data}|\text{theo})$
future repeated identical experiments

*“If you’d repeat this experiment identically many times,
in a fraction P you will observe the same outcome”*

Interpreting probabilities

- Frequentist:
Constants of nature are fixed – you cannot assign a probability to these. Probability are restricted to observable experimental results
 - “The Higgs either exists, or it doesn’t” – you can’t assign a probability to that
- Bayesian:
Probabilities can be assigned to constants of nature
 - Quantify your *belief* in the existence of the Higgs – can assign a probability
- **Example of weather forecast**

Bayesian: “*The probability it will rain tomorrow is 95%*”

- Assigns probability to constant of nature (“rain tomorrow”)
 $P(\text{rain-tomorrow}|\text{weather-data}) = 95\%$

Frequentist: “*It will rain tomorrow^(*)*”

(*) 95% of the forecast are correct.

- Only states $P(\text{weather-data}|\text{rain-tomorrow})$

Bayesians and Frequentists

- A slide from a professional statistician found when Googling...

ACCP 37th Annual Meeting, Philadelphia, PA [2]

Differences Between Bayesians and Non-Bayesians According to my friend Jeff Gill



Typical Bayesian



Typical Non-Bayesian

Bayesians and Frequentists

- Another slide from a particle physicist..

Why isn't everyone a Bayesian ?

My suspicion: it is because most people do not understand the frequentist approach. Frequentist statements and Bayesian statements are thought to be about the same logical concept, and the frequentist statement does not require a prior, so ...

A. L. Read, *Presentation of search results: the CL_s technique*, J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693-2704.

nearly all physicists tend to misinterpret frequentist results as statements about the theory given the data.

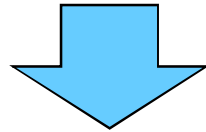
Frequentist statements are not statements about the model – only about the data in the context of the model. This is not what we wanted to know ... At least not the ultimate statement.

Formulating evidence for discovery

- Given a scenario with exactly two competing hypotheses
- In the Bayesian school you can cast evidence as an odd-ratio

$$O_{prior} \equiv \frac{P(H_{sb})}{P(H_b)} = \frac{P(H_{sb})}{1 - P(H_{sb})}$$

If $p(H_{sb})=p(H_b)$ Odds are 1:1



'Bayes Factor' K multiplies prior odds

$$O_{posterior} \equiv \frac{L(x | H_{sb})P(H_{sb})}{L(x | H_b)P(H_b)} = \overbrace{\frac{L(x | H_{sb})}{L(x | H_b)}}^{K} O_{prior}$$

If $\begin{matrix} P(\text{data}|H_b)=10^{-7} \\ P(\text{data}|H_{sb})=0.5 \end{matrix}$ $K=2.000.000$ Posterior odds are 2.000.000 : 1

Formulating evidence for discovery

- In the frequentist school you restrict yourself to $P(\text{data}|\text{theory})$ and there is no concept of ‘priors’
 - But given that you consider (exactly) 2 competing hypothesis, very low probability for data under H_b lends credence to ‘discovery’ of H_{sb} (since H_b is ‘ruled out’). Example

$$\begin{array}{l} P(\text{data}|H_b)=10^{-7} \\ P(\text{data}|H_{sb})=0.5 \end{array} \Rightarrow \text{“}H_b \text{ ruled out”} \Rightarrow \text{“Discovery of } H_{sb}\text{”}$$

- Given importance to interpretation of the lower probability, it is customary to quote it in “physics intuitive” form: Gaussian σ .
 - E.g. ‘5 sigma’ \Rightarrow probability of 5 sigma Gaussian fluctuation $=2.87 \times 10^{-7}$
- No formal rules for ‘discovery threshold’
 - Discovery also assumed is not too unlikely under H_{sb} . If not, no discovery, but again no formal rules (“your good physics judgment”)
 - NB: In Bayesian case, both likelihoods low reduces Bayes factor K to $O(1)$

Taking decisions based on your result

- What are you going to do with the results of your measurement?
- Usually basis for a decision
 - **Science**: declare discovery of Higgs boson (or not), make press release, write new grant proposal
 - **Finance**: buy stocks or sell
- Suppose you believe $P(\text{Higgs}|\text{data})=99\%$.
- **Should declare discovery, make a press release?**
A: Cannot be determined from the given information!
- Need in addition: the utility function (or cost function),
 - The cost function specifies the relative costs (to You) of a Type I error (declaring model false when it is true) and a Type II error (not declaring model false when it is false).

Taking decisions based on your result

- Thus, your *decision*, such as where to invest your time or money, requires two subjective inputs:

Your prior probabilities, and

the relative costs to You of outcomes.

- Statisticians often focus on decision-making; in HEP, the tradition thus far is to communicate experimental results (well) short of formal decision calculations.
- Costs can be difficult to quantify in science.
 - What is the cost of declaring a false discovery?
 - Can be high (“Fleischman and Pons”), but hard to quantify
 - What is the cost of missing a discovery (“Nobel prize to someone else”), but also hard to quantify

How a theory becomes text-book physics

Frequentist

Information from experiment

$$P(\text{data}|H_b)=10^{-7}$$
$$P(\text{data}|H_{sb})=0.5$$

*P-value threshold from “prior”
(judgment call – no formal theory!)*

A: declare discovery at 3σ
B: declare discovery at 5σ

*Recent judgements
on of 5σ effects:
Higgs – text book
 $v(\beta>1)$ – rejected*

Press release, accept as new
‘text book physics’
OR
Wait for more data

Potentially fuzzy information

*Prior belief in theory
(can be hard to quantify)*

$$A: P(H_{sb})=50\%$$

$$B: P(H_{sb})=0.000001\%$$

*Cost of wrong decision
(can be hard to quantify)*

Cost(FalseDiscovery)
= EternalRidicule/Fired

Cost(UnclaimedDiscovery)
= MissedNobelPrize

Bayesian

Information from experiment

$$P(\text{data}|H_b)=10^{-7}$$
$$P(\text{data}|H_{sb})=0.5$$

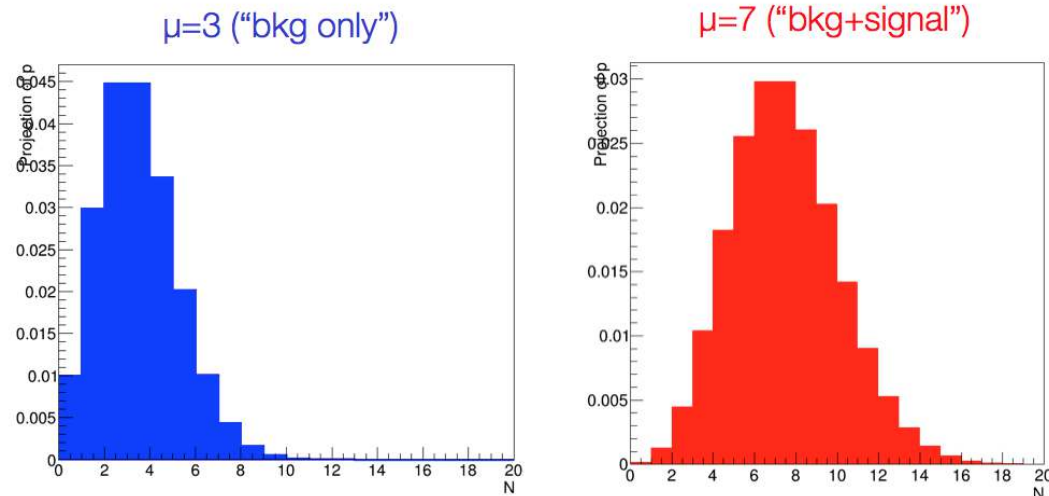
*Posterior from expt and prior
following Bayesian paradigm*

$$A: P(H_{sb}|\text{data})=0.9999998$$
$$B: P(H_{sb}|\text{data}) = 83\%$$

Press release, accept as new
‘text book physics’
or
Wait for more data

Summary on statistical test with simple hypotheses

- So far we considered simplest possible experiment we can do: counting experiment
- For a set of 2 or more completely specified (i.e. simple) hypotheses



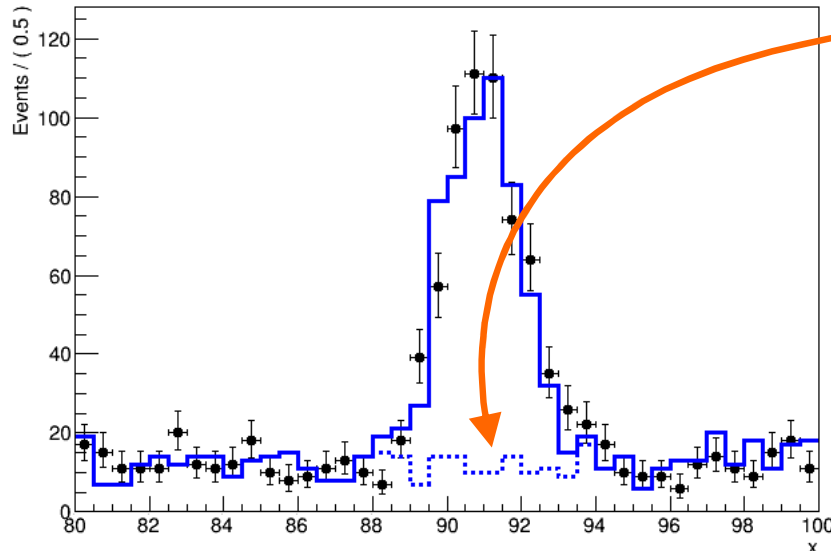
□ Given probability models $P(N|\text{bkg})$, and $P(N|\text{sig})$
we can calculate $P(\text{Nobs}|\text{H}_x)$ under both hypothesis

□ With additional information on $P(H_i)$ we can also calculate $P(\text{H}_x|\text{Nobs})$

- In principle, *any potentially complex measurement (for Higgs, SUSY, top quarks) can ultimately take this a simple form.*
But there is some 'pre-work' to get here – examining (multivariate) discriminating distributions □ Now try to incorporate that

Practical statistics – (Multivariate) distributions

- Most realistic HEP analysis are not like simple counting expts at all
 - Separation of signal-like and background-like is a complex task that involves study of many observable distributions
- How do we deal with distributions in statistical inference?
 - Construct a probability model for the distribution
- Case 1 – Signal and background distributions from MC simulation
 - Typically have *histograms* for signal and background



counting experiment

product of Likelihoods for each bin

$$L(\vec{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$$

$$L(\vec{N} | H_{s+b}) = \prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)$$

Working with Likelihood functions for distributions

- How do the statistical inference procedures change for Likelihoods describing distributions?
- Bayesian calculation of $P(\text{theo}|\text{data})$ they are *exactly the same*.
 - Simply substitute counting model with binned distribution model

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

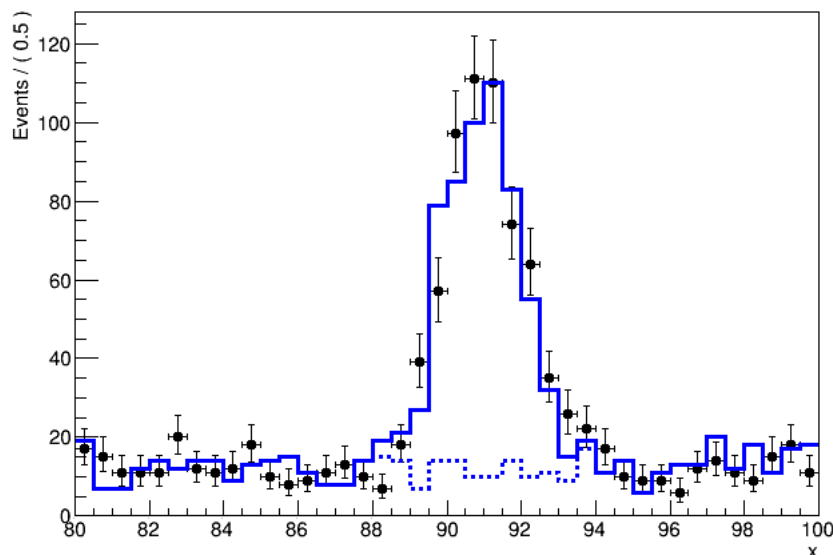


Simply fill in new Likelihood function
Calculation otherwise unchanged

$$P(H_{s+b} | \vec{N}) = \frac{\prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b})}{\prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b}) + \prod_i \text{Poisson}(N_i | \tilde{b}_i)P(H_b)}$$

Working with Likelihood functions for distributions

- Frequentist calculation of $P(\text{data}|\text{hypo})$ also unchanged, but question arises if $P(\text{data}|\text{hypo})$ is still relevant?



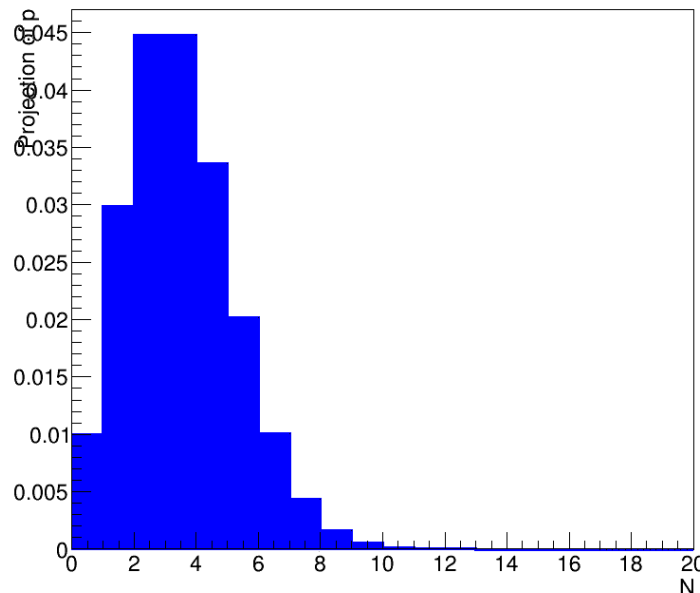
$$L(\vec{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$$

$$L(\vec{N} | H_{s+b}) = \prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)$$

- $L(N|H)$ is probability to obtain *exactly* the histogram observed.
- Is that what we want to know?* Not really.. We are interested in probability to observe any ‘similar’ dataset to given dataset, or in practice dataset ‘similar or more extreme’ that observed data
- Need a way to quantify ‘similarity’ or ‘extremity’ of observed data

Working with Likelihood functions for distributions

- *Definition:* a test statistic $T(x)$ is any function of the data
- We need a test statistic that will **classify ('order') all possible observations** in terms of 'extremity' (definition to be chosen by physicist)
- NB: For a counting measurement the count itself is already a useful test statistic for such an ordering (i.e. $T(x) = x$)



Test statistic $T(N) = N$ orders observed events count by estimated signal yield

Low N \Rightarrow low estimated signal

High N \Rightarrow large estimated signal

Ordering distributions by 'signal-likeness' aka 'extremity'

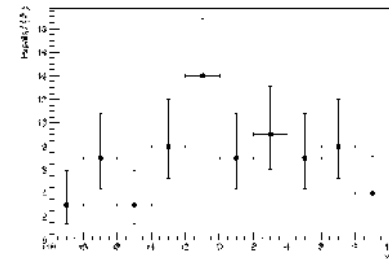
- How to define 'extremity' if observed data is a distribution

Counting

Histogram

Observation

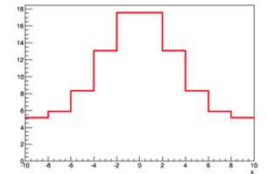
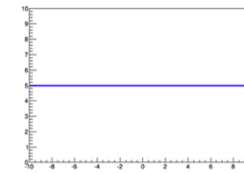
$$N_{\text{obs}}=7$$



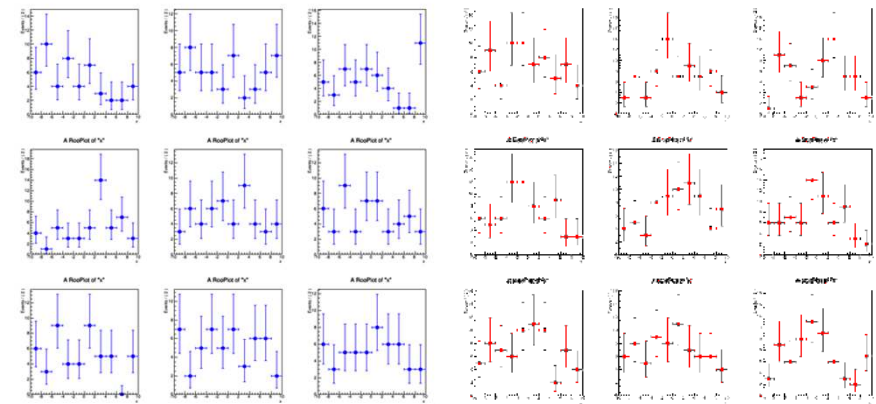
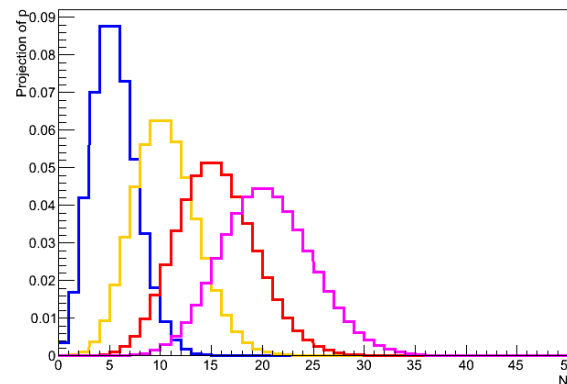
Median expected
by hypothesis

$$N_{\text{exp}}(s=0) = 5$$

$$N_{\text{exp}}(s=5) = 10$$



Predicted distribution
of observables



Which histogram is more 'extreme'?

The Likelihood Ratio as a test statistic

- Given two hypothesis H_b and H_{s+b} the ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$

- Intuitive picture:

□ If data is likely under H_b ,
 $L(N|H_b)$ is large,
 $L(N|H_{s+b})$ is smaller

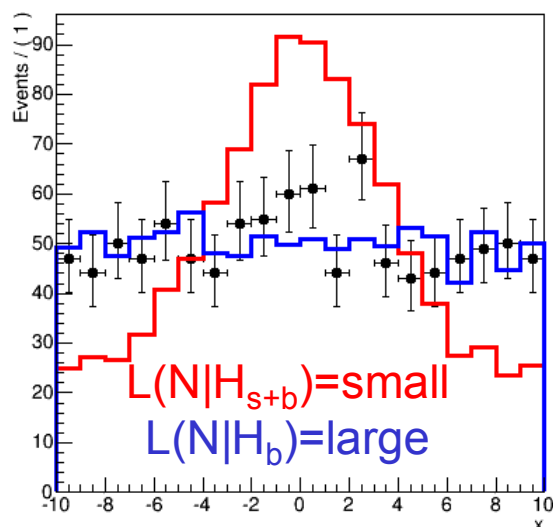
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

□ If data is likely under H_{s+b}
 $L(N|H_{s+b})$ is large,
 $L(N|H_b)$ is smaller

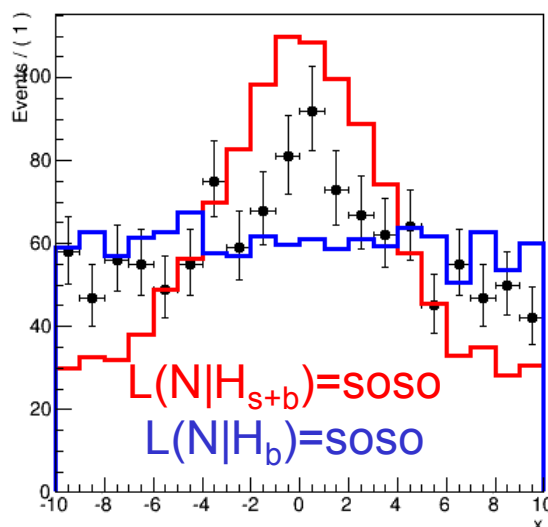
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

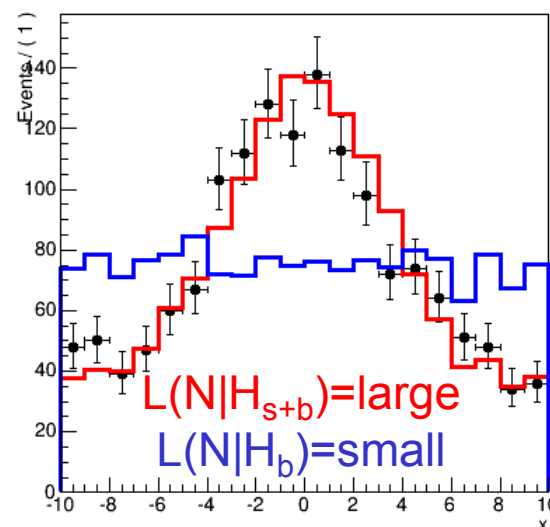
- The Likelihood ratio as ordering principle



$\lambda(N)=0.0005$



$\lambda(N)=0.47$

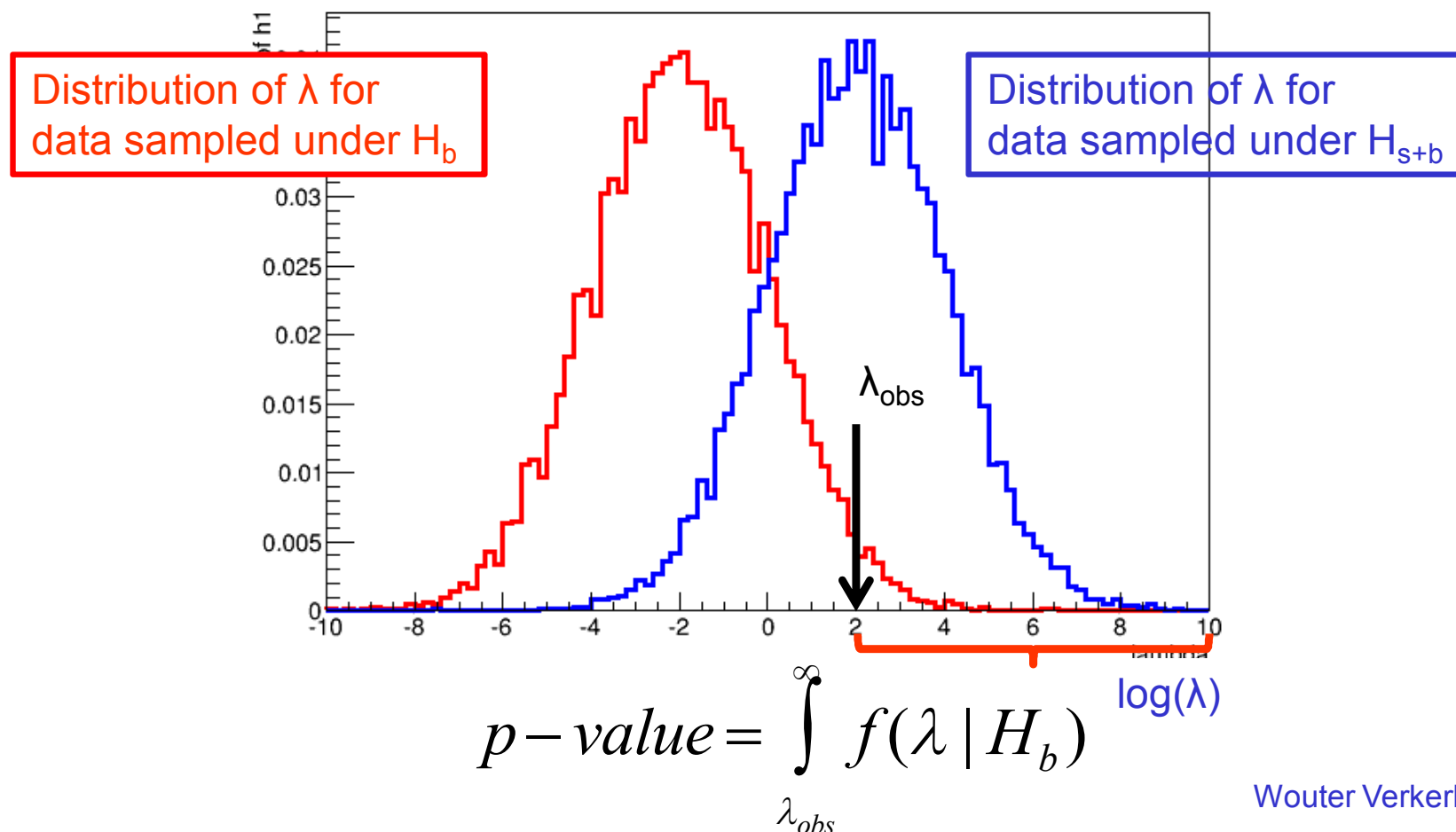


$\lambda(N)=5000$

- Frequentist solution to 'relevance of $P(\text{data}|\text{theory})$ ' is to classify all observed data using a (Likelihood Ratio) test statistic
 - Probability to observe 'similar data or more extreme' then amounts to calculating 'probability to observe test statistic $\lambda(N)$ as large or larger than the observed test statistic $\lambda(N_{\text{obs}})$ '

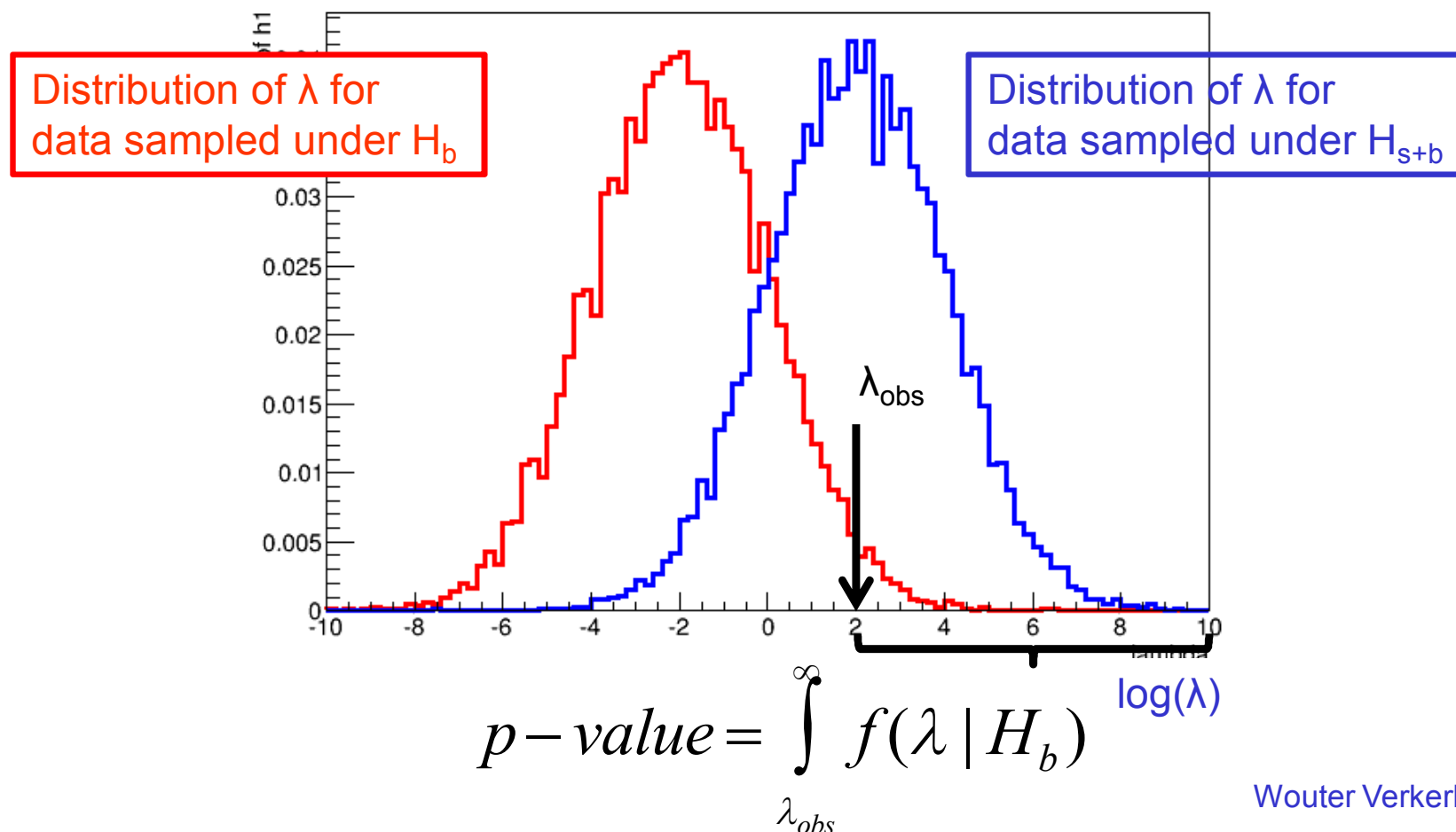
The distribution of the test statistic

- Distribution of a test statistic is generally not known
- Use toy MC approach to approximate distribution
 - Generate many toy datasets N under H_b and H_{s+b} and evaluate $\lambda(N)$ for each dataset



The distribution of the test statistic

- Definition: p-value:
probability to obtain the observed data, or more extreme
in future repeated identical experiments
(extremity define in the precise sense of the (LR) ordering rule)



Likelihoods for distributions - summary

- Bayesian inference unchanged

□ simply insert L of distribution to calculate $P(H|\text{data})$

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

- Frequentist inference procedure *modified*

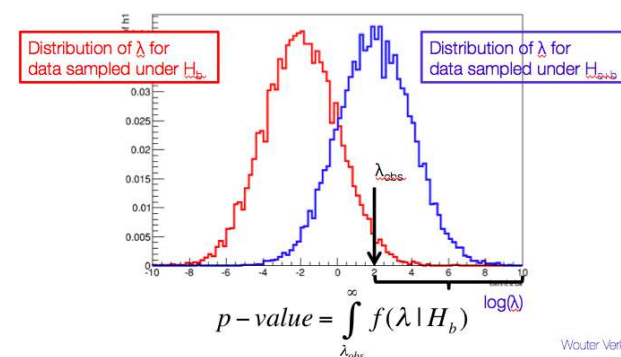
□ Pure $P(\text{data}|\text{hypo})$ not useful if data is a distribution

□ Order all possible data with a (LR) test statistic in 'extremity'

□ Quote $p(\text{data}|\text{hypo})$ as 'p-value' for hypothesis

Probability to obtain observed data, *or more extreme*, is X%

- Definition: p-value



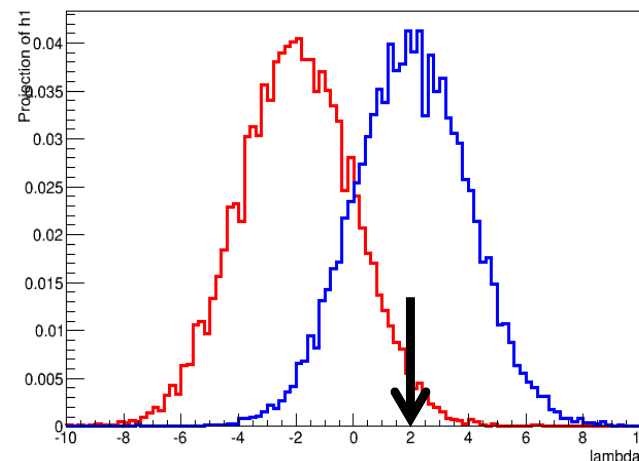
The likelihood principle

- Note that 'ordering procedure' introduced by test statistic also has a profound implication on procedure
- Bayesian inference only uses the Likelihood of the observed data

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b}) P(H_{s+b})}{L(\vec{N} | H_{s+b}) P(H_{s+b}) + L(\vec{N} | H_b) P(H_b)}$$

- While the observed Likelihood Ratio also only uses likelihood of observed data.

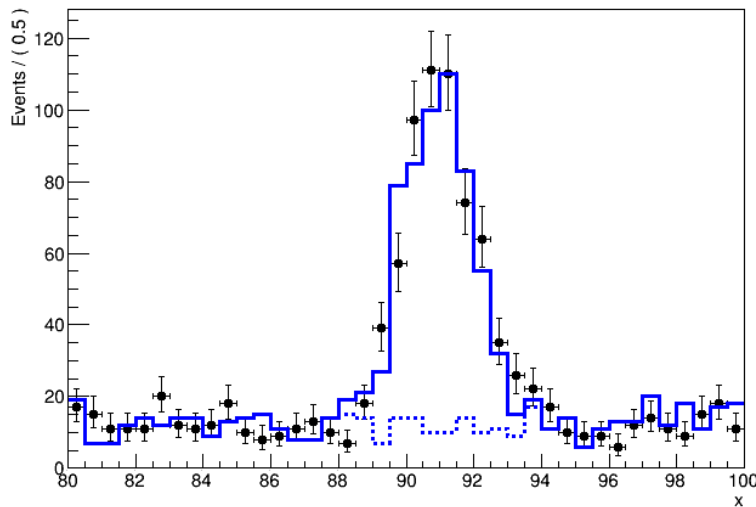
$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$



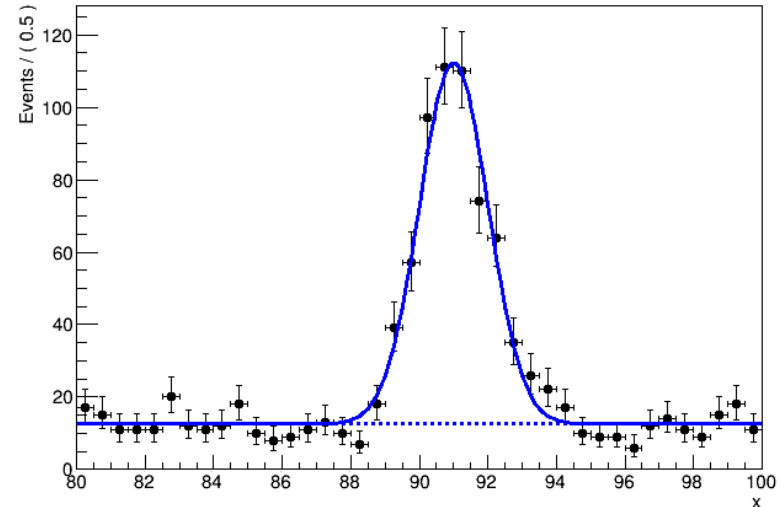
- Distribution $f(\lambda|N)$, and thus p-value, also uses likelihood of non-observed outcomes (in fact Likelihood of every possible outcome is used)

Generalizing to continuous distributions

- Can generalize likelihood to described continuous distributions



$$L(\vec{N}) = \prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)$$

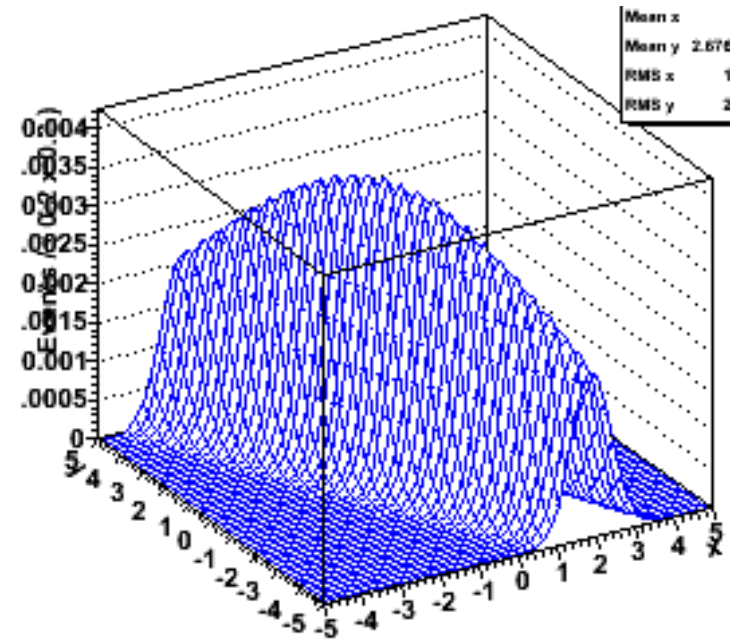
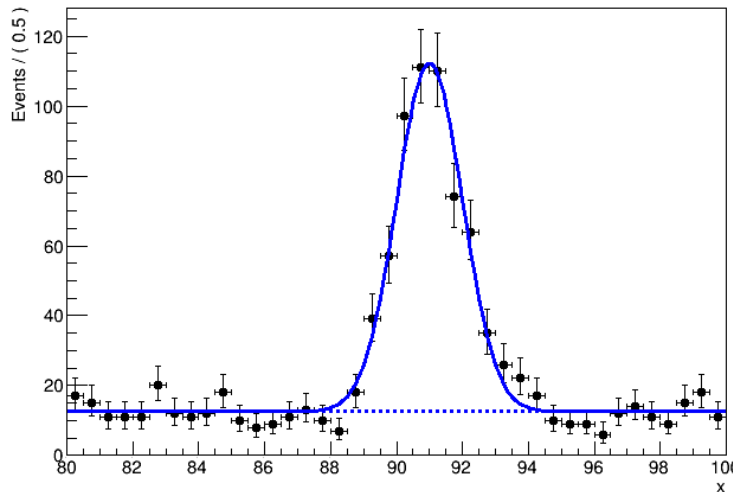


$$L(\vec{m}_{ll}) = \prod_i \left[\tilde{f}_{sig} \text{Gauss}(m_{ll}^{(i)}, 91, 1) + (1 - \tilde{f}_{sig}) \cdot \text{Uniform}(m_{ll}^{(i)}) \right]$$

- **Probability model becomes a probability *density* model**
 - Integral of probability density model over full space of observable is always 1 (just like sum of bins of a probability model is always 1)
 - Integral of p.d.f. over a range of observable results in a probability
- Probability density models have (in principle) more analyzing power
 - But relies on your ability to formulate an analytical model (e.g. hard at LHC)

Generalizing to multiple dimensions

- Can also generalize likelihood models to distributions in *multiple* observables



$$L(\vec{x}) = \prod_i f(x_i)$$

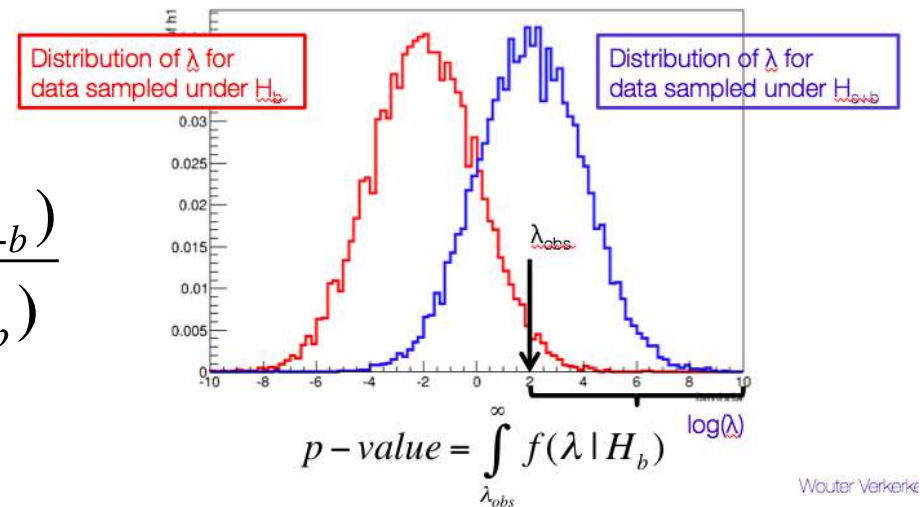
$$L(\vec{x}, \vec{y}) = \prod_i f(x_i, y_i)$$

- Neither generalization (binned→continuous, one→multiple observables) has any further consequences for Bayesian or Frequentist inference procedures

The Likelihood Ratio test statistic as tool for event selection

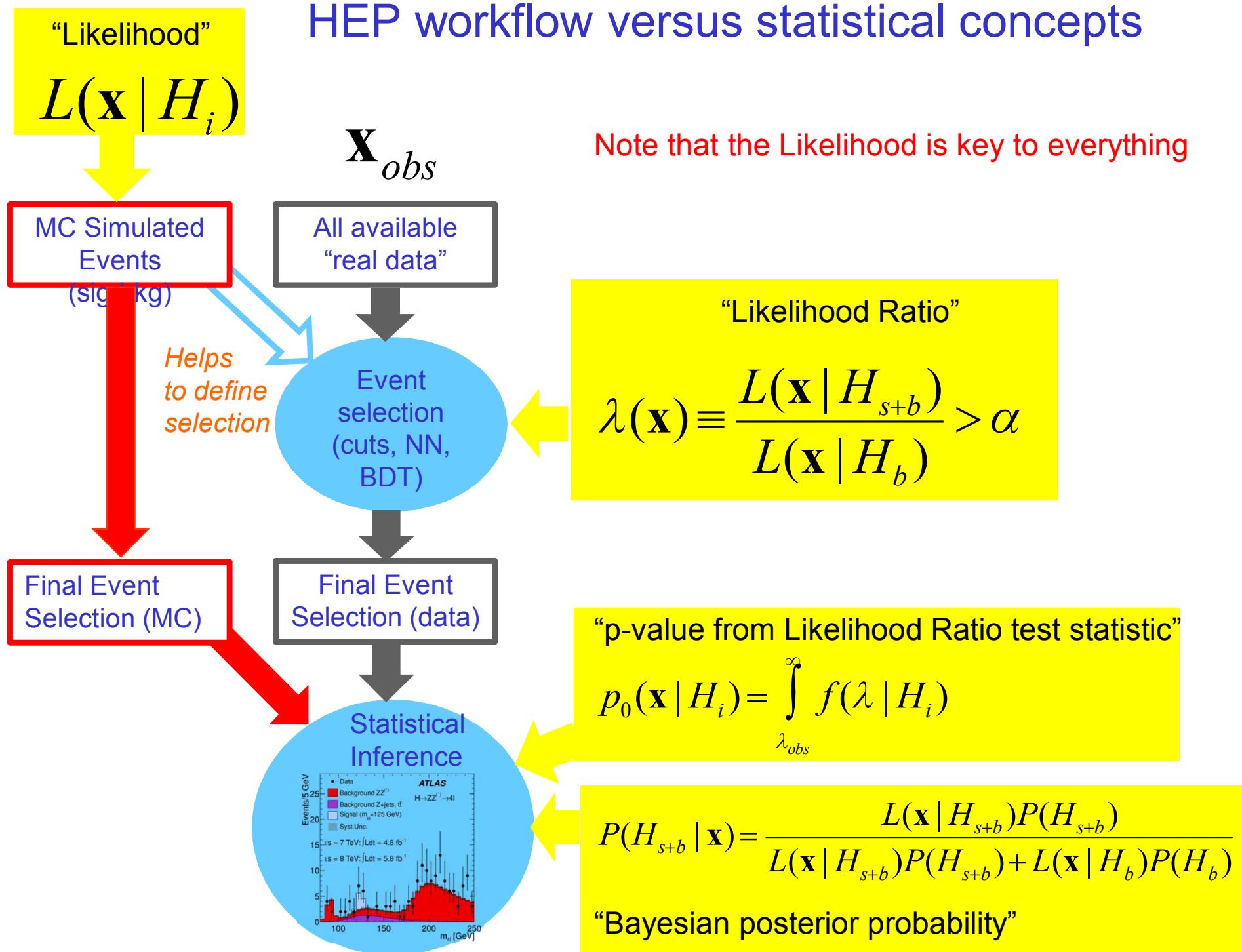
- Note that hypothesis testing with two simple hypotheses for observable distributions, exactly describes ‘event selection’ problem
- In fact we have already ‘solved’ the optimal event selection problem! Given two hypothesis H_{s+b} and H_b that predict an complex multivariate distribution of observables, you can always classify all events in terms of ‘signal-likeness’ (a.k.a ‘extremity’) with a likelihood ratio

$$\lambda(\vec{x}, \vec{y}, \vec{z}, \dots) = \frac{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_{s+b})}{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_b)}$$



- So far we have exploited λ to calculate a frequentist p-value will now explore properties ‘cut on λ ’ as basis of (optimal) event selection

HEP workflow versus statistical concepts



Event selection

- The event selection problem:
 - Input: Two classes of events “signal” and “background”
 - Output: Two categories of events “selected” and “rejected”
- Goal: select as many signal events as possible,
reject as many background events as possible
- Note that optimization goal as stated is ambiguous.
 - But can choose a well-defined by optimization goal by e.g. fixing desired background acceptance rate, and then choose procedure that has highest signal acceptance.
- Relates to “classical hypothesis testing”
 - Two competing hypothesis (traditionally named ‘null’ and ‘alternate’)
 - Here null = background, alternate = signal

Terminology of classical hypothesis testing

- Definition of terms

- Rate of type-I error = α
- Rate of type-II error = β
- Power of test is $1-\beta$

		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) Type I error
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) Type II error	True Negative

- Treat hypotheses asymmetrically

- Null hypo is usually special \Rightarrow Fix rate of type-I error
- Criminal convictions: Fix rate of unjust convictions
- Higgs discovery: Fix rate of false discovery
- Event selection: Fix rate of background that is accepted

- Now can define a well stated goal for optimal testing

- Maximize the power of test (minimized rate of type-II error) for given α
- Event selection: Maximize fraction of signal accepted

The Neyman-Pearson lemma

- In 1932-1938 Neyman and Pearson developed a theory in which one must consider competing hypotheses
 - Null hypothesis (H_0) = Background only
 - Alternate hypotheses (H_1) = e.g. Signal + Background

and proved that

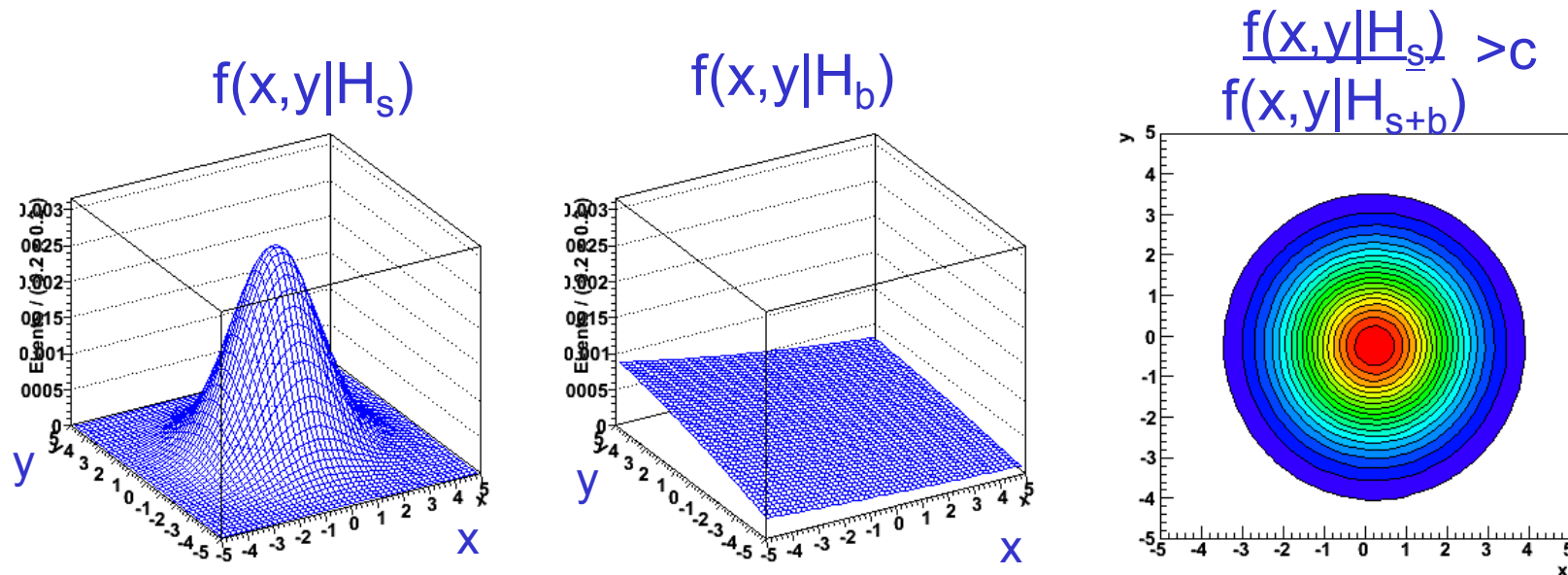
- The region W that minimizes the rate of the type-II error (not reporting true discovery) is a contour of the Likelihood Ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

- Any other region of the same size will have less power

The Neyman-Pearson lemma

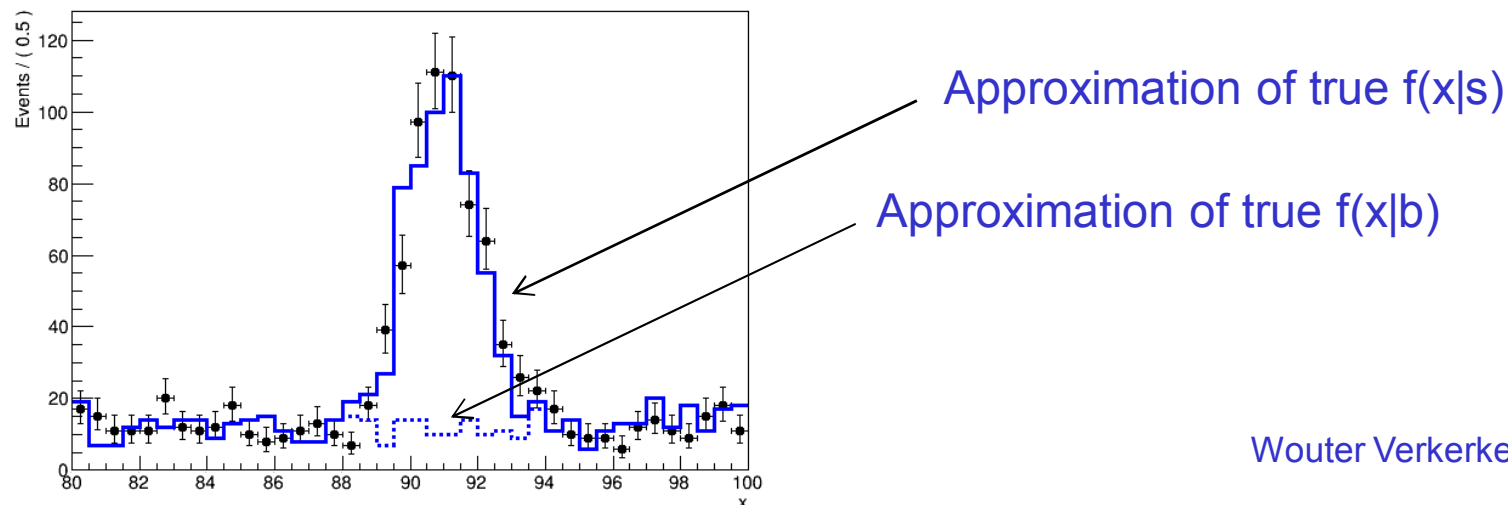
- Example of application of NP-lemma with two observables



- Cut-off value c controls type-I error rate ('size' = bkg rate)
Neyman-Pearson: LR cut gives best possible 'power' = signal eff.
- So why don't we *always* do this? (instead of training neural networks, boosted decision trees etc)

Why Neyman-Pearson doesn't always help

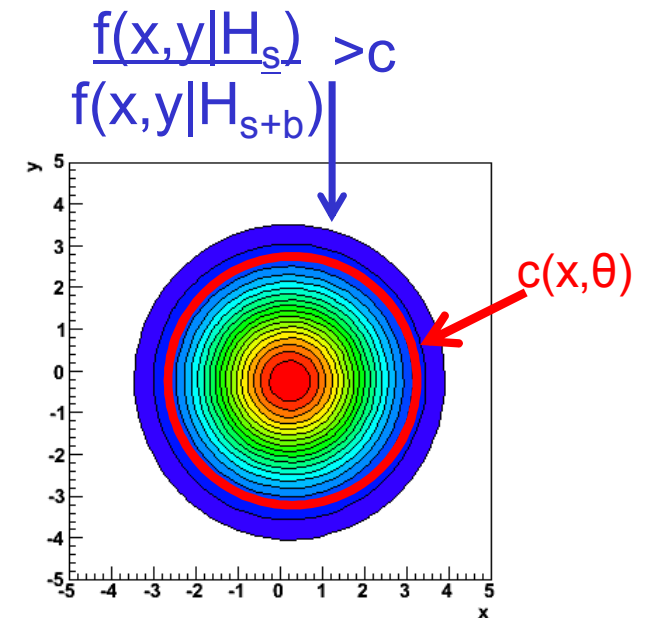
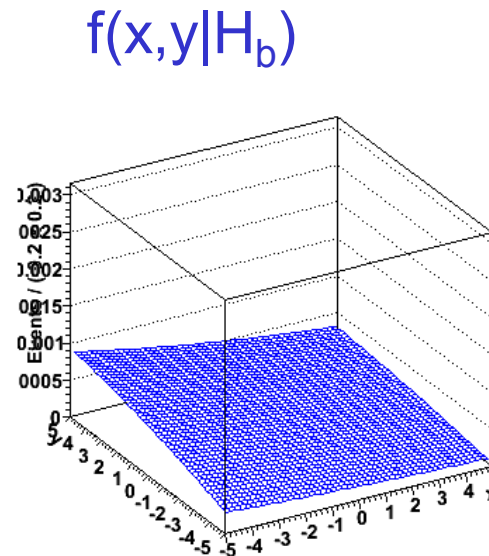
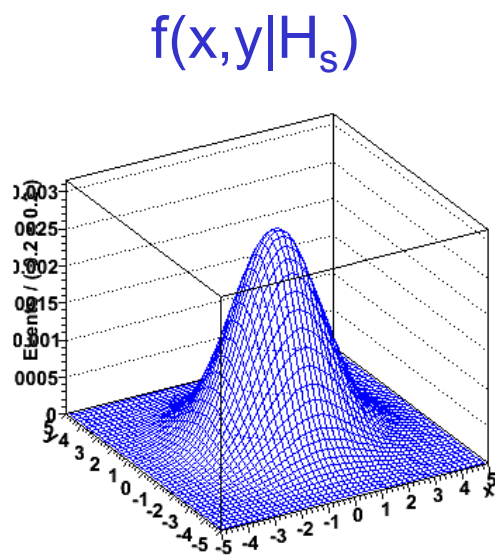
- The problem is that we usually don't have explicit formulae for the pdfs $f(\vec{x}|s)$, $f(\vec{x}|b)$.
- Instead we may have Monte Carlo samples for signal and background processes
 - Difficult to reconstruct analytical distributions of pdfs from MC samples, especially if number of dimensions is large
- If physics problem has only few observables can still estimate pdfs with histograms or kernel estimation,
 - But in such cases one can also forego event selection and go straight to hypothesis testing / parameter estimation with all events



Hypothesis testing with a large number of observables

- When number of observables is large follow different strategy
- Instead of aiming at approximating p.d.f.s $f(x|s)$ and $f(x|b)$ aim to approximate decision boundary with an empirical parametric form

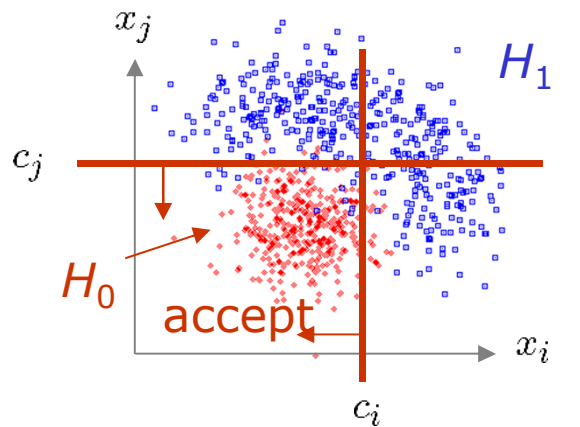
$$A_\alpha(\vec{x}) = \left[\frac{f(\vec{x} | s)}{f(\vec{x} | s + b)} > \alpha \right] \Rightarrow A_\alpha(\vec{x}) = c(\vec{x}, \vec{\theta})$$



Empirical parametric forms of decision boundaries

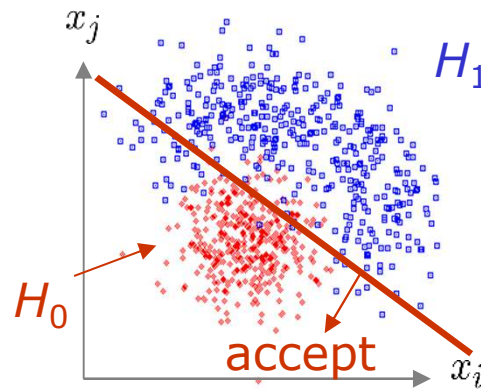
- Can in principle choose any type of Ansatz parametric shape

Rectangular cut



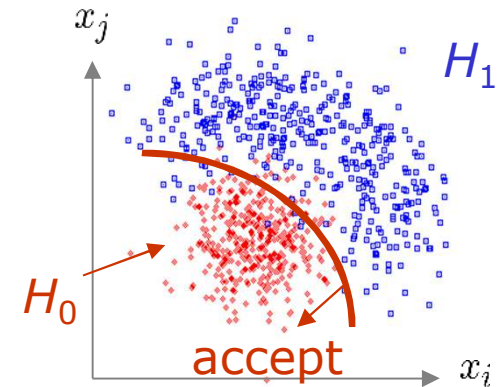
$$t(x) = \theta(x_j - c_j)\theta(x_i - c_i)$$

Linear cut



$$t(x) = a_j \cdot x_j + a_i \cdot x_i$$

Non-linear cut



$$t(x) = a \cdot x + xAx + \dots$$

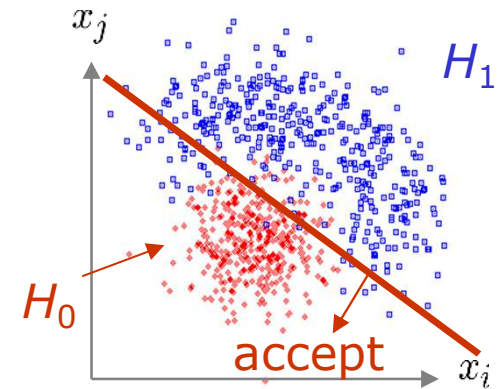
- Goal of Ansatz form is estimate of a 'signal probability' for every event in the observable space x (just like the LR)
- Choice of desired type-I error rate (selected background rate), can be set later by choosing appropriate cut on Ansatz test statistic.

The simplest Ansatz – A linear discriminant

- A **linear discriminant** constructs $t(x)$ from a linear combination of the variables x_i

$$t(\vec{x}) = \sum_{i=1}^N a_i x_i = \vec{a} \cdot \vec{x}$$

– A cut on $t(x)$ results in a linear decision plane in x -space



- What is optimal choice of direction vector a ?
- Solution provided by the **Fisher** – The Fisher discriminant

$$F(x) = (\mu_S - \mu_B)^T V^{-1} x$$

Mean values in
 x_i for sig,bkg

Inverse of variance matrix
of signal/background
(assumed to be the same)

R.A. Fisher
Ann. Eugen. 7(1936) 179.

The simplest Ansatz – A linear discriminant

- Operation advantage of Fisher discriminant is that test statistic parameters can be *calculated* (no iterative estimation is required)

$$F(x) = (\mu_S - \mu_B)^T V^{-1} x$$

R.A. Fisher
Ann. Eugen. 7(1936) 179.

- Fisher discriminant is optimal test statistic (i.e. maps to Neyman Pearson Likelihood Ratio) for case where both hypotheses are multivariate Gaussian distributions with the same variance, but different means

$$\left. \begin{aligned} f(x|s) &= \text{Gauss}(\vec{x} - \vec{\mu}_s, V) \\ f(x|b) &= \text{Gauss}(\vec{x} - \vec{\mu}_b, V) \end{aligned} \right\} \begin{array}{l} \text{Multivariate Gaussian distributions} \\ \text{with **different means** but **same width**} \\ \text{for signal and background} \end{array}$$

The simplest Ansatz – A linear discriminant

- How the Fisher discriminant follows from the LR test statistic

$$\begin{aligned}
 -\log\left(\frac{f(x|s)}{f(x|b)}\right) &= 0.5\left(\frac{x-\mu_s}{\sigma^2}\right)^2 - 0.5\left(\frac{x-\mu_b}{\sigma^2}\right)^2 + C \\
 &= 0.5\frac{x^2 - 2x\mu_s + \mu_s^2 - x^2 + 2x\mu_b - \mu_b^2}{\sigma^2} + C \\
 &\rightarrow = \frac{x(\mu_s - \mu_b)}{\sigma^2} + C'
 \end{aligned}$$

- Generalization for multidimensional Gaussian distributions

$$\log \lambda(x) = \frac{x(\mu_s - \mu_b)}{\sigma^2} + C' \xrightarrow{\sigma^2 \rightarrow V} \lambda(x) = \vec{x}(\vec{\mu}_s - \vec{\mu}_b)V^{-1} + C'$$

- Note that since we took -log of λ , **F(x) is not signal probability**, but we can trivially recover this

$$P_s(F) = \frac{1}{1 + e^{-F}}$$

If $\lambda=1$, x is equally likely under s,b
 Then $F = -\log(\lambda)=0 \Rightarrow P = 50\%$

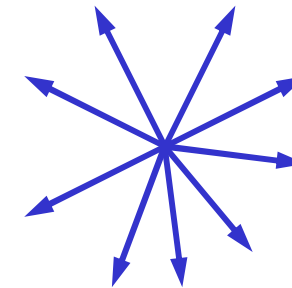
Wouter Verkerke, NIKHEF

“Logistic sigmoid function”

Example of Fisher discriminant use in HEP

- The “CLEO” Fisher discriminant

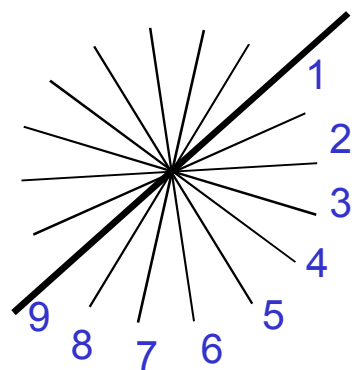
- **Goal:** distinguish between $e^+e^- \rightarrow Y4s \rightarrow b\bar{b}$ and $u\bar{u}, d\bar{d}, s\bar{s}, c\bar{c}$
- **Method:** Measure energy flow in 9 concentric cones around direction of B candidate



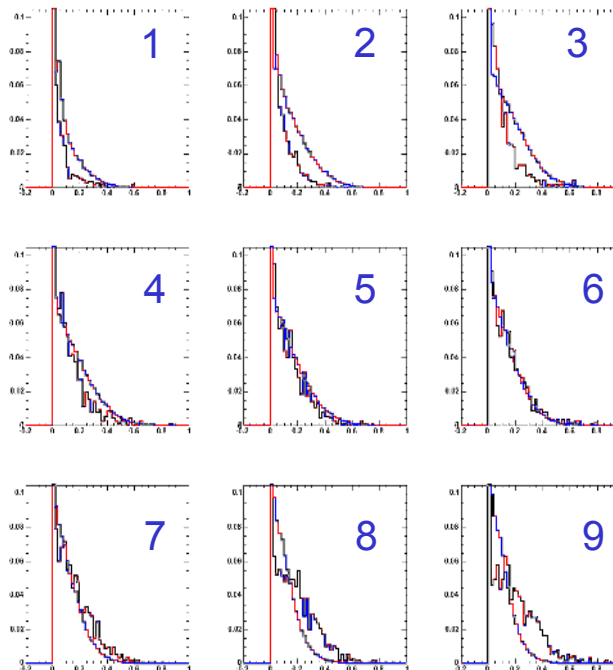
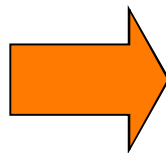
Energy flow
in $b\bar{b}$



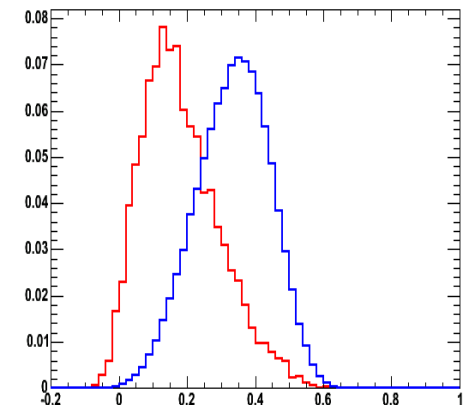
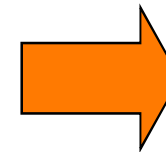
Energy flow
in u, d, s, c



Cone
Energy
flows

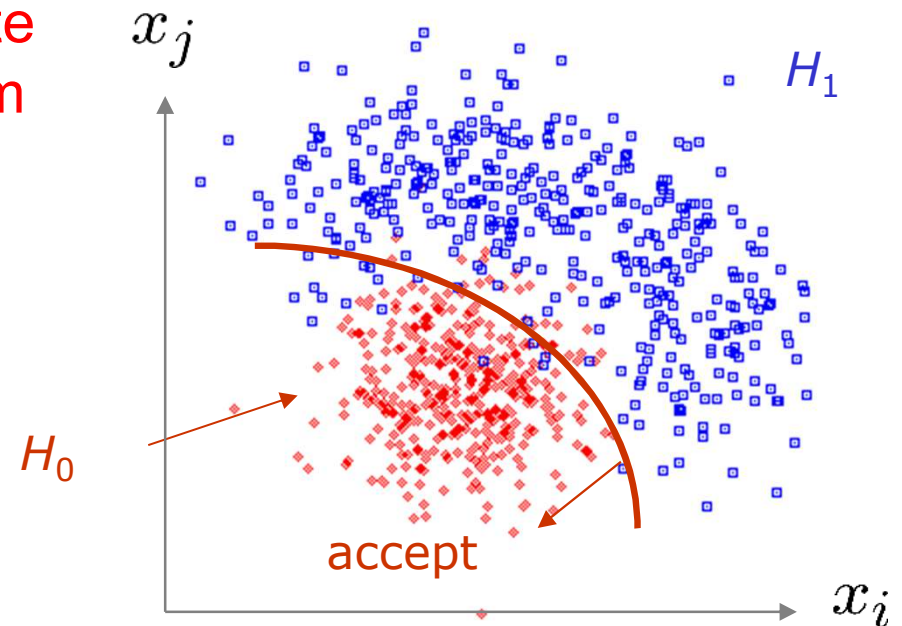


$F(x)$



Non-linear test statistics

- In most real-life HEP applications signal and background are not multi-variate Gaussian distributions with different means
- Will need more complex Ansatz shapes than Fisher discriminant
- Loose ability analytically calculate parameters of Ansatz model from Likelihood Ratio test statistic (as was done for Fisher)
- Choose an Ansatz shapes with tunable parameters
 - Artificial Neural Networks
 - Decision Trees
 - Support Vector Machines
 - Rule Ensembles
- Need numeric procedure to estimate Ansatz parameters □
Machine learning or Bayesian Learning




Machine Learning – General Principles

- Given a Ansatz parametric test statistic $T(x|\theta)$, quantify ‘risk’ due ‘loss of performance’ due to misclassifications by T as follows


Loss function (\sim log of Gaussian Likelihood)

$$R(\theta) = \int \underbrace{\left(T(\vec{x}|\theta) - 0\right)^2}_{\text{Target value of } T \text{ for background classification}} f(\vec{x}|b) d\vec{x} + \int \underbrace{\left(T(\vec{x}|\theta) - 1\right)^2}_{\text{Target value of } T \text{ for signal classification}} f(\vec{x}|s) d\vec{x}$$

 Risk function

- Practical issue: *since $f(x|s,b)$ not analytically available, cannot evaluate risk function.* Solution \square Substitute risk with ‘empirical risk’ which substitutes integral with Monte Carlo approximation

$$E(\theta) = \frac{1}{N_b} \sum_{D(x|b)} \left(T(\vec{x}_i|\theta) - 0\right)^2 + \frac{1}{N_s} \sum_{D(x|s)} \left(T(\vec{x}_i|\theta) - 1\right)^2$$

 Empirical Risk function

x_i is a set of points sampled from $f(x|b)$

x_i is a set of points sampled from $f(x|s)$

Machine Learning – General Principles

- Minimization of empirical risk $E(\theta)$ can be performed with numerical methods (many tools are available, e.g. TMVA)
- But approximation of empirical risk w.r.t analytical risk introduces possibility for ‘overtraining’:

If MC samples for signal and background are small, and number of parameters θ , one can always reduce empirical risk to zero (‘perfect selection’)

(Conceptually similar to χ^2 fit : if you fit a 10th order polynomial to 10 points – you will always perfectly describe the data. You will however not perfectly describe an independent dataset sampled from the same parent distribution)

- **Even if empirical risk is not reduced to zero by training, it may still be smaller than true risk** □ Control effect by evaluating empirical risk also on independent validation sample during minimization.

If ER on samples start to diverge, stop minimization

Bayesian Learning – General principles

- Can also applied Bayesian methodology to learning process of decision boundaries
- Given a dataset $D(x,y)$ and a Ansatz model with parameters w , aim is to estimate parameters w

$P(w)$ = posterior density on parameters of discriminant

Likelihood of the data under hypothesis w

$$P(w | \vec{x}, y) = \frac{L(\vec{x}, y | w) P(w)}{P(\vec{x}, y)}$$

$$= \frac{L(y | w, \vec{x}) L(x | w) P(w)}{\int L(y | w, \vec{x}) dw L(\vec{x})}$$

Training data
x: inputs
y: class label
(S/B) typically

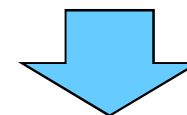
$$= \frac{L(y | w, \vec{x}) P(w)}{\int L(y | w, \vec{x}) dw L(\vec{x})}$$

$L(x|w)=1$ since
input observables
independent of model

Bayesian Learning – General principles

- Inserting a binomial likelihood function to model classification the classification problem

$$L(y | x, w) = \prod_i T(x_i, w)^y [1 - T(x_i, w)]^{1-y}$$



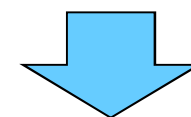
- The parameters w are thus estimated from the Bayesian posteriors densities

$$P(w | \vec{x}, y) = \frac{L(y | w, \vec{x}) P(w)}{\int L(y | w, \vec{x}) dw L(\vec{x})}$$

- No iterative minimization, but Note that integrals over ‘w-space’ can usually only be performed numerically and if w contains many parameters, this is computationally challenging
- If class of function $T(x, w)$ is large enough it will contain a function $T(x, w^*)$ that represents the true minimum in $E(w)$
 - I.e. $T(x, w^*)$ is the Bayesian equivalent of of Frequentist TS that is NP L ratio
 - In that case the test statistic is

$$T(x, w^*) = \int y L(y | x) dy$$

$$L(y | x, w) = \prod_i T(x_i, w)^y [1 - T(x_i, w)]^{1-y}$$



With $y=0,1$ only

$$= L(y = 1 | x) = \frac{L(x | y = 1) P(y = 1)}{L(x | y = 0) P(y = 0) + L(x | y = 1) P(y = 1)}$$

Machine/Bayesian learning – Non-linear Ansatz functions

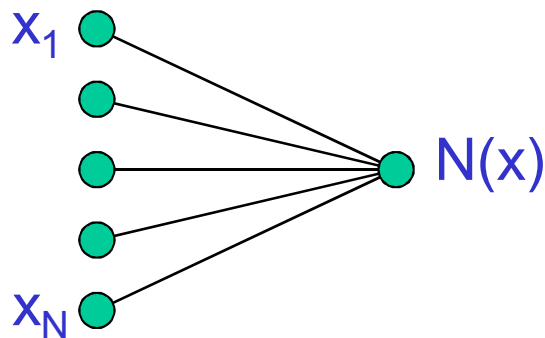
- Artificial Neural Network is one of the most popular non-linear ansatz forms. In its simplest incarnation the classifier function is

$$N(x) = s\left(a_0 + \sum_i a_i x_i\right)$$

$s(t)$ is the activation function, usually a logistic sigmoid

$$s(t) = \frac{1}{1 + e^{-t}}$$

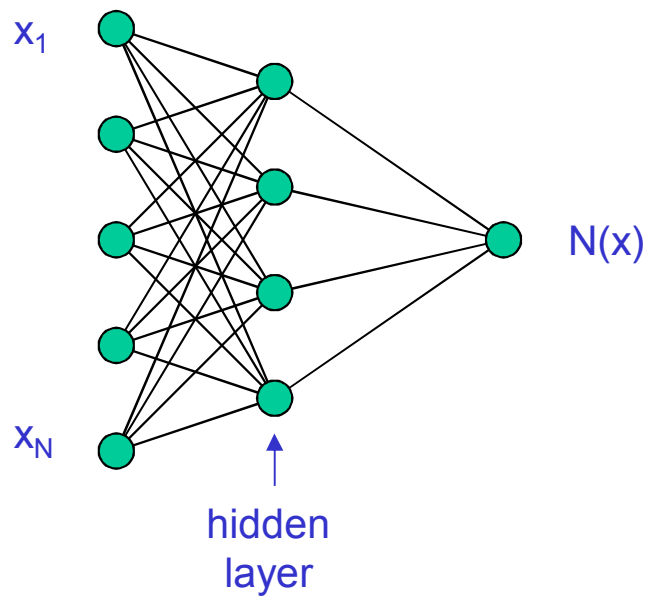
- This formula corresponds to the ‘single layer perceptron’
 - Visualization of single layer network topology



Since the activation function $s(t)$ is monotonic, a single layer $N(x)$ is equivalent to the Fisher discriminant $F(x)$

Neural networks – general structure

- The single layer model can easily be generalized to a **multilayer** perceptron



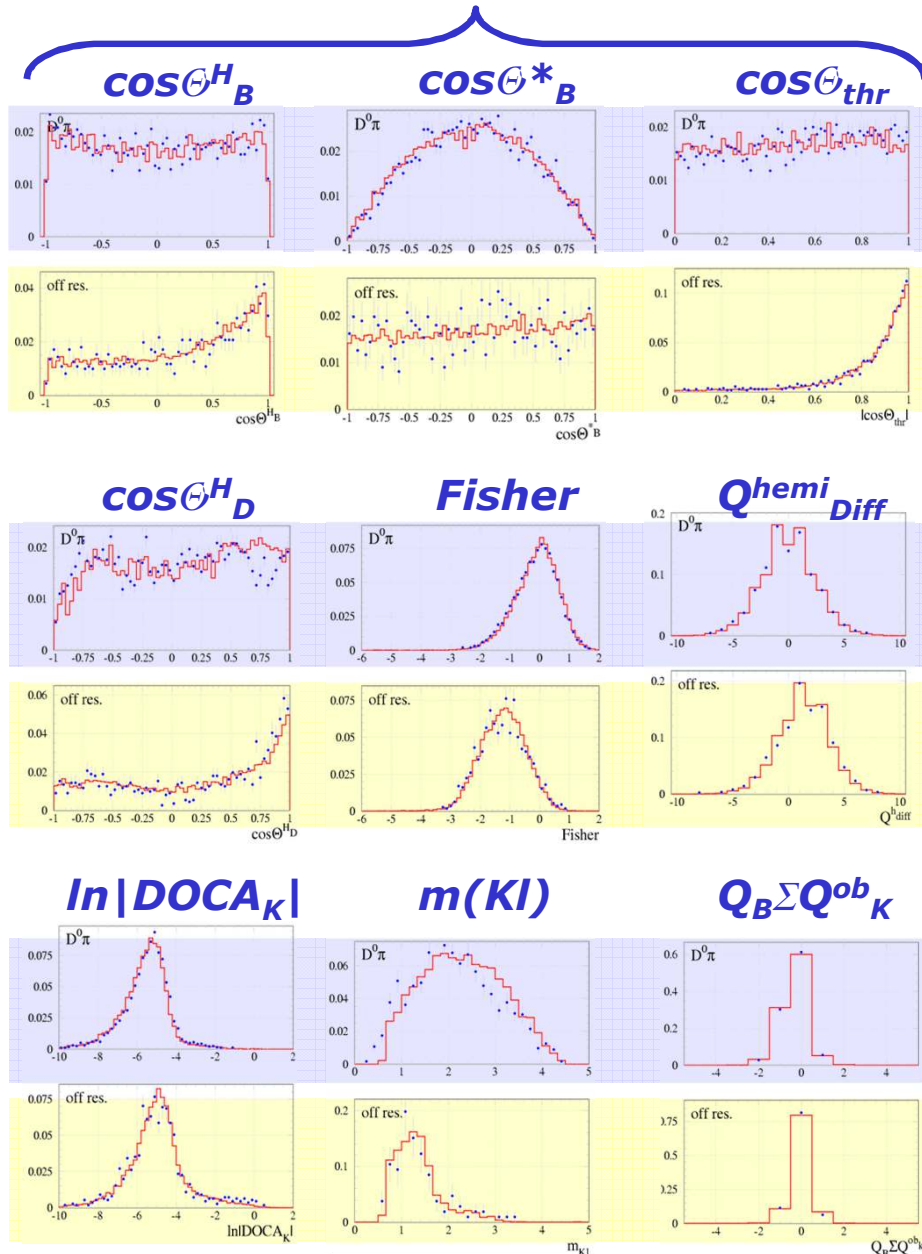
$$N(x) = s\left(a_0 + \sum_{i=1}^m a_i h_i(x)\right)$$
$$\text{with } h_i(x) = s\left(w_{i0} + \sum_{j=1}^n w_{ij} x_j\right)$$

with a_i and w_{ij} weights
(connection strengths)

- Easy to generalize to **arbitrary number of layers**
- **Feed-forward net**: values of a node depend only on earlier layers (usually only on preceding layer) 'the network architecture'
- More nodes bring $N(x)$ allow it to be closer to optimal (Neyman Pearson / Bayesian posterior) but with much more parameters to be determined

Neural networks – training example

Input Variables (9)



Signal

Background

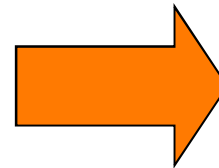
Signal

Background

Signal

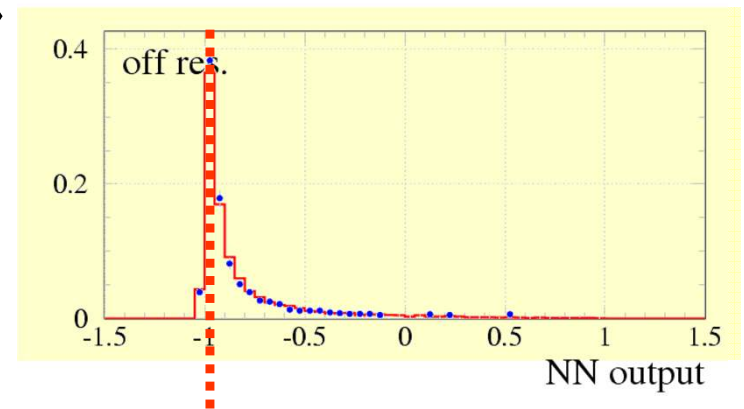
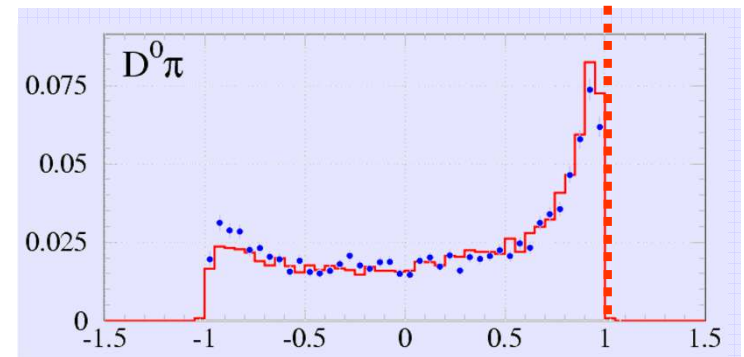
Background

$N(x)$



Output Variables (1)

Signal MC Output



Background MC Output

Wouter Verkerke, UCSB

Practical aspects of machine learning

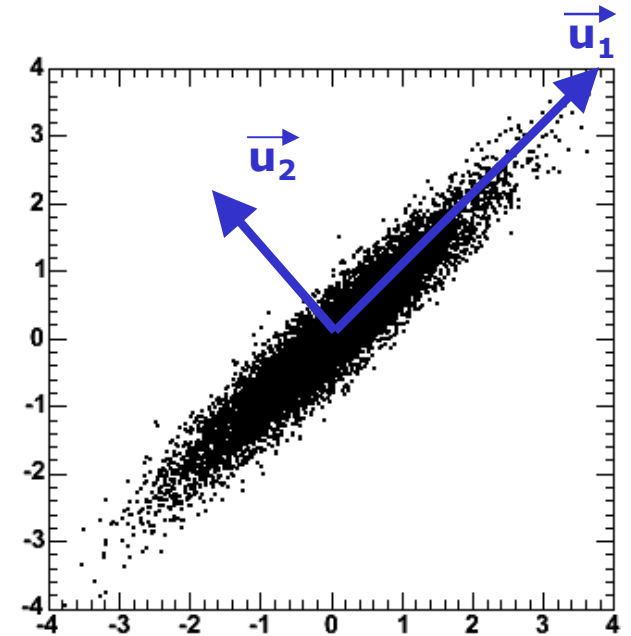
- Choose input variables sensibly
 - Don't include badly understood observables (such as #tracks/evt), variables that are not expected carry useful information
 - Generally: “Garbage in = Garbage out”
- Traditional Machine learning provides no guidance of useful complexity of test statistic (e.g. NN topology, layers)
 - Usually better to start simple and gradually increase complexity and see how that pays off
- Bayesian learning can (in principle) provide guidance on model complexity through Bayesian model selection
 - Bayes factors automatically includes a penalty for including too much model structure.

$$K = \frac{P(D | H_1)}{P(D | H_2)} = \frac{\int L(D | \theta_1, H_1) P(\theta_2 | H_1) d\theta_2}{\int L(D | \theta_2, H_2) P(\theta_2 | H_2) d\theta_2}$$

- But availability of Bayesian model selection depends in practice on the software that you use.

Practical aspects of machine learning

- Don't make the learning problem unnecessarily difficult for the machine
- E.g. remove strong correlation with explicit decorrelation before learning step
 - Can use Principle Component Analysis
 - Or Cholesky decomposition (rotate with square-root of covariance matrix)
- Also: remember that for 2-class problem (sig/bkg) that each have multivariate Gaussian distributions with different means, the optimal discriminant is can be calculated analytically
 - Fisher discriminant is analytical solution. NN solution reduces to single-layer perceptron
- Thus, you can help your machine by transforming your inputs in a form as close as possible to the Gaussian form by transforming your input observables



Gaussianization of input observables

- You can transform *any* distribution in a Gaussian distribution in two steps

- 1 – Probability integral transform

$$y(x) = \int_{-\infty}^x f(x' | H) dx'$$

"...seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years"
–Egon Pearson (1938)

turns any distribution $f(x)$ into a flat distribution in $y(x)$

- 2 – Inverse error function

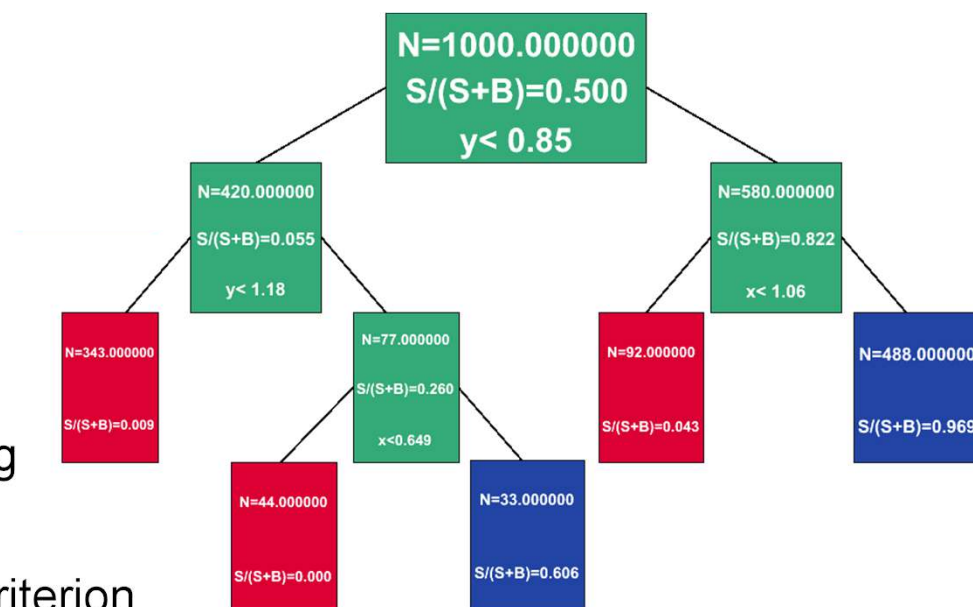
$$x^{\text{Gauss}} = \sqrt{2} \cdot \text{erf}^{-1}(2x^{\text{flat}} - 1) \quad \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

turns flat distribution into a Gaussian distribution

- Note that you can make either signal or background Gaussian, but usually not *both*

A very different type of Ansatz - Decision Trees

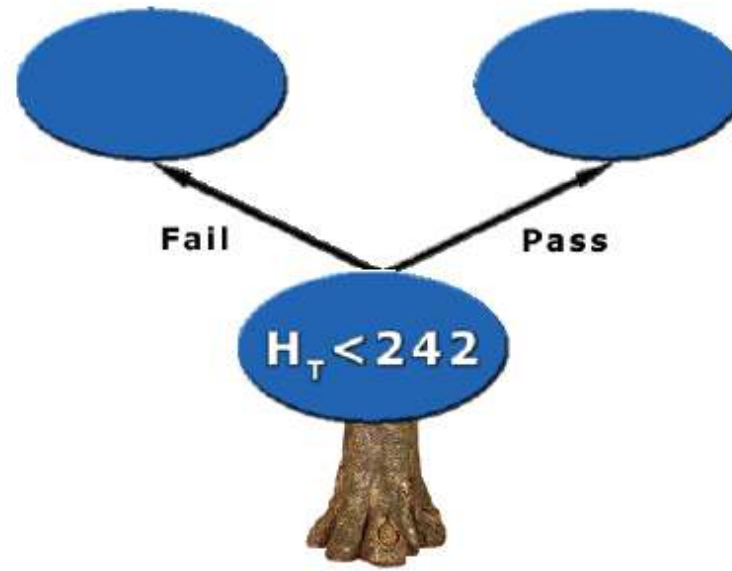
- A **Decision Tree** encodes sequential rectangular cuts
 - But with a lot of underlying theory on training and optimization
 - Machine-learning technique, widely used in social sciences
 - L. Breiman et al., “Classification and Regression Trees” (1984)



- Basic principle
 - Extend cut-based selection
 - Try not to rule out events failing a particular criterion
 - Keep events rejected by one criterion and see whether other criteria could help classify them properly

Building a tree – splitting the data

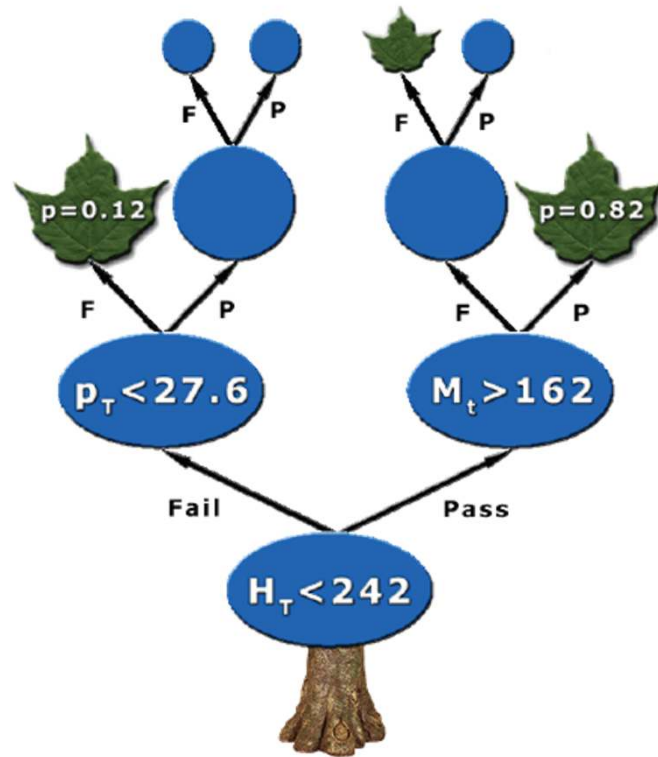
- Essential operation :
splitting the data in 2 groups using a single cut, e.g. $H_T < 242$



- Goal: find 'best cut' as quantified through **best separation of signal and background** (requires some metric to quantify this)
- Procedure:
 - 1) Find cut value with best separation for *each* observable
 - 2) Apply **only** cut on observable that results in best separation

Building a tree – recursive splitting

- Repeat splitting procedure on sub-samples of previous split



- Output of decision tree:
 - ‘signal’ or ‘background’ (0/1) or
 - probability based on *expected purity* of leaf ($s/s+b$)

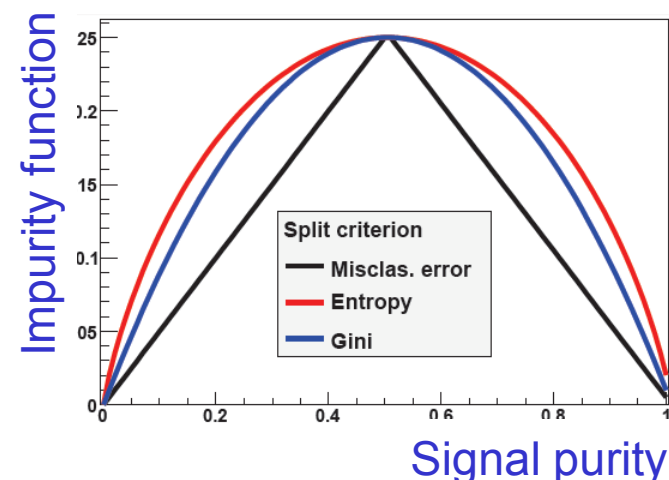
Parameters in the construction of a decision tree

- Normalization of signal and background before training
 - Usually *same total weight* for signal and background events
- In the selection of splits
 - list of questions ($var_i < cut_i$) to consider
 - Separation metric (quantifies how good the split is)
- Decision to stop splitting (declare a node terminal)
 - Minimum leaf size (e.g. 100 events)
 - Insufficient improvement from splitting
 - Perfect classification (all events in leaf belong to same class)
- Assignment of terminal node to a class
 - Usually: $purity > 0.5$ = signal, $purity < 0.5$ = background

Machine learning with Decision Trees

- Instead of 'Empirical Risk' minimize 'Impurity Function' of leaves

- Impurity function $i(t)$ quantifies (im)purity of a sample, but is not uniquely defined
- Simplest option: $i(t)$ = misclassification rate



- For a proposed split s on a node t , decrease of impurity is

$$\Delta i(s, t) = i(t) - p_L \cdot i(t_L) - p_R \cdot i(t_R)$$

Impurity
of sample
before split

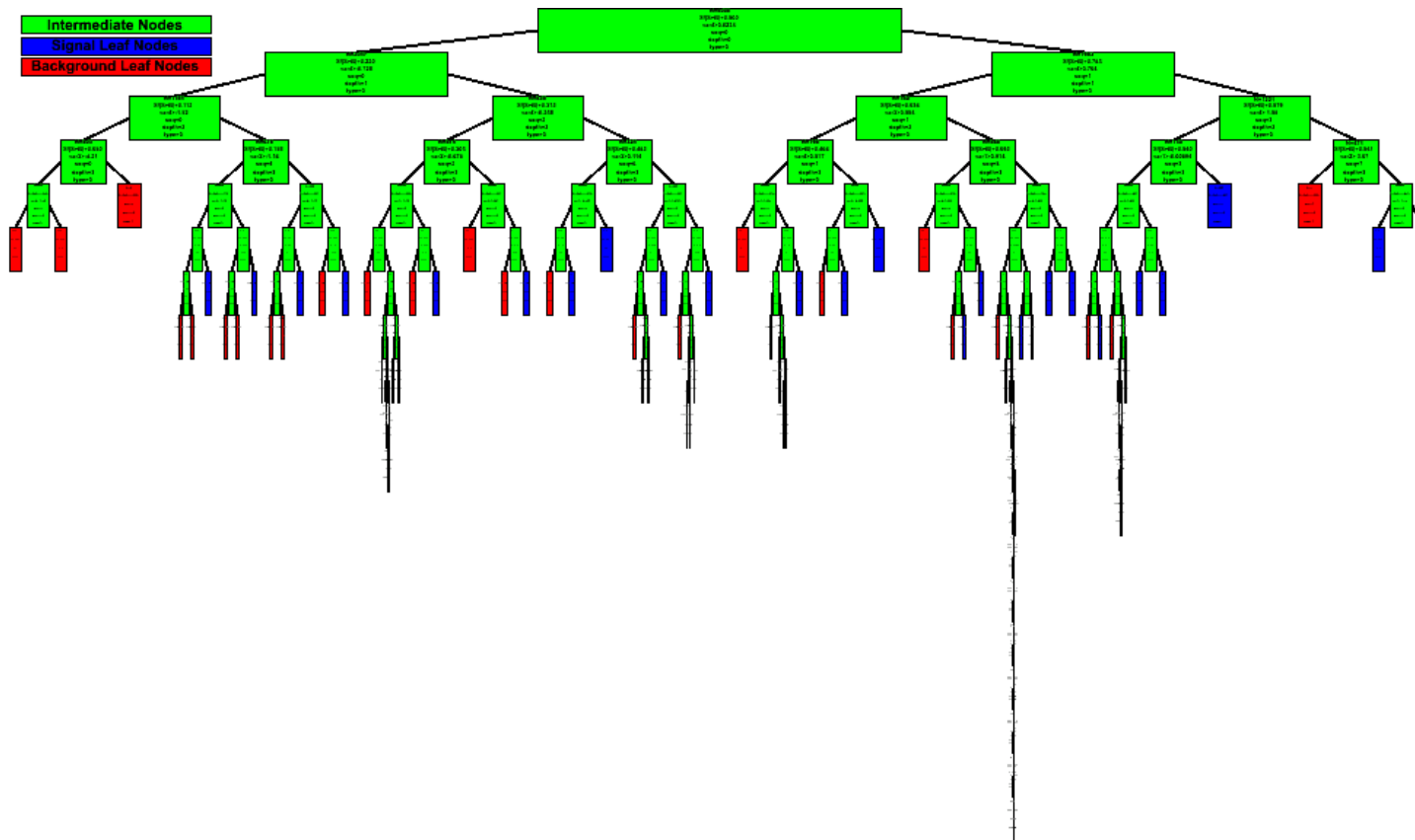
Impurity
of 'left'
sample

Impurity
of 'right'
sample

- Take split that results in largest Δi

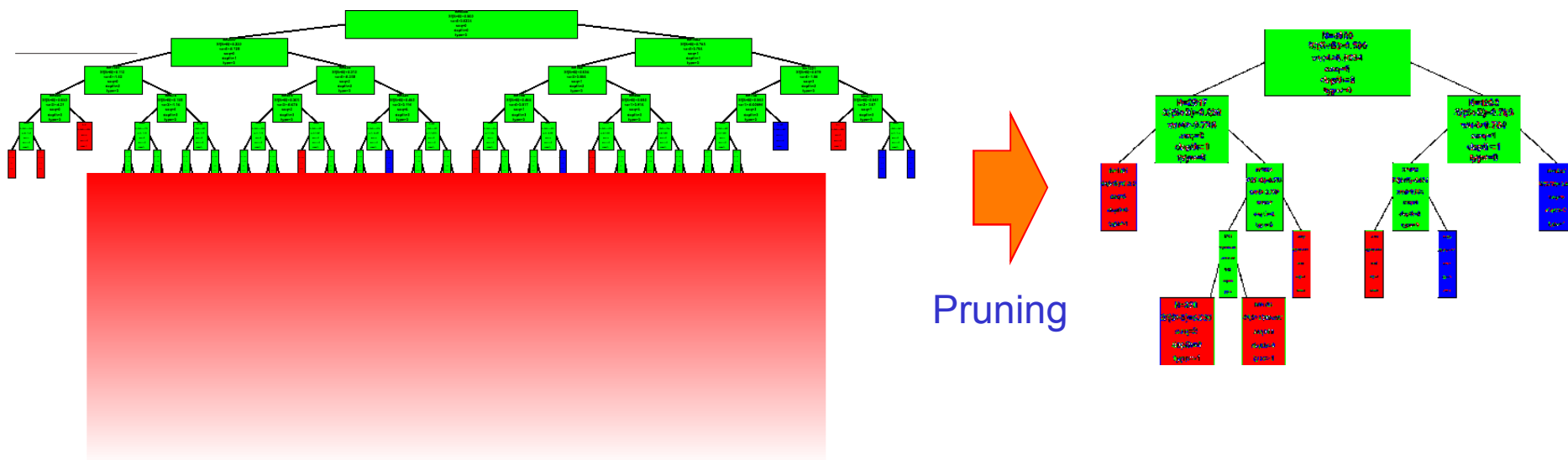
Machine learning with Decision Trees

- Stop splitting when
 - not enough improvement (introduce a cutoff Δi)
 - not enough statistics in sample, or node is pure (signal or background)
- Example decision tree from learning process



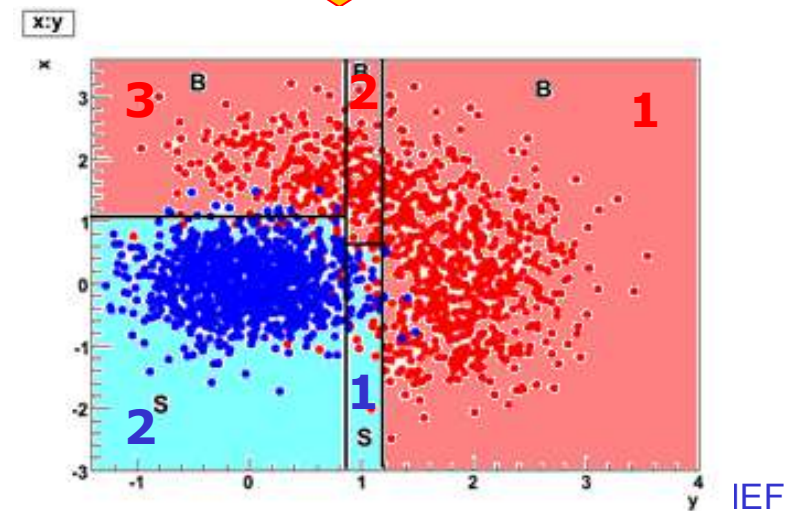
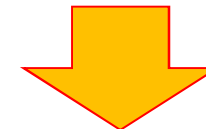
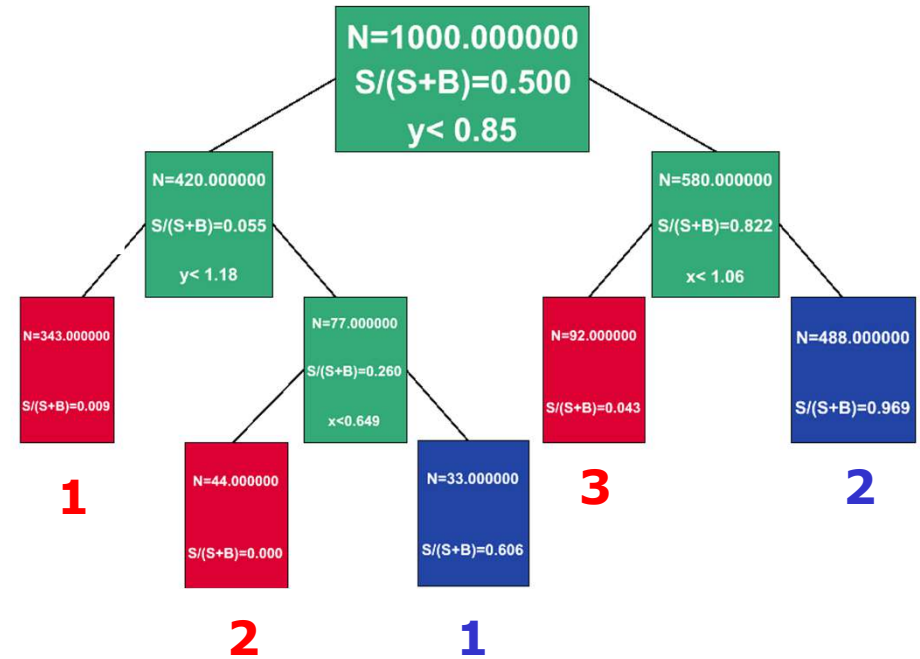
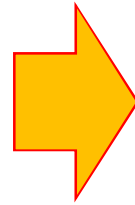
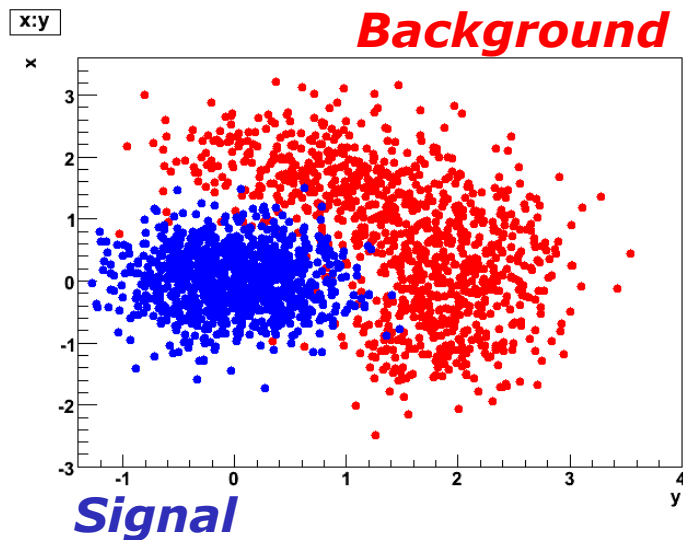
Machine learning with Decision Trees

- Given that analytical pdfs $f(x|s)$ and $f(x|b)$ are usually not available, **splitting decisions are based on 'empirical impurity'** rather than true 'impurity' \square **risk of overtraining exists**



- Can mitigate effects of **overtraining** by 'pruning' tree *a posteriori*
 - Expected error pruning (prune weak splits that are consistent with original leaf within statistical error of training sample)
 - Cost/Complexity pruning (generally strategy to trade tree complexity against performance)

Concrete example of a trained Decision Tree

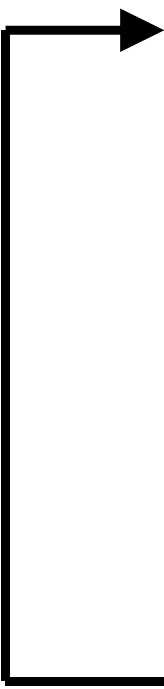


Boosted Decision trees

- Decision trees largely used with ‘boosting strategy’
- Boosting = strategy to combine multiple weaker classifiers into a single strong classifier
- First provable boosting algorithm by Schapire (1990)
 - Train classifier T_1 on N events
 - Train T_2 on new N -sample, half of which misclassified by T_1
 - Build T_3 on events where T_1 and T_2 disagree
 - **Boosted classifier**: $\text{MajorityVote}(T_1, T_2, T_3)$
- **Most used: AdaBoost** = Adaptive Boosting (Freund & Shapire ‘96)
 - Learning procedure adjusts to training data to classify it better
 - Many variations on the same theme for actual implementation

AdaBoost

- Schematic view of *iterative* algorithm

- 
- Train Decision Tree on (weighted) signal and background training samples
 - Calculate misclassification rate for Tree K (initial tree has k=1)

$$\epsilon_k = \frac{\sum_{i=1}^N w_i^k \times \text{isMisclassified}_k(i)}{\sum_{i=1}^N w_i^k}$$

“Weighted average of isMisclassified over all training events”

- Calculate weight of tree K in ‘forest decision’ $\alpha_k = \beta \times \ln((1 - \epsilon_k)/\epsilon_k)$
- **Increase weight of misclassified events** in Sample(k) to create Sample(k+1)

$$w_i^k \rightarrow w_i^{k+1} = w_i^k \times e^{\alpha_k}$$

- Boosted classifier is result is performance-weighted ‘forest’

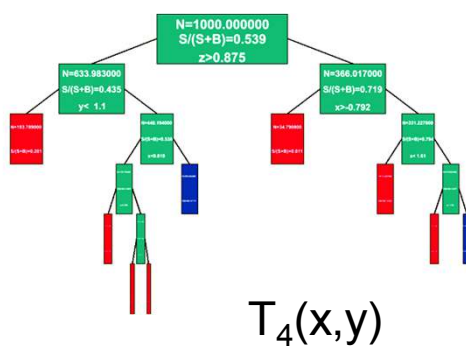
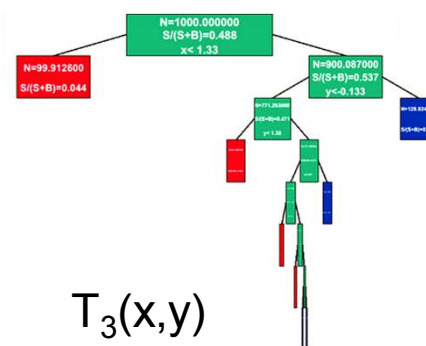
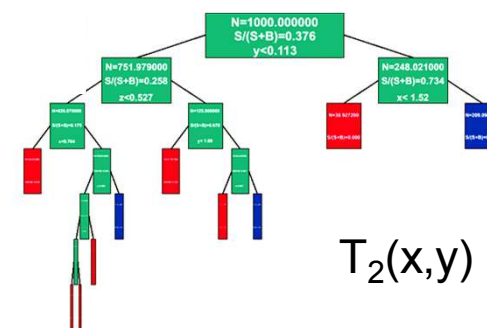
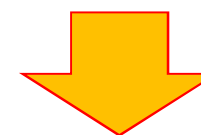
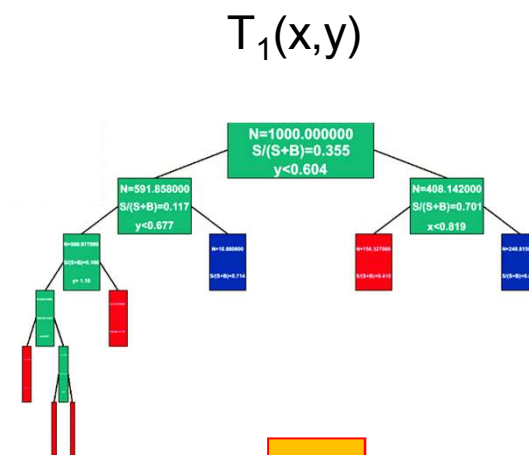
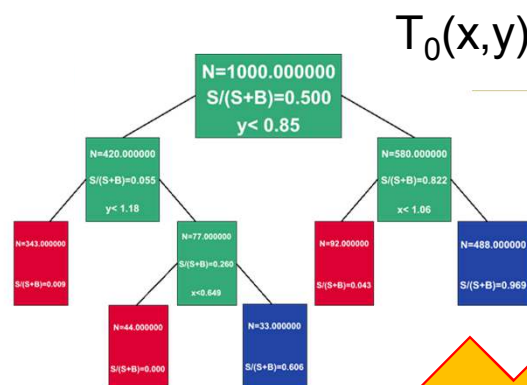
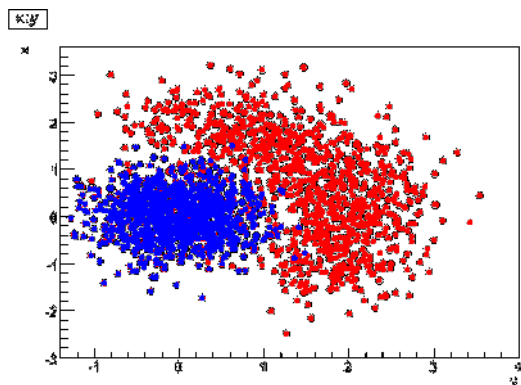
$$T(i) = \sum_{k=1}^{N_{\text{tree}}} \alpha_k T_k(i)$$

“Weighted average of Trees by their performance”

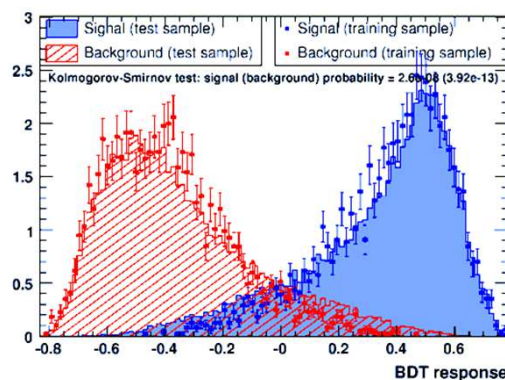
AdaBoost by example

- So-so classifier (Error rate = 40%) $\alpha = \ln \frac{1-0.4}{0.4} = 0.4$
 - Misclassified events get their weight multiplied by **exp(0.4)=1.5**
 - Next tree will have to work a bit harder on these events
- Good classifier (Error rate = 5%) $\alpha = \ln \frac{1-0.05}{0.05} = 2.9$
 - Misclassified events get their weight multiplied by **exp(2.9)=19** (!!)
 - Being failed by a good classifier means a big penalty: must be a difficult case
 - Next tree will have to pay much more attention to this event and try to get it right
- Note that boosting usually results in (strong) overtraining
 - Since with misclassification rate will ultimately go to zero

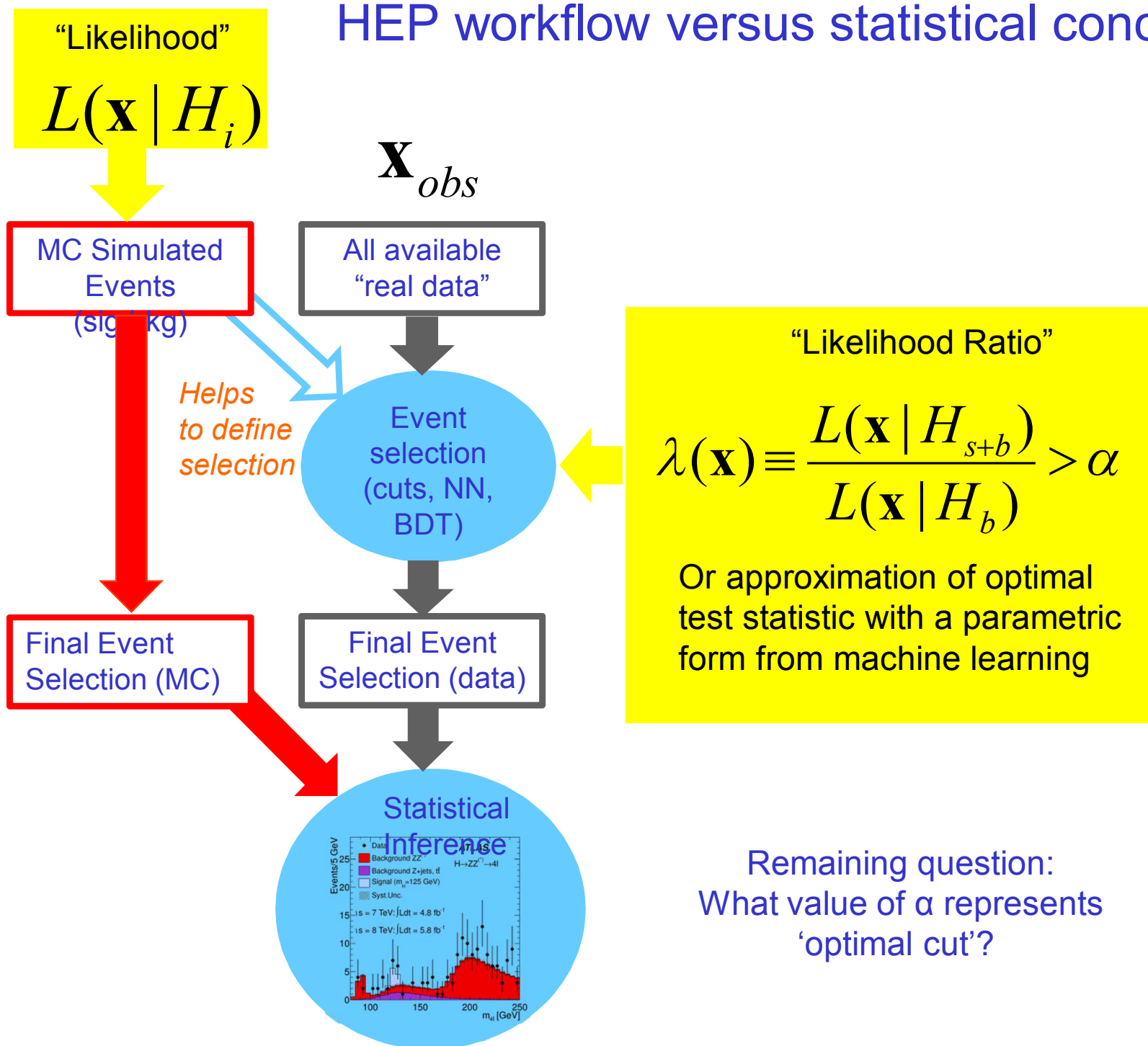
Example of Boosting



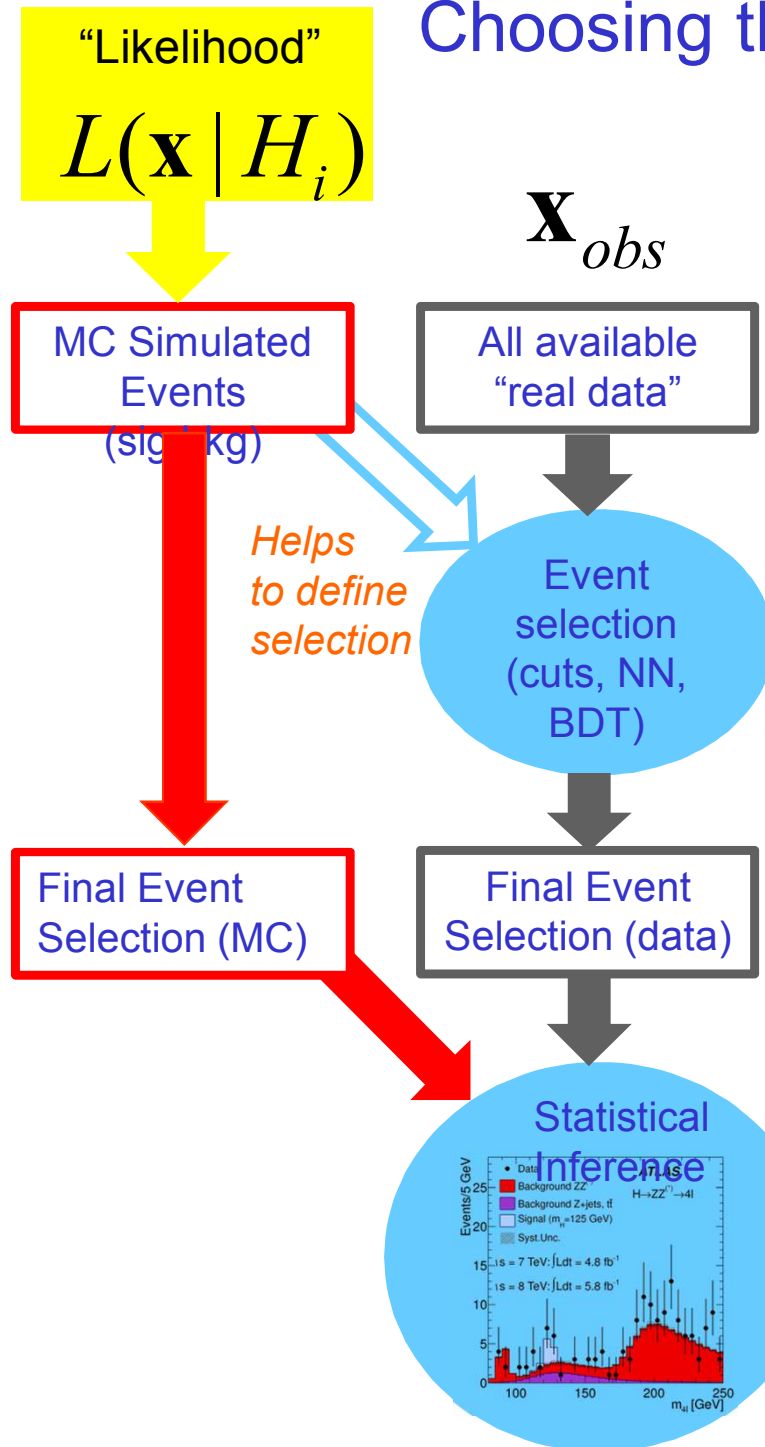
$$B(x,y) = \sum_{i=0}^4 \alpha_i T_i(x,y)$$



HEP workflow versus statistical concepts



Choosing the optimal cut on the test statistic



Note that in the limit of an optimal test statistic, and when subsequent using LR hypothesis test, *the cut on α has no influence on the statistical inference!*

□ Purely operational decision (ntuple-sizes etc...)

"Likelihood Ratio"

$$\lambda(\mathbf{x}) \equiv \frac{L(\mathbf{x} | H_{s+b})}{L(\mathbf{x} | H_b)} > \alpha$$

"p-value from Likelihood Ratio test statistic"

$$p_0(\mathbf{x} | H_i) = \int_{\lambda_{obs}}^{\infty} f(\lambda | H_i)$$

Choosing the optimal cut on the test statistic

- But reality is usually more complex:
 - Test statistics are usually not optimal,
 - Ingredients to test statistics, i.e. the event selection, are usually not perfectly known (systematic uncertainties)
- *In the subsequent statistical test phase we can account for (systematic) uncertainties in signal and background models in a detailed way. In the event selection phase we cannot*
- Pragmatically considerations in design of event selection criteria
 - Ability to estimate level of background from the selected data
 - Small sensitivity of signal acceptance to selection criteria used
- Result is that Likelihood Ratio used for event selection and final hypothesis test are different ($\lambda_{\text{selection}} \neq \lambda_{\text{hypotest}}$)
 - Cut on $\lambda_{\text{selection}}$ will influence statistical test with $\lambda_{\text{hypotest}}$
- To be able decide on optimal cut on $\lambda_{\text{selection}}$ you need a figure merit that approximates behavior of statistical test using $\lambda_{\text{hypotest}}$

Traditional approximate Figures of Merit

- Traditional choices for Figure of Merit

$$F(\alpha) = \frac{S(\alpha)}{\sqrt{B(\alpha)}}$$

‘discovery’

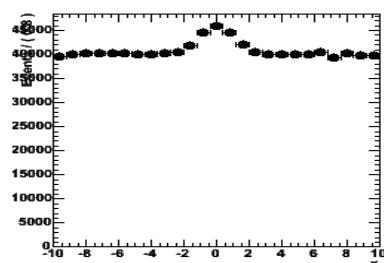
$$F(\alpha) = \frac{S(\alpha)}{\sqrt{S(\alpha) + B(\alpha)}}$$

‘measurement’

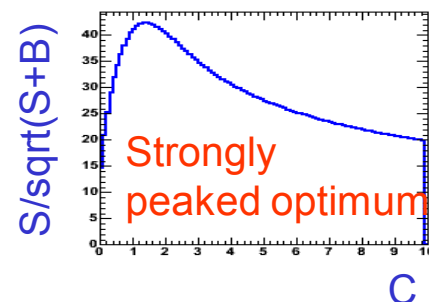
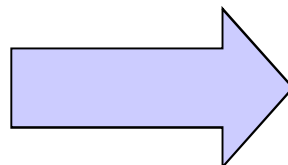
Note that position of optimum depends on a priori knowledge of signal cross section

- Note: these FOMs quantify best signal significance for a counting experiment with an known level of background, and not e.g. ‘strongest upper limit’, no accounting for systematic uncertainties

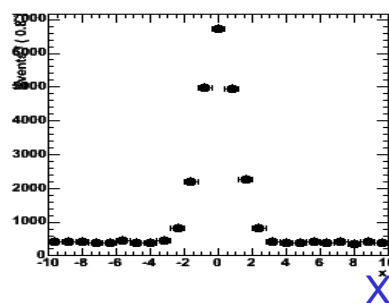
Large Bkg Scenario



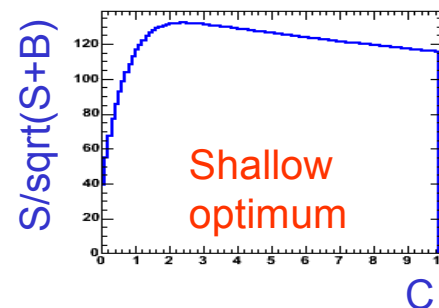
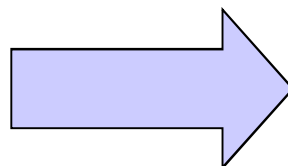
Make cut $|x| < C$



Small Bkg Scenario

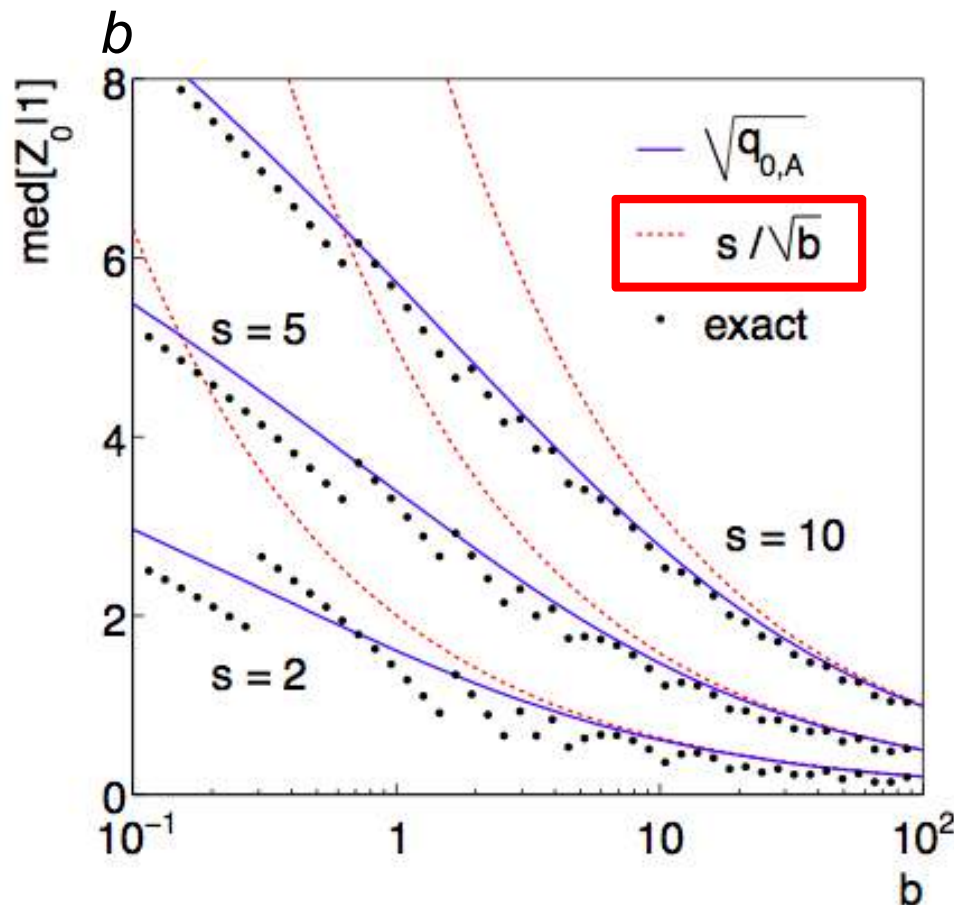


Make cut $|x| < C$



Validity of approximations in Figures of Merit

- Note that approximations made in ‘traditional’ figure of merit are not always good.
- E.g. for ‘discovery FOM’ s/\sqrt{b} illustration of approximation for $s=2,5,10$ and b in range $[0.01-100]$ shows significant deviations of s/\sqrt{b} from actual significance at low b



Improved discovery F.O.M
("Asimov Z") suggested for
situations where $s \ll b$ is not true

$$\begin{aligned}\sqrt{q_{0,A}} &= \sqrt{2((s+b)\ln(1+s/b) - s)} . \\ &= \frac{s}{\sqrt{b}} (1 + \mathcal{O}(s/b)) .\end{aligned}$$

Final comments on event selection

- Main issue with event selection is usually, sensitivity of selection criteria to systematic uncertainties
- What you'd like to avoid is your BDT/NN that is trained to get a small statistical uncertainty has a large sensitivity to a systematic uncertainties
- No easy way to incorporate effect of systematic uncertainties in training process
 - Can insert some knowledge of systematic uncertainties included in figure of merit when deciding where to cut in BDT/NN, but proper calculation usually requires much more information than signal and background event counts and is time consuming
- Use your physics intuition...

Roadmap for this course

- Tomorrow we will start with *hypothesis with parameters*

