Machine Learning and Multivariate Statistical Methods in Particle Physics

Glen Cowan
RHUL Physics
www.pp.rhul.ac.uk/~cowan

RHUL Computer Science Seminar

17 March, 2009

Outline

Quick overview of particle physics at the Large Hadron Collider (LHC)

Multivariate classification from a particle physics viewpoint

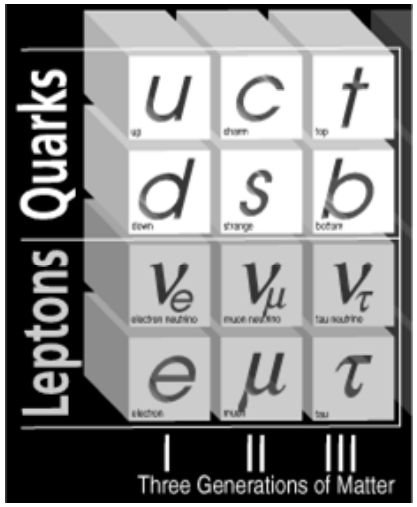
Some examples of multivariate classification in particle physics

- Neural Networks
- Boosted Decision Trees
- Support Vector Machines

Summary, conclusions, etc.

The Standard Model of particle physics

Matter...



+ gauge bosons...

photon (γ), $W^{\pm}$, Z, gluon (g)

+ relativity + quantum mechanics + symmetries... = Standard Model

25 free parameters (masses, coupling strengths,...).

Includes Higgs boson (not yet seen).

Almost certainly incomplete (e.g. no gravity).

Agrees with all experimental observations so far.

Many candidate extensions to SM (supersymmetry, extra dimensions,...)

The Large Hadron Collider

Counter-rotating proton beams
in 27 km circumference ring

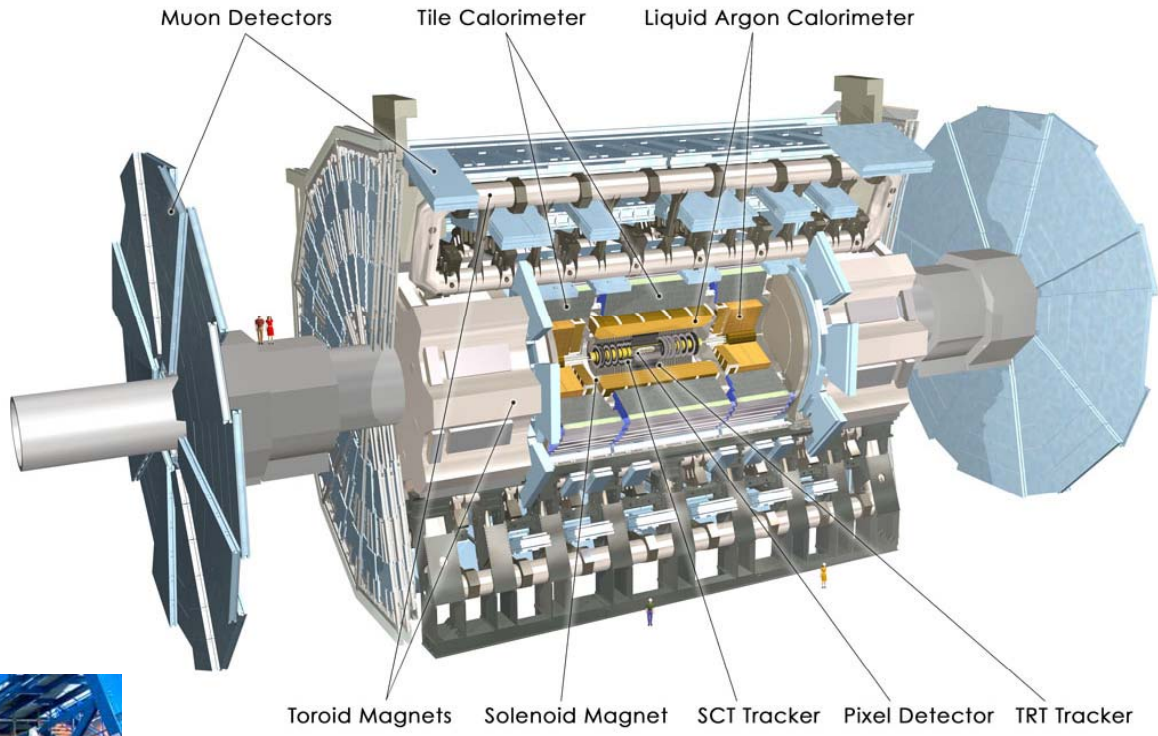
pp centre-of-mass energy 14 TeV

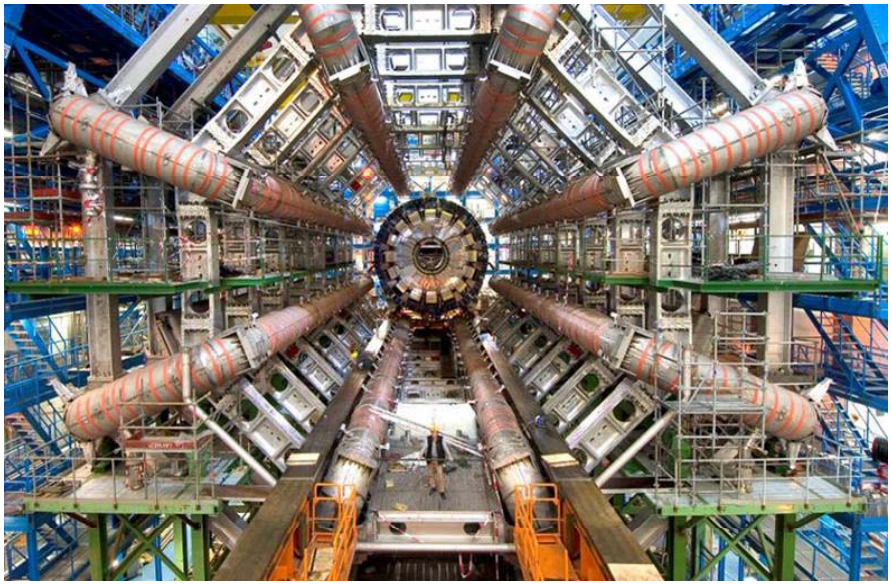
Detectors at 4 pp collision points:

ATLAS
CMS ← general purpose
LHCb ← (b physics)
ALICE (heavy ion physics)

The ATLAS detector

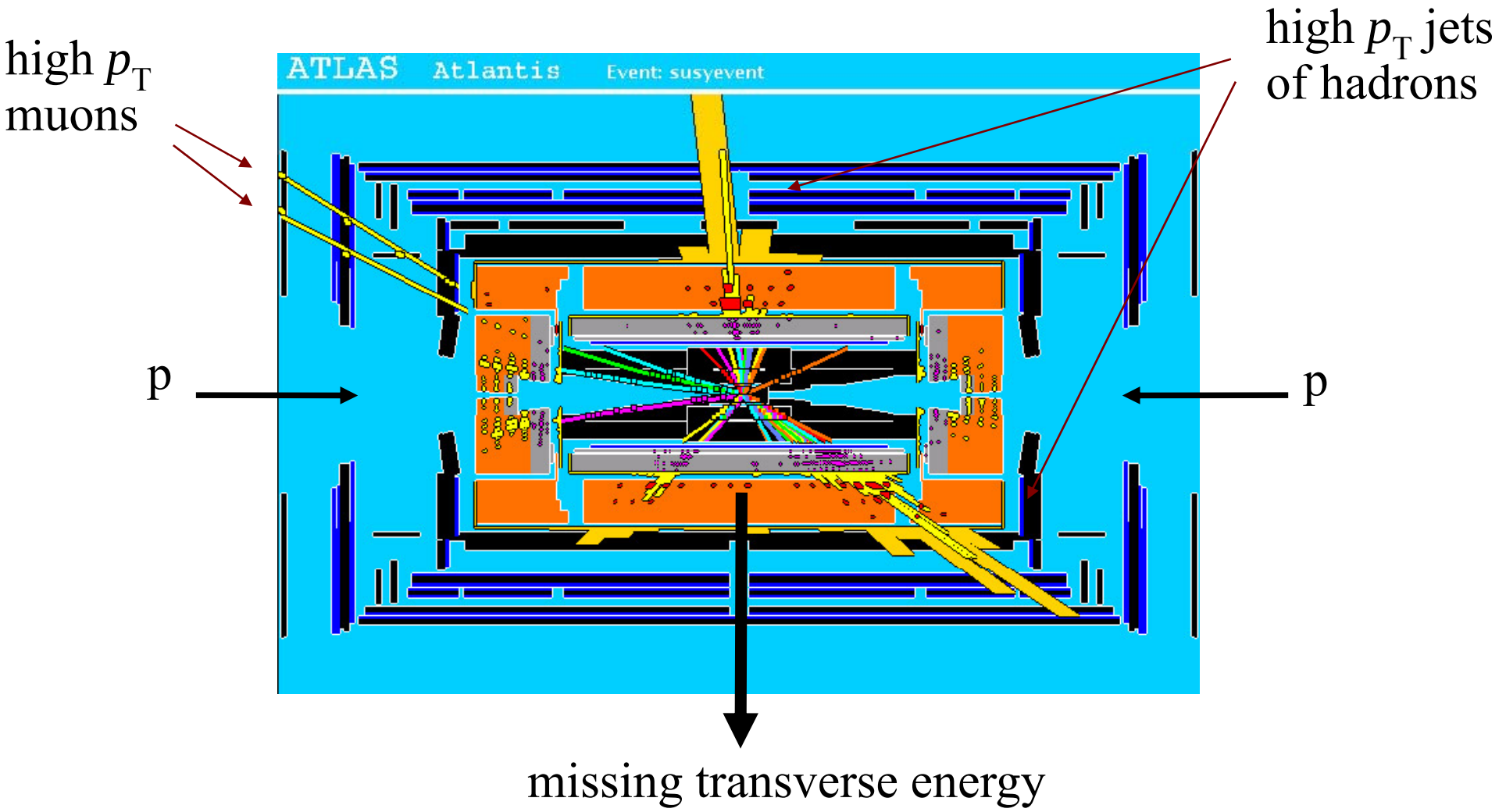
2100 physicists
37 countries
167 universities/labs



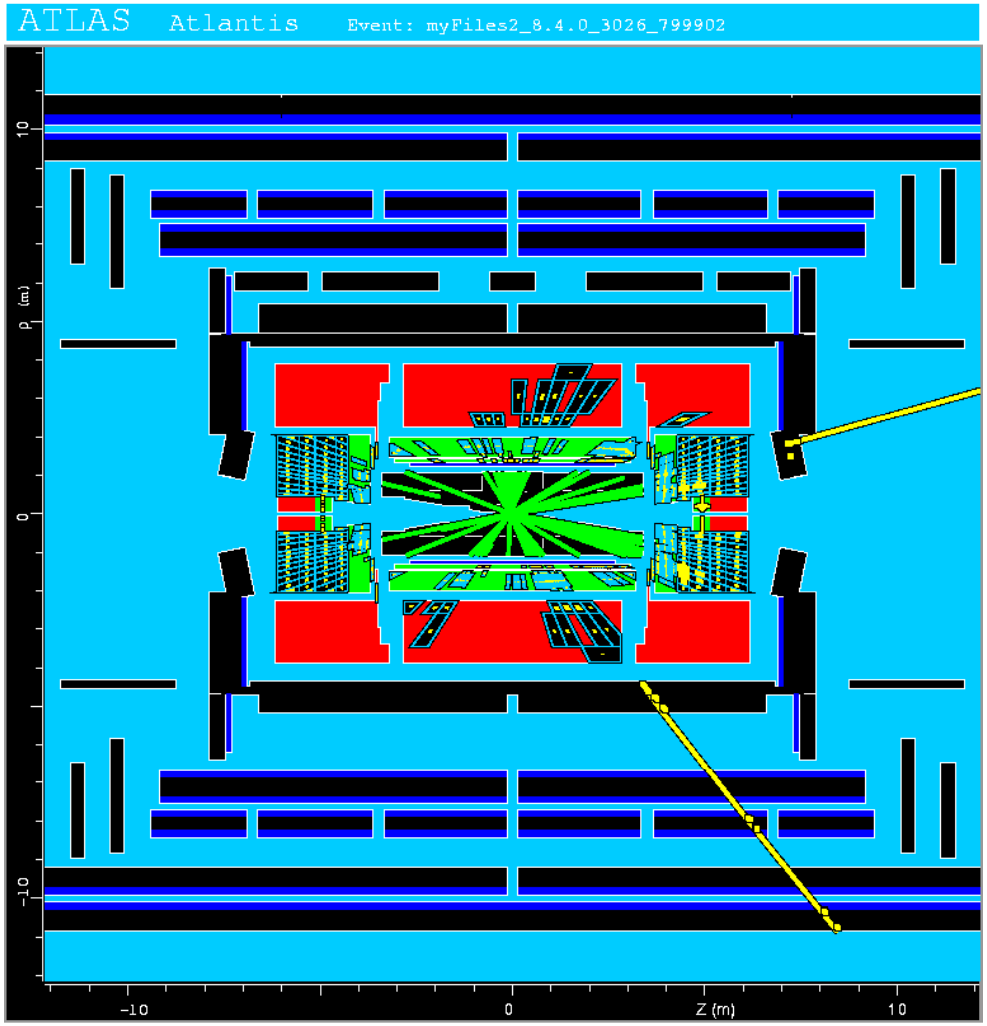


25 m diameter
46 m length
7000 tonnes
~10^8 electronic channels

A simulated SUSY event in ATLAS



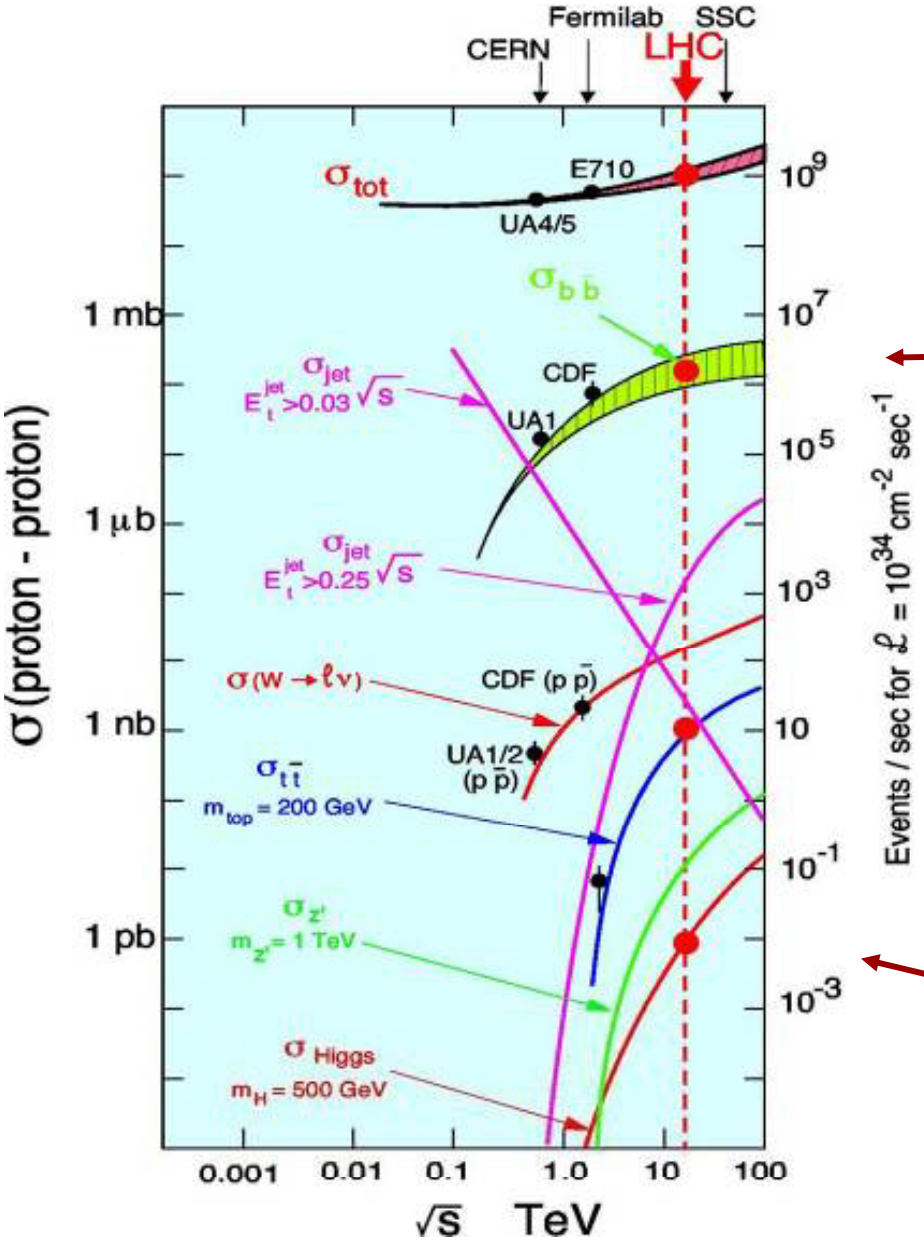
Background events



This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

LHC event production rates



most events (boring)

mildly interesting

interesting

very interesting
(~1 out of every 10^{11})

LHC data

At LHC, ~10^9 pp collision events per second, mostly uninteresting

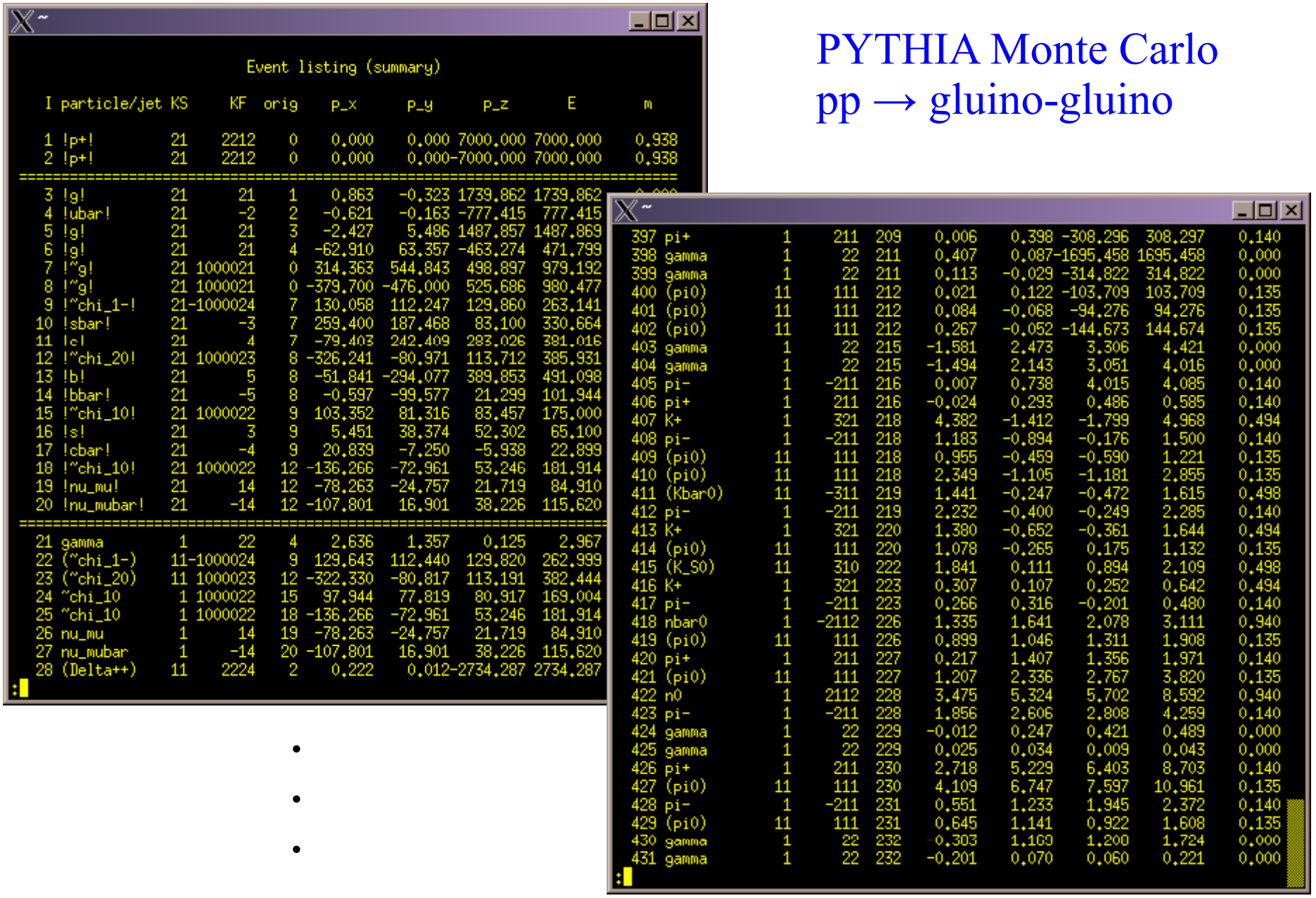
do quick sifting, record ~200 events/sec
single event ~ 1 Mbyte
1 "year" ≈ 10^7 s, 10^{16} pp collisions / year
2×10^9 events recorded / year (~2 Pbyte / year)

For new/rare processes, rates at LHC can be vanishingly small

e.g. Higgs bosons detectable per year could be ~10^3
→ 'needle in a haystack'

For Standard Model and (many) non-SM processes we can generate simulated data with Monte Carlo programs (including simulation of the detector).

A simulated event



Event listing (summary)

I	particle/jet	KS	KF	orig	p_x	p_y	p_z	E	m
1	!p+!	21	2212	0	0.000	0.000	7000.000	7000.000	0.938
2	!p+!	21	2212	0	0.000	0.000	-7000.000	7000.000	0.938
3	!g!	21	21	1	0.863	-0.323	1739.862	1739.862	
4	!ubar!	21	-2	2	-0.621	-0.163	-777.415	777.415	
5	!g!	21	21	3	-2.427	5.486	1487.857	1487.869	
6	!g!	21	21	4	-62.910	63.357	-463.274	471.799	
7	!~g!	21	1000021	0	314.363	544.843	498.897	979.192	
8	!~g!	21	1000021	0	-379.700	-476.000	525.686	980.477	
9	!~chi_1-!	21	-1000024	7	130.058	112.247	129.860	263.141	
10	!sbar!	21	-3	7	259.400	187.468	83.100	330.664	
11	!c!	21	4	7	-79.403	242.409	283.026	381.016	
12	!~chi_20!	21	1000023	8	-326.241	-80.971	113.712	385.931	
13	!b!	21	5	8	-51.841	-294.077	389.853	491.098	
14	!bbar!	21	-5	8	-0.597	-99.577	21.299	101.944	
15	!~chi_10!	21	1000022	9	103.352	81.316	83.457	175.000	
16	!s!	21	3	9	5.451	38.374	52.302	65.100	
17	!cbar!	21	-4	9	20.839	-7.250	-5.938	22.899	
18	!~chi_10!	21	1000022	12	-136.266	-72.961	53.246	181.914	
19	!nu_mu!	21	14	12	-78.263	-24.757	21.719	84.910	
20	!nu_mubar!	21	-14	12	-107.801	16.901	38.226	115.620	
21	gamma	1	22	4	2.636	1.357	0.125	2.967	
22	(~chi_1-)	11	-1000024	9	129.643	112.440	129.820	262.999	
23	(~chi_20)	11	1000023	12	-322.330	-80.817	113.191	382.444	
24	~chi_10	1	1000022	15	97.944	77.819	80.917	169.004	
25	~chi_10	1	1000022	18	-136.266	-72.961	53.246	181.914	
26	nu_mu	1	14	19	-78.263	-24.757	21.719	84.910	
27	nu_mubar	1	-14	20	-107.801	16.901	38.226	115.620	
28	(Delta++)	11	2224	2	0.222	0.012	-2734.287	2734.287	

397	pi+	1	211	209	0.006	0.398	-308.296	308.297	0.140
398	gamma	1	22	211	0.407	0.087	-1695.458	1695.458	0.000
399	gamma	1	22	211	0.113	-0.029	-314.822	314.822	0.000
400	(pi0)	11	111	212	0.021	0.122	-103.709	103.709	0.135
401	(pi0)	11	111	212	0.084	-0.068	-94.276	94.276	0.135
402	(pi0)	11	111	212	0.267	-0.052	-144.673	144.674	0.135
403	gamma	1	22	215	-1.581	2.473	3.306	4.421	0.000
404	gamma	1	22	215	-1.494	2.143	3.051	4.016	0.000
405	pi-	1	-211	216	0.007	0.738	4.015	4.085	0.140
406	pi+	1	211	216	-0.024	0.293	0.486	0.585	0.140
407	K+	1	321	218	4.382	-1.412	-1.799	4.968	0.494
408	pi-	1	-211	218	1.183	-0.894	-0.176	1.500	0.140
409	(pi0)	11	111	218	0.955	-0.459	-0.590	1.221	0.135
410	(pi0)	11	111	218	2.349	-1.105	-1.181	2.855	0.135
411	(Kbar0)	11	-311	219	1.441	-0.247	-0.472	1.615	0.498
412	pi-	1	-211	219	2.232	-0.400	-0.249	2.285	0.140
413	K+	1	321	220	1.380	-0.652	-0.361	1.644	0.494
414	(pi0)	11	111	220	1.078	-0.265	0.175	1.132	0.135
415	(K_S0)	11	310	222	1.841	0.111	0.894	2.109	0.498
416	K+	1	321	223	0.307	0.107	0.252	0.642	0.494
417	pi-	1	-211	223	0.266	0.316	-0.201	0.480	0.140
418	nbar0	1	-2112	226	1.335	1.641	2.078	3.111	0.940
419	(pi0)	11	111	226	0.899	1.046	1.311	1.908	0.135
420	pi+	1	211	227	0.217	1.407	1.356	1.971	0.140
421	(pi0)	11	111	227	1.207	2.336	2.767	3.820	0.135
422	n0	1	2112	228	3.475	5.324	5.702	8.592	0.940
423	pi-	1	-211	228	1.856	2.606	2.808	4.259	0.140
424	gamma	1	22	229	-0.012	0.247	0.421	0.489	0.000
425	gamma	1	22	229	0.025	0.034	0.009	0.043	0.000
426	pi+	1	211	230	2.718	5.229	6.403	8.703	0.140
427	(pi0)	11	111	230	4.109	6.747	7.597	10.961	0.135
428	pi-	1	-211	231	0.551	1.233	1.945	2.372	0.140
429	(pi0)	11	111	231	0.645	1.141	0.922	1.608	0.135
430	gamma	1	22	232	-0.303	1.169	1.200	1.724	0.000
431	gamma	1	22	232	-0.201	0.070	0.060	0.221	0.000

PYTHIA Monte Carlo
pp → gluino-gluino

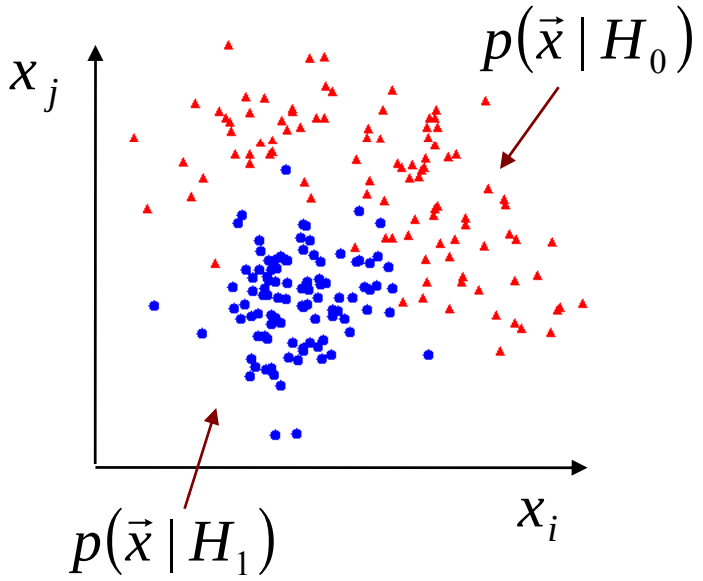
-
-
-

Multivariate analysis in particle physics

For each event we measure a set of numbers: $\vec{x} = (x_1, \ldots, x_n)$

x_1 = jet p_T
x_2 = missing energy
x_3 = particle i.d. measure, ...

$\vec{x}$ follows some n-dimensional joint probability density, which depends on the type of event produced, i.e., was it $pp \to t\bar{t}$, $pp \to \tilde{g}\tilde{g}$, ...

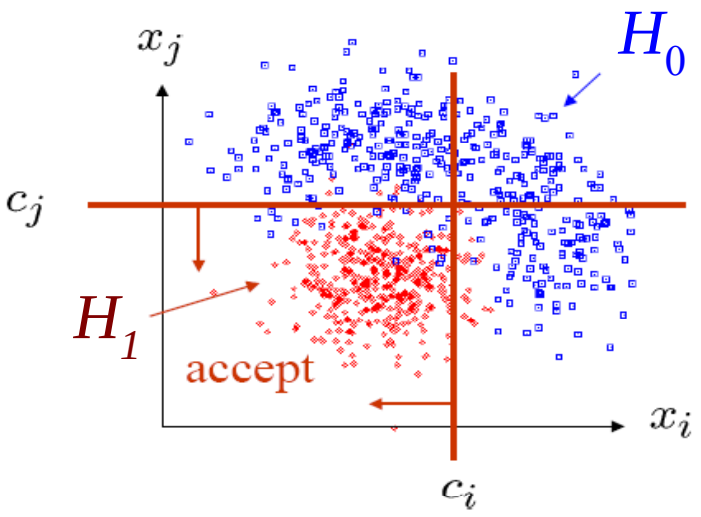


E.g. hypotheses H_0, H_1, ...
Often simply "signal", "background"

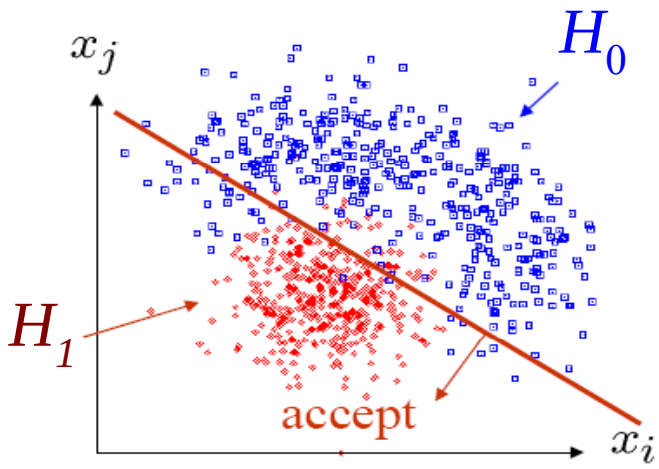
Finding an optimal decision boundary

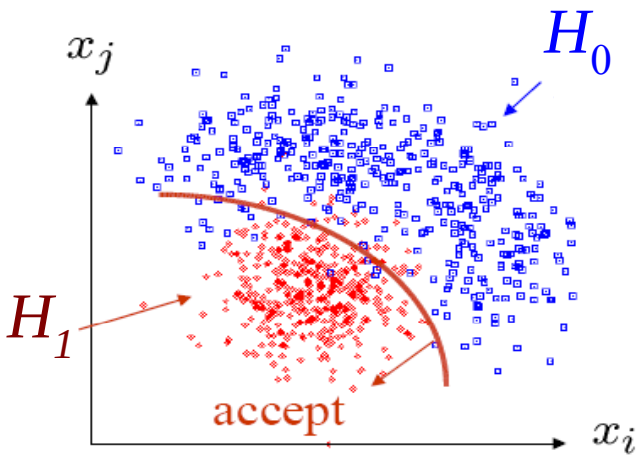
In particle physics usually start by making simple "cuts":

$$x_i < c_i$$
$$x_j < c_j$$



Maybe later try some other type of decision boundary:





The optimal decision boundary

Try to best approximate optimal decision boundary based on likelihood ratio:

$$y(\mathbf{x}) = \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} = \text{const.}$$

or equivalently think of the likelihood ratio as the optimal statistic for a test of H_0 vs H_1.

In general we don't have the pdfs $p(\boldsymbol{x}|H_0)$, $p(\boldsymbol{x}|H_1)$,...
Rather, we have Monte Carlo models for each process.

Usually training data from the MC models is cheap.

But the models contain many approximations:
predictions for observables obtained using perturbation theory (truncated at some order); phenomenological modeling of non-perturbative effects; imperfect detector description,...

Two distinct event selection problems

In some cases, the event types in question are both known to exist.

Example: separation of different particle types (electron vs muon)
Use the selected sample for further study.

In other cases, the null hypothesis H_0 means "Standard Model" events, and the alternative H_1 means "events of a type whose existence is not yet established" (to do so is the goal of the analysis).

Many subtle issues here, mainly related to the heavy burden of proof required to establish presence of a new phenomenon.

Typically require p-value of background-only hypothesis below $\sim 10^{-7}$ (a 5 sigma effect) to claim discovery of "New Physics".

Discovering "New Physics"

The LHC experiments are expensive

~ 10^{10} (accelerator and experiments)

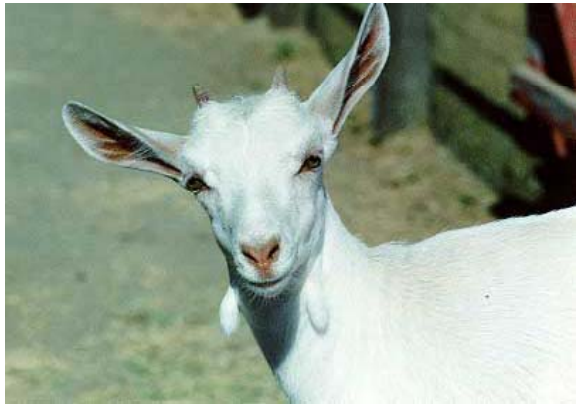
the competition is intense

(ATLAS vs. CMS) vs. Tevatron

and the stakes are high:

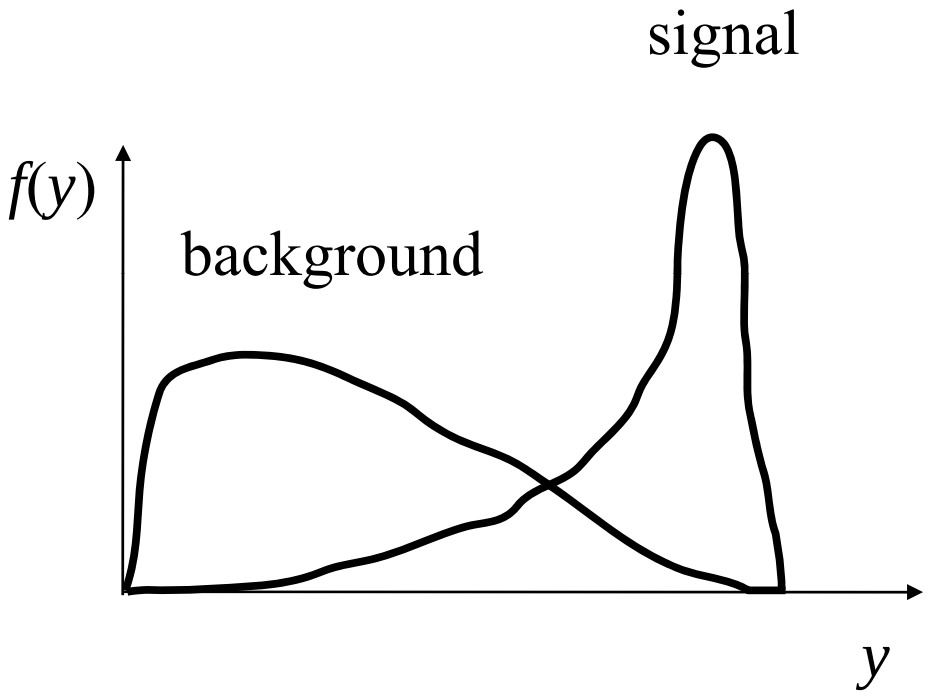
4 sigma effect

5 sigma effect

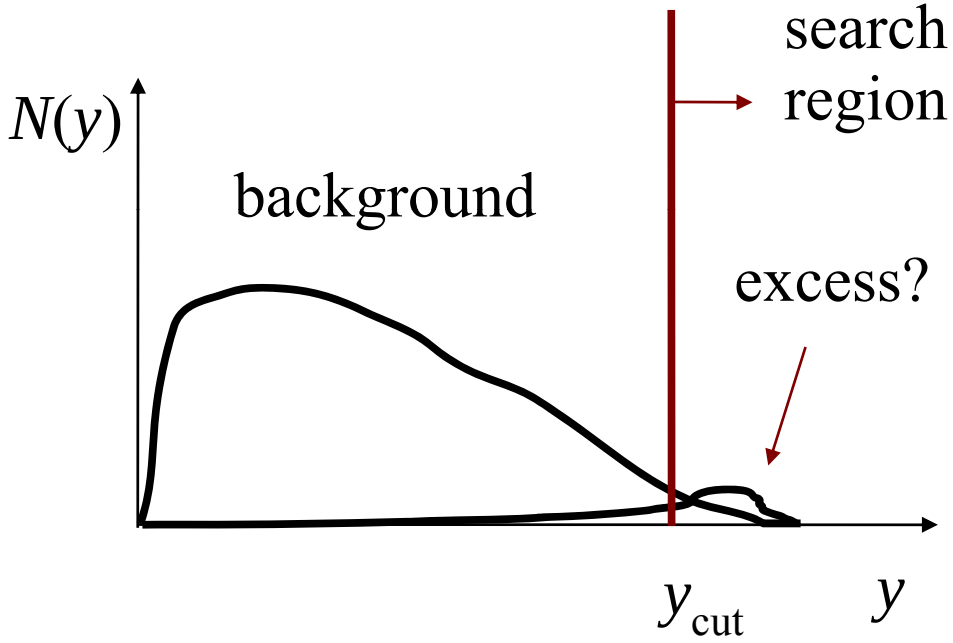


So there is a strong motivation to extract all possible information from the data.

Using classifier output for discovery



Normalized to unity



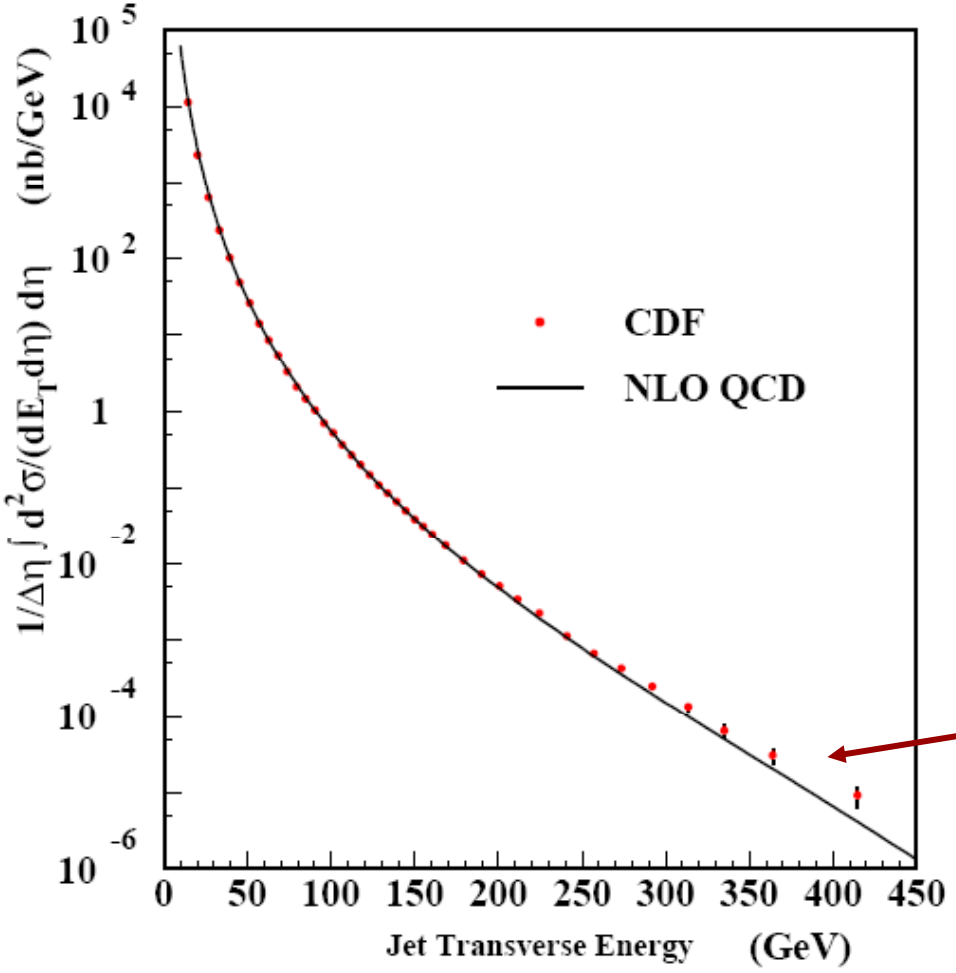
Normalized to expected number of events

Discovery = number of events found in search region incompatible with background-only hypothesis.

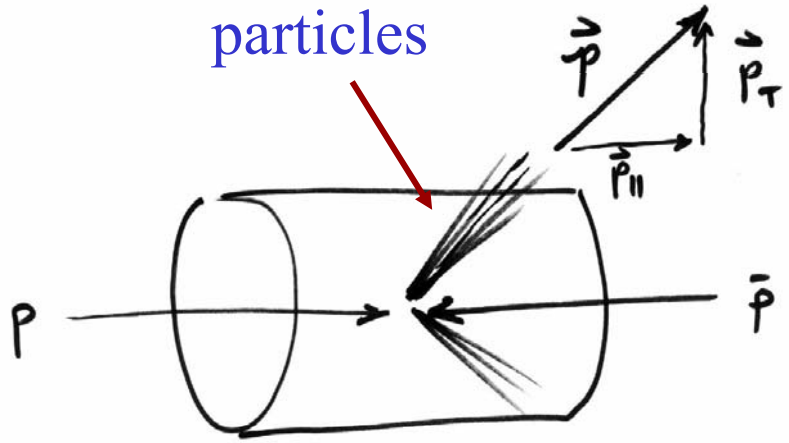
p-value of background-only hypothesis can depend crucially distribution $f(y|b)$ in the "search region".

Example of a "cut-based" study

In the 1990s, the CDF experiment at Fermilab (Chicago) measured the number of hadron jets produced in proton-antiproton collisions as a function of their momentum perpendicular to the beam direction:



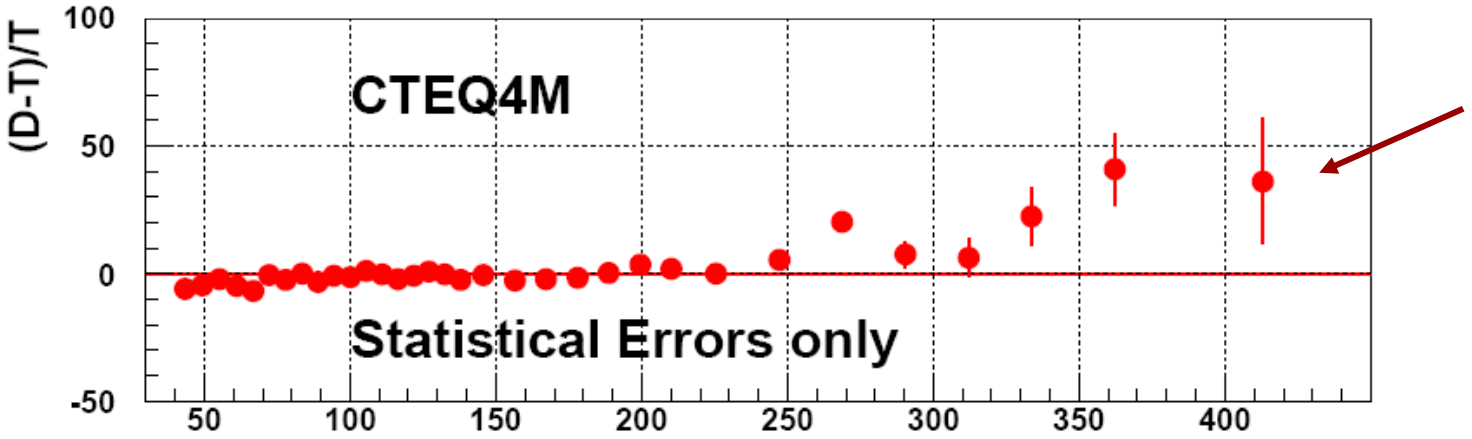
"jet" of particles



Prediction low relative to data for very high transverse momentum.

High p_T jets = quark substructure?

Although the data agree remarkably well with the Standard Model (QCD) prediction overall, the excess at high p_T appears significant:



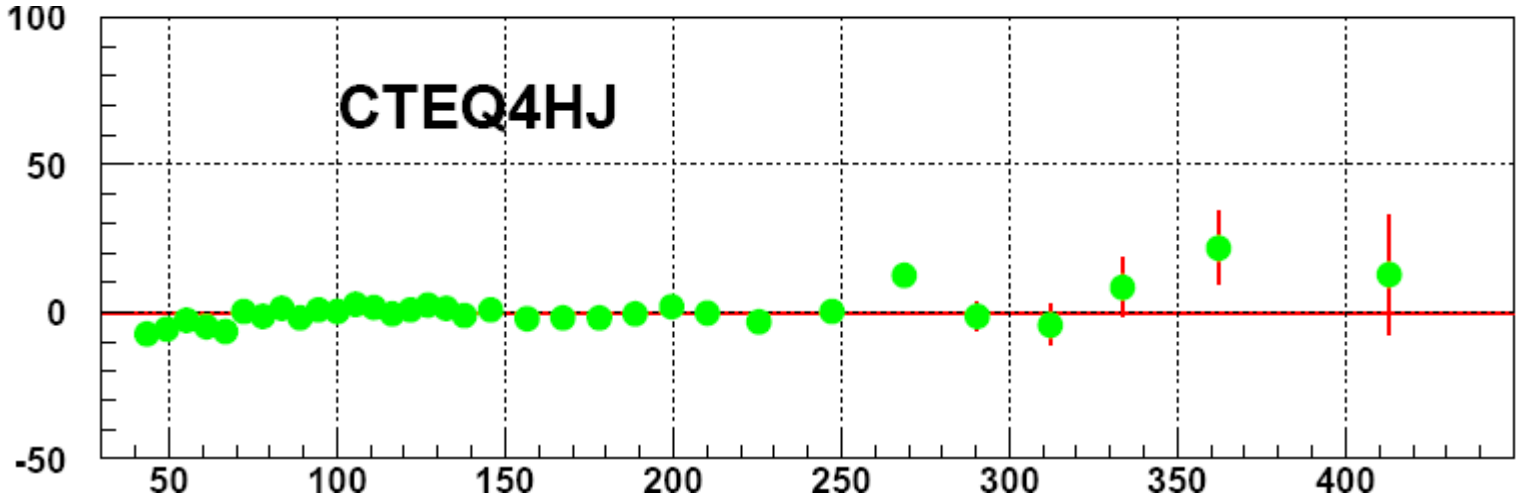
The fact that the variable is "understandable" leads directly to a plausible explanation for the discrepancy, namely, that quarks could possess an internal substructure.

Would not have been the case if the variable plotted was a complicated combination of many inputs.

High p_T jets from parton model uncertainty

Furthermore the physical understanding of the variable led one to a more plausible explanation, namely, an uncertain modelling of the quark (and gluon) momentum distributions inside the proton.

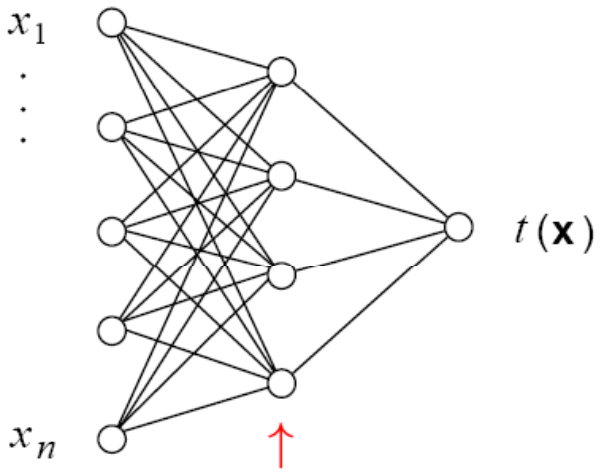
When model adjusted, discrepancy largely disappears:



Can be regarded as a "success" of the cut-based approach. Physical understanding of output variable led to solution of apparent discrepancy.

Neural networks in particle physics

For many years, the only "advanced" classifier used in particle physics.

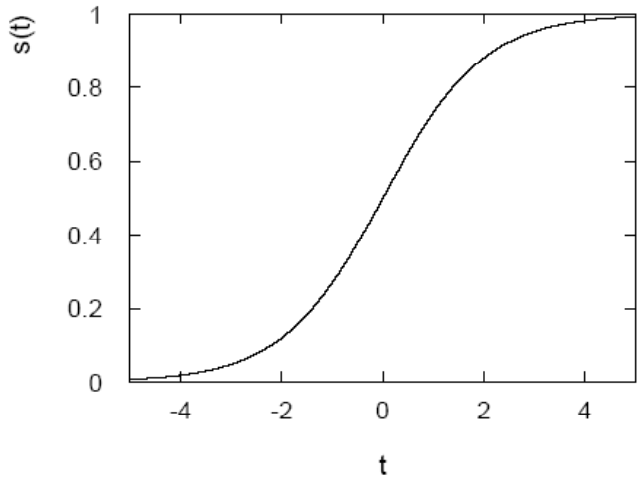


$$h_i(\vec{x}) = s\left(w_{i0} + \sum_{j=1}^{n} w_{ij}x_j\right) ,$$

$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i h_i(\vec{x})\right) .$$

Usually use single hidden layer,
logistic sigmoid activation function:

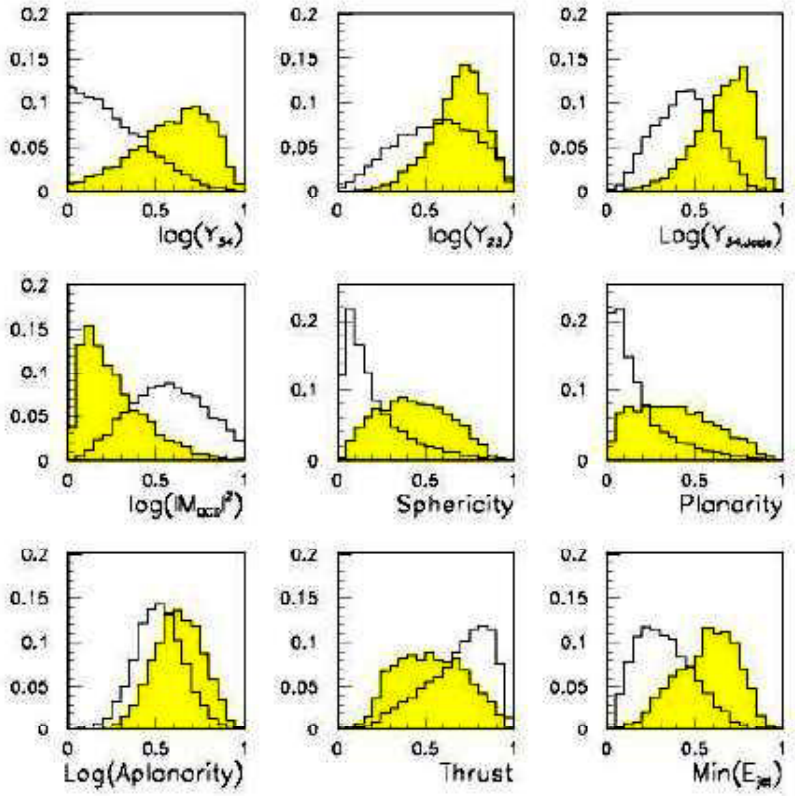
$$s(u) \equiv (1 - e^{-u})^{-1} .$$



Neural network example from LEP II

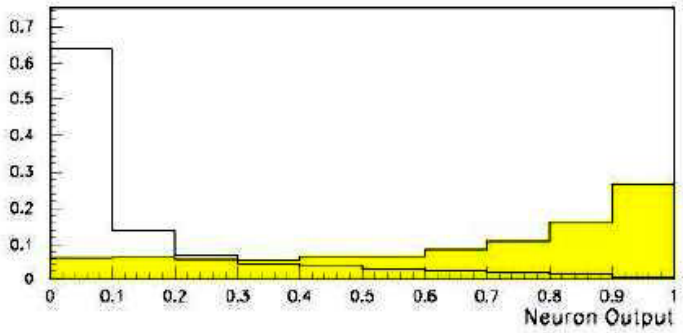
Signal: $e^+e^- \to W^+W^-$ (often 4 well separated hadron jets)

Background: $e^+e^- \to qqgg$ (4 less well separated hadron jets)



← input variables based on jet structure, event shape, ...
none by itself gives much separation.

Neural network output:



(Garrido, Juste and Martinez, ALEPH 96-144)

Some issues with neural networks

In the example with WW events, goal was to select these events so as to study properties of the W boson.

> Needed to avoid using input variables correlated to the properties we eventually wanted to study (not trivial).

In principle a single hidden layer with an sufficiently large number of nodes can approximate arbitrarily well the optimal test variable (likelihood ratio).

> Usually start with relatively small number of nodes and increase until misclassification rate on validation data sample ceases to decrease.

Usually MC training data is cheap -- problems with getting stuck in local minima, overtraining, etc., less important than concerns of systematic differences between the training data and Nature, and concerns about the ease of interpretation of the output.

Decision trees

Out of all the input variables, find the one for which with a single cut gives best improvement in signal purity:

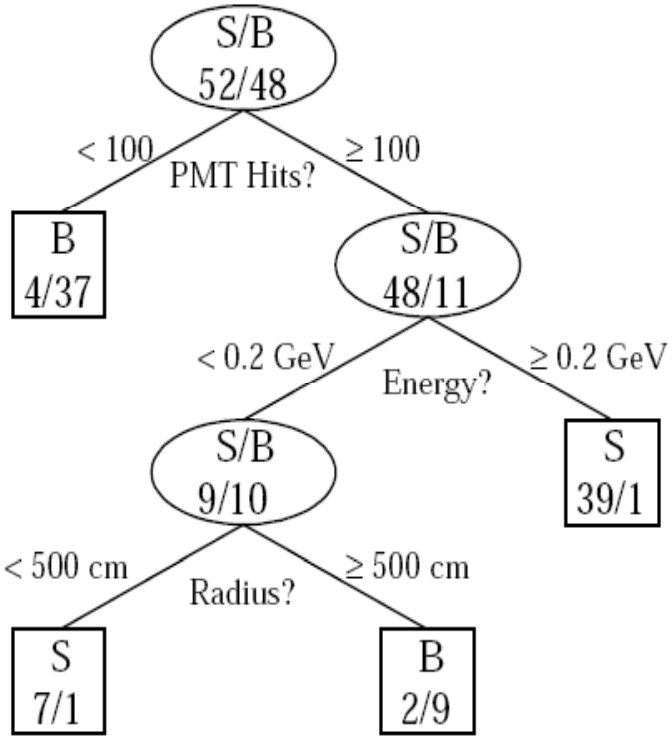
$$P = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}$$

where w_i. is the weight of the ith event.

Resulting nodes classified as either signal/background.

Iterate until stop criterion reached based on e.g. purity or minimum number of events in a node.

The set of cuts defines the decision boundary.



Example by MiniBooNE experiment,
B. Roe et al., NIM 543 (2005) 577

Boosting

The resulting classifier is usually very sensitive to fluctuations in the training data. Stabilize by boosting:

Create an ensemble of training data sets from the original one by updating the event weights (misclassified events get increased weight).

Assign a score α_k to the classifier from the kth training set based on its error rate ε_k:

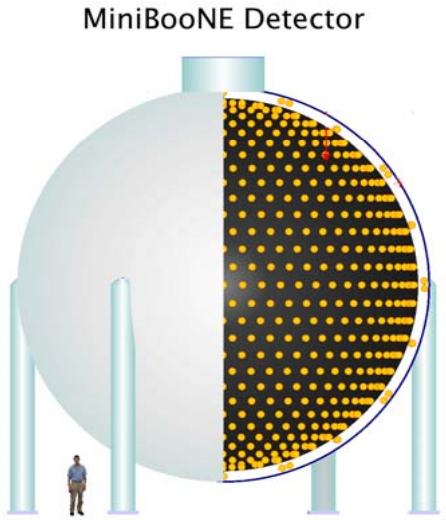
$$\alpha_k = \ln \frac{1-\varepsilon_k}{\varepsilon_k}$$

Final classifier is a weighted combination of those from the ensemble of training sets:

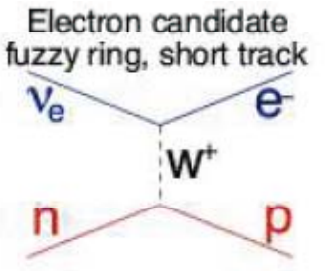
$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{x}, T_k)$$

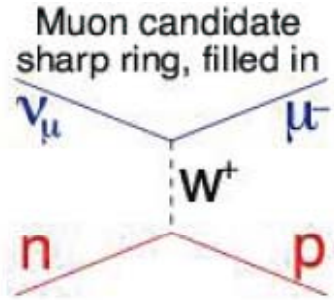
Particle i.d. in MiniBooNE

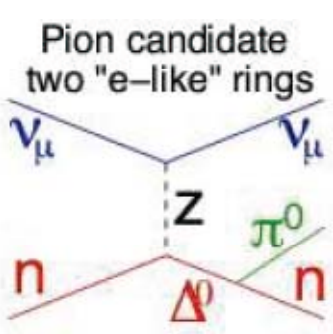
Detector is a 12-m diameter tank of mineral oil exposed to a beam of neutrinos and viewed by 1520 photomultiplier tubes:

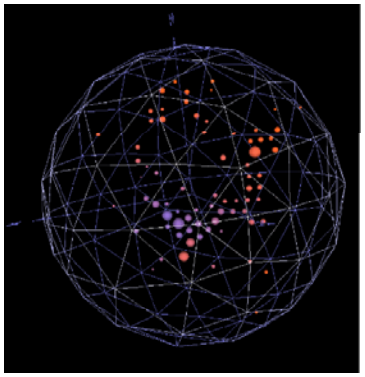


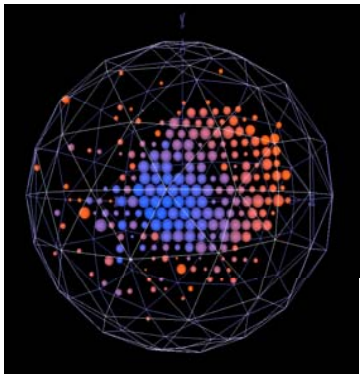
Search for ν_μ to ν_e oscillations required particle i.d. using information from the PMTs.

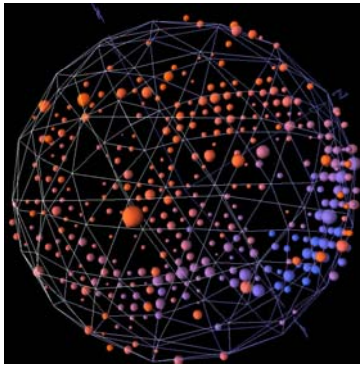










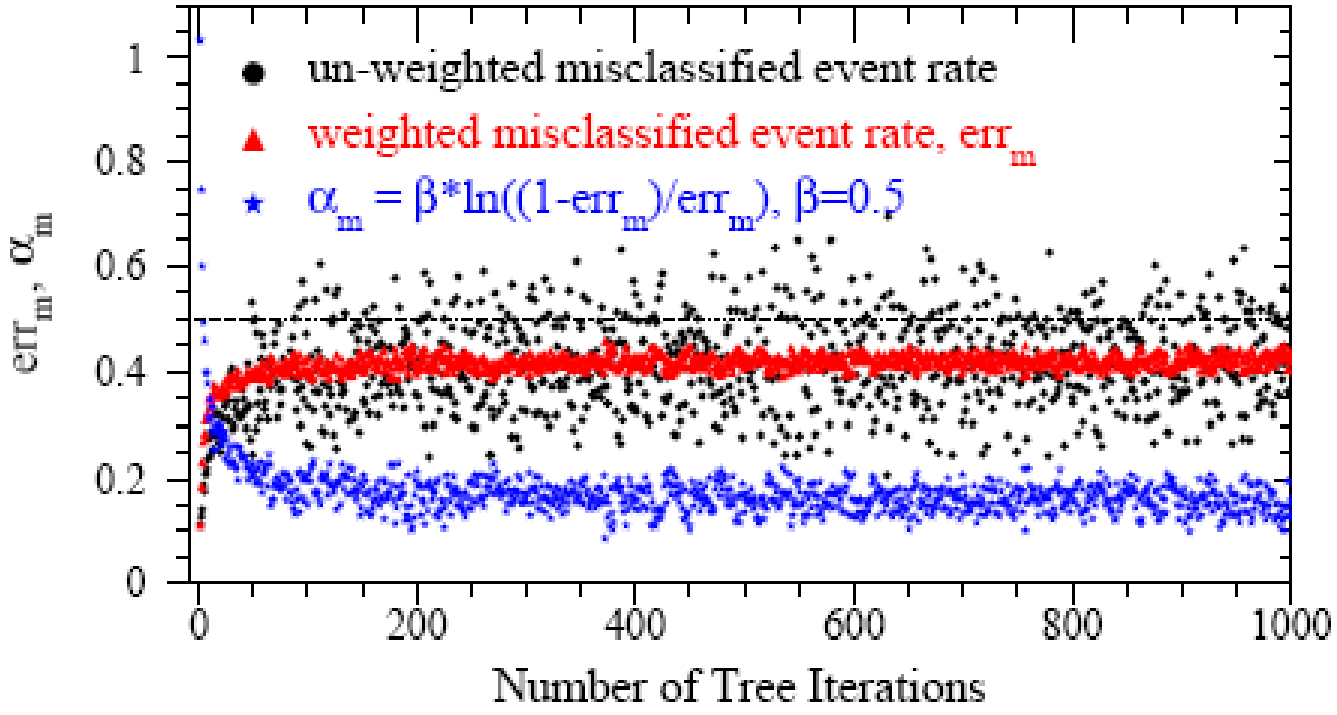


H.J. Yang, MiniBooNE PID, DNP06

BDT example from MiniBooNE

~200 input variables for each event (ν interaction producing e, μ or π).

Each individual tree is relatively weak, with a misclassification error rate ~ 0.4 – 0.45

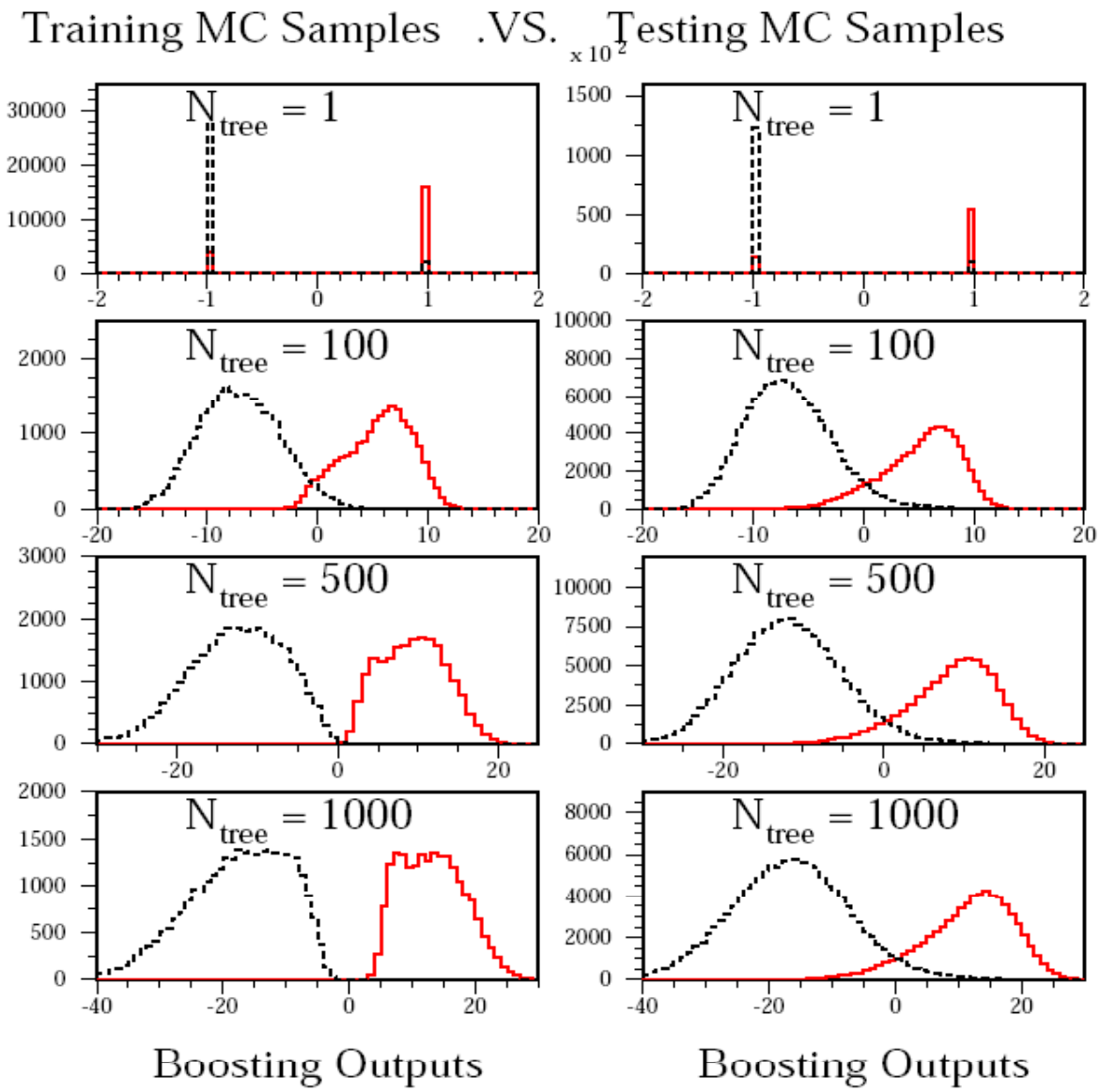


B. Roe et al., NIM 543 (2005) 577

Monitoring overtraining

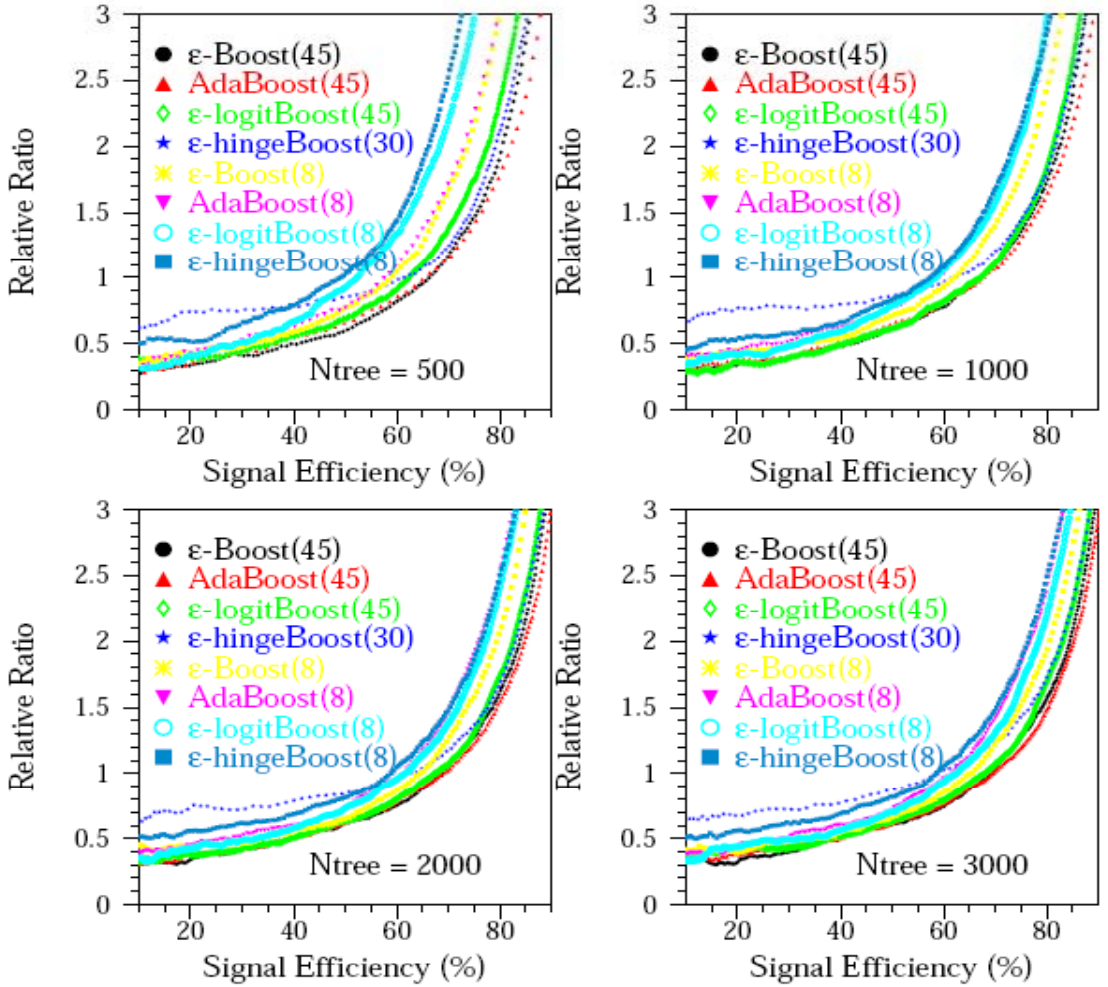
From MiniBooNE example:

Performance stable after a few hundred trees.



Comparison of boosting algorithms

A number of boosting algorithms on the market; differ in the update rule for the weights.



Boosted decision tree comments

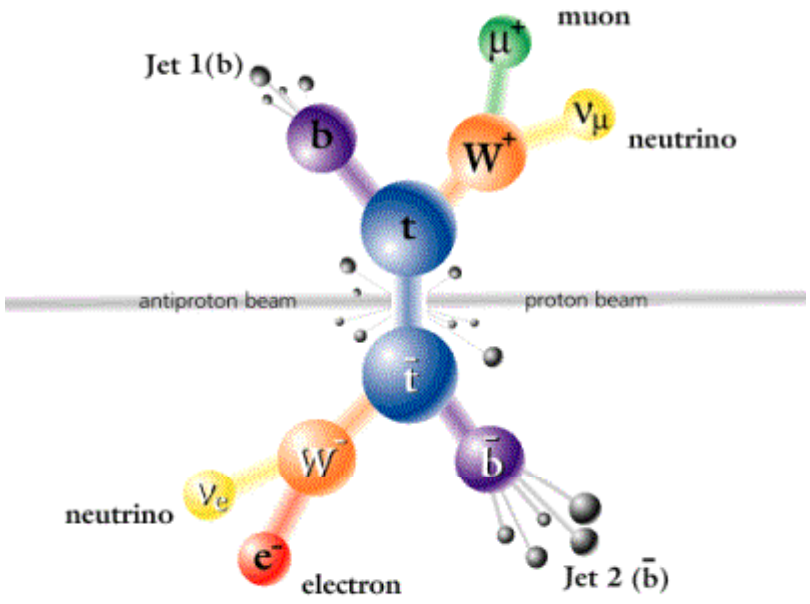
Boosted decision trees have become popular in particle physics because they can handle many inputs without degrading; those that provide little/no separation are rarely used as tree splitters are effectively ignored.

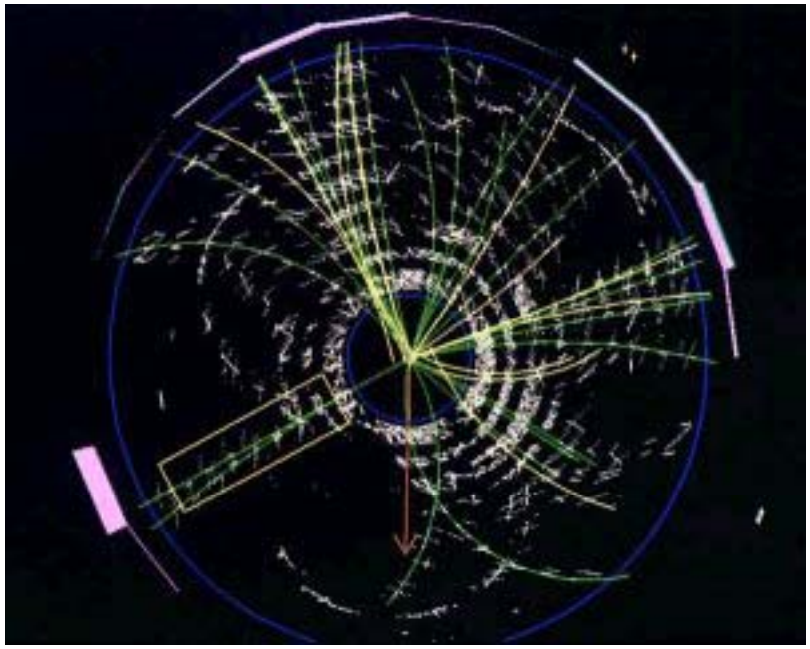
A number of boosting algorithms have been looked at, which differ primarily in the rule for updating the weights (ε-Boost, LogitBoost,...).

Some studies have looked at other ways of combining weaker classifiers, e.g., Bagging (Boostrap-Aggregating), generates the ensemble of classifiers by random sampling with replacement from the full training sample. Not much experience yet with these.

The top quark

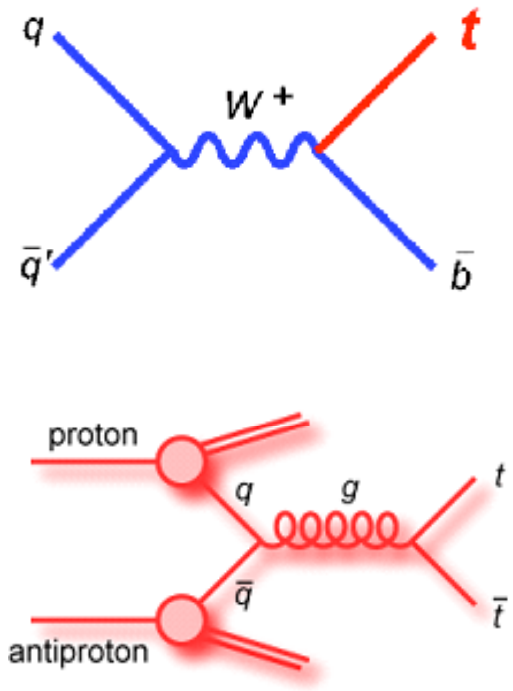
Top quark is the heaviest known particle in the Standard Model.
Since mid-1990s has been observed produced in pairs:



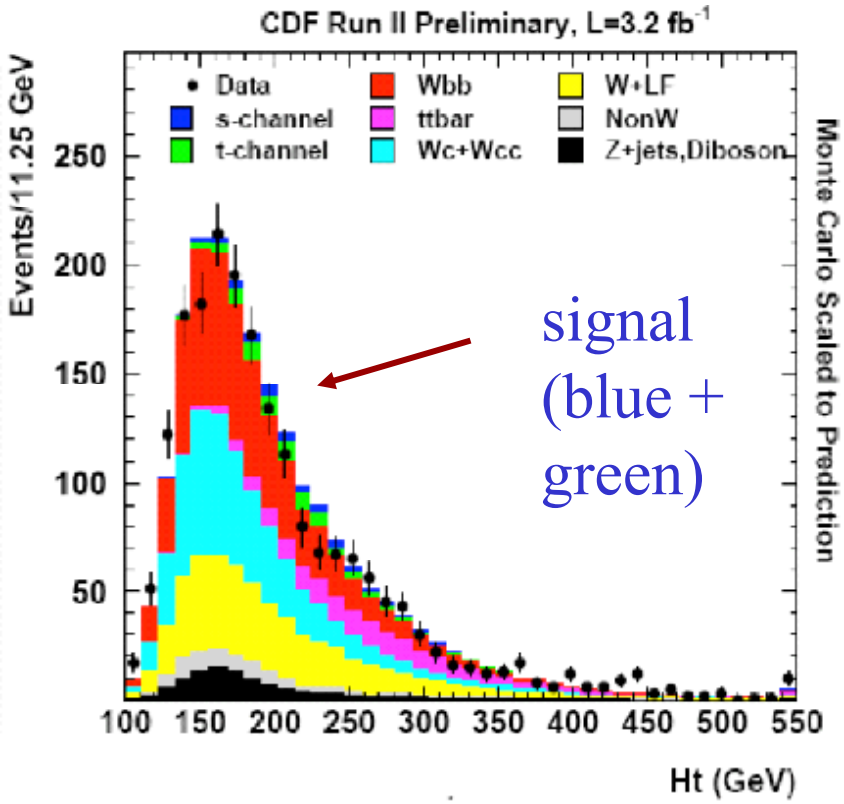


Single top quark production

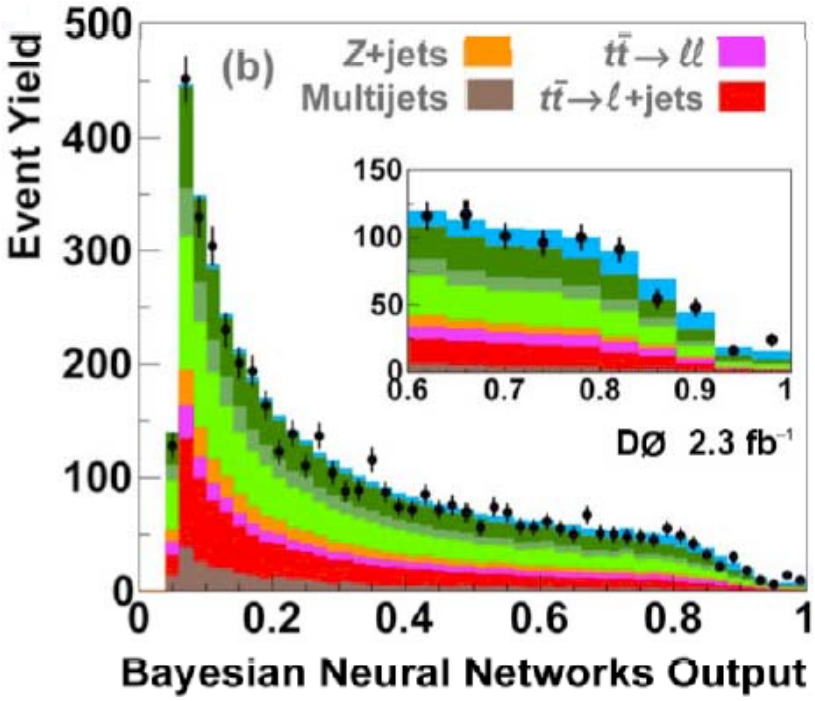
One also expected to find singly produced top quarks; pair-produced tops are now a background process.

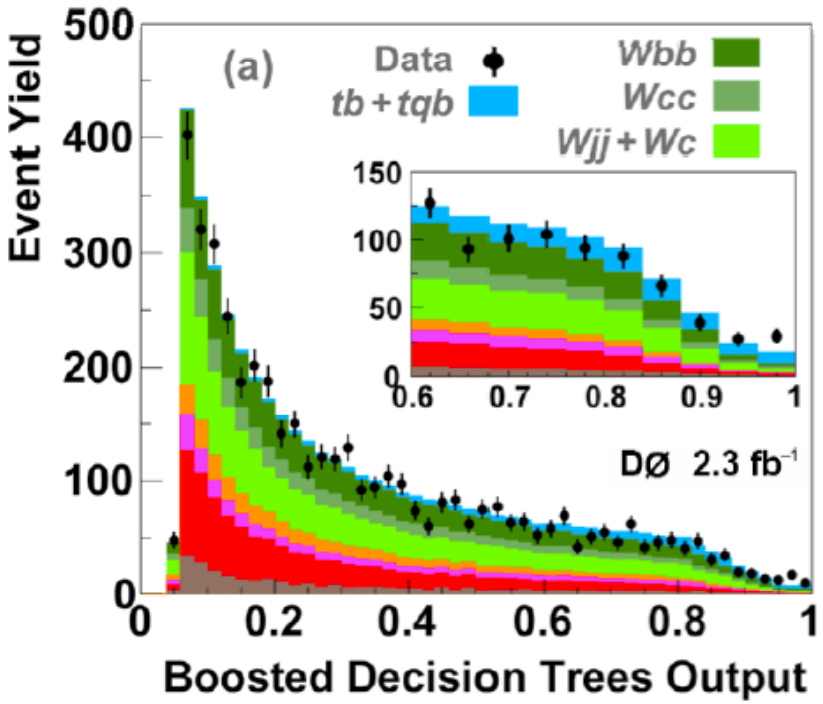


Use many inputs based on jet properties, particle i.d., ...



Different classifiers for single top





Also Naive Bayes and various approximations to likelihood ratio,....

Final combined result is statistically significant (>5σ level) but not easy to understand classifier outputs.

Support Vector Machines

Map input variables into high dimensional feature space: $\boldsymbol{x} \rightarrow \boldsymbol{\phi}$

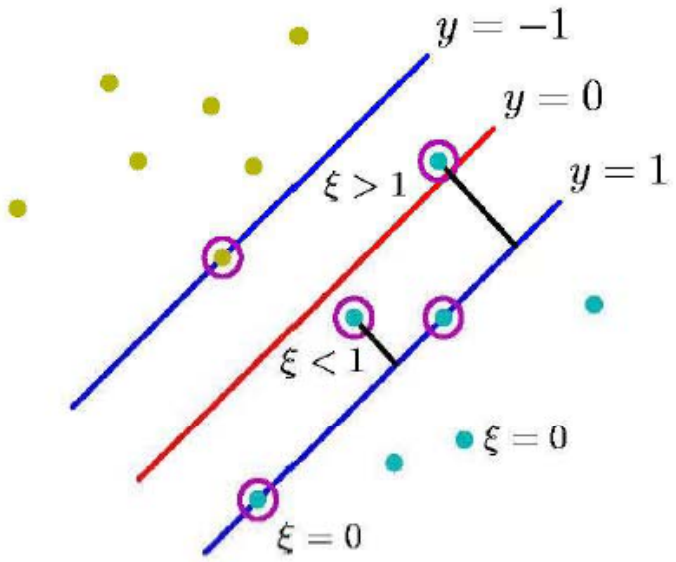
Maximize distance between separating hyperplanes (margin) subject to constraints allowing for some misclassification.

Final classifier only depends on scalar products of $\boldsymbol{\phi}(\boldsymbol{x})$:

$$y(\boldsymbol{x}) = \text{sign}\left(\sum_i \alpha_i y_i \vec{\varphi}(\boldsymbol{x}) \cdot \vec{\varphi}(\boldsymbol{x}_i) + b \right)$$

So only need kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \vec{\varphi}(\boldsymbol{x}) \cdot \vec{\varphi}(\boldsymbol{x}')$$



Bishop ch 7

Using an SVM

To use an SVM the user must as a minimum choose

a kernel function (e.g. Gaussian)
any free parameters in the kernel (e.g. the σ of the Gaussian)
the cost parameter C (plays role of regularization parameter)

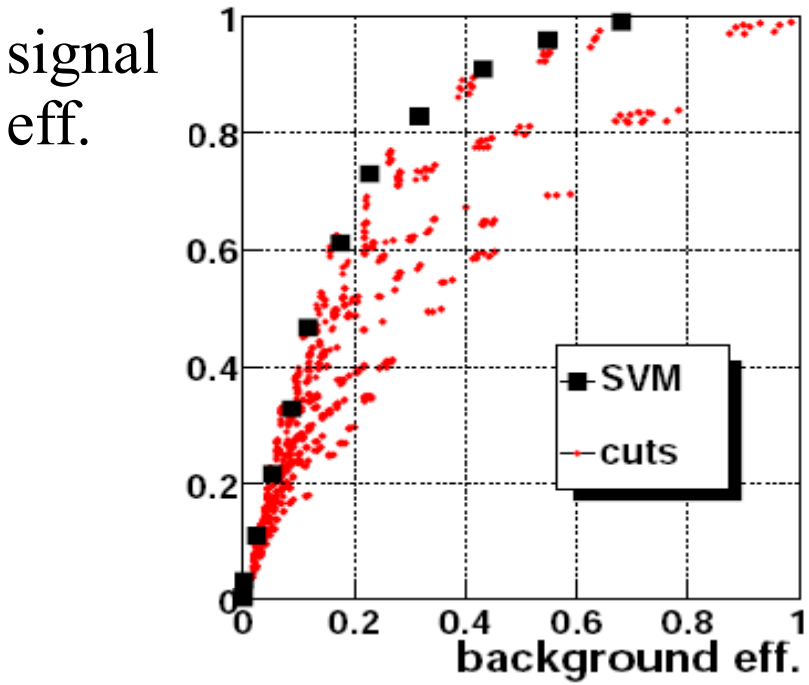
The training is relatively straightforward because, in contrast to neural networks, the function to be minimized has a single global minimum.

Furthermore evaluating the classifier only requires that one retain and sum over the support vectors, a relatively small number of points.

The advantages/disadvantages and rationale behind the choices above is not always clear to the particle physicist -- help needed here.

SVM in particle physics

SVMs are very popular in the Machine Learning community but have yet to find wide application in HEP. Here is an early example from a CDF top quark anlaysis (A. Vaiciulis, contribution to PHYSTAT02).



Summary, conclusions, etc.

Particle physics has used several multivariate methods for many years:

linear (Fisher) discriminant
neural networks
naive Bayes

and has in the last several years started to use a few more

k-nearest neighbour
boosted decision trees
support vector machines

The emphasis is often on controlling systematic uncertainties between the modeled training data and Nature to avoid false discovery.

Although many classifier outputs are "black boxes", a discovery at 5σ significance with a sophisticated (opaque) method will win the competition if backed up by, say, 4σ evidence from a cut-based method.

Quotes I like

*"Alles sollte so einfach wie möglich sein,
aber nicht einfacher."*

– A. Einstein

*"If you believe in something
you don't understand, you suffer,..."*

– Stevie Wonder

Extra slides

Software for multivariate analysis

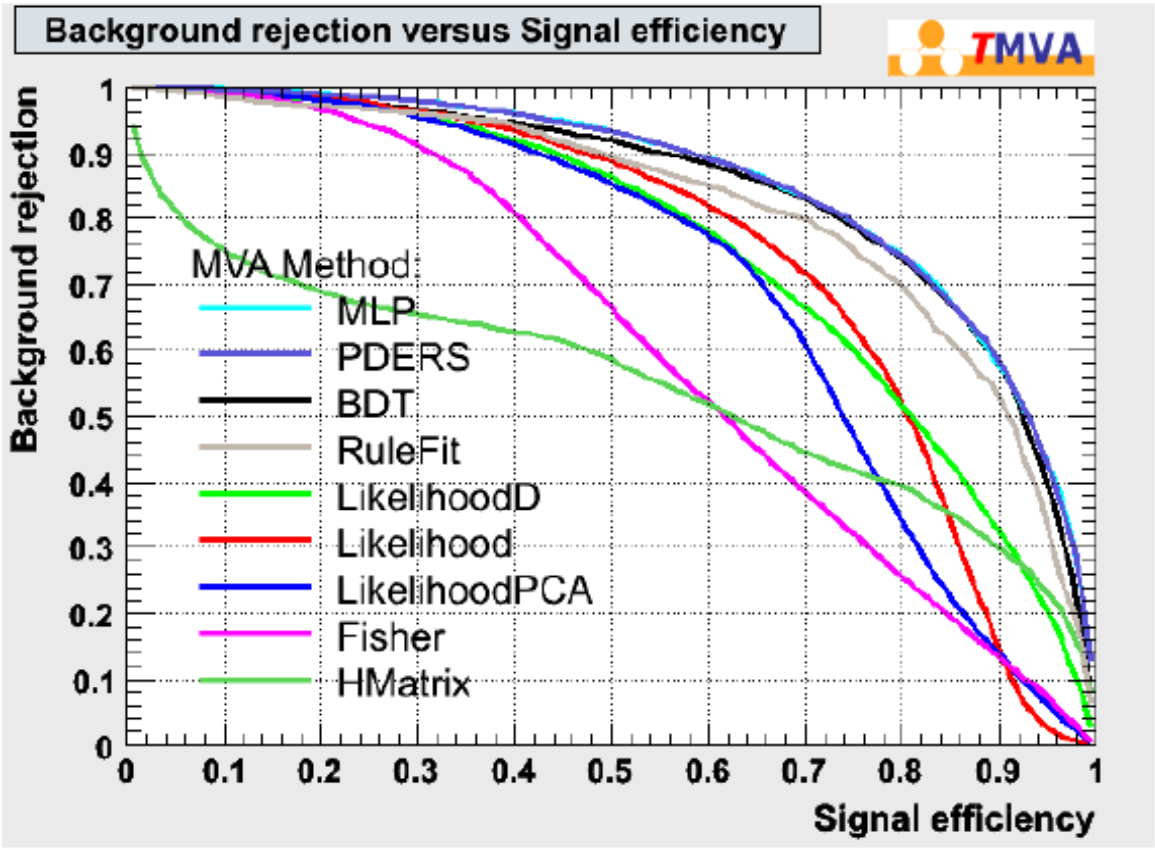
TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

From **tmva.sourceforge.net**, also distributed with ROOT
Variety of classifiers
Good manual

StatPatternRecognition, I. Narsky, physics/0507143

Further info from **www.hep.caltech.edu/~narsky/spr.html**
Also wide variety of methods, many complementary to **TMVA**
Currently appears project no longer to be supported

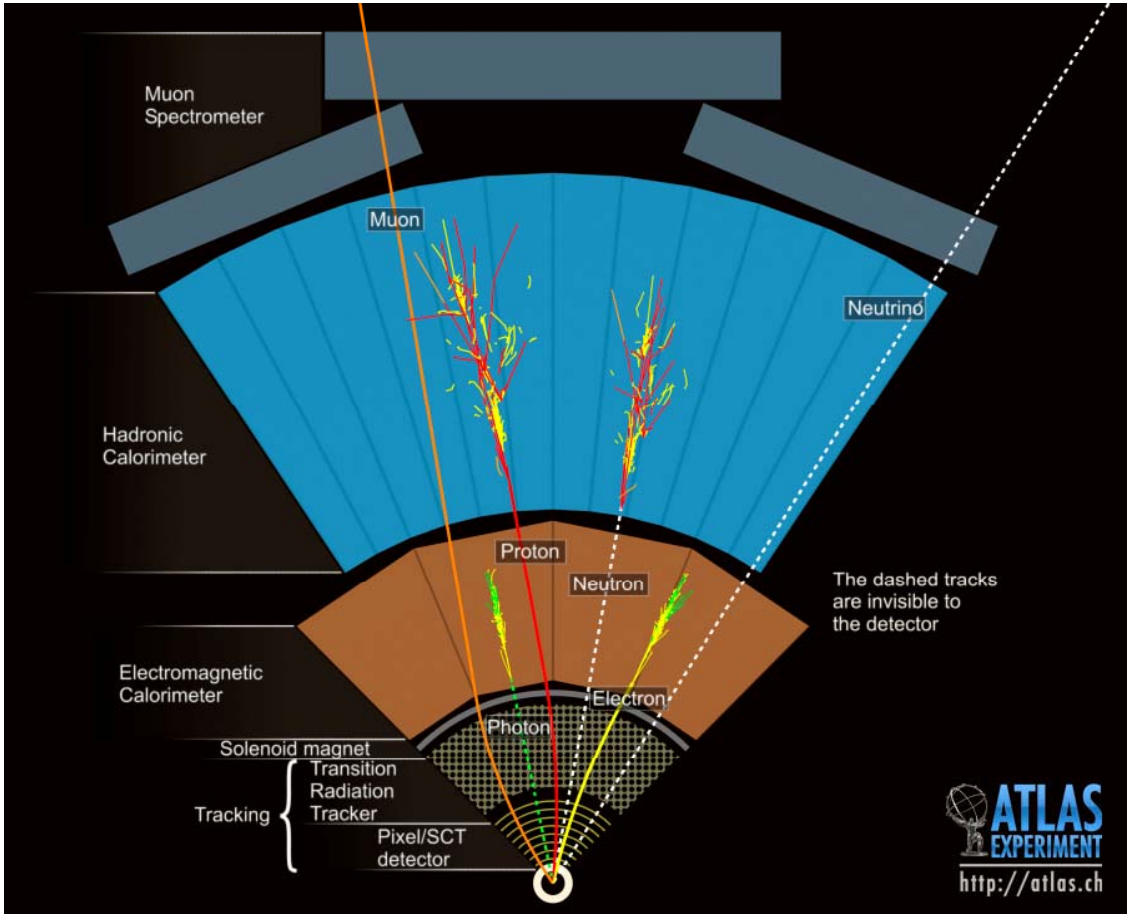
Comparing multivariate methods (TMVA)



Choose the best one!

Identifying particles in a detector

Different particle types (electron, pion, muon,...) leave characteristically distinct signals as in the particle detector:



But the characteristics overlap, hence the need for multivariate classification methods.

Goal is to produce a list of "electron candidates", "muon candidates", etc. with well known acceptance probabilities for all particle types.

Example of neural network for particle i.d.

For every particle measure pattern of energy deposit in calorimeter

~ shower width, depth

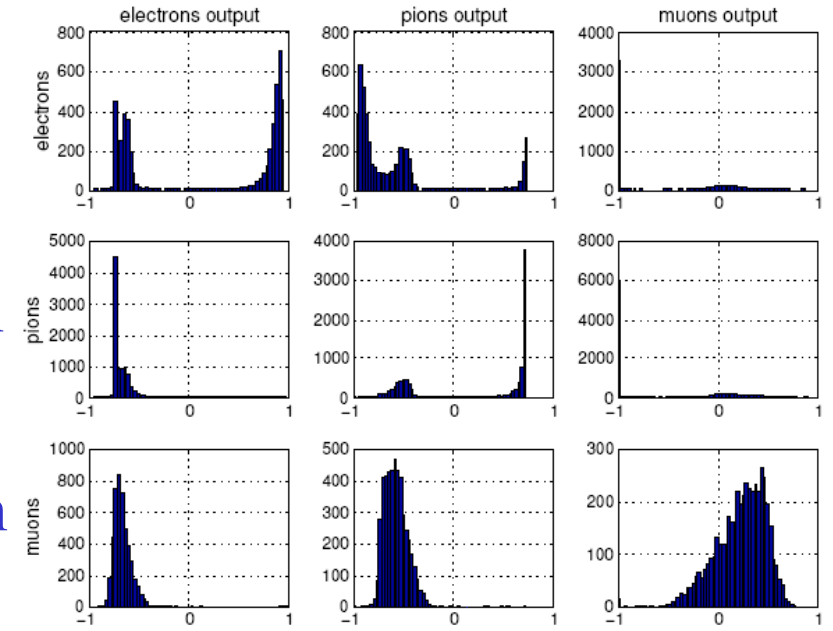
Get training data by placing detector in test beam of pions, muons, etc.

here muon beam essentially "pure"; electron and pion beams both have significant contamination.

ATLAS Calorimeter test

NN architecture:

10 input nodes
8 nodes in 1 hidden layer
3 output nodes



Damazio and de Seixas