

Modern Data Analysis Techniques for High Energy Physics

Kyle Cranmer
Brookhaven National Laboratory

April 11, 2005
EFI High Energy Physics Seminar

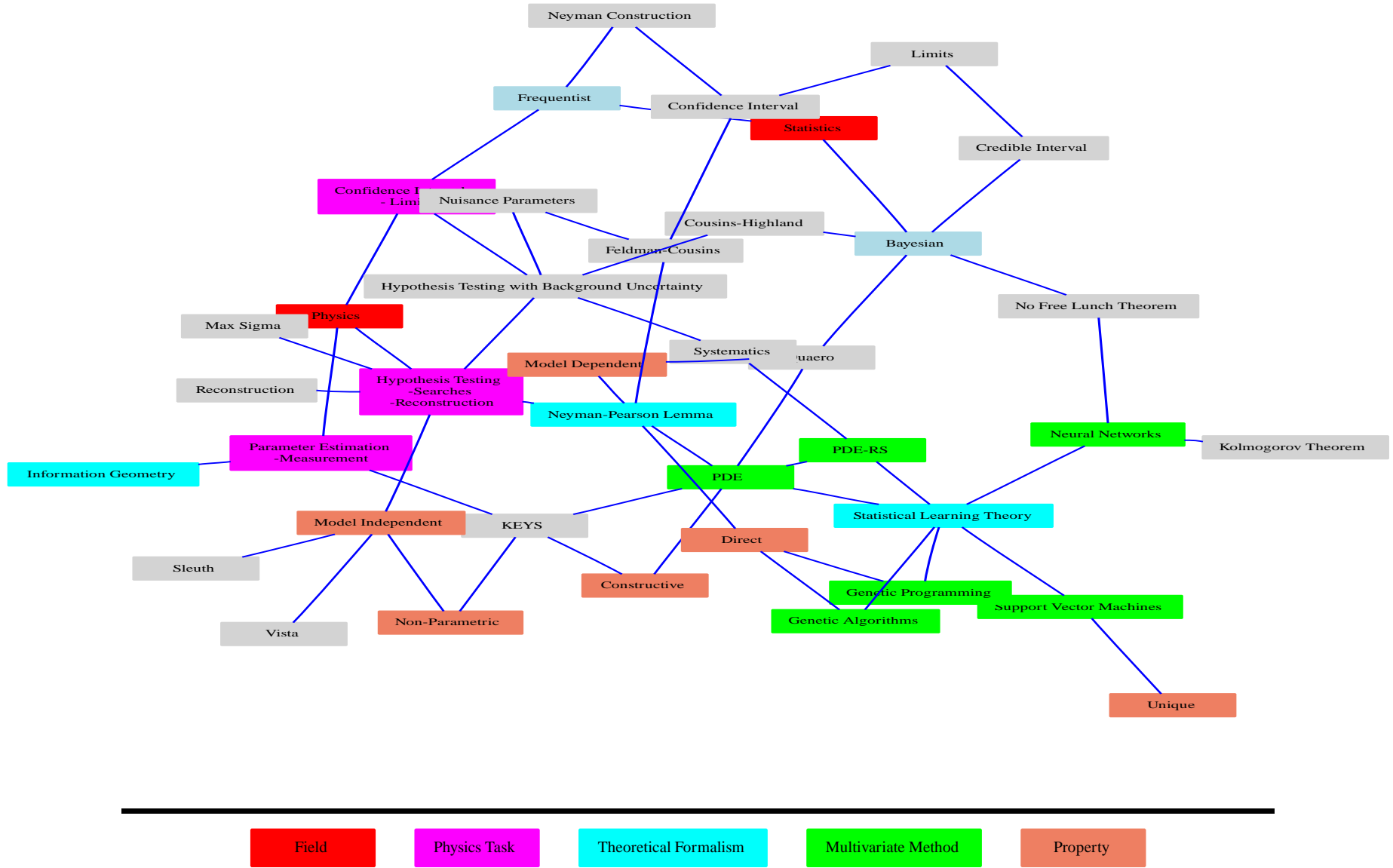
In this talk I would like to talk about

- New multivariate analysis techniques and insights
- Model-Independent vs. Model-Dependent Searches
- Event Weighting & Incorporation of Systematics
- Challenges for the LHC

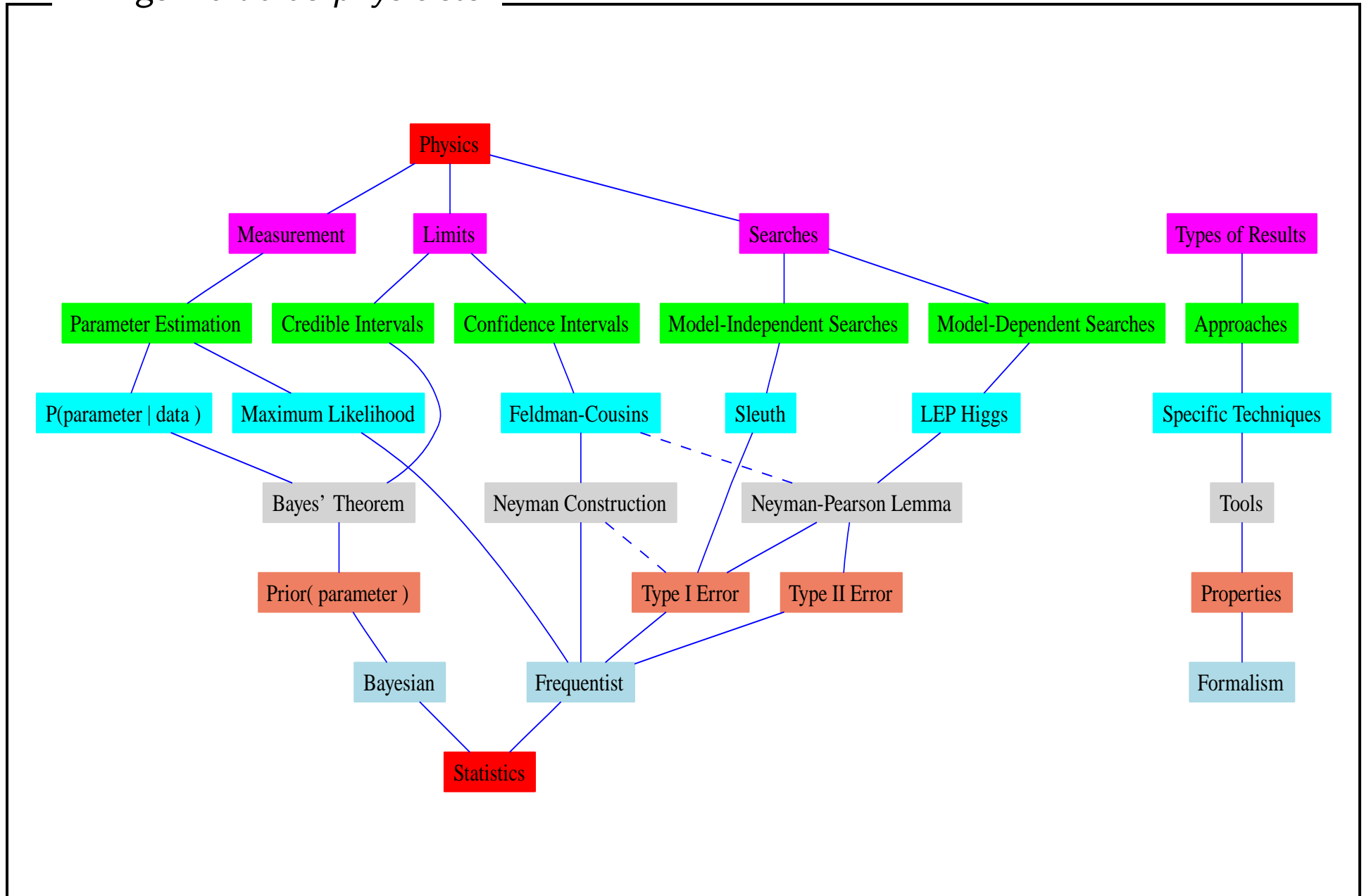
I will do my best to

- Discuss different topics in a common framework/formalism
- Motivate a new analysis technique with clear physics goal
- Fairly assess pro's / con's of different techniques

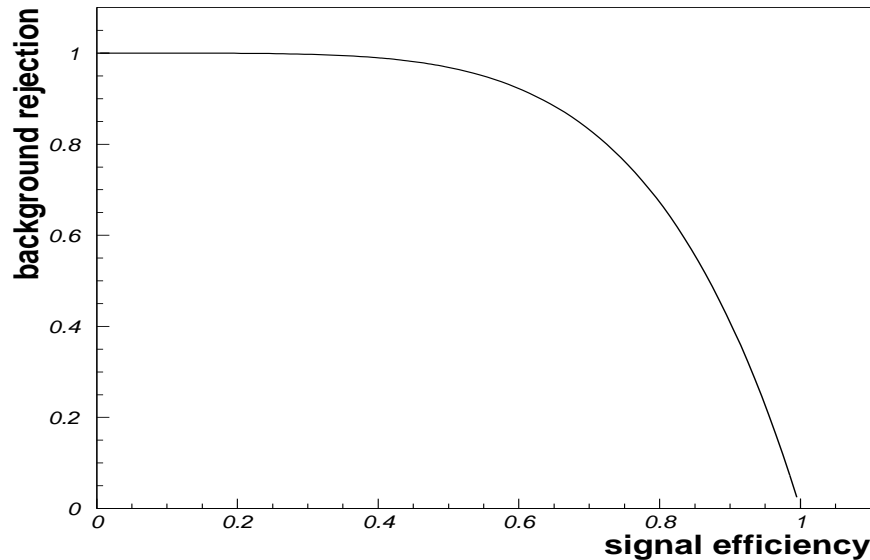
Overview



Things we do as physicists



Hypothesis Testing



Particle Identification and (model-dependent) Searches for new particles are both examples of Hypothesis Testing.

Hypothesis Testing	Particle ID	Searches
Null-Hypothesis (H_0)	π^-	Standard Model
Alternate Hypothesis (H_1)	e^-	Higgs
Type I Error	mis-tag $\pi \rightarrow e$	False Discovery
Type II Error	mis-tag $e \rightarrow \pi$	Missed Discovery
Observation	An Event	An Experiment
Truth	Property of Event	Property of Nature

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:

- the Null Hypothesis H_0 (background only)
- the Alternate Hypothesis H_1 (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

Find the region W such that we minimize the probability of wrongly accepting the H_0 (when H_1 is true)

$$\beta = P(x \in W | H_1)$$

The region W that minimizes the probability of wrongly accepting the H_0 is just a contour of the Likelihood Ratio:

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

This is the goal!

The problem is we don't have access to $L(x|H_0)$ & $L(x|H_1)$

Given our stated goal, what should be priorities be?

- 1 Maximize Power (minimize Type II error for fixed Type I error)
- 2 Approximate W
- 3 Approximate the Likelihood Ratio
- 4 Approximate $L(x|H_0)$ & $L(x|H_1)$
- 5 More indirect methods

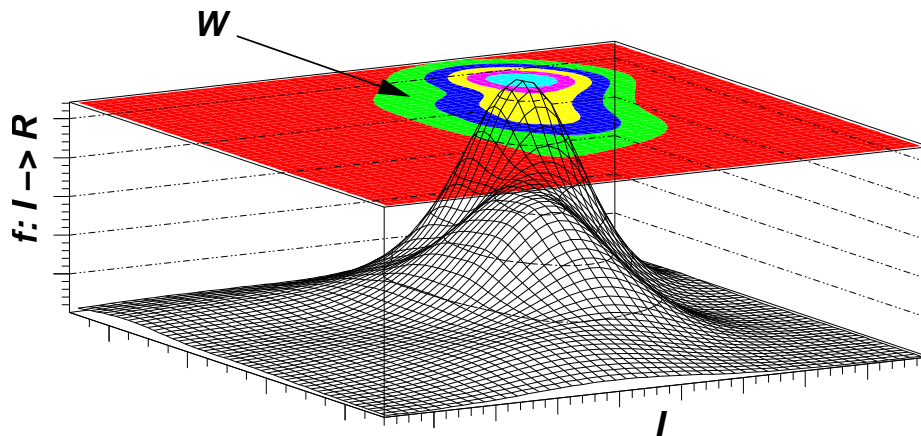
What do our algorithms actually do?

Algorithm	Priority	Goal of Algorithm
Cuts	2	Approximate W by looking at 1-d histos
Neural Nets	5	Approximate an auxiliary function
Kernel Methods	4-5	Approximate $P(x H_0)$ & $P(x \text{signal})$
Genetic Programming	1	Directly optimizes user-defined performance

Holmström, Sain, Miettinen *Comp. Phys. Comm.* **88** 1995 Cranmer, *Comp. Phys. Comm.* **136** 2001

Kernel Estimation Techniques used to construct p.d.f.'s from Monte Carlo

$$f(x) = \frac{1}{l} \sum_{i=1}^l K(x - x_i)$$



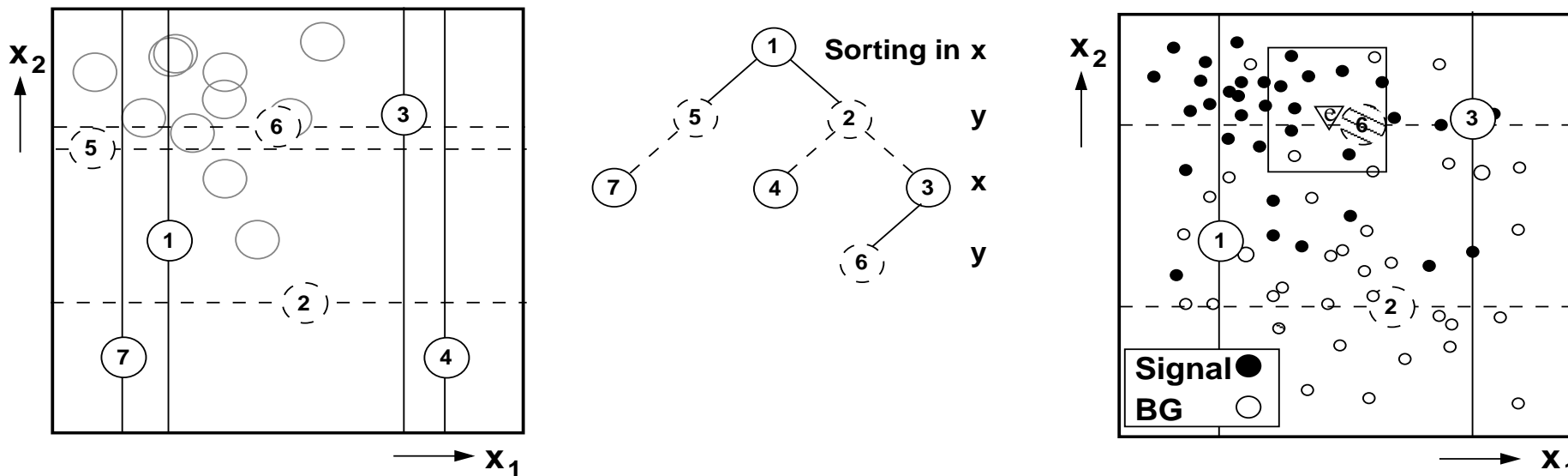
Construct $D(x)$, a Discriminant function, instead of the Likelihood Ratio:

$$D(x) = \frac{f_S(x)}{f_S(x) + f_b(x)}$$

Signal $\rightarrow 1$, Background $\rightarrow 0$

Most of the theory of Kernel Estimation is based on asymptotic properties ($l \rightarrow \infty$), but in reality we only have limited data.

PDE methods are intuitive and powerful, but slow to evaluate



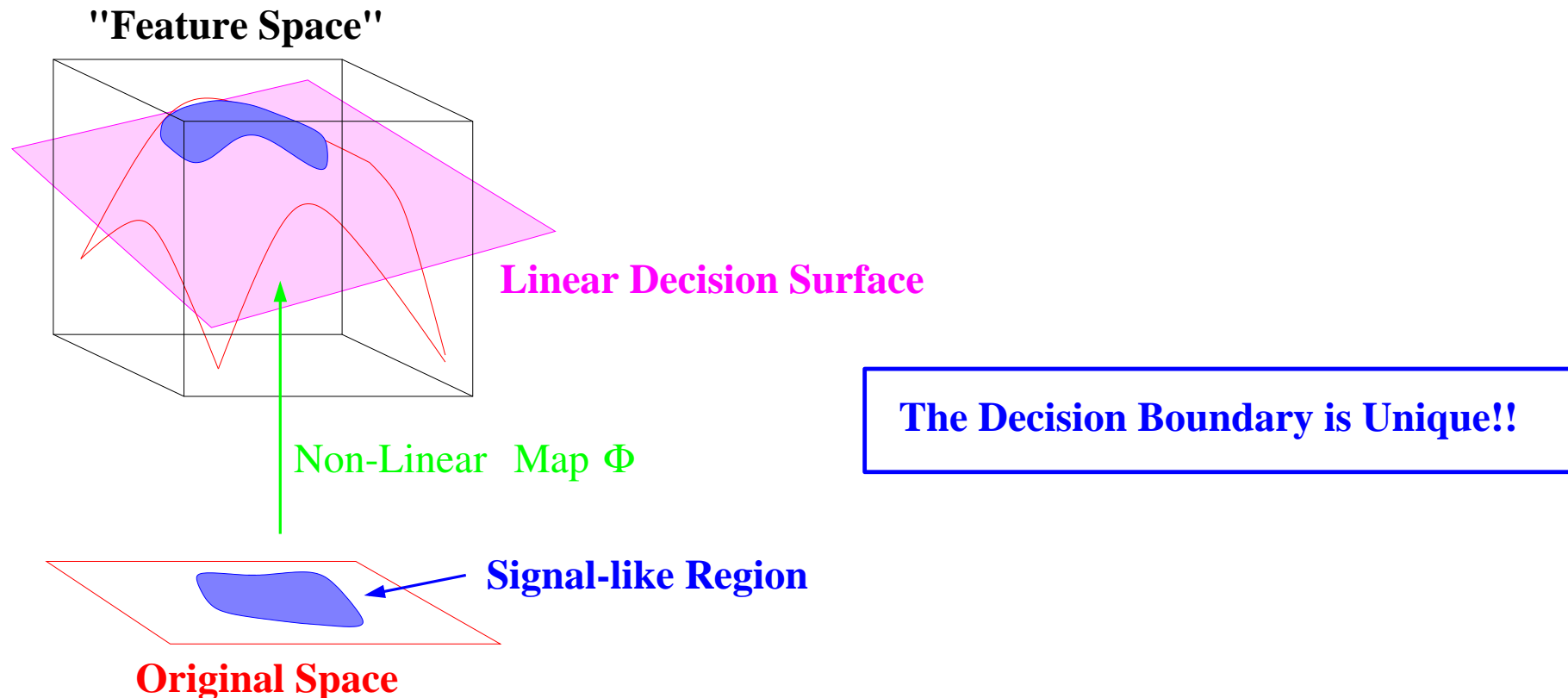
T.Carli & B. Koblitz proposed a different implementation in *NIM A 501* (2003) which uses Range Searching

Provides a new handle to assess systematics or remove under-trained regions. **And It's Fast!**

Support Vector Machines

Support Vector Machines find an “optimal” decision hyperplane in a high-dimensional “Feature Space”.

A non-linear map from the original space to the “feature space” is what allows the signal region to be of arbitrary complexity.



Evolutionary Computing: The Metaphor

Evolutionary Computing is usually discussed within the context of a Metaphor...

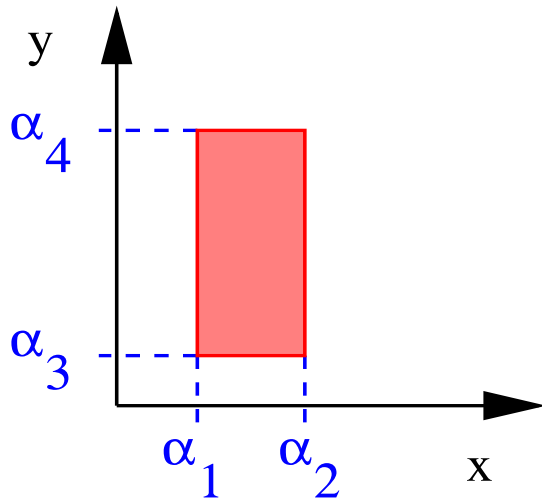
Population	↔	Set of cuts
Individual	↔	A particular cut
Evolution	↔	Optimization of a set of cuts
Generation	↔	A training epoch
Mutation	↔	Stochastic search
Fitness	↔	Significance (in “sigma”)
Competition	↔	Sampling a Fitness Distribution

CAUTION! Genetic Programming \neq Genetic Algorithms

Genetic Programming and Genetic Algorithms rest on a similar Metaphor, but the techniques are quite different.

While Genetic Algorithms have been used in HEP, Genetic Programming seems to be new technique for event selection!

- Cranmer & Bowman *Comput. Phys. Commun.* **167** (2005) 165 (physics/0402030)
- The FOCUS Collaboration *submitted to NIM A* (hep-ex/0503007)



Cuts can be viewed as an individual.

$$f = \begin{cases} 1 & \alpha_1 < x < \alpha_2 \text{ and } \alpha_3 < y < \alpha_4 \\ 0 & \text{else} \end{cases}$$

Thresholds form genotype for an individual.

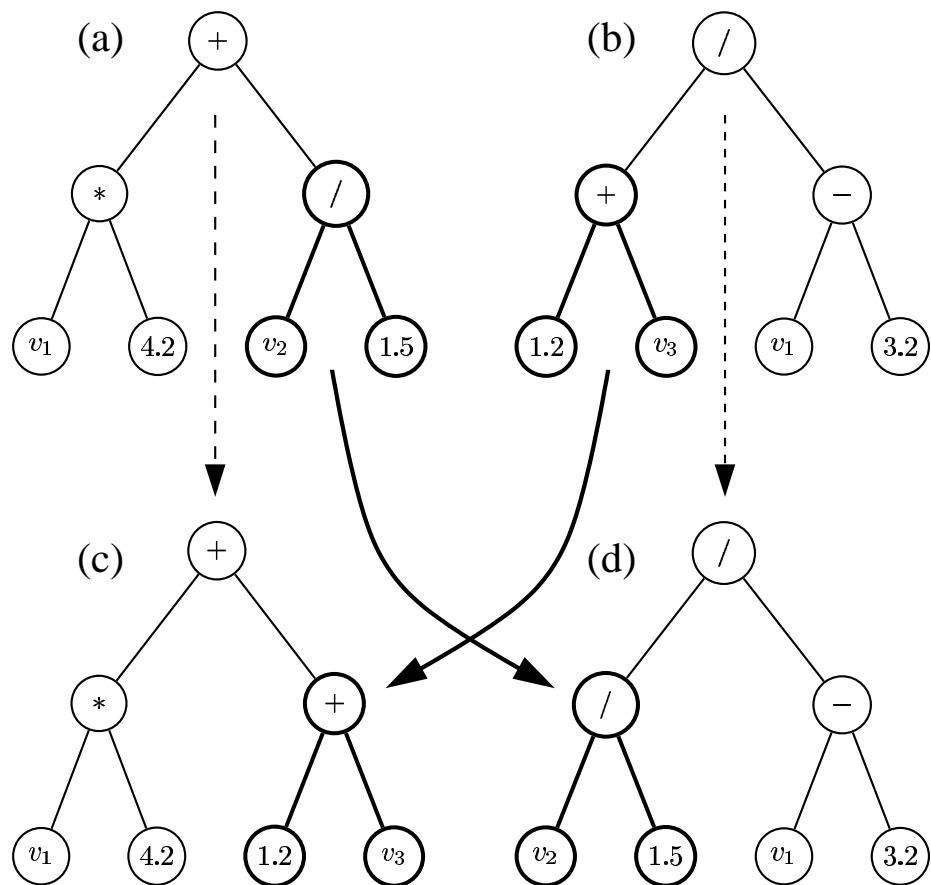
α_1	α_2	α_3	α_4
------------	------------	------------	------------

Evolution of the genotype = optimizing cuts.

Population is less prone to finding local minima.

The form of the cut is hard-coded.

Genetic Programming



(physics/0402030)

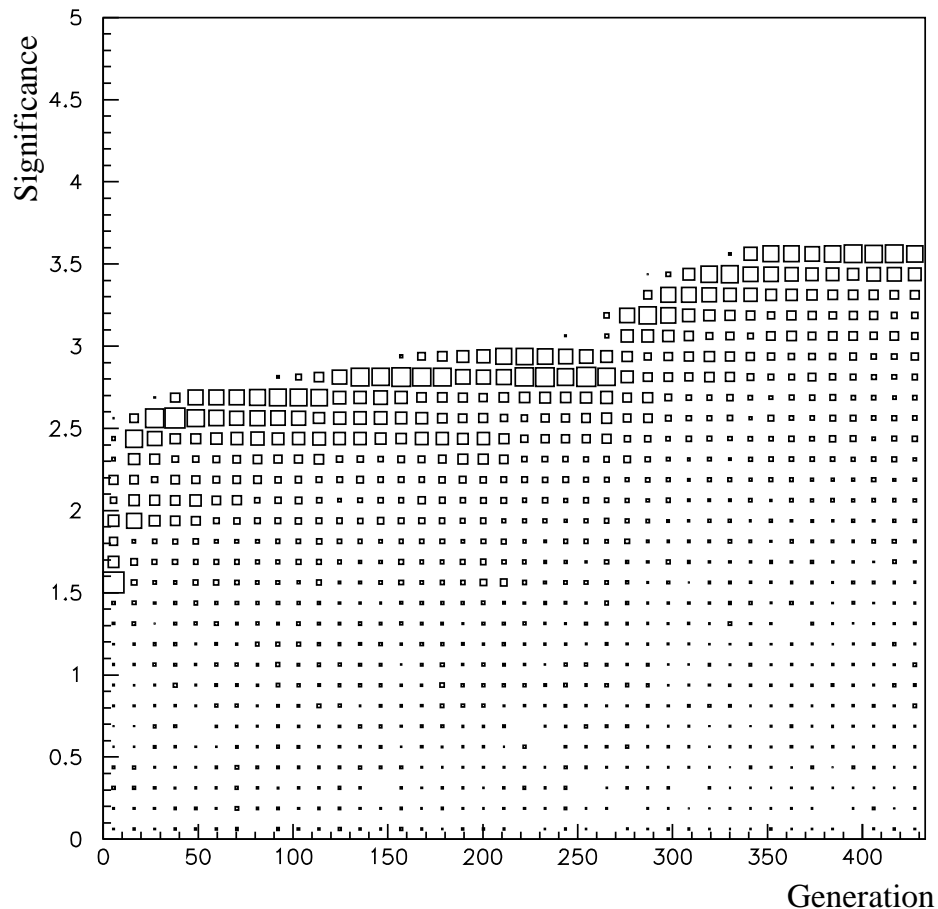
Represent function as a tree:
 $f(v_1, v_2) = (v_1 * 4.2) + (v_2/1.5)$

An event is accepted if
 $-1 < f(v_1, v_2) < 1$

An individual can consist of
Boolean conjunction of cuts

Construct a population of
individuals, let them compete
w.r.t. a user-defined performance
measure (e.g. s/\sqrt{b})

Evolution consists of node muta-
tions and individuals cross-over



Neural Networks:

- Hard-coded performance measure
- Deterministic optimization of a single function
- Sensitive to local Minima
- “Black Box” property

PhysicsGP:

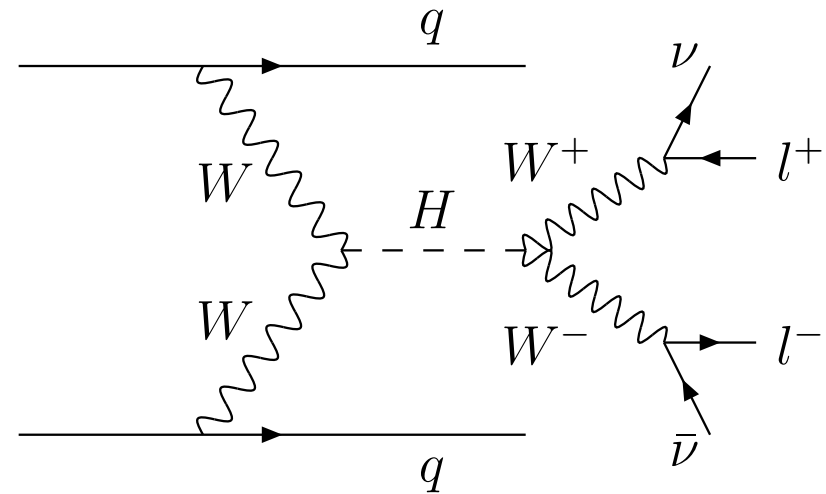
- User-defined performance measure
- Stochastic optimization of an entire population
- Relatively Insensitive to local minima
- Cuts are human-readable in principle

Comparison of Multivariate Methods

Compared Neural Networks, PhysicsGP, and Support Vector Regression

Used 7 variables: $\Delta\eta_u, \Delta\phi_u, M_u, \Delta\eta_{jj}, \Delta\phi_{jj}, M_{jj}, M_T$

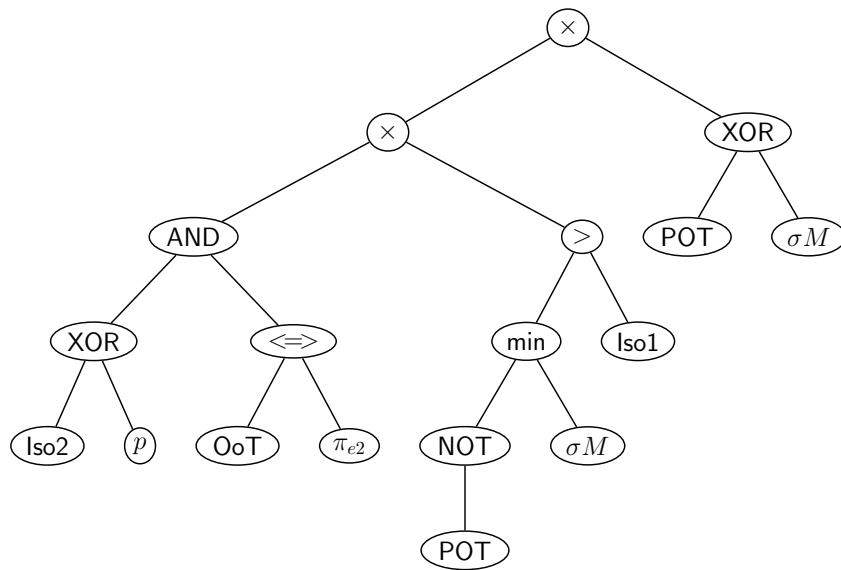
Three methods have similar results, PhysicsGP a novel, powerful method



	Ref. Cuts	low- m_H Cuts	NN	GP	SVR
120 ee	0.87	1.25	1.72	1.66	1.44
120 $e\mu$	2.30	2.97	3.92	3.60	3.33
120 $\mu\mu$	1.16	1.71	2.28	2.26	2.08
Combined	2.97	3.91	4.98	4.57	4.26
130 $e\mu$	4.94	6.14	7.55	7.22	6.59

Table 1: Expected significance for two cut analyses and three multivariate analyses for different Higgs masses and final state topologies. (physics/0402030)

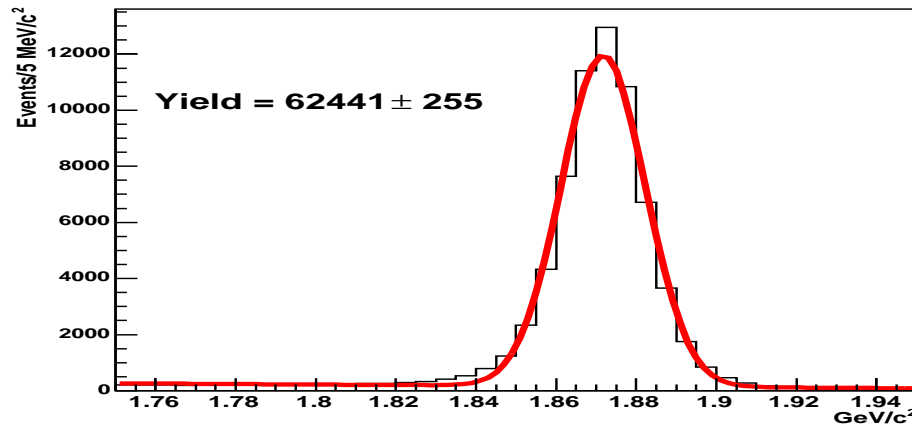
Applications of Genetic Programming



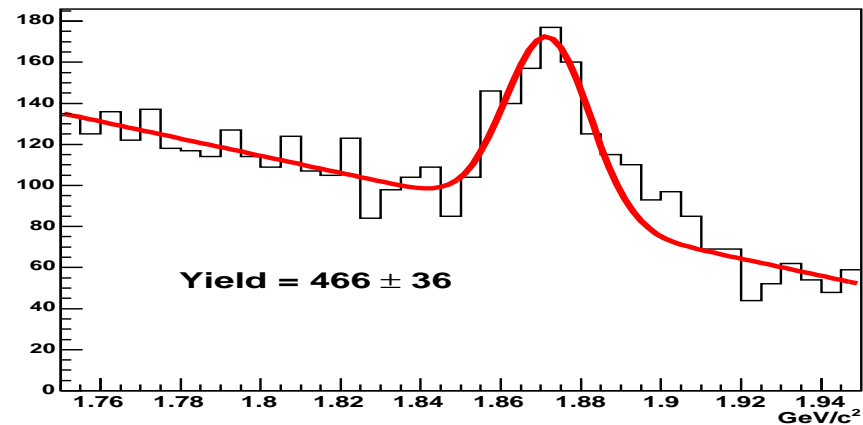
The FOCUS collaboration has recently used Genetic Programming to study doubly Cabibbo suppressed decay of $D^+ \rightarrow K^+ \pi^+ \pi^-$ relative to Cabibbo favored $D^+ \rightarrow K^- \pi^+ \pi^+$

hep-ex/0503007

a) Selected CF



b) Selected DCS



A Multivariate Analysis Scorecard

	Neural Nets	SVM	PDE	Gen. Prog.	Cuts
“multivariate-ness”	++++	++++	+++	++++	+
Speed: Training	Slow	Slow	N/A	Slow	Manual
Speed: Evaluation	Fast	Fast	Slow	Fast	Fast
Constructive	No	No	Yes	No	No
Uniqueness	No	Yes	Yes	No	No
Direct / Indirect	Indirect	Indirect	Indirect	Direct	Indirect

	PDE-RS	Gen. Algs	Likelihood
“multivariate-ness”	+++	+	++
Speed: Training	Fast	Slow	Manual
Speed: Evaluation	Fast	Fast	Fast
Constructive	Yes	No	Yes
Uniqueness	Yes	No	Yes
Direct / Indirect	Indirect	Direct	Indirect

Statistical Learning Theory

*When solving a given problem,
try to avoid solving a more general problem as an intermediate step.*

-V.N. Vapnik

Statistical Learning Theory is a fairly new field which:

Is general enough to encompass all our multivariate algorithms

Sheds light on issues of over-training & # training samples needed

Summarized by Occam's Razor:

Pluralitas non est ponenda sine neccesitate

Multivariate Algorithms / Learning Machines
are essentially Black Boxes with some parameters.

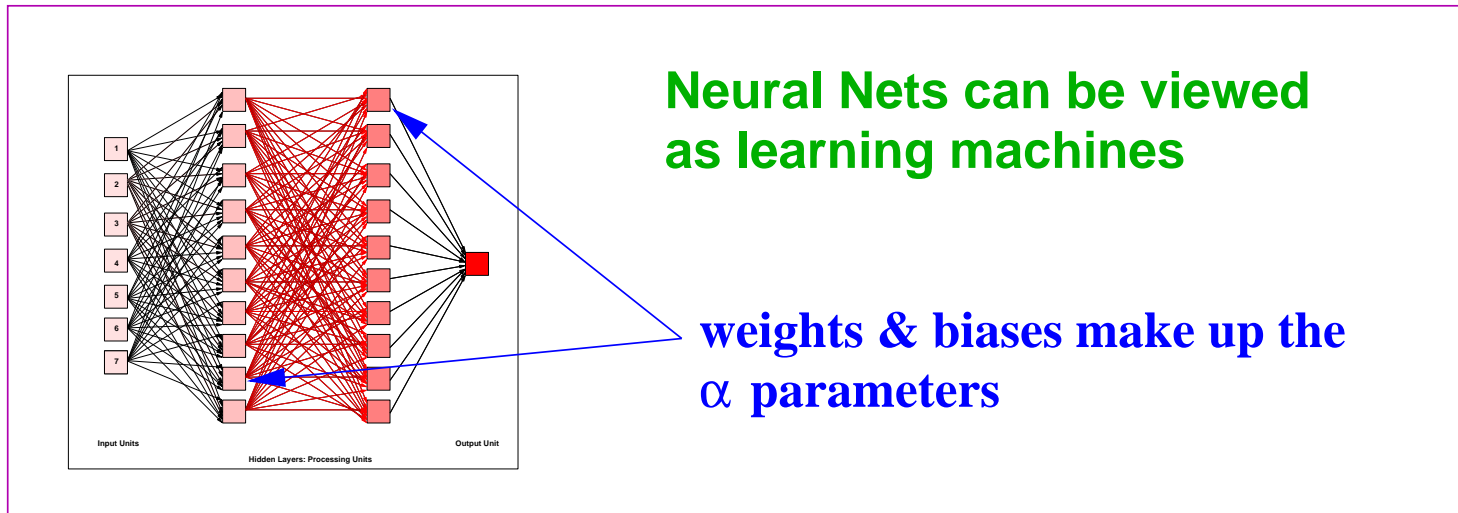
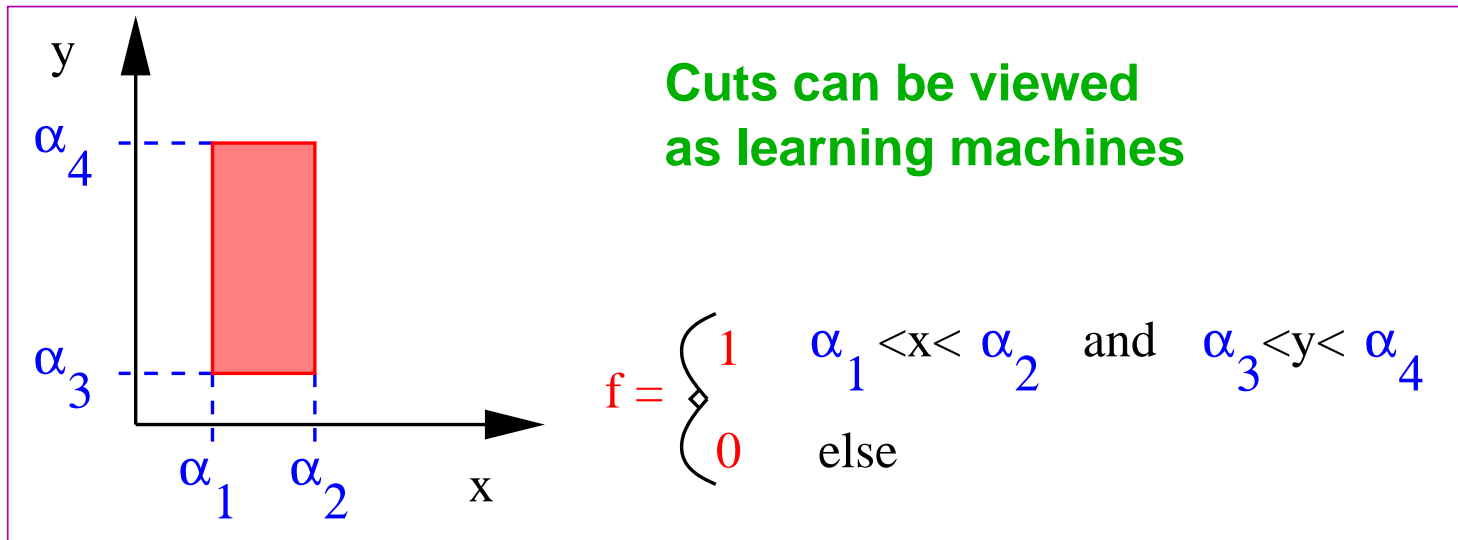
Formally, a learning machine looks like a family of functions
from an input space I to an output space O ,
each specified by some parameters α .

$$f(x \in I; \alpha) = y \in O$$

Training Data is a set of pairs $\{x_i, y_i\}$

The way in which the function's parameters are determined from
training data is associated *learning*.

An Example



Goal of Learning = minimizing some notion of Risk.

$$R(\alpha) = \int Q(x, y; \alpha) p(x, y) dx dy$$

- Use different $Q(x, y; \alpha)$ for different problems
- Note: in general we don't know $p(x, y)$.

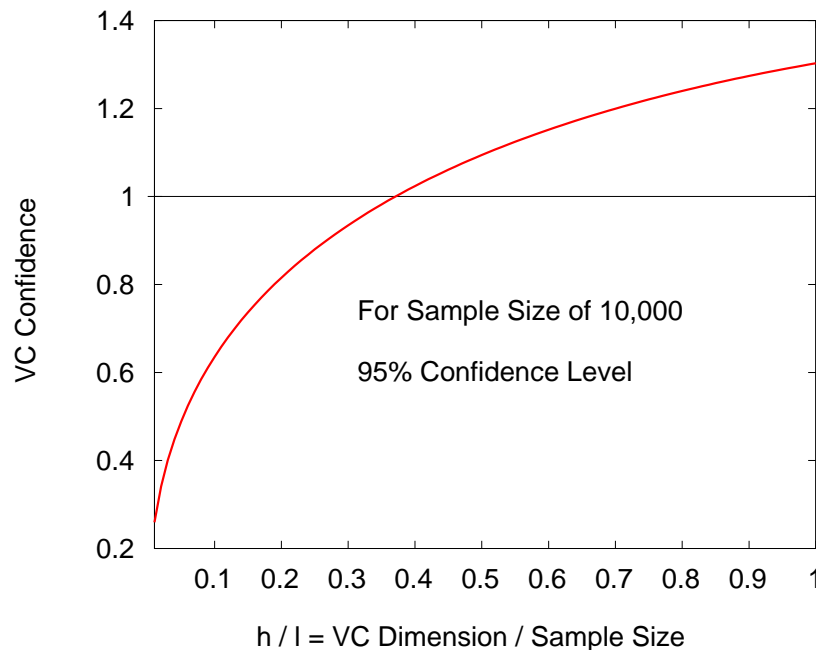
In practice, we only have the *Empirical Risk*

$$R_{\text{emp}}(\alpha) = \sum_{i=1}^l Q(x_i, y_i; \alpha).$$

Bounds on Risk

Surprisingly, there are general bounds on the generalization performance of a Pattern Recognition Algorithm, given by:

$$R(\alpha) = \int \frac{1}{2} |y - f(x; \alpha)| p(x, y) dx dy$$
$$\leq R_{\text{emp}}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l} \right)}$$



$h \rightarrow$ the Vapnik Chervonenkis (VC) dimension

$l \rightarrow$ the sample size

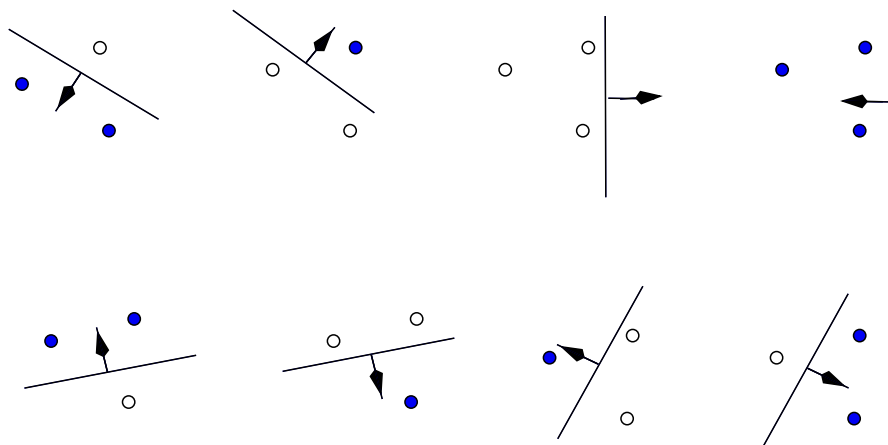
$1 - \eta \rightarrow$ the confidence the bound holds.

Independent of $p(x, y)$!

Vapnik Chervonenkis Dimension

The VC dimension h is equal to the maximal number of points that can be *shattered* by the learning machine $f(x; \alpha)$.

“A set $\{x_i\}$ is shattered by $f(x; \alpha)$ ” means that for every permutation of classifications $\{x_i, y_i\}$, there is an α such that $f(x_i; \alpha) = y_i$.



Examples:

An oriented line can shatter
3 points in \mathbb{R}^2

A Hyperplane can shatter
 $d + 1$ points in \mathbb{R}^d

Note: Not every set of h elements must be shattered by $f(x; \alpha)$, but just one.

Significance of VC Dimension

Suggests: $(\# \text{ Training Samples}) > (20 \times \text{VC dim})$

Higher VC dimension \rightarrow Higher Generalization Capacity \rightarrow Higher Risk

The Risk bound essentially describes potential for over-training.

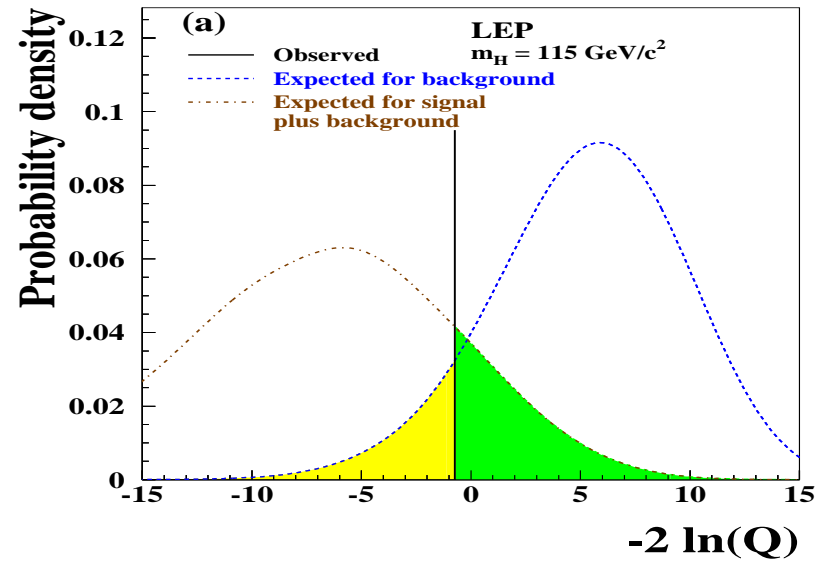
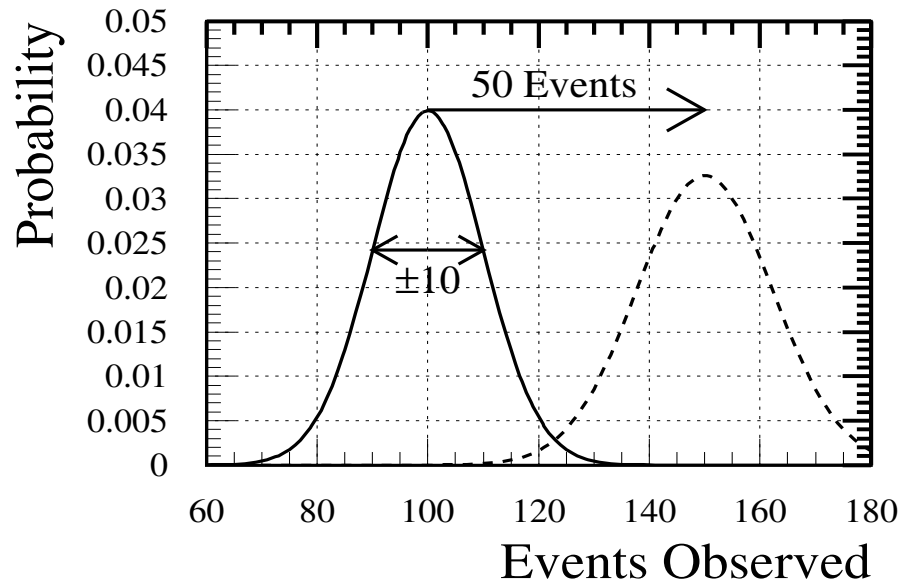
Tighter bounds are possible with an independent testing set.

Algorithm	VC Dim	Equivalent # Training Samples
cuts (7-d)	$h = 14$	$\approx 1,000$
Genetic Programming	$h \approx 100$	$\approx 7,000$
NN (7-10-10-1)	$400 \lesssim h < 1.6 \cdot 10^6$	$\approx 25,000$

Multivariate Analysis vs. Event Weighting

Multivariate Analysis vs. Event Weighting

The most powerful search technique considers likelihood of experiment

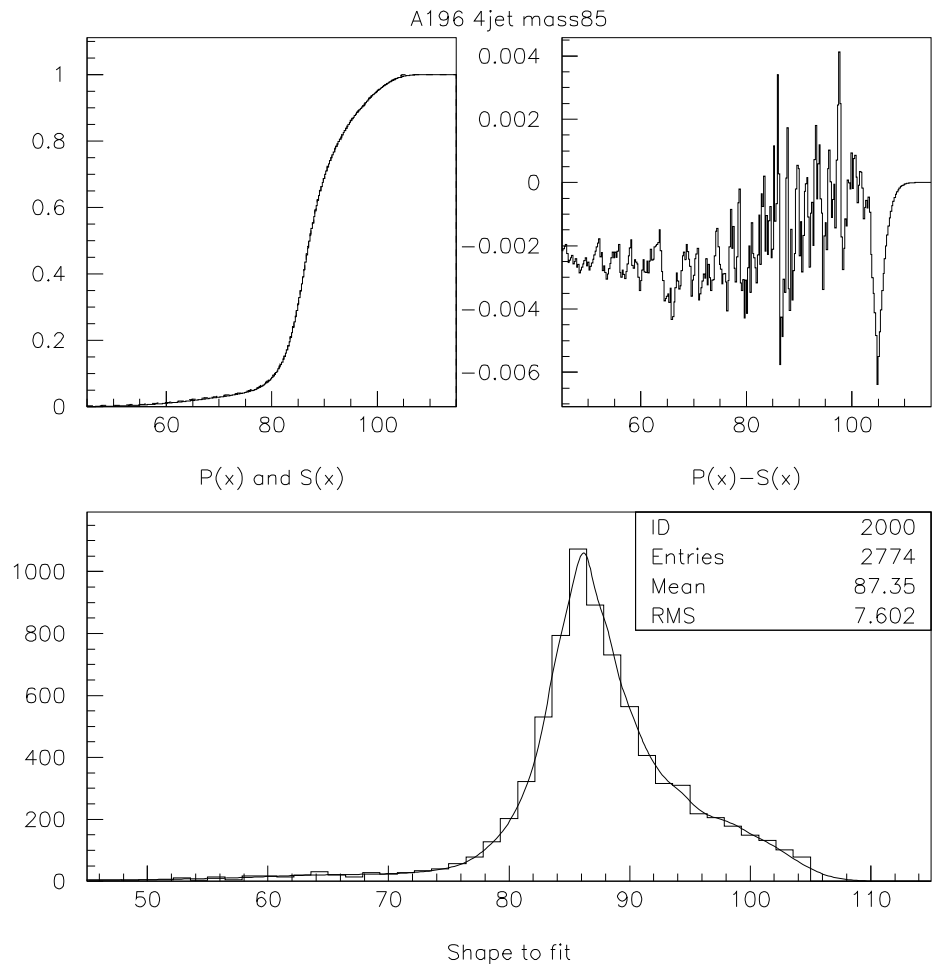


Not all events are equal.

LEP Higgs used likelihood ratio as a *test statistic*

KEYS

I was only peripherally involved in the LEP Higgs effort, but I did make one contribution: KEYS.



Each event was given a weight $\log\left(1 + \frac{sf_s(x)}{bf_b(x)}\right)$

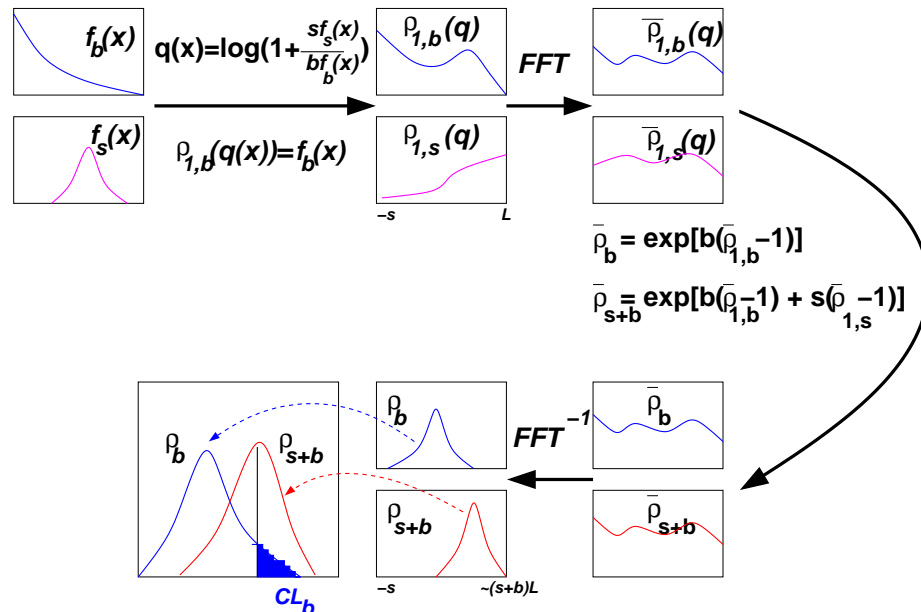
There were problems with estimating the Probability Density Functions $f_s(x)$ and $f_b(x)$

I used Kernel Estimation techniques to solve the problem

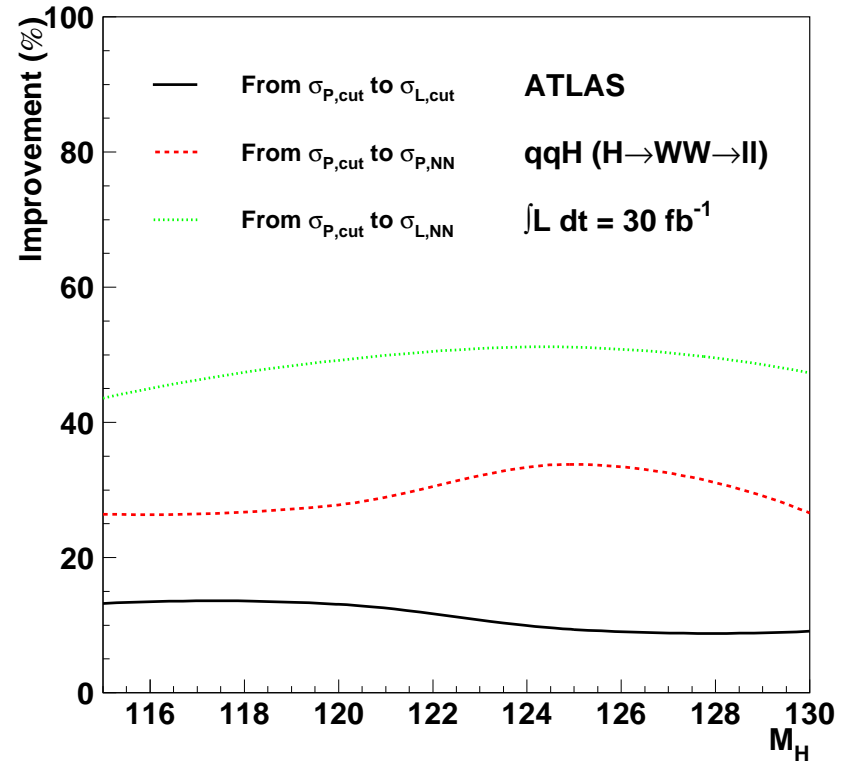
KEYS was used by LEP Higgs working group and BaBar

Migrating LEP Statistics to the LHC

LEP Higgs Working group developed formalism to combine channels and take advantage of discriminating variables



At the LHC this is numerically quite tricky



Incorporating Systematics

Why not use Bayesian Techniques?



Archbishop of Canterbury Thomas Cranmer
born: 1489 executed: 21 March 1556
author of the "Book of Common Prayer"



Two centuries later, (when this Book had become an official prayer book of the Church of England) Thomas Bayes was a Non-conformist minister (Presbyterian) who refused to use Cranmer's book.

Therefore, I prefer Frequentist Methods.

Have “Compound Hypothesis” $L(x|\mu)$ where μ is an unknown physics parameter and μ_t is the “true value”

No notion of $L(\mu)$ – that’s Bayesian!

Frequentist ask for a “Confidence Interval”:
an interval of μ which will contain μ_t with probability $1 - \alpha$.

The Interval is not unique!

The *ordering rule* defines the interval

Feldman & Cousins “Unified Approach” looks like this:

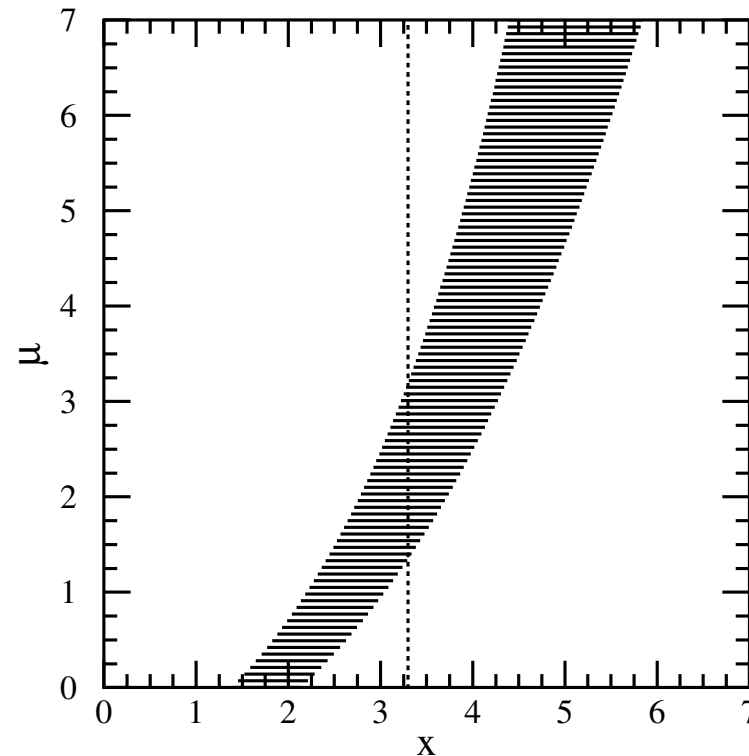
Neyman Construction

- For each μ : find region R_μ with probability $1 - \alpha$
- Confidence Interval includes all μ consistent with observation at x_0

Ordering Rule specifies what region

F-C ordering rule is the Likelihood Ratio

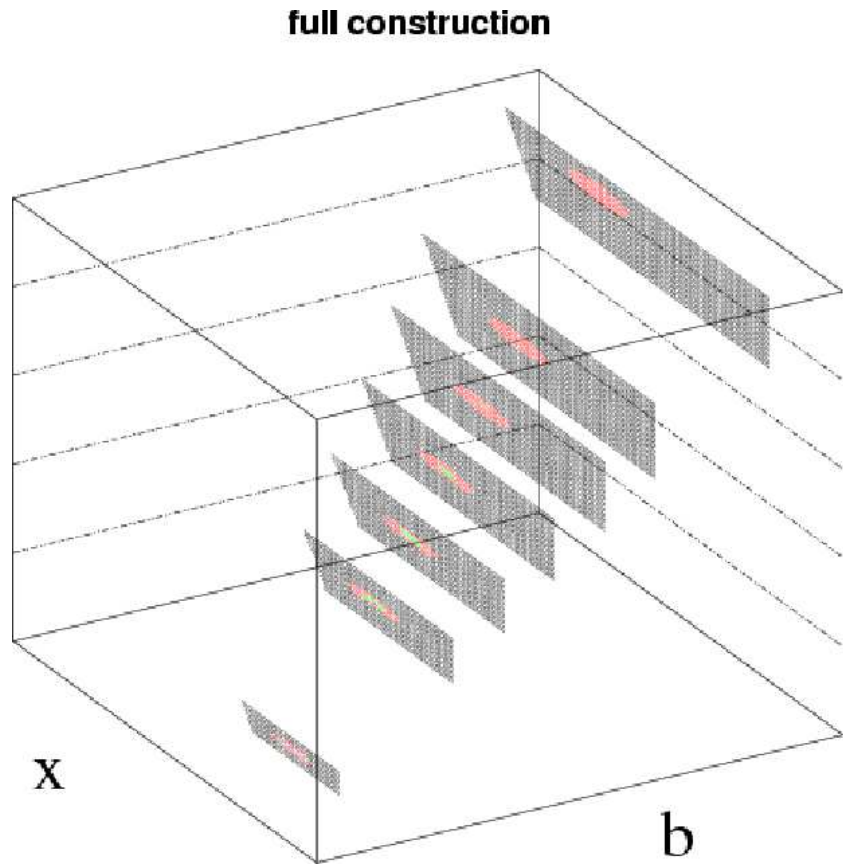
$$R_\mu = \left\{ x \mid \frac{L(x|\mu)}{L(x|\mu_{\text{best}})} > k_\alpha \right\}$$



Phys. Rev. D57:3873 (1998)

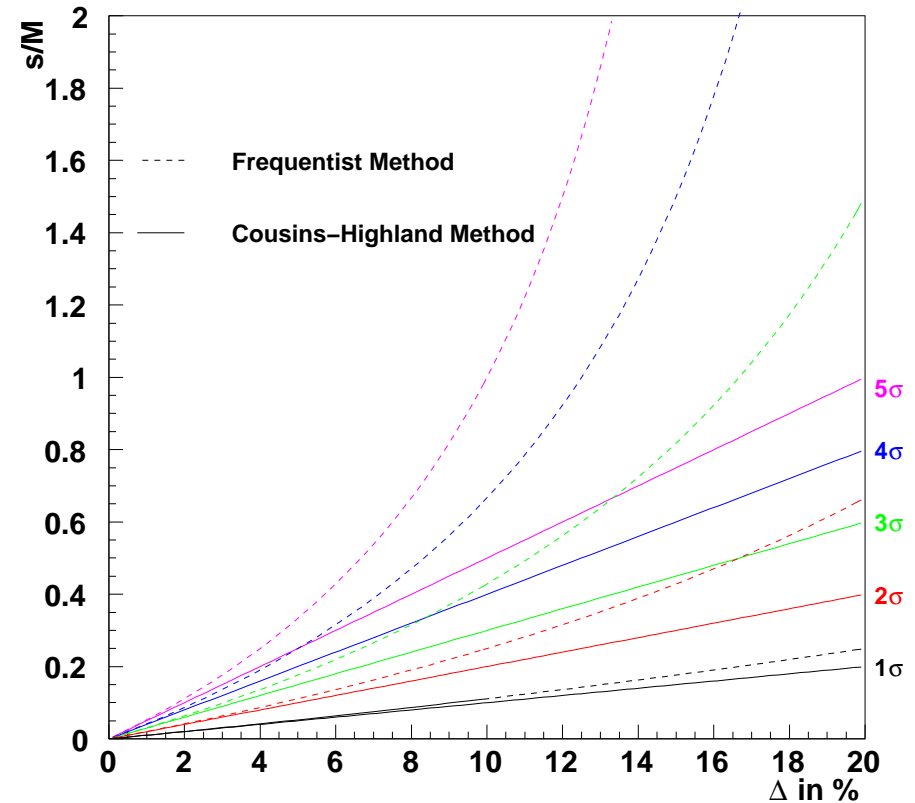
The F-C ordering rule follows naturally from Neyman-Pearson Lemma

Frequentist Hypothesis Testing with Background Uncertainty



At LEP, uncertainty on background included via smearing

That requires an implicit prior on background rate b



I developed a Frequentist technique based on Neyman Construction and side-band measurements

Big difference when background uncertainty is high. (physics/0310108)

Model Dependent vs. Model Independent

The fundamental difference for Model Independent searches is the lack of an alternate hypothesis H_1 .

Which Implies:

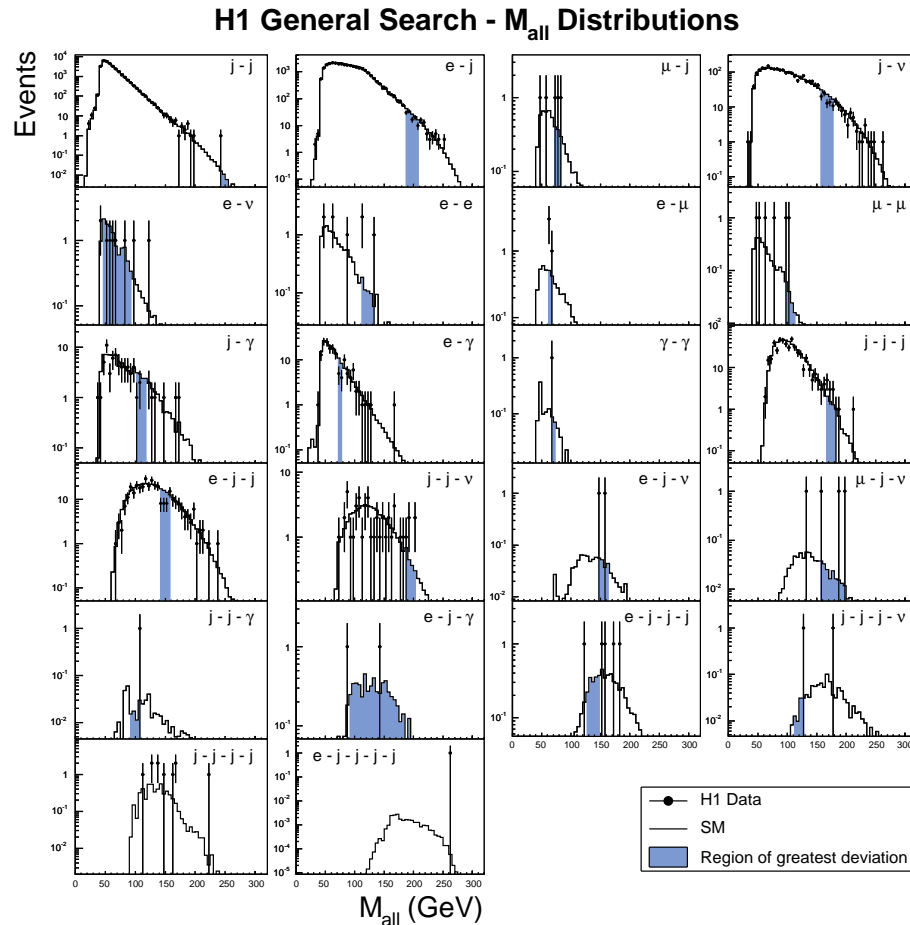
- ⇒ No notion of Type II error
- ⇒ No Neyman-Pearson lemma
- ⇒ No unique choice of acceptance region W .

There are two popular model independent approaches now:

- Data-driven “bump hunters” that are careful with statistics
- Signature-based searches custom to physics expectations

Model Independent: Bump Hunters

The H1 General Search and SLEUTH are both “bump hunters” with statistically meaningful results



Look in data for region with biggest discrepancy from data.

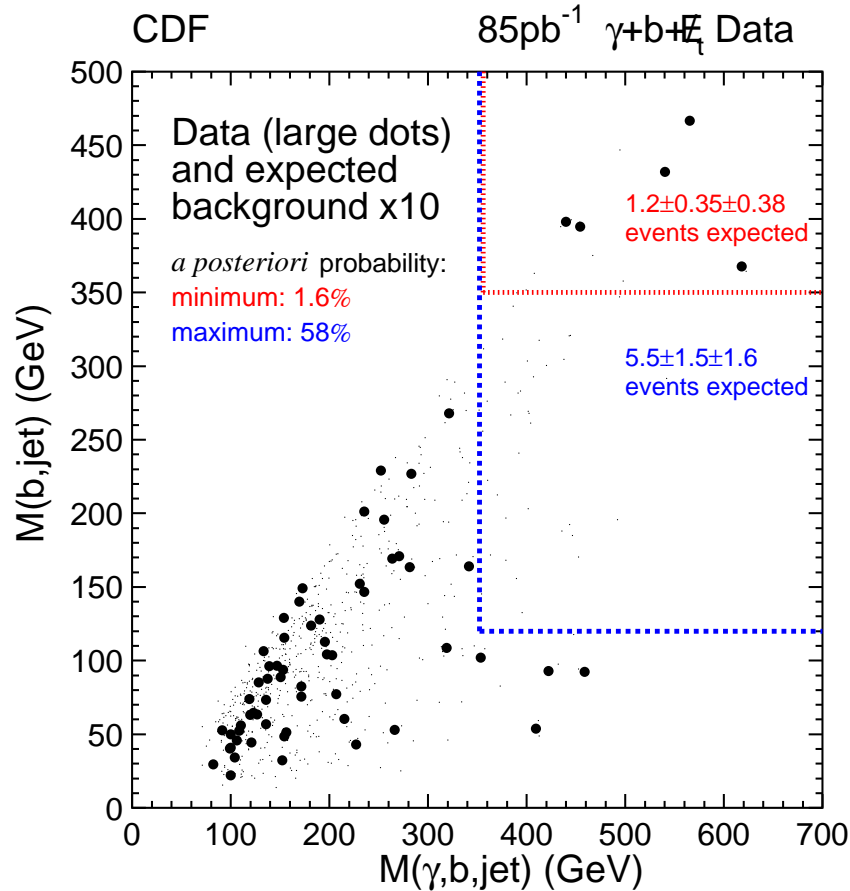
Repeat many toy experiments based on Standard Model predictions to estimate chance of a discrepancy of that size.

See

- $D\bar{D}$ (hep-ex/0006011)
- H1 (hep-ex/0408044)

Model Independent: Signature-Based

Signature-Based Searches fix region based on some physics expectation



Choice of search region is not data driven

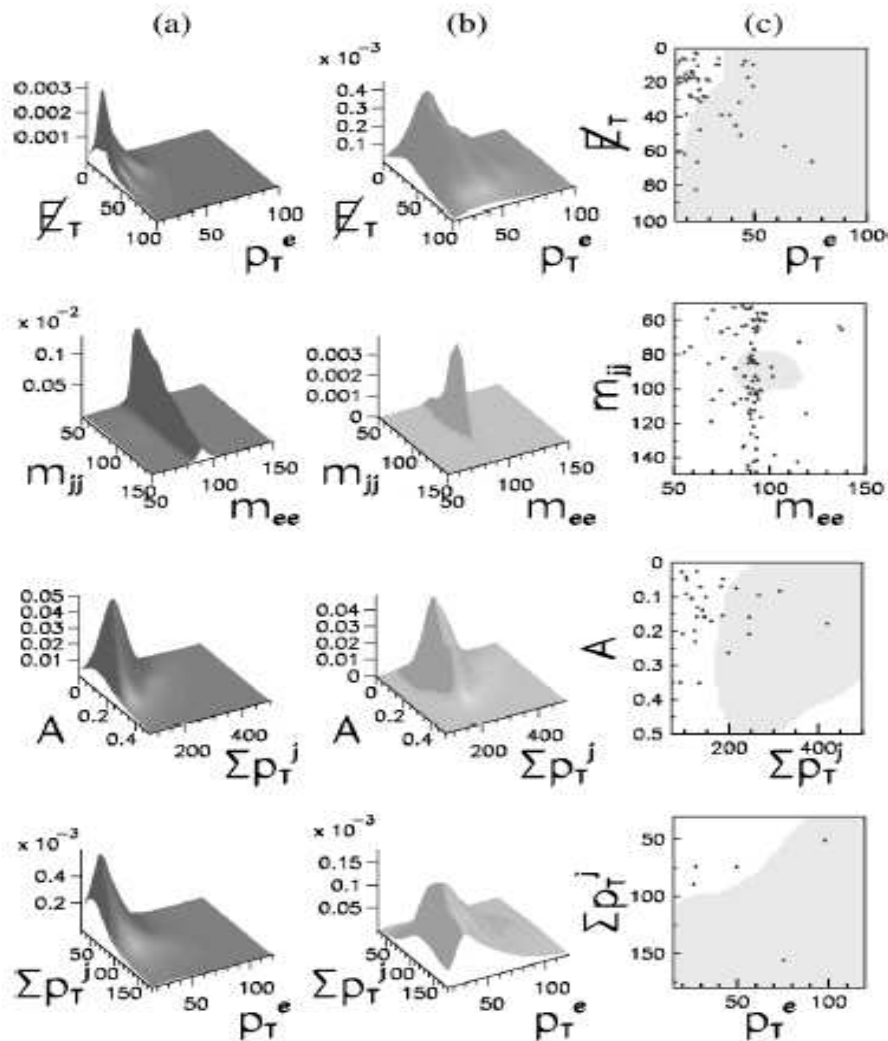
There is a notion of significance, but no notion of power

Need to be careful about the coupling of significance and the number of searches are performed

See

- Searches for New Physics in Events with a Photon and b -quark Jet at CDF (2001)

Automated Methods



Bruce Knuteson has developed an automated analysis procedure called QUAERO.

VISTA is a related tool for comparing data to Standard Model predictions

$D\emptyset$ results published in *Phys. Rev. Lett.* **87**, 231801 (2001)

Given signal and background Monte Carlo, QUAERO constructs a set of cuts tailored to the signal.

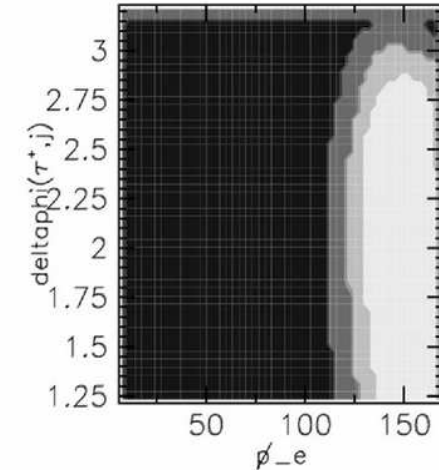
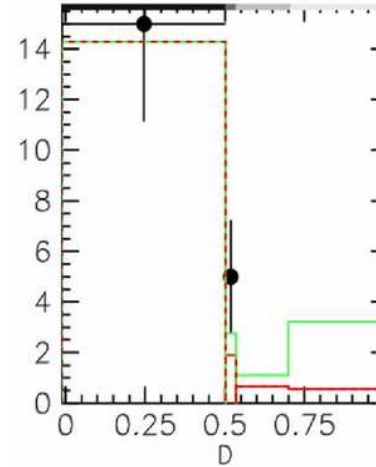
I applied this technique to ALEPH's LEP2 data in order to assess the method for the LHC environment

QUAERO does not attempt to optimize reconstruction and is limited to a pre-defined set of kinematic variables for discrimination

Relevant Question: Is QUAERO's analysis procedure powerful enough?

After a user provides signal events, QUAERO uses 0-3 of the most powerful variables to construct

$$D(x) = \frac{f_s(x)}{f_s(x) + f_b(x)}$$



Contours of $D(x)$ define bins used to calculate the likelihood ratio Q , which is QUAERO's final result

$$Q = \frac{L(\mathcal{D}|H_1)}{L(\mathcal{D}|H_0)}$$

Interpretation of QUAERO's Results

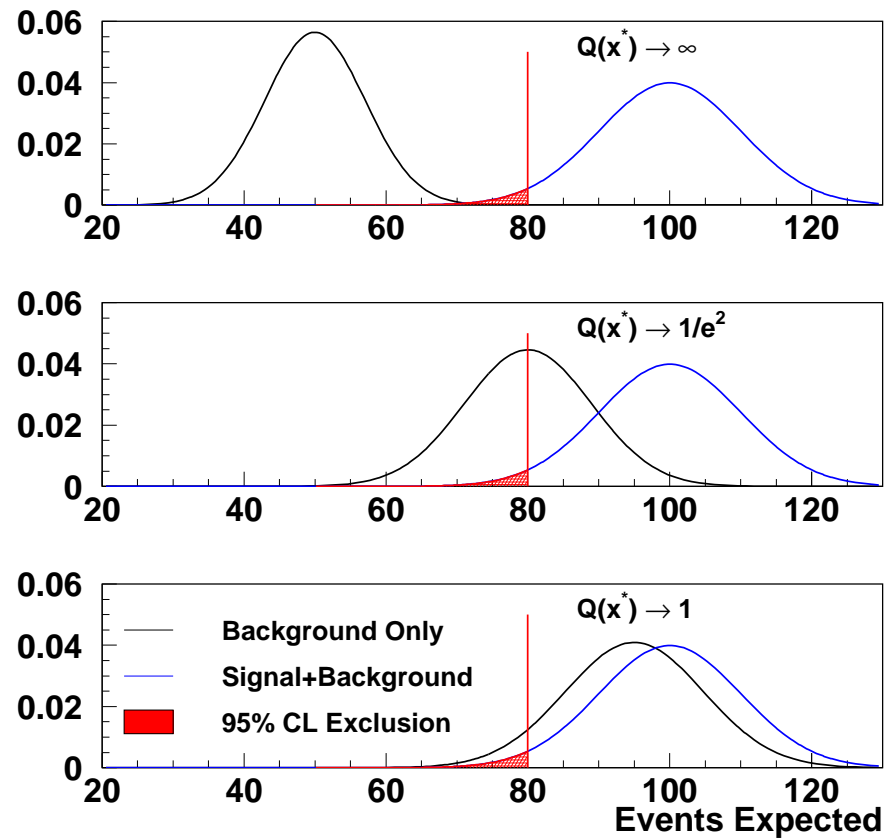
QUAERO result is the likelihood ratio

$$Q = \frac{L(\mathcal{D}|H_1)}{L(\mathcal{D}|H_0)}$$

Can't translate to 95% exclusion

Depending on s/b , the likelihood ratio at exclusion threshold may be large or small

The Likelihood ratio does make sense in a Bayesian framework



Nice Features:

- ❖ automatically spans final states & combines results
- ❖ tuned for the model point in question
- ❖ systematic errors and can be incorporated

Requires:

- ❖ sample of *reconstructed* signal events
- ❖ a reasonable sample of background events in the signal-like region

Biggest Challenges:

- ❖ creating a general-purpose background Monte Carlo (that you believe)
- ❖ creating a fast-simulation for the signal events (that you believe)
- ❖ estimating systematics for the entire phase-space (that you believe)

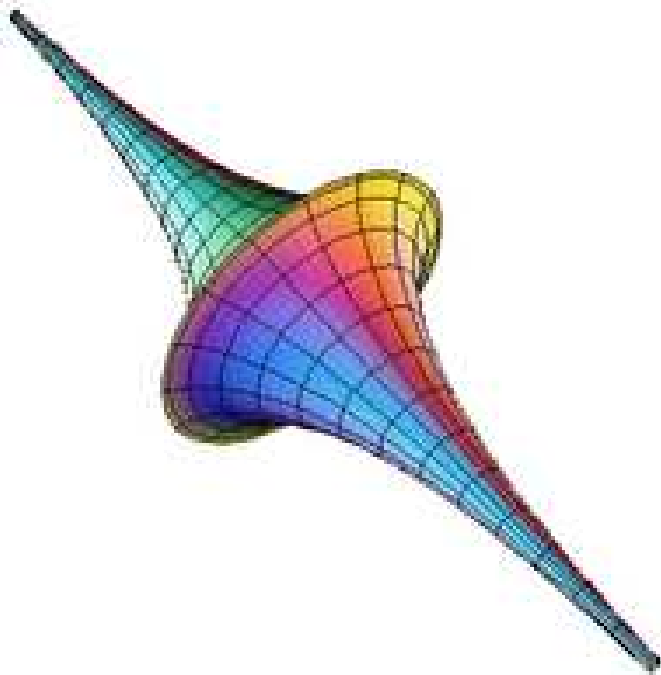
Ways to Improve on the Current QUAERO Implementation:

- ❖ streamline event format and storage, think of more extensible event data
- ❖ more modular structure (identify variables → optimize cuts → final result)
- ❖ high-level API for automated QUAERO submissions and result processing

Where Things May Be Going

Amari considered the *Fisher Information Matrix* g_{ij} as a metric on a *Manifold* M parametrized by α :

$$g_{ij}(\alpha) = \int dx f_{\alpha}(x) \left[\frac{\partial \log f_{\alpha}(x)}{\partial \alpha_i} \right] \left[\frac{\partial \log f_{\alpha}(x)}{\partial \alpha_j} \right]$$



Example:

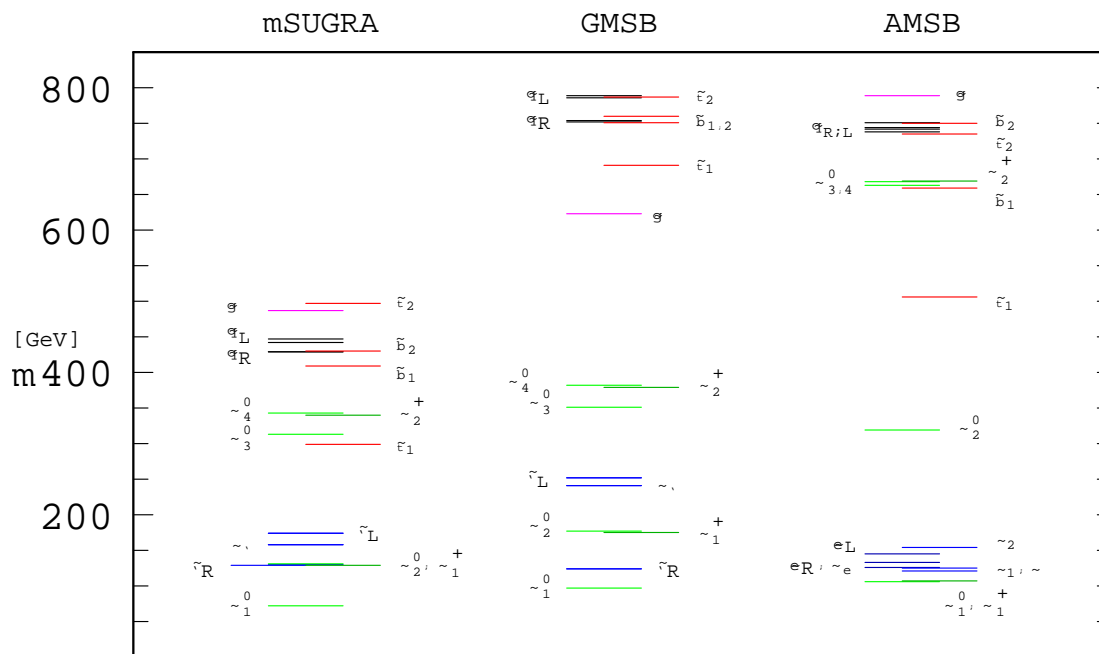
Consider Gaussians $G(x; \mu, \sigma)$ as a 2-d Manifold parametrized by $\alpha = (\mu, \sigma)$

the geometry is isotropic and negatively curved

Natural Learning Rules correspond to *geodesics* on the Manifold M .

Can lead to exponentially faster rates of convergence!

Information Geometry of MSSM



We could try to use Information Geometry to improve how we sample the model space

An example use of Information Geometry for the MSSM:

- ❖ $\alpha = 105$ model parameters
- ❖ $x =$ measured mass spectrum
- ❖ $f_\alpha(x) =$ probability to measure that spectrum given model parameters

Neyman-Pearson lemma gives us *the most powerful* search strategy.

The problem for experimentalists is we don't know $L(x|H_0)$ & $L(x|H_1)$. It's a convolution of QFT with detector

But theorists do know $L(x|H_0)$ (at least to leading order)

Can analytically calculate an upper limit on the expected significance of a hypothesized particle.

$$q(x) = \ln \left(\frac{L(x|H_1)}{L(x|H_0)} \right) = \ln \left(1 + \frac{|\mathcal{M}_H|^2 \cdot d\text{LIPS}}{|\mathcal{M}_Z|^2 \cdot d\text{LIPS}} \right)_x$$

$$\rho_{1,s}(q_0) = \frac{1}{\sigma_H} \int_x d\text{LIPS} |\mathcal{M}_H|^2 \cdot \delta(q_0 - q(x))$$

$$\rho_{1,b}(q_0) = \frac{1}{\sigma_Z} \int_x d\text{LIPS} |\mathcal{M}_Z|^2 \cdot \delta(q_0 - q(x))$$

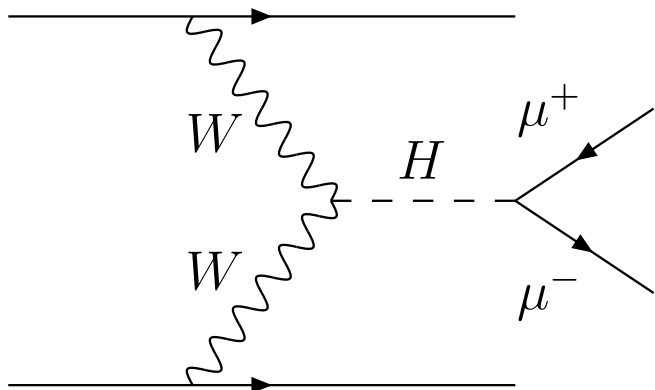
$$\rho_{N,i}(q) = \underbrace{\rho_{N,i}(q) \oplus \dots \oplus \rho_{N,i}(q)}_{N \text{ times}} = \mathcal{F}^{-1} \left\{ [\mathcal{F}(\rho_{1,i})]^N \right\}$$

$$\rho_i(q) = \sum_{N=0}^{\infty} P(N; L\sigma_i) \cdot \rho_{N,i}(q) = \mathcal{F}^{-1} \left\{ e^{L\sigma_i [\mathcal{F}(\rho_{1,i}(q)) - 1]} \right\}$$

A theoretical example: $VBF H \rightarrow \mu\mu$

I'm working with Tilman Plehn to re-use MC used in hep-ph/0107180

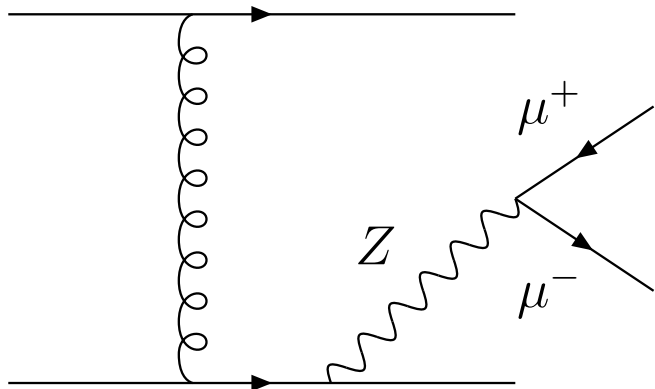
With basic cuts, only need to consider signal and irreducible backgrounds



Phase Space:

2	for incoming quarks
$+(3 \times 4)$	for outgoing fermions
-4	for 4-momentum conservation
<hr/>	
10	phase space dimensions

All other observables are a function of these. There is more information available.



Changed Higgs width to 2.4 GeV to simulate mass resolution

Preliminary result:

- Original cuts give 1.8σ after 300 fb^{-1}
- Upper-bound on significance $\sim 6\sigma$ after 300 fb^{-1}

Conclusions

Summary

I have reviewed several multivariate algorithms used within HEP

- distinguished between direct / indirect methods
- introduced new genetic programming approach
- provided a scorecard for multivariate techniques
- established relationship to Statistical Learning Theory

I have discussed developments in Frequentist techniques

- using event weighting / combining results
- incorporating systematics

I have compared Model Dependent and Model Independent Searches

I have commented on Automated Methods (QUAERO) for the LHC

Conclusions

Several new multivariate analysis techniques on the market with complementary approaches.

Large systematic uncertainties at LHC will stress our advanced techniques

Model Independent search strategies have a big role at the LHC

Automated methods have most promise in a more restricted setting

There is a lot to gain from collaboration with phenomenologists

Backup Slides

Calculating the VC Dimension for Neural Networks

Use the following Theorems:

- 1) VCD for Boolean connections $VCD(b(f_1, \dots, f_k)) \leq c_k \max_i VCD(f_i)$
- 2) VCD for composition
- 3) Define $\rho = \#$ weights and biases in the network

Google for Eduardo D. Sontag's "VC dimension of neural networks"

Best known bounds $\rho^2 < h < \rho^4$

Shattered Sets with more than $\mu = 2\rho - 1$ elements are "special" (i.e. measure 0). Maybe μ is more practical.

For Learning Machines that form a Vector Space

$$\text{i.e. } \exists \alpha \ni f(x; \alpha) = a f(x; \alpha_1) + b f(x; \alpha_2)$$

The VC dimension is given by $h = \dim(\text{span}(\alpha))$

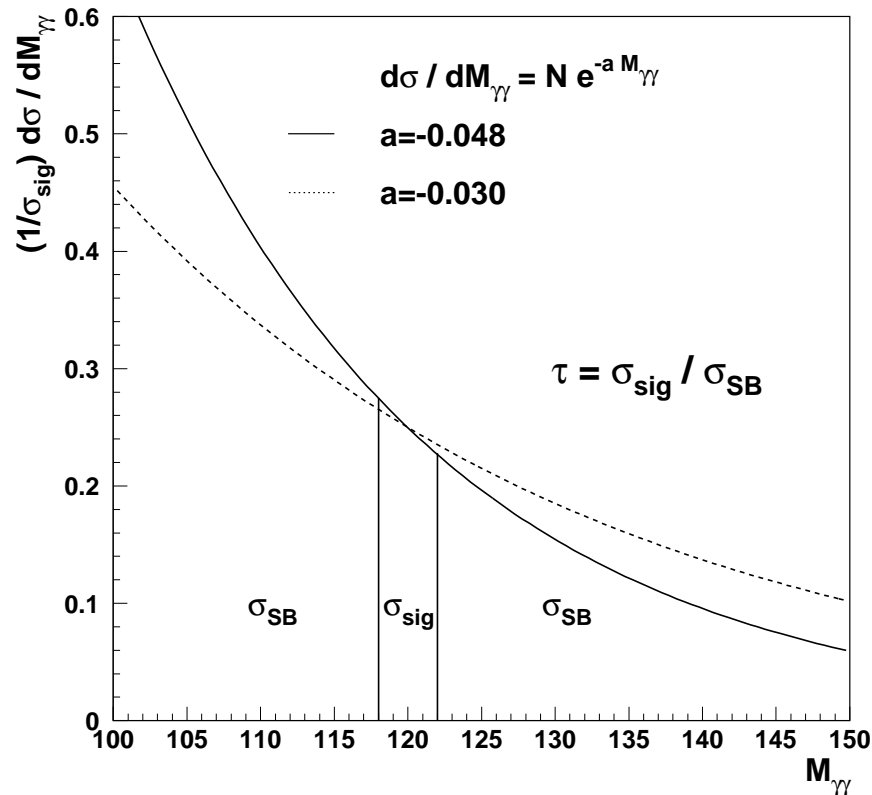
Genetic Programming approach looks like polynomials \rightarrow a vector space

Example: For a recent ATLAS Higgs Search with 7 variables we compared

Algorithm	VC Dim h	Significance
Cuts	$h = 14$	3.91σ
Neural Nets	$400 \lesssim h < 1.6 \cdot 10^6$	4.98σ
Genetic Programming	$h \approx 100$	4.57σ

Toy Monte Carlo Approach

The goal of this study was not to focus on the $M_{\gamma\gamma}$ spectrum, but to study the coverage of statistical methods.



We used $d\sigma/dM_{\gamma\gamma} = N e^{-a M_{\gamma\gamma}}$

We used a Toy MC to generate experiments with $\overline{N}_{\text{sig}} = 16000$ ($\approx 30\text{fb}^{-1}$).

We sampled the range $100 < M_{\gamma\gamma} < 150$ (about 200K events/exper).

We varied the exponent a in the range $[-.048, -.030]$, generating nearly 1M experiments per exponent tested.

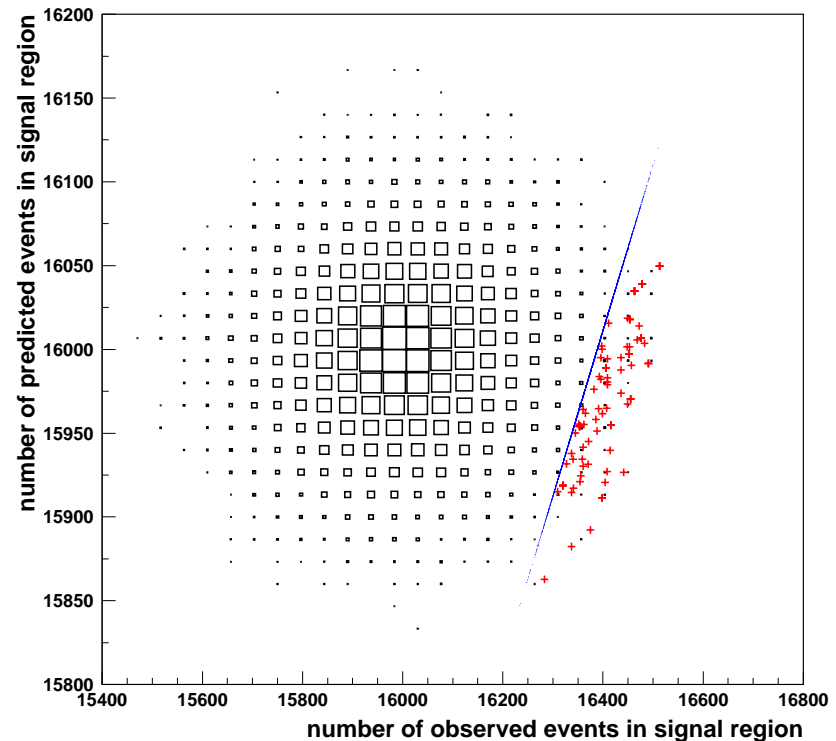
Systematic Uncertainty on Background Prediction

To include systematic errors, we must know the uncertainty on the background prediction.

So far, we have only guessed for $H \rightarrow \gamma\gamma$

After performing many fits, we find that $\Delta b \approx 38$ events or 0.25%.

A good guess for error is $\tau\sqrt{N_{SB}}$, which predicts 36 events!



Note systematic uncertainty depends on τ

The Cousins-Highland Method

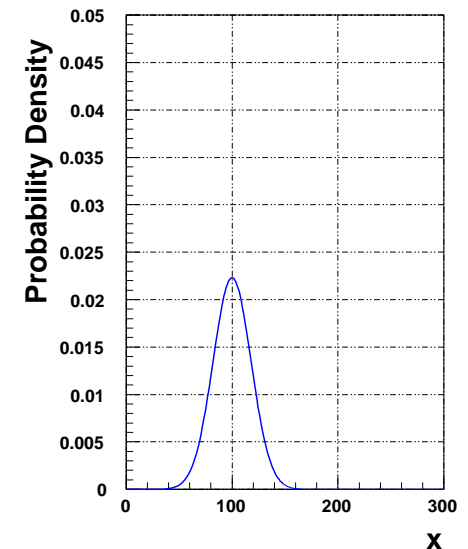
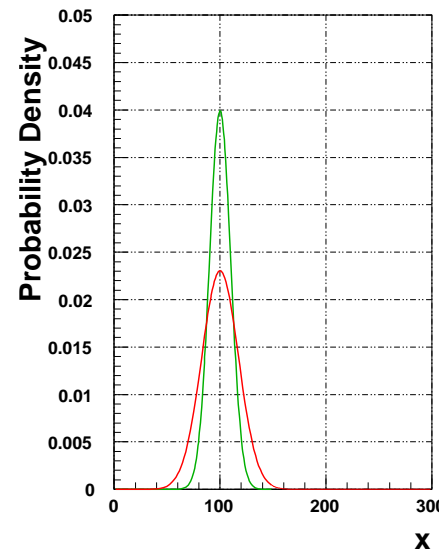
At LEP the Cousins-Highland Method was used for Systematics

The Cousins-Highland method
integrates-out b

$$L(x|H_0, M) = \int_b L(x|b)L(b|M)$$

But it uses a Bayesian notion $L(b)$

$$L(b|M) = \frac{L(M|b) L(b)}{L(M)}$$



For a 5% systematic uncertainty on background
and a measurement at $M_0 = 100$
(with an expectation of 50 signal events)

Must observe 167 events to claim a discovery

With the Cousins-Highland approach only need 161 events