W

un I: DØ ~20%, CDF ~53%
neasure the top quark mass
ching that of CDF by using
to separate signal and
nizing the correlation of
the selection with the top quark mass

W

DØ :

l neural networks (NN)

uction used NN

nt benefited from a multivariate

ent uses NN

ved measurement

is

o

experiments will be a lot more
vious in data handling at all stages

dure of choosing and applying cuts
time is rarely optimal!

multivariate, the optimal methods

ition multidimensional variable
etween classes of objects

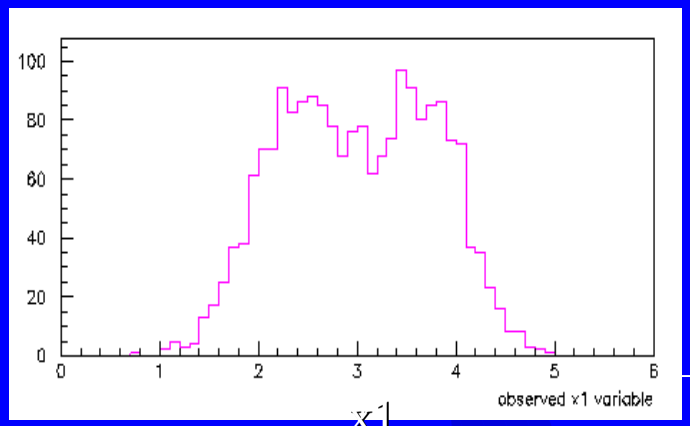
ects to groups based on

Data A

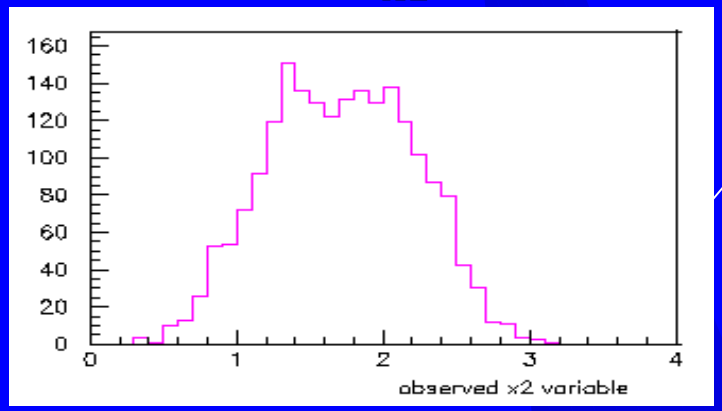
- Particle Identification
  - e-ID, τ-ID, b-ID, γ/π^0, q/g
- Signal/Background Event Classification
  - Signals of new physics are rare and small

    (Finding a "jewel" in a hay-stack)
- Parameter Estimation
  - t mass, H mass, track parameters, for example
- Function Approximation
  - Correction functions, tag rates, fake rates
- Data Exploration
  - Data-driven extraction of information, latent structure analysis

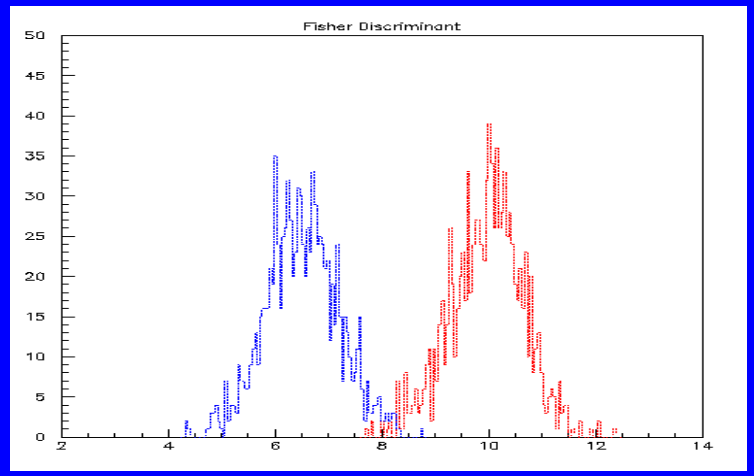
Why Multivariate Methods?



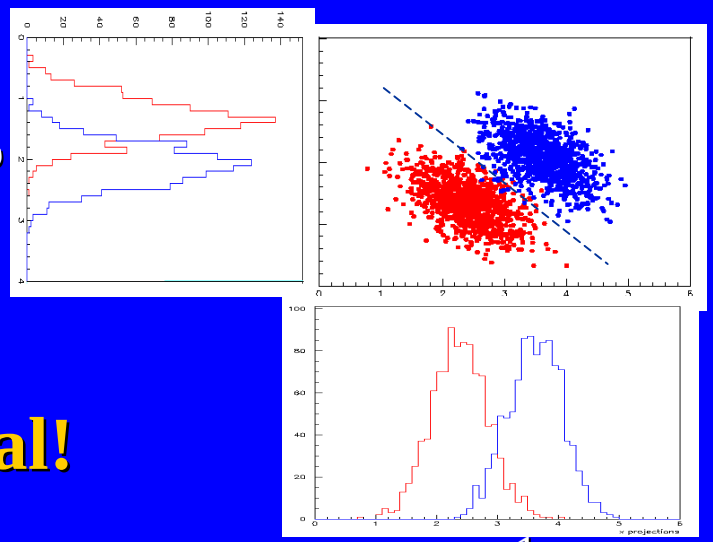
x1



x2



D(x1,x2)=2.014x1 + 1.592x2



x2

x1

- **Because they are optimal!**

Optimal Event Selection

$$r(\vec{x}) = \frac{p(s|\vec{x})}{p(b|\vec{x})} = \frac{p(x \mid s)p(s)}{p(x|b)p(b)}$$

defines decision boundaries that minimize the probability of misclassification

Posterior probability

$$p(s \mid x) = \frac{r}{1+r} = \frac{p(x \mid s)}{p(x|b) + p(x \mid s)}$$

So, the problem mathematically reduces to that of calculating *r(x)*, the Bayes Discriminant Function or probability densities

Probability Density Estimators

- Histogramming:

  The basic problem of non-parametric density estimation is very simple!

  - **Histogram** data in M bins in each of the d feature variables
  - M^d bins ← **Curse Of Dimensionality**
  - **In high dimensions, we would either require a huge number of data points or most of the bins would be empty leading to an estimated density of zero.**
  - **But, the variables are generally correlated and hence tend to be restricted to a sub-space ← Intrinsic Dimensionality**

K

but adopts importance sampling

ace a hypercube of side *h* centered o

x^n

N = Number of data points

$H(u)$ = 1 if x^n in the hypercube

= 0 otherwise

uities

ent forms for kernel functions

A

rnel approach

ion is a linear combination of

of basis functions or

s

$$p(j) \qquad M << N_{data}$$

parameters and p(x|j) a

multivariate kernel function. For example

K nearest-neighbor Method

- **Place a hyper-sphere centered at each data point *x* and allow the radius to grow to a volume *V* until it contains *K* data points. Then, density at *x***

$$p(x) = \frac{K}{NV}$$

N = Number of data points

- **If our data set contains N_k points in class C_k and *N* points in total, then**

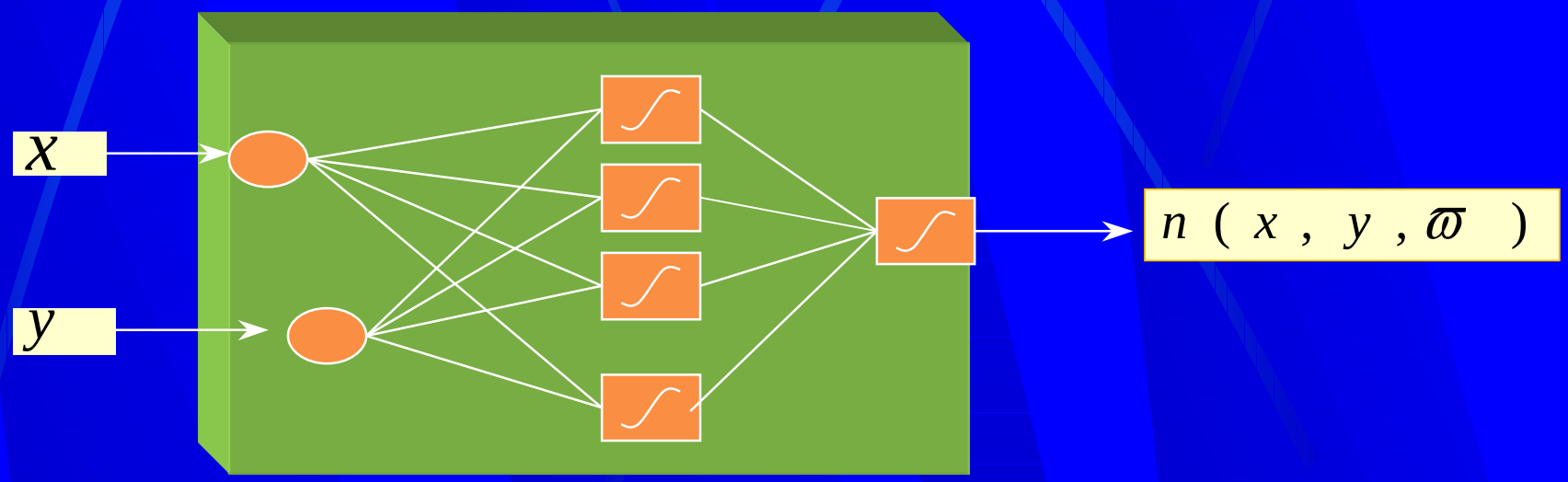
$$p(x \mid C_k) = \frac{K_k}{N_k V}$$

K_k = # of points in volume V for class C_k

$$p(C_k \mid x) = \frac{P(x \mid C_k)\, p(C_k)}{p(x)} = \frac{K_k}{K}$$

Discriminant Approximation with Neural Networks

Output of a feed forward neural network can approximate the Bayesian posterior probability *p(s|x,y)*
Directly without estimating class-conditional probabilities



$$n(x, y, \varpi) \approx p(s \mid x, y) = \frac{r}{1 + r}$$

Calculating the Discriminant

Consider the sum

$$E(x, y, \omega) = \sum_i [n(x_i, y_i, \omega) - d_i]^2$$

Where

d_i = 1 for signal
= 0 for background

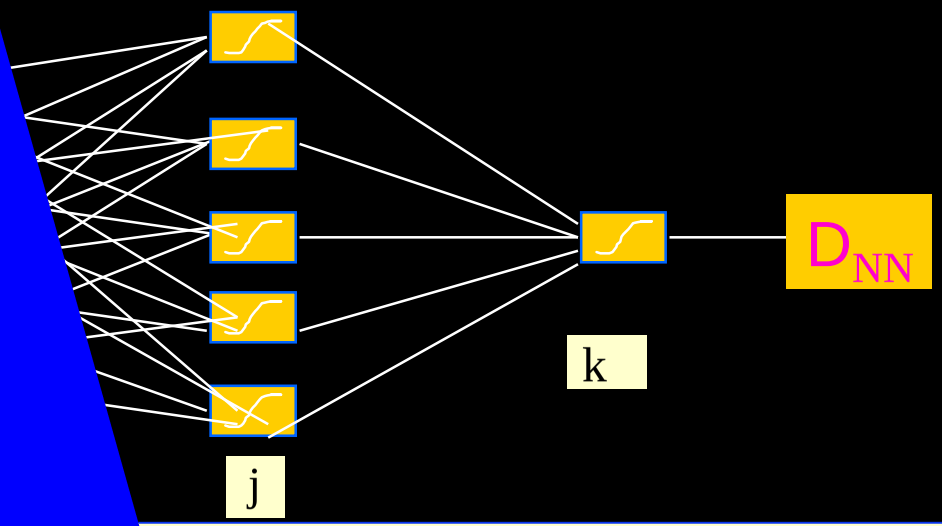
ω = vector of parameters

Then

$$\frac{dE(x, y, \omega)}{d\omega} = 0 \Rightarrow n(x, y, \varpi) = p(s \mid x, y) = \frac{r}{1 + r}$$

in the limit of large data samples and provided that the function *n(x,y,ω)* is flexible enough.

ral Networks



$_i$)};

$$g(a) = \frac{1}{1 + e^{-a}}$$

- ... mapping function without requiring a mathematical

W

universal approximators

ge NN, you can

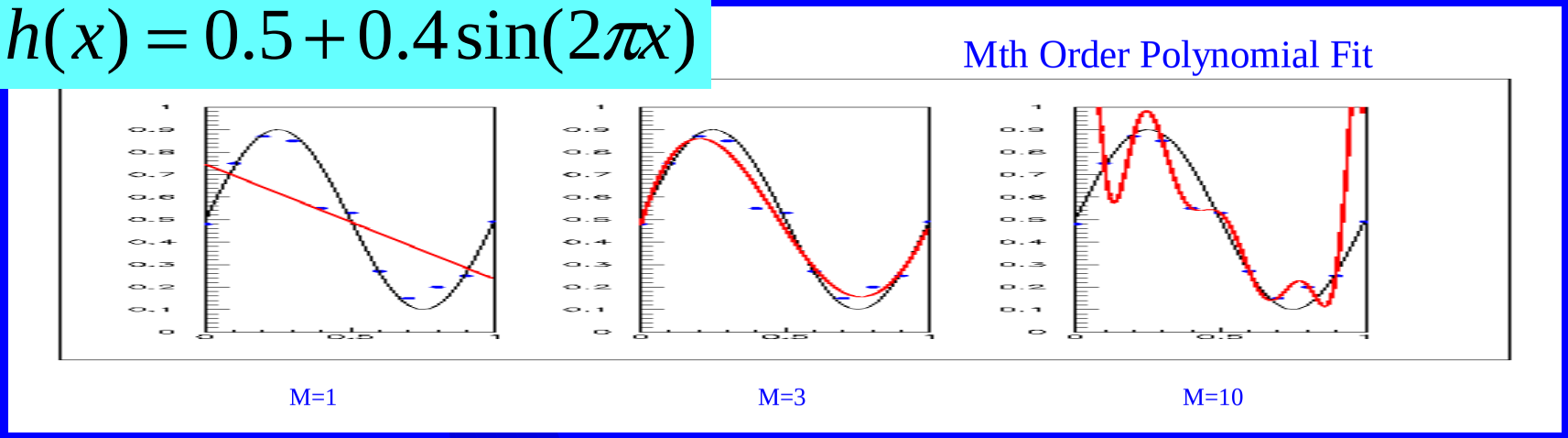
n to arbitrary accuracy

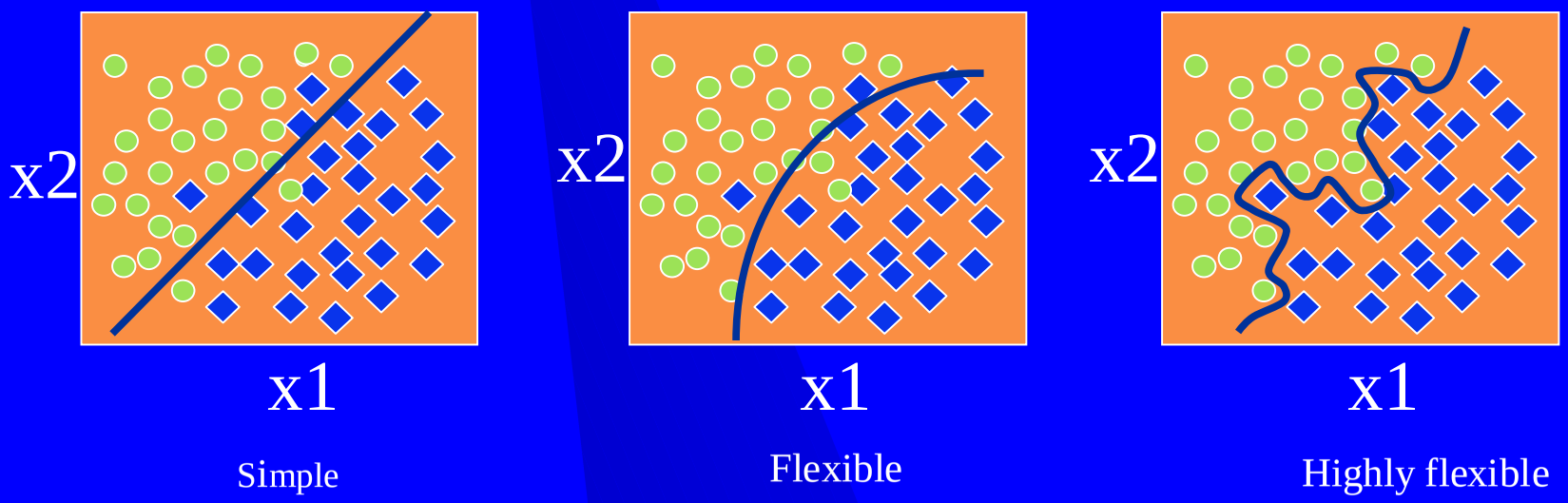
imation is rapid

- High dimensionality is not a curse any more!

- Also, they need to have optimal flexibility/complexity

$$h(x) = 0.5 + 0.4\sin(2\pi x)$$





Simple

Flexible

Highly flexible

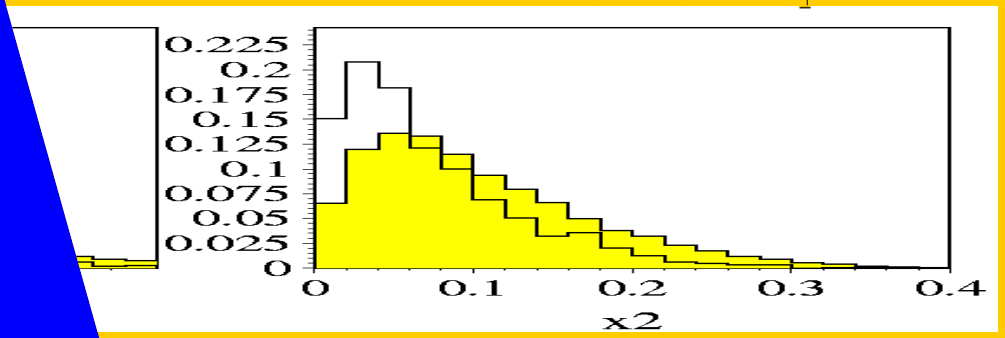
The Golden Rule

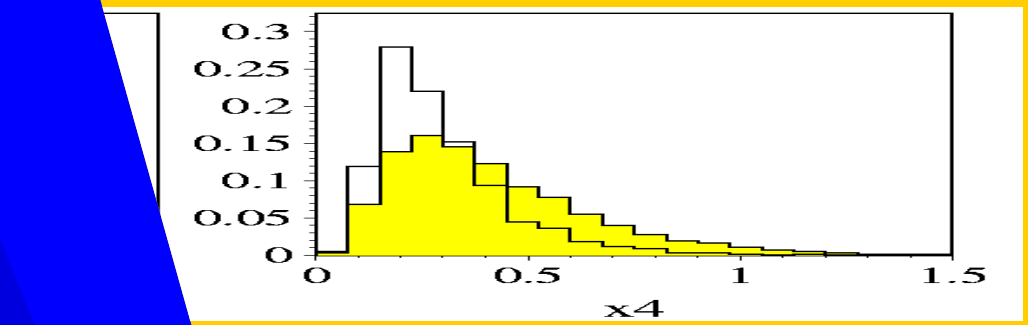
Keep it simple
As simple as possible
Not any simpler
- Einstein

M

inant variables

shaded = top

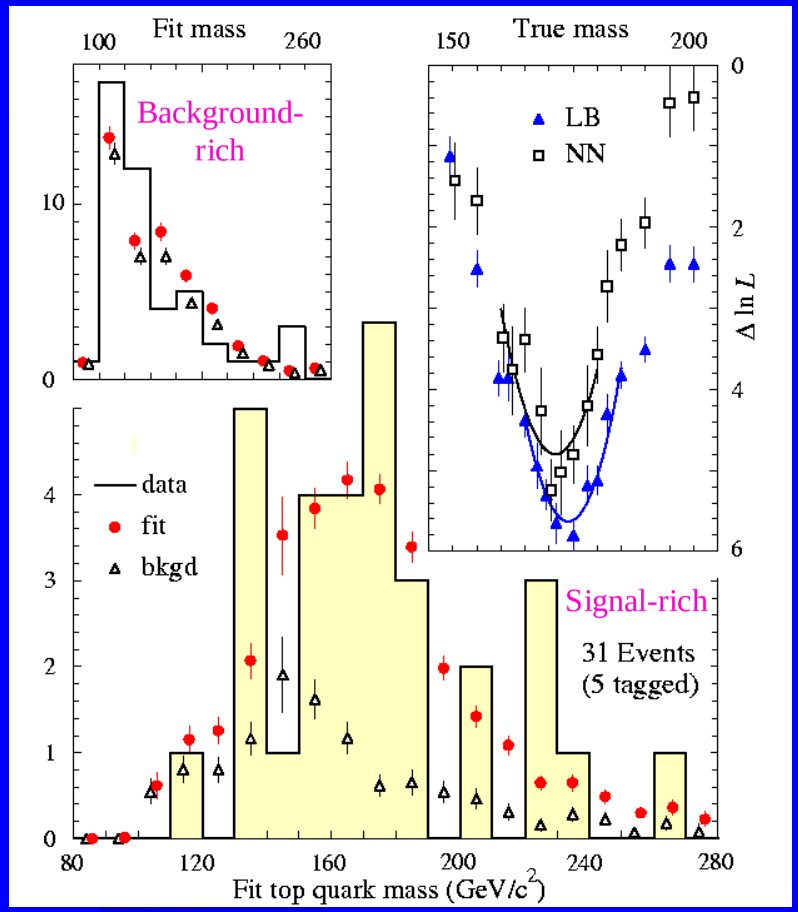




The Discriminants

Measuring the Top Quark Mass

DØ Lepton+jets



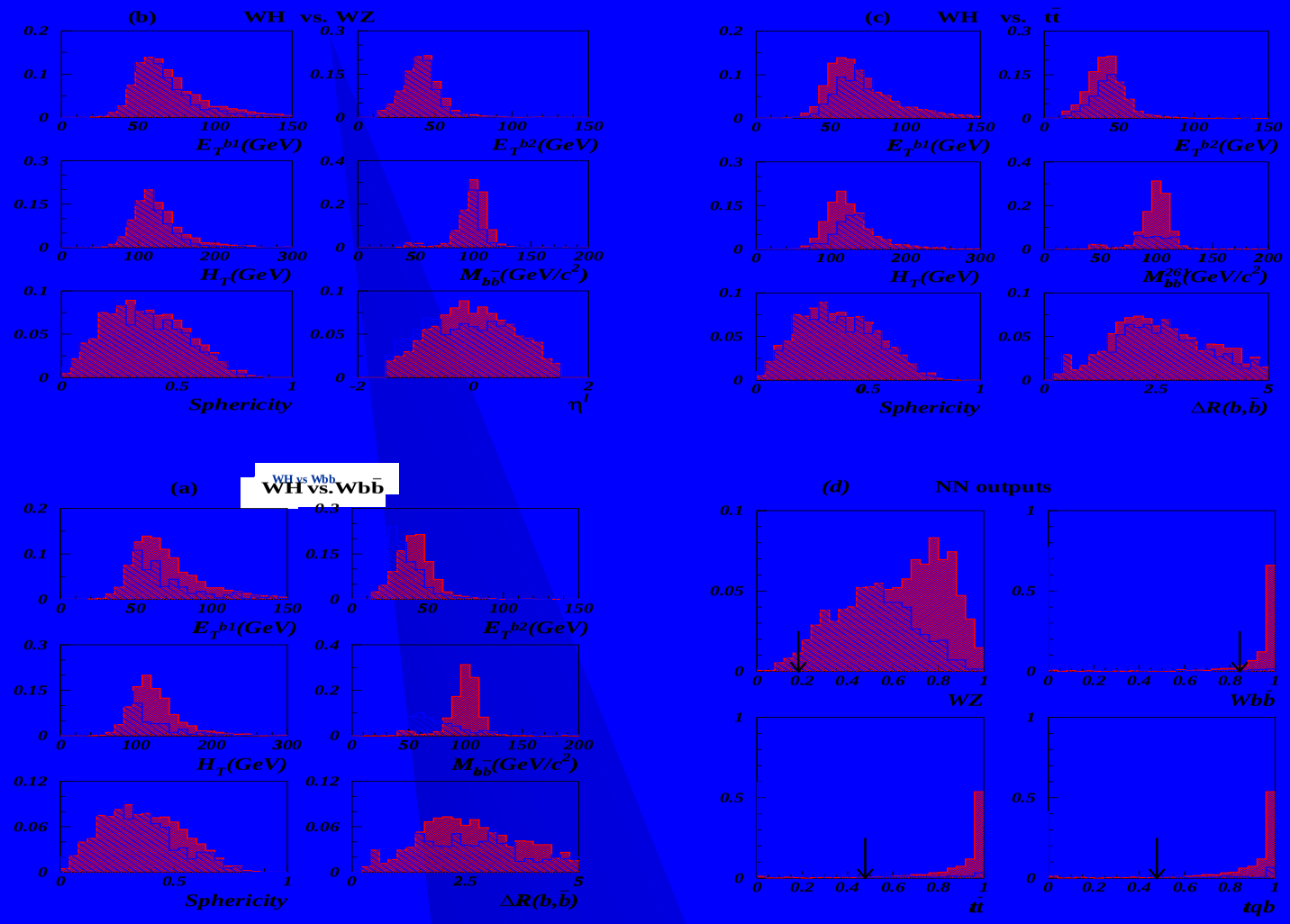
$m_t = 173.3 \pm 5.6(\text{stat.}) \pm 6.2\ (\text{syst.})\ \text{GeV}/c^2$

Strategy for Discovering the Higgs Boson at the Tevatron

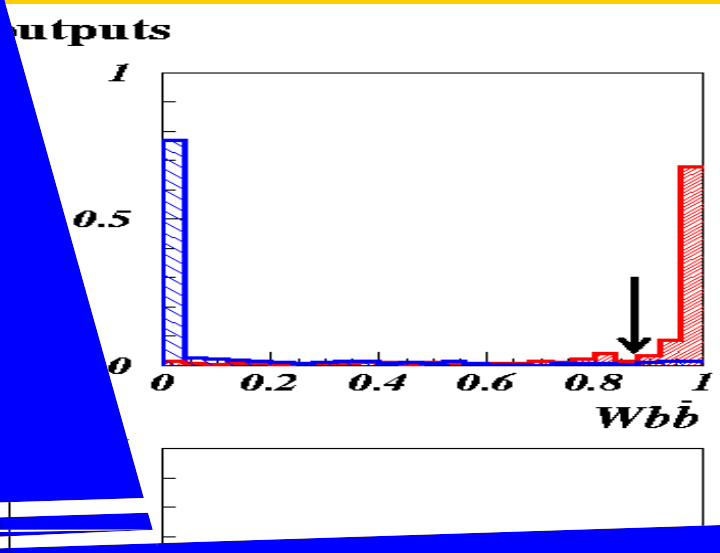
P.C. Bhat, R. Gilmartin, H. Prosper, **PRD 62 (2000)**
hep-ph/0001152

WH Results from NN Analysis

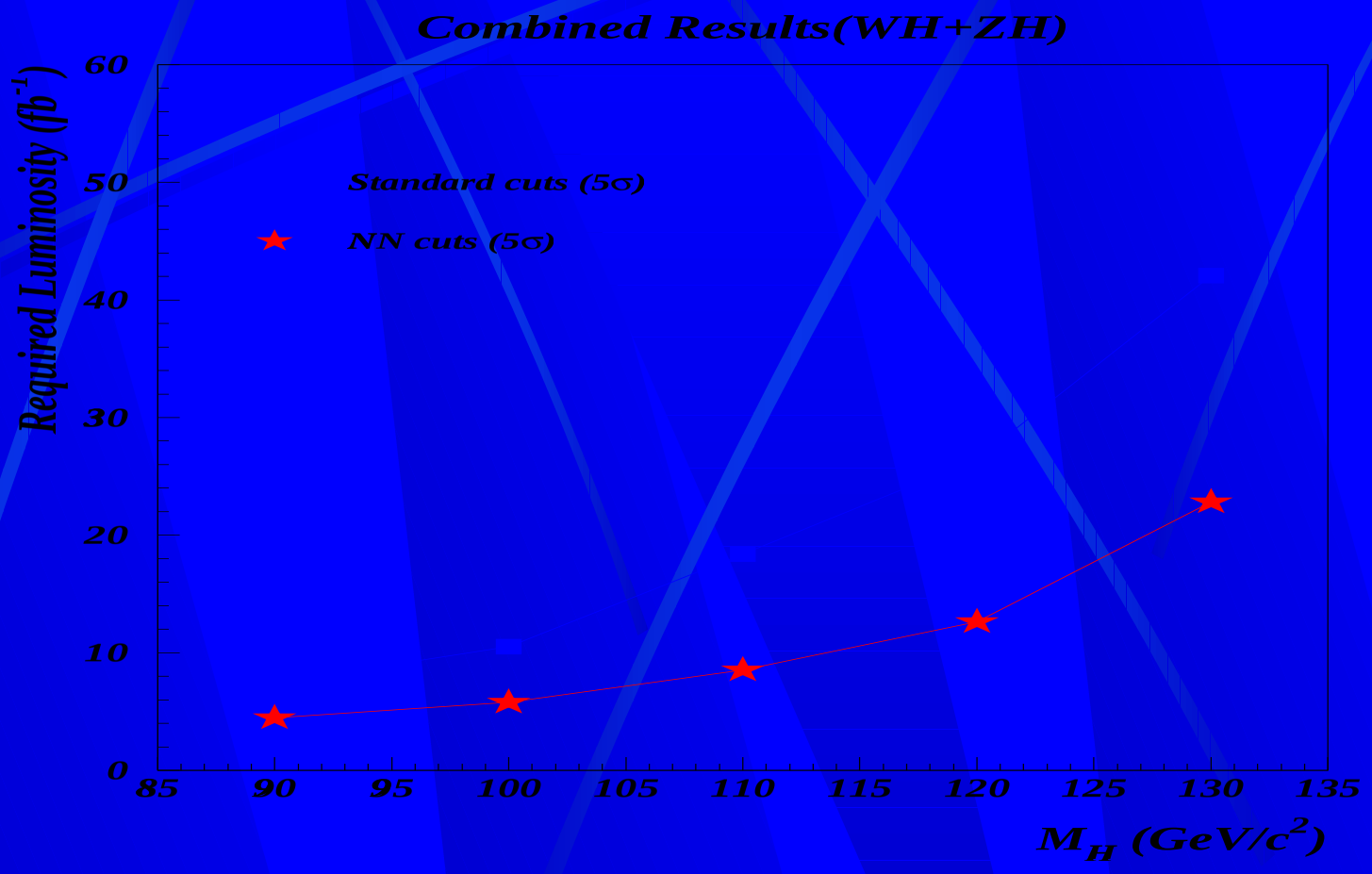
$M_H = 100\ GeV/c^2$



W



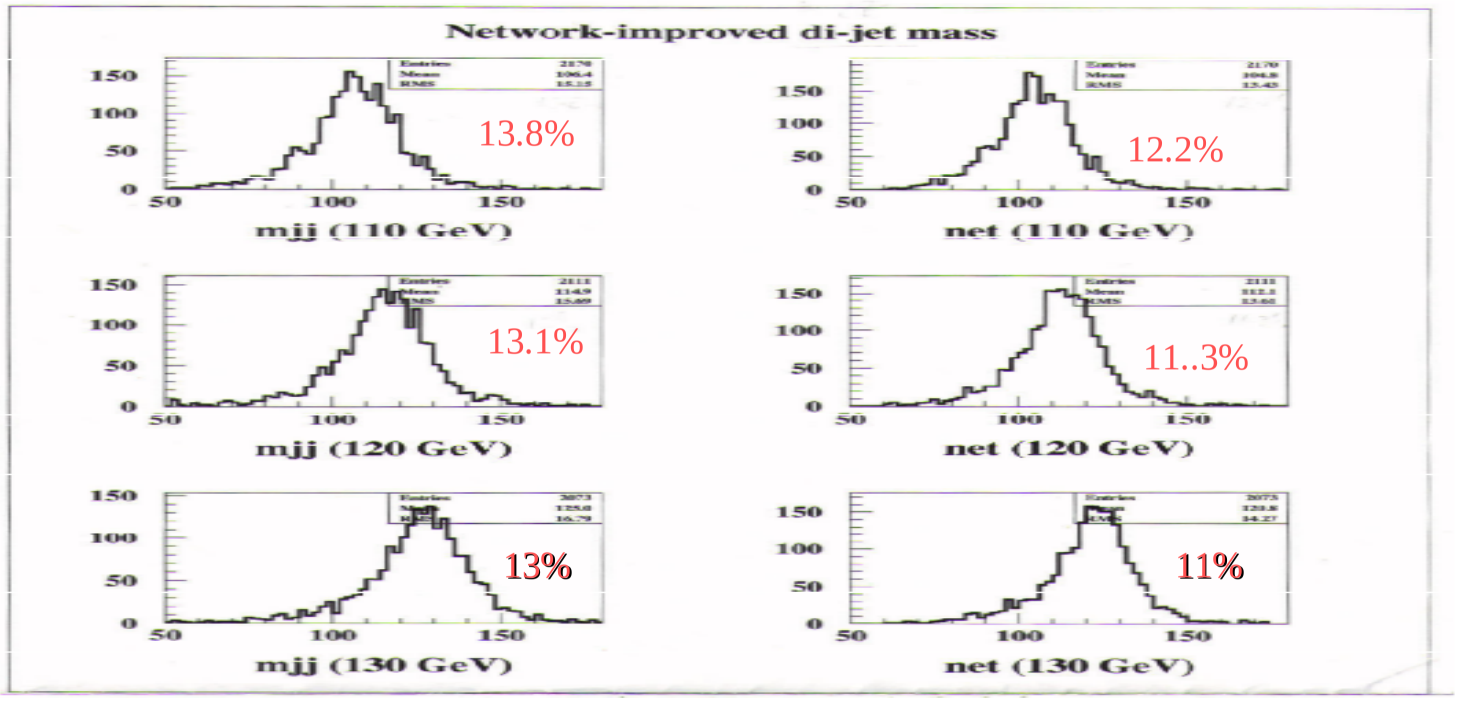
Results, Standard vs. NN



A good chance of discovery up to M_H= 130 GeV/c^2 with 20-30fb^{-1}

Improving the Higgs Mass Resolution

Use m_{jj} and H_T ($=\sum E_t^{jets}$) to train NNs to predict the Higgs boson mass



N

etworks

be better than the best single

d variance

B

ach provides a well-founded
ure to make straight-forward and
nparisons. It also allows treatment of
nsistent manner.

lications:

lti-source models

o survival probability

adaptive algorithms such as
Neural Networks (NN)

S

s have already made impact
ion measurements and will be
n future analyses.

the surface in our use of
thms.

ng "intelligent" algorithms