Advanced event analysis methods at the energy frontier

Reinhard Schwienhorst

MICHIGAN STATE UNIVERSITY

CPPM Seminar, July 5 2007

Outline

- Introduction
  - Physics at the energy frontier
- Analysis procedure
- Event analysis methods
  - Decision trees
    - Boosting
    - Random forest
  - Bayesian neural networks
  - Classifier comparisons
- Conclusions

Disclaimer: highlight general principles and guiding ideas

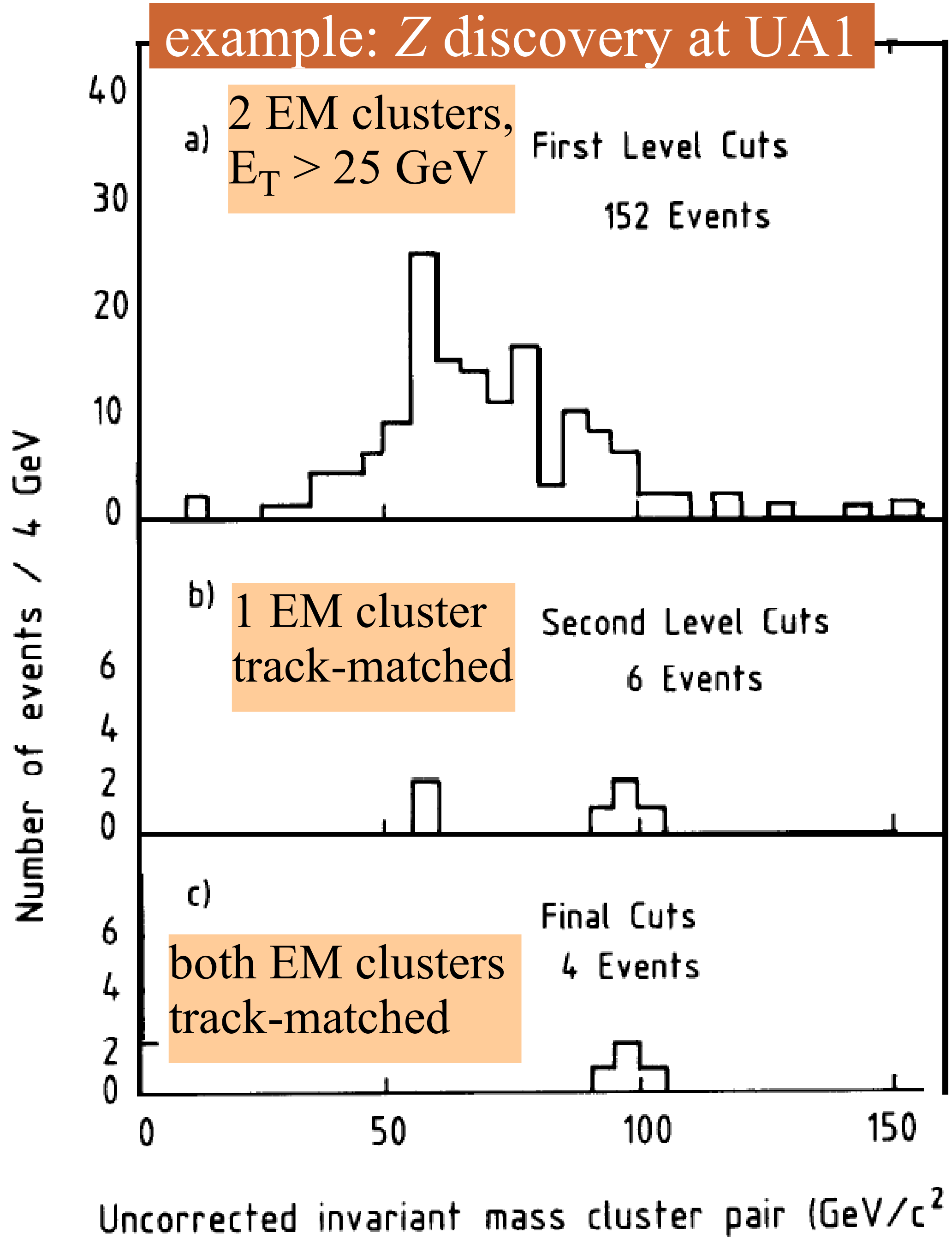
- Not necessarily mathematically rigorous
- Some of the same topics were addressed at PhyStat conferences

Typical event analysis procedures

1) Cut-based event counting
2) Peak in a characteristic distribution

Event counting

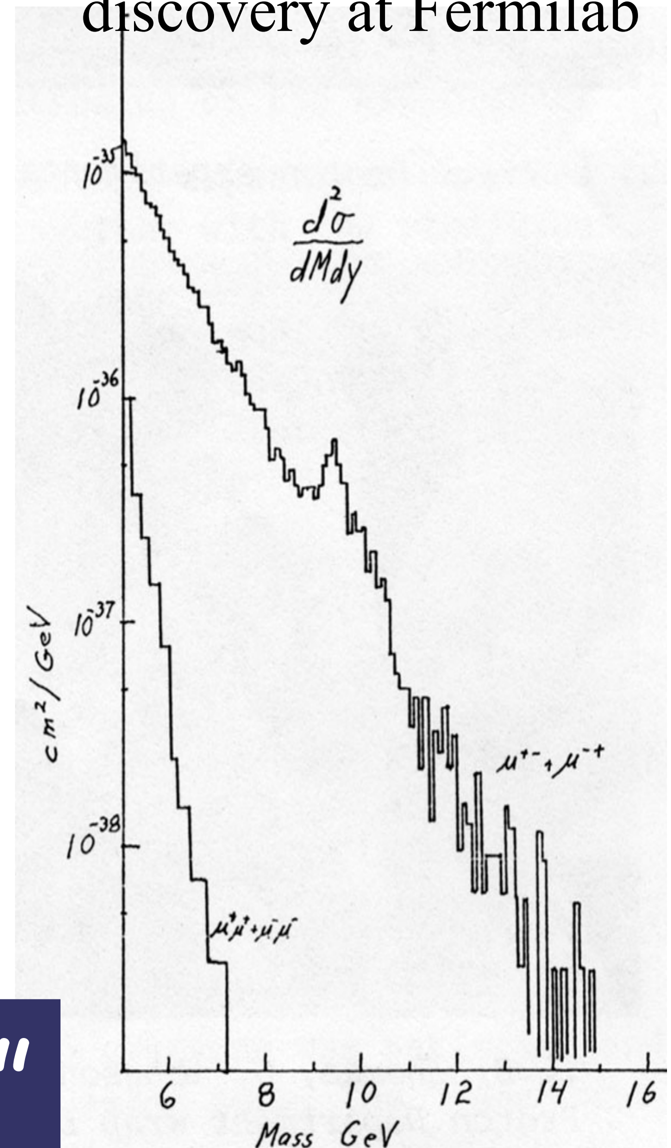
- Apply cuts to variables describing the event
  - Object identification
  - Kinematic cuts on objects
  - Event kinematics
- Goal: cut until the signal is visible
  - No background left
  - Or large $S/\sqrt{B}$
- Sensitive to any signal with this final state
- Requires understanding of background



Peak in a characteristic distribution

- Find a variable that has a smooth distribution for background
  - Typically invariant mass
- Measure this distribution over a large range of possible values
- Look for possible resonance peaks
- Sensitive to any resonance with this final state
- Background estimate for sidebands

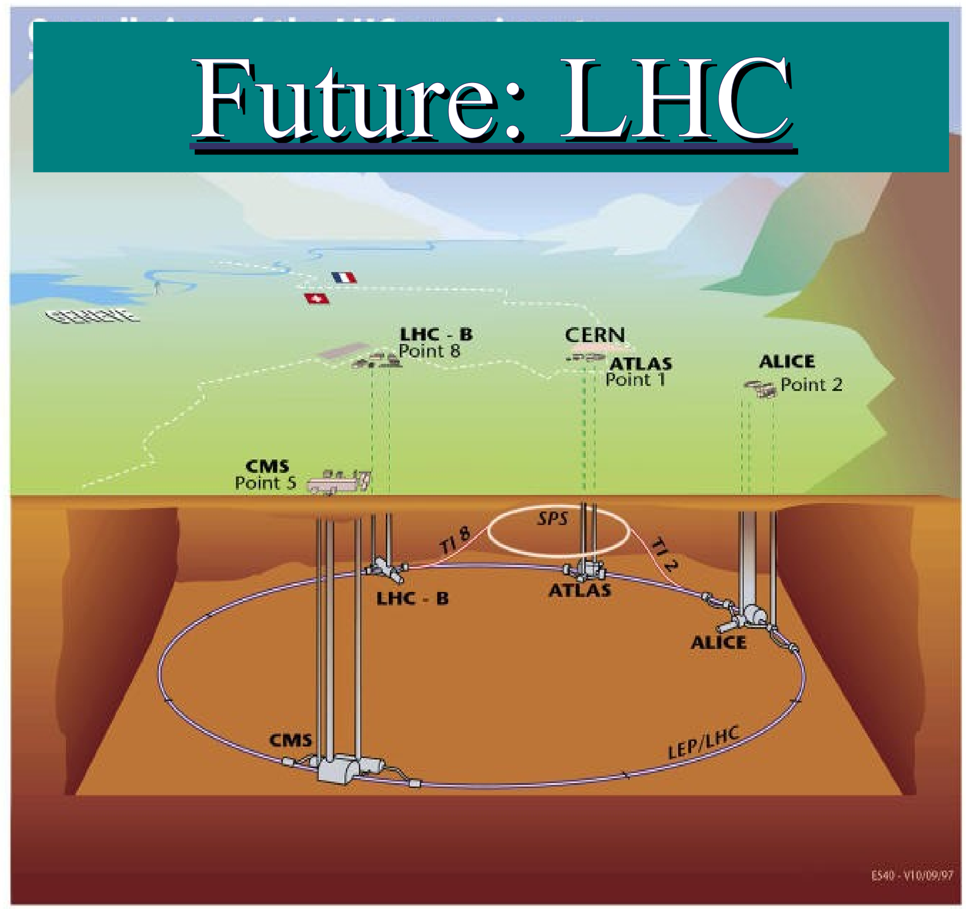
Example: b-quark discovery at Fermilab



"Bump Hunting"

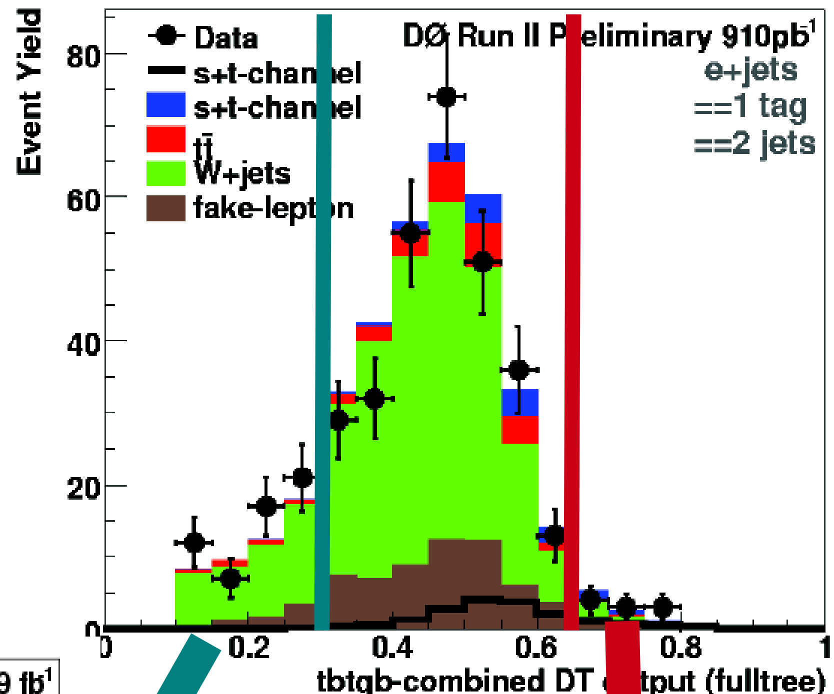
The energy frontier

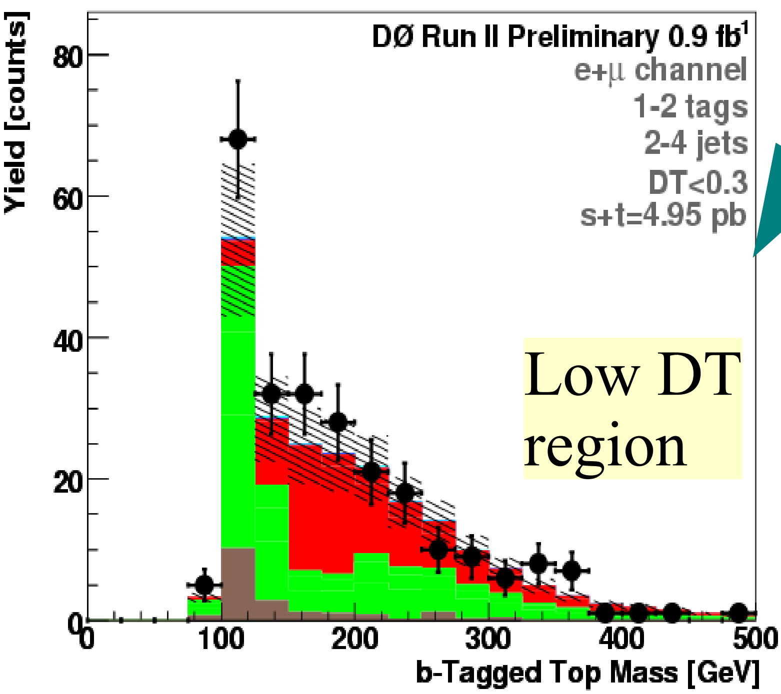
- Colliding particles at the highest available energies
    - Probe structure of matter at the most fundamental level
  - Observe interactions at the smallest possible distances
  - Produce never-before-seen particles

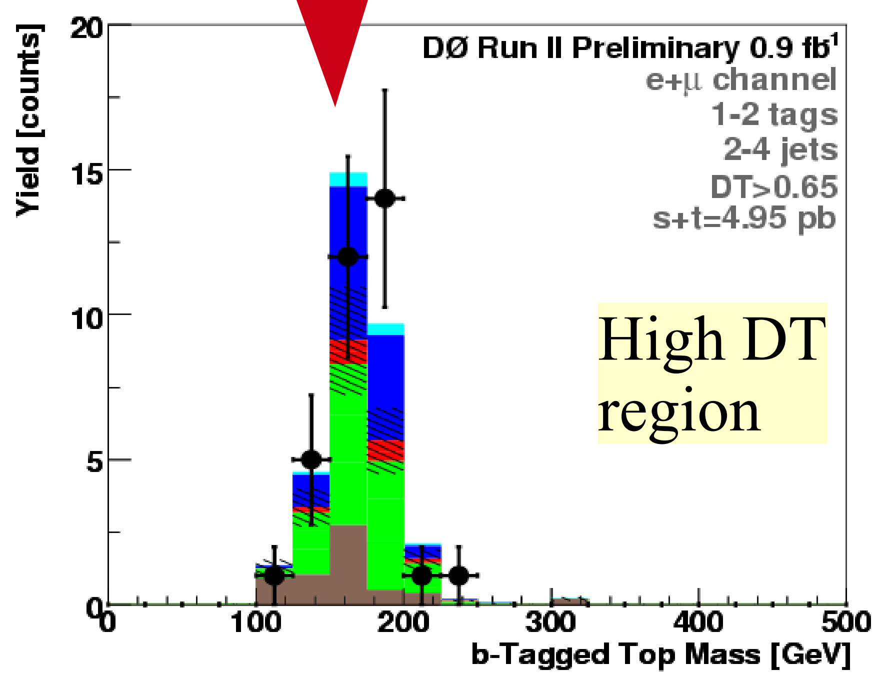


Searches at the energy frontier

- Tevatron:
  - Single top quark production

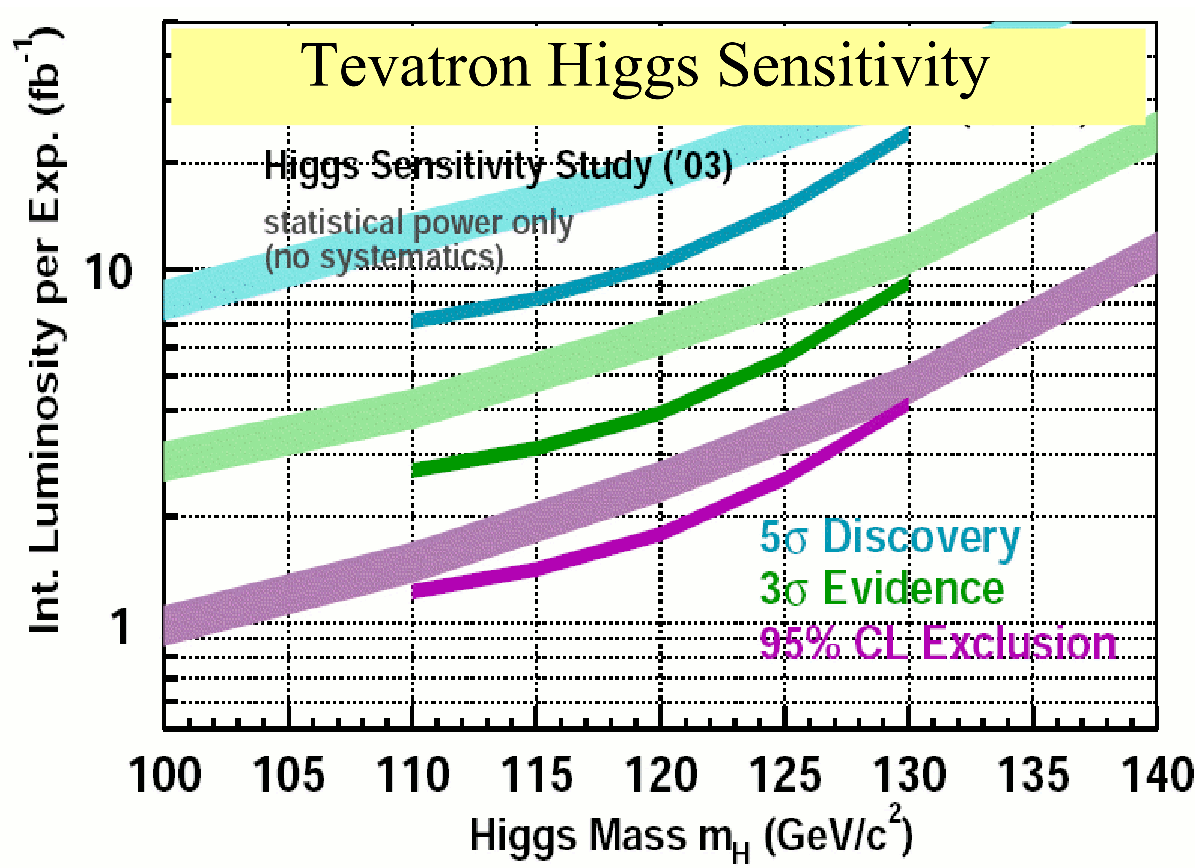






Searches at the energy frontier

- Searches for new particles, phenomena, couplings
  - Tevatron:
    - Single top quark production
    - Higgs boson search
    - SUSY
    - Extra dim
    - ...
    - Statistics-limited

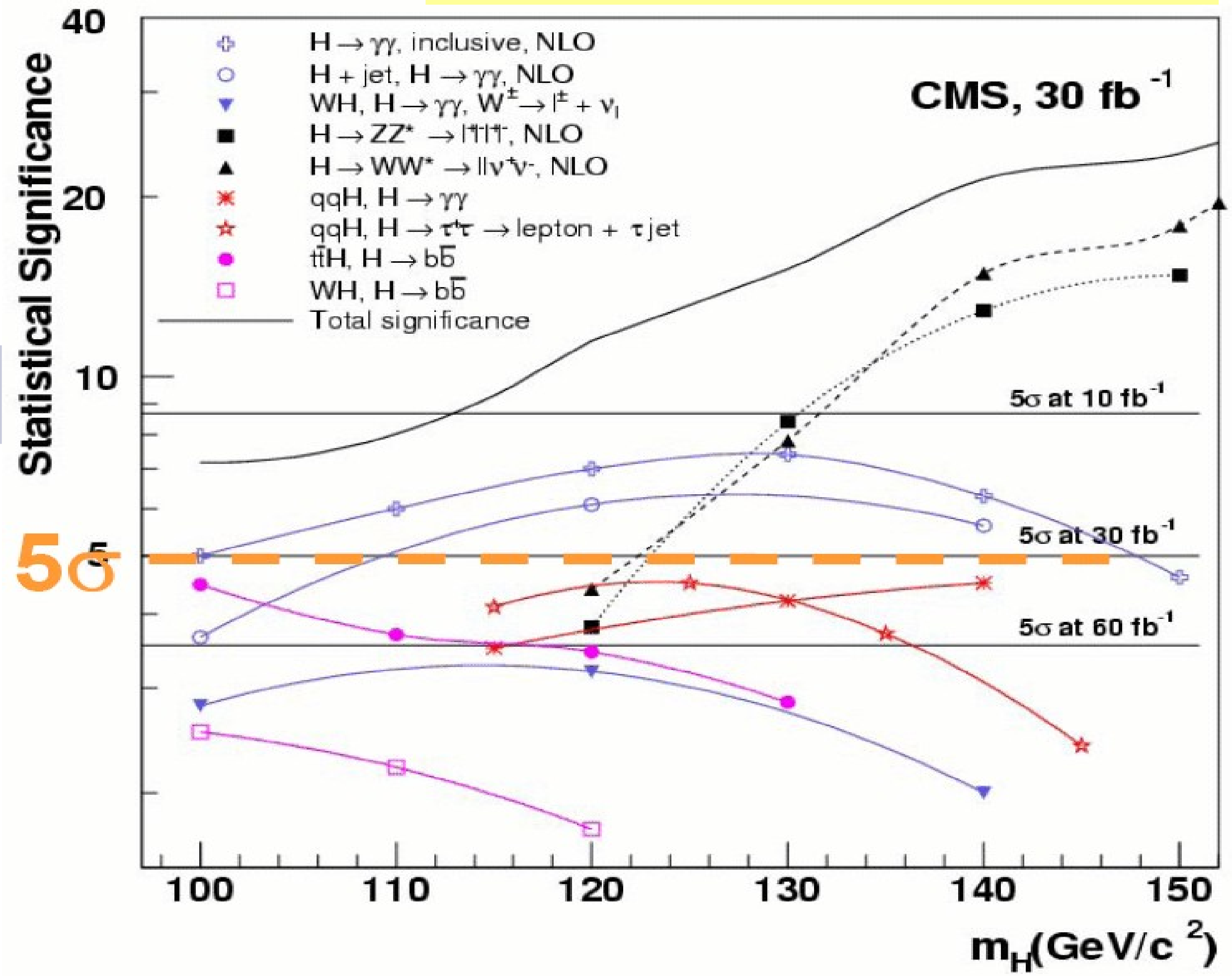


Multivariate techniques make this level of sensitivity possible

Searches at the energy frontier

- Searches for new particles, phenomena, couplings
  - Tevatron:
    - Single top quark production
    - Higgs boson search
    - SUSY
    - Extra dim
    - …
  - LHC:
    - Higgs searches

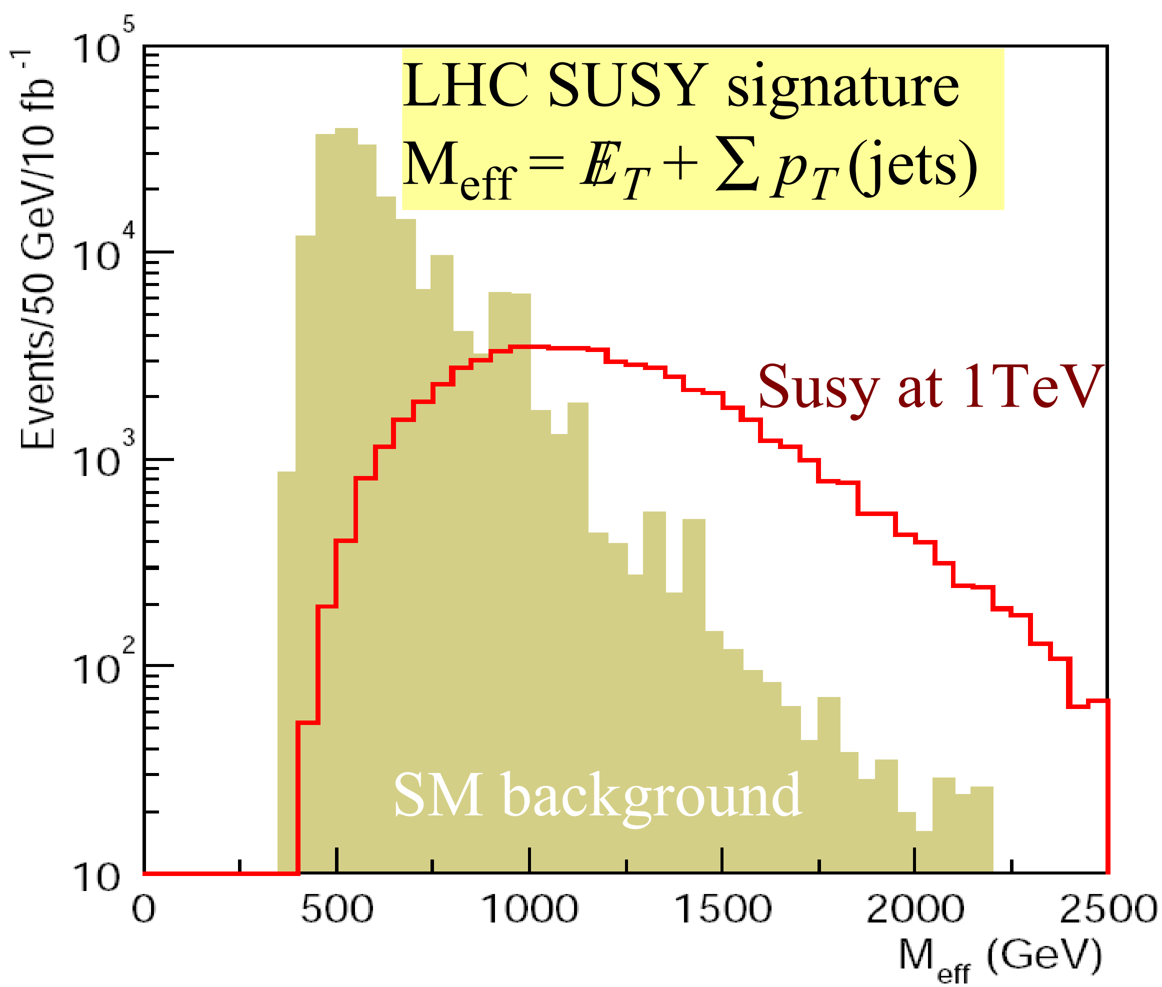
LHC Higgs Sensitivity



Multivariate techniques will be required to reach this level of sensitivity

Searches at the energy frontier

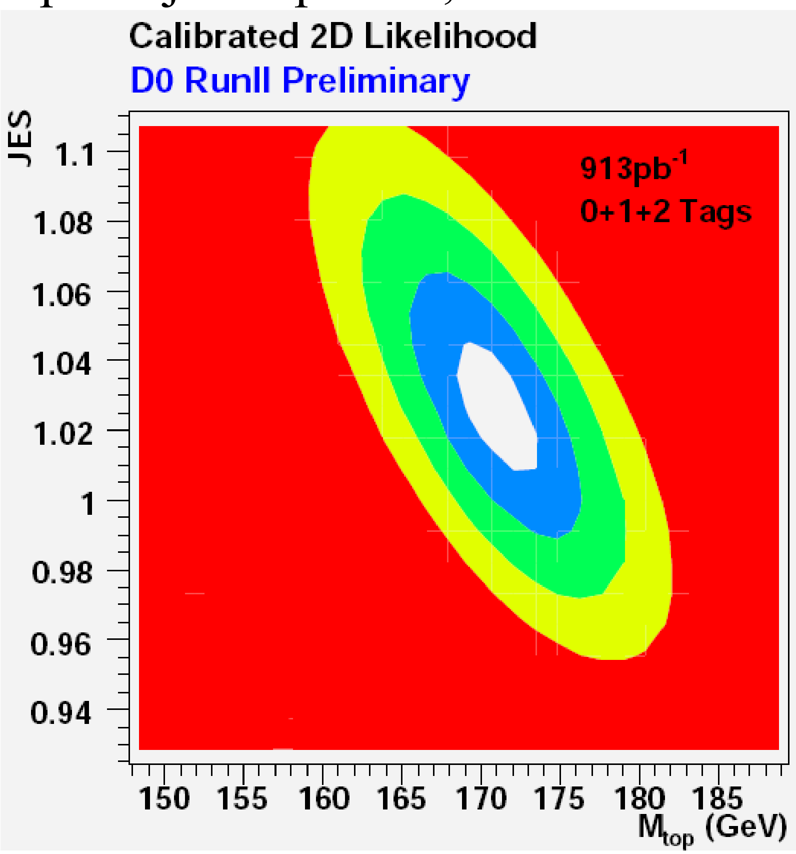
- Searches for new particles, phenomena, couplings
  - Tevatron:
    - Single top quark production
    - Higgs boson search
    - SUSY
    - Extra dim
    - ...
  - LHC:
    - Higgs searches
    - SUSY
    - Extra dim
    - ...
    -

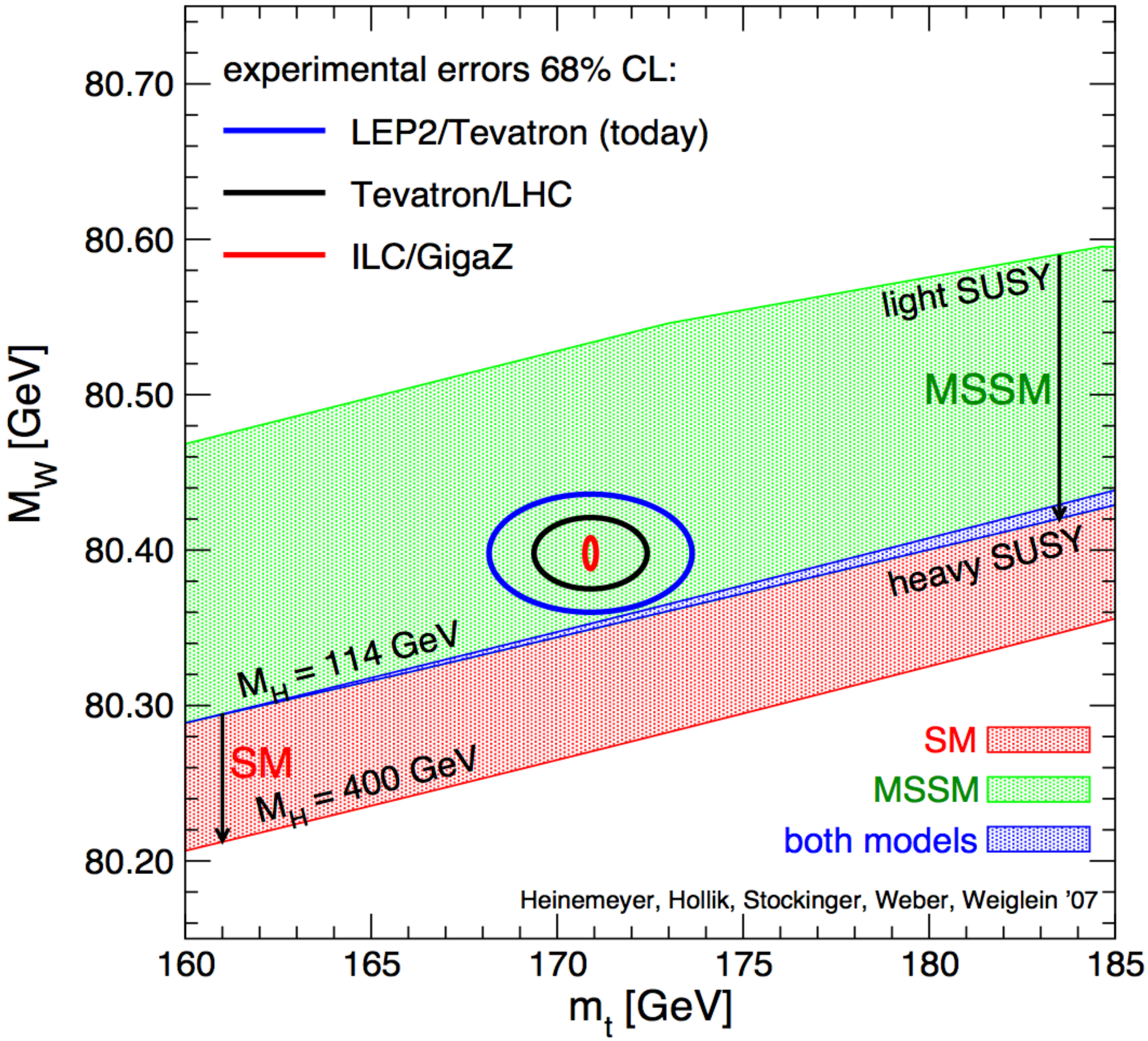


Measurements at the energy frontier

- First measurements of properties, couplings
  - With samples of limited size
  - Example: Top quark mass
    - $\sigma \sim 3$ GeV with 1 fb^{-1}
    - Was the goal for 2 fb^{-1}

lepton+jets top mass, matrix element

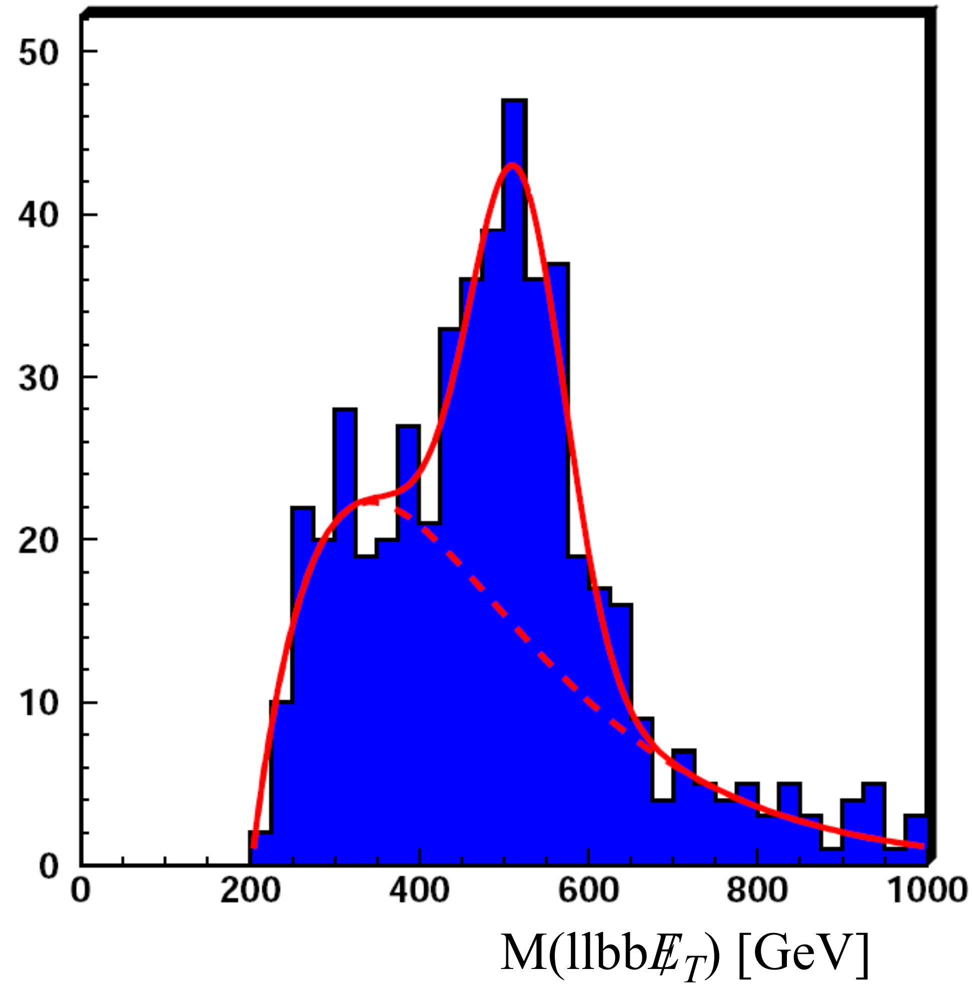




Measurements at the energy frontier

- First measurements of properties, couplings
  - With samples of limited size
  - Example:
    LHC SUSY particle
    masses

LHC $\tilde{b}$ mass:
100 signal
events in 30 fb^{-1}



Physics at the energy frontier

- Searches for new particles, phenomena, couplings
- First measurements of properties, couplings
- Multivariate techniques $\leftrightarrow$ Adding more data

Making the most out of small samples of events

How to improve upon

Event counting

and

Bump hunting

Bayesian limit

- For each analysis, there exists a fully optimized signal-background separation
  - Target function, also called Bayes discriminant or Bayesian limit

$$B(x) = \frac{L(S|x)}{L(B|x)}$$

- For a single discriminating variable, this ratio of signal and background likelihoods is easy to calculate
  - Monte Carlo procedure:
    - Generate signal and background MC events
    - Fill histograms for signal and background
    - Divide the two histograms

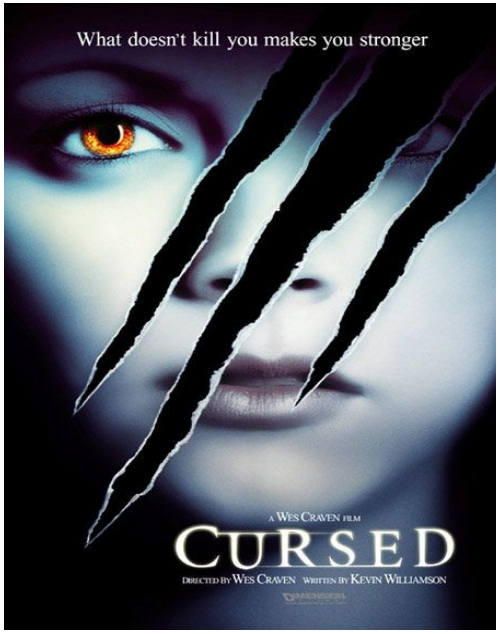
Bayesian limit

- For each analysis, there exists a fully optimized signal-background separation
  - Target function, also called Bayes discriminant or Bayesian limit

$$B(x) = \frac{L(S|x)}{L(B|x)}$$

- For a single discriminating variable, this ratio of signal and background likelihoods is easy to calculate
- In case of more than one variables, this isn't possible anymore
  - Not enough MC statistic to compute a many-dimensional likelihood

Curse of dimensionality



Optimized event analysis

Optimized = $\left\{ \begin{array}{l} \text{Optimize signal-background separation} \\ \text{Exploit full event information} \\ \quad \text{Event kinematics, angular correlations, ...} \\ \text{Take all correlations into account} \end{array} \right.$

Goal: Reach the Bayesian limit

- Requires detailed understanding of signal and background
  - Only applicable to searches for a specific signal or measurements of a specific process

Optimized event analysis

Optimized = { Optimize signal-background separation
Exploit full event information
Event kinematics, angular correlations, ...
Take all correlations into account }

Goal: Reach the Bayesian limit

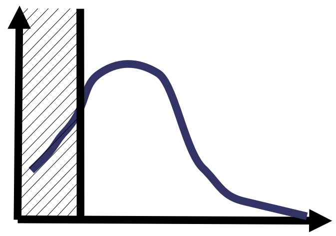
- Requires detailed understanding of signal and background
  - Only applicable to searches for a specific signal or measurements of a specific process
- Limited by background and signal modeling
  - MC statistics, MC model, background composition, shape,

<u>If signal model is wrong:</u> search is not sensitive 😐

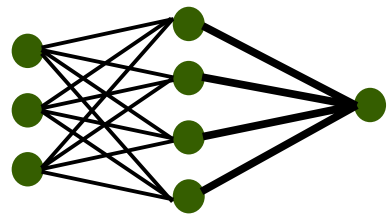
<u>If background model is wrong:</u> find something that isn't there ☹

Event analysis techniques

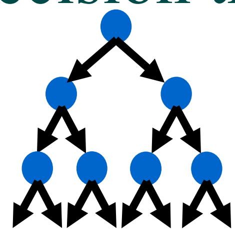
Cut-Based



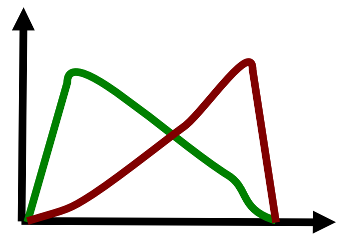
Neural networks



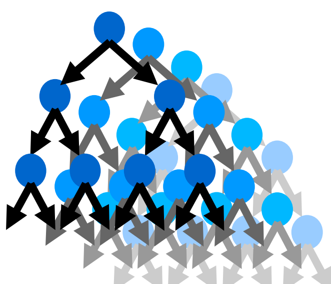
Decision trees



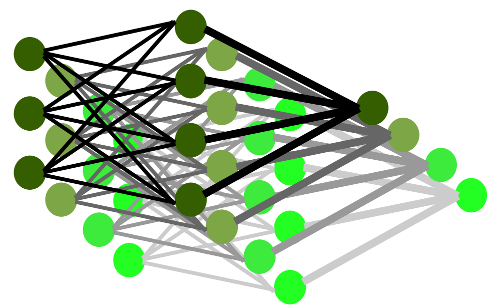
Likelihood



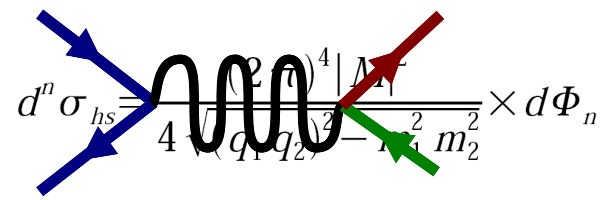
Boosted decision trees, random forest



Bayesian neural networks



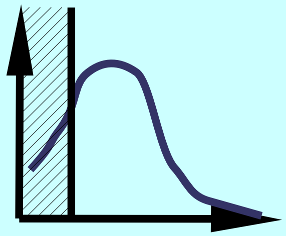
Matrix Elements



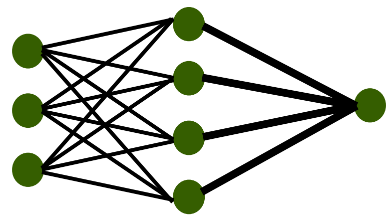
Many others: Kernel methods, support vector machines, ...

Event analysis techniques

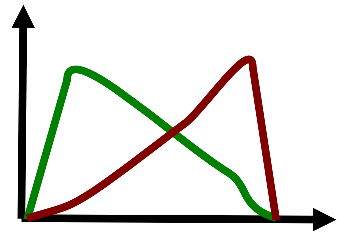
Cut-based



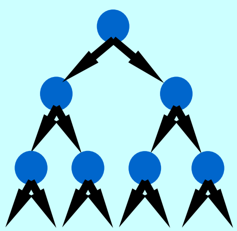
Neural networks



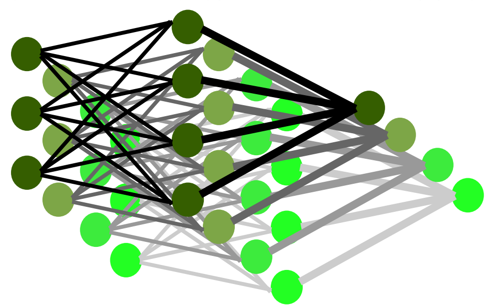
Likelihood



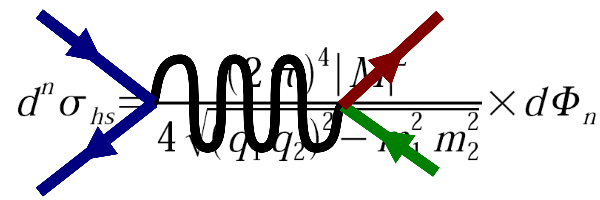
decision trees



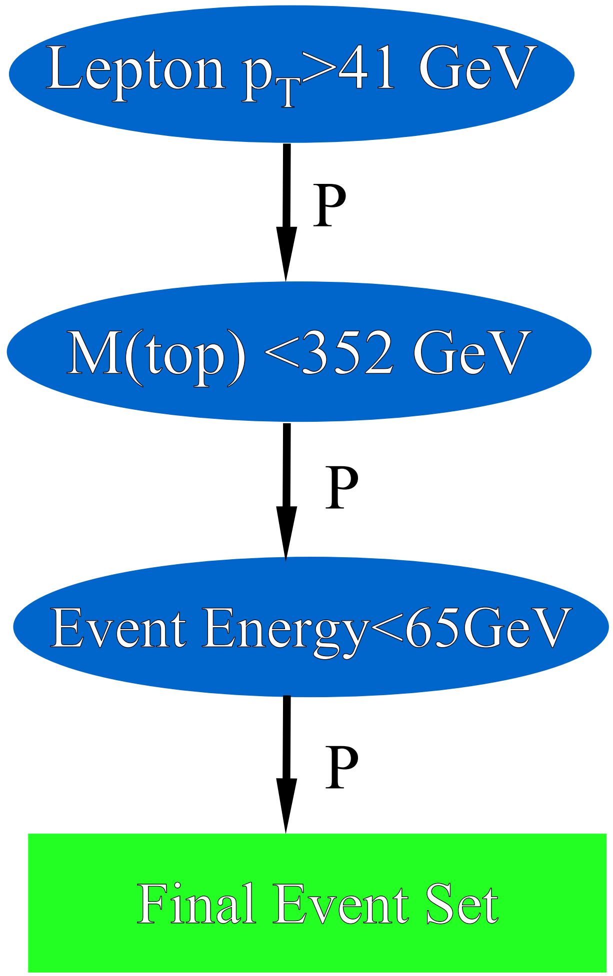
Bayesian neural networks



Matrix Elements



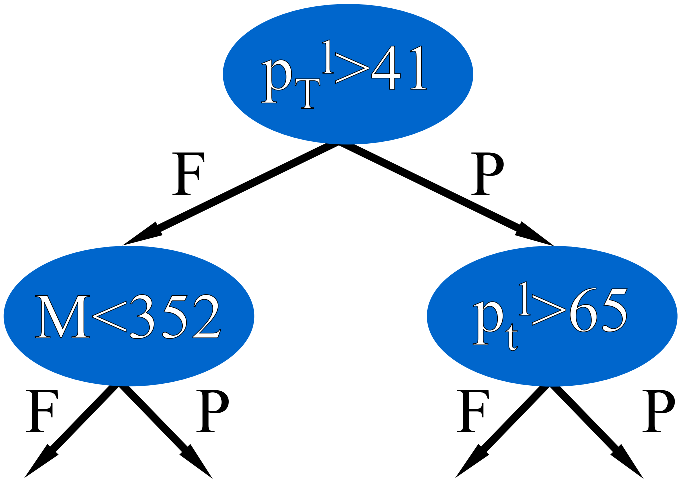
Cut-based analysis



In the final event set

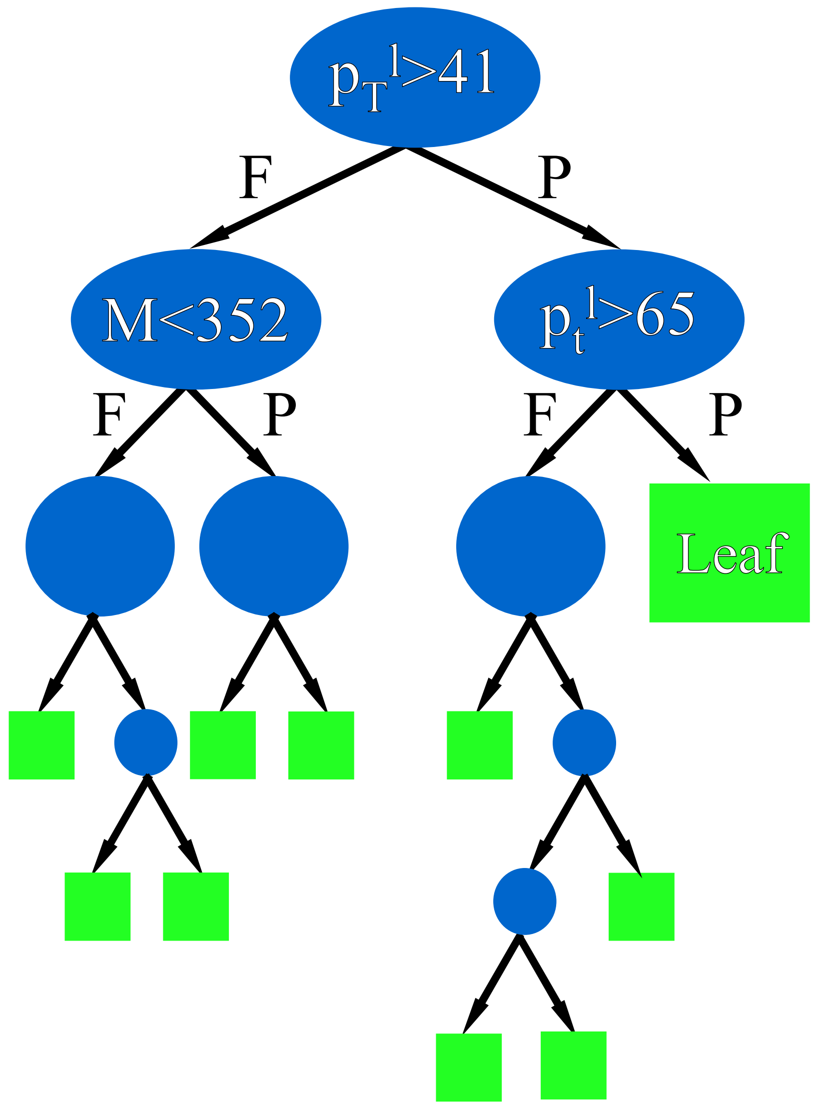
- Estimate background yield
- Compare to data
  $N_{obs} = N_{data} - N_B$
- Calculate signal acceptance
  $\sigma = N_{obs} / (A*L)$

Including events that fail a cut



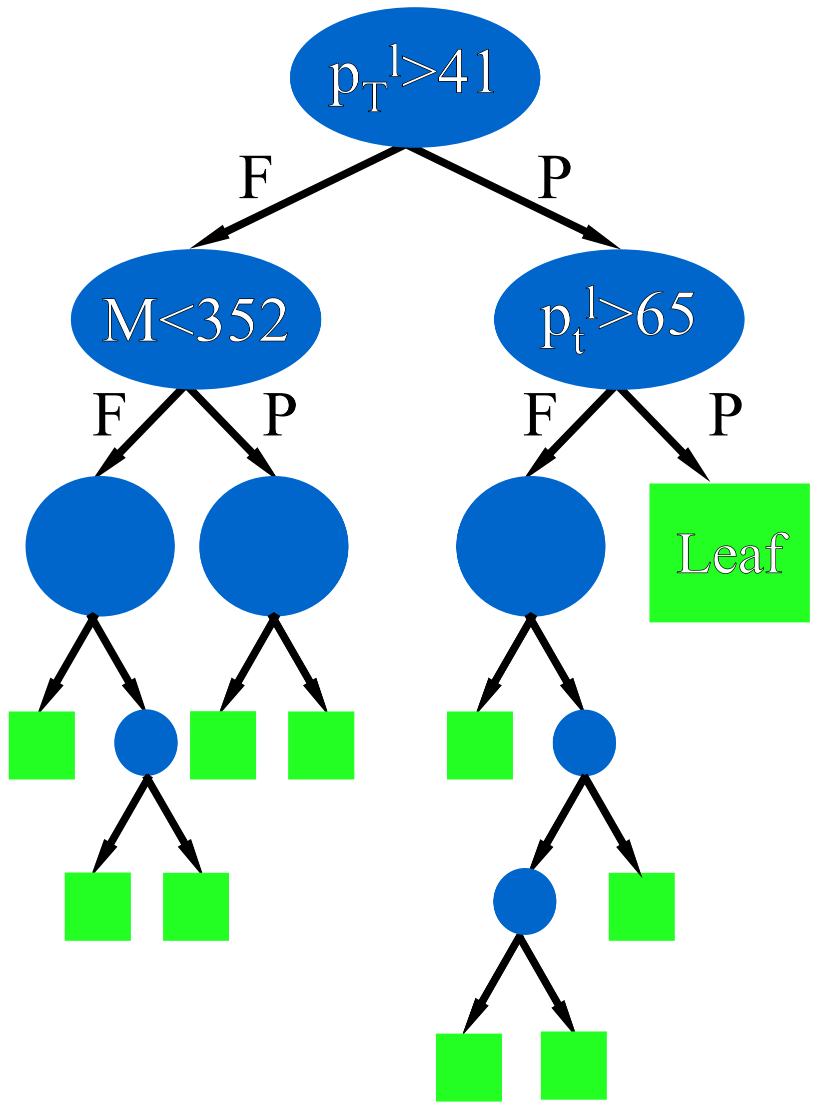
- Create a tree of cuts
- Divide sample into "pass" and "fail" sets
- Each node corresponds to a cut (branch)

Trees and leafs



- Create a tree of cuts
- Divide sample into "pass" and "fail" sets
- Each node corresponds to a cut (branch)
- A leaf corresponds to an end-point
- For each leaf, calculate purity (from MC):
  purity = $N_S/(N_S+N_B)$

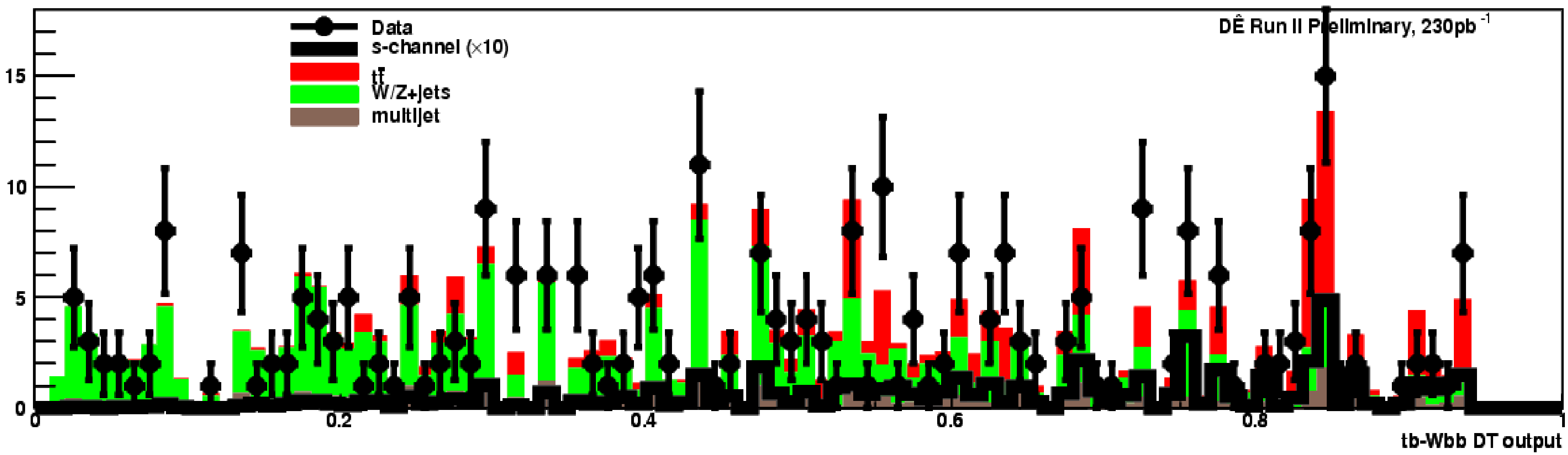
Decision tree



- Create a tree of cuts
- Divide sample into "pass" and "fail" sets
- Each node corresponds to a cut (branch)
- A leaf corresponds to an end-point
- For each leaf, calculate purity (from MC): purity = $N_S/(N_S+N_B)$
- Train the tree by optimizing the Gini improvement:
  - Gini = $2\, N_S N_B /(N_S + N_B)$
  - Each leaf will be either background- or signal-enhanced

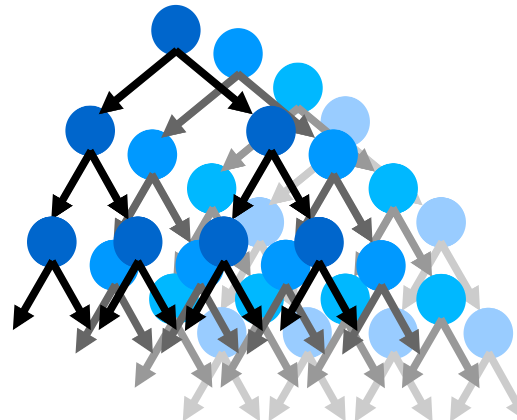
Decision tree output

- Train on signal and background models (MC)
  - Stop and create leaf when $N_{MC}<100$
- Compute purity value for each leaf
- Send data events through tree
  - Assign purity value corresponding to the leaf to the event
- Result approximates a probability density distribution



Boosting

Boosting

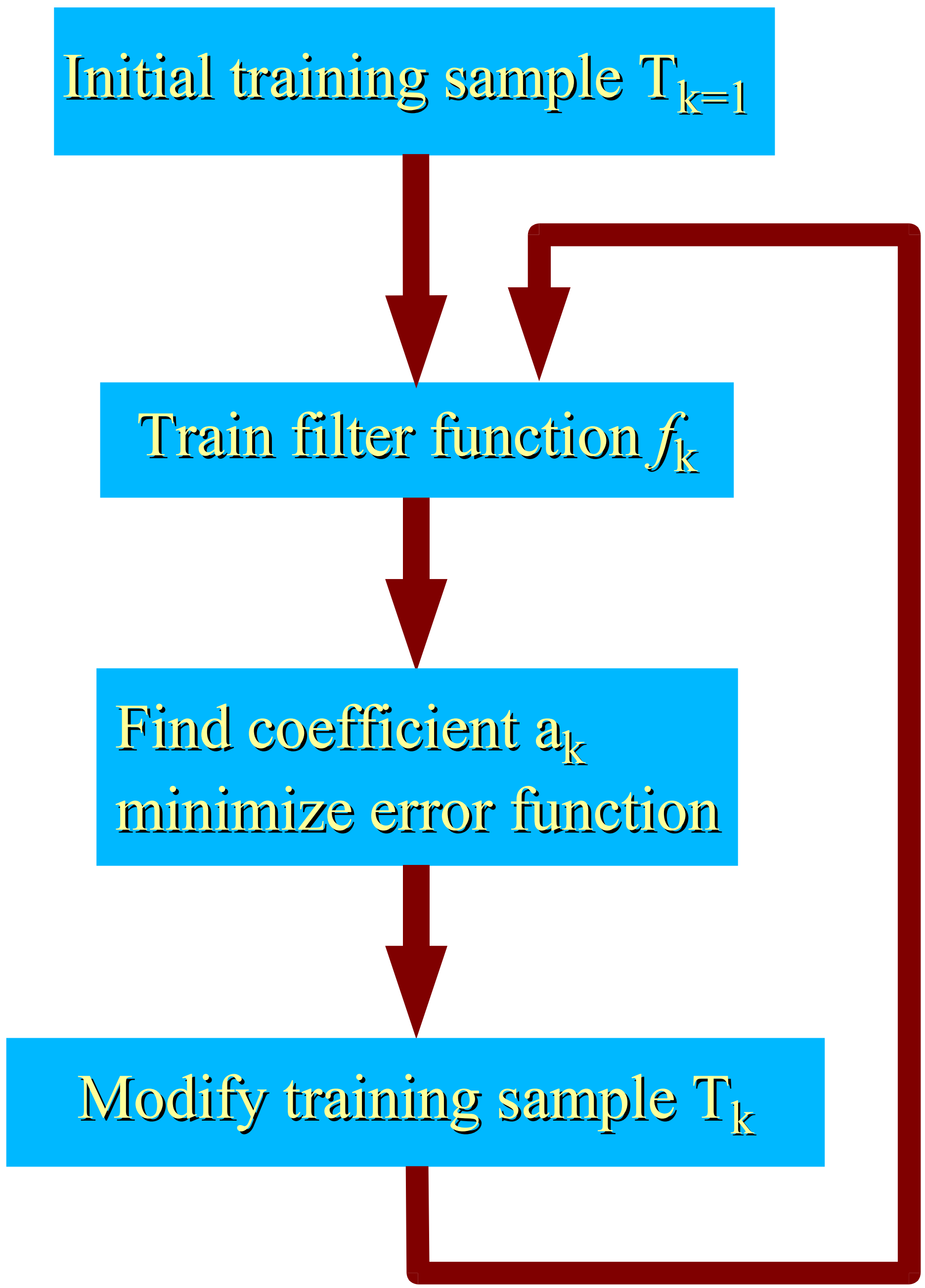


- A general method to improve the performance of any weak qualifier
  - Decision trees, neural networks, ...
- Linear combination of many filter functions

$$F(x) = \sum_k a_k \, f_k(x)$$

  - a_k: coefficient, typically result of minimization of error function

Boosting procedure

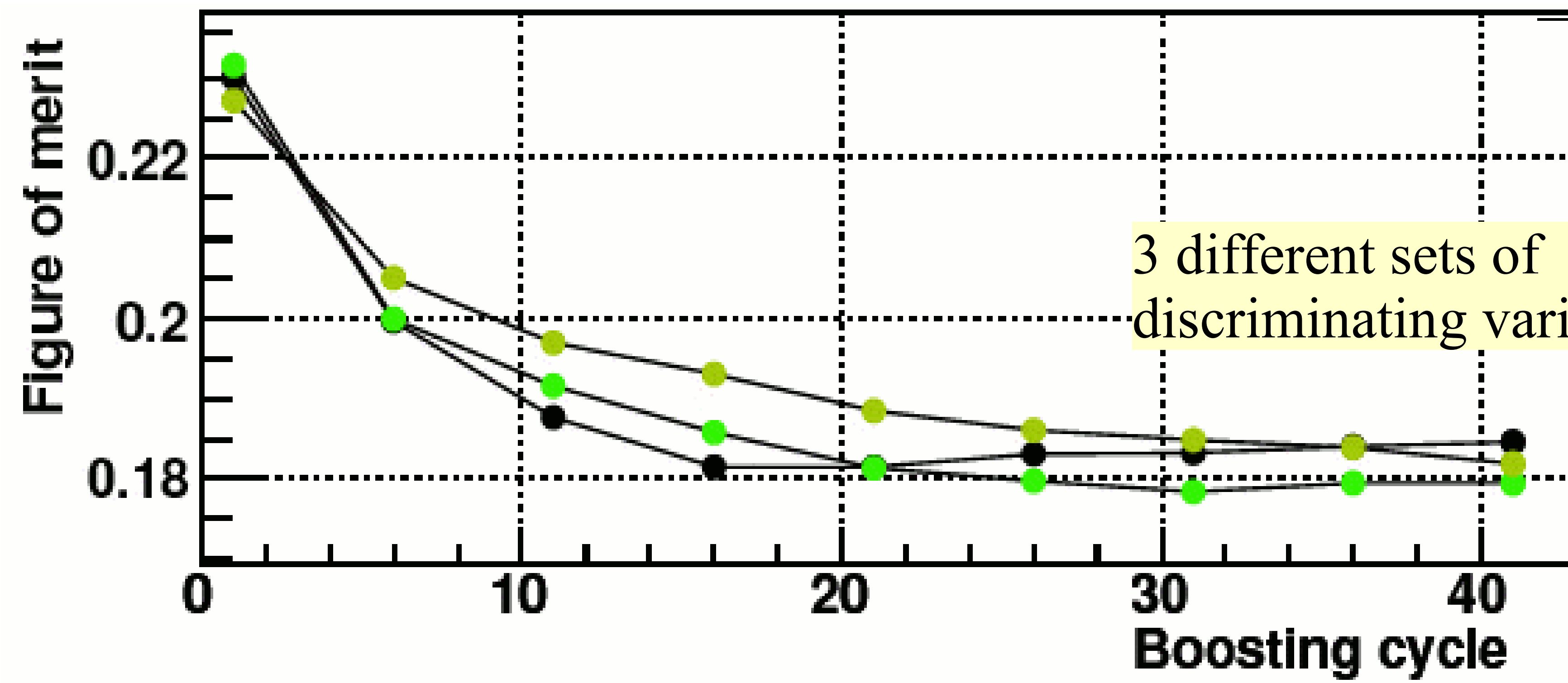


$$F = \sum_k a_k f_k$$

Adaptive boosting

- In each iteration, update coefficient a_k
  - From minimizing error function
  - coefficients decrease at each iteration
- Update weight for each event in training sample T_k
  - Figure out which events have been misclassified
    - Signal events should have purity ≥ 0.5
    - Background should have purity <0.5
  - Increase event weight for those events that have been misclassified

Boosting performance

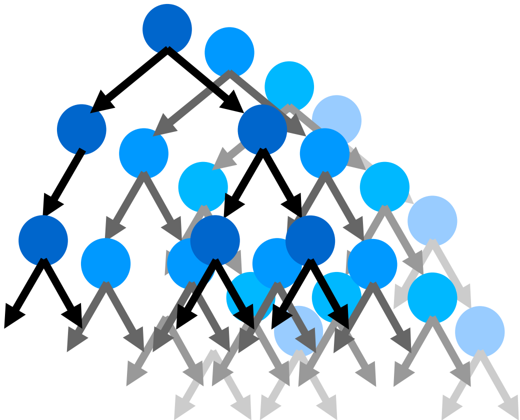


3 different sets of discriminating variables

DØ single top search
with decision trees

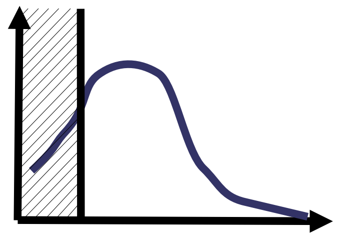
Random forest

- Average over many decision trees
  - Typically O(100)
- Each tree is grown using m variables
  - For N total variables, m<<N
- Very fast algorithm
  - Even with large number of variables
- Very few parameters to adjust
  - Typically only m

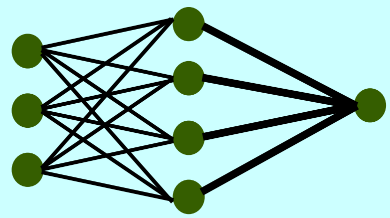


Event analysis techniques

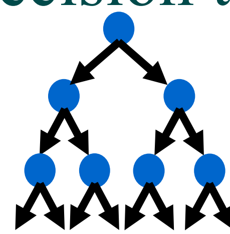
Cut-Based



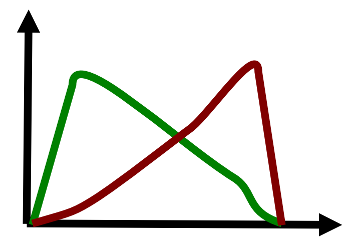
Neural networks



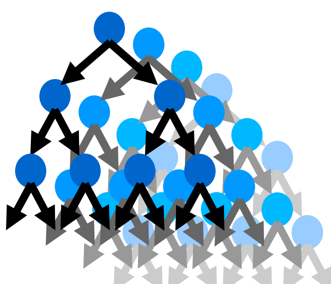
Decision trees



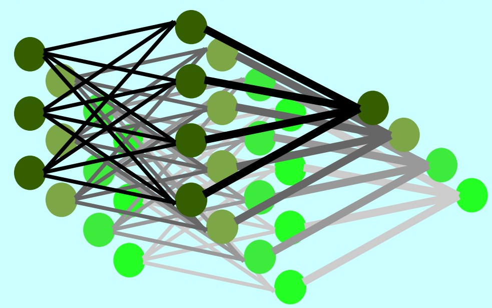
Likelihood



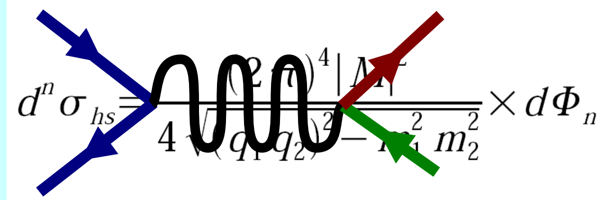
Boosted decision trees,
random forest



Bayesian neural networks

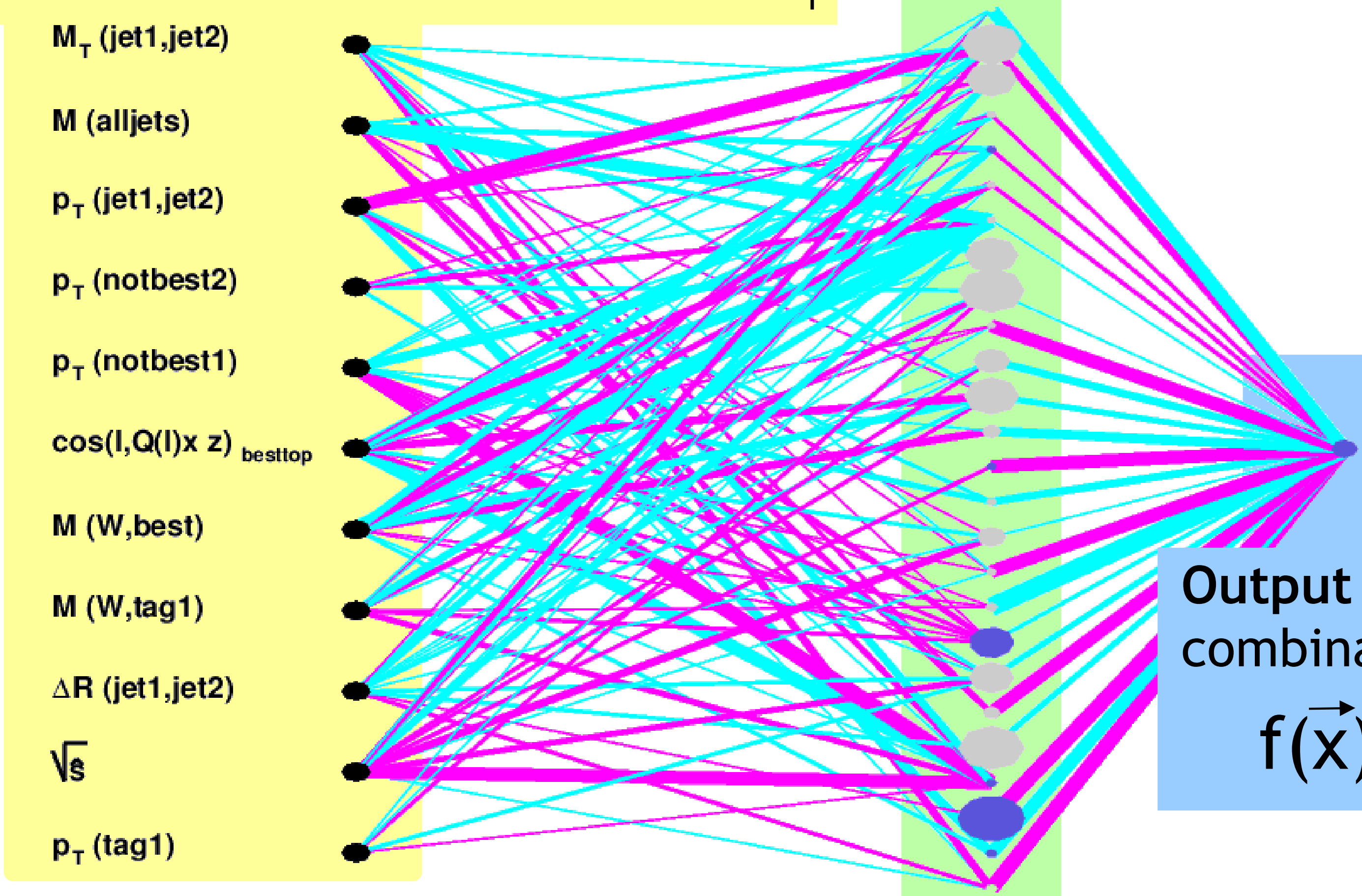


Matrix Elements



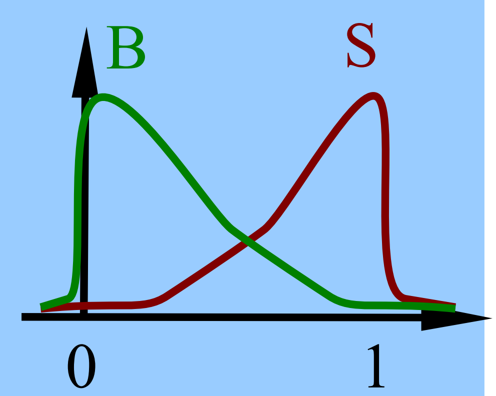
Neural networks

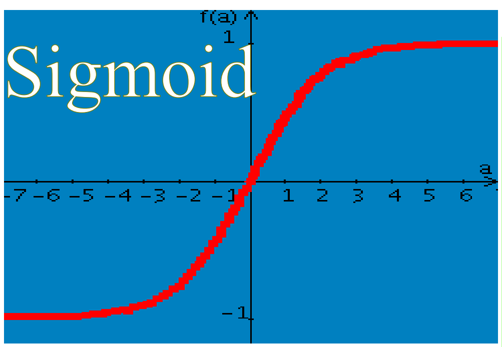
Input Nodes: One for each variable x_i



Output Node: linear combination of hidden nodes

$$f(\vec{x}) = \Sigma\, w'_k\, n_k(\vec{x},\vec{w}_k)$$



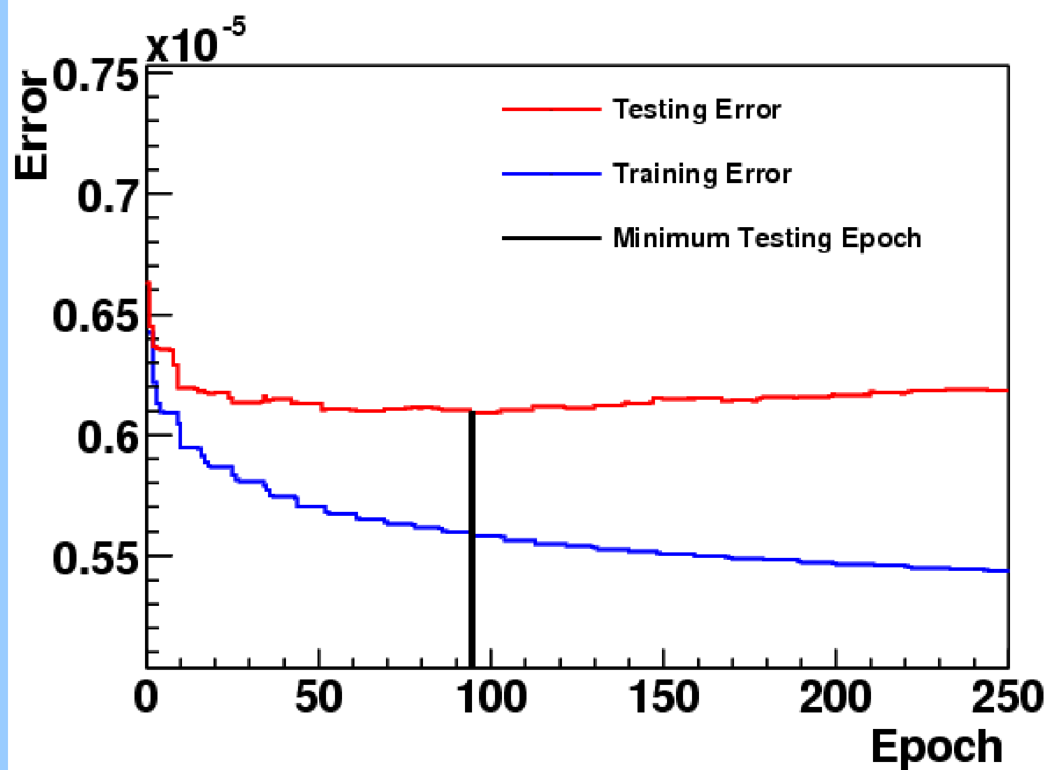


Hidden Nodes: Each is a sigmoid dependent on the input variables

$$n_k(\vec{x},\vec{w}_k) = \frac{1}{1 + e^{-\Sigma w_{ik} x_i}}$$

Neural Network Training

- Find optimum NN parameters on training signal/background events
- Apply NN to independent set of signal and background
  - Testing sample
- Stop training when error from testing sample starts increasing
  - Overfitting



DØ single top search

Bayesian neural networks

- Bayesian idea:
  - Rather than finding one value for each weight, determine the posterior probability for each weight
- Form many networks by sampling from the posterior
- Typical case: ~100 individual neural networks
  - Each network gets a weight based on training performance
- Avoids overfitting
- But: very slow due to integration required to determine the posterior

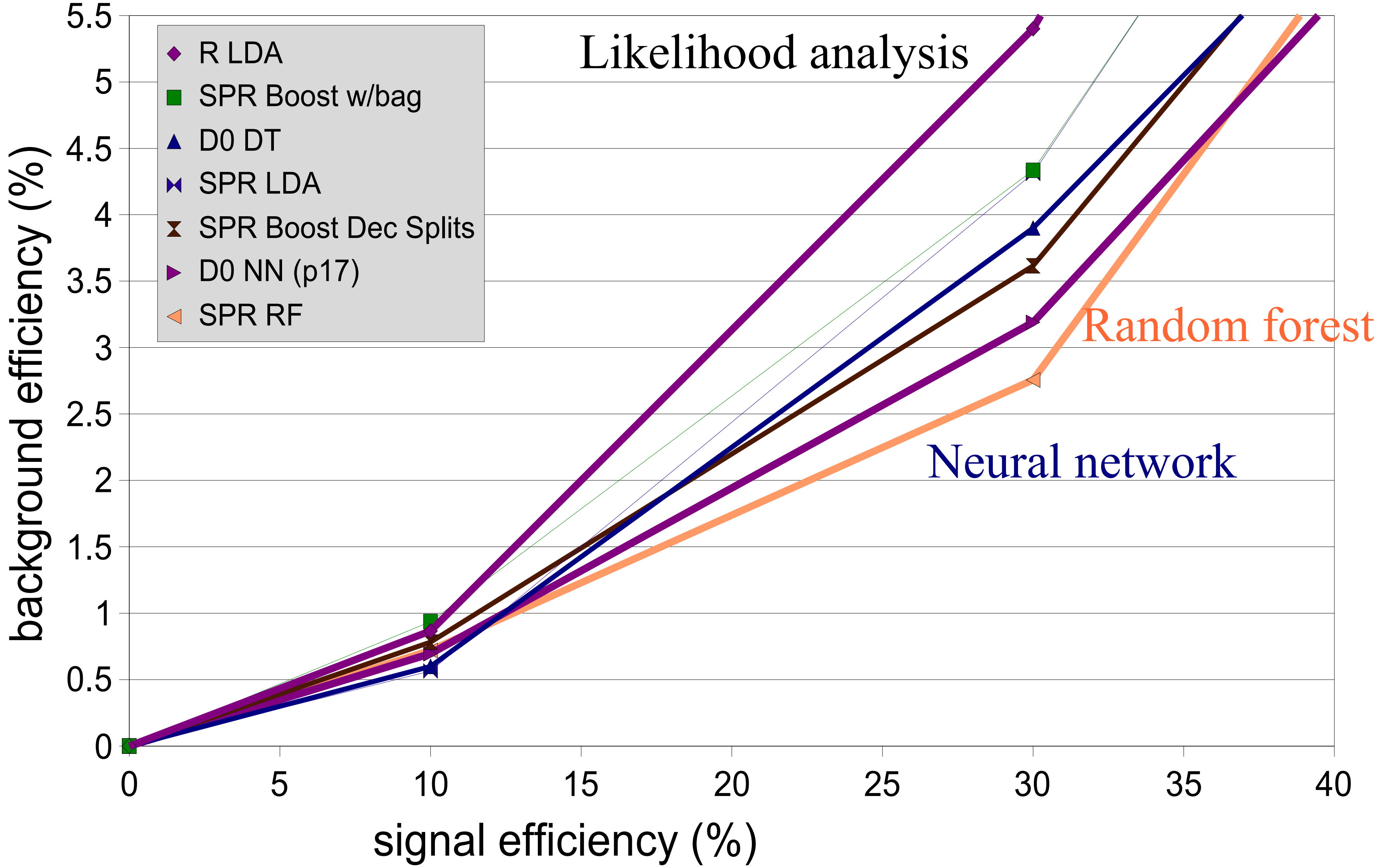
Comparing multivariate methods

How optimal can an
optimal event analysis be?

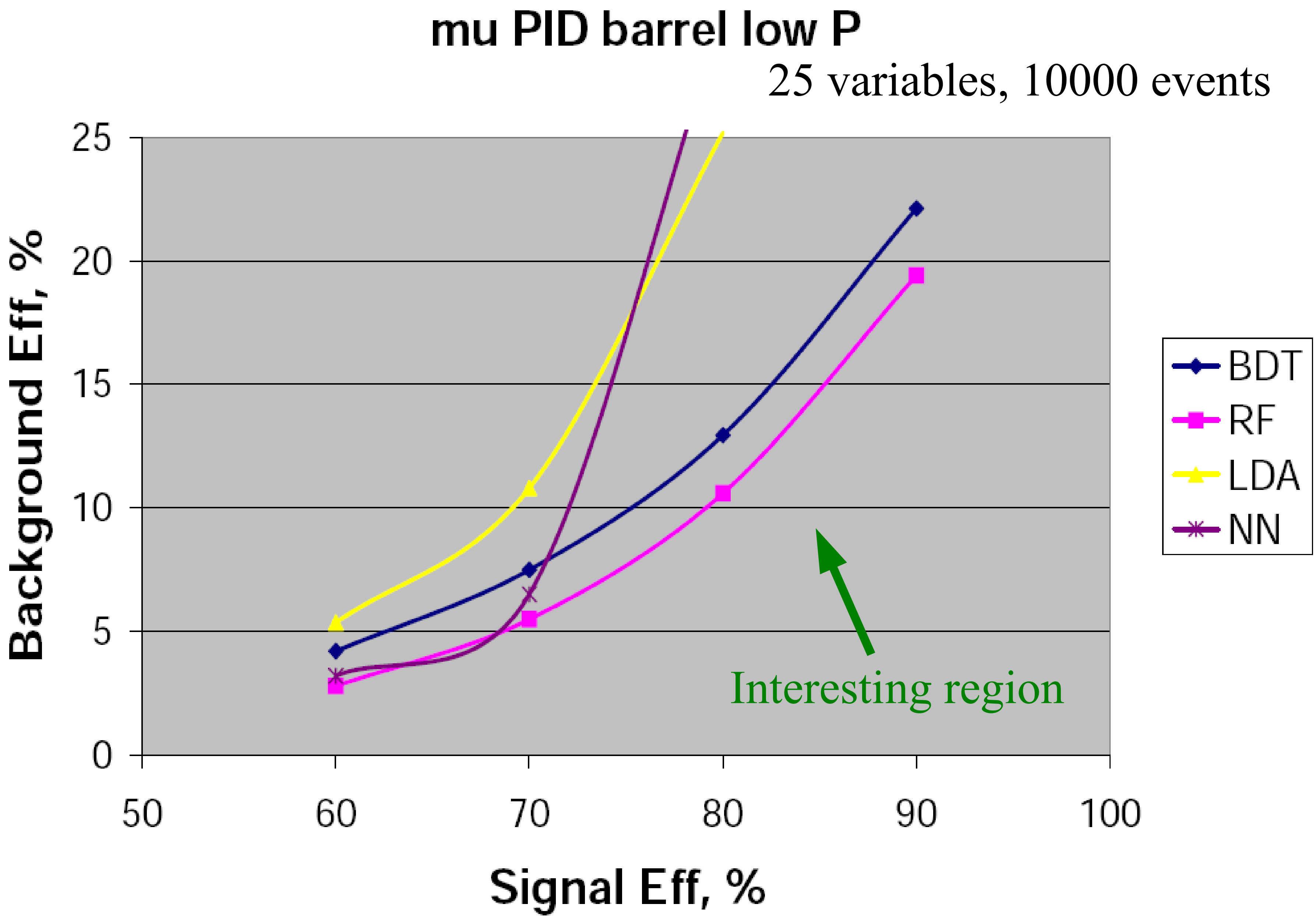
Classifier comparison, DØ single top

Classifiers are evaluated at fixed points, lines connect points for better visibility

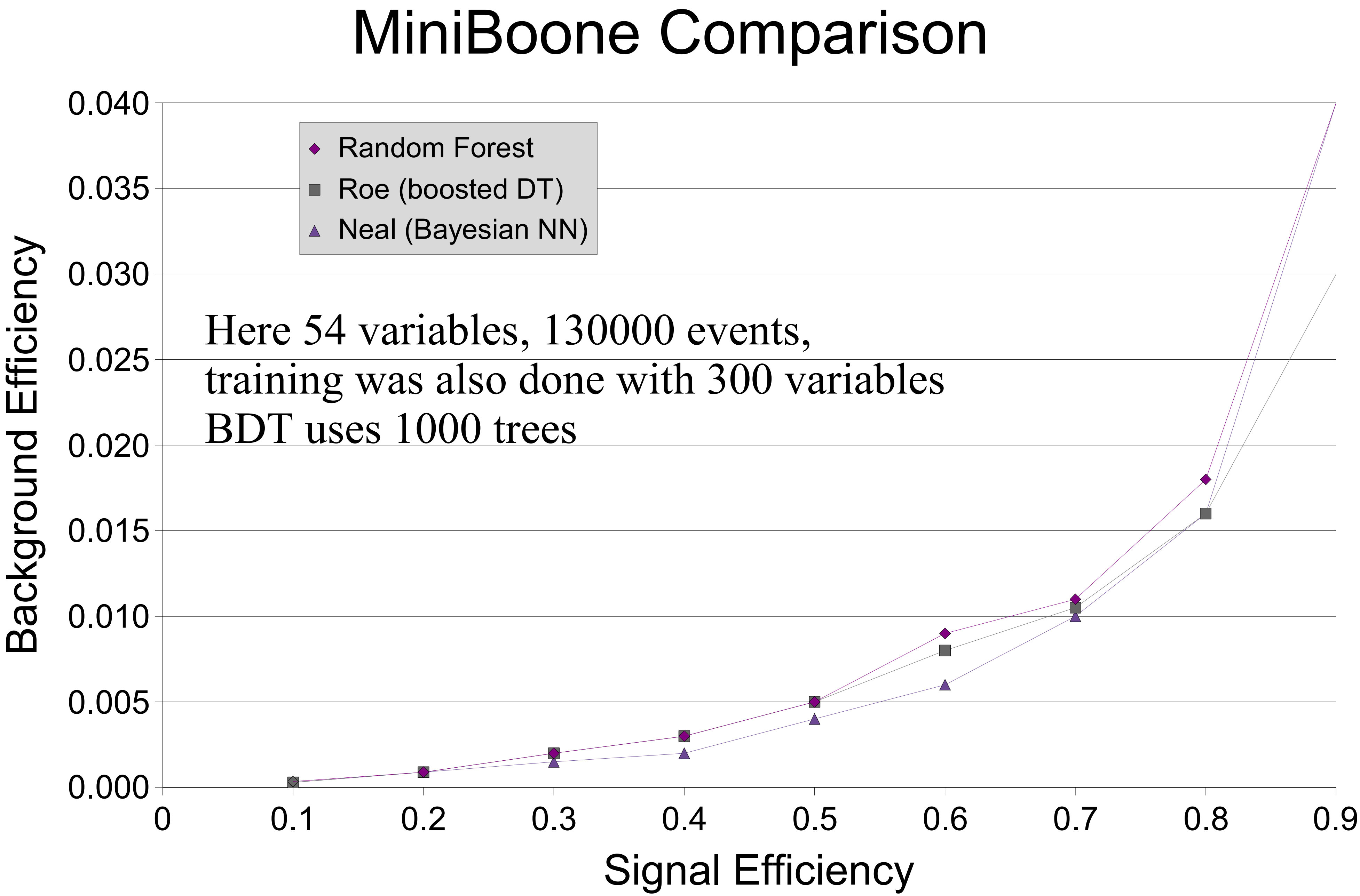
25 variables, 10000 events



Babar Muon ID

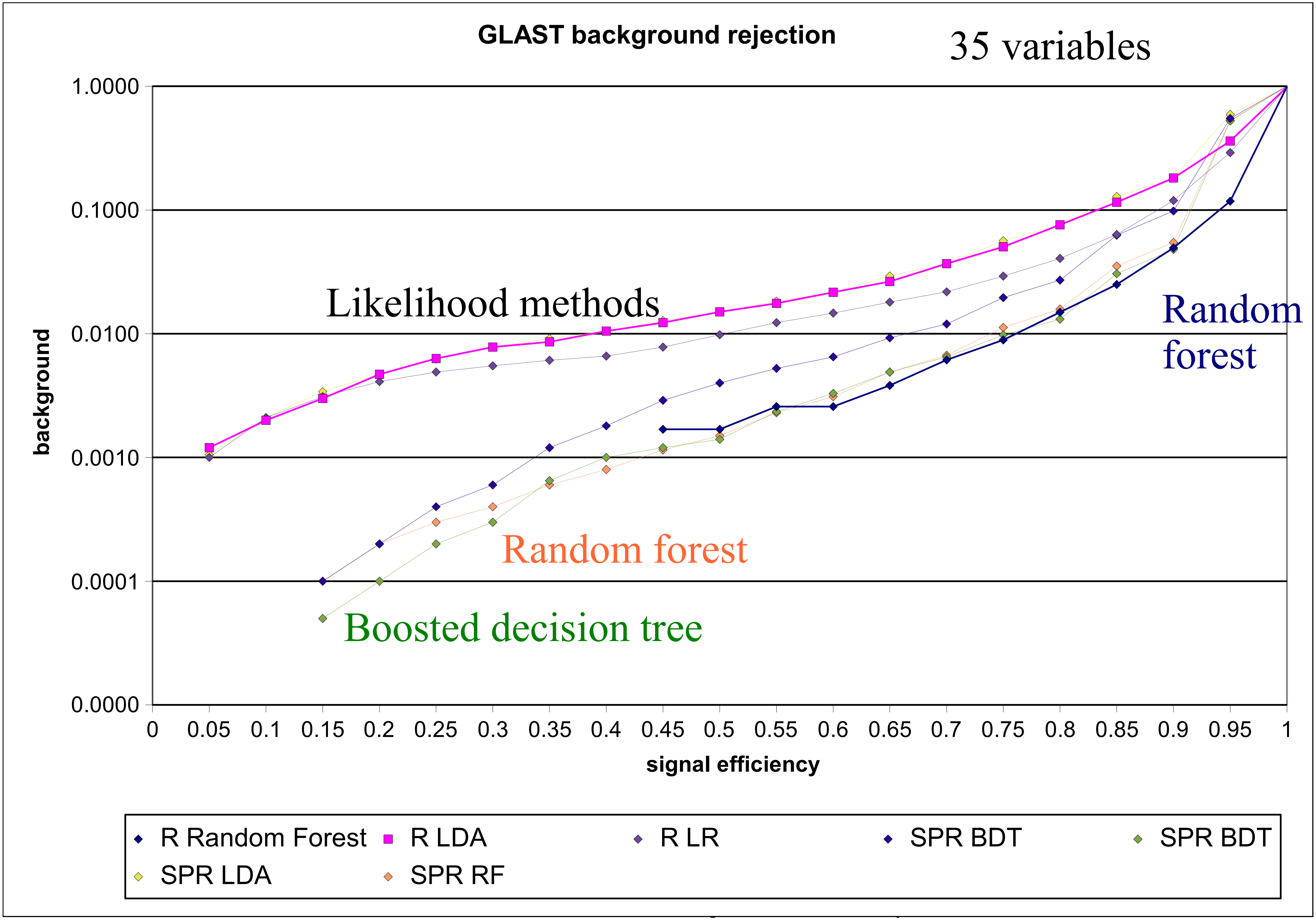


Mini-Boone



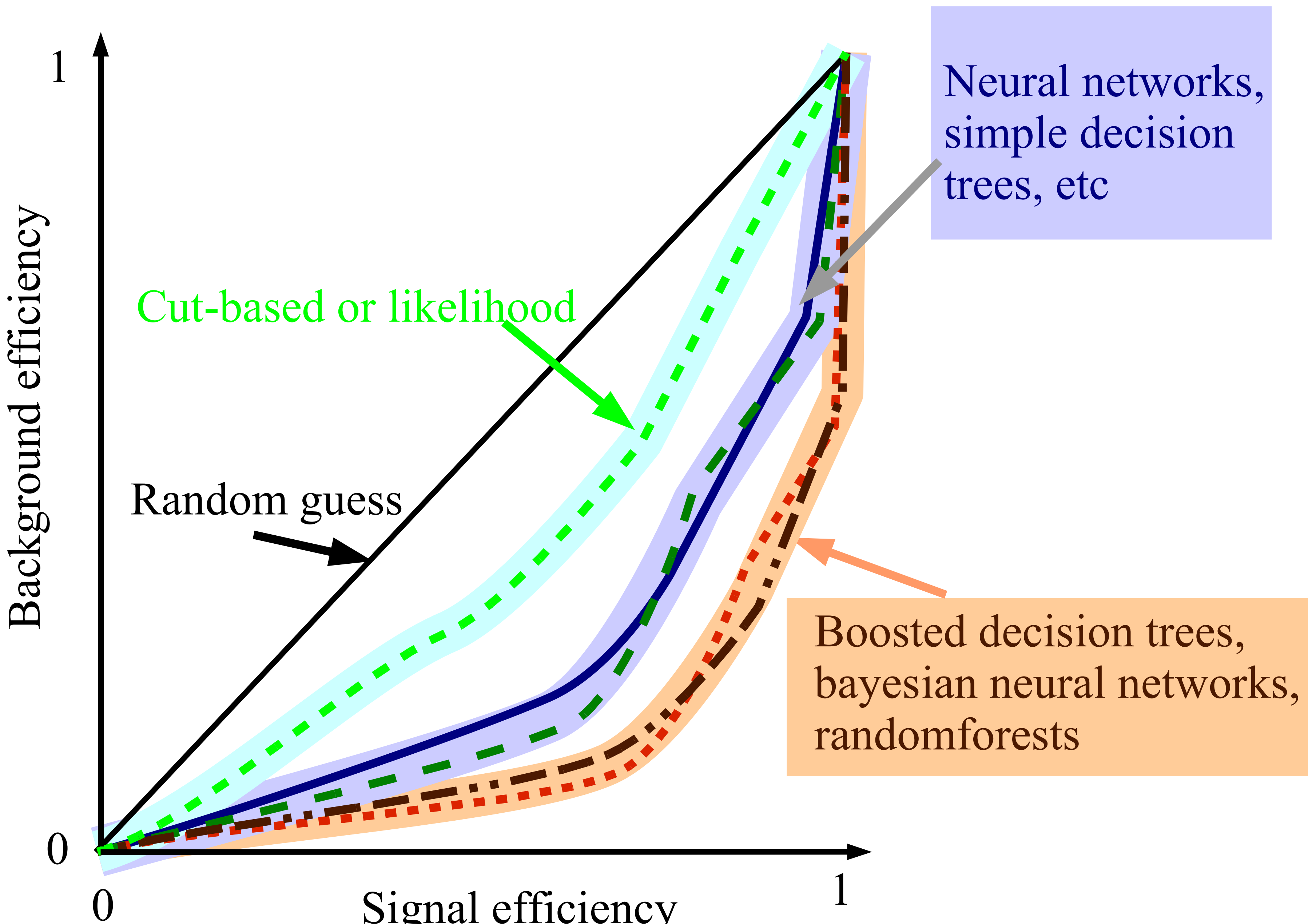
Here 54 variables, 130000 events,
training was also done with 300 variables
BDT uses 1000 trees

Glast



GLAST background rejection
35 variables
1.0000
0.1000
0.0100
0.0010
0.0001
0.0000
background
Likelihood methods
Random forest
Random forest
Boosted decision tree
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
signal efficiency
R Random Forest
R LDA
R LR
SPR BDT
SPR BDT
SPR LDA
SPR RF

Summary



1
0
Background efficiency
0
1
Signal efficiency
Cut-based or likelihood
Random guess
Neural networks, simple decision trees, etc
Boosted decision trees, bayesian neural networks, randomforests

Conclusions

- Multivariate event analysis techniques are now a common tool in HEP
  - In the past mostly neural networks, now also decision tree-related methods
    - Glast, MiniBoone, Atlas, Dzero
- Modern classification tools make life easy
  - Very few parameters to adjust
  - Can use many variables
    - Ranking of variables automatically provided
  - Implemented in several software packages

Conclusions

- Multivariate event analysis techniques are now a common tool in HEP
  - In the past mostly neural networks, now also decision tree-related methods
    - Glast, MiniBoone, Atlas, Dzero
- Modern classification tools make life easy
  - Very few parameters to adjust
  - Can use many variables
    - Ranking of variables automatically provided
  - Implemented in several software packages

Advanced event analysis enables discoveries

Resources

- PhyStat code repository
  https://plone4.fnal.gov:4430/P0/phystat/
- PhyStat 2007 conference
  http://phystat-lhc.web.cern.ch/phystat-lhc/
- Jim Linnemann's collection of statistics links:
  http://www.pa.msu.edu/people/linnemann/stat_resources.html
- Statistical analysis tool R
  http://www.r-project.org/
- TMVA (multivariate analysis tools in root)
  http://tmva.sourceforge.net/
- Neural Networks in Hardware
  http://neuralnets.web.cern.ch/NeuralNets/nnwInHep.html
- Boosted Decision Trees in MiniBoone
  http://arxiv.org/abs/physics/0508045
- Decision Tree Introduction
  http://www.statsoft.com/textbook/stcart.html
- GLAST Decision Trees
  http://scipp.ucsc.edu/~atwood/Talks%20Given/CPAforGLAST.ppt