

Comparing Classifier Performance

Toby Burnett, Washington

Nicolai Meinshausen, Berkeley

Ilya Narsky, Cal Tech

Reinhard Schwienhorst, MSU

Brigitte Vachon, McGill

Goal

- Comparison of different classifiers for as many datasets as possible
- Using standard tools
- In one evening

Datasets

- MiniBoone particle ID (Byron Roe)
- D0 single top quark search
 - T-channel vs $t\bar{t}$ background
 - T-channel vs W^+ jets background
- Babar muon ID
 - Low p_T
 - High p_T
- Glast
 - Not yet started

Classifiers

- Random Forest
 - In R and in StatPatternRecognition
- Linear Discriminant Analysis
- Logistic Regression
- Bayesian Neural Networks
- Boosted Decision Trees
 - Optimized (Byron Roe)
 - “out of the box” (Ilya Narsky)
- Neural networks and single decision tree
 - Optimized for specific analysis

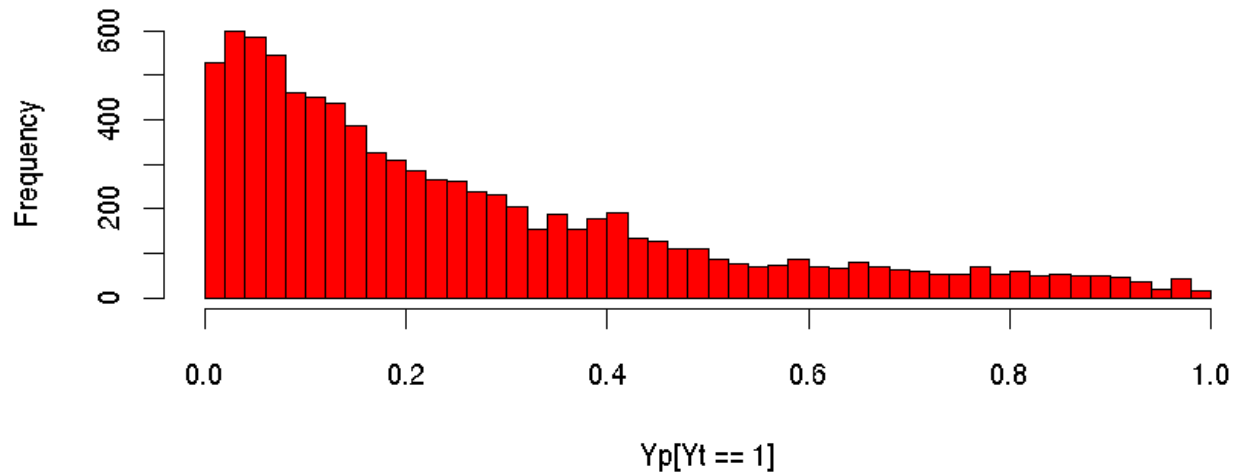
Classifier Packages

- R
 - Easy to install, use
 - Lots of packages
 - Good documentation
- StatPatternRecognition
 - Installable with some effort
 - Example programs (not interactive)
 - Fast
- Individually optimized packages
 - MiniBoone: Byron Roe
 - D0 single top quark search: NN (MLPFit), 1DT

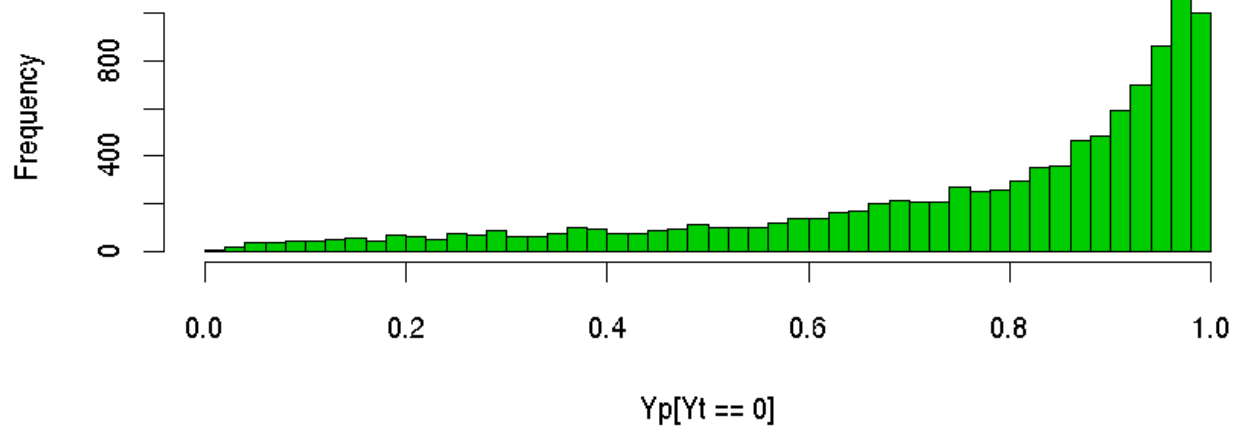
Random Forest in R, D0 single top – ttbar

proof that a physicist (with some help) can learn how to use R in a few hours

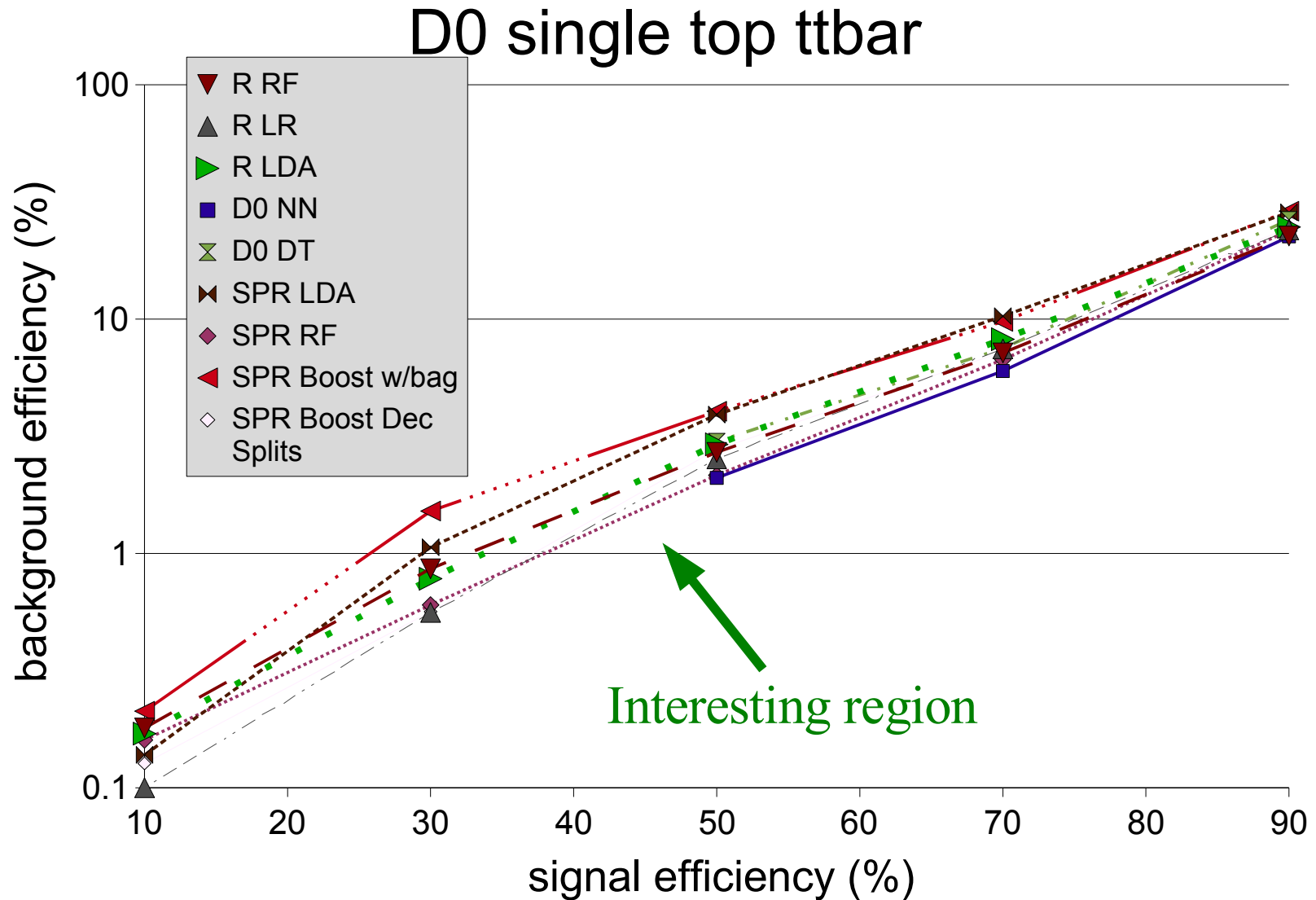
background



signal

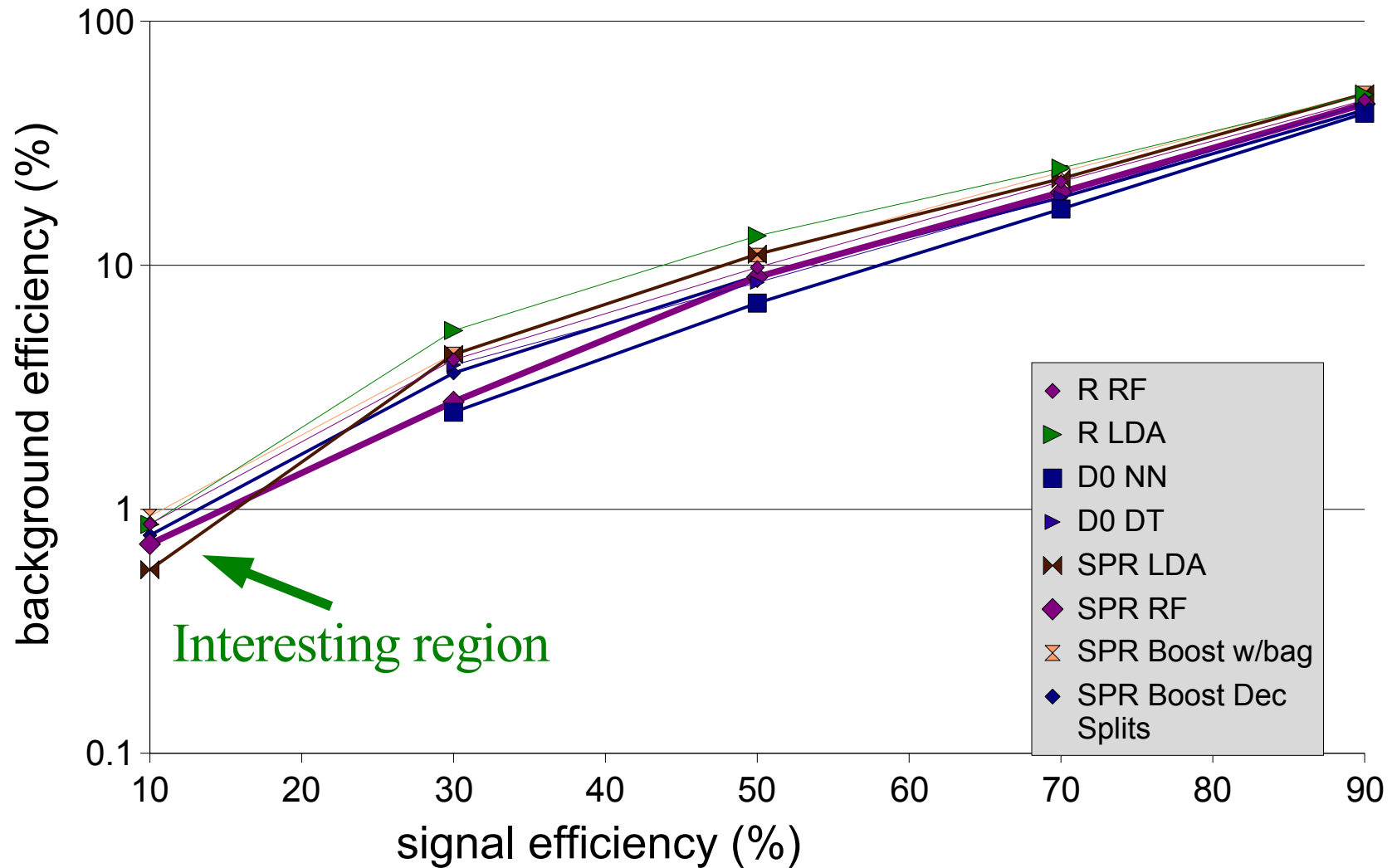


Comparison for the D0 single top tt sample



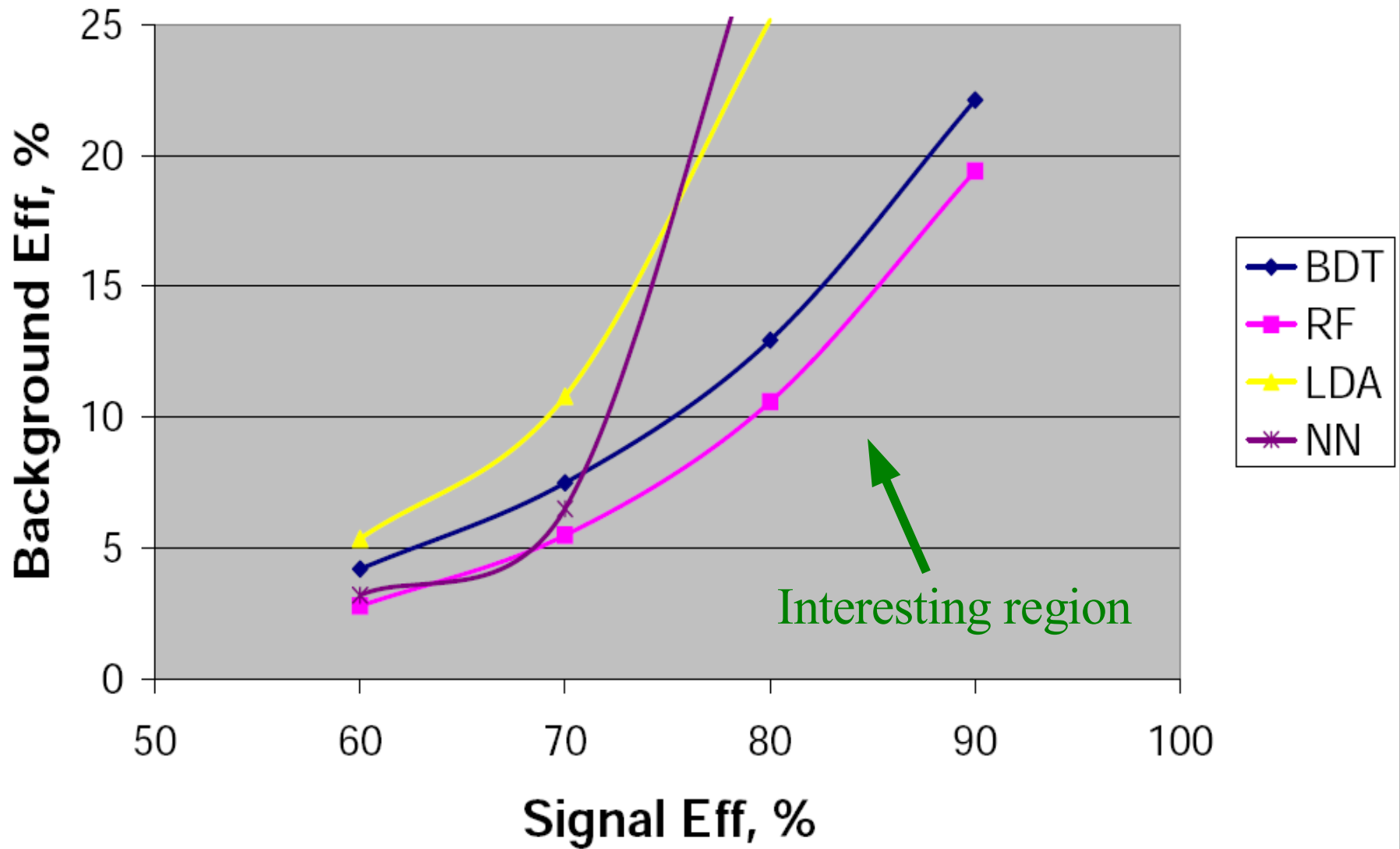
Comparison for the D0 single top Wbb sample

D0 single top Wbb



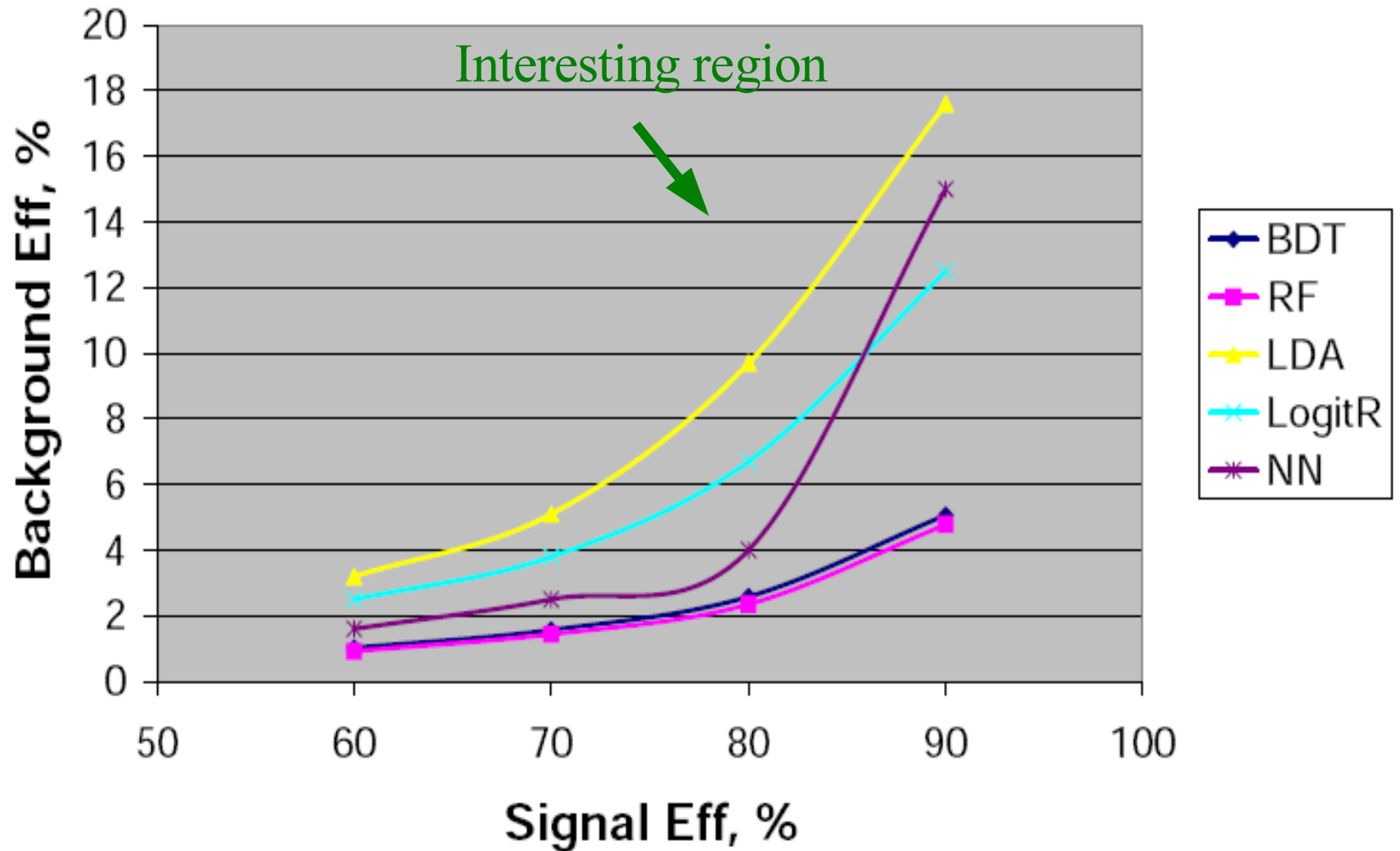
Babar Muon ID

mu PID barrel low P

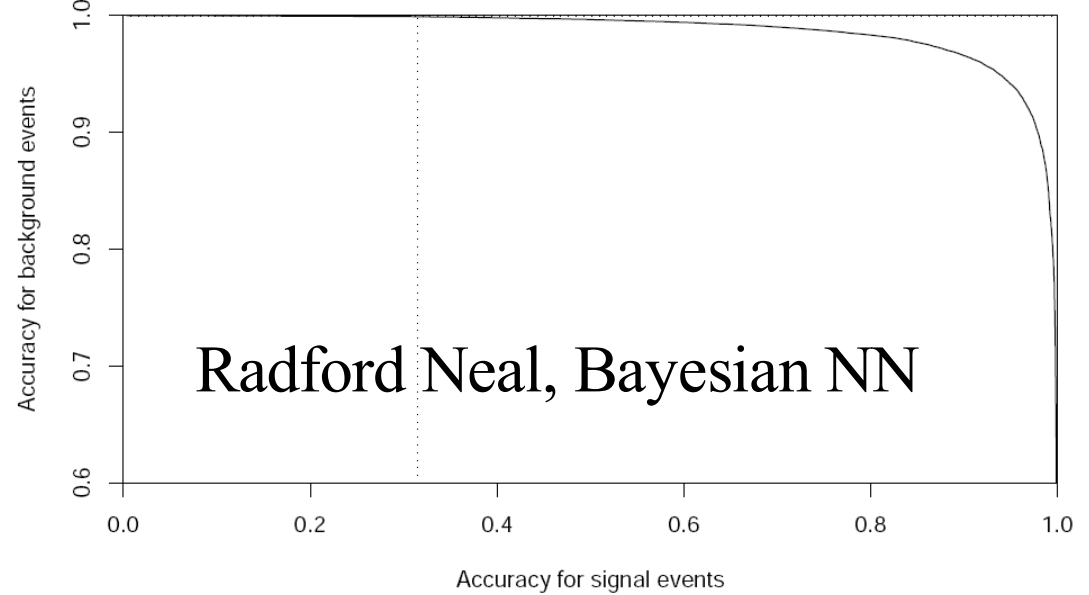
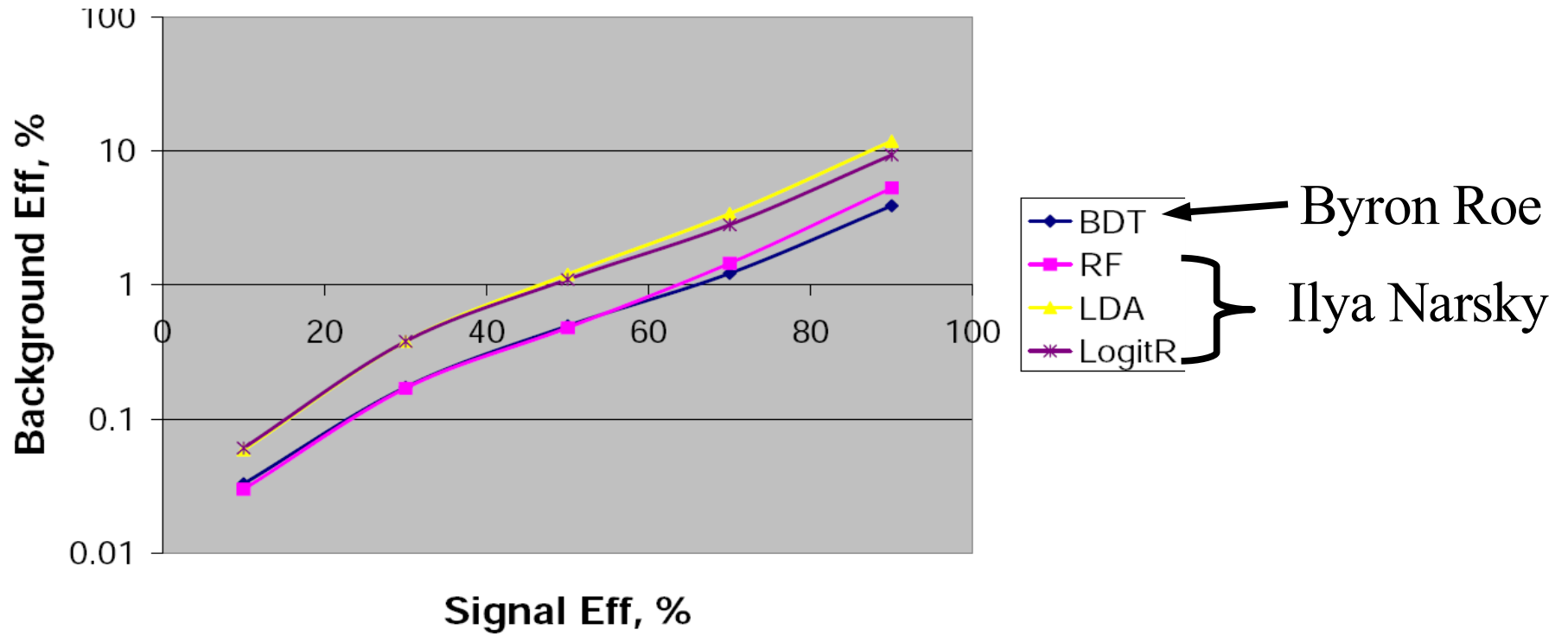


Babar Muon ID

mu PID barrel high P



MiniBoone



Notes

- Quick optimization steps can yield improvements
 - 30% improvement from leaf size 5 -> 10
- Some samples come with weights for individual events
 - R random forest cannot deal with them
 - But performance is still decent

Summary

- We have tested many classifiers on several datasets
 - It is possible for physicists to use statistical tools
 - Each classifier typically takes a few minutes to run
- Typically, boosted decision trees and random forests do best
 - Bayesian NN similar
 - Logistic regression slightly worse
 - Linear discriminant analysis typically worse
- Carefully optimized methods (NN, (b)DT) perform as well or better than out-of-the-box methods shown here