Paradigm – Decision Making Framework for HEP+

Sergei V. Gleyzer,

Harrison B. Prosper

Florida State University, Tallahasee Florida

Introduction

PARADIGM is a ***TOOL*** that helps make decisions based on critical information

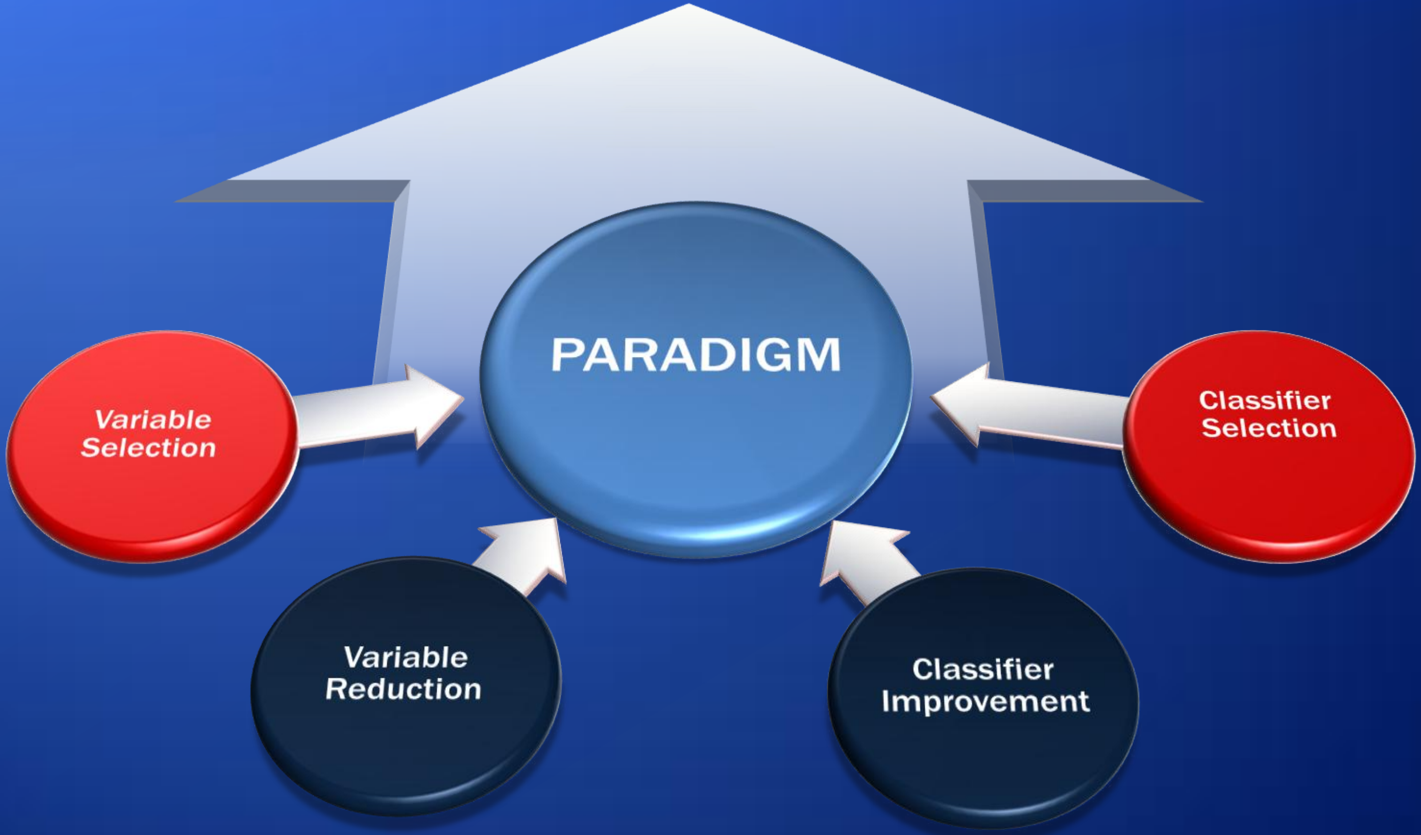
- NO EXPERT KNOWLEDGE REQUIRED

What PARADIGM is NOT:

- Something that makes decisions for you blindly

Choose your

Decision Making Tools



Classifiers and PARADIGM

- PARADIGM is classifier-choice independent
  - Any classifier type can be chosen (Neural Nets, Decision Trees, Rule Ensemble and so on) as long as some form of a performance measure can be assigned to all classifiers
    - For Ex. area under the ROC curve
- We encourage the use of many different classifiers
- PARADIGM provides a way to compare and select an optimal choice for one's analysis
  - More about this topic later in the talk

Tasks: Divide and Conquer

Variable Selection and Reduction

Do better with less

Classifier Relevance, Performance & Improvement

Improve what you have

Analysis Optimization

Are you ready to do your best analysis?

Variable Selection & Reduction

Criteria developed for better:

Variable selection

Crucial to improvement of an on-going analysis

Variable reduction

Safety First

Minimize Loss to Analysis

Variable Selection & Reduction

Relevant PARADIGM criteria:

Relative Variable Importance RVI

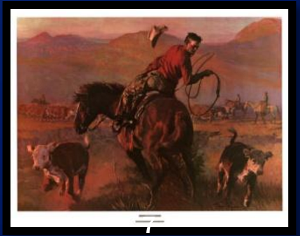
Useful for :

Variable Selection
Classifier Improvement

Global Loss Function GF

Useful for :

Variable Reduction
Classifier Selection



Relative Variable Importance

- **Definition:** The quality (positive or negative) that renders variables desirable or valuable relative to other variables for a particular goal.
  - One possible goal: separation of signal and background
- Proportional to performance of classifiers in which variables participate
- **Given:**

Full set of variables {V}

Variable subsets {S}

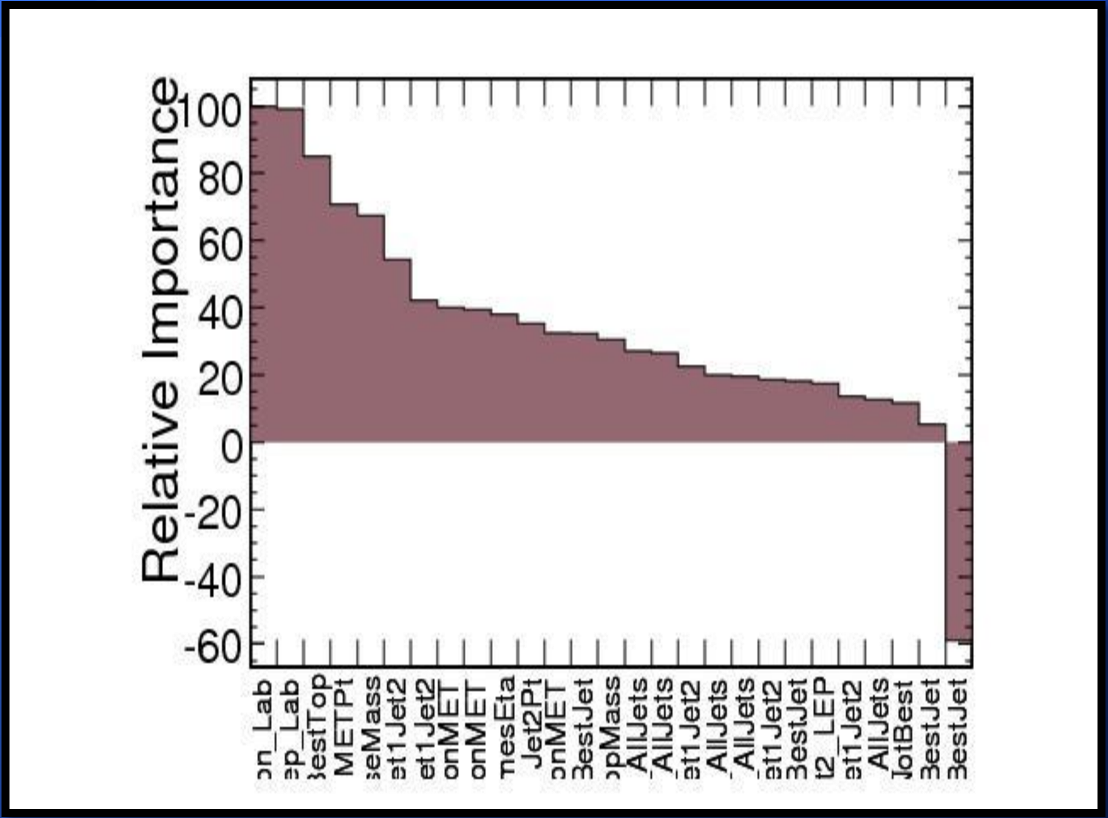
Classifier performance measure F(S)

$$RVI(i) \equiv \sum_{S \subset V : i \in S} F(S) * W_i(S)$$

$$W_i(S) \equiv 1 - \frac{F(S - \{i\})}{F(S)}$$

Amount of classifier performance loss (or gain) if variable i is removed

Typical RVI Plot



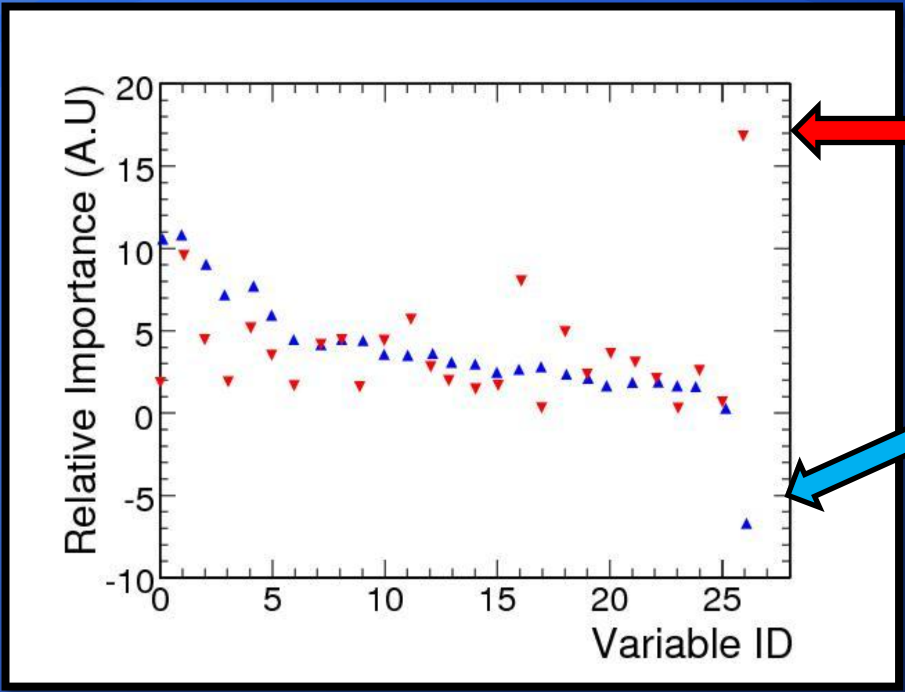
RVI shown in descending order

Rulefit Variable Importance

- Rulefit: rule based binary classification and regression (J. Friedman)
  - Transforms decision trees into rule ensembles
  - A powerful classifier even if some rules are poor
- Variable Importance:
  - Proportional to performance of classifiers in which variables participate (similar to RVI)
  - MAJOR DIFFERENCE:
    - No $W_i(S)$
    - Individual Classifier Performance evenly divided among all participating variables

Rulefit and RVI Comparison

RULEFIT
RVI



Important but **NOT** in a good way

Adverse Variable Identification

Differences:

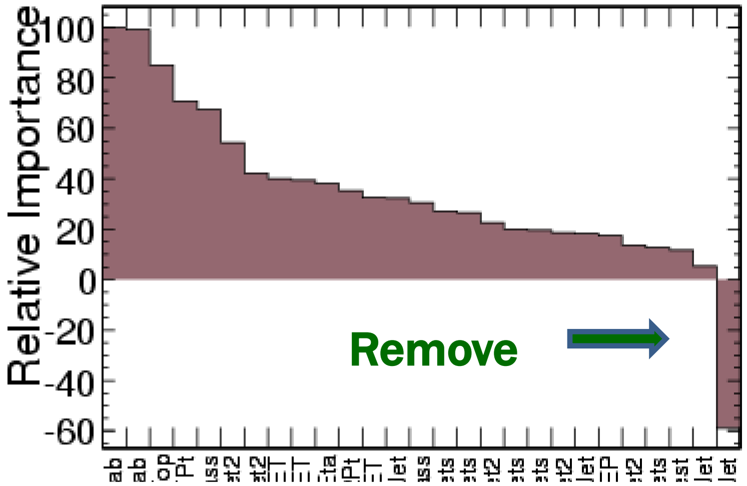
- RULEFIT ⟹ Absolute value importance
- RVI ⟹ True importance
  (if the variable hinders the classification process it is given negative importance)

Caveats for Variable Reduction

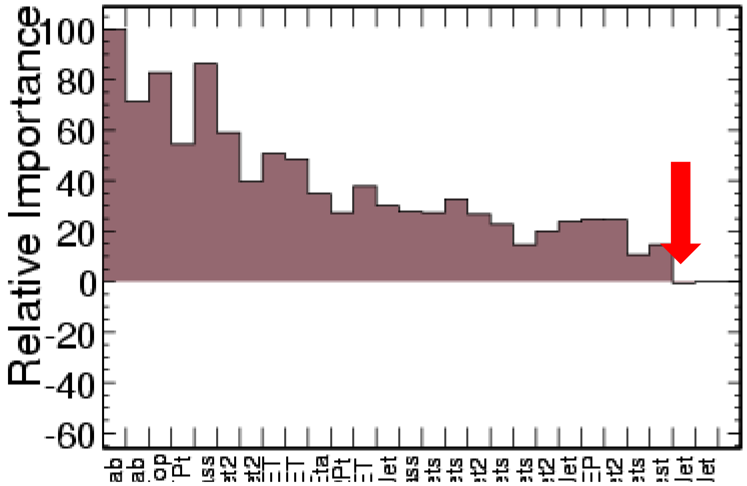
- Variables often strongly interact with others in the classification process. Their removal affects the performance of remaining interacting partners.
  - Strength of interactions quantified by both RULEFIT and PARADIGM
- In some classifiers variables can be overlooked (or shadowed) by their interacting partners

Beware of the hidden reefs

Caveats for Variable Reduction



Before the removal of the adverse variable



After the removal of the adverse variable

Importance Landscape Has Changed

Caveats for Variable Reduction

- RVI and RULEFIT Variable Importance criteria (and others like them) are not consistent enough to be used for variable reduction because of this subtlety
  - Unless a given analysis has negligible variable interactions
- This holds for any such criterion that does not incorporate interactions.

Global Loss Function

- Gloss Function → global measure of loss
- Selects variable subsets for global removal

$$GF(S') \equiv 1 - \frac{\sum\limits_{S \subset (V-S')} F(S)}{2^{|V-S'|}}$$

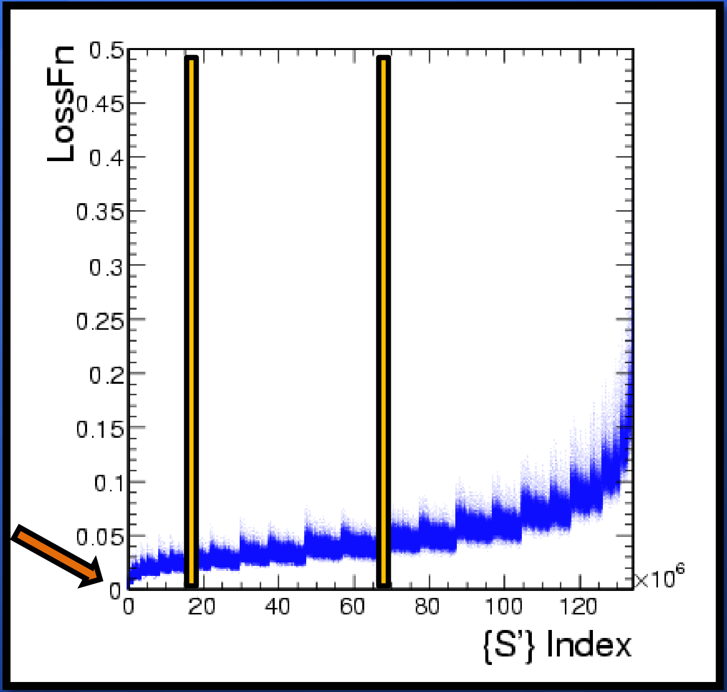
S' is the subset to be removed

- Shows the amount of predictive power loss relative to the upper bound of performance of remaining classifiers

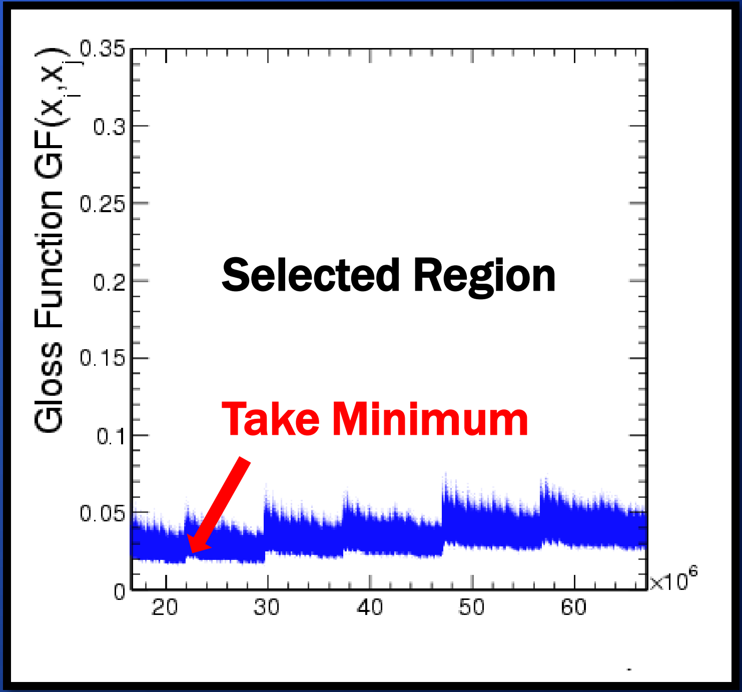
$$\sum_{S \subset (V-S')} F(S)_{\max} = 2^{|V-S'|}$$

Global Loss Function

- The lower the GF ⟶ the lower loss of classification power from removing subset {S'} from {V}



Full Set

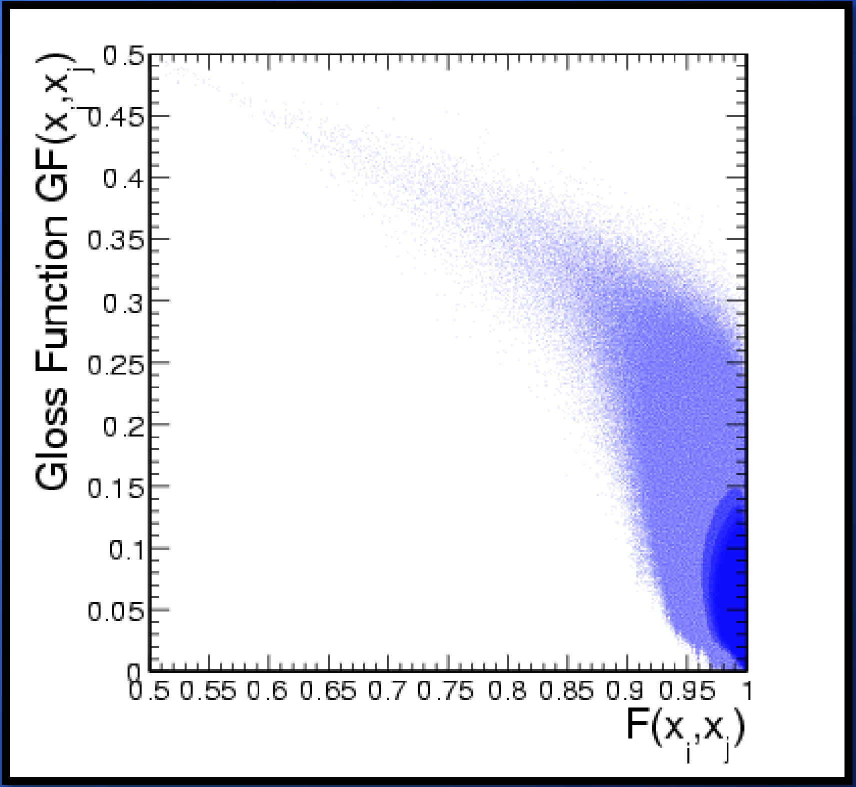


Selected Region

Take Minimum

- Global optimization of predictive performance
- Implicitly incorporates variable interactions

Global Loss and Classifier Performance



Minimization of the Gloss Function NOT EQUIVALENT to Maximization of F(S) – i.e. finding the highest performing classifier and its constituent variables

A Word about Search Algorithms

- Some algorithms attempt a fast search for "high-performing" classifiers by adding/removing variables
- Popular implementations:
  - **Forward Selection** - start with an empty set then add
  - **Backward Elimination** - start with full set then subtract
  - Variations of this sort (add X take away Y)
    - Stop when performance improvement drops
- It is a way of searching for well-performing classifiers but
- Does not amount to finding optimal parameter space for further analysis
- Note: Variable interactions affect search-based criteria as well

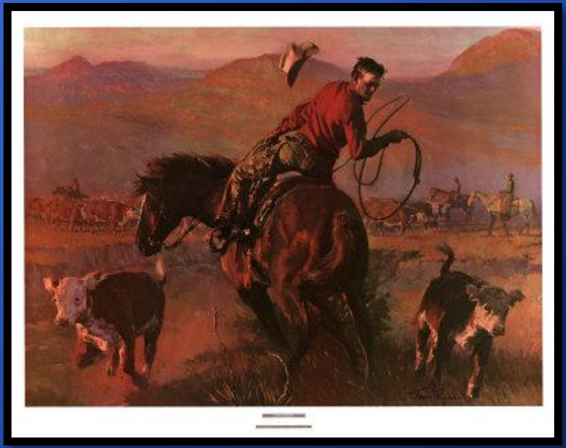
  Same subtlety applies as before – i.e. unsuitable for reduction

Why

- Often in HEP one searches for new phenomena and applies classifiers trained on MC for at least one of the classes (signal) or sometimes both to real data. Limiting to just one classifier (albeit well-performing and possibly boosted at that stage) is an undesirable restriction.
  - Flexibility is KEY for any search
- It is more beneficial to choose a reduced parameter space that consistently produces strong performing classifiers out of its constituents (as in the Gloss Fn approach) with greater flexibility at actual analysis time.

Classifier Selection & Improvement

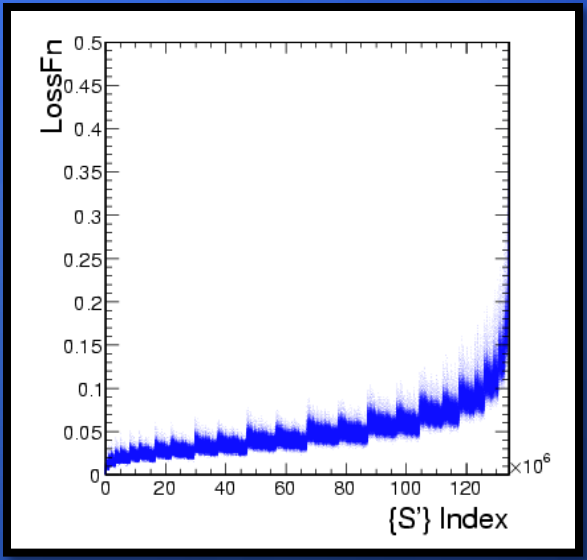
- Quantitatively select optimal classifier for a given task (i.e. search for some new HEP phenomenon)



- Improve classifier performance for this task

Classifier Selection

On the basis of the Gloss Function criteria



Integrate under the Gloss Function

The lower the area under the curve the better the classifier for this analysis task – select classifier

Classifier Improvement

How does one know one relative variable importance criteria is better than the other?

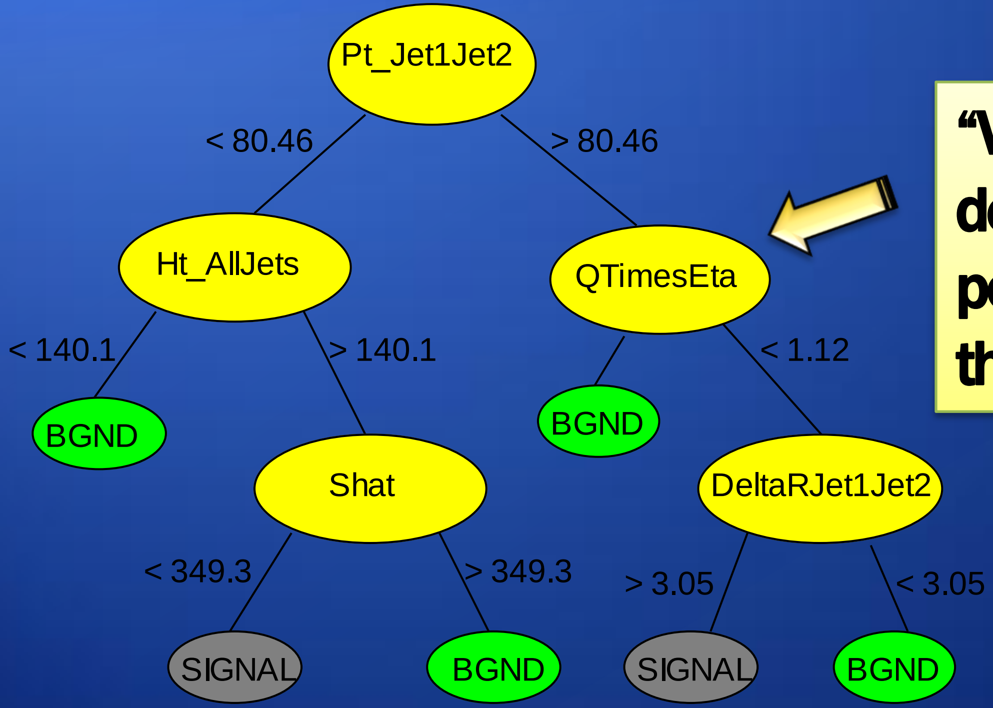
Use it for actual decision making!

Go back to the process of building the classifier and use Relative Variable Importance information to improve it

Classifier Improvement

- Example with decision trees
  - same idea applies to NN and many others

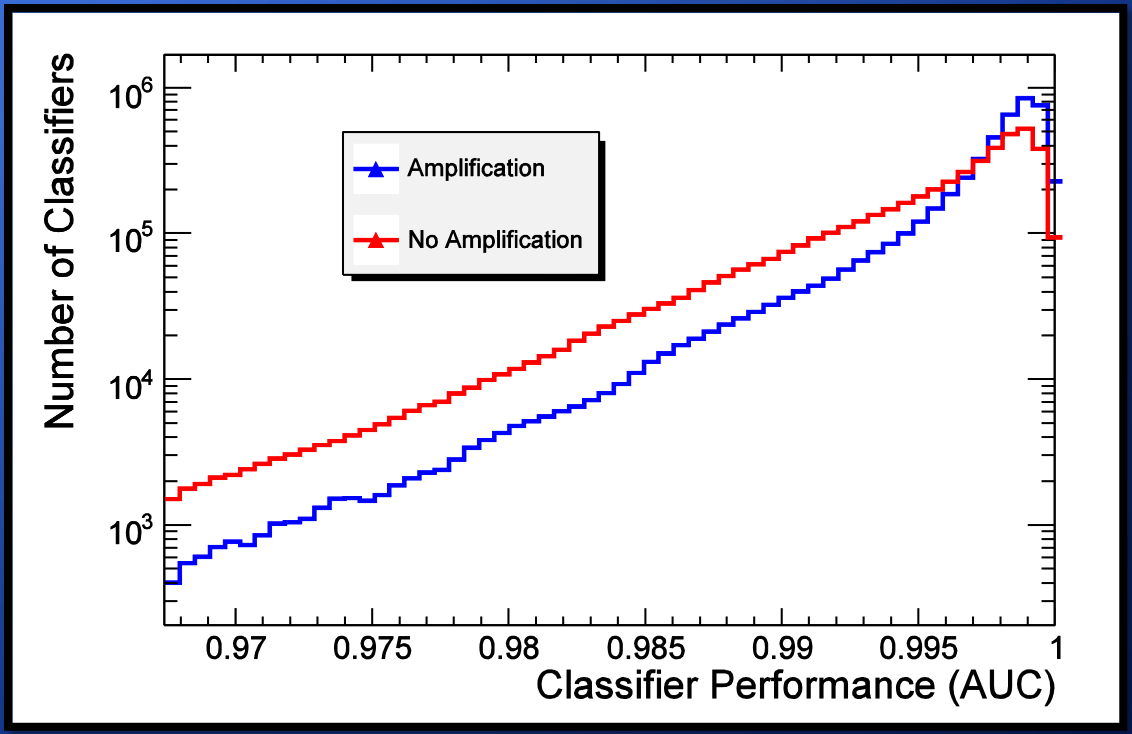
- Decision Tree Reminder:



"Votes" taken at each decision junction on possible splits among the attributes

Classifier Improvement

Next introduce RVI information into the decision making process (the voting process is now weighted by RVI)



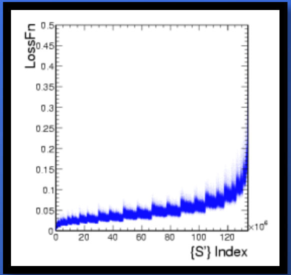
Variable Amplification

- Adding the RVI component to the decision making process quantitatively improves the classifiers (a form of boosting)

- That shows that RVI information is not only useful in its own right but beneficial for further improvement of classifiers used in analysis

Decision Making MAP

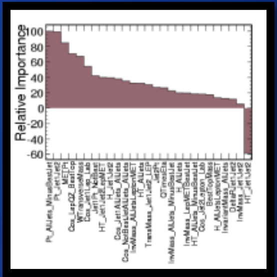
Global
Loss
Function



Reduce Variables
Choose Classifier

Reduced Parameter Space

Relative
Variable
Importance



Improve
Classifier

Solid MVA Analysis

PARADIGM Framework

- Completely parallelized
- Adjusts to user resources for optimization
- Fast algorithms and custom data structures for internal computation
- Other Features:
  - Variable interactions
  - Fast algorithms for
    - RVI Computation
    - Gloss Function Computation
- Let me know if you would like to try it on your favorite analysis

Summary

- PARADIGM is a very robust parallelized framework that provides decision-making information to help with and improve modern day multivariate HEP analyses.

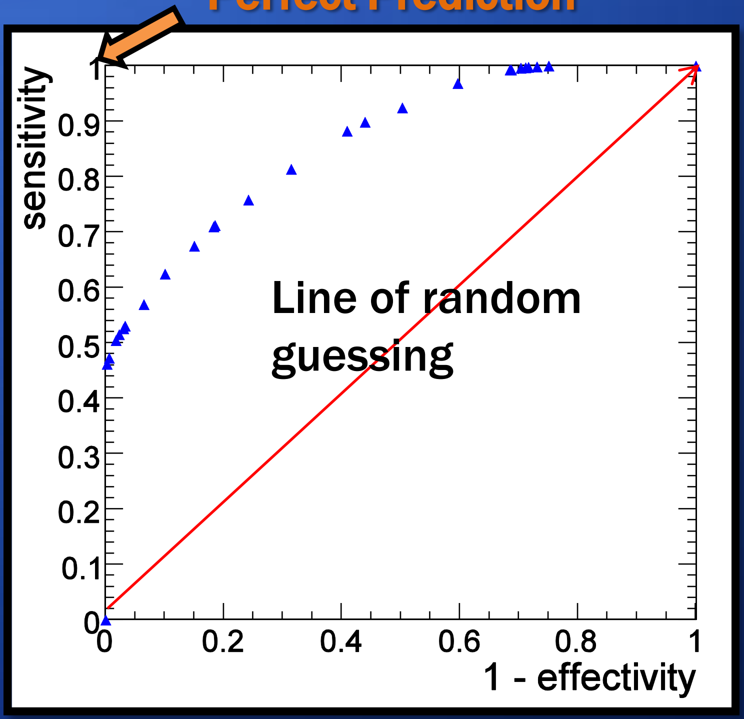
- Areas of application are :
  - Classifier selection
  - Classifier improvement
  - Variable selection
  - Variable reduction

Backups

Reciever Operating Characteristic

- Commonly known as an ROC curve
  - Shows the relationship between correctly classified positive cases (sensitivity) and incorrectly classified negative cases (1 – effectivity)
  - Pioneered in radar signal analysis during WWII

Perfect Prediction



❑AREA UNDER THE ROC Curve (AUC)

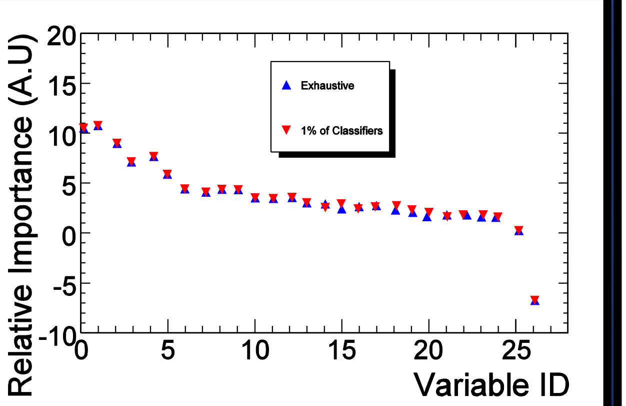
GF and Interactions

- When a subset is selected for removal in the GF procedure all the interaction effects are implicitly accounted for and the loss of predictive power associated with the removal of this subset is known exactly

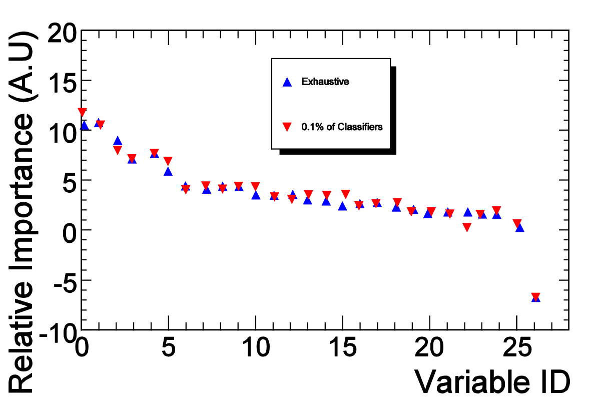
- This is in contrast to making a cut on the RVI or Rulefit Variable importance criteria, where one can not predict the changes to the variable importance landscape after the reduction

Fast Algorithms

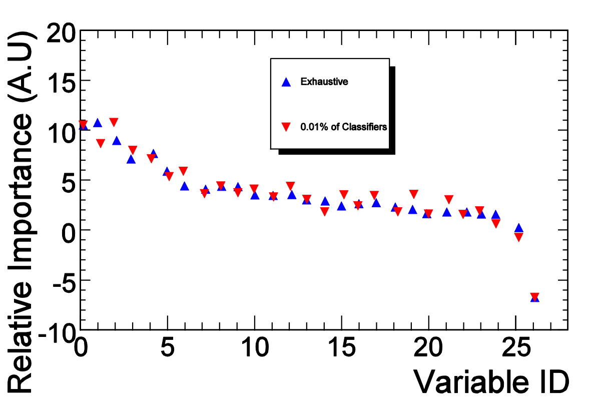
- PARADIGM implements fast algorithms for both RVI and the Gloss Function computation
  - Using random seeds
  - Validated with the full exhaustive procedure



1% of classifiers



0.1% of classifiers

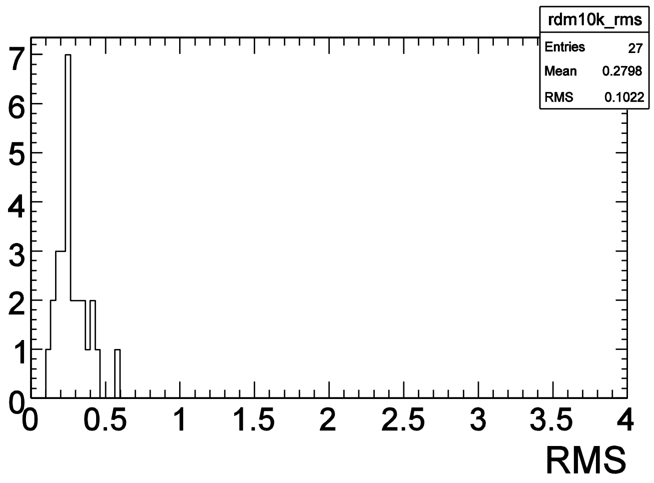


0.01% of classifiers

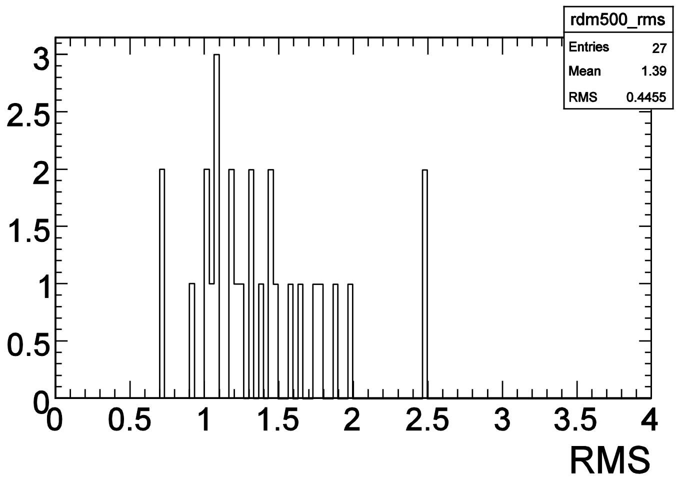
  - Fast algorithms show very good agreement even when a low number of seeds is chosen

10k vs 500 seeds

- Fast RVI algorithms were additionally validated by repeating each procedure 10 times for different number of seeds.



0.1% of classifiers

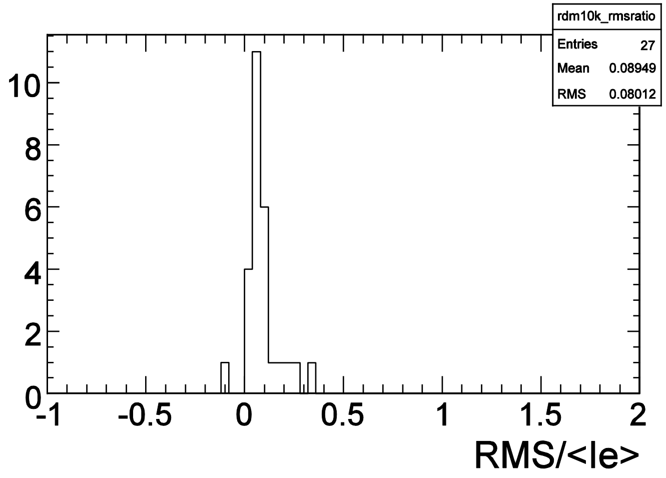


0.005% of classifiers

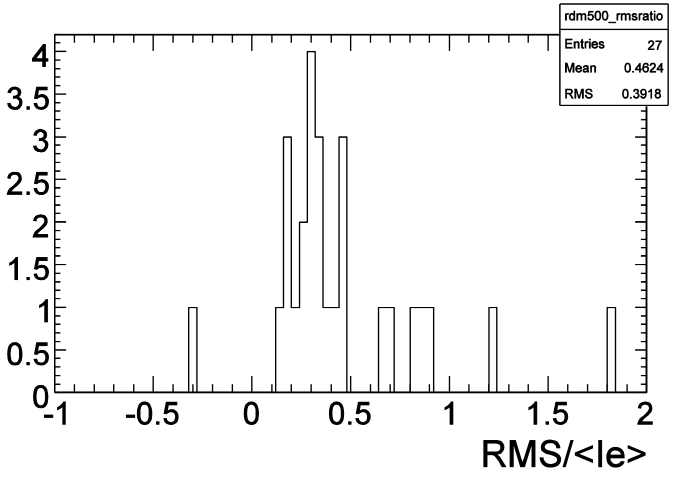
- Above figure shows the rms values for the variations in variable importance.

10k vs 500 seeds

- These two figures show the ratio of the rms to the average RVI value for all variables for 10k and 500 seeds (after 10 repetitions)

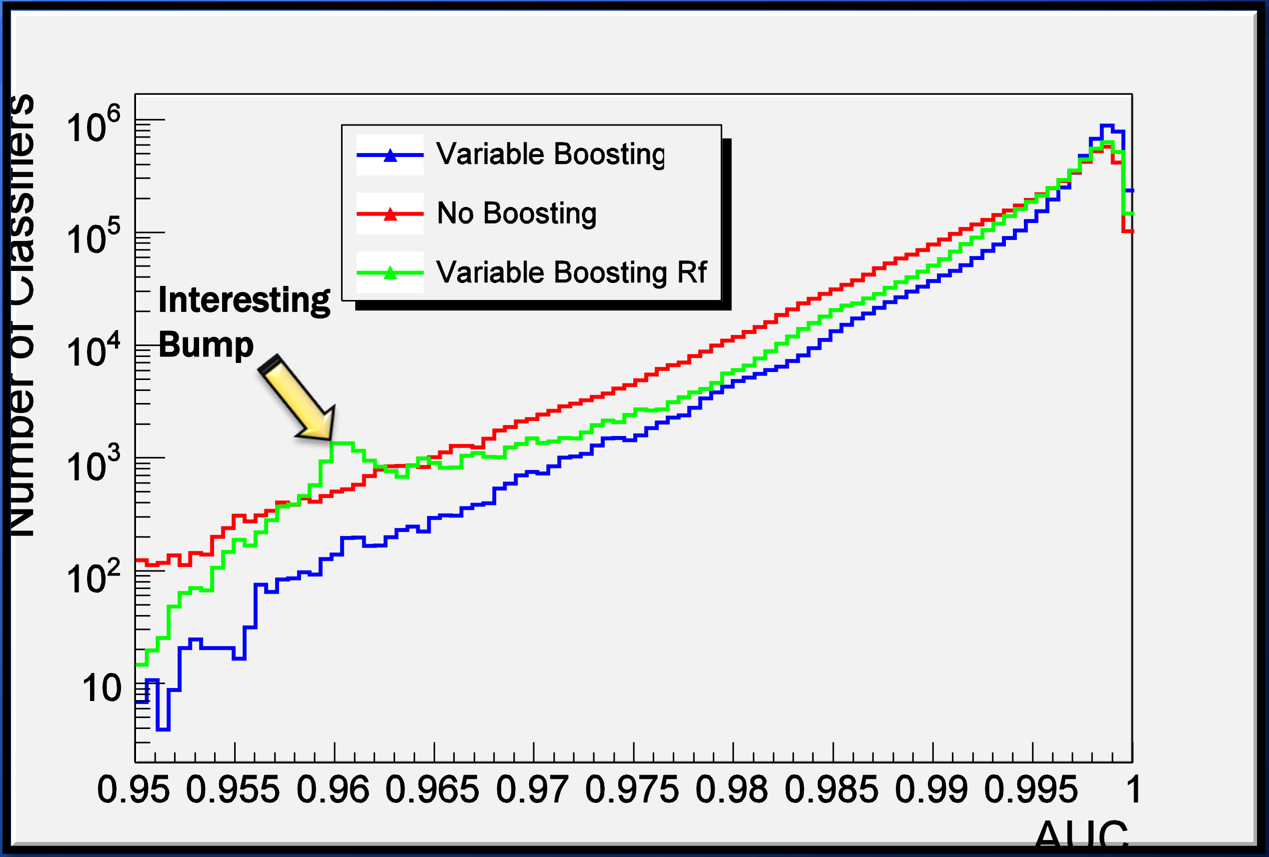


0.1% of classifiers



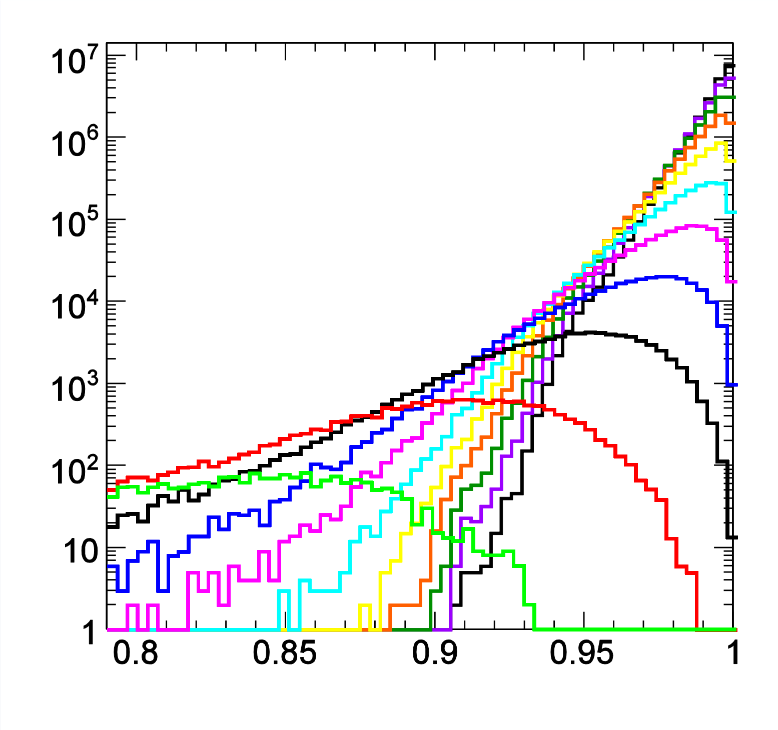
0.005% of classifiers

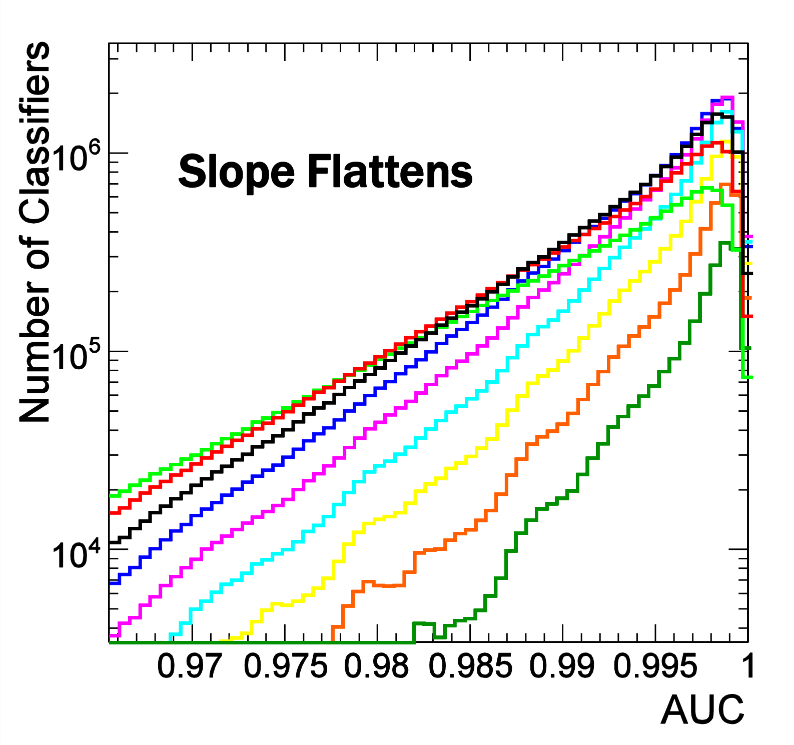
Comparison of Rulefit and RVI



Using Rulefit Variable Importance Clearly Worse than using RVI

Classifier Performance and Cardinality





Cardinality 2-12

Cardinality 10-18