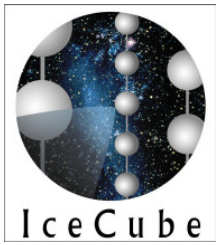
technische universität dortmund

Fakultät Physik
Experimentelle Physik V

Data mining on the rocks



T. Ruhe for the IceCube collaboration,

K. Morik

GREAT workshop on Astrostatistics

and data mining 2011

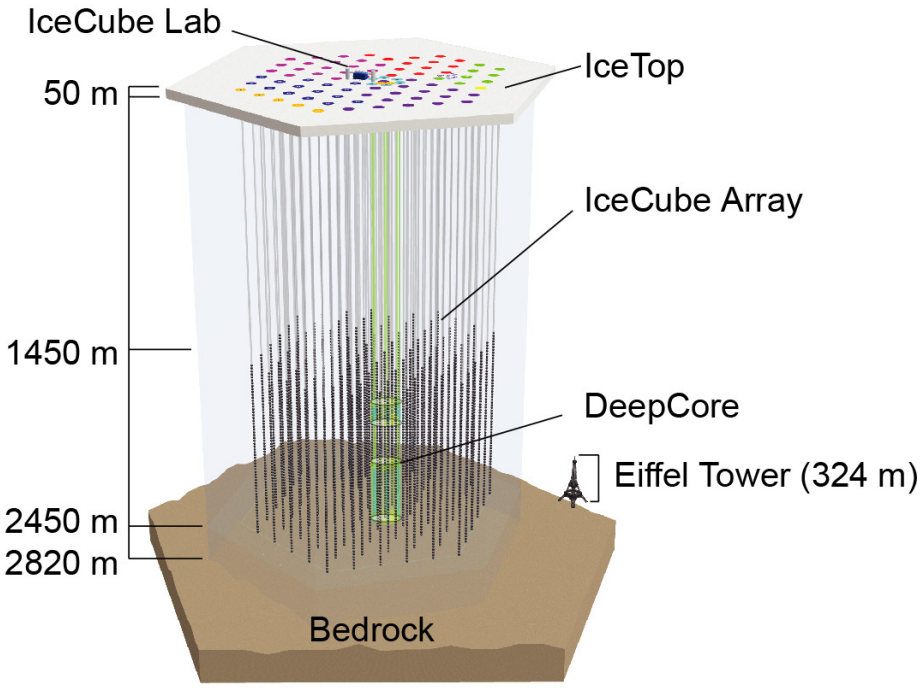
bmb+f - Förderschwerpunkt
Astro-Teilchenphysik
Großgeräte der physikalischen Grundlagenforschung

Outline:

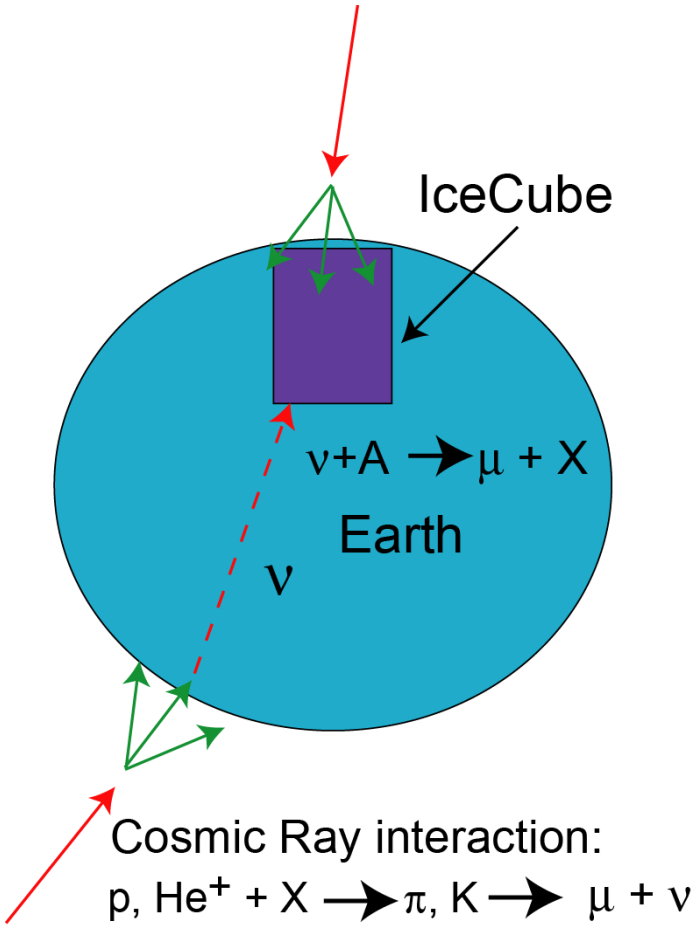
- IceCube, detector and detection principle
- Signal and Background
- Feature Selection and Performance
- Feature Selection stability
- Random Forest, training and testing
- Summary and Outlook

IceCube: The detector

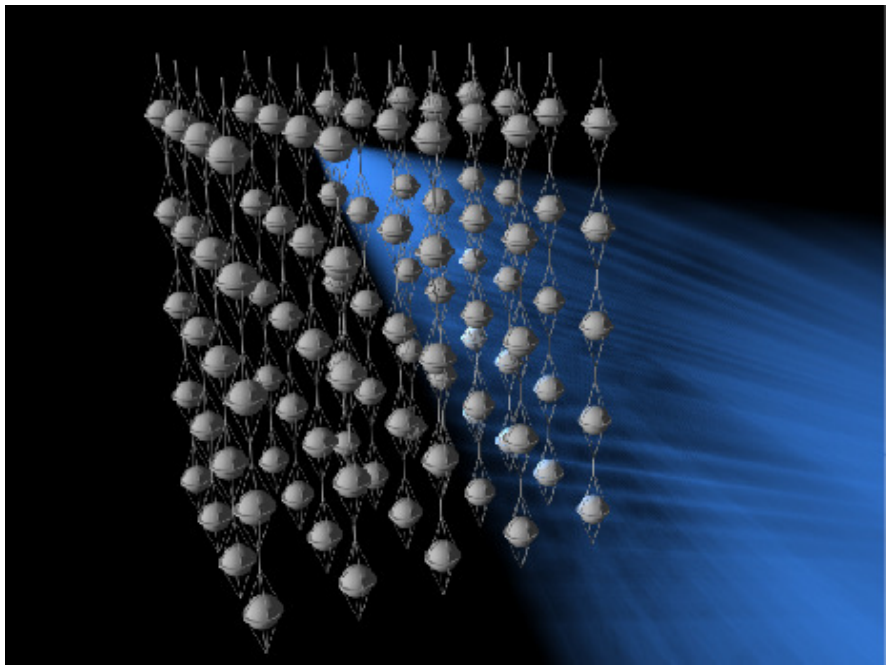
- completed in December 2010
- located at the geographic South Pole
- 5160 Digital Optical Modules on 86 strings
- instrumented volume of 1 km^3
- subdetectors DeepCore and IceTop



IceCube detection principle:

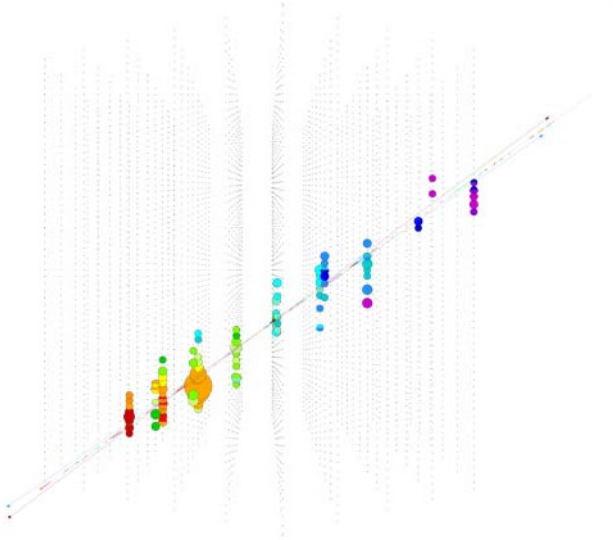


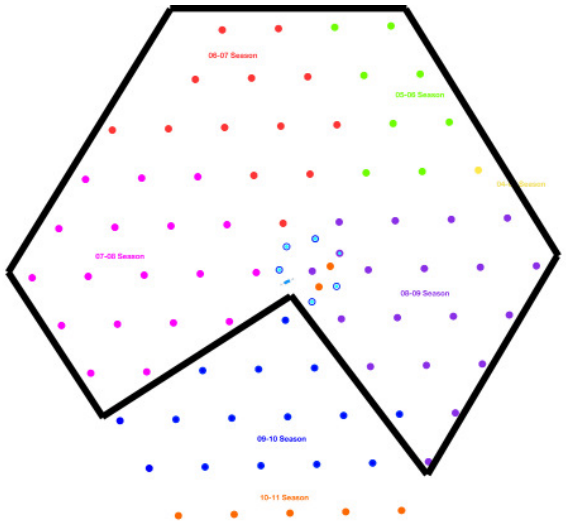
Cherenkov light detected by Digital Optical Modules (DOMs)



IceCube-operation:

- IceCube has taken data in various configurations
- This study: 59-strings
- “Pacman-like“ geometry
- reconstruction based on timing information by the DOMS





Signal and Background:

Signal: atmospheric ν (+ ν from astrophysical sources)

~ 14180 events in 33.28 days of IC-59

Backgr.: (misreconstructed) atmospheric muons

~ 9.699×10^6 events in 33.28 days of IC-59

→ Signal to Background ratio: 1.46×10^{-3}

Application of Precuts:

Signal: 10419 events in 33.28 days of IC-59

Backgr.: 1.68×10^6 events in 33.28 days of IC-59

→ Ratio = 6.2×10^{-3}

RapidMiner:

- Formerly known as YALE
  (**Y**et **A**nother **L**earning **E**nvironment)
- Developed at TU Dortmund University
  (group of K. Morik)
- Operator based
- Open Source, written in java
- Weka operators are fully included
- Quite intuitive to use (personal opinion)

Preselection of parameters:

1. Check for consistency (data vs. simulation)

→ Eliminate if missing in one (reduction ~ 10 – 20 out of ~2600)

2. Check for missing values (nans, infs)

→ Eliminate if number of missing values exceeds 30%
(reduction to 1408 attributes)

3. Eliminate the “obvious“ (Azimuth, Time, galactic coordinates...)
(reduction to 612 attributes)

4. Eliminate highly correlated ($\rho = 1.0$) and constant parameters

→ Final set of 477 parameters

MRMR-Feature Selection:

- **M**aximum-**R**elevance-**M**inimum-**R**edundancy
- iteratively adds to a set F_j a feature according to a quality criterion Q:

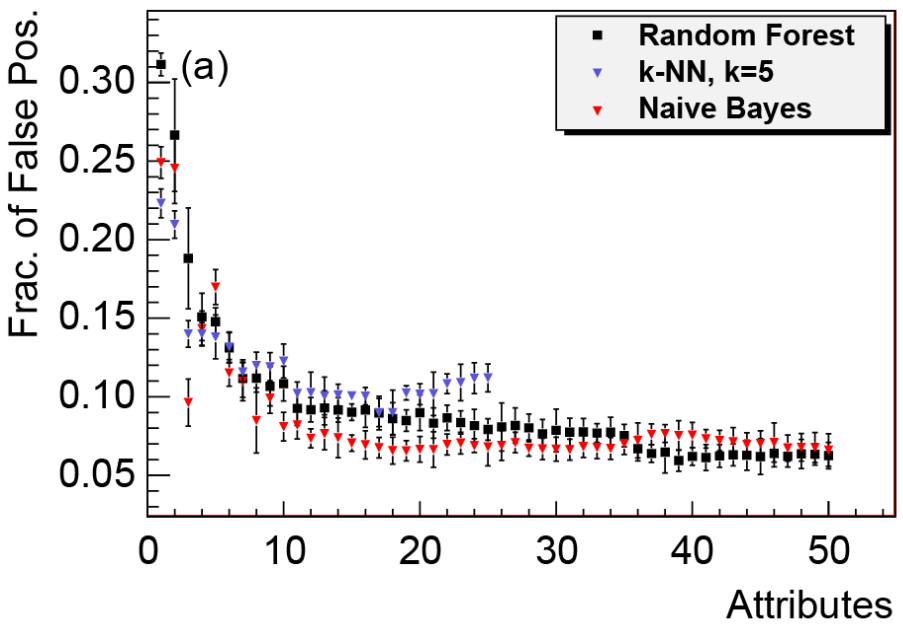
$$Q = rel(y,x) - \sum_{x' \in F_j} red(x,x') \qquad Q = \frac{rel(y,x)}{\sum_{x' \in F_j} red(x,x')}$$

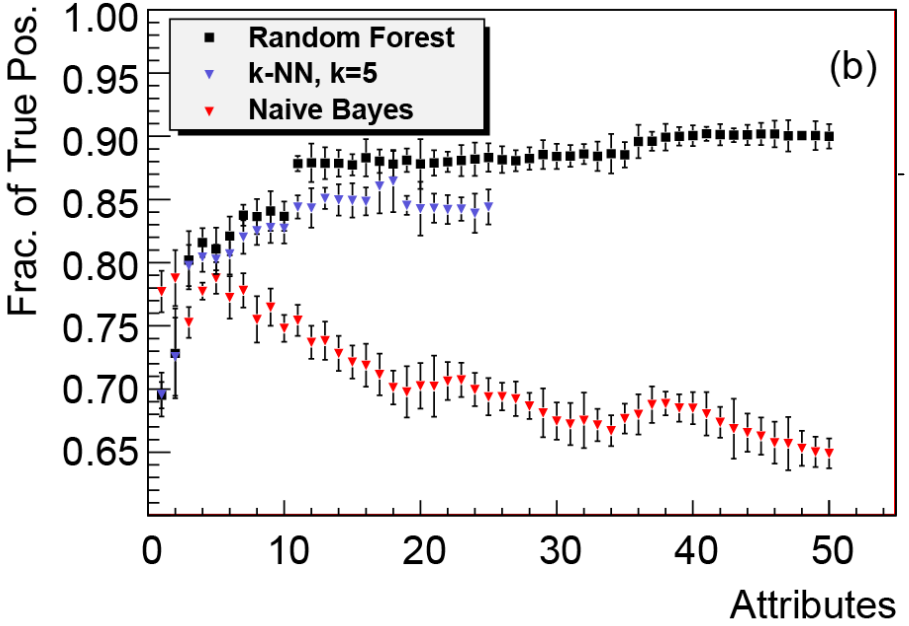
- Relevance and Redundancy automatically map to linear correlation, F-test score or mutual information (depending on attribute and class type)

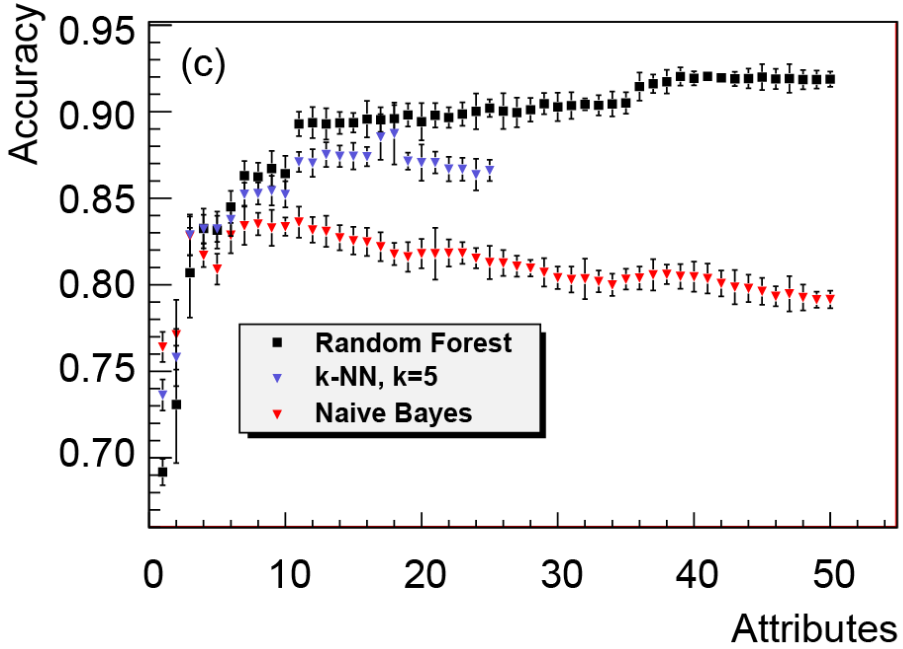
Feature Selection and Performance:

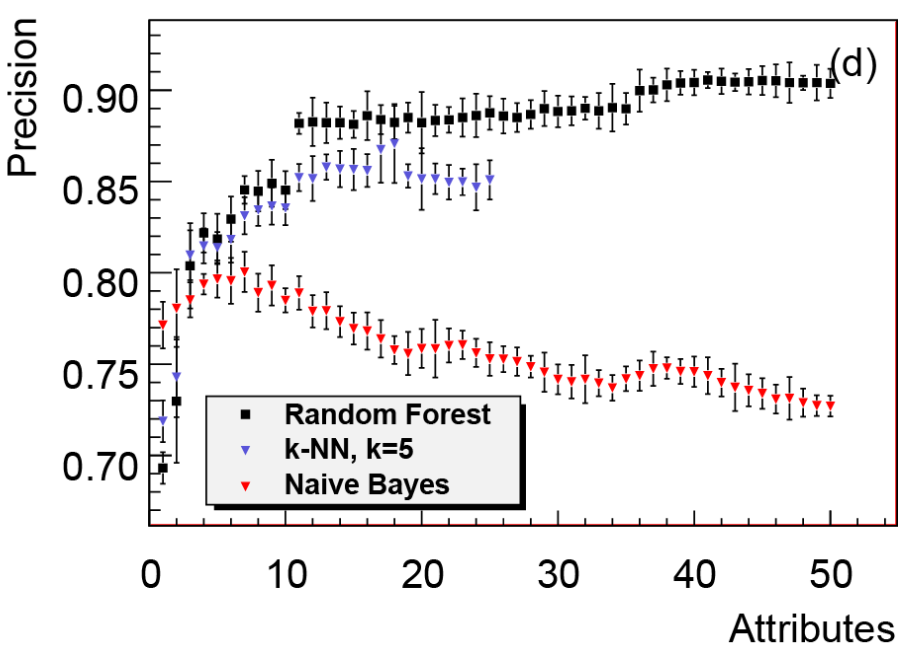
- MRMR Feature Selection
- Random Forest (from the Weka package)

  n_{trees} matched to $n_{Attr.}$ such that: $n_{trees} = 10 \times n_{Attr}$
- Naive Bayes
- k-NN, k =5, weighted vote, mixed Eucledian distance
- 10-fold cross validation
- 3.4×10^5 simulated signal and background events


(a)
Random Forest
k-NN, k=5
Naive Bayes
Frac. of False Pos.
0.30
0.25
0.20
0.15
0.10
0.05
0
10
20
30
40
50
Attributes


(b)
Random Forest
k-NN, k=5
Naive Bayes
Frac. of True Pos.
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0
10
20
30
40
50
Attributes


(c)
Random Forest
k-NN, k=5
Naive Bayes
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0
10
20
30
40
50
Attributes


(d)
Random Forest
k-NN, k=5
Naive Bayes
Precision
0.90
0.85
0.80
0.75
0.70
0
10
20
30
40
50
Attributes

Stability of the MRMR Selection:

Jaccard Index:

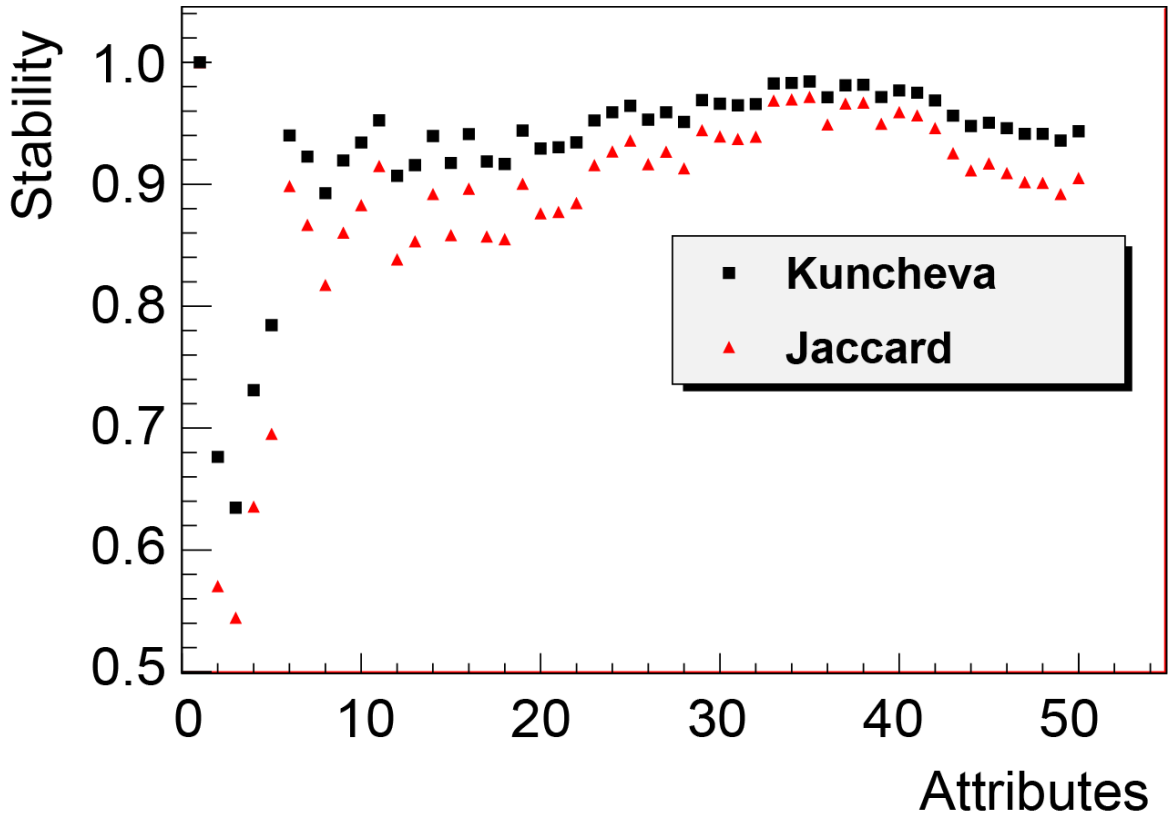
$$J = \frac{|A \cap B|}{|A \cup B|}$$

Kuncheva's Index:

$$I_C(A,B) = \frac{rn - k^2}{k(n-k)}$$

$$|A| = |B| = k$$

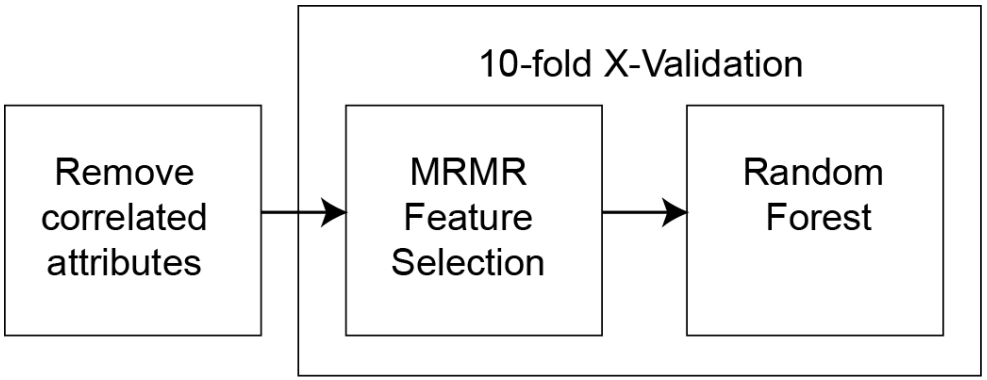
$$r = |A \cap B|$$

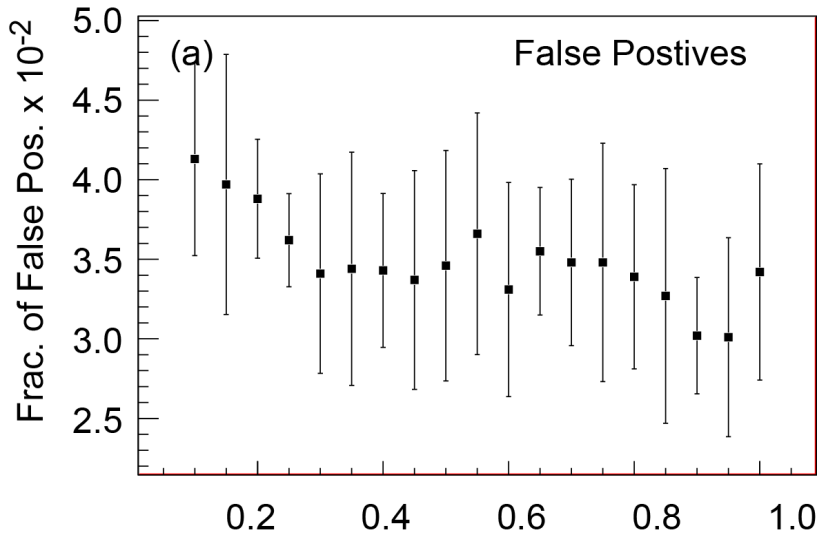


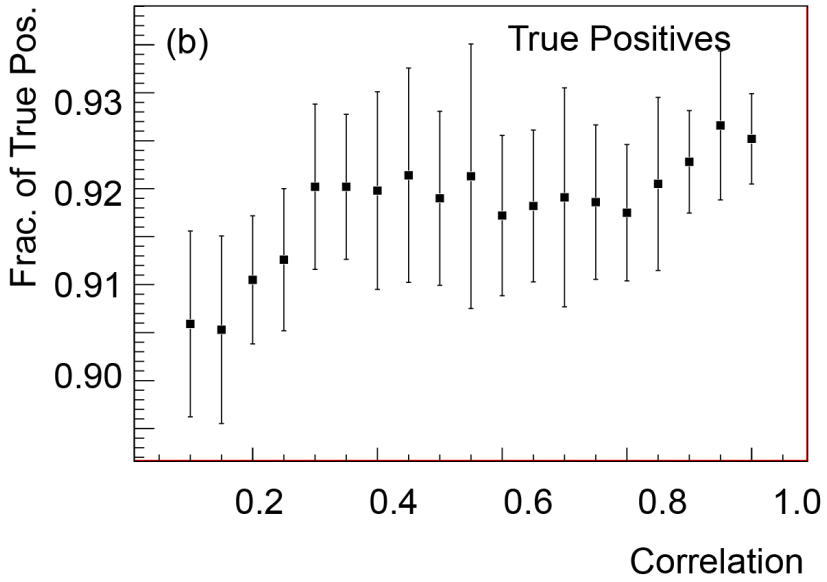
Removing further correlations:

- after MRMR still some correlated attributes

→ Study the performance if further correlations are removed



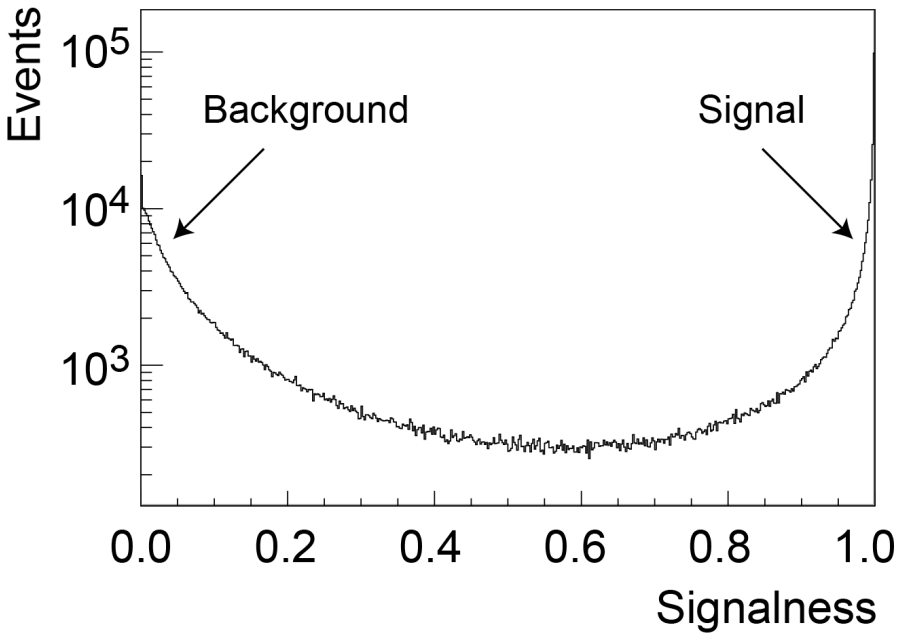


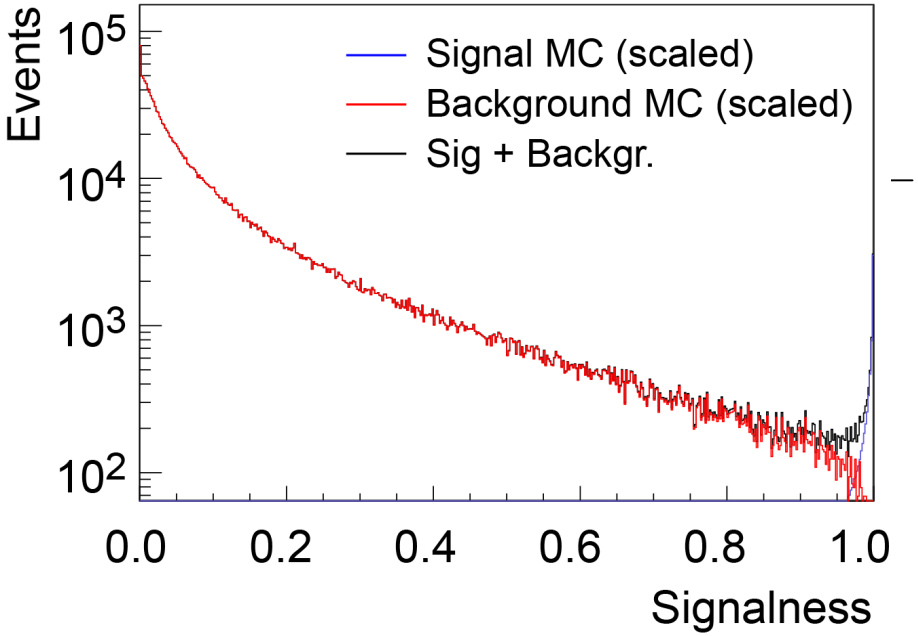


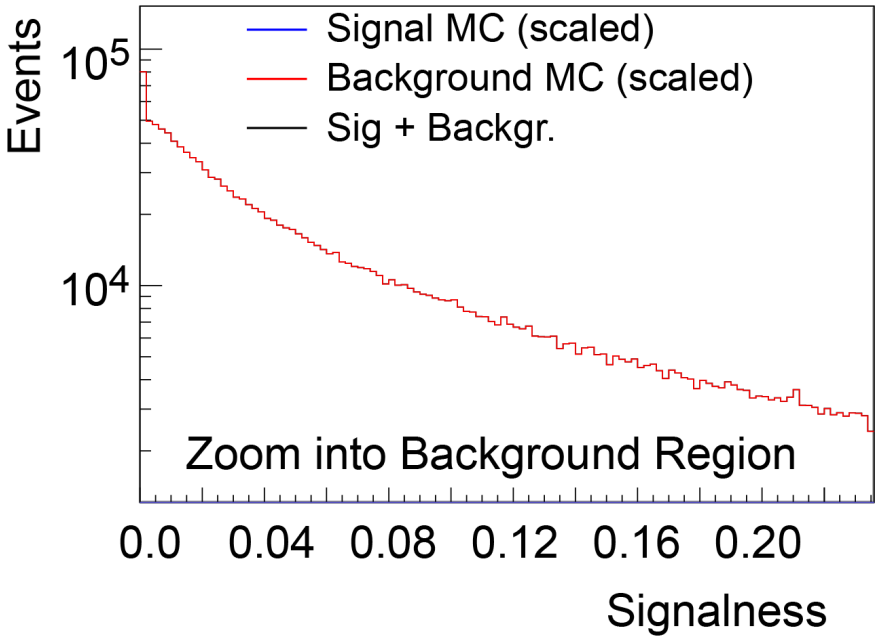
Training and Testing a Random Forest:

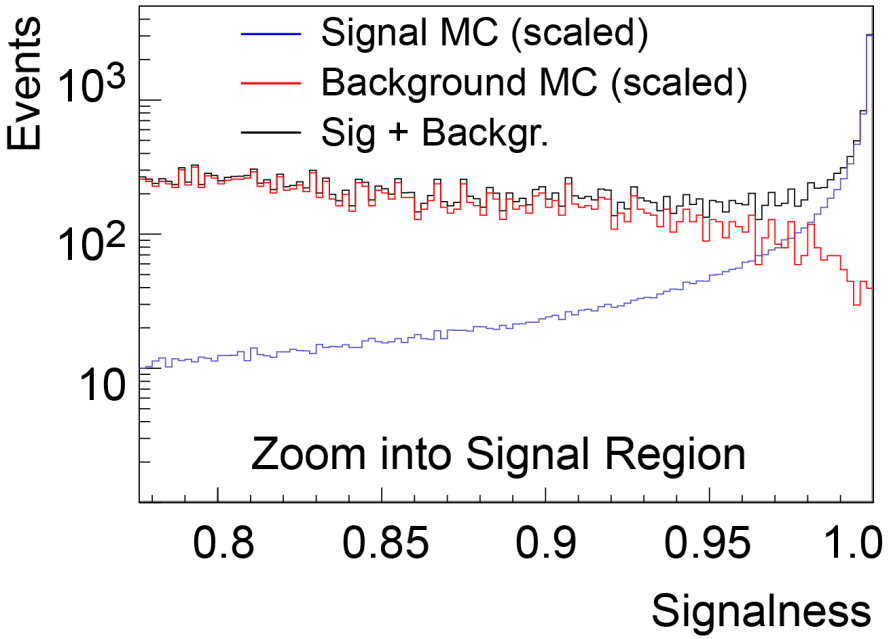
- 500 trees, Weka Random Forest
- 3.4 x 10^5 simulated signal events
- 3.4 x 10^5 simulated background events
- 5-fold cross validation
- 2.8 x 10^4 signal and background events for training

→ Avoid possible overfitting


Events
10^5
10^4
10^3
Background
Signal
0.0
0.2
0.4
0.6
0.8
1.0
Signalness


Events
10^5
10^4
10^3
10^2
Signal MC (scaled)
Background MC (scaled)
Sig + Backgr.
0.0
0.2
0.4
0.6
0.8
1.0
Signalness


Events
10^5
10^4
Signal MC (scaled)
Background MC (scaled)
Sig + Backgr.
Zoom into Background Region
0.0
0.04
0.08
0.12
0.16
0.20
Signalness


Events
10^3
10^2
10
Signal MC (scaled)
Background MC (scaled)
Sig + Backgr.
Zoom into Signal Region
0.8
0.85
0.9
0.95
1.0
Signalness

Forest Performance and Results:

Background computed using a conservative estimate

Cut	Exp. Bckgr. Ev.	Exp. Sig. Ev.	Est. Pur. [%]
0.990	311	5079	94.2
0.992	263	4864	94.9
0.994	215	4606	95.5
0.996	139	4271	96.8
0.998	118	3804	97.0
1.000	77	3017	97.5

Summary and Outlook:

- influence of detailed FS on training of multivariate classifiers has been studied
- MRMR is stable if $n_{Attr.} \geq 20$
- most stable performance of the forest when attributes with $\rho > 0.9$ removed prior to MRMR
- Random Forest trained and tested
- > 3000 events are found, Purity routinely > 95 %
- Not fully optimized
- hope to achieve even better with a fully optimized classifier

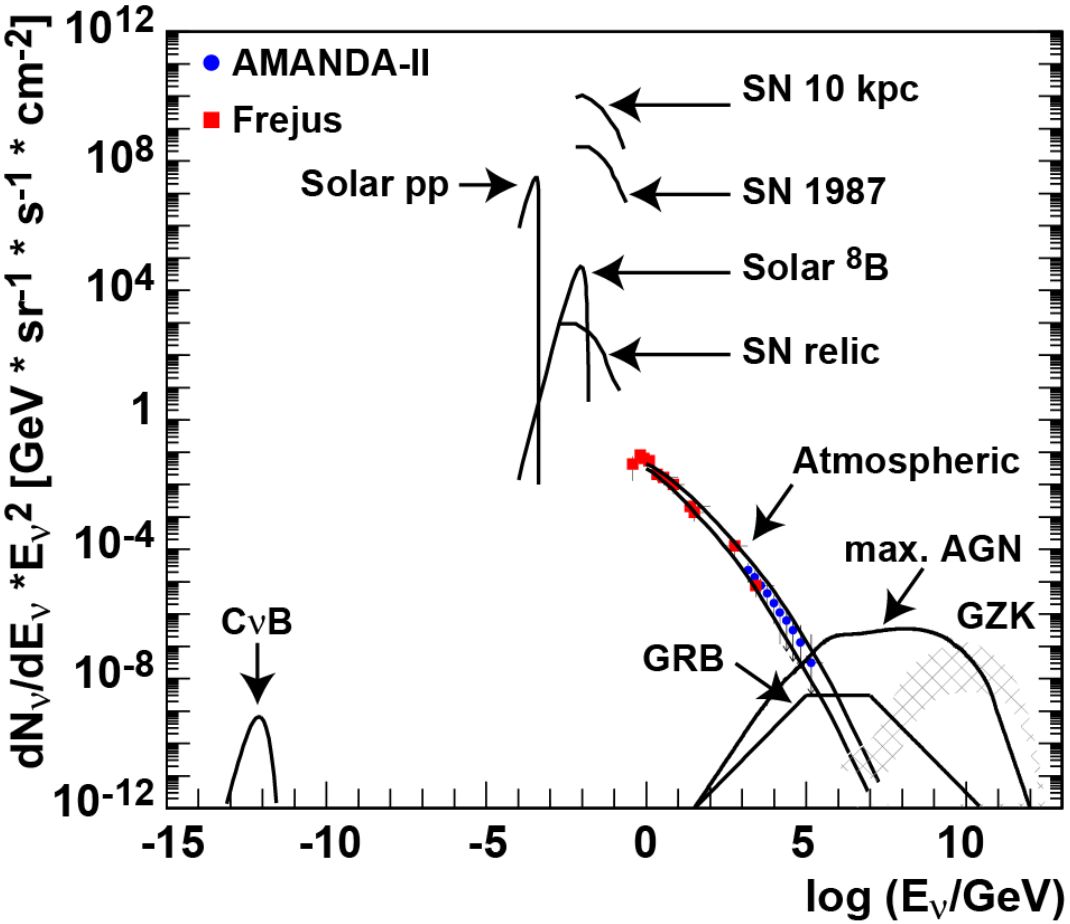
Backup Slides

The Random Forest:

- Developed by Leo Breiman (2001)
- Utilizes ensemble of decision trees

  → Forest
- No boosting between individual trees
- Average used for final classification
- No overtraining! (see: Machine Learning, 45, 5 – 32 (2001))
- **Here:** → **Rapidminer toolkit** (developed in Dortmund)

Atmospheric neutrinos:



A detailed understanding of the atmospheric neutrino spectrum is crucial for the detection of an astrophysical flux.