Data Mining Ice Cubes

Tim Ruhe, Katharina Morik

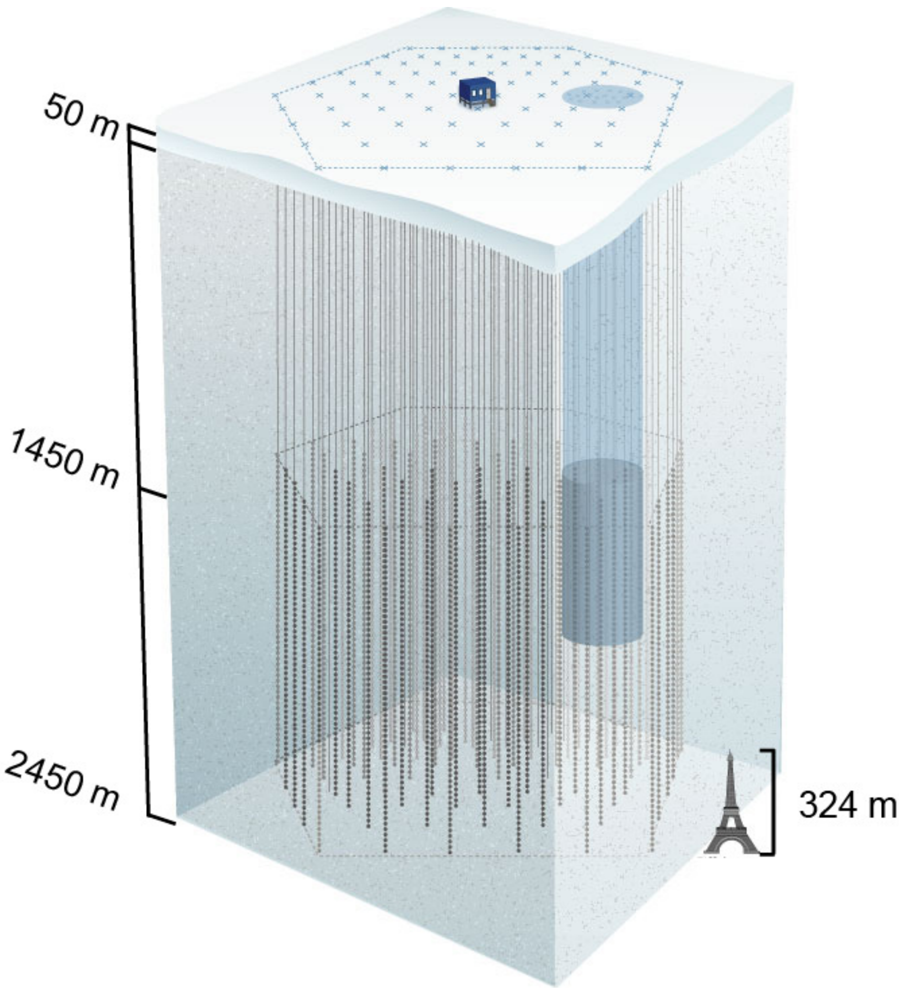
ADASS XXI, Paris 2011

IceCube

UNIVERSITÄT DORTMUND

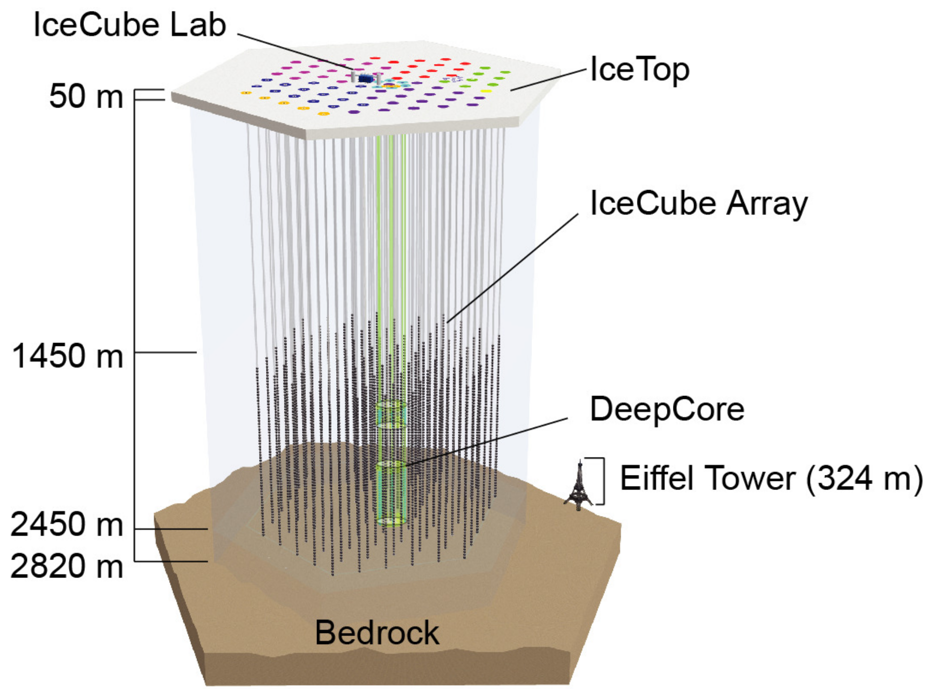
Outline:

- IceCube
- RapidMiner
- Feature Selection
- Random Forest training and application
- Summary and outlook



The IceCube detector:

- Completed in December 2010
- Located at the geographic South Pole
- 5160 Digital Optical Modules on 86 strings
- Instrumented volume of 1 km^3
- Has taken data in various string configurations (this work: 59 strings)



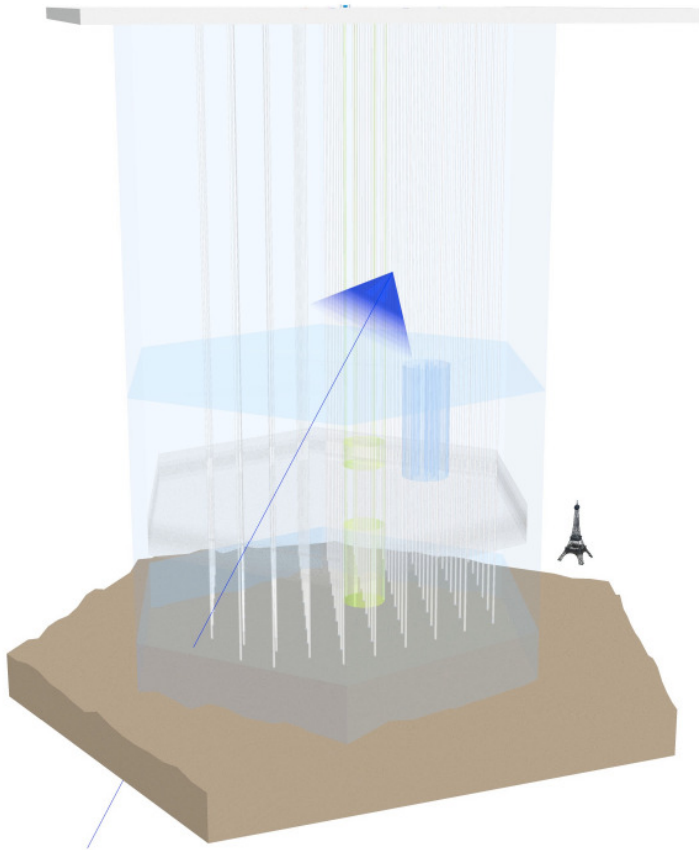
The IceCube detector:

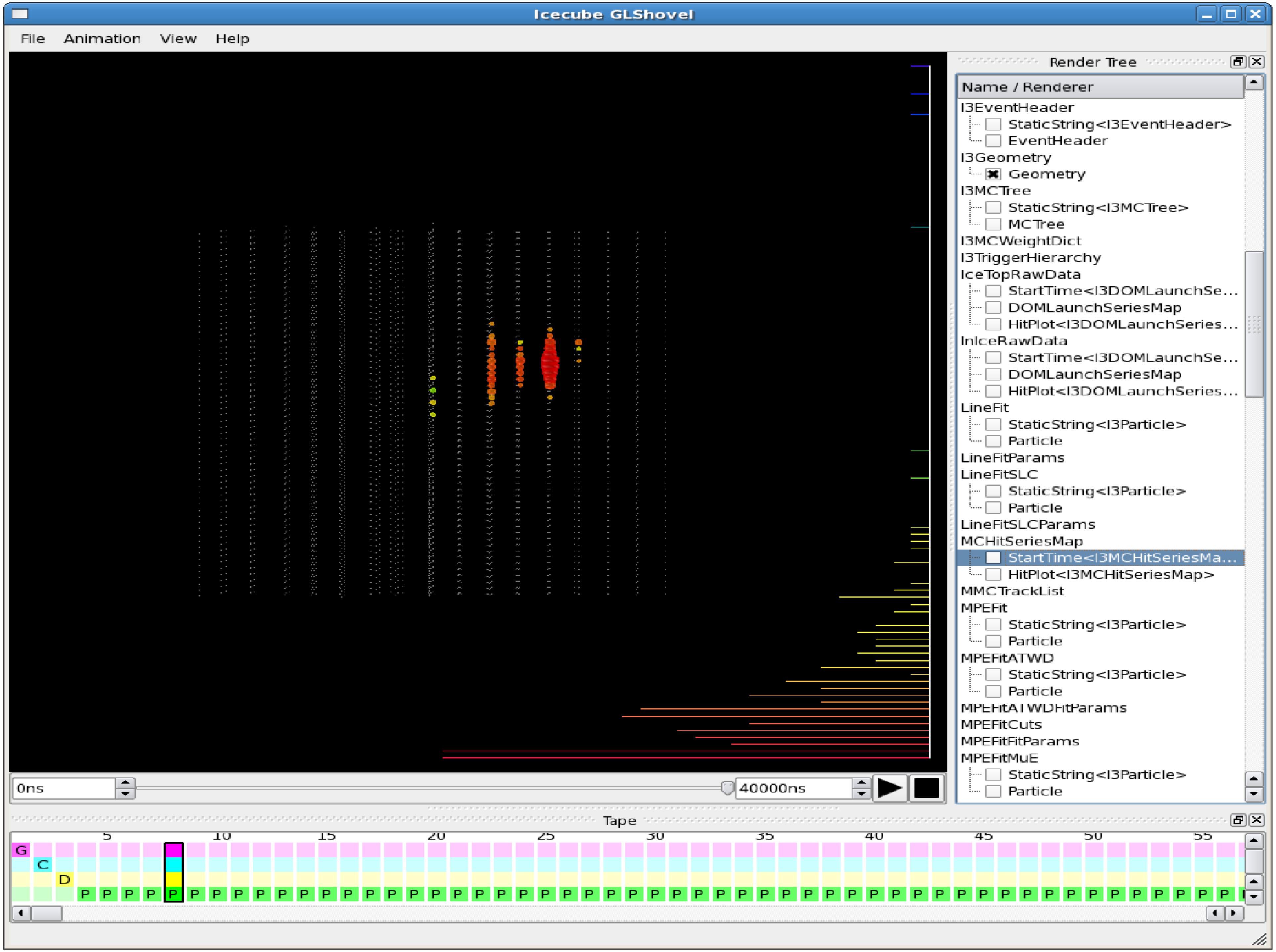
- Detection principle: Cherenkov light
- Look for events of the form:

$\nu + X \rightarrow e, \mu, \tau$

- Dominant background of atm. μ

→ Use earth as a filter (select upgoing events only)




Icecube GLShovel
File
Animation
View
Help
Render Tree
Name / Renderer
I3EventHeader
StaticString<I3EventHeader>
EventHeader
I3Geometry
Geometry
I3MCTree
StaticString<I3MCTree>
MCTree
I3MCWeightDict
I3TriggerHierarchy
IceTopRawData
StartTime<I3DOMLaunchSe...
DOMLaunchSeriesMap
HitPlot<I3DOMLaunchSeries...
InIceRawData
StartTime<I3DOMLaunchSe...
DOMLaunchSeriesMap
HitPlot<I3DOMLaunchSeries...
LineFit
StaticString<I3Particle>
Particle
LineFitParams
LineFitSLC
StaticString<I3Particle>
Particle
LineFitSLCParams
MCHitSeriesMap
StartTime<I3MCHitSeriesMa...
HitPlot<I3MCHitSeriesMap>
MMCTrackList
MPEFit
StaticString<I3Particle>
Particle
MPEFitATWD
StaticString<I3Particle>
Particle
MPEFitATWDFitParams
MPEFitCuts
MPEFitFitParams
MPEFitMuE
StaticString<I3Particle>
Particle
0ns
40000ns
Tape
5
10
15
20
25
30
35
40
45
50
55
G
C
D
P

Data Mining in IceCube:

- App. 2600 reconstructed attributes
- Data and MC do not necessarily agree
- Signal/background ratio ~ 10^{-3}

→Interesting for studies within the scope of machine learning

RapidMiner:

- Data Mining environment, Open Source, Java
- Developed at the Department of Computer Science
  at TU Dortmund (group of K. Morik)
- Operator based
- Quite intuitive to handle (personal opinion)

Preselection of parameters: (After application of precuts)

1. Check for consistency (data vs. nu MC vs. background MC)

→ 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, DelAng, GalLong, Time...)
(reduction to 612 attributes)

4. Eliminate highly correlated (ρ = 1.0) and constant parameters

→ Final set of 477 parameters

Mininmum Redundancy Maximum Relevance (MRMR):

- Iteratively add features with biggest relevance and least redundancy
- Quality criterion Q:

$$Q = R(x, y) - \frac{1}{j} \sum_{x' \, in \, F_j} D(x', x)$$

R: Relevance; D: Redundancy; F_j = already selected features

Stability of the MRMR Selection:

Jaccard Index:

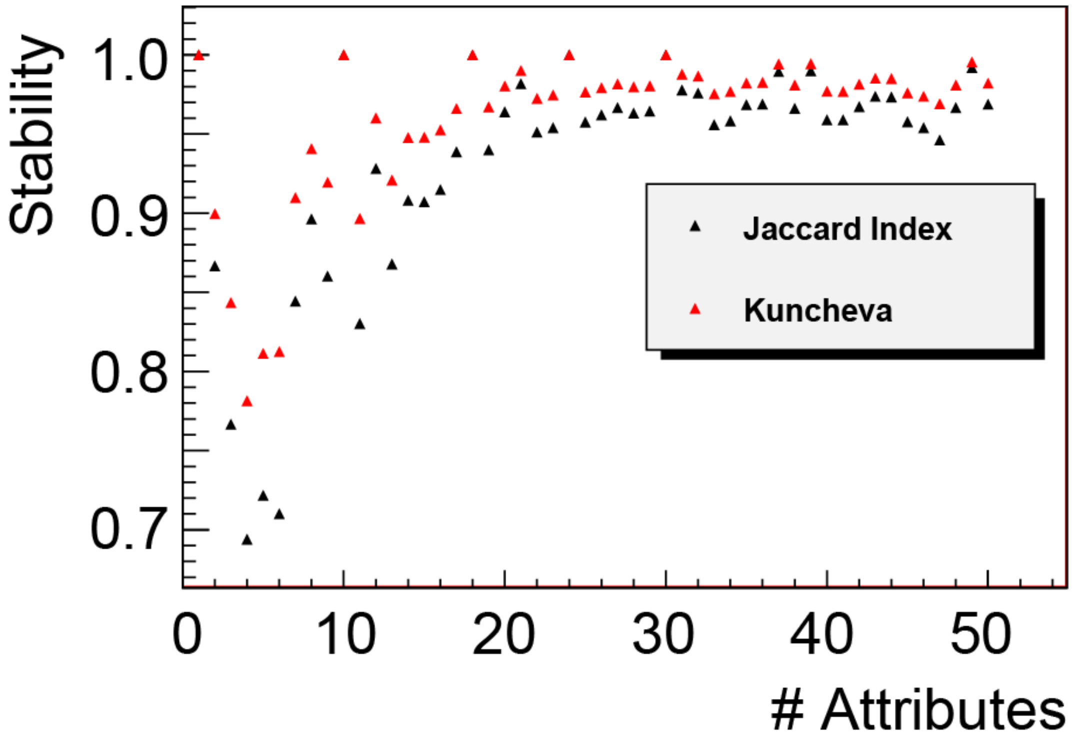
$$J = \frac{|A \cap B|}{|A \cup B|}$$

Kuncheva's Index:

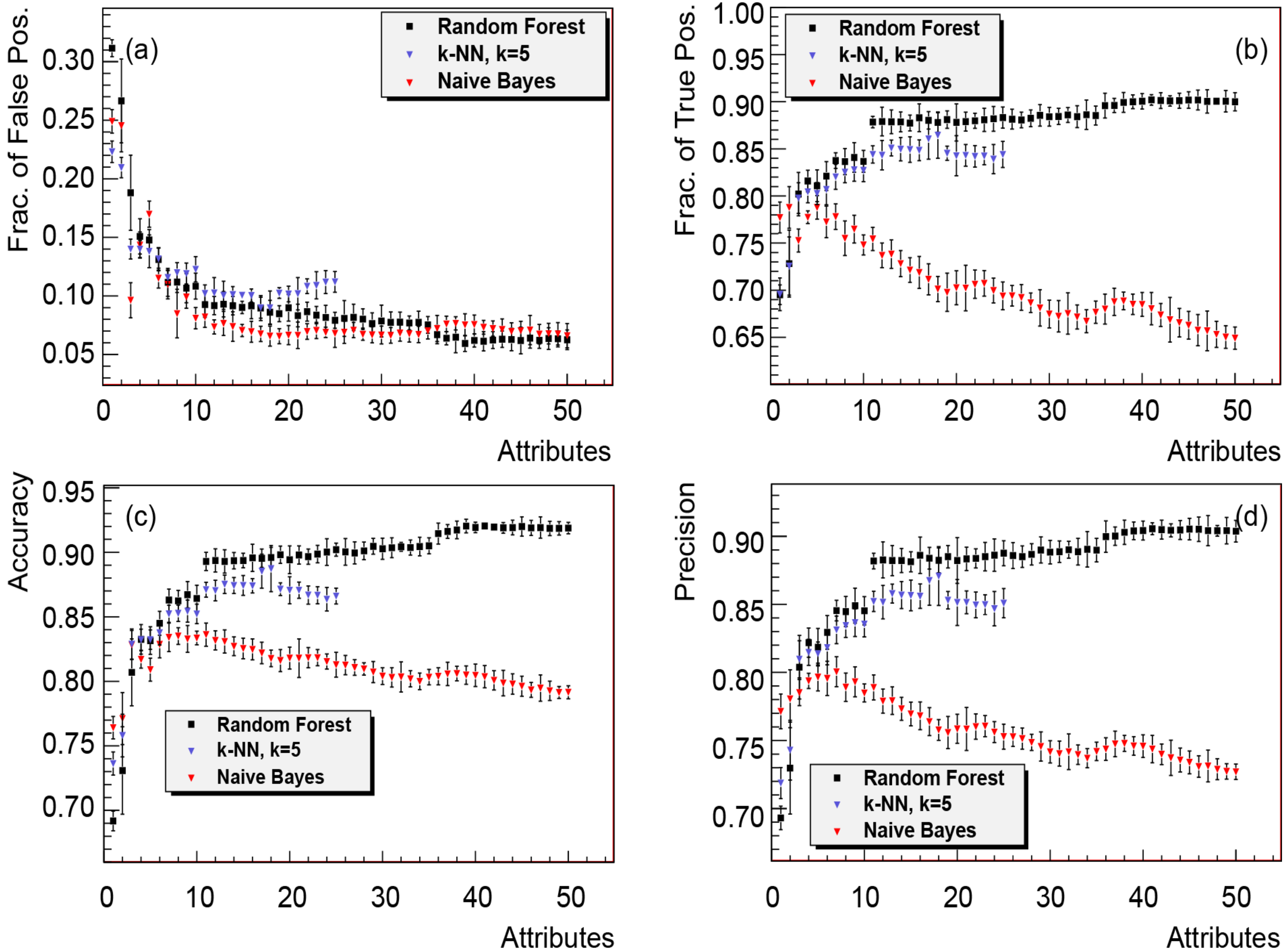
$$I_C(A,B) = \frac{rn - k^2}{k(n-k)}$$

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

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

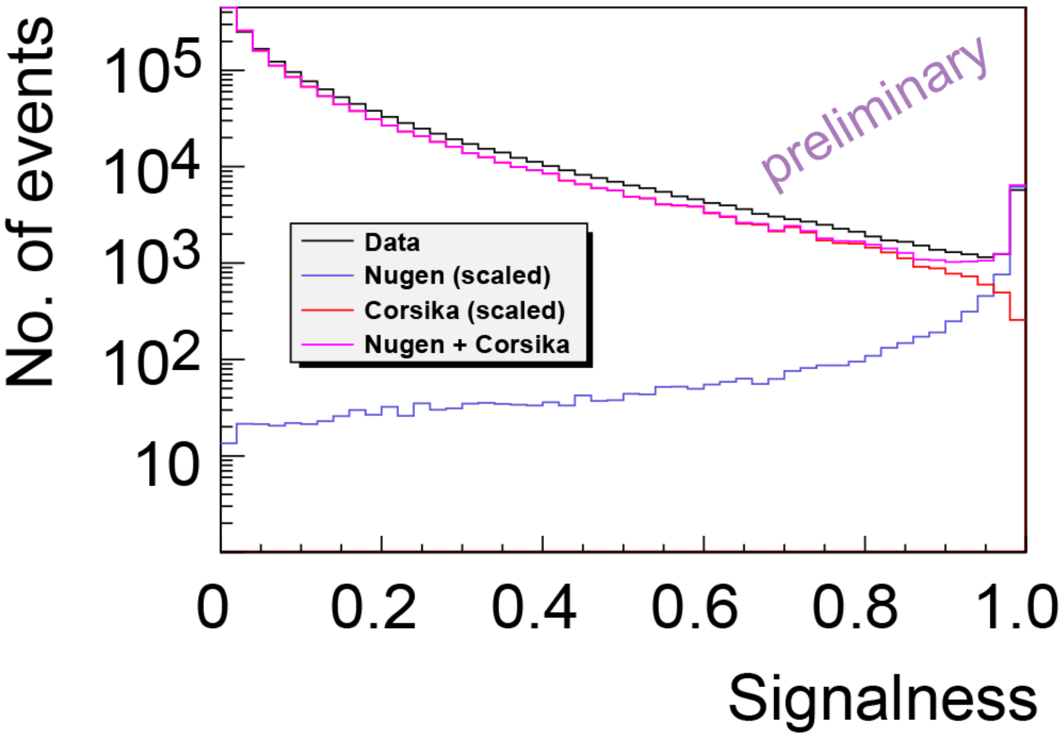


http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.6458&rep=rep1&type=pdf



Frac. of False Pos.
(a)
Random Forest
k-NN, k=5
Naive Bayes
0.30
0.25
0.20
0.15
0.10
0.05
0
10
20
30
40
50
Attributes
Frac. of True Pos.
(b)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Accuracy
(c)
0.95
0.90
0.85
0.80
0.75
0.70
Precision
(d)

Random Forest output:



Forest parameters:

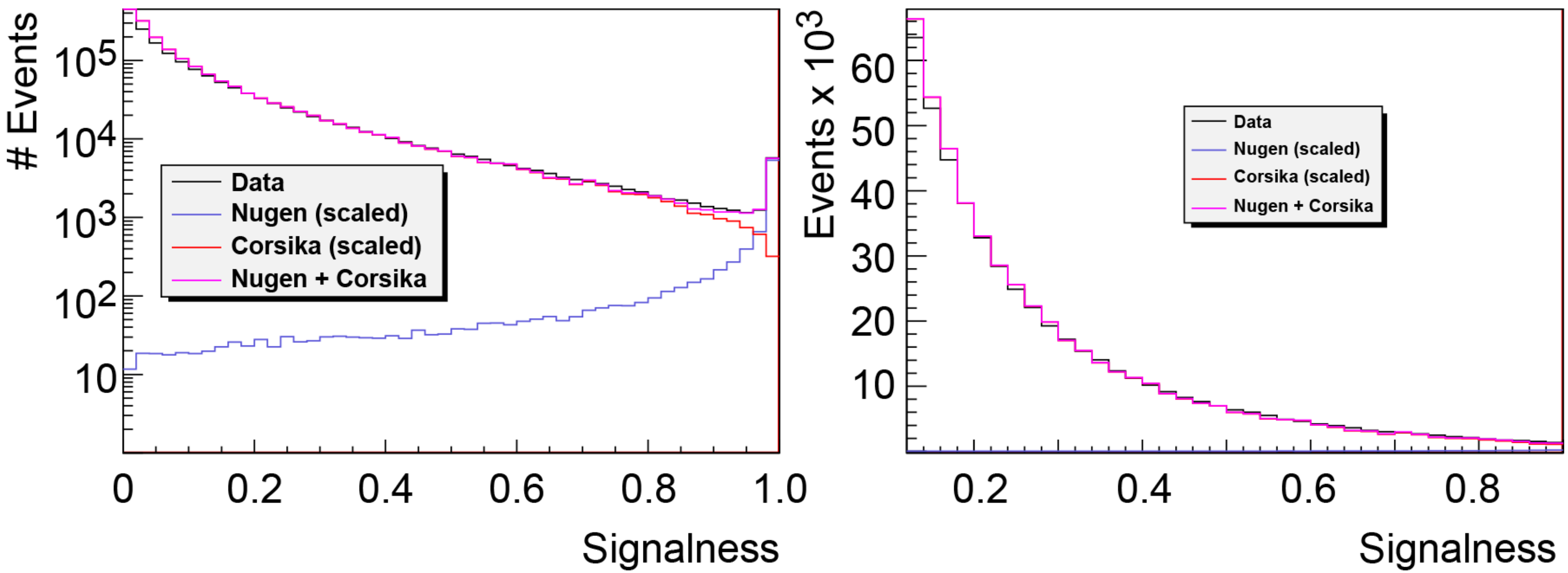
- 500 trees
- 3.8×10^5 backgr. events
- 7.0×10^4 signal events
- 5 fold X-Validation
- 28×10^4 of each class used for training

Data/MC mismatch → underestimation of background

Change the Scaling of the Background:

→ such that it matches data for Signalness > 0.2

preliminary



Expected Numbers: With Rescaled Background

preliminary

Cut	Nugen	Corsika	Sum	Data
0.990	4817 ± 44	114 ± 47	4931 ± 64	4988
0.992	4633 ± 43	98 ± 37	4731 ± 57	4757
0.994	4414 ± 41	71 ± 37	4485 ± 55	4476
0.996	4122 ± 32	60 ± 32	4182 ± 45	4134
0.998	3695 ± 44	22 ± 20	3717 ± 50	3638
1.000	2932 ± 33	5 ± 11	2937 ± 35	2833

Summary and Outlook:

- IceCube is well suited for a detailed study within machine learning
- Random Forest outperforms simpler classifiers
- Feature Selection shows stable performance
- Application on data matches MC expectations
- Increase in performance expected for full optimization

Backup Slides