Data Science in Astroparticle Physics

--Project C3 of the Collaborative Research Center 876 --

Tim Ruhe

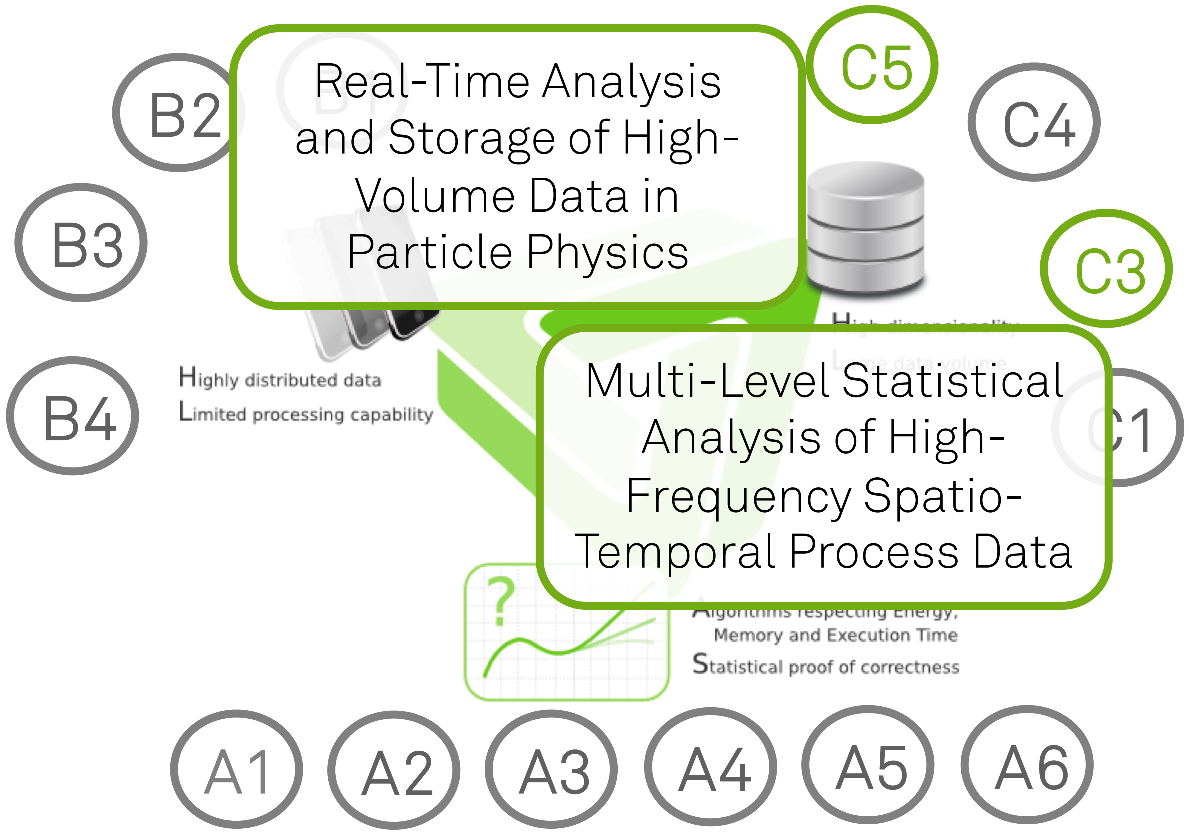
HAP Workshop Aachen 2019

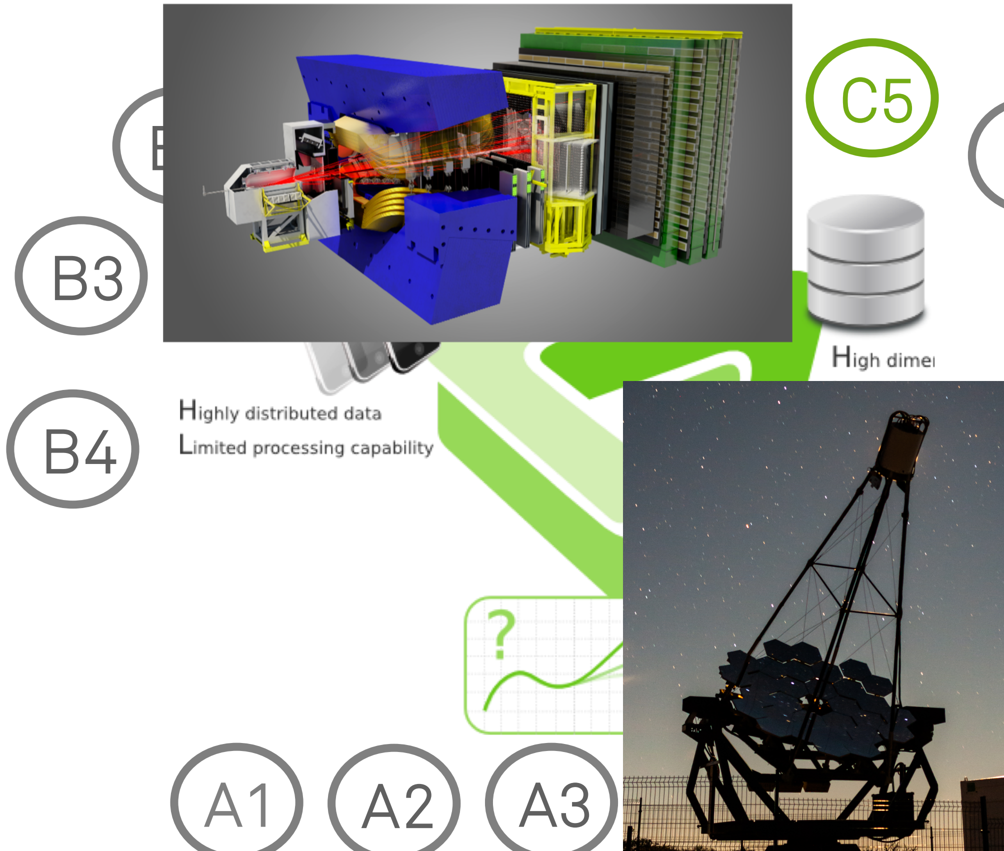
tim.ruhe@tu-dortmund.de

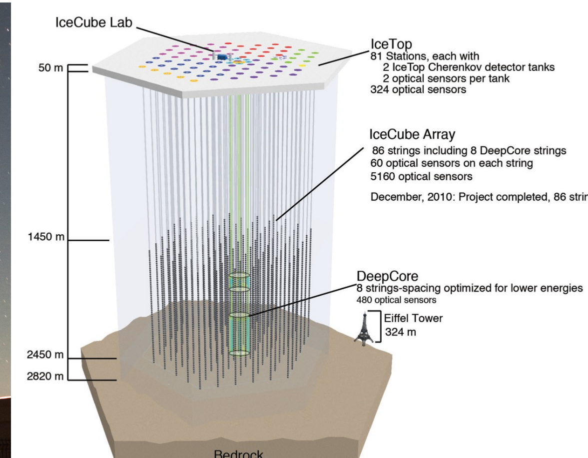
Speaker: K. Morik

Real-Time Analysis and Storage of High-Volume Data in Particle Physics

Multi-Level Statistical Analysis of High-Frequency Spatio-Temporal Process Data



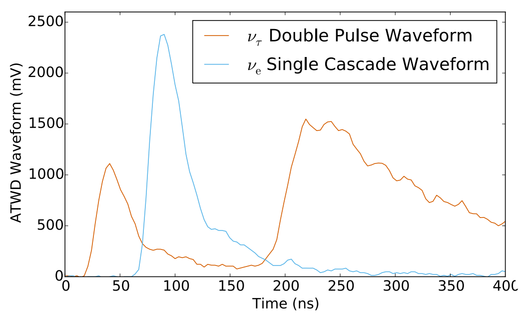

C5
C4
B3
C3
B4
Highly distributed data
Limited processing capability
A1
A2
A3

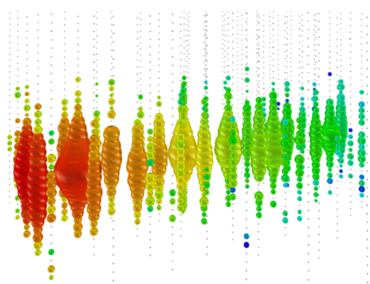

IceCube Lab
IceTop
81 Stations, each with
2 IceTop Cherenkov detector tanks
2 optical sensors per tank
324 optical sensors
50 m
IceCube Array
86 strings including 8 DeepCore strings
60 optical sensors on each string
5160 optical sensors
December, 2010: Project completed, 86 strings
1450 m
DeepCore
8 strings-spacing optimized for lower energies
480 optical sensors
Eiffel Tower
324 m
2450 m
2820 m
Bedrock

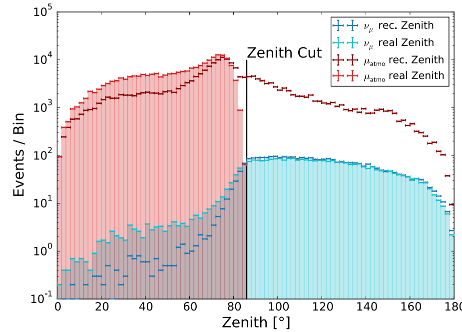
Data Analysis in APP

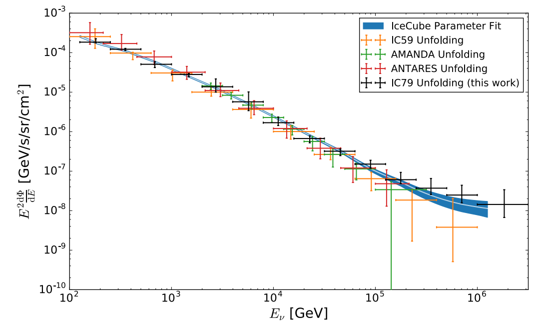
IceCube

DAQ → Reconstruction → Background Rejection → Unfolding

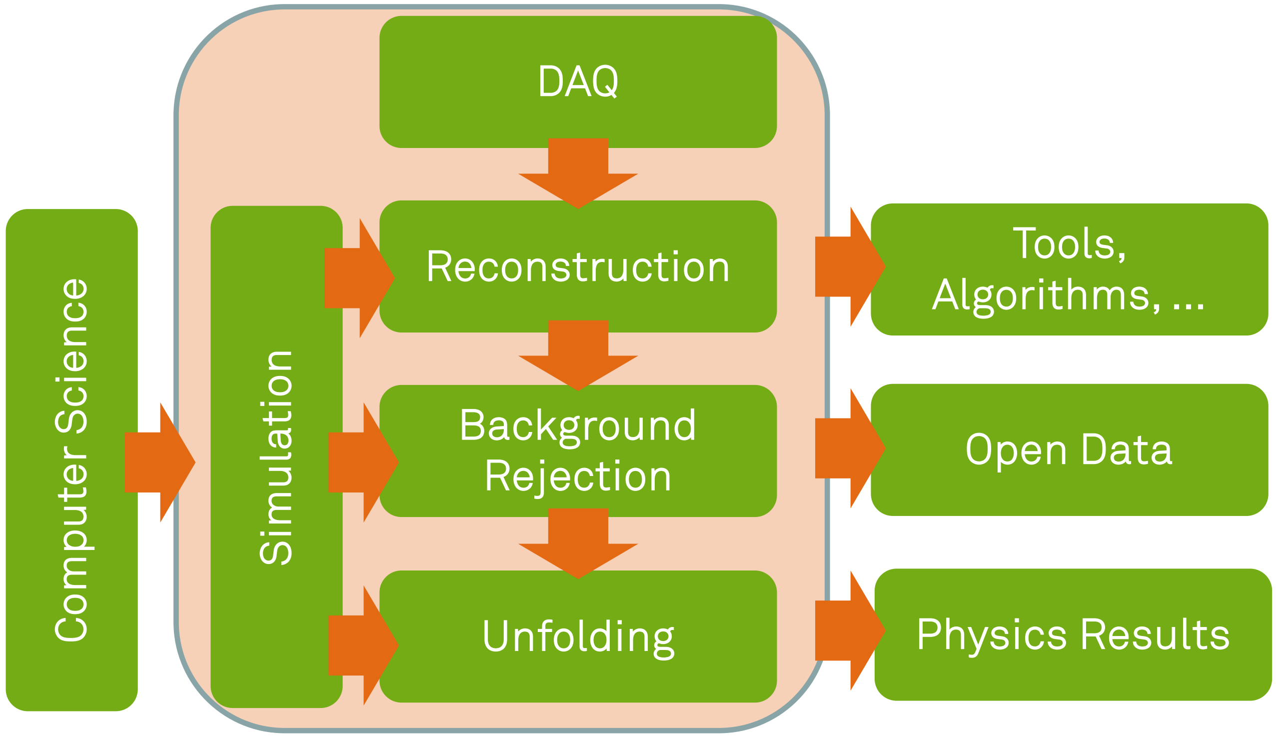




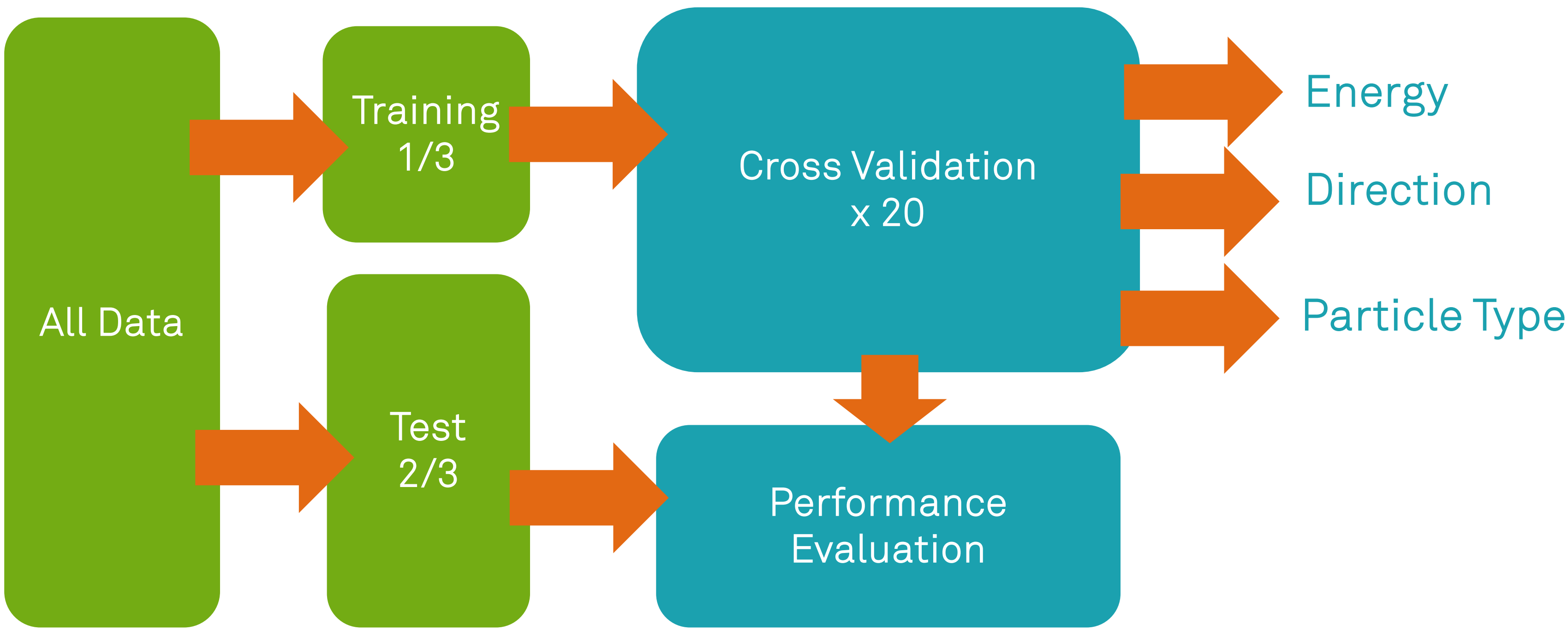




APP and Computer Science



Event Reconstruction with Machine Learning



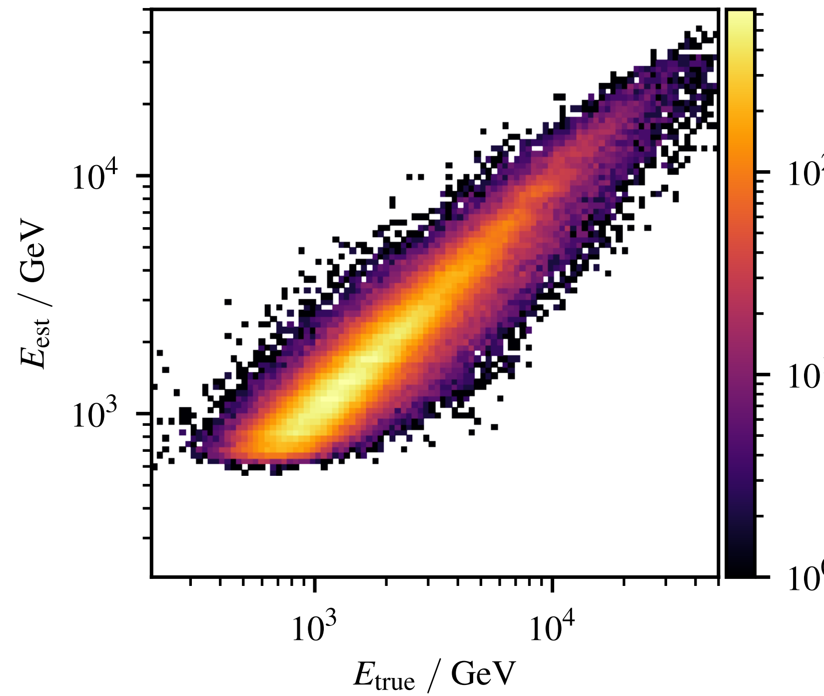
M. Nöthe, Kai Brügge

Energy and Particle Type

M. Nöthe, Kai Brügge

- Energy:
  - Random Forest Regressor (200 trees, max. depth 15)
  - Cross validdated r^2-score: 0.785 +/- 0.017

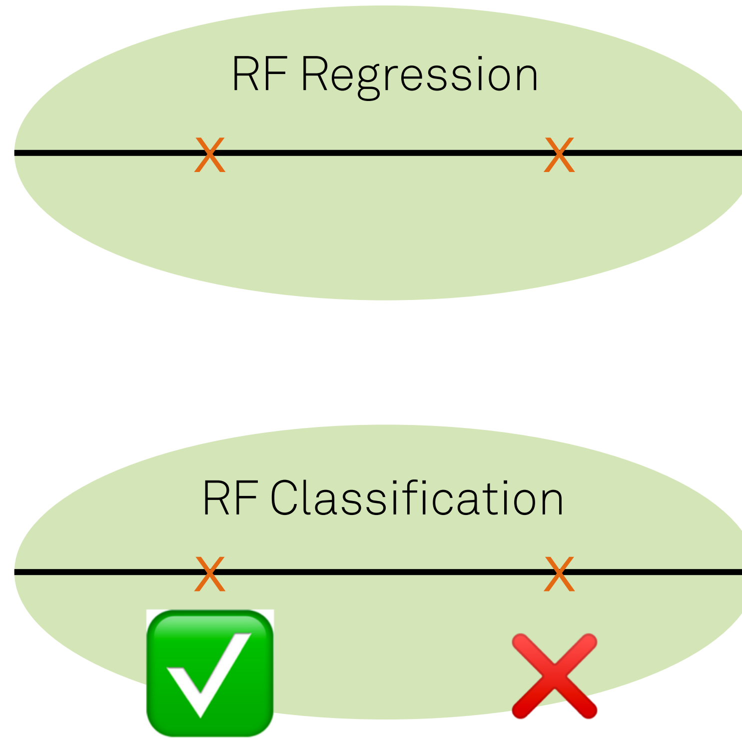
- Particle type:
  - Random Forest Classifier (200 trees, max. depth 15)
  - AUC: 0.827 +/- 0.003



Directional Reconstruction: Disp Method

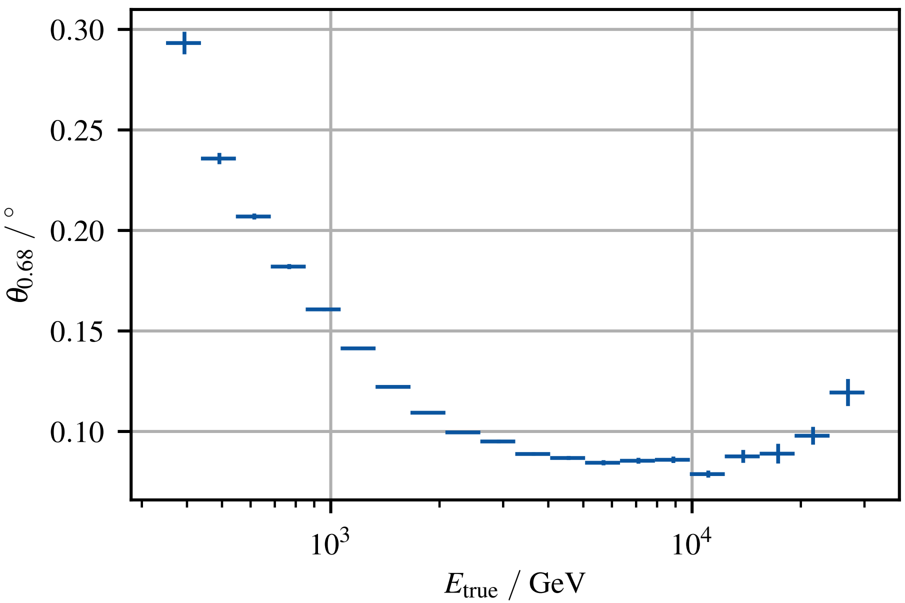
M. Nöthe, Kai Brügge

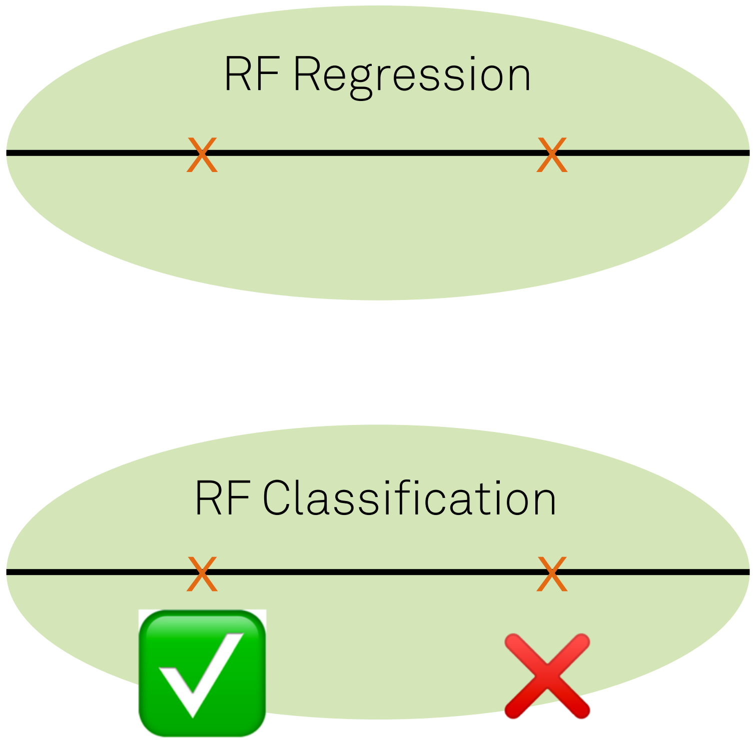
- Simplifies 2D regression to 1D regression plus binary classification
- True source position is somewhere on main shower axis
- Use Random Forest Regressor to estimate distance to cog of light distribution (2 solutions)
- Use Random Forest Classifier to pick correct solution



Directional Reconstruction: Disp Method

M. Nöthe, Kai Brügge

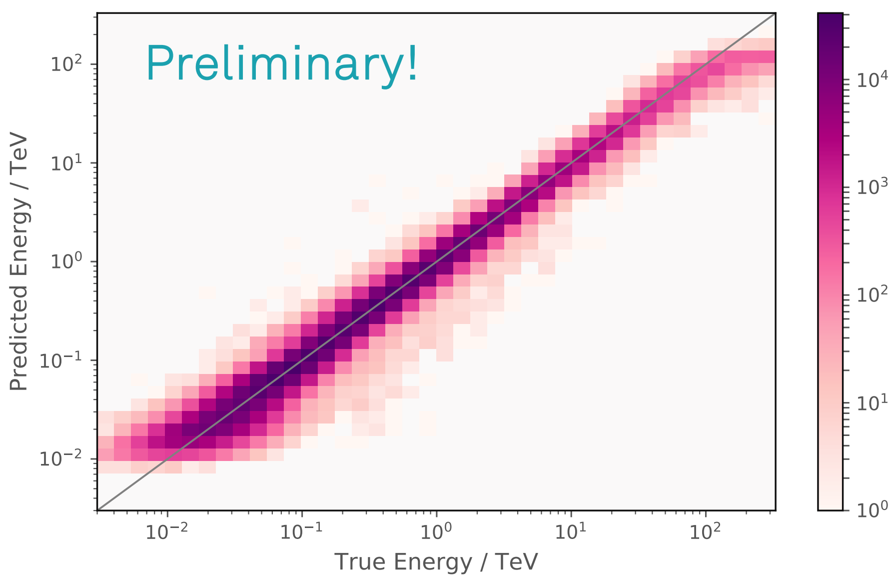


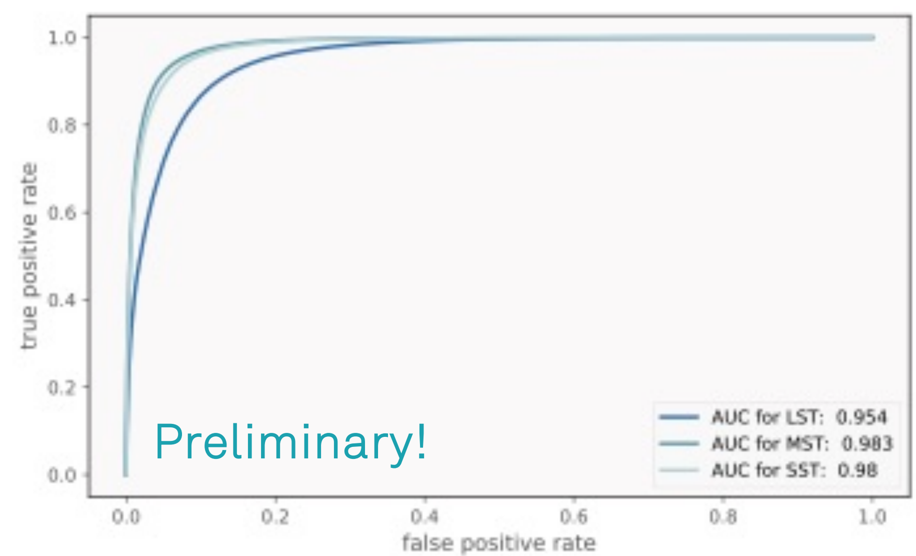


First Results for CTA

Kai Brügge

simtel ctapipe HDF5 aicttools FITS gammapy





Classification: Approach and Challenges

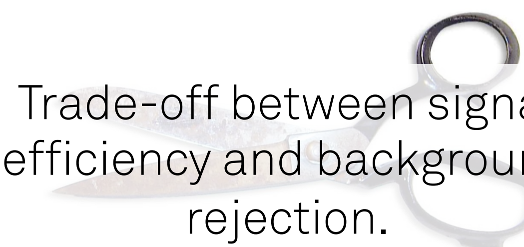
Variable Selection → Classifier Training → Cut on Classifier Output

Classification: Approach and Challenges

Variable Selection → Classifier Training → Cut on Classifier Output

~1200 Features

Signal to Background Ratio: 10^{-3}



Trade-off between signal efficiency and background rejection.

Feature Selection: MRMR

Initially	1219
blacklisted	1129
constant & useless	855
Correlation cut	323
Data/MC Clf	311
mRMR	60

M. Börner, PhD thesis (2018)

- Select features according to relevance and redundancy
- Feature set is built by iteratively adding features that fulfill the following criterion

$$\max_{x_j \,\epsilon\, X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \,\epsilon\, S_{m-1}} I(x_i, x_j) \right]$$

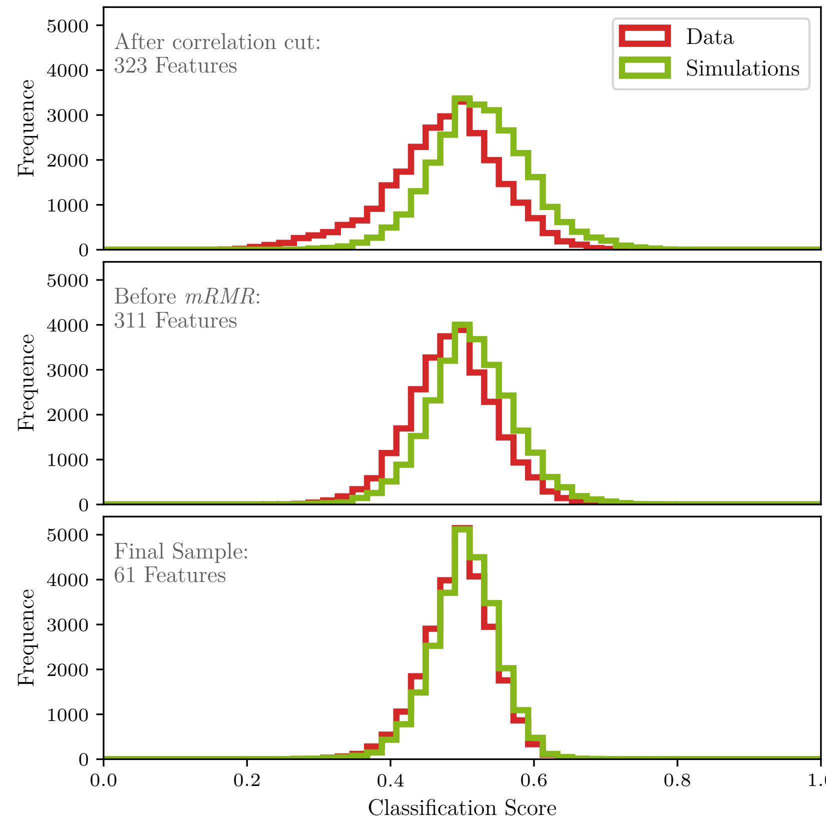
Peng, H.C., Long, F., and Ding, C., IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238, 2005.

Ding, C., & Peng, H., *Journal of bioinformatics and computational biology*, *3*(02), 185-205. (2005)

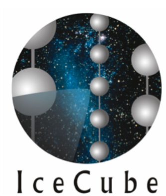
Excluding Data-MC Mismatches

- Train classifier to distinguish data and Monte Carlo
- Inspect feature importance
- Dismiss important feature

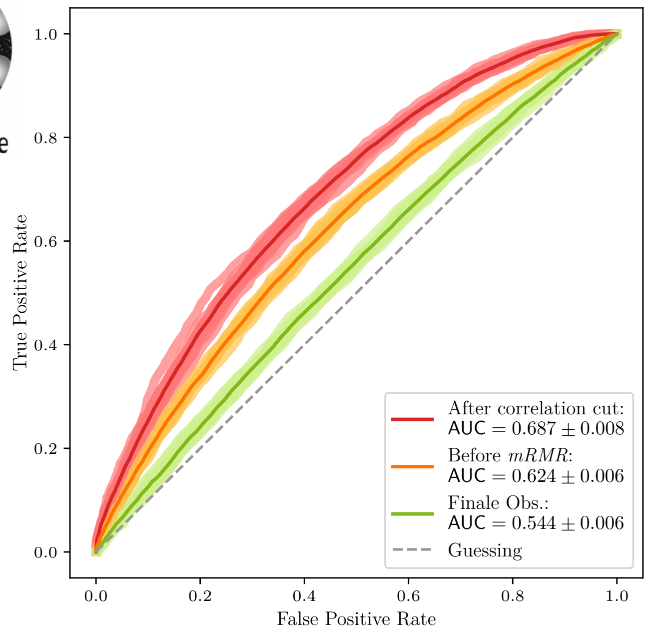
- Useful for feature selection and verification

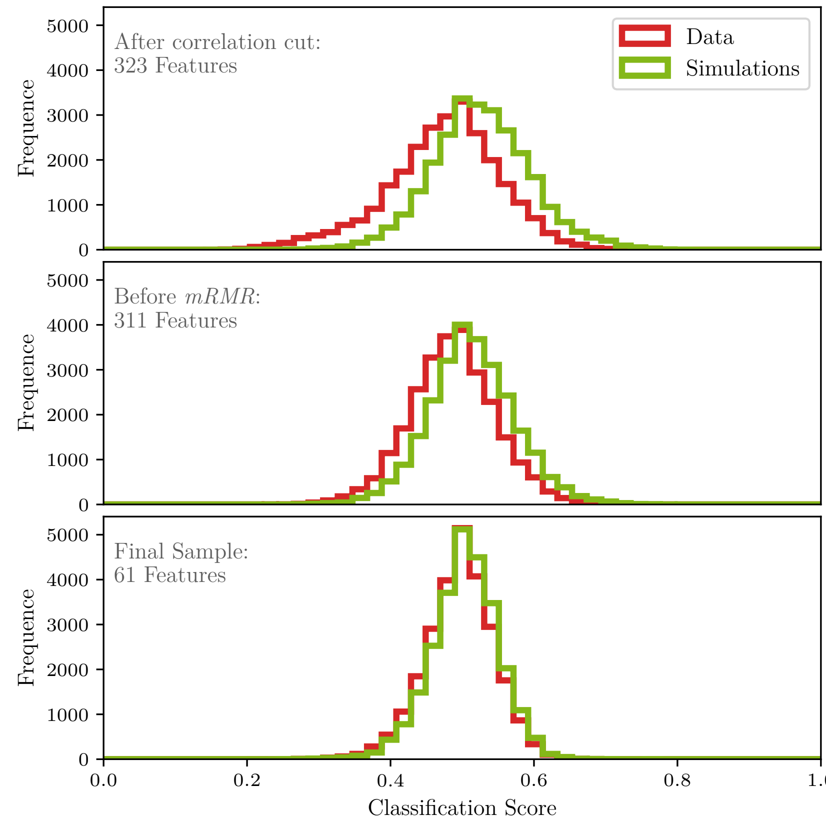


M. Börner, PhD thesis (2018)



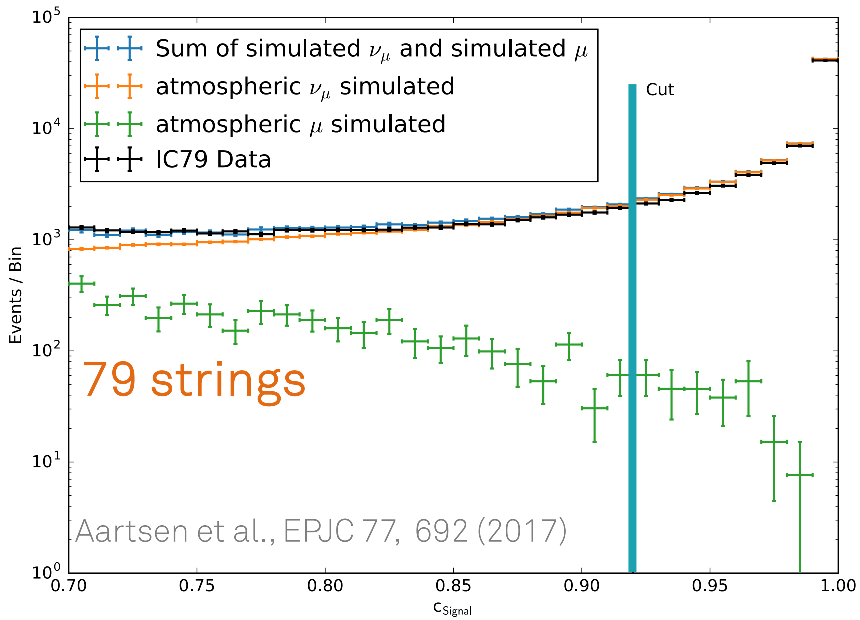
Excluding Data-MC Mismatches



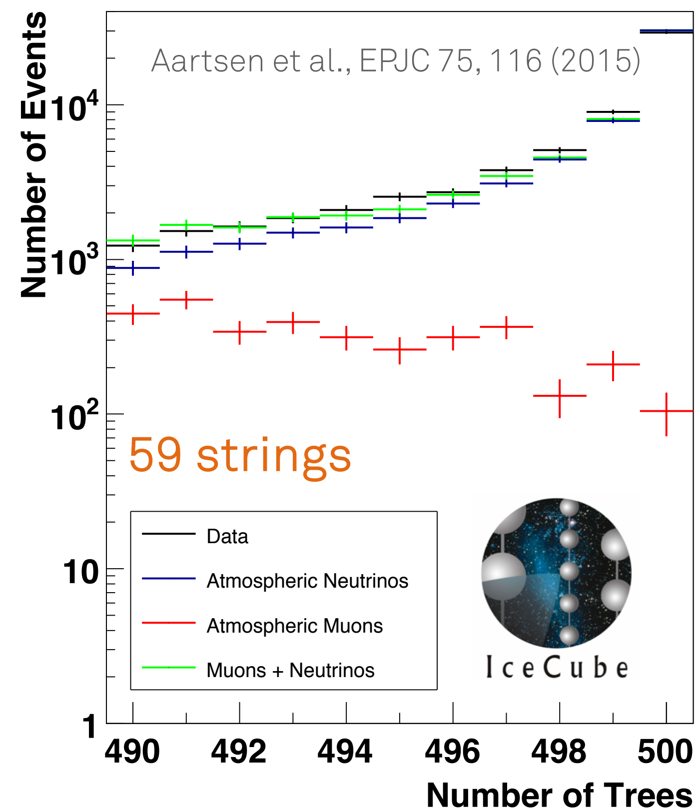


M. Börner, PhD thesis (2018)

Classifier Training and Output

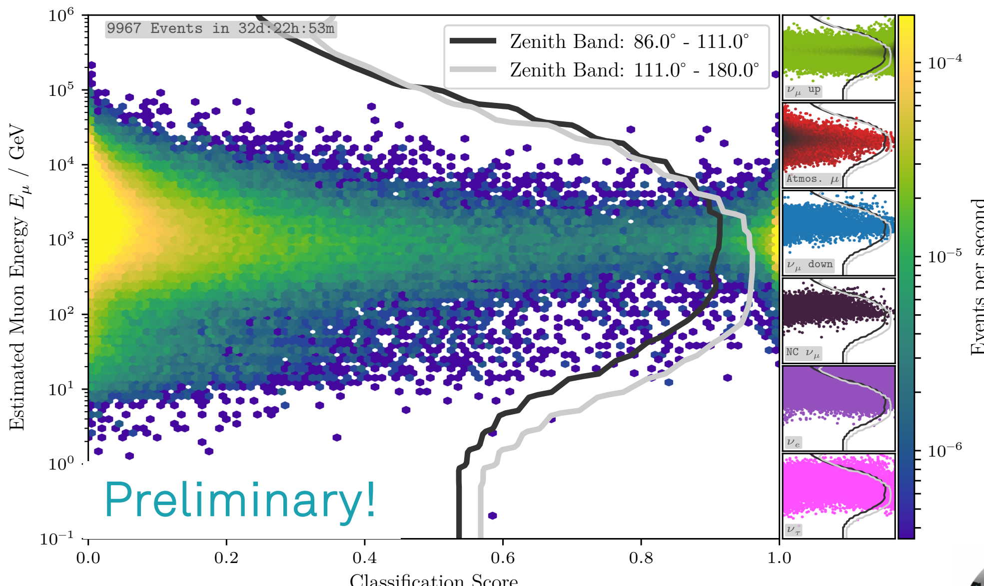


~ 200 neutrino candidates per day

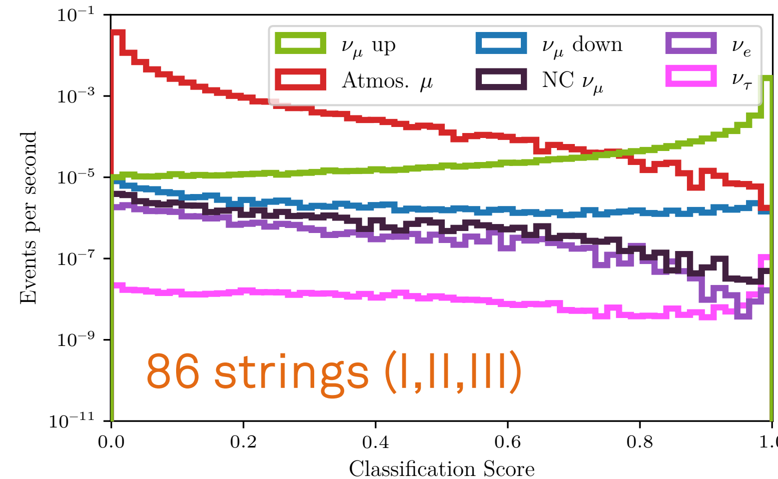


~ 80 neutrino candidates per day

Classifier Training and Output



~ 300 neutrino candidates per day



Classifier output is energy and zenith dependent.

Score cut as a function of energy and zenith.

M. Börner, PhD thesis (2018)

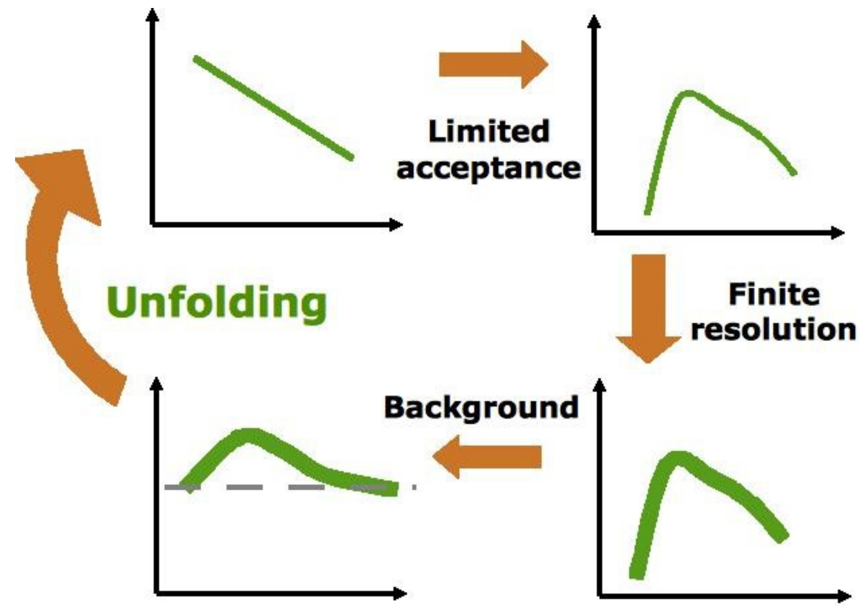
Why Unfold?

The production of muons from muon neutrinos is a stochastic process:

$$\frac{dN_\mu}{dE_\mu} = \int_{E_\mu}^{\infty} dE_\nu \left(\frac{dN_\nu}{dE_\nu}\right)\left(\frac{dP(E_\nu)}{dE_\mu}\right)$$

Neutrino energy spectrum

Physics of neutrino interaction



Why Unfold?

$$\frac{dN_\mu}{dE_\mu} = \int_{E_\mu}^{\infty} dE_\nu \left(\frac{dN_\nu}{dE_\nu}\right)\left(\frac{dP(E_\nu)}{dE_\mu}\right)$$

Fredholm integral equation of the first kind

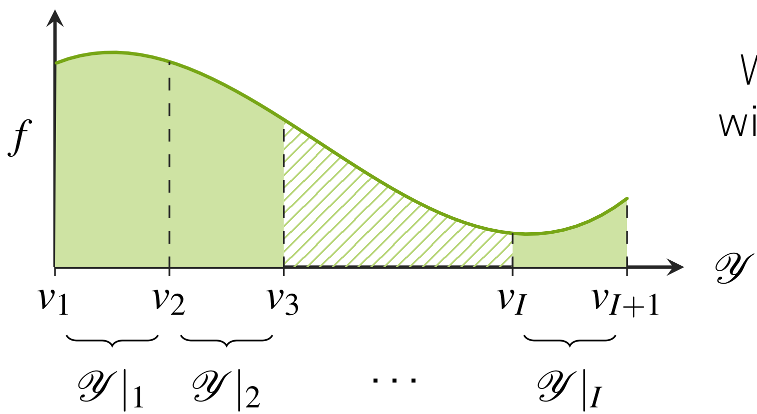
$$g(y) = \int_{E_{min}}^{E_{max}} A(E,y) f(E) dE$$

$A(E,y)$ also includes muon propagation and additional smearing introduced by the detector itself

$$\vec{g}(y) = \underline{A}(E,y)\vec{f}(x)$$

Generally solved as a matrix equation, matrix $\underline{A}(E,y)$ obtained from simulation

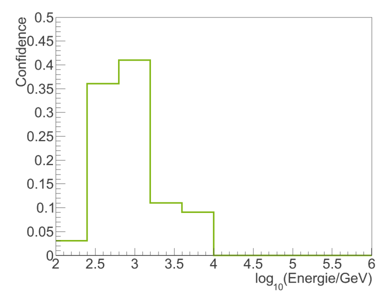
The Dortmund Spectrum Estimation Algorithm (DSEA)

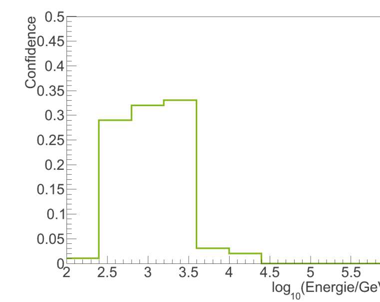


We are generally happy with a discretized version of the result.

Interpret every bin as a class of event and solve with classifier

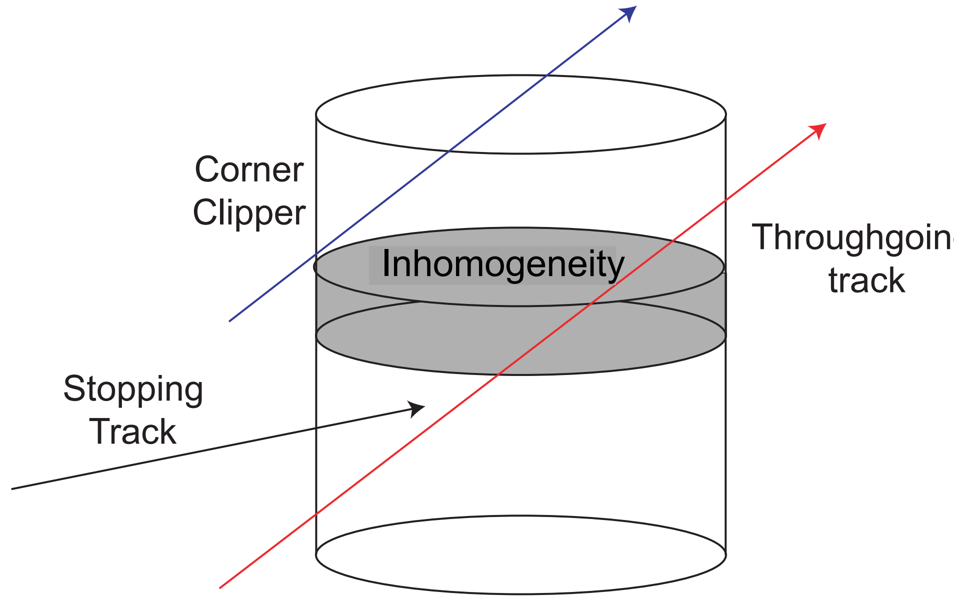
Interpret classifier output as pdf, obtain estimator for by summation over confidence distributions

$\hat{f} =$ 

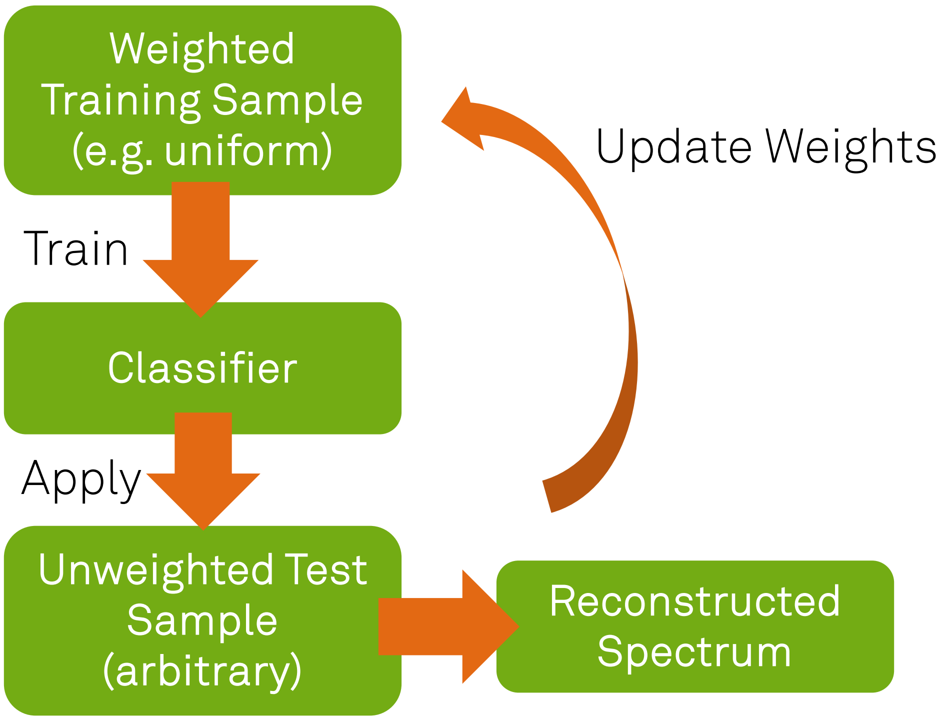
$+ \cdots +$ 

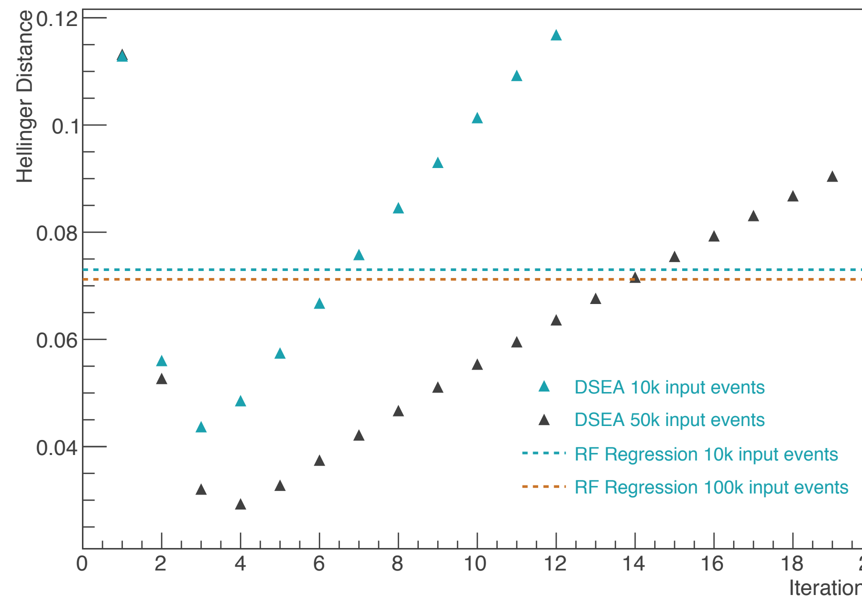
Motivation for DSEA

- Muons of the exact same energy, may create very different patterns, depending on their geometry
- Geometric information might increase accuracy
- Many existing algorithms are limited w.r.t. the number of input variables
- Spectra are returned accurately, but information on individual events is lost



DSEA: Iterative Update of Weights

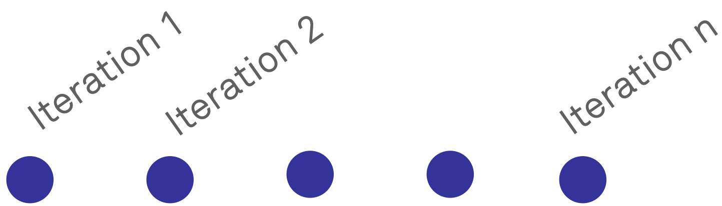




Ruhe et al., Proc. of ADASS XXVI (2016)

DSEA+: Adaptive Step-Width

Fixed Stepsize (DSEA)



Adaptive Stepsize (DSEA+)

$$p^{(k)} = f^{(k)} - f^{(k-1)}$$

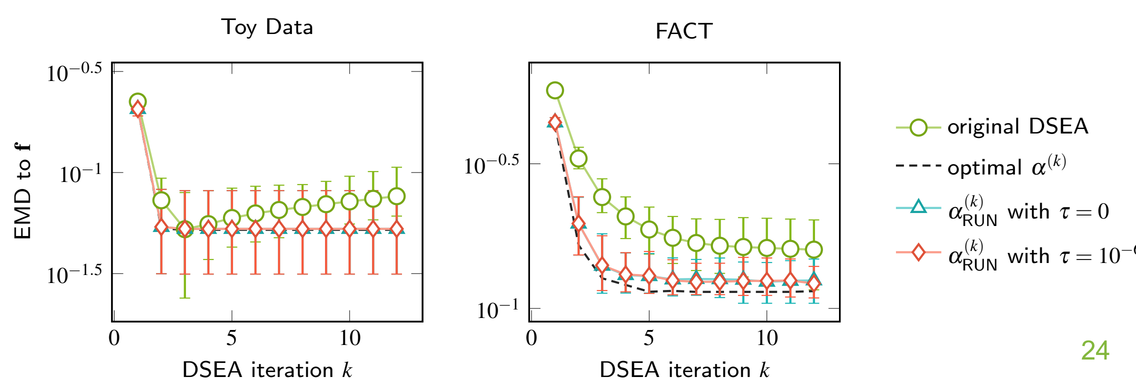
$$f^{(k)} = f^{(k-1)} + \alpha^{(k)} p^{(k)}$$

$$\alpha^{(k)} = \alpha^{\eta-1} \quad \text{(multiplicative decay)}$$

$$\alpha^{(k)} = \eta^{k-1} \quad \text{(exponential decay)}$$

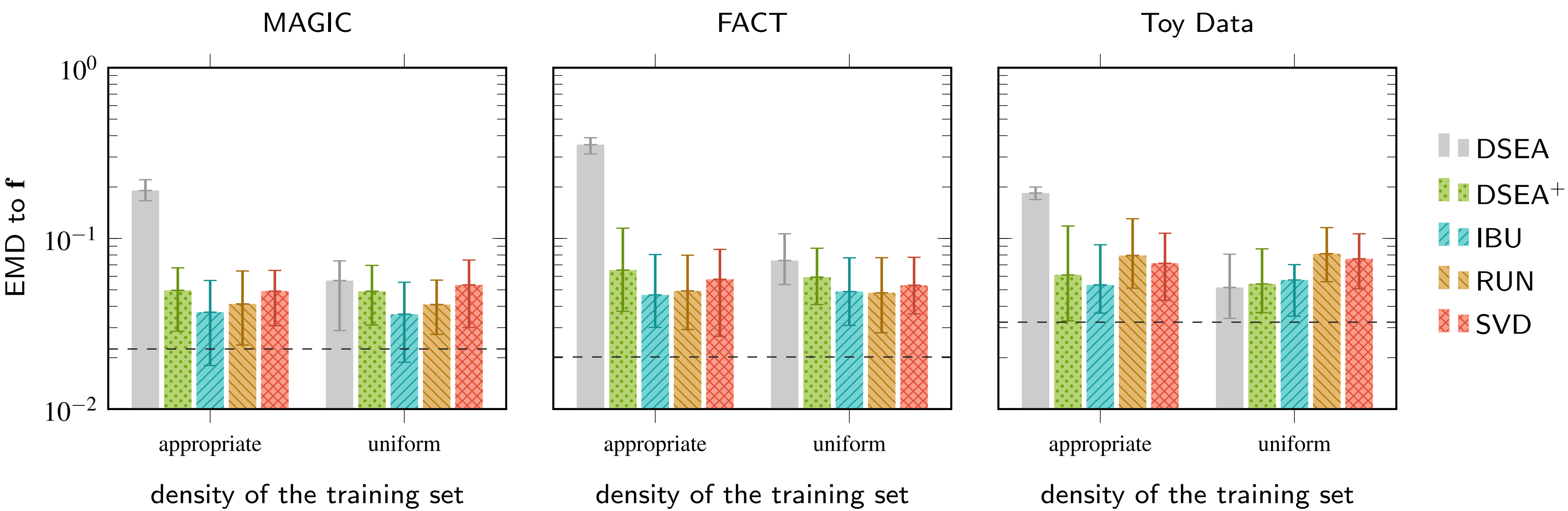
$$0 < \eta < 1$$

M. Bunse, Master thesis 2018.
Bunse, Ruhe et al., *Proc. of the 5th IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA)* 2018.

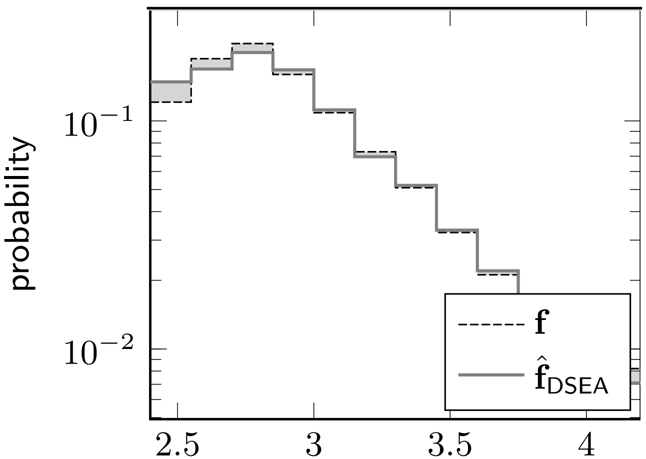


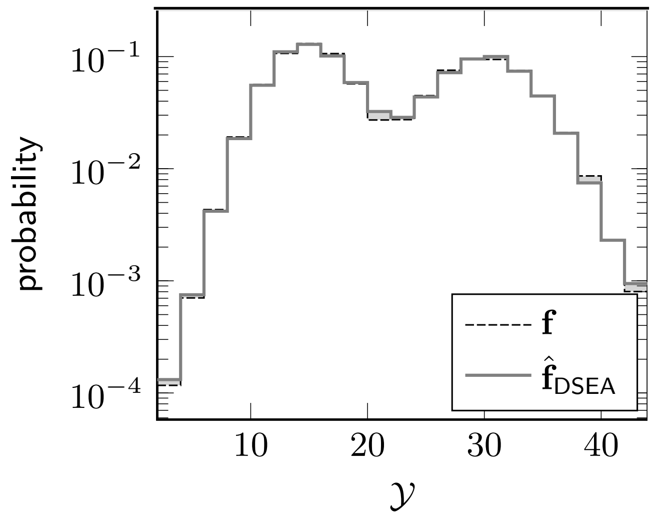
SFB 876 Providing Information by Resource-Constrained Data Analysis

DSEA+: Comparison of Different Algorithms



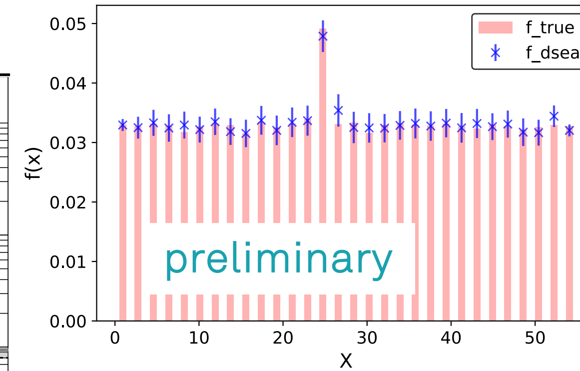
DSEA+: Preliminary Results

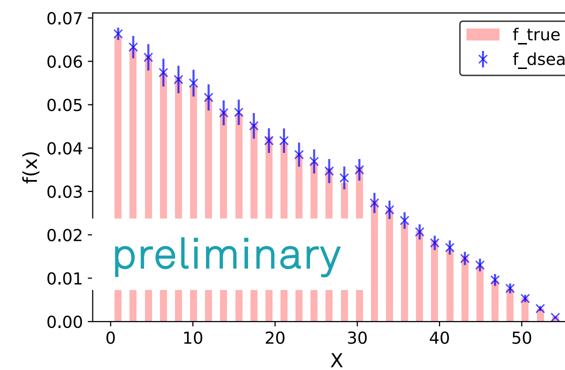




FACT

M. Bunse

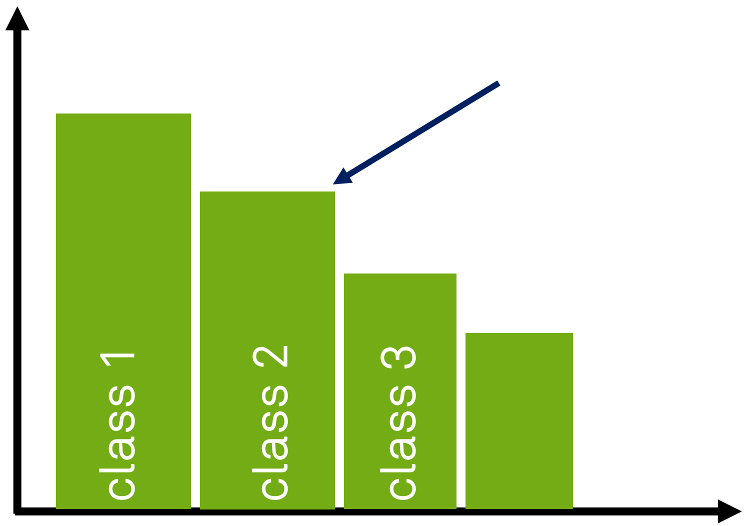




P. Schäfers

https://sfb876.tu-dortmund.de/deconvolution/index.html

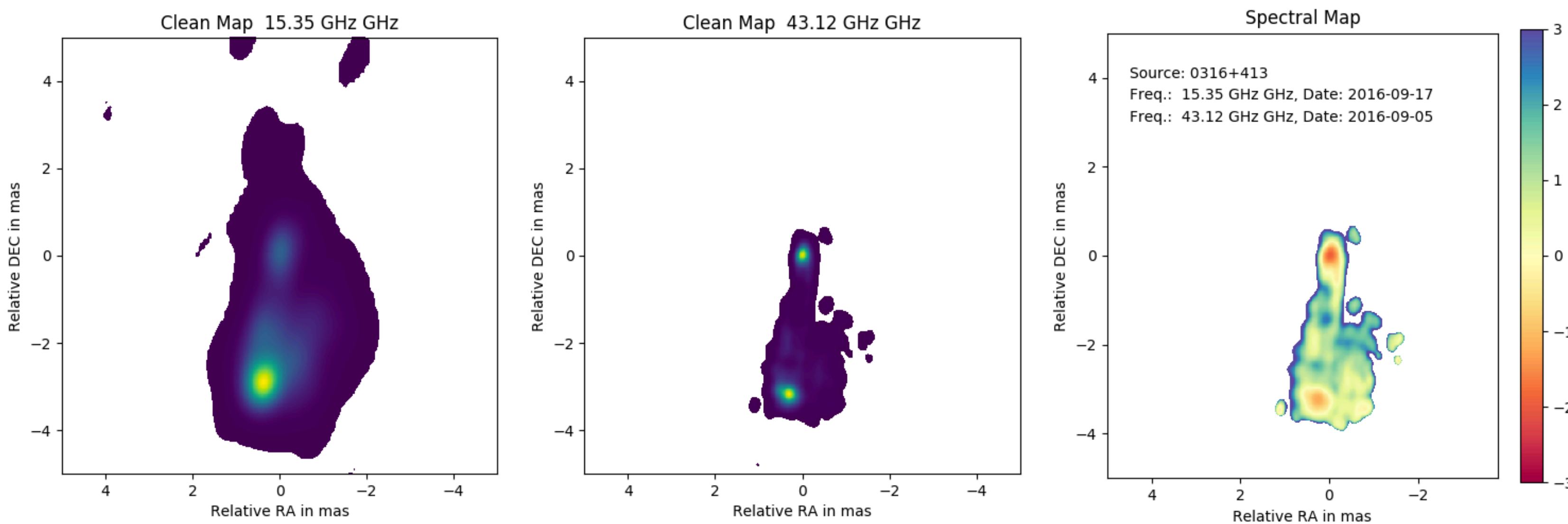
DSEA+ Future Plans: Neighbourhood Relations



- Binning is somewhat arbitrary
- Events located near upper and lower bin edges might be not just somewhat different
- Events near edges may be more like events in adjacent bins
- Class label is ordinal
- Take this into account appropriatly...

Radio Image Segmentation with Random Walks

L. Linhoff

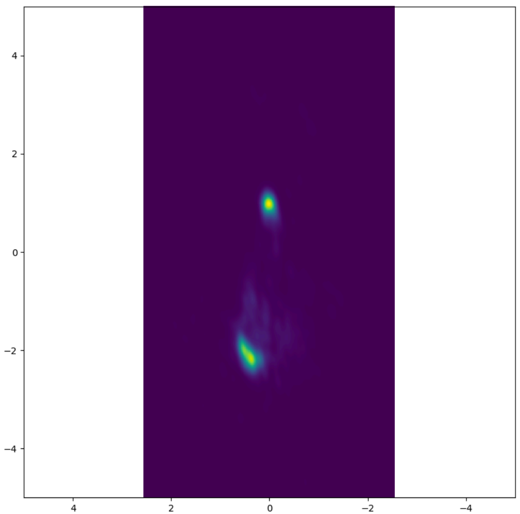


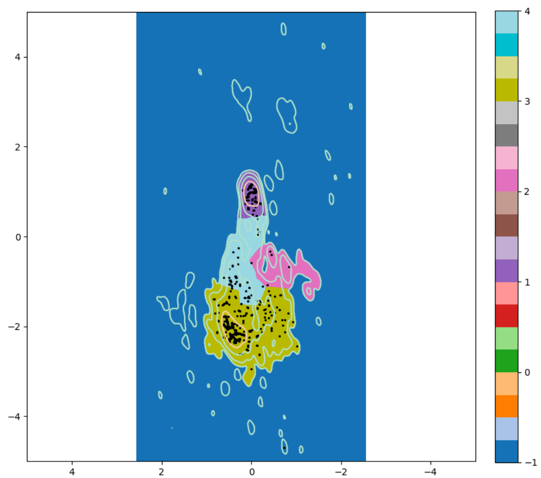
technische universität dortmund

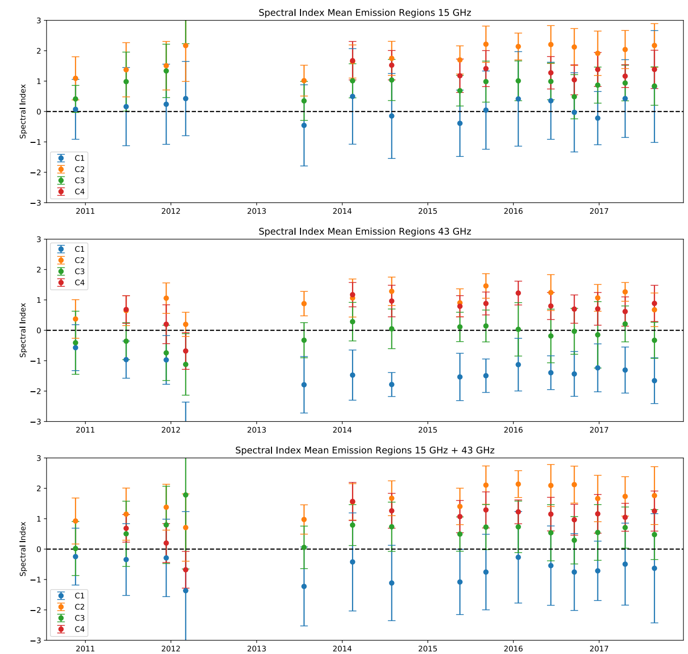
lehrstuhl physik e5

Radio Image Segmentation with Random Walks

L. Linhoff







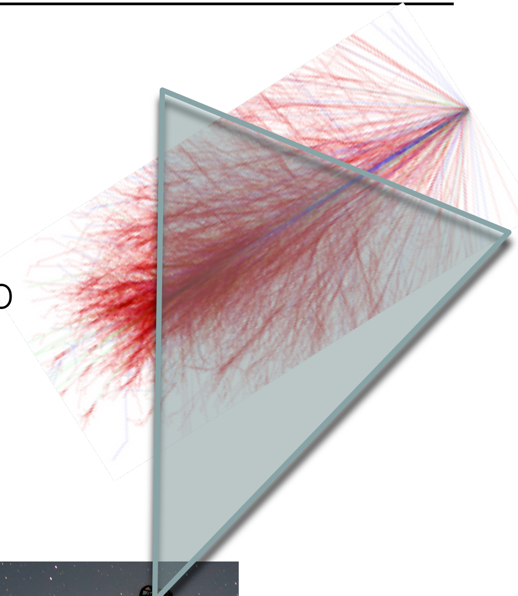
SFB 876 Providing Information by Resource-Constrained Data Analysis

Simulation: CORSIKA-Extension

D. Baack

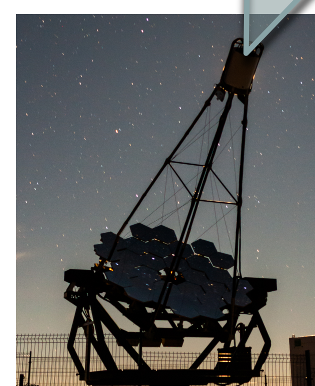
CORSIKA	100 %
Telescope Sim.	0.3 %
Cleaning	0.18 %
Precuts	0.11 %

Majority of particles in a shower does not contribute to Cherenkov light seen by the telescope.



FACT

First G-APD Cherenkov Telescope

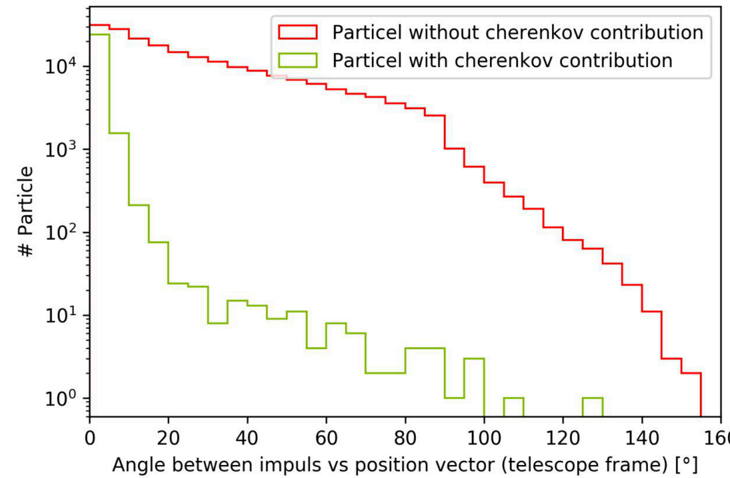


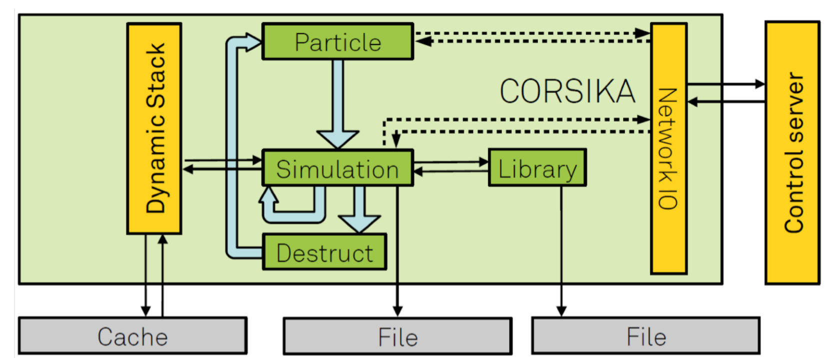
D. Baack, Technical Report,
https://sfb876.tu-dortmund.de/PublicPublicationFiles/baack_2016a.pdf

Simulation: CORSIKA-Extension

D. Baack

- Replace FILO stack with Dynamic Stack
- Stop simulating particles w.o. Cherenkov contribution
- 70% performance increase for FACT simulations
- Official part of CORSIKA as of March 2017

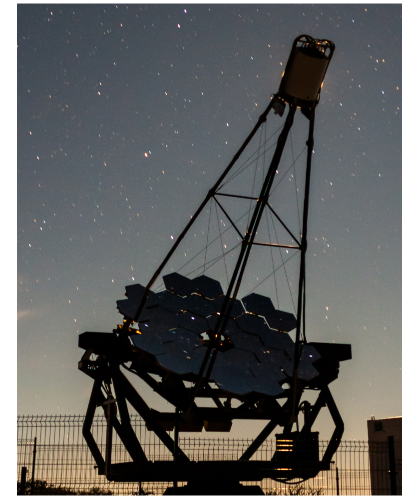




D. Baack, Technical Report, https://sfb876.tu-dortmund.de/PublicPublicationFiles/baack_2016a.pdf

The FACT Open Data Project

FACT
First G-APD Cherenkov Telescope



- Observations of the Crab nebula
- Point source gamma-ray simulations
- Diffuse gamma-ray simulations
- Diffuse proton simulations
- Data are available in multiple formats and at various stages of the analysis

- https://fact-project.org/data/

FACT Open Data

You can find an overview talk about the FACT open data release as PDF on github.

Crab Nebula Observations

In November 2017 the FACT Collaboration decided to make a first step into the direction of Open Data and release a sample of 17.7 hours of Crab Nebula observations measured in November 2013 available to the general public, along with simulations needed to perform analysis of this data sample.

We encourage to use this data for training, education and outreach for FACT and gamma-ray astronomy in general.

Please cite [1] and [2] if you use the data provided here.

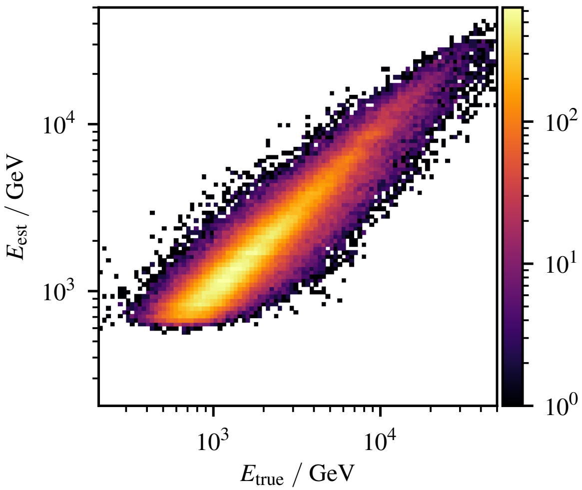
Links

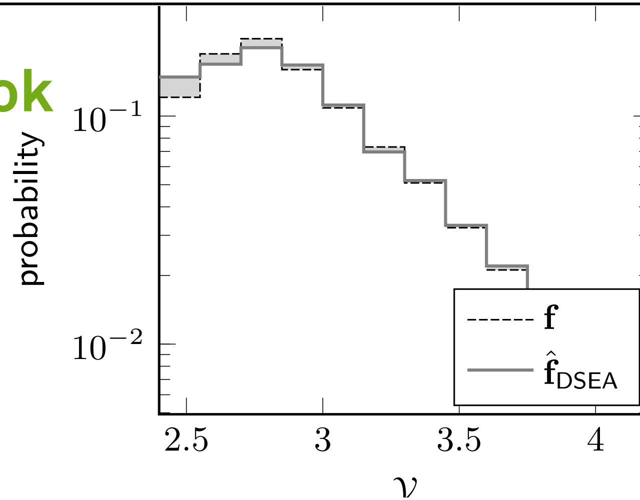
- FACT Open Data: https://fact-project.org/data/
- FACT Tools: https://github.com/fact-project/fact-tools
- aicttools: https://github.com/fact-project/aict-tools
- DSEA and DSEA+: https://sfb876.tu-dortmund.de/deconvolution/index.html

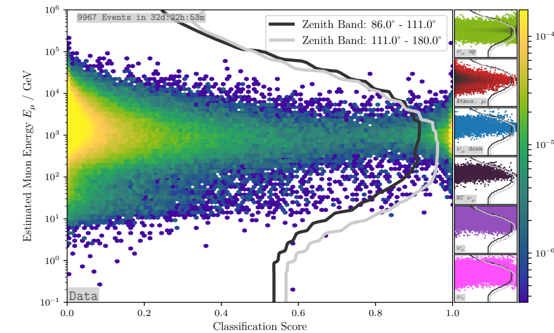
technische universität dortmund

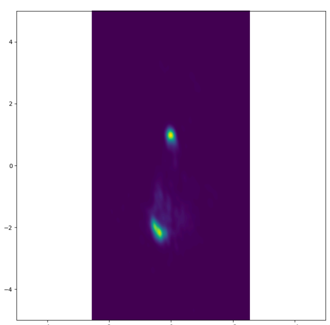
lehrstuhl physik e5

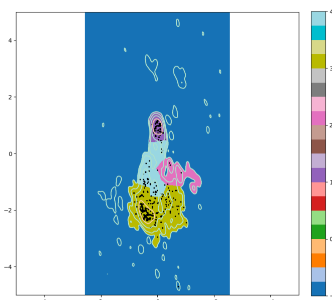
Summary and Outlook

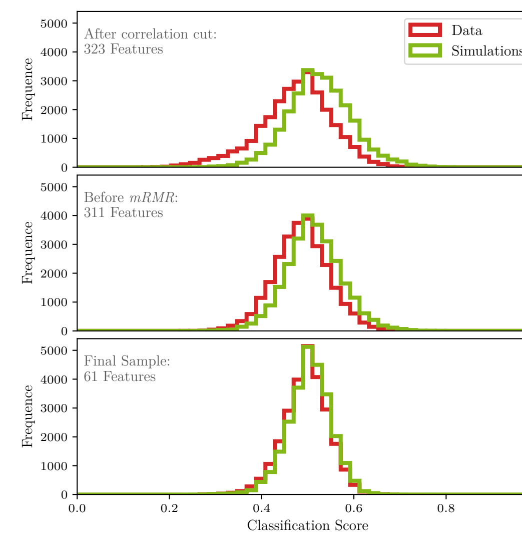












ctapipe

aicttools

gammapy

SFB 876 Providing Information by Resource-Constrained Data Analysis