technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

Project C3

Multi-level statistical analysis of high-frequency spatio-temporal process data

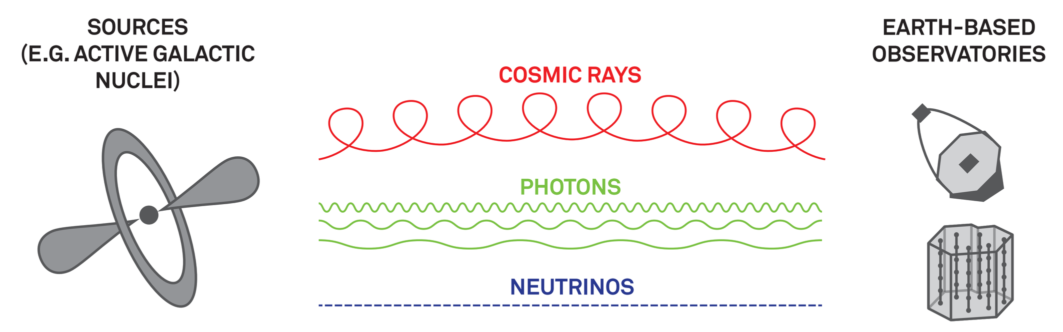
Prof. Dr. Katharina Morik, Prof. Dr. Dr. Wolfgang Rhode, Dr. Tim Ruhe

Problem

Big Data in Multi-Messenger Astronomy

Physics questions: origins of cosmic rays | cosmic particle acceleration | dark matter | fundamental physics at high energies

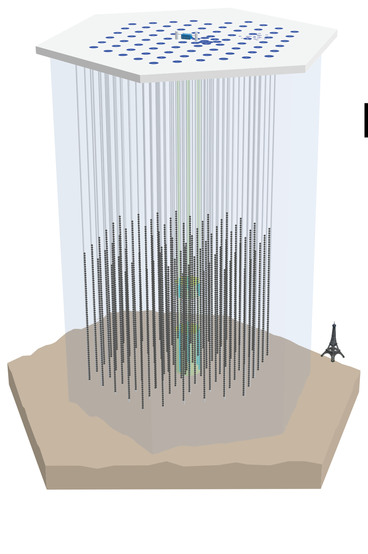
Common Challenges



- Extreme amounts of data
- Real-time analyses required to trigger follow-up observations
- Indirect measurement processes for the quantities of interest
- Signal-to-background ratios from 1 : 10^3 to 1 : 10^{10}
- No annotated samples from the real world

The need for and the possibility of extensive physical simulations are both a blessing and a challenge.

Neutrino Astronomy



- Extreme signal-to-background ratios ν_μ: 1 : 10^6, ν_τ: 1 : 10^{10}

IceCube

- Below the geographic South Pole
- 5160 light sensors in 1 km^3 of ice
- Complex high-dimensional and variable data
- Limited computational resources and bandwidth (100 GB/d)
- 3000 Events/s ⇒ **1 TB/d**

Gamma-Ray Astronomy

- Cherenkov radiation measured
- Atmosphere used as the detector volume
- 1 gamma-ray : 10^4 hadrons

FACT

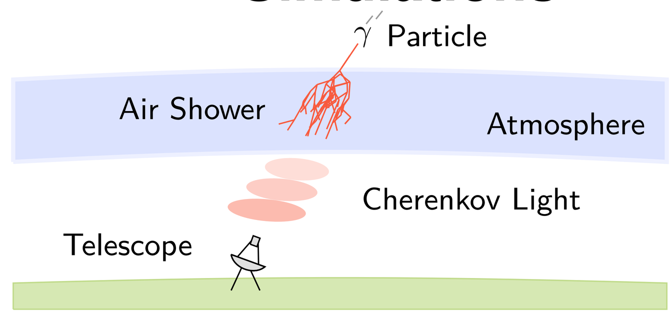
- New detector technology
- ~70 Events/s ⇒ **750 GB/d**
- Monitoring ⇒ fast alerts

MAGIC

- Stereoscopic system
- ~300 Events/s ⇒ **1 TB/d**

Simulations

- Particle interactions
- Propagation of leptons
- Cherenkov light production
- Propagation of photons
- Detector optics and electronics



- Several cluster years of computations
- More than 5 PB simulated data
- Many events are discarded in the first analysis steps
- **Great potential for optimization**

Methodology & Results

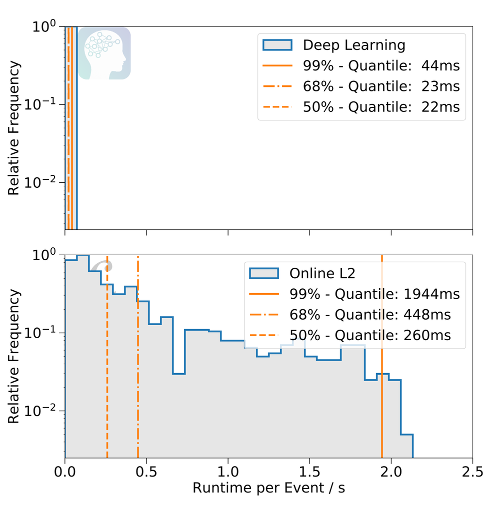
Real-Time Analysis

FACT [Bockermann et al. 2015] (A1)

- Full real-time analysis chain for FACT
- Online application of pre-trained machine learning models
- **Best Paper Award ECML 2015**

IceCube [Hünnefeld 2017]

- Deep learning improves reconstruction while being orders of magnitude faster
- Constant application time prevents event pileup



Improved Event Reconstruction through Machine Learning

FACT [Nöthe 2017]

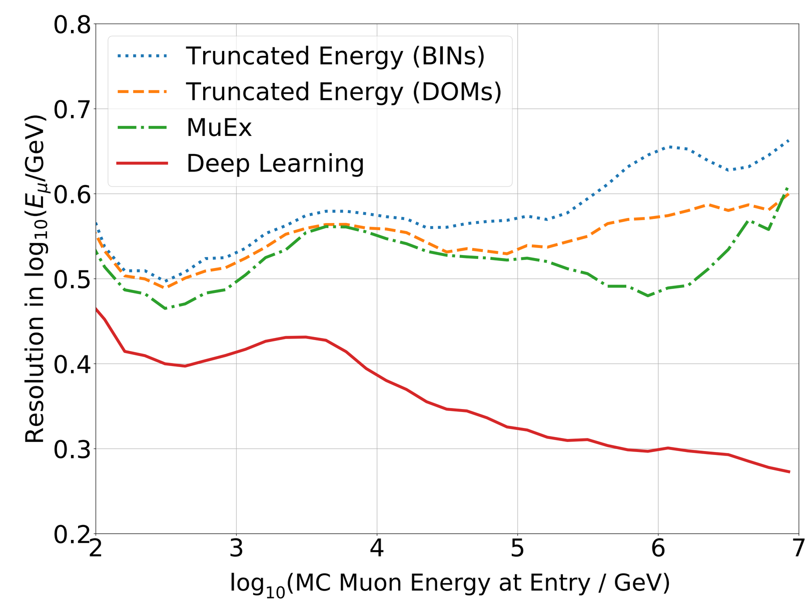
- All reconstruction tasks are now solved using random forests (classification & regression)
- Greatly improved angular resolution
- Background suppression improved by a factor of 2
- 50 % gain in sensitivity
- Evaluation of deep learning started

IceCube [Hünnefeld 2017]

- Deep learning improved reconstruction in many areas
- Random forests improved sensitivity of search for rare tau neutrino events by a factor of 2

MAGIC [Mielke 2017]

- Deep learning improved energy resolution by a factor of 2 in the higher energies



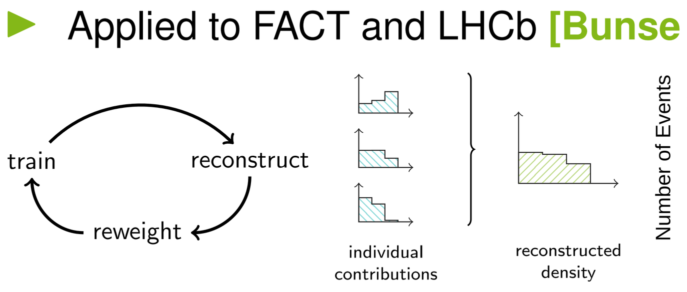
Resolution of energy reconstruction algorithms for IceCube. The novel deep learning reconstruction outperforms other methods by large margins.

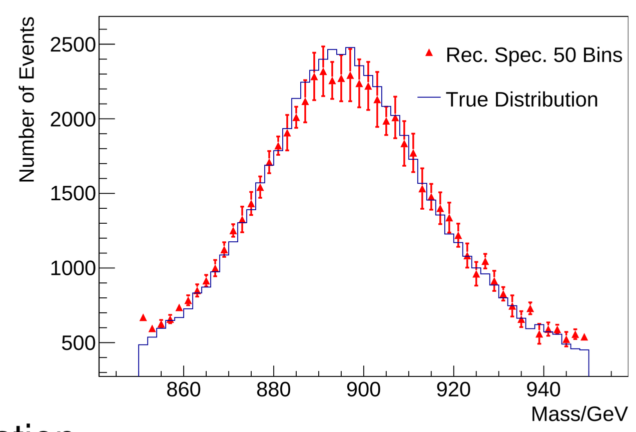
Novel Unfolding Techniques

- Indirect measurements result in limited resolution and acceptance
- Unfolding methods **reconstruct the probability distribution** of energies (spectrum) from the observed quantities

Dortmund Spectrum Estimation Algorithm (A1) (C5)

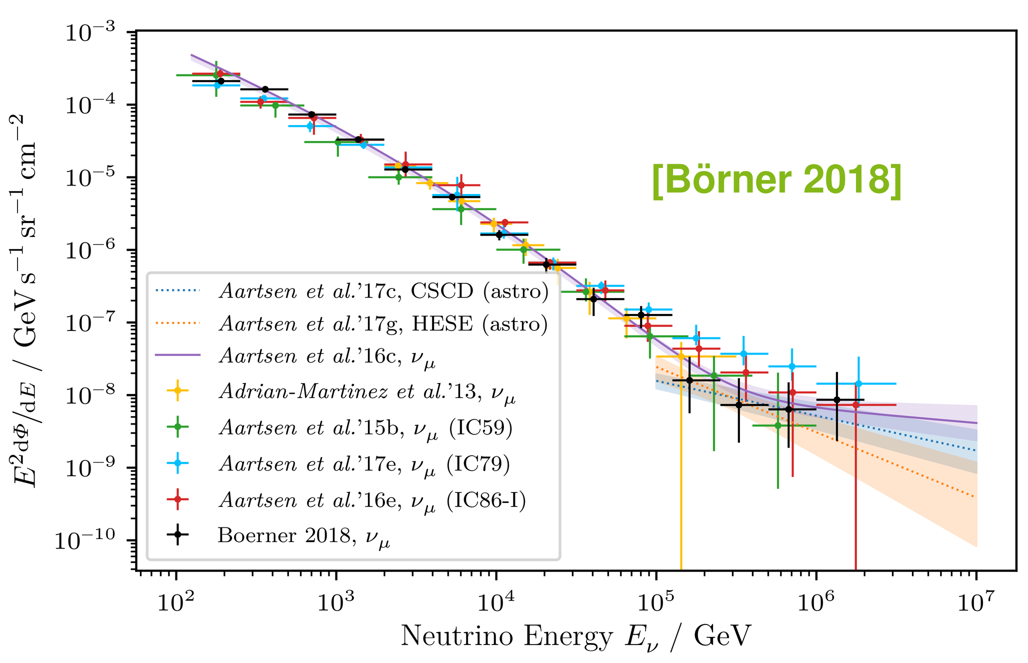
- Individual information by multivariate estimation of class probabilities
- Applied to FACT and LHCb [Bunse et al. 2018] [Ruhe et al. 2016]





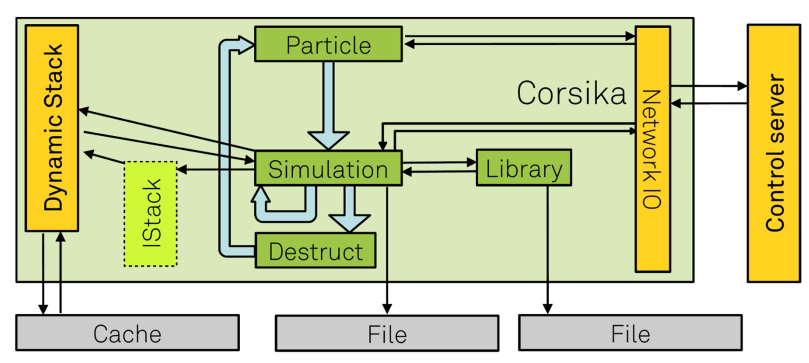
MCMC-Likelihood-Sampling

- Estimates uncertainty and correlation
- Accounts for systematic uncertainties as nuisance parameters

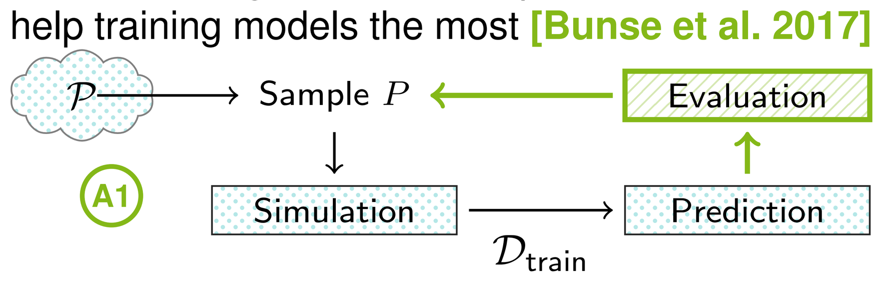


Increasing Simulation Efficiency

- **Abort unnecessary simulations** as early as possible [Baack 2016a]
- Flexible stack implementation allows for prioritization and discarding of particles



- Active learning for simulation parameters that help training models the most [Bunse et al. 2017]



- Network interface allows for parallelization of single shower simulations and optimal usage of cluster resources

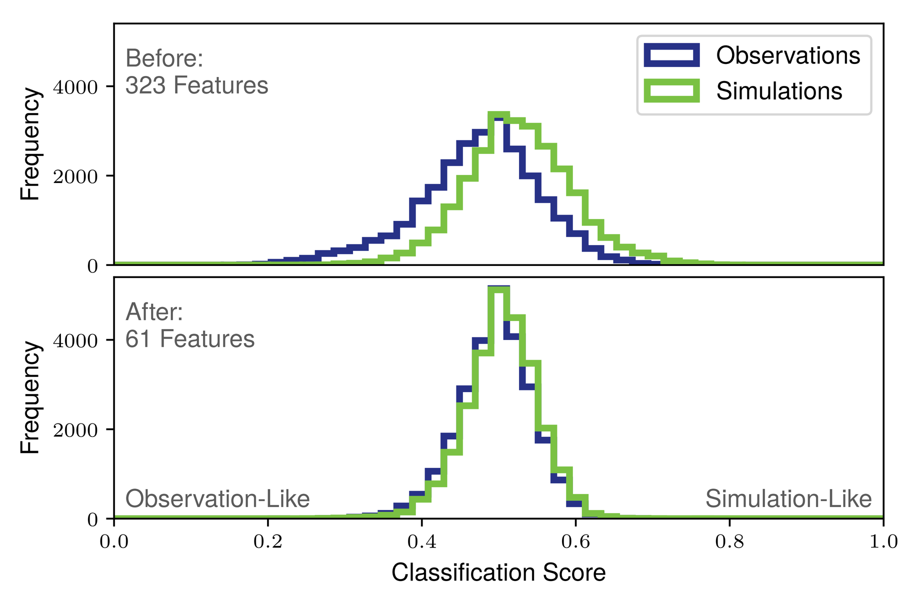
IceCube

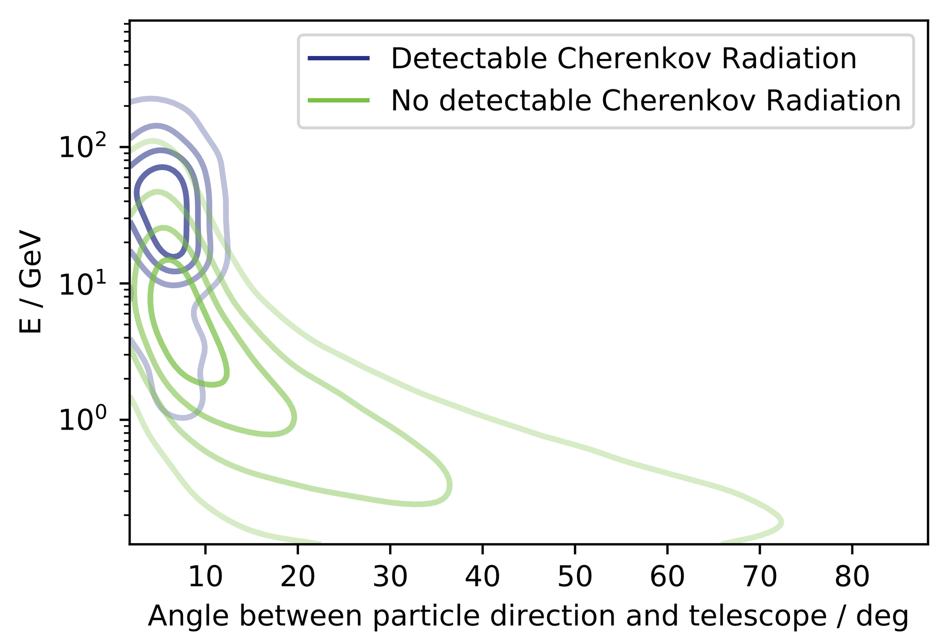
- Showers without high-energy muons are aborted
- Run-time reduced from 169.5 ms to 2.5 ms per shower

FACT [Baack 2016b]

- Early abortion of particles unlikely to produce detectable Cherenkov radiation
- Run-time reduction of ≈ 80 % per shower

- Perfect agreement between observations and simulations is hard to achieve
- Improving agreement by feature selection
- **Remove features that distinguish observations from simulations** [Börner et al. 2017]





lehrstuhl exp. physik 5

Artificial Intelligence Group