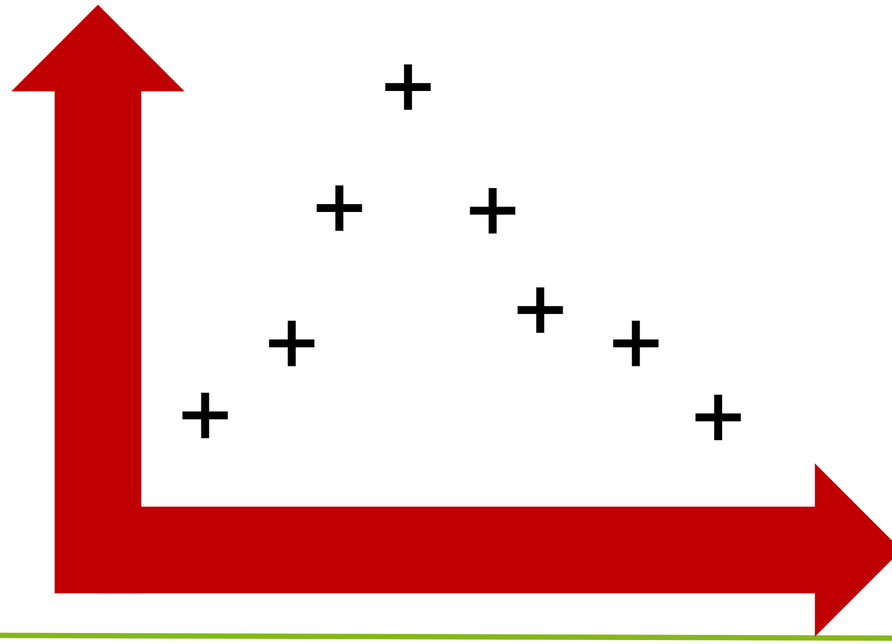
technische universität dortmund

lehrstuhl physik e5



# DSEA – A Data Mining Approach to Unfolding

Tim Ruhe

12/08/2016

# Outline

- Why unfold?
- The Dortmund Spectrum Estimation Algorithm (DSEA)
- Performance Studies using Toy MC simulation
- Summary & Outlook

**SFB 876** Providing Information by Resource-Constrained Data Analysis

Other people who contribute to DSEA:

Max Wornowizki, Tobias Voigt, Mathis Börner, Martin Schmitz

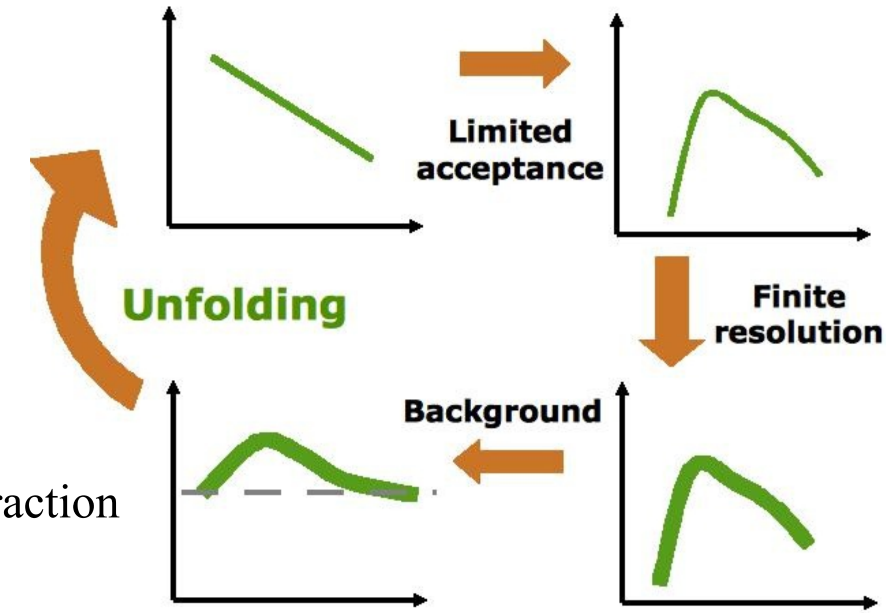
# Why Unfold?

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 spectrum

Physics of neutrino interaction




Additional smearing due to limited acceptance and finite resolution.

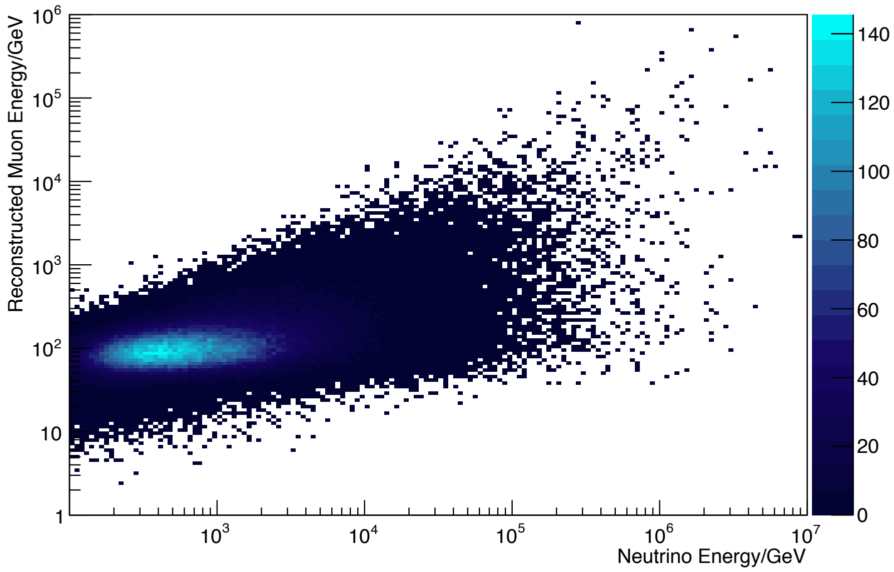
$$g(y) = \int_{E_{\min}}^{E_{\max}} A(E,y) f(E) dE$$

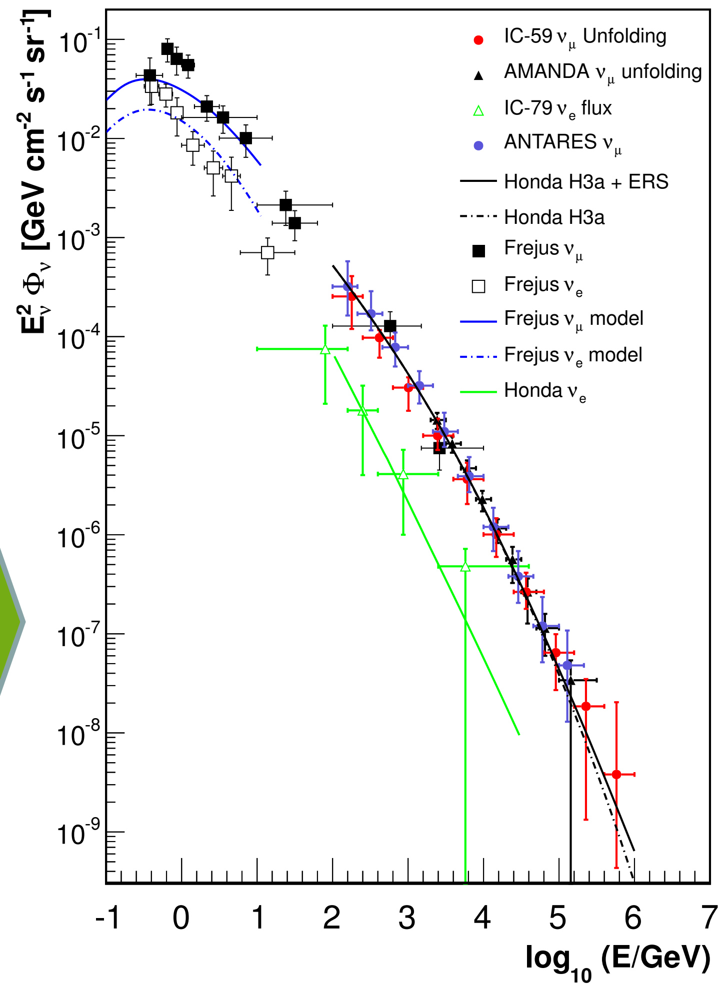
Fredholm Integral equation of the first kind

# Unfolding Spectra of Atmospheric Neutrinos

- Example taken from atm. $\nu_\mu$ measurement with IceCube.

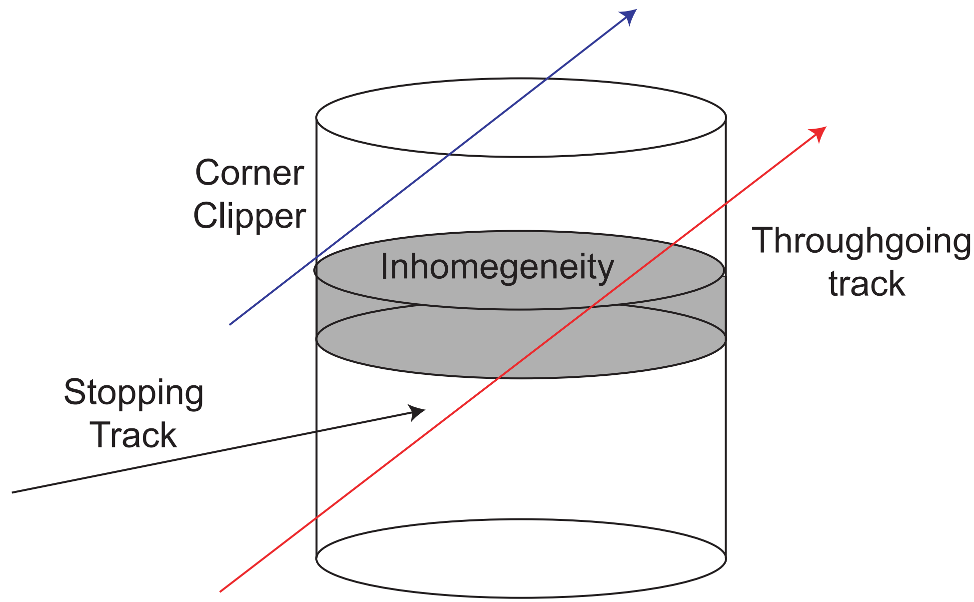
Aartsen et al., EPJC 75, 116 (2015)



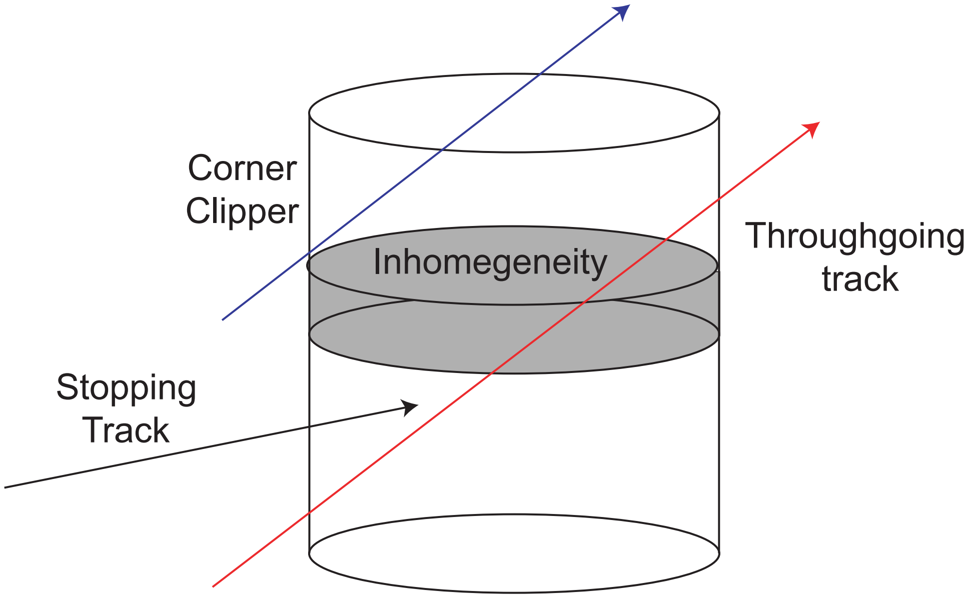
# Why do we need to improve on existing algorithms?

- Events with identical energy may cause very different patterns inside the detector
- Geometrical information will increase precision of the measurement
- Existing algorithms are limited in the number of input variables („curse of dimensionality“)
- Algorithms return spectra accurately, but information on individual events is lost
- Generally require a certain amount of unfolded events to obain a spectrum



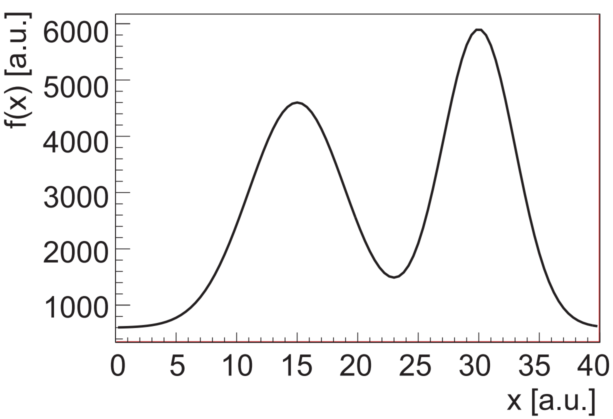
## Why do we need to improve on existing algorithms?

- Events with identical energy may cause very different patterns inside the detector
- Geometrical information will increase precision of the measurement
- Existing algorithms are limited in the number of input variables (“curse of dimensionality“)
- Algorithms return spectra accurately, but information on individual events is lost
- Generally require a certain amount of unfolded events to obain a spectrum

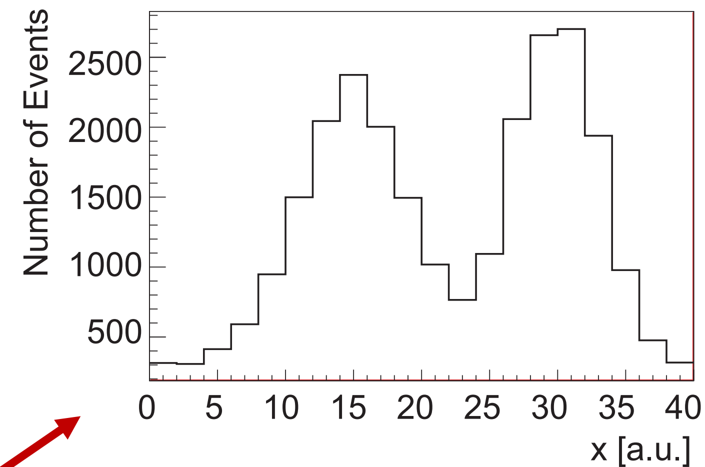



We need algorithms that can utilize an arbitrary number of input variable and fully retain the information on individual events.

# DSEA – The General Idea




Discretize

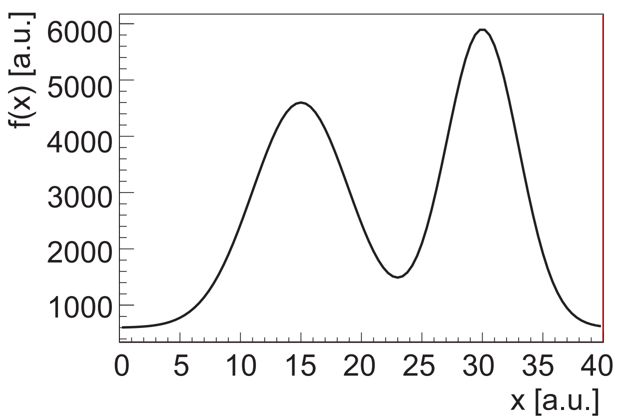



Classify

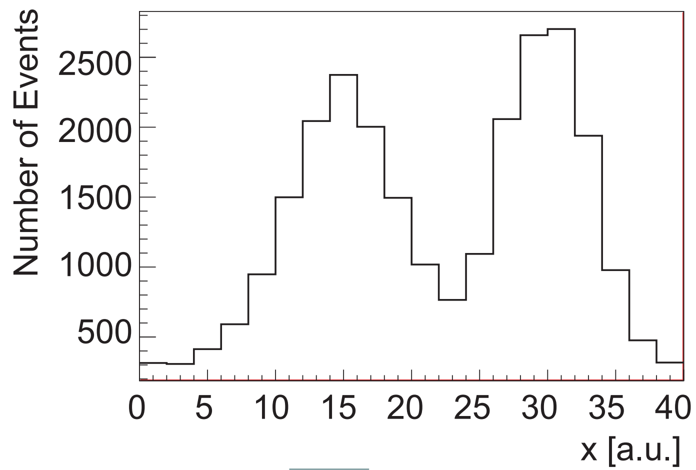
Inverse Problem is transformed into a multinominal classification task that can be solved by an arbitrary classifier.

Really beautiful spectrum

# DSEA – The General Idea




Discretize




Classify

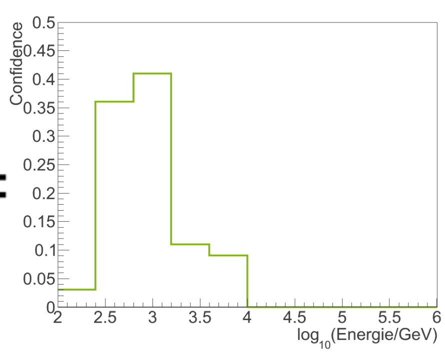
- Arbitrary number of input variables
- Information on individual events is fully retained

Really beautiful spectrum

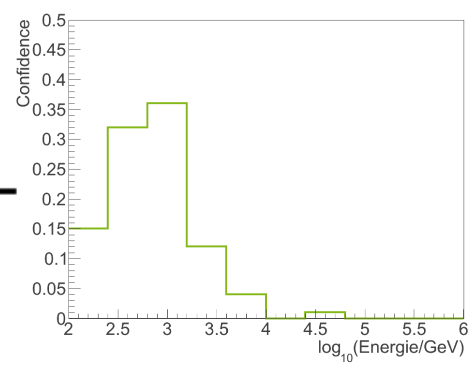
# However, it is not really that simple...

- Maximum of confidence distribuition generally not very distinct
- Therefore, classification also not very distinct
- Results in poorly reconstructed spectrum
- Solution: Interpret confidence distribution as estimation of pdf

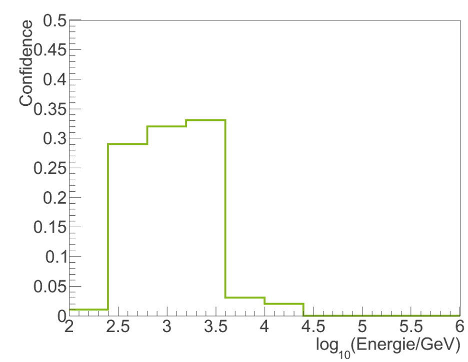
$\hat{f} =$




$+$

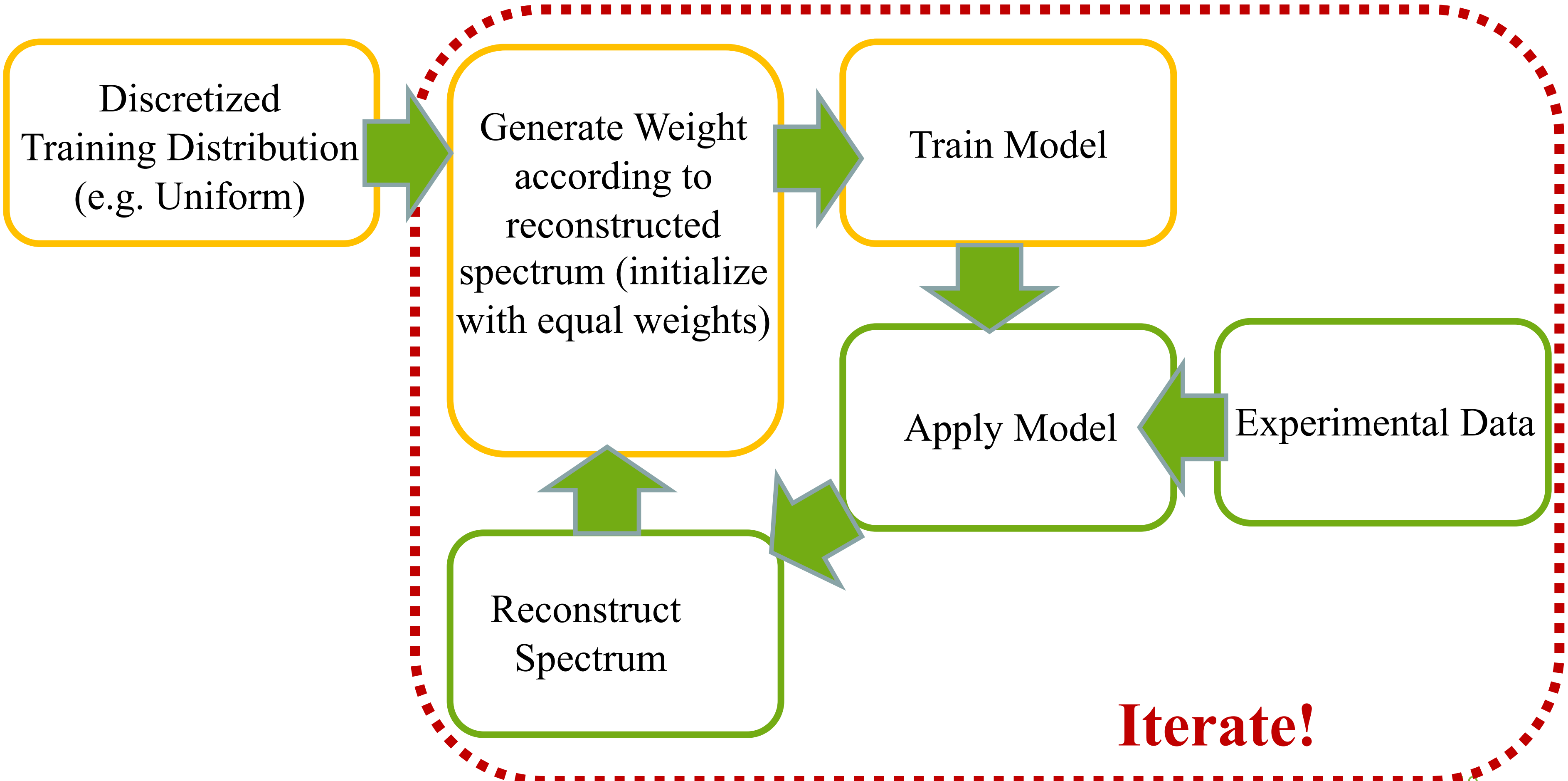



$+ \cdots +$



## Don‘t you only get out what you put in?

**NO!**



# DSEA – Full Size Technical Description

1. **Discretize** $f(x) \longmapsto \vec{f}(x) = (f_1, \ldots, f_m)$. **(Initialize)**
2. **Train Model** A subset of $n$ examples $(\underline{A}, W, L) = \{(\vec{a}, w, l)_1; \ldots; (\vec{a}, w, l)_n\}$ is used to train a model $M(\underline{A}, W, L)$. Each example consists of a label $l$, a weight $w$ and $h$ attributes $\vec{a} = (a_1, \ldots, a_h)$.
3. **Apply Model** The Model $M(\underline{A}, W, L)$ is applied to a set off $\tilde{n}$ unlabeled examples $\underline{\tilde{A}} = (\vec{\tilde{a}}_1, \ldots, \vec{\tilde{a}}_{\tilde{n}})$ yielding a confidence $c_{i,j} = g(M(\underline{A}, W, L), \vec{\tilde{a}}_i)$ for the $i$-th example t o belong to the $j$-th bin in $\vec{f}(x)$.
4. **Reconstruct Spectrum** For the $k$-th iteration, the bin content $\hat{f}_{j,k}$ of the $j$-th bin is estimated as $\hat{f}_{j,k} = \sum_{i=1}^{\tilde{n}} c_{i,j}$.
5. **Update weights** The example weights for the $(k+1)$-th iteration are updated according to $w_{i,k+1} = \frac{\hat{f}_{k,l_i}}{\tilde{n}}$. **(Continue with Step 2)**

Cute little puppy. Makes slide look less scary.

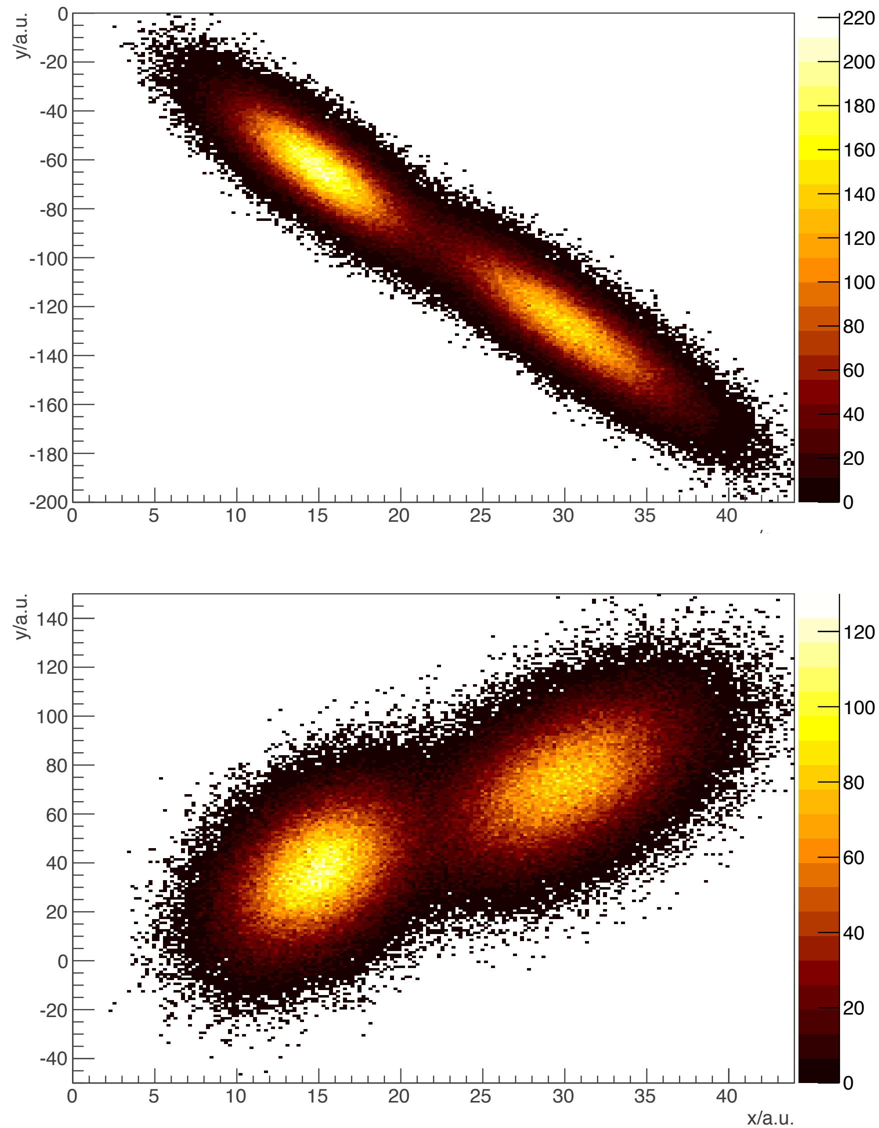
technische universität dortmund

lehrstuhl physik e5

# Performance Studies using Toy MC

- Generate function with two-peak structure (2 Gaussian with different mean and width, test set)
- Generate uniform distribution using identical settings (training set)
- Generate 10 correlated variables, apply Gaussian smearing (amount of smearing is randomized within pre-defined range, but kept constant during simulation)

- Statistical uncertainties are obtained by running DSEA 10 times, varying the random seed for the sampling of the training set



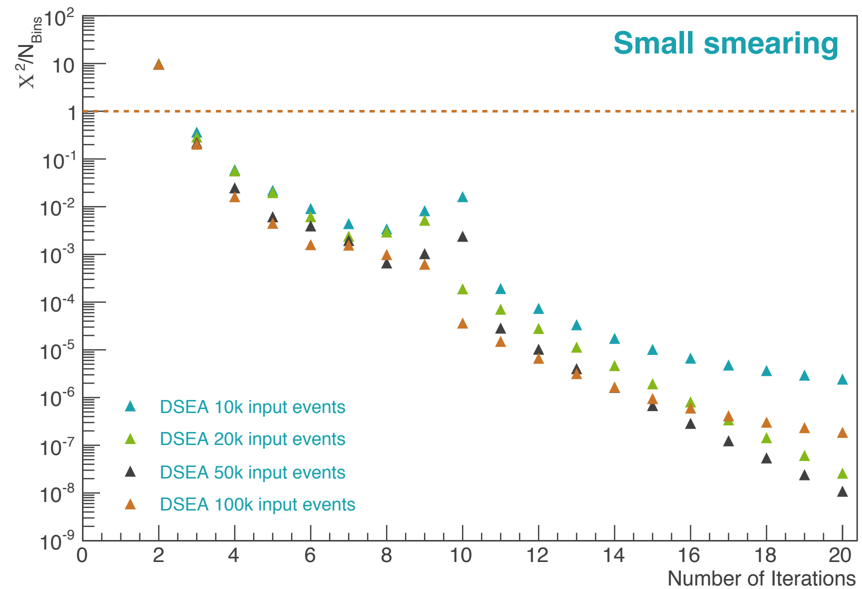
# Convergence – when to stop iterating?

General rule from Iterative Bayesian Unfolding:

$$\frac{\chi^2}{m} \leq 1$$

Where

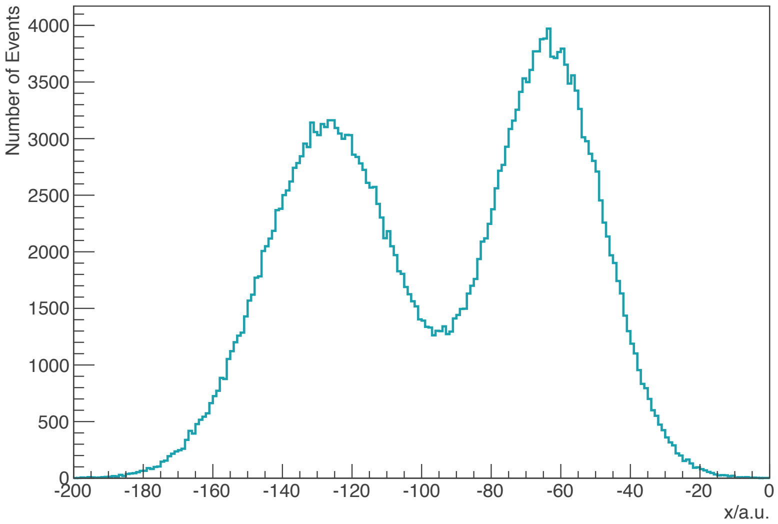
$$\chi^2 = \sum_{j=1}^{m} \left( \frac{\hat{f}_{j,i-1} - \hat{f}_{j,i}}{\sqrt{\hat{f}_{j,i-1}}} \right)^2$$

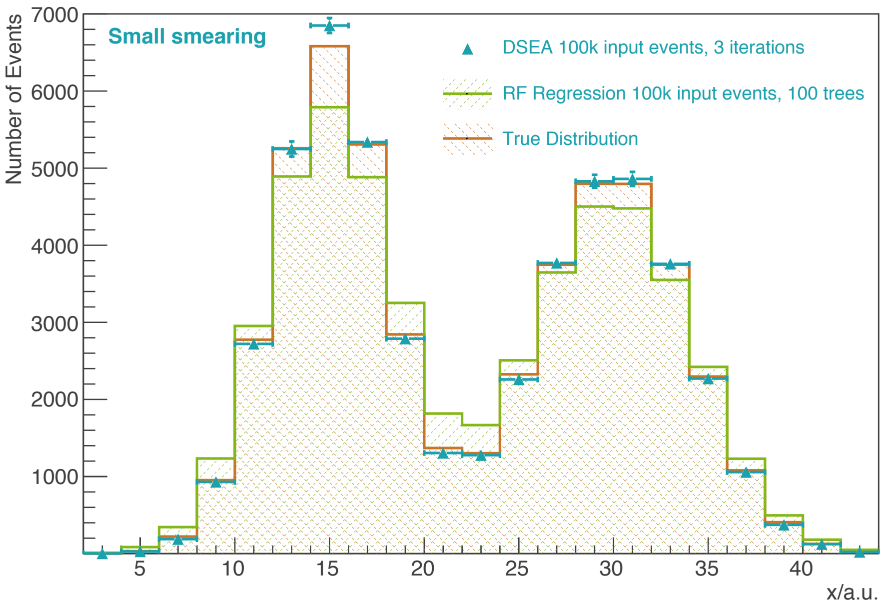


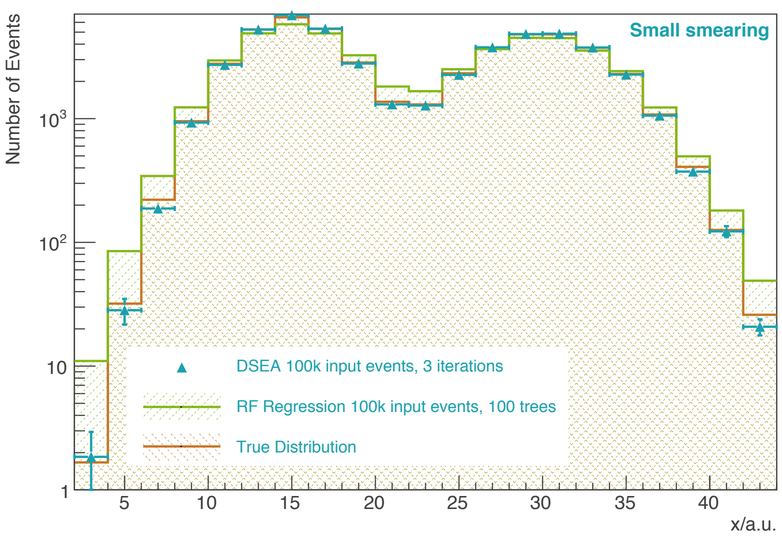

- Convergence criterion reached after 3 – 5 iterations
- Some dependence on the number of input examples

# Performance – Spectral Reconstruction

Spectral Reconstruction for small smearing using DSEA (blue), compared to a Random Forest regression (green).

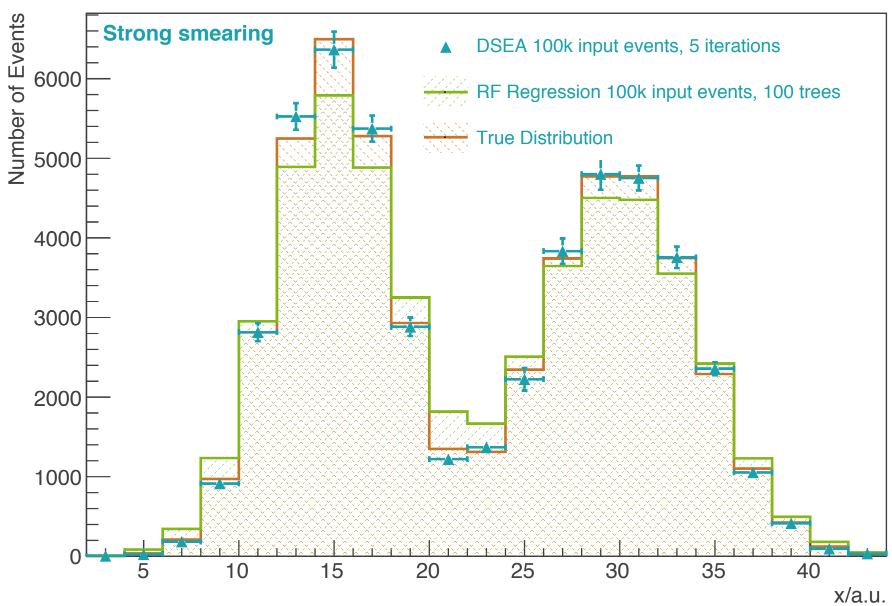


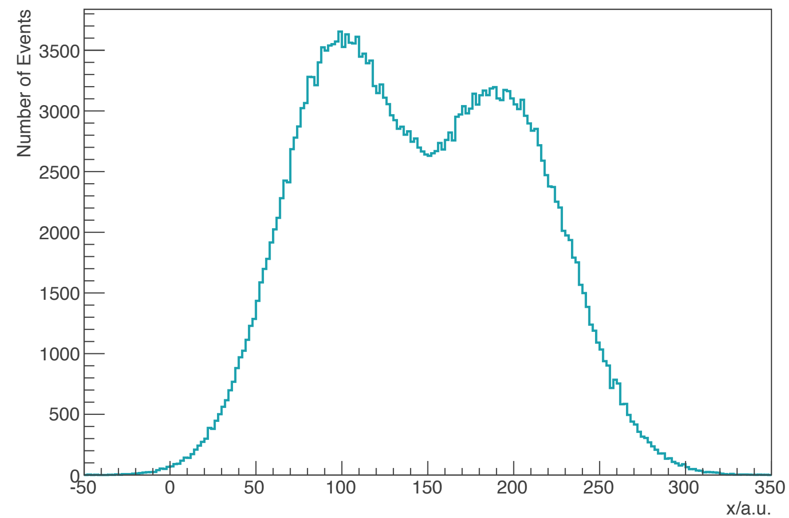



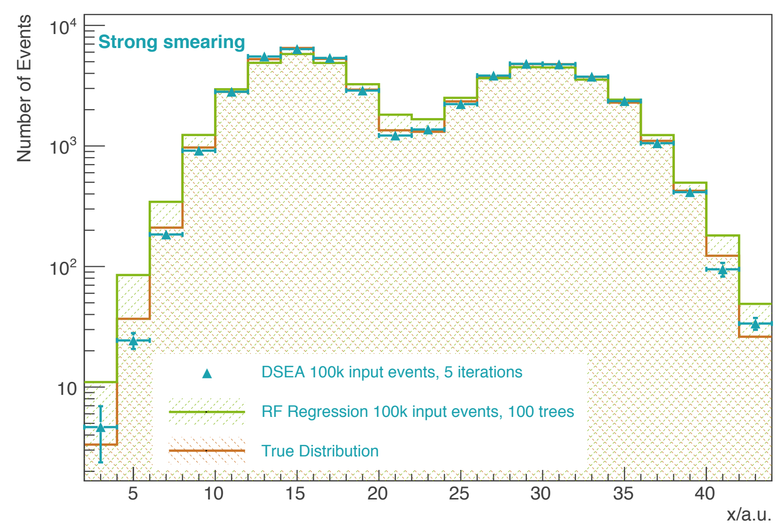



# Performance – Spectral Reconstruction

Spectral Reconstruction for strong smearing using DSEA (blue), compared to a Random Forest regression (green).

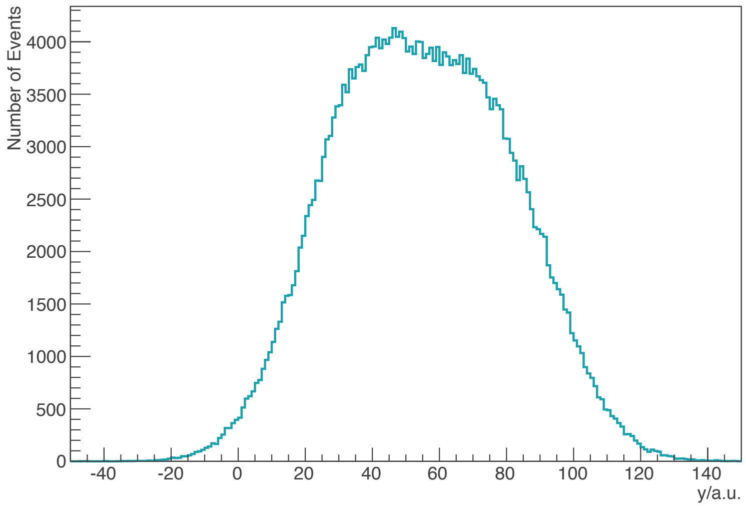


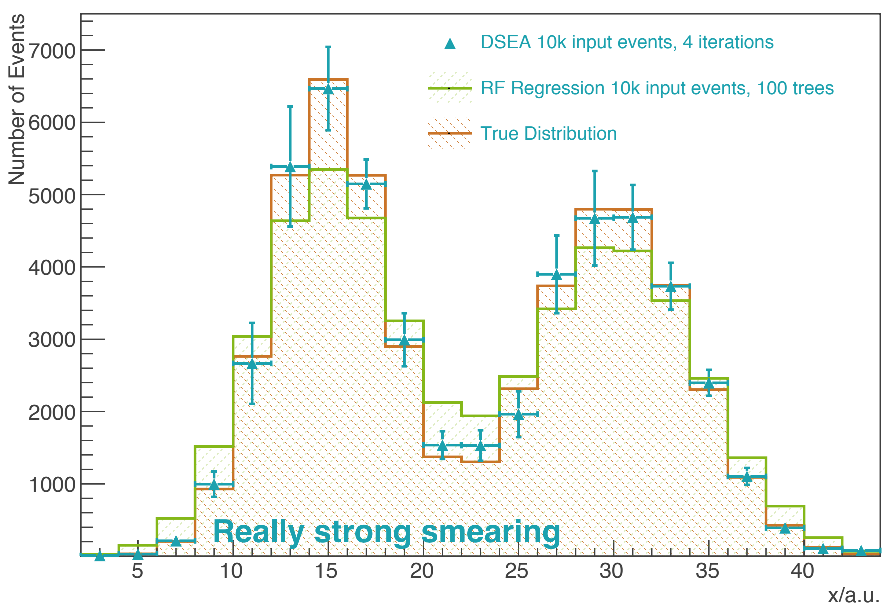



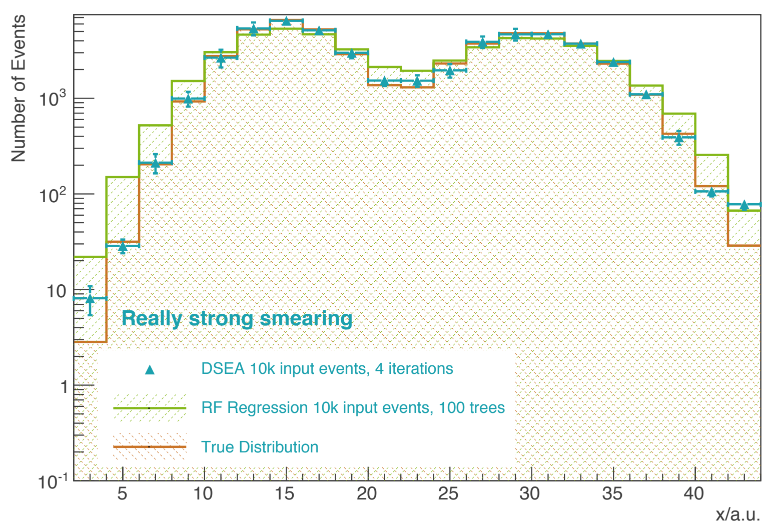



# Performance – Spectral Reconstruction

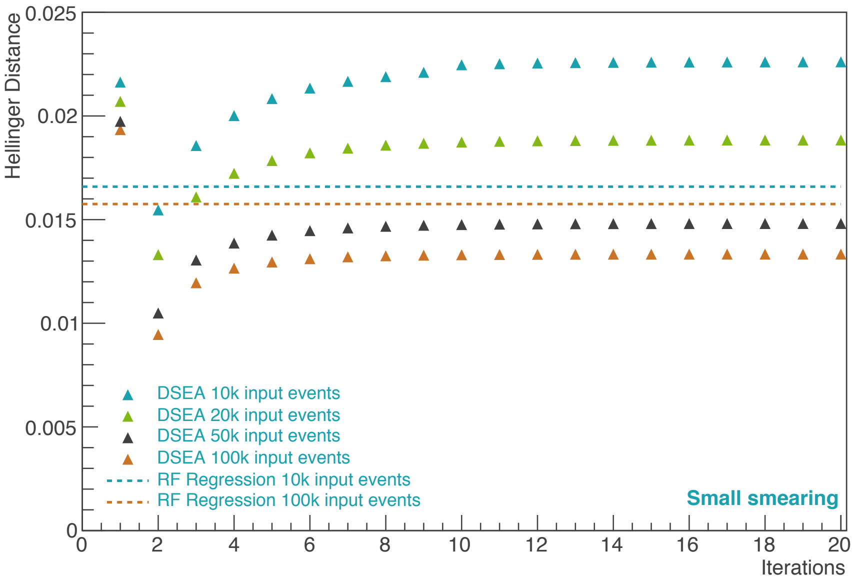
Spectral shape (two peaks) no longer visible in input variables, but spectrum correctly reconstructed using only 10k examples!



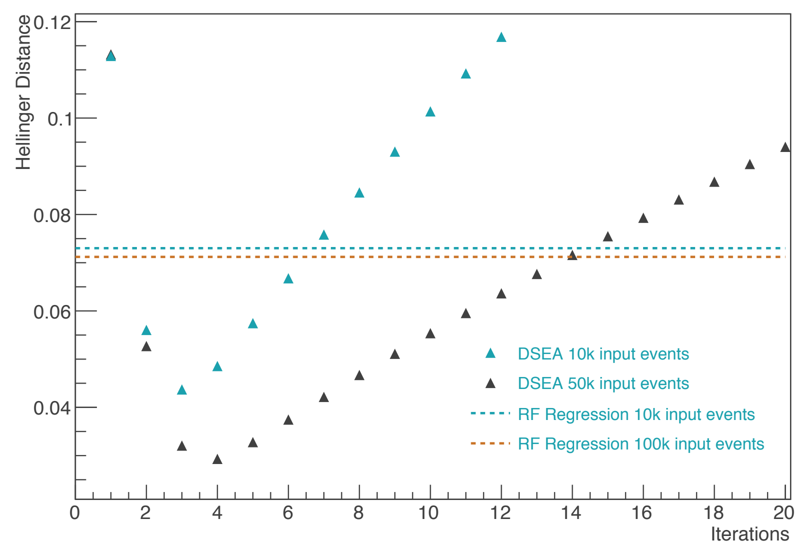
# Performance – Agreement with Underlying Distribution




Hellinger Distance:

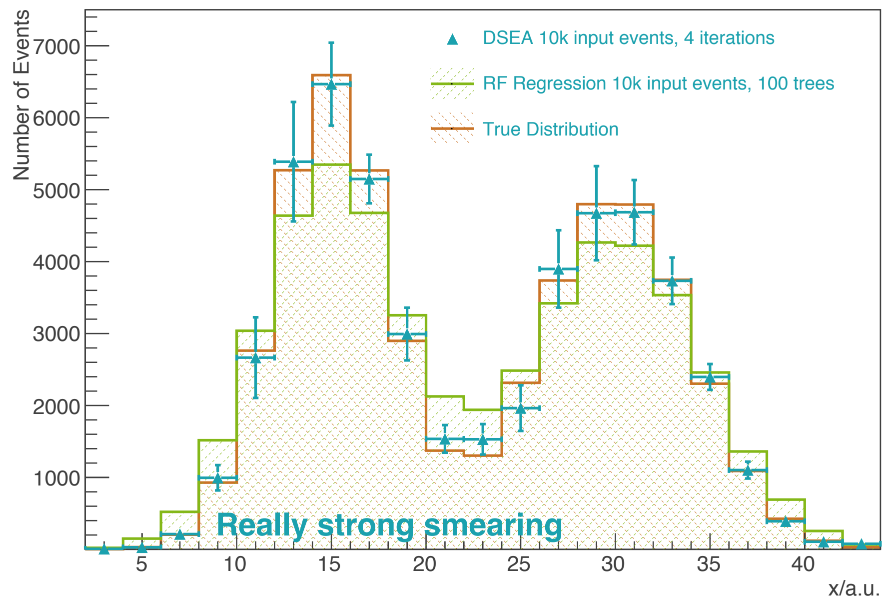
$$H^2(f,\hat{f}) = \frac{1}{2}\sum_{j=1}^{m}\sqrt{f_j} - \sqrt{\hat{f}_j}$$

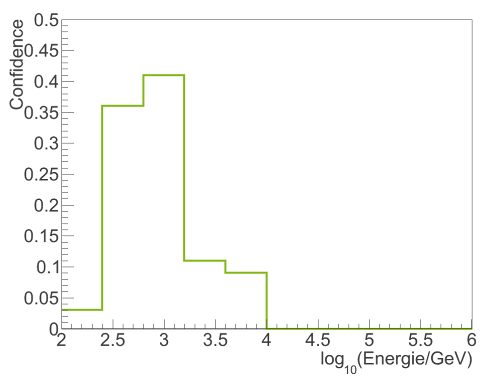
Agreement tends to be optimal for 2 to 5 iterations, gets worse if too many iterations are used.

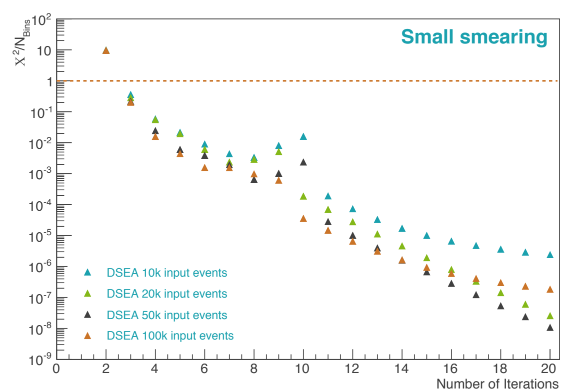


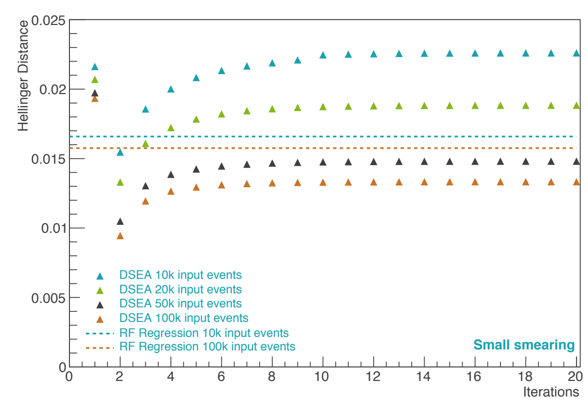
# Summary and Outlook

- Apply DSEA to other types of function (saw tooth, steeply falling, steeply rising,...)
- Obtain fair comparison to existing approaches
- Apply DSEA in spectral measurement on simulated experimental data



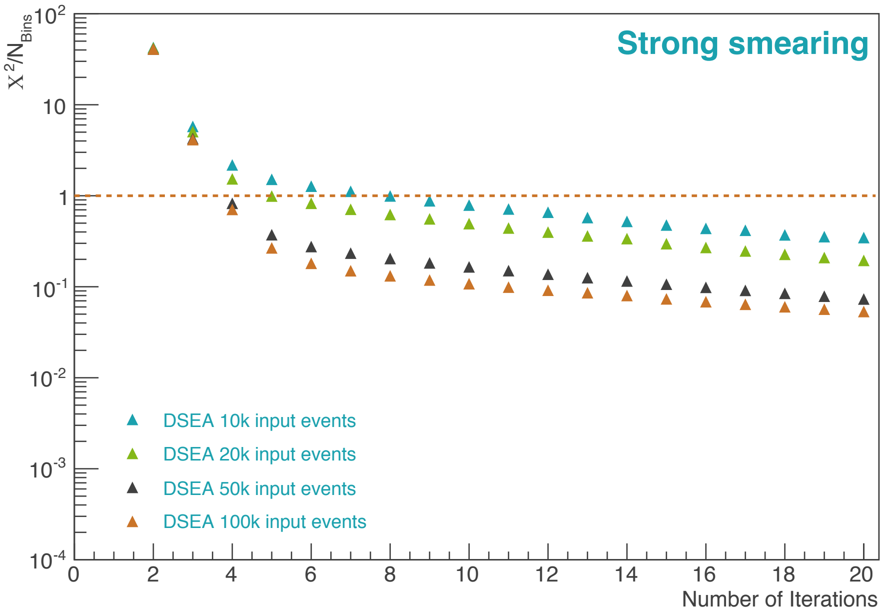
# Backup Slides

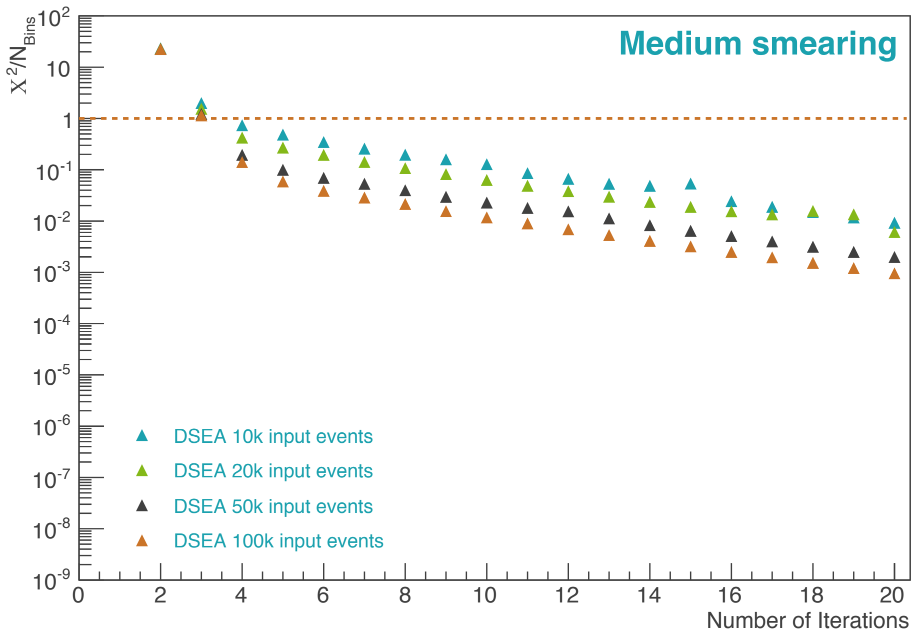
Here be dragons.

# Overlap with Existing Algorithm

- Some overlap with Kernel Density Estimation, but...
  - ... Kernel function not predefined, so confidence distribution can be anything
  - ... „Kernel“ in DSEA is discrete
- Some overlap with Iterative Bayesian Unfoling, especially if Naive Bayes is used as a classifier, but...
  - ... Event-by-event, rather than bin-by-bin
  - ... Fully retains information on individual events

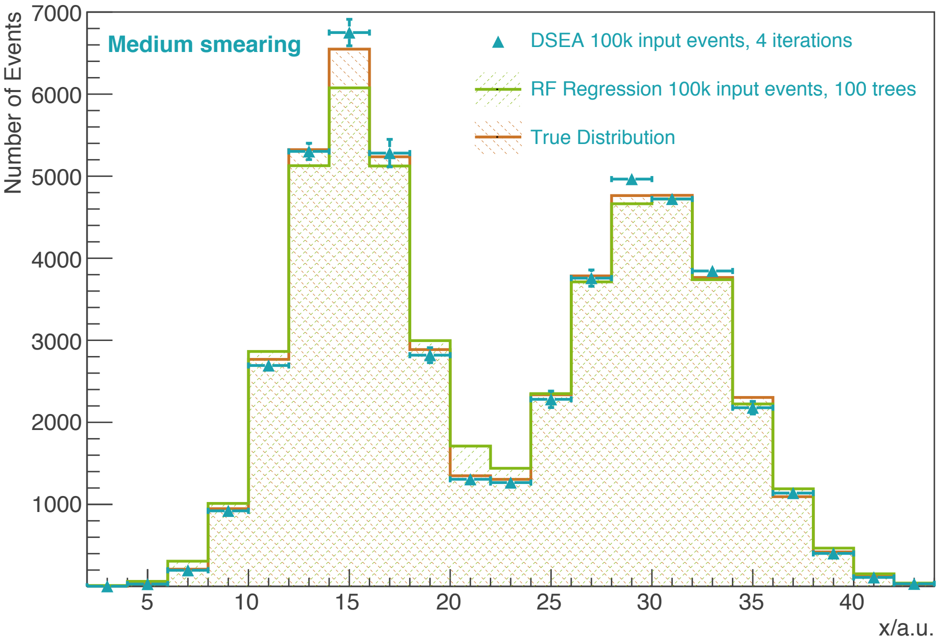
# Convergence for Strong and Medium Smearing

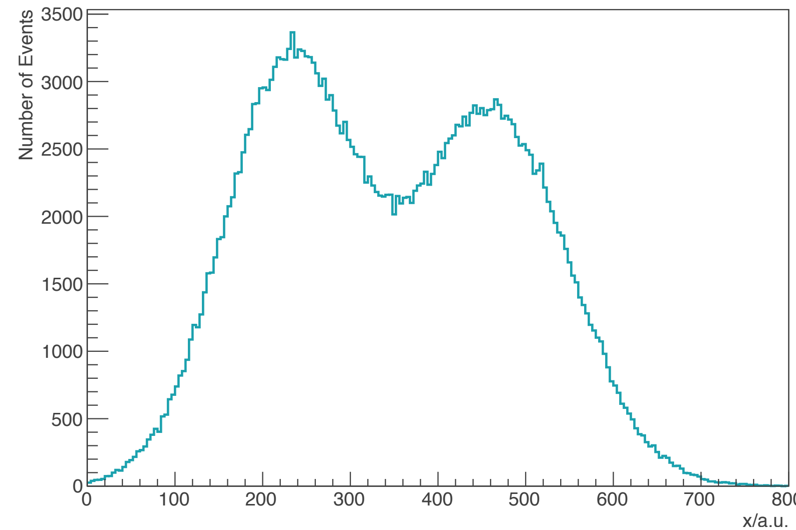


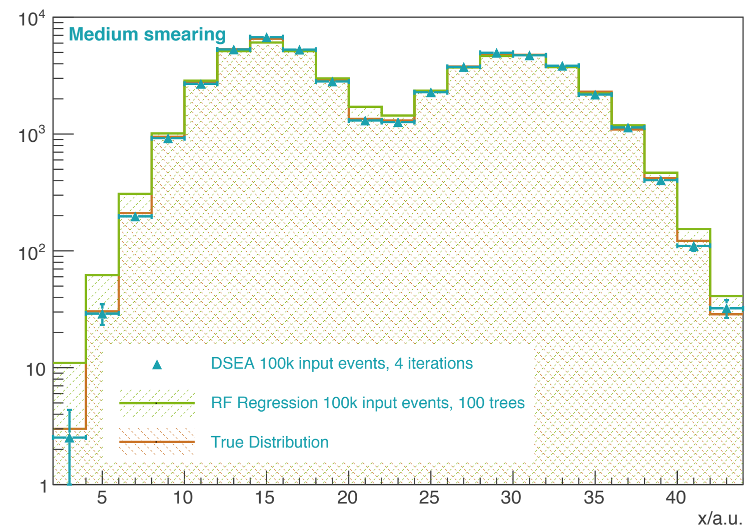



# Performance – Spectral Reconstruction

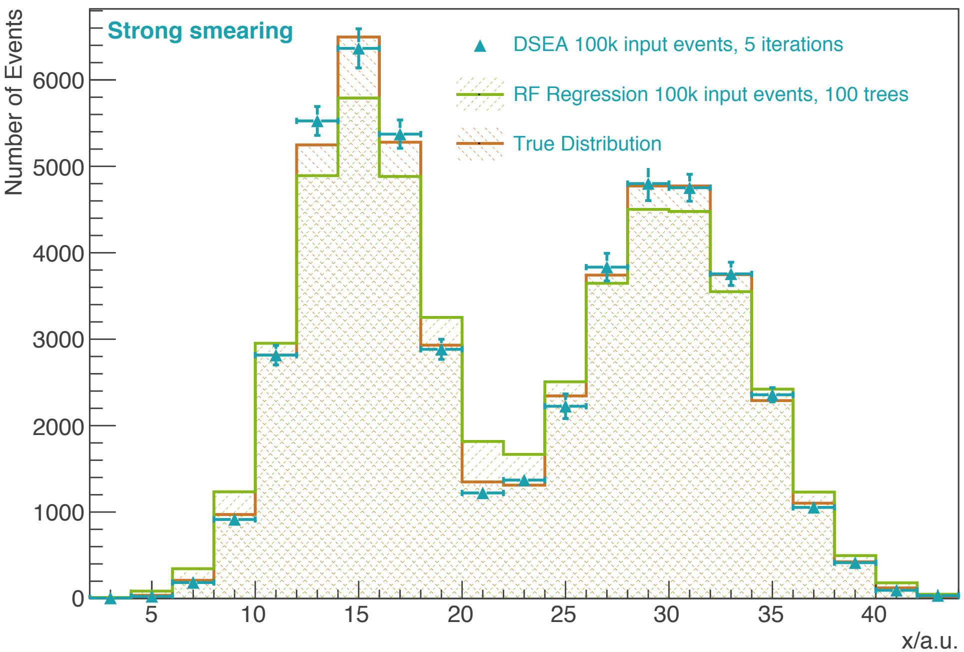
Spectral Reconstruction for strong medium using DSEA (blue), compared to a Random Forest regression (green).

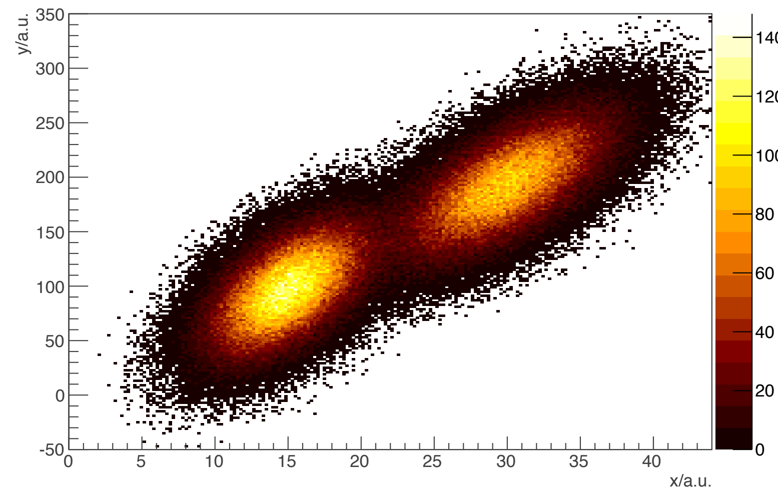


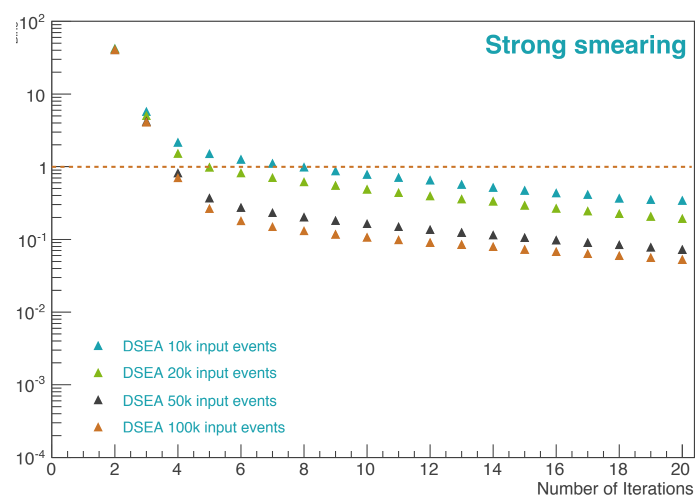



# The Dortmund Spectrum Estimation Algorithm (DSEA)

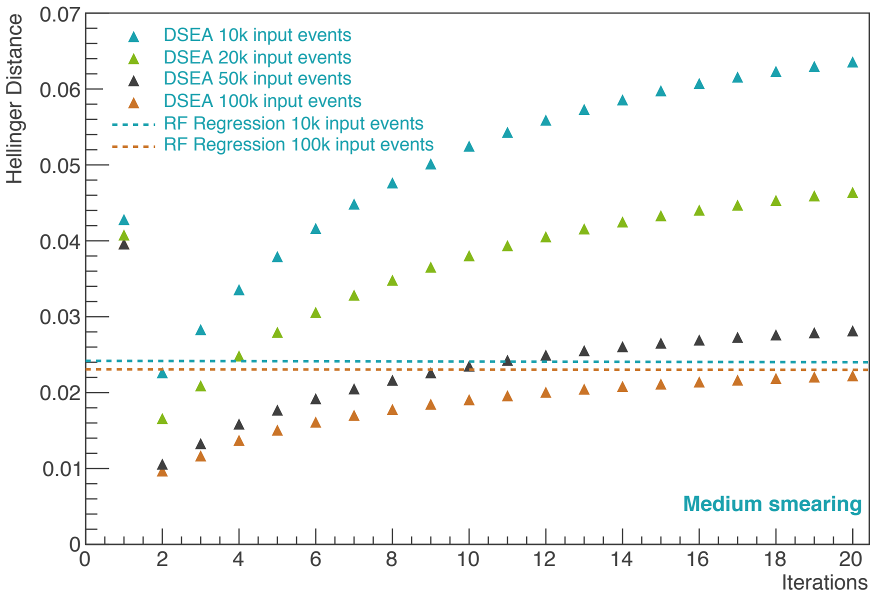


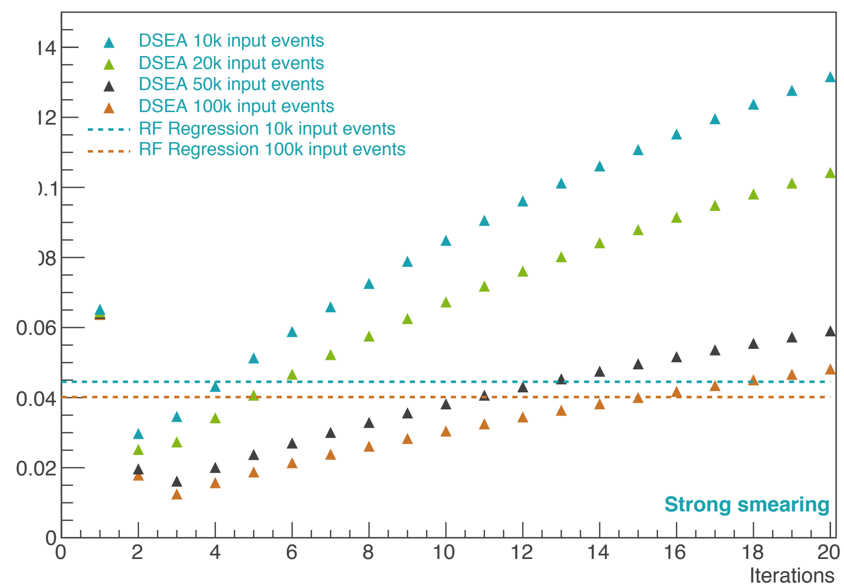




T. Ruhe et al., Proc. of the ADASS XXVI (2016)

# Performance – Agreement with Underlying Distribution




Hellinger distance for strong an medium smearing.

# More Information is Good for You!