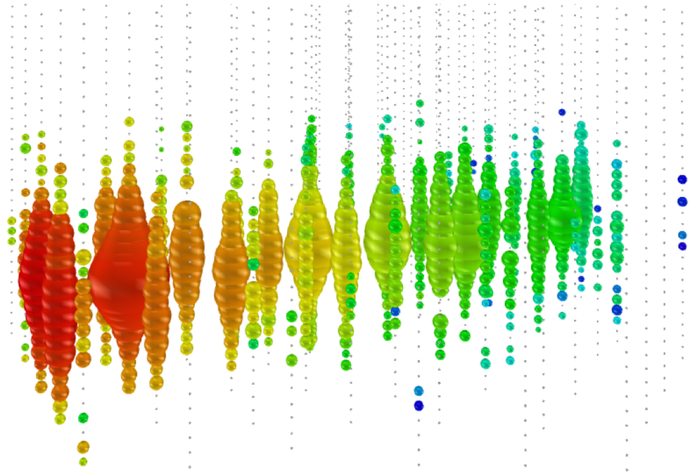
technische universität dortmund

lehrstuhl physik e5

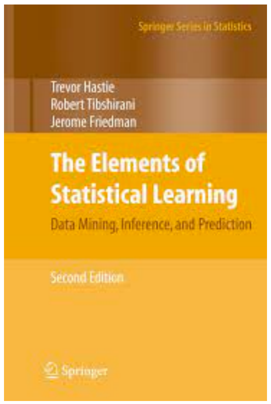


Data Science on Ice

Machine Learning Tools for the IceCube Neutrino Telescope

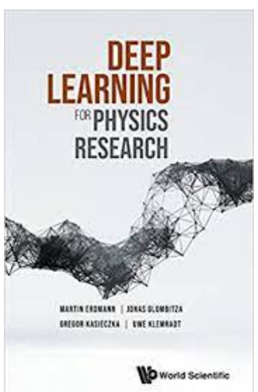
Tim Ruhe, TU Dortmund University
tim.ruhe@tu-dortmund.de

Where to Find Help



- General Introduction to Statistical Learning
- Good start to get an overview
- A lot of extra material: https://hastie.su.domains/ElemStatLearn/
- (I believe you can also download the pdf there....)
- Mathematical and statistical foundations of machine learning

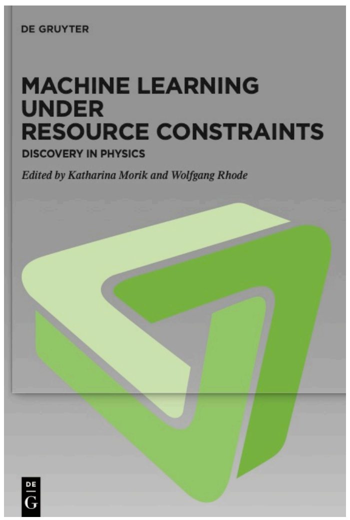
Source: https://hastie.su.domains/ElemStatLearn/



- Focus on Deep Learning and Neural Networks
- Nice pedagogic approach
- Relatively expensive (by no fault of the authors)
- Focus on physics application!

Source: amazon

Where to Find Help



- Focus on astroparticle and particle physics
- Contains a lot of topics also covered in this talk
- Open access
- published by the end of 2022
- Open Access!!!
- https://www.degruyter.com/serial/mlrc-b/html?lang=en
- Many physics examples (CTA, IceCube, FACT, LHCb, ATLAS)

Source: Von The scikit-learn developers - github.com/scikit-learn/scikit-learn/blob/master/doc/logos/scikit-learn-logo.svg, BSD, https://commons.wikimedia.org/w/index.php?curid=71445288

THE ICECUBE COLLABORATION

AUSTRALIA
University of Adelaide

BELGIUM
Université libre de Bruxelles
Universiteit Gent
Vrije Universiteit Brussel

CANADA
SNOLAB
University of Alberta–Edmonton

DENMARK
University of Copenhagen

GERMANY
Deutsches Elektronen-Synchrotron
ECAP, Universität Erlangen-Nürnberg
Humboldt–Universität zu Berlin
Ruhr-Universität Bochum
RWTH Aachen University
Technische Universität Dortmund
Technische Universität München
Universität Mainz
Universität Wuppertal
Westfälische Wilhelms-Universität Münster

JAPAN
Chiba University

NEW ZEALAND
University of Canterbury

REPUBLIC OF KOREA
Sungkyunkwan University

SWEDEN
Stockholms universitet
Uppsala universitet

SWITZERLAND
Université de Genève

UNITED KINGDOM
University of Oxford

UNITED STATES
Clark Atlanta University
Drexel University
Georgia Institute of Technology
Lawrence Berkeley National Lab
Marquette University
Massachusetts Institute of Technology
Michigan State University
Ohio State University
Pennsylvania State University
South Dakota School of Mines and Technology
Southern University and A&M College
Stony Brook University
University of Alabama
University of Alaska Anchorage
University of California, Berkeley
University of California, Irvine
University of California, Los Angeles
University of Delaware
University of Kansas
University of Maryland
University of Rochester
University of Texas at Arlington
University of Wisconsin–Madison
University of Wisconsin–River Falls
Yale University

FUNDING AGENCIES

Fonds de la Recherche Scientifique (FRS-FNRS)
Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO-Vlaanderen)
Federal Ministry of Education and Research (BMBF)
German Research Foundation (DFG)
Deutsches Elektronen-Synchrotron (DESY)
Japan Society for the Promotion of Science (JSPS)
Knut and Alice Wallenberg Foundation
Swedish Polar Research Secretariat
The Swedish Research Council (VR)
University of Wisconsin Alumni Research Foundation (WARF)
US National Science Foundation (NSF)

ICECUBE
SOUTH POLE NEUTRINO OBSERVATORY
icecube.wisc.edu

Outline of the Lecture

- Take away messages
- Very quick motivation for astroparticle physics
- Brief introduction to IceCube
- Applications of Data Science Techniques
- Deep Learning Applications

Understand your Input

Data → Model → Predictions

In machine learning, one uses data to build a model, which will then generate an output (generally some sort of prediction)

If the data are biased, then there‘s a good chance that the model and also the output will be biased!

Hiring models that disfavour e.g. women are the result of input data that disfavour women.

Validate!

Data set 1

Data set 2

Data set 3

One big data set!!!
(data set 4)

If data sets 1 to 3 have overlappping examples, any classifier, build using data set 4, will appear much more accurate than it actually is!

The same might be true if you do not cross validate!

Understand your Output

During the COVID-19 pandemic a lot of hope was put on AI tools to assist doctors to diagnose and treat patients. But none of them succeeded. Instead

- AI models learned to distinguish kids from adults based on chest scans
- Identified serious cases of COVID based on the font a hospital used to label the chest scans
- ...

This is where your scientic expertise as a physicist (your expert knowledge) becomes incredibly valuable.

Machine Learning is a Tool

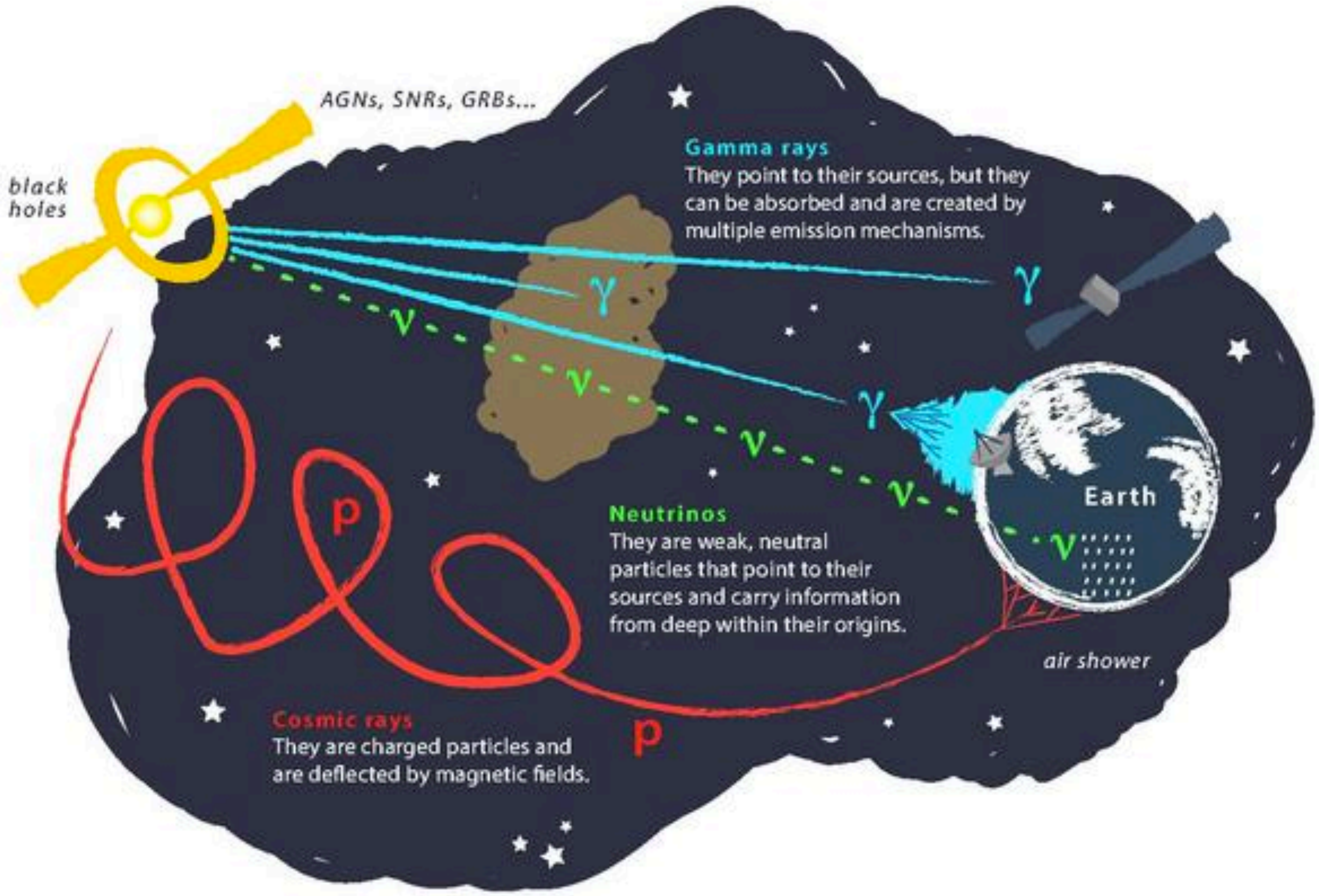
Von Banffy - Eigenes Werk, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=11657709

Source: CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=577886

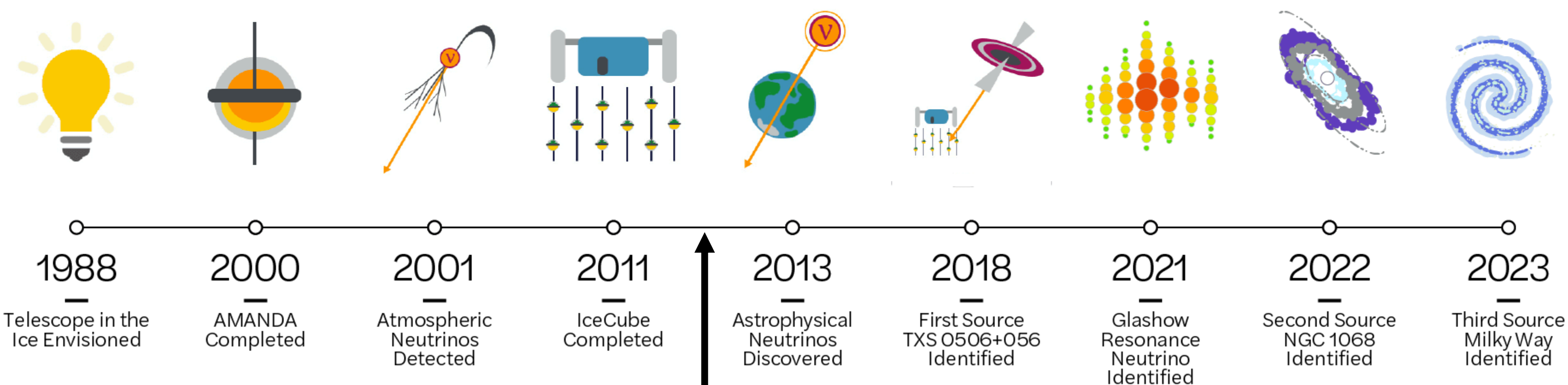
Machine Learning provides tools to accomplish an analysis task faster and more accurately (when used correctly).

Source: Von smial (talk) - Eigenes Werk, FAL, https://commons.wikimedia.org/w/index.php?curid=6028669

Astroparticle Physics and Neutrino Astronomy

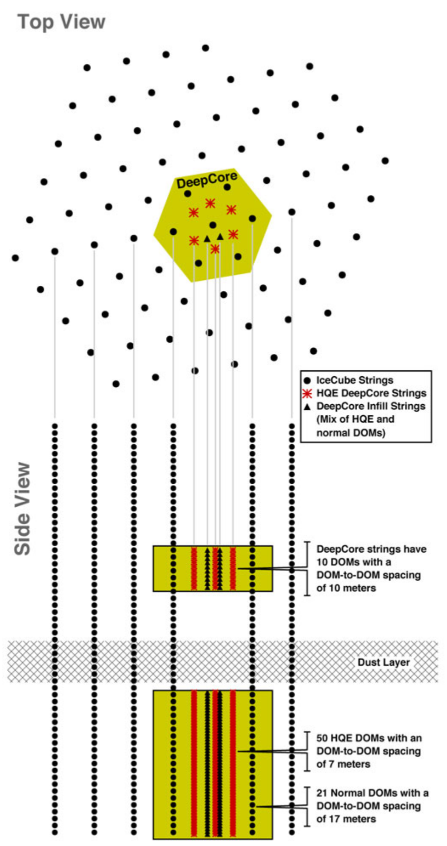


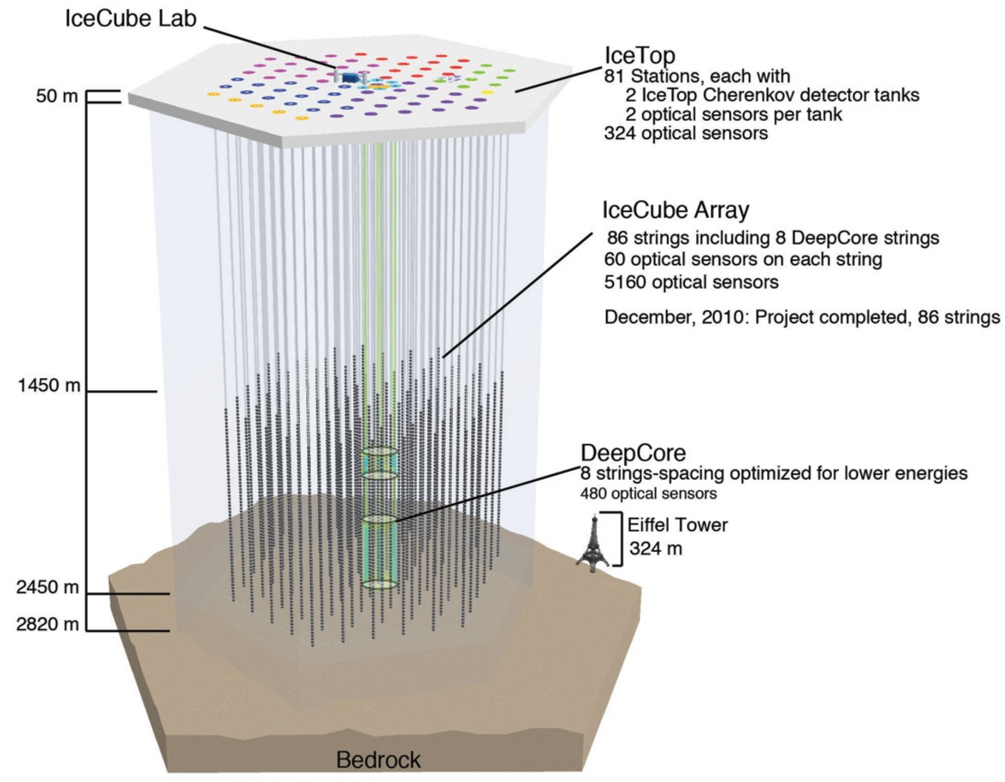
A History of Neutrino Astronomy in Antarctica



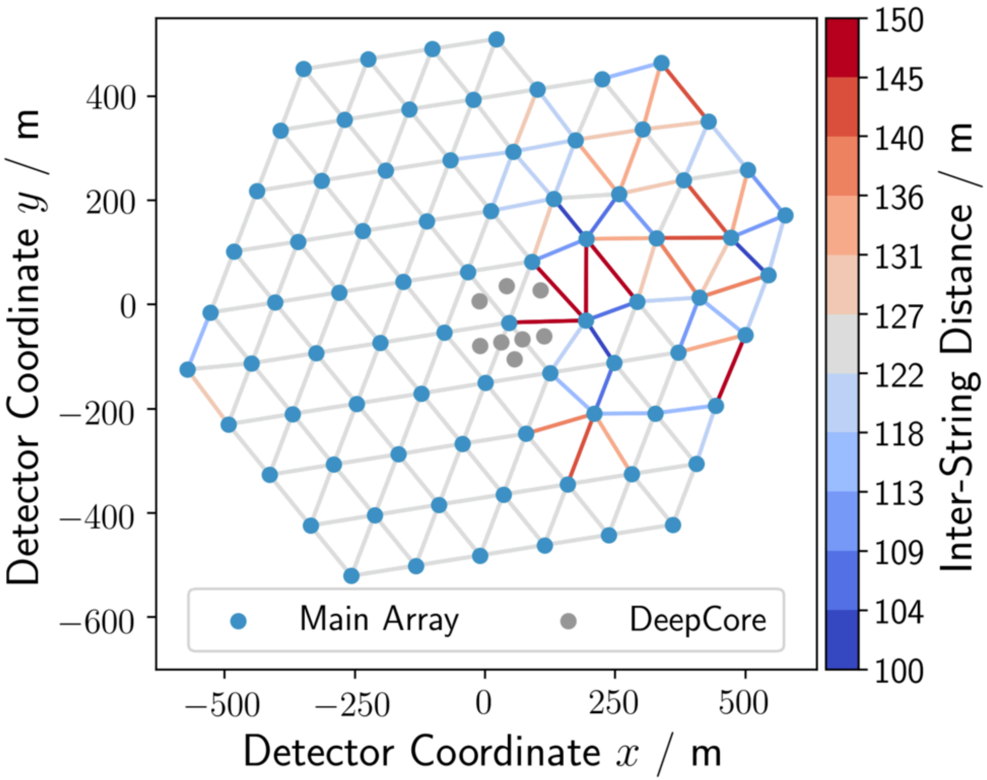
2012: People getting itchy about missing astrophysical neutrinos

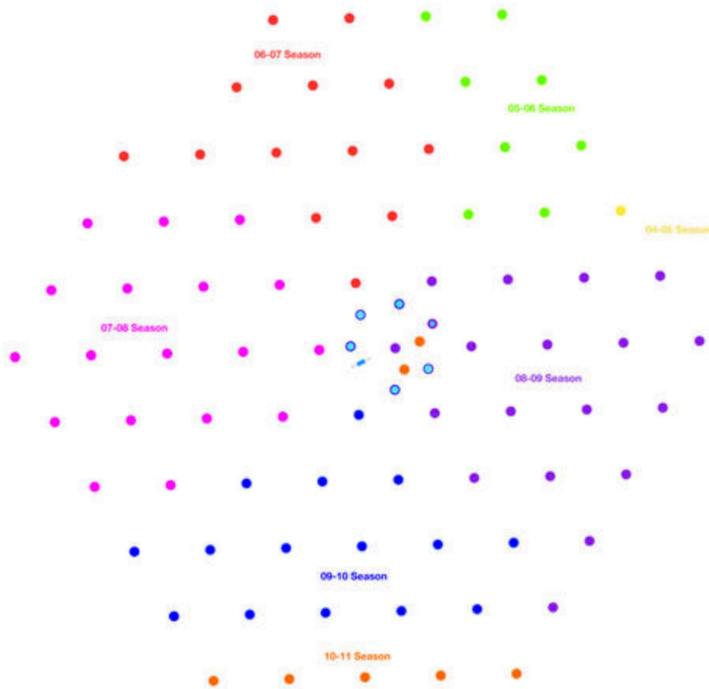
The IceCube Neutrino Observatory



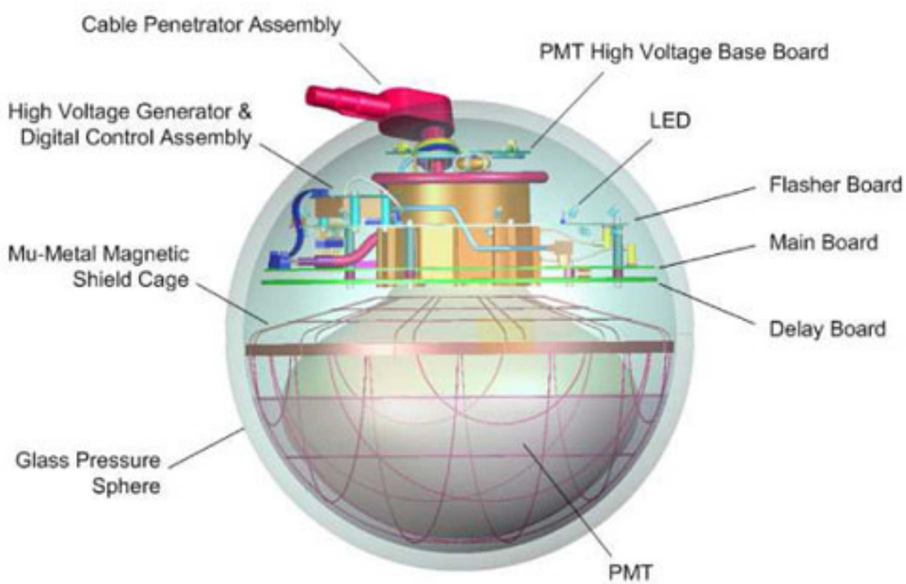


IceCube Geometry and Drill Seasons



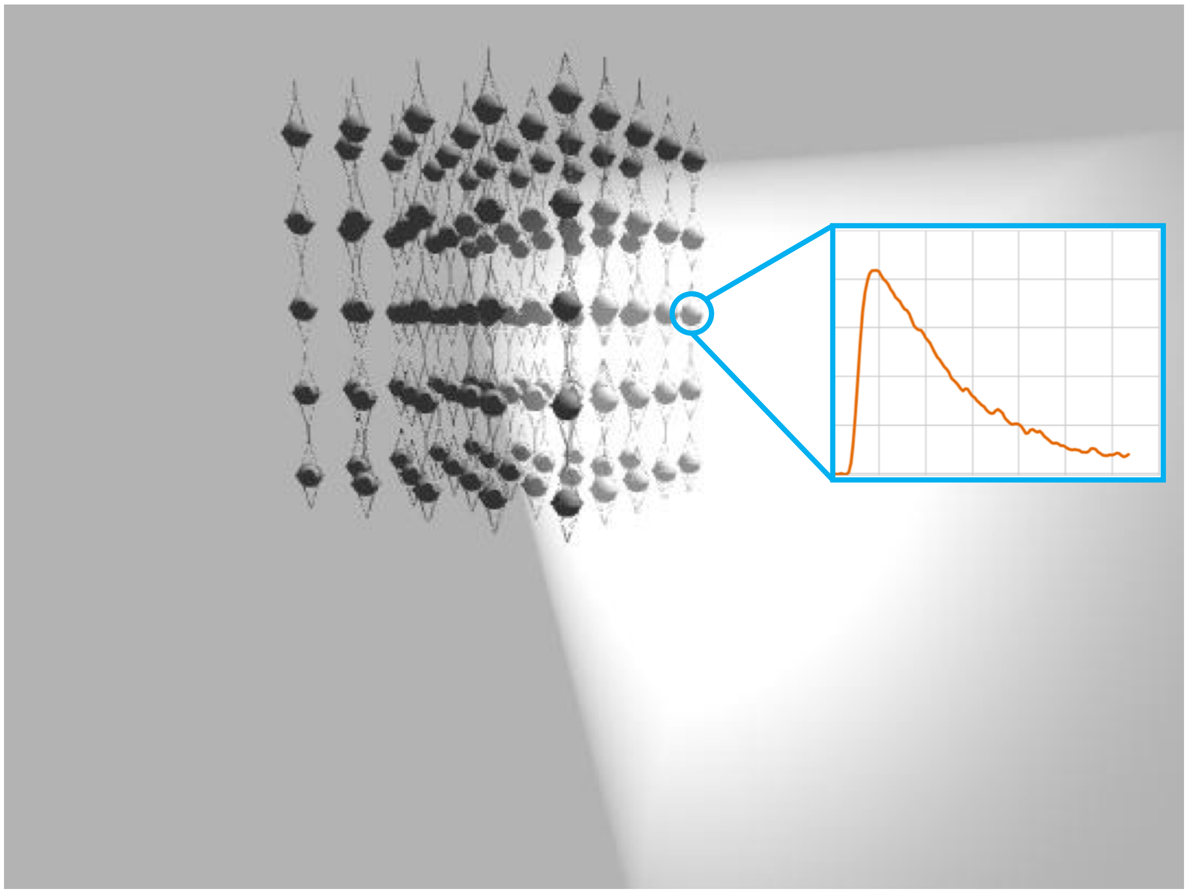


The Fundamental Unit of IceCube: The DOM

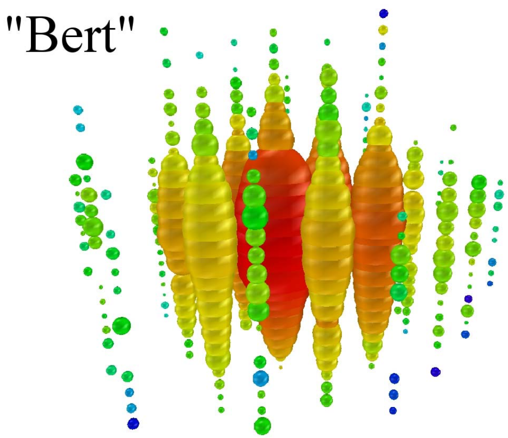


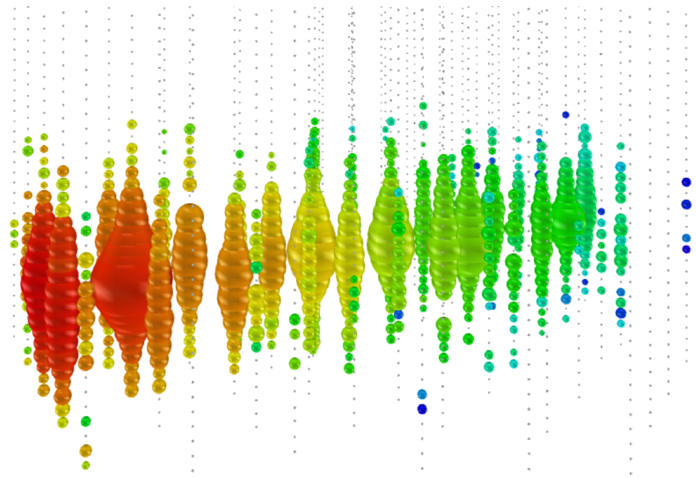
- Downward facing 10“ PMT (Hammamatsu R7081-02), 25% Peak QE
- High Voltage Supply
- Electronics
- Flasher LEDs
- Higher QE (34%) for DeepCore DOMs (Hammamatsu R7081MOD)
- Very few DOM failures (mostly during deployment)
- Slightly larger fraction of DOMs with *issues* (mostly non-standard Local Coincidence)

The Detection Principle: Cherenkov Light



IceCube Events: Tracks and Cascades





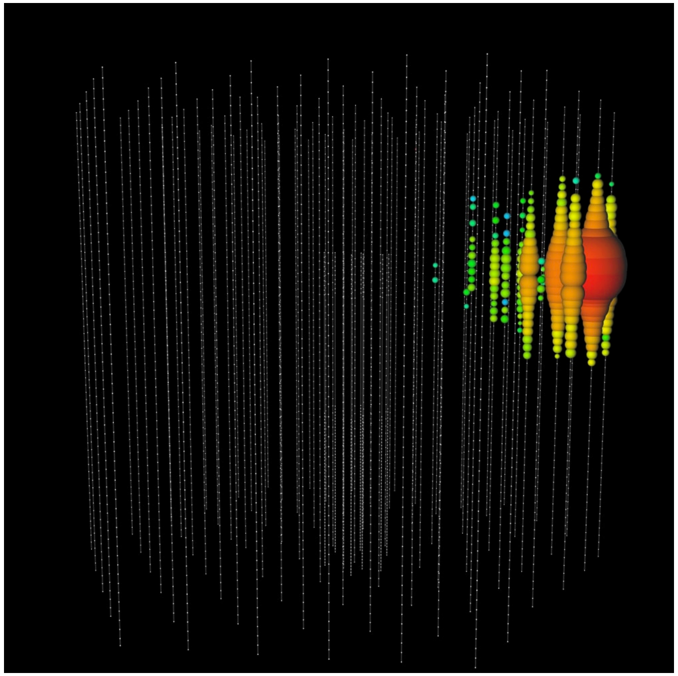
Cascade like events:

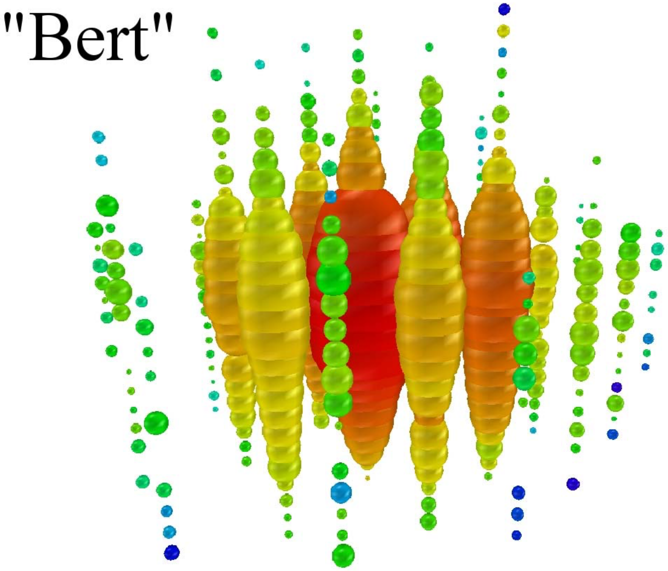
- ν_e - CC and all flavour NC interactions
- Interaction inside instrumented volume
- Poor angular resolution $\approx 15°$
- Good energy resolution

Track like events:

- ν_μ - CC interactions
- Interaction may happen outside instrumented volume
- Good angular resolution $\approx 1°$
- Poor energy resolution

Fully and Partially Contained Cascades

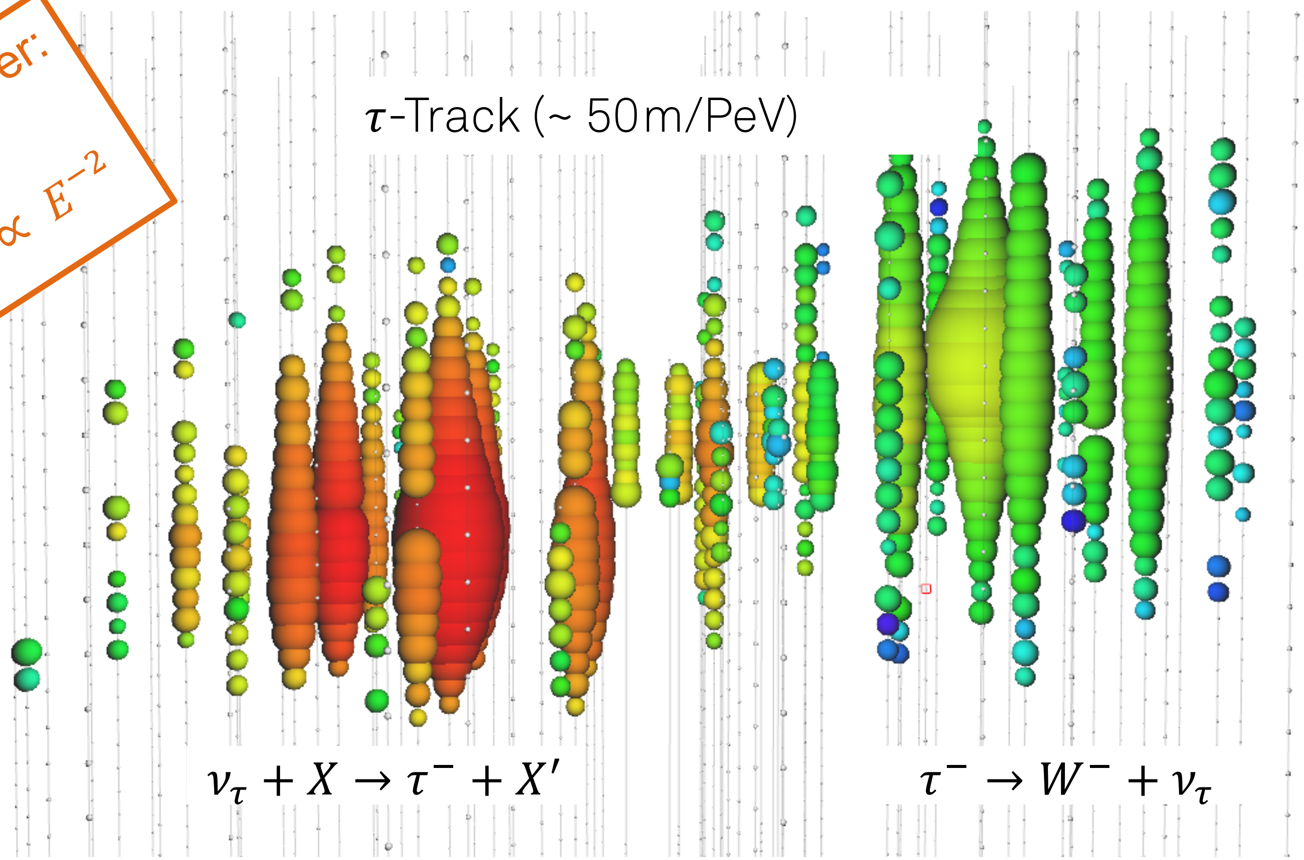




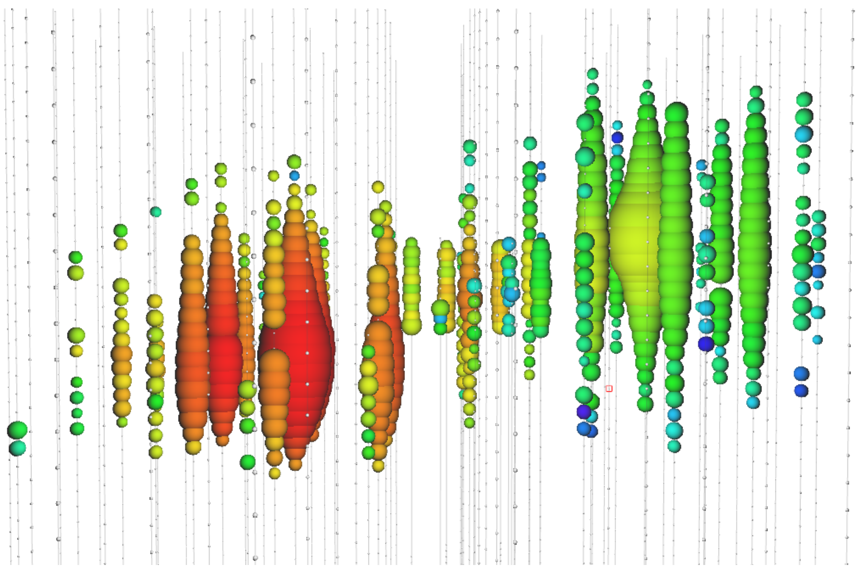
IceCube Events: Double Cascades

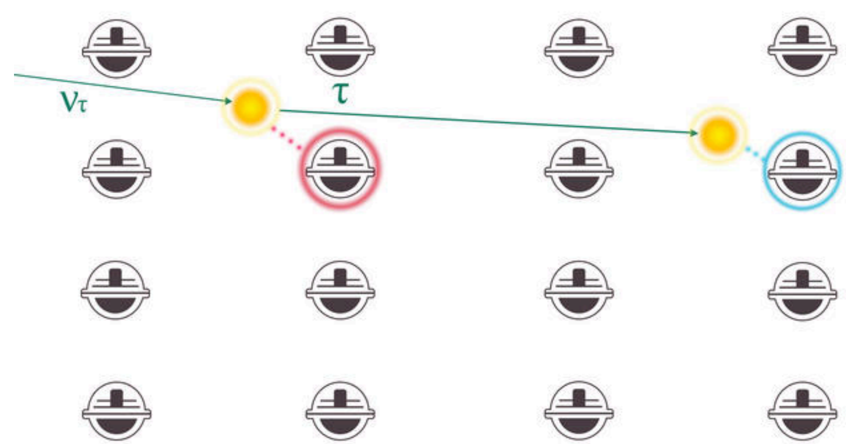
Remember:

$$\frac{d\Phi}{dE} \propto E^{-2}$$

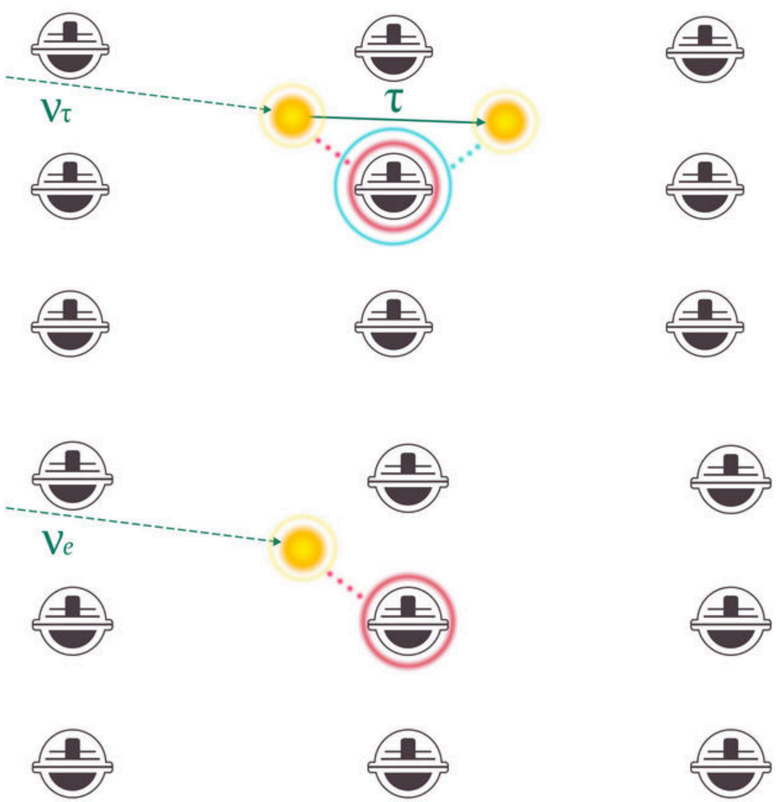


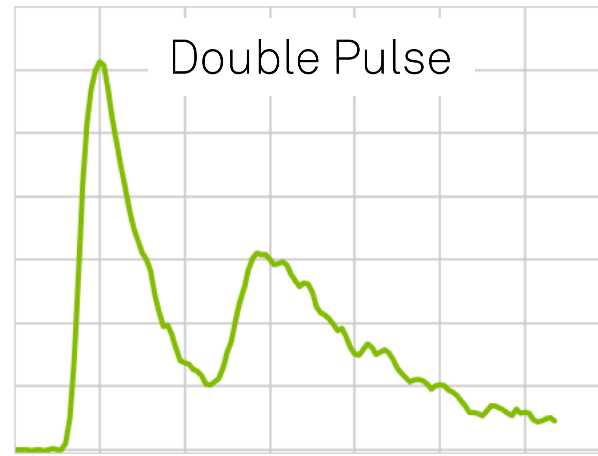
From Double Cascades to Double Pulses

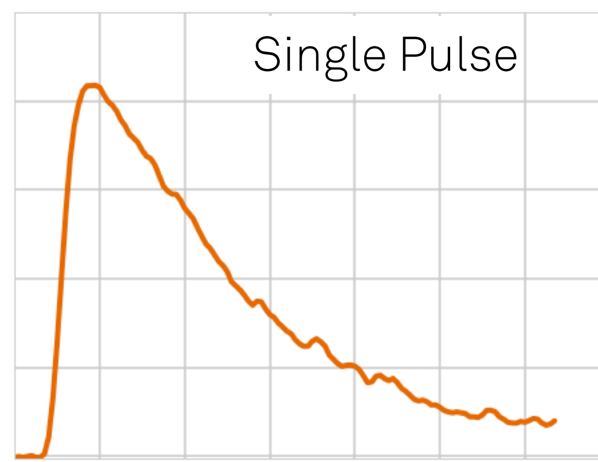




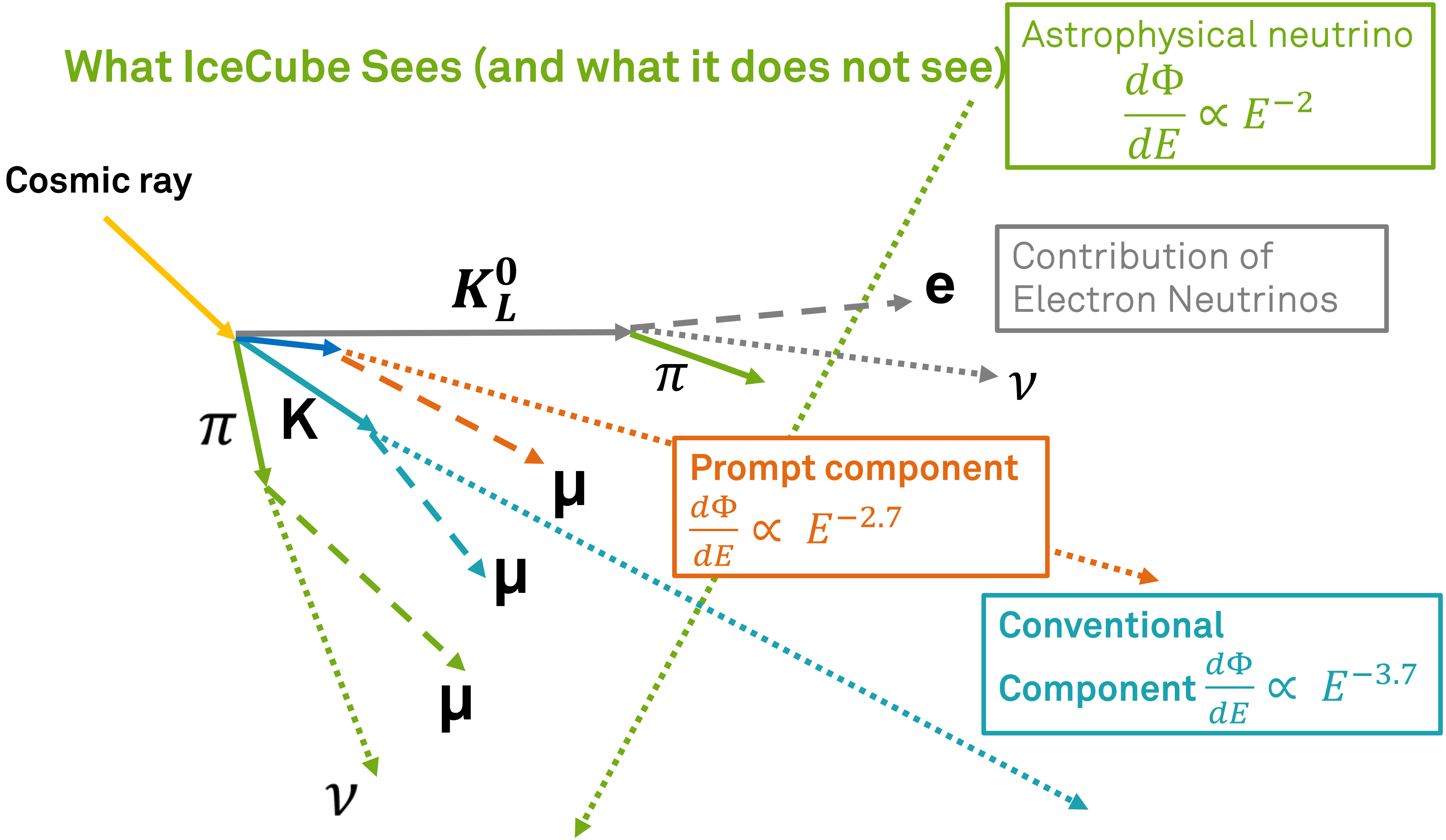
From Double Cascades to Double Pulses







What IceCube Sees (and what it does not see)

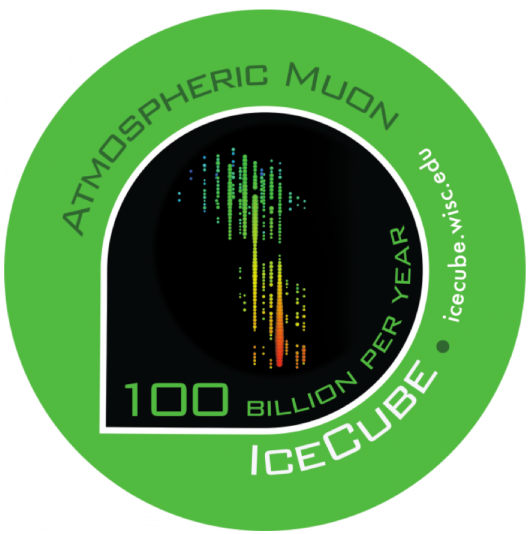


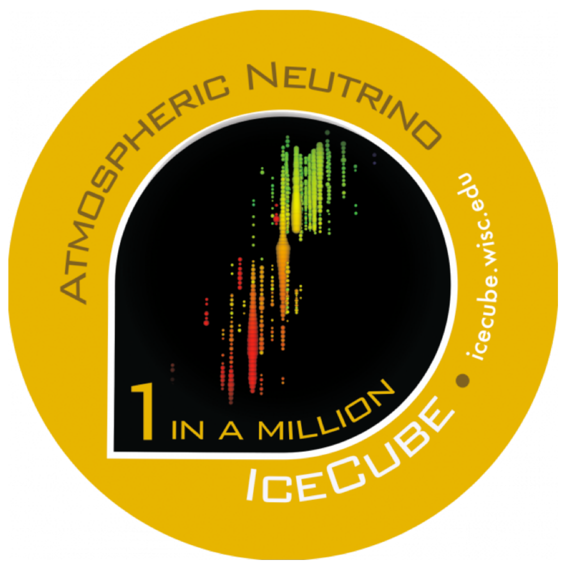
What IceCube Sees

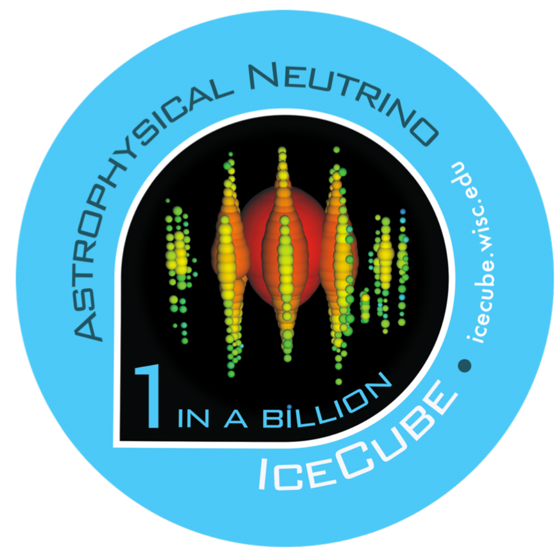
Atmospheric Muons
$$\frac{d\Phi}{dE} \propto E^{-3.7}$$

Conventional Component $\frac{d\Phi}{dE} \propto E^{-3.7}$

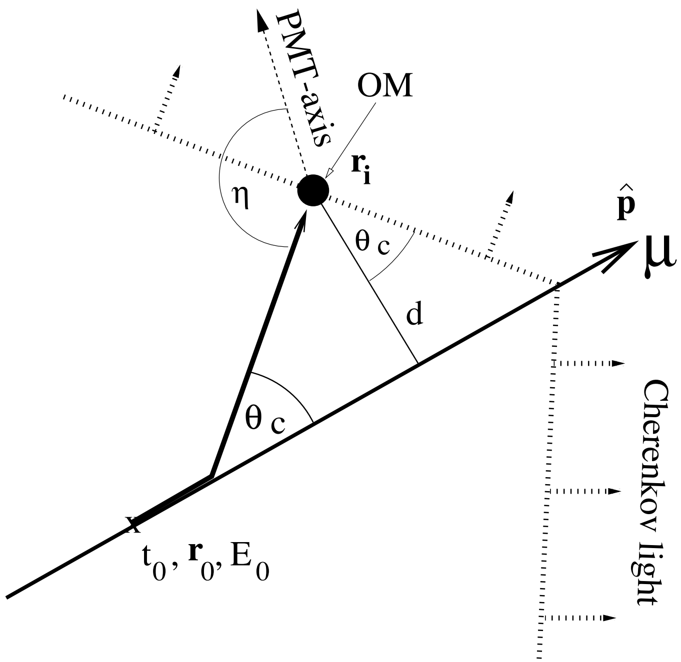
Astrophysical neutrino
$$\frac{d\Phi}{dE} \propto E^{-2}$$







IceCube Reconstructions



Ahrens et al., NIM A 524 (1-3), 169 -194 (2004)

1. Simple features, which do not require sophisticated reconstruction, e.g. total charge acquired in an event.

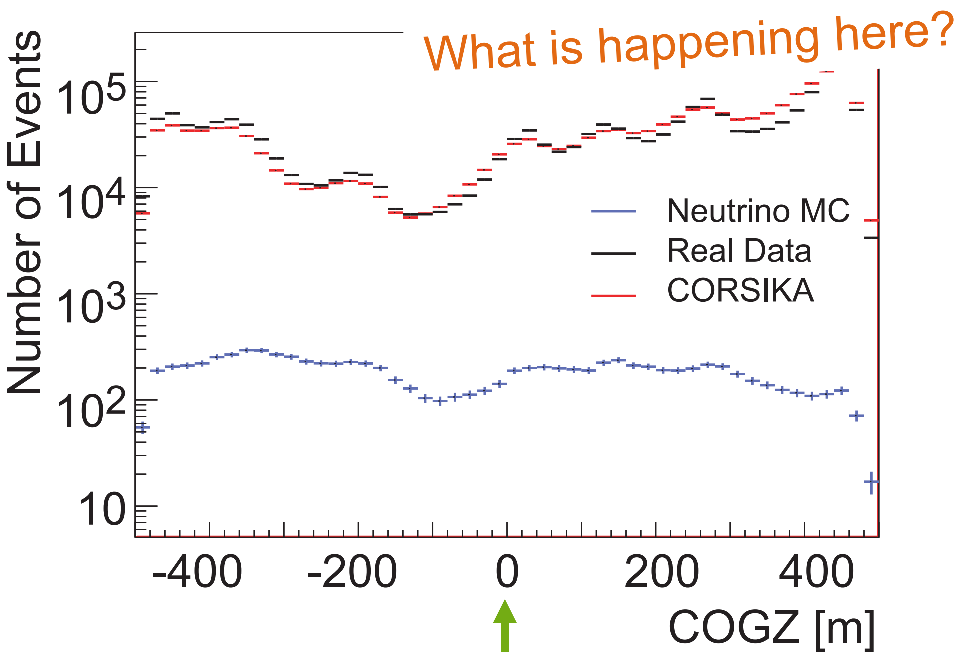
2. Simple Fits, assuming a straight line and minimizing a X2:

$$\chi^2 = \sum_{i=1}^{N} (r_i - r_{Linefit} - c_{Medium} \cdot t)^2$$

3. Likelihood-based reconstructions:

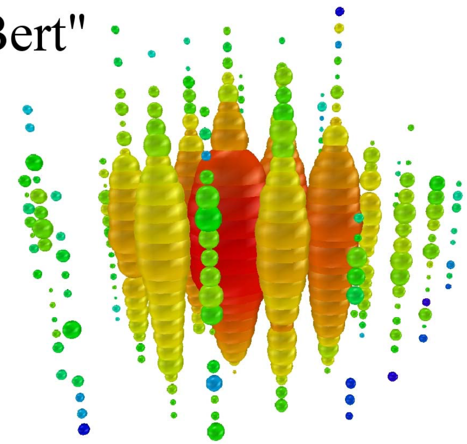
$$\mathcal{L} = \prod_i p(x_i|\vec{p})$$

COGX, -Y and -Z



Center of the detector.

"Bert"



Center of Gravity of the charge distribution in the detector.

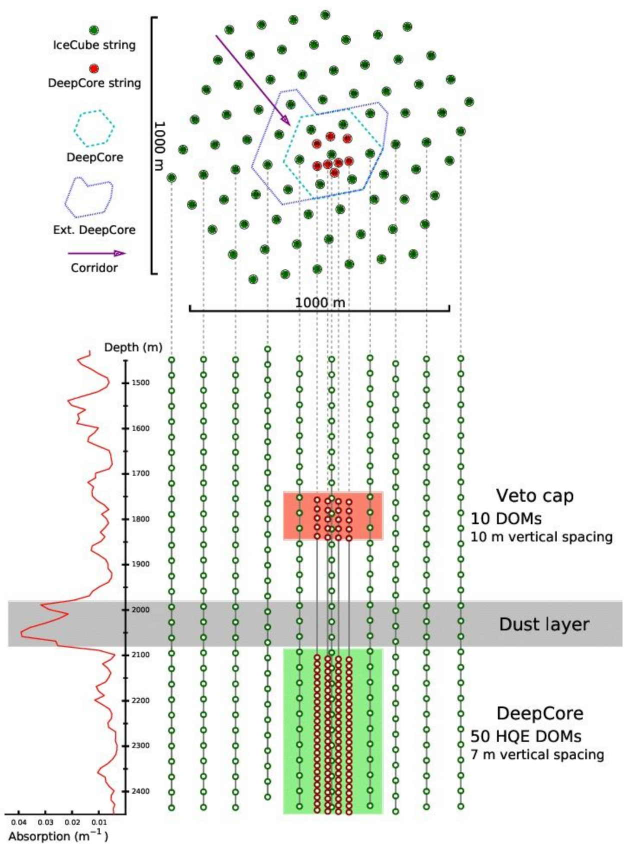
For cascades this is a relatively good estimator for the vertex position of the neutrino interacation.

Can be used e.g. for containment cuts.

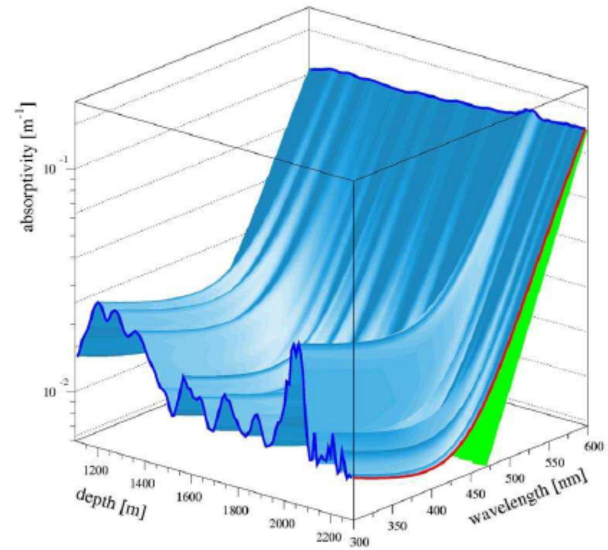
Interpretation is less intuitive for tracks

South Pole Ice as a Detection Medium

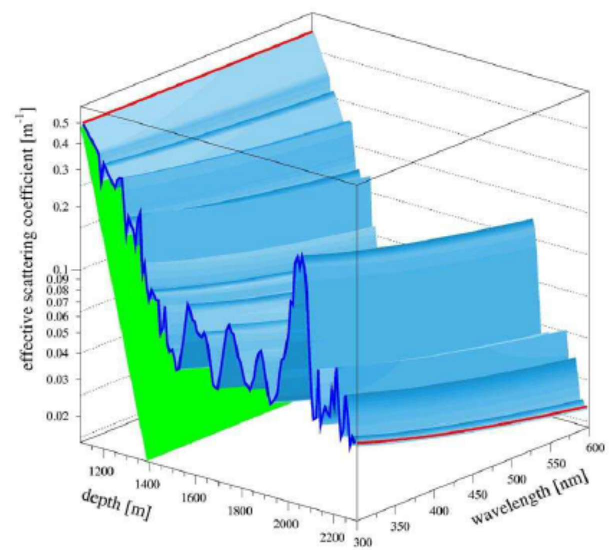
Graphics: Aartsen et al., *PRD* 99.3 (2019): 032007.



Absoprtion



Scattering



Ice Model Evolution

AHA

SPICE-1

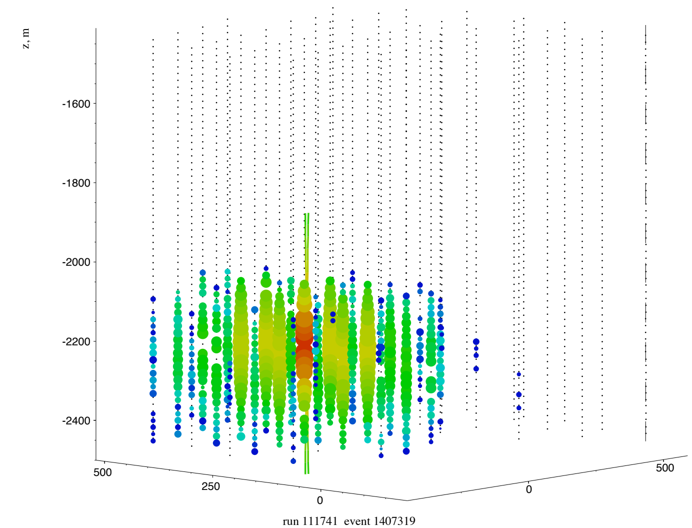
SPICE:Mie

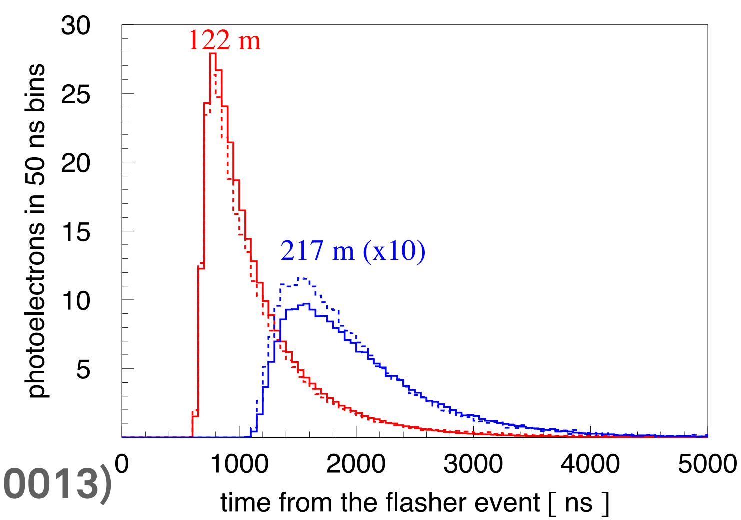
SPICE-Lea

...

Ice core measurements

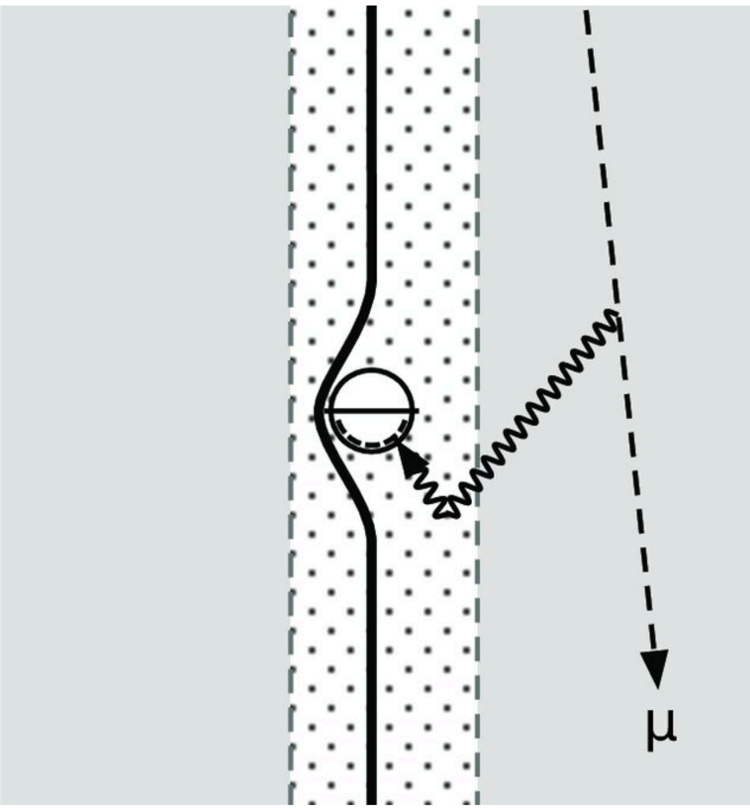
In situ measurements using flasher LEDs

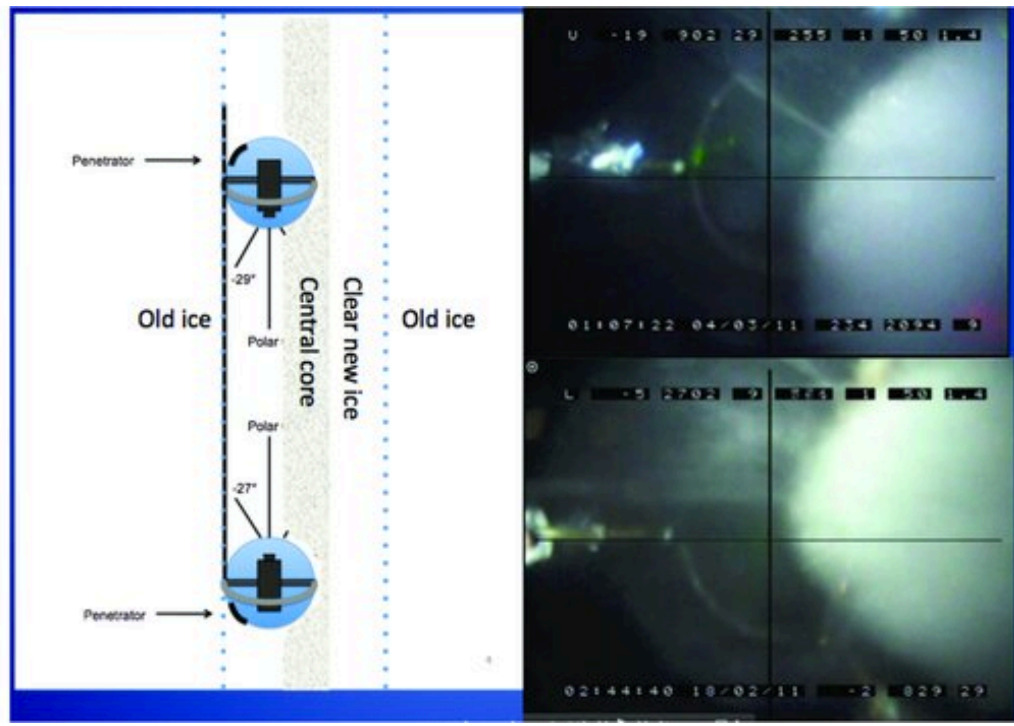




Aartsen et al., NIMA 711, 73 – 89 (20013)

Bulk Ice vs. Hole Ice





Aartsen, Mark G., et al., *Journal of Physics G,* 44.5 (2017): 054006.

IceCube Data Levels

Level 0 (trigger level)

Supposed to be very fast. The aim is to identify and extract particle interactions from noise.

Level 1 (filter level)

Level 2

Level 3

IceCube Data Levels

Level 0 (trigger level)

Level 1 (filter level)

Data rate of approx. 3 kHz. Dominated by atmospheric muons. Some degree of background rejection, mainly by selecting events with a certain topology or energy (e.g. DeepCore Filter).

Level 2

Level 3

IceCube Data Levels

Level 0 (trigger level)

Level 1 (filter level)

Level 2

No events are discarded at this level. Sophisticated reconstructions are applied.

Level 3

IceCube Data Levels

Level 0 (trigger level)

Level 1 (filter level)

Level 2

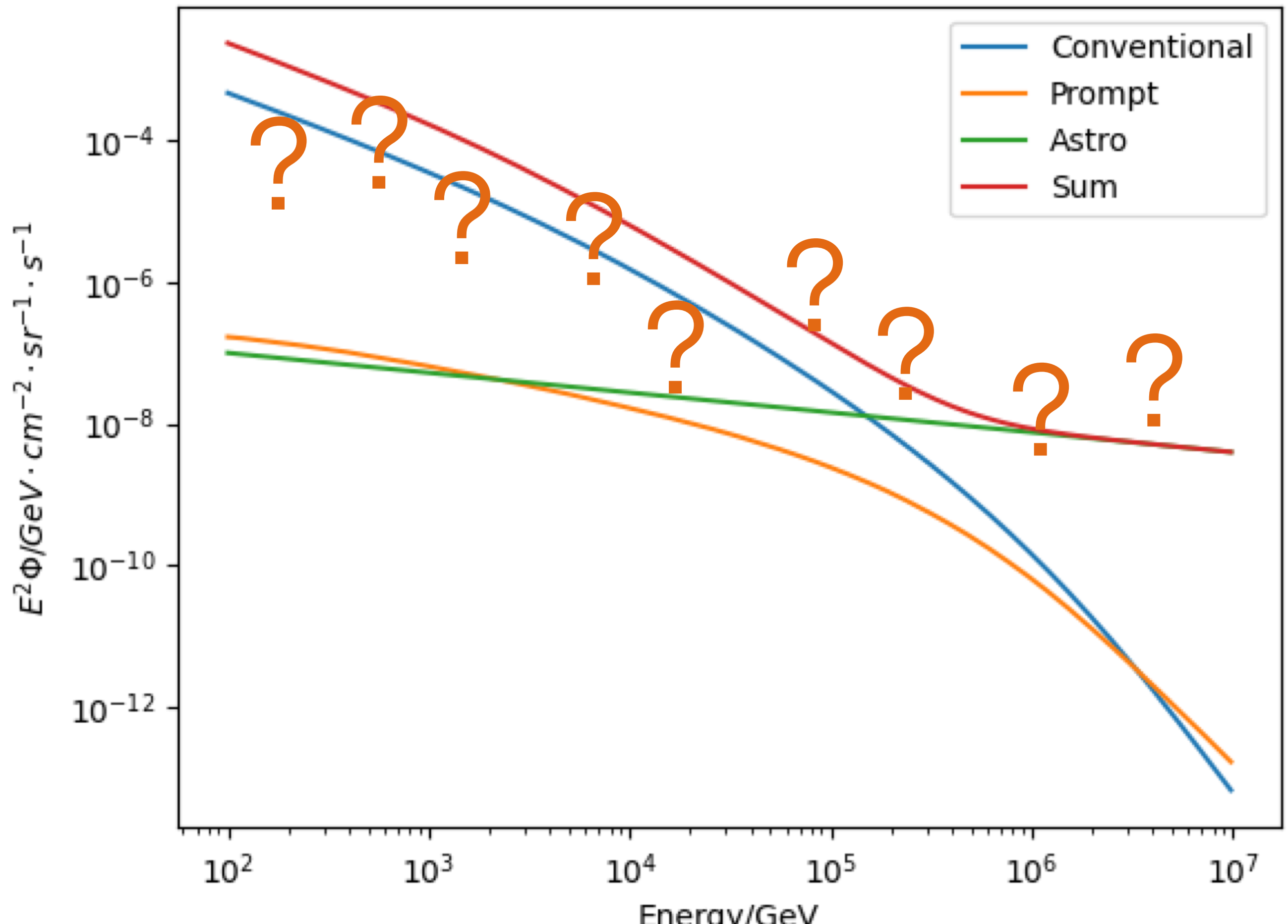
Level 3

Reconstructions are continued and become working group specific. Some cuts are applied to events that pass a subset of filters (cascades, muons, low energy).

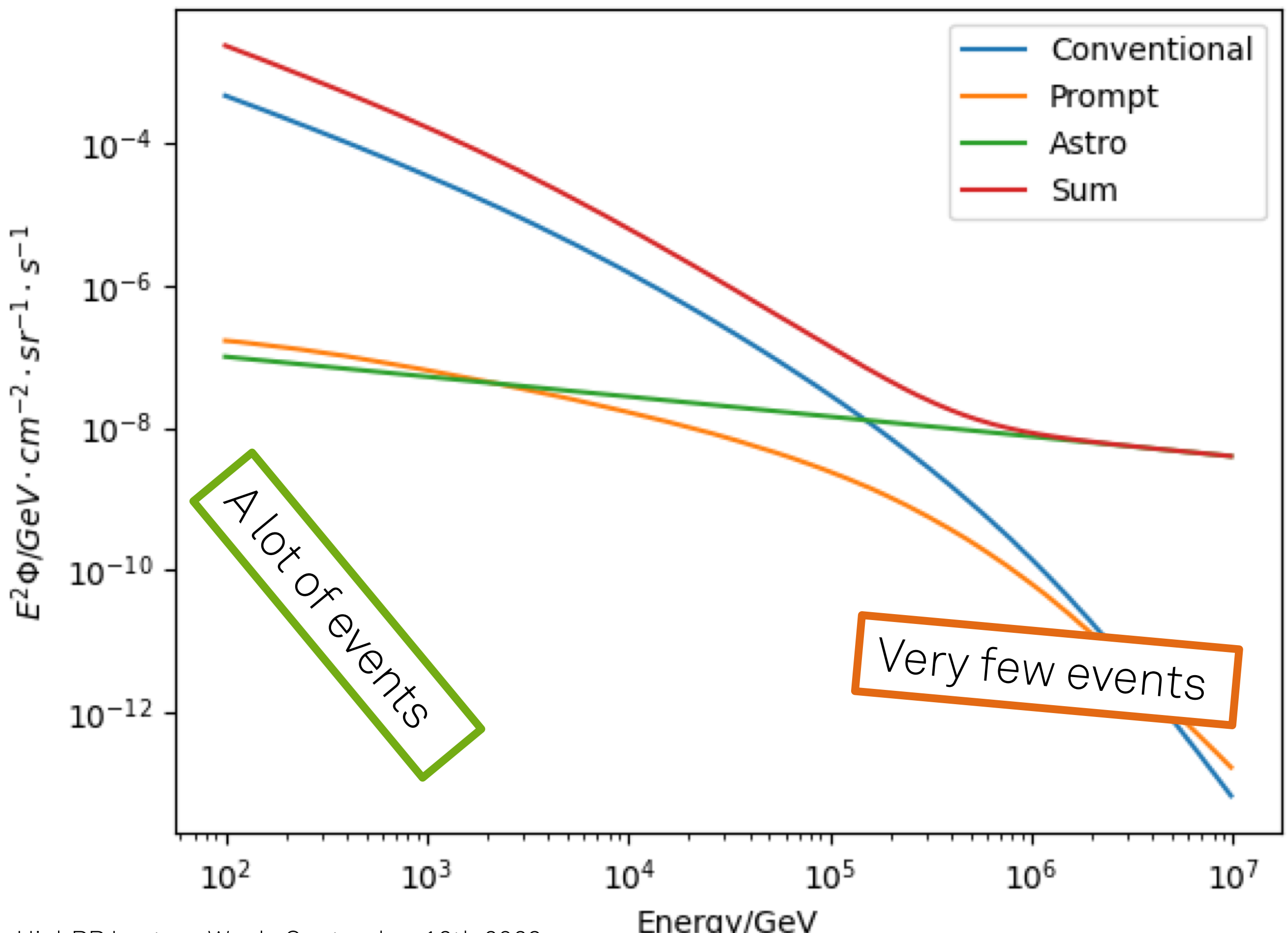
Data rate is reduced to approx. 1 Hz.

Starting point for the majority of the IceCube analyses.

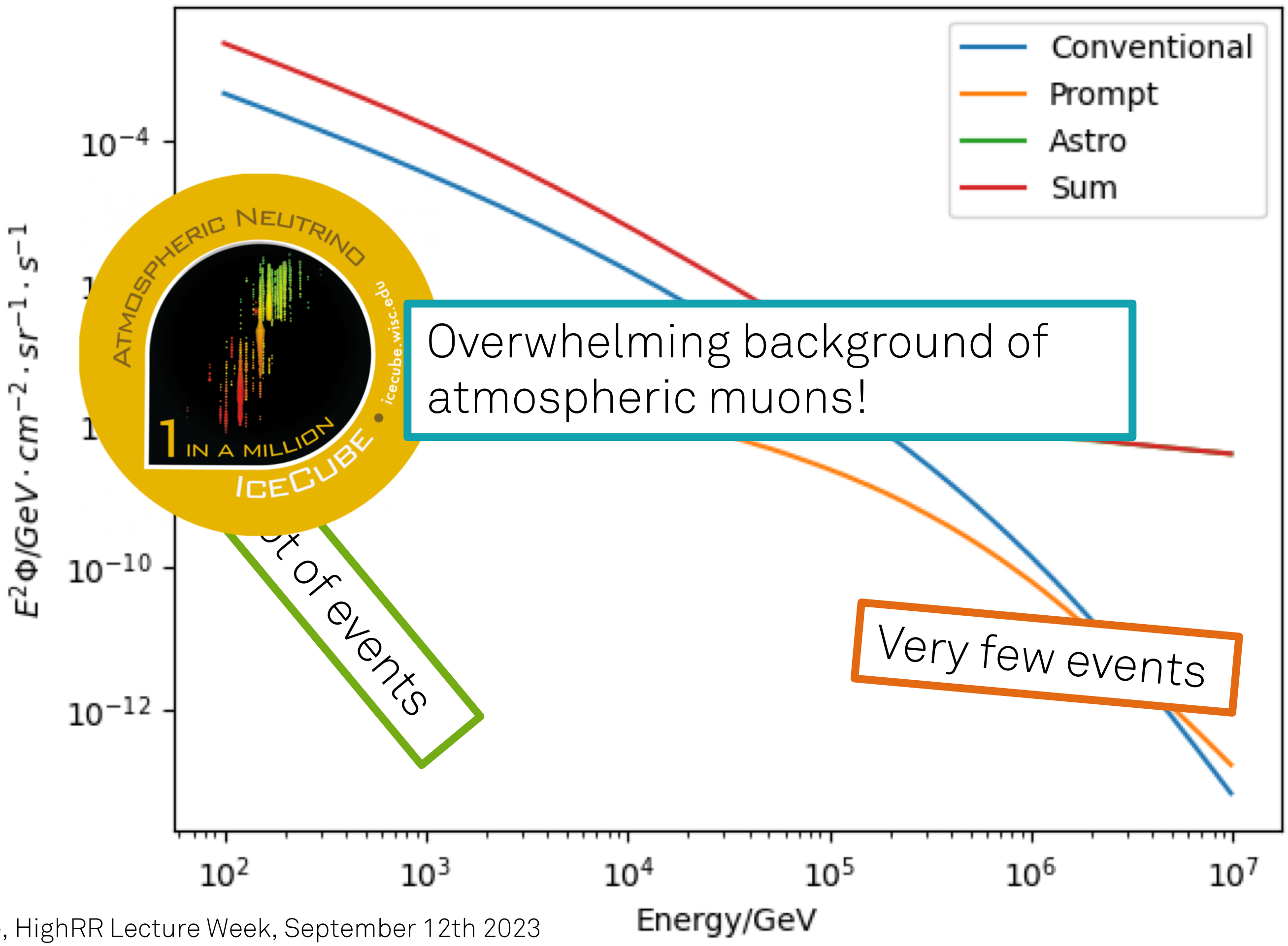
Example Analysis: Reconstruction of Neutrino Energy Spectra



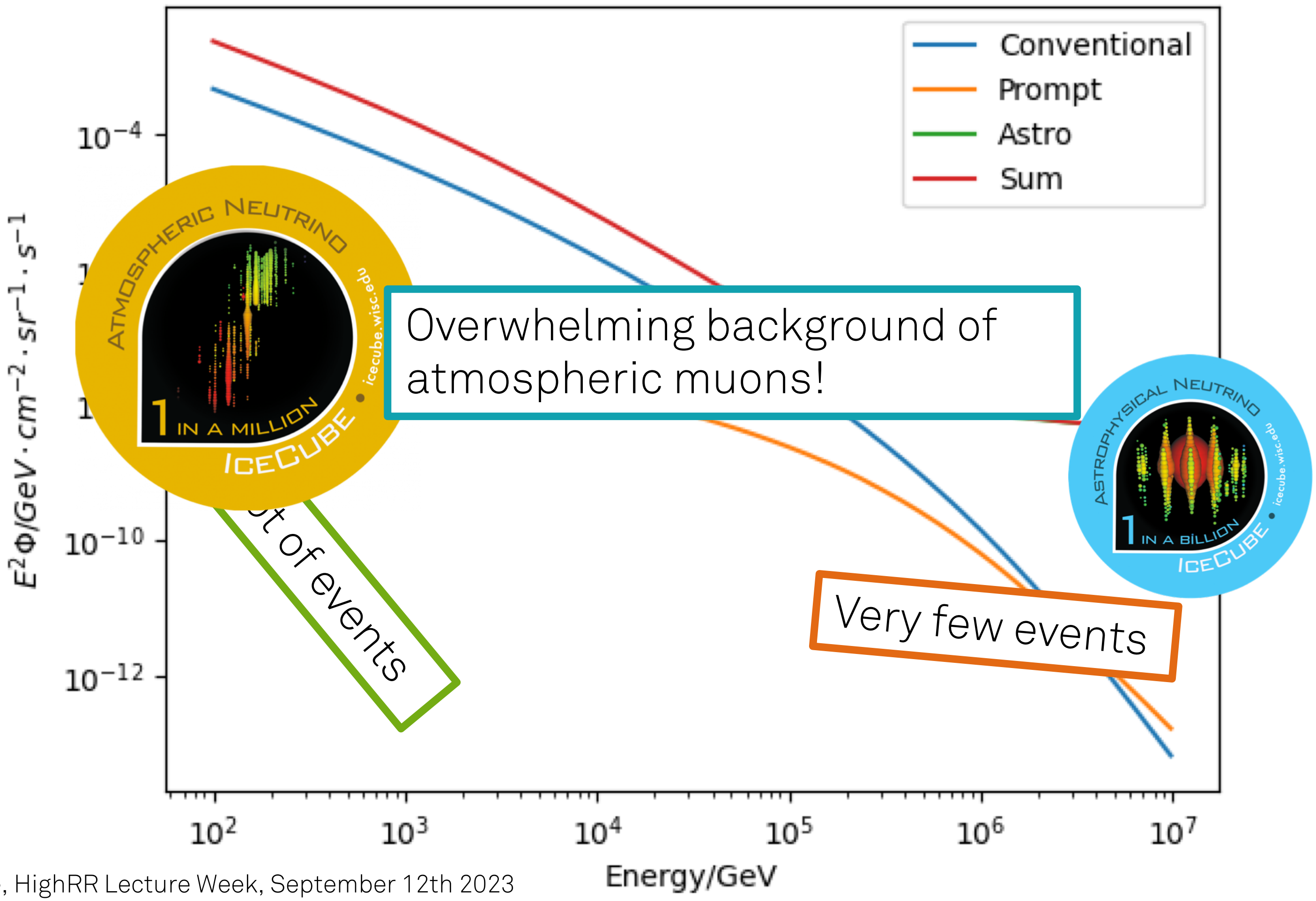
Example Analysis: Challenges



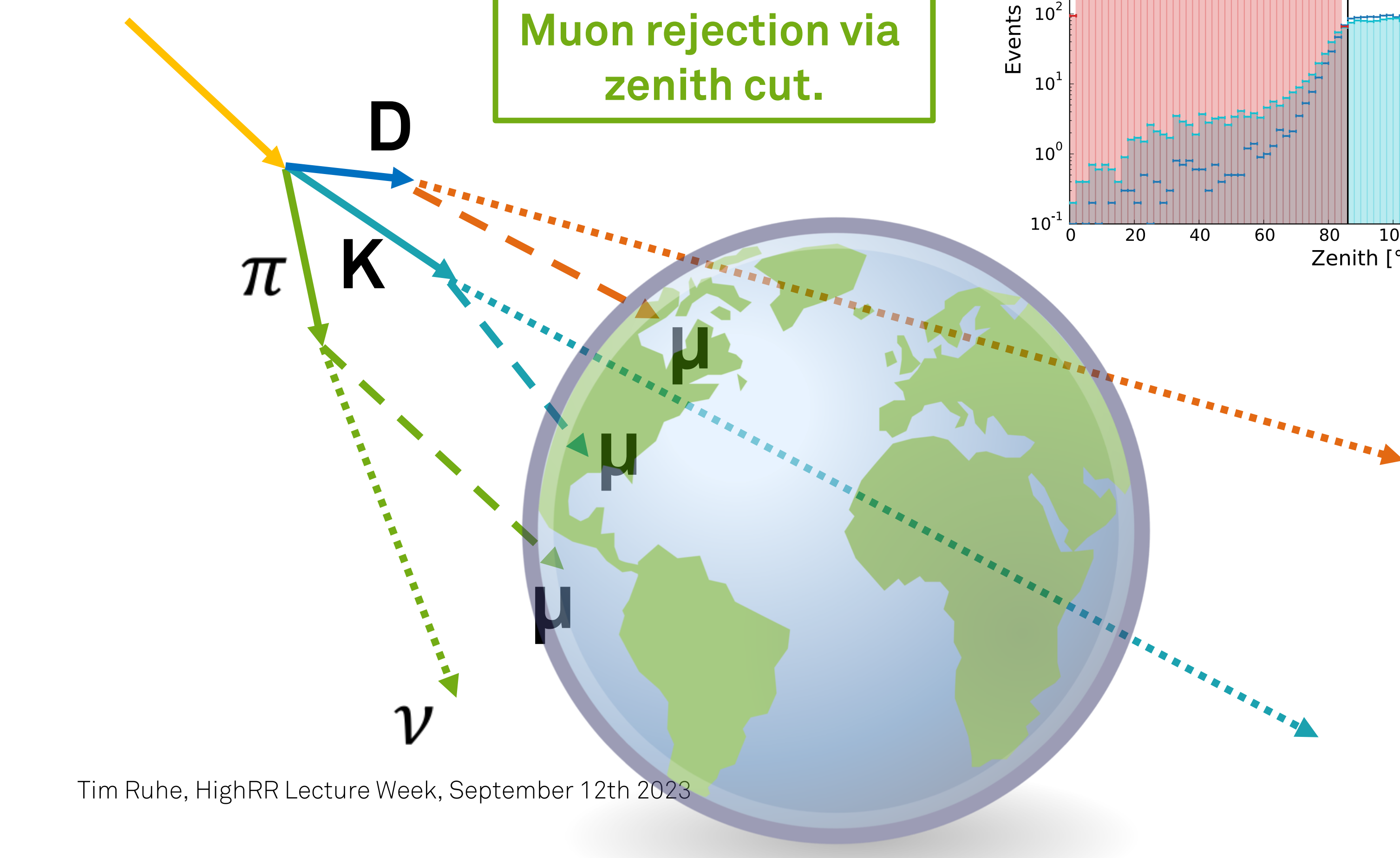
Example Analysis: Challenges



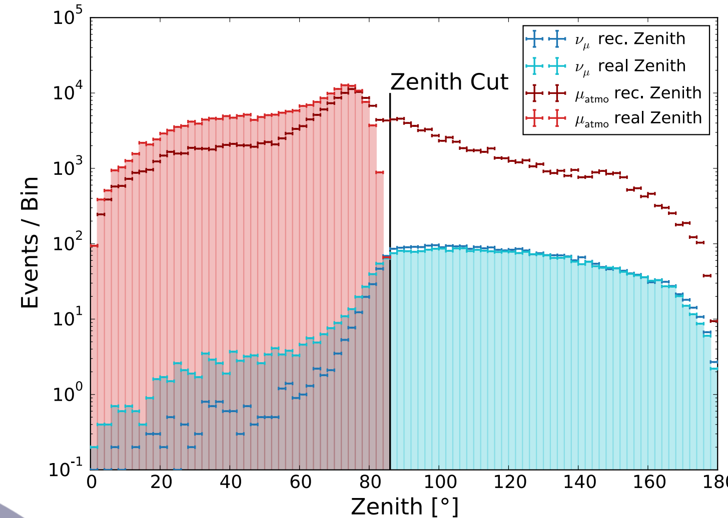
Example Analysis: Challenges



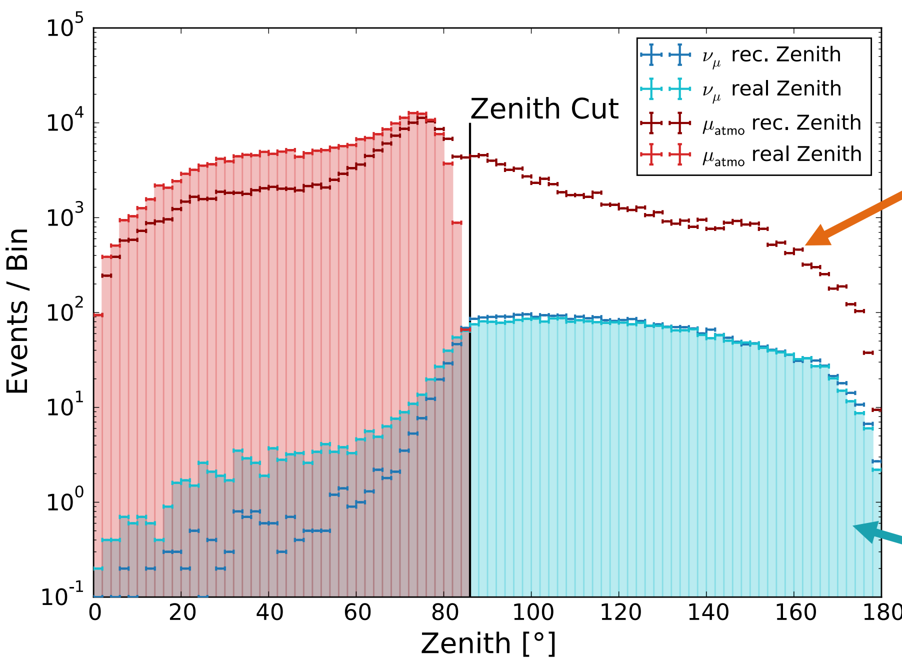
Using the Earth as Muon Shield



Muon rejection via zenith cut.



Defining Signal and Background



M. Börner, PhD thesis (2018)

Background: Muon tracks falsely reconstructed as upward going. (Encoded as 0)

Signal: Muon tracks correctly reconstructed as upward going. (Encoded as 1)

How to handle Missing Values

Name	Age	Gender	Job	Salary
Joey	23	male	actor	1000,00
Rachel	24	female	buyer	1340,00
Ross	28	male	???	1200,00
Chandler	28	male	accountant	nan
Monica	24	female	chef	1280,00
Phoebe	nan	female	???	4300,00

- Some algorithms can only handle numerical values
- Missing values (nans, infs, …) can be replaced
  - Average
  - Median
  - Constant
  - …
- Features with too many missing values can be excluded
- Missing value can actually provide valuable information, e.g. reconstruction algorithm failed because this is an event with poor information

How to handle Missing Values

Name	Age	Gender	Job	Salary
Joey	23	male	actor	1000,00
Rachel	24	female	buyer	1340,00
Ross	28	male	???	1200,00
Chandler	28	male	accountant	nan
Monica	24	female	chef	1280,00
Phoebe	nan	female	???	4300,00

Given that we understand the data fairly well, it is probably safe to replace this values with mean or median.

How to handle Missing Values

Name	Age	Gender	Job	Salary
Joey	23	male	actor	1000,00
Rachel	24	female	buyer	1340,00
Ross	28	male	???	1200,00
Chandler	28	male	accountant	nan
Monica	24	female	chef	1280,00
Phoebe	nan	female	???	4300,00

Given that we understand the data fairly well, it is probably safe to replace this values with mean or median.

This might be more problematic, because Phoebe‘s salary is an outlier here, that might create a bias.

How to handle Missing Values

Name	Age	Gender	Job	Salary
Joey	23	male	actor	1000,00
Rachel	24	female	buyer	1340,00
Ross	28	male	???	1200,00
Chandler	28	male	accountant	nan
Monica	24	female	chef	1280,00
Phoebe	nan	female	???	4300,00

Given that we understand
fairly well, it is probably s
replace this values with m
median.

This is even harder. The feature is not numerical, so it cannot be easily replaced.

One could encode it, e.g. „unknown" or one could discard the entire feature.

re problematic,
s salary is an outlier
create a bias.

Feature Selection

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

Exclude features that either bias the selection or are only present in simulation.

M. Börner, PhD thesis (2018)

Feature Selection

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

M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

Constant features do not carry information.

*Useless in this case includes features that exceed a certain threshold in the relative number of missing values.

**One could go one step further and also exclude features with small variance.

Feature Selection

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

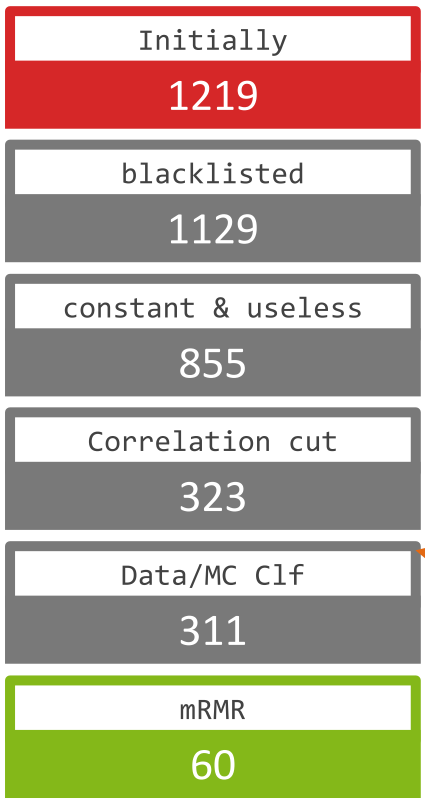
M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Feature Selection



M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Simulated and experimental data should agree to not bias the result.

technische universität dortmund

lehrstuhl physik e5

Feature Selection

Initially
1219

blacklisted
1129

constant & useless
855

Correlation cut
323

Data/MC Clf
311

mRMR
60

M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

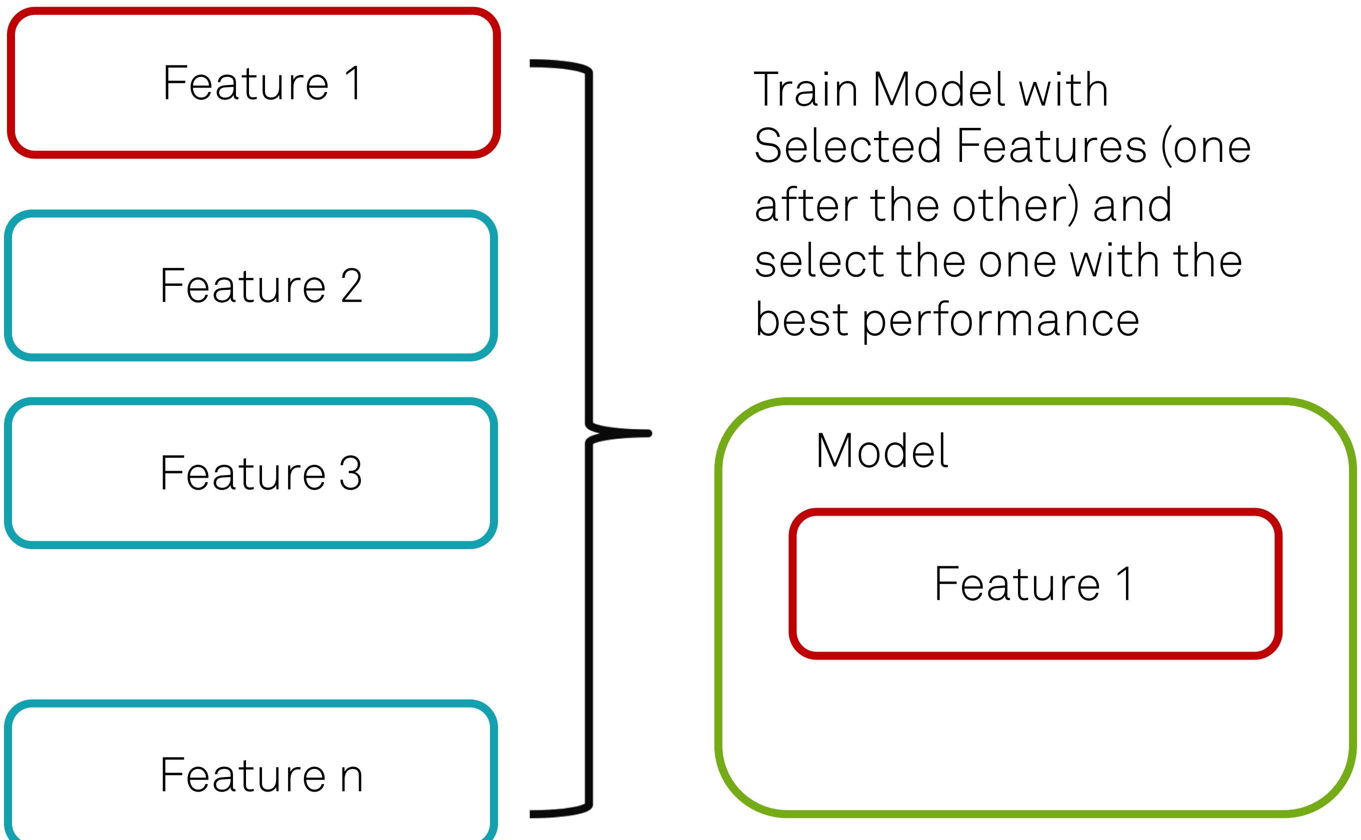
Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Simulated and experimental data should agree to not bias the result.

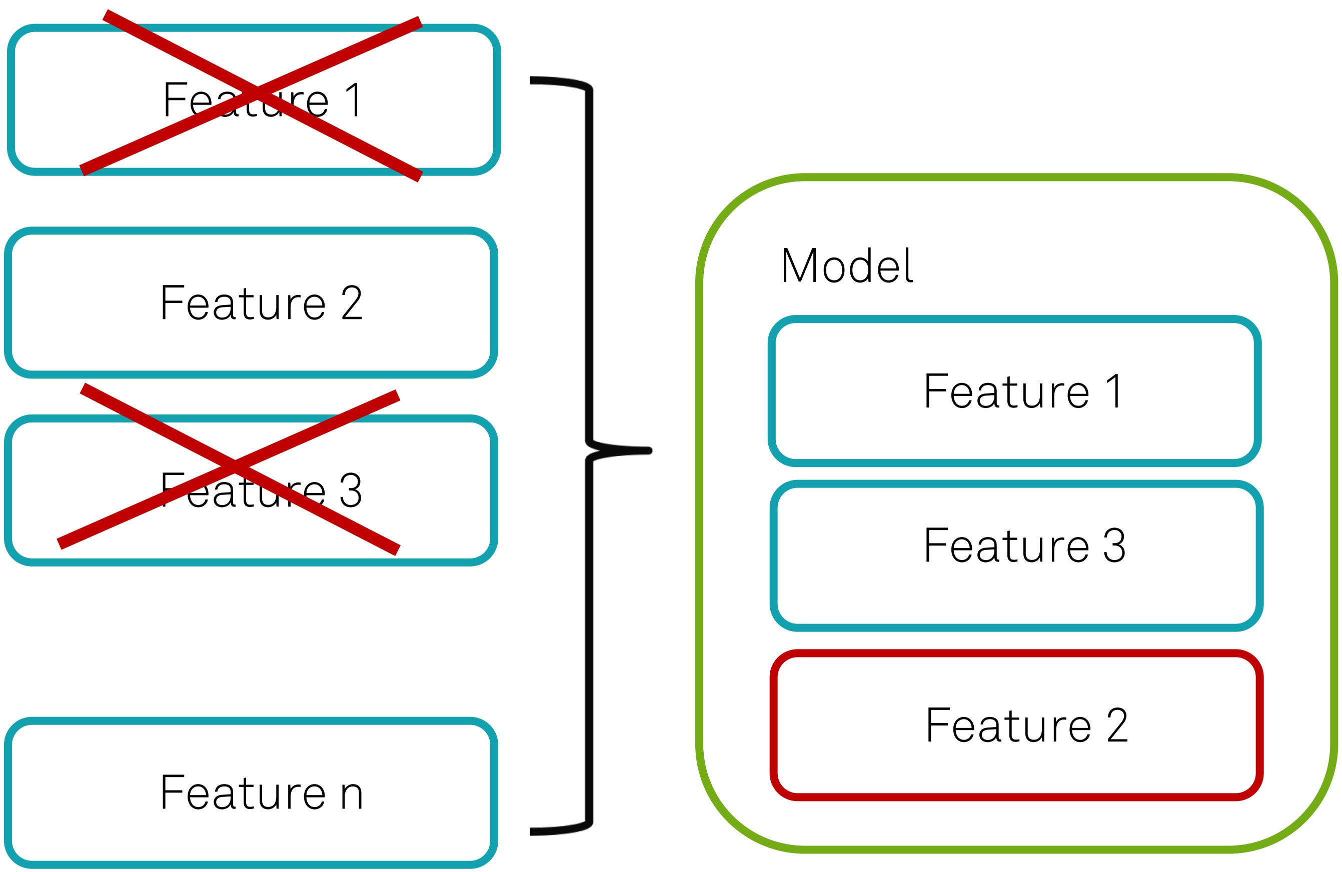
Automated selection by a feature selection algorithm.

Feature Selection: Forward Selection



Train Model with Selected Features (one after the other) and select the one with the best performance

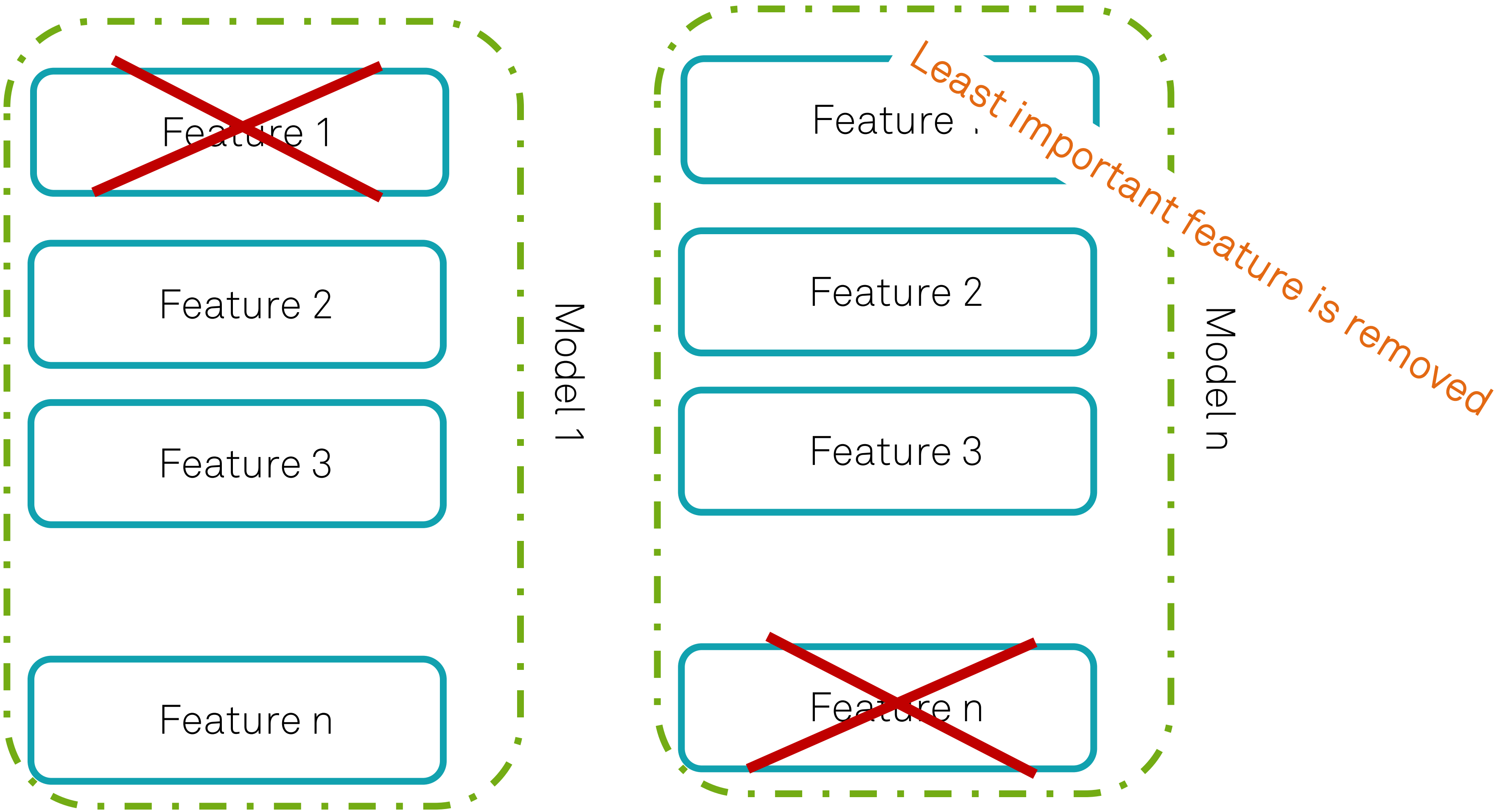
Feature Selection: Forward Selection



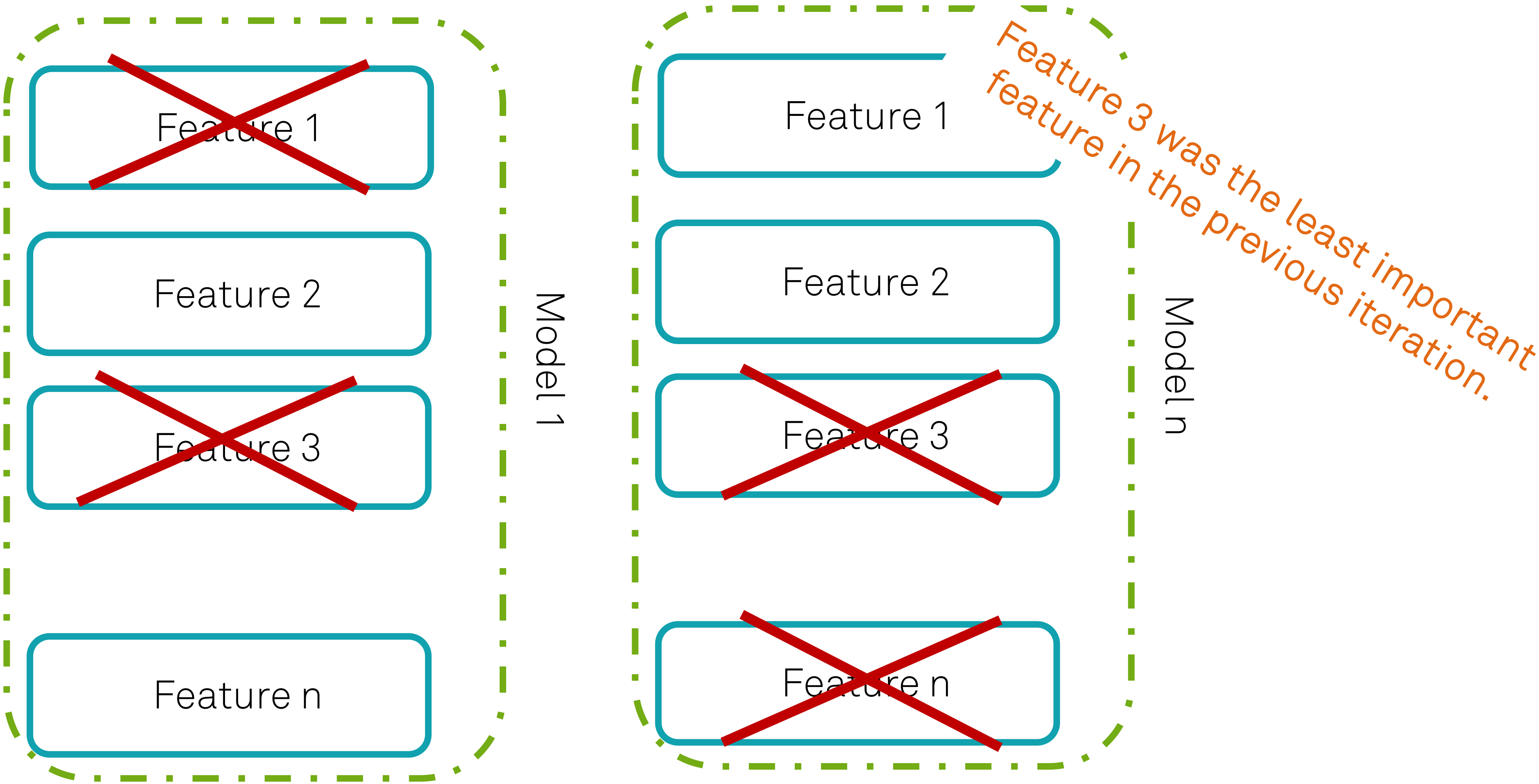
Add Features (one after another) and train a classifier.

Select the features that give the best performance.

Feature Selection: Backward Elimination (1st Iteration)



Feature Selection: Backward Elimination (2nd Iteration)



Minimum Redundancy Maximum Relevance (1st Iteration)

Feature 1

Feature 2

Feature 3

Feature n

Compute correlation of all features with respect to the target class.

Select the feature with the largest correlation.

Minimum Redundancy Maximum Relevance (3rd Iteration)

~~Feature 1~~

Feature 2

~~Feature 3~~

Feature n

Compute correlation of all features with respect to the target class.

Compute Correlation with already selected features.

Select the feature which maximizes

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

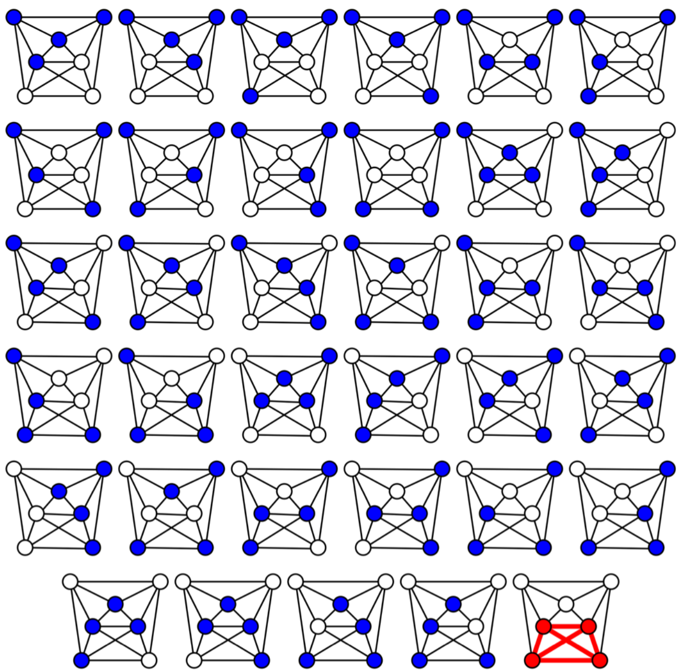
Feature 1 (Selected)

Feature 3 (Selected)

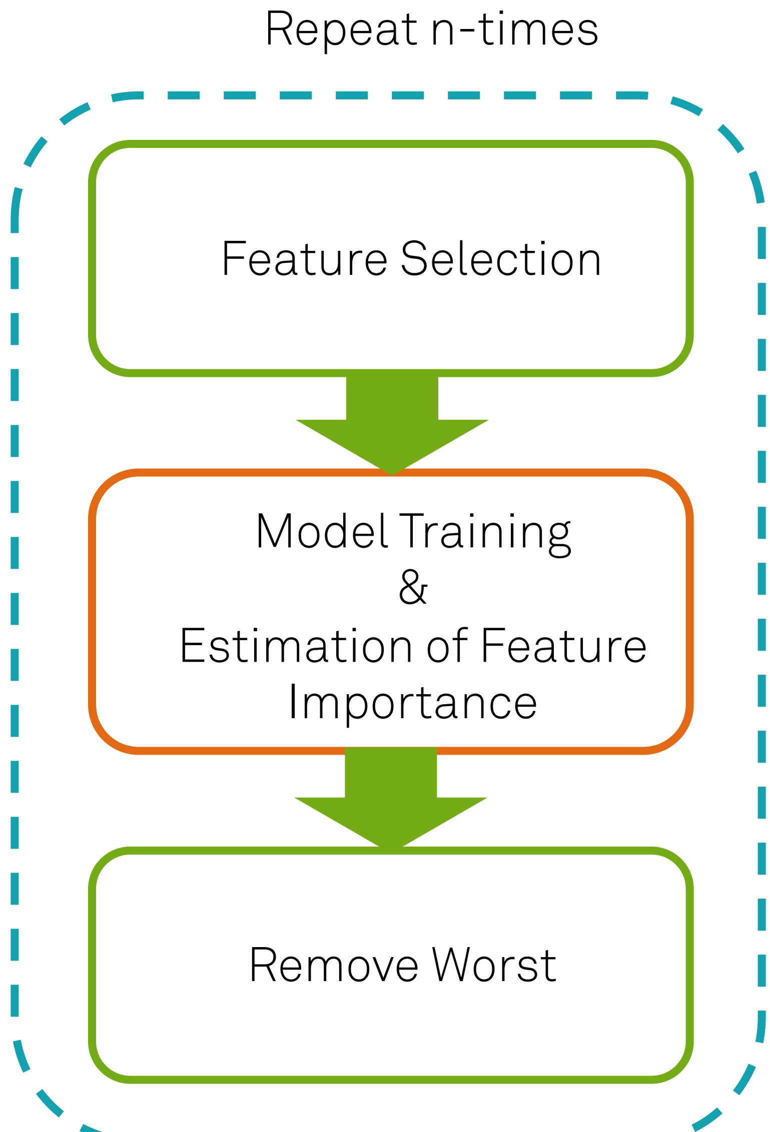
Peng, H.C., Long, F., and Ding, C., IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238, 2005.

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

Why is this Preferrable?



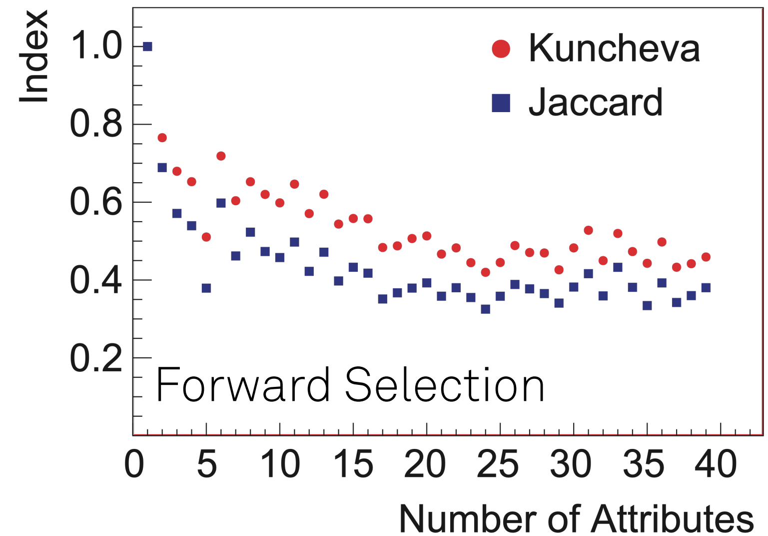
Source: By Thore Husfeldt at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31823619



Feature Selection Stability

Jaccard Index:

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



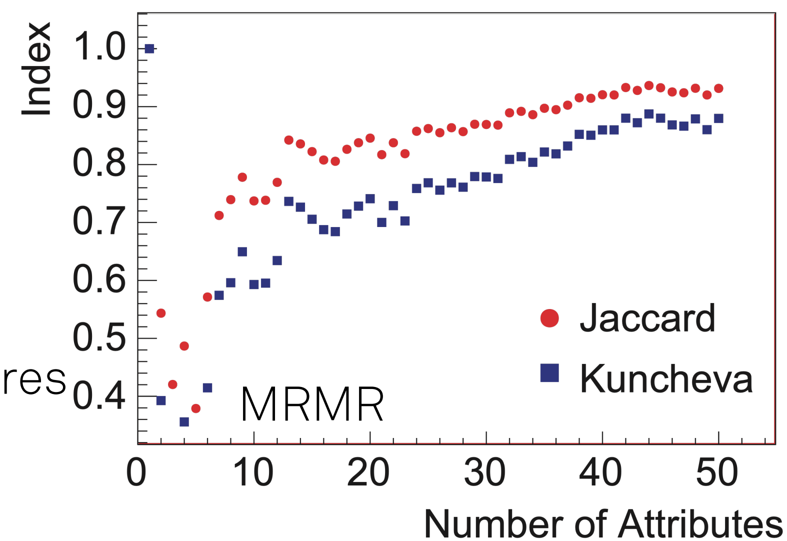
Kuncheva's Index:

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

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

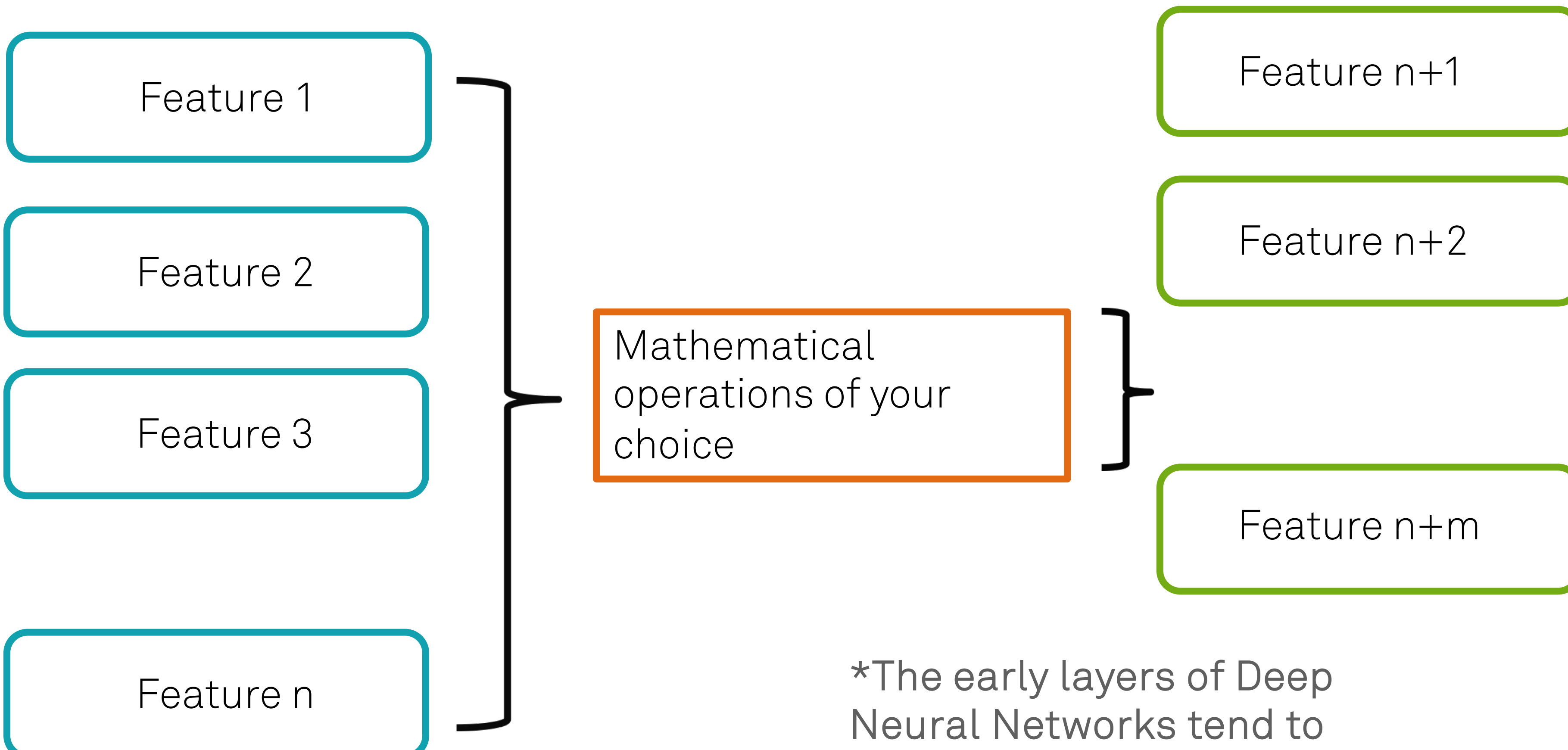
$$r = |A \cap B|$$

n: number of features



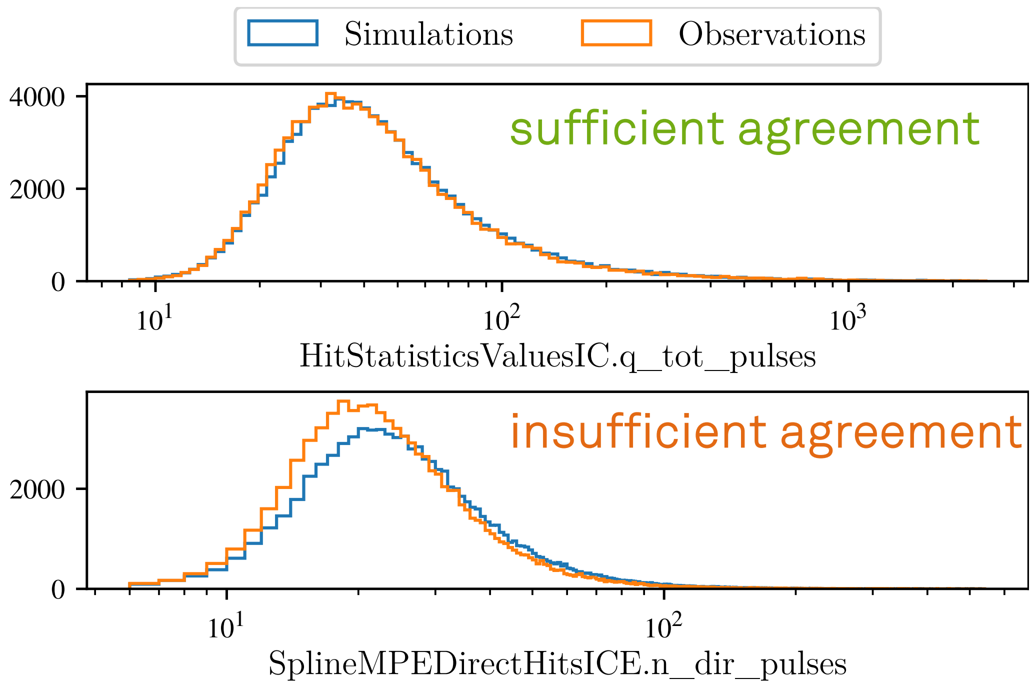
Ludmila I. Kuncheva: A STABILITY INDEX FOR FEATURE SELECTION

Feature Engineering



*The early layers of Deep Neural Networks tend to take care of this automatically.

Data/MC Disagreement

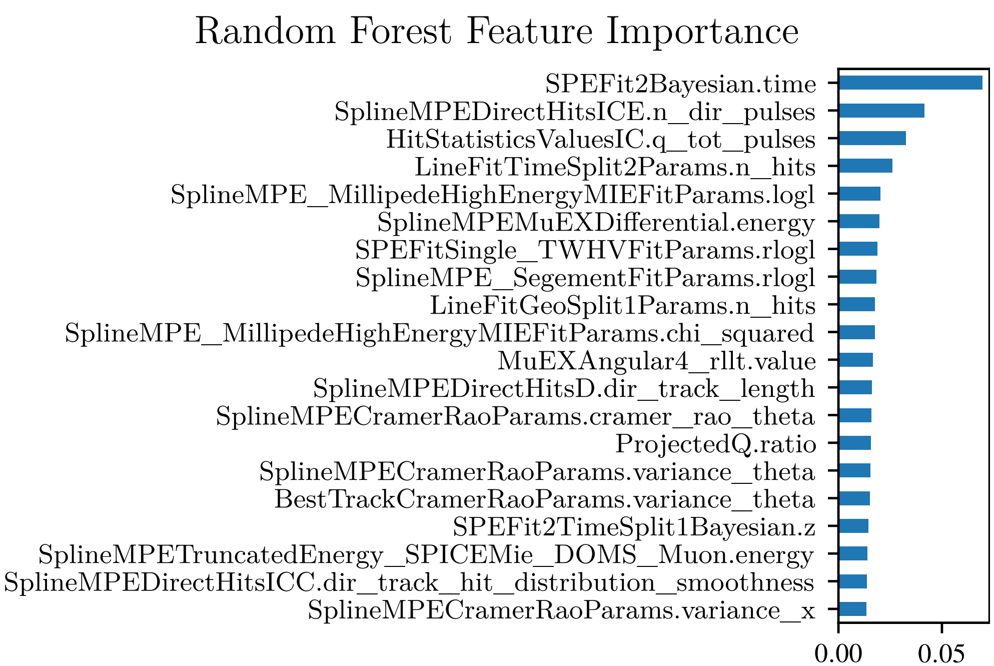


Graphics: M. Linhoff

Challenges when inspecting distributions by eye:

- only looking at one-dimensional distributions
- Systematic errors in simulation will also affect correlations between features
- Which metric ???
- Which threshold ???

Quantifying Disagreements

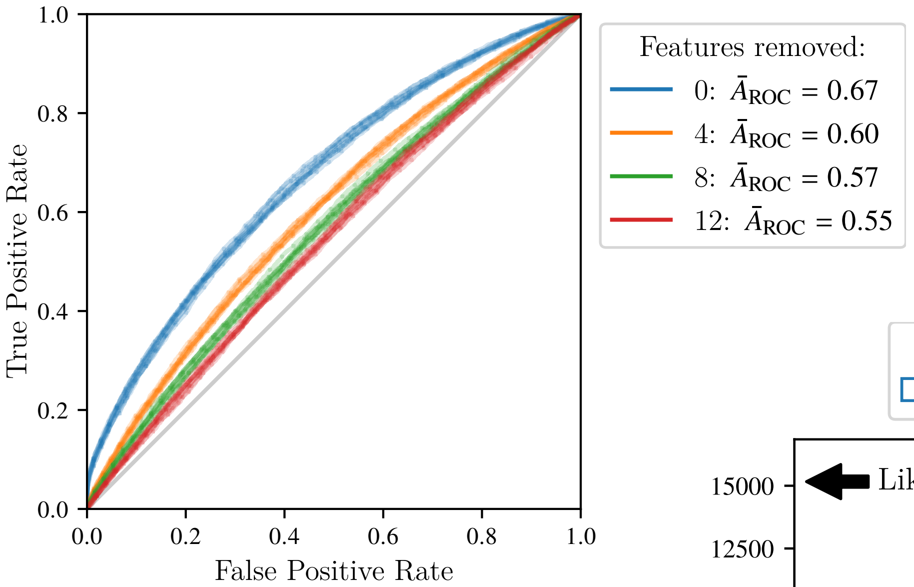


Graphics: M. Linhoff

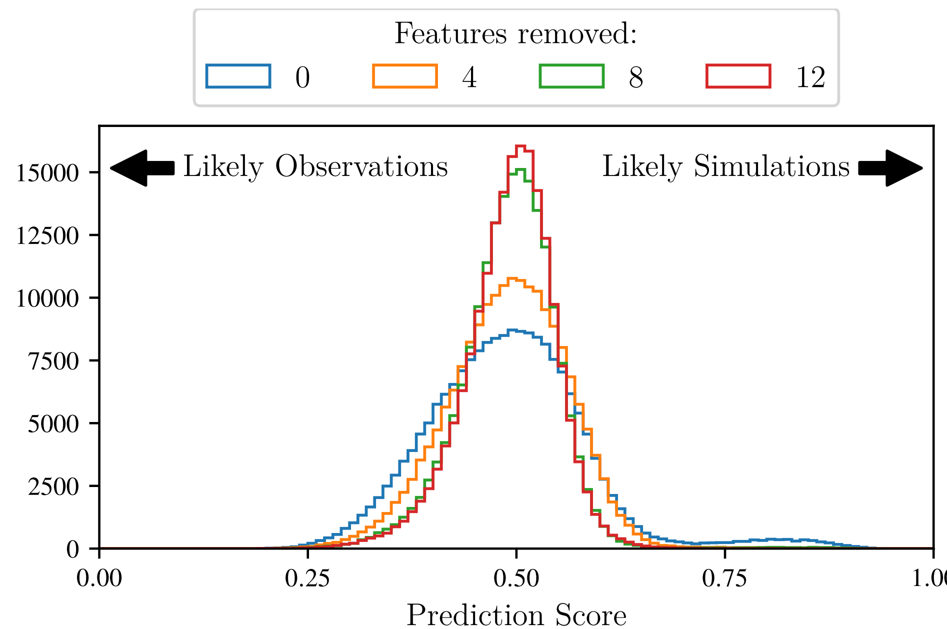
General Idea:

- Train classifier to distinguish simulated and experimental data
- Hard to impossible for a perfect agreement
- Sort features according to their importance
- Discard to n features
- Advantage: Extent to which mismatches can be tolerated is set by the classifier

Separating Data from MC



Area under Curve is close to 0.5 (close to random guess).



Prediction score centered around 0.5 (close to random guess).

Graphics: M. Linhoff

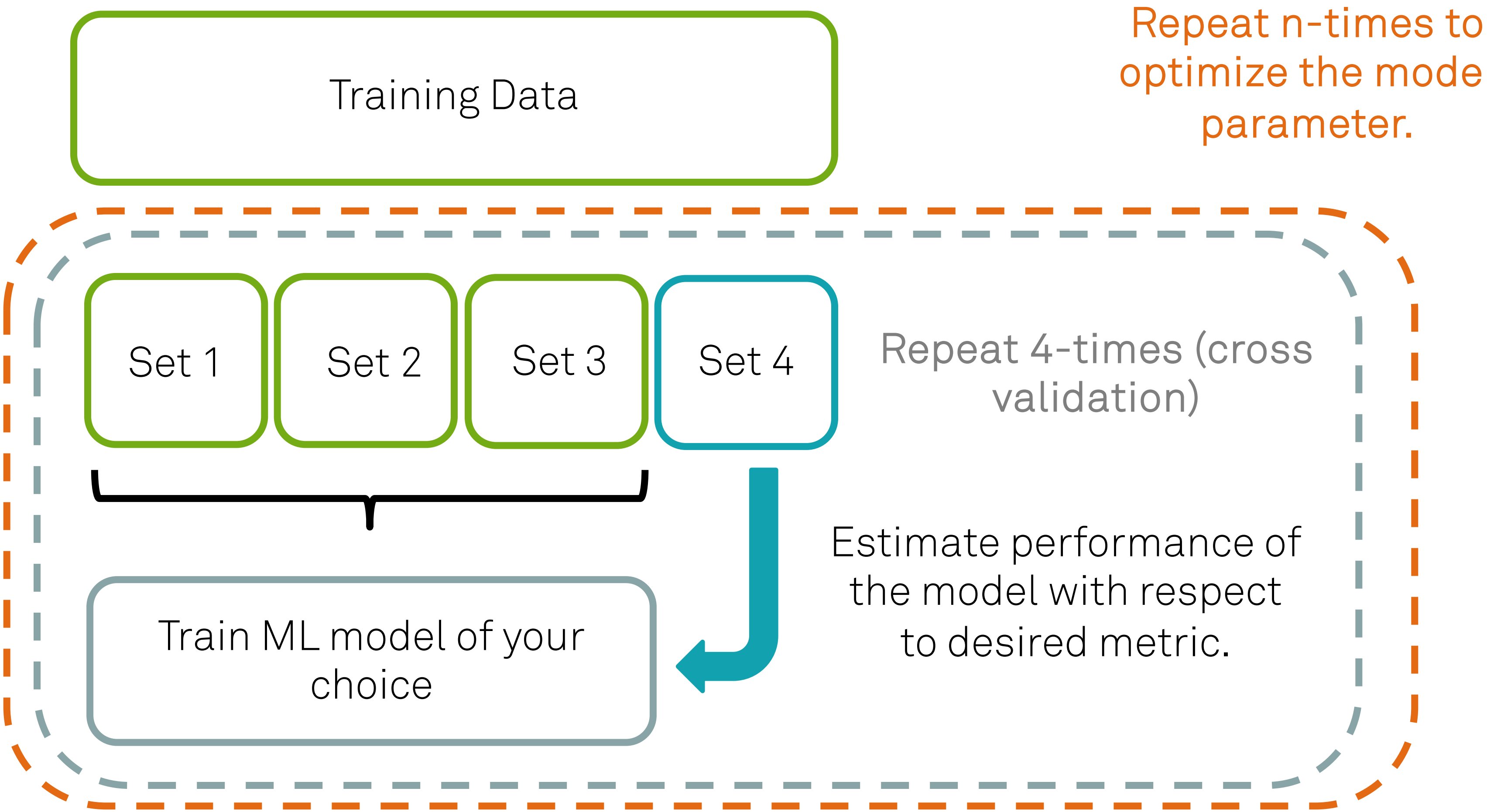
Model Building and Model Performance (Sketch)

Events available for model building (e. g. Monte Carlo simulations).

Training Data

Holdout Set

Model Building and Model Performance (Sketch)



Model Building and Model Performance (Sketch)

Events available for model building (e. g. Monte Carlo simulations).

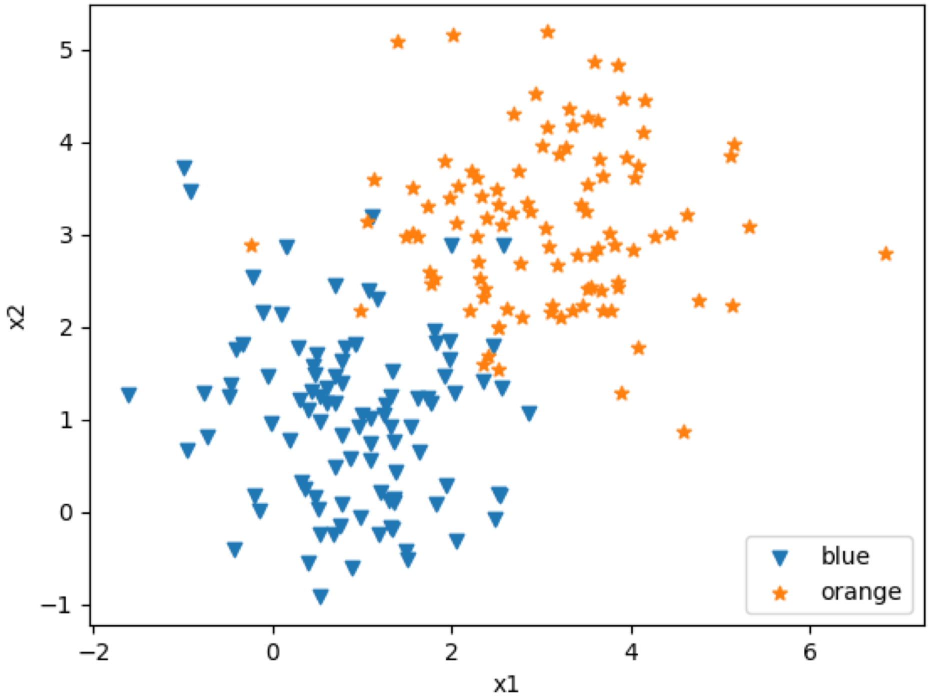
Training Data

Holdout Set

Fully optimized ML model

Re-check classifier performance on unseen data.

Nomenclature



N $(\vec{X}, y)$ pairs are referred to as training set
Or annotated data

Events (Examples) are characterized by a feature vector:

$$\vec{X} = (x_1 \ \dots \ x_n)$$

In this example

$$\vec{X} = (x_1, x_2)$$

And a class variable

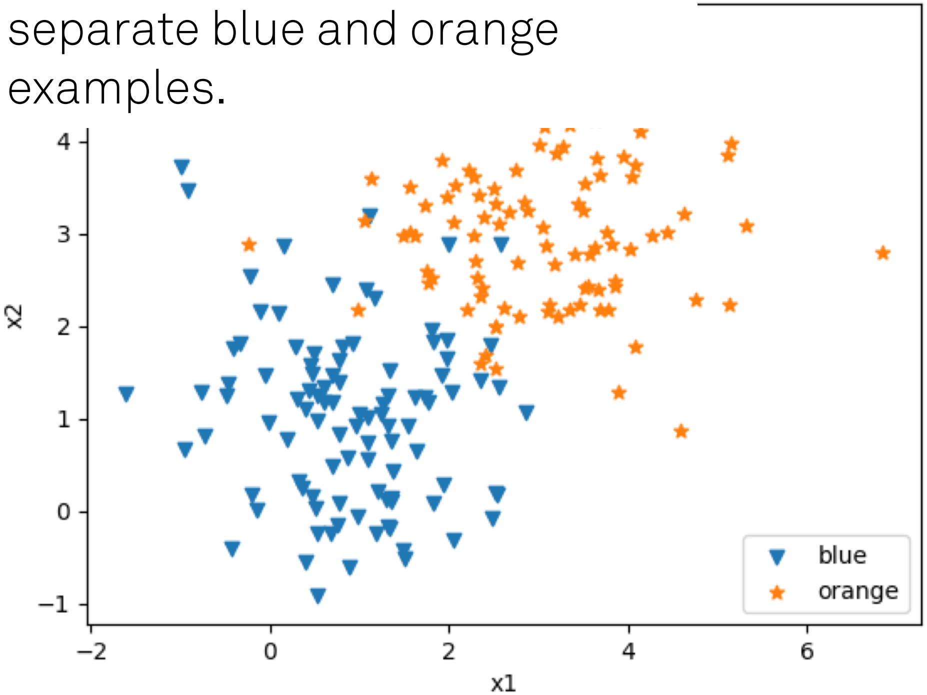
$$y \ \in \ [y_1 \ \dots \ y_n]$$

In this example

$$y \ \in [blue, orange]$$

Nomenclature

Task: Build a model to separate blue and orange examples.



Events (Examples) are characterized by a feature vector:

$$\vec{X} = (x_1 \ \dots \ x_n)$$

In this example

$$\vec{X} = (x_1, x_2)$$

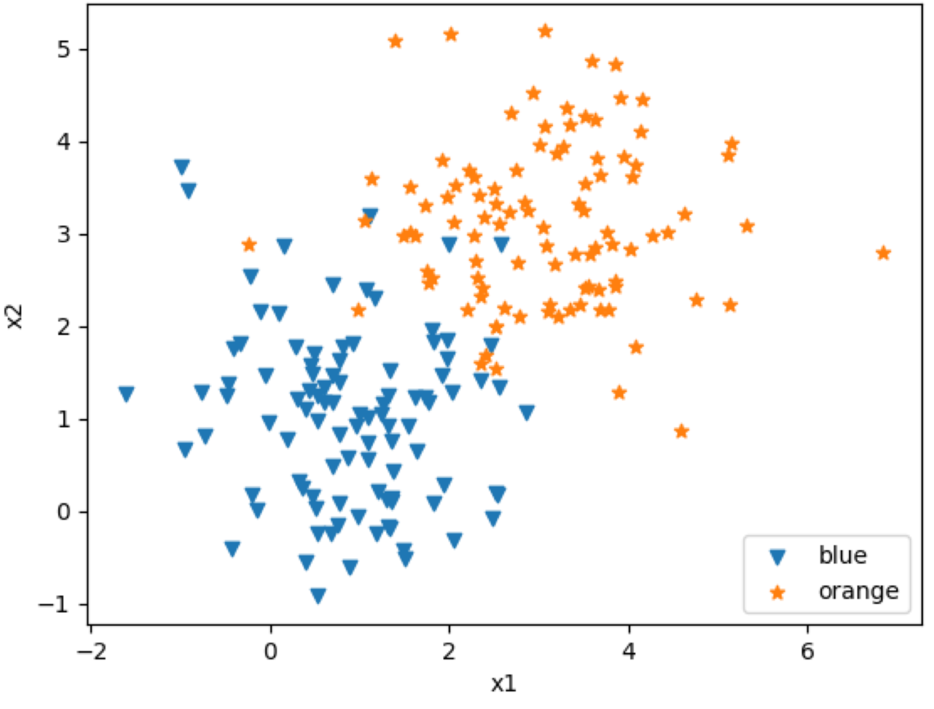
And a class variable

$$y \ \in \ [y_1 \ \dots \ y_n]$$

In this example

$$y \ \in [blue, orange]$$

The Linear Model



$$\hat{y} = \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

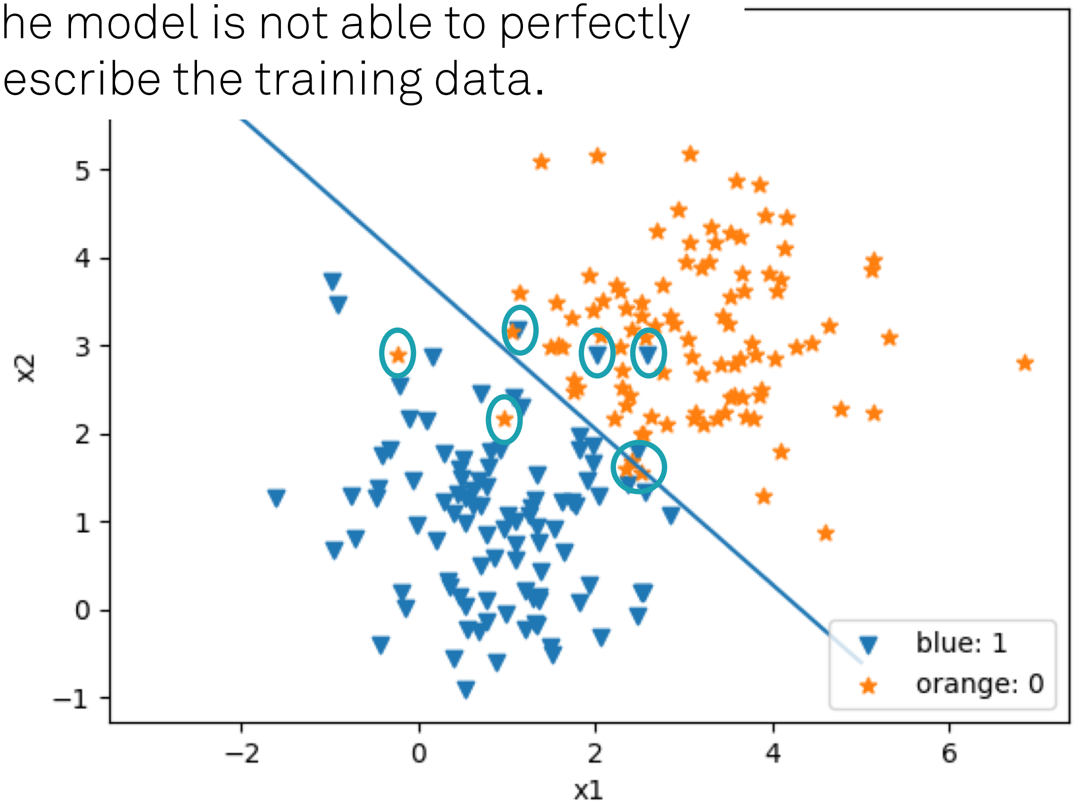
$$\hat{y} = \begin{cases} \text{orange: } 0 \\ \text{blue: } 1 \end{cases}$$

Solve e.g. by least squares fit

The Linear Model: Graphical Representation of the Model

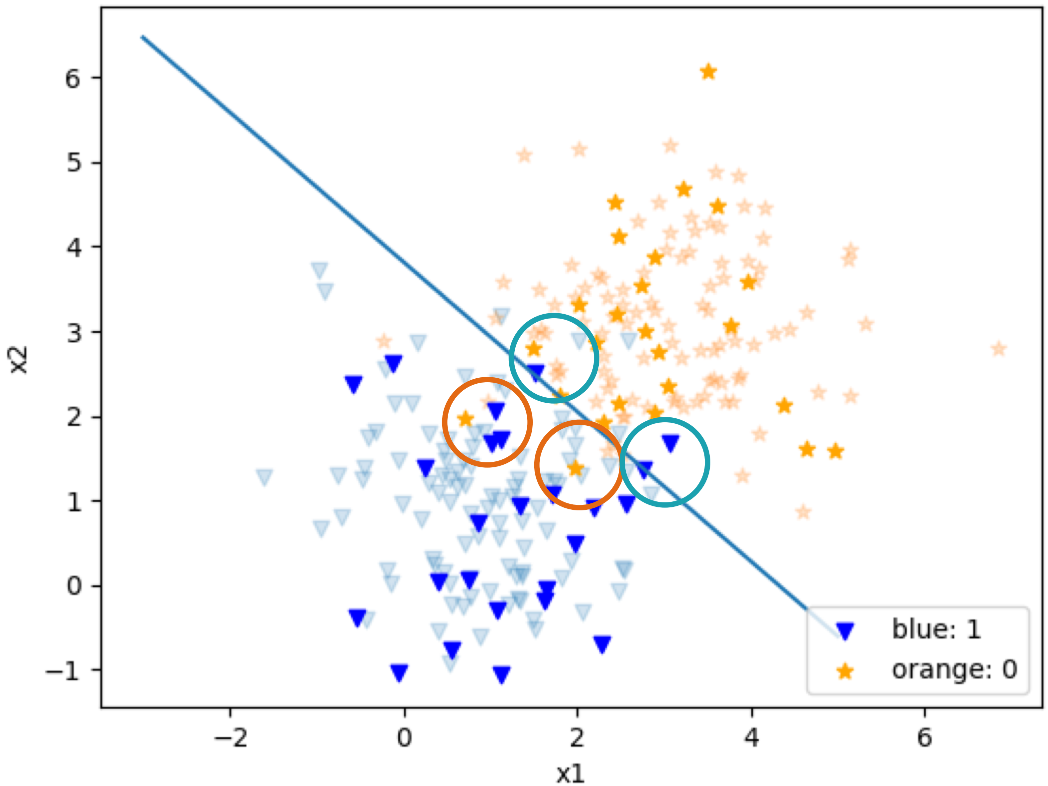
The model is not able to perfectly
describe the training data.

Above line:
Classify as orange



Below line:
Classify as blue

Application to Unseen Data

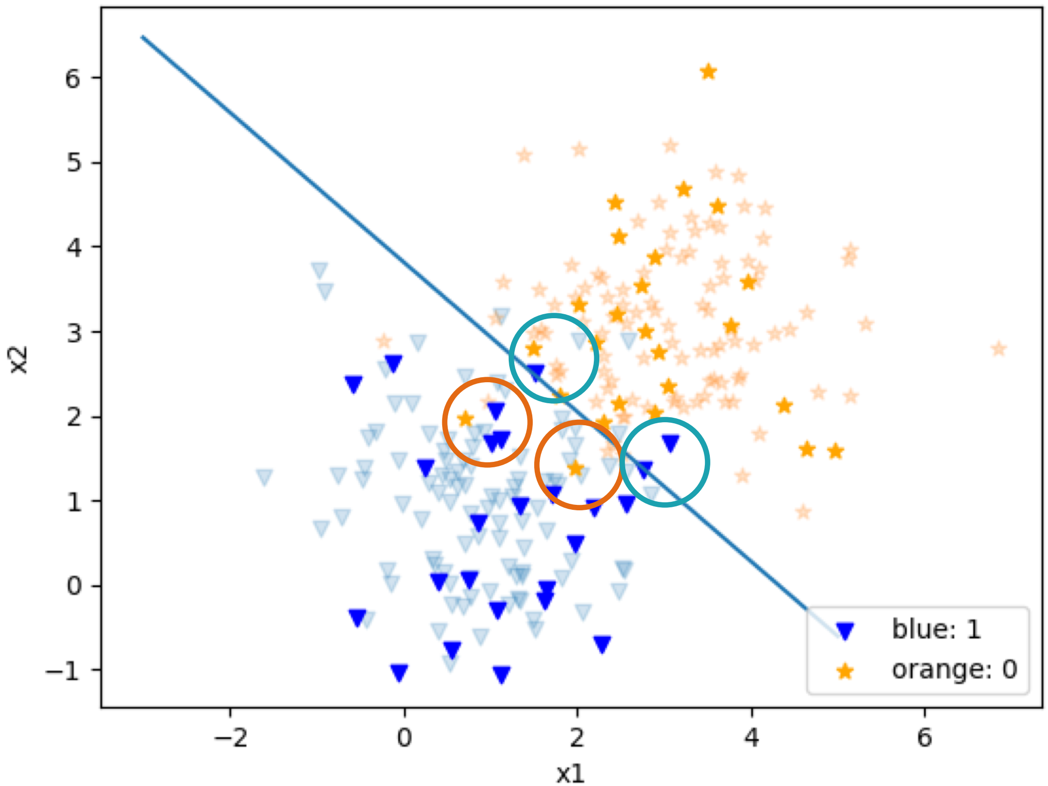


Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

*I defined that border will be part of the orange class.

True and False Negatives and Postives



True Positives (TP)

False Positives (FP)

Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

False Negatives (FN)

True Negatives (TN)

*I defined that border will be part of the orange class.

TPR, FPR, Accuracy and All That

Accuracy:

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision:

$$PREC = \frac{TP}{TP + FP}$$

Recall:

$$REC = \frac{TP}{TP + FP}$$

* These measures can sometimes have different names

True Positives (TP)

False Positives (FP)

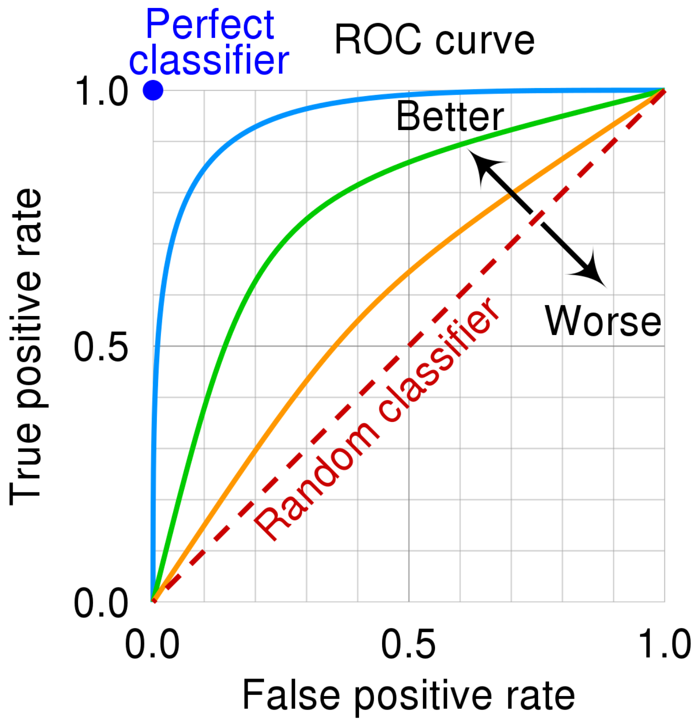
Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

False Negatives (FN)

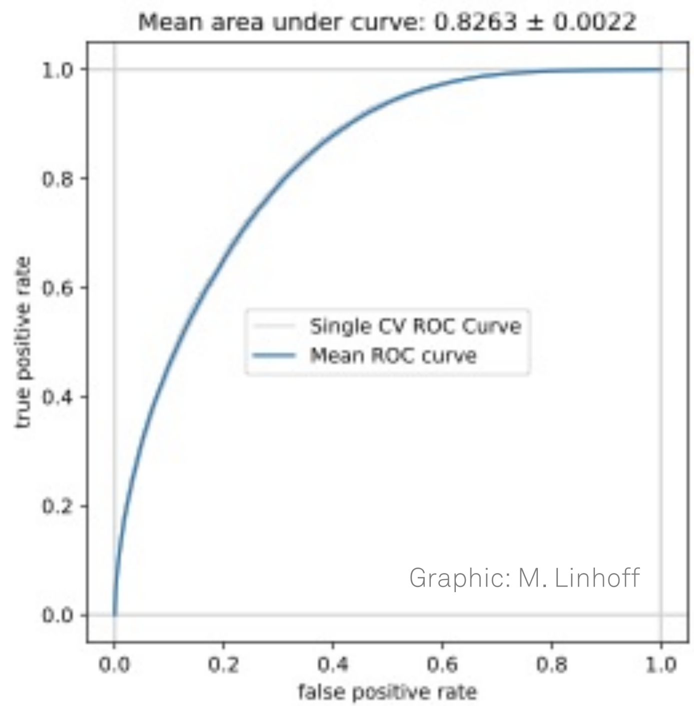
True Negatives (TN)

Area Under Curve



Source: By cmglee, MartinThoma - Roc-draft-xkcd-style.svg, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=109730045

Graphics: M. Linhoff [Learning Under Resource Constraints – Discovery in Physics] (in preparation)



ROC characteristic for the FACT Open Crab data set

Naive Bayes

$$p(k|\vec{x}) = \frac{p(k) \cdot p(\vec{x}|k)}{p(\vec{x})}$$

Naive assumption that all features are independent.

$$p(k|\vec{x}) = \frac{1}{Z} p(k) \prod_{i=1}^{n} p(x_i|k)$$

Use Naive Bayes here, and replace with actualy algorithm after debugging.

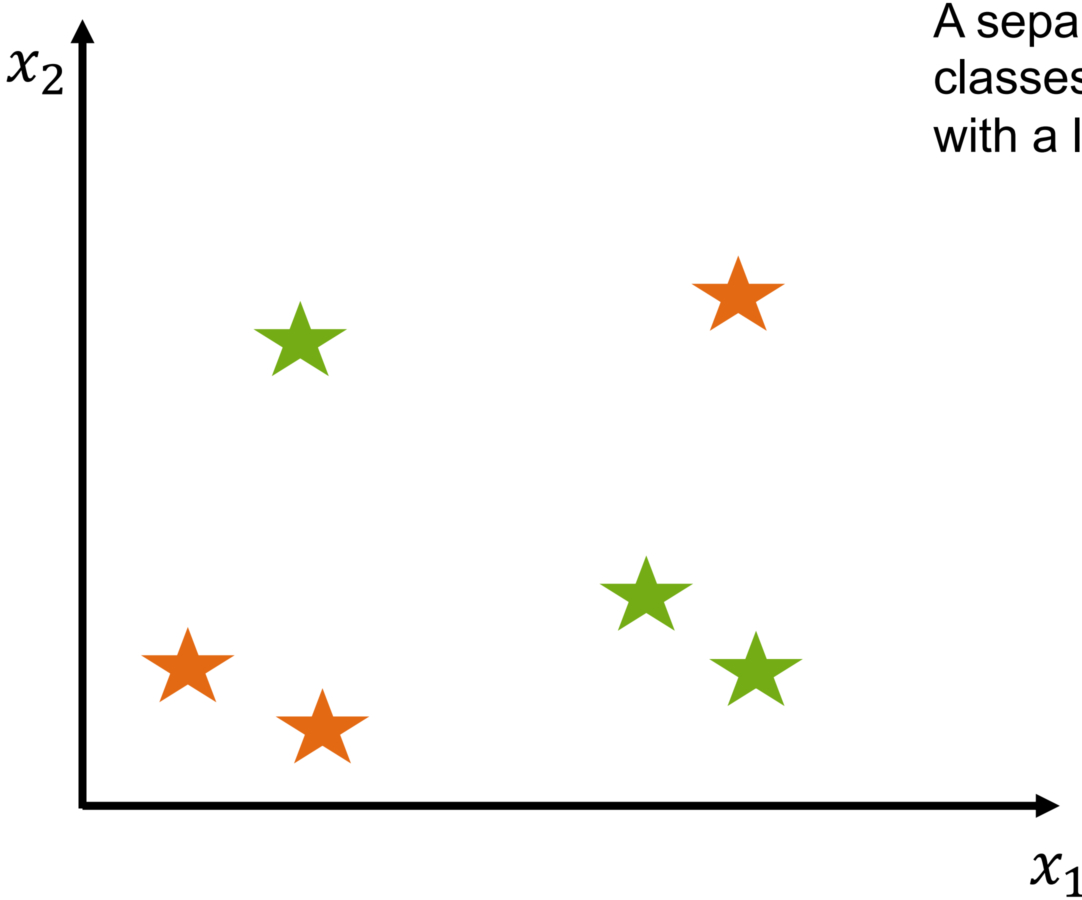
Some sort of preprocessing

Super fancy machine learning algorithm that'll take days to finish

Some sort of postprocessing

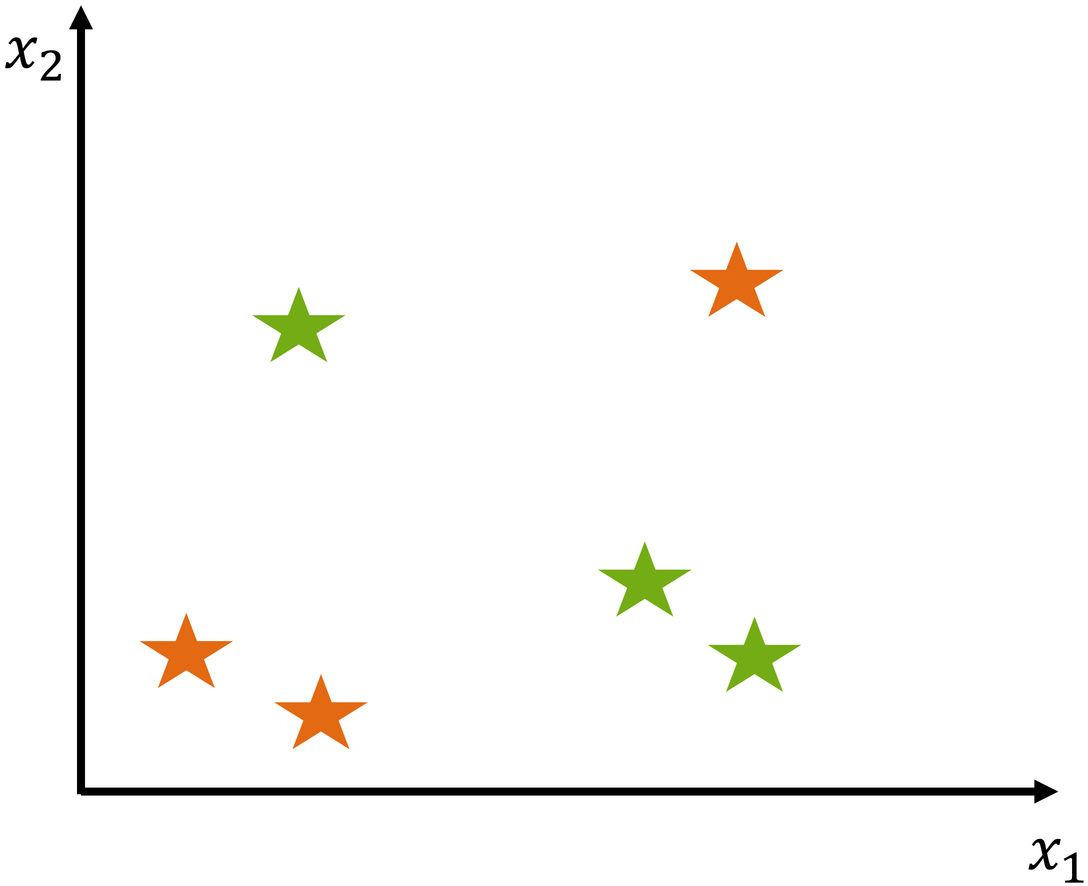
Great for debugging and benchmarking!

Decision Trees

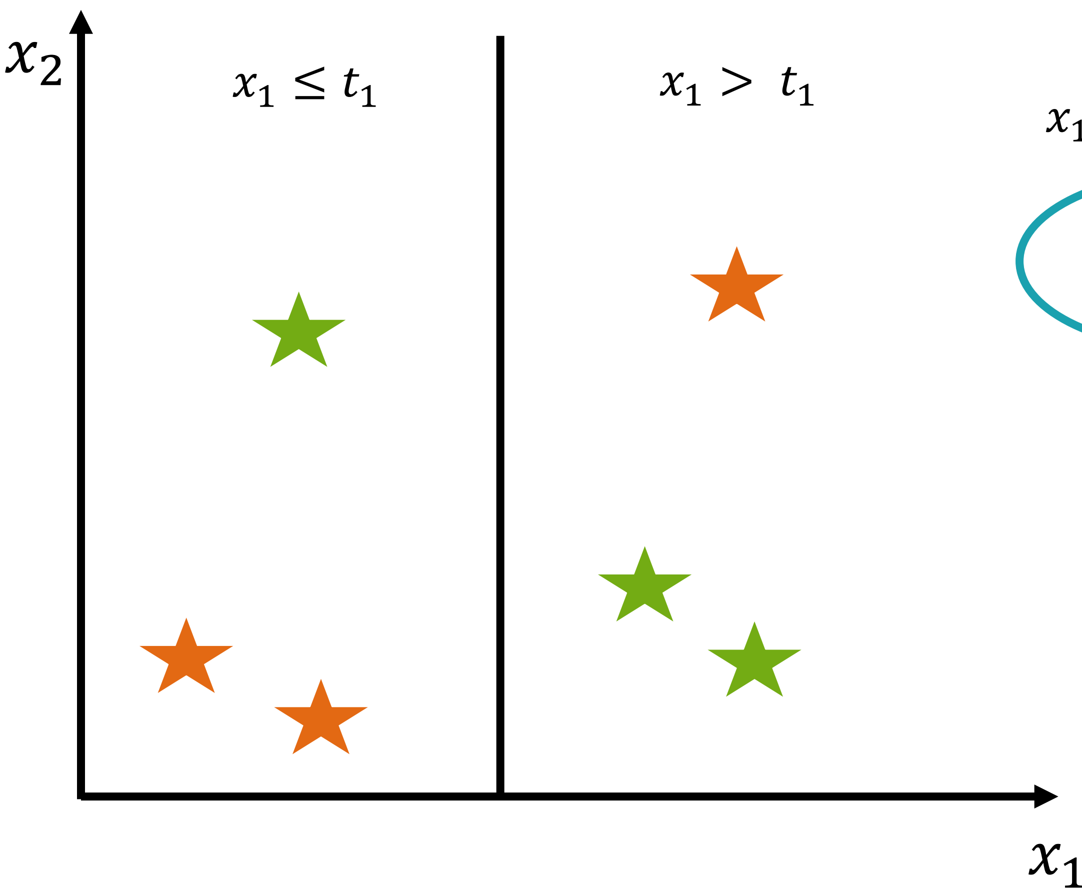


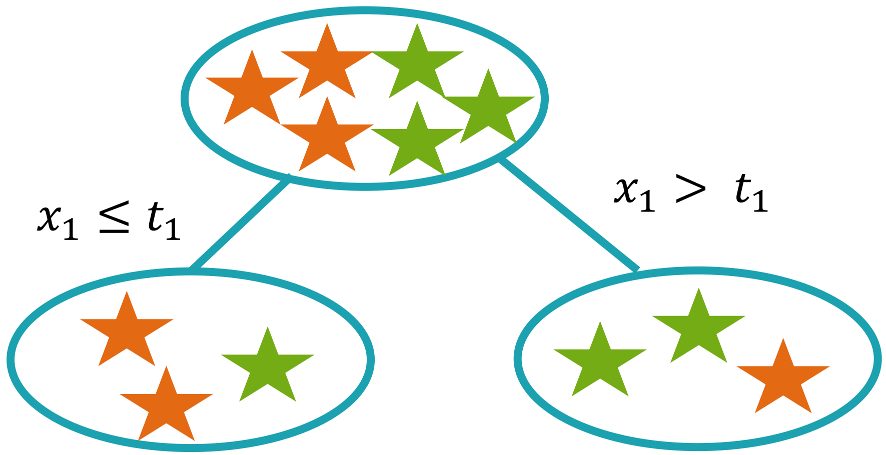
A separation of the two classes is difficult to achieve with a linear model.

Building a Decision Tree



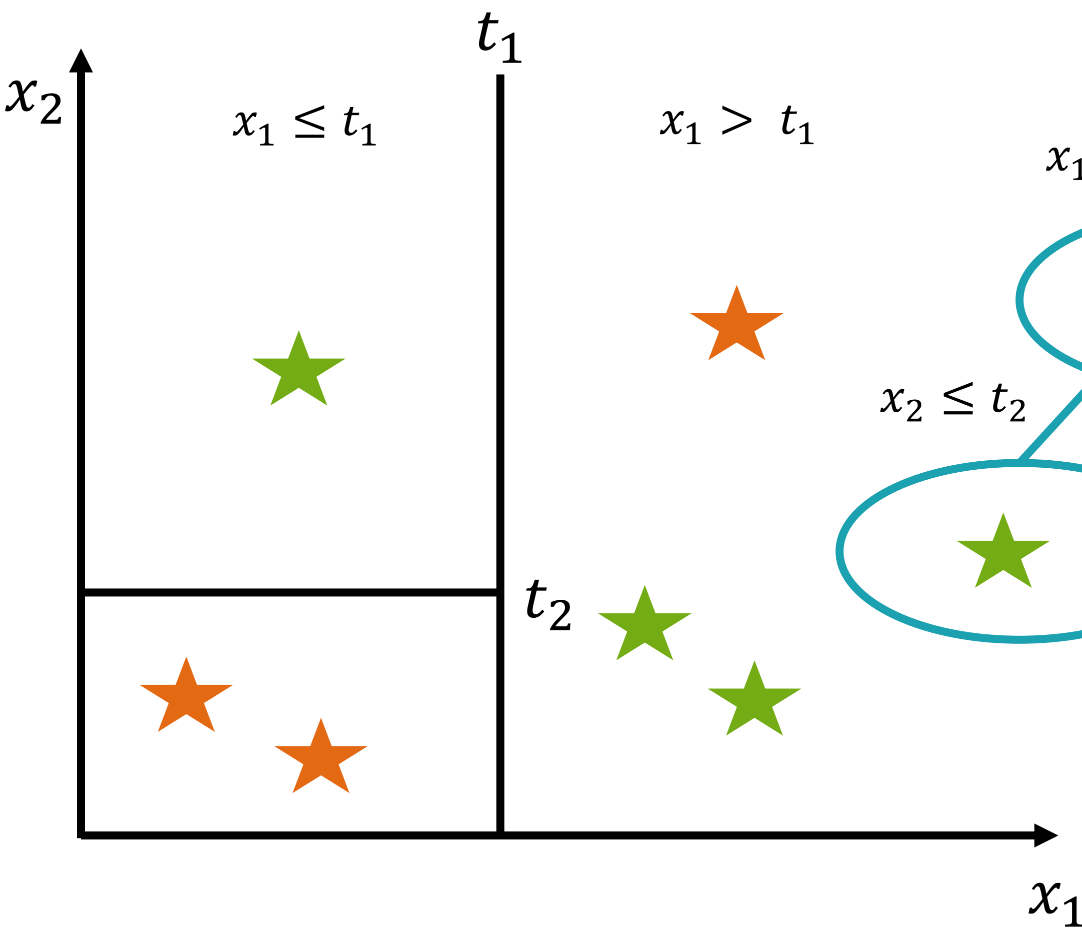
Building a Decision Tree

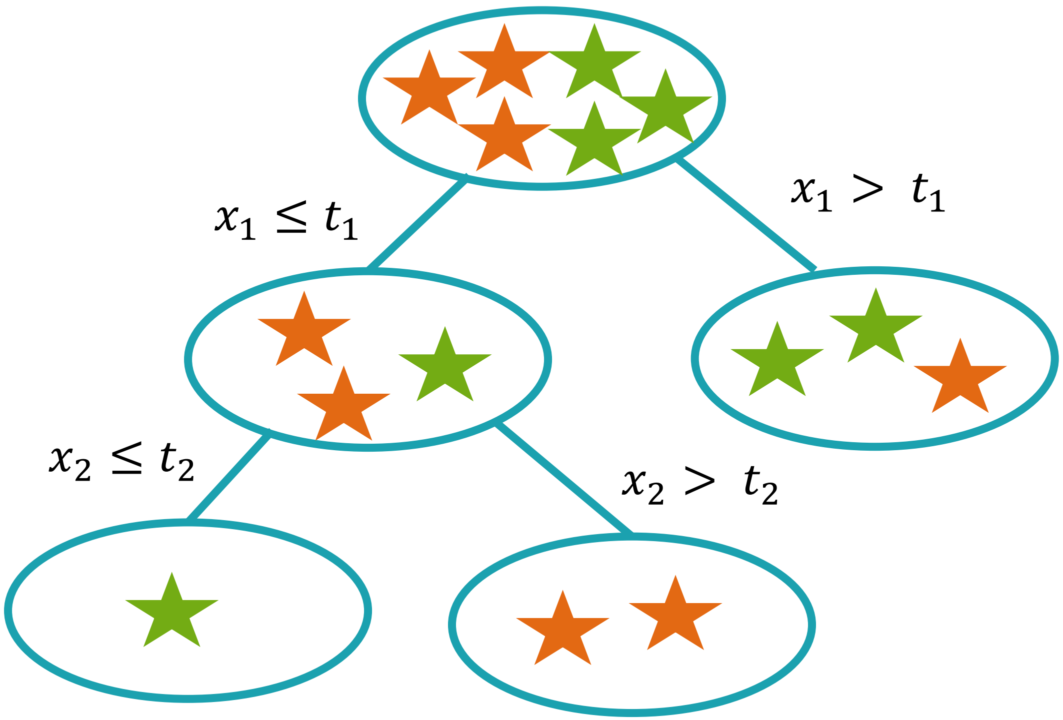




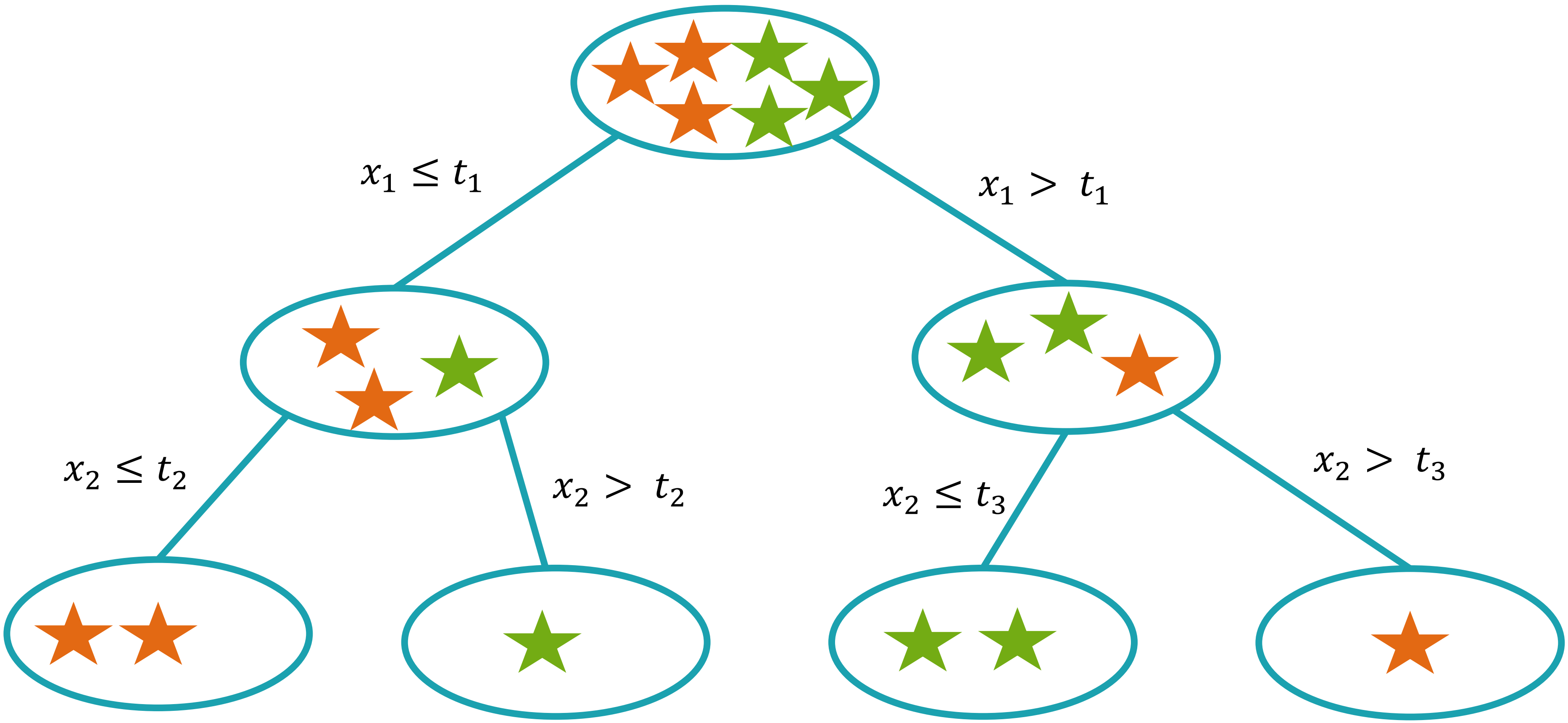
A decision tree with only a single split is sometimes referred to as a decision stump.

Building a Decision Tree

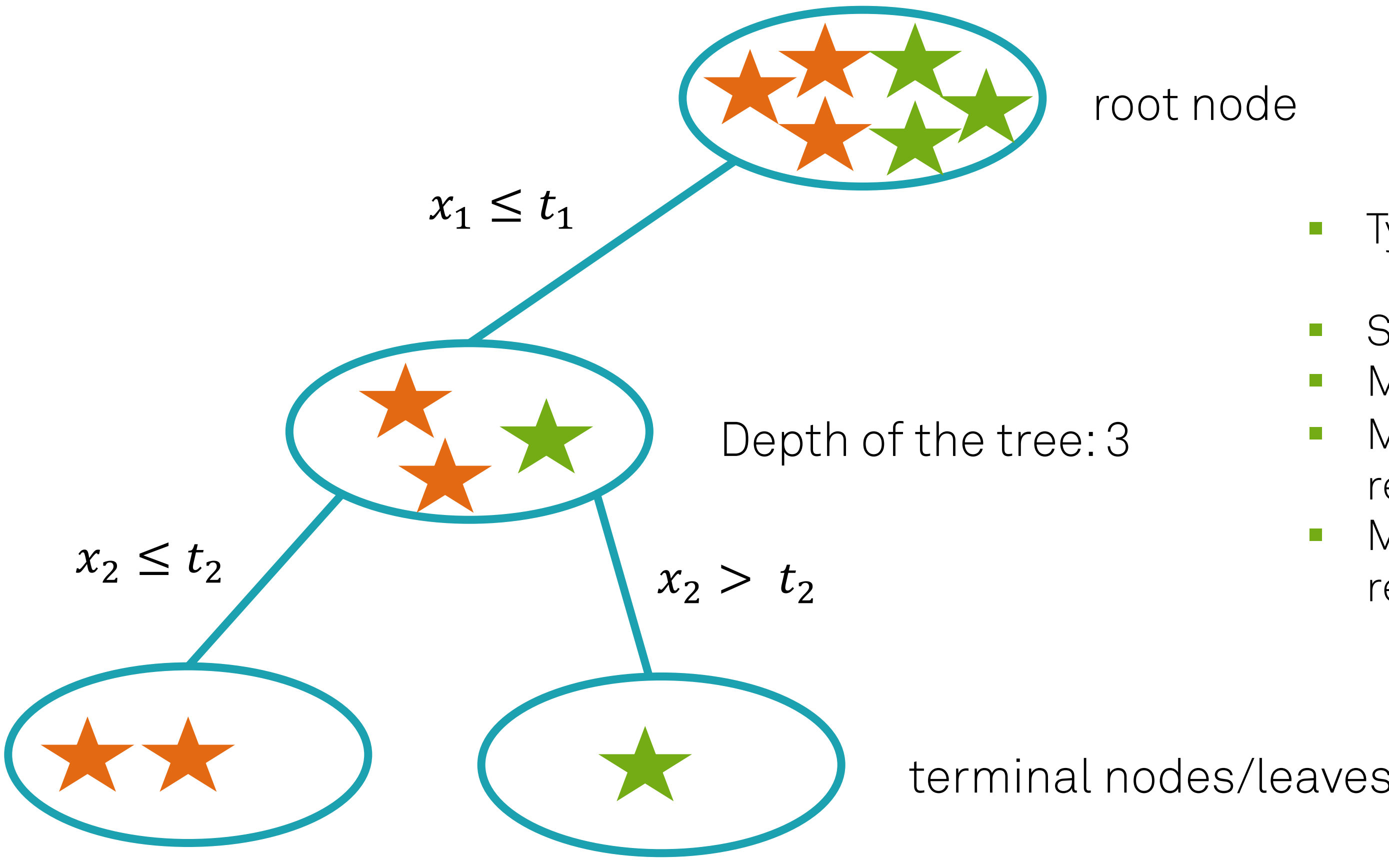




Building a Decision Tree



Decision Trees: Parameters and Nomenclature



- Typical Settings:

- Split criterion
- Maximum depth
- Minimum samples required for split
- Minimum samples required for a leaf

No Further Splits

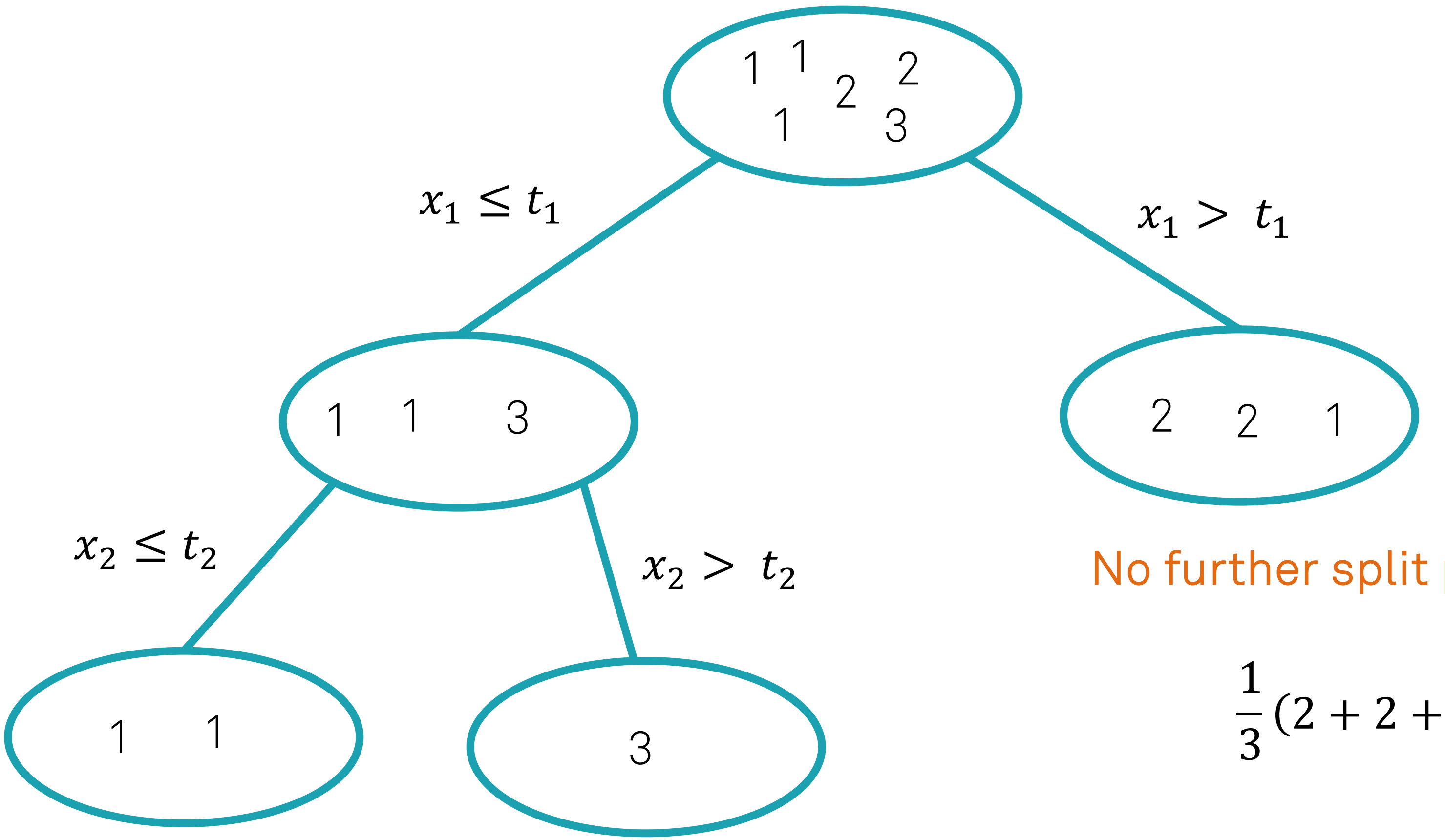
No further split possible

→ Majority vote: orange
→ Average:
confidence(green)= 1/3,
confidence(orange) = 2/3

No further split possible

→ Majority vote: green
→ Average:
confidence(green)= 2/3,
confidence(orange) = 1/3

Decision Trees for Regression



No further split possible

$$\frac{1}{3}(2+2+1) = \frac{5}{3}$$

How to Decide Where to Split?

$$\Delta i(N) = i(N) - P_L \, i(N_L) - (1 - P_L) \cdot i(N_R)$$

Impurity at the current node, e.g. the root node

Impurity at the left node, after a certain split

Impurity at the right node after a certain split.

The goal of a split is to find the combination of feature and cut-value that maximizes the decrease in impurity.

The nodes at the next step should be as pure as possible.

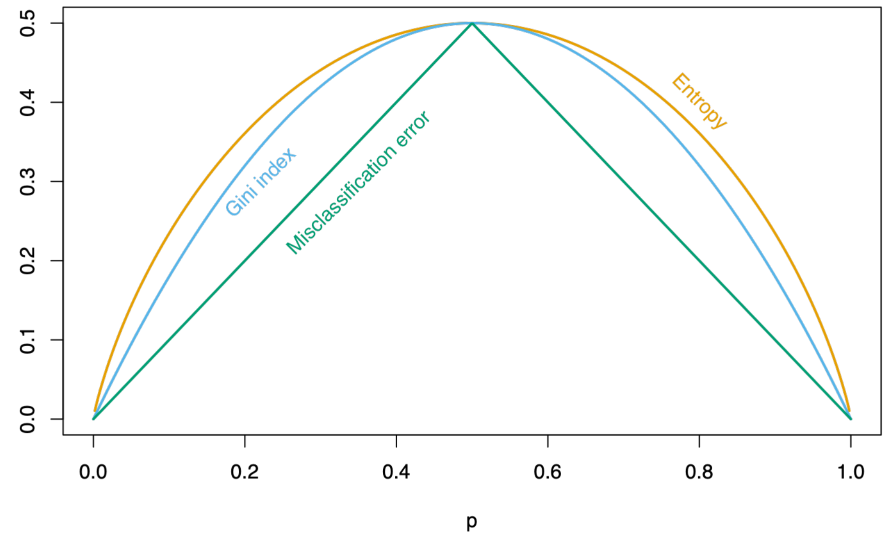
Different Measures of Impurity

p_{mk}: Proportion of class k in node m

$$i_{Misclass} = 1 - p_{mk(m)}$$

$$i_{Cros\ Entropy} = -\sum_{k=1}^{K} p_{mk} \log(p_{mk})$$

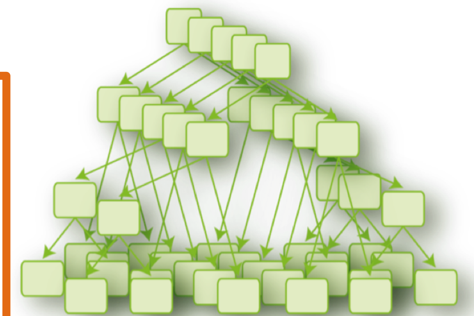
$$i_{Gini} = \sum\nolimits_{k \neq k'} p_{mk} p_{mk'} = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$



**Definitions are from Elements of statistical learning.

Random Forests

A Decision Tree is a weak classifier, but it can be strengthened by using ensembles of decision trees.



Random subset of examples to build each tree.

$x_1 \leq t_1$

Random subset features to determine the optimal split

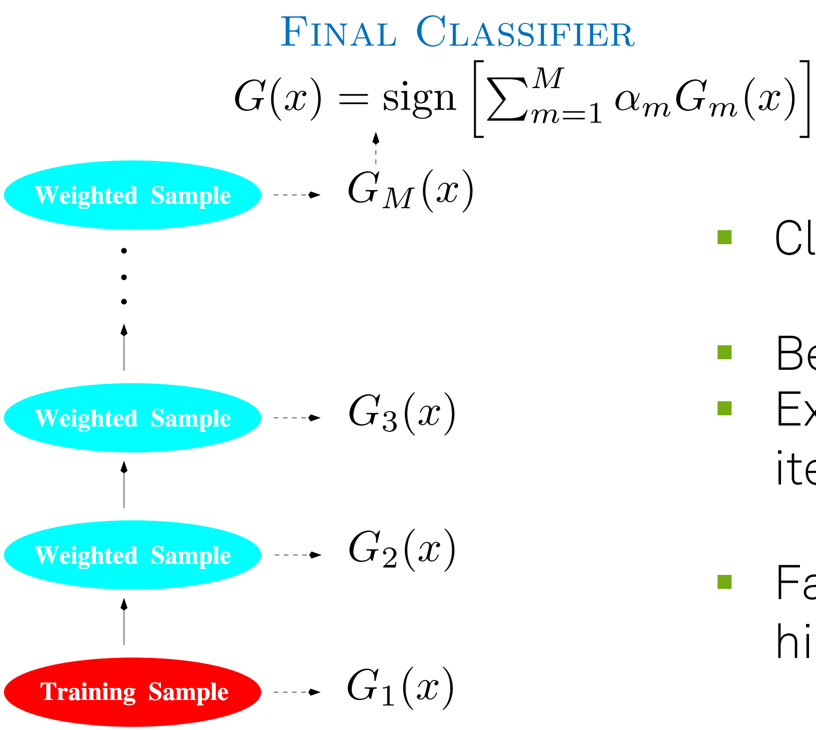
Random forests utilize an ensemble of independent weak classifiers (decision trees) to obtain a better classification.

Final classification is achieved via:

$$c_j = \frac{1}{n_{tress}} \sum_{i=1}^{n_{trees}} c_{ij}$$

c_{ij}: Classification for example j by tree i

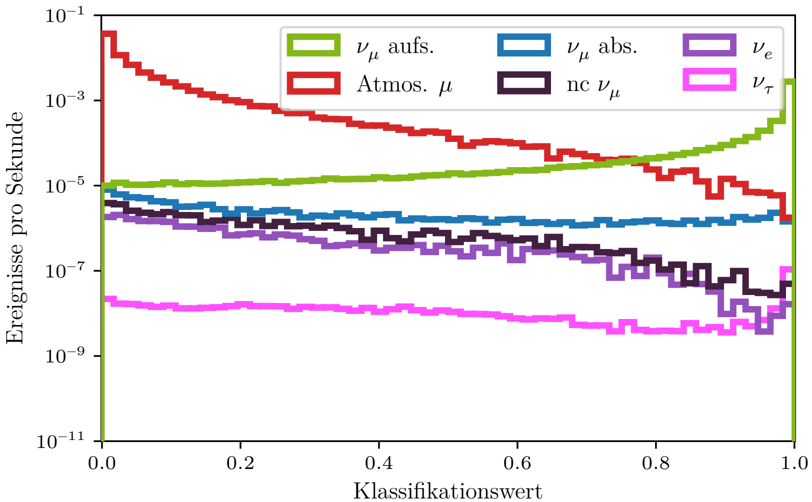
Boosting



- Classifiers are weighted by
  $$\alpha_m = \log((1 - err_m)/err_m)$$
- Better classifiers obtain higher weights
- Example weights are updated in every iteration
  $$w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot I(y_i \neq G(x_i))$$
- Falsely classified examples obtain higher weights in the next iteration

Source: Elements of Statistical Learning, Figure 10.1

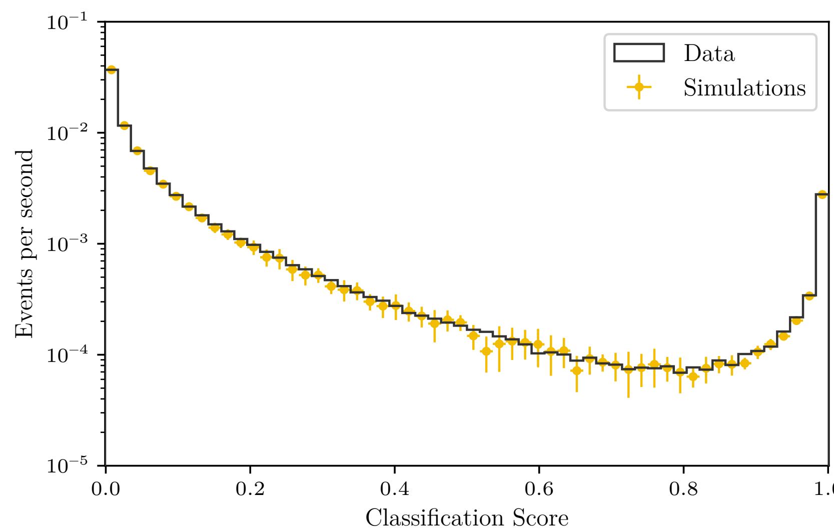
Model Output



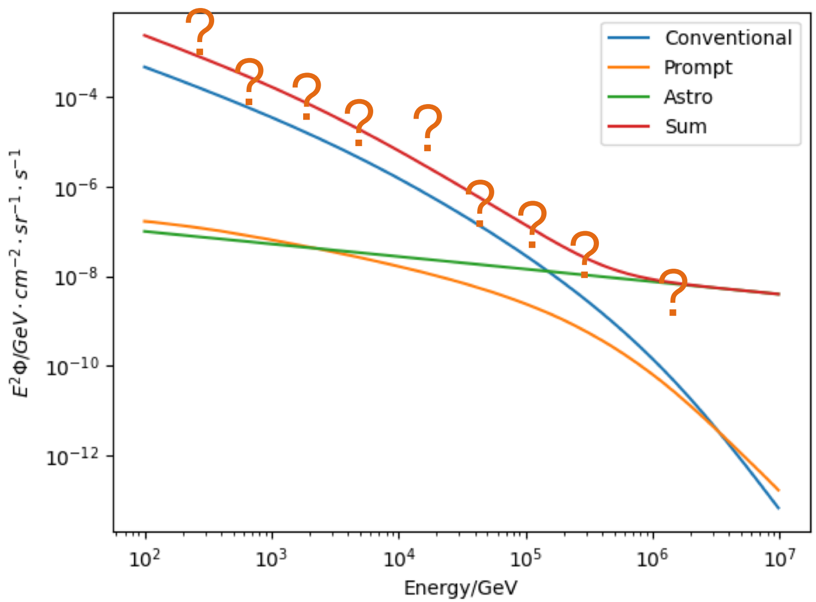
Reasonable agreement between simulated and experimental data.

The classifier will declare everything with $c > 0.5$ as signal.
Direct usage of the classifier output is not sufficient.

Extra cut is required.

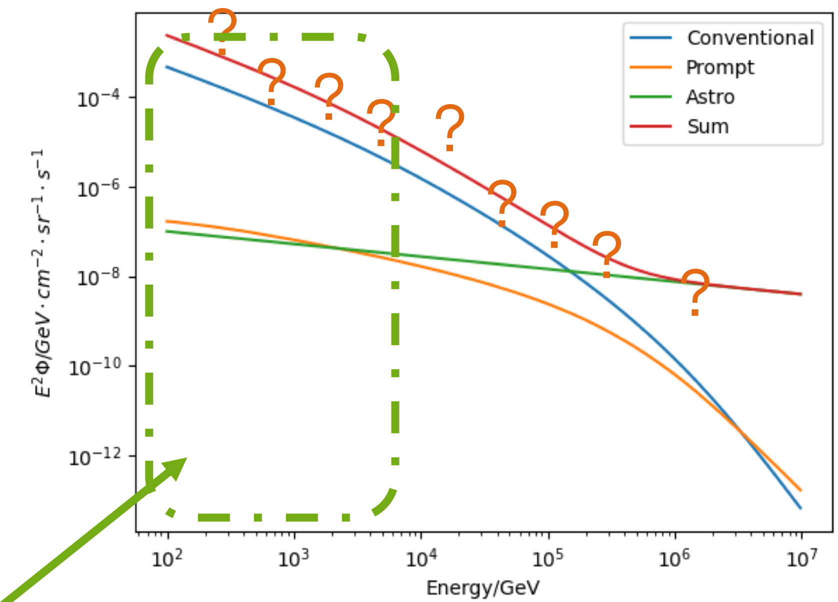


What to Consider Bevore Applying Additional Cuts



The analysis goal is to reconstruct a muon neutrino energy spectrum and to observe a flattening of the flux at high energies.

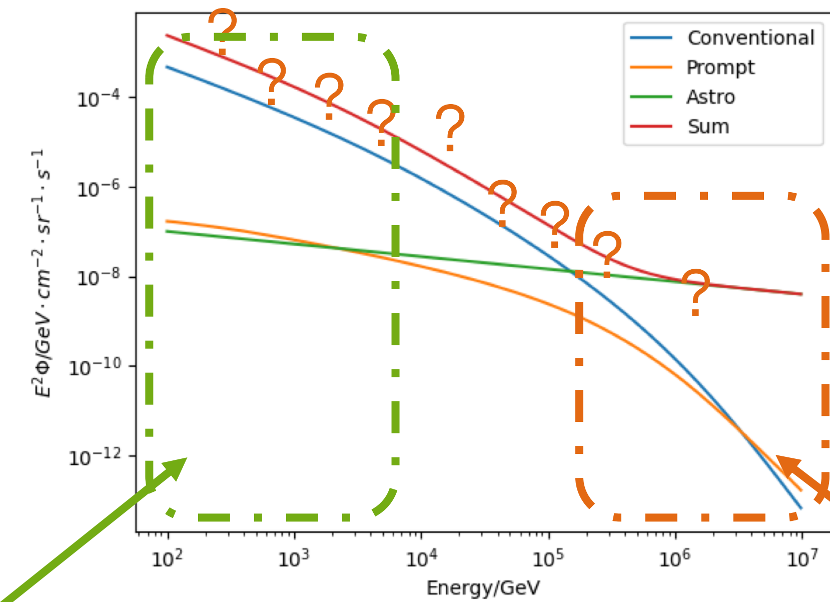
What to Consider Bevore Applying Additional Cuts



The analysis goal is to reconstruct a muon neutrino energy spectrum and to observe a flattening of the flux at high energies.

We can probably tolerate some muon backround in this region.

What to Consider Bevore Applying Additional Cuts



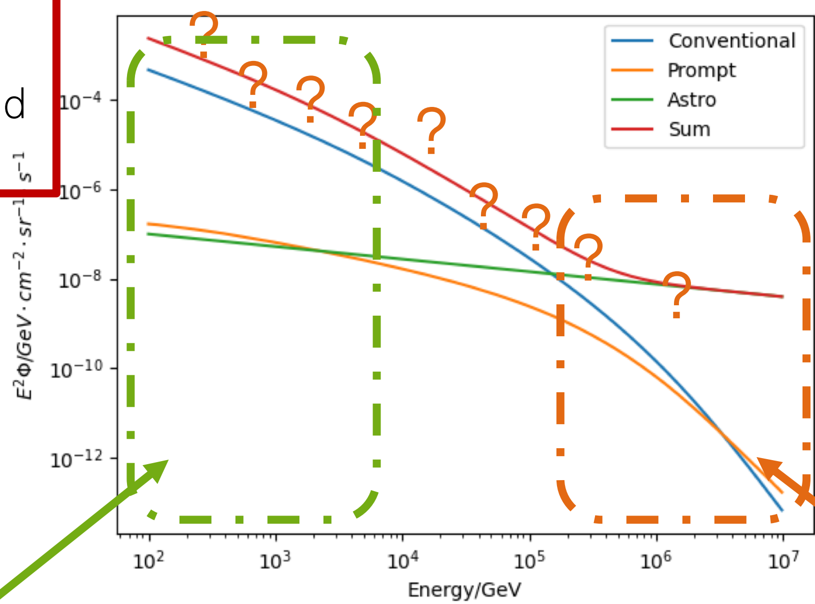
The analysis goal is to reconstruct a muon neutrino energy spectrum and to observe a flattening of the flux at high energies.

We can probably tolerate some muon backround in this region.

Muon background in this region will alter the physical interpretation of the result.

What to Consider Bevore Applying Additional Cuts

The caveat is that there might not be sufficient background simulation.

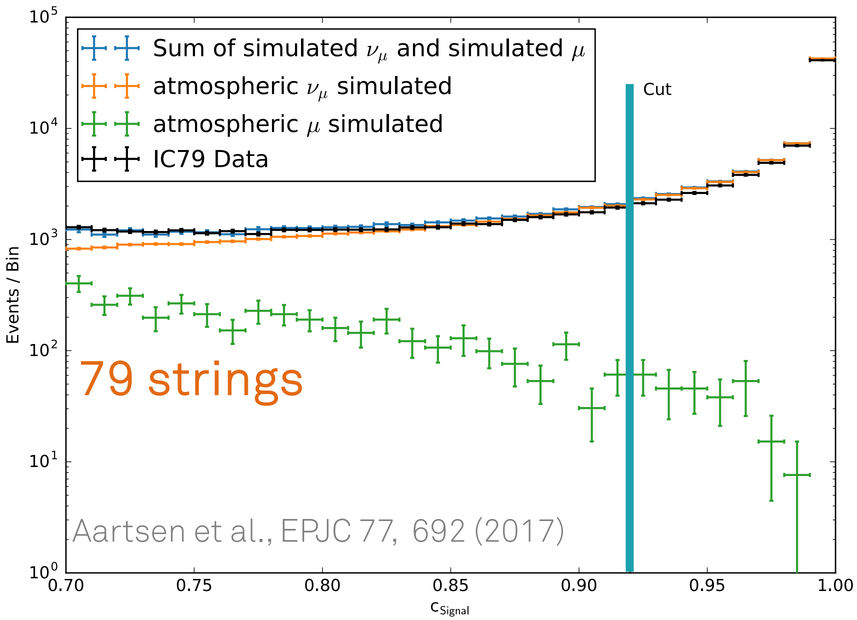


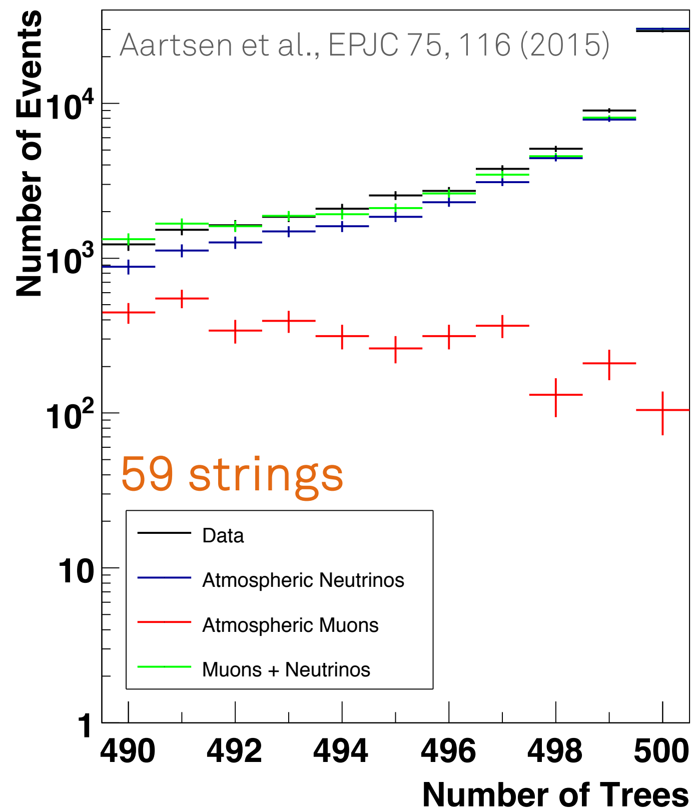
The analysis goal is to reconstruct a muon neutrino energy spectrum and to observe a flattening of the flux at high energies.

We can probably tolerate some muon backround in this region.

Muon background in this region will alter the physical interpretation of the result.

Additional Confidence Cut



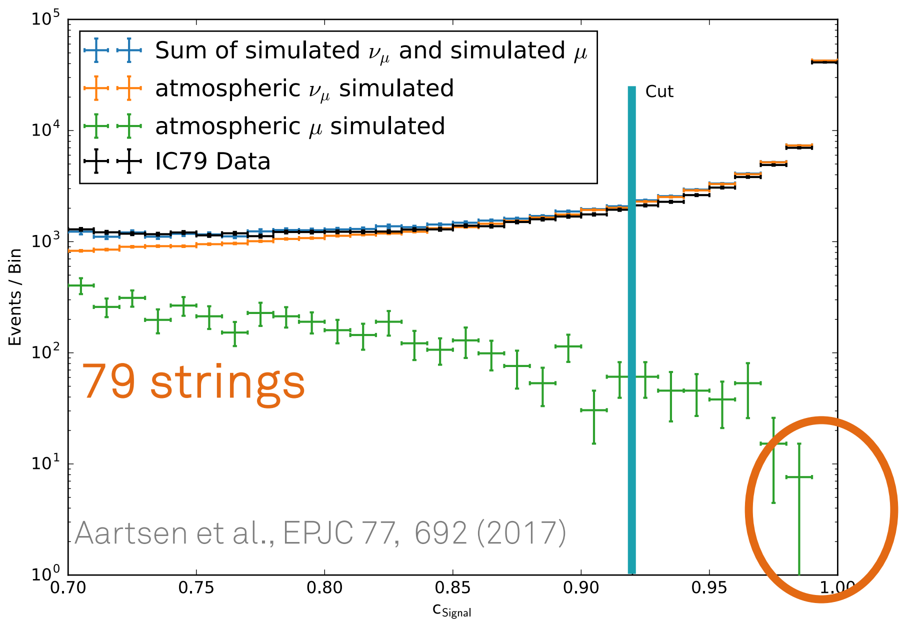


~ 200 neutrino candidates per day

~ 80 neutrino candidates per day

Expected Purity well above 99.5% for both analyses.

The Point of Cross Validation Exemplified



A not so untypical siuation:

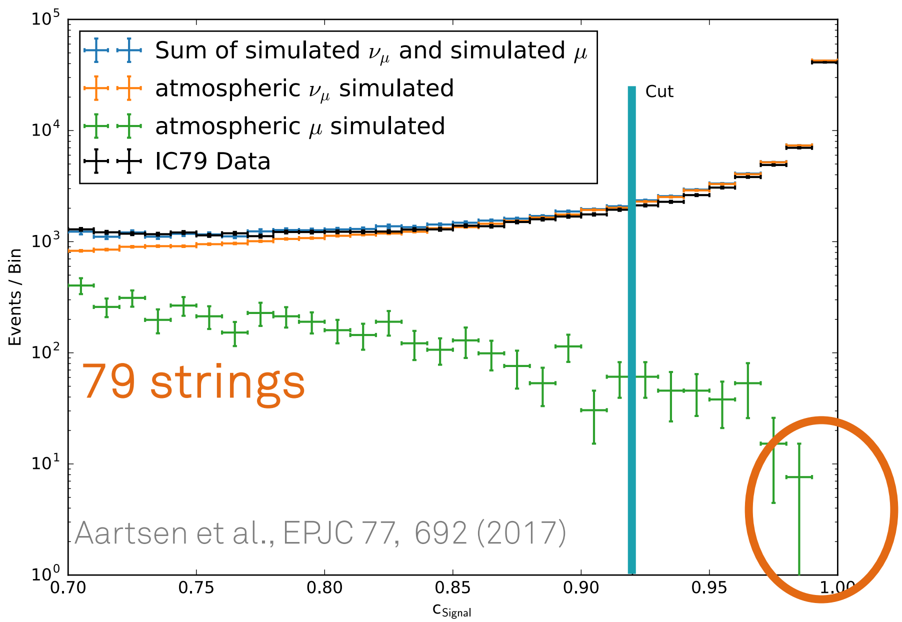
Signal	Backgr.	Total	Data
4	1	5	10

For very high confidence scores we run out of background simulation!!!

Is this a problem???

This might be a fluctuation or it might be that the classifier does not generalize well.

The Point of Cross Validation Exemplified

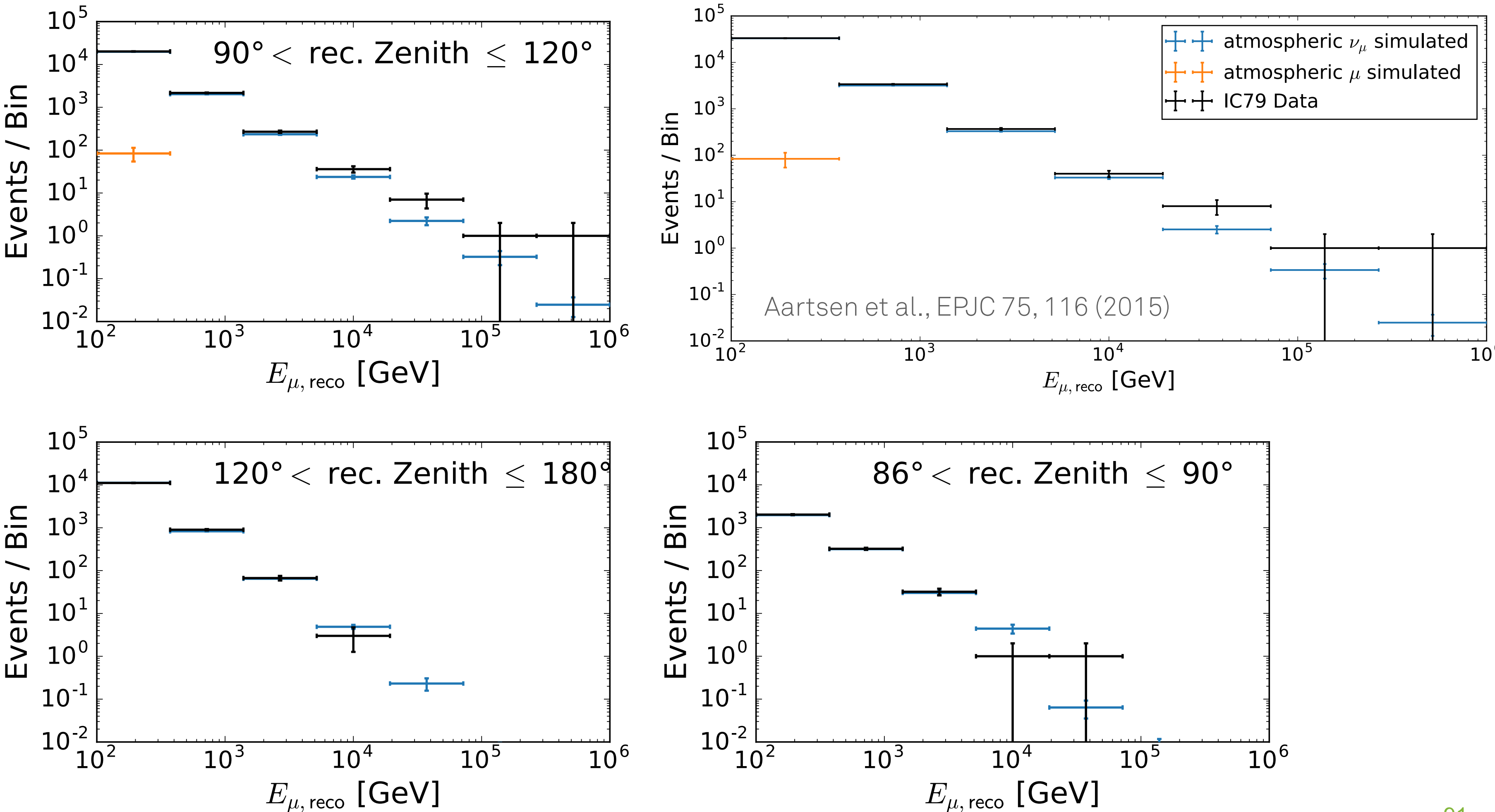


For very high confidence scores we run out of background simulation!!!

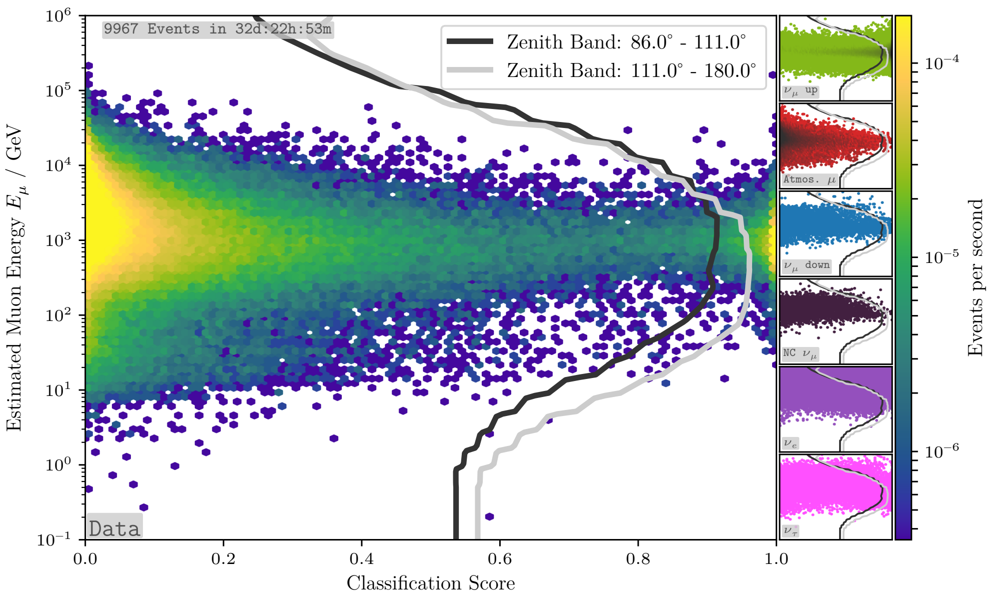
It.	Signal	Backgr.	Total
1	4	1	5
2	3	0	3
3	6	2	8
4	4	0	4
5	5	1	6

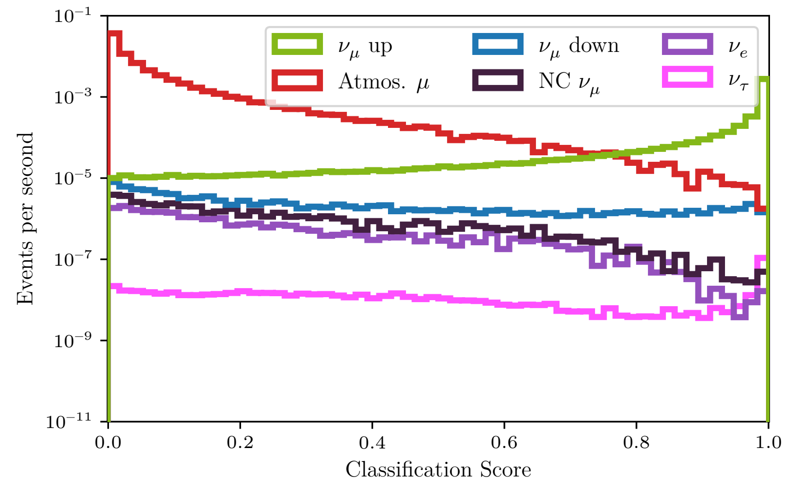
Using cross validtion will not make this problem disappear, but it will provide you with more information.

Where ist the Muon Background?



Additional Improvements: 2D Cuts





~ 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)

How to continue

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

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

Fredholm Integral equation

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

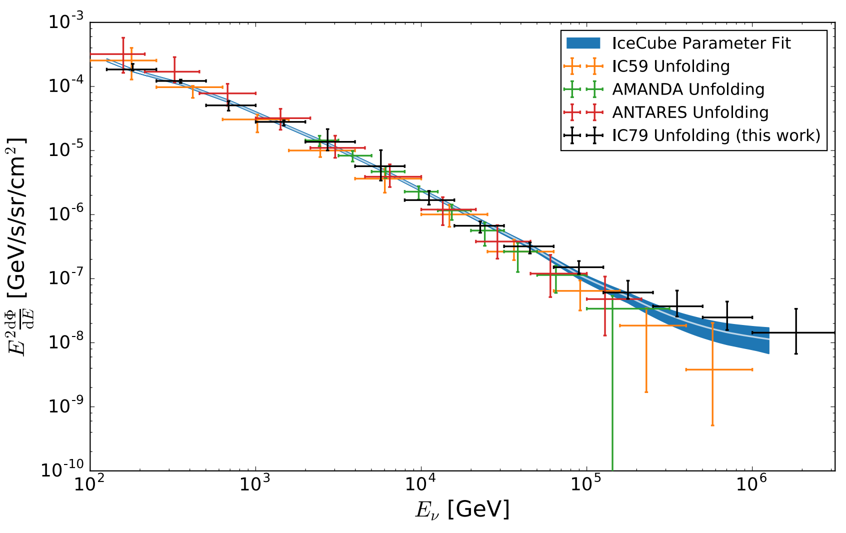
Matrix equation

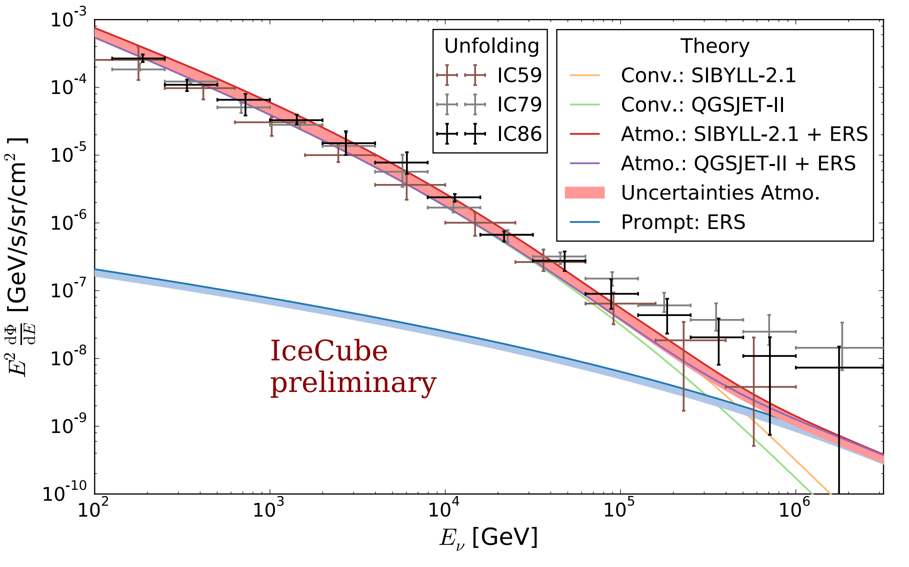
See for example: https://sfb876.tu-dortmund.de/deconvolution/index.html

Sought-after neutrino spectrum

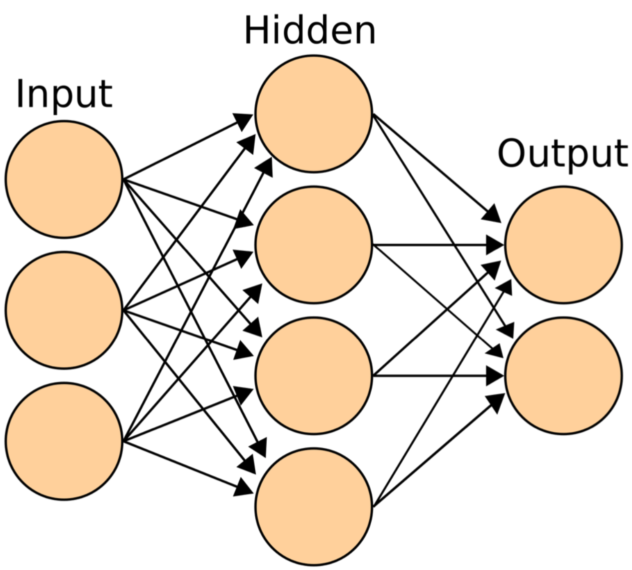
Obtained from simulations

(Atmospheric) Neutrino Energy Spectra





Deep Neural Networks for Reconstruction in IceCube

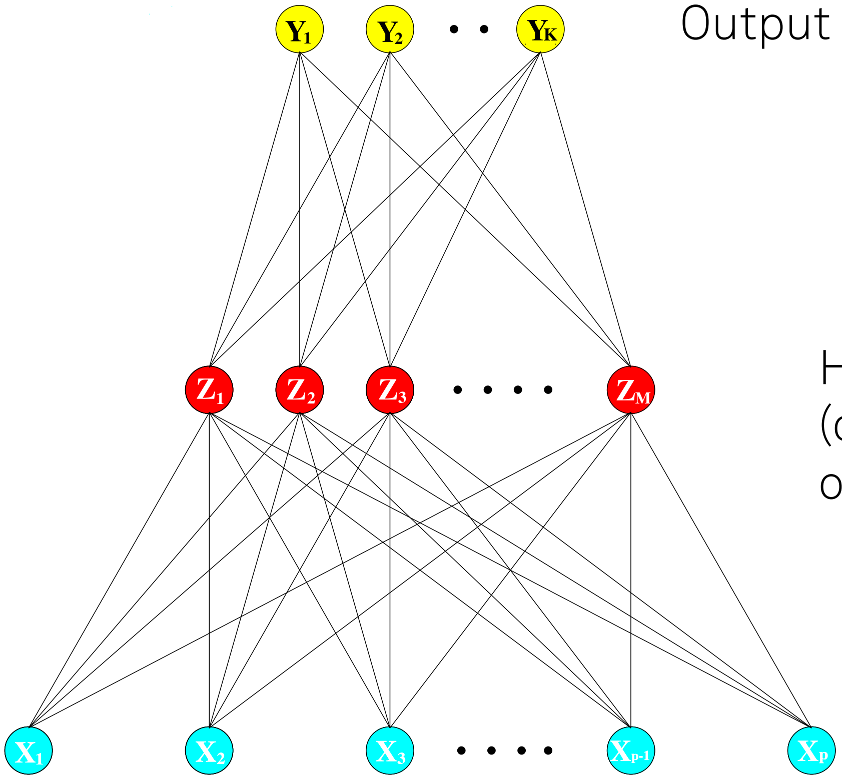


Source: By en:User:Cburnett - This W3C-unspecified vector image was created with Inkscape., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1496812

Source: By Alvesgaspar - Top left: File:Cat August 2010-4.jpg by AlvesgasparTop middle: File:Gustav chocolate.jpg by Martin BahmannTop right: File:Orange tabby cat sitting on fallen leaves-Hisashi-01A.jpg by HisashiBottom left: File:Siam lilacpoint.jpg by Martin BahmannBottom middle: File:Felis catus-cat on snow.jpg by Von.grzankaBottom right: File:Sheba1.JPG by Dovenetel, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17960205

Highly successful in many applications, including image classification.

Neural Network Basics: The General Idea



Output layer

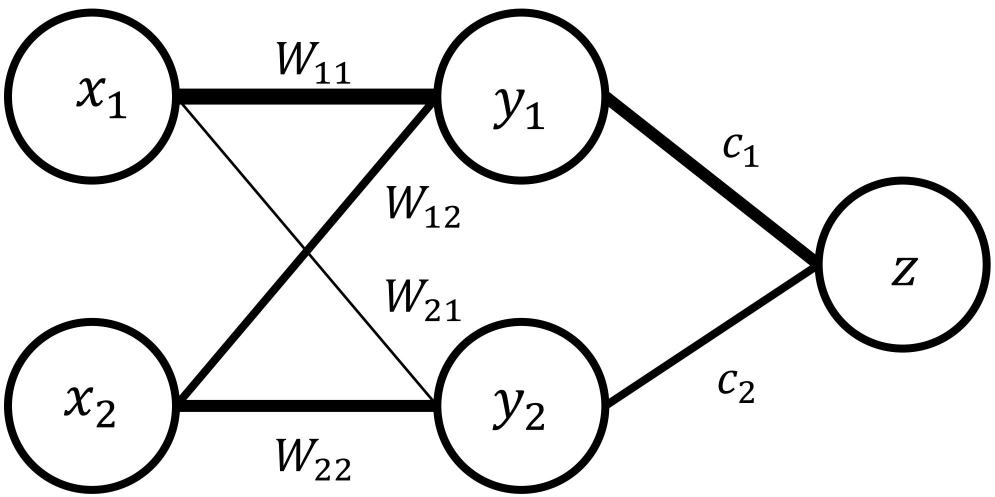
Hidden layer (cannot be observed directly)

Input layer

Originally developed to mimic the human brain.

Nodes are sometimes called neurons, and connections are sometimes called synapses.

Neural Network Basics: The General Idea



Nodes are linear combinations of nodes from previous layers.

The task is to optimize the weights such that the estimated label z matches the true label $\hat{z}$.

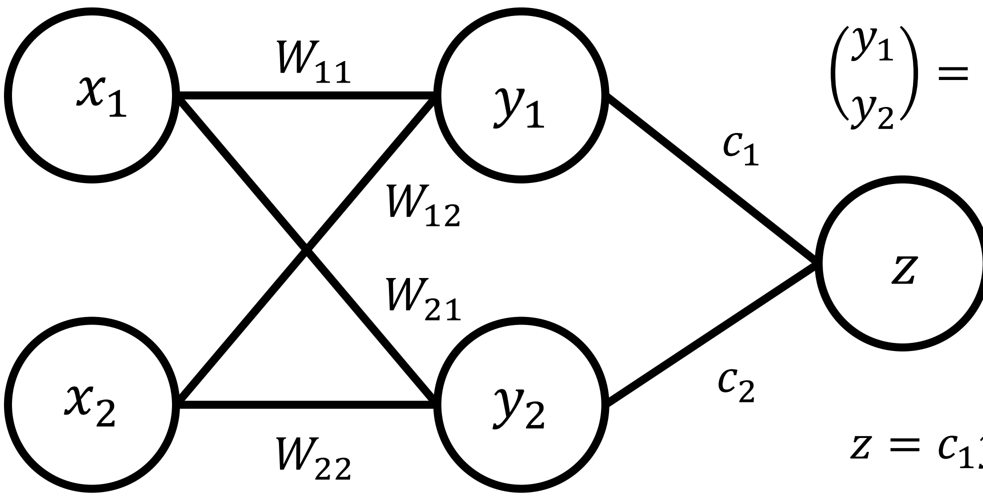
A feed-forward neural network with linear output and at least one hidden layer with a finite number of nodes can approximate any of the above* functions with arbitrary precision**.

*Continuous functions on closed bounded subsets of the Eucledian space $\mathbb{R}^n$.

**Figure and definition adopted from Erdmann et al.

Neural Network Basics: More mathematically speaking

$$y_1 = W_{11}x_1 + W_{12}x_2 + b_1$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{11} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

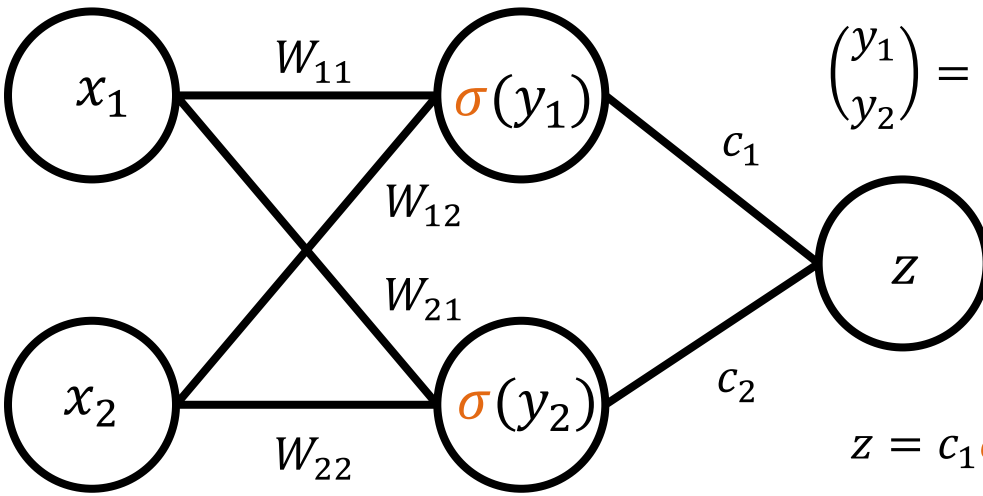
$$z = c_1 y_1 + c_2 y_2 + d$$

$$y_2 = W_{21}x_1 + W_{22}x_2 + b_2$$

**Figure adopted from Erdmann et al.

Neural Network Basics: Adding Non-Linearity

$$y_1 = W_{11}x_1 + W_{12}x_2 + b_1$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{11} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

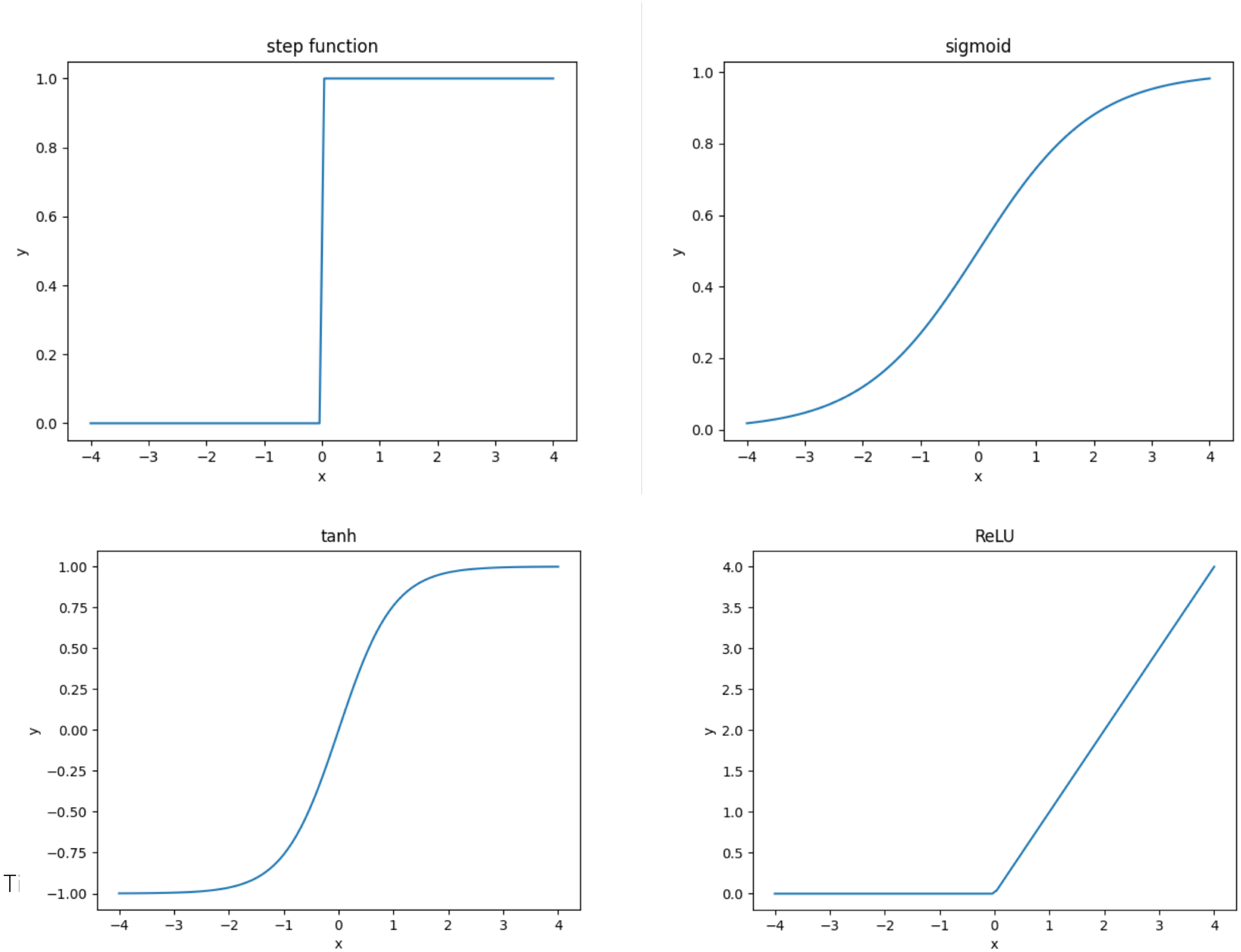
$$z = c_1\sigma(y_1) + c_2\sigma(y_2) + d$$

$$y_2 = W_{21}x_1 + W_{22}x_2 + b_2$$

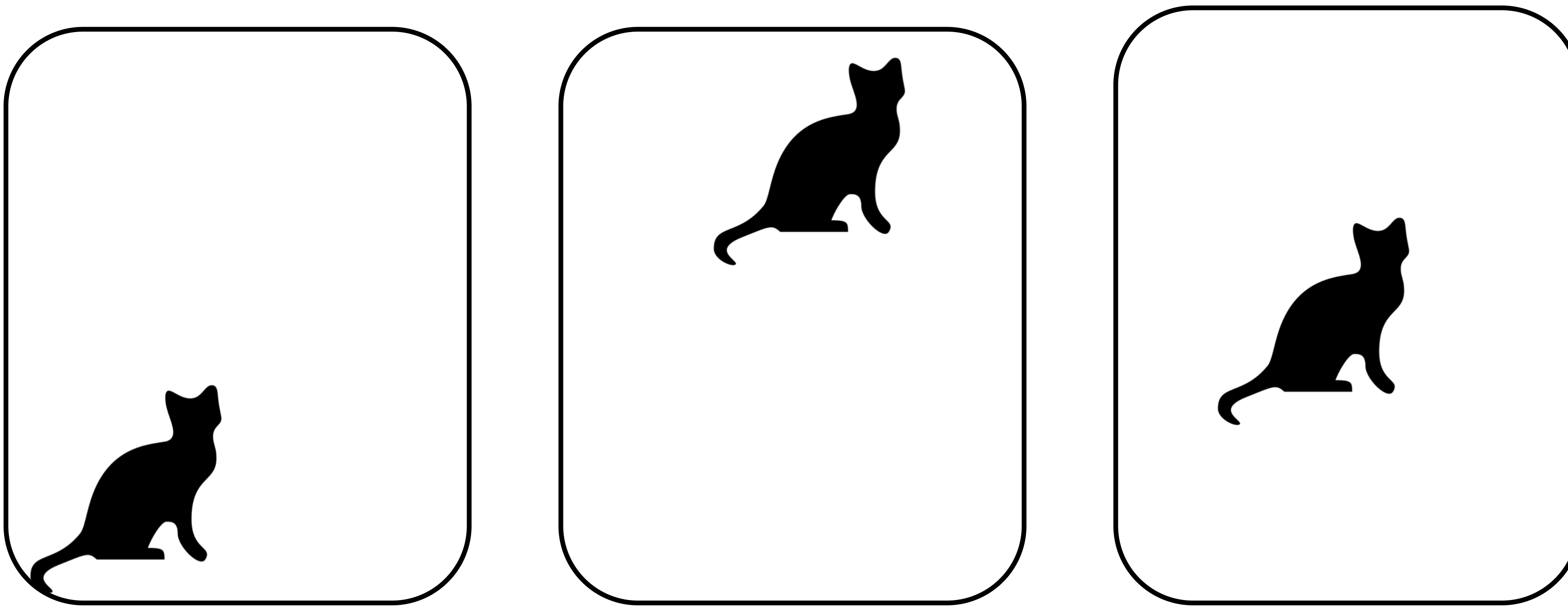
**Figure adopted from Erdmann et al.

σ is generally referred to as the activation function.

Some Popular Activation Functions



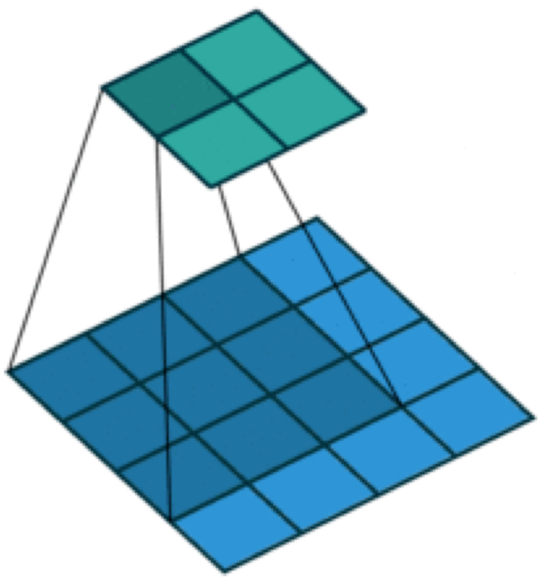
Deep Neural Networks can Exploit Spatial Invariance



A Deep Neural Network will classify this as a cat independent of its position in the picture.

The phyics of a neutrino interaction is also spatially invariant.

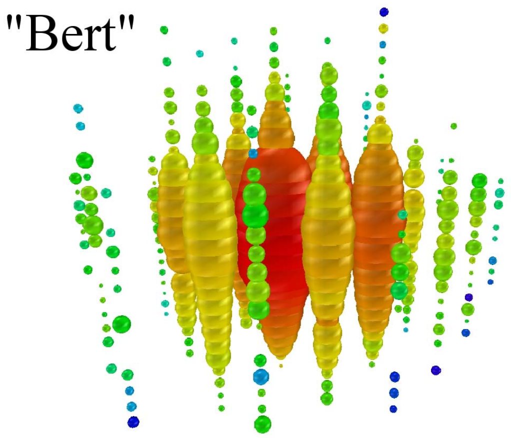
Convolutional Layers



- Considering all pixels in an image in a fully connected network, results in too many parameters to be optimized
- The position of an object in an image should not alter the prediction (translational invariance)
- The convolutional operation exploits the neighbourhood of each pixel

Source: By Vincent Dumoulin, Francesco Visin - https://github.com/vdumoulin/conv_arithmetic, MIT, https://commons.wikimedia.org/w/index.php?curid=78003423

Challenges in Cascade Reconstruction



- Missing lever arm due to spherical light distribution
- Local events and therefore more susceptible to ice properties
- Quantities of interest: Deposited energy and direction of incoming neutrino

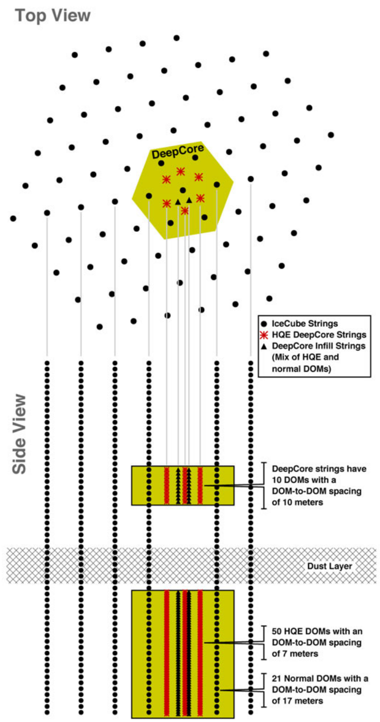
Why cascades?

- About 2/3 of IceCube's HESE starting events are cascades
- Directional resolution is awful → huge potential for improvements

Reasons for Using Neural Networks in IceCube

- Improved reconstruction methods will lead to increased sensitivity for the detection of sources
- Hardware limitations at the South Pole
- Events need to be processed in a given time frame to prevent pileup
- Limitations call for robust method that can handle raw data in constant time
- Neural networks are compuattionally inexpensive once the network is trained
- Fixed amount of operations, runtime is (largely) independent of the input
- Translational invariance (position of the classified object does not impact the class)
- Physics of neutrino interaction is invariant in time and space

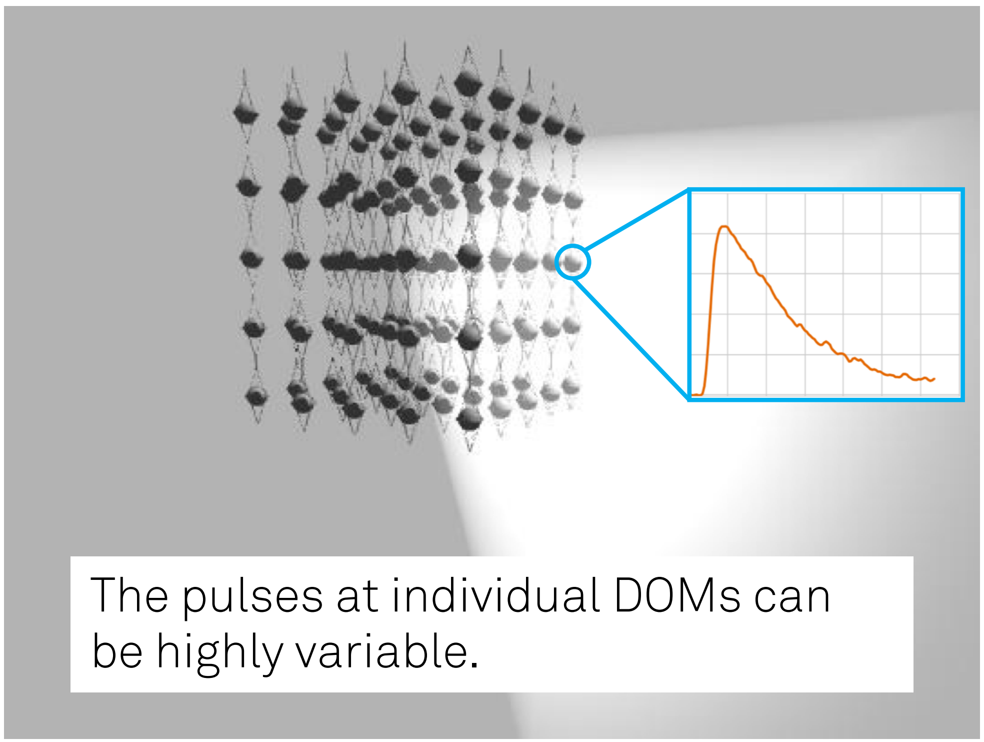
Difference Between IceCube and Images



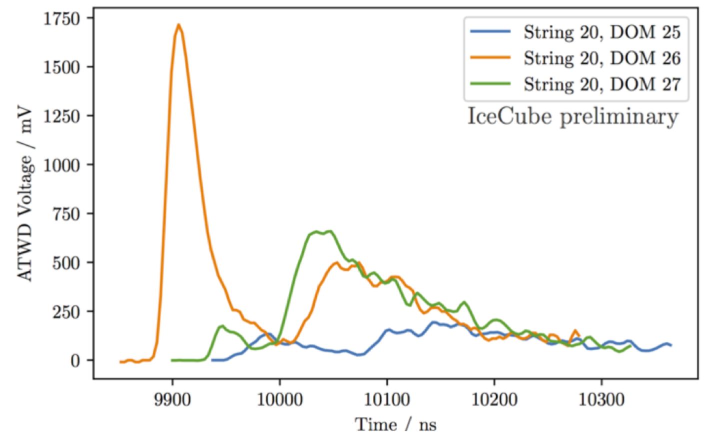
Source: By Alvesgaspar - Top left: File:Cat August 2010-4.jpg by AlvesgasparTop middle: File:Gustav chocolate.jpg by Martin BahmannTop right: File:Orange tabby cat sitting on fallen leaves-Hisashi-01A.jpg by HisashiBottom left: File:Siam lilacpoint.jpg by Martin BahmannBottom middle: File:Felis catus-cat on snow.jpg by Von.grzankaBottom right: File:Sheba1.JPG by Dovenetel, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17960205

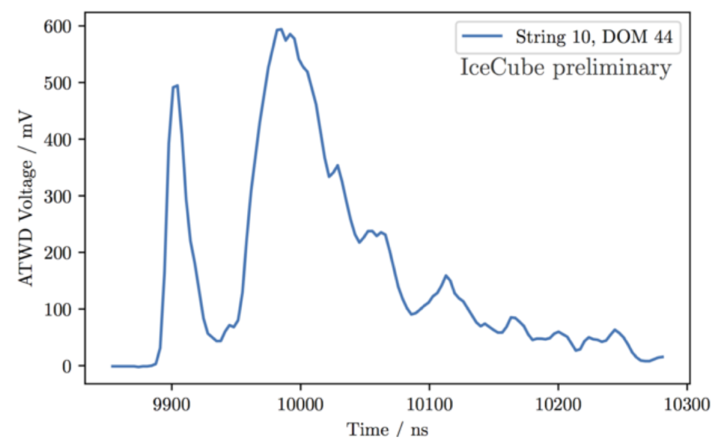
Key challenges: hexagonal grid, high dimensionality and variability

IceCube Pulses



The pulses at individual DOMs can be highly variable.

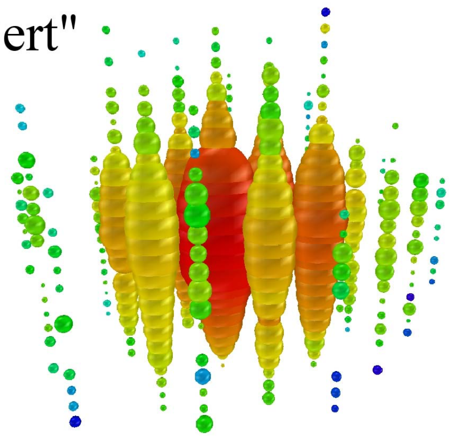




Defining the Task

Energy???

"Bert"



Azimuth???

Zenith???

Decomposed into directional unit vector

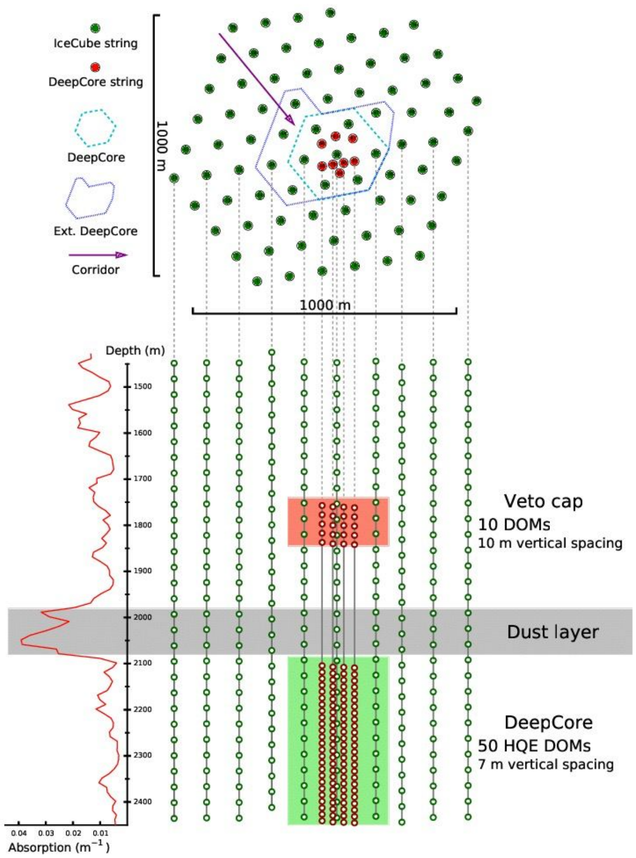
$$\vec{d} = (d_x, d_y, d_z)$$

Mainly due periodicity of the azimuth

Angular Uncertainty???

This is a multi-label regression task!

A Closer Look at DeepCore

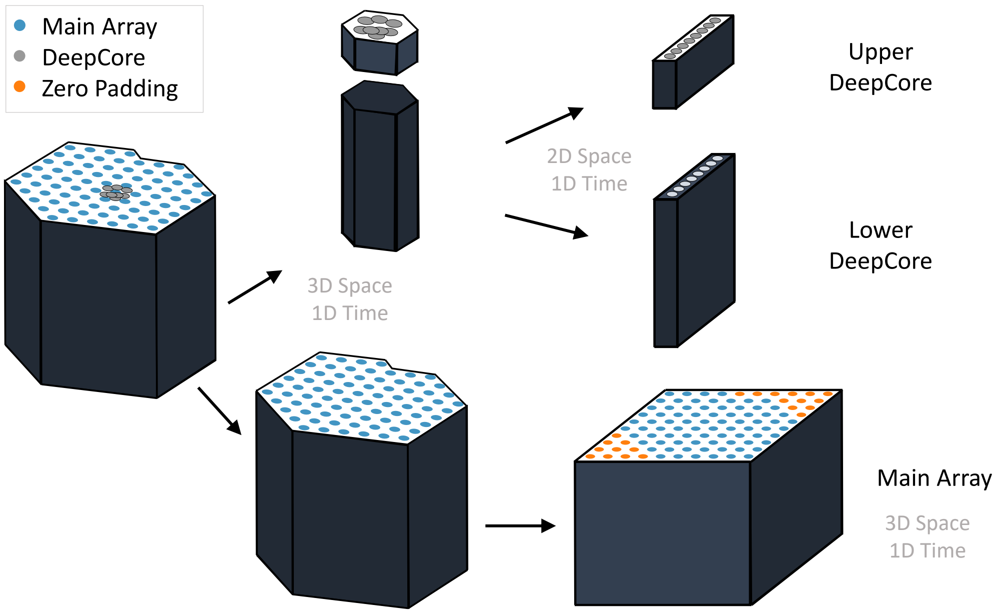


The DeepCore Sub-Array consists of two parts:

- 10 DoMs with 10 m vertical spacing, upper part (red), veto cap
- 50 HQE DOMs with 7 m vertical spacing, Deep Core

Horizontal and vertical spacings differ from each other and from the main array → Needs to be considered in the DNN!

Handling the Sub-Arrays



(8 x 10 x n)

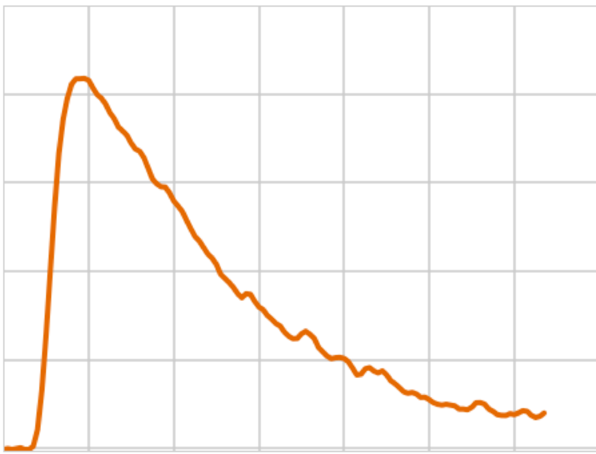
(8 x 50 x n)

(10 x 10 x 60 x n)

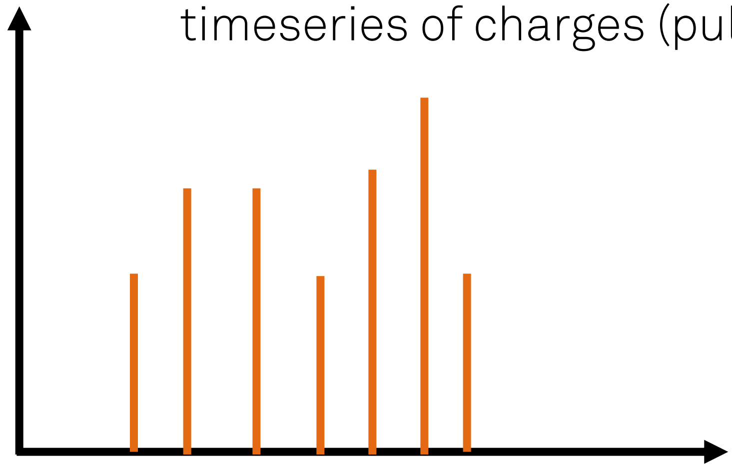
Abbasi et al., JINST, 16 (7) (2021).

Allows for convolution over z-dimension for DeepCore and over all spatial dims for the main array.

(Im)Possibility of Discretizing Pulses



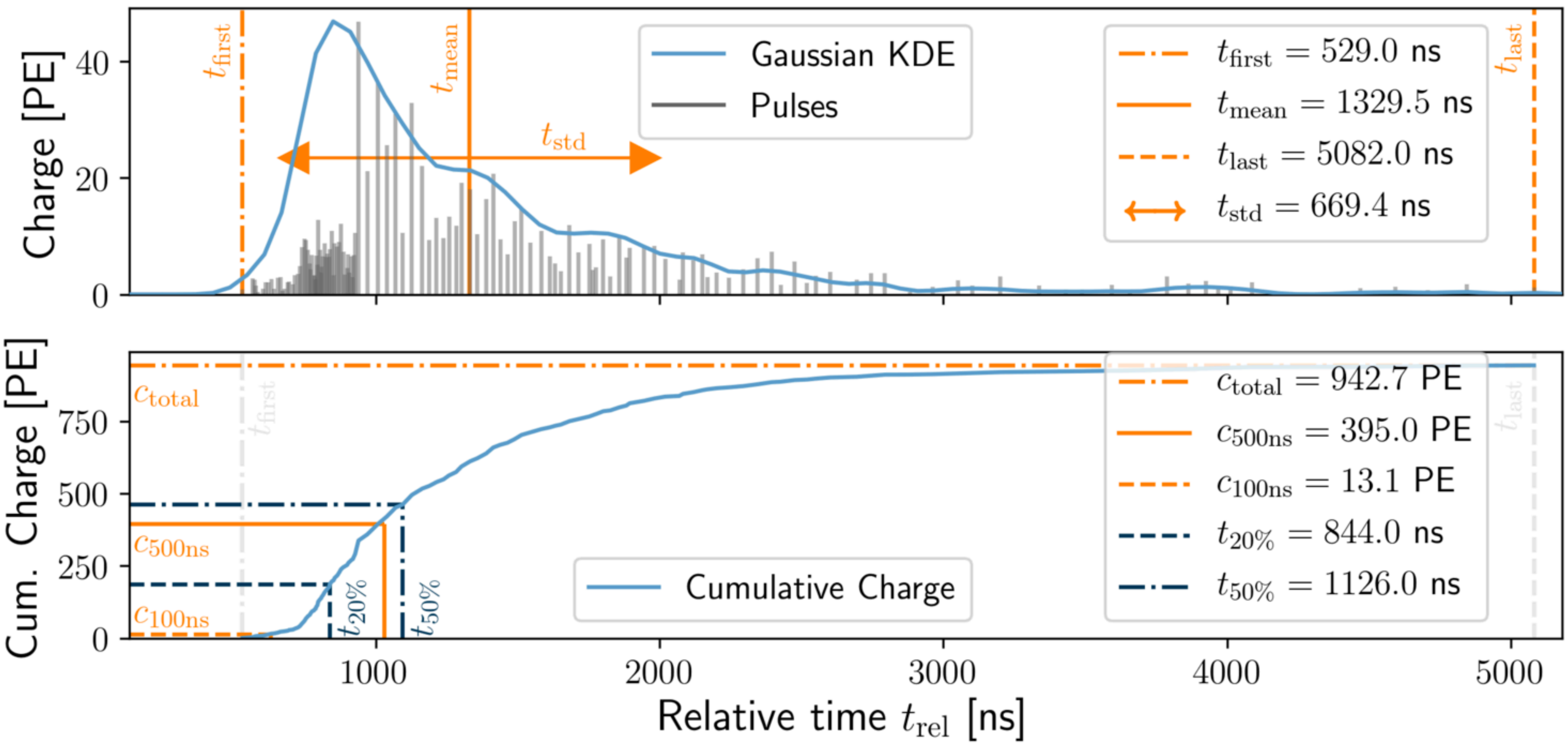
Waveform is transformed into timeseries of charges (pulses).



Number of pulses per DOM is highly variable, but NN requires uniform and constant input size.

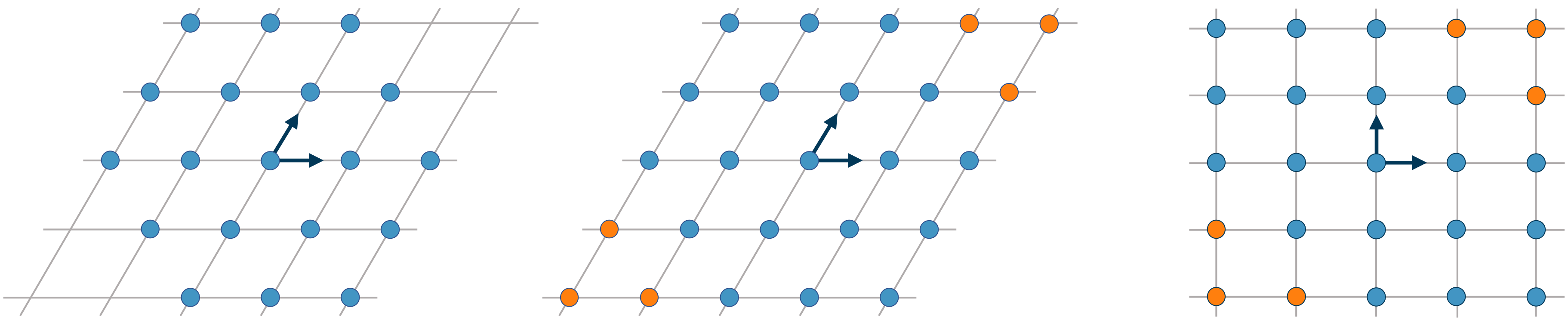
Binning requires thousands of bins to achieve desired timing resolution. → Infeasible due to computational complexity.

Input Variables



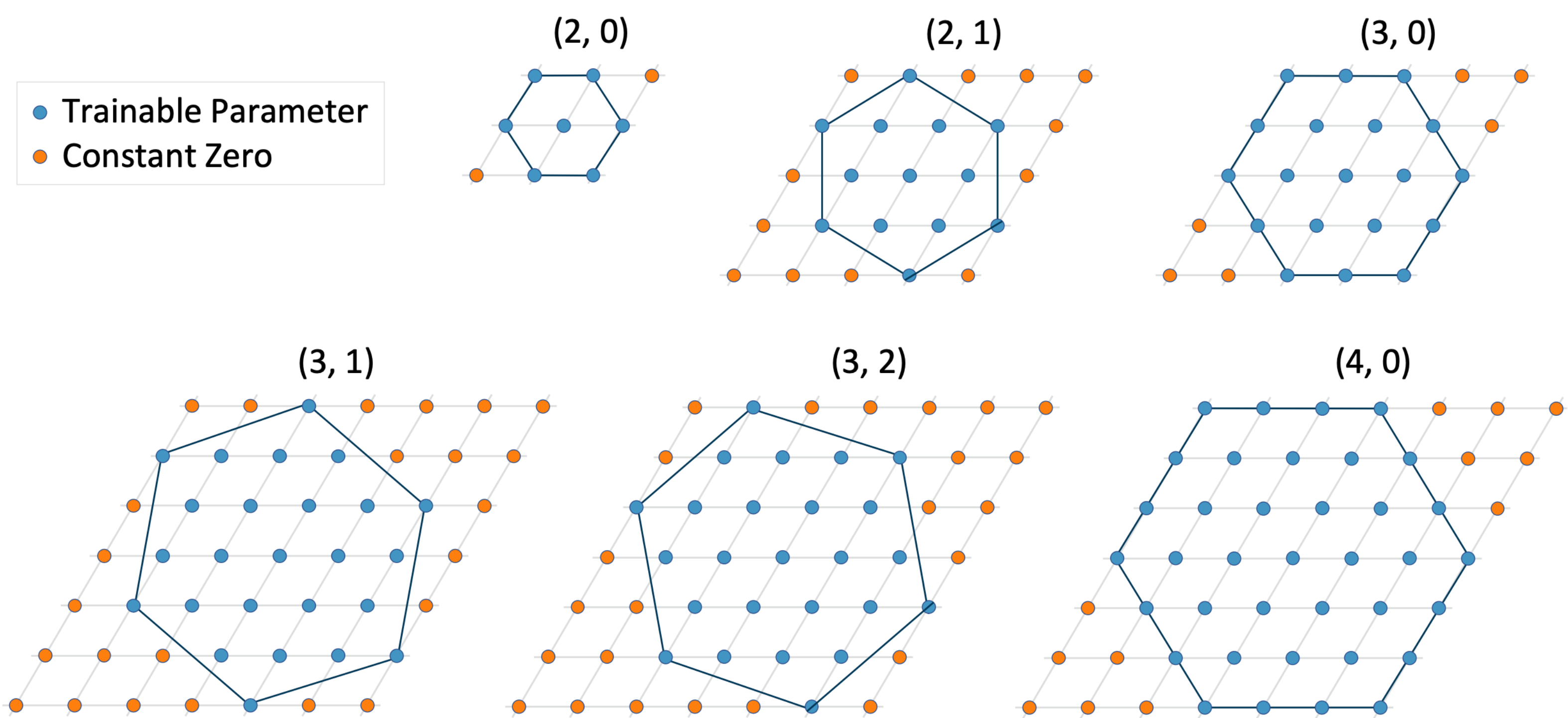
Abbasi et al., JINST, 16 (7) (2021).

Zero Padding



Abbasi et al., JINST, 16 (7) (2004)

Hexagonal Kernels



Preparing the Input

$$X' = \ln(1.0 + X)$$

$$Y' = \ln(1.0 + Y)$$

Applied to total charge and charge collected within the first 100 and 500 ns

$$X' = X$$

$$Y' = Y$$

All other features.

NNs can handle data of basically any scaling, but activation functions are typically centered around zero.

Preparing the Input

$$X'' = \frac{X' - \overline{X}'}{\sigma_{X'} + \epsilon}$$

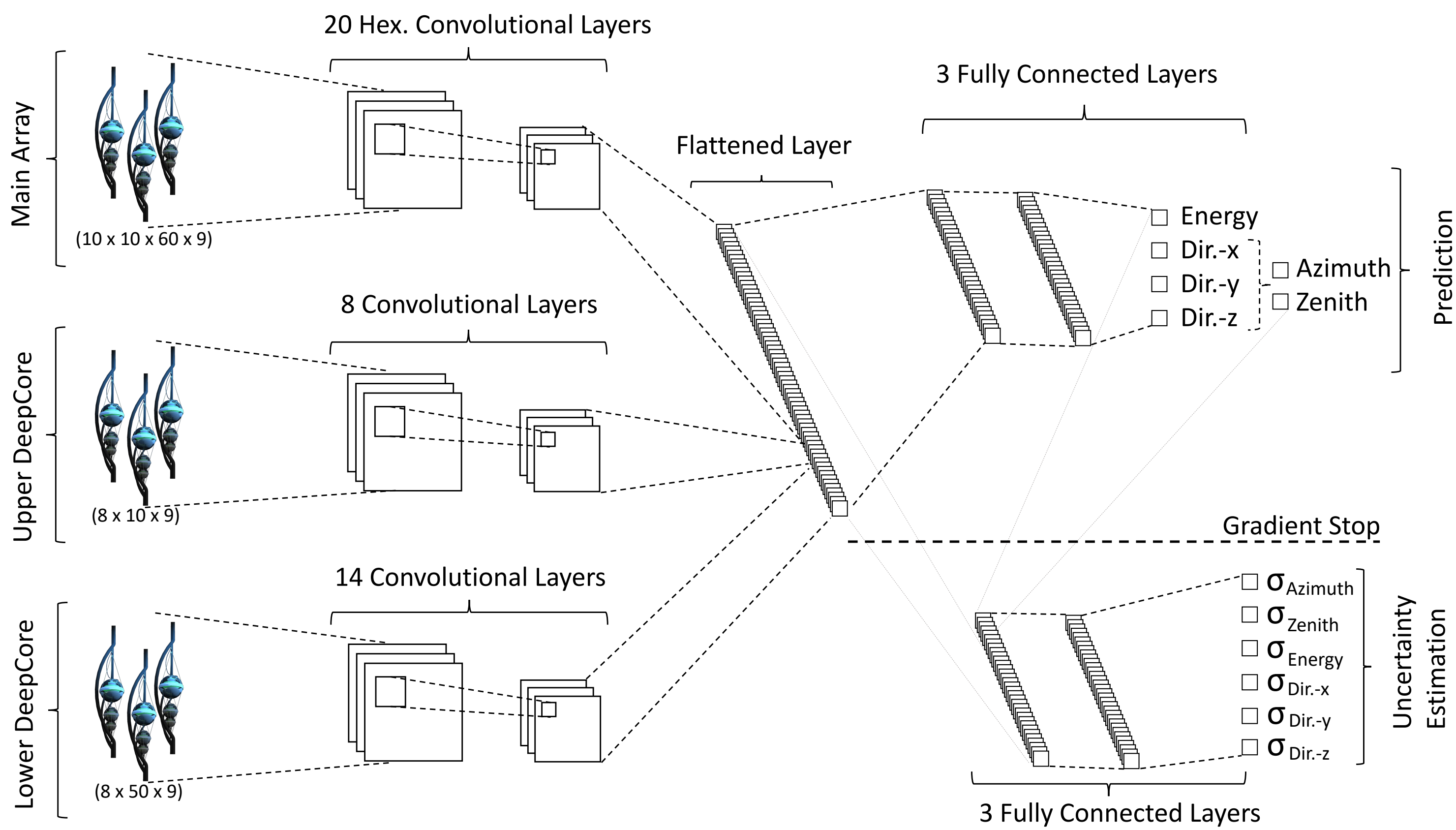
$$Y'' = \frac{Y' - \overline{Y}'}{\sigma_{Y'} + \epsilon}$$

Normalize input data and labels to zero mean and unit variance.

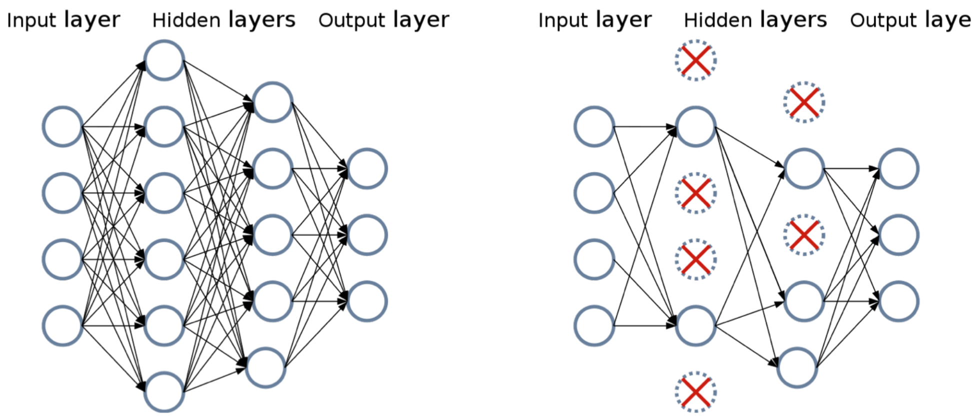
Small constante (10^{-4}) is added to prevent division by zero

NNs can handle data of basically any scaling, but activation functions are typically centered around zero.

Network Architecture



Regularization

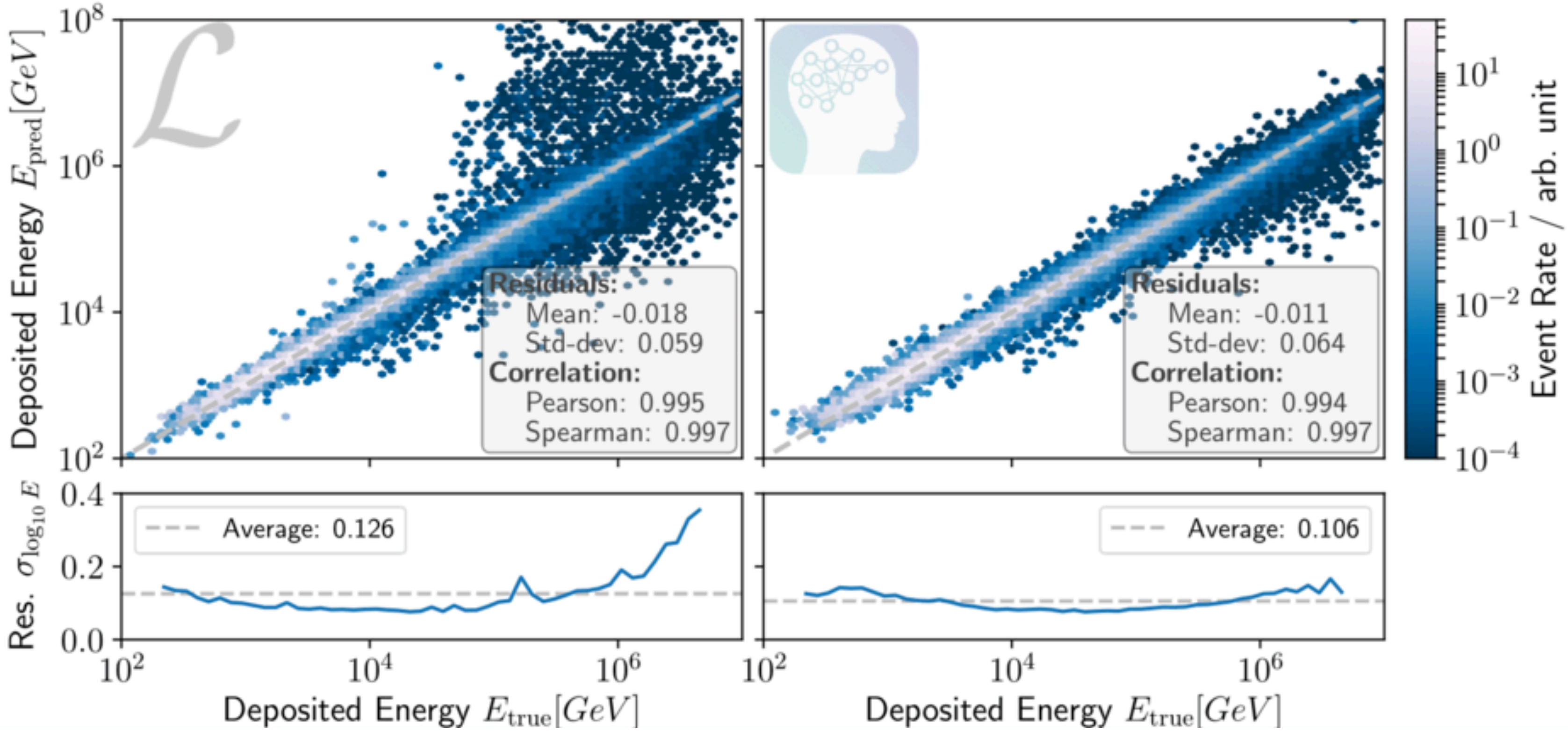


Source: By Mads Dyrmann - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=110373041

Regularization via Dropout, but gradually decreased at later layers.

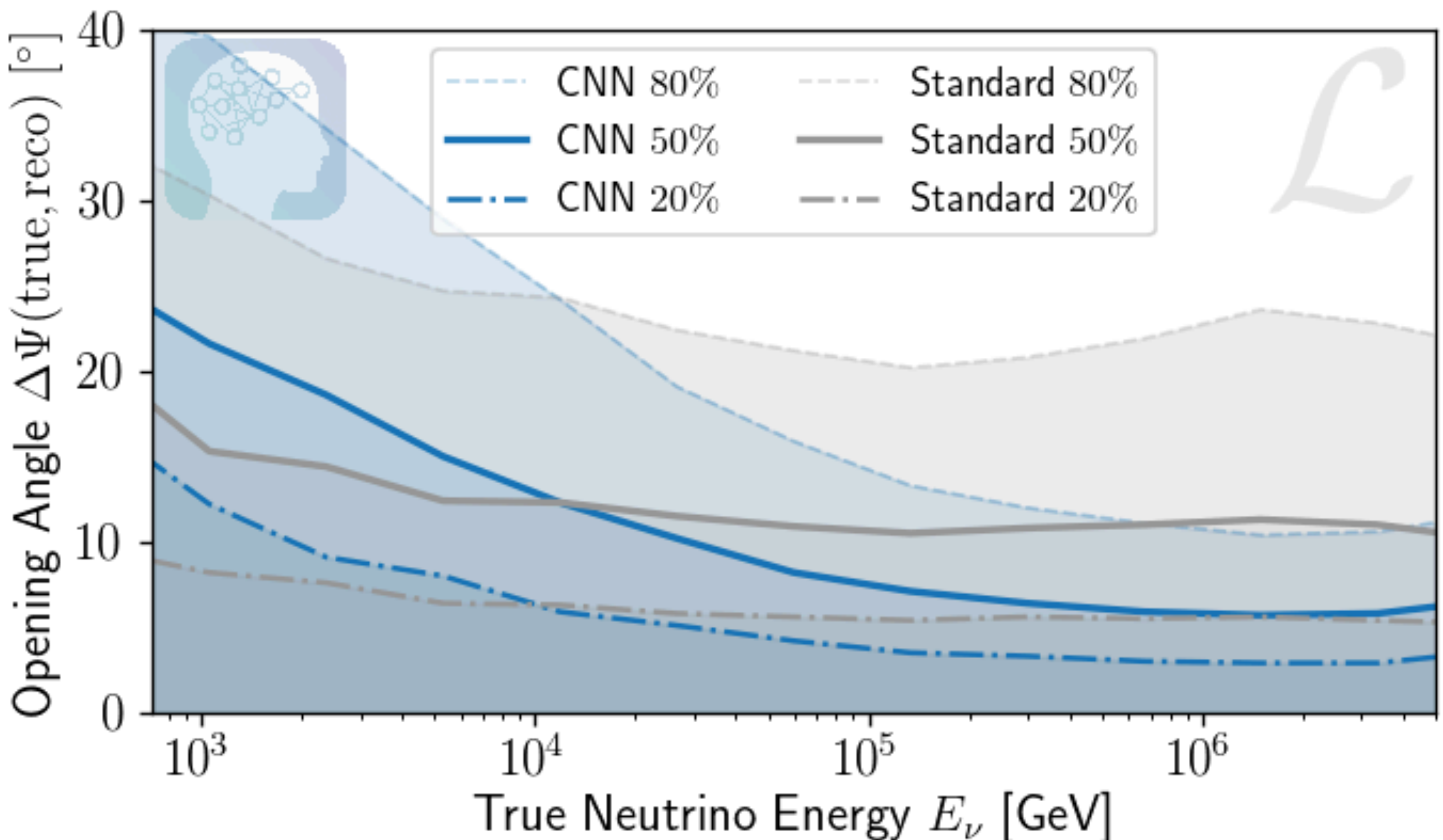
In addition, individual DOMs are randomly dropped for increased stability.

Energy Reconstruction Results



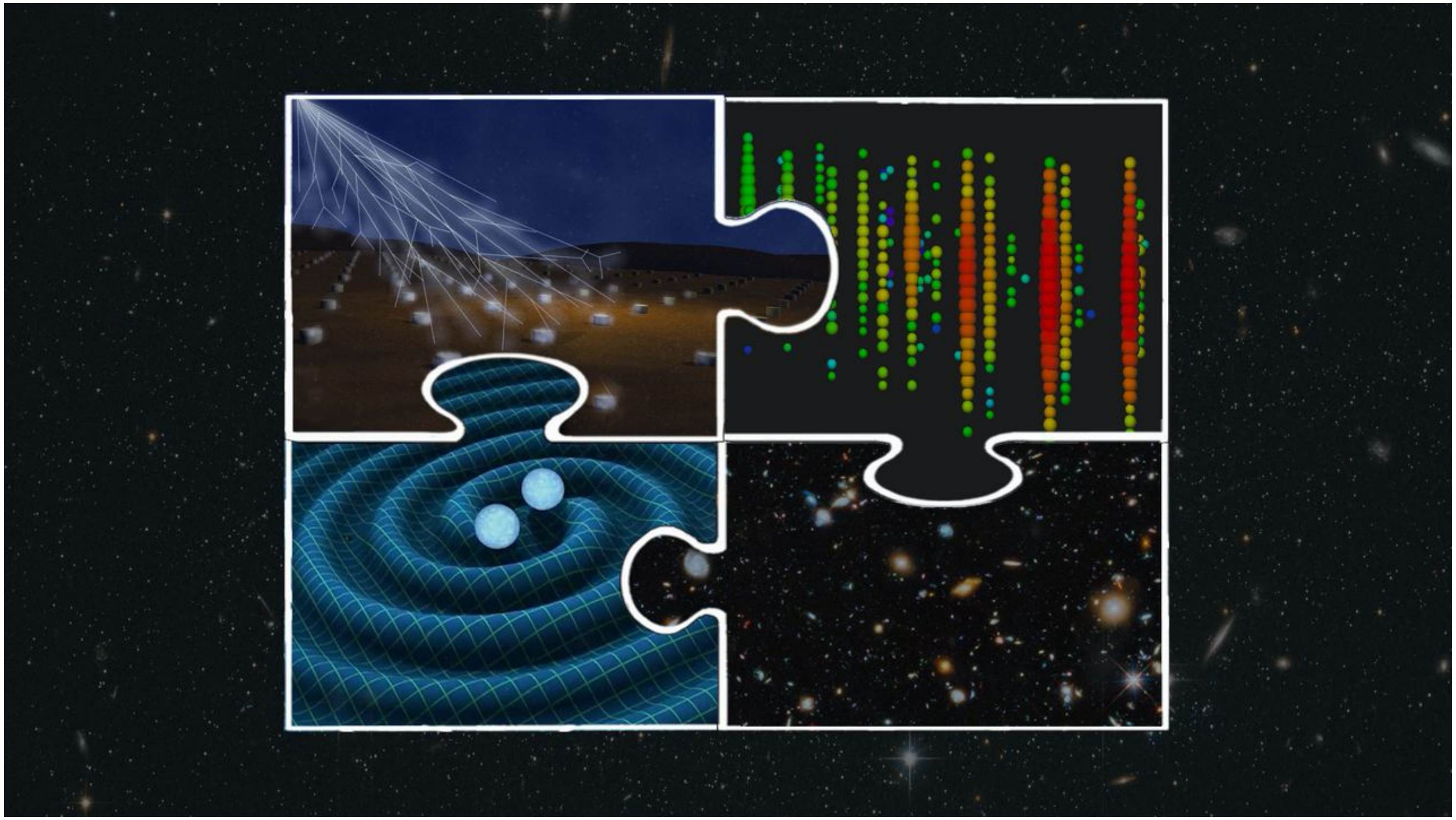
Abbasi et al., JINST, 16 (7) (2021).

Directional Reconstruction Results

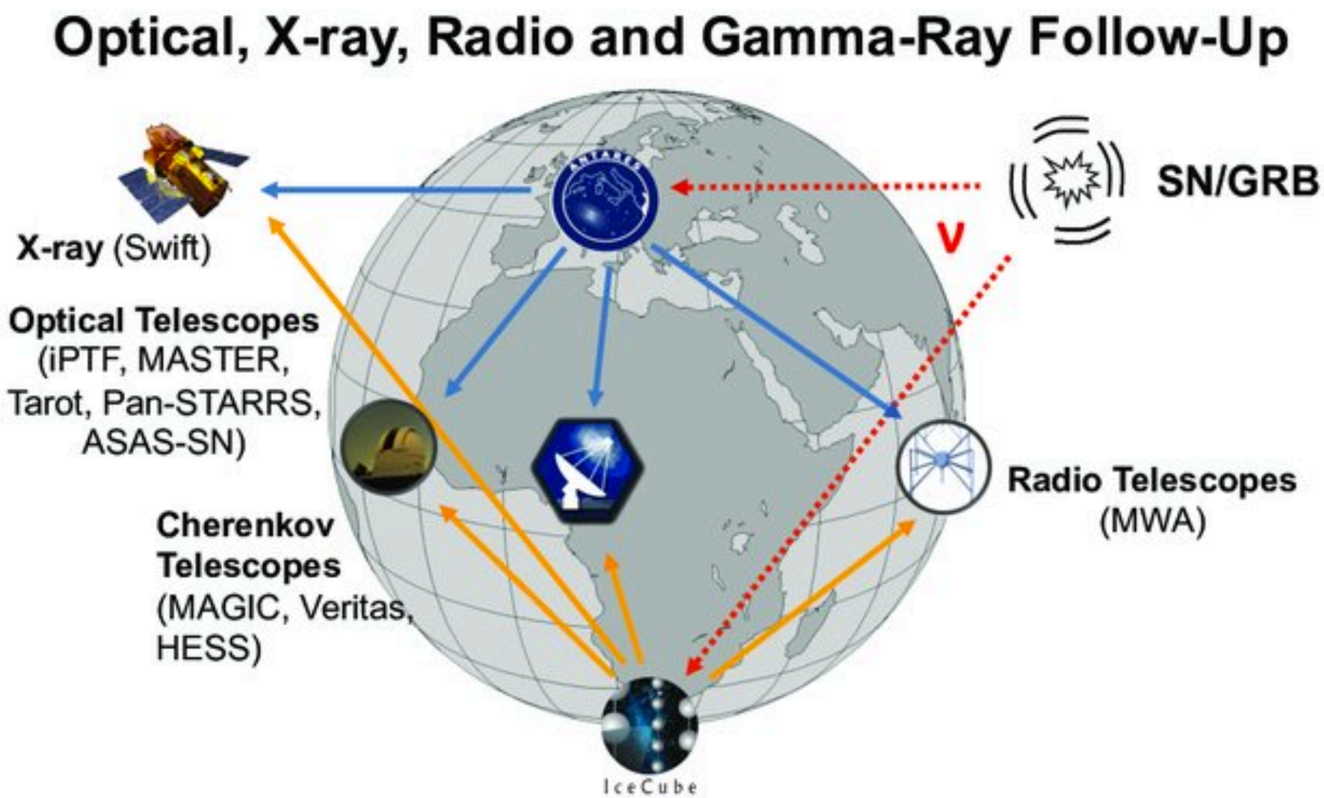


Abbasi et al., JINST, 16 (7) (2021).

The Need To Reconstruct the Uncertainty

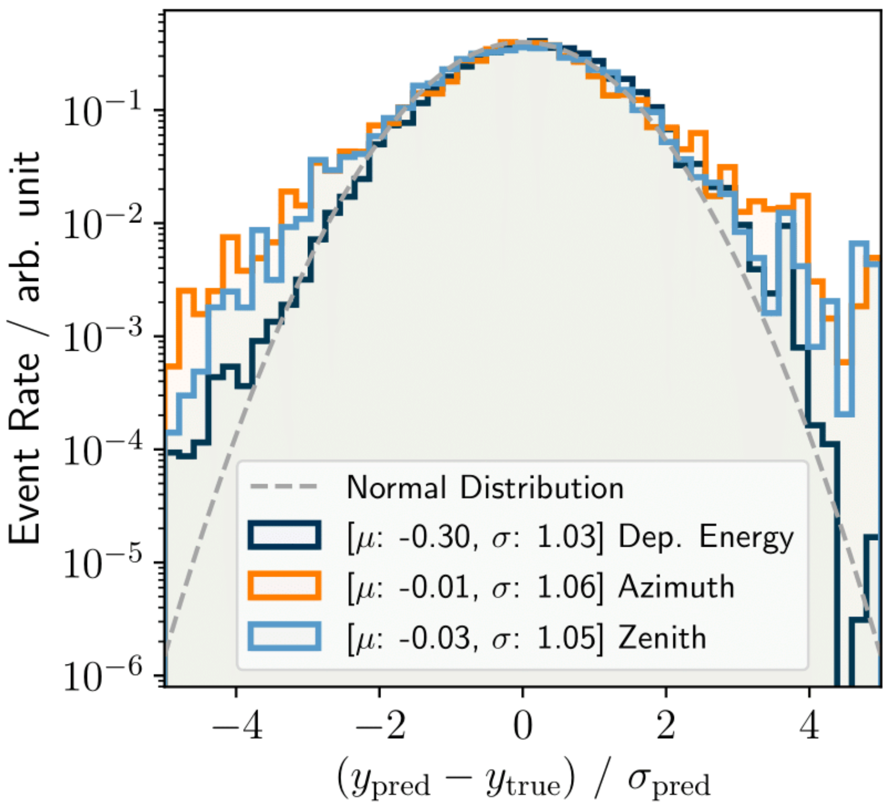


The Need to Reconstruct the Uncertainty



Franckowiak, Anna. "Multimessenger astronomy with neutrinos." *Journal of Physics: Conference Series*. Vol. 888. No. 1

Uncertainty Estimation



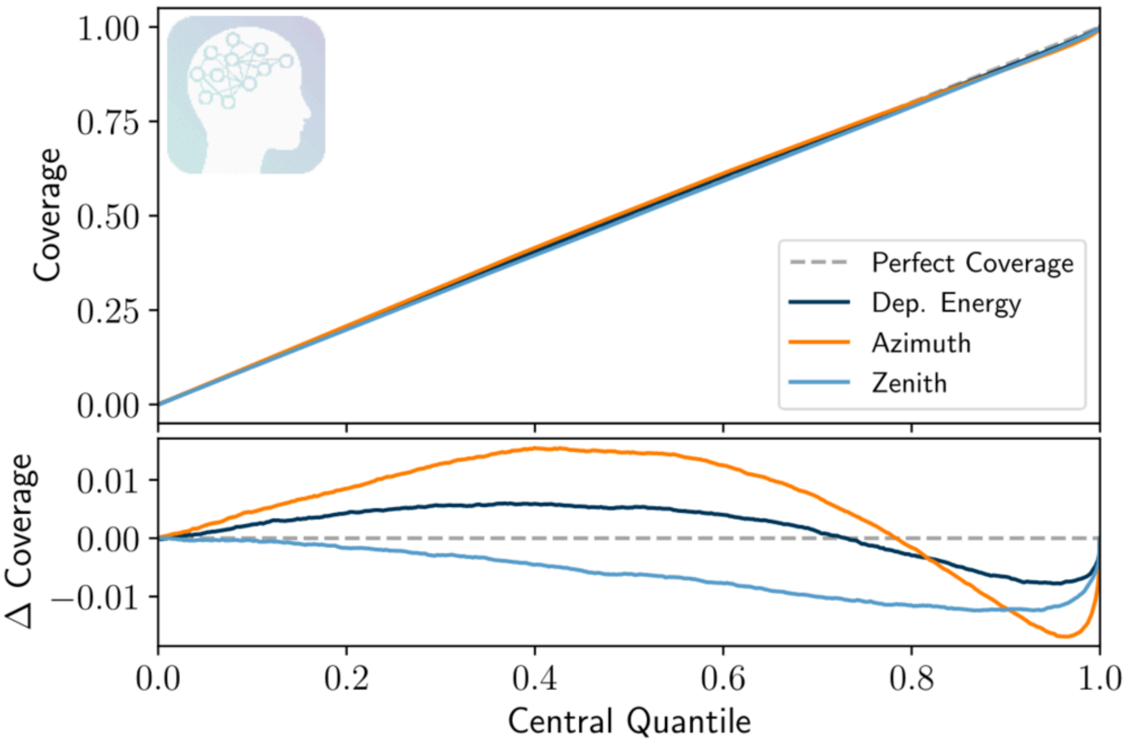
Works well in case the pulls follow a Gaussian distribution.

Pulls in this case are well described by a Gaussian distribution, except for the tails.

Deviations in the tails are driven by rare outlier events

This can likely be correctec by additional training iterations.

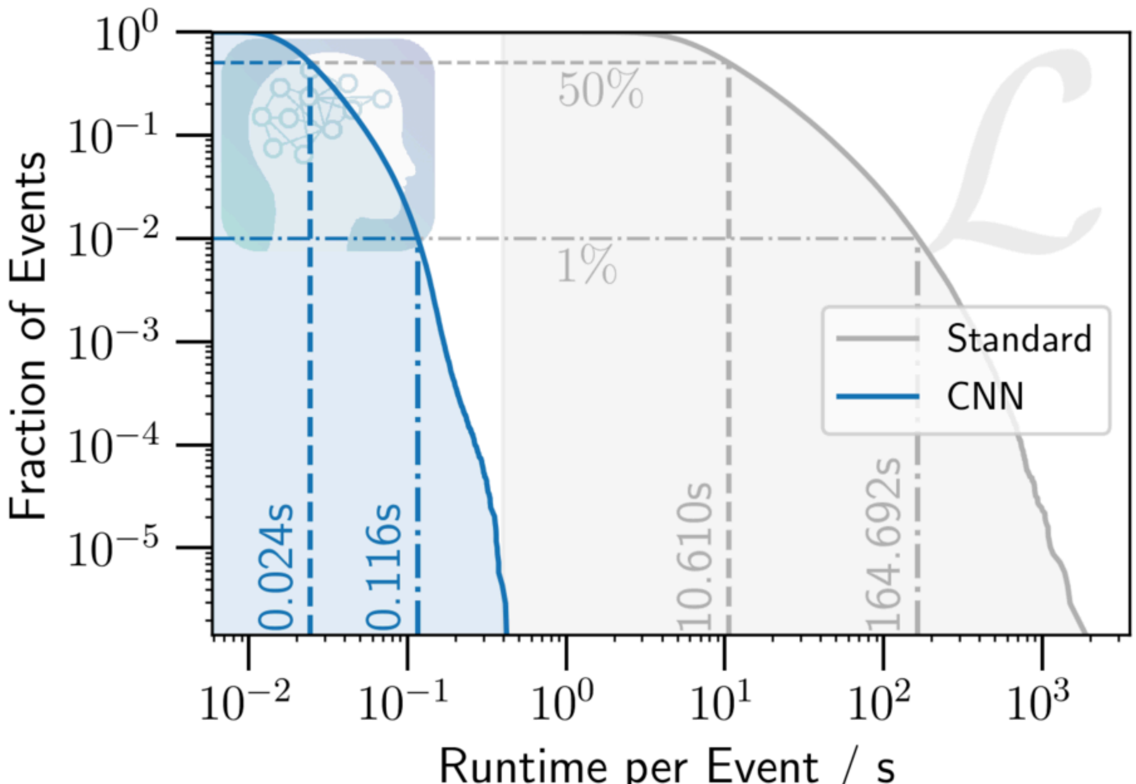
Uncertainty Estimation: Coverage Test



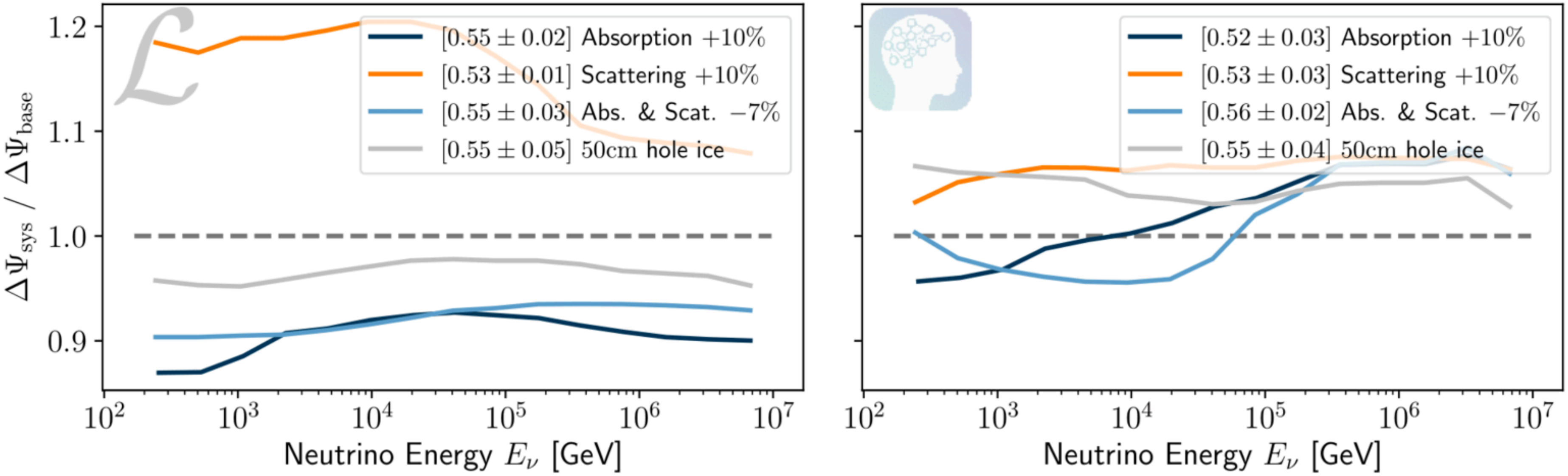
Based on assumption of Gaussian with width estimated by the network, one can compute the number of events within a certain quantile.

For perfect coverage this results in a 1:1 relationship when compared with actual results.

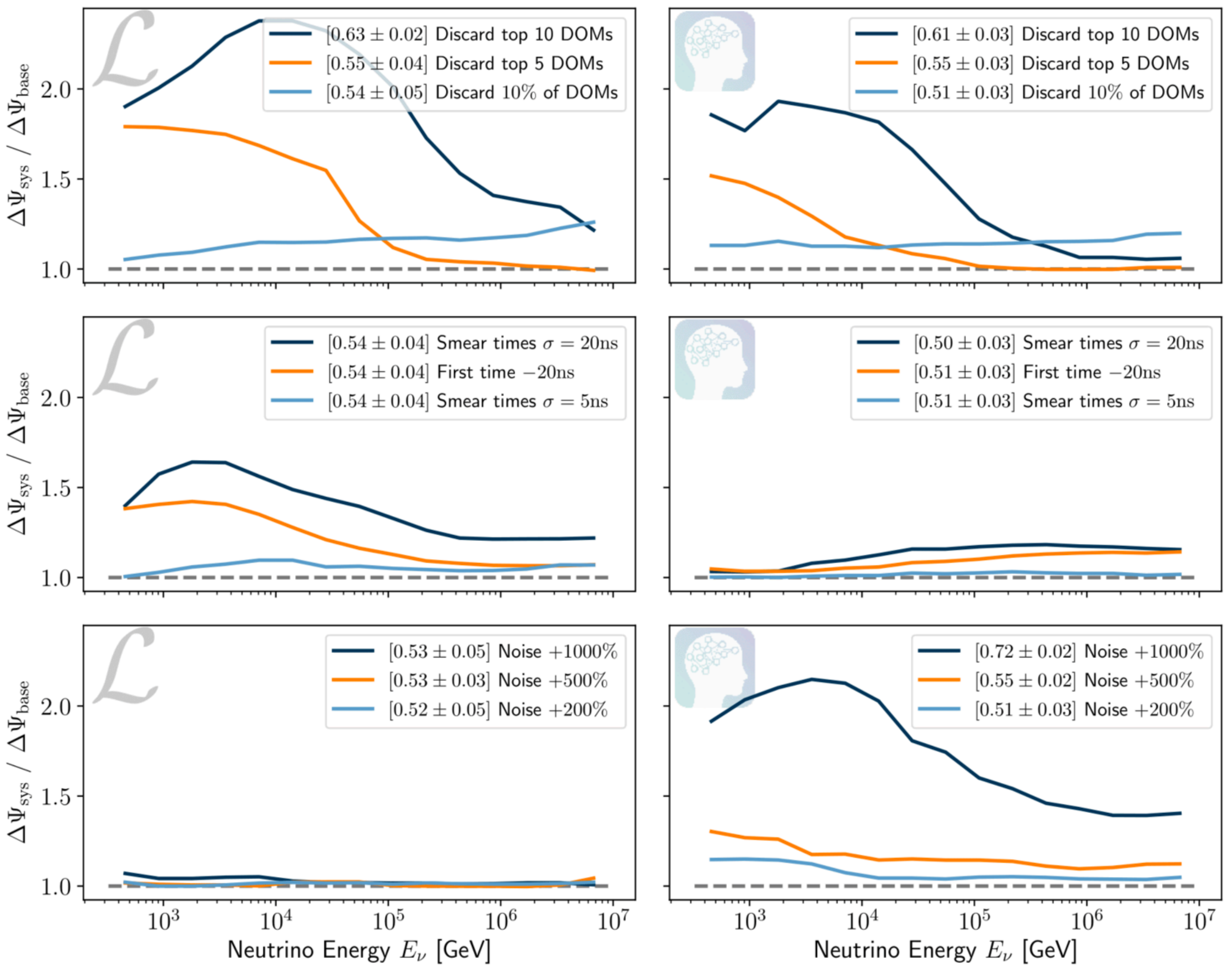
DNN Runtime


Fraction of Events
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
50%
1%
0.024s
0.116s
10.610s
164.692s
Standard
CNN
10^{-2}
10^{-1}
10^0
10^1
10^2
10^3
Runtime per Event / s

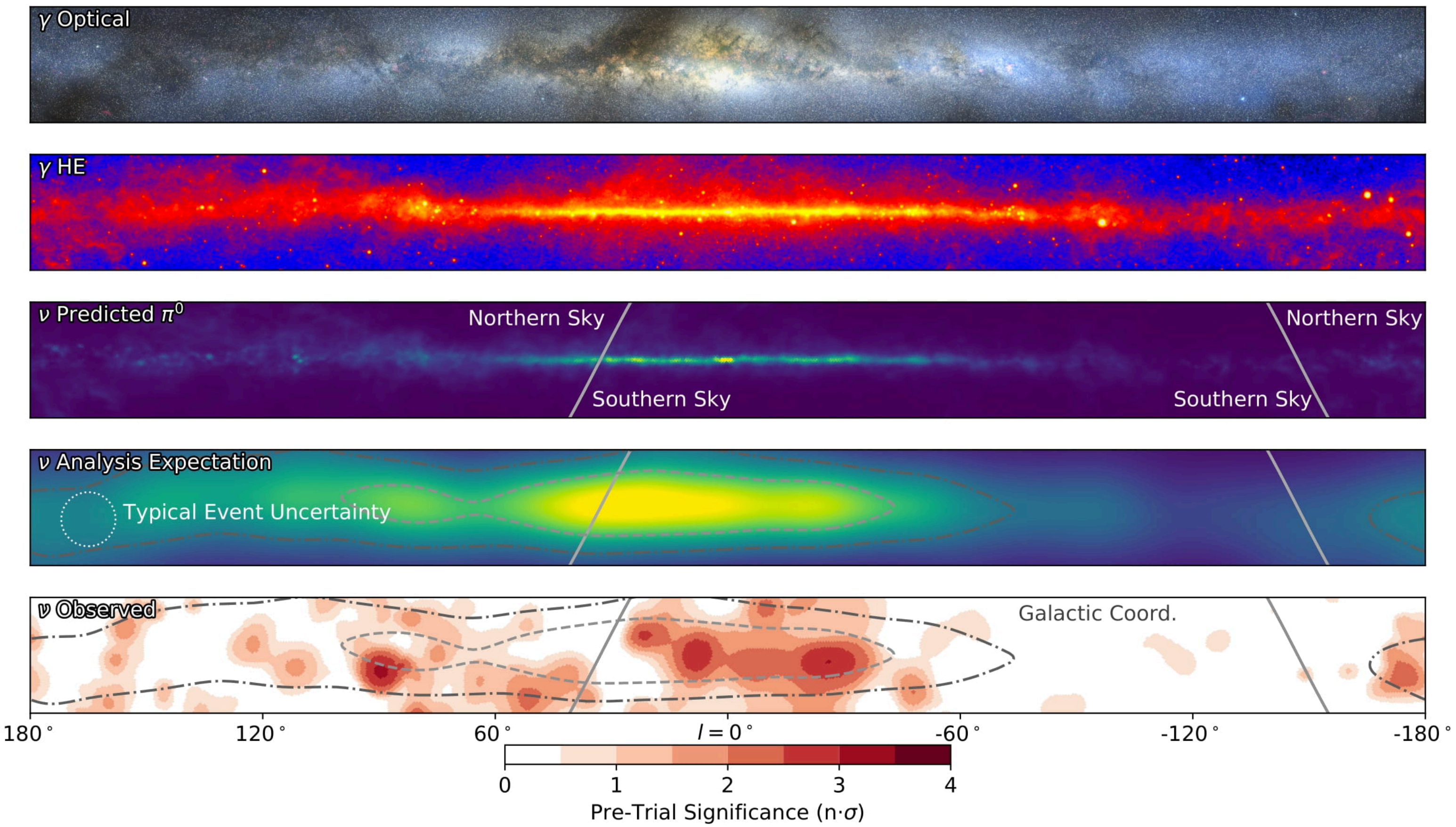
Robustness with Respect to Systematics



Additional Robustness Tests



A New Window to the Milky Way

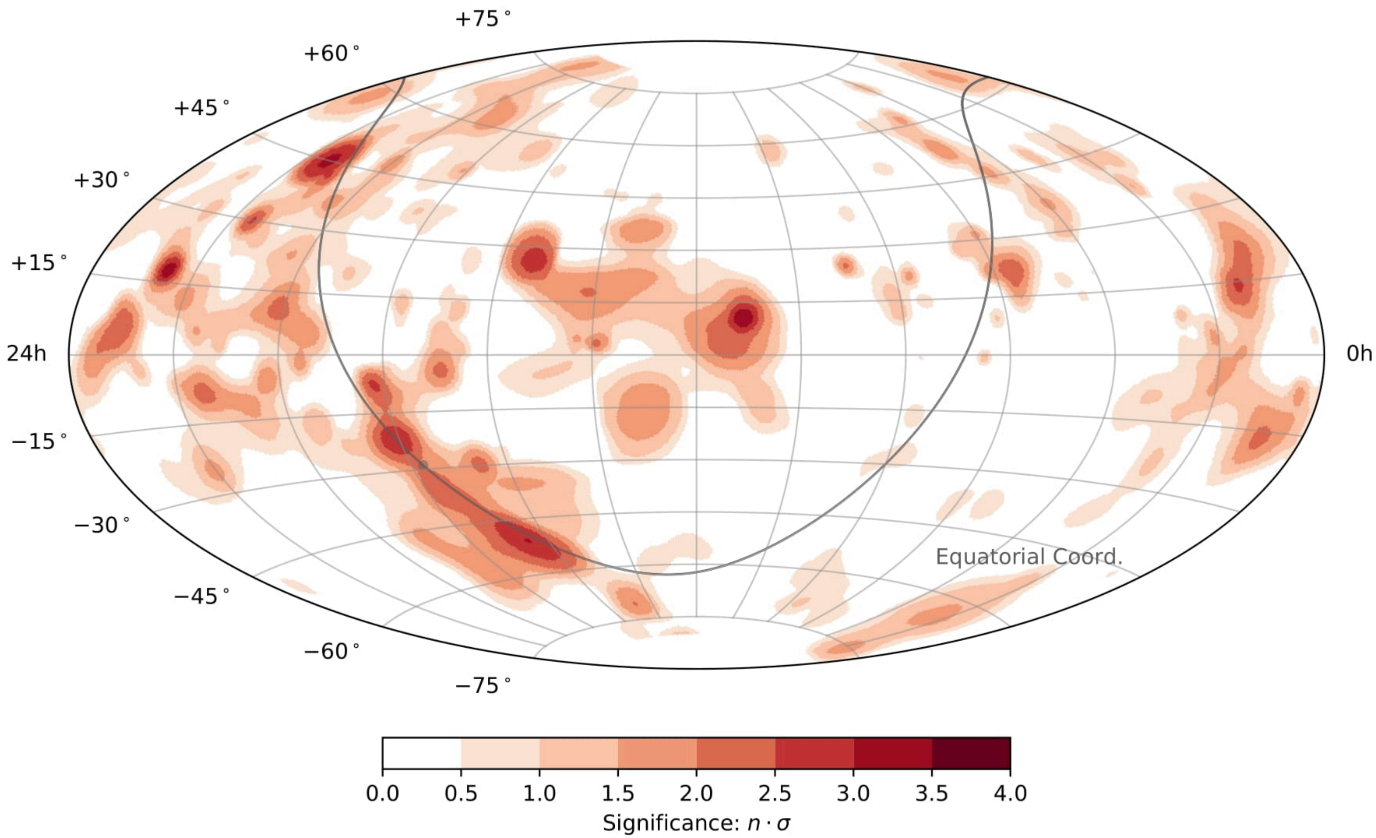


The Role of Deep Learning

- We deep learning-based tools to reject the overwhelming background of atmospheric muons and neutrinos
- High inference speed of NNs allows us to use more complex reconstructions at earlier stages of the event selection
- This allows us to retain more astrophysical cascade events in the sample, including events that are either difficult to reconstruct or hard to distinguish from background
- We retain a factor of 20 more events in the sample compared to previous analyses**
- We also utilized a GAN to parameterize the relationship between event hypothesis an expected light yield.

** not all of them are due to Deep Learning.

Discovering Neutrinos From the Galactic Plane



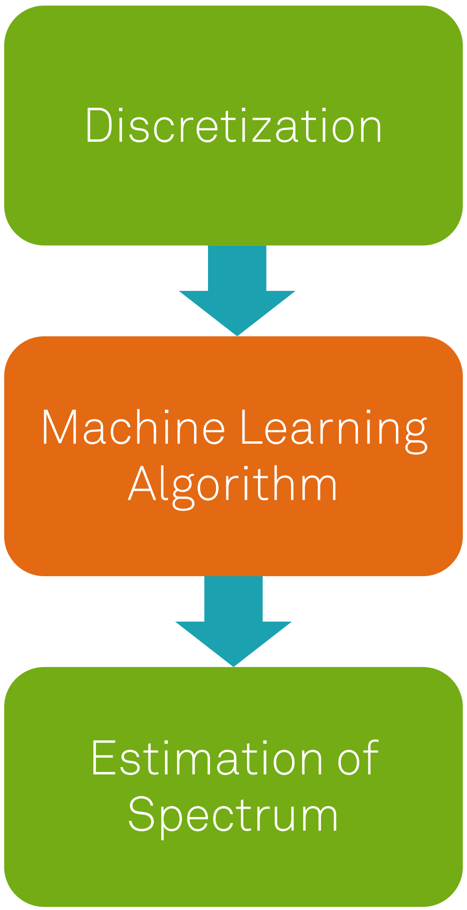
Summary

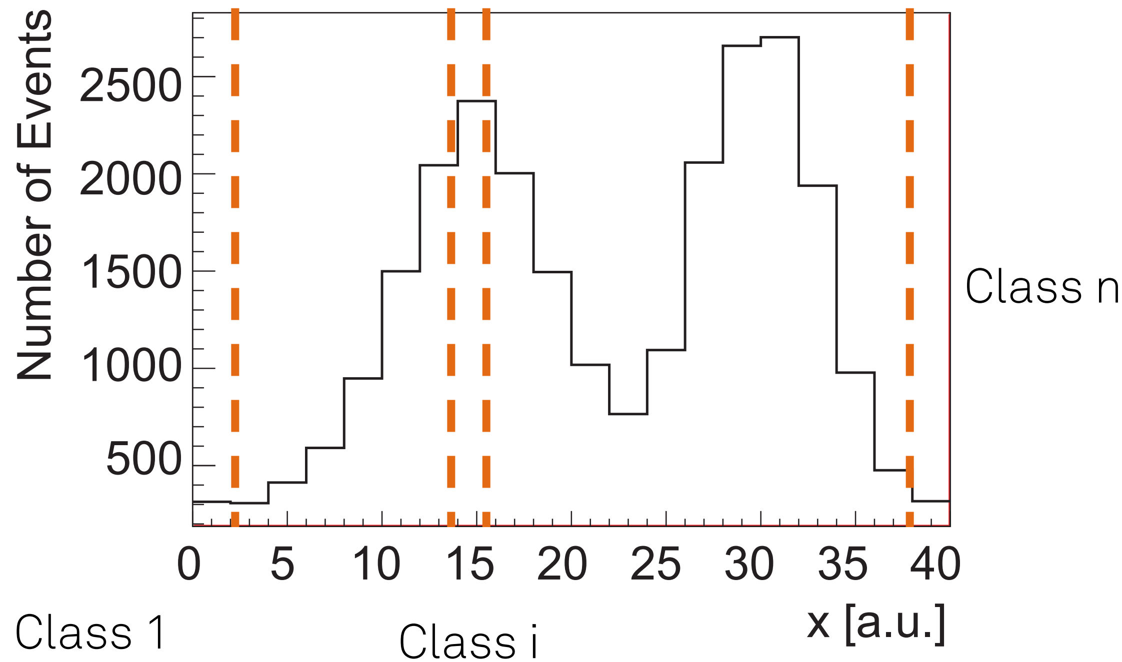
- Understand your input!
- Use the right model for the task!
- Understand your ouput!
- Cross validate!
- Machine learning is a tool, use it wisely!

Backup Slides

Feature Importance

Unfolding via Machine Learning

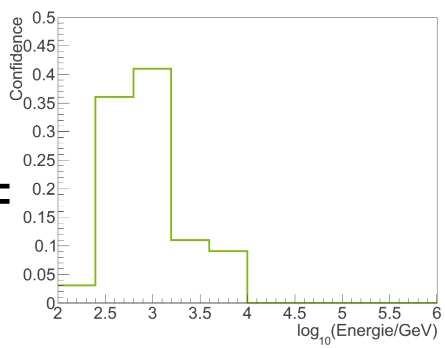




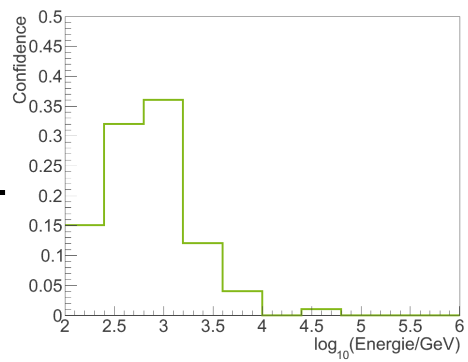
Inverse problem is transferred into multinominal classification problem.

DSEA in greater detail

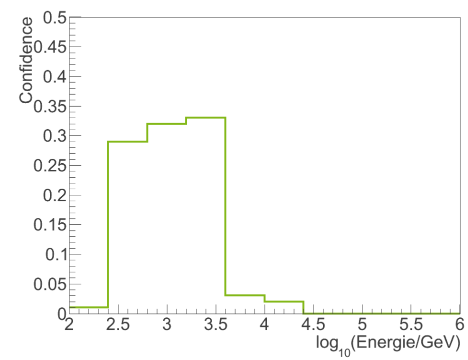
$\hat{f} =$



$+$



$+ \ldots +$



Iterate:
1. Discretize
2. Train Model
3. Apply Model
4. Reconstruct spectrum
5. Update weights according to unfolding result

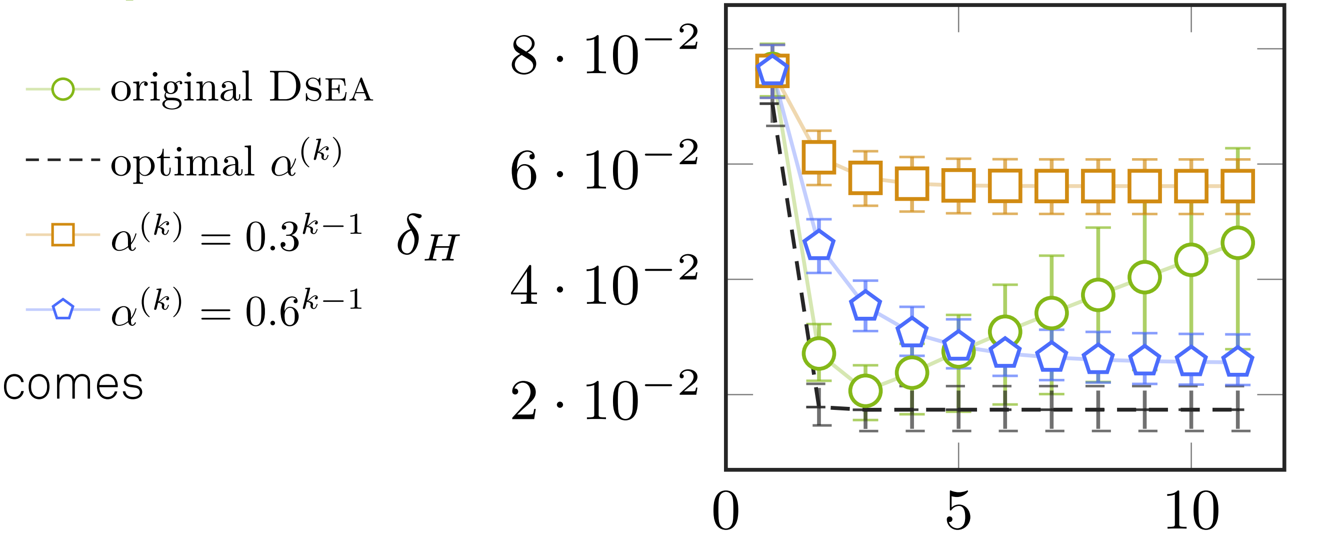
Choice of learning algorithm largely arbitrary (and probably somewhat problem dependent).

Some overlap with IBU in case Naive Bayes is used as a learner.

DSEA+: Variable step width

Step Width

$$p_k = f_k - f_{k-1}$$



Next estimate then becomes

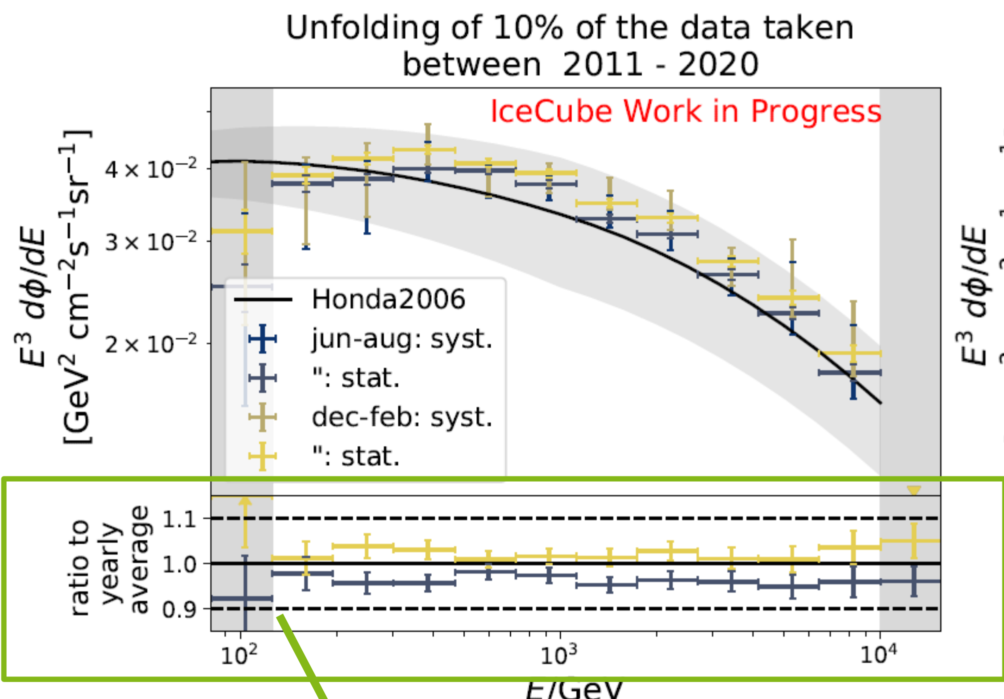
$$f_k = f_{k-1} + \alpha_k p_k$$

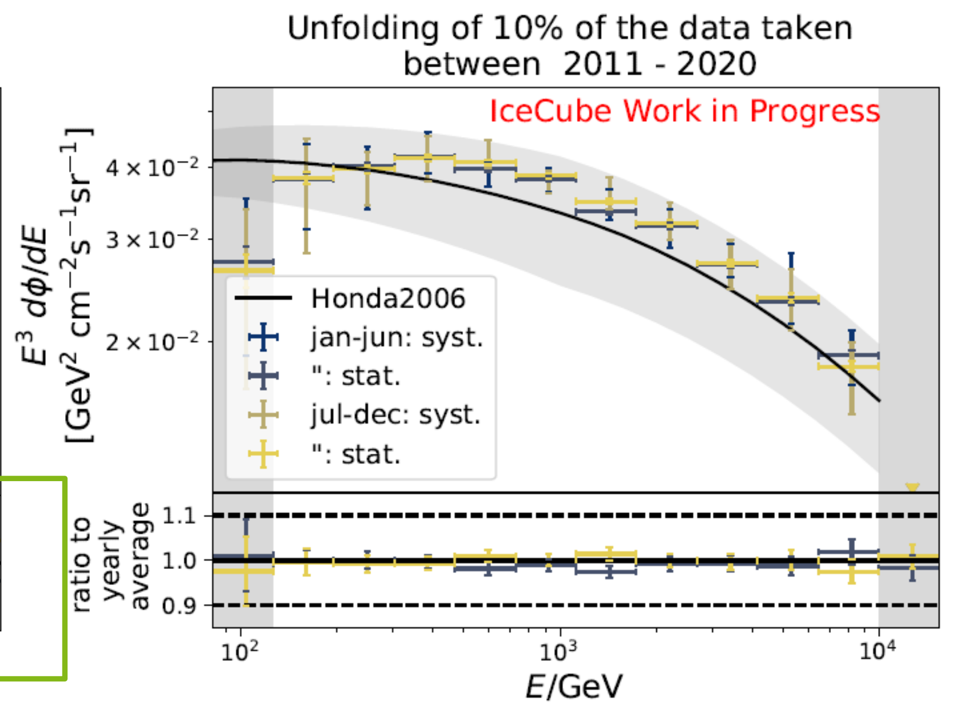
Find optimal α via:

$$\alpha = \arg\min_{\alpha \geq 0} l(f_{k-1} + \alpha_k p_k)$$

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

Very Preliminary Results





only dependent on statistics – systematics are independent on season!

Loss Functions

MSE (first training steps, robustness)

$$(\mathrm{L})_{b,k} = \underbrace{\left(\left(Y''_{\mathrm{true}} - Y''_{\mathrm{pred}}\right)_{b,k}\right)^2}_{(\mathrm{L}_{\mathrm{pred}})_{b,k}} + \underbrace{\left[(Y''_{\mathrm{unc}})_{b,k} - \texttt{gradient\_stop}\left(\left|\left(Y''_{\mathrm{true}} - Y''_{\mathrm{pred}}\right)_{b,k}\right|\right)\right]^2}_{(\mathrm{L}_{\mathrm{unc}})_{b,k}}$$

Gaussian Likelihood (later stages)

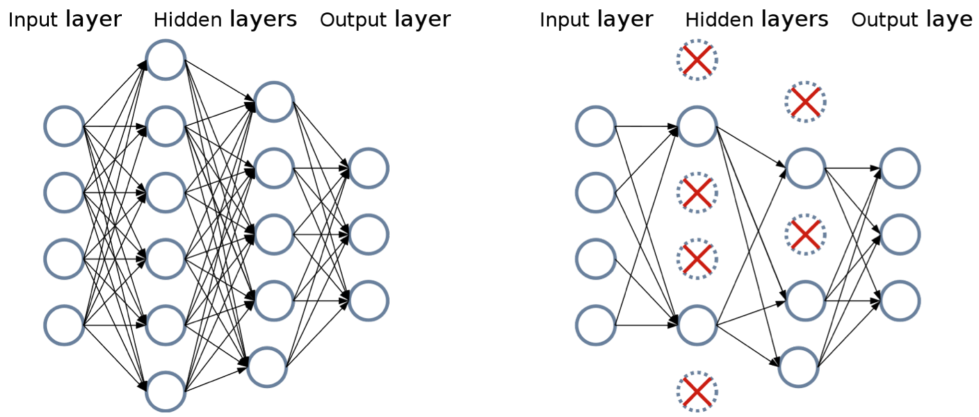
$$(\mathrm{L})_{b,k} = 2 \cdot \ln\left((Y''_{\mathrm{unc}})_{b,k}\right) + \left(\frac{\left(Y''_{\mathrm{true}} - Y''_{\mathrm{pred}}\right)_{b,k}}{(Y''_{\mathrm{unc}})_{b,k}}\right)^2$$

Epoch and (mini)batch

- Minibatch, Batch: Using all examples can be infeasible in case many parameters need to be optimized, instead random subsets (batches) of examples are used. The optimal size of the batch depends on the problem to be solved. Popular choices are 2^k

- Epoch: Complete use of all examples.

Regularization



Source: By Mads Dyrmann - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=110373041

Regularization via Dropout, but gradually decreased at later layers.

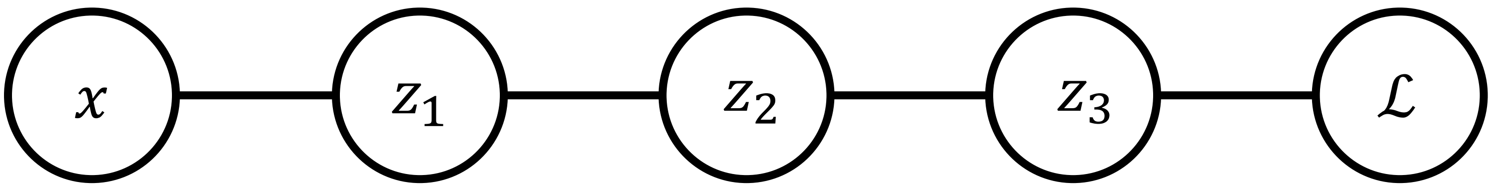
In addition, individual DOMs are randomly dropped for increased stability.

Input Preparation

- **Zero-centered:** ReLU changes drastically around 0, $x_i - \langle x_i \rangle$ to include positive and negative values
- **Order of magnitude:** Large variables could be preferred in the network training $x'_i = \frac{x_i - \langle x_i \rangle}{\sigma_i}$
- **Logarithm** to achieve more evenly distributed data
- **Decorrelation:** highly correlated variables should be decorrelated

Weight Updates

$$z_1 = W \cdot x \qquad z_3 = \sigma(z_2)$$



$$z_2 = z_1 + b$$

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial z_2} \cdot \frac{\partial z_2}{\partial z_1} \cdot \frac{\partial z_1}{\partial W}$$

$$W_{t+1} = W_t - \alpha \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial W}\right]_t$$

$$\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial W}\right] = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\partial \mathcal{L}}{\partial W}\right)_i$$

This is the basic idea, this will most likely be handled by an optimizer.

Take-Away Messages

- Machine Learning and esp. Deep Learning is not magic
- Machine Learning and Deep Learning are tools that will help you to accomplish an analysis task faster and more accurately (when used correctly)
- The preprocessing of data is part of machine learning (and very important)
- Not every classifier is suited for every problem (consider runtime)
- If something fast and simple does the job: use it
- Make sure simulated and experimental data agree
- ...