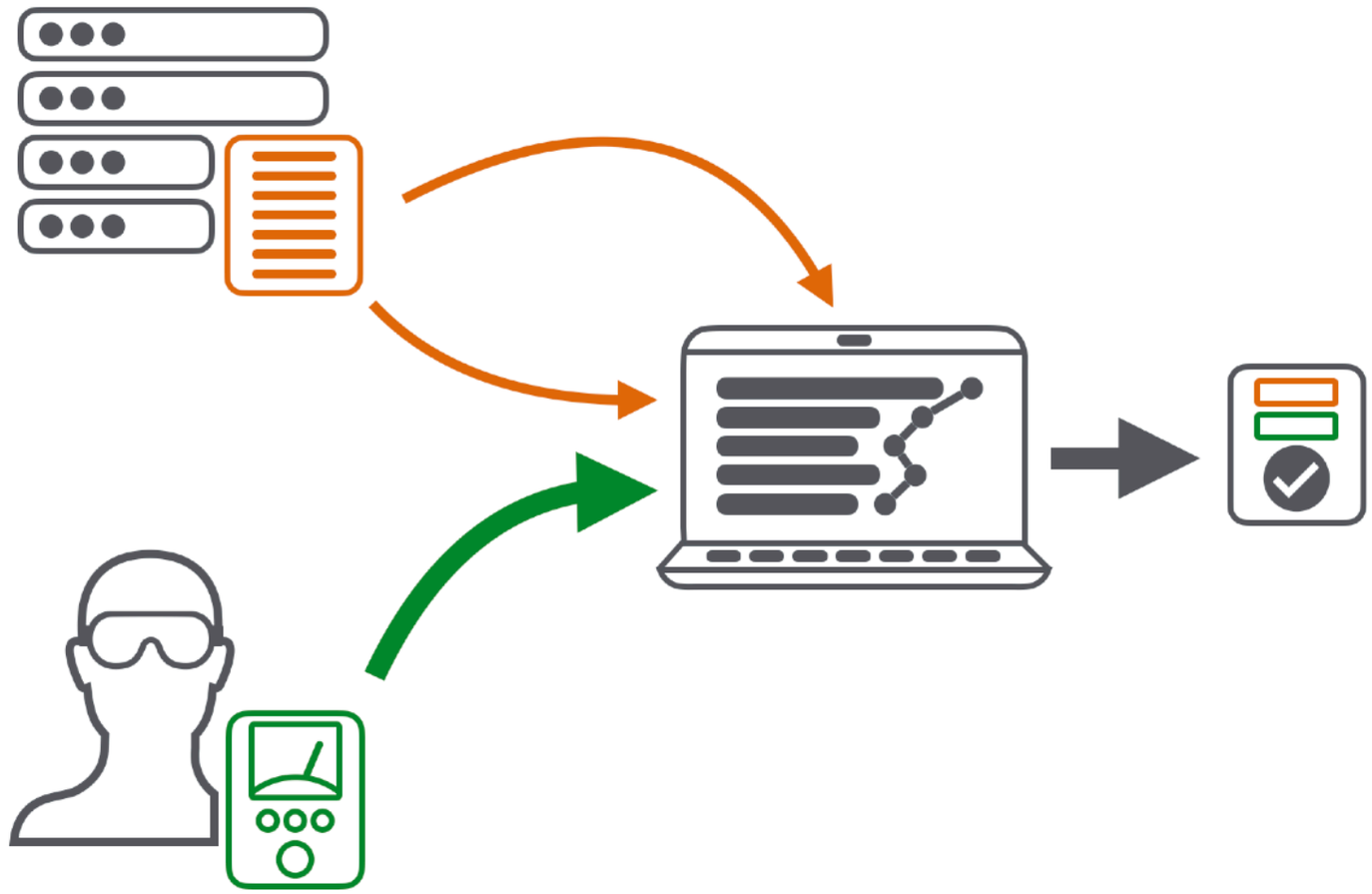


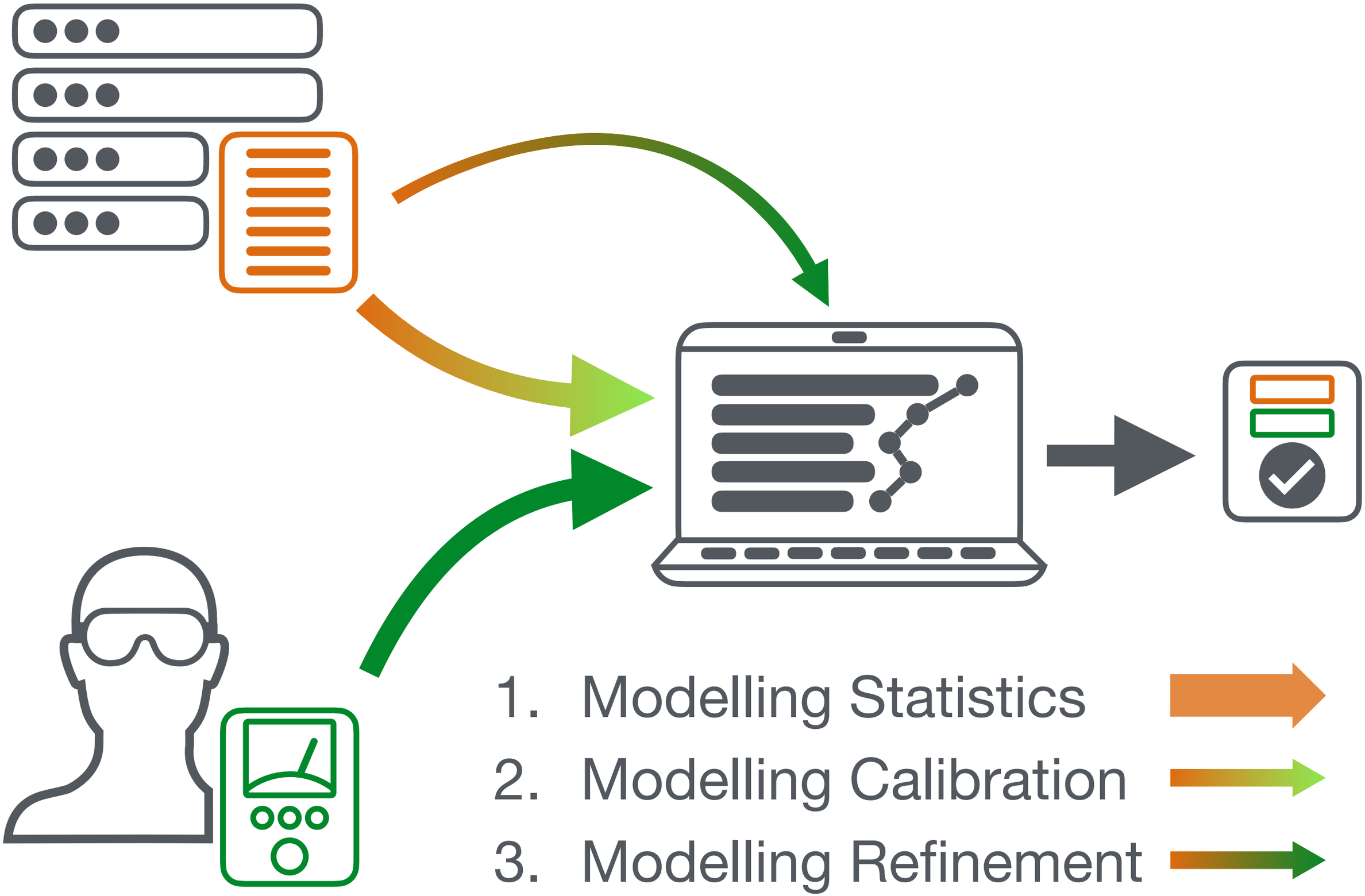
Improved Data Modelling through Machine Learning

Dennis Noll

PCD Lecture

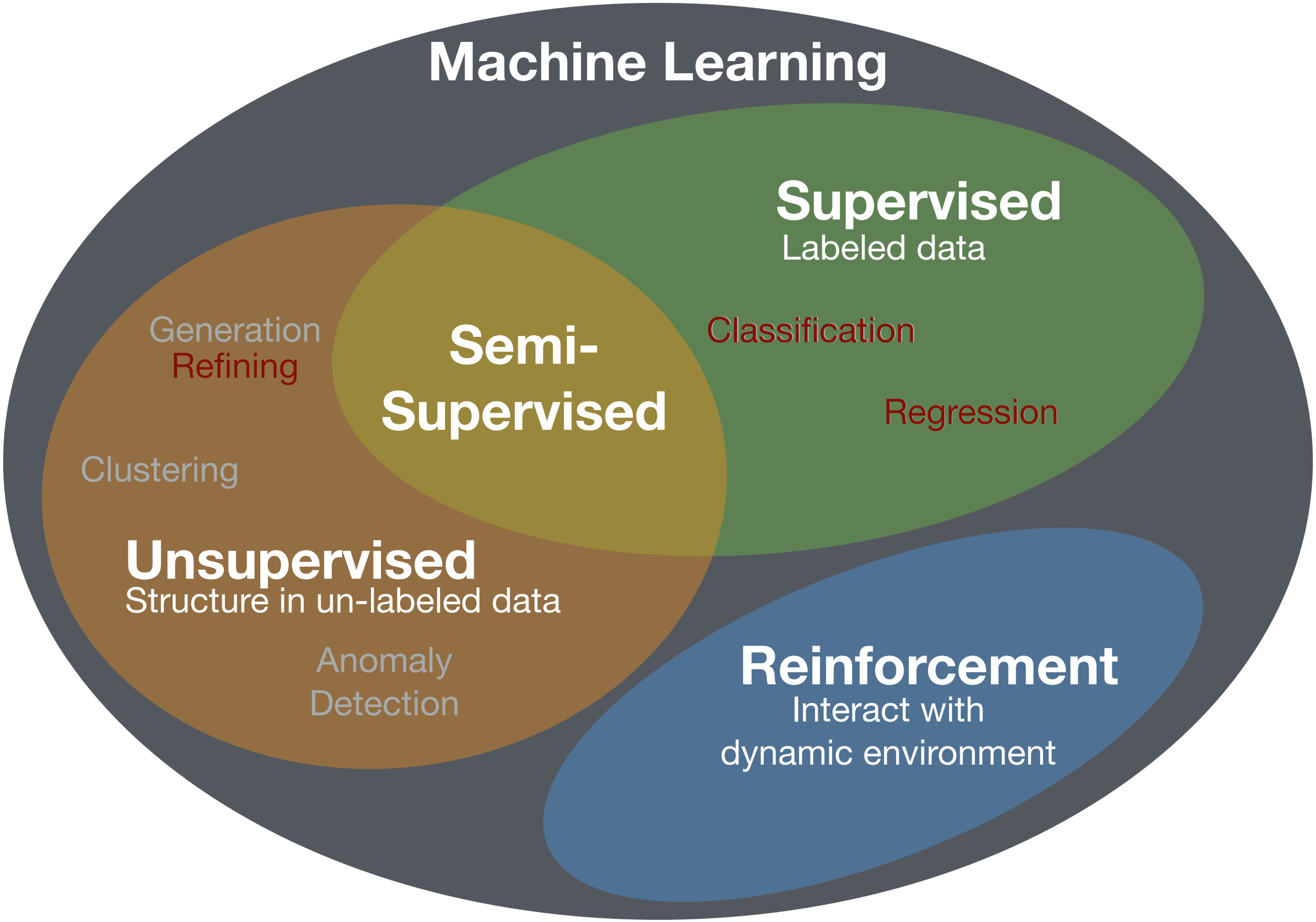
05.07.22

RWTH AACHEN UNIVERSITY


1. Modelling Statistics
2. Modelling Calibration
3. Modelling Refinement

Definition of Machine Learning

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



Primer: Linear Regression

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

Define terminology based on linear regression (supervised learning):

Dataset

$$\{x_i, y_i\} \;\; i = 1 \ldots N$$

Model

$$y(x, \theta) = Wx + b$$

Parameters $\theta = (W, b)$

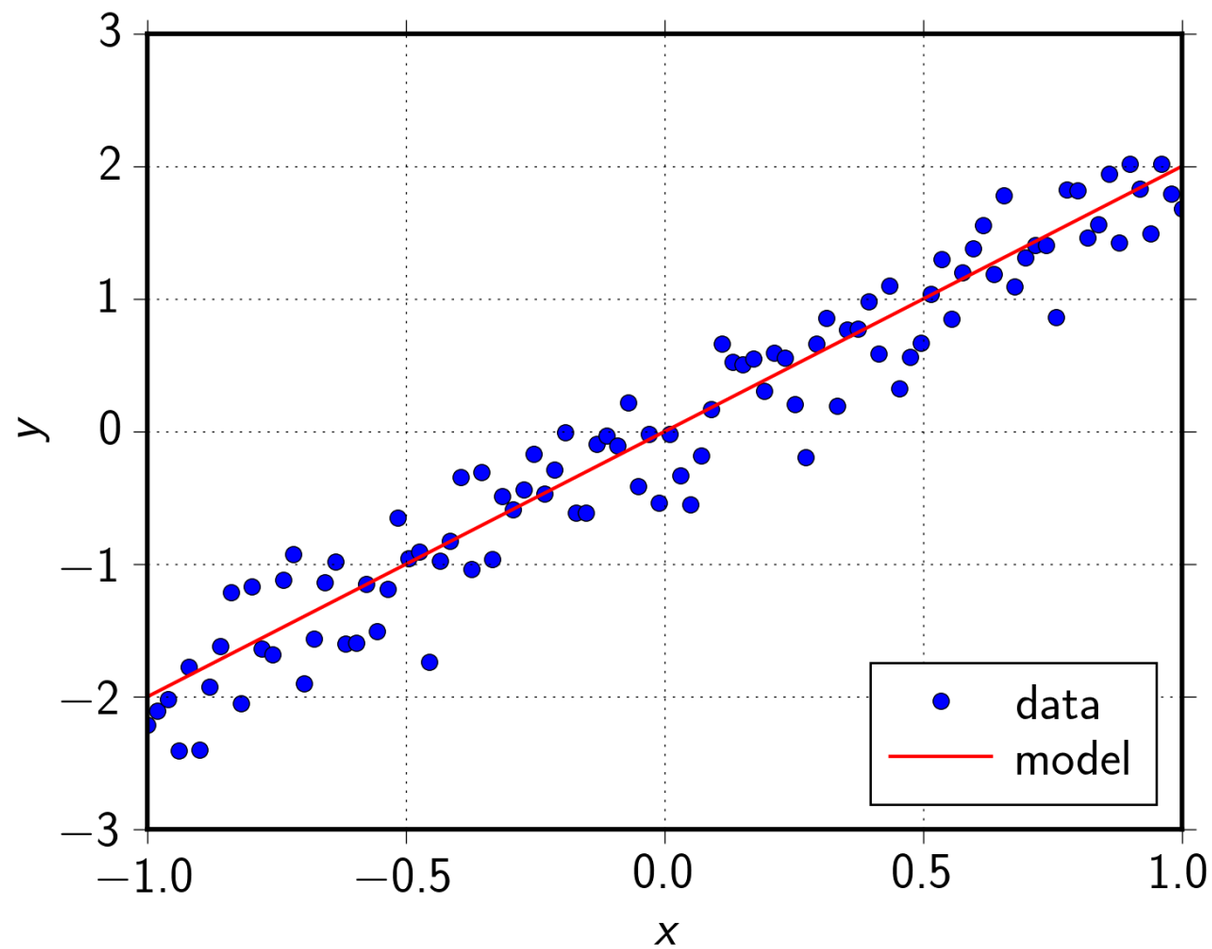
Objective Function (Loss)

$$\mathscr{L}(\theta) = \sum_{i=1}^{N} (y(x_i, \theta) - y_i)^2$$

Training

Find parameters:

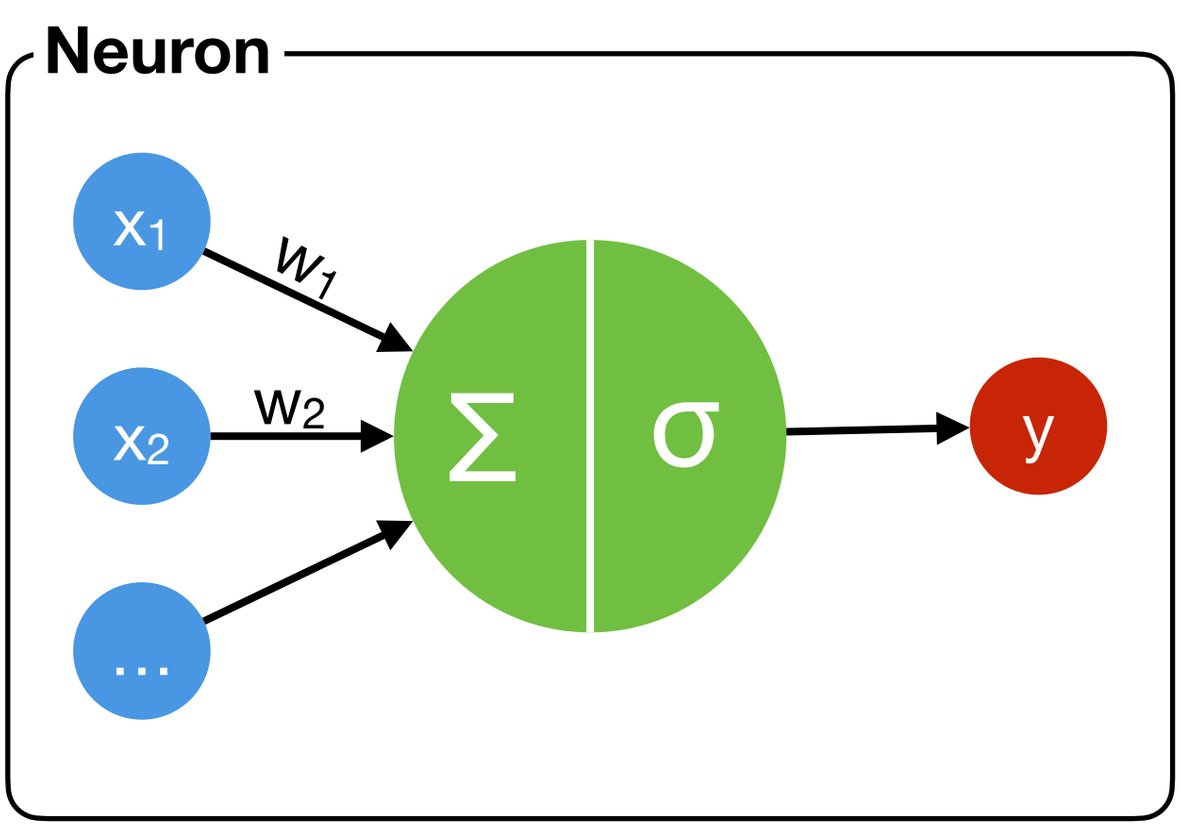
$$\hat{\theta} = \underset{\theta}{argmin} \; \mathscr{L}(\theta)$$

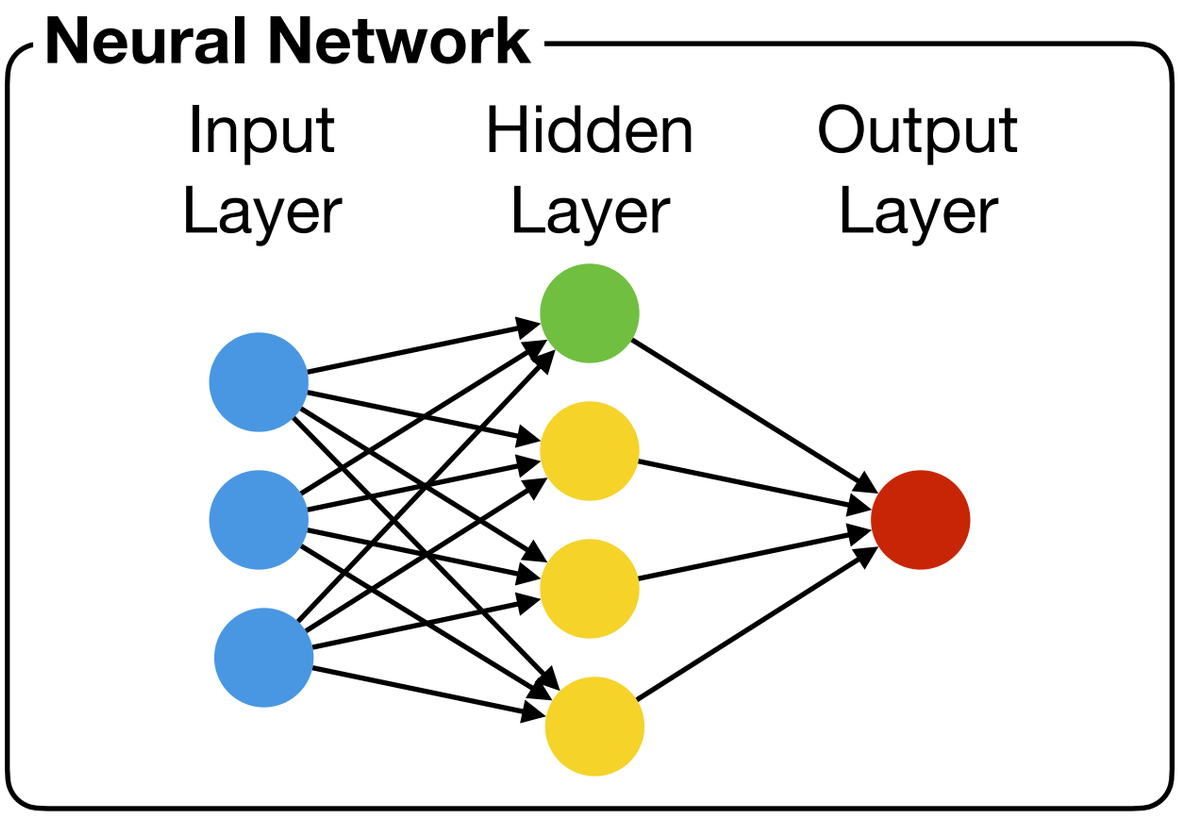


Neural Networks: Basics

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Want: Approximate arbitrary function
- Need: General parametrisable model
- Solution: Network out of many neurons



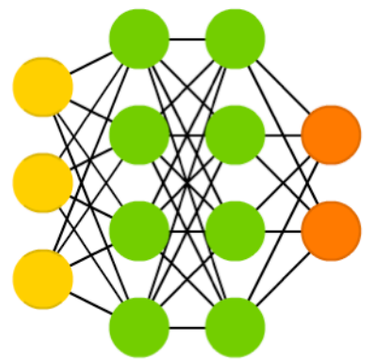


Neural Networks: Architectures

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

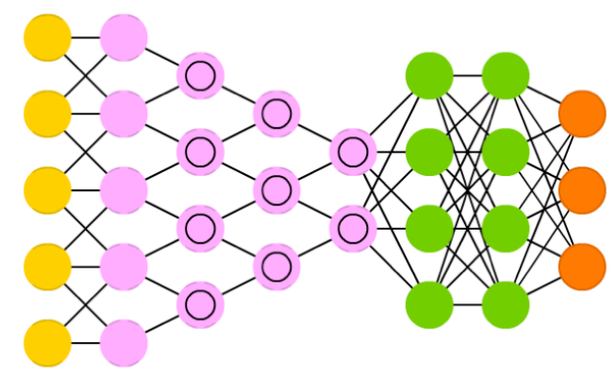
- Many different ways to arrange nodes and layers

Fully Connected



No particular symmetries
Many parameters (e.g. DenseNet)

Convolutional

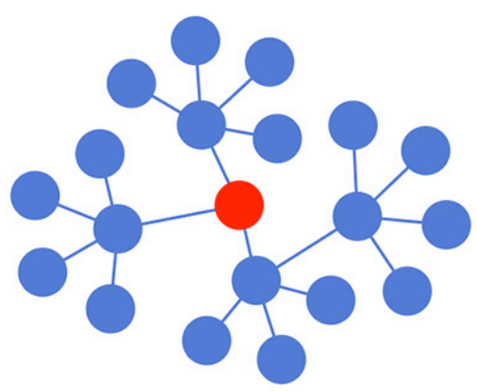


Spatial Symmetries

Legend:

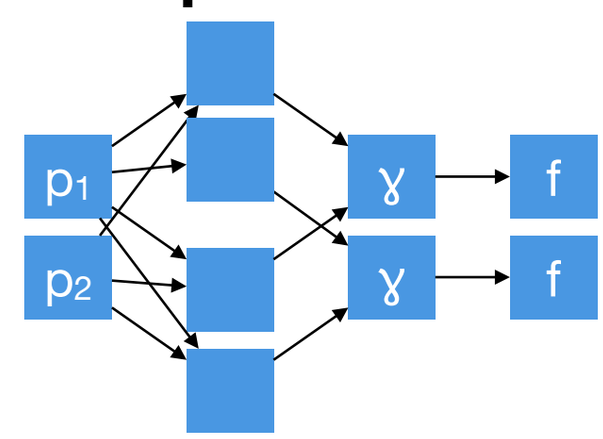
- Input Cell
- Output Cell
- Hidden Cell
- Kernel
- Convolution
- Graph Node

Graph Networks



Graph Symmetries

Physics Inspired



Physics Symmetries (e.g. Lorentz)

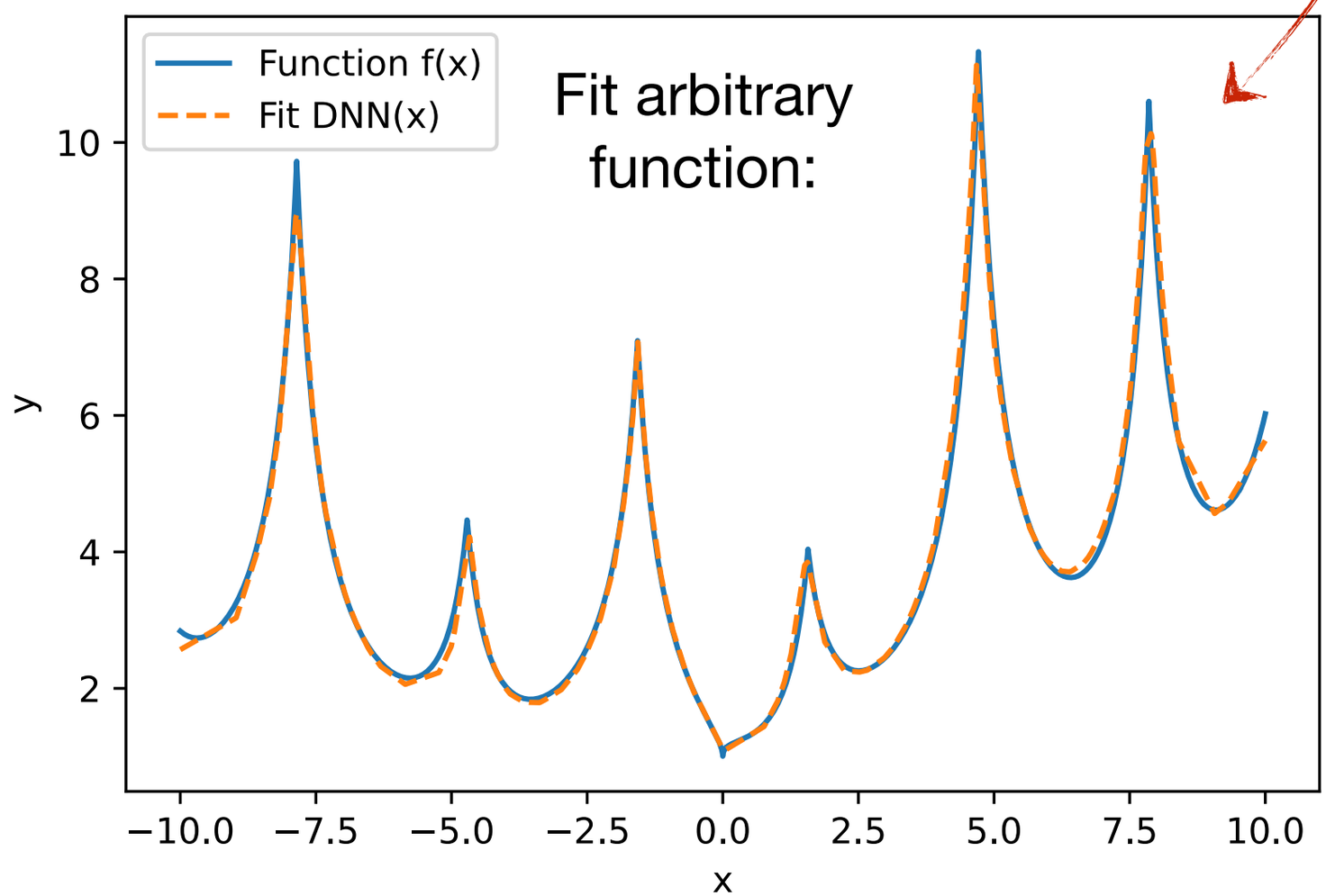
…many more

Graphics taken from [here] and [here]

Neural Networks: Objective and Training

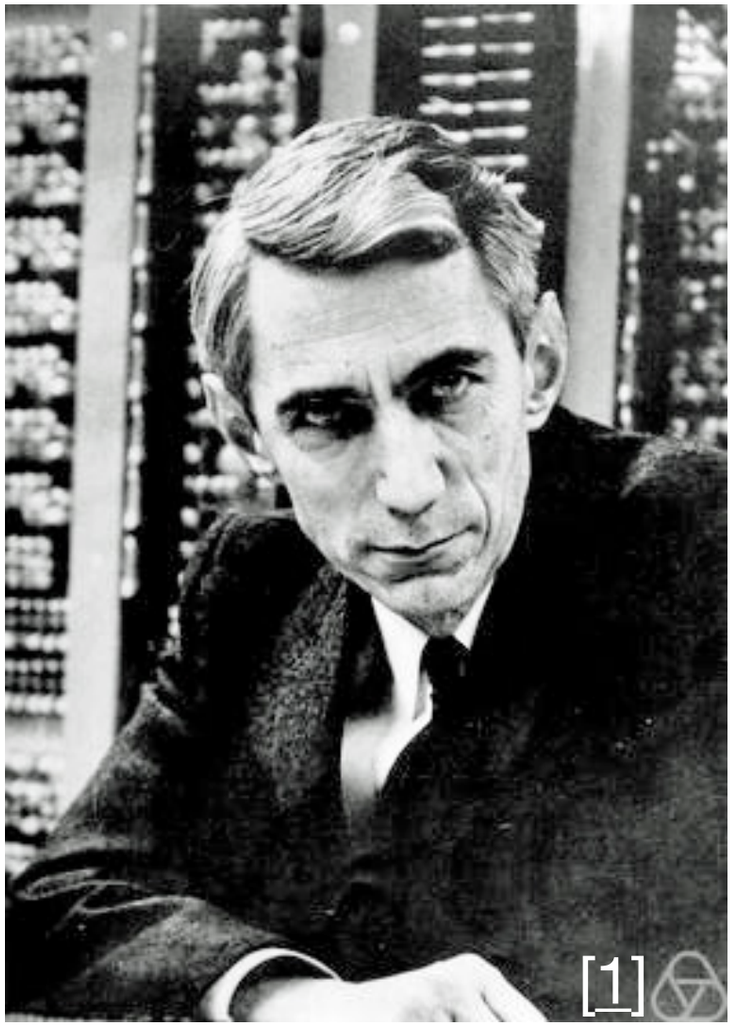
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Loss Functions (What do I want my model to do?):
  - Two standard loss functions:
    - Mean squared error (Regression)
    - Cross-Entropy (Classification)
- Training neural network to find best model parameters:
  - Many Optimizers available: SGD, Adam, …



Information Theory

- Founded in 1948 by Claude Shannon
- What is a bit:
  - Bit = 0 or 1
  - Bit = Uncertainty reduced 50%
- Deep connection between:
  - Information theory
  - (Cross-)Entropy
  - Kullback-Leibler Divergence



The Bell System Technical Journal,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

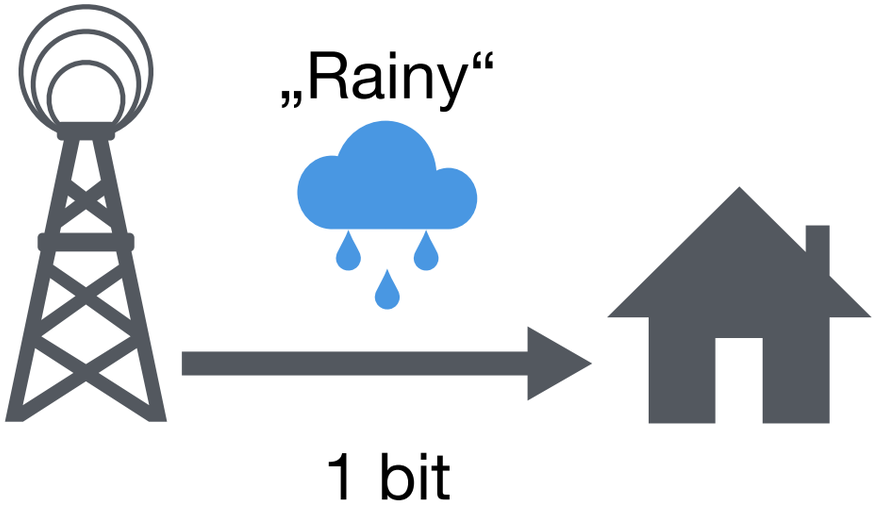
A Mathematical Theory of Communication

By C. E. SHANNON

Following slides adapted from Aurélien Géron [2]

Measuring Information

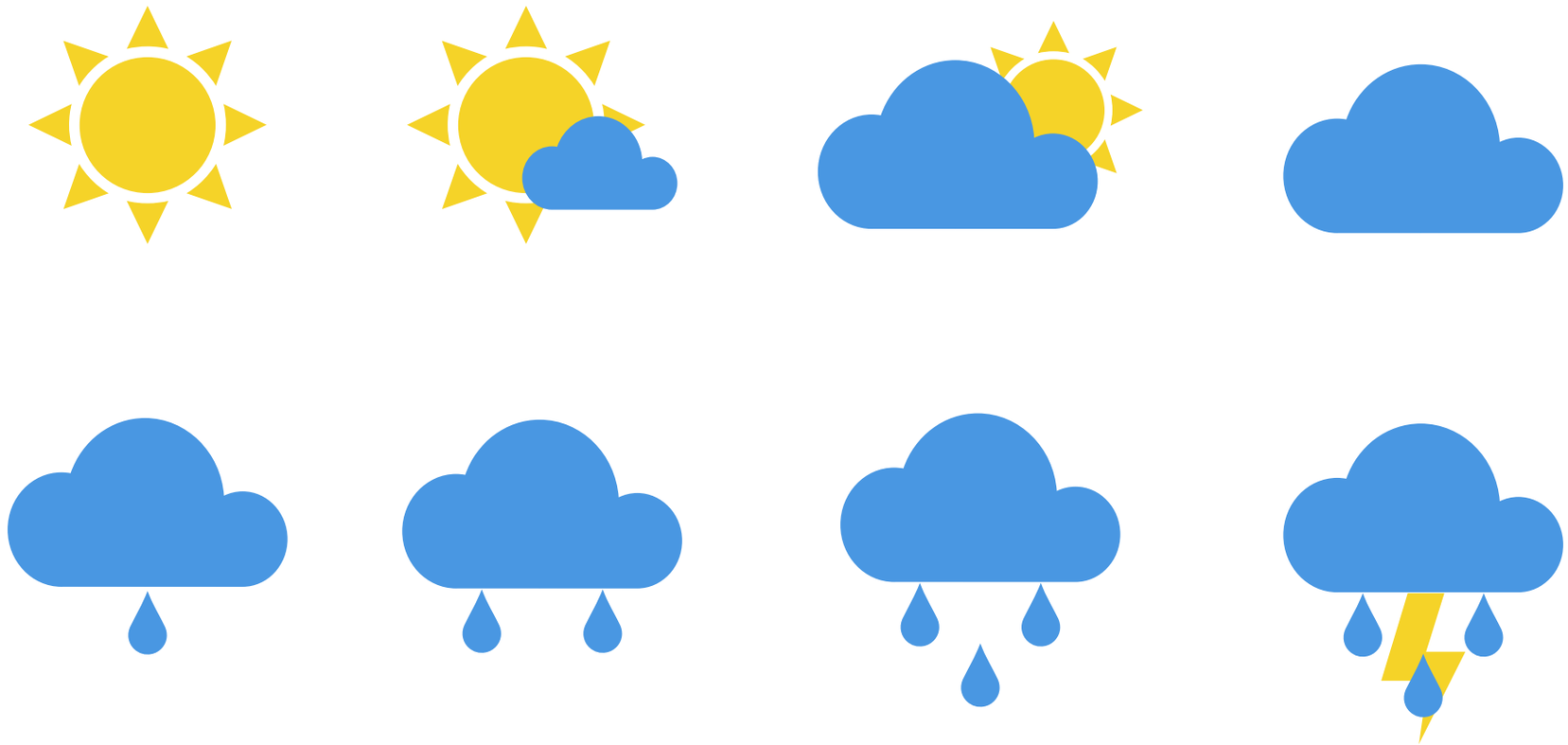
Dennis Noll - 05.07.22



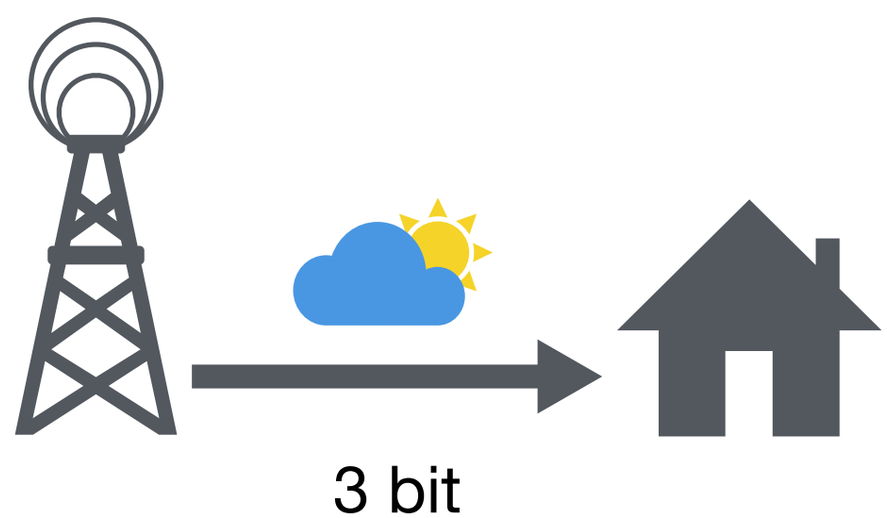
[2]

Measuring Information

Dennis Noll - 05.07.22

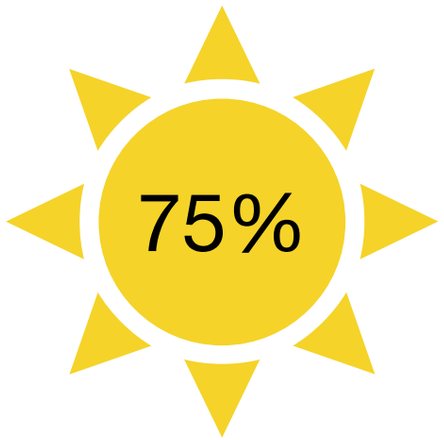


Uncertainty reduced: $8 = 2^3$
Number of bits: $\log_2(8) = 3$



[2]

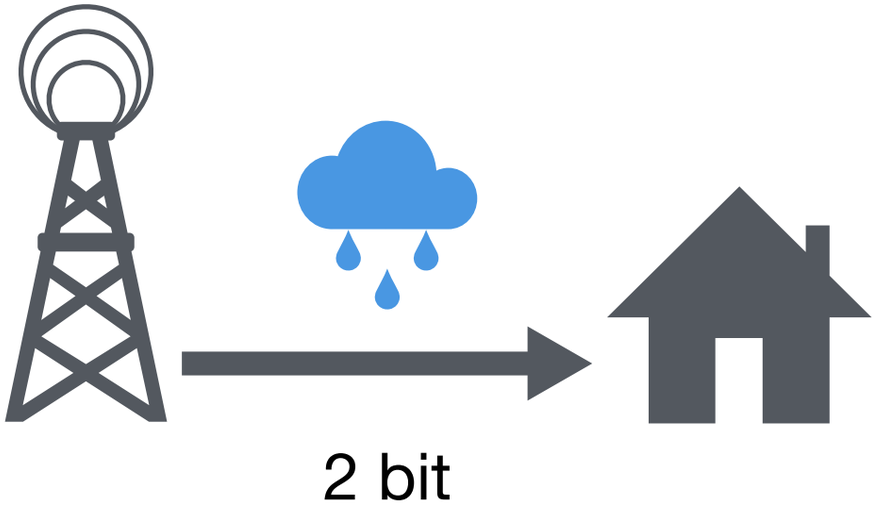
Dennis Noll - 05.07.22



Uncertainty reduced: 4

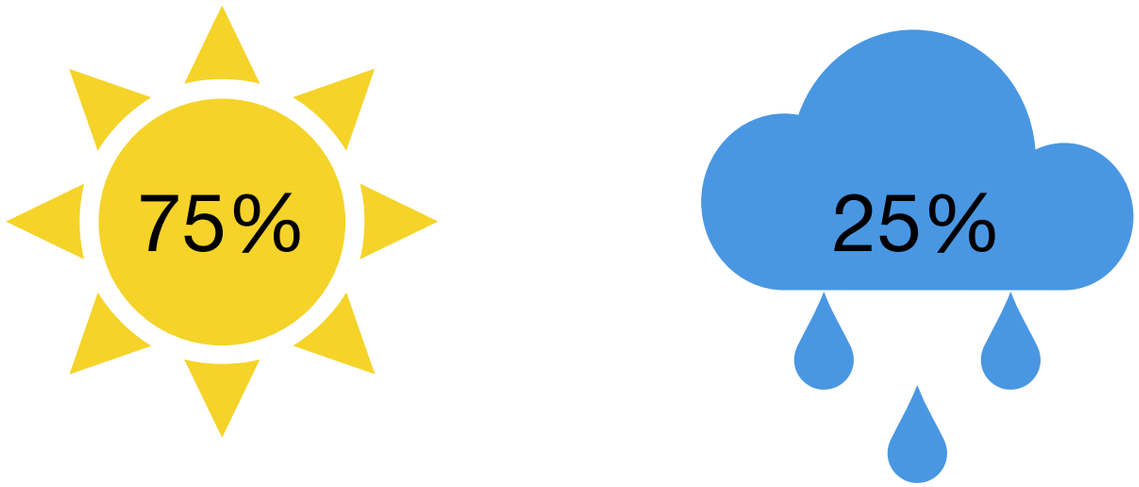
Number of bits: $-\log_2(25\%) = 2$

Probability

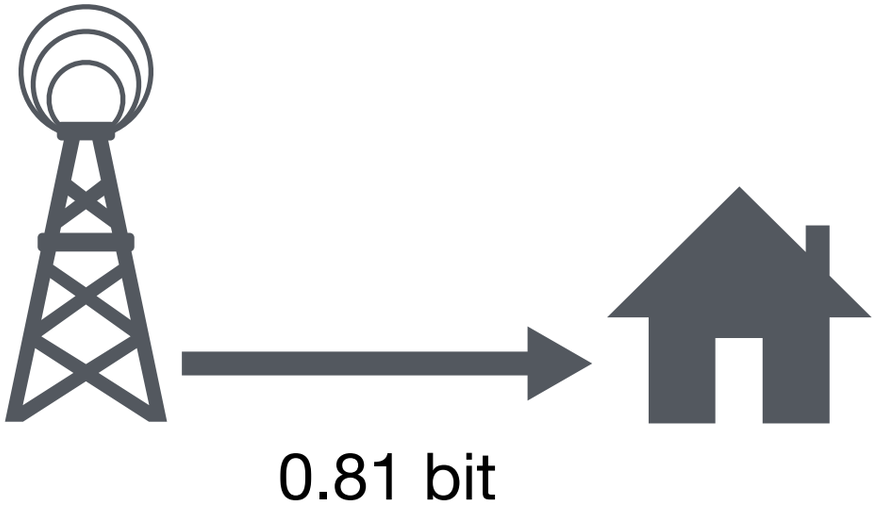


[2]

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



Desert = low
Alps = high



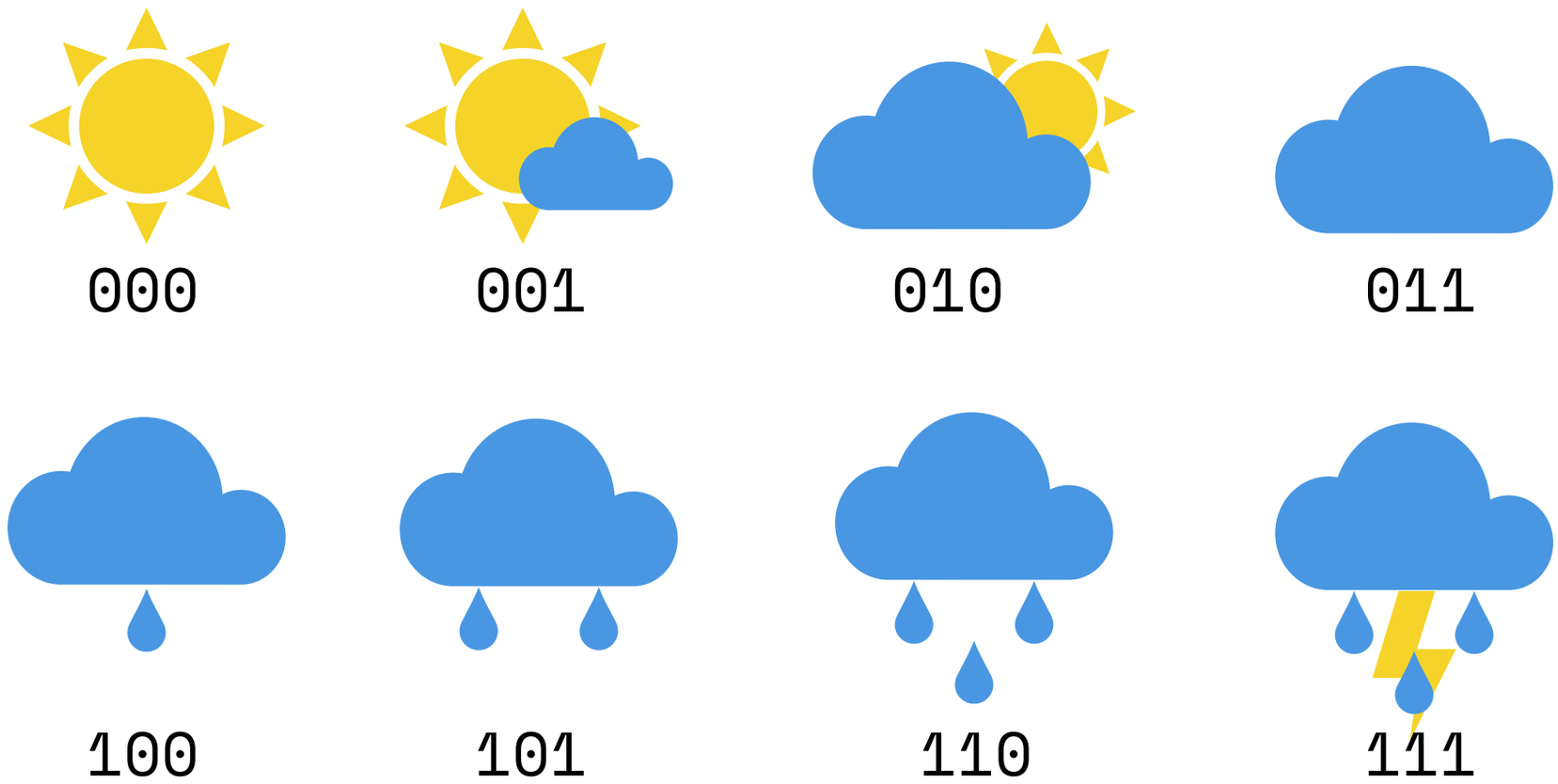
Entropy (Average Information):

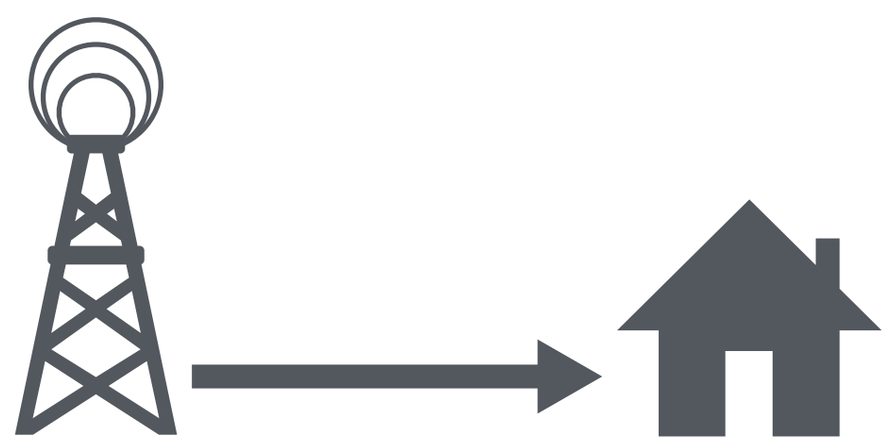
$H(p) = -\Sigma_i\, p_i \times \log_2(p_i) = 75\% \times 0.41 + 25\% \times 2 = 0.81$

[2]

Cross Entropy

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



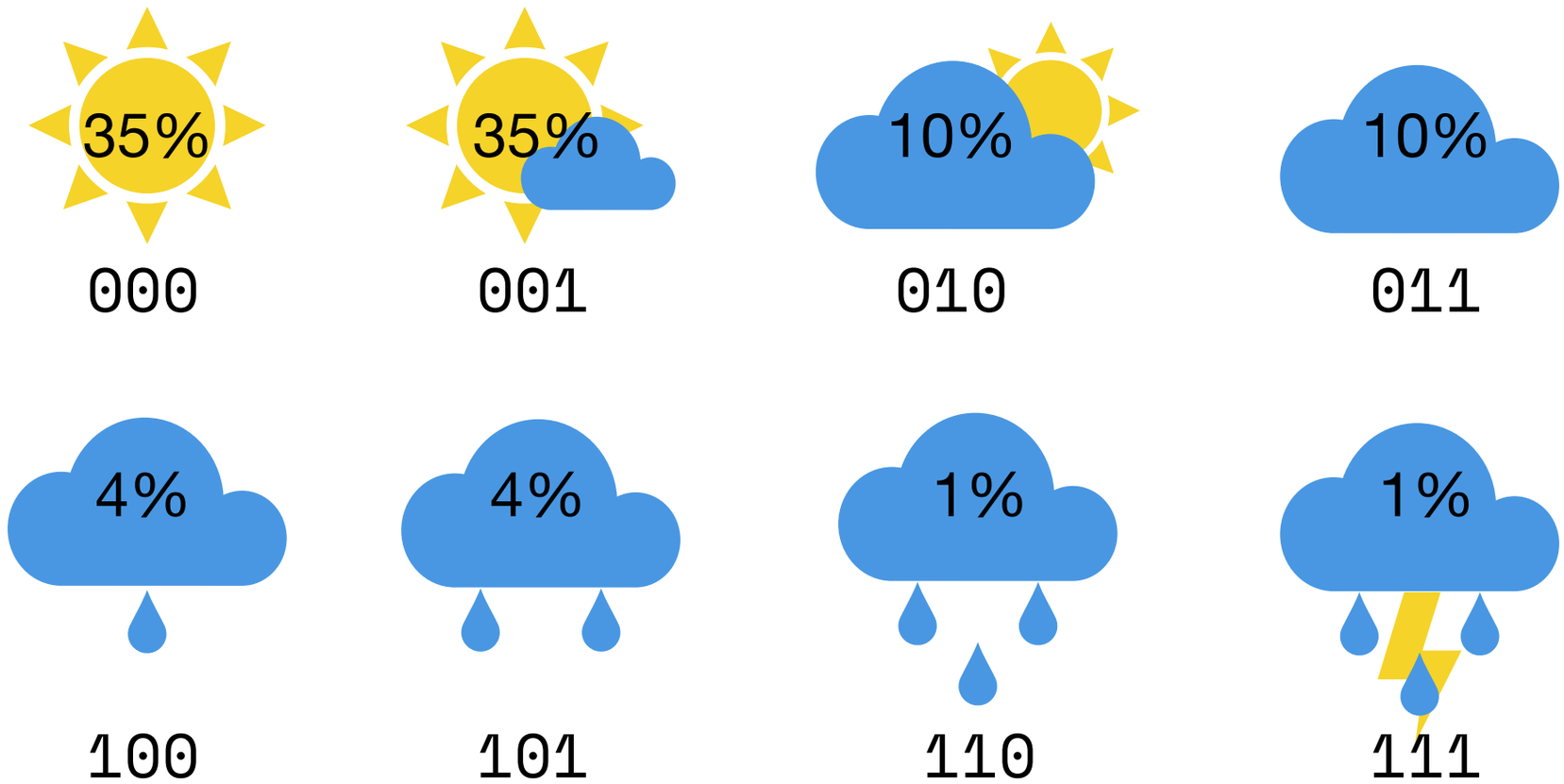


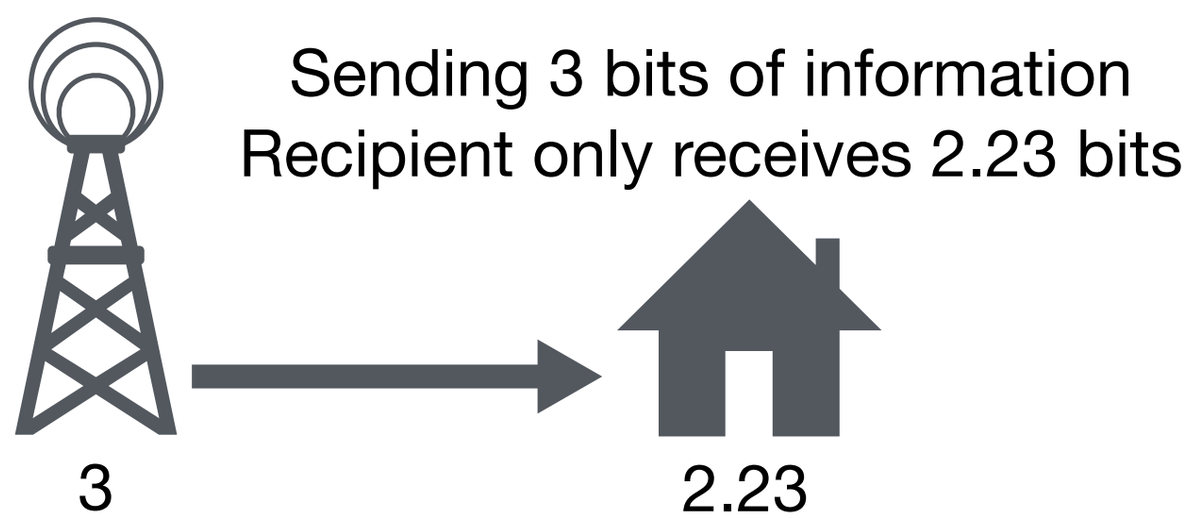
Cross Entropy (Average Message Length) = 3

[2]

Cross Entropy

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



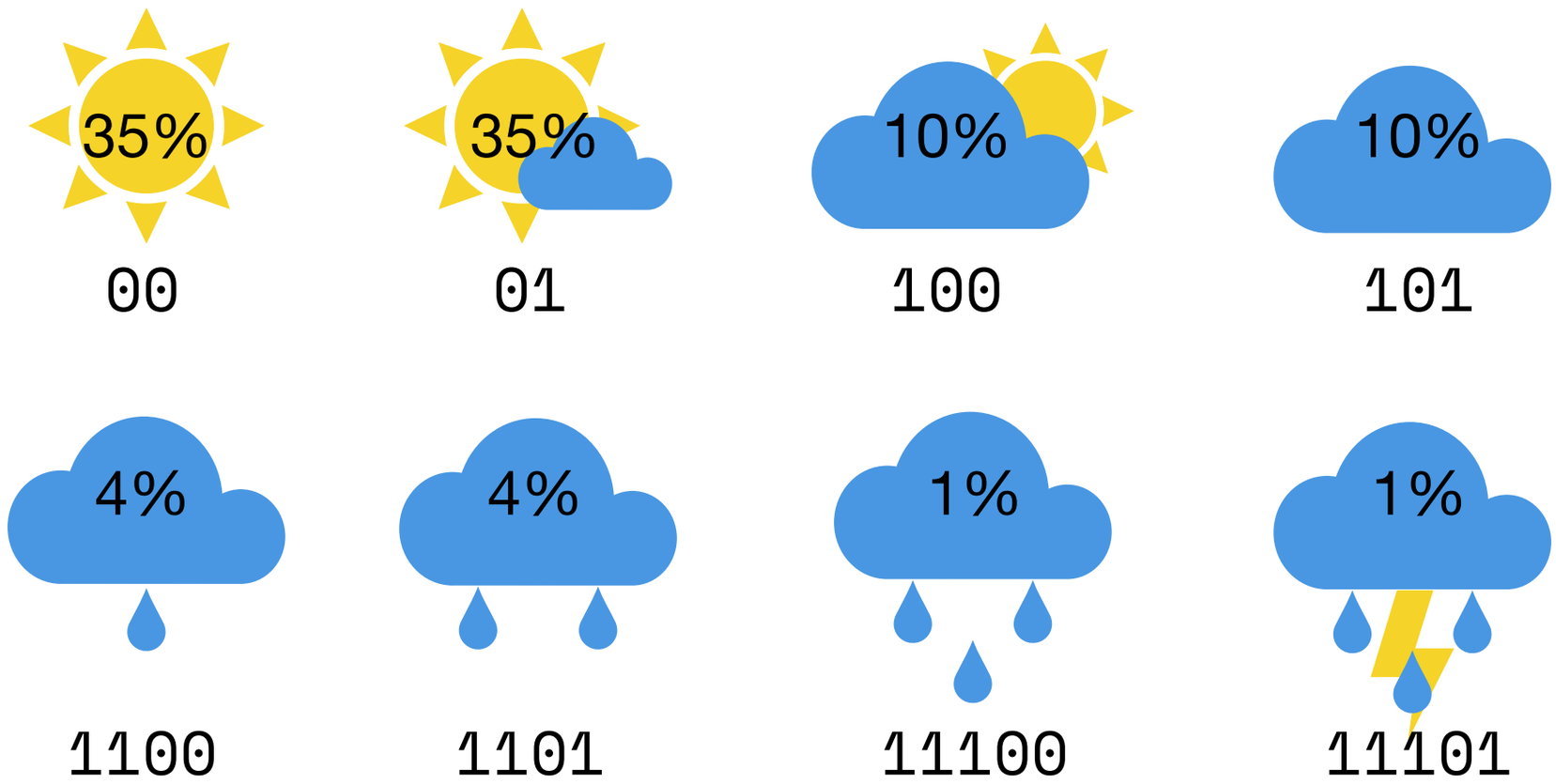


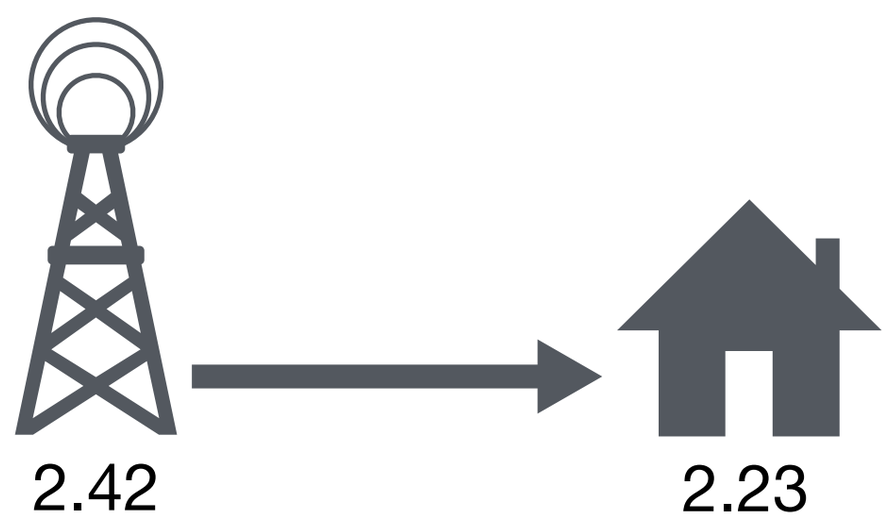
Cross Entropy (Average Message Length) = 3
Entropy = 35% x $\log_2$(35%) + … = 2.23 bits

[2]

Cross Entropy

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



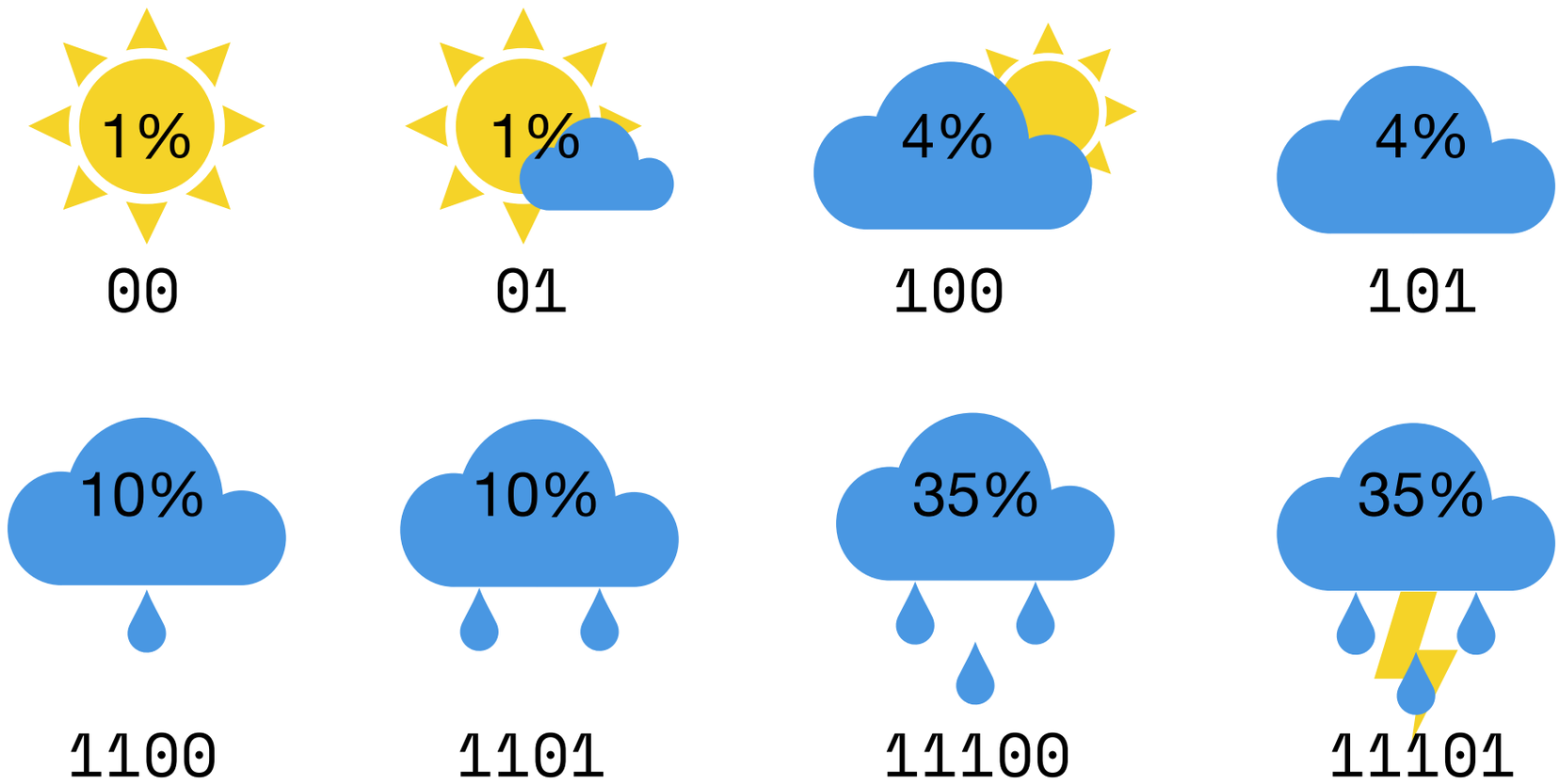


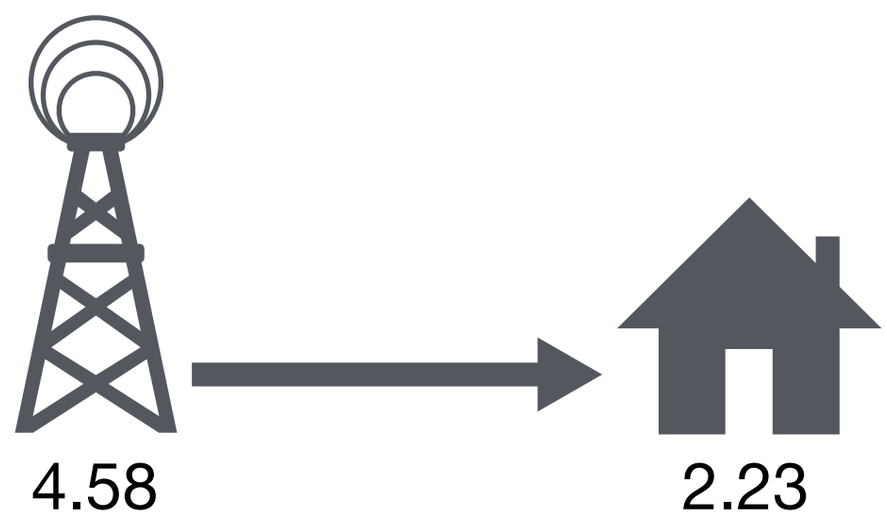
Cross Entropy = 35% x 2 + 35% x 2 + 10% x 3 + 10% x 3 + 4% x 4 + 4% x 4 + 1% x 4 + 1% x 4
= 2.42 bits

[2]

Cross Entropy

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22





Cross Entropy = 1% x 2 + 1% x 2 + 4% x 3 + 4% x 3 + 10% x 4 + 10% x 4 + 35% x 4 + 35% x 4
= 4.58 bits

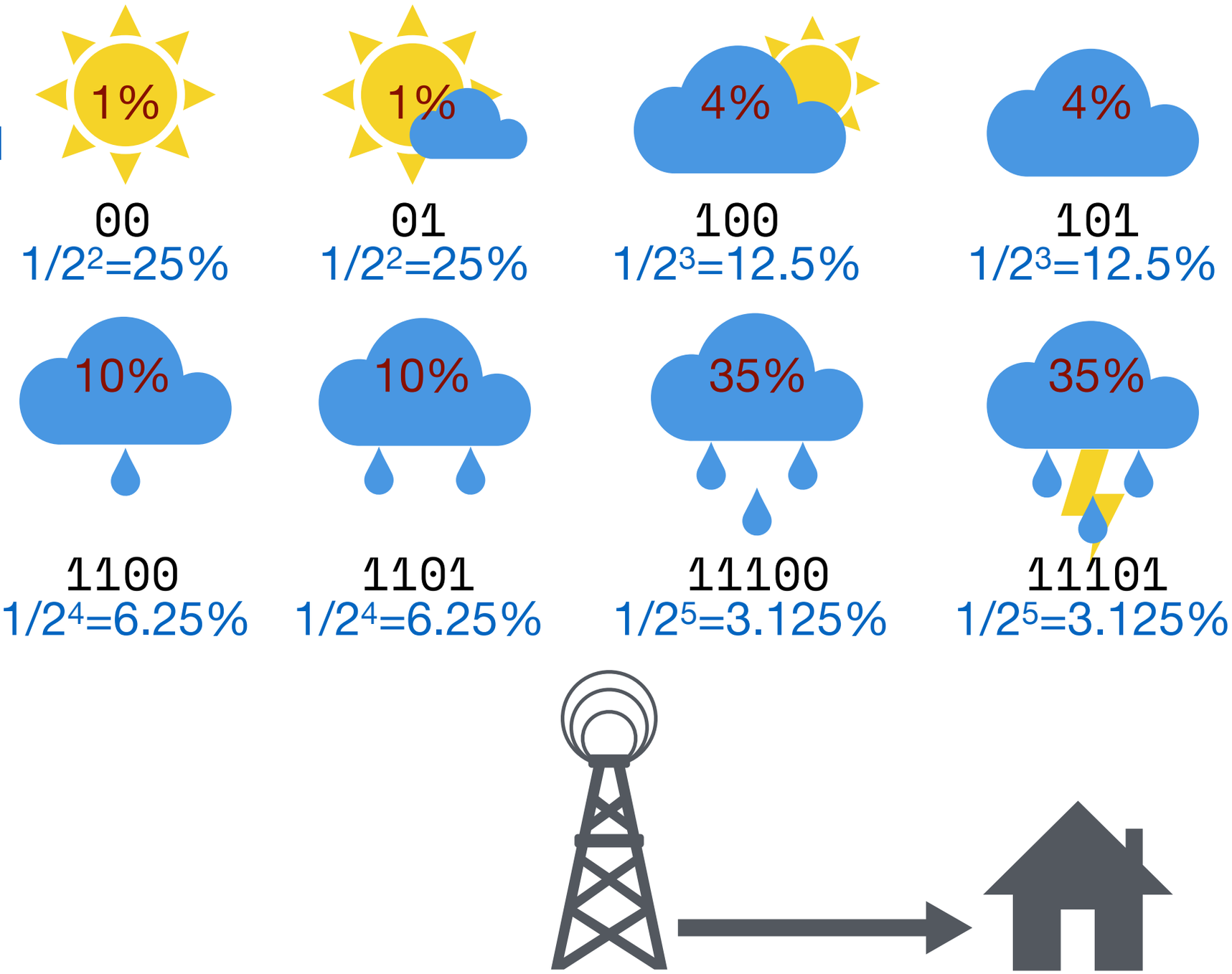
Code makes (wrong) implicit assumptions about distribution!

[2]

Cross Entropy

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

p = true
q = predicted

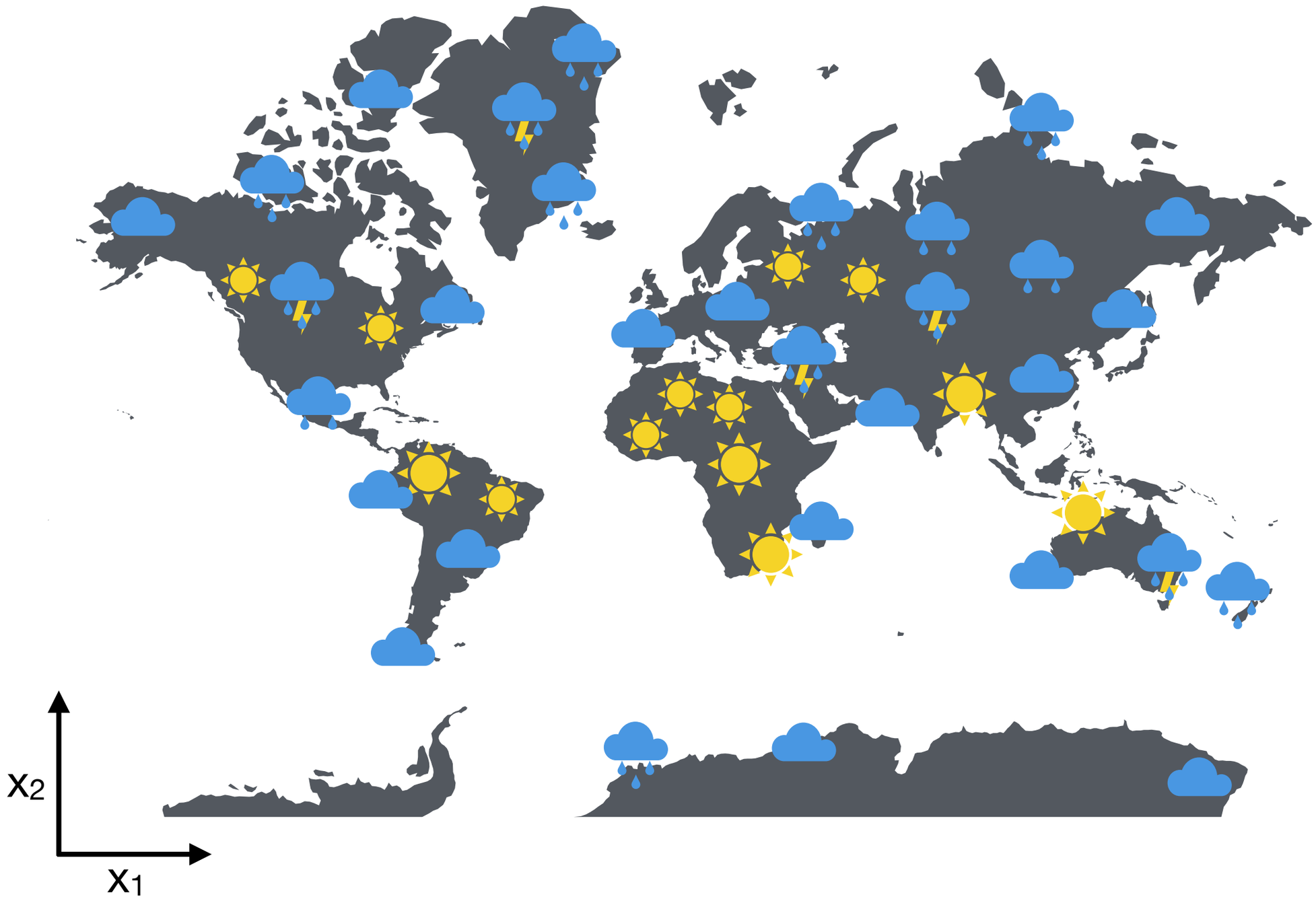


Cross Entropy: $H(p, q) = -\Sigma_i\, p_i \times \log_2(q_i)$
KL Divergence = Cross Entropy - Entropy

[2]

Approximate Probabilities Everywhere

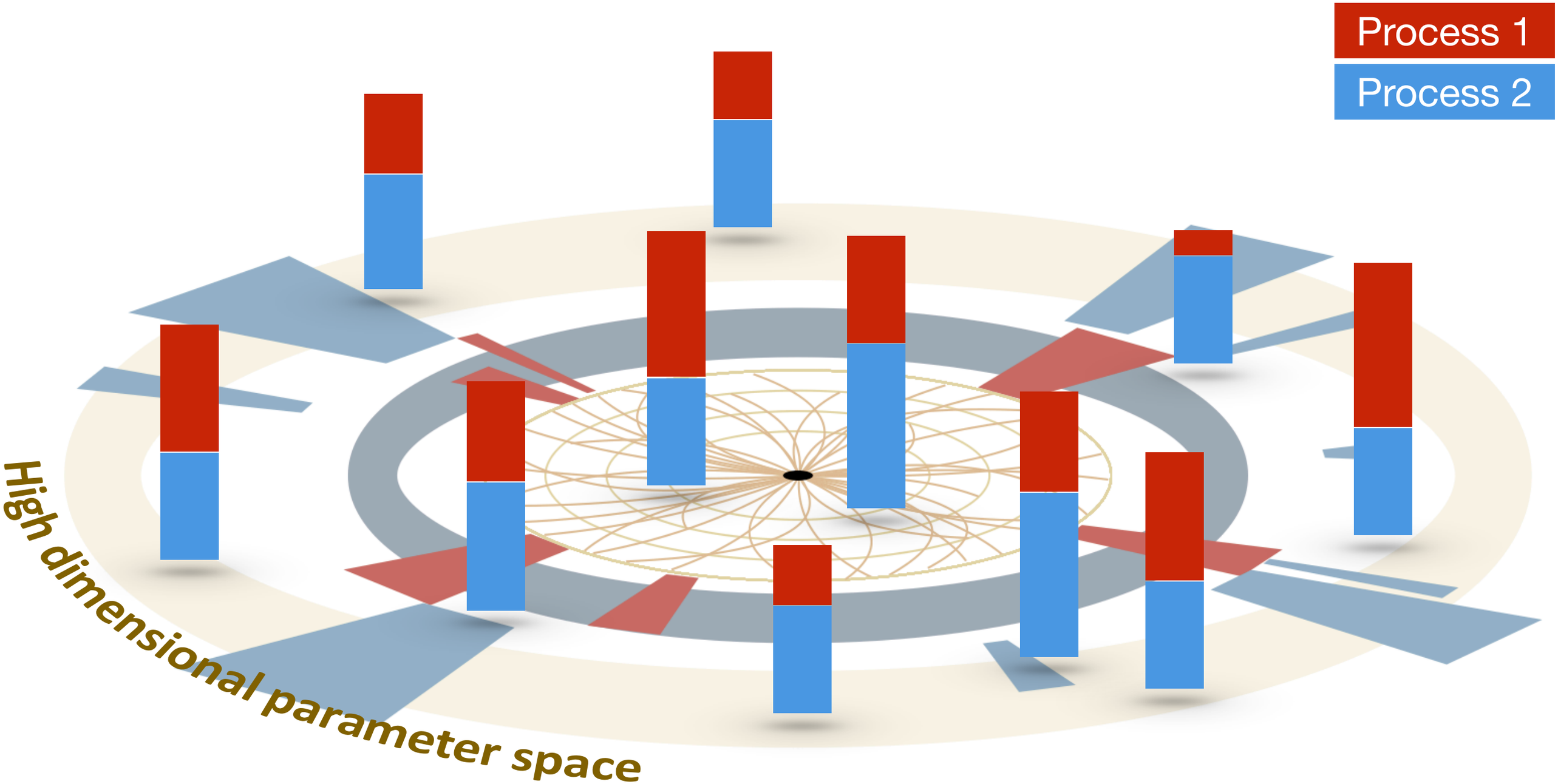
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



Weather probabilities = DNN(x_1, x_2)

Approximate Probabilities Everywhere

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

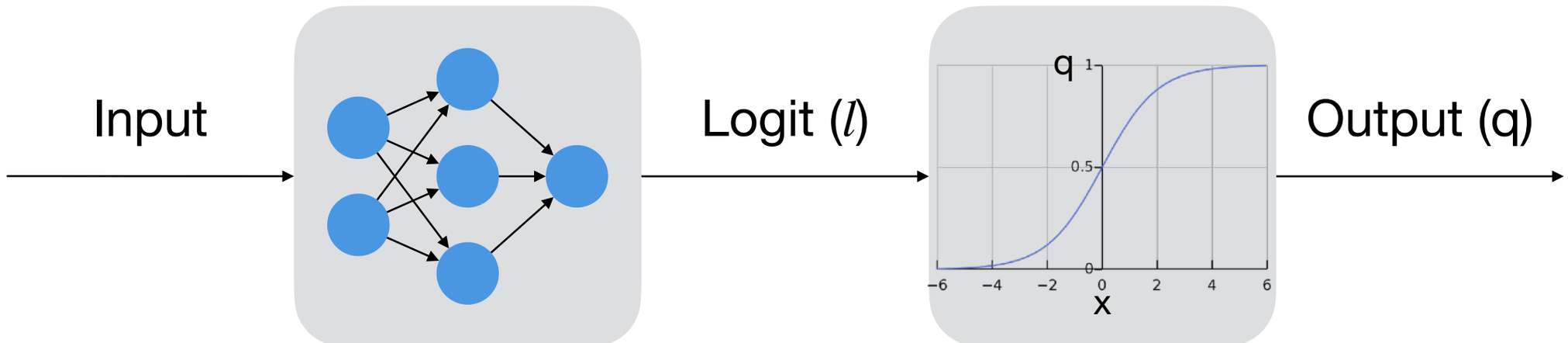


Process probabilities = DNN( )

From Classification to Ratio

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Network with probability like output



- Training minimises cross entropy loss H(p, q)

Loss:

$$\mathscr{L} = \sum_i p_i \log q_i$$

$$= p_1 \log q_1 + p_2 \log q_2 \qquad q = q_1 = (1 - q_2)$$

$$= p_1 \log q + p_2 \log(1 - q)$$

Minimization:

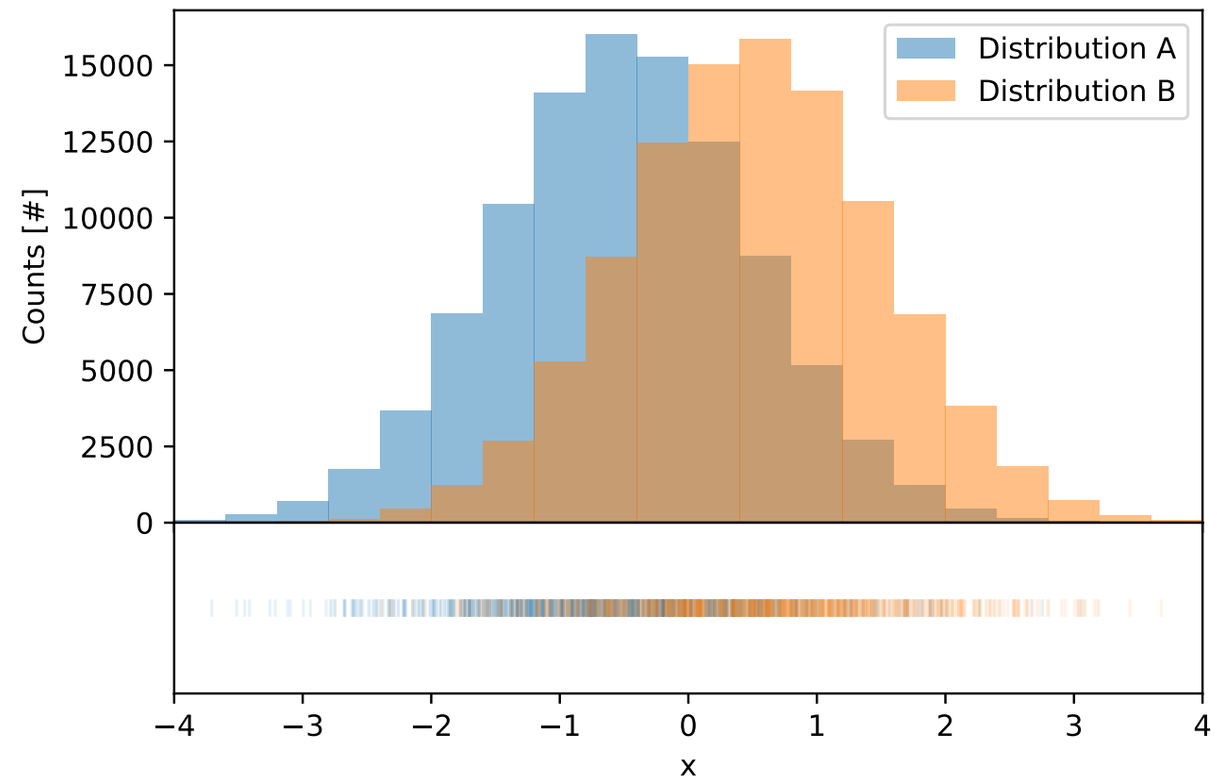
$$\frac{\partial \mathscr{L}}{\partial q} = -\frac{p_1}{q} - \frac{p_2}{1-q} \overset{!}{=} 0$$

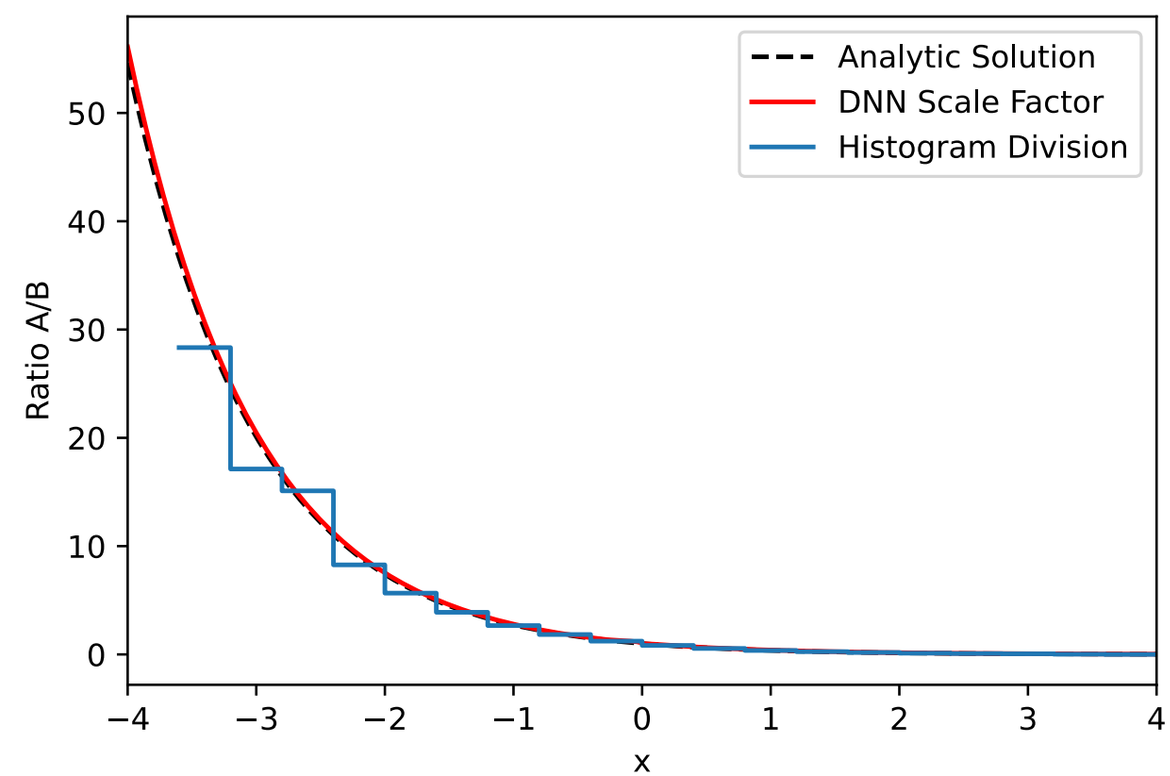
$$\boxed{\Rightarrow r = \frac{p_1}{p_2} = \frac{q}{1-q} = e^l}$$

Measuring Ratios

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Classic solution:
  - Bin events
  - Divide bin contents
  - Bad approximation:
    - Step function
    - Spikes & Infinities
    - N(dim) limited
- DNN solution:
  - DNN as classification (A vs B)
  - Approximate ratio via
    - r = A/B = q / (1-q) = e^l
  - Good approximation:
    - Smooth distribution
    - No statistical spikes
    - N(dim) basically unlimited

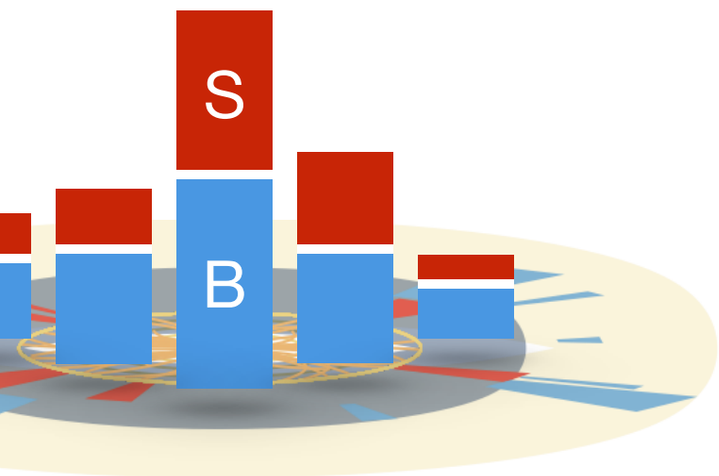


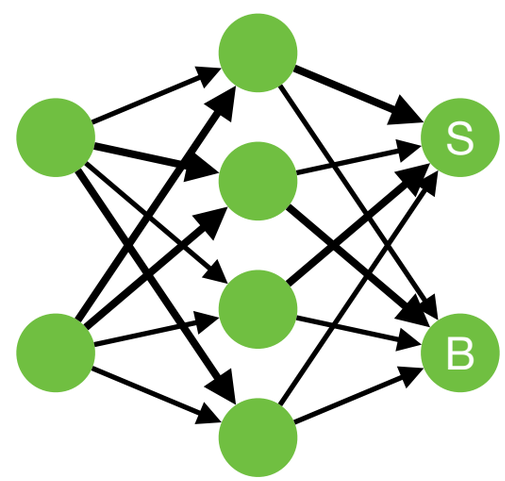


Connection to Your Daily Work

HEP search: Measuring strength μ of signal s over background b

Using Likelihood function: $$\mathscr{L}(\mu) = \prod_{i}^{bins} Poisson(x_i \,|\, \mu \cdot s_i + b_i)$$

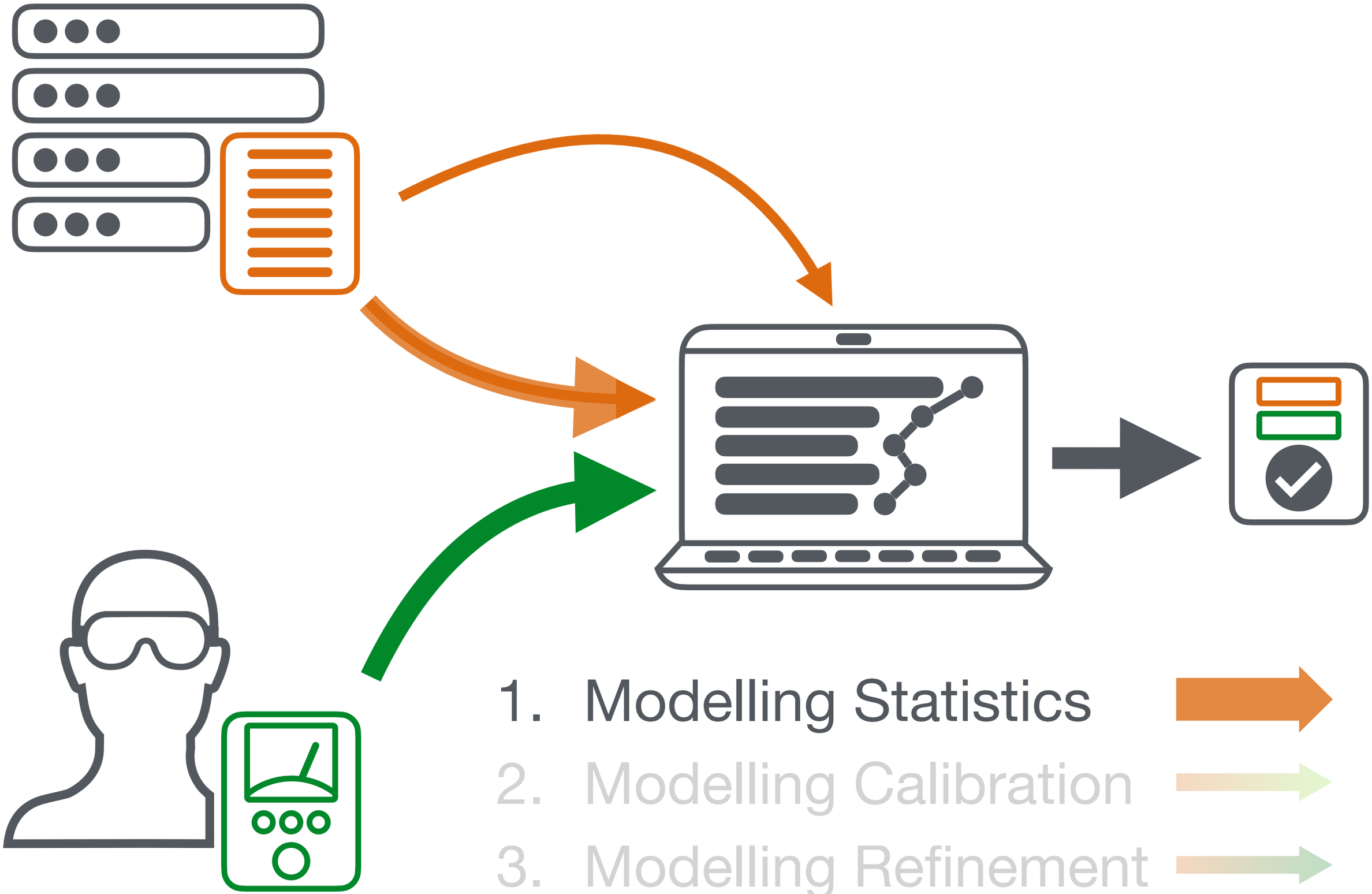




$$\frac{S}{S+B}$$

Signal Classifier trained with cross entropy

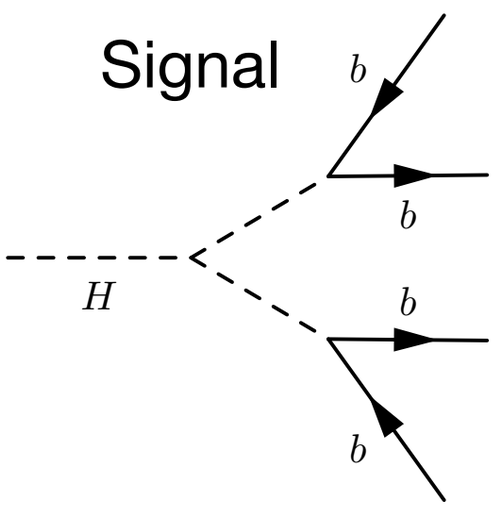
Neyman-Pearson
Optimal variable
(summary statistic)

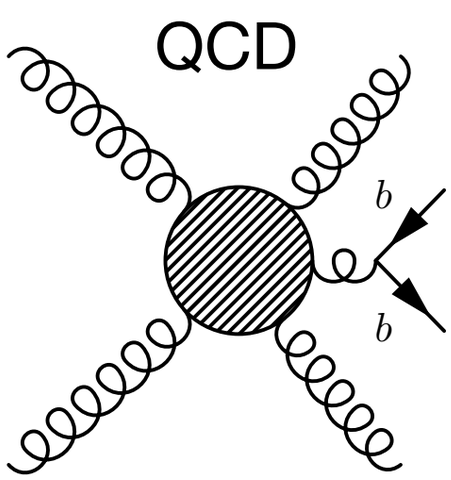

1. Modelling Statistics
2. Modelling Calibration
3. Modelling Refinement

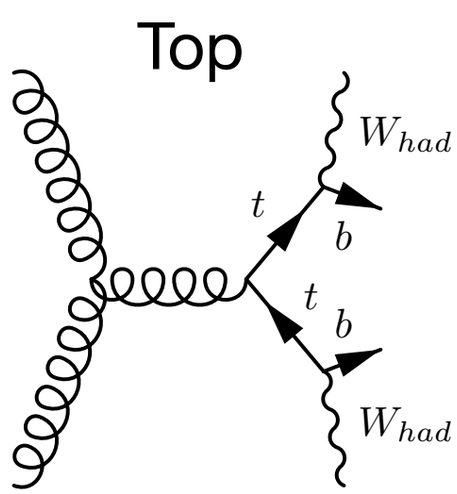
Physics Case: ATLAS HH→4b Analysis

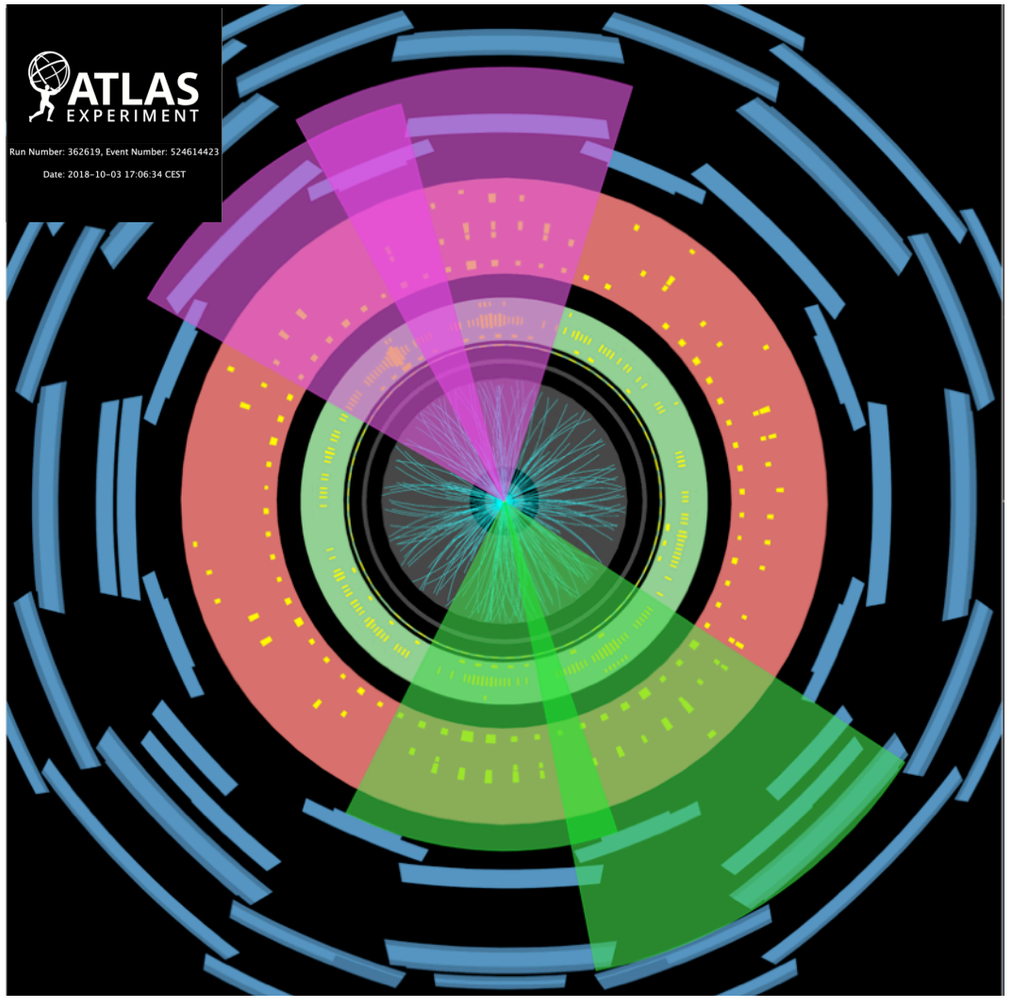
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- ATLAS experiment at LHC (multi purpose detector)
- Search for di-Higgs boson production with 4b final state:
  - Signal signature: 4 b-jets (77% identification efficiency)
  - Backgrounds: 90% QCD, 10% Top
- Challenge: Not enough simulations of QCD
- Solution: Data-driven method (ABCD)



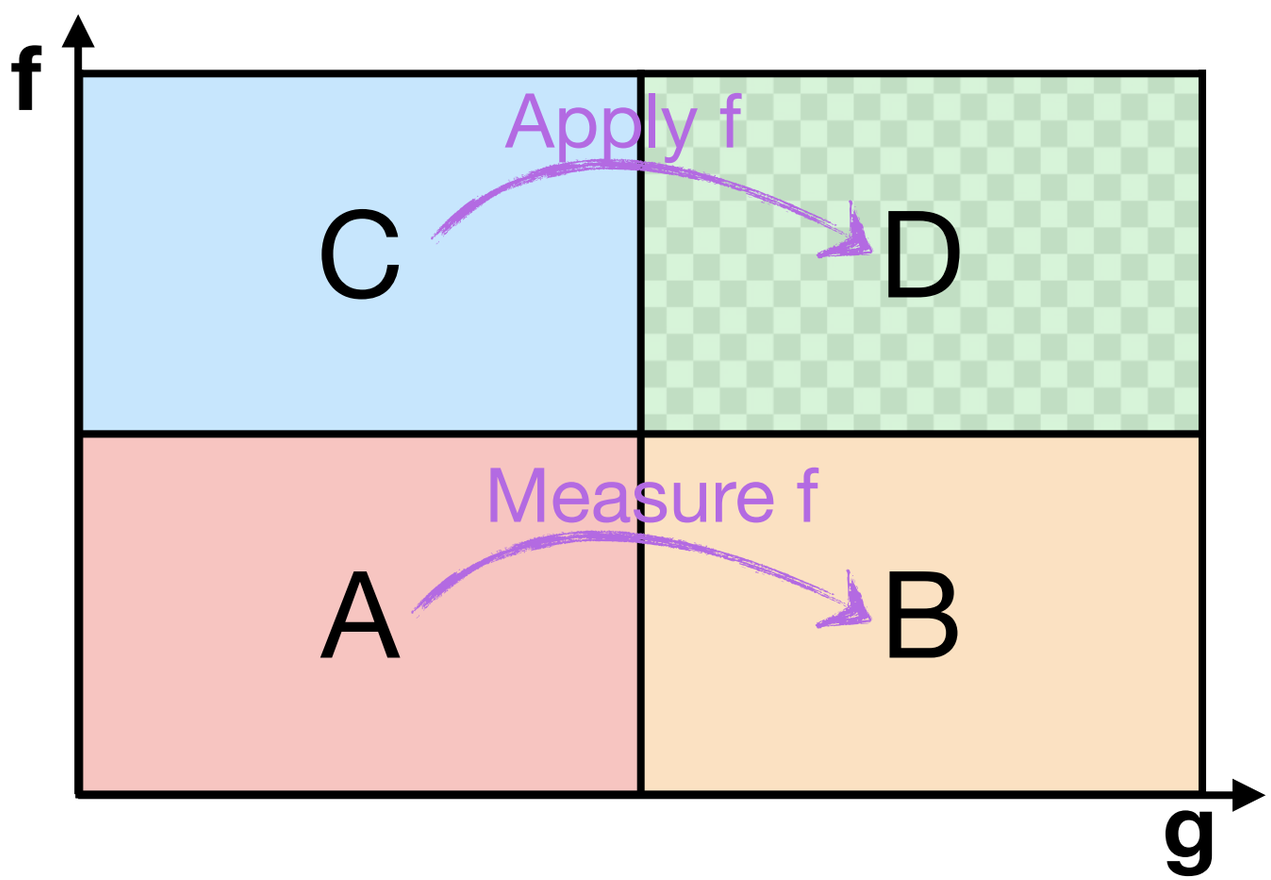






ABCD Method

- Three steps:
  - Divide events into four parts according to two uncorrelated variables (f, g)
  - Measure transfer function: $f = N_B/N_A$
  - Apply transfer function: $N_D = f \times N_C$
- Transfer function f given by ratio (N_B/N_A):
  - Classic: inclusive or binned function (1D $f(x)$ or 2D $f(x_1, x_2)$)
  - New: Un-binned function $DNN(x_1, \ldots, x_N)$

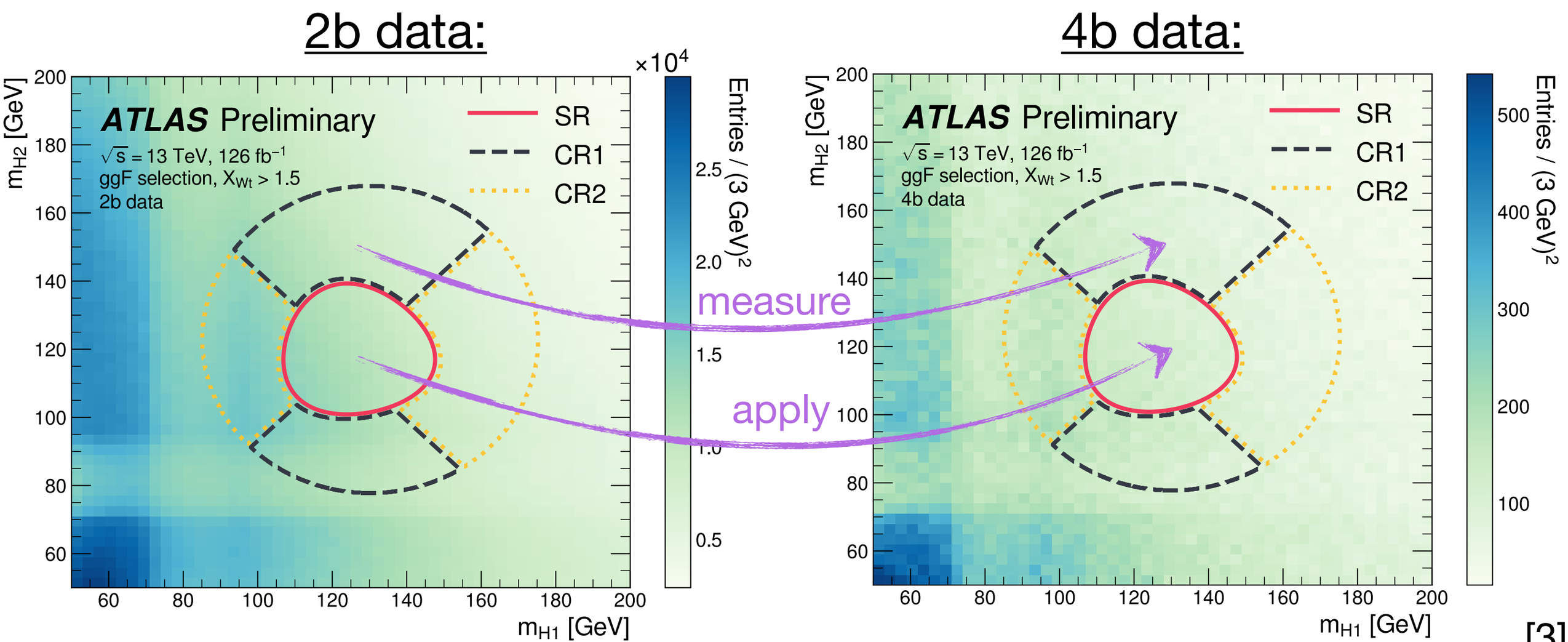


ABCD Method in ATLAS HH→4b Analysis

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Problem: Not enough QCD 4b events
- Solution: ABCD 2b → 4b increases statistics O(100)
  - Dimensions f = 2b/4b, g = SR/CR1
  - Median of 100 trainings (slide 29)
  - CR2 for systematic uncertainties (slide 30)

$$w(\vec{x}) = \frac{p_{4b}(\vec{x})}{p_{2b}(\vec{x})}$$



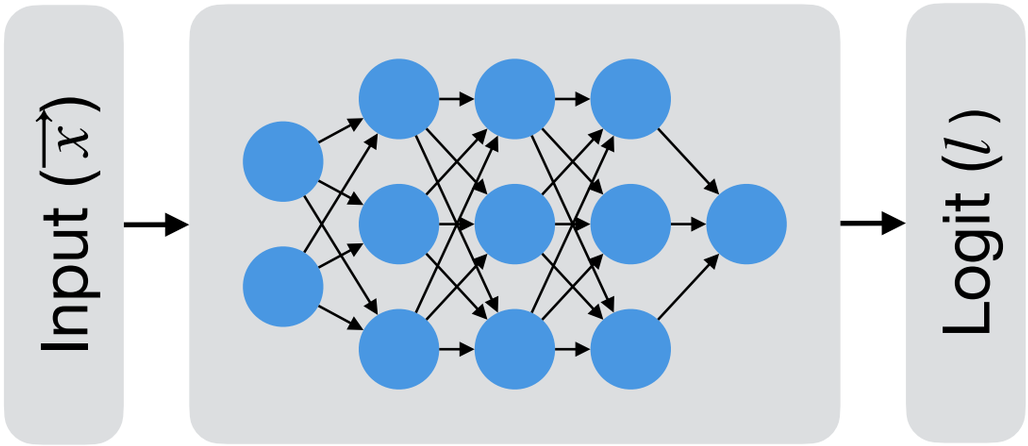
[3]

Architecture, Inputs and Training

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

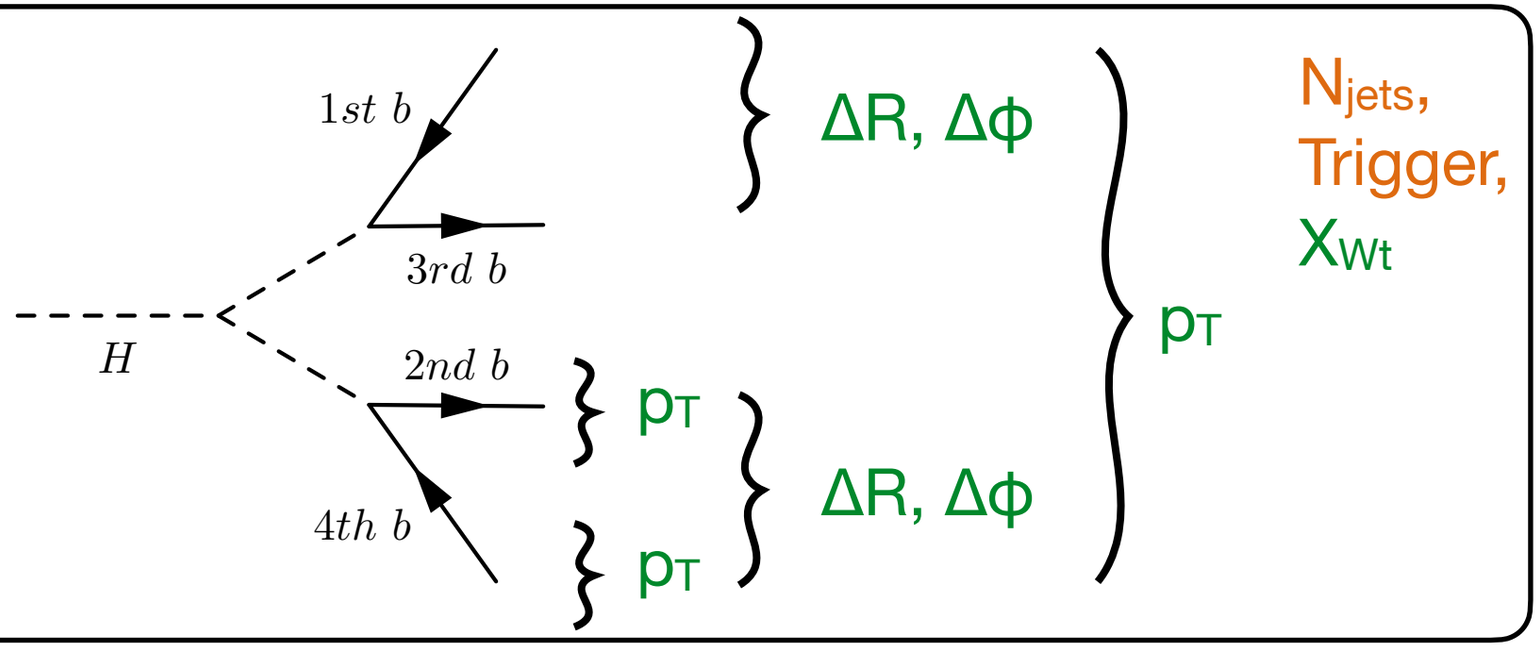
Architecture

- 3 layers:
  - 50 (20) nodes in GGF (VBF) region
  - ReLU activation
- Single node linear output (logit)



Input variables ($\vec{x}$)

- Kinematic
- Event-level
- DNN can understand physics of the process



Loss function

- Exponential loss defined on logit output
- Related to cross-entropy loss (backup)
- More information: 1911.00405

$$\mathscr{L} = p_1 e^{-l/2} + p_2 e^{l/2}$$

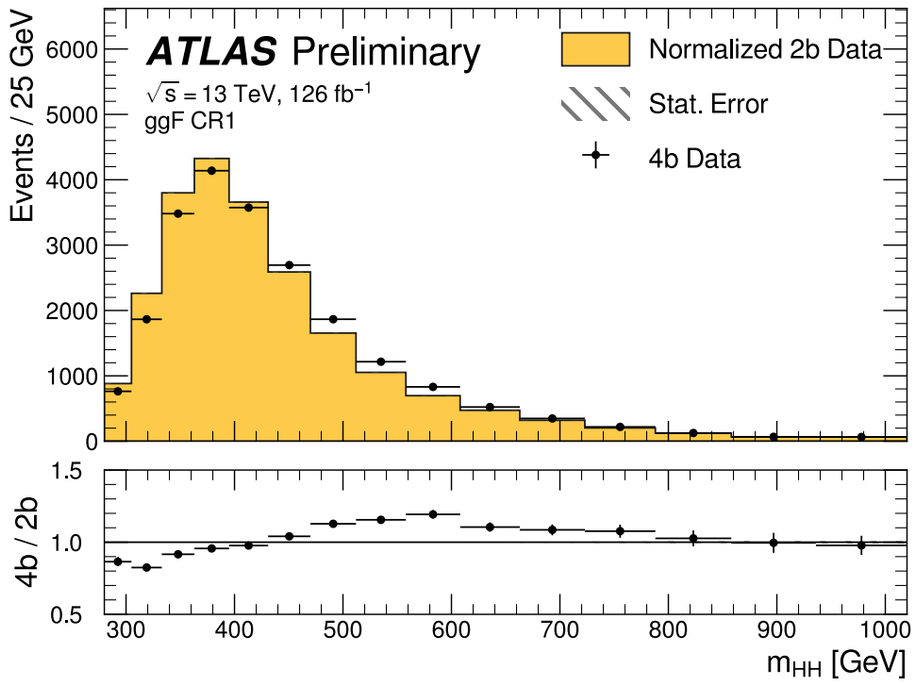
Validation in Control Region (CR1)

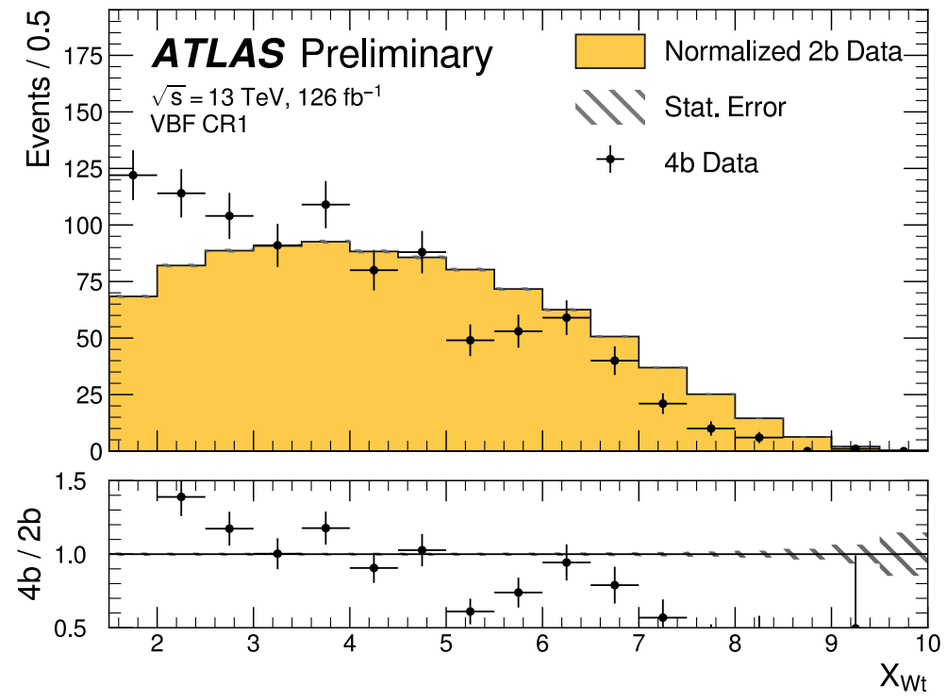
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

HH invariant Mass

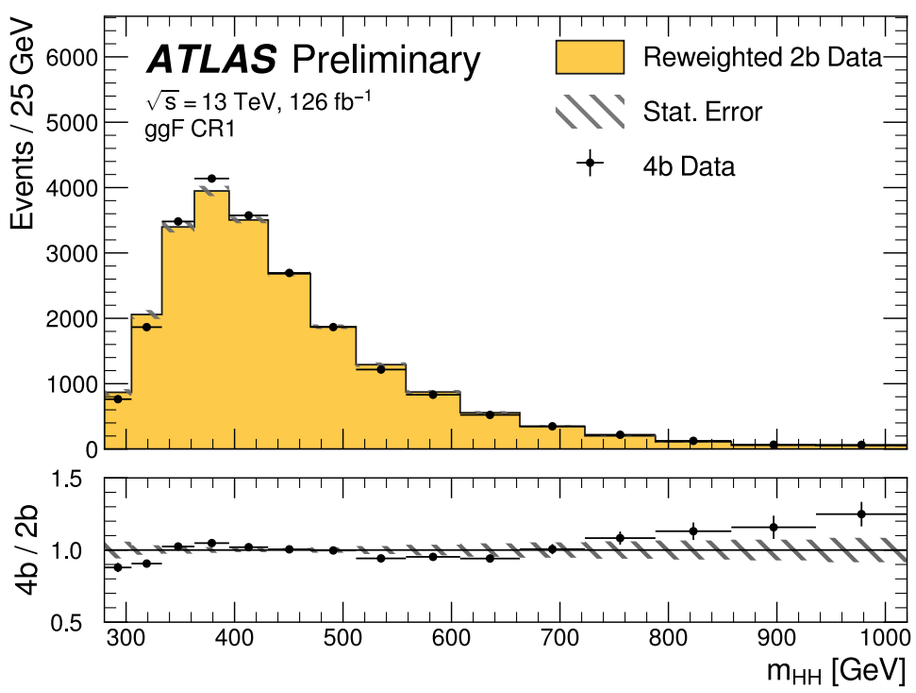
$$X_{W_t} = \sqrt{\left(\frac{m_W - 80.4\ \text{GeV}}{0.1\ m_W}\right)^2 + \left(\frac{m_t - 172.5\ \text{GeV}}{0.1\ m_t}\right)^2}$$

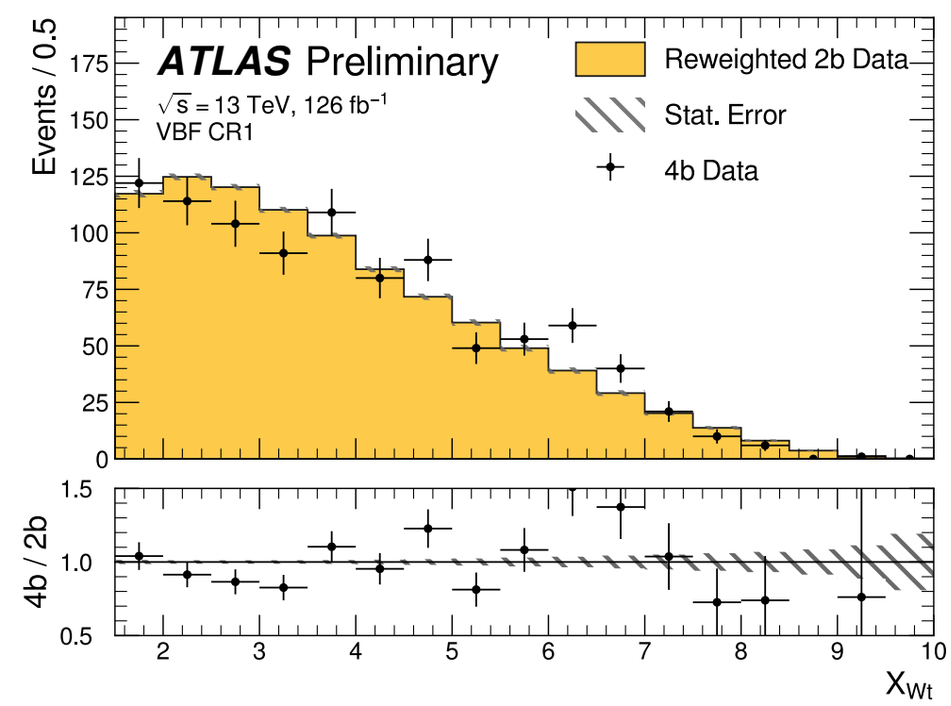
Before:





After





[3]

Systematic Uncertainties (Statistical Origin)

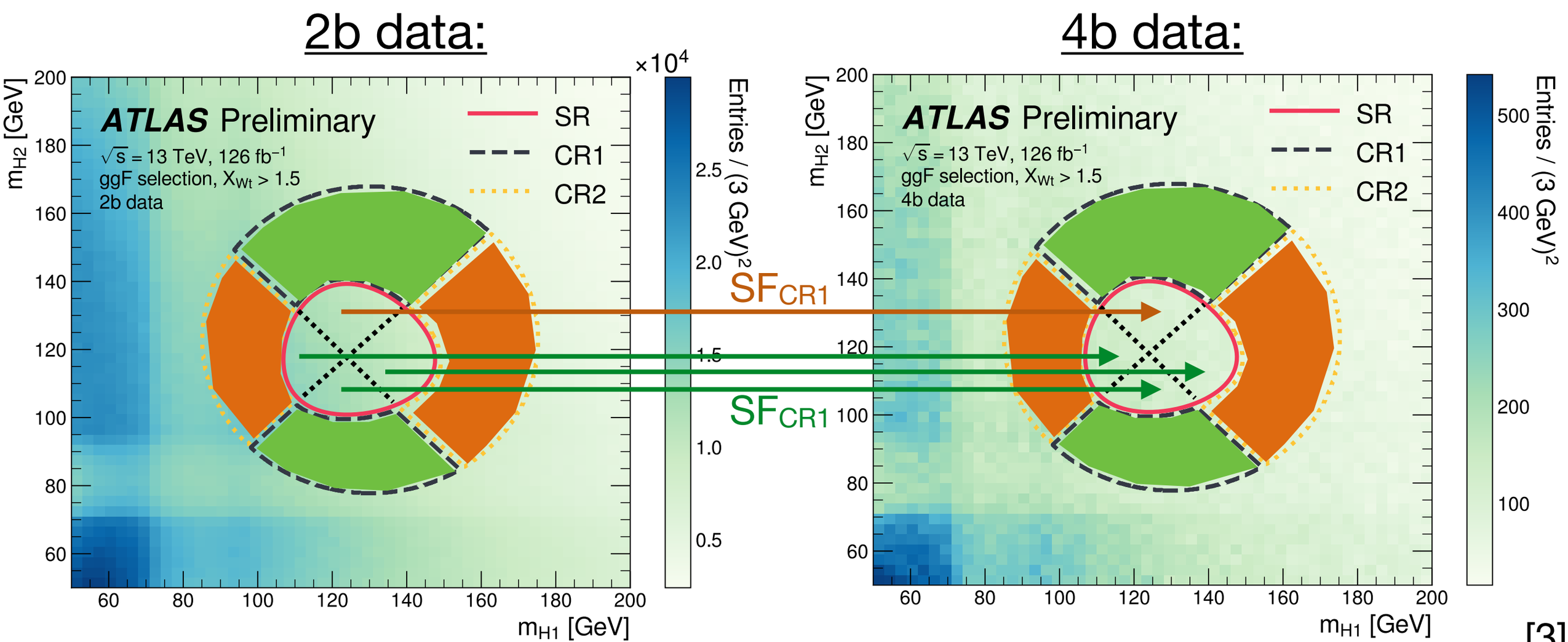
- Measurement region (A → B) has limited statistics
- Classic way: Propagate poisson uncertainties through A/B
- DNN way:
  - Train Deep Ensemble of 100 Networks
    - Using weighted Loss function: $\mathscr{L} = \sum_{e \in events} w_e \cdot \mathscr{L}_e$
    - In each training draw w_e from poisson with mean 1, total yield constant
  - Results in 100 weighted shapes: mean as nominal, std. dev. as uncertainty

[4]

Systematic Uncertainties (Systematic Origin)

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Split SR in 4 quadrants
- Measurement: CR1 produces SF_{CR1} and CR2 produces SF_{CR2}
- Application:
  - Apply SF_{CR1} to three quadrants
  - Apply SF_{CR2} to the last quadrant
- Difference to nominal distribution (signal classifier m_{HH}) as additional uncertainty

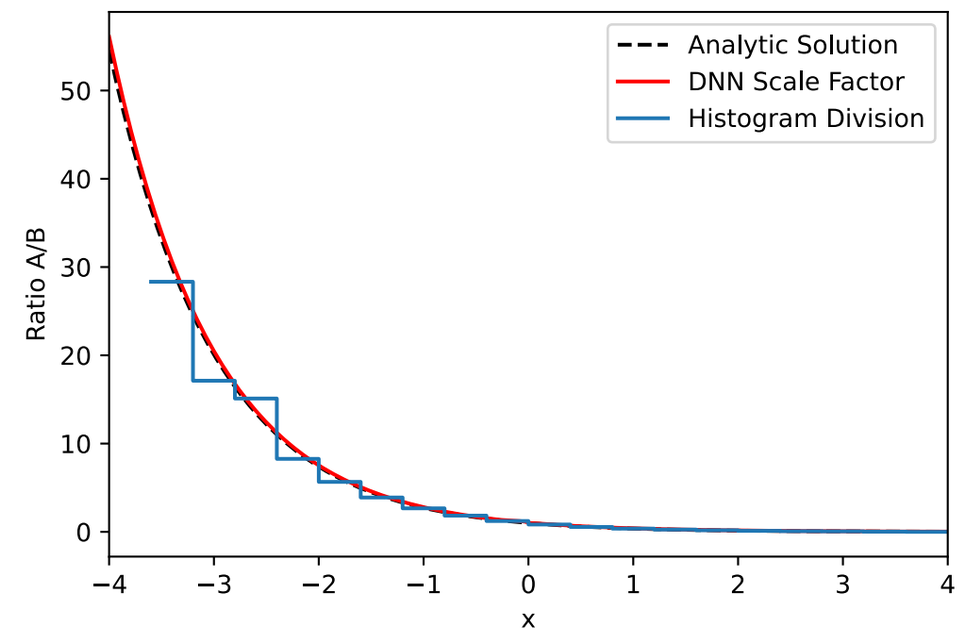


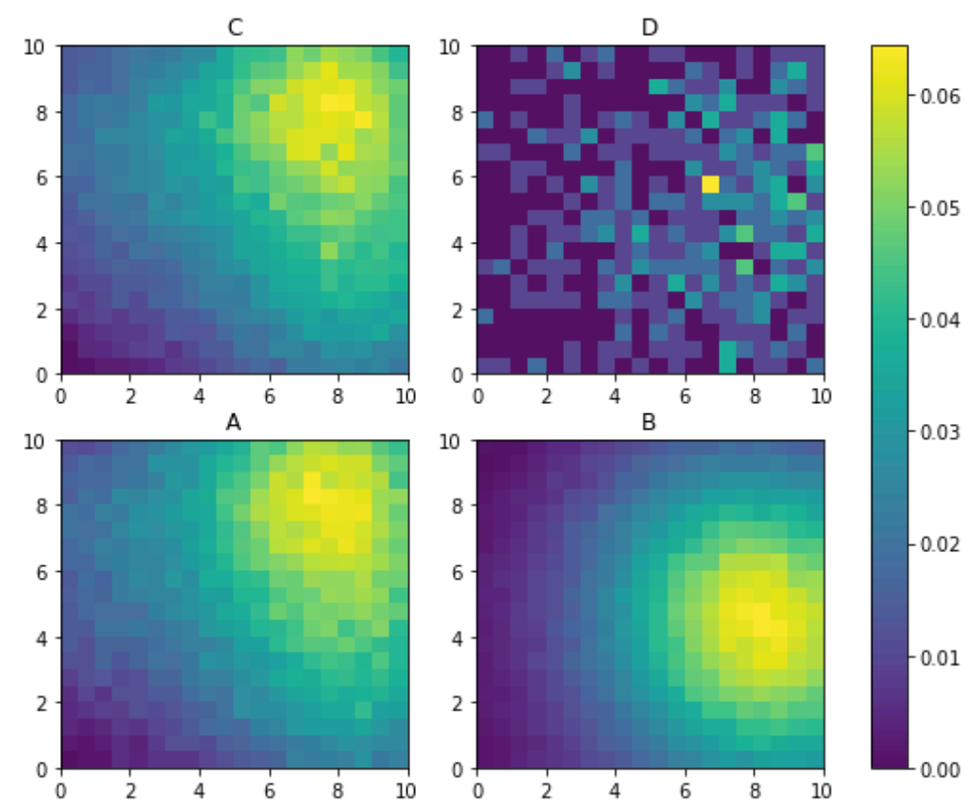
[3]

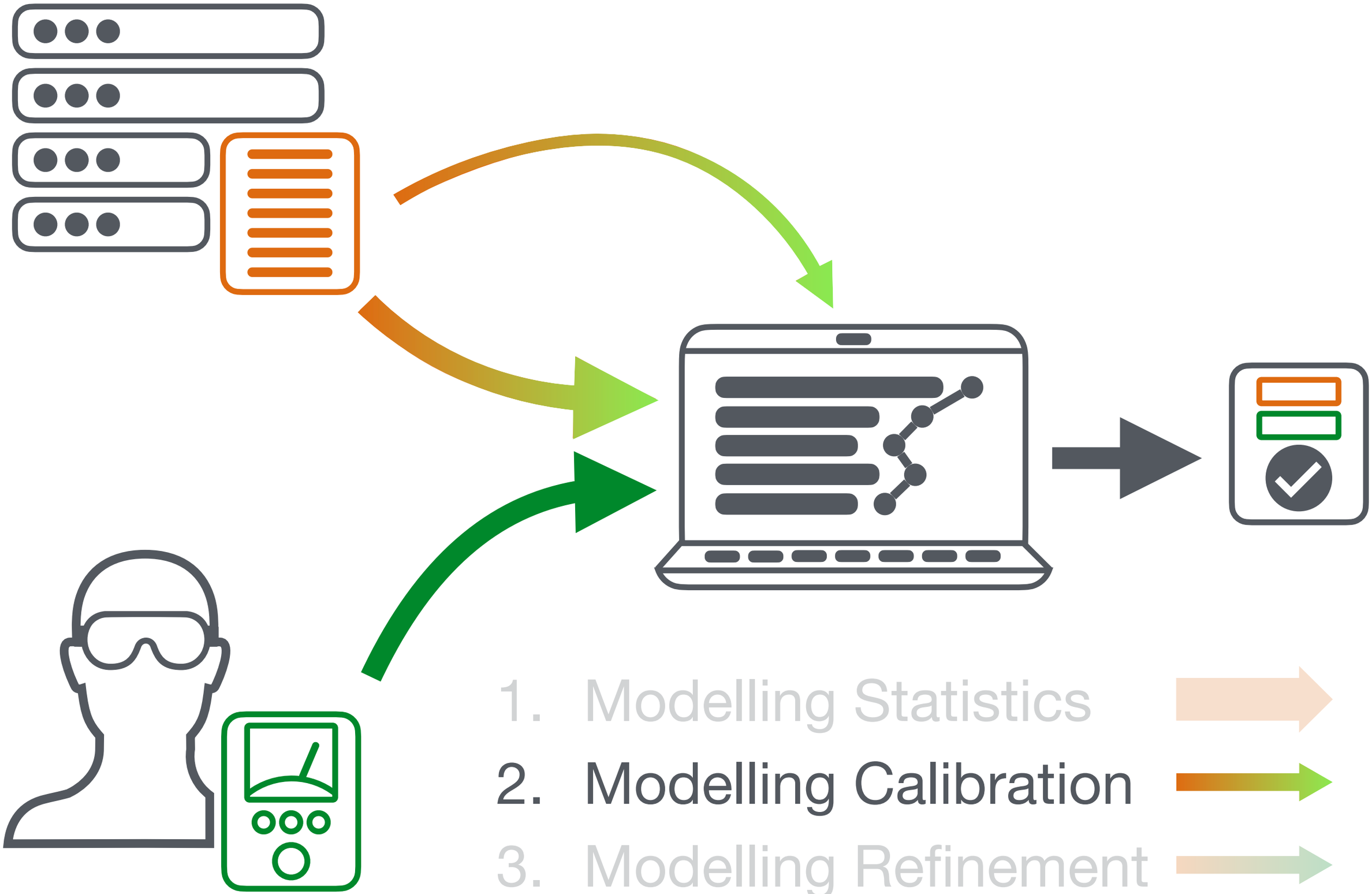
Exercises

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Link to all exercises
- Exercise 0 - Basics (ratio):
  - Measuring a ratio with a DNN
  - Todos:
    - Setup Model
    - Compile Model
    - Train Model
    - Extract ratio
  - Solution
- Exercise 1 - ABCD:
  - Multiple n-dimensional Gaussians
  - Measure A→C and apply to B→D
  - Todos - What happens for?
    - Very complicated functions
    - Low statistics
    - Many dimensions

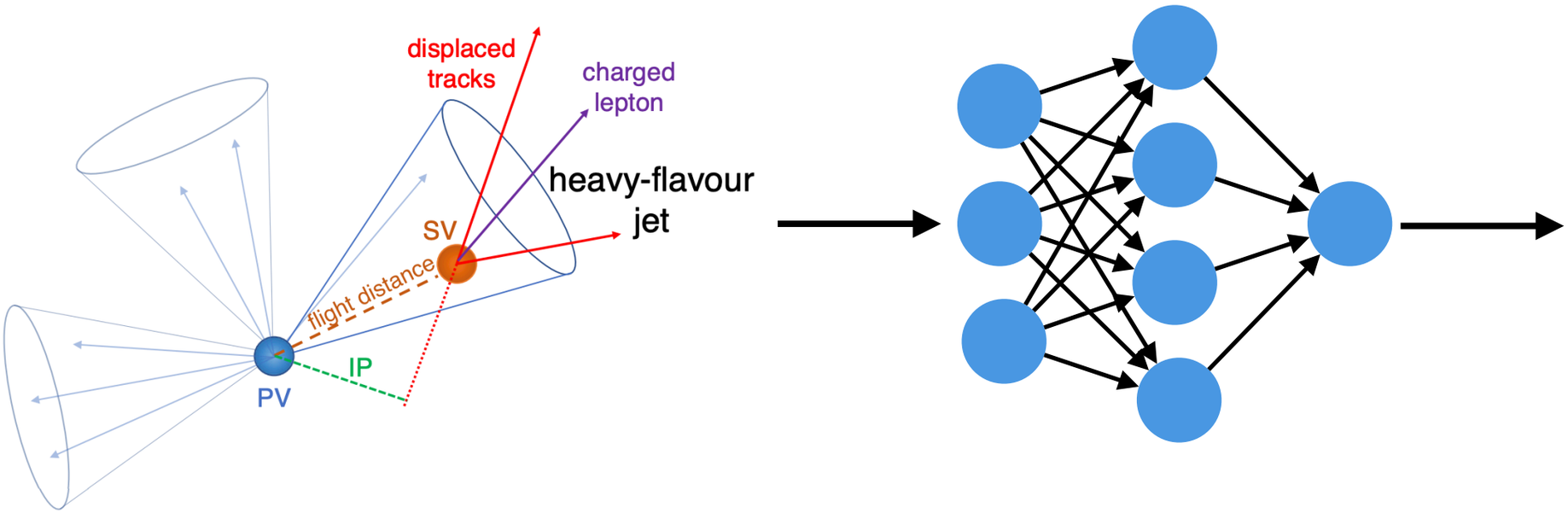





1. Modelling Statistics
2. Modelling Calibration
3. Modelling Refinement

Problem Introduction (b-tagging)

- Many HEP analyses are using b jets (Higgs analyses, top analyses, …)
- Identification of b via b-tagging:
  - b mesons are special:
    - higher mass → higher pT
    - longer lifetime → secondary vertex (SV)
    - weak decay → often lepton in jet
  - MVA Algorithm (e.g. DeepCSV):
    - Assign score to every jet:
      - P(B): probability to originate from b
    - Trained on simulation

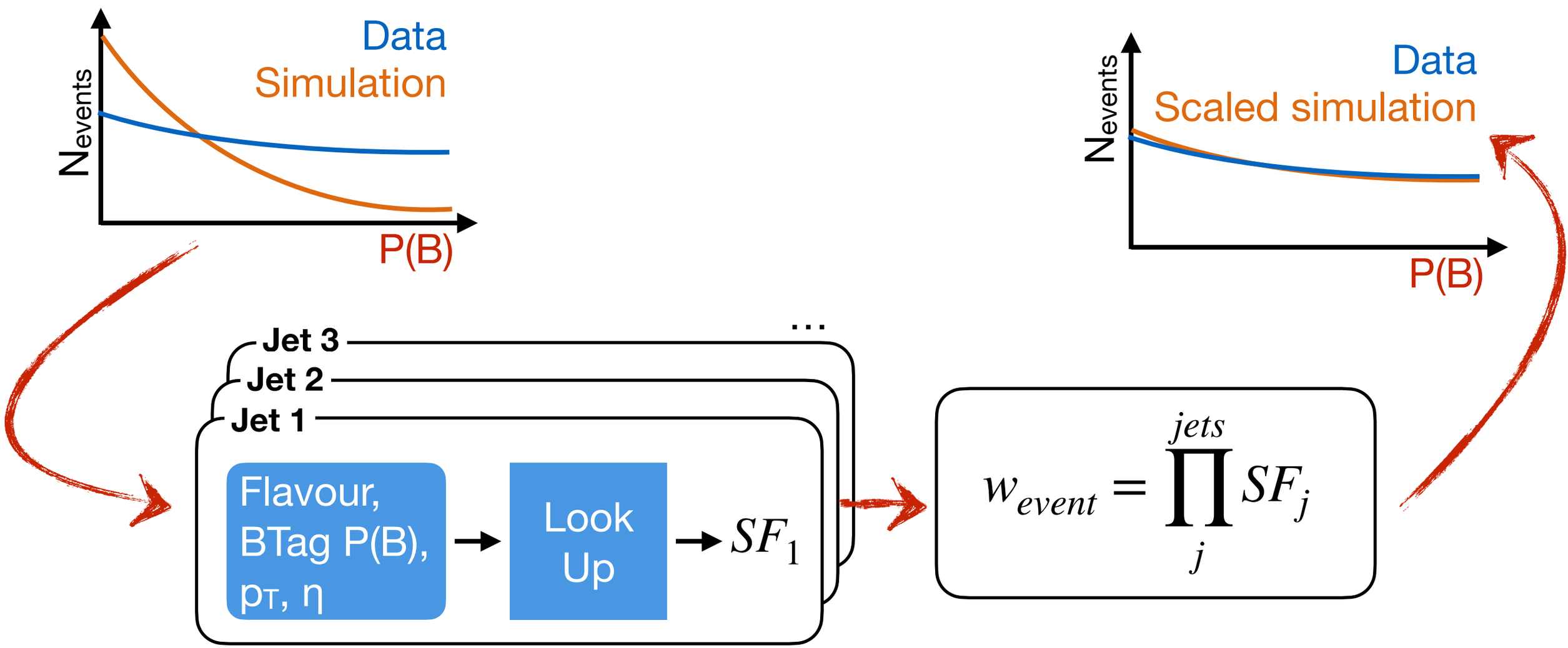


Thats a b-jet, I am sure with P(B)=77%!

Problem Introduction (Calibration)

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Challenge: Calibration needed!
  - Insufficient modelling of data
  - Differences in b-tagging between data and simulation
- Solution: Shape calibration of b-tagging
  - Scale factors (SF_j) for every jet (j) build up event weight (w_{event})
  - Careful: Do not change (sculpt) other variables (i.e. pT, η)

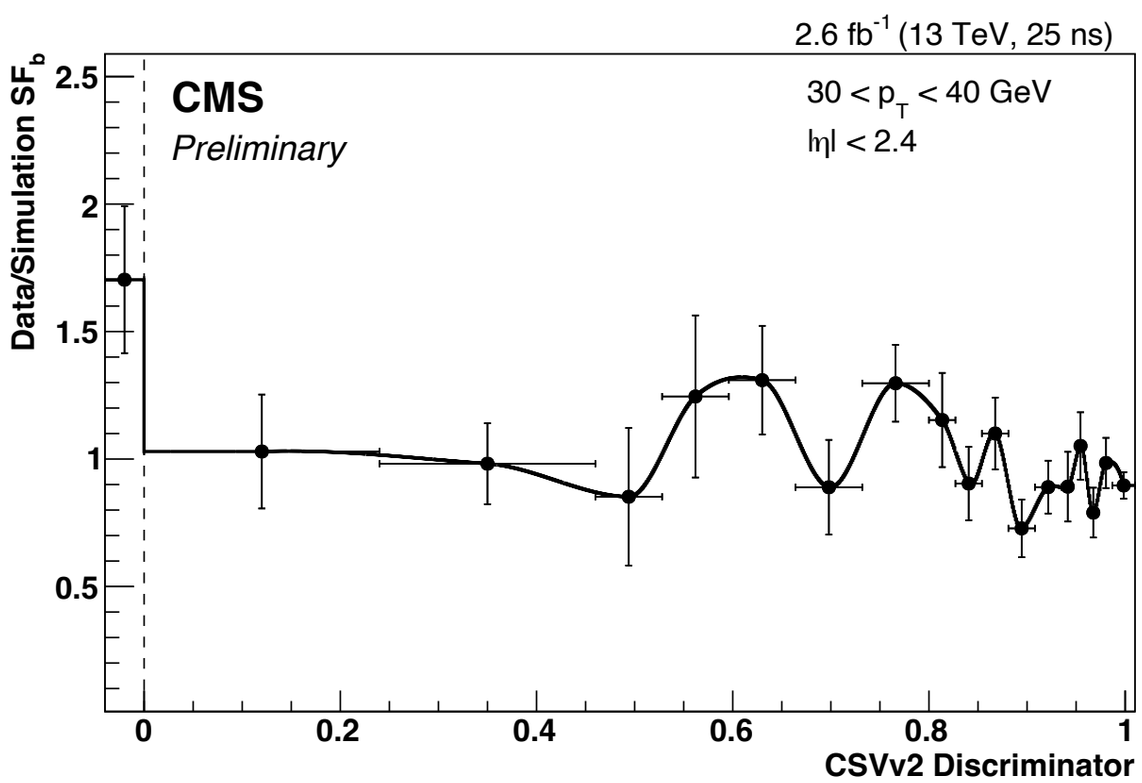


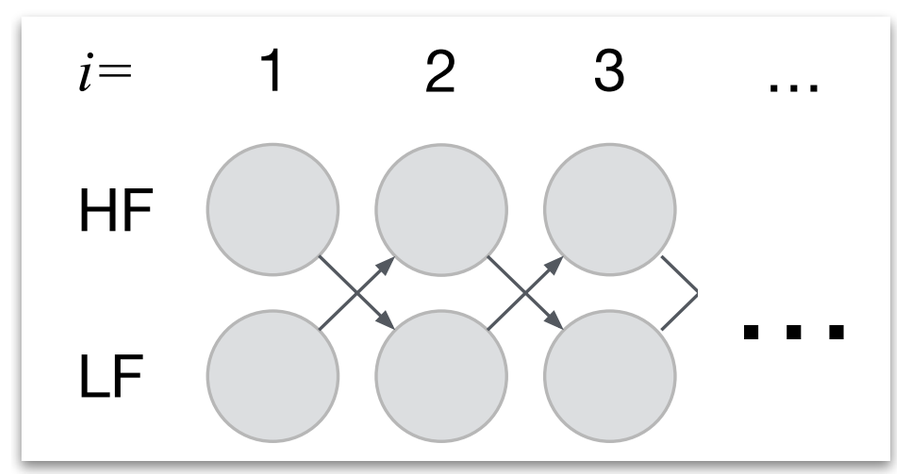
Classic Method: Iterative Fit

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Method:
  1. Tag & probe 2 jets in flavour f (heavy & light) enriched regions R_f
  2. Bin events in pT, η and b-tag discriminator of probe jet (3D)
  3. Infer jet scale factor per bin as: $$SF_{f,i+1} = \left.\frac{Data - \omega_i \cdot MC_{\neg f}}{MC_f}\right|_{R_f}$$
  4. Obtain event weight via $$\omega_{i+1} = \prod_i^{jets} SF_{j,i+1}$$

iterate





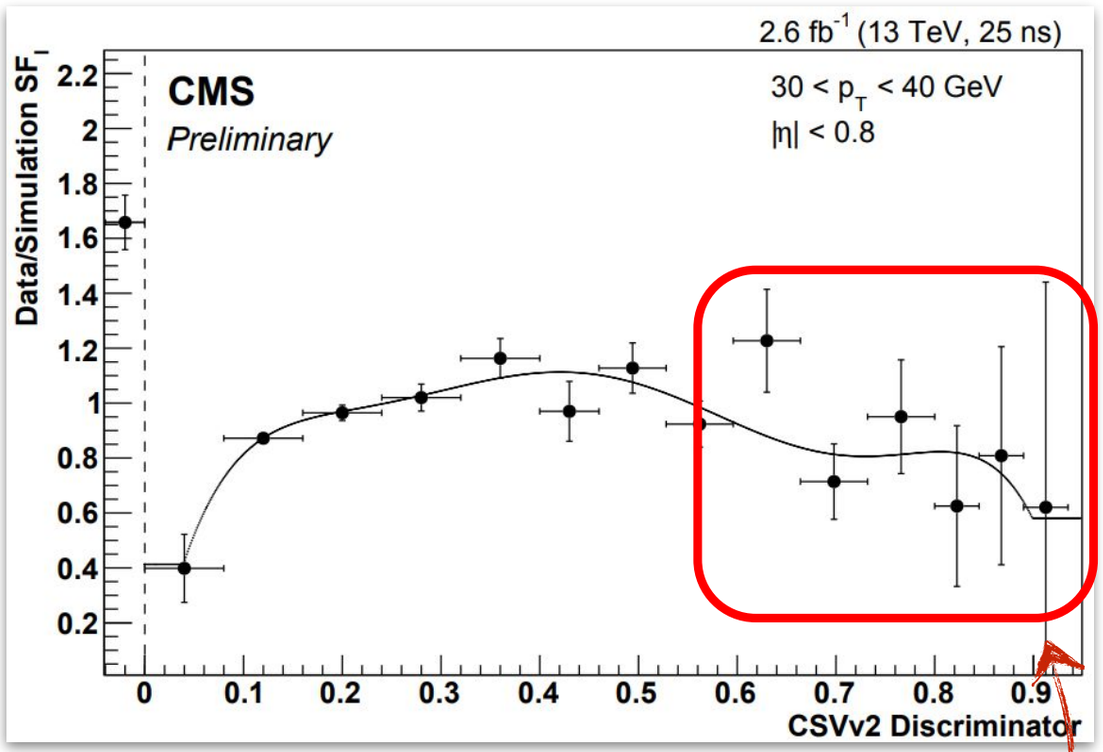
$$SF_f(p_T, \eta, \text{b-tag})$$

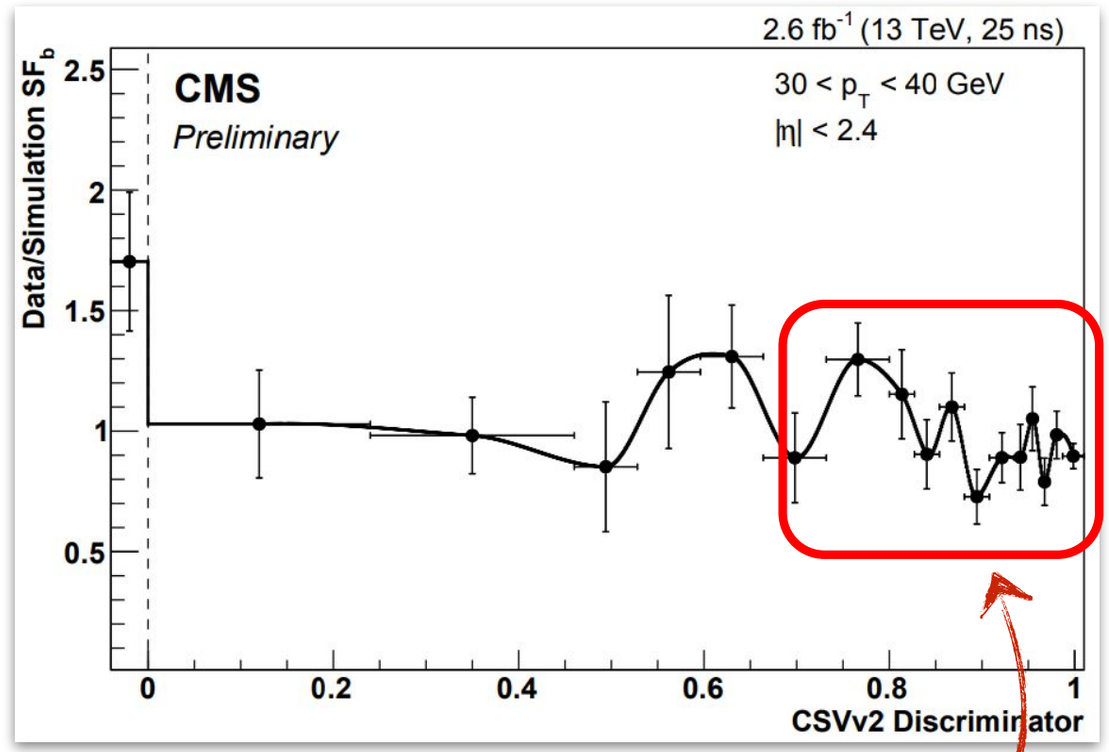
[5]

Traditional Scale Factors Implications

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Binned in pT, η, b-tag value of probe jet:
  - bins must have sufficient statistics
  - number of input variables limited
- Only two regions (heavy/light)
- Needs to be smoothened by fitting a (hand-tuned) function





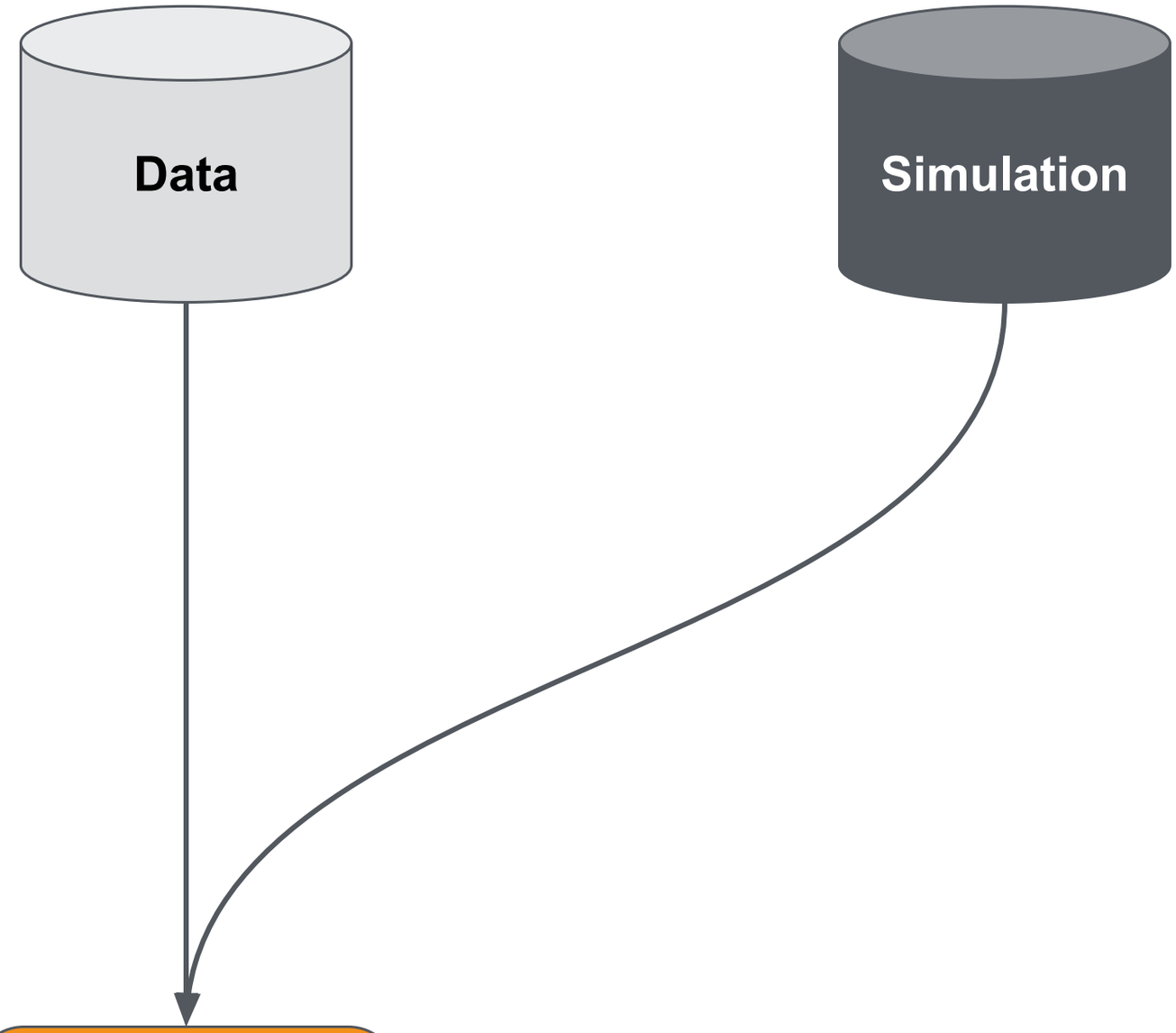
Large uncertainties due to low statistic bins

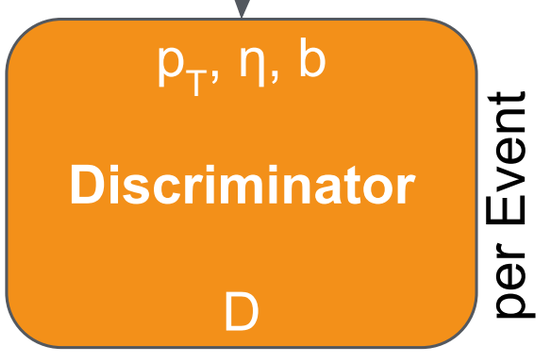
Features of fitting function

[5]

Deep Learning Approach Architecture

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

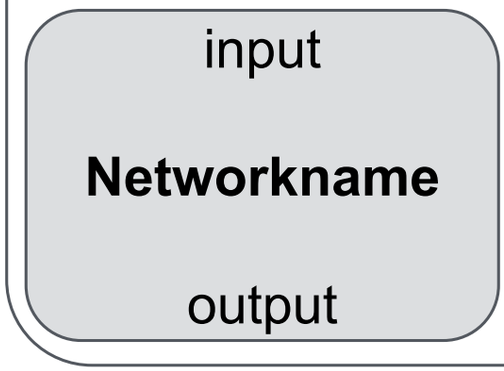




Variables:

- p_T
- η
- b: BTag discr.
- f: gen. flavor

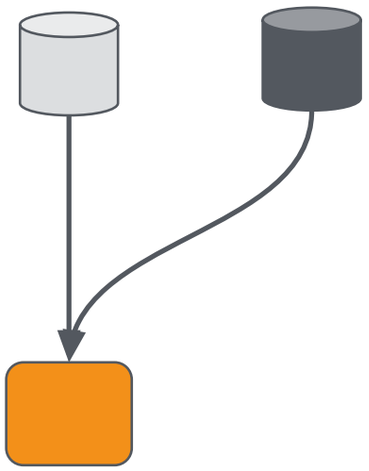
Legend:



Deep Learning Approach Loss 1

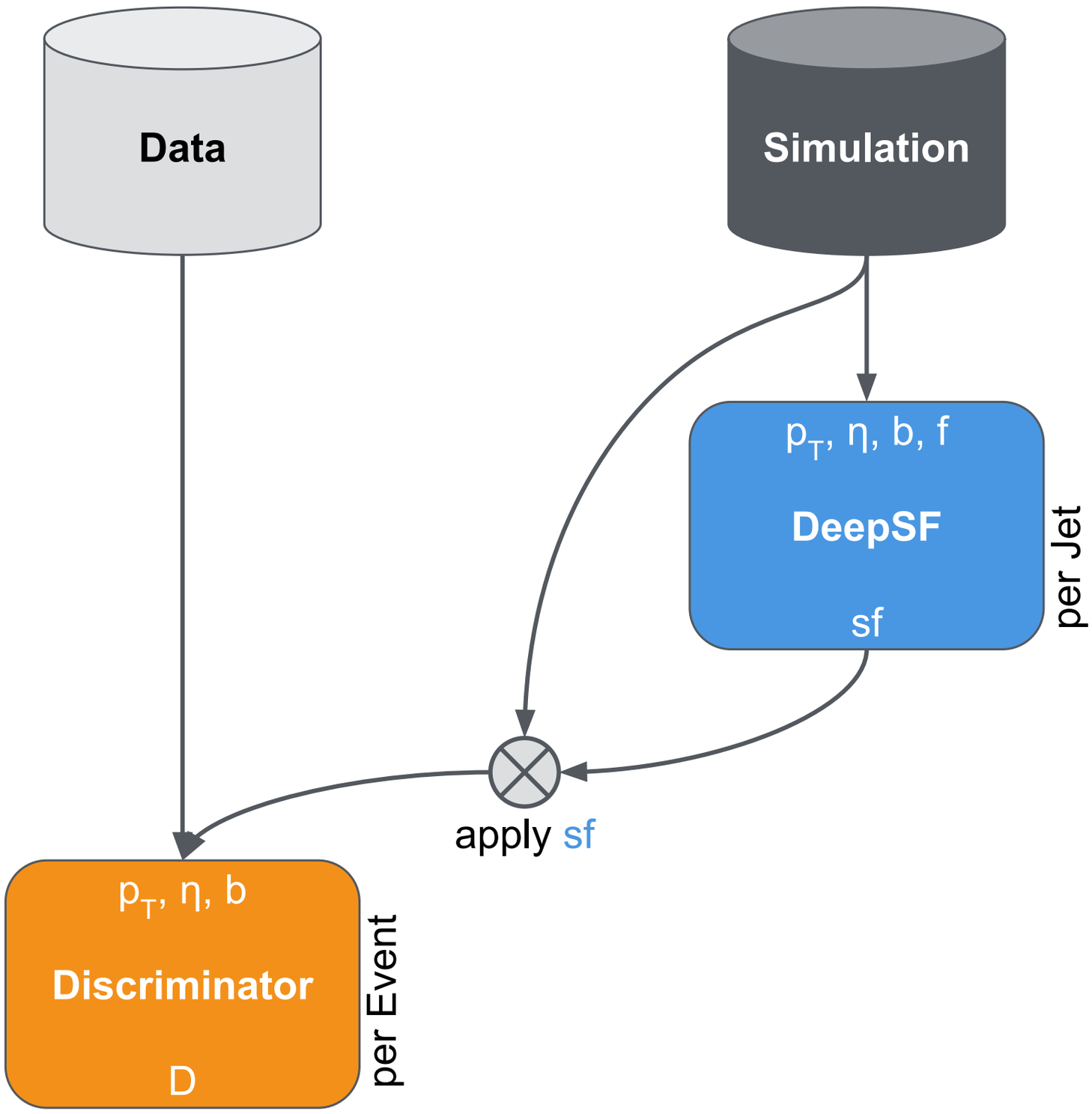
Dennis Noll - 05.07.22

- Total loss: $L = L_D$
- Discriminator loss: $L_D = \mathrm{XEnt}(D, D_{true})$
  - Binary cross entropy between Data and weighted Sim.
  - Models event scale factor: $w = D / (1 - D)$



Deep Learning Approach Architecture

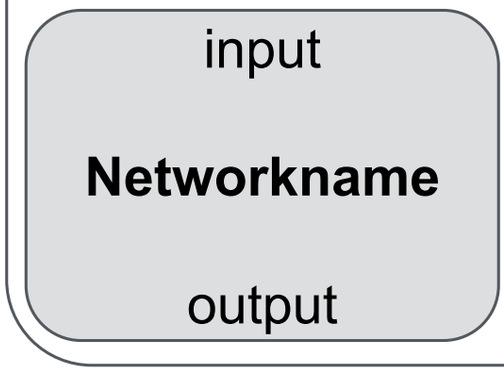
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



Variables:

- p_T
- η
- b: BTag discr.
- f: gen. flavor

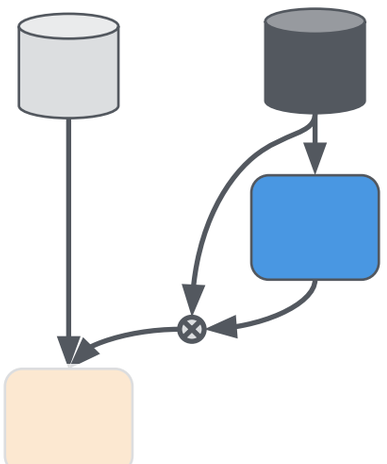
Legend:



Deep Learning Approach Loss 3

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

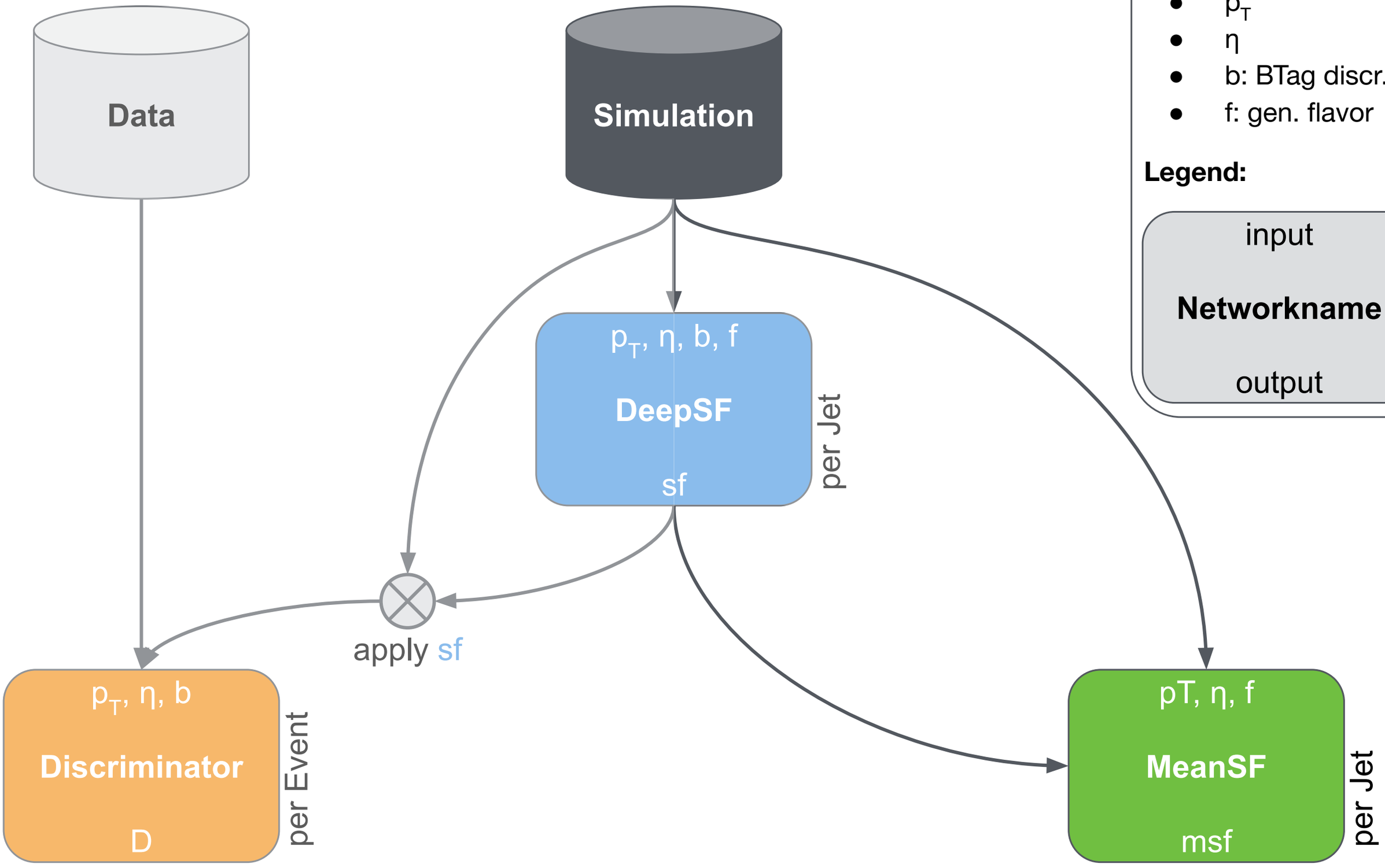
- Total loss: $L = L_D + L_{SF}$
- Discriminator loss: $L_D = \mathrm{XEnt}(D, D_{true})$
  - Binary cross entropy between Data and weighted Sim.
  - Models event scale factor: $w = D / (1 - D)$
- DeepSF Loss: $L_{SF} = \mathrm{MSE}(w^0, \omega/\omega^0)$, $\omega = \prod_{jets} sf$
  - w: remaining Data-Sim. difference
  - ω/ω^0: 1 but with gradient for DeepSF Network



Gradient stop: $x^0 = x$ with $d(x^0)/dx = 0$

Deep Learning Approach Architecture

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



Variables:

- p_T
- η
- b: BTag discr.
- f: gen. flavor

Legend:

input

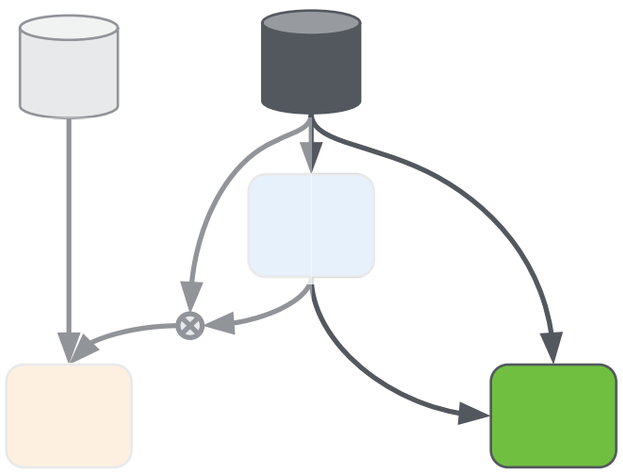
Networkname

output

Deep Learning Approach Loss 3

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

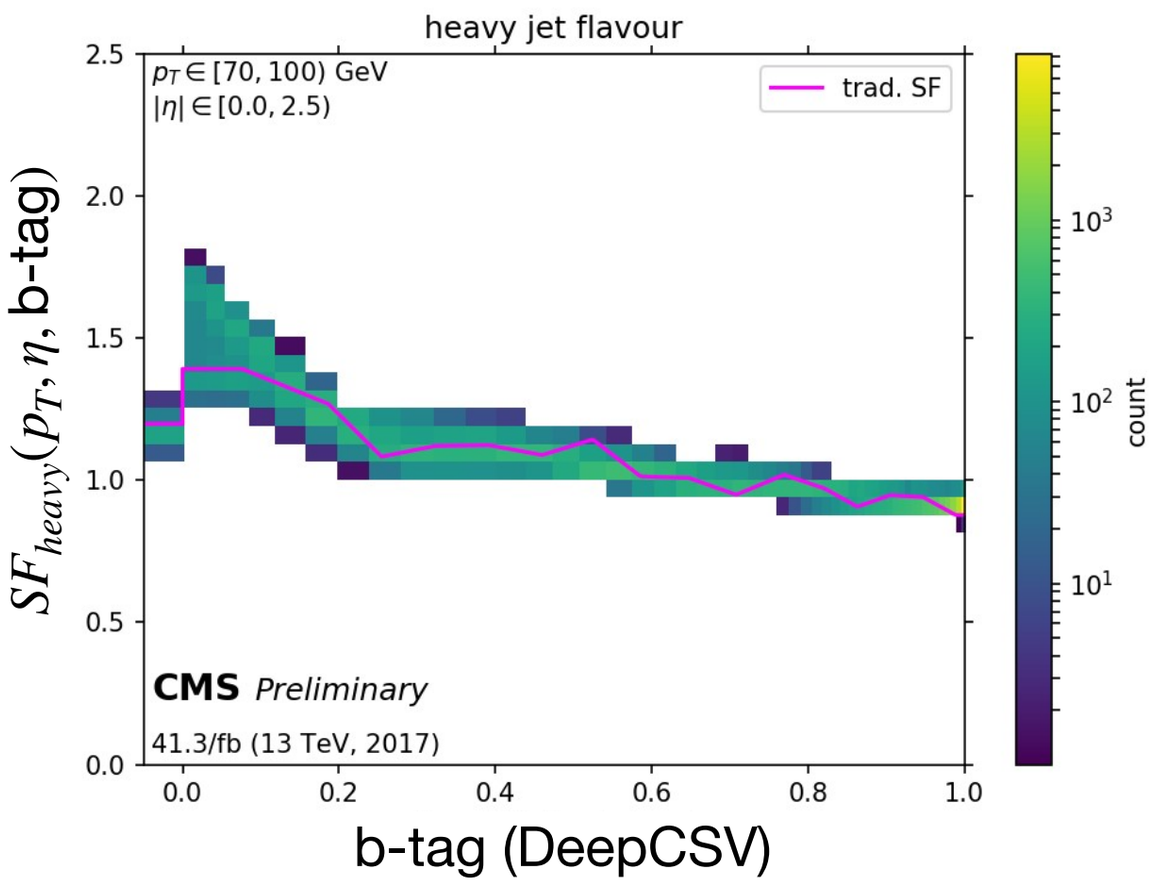
- Total loss: $L = L_D + L_{SF} + L_{MSF} + L_{prod}$
- Discriminator loss: $L_D = \mathrm{XEnt}(D, D_{true})$
  - Binary cross entropy between Data and weighted Sim.
  - Models event scale factor: $w = D / (1 - D)$
- DeepSF Loss: $L_{SF} = \mathrm{MSE}(w^0, \omega/\omega^0)$, $\omega = \prod_{jets} sf$
  - w: remaining Data-Sim. difference
  - ω/ω^0: 1 but with gradient for DeepSF Network
- MeanSF Loss: $L_{MSF} = \mathrm{MSE}(sf^0, msf)$
  - Reference point for sculpting suppression
  - Produces mean SF across b-tag discriminant for any p_T, η point
- Product Loss: $L_{prod} = \log(sf) \cdot \log(msf^0)$
  - Pulls sf such that msf gets close to 1
  - Performs sculpting suppression

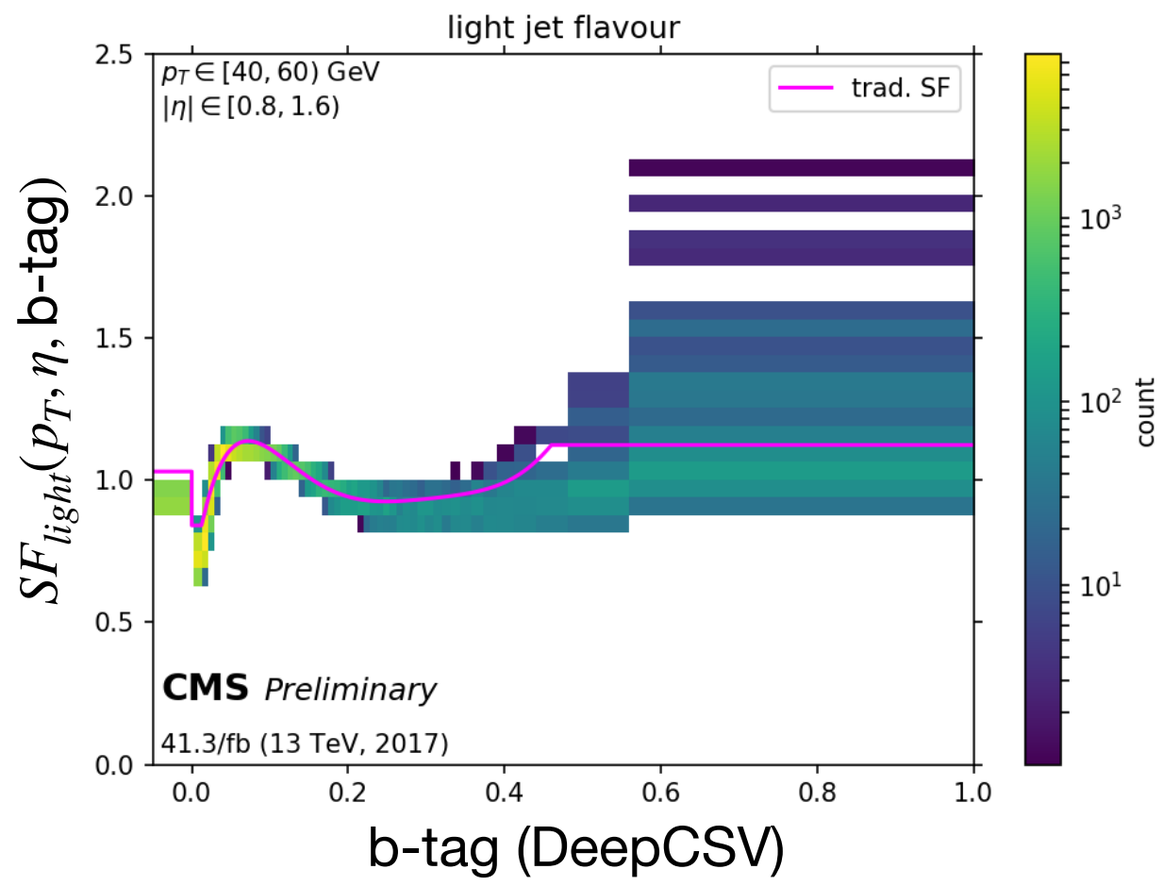


Gradient stop: $x^0 = x$ with $d(x^0)/dx = 0$

Results - Histograms

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



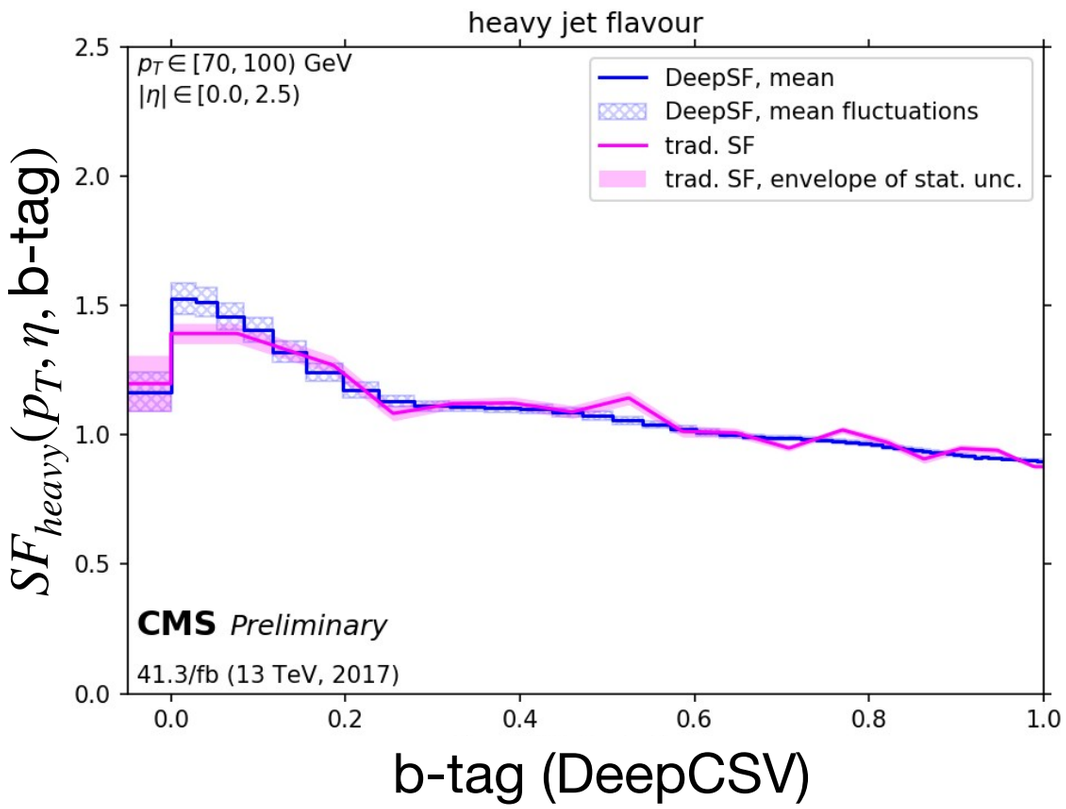


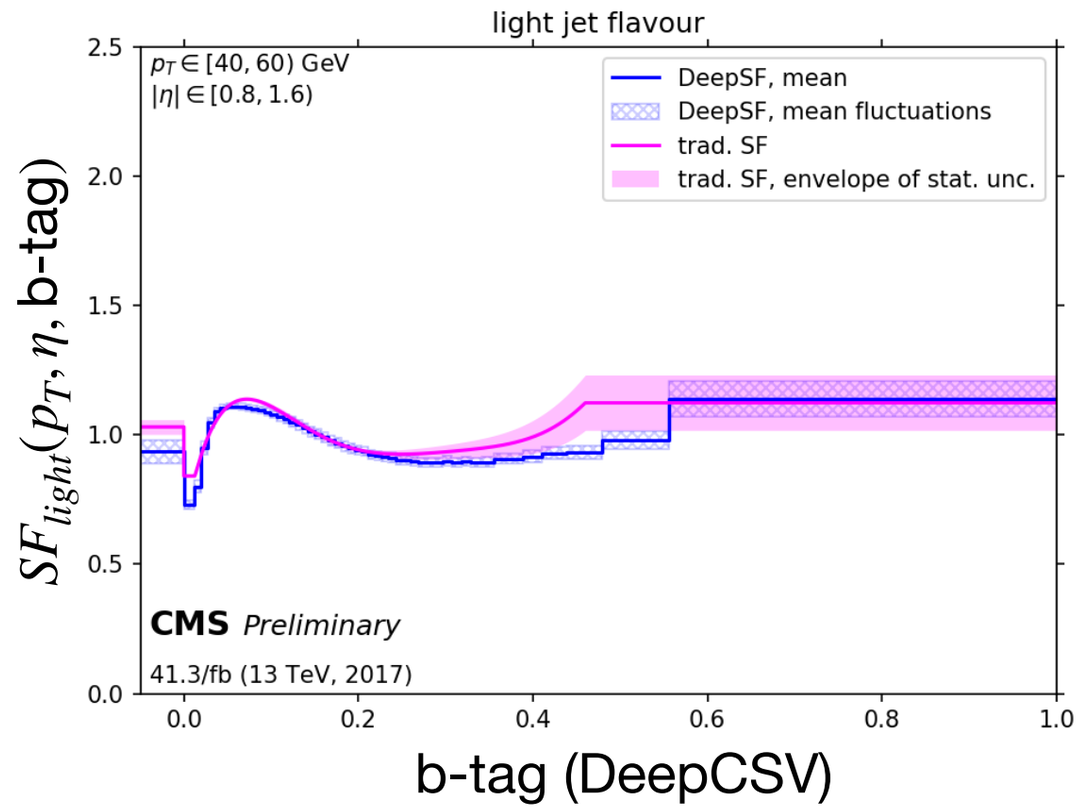
- Scale factors for probe jet in (pT, η)-bin
- Observations:
  - Traditional SF fixed value
  - DeepSF one value per jet:
    - distribution over pT and η

[6]

Results - Means

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22



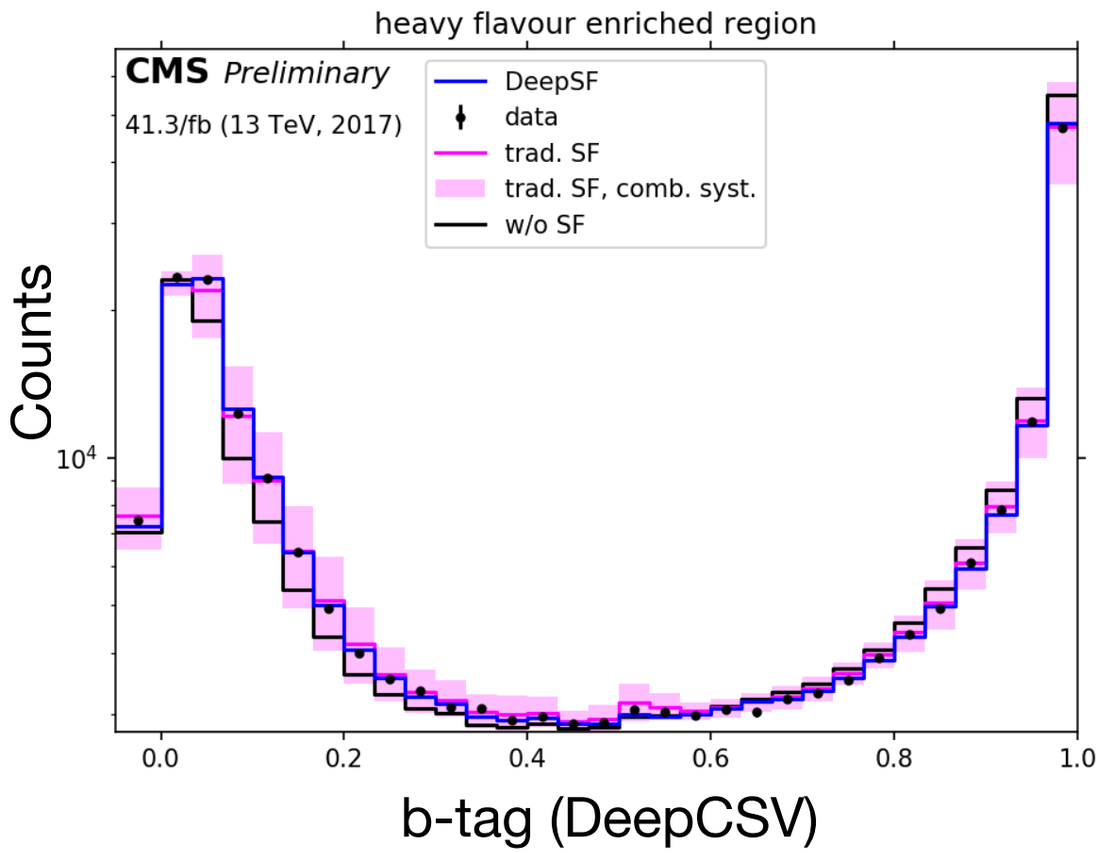


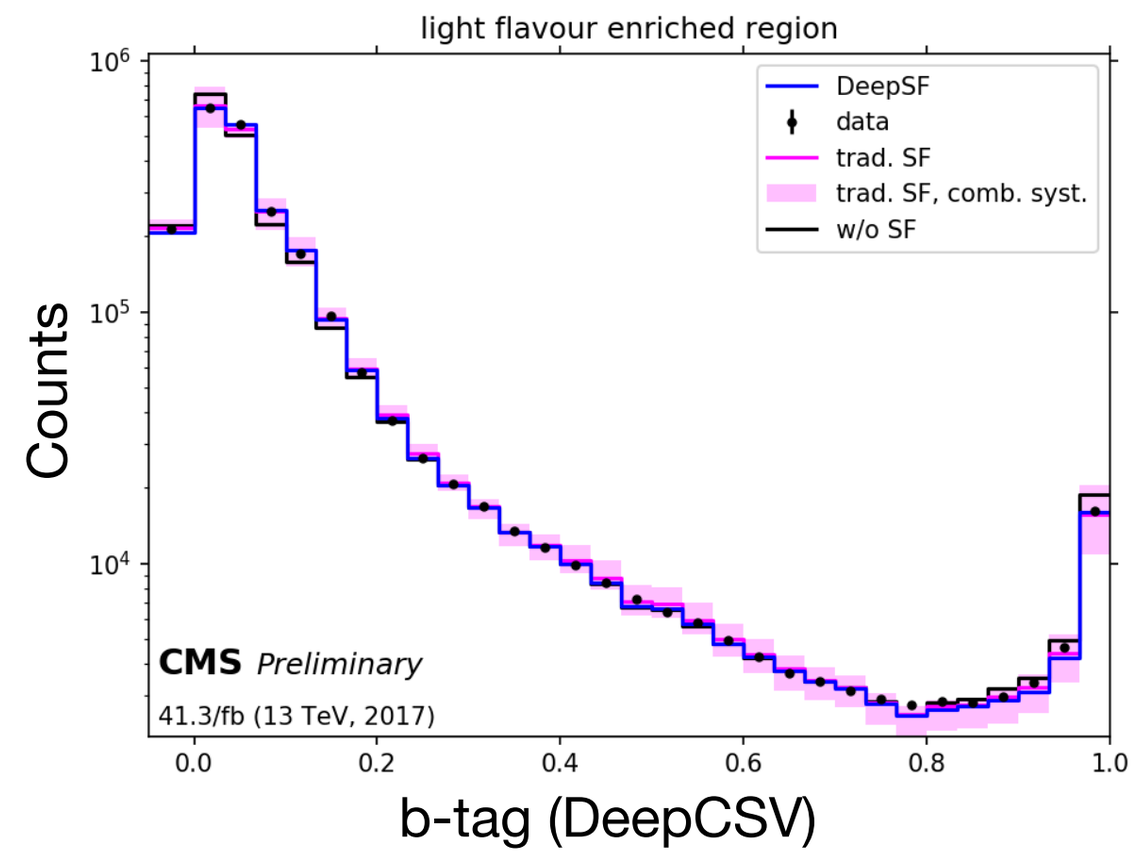
- Fluctuations
  - Traditional SF limited statistics in bins
  - DeepSF ensemble fluctuations due to limited training statistics

[6]

Results - Application

Dennis Noll - 05.07.22





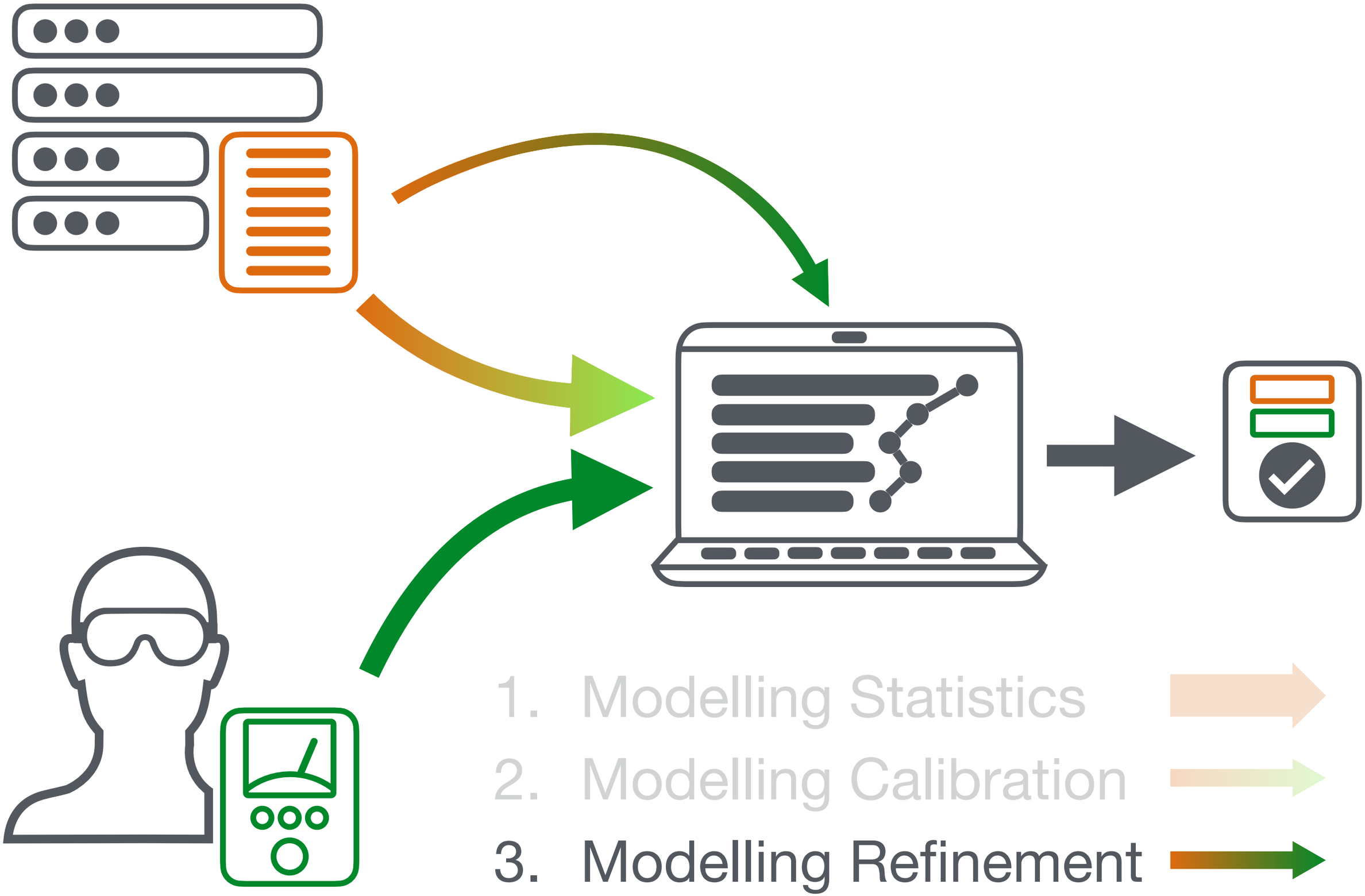
- Test if the method works:
  - Apply $SF_f(p_T, \eta, \text{b-tag})$ to probe jets in all regions f
  - Does not use sculpting suppression
- Shows overall good agreement!

[6]

Comparison of Traditional and DeepSF Approach

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

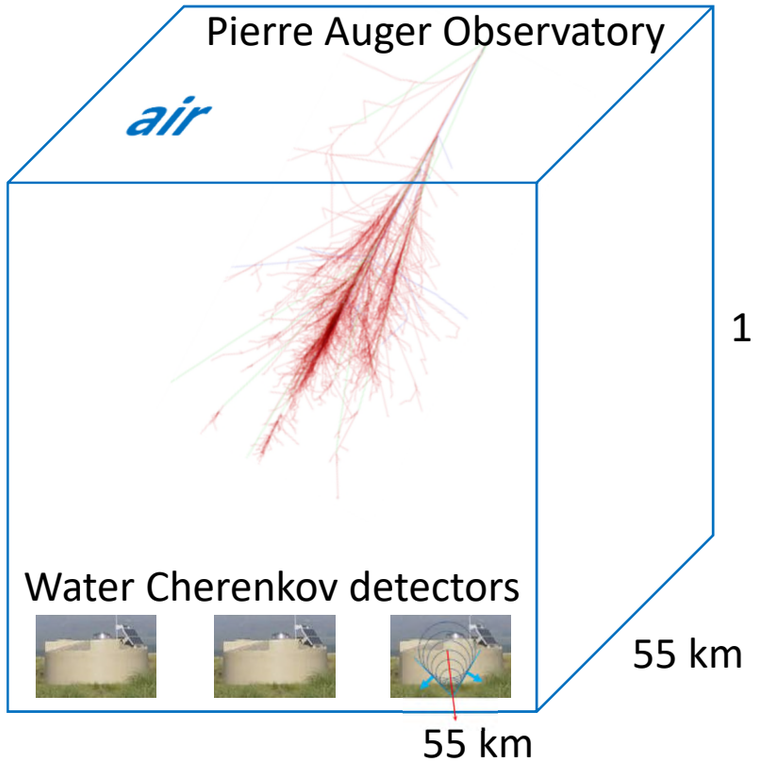
Features	Traditional	DeepSF
Calibration	✔	✔
Continuous	✘ (binned in pT, η)	✔
Un-biased	✘ (polynomial)	✔
Many Inputs	✘ (pT, η)	✔
Many Regions	✘ (only pure LF/HF)	✔
Many Flavours	✘ (only LF & HF)	✔
Active Sculpting Suppresion	✘ (only passive)	✔

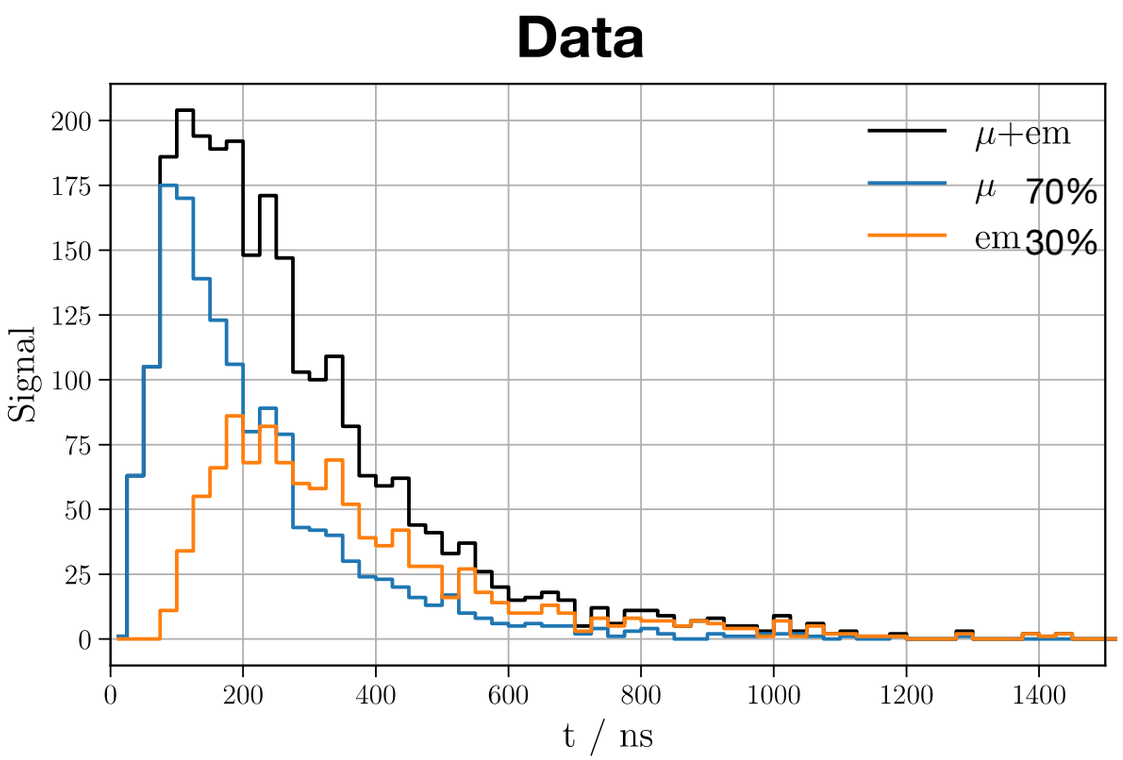

1. Modelling Statistics
2. Modelling Calibration
3. Modelling Refinement

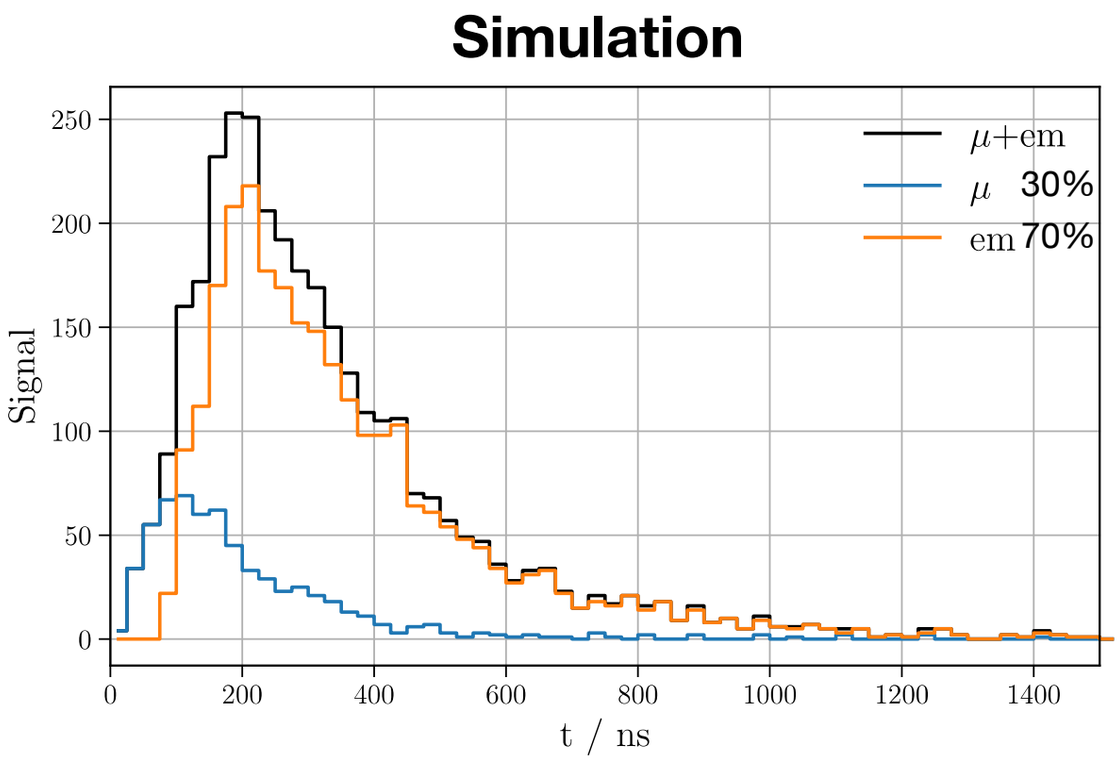
Auger Signal Traces

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Cosmic ray observatory in Argentina
- Ground-based water Cherenkov detectors
- Basically one-layer calorimeter
- Key measurement: Time traces
  - Used to regress energy of initial particle
  - Built of two components:
    - Muon (fast), Electron (slow)
    - Mixture unclear (Data/Sim. arbitrary)





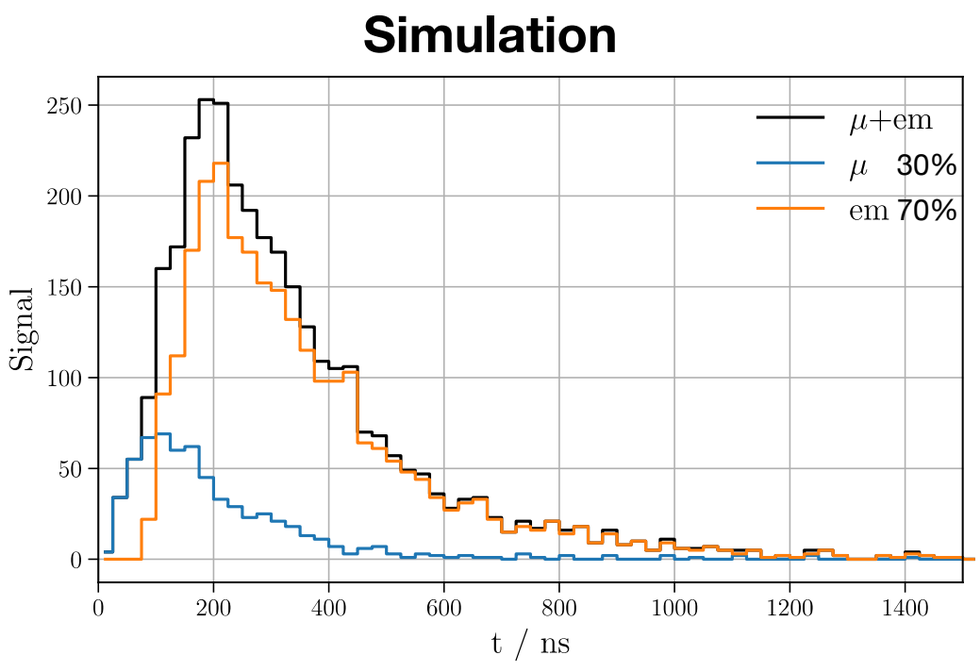


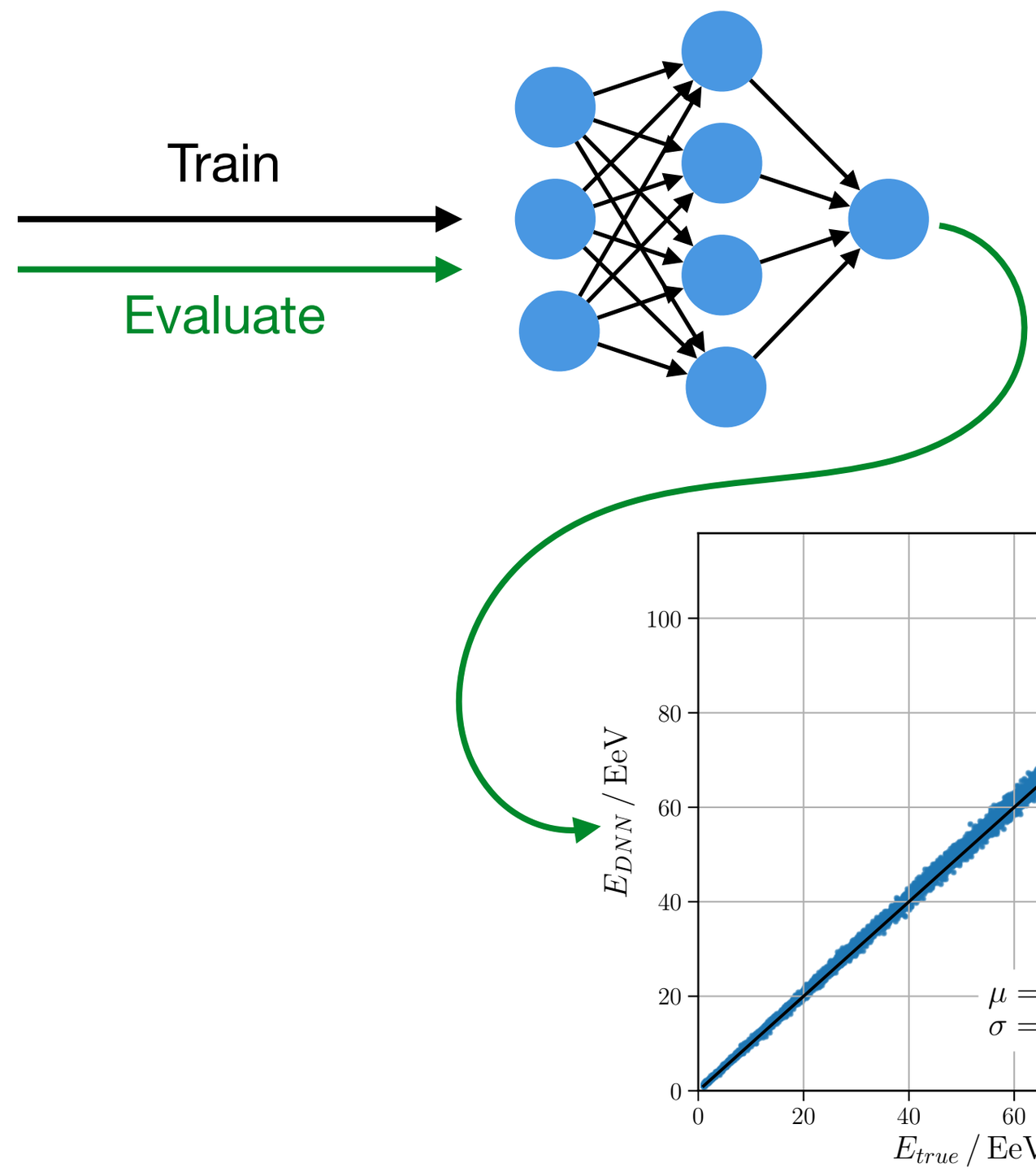
[7]

Auger Signal Traces - Challenge

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Regression of shower energy from signal shapes



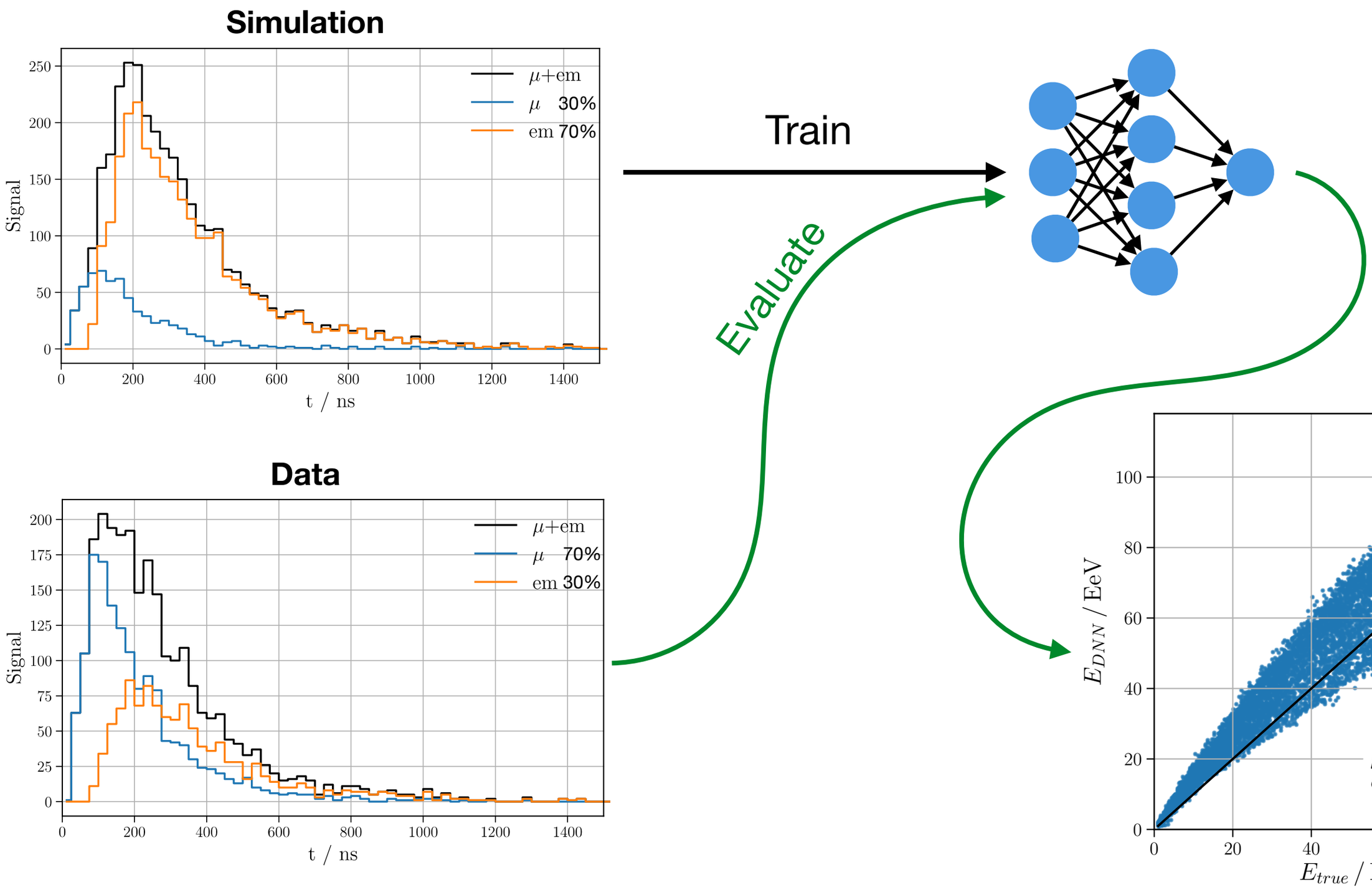


[7]

Auger Signal Traces - Challenge

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Regression of shower energy from signal shapes
- If mixture unseen before, quality of regression strongly degrades

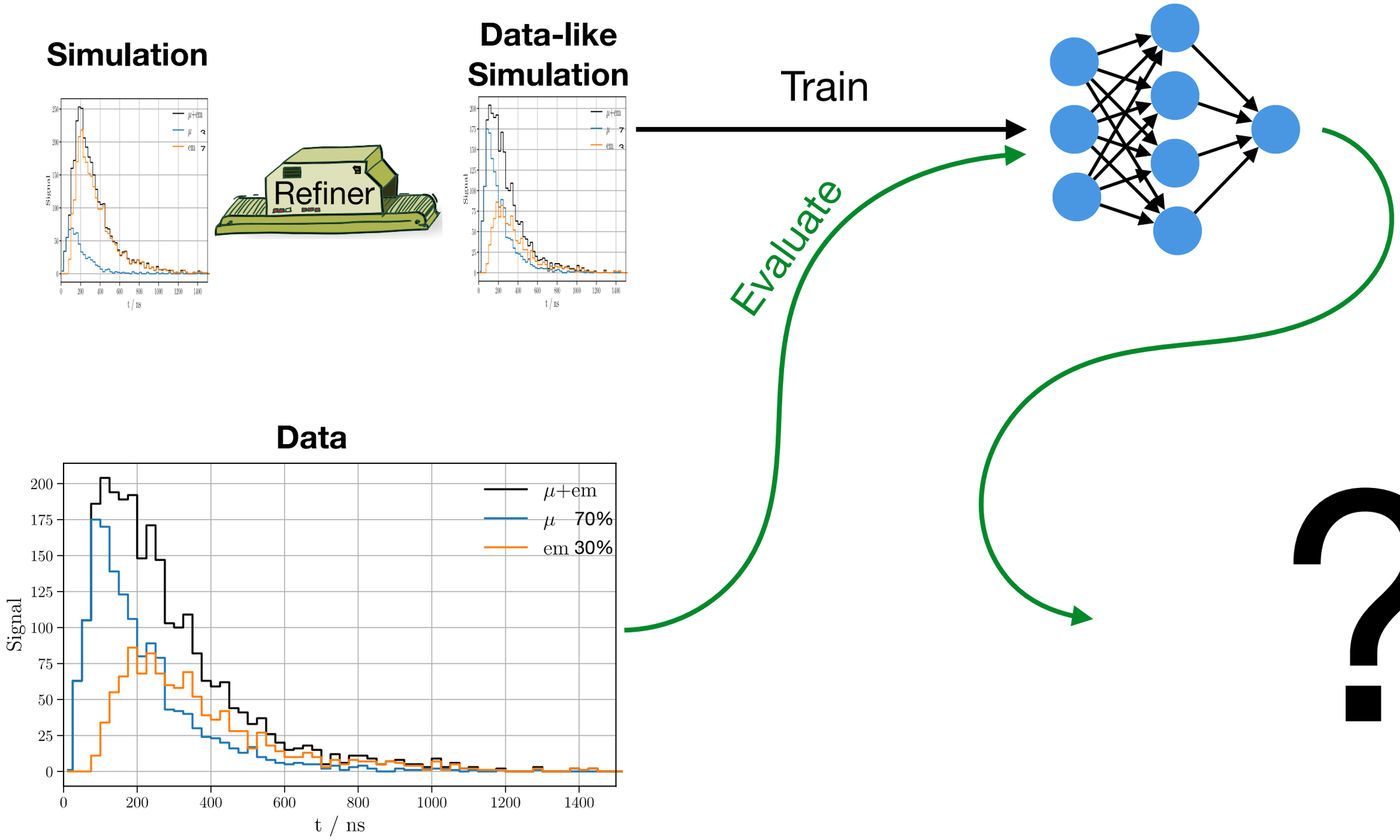


[7]

Auger Signal Traces - Idea

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Regression of shower energy from signal shapes
- If mixture unseen before, quality of regression strongly degrades
- Idea: Use Refiner DNN to make training distribution more evaluation-like

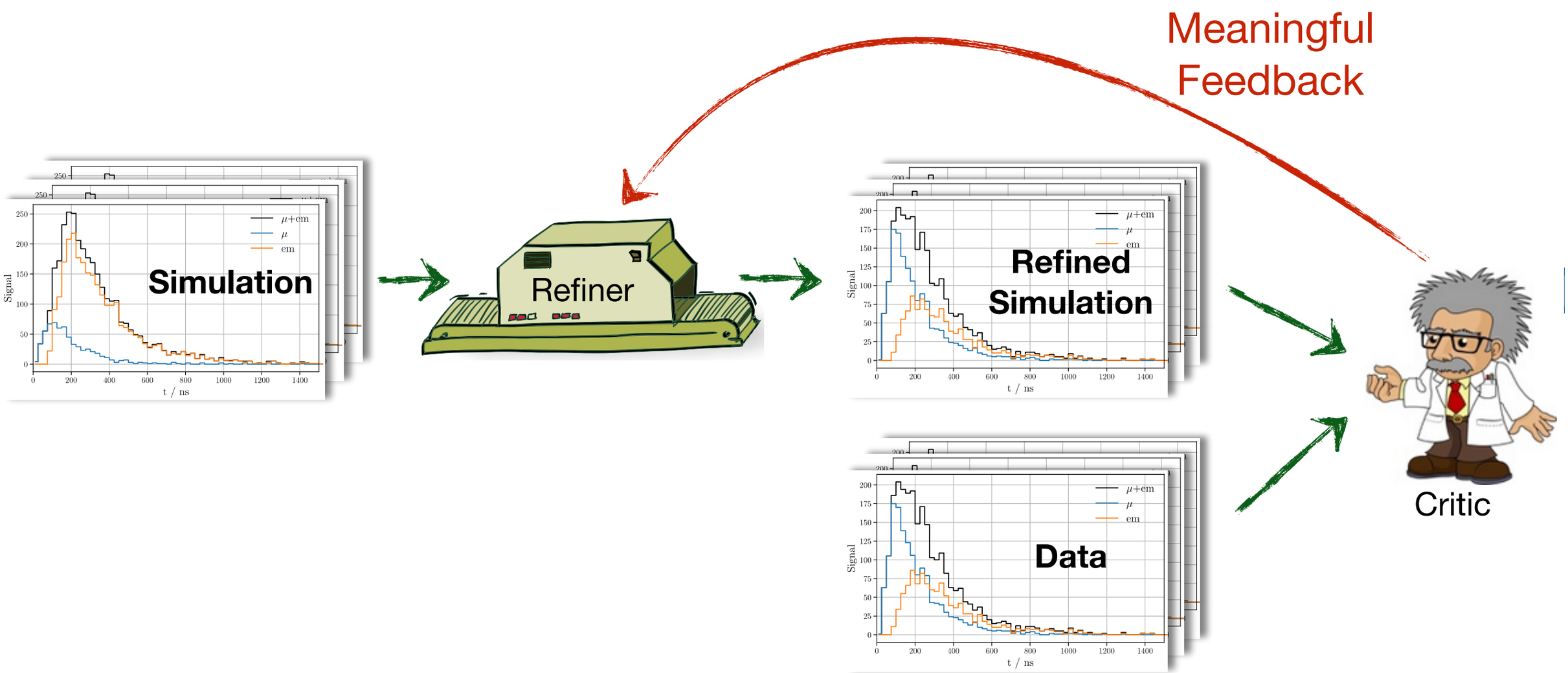


[7]

Simulation Refinement: Architecture

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Mitigate data-simulation mismatches
- Adversarial approach:
  - Refiner network refines simulations
  - Critic network provides feedback about quality (next slides)



Compare PCD lecture Generative Networks (22.06.21) by Anja Butter [7]

Meaningful Feedback from Critic

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Wasserstein distance:
  - Quantifies difference of the two distributions P_d (data) and P_r (refined)
  - Describes minimal cost (mass x distance) to move distribution P_r onto P_d

Ensures minimal cost

Distance

$$D_W(P_d||P_r) = \inf_{\gamma \in \prod(P_d, P_r)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||]$$

Transportation plans

[8]

Critic Training

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

Kantorovich-Rubinstein Duality (rewrite Wasserstein distance)

Critic

$$D_W(P_d||P_r) = \sup_{f_w \in Lip_1} \mathbb{E}_{x_d \sim P_d}[f_w(x_d)] - \mathbb{E}_{x_r \sim P_r}[f_w(x_r)]$$

Data

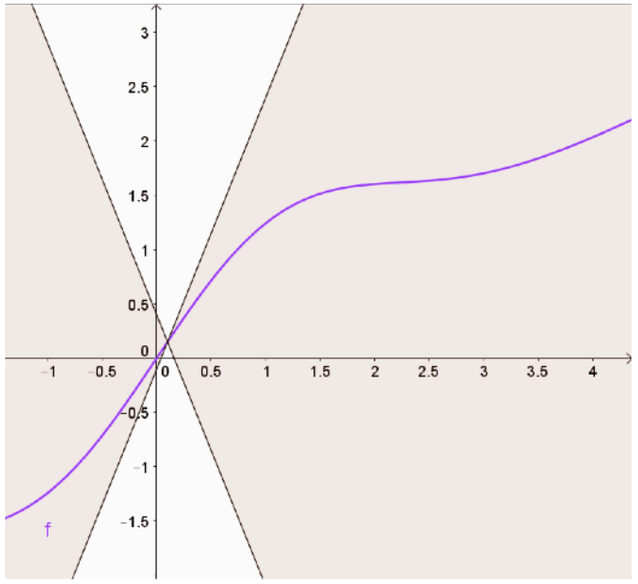
Refined simulation
x_r = Refiner(x_{sim})

Lipschitz Constrain

- Lipschitz-1 (Lip_1) functions have slope ≤ 1
- Constrain through gradient penalty (penalize ≠ 1):

$$\mathscr{L}_{GP} = \lambda \mathbb{E}_{u \sim P_u}[(||\nabla_u f_w(u)||_2 - 1)^2]$$

$$u = \epsilon x_d + (1 - \epsilon) x_r \qquad 0 \leq \epsilon \leq 1$$



[8]

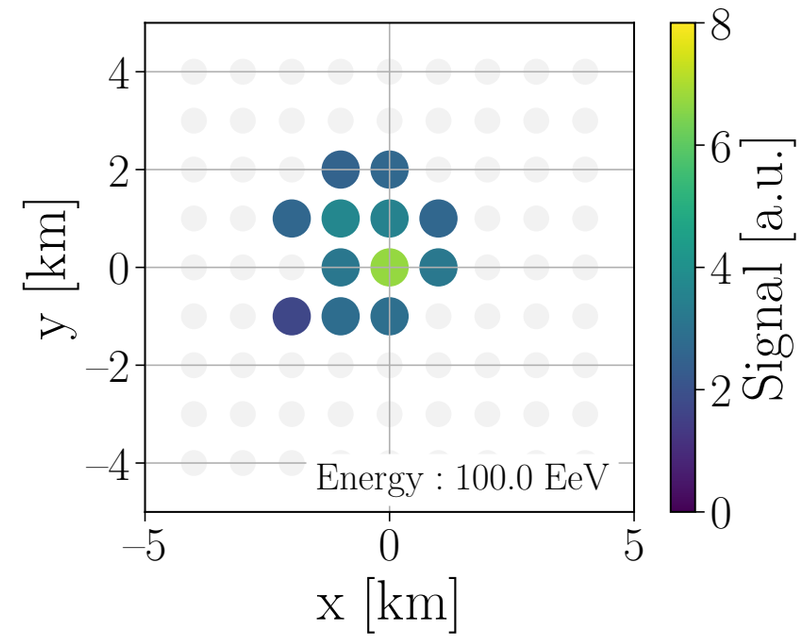
Refiner Architecture and Training

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

[7]

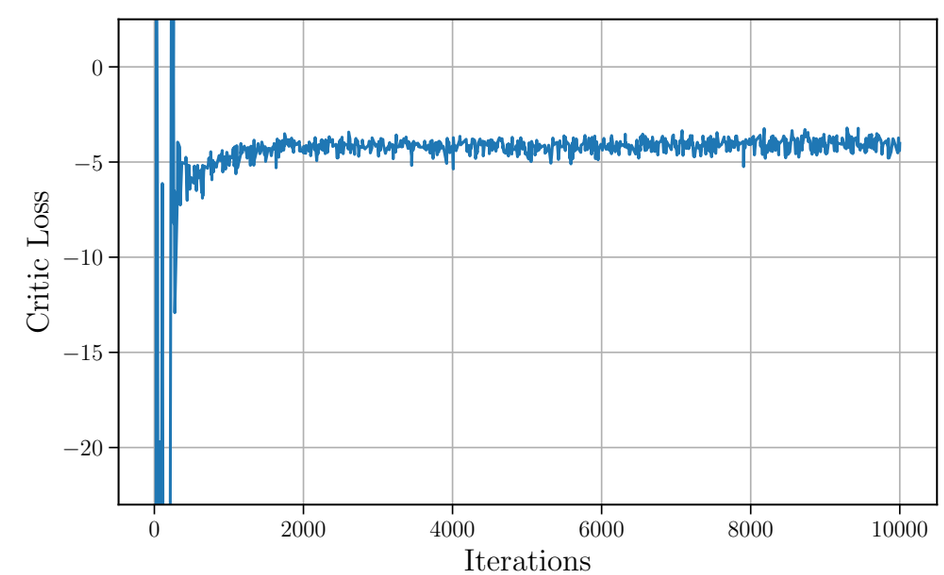
Architecture

- Networks acting on multiple stations
  - 3D: 2D (Detectors) + 1D (Time)
- Refiner:
  - Residual Network
  - 4 layers of 3D Convolution ReLU
- Critic:
  - AixNet (1708.00647)
  - 4 layers of 3D Convolution LeakyReLU



Training

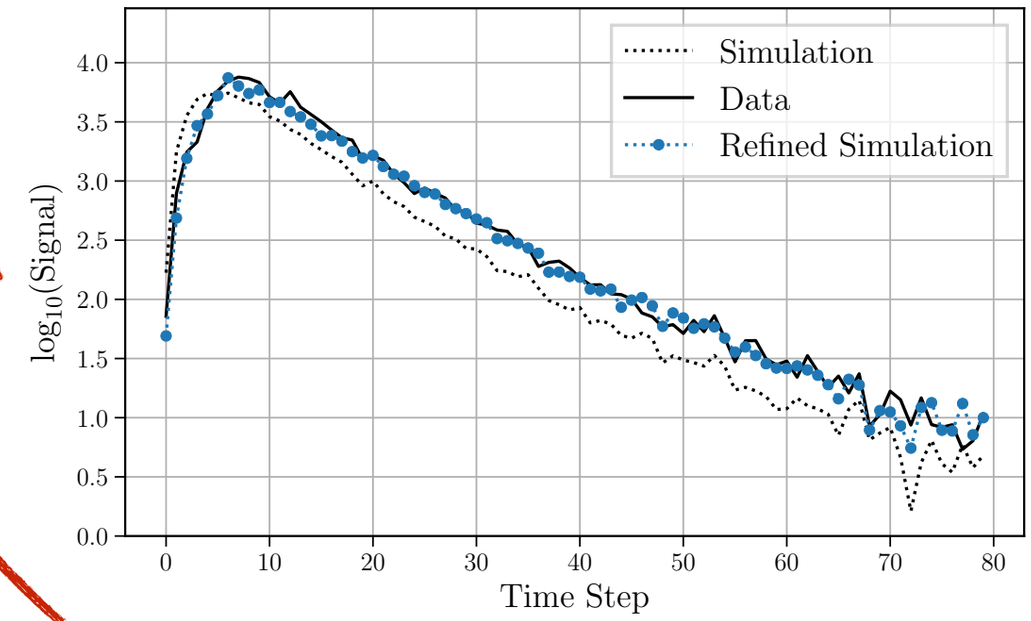
- Total: 10000 Refiner iterations x 100 events
- Alternating schedule 1 x Refiner = 10 x Critic
- Optimizer: Adam(lr=10^{-4}, β_1=0.5, β_2=0.999)
- Gradient penalty: λ=5



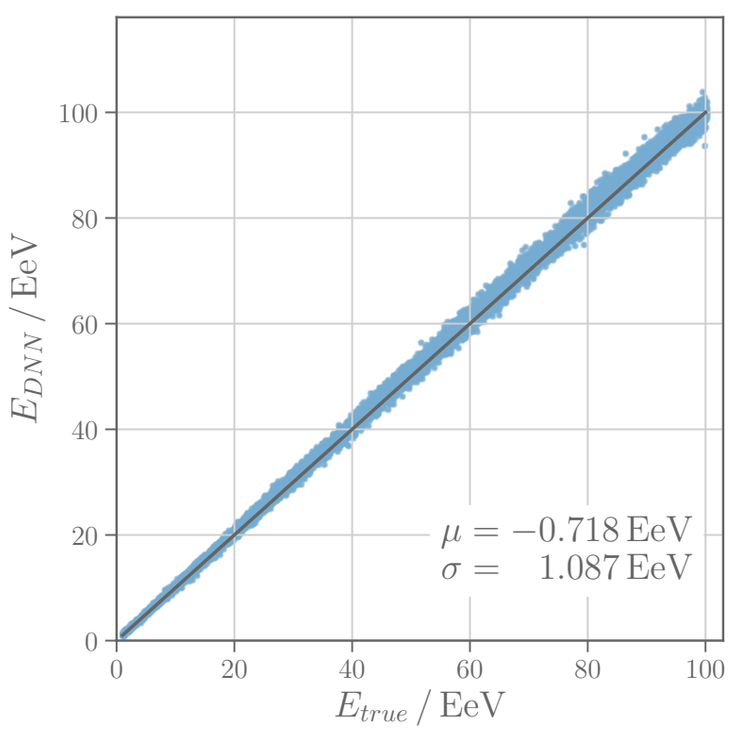
Refiner Results

RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

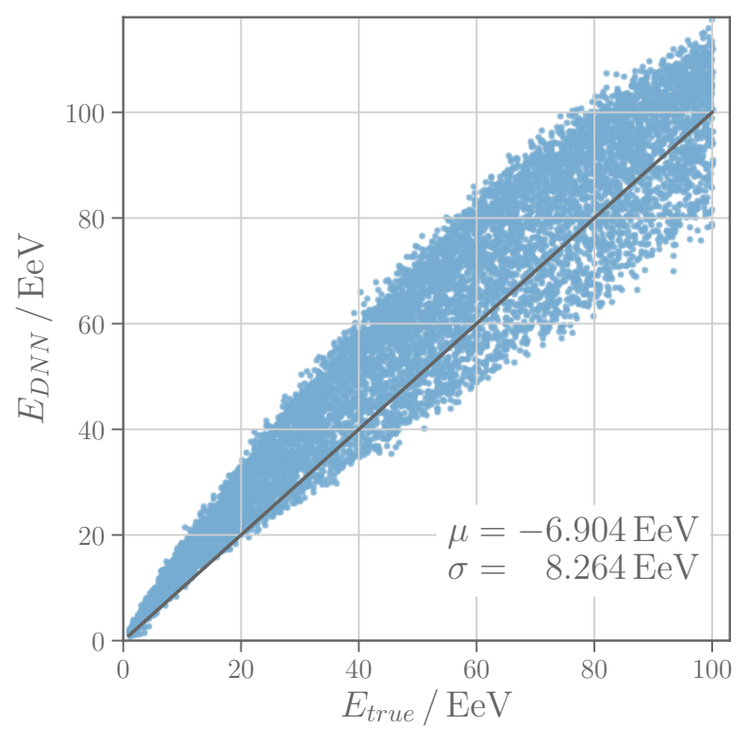
- Refinement of signal time trace works
- Refinement improves regression:
  - No bias
  - Better resolution



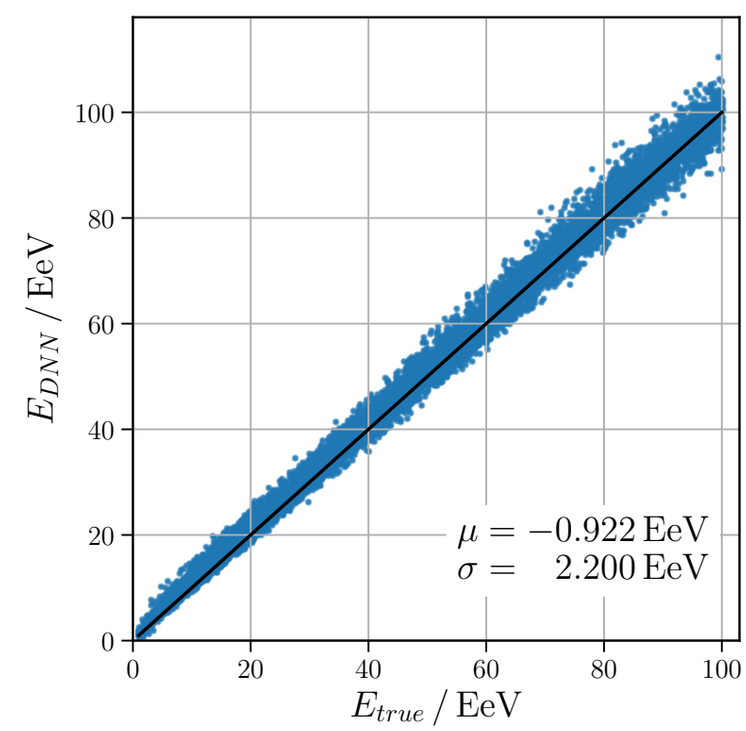
Train on Sim. Eval on Sim.



Train on Sim. Eval on Data



Train on Refined Sim. Eval on Data

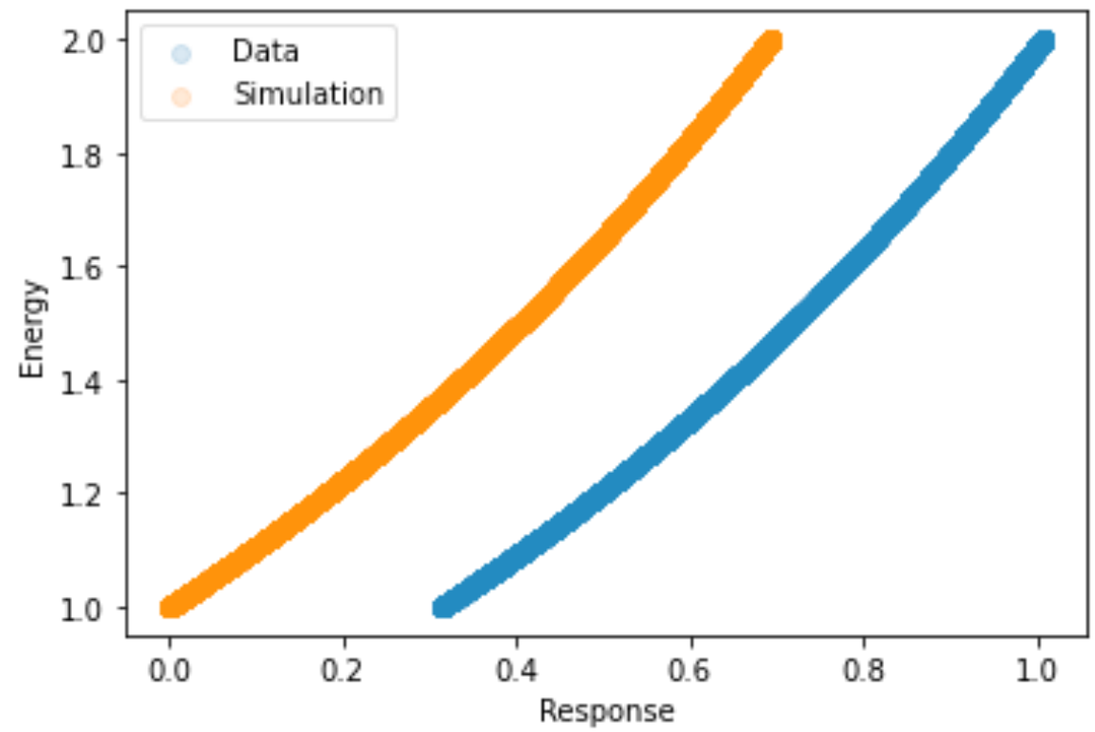


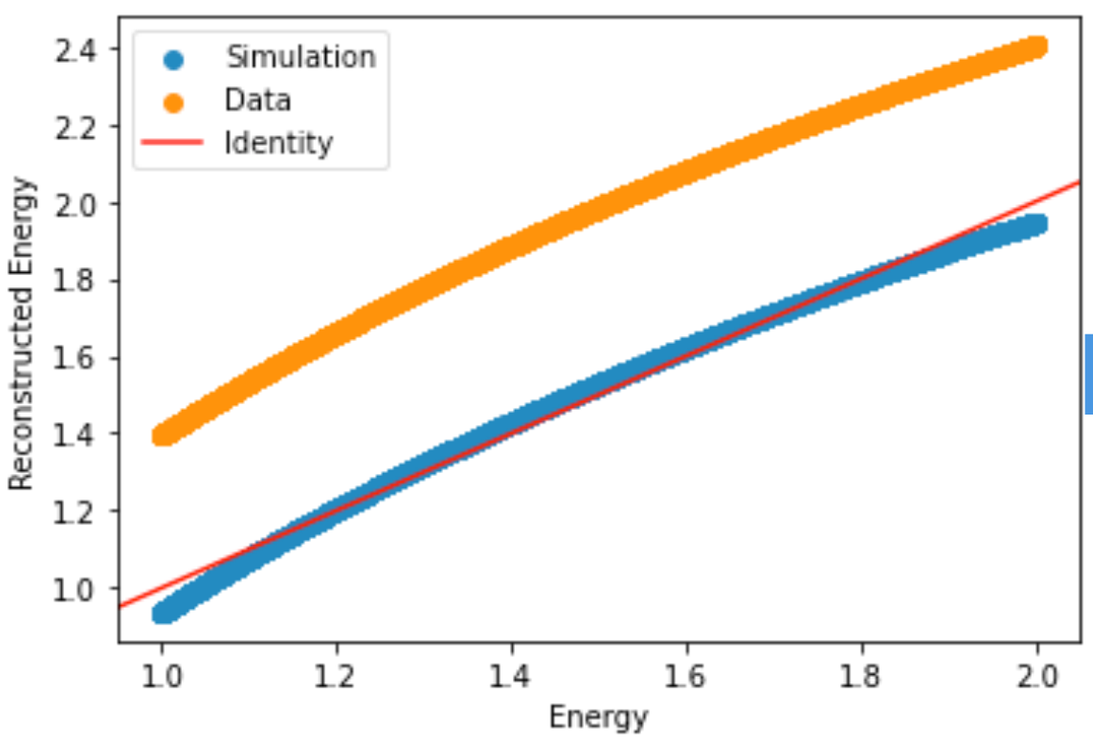
[7]

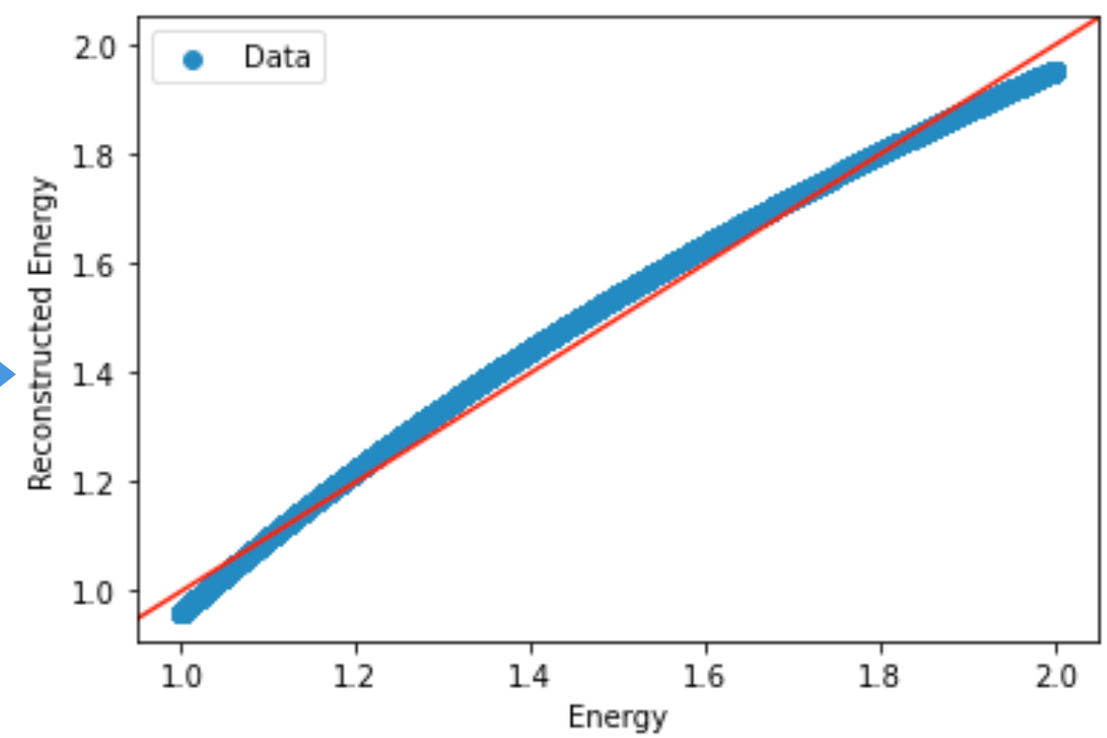
Exercises

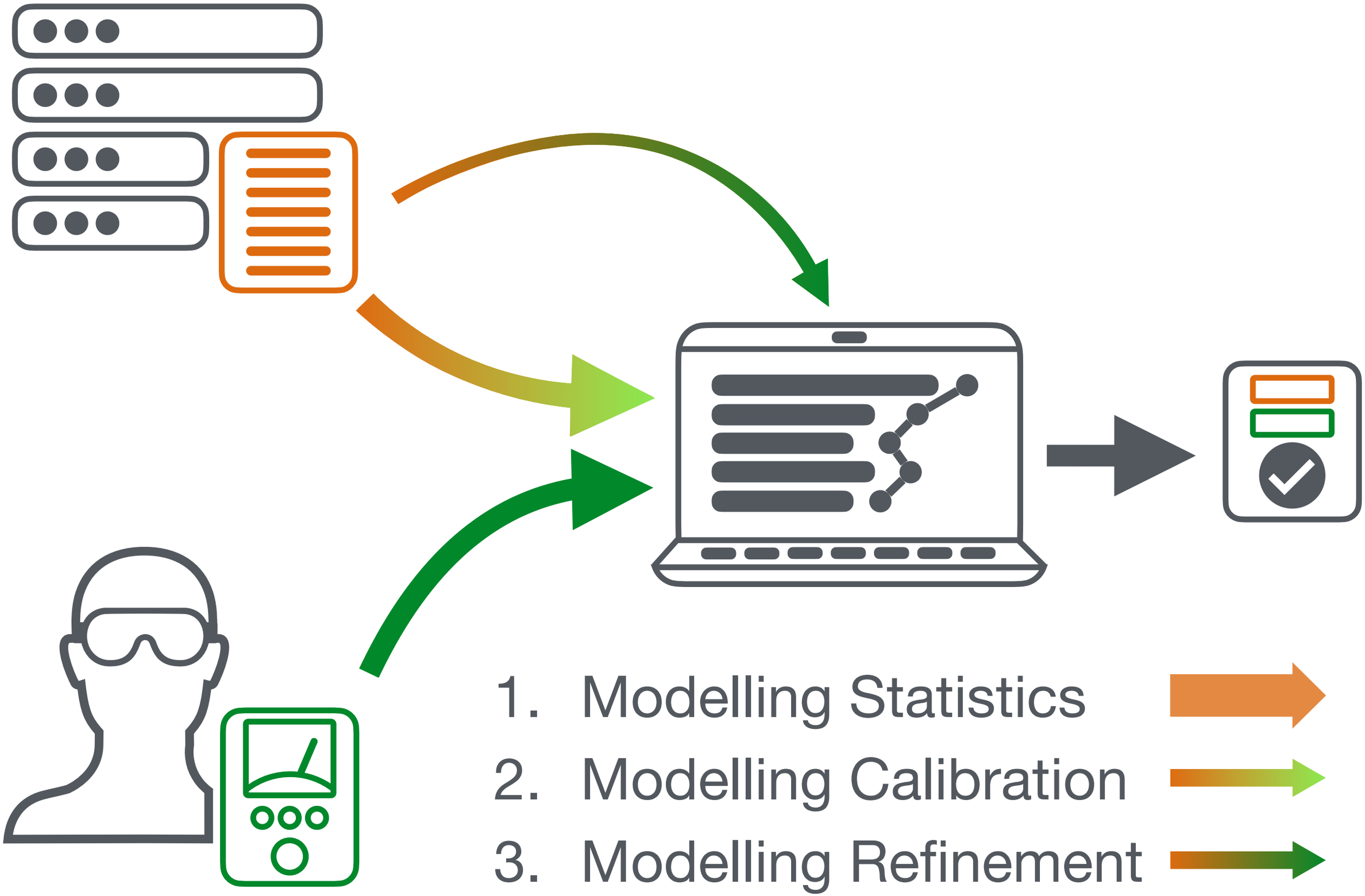
RWTH AACHEN UNIVERSITY
Dennis Noll - 05.07.22

- Link to all exercises
- Exercise 2 - Refiner:
  - Data and Simulation with different responses
  - Train refiner to make simulation more data-like








1. Modelling Statistics
2. Modelling Calibration
3. Modelling Refinement

Citations and Acknowledgements

- Citations:
  1. **Shannon, Claude**, Author: Jacobs, Konrad, Source: Konrad Jacobs, Erlangen, Copyright: MFO
  2. **A Short Introduction to Entropy, Cross-Entropy and KL-Divergence**, Aurélien Géron, February 2018, link
  3. **Search for non-resonant pair production of Higgs bosons in the bbbb final state in pp collisions at √s=13 TeV with the ATLAS detector**, The ATLAS collaboration, ATLAS-CONF-2022-035
  4. **Non-resonant 4b in ATLAS and CMS**, Daniel Guerrero, Rafael Teixeira de Lima, Higgs Pairs Workshop Dubrovnik 2022, link
  5. **Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV**, The CMS Collaboration, Published in: JINST 13 (2018) 05, P05011, 1712.07158
  6. **Adversarial Neural Network-based data-simulation corrections for jet-tagging at CMS**, Martin Erdmann et al 2020 J. Phys.: Conf. Ser. 1525 012094, CMS-DP-2019-003
  7. **Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks**, Martin Erdmann et al., Comput.Softw.Big Sci. 2 (2018) 1, 4, 1802.03325
  8. **Deep Learning in Physics Research Lecture**, Martin Erdmann, Jonas Glombitza, Uwe Klemradt, Dennis Noll, RWTH Aachen University
- Acknowledgements:
  - Prof. Martin Erdmann for fruitful discussion and great mentoring on the choice of the topics
  - Benjamin Fischer for great help with the ABCD exercise, Figures on the DeepSF Architecture and many fruitful discussions across all topics
  - Maximilian J Swiatlowski and Max and Rafael Teixeira de Lima for valuable comments on the ATLAS ABCD Method
  - Erwin Rudi for valuable insights into the DeepSF architecture
  - Jonas Glombitza for great insights into the Auger Refinement challenge