

# Uncertainties in ML for HEP: Workshop Summary

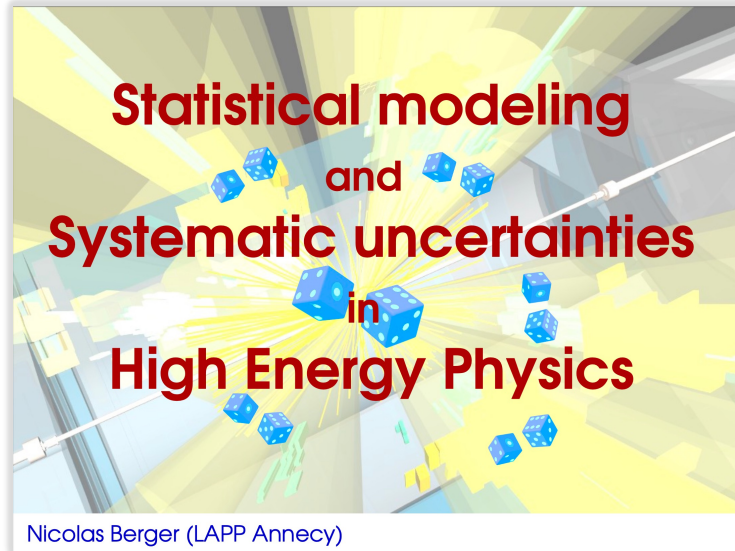
Michael Kagan  
SLAC

April 27, 2022



# Thanks to the Presenters for their Excellent Talks!

2



Quantmetry

## MAPIE

Model Agnostic Prediction Interval Estimator :  
a [scikit-learn-contrib](#) library

LEARNING TO DISCOVER  
from April 19 to  
April 22, 2022  
Institut Pasteur, Université Paris-Saclay  
Orsay

21/04/2022

Nicolas Brunel  
Laboratoire de Mathématiques et  
Modélisation d'Evry,  
ENSIEE, Univ. Paris Saclay  
[nbrunel@quantmetry.com](mailto:nbrunel@quantmetry.com)


Vianney Taquet  
Senior Data Scientist  
[vttaquet@quantmetry.com](mailto:vttaquet@quantmetry.com)

Vincent Blot  
Data Scientist  
[yblot@quantmetry.com](mailto:yblot@quantmetry.com)


Thomas Morzadec  
Data Scientist  
[tmorzadec@quantmetry.com](mailto:tmorzadec@quantmetry.com)

Q

© Quantmetry 2020 | Diffusion Interdite sans accord - [Quantmetry.com](#) - 52 rue d'Anjou, 75008 Paris - RCS Paris 531172393 - TVA : FR27531172393



DE LA RECHERCHE À L'INDUSTRIE



Uncertainty quantification in  
Deep Learning


Geoffrey DANIEL, CEA/DES/ISAS/DM2S/STMF/LGLS

Commissariat à l'énergie atomique et aux énergies alternatives - [www.cea.fr](http://www.cea.fr)

## Simulation-based inference: Proceed with caution!

Learning to Discover  
April 22, 2022

Gilles Louppe  
[g.louppe@uliege.be](mailto:g.louppe@uliege.be)



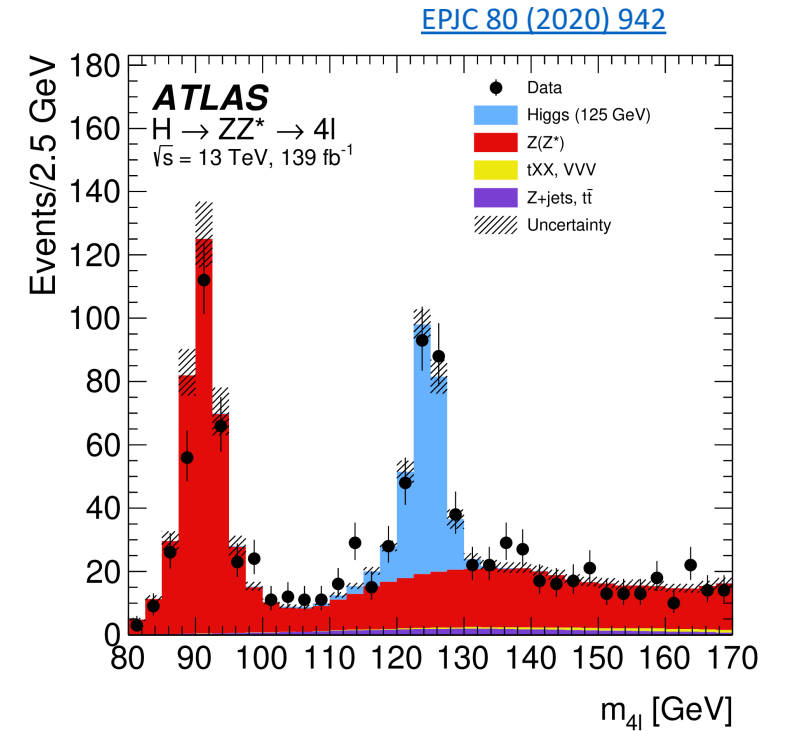
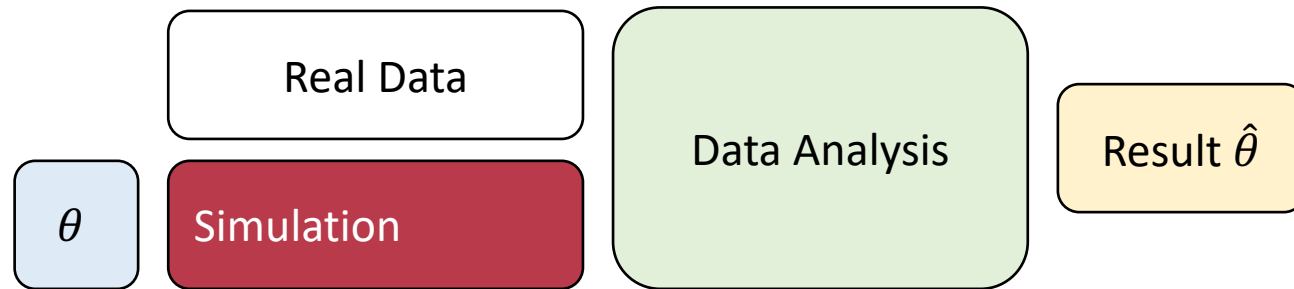
# The Setting: Data Analysis in HEP

Talk:

1) N. Berger

# Data Analysis Pipeline in HEP

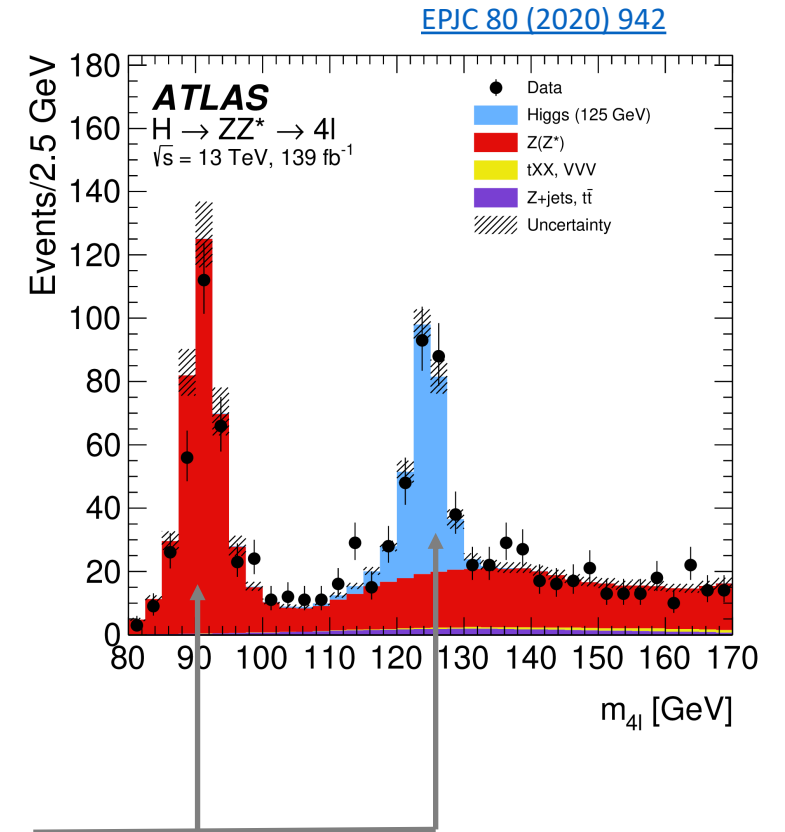
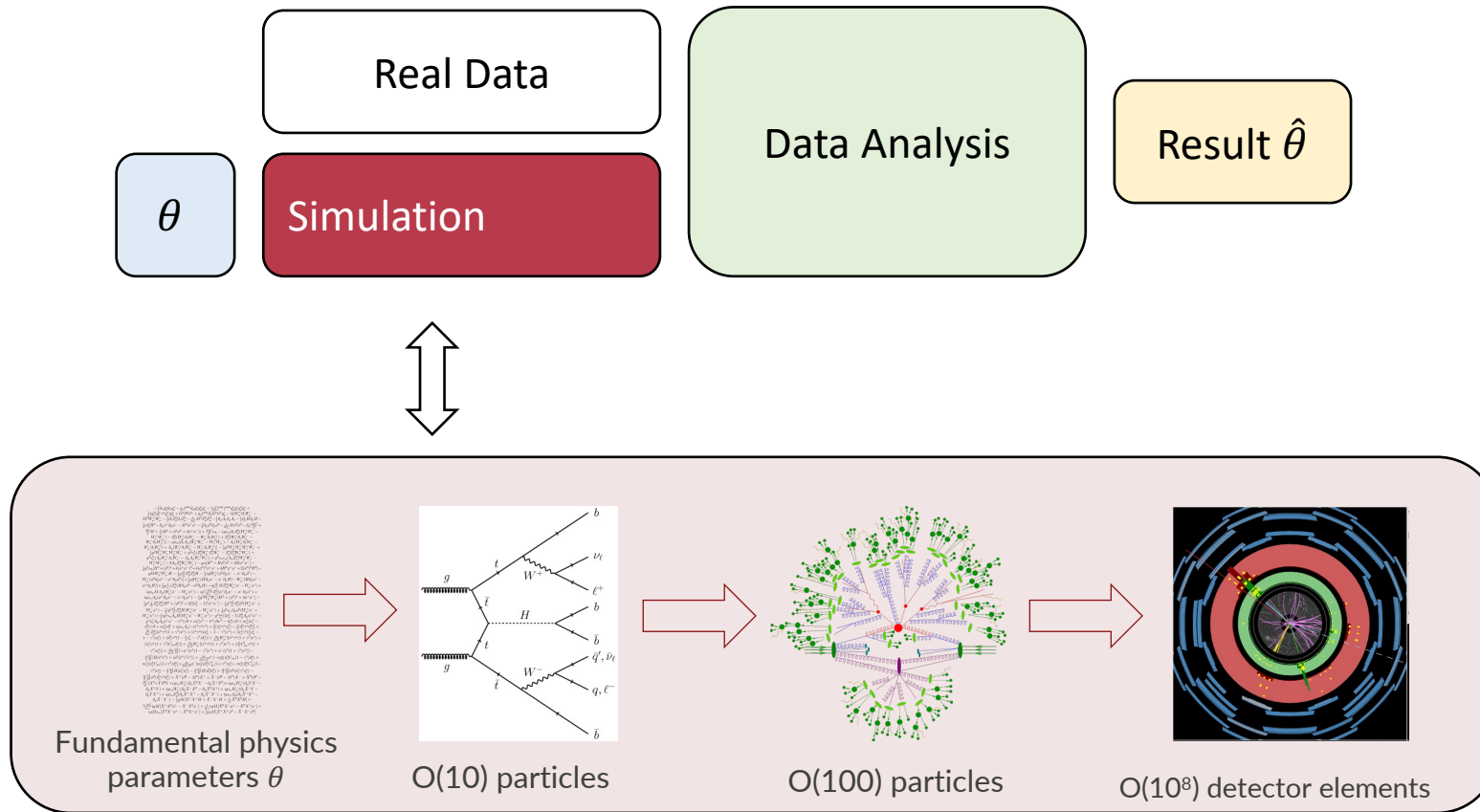
4





# Data Analysis Pipeline

5

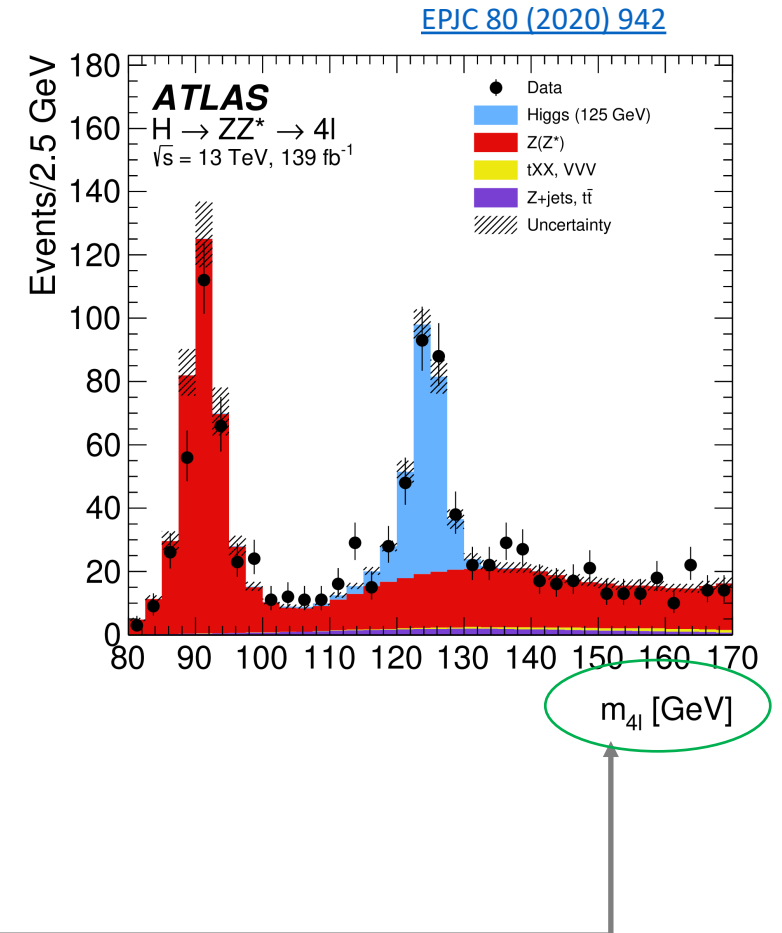
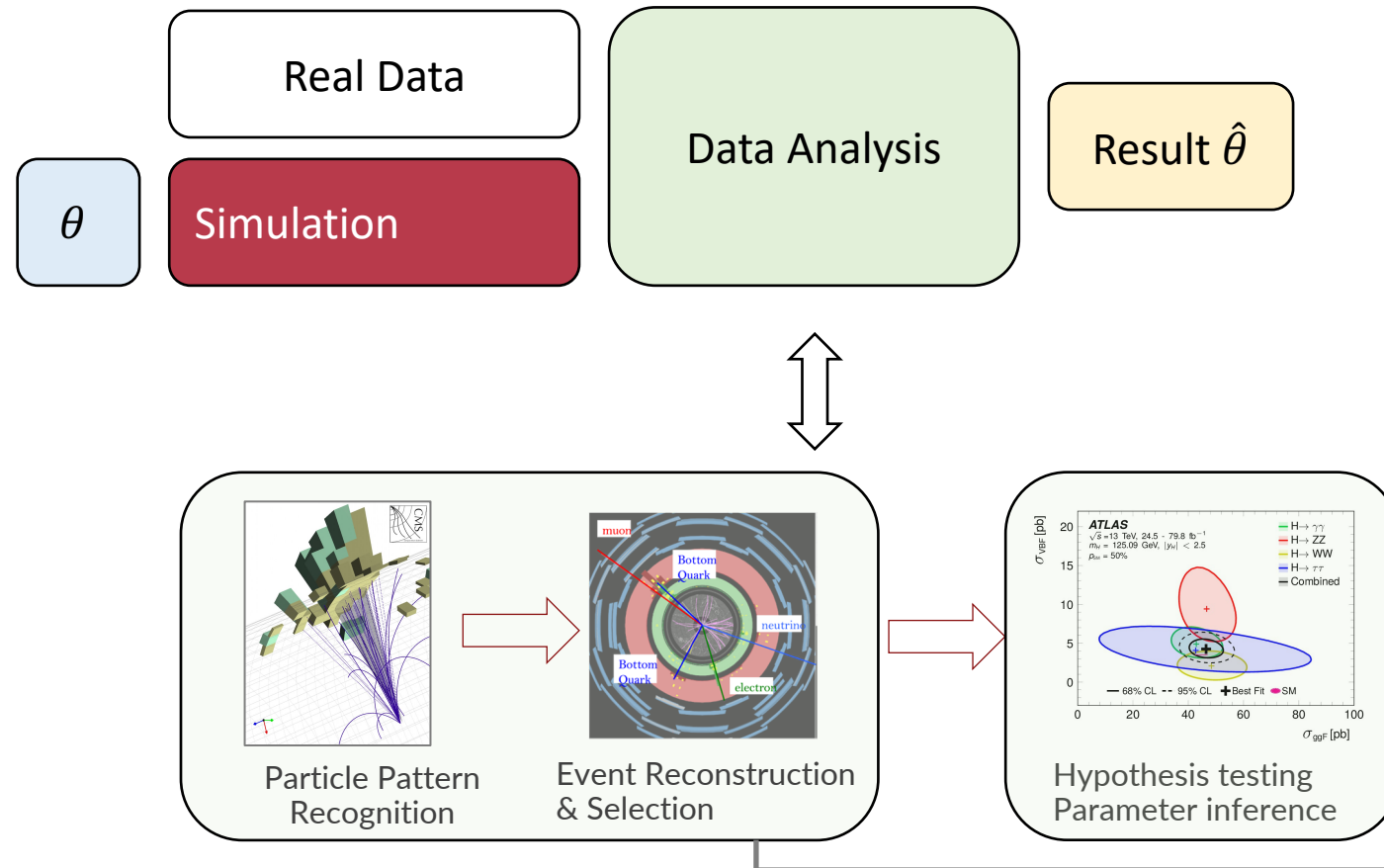


ML efforts growing to develop fast, approximate surrogates of simulators

- See the Generative Models Workshop Summary

# Data Analysis Pipeline in HEP

6

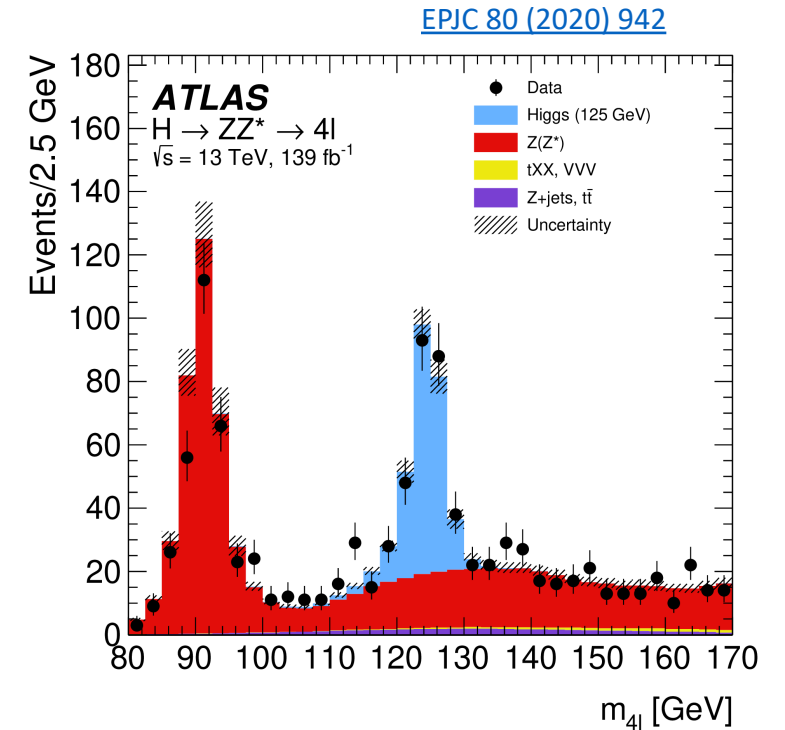
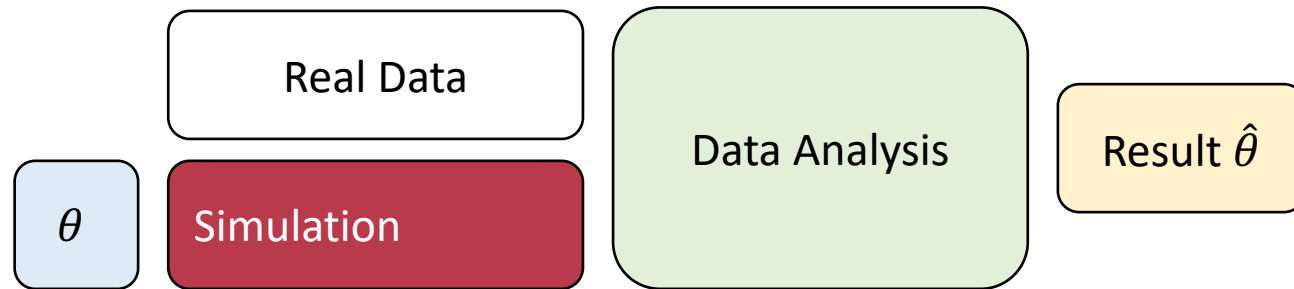


Long history of ML to improve particle pattern recognition and event classification

- See the Representations Workshop Summary

# Data Analysis Pipeline in HEP

7



Data analysis: “Invert” generation process

$$p(f(x) | \theta)$$

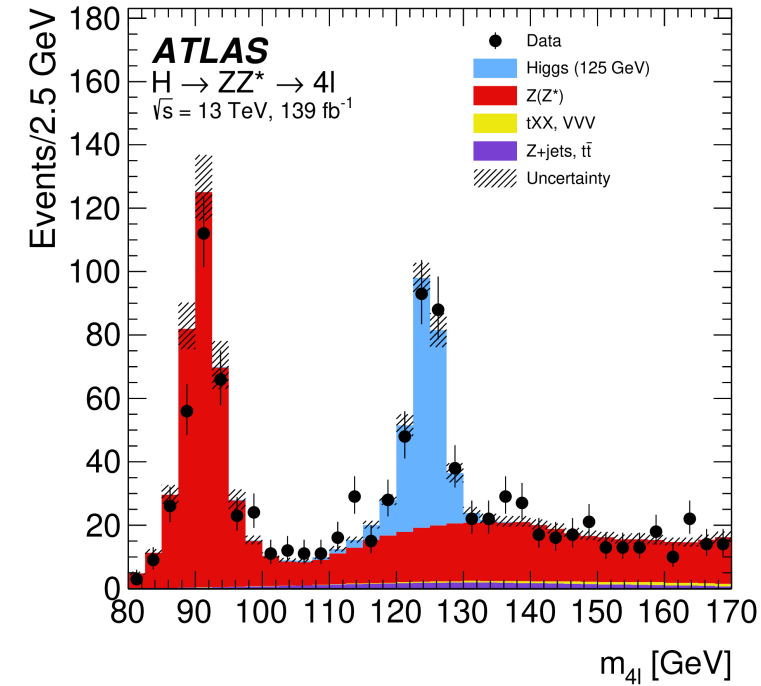
Inference: Reduce 100M  $\rightarrow$  1 informative number

# Poisson Likelihood over Bins

Slide Credit: N. Berger 8

$$P(\mu, \{\theta_j\}_{j=1 \dots n_{NP}}; \{n_i^{(k)}\}_{i=1 \dots n_{data}^{(k)}}, \{\theta_j^{obs}\}_{j=1 \dots n_{NP}}) =$$

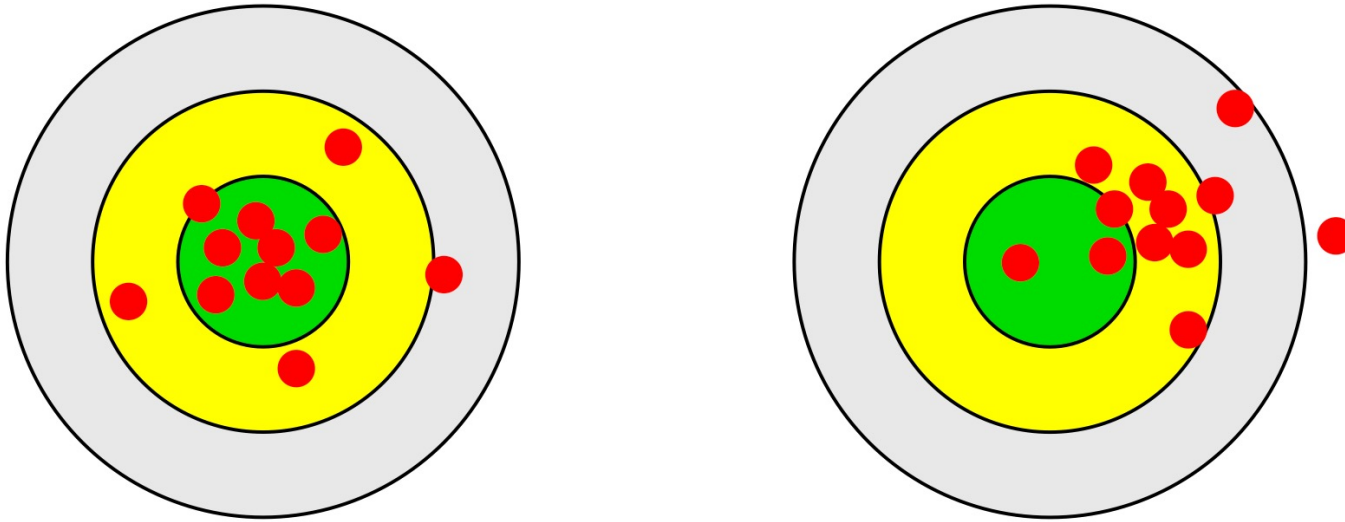
$$\prod_{k=1}^{n_{cats}} P[n_i; \mu \epsilon_{i,k}(\vec{\theta}) N_{S,i,k}(\vec{\theta}) + B_{i,k}(\vec{\theta})] \prod_{j=1}^{n_{syst}} G(\theta_j^{obs}; \theta_j; 1)$$



# Systematic Uncertainties

Slide Credit: N. Berger 9

The statistical model (PDF) is a way to express **uncertainty** on the outcome of an experiment. e.g. 2D Gaussian :



These uncertainties are also called **Statistical Uncertainties** – they are the ones encoded in the model PDF.

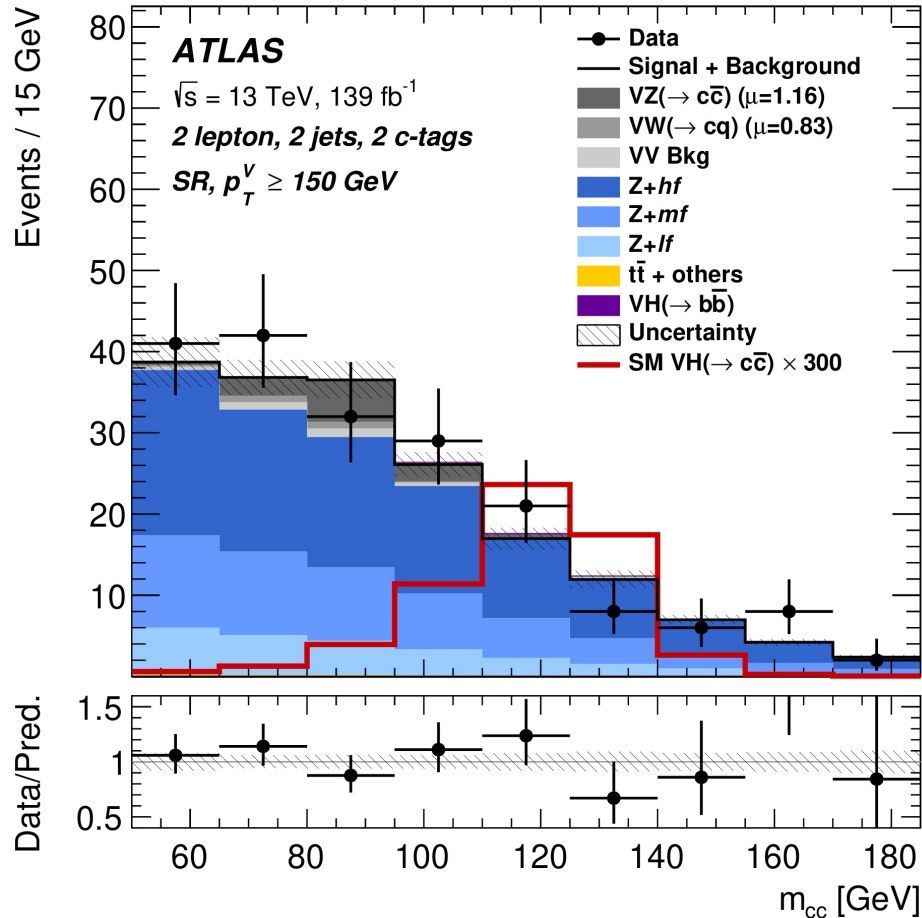
However **the model itself may be wrong** : this is a *systematic error*  
→ To account for them, need a set of **Systematic uncertainties**

## Domain Shift in ML

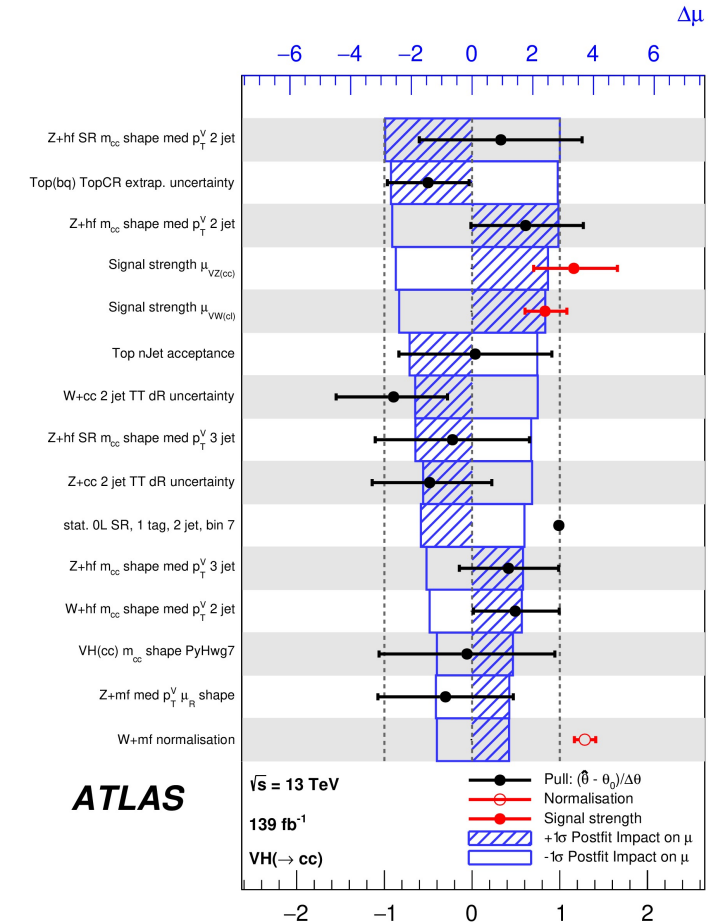
The distribution of training data is different from the distribution of data on which we will apply the model

# Systematic Uncertainties

10



Source of uncertainty	$\mu_{VH(cc)}$
Total	21.5
Statistical	16.2
Systematics	14.0
Statistical uncertainties	
Data statistics only	13.0
Floating normalisations	7.2
Theoretical and modelling uncertainties	
VH( $\rightarrow c\bar{c}$ )	2.1
Z+jets	7.7
Top-quark	5.6
W+jets	3.4
Diboson	0.8
VH( $\rightarrow b\bar{b}$ )	0.8
Multi-Jet	1.0
Simulation statistics	
	5.1
Experimental uncertainties	
Jets	3.7
Leptons	0.4
$E_T^{\text{miss}}$	0.5
Pile-up and luminosity	0.4
Flavour tagging	
c-jets	2.3
b-jets	1.2
light-jets	0.7
$\tau$ -jets	0.4
Truth-flavour tagging	
$\Delta R$ correction	3.0
Residual non-closure	1.4





# Approaches for Handling Systematic Uncertainty in HEP-ML Models

11

- **propagation of errors:** one works with a model  $f(x)$  and simply characterizes how uncertainty in the data distribution propagate through the function to the down-stream task irrespective of how it was trained.
- **domain adaptation:** one incorporates knowledge of the distribution for domains (or the parameterized family of distributions  $p(x|y, \nu)$ ) into the training procedure so that the performance of  $f(x)$  for the down-stream task is robust or insensitive to the uncertainty in  $\nu$ .
- **parameterized models:** instead of learning a single function of the data  $f(x)$ , one learns a family of functions  $f(x; \nu)$  that is explicitly parameterized in terms of nuisance parameters and then accounts for the dependence on the nuisance parameters in the down-stream task.
- **data augmentation:** one trains a model  $f(x)$  in the usual way using training dataset from multiple domains by sampling from some distribution over  $\nu$ .

From nice new [PDG review of ML in HEP](#), including discussion of uncertainties

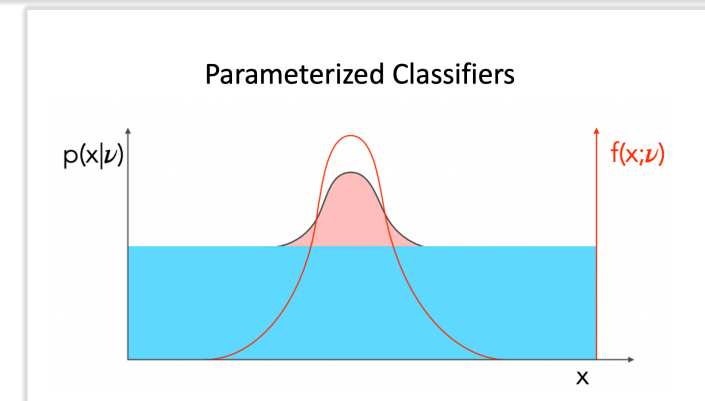
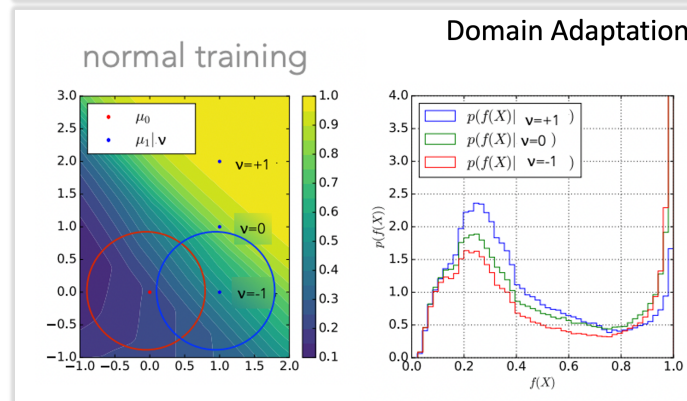
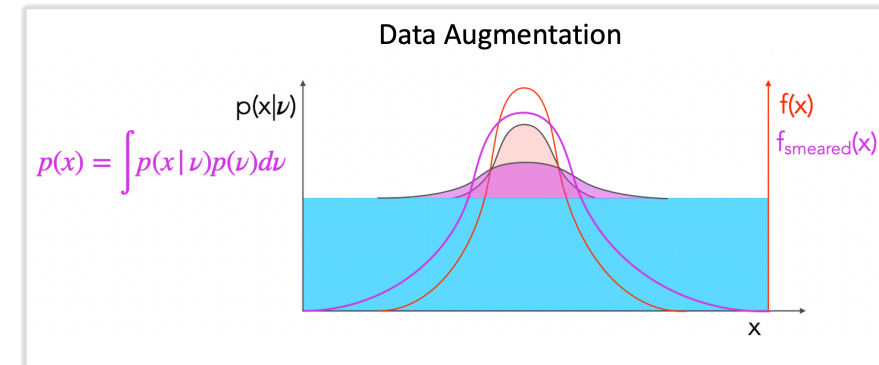
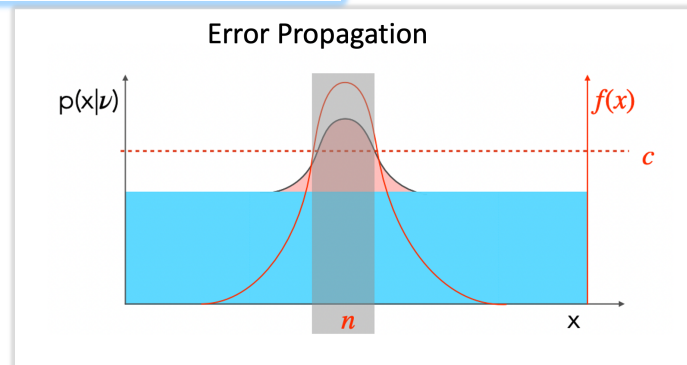


Image credit:  
[K. Cranmer](#)

# Approaches for Handling Systematic Uncertainty in HEP-ML Models

12

- **propagation of errors:** one works with a model  $f(x)$  and simply characterizes how uncertainty in the data distribution propagate through the function to the down-stream task irrespective of how it was trained.
- **domain adaptation:** one incorporates knowledge of the distribution for domains (or the parameterized family of distributions  $p(x|y, \nu)$ ) into the training procedure so that the performance of  $f(x)$  for the down-stream task is robust or insensitive to the uncertainty in  $\nu$ .
- **parameterized models:** instead of learning a single function of the data  $f(x)$ , one learns a family of functions  $f(x; \nu)$  that is explicitly parameterized in terms of nuisance parameters and then accounts for the dependence on the nuisance parameters in the down-stream task.
- **data augmentation:** one trains a model  $f(x)$  in the usual way using training dataset from multiple domains by sampling from some distribution over  $\nu$ .

From nice new [PDG review of ML in HEP](#), including discussion of uncertainties

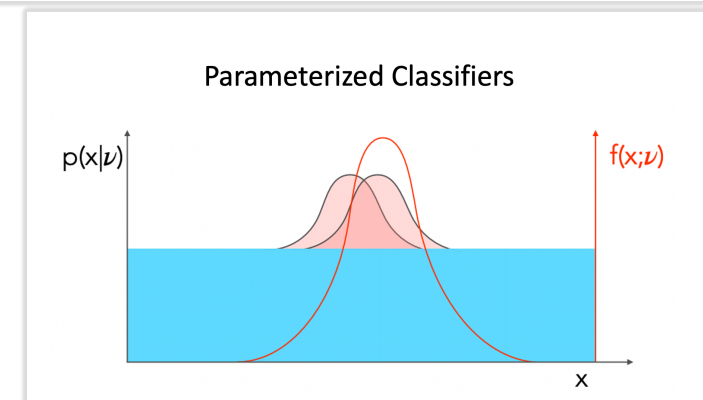
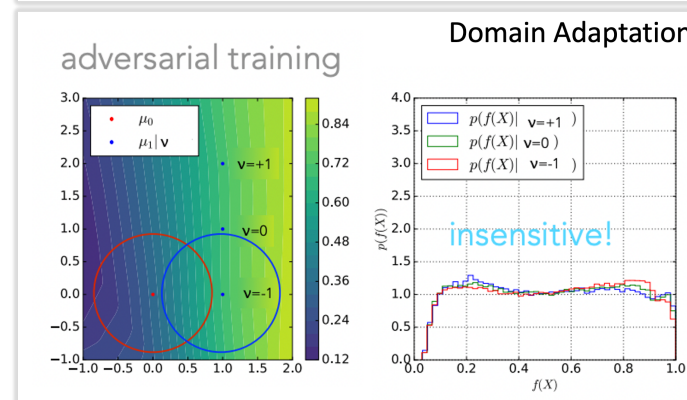
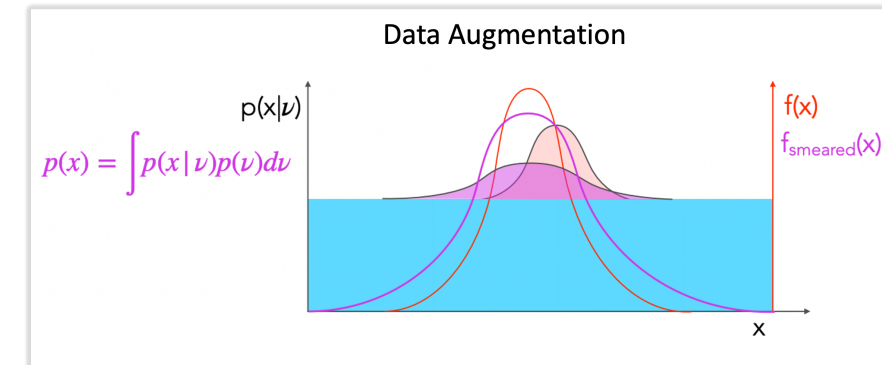
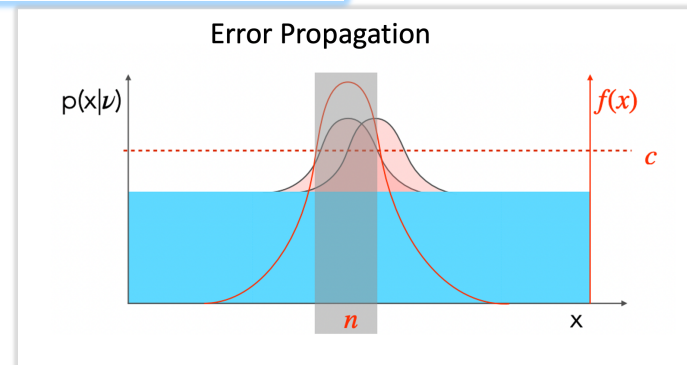


Image credit:  
[K. Cranmer](#)



# Poisson Likelihood over Bins

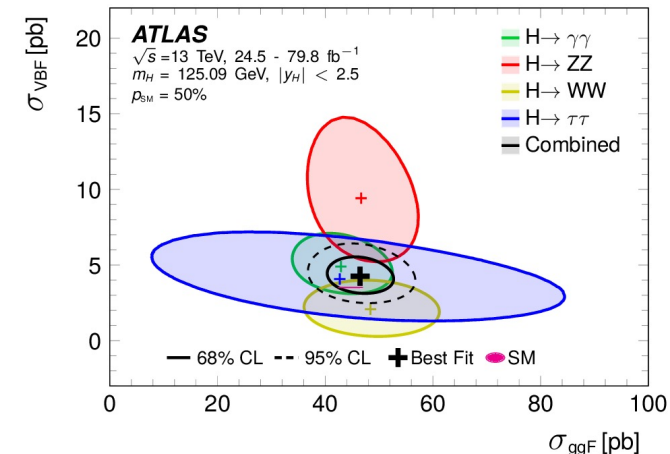
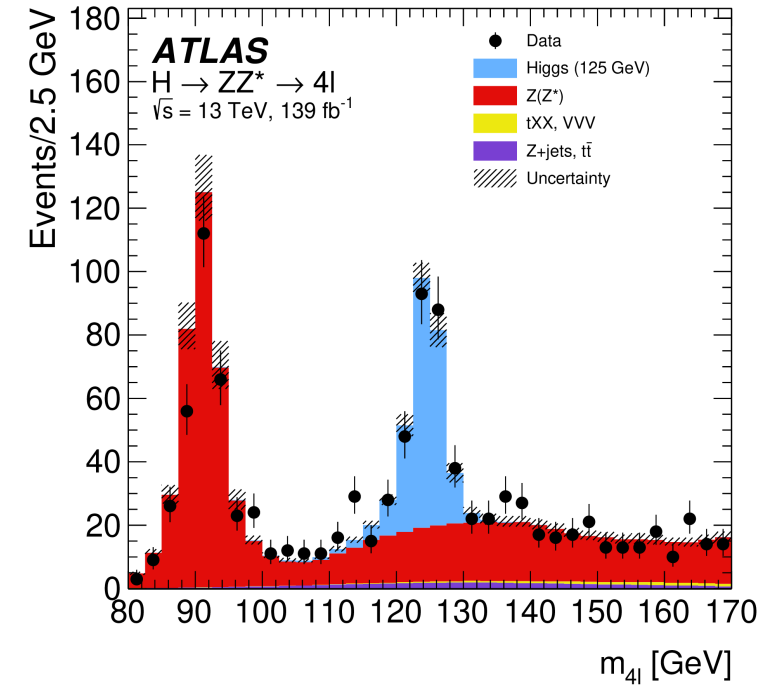
Slide Credit: N. Berger 13

$$P(\boldsymbol{\mu}, \{\boldsymbol{\theta}_j\}_{j=1 \dots n_{NP}}; \{\mathbf{n}_i^{(k)}\}_{i=1 \dots n_{data}^{(k)}}^{k=1 \dots n_{cat}}, \{\boldsymbol{\theta}_j^{obs}\}_{j=1 \dots n_{NP}}) =$$

$$\prod_{k=1}^{n_{cats}} P[\mathbf{n}_i; \boldsymbol{\mu} \boldsymbol{\epsilon}_{i,k}(\vec{\boldsymbol{\theta}}) N_{S,i,k}(\vec{\boldsymbol{\theta}}) + \mathbf{B}_{i,k}(\vec{\boldsymbol{\theta}})] \prod_{j=1}^{n_{syst}} G(\boldsymbol{\theta}_j^{obs}; \boldsymbol{\theta}_j; 1)$$



$$t(\boldsymbol{\mu}) = -2 \log \frac{L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}))}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}$$



# Uncertainty in ML

Talks:

- 1) N. Brunel, T. Morzadec, V. Taquet, V. Blot
- 2) G. Daniel

# Predictive, aleatoric and epistemic uncertainties

15

Let  $x$  an input point,  $f_\omega$  a predictive model with parameters  $\omega$

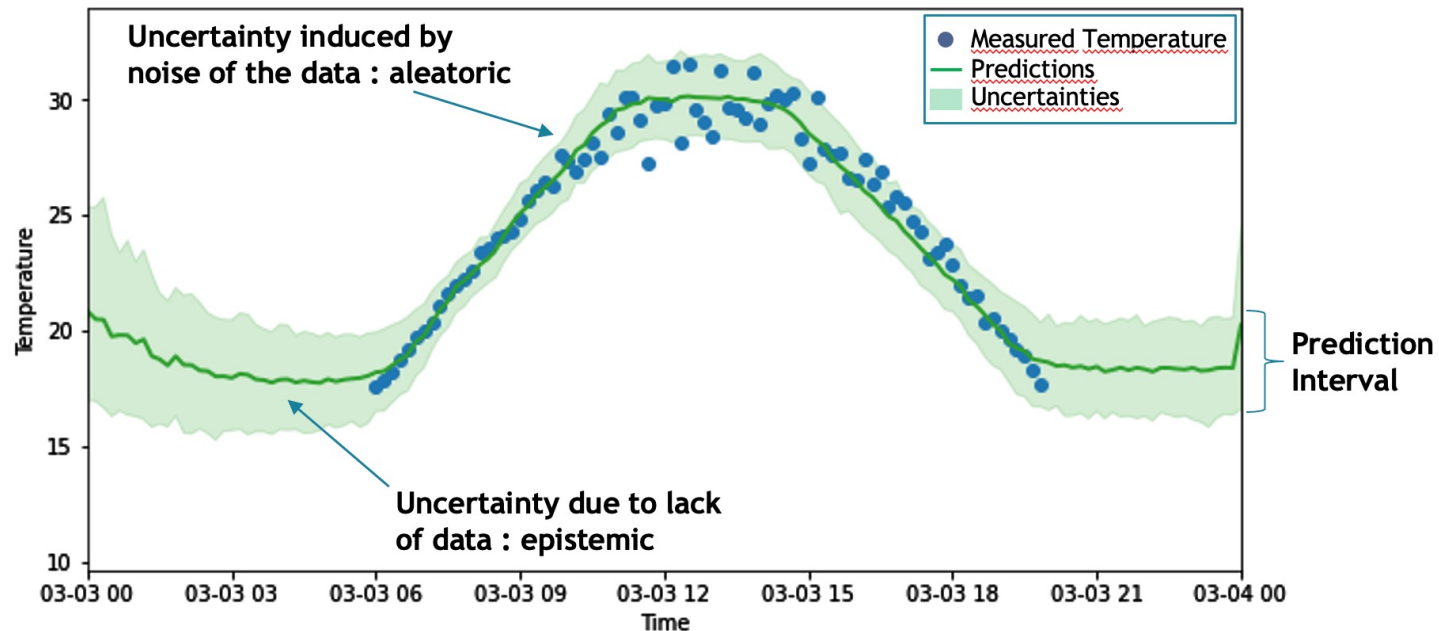
**Objective:** Quantifying the uncertainty on the prediction  $f_\omega(x)$   
→ **Predictive uncertainty**

**Aleatoric uncertainty**

→ Uncertainty related to the data/the phenomenon

**Epistemic uncertainty**

→ Uncertainty related to the model



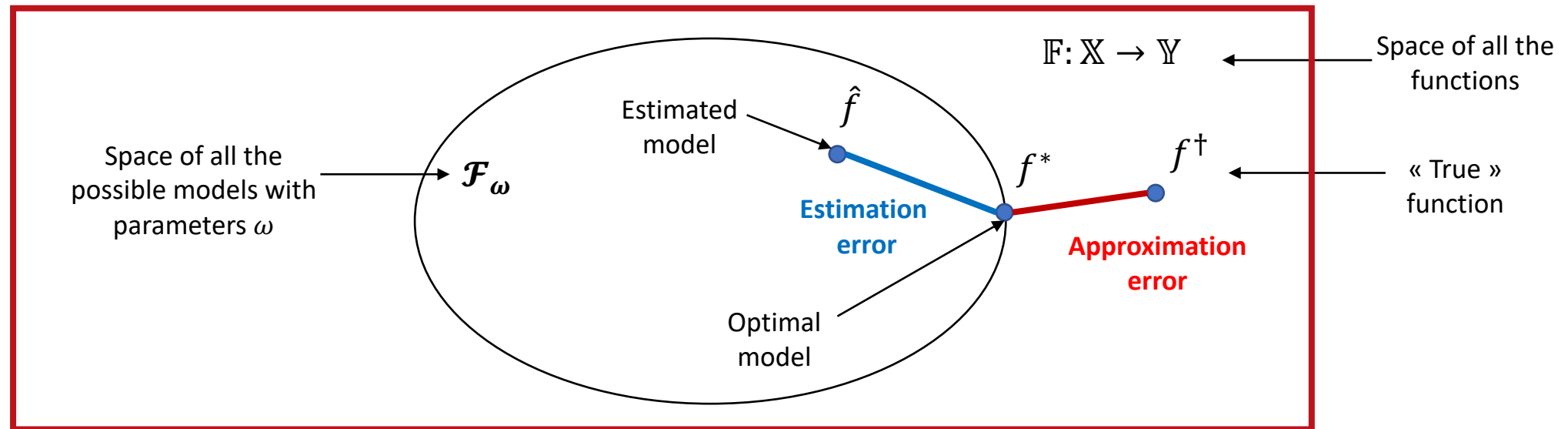
# Epistemic Uncertainty

16

Represents the lack of « knowledge » or « understanding » of a model on a specific input data point

Two main origins of epistemic uncertainty for machine learning models:

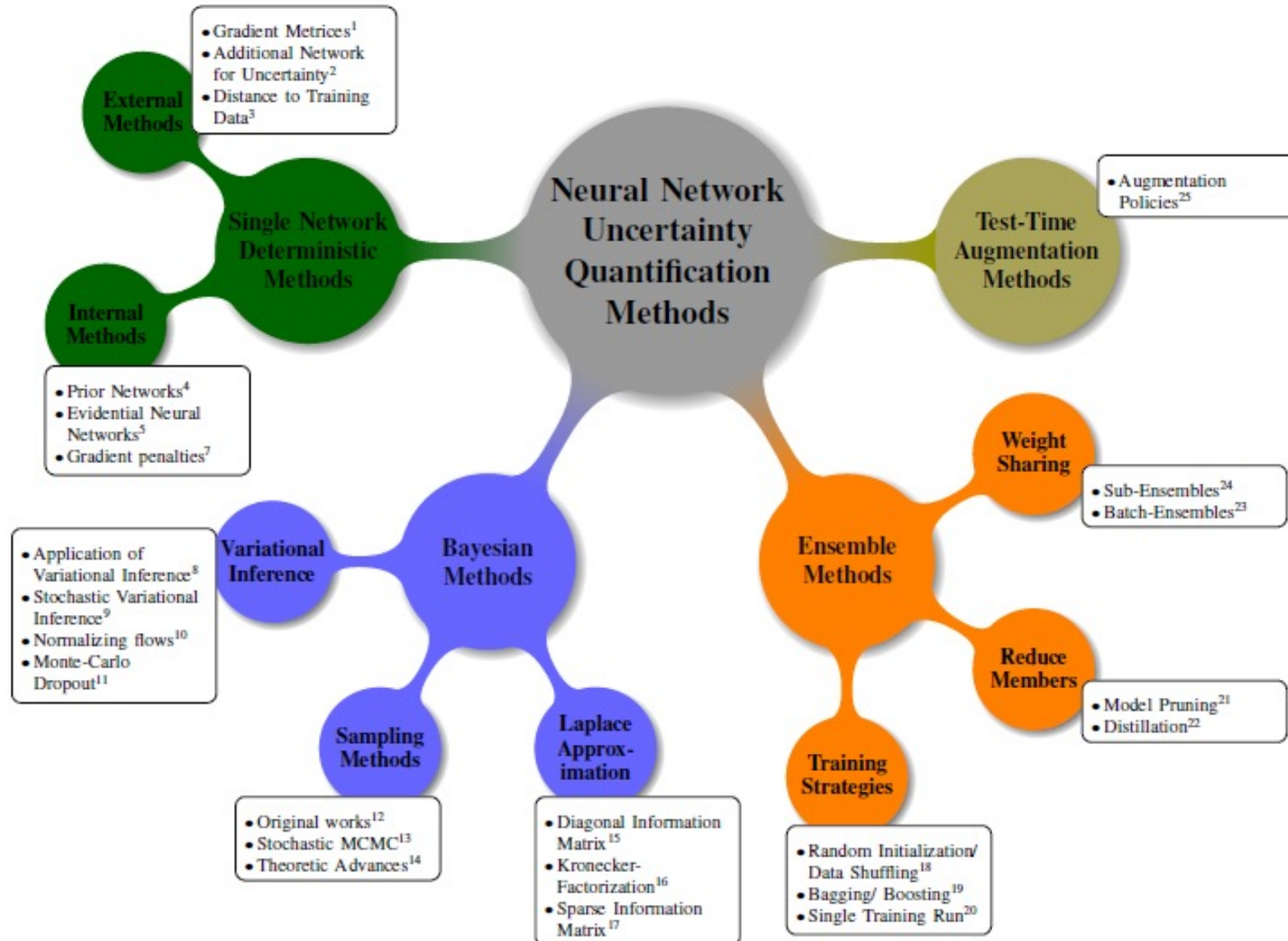
- **Estimation error:** the training dataset is just a sample of all the possible observable data
- **Approximation error:** no model can approximate perfectly the unknown « true » function



It can be possible to reduce epistemic uncertainty by using more data and increasing the model complexity

# Uncertainty Estimation Approaches in Deep Learning

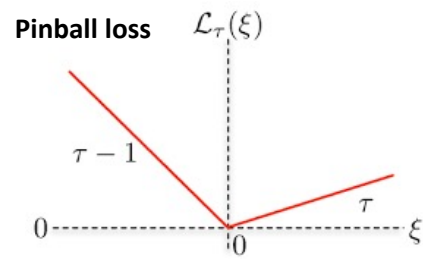
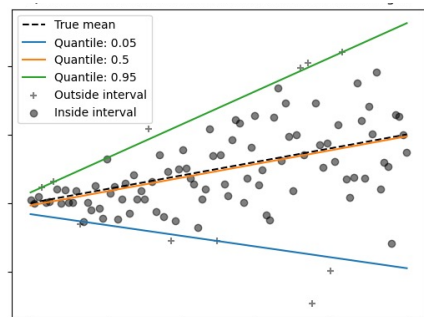
17



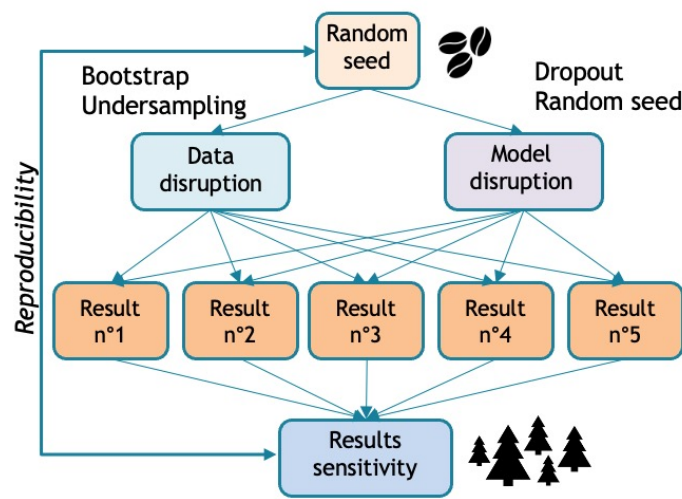
# Common approaches for computing Confidence Intervals

18

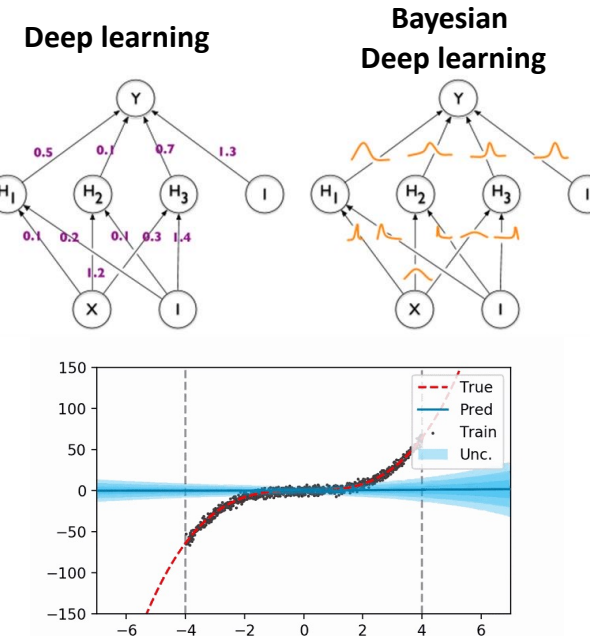
## Quantile Regression



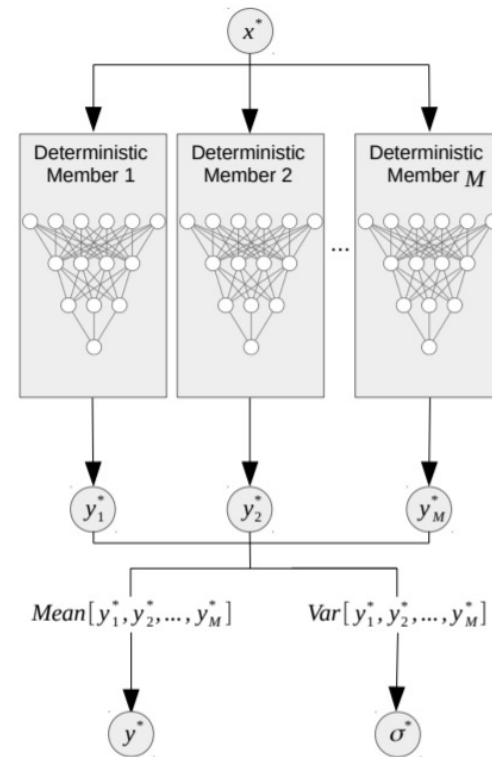
## Model and Data perturbations



## Bayesian Inference



## Ensemble Methods



A Survey of Uncertainty in Deep Neural Networks, J. Gawlikowski et al.,  
[arXiv:2107.03342](https://arxiv.org/abs/2107.03342)

# MAPIE – Conformal Prediction

19

Model Agnostic Prediction Interval Estimator: a scikit-learn-contrib library

- ▶ For  $\mathcal{Y} = \mathbb{R}$ , consider a pre-fitted model  $\hat{\mu} = \mathcal{A}(\mathcal{D}_n^{Tr})$  ( $\mathcal{A}$  : any ML algo = Random Forest, Boosting, Neural Network...)
- ▶ Define a non-conformity score function  $s(x, y) \in \mathbb{R}$

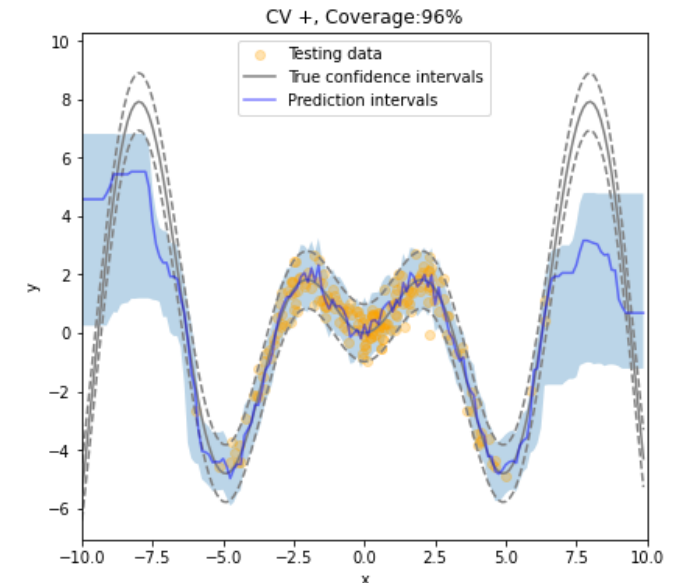
$$s(x, y) = |\hat{\mu}(x) - y|$$

- ▶ Consider a **holdout calibration** data set  $\mathcal{D}_N^{Cal} = \{Z_i, i = 1, \dots, N\}$  and compute the (almost)  $(1 - \alpha)$ -quantile

$$\hat{q}_{1-\alpha} = \text{Quantile} \left( \left\{ S_i \triangleq s(X_i, Y_i) \right\}; \frac{\lceil (N + 1)(1 - \alpha) \rceil}{N} \right)$$

- ▶ The prediction set for new  $X_{N+1}$  is

$$\hat{C}_{N,\alpha}(X_{N+1}) = \{y \in \mathbb{R} \mid s(X_{N+1}, y) \leq \hat{q}_{1-\alpha}\} = [\hat{\mu}(X_{N+1}) \pm \hat{q}_{1-\alpha}]$$





# Model Calibration – Do Output Intervals Make Sense?

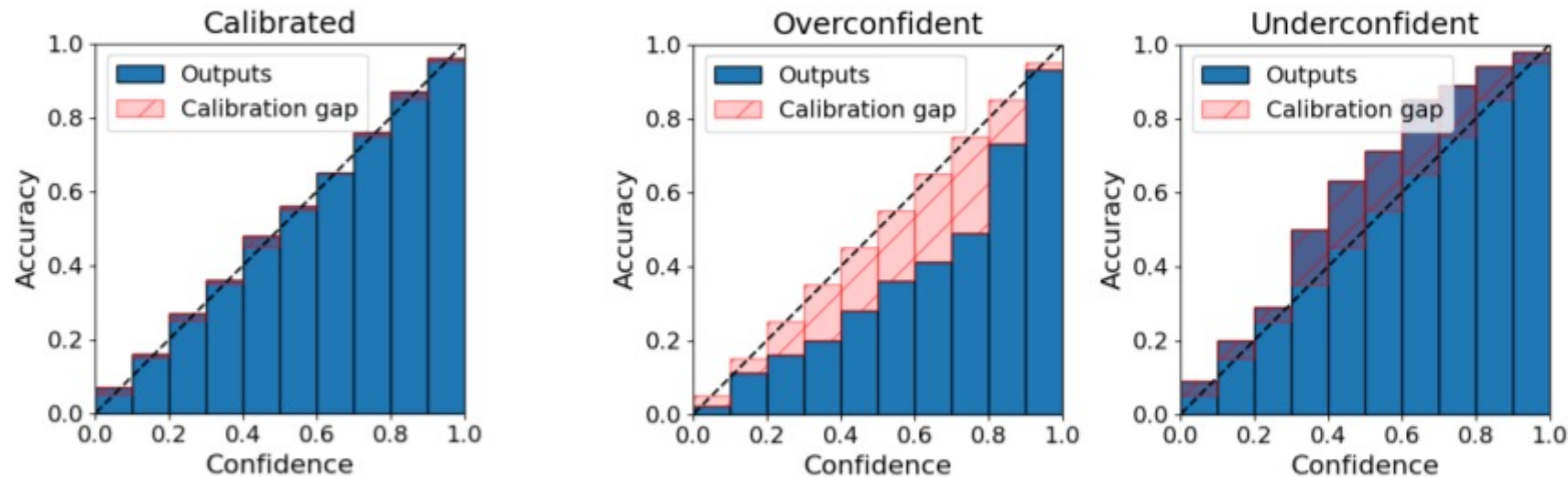
20

Main idea: probabilistic interpretation of the neural network output:

$$f_{\omega}^{(i)} = p(y^{(i)}|x)$$

Comparison in the test set between the frequency of correct answers and the associated probabilities

Example: the frequency of correct classifications with a predicted probability 40% should be 40%



A Survey of Uncertainty in Deep Neural Networks, J. Gawlikowski et al.,  
[arXiv:2107.03342](https://arxiv.org/abs/2107.03342)

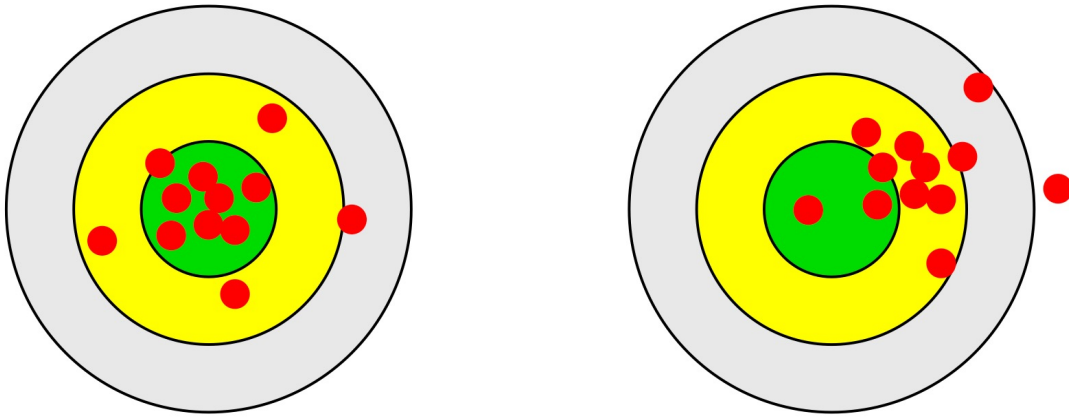


Do we need to worry about Epistemic Uncertainty in HEP?

Answer likely depends on the task, and the statistical inference method

# Different Kinds of Uncertainty in Focus

22

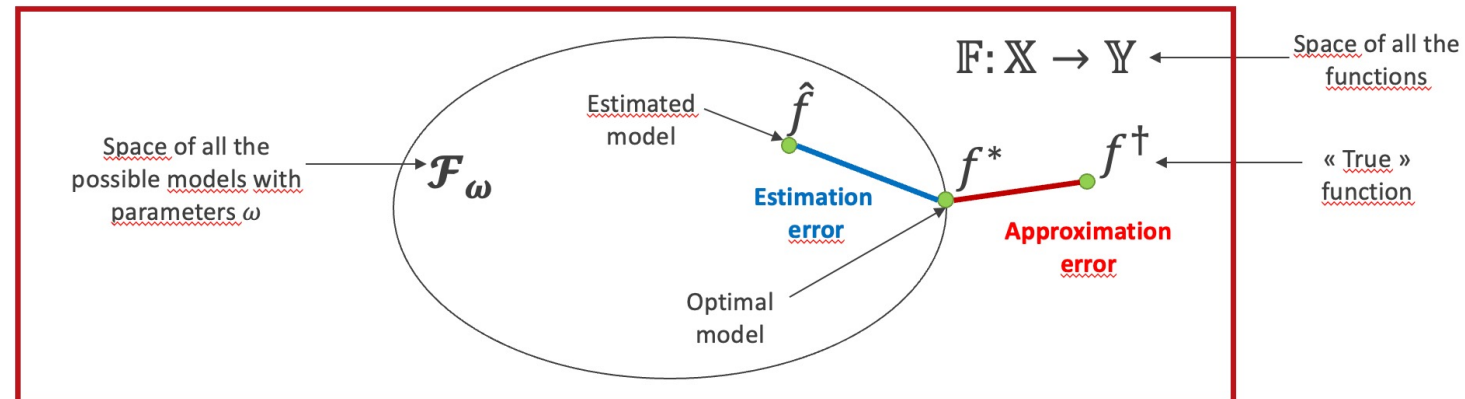


Systematic Uncertainties  $\rightarrow$  Domain Shift

How wrong is our model when the data changes a little bit

## Epistemic Uncertainties

For fixed data, does our model represent the underlying relationships



# ML in Reconstruction and Event Classification

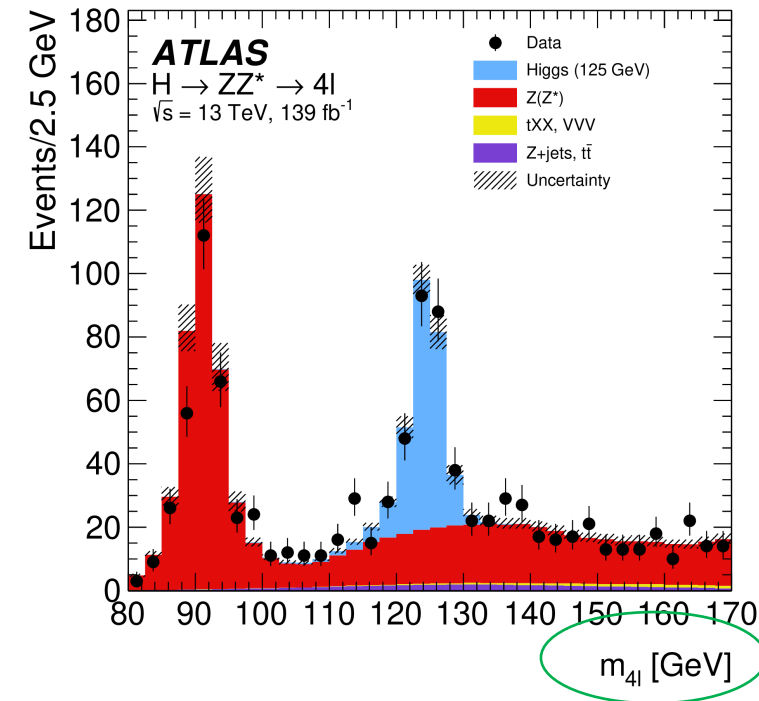
23

For the moment, let's ignore data / simulation differences (domain shift)

Does it matter what model we use for electron energy estimation, or for classifying Higgs bosons?

→ *This determines our summary feature*

- Seems more a question of *optimality* rather than uncertainty
- Fitting training data imperfectly, leads to suboptimal results but not wrong results
- Epistemic uncertainty may not be important in this case, but mainly a matter of systematic uncertainties



# ML in modeling Signal and Background

24

Need epistemic uncertainty when using ML to model signal & backgrounds?

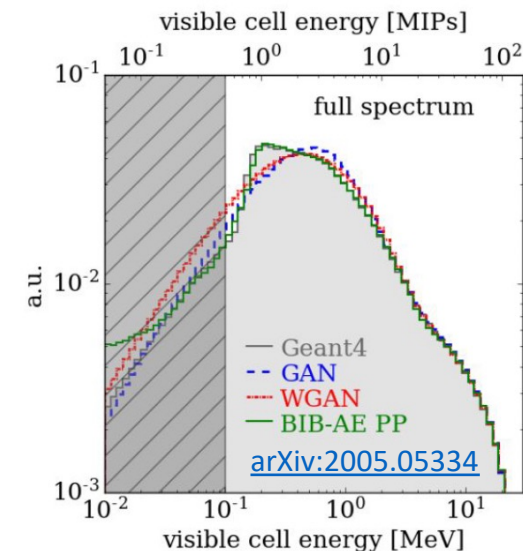
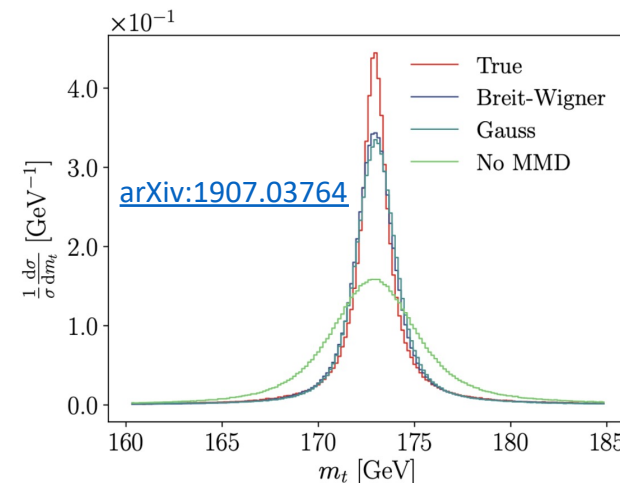
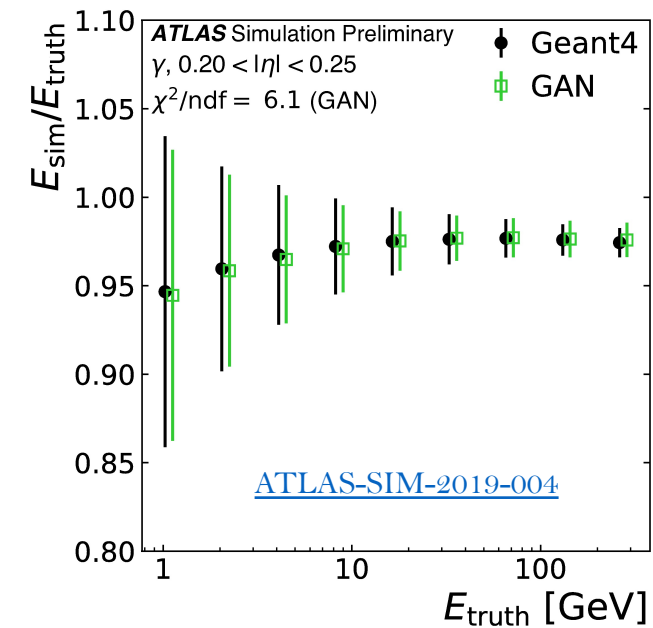
- ML for fast simulations of detector or theory predictions
- ML for reweighting backgrounds from control regions
- Data-driven ML-based density estimation

Bad signal / background predictions means statistical model is wrong

- Need to account for how wrong ML model could be
- Or could this be covered by estimating systematic uncertainties?

These are open questions, likely needing some R&D

- Not clear how to perform uncertainty estimation for ML-based density estimation



# Different Inference Procedure?

---

25

Story will change for different statistical model (rather than histograms)...

What about Matrix Element Methods, or other methods that try to estimate a per-event likelihood?

Bayesian method often assign priors and consider uncertainty on NN weights

- E.g. Bayesian Neural Network, Variational Inference, ...
- Ideas only beginning to be explored in HEP
- Can learn a lot from our Cosmology colleagues who focus on Bayesian Inference

Also open questions for HEP, likely in need of R&D

# Beyond Histograms → Simulation Based Inference (SBI)

Talks:

1) G. Louppe

# Simulation Based Inference (SBI)

27

Start with

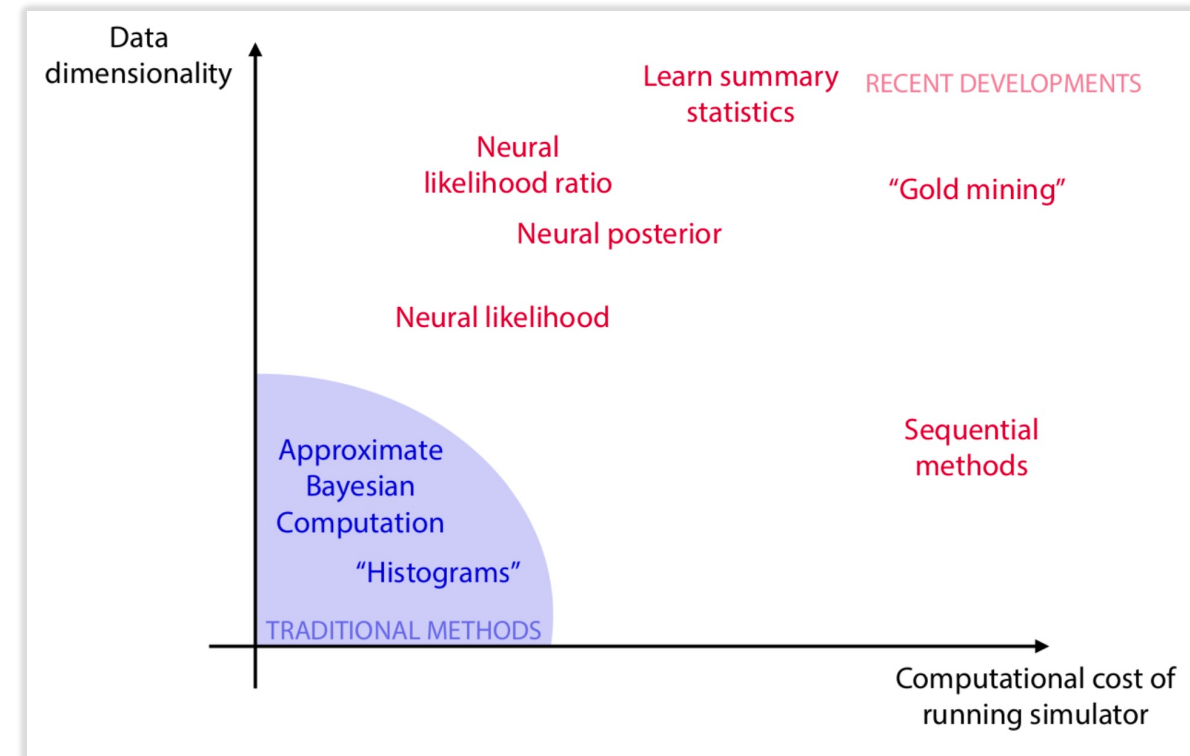
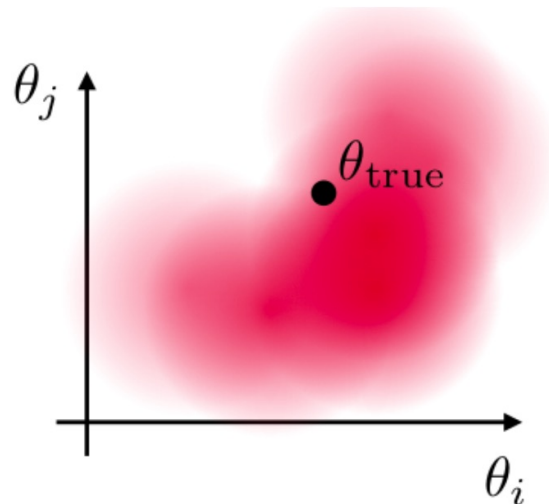
- a simulator that can generate  $N$  samples  $x_i \sim p(x_i | \theta_i)$ ,
- a prior model  $p(\theta)$ ,
- observed data  $x_{\text{obs}} \sim p(x_{\text{obs}} | \theta_{\text{true}})$ .

Then, estimate the posterior

$$p(\theta | x_{\text{obs}}) = \frac{p(x_{\text{obs}} | \theta) p(\theta)}{p(x_{\text{obs}})}$$

Or a likelihood ratio

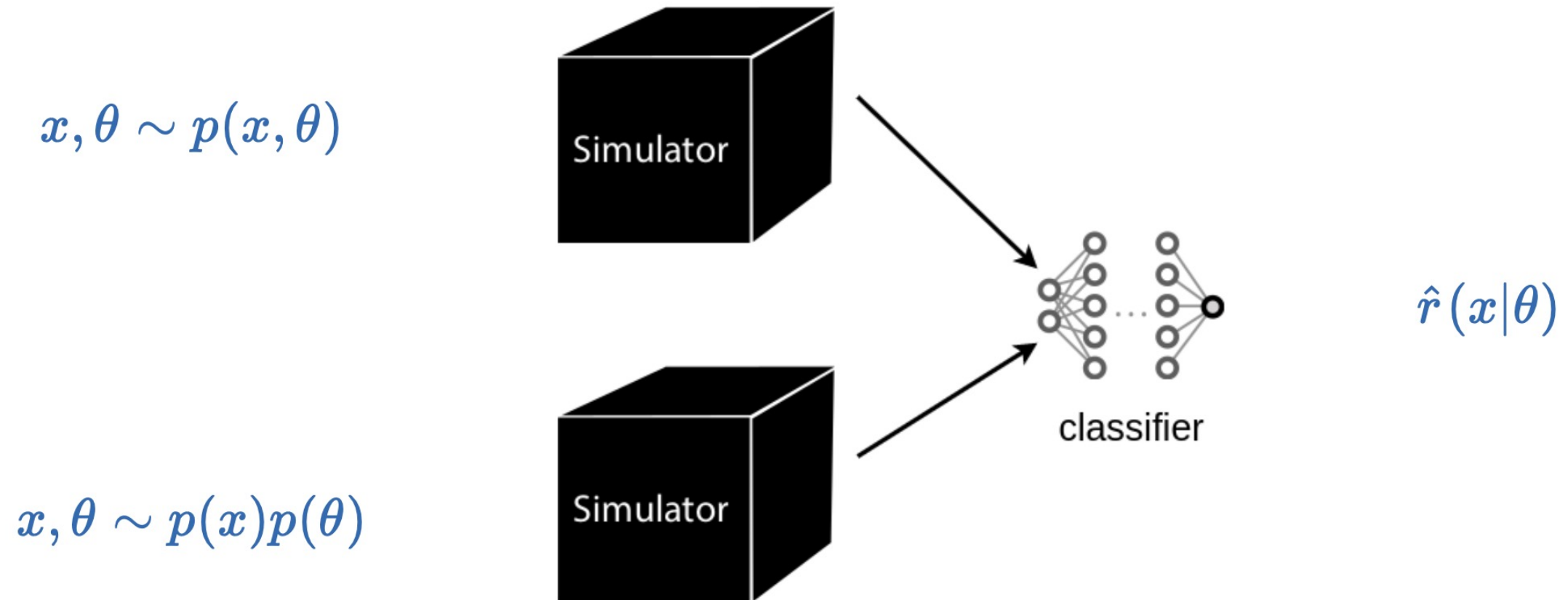
$$r(\theta) = \frac{p(x_{\text{obs}} | \theta)}{p(x_{\text{obs}} | \theta_0)}$$



# Neural Ratio Estimation

28

The likelihood-to-evidence  $r(x|\theta) = \frac{p(x|\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)p(\theta)}$  ratio can be learned, even if neither the likelihood nor the evidence can be evaluated:



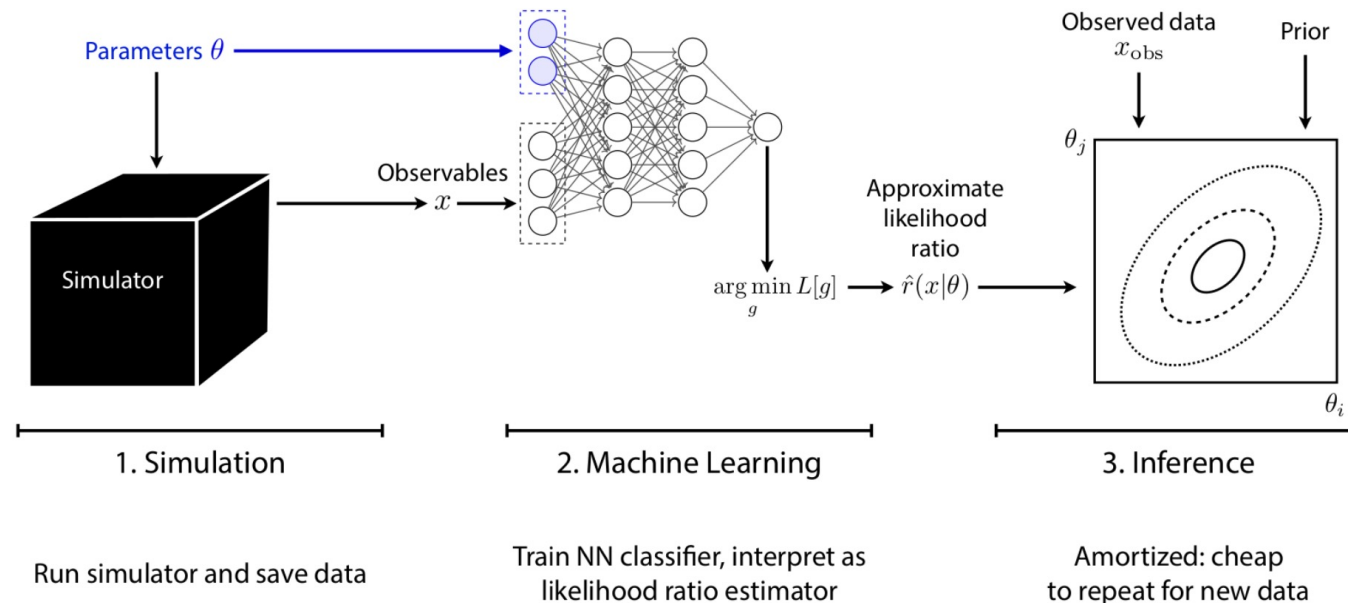


# Neural Ratio Estimation

29

The likelihood-to-evidence  $r(x|\theta) = \frac{p(x|\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)p(\theta)}$  ratio can be learned, even if neither the likelihood nor the evidence can be evaluated:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \approx \hat{r}(x|\theta)p(\theta)$$

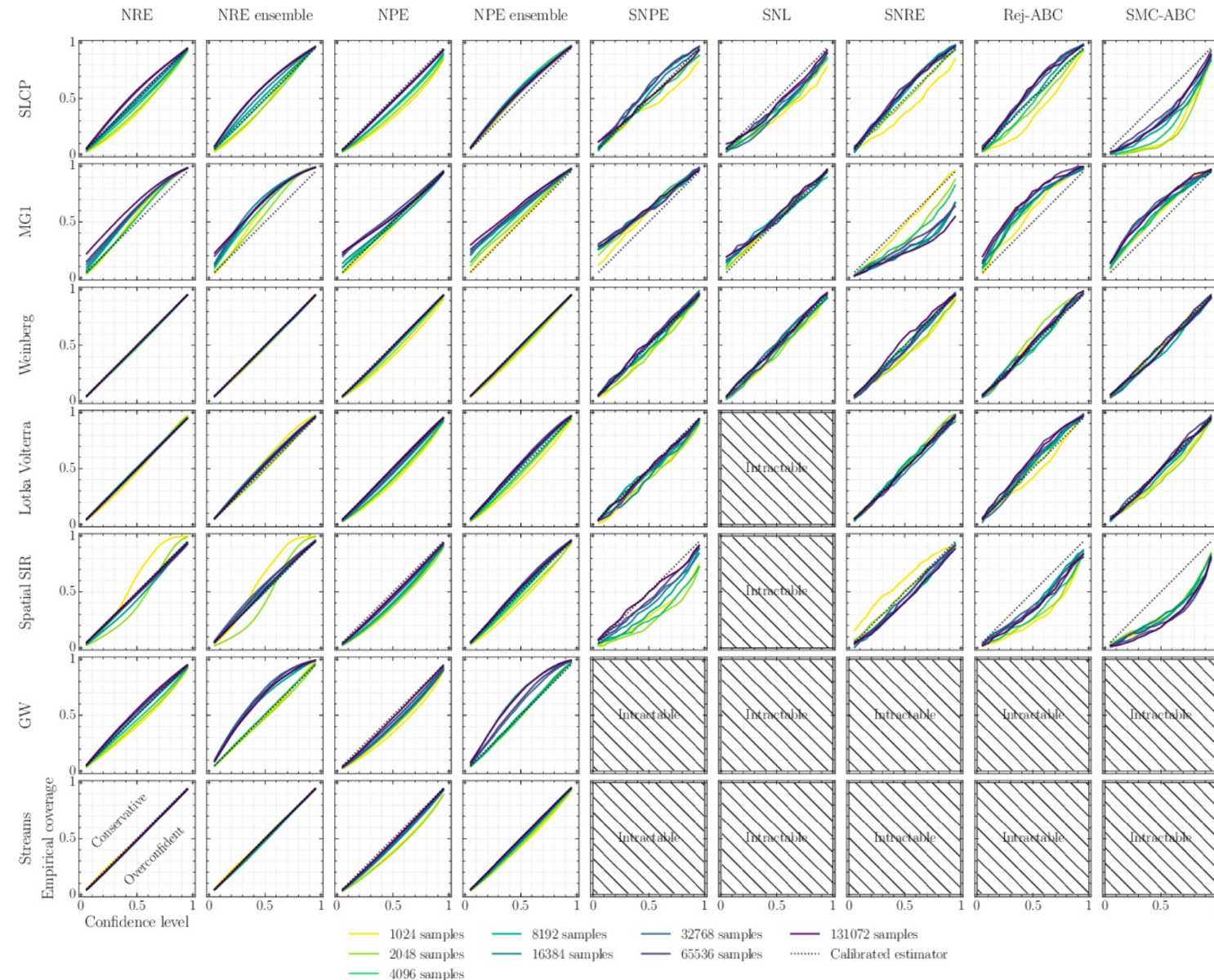
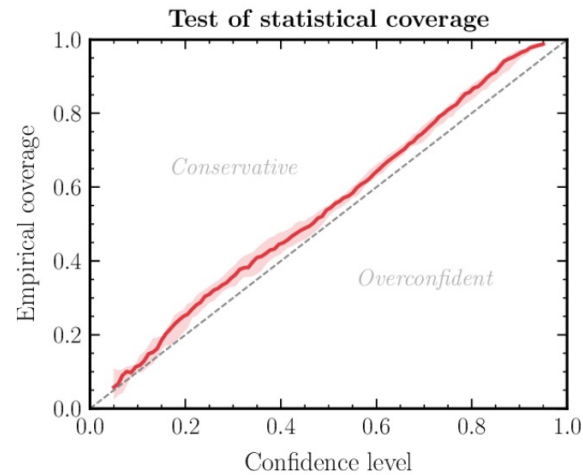


# But proceed with caution! ... model checking, evaluation, and criticism

30

## Coverage diagnostic:

- For  $x, \theta \sim p(x, \theta)$ , compute the  $1 - \alpha$  credible interval based on  $\hat{p}(\theta|x)$ .
- If the fraction of samples for which  $\theta$  is contained within the interval is larger than the nominal coverage probability  $1 - \alpha$ , then the approximate posterior  $\hat{p}(\theta|x)$  has coverage.



# Data Sets

# Datasets for Studying Uncertainties

32

Nice discussion on what data sets are available / needed, for a typical cross-section measurement (measuring a signal which shape is known)

Available:

- Higgs ML:  $H \rightarrow \tau\tau \rightarrow lh$ ,
  - script <https://zenodo.org/record/1887847> to create 5 different systematics) :
  - realistic but too few events for a precise evaluation of systematic uncertainties
- UCI Higgs  $H \rightarrow \tau\tau \rightarrow ll$  :
  - Large dataset, but less handles to introduce systematics.
- CMS open data for Upcoming analysis using Inferno
  - H cross section will make public a script to generate data from CMS open data
  - Should be complex enough but is statistics large enough ?

Request from ML experts for a simplified system that maintains key aspects

- Useful to explore system simpler than LHC events and more realistic than gaussian

Figure of merit :

- Total uncertainty (stat and syst), or confidence intervals?

Large focus on uncertainty modeling both in HEP and in ML

- HEP focus on systematics uncertainties from data / simulation differences
- ML focus on aleatoric / epistemic uncertainty on if we fit the correct model

When do we need these ML uncertainties?

- Will depend on modeling task and inference method
- Dedicated work needed to understand this in HEP

Great deal of work to move beyond histograms and modeling likelihood, likelihood ratios, and posteriors

- Capture the variations in data
- Can include nuisance parameters for systematic uncertainties
- But they are approximations... Ongoing work on how to validate these models

# Backup



# Systematic Uncertainties

35

Statistical models include:

- **Parameters of interest** (POIs) :  $\mathbf{S}, \sigma \times \mathbf{B}, m_W, \dots$
- **Nuisance parameters** (NPs) : other parameters needed to define the model  
→ Ideally, **constrained by data** like the POI

**And systematics ?**

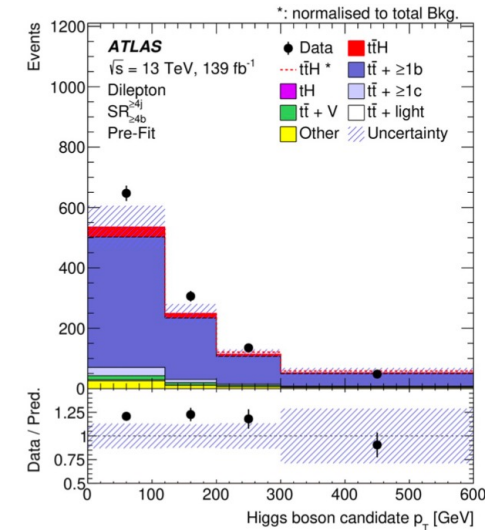
= Cover what we don't know about the random process.

⇒ **Parameterize using additional NPs**

→ Can't be constrained by the data ⇒ Add constraints in the likelihood

$$L(\underbrace{\mu}_{\text{POI}}, \underbrace{\theta}_{\text{Systematics NP}}; \text{data}) = \underbrace{L_{\text{measurement}}(\mu, \theta; \text{data})}_{\text{Measurement Likelihood}} \underbrace{C(\theta)}_{\text{NP Constraint term}}$$

$C(\theta)$  represents **external knowledge** about the NP



"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables.

G. Punzi, *What is systematics ?*

# MAPIE – Conformal Prediction

36

We aim at building a Python library that computes confidence sets with three objectives :

Theoretical  
guarantees on  
the coverage  
probabilities

Model and use  
case agnostic

Algorithmic  
transparency  
for trustworthy  
AI

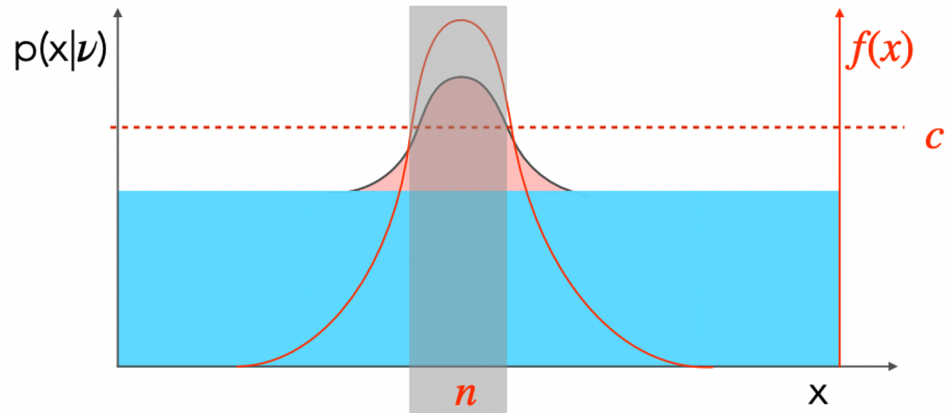
Method	Theoretical guarantees	Model Agnostic	Open-Source Implementation
Quantile Regression	✗	✗	✓
Data Perturbation (Bootstrap, Jackknife)	✗	✓	✓
Conformal Prediction	✓	✓	✗
Model Perturbation (Random seed, MC Dropout)	✗	✗	✓
Bayesian inference	✓	✗	✓

MAPIE

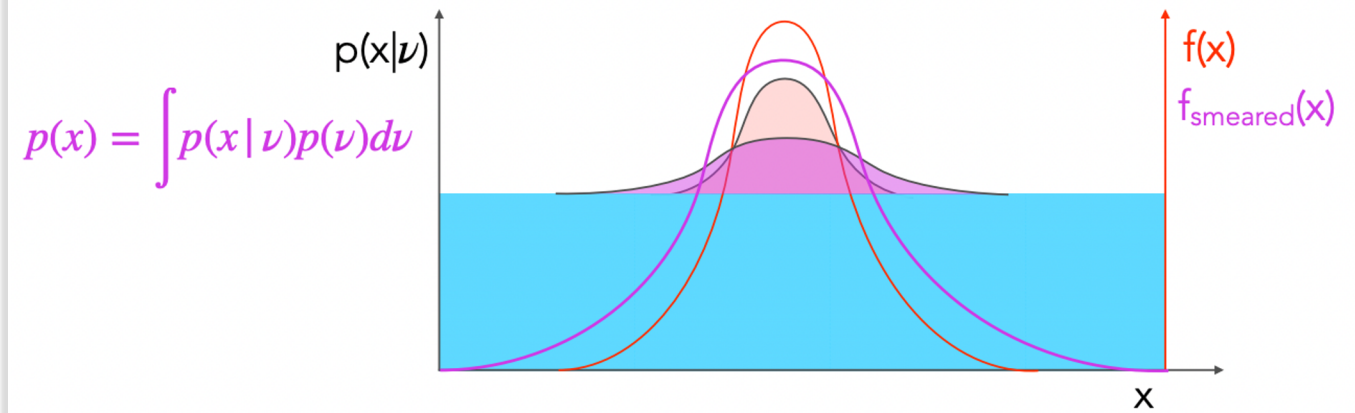


- **propagation of errors:** one works with a model  $f(x)$  and simply characterizes how uncertainty in the data distribution propagate through the function to the down-stream task irrespective of how it was trained.
- **domain adaptation:** one incorporates knowledge of the distribution for domains (or the parameterized family of distributions  $p(x|y, \nu)$  ) into the training procedure so that the performance of  $f(x)$  for the down-stream task is robust or insensitive to the uncertainty in  $\nu$ .
- **parameterized models:** instead of learning a single function of the data  $f(x)$ , one learns a family of functions  $f(x; \nu)$  that is explicitly parameterized in terms of nuisance parameters and then accounts for the dependence on the nuisance parameters in the down-stream task.
- **data augmentation:** one trains a model  $f(x)$  in the usual way using training dataset from multiple domains by sampling from some distribution over  $\nu$ .

### Error Propagation

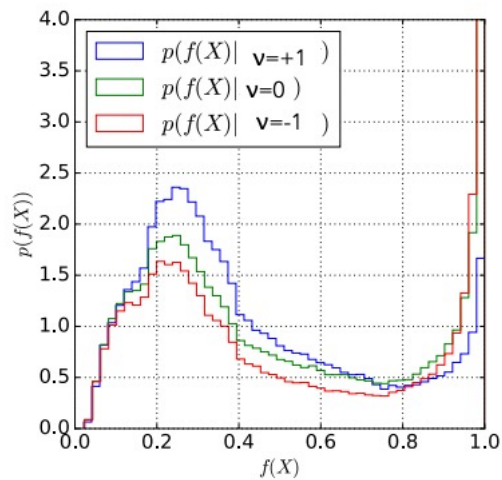
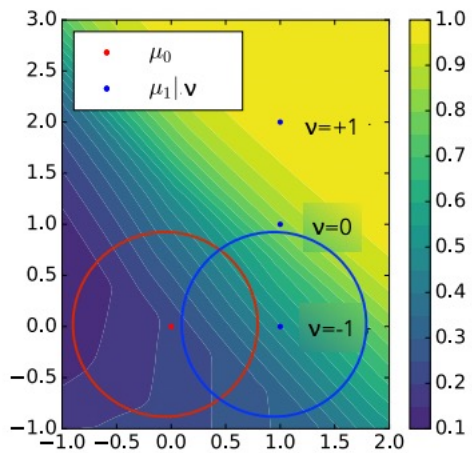


### Data Augmentation

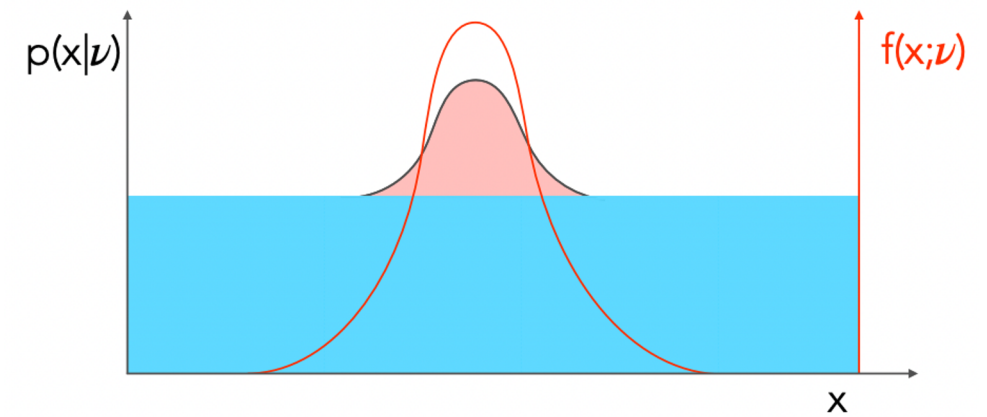


### Domain Adaptation

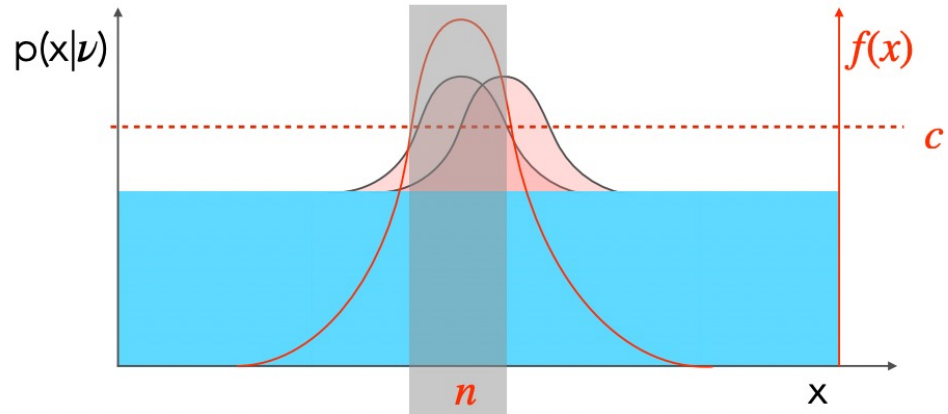
normal training



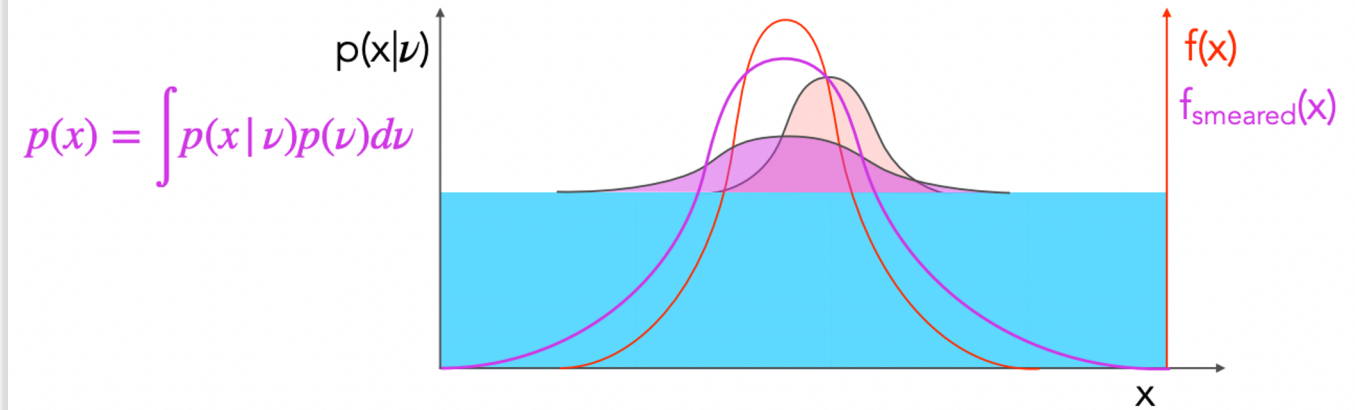
### Parameterized Classifiers



## Error Propagation

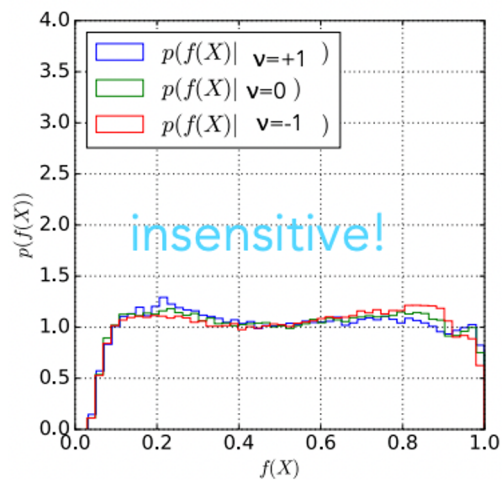
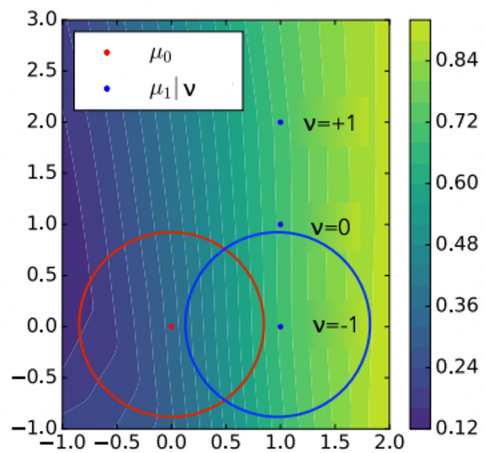


## Data Augmentation



## Domain Adaptation

### adversarial training



## Parameterized Classifiers

