Simulation-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms or Posterior Estimators

Luca Masserano[1]

Joint work with:

Tommaso Dorigo[2], Rafael Izbicki[3],

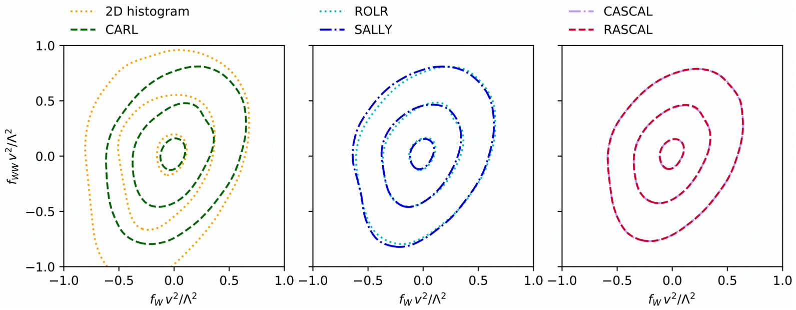
Mikael Kuusela[1], Ann B. Lee[1]

1. Department of Statistics and Data Science, Carnegie Mellon University
2. Italian Institute for Nuclear Physics and CERN
3. Department of Statistics, Federal University of Sao Carlos

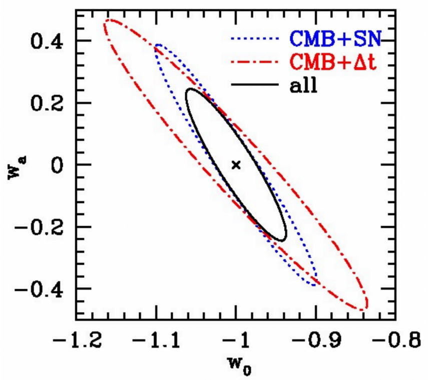
Carnegie Mellon University

Constraining Parameters → Uncertainty Quantification

- Much of modern Machine Learning targets prediction problems
- In many science applications, however, the interest is more on uncertainty quantification than in point estimation
- All the examples on the right are <u>*inverse*</u> problems. The interest is on internal parameters θ, i.e. the "causes" of $\mathbf{x}$



Particle Physics

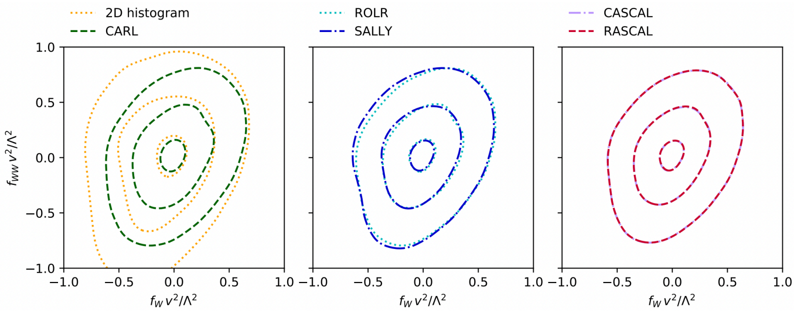


Cosmology

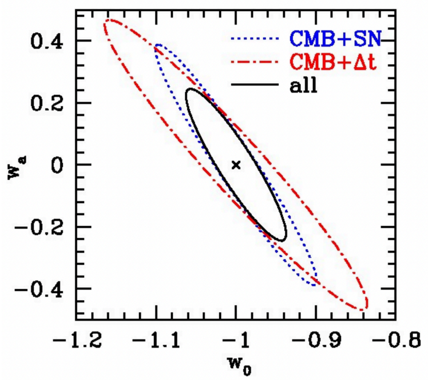
Image credits: Brehmer, J. et al. (2018); Abell, Paul A. et al. (2009)

Constraining Parameters → Uncertainty Quantification

- Much of modern Machine Learning targets prediction problems
- In many science applications, however, the interest is more on uncertainty quantification than in point estimation
- All the examples on the right are <u>*inverse*</u> problems. The interest is on internal parameters θ, i.e. the "causes" of $\mathbf{x}$



Particle Physics



Cosmology

Goal: constraining parameters of interest using theoretical (or simulation) models and experimental data, while guaranteeing coverage

Image credits: Brehmer, J. et al. (2018); Abell, Paul A. et al. (2009)

Science relies heavily on high-fidelity simulators

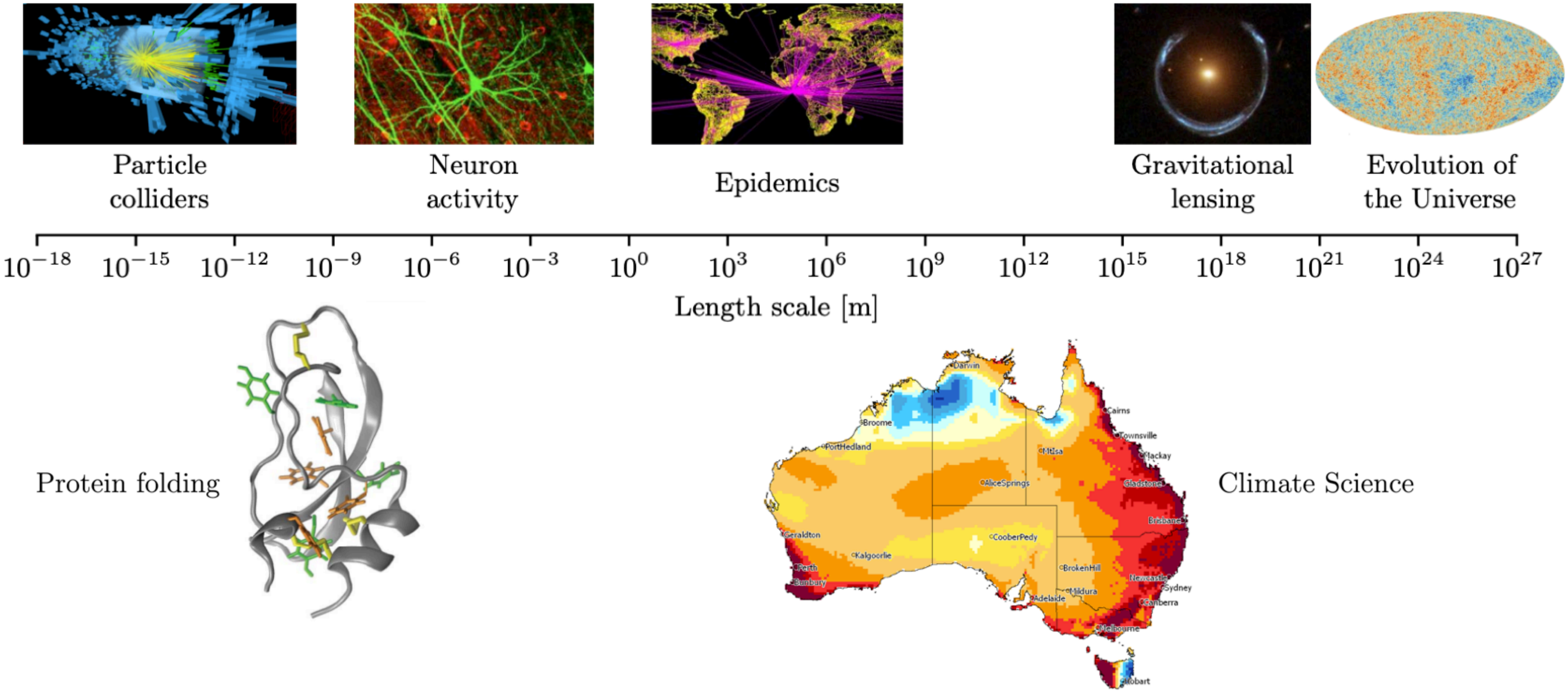


Image adapted from Cranmer K., Brehmer J., Louppe G., PNAS (2020)

Science relies heavily on high-fidelity simulators

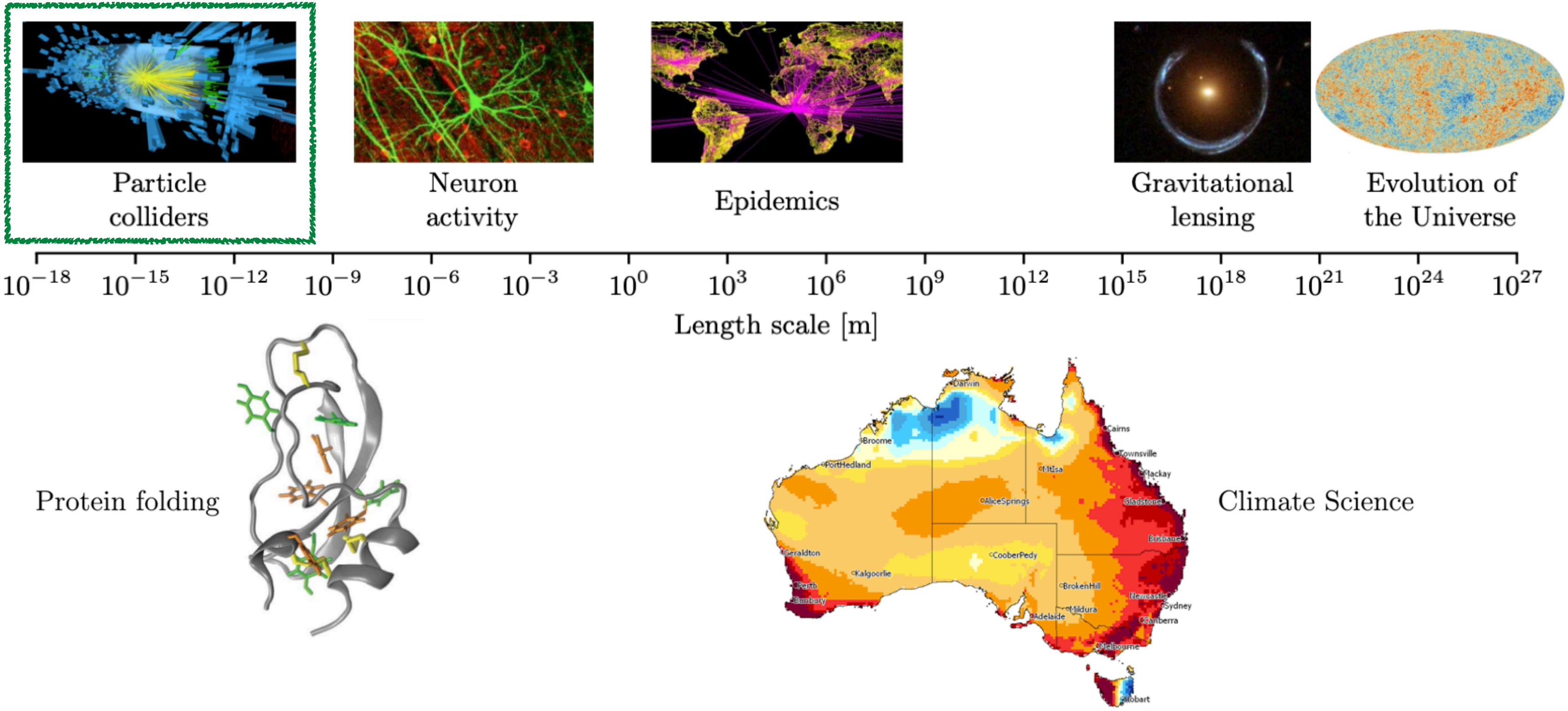
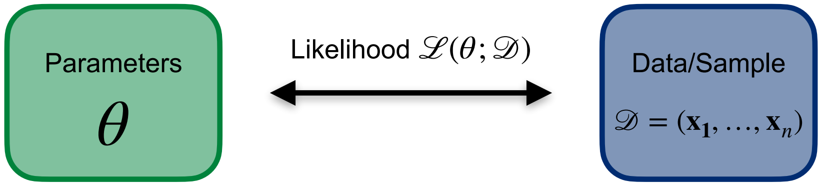
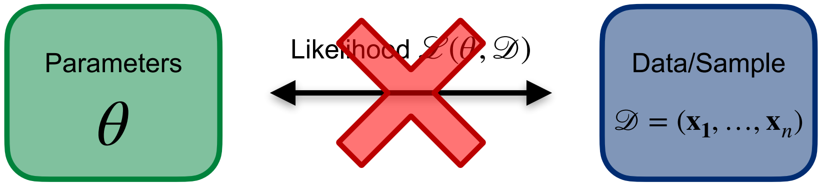


Image adapted from Cranmer K., Brehmer J., Louppe G., PNAS (2020)

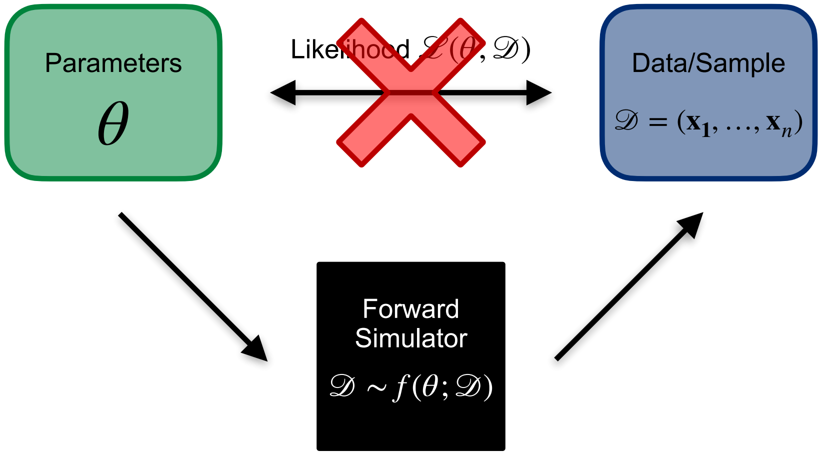
Likelihood-based Inference



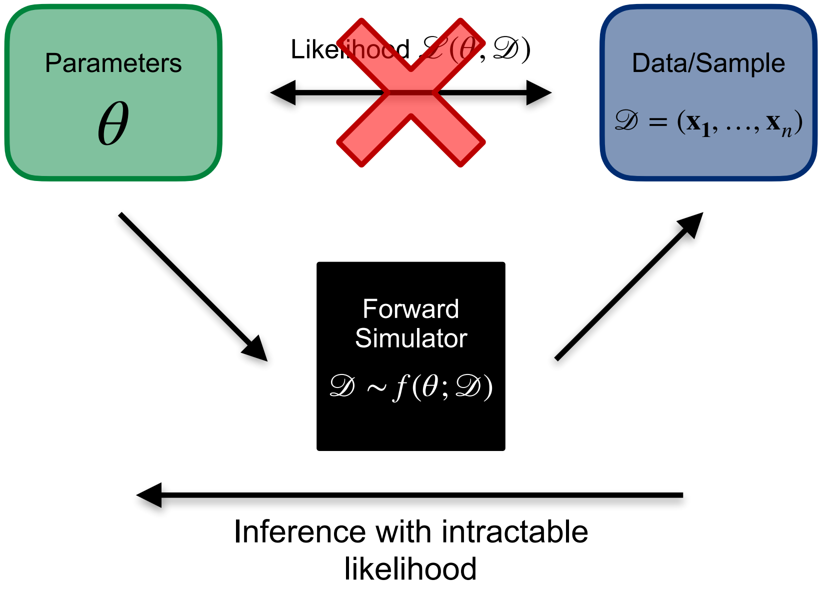
Likelihood-based Inference



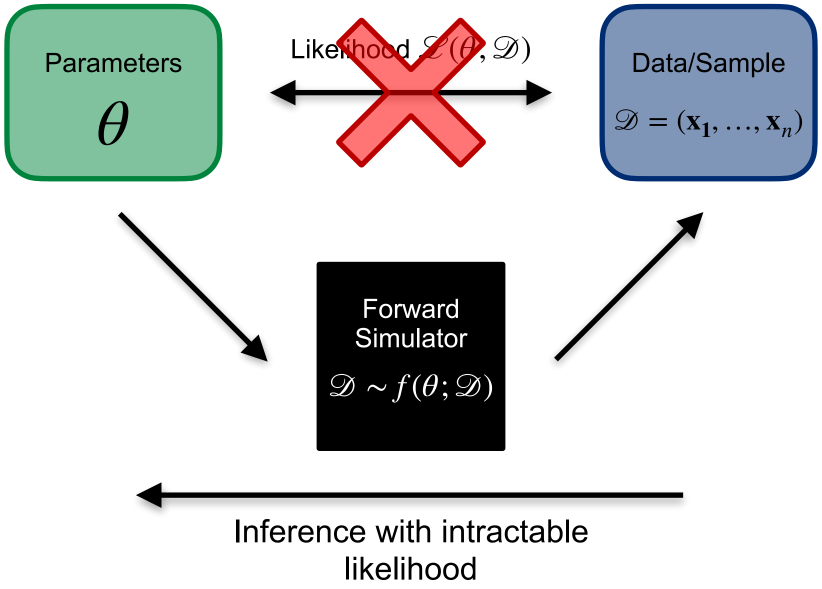
Likelihood-based Inference



Likelihood-based Inference



Likelihood-based Inference



→ **Likelihood-Free Inference (LFI)**

Inference via predictions and posteriors: bias and overconfidence

- **Recent advances in LFI[1].** Use ML algorithms and simulated data to directly estimate key inferential quantities:

use $\{(\theta_1, \mathscr{D}_1), \ldots, (\theta_B, \mathscr{D}_B)\}$, where $\theta \sim \pi_\theta,\ \mathscr{D} \sim F_\theta$ $\longrightarrow$ $\underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid \mathscr{D})}_{Posteriors}, \underbrace{\mathscr{L}(\theta; \mathscr{D})}_{Likelihoods}, \underbrace{\mathscr{L}(\theta_1; \mathscr{D}) / \mathscr{L}(\theta_2; \mathscr{D})}_{Likelihood\ ratios}$

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014) 2. Hermans et al. (2021)

Inference via predictions and posteriors: bias and overconfidence

- **Recent advances in LFI[1].** Use ML algorithms and simulated data to directly estimate key inferential quantities:

$$\text{use } \{(\theta_1, \mathscr{D}_1), \ldots, (\theta_B, \mathscr{D}_B)\}, \text{ where } \theta \sim \pi_\theta,\ \mathscr{D} \sim F_\theta \longrightarrow \underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid \mathscr{D})}_{Posteriors}, \underbrace{\mathscr{L}(\theta; \mathscr{D})}_{Likelihoods}, \underbrace{\mathscr{L}(\theta_1; \mathscr{D}) / \mathscr{L}(\theta_2; \mathscr{D})}_{Likelihood\ ratios}$$

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014) 2. Hermans et al. (2021)

Inference via predictions and posteriors: bias and overconfidence

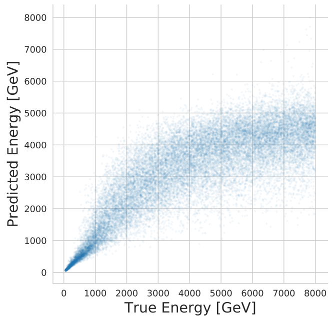
- **Recent advances in LFI[1].** Use ML algorithms and simulated data to directly estimate key inferential quantities:

use $\{(\theta_1, \mathcal{D}_1), \ldots, (\theta_B, \mathcal{D}_B)\}$, where $\theta \sim \pi_\theta$, $\mathcal{D} \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid \mathcal{D})}_{Posteriors}, \underbrace{\mathcal{L}(\theta; \mathcal{D})}_{Likelihoods}, \underbrace{\mathcal{L}(\theta_1; \mathcal{D}) / \mathcal{L}(\theta_2; \mathcal{D})}_{Likelihood\ ratios}$

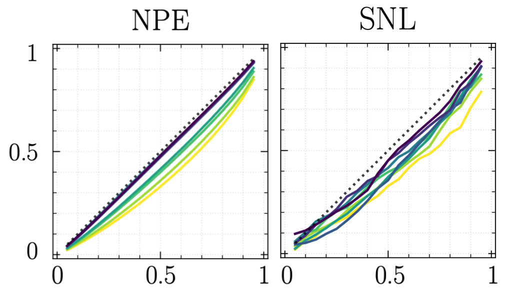
- **Do these methods give reliable measures of uncertainty around parameters of interest?**

Prediction algorithms are biased

$$\hat{\theta} \neq \theta$$



Posterior estimators are overconfident[2]



1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014) 2. Hermans et al. (2021)

Inference via predictions and posteriors: bias and overconfidence

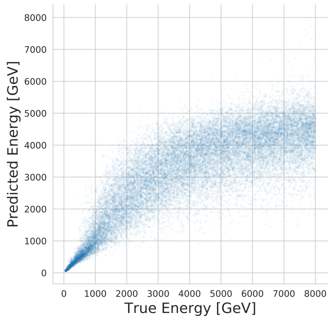
- **Recent advances in LFI[1].** Use ML algorithms and simulated data to directly estimate key inferential quantities:

use $\{(\theta_1, \mathscr{D}_1), \ldots, (\theta_B, \mathscr{D}_B)\}$, where $\theta \sim \pi_\theta$, $\mathscr{D} \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid \mathscr{D})}_{Posteriors}, \underbrace{\mathscr{L}(\theta; \mathscr{D})}_{Likelihoods}, \underbrace{\mathscr{L}(\theta_1; \mathscr{D})/\mathscr{L}(\theta_2; \mathscr{D})}_{Likelihood\ ratios}$

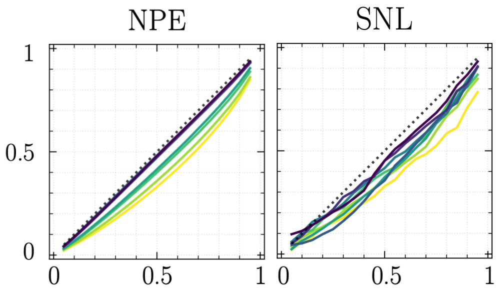
- **Do these methods give reliable measures of uncertainty around parameters of interest?**

Prediction algorithms are biased

$\hat{\theta} \neq \theta$



Posterior estimators are overconfident[2]



Problem: both approaches rely on $\theta \sim \pi_\theta$, which introduces a bias that might or might not be consistent with the data

→ Hinders the reliability of scientific conclusions

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014) 2. Hermans et al. (2021)

Constraining parameters while guaranteeing coverage

- **Reliable inference should achieve confidence sets whose coverage guarantees are independent of**

    1. the choice of the prior π_θ, so that good priors lead to tighter constraints, but bad priors do not degrade coverage;

    2. the specific value of θ : coverage guarantees should hold everywhere, not in expectation;

    3. the size of the observed sample: no asymptotics

Constraining parameters while guaranteeing coverage

- **Reliable inference should achieve confidence sets whose coverage guarantees are independent of**

  1. the choice of the prior π_θ, so that good priors lead to tighter constraints, but bad priors do not degrade coverage;
  2. the specific value of θ : coverage guarantees should hold everywhere, not in expectation;
  3. the size of the observed sample: no asymptotics

- **How?**

  1. Leverage predictions and posteriors and use Neyman inversion to achieve correct conditional coverage

$$\mathbb{P}(\theta \in \mathscr{R}(D) \,|\, \theta) = 1 - \alpha \quad \forall \theta \in \Theta$$

  2. Independent diagnostics: check actual coverage across the whole Θ, without costly Monte-Carlo simulations

Neyman construction of confidence sets

- **Ingredients:**

1. Data $\mathscr{D} \sim F_\theta$
2. Test statistic $\tau(\mathscr{D};\theta)$
3. Critical values $C_{\theta,\alpha}$

Theorem (Neyman 1937)

Constructing a $1-\alpha$ *confidence set for* θ *is equivalent to testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $\mathscr{D} \sim F_\theta$
  2. Test statistic $\tau(\mathscr{D};\theta)$
  3. Critical values $C_{\theta,\alpha}$

Theorem (Neyman 1937)

Constructing a $1-\alpha$ *confidence set for* θ *is equivalent to testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

i. Rejection region for $\tau(\mathscr{D};\theta),\ \forall\theta \in \Theta$

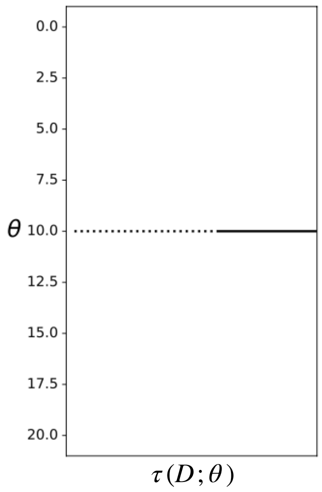


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

1. Data $\mathscr{D} \sim F_\theta$
2. Test statistic $\tau(\mathscr{D};\theta)$
3. Critical values $C_{\theta,\alpha}$

Theorem (Neyman 1937)

Constructing a $1-\alpha$ *confidence set for* θ *is equivalent to testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

i. Rejection region for $\tau(\mathscr{D};\theta),\ \forall\theta \in \Theta$

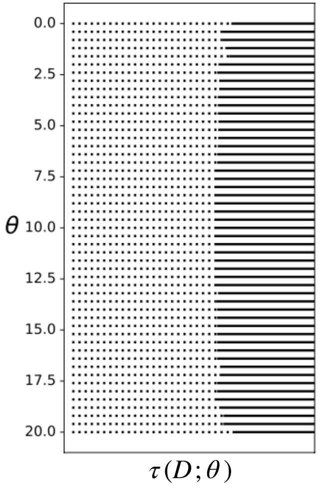


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $\mathscr{D} \sim F_\theta$
  2. Test statistic $\tau(\mathscr{D};\theta)$
  3. Critical values $C_{\theta,\alpha}$

Theorem (Neyman 1937)

Constructing a $1-\alpha$ confidence set for θ is equivalent to testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

i. Rejection region for $\tau(\mathscr{D};\theta),\ \forall\theta \in \Theta$

ii. $\tau(D;\theta),\ \forall\theta \in \Theta$

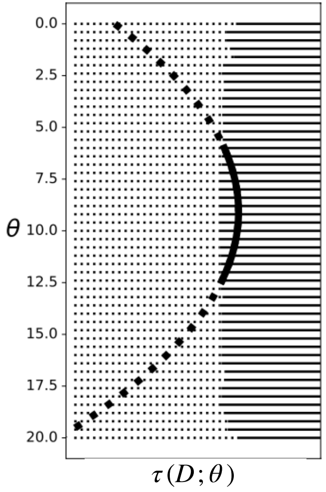


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $\mathscr{D} \sim F_\theta$
  2. Test statistic $\tau(\mathscr{D};\theta)$
  3. Critical values $C_{\theta,\alpha}$

Theorem (Neyman 1937)

Constructing a $1-\alpha$ *confidence set for* θ *is equivalent to testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

for every $\theta_0 \in \Theta$.

i. Rejection region for $\tau(\mathscr{D};\theta),\ \forall\theta \in \Theta$

ii. $\tau(D;\theta),\ \forall\theta \in \Theta$

iii. $(1-\alpha)$ confidence set for θ

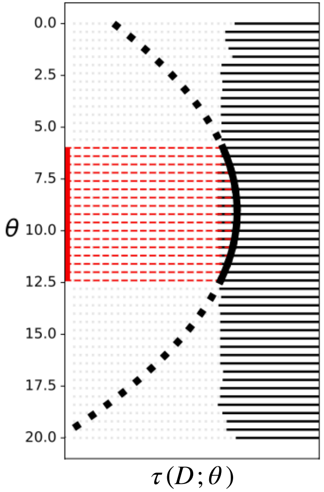


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

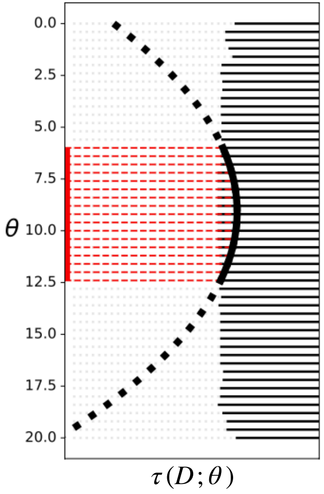
  1. Data $\mathscr{D} \sim F_\theta$
  2. Test statistic $\tau(\mathscr{D};\theta)$
  3. Critical values $C_{\theta,\alpha}$

> Theorem (Neyman 1937)
>
> *Constructing a $1-\alpha$ confidence set for θ is equivalent to testing*
>
> $$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$
>
> *for every $\theta_0 \in \Theta$.*

i. Rejection region for $\tau(\mathscr{D};\theta), \ \forall\theta \in \Theta$

ii. $\tau(D;\theta), \ \forall\theta \in \Theta$

iii. $(1-\alpha)$ confidence set for θ



- **Wald** test statistic (1D case):

$$\tau^{Wald}(\mathscr{D};\theta_0) := \frac{(\theta^{MLE}-\theta_0)^2}{\mathbb{V}[\theta^{MLE}]}$$

Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

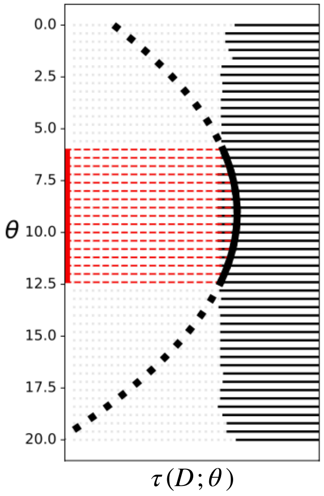
  1. Data $\mathscr{D} \sim F_\theta$
  2. Test statistic $\tau(\mathscr{D};\theta)$
  3. Critical values $C_{\theta,\alpha}$

> **Theorem (Neyman 1937)**
>
> *Constructing a $1-\alpha$ confidence set for θ is equivalent to testing*
>
> $$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$
>
> *for every $\theta_0 \in \Theta$.*

i. Rejection region for $\tau(\mathscr{D};\theta),\ \forall\theta \in \Theta$

ii. $\tau(D;\theta),\ \forall\theta \in \Theta$

iii. $(1-\alpha)$ confidence set for θ



- **Wald** test statistic (1D case):

$$\tau^{Wald}(\mathscr{D};\theta_0) := \frac{(\theta^{MLE} - \theta_0)^2}{\mathbb{V}[\theta^{MLE}]}$$

→

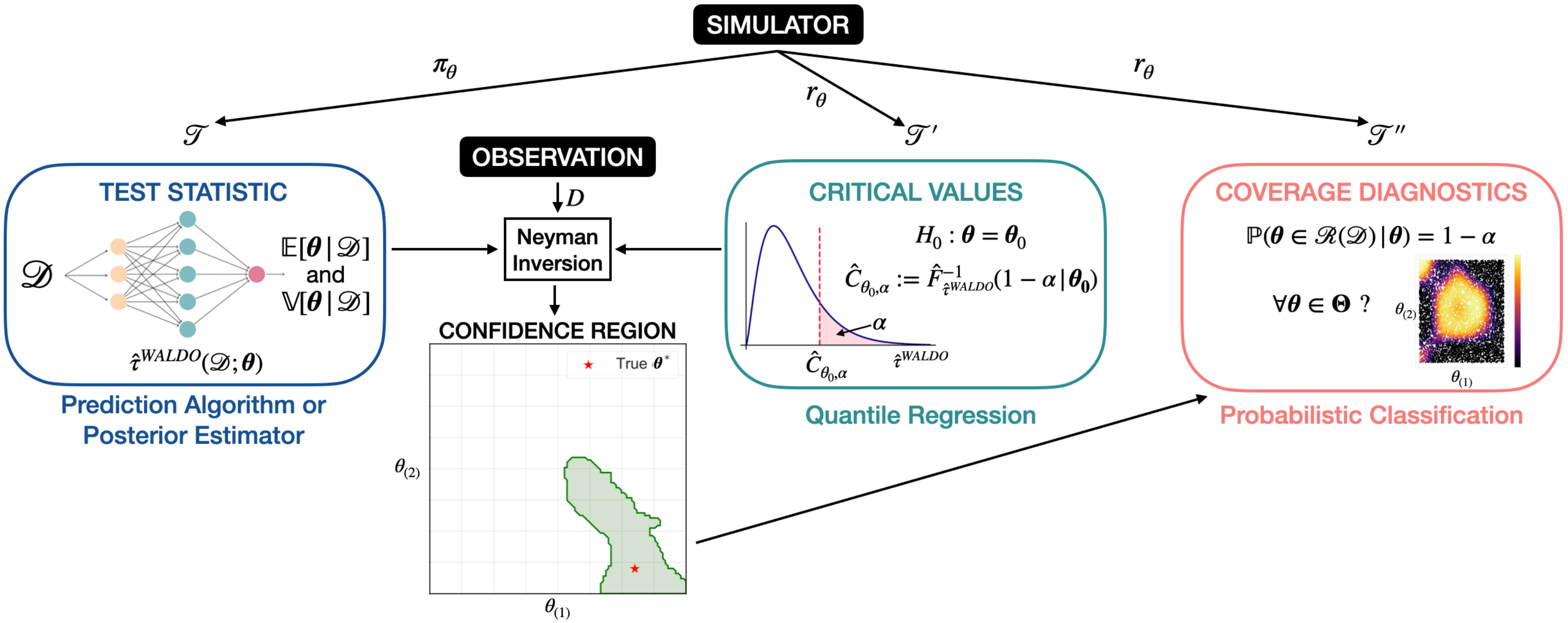
- **Waldo** test statistic (1D and p-D case):

$$\tau^{Waldo}(\mathscr{D};\theta_0) := \frac{(\mathbb{E}[\theta \mid \mathscr{D}] - \theta_0)^2}{\mathbb{V}[\theta \mid \mathscr{D}]}$$

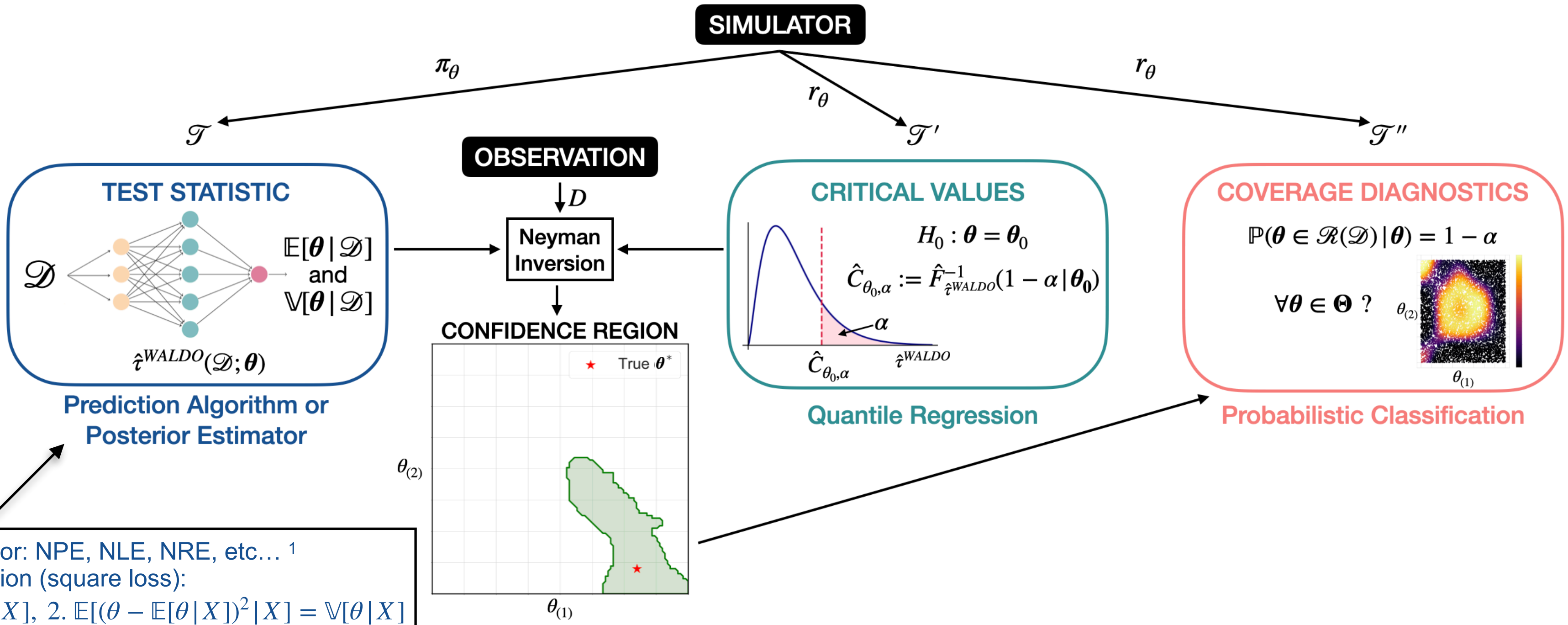
$$\tau^{Waldo}(\mathscr{D};\theta_0) := (\mathbb{E}[\theta \mid \mathscr{D}] - \theta_0)^T \mathbb{V}[\theta \mid \mathscr{D}]^{-1} (\mathbb{E}[\theta \mid \mathscr{D}] - \theta_0)$$

Image adapted from Dalmasso et al. (2020)

Waldo


SIMULATOR
π_θ
r_θ
r_θ
$\mathscr{T}$
$\mathscr{T}'$
$\mathscr{T}''$
OBSERVATION
D
TEST STATISTIC
$\mathscr{D}$
$\mathbb{E}[\boldsymbol{\theta} \mid \mathscr{D}]$
and
$\mathbb{V}[\boldsymbol{\theta} \mid \mathscr{D}]$
$\hat{\tau}^{WALDO}(\mathscr{D}; \boldsymbol{\theta})$
Prediction Algorithm or Posterior Estimator
Neyman Inversion
CONFIDENCE REGION
True $\boldsymbol{\theta}^*$
$\theta_{(2)}$
$\theta_{(1)}$
CRITICAL VALUES
$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$
$\hat{C}_{\theta_0,\alpha} := \hat{F}^{-1}_{\hat{\tau}^{WALDO}}(1 - \alpha \mid \boldsymbol{\theta}_0)$
α
$\hat{C}_{\theta_0,\alpha}$
$\hat{\tau}^{WALDO}$
Quantile Regression
COVERAGE DIAGNOSTICS
$\mathbb{P}(\boldsymbol{\theta} \in \mathscr{R}(\mathscr{D}) \mid \boldsymbol{\theta}) = 1 - \alpha$
$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$?
$\theta_{(2)}$
$\theta_{(1)}$
Probabilistic Classification

Waldo



Posterior: NPE, NLE, NRE, etc… [1]
Prediction (square loss):
$1.\ \mathbb{E}[\theta\,|\,X],\ 2.\ \mathbb{E}[(\theta-\mathbb{E}[\theta\,|\,X])^2\,|\,X] = \mathbb{V}[\theta\,|\,X]$

1. Neural Posterior Estimator, Neural Likelihood Estimator, Neural Ratio Estimator.

An example leveraging posterior estimators

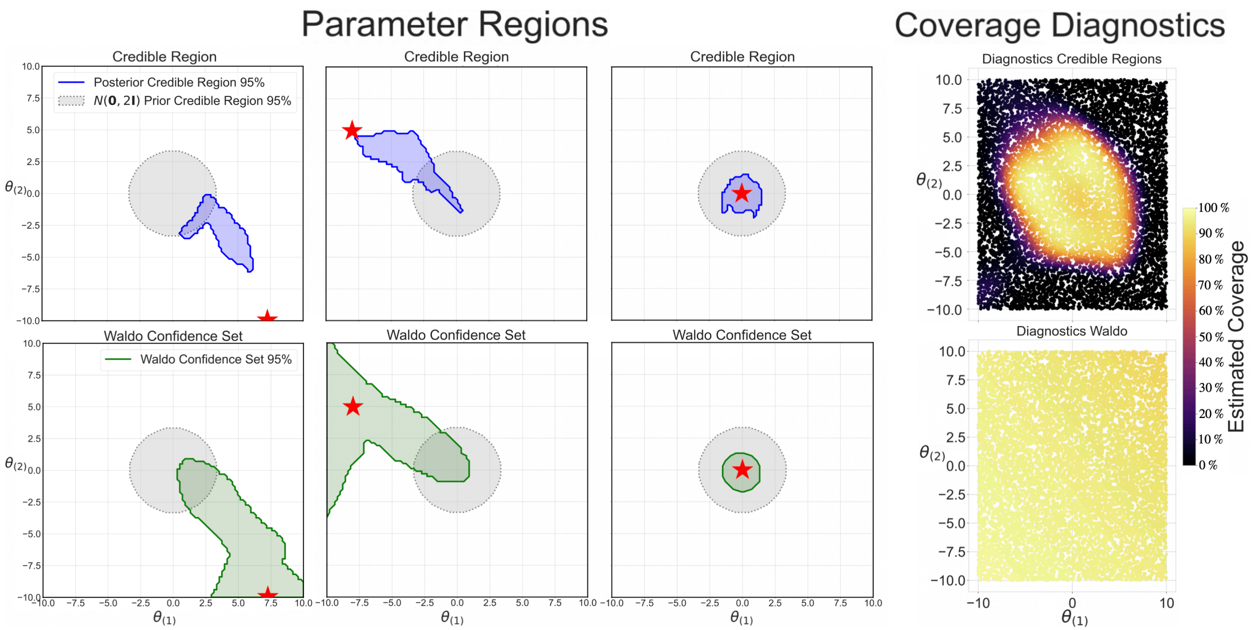
- **Synthetic example:** estimate the common mean of the components of a Gaussian mixture

$$\mathcal{D} \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, \underline{\mathbf{I}}) + \frac{1}{2}\mathcal{N}(\theta, \underline{0.01\mathbf{I}}), \quad \theta \in \mathbb{R}^2, \ n = 1$$

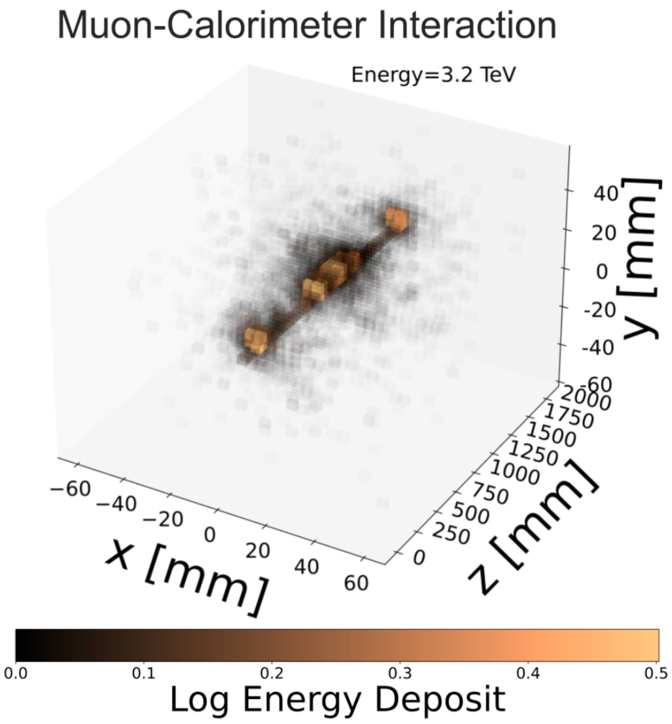
An example leveraging posterior estimators

- **Synthetic example:** estimate the common mean of the components of a Gaussian mixture

$$\mathscr{D} \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, \underline{\mathbf{I}}) + \frac{1}{2}\mathcal{N}(\theta, \underline{0.01\mathbf{I}}), \quad \theta \in \mathbb{R}^2, \ n = 1$$

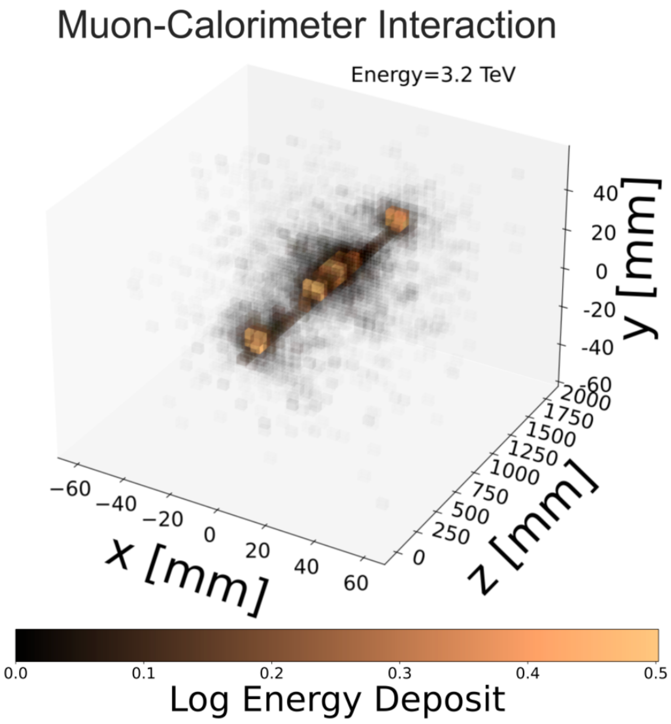


Inference on Muon energies using CNN predictions



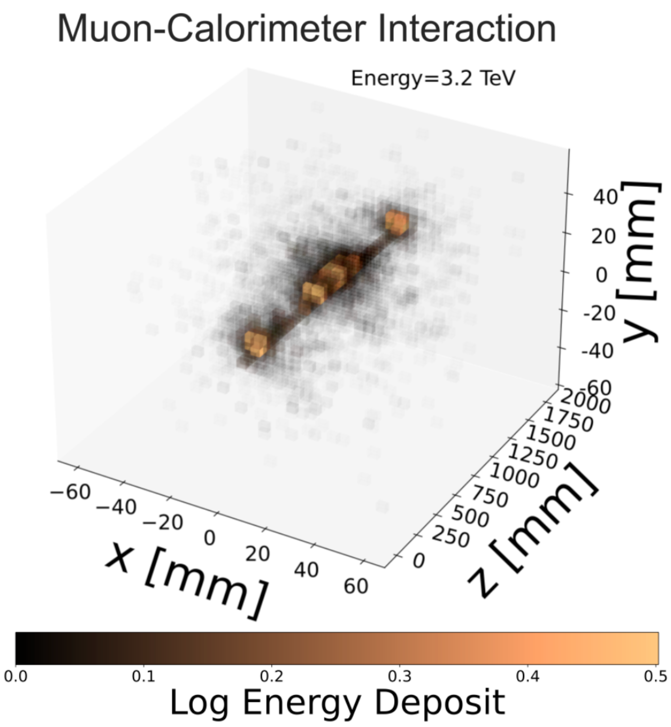
- **Goal:** alternative to traditional way of measuring muons

Inference on Muon energies using CNN predictions



- **Goal:** alternative to traditional way of measuring muons
- Data obtained from Geant4[1] with incoming energy between 50 GeV and 8000 GeV

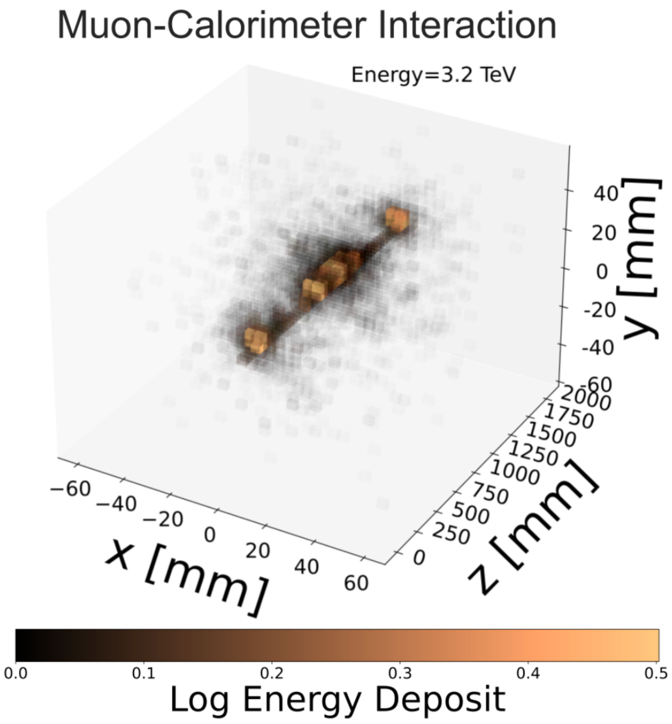
Inference on Muon energies using CNN predictions



- **Goal:** alternative to traditional way of measuring muons
- Data obtained from Geant4[1] with incoming energy between 50 GeV and 8000 GeV
- finely segmented calorimeter with 50 layers in z, each divided in a 32×32 grid $\rightarrow$ $51{,}200$ cells

1. Agostinelli et al. (2003); 2. From Kieseler et al. (2022)

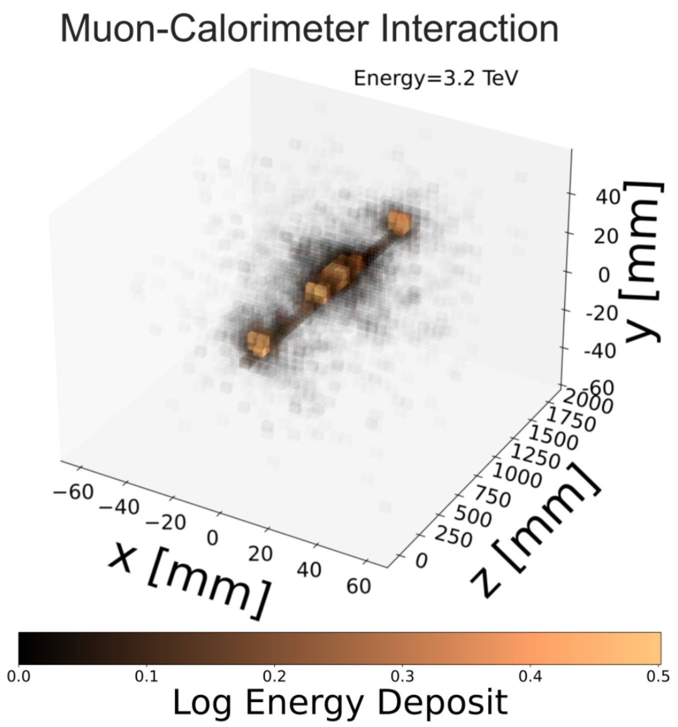
Inference on Muon energies using CNN predictions



- **Goal:** alternative to traditional way of measuring muons
- Data obtained from Geant4[1] with incoming energy between 50 GeV and 8000 GeV
- finely segmented calorimeter with 50 layers in z, each divided in a 32×32 grid → $51{,}200$ cells
- 28 features[2] extracted from the spatial and energy information of the calorimeters cells. Three main groups:
    1. general properties of the energy deposition (e.g. sum of energy above/below a threshold)
    2. more fine-grained information (e.g. moments of the energy distributions in different regions over z)
    3. custom procedure that isolates clusters of deposited energy along the track

1. Agostinelli et al. (2003); 2. From Kieseler et al. (2022)

Inference on Muon energies using CNN predictions



- **Goal:** alternative to traditional way of measuring muons
- Data obtained from Geant4[1] with incoming energy between 50 GeV and 8000 GeV
- finely segmented calorimeter with 50 layers in z, each divided in a 32×32 grid $\rightarrow$ $51{,}200$ cells
- 28 features[2] extracted from the spatial and energy information of the calorimeters cells. Three main groups:
    1. general properties of the energy deposition (e.g. sum of energy above/below a threshold)
    2. more fine-grained information (e.g. moments of the energy distributions in different regions over z)
    3. custom procedure that isolates clusters of deposited energy along the track
- sum energy deposits over 0.1 GeV to get one-dimensional energy-sum data

1. Agostinelli et al. (2003); 2. From Kieseler et al. (2022)

Can we do frequentist inference for muon energy?

We are mainly interested in **two questions**:

1. Infer, from the pattern of the energy deposits in the calorimeter, how much energy the incoming muon had *and* construct a **confidence set for it with proper coverage**

   → **goal**: Reconstruct muon properties with rigorous uncertainties for downstream analyses

Can we do frequentist inference for muon energy?

We are mainly interested in **two questions**:

1. Infer, from the pattern of the energy deposits in the calorimeter, how much energy the incoming muon had *and* construct a **confidence set for it with proper coverage**

   → **goal**: Reconstruct muon properties with rigorous uncertainties for downstream analyses

2. How much added value does a **high granularity of the calorimeter** cells offer over the 1D and 28D representations?

   → **goal**: devise better and more cost-effective calorimeters for future particle colliders

Prediction algorithms used

Three "nested" datasets:

1. One-dimensional energy sum: best predictor wrt Cross-Validation MSE loss (XGBoost)
2. 27 features + 1D energy sum: best predictor wrt Cross-Validation MSE loss (XGBoost)
3. Full calorimeter (51200-D) + 28 features: custom CNN (with MSE loss) from Kieseler et al. (2022)

→ We estimate $\mathbb{E}[\theta \,|\, \mathscr{D}]$ and $\mathbb{V}[\theta \,|\, \mathscr{D}]$ for each of these. Muon energy is θ

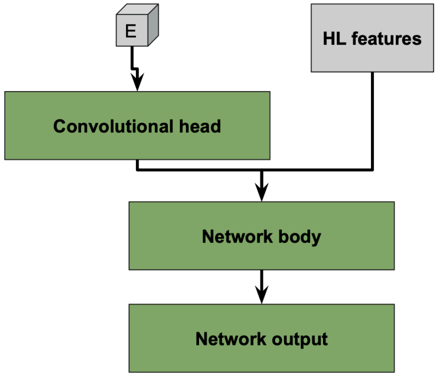
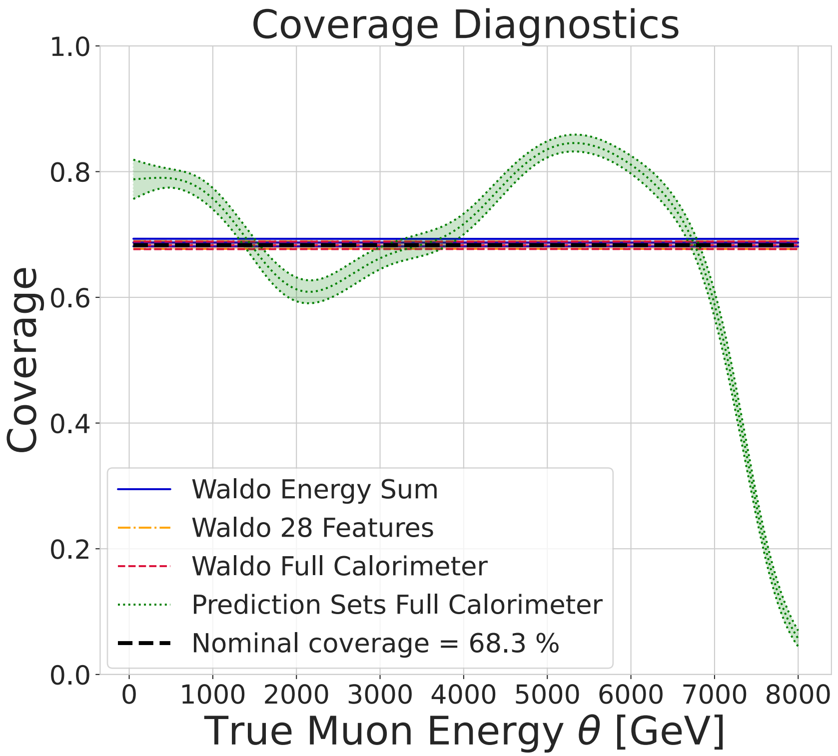


Image credit: Kieseler et al. (2022)

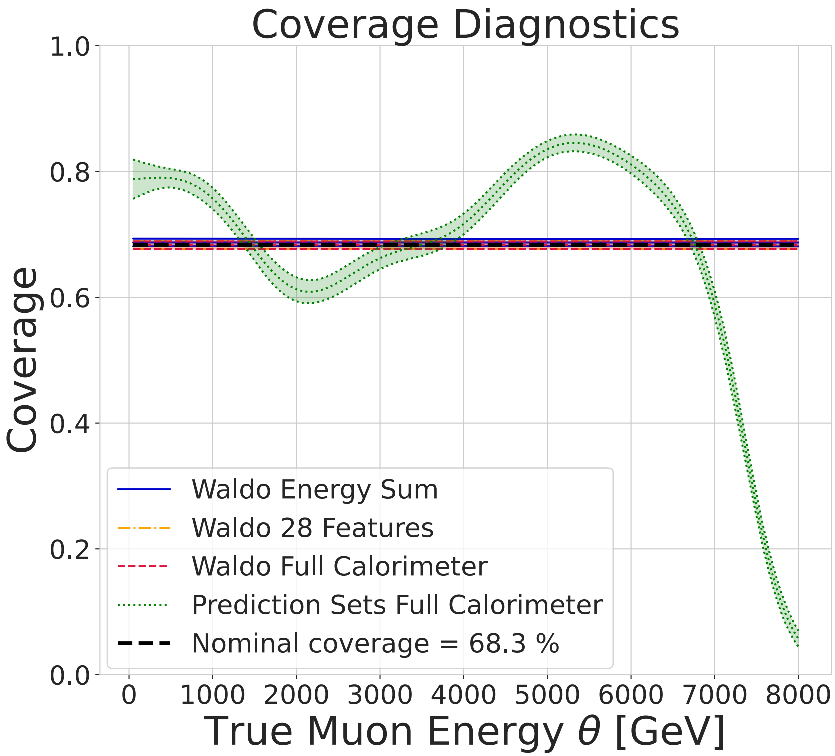
Confidence sets for muon energy have proper coverage

- Nominal coverage is achieved regardless of the dataset used
- Prediction sets ($\mathbb{E}[\theta \mid x] \pm \sqrt{\mathbb{V}[\theta \mid x]}$) do not achieve the desired level of coverage

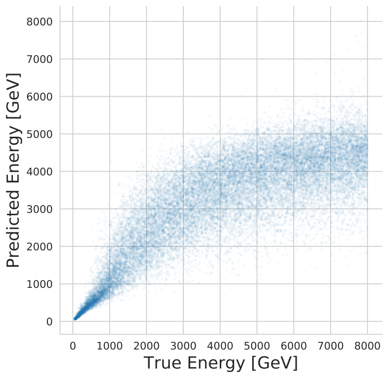


Confidence sets for muon energy have proper coverage

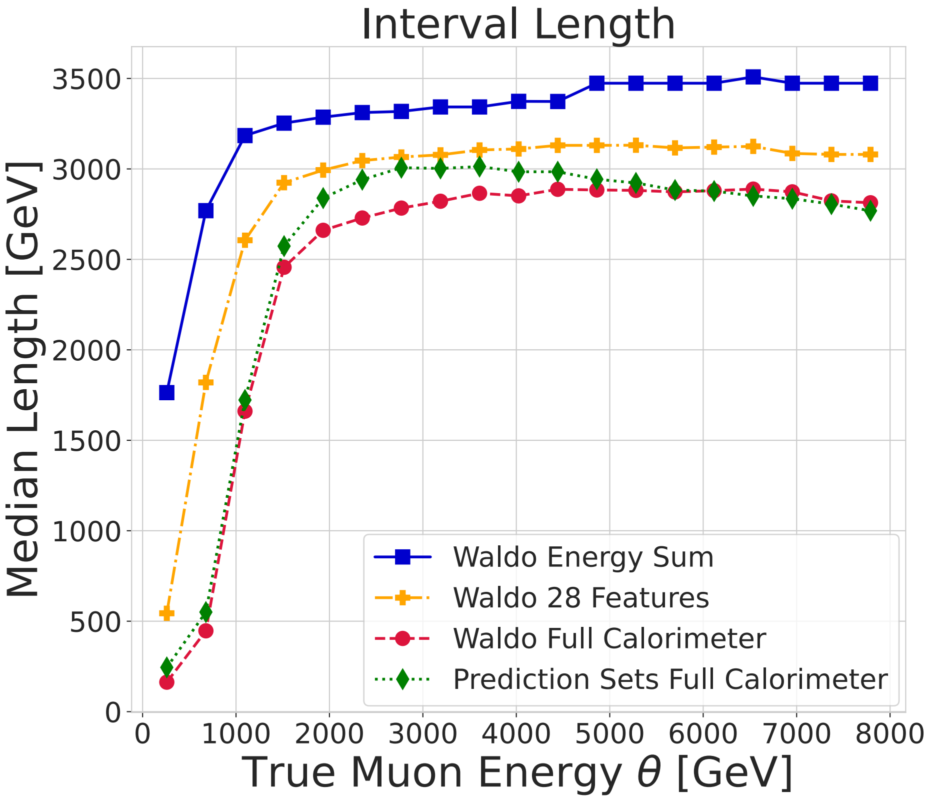
- Nominal coverage is achieved regardless of the dataset used
- Prediction sets ($\mathbb{E}[\theta \mid x] \pm \sqrt{\mathbb{V}[\theta \mid x]}$) do not achieve the desired level of coverage



Why?



Valuable information in high-granularity calorimeter



- Intervals are shorter as the data becomes higher-dimensional
- Prediction sets can even be larger than Waldo confidence sets (while also not guaranteeing coverage)

Summary

- WALDO, a method to construct confidence regions with correct conditional coverage for parameters in *inverse* problems by leveraging any prediction algorithm or posterior estimator

- WALDO disentangles the coverage guarantees of the confidence region from the choice of the prior distribution. To increase power, one may be able to leverage domain-specific knowledge, take advantage of the internal structure of the simulator, or exploit active learning strategies

- We demonstrated its effectiveness estimating the energy of muons at a future particle collider. Calorimeter data represents a viable alternative for the measurement of muons of very high energy

Useful Links:

ArXiv:

- WALDO (under review): https://arxiv.org/abs/2205.15680
- LF2I (under review): https://arxiv.org/abs/2107.03920

Code

- https://github.com/lee-group-cmu/lf2i ← We are looking for interested users to gather feedback on the package!

Thanks!

Carnegie Mellon University

Bias and coverage of prediction intervals

- Train on $(\mathscr{D}_1, \theta_1), \ldots, (\mathscr{D}_B, \theta_B) \sim f(\mathscr{D}, \theta)$ and output $\hat{\theta} = \hat{\mathbb{E}}[\theta \mid \mathscr{D}]$
  - → posterior mean, which depends on **marginal** since $f(\mathscr{D}, \theta) = f(\mathscr{D} \mid \theta) f(\theta)$

- What about coverage of standard prediction intervals? Construct a $1-\alpha$ interval of the form $\hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}$
  - → Coverage is a strictly decreasing function of $|\text{bias}(\hat{\theta})| = |\mathbb{E}[\hat{\theta}] - \theta|$
  - → Prediction intervals over-cover when $\text{bias}(\hat{\theta}) = 0$ and under-cover for large bias values

- Simple univariate Gaussian example:

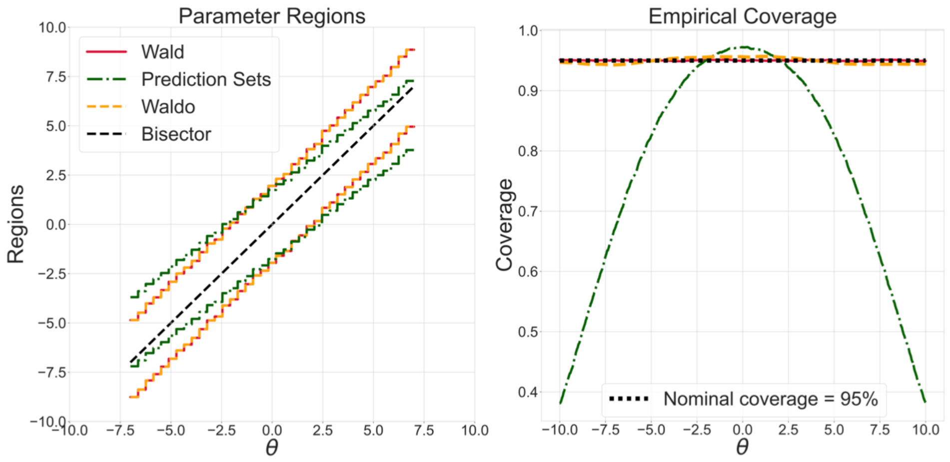
$$\theta \sim \mathcal{N}(\mu = 0, \sigma = 2)$$
$$\mathscr{D} \mid \theta \sim \mathcal{N}(\theta, \sigma = 1)$$

Construct confidence sets via

- Wald test
- Waldo

and

- Prediction sets



Left: means of upper and lower bounds of interval estimates for 100,000 observations divided in 38 bins over the true parameter.

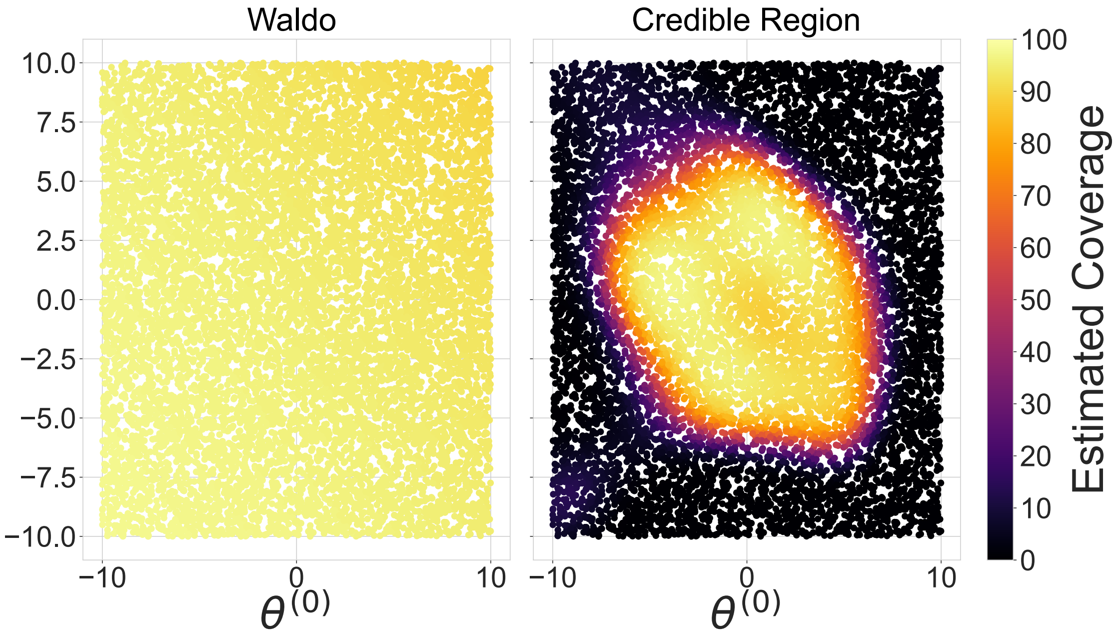
Right: empirical coverage of the intervals on the left as a function of the true parameter.

Statistical Properties (coverage diagnostics)

- **Synthetic example:** estimate the common mean of the components of a Gaussian mixture

$$\mathcal{D} \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, \mathbf{I}) + \frac{1}{2}\mathcal{N}(\theta,\ 0.01\mathbf{I}), \quad \theta \in \mathbb{R}^2$$

Prior: $N(\mathbf{0}, 2\mathbf{I})$



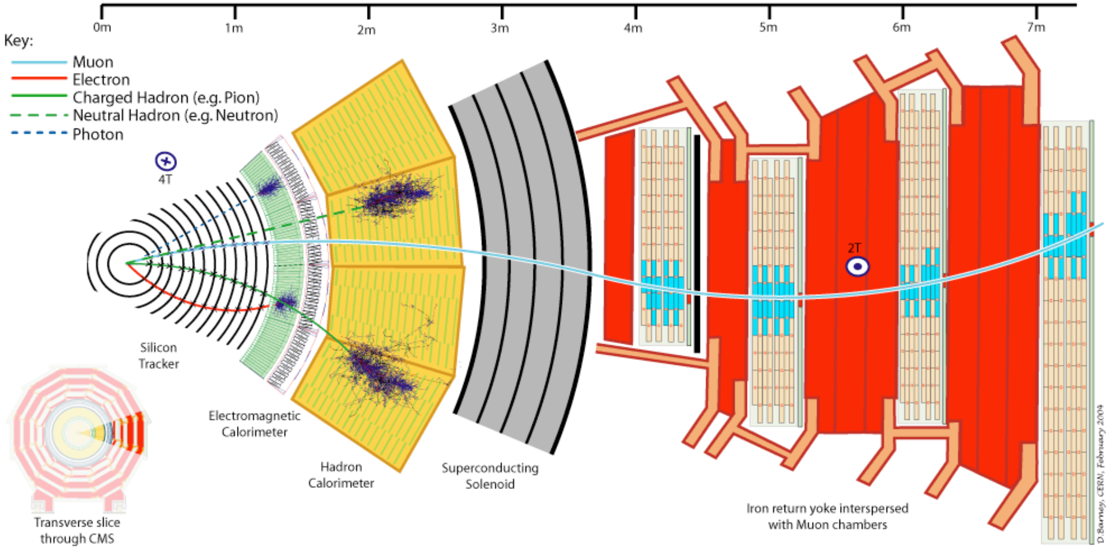
Inference for calorimetric muon energy measurements

Muons are one of the elementary particles described by the **Standard Model**.

Their importance is mainly due to two facts: **first**, they emerge as a signature in processes which could signal the existence of new physics, and **second**, they are (relatively) easy to identify.

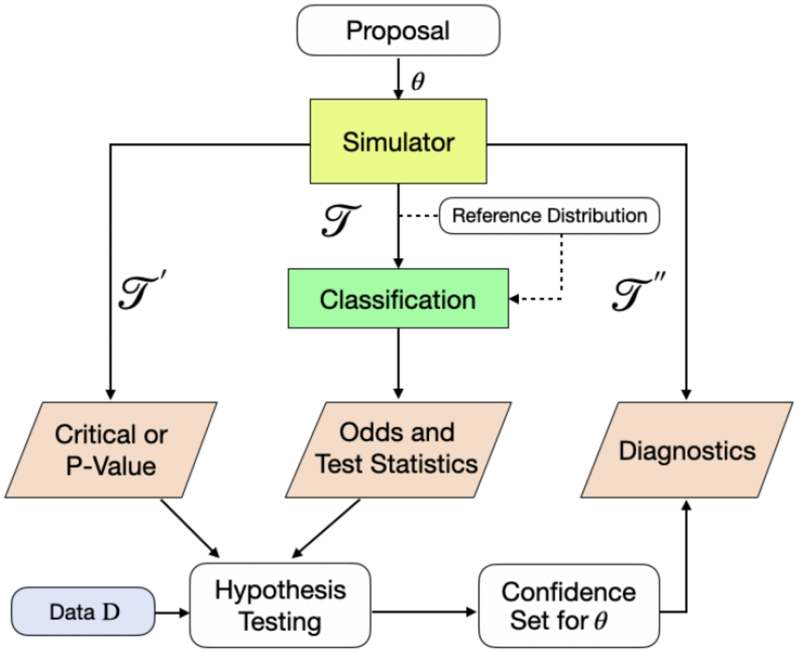
Above: Aerial view of the position of the LHC.

Right: transverse slice of CMS, one of the particle detectors at the LHC in Geneva.



Likelihood-free Frequentist Inference (LF2I)

https://arxiv.org/pdf/2107.03920.pdf



A modular framework:

1. **central branch:** parameterized odds

$$\mathbb{O}(X;\theta) := \frac{\mathbb{P}(Y=1\,|\,\theta,\mathbf{x})}{\mathbb{P}(Y=0\,|\,\theta,\mathbf{x})}$$

used to construct test statistics $\tau(\mathcal{D};\theta_0)$

2. **left branch:** quantile regression to estimate critical values C_{θ_0} for $\tau(\mathcal{D};\theta_0)$ for hypothesis tests

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0, \quad \forall \theta \in \Theta$$

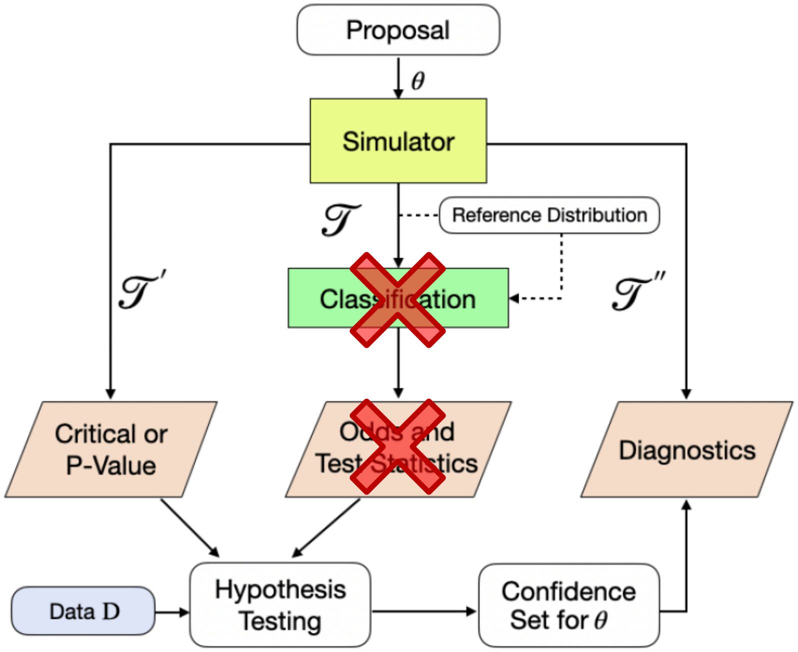
→ **(1 + 2) use Neyman inversion:**

$$\left\{\theta_0 \in \Theta \,|\, \hat{\tau}(\mathcal{D} = D;\theta_0) \textbf{ in acceptance region}\right\}$$

3. **right branch:** assess empirical coverage across Θ by regressing $\mathbb{I}\{\theta \in \mathcal{C}(\mathcal{D})\,|\,\theta\}$ against θ

Image credit: Dalmasso et al. (2021)

Likelihood-free Frequentist Inference (LF2I)

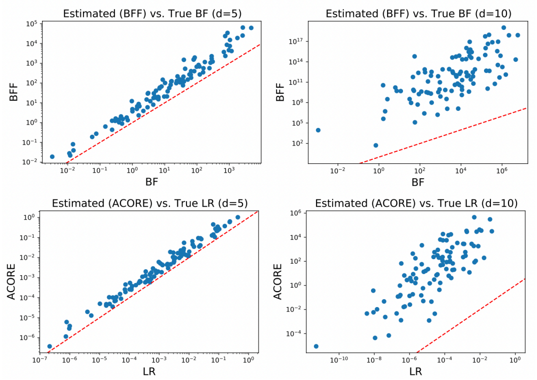


1. Image adapted from Dalmasso et al. (2021)

- Left branch guarantees coverage provided that the quantile regressor is well estimated

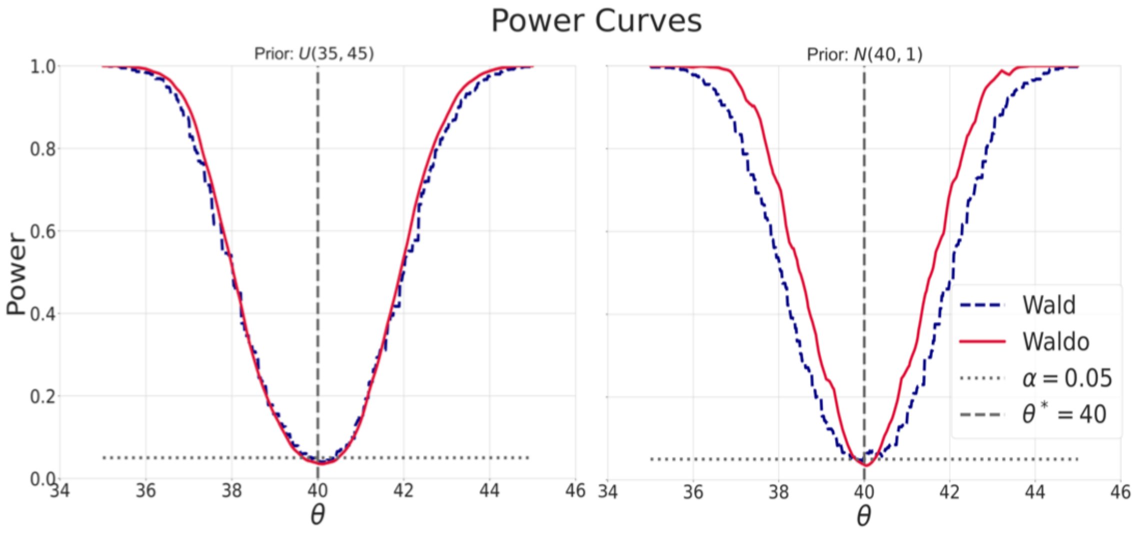
- Computing the test statistics involves optimization/ integration procedures that negatively affect the power of the resulting test;

$$\mathrm{LR}(\mathcal{D};\Theta_0) = \log \frac{\sup_{\theta\in\Theta_0}\mathcal{L}(\mathcal{D};\theta)}{\sup_{\theta\in\Theta}\mathcal{L}(\mathcal{D};\theta)} \longrightarrow \Lambda(\mathcal{D};\Theta_0) := \log \frac{\sup_{\theta_0\in\Theta_0}\prod_{i=1}^{n}\mathbb{O}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\sup_{\theta\in\Theta}\prod_{i=1}^{n}\mathbb{O}(\mathbf{X}_i^{\text{obs}};\theta)}$$



Combining frequentist coverage with prior knowledge

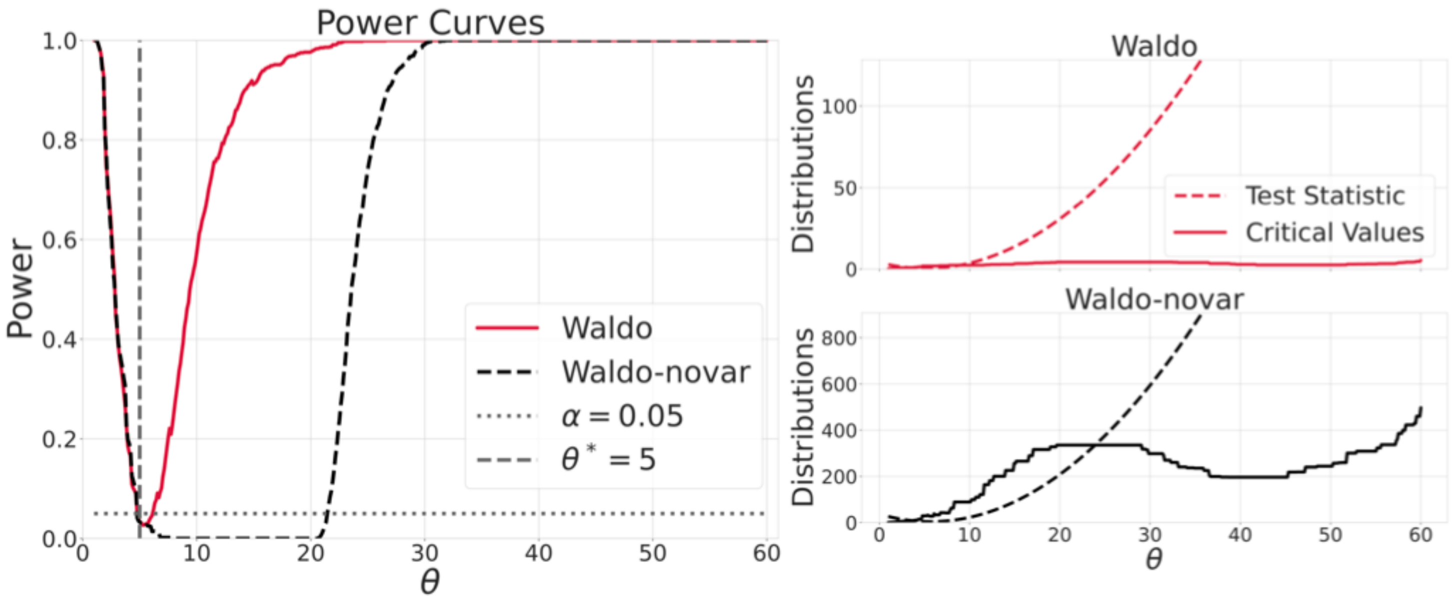
$\mathscr{D} \mid \theta \sim \mathscr{N}(\theta, \ 1);$ LEFT: $\theta \sim \mathscr{U}(35,45)$, RIGHT: $\theta \sim \mathscr{N}(40, \ 1)$



Is it useful to divide by $\mathbb{V}[\theta \mid X]$?

- **Waldo** requires to estimate $\mathbb{V}[\theta \mid \mathscr{D}]$. Why not simply use $\tau^{Waldo-novar}(\mathscr{D};\theta) := (\mathbb{E}[\theta \mid \mathscr{D}] - \theta)^T(\mathbb{E}[\theta \mid \mathscr{D}] - \theta)$?
- Reject H_0 if $\mathscr{D} \in Rej$. Let $\mathscr{P}^{Waldo} = \mathbb{P}_\theta[\mathscr{D} \in Rej]$ be the **power function** of the Waldo test statistics

  → setting: inference on the shape of a **Pareto** likelihood $\mathscr{D} \sim Pareto(\theta, x_{min} = 1), \quad \theta \sim \mathscr{U}(0{,}60)$



Coverage guarantees

Assumption 1 (Uniform consistency) *Let $F(\cdot|\boldsymbol{\theta})$ be the cumulative distribution function of the test statistic $\tau(\mathcal{D};\boldsymbol{\theta}_0)$ conditional on $\boldsymbol{\theta}$, where $\mathcal{D} \sim F_{\boldsymbol{\theta}}$. Let $\widehat{F}_{B'}(\cdot|\boldsymbol{\theta})$ be the estimated conditional distribution function, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_{\boldsymbol{\theta}}$. Assume that the quantile regression estimator is such that*

$$\sup_{\tau \in \mathbb{R}} |\widehat{F}_{B'}(\tau|\boldsymbol{\theta}_0) - F(\tau|\boldsymbol{\theta}_0)| \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} 0.$$

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a strictly continuous statistic $\tau(\mathcal{D};\boldsymbol{\theta}_0)$ chosen according to step **(ii)** for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 1, then,*

$$\mathbb{P}_{\mathcal{D}|\boldsymbol{\theta}_0, C_{B'}}(\tau(\mathcal{D};\boldsymbol{\theta}_0) \geq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\boldsymbol{\theta}_0, C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\boldsymbol{\theta}_0}$ and conditional on the random variable $C_{B'}$.

From Dalmasso, Masserano, Zhao, Izbicki, Lee (2021).

Coverage Diagnostics Gaussian Mixture, $\pi_\theta \equiv \mathcal{U}([-10,10]^2)$