Likelihood-Free Frequentist Inference for Calorimetric Muon Energy Measurement

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

Science relies heavily on high-fidelity simulators

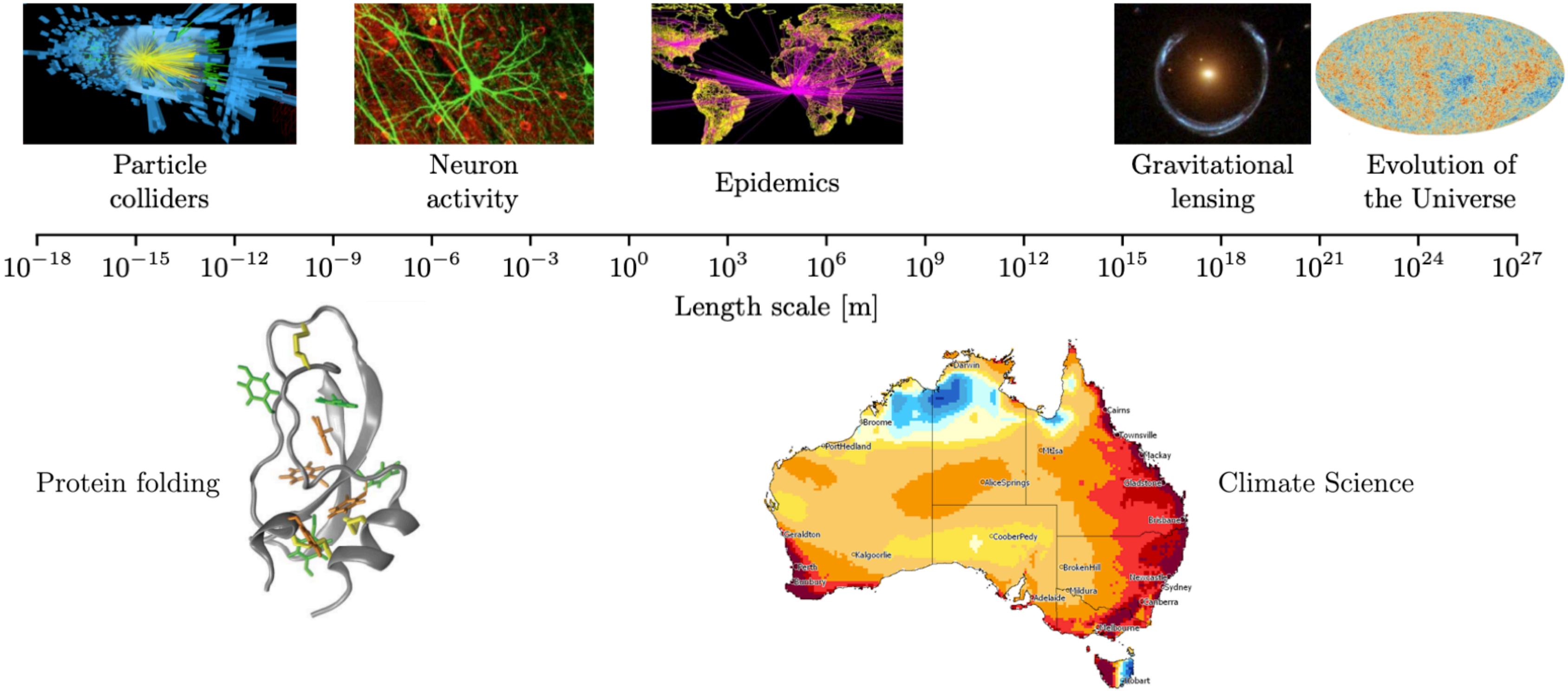


Image adapted from Cranmer K., Brehmer J., Louppe G., PNAS (2020)

Science relies heavily on high-fidelity simulators

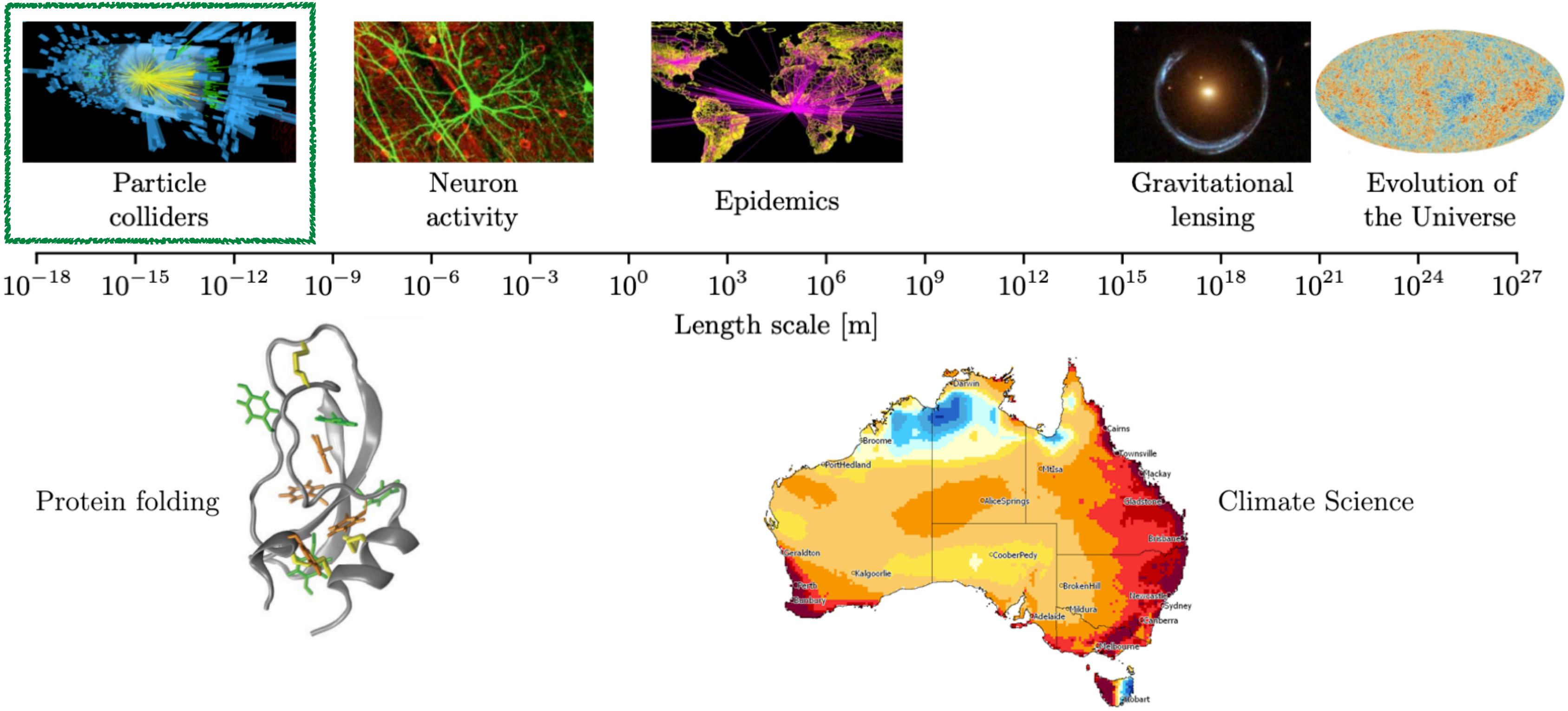


Image adapted from Cranmer K., Brehmer J., Louppe G., PNAS (2020)

We are interested in stochastic simulators

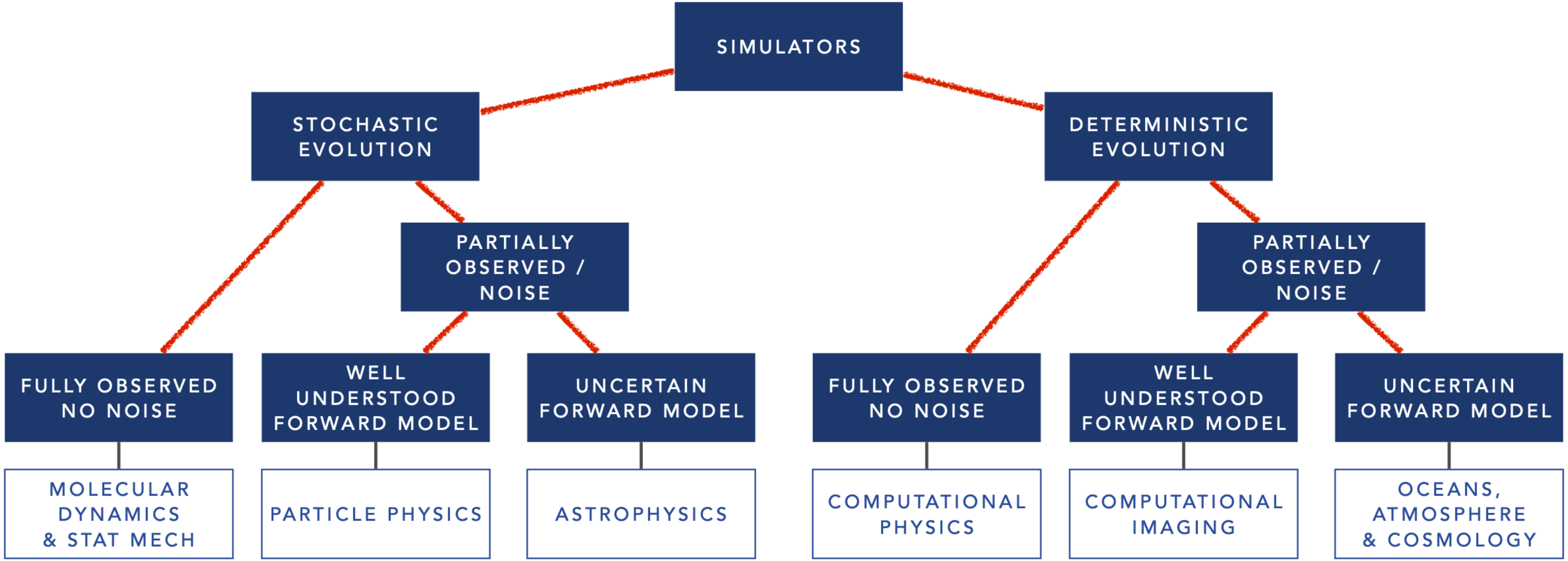
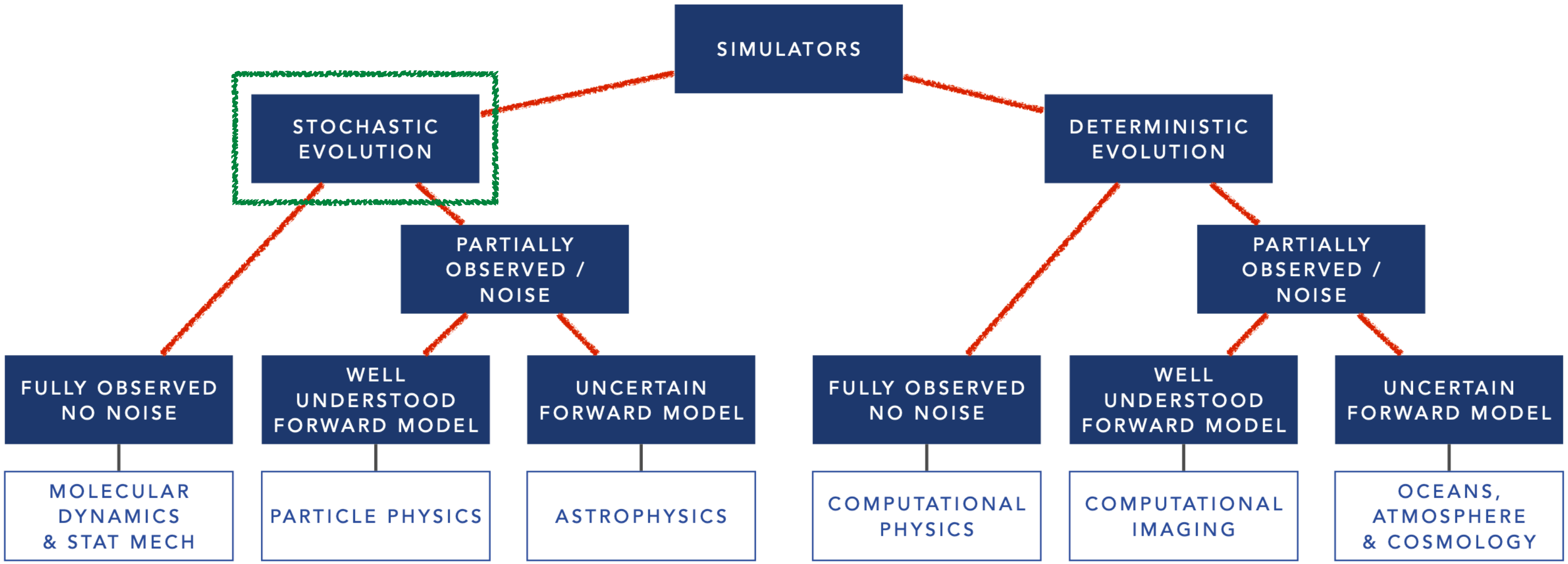


Image credit: Kyle Cranmer

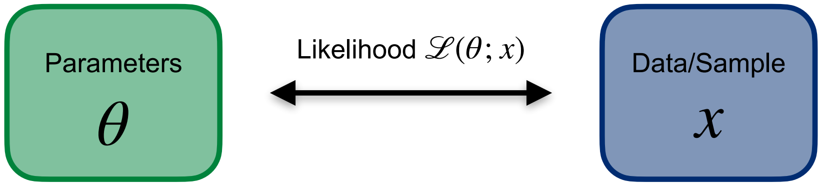
We are interested in stochastic simulators



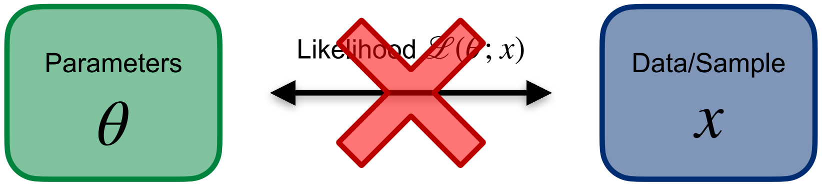
→ **Simulators that encode a likelihood and generate observable data**

Image credit: Kyle Cranmer

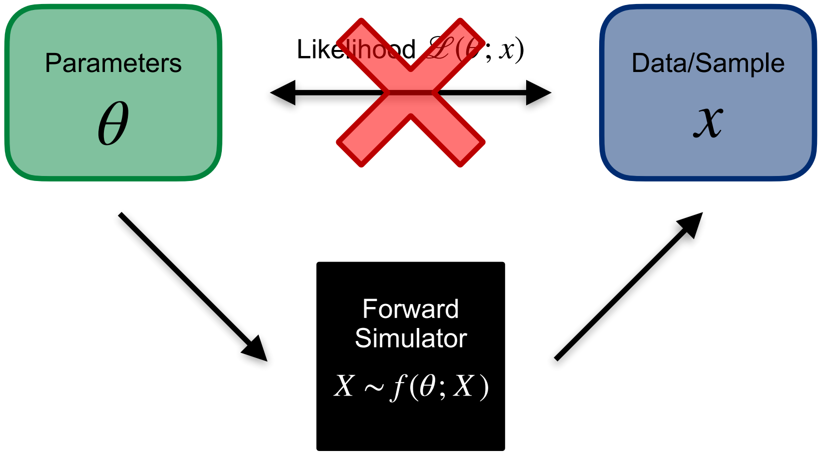
Likelihood-based Inference



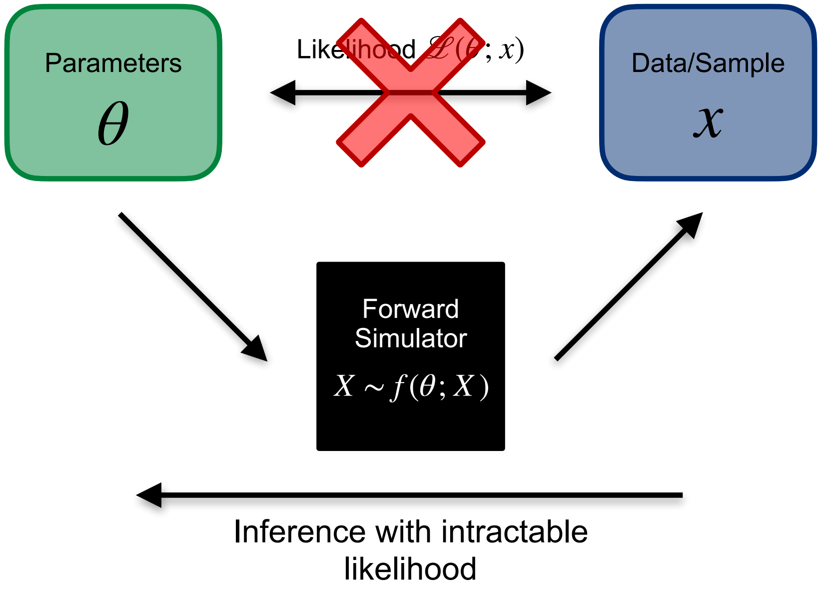
Likelihood-based Inference



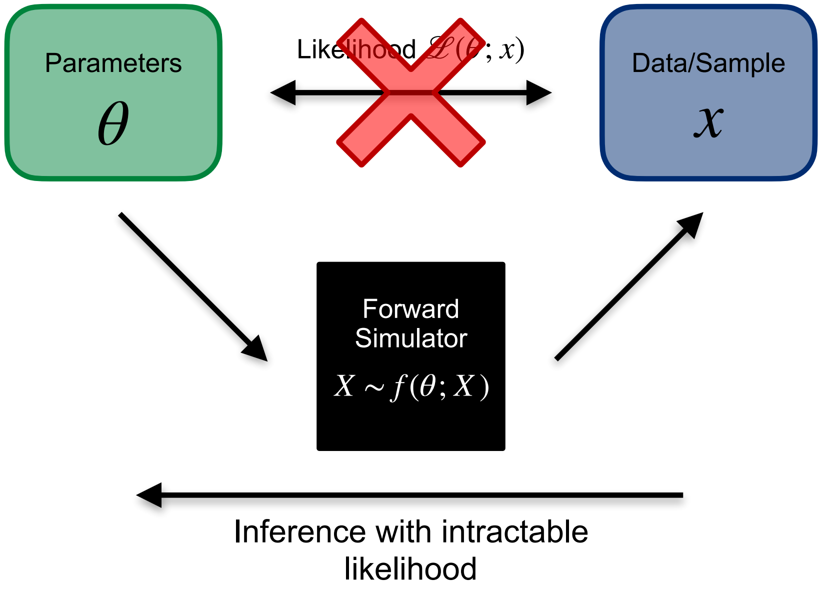
Likelihood-based Inference



Likelihood-based Inference



Likelihood-based Inference



→ **Likelihood-Free Inference (LFI)**

Constraining parameters while guaranteeing coverage

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

use $\{(\theta_1, X_1), \ldots, (\theta_B, X_B)\}$, where $\theta \sim \pi_\theta, X \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid x)}_{Posteriors}, \underbrace{\mathscr{L}(\theta; x)}_{Likelihoods}, \underbrace{\mathscr{L}(\theta_1; x)/\mathscr{L}(\theta_2; x)}_{Likelihood\ ratios}$

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014); Thomas et al (2014); Cranmer et al. (2015)

Constraining parameters while guaranteeing coverage

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

use $\{(\theta_1, X_1), \ldots, (\theta_B, X_B)\}$, where $\theta \sim \pi_\theta, X \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}, \underbrace{f(\theta \mid x)}_{Posteriors}, \underbrace{\mathscr{L}(\theta; x)}_{Likelihoods}, \underbrace{\mathscr{L}(\theta_1; x)/\mathscr{L}(\theta_2; x)}_{Likelihood\ ratios}$

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014); Thomas et al (2014); Cranmer et al. (2015)

Constraining parameters while guaranteeing coverage

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

  use $\{(\theta_1, X_1), \ldots, (\theta_B, X_B)\}$, where $\theta \sim \pi_\theta, X \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}$, $\underbrace{f(\theta \mid x)}_{Posteriors}$, $\underbrace{\mathcal{L}(\theta; x)}_{Likelihoods}$, $\underbrace{\mathcal{L}(\theta_1; x)/\mathcal{L}(\theta_2; x)}_{Likelihood\ ratios}$

- **Do these methods give reliable measures of uncertainty around parameters of interest?**
  - → Hermans et al. (2021) showed all algorithms produce overconfident approximations: $\mathbb{E}\left[\mathbb{I}[\theta \in \Theta_{\hat{p}(\theta|x)}(1-\alpha)]\right] < 1 - \alpha$
  - → 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); Thomas et al (2014); Cranmer et al. (2015)

Constraining parameters while guaranteeing coverage

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

  use $\{(\theta_1, X_1), \ldots, (\theta_B, X_B)\}$, where $\theta \sim \pi_\theta, X \sim F_\theta$ $\rightarrow$ $\underbrace{\theta}_{Parameters}$, $\underbrace{f(\theta \mid x)}_{Posteriors}$, $\underbrace{\mathscr{L}(\theta; x)}_{Likelihoods}$, $\underbrace{\mathscr{L}(\theta_1; x)/\mathscr{L}(\theta_2; x)}_{Likelihood\ ratios}$

- **Do these methods give reliable measures of uncertainty around parameters of interest?**
  - → Hermans et al. (2021) showed all algorithms produce overconfident approximations: $\mathbb{E}\left[\mathbb{I}[\theta \in \Theta_{\hat{p}(\theta|x)}(1-\alpha)]\right] < 1 - \alpha$
  - → Hinders the reliability of scientific conclusions

- **Goals:**
  1. Recalibrate predictions and/or posteriors → confidence sets with frequentist guarantees for finite n across Θ

  $$\mathbb{P}(\theta \in \mathscr{R}(X_o) \mid \theta) = 1 - \alpha \quad \forall \theta \in \Theta, \quad X_o = (X_1, \ldots, X_n)$$

  2. Check actual coverage across the whole Θ, without costly Monte-Carlo simulations

1. E.g. Heinrich (2022); Miller et al. (2021); Papamakarios et al. (2016); Lueckmann et al (2016); Izbicki et al. (2014); Thomas et al (2014); Cranmer et al. (2015)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  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 $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  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 $\tau(X;\theta_0),\ \forall\theta_0 \in \Theta$

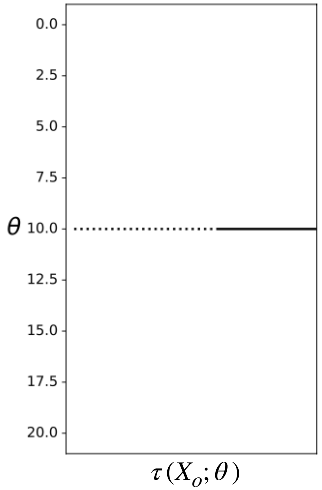


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  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 $\tau(X;\theta_0),\ \forall\theta_0 \in \Theta$

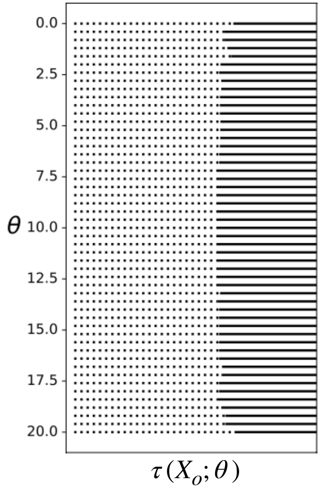


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  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 $\tau(X;\theta_0),\ \forall\theta_0 \in \Theta$

ii. $\tau(X_o;\theta_0),\ \theta_0 \in \Theta$

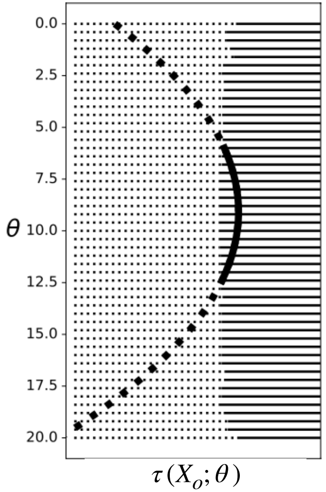


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  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 $\tau(X;\theta_0),\ \forall \theta_0 \in \Theta$

ii. $\tau(X_o;\theta_0),\ \theta_0 \in \Theta$

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

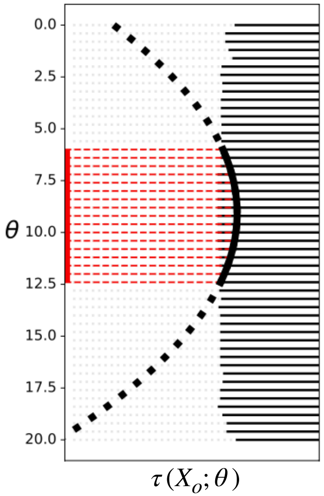


Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

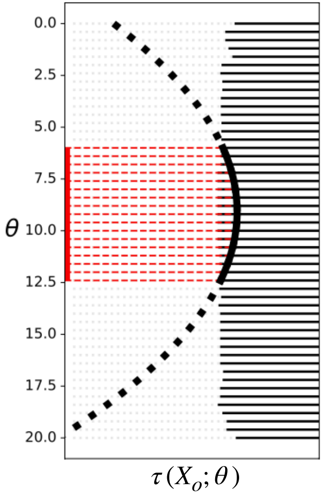
  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  3. Critical values $C_{\theta,\alpha}$

> **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 $\tau(X;\theta_0),\ \forall \theta_0 \in \Theta$

ii. $\tau(X_o;\theta_0),\ \theta_0 \in \Theta$

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



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

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

Image adapted from Dalmasso et al. (2020)

Neyman construction of confidence sets

- **Ingredients:**

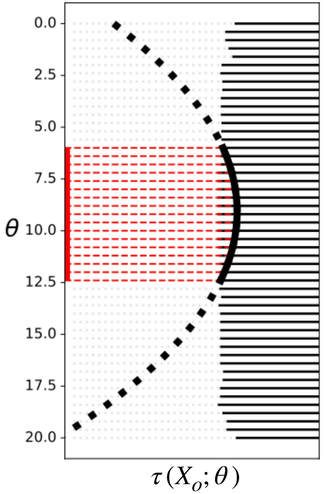
  1. Data $X \sim F_\theta$
  2. Test statistic $\tau(X;\theta)$
  3. Critical values $C_{\theta,\alpha}$

> **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 $\tau(X;\theta_0),\ \forall\theta_0 \in \Theta$

ii. $\tau(X_o;\theta_0),\ \theta_0 \in \Theta$

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



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

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

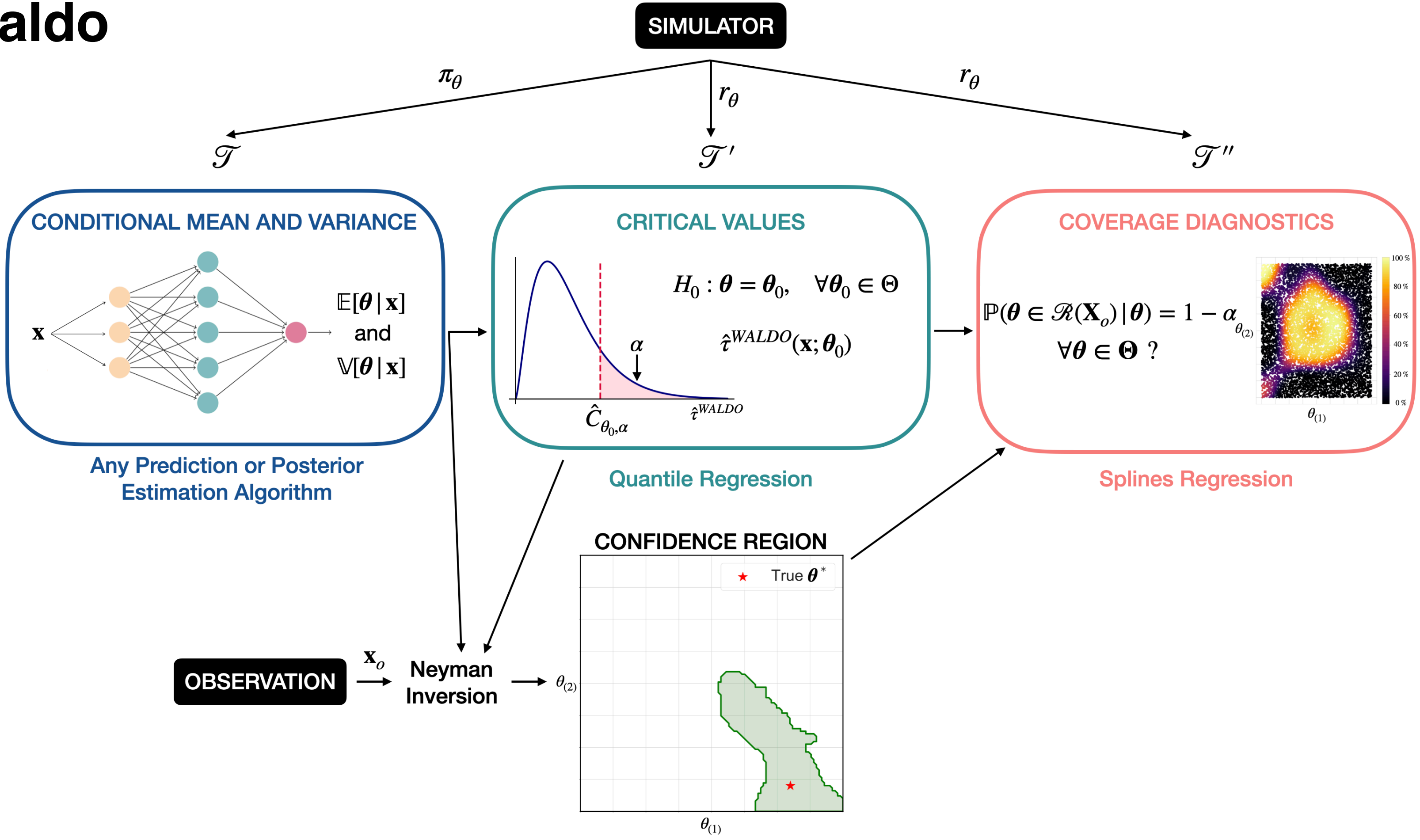
→

- **Waldo** test statistic:

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

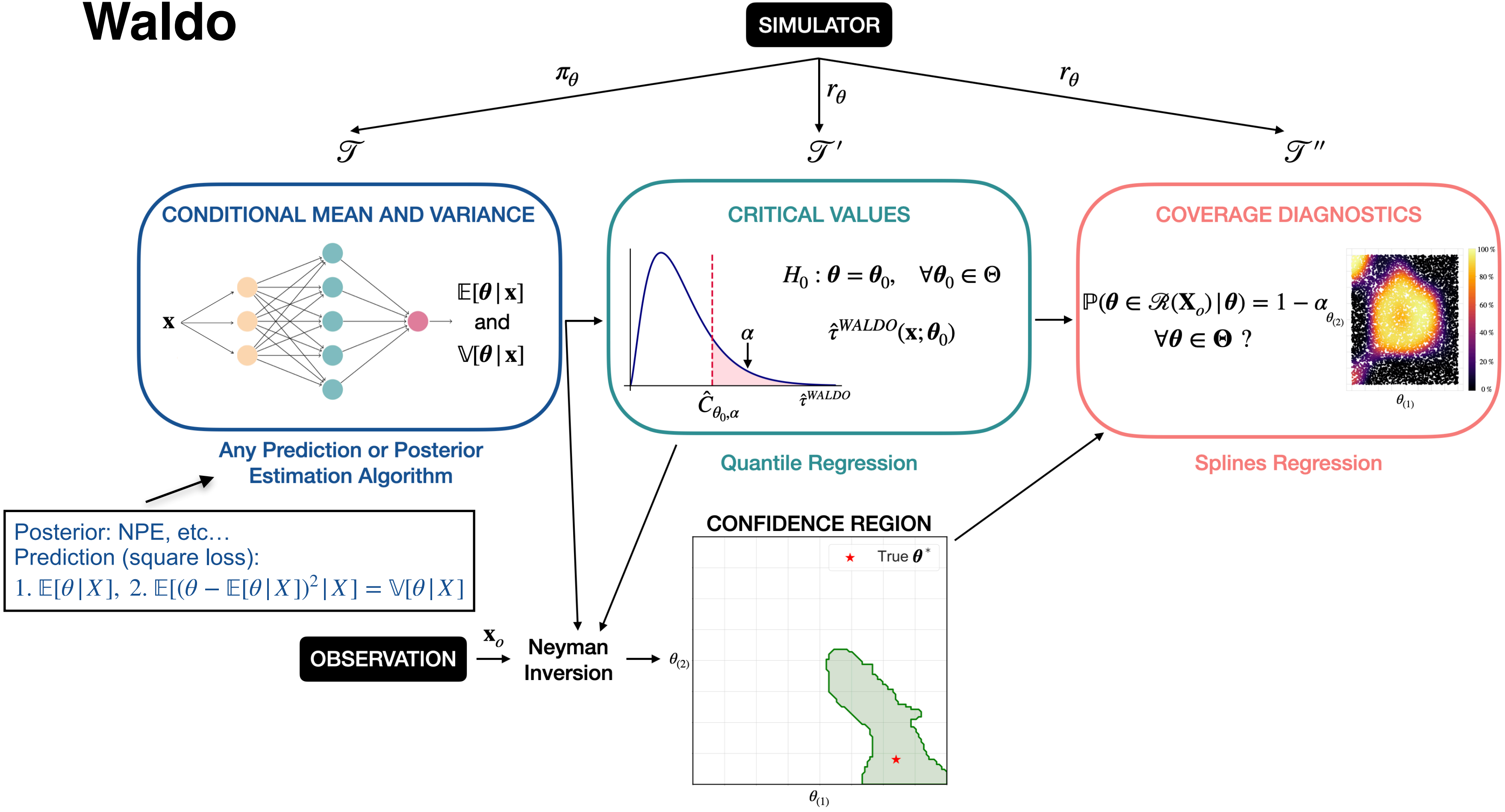
Image adapted from Dalmasso et al. (2020)

Waldo



SIMULATOR
π_θ
r_θ
r_θ
$\mathscr{T}$
$\mathscr{T}'$
$\mathscr{T}''$
CONDITIONAL MEAN AND VARIANCE
$\mathbf{x}$
$\mathbb{E}[\boldsymbol{\theta} \,|\, \mathbf{x}]$
and
$\mathbb{V}[\boldsymbol{\theta} \,|\, \mathbf{x}]$
Any Prediction or Posterior Estimation Algorithm
CRITICAL VALUES
$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \quad \forall \boldsymbol{\theta}_0 \in \Theta$
α
$\hat{\tau}^{WALDO}(\mathbf{x}; \boldsymbol{\theta}_0)$
$\hat{C}_{\theta_0,\alpha}$
$\hat{\tau}^{WALDO}$
Quantile Regression
COVERAGE DIAGNOSTICS
$\mathbb{P}(\boldsymbol{\theta} \in \mathscr{R}(\mathbf{X}_o) \,|\, \boldsymbol{\theta}) = 1 - \alpha$
$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$?
$\theta_{(2)}$
$\theta_{(1)}$
100 %
80 %
60 %
40 %
20 %
0 %
Splines Regression
CONFIDENCE REGION
True $\boldsymbol{\theta}^*$
$\theta_{(2)}$
$\theta_{(1)}$
OBSERVATION
$\mathbf{x}_o$
Neyman Inversion

Waldo



SIMULATOR
π_θ
r_θ
r_θ
$\mathscr{T}$
$\mathscr{T}'$
$\mathscr{T}''$
CONDITIONAL MEAN AND VARIANCE
$\mathbf{x}$
$\mathbb{E}[\boldsymbol{\theta}\,|\,\mathbf{x}]$
and
$\mathbb{V}[\boldsymbol{\theta}\,|\,\mathbf{x}]$
Any Prediction or Posterior Estimation Algorithm
Posterior: NPE, etc…
Prediction (square loss):
1. $\mathbb{E}[\theta\,|\,X]$, 2. $\mathbb{E}[(\theta - \mathbb{E}[\theta\,|\,X])^2\,|\,X] = \mathbb{V}[\theta\,|\,X]$
CRITICAL VALUES
$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \quad \forall \boldsymbol{\theta}_0 \in \Theta$
α
$\hat{\tau}^{WALDO}(\mathbf{x}; \boldsymbol{\theta}_0)$
$\hat{C}_{\theta_0,\alpha}$
$\hat{\tau}^{WALDO}$
Quantile Regression
COVERAGE DIAGNOSTICS
$\mathbb{P}(\boldsymbol{\theta} \in \mathscr{R}(\mathbf{X}_o)\,|\,\boldsymbol{\theta}) = 1 - \alpha$
$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$?
$\theta_{(2)}$
$\theta_{(1)}$
100 %
80 %
60 %
40 %
20 %
0 %
Splines Regression
OBSERVATION
$\mathbf{x}_o$
Neyman Inversion
CONFIDENCE REGION
True $\boldsymbol{\theta}^*$
$\theta_{(2)}$
$\theta_{(1)}$

Statistical Properties

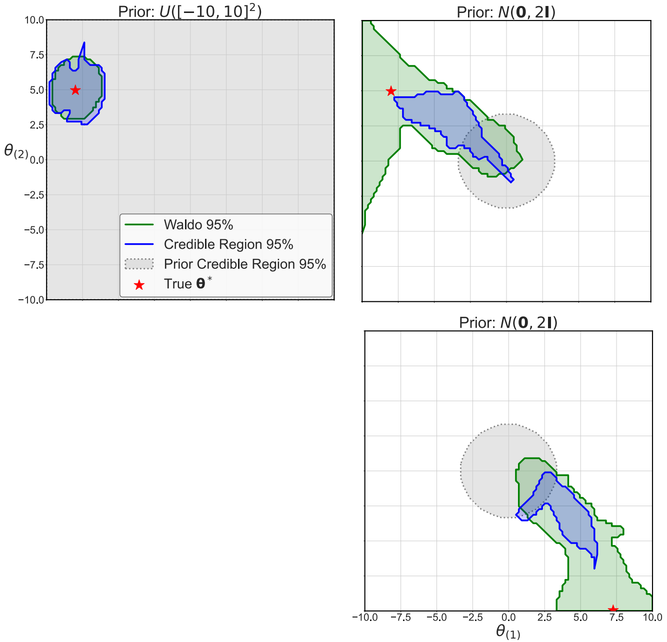
- **Synthetic example:** estimate mean of components of a Gaussian mixture (as in Lueckmann et al. 2021)

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

Statistical Properties

- **Synthetic example:** estimate mean of components of a Gaussian mixture (as in Lueckmann et al. 2021)

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

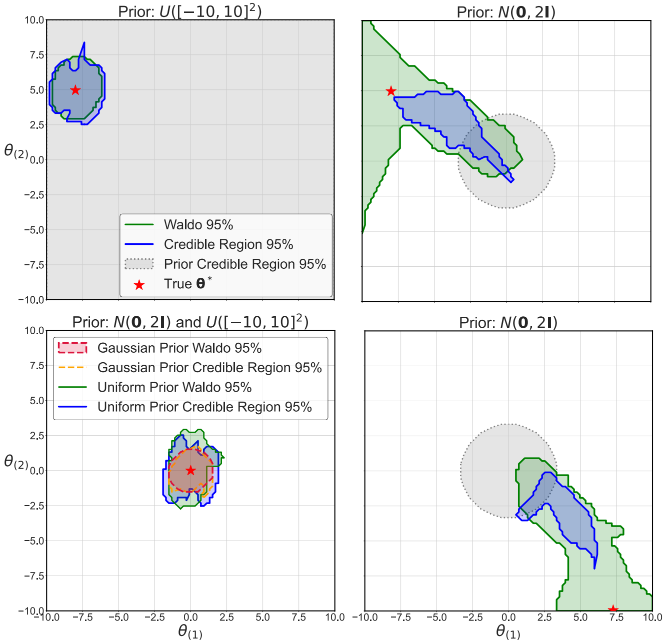


- **Coverage:** Waldo recalibrates posterior credible regions to account for estimation error and/or bias, regardless of prior and sample size

Statistical Properties

- **Synthetic example:** estimate mean of components of a Gaussian mixture (as in Lueckmann et al. 2021)

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



- **Coverage:** Waldo recalibrates posterior credible regions to account for estimation error and/or bias, regardless of prior and sample size

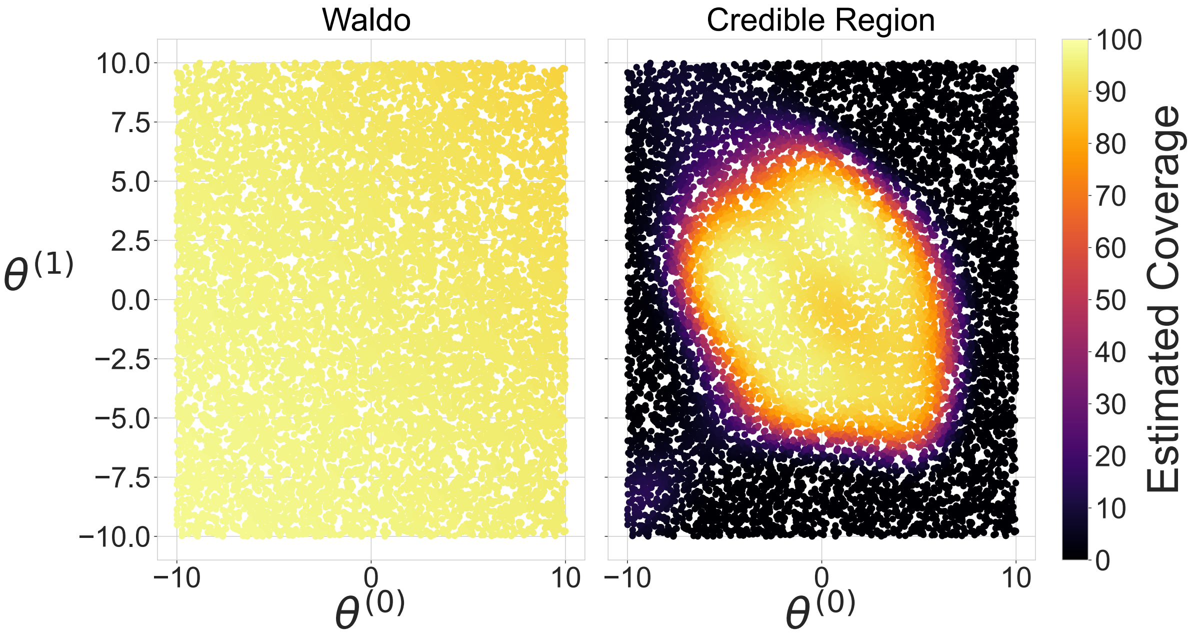
- **Power (expected size):** if the prior is correctly specified, Waldo still benefits from the additional information

Statistical Properties (coverage diagnostics)

- **Synthetic example:** estimate mean of a Gaussian mixture (as in Lueckmann et al. 2021)

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

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.

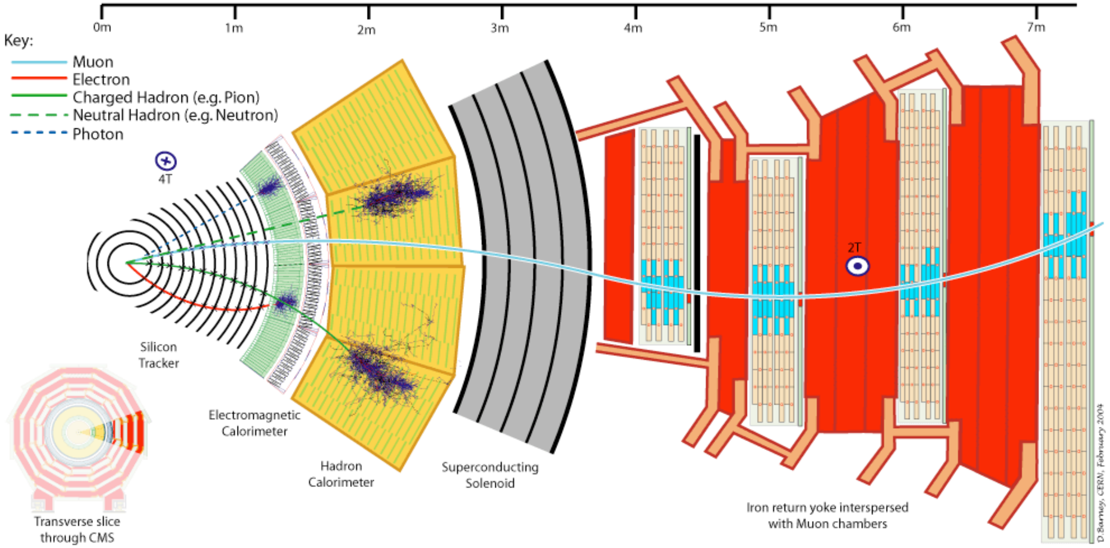
Inference for calorimetric muon energy measurements

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

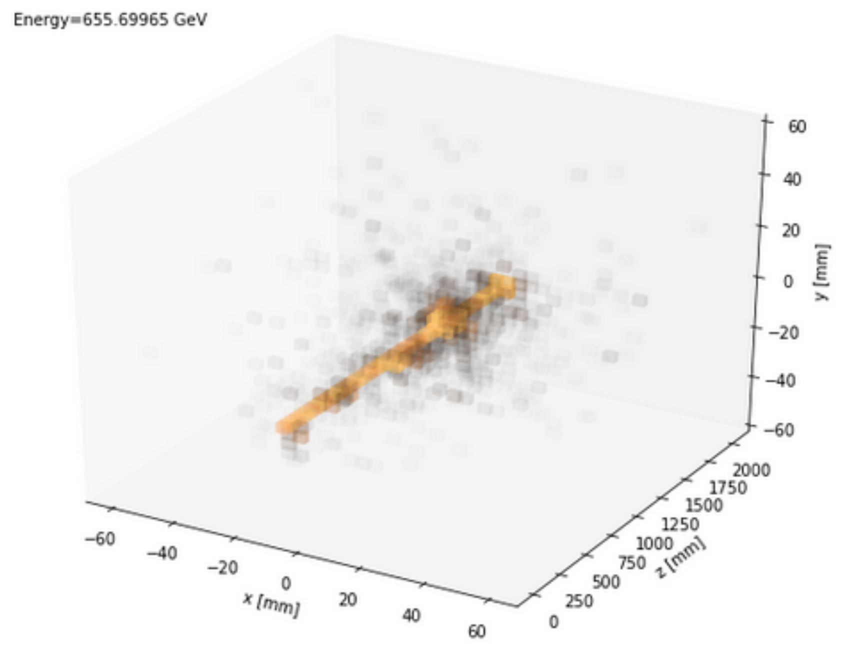
Their importance is mainly due to two facts: **first**, they emerge as a signature in processes which could signal the existence of new physics, and **second**, they are (relatively) easy to identify.

Above: Aerial view of the position of the LHC.

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



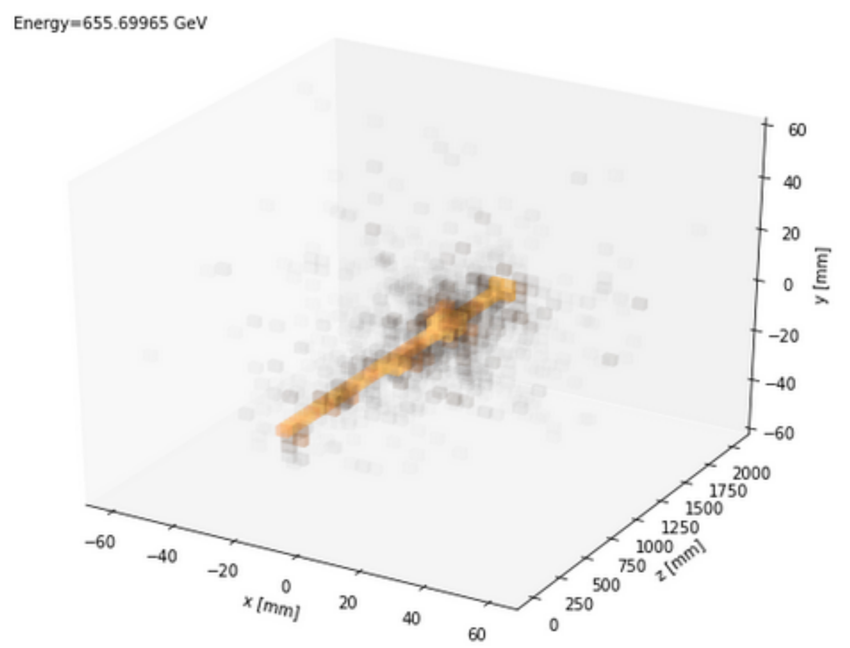
Inputs: 1D energy-sum, 28 features and full calorimeter



Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- Data obtained from Geant4[1] with incoming energy between 50 GeV and 8000 GeV

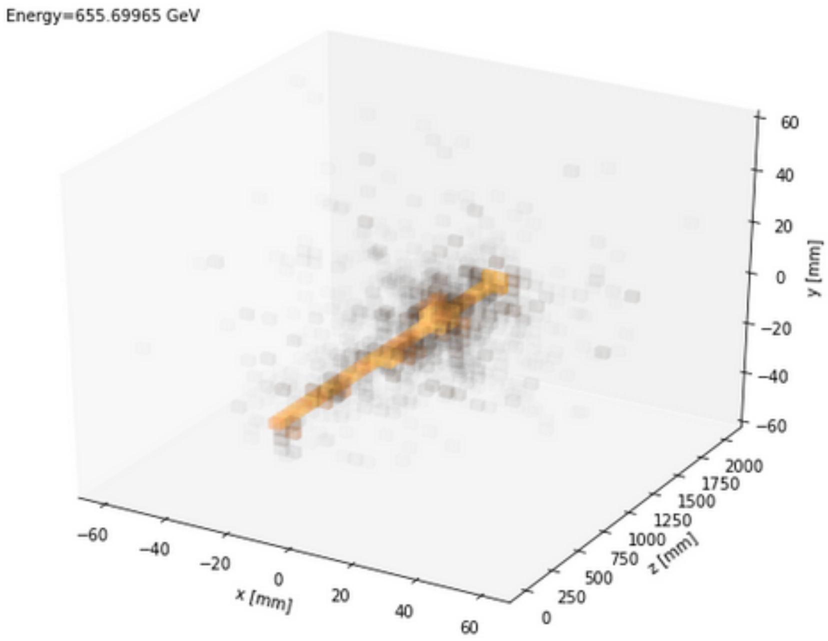
Inputs: 1D energy-sum, 28 features and full calorimeter



Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- 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

Inputs: 1D energy-sum, 28 features and full calorimeter

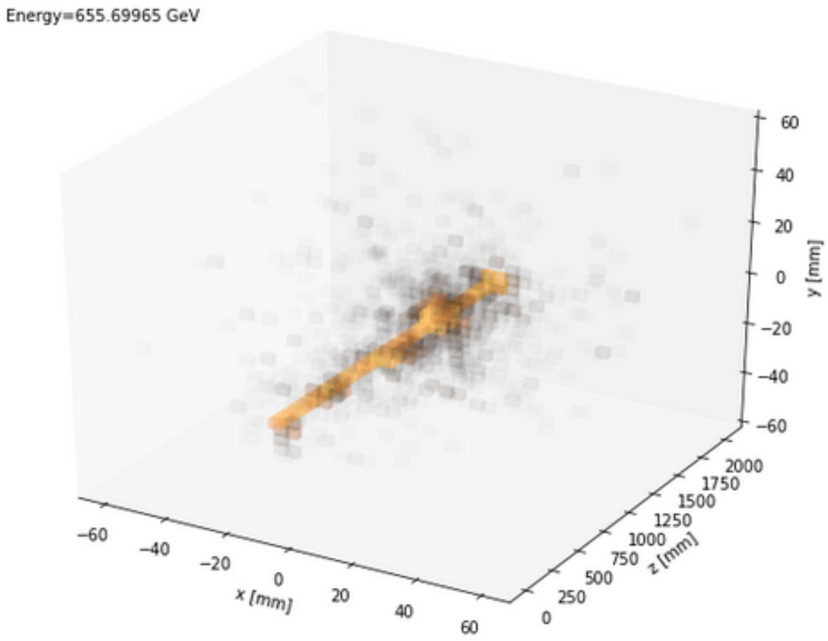


Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

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

Inputs: 1D energy-sum, 28 features and full calorimeter



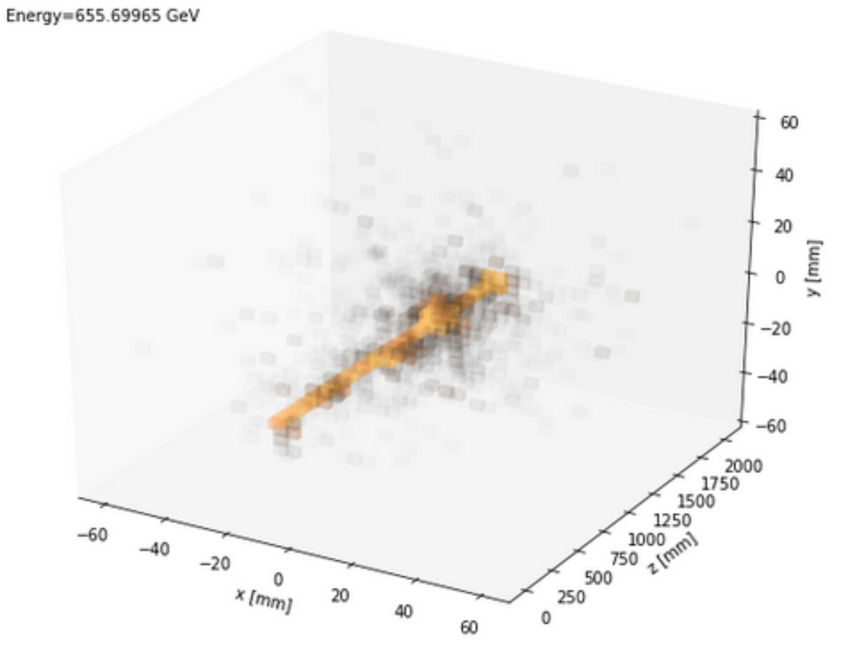
Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- 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)

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

Inputs: 1D energy-sum, 28 features and full calorimeter

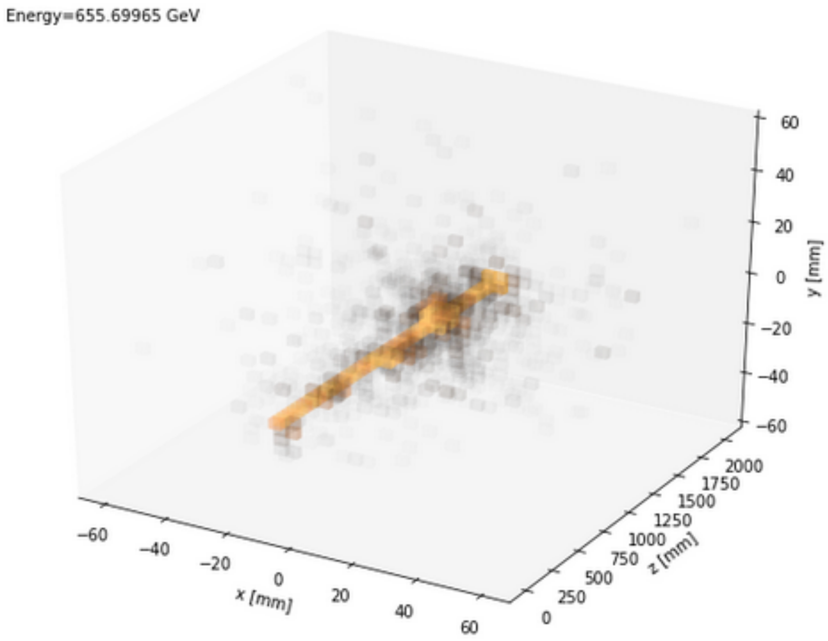


Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- 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)

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

Inputs: 1D energy-sum, 28 features and full calorimeter

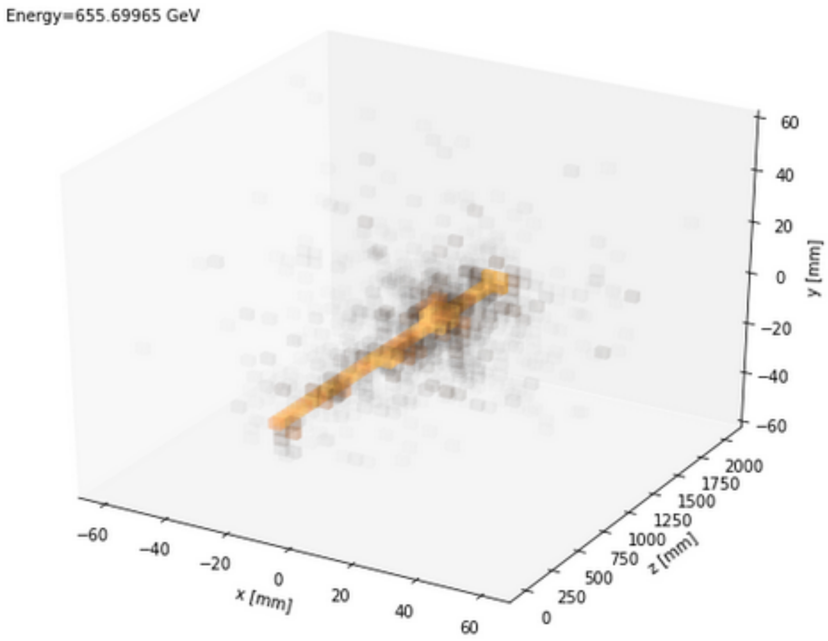


Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- 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)

Inputs: 1D energy-sum, 28 features and full calorimeter



Muon entering the calorimeter in z direction. The colour gradient indicates the logarithmic energy deposits of a muon with incoming energy ≈ 655.7 GeV. Black corresponds to zero, orange to intermediate, and white to maximum energy.

- 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**:

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 physical process that produced a muon and discover new physics

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 physical process that produced a muon and discover new physics

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

Prediction algorithms

Three “nested” datasets:

Prediction algorithms

Three “nested” datasets:

1. One-dimensional energy sum: minimizer of Cross-Validation MSE loss (XGBoost)

Prediction algorithms

Three "nested" datasets:

1. One-dimensional energy sum: minimizer of Cross-Validation MSE loss (XGBoost)
2. 27 features + 1D energy sum: minimizer of Cross-Validation MSE loss (XGBoost)

Prediction algorithms

Three "nested" datasets:

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

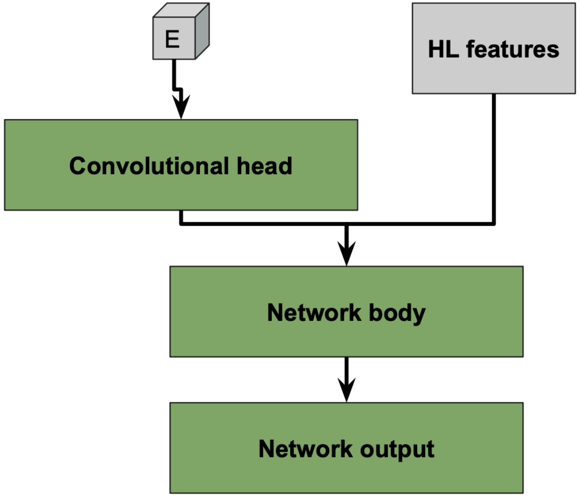


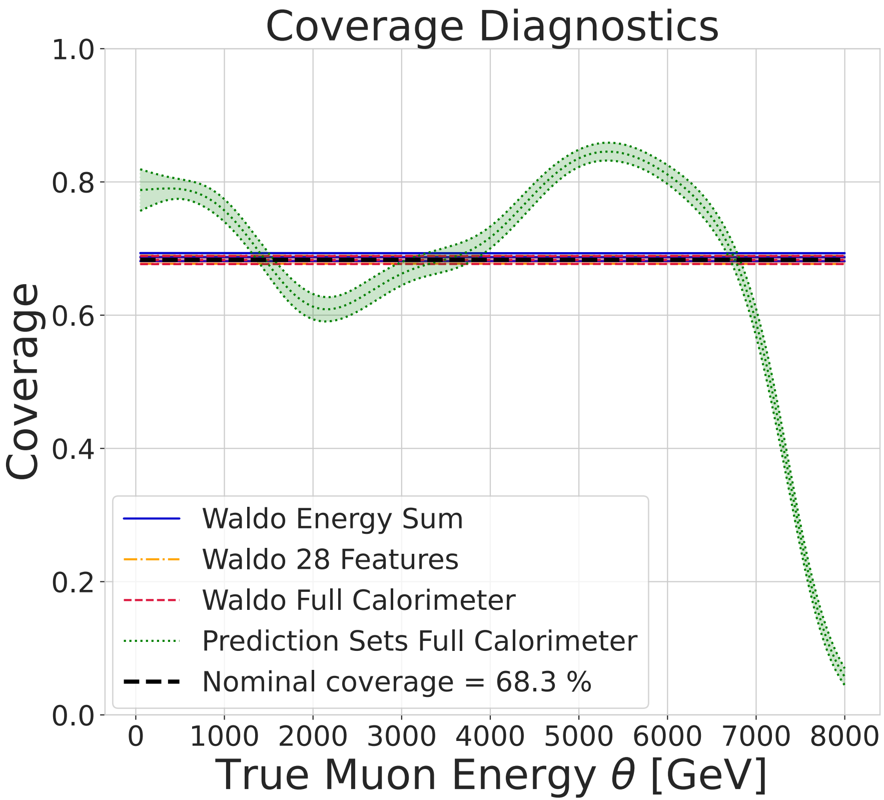
Image credit: Kieseler et al. (2022)

Confidence sets for muon energy have proper coverage

- Nominal coverage is achieved regardless of the dataset used

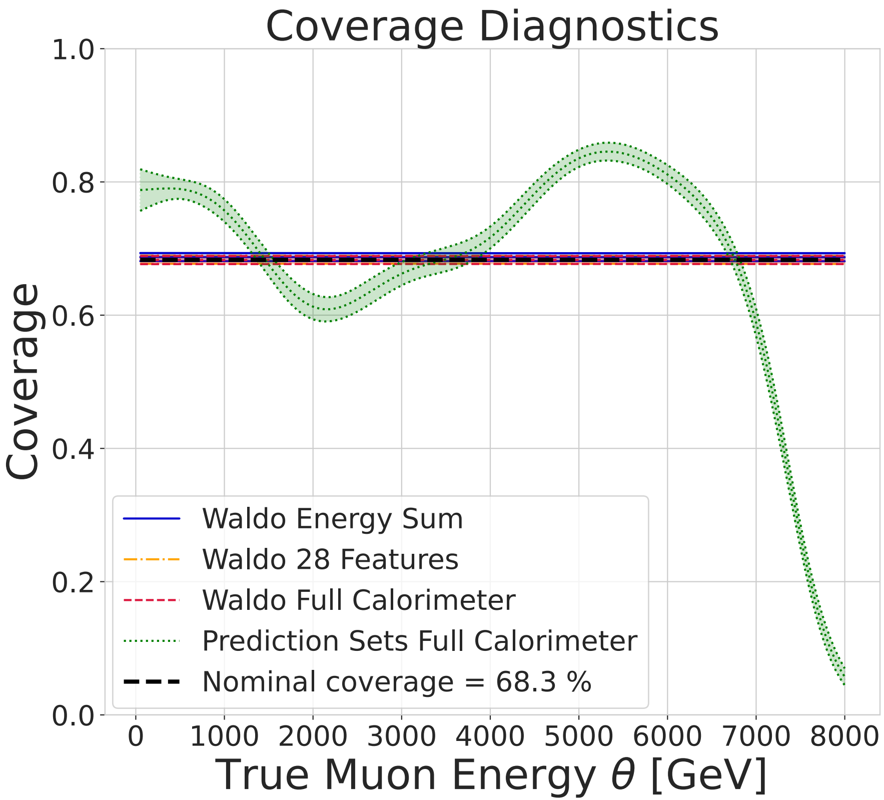
Confidence sets for muon energy have proper coverage

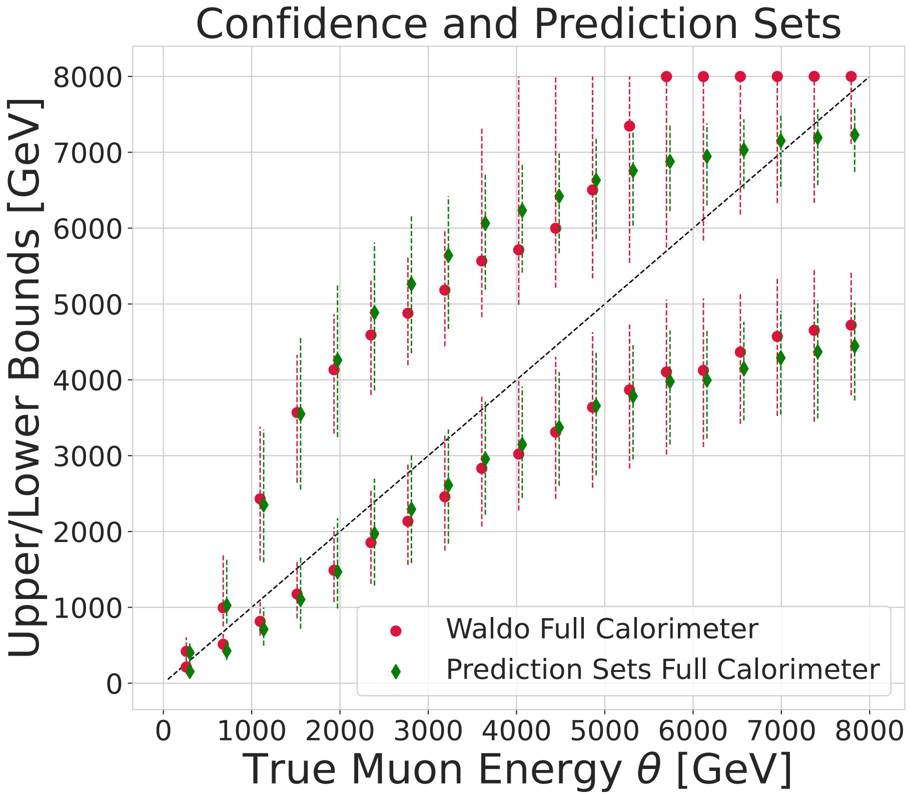
- Nominal coverage is achieved regardless of the dataset used



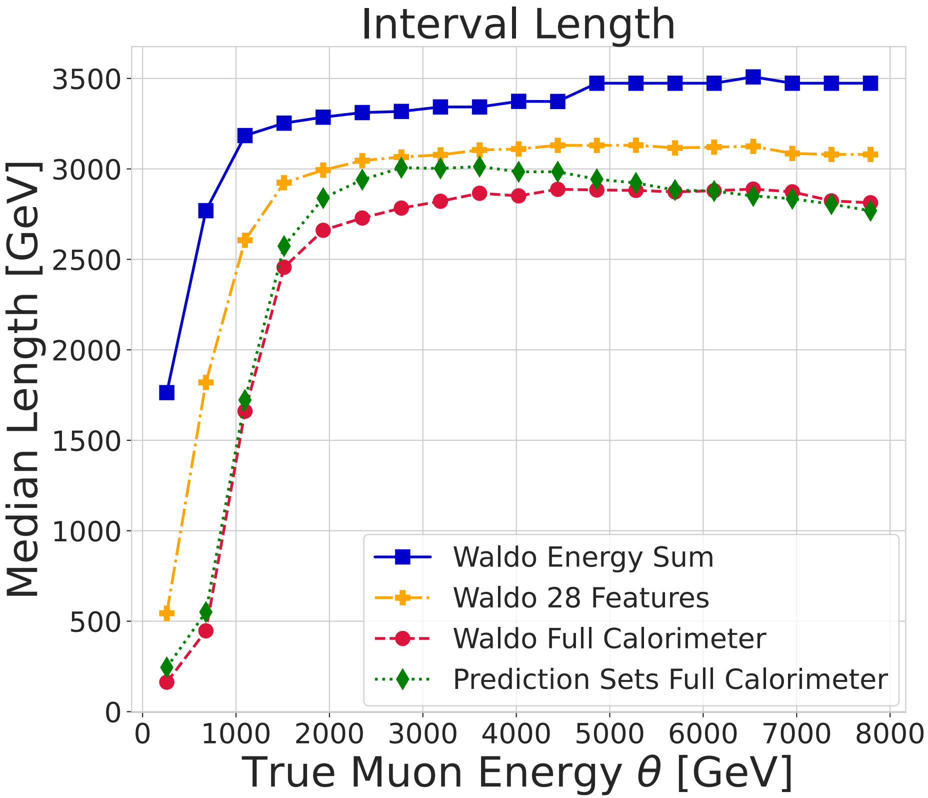
Confidence sets for muon energy have proper coverage

- Nominal coverage is achieved regardless of the dataset used





Valuable information in high-granularity calorimeter



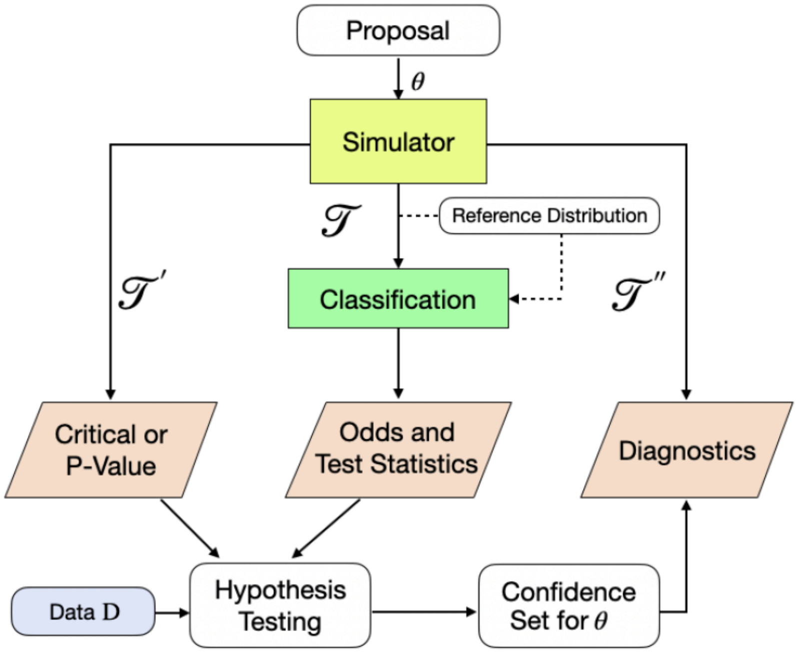
- Intervals are longer as the data becomes lower-dimensional
- Prediction sets can even be larger than Waldo confidence sets (while also not guaranteeing coverage)

Thanks!

Carnegie Mellon University

Likelihood-free Frequentist Inference (LF2I)

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



A modular framework:

1. **central branch:** parameterized odds

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

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

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

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

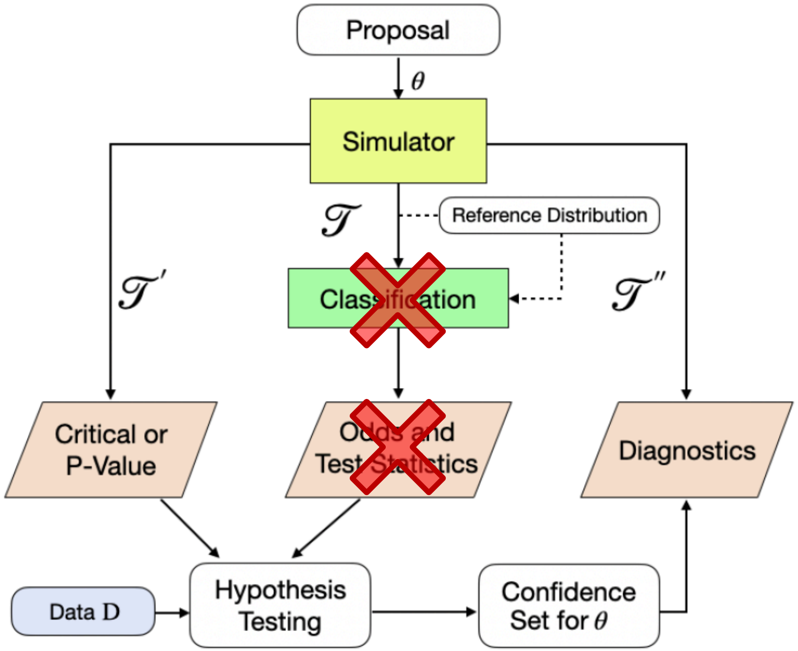
→ **(1 + 2) use Neyman inversion:**

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

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

Image credit: Dalmasso et al. (2021)

Likelihood-free Frequentist Inference (LF2I)



- 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;

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

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

Bias and coverage of prediction intervals

- Train on $(X_1, \theta_1), \ldots, (X_B, \theta_B) \sim f(X, \theta)$ and output $\hat{\theta} = \hat{\mathbb{E}}[\theta \mid X]$
  - → posterior mean, which depends on **marginal** since $f(X, \theta) = f(X \mid \theta) f(\theta)$

- 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:

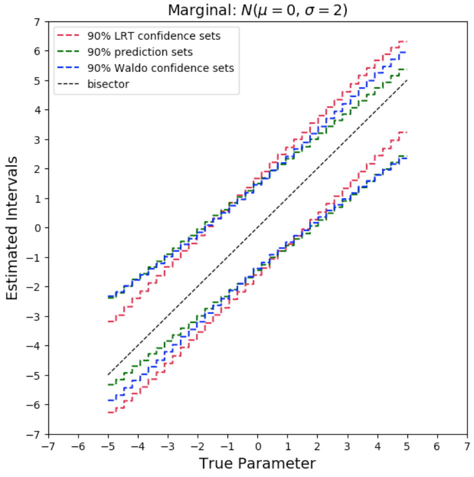
$$\theta \sim \mathcal{N}(\mu = 0, \sigma = 2)$$
$$X \mid \theta \sim \mathcal{N}(\theta, \sigma = 1)$$

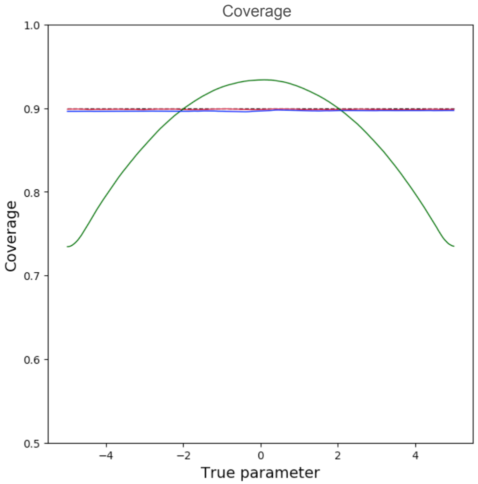
Construct confidence sets via

- Likelihood-ratio test
- Waldo

and

- Prediction sets





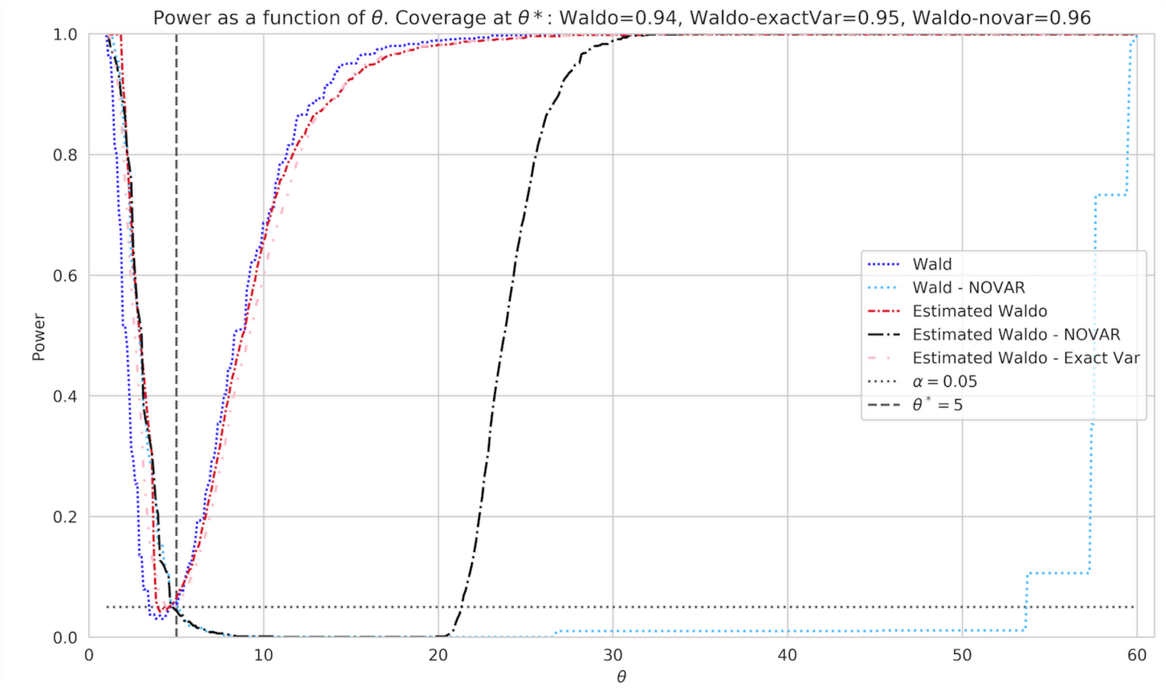
Left: means of upper and lower bounds of interval estimates for 100,000 observations divided in 38 bins over the true parameter.

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

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

- **Waldo** requires to estimate $\mathbb{V}[\theta \,|\, X]$. Why not simply use $\tau^{Waldo-novar}(X;\theta) := (\mathbb{E}[\theta \,|\, X] - \theta)^2$?
- Reject H_0 if $(X_1, \ldots, X_n) \in R$. Let $\mathscr{P}^{Waldo} = \mathbb{P}_\theta[(X_1, \ldots, X_n) \in R]$ be the **power function** of the Waldo test

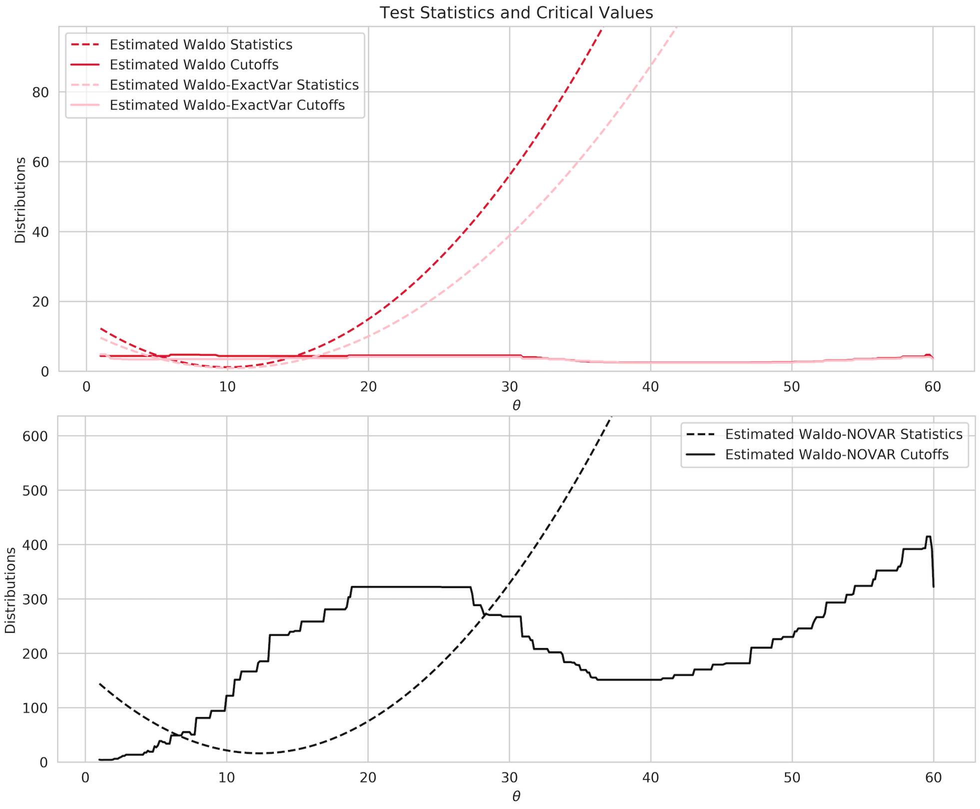
→ setting: inference on the shape of a **Pareto** likelihood $X \sim Pareto(\theta, x_{min} = 1), \theta \sim \mathscr{U}(0{,}60)$



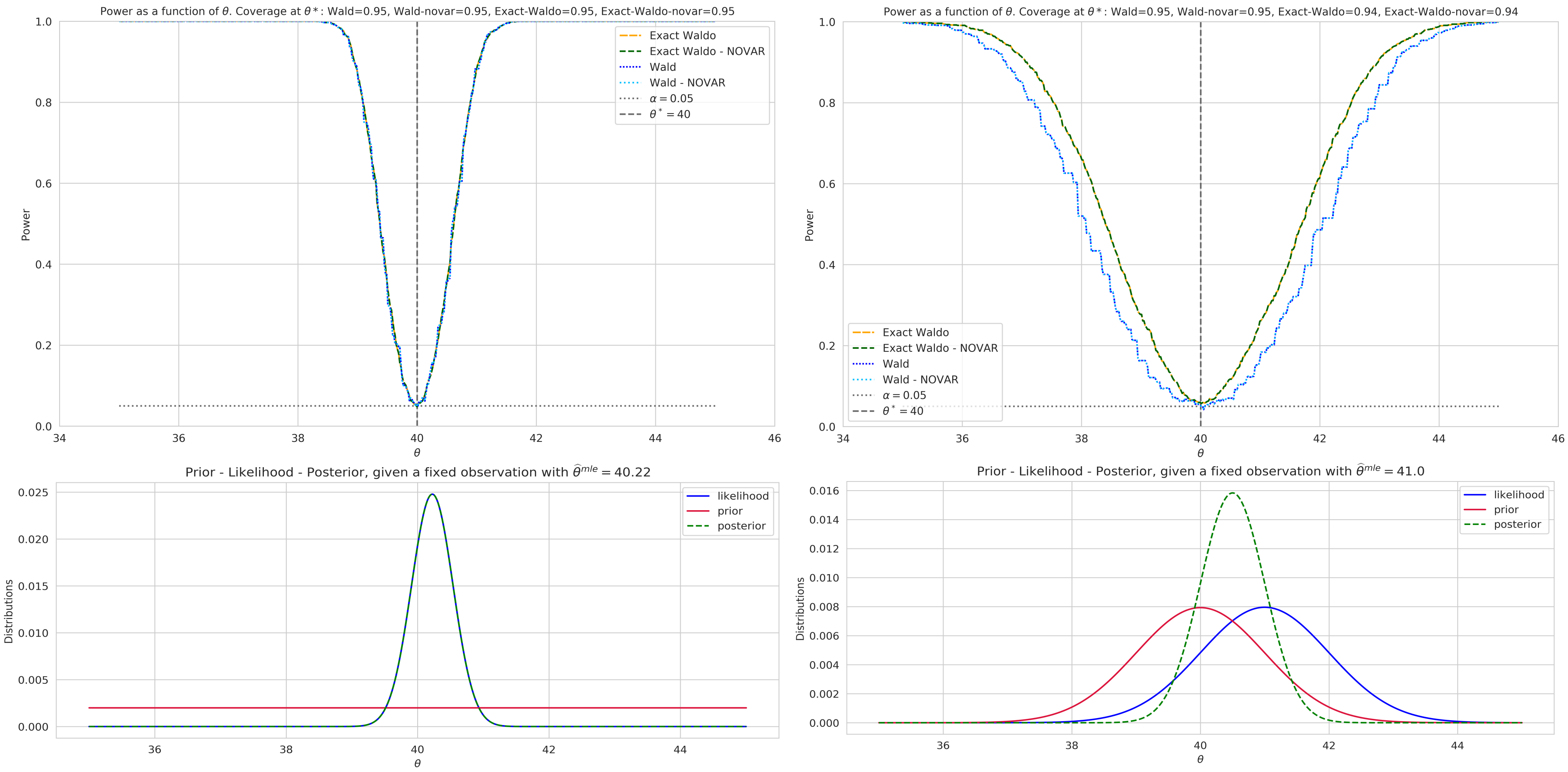
$$\mathscr{P}^{Waldo} \gg \mathscr{P}^{Waldo-novar}$$

Waldo appears to be a pivotal test statistic

- A **pivot** is a function of the data and the unknown parameter θ, whose distribution does not depend on θ.



Combining frequentist coverage with prior knowledge



Coverage guarantees

Assumption 1 (Uniform consistency) *Let $F(\cdot|\theta)$ be the cumulative distribution function of the test statistic $\lambda(\mathcal{D};\theta_0)$ conditional on θ, where $\mathcal{D} \sim F_\theta$. Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated conditional distribution function, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_\theta$. Assume that the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} 0.$$

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a strictly continuous statistic $\lambda(\mathcal{D};\theta_0)$ chosen according to Algorithm 1 for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 1, then,*

$$\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D};\theta_0) \leq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\theta_0}$ and conditional on the random variable $C_{B'}$.

Coverage guarantees

Assumption 1 (Uniform consistency) *Let $F(\cdot|\theta)$ be the cumulative distribution function of the test statistic $\lambda(\mathcal{D};\theta_0)$ conditional on θ, where $\mathcal{D} \sim F_\theta$. Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated conditional distribution function, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_\theta$. Assume that the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} 0.$$

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a strictly continuous statistic $\lambda(\mathcal{D};\theta_0)$ chosen according to Algorithm 1 for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 1, then,*

$$\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D};\theta_0) \leq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\theta_0}$ and conditional on the random variable $C_{B'}$.

From Dalmasso, Masserano, Zhao, Izbicki, Lee (2021). Available at https://arxiv.org/pdf/2107.03920.pdf