DIRECT LEARNING OF SYSTEMATICS-AWARE SUMMARY STATISTICS

Pablo de Castro (@pablodecm) and Tommaso Dorigo (@dorigo)

6th August 2018 @ XIIIth QCHS Conference (Maynooth University - Ireland)

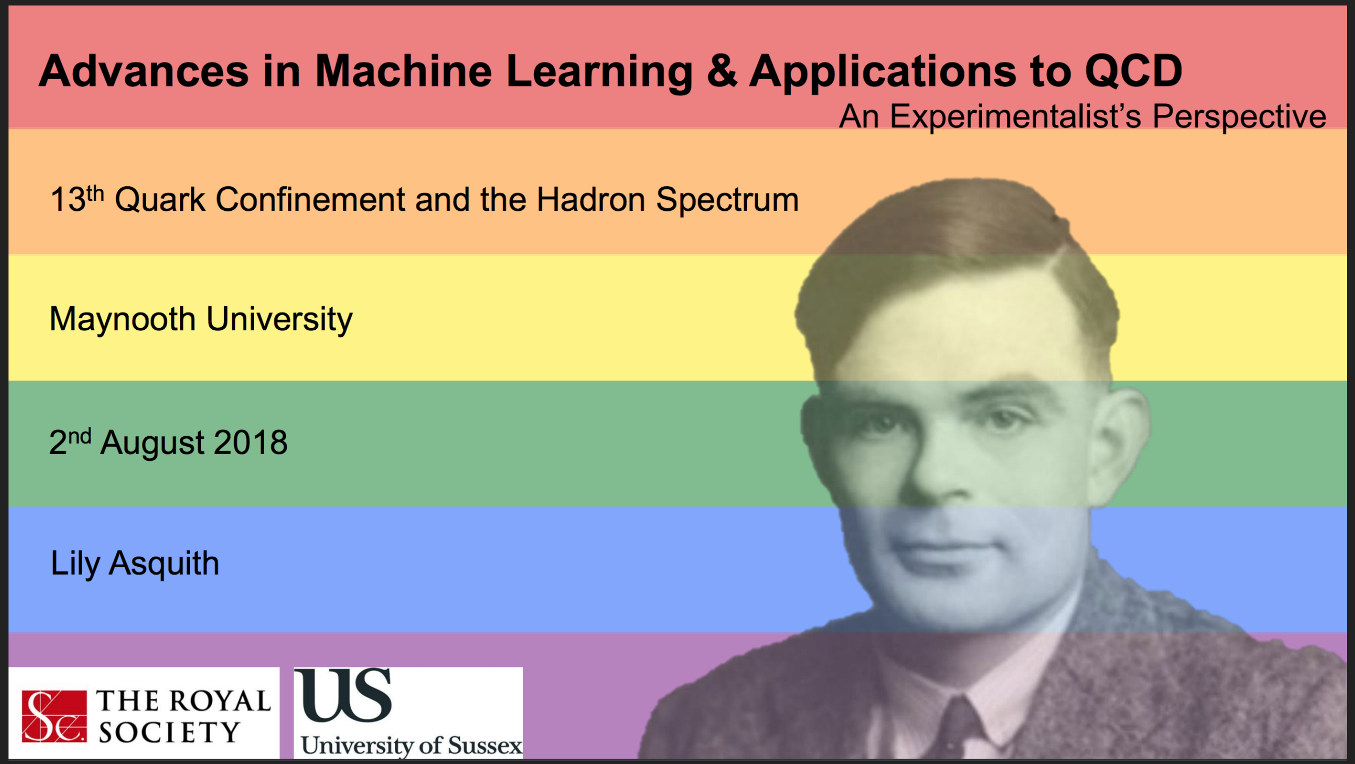
Poster Award Lightning Talk

Poster & *Parallel* - *Statistical Methods for Physics Analysis in the XXI Century*

AMVA4NewPhysics has received funding from European Union's Horizon 2020 Programme under Grant Agreement number 675440

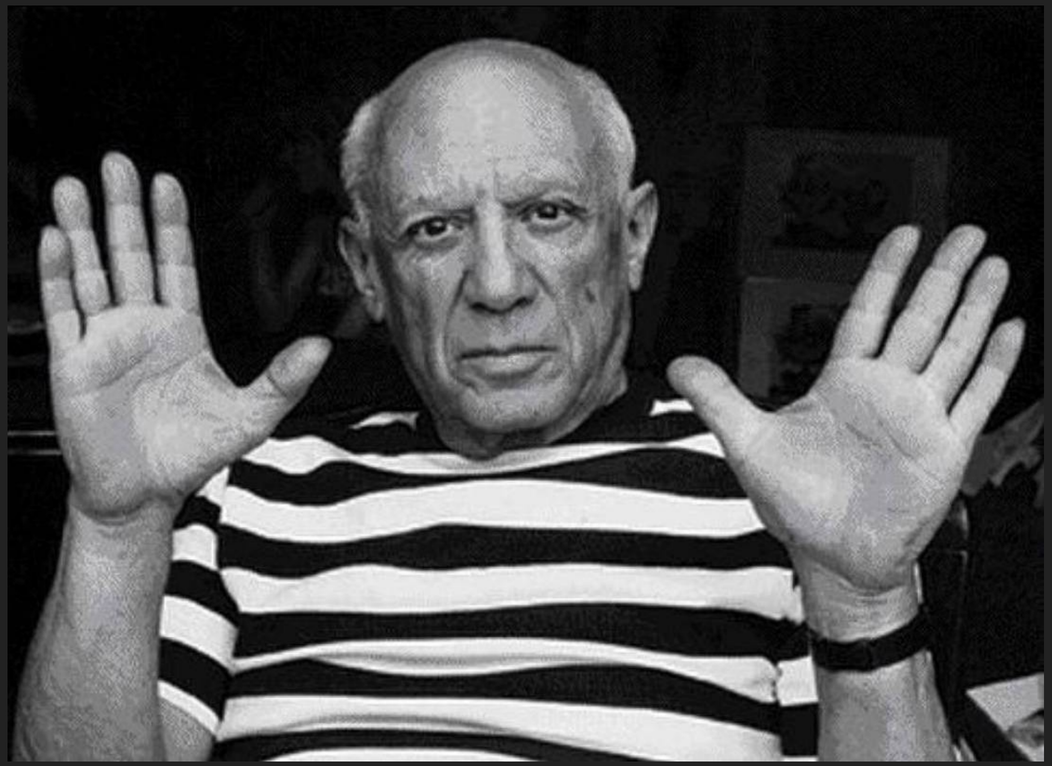
USE OF MACHINE LEARNING IN PHYSICS DATA ANALYSIS

Currently undergoing inflation



Review by Lily Asquith on her Advances in ML and applications to QCD plenary talk

But beware ...



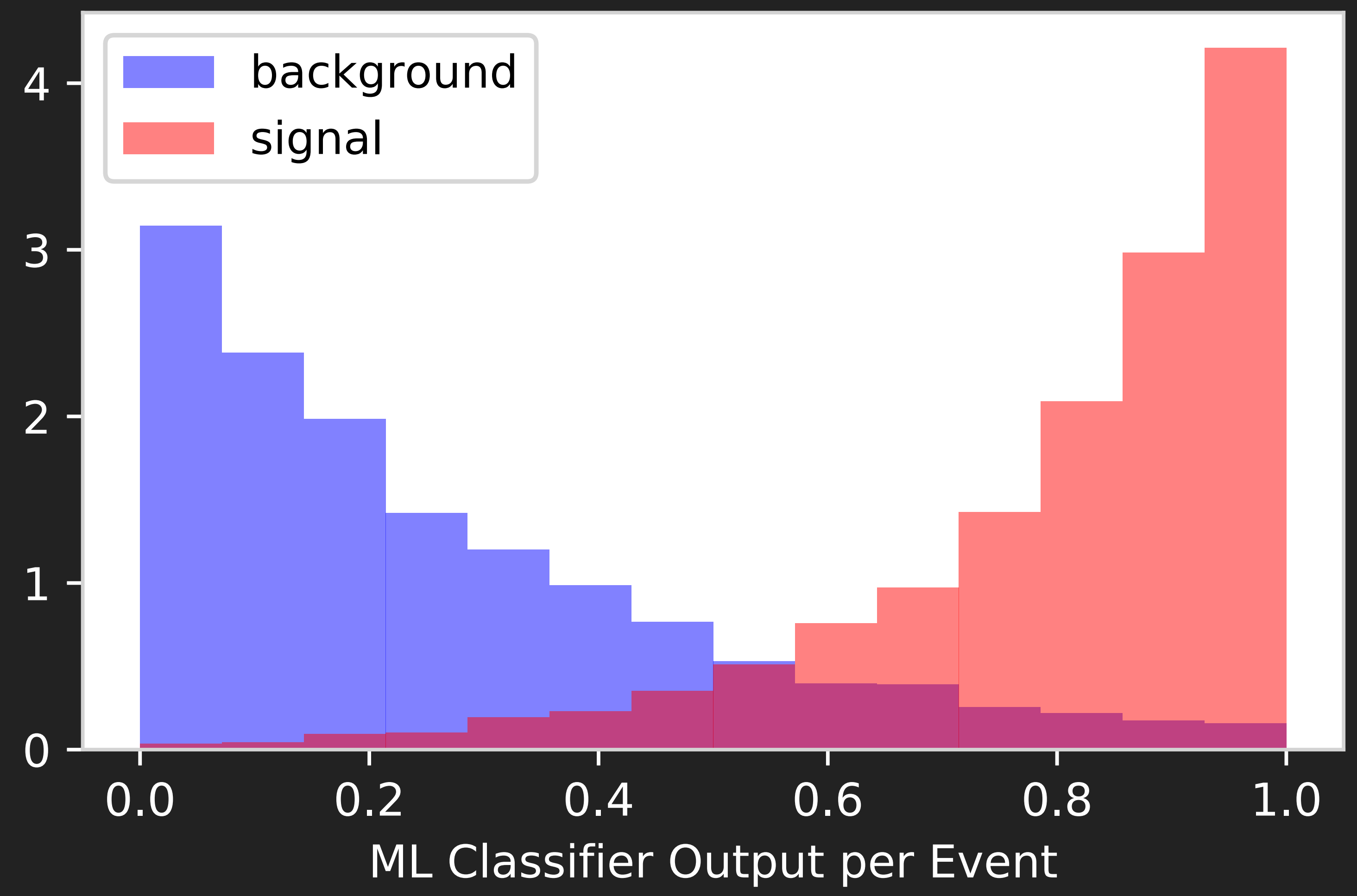
Computers are useless. They can only give you answers.

Pablo Picasso (1960s)

ARE ASKING THE RIGHT QUESTIONS AT THE LHC ANALYSES?

MACHINE LEARNING WITHIN LHC ANALYSES

Most common approach → supervised learning classification trained on simulation



Event-by-Event Signal vs Background

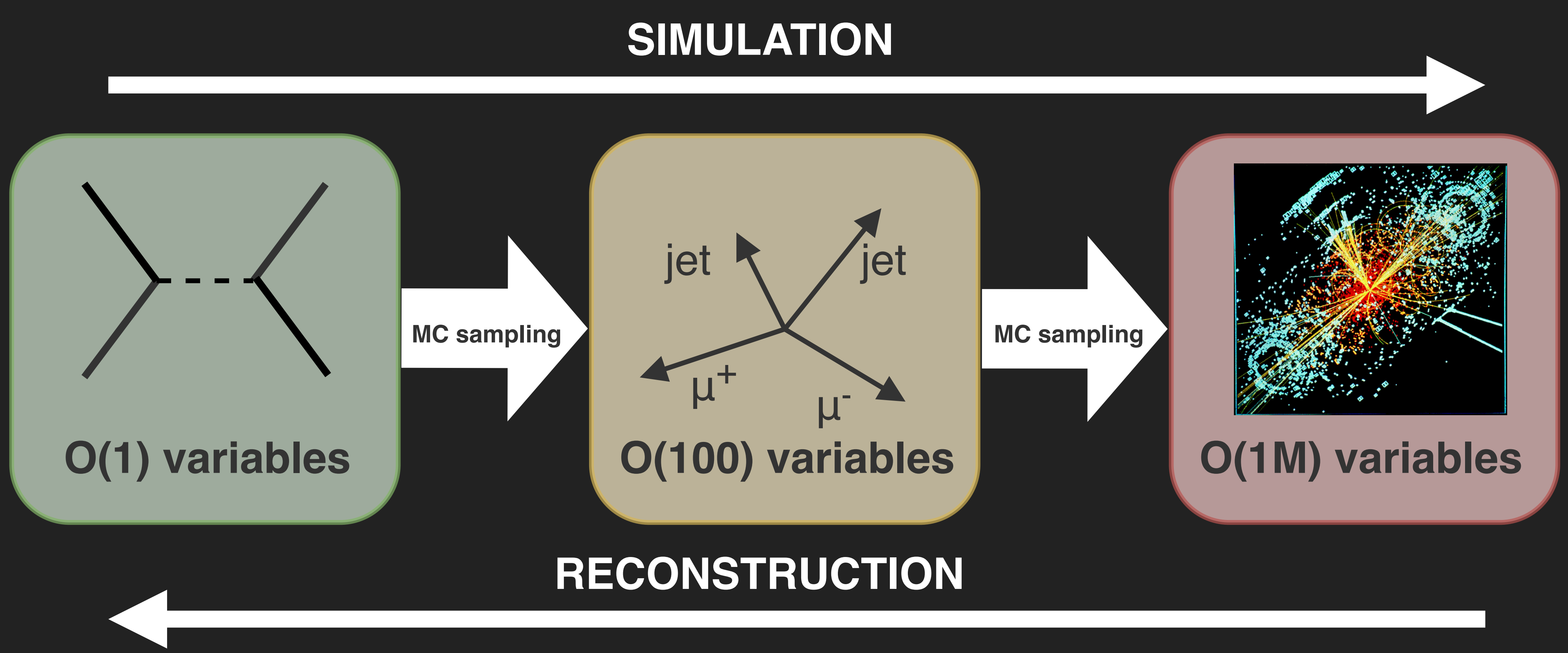
- Higgs Kaggle challenge (2014)
- Almost every other LHC analysis

IS IT REALLY A CLASSIFICATION PROBLEM?

NO! IT IS A STATISTICAL INFERENCE PROBLEM

p(x|model) IS NOT KNOWN AT LHC EXPERIMENTS

Samples under different hypotheses can be simulated via complex physics-based MC programs but $p(\mathbf{x})$ cannot be directly evaluated → **LIKELIHOOD-FREE INFERENCE**



good approximations of $p(\mathbf{x})$ are unachievable due to curse of dimensionality

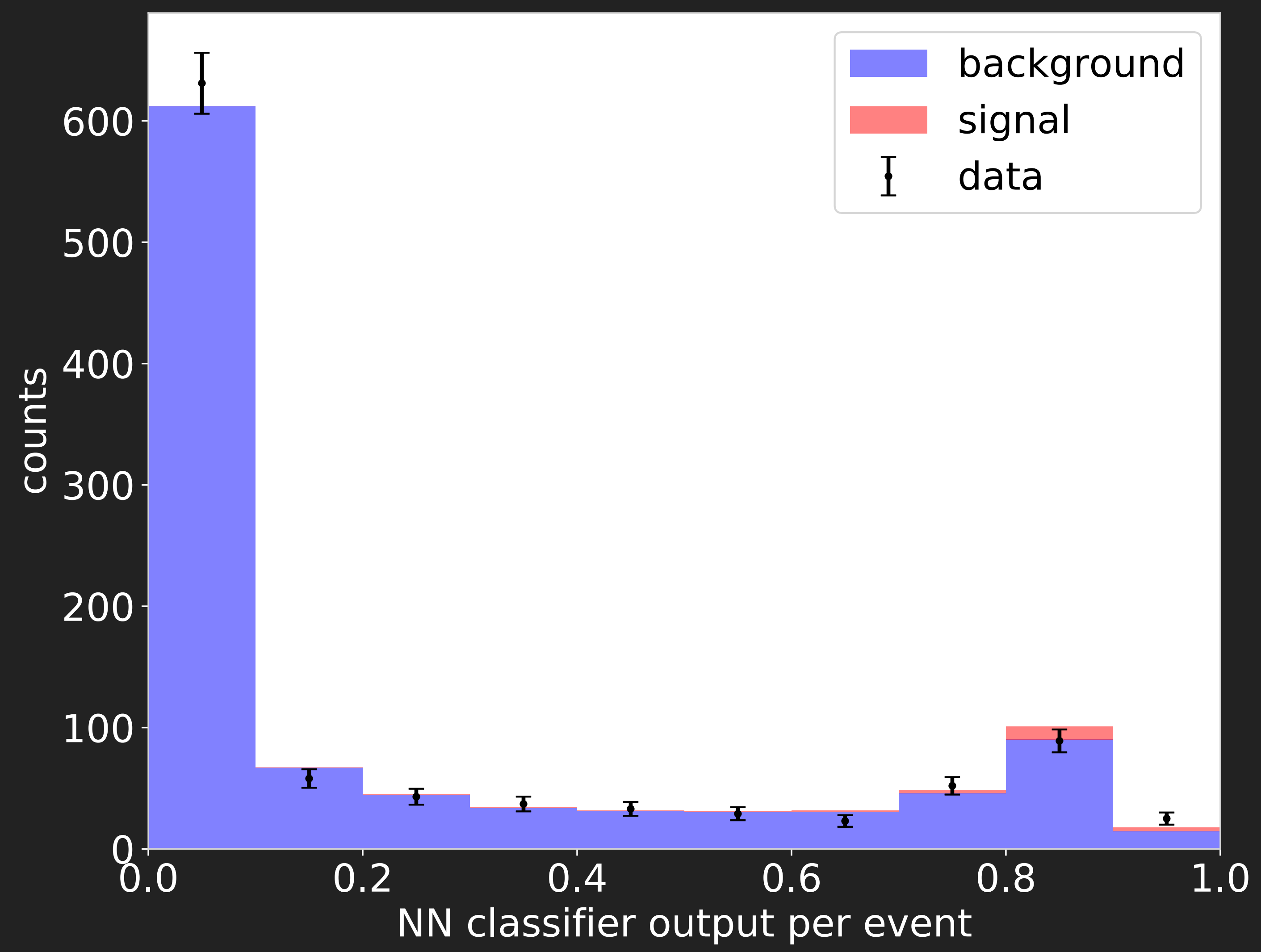
DIMENSIONALITY REDUCTION $\mathbf{R}^{n} \rightarrow \mathbf{R}^{O(1)}$ (SUMMARY STATISTIC)

KEEPING AS MUCH USEFUL INFORMATION FOR INFERENCE AS POSSIBLE

CLASSIFIER-BASED INFERENCE

A ML classifer trained on simulation $d(\mathbf{x})$ is an approximation of $p_s(\mathbf{x})/p_b(\mathbf{x})$

How can it be used for statistical inference from observed data $\mathcal{D}$?



1-D → **cut or histogram to build a Poisson counts non-parametric likelihood**

$$\mathcal{L}(\mu) = \prod_{\mathrm{i} \in \mathrm{bins}} \mathrm{Pois}(n_\mathrm{i} | \mu \cdot s_\mathrm{i} + b_\mathrm{i})$$

which can be used for further inference, such as measuring μ given observed $\mathcal{D}$

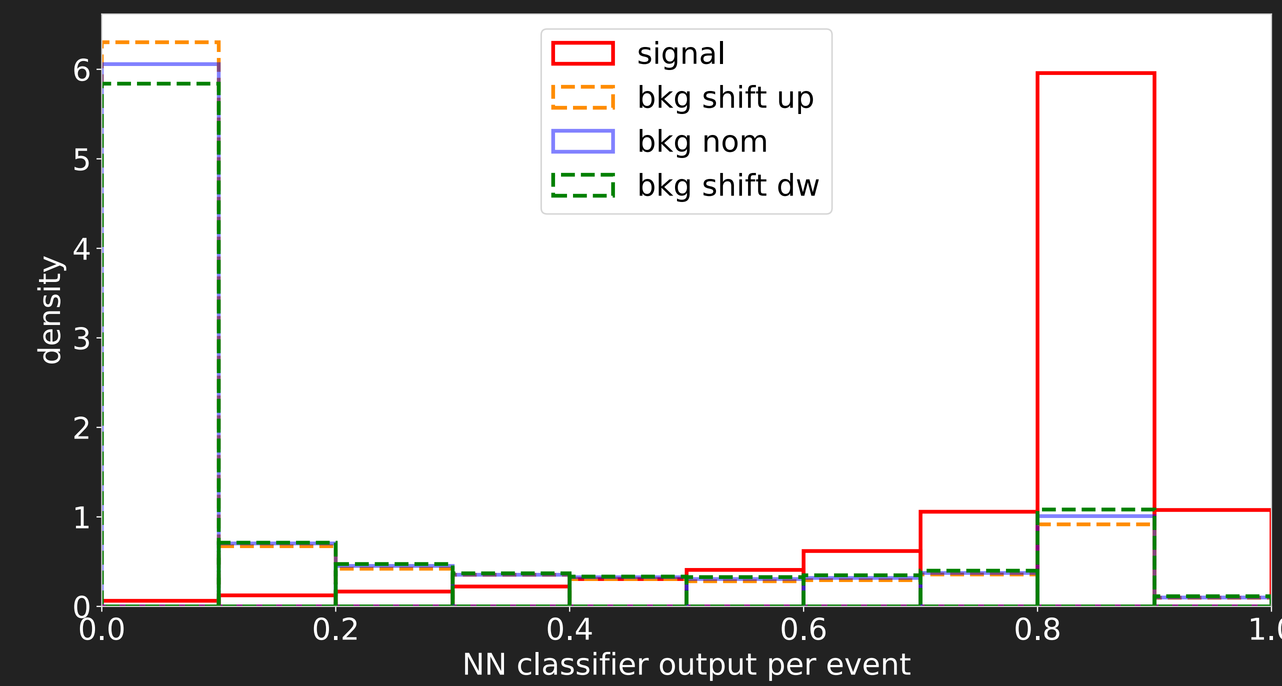
A more direct exploitation of the likelihood ratio approximation for inference is carried out in "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers" by K. Cranmer et al. That work was further extendend in arXiv:1805.12244 (and cited articles therein) by J. Brehmer et al to use also the joint score.

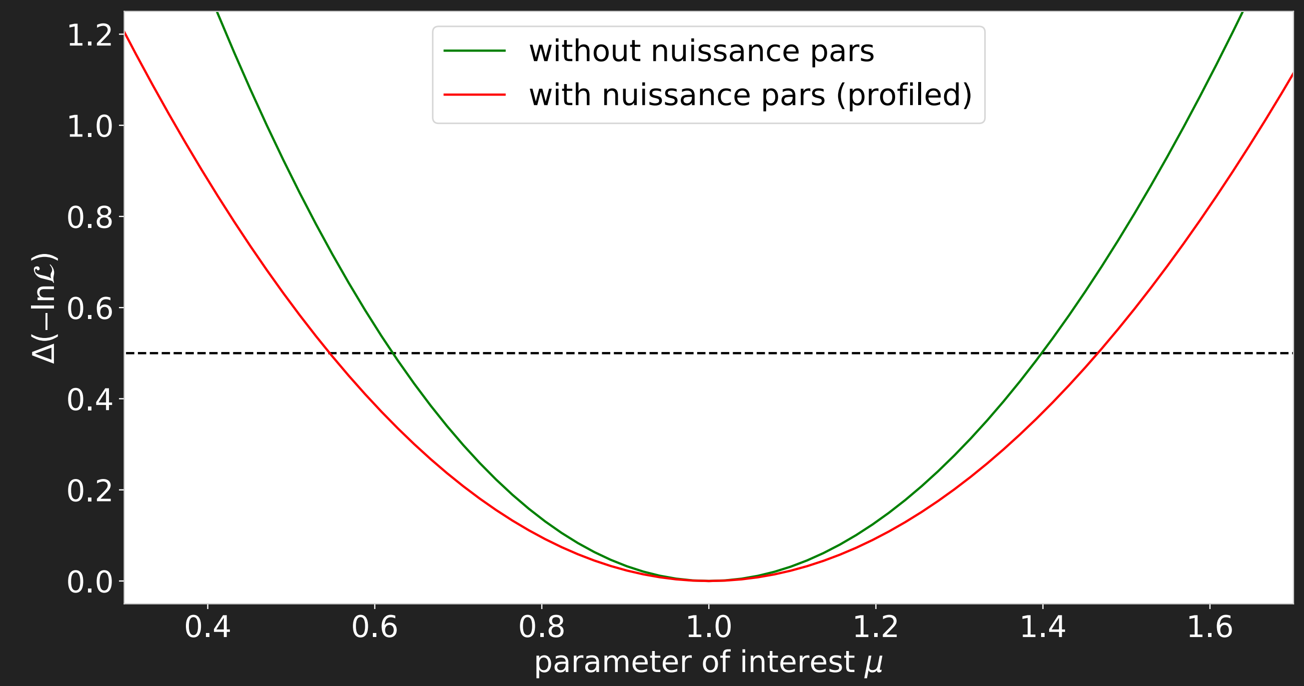
MODELLING UNCERTAINTIES DEGRADE INFERENCE

Simulations are imperfect, mainly due to the limited information of the system being modelled

Lack of knowledge for inference accounted by additional uncertain parameters (nuisance parameters ν)

Causes a degradation of classifier-based inference, leading to larger measurement error → **systematic uncertainties**

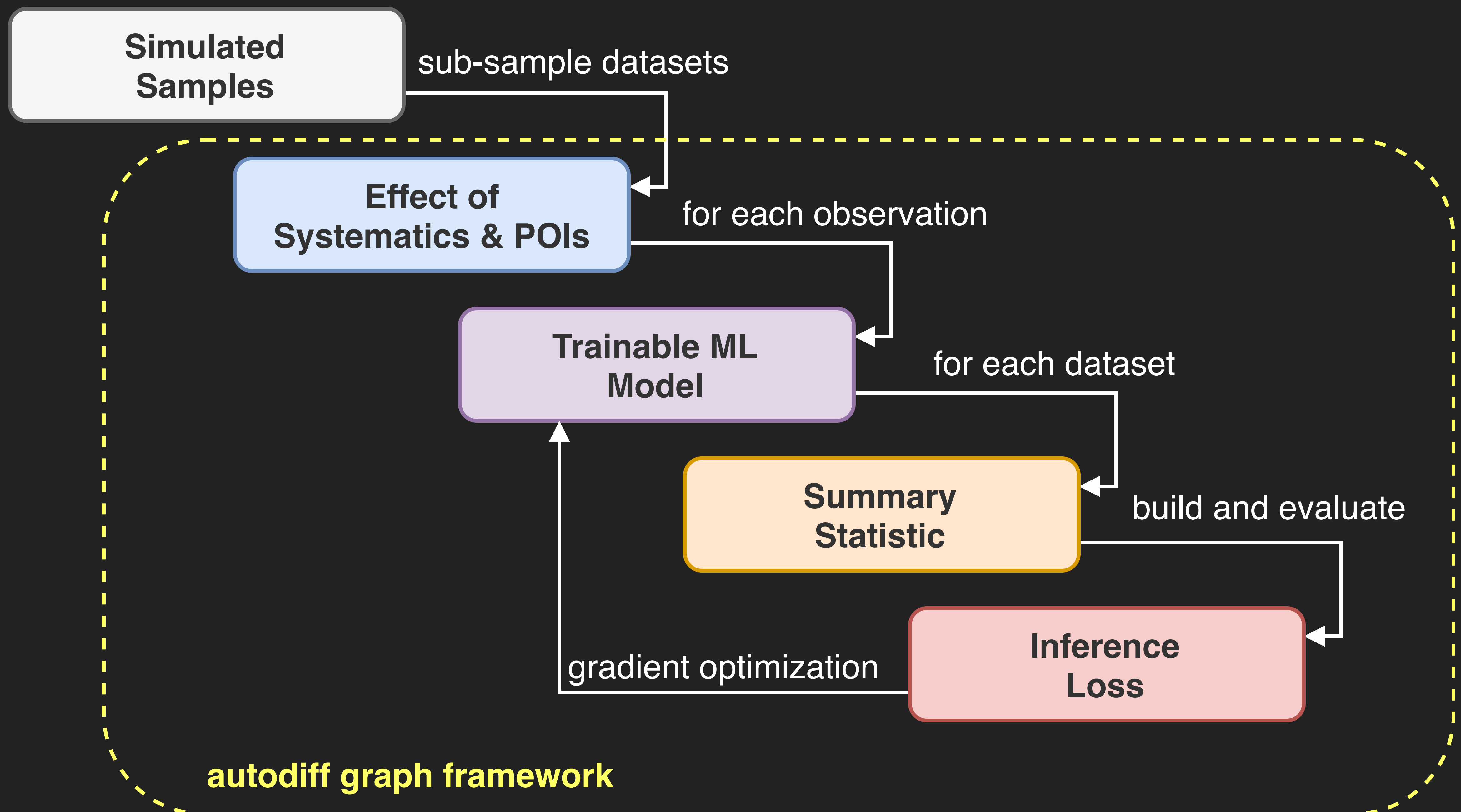




UPPER LIMIT OF ML USEFULNESS IN LHC ANALYSES

Classifiers can be made pivotal as described in "Learning to Pivot" by G. Louppe et al. A review/benchmarks on how to deal with systematics when using machine learning can be found in Adversarial learning to eliminate systematic errors: a case study in High Energy Physics by Victor Estrade et al NIPS2017.

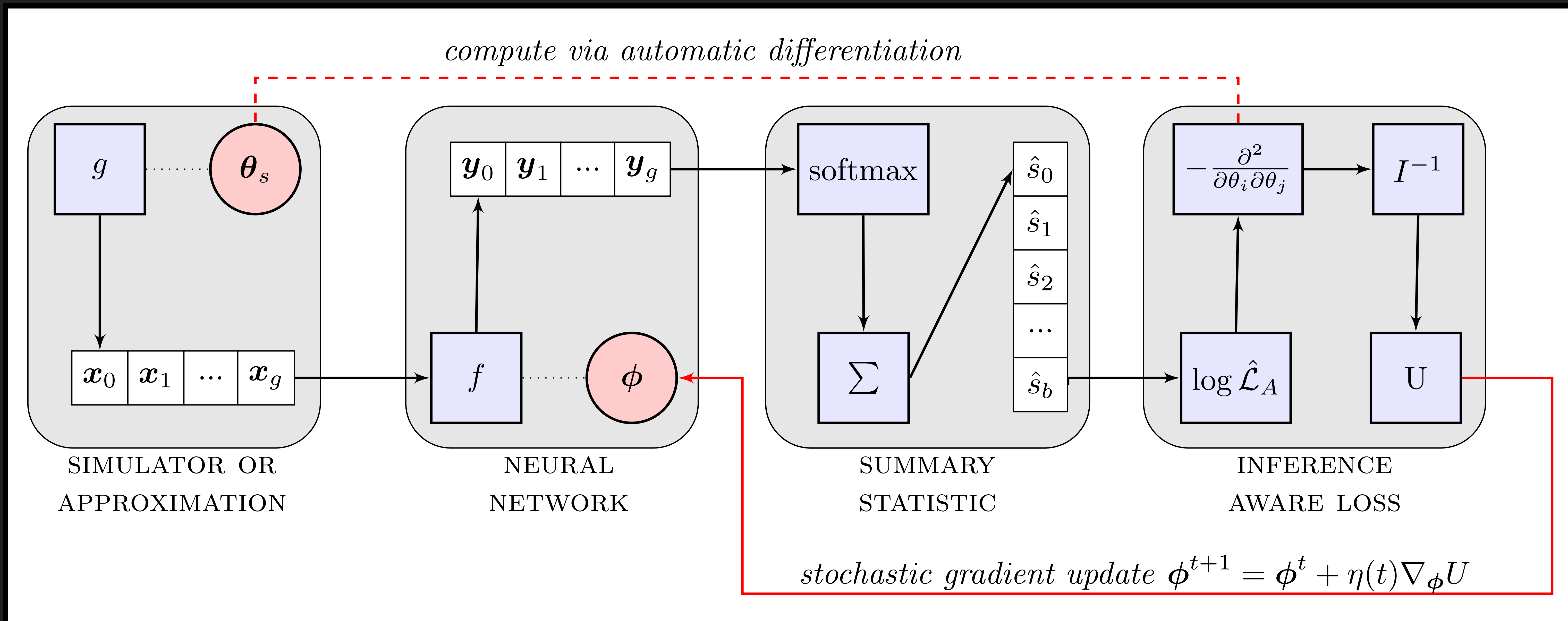
END-TO-END DIFFERENTIABILITY FOR LHC ANALYSES



Within this general framework, several approaches are possible, focus here is

DIRECT LEARNING OF SYSTEMATICS-AWARE SUMMARY STATISTICS

INFERNO: INFERENCE-AWARE NEURAL OPTIMISATION



differentiable approx. covariance matrix of statistical model → SGD optimisation

check arxiv.org/abs/1806.04743 for a detailed mathematical description

INFERENCE-AWARE LOSS FUNCTION

Inference-Aware loss:

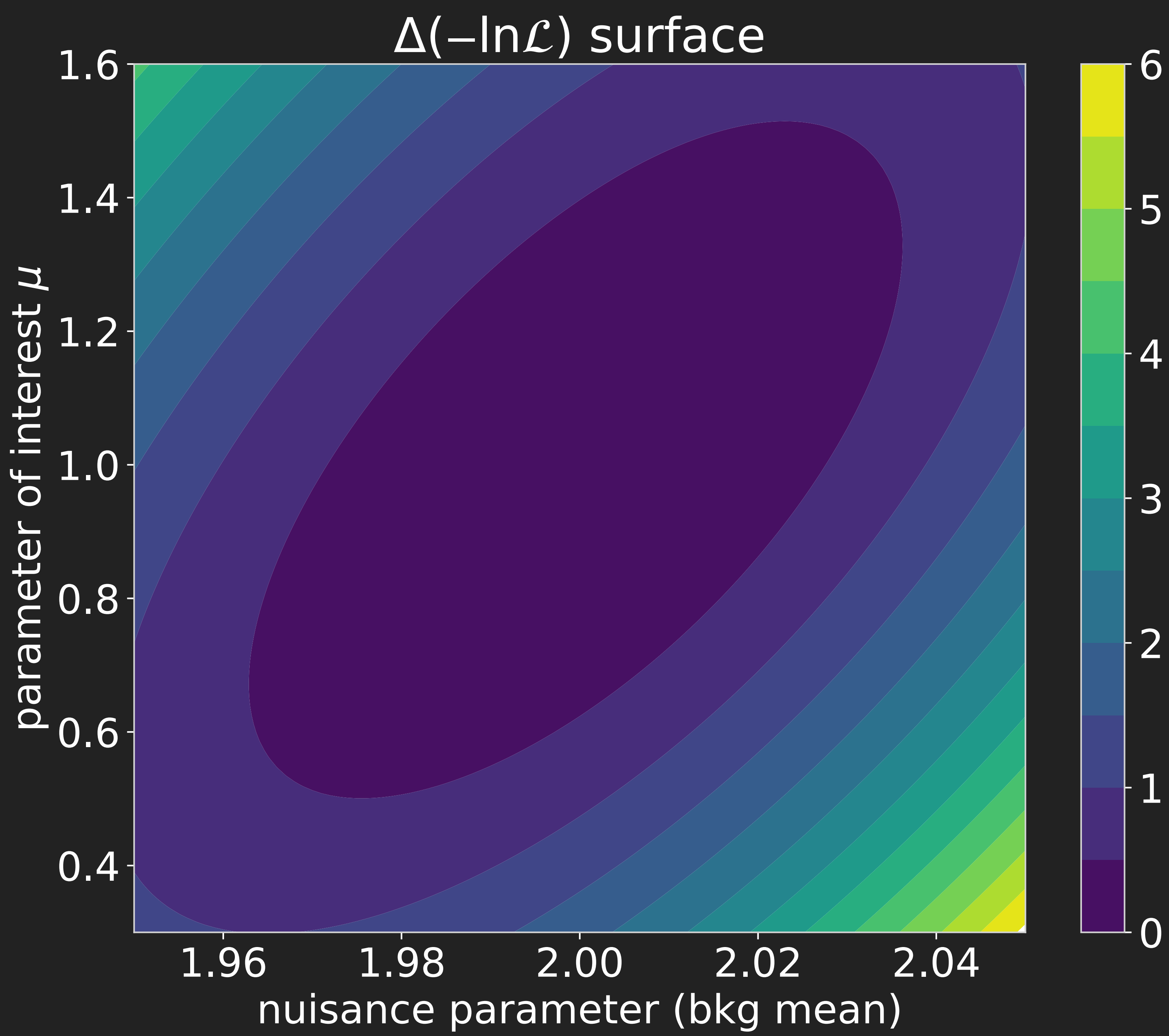
$$\text{loss} \approx \text{Var}(\mu) \quad \text{(expected)}$$

- final analysis figure-of-merit
- accounts for systematics unc.
- extended for other parameters

Classification-based loss:

$$\text{loss} = -\sum_i k_i \log y_i \quad \text{(c. entropy)}$$

- approximates $p_\text{s}(\mathbf{x})/p_\text{b}(\mathbf{x})$
- does not account for systematic unc.



SYNTHETIC EXAMPLE RESULTS → IT WORKS!

Applied on 2D Gaussian two-component mixture toy problem, with an unknown background mean in one of the coordinates → **one nuisance parameter ν**

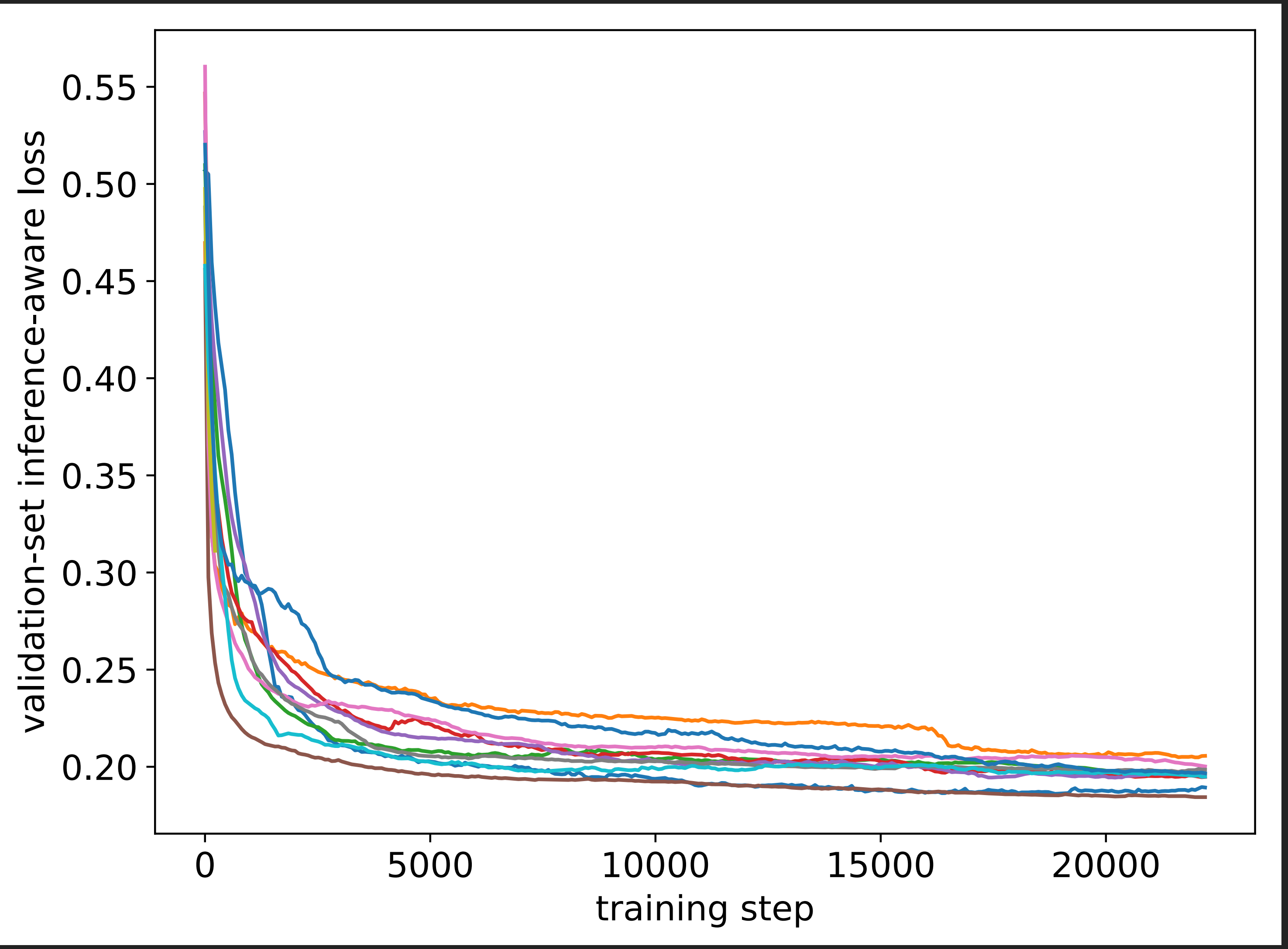
Approach converges consistently to low $\mathrm{Var}(\mu)$ solutions ind. of initialisation

Expected signal strength uncertainties computed using the validation set (average of 10 random initialisations):

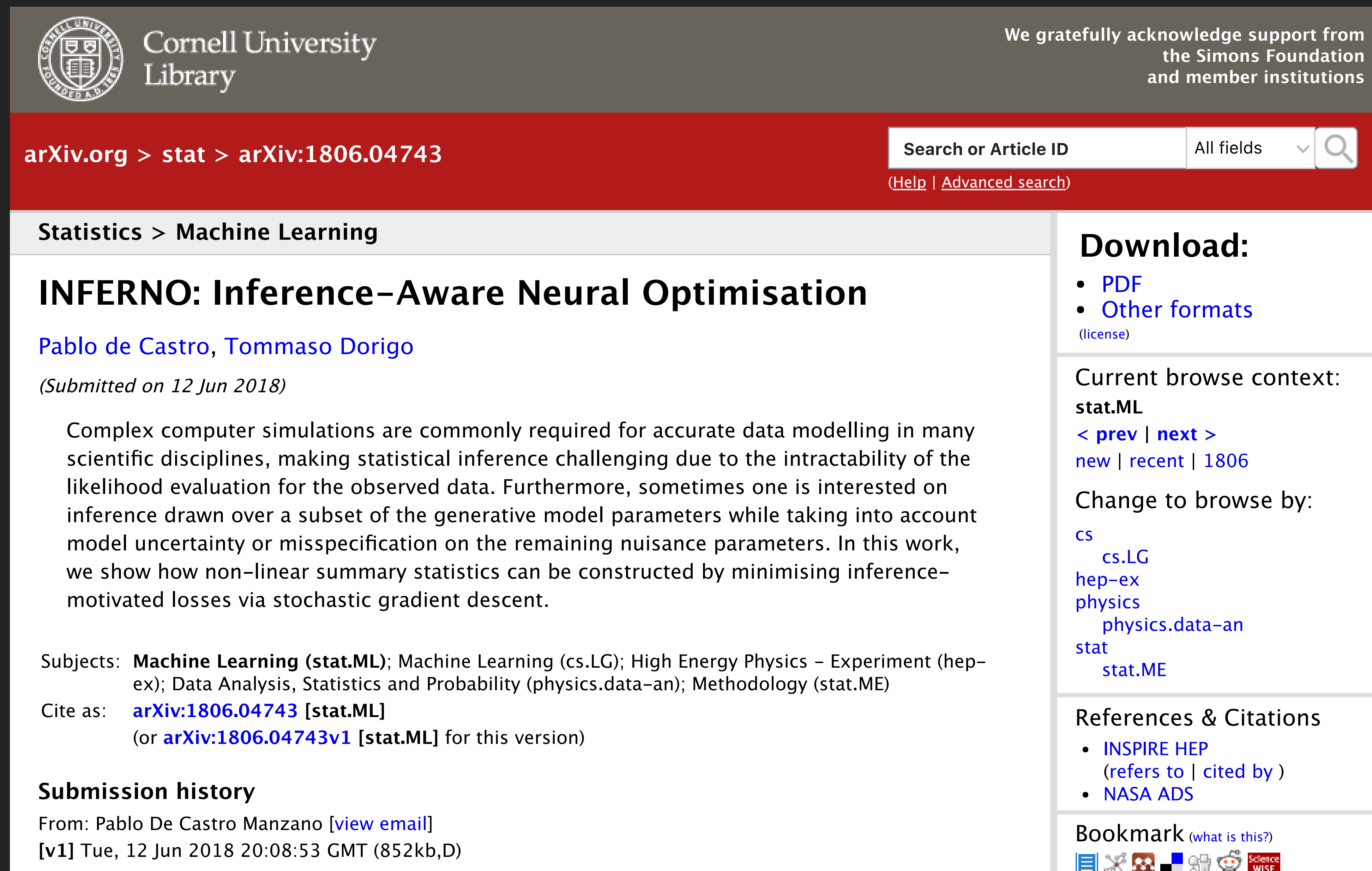
- cross-entropy: 0.444 ± 0.003
- **inference-aware: $\mathbf{0.437 \pm 0.008}$**

Now working on higher-dimensional problems with more nuisance parameters

BETTER/EQUAL THAN CLASSIFICATION



THANK YOU FOR YOUR ATTENTION!



Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > stat > arXiv:1806.04743

Search or Article ID | All fields

(Help | Advanced search)

Statistics > Machine Learning

INFERNO: Inference-Aware Neural Optimisation

Pablo de Castro, Tommaso Dorigo

(Submitted on 12 Jun 2018)

Complex computer simulations are commonly required for accurate data modelling in many scientific disciplines, making statistical inference challenging due to the intractability of the likelihood evaluation for the observed data. Furthermore, sometimes one is interested on inference drawn over a subset of the generative model parameters while taking into account model uncertainty or misspecification on the remaining nuisance parameters. In this work, we show how non-linear summary statistics can be constructed by minimising inference-motivated losses via stochastic gradient descent.

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG); High Energy Physics - Experiment (hep-ex); Data Analysis, Statistics and Probability (physics.data-an); Methodology (stat.ME)

Cite as: arXiv:1806.04743 **[stat.ML]**
(or arXiv:1806.04743v1 **[stat.ML]** for this version)

Submission history

From: Pablo De Castro Manzano [view email]
[v1] Tue, 12 Jun 2018 20:08:53 GMT (852kb,D)

Download:

- PDF
- Other formats

(license)

Current browse context:
stat.ML
< prev | next >
new | recent | 1806

Change to browse by:
cs
cs.LG
hep-ex
physics
physics.data-an
stat
stat.ME

References & Citations

- INSPIRE HEP
(refers to | cited by)
- NASA ADS

Bookmark (what is this?)

if interested on technique, more details on preprint arxiv.org/abs/1806.04743
feedback is greatly welcomed (DM @pablodecm or pablo.decastro@cern.ch)