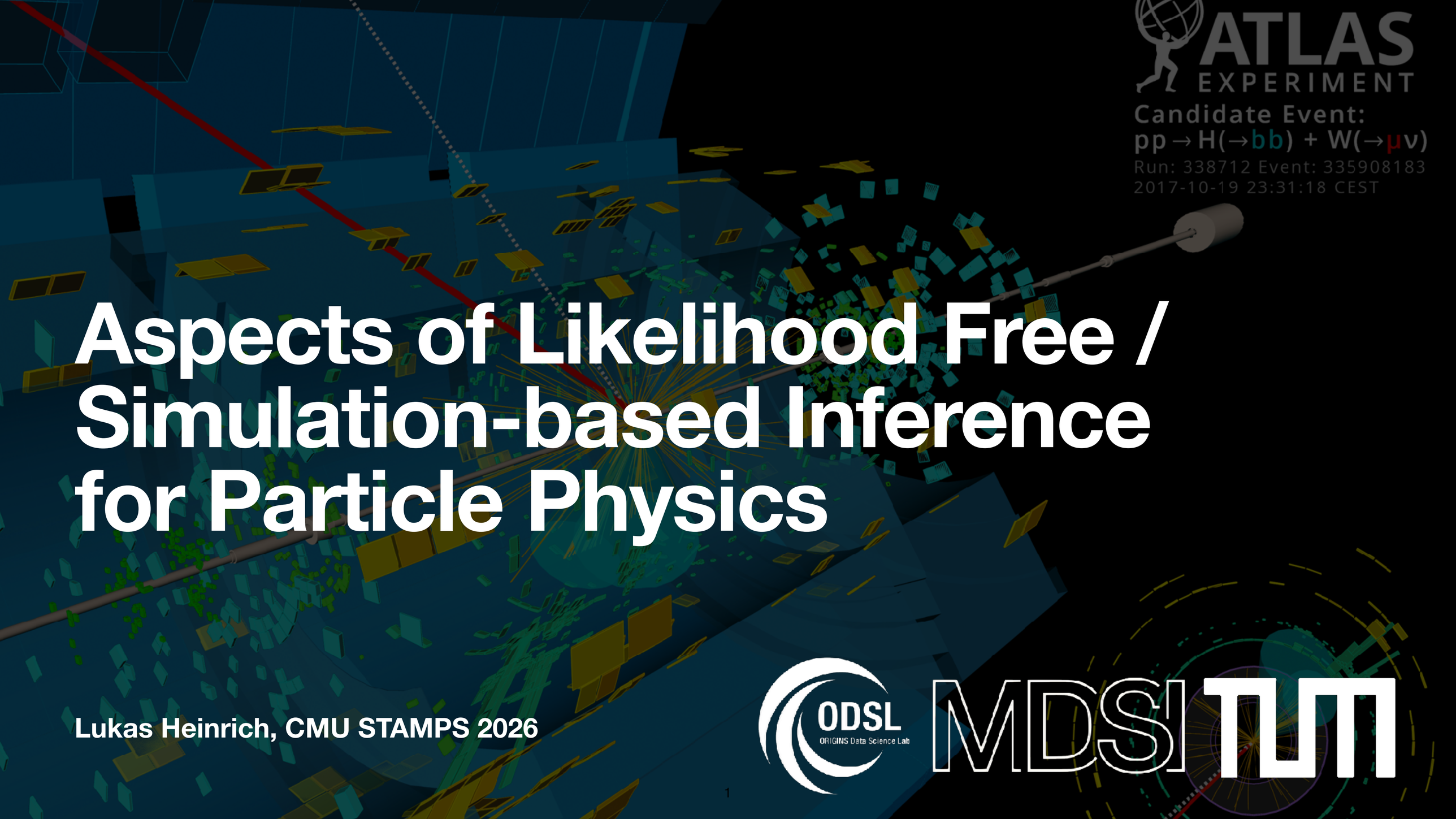


Aspects of Likelihood Free / Simulation-based Inference for Particle Physics

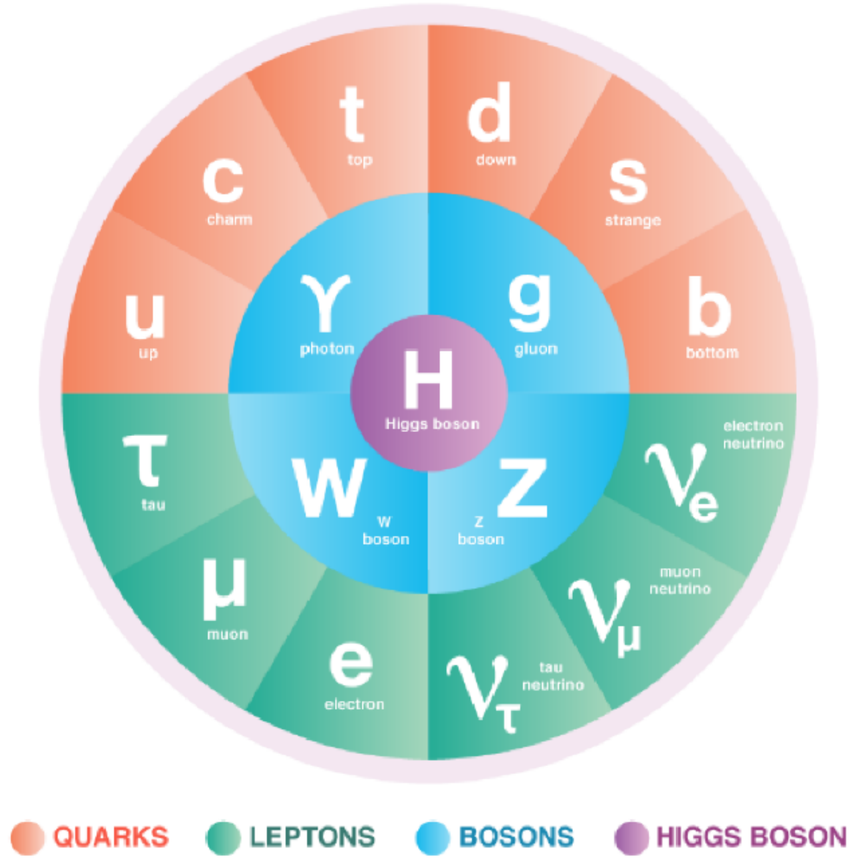
Lukas Heinrich, CMU STAMPS 2026

ODSL
ORIGINS Data Science Lab

MDSI

TUM

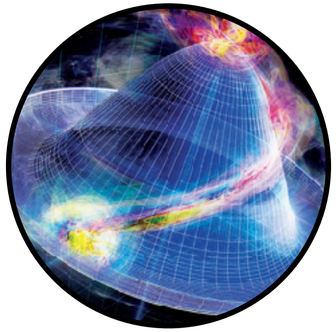
My Physics Context



Dark Matter

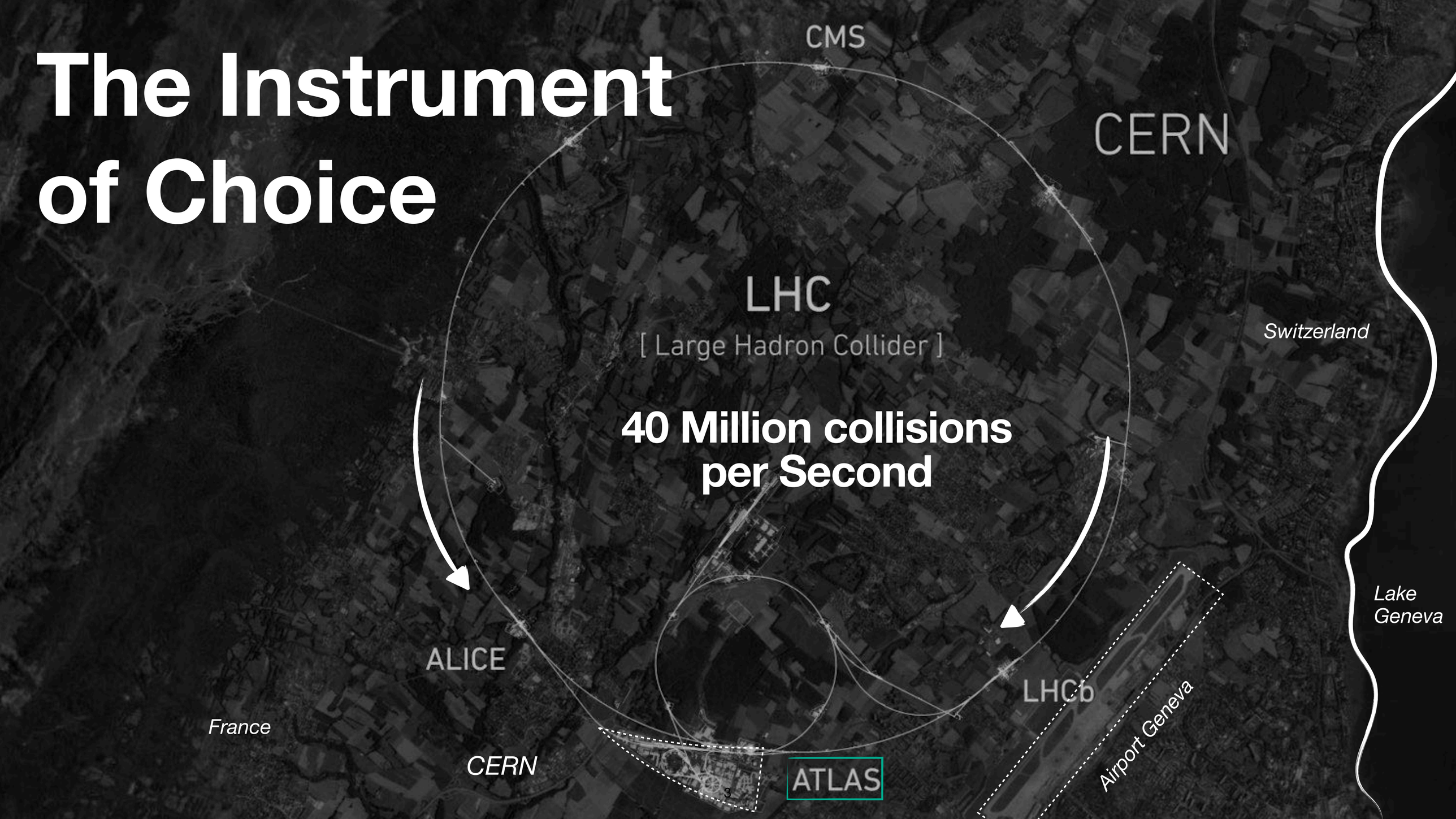
Nature of Neutrinos

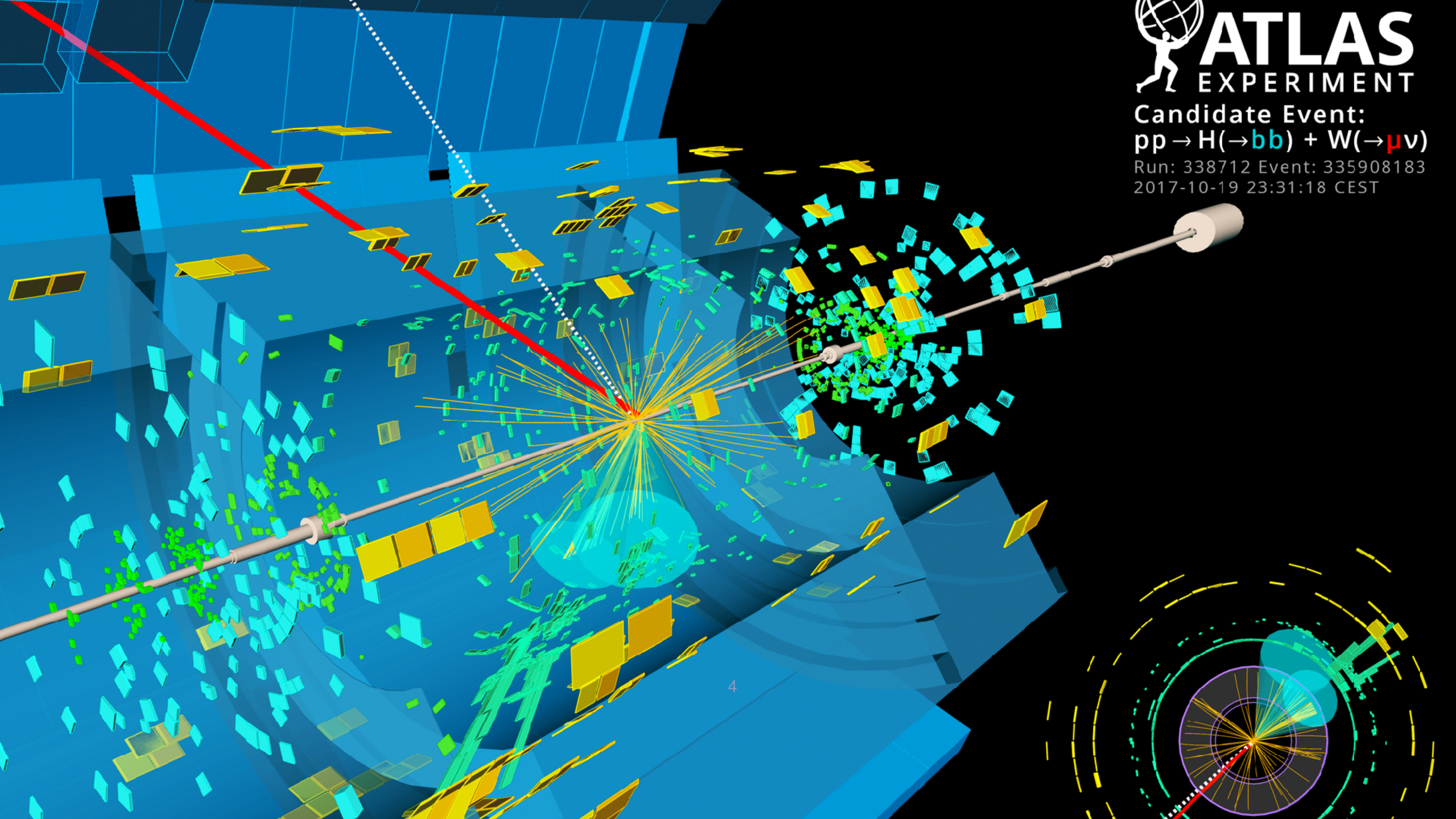
Matter & Antimatter



Origin of Mass

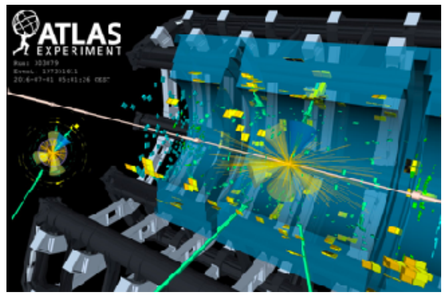
The Instrument of Choice

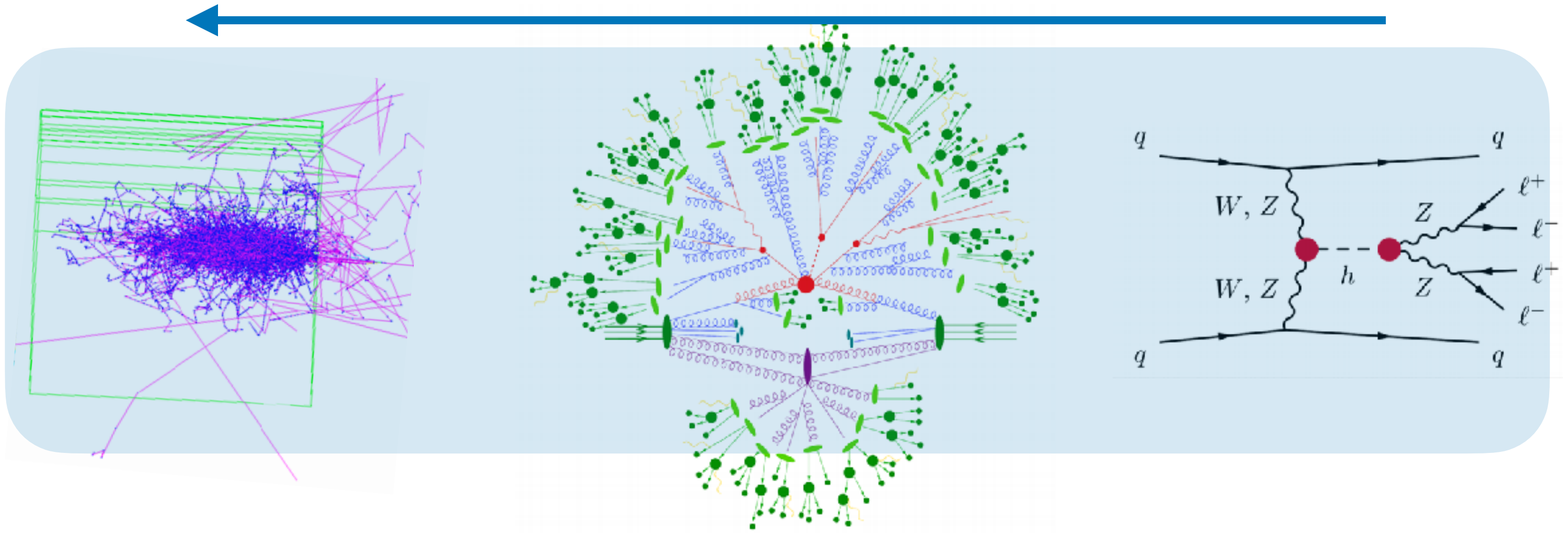



ATLAS
EXPERIMENT
Candidate Event:
pp → H(→bb) + W(→μν)
Run: 338712 Event: 335908183
2017-10-19 23:31:18 CEST

HEP is Likelihood-Free

The data-generating process in particle physics is defined by a long sequence of complicated stochastic transitions





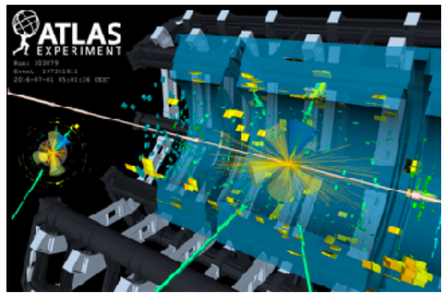
x

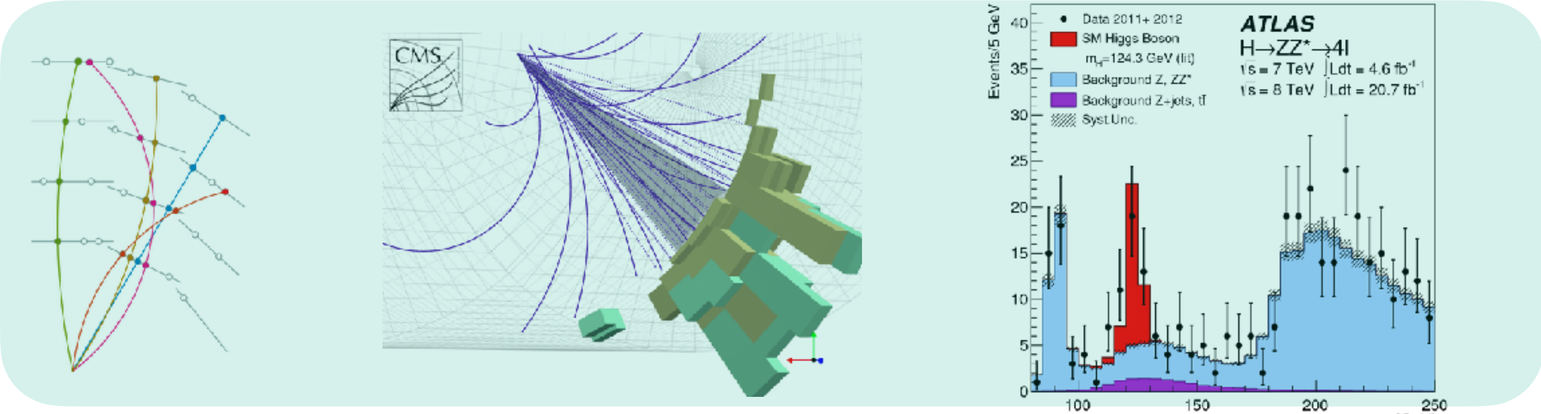
θ

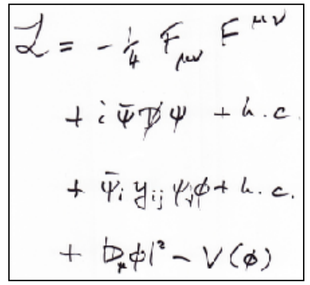
$$p(x|\theta) = \int dz \;\; p(x|z_h)\; p(z_h|z_p)\; p(z_p|\theta)$$

HEP is Likelihood-Free

Traditionally, inference is done by a series of per-event manually derived summary statistics







x

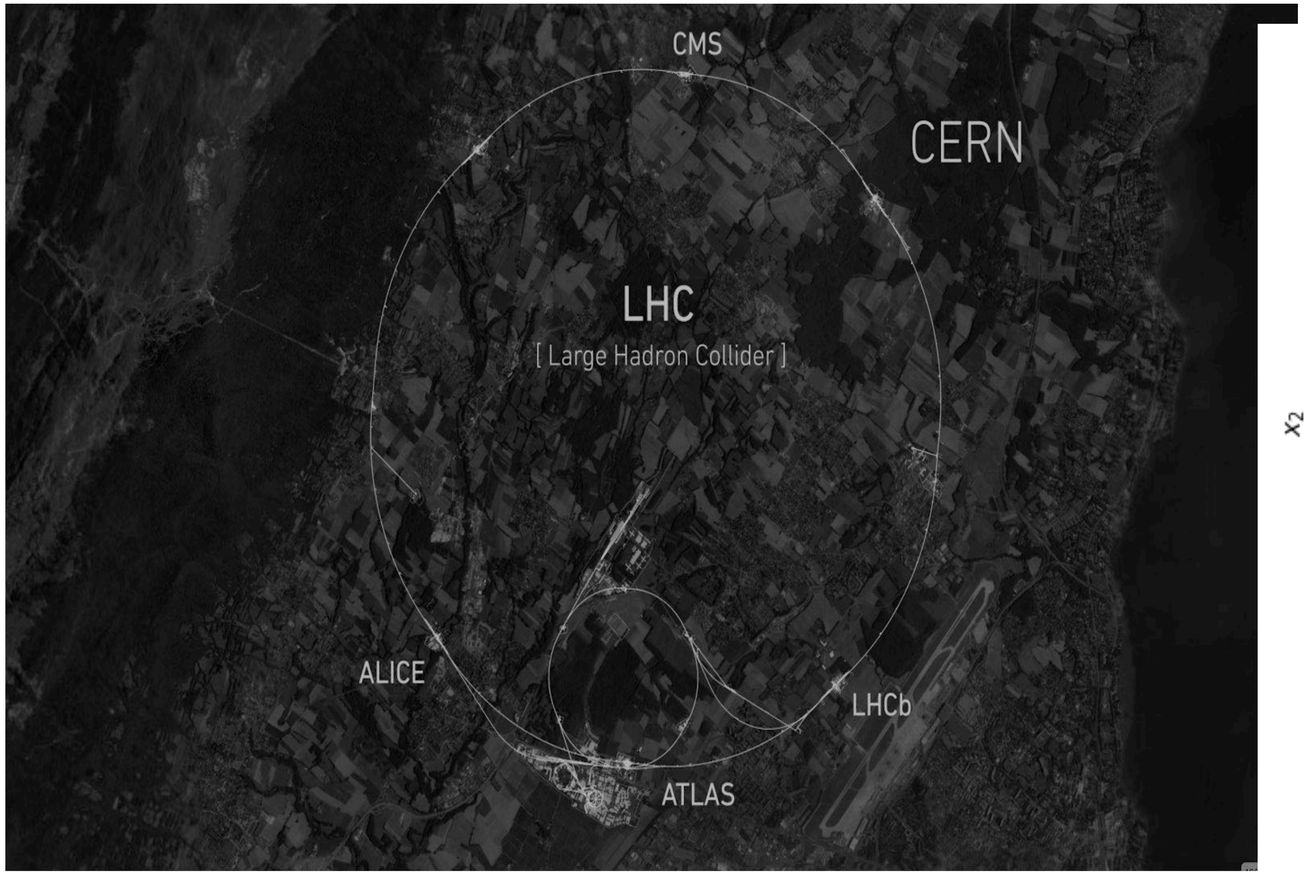
θ

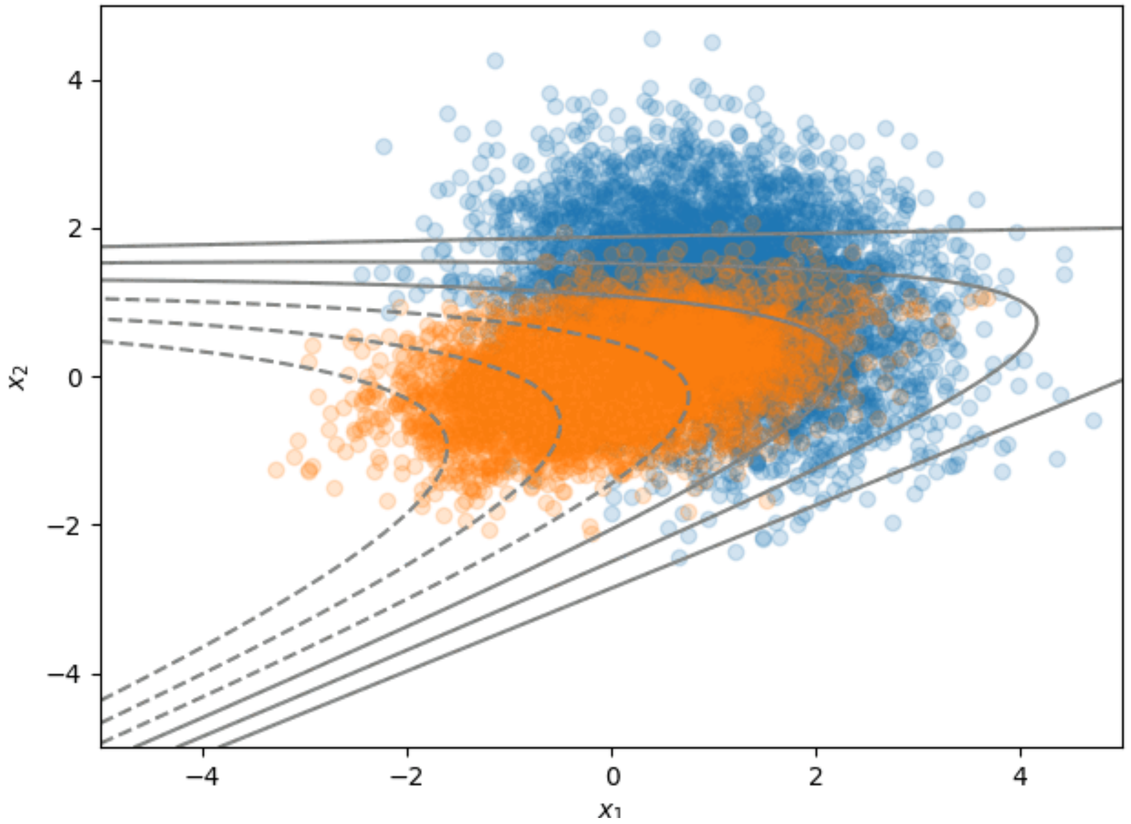
approximate inference: $p(\theta \mid x) \approx p(\theta \mid \lambda_\phi(x))$

what's the best way to compress the data?

ML-driven SBI

More recently, ML techniques have changed how we think about this simulation-based inference workflow

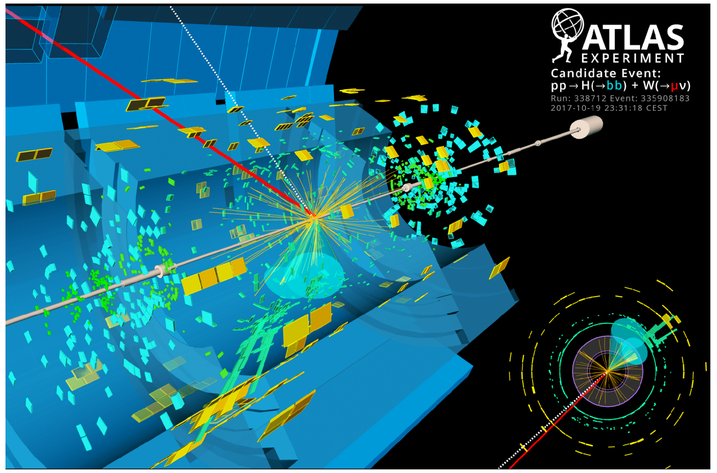




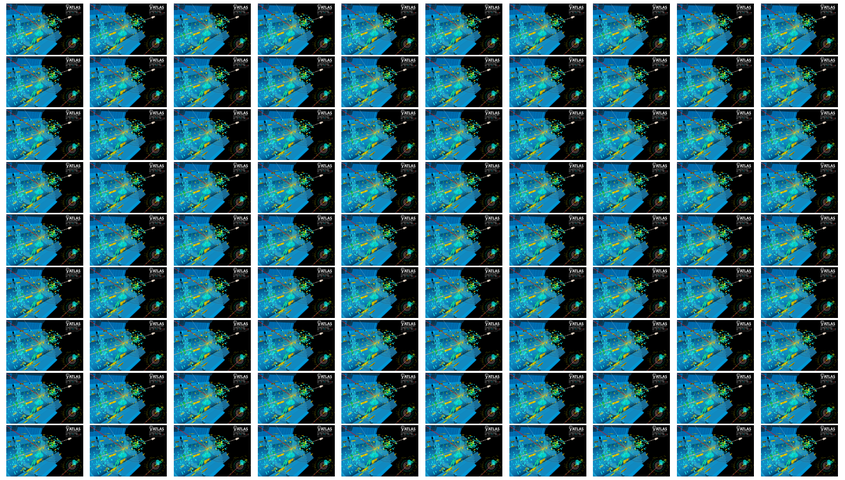
Likelihood-Ratio Trick

Three Types of High Dimension

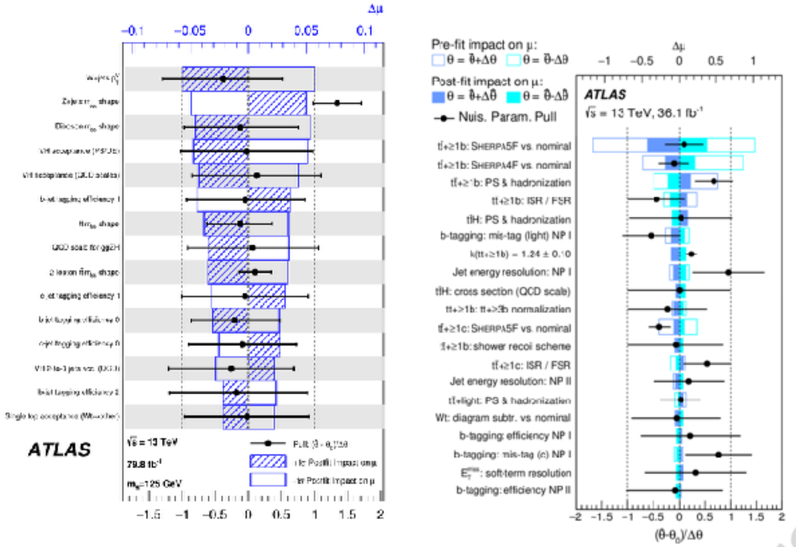
A catch: the data and parameters in particle physics is high-dim. in three different ways, which impacts the SBI workflow



high-dimensional data



extremely large datasets



large(ly nuisance) parameter spaces

$x \in \mathbb{R}^{10^8}$ $\mathscr{D} = \{x_1, \ldots, x_k\},\ k = 10^3, 10^6, 10^9, \ldots$ $\theta \in \mathbb{R}^{10^{2-3}}$ $\mu \in \mathbb{R}^{1-2}$

Three Types of High Dimension

In particle physics we want to do inference on the full set of particle events we record (potentially billions!)

$$p(\theta \mid \mathscr{D})$$

Bayes

$$\frac{p(\mathscr{D}) \mid \theta)}{p(\mathscr{D} \mid \hat{\theta})}$$

Frequentist

$$\mathscr{D} = \{x_1, x_2, x_3 \ldots x_n\}$$

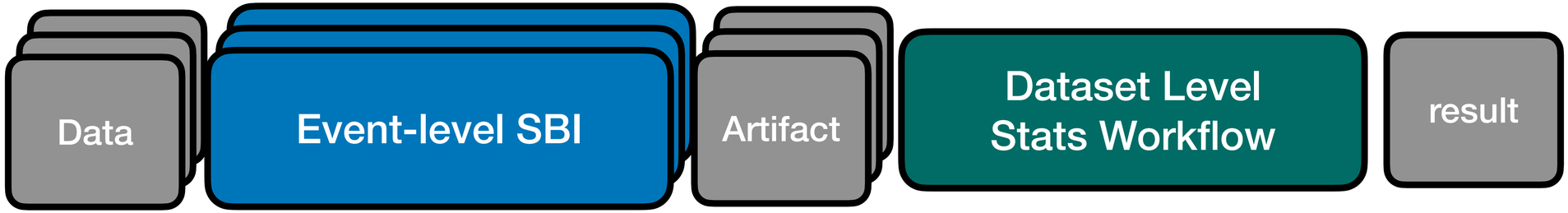
where we assume the data emerges from an i.i.d process

$$x_i \in \mathscr{D} \quad x_i \sim p(x \mid \theta)$$

potentially set cardinality depends on θ (in HEP-speak "extended likelihood")

A stymied situation

The SBI workflow in HEP was therefore never really fully “end-to-end” but always had a “Traditional Stats” component



vs.

Example

1) Train an event-level L'hood-ratio estimator $r_\phi(x, \theta) = \dfrac{p(x\,|\,\theta)}{p(x\,|\,\theta_0)}$

2) Full-dataset likelihood ratio

Important:
exploiting the fact that
Lhood ratios compose

$$R(\theta) = \frac{p(\mathscr{D}\,|\,\theta)}{p(\mathscr{D}\,|\,\theta_0)} = \prod_i \frac{p(x_i\,|\,\theta)}{p(x\,|\,\theta_0)}$$

3) Run inference (MCMC, MLE, …) on the dataset-level $R(\theta)$

Can we get to a fully end-to-end setup?

Another Problem

A lot of inference targets the subset of the full parameter space: parameters of interest vs nuisance parameters

$$p(\mu \mid \mathscr{D}) = \int_{\theta \neq \mu} p(\theta \mid \mathscr{D})$$

Bayes

$$\frac{p(\mathscr{D} \mid \mu, \hat{\hat{\nu}})}{p(\mathscr{D} \mid \hat{\mu}, \hat{\nu})} = \frac{\sup_{\theta \neq \mu} p(\ldots)}{\sup_{\theta} p(\ldots)}$$

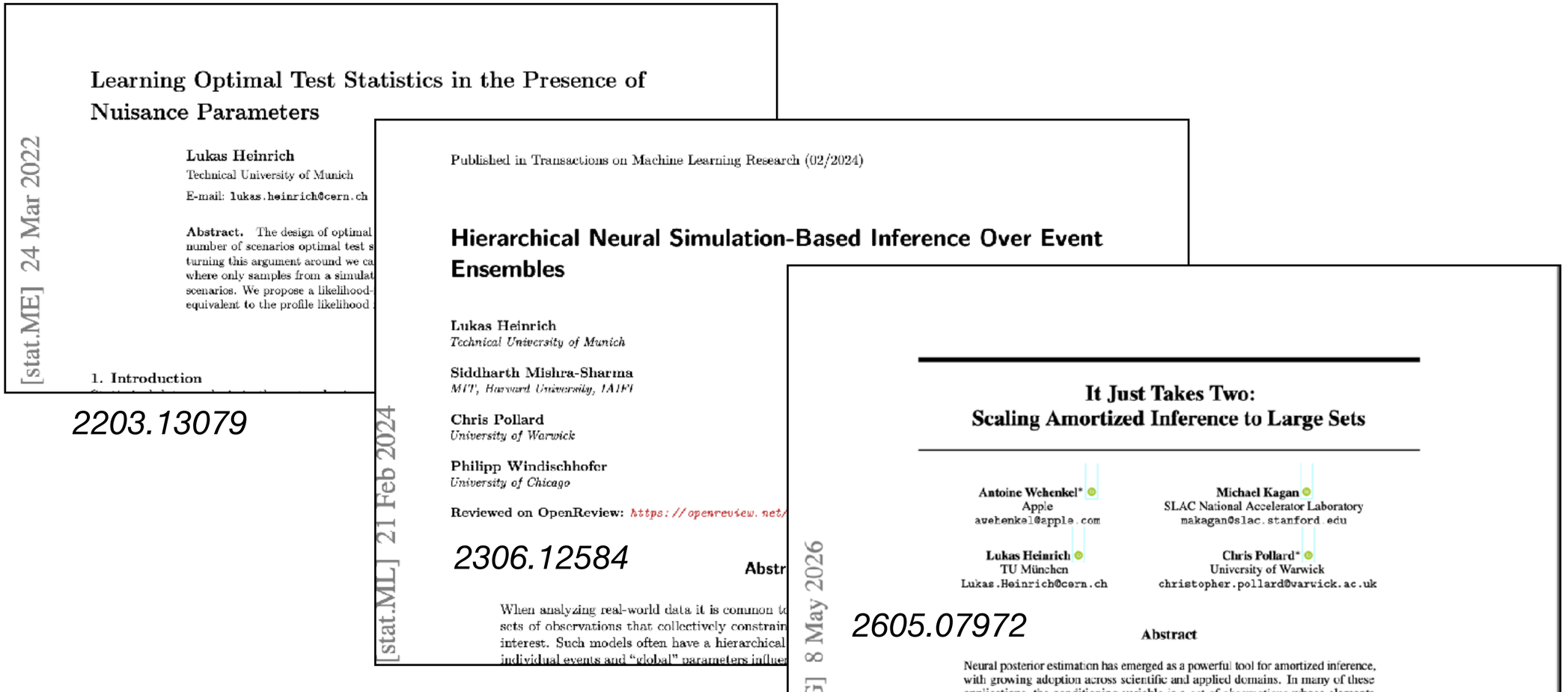
Frequentist

But to compose full-dataset likelihood-(ratio) we need to train SBI methods on the full parameter space → **expensive!**

profile likelihoods / marginal posteriors do not compose

Dataset-level SBI

With data-set level SBI you can target the **low-dimensional** inference result directly & quickly without intermediate steps



Learning Optimal Test Statistics in the Presence of Nuisance Parameters

Lukas Heinrich
Technical University of Munich
E-mail: lukas.heinrich@cern.ch

Abstract. The design of optimal number of scenarios optimal test s turning this argument around we ca where only samples from a simulat scenarios. We propose a likelihood- equivalent to the profile likelihood

1. Introduction

[stat.ME] 24 Mar 2022

2203.13079

Published in Transactions on Machine Learning Research (02/2024)

Hierarchical Neural Simulation-Based Inference Over Event Ensembles

Lukas Heinrich
Technical University of Munich

Siddharth Mishra-Sharma
MIT, Harvard University, IAIFI

Chris Pollard
University of Warwick

Philipp Windischhofer
University of Chicago

Reviewed on OpenReview: https://openreview.net/

Abstr

When analyzing real-world data it is common t sets of observations that collectively constrain interest. Such models often have a hierarchical individual events and "global" parameters influe

[stat.ML] 21 Feb 2024

2306.12584

It Just Takes Two:
Scaling Amortized Inference to Large Sets

Antoine Wehenkel*
Apple
awehenkel@apple.com

Michael Kagan
SLAC National Accelerator Laboratory
makagan@slac.stanford.edu

Lukas Heinrich
TU München
Lukas.Heinrich@cern.ch

Chris Pollard*
University of Warwick
christopher.pollard@warwick.ac.uk

Abstract

Neural posterior estimation has emerged as a powerful tool for amortized inference, with growing adoption across scientific and applied domains. In many of these

[stat.G] 8 May 2026

2605.07972

Example: Learning to Profile

The likelihood ratio trick is a simple argument

> *The likelihood ratio is the optimal statistic (a function) for hypo-testing, so function optimization (ML) for discrimination power must converge to it*

The low-dimensional equivalent is the profile likelihood ratio

> *If we can find an argument why / how the profile likelihood statistic (a function) is optimal, we can find it with machine learning!*

Example: Learning to Profile

Wald (1943) showed that there is a sense in which the profile likelihood is optimal (and thus findable with optimization)

TESTS OF STATISTICAL HYPOTHESES CONCERNING SEVERAL PARAMETERS WHEN THE NUMBER OF OBSERVATIONS IS LARGE(1)

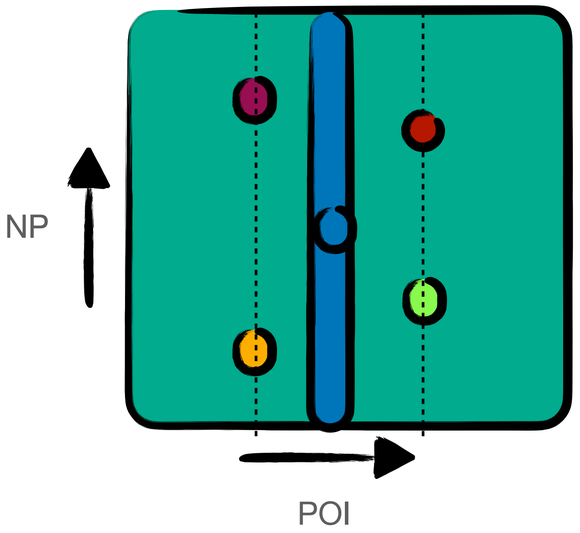
BY

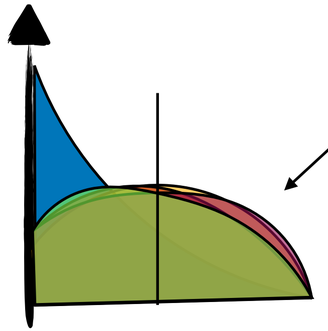
ABRAHAM WALD

TABLE OF CONTENTS

1. **Introduction.** In this paper we shall deal with the following general problem: Let $f(x^1, x^2, \cdots, x^r, \theta^1, \cdots, \theta^k)$ be the joint probability density func-

*pLR has **best average power** for hypos equidistant to a given POI value*



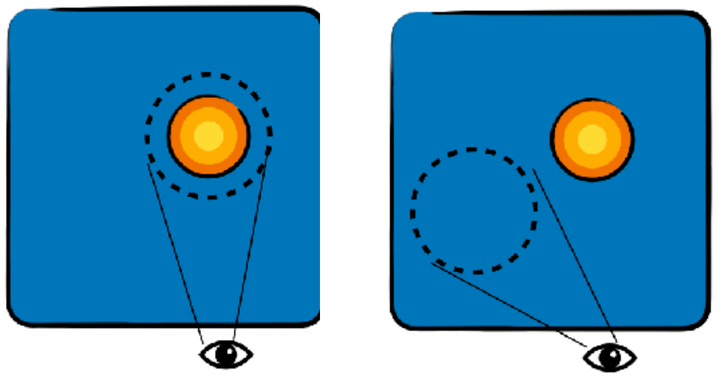


$$t(x|\mu) = -2\log\frac{p(x|\mu,\hat{\hat{\nu}})}{p(x|\hat{\mu},\hat{\nu})}$$

Example: Learning to Profile

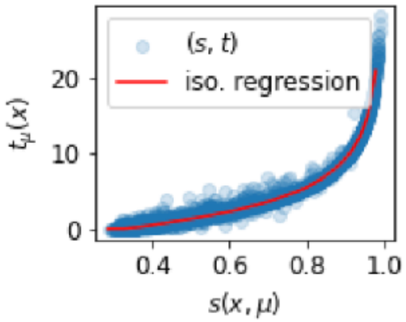
For counting experiments, this was easy to show, i.e. sets where we ~only consider the cardinality!

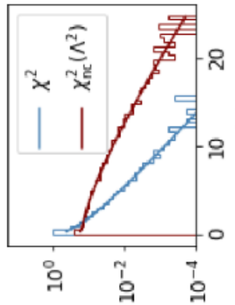
$$f(\mathscr{D}) \to c_1, c_2, \ldots c_n$$

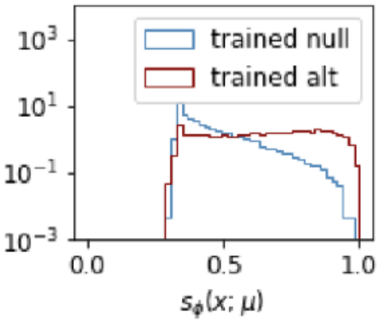


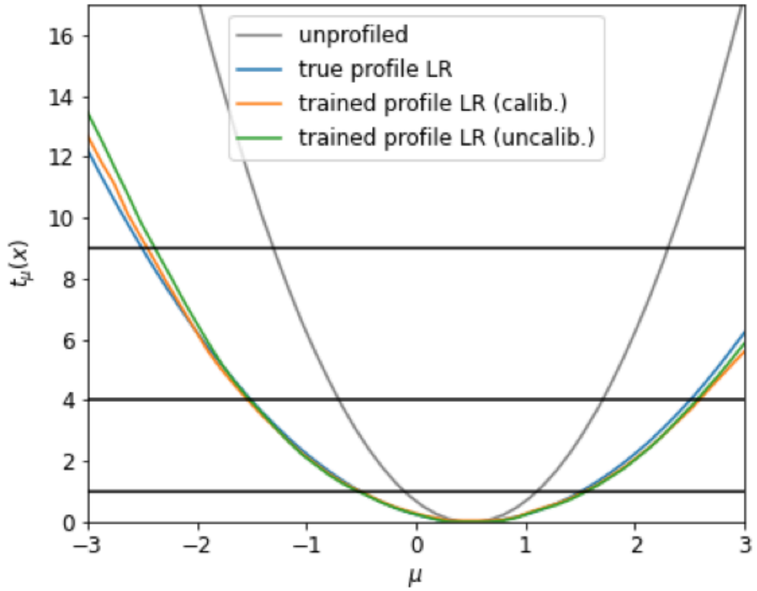
main measurement *side-band*

$$p(c_1, c_2 \,|\, \mu, \nu) = \mathrm{Pois}(c_1 \,|\, \mu s + \nu b)\,\mathrm{Pois}(c_2 \,|\, \nu)$$









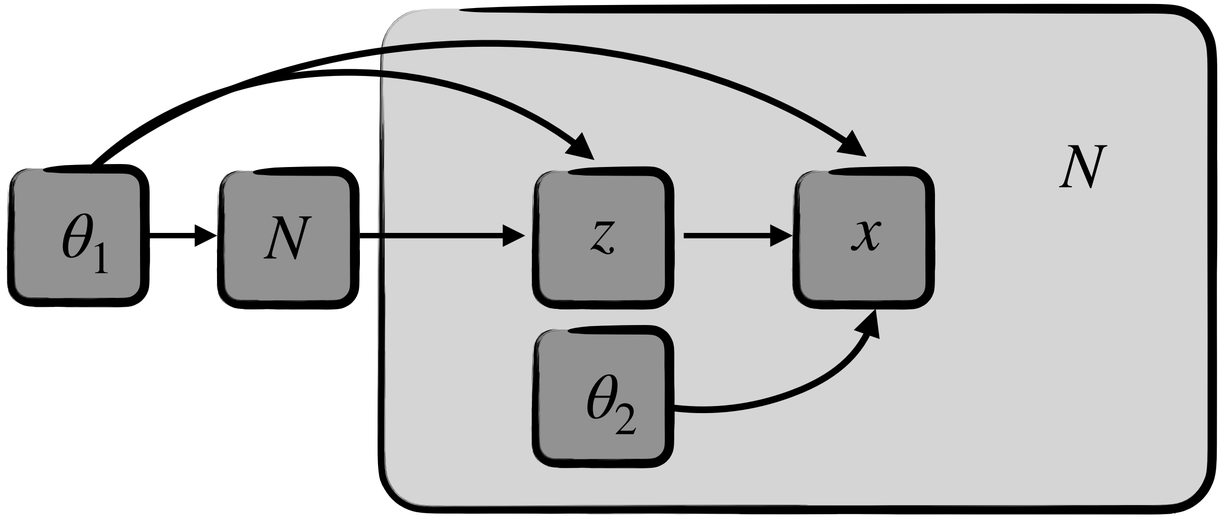
Beyond Cardinalities

If we want to go beyond simple examples, the SBI workflow at the dataset level becomes more complicated

The simulator is now a **simulator of variable-sized sets**

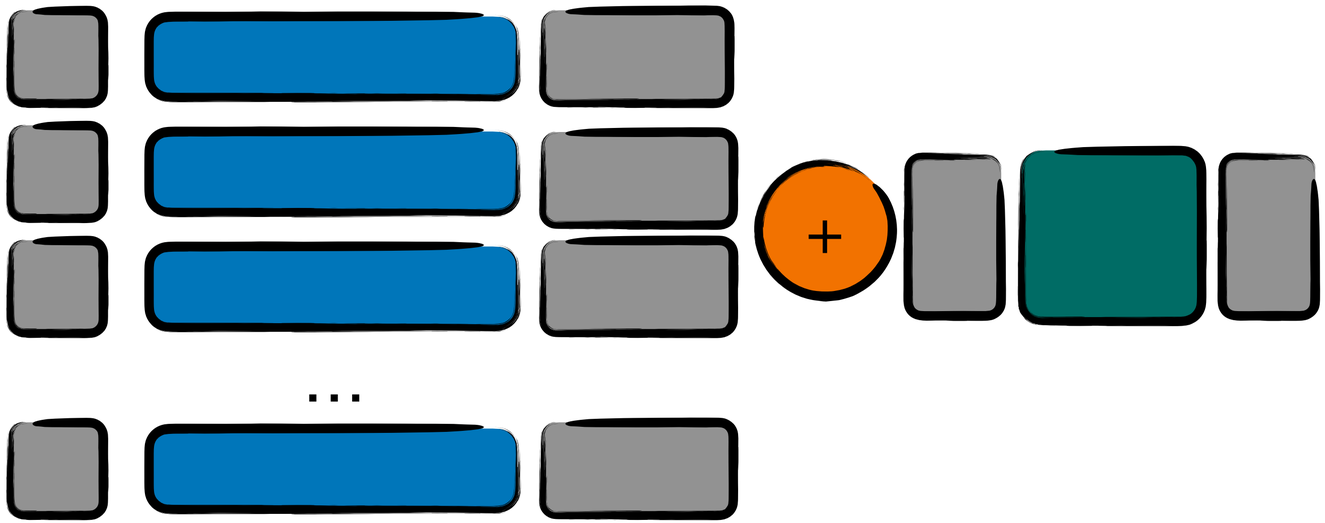
→ some θ impact the set itself, some only set members

$$(\mathscr{D}, \theta) \sim p(\mathscr{D} \mid \theta) p(\theta)$$



The naive approach

There is a rich literature for set embeddings in ML literature
→ foundational method Deep Sets, then Transformers …

$\mathscr{D}$ 

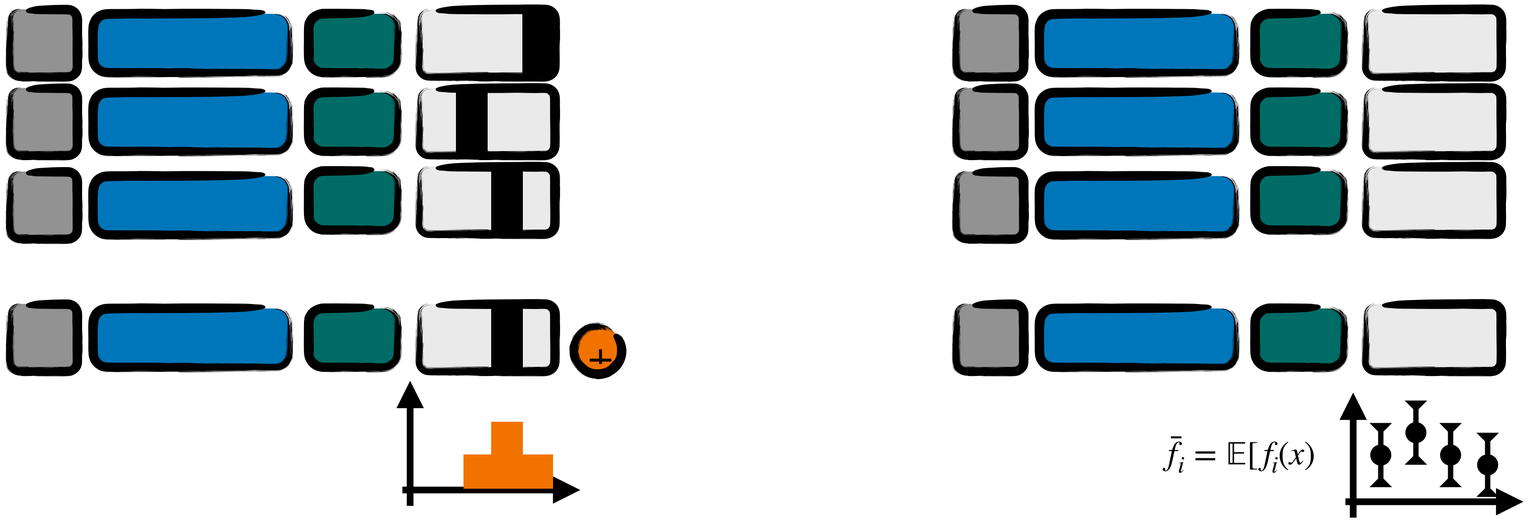
$p(\theta \mid \mathscr{D}), \frac{p(\mathscr{D}, \mu, \hat{\hat{\nu}})}{p(\mathscr{D}) \mid \hat{\mu}, \hat{\nu})}, \ldots$

This structure is sufficient to model any functions of a set

Small Diversion

We can see that Deep Sets are sufficient, because they are either collections of moments or essentially histograms

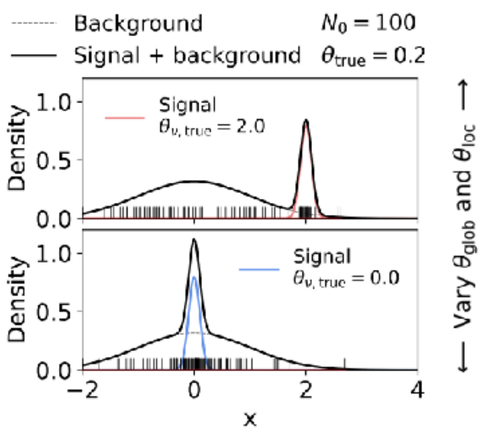
→ Riesz-Markov-Kakutani Theorem and related

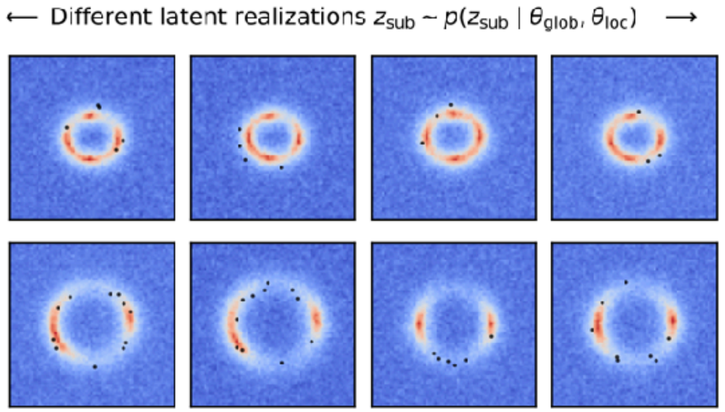


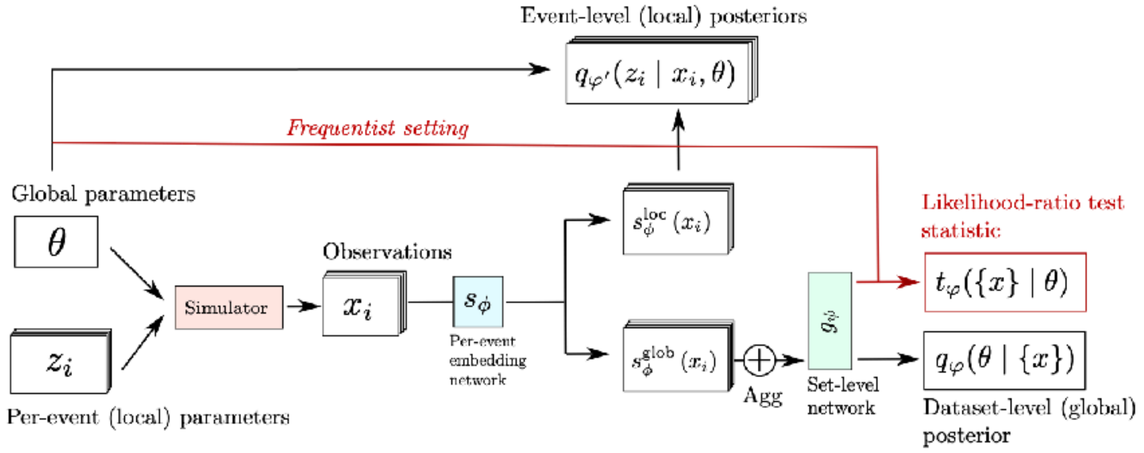
During learning, Deep Sets choose what projection to histogram what expectations to measure in order to fully characterize $p(x \mid \theta)$

Unsurprisingly it works

The standard SBI workflow on sets (of course) works







.. but it doesn't scale to very large sets (thousands, millions)

→ every forward pass processes the full set, scales with $|\mathscr{D}|$

Scaling to Large Sets

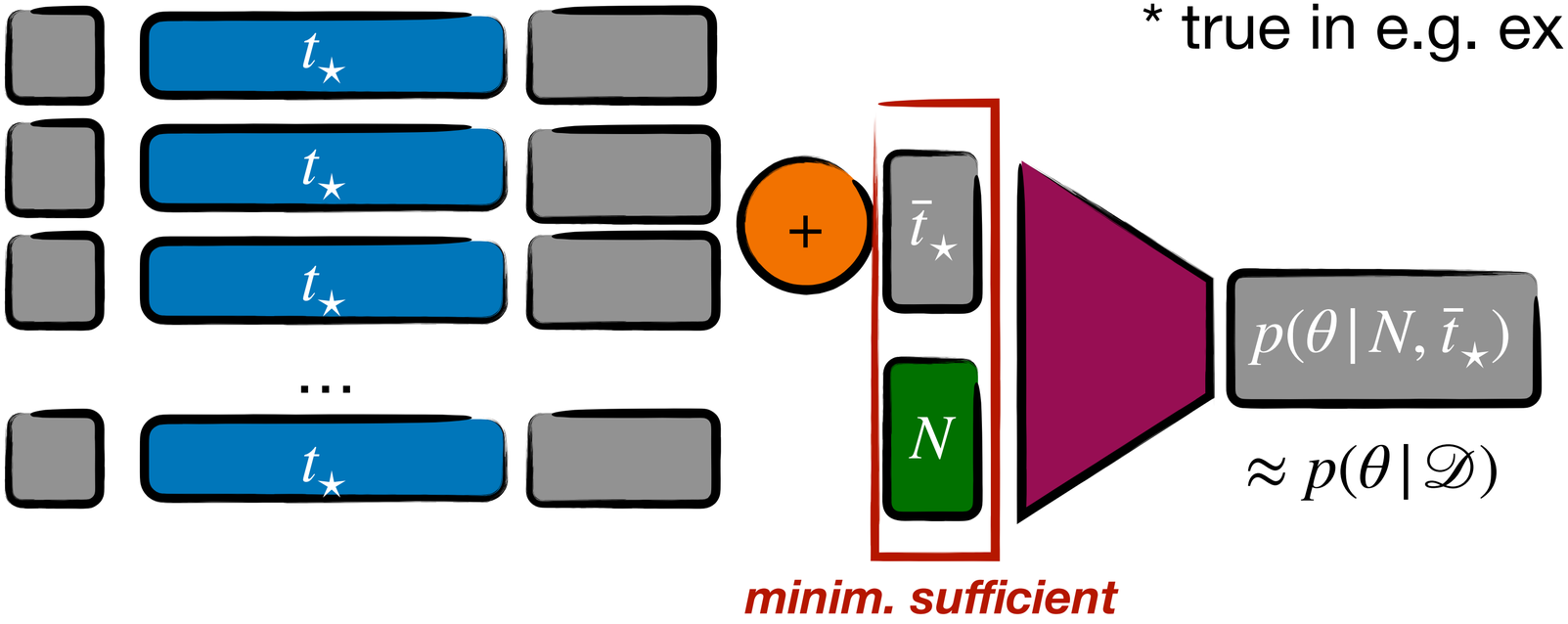
To scale to large sets, we're looking for embeddings of $\mathscr{D}$ that are approximately (Bayesian) sufficient for

$$p(\theta \mid \mathscr{D}) \approx p(\theta \mid E_\phi(\mathscr{D}))$$

Question: Can we find a training procedure that produces a sufficient embedding of datasets without training on full $\mathscr{D}$

Scaling to Large Sets

Assumption*: for inferring θ ther eis a per-sample embedder $t_\star(x)$ whose pooling + cardinality $|\mathscr{D}|$ is minim. sufficient



* true in e.g. exponential family

$$(N, \bar{t}_\star) = \left(N, \sum_i t_\star(x_i) \right)$$

How can we efficiently find the embedder $t_\star(x)$?

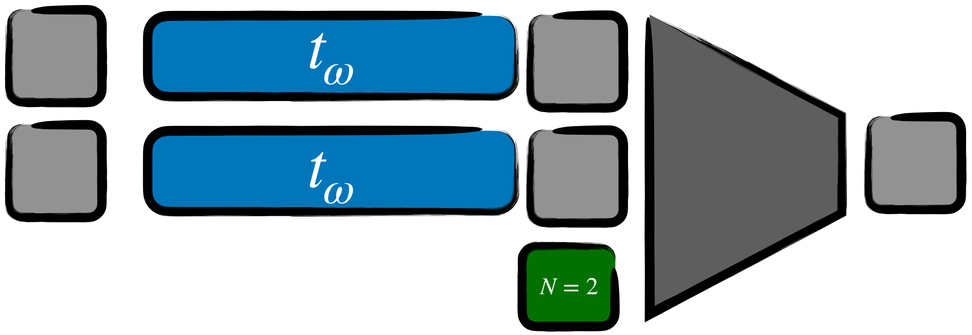
PAIRS*: Small-Cardinality Pretraining

Simple result: we can find a sufficient embedder by pre-training on small-cardinalities (N=2)

** Pretraining Aggregators for Inference at Arbitrary Set-Sizes*

PAIRS*: Small-Cardinality Pretraining

Simple result: we can find a sufficient embedder by pre-training on small-cardinalities (N=2)



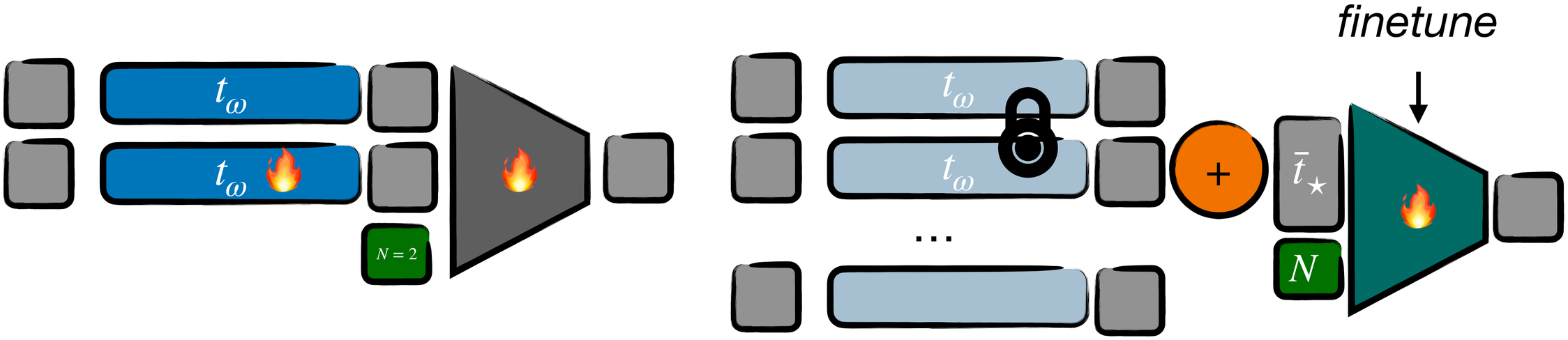
pretrain at small N

$(N = 1,2)$

** Pretraining Aggregators for Inference at Arbitrary Set-Sizes*

PAIRS*: Small-Cardinality Pretraining

Simple result: we can find a sufficient embedder by pre-training on small-cardinalities (N=2)



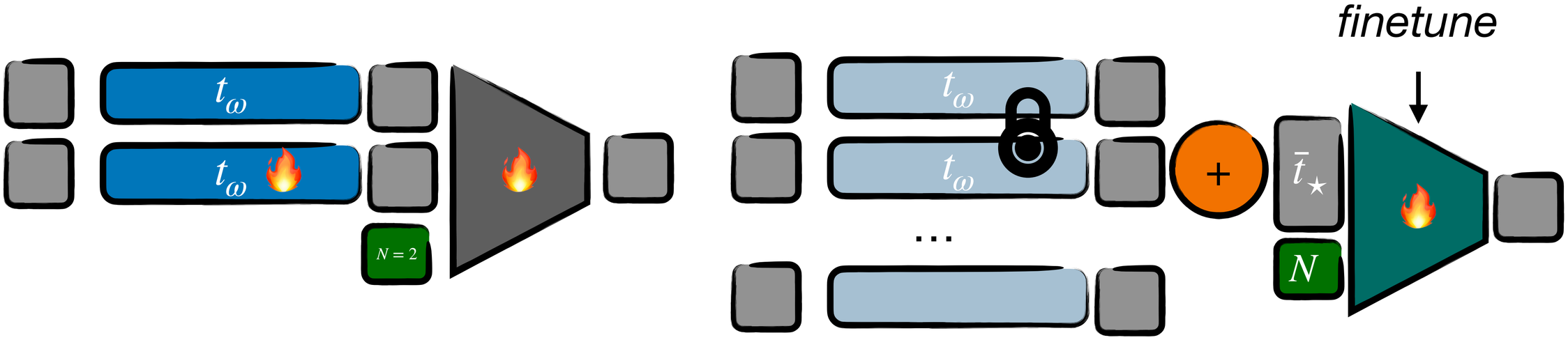
1) pretrain at small N

2) finetune the head to re-optimize to arbitary N

** Pretraining Aggregators for Inference at Arbitrary Set-Sizes*

PAIRS*: Small-Cardinality Pretraining

Simple result: we can find a sufficient embedder by pre-training on small-cardinalities (N=2)



1) pretrain at small N

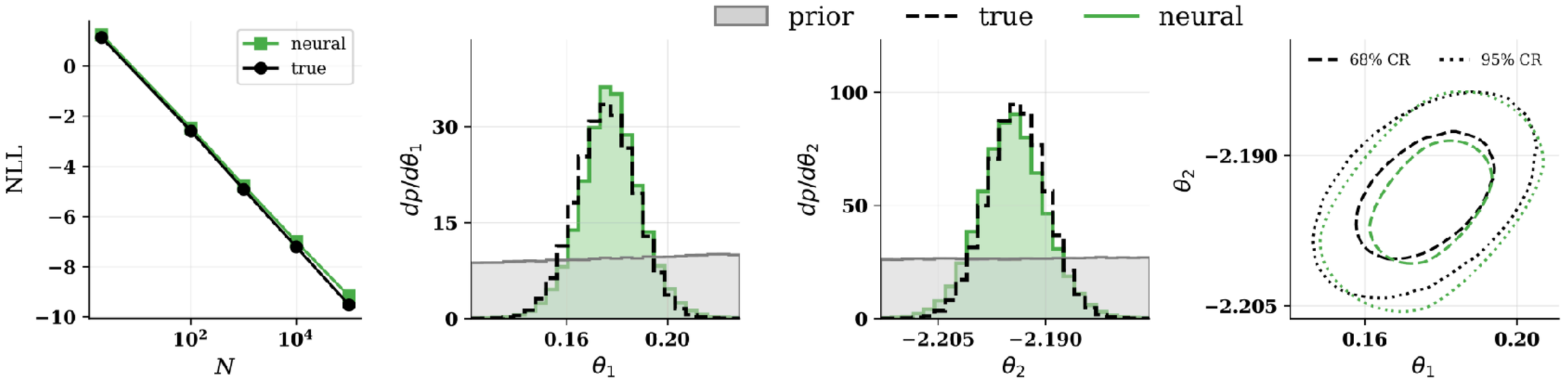
2) finetune the head to re-optimize to arbitary N

Large-N Set Embed can be cached! Not part of forward pass

** Pretraining Aggregators for Inference at Arbitrary Set-Sizes*

PAIRS*: Small-Cardinality Pretraining

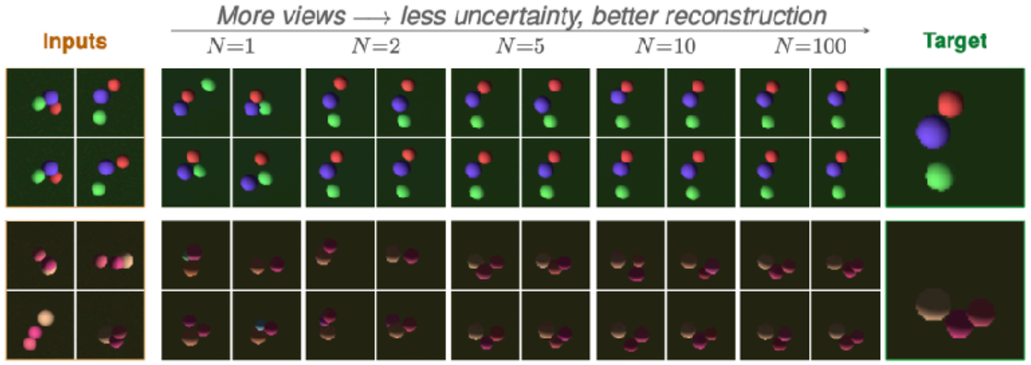
Consequence: we can do a fully neural end-to-end inference pipeline with no explicit MCMC / Profiling, etc
→ without ever training on full-scale sets

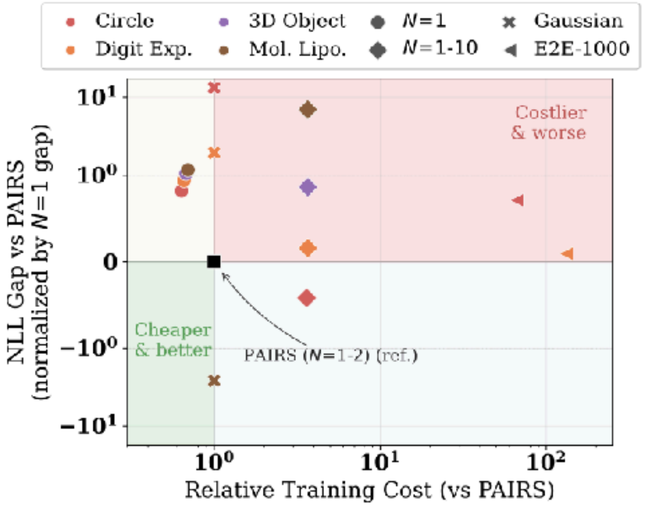


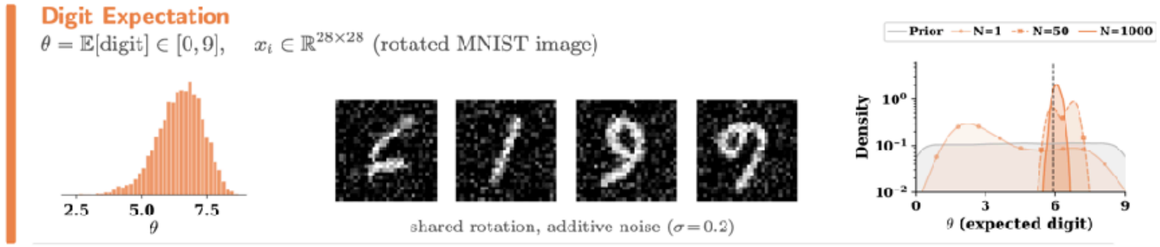
** Pretraining Aggregators for Inference at Arbitrary Set-Sizes*

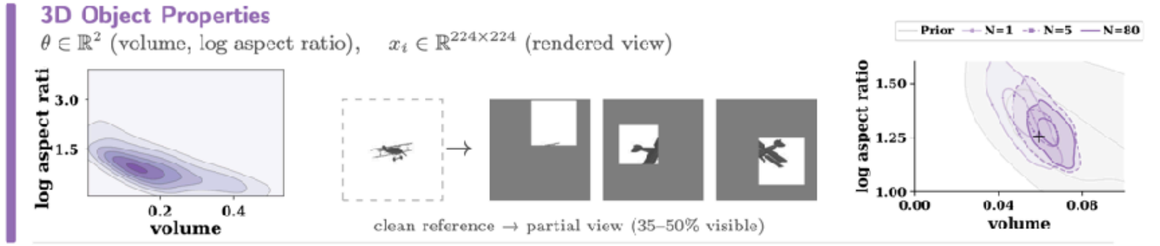
Applications

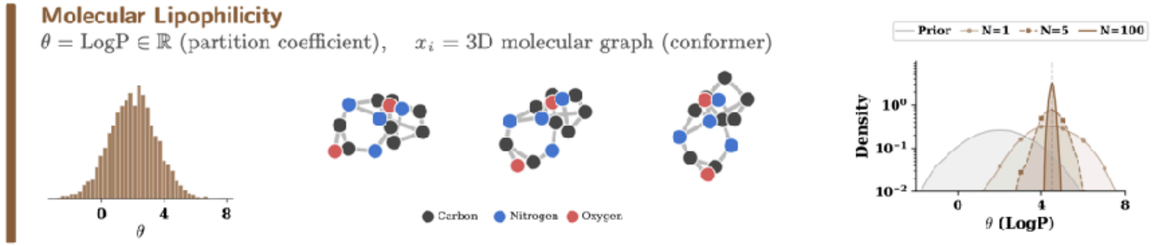
Very motivated by HEP, but set-level inference appears in many appliations in biology, astronomy, vision, ...











Summary & Outlook

SBI in Particle Physics is tough because of three “high-D”:
→ instance size, dataset cardinality, parameter space

Problem can be generalized / formalized as “end-to-end SBI on large-scale i.i.d. sets”

For a large-class of problems pretrain-finetune is an efficient strategy to produces an embedder that is still ~optimal

Can do dataset-level SBI without O(N) forward pass and without a secondary stats. step

Proof Sketch

To know a iid dataset $\mathscr{D}$ “sufficiently” you “just” need to know

i) know $p(x)$
ii) know N

ad i) if you know sufficiently many moments you know $p(x)$ (Riesz)

$$\bar{f}_k = \int f_k(x) p(x) \quad \text{with} \quad f_k = \mathscr{F} \qquad\qquad \bar{f}_k \approx \frac{1}{N} \sum_k f_k(x_i)$$

ergo: knowing $\left(\frac{1}{N} \sum_k f_k(x_i), N \right) \approx$ sufficient stat $f(\mathscr{D})$