The frontiers of simulation-based inference (Part I)

PHYSTAT-SBI 2024
May 15, 2024

Gilles Louppe
g.louppe@uliege.be

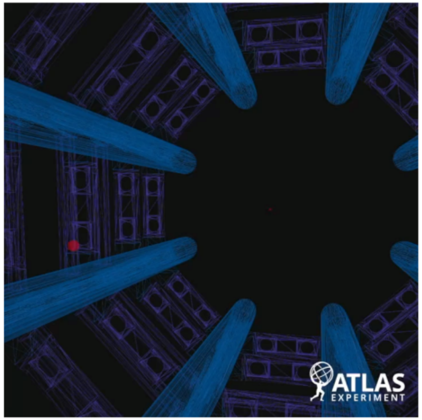
Simulation-based inference

Simulators as generative models

A simulator prescribes a generative model that can be used to simulate data $\mathbf{x}$.

Collider data

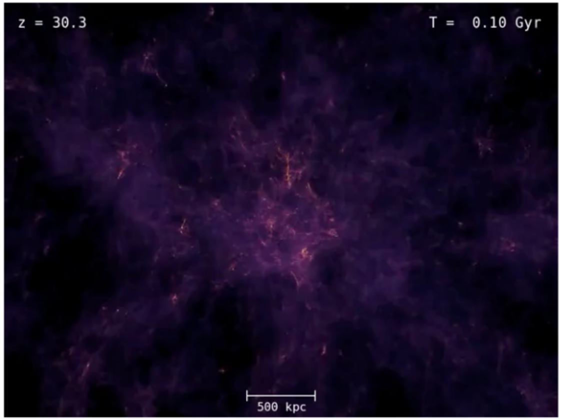
$\text{particles} \sim p(\text{particles})$



[C. Cesarotti with ATLAS]

Cosmology data

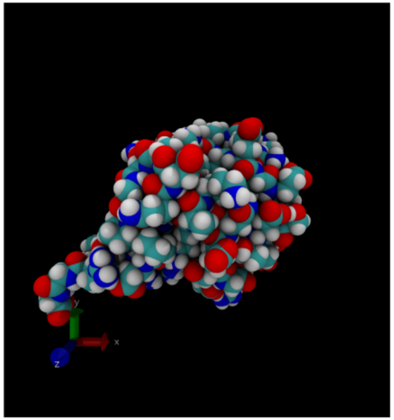
$\text{particles} \sim p(\text{particles})$



[Aquarius simulation]

Molecular dynamics

$\text{configurations} \sim p(\text{configurations})$



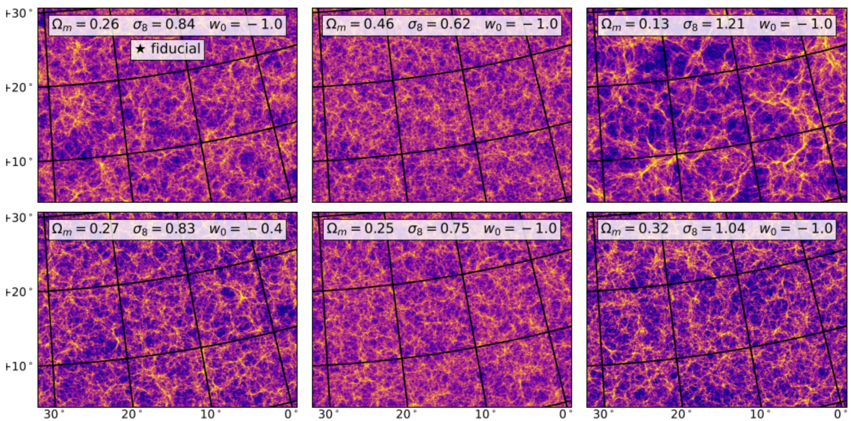
[E. Cances et al]

Conditional simulators

A conditional simulator prescribes a way to sample from the likelihood $p(\mathbf{x}|\theta)$, where θ is a set of conditioning variables or parameters.

Cosmology data

$$\text{map} \sim p(\text{map} \mid \{\Omega_m, \sigma_8, w_0\})$$



[Kacprzak et al 2022]

$$x \sim p(x; \mathscr{M})$$

Model

or

$$x \sim p(x \mid \theta)$$

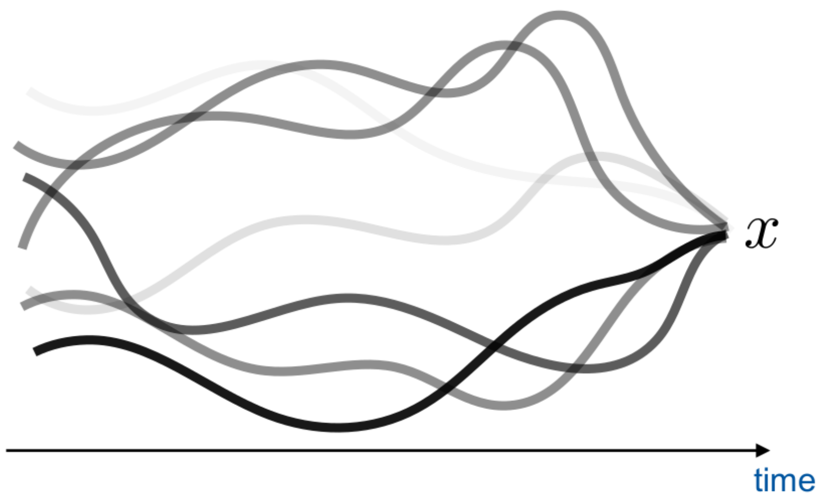
Model parameters

Intractable likelihoods

The (modeled) data generating process may involve additional latent variables $\mathbf{z}$ that are not observed, leading to likelihoods

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z}.$$

In this case, evaluating the likelihood becomes intractable.



$p(\mathbf{z}_p|\theta)$

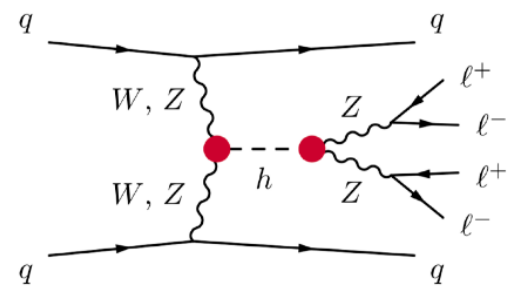
Latent variables

Parameters of interest

Parton-level momenta

Theory parameters

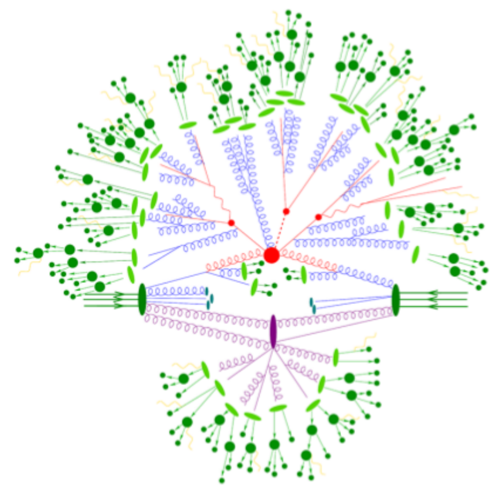
$z_p \longleftarrow \theta$



$$p(\mathbf{z}_s|\theta) = \int p(\mathbf{z}_p|\theta)p(\mathbf{z}_s|\mathbf{z}_p)d\mathbf{z}_p$$

Latent variables		Parameters of interest
Shower splittings	Parton-level momenta	Theory parameters

$$z_s \longleftarrow z_p \longleftarrow \theta$$



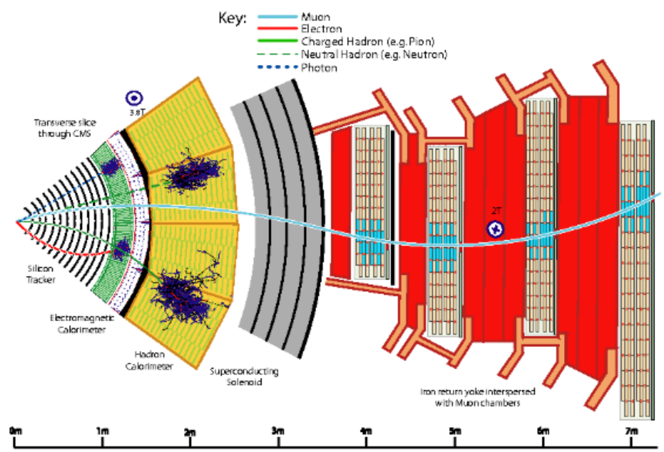
$$p(\mathbf{z}_d|\theta) = \iint p(\mathbf{z}_p|\theta)p(\mathbf{z}_s|\mathbf{z}_p)p(\mathbf{z}_d|\mathbf{z}_s)d\mathbf{z}_p d\mathbf{z}_s$$

Latent variables

Parameters of interest

Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
z_d	z_s	z_p	θ

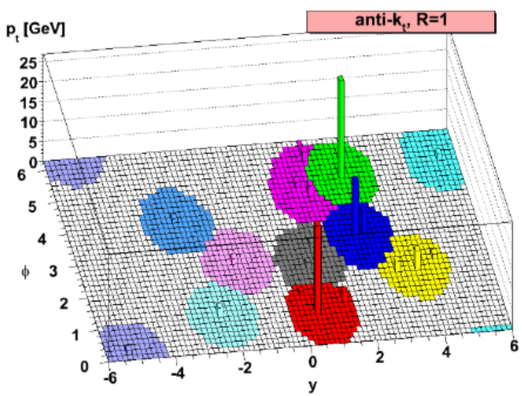
$$z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$$



$$p(\mathbf{x}|\theta) = \underbrace{\iiint}_{\text{yikes!}} p(\mathbf{z}_p|\theta)p(\mathbf{z}_s|\mathbf{z}_p)p(\mathbf{z}_d|\mathbf{z}_s)p(\mathbf{x}|\mathbf{z}_d)d\mathbf{z}_p d\mathbf{z}_s d\mathbf{z}_d$$

Features		Latent variables		Parameters of interest
Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters

$$x \longleftarrow z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$$



[Image source: M. Cacciari, G. Salam, G. Soyez 0802.1189]

What can we do with generative models?

Produce samples and make predictions

$$\mathbf{x} \sim p(\mathbf{x}|\theta)$$

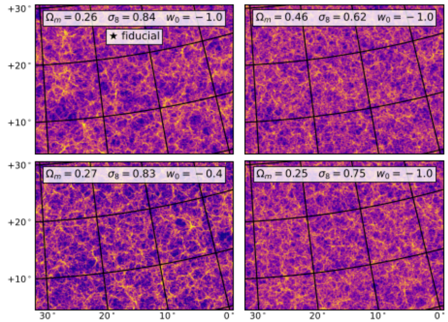
Evaluate densities

$$p(\mathbf{x}|\theta)$$

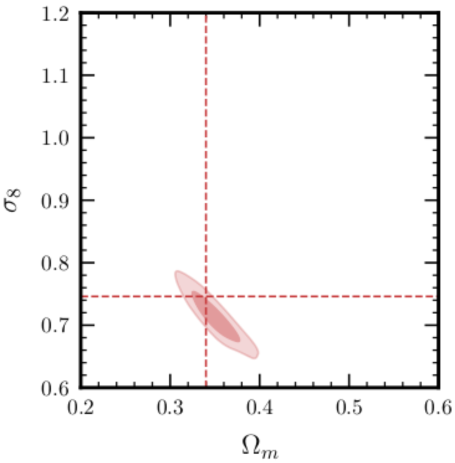
$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

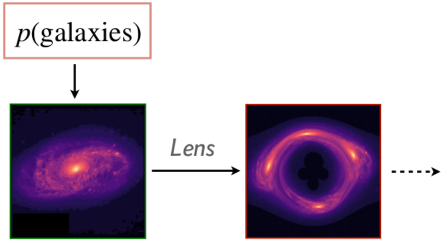
Encode complex priors

$$p(\mathbf{x})$$

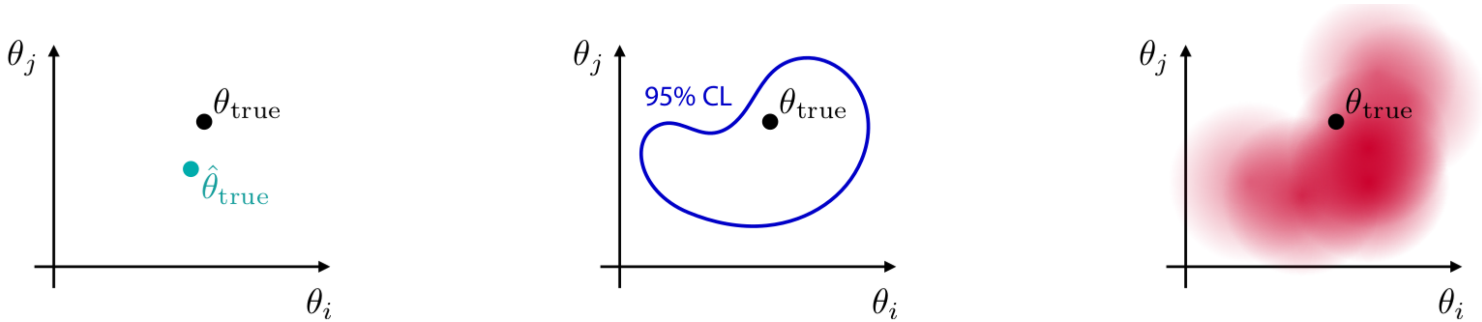


[Kacprzak et al 2022]



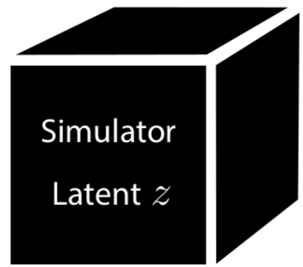


Inference



- Frequentist inference: find $\hat{\theta}$ that maximizes the likelihood $p(\mathbf{x}|\theta)$ or build a confidence interval thereof.
- Bayesian inference: compute the posterior distribution $p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$.

Statistical inference becomes challenging when the likelihood $p(\mathbf{x}|\theta)$ is implicit or intractable. **Simulation-based inference algorithms are needed.**

Parameters θ → 

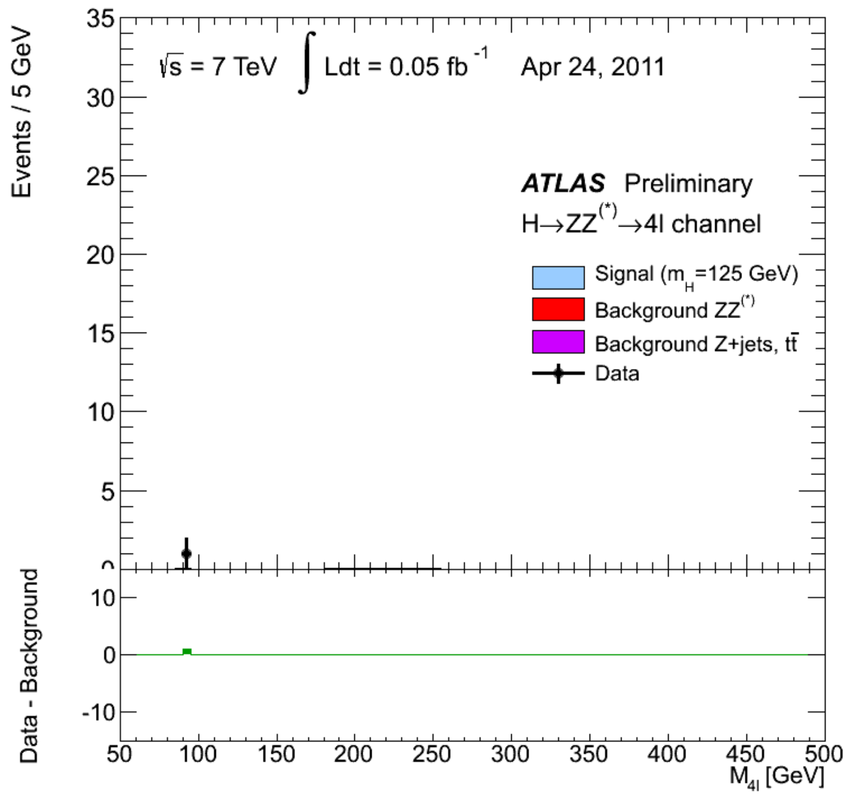
→ Observables x

Prediction:
- Well-motivated mechanistic, causal model
- Simulator can generate samples $x \sim p(x|\theta)$

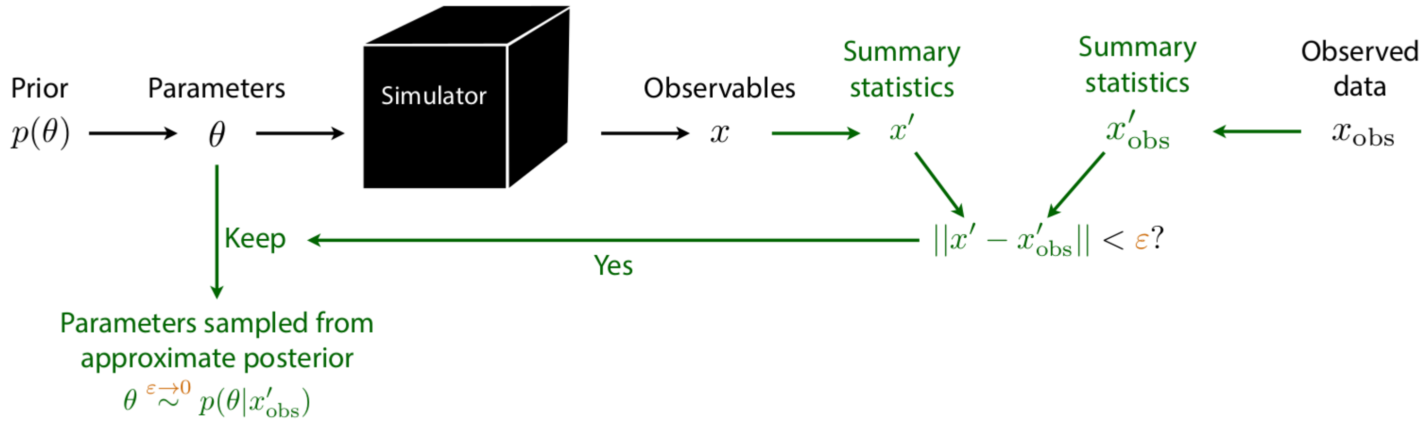
Inference:
- Interactions between low-level components lead to challenging inverse problems
- Likelihood $p(x|\theta) = \int \mathrm{d}z \; p(x, z|\theta)$ is intractable

The frontiers

pre-2019



(Frequentist) Approximate the likelihood $p(\mathbf{x}|\theta)$ as
$p(\mathbf{x}|\theta) \approx \hat{p}(\mathbf{x}|\theta) = p(s(\mathbf{x})|\theta)$ for some (well-chosen) summary statistic $s(\cdot)$.



(Bayesian) Approximate the posterior $p(\theta|\mathbf{x})$ using Approximate Bayesian Computation.

Issues:

- How to choose $\mathbf{x}' = s(\mathbf{x})$? ϵ? $||\cdot||$?
- No tractable posterior.
- Need to run new simulations for new data or new prior.

2019

COLLOQUIUM PAPER

The frontier of simulation-based inference

Kyle Cranmer[a,b,1], **Johann Brehmer**[a,b], **and Gilles Louppe**[c]

[a]Center for Cosmology and Particle Physics, New York University, New York, NY 10003; [b]Center for Data Science, New York University, New York, NY 10011; and [c]Montefiore Institute, University of Liège, B-4000 Liège, Belgium

Edited by Jitendra Malik, University of California, Berkeley, CA, and approved April 10, 2020 (received for review November 4, 2019)

Many domains of science have developed complex simulations to describe phenomena of interest. While these simulations provide high-fidelity models, they are poorly suited for inference and lead to challenging inverse problems. We review the rapidly developing field of simulation-based inference and identify the forces giving additional momentum to the field. Finally, we describe how the frontier is expanding so that a broad audience can appreciate the profound influence these developments may have on science.

statistical inference | implicit models | likelihood-free inference | approximate Bayesian computation | neural density estimation

Mechanistic models can be used to predict how systems will behave in a variety of circumstances. These run the gamut of distance scales, with notable examples including particle physics, molecular dynamics, protein folding, population genetics, neuroscience, epidemiology, economics, ecology, climate science, astrophysics, and cosmology. The expressiveness of programming languages facilitates the development of complex, high-fidelity simulations and the power of modern computing provides the ability to generate synthetic data from them. Unfortunately, these simulators are poorly suited for statistical inference. The source of the challenge is that the probability density (or likelihood) for a given observation—an essential ingredient for both frequentist and Bayesian inference methods—is typically intractable. Such models are often referred to as implicit models and contrasted against prescribed models where the likelihood for an observation can be explicitly calculated (1). The problem setting of statistical inference under intractable likelihoods has been dubbed likelihood-free inference—although it is a bit of a misnomer as typically one attempts to estimate the intractable likelihood, so we feel the term simulation-based inference is more apt.

The intractability of the likelihood is an obstruction for scientific progress as statistical inference is a key component of the scientific method. In areas where this obstruction has appeared, scientists have developed various ad hoc or field-specific meth-

the simulator—is being recognized as a key idea to improve the sample efficiency of various inference methods. A third direction of research has stopped treating the simulator as a black box and focused on integrations that allow the inference engine to tap into the internal details of the simulator directly.

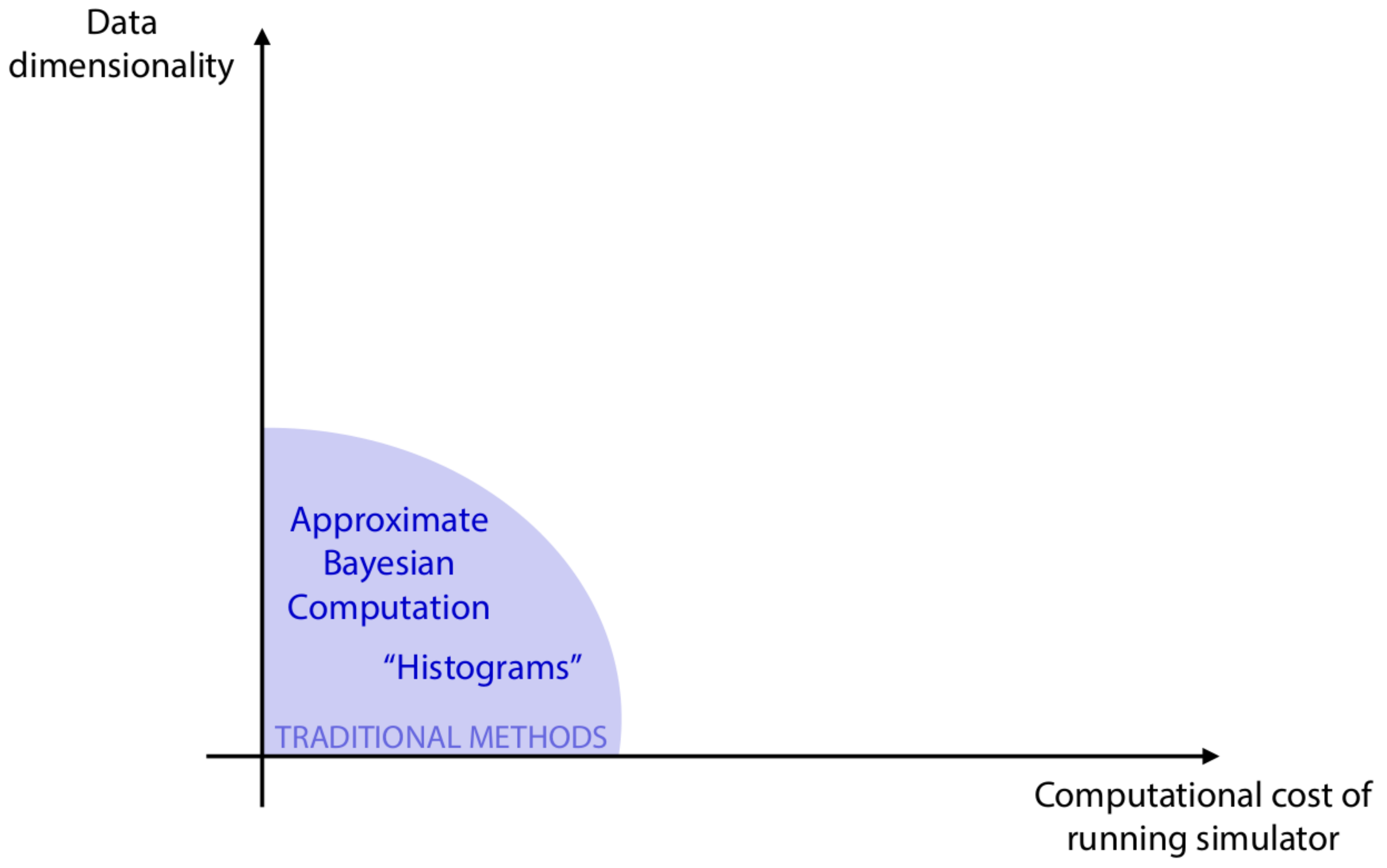
Amidst this ongoing revolution, the landscape of simulation-based inference is changing rapidly. In this review we aim to provide the reader with a high-level overview of the basic ideas behind both old and new inference techniques. Rather than discussing the algorithms in technical detail, we focus on the current frontiers of research and comment on some ongoing developments that we deem particularly exciting.

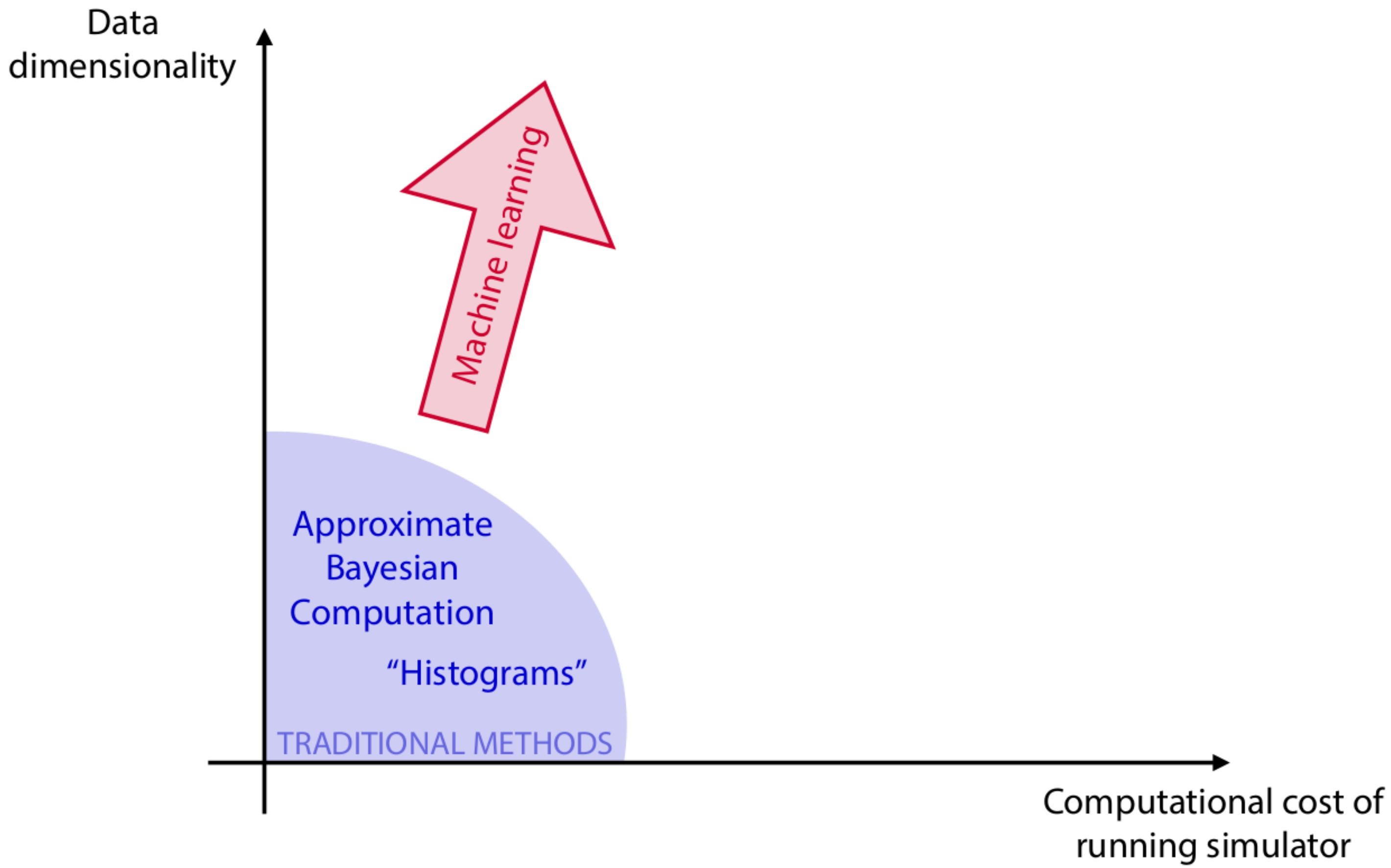
Simulation-Based Inference

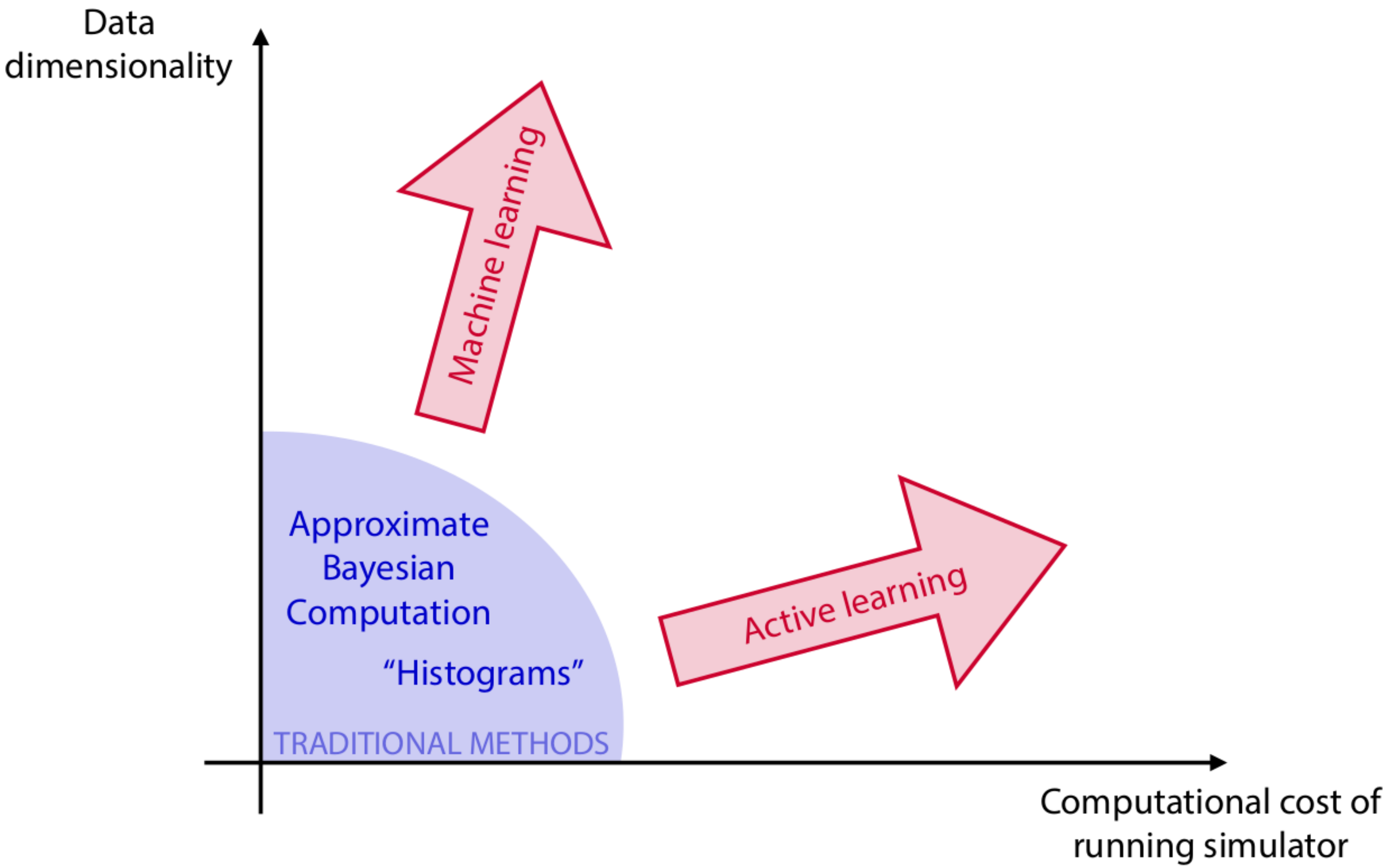
Simulators. Statistical inference is performed within the context of a statistical model, and in simulation-based inference the simulator itself defines the statistical model. For the purpose of this paper, a simulator is a computer program that takes as input a vector of parameters θ, samples a series of internal states or latent variables $z_i \sim p_i(z_i|\theta, z_{<i})$, and finally produces a data vector $x \sim p(x|\theta, z)$ as output. Programs that involve random samplings and are interpreted as statistical models are known as probabilistic programs, and simulators are an example. Within this general formulation, real-life simulators can vary substantially:

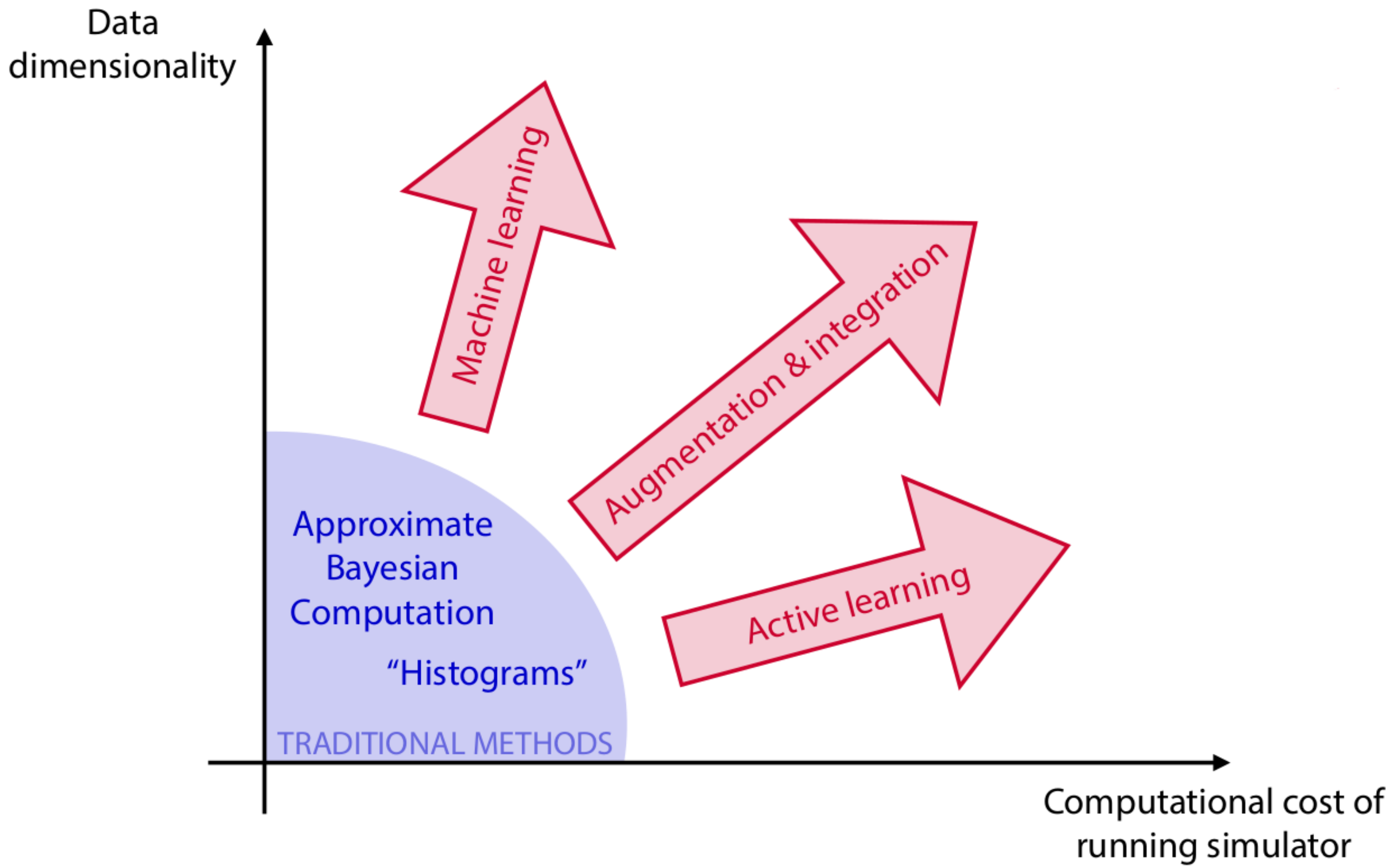
STATISTICS

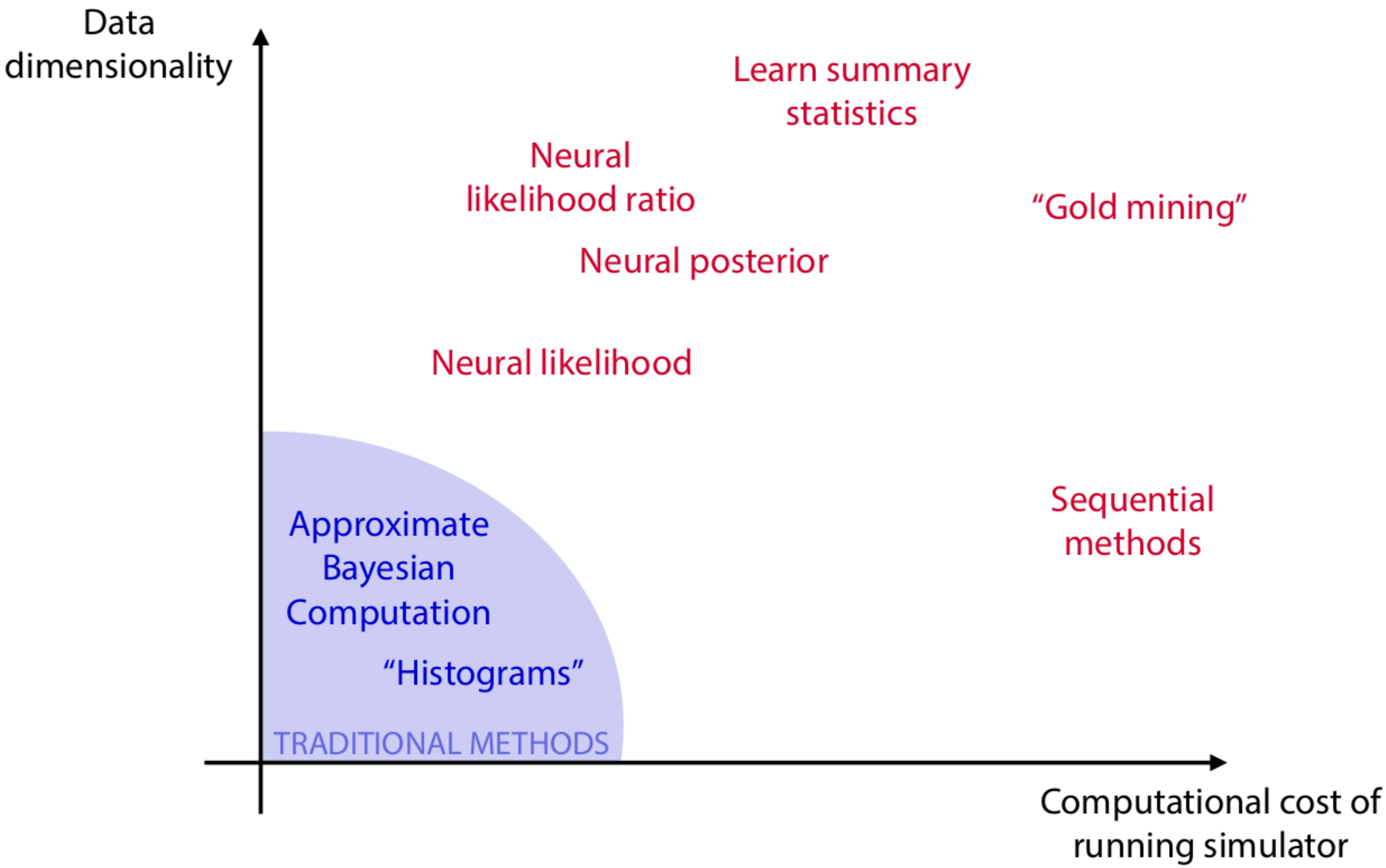
- The parameters θ describe the underlying mechanistic model and thus affect the transition probabilities $p_i(z_i|\theta, z_{<i})$. Typically the mechanistic model is interpretable by a domain scientist and θ has relatively few components and a fixed dimensionality. Examples include coefficients found in the Hamiltonian of a physical system, the virulence and incubation rate of a pathogen, or fundamental constants of Nature.
- The latent variables z that appear in the data-generating process may directly or indirectly correspond to a physically meaningful state of a system, but typically this state is unobservable in practice. The structure of the latent space varies substantially between simulators. The latent variables may be continuous or discrete and the dimensionality of the latent space may be


Data
dimensionality
Approximate
Bayesian
Computation
"Histograms"
TRADITIONAL METHODS
Computational cost of
running simulator


Data
dimensionality
Machine learning
Approximate
Bayesian
Computation
“Histograms”
TRADITIONAL METHODS
Computational cost of
running simulator

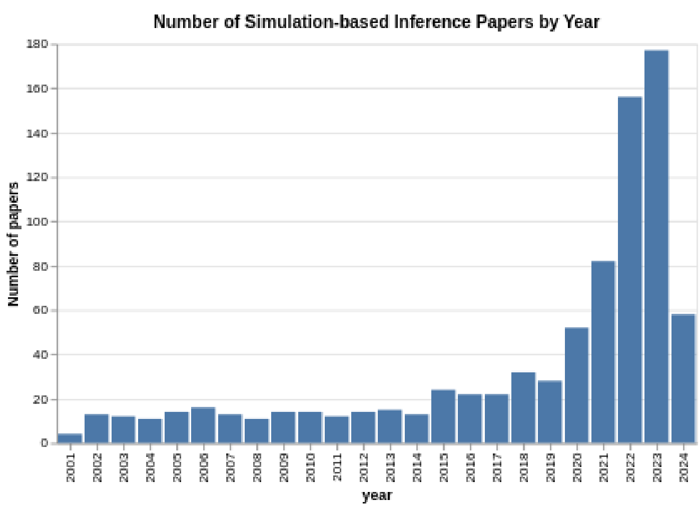

Data dimensionality
Machine learning
Approximate Bayesian Computation
"Histograms"
TRADITIONAL METHODS
Active learning
Computational cost of running simulator

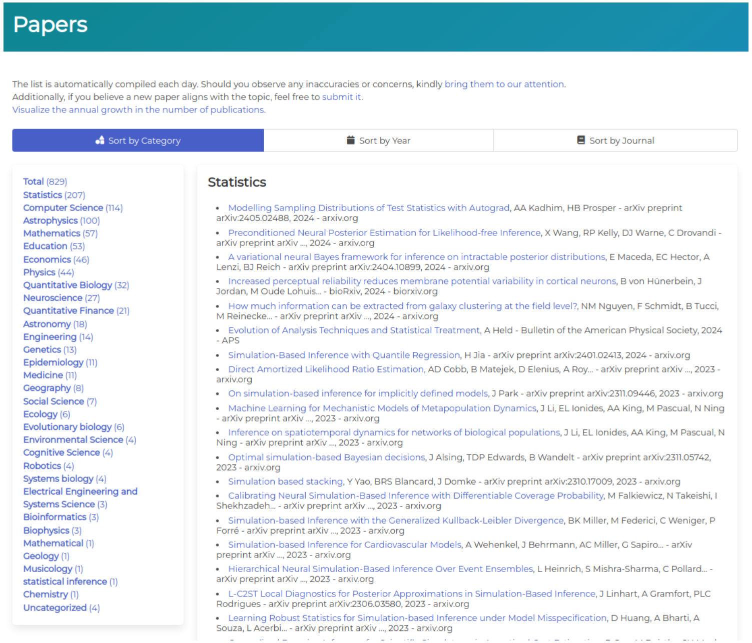

Data dimensionality
Machine learning
Augmentation & integration
Active learning
Approximate Bayesian Computation
"Histograms"
TRADITIONAL METHODS
Computational cost of running simulator


Data
dimensionality
Learn summary
statistics
Neural
likelihood ratio
"Gold mining"
Neural posterior
Neural likelihood
Sequential
methods
Approximate
Bayesian
Computation
"Histograms"
TRADITIONAL METHODS
Computational cost of
running simulator

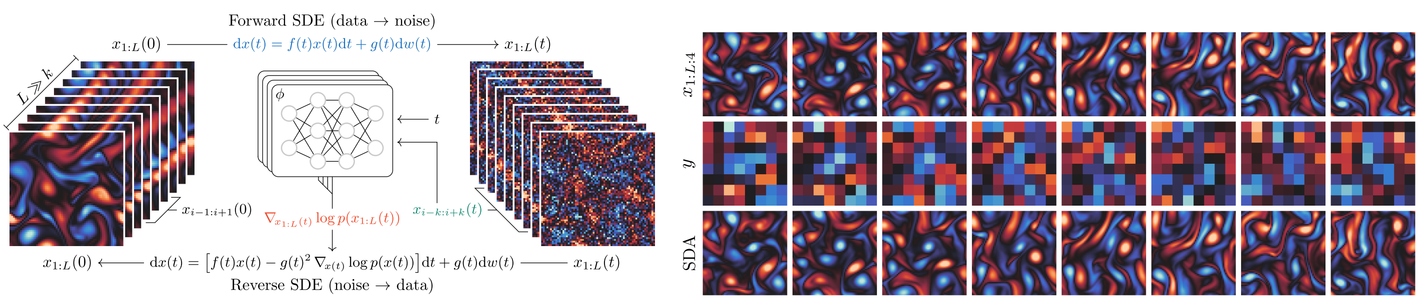
2024

The adoption of simulation-based inference has been growing steadily since then, with new algorithms and applications pushing the boundaries of what is possible.





Developments in deep learning (e.g., diffusion models, transformers, GNNs, etc) have continued to **scale up simulation-based inference to higher dimensional simulation models** (both in the number of parameters θ and size of the data $\mathbf{x}$).



Rozet and Louppe et al (2023): "We introduce score-based data assimilation for trajectory inference. We learn a score-based generative model of state trajectories of a high-dimensional dynamical system and use it for the assimilation of noisy observations."

—

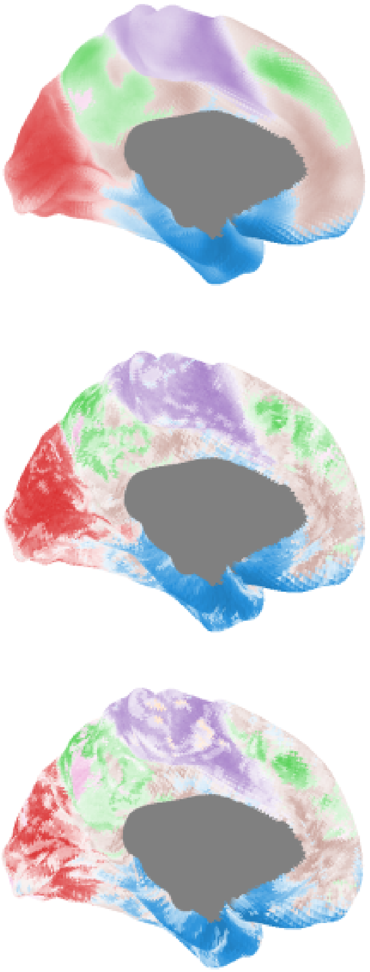
Credits: Rozet and Louppe, 2023.

Active learning remains largely unexplored. Beyond greedy strategies, little attention has been given to the informed selection of simulations for building a training set.

Current paradigms cannot deal with expensive simulators (e.g., climate models, cosmological simulations).

Extracting side information based on $\boldsymbol{\theta}, \mathbf{z}, \mathbf{x}$ remains challenging due to implementation constraints.

However, designing inference networks that **leverage domain knowledge** (e.g., symmetries, conservation laws) **or the structure of the simulator** (e.g., hierarchical models) has shown promising results.



Rouillard et al (2024): "We demonstrate the ability of PAVI to tackle large neuroimaging hierarchical inference problems. For each of the 59000 vertices of every of the 1000 subjects, we infer a probabilistic label to belong to one of the 7 functional networks. This amounts to inferring over 400 million latent variables."

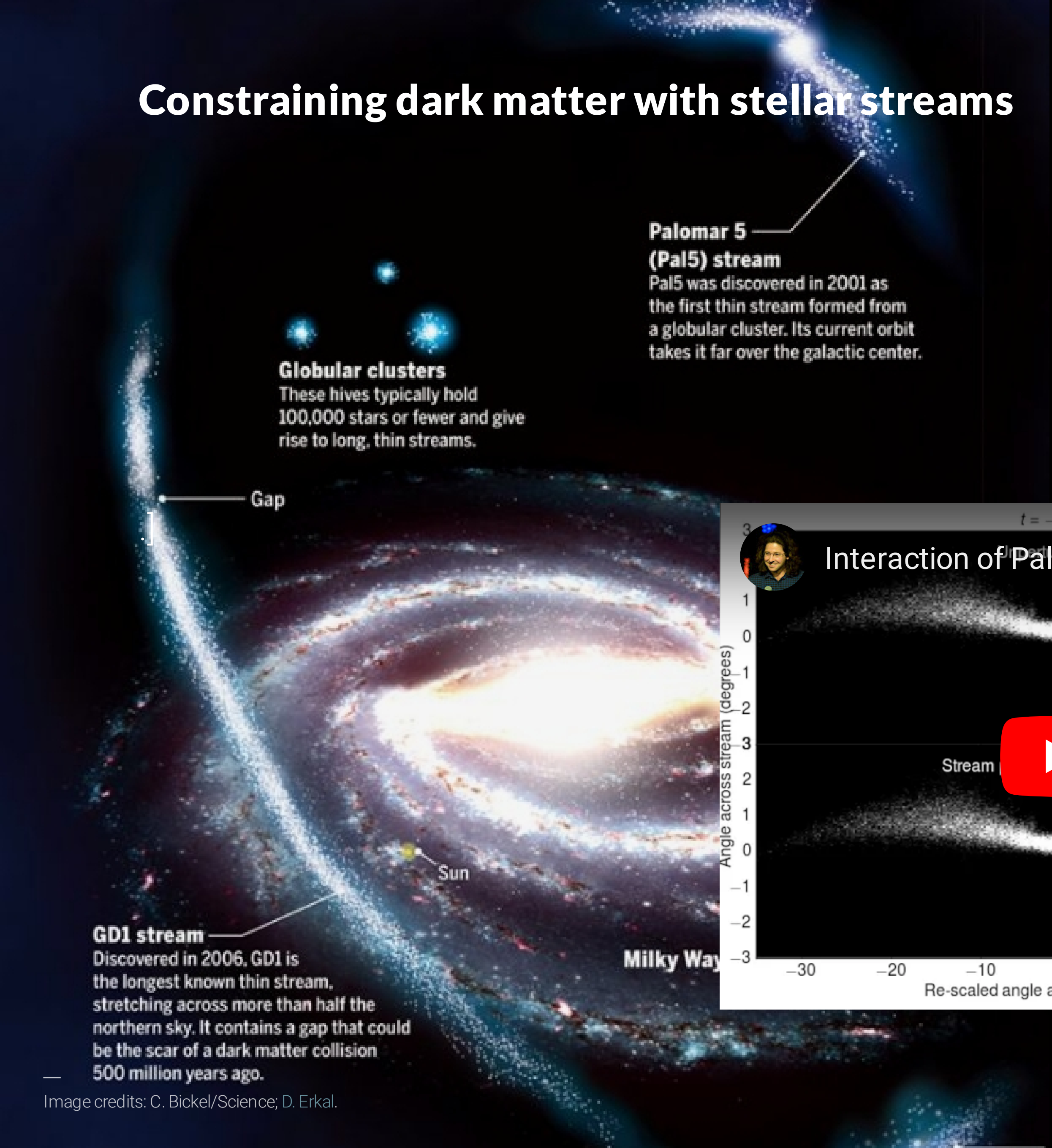
—
Credits: Rouillard et al, 2023; Rouillard, 2024.

A case study

Hermans et al, "Constraining dark matter with stellar streams", 2021.

Can we constrain the nature of dark matter from cosmological observations?

Constraining dark matter with stellar streams



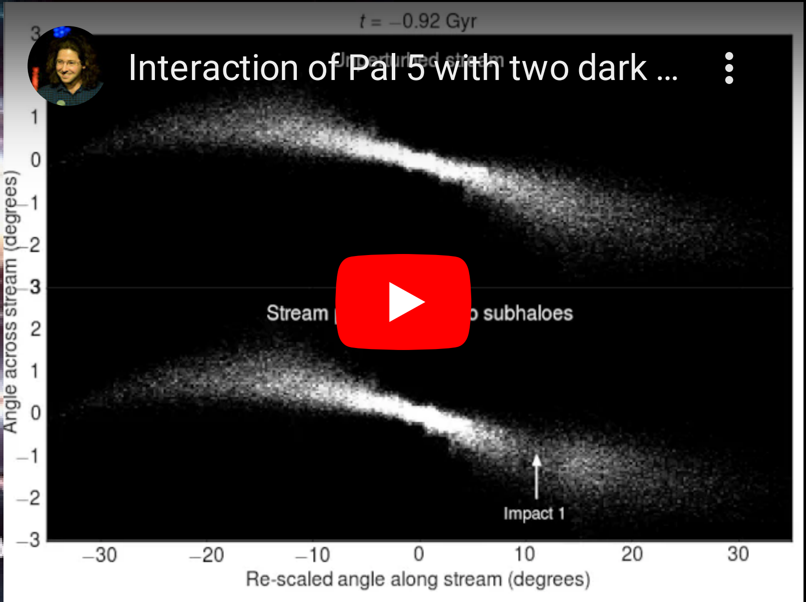
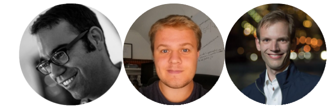


Image credits: C. Bickel/Science; D. Erkal.

$$p(m_{\mathrm{WDM}}, t_{\mathrm{age}}|\mathrm{GD1}) = \frac{p(\mathrm{GD1}|m_{\mathrm{WDM}}, t_{\mathrm{age}})p(m_{\mathrm{WDM}}, t_{\mathrm{age}})}{p(\mathrm{GD\text{-}1})}$$

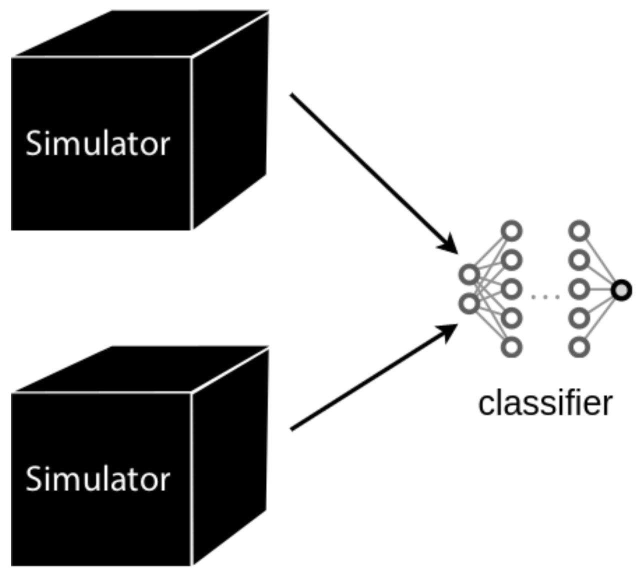


Neural ratio estimation (NRE)

The likelihood-to-evidence $r(\mathbf{x}|\theta) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x})} = \frac{p(\theta,\mathbf{x})}{p(\theta)p(\mathbf{x})}$ ratio can estimated from a binary classifier $d(\theta, \mathbf{x})$, even if neither the likelihood nor the evidence can be evaluated.

$\theta, \mathbf{x} \sim p(\theta, \mathbf{x})$

$\theta, \mathbf{x} \sim p(\theta)p(\mathbf{x})$



$\hat{r}(\mathbf{x}|\theta)$

—

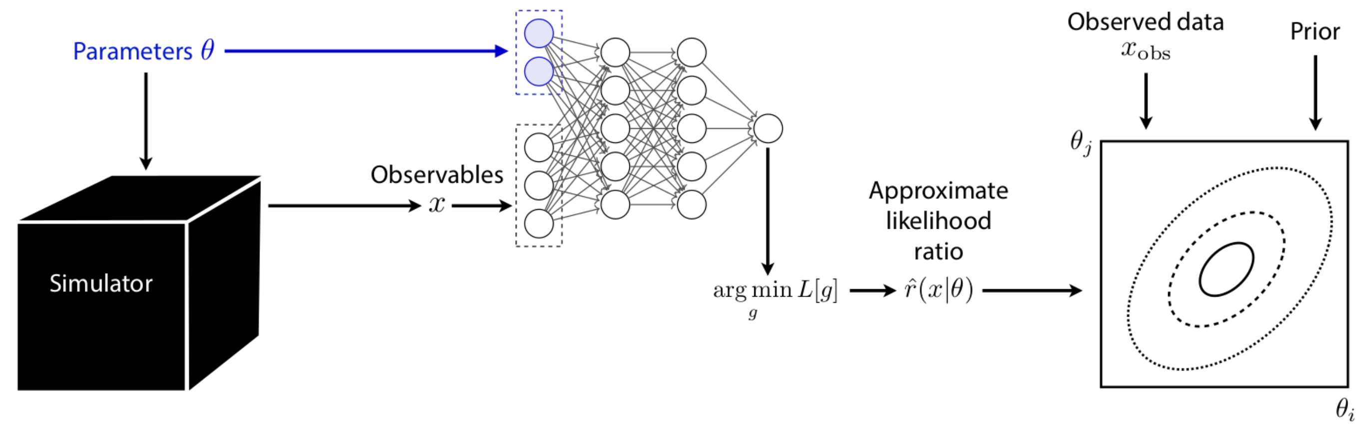
Credits: Cranmer et al, 2015; Hermans et al, 2020.

The solution d found after training approximates the optimal classifier

$$d(\theta, \mathbf{x}) \approx d^*(\theta, \mathbf{x}) = \frac{p(\theta, \mathbf{x})}{p(\theta, \mathbf{x}) + p(\theta)p(\mathbf{x})}.$$

Therefore,

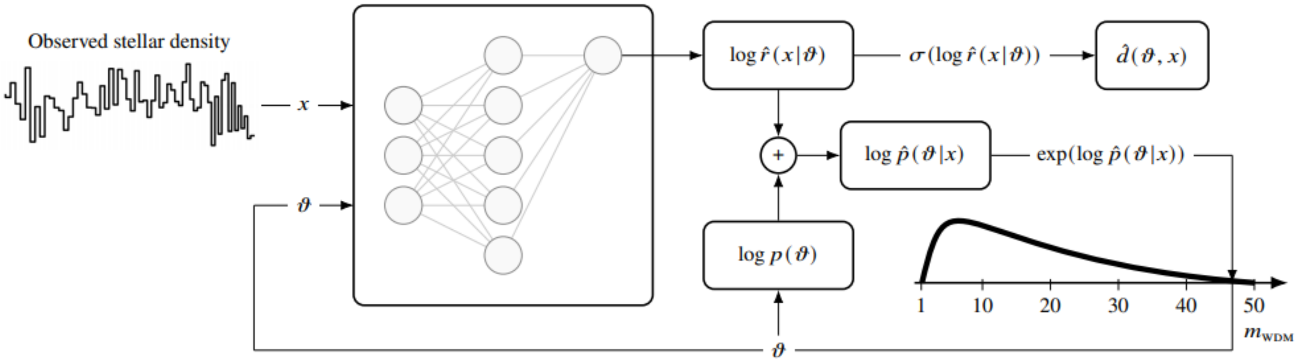
$$r(\mathbf{x}|\theta) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x})} = \frac{p(\theta, \mathbf{x})}{p(\theta)p(\mathbf{x})} \approx \frac{d(\theta, \mathbf{x})}{1 - d(\theta, \mathbf{x})} = \hat{r}(\mathbf{x}|\theta).$$

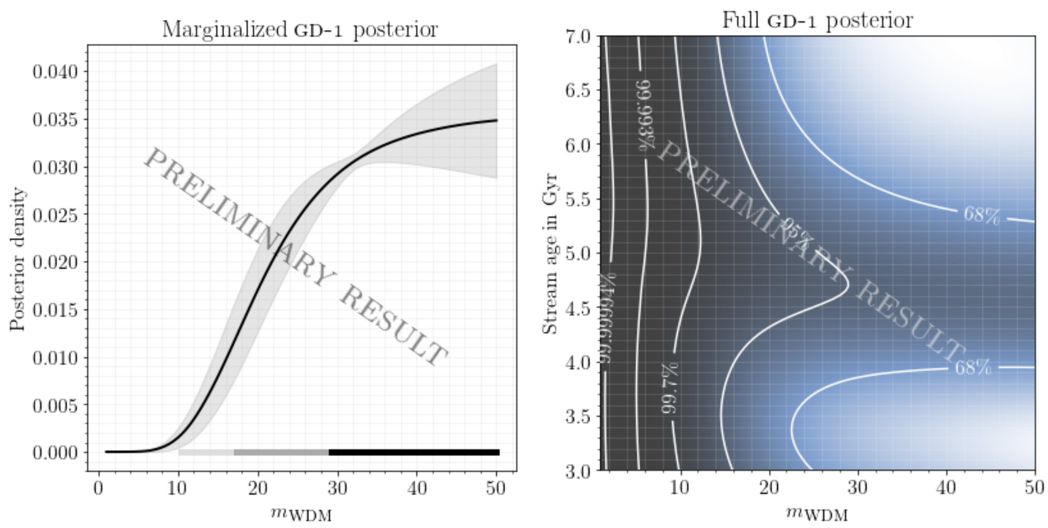


1. Simulation	2. Machine Learning	3. Inference
Run simulator and save data	Train NN classifier, interpret as likelihood ratio estimator	Amortized: cheap to repeat for new data

$$p(\theta|\mathbf{x}) \approx \hat{r}\,(\mathbf{x}|\theta)p(\theta)$$

NRE for stellar streams





—
Credits: Hermans et al, 2021.

Preliminary results for GD-1 suggest a preference for CDM over WDM.

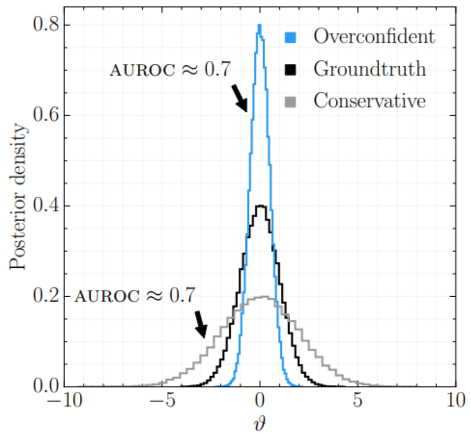
Wait a minute Gilles...
I can't claim that in a paper!
Your neural network must be wrong!

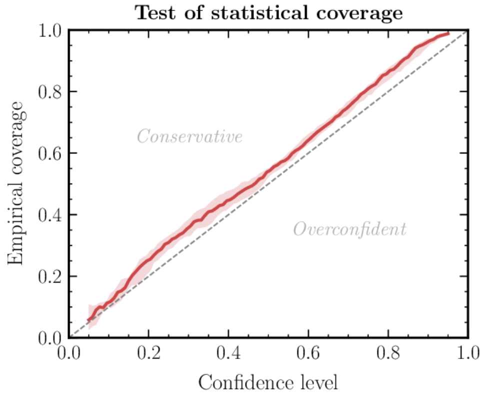
Expected coverage

$$\text{EC}(\hat{p}, \alpha) = \mathbb{E}_{p(\theta,\mathbf{x})}[\theta \in \Theta_{\hat{p}(\theta|\mathbf{x})}(\alpha)]$$

If the expected coverage is close to the nominal coverage probability α, then the approximate posterior $\hat{p}$ is calibrated.

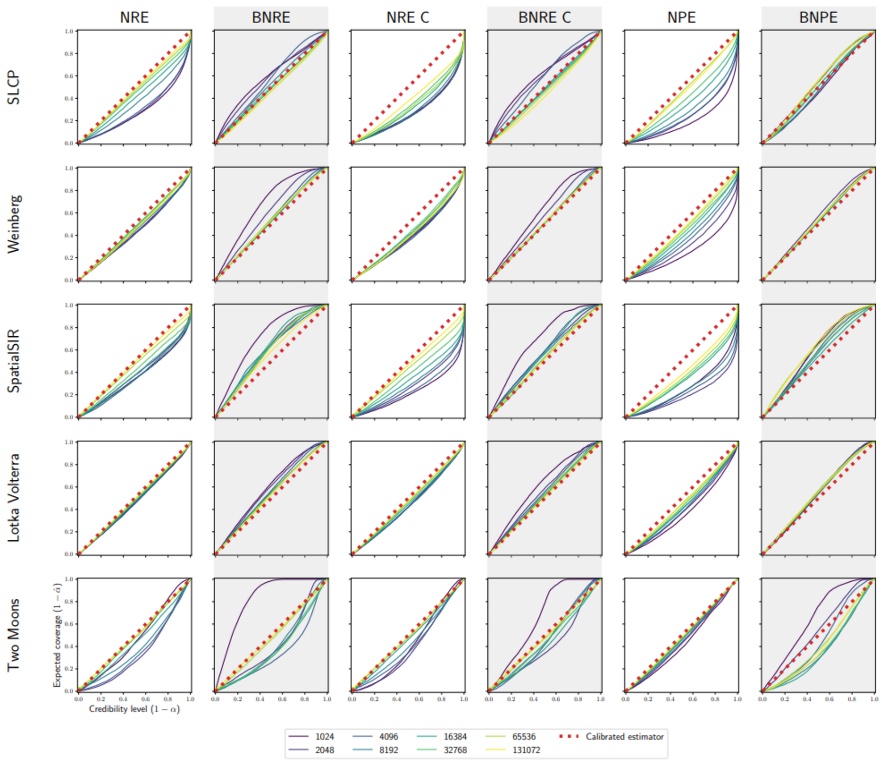
- If $\text{EC} < \alpha$, then the posterior is underdispersed and overconfident.
- If $\text{EC} > \alpha$, then the posterior is overdispersed and conservative.





—
Credits: Hermans et al, 2021; Siddharth Mishra-Sharma, 2021.

Balancing inference
for conservative posteriors



Conservative posteriors can be obtained by enforcing d to be balanced, i.e. such that $\mathbb{E}_{p(\theta,\mathbf{x})}\left[d(\theta,\mathbf{x})\right] = \mathbb{E}_{p(\theta)p(\mathbf{x})}\left[1 - d(\theta,\mathbf{x})\right]$.

—

Credits: Delaunoy et al, 2022; Delaunoy et al, 2023.

Summary

Simulation-based inference is a major evolution in the statistical capabilities for science, as it enables the analysis of complex models and data without simplifying assumptions.

Obstacles remain to be overcome, such as the curse of dimensionality, the need for large amounts of data, or the necessary robustness of the inference network.

The next frontiers? Let's find out this week!