Physics 361 - Machine Learning in Physics

Lecture 23 – Simulation-Based Inference

April 16th 2024

Moritz Münchmeyer

PHYSICS 406

INTRO TO PARTICLE ASTROPHYSICS

- A Brief Introduction to Particle Physics
- The Highest-Energy Particles
- Neutrinos and Neutrino Astronomy
- Gravitational Waves
- Gamma-Rays
- Dark Matter
- Multi-Messenger Astronomy
- Particle Interactions and Detection Principles
- Data Analysis Opportunities from the IceCube Neutrino Observatory

MW 14:30 - 15:45, 3 Cr

Final project

- You will write **a paper on an application of machine learning to physics** of your choice. Your paper needs to contain a computational analysis, which generally will mean applying a machine learning method to some data set.
- You can **work alone or in groups of two.**
- The paper should be **5 to 10 pages**.
- If you are already doing research in physics or a related field, you can write the paper on this topic if you wish.
- Your **paper will be due on Sunday May 5th at midnight**.
- We want to **know your topic by April 16th at latest**. We will make a list today at the end of the lecture.

- We will have a brief presentation of your results in the last lecture, May 2nd. We should have about 5 minutes per student / pair.

Simulation-Based Inference

Overview

References

- This presentation is based on the review “The frontier of simulation-based inference” (Cranmer et al 2019)
  - https://arxiv.org/abs/1911.01429

- Software packages for SBI include
  - https://sbi-dev.github.io/sbi/ simulation-based inference in python
  - https://docs.pyro.ai/en/stable/index.html Pyro - probabilistic machine learning with tensor flow

Inference with Simulations

- **Complex phenomena in physics and other sciences can often be described with simulations** (while analytic treatment is often impossible)
- Examples:
  - Formation of the universe
  - Particle shower induced by a cosmic ray in the atmosphere
  - Particle tracks generated by colliding protons at the LHC
  - Climate, Weather, Epidemics etc.
- We want to **infer physical parameters**, that go into the simulation (such as the age of the universe or the mass of the Higgs) by **“comparing” the outcome of simulations with measured data**.
- This is usually not straight forward to do. Problems include:
  - Simulations usually don’t provide explicit likelihoods.
  - We need a way to get statistical and systematic error bars.
  - It may be difficult to run enough simulations.
  - Simulations may have uncertainties themselves.

Implicit vs Explicit inference

- In most situations a simulation does not provide a probability density (likelihood) $\mathscr{L}(x \,|\, \theta)$ of observations given parameters. Such simulations are sometimes called **implicit models**.
- Implicit means that their **likelihood cannot be computed explicitly**, i.e. it is not computationally tractable. We only get samples of the simulation.
- On the other hand, models or simulations that do provide a likelihood are called **explicit models**. Recall for example Gaussian likelihoods.
- In this section we focus on implicit models, i.e. most simulations. Implicit inference is also called **simulation-based inference** or **likelihood-free inference** (a bit of a misnomer since we learn the likelihood from the simulation). These terms usually mean the same.

Simulators

- A simulator is a computer program that takes as input a vector of **parameters** θ, samples a series of internal states or **latent variables** $z_i \sim p_i(z_i \mid \theta, z_{<i})$, and finally produces **a data vector** $x \sim p(x \mid \theta, z)$ as output.
  - θ: the parameters of interest. **z**: Typically unobservable variables of the data generating process (including e.g. initial conditions). **x**: Observations.
- Programs that involve random samplings and are interpreted as statistical models are known as probabilistic programs, and simulators are an example.

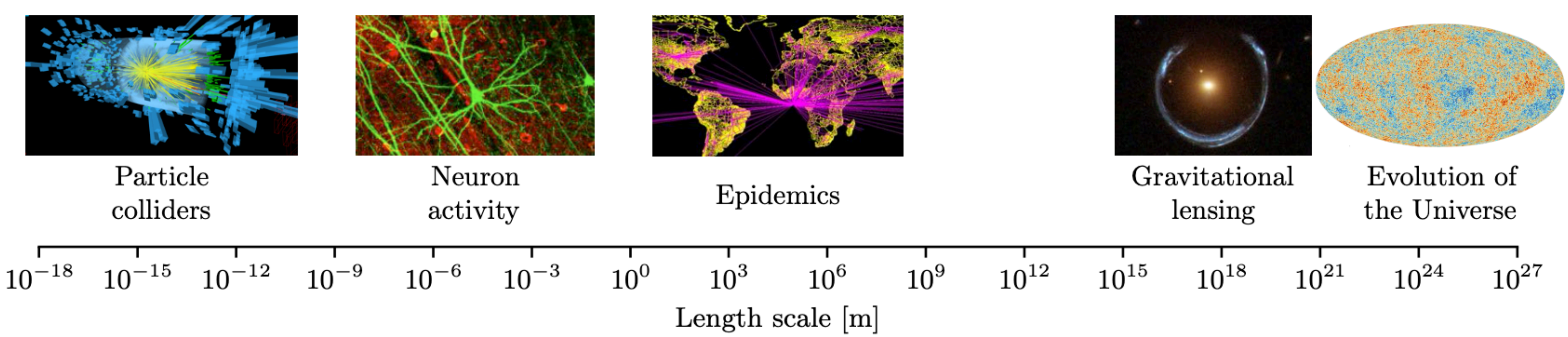
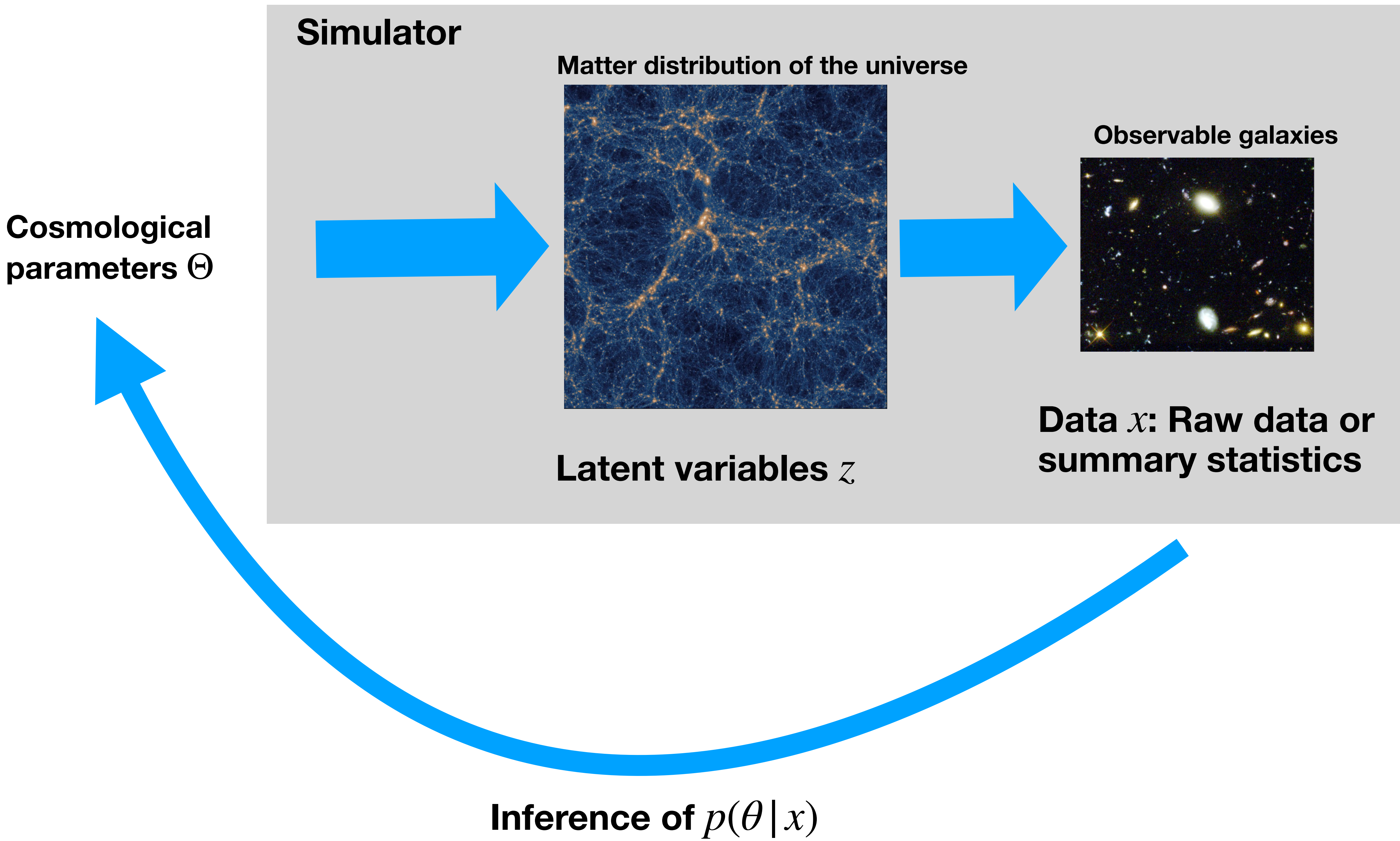


Fig. 1. Examples of phenomena at various length scales described by a diverse set of simulators, each with an intractable likelihood. Contains image material from Refs. (5–9).

Example from cosmology



Inference

- In a Bayesian data analysis, the goal is to find the posterior over the parameters given the data:

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int \mathrm{d}\theta'\, p(x|\theta')\, p(\theta')}$$

- The likelihood is given by an intractable integral over the latent space of the simulator (e.g. we cannot simulate all possible initial conditions)

$$p(x|\theta) = \int \mathrm{d}z\; p(x, z|\theta)$$

- The fundamental challenge of SBI is thus to perform Bayesian (or Frequentist) inference despite this intractability.

Summary statistics

- Often there are many possible choices what the observed data x should be.
- A **summary statistic x'=f(x)** is any compression of the raw data x.
  - Example: Instead of analyzing the raw electric field data from an antenna, we may chose to analyze the measured power spectrum.
- **Low-dimensional summary statistics are usually require**d to analyze high dimensional data to make the computation tractable.
- Traditionally the **choice of good summary statistics was left to domain experts**, which (should) know which parts of the data are important.
- However, **summary statistics are almost always lossy**, i.e. they contain less information on the parameters than the uncompressed data.
- In principle with machine learning one can **learn optimal summary** statistics directly from the simulations. SBI works both with classical and with learned summary statistics.

Approximate Bayesian Computation (ABC)

- There are two general traditional approaches of SBI, Approximate Bayesian Computation (ABC) and Density Estimation. The simplest **rejection sampling ABC algorithm** works as follows:
  - **Purpose**: Estimate parameters when likelihoods are intractable.
  - **Process**:
    1. **Draw** θ from prior.
    2. **Simulate** data x_{sim} using θ.
    3. **Measure** distance $\rho(x_{\mathrm{sim}}, x_{\mathrm{obs}})$.
    4. **Accept** θ if $\rho \leq \epsilon$ (tolerance level).
  - **Repeat**: Until a sufficient sample from the approximate posterior is obtained.
  - **Result**: A sample of parameters that likely generate the observed data.
- The tolerance ϵ controls the trade-off between the approximation quality and the computational feasibility: smaller ϵ leads to a better approximation but requires more simulations and thus more computational effort.
- **Problems**: Need a distance metric between data and simulation. No amortized inference: Need to re-run simulations when we get new data. “Throws away” simulations that are not within the tolerance level.
- There are many versions of ABC that improve its performance, e.g. Population Monte Carlo.

Classical density estimation of the likelihood

- Instead of ABC we can model the likelihood by estimating the distribution of simulated data given parameters.
- The **density estimation** algorithm works as follows:
  - Define a proposal density $\tilde{p}(\theta)$ (not necessarily the prior)
  - Draw parameters from this proposal and run many simulations to generate a data set of N pairs of parameters θ and data x: $\{(\theta_n, x_n)\}_{n=1}^{N}$
  - These are samples of the joint pdf $p(\theta, x) = p(x \mid \theta)\tilde{p}(\theta)$
  - From the samples we can estimate the likelihood $p(x \mid \theta)$ with histograms or kernel density estimation.
- **Advantage**: Amortized inference, i.e. when we take new data we can directly evaluate the likelihood without running new simulations.
- **Problem**: Potentially need more simulations than we can afford to compute (given computational limits).

Simulation-Based Inference

SBI with Neural Density Estimators

Improving SBI with machine learning

- Classical SBI has several shortcomings:
  - **Sample efficiency:** Both ABC and classical density estimation techniques suffer from the curse of dimensionality. The poor scaling means that the number of simulated samples needed to provide a good estimate of the likelihood or posterior can be prohibitively expensive.
  - **Quality of inference:** The reduction of the data to low- dimensional summary statistics invariably discards some of the information in the data about θ, which results in a loss in statistical power

- **Machine learning can improve SBI** in several ways:
  - We can **learn PDFs** with neural networks rather than using histograms or kernel density estimation. These techniques work in higher dimensions.
  - Active learning methods can systematically improve sample efficiency. Draw new simulations where they help the most.
  - Neural networks can work with very high-dimensional data. In particular we can learn optimal summary statistics of the data.

Neural density estimators (NDEs)

- Machine learning offers several methods that can be used to learn the PDF underlying a data set.
- Both unconditional and conditional PDFs can be learned.

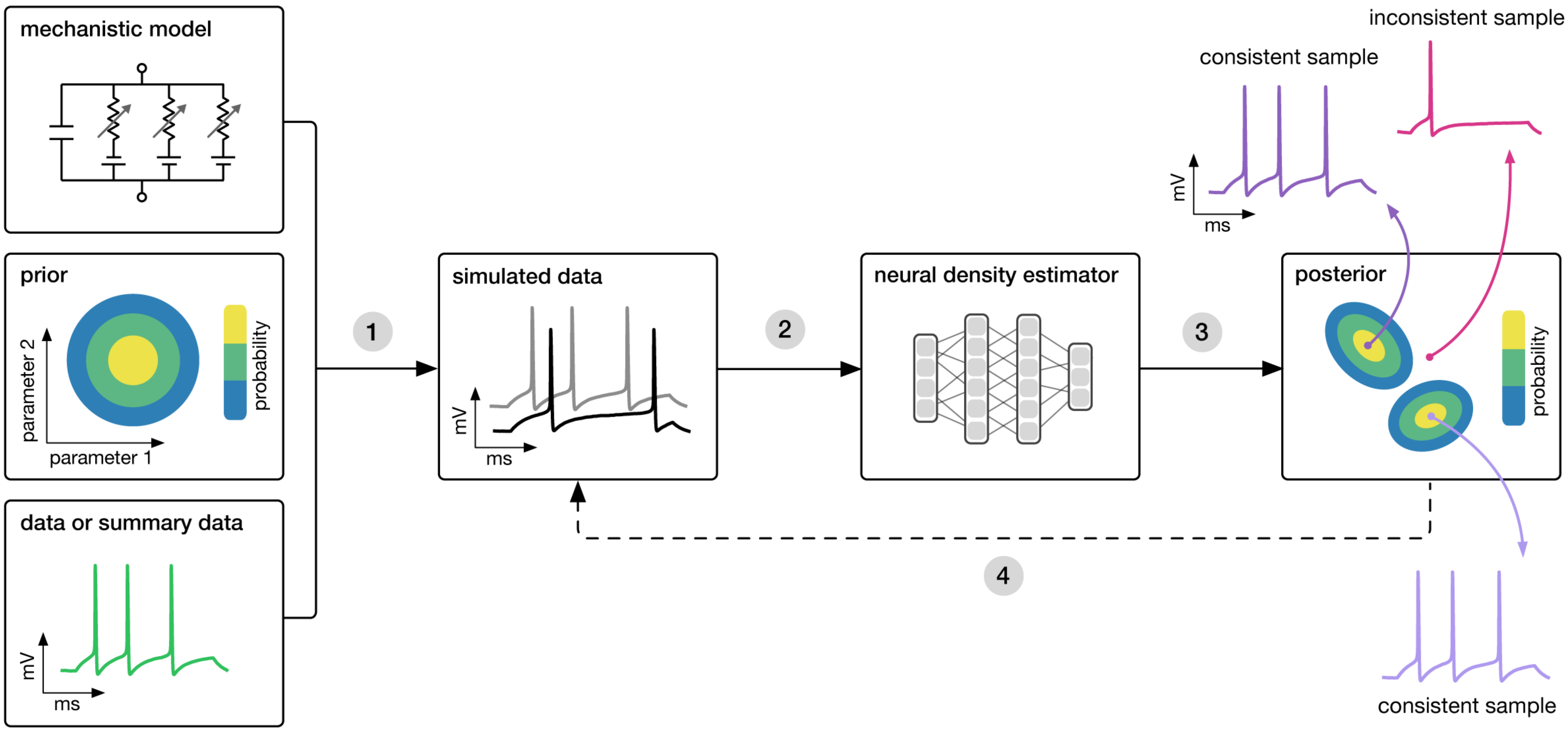
$$\{x_n\}_{n=1}^{N} \Rightarrow p(x)$$

$$\{(\theta_n, x_n)\}_{n=1}^{N} \Rightarrow p(x \mid \theta)$$

- An **NDE learns a PDF** $p(x)$ that, when sampled from, makes samples that "look like the training data". It is trained to make the training data likely under the model. Normal neural networks learn functions, NDEs learn PDFs.
- The dominant NDEs in SBI are **Normalizing Flows**. We will **discuss them in detail in the next lecture**. For now assume that we have some model that can learn PDFs from data.
- A different older NDE is a **mixture density network.** In this model a neural network outputs the parameters (means, variances, and mixture coefficients) of the mixture model (e.g. a collection of Gaussians).

Simulation-based inference with NDEs

inputs: A candidate mechanistic model, prior knowledge or constraints on model parameters, and observational data (or summary statistics thereof).



Goal: Algorithmically identify mechanistic models (simulators) which are consistent with data.

Source: https://sbi-dev.github.io/sbi/

Flavors of simulation-based inference with NDEs

- There are a number of slightly different approaches for SBI. These are (each with several variants):
  - Neural Likelihood Estimation
  - Neural Posterior Estimation
  - Neural Ratio Estimation

- These have different properties and in some situations one is more suitable than the other. We will briefly discuss the first two.
- Mathematical details will follow after we discuss normalizing flows.

Neural Likelihood Estimation (NLE)

- In neural likelihood estimation we learn the likelihood from simulated pairs of model parameters and data:

$$\{(\theta_n, x_n)\}_{n=1}^{N} \rightarrow \mathscr{L}(x\,|\,\theta)$$

- After training the NDE on our simulated data, we can then evaluate the likelihood of observed data from our measurement.

$$\mathscr{L}(x^{obs}\,|\,\theta)$$

- Now we can proceed with normal Bayesian data analysis. That usually means that we sample from the posterior with MCMC:

$$p(\theta\,|\,x^{obs}) \propto \mathscr{L}(x^{obs}\,|\,\theta)p(\theta)$$

- NLE is amortized (no new sims needed for new data). However sometimes it takes too much training data to learn the likelihood everywhere. Sequential Neural Likelihood Estimation (S-NLE) only learns the likelihood near the data and thus saves samples, at the cost of not being amortized anymore.

Neural Posterior Estimation (NPE)

- You might wonder why why learn the likelihood and not the posterior which is our ultimate goal. Learning the posterior is indeed a possibility.
- From a simulated data set

$$\{(\theta_n, x_n)\}_{n=1}^{N}$$

drawn from a proposal density $\tilde{p}(\theta)$ it is possible to directly learn the posterior

$$p(\theta | x)$$

- An advantage of learning the posterior directly is that we do not need to run an MCMC anymore. The model directly outputs the desired posterior, i.e. our parameter measurement. A disadvantage is that it is difficult to explore different prior distributions of the parameters.

Testing the neural density estimator

- After training the NLE or NPE, it is important to **assess that the learned PDFs are correct**, and we thus do not over- or under-estimate parameter uncertainties. Most importantly, we want to be sure that our posterior is not overconfident.
- For example, if we did not train the NDE on enough simulations, we will get incorrect likelihoods and posteriors.
- It is not possible to formally guarantee that our machine learning model is correct. However, one can **run a series of tests**. In general, the more test simulations we have (independent from the training simulations), the more convincing our tests can be.
- Typical tests include in particular
    - **Posterior Predictive Checks (PPC)**
    - **Simulation-based calibration**
    - For more details, see https://sbi-dev.github.io/sbi/ diagnostics.

Simulation-Based Inference

Examples of SBI

Gaussian Toy example

- We will first have a look at the Gaussian demo from https://github.com/sbi-dev/sbi/blob/main/tutorials/00_getting_started_flexible.ipynb

- Model parameters θ: 3 parameters
- Output parameters x: 3 parameters

- This example uses Neural Posterior Estimation.

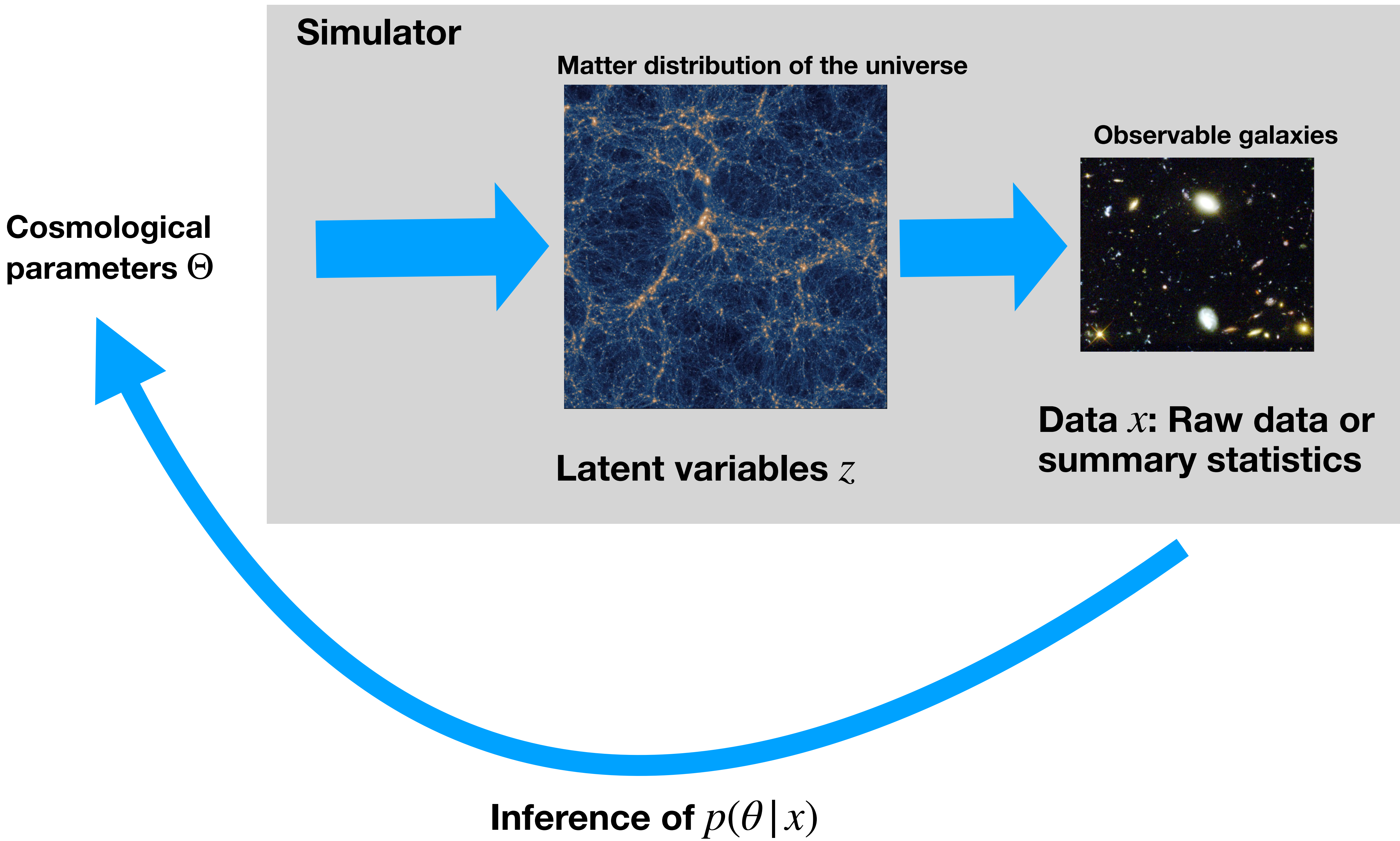
- (Discussion on Colab)

SBI in Cosmology

- SBI is used in cosmology to extract more information about the fundamental parameters of the universe from the observed galaxy distribution.
- We will have quick look at these papers:
    - https://arxiv.org/pdf/2211.00723.pdf SIMBIG: A Forward Modeling Approach To Analyzing Galaxy Clustering
    - https://arxiv.org/pdf/2310.15246.pdf SIMBIG: The First Cosmological Constraints from Non-Gaussian and Non-Linear Galaxy Clustering

- The SBI method these papers use is **NPE with a "Masked Autoregressive Flow"**. We will see in the **next lecture** how this method works in detail.

Recall example from cosmology



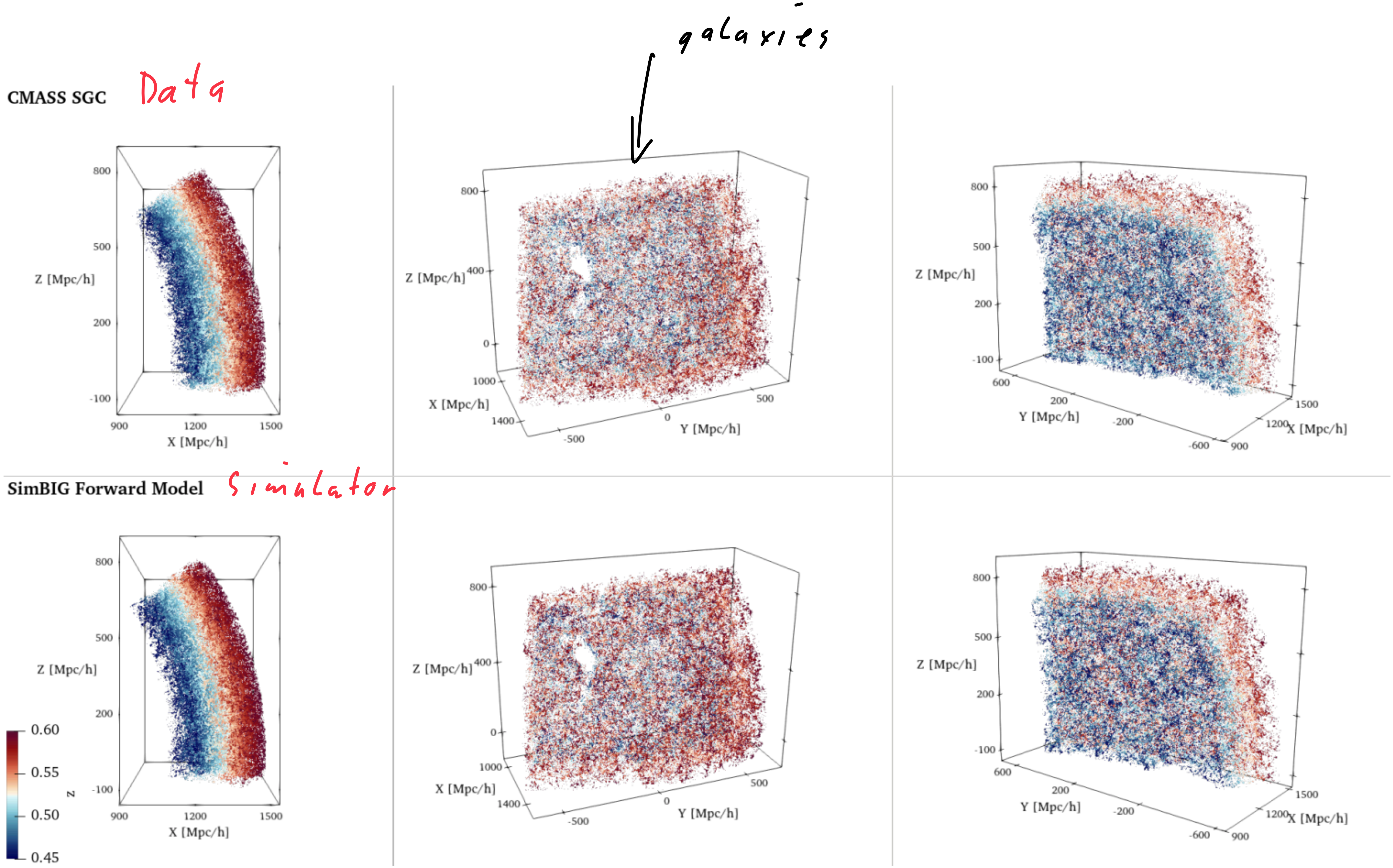


Fig. 1. The SIMBIG forward model produces simulated galaxy samples with the same survey geometry and observational systematics as the observed BOSS CMASS SGC galaxy sample. We present the 3D distribution of the galaxies from three different viewing angles. The colormap represents the redshift of the galaxies. In the top set of panels, we present the distribution of galaxies in the CMASS sample. In the bottom, we present the distribution of a simulated galaxy sample, generated from our forward model. The SIMBIG galaxy samples are constructed from QUIJOTE N-body dark matter simulations using an HOD model that populates dark matter halos identified using the ROCKSTAR algorithm. The 3D distributions illustrate that our forward model is able to generate galaxy distributions that are difficult to statistically distinguish from observations. For more comparisons of the 3D distributions, we refer readers to .

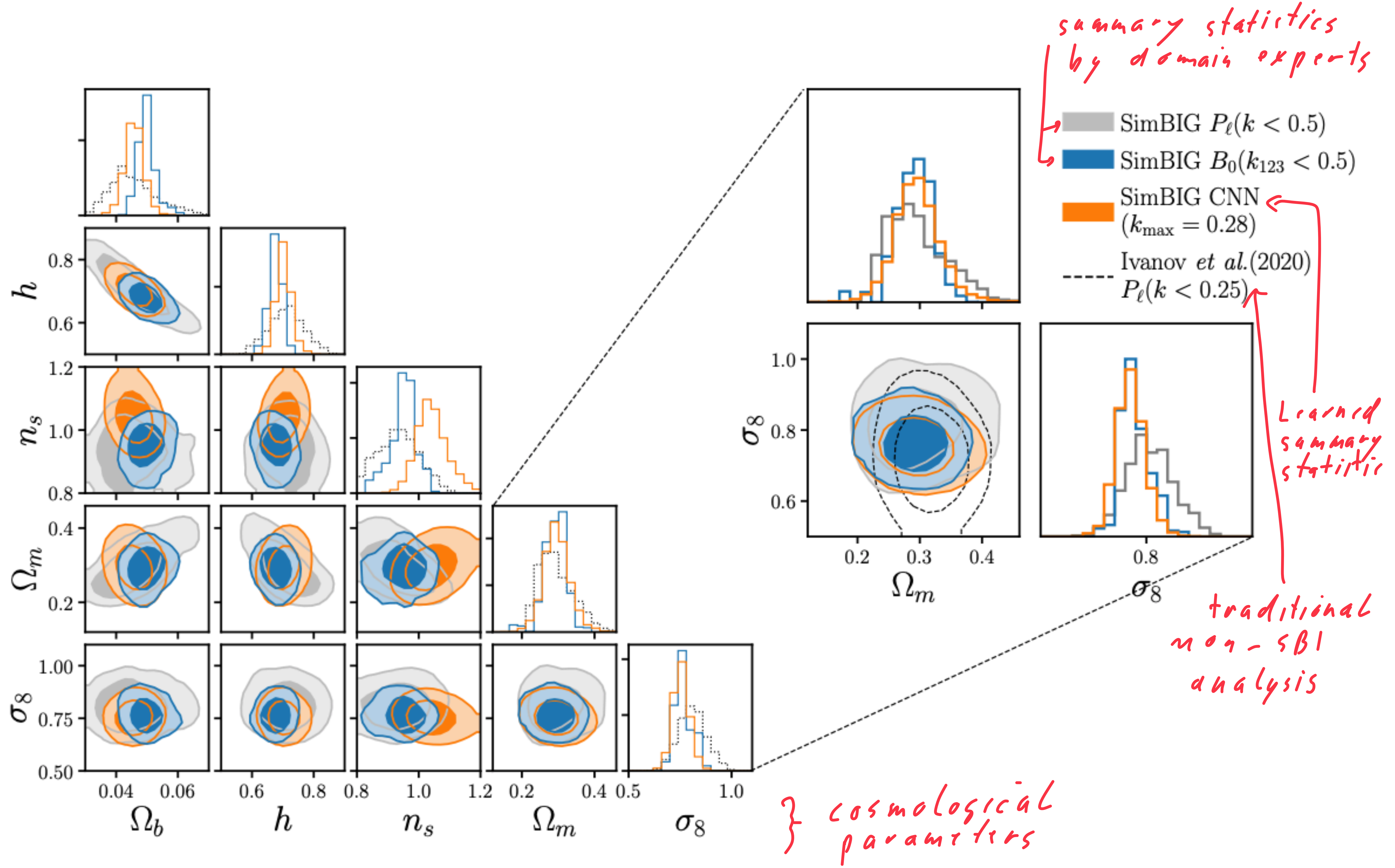


Figure 1. *Left*: Posteriors of cosmological parameters inferred from B_0 (blue) and CNN (orange) using SIMBIG. All posteriors include a ω_b prior from BBN studies. The contours mark the 68 and 95 percentiles. For comparison, we include the posterior from the SIMBIG P_ℓ analysis (gray). *Right*: We focus on the posteriors of Ω_m and σ_8, the parameters that can be most significantly constrained by galaxy clustering. We include the posterior from the Ivanov et al. (2020) PT-based $P_\ell(k < 0.25\,h/\text{Mpc})$ analysis for reference (black dashed). The SIMBIG B_0 and CNN constraints are significantly tighter yet consistent with P_ℓ constraints.

Course logistics

- **Reading for this lecture:**
    - This lecture was based mostly on https://arxiv.org/abs/1911.01429