Towards reliable simulation-based inference and beyond

Dagstuhl Seminar
Machine Learning for Science

September 20, 2022

Gilles Louppe
g.louppe@uliege.be

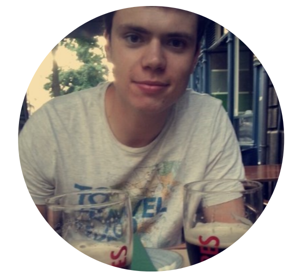
LIÈGE université

Kyle Cranmer

Johann Brehmer

Joeri Hermans

Antoine Wehenkel



Norman Marlier

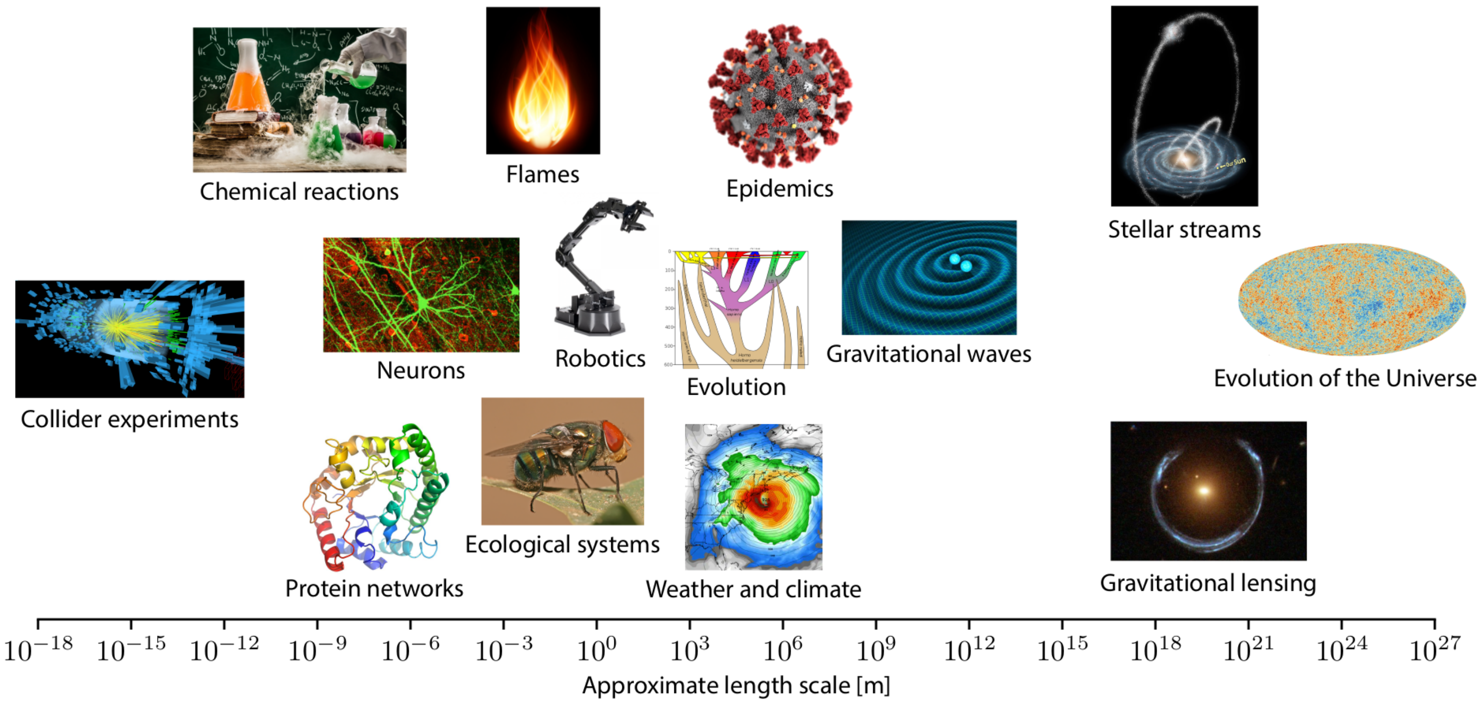
Siddharth Mishra-Sharma

Christoph Weniger

Arnaud Delaunoy

Malavika Vasist

Francois Rozet


Chemical reactions
Flames
Epidemics
Stellar streams
Collider experiments
Neurons
Robotics
Evolution
Gravitational waves
Evolution of the Universe
Protein networks
Ecological systems
Weather and climate
Gravitational lensing
10^{-18} 10^{-15} 10^{-12} 10^{-9} 10^{-6} 10^{-3} 10^{0} 10^{3} 10^{6} 10^{9} 10^{12} 10^{15} 10^{18} 10^{21} 10^{24} 10^{27}
Approximate length scale [m]

$$v_x = v\cos(\alpha), \;\; v_y = v\sin(\alpha),$$

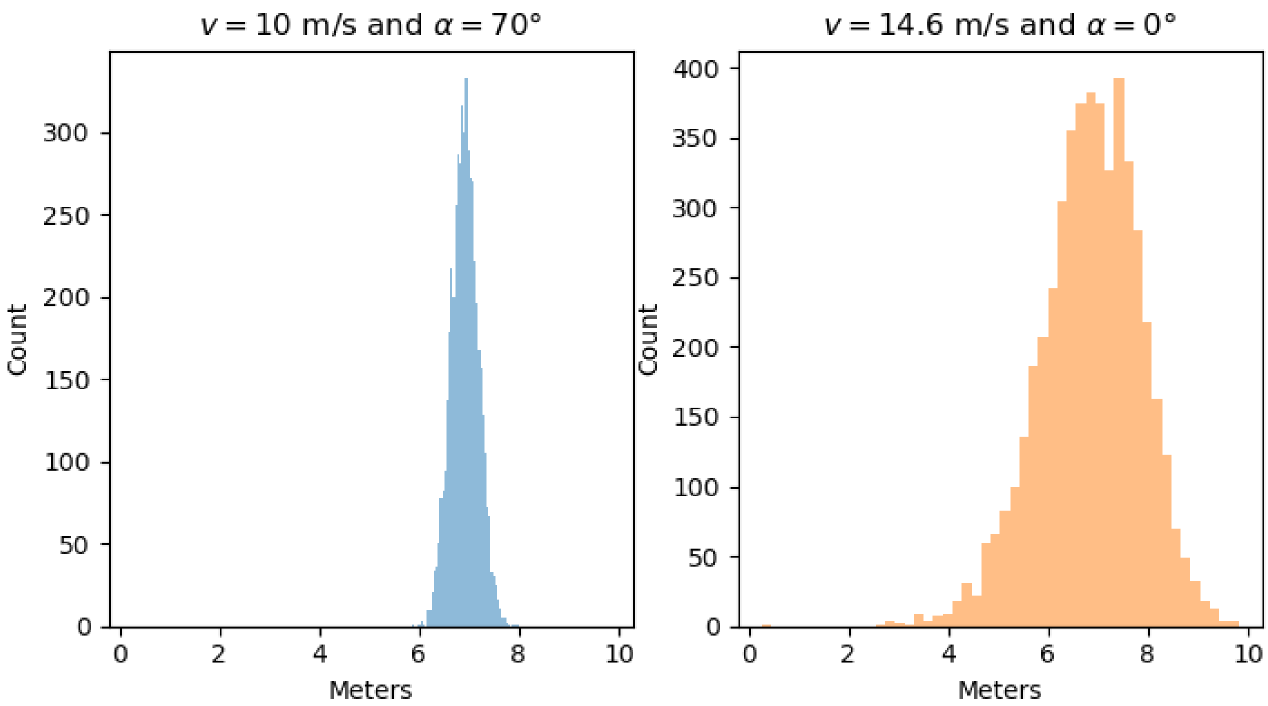
$$\frac{dx}{dt} = v_x\,, \;\; \frac{dy}{dt} = v_y, \frac{dv_y}{dt} = -G.$$

```
def simulate(v, alpha, dt=0.001):
    v_x = v * np.cos(alpha)  # x velocity m/s
    v_y = v * np.sin(alpha)  # y velocity m/s
    y = 1.1 + 0.3 * random.normal()
    x = 0.0

    while y > 0: # simulate until ball hits floor
        v_y += dt * -G  # acceleration due to gravity
        x += dt * v_x
        y += dt * v_y

    return x + 0.25 * random.normal()
```

The computer simulator defines the likelihood function $p(x|\theta)$ implicitly.



What parameter values θ are the most plausible?

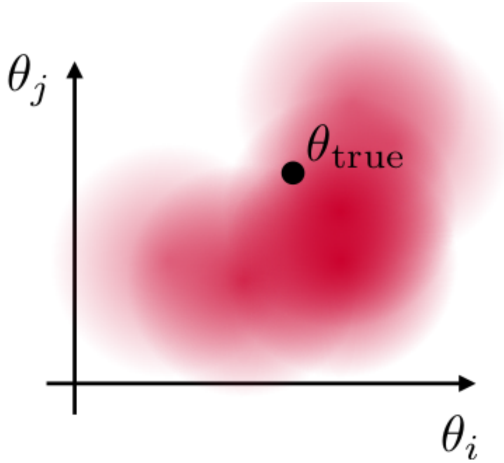
Bayesian inference

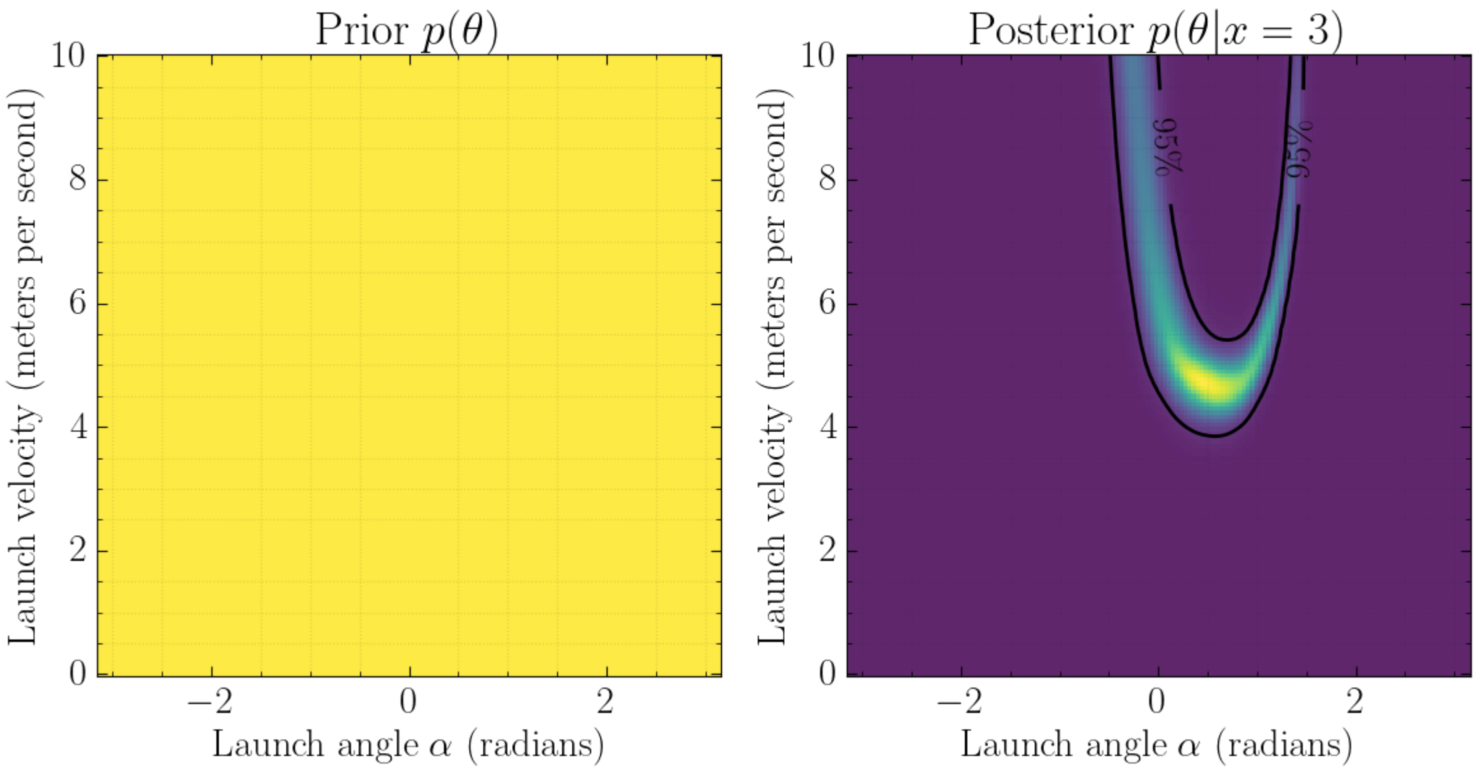
Start with

- a simulator that can generate N samples $x_i \sim p(x_i|\theta_i)$,
- a prior model $p(\theta)$,
- observed data $x_{\text{obs}} \sim p(x_{\text{obs}}|\theta_{\text{true}})$.

Then, estimate the posterior

$$p(\theta|x_{\text{obs}}) = \frac{p(x_{\text{obs}}|\theta)p(\theta)}{p(x_{\text{obs}})}$$





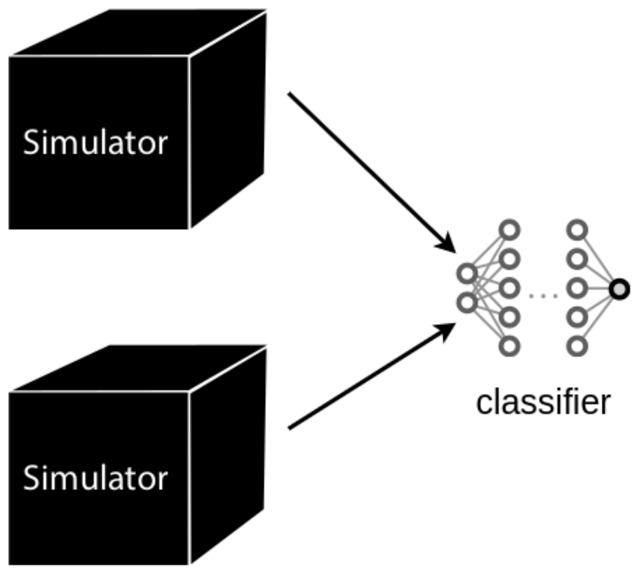
Prior $p(\theta)$
Launch velocity (meters per second)
Launch angle α (radians)
Posterior $p(\theta|x = 3)$
Launch velocity (meters per second)
Launch angle α (radians)
95%
95%

Neural ratio estimation (NRE)

The likelihood-to-evidence $r(x|\theta) = \frac{p(x|\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)p(\theta)}$ ratio can be learned, even if neither the likelihood nor the evidence can be evaluated:

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

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



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

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

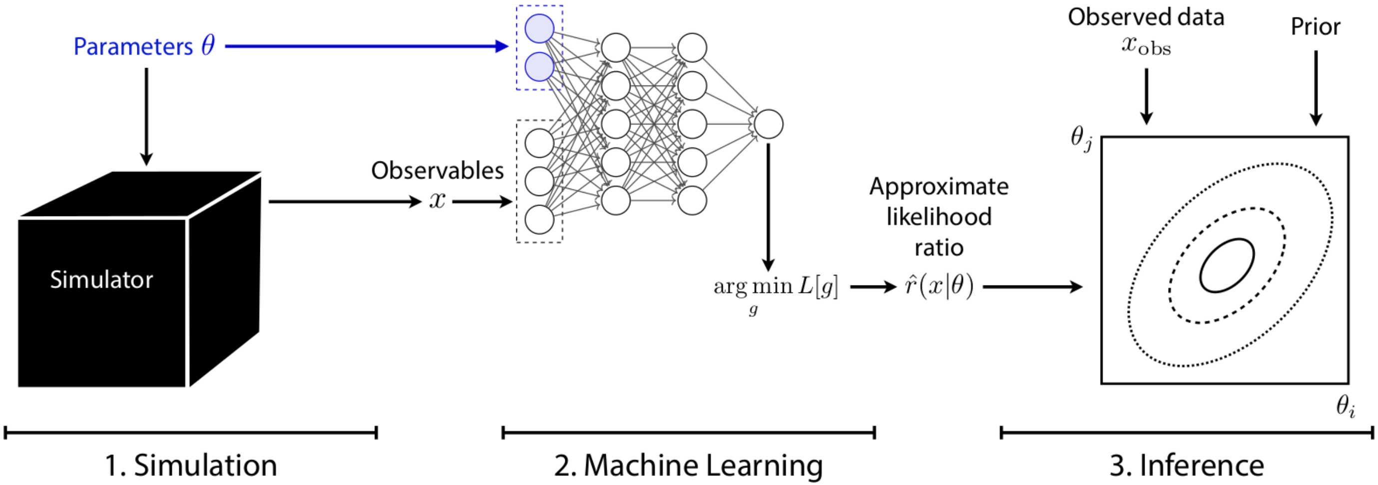
The solution d found after training approximates the optimal classifier

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

Therefore,

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \approx \hat{r}(x|\theta)p(\theta)$$



Run simulator and save data

Train NN classifier, interpret as likelihood ratio estimator

Amortized: cheap to repeat for new data

Constraining dark matter with stellar streams

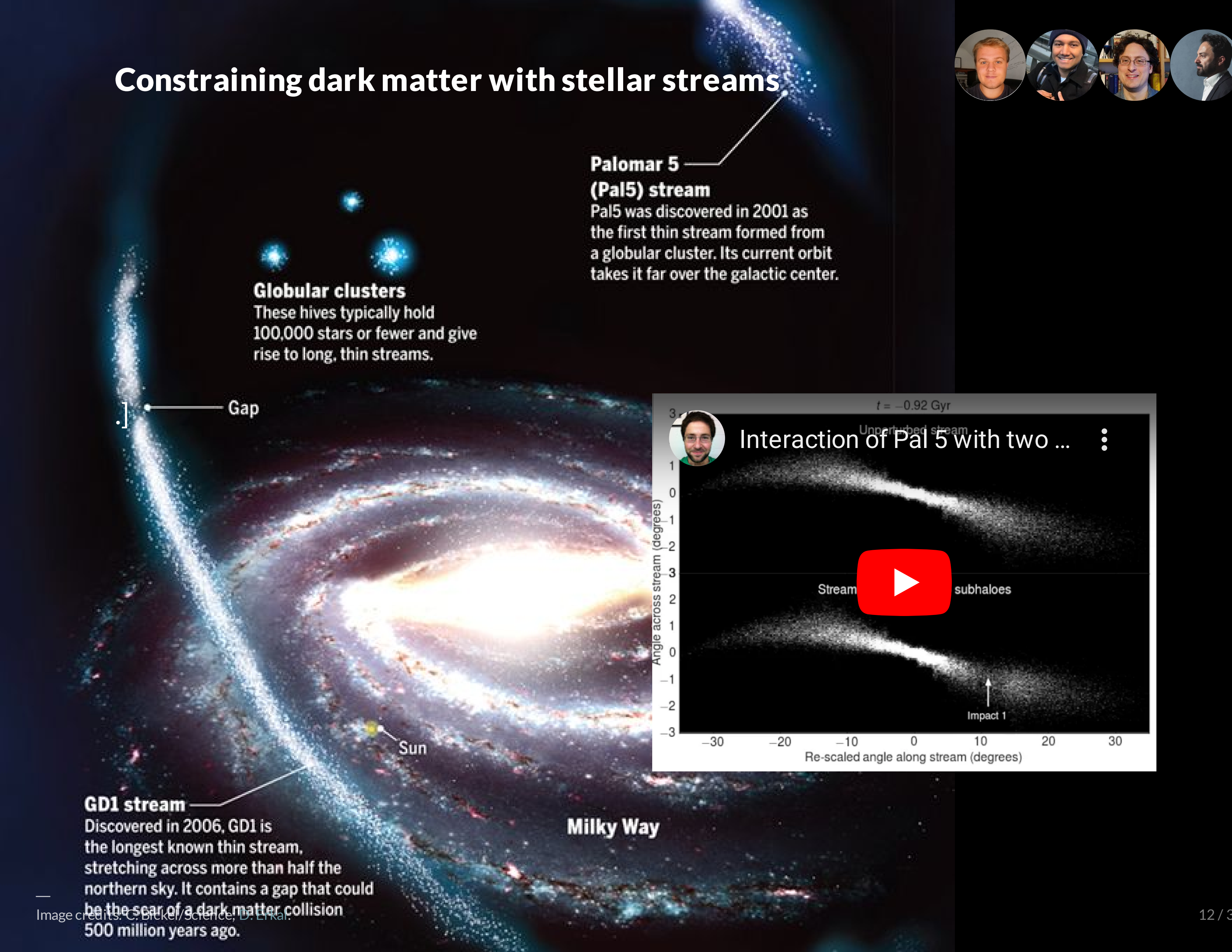
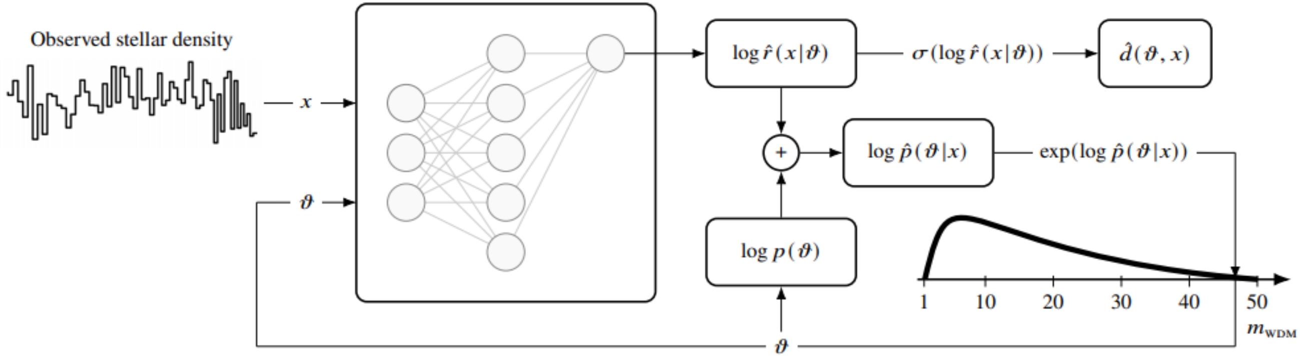
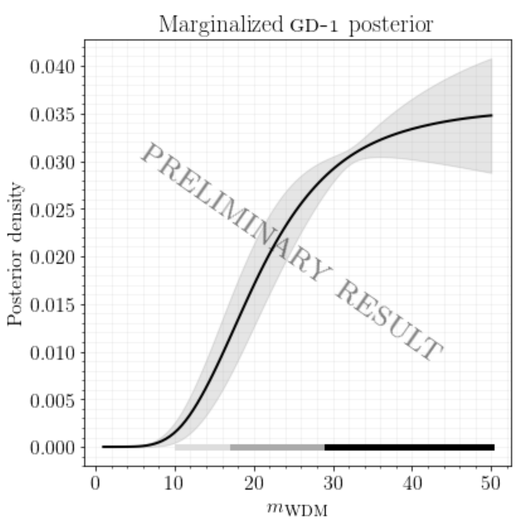
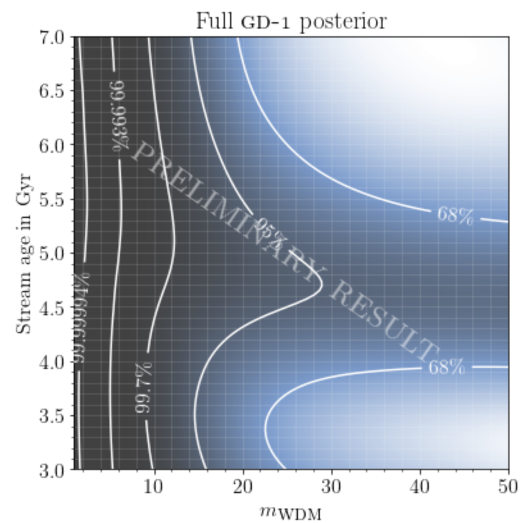


Image credits: C. Bickel/Science; D. Erkal.







—

Credits: Hermans et al, 2021.

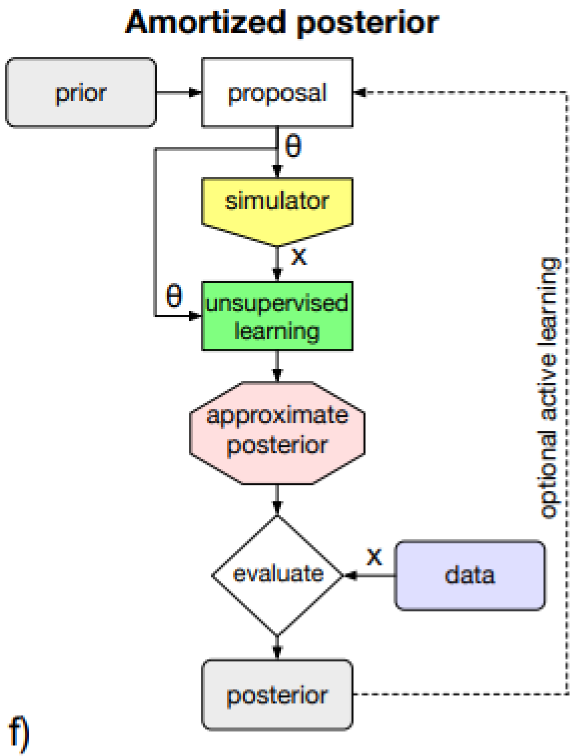
Preliminary results for GD-1 suggest a **preference for CDM over WDM**.

Neural Posterior Estimation (NPE)

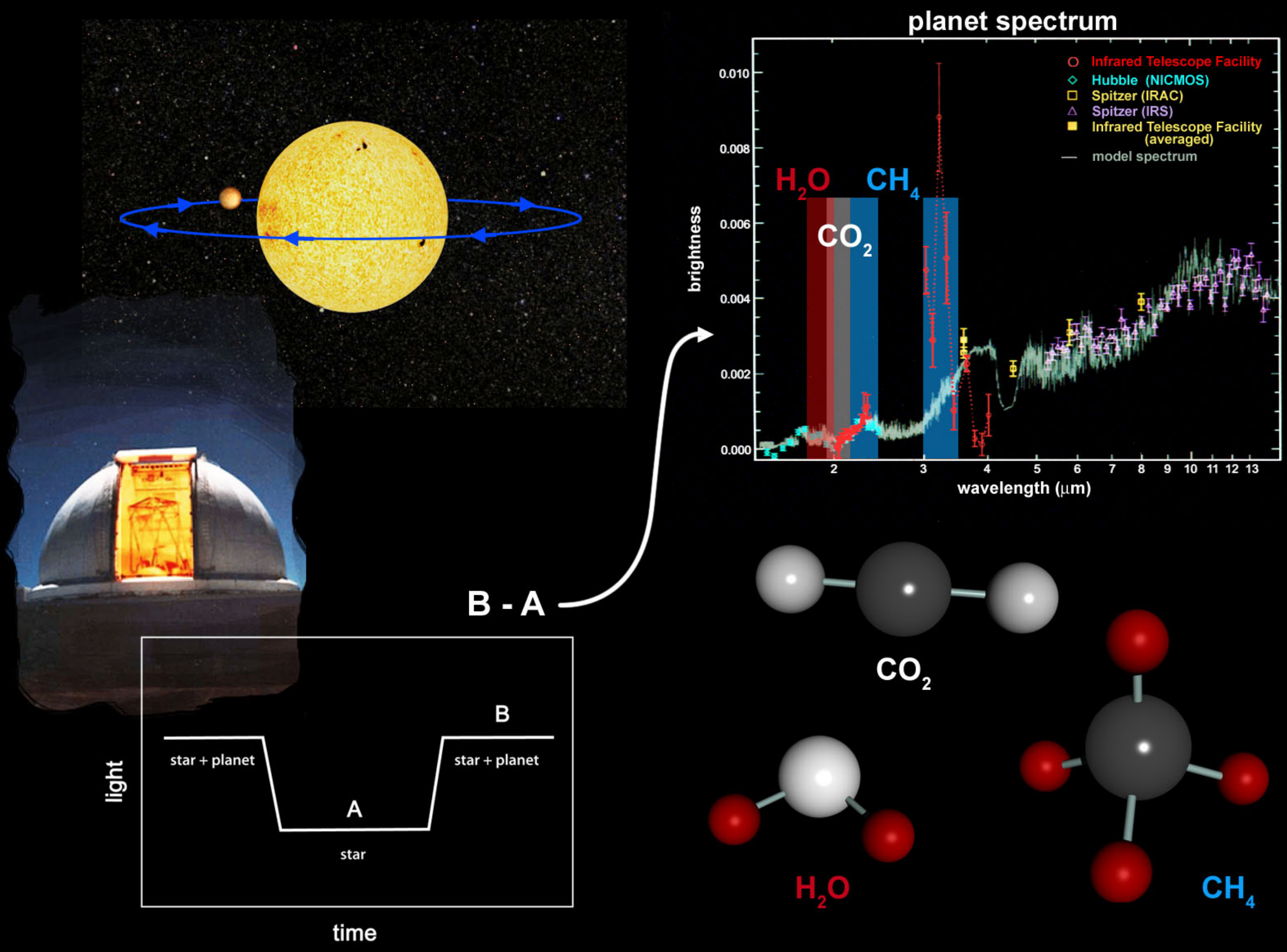
Use variational inference to directly estimate the posterior, by solving

$$\min_{q_\phi} \mathbb{E}_{p(x)} \left[\mathrm{KL}(p(\theta|x) || q_\phi(\theta|x)) \right]$$

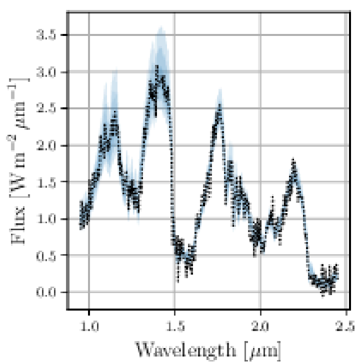
where q_ϕ is a neural density estimator, such as a normalizing flow.



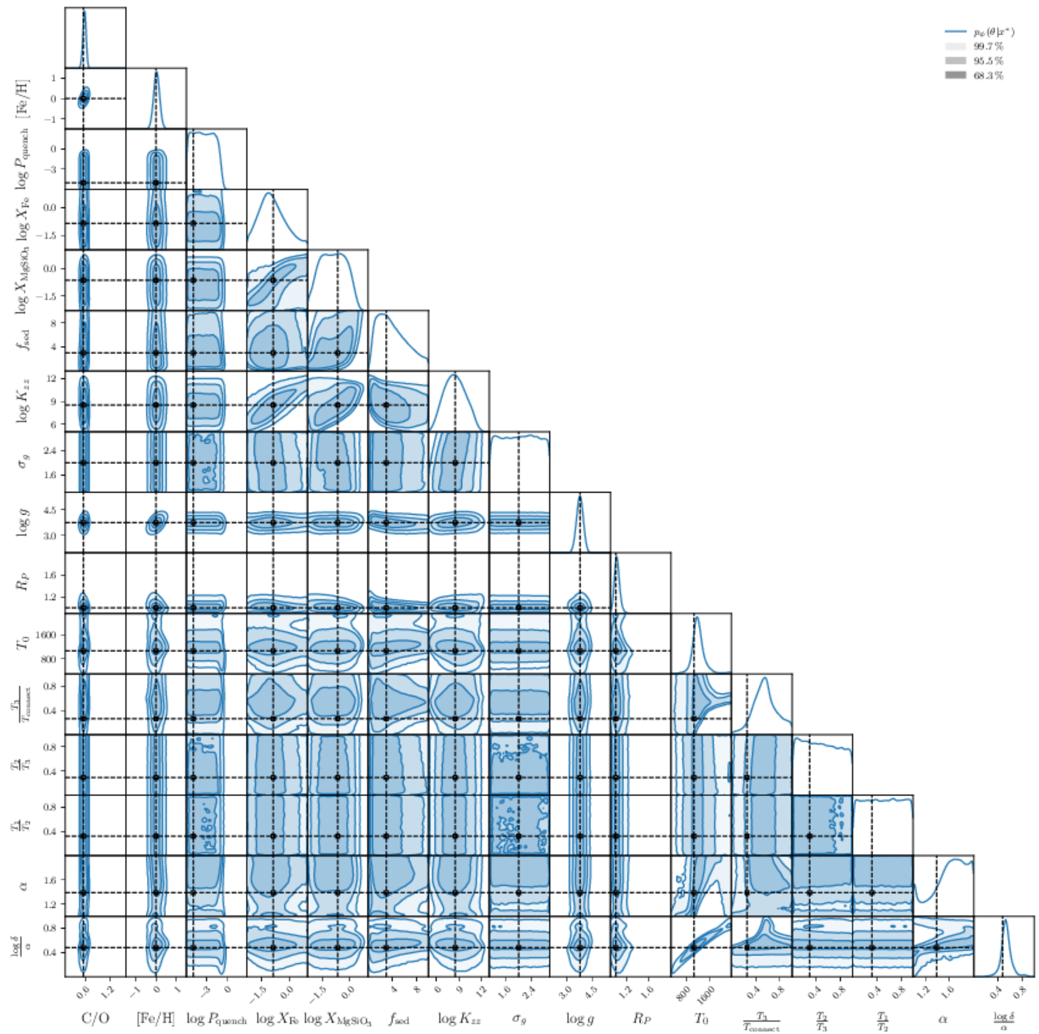
Exoplanet atmosphere characterization



—
Credits: NASA/JPL-Caltech, 2010.



Flux [W m^-2 μm^-1]
Wavelength [μm]
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1.0
1.5
2.0
2.5

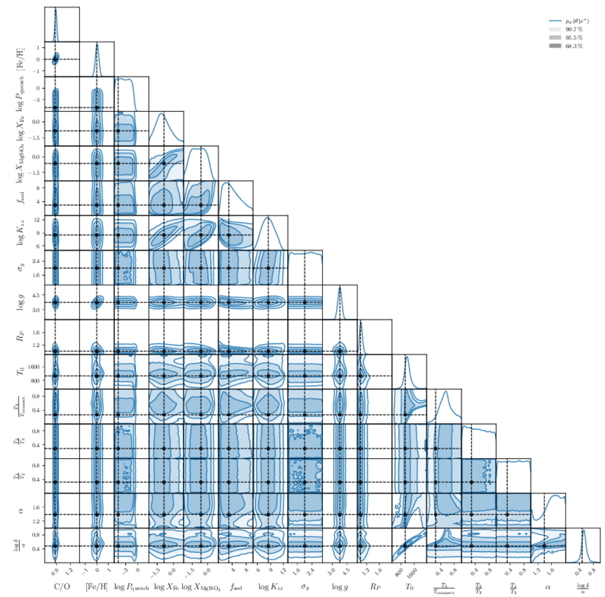


C/O
[Fe/H]
log P_quench
log X_Fe
log X_MgSiO3
f_sed
log K_zz
σ_g
log g
R_P
T_0
α

Computational faithfulness

$$\hat{p}(\theta|x) = \text{sbi}(p(x|\theta), p(\theta), x)$$

We must make sure our approximate simulation-based inference algorithms can (at least) actually realize faithful inferences on the (expected) observations.

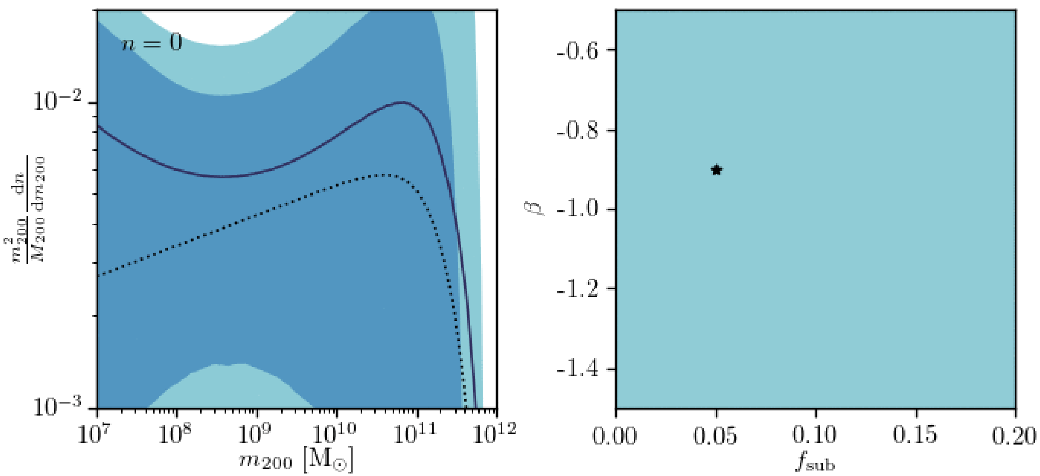


How do we know this is good enough?

Mode convergence:

The maximum a posteriori estimate converges towards the nominal value θ^* for an increasing number of independent and identically distributed observables $x_i \sim p(x|\theta^*)$:

$$\lim_{N\to\infty} \arg\max_\theta p(\theta|\{x_i\}_{i=1}^N)$$
$$= \lim_{N\to\infty} \arg\max_\theta p(\theta) \prod_{x_i} r(x_i|\theta) = \theta^*$$



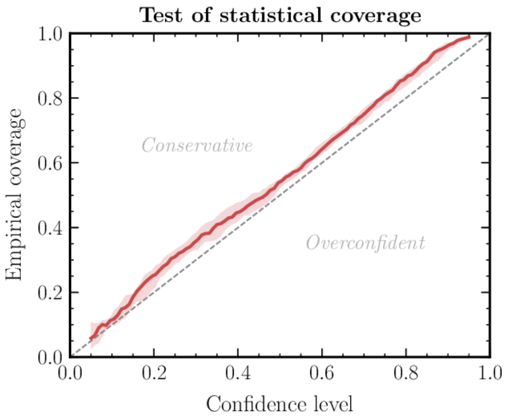
—
Credits: Brehmer et al, 2019.

A common observation at the root of several other diagnostics is to check for the **self-consistency** of the Bayesian joint distribution,

$$p(\theta) = \int p(\theta')p(x|\theta')p(\theta|x)d\theta'\, dx.$$

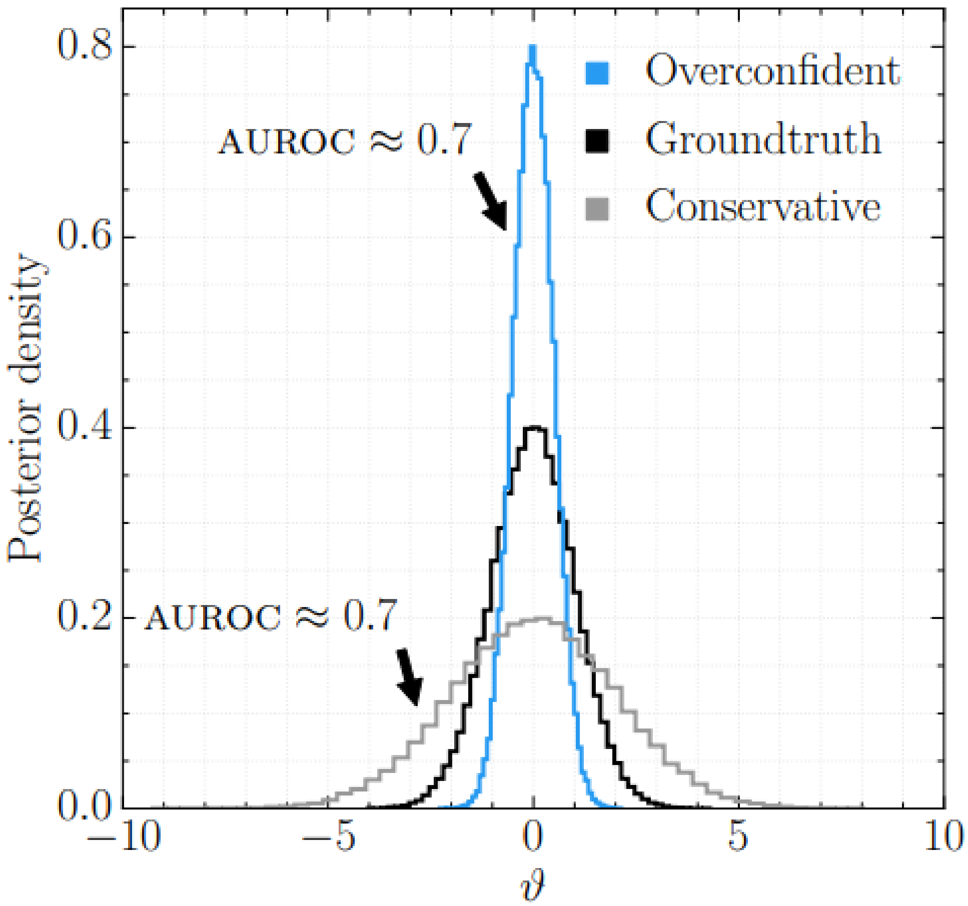
Coverage diagnostic:

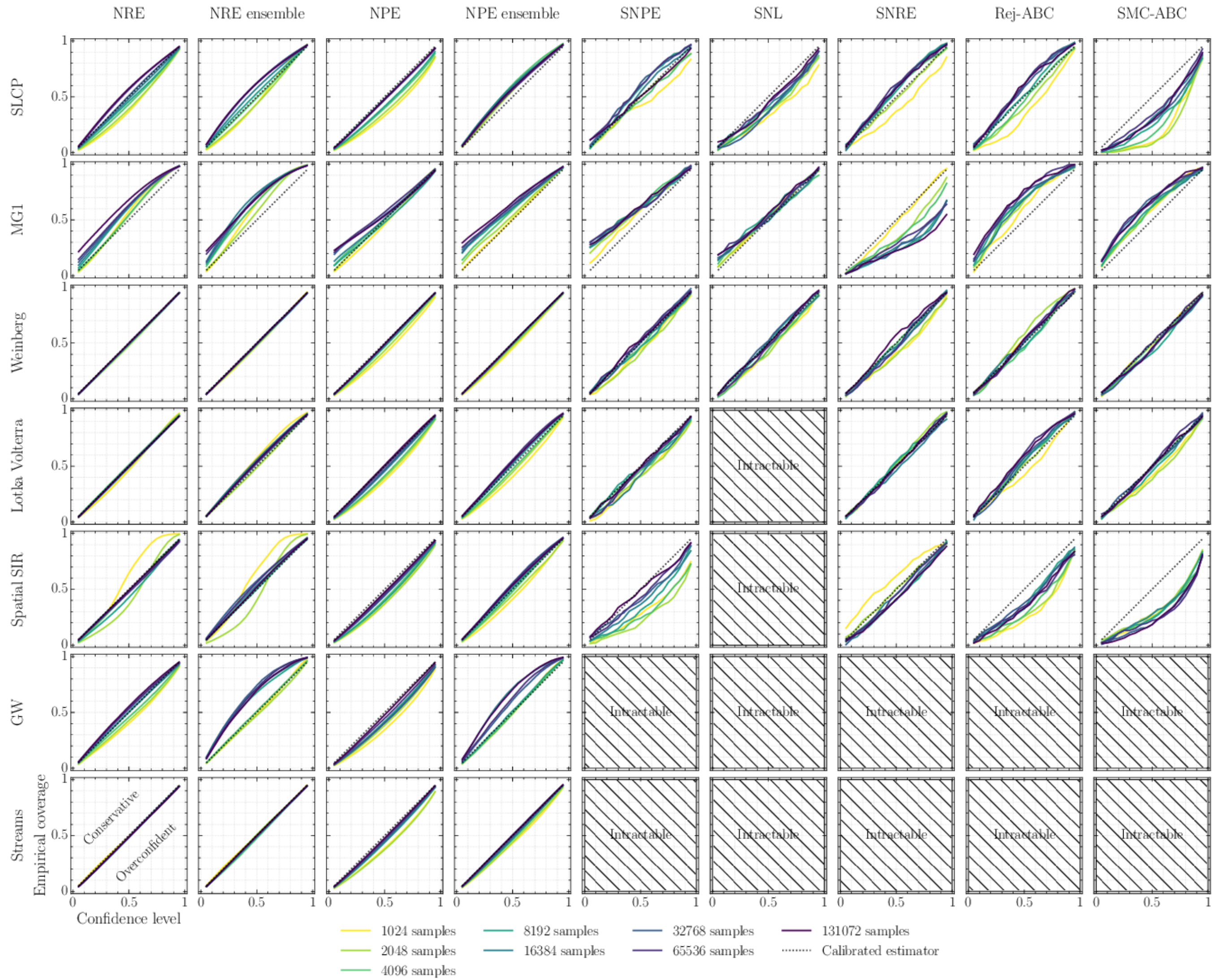
- For $x, \theta \sim p(x, \theta)$, compute the $1-\alpha$ credible interval based on $\hat{p}(\theta|x)$.
- If the fraction of samples for which θ is contained within the interval is larger than the nominal coverage probability $1-\alpha$, then the approximate posterior $\hat{p}(\theta|x)$ has coverage.



—

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


Overconfident
Groundtruth
Conservative
AUROC ≈ 0.7
AUROC ≈ 0.7
Posterior density
0.0
0.2
0.4
0.6
0.8
−10
−5
0
5
10
ϑ



—

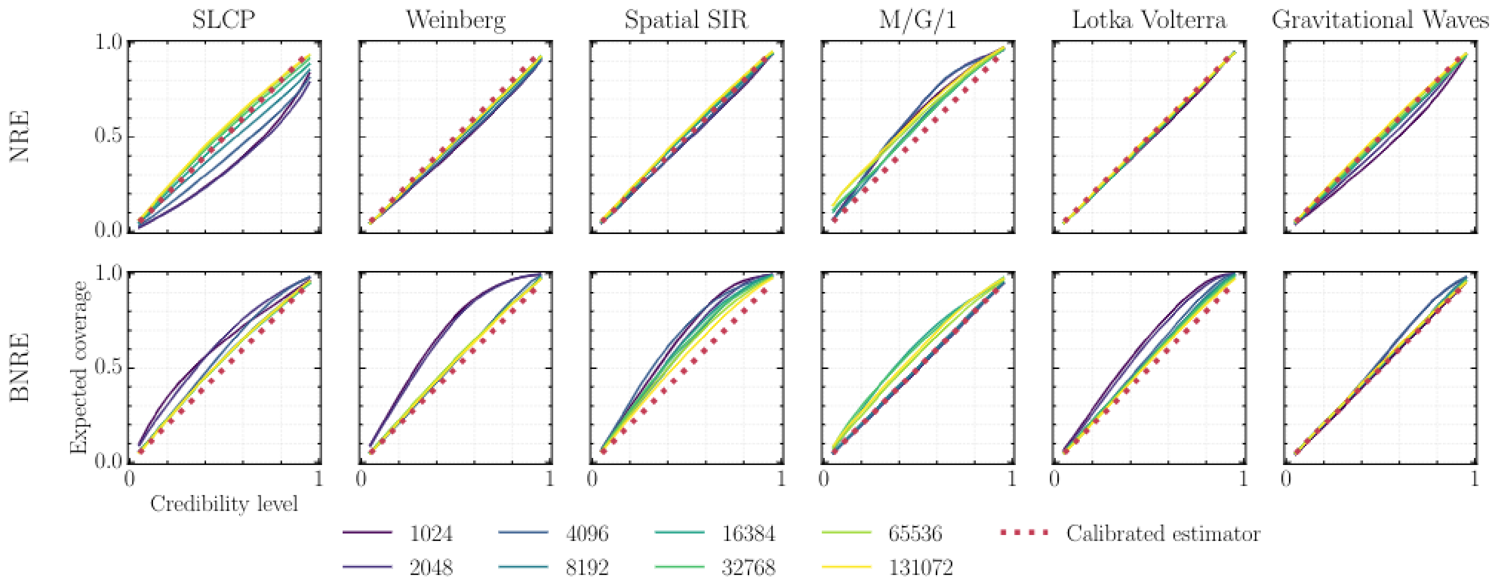
Credits: Hermans et al, 2021.

What if diagnostics fail?

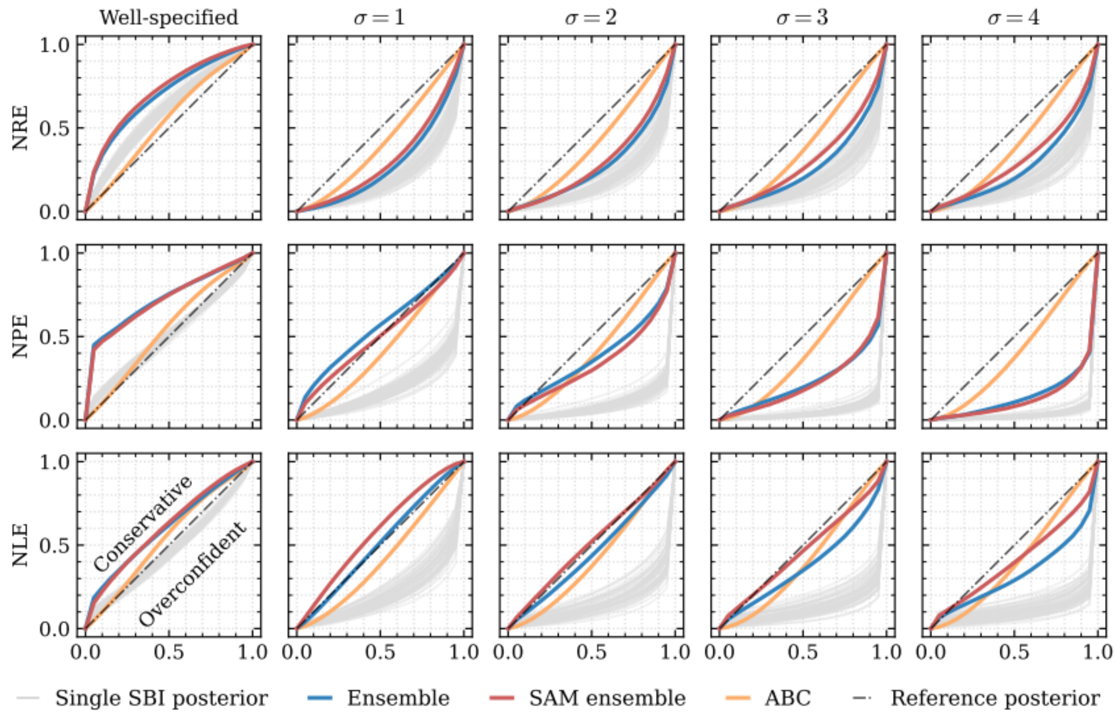
Balanced NRE

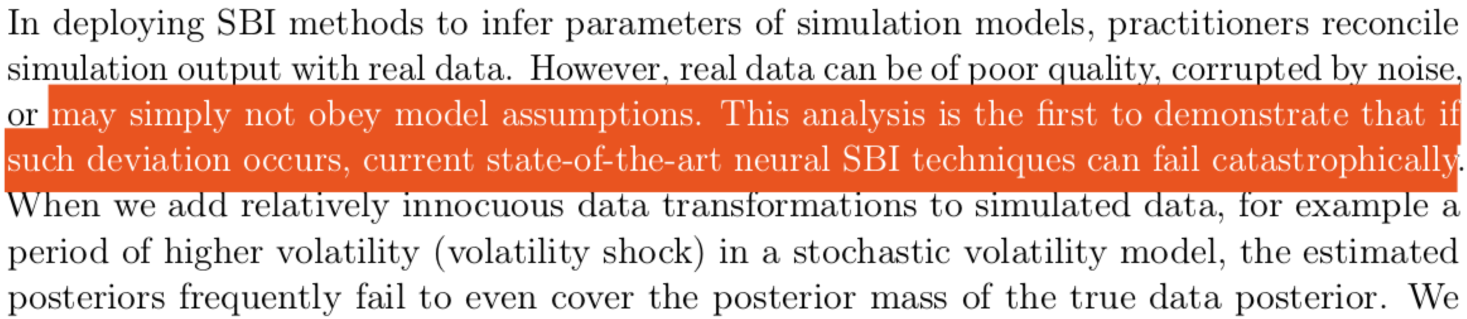
Enforce neural ratio estimation to be **conservative** by using binary classifiers $\hat{d}$ that are balanced, i.e. such that

$$\mathbb{E}_{p(\theta,x)}\left[\hat{d}\left(\theta,x\right)\right]=\mathbb{E}_{p(\theta)p(x)}\left[1-\hat{d}\left(\theta,x\right)\right].$$



Wait a minute... What if your model is wrong?


Well-specified
σ = 1
σ = 2
σ = 3
σ = 4
NRE
NPE
NLE
Conservative
Overconfident
1.0
0.5
0.0
Single SBI posterior
Ensemble
SAM ensemble
ABC
Reference posterior


In deploying SBI methods to infer parameters of simulation models, practitioners reconcile simulation output with real data. However, real data can be of poor quality, corrupted by noise, or may simply not obey model assumptions. This analysis is the first to demonstrate that if such deviation occurs, current state-of-the-art neural SBI techniques can fail catastrophically. When we add relatively innocuous data transformations to simulated data, for example a period of higher volatility (volatility shock) in a stochastic volatility model, the estimated posteriors frequently fail to even cover the posterior mass of the true data posterior. We

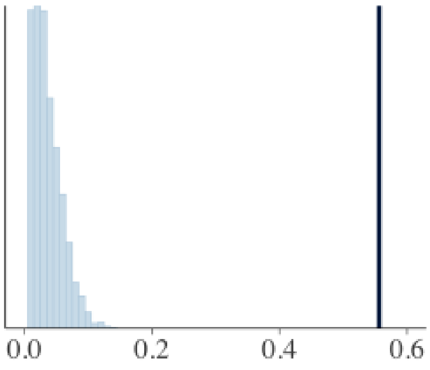
Posterior predictive checks

If a model is a good fit, then we should be able to use it to generate data that resemble the data we observe.

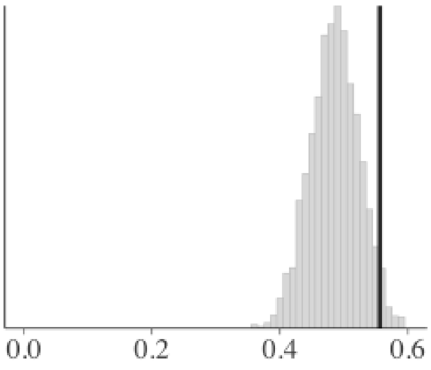
Formally, this can be diagnosed with posterior predictive checks that generates data x^{sim} according to the posterior predictive distribution

$$p(x^{\text{sim}}|x) = \int p(x^{\text{sim}}|\theta)p(\theta|x)d\theta,$$

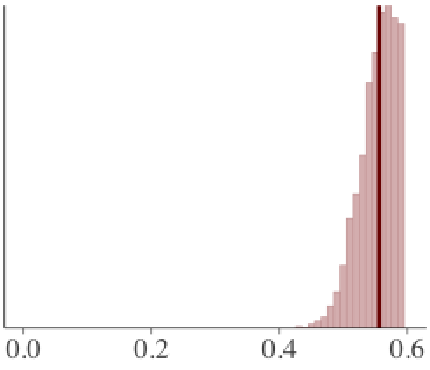
or summary statistics $T(x^{\text{sim}})$ thereof.



(a) Model 1



(b) Model 2

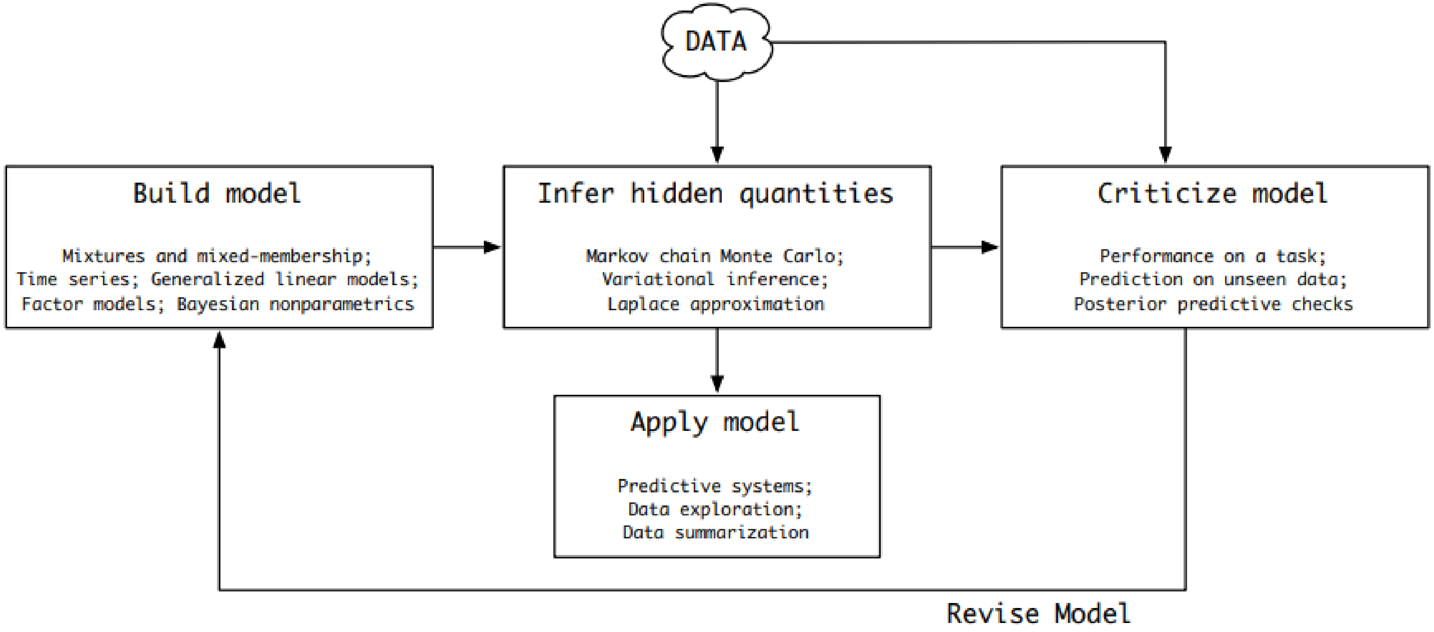


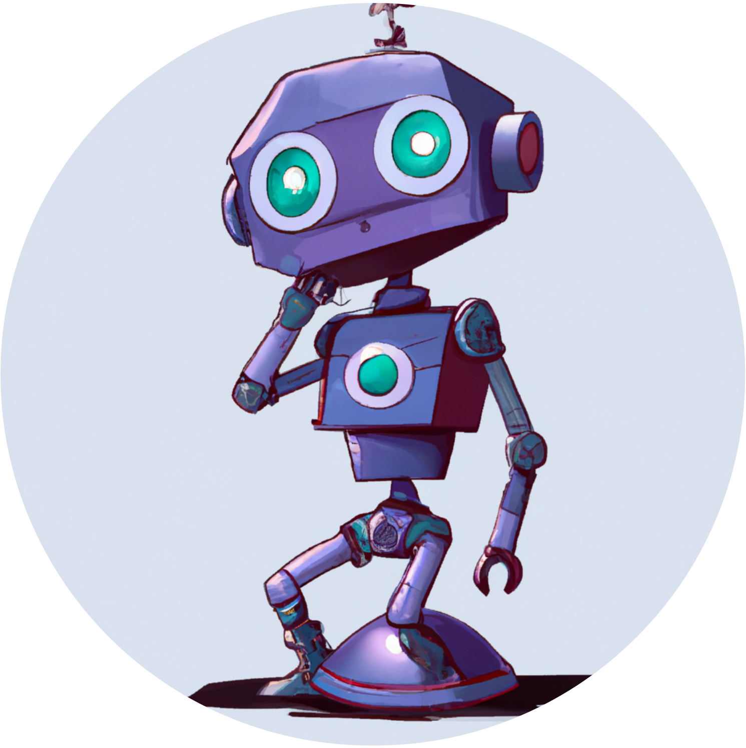
(c) Model 3

Fig. 7: *Histograms of statistics* skew(y_{rep}) *computed from 4000 draws from the posterior predictive distribution. The dark vertical line is computed from the observed data. These plots can be produced using* `ppc_stat` *in the* `bayesplot` *package.*

Credits: Gabry et al, 2017.

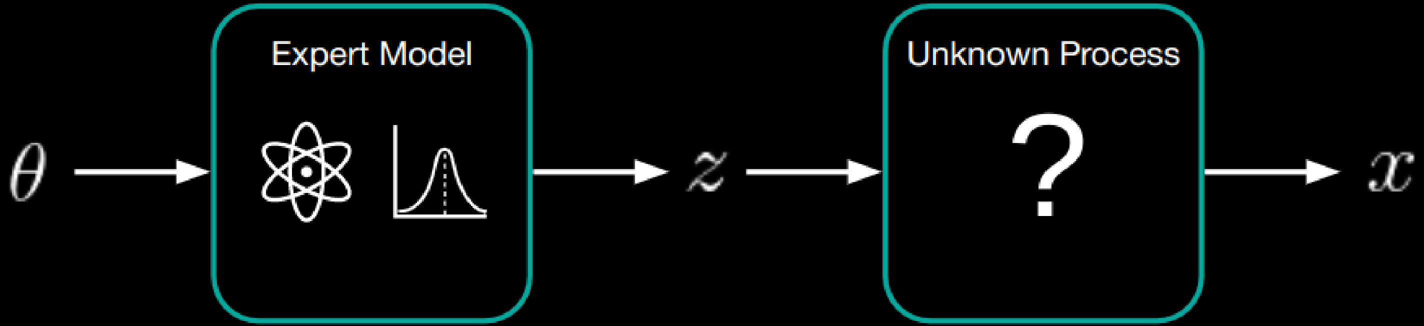
Box's loop: build, compute, critique, repeat





Wait a minute... Can't I machine learn the model discrepancy?

Hybrid models

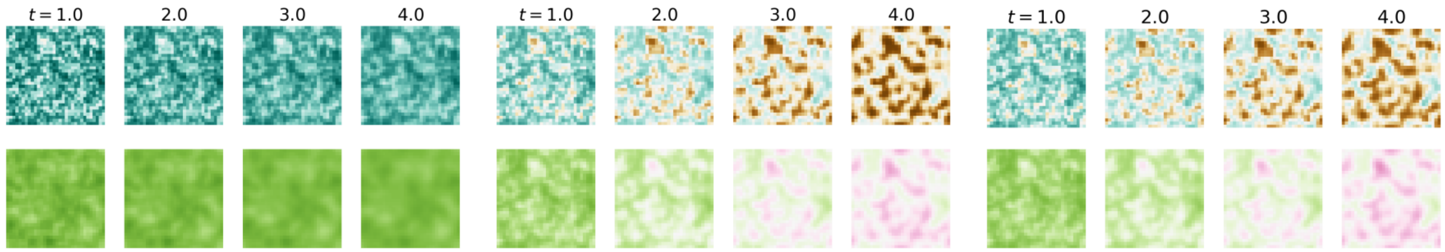


APHYNITY (Yin et al, 2021)

$$\min_{F_p, F_a} ||F_a||$$

subject to

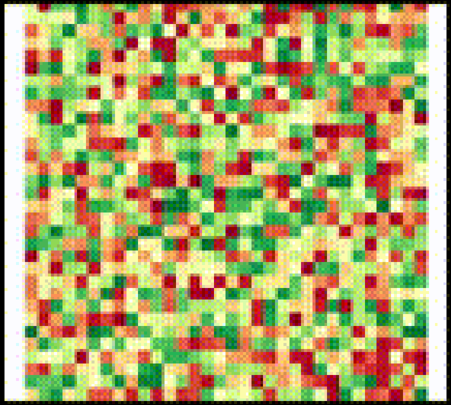
$$\forall t \quad \frac{dX_t}{dt} = (F_p + F_a)(X_t)$$



(a) Param PDE (a, b), diffusion-only (b) APHYNITY Param PDE (a, b) (c) Ground truth simulation

—
Credits: Yin et al, 2021.

Watch out for out-of-distribution data!



HVAE Groundtruth Ours

—
Credits: Wehenkel et al, 2022.

Summary

Simulation-based inference is a major evolution in the statistical capabilities for science, enabled by advances in machine learning.

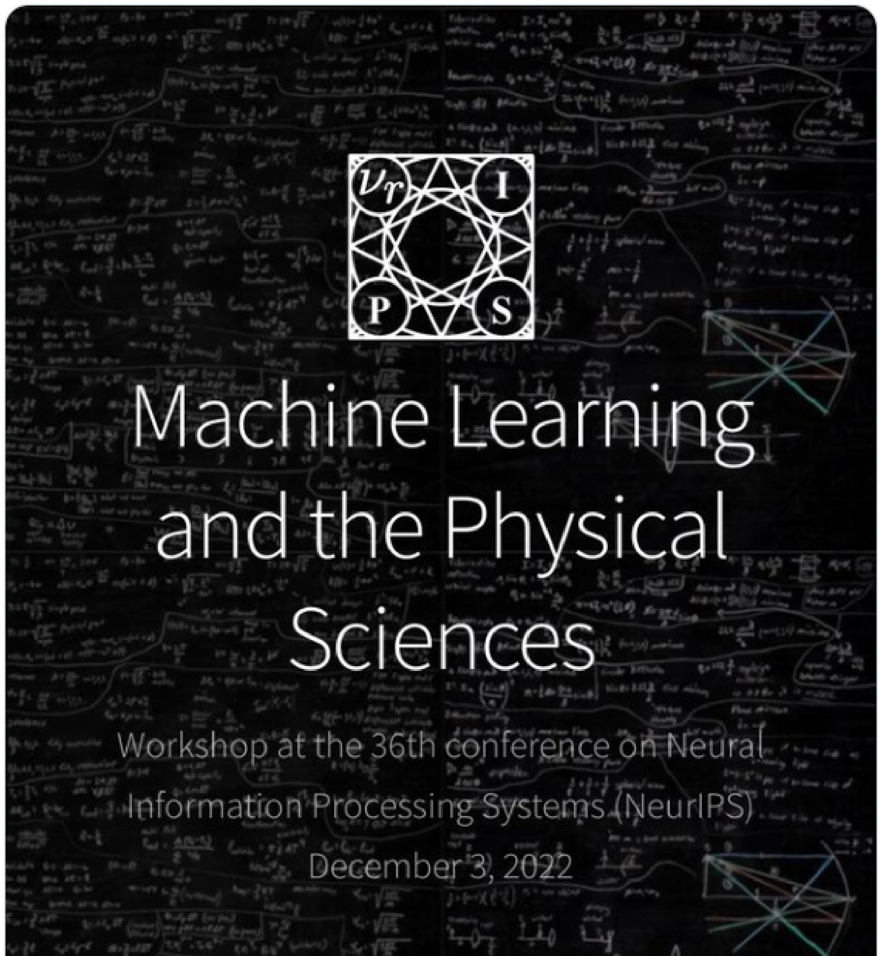
Need to reliably and efficiently evaluate the quality of the posterior approximations.

Further advances will eventually augment incomplete physical models with AI.

Kyle Cranmer
@KyleCranmer

Counting down

Important Update: We have extended the submission deadline for the Machine Learning and Physical Sciences workshop at #NeurIPS2022 to September 29 #ml4ps2022



The end.