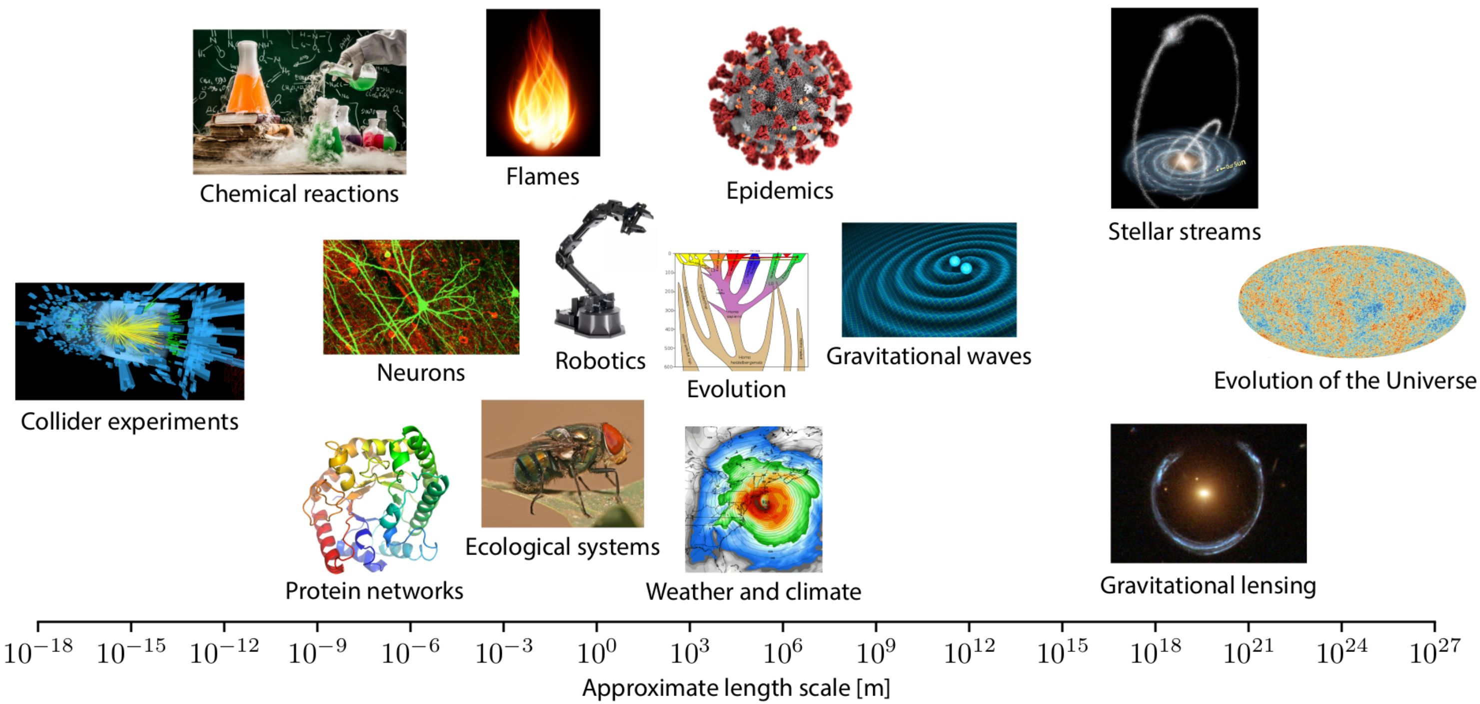
The frontier of simulation-based inference

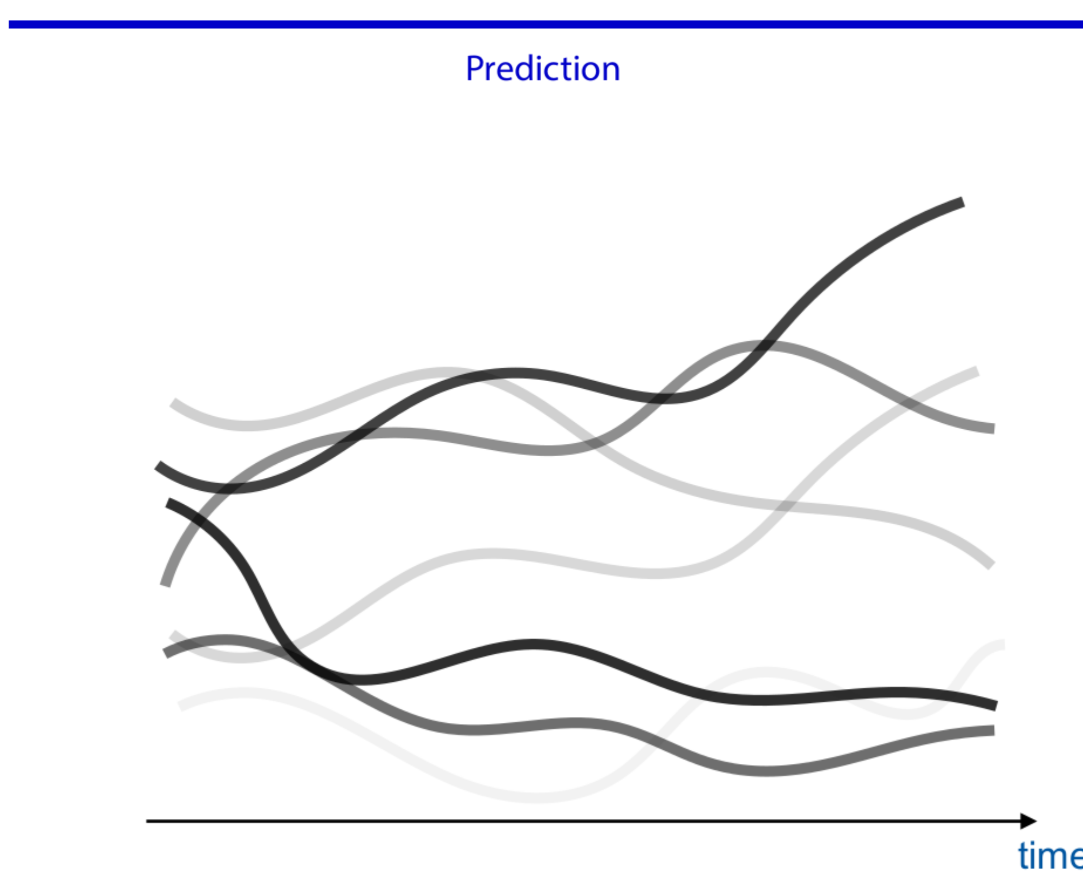
DataLearning seminars, Imperial College London
February 8, 2022

Gilles Louppe
g.louppe@uliege.be

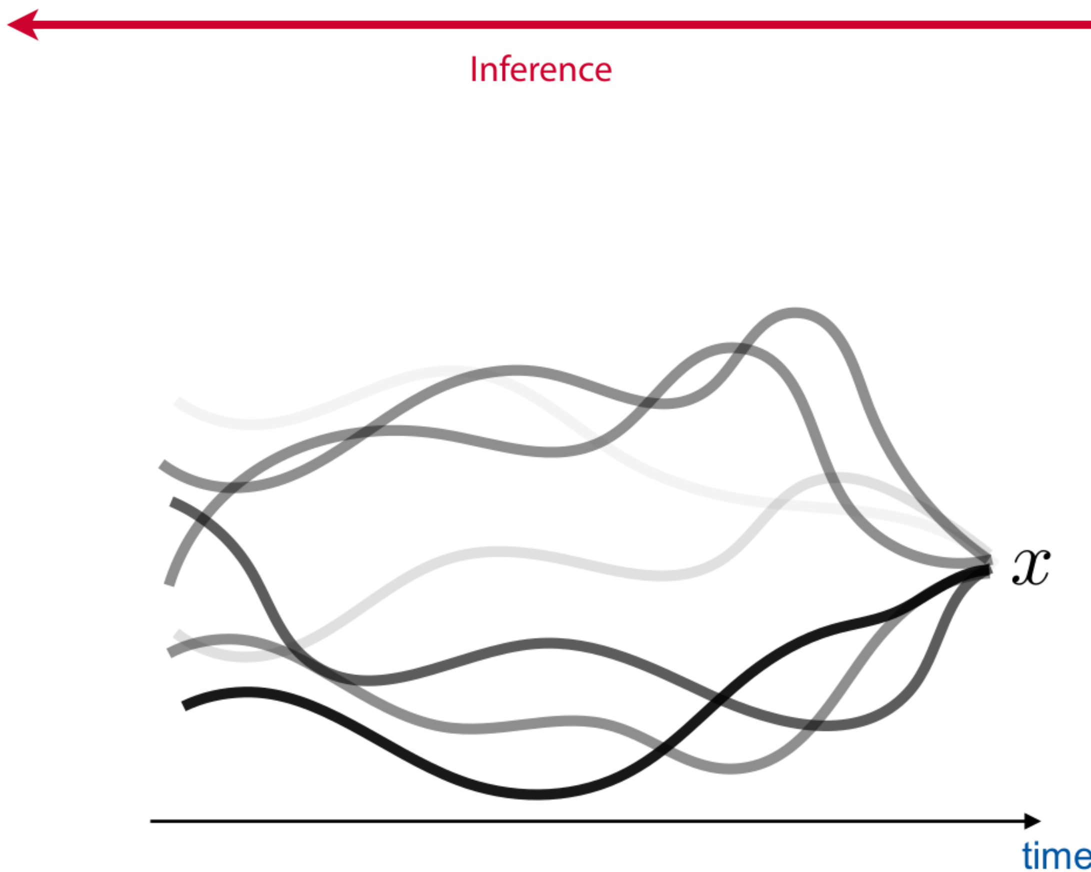
LIÈGE université

Scientific simulators





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



This results in the likelihood $p(x|\theta) = \int p(x, z|\theta) dz$ to be intractable.

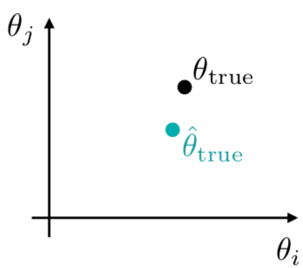
Problem statement

Start with

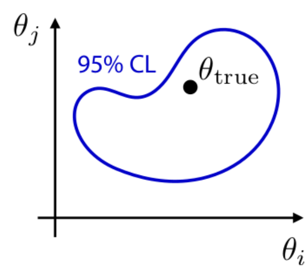
- a simulator that lets you generate N samples $x_i \sim p(x_i|\theta_i)$,
- observed data $x_{\text{obs}} \sim p(x_{\text{obs}}|\theta_{\text{true}})$,
- a prior $p(\theta)$.

Then,

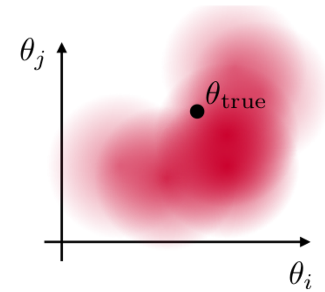
a) estimate θ_{true} (e.g., MLE)



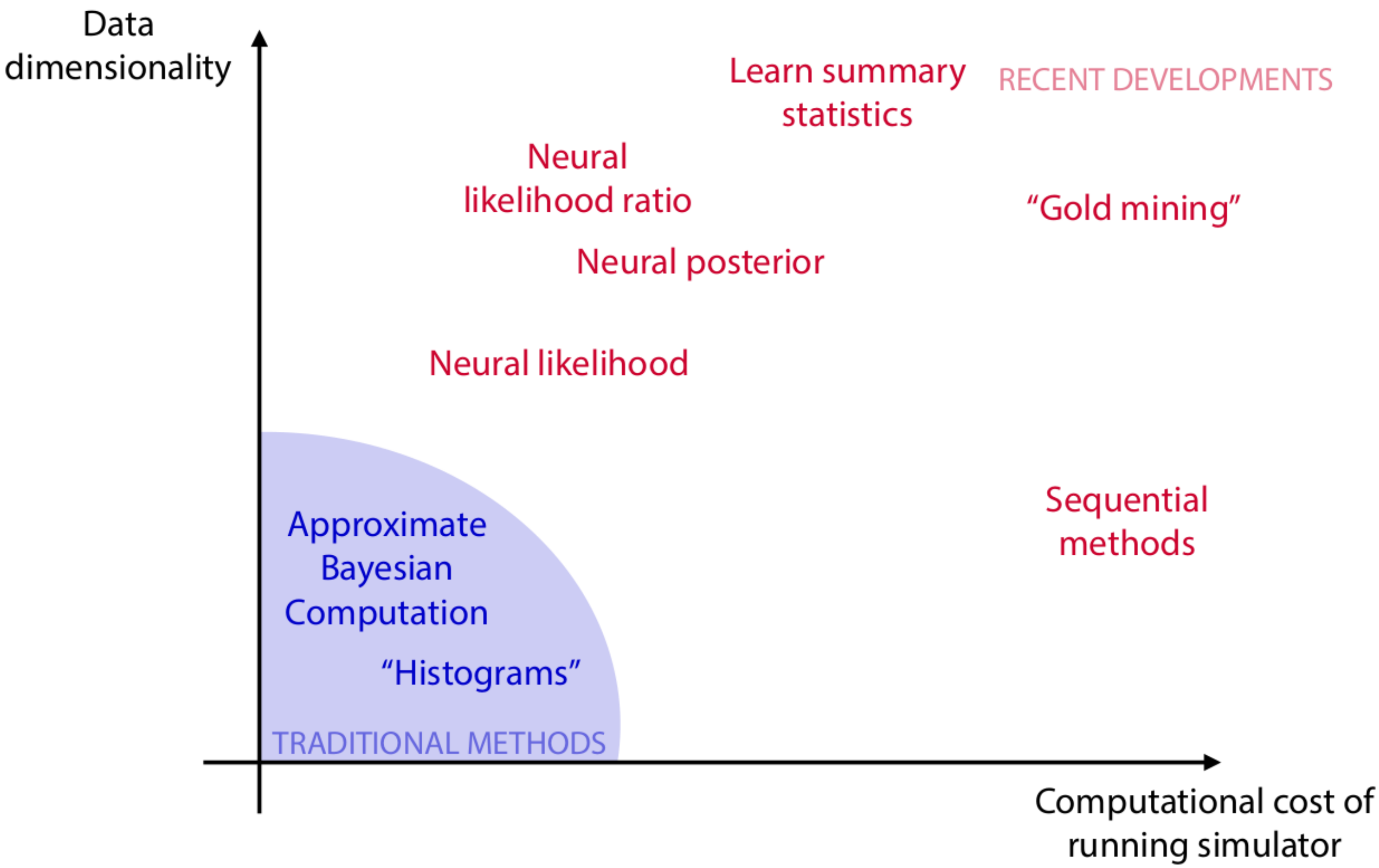
b) construct confidence sets



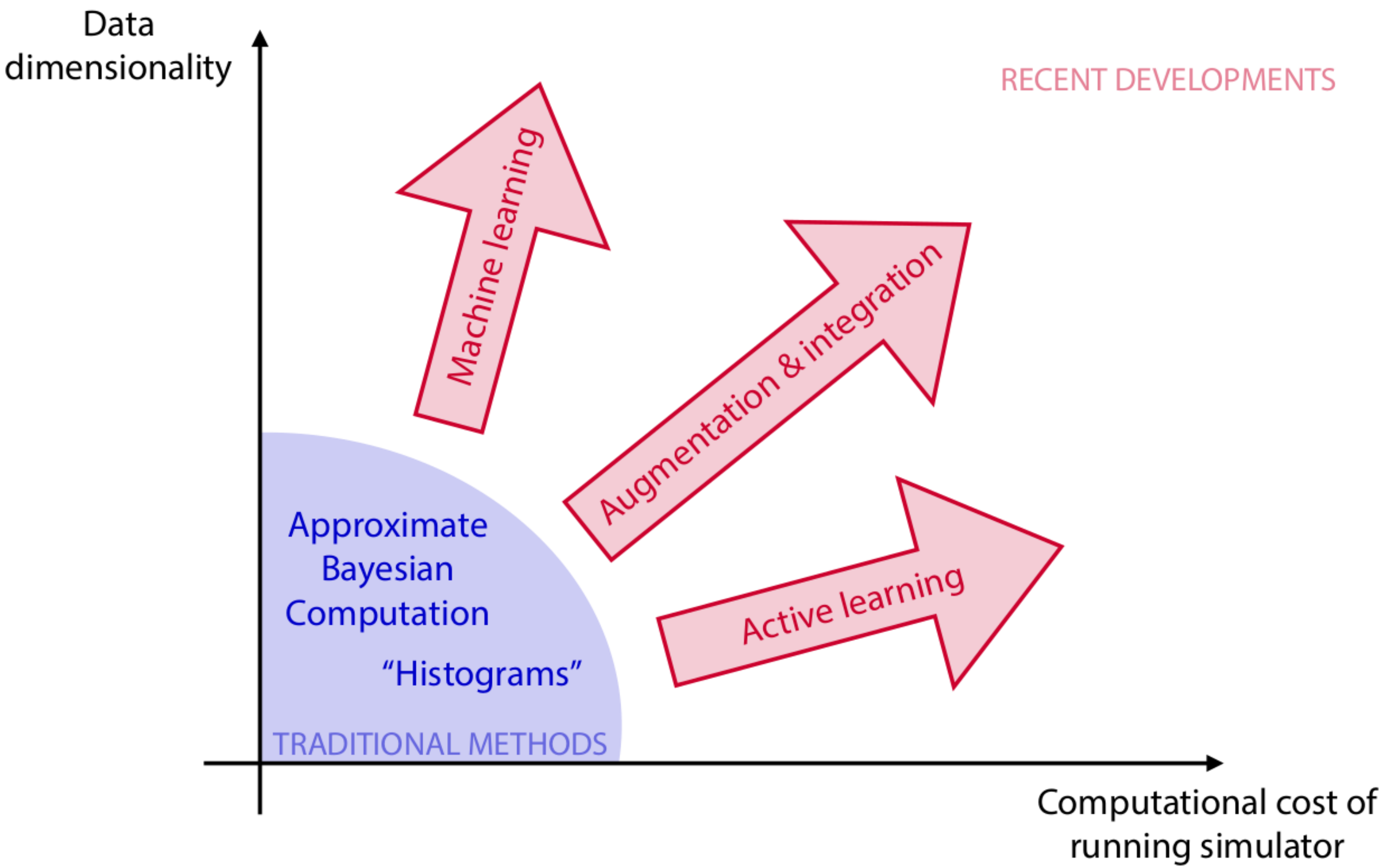
c) estimate the posterior $p(\theta|x_{\text{obs}})$



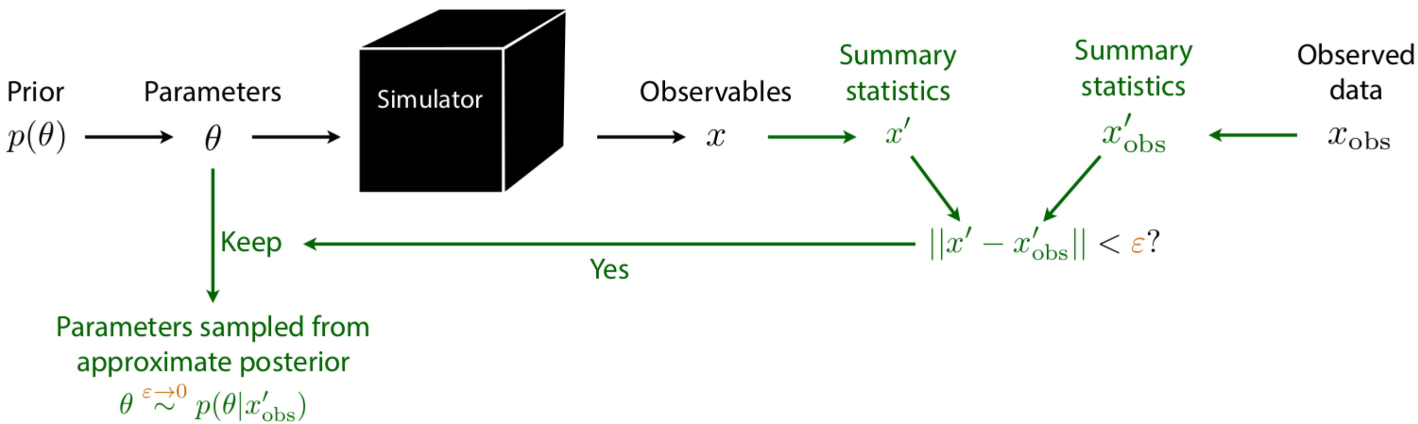
Inference algorithms



Inference algorithms



Approximate Bayesian Computation (ABC)



Issues

- How to choose x'? ϵ? $|| \cdot ||$?
- No tractable posterior.
- Need to run new simulations for new data or new prior.

Neural Ratio Estimation (NRE)

The Bayes rule can be rewritten as

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

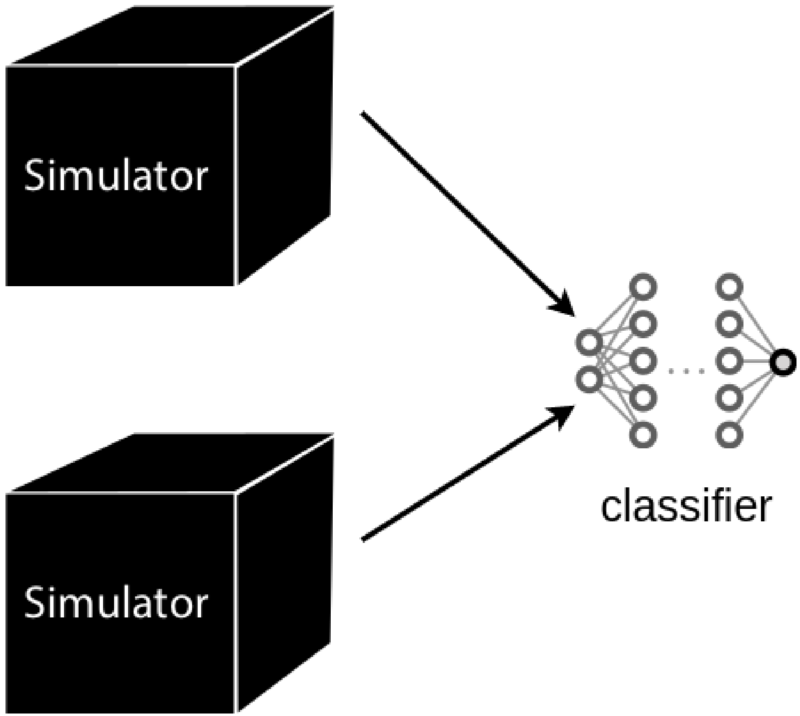
where $r(x|\theta) = \frac{p(x|\theta)}{p(x)}$ is the likelihood-to-evidence ratio.

The ratio can be learned with machine learning, even neither the likelihood nor the evidence can be evaluated!

The likelihood ratio trick

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

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



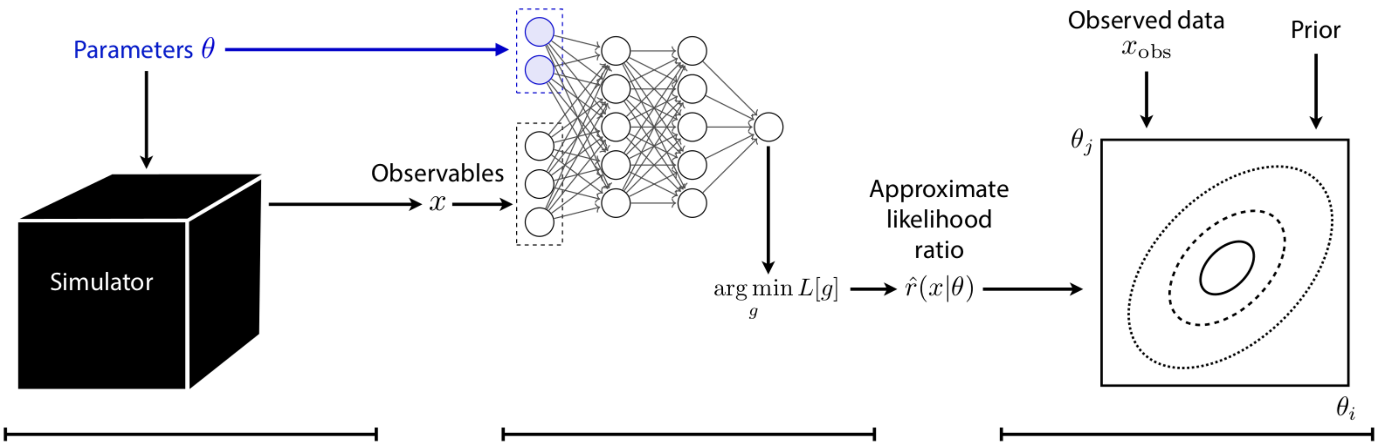
The solution d found after training approximates the optimal classifier

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

Therefore,

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

Inference



1. Simulation

Run simulator and save data

2. Machine Learning

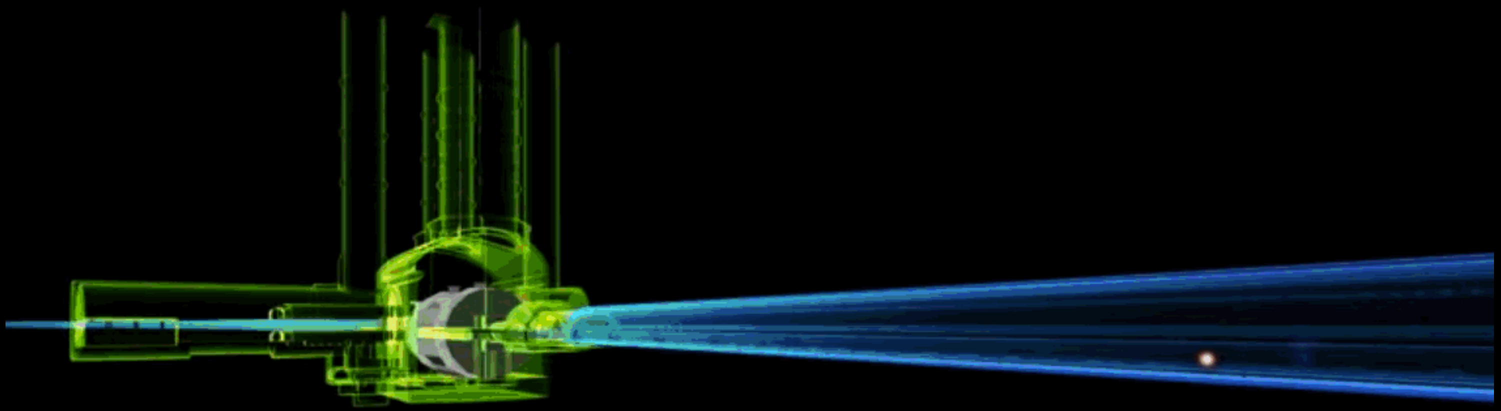
Train NN classifier, interpret as likelihood ratio estimator

3. Inference

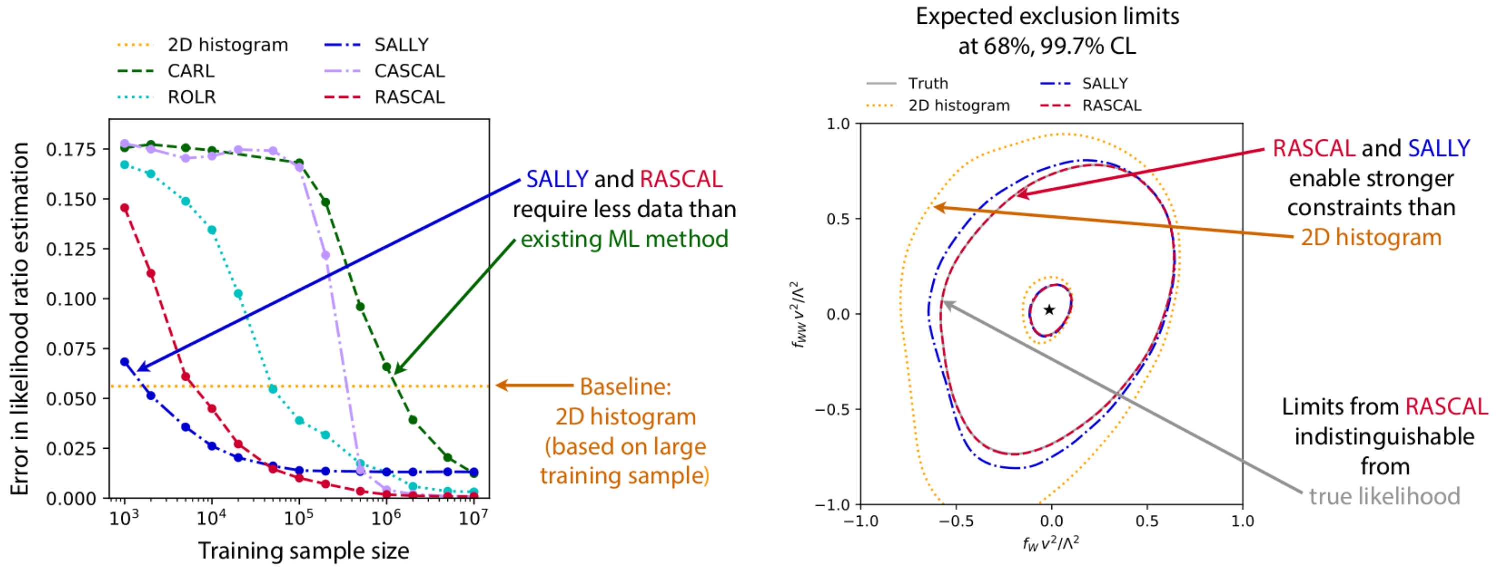
Amortized: cheap to repeat for new data

Showtime!

Some applications of physics and astrophysics.



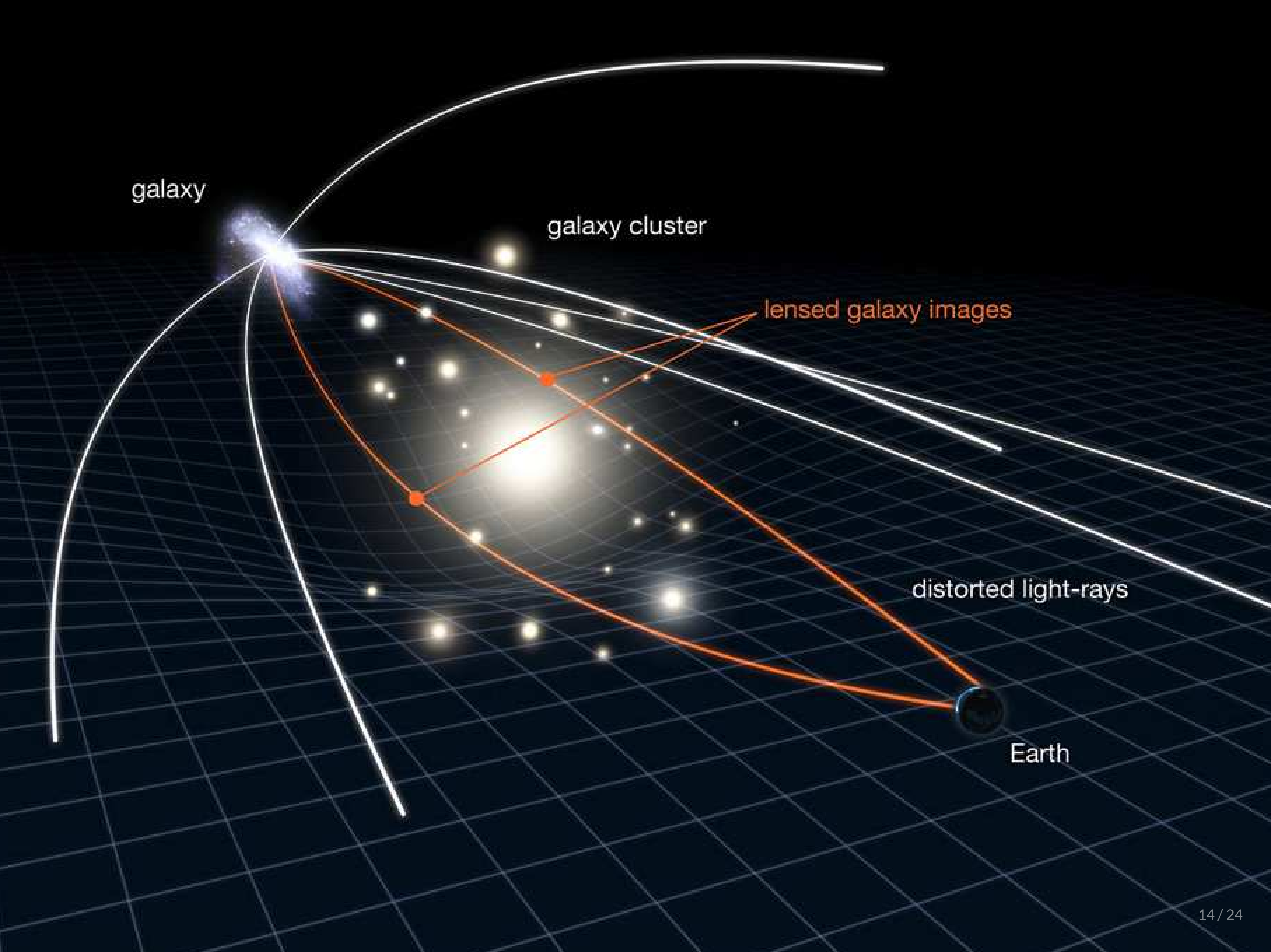
Case 1: Hunting new physics at particle colliders

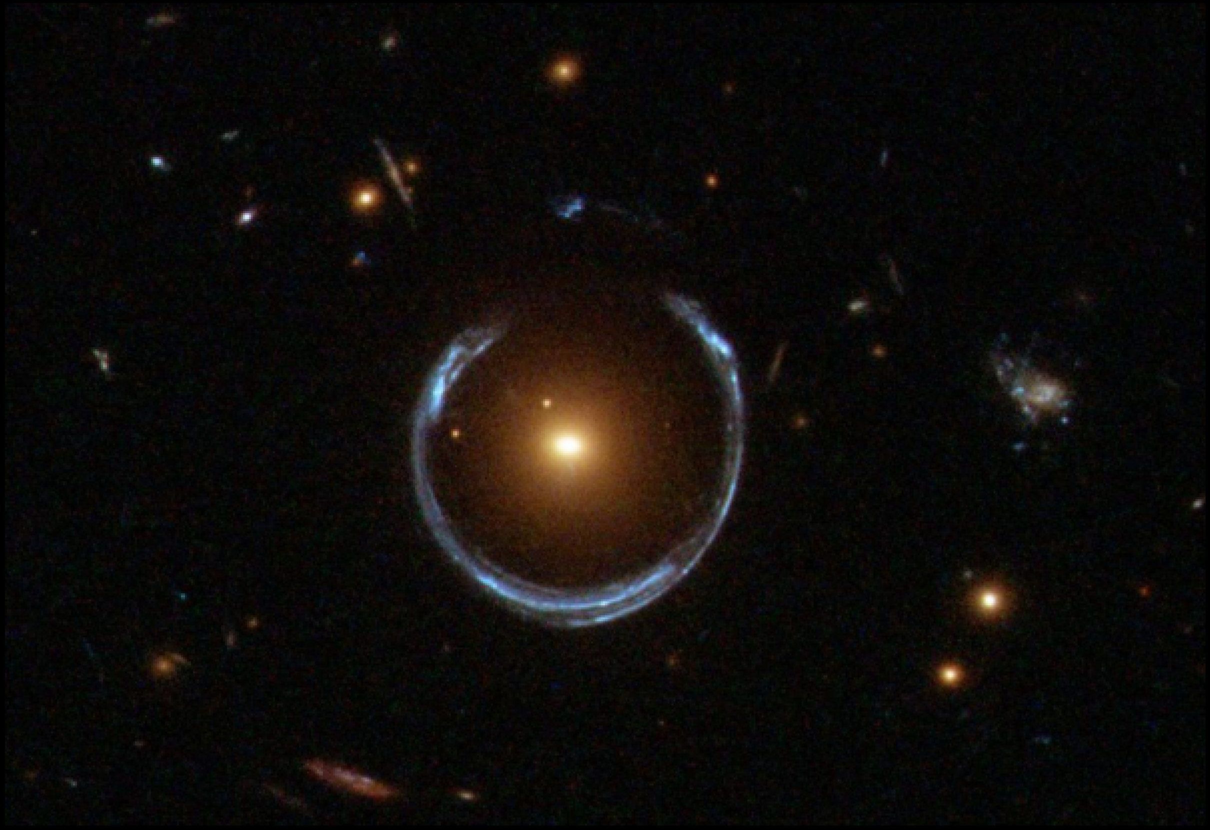


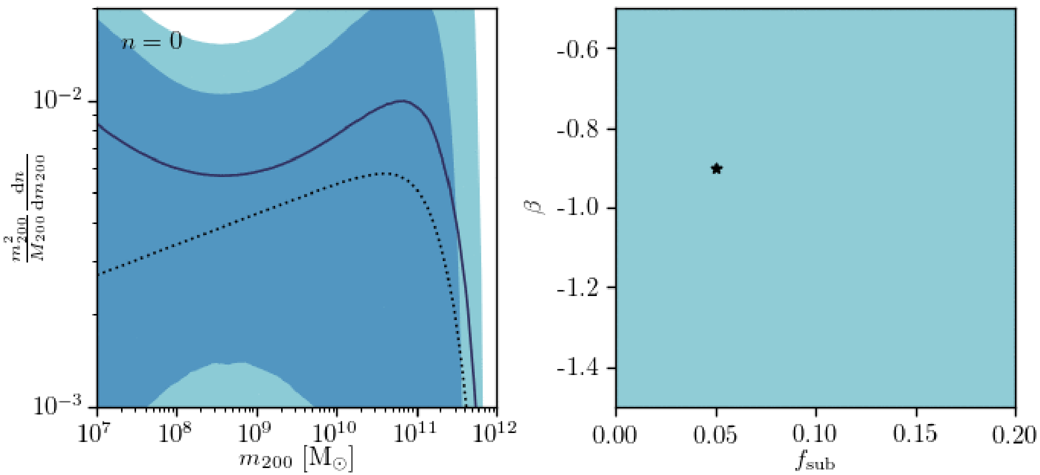
With enough training data, NRE gets the likelihood-ratio statistic right.

Using more information from the simulator improves sample efficiency substantially.

Case 2: Dark matter substructure from gravitational lensing

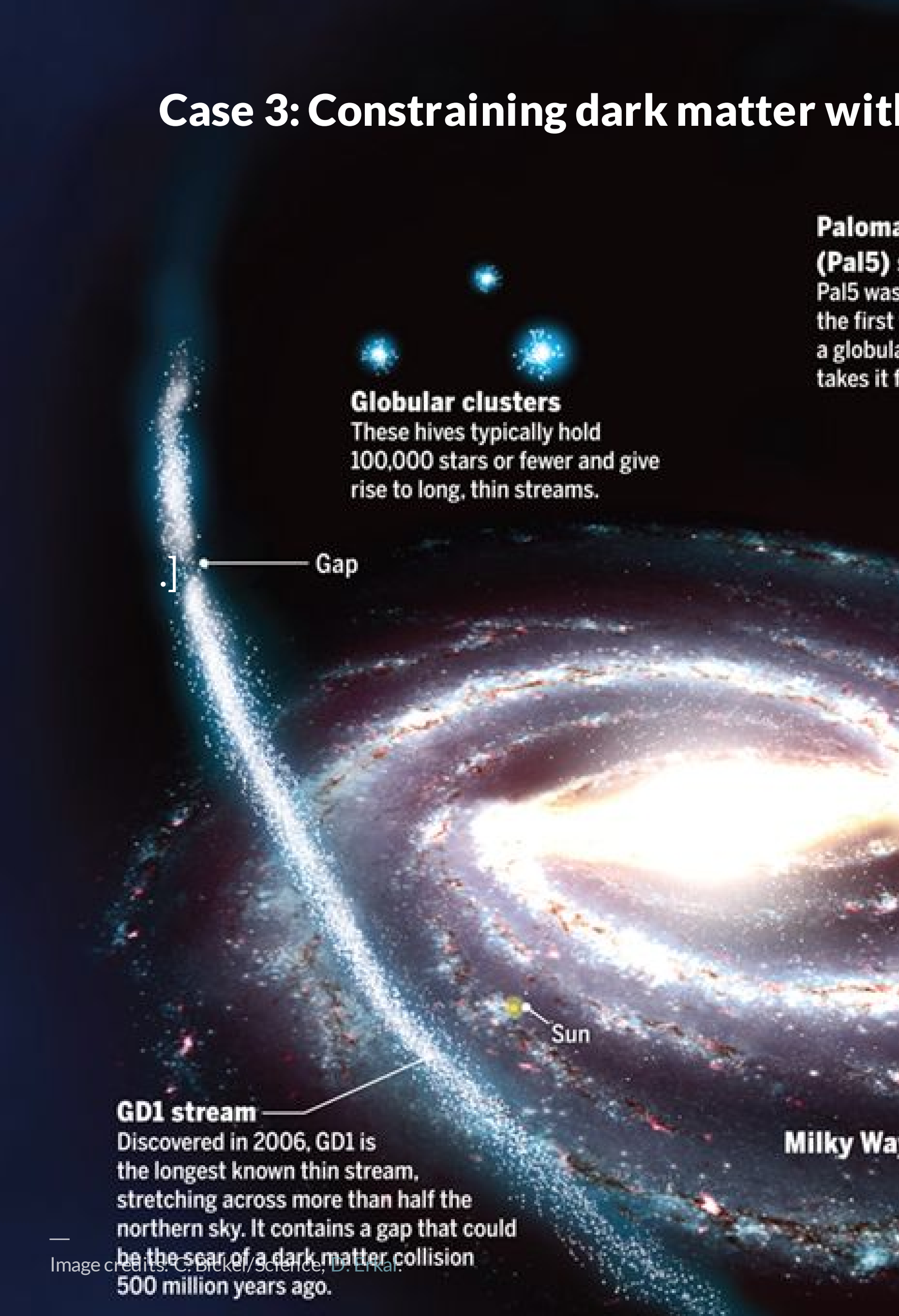

galaxy
galaxy cluster
lensed galaxy images
distorted light-rays
Earth

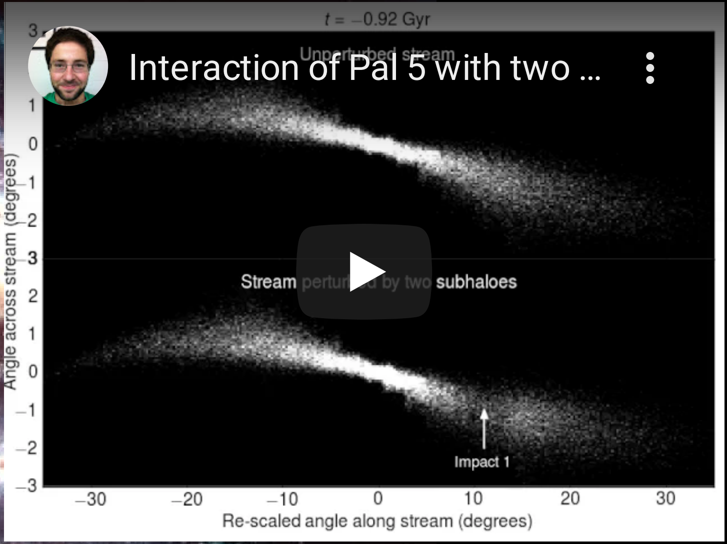


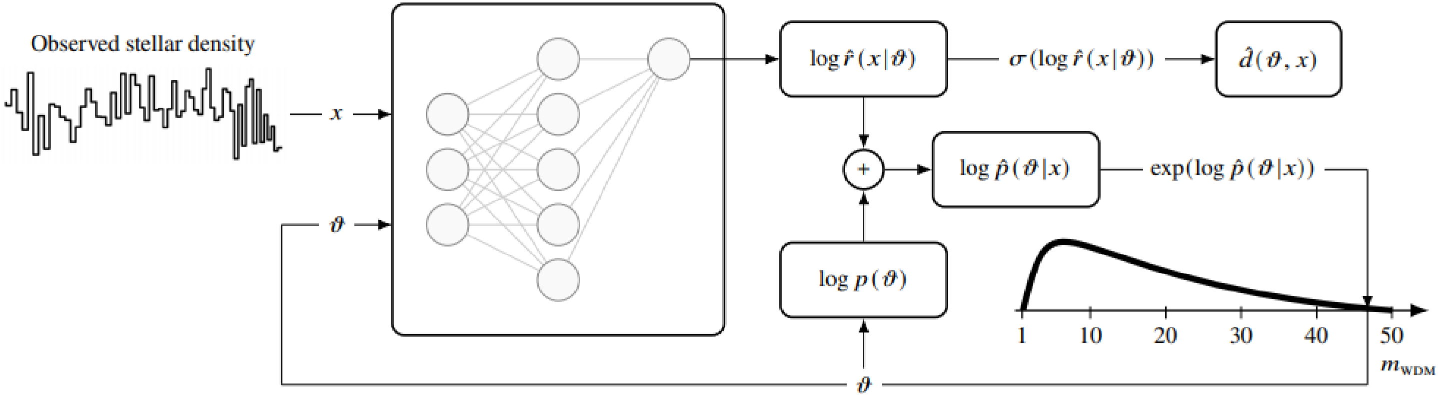


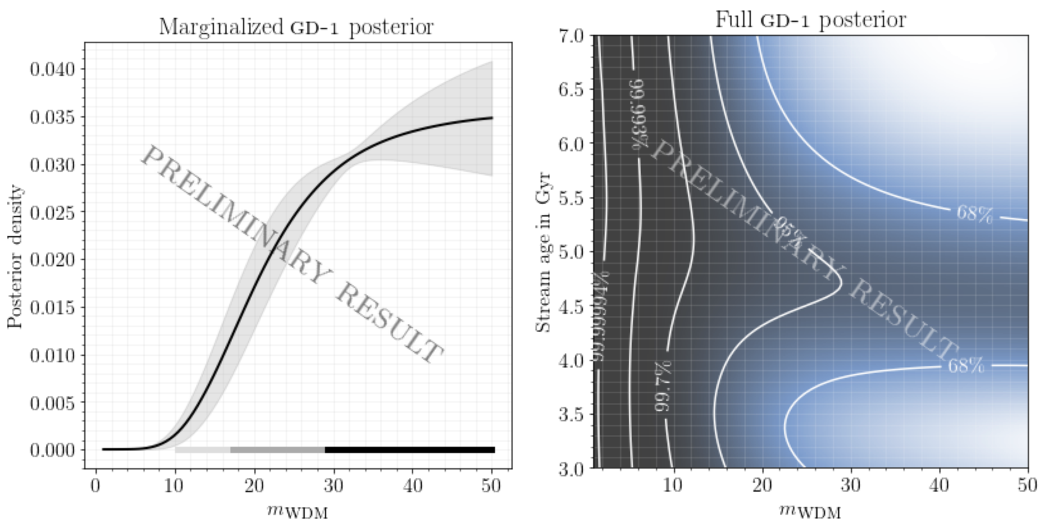
$n = 0$
10^{-2}
10^{-3}
$\frac{m_{200}^2}{M_{200}} \frac{\mathrm{d}n}{\mathrm{d}m_{200}}$
10^7 10^8 10^9 10^{10} 10^{11} 10^{12}
m_{200} [$\mathrm{M}_\odot$]
-0.6
-0.8
-1.0
-1.2
-1.4
β
0.00 0.05 0.10 0.15 0.20
f_{sub}

Case 3: Constraining dark matter with stellar streams









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

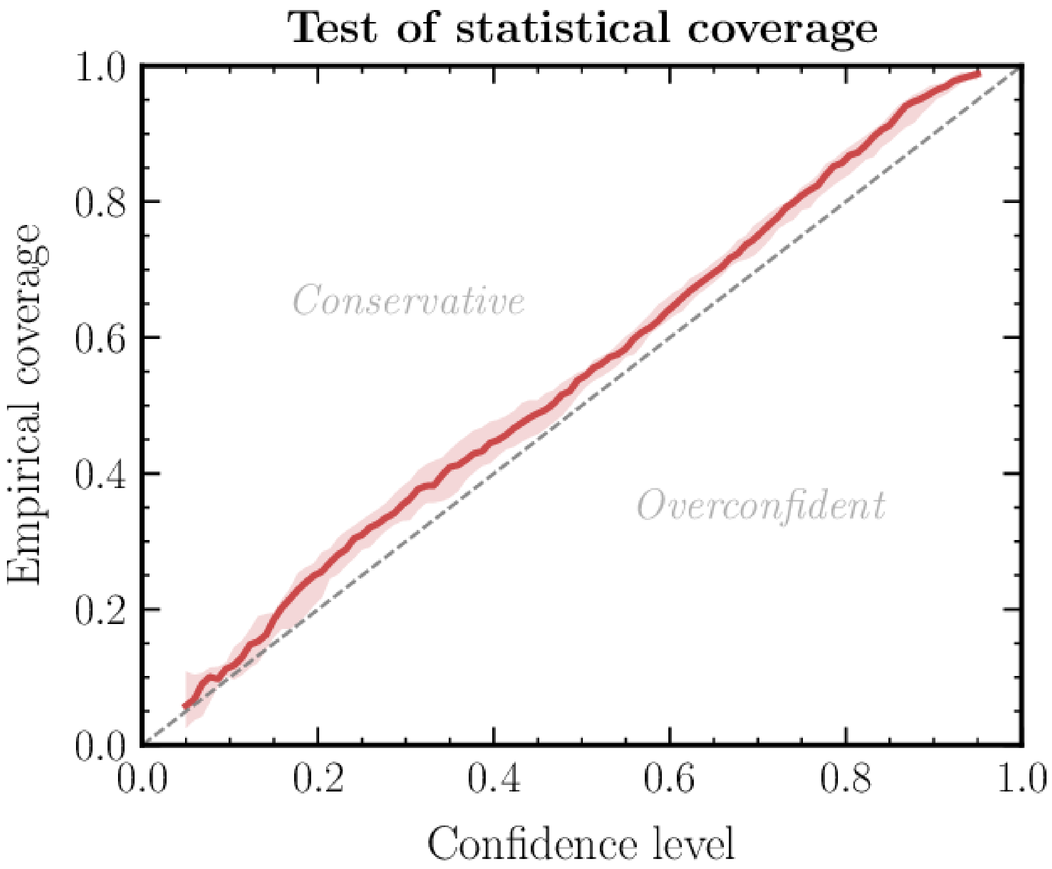
Diagnosing inference

How to assess that approximate posteriors are not too wrong?

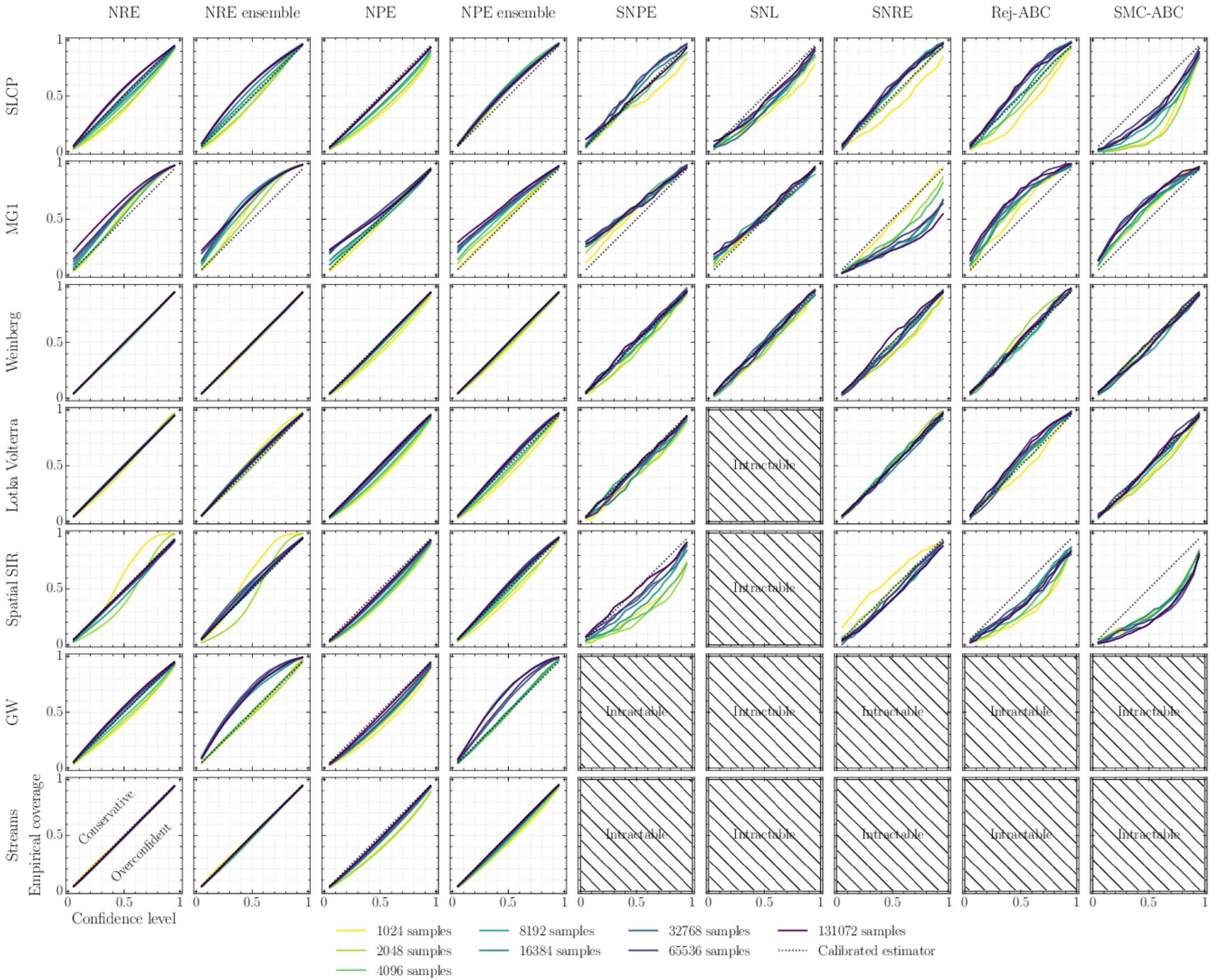
Coverage

- For every $x, \theta \sim p(x, \theta)$ in a validation set, compute the $1 - \alpha$ credible interval based on $\hat{p}(\theta|x) = \hat{r}(x|\theta)p(\theta)$.
- The fraction of samples for which θ is contained within the interval corresponds to the empirical coverage probability.

If the empirical coverage is larger that the nominal coverage probability $1 - \alpha$, then the ratio estimator $\hat{r}$ passes the diagnostic.



—
Image credits: Siddharth Mishra-Sharma, 2021.



—

Image credits: Hermans et al, 2021.

All benchmarked algorithms can produce non-conservative posterior approximations.

~~The frontier of simulation-based inference~~

Averting a crisis in simulation-based inference?

DataLearning seminars, Imperial College London
February 8, 2022

Gilles Louppe
g.louppe@uliege.be

Summary

- Much of modern science is based on simulators making precise predictions, but in which inference is challenging.
- Machine learning enables powerful inference methods, which work in problems from the smallest to the largest scales.
- Advances in simulation-based inference will translate into scientific progress.
- However, further work is needed to make these methods more robust and reliable.

The end.