Simulation-based inference: Proceed with caution!

May 23, 2022

Gilles Louppe
g.louppe@uliege.be

LIÈGE université

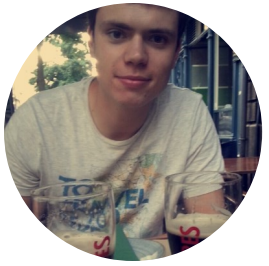
Kyle Cranmer

Johann Brehmer

Michael Kagan

Joeri Hermans

Antoine Wehenkel



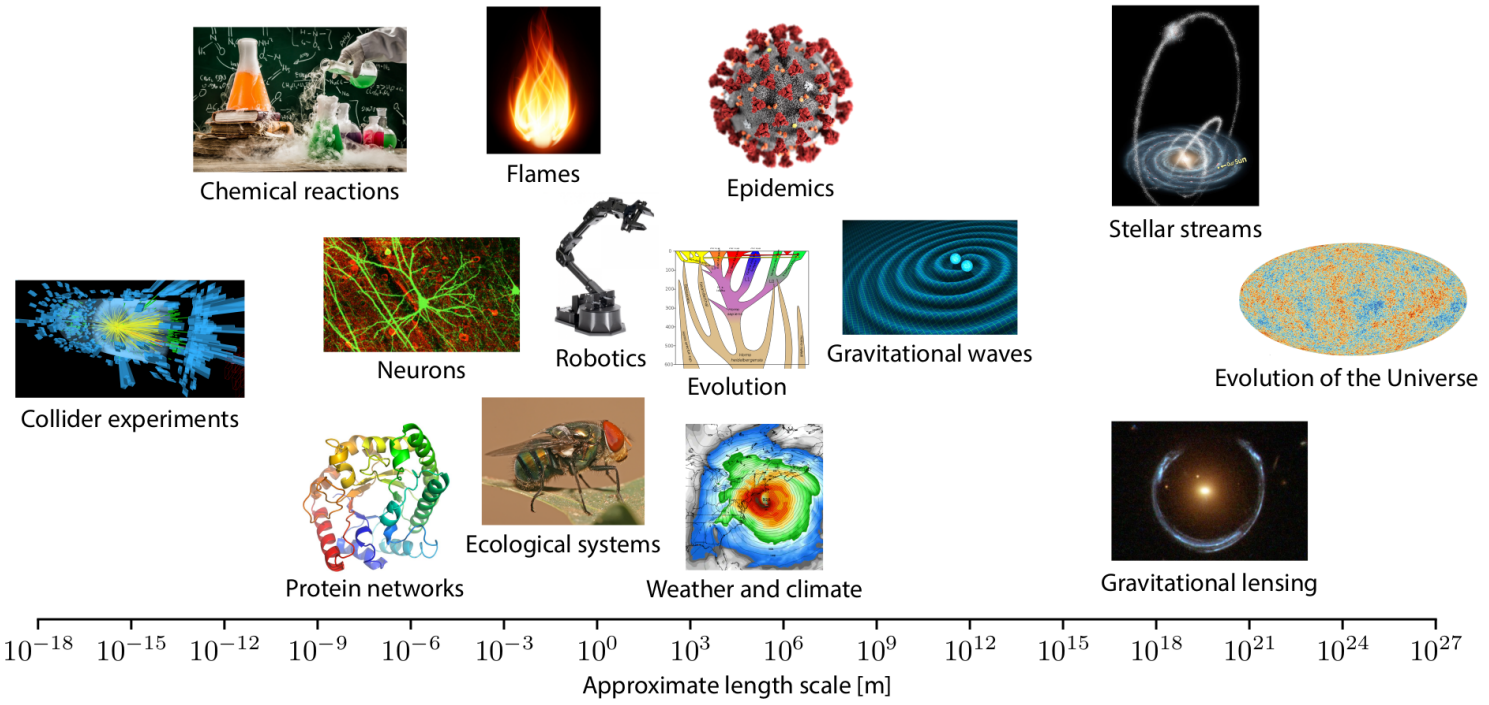
Norman Marlier

Arnaud Delaunoy

Maxime Vandegar

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]

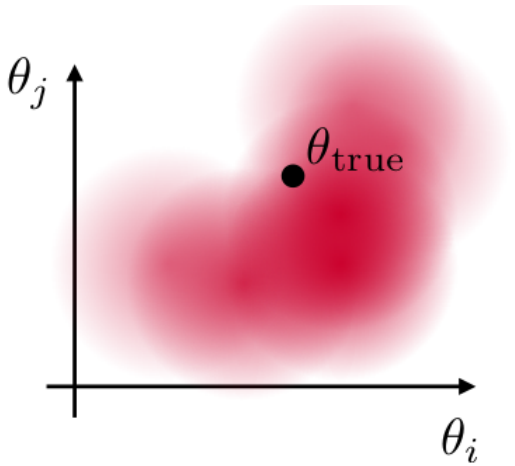
Simulation-based inference

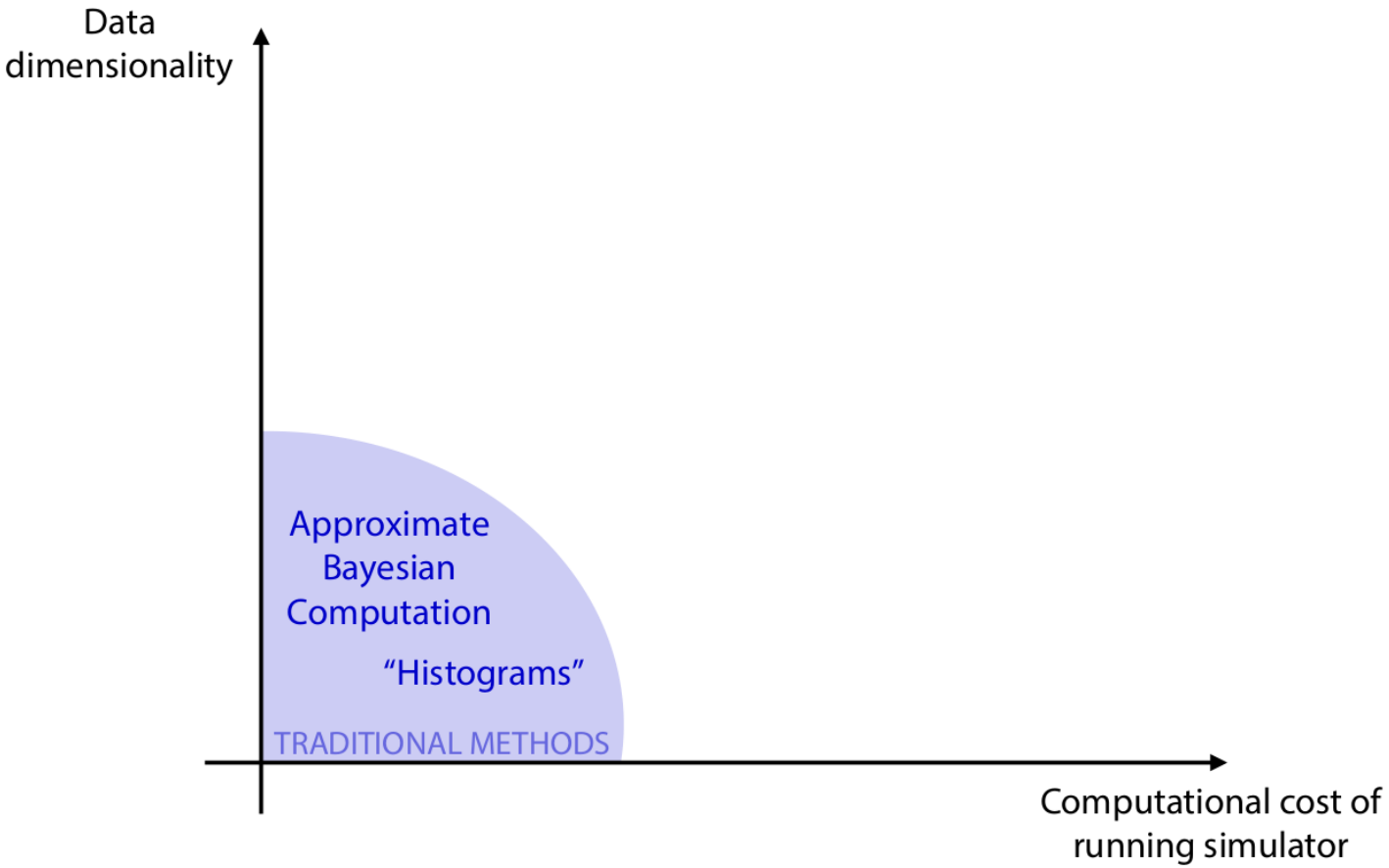
Start with

- 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

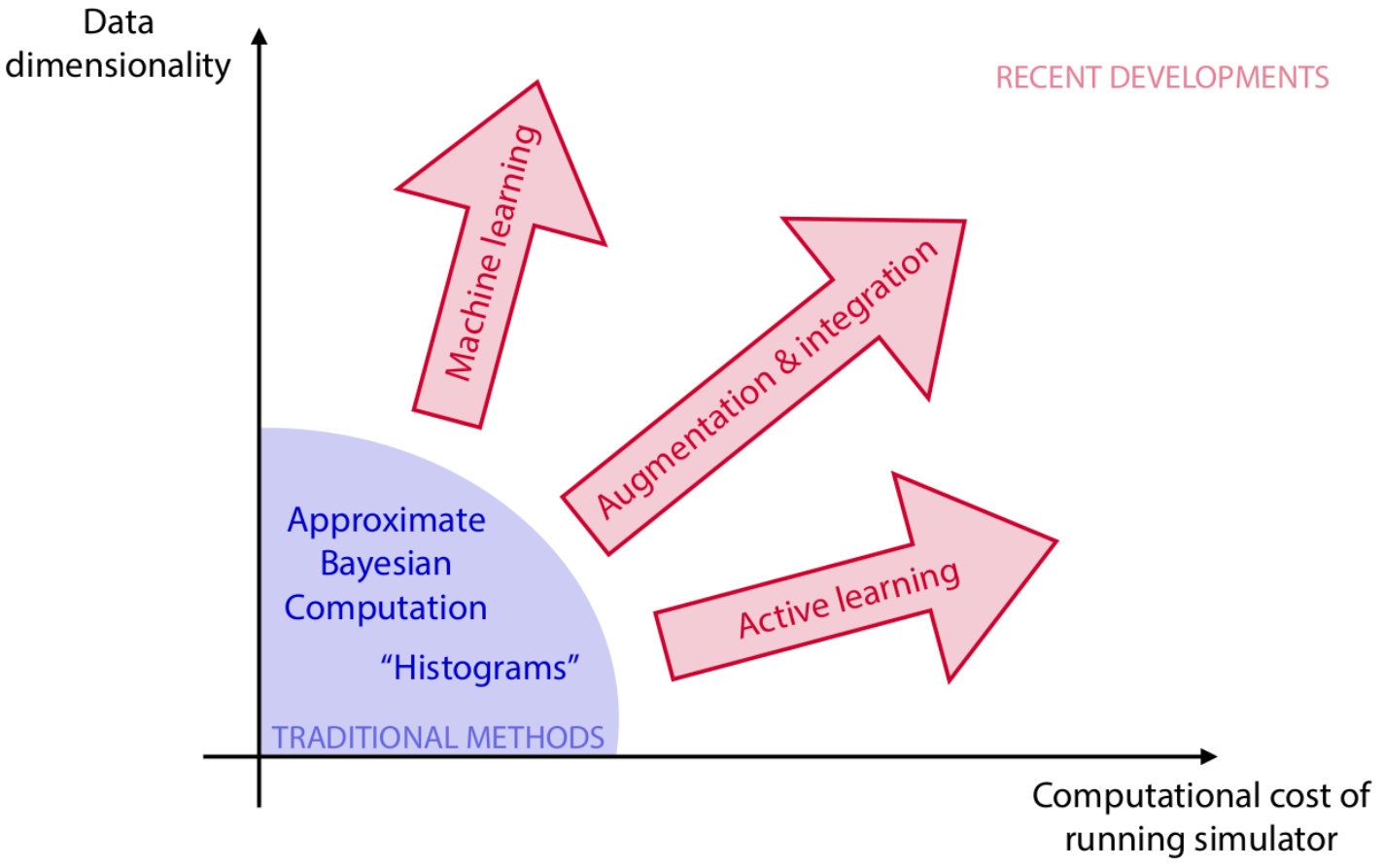
$$p(\theta|x_{\text{obs}}) = \frac{p(x_{\text{obs}}|\theta)p(\theta)}{p(x_{\text{obs}})}$$





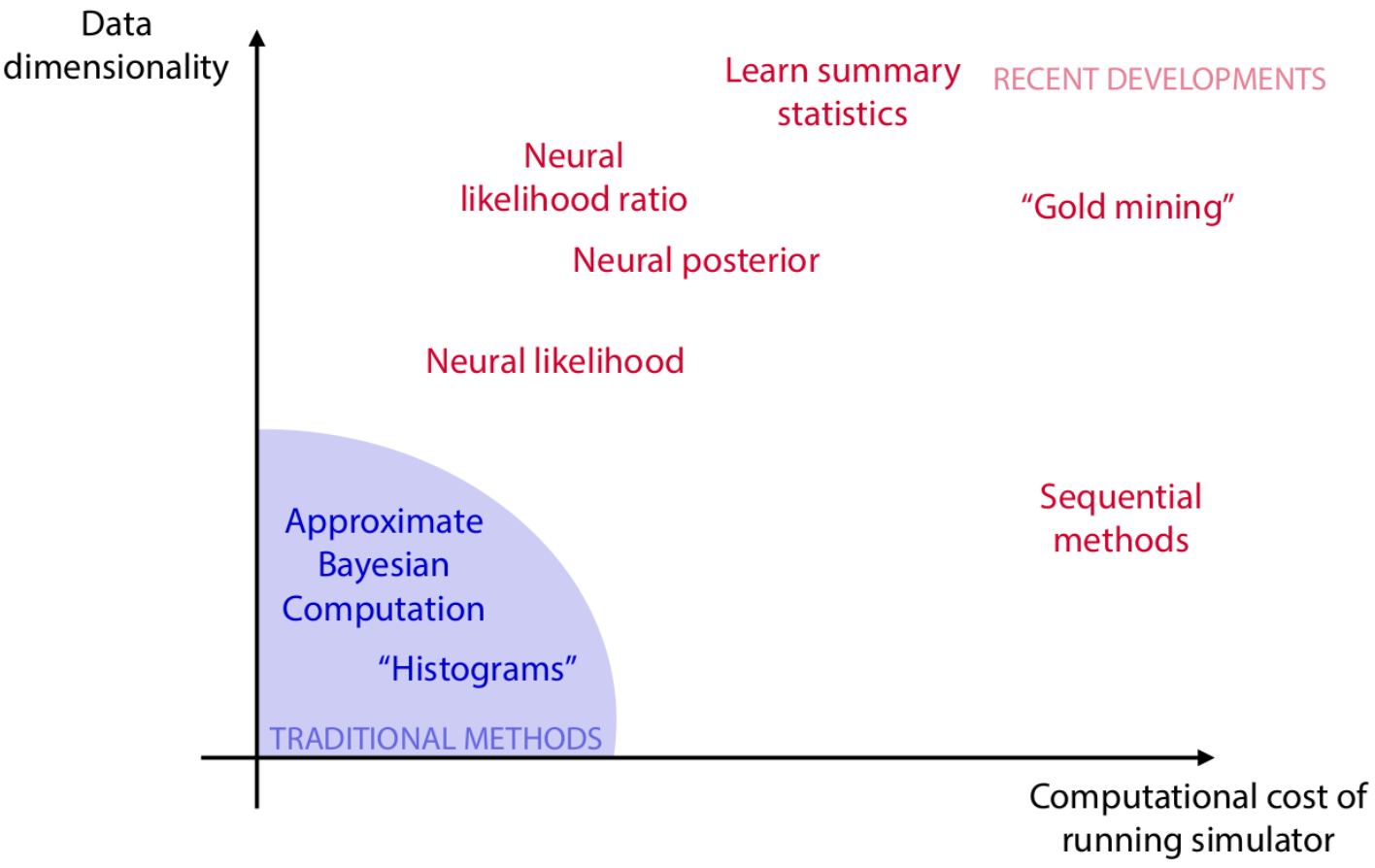
—

Credits: Cranmer, Brehmer and Louppe, 2020.



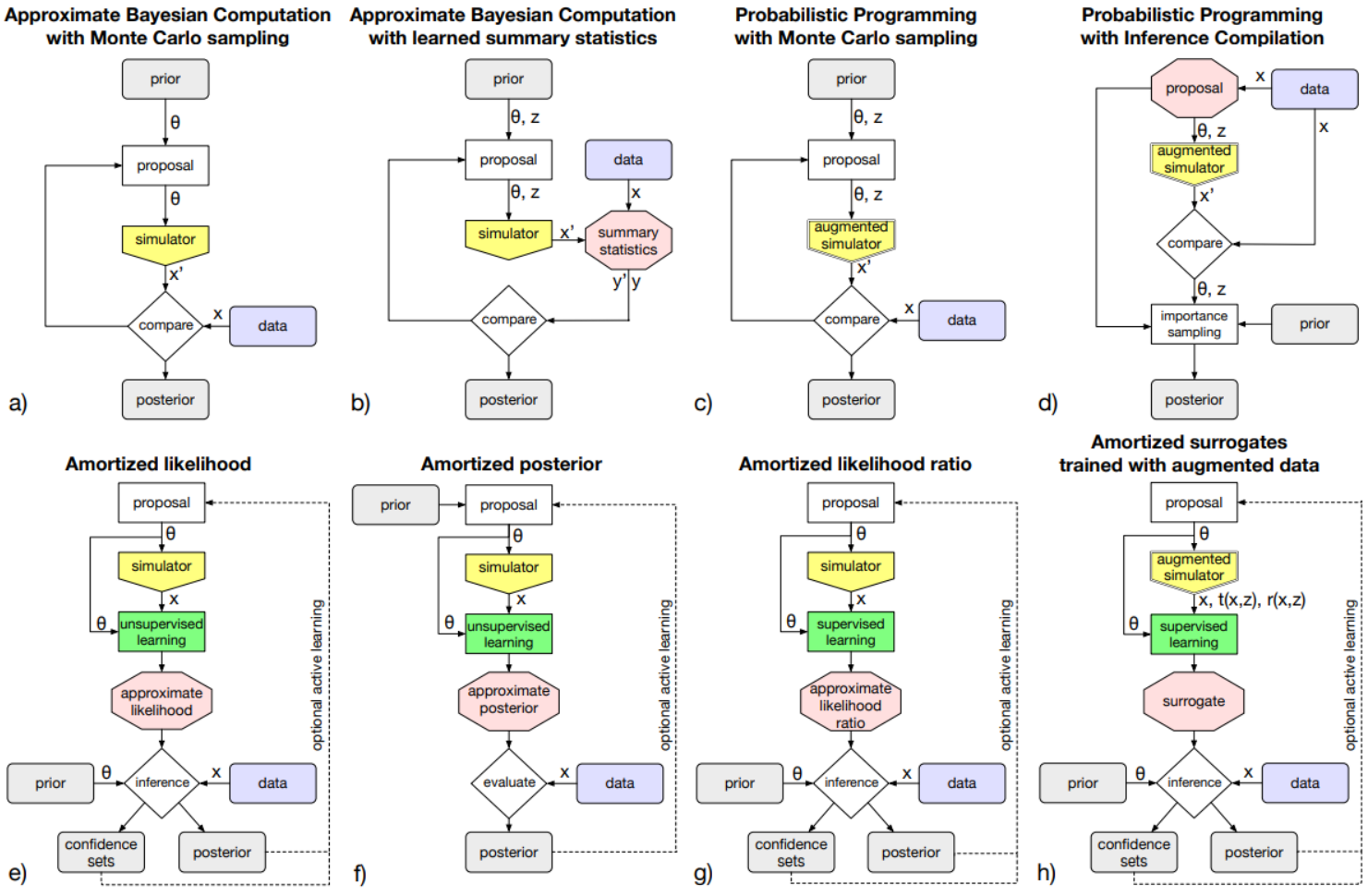
—

Credits: Cranmer, Brehmer and Louppe, 2020.

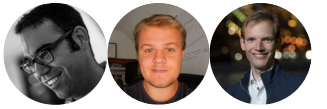


—

Credits: Cranmer, Brehmer and Louppe, 2020.



Credits: Cranmer, Brehmer and Louppe, 2020.

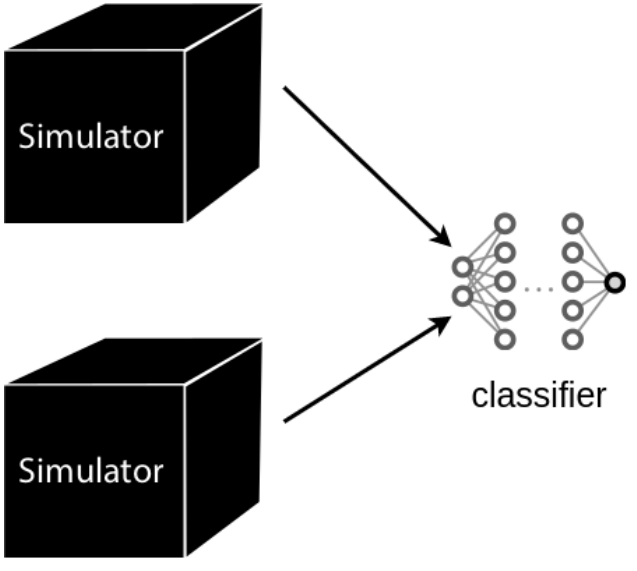


Neural ratio estimation (NRE)

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:

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

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



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

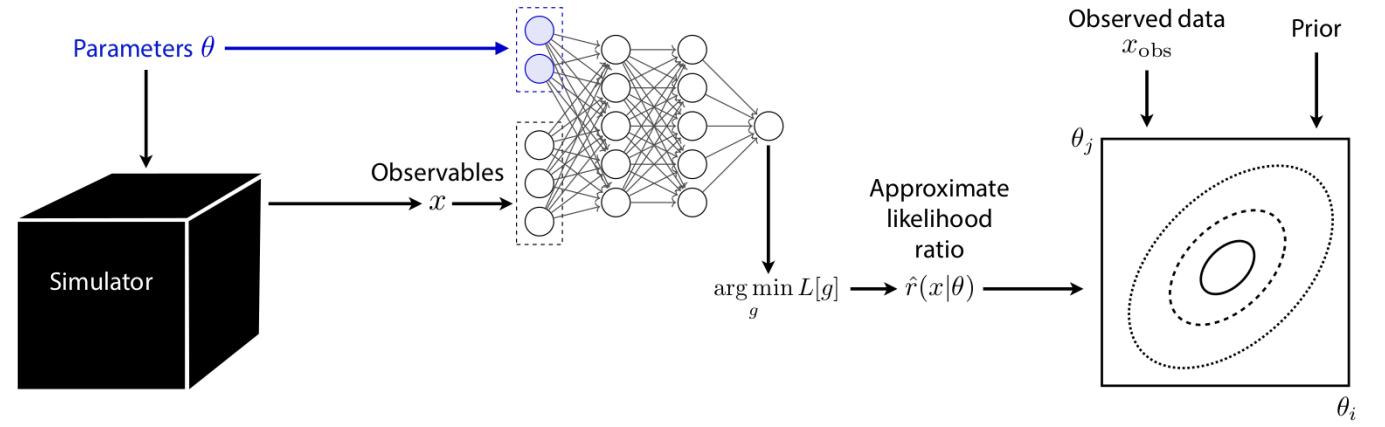
The solution d found after training approximates the optimal classifier

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

Therefore,

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

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

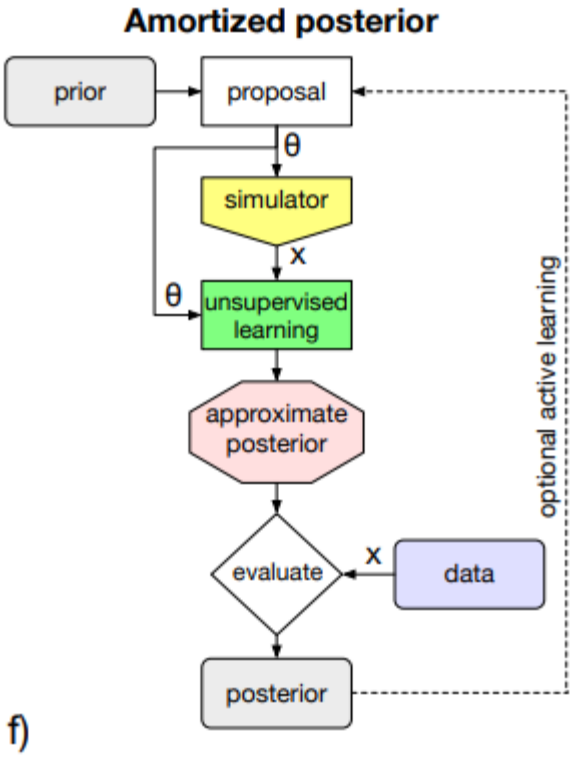


Neural Posterior Estimation (NPE)

The posterior density can approximated directly using variational inference, solving

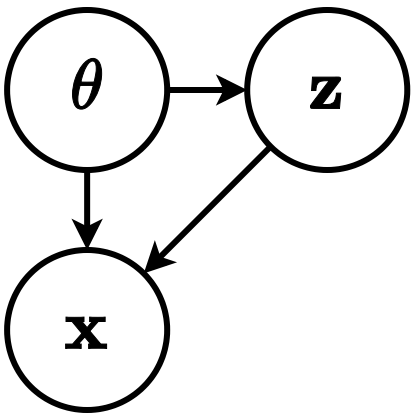
$$\begin{aligned}
&\min_{q_\phi} \mathbb{E}_{p(x)} \left[\mathrm{KL}(p(\theta|x) || q_\phi(\theta|x)) \right] \\
&= \min_{q_\phi} \mathbb{E}_{p(x)} \mathbb{E}_{p(\theta|x)} \left[\log \frac{p(\theta|x)}{q_\phi(\theta|x)} \right] \\
&= \max_{q_\phi} \mathbb{E}_{p(x,\theta)} \left[\log q_\phi(\theta|x) \right]
\end{aligned}$$

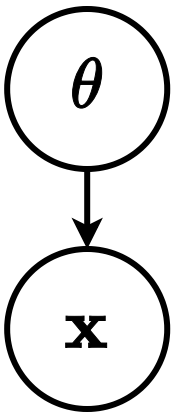
where q_ϕ is a neural density estimator, such as a normalizing flow.



... but proceed with caution!

aka model checking, evaluation, and criticism.


θ
z
x



Prior model $p(\theta)$

Observational model $p(x|\theta)$

"All models are wrong, but some are useful" - George Box

The observational model $p(x|\theta)$

$p(x|\theta)$ should capture the pertinent structure of the true data generating process for the inference results to be useful.

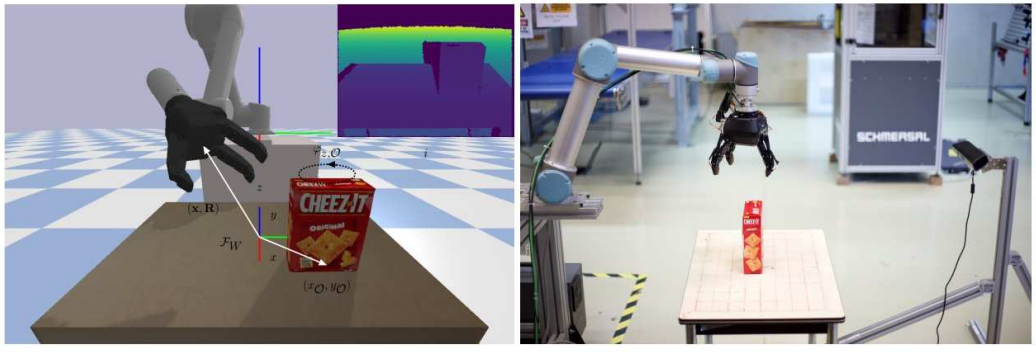
A model that does not capture every precise detail of the true data generating process can still be useful if it captures the details relevant to the particular analysis goals.

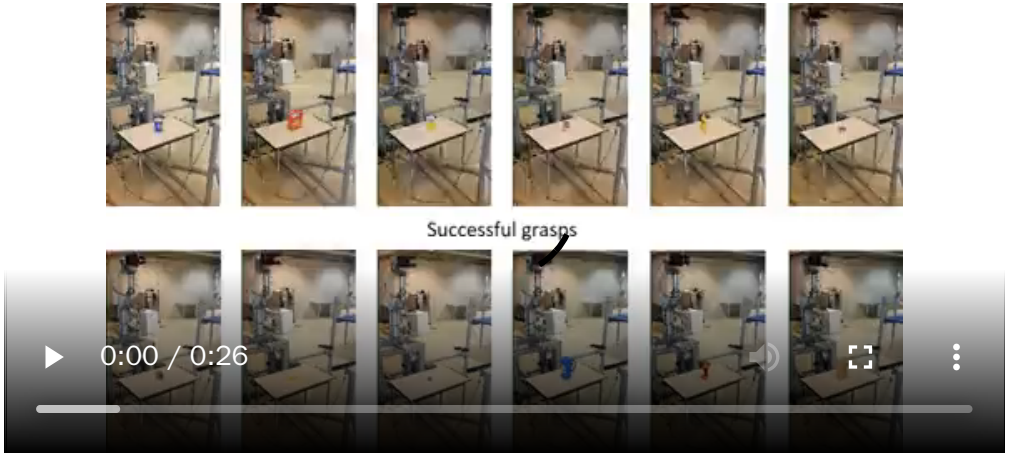
The observational model can often be made richer by including in it additional **nuisance parameters** ν that capture known unknowns.

In this case, the likelihood becomes

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

Although nuisance parameters can reduce model misspecification, their presence and marginalization will result in increased uncertainties for the parameters θ of interest.





Nuisance parameters are used to model known unknowns in a robotic setup (e.g., camera position, table position, etc).

—
Credits: Marlier et al, 2021.

The prior model $p(\theta)$

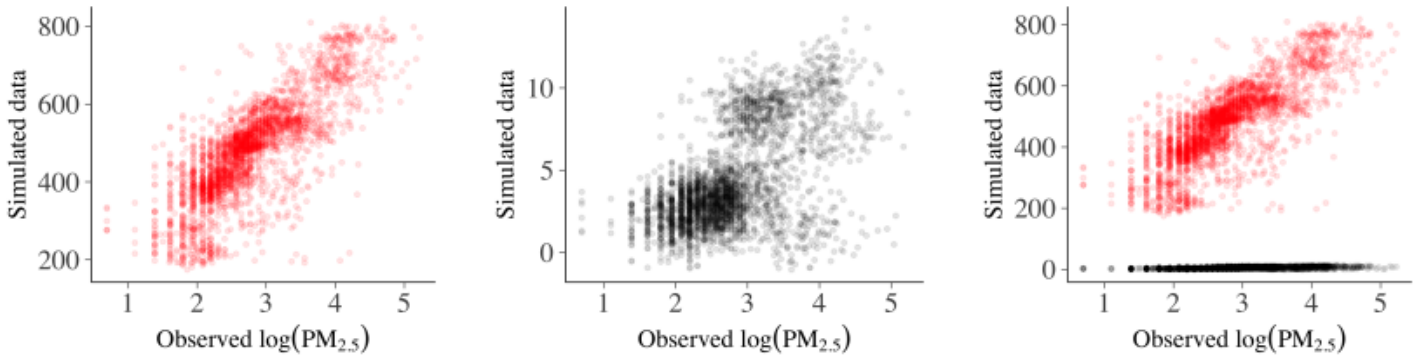
The prior model $p(\theta)$ specifies one's beliefs about the model parameters. It should reflect domain expertise.

The consequences of the prior model in the context of the observational model can be diagnosed with **prior predictive checks** to evaluate what data sets would be consistent with the prior.

A prior predictive check generates data x^{sim} according to the prior predictive distribution $p(x)$ as

$$\begin{aligned} \theta^{\text{sim}} &\sim p(\theta) \\ x^{\text{sim}} &\sim p(x|\theta^{\text{sim}}), \end{aligned}$$

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



(a) Vague priors (b) Weakly informative priors (c) Comparison

Fig. 4: *Visualizing the prior predictive distribution. Panels (a) and (b) show realizations from the prior predictive distribution using priors for the β's and τ's that are vague and weakly informative, respectively. The same $N_+(0,1)$ prior is used for σ in both cases. Simulated data are plotted on the y-axis and observed data on the x-axis. Because the simulations under the vague and weakly informative priors are so different, the y-axis scales used in panels (a) and (b) also differ dramatically. Panel (c) emphasizes the difference in the simulations by showing the red points from (a) and the black points from (b) plotted using the same y-axis.*

—

Credits: Gabry et al, 2017.

In the absence of a good prior, **neural empirical Bayes** can be used to estimate a prior distribution $p_\phi(\theta)$ by maximizing the (log) evidence of a set of observations

$$\log p_\phi(\{x_i\}_{i=1}^N) = \sum_{i=1}^N \log \int p(x_i|\theta) p_\phi(\theta) d\theta.$$

—

Credits: Vandegar et al, 2021.

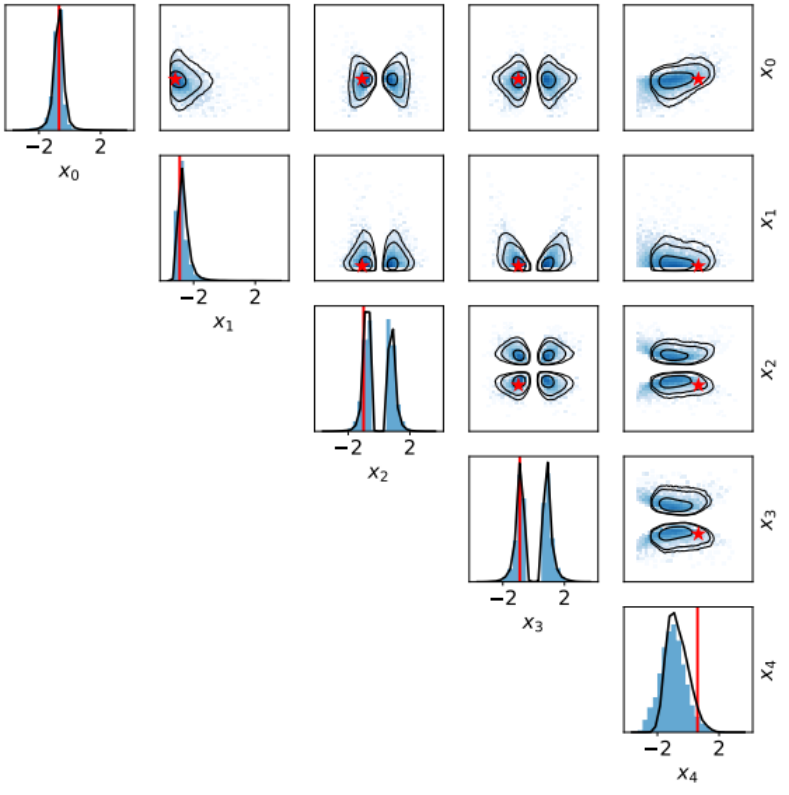


Figure 4: Posterior distribution obtained from MCMC with the exact source distribution and the exact likelihood function on SLCP in blue against the posterior distribution obtained with $q_\phi(\mathbf{y}|\mathbf{x})$ and $q_\theta(\mathbf{x})$ learned from $\mathcal{L}_{1024}$ in black (the 68-95-99.7% contours are shown). Generating source sample $\mathbf{x}$ are indicated in red. *The approximated posterior distribution closely matches the ground truth.*

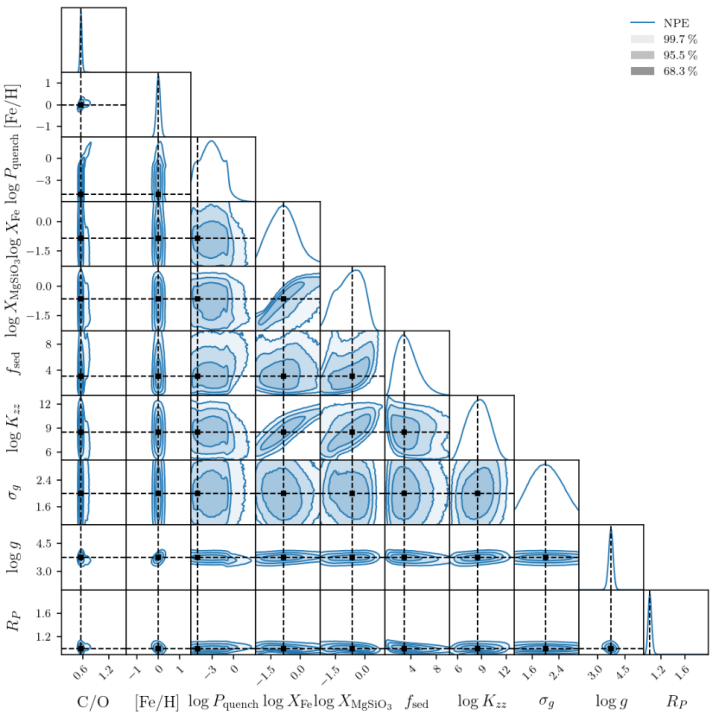
—

Credits: Vandegar et al, 2021.

Computational faithfulness

$$\hat{p}\left(\theta|x\right) = \mathrm{sbi}(p(x|\theta), p(\theta), x)$$

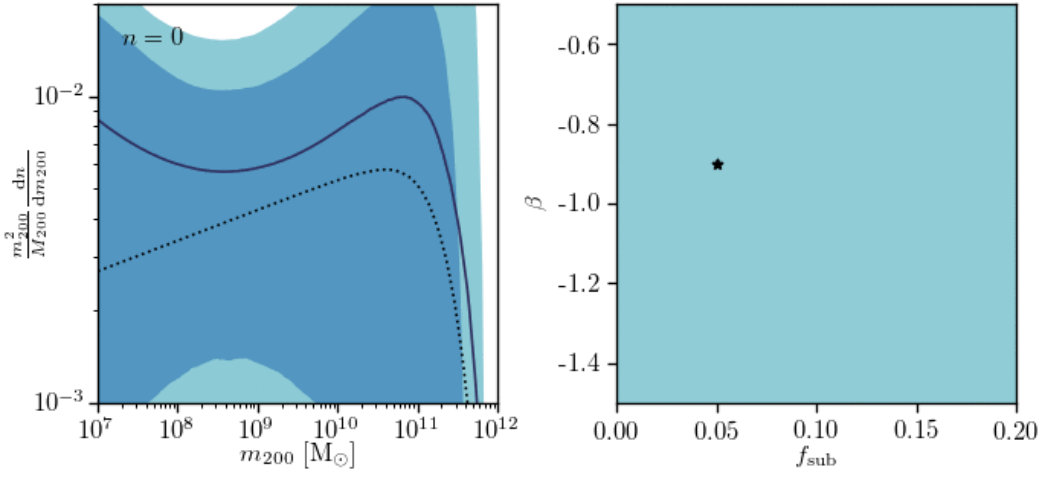
We must make sure our approximate simulation-based inference algorithms can (at least) actually realize faithful inferences on the observations we expect a priori -- i.e. those $x^{\mathrm{sim}} \sim p(x)$.



Mode convergence:

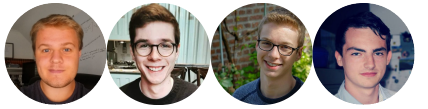
The maximum a posteriori estimate converges towards the nominal value θ^* for an increasing number of independent and identically distributed observables $x_i \sim p(x|\theta^*)$:

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



—

Credits: Brhemer et al, 2019.

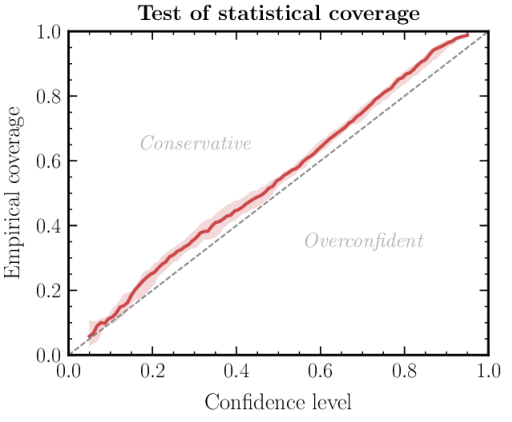


A common observation at the root of several other diagnostics is to check for the **self-consistency** of the Bayesian joint distribution,

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

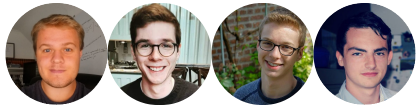
Coverage diagnostic:

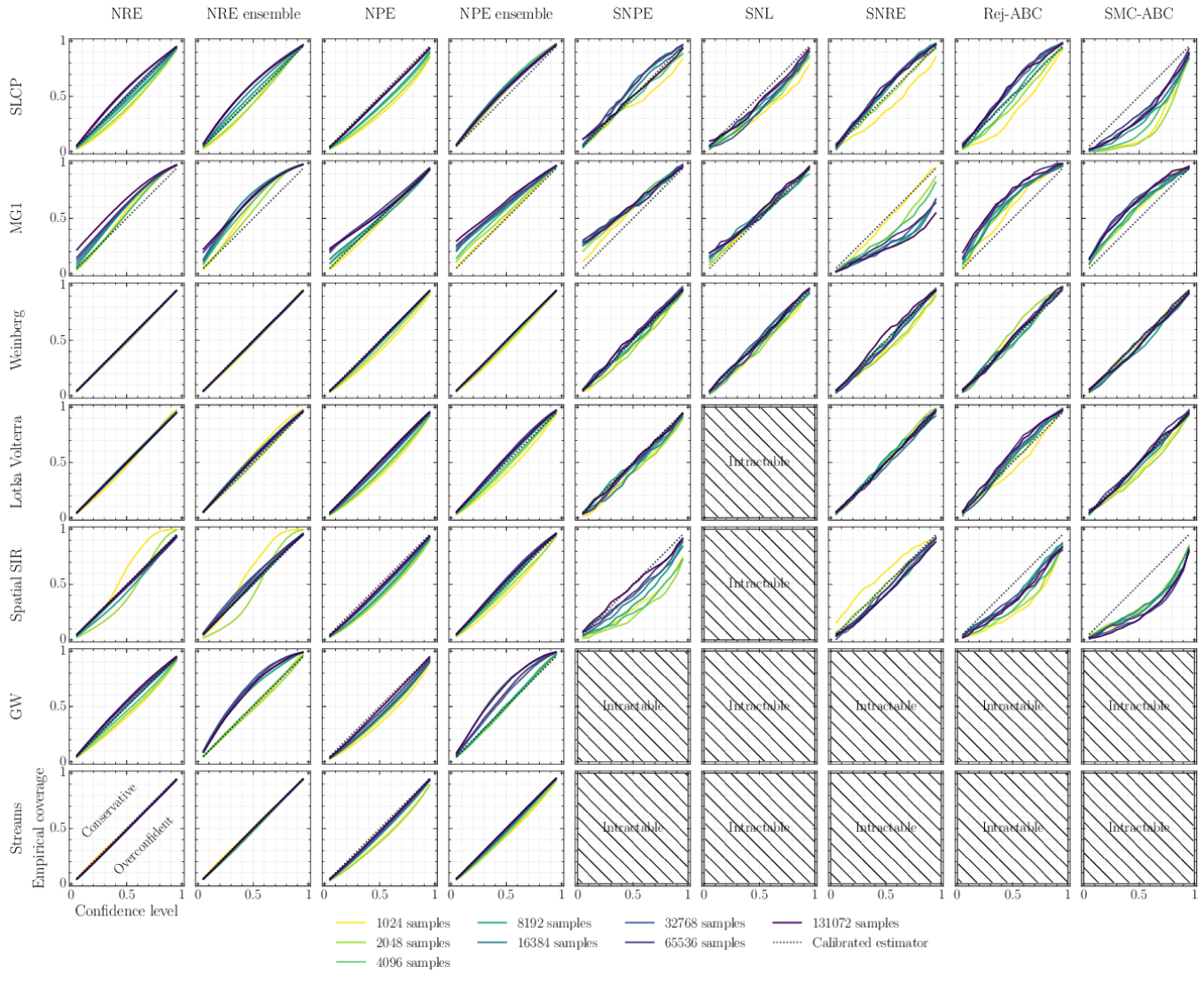
- For $x, \theta \sim p(x, \theta)$, compute the $1 - \alpha$ credible interval based on $\hat{p}\left(\theta|x\right)$.
- If the fraction of samples for which θ is contained within the interval is larger than the nominal coverage probability $1 - \alpha$, then the approximate posterior $\hat{p}\left(\theta|x\right)$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



—

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

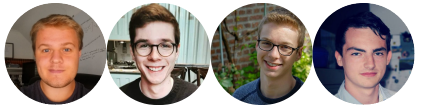




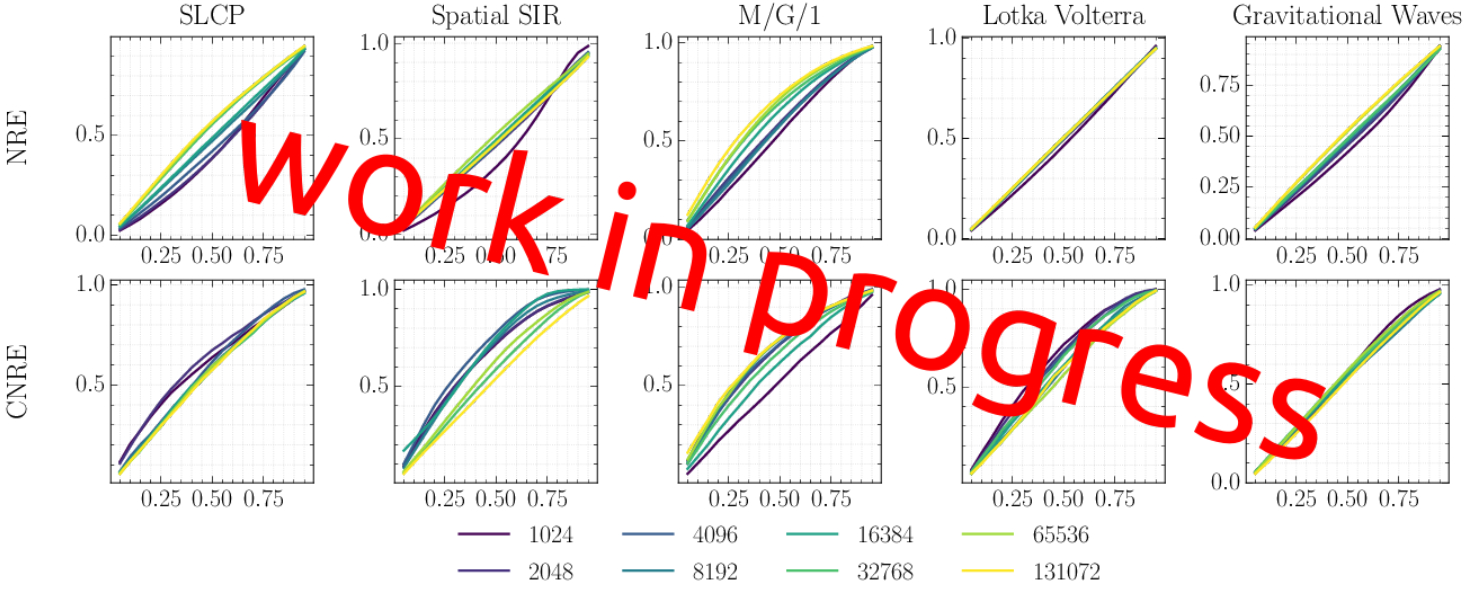
Credits: Hermans et al, 2021.

Faithfulness diagnostics require the ability to repeatedly compute $\hat{p}\left(\theta \middle| x\right)$, which is immediate for amortized approaches but **computationally very heavy for sequential inference algorithms**.

What if the diagnostic fails?



Neural ratio estimation can be forced to be more **conservative**, hence increasing the reliability of the approximate posteriors and reducing the risk of false inferences.



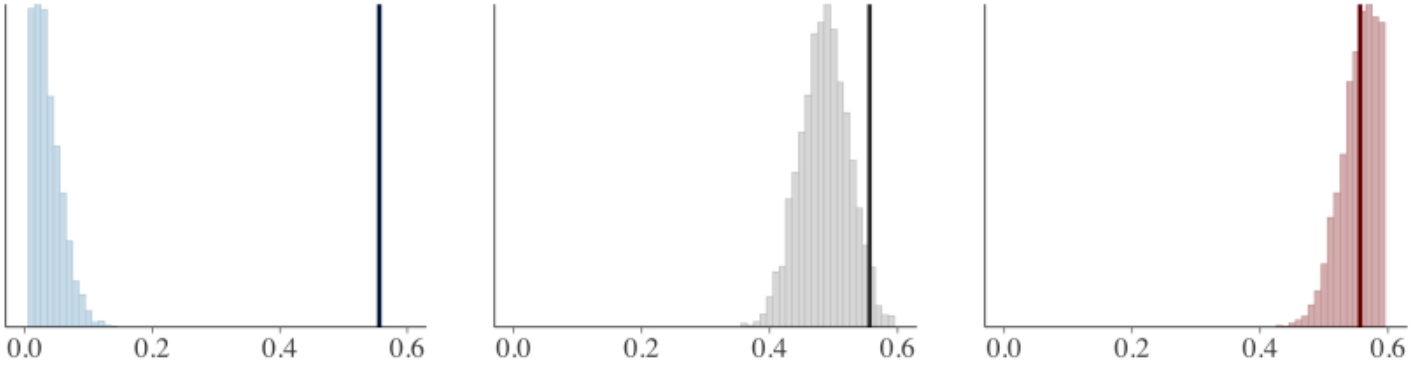
Posterior predictive checks

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.



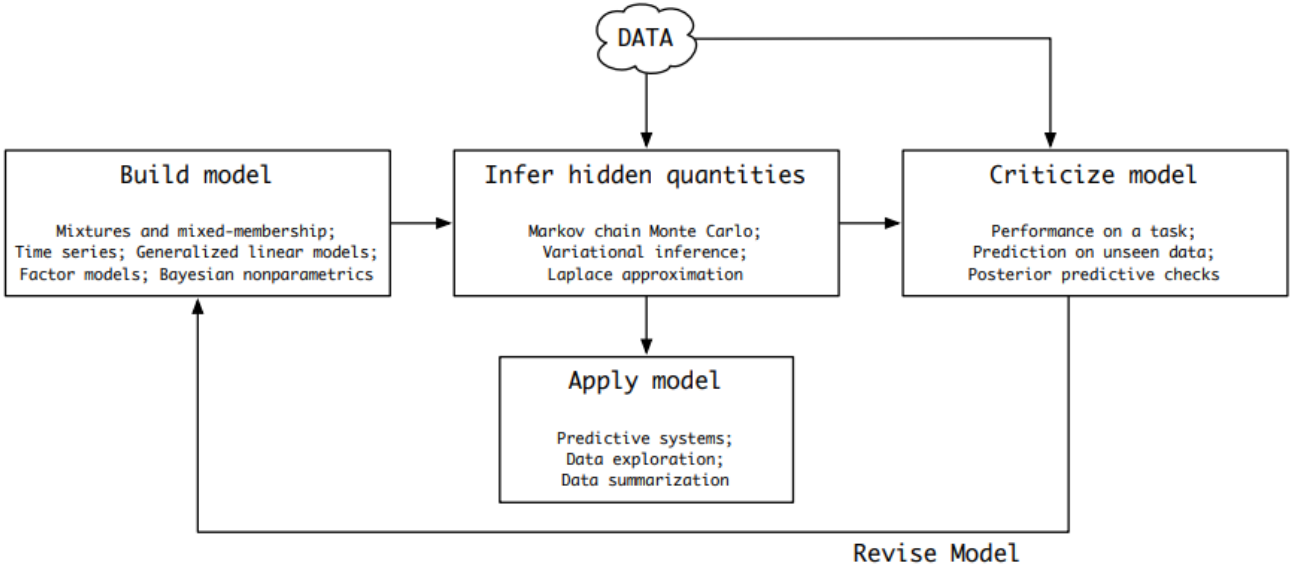
(a) Model 1 (b) Model 2 (c) Model 3

Fig. 7: *Histograms of statistics* $\text{skew}(y_{\text{rep}})$ *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



Science does not end at the inference results. Instead, they should inform the next revision of the model.

—

Credits: Blei, 2014.

Proceed with caution!

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

Need to reliably and efficiently assess the adequacy of the full Bayesian model.

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

Need to efficiently generate simulated data and use it to train ML components.

The end.