An introduction to simulation-based inference

Advanced Machine Learning
March 10, 2022

Gilles Louppe
g.louppe@uliege.be

LIÈGE université

Kyle Cranmer

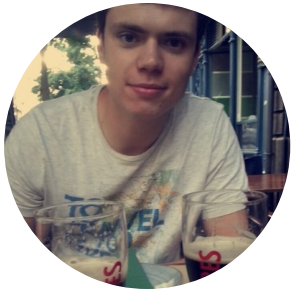
Juan Pavez

Johann Brehmer

Joeri Hermans

Antoine Wehenkel

Arnaud Delaunoy



Norman Marlier

Francois Rozet

Malavika Vasist

Christophe Weniger

Siddarth Mishra-Sharma

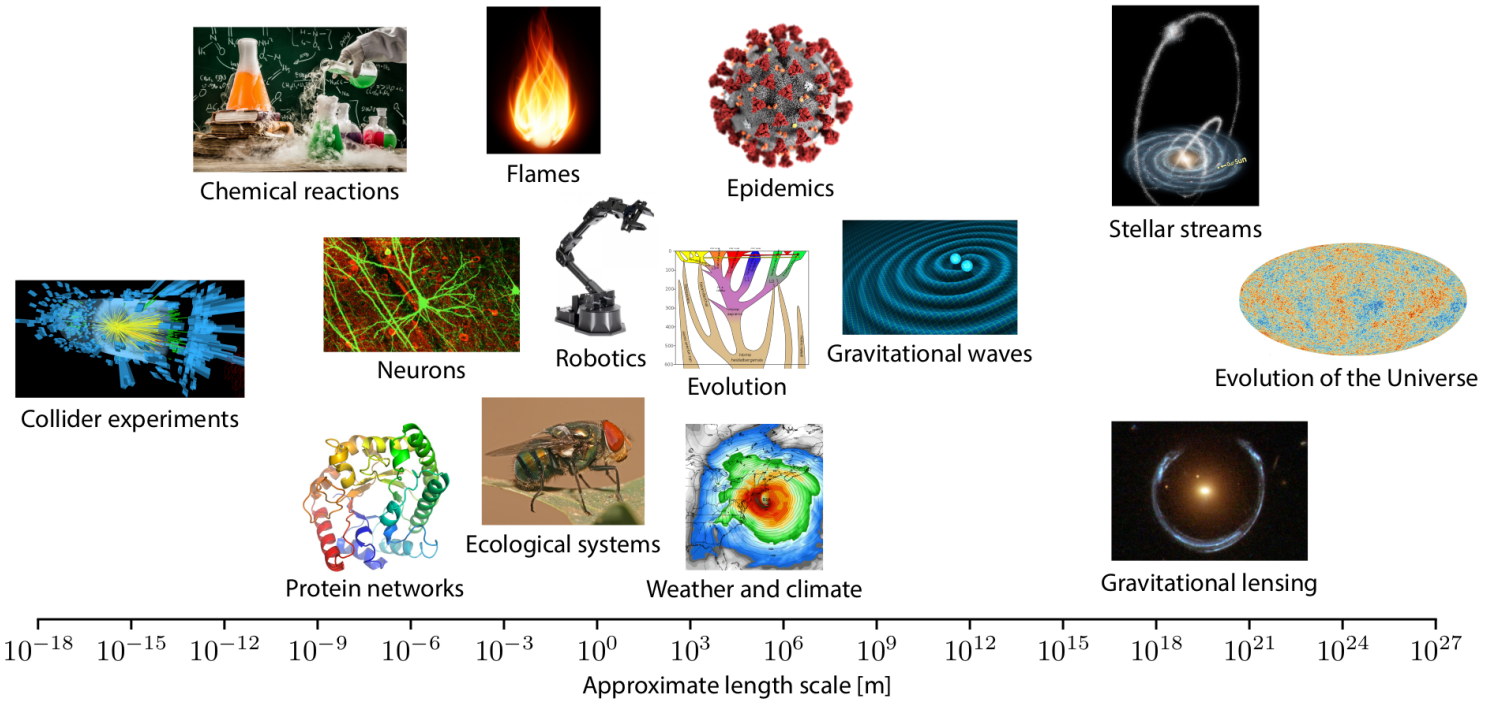
Lukas Heinrich

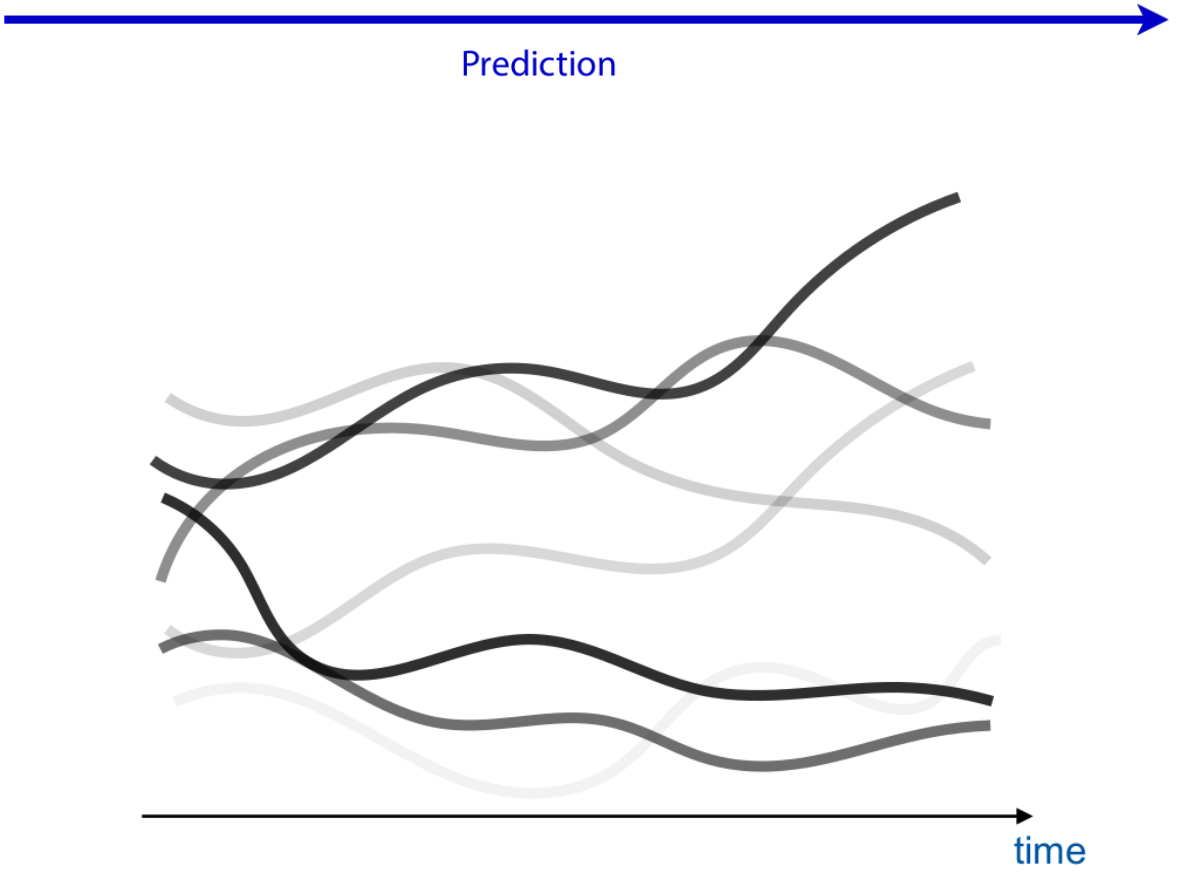
Atılım Güneş Baydin

Gianfranco Bertone

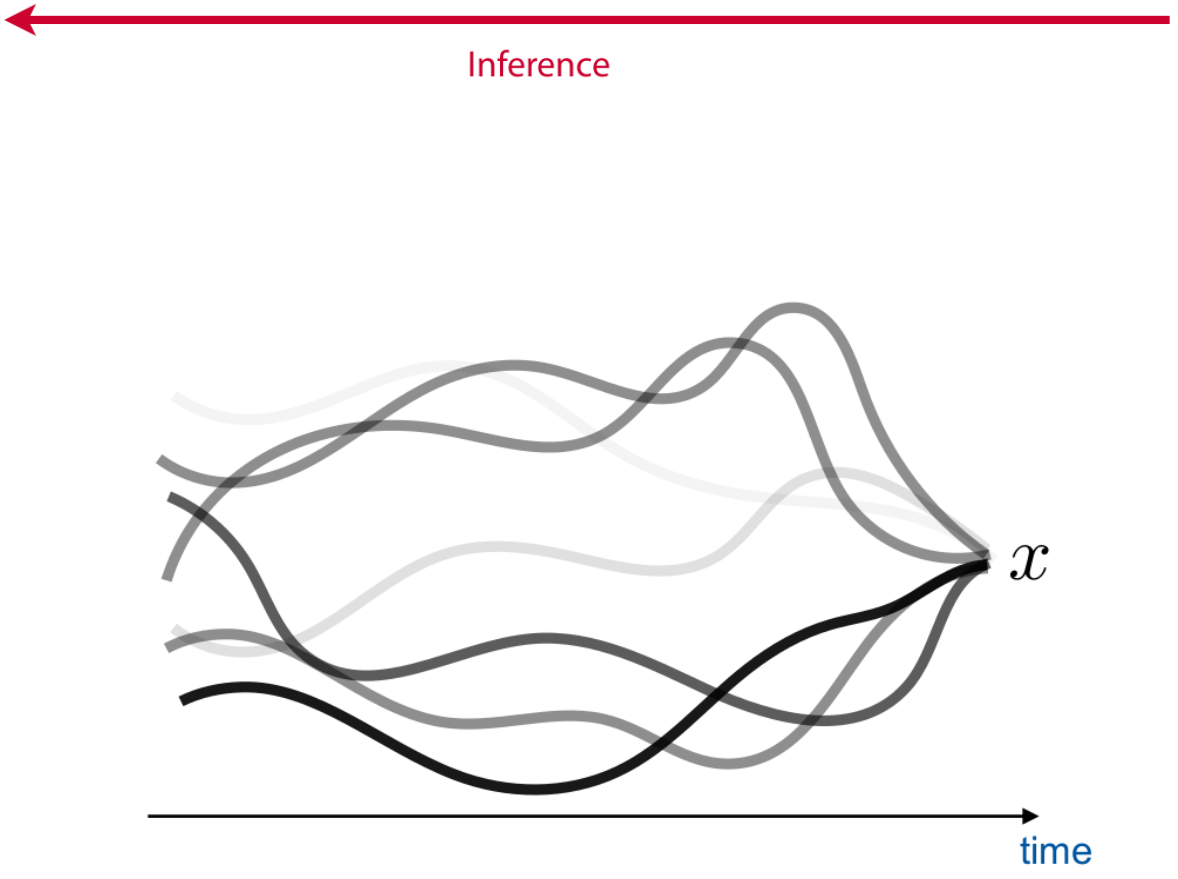
Nil Banik

Scientific simulators





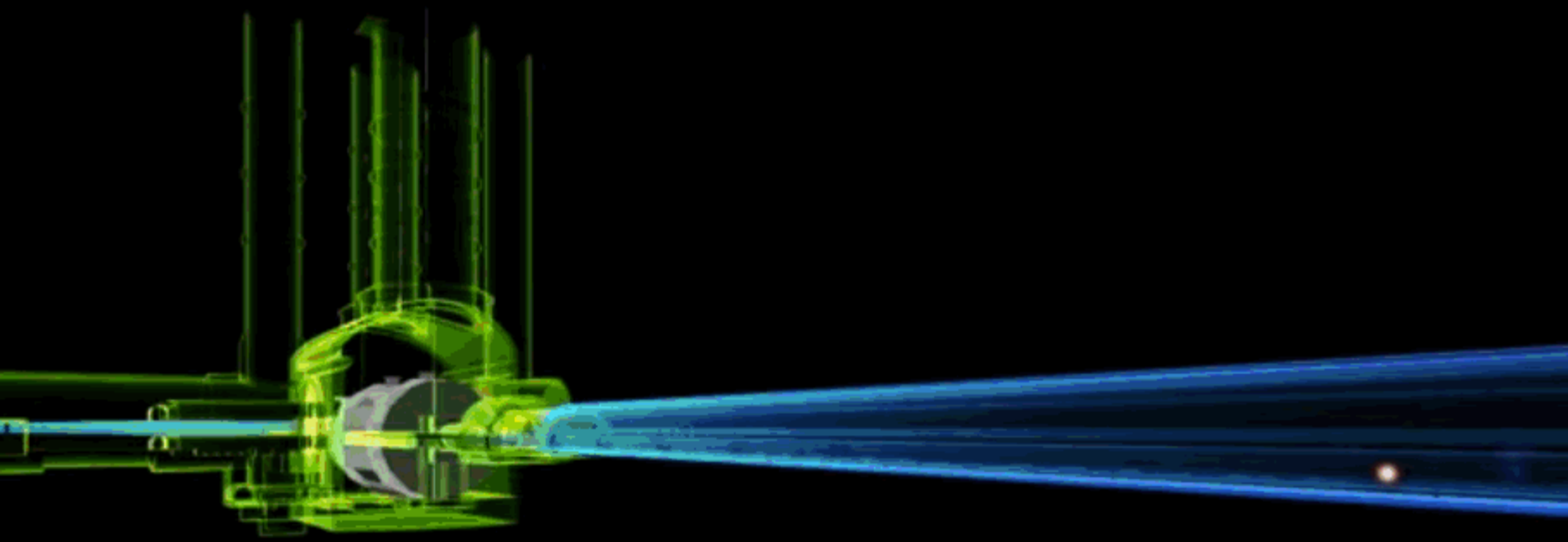
$$\theta, z, x \sim p(\theta, z, x)$$

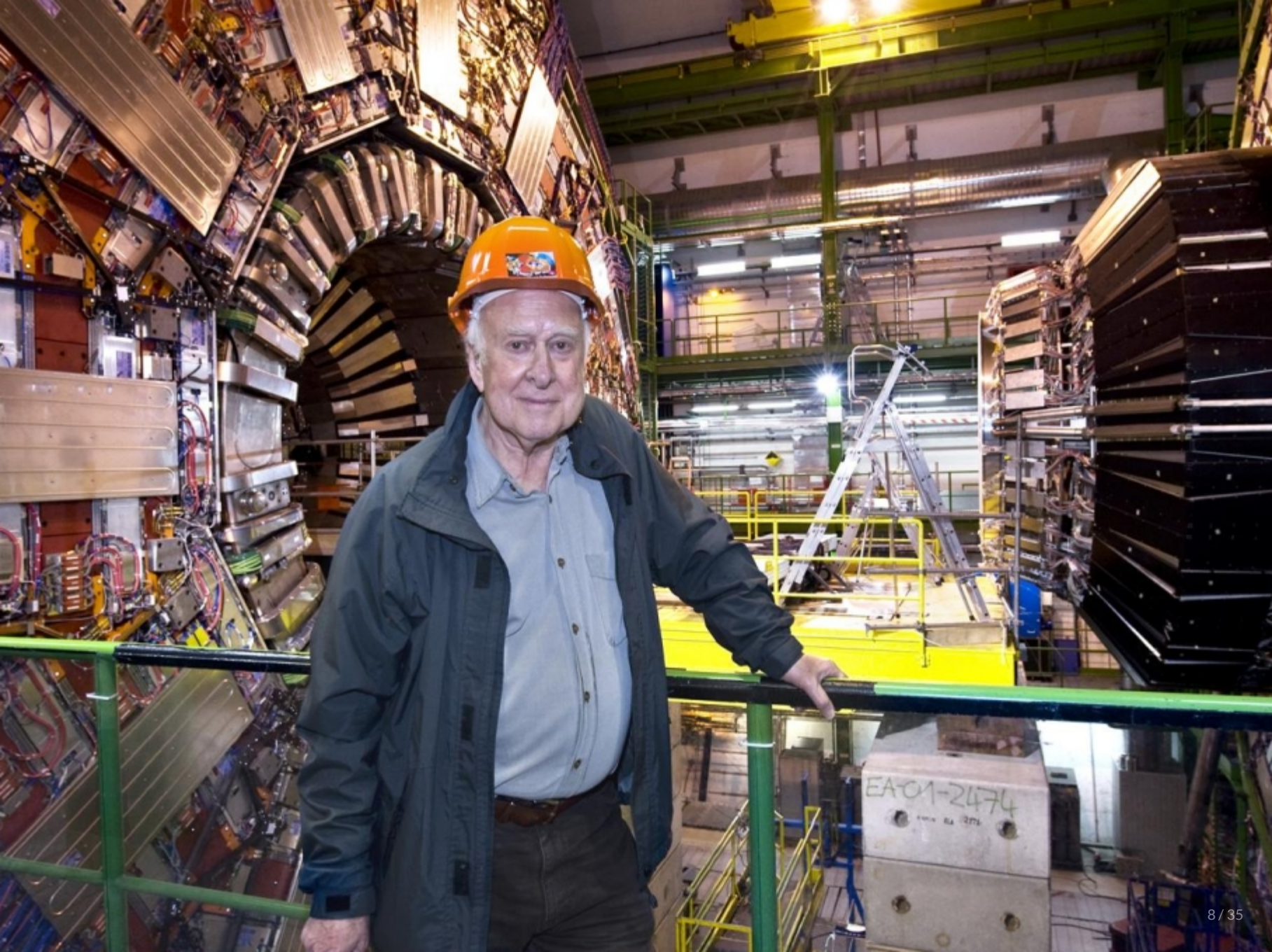


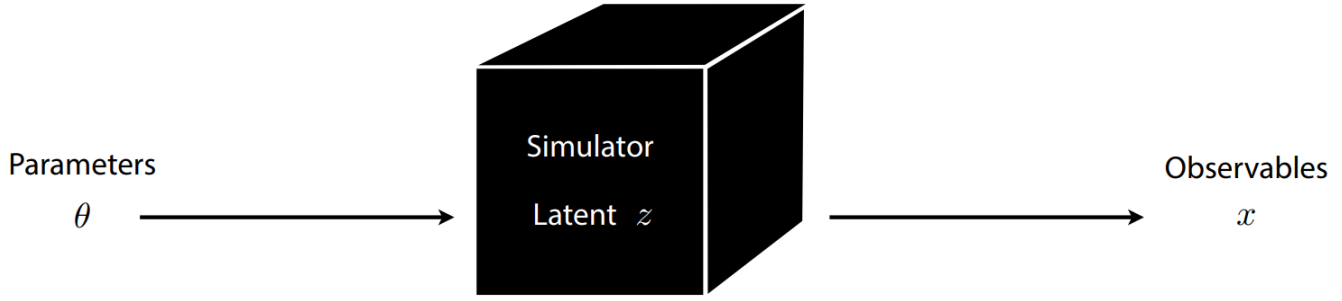
This results in the likelihood $p(x|\theta) = \int p(x, z|\theta)dz$ to be intractable.

The case of particle physics

$$\mathcal{L}_{SM} = -\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc}f^{ade}g^b_\mu g^c_\nu g^d_\mu g^e_\nu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w(\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - Z^0_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - igs_w(\partial_\nu A_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2 (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + g^2 s_w^2 (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w (A_\mu Z^0_\nu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu) - \tfrac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \tfrac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \beta_h \left(\tfrac{2M^2}{g^2} + \tfrac{2M}{g}H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-) \right) + \tfrac{2M^4}{g^2}\alpha_h - g\alpha_h M (H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-) - \tfrac{1}{8}g^2 \alpha_h (H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2 H^2) - gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac{1}{2}ig\left(W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)\right) + \tfrac{1}{2}g\left(W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) + W^-_\mu(H\partial_\mu\phi^+ - \phi^+\partial_\mu H)\right) + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) + M\left(\tfrac{1}{c_w}Z^0_\mu\partial_\mu\phi^0 + W^+_\mu\partial_\mu\phi^- + W^-_\mu\partial_\mu\phi^+\right) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + igs_wMA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + igs_wA_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac{1}{4}g^2W^+_\mu W^-_\mu(H^2 + (\phi^0)^2 + 2\phi^+\phi^-) - \tfrac{1}{8}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu(H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2\phi^+\phi^-) - \tfrac{1}{2}g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) + \tfrac{1}{2}ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - g^2 s_w^2 A_\mu A_\mu\phi^+\phi^- + \tfrac{1}{2}ig_s\lambda^a_{ij}(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu - \bar{e}^\lambda(\gamma\partial + m^\lambda_e)e^\lambda - \bar{\nu}^\lambda(\gamma\partial + m^\lambda_\nu)\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + m^\lambda_u)u^\lambda_j - \bar{d}^\lambda_j(\gamma\partial + m^\lambda_d)d^\lambda_j + igs_wA_\mu\left(-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)\right) + \tfrac{ig}{4c_w}Z^0_\mu\{(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{d}^\lambda_j\gamma^\mu(\tfrac{4}{3}s_w^2 - 1 - \gamma^5)d^\lambda_j) + (\bar{u}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 + \gamma^5)u^\lambda_j)\} + \tfrac{ig}{2\sqrt{2}}W^+_\mu\left((\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)U^{lep}_{\lambda\kappa}e^\kappa) + (\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)\right) + \tfrac{ig}{2\sqrt{2}}W^-_\mu\left((\bar{e}^\kappa U^{lep\dagger}_{\kappa\lambda}\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\kappa\lambda}\gamma^\mu(1+\gamma^5)u^\lambda_j)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_e(\bar{\nu}^\lambda U^{lep}_{\lambda\kappa}(1-\gamma^5)e^\kappa) + m^\lambda_\nu(\bar{\nu}^\lambda U^{lep}_{\lambda\kappa}(1+\gamma^5)e^\kappa)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_e(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1+\gamma^5)\nu^\kappa) - m^\kappa_\nu(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1-\gamma^5)\nu^\kappa)\right) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{\nu}^\lambda\nu^\lambda) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{e}^\lambda e^\lambda) + \tfrac{ig}{2}\tfrac{m^\lambda_e}{M}\phi^0(\bar{\nu}^\lambda\gamma^5\nu^\lambda) - \tfrac{ig}{2}\tfrac{m^\lambda_e}{M}\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda) - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_d(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + m^\lambda_u(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_d(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m^\kappa_u(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1-\gamma^5)u^\kappa_j\right) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m^\lambda_u}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \tfrac{ig}{2}\tfrac{m^\lambda_d}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j) + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar{G}^a G^b g^c_\mu + \bar{X}^+(\partial^2 - M^2)X^+ + \bar{X}^-(\partial^2 - M^2)X^- + \bar{X}^0(\partial^2 - \tfrac{M^2}{c_w^2})X^0 + \bar{Y}\partial^2 Y + igc_w W^+_\mu(\partial_\mu\bar{X}^0X^- - \partial_\mu\bar{X}^+X^0) + igs_w W^+_\mu(\partial_\mu\bar{Y}X^- - \partial_\mu\bar{X}^+Y) + igc_w W^-_\mu(\partial_\mu\bar{X}^-X^0 - \partial_\mu\bar{X}^0X^+) + igs_w W^-_\mu(\partial_\mu\bar{X}^-Y - \partial_\mu\bar{Y}X^+) + igc_w Z^0_\mu(\partial_\mu\bar{X}^+X^+ - \partial_\mu\bar{X}^-X^-) + igs_w A_\mu(\partial_\mu\bar{X}^+X^+ - \partial_\mu\bar{X}^-X^-) - \tfrac{1}{2}gM\left(\bar{X}^+X^+H + \bar{X}^-X^-H + \tfrac{1}{c_w^2}\bar{X}^0X^0H\right) + \tfrac{1-2c_w^2}{2c_w}igM\left(\bar{X}^+X^0\phi^+ - \bar{X}^-X^0\phi^-\right) + \tfrac{1}{2c_w}igM\left(\bar{X}^0X^-\phi^+ - \bar{X}^0X^+\phi^-\right) + igMs_w\left(\bar{X}^0X^-\phi^+ - \bar{X}^0X^+\phi^-\right) + \tfrac{1}{2}igM\left(\bar{X}^+X^+\phi^0 - \bar{X}^-X^-\phi^0\right).$$

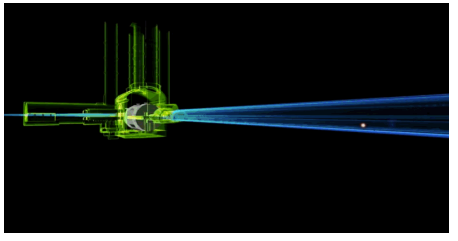



EA01-2474

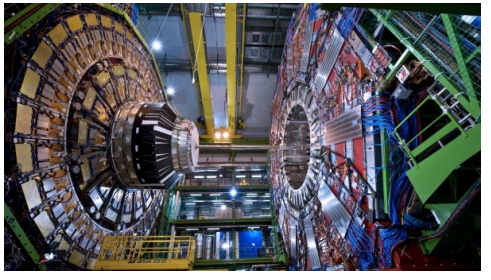


SM with parameters θ

Simulated observables x



Real observations x_{obs}



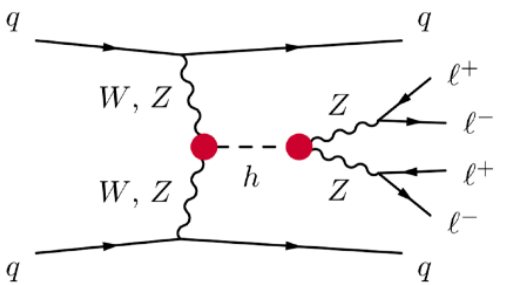
Latent variables

Parameters of interest

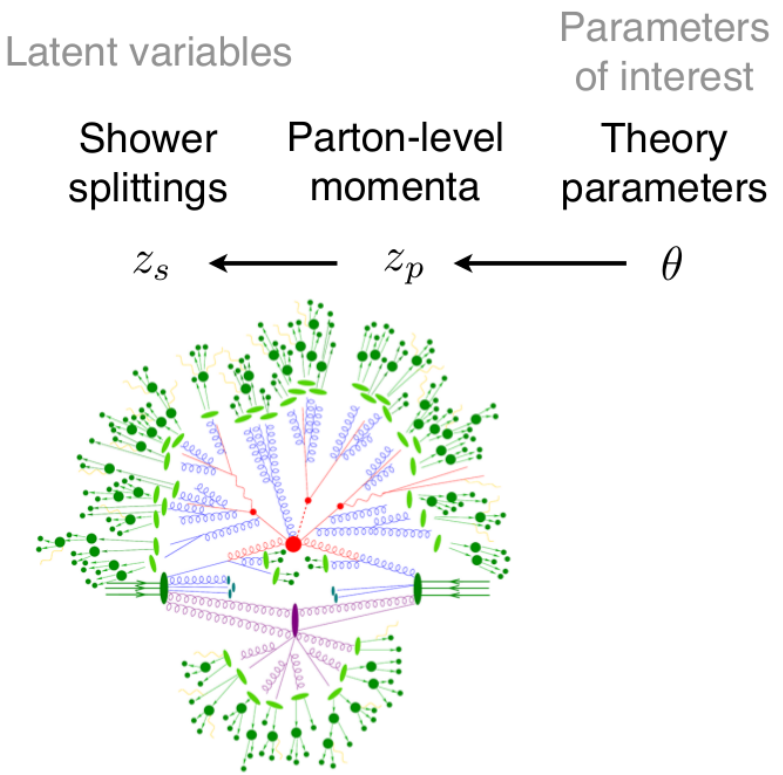
Parton-level momenta

Theory parameters

$z_p \longleftarrow \theta$



Credits: Johann Brehmer.



Credits: Johann Brehmer.

Latent variables

Parameters of interest

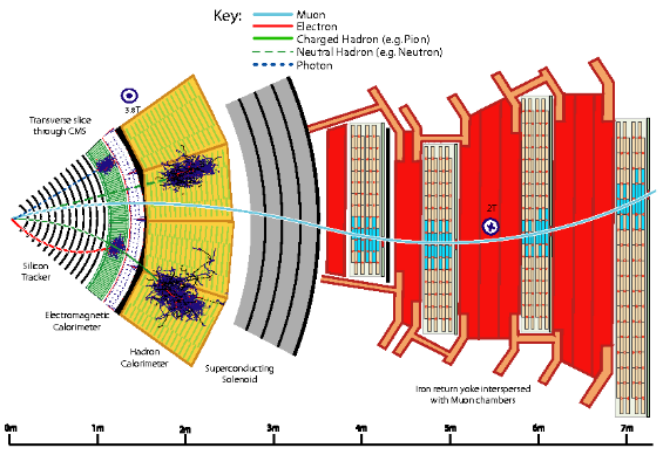
Detector interactions

Shower splittings

Parton-level momenta

Theory parameters

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



Credits: Johann Brehmer.

Features

Latent variables

Parameters of interest

Observables

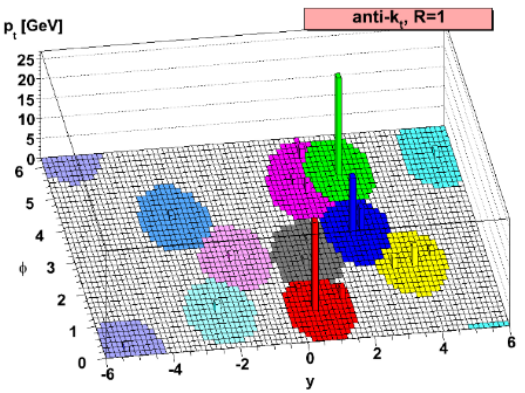
Detector interactions

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]

Credits: Johann Brehmer.

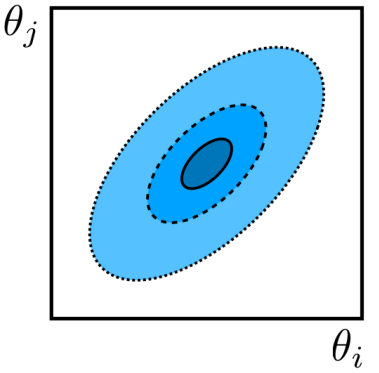
$$p(x|\theta) = \underbrace{\iiint}_{\text{intractable}} p(z_p|\theta)p(z_s|z_p)p(z_d|z_s)p(x|z_d)dz_p dz_s dz_d$$

Inference, the frequentist (physicist's) way

The Neyman-Pearson lemma states that the likelihood ratio

$$r(x|\theta_0, \theta_1) = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

is the most powerful test statistic to discriminate between a null hypothesis θ_0 and an alternative θ_1.



IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. NEYMAN, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. PEARSON, *Department of Applied Statistics, University College, London.*

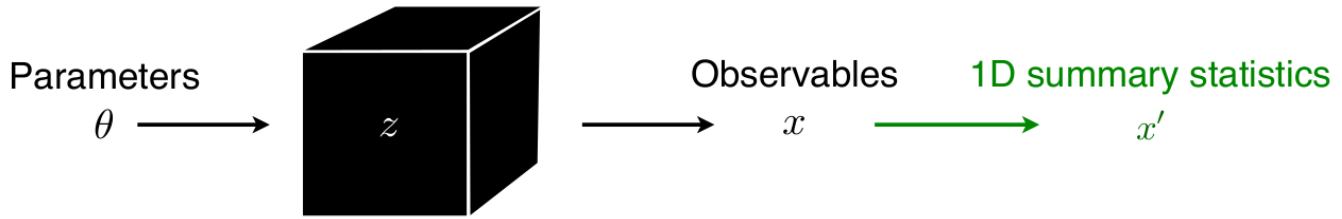
(*Communicated by* K. PEARSON, *F.R.S.*)

(Received August 31, 1932.—Read November 10, 1932.)

CONTENTS.

PAGE.

I. Introductory 289
II. Outline of General Theory 293



Define a projection function $s : \mathcal{X} \to \mathbb{R}$ mapping observables x to a summary statistic $x' = s(x)$.

Then, approximate the likelihood $p(x|\theta)$ with the surrogate $\hat{p}(x|\theta) = p(x'|\theta)$.

From this it comes

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} \approx \frac{\hat{p}(x|\theta_0)}{\hat{p}(x|\theta_1)} = \hat{r}(x|\theta_0, \theta_1).$$

Wilks theorem

Consider the test statistic

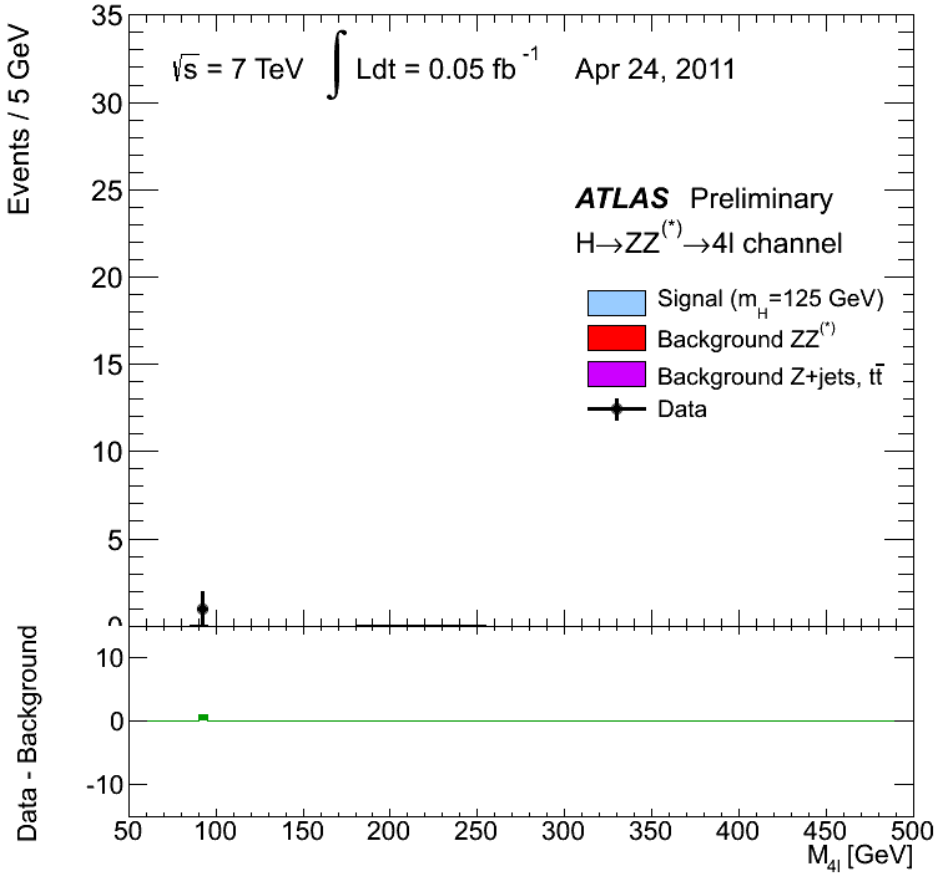
$$q(\theta) = -2 \sum_x \log \frac{p(x|\theta)}{p(x|\hat{\theta})} = -2 \sum_x \log r(x|\theta, \hat{\theta})$$

for a fixed number N of observations $\{x\}$ and where $\hat{\theta}$ is the maximum likelihood estimator.

When $N \to \infty, q(\theta) \sim \chi_2$.

Therefore (and provided the assumptions apply!), an observed value $q_{\text{obs}}(\theta)$ translates directly to a p-value that measures the confidence with which θ can be excluded:

$$p_\theta \equiv \int_{q_{\text{obs}}(\theta)}^{\infty} p(q|\theta) \mathrm{d}q = 1 - F_{\chi_2}(q_{\text{obs}}(\theta)).$$



Discovery of the Higgs boson

Inference algorithms

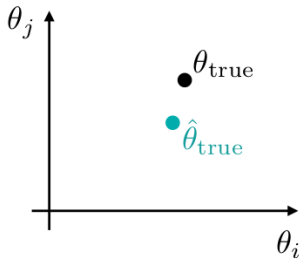
Problem statement

Start with

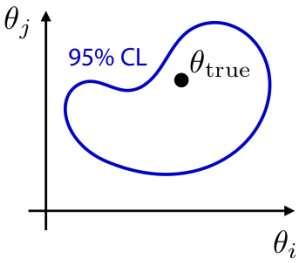
- a simulator that lets you generate N samples $x_i \sim p(x_i|\theta_i)$,
- observed data $x_{\rm obs} \sim p(x_{\rm obs}|\theta_{\rm true})$,
- a prior $p(\theta)$.

Then,

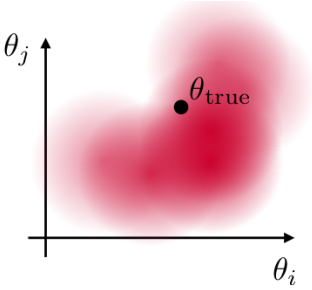
a) estimate $\theta_{\rm true}$ (e.g., MLE)

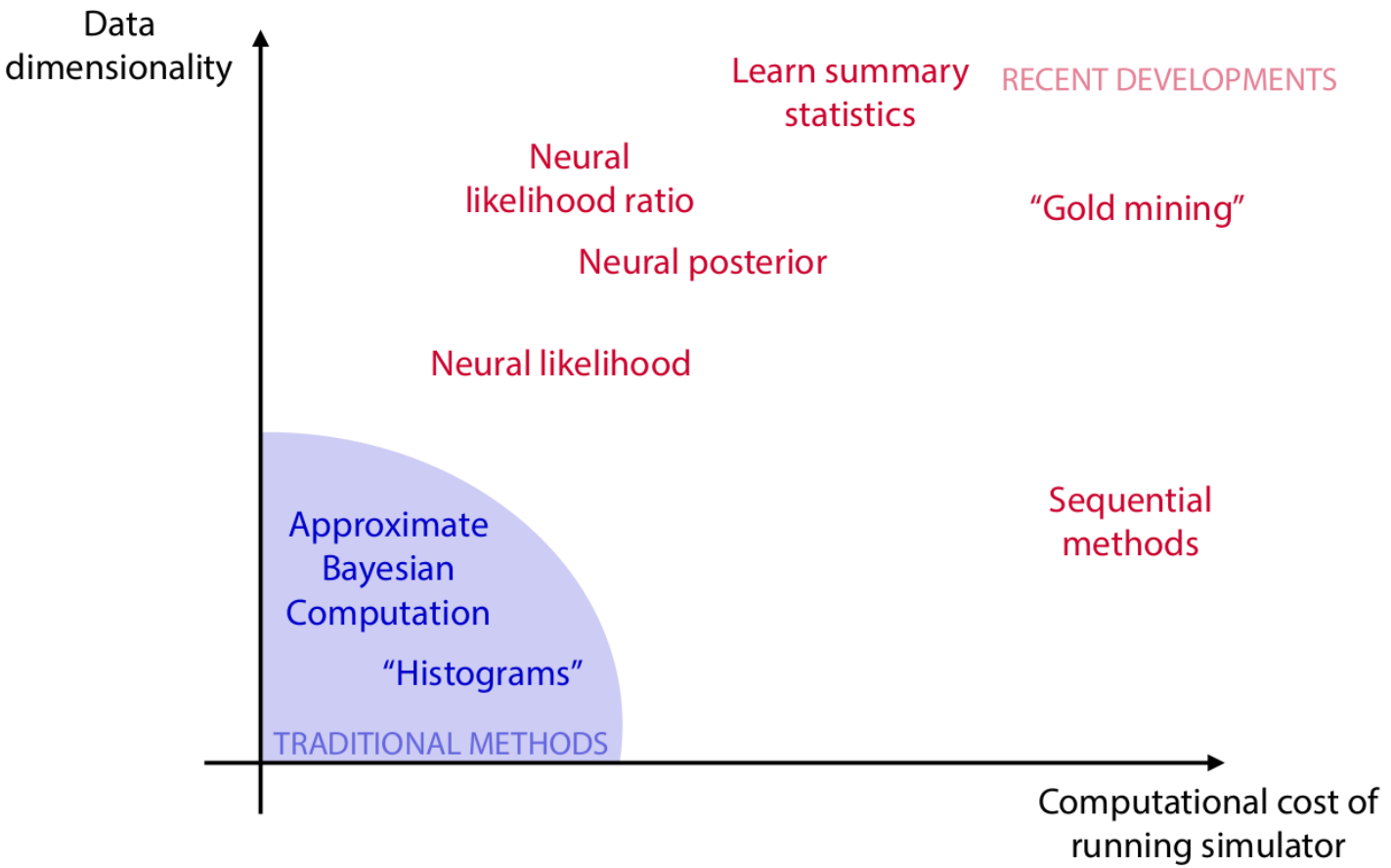


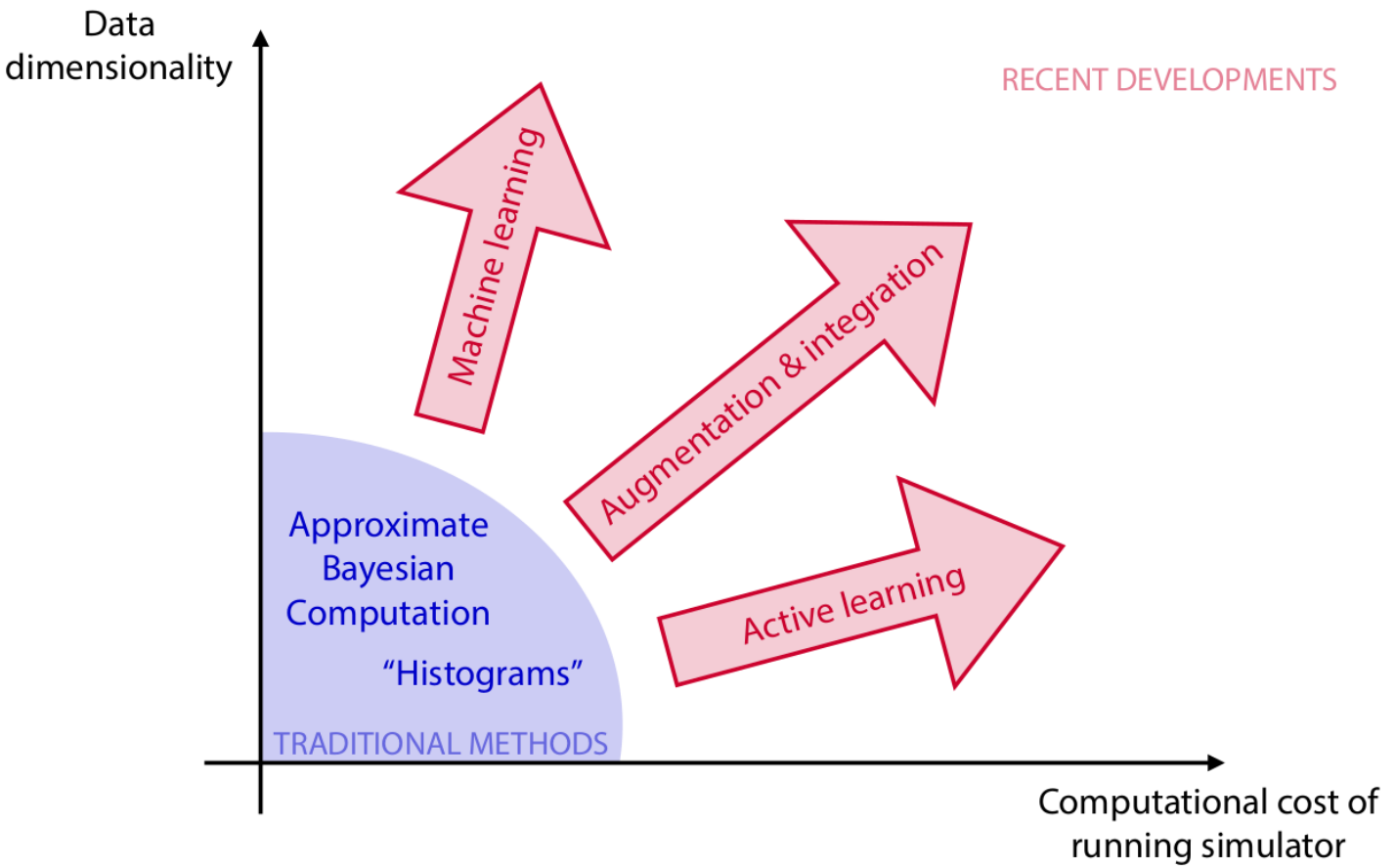
b) construct confidence sets

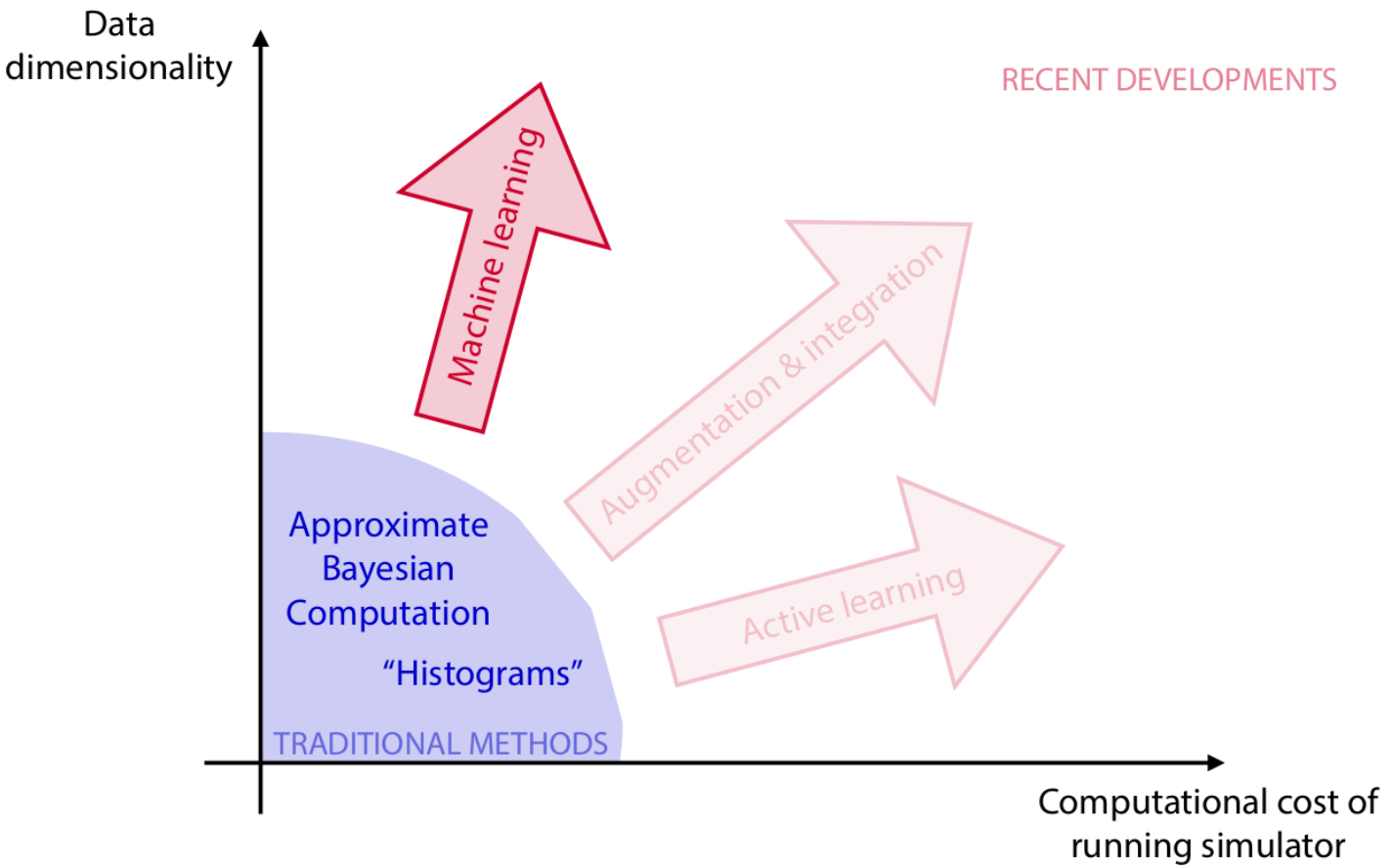


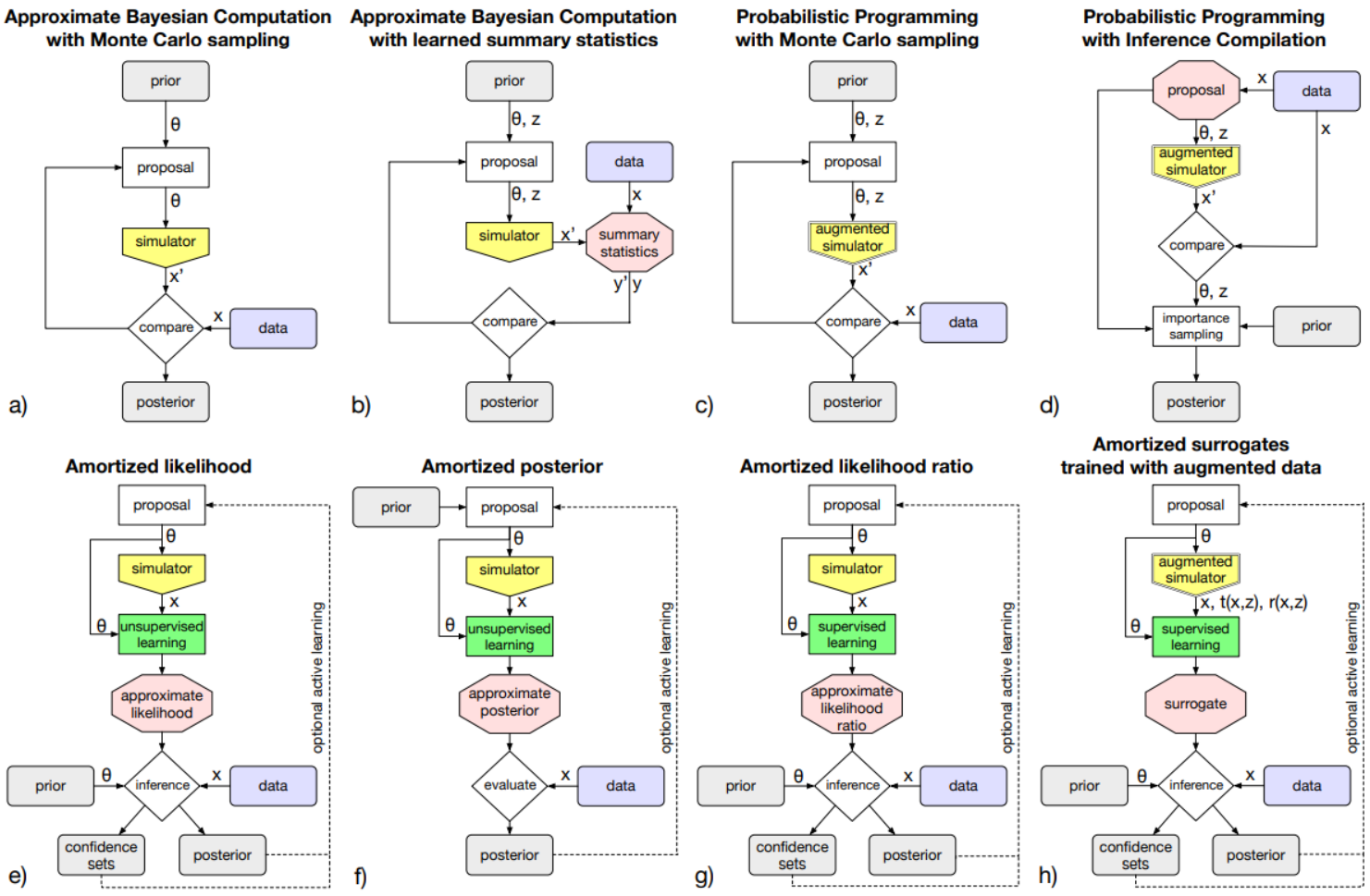
c) estimate the posterior $p(\theta|x_{\rm obs})$




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


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


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

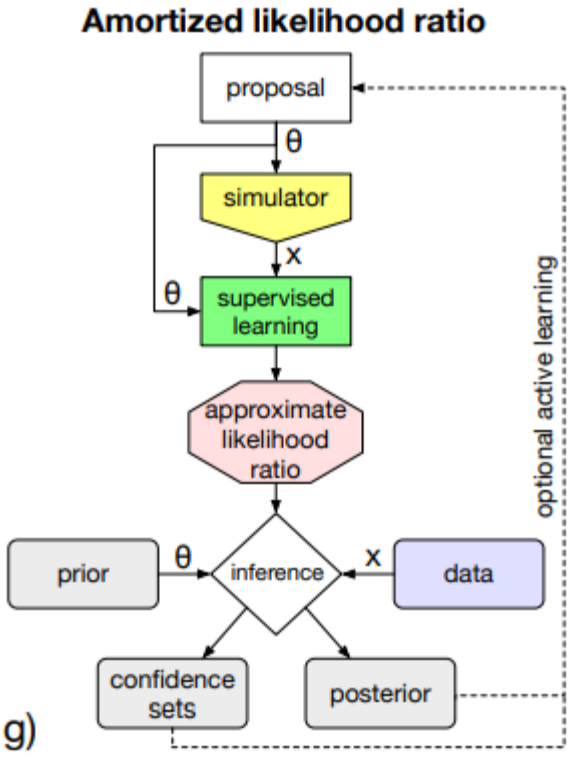

Approximate Bayesian Computation with Monte Carlo sampling
prior
θ
proposal
θ
simulator
x'
compare
x
data
posterior
a)
Approximate Bayesian Computation with learned summary statistics
prior
θ, z
proposal
data
x
θ, z
simulator
x'
summary statistics
y'
y
compare
posterior
b)
Probabilistic Programming with Monte Carlo sampling
prior
θ, z
proposal
θ, z
augmented simulator
x'
compare
x
data
posterior
c)
Probabilistic Programming with Inference Compilation
proposal
x
data
x
θ, z
augmented simulator
x'
compare
θ, z
importance sampling
prior
posterior
d)
Amortized likelihood
proposal
θ
simulator
x
θ
unsupervised learning
approximate likelihood
prior
θ
inference
x
data
confidence sets
posterior
optional active learning
e)
Amortized posterior
prior
proposal
θ
simulator
x
θ
unsupervised learning
approximate posterior
evaluate
x
data
posterior
optional active learning
f)
Amortized likelihood ratio
proposal
θ
simulator
x
θ
supervised learning
approximate likelihood ratio
prior
θ
inference
x
data
confidence sets
posterior
optional active learning
g)
Amortized surrogates trained with augmented data
proposal
θ
augmented simulator
x, t(x,z), r(x,z)
θ
supervised learning
surrogate
prior
θ
inference
x
data
confidence sets
posterior
optional active learning
h)

Neural Ratio Estimation (NRE)

The Bayes rule can be rewritten as

$$
\begin{aligned}
p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\
&= r(x|\theta)p(\theta) \\
&\approx \hat{r}(x|\theta)p(\theta)
\end{aligned}
$$

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

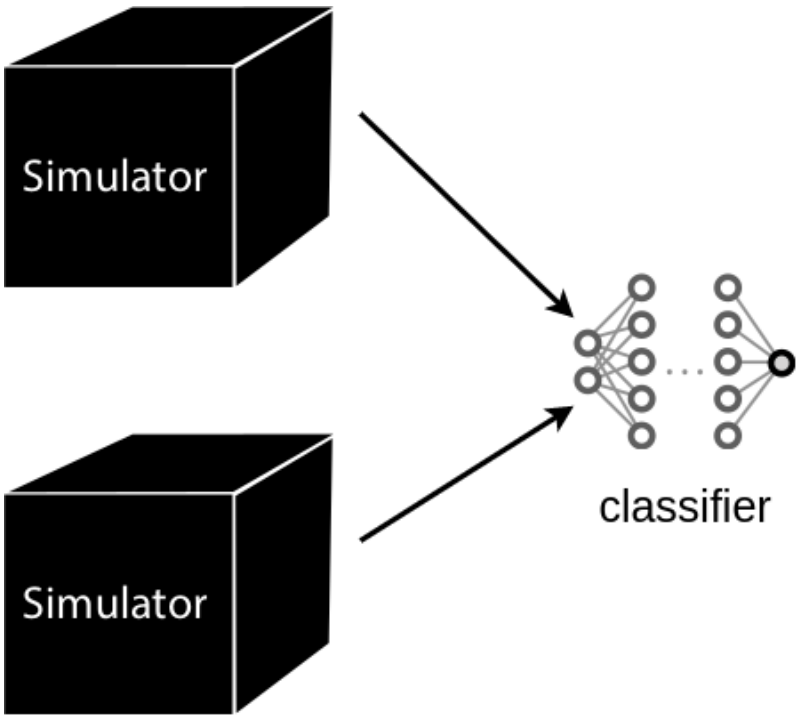


The likelihood ratio trick

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

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

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



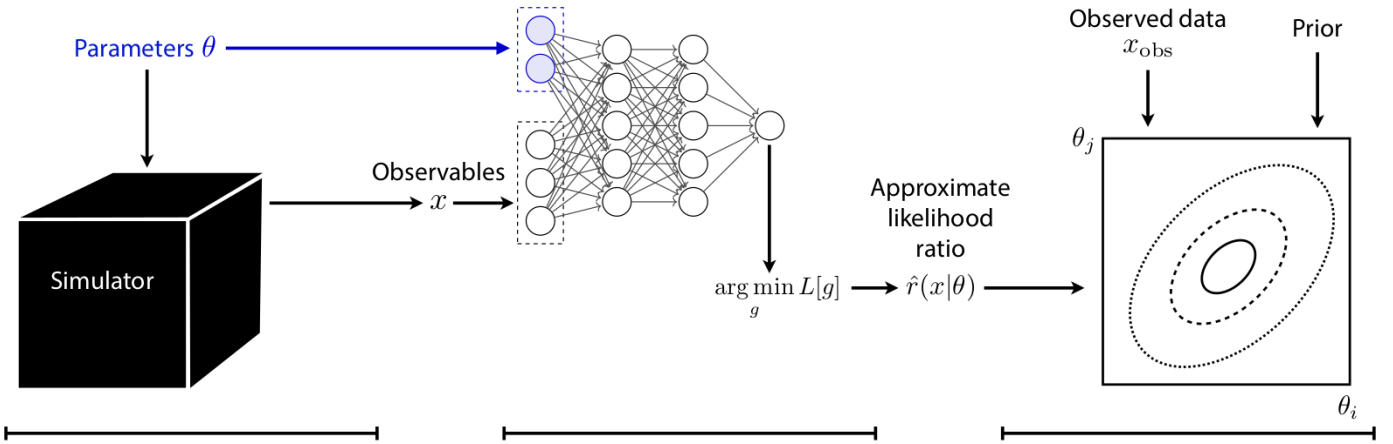
The solution d found after training approximates the optimal classifier

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

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



Run simulator and save data

Train NN classifier, interpret as likelihood ratio estimator

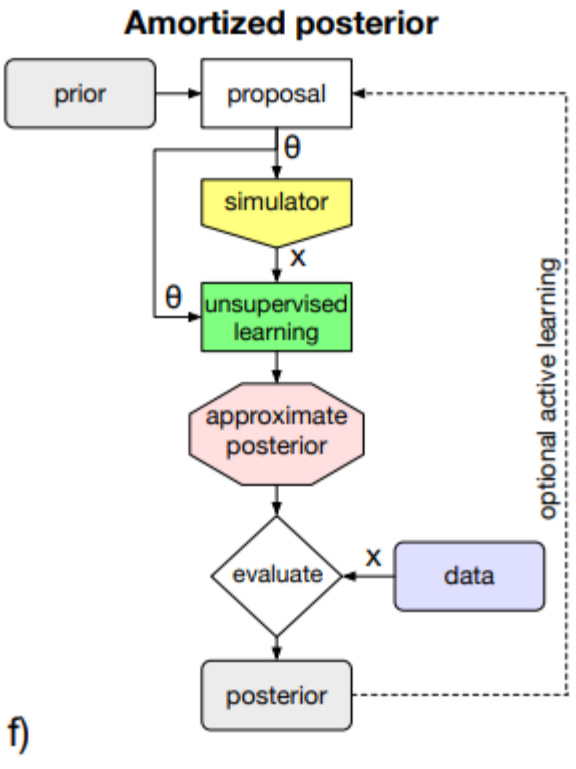
Amortized: cheap to repeat for new data

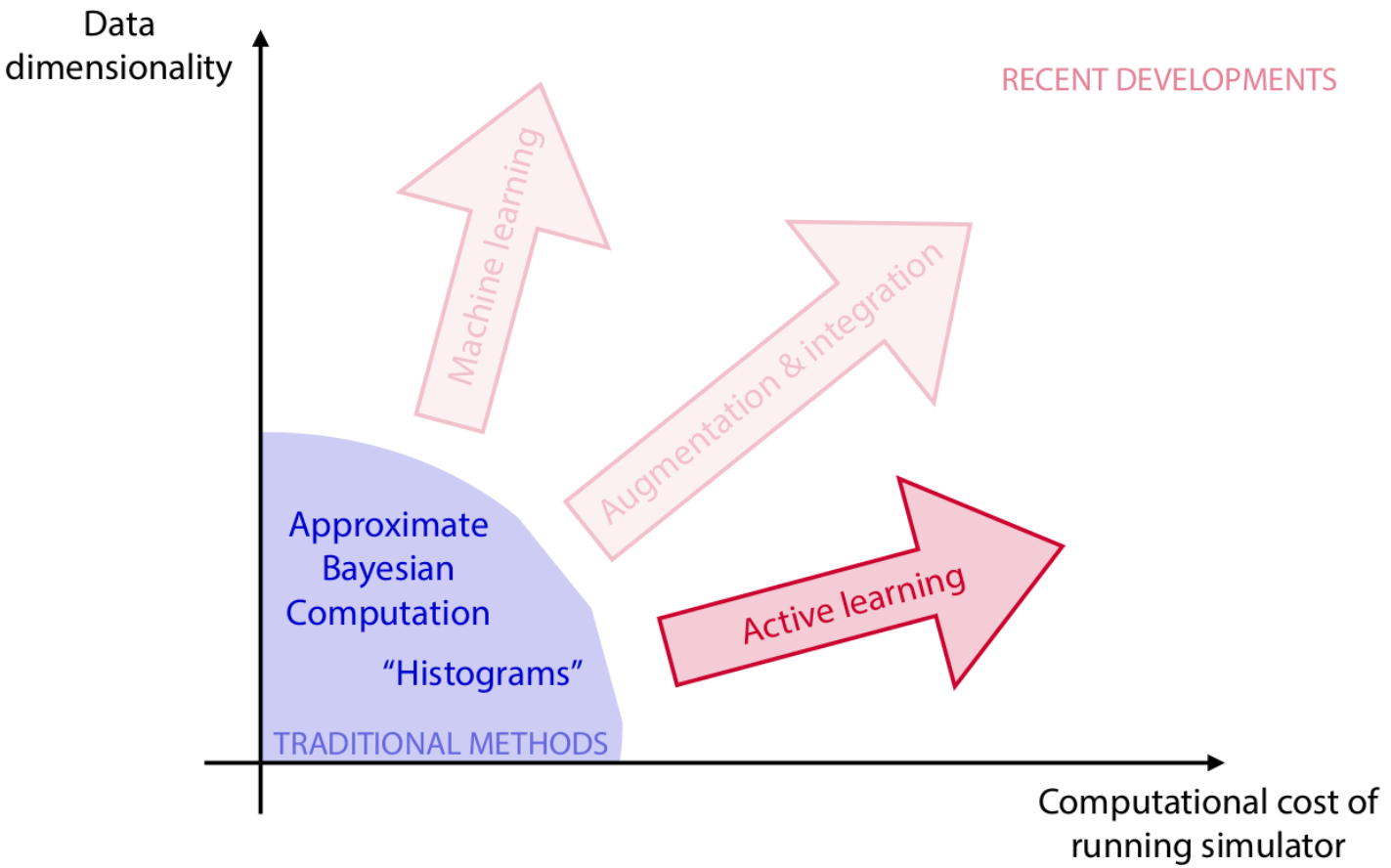
Neural Posterior Estimation (NPE)

Use variational inference to directly estimate the posterior:

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

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

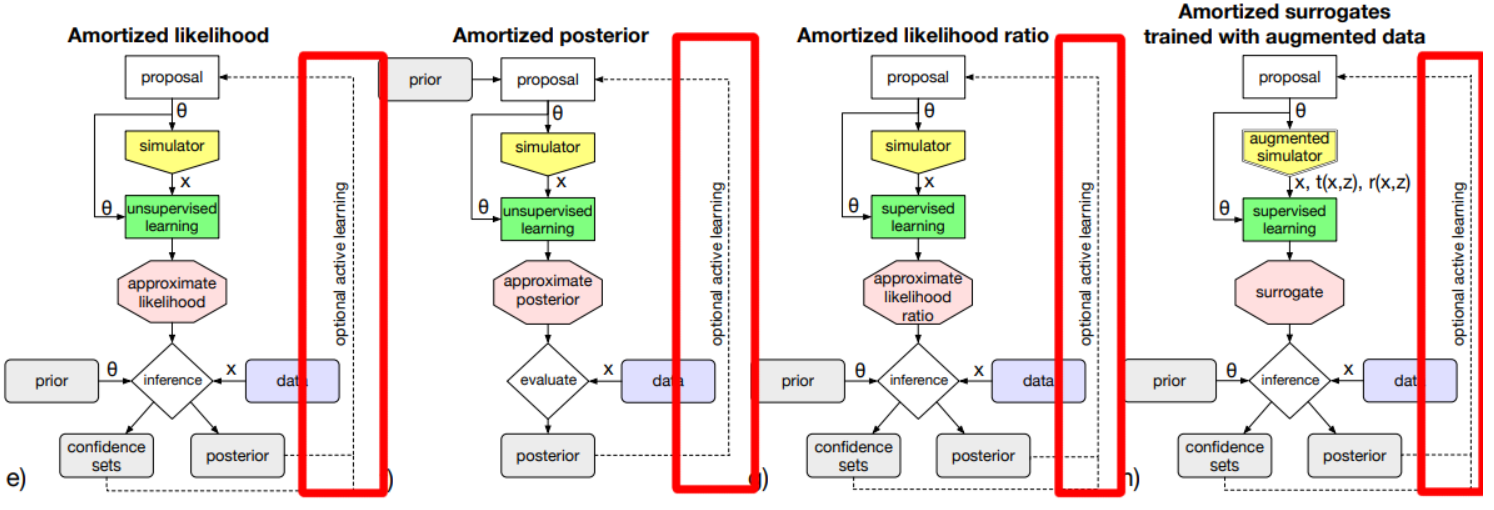



Data
dimensionality
RECENT DEVELOPMENTS
Machine learning
Augmentation & integration
Active learning
Approximate
Bayesian
Computation
"Histograms"
TRADITIONAL METHODS
Computational cost of
running simulator

Sequential estimation

When the posterior concentrates significantly compared to the prior, then we do not need to estimate the likelihood or posterior accurately everywhere:

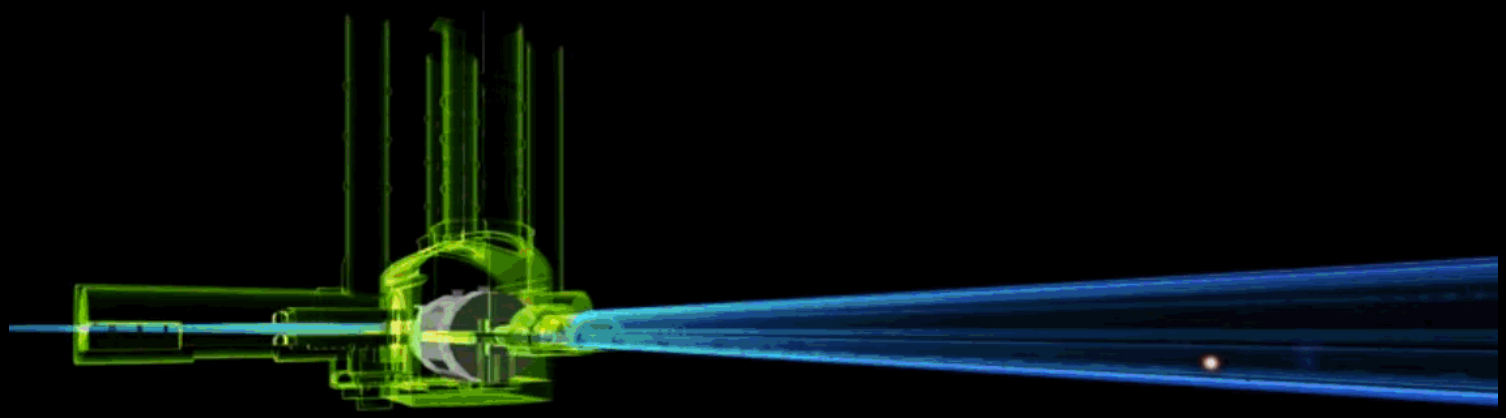
- Instead, we can approximate the likelihood or posterior only in relevant regions of the parameter or data space.
- Active learning: iteratively estimate the likelihood or the posterior $p_t(\theta|x)$, sample $\theta \sim p_t(\theta|x), x \sim p(x|\theta)$ and then refine.



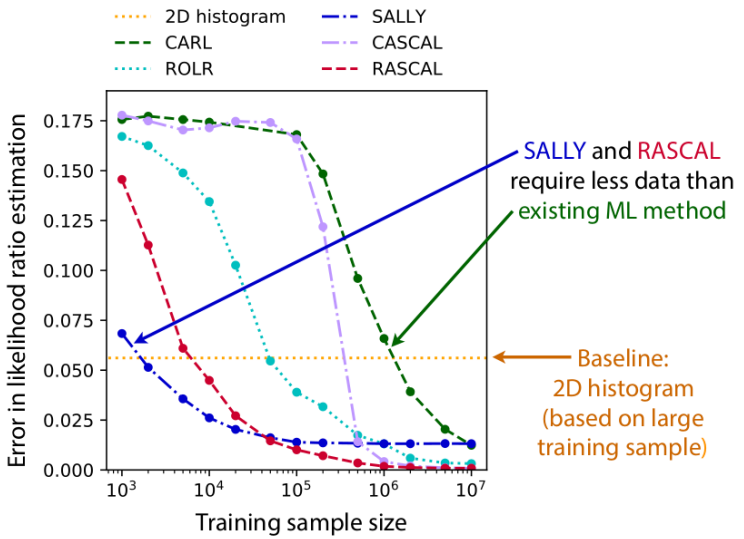
—

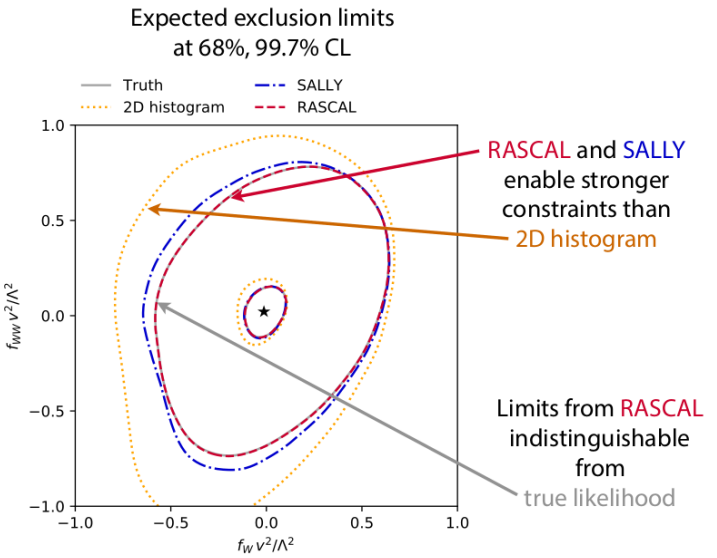
Credits: Kyle Cranmer.

Showtime!



Case 1: Hunting new physics at particle colliders

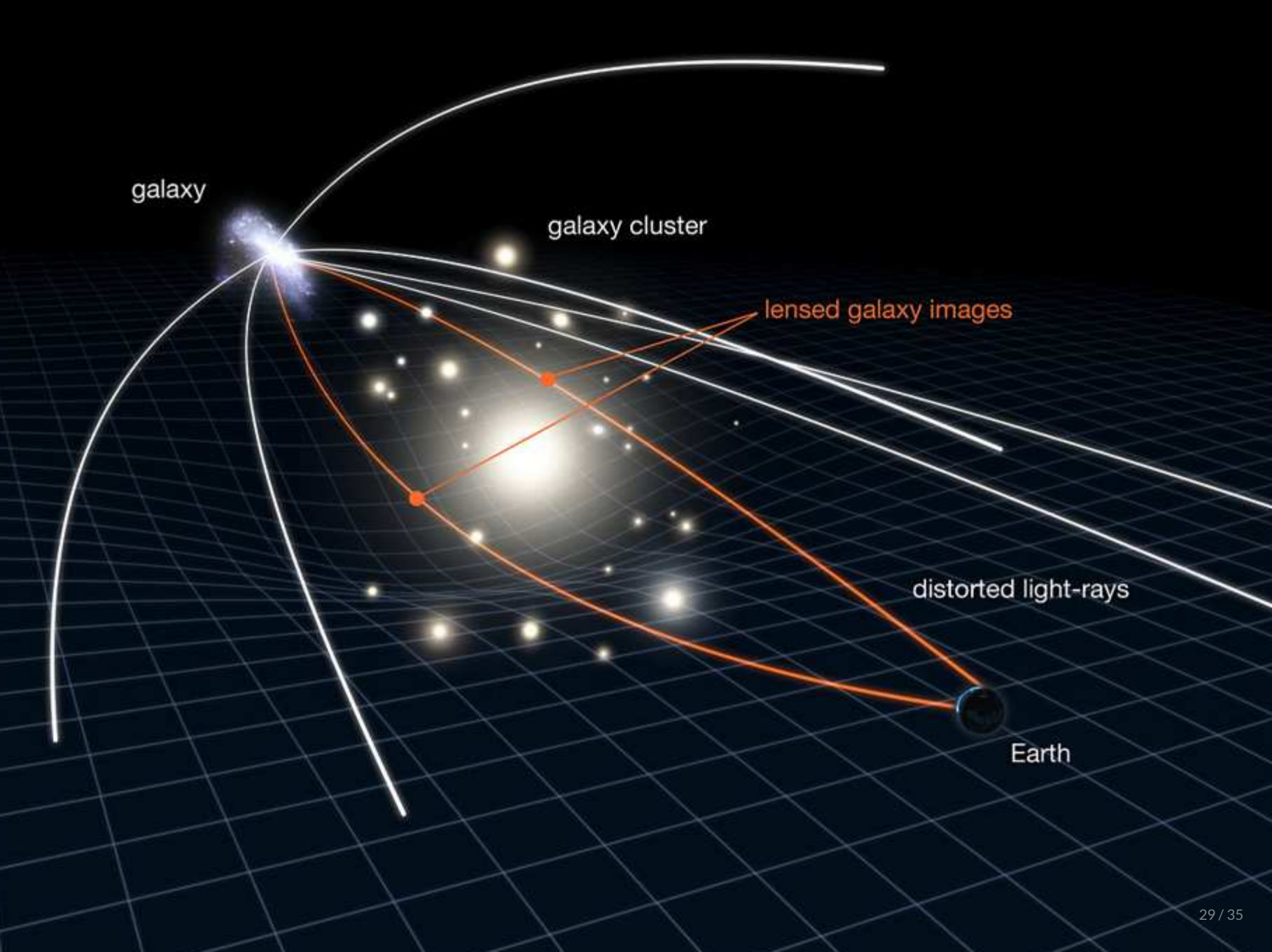


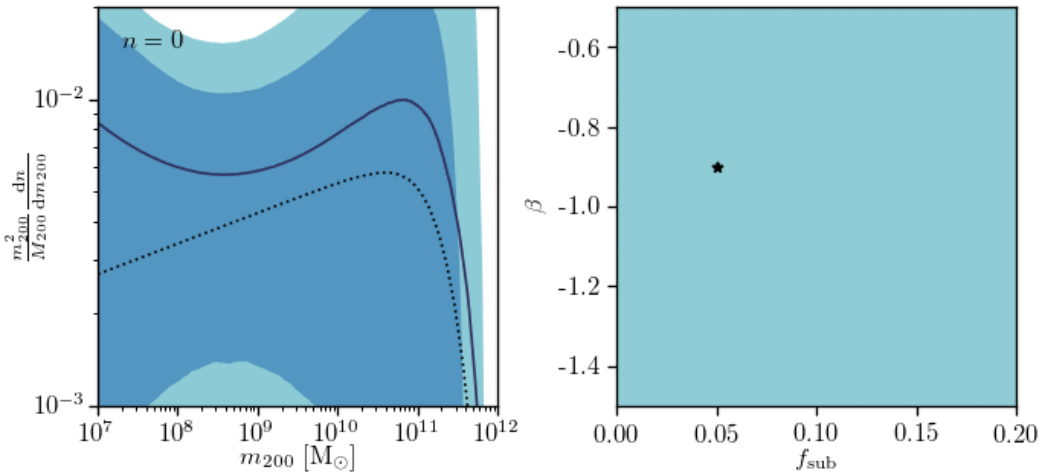


With enough training data, NRE gets the likelihood-ratio statistic right.

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

Palomar 5 (Pal5) stream
Pal5 was discovered in 2001 as the first thin stream formed from a globular cluster. Its current orbit takes it far over the galactic center.

Globular clusters
These hives typically hold 100,000 stars or fewer and give rise to long, thin streams.

Gap

Sun

Milky Way

GD1 stream
Discovered in 2006, GD1 is the longest known thin stream, stretching across more than half the northern sky. It contains a gap that could be the scar of a dark matter collision 500 million years ago.

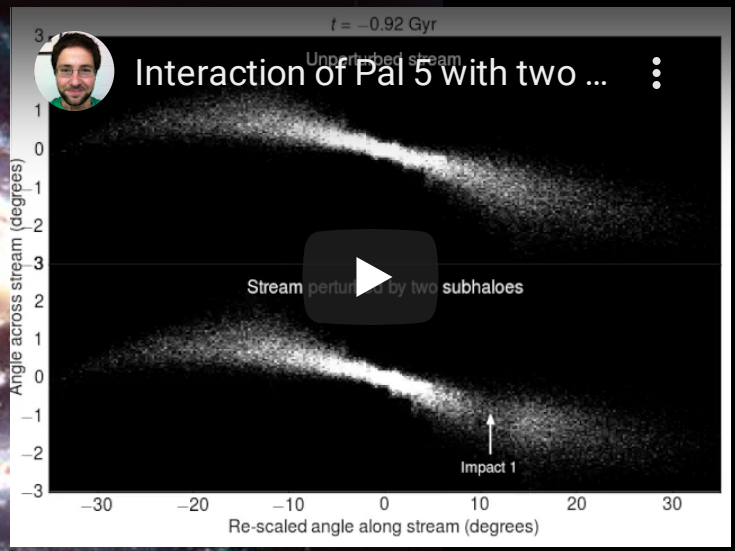
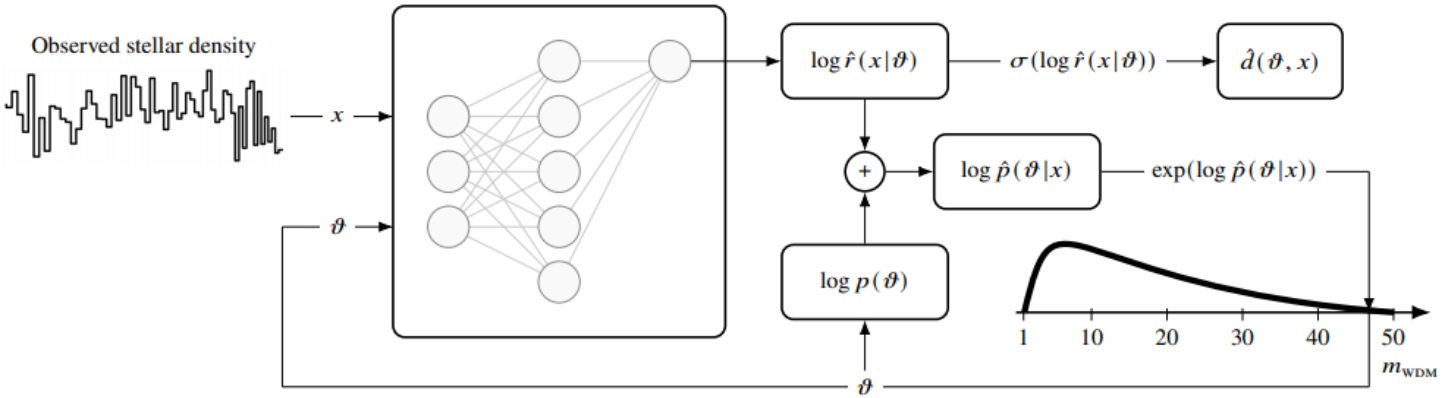
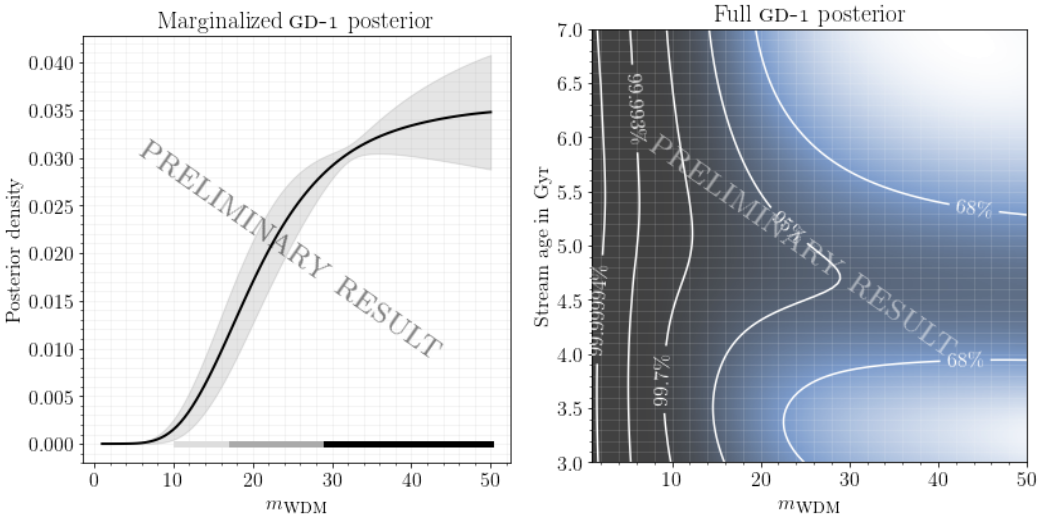


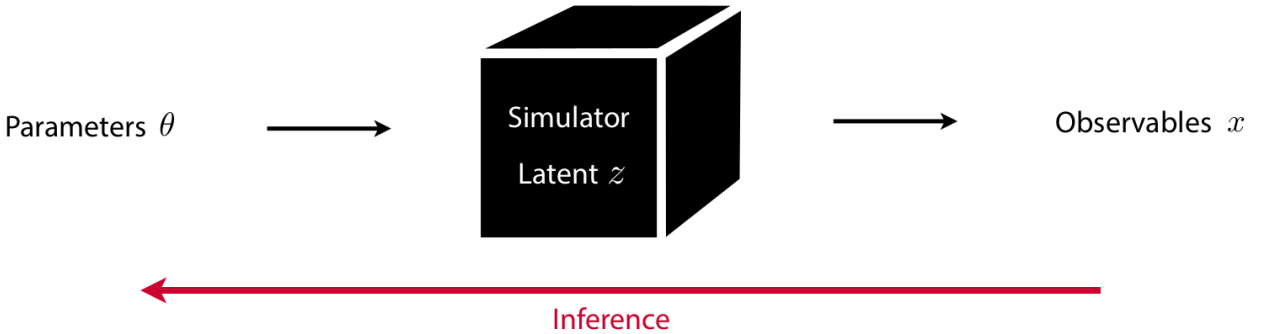
Image credits: C. Bickel/Science; D. Erkal.

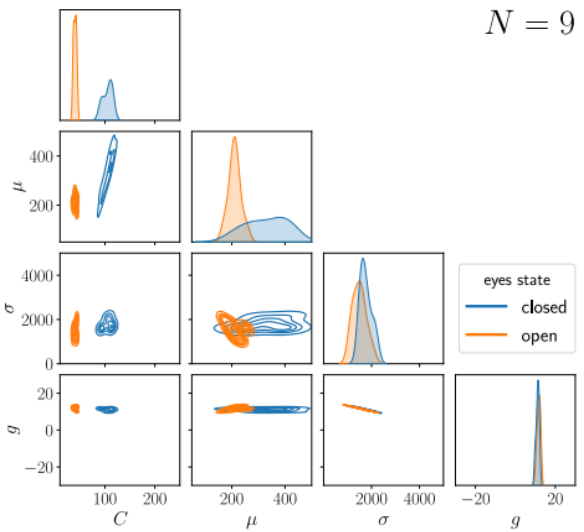


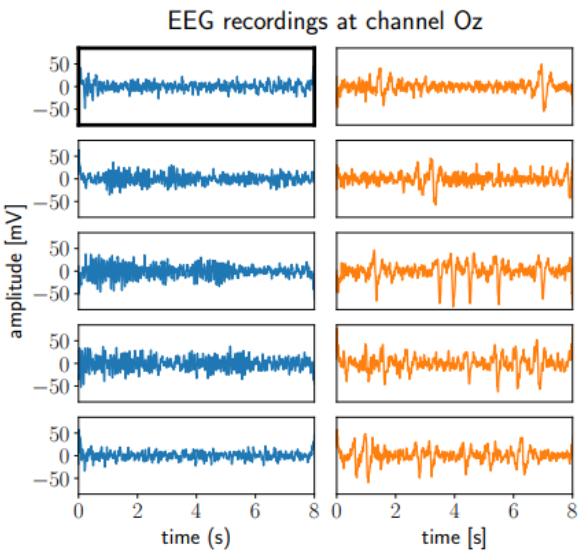


Preliminary results for GD-1 suggest a preference for CDM over WDM.

Case 4: Inverting a neural mass model (Jansen and Rit)







Summary

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

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

The end.