Likelihood-free inference

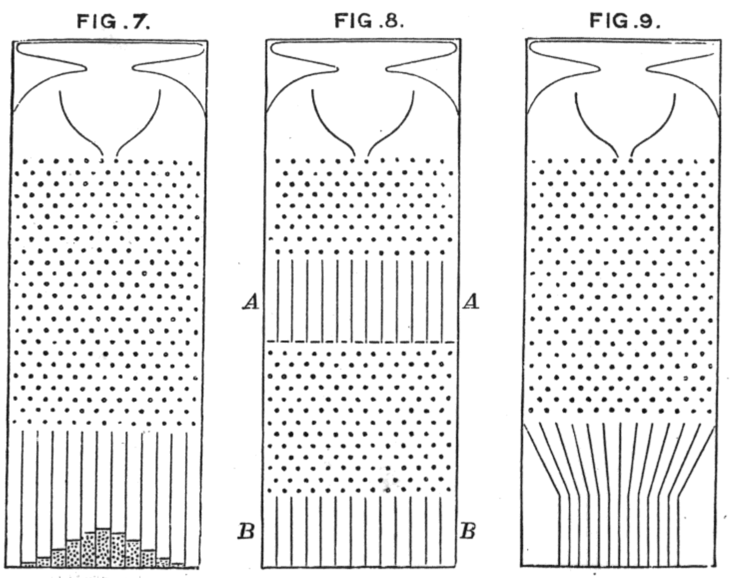
1st Terascale School of Machine Learning

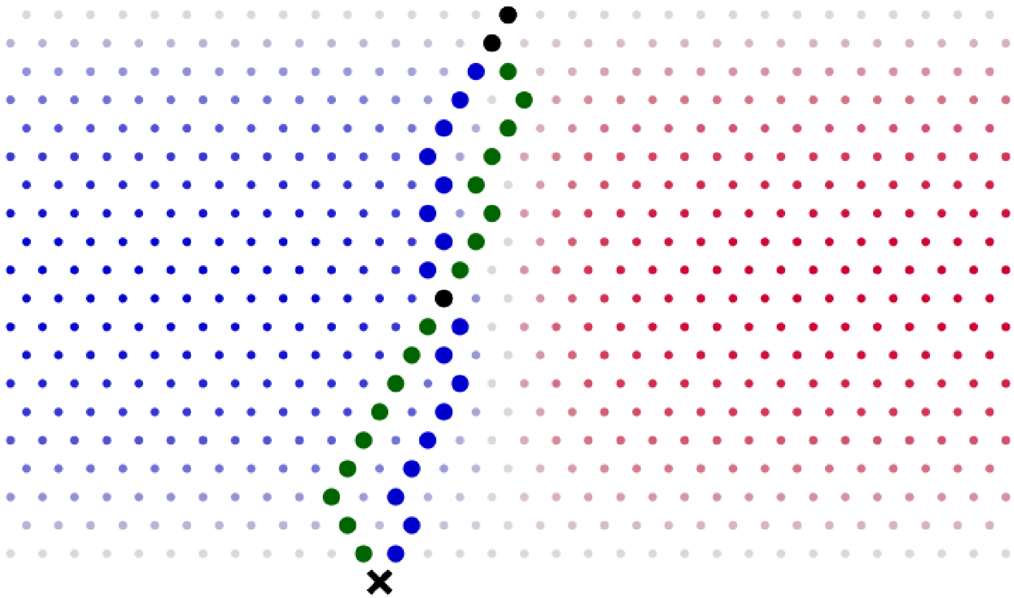
Gilles Louppe
g.louppe@uliege.be


GALTON BOARD
@physicsfun

Sir Francis Galton saw his bean machine as an analogy for the inheritance of genetic traits.

- The pinballs accumulate in a bell-shaped curve that is similar to the distribution of human heights.
- The puzzle of why human heights do not spread out from one generation to the next, as the balls would, led him to the discovery of "regression to the mean".





The probability of ending in bin x corresponds to the total probability of all the paths z from start to x.

$$p(x|\theta) = \int p(x, z|\theta) dz = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

Inference

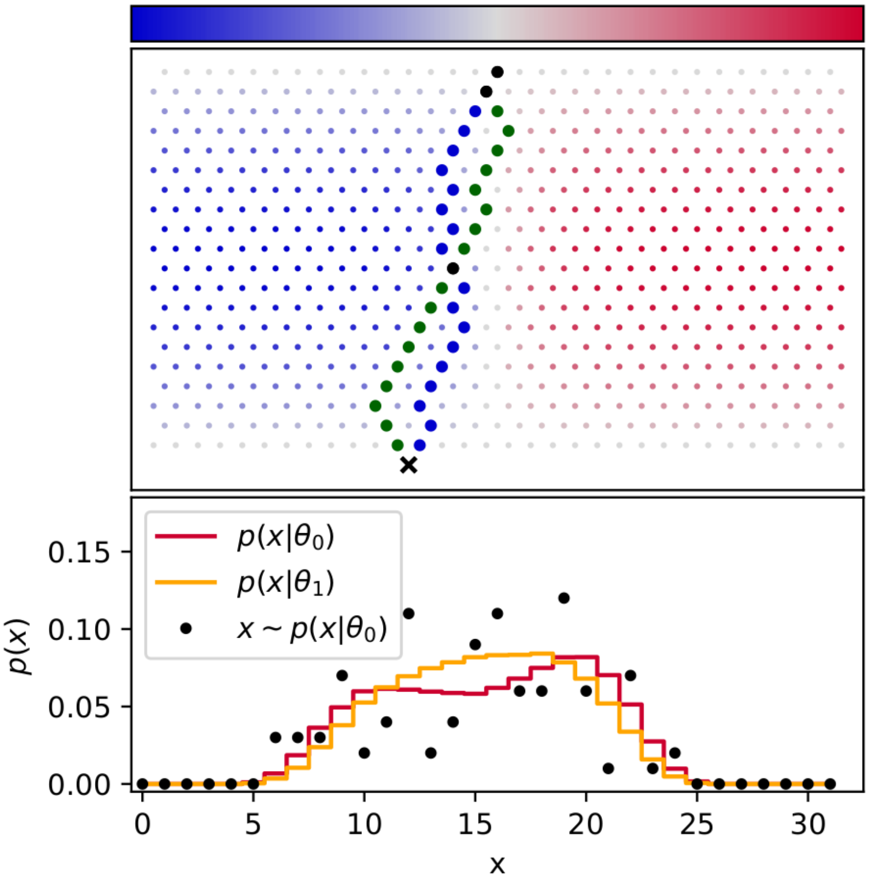
Given a set of realizations $\mathbf{d} = \{x_i\}$ at the bins, inference consists in determining the value of θ that best describes these observations.

For example, following the principle of maximum likelihood estimation, we have

$$\hat{\theta} = \arg\max_{\theta} \prod_{x_i \in \mathbf{d}} p(x_i|\theta).$$

In general, when $p(x_i|\theta)$ can be evaluated, this problem can be solved either analytically or using optimization algorithms.

What if we shift or remove some of the pins?



The probability of ending in bin x still corresponds to the cumulative probability of all the paths from start to x:

$$p(x|\theta) = \int p(x, z|\theta) dz$$

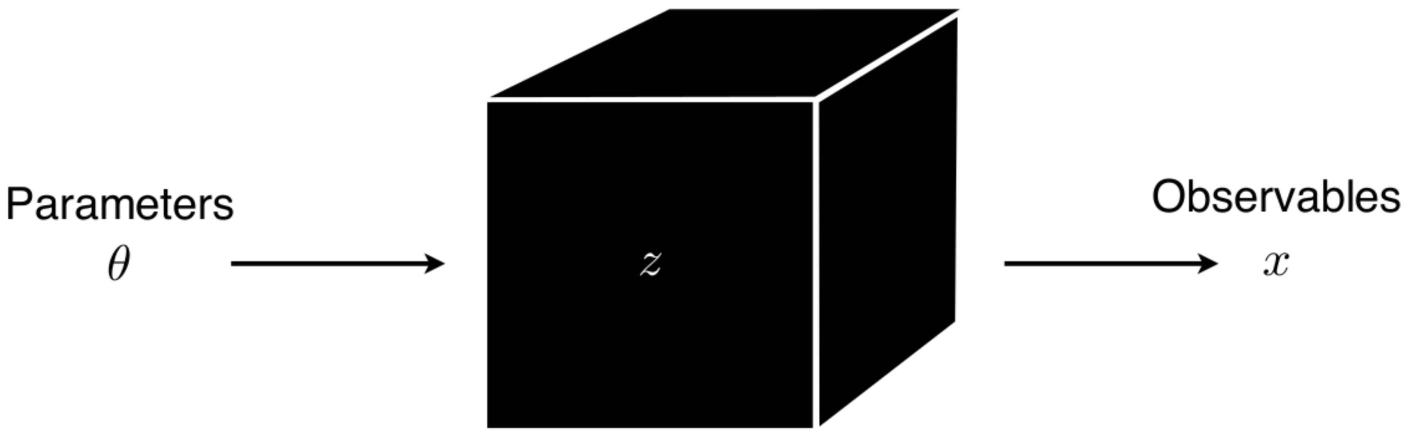
- But this integral can no longer be simplified analytically!
- As n grows larger, evaluating $p(x|\theta)$ becomes intractable since the number of paths grows combinatorially.
- Generating observations remains easy: drop the balls.

Since $p(x|\theta)$ cannot be evaluated, does this mean inference is no longer possible?

The Galton board is a metaphore of simulation-based science:

- the Galton board device is the equivalent of the scientific simulator
- x are observables
- θ are parameters of interest
- z are stochastic execution traces through the simulator

Inference in this context requires likelihood-free algorithms.



Prediction (simulation):

- Well-understood mechanistic model
- Simulator can generate samples

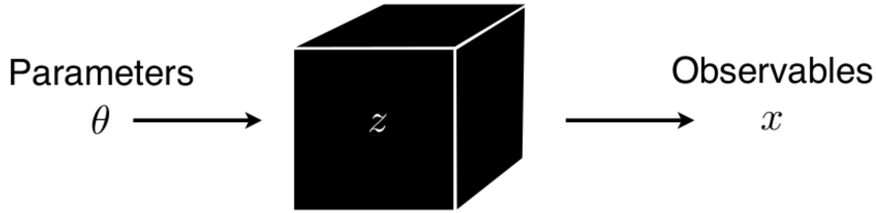
Inference:

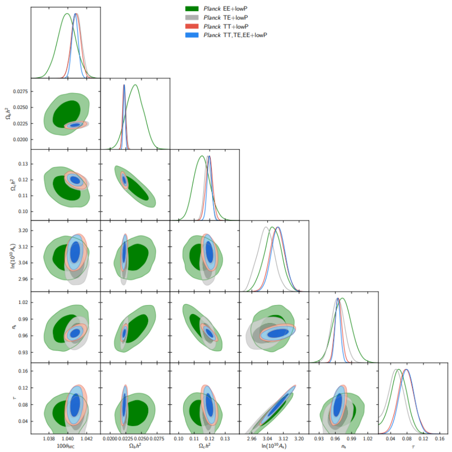
- Likelihood function $p(x|\theta)$ is intractable
- Goal: estimator $\hat{p}(x|\theta)$

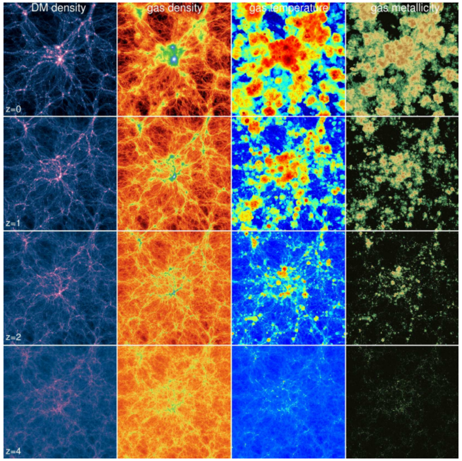
Credits: Johann Brehmer

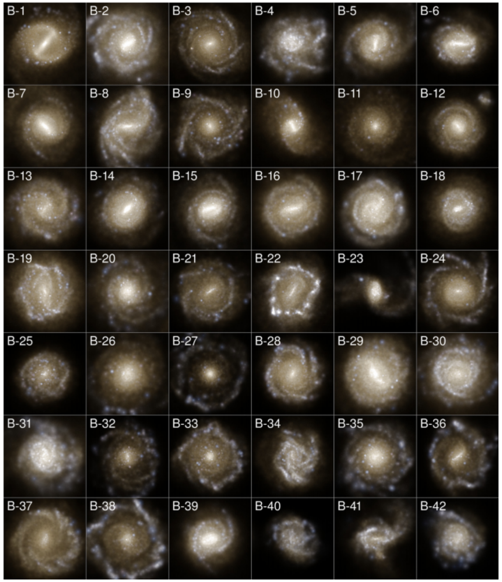
Applications

Cosmological N-body simulations



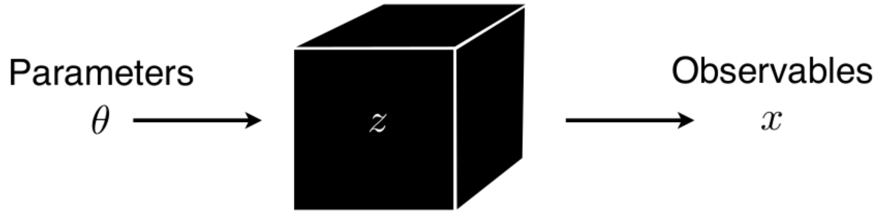


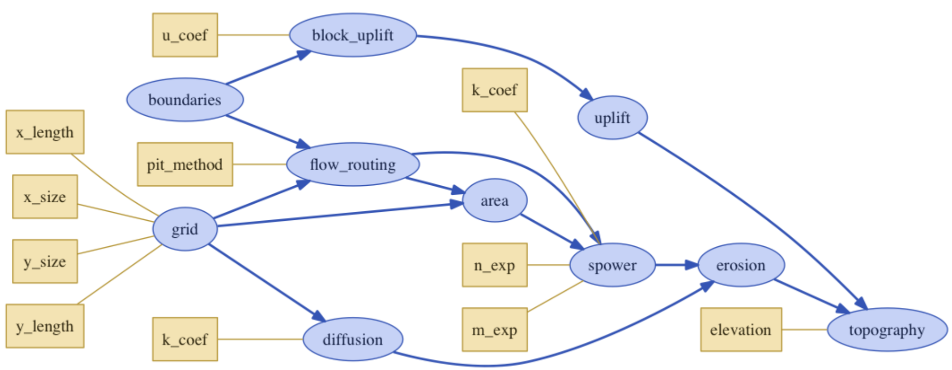


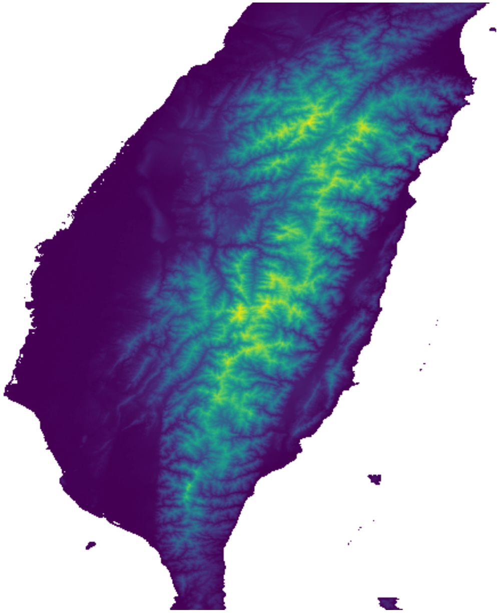


Refs: Planck Collaboration, 2015 (arXiv:1502.01589); Vogelsberger et al, 2014 (arXiv:1405.2921)

Computational topography

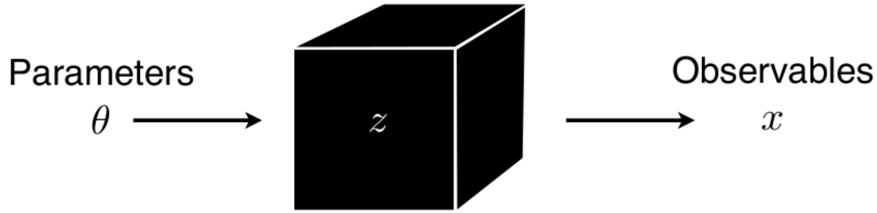


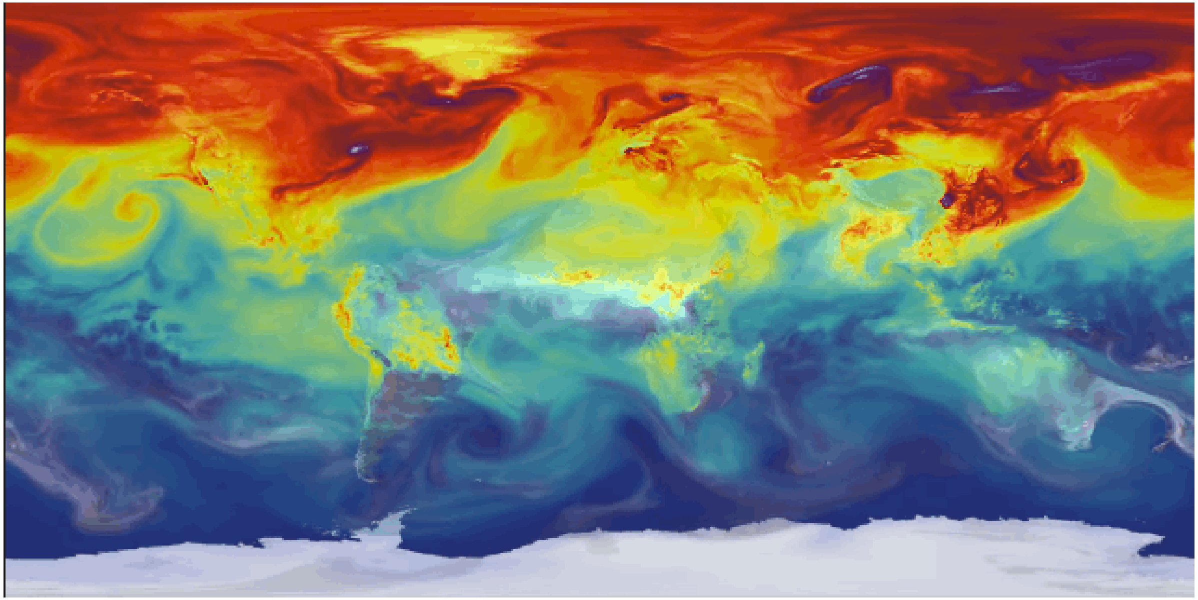




Refs: Benoit Bovy (xarray-simlab)

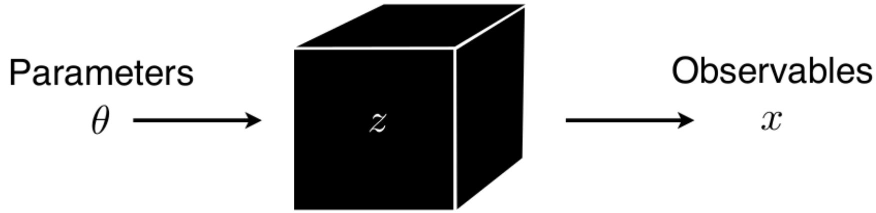
Climatology

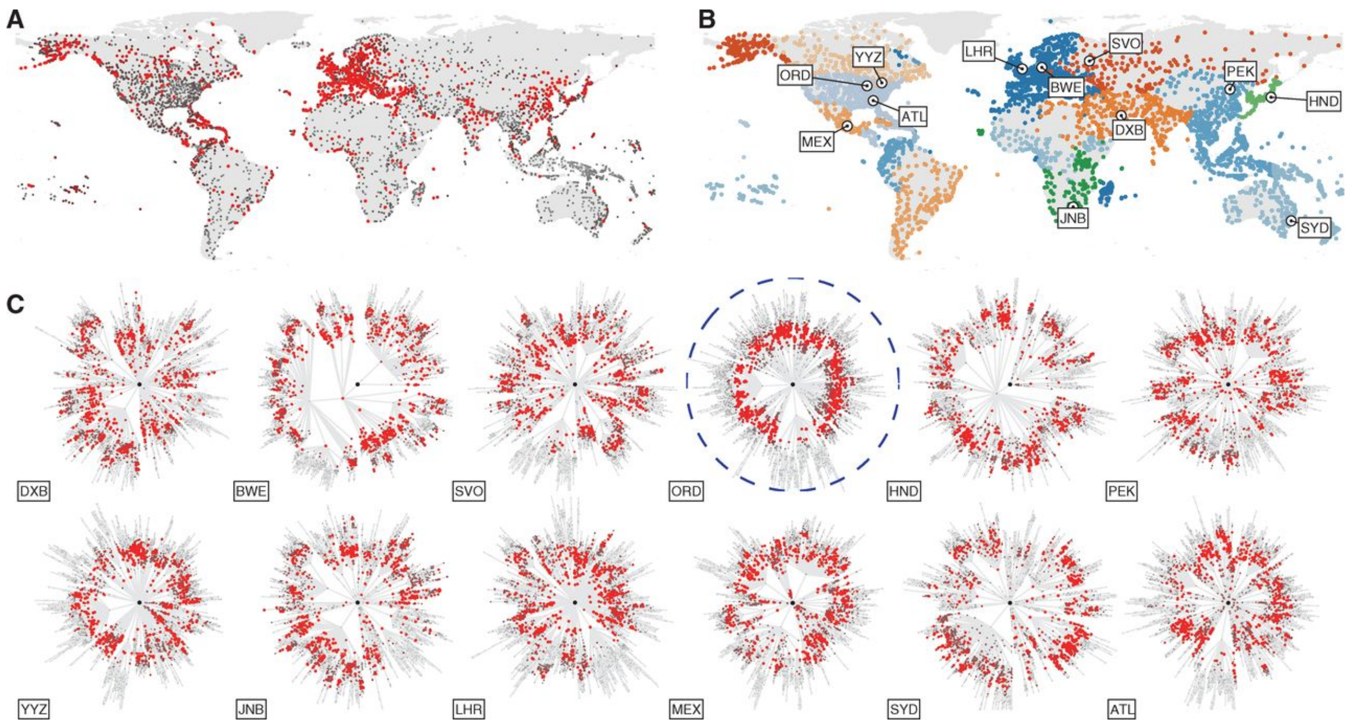




Refs: NASA's Goddard Space Flight Center / B. Putman, 2014 (press release)

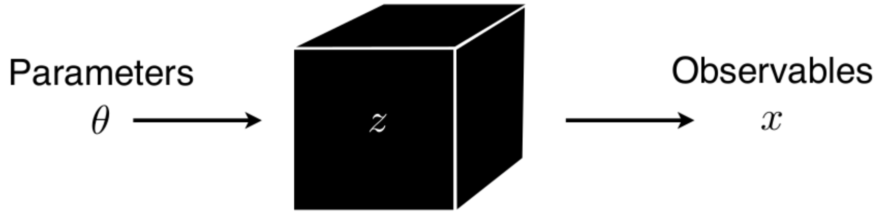
Epidemiology

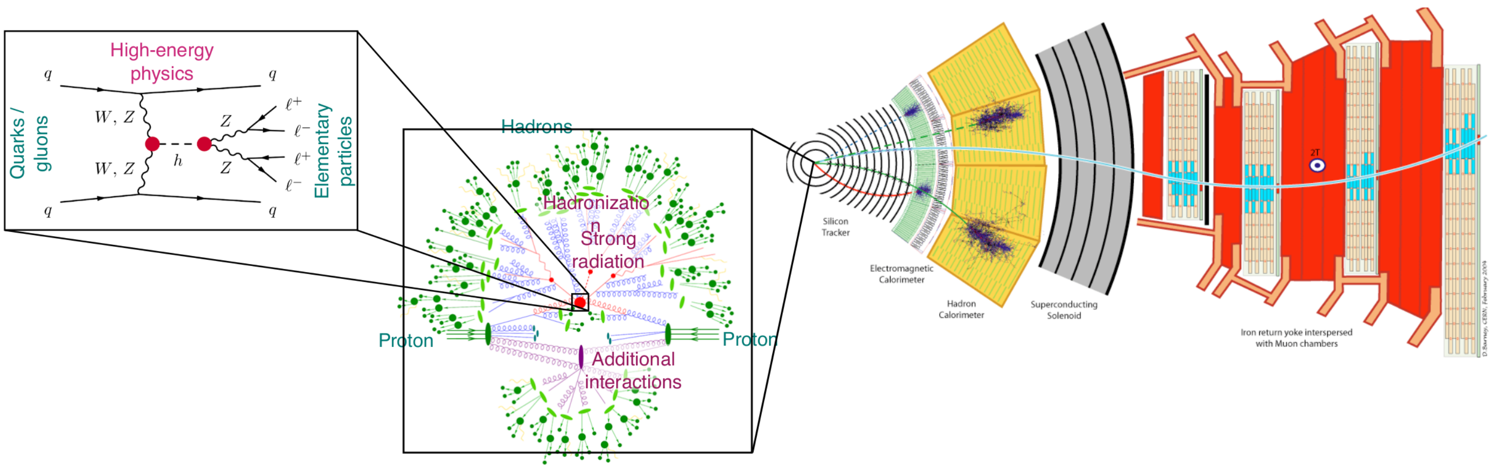




Refs: Brockmann and Helbing, 2013 (doi:10.1126/science.1245200)

Particle physics



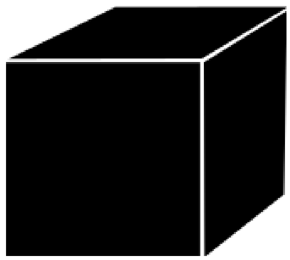


The Galton board of particle physics

Algorithms

Treat the simulator as a black box

- Histograms of observables
- Approximate Bayesian computation
- Neural density (ratio) estimation
- Adversarial Variational Optimization

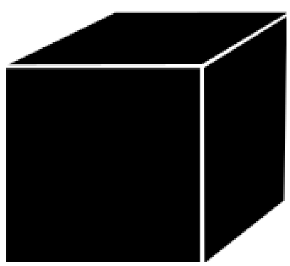


Use latent structure

- Matrix Element Method
- Optimal Observables
- Shower deconstruction, event Deconstruction
- Mining gold from the simulator
- Probabilistic programming

Treat the simulator as a black box

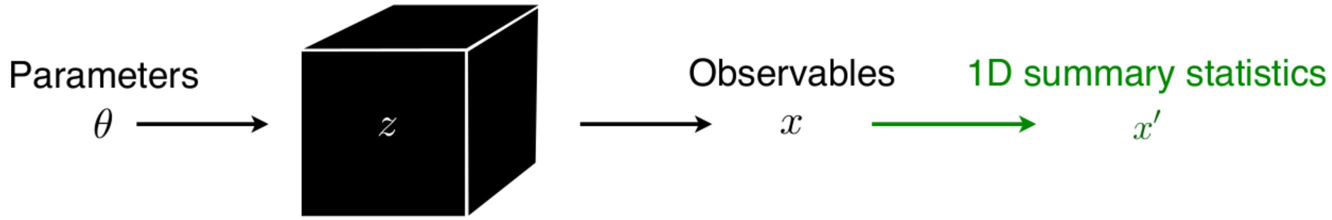
- Histograms of observables
- Approximate Bayesian computation
- Neural density (ratio) estimation
- Adversarial Variational Optimization



Use latent structure

- Matrix Element Method
- Optimal Observables
- Shower deconstruction, event Deconstruction
- Mining gold from the simulator
- Probabilistic programming

The physicist's way

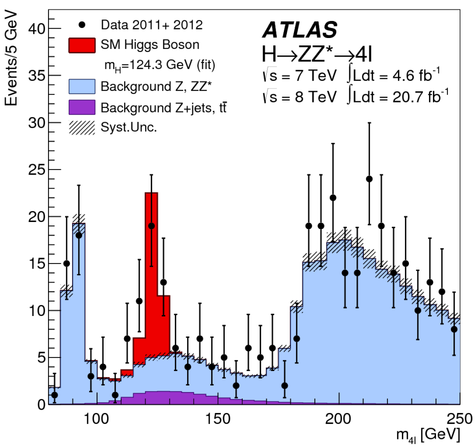


Define a projection function $s : \mathcal{X} \to \mathbb{R}$ mapping observables x to a summary statistics $x' = s(x)$.

Then, approximate the likelihood $p(x|\theta)$ as

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

where $p(x'|\theta)$ can be estimated by running the simulator for different parameter values θ and filling histograms.

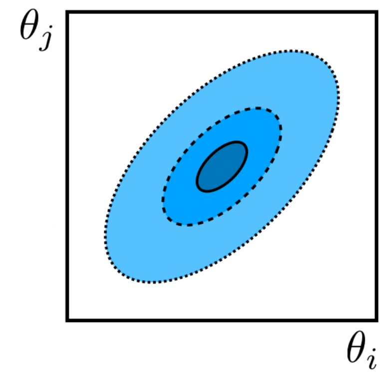


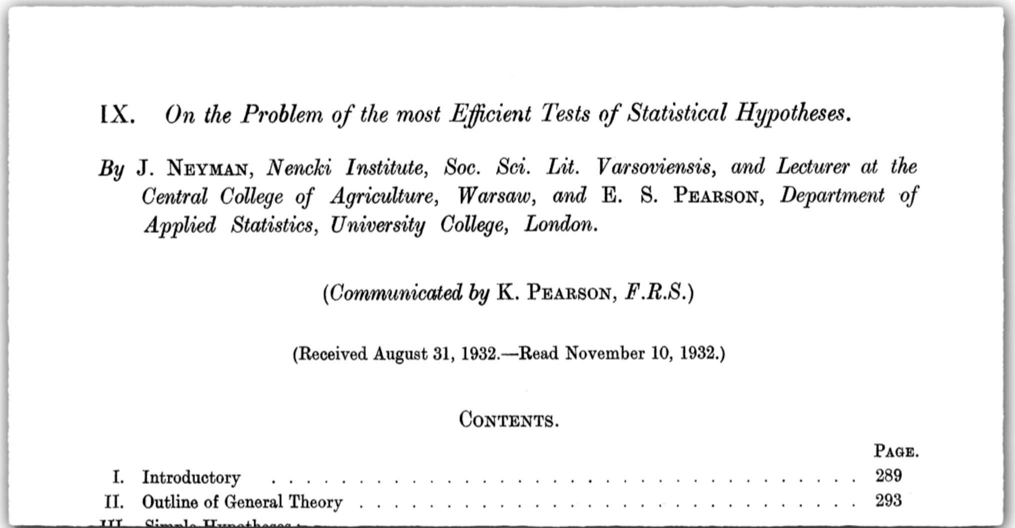
Hypothesis testing

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.)

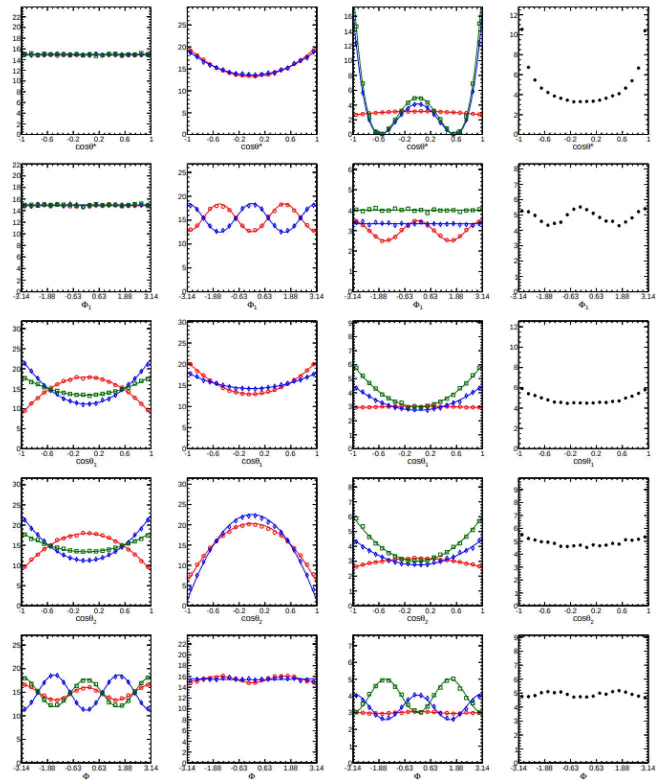
CONTENTS.

In the likelihood-free setup, the ratio is difficult to compute. However, using the approximate likelihood we can define

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} \approx \frac{\hat{p}(x|\theta_0)}{\hat{p}(x|\theta_1)}$$

This methodology has worked great for physicists for the last 20-30 years, but ...

- Choosing the projection s is difficult and problem-dependent.
- Often there is no single good variable: compressing to any x' loses information.
- Ideally: analyse high-dimensional x', including all correlations.



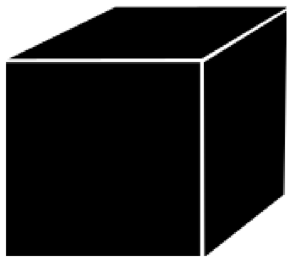
Unfortunately, because of the curse of dimensionality, filling high-dimensional histograms is not tractable.

Who you gonna call? Machine learning!

Refs: Bolognesi et al, 2012 (arXiv:1208.4018)

Treat the simulator as a black box

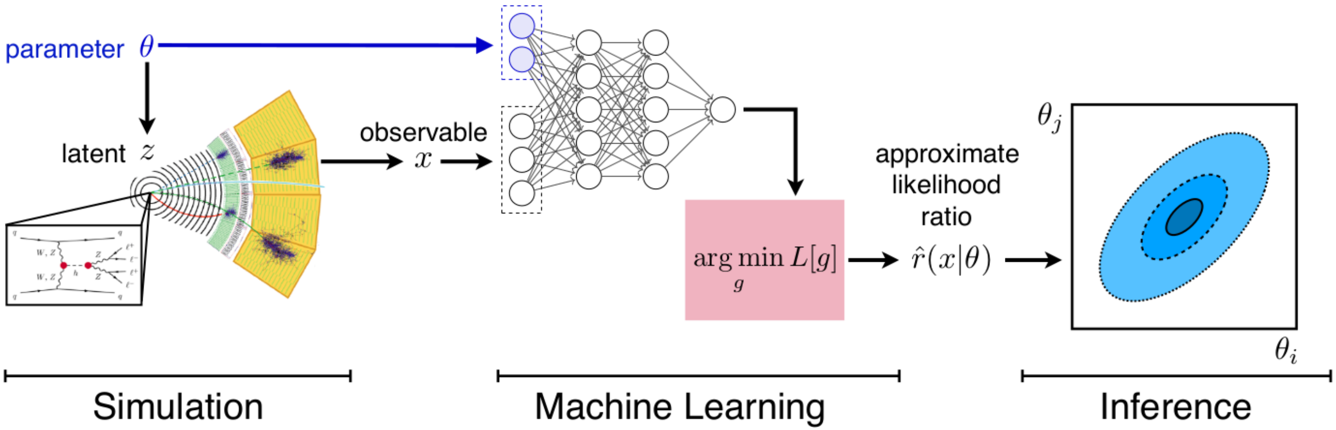
- Histograms of observables
- Approximate Bayesian computation
- Neural density (ratio) estimation
- Adversarial Variational Optimization



Use latent structure

- Matrix Element Method
- Optimal Observables
- Shower deconstruction, event Deconstruction
- Mining gold from the simulator
- Probabilistic programming

Carl



Key insights

- The likelihood ratio is sufficient for maximum likelihood estimation.
- Evaluating the likelihood ratio does not require evaluating the individual likelihoods.
- Machine learning can be used to learn the likelihood ratio.

Refs: Cranmer et al, 2016 (arXiv:1506.02169)

Theorem. The likelihood ratio is invariant under the change of variable $U = s(X)$, provided $s(x)$ is monotonic with $r(x)$.

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

- Note that the equality is strict.
- No information relevant for determining the ratio is lost.
- Although information about x may be lost through s.

Supervised learning provides a way to automatically construct s:

- Let us consider a binary classifier $\hat{s}$ (e.g., a neural network) trained to distinguish $x \sim p(x|\theta_0)$ from $x \sim p(x|\theta_1)$.
- $\hat{s}$ is trained by minimizing the cross-entropy loss

$$\begin{aligned} L_{XE}[\hat{s}] = -\mathbb{E}_{p(x|\theta)\pi(\theta)}[&1(\theta = \theta_0) \log \hat{s}(x) + \\ &1(\theta = \theta_1) \log(1 - \hat{s}(x))] \end{aligned}$$

The solution $\hat{s}$ found after training approximates the optimal classifier

$$\hat{s}(x) \approx s^*(x) = \frac{p(x|\theta_1)}{p(x|\theta_0) + p(x|\theta_1)},$$

which is monotonic with r.

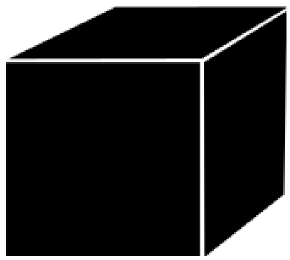
Therefore,

$$r(x|\theta_0, \theta_1) \approx \hat{r}(x|\theta_0, \theta_1) = \frac{1 - \hat{s}(x)}{\hat{s}(x)}$$

That is, supervised classification is equivalent to likelihood ratio estimation and can therefore be used for MLE inference.

Treat the simulator as a black box

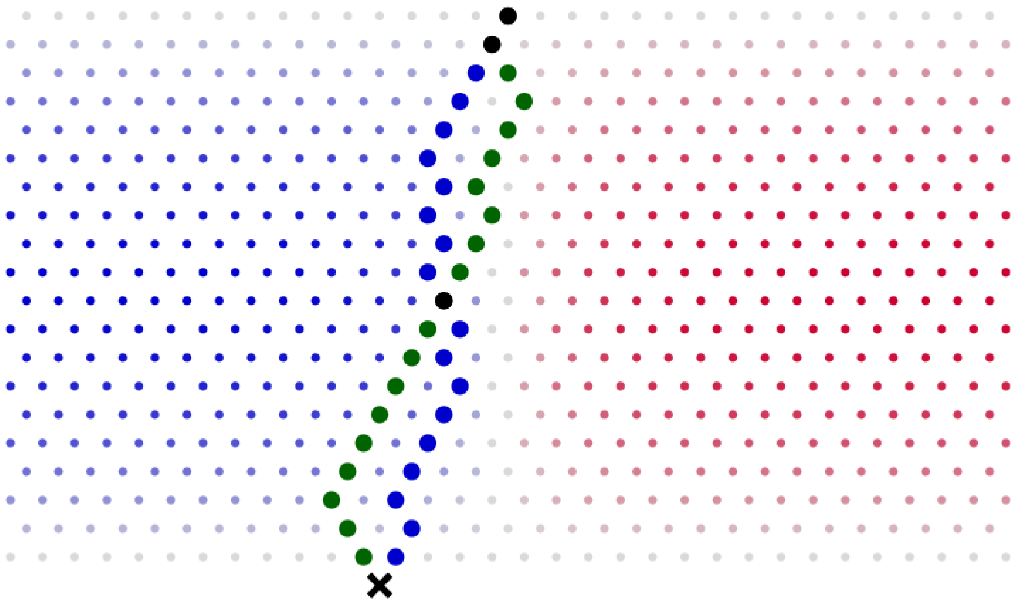
- Histograms of observables
- Approximate Bayesian computation
- Neural density (ratio) estimation
- Adversarial Variational Optimization



Use latent structure

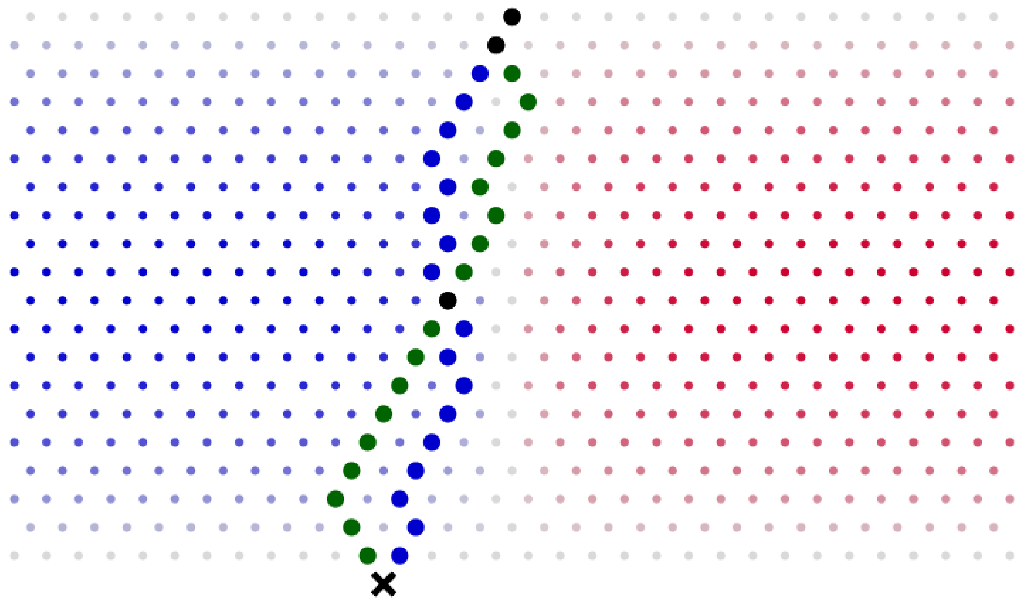
- Matrix Element Method
- Optimal Observables
- Shower deconstruction, event Deconstruction
- Mining gold from the simulator
- Probabilistic programming

Mining gold from simulators



$p(x|\theta)$ is usually intractable.

What about $p(x, z|\theta)$?



$$
\begin{aligned}
p(x, z|\theta) &= p(z_1|\theta)p(z_2|z_1, \theta)\dots p(z_T|z_{<T}, \theta)p(x|z_{\leq T}, \theta) \\
&= p(z_1|\theta)p(z_2|\theta)\dots p(z_T|\theta)p(x|z_T) \\
&= p(x|z_T)\prod_t \theta^{z_t}(1-\theta)^{1-z_t}
\end{aligned}
$$

This can be computed as the ball falls down the board!

As the trajectory $z_1, ..., z_T$ and the observable x are emitted, it is often possible:

- to calculate the joint likelihood $p(x, z|\theta)$;
- to calculate the joint likelihood ratio $r(x, z|\theta_0, \theta_1)$;
- to calculate the joint score $t(x, z|\theta_0) = \nabla_\theta \log p(x, z|\theta)\big|_{\theta_0}$.

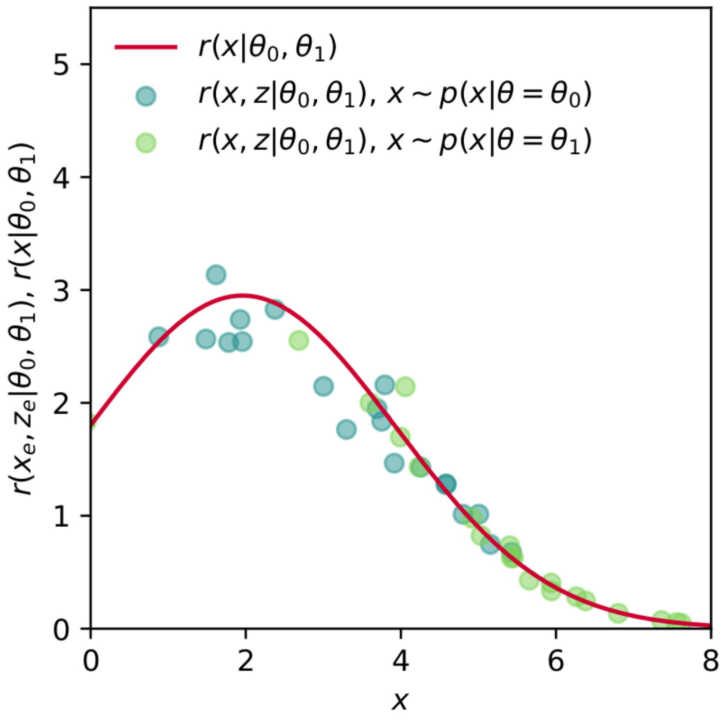
We call this process mining gold from your simulator!

Observe that the joint likelihood ratios

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

are scattered around $r(x|\theta_0, \theta_1)$.

Can we use them to approximate $r(x|\theta_0, \theta_1)$?



Key insights

Consider the squared error of a function $\hat{g}(x)$ that only depends on x, but is trying to approximate a function $g(x, z)$ that also depends on the latent z:

$$L_{MSE} = \mathbb{E}_{p(x,z|\theta)}\left[\left(g(x, z) - \hat{g}(x)\right)^2\right].$$

Via calculus of variations, we find that the function $g^*(x)$ that extremizes $L_{MSE}[g]$ is given by

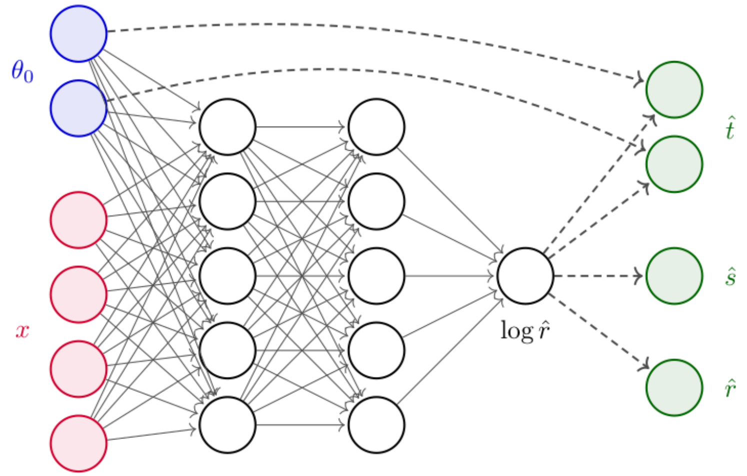
$$\begin{aligned} g^*(x) &= \frac{1}{p(x|\theta)} \int p(x, z|\theta) g(x, z) dz \\ &= \mathbb{E}_{p(z|x,\theta)}\left[g(x, z)\right] \end{aligned}$$

Therefore, by identifying the $g(x, z)$ with the joint likelihood ratio $r(x, z|\theta_0, \theta_1)$ and θ with θ_1, we define

$$L_r = \mathbb{E}_{p(x,z|\theta_1)} \left[(r(x, z|\theta_0, \theta_1) - \hat{r}(x))^2 \right],$$

which is minimized by

$$\begin{aligned} r^*(x) &= \frac{1}{p(x|\theta_1)} \int p(x, z|\theta_1) \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)} dz \\ &= \frac{p(x|\theta_0)}{p(x|\theta_1)} \\ &= r(x|\theta_0, \theta_1). \end{aligned}$$



How does one find r^*?

$$r^*(x|\theta_0, \theta_1) = \arg\min_{\hat{r}} L_r[\hat{r}]$$

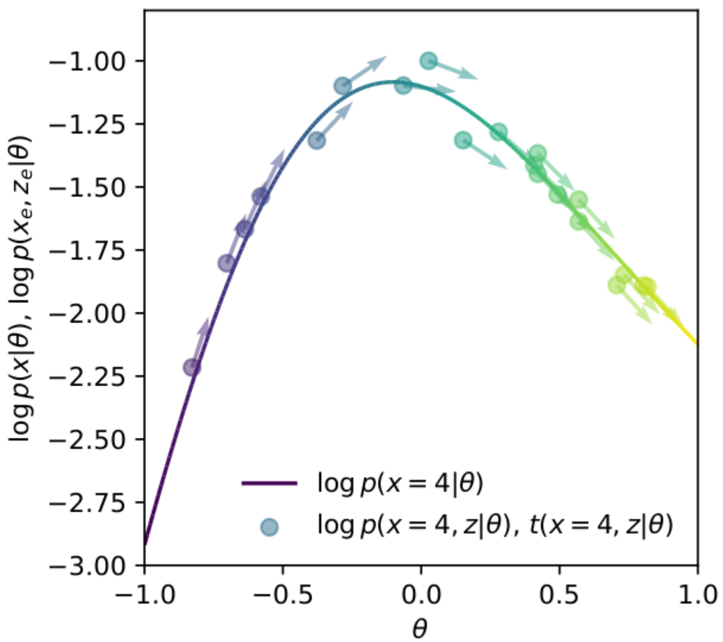
Minimizing functionals is exactly what machine learning does. In our case,

- $\hat{r}$ are neural networks (or the parameters thereof);
- L_r is the loss function;
- minimization is carried out using stochastic gradient descent from the data extracted from the simulator.

Similarly, we can mine the simulator to extract the joint score

$$t(x, z|\theta_0) = \nabla_\theta \log p(x, z|\theta)\big|_{\theta_0},$$

which indicates how much more or less likely x, z would be if one changed θ_0.



Using the same trick, by identifying $g(x, z)$ with the joint score $t(x, z|\theta_0)$ and θ with θ_0, we define

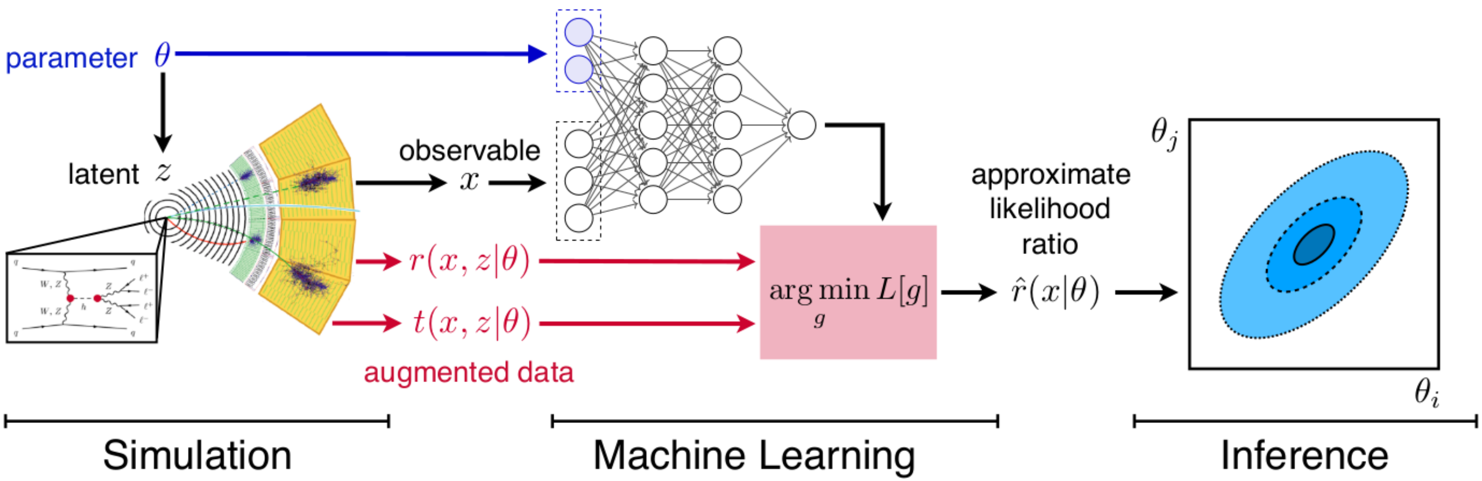
$$L_t = \mathbb{E}_{p(x,z|\theta_0)}\left[(t(x, z|\theta_0) - \hat{t}\,(x))^2\right],$$

which is minimized by

$$\begin{aligned}
t^*(x) &= \frac{1}{p(x|\theta_0)} \int p(x, z|\theta_0)(\nabla_\theta \log p(x, z|\theta)\big|_{\theta_0})dz \\
&= \frac{1}{p(x|\theta_0)} \int p(x, z|\theta_0) \frac{\nabla_\theta p(x, z|\theta)\big|_{\theta_0}}{p(x, z|\theta_0)} dz \\
&= \frac{\nabla_\theta p(x|\theta)\big|_{\theta_0}}{p(x|\theta_0)} \\
&= t(x|\theta_0).
\end{aligned}$$

RASCAL

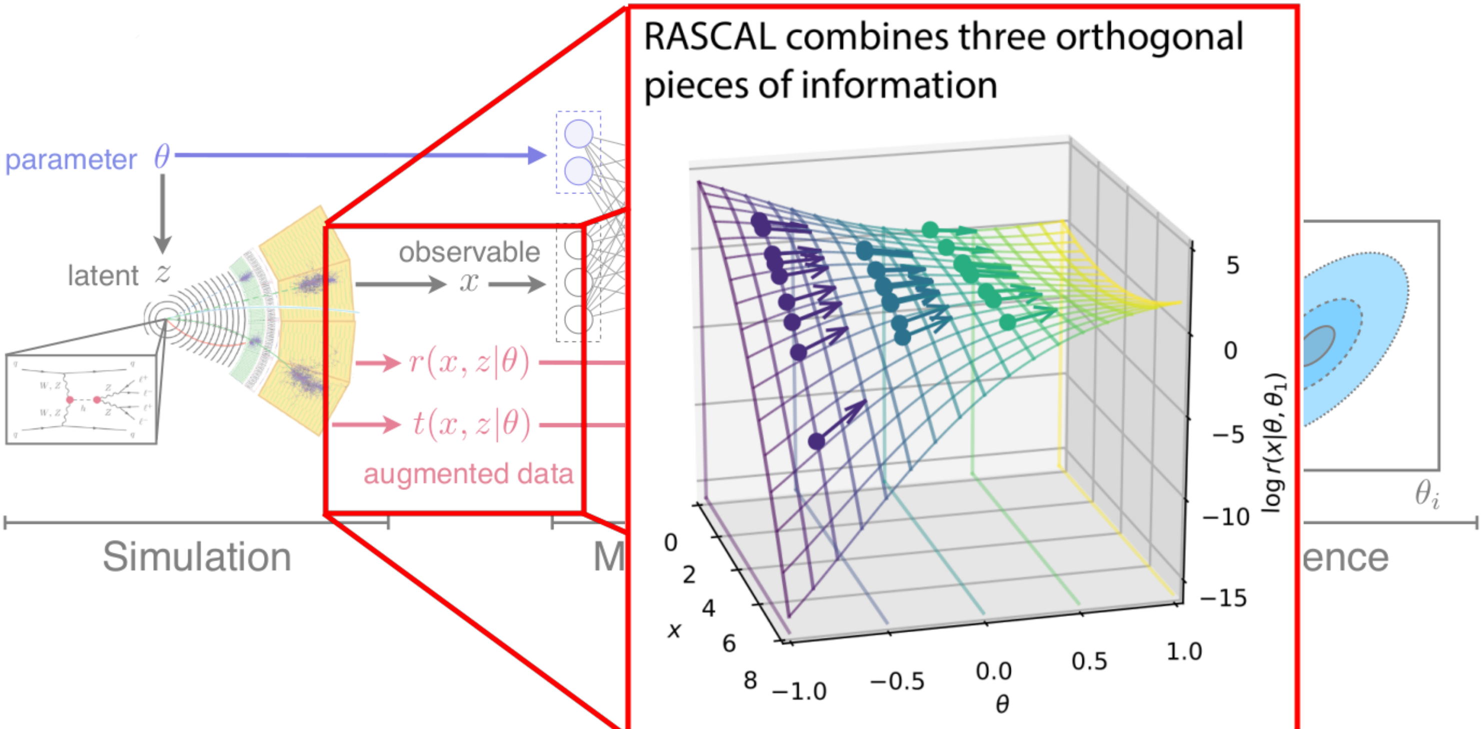
$$L_{RASCAL} = L_r + L_t$$



Refs: Brehmer et al, 2018 (arXiv:1805.12244)

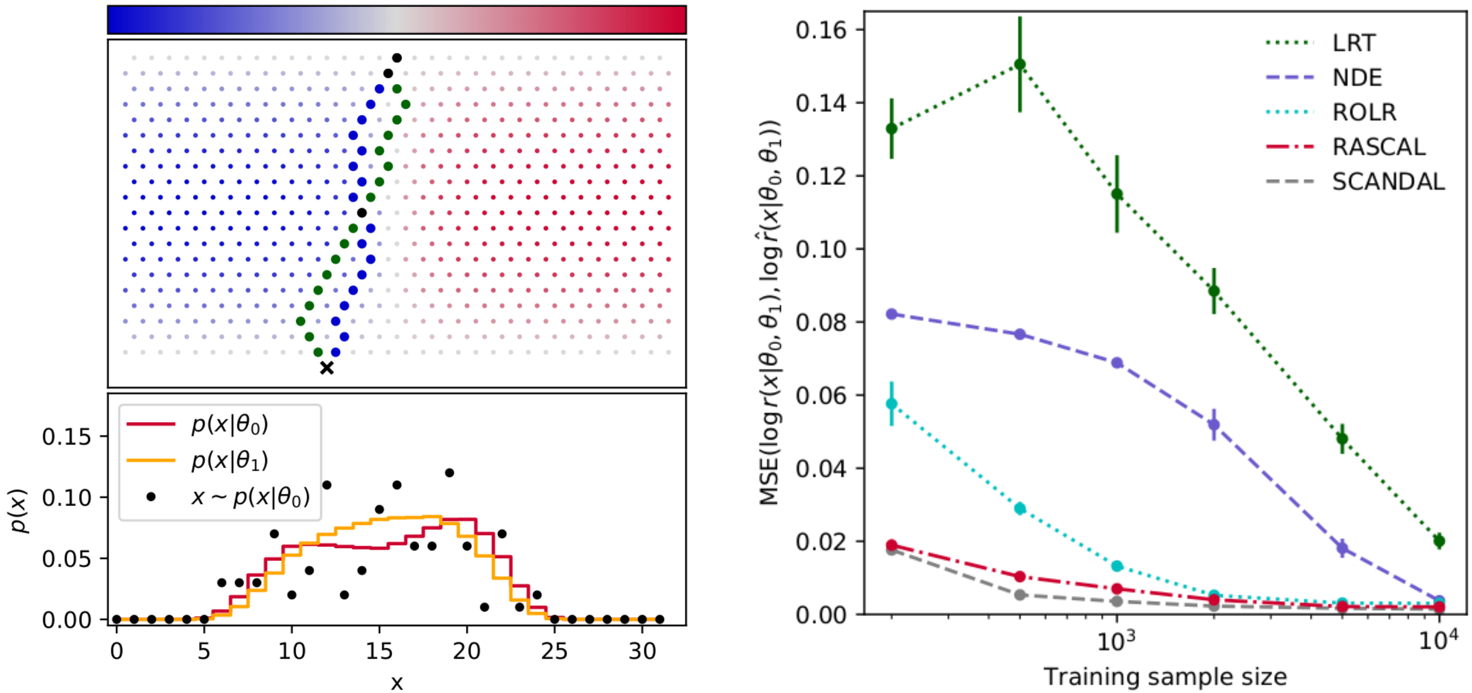
Rascal

$$L_{RASCAL} = L_r + L_t$$



Refs: Brehmer et al, 2018 (arXiv:1805.12244)

Effective inference



Toy experiment on the Galton board.

Constraining Effective Field Theories, effectively

LHC processes

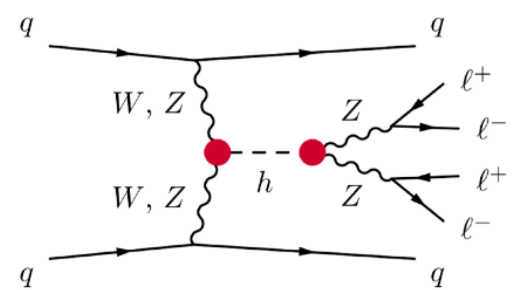
Latent variables

Parameters of interest

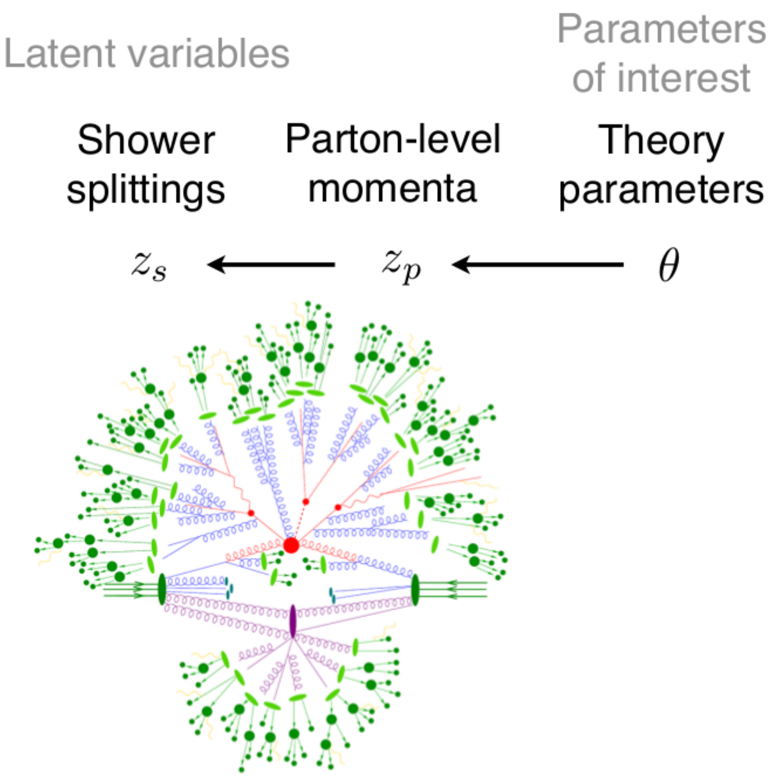
Parton-level momenta

Theory parameters

$$z_p \longleftarrow \theta$$



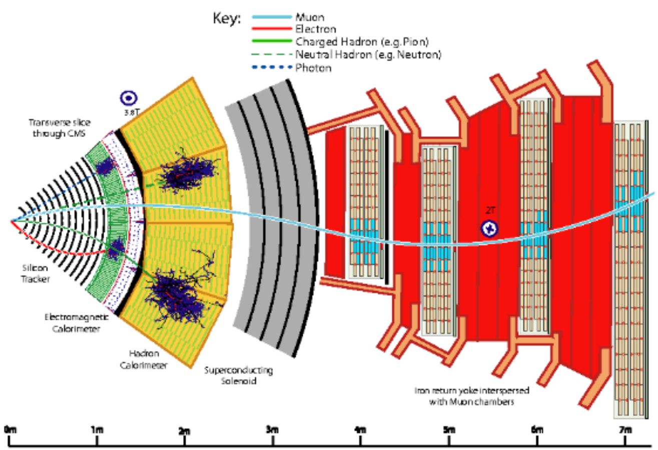
LHC processes



LHC processes

	Latent variables		Parameters of interest
Detector interactions	Shower splittings	Parton-level momenta	Theory parameters

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

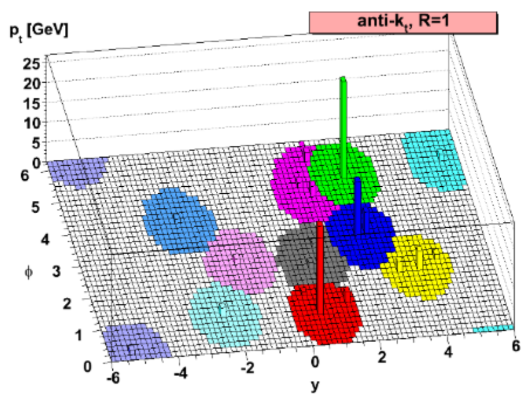


Credits: Johann Brehmer

LHC processes

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]

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

Key insights

- The distribution of parton-level momenta

$$p(z_p|\theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma(\theta)}{dz_p},$$

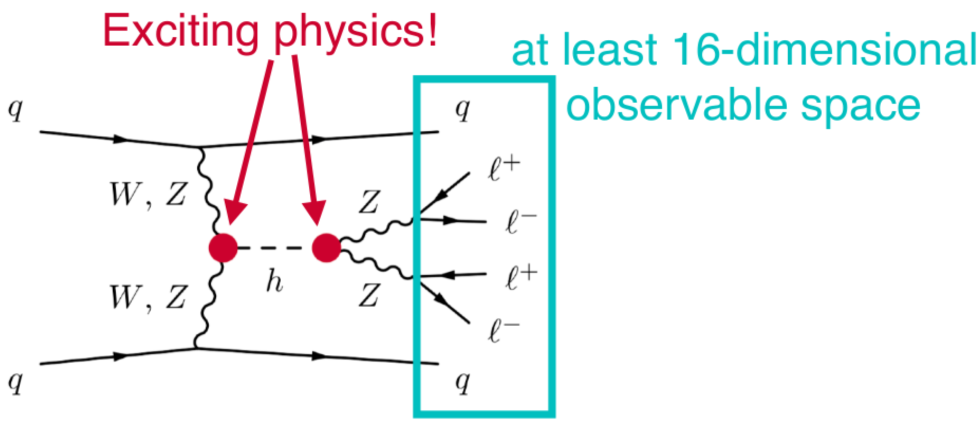
  where $\sigma(\theta)$ and $\frac{d\sigma(\theta)}{dz_p}$ are the total and differential cross sections, is tractable.

- Downstream processes $p(z_s|z_p)$, $p(z_d|z_s)$ and $p(x|z_d)$ do not depend on θ.

$\Rightarrow$ This implies that both $r(x, z|\theta_0, \theta_1)$ and $t(x, z|\theta_0)$ can be mined. E.g.,

$$r(x, z|\theta_0, \theta_1) = \frac{p(z_p|\theta_0)}{p(z_p|\theta_1)} \frac{p(z_s|z_p)}{p(z_s|z_p)} \frac{p(z_d|z_s)}{p(z_d|z_s)} \frac{p(x|z_d)}{p(x|z_d)} = \frac{p(z_p|\theta_0)}{p(z_p|\theta_1)}$$

Proof of concept

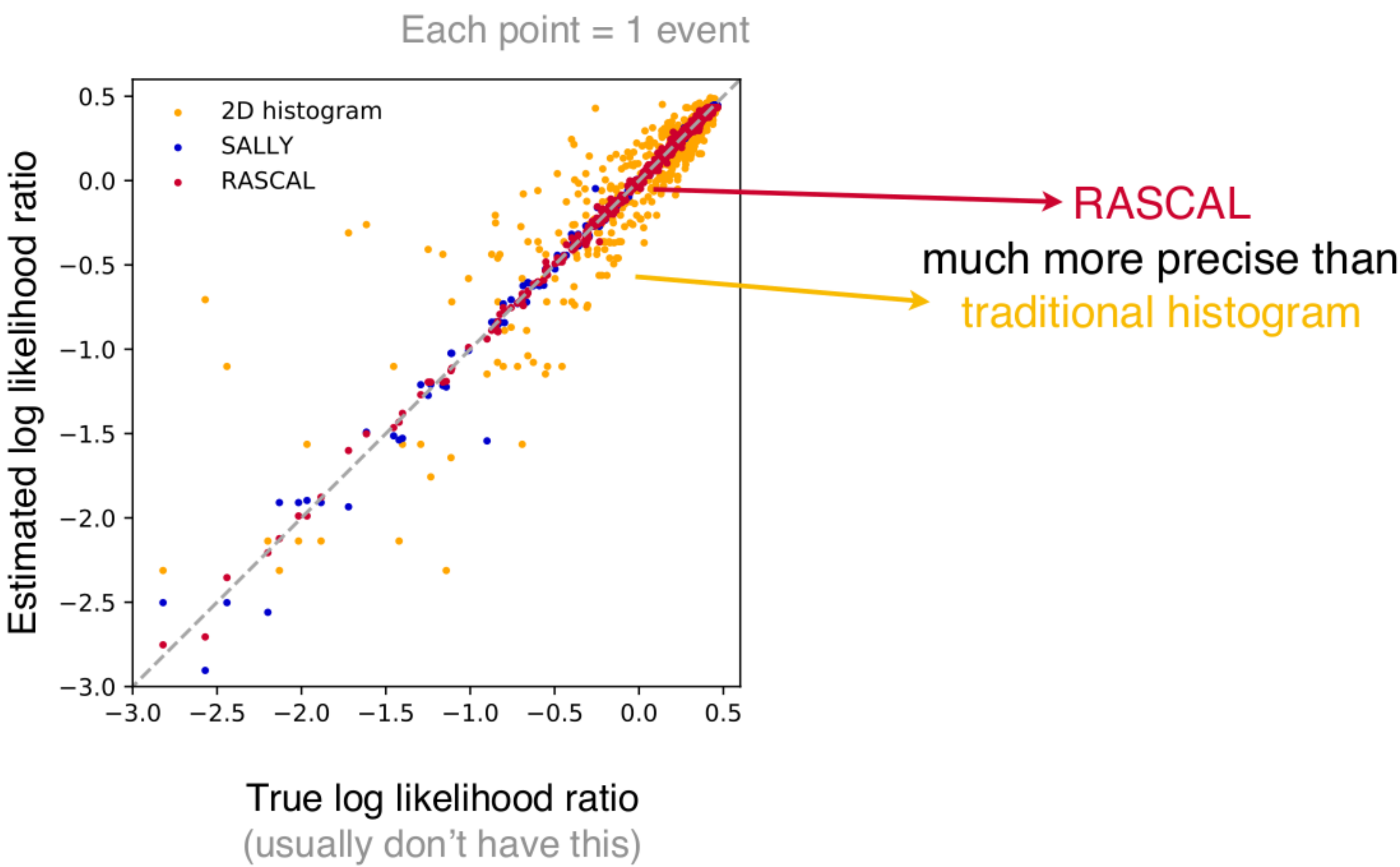


Higgs production in weak boson fusion

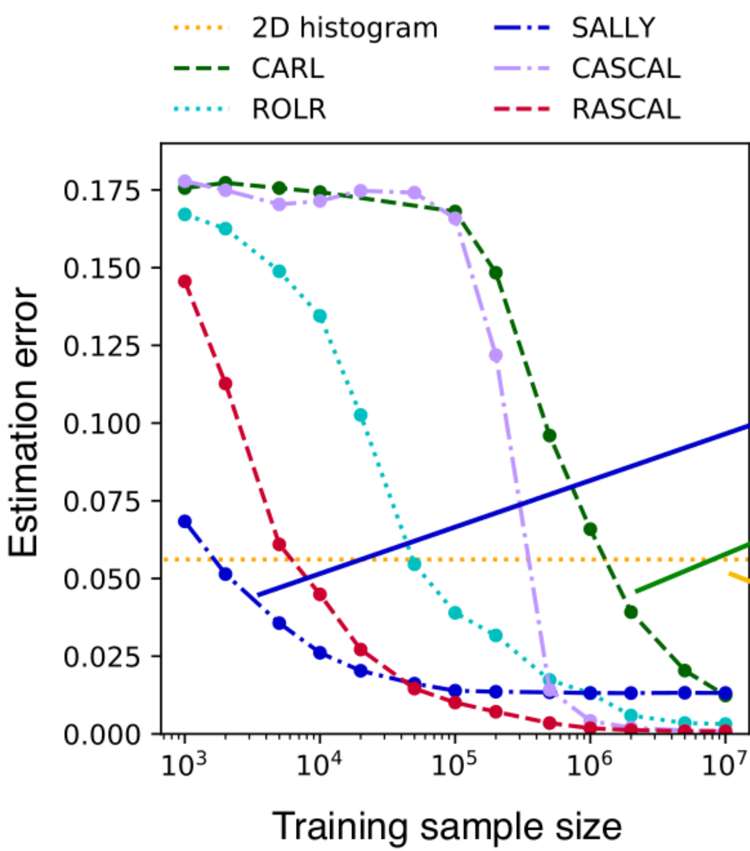
Goal: Constraints on two theory parameters:

$$\mathcal{L} = \mathcal{L}_{SM} + \underbrace{\frac{f_W}{\Lambda^2}} \frac{ig}{2} (D^\mu \phi)^\dagger \sigma^a D^\nu \phi \, W^a_{\mu\nu} - \underbrace{\frac{f_{WW}}{\Lambda^2}} \frac{g^2}{4} (\phi^\dagger \phi) \, W^a_{\mu\nu} W^{\mu\nu\, a}$$

Precise likelihood ratio estimates



Increased data efficiency

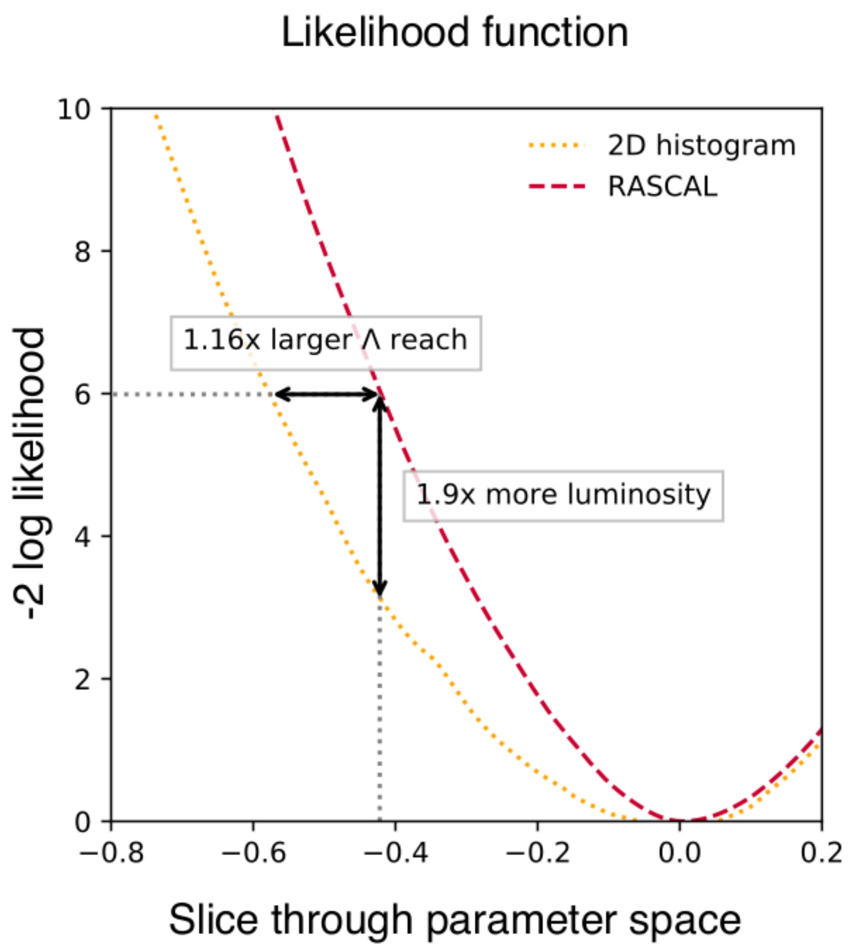


New techniques
require less data than
generic ML method
[CARL, see K. Cranmer, G. Louppe, J. Pavez 1506.02169]

Traditional Histogram

Credits: Johann Brehmer

Better sensitivity

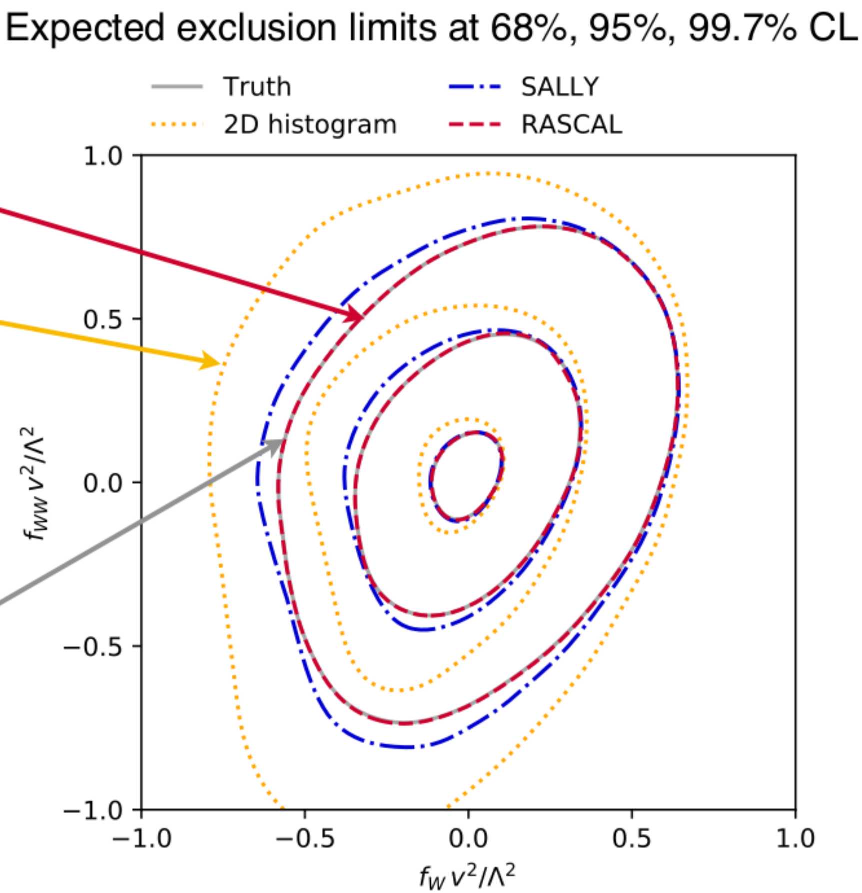


36 events, assuming SM

Credits: Johann Brehmer

Stronger bounds

RASCAL enables stronger limits than traditional histogram



Limits from RASCAL virtually indistinguishable from true likelihood
(usually we don't have that)

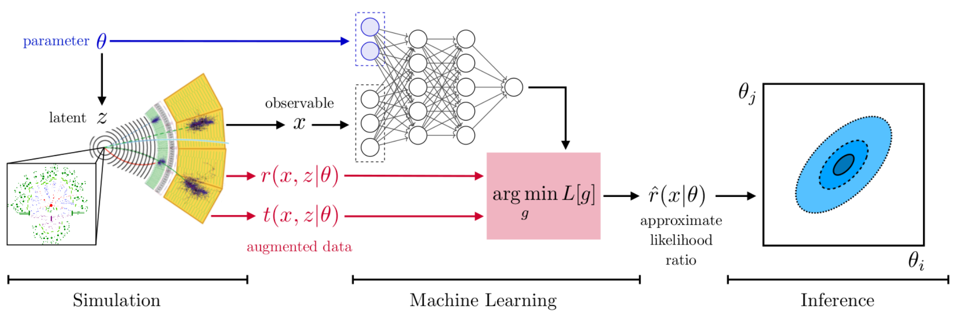
36 events, assuming SM

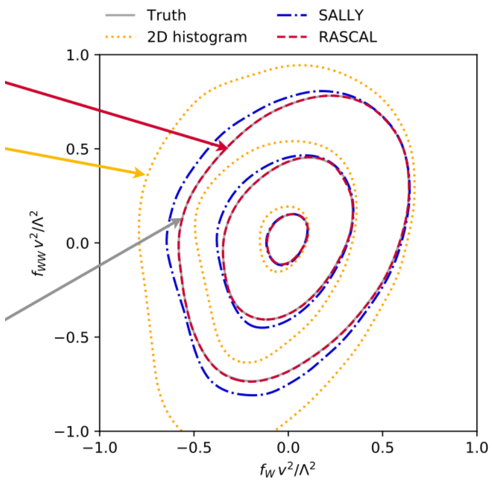
Credits: Johann Brehmer

Summary

Summary

- Many LHC analysis (and much of modern science) are based on "likelihood-free" simulations.
- New inference algorithms:
  - Leverage more information from the simulator
  - Combine with the power of machine learning
- First application to LHC physics: stronger EFT constraints with less simulations.





Collaborators



References

- Stoye, M., Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. (2018). Likelihood-free inference with an improved cross-entropy estimator. arXiv preprint arXiv:1808.00973.
- Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. (2018). Mining gold from implicit models to improve likelihood-free inference. arXiv preprint arXiv:1805.12244.
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. (2018). Constraining Effective Field Theories with Machine Learning. arXiv preprint arXiv:1805.00013.
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. (2018). A Guide to Constraining Effective Field Theories with Machine Learning. arXiv preprint arXiv:1805.00020.
- Cranmer, K., Pavez, J., & Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. arXiv preprint arXiv:1506.02169.

The end.