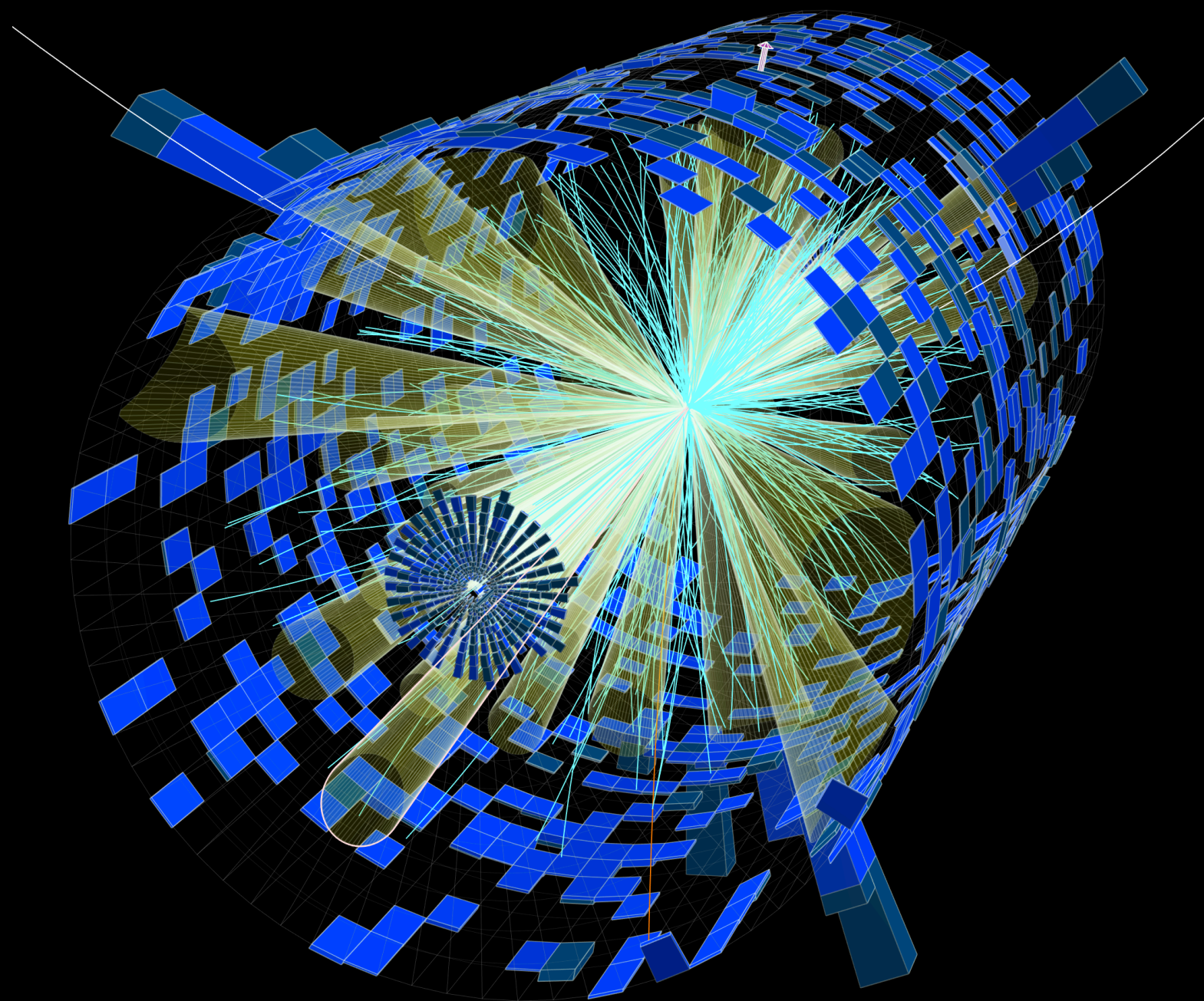




# SIMULATION-BASED INFERENCE, INTERPRETABILITY, AND EXPERIMENTAL DESIGN



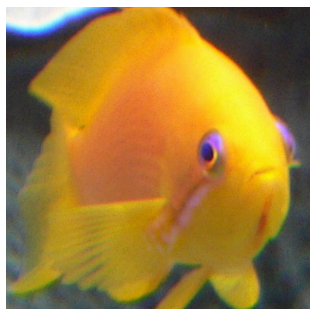
**@KyleCranmer**

New York University  
Department of Physics  
Center for Data Science  
CILVR Lab

SUPPORT



**The SCAILFIN Project**  
[scailfin.github.io](https://scailfin.github.io)





# COLLABORATORS (+ MANY MORE)



Gilles Louppe  
U. Liège



Kyunghyun Cho



Joan Bruna



Brenden Lake



Meghan Frate



Juan Pavez



Sid Mishra-Sharma



Johann Brehmer



Felix Kling



Lukas Heinrich



Markus Store



Tim Head



Peter Battaglia



Irina Espejo



Peter Sadowski



Daniel Whiteson



Pierre Baldi



Lezcano Casado



Atılım Güneş Baydin  
University of Oxford



Prabhat  
NERSC, Berkeley Lab



Wahid Bhimji  
NERSC, Berkeley Lab



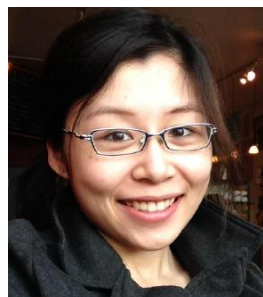
Frank Wood  
University of Oxford



Phiala Shanahan



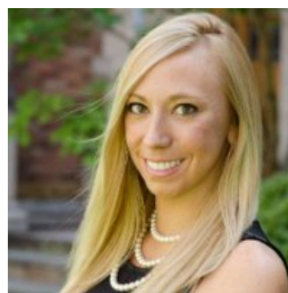
William Detmold



Karen Ng



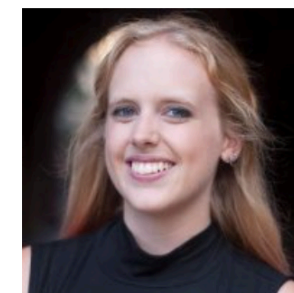
Tuan Anh Le



Michela Paganini  
Yale University



Daniela Huppenkothen  
New York University



Savannah Thais  
Yale University



Ruth Angus  
Columbia University



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



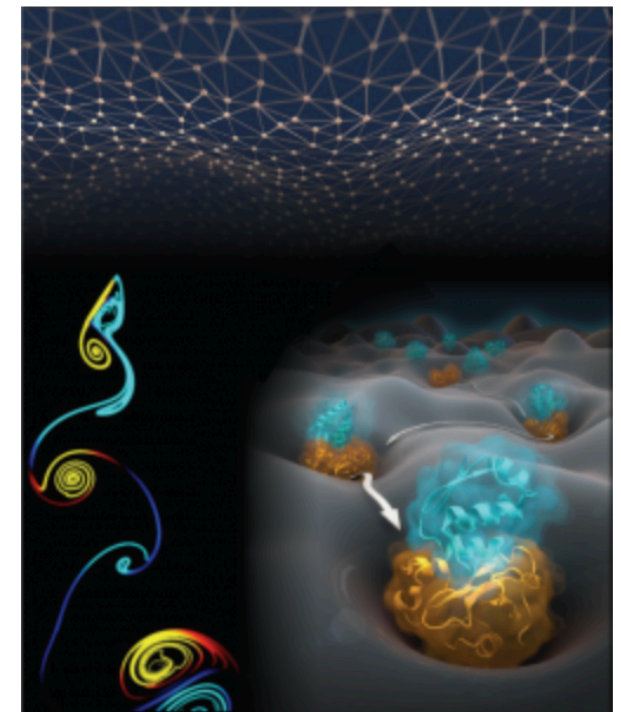
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.





# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



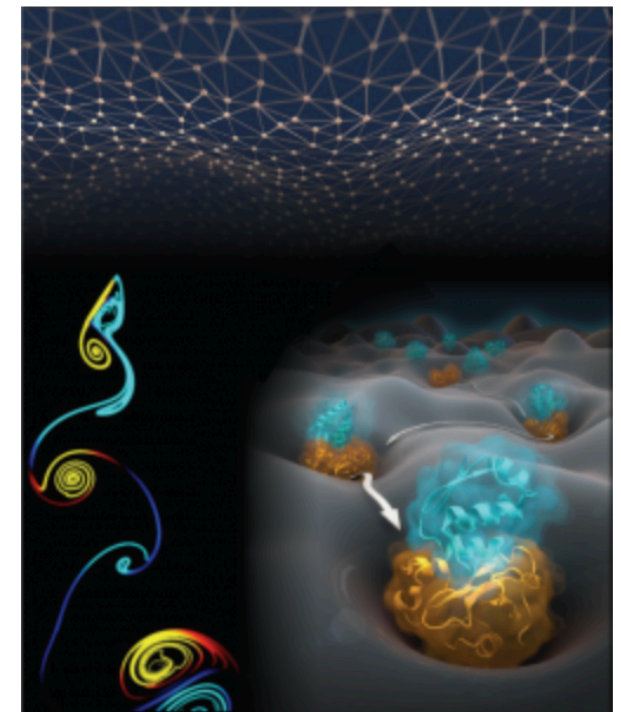
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



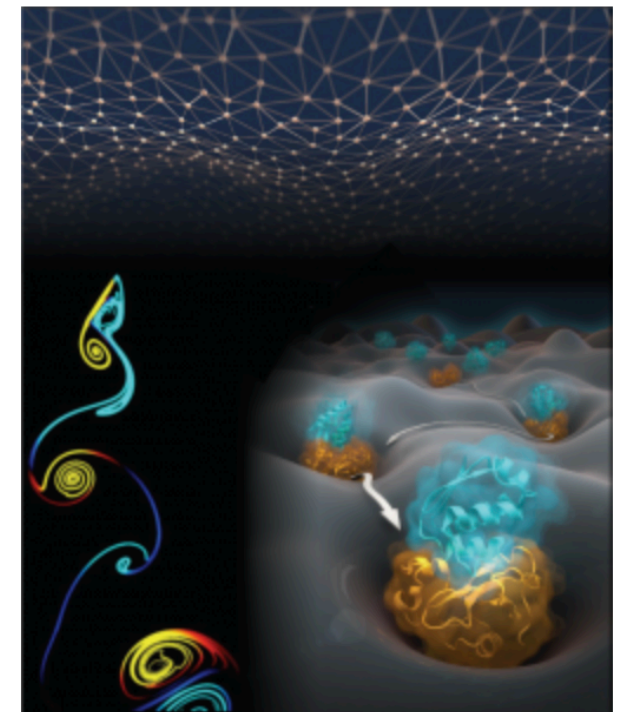
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.





# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



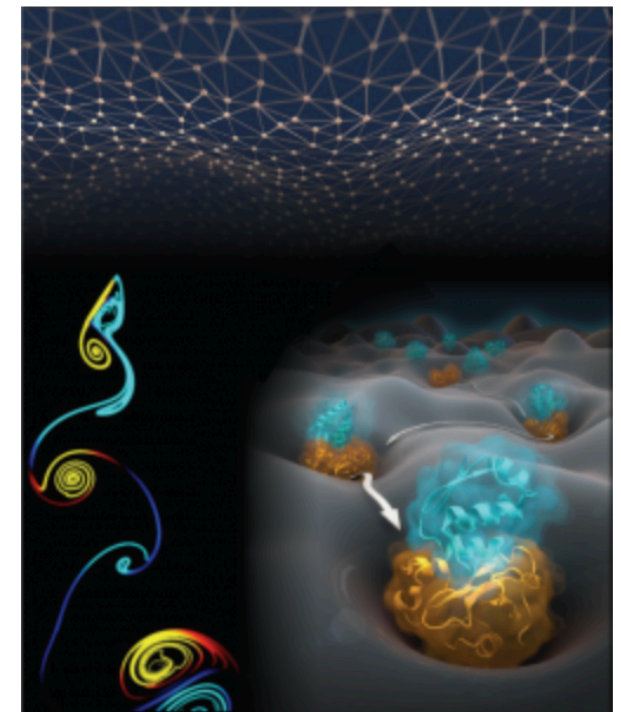
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



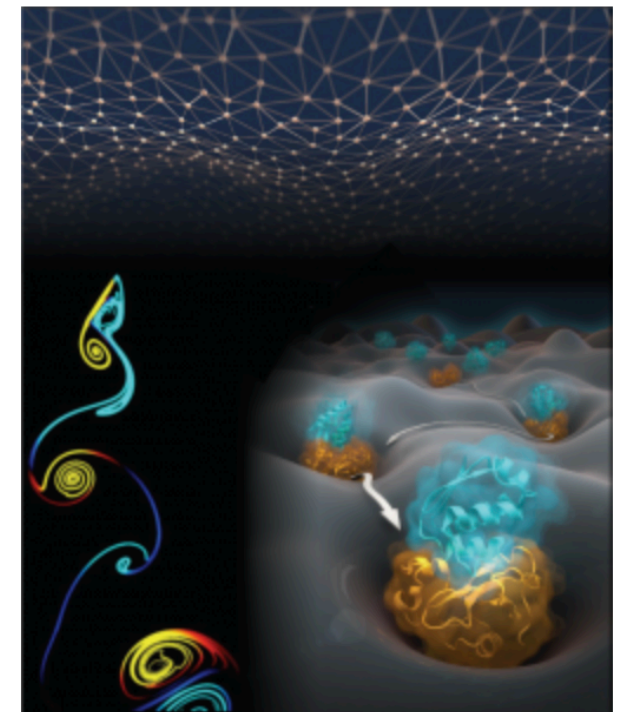
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.





# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



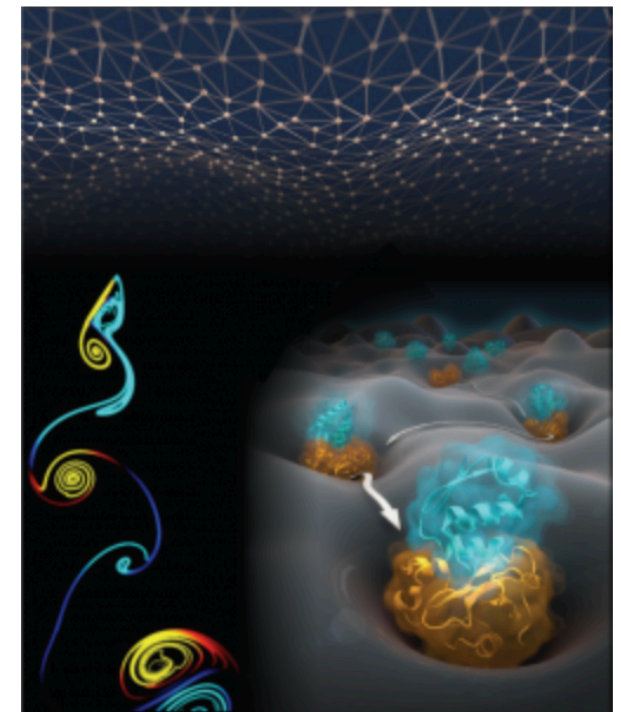
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



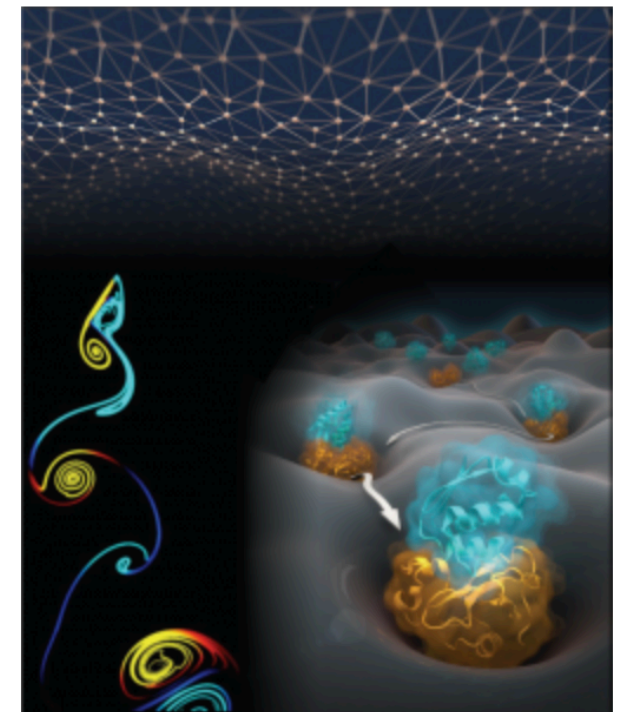
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.





# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



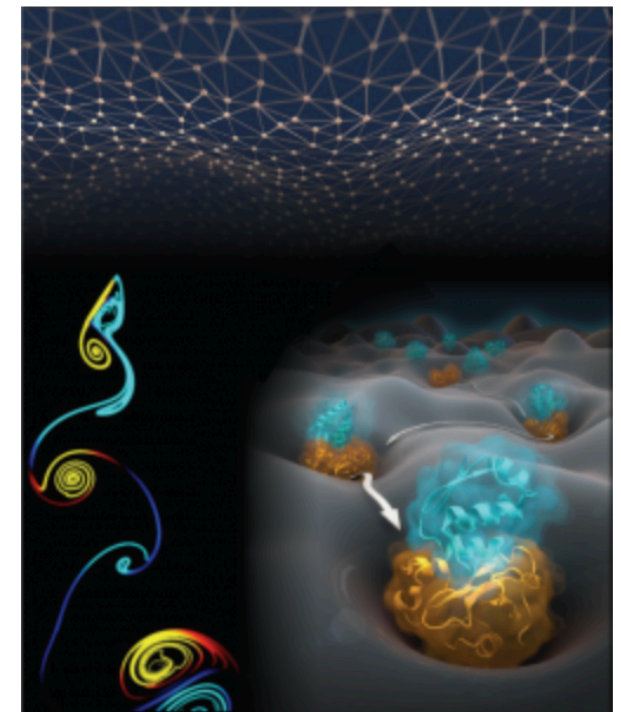
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



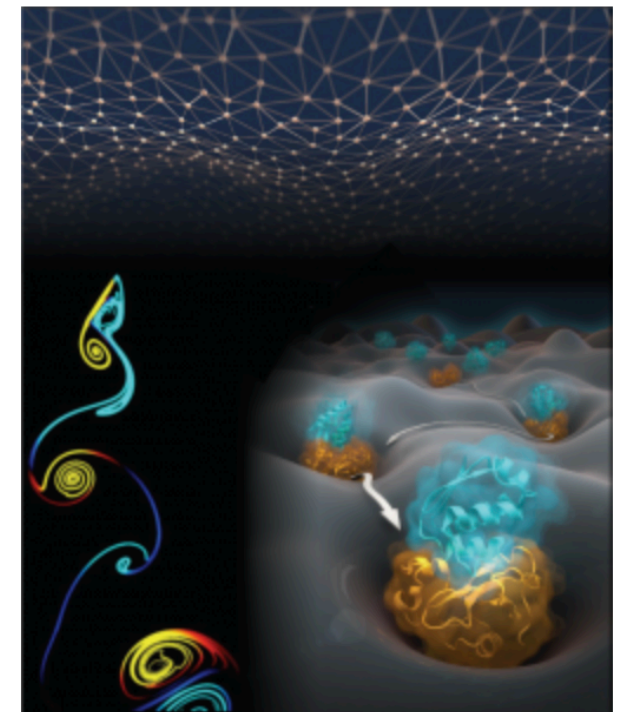
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

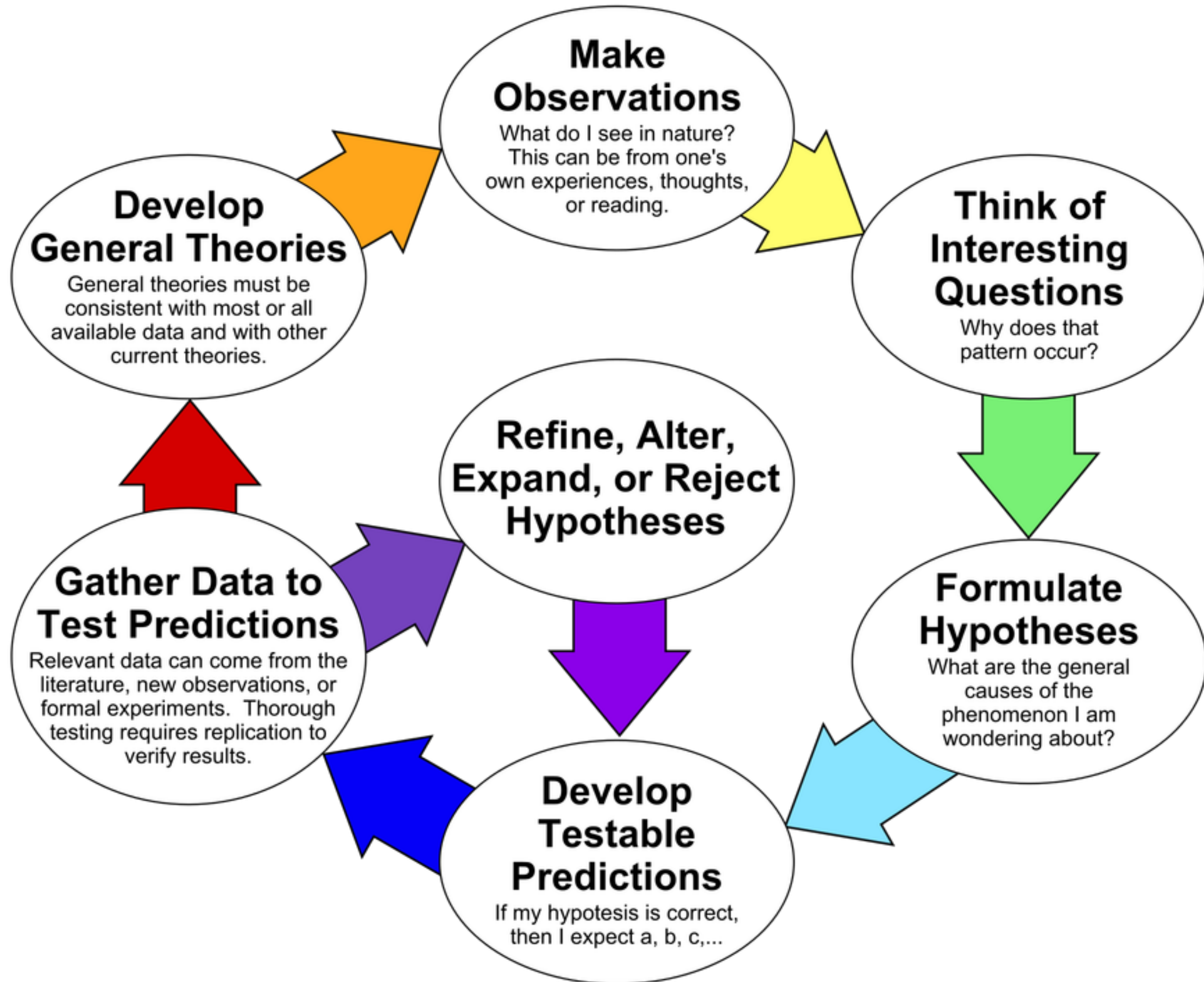
As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.

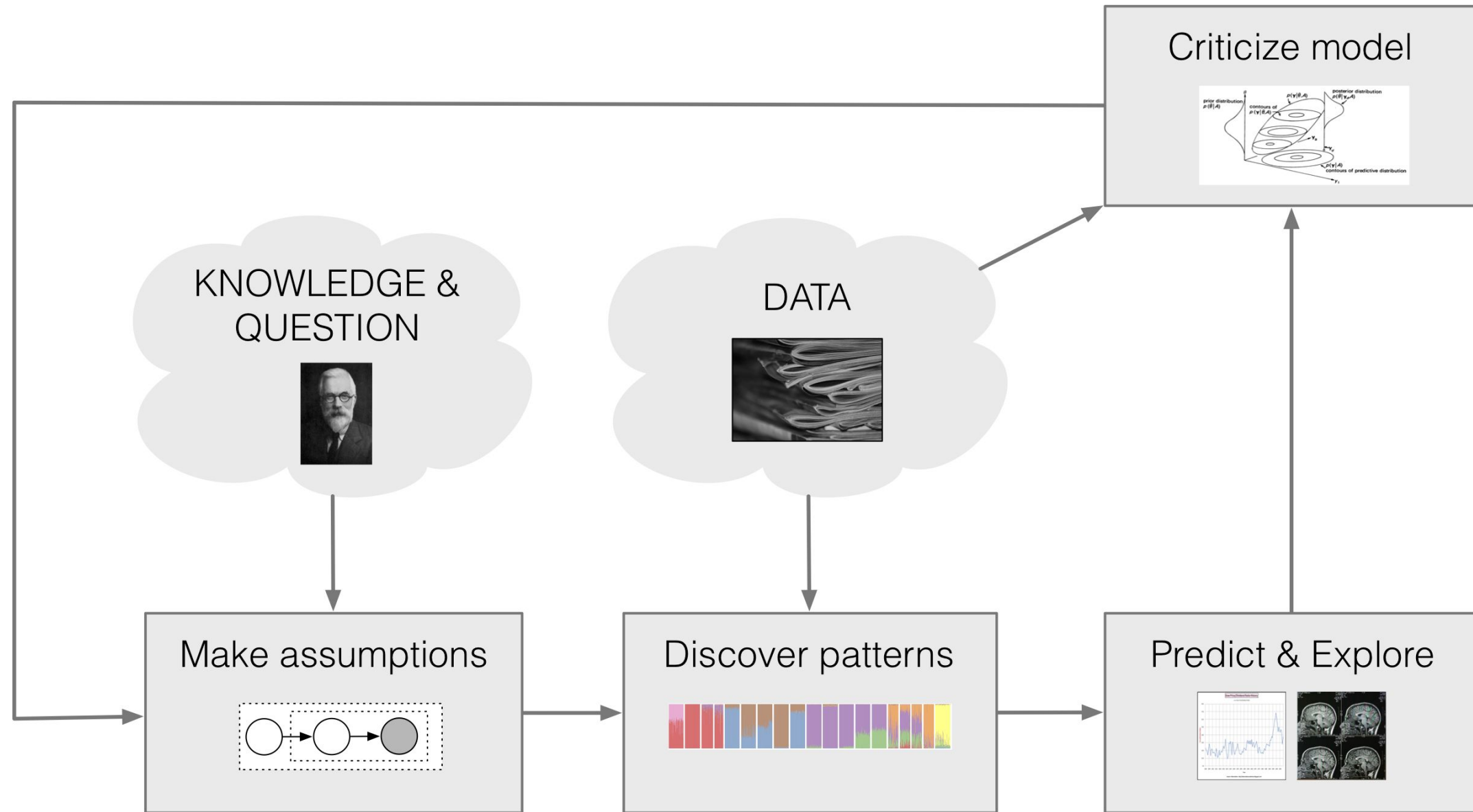




# The Scientific Method as an Ongoing Process



# A PROBABILISTIC PIPELINE

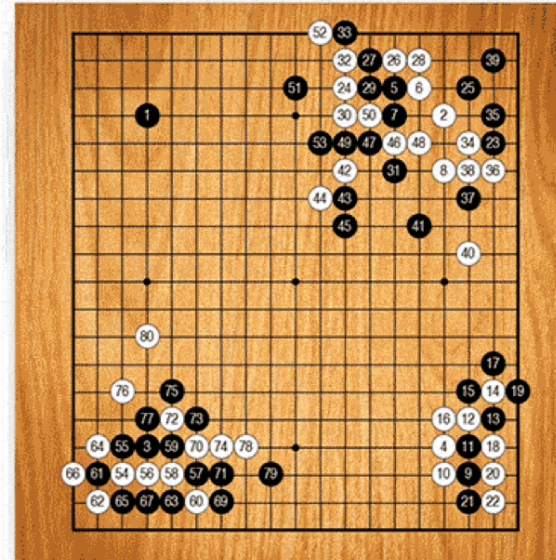


[Box, 1980; Rubin, 1984; Gelman et al., 1996; Blei, 2014]



# REINFORCEMENT LEARNING & SCIENTIFIC METHOD

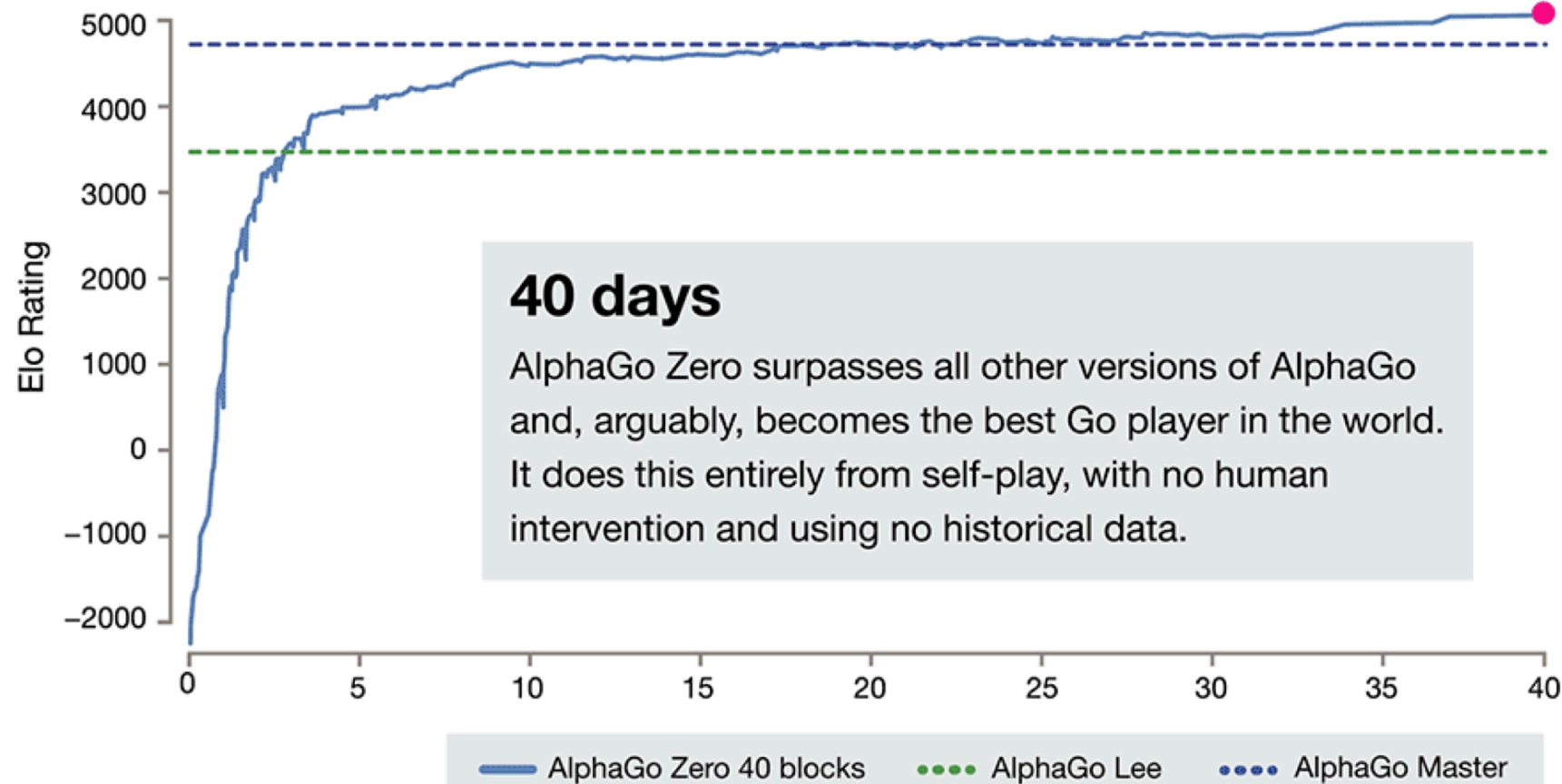
## Scientist trying to decide what experiment to do next



68 at 61  
Captured Stones

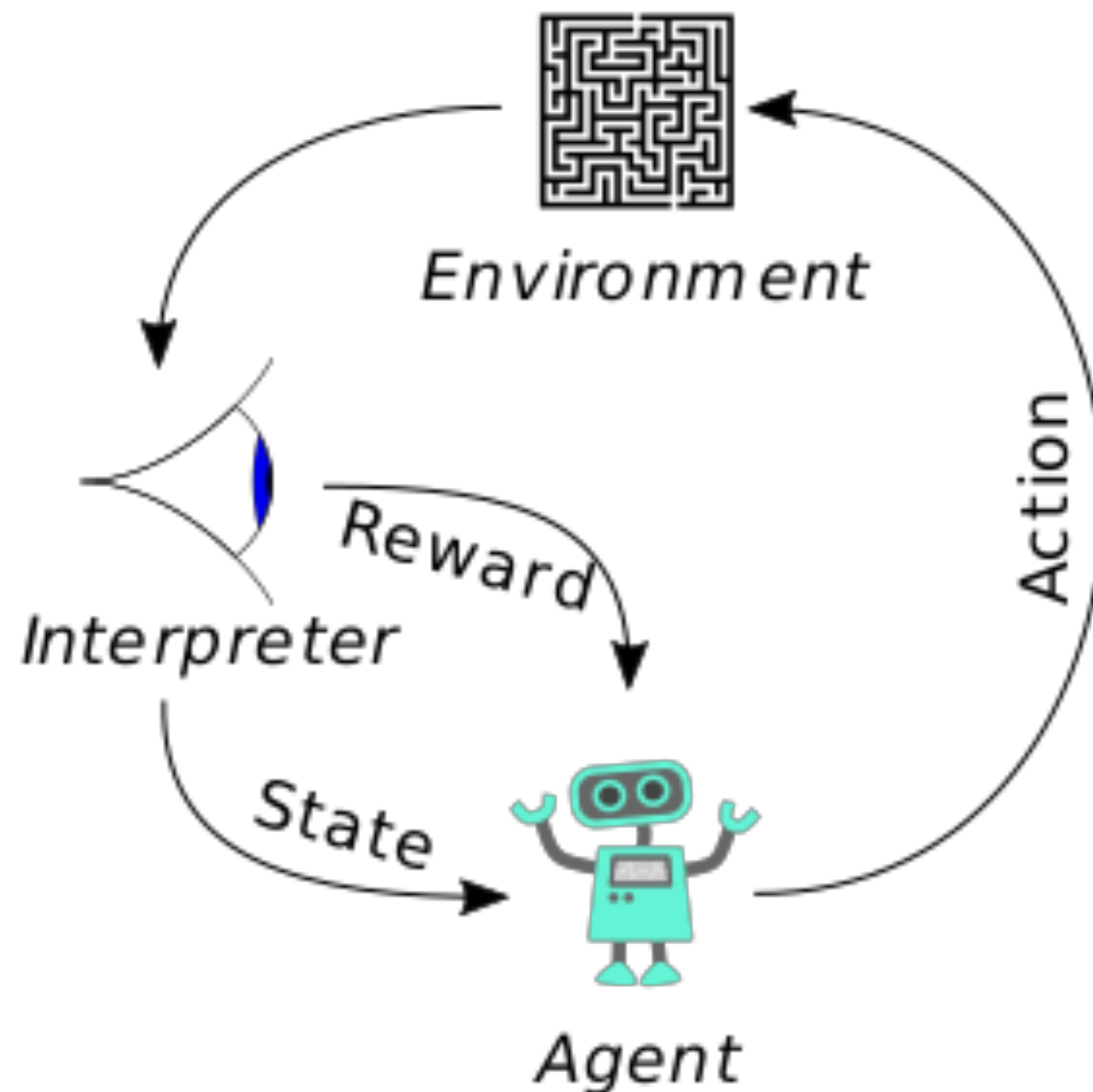
### 70 hours

AlphaGo Zero plays at super-human level. The game is disciplined and involves multiple challenges across the board.



# REINFORCEMENT LEARNING & SCIENTIFIC METHOD

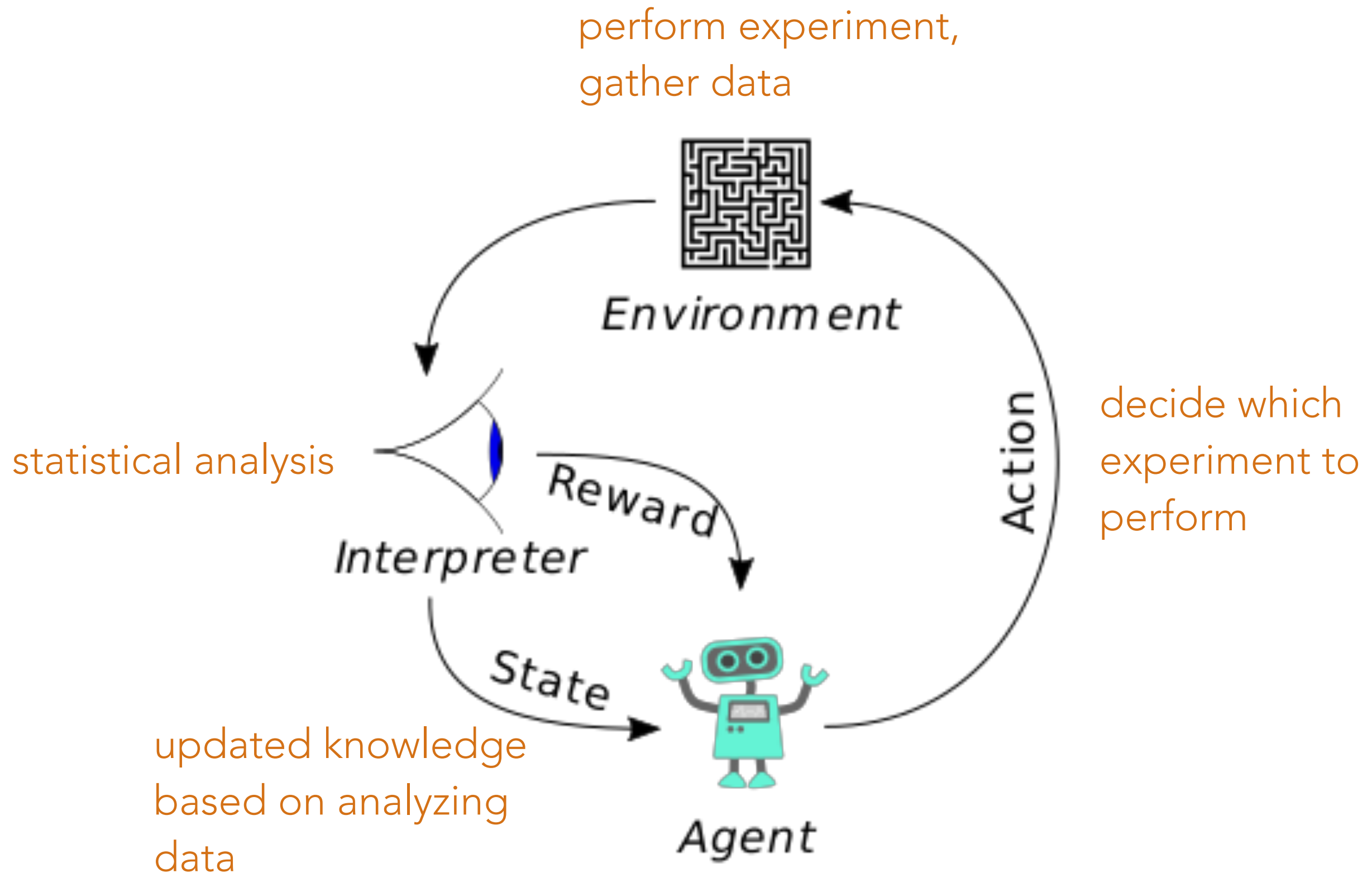
Scientist trying to decide what experiment to do next



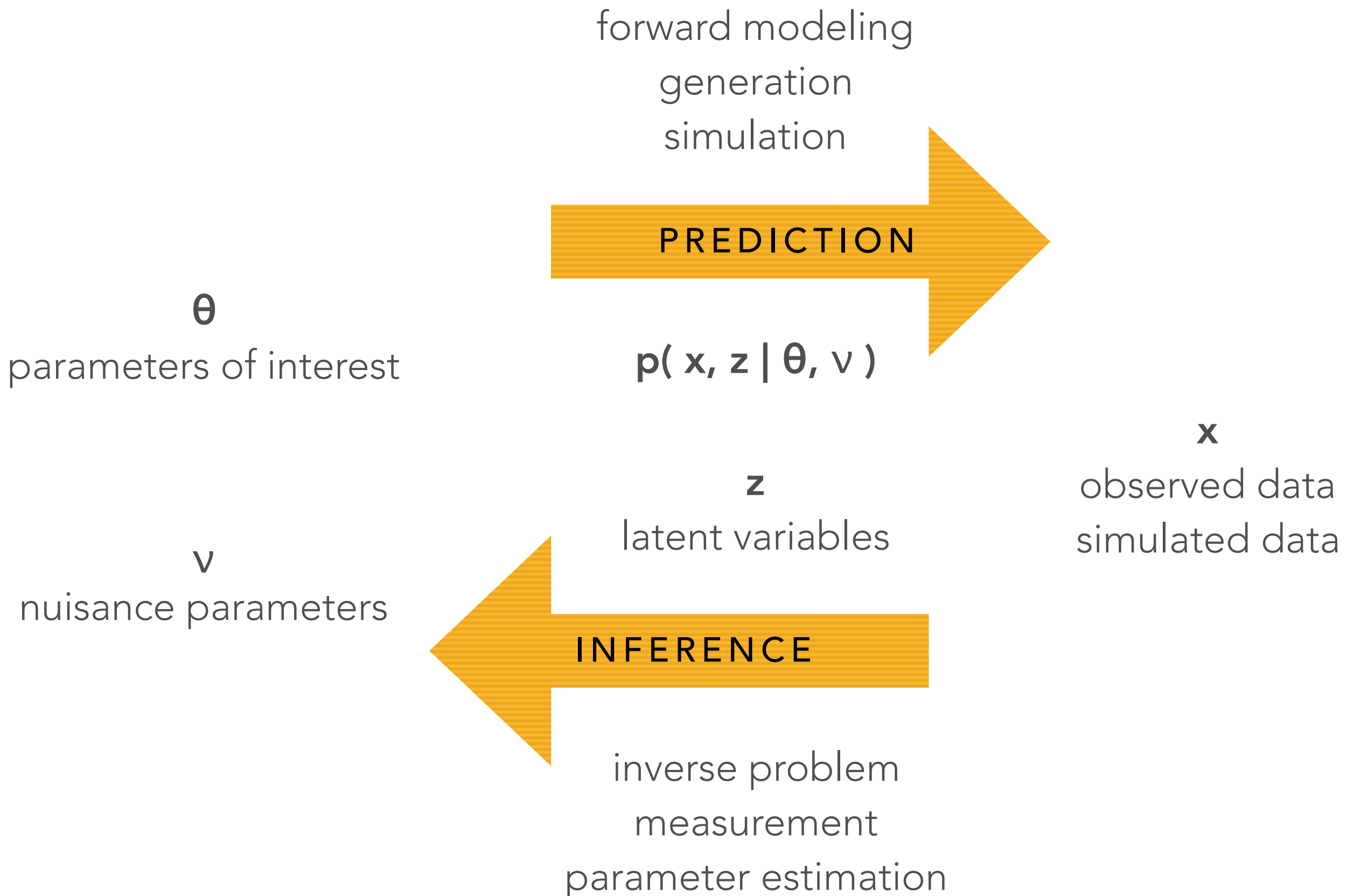


# REINFORCEMENT LEARNING & SCIENTIFIC METHOD

Scientist trying to decide what experiment to do next



# NOTATION / TERMINOLOGY





# STATISTICAL DECISION THEORY IN 1 SLIDE

$\Theta$  - States of nature;  $X$  - possible observations;  $A$  - action to be taken

$p(x|\theta)$  - statistical model (likelihood);  $\pi(\theta)$  - prior

$\delta: X \rightarrow A$  - **decision rule** (take some action based on observation)

$L: \Theta \times A \rightarrow \mathbb{R}$  - **loss function**, real-valued function true parameter and action

$R(\theta, \delta) = E_{p(x|\theta)}[L(\theta, \delta)]$  - **risk**

$r(\pi, \delta) = E_{\pi(\theta)}[R(\theta, \delta)]$  - **Bayes risk** (expectation over  $\theta$  w.r.t. prior and possible observations)

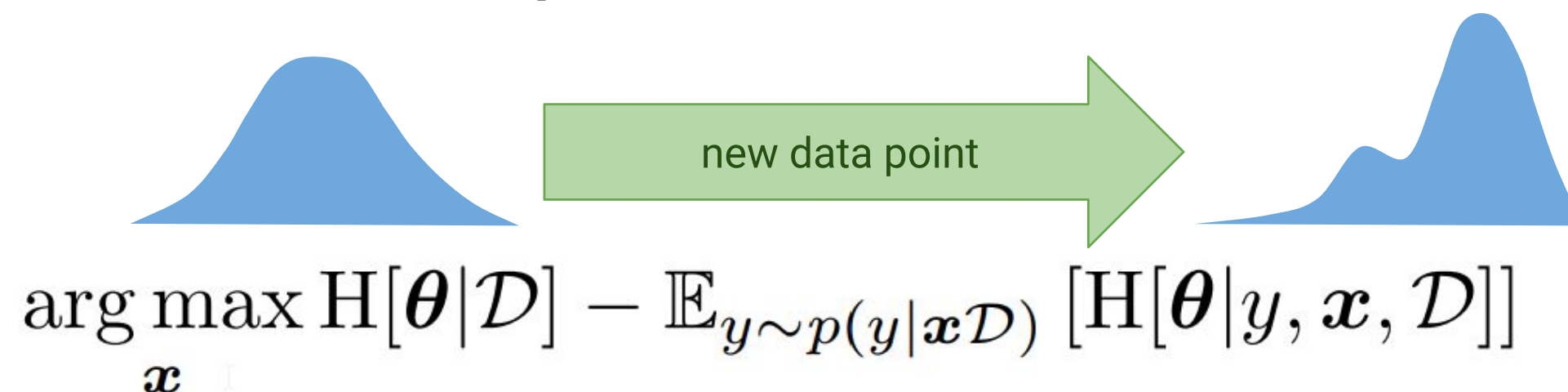
$\rho(\pi, \delta | x) = E_{\pi(\theta|x)}[L(\theta, \delta(x))]$  - **expected loss** (expectation over  $\theta$  w.r.t. posterior  $\pi(\theta|x)$ )



## Active Learning & Control

Given data points  $\{x, y\}$ , how to select the next data point to fit the model?

**Ex.** Select data points which maximize expected information gain. [Lindley et al. 1956; Mackay 1992; Houthoof et al. 2016]

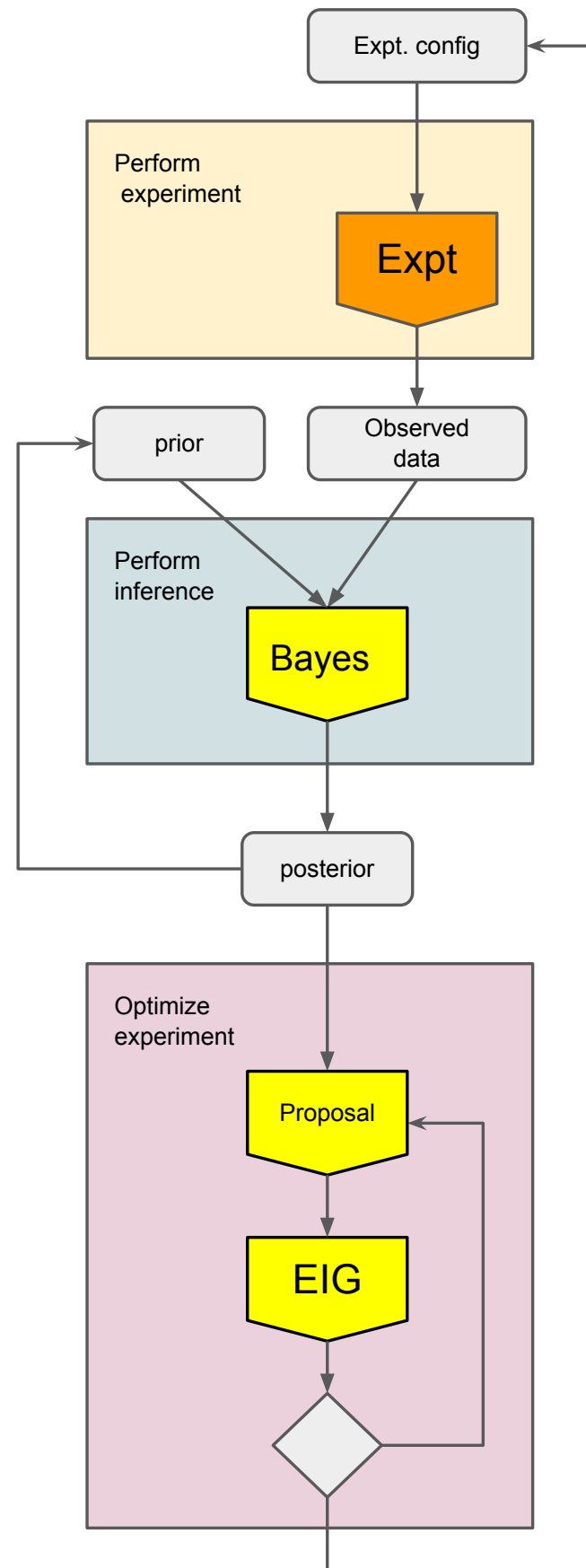


Uncertainty determines which  $\mathbf{x}$  is most informative and, therefore, the model's success.

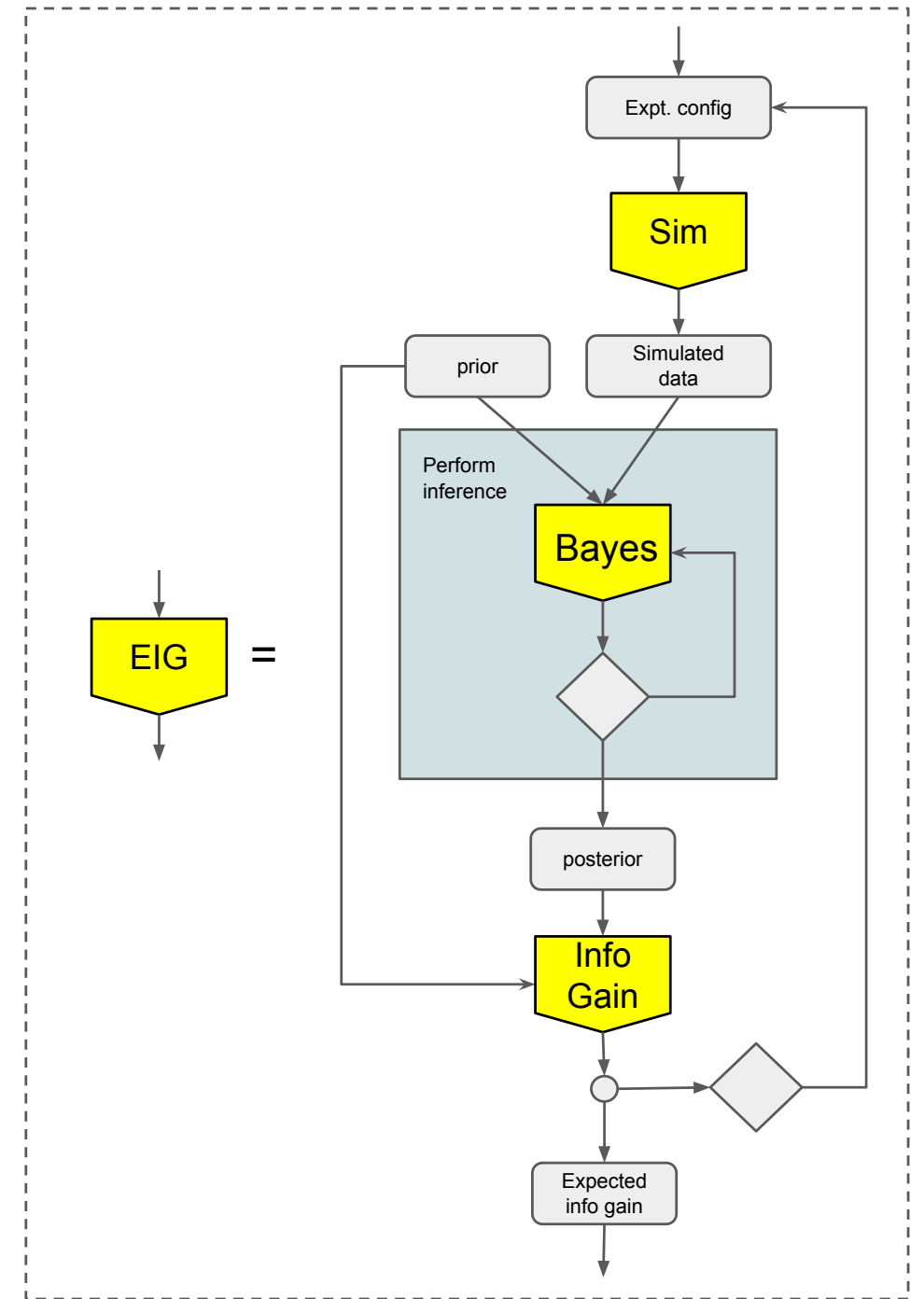
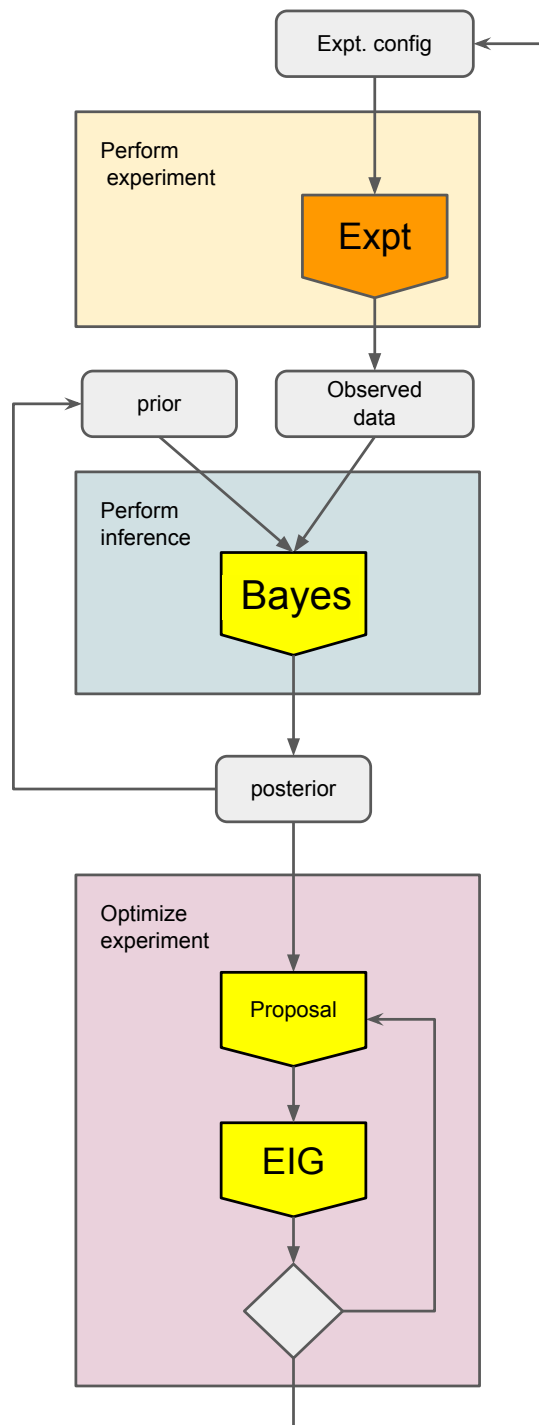
[Hafner et al., 2019]



# "ACTIVE SCIENCING"



# "ACTIVE SCIENCING"



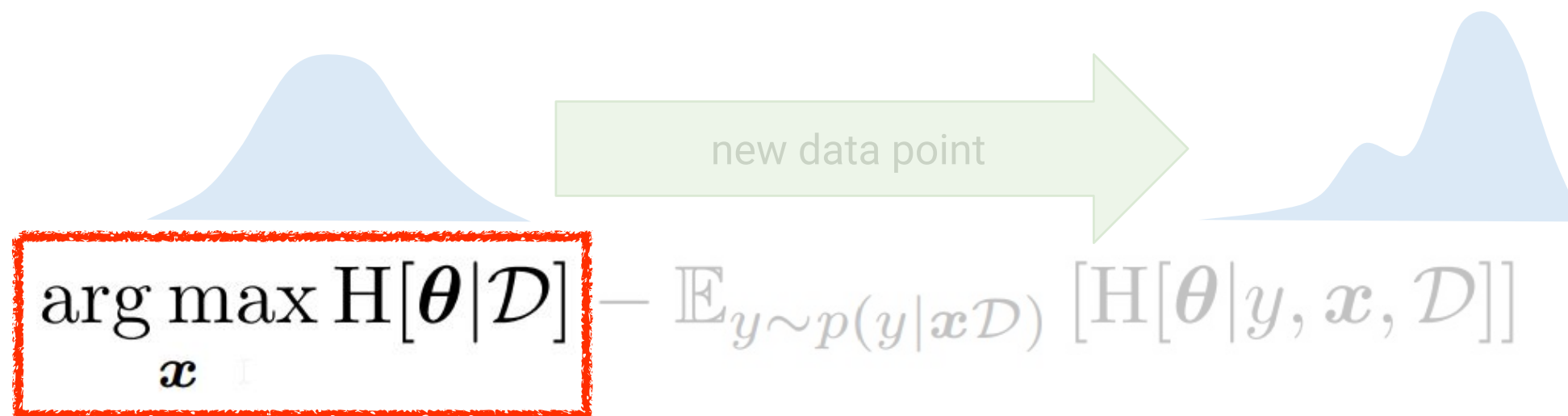




## Active Learning & Control

Given data points  $\{x, y\}$ , how to select the next data point to fit the model?

**Ex.** Select data points which maximize expected information gain. [Lindley et al. 1956; Mackay 1992; Houthooft et al. 2016]



Uncertainty determines which  $\mathbf{x}$  is most informative and, therefore, the model's success.

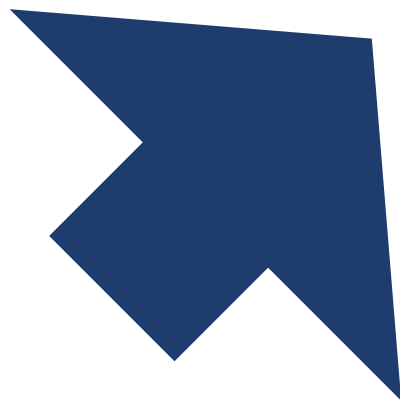
[Hafner et al., 2019]

# SYNTHESIS

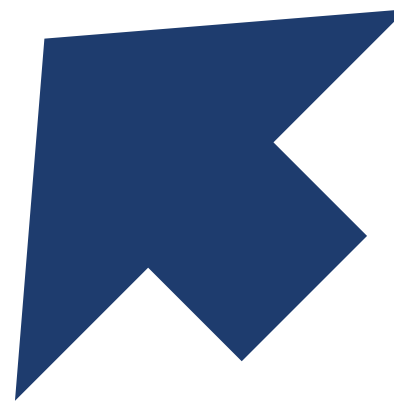
active learning / sequential design / black box optimization



## Active Sciencing



reusable workflows

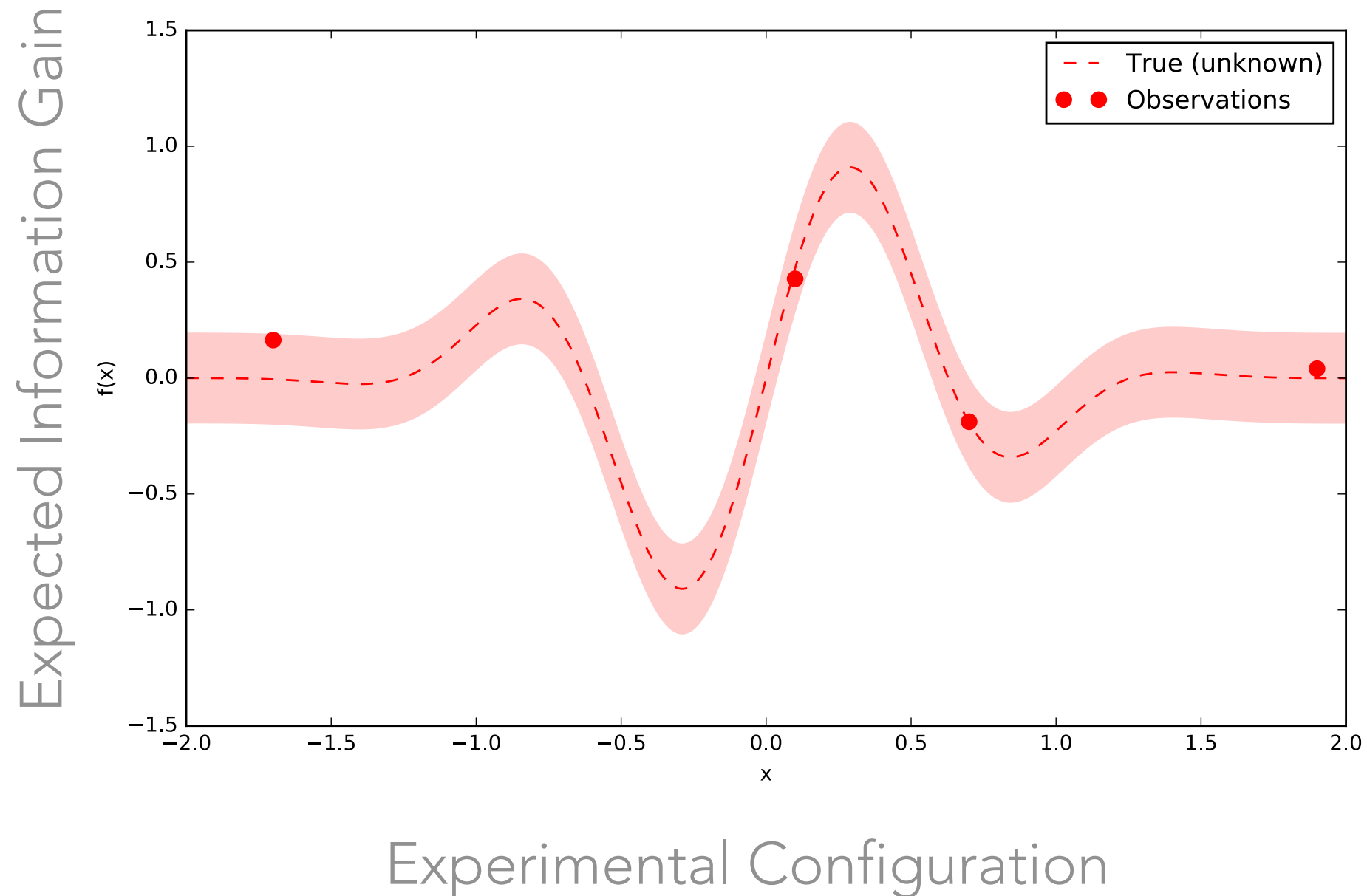


simulation-based /  
likelihood-free  
inference engines



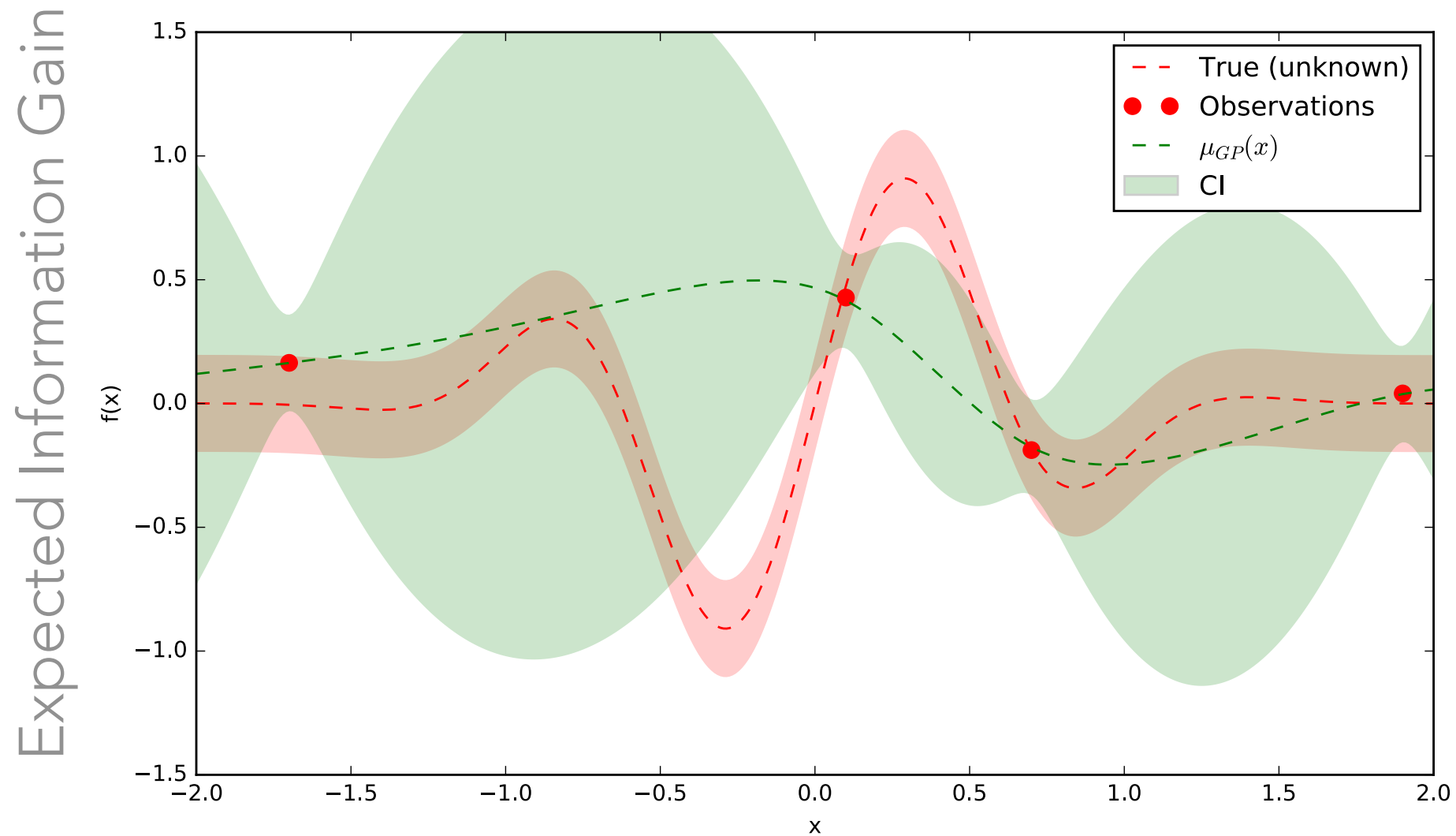
# BAYESIAN OPTIMIZATION

What experimental configuration should we evaluate next?



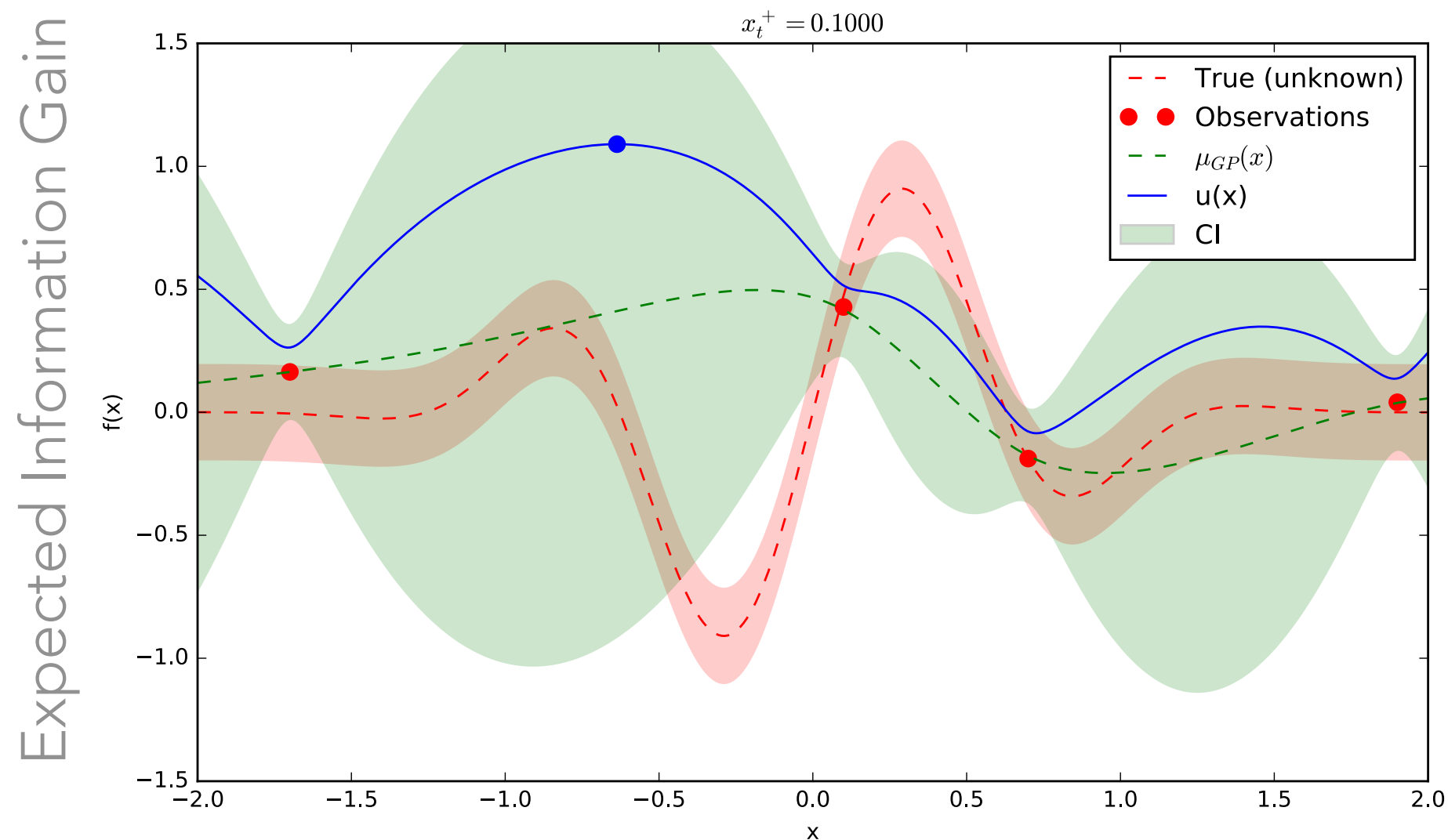
# BAYESIAN OPTIMIZATION

Build a probabilistic model for the EIG objective function



# BAYESIAN OPTIMIZATION

Use it to define an inexpensive acquisition function, and optimize that.  
Use this as next configuration to evaluate expensive EIG

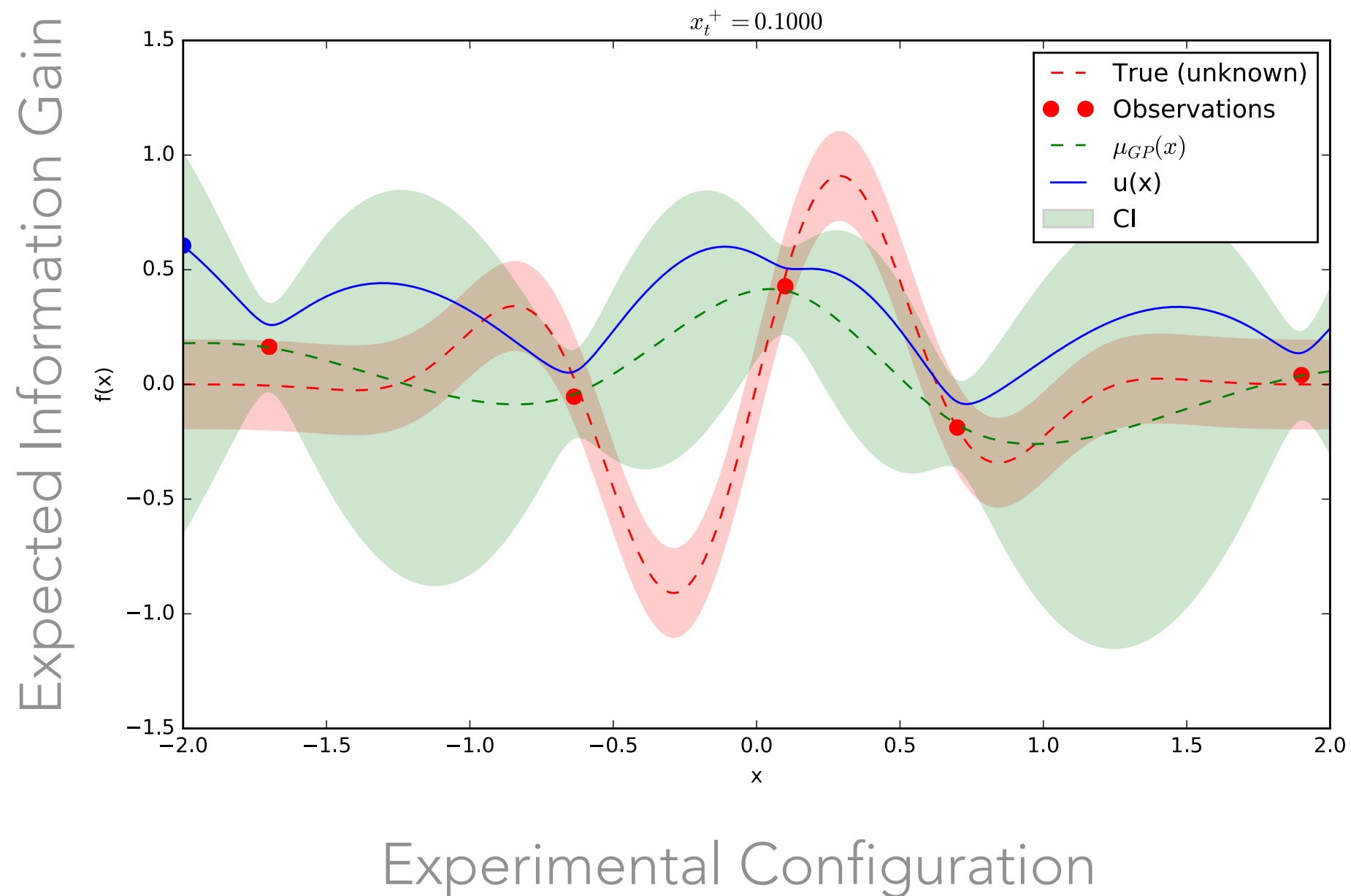


$$x_{t+1} = \arg \max_x \text{UCB}(x)$$



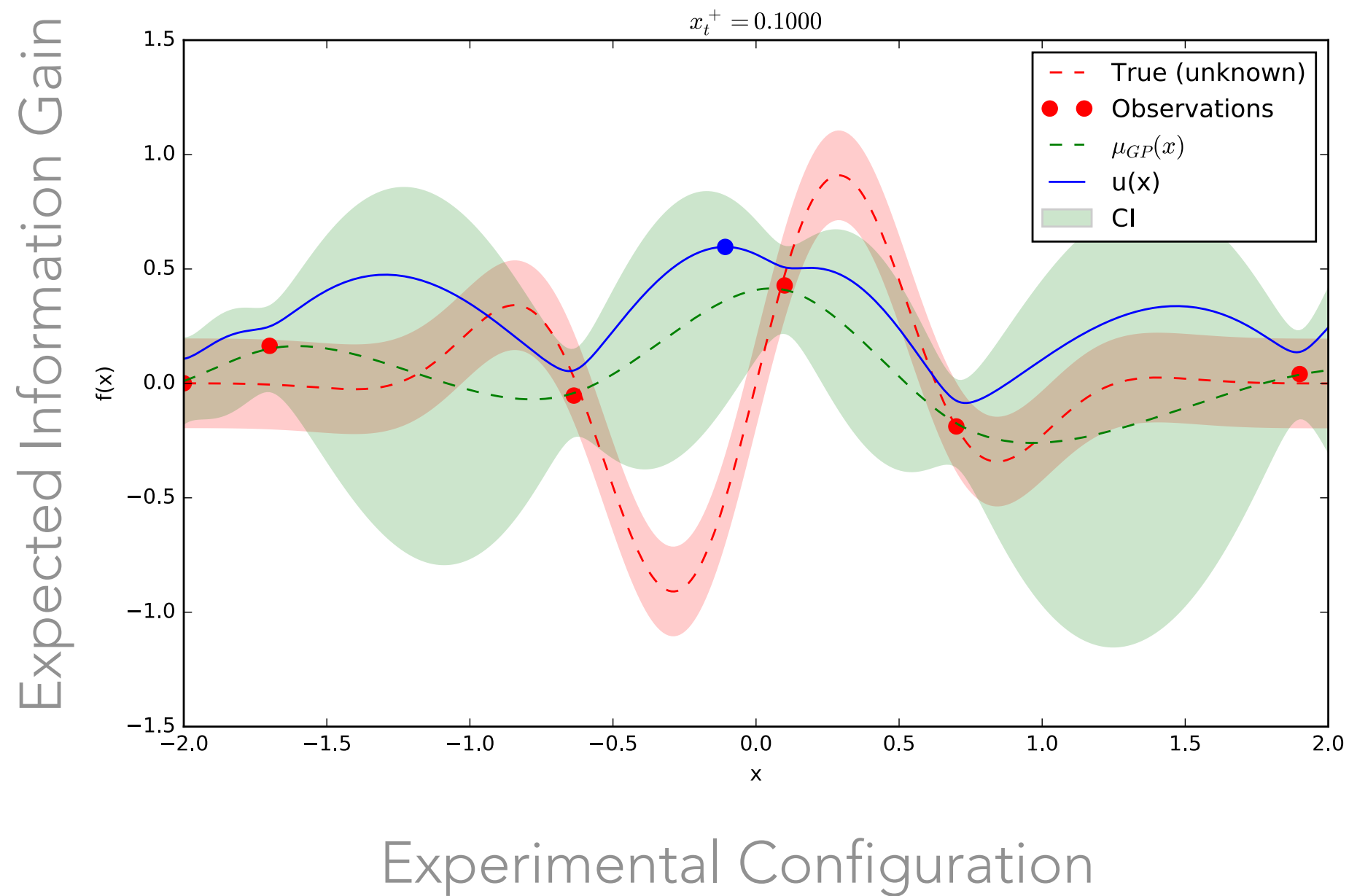
# BAYESIAN OPTIMIZATION

Then evaluate EIG for that configuration. Repeat



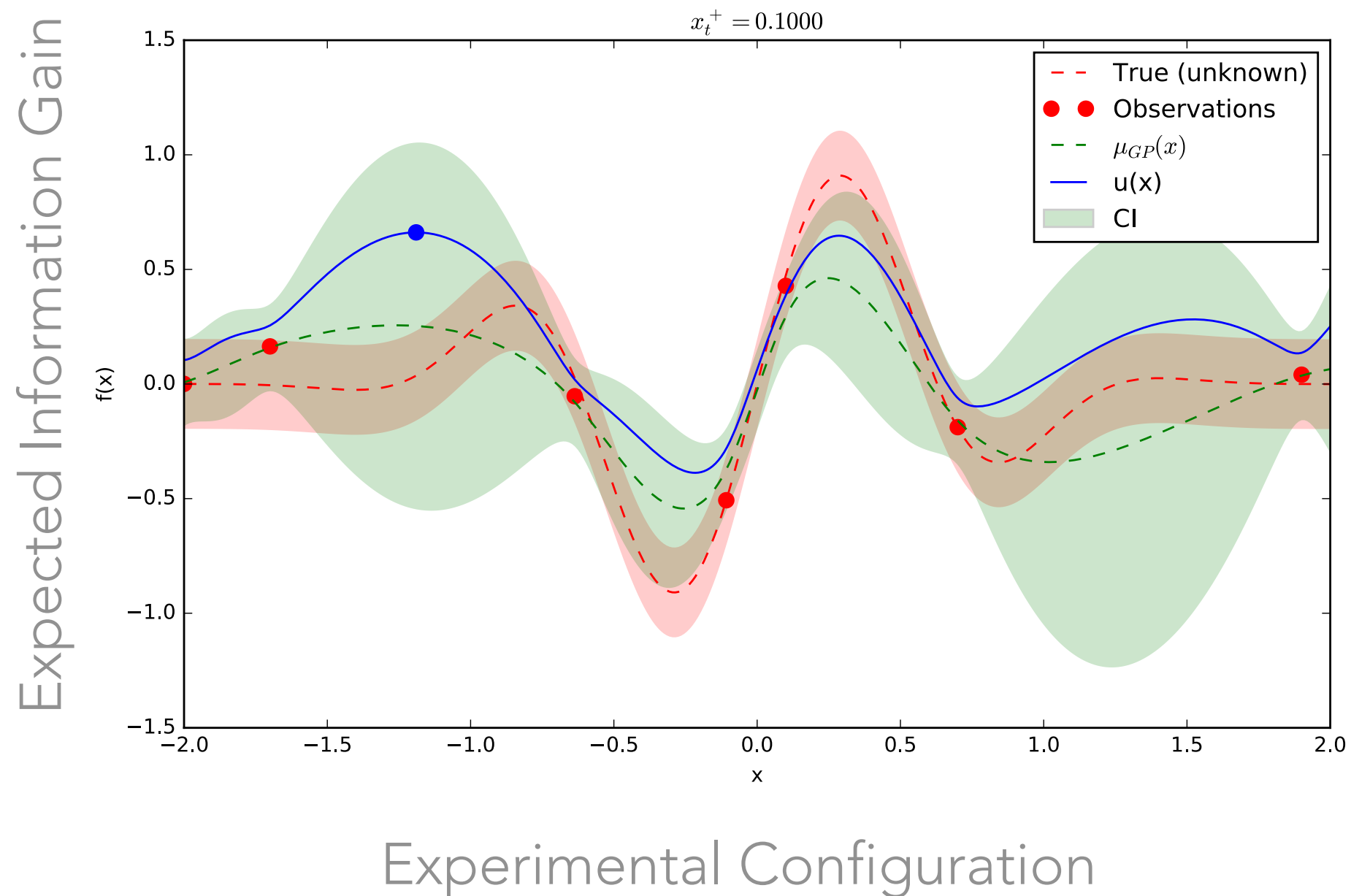
# BAYESIAN OPTIMIZATION

Then evaluate EIG for that configuration. Repeat



# BAYESIAN OPTIMIZATION

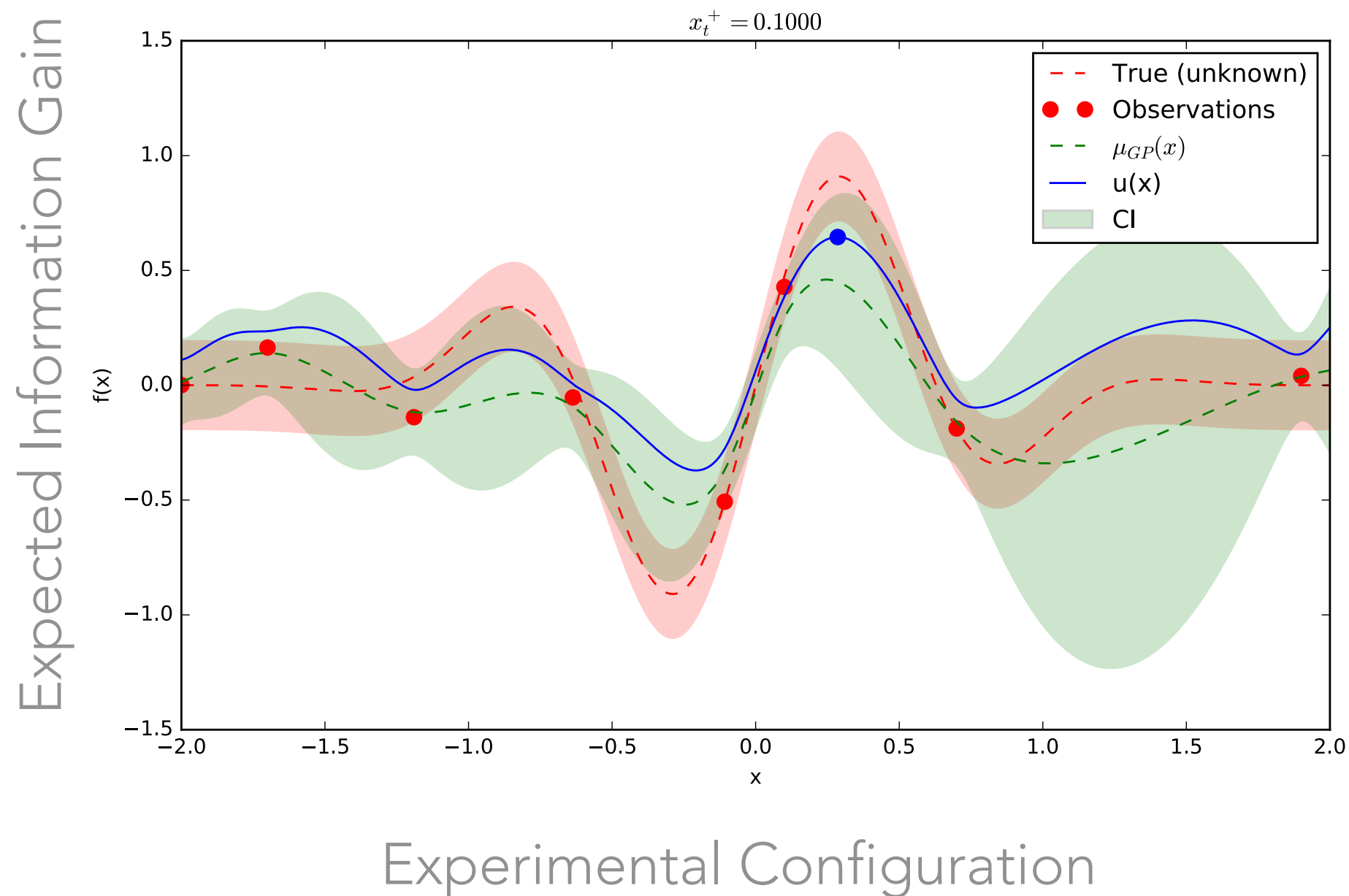
Then evaluate EIG for that configuration. Repeat





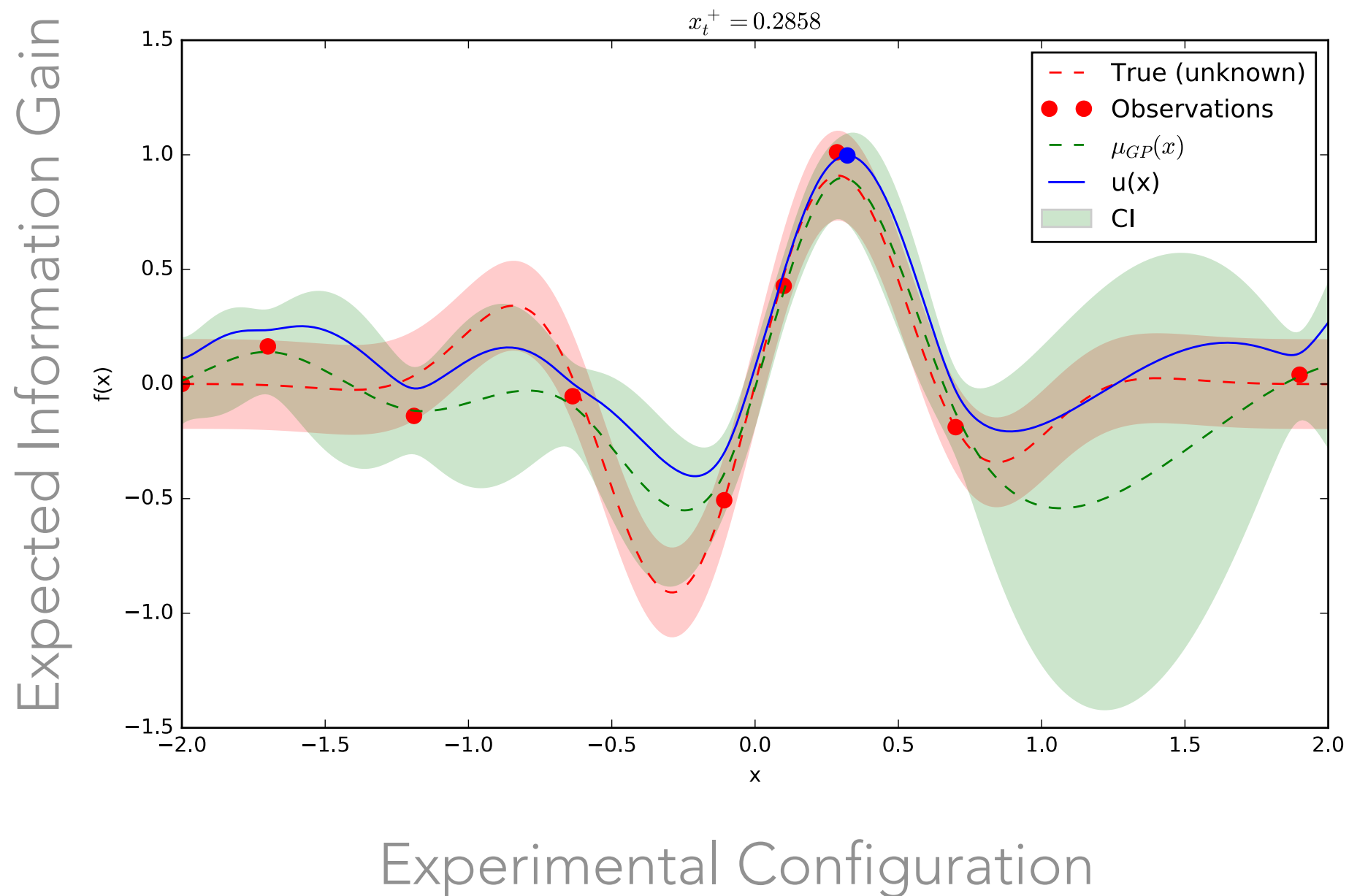
# BAYESIAN OPTIMIZATION

Then evaluate EIG for that configuration. Repeat



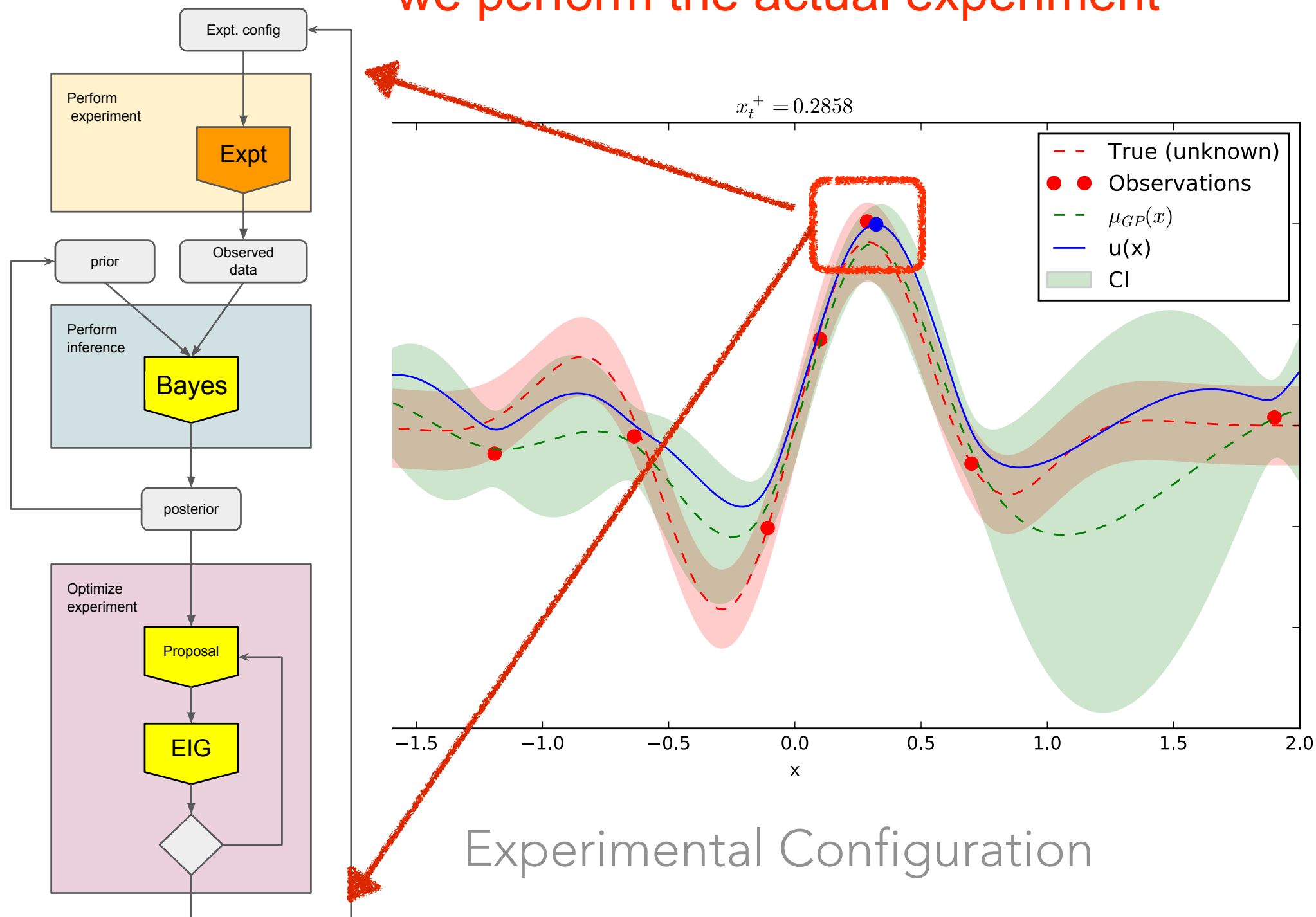
# BAYESIAN OPTIMIZATION

Then evaluate EIG for that configuration. Repeat



# BAYESIAN OPTIMIZATION

Now that this experimental configuration is optimized (with simulation), we perform the actual experiment



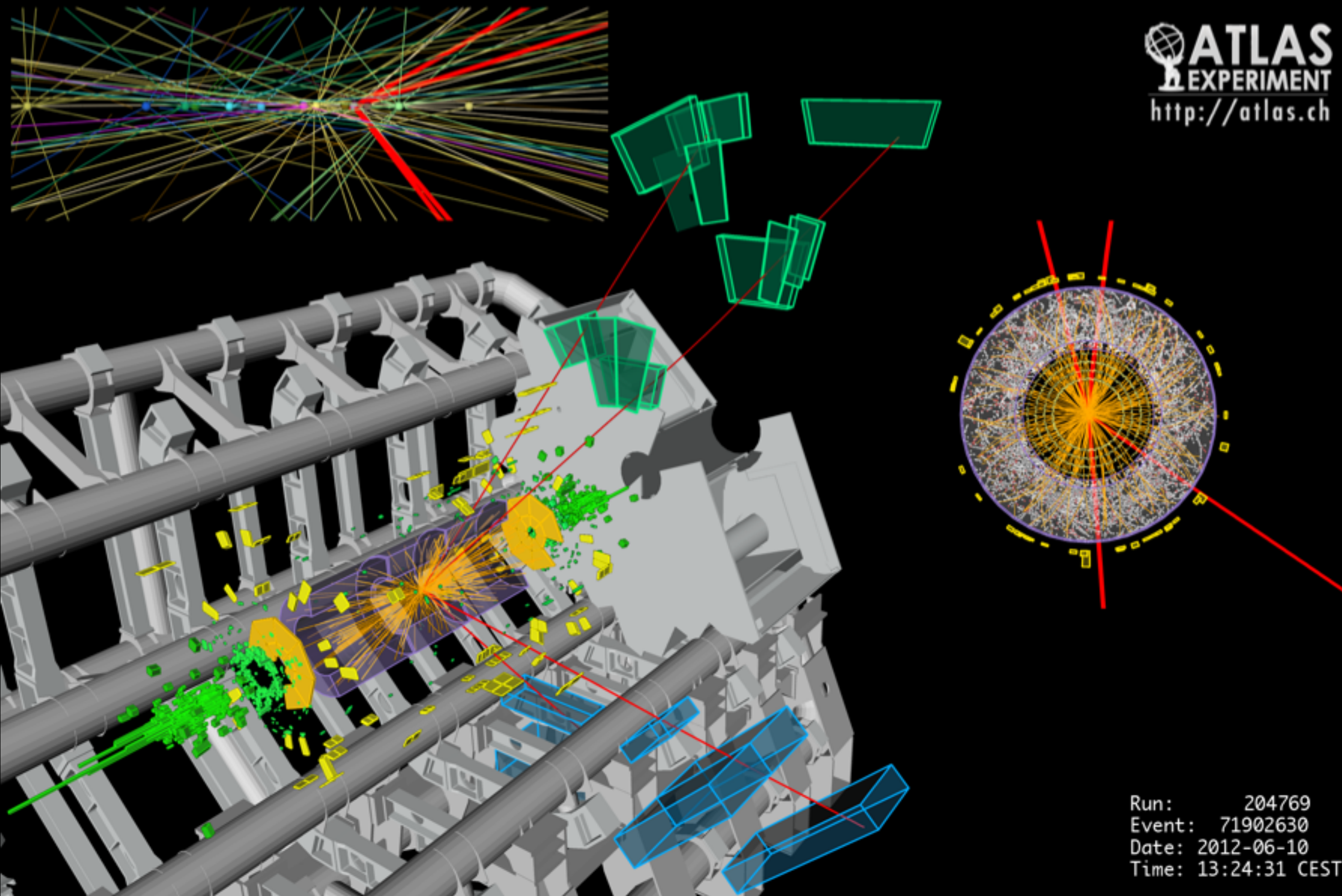


What is the probability model used for inference?

Shouldn't we use the same simulator we used to  
evaluate the expected information gain?

# PARTICLE PHYSICS

 **ATLAS**  
EXPERIMENT  
<http://atlas.ch>

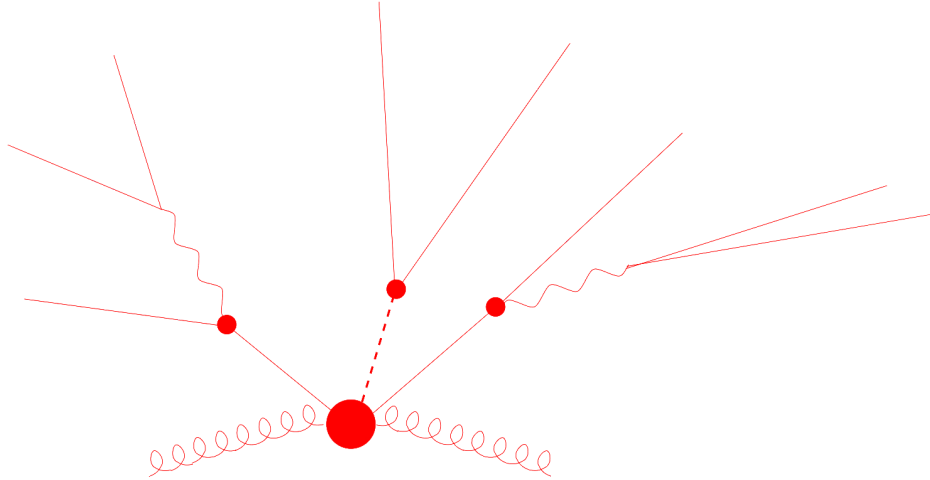


# THE CAUSAL, GENERATIVE MODEL

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

# THE CAUSAL, GENERATIVE MODEL

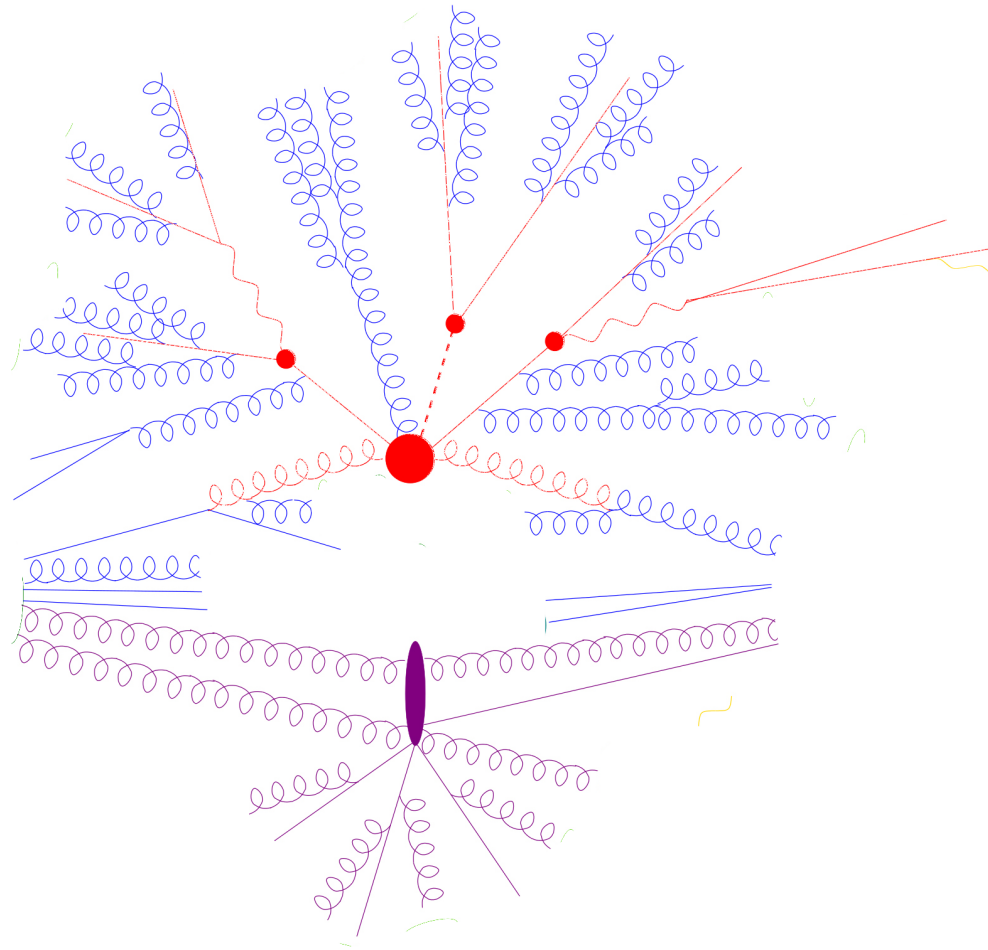
$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$



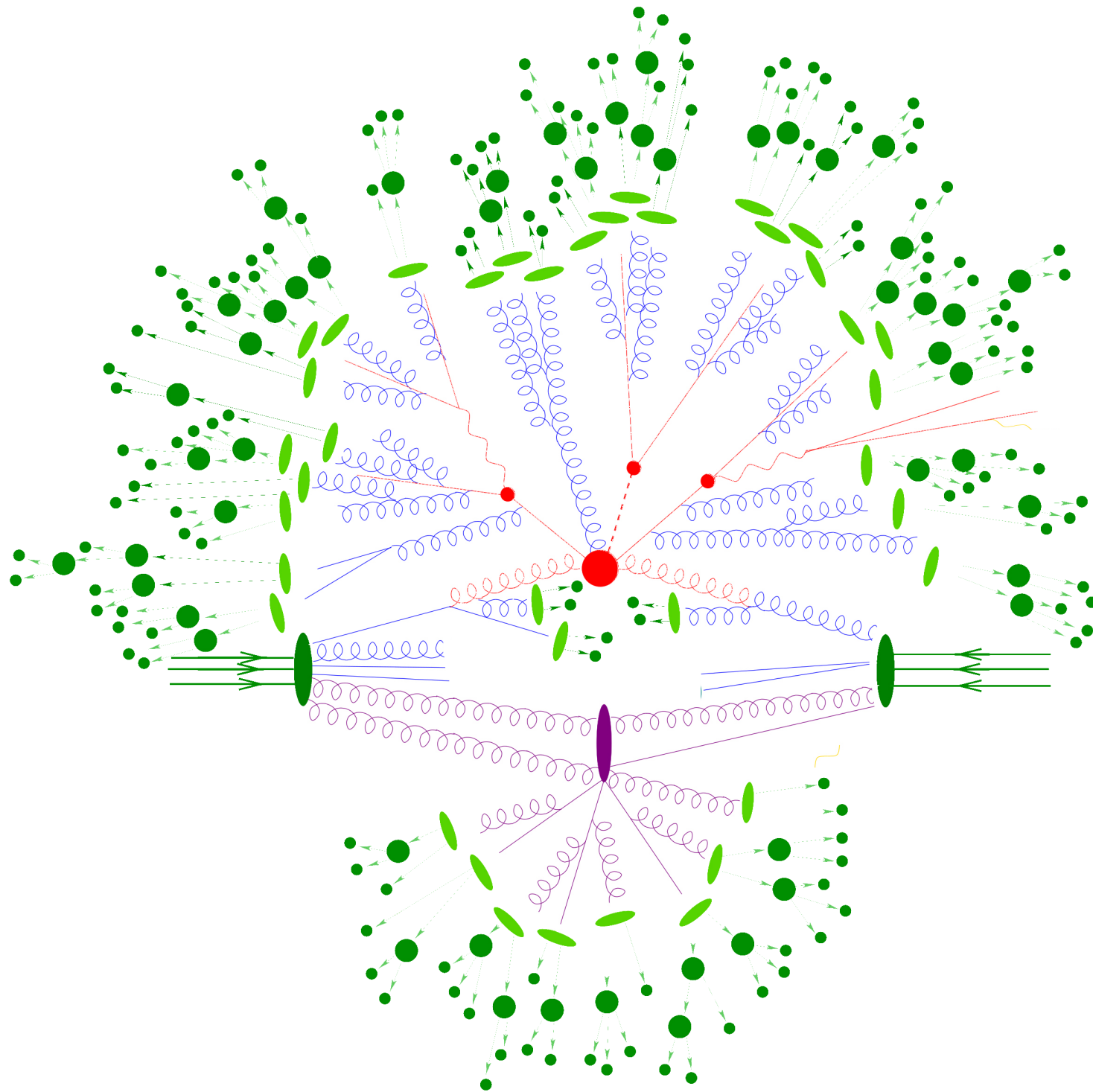


# THE CAUSAL, GENERATIVE MODEL

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$



# THE CAUSAL, GENERATIVE MODEL



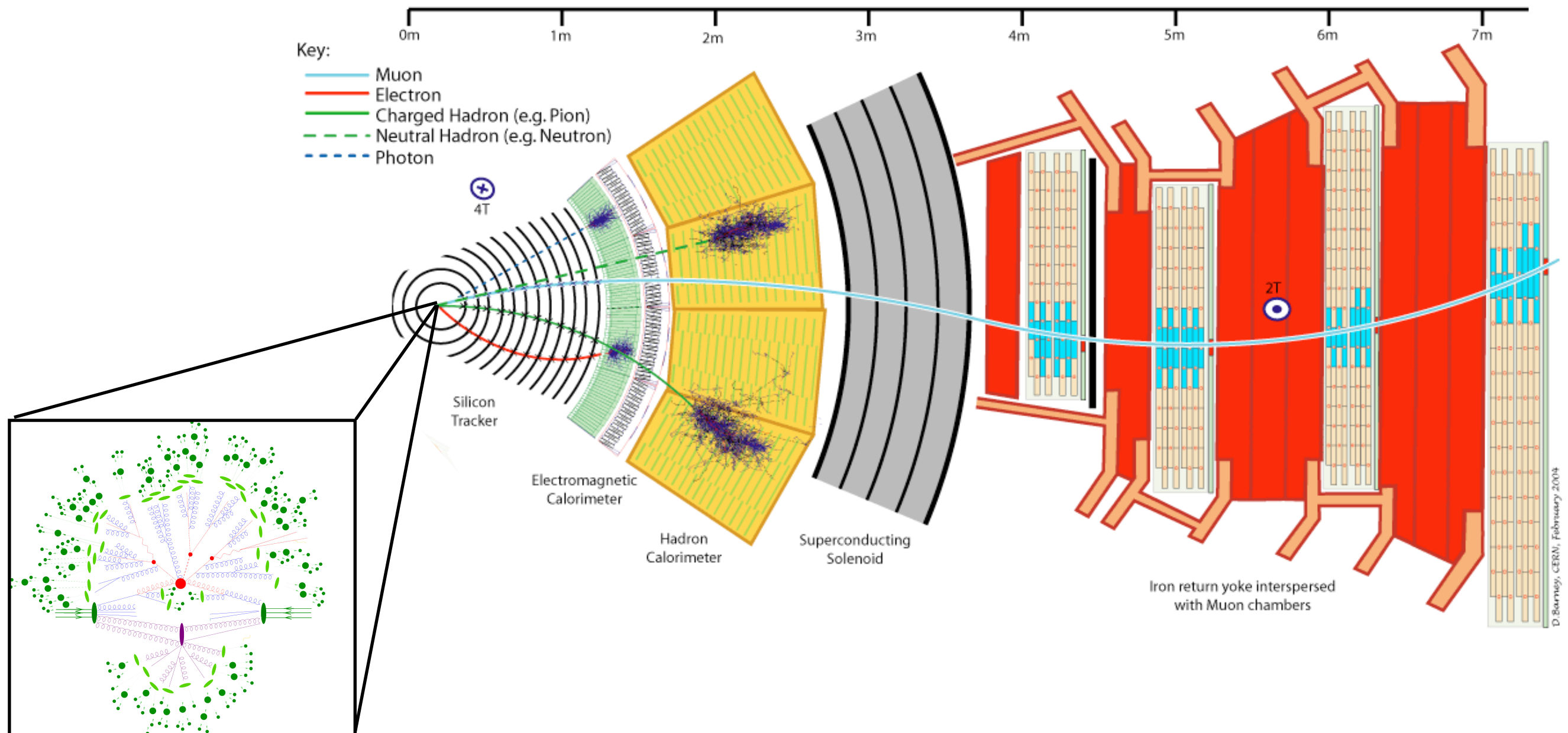
$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

# THE CAUSAL, GENERATIVE MODEL

**Conceptually:**  $\text{Prob}(\text{detector response} \mid \text{particles})$

**Implementation:** Monte Carlo integration over micro-physics

**Consequence:** evaluation of the likelihood is intractable



# GALTON BOARD

Say we want to infer  $\theta$ , the probability to bounce right



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

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



# GALTON BOARD

Say we want to infer  $\theta$ , the probability to bounce right



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

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

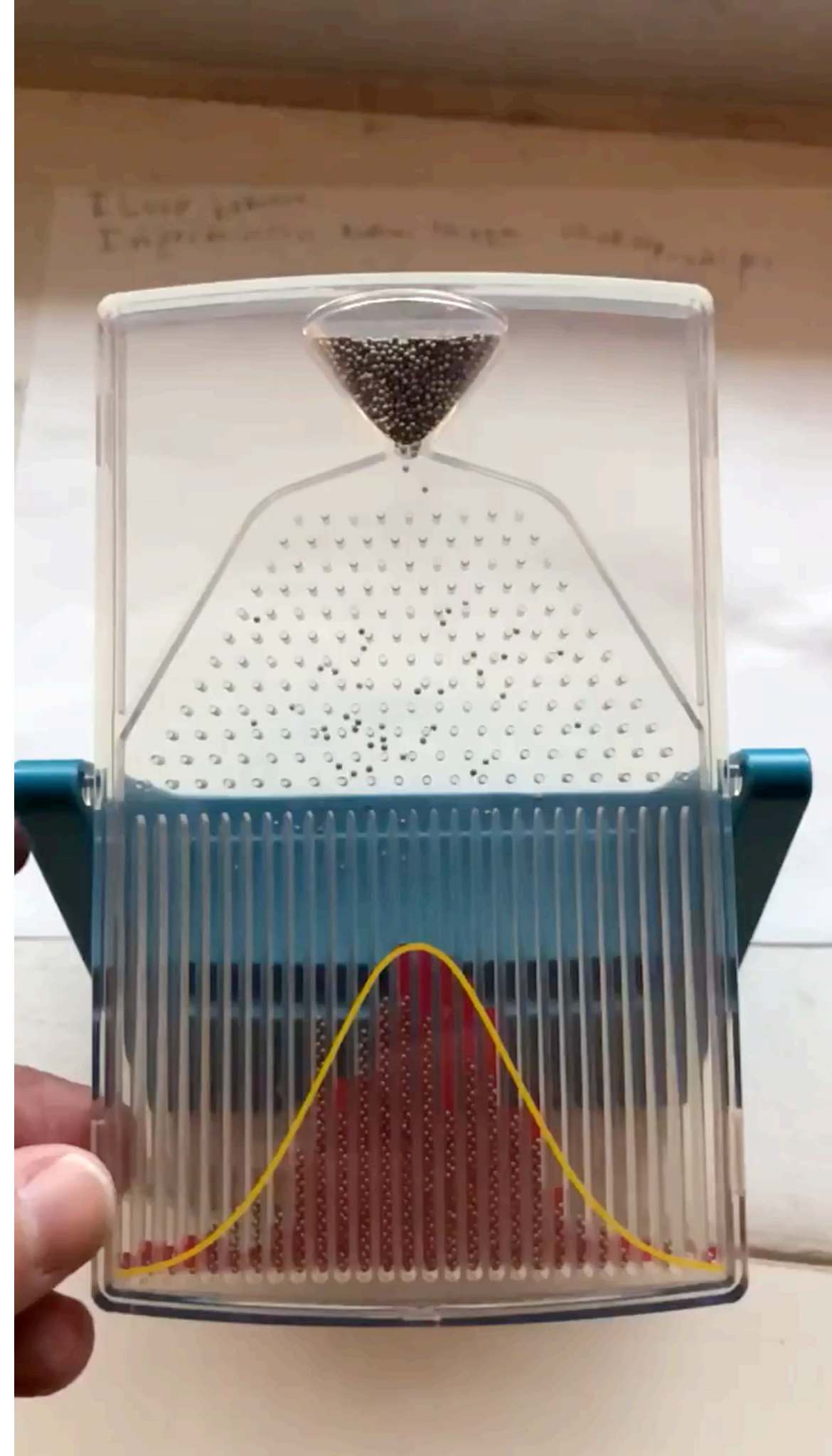
UH OH!

The actual situation is much more complicated.

It's not a Binomial distribution!

What is it?

I have no idea, but I could simulate it!



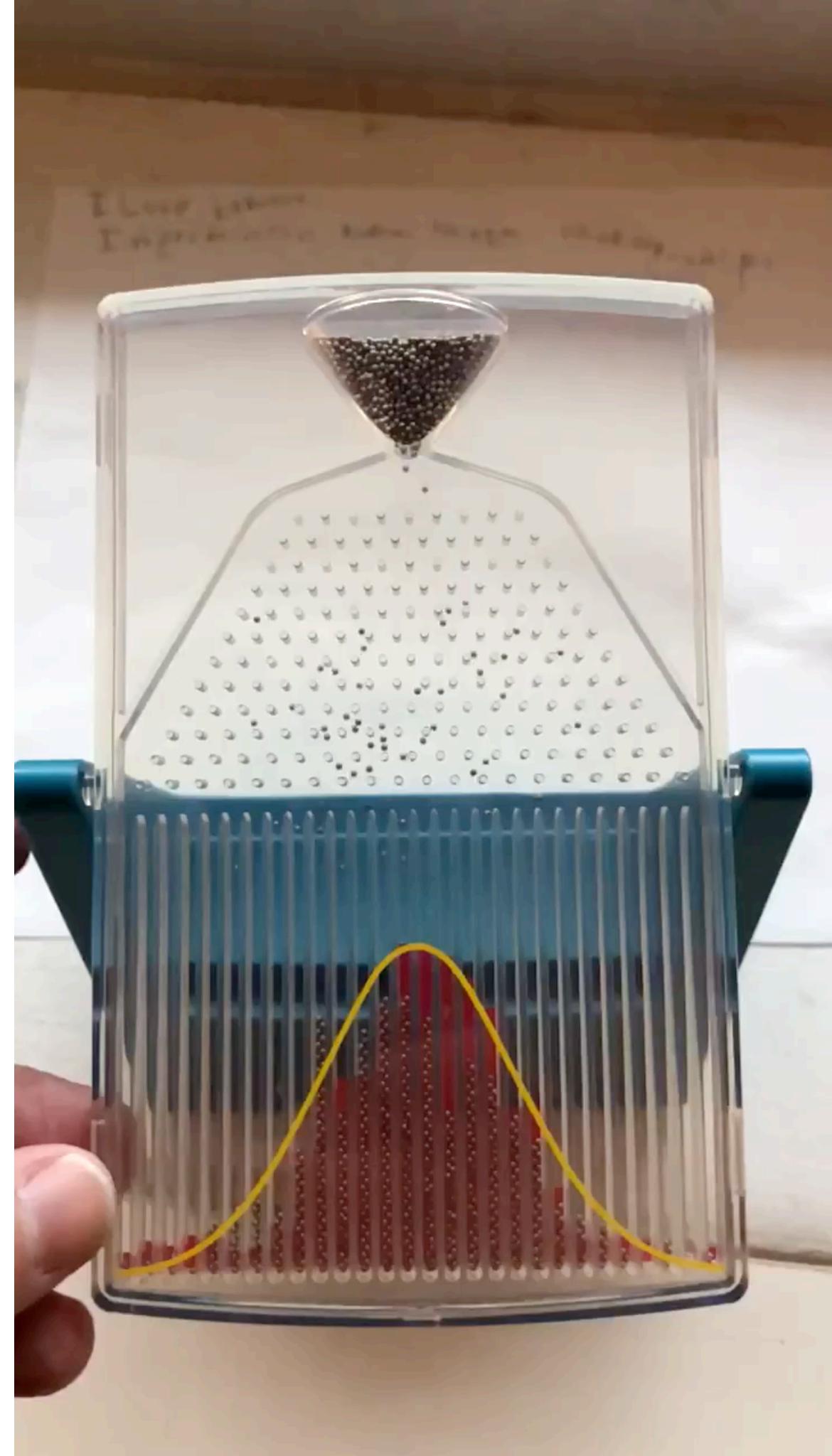
UH OH!

The actual situation is much more complicated.

It's not a Binomial distribution!

What is it?

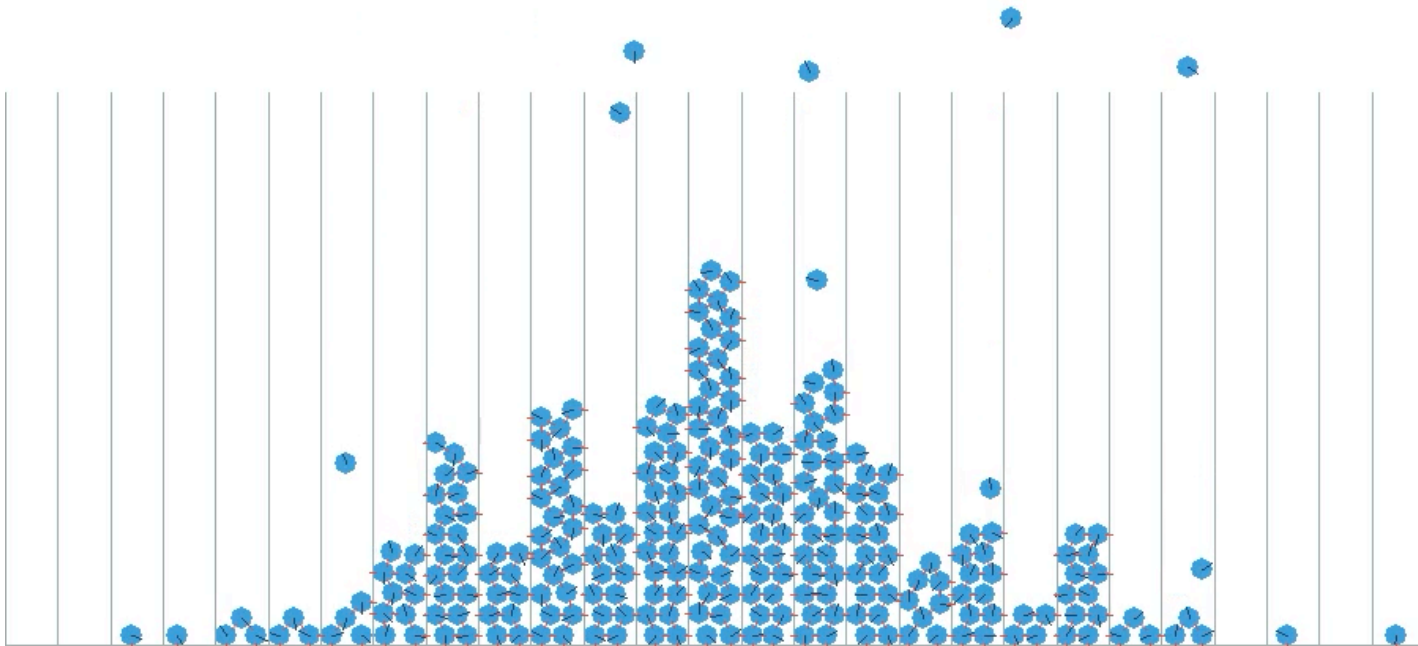
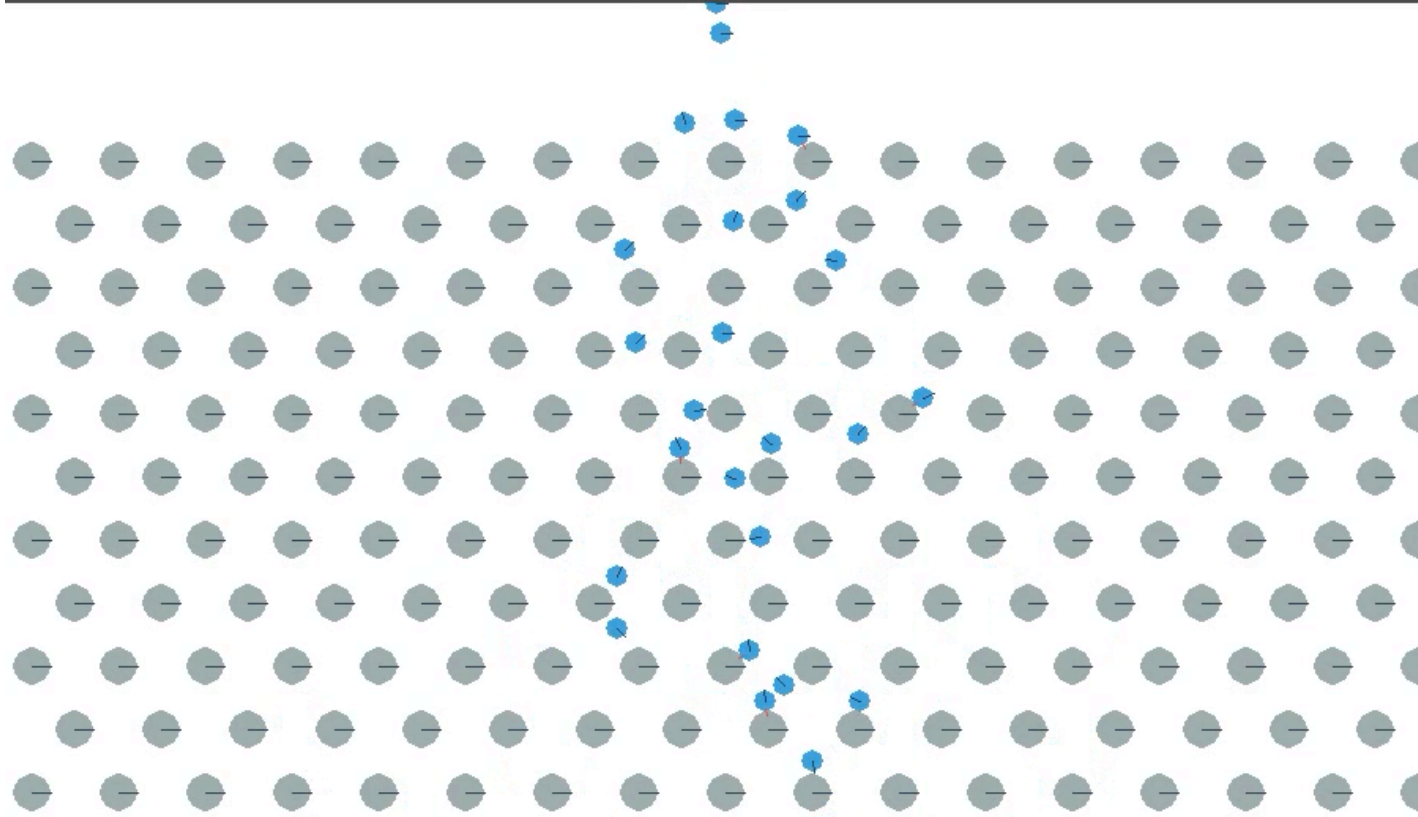
I have no idea, but I could simulate it!



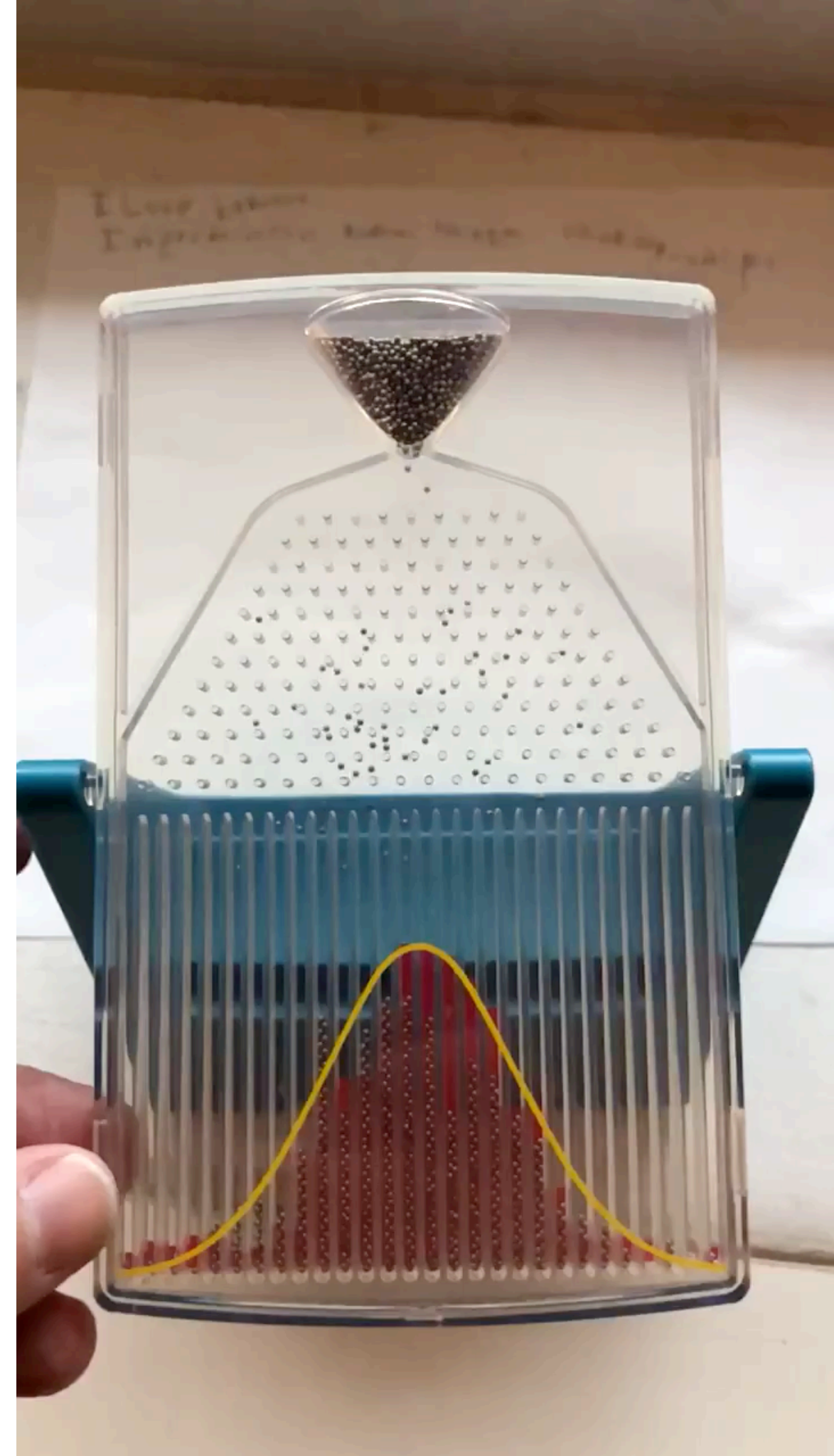


UH OH!

fps: 63.3, balls: 298



ANIMATION BY ATILIM GÜNEŞ BAYDIN

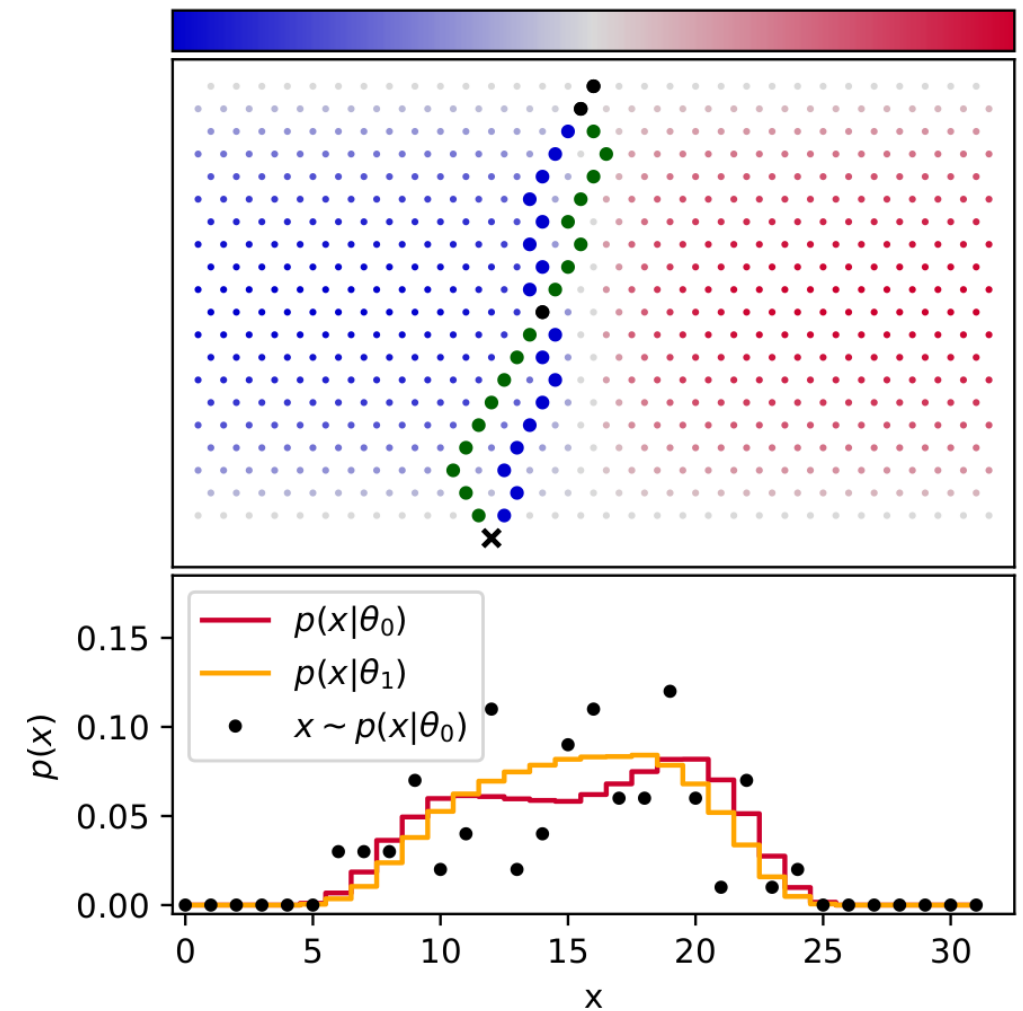




# AN INTRACTABLE INTEGRAL

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

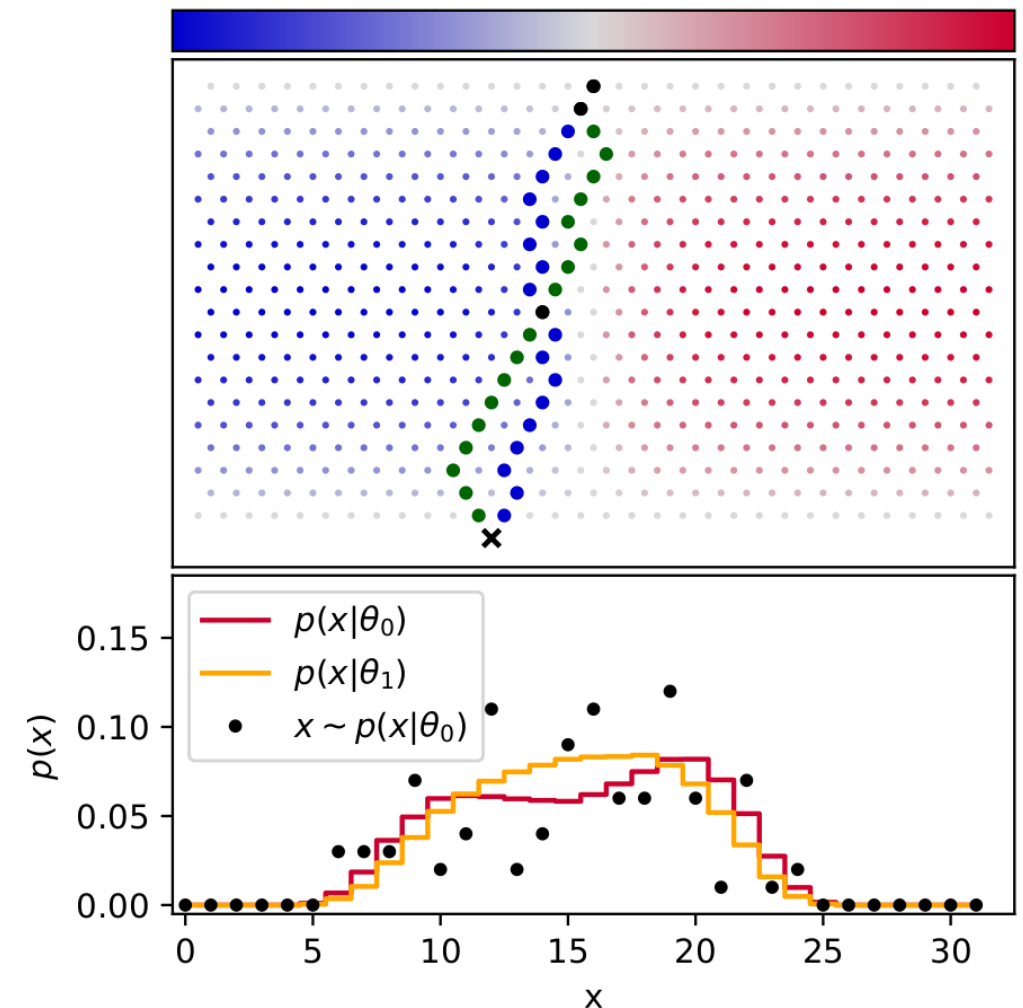
$$p(x|\theta) = \int p(x, z|\theta) dz$$



# AN INTRACTABLE INTEGRAL

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

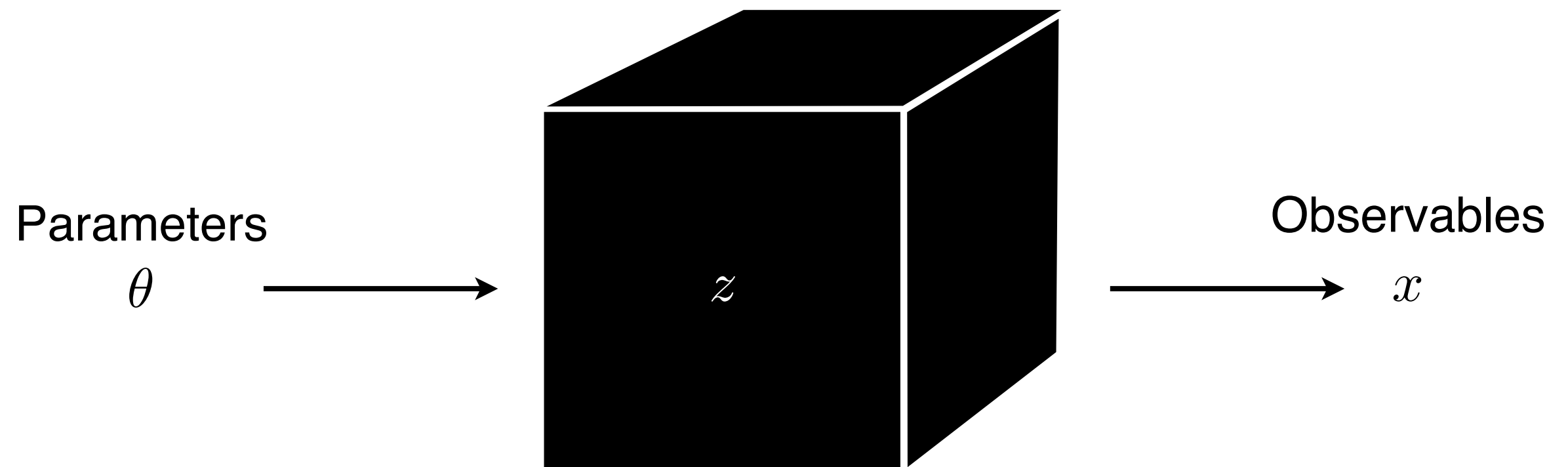
$$p(x|\theta) = \int p(x, z|\theta) dz$$



- 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.

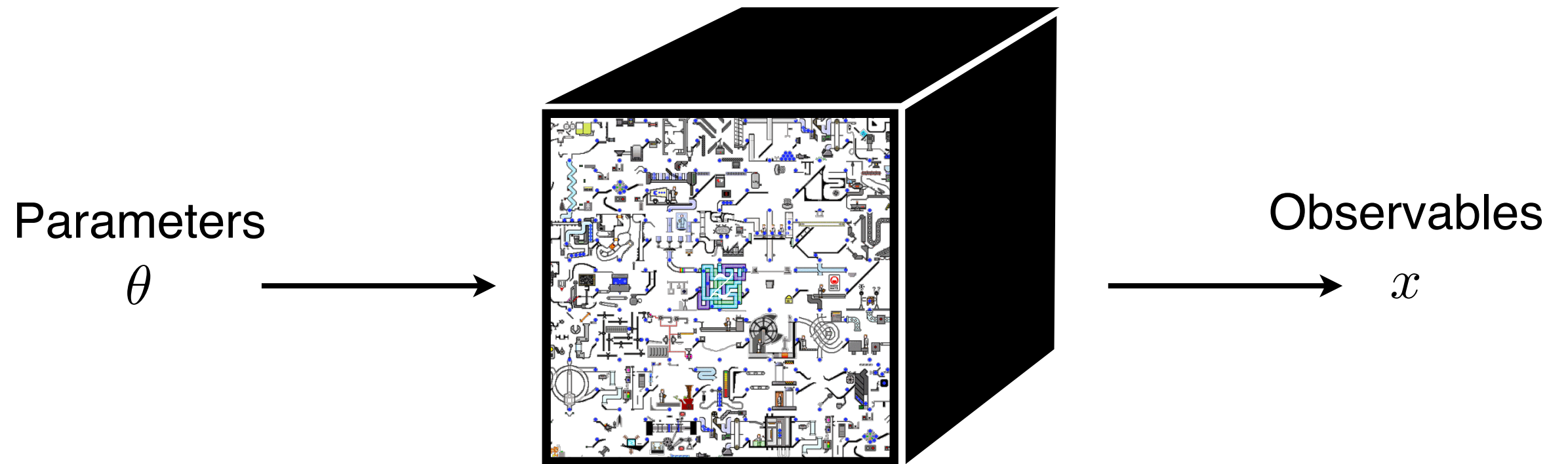
Since  $p(x|\theta)$  cannot be evaluated, does this mean inference is no longer possible?

# LIKELIHOOD-FREE INFERENCE



- 
- Prediction (simulation):
- Well-understood mechanistic model
  - Simulator can generate samples

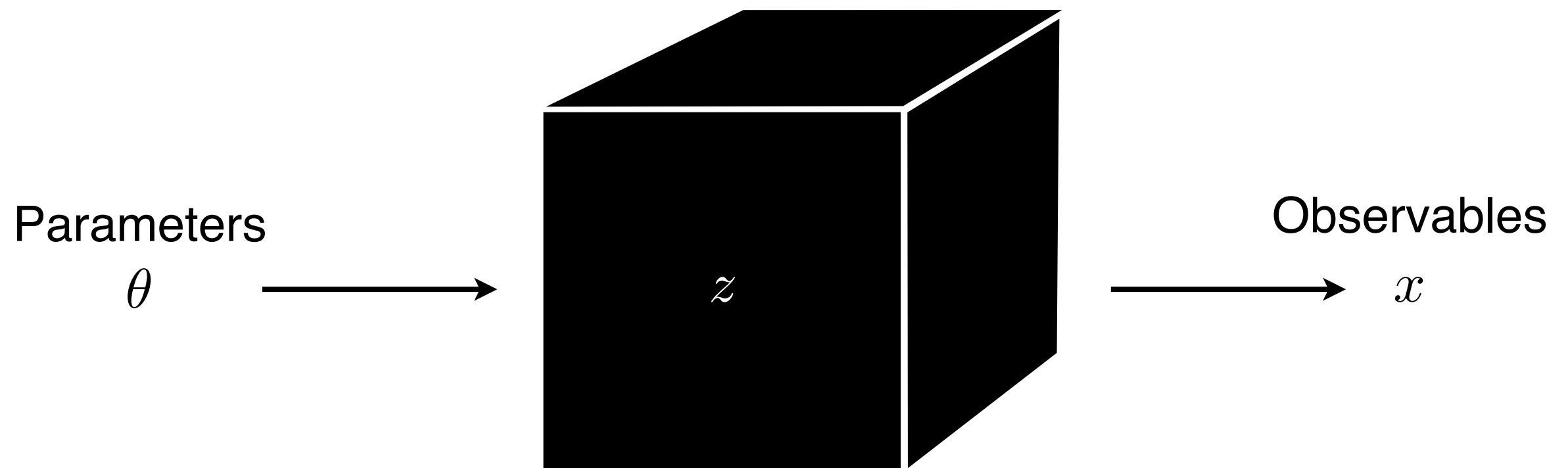
# LIKELIHOOD-FREE INFERENCE



- 
- Prediction (simulation):
- Well-understood mechanistic model
  - Simulator can generate samples



# LIKELIHOOD-FREE INFERENCE



- 
- Prediction (simulation):
- Well-understood mechanistic model
  - Simulator can generate samples

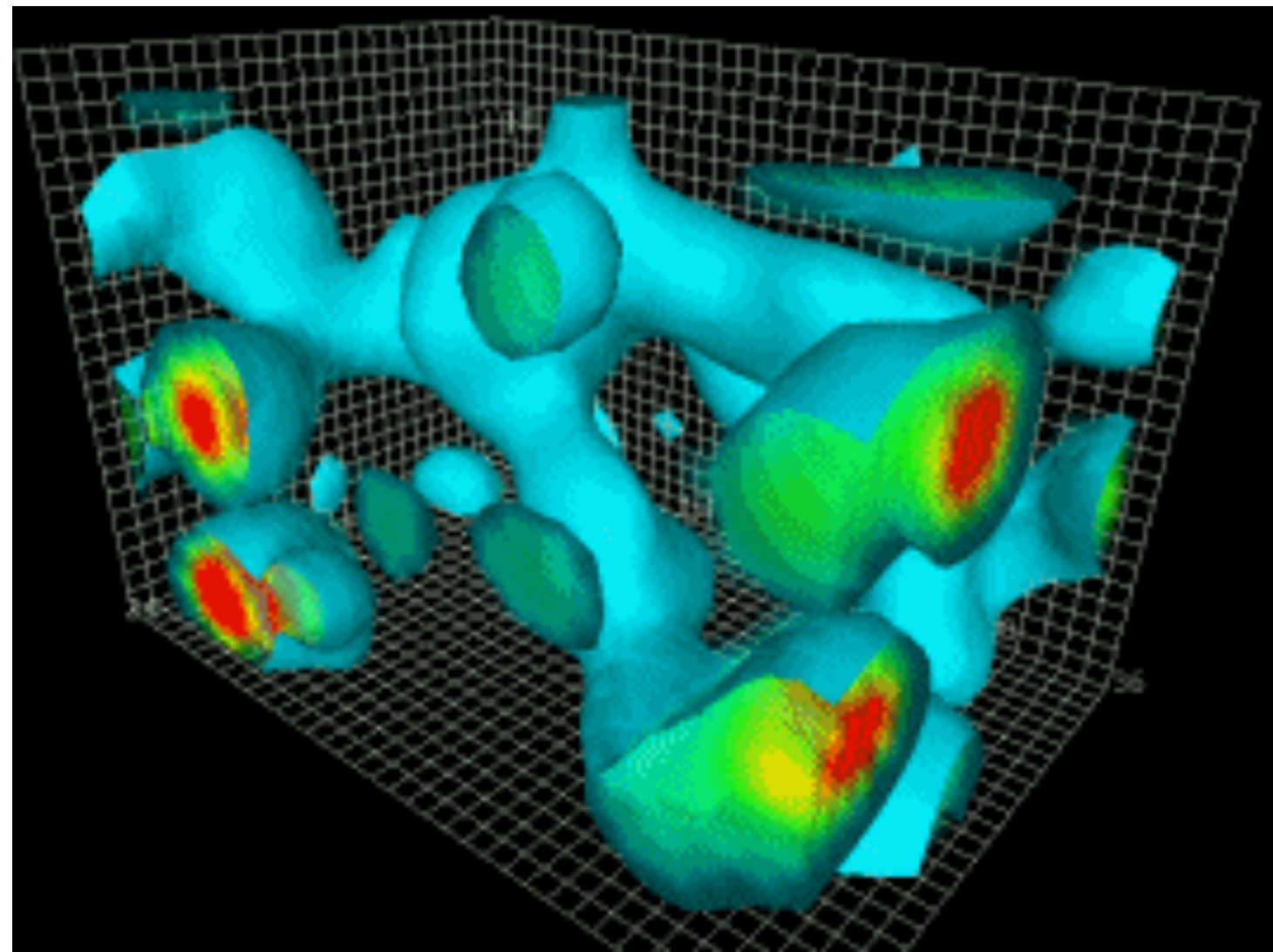
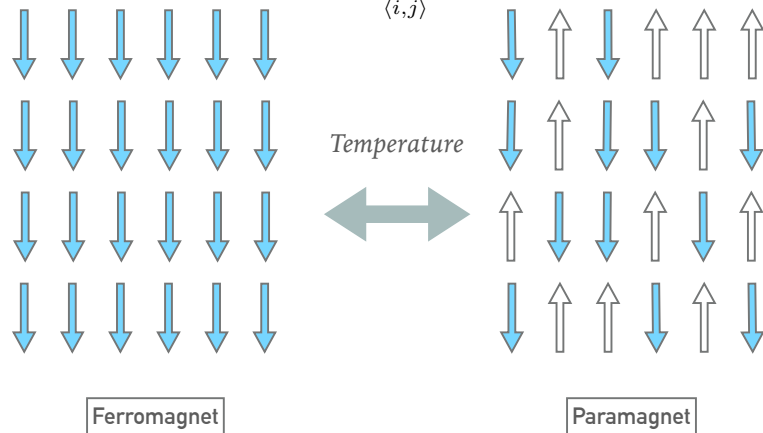
- 
- Inference:
- Likelihood function  $p(x|\theta)$  is intractable
  - Goal: estimator  $\hat{p}(x|\theta)$

# LATTICE FIELD THEORY

## PHASES, PHASE TRANSITIONS, AND THE ORDER PARAMETER

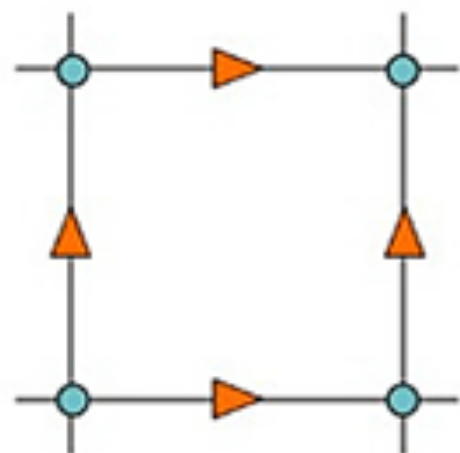
Ising ferromagnet in two dimensions

$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j$$

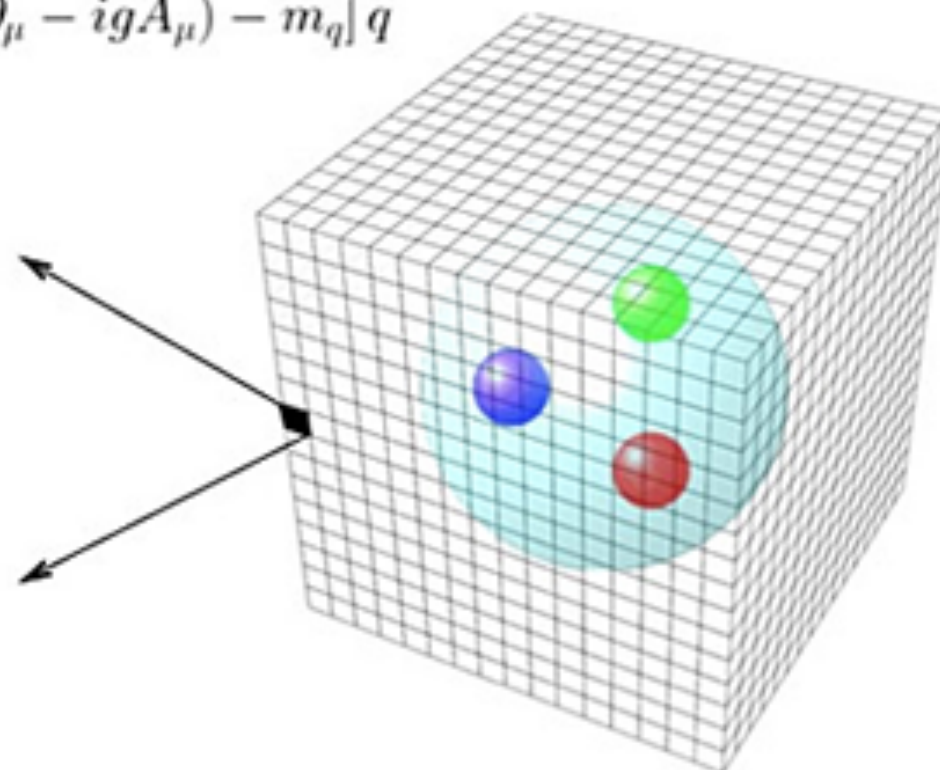


## QCD Lagrangian

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \sum_{q=u,d,s,c,b,t} \bar{q} [i\gamma^\mu (\partial_\mu - igA_\mu) - m_q] q$$



● quark    ▲ gluon

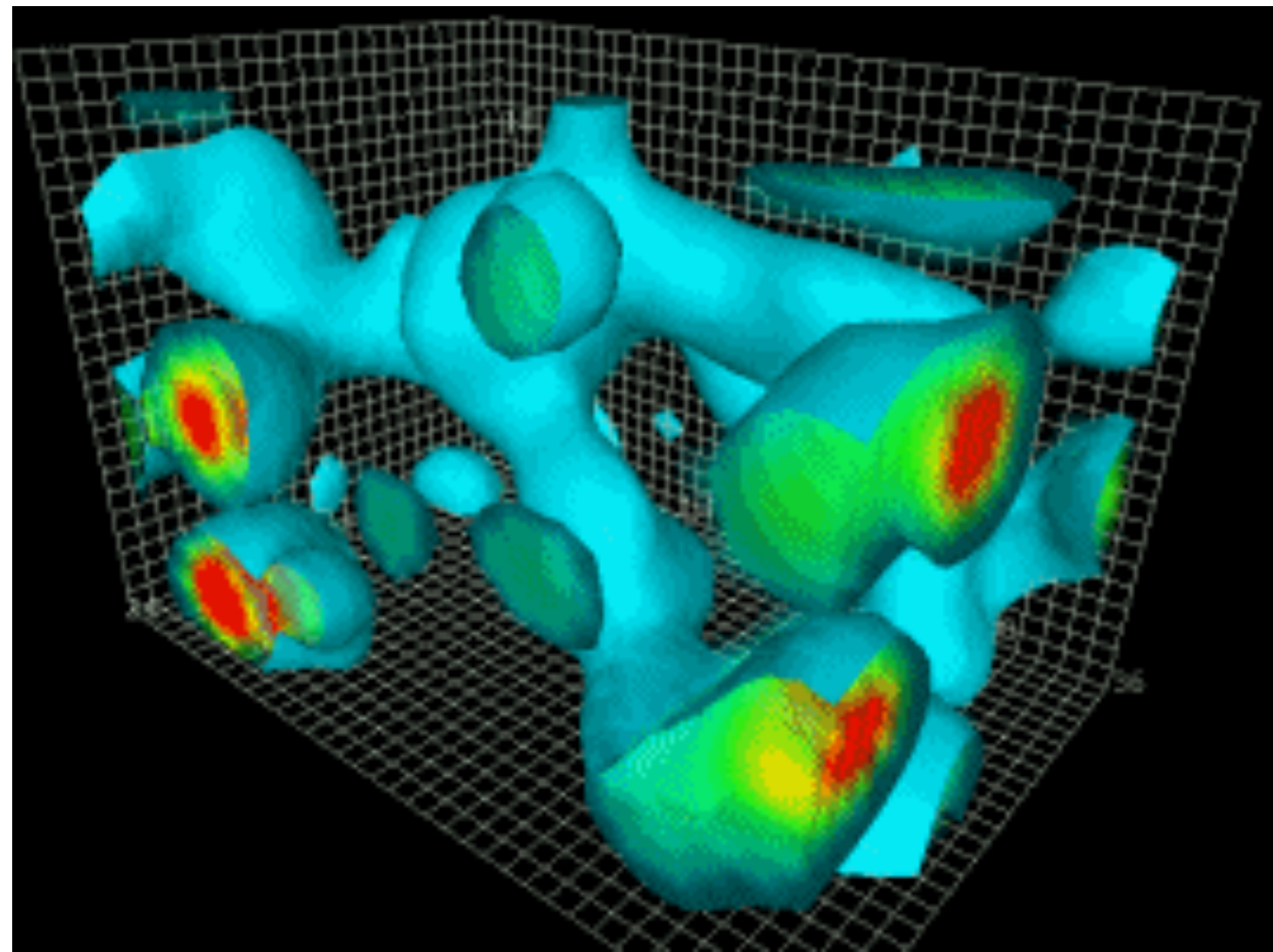
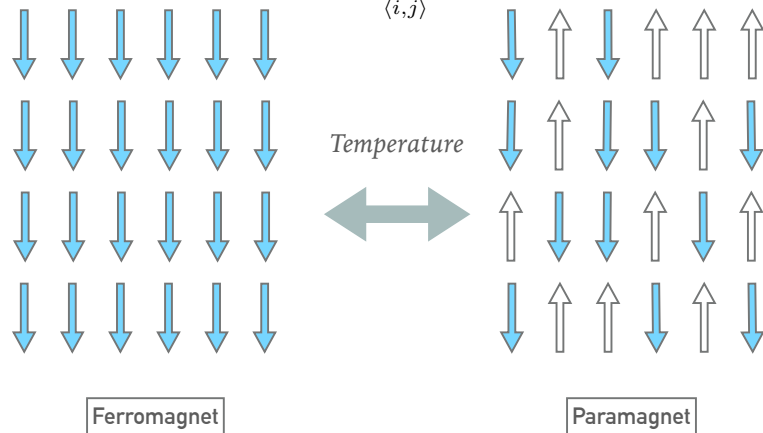


# LATTICE FIELD THEORY

## PHASES, PHASE TRANSITIONS, AND THE ORDER PARAMETER

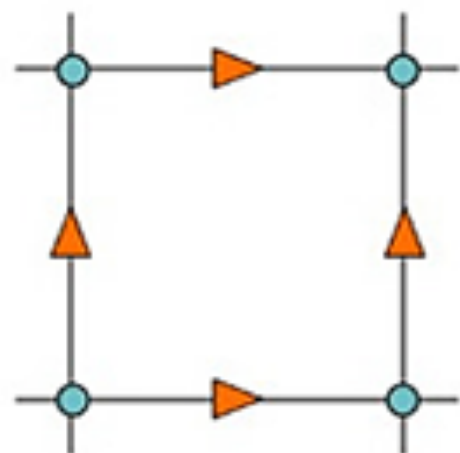
Ising ferromagnet in two dimensions

$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j$$

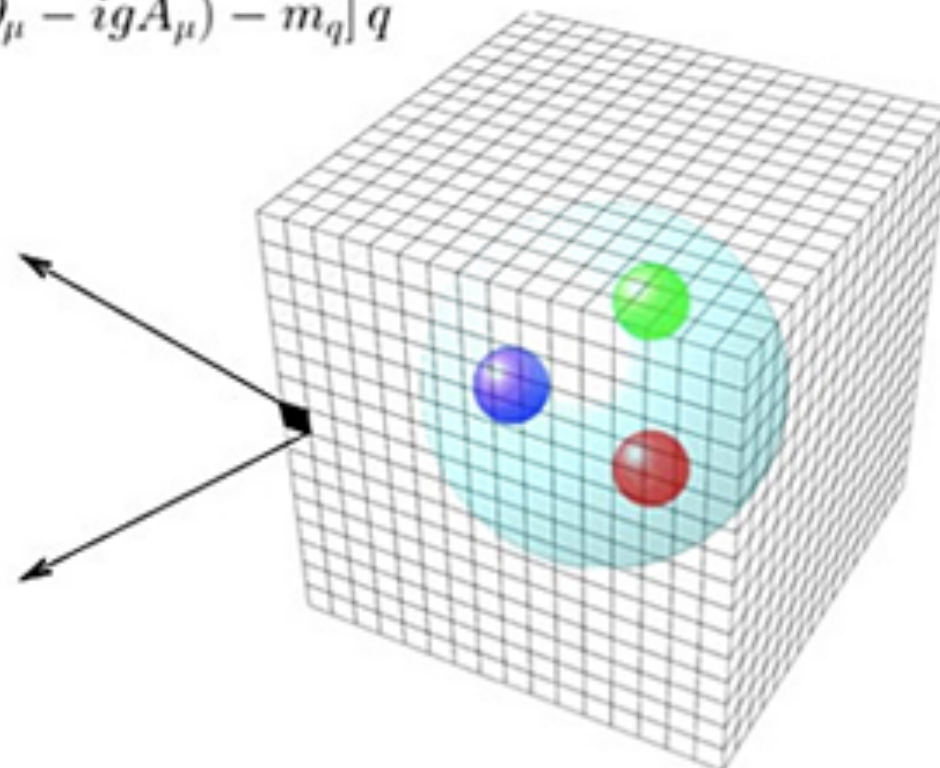


## QCD Lagrangian

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \sum_{q=u,d,s,c,b,t} \bar{q} [i\gamma^\mu (\partial_\mu - igA_\mu) - m_q] q$$

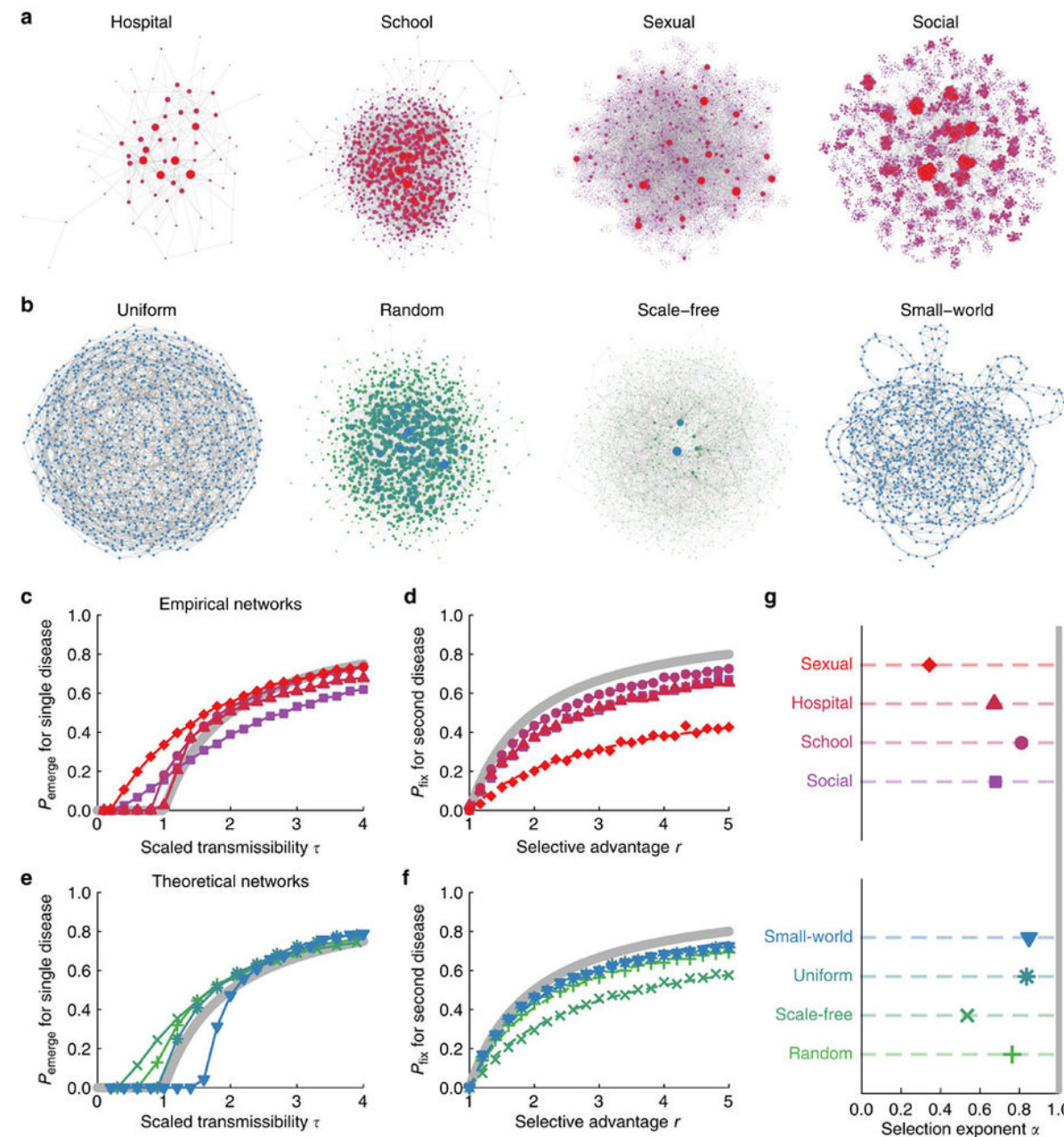
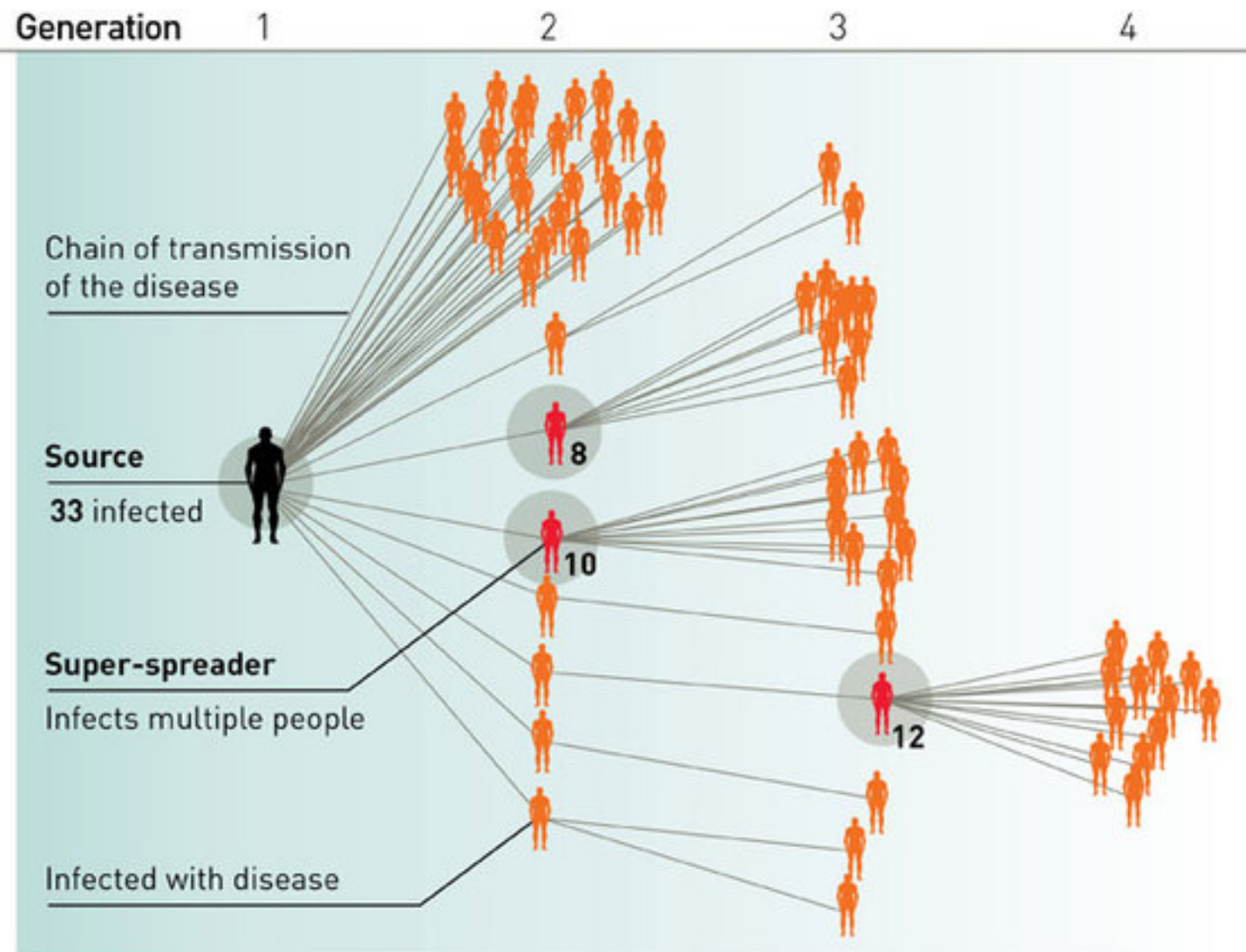


● quark    ▲ gluon





# EPIDEMIOLOGY & POPULATION GENETICS



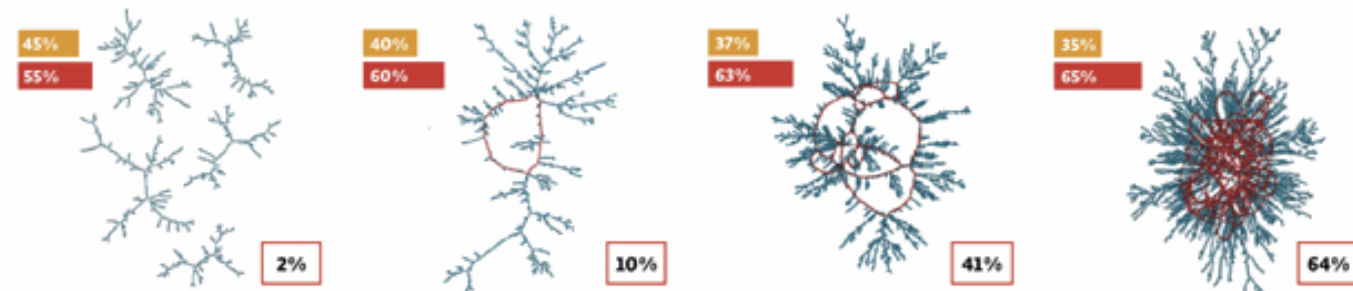
## Small Change, Big Effects

**KEY**  
1 partner  
2 or 3 partners

Percent of people that are connected in the network through their sexual partnerships

Modest variations in the concurrency rate—the proportion of people in overlapping sexual partnerships—can have a dramatic effect on a population's vulnerability to HIV.

When the concurrency rate is 55%, only 2% of this population is connected to the broader sexual network required for HIV transmission (top). But when concurrency reaches 65%, an astonishing 64% of the population is vulnerable, even though the number of sexual partners remains constant.



Source: Morris, et al. The Relationship Between Concurrent Partnerships and HIV Transmission, 2008. See [www.aidsstar-one.com/](http://www.aidsstar-one.com/).



# TAXONOMY FOR SIMULATION

**Deterministic:** fluid mechanics, quantum state evolution, ODEs and PDEs

- Often differentiable (at least in principle)

**Stochastic:** statistical physics (Ising model, etc.); particle scattering process, ...

- Non-differentiable elements due to probabilistic control flow (eg. if/then/else conditions)

**Measurement noise:** may or may not be included

- eg. Use of ML for theoretical physics often treats system as if it can be exactly, directly observed

# A COMMON THEME, A COMMON LANGUAGE

## ABC

resources on approximate  
Bayesian computational  
methods

 Search

Home

## Home

This website keeps track of developments in approximate Bayesian computation (ABC) (a.k.a. likelihood-free), a class of computational statistical methods for Bayesian inference under intractable likelihoods. The site is meant to be a resource both for biologists and statisticians who want to learn more about ABC and related methods. Recent publications are under Publications 2012. A comprehensive list of publications can be found under Literature. If you are unfamiliar with ABC methods see the Introduction. Navigate using the menu to learn more.

[ABC in Montreal](#)

[ABC in Montreal \(2014\)](#)

## ABC in Montreal

Approximate Bayesian computation (ABC) or likelihood-free (LF) methods have developed mostly beyond the radar of the machine learning community, but are important tools for a large and diverse segment of the scientific community. This is particularly true for systems and population biology, computational neuroscience, computer vision, healthcare sciences, but also many others.

Interaction between the ABC and machine learning community has recently started and contributed to important advances. In general, however, there is still significant room for more intense interaction and collaboration. Our workshop aims at being a place for this to happen.

# Markov chain Monte Carlo without likelihoods

Paul Marjoram\*, John Molitor\*, Vincent Plagnol†, and Simon Tavaré†‡

\*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and †Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

**Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.**

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data  $\mathcal{D}$  generated from a model  $\mathcal{M}$  determined by parameters  $\theta$ , the prior density of which is denoted by  $\pi(\theta)$ . We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is  $f(\theta|\mathcal{D})$ , which is given by

$$f(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)/\mathbb{P}(\mathcal{D}), \quad [1]$$

where  $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta$  is the normalizing constant.

In most scientific contexts, explicit formulae for such posterior densities are few and far between, and we usually resort to stochastic simulation to generate observations from  $f$ . Perhaps the simplest approach for this is the rejection method:

- A1. Generate  $\theta$  from  $\pi(\cdot)$ .
- A2. Accept  $\theta$  with probability  $h = \mathbb{P}(\mathcal{D}|\theta)$ ; return to A1.

of  $\varepsilon$  therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given  $\rho$  and  $\varepsilon$ , we are generating independent and identically distributed observations from  $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$ .

When  $\mathcal{D}$  is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of  $\mathcal{D}'$  with  $\mathcal{D}$  can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics  $S = (S_1, \dots, S_p)$  is sufficient for  $\theta$ , in that  $\mathbb{P}(\mathcal{D}|S, \theta)$  is independent of  $\theta$ , then  $f(\theta|\mathcal{D}) = f(\theta|S)$ . The normalizing constant  $\mathbb{P}(S)$  is typically larger than  $\mathbb{P}(\mathcal{D})$ , resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about  $\theta$ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data  $\mathcal{D}$  summarized by  $S$ :

- D1. Generate  $\theta$  from  $\pi(\cdot)$ .
- D2. Simulate  $\mathcal{D}'$  from stochastic model  $\mathcal{M}$  with parameter  $\theta$ , and compute the corresponding statistics  $S'$ .
- D3. Calculate the distance  $\rho(S, S')$  between  $S$  and  $S'$ .
- D4. Accept  $\theta$  if  $\rho \leq \varepsilon$ , and return to D1.

There are several advantages to these rejection methods, among them the fact that they are usually easy to code, they generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-

# Markov chain Monte Carlo without likelihoods

Paul Marjoram\*, John Molitor\*, Vincent Plagnol†, and Simon Tavaré†‡

\*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and †Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

- D1. Generate  $\theta$  from  $\pi(\cdot)$ .
- D2. Simulate  $\mathcal{D}'$  from stochastic model  $\mathcal{M}$  with parameter  $\theta$ , and compute the corresponding statistics  $S'$ .
- D3. Calculate the distance  $\rho(S, S')$  between  $S$  and  $S'$ .
- D4. Accept  $\theta$  if  $\rho \leq \varepsilon$ , and return to D1.

## discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data  $\mathcal{D}$  generated from a model  $\mathcal{M}$  determined by parameters  $\theta$ , the prior density of which is denoted by  $\pi(\theta)$ . We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is  $f(\theta|\mathcal{D})$ , which is given by

$$f(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)/\mathbb{P}(\mathcal{D}), \quad [1]$$

where  $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta$  is the normalizing constant.

In most scientific contexts, explicit formulae for such posterior densities are few and far between, and we usually resort to stochastic simulation to generate observations from  $f$ . Perhaps the simplest approach for this is the rejection method:

- A1. Generate  $\theta$  from  $\pi(\cdot)$ .
- A2. Accept  $\theta$  with probability  $h = \mathbb{P}(\mathcal{D}|\theta)$ ; return to A1.

typically larger than  $\mathbb{P}(\mathcal{D})$ , resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about  $\theta$ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data  $\mathcal{D}$  summarized by  $S$ :

- D1. Generate  $\theta$  from  $\pi(\cdot)$ .
- D2. Simulate  $\mathcal{D}'$  from stochastic model  $\mathcal{M}$  with parameter  $\theta$ , and compute the corresponding statistics  $S'$ .
- D3. Calculate the distance  $\rho(S, S')$  between  $S$  and  $S'$ .
- D4. Accept  $\theta$  if  $\rho \leq \varepsilon$ , and return to D1.

There are several advantages to these rejection methods, among them the fact that they are usually easy to code, they generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-



# Markov chain Monte Carlo without likelihoods

Paul Marjoram\*, John Molitor\*, Vincent Plagnol<sup>†</sup>, and Simon Tavaré<sup>†‡</sup>

\*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and <sup>†</sup>Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data  $\mathcal{D}$  generated from a model  $\mathcal{M}$  determined by parameters  $\theta$ , the prior density of which is denoted by  $\pi(\theta)$ . We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is  $f(\theta|\mathcal{D})$ , which is given by

of  $\varepsilon$  therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given  $\rho$  and  $\varepsilon$ , we are generating independent and identically distributed observations from  $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$ .

When  $\mathcal{D}$  is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of  $\mathcal{D}'$  with  $\mathcal{D}$  can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics  $S = (S_1, \dots, S_p)$  is sufficient for  $\theta$ , in that  $\mathbb{P}(\mathcal{D}|S, \theta)$  is independent of  $\theta$ , then  $f(\theta|\mathcal{D}) = f(\theta|S)$ . The normalizing constant  $\mathbb{P}(S)$  is typically larger than  $\mathbb{P}(\mathcal{D})$ , resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about  $\theta$ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data  $\mathcal{D}$  summarized by  $S$ :

When  $\mathcal{D}$  is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of  $\mathcal{D}'$  with  $\mathcal{D}$  can be made by using lower-dimensional summaries of the data. The

- A1. Generate  $\theta$  from  $\pi(\cdot)$ .
- A2. Accept  $\theta$  with probability  $h = \mathbb{P}(\mathcal{D}|\theta)$ ; return to A1.

generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-

# Markov chain Monte Carlo without likelihoods

Paul Marjoram\*, John Molitor\*, Vincent Plagnol<sup>†</sup>, and Simon Tavaré<sup>†‡</sup>

\*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and <sup>†</sup>Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data  $\mathcal{D}$  generated from a model  $\mathcal{M}$  determined by parameters  $\theta$ , the prior density of which is denoted by  $\pi(\theta)$ . We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is  $f(\theta|\mathcal{D})$ , which is given by

of  $\varepsilon$  therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given  $\rho$  and  $\varepsilon$ , we are generating independent and identically distributed observations from  $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$ .

When  $\mathcal{D}$  is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of  $\mathcal{D}'$  with  $\mathcal{D}$  can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics  $S = (S_1, \dots, S_p)$  is sufficient for  $\theta$ , in that  $\mathbb{P}(\mathcal{D}|S, \theta)$  is independent of  $\theta$ , then  $f(\theta|\mathcal{D}) = f(\theta|S)$ . The normalizing constant  $\mathbb{P}(S)$  is typically larger than  $\mathbb{P}(\mathcal{D})$ , resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about  $\theta$ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data  $\mathcal{D}$  summarized by  $S$ :

practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach.

- A1. Generate  $\theta$  from  $\pi(\cdot)$ .
- A2. Accept  $\theta$  with probability  $h = \mathbb{P}(\mathcal{D}|\theta)$ ; return to A1.

generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-

# ICML 2017 Workshop on Implicit Models

## Workshop Aims

Probabilistic models are an important tool in machine learning. They form the basis for models that generate realistic data, uncover hidden structure, and make predictions. Traditionally, probabilistic models in machine learning have focused on prescribed models. Prescribed models specify a joint density over observed and hidden variables that can be easily evaluated. The requirement of a tractable density simplifies their learning but limits their flexibility --- several real world phenomena are better described by simulators that do not admit a tractable density. Probabilistic models defined only via the simulations they produce are called implicit models.

Arguably starting with generative adversarial networks, research on implicit models in machine learning has exploded in recent years. This workshop's aim is to foster a discussion around the recent developments and future directions of implicit models.

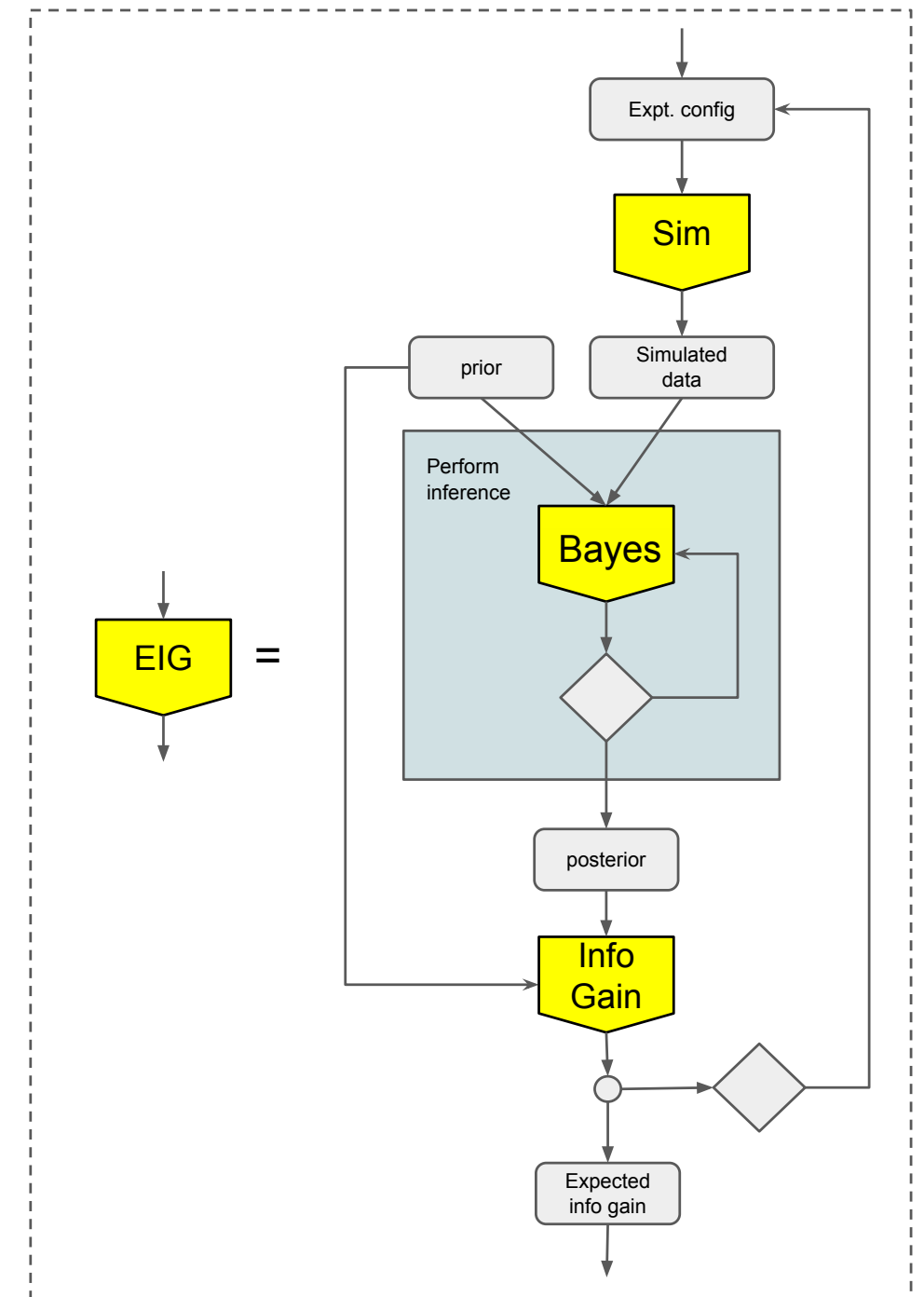
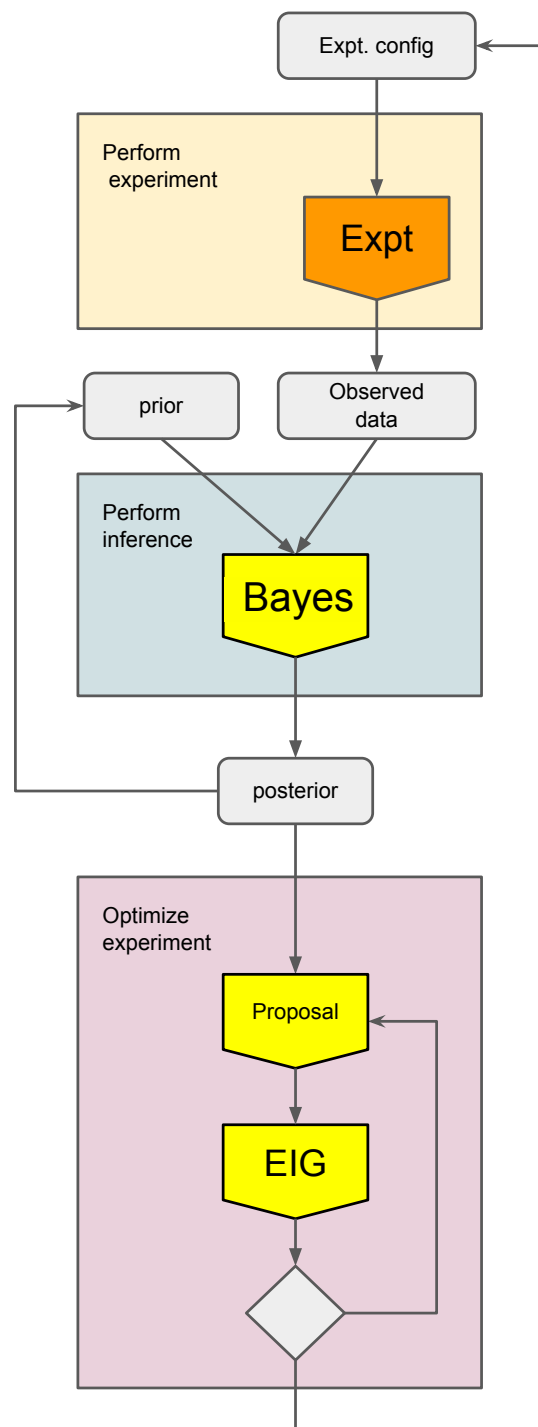
Implicit models have many applications. They are used in ecology where models simulate animal populations over time; they are used in phylogeny, where simulations produce hypothetical ancestry trees; they are used in physics to generate particle simulations for high energy processes. Recently, implicit models have been used to improve the state-of-the-art in image and content generation. Part of the workshop's focus is to discuss the commonalities among applications of implicit models.

Of particular interest at this workshop is to unite fields that work on implicit models. For example:

- **Generative adversarial networks** (a NIPS 2016 workshop) are implicit models with an adversarial training scheme.
- Recent advances in **variational inference** (a NIPS 2015 and 2016 workshop) have leveraged implicit models for more accurate approximations.
- **Approximate Bayesian computation** (a NIPS 2015 workshop) focuses on posterior inference for models with implicit likelihoods.
- Learning implicit models is deeply connected to **two sample testing, density ratio and density difference** estimation.

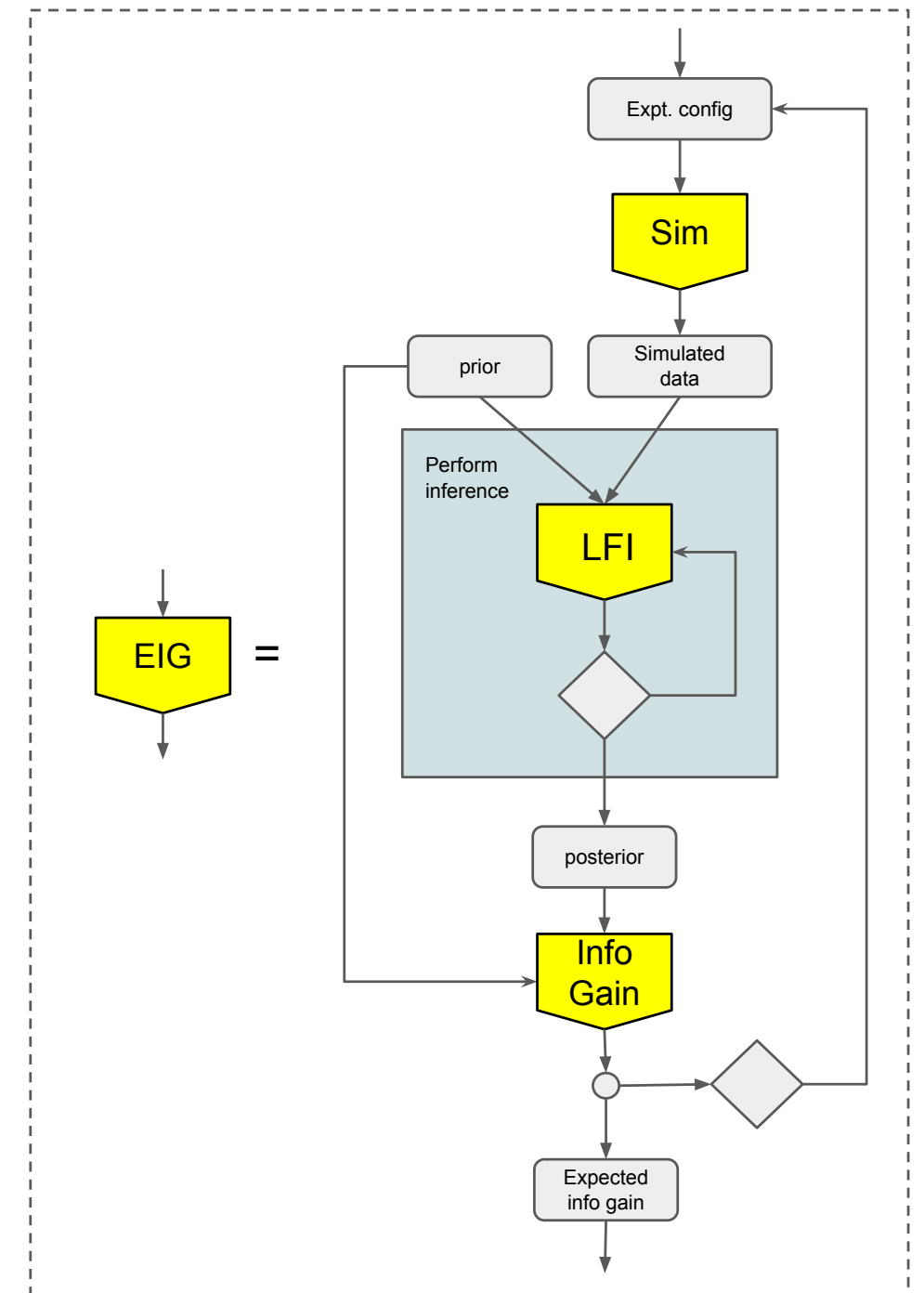
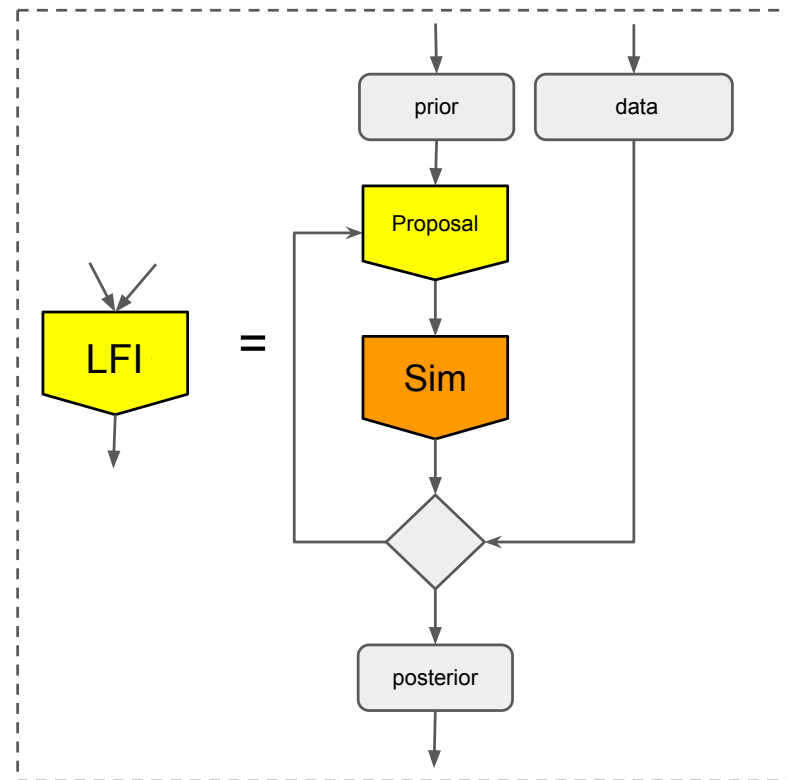
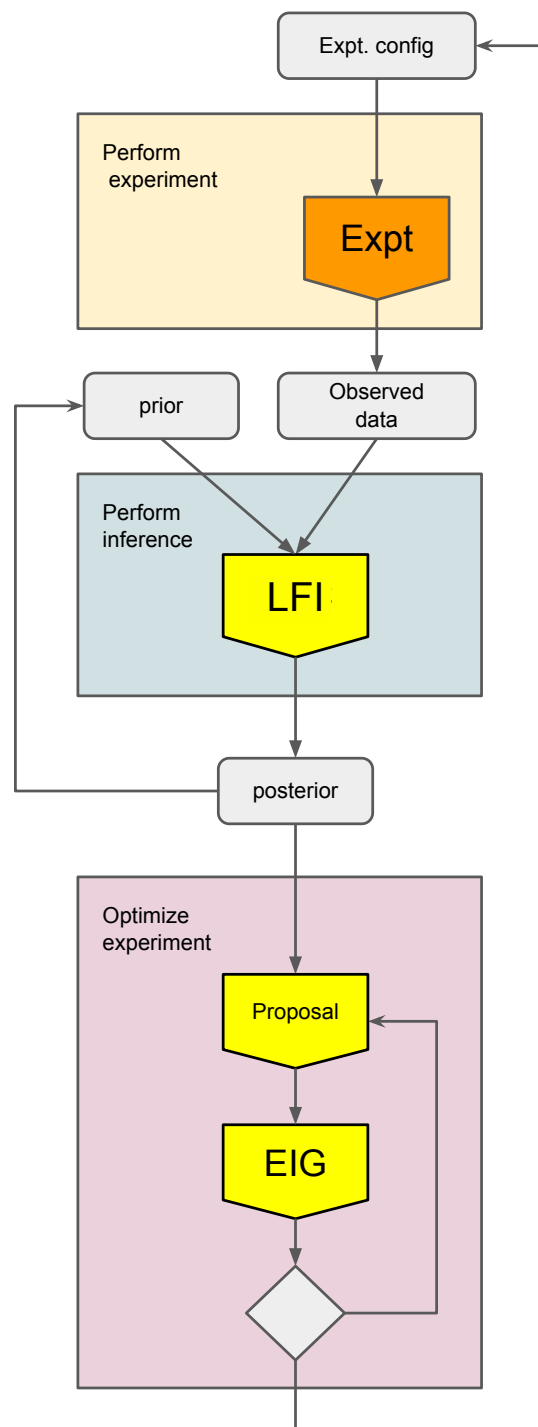
We hope to bring together these different views on implicit models, identifying their core challenges and combining their innovations.

# "ACTIVE SCIENCING"





# "ACTIVE SCIENCING"



# SYNTHESIS

active learning / sequential design / black box optimization



## Active Sciencing



simulation-based /  
likelihood-free  
inference engines

# reana

## Reproducible research data analysis platform

### Flexible

Run many computational workflow engines.



COMMON  
WORKFLOW  
LANGUAGE



### Scalable

Support for remote compute clouds.



kubernetes

### Reusable

Containerise once, reuse elsewhere. Cloud-native.

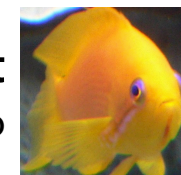


### Free

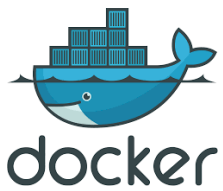
Free Software. MIT licence.  
Made with ❤ at CERN.



The SCAILFIN Project  
[scailfin.github.io](https://scailfin.github.io)



# ENCAPSULATING THE SIMULATION



<https://github.com/lukasheinrich/weinberg-test>

README.md

## Run HEP workflows from the web.

by [Kyle Cranmer](#) and [Lukas Heinrich](#)

An example notebook on how to generate simulated high energy physics collision events using the generator package MadGraph. Simulated datasets obtained from this notebook can then be used to train and evaluate the performance of generative models for physics.

## Usage:

This repository has been equipped with a Dockerfile to encapsulate its software environment. It can be used with the [mybinder](#) service to launch an ephemeral jupyter notebook server to run the notebook.

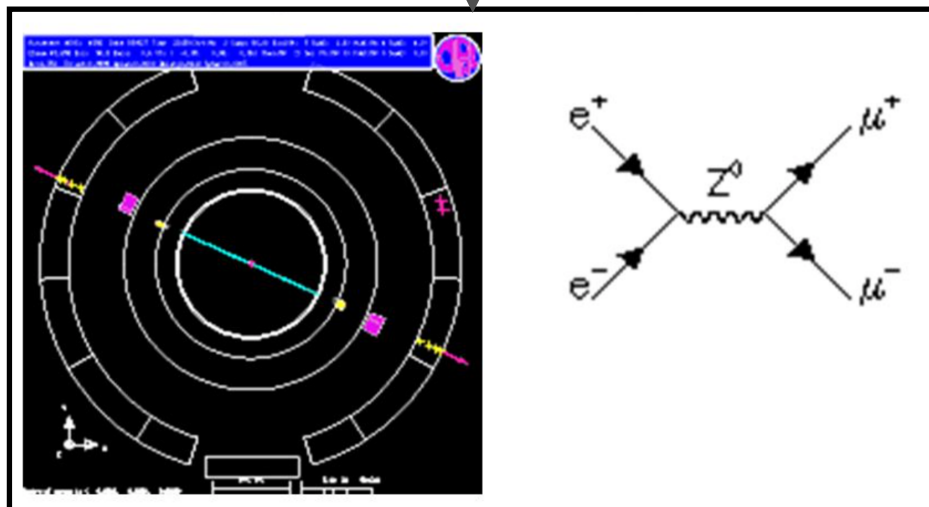
Click on the below badge and open the notebook `adage.ipynb`.

[launch binder](#)

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \tau \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} |(i \partial_\mu - \frac{1}{2} g \tau \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi|^2 - V(\phi)}_{W^\pm, Z, \gamma \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

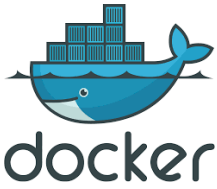
other electroweak parameters. This can be shown with Eq. (2.96), giving

$$A_{FB}^f(s) \simeq A_{FB}^f(m_Z^2) + \frac{(s - m_Z^2)}{s} \frac{3\pi\alpha(s)}{\sqrt{2}G_F m_Z^2} \frac{2Q_e Q_f g_{Ae} g_{Af}}{(g_{Ve}^2 + g_{Ae}^2)(g_{Vf}^2 + g_{Af}^2)} . \quad (8.30)$$





# ENCAPSULATING THE SIMULATION



<https://github.com/lukasheinrich/weinberg-test>

README.md

## Run HEP workflows from the web.

by [Kyle Cranmer](#) and [Lukas Heinrich](#)

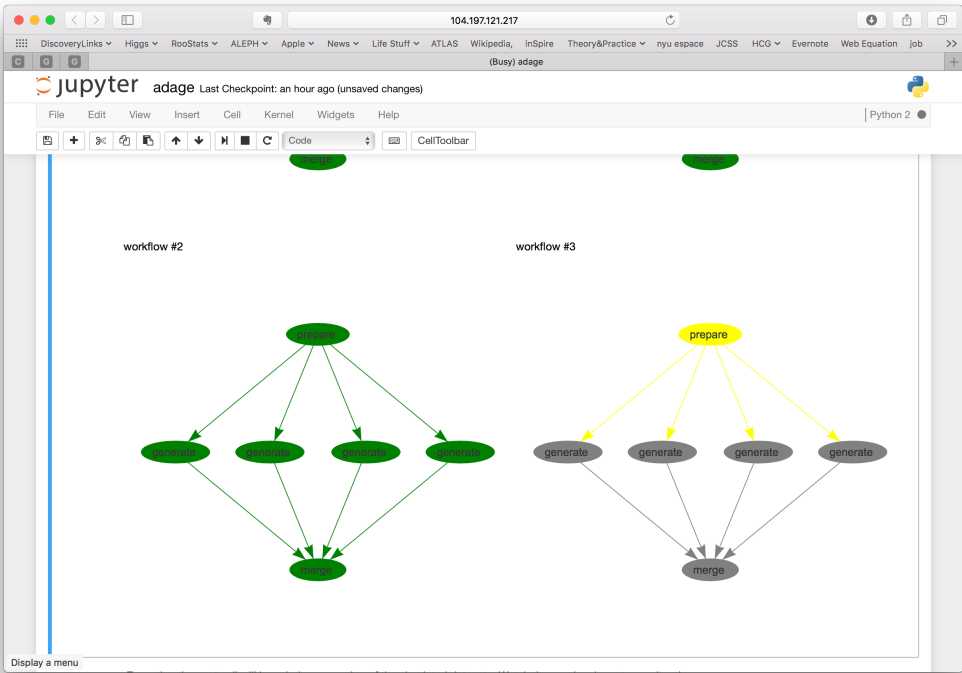
An example notebook on how to generate simulated high energy physics collision events using the generator package MadGraph. Simulated datasets obtained from this notebook can then be used to train and evaluate the performance of generative models for physics.

### Usage:

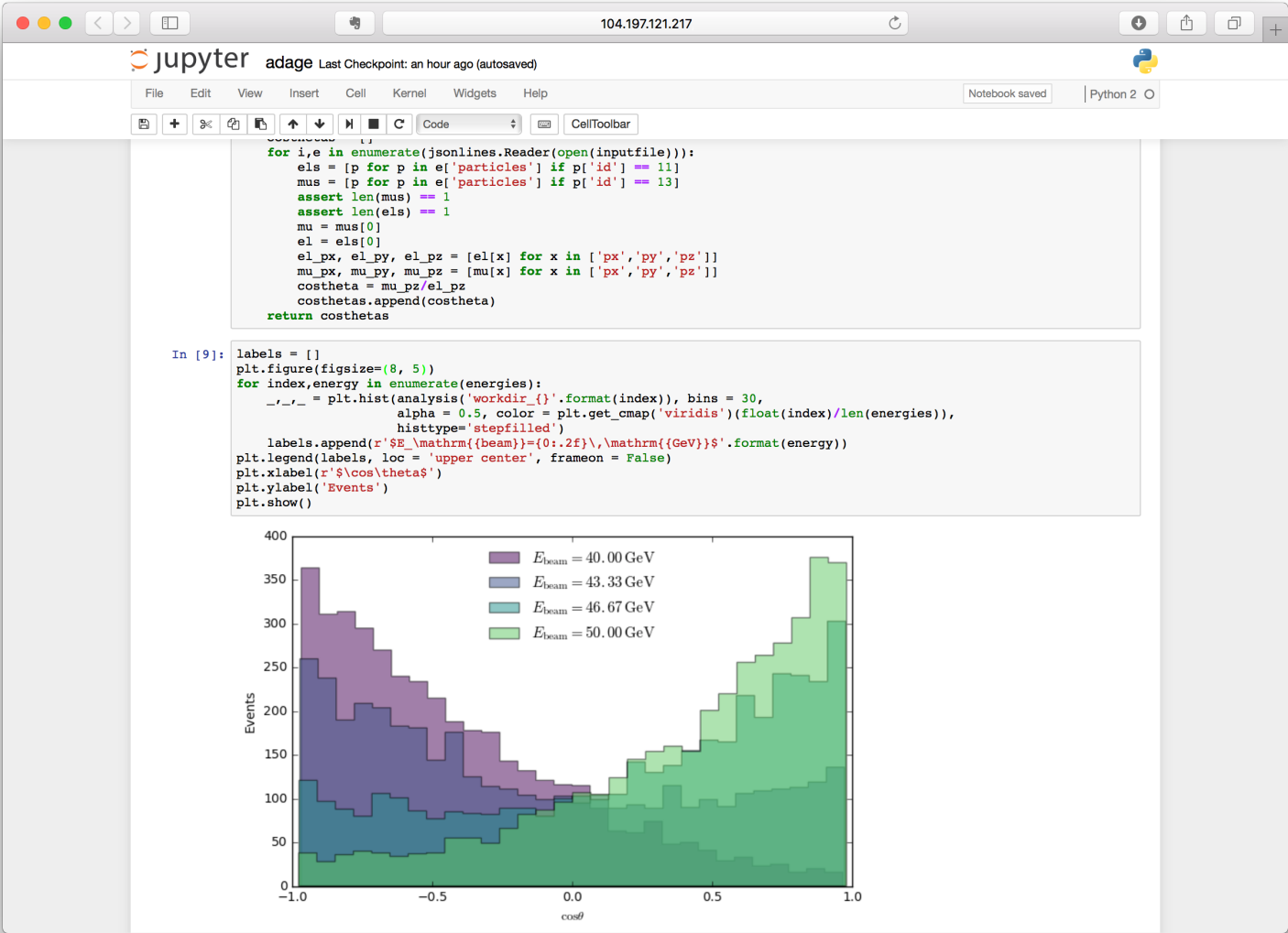
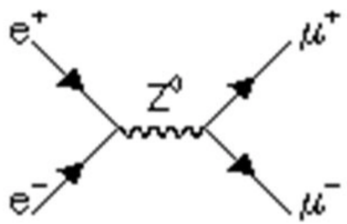
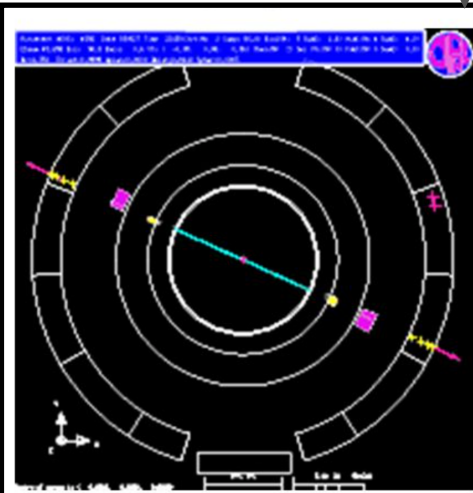
This repository has been equipped with a Dockerfile to encapsulate its software environment. It can be used with the [mybinder](#) service to launch an ephemeral jupyter notebook server to run the notebook.

Click on the below badge and open the notebook `adage.ipynb`.

[launch binder](#)



$$\begin{aligned} \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\ & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \tau \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\ & + \underbrace{\frac{1}{2} |(i \partial_\mu - \frac{1}{2} g \tau \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi|^2 - V(\phi)}_{W^\pm, Z, \gamma \text{ and Higgs masses and couplings}} \\ & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}} \end{aligned}$$



# SYNTHESIS

active learning / sequential design / black box optimization



## Active Sciencing



reusable workflows



simulation-based /  
likelihood-free  
inference engines

# A DEMO

## Proof-of-principle algorithm can:

- measure parameter of theory (eg. Weinberg angle in Standard Model of particle Physics) from raw data
- optimize experiment (eg. beam energy) for most sensitive measurement

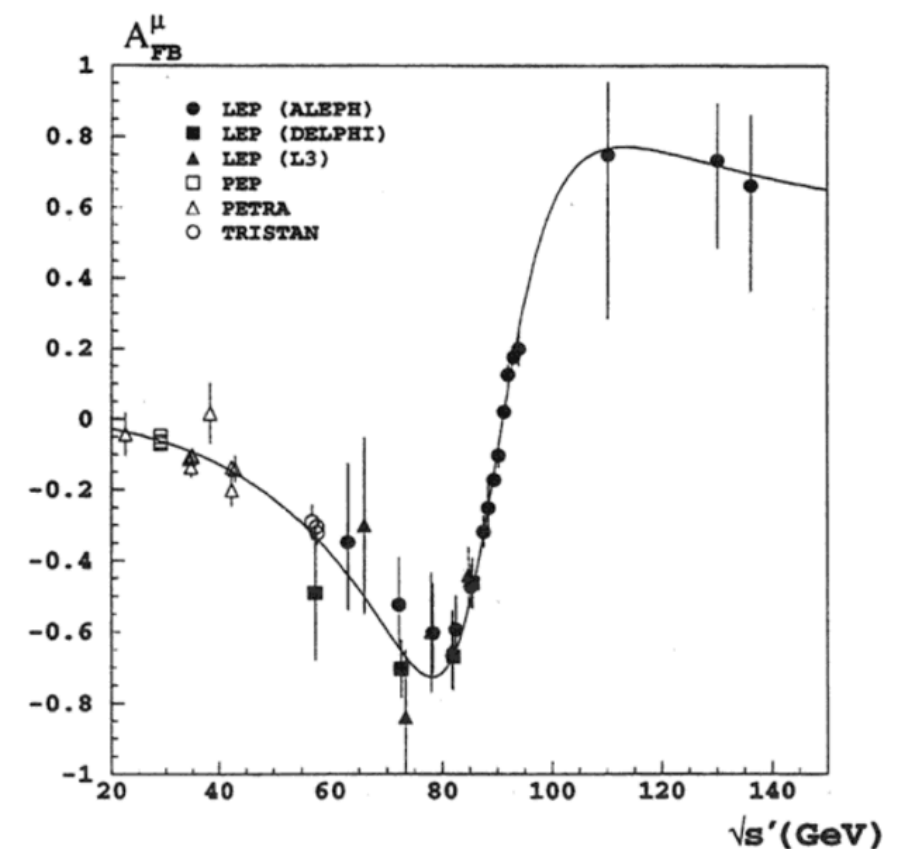
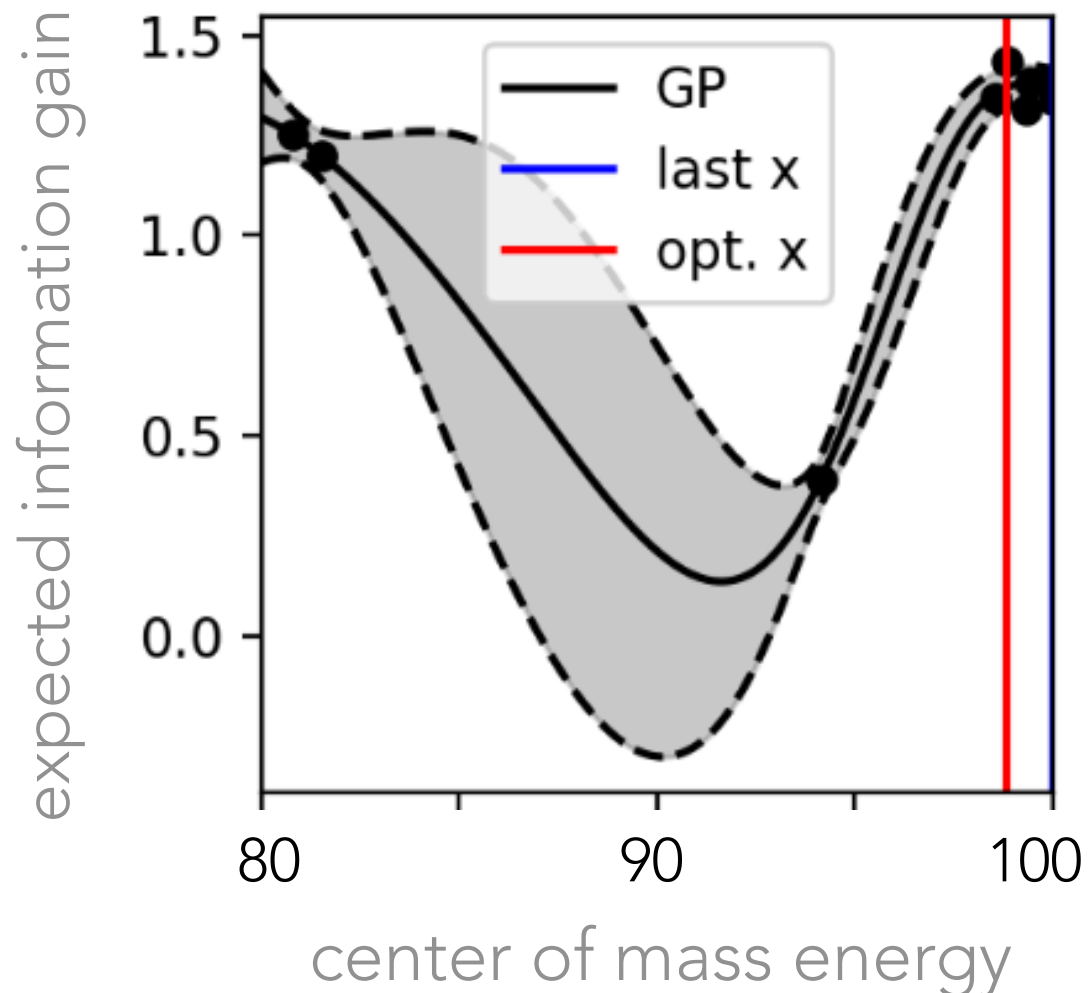


Figure 2: Measured forward-backward asymmetries of muon-pair production compared with the model independent fit results.



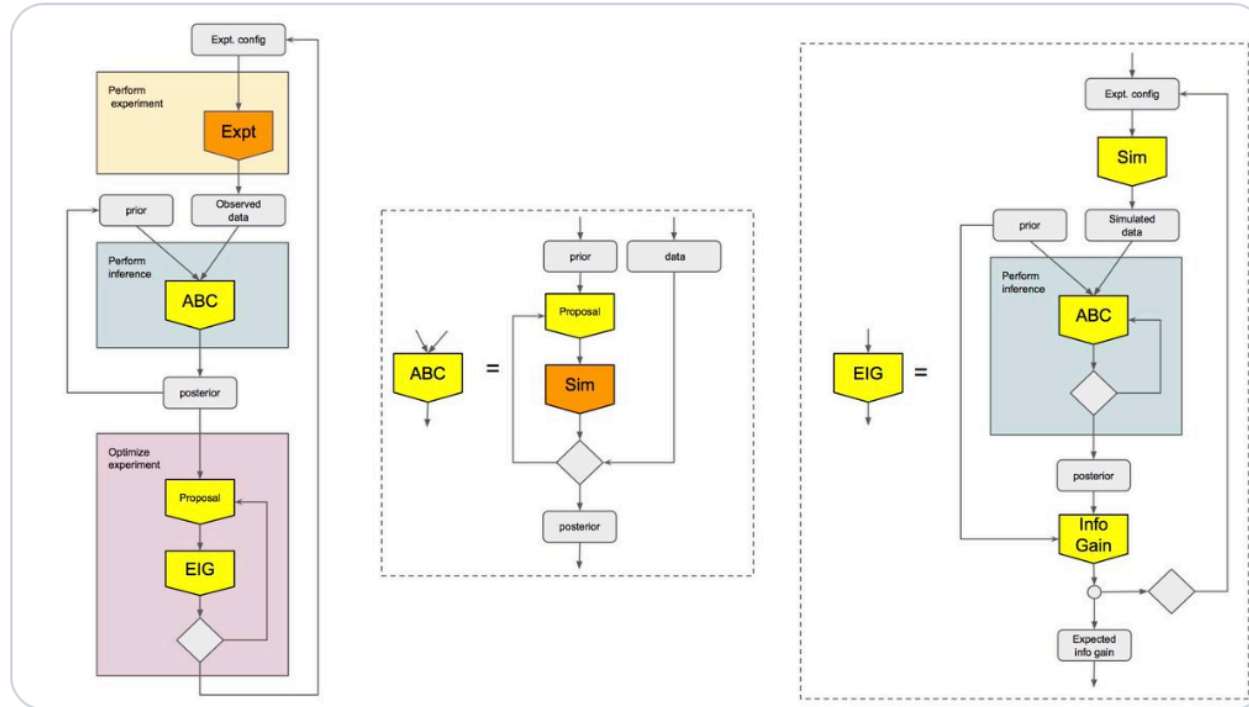
**Kyle Cranmer** @KyleCranmer · Jun 11, 2017

Demo for YComb research

active learning + workflows + implicit models = [#ActiveSciencing](#)

[@lukasheinrich\\_](#) [@glouppe](#)

[github.com/cranmer/active...](https://github.com/cranmer/active...)



4

22

61



**Danilo J. Rezende** @DeepSpiker · Jul 19, 2017

This is great!

1



3



**Kyle Cranmer** @KyleCranmer · Jul 19, 2017

Thanks!!!

1



2



**Danilo J. Rezende**

@DeepSpiker

Replying to [@KyleCranmer](#) [@lukasheinrich\\_](#) and [@glouppe](#)

You have the full loop of the scientific method in a python notebook :)

3:12 PM · Jul 19, 2017 · [Twitter for iPhone](#)





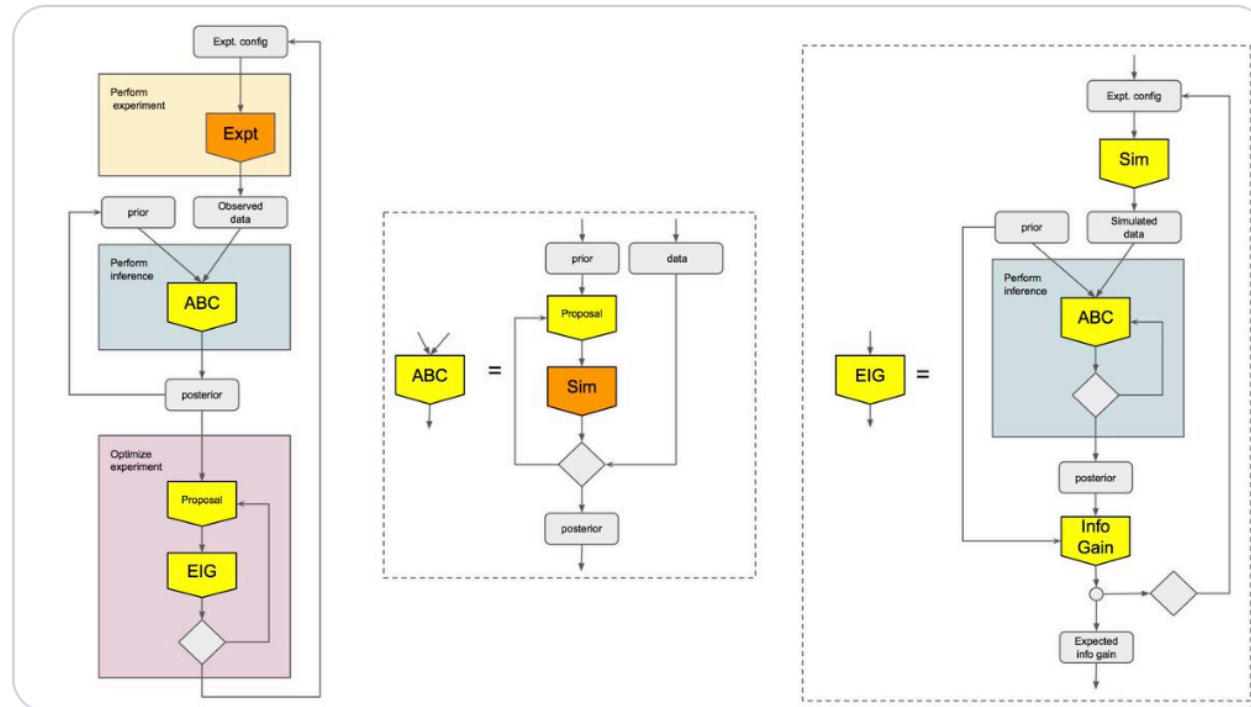
**Kyle Cranmer** @KyleCranmer · Jun 11, 2017

Demo for YComb research

active learning + workflows + implicit models = [#ActiveSciencing](#)

[@lukasheinrich\\_](#) [@glouppe](#)

[github.com/cranmer/active...](https://github.com/cranmer/active...)



4

22

61



**Danilo J. Rezende** @DeepSpiker · Jul 19, 2017

This is great!

1



3



**Kyle Cranmer** @KyleCranmer · Jul 19, 2017

Thanks!!!

1



2



**Danilo J. Rezende**

@DeepSpiker

Replying to [@KyleCranmer](#) [@lukasheinrich\\_](#) and [@glouppe](#)

You have the full loop of the scientific method in a python notebook :)

3:12 PM · Jul 19, 2017 · [Twitter for iPhone](#)

Reality check...

Keep in mind that

- the simulator model was specified
- the space of experimental configurations was well specified

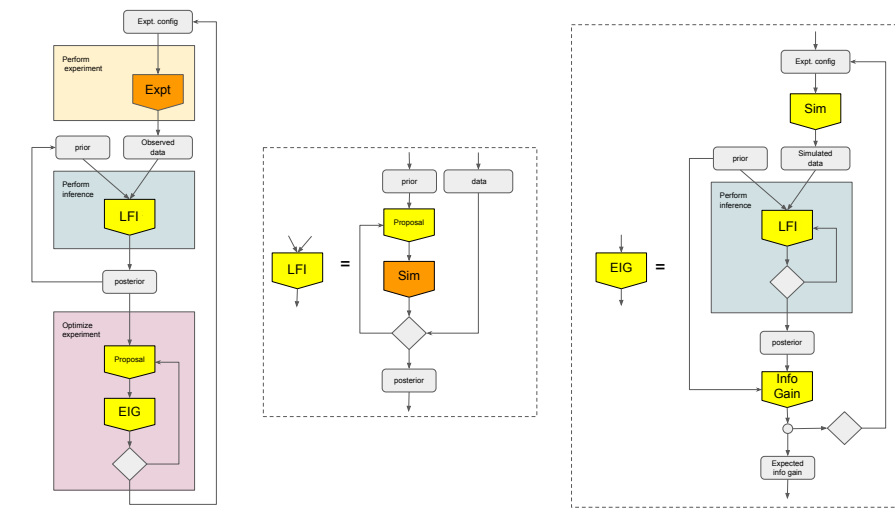
Still it was hard enough!

Going to open world of experimental configurations and potential models much harder.

Hypothesis generation also hard.

# CONSIDERATIONS

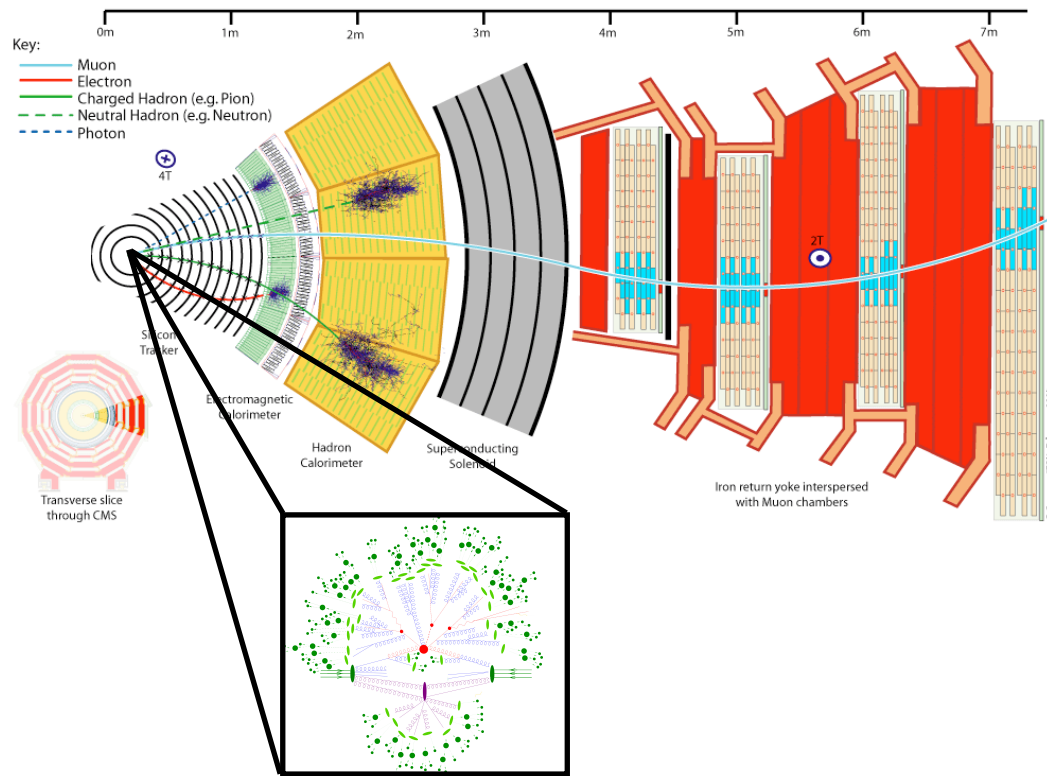
The computational cost of this workflow scales rapidly with computational cost of the LFI engine (c.f. nested loops)



- Benefit to **amortized** approaches that pay an up-front training cost in return for faster repeated inference
- Want the training techniques to be as **sample efficient** as possible when the simulator is computationally expensive
- Anticipate a hierarchy of surrogate models:
  - likelihood, posterior, utility (EIG) surface, acquisition, ...

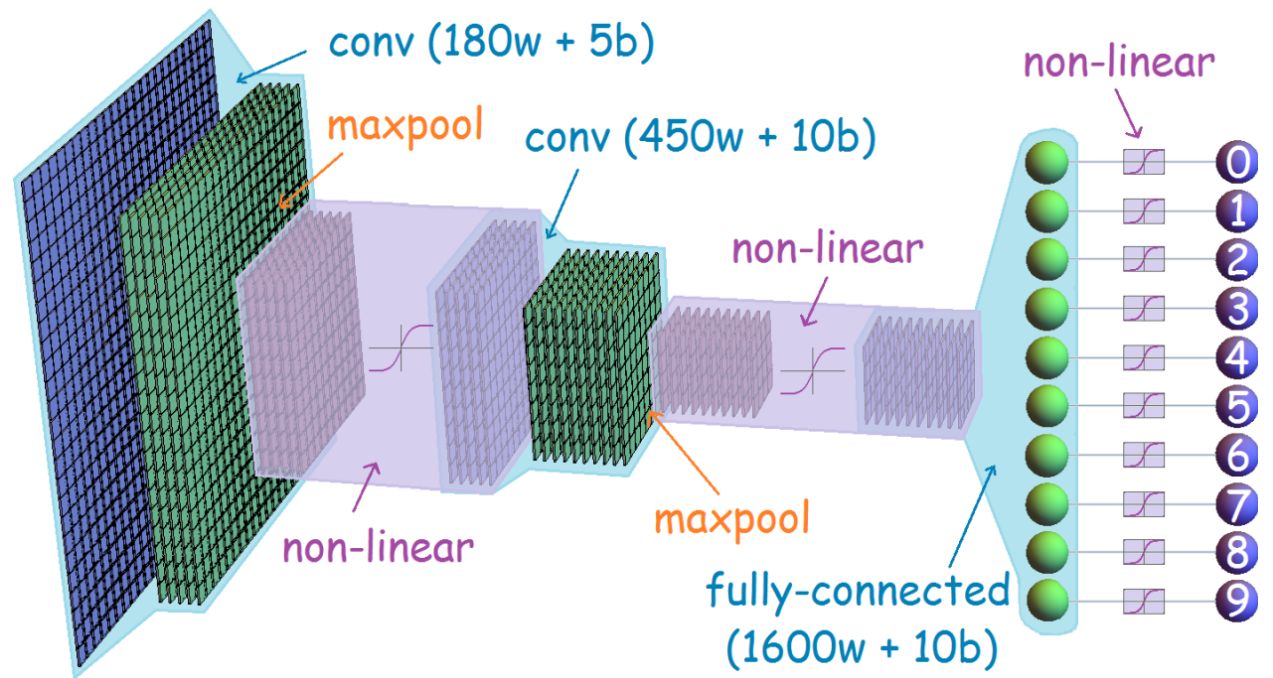
# APPROACHES TO LIKELIHOOD-FREE INFERENCE

**Use simulator**  
(much more efficiently)



- Approximate Bayesian Computation (ABC)
- Probabilistic Programming
- Adversarial Variational Optimization (AVO)

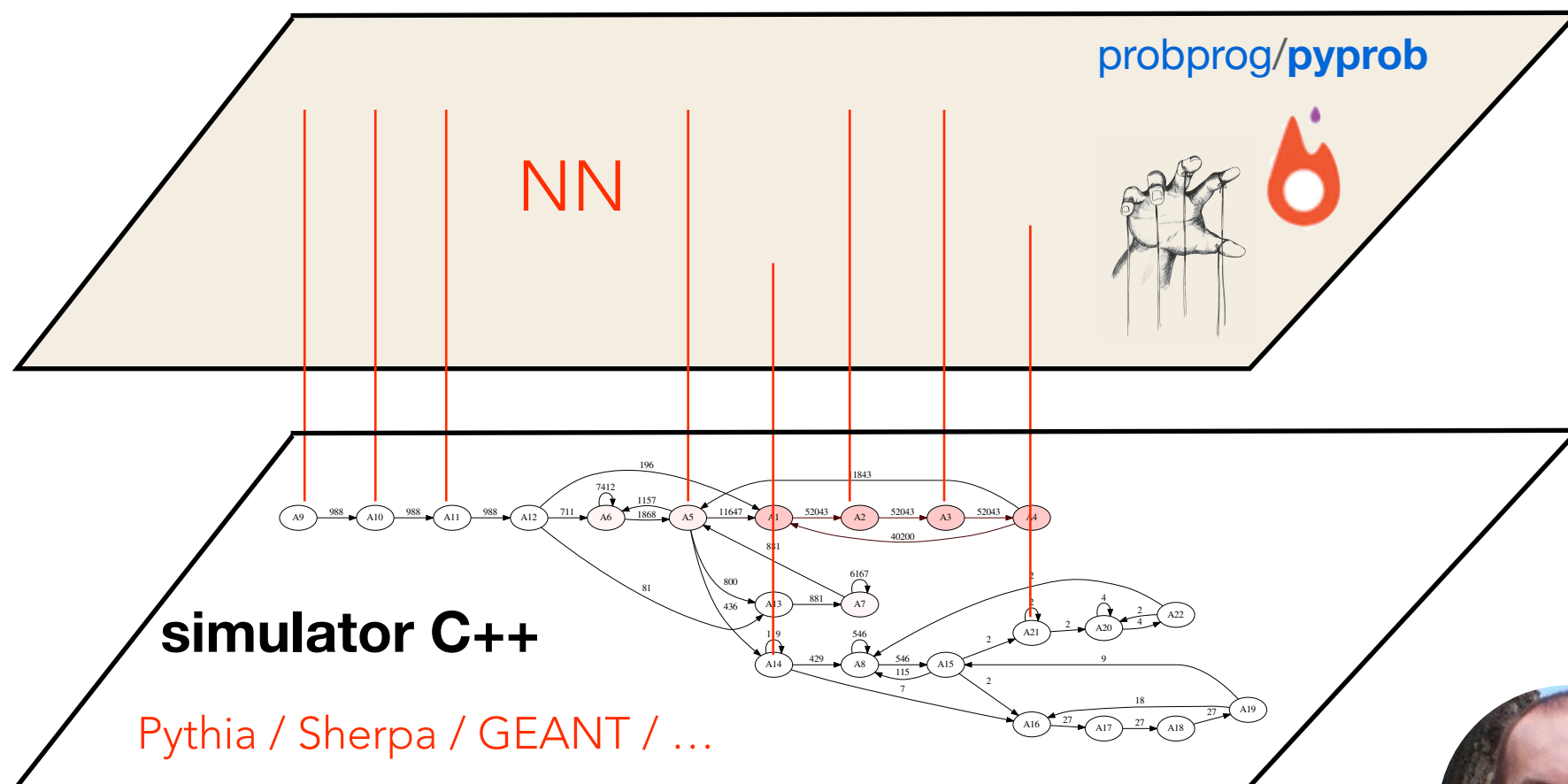
**Learn simulator**  
(with deep learning)



- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autoregressive models, Normalizing Flows

# PROBABILISTIC PROGRAMMING

**Idea:** hijack the random number generators and use Neural Network to perform a very fancy type of importance sampling

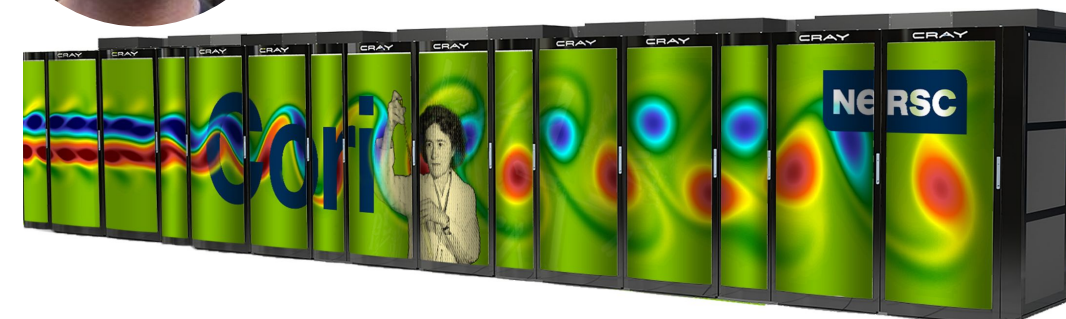
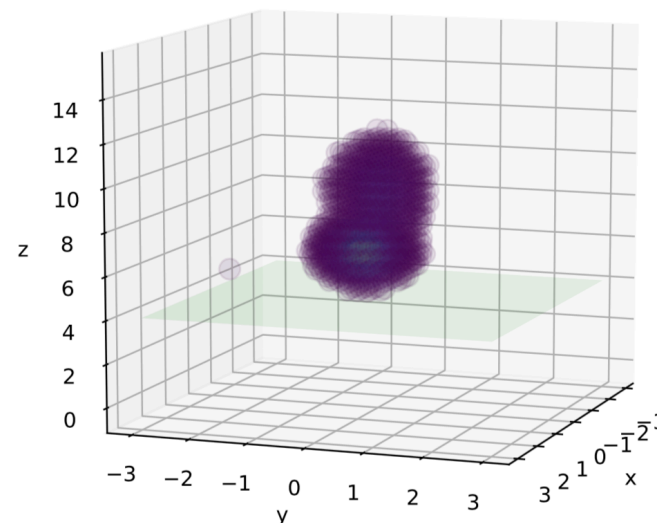
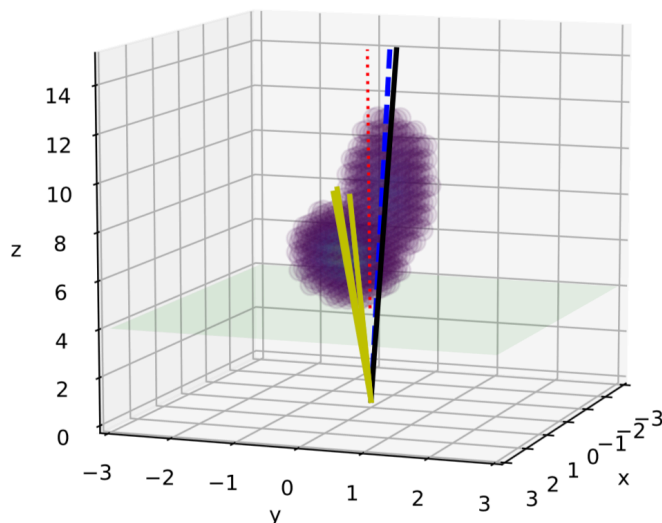


- Neural Network powered inference engine (python)
- real-world scientific simulator (C++)



Observation

Mean Simulated Observation

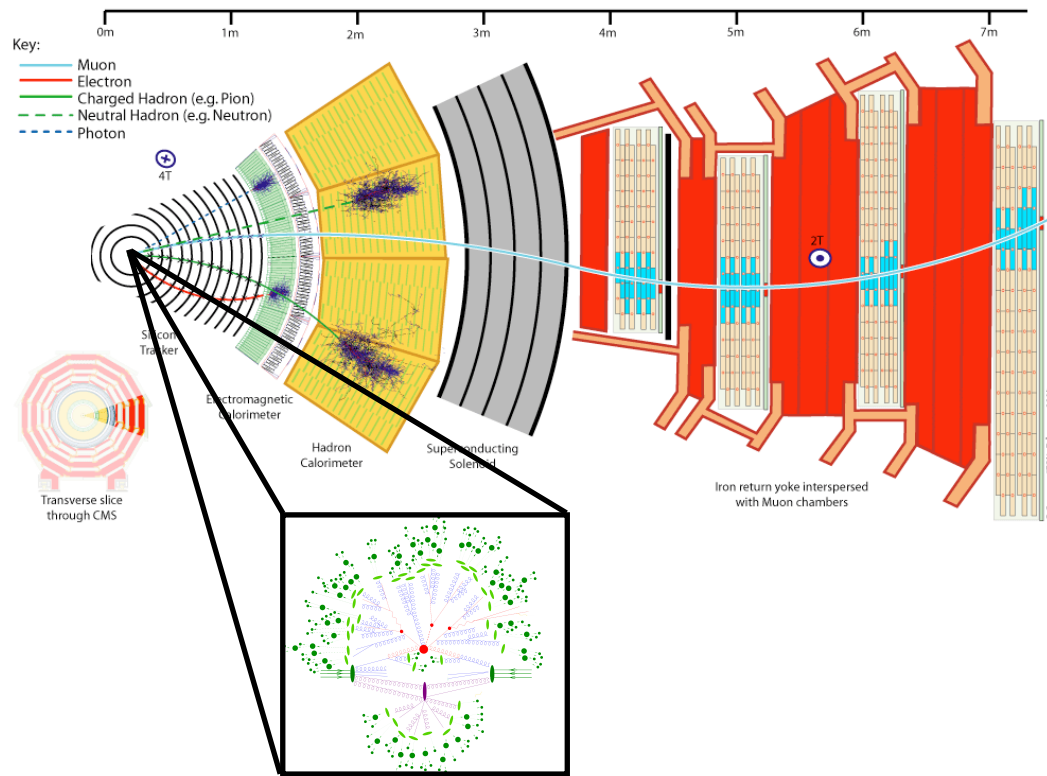


NERSC, Lawrence Berkeley National Lab



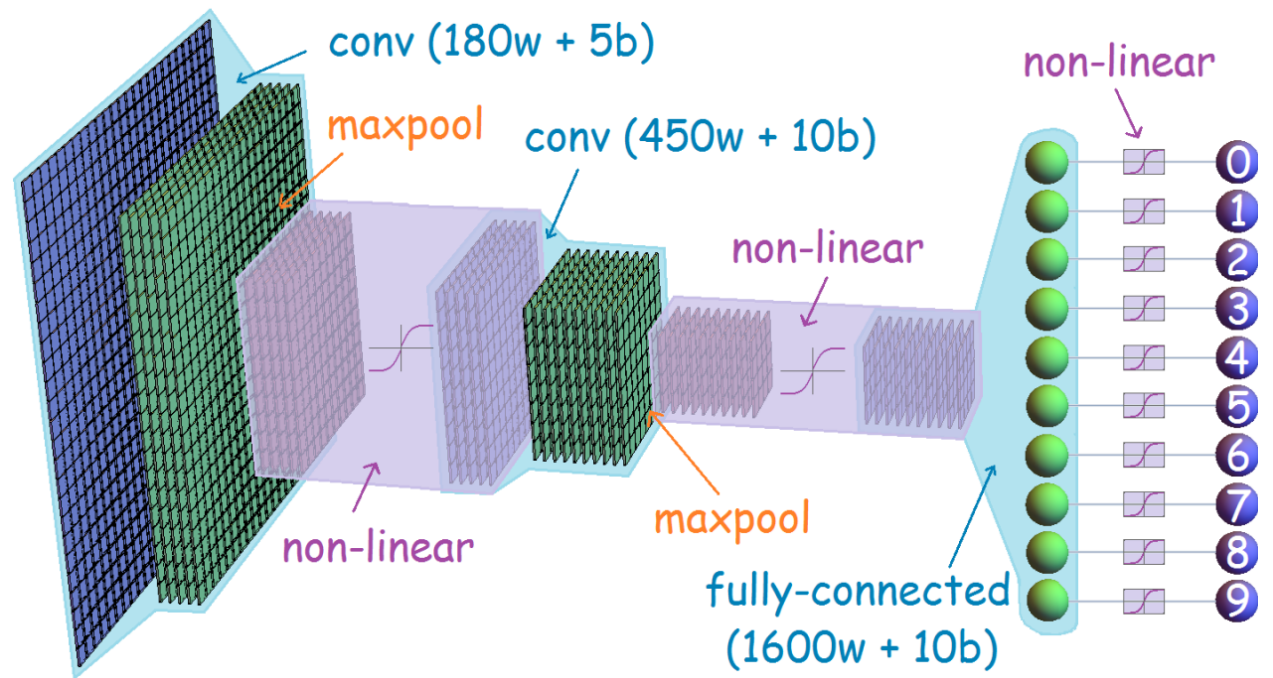
# APPROACHES TO LIKELIHOOD-FREE INFERENCE

**Use simulator**  
(much more efficiently)



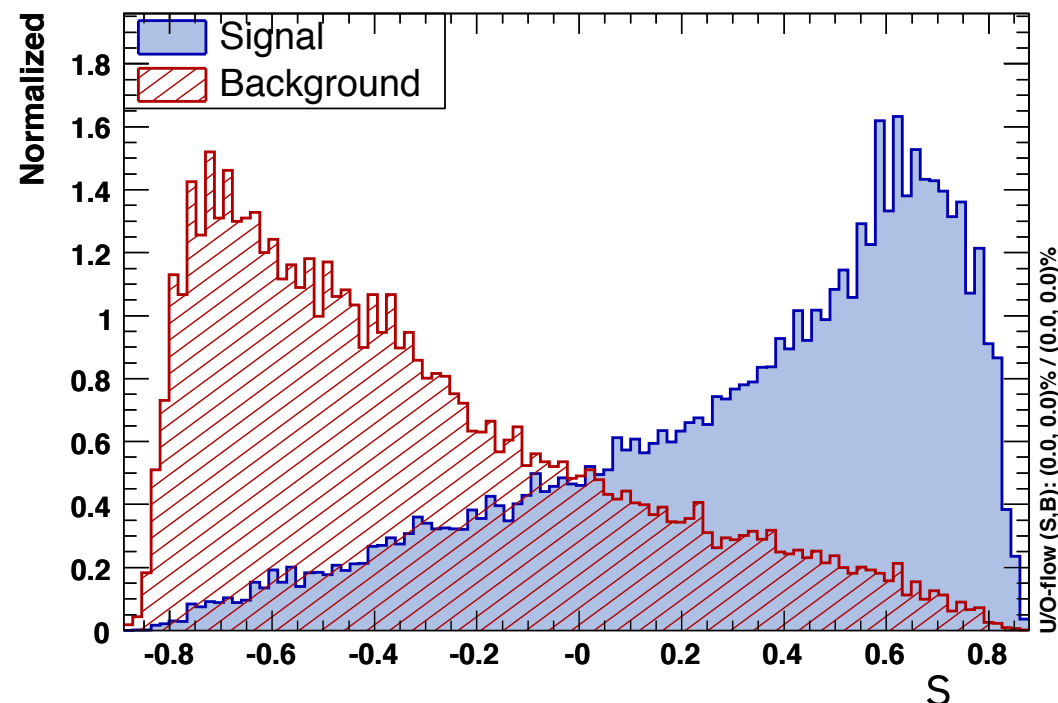
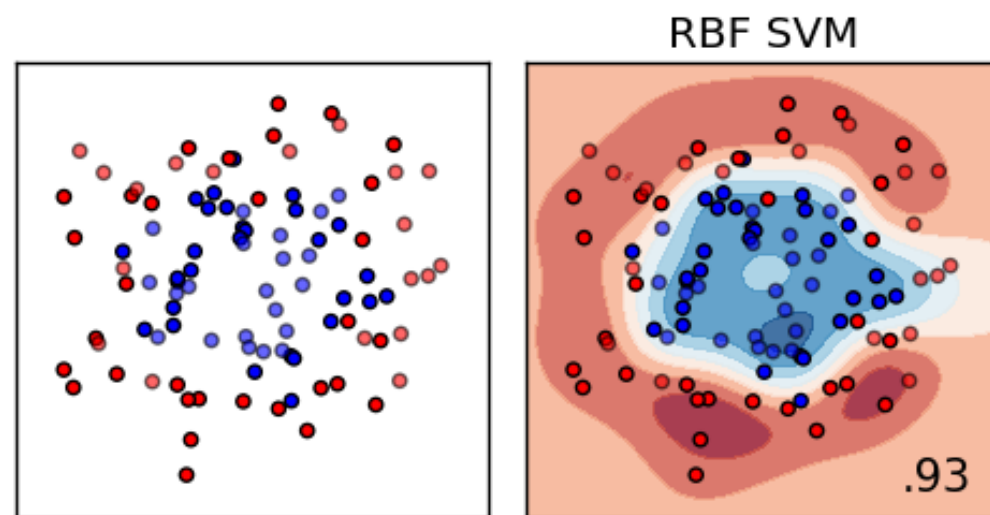
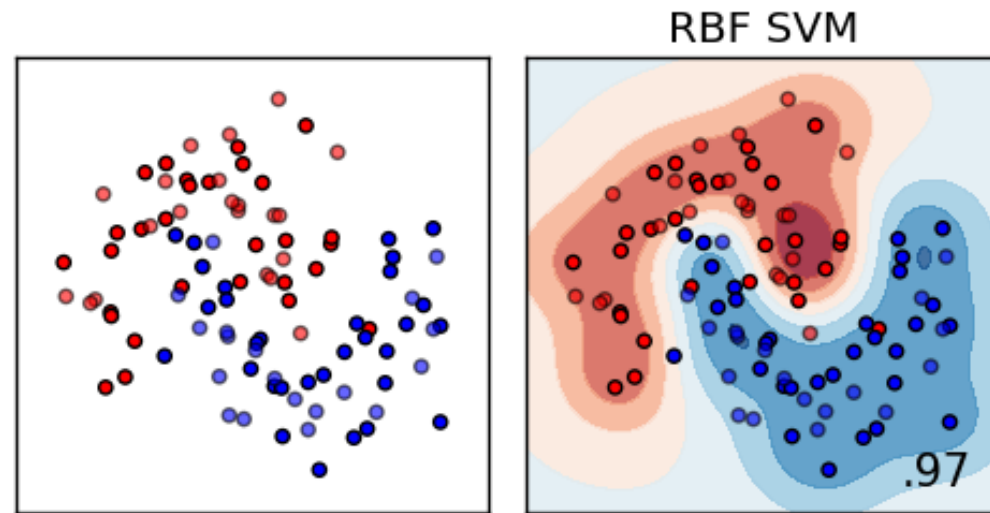
- Approximate Bayesian Computation (ABC)
- Probabilistic Programming
- Adversarial Variational Optimization (AVO)

**Learn simulator**  
(with deep learning)



- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autoregressive models, Normalizing Flows

# LIKELIHOOD RATIO TRICK



- **binary classifier**: find function  $s(x)$  that minimizes **loss**:

$$L[s] = \int p(x|H_0) (0 - s(x))^2 dx + \int p(x|H_1) (1 - s(x))^2 dx$$

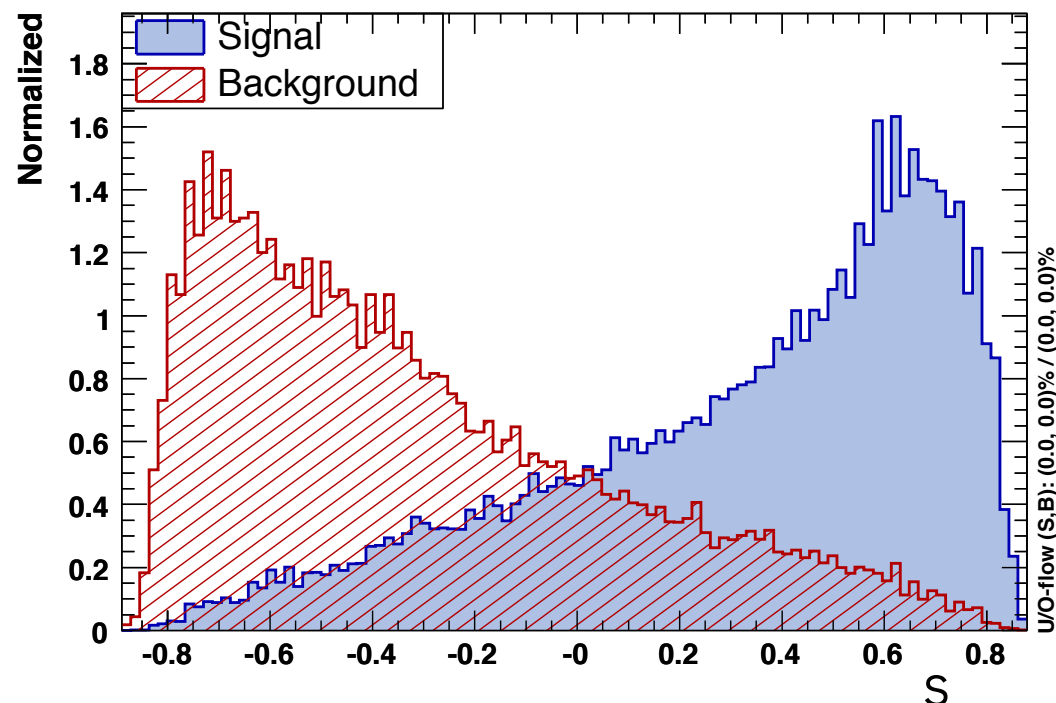
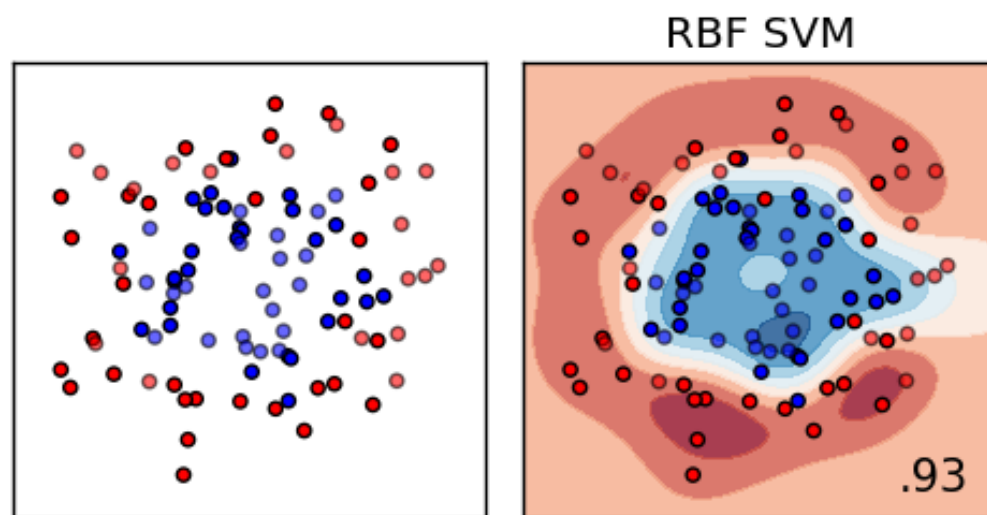
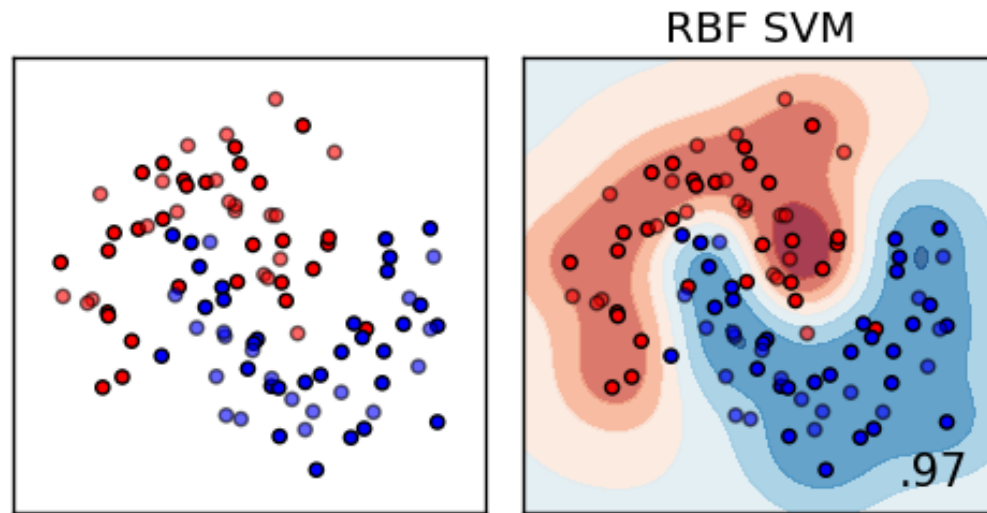
- i.e. approximate the Bayes optimal classifier

$$s(x) = \frac{p(x|H_1)}{p(x|H_0) + p(x|H_1)}$$

- which is 1-to-1 with the likelihood ratio

$$\frac{p(x|H_1)}{p(x|H_0)}$$

# LIKELIHOOD RATIO TRICK



- **binary classifier**: find function  $s(x)$  that minimizes **loss**:

$$L[s] = \int p(x|H_0) (0 - s(x))^2 dx + \int p(x|H_1) (1 - s(x))^2 dx$$

$$\approx \frac{1}{N} \sum_{i=1}^N (y_i - s(x_i))^2$$

- i.e. approximate the Bayes optimal classifier

$$s(x) = \frac{p(x|H_1)}{p(x|H_0) + p(x|H_1)}$$

- which is 1-to-1 with the likelihood ratio

$$\frac{p(x|H_1)}{p(x|H_0)}$$

# GANs AND THE LIKELIHOOD RATIO TRICK

The discriminator of a GAN approximates

$$s(x) = \frac{p(x|G)}{p(x|D) + p(x|G)}$$

Which is one-to-one with the likelihood ratio

$$\frac{p(x|D)}{p(x|G)} = 1 - \frac{1}{s(x)}$$

Can do the same thing for any two points  $\theta_0$  &  $\theta_1$  in parameter space  $\Theta$ . I call this a **parametrized classifier**

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



# LIKELIHOOD-FREE INFERENCE

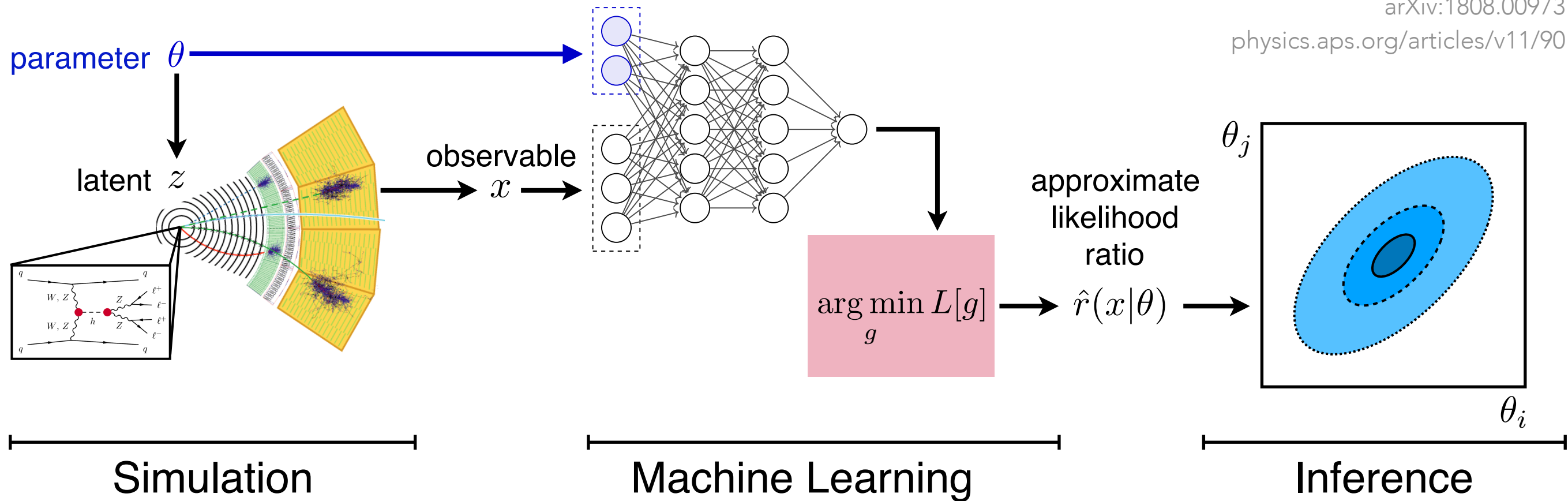
arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90



The **surrogate for the likelihood (ratio)** used for inference

Currently a 2-stage process:

1. learning surrogate
2. Inference on parameters of simulator

Wanted: theory with **joint treatment** of the two stages

# LEARNING THE LIKELIHOOD RATIO

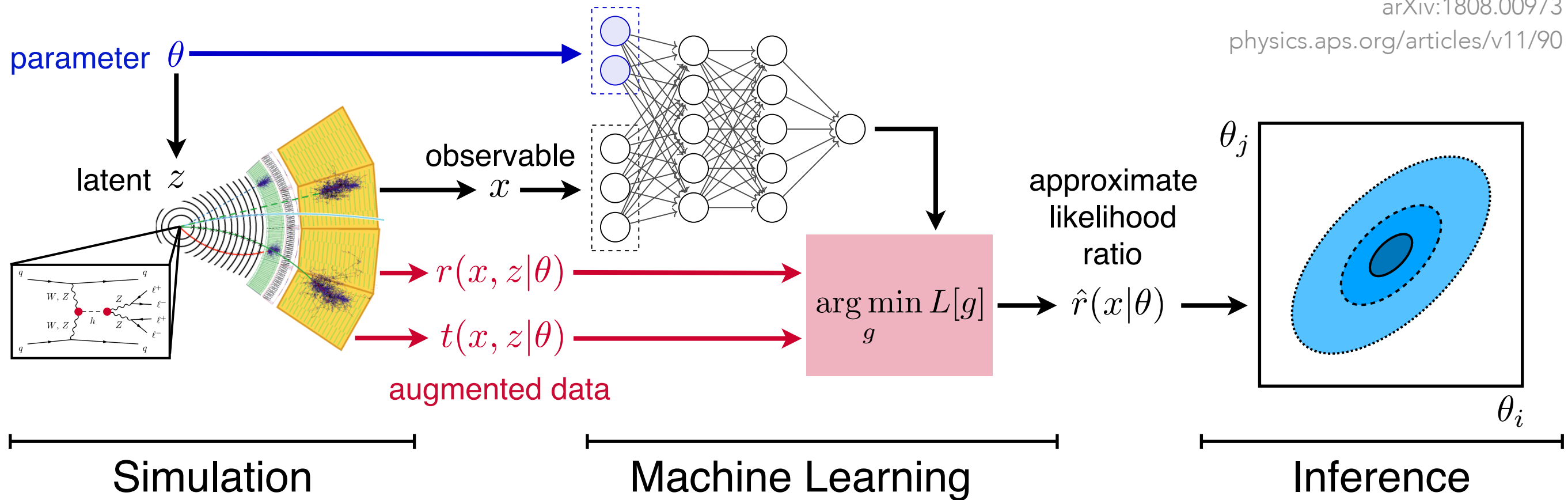
arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90



Recently, we realized we can **extract more from the simulator**.  
We can use **augmented data** to improve training



# MINING GOLD

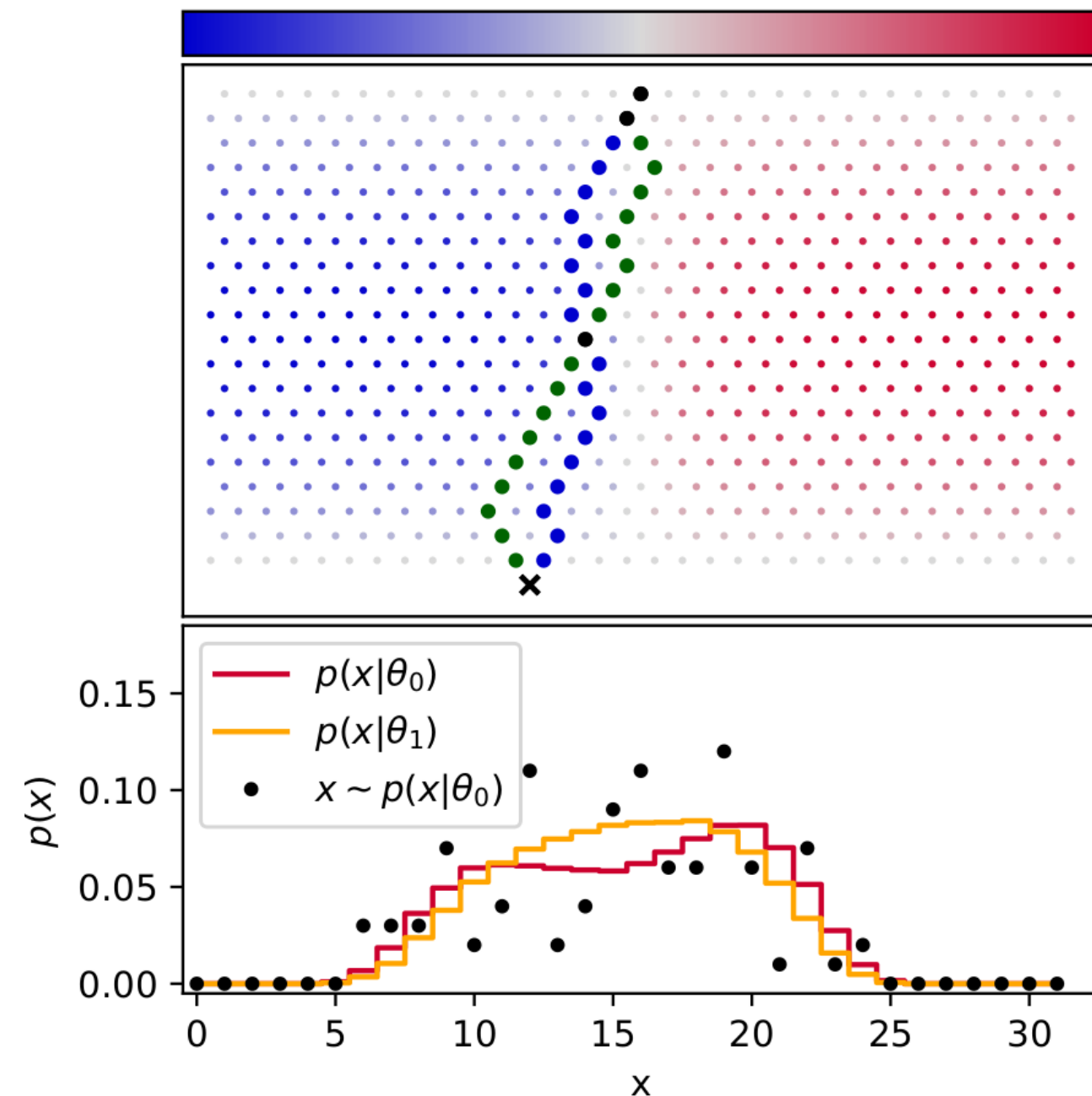
While implicit density is intractable

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

Some quantities conditioned on latent  $z$  are tractable:

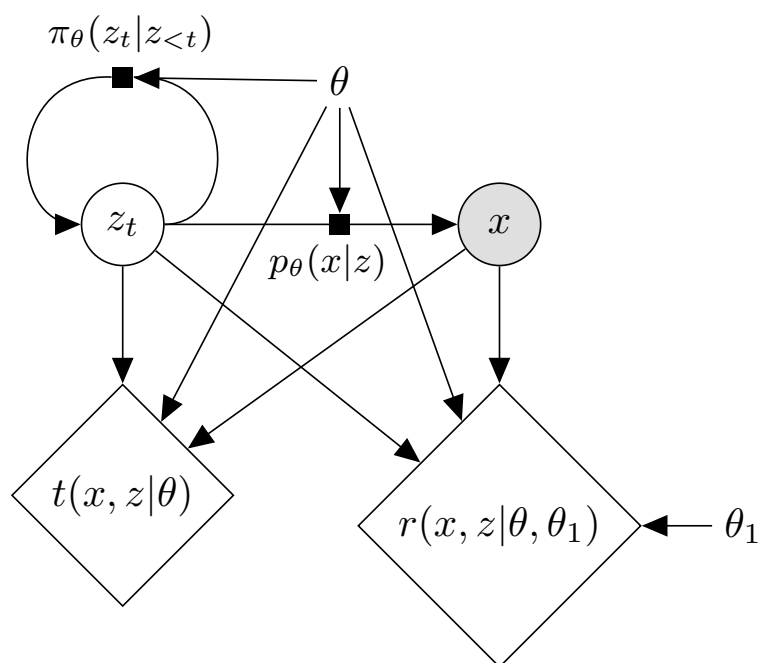
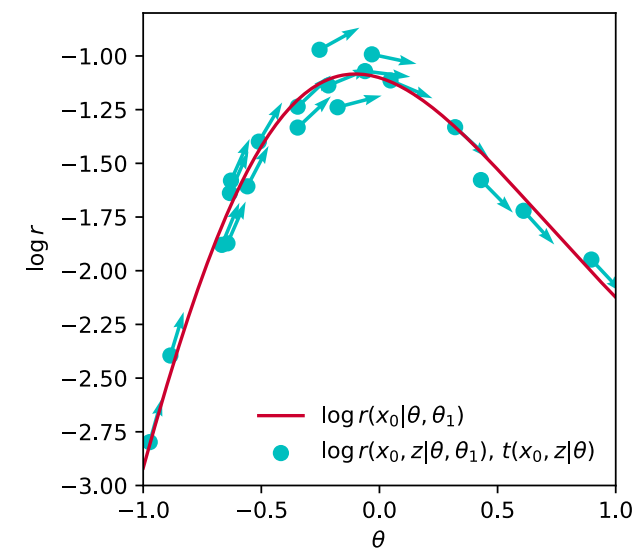
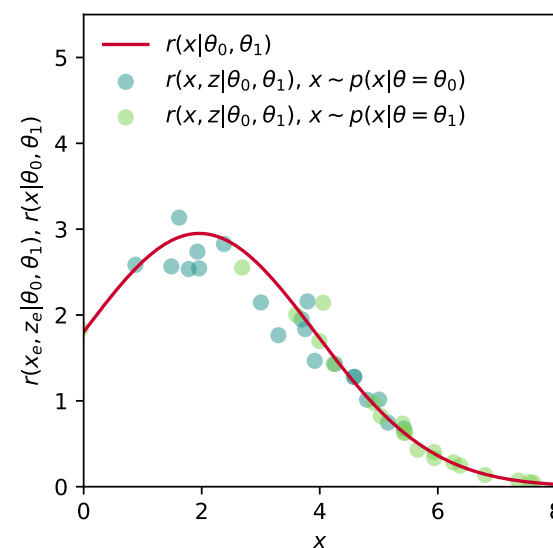
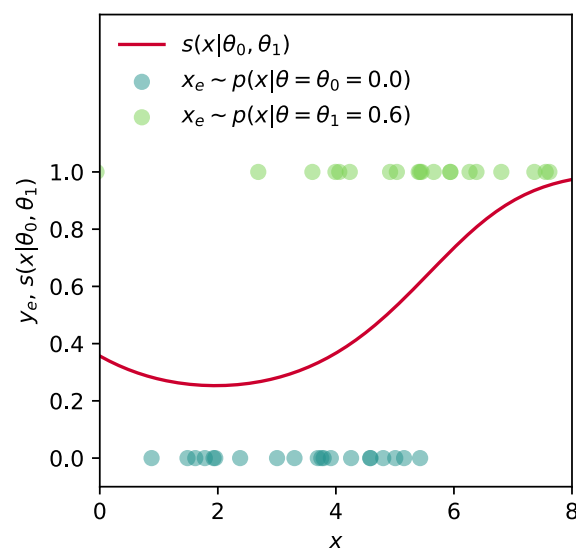
$$r(x, z|\theta_0, \theta_1) = \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)}$$

and similar to REINFORCE policy gradient

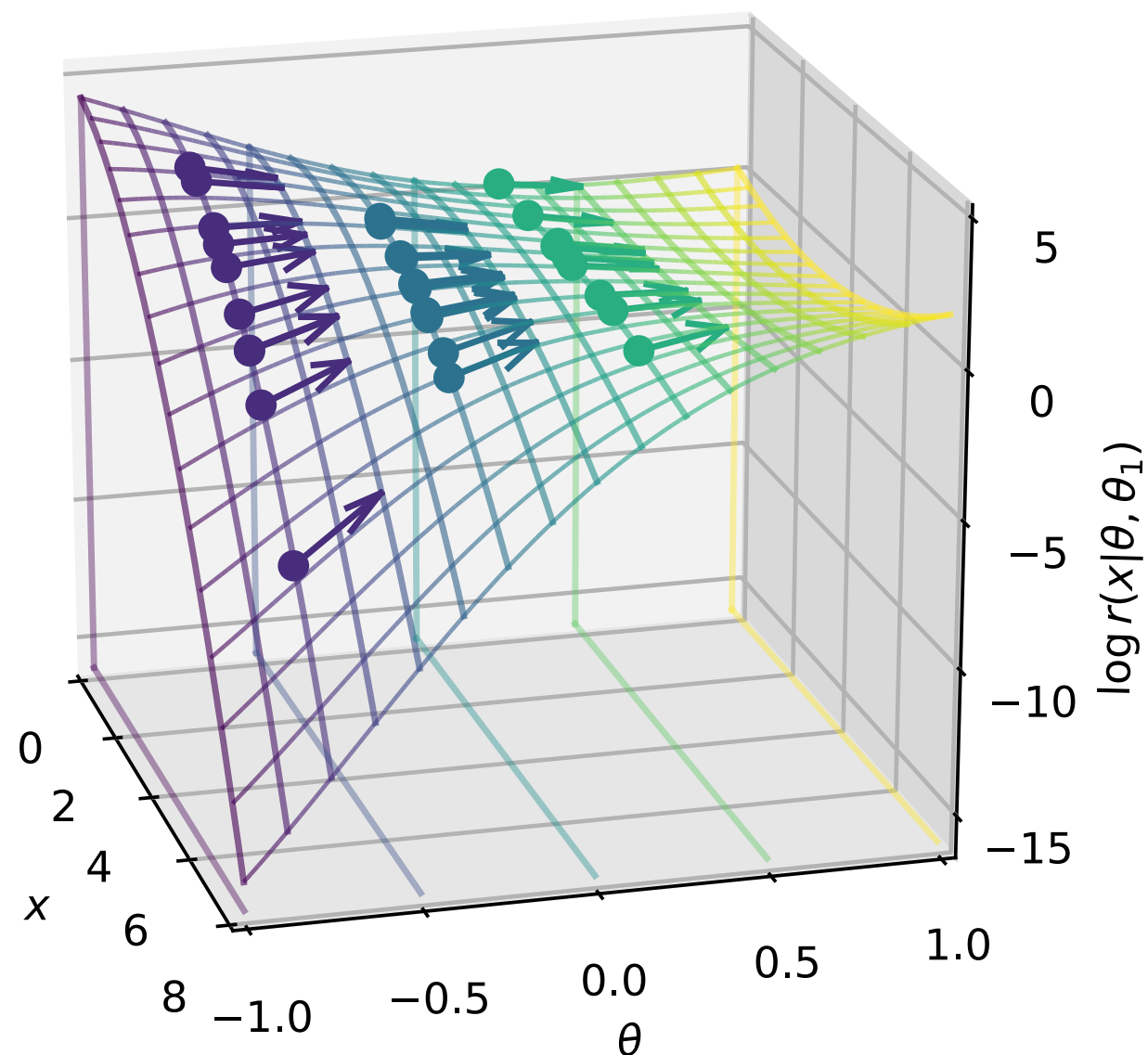


$$t(x, z|\theta_0) = \frac{\nabla_{\theta} p(x, z|\theta)|_{\theta_0}}{p(x, z|\theta_0)} = \nabla_{\theta} \log p(x, z|\theta)|_{\theta_0}$$

# PUTTING IT ALL TOGETHER



can think of simulator  
as policy  $\pi_\theta$  in language of  
reinforcement learning





# LEARNING THE LIKELIHOOD RATIO

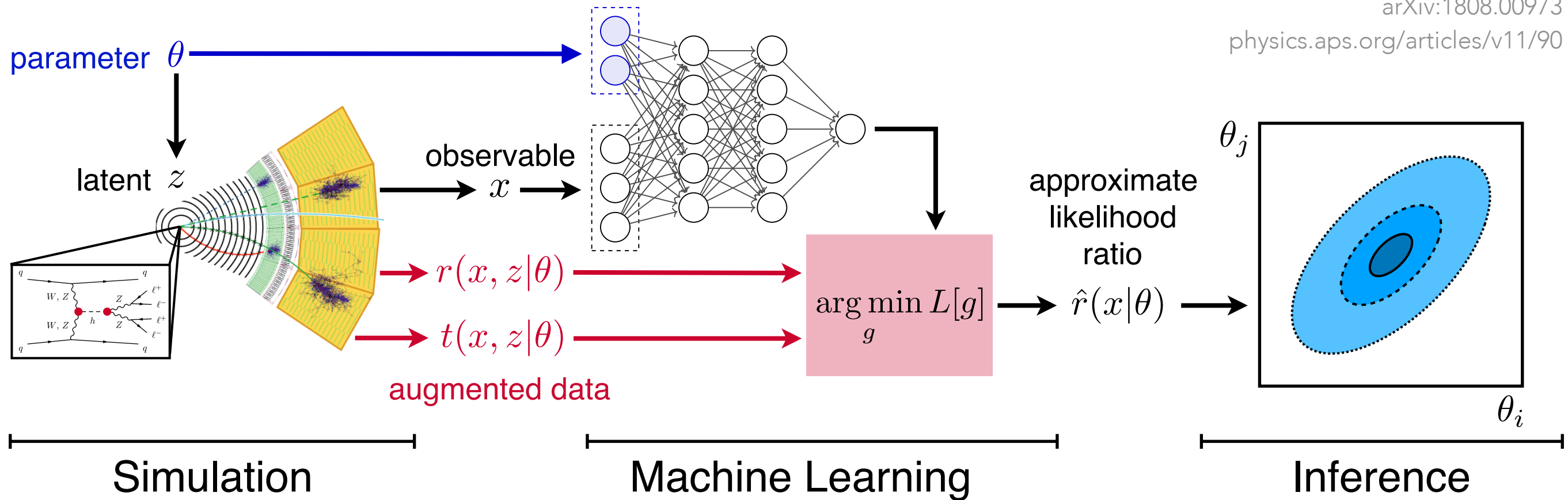
arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90



# LEARNING THE LIKELIHOOD RATIO

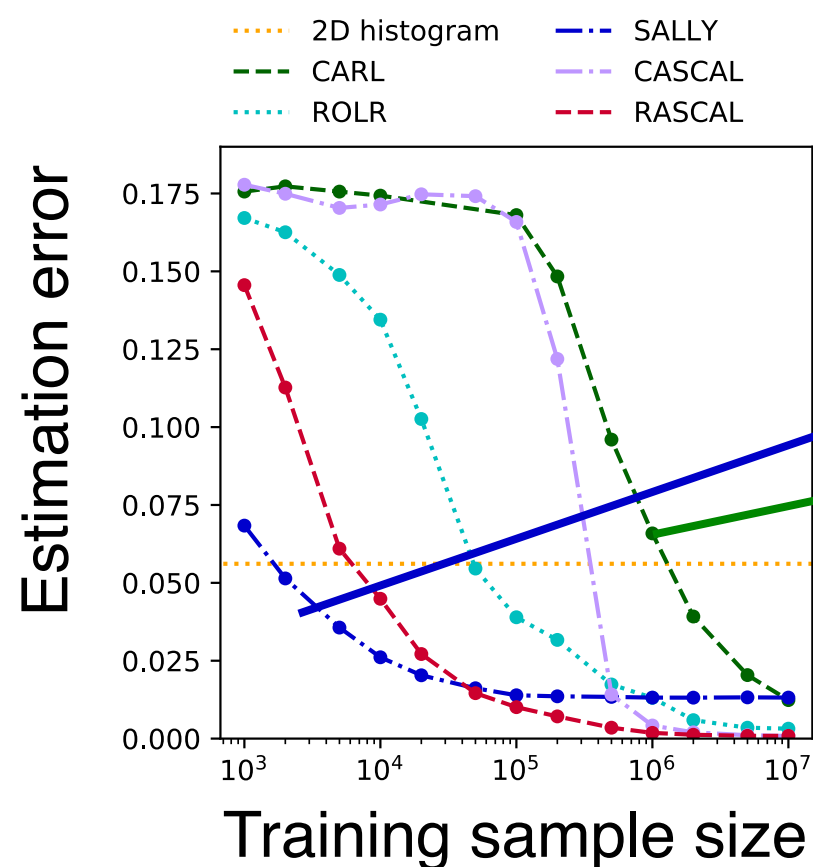
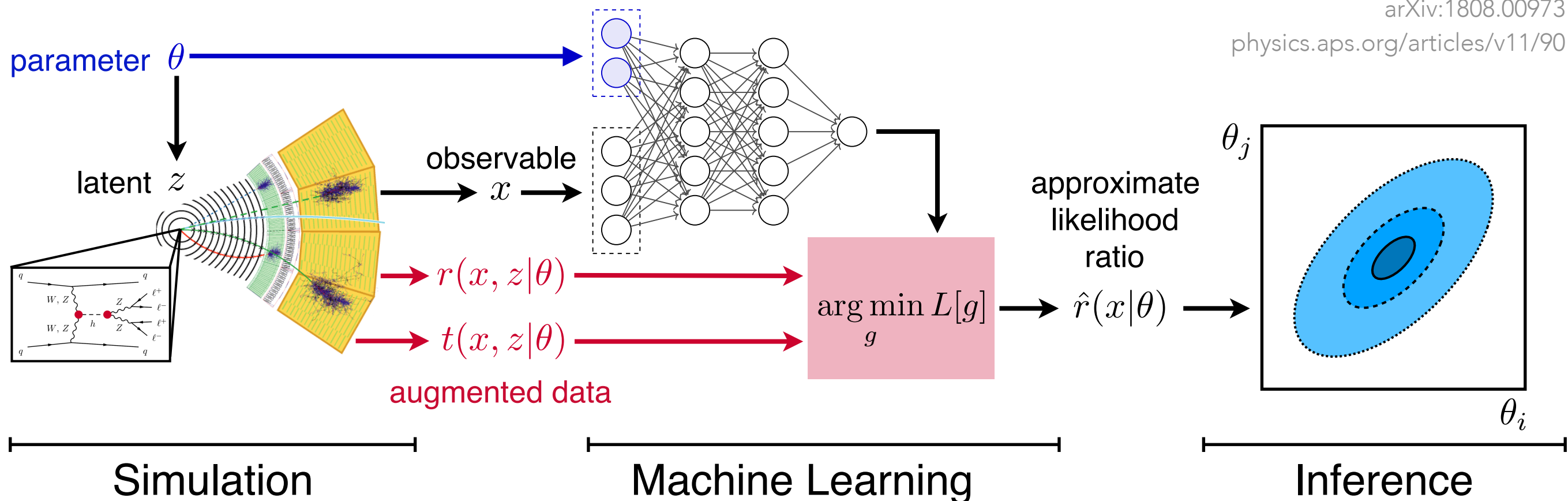
arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90



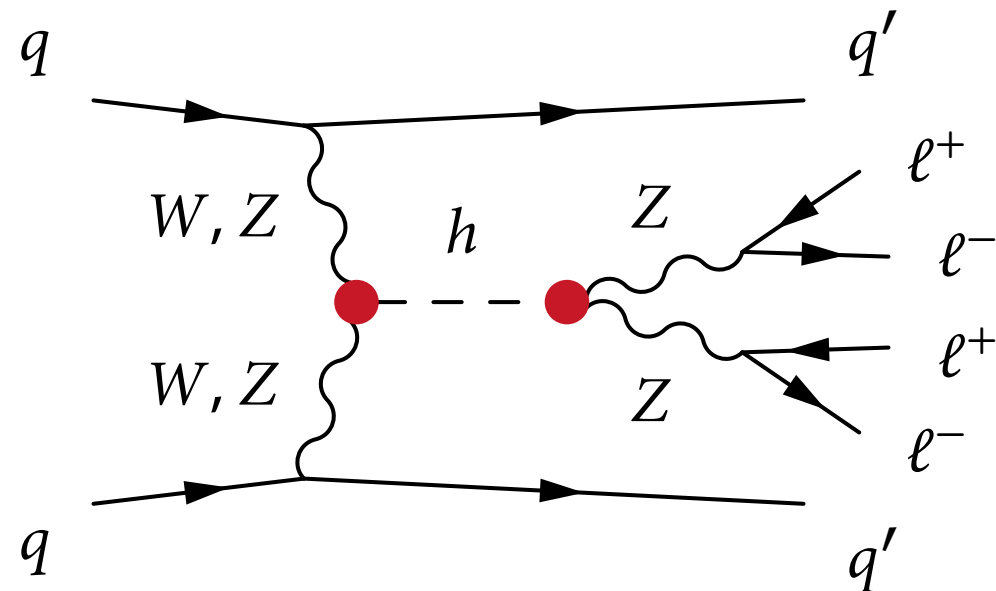
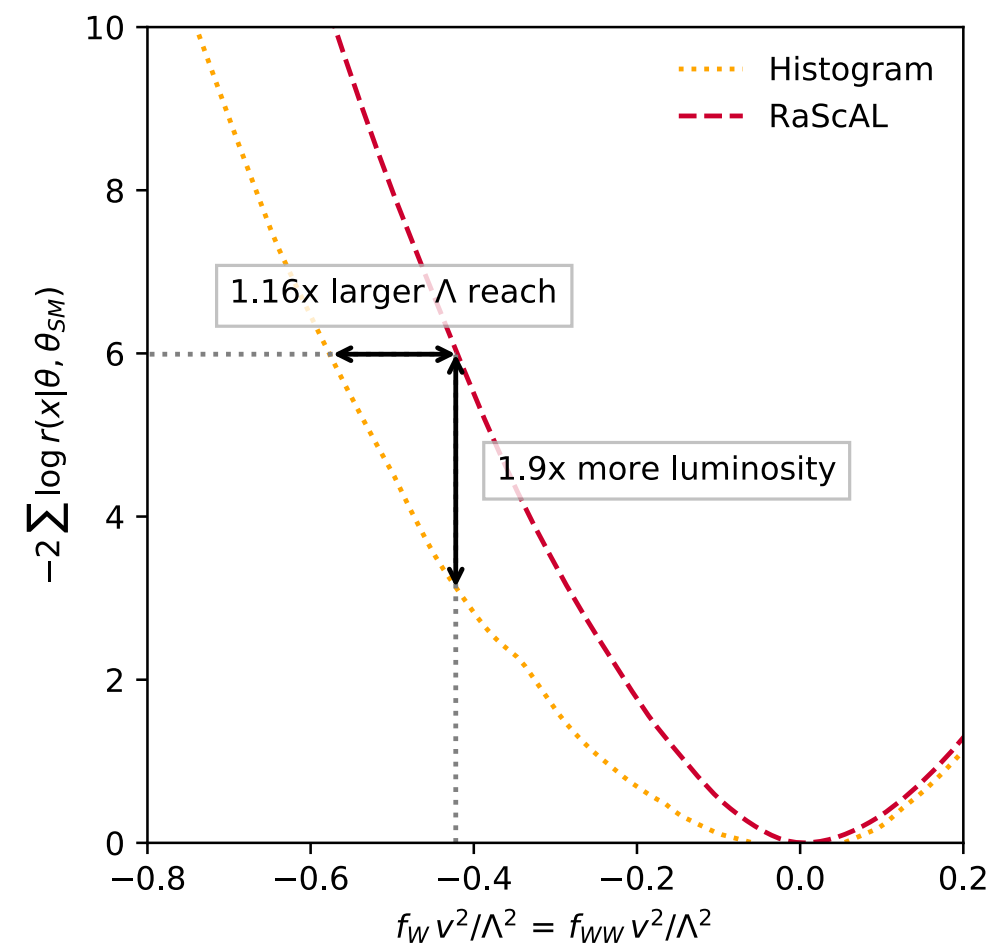
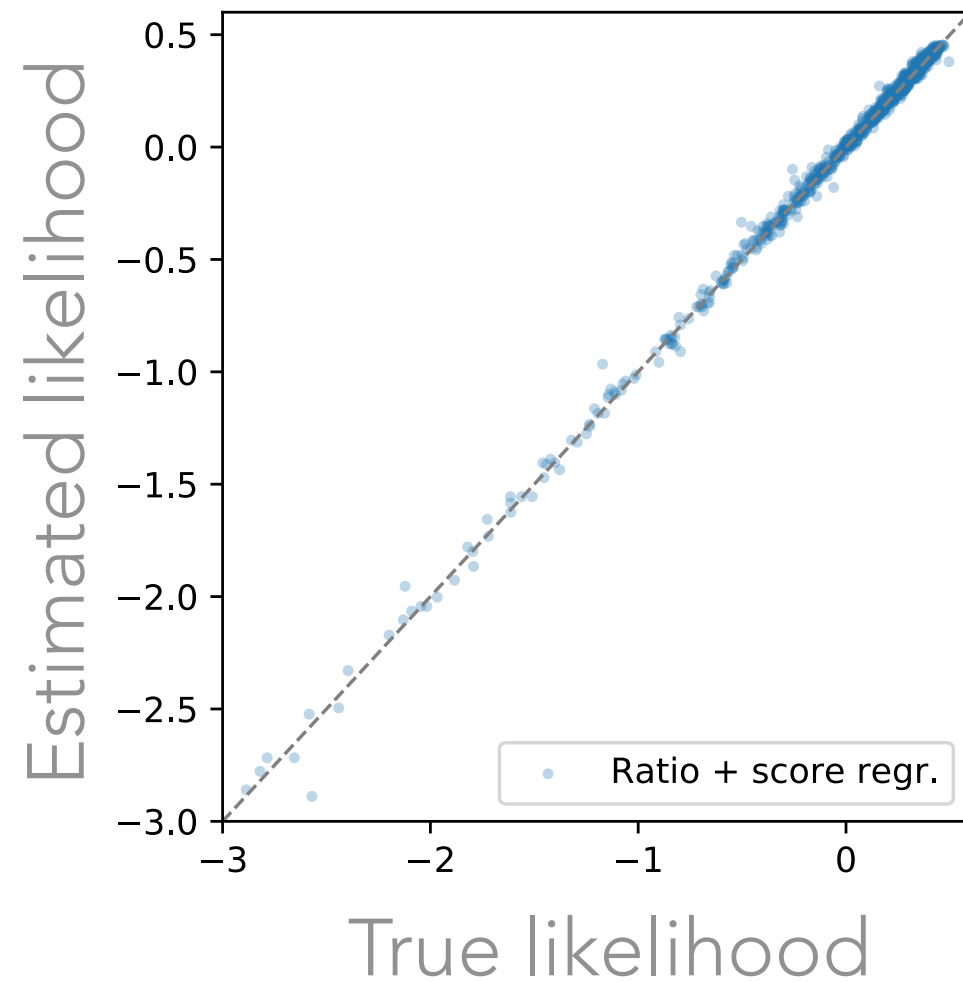
New techniques  
require less data than  
without augmented data

Traditional Approach no NN

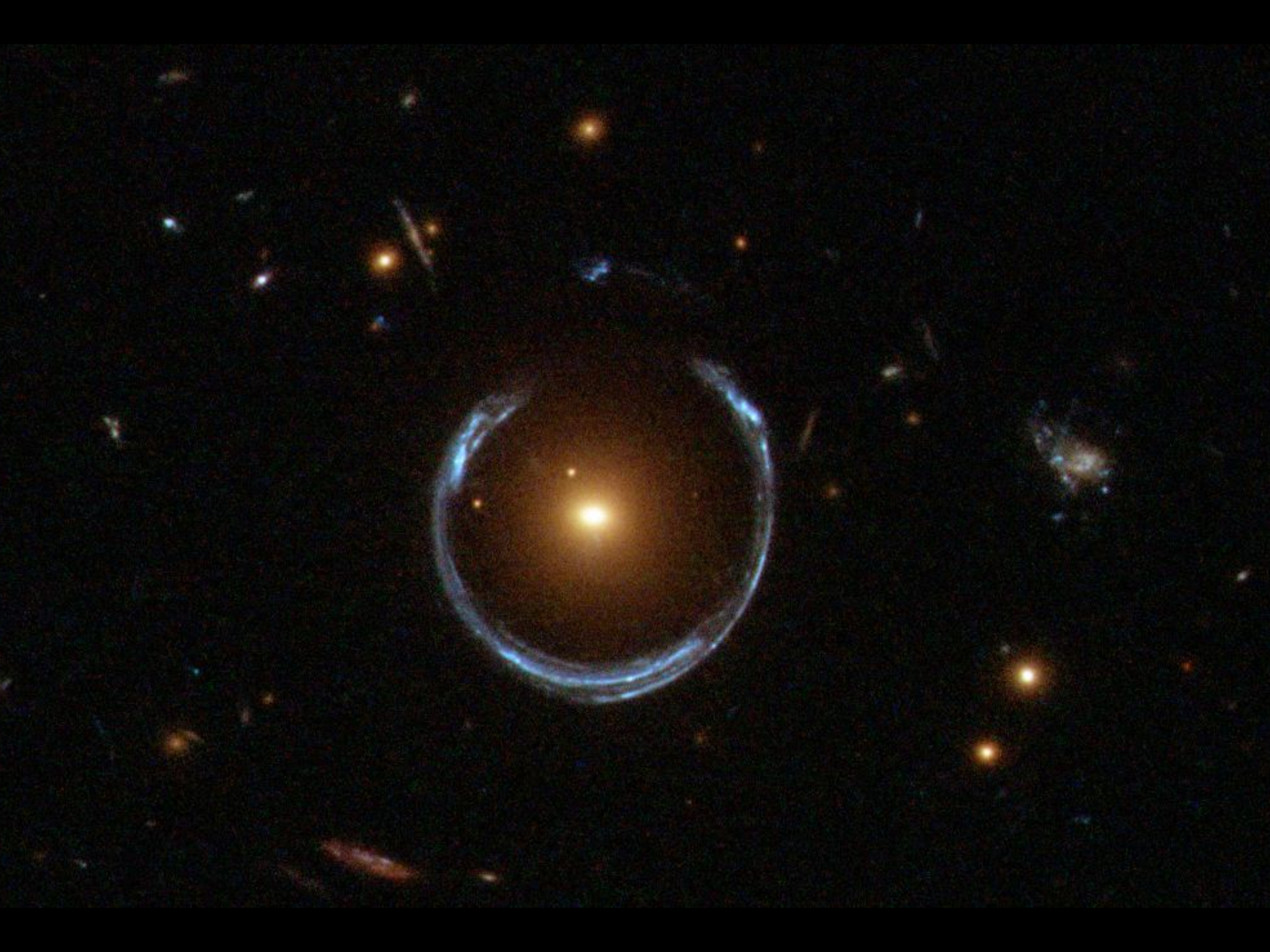
Two Examples

# IMPACT ON STUDIES OF THE HIGGS BOSON

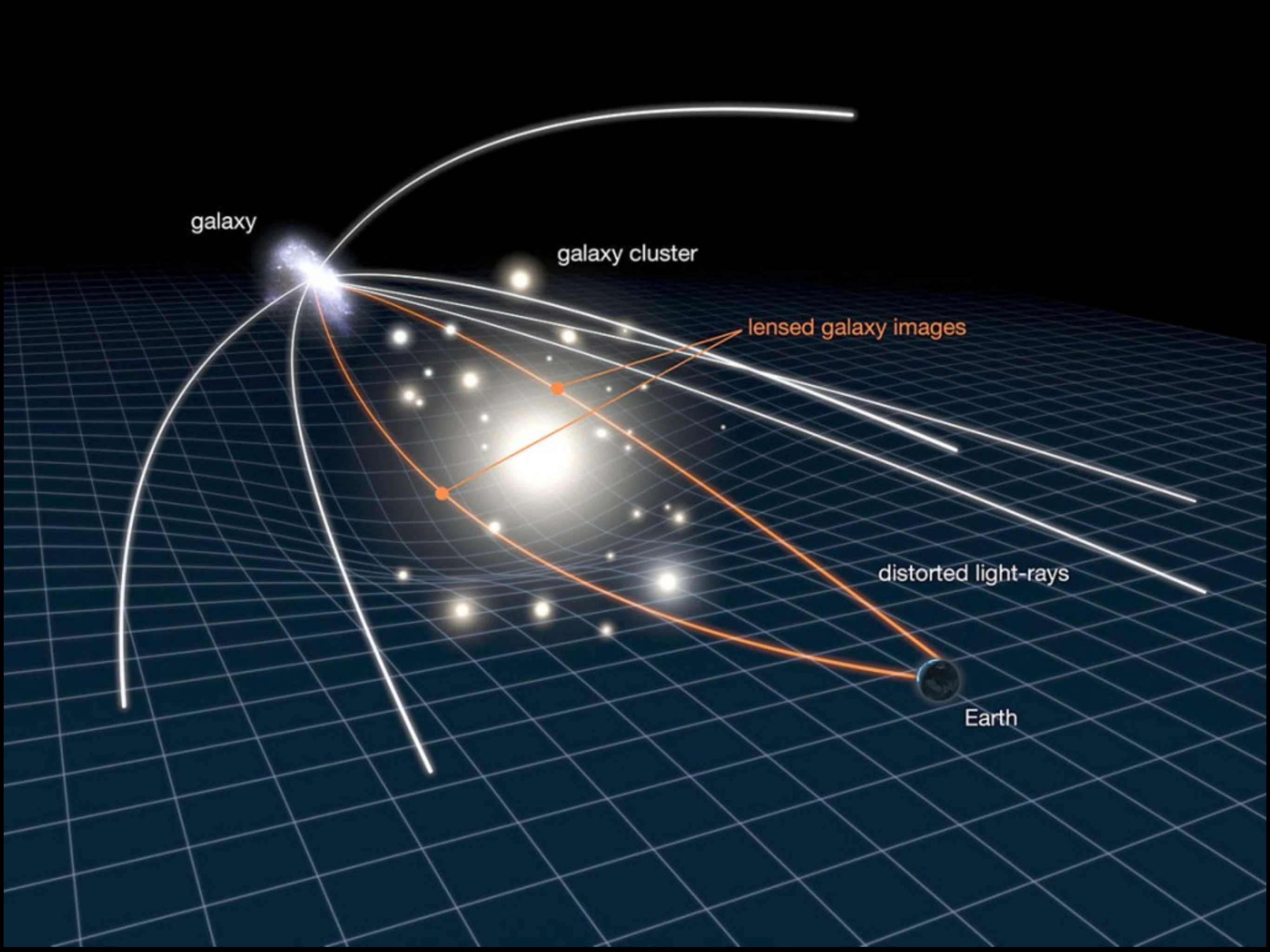
(based on a 42-Dim observation  $\mathbf{x}$ )











galaxy

galaxy cluster

lensed galaxy images

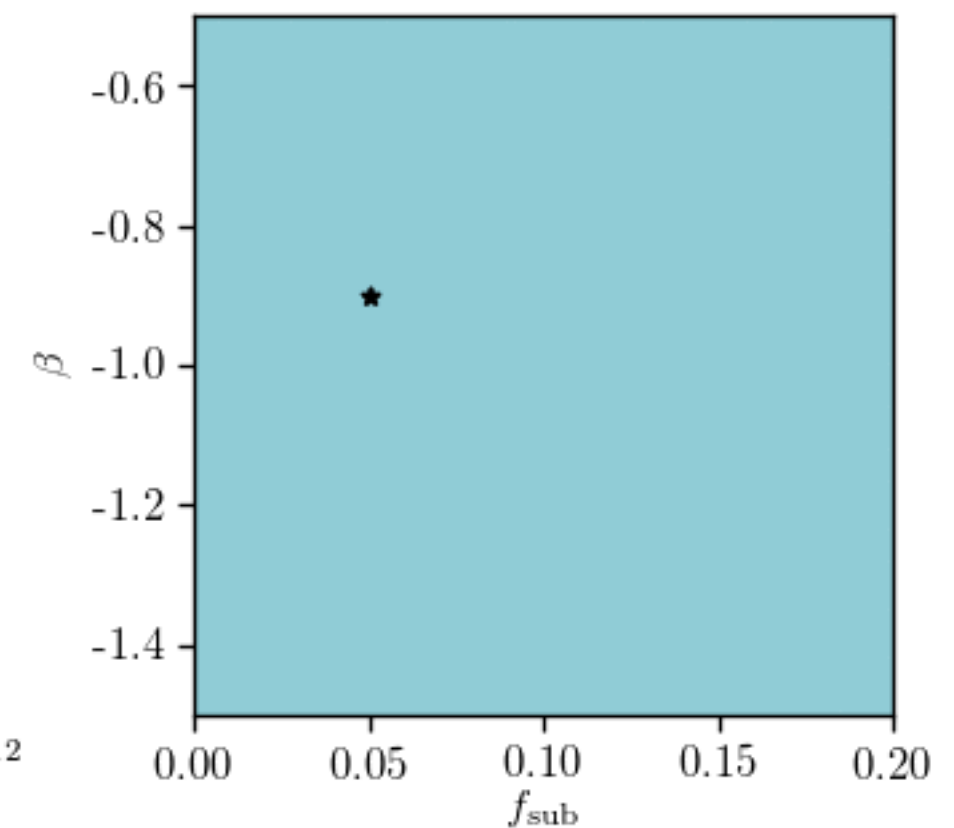
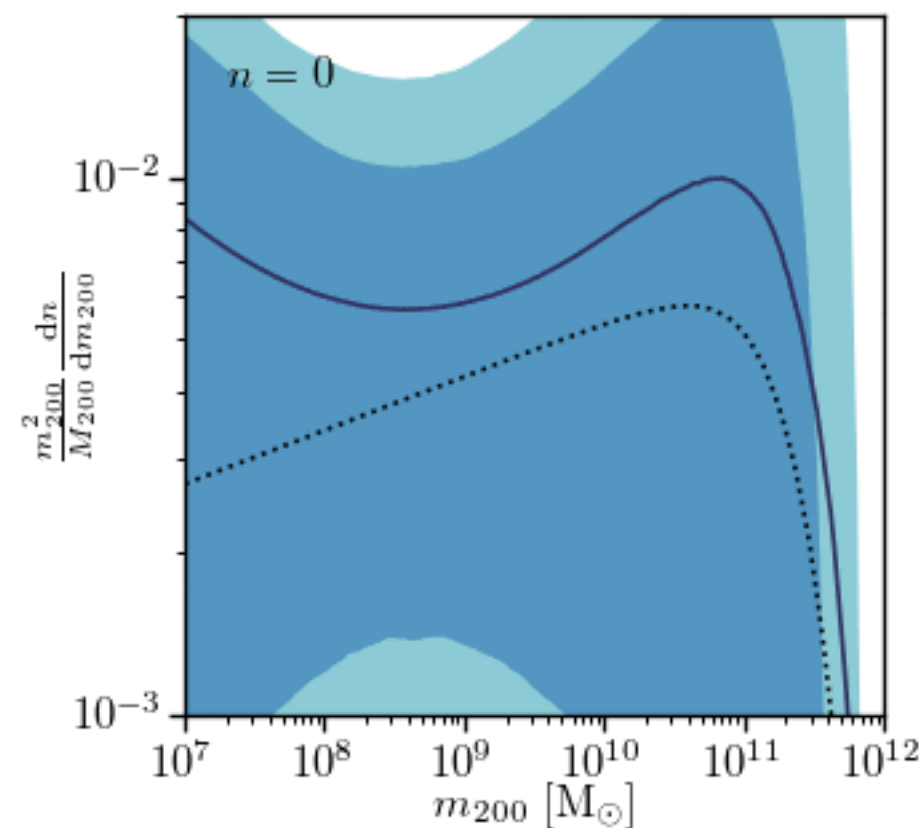
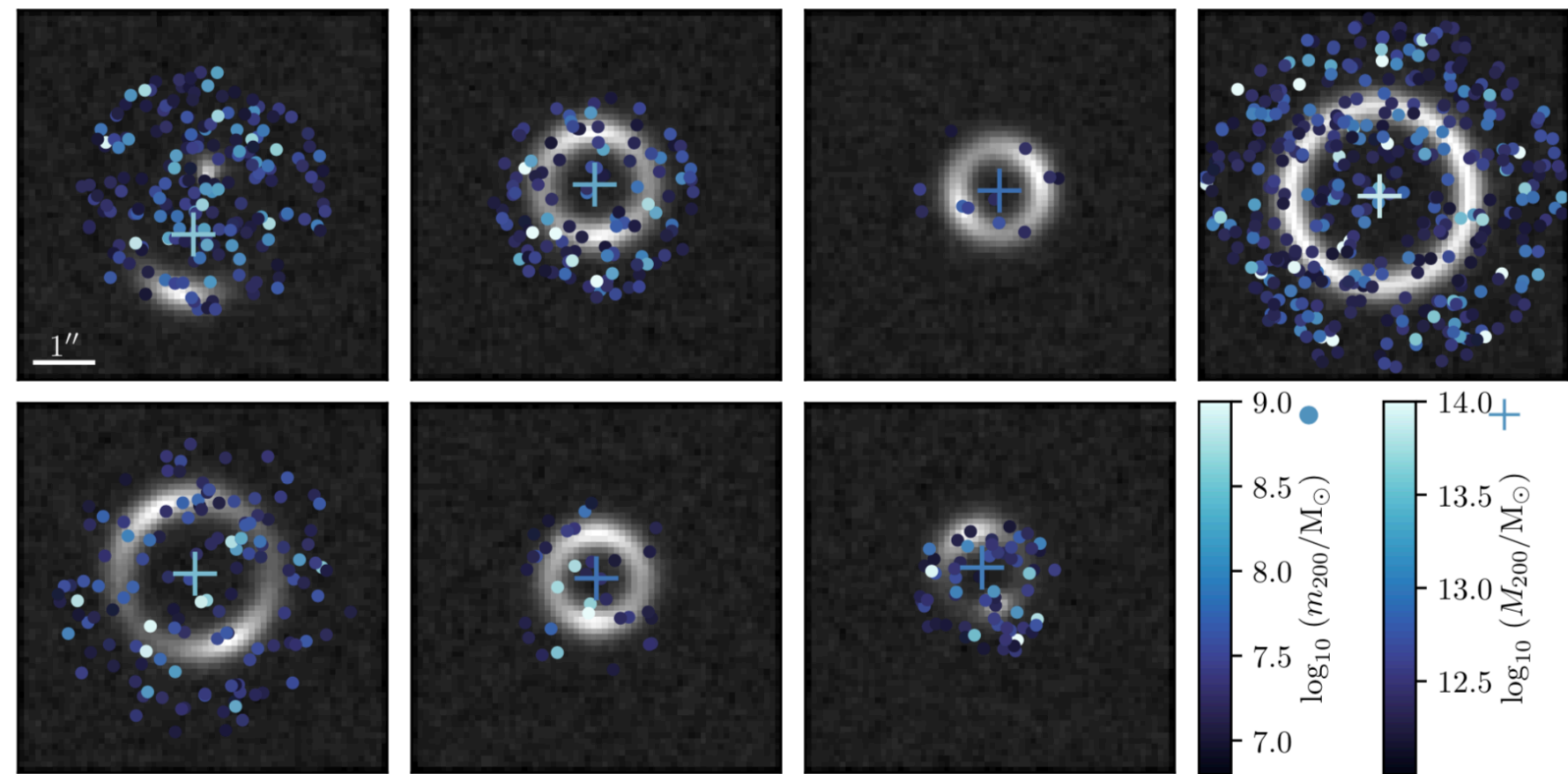
distorted light-rays

Earth

**Latent space  $\mathbf{Z}$ :**

Number of dark matter sub halos and their mass and location lead to complex latent space for each image.

Goal is inference at the population-level

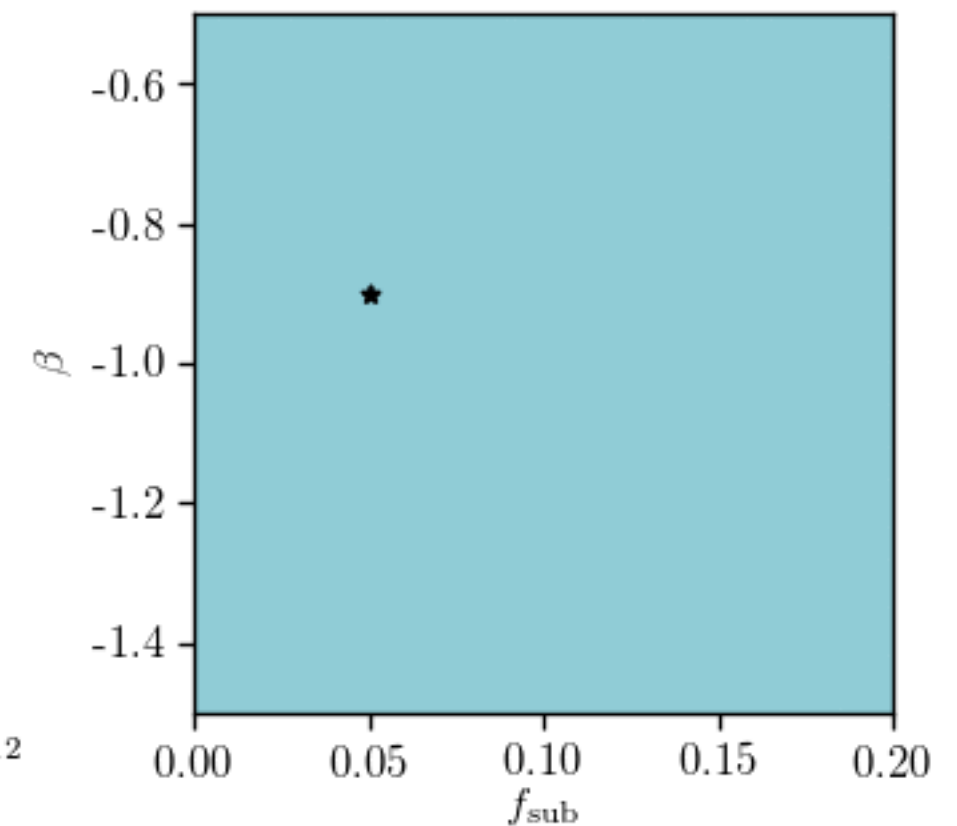
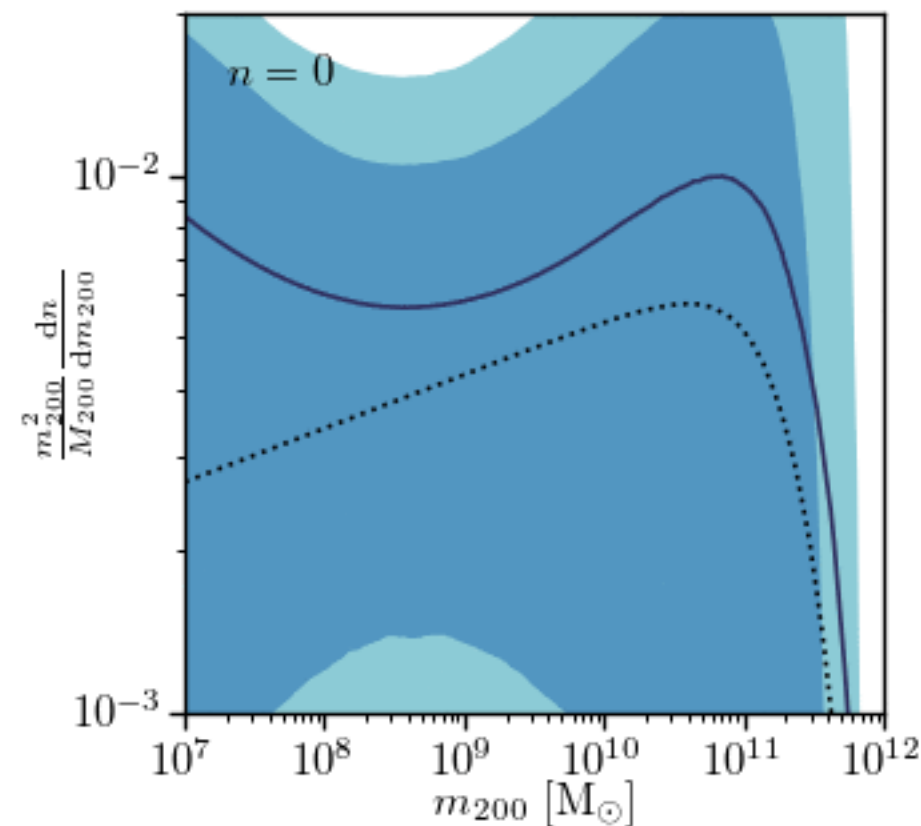
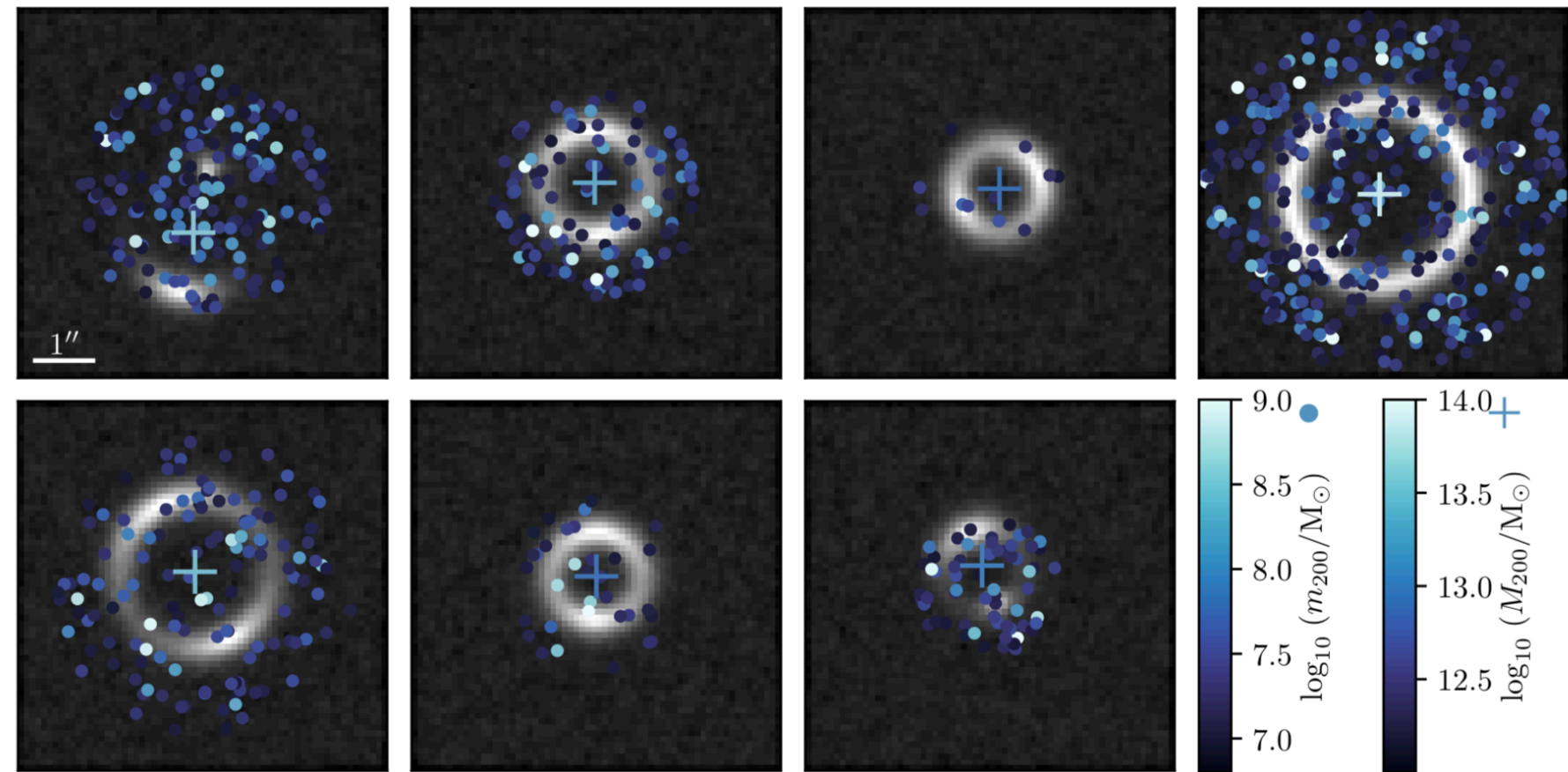




**Latent space Z:**

Number of dark matter sub halos and their mass and location lead to complex latent space for each image.

Goal is inference at the population-level

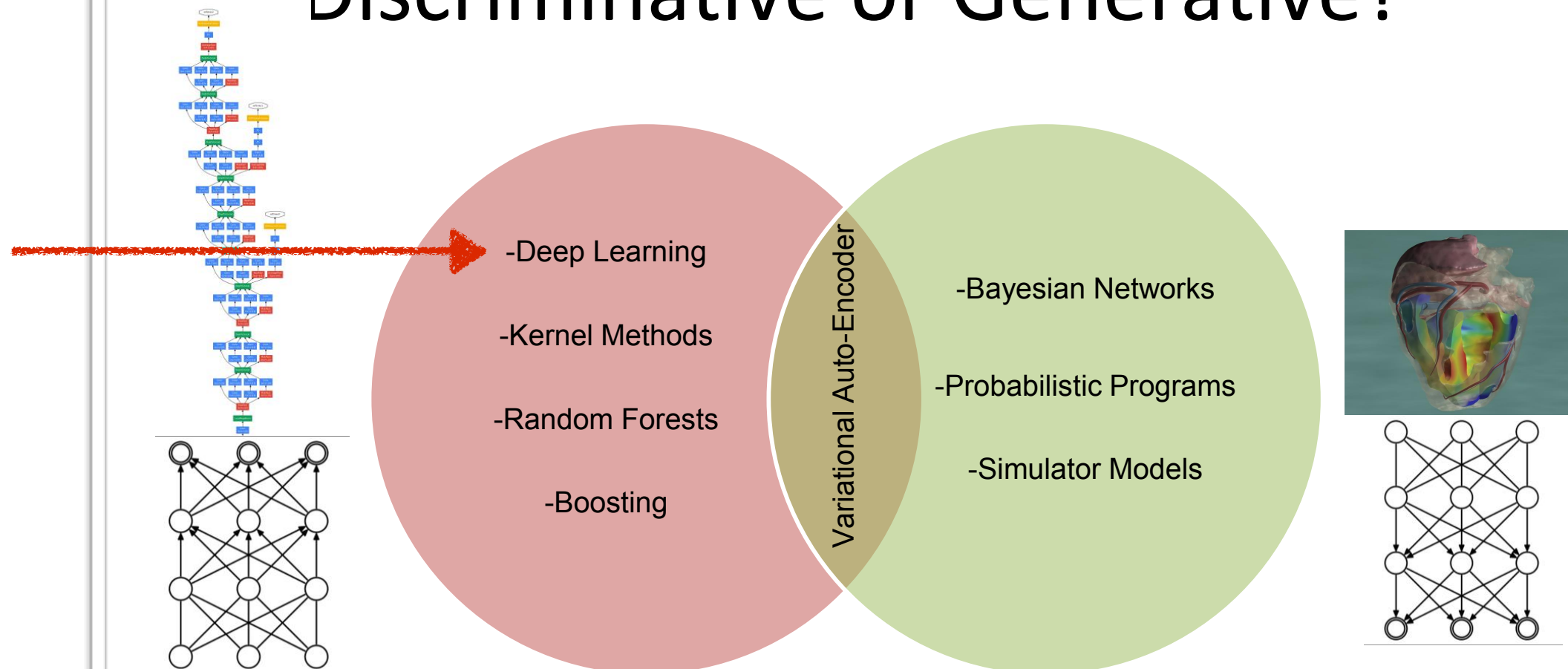






Max Welling

## Discriminative or Generative?



- Advantages discriminative models:
  - Flexible map from input to target (low bias)
  - Efficient training algorithms available
  - Solve the problem you are evaluating on.
  - Very successful and accurate!

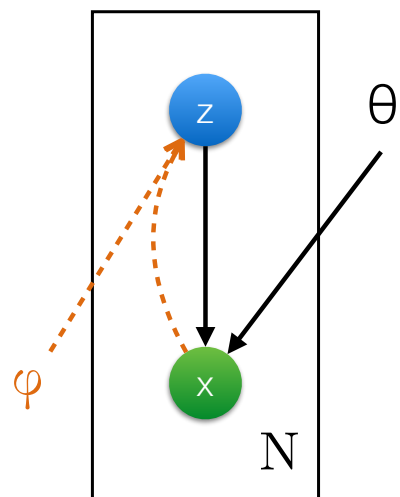
- Advantages generative models:
  - Inject expert knowledge
  - Model causal relations
  - Interpretable
  - Data efficient
  - More robust to domain shift
  - Facilitate un/semi-supervised learning

# Deep Generative Models

## Auto-Encoding Variational Bayes

[Kingma and Welling, 2013/2014]

[Rezende et al, 2014]



- $q_{\phi}(z|x) = \mathcal{N}(\mu, \sigma^2)$   
 $[\mu, \sigma^2] = f^{(z|x)}(x, \phi) = \text{multilayer neural net}$
- Objective: lower bound of  $\log p(x)$ .
  - Jointly optimized w.r.t.  $\phi$  and  $\theta$
  - This is approx. maximum likelihood
  - Simple SGD:
    - Sampling small minibatches of data
    - Sampling from approx. posterior
- This also minimizes an expected KL divergence  
 $D_{\text{KL}}(q_{\phi}(z|x) || p(z|x))$   
 -> gives us cheap approx. inference for new datapoints



Diederik (Durk)  
Kingma



Max  
Welling



Danilo J. Rezende

### Conv. net as encoder/decoder, trained on faces



Kingma and Welling, Auto-encoding Variational Bayes, ICLR 2014

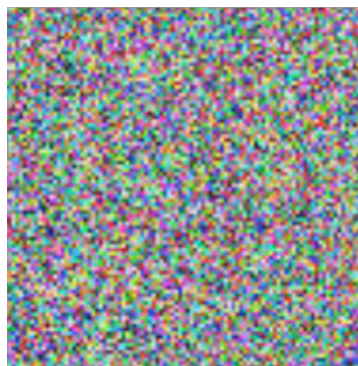
Rezende, Mohamed and Wierstra, Stochastic back-propagation and variational inference in deep latent Gaussian models, ICML 2014

# GENERATIVE ADVERSARIAL NETWORKS

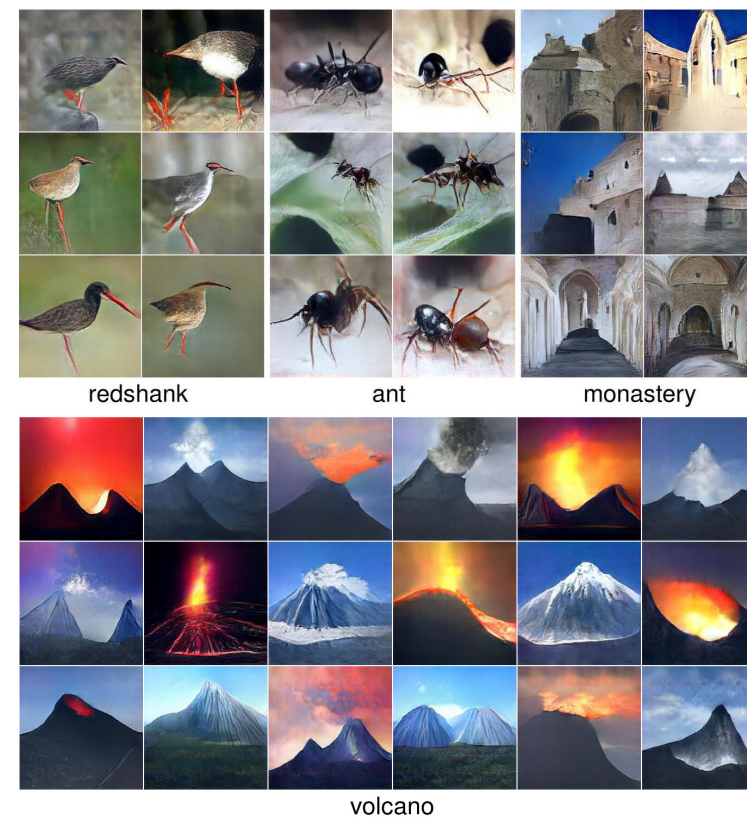
Z

X

Noise  $\sim N(0,1)$



Generative  
Model



catch me

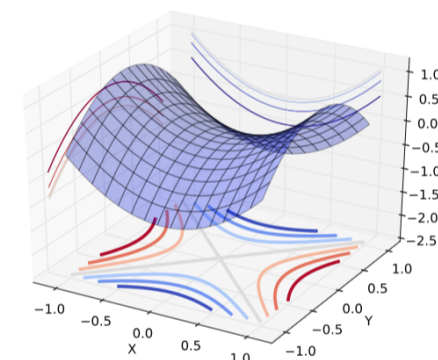
if you can

Leo is  $G$

Tom is  $D$

- We want to
  - For fixed  $G$ , find  $D$  which **maximizes**  $V(D, G)$ ,
  - For fixed  $D$ , find  $G$  which **minimizes**  $V(D, G)$ ;
- In other words, we are looking for the *saddle point*

$$(D^*, G^*) = \max_D \min_G V(D, G).$$





# TWO OBSERVATIONS

GANs and VAEs use deep neural network to transform latent  $Z$  to observed  $X$

- But resulting density  $p(x)$  is intractable
  - Say the density is “implicit”
  - Not directly useful for likelihood-free inference
- ... and latent space  $z$  for GAN and VAE has no specific meaning or interpretation

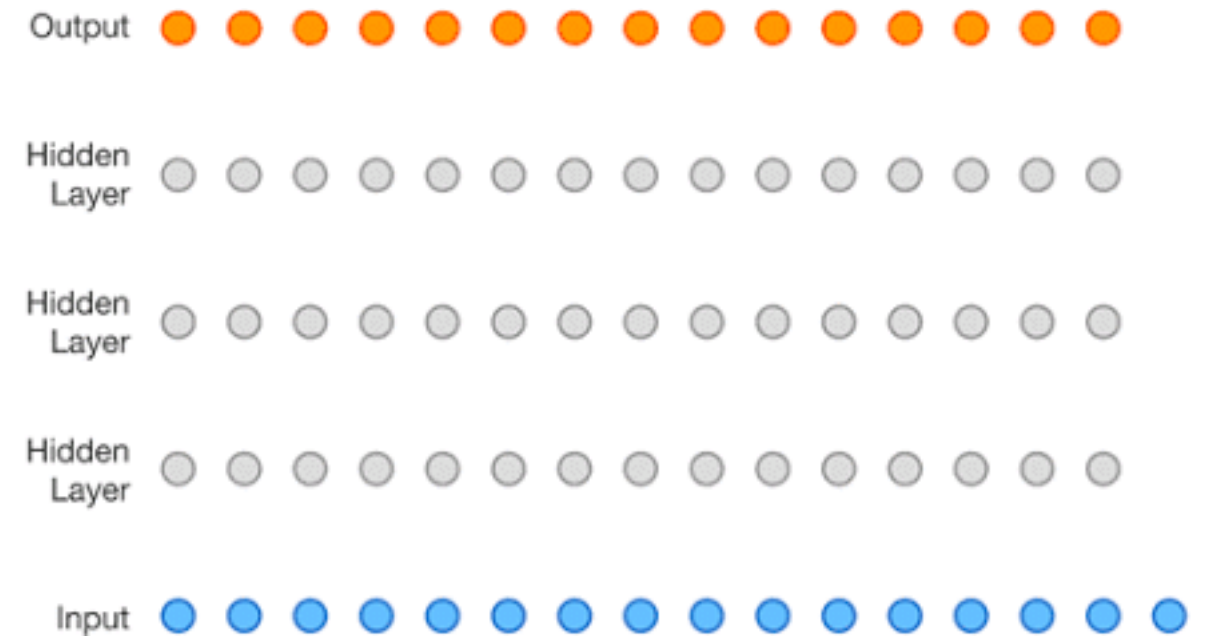
Are there other deep generative models that can help?

# WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

$$p(x) = \prod_{t=1}^T p(x_t \mid x_{t-1}, \dots, x_1).$$

have a tractable density. Train via maximum likelihood



1 Second

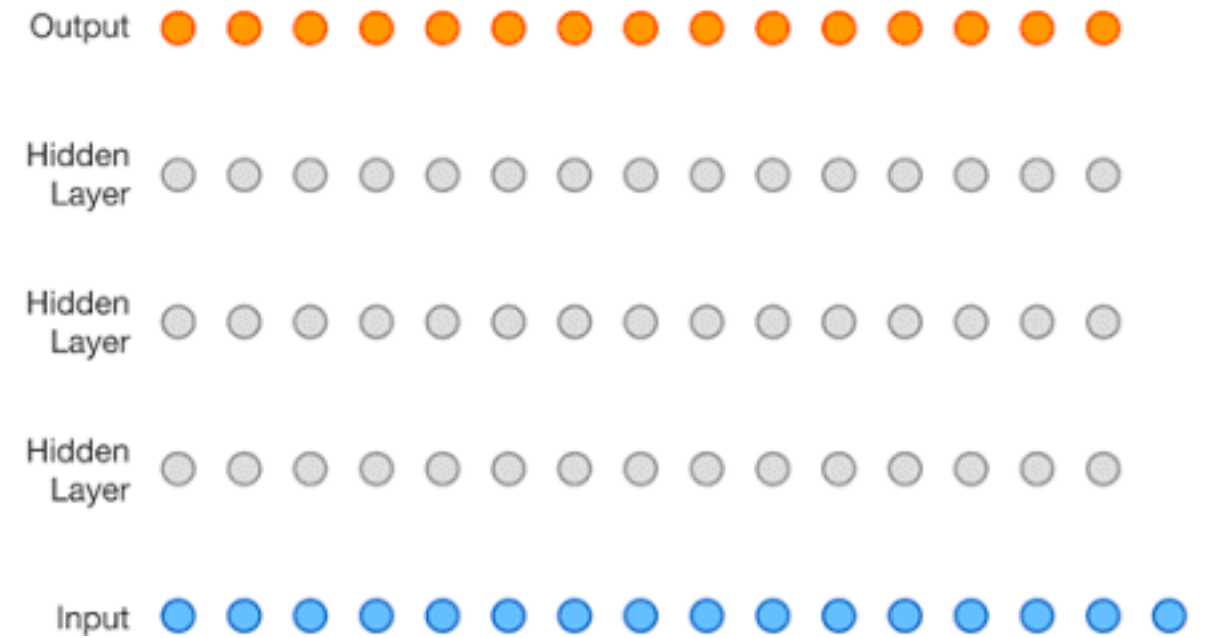


# WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

$$p(x) = \prod_{t=1}^T p(x_t \mid x_{t-1}, \dots, x_1).$$

have a tractable density. Train via maximum likelihood



1 Second

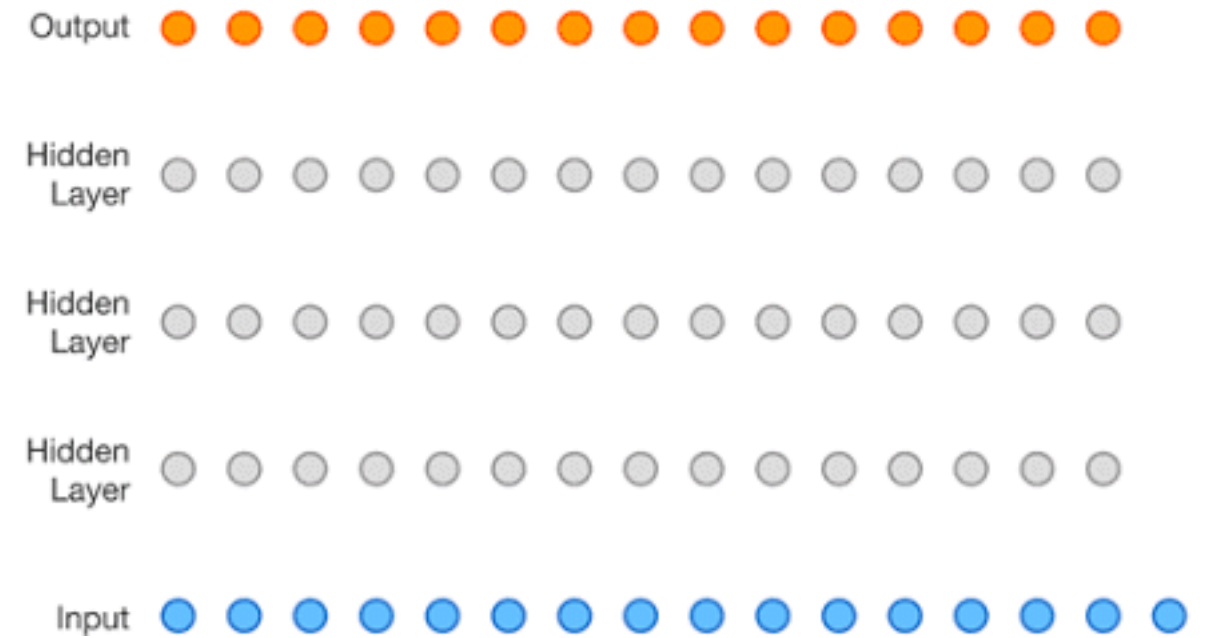


# WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

$$p(x) = \prod_{t=1}^T p(x_t \mid x_{t-1}, \dots, x_1).$$

have a tractable density. Train via maximum likelihood



1 Second





## Approximations using Change-of-variables

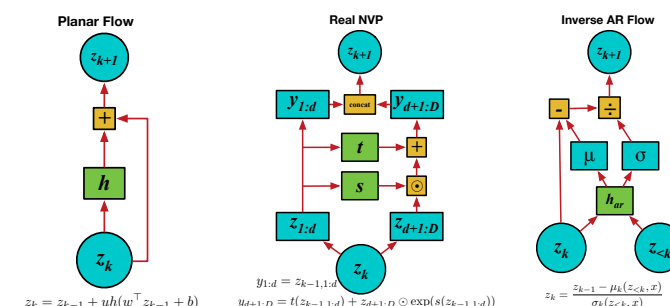
Exploit the rule for change of variables for random variables:

- Begin with an initial distribution  $q_0(\mathbf{z}_0|\mathbf{x})$ .
- Apply a sequence of  $K$  invertible functions  $f_k$ .

### Choice of Transformation Function

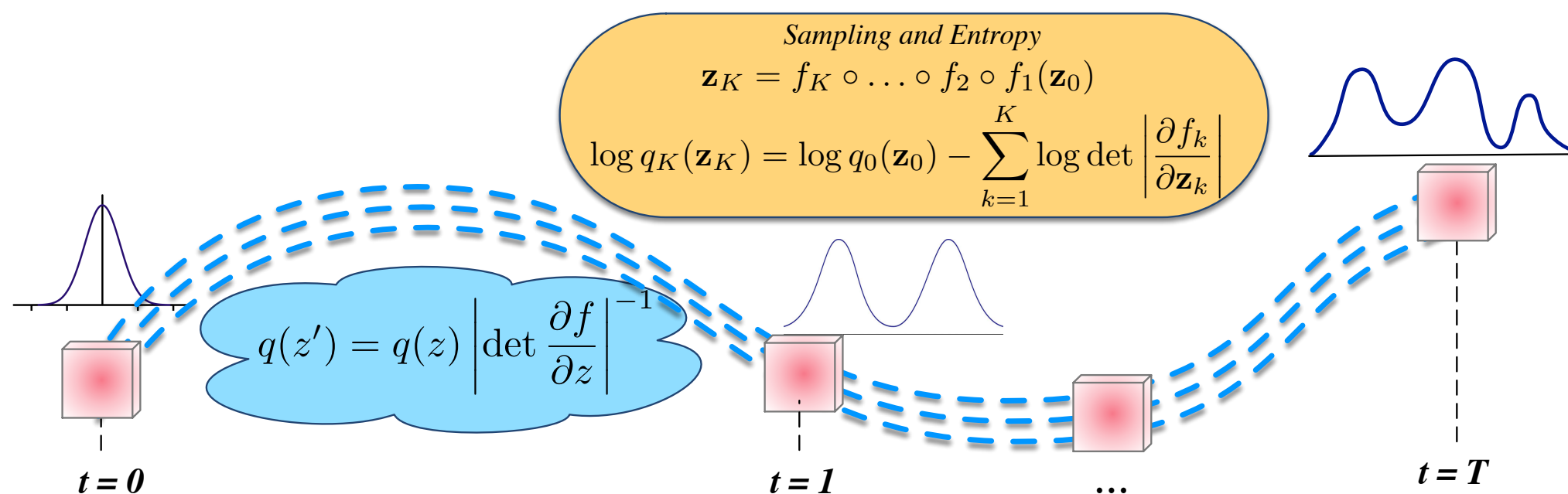
$$\mathcal{L} = \mathbb{E}_{q_0(\mathbf{z}_0)}[\log p(\mathbf{x}, \mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z}_0)}[\log q_0(\mathbf{z}_0)] - \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\sum_{k=1}^K \log \det \left| \frac{\partial f_k}{\partial \mathbf{z}_k} \right| \right]$$

- Begin with a fully-factorised Gaussian and improve by change of variables.
- Triangular Jacobians allow for computational efficiency.



[Rezende and Mohamed, 2016; Dinh et al., 2016; Kingma et al., 2016]

*Linear time computation of the determinant and its gradient.*



*Distribution flows through a sequence of invertible transforms*

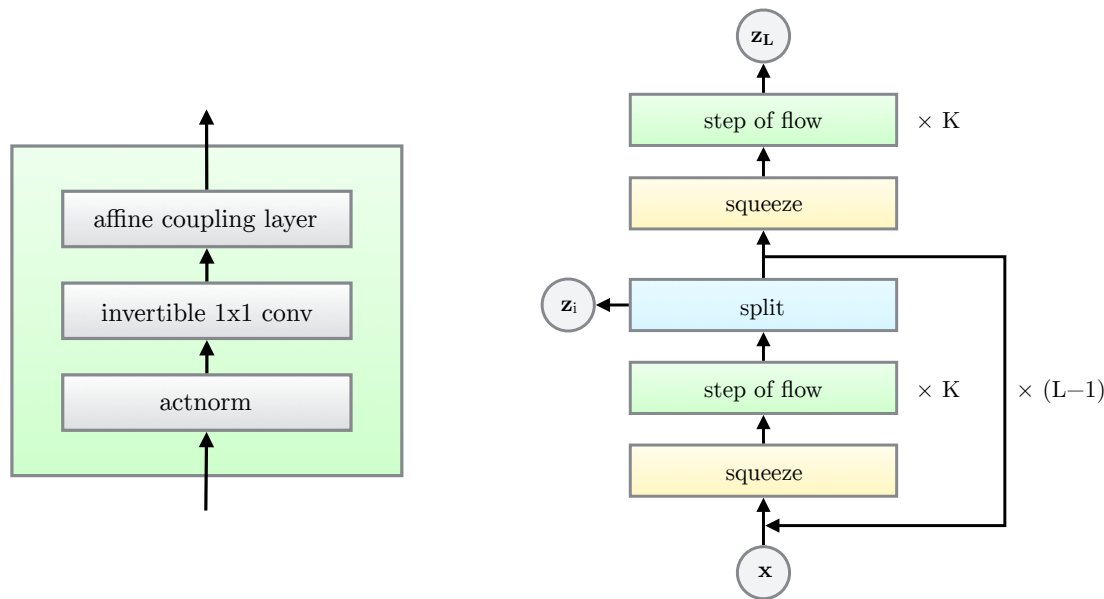
[Rezende and Mohamed, 2015]

# FLOWs WITH INVERTIBLE CONVOLUTIONS

<https://arxiv.org/abs/1807.03039>

## Glow: Generative Flow with Invertible $1 \times 1$ Convolutions

Diederik P. Kingma\*, Prafulla Dhariwal\*  
OpenAI, San Francisco



(a) One step of our flow.

(b) Multi-scale architecture (Dinh et al., 2016).



<https://arxiv.org/abs/1901.11137>

## Emerging Convolutions for Generative Normalizing Flows

Emiel Hooeboom<sup>1,2</sup> Rianne van den Berg<sup>1</sup> Max Welling<sup>1,3</sup>

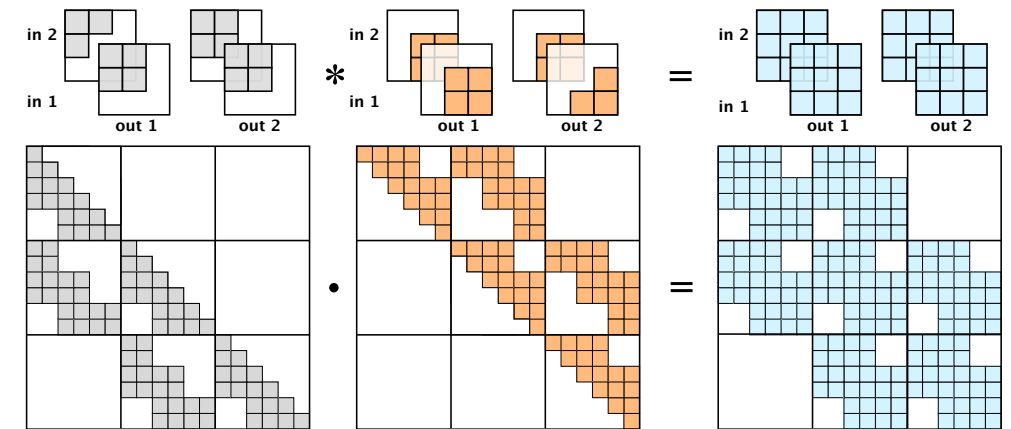


Table 3. Performance of Emerging convolutions on CIFAR10, ImageNet 32x32 and ImageNet 64x64 in bits per dimension (negative  $\log_2$ -likelihood), and  $\pm$  reports standard deviation.

|          | CIFAR10                            | ImageNet<br>32x32 | ImageNet<br>64x64 |
|----------|------------------------------------|-------------------|-------------------|
| Real NVP | 3.51                               | 4.28              | 3.98              |
| Glow     | $3.36 \pm 0.002$                   | <b>4.09</b>       | <b>3.81</b>       |
| Emerging | <b><math>3.34 \pm 0.002</math></b> | <b>4.09</b>       | <b>3.81</b>       |

# FLOWS WITH CONTINUOUS TIME

## FFJORD: FREE-FORM CONTINUOUS DYNAMICS FOR SCALABLE REVERSIBLE GENERATIVE MODELS

Will Grathwohl<sup>\*†‡</sup>, Ricky T. Q. Chen<sup>\*†</sup>, Jesse Bettencourt<sup>†</sup>, Ilya Sutskever<sup>‡</sup>, David Duvenaud<sup>†</sup>

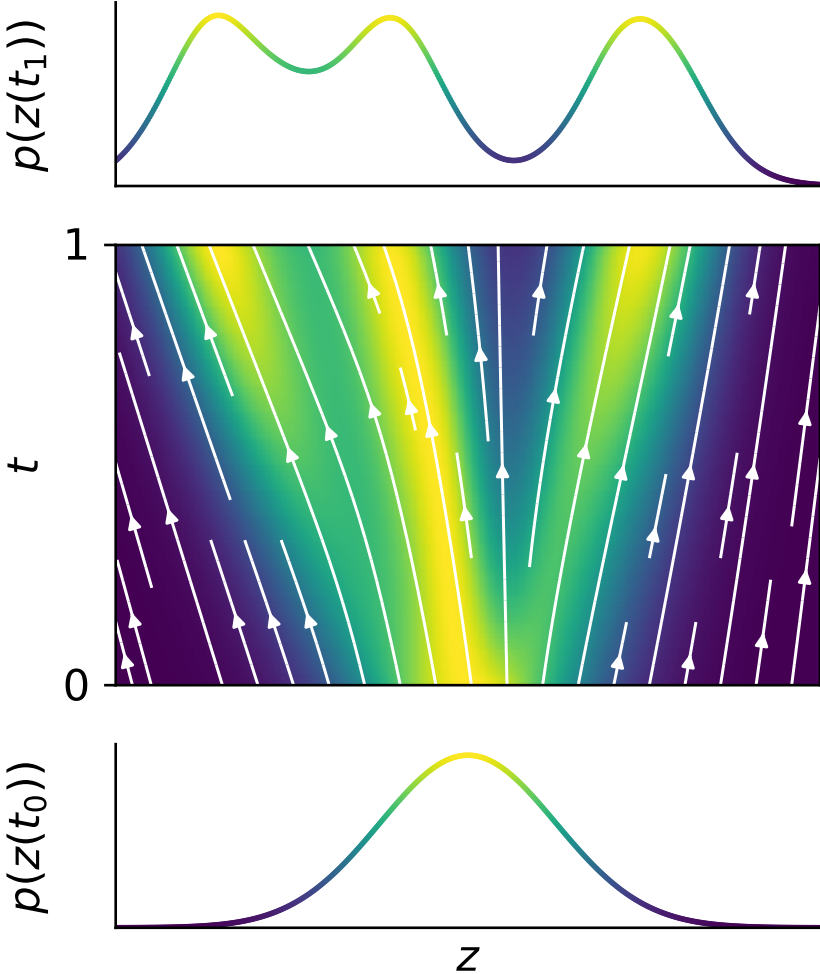
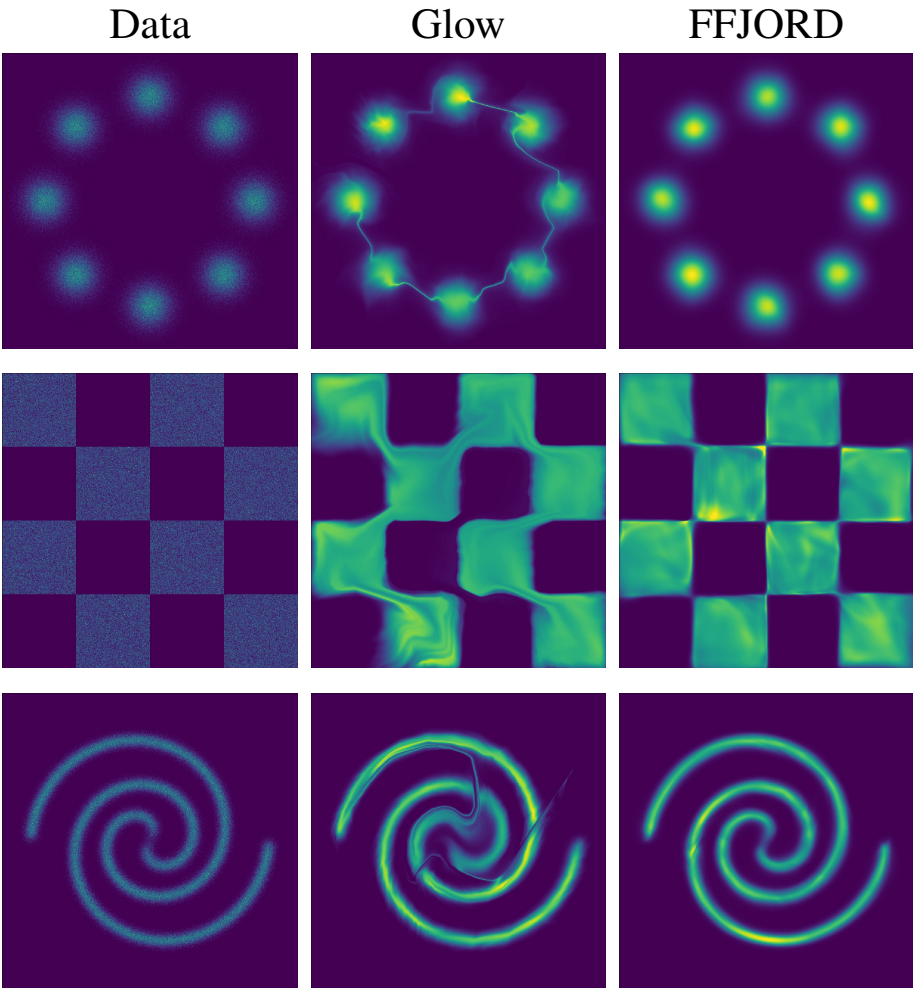


Figure 1: FFJORD transforms a simple base distribution at  $t_0$  into the target distribution at  $t_1$  by integrating over learned continuous dynamics.



| Method              |                                  | Train on data | One-pass Sampling | Exact log-likelihood | Free-form Jacobian |
|---------------------|----------------------------------|---------------|-------------------|----------------------|--------------------|
| Change of Variables | Variational Autoencoders         | ✓             | ✓                 | ✗                    | ✓                  |
|                     | Generative Adversarial Nets      | ✓             | ✓                 | ✗                    | ✓                  |
|                     | Likelihood-based Autoregressive  | ✓             | ✗                 | ✓                    | ✗                  |
|                     | Normalizing Flows                | ✗             | ✓                 | ✓                    | ✗                  |
|                     | Reverse-NF, MAF, TAN             | ✓             | ✗                 | ✓                    | ✗                  |
|                     | NICE, Real NVP, Glow, Planar CNF | ✓             | ✓                 | ✓                    | ✗                  |
|                     | <b>FFJORD</b>                    | ✓             | ✓                 | ✓                    | ✓                  |

Table 1: A comparison of the abilities of generative modeling approaches.



# EQUIVARIANT FLOWS

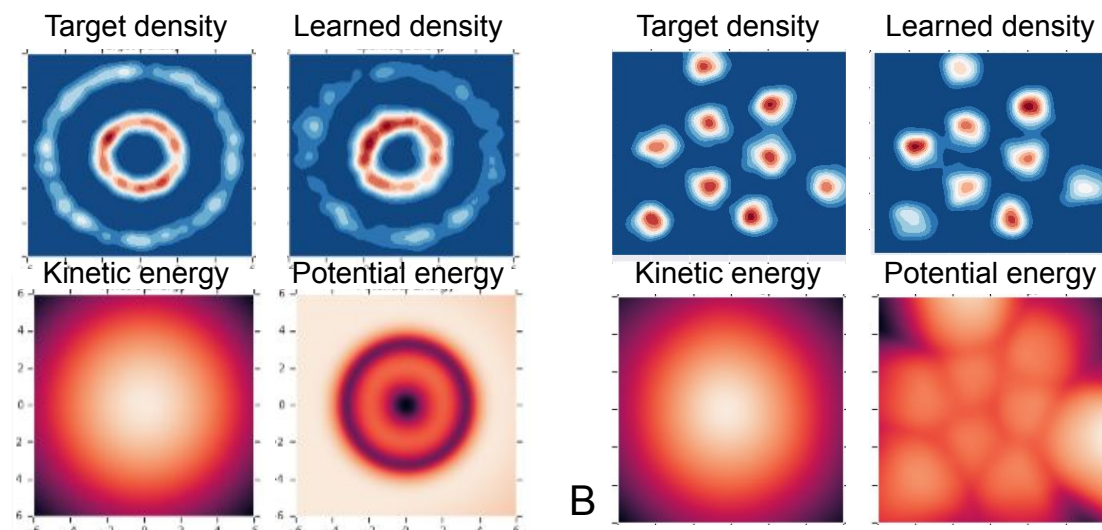
## Equivariant Hamiltonian Flows

Danilo J. Rezende\* Sébastien Racanière\* Irina Higgins\* Peter Toth\*

\*{danilor, sracaniere, irinah, petertoth}@google.com

### Abstract

This paper introduces equivariant hamiltonian flows, a method for learning expressive densities that are invariant with respect to a known Lie-algebra of local symmetry transformations while providing an equivariant representation of the data. We provide proof of principle demonstrations of how such flows can be learnt, as well as how the addition of symmetry invariance constraints can improve data efficiency and generalisation. Finally, we make connections to disentangled representation learning and show how this work relates to a recently proposed definition.



## Equivariant Flows: sampling configurations for multi-body systems with symmetric energies

Jonas Köhler \*<sup>†</sup>

Leon Klein \*<sup>†</sup>

Frank Noé <sup>†‡§</sup>

<sup>†</sup> Freie Universität Berlin, Department of Mathematics and Computer Science.

<sup>‡</sup> Freie Universität Berlin, Department of Physics.

<sup>§</sup> Rice University, Department of Chemistry.

{jonas.koehler, leon.klein, frank.noel}@fu-berlin.de

\* Authors contributed equally to this work.

1. Permutation invariance: Swapping the labels of any two interchangeable particles  $(x_1, \dots, x_K) \rightarrow (x_{\sigma(1)}, \dots, x_{\sigma(K)})$ .
2. Rotation invariance: Any 2D/3D rotation of the system  $Rx = (Rx_1, \dots, Rx_K)$ .
3. Translation invariance: Any 2D/3D translation  $x + v = (x_1 + v, \dots, x_K + v)$ .





Can we use the idea for flows directly on **an orthonormal basis** of complex quantum wave functions?

- instead of  $p(z) \rightarrow p(x)$  can we do  $\phi_i(z) \rightarrow \psi_i(x)$  ?

- yes!

1) Start with:

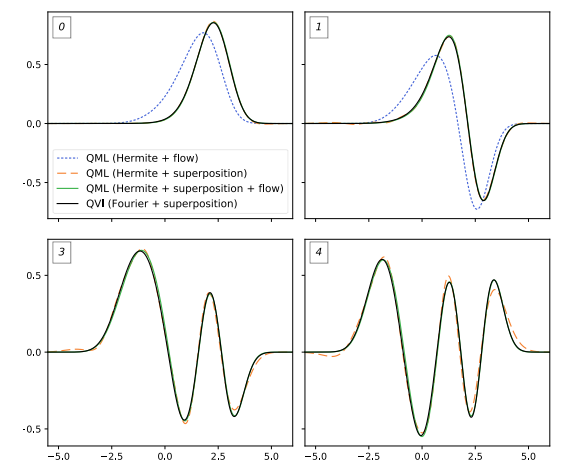
$$\int dz \phi_i(z) \phi_j^*(z) = \delta_{ij}$$

2) Change variables:  $\int dz \phi_i(z) \phi_j^*(z) = \int dx \left| \det \frac{\partial f}{\partial x} \right| \phi_i(f(x)) \phi_j^*(f(x)) = \delta_{ij}$

3) Profit:

$$\psi_i(x) = \phi_i(f(x)) \left| \det \frac{\partial f}{\partial x} \right|^{\frac{1}{2}}$$

complex!  
&  
orthonormal!

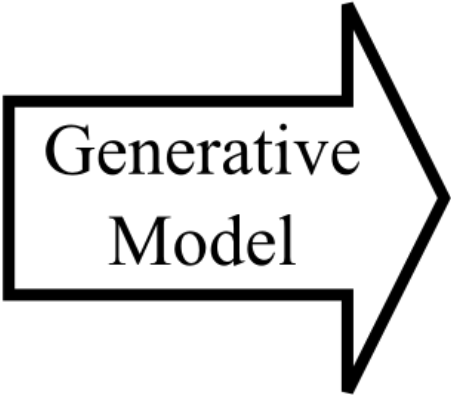


Correlation  $\neq$  Causation

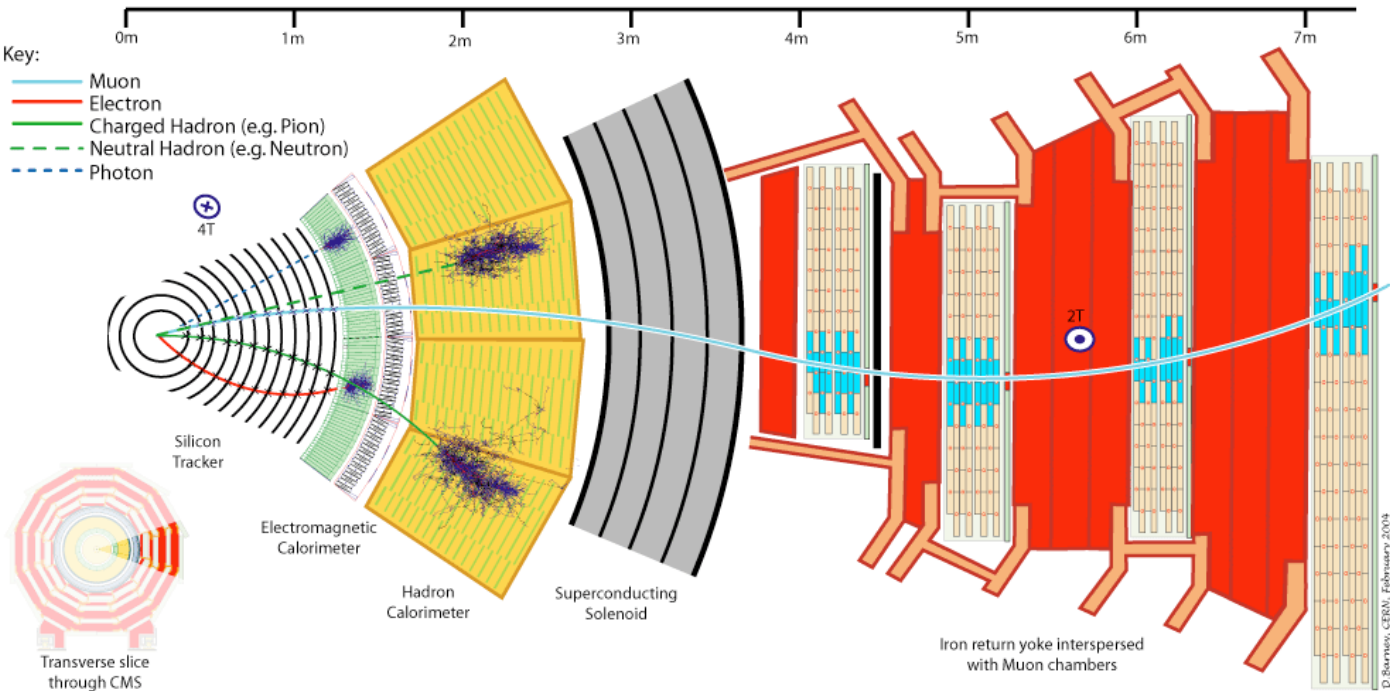
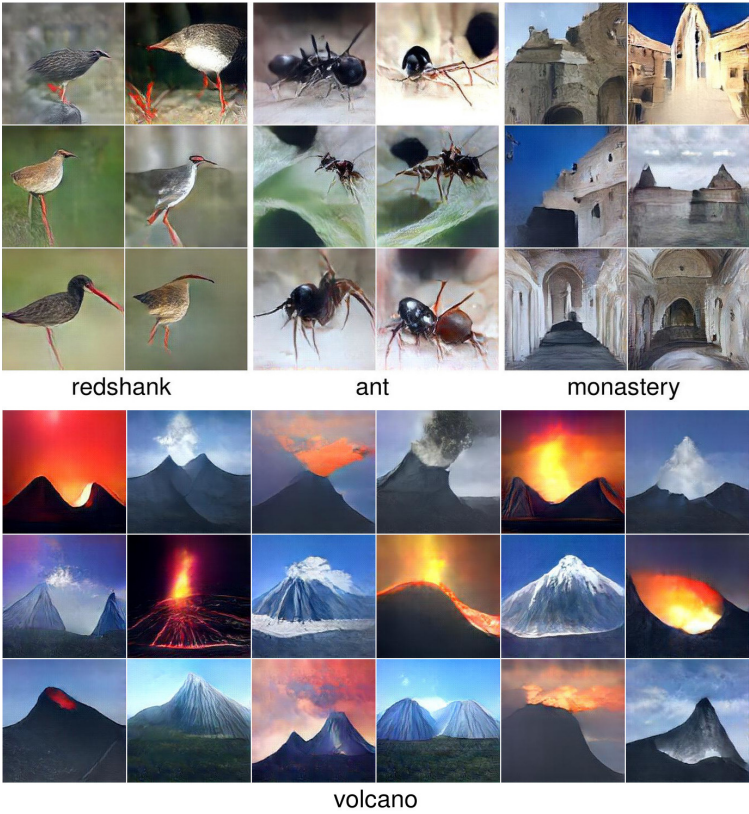
# DEEP GENERATIVE MODEL VS. SIMULATION

Z

Noise  $\sim N(0,1)$



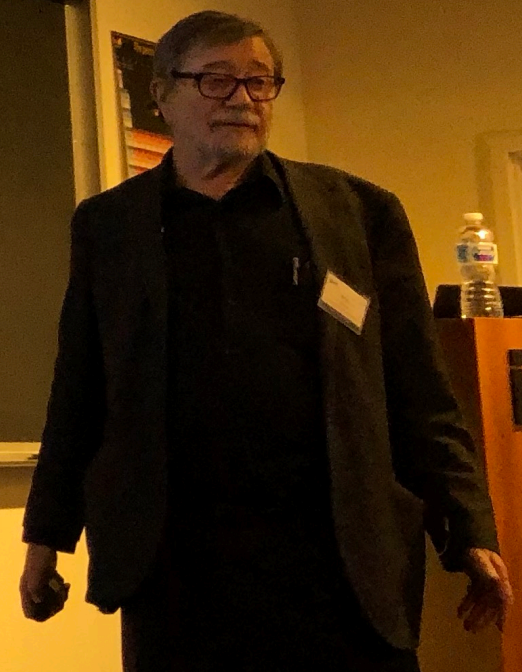
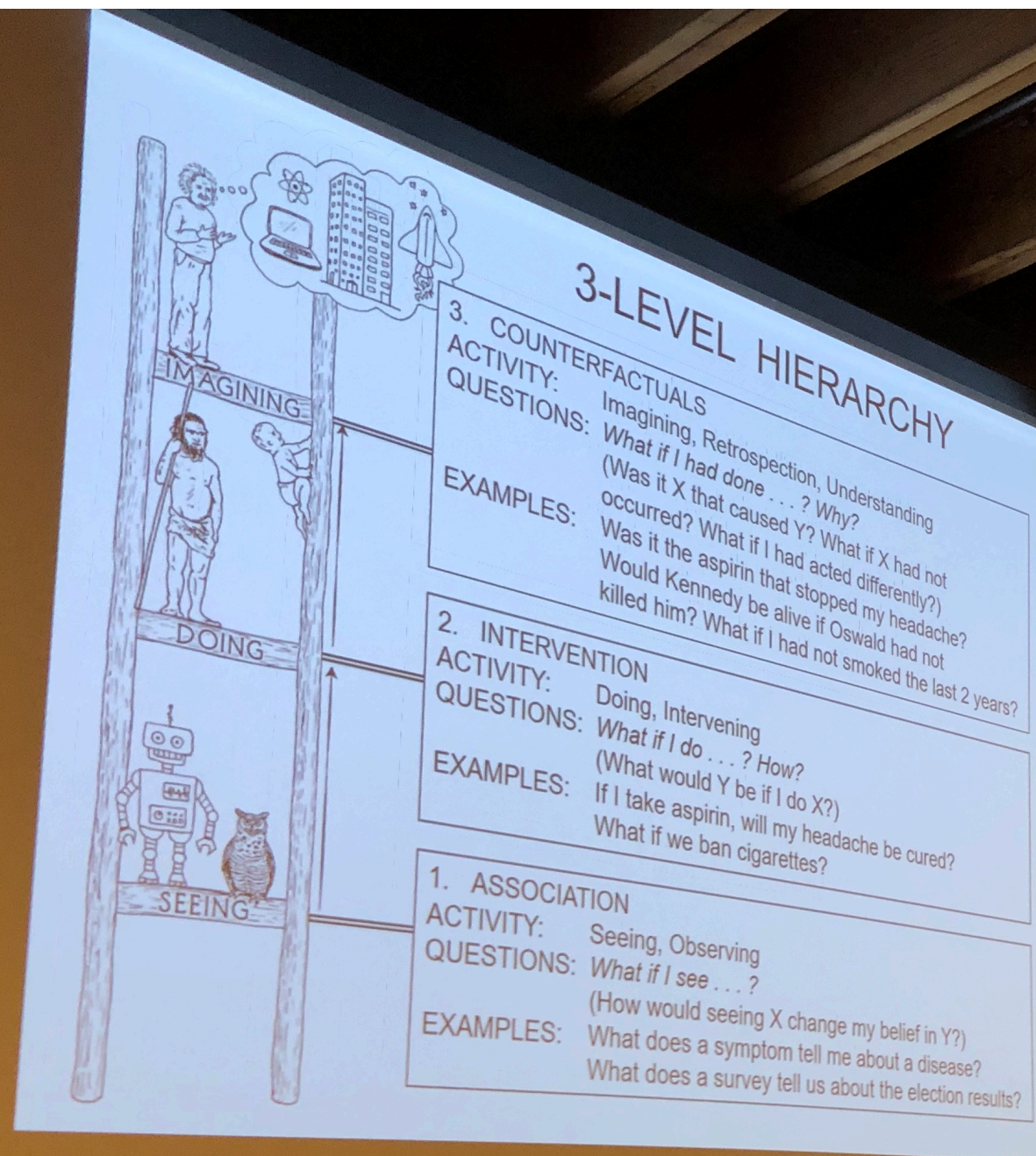
X





# THE CAUSAL HIERARCHY

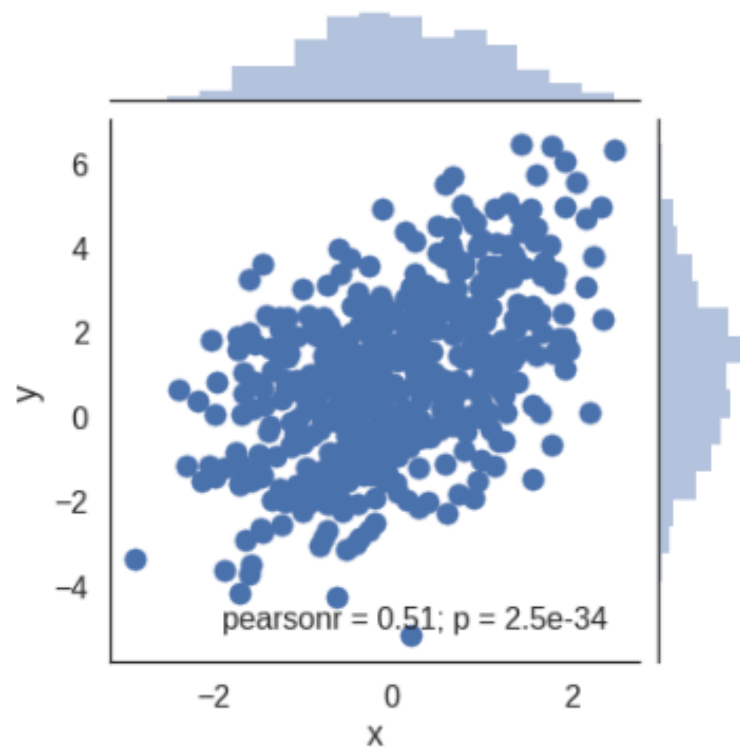
Judea Pearl



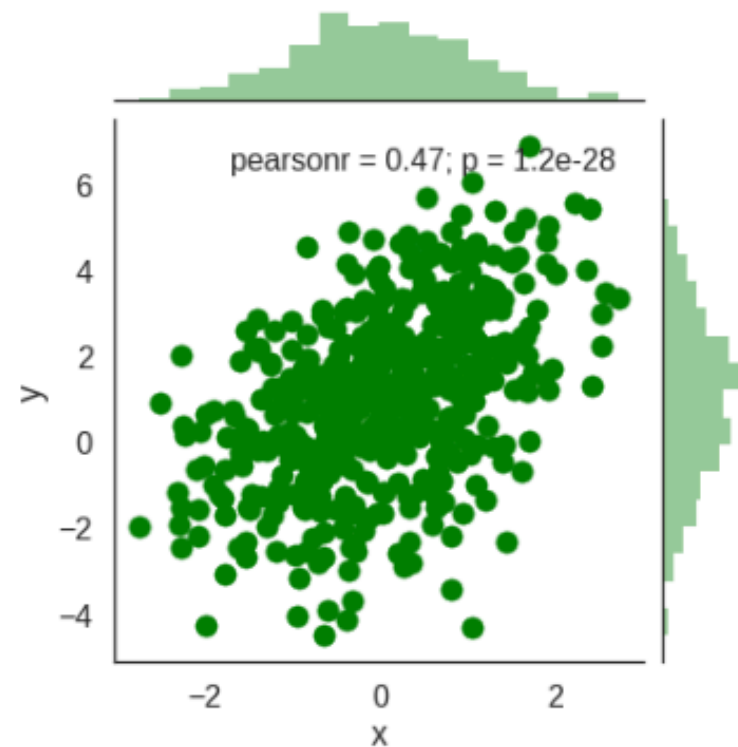


# SAME JOINT DISTRIBUTION, DIFFERENT CAUSAL MODEL

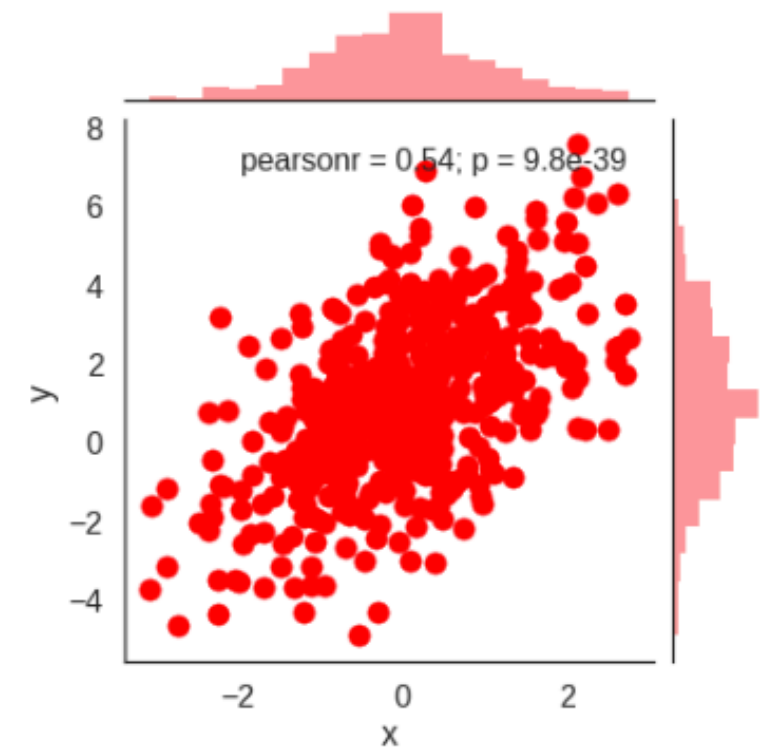
```
x = randn()  
y = x + 1 + sqrt(3)*randn()
```



```
y = 1 + 2*randn()  
x = (y-1)/4 + sqrt(3)*randn()/2
```

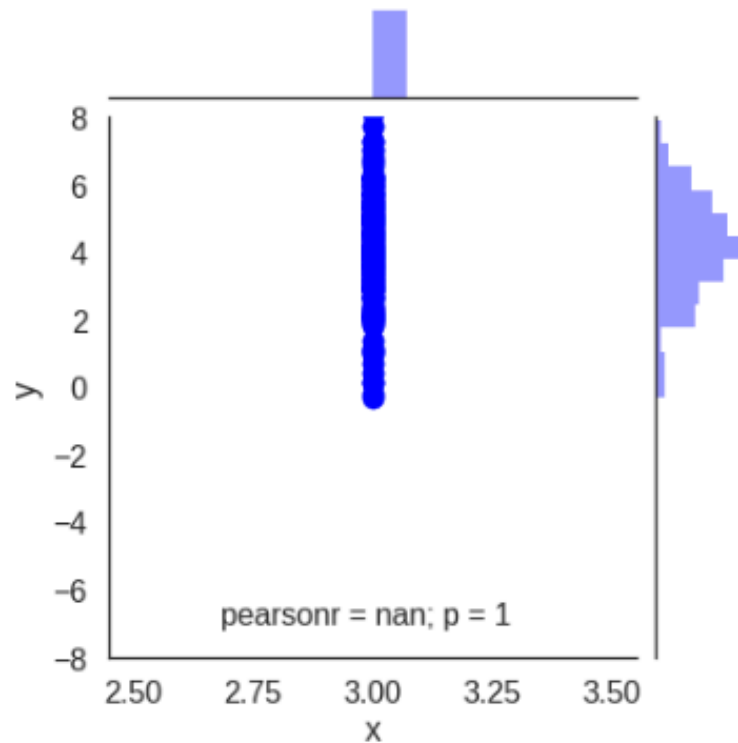


```
z = randn()  
y = z + 1 + sqrt(3)*randn()  
x = z
```

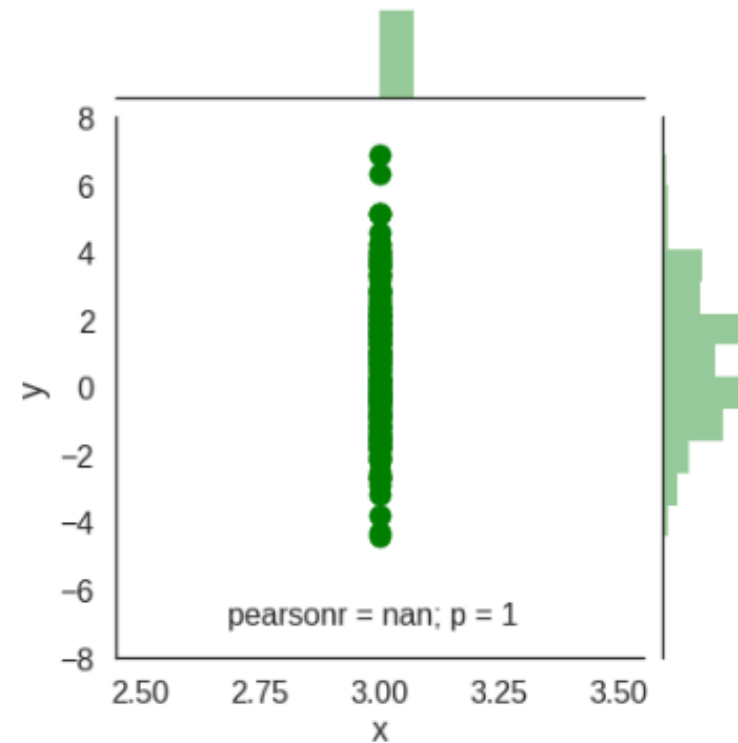


# CAUSATION > CORRELATION

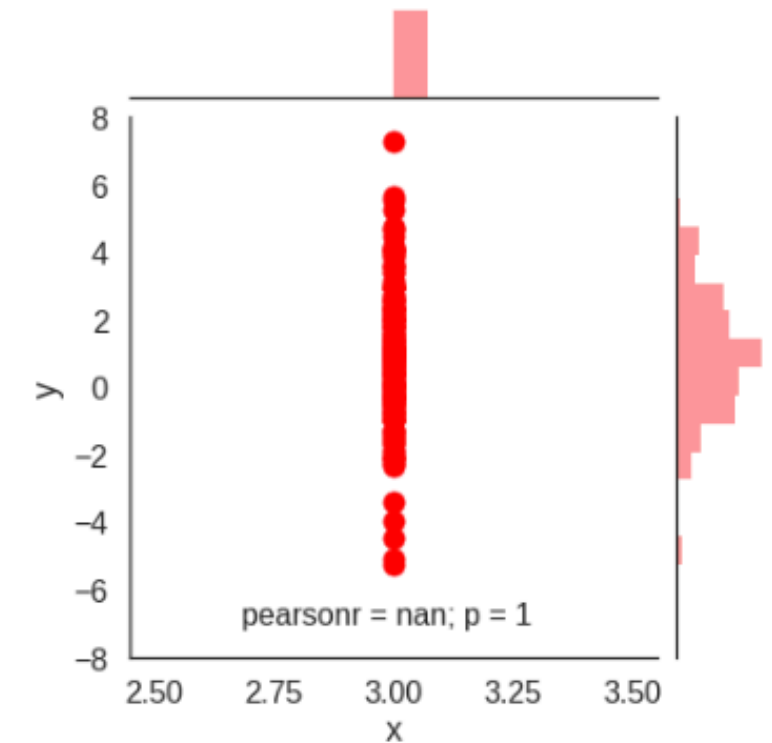
```
x = randn()  
x = 3  
y = x + 1 + sqrt(3)*randn()  
x = 3
```



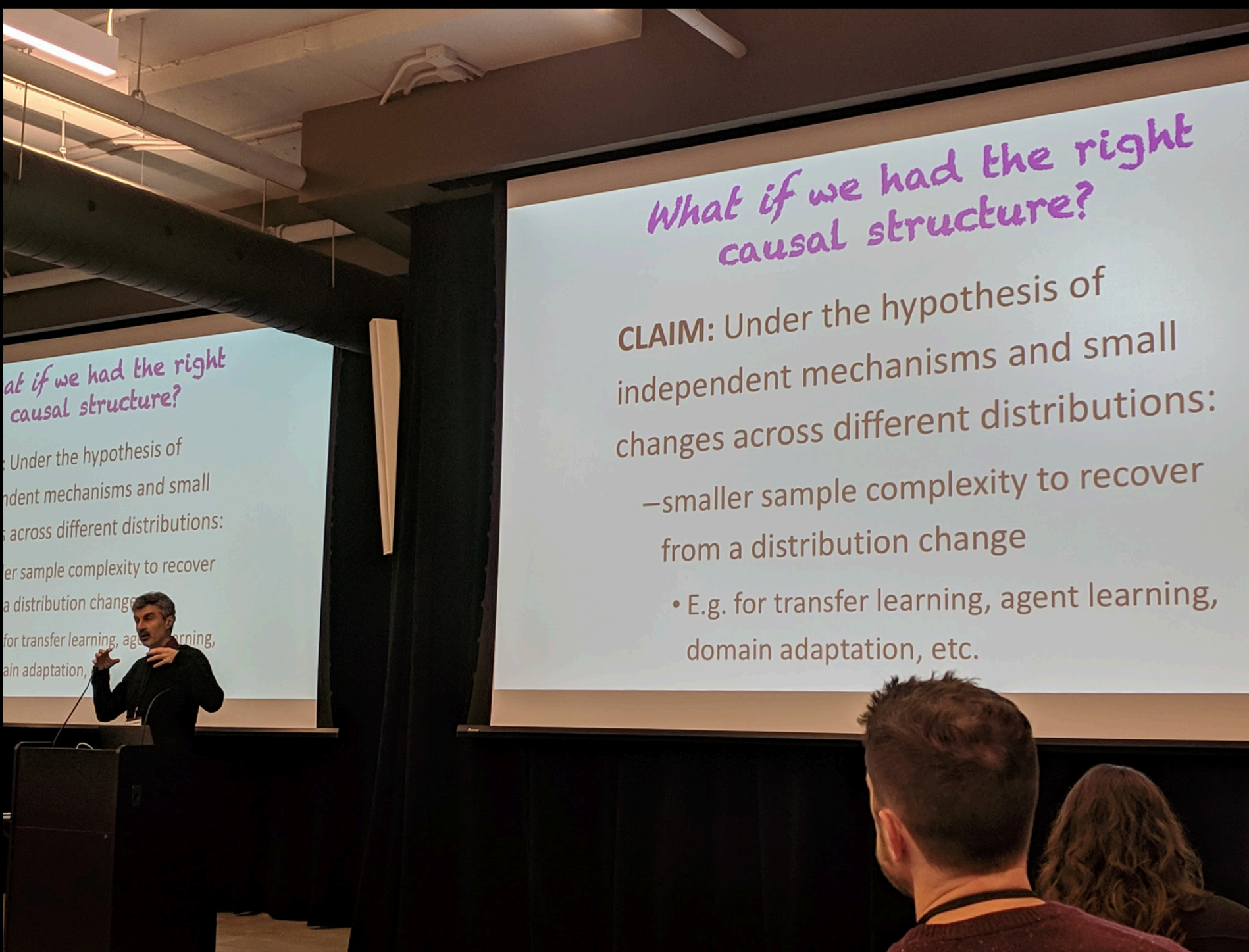
```
y = 1 + 2*randn()  
x = 3  
x = (y-1)/4 + sqrt(3)*randn()/2  
x = 3
```



```
z = randn()  
x = 3  
x = z  
x = 3  
y = z + 1 + sqrt(3)*randn()  
x = 3
```







## What if we had the right causal structure?

**CLAIM:** Under the hypothesis of independent mechanisms and small changes across different distributions:

—smaller sample complexity to recover from a distribution change

- E.g. for transfer learning, agent learning, domain adaptation, etc.



**Max Welling** Isn't this what Bernhard Schoelkopf has been saying for a while?

Like · Reply · 6w



**Yann LeCun** ...and Leon Bottou ?

Like · Reply · 6w



**Leon Bottou** Yoshua's paper says: if you observe a distribution change that comes from a causal effect, then you'll adapt faster if your generative model matches the causal model.

Another way of seeing it is : the right causal graph suggests a particular factorization of the joint distribution (a directed bayesian network). A causal intervention means that you only change one of these factors (or a few factors) while leaving the other ones unchanged. Therefore if your generative model is the right causal model, meaning that it factorizes the joint in the same way, it will be easy to adapt it to the change because only a few parameters need changing (those associated with the factors that actually changed).

Said like this, it feels pretty trivial. Yoshua proposes to use this to infer the right causal model from a plurality of observed distributions.



**Dan Roy** **Max Welling** yes. He's been arguing for generative models with causal structure for years as the way to extract information for rich environments. So not this



**Max Welling** **Dan Roy** I am, and I think most of us, are keenly aware that Josh has been the big proponent of this view. And I think most people agree with him on this view. Integrating this view with deep learning for more narrowly defined tasks seems to me an interesting intellectual pursuit though. I think that's what's happening here but I was not at the talk 😊

Yoshua Bengio on [arXiv:1901.10912]  
and public FB discussion

# Inductive Bias

## Compositionality

## Symmetry

## Causality

separation



## The message from human cognition:

Richly structured models of objects and their relations are a powerful tool for reasoning about, and interacting with, the world.

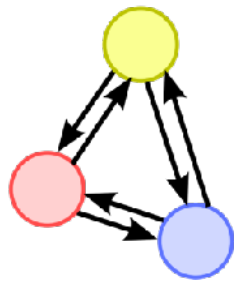
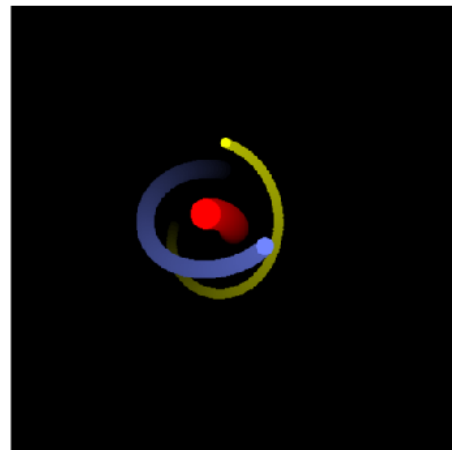
- Objects and relations reflect *decisions* made by evolution, experience, and task demands about how to represent the world in an *efficient and useful way*
- Intelligence is about *model-building*, beyond just recognizing patterns (Tenenbaum)
- *Combinatorial generalization* via abstraction and compositionality ("infinite use of finite means")



# Insight of data generating process informs inductive bias on architecture

## Physical systems as graphs

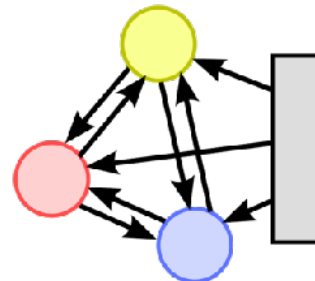
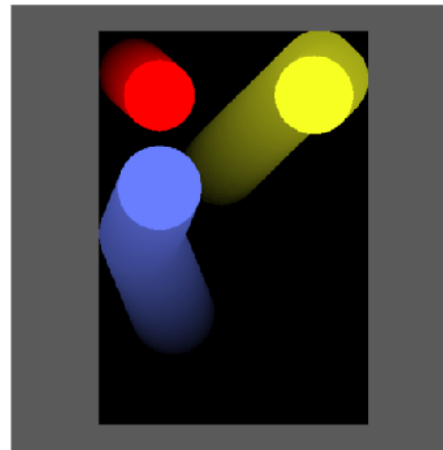
n-body



Nodes: bodies

Edges: gravitational forces

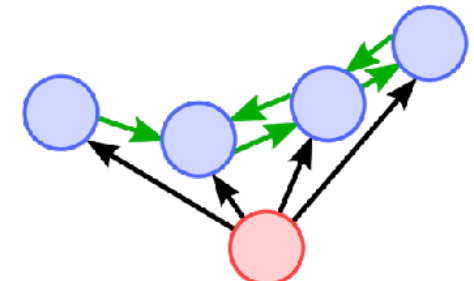
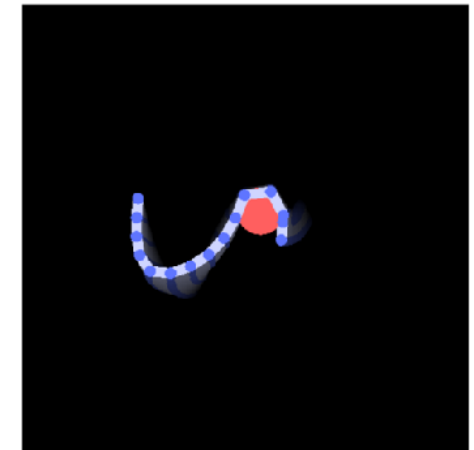
Balls



Nodes: balls

Edges: rigid collisions between balls, and walls

String

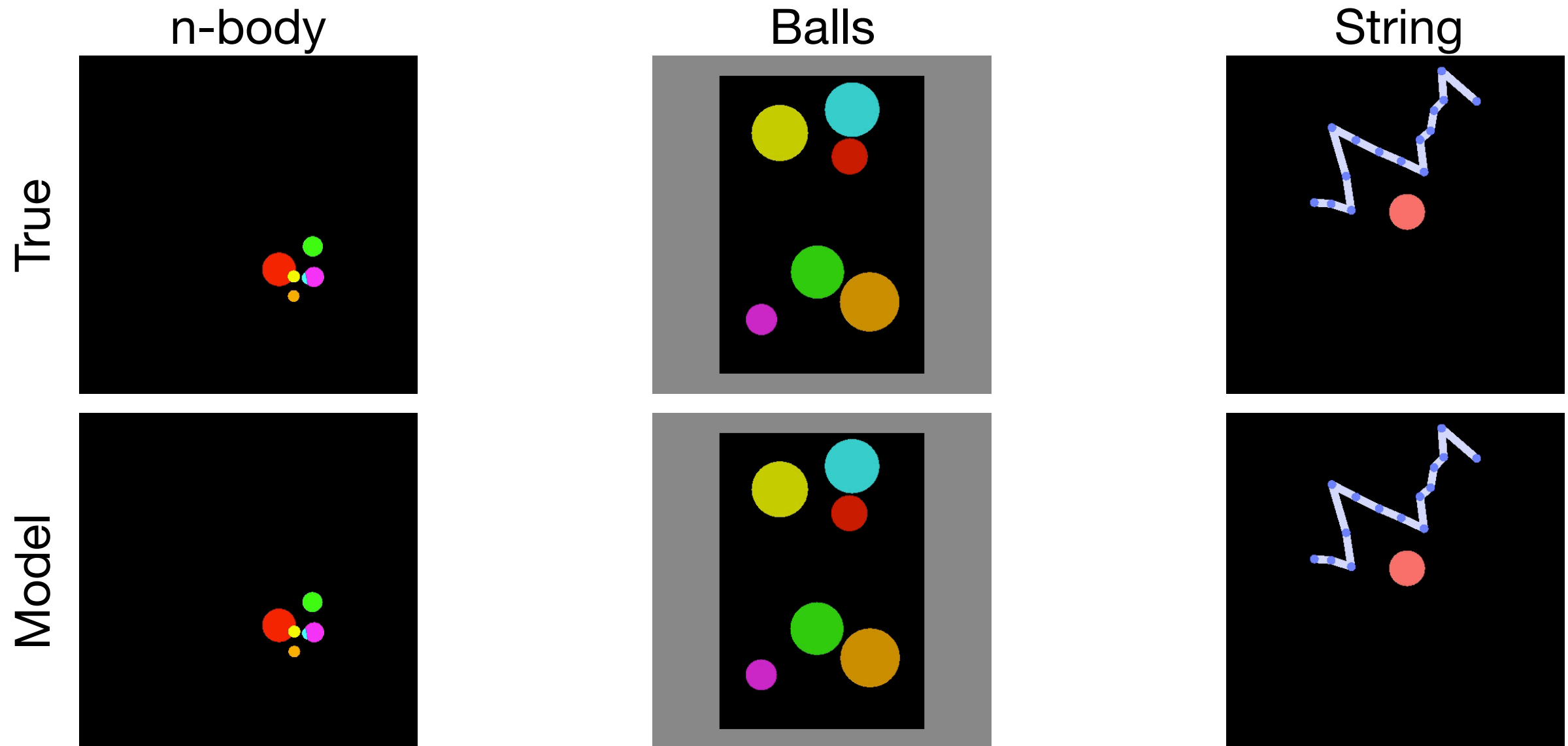


Nodes: masses

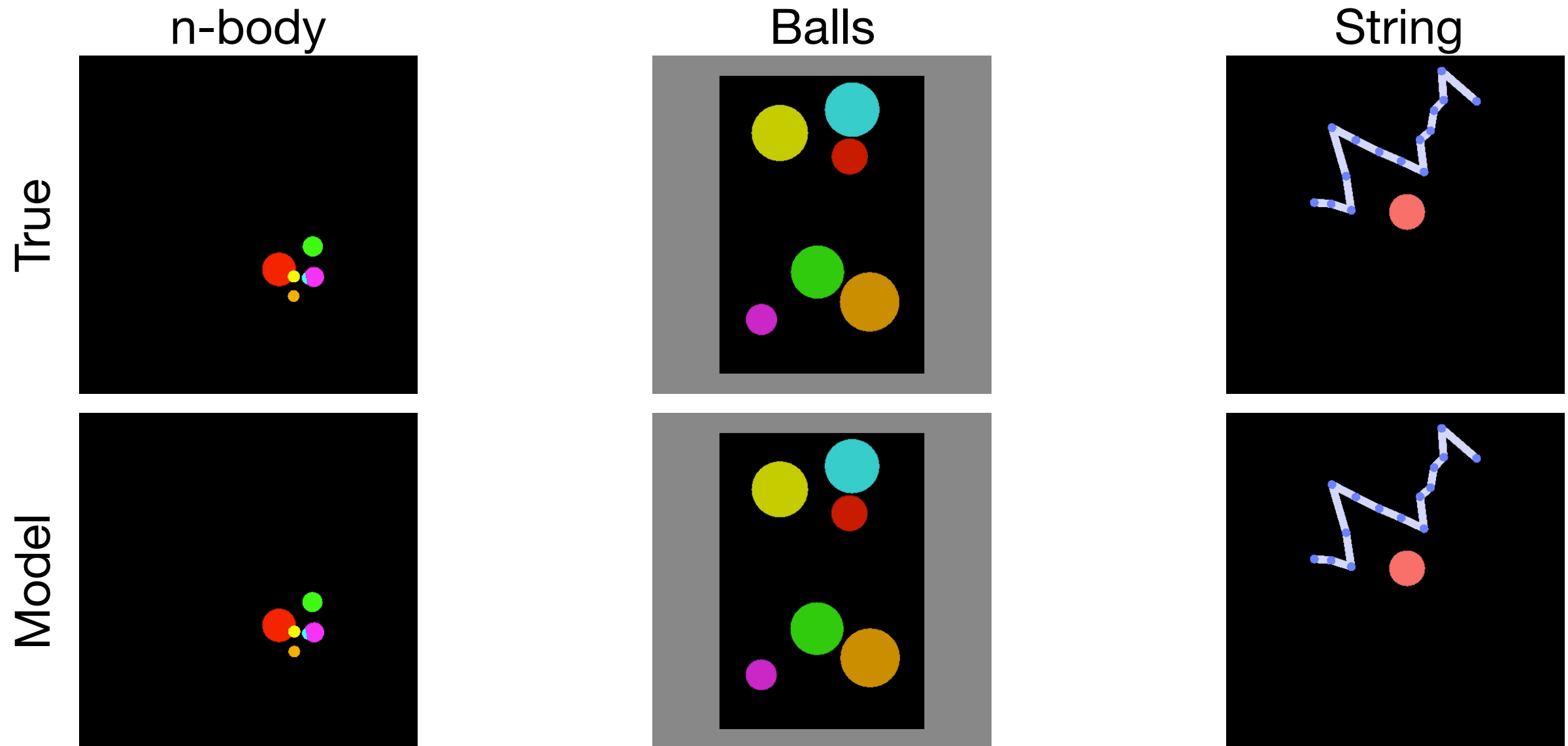
Edges: springs and rigid collisions

Battaglia et al., 2016, NeurIPS

1000-step rollouts of true (top row) vs predicted (bottom row)

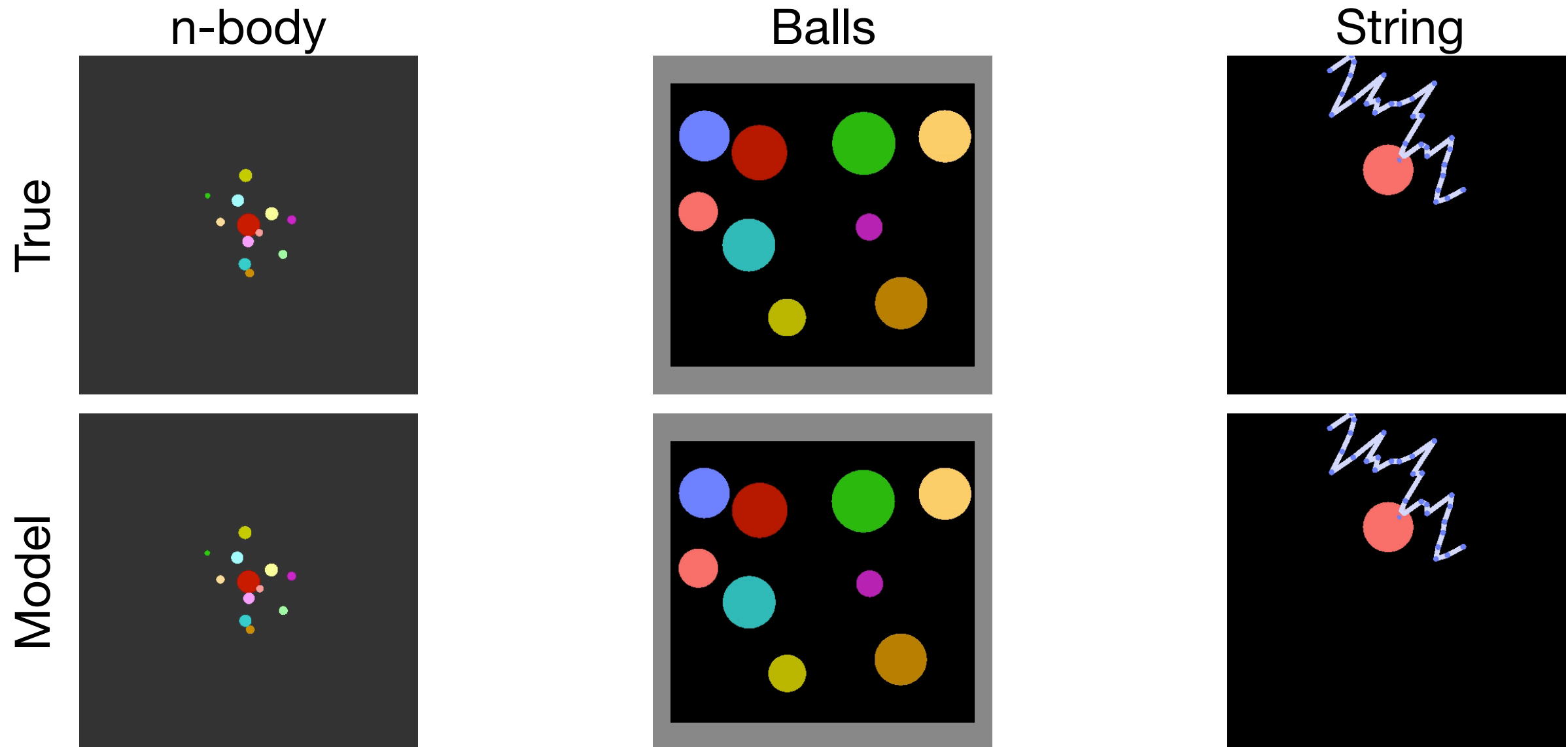


1000-step rollouts of true (top row) vs predicted (bottom row)

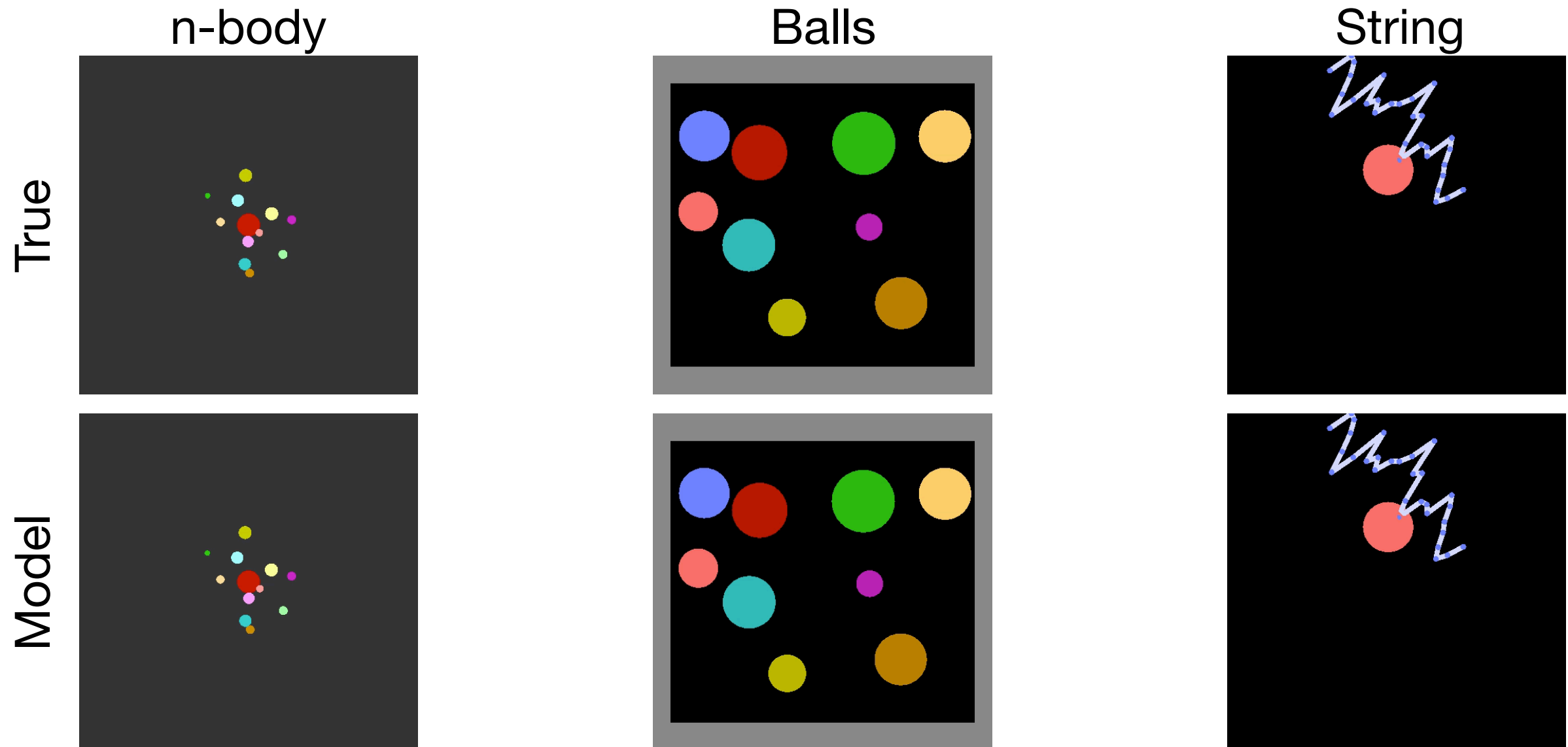




# Zero shot generalisation to larger systems



# Zero shot generalisation to larger systems



# NEW!

We incorporated two physically informed inductive biases

- ODE integrators
- Hamiltonian mechanics

into graph networks for learning simulation, and found they could improve performance, energy, and zero-shot time-step generalization.

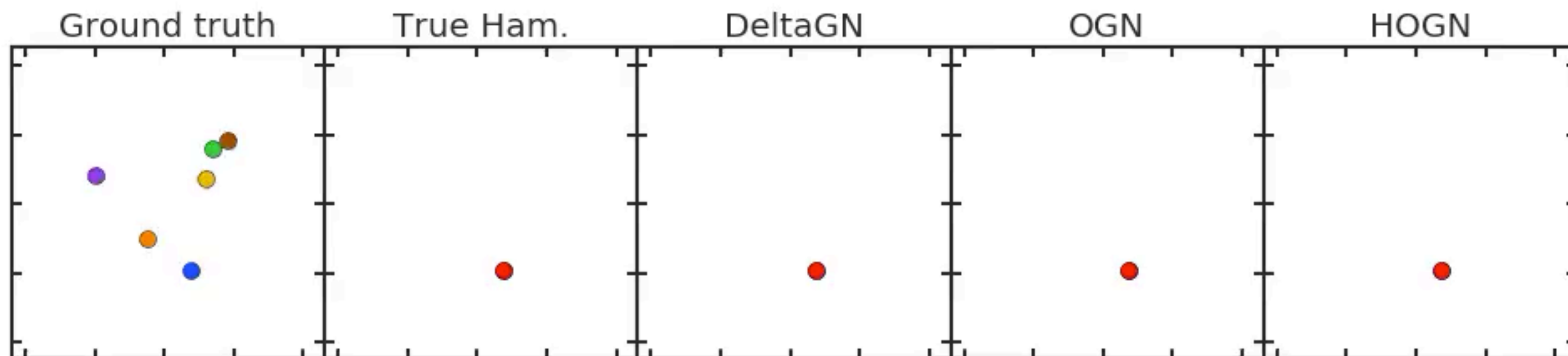
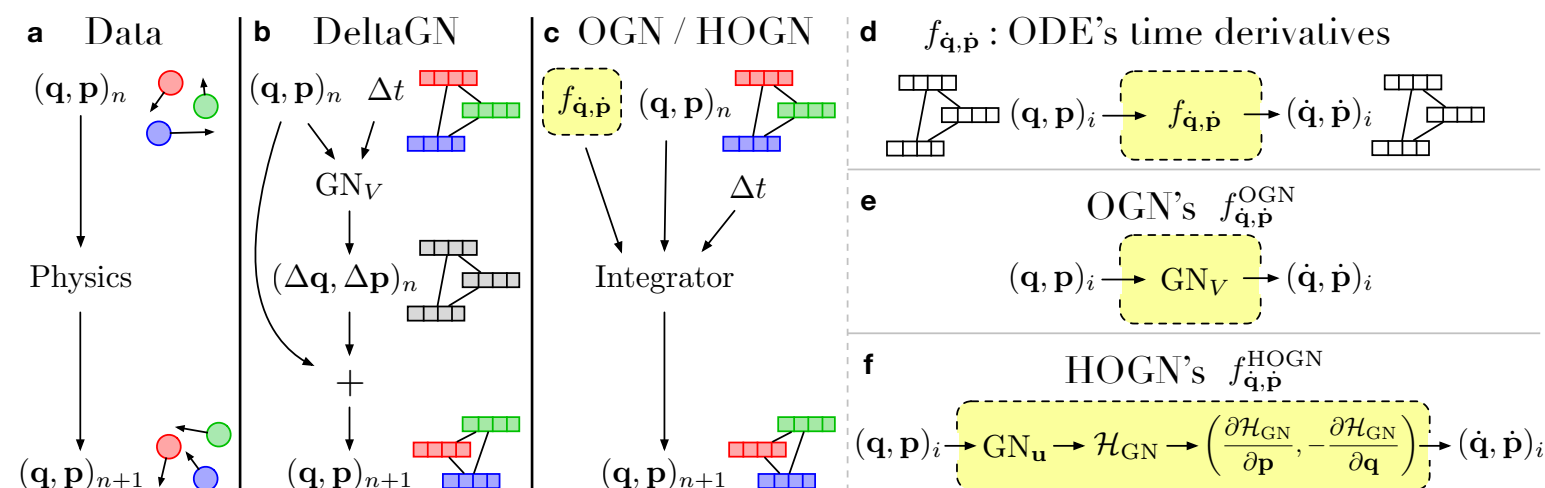
## Hamiltonian Graph Networks with ODE Integrators

**Alvaro Sanchez-Gonzalez**  
DeepMind  
London, UK  
alvarosg@google.com

**Victor Bapst**  
DeepMind  
London, UK  
vbapst@google.com

**Kyle Cranmer**  
NYU  
New York, USA  
kc90@nyu.edu

**Peter Battaglia**  
DeepMind  
London, UK  
peterbattaglia@google.com



# NEW!

We incorporated two physically informed inductive biases

- ODE integrators
- Hamiltonian mechanics

into graph networks for learning simulation, and found they could improve performance, energy, and zero-shot time-step generalization.

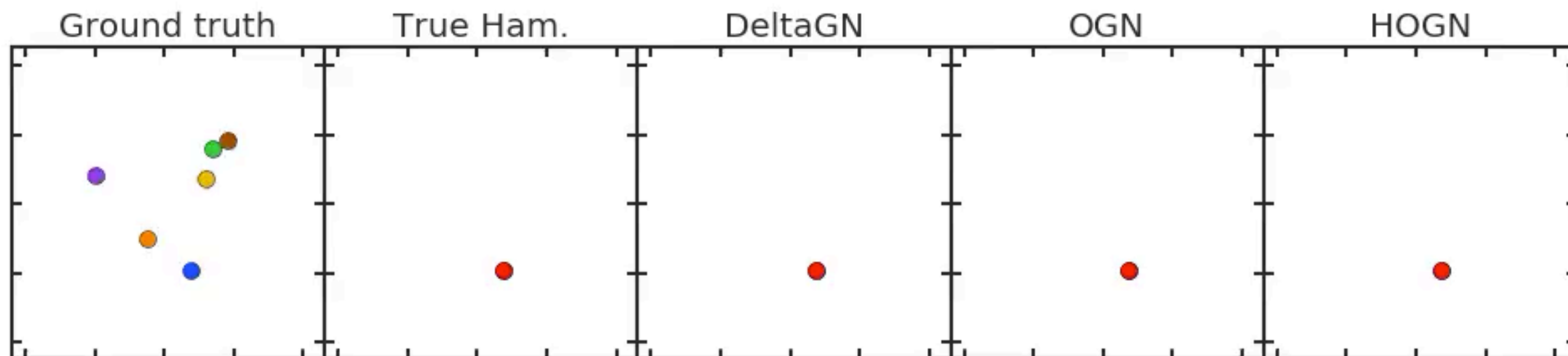
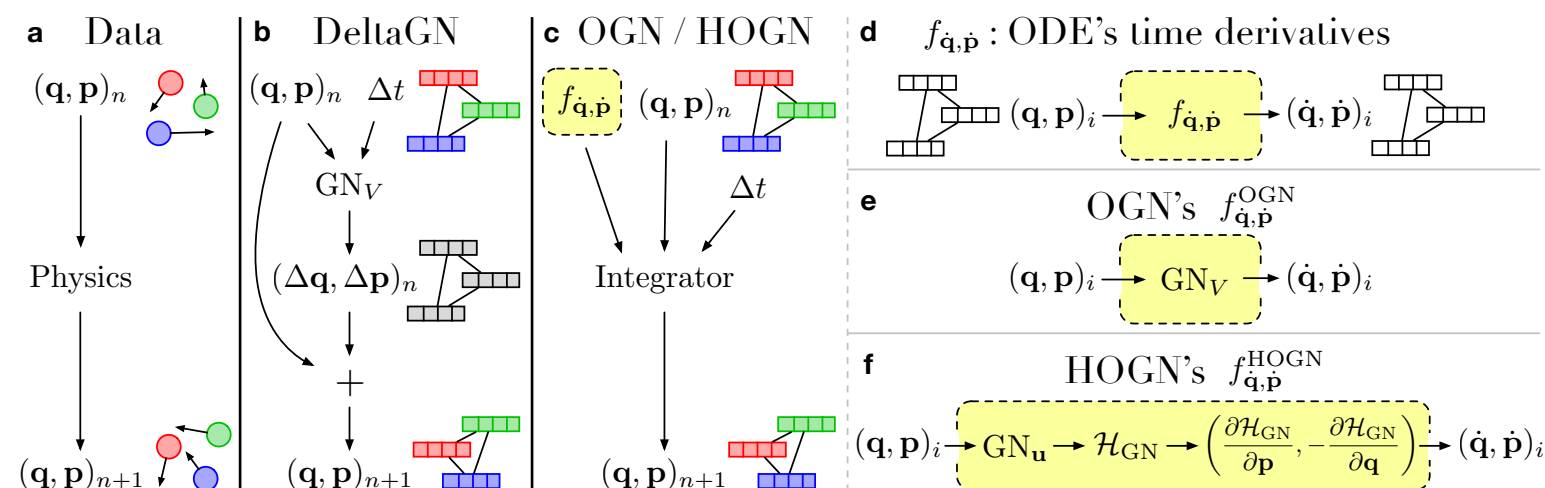
## Hamiltonian Graph Networks with ODE Integrators

**Alvaro Sanchez-Gonzalez**  
DeepMind  
London, UK  
alvarosg@google.com

**Victor Bapst**  
DeepMind  
London, UK  
vbapst@google.com

**Kyle Cranmer**  
NYU  
New York, USA  
kc90@nyu.edu

**Peter Battaglia**  
DeepMind  
London, UK  
peterbattaglia@google.com





JETS

Run: 329716

Event: 857582452

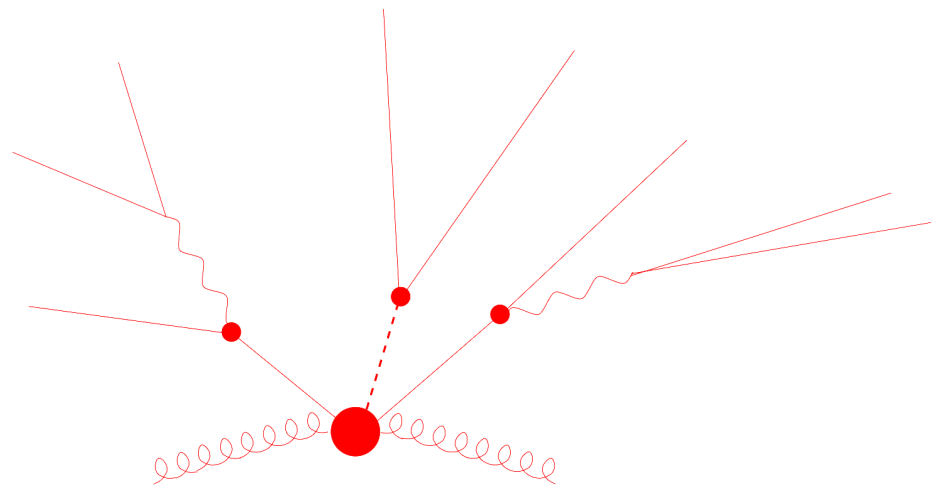
2017-07-14 10:48:51 CEST



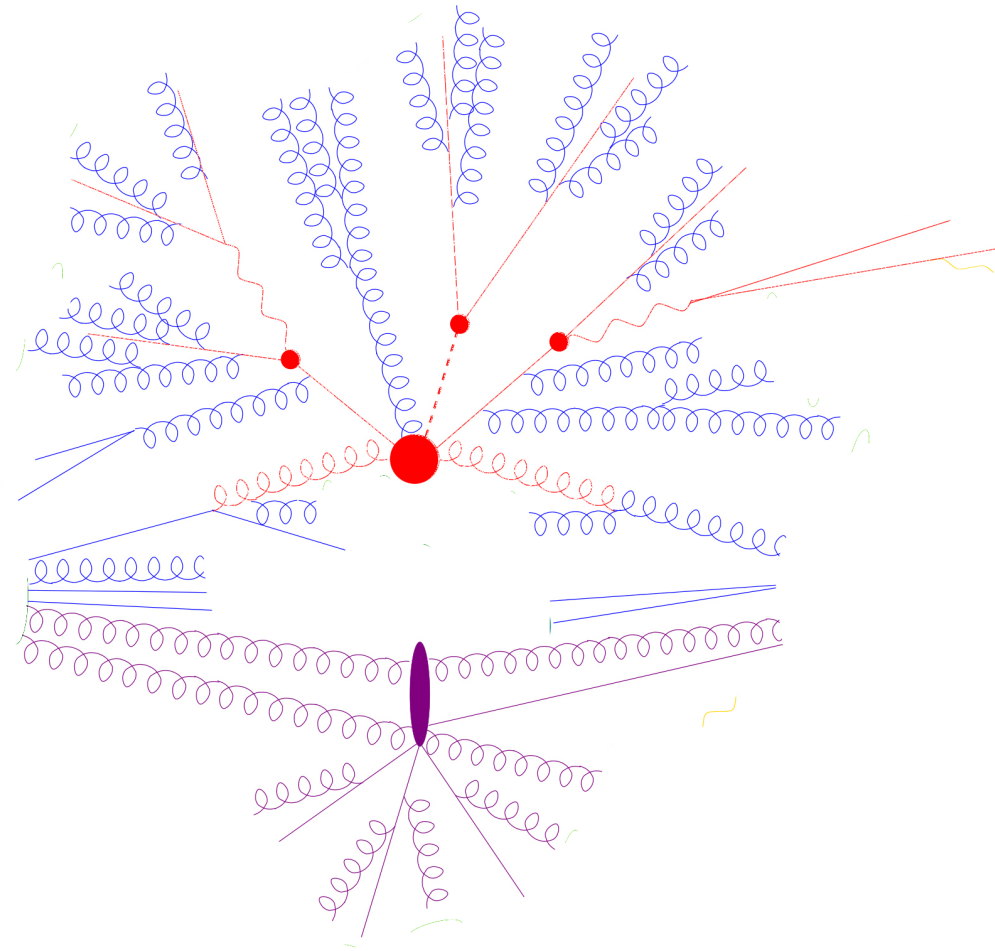
**ATLAS**  
EXPERIMENT

# CAUSAL, GENERATIVE MODEL FOR JETS

# CAUSAL, GENERATIVE MODEL FOR JETS

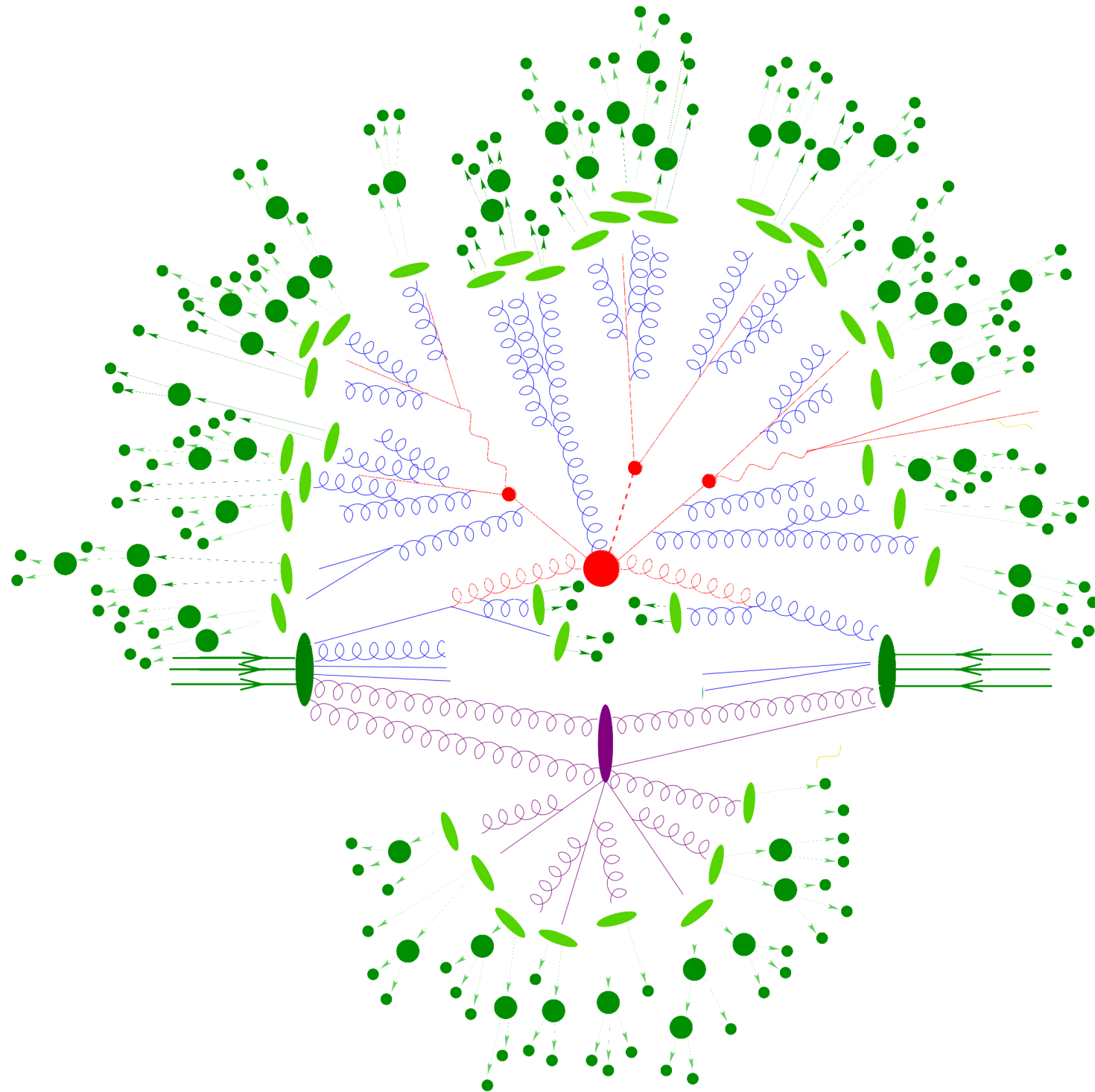


# CAUSAL, GENERATIVE MODEL FOR JETS



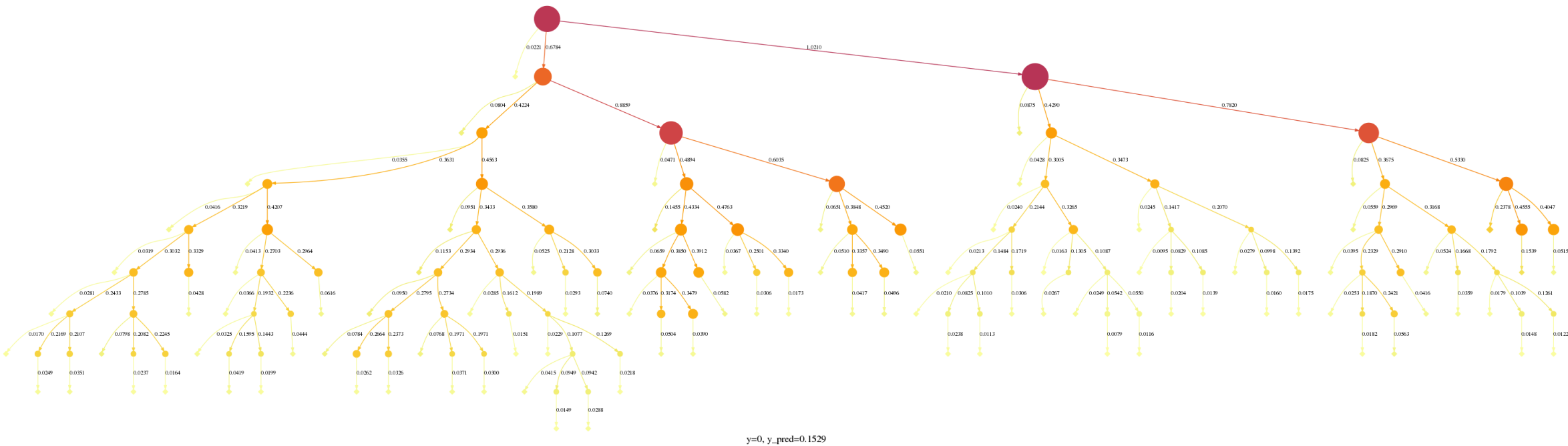


# CAUSAL, GENERATIVE MODEL FOR JETS

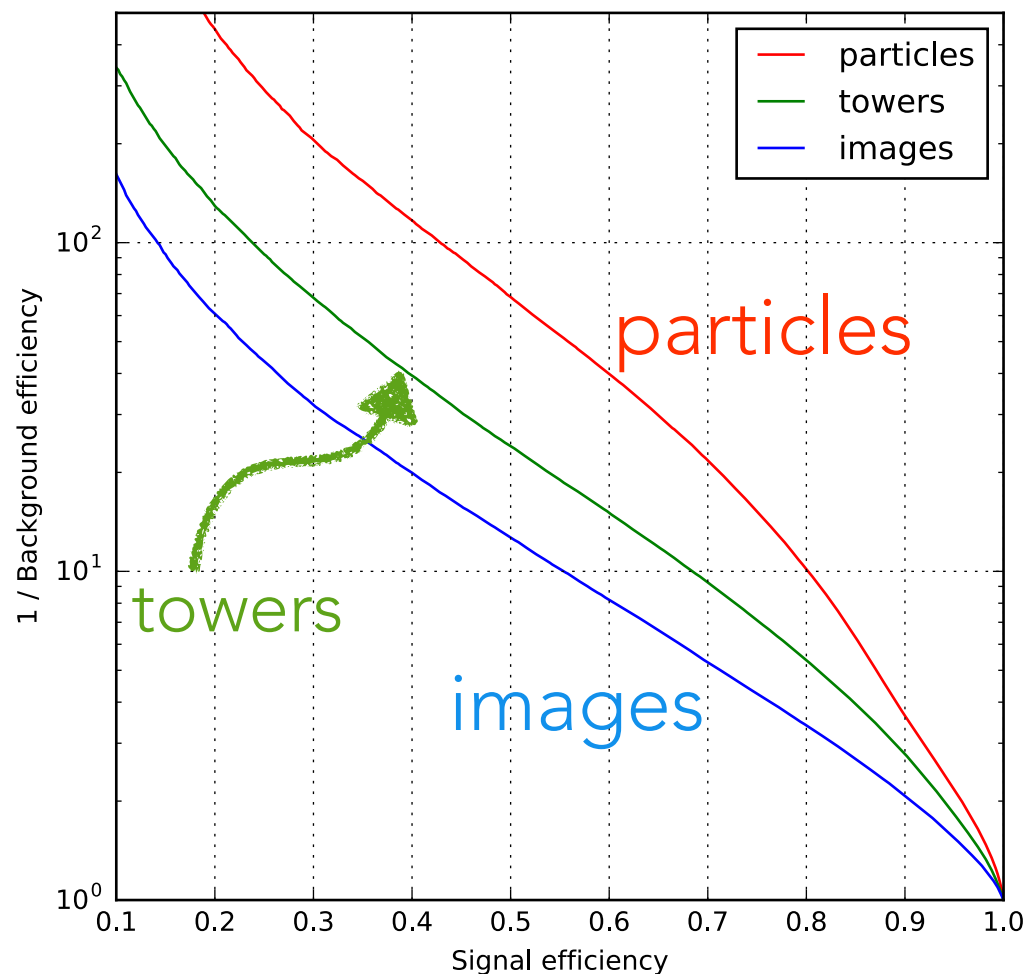


# QCD-INSPIRED RECURSIVE NEURAL NETWORKS

Insight of data generating process informs inductive bias on architecture



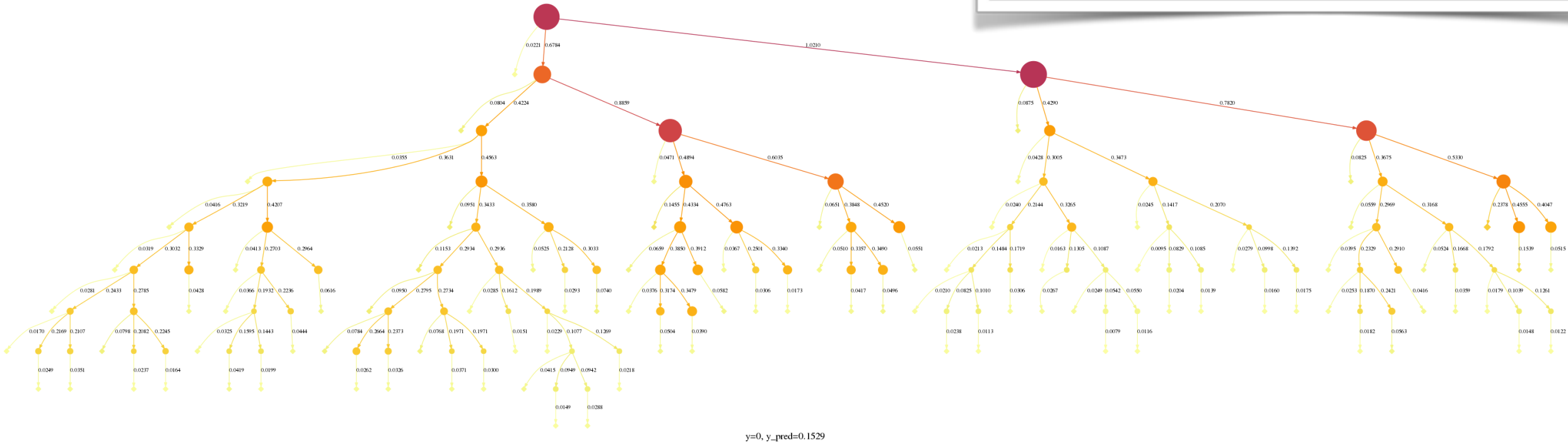
y=0, y\_pred=0.1529



- Generative process is a tree-like, ~stationary Markov Process
- Physics algorithms exist to estimate the tree
- Tree-RNN needs much less data to train!

The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)<sup>1</sup>, T. Plehn (ed)<sup>2</sup>, A. Butter<sup>2</sup>, K. Cranmer<sup>3</sup>, D. Debnath<sup>4</sup>,  
M. Fairbairn<sup>5</sup>, W. Fedorko<sup>6</sup>, C. Gay<sup>6</sup>, L. Gouskos<sup>7</sup>, P. T. Komiske<sup>8</sup>, S. Leiss<sup>1</sup>, A. Lister<sup>6</sup>,  
S. Macaluso<sup>3,4</sup>, E. M. Metodiev<sup>8</sup>, L. Moore<sup>9</sup>, B. Nachman,<sup>10,11</sup> K. Nordström<sup>12,13</sup>,  
J. Pearkes<sup>6</sup>, H. Qu<sup>7</sup>, Y. Rath<sup>14</sup>, M. Rieger<sup>14</sup>, D. Shih<sup>4</sup>, J. M. Thompson<sup>2</sup>, and S. Varma<sup>5</sup>

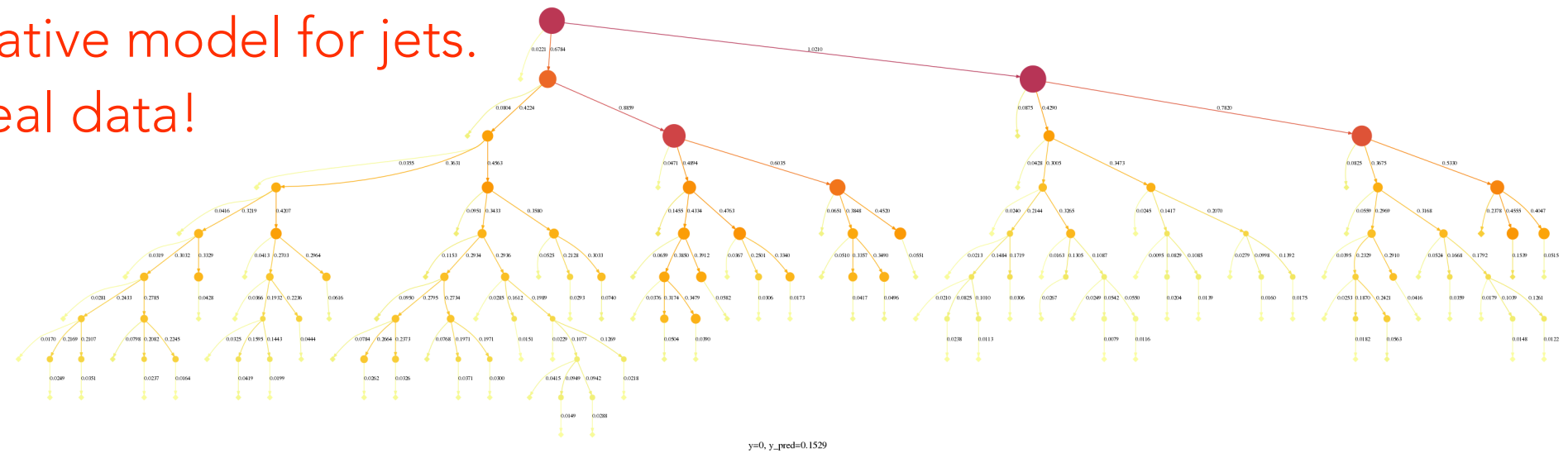


|  | AUC   | Acc   | 1/ε <sub>B</sub> (ε <sub>S</sub> = 0.3) |         |          | #Param |
|--|-------|-------|---|---------|----------|--------|
|  |       |       | single                                  | mean    | median   |        |
| CNN [16]                                 | 0.981 | 0.930 | 914±14                                  | 995±15  | 966±18   | 610k   |
| ResNeXt [30]                             | 0.984 | 0.936 | 1122±47                                 | 1246±28 | 1286±31  | 1.46M  |
| TopoDNN [18]                             | 0.972 | 0.916 | 295±5                                   | 378± 5  | 391 ± 8  | 59k    |
| Multi-body <i>N</i> -subjettiness 6 [24] | 0.979 | 0.922 | 792±18                                  | 802±12  | 783±13   | 57k    |
| Multi-body <i>N</i> -subjettiness 8 [24] | 0.981 | 0.929 | 867±15                                  | 926±20  | 886±18   | 58k    |
| TreeNiN [43]                             | 0.982 | 0.933 | 1025±11                                 | 1209±23 | 1167±24  | 34k    |
| P-CNN                                    | 0.980 | 0.930 | 732±24                                  | 838±13  | 841±14   | 348k   |
| ParticleNet [47]                         | 0.985 | 0.938 | 1298±46                                 | 1383±45 | 1374±41  | 498k   |
| LBN [19]                                 | 0.981 | 0.931 | 836±17                                  | 852±67  | 971±20   | 705k   |
| LoLa [22]                                | 0.980 | 0.929 | 722±17                                  | 768±11  | 751±11   | 127k   |
| Energy Flow Polynomials [21]             | 0.980 | 0.932 | 384                                     |         |          | 1k     |
| Energy Flow Network [23]                 | 0.979 | 0.927 | 633±31                                  | 734±13  | 729±11   | 82k    |
| Particle Flow Network [23]               | 0.982 | 0.932 | 891±18                                  | 1005±21 | 1005±29  | 82k    |
| GoaT                                     | 0.985 | 0.939 | 1368±140                                |         | 1549±208 | 35k    |



# CAUSAL, GENERATIVE MODELS FOR JETS

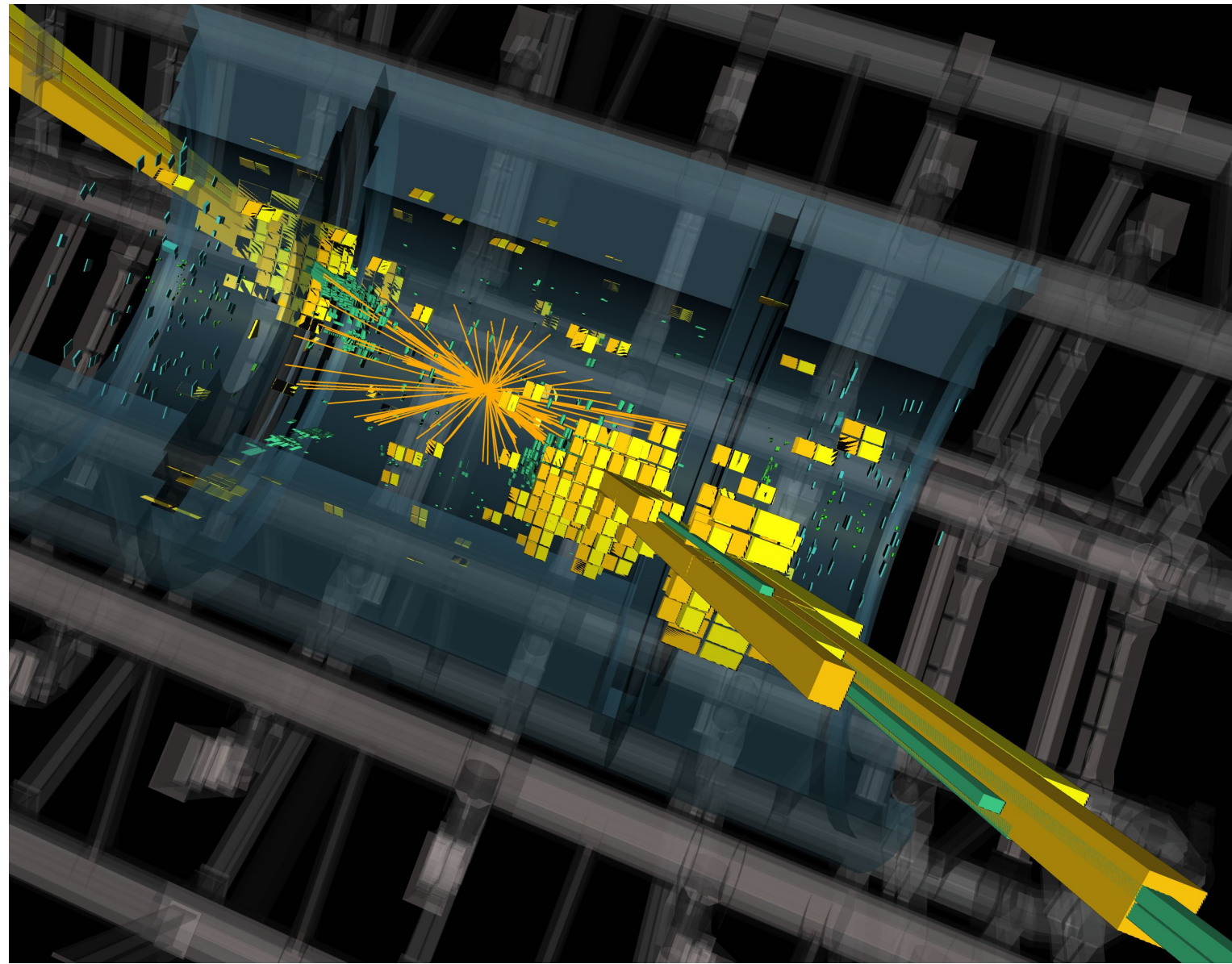
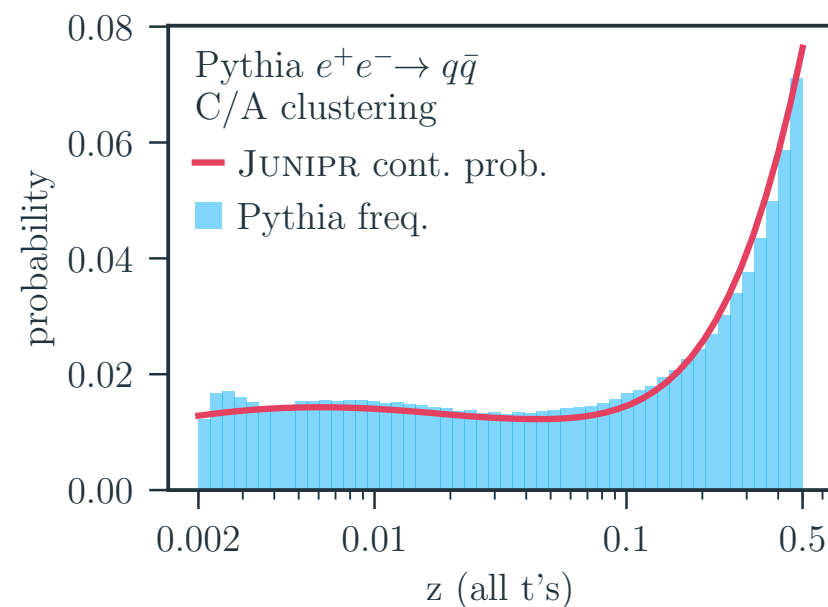
JUNIPR is a causal, generative model for jets.  
Can train on real data!



tractable likelihood

$$P_{\text{jet}}(\{p_1, \dots, p_n\}) = \left[ \prod_{t=1}^{n-1} P_t(k_1^{(t+1)}, \dots, k_{t+1}^{(t+1)} | k_1^{(t)}, \dots, k_t^{(t)}) \right] \times P_n(\text{end} | k_1^{(n)}, \dots, k_n^{(n)}).$$

... and it is interpretable

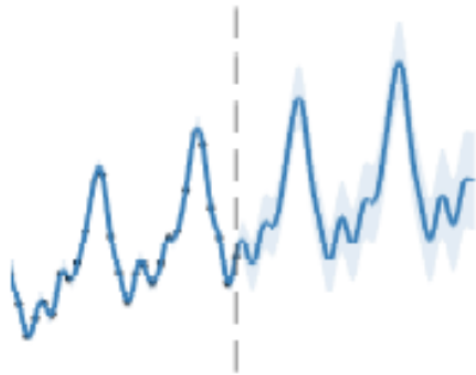




# COMPOSITIONALITY

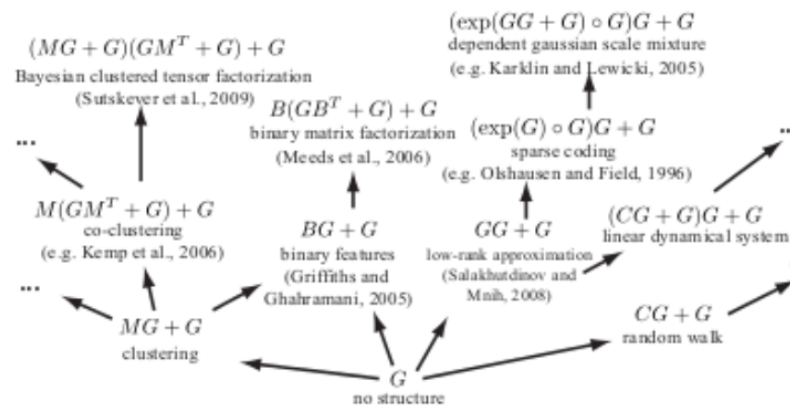
## Vocabulary of kernels + grammar for composition

- physics goes into the construction of a "Kernel" that describes covariance of data



### Structure Discovery in Nonparametric Regression through Compositional Kernel Search

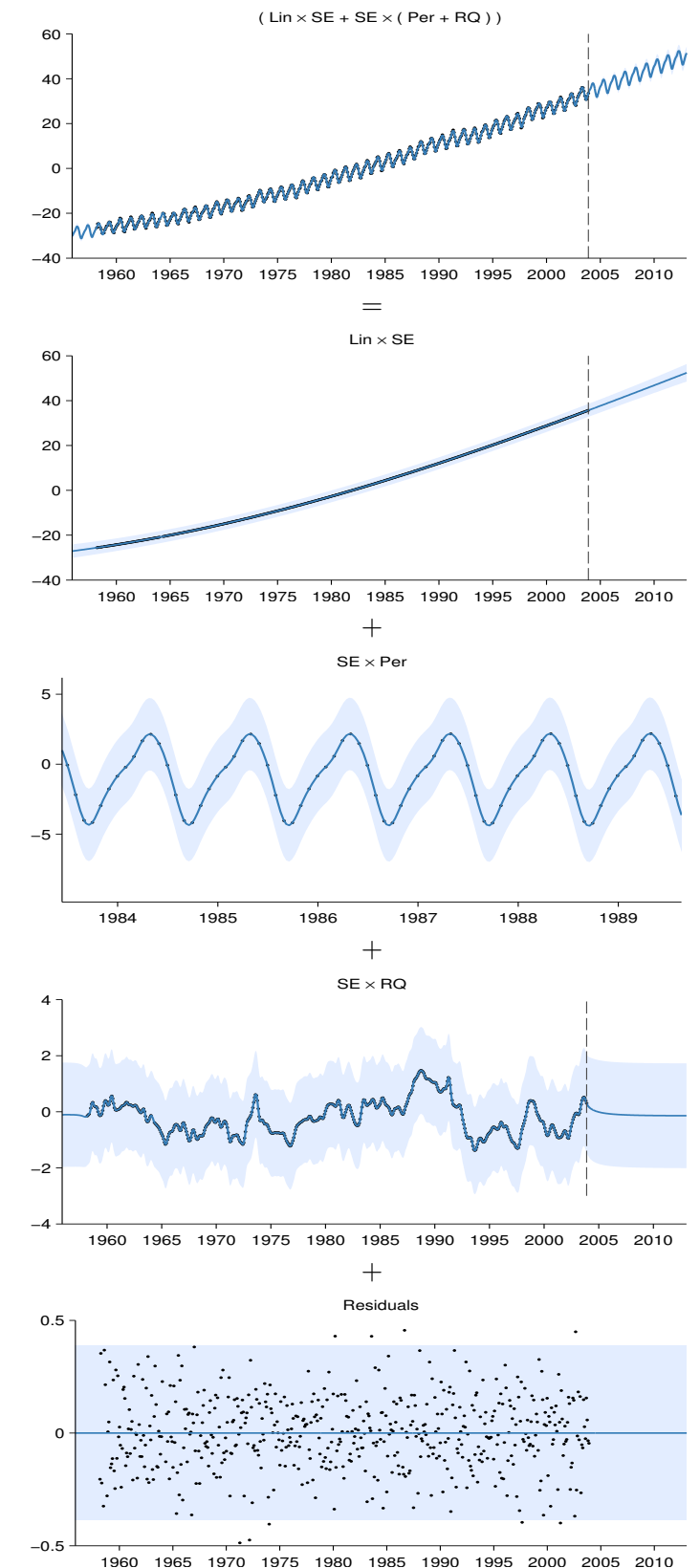
David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, Zoubin Ghahramani  
*International Conference on Machine Learning, 2013*  
[pdf](#) | [code](#) | [poster](#) | [bibtex](#)



### Exploiting compositionality to explore a large space of model structures

Roger Grosse, Ruslan Salakhutdinov, William T. Freeman, Joshua B. Tenenbaum  
*Conference on Uncertainty in Artificial Intelligence, 2012*  
[pdf](#) | [code](#) | [bibtex](#)

## Mauna Loa atmospheric CO<sub>2</sub>



# Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019



OVERVIEW



PARTICIPANT LIST



ACTIVITIES



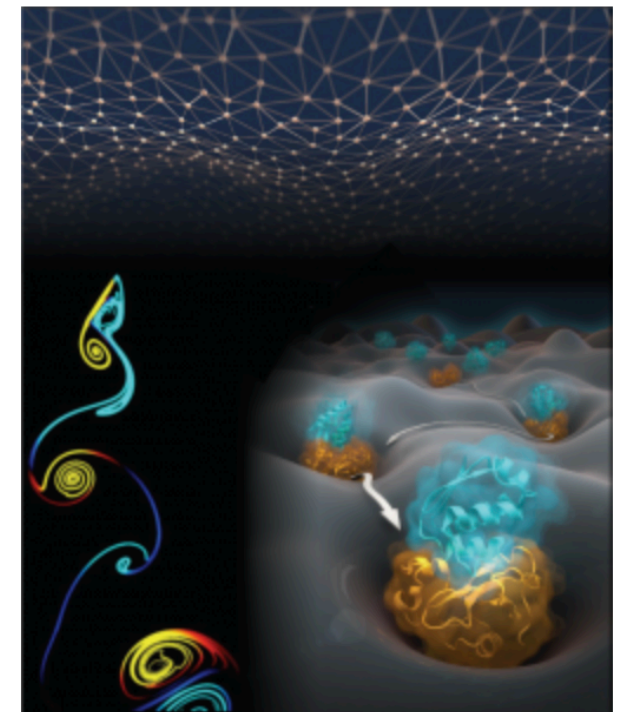
APPLICATION

## Overview

Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and applying the methods of machine learning to interrogate high-dimensional complex data in a way that has not been possible before.

As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity to learn the models themselves—that is, to learn the physical principles and structures underlying the data—and that with more realistic constraints, machine learning will also be able to generate and design complex and novel physical structures and objects. Finally, physicists would not just like to fit their data, but rather obtain models that are physically understandable; e.g., by maintaining relations of the predictions to the microscopic physical quantities used as an input, and by respecting physically meaningful constraints, such as conservation laws or symmetry relations.

The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information theory and statistics are intimately related in their goal to extract valid information from noisy data, and we want to push the cross-pollination further in the specific context of discovering physical principles from data.



# BABY STEPS

Before we are able to discover new models on experimental data, should be able to recover model from simulation

- should be able to recover ground truth with increasingly fewer hints (in less restricted model space)
- Simulators have causal structure, can perform interventions and test different approaches to causal discovery

The ability to systematically improve on an existing simulator model with real data may be easier than discovering new model from scratch, and may be even more valuable in practice

# WHY DO WE CARE ABOUT INTERPRETABILITY?

For a fixed task, one might not care about interpretability as long as the performance on the task is good

- Depending on context, “good” may mean that it generalizes well, is robust to domain shift, performance can be characterized and validated to be within some tolerance, etc...

But for progress in science, we don't just want to solve today's task well.

- For science to progress we need to be able to generate new hypotheses, design experiments, etc.





# Machine Learning and the Physical Sciences

Workshop at the 33rd Conference on Neural Information Processing Systems

(NeurIPS)

December 13 or 14, 2019

[ml4physicalsciences.github.io](https://ml4physicalsciences.github.io)

Backup

# GENERALIZATION

**Teacher → Causal, Generative Model (Simulator)**

Richer set of problems can be investigated.

Insight of data generating process informs inductive bias on architecture



We have **joint likelihood ratio**

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

With  $r(x, z|\theta_0, \theta_1)$ , we define the functional

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

One can show it is minimized by

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

We want **likelihood ratio**

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



# LEARNING THE SCORE

arXiv:1805.12244


PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90

Similar to the joint likelihood ratio,  
we can calculate the **joint score**

$$t(x, z|\theta_0) \equiv \nabla_{\theta} \log p(x, z_d, z_s, z|\theta) \Big|_{\theta_0}$$


We want **score**

$$t(x|\theta_0) \equiv \nabla_{\theta} \log p(x|\theta) \Big|_{\theta_0}$$

Similar to the joint likelihood ratio,  
we can calculate the **joint score**

$$t(x, z|\theta_0) \equiv \nabla_{\theta} \log p(x, z_d, z_s, z|\theta) \Big|_{\theta_0}$$

Given  $t(x, z|\theta_0)$ ,  
we define the functional

$$L_t[\hat{t}(x|\theta_0)] = \int dx \int dz p(x, z|\theta_0) \left[ \left( \hat{t}(x|\theta_0) - t(x, z|\theta_0) \right)^2 \right]$$

One can show it is minimized by

$$t(x|\theta_0) = \arg \min_{\hat{t}(x|\theta_0)} L_t[\hat{t}(x|\theta_0)]$$

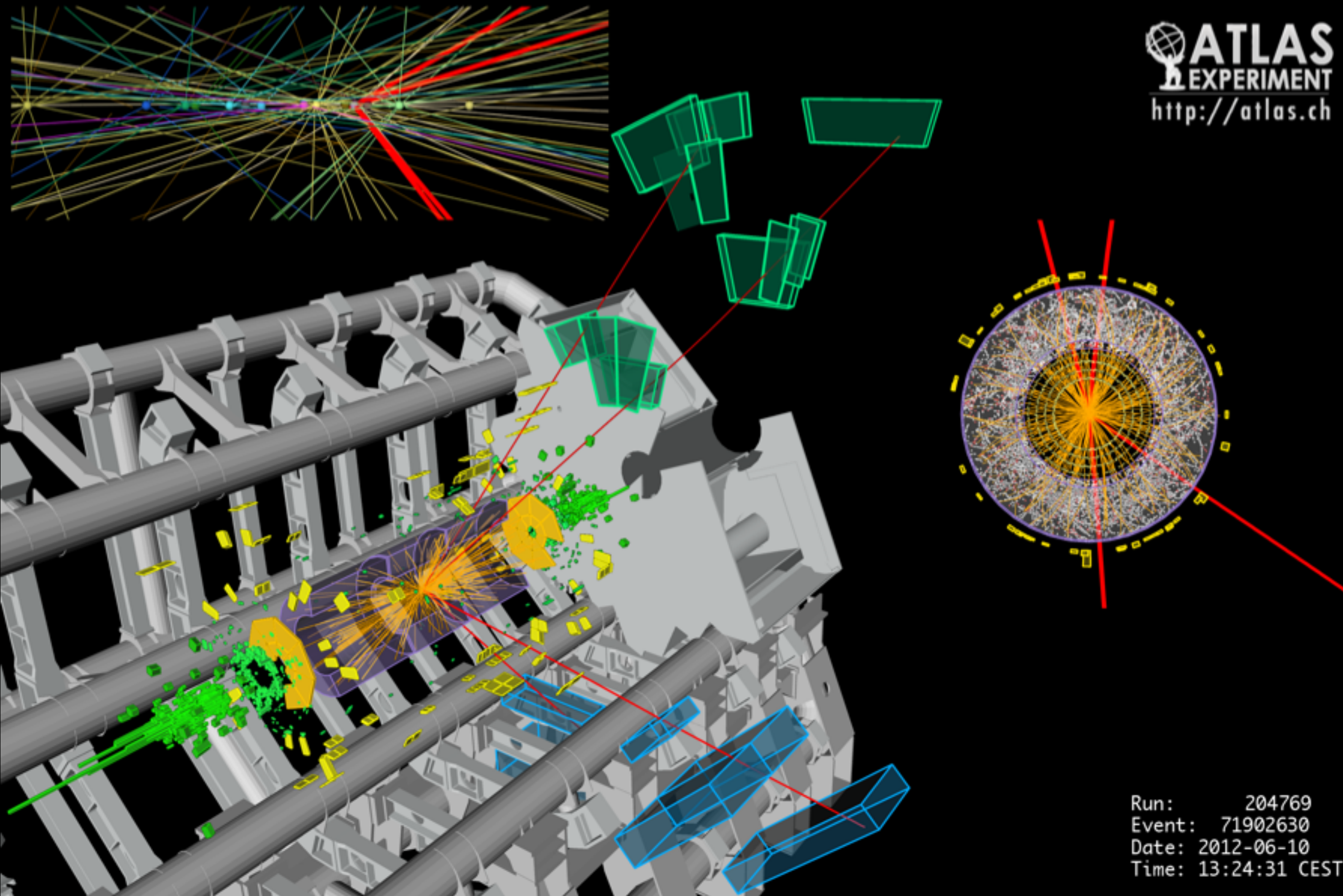
We want **score**

$$t(x|\theta_0) \equiv \nabla_{\theta} \log p(x|\theta) \Big|_{\theta_0}$$

Again, we implement this minimization  
through machine learning

# THE HIGGS BOSON

**ATLAS**  
EXPERIMENT  
<http://atlas.ch>



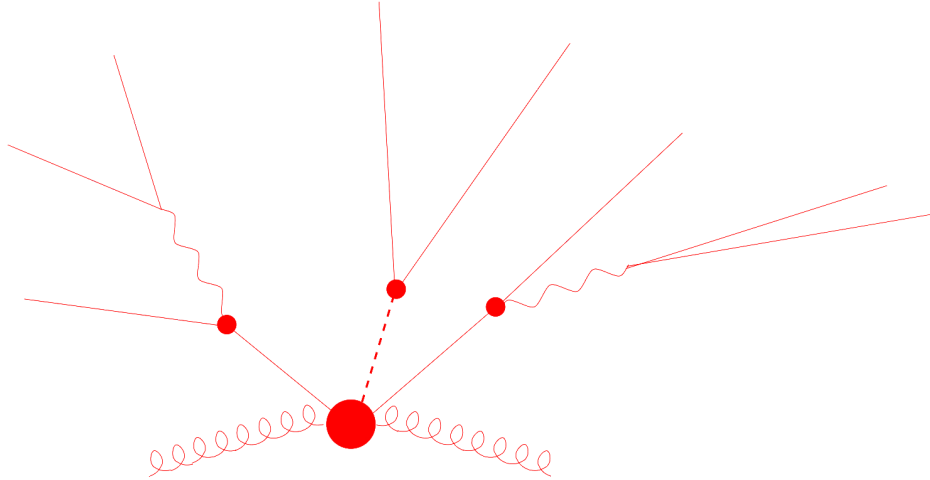
# THE CAUSAL, GENERATIVE MODEL

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$



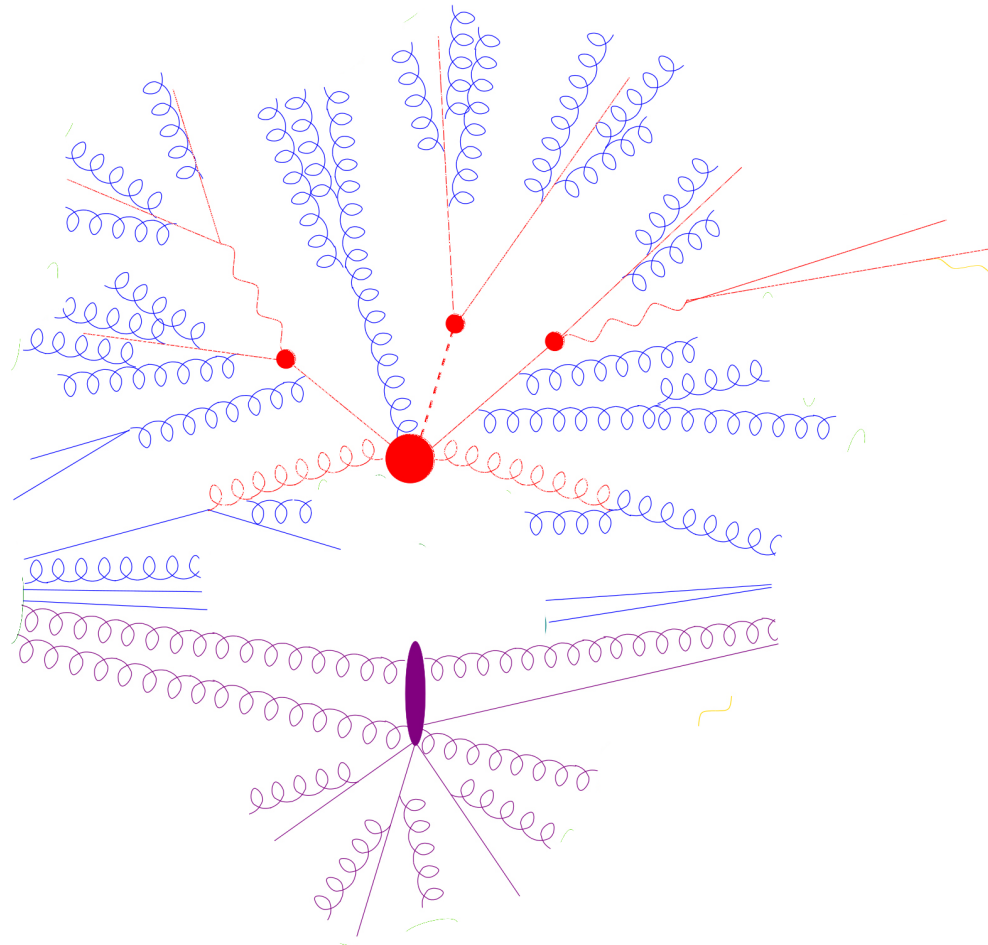
# THE CAUSAL, GENERATIVE MODEL

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

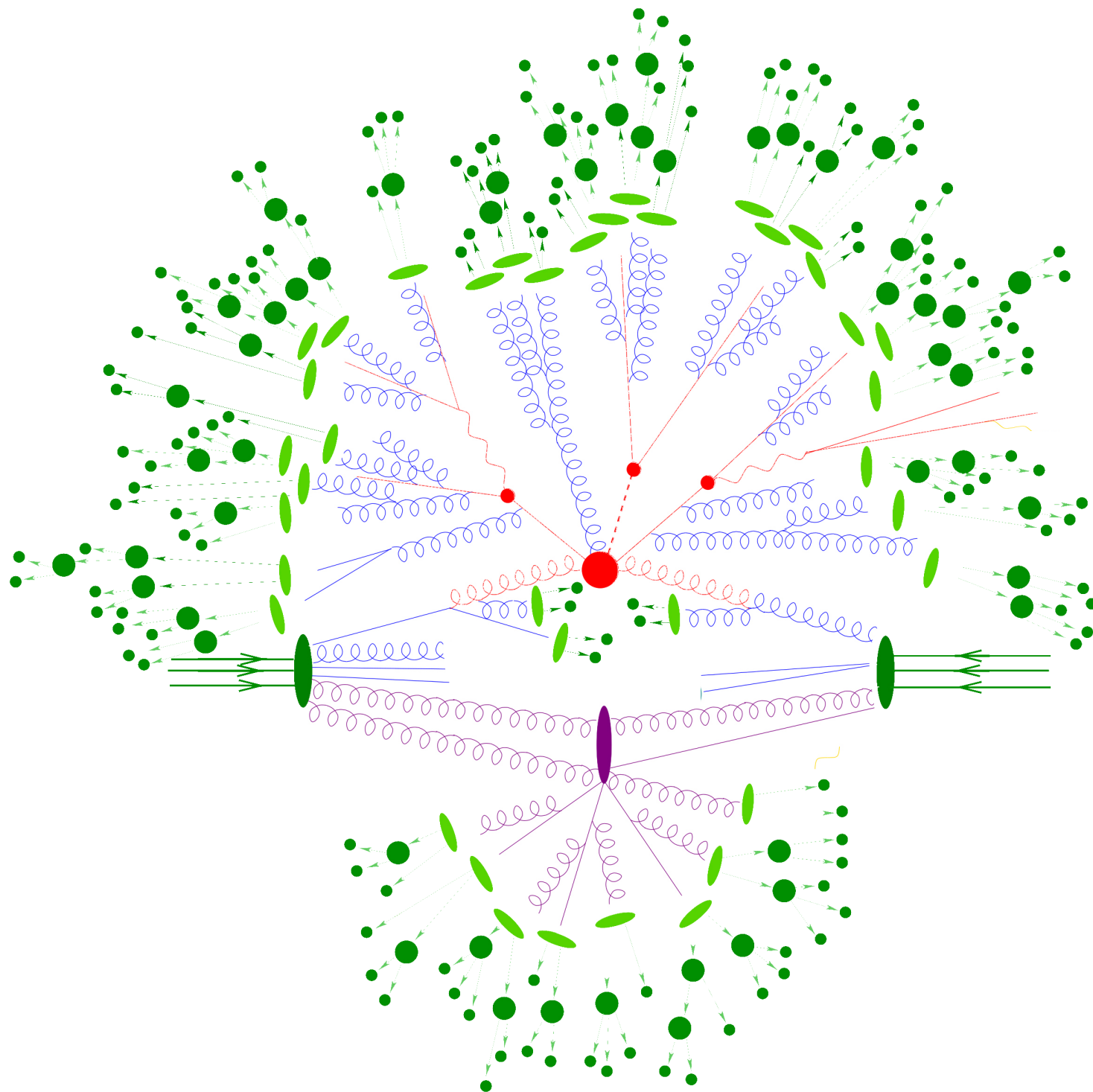


# THE CAUSAL, GENERATIVE MODEL

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$



# THE CAUSAL, GENERATIVE MODEL



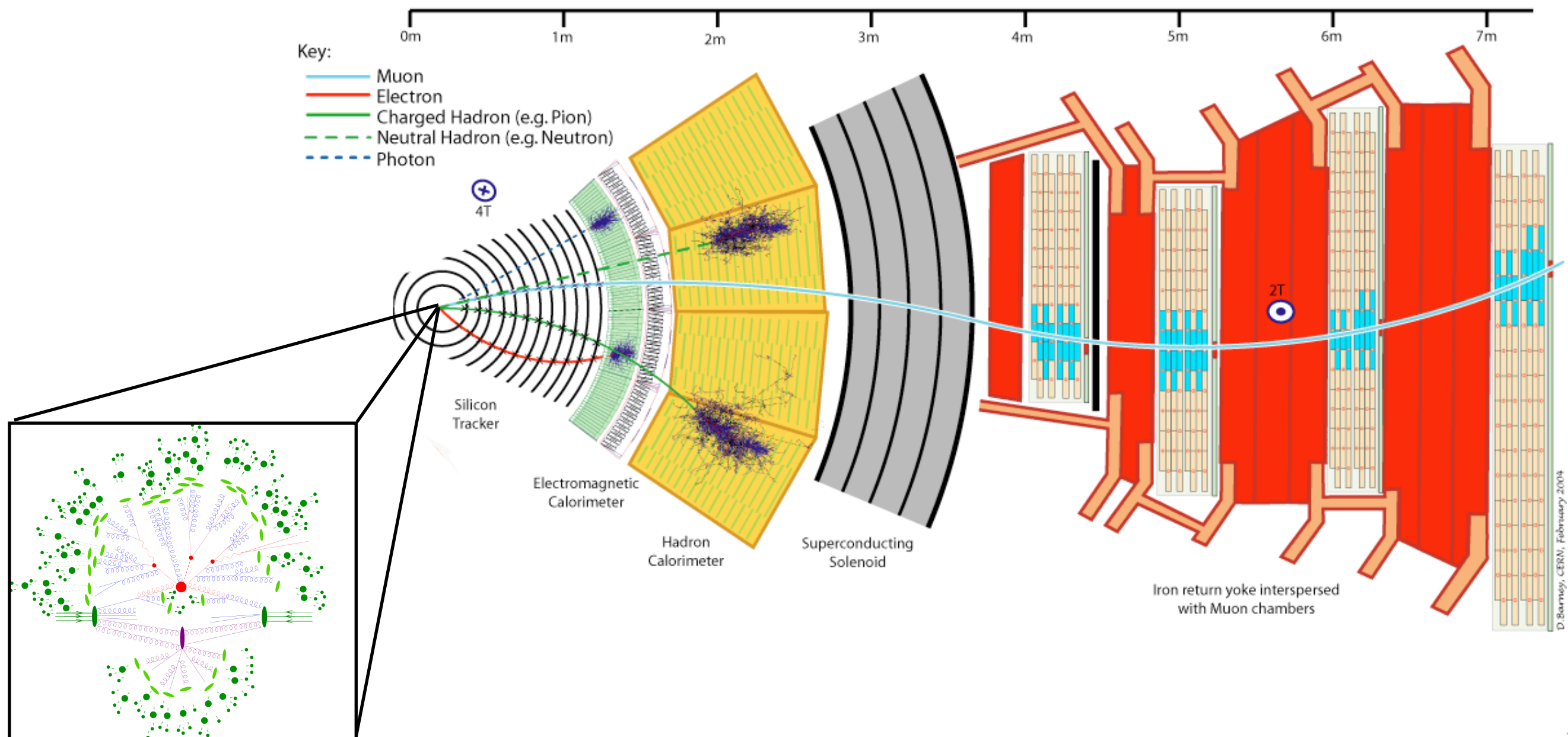
$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} \left| (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi \right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

# THE CAUSAL, GENERATIVE MODEL

**Conceptually:**  $\text{Prob}(\text{detector response} \mid \text{particles})$

**Implementation:** Monte Carlo integration over micro-physics

**Consequence:** evaluation of the likelihood is intractable

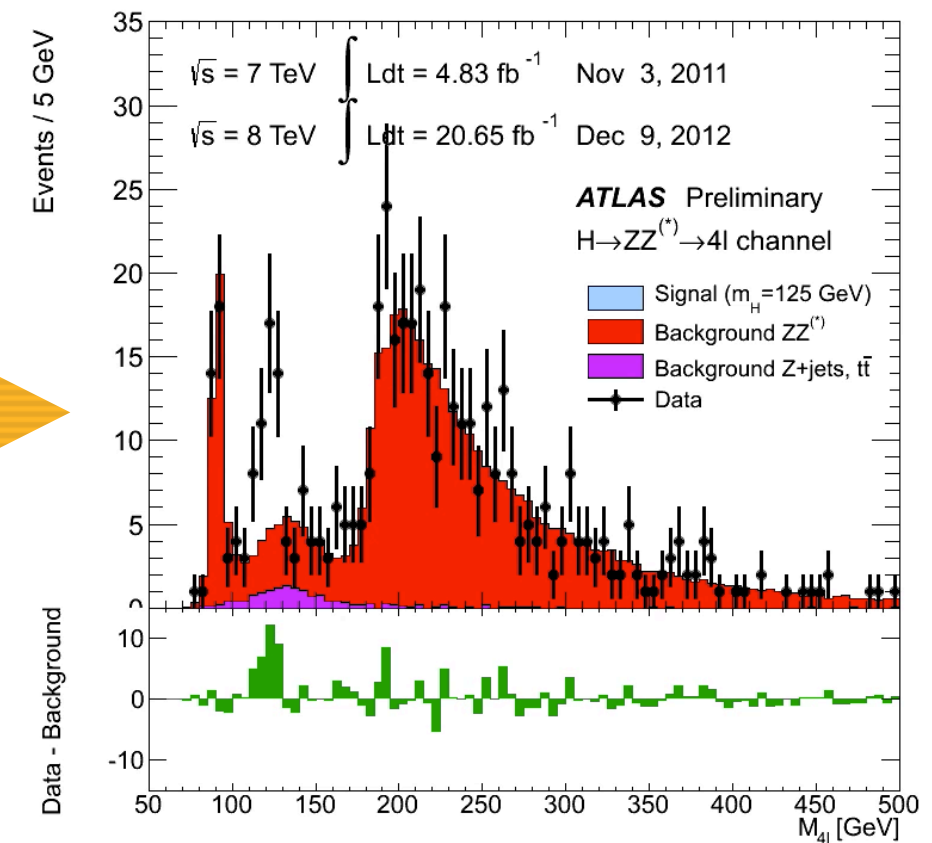
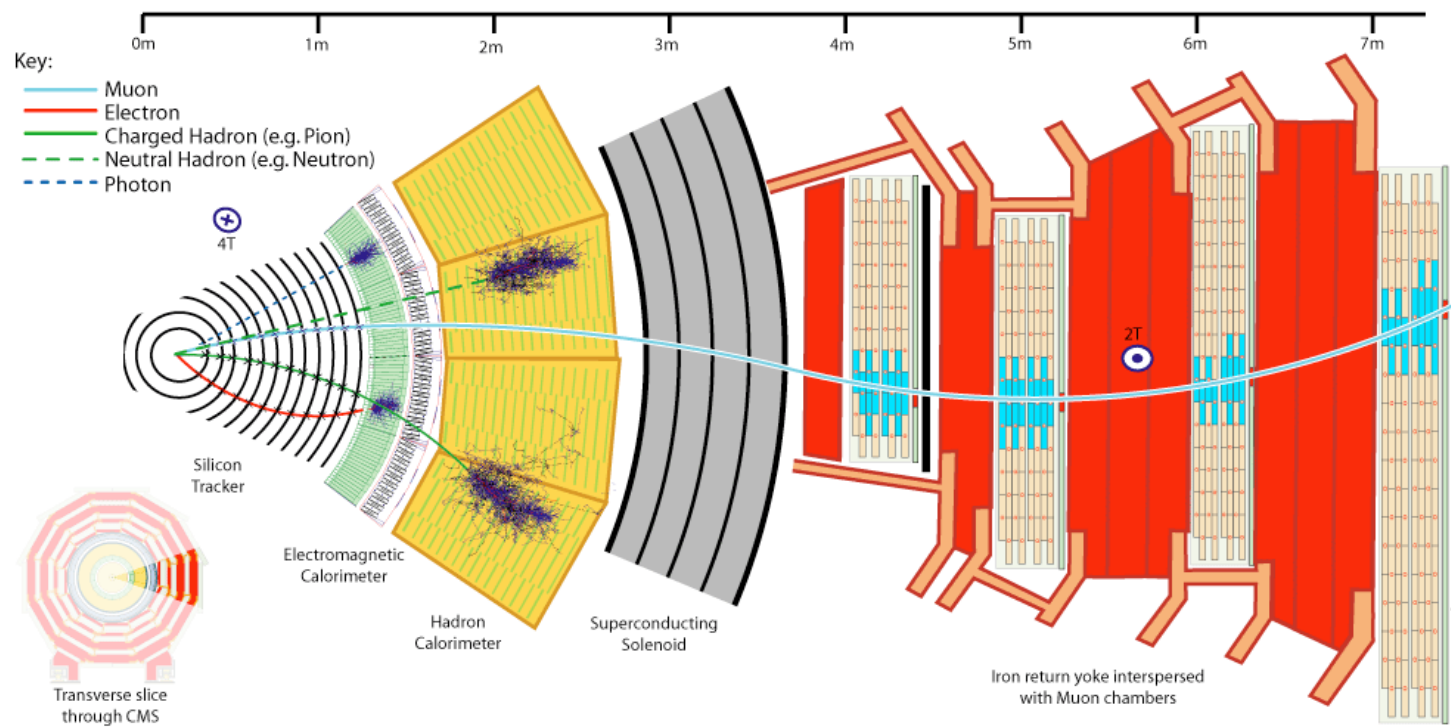




# $10^8$ SENSORS $\rightarrow$ 1 REAL-VALUED QUANTITY

Most measurements and searches for new particles at the LHC are based on the distribution of a single summary statistic

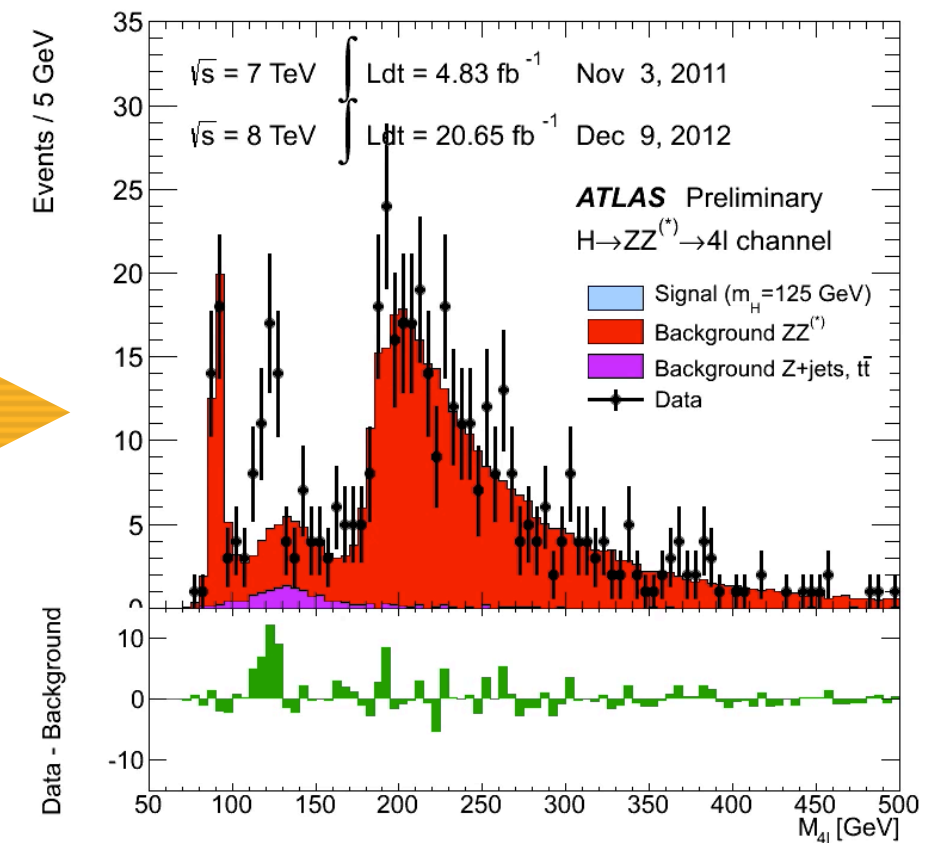
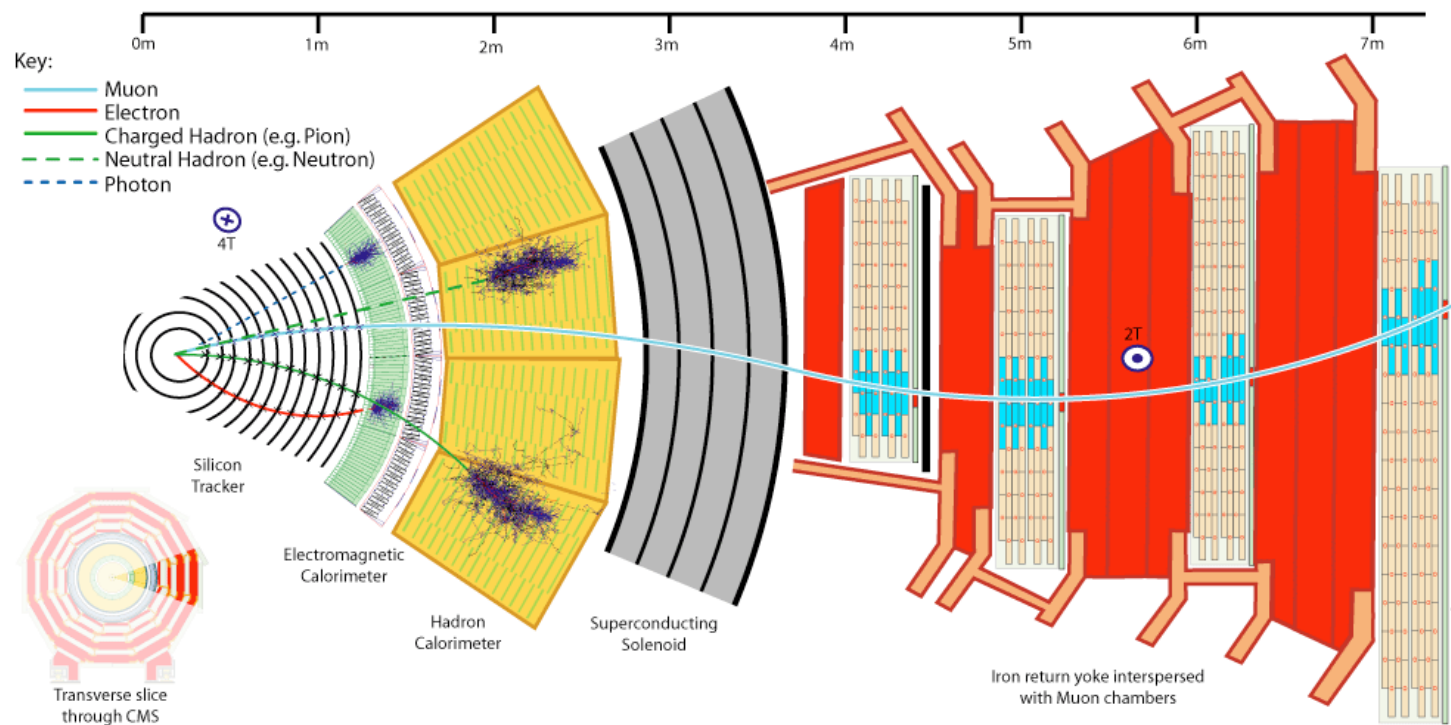
- choosing a good summary statistic (feature engineering) is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood  $p(x|\theta)$  **approximated** using histograms (univariate density estimation)



# $10^8$ SENSORS $\rightarrow$ 1 REAL-VALUED QUANTITY

Most measurements and searches for new particles at the LHC are based on the distribution of a single summary statistic

- choosing a good summary statistic (feature engineering) is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood  $p(x|\theta)$  **approximated** using histograms (univariate density estimation)

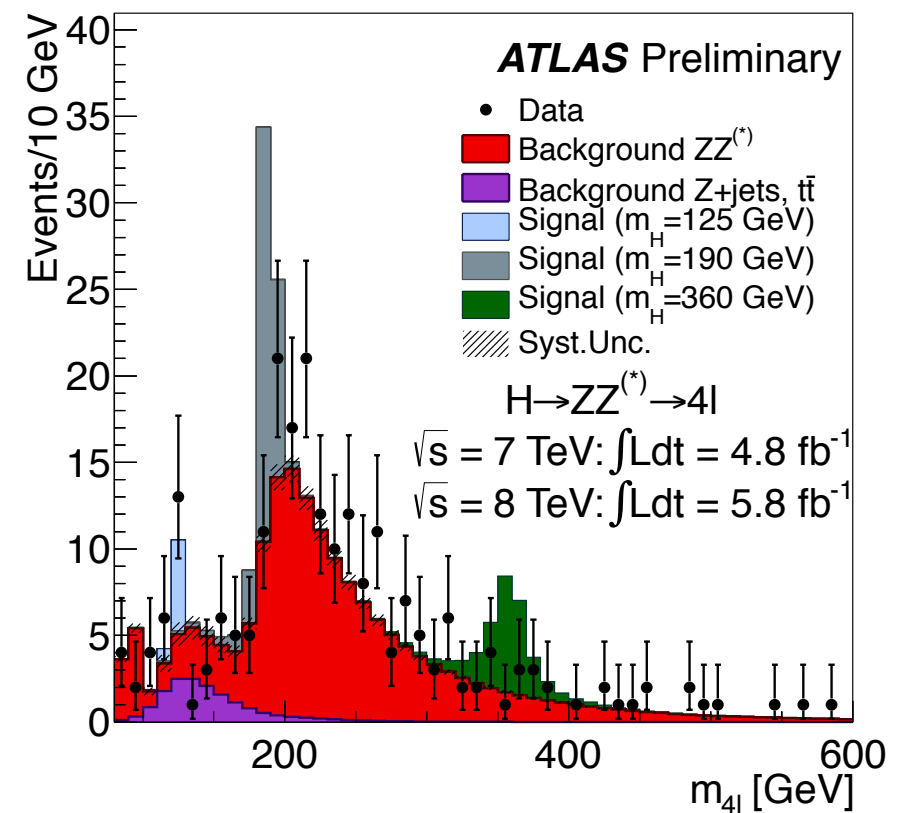


# THE CRUX, AN INTRACTABLE INTEGRAL

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

observed
MC Sampling
simulation

$\hat{p}(x|\theta)$   
↑
 histogram  
 approximation



# THE CRUX, AN INTRACTABLE INTEGRAL

observed

MC Sampling

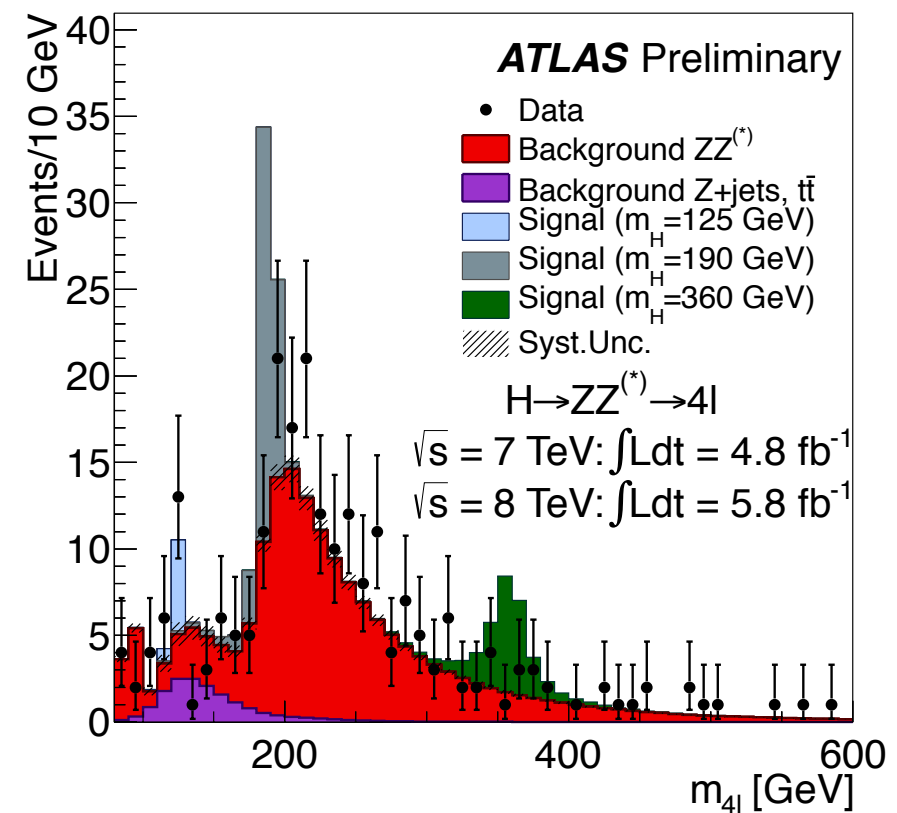
simulation

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

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

↑

histogram  
approximation



**This doesn't scale if  $x$  is high dimensional!**



# HIGH DIMENSIONAL EXAMPLE

When looking for deviations from the standard model  
Higgs, we would like leverage subtle kinematic correlations

- thus each observation  $\mathbf{x}$  is high-dimensional

