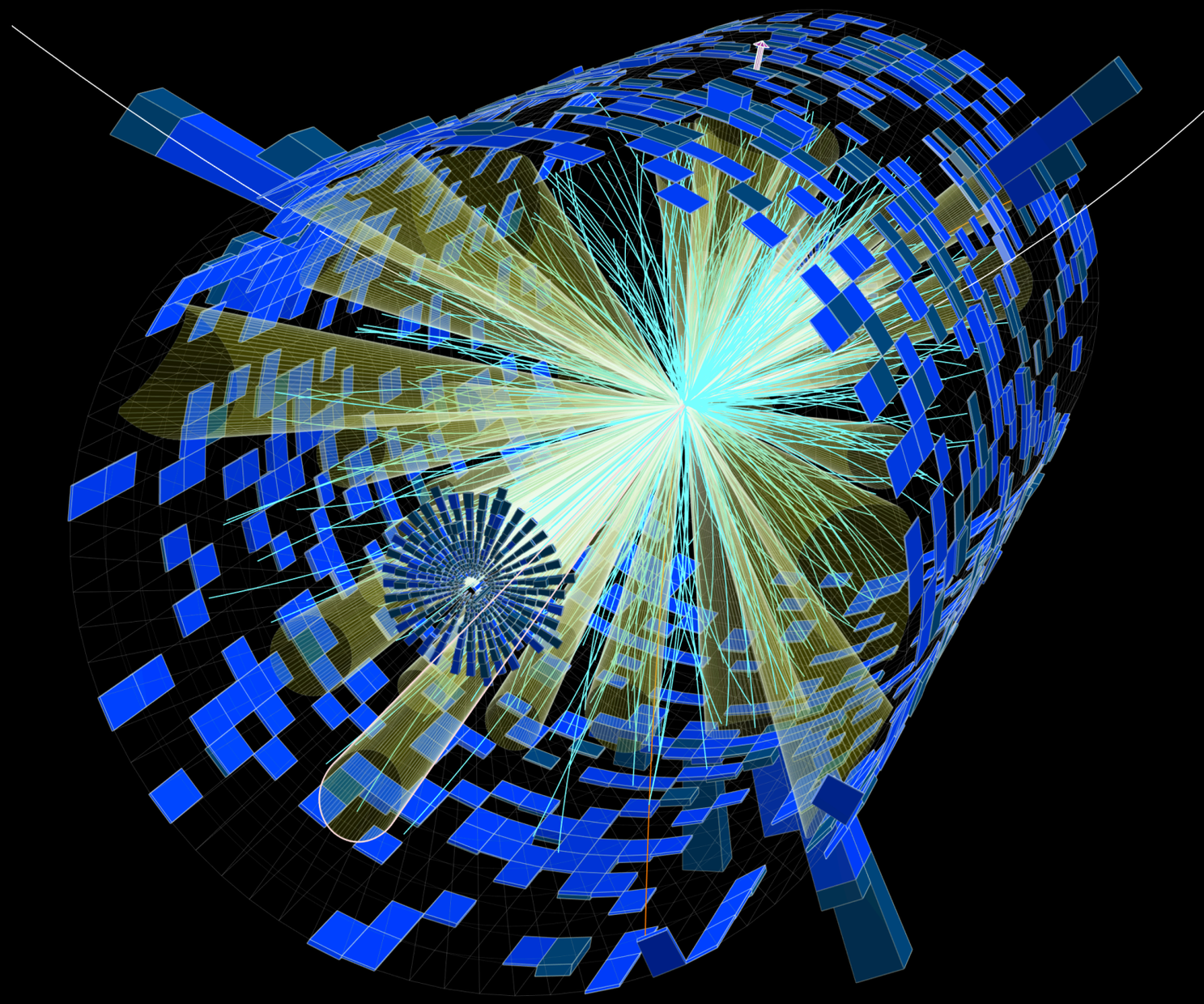
NYU Center for Data Science

Center for Cosmology and Particle Physics

SIMULATION-BASED INFERENCE,

CAUSALITY, AND ACTIVE LEARNING



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

SUPPORT





MSDSE

Moore - Sloan
Data Science Environments

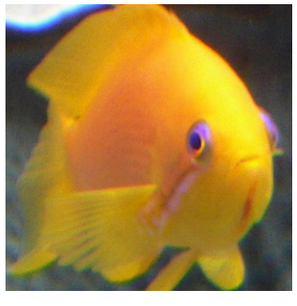




Institute for Research and Innovation
in Software for High Energy Physics

The SCAILFIN Project

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



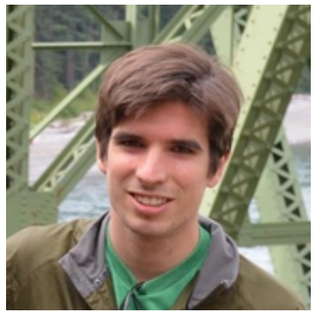
Tim Head



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



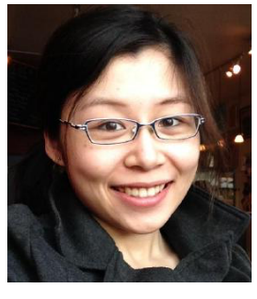
Frank Wood
University of Oxford



Phiala Shanahan



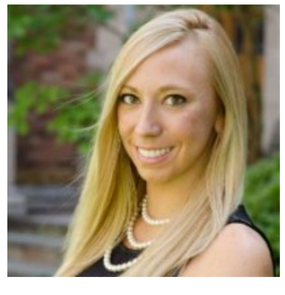
William Detmold



Karen Ng



Tuan Anh Le



Michela Paganini
Yale University



Daniela Huppenkothen
New York University



Savannah Thais
Yale University



Ruth Angus
Columbia University



Machine Learning and the Physical Sciences

Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS)

December 13 or 14, 2019

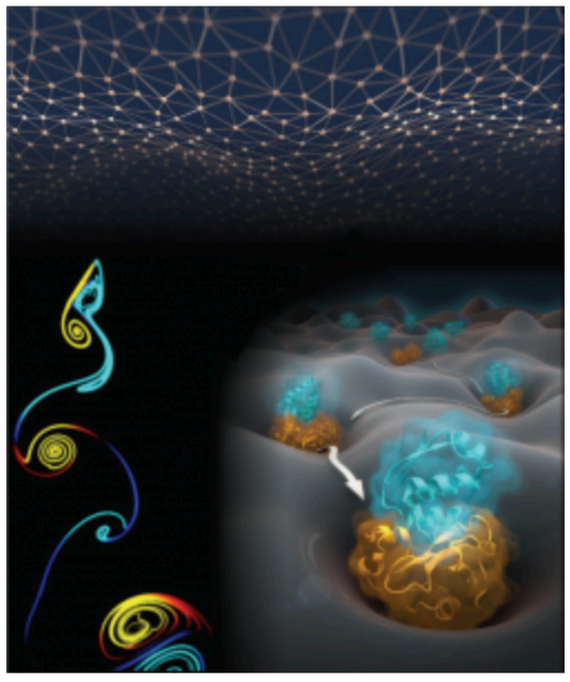
ml4physicalsciences.github.io

Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019

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.



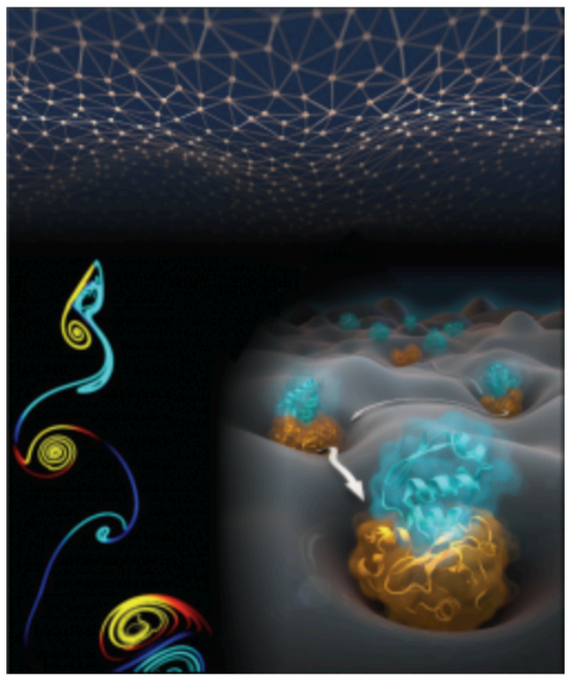
As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity 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



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.

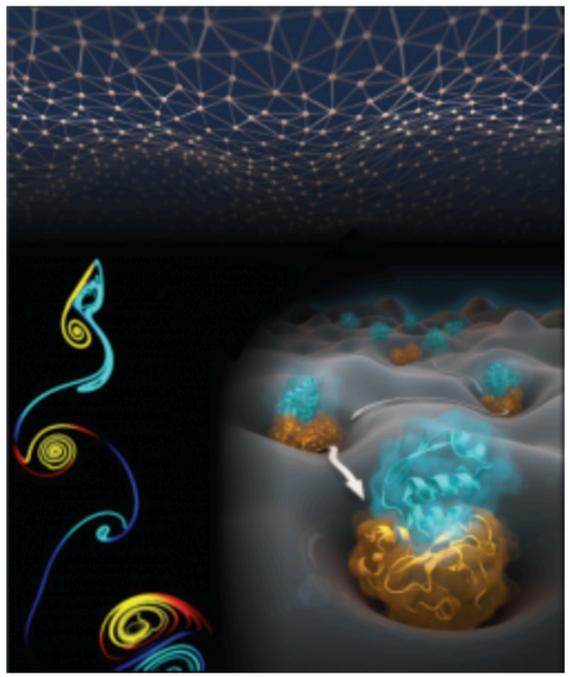
The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information 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

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.

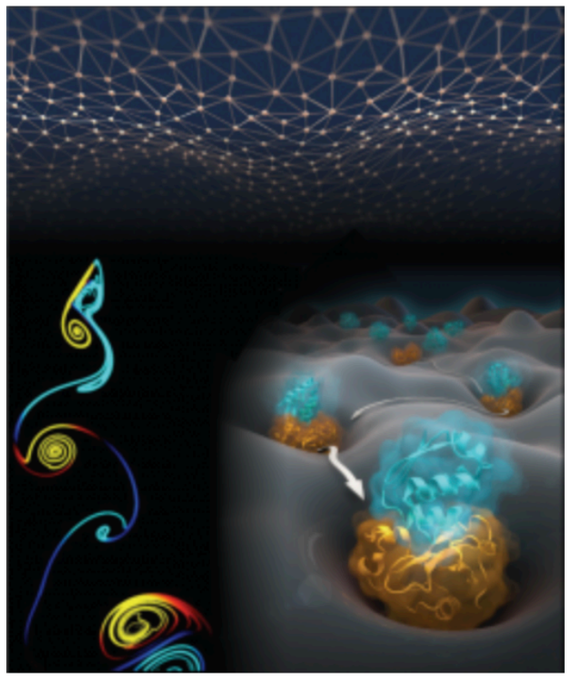
The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information 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

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.



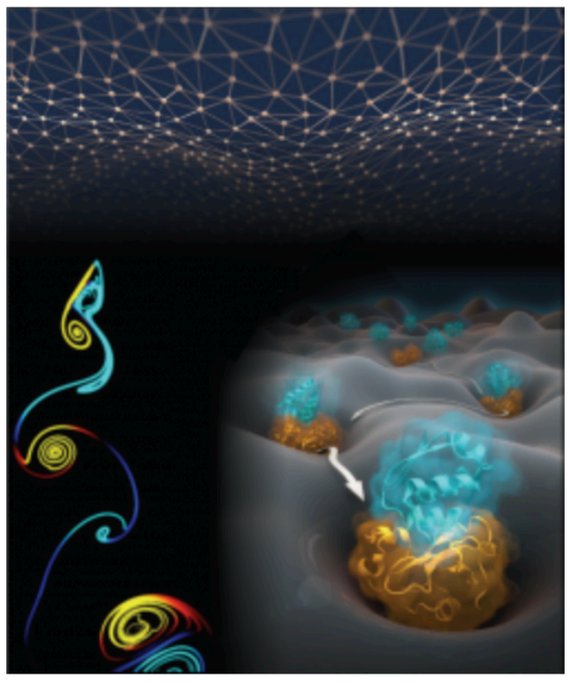
As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity 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



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.

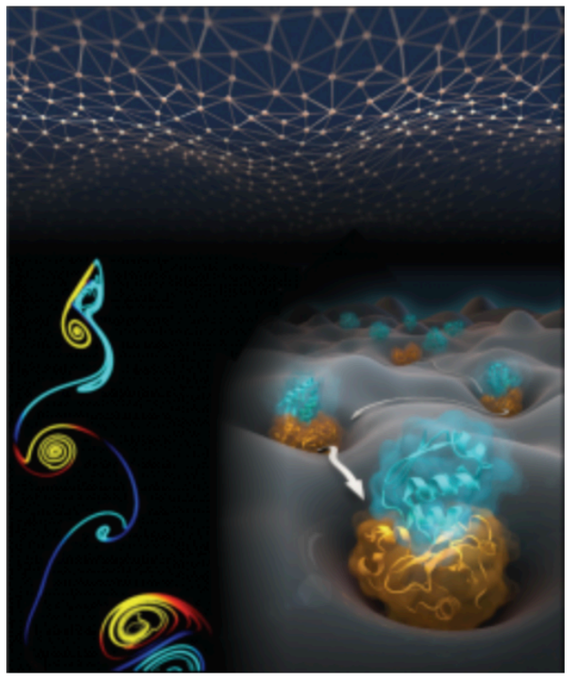
The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information 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

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.



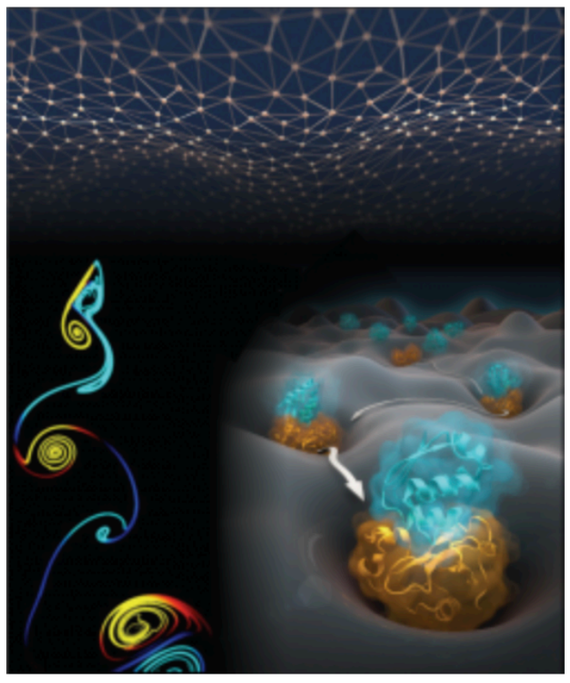
As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity 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



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.

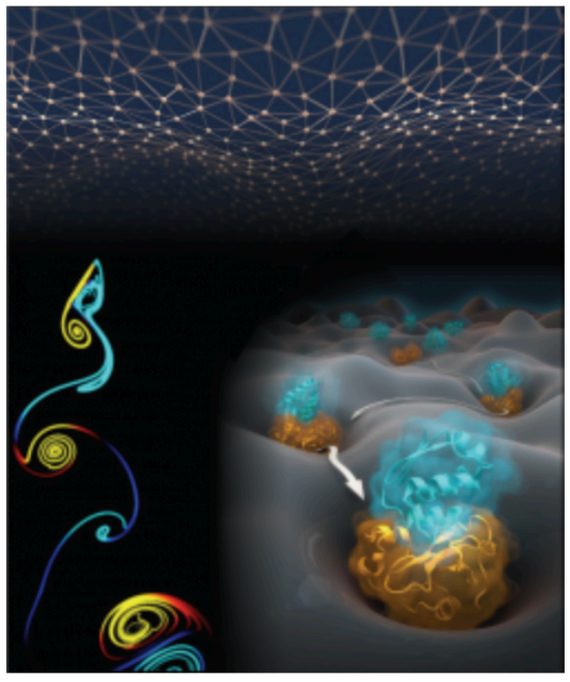
The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information 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

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.

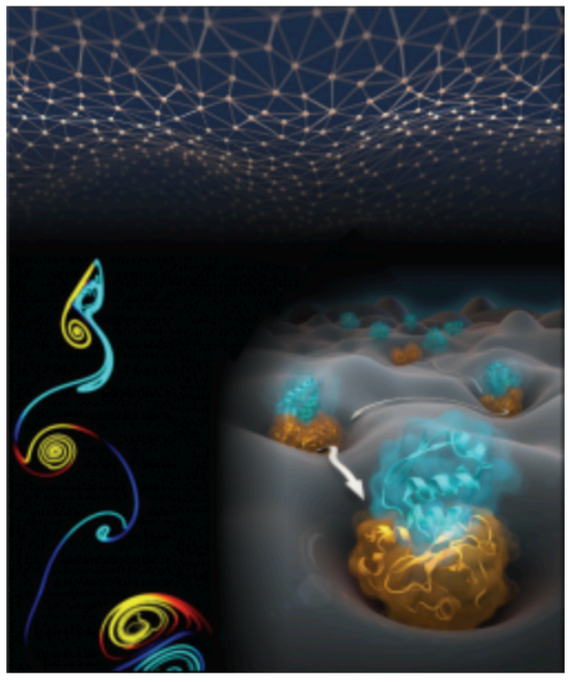
The exchange between fields can go in both directions. Since its beginning, machine learning has been inspired by methods from statistical physics. Many modern machine learning tools, such as variational inference and maximum entropy, are refinements of techniques invented by physicists. Physics, information 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

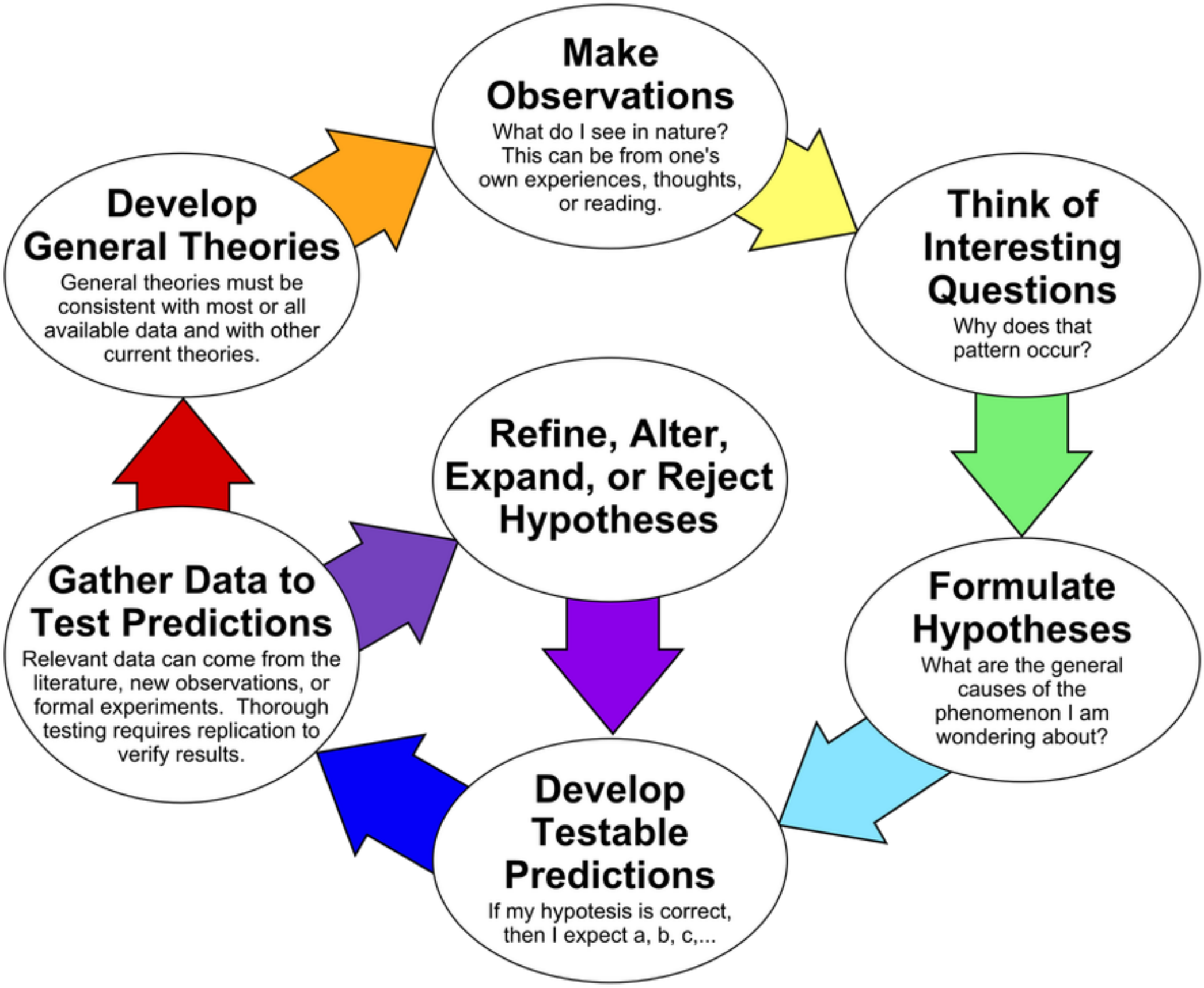
Machine Learning (ML) is quickly providing new powerful tools for physicists and chemists to extract essential information from large amounts of data, either from experiments or simulations. Significant steps forward in every branch of the physical sciences could be made by embracing, developing and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



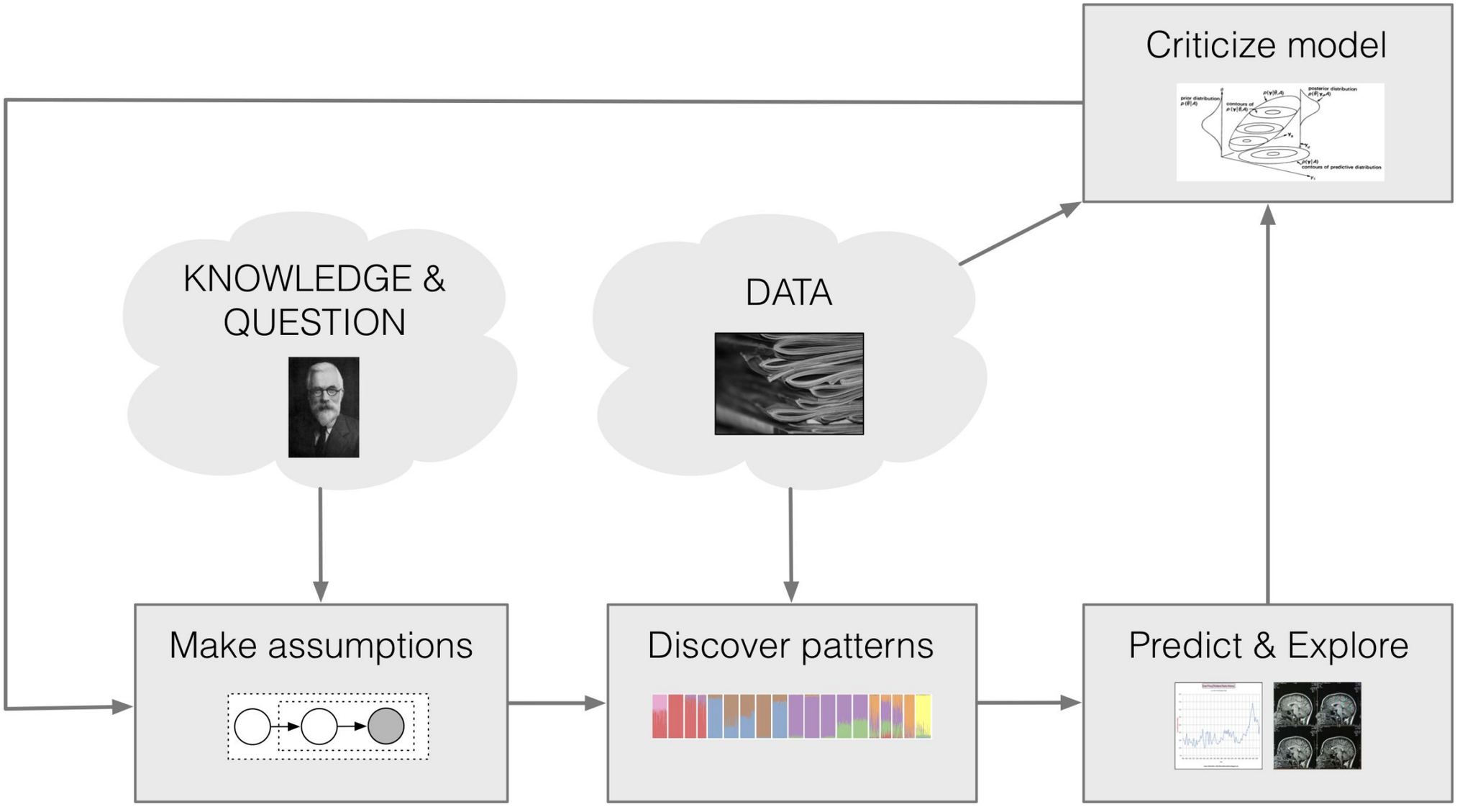
As yet, most applications of machine learning to physical sciences have been limited to the “low-hanging fruits,” as they have mostly been focused on fitting pre-existing physical models to data and on discovering strong signals. We believe that machine learning also provides an exciting opportunity 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



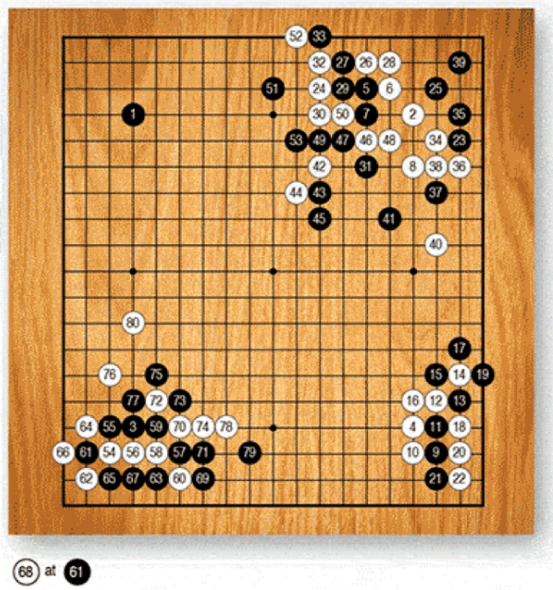
A PROBABILISTIC PIPELINE



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

Slide from David Blei, Rajesh Ranganath, Shakir Mohamed, Dustin Tran

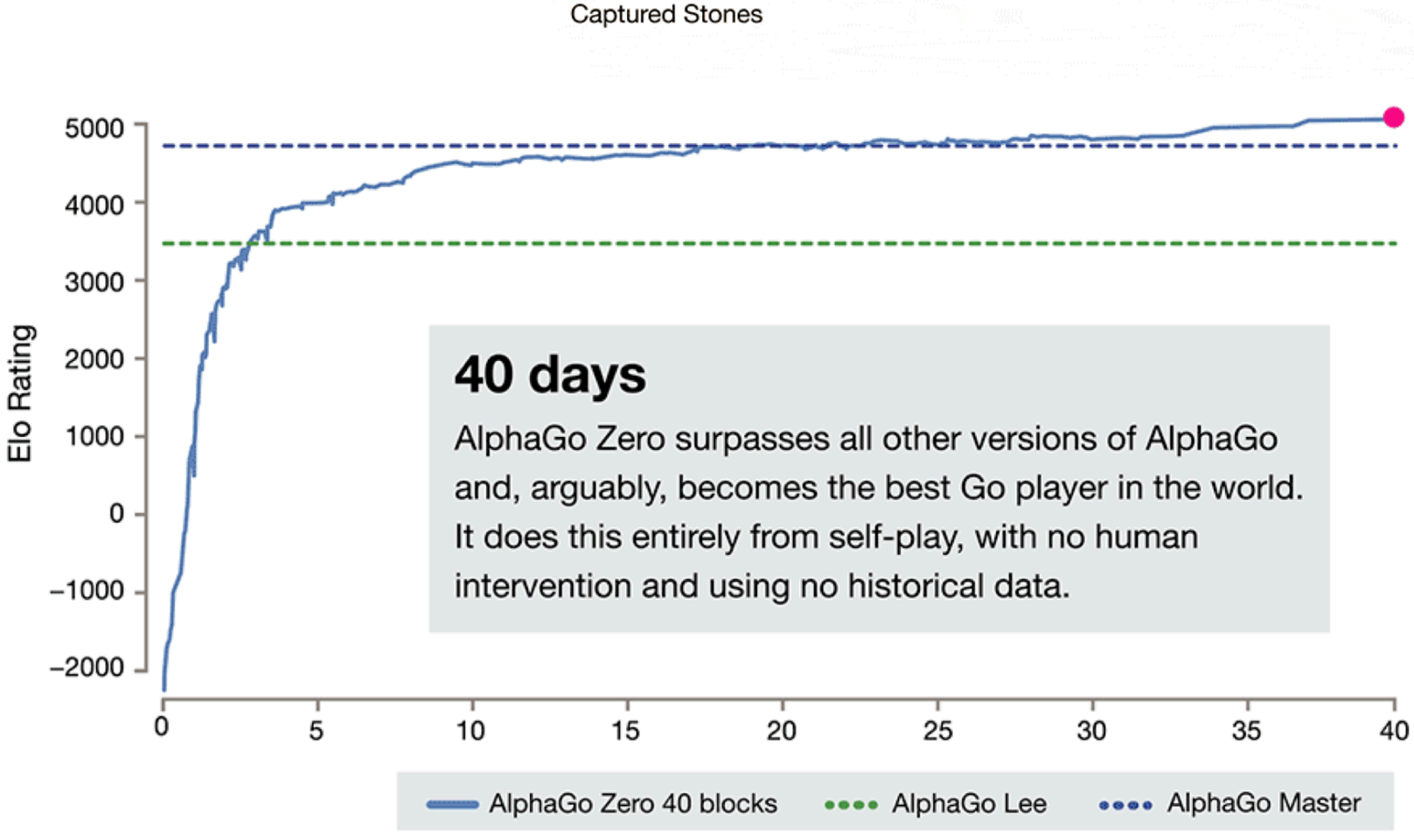
Scientist trying to decide what experiment to do next



(68) at (61)

70 hours

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

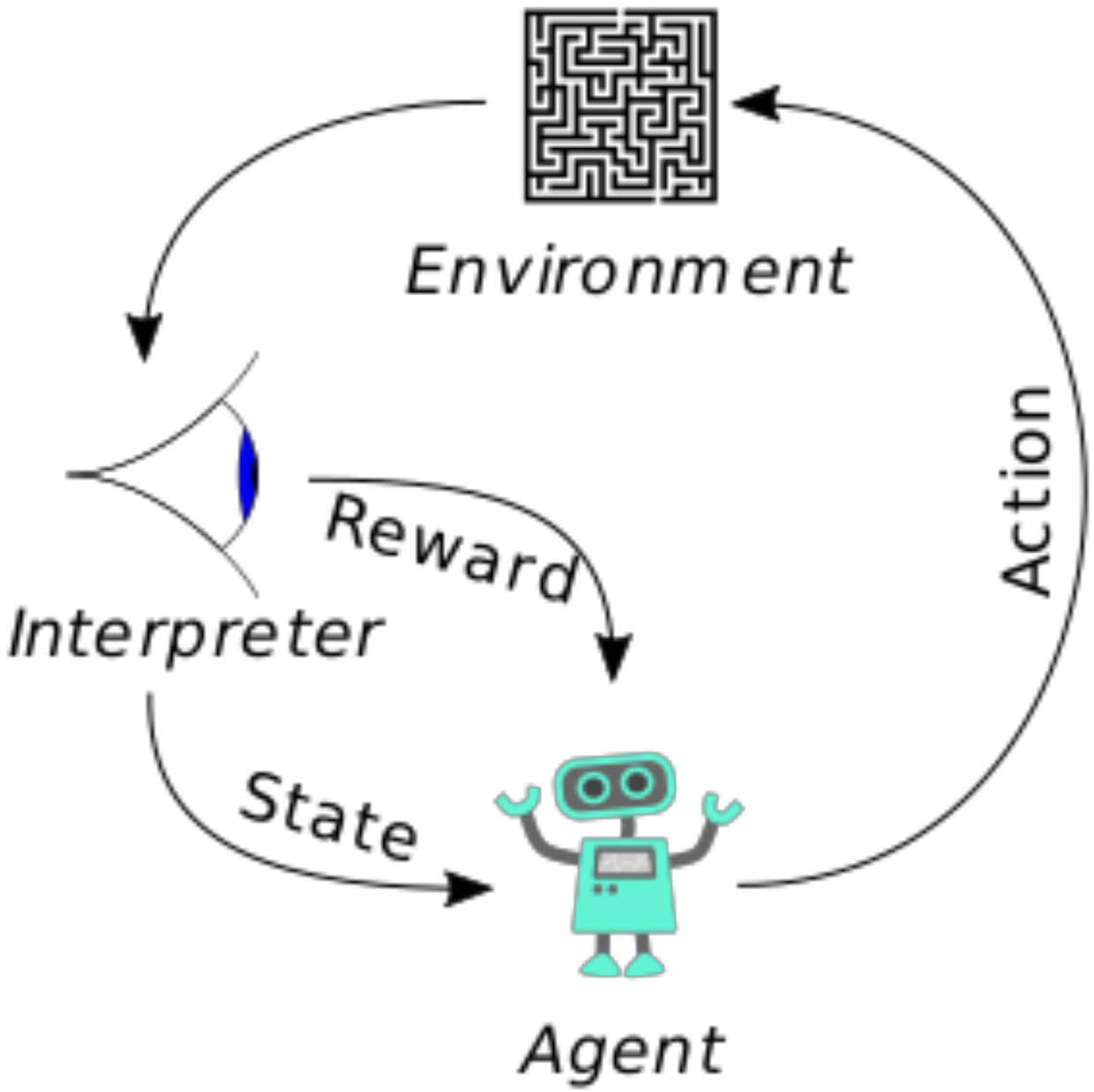


40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.

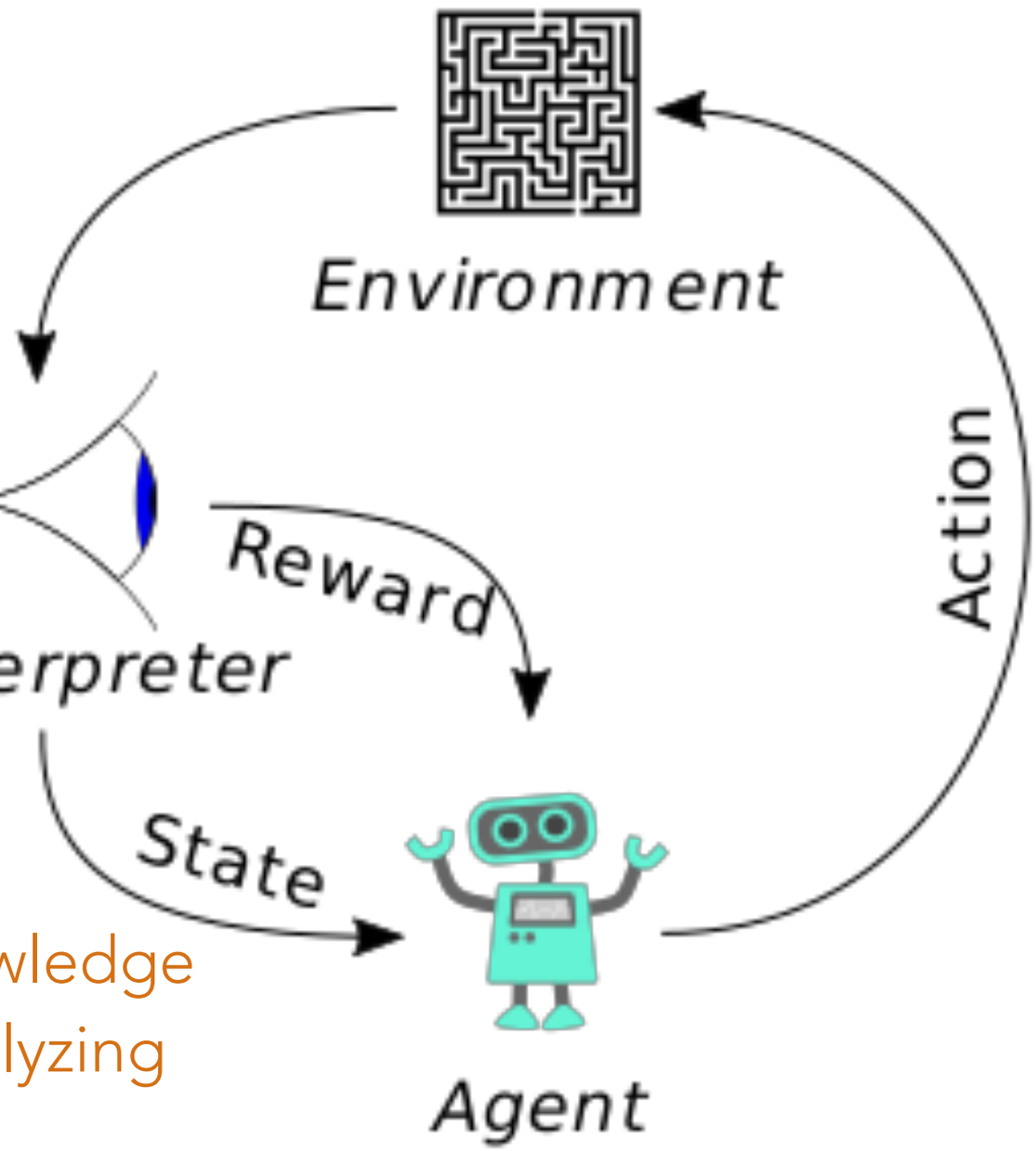
REINFORCEMENT LEARNING & SCIENTIFIC METHOD

Scientist trying to decide what experiment to do next



Scientist trying to decide what experiment to do next

perform experiment,
gather data



statistical analysis

decide which experiment to perform

updated knowledge based on analyzing data

NOTATION / TERMINOLOGY

forward modeling
generation
simulation

PREDICTION

θ
parameters of interest

p(x, z | θ, ν)

x
observed data
simulated data

z
latent variables

ν
nuisance parameters

INFERENCE

inverse problem
measurement
parameter estimation

STATISTICAL DECISION THEORY IN 1 SLIDE

Θ - States of nature; X - possible observations; A - action to be taken

$f(x|\theta)$ - 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

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

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

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

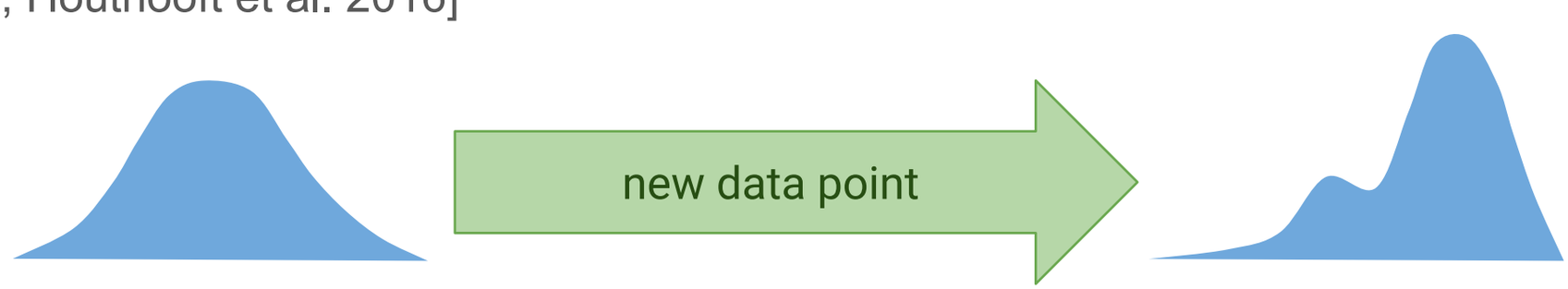


Dustin Tran
Research Scientist at Google Brain
trandustin@google.com
Blog

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]



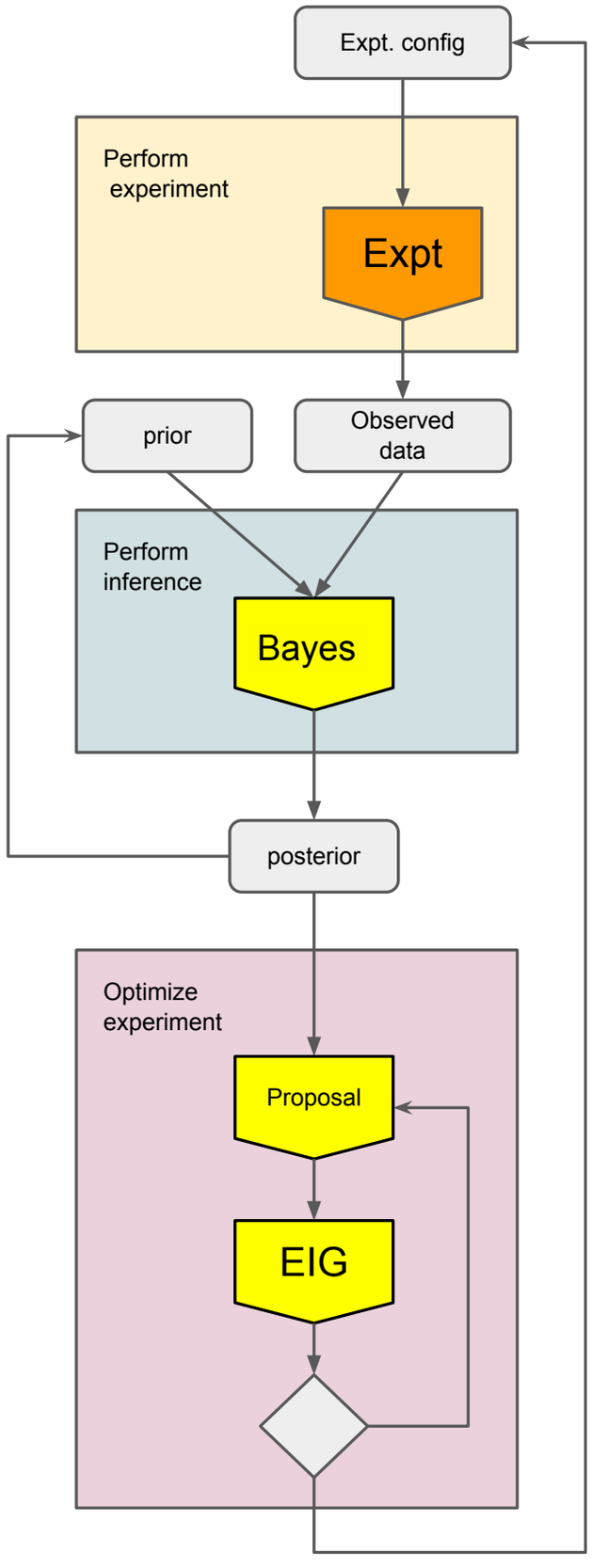
$$\arg\max_{\boldsymbol{x}} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y\sim p(y|\boldsymbol{x}\mathcal{D})}\left[\mathrm{H}[\boldsymbol{\theta}|y, \boldsymbol{x}, \mathcal{D}]\right]$$

Uncertainty determines which **x** is most informative and, therefore, the model's success.

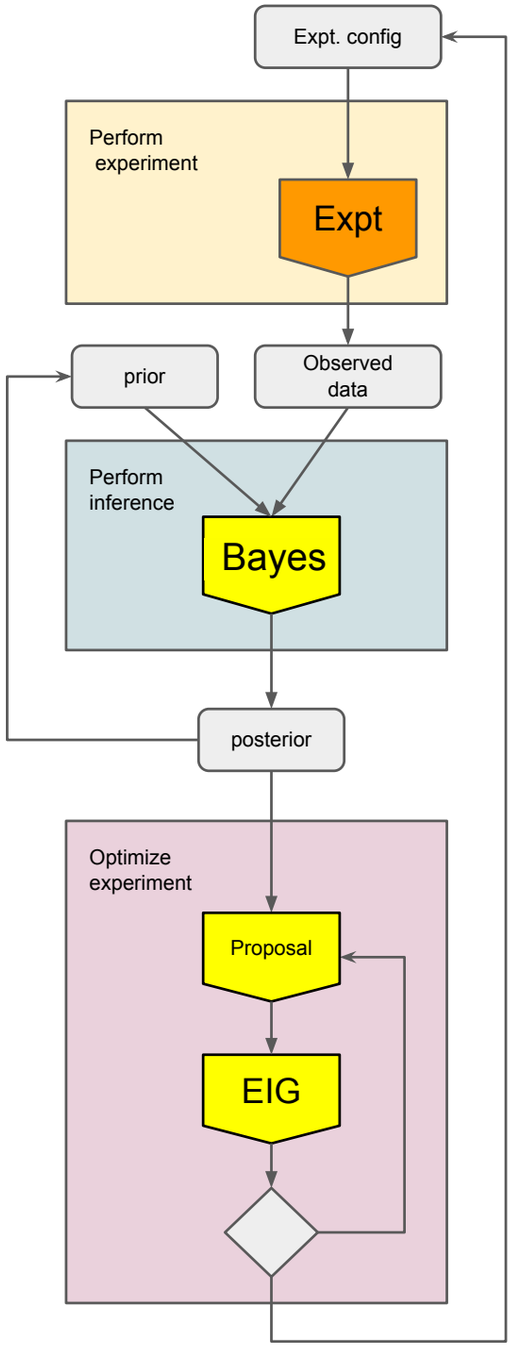
[Hafner et al., 2019]

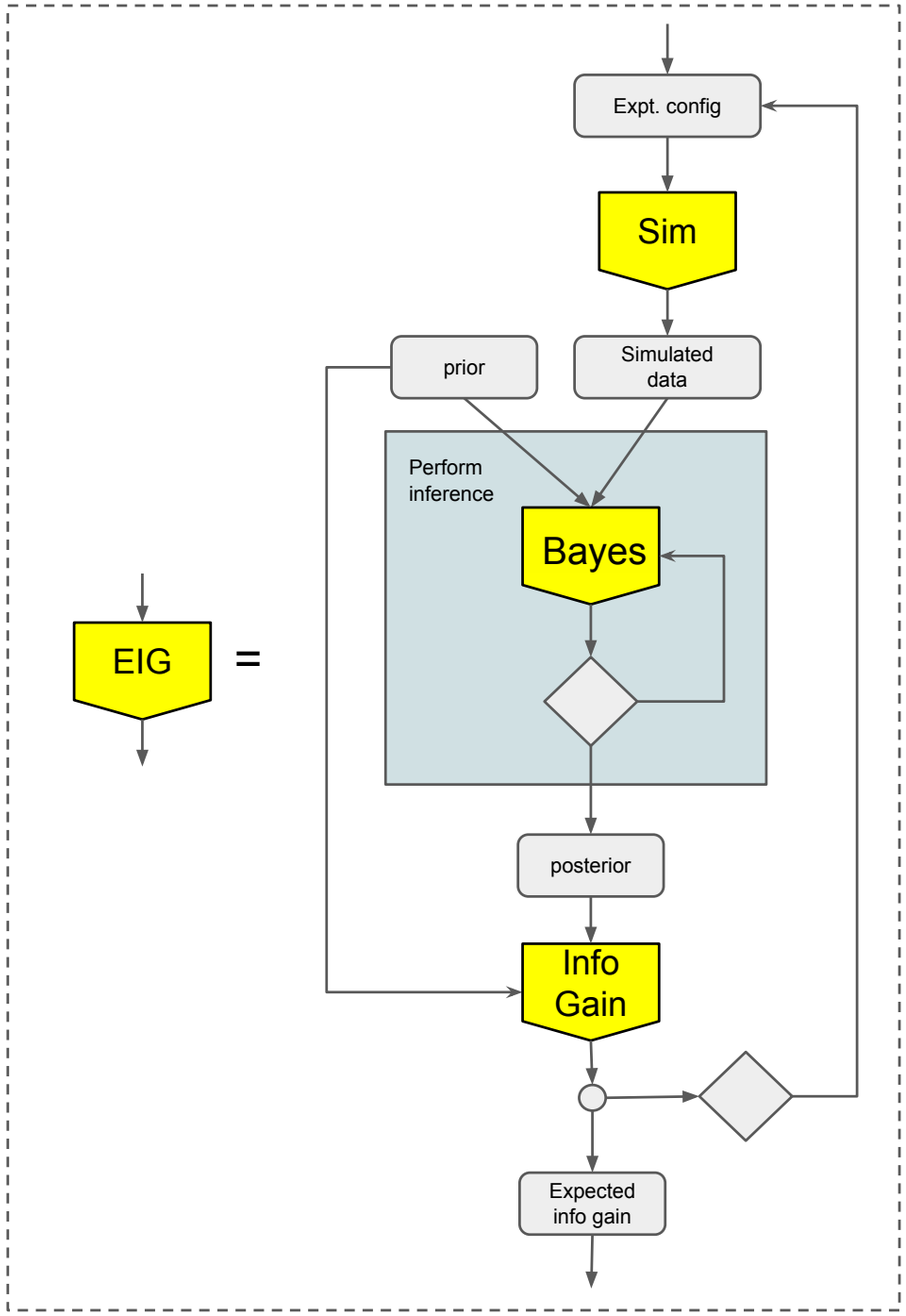
"ACTIVE SCIENCING"

https://github.com/cranmer/active_sciencing



"ACTIVE SCIENCING"





https://github.com/cranmer/active_sciencing

EXPECTED INFORMATION GAIN

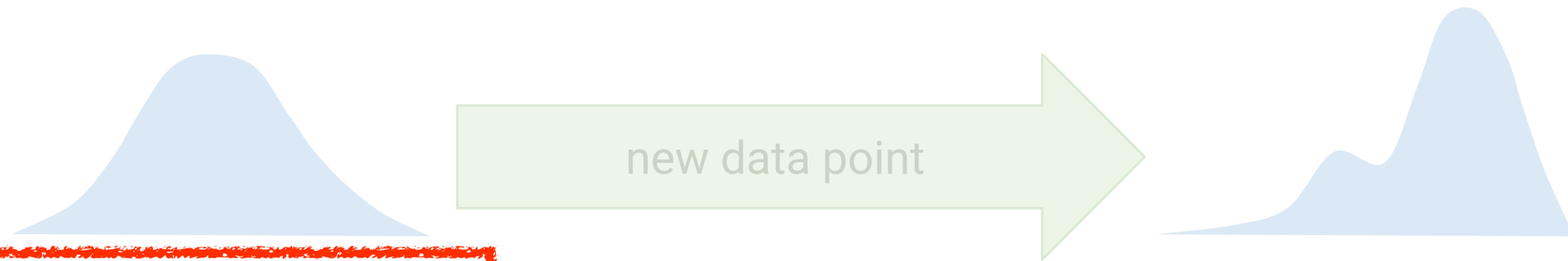


Dustin Tran
Research Scientist at Google Brain
trandustin@google.com
Blog

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]



$$\arg\max_{\boldsymbol{x}} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y\sim p(y|\boldsymbol{x}\mathcal{D})}\left[\mathrm{H}[\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D}]\right]$$

Uncertainty determines which **x** is most informative and, therefore, the model's success.

[Hafner et al., 2019]

SYNTHESIS

active learning / sequential design / black box optimization



Active Sciencing



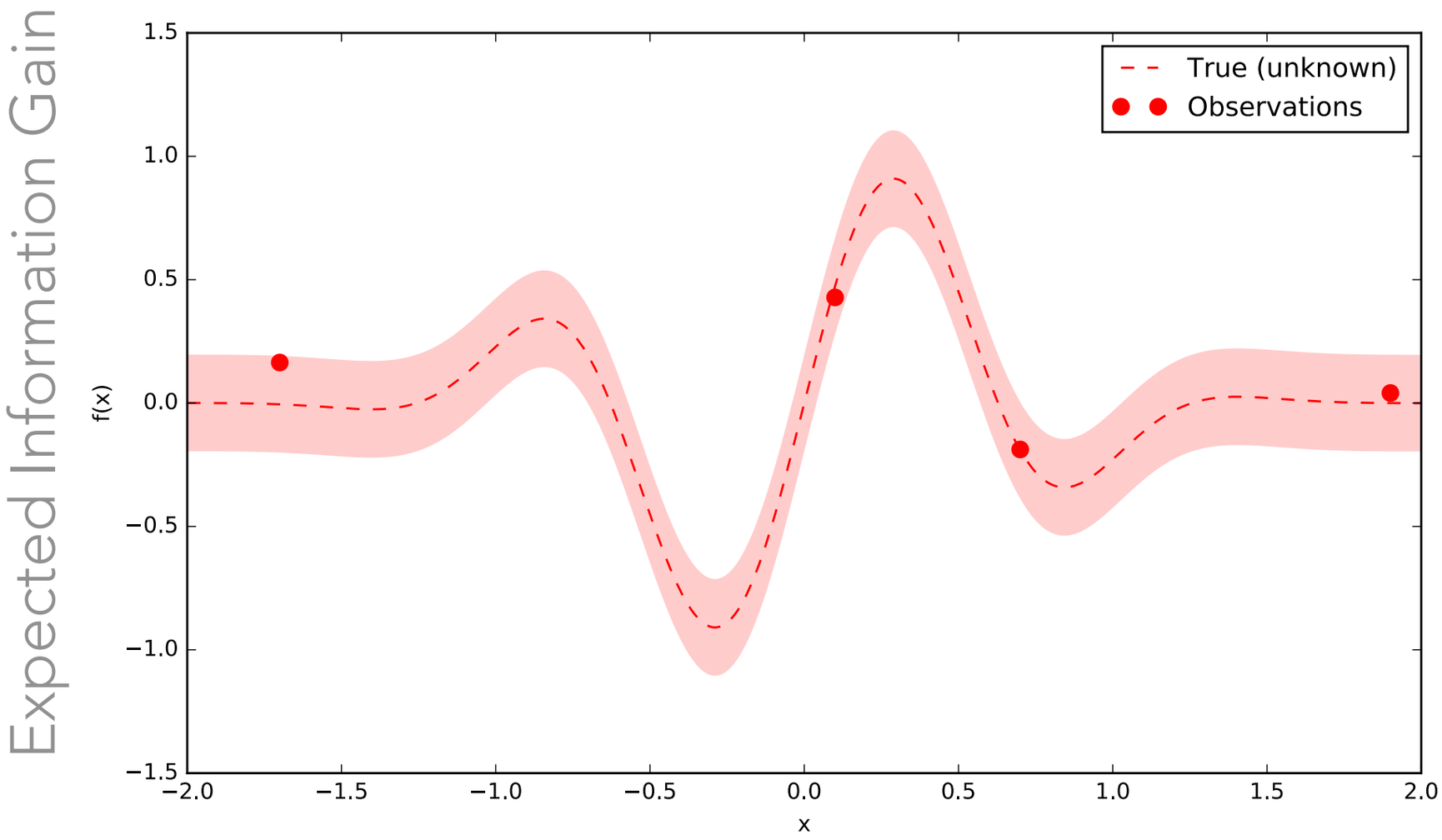
reusable workflows



simulation-based / likelihood-free inference engines

BAYESIAN OPTIMIZATION

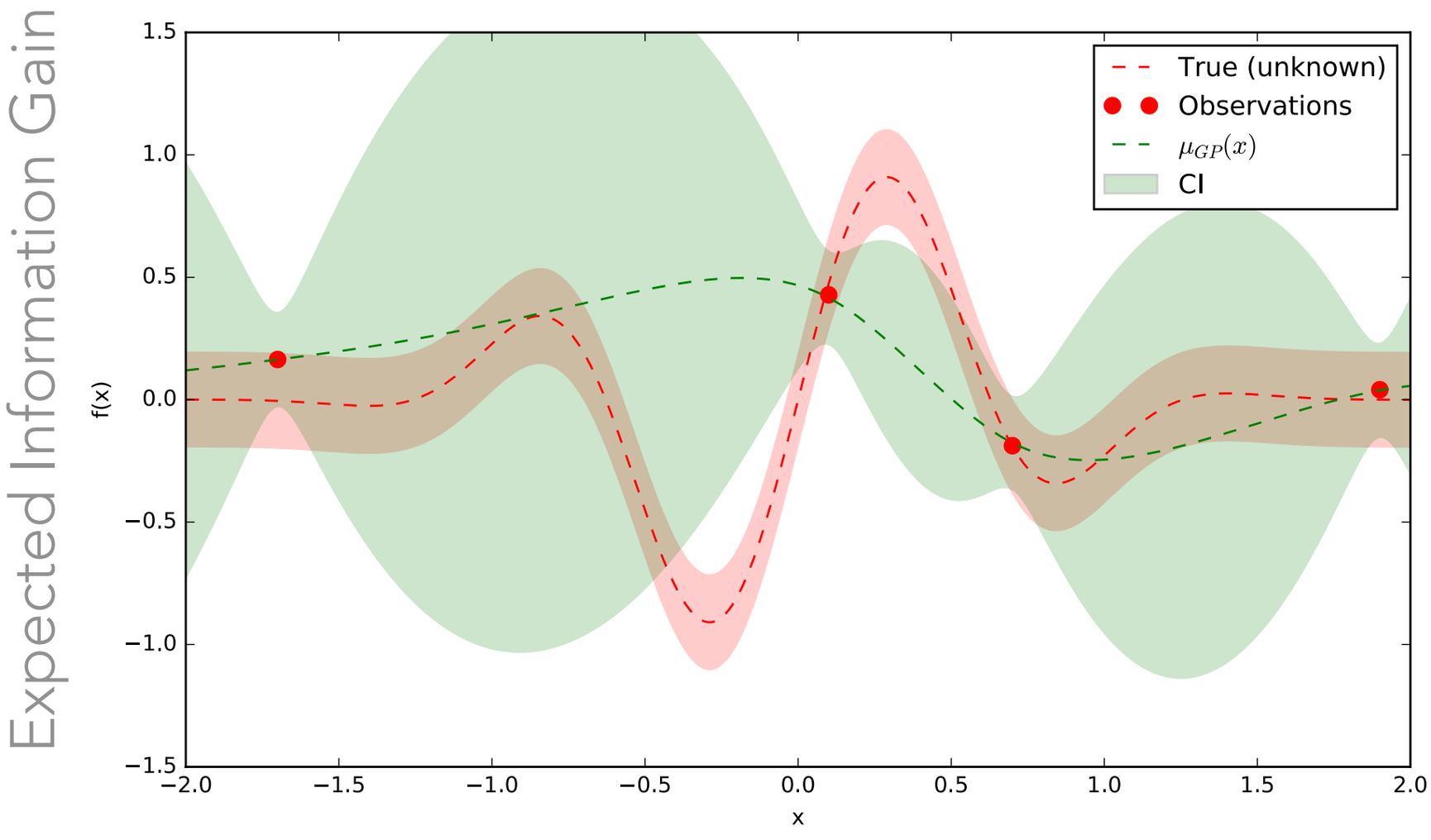
What experimental configuration should we evaluate next?



Experimental Configuration

Slides from Gilles Louppe. See scikit-optimize project

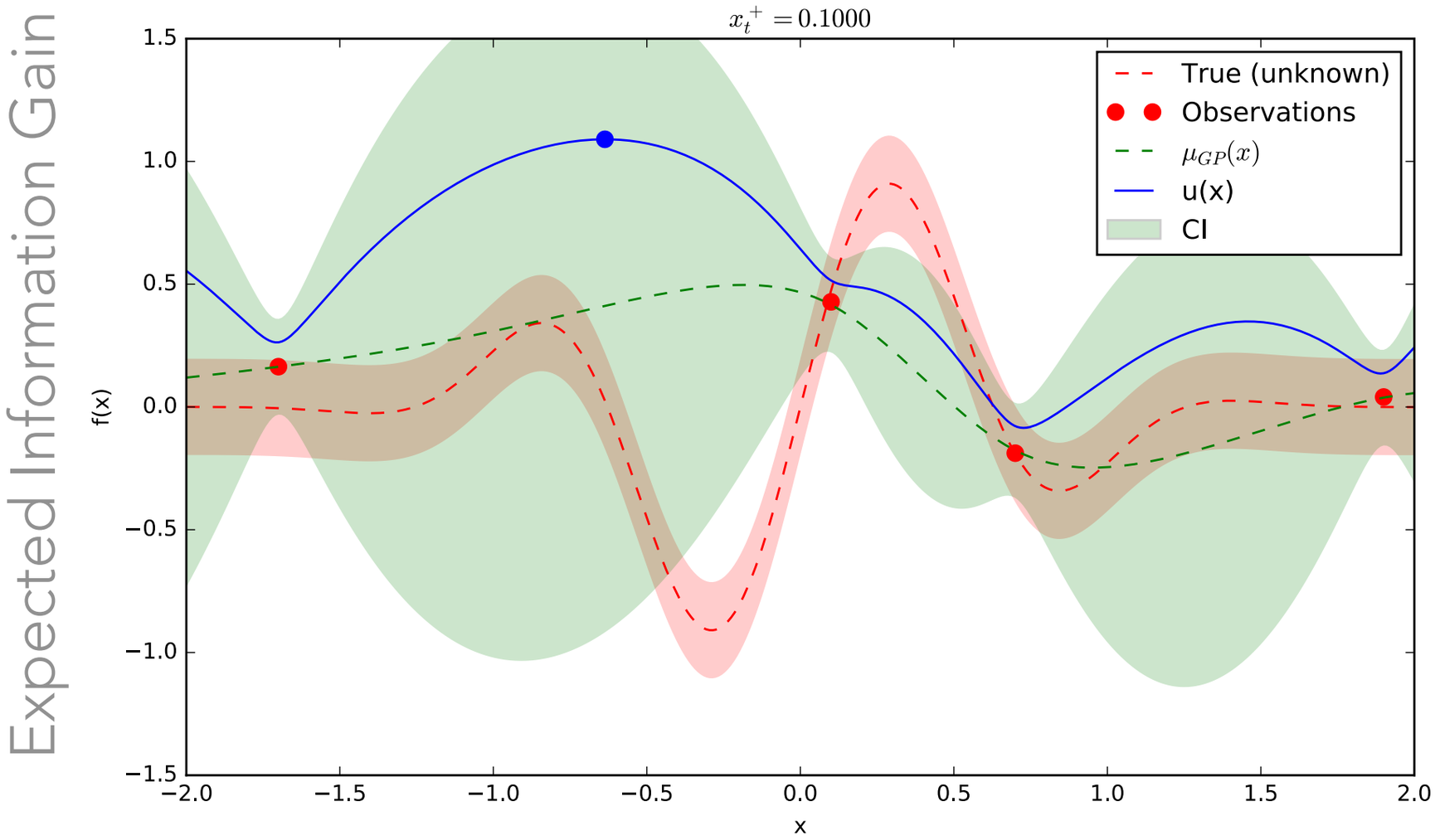
Build a probabilistic model for the EIG objective function



Slides from Gilles Louppe. See scikit-optimize project

BAYESIAN OPTIMIZATION

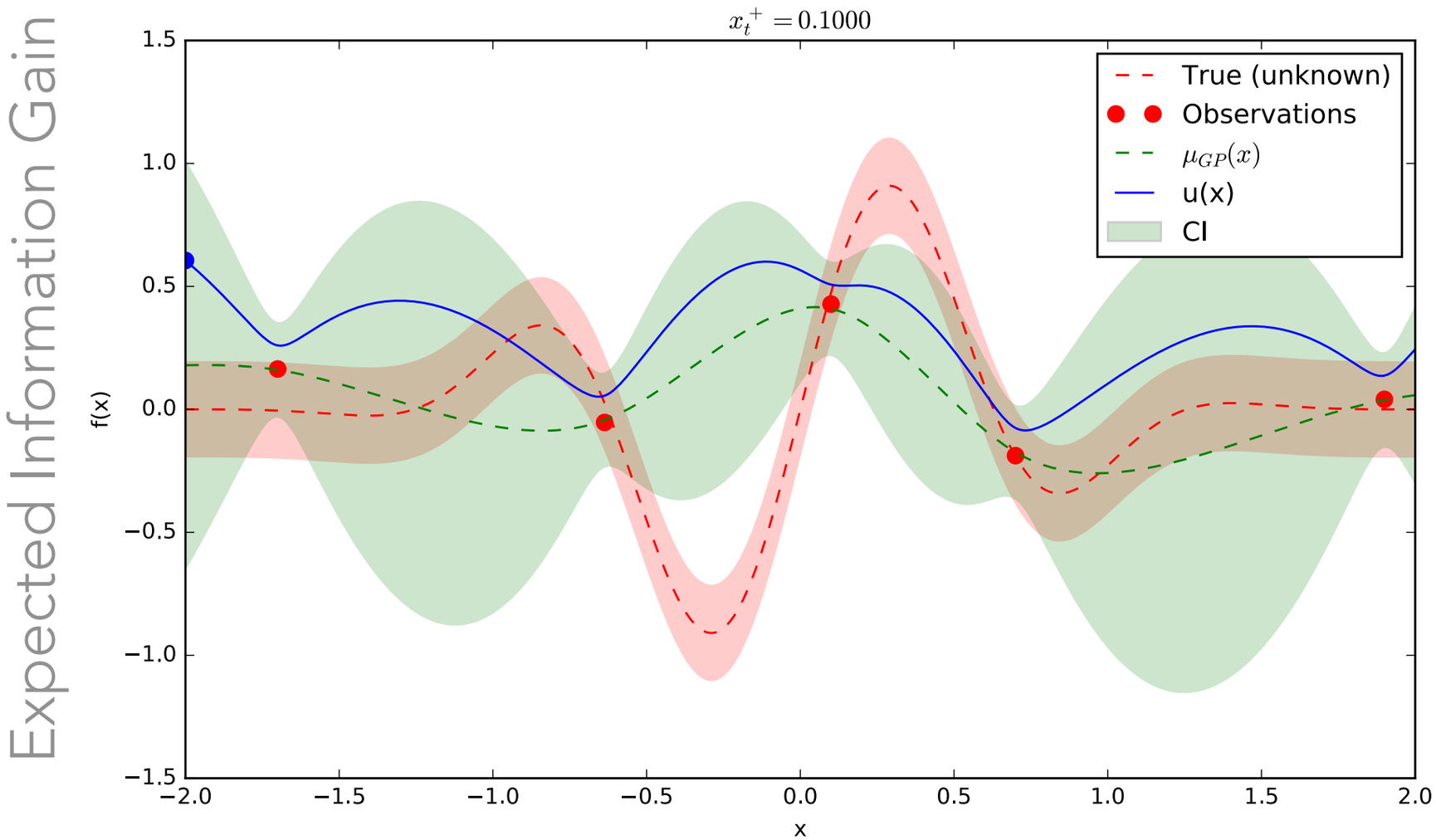
Use it to define an inexpensive acquisition function, and optimize that.
Use this as next configuration to evaluate expensive EIG



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

Slides from Gilles Louppe. See scikit-optimize project

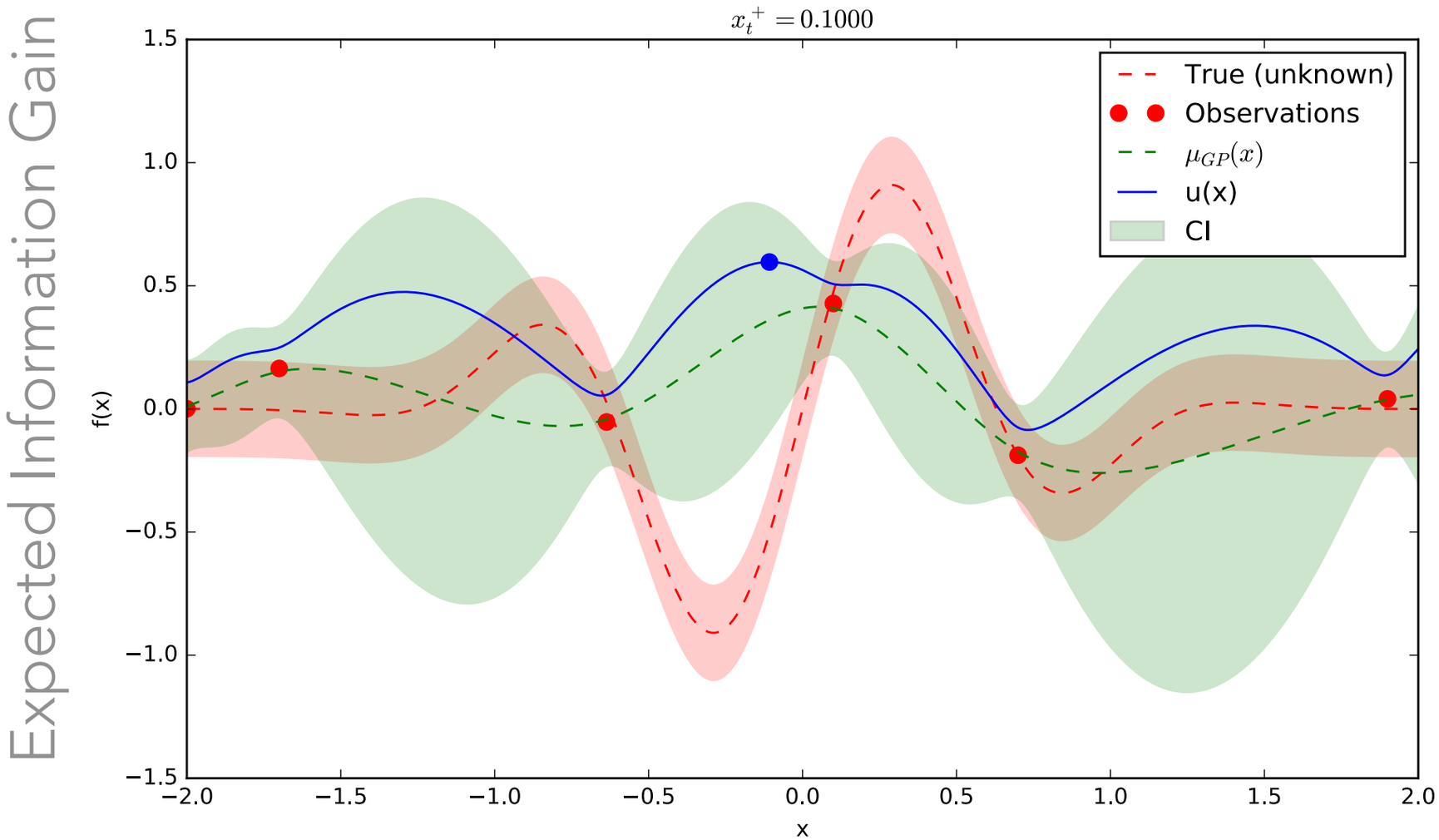
Then evaluate EIG for that configuration. Repeat



Experimental Configuration

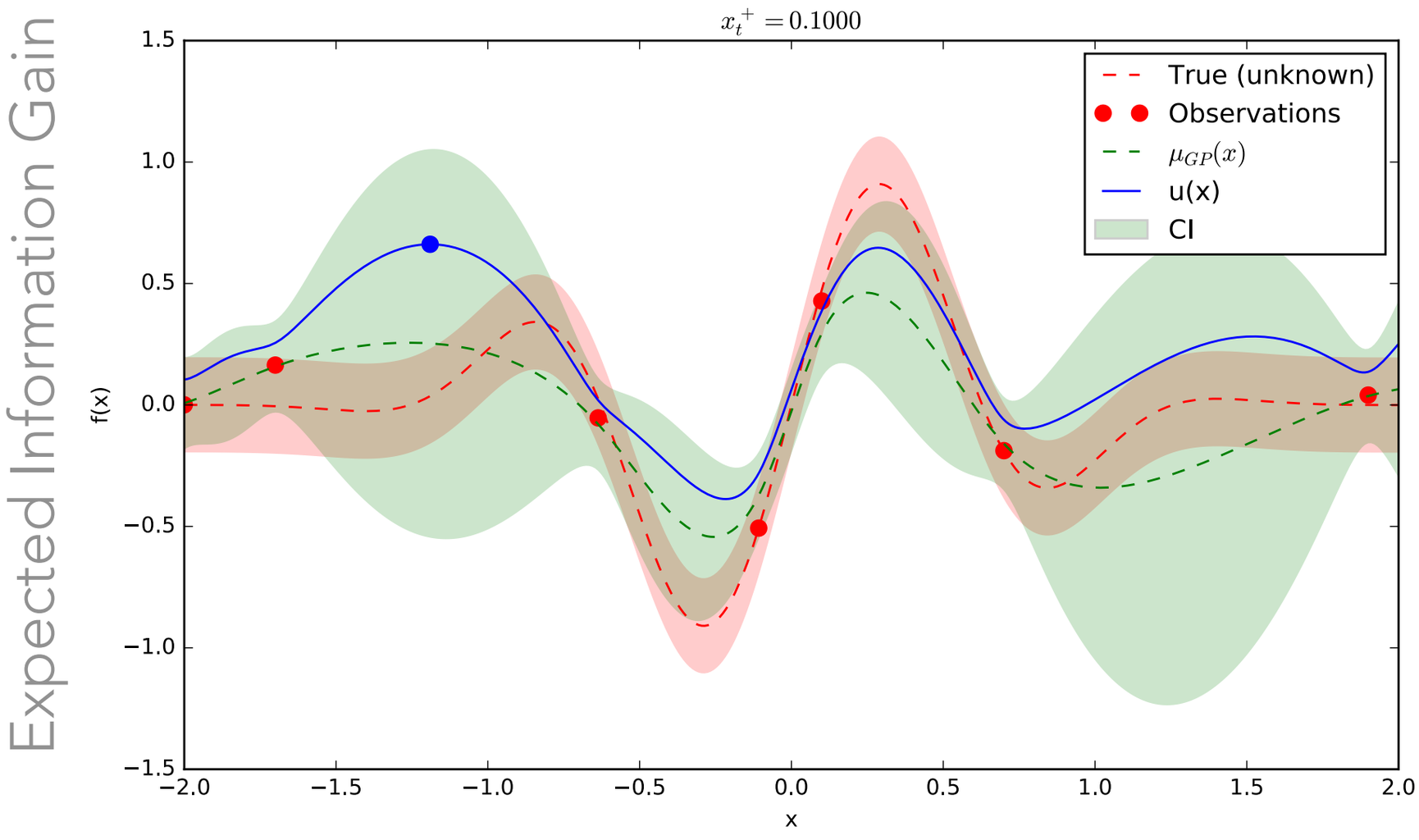
Slides from Gilles Louppe. See scikit-optimize project

Then evaluate EIG for that configuration. Repeat



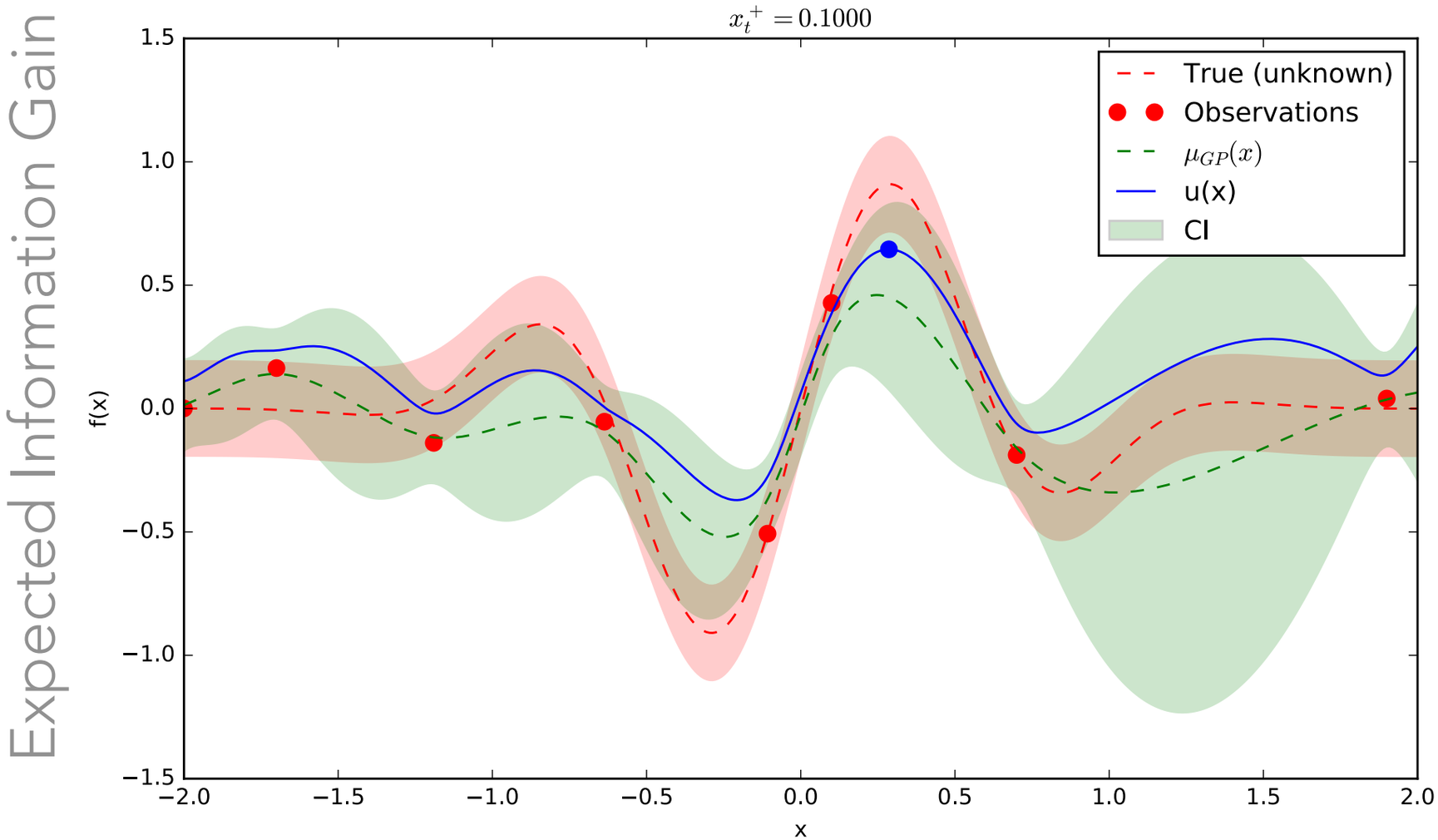
Experimental Configuration

Then evaluate EIG for that configuration. Repeat



Experimental Configuration

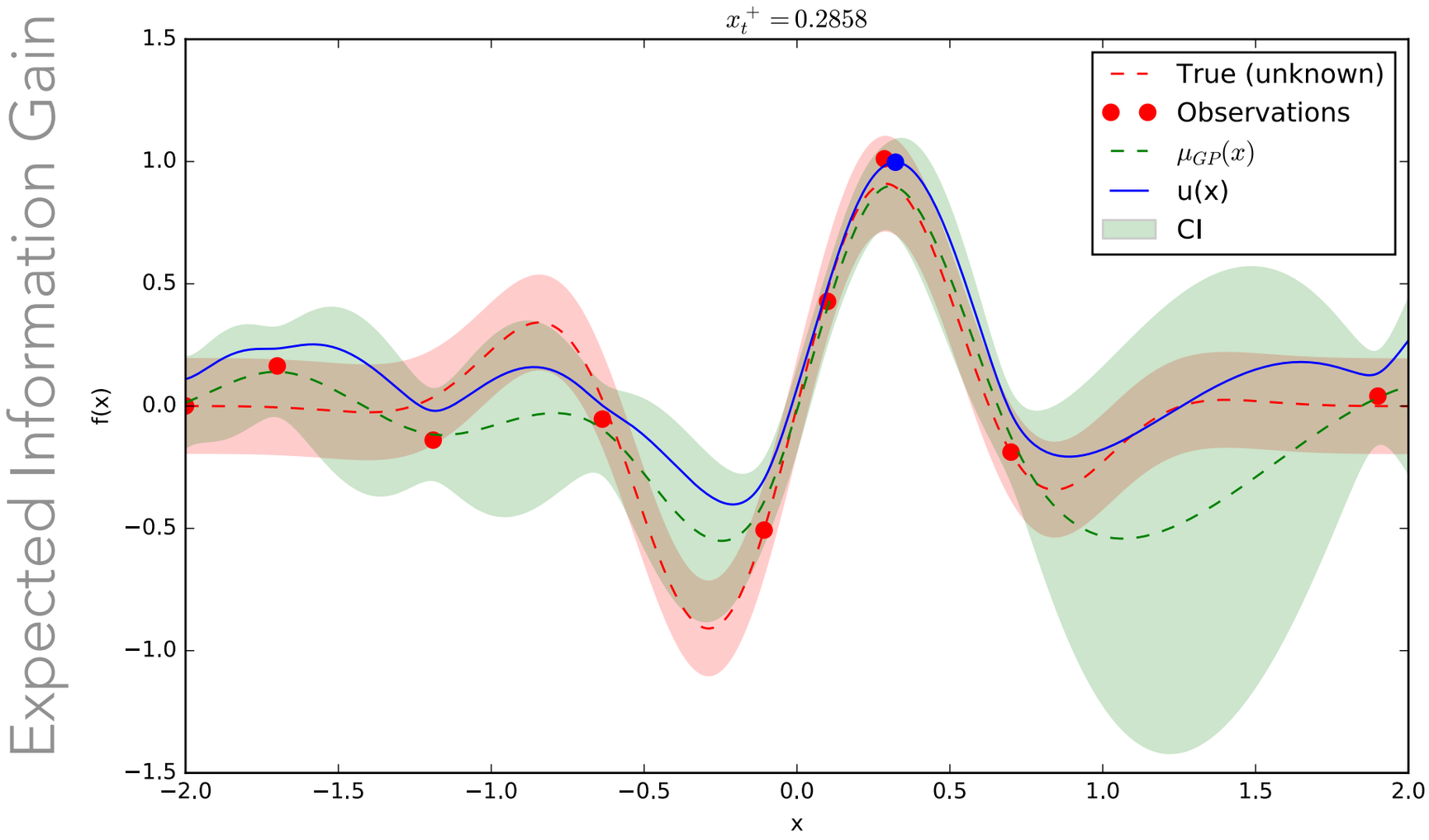
Then evaluate EIG for that configuration. Repeat



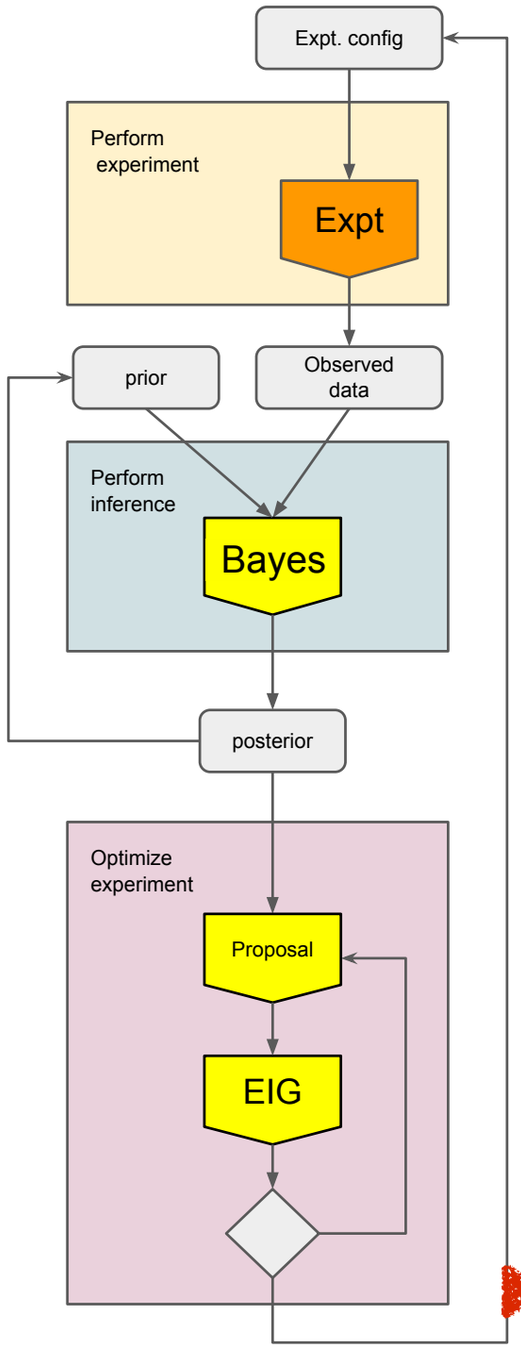
Experimental Configuration

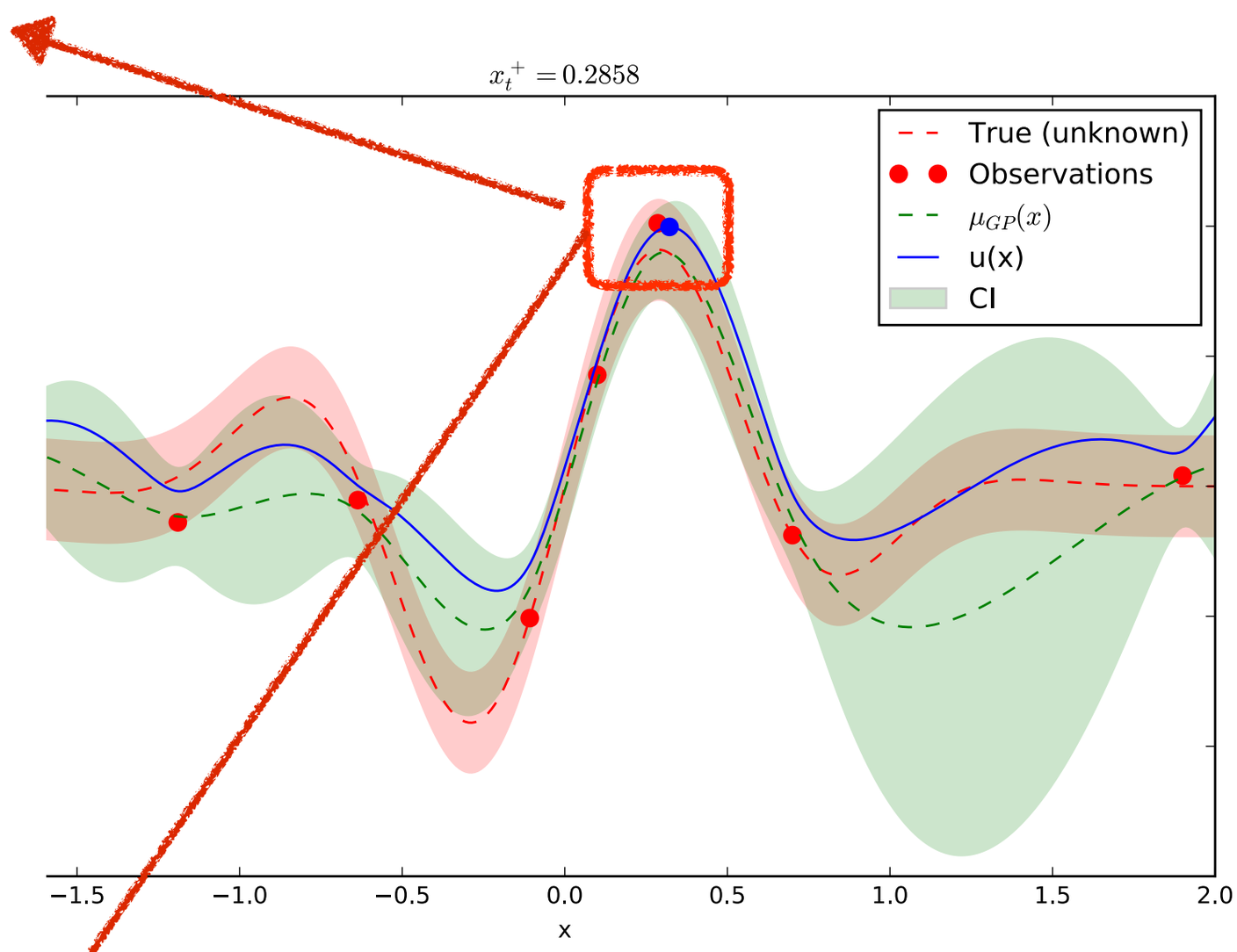
Slides from Gilles Louppe. See scikit-optimize project

Then evaluate EIG for that configuration. Repeat



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



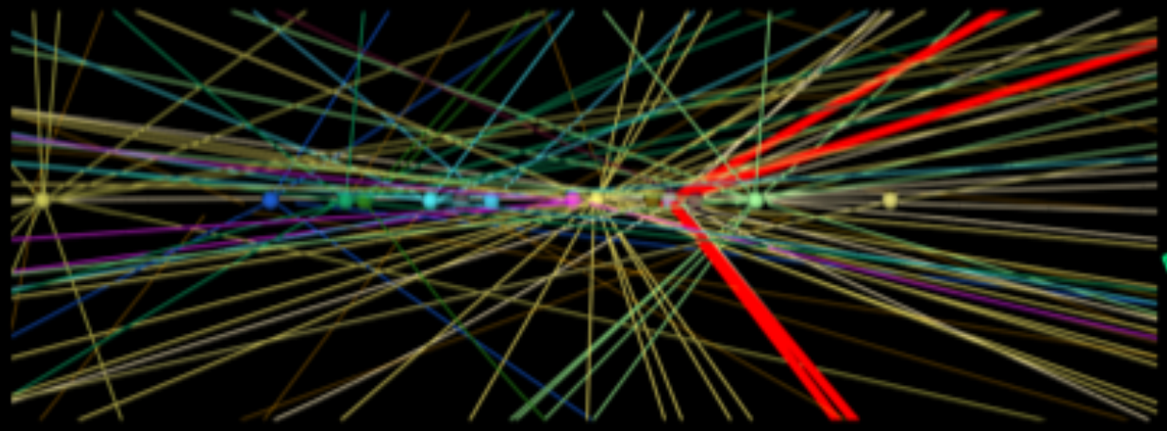


Experimental Configuration

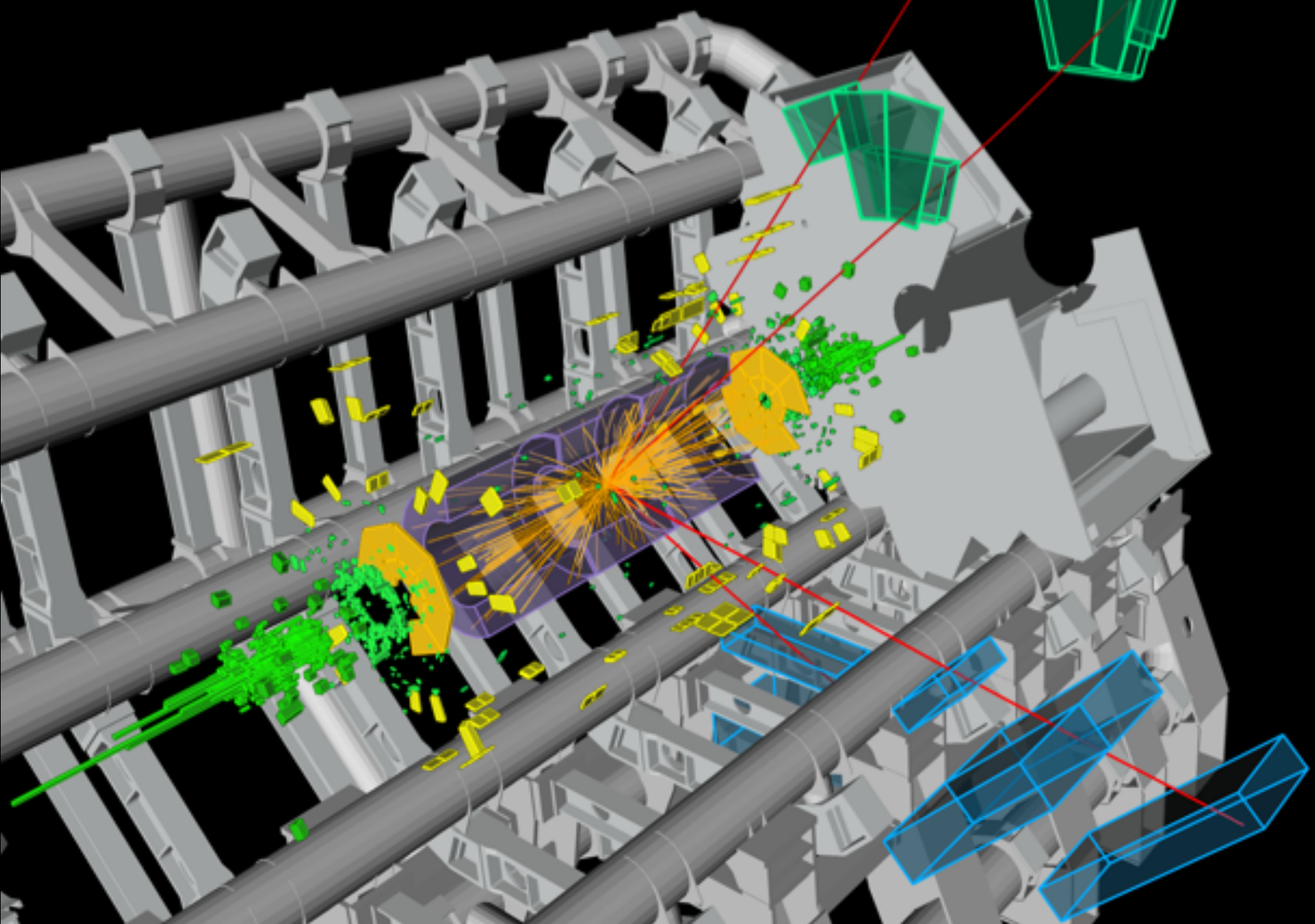
What is the probability model used for inference?

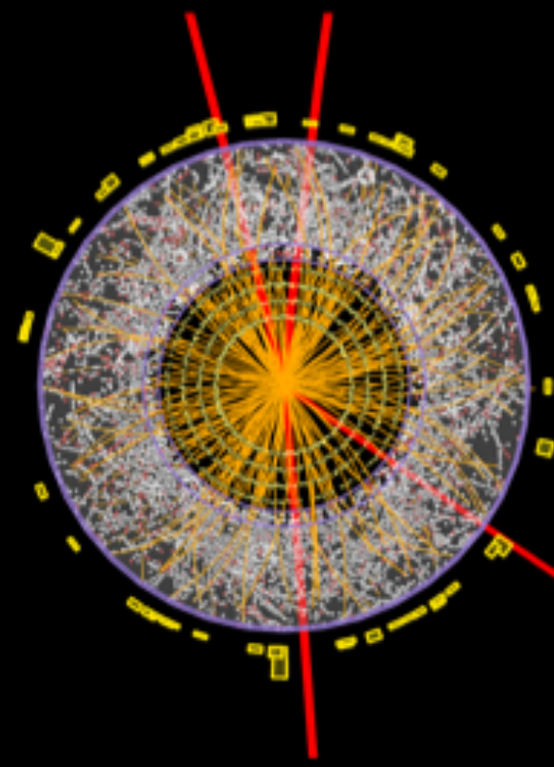
Shouldn’t we use the same simulator we used to evaluate the expected information gain?

PARTICLE PHYSICS



ATLAS EXPERIMENT
http://atlas.ch





Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

THE CAUSAL, GENERATIVE MODEL

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

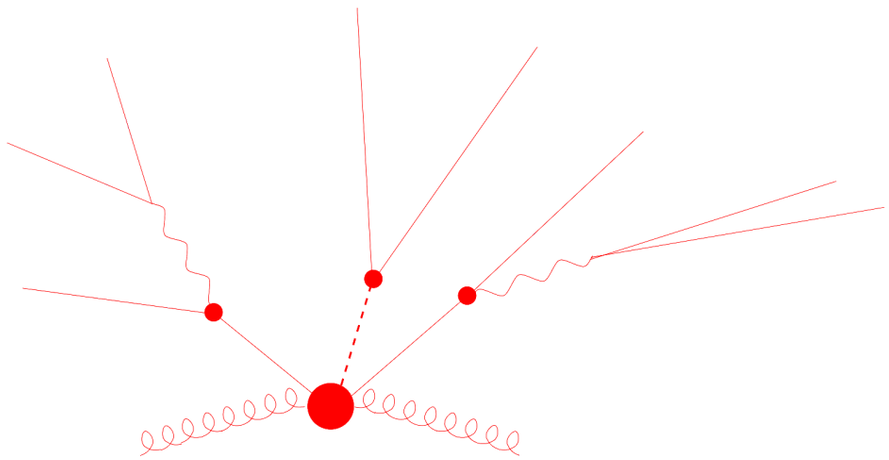
THE CAUSAL, GENERATIVE MODEL

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L \;+\; \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} \quad + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



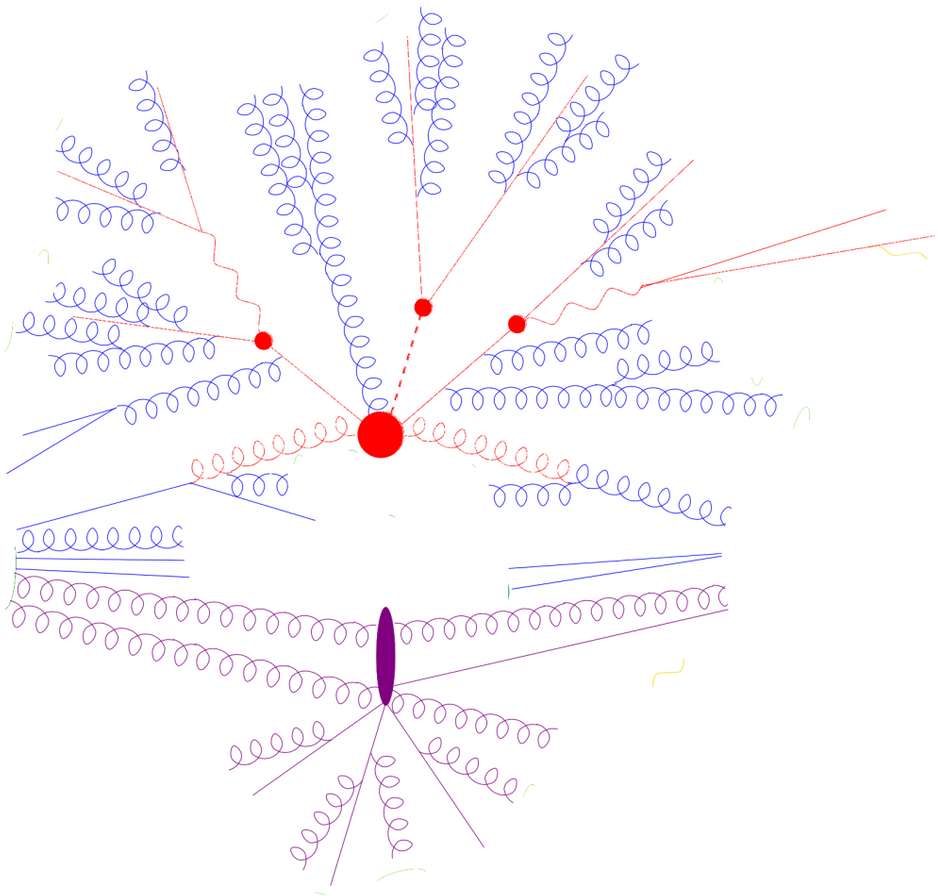
THE CAUSAL, GENERATIVE MODEL

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

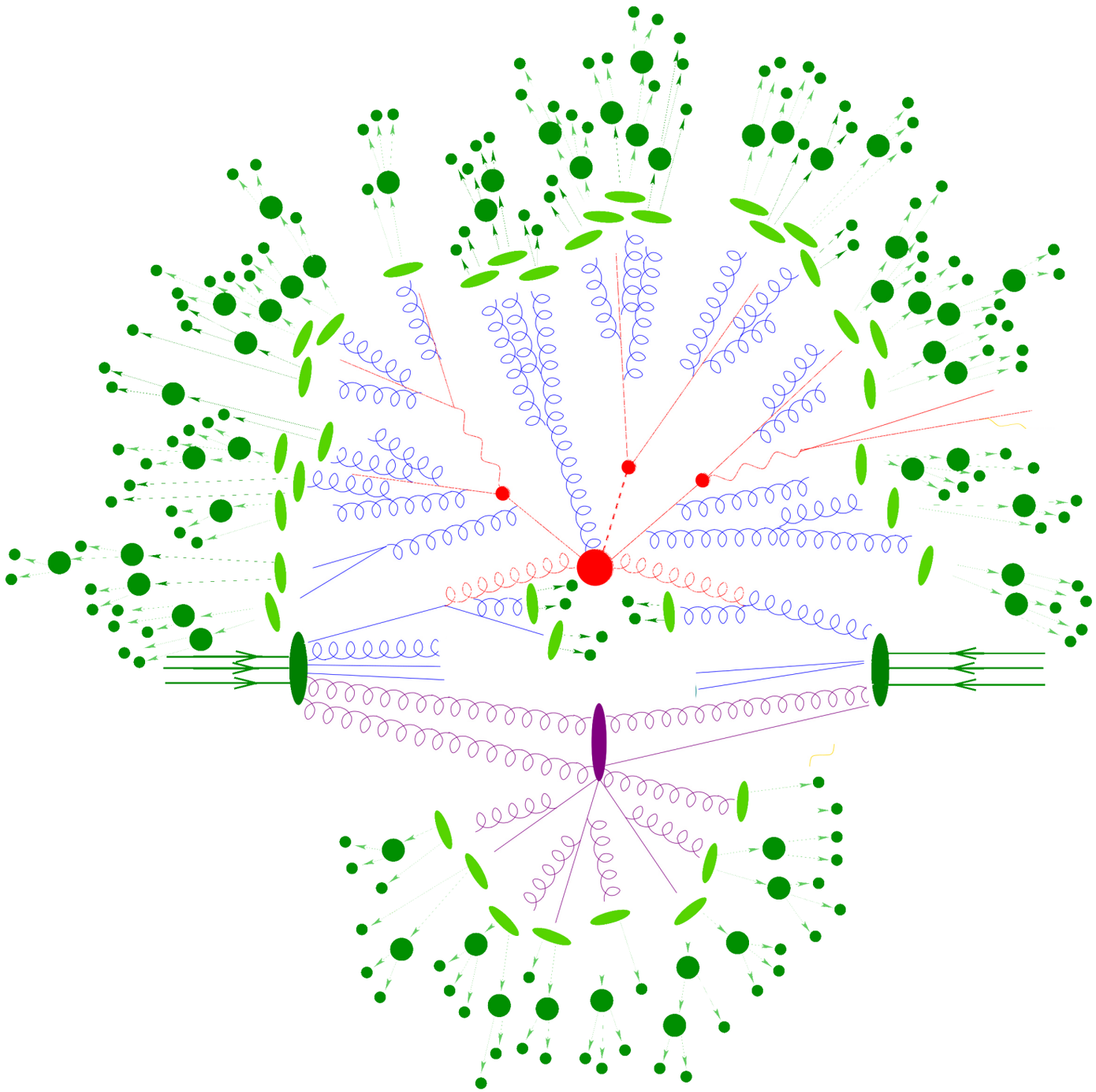
$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



THE CAUSAL, GENERATIVE MODEL



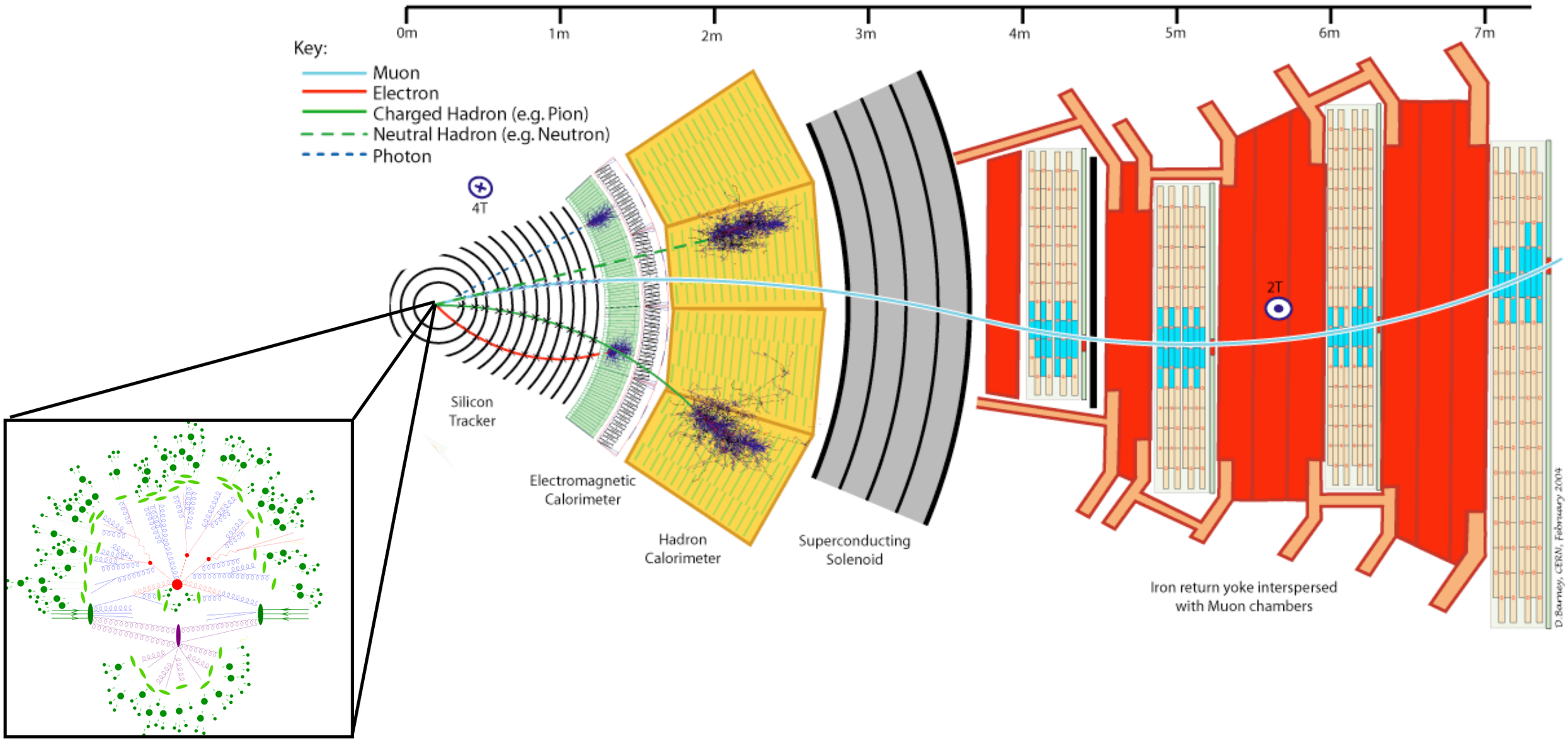
$$
\begin{aligned}
\mathcal{L}_{SM} = & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu}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

THE CAUSAL, GENERATIVE MODEL

Conceptually: Prob(detector response | particles)

Implementation: Monte Carlo integration over micro-physics

Consequence: evaluation of the likelihood is intractable



GALTON BOARD

Say we want to infer θ, 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 θ, 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.

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

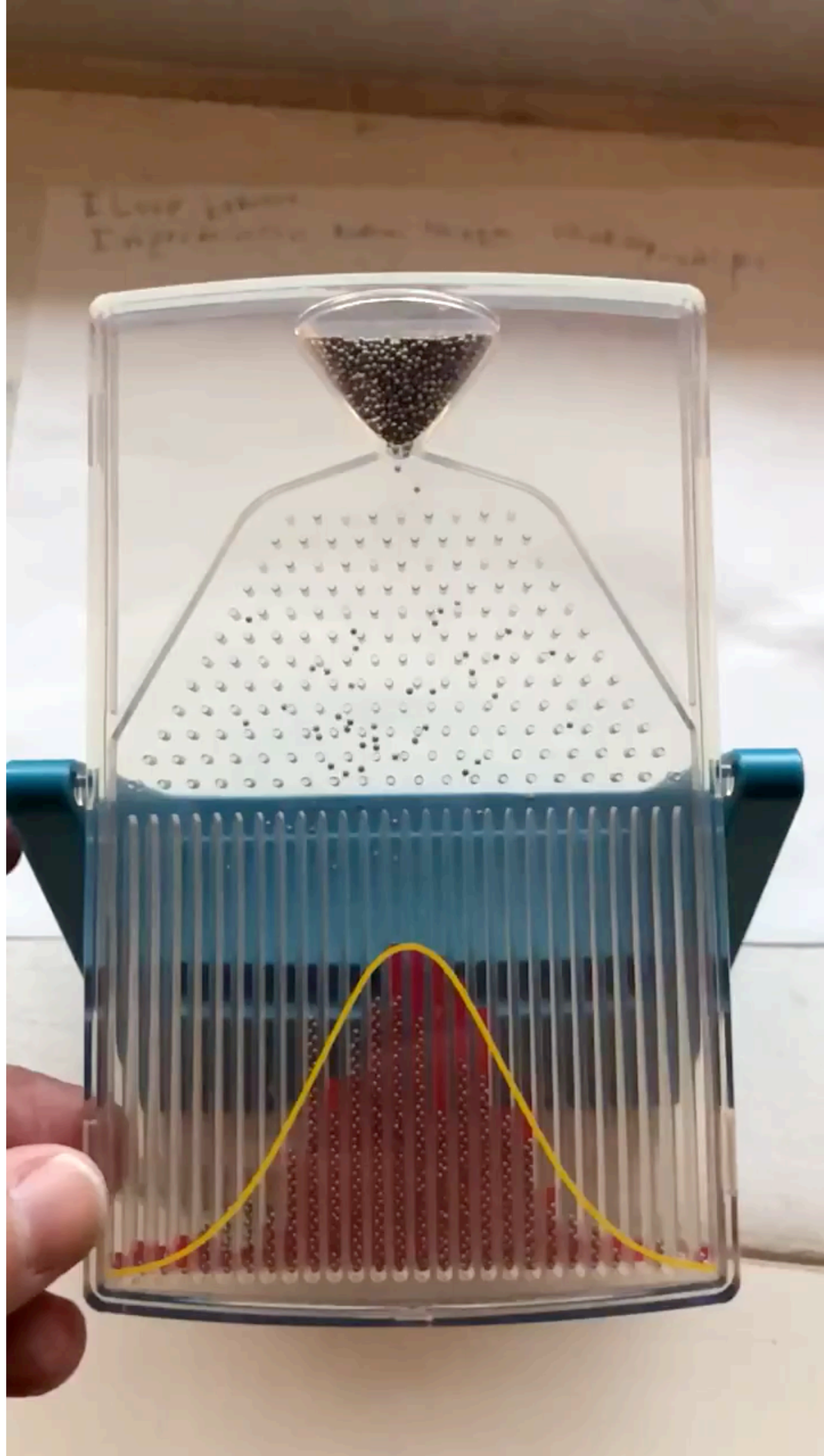
UH OH!

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!



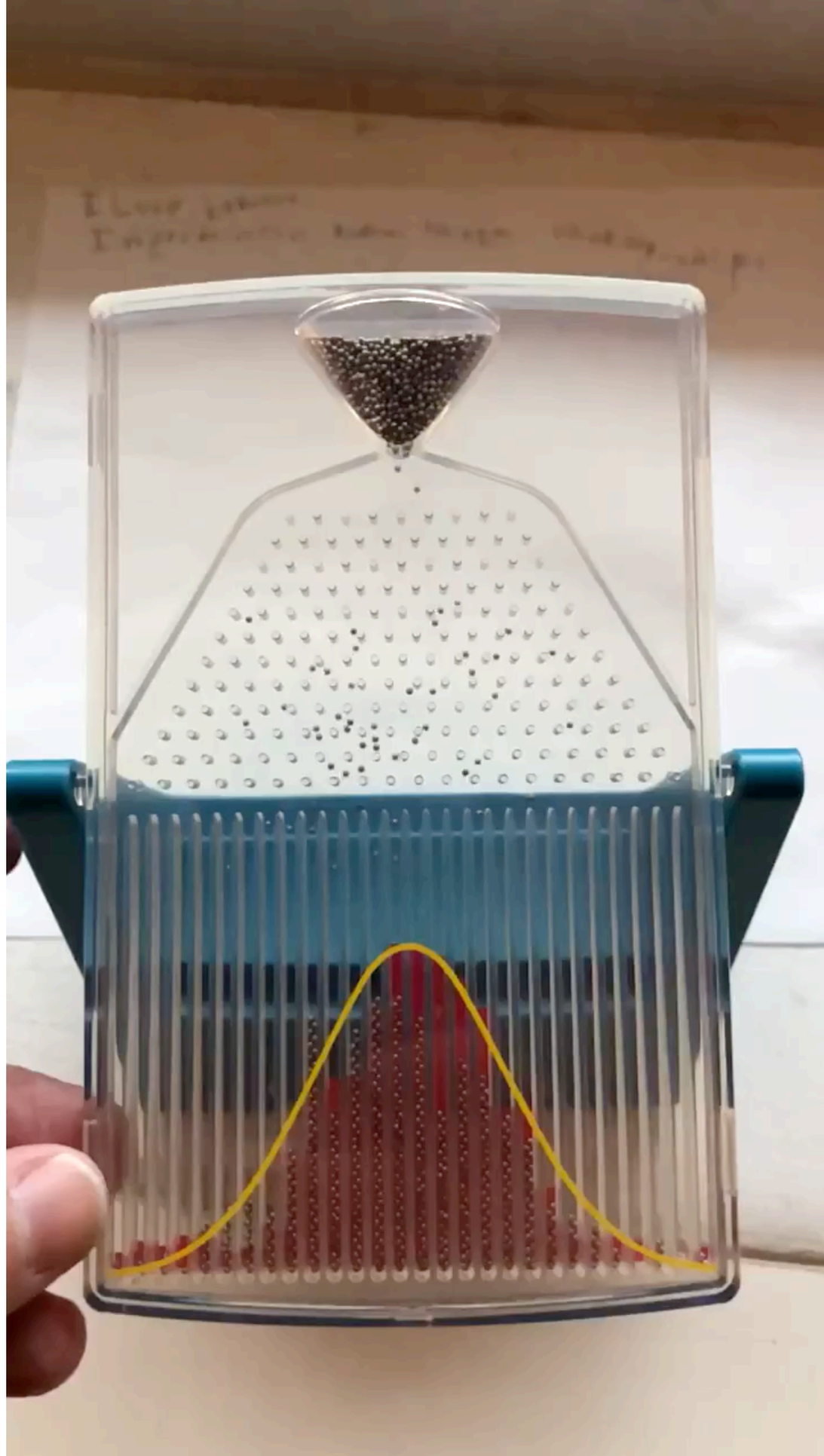
UH OH!

The actual situation is much more complicated.

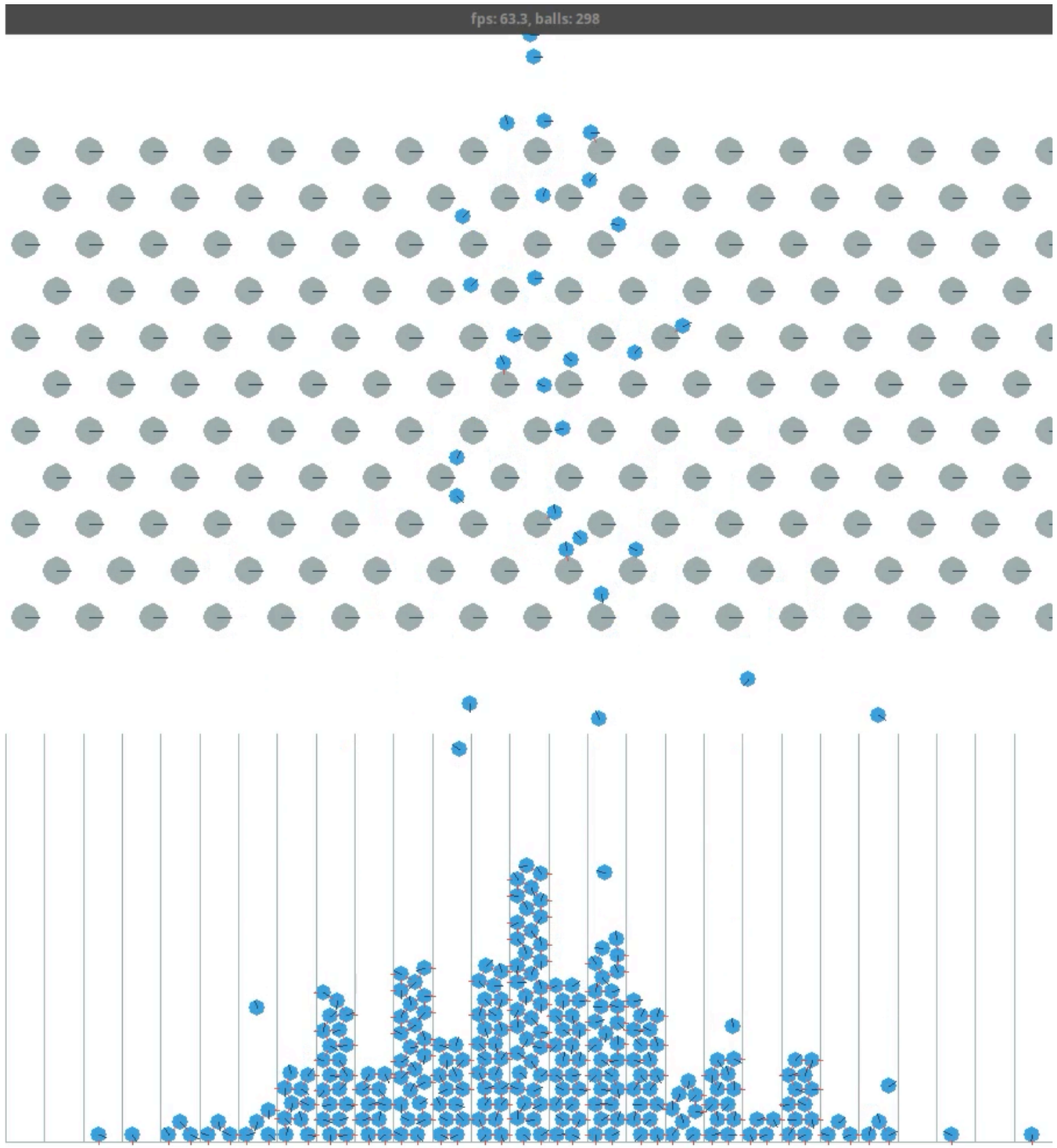
It's not a Binomial distribution!

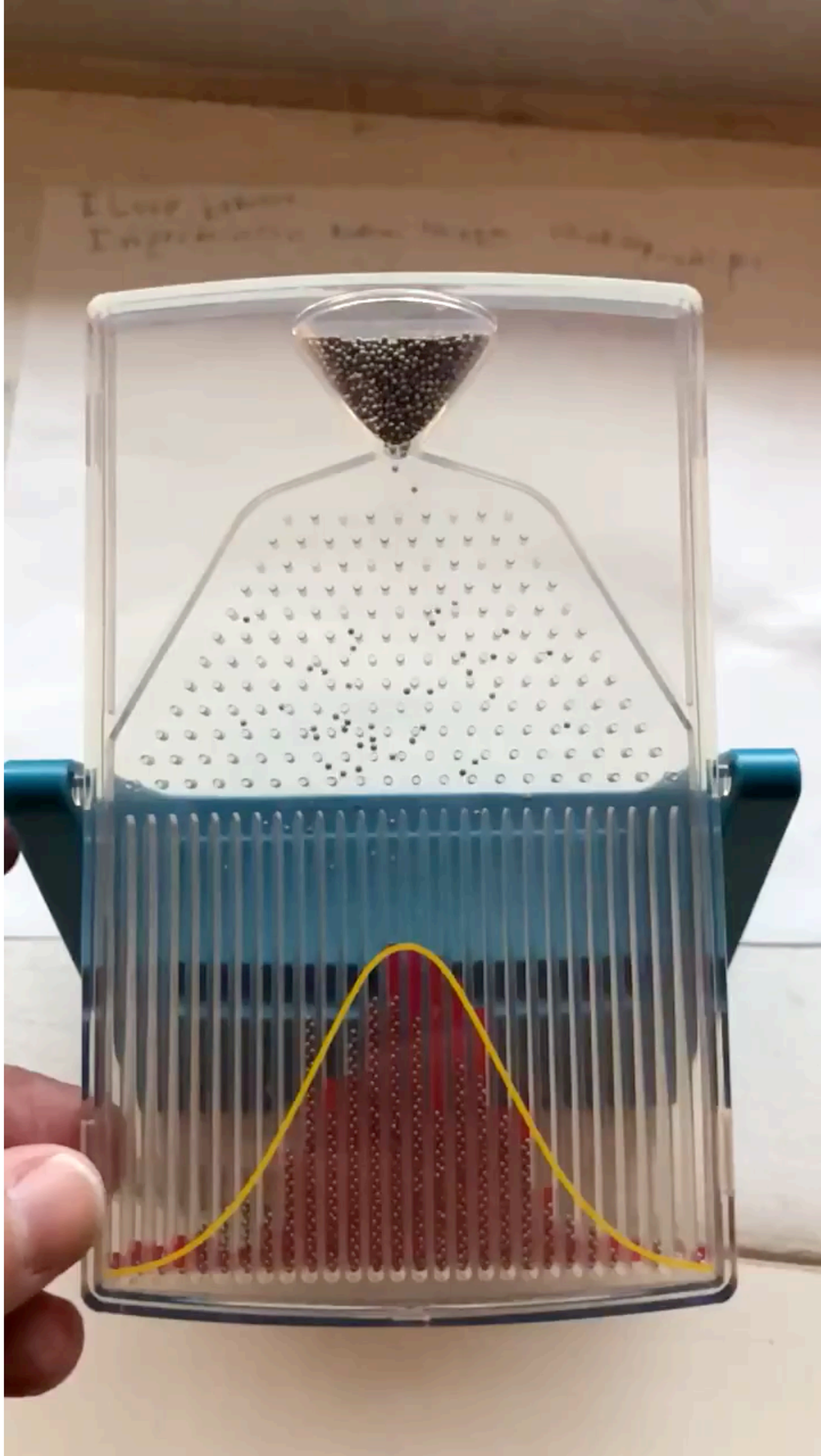
What is it?

I have no idea, but I could simulate it!



UH OH!



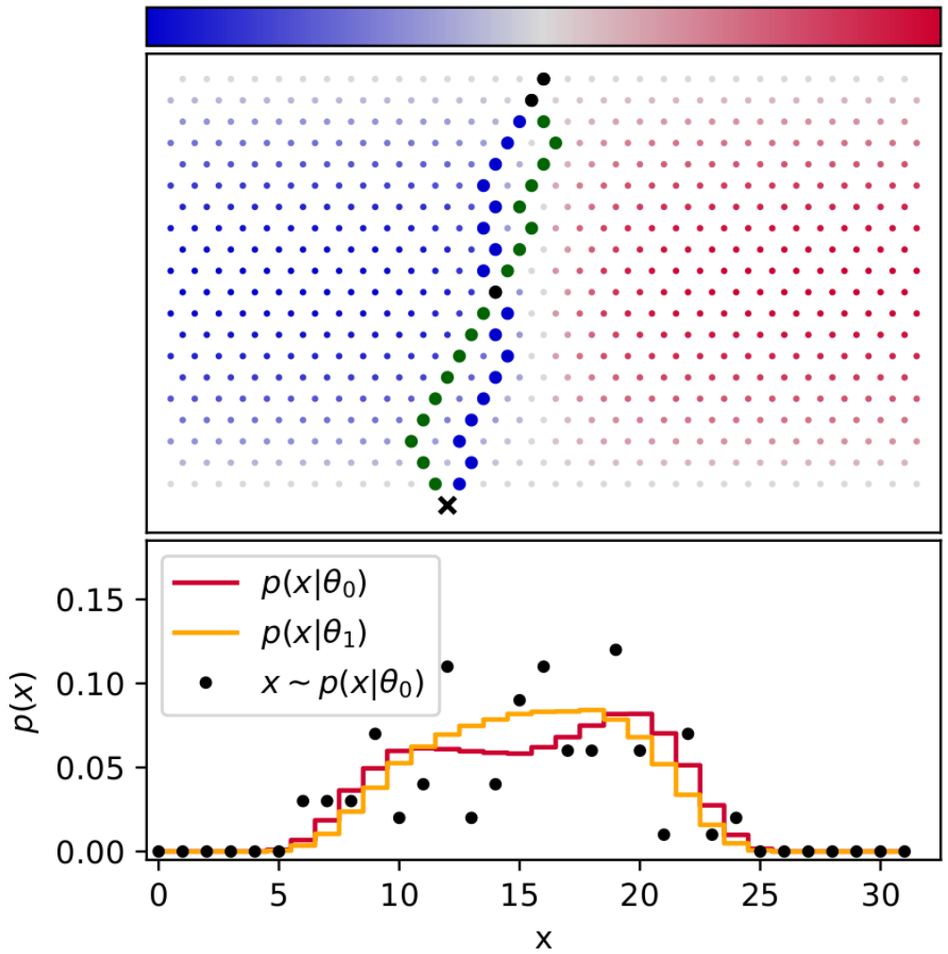


ANIMATION BY ATILIM GÜNEŞ BAYDIN

AN INTRACTABLE INTEGRAL

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

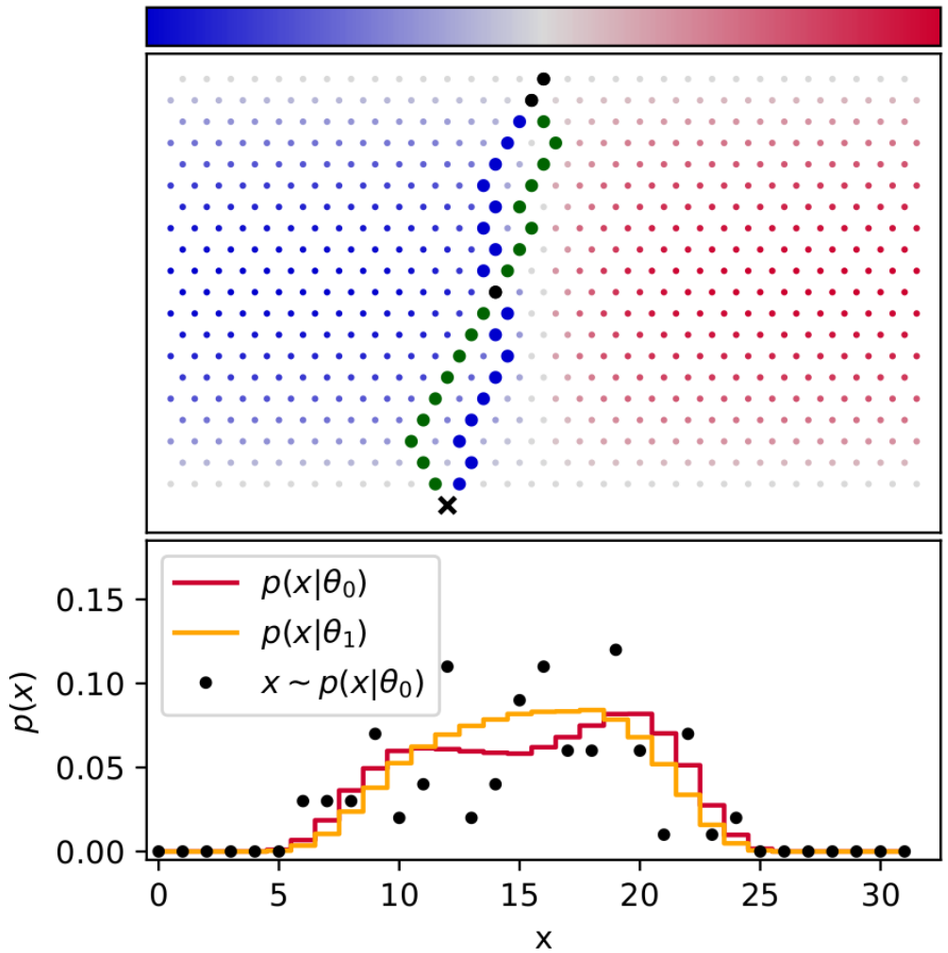
$$p(x|\theta) = \int p(x, z|\theta) dz$$



AN INTRACTABLE INTEGRAL

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

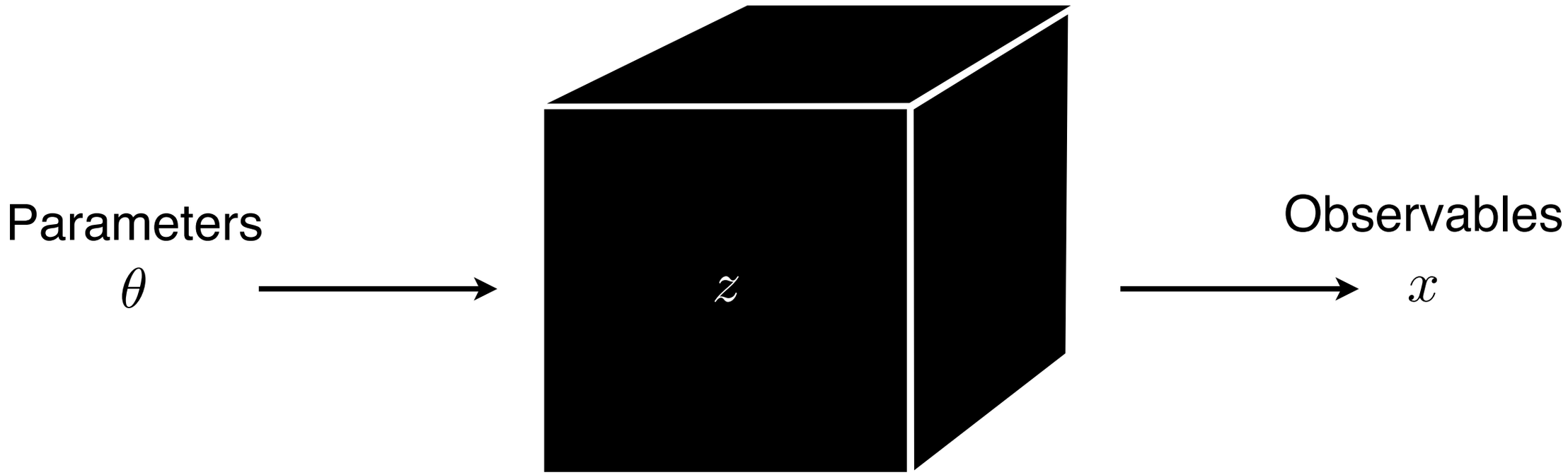
$$p(x|\theta) = \int p(x, z|\theta) dz$$



- 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



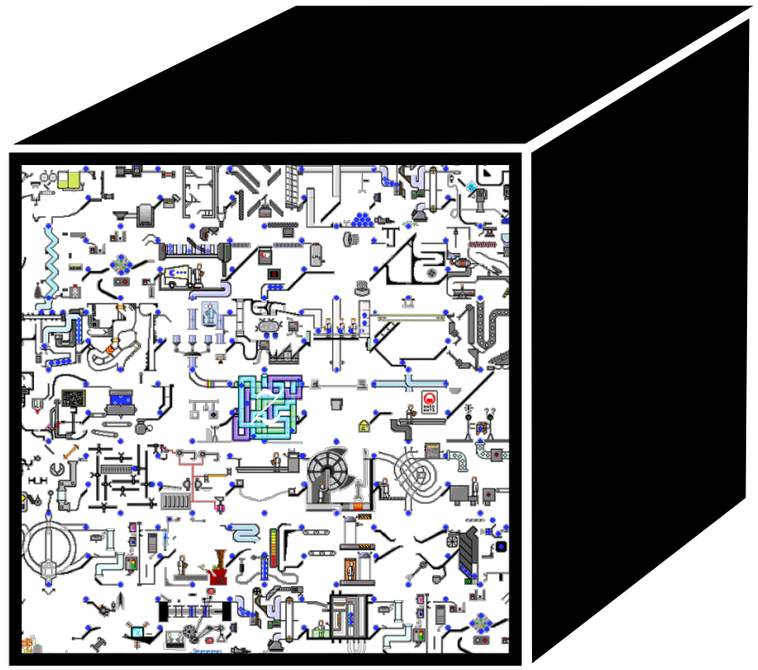
Prediction (simulation):

- Well-understood mechanistic model
- Simulator can generate samples

LIKELIHOOD-FREE INFERENCE

Parameters

θ



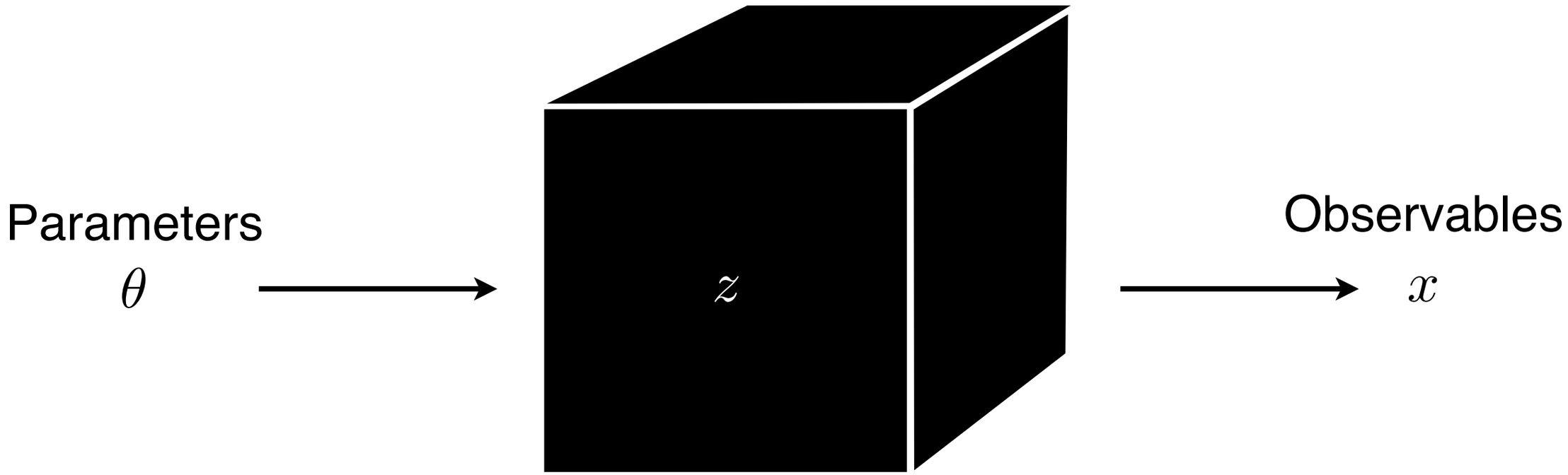
Observables

x

Prediction (simulation):

- Well-understood mechanistic model
- Simulator can generate samples

LIKELIHOOD-FREE INFERENCE



Prediction (simulation):

- Well-understood mechanistic model
- Simulator can generate samples

Inference:

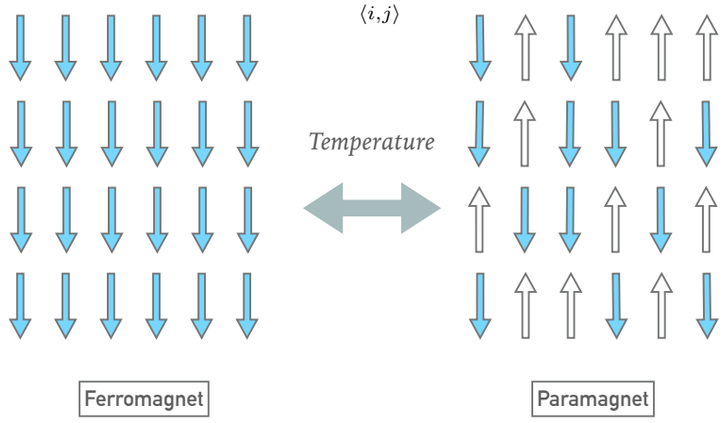
- Likelihood function $p(x|\theta)$ is intractable
- Goal: estimator $\hat{p}(x|\theta)$

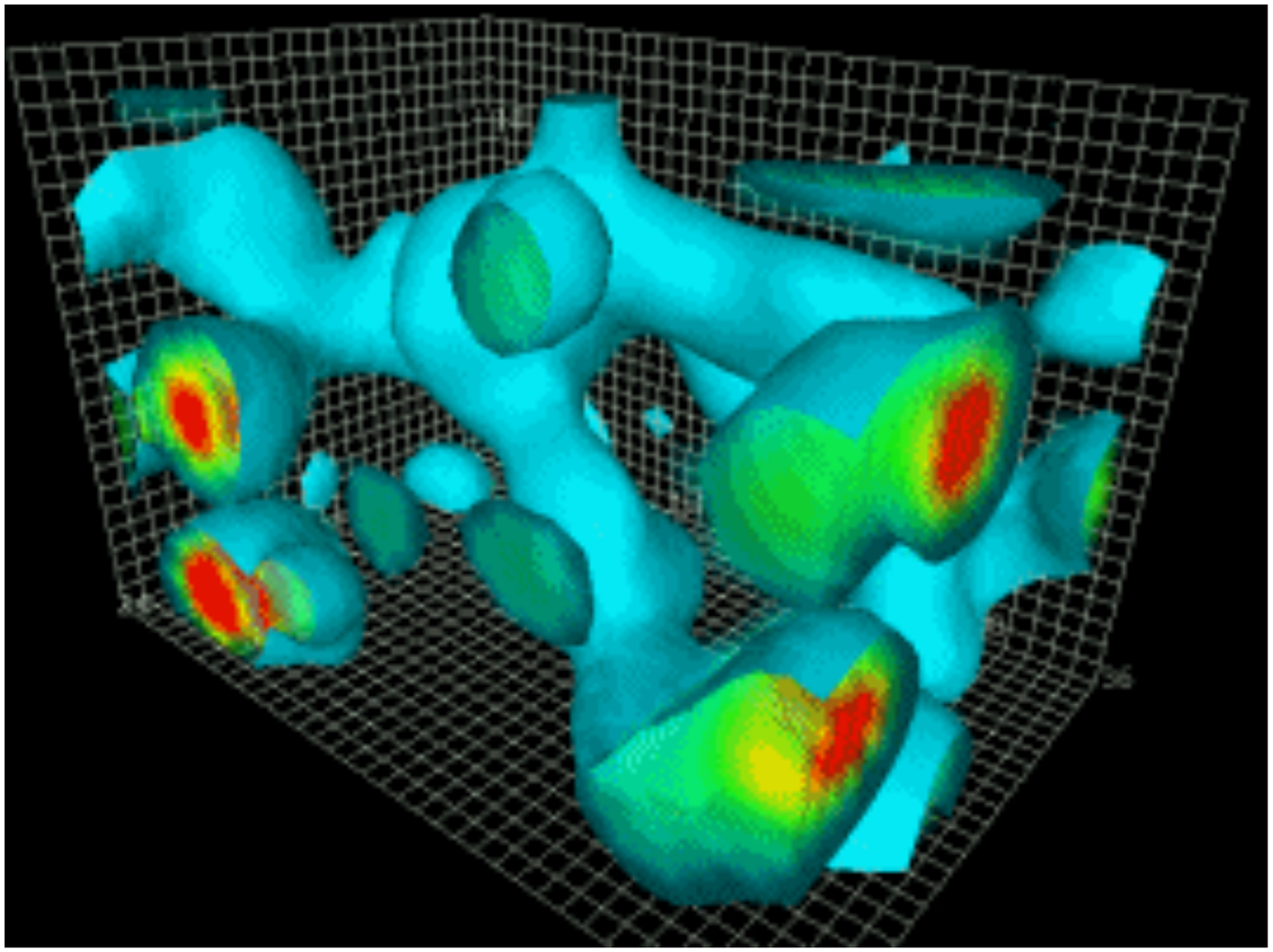
LATTICE FIELD THEORY

PHASES, PHASE TRANSITIONS, AND THE ORDER PARAMETER

Ising ferromagnet in two dimensions

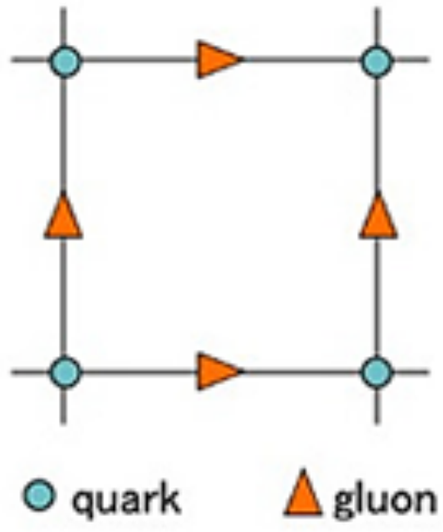
$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j$$

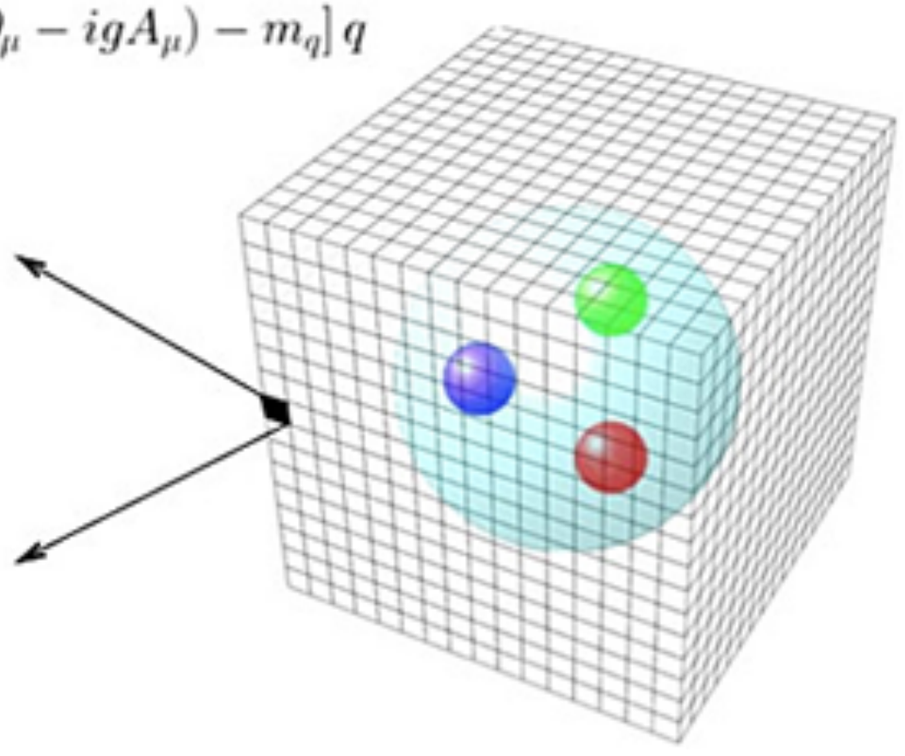




QCD Lagrangian

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



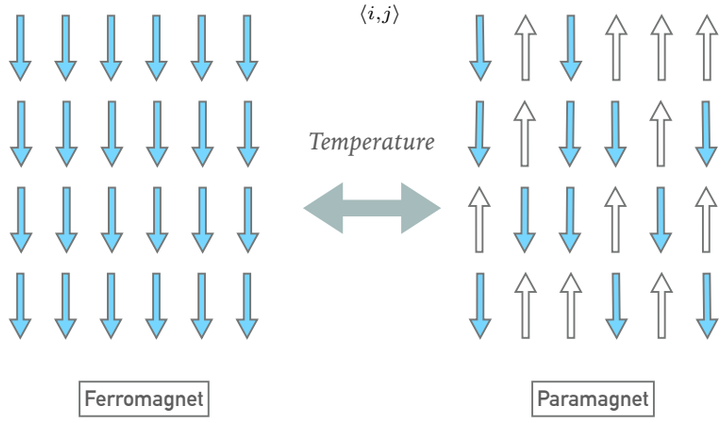


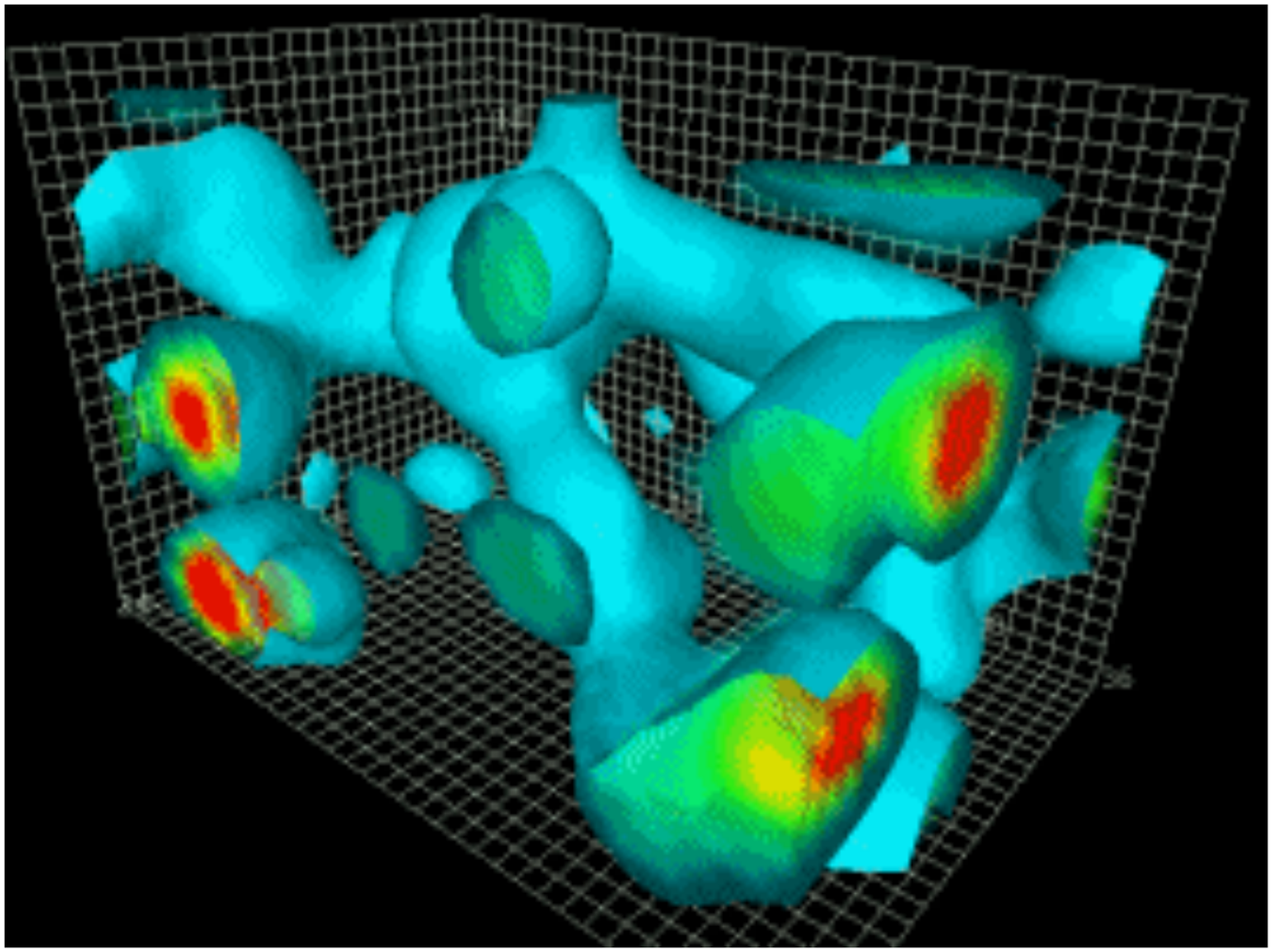
LATTICE FIELD THEORY

PHASES, PHASE TRANSITIONS, AND THE ORDER PARAMETER

Ising ferromagnet in two dimensions

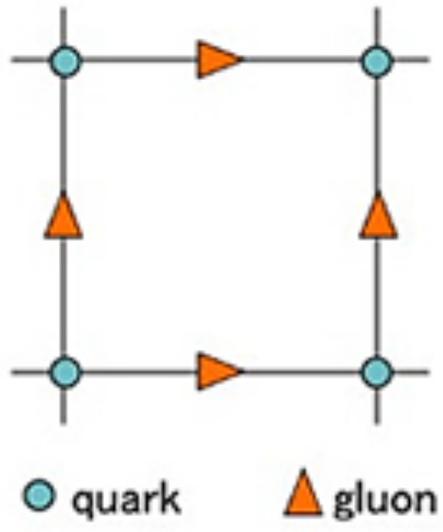
$$E = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j$$

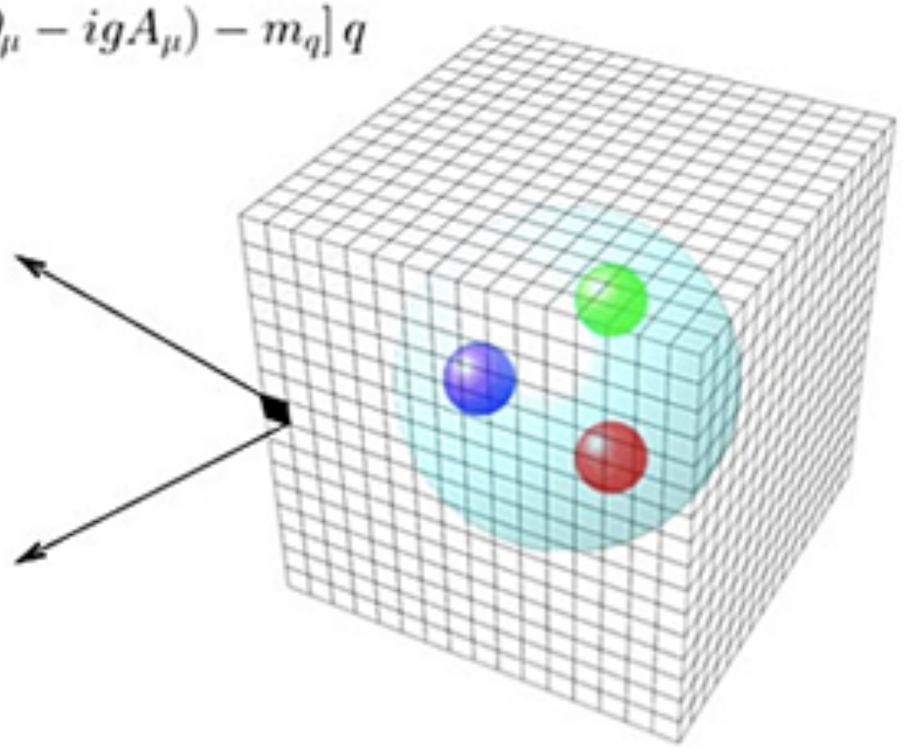




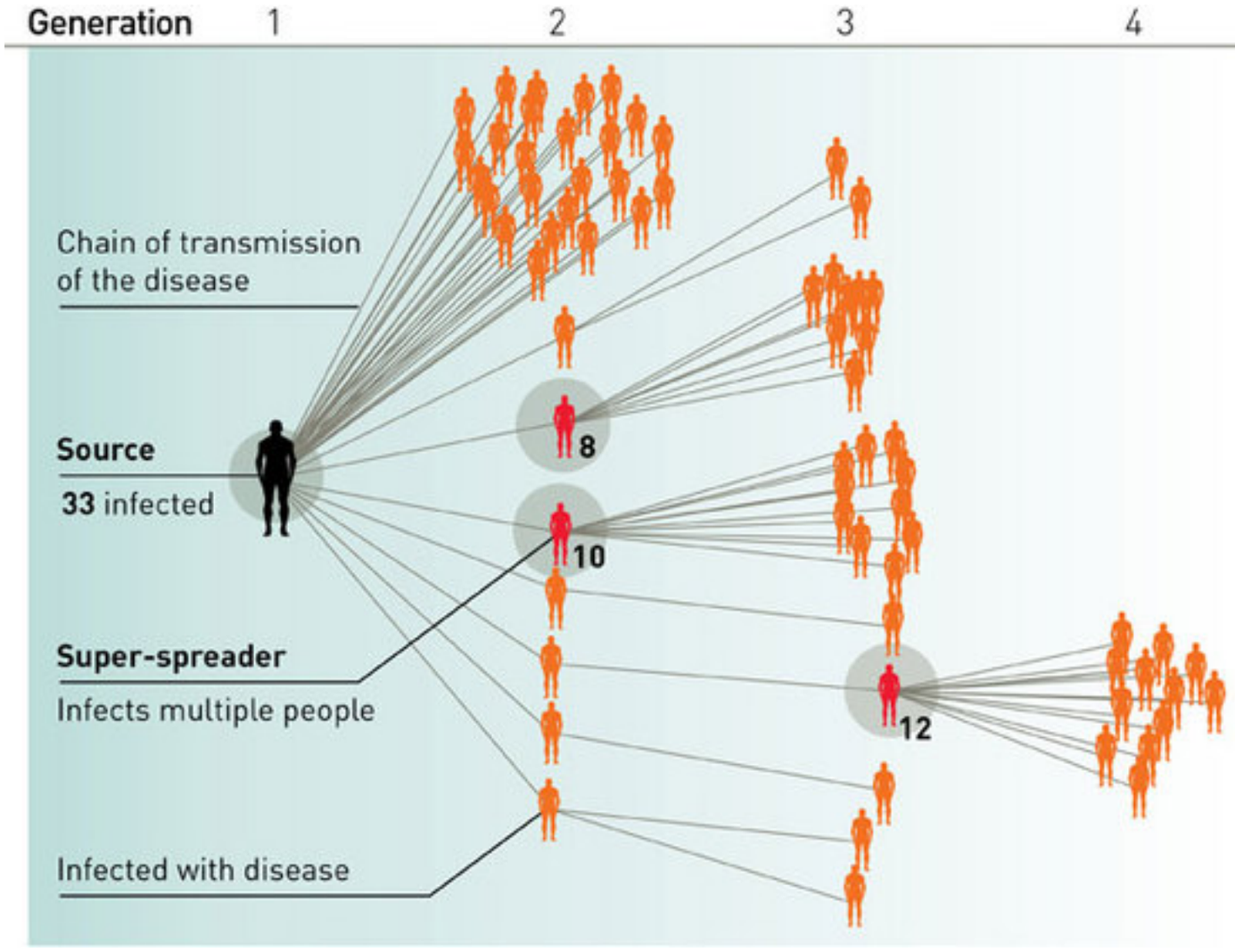
QCD Lagrangian

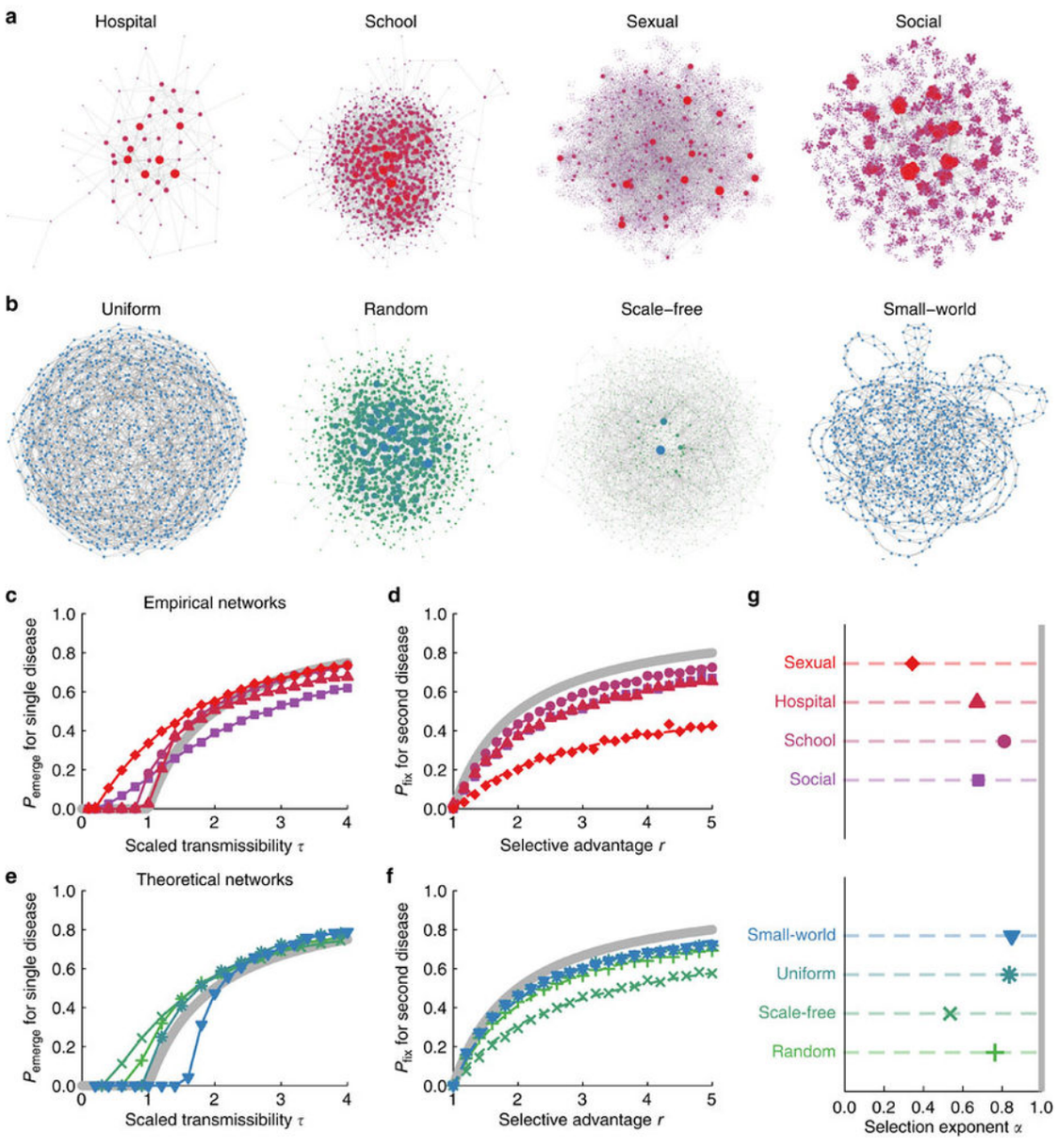
$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} + \sum_{q=u,d,s,c,b,t} \bar{q} \left[i\gamma^\mu (\partial_\mu - igA_\mu) - m_q \right] q$$

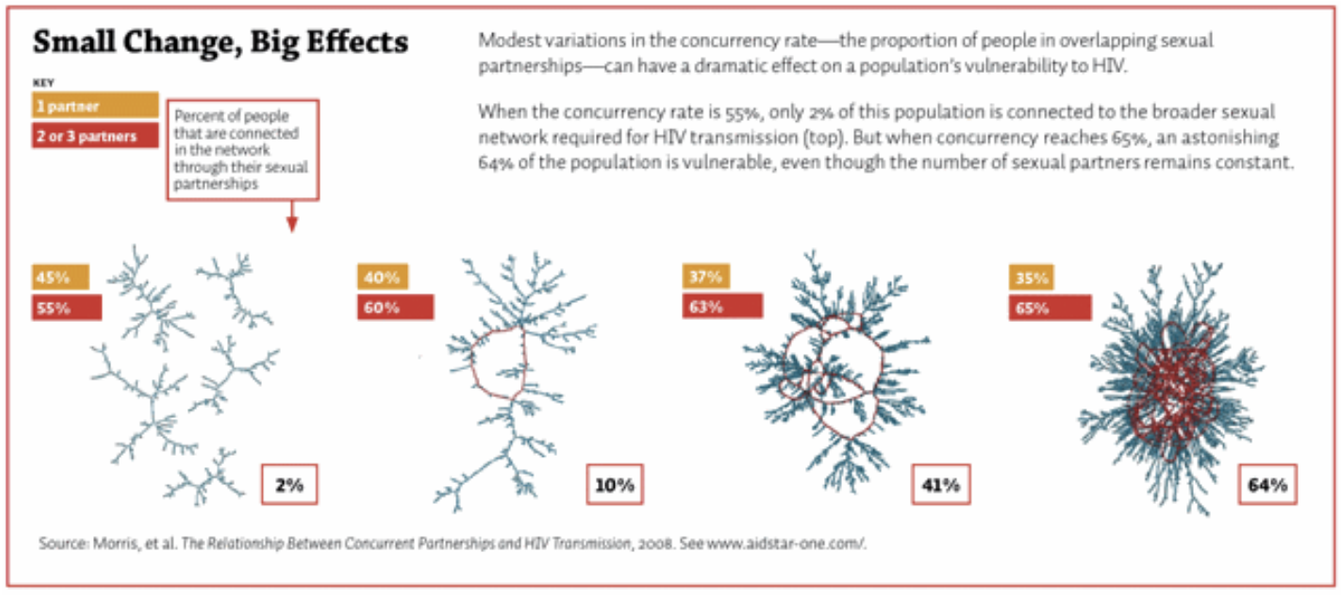




EPIDEMIOLOGY & POPULATION GENETICS







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 θ, 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

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

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

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε, 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 $S = (S_1, \ldots, S_p)$ is sufficient for θ, in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ, 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 identity 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 θ. With these statistics in hand, we have the following approximate Bayesian computation scheme for data $\mathcal{D}$ summarized by S:

D1. Generate θ from $\pi(\cdot)$.
D2. Simulate $\mathcal{D}'$ from stochastic model $\mathcal{M}$ with parameter θ, and compute the corresponding statistics S'.
D3. Calculate the distance $\rho(S, S')$ between S and S'.
D4. Accept θ 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 Baves factors that can be used for model com-

ABC

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 θ from $\pi(\cdot)$.
D2. Simulate $\mathcal{D}'$ from stochastic model $\mathcal{M}$ with parameter θ, and compute the corresponding statistics S'.
D3. Calculate the distance $\rho(S, S')$ between S and S'.
D4. Accept θ 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 θ, 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 θ from $\pi(\cdot)$.
A2. Accept θ 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 identity 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 θ. With these statistics in hand, we have the following approximate Bayesian computation scheme for data $\mathcal{D}$ summarized by S:

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

ABC

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 θ, 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

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

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε, 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 $S = (S_1, \ldots, S_p)$ is sufficient for θ, in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ, 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 identity 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 θ. 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

θ, and

hods,

, they

generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Baves 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

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 θ, 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

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

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε, 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 $S = (S_1, \ldots, S_p)$ is sufficient for θ, in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ, 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 identity 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 θ. 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 identity a suitable set of sufficient statistics, and we then might resort to a more heuristic approach.

θ, and

hods,
, they
generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Baves 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.

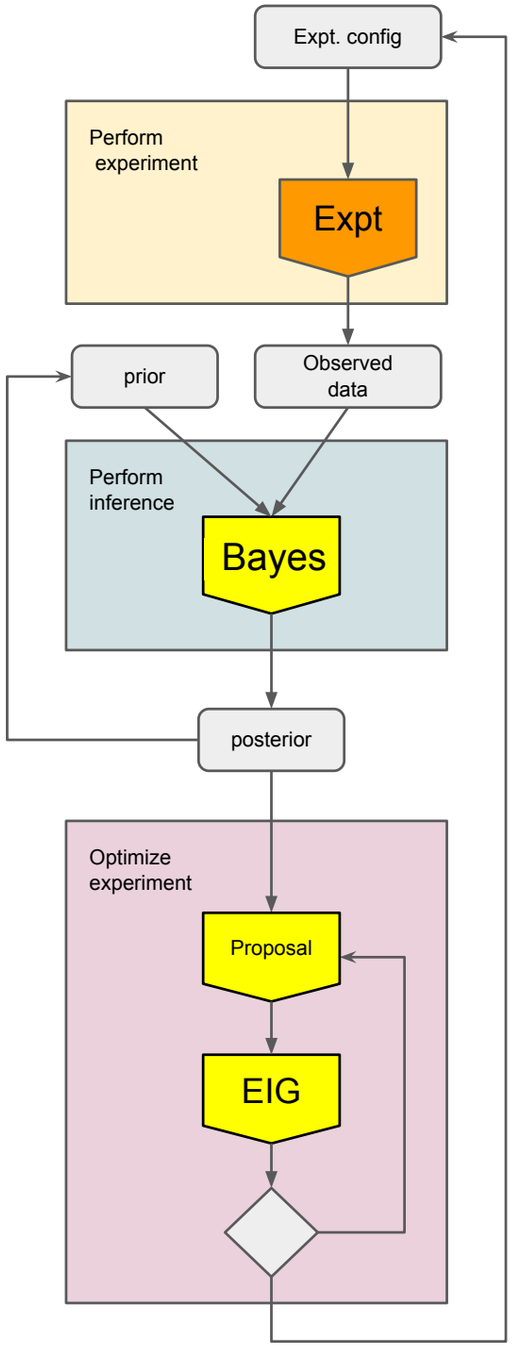
Implicit models have many applications. They are used in ecology where models simulate animal populations over time; they are used in phylogeny, where simulations produce hypothetical ancestry trees; they are used in physics to generate particle simulations for high energy processes. Recently, implicit 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.

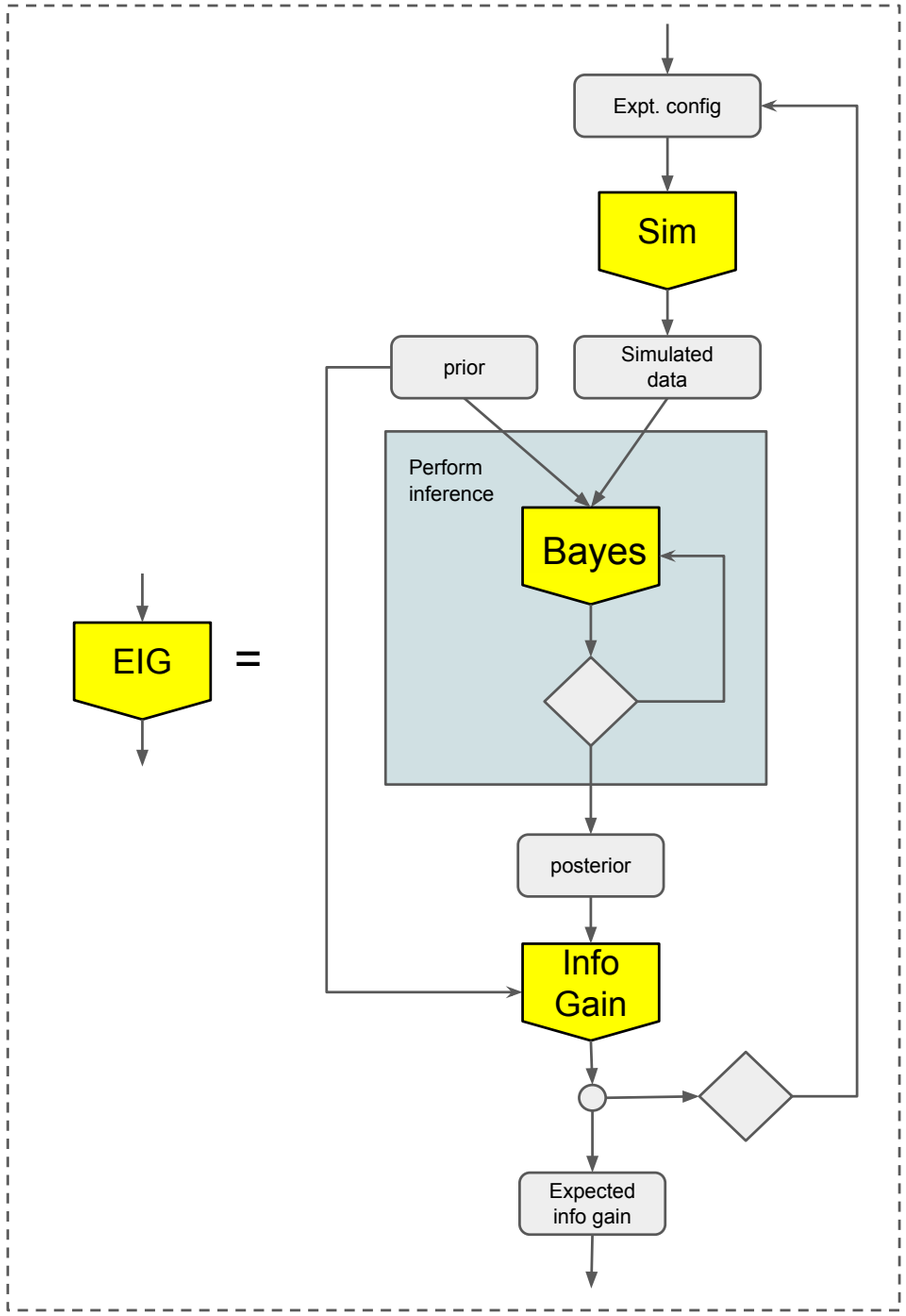
Of particular interest at this workshop is to unite fields that work on implicit models. For example:

- **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.

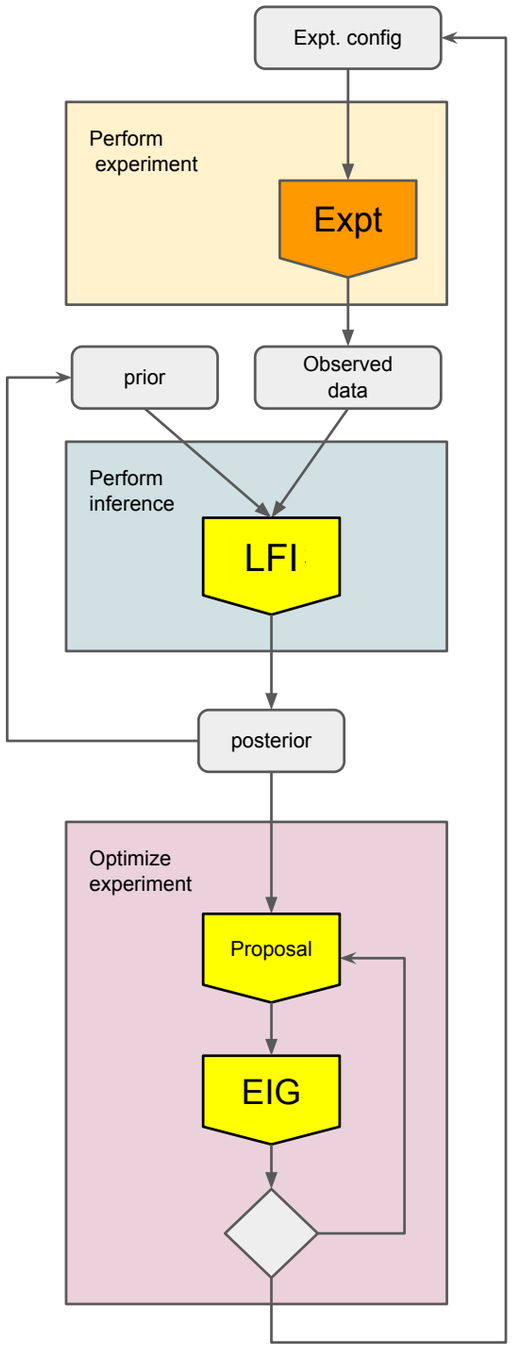
"ACTIVE SCIENCING"

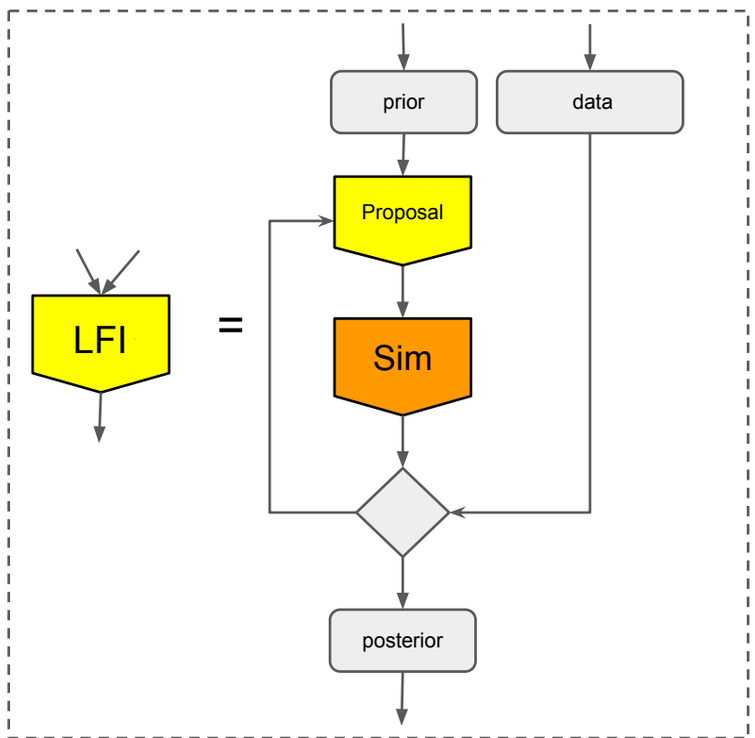


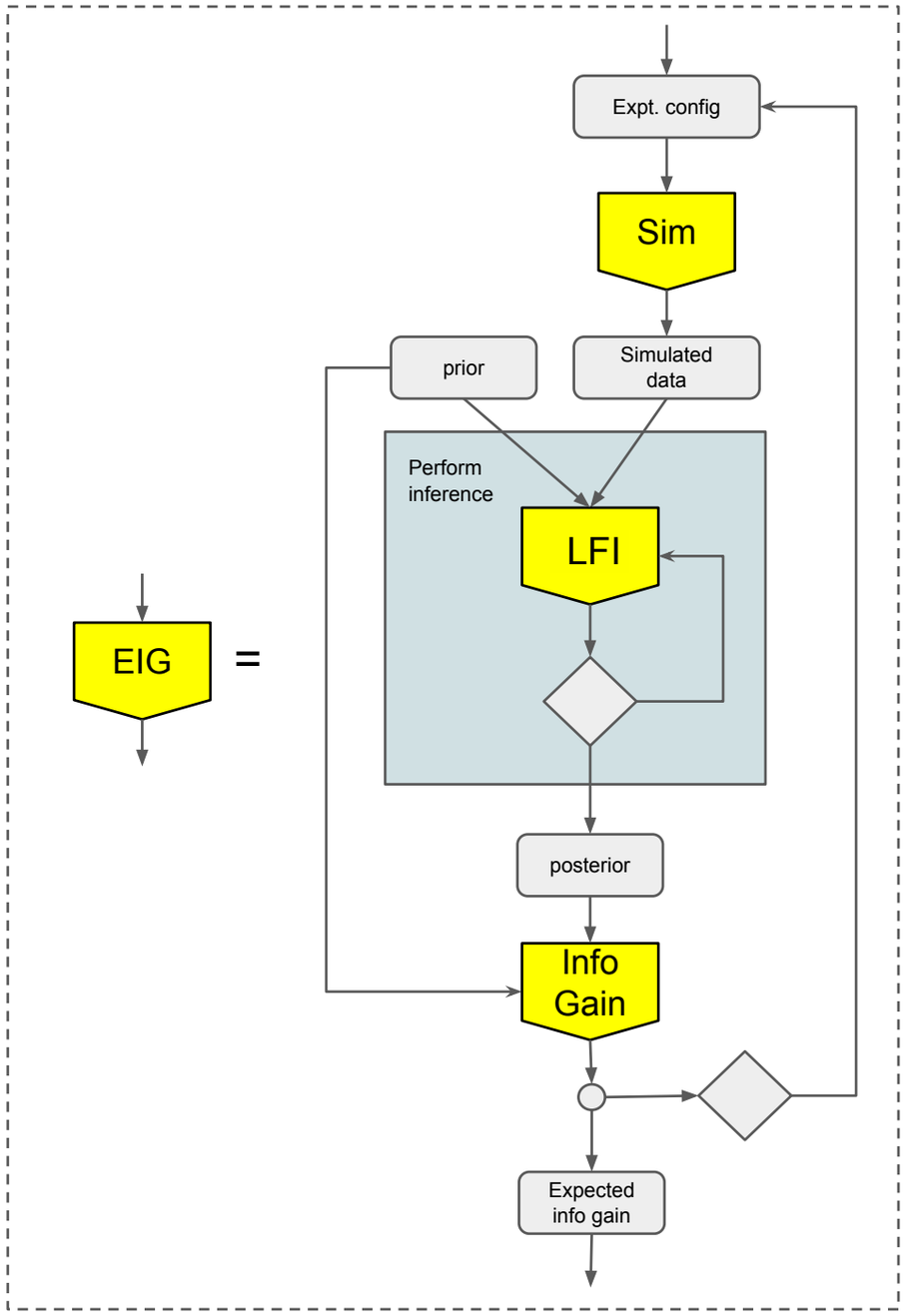


https://github.com/cranmer/active_sciencing

"ACTIVE SCIENCING"







https://github.com/cranmer/active_sciencing

active learning / sequential design / black box optimization



Active Sciencing

reusable workflows



simulation-based /
likelihood-free
inference engines

http://reanahub.io

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.

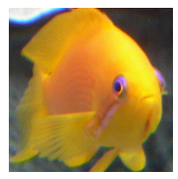
CERN

NSF

diana hep

iris hep
Institute for Research and Innovation in Software for High Energy Physics

The SCAILFIN Project
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`.

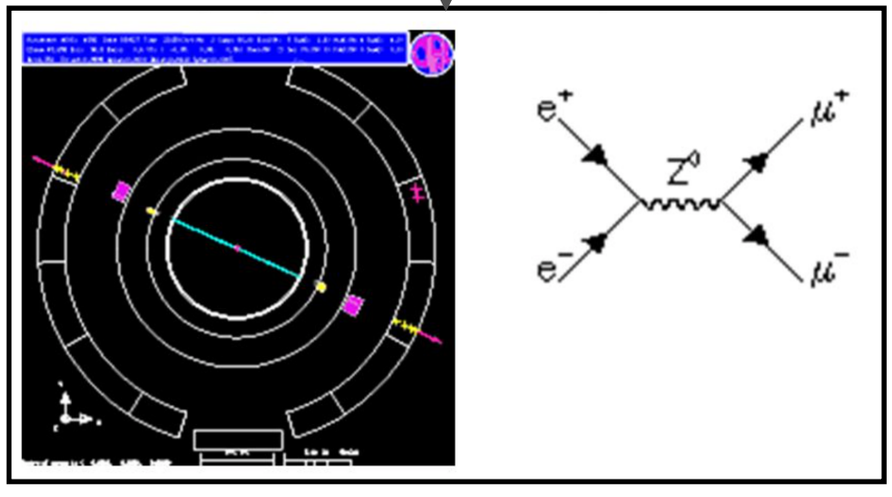
launch binder

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



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

$$A^f_{FB}(s) \simeq A^f_{FB}(m_Z^2) + \frac{(s-m_Z^2)}{s}\,\frac{3\pi\alpha(s)}{\sqrt{2}\mathrm{G_F}m_Z^2}\,\frac{2Q_eQ_fg_{Ae}g_{Af}}{(g_{Ve}^2+g_{Ae}^2)(g_{Vf}^2+g_{Af}^2)}\ . \quad (8.30)$$

ENCAPSULATING THE SIMULATION

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

README.md

Run HEP workflows from the web.

by Kyle Cranmer and Lukas Heinrich

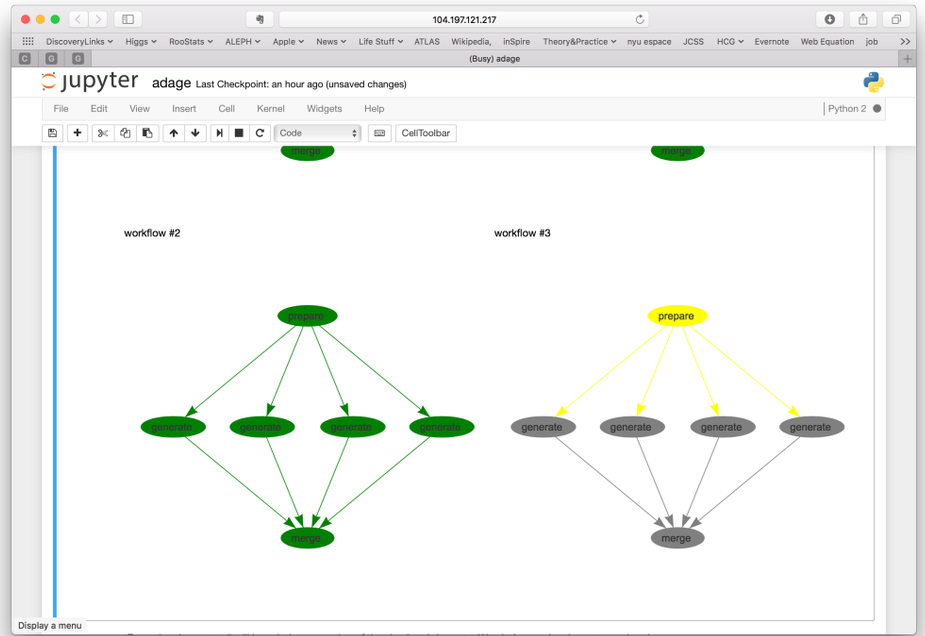
An example notebook on how to generate simulated high energy physics collision events using the generator package MadGraph. Simulated datasets obtained from this notebook can then be used to train and evaluate the performance of generative models for physics.

Usage:

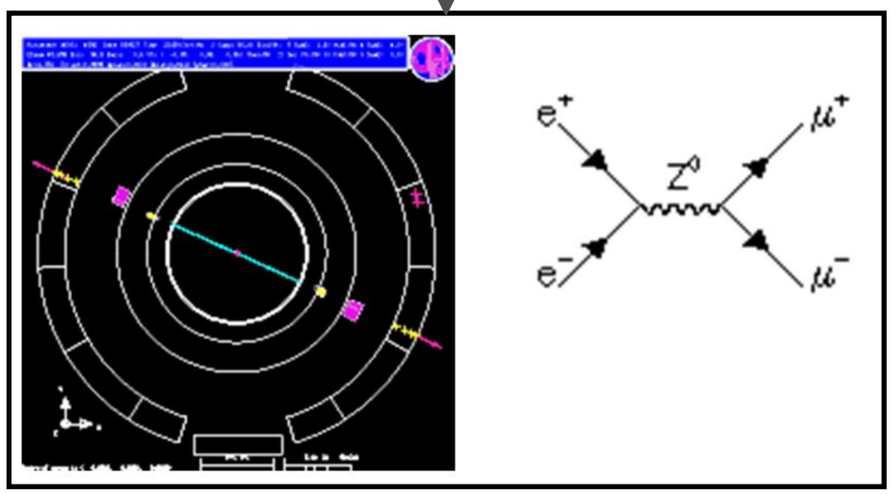
This repository has been equipped with a Dockerfile to encapsulate its software environment. It can be used with the mybinder service to launch an ephemeral jupyter notebook server to run the notebook.

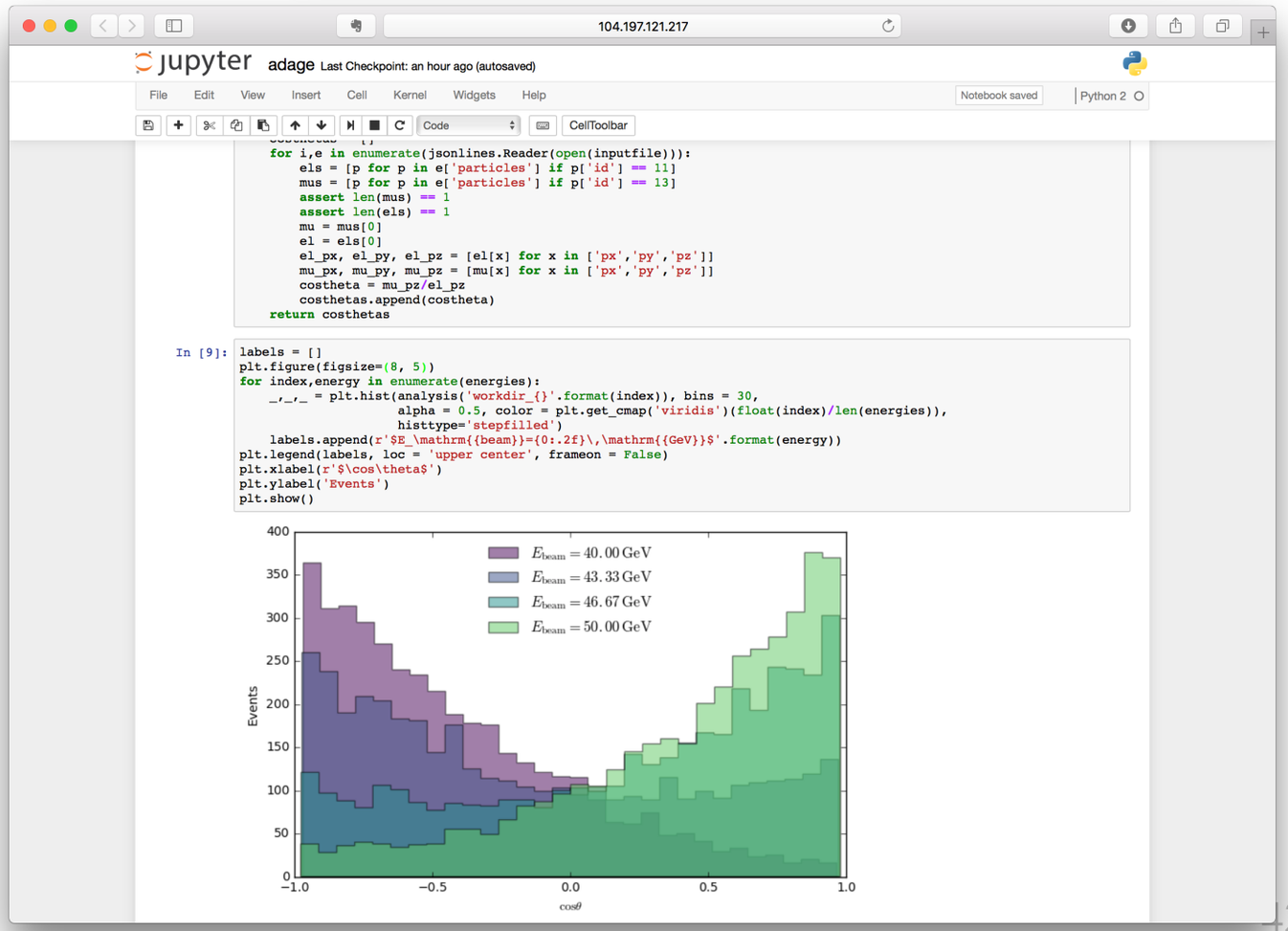
Click on the below badge and open the notebook `adage.ipynb`.

launch binder



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





SYNTHESIS

active learning / sequential design / black box optimization



Active Sciencing



reusable workflows



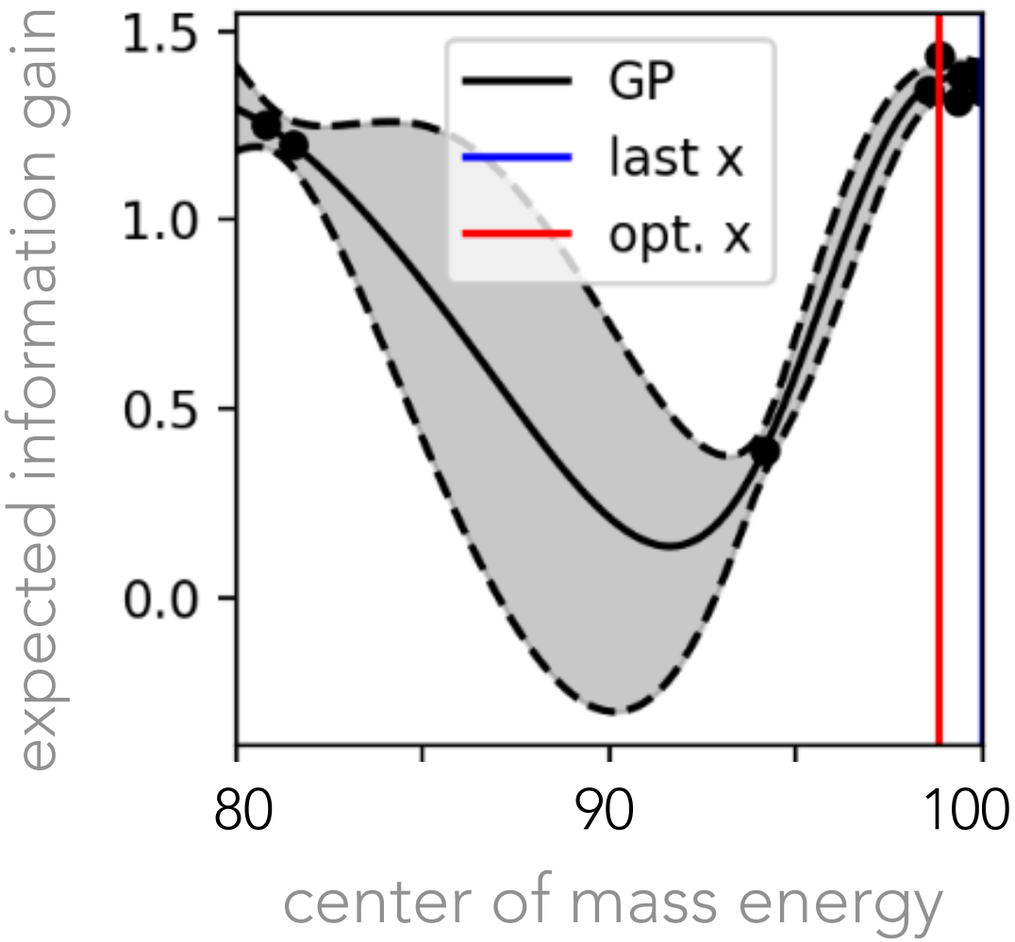
simulation-based / likelihood-free inference engines

https://github.com/cranmer/active_sciencing

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



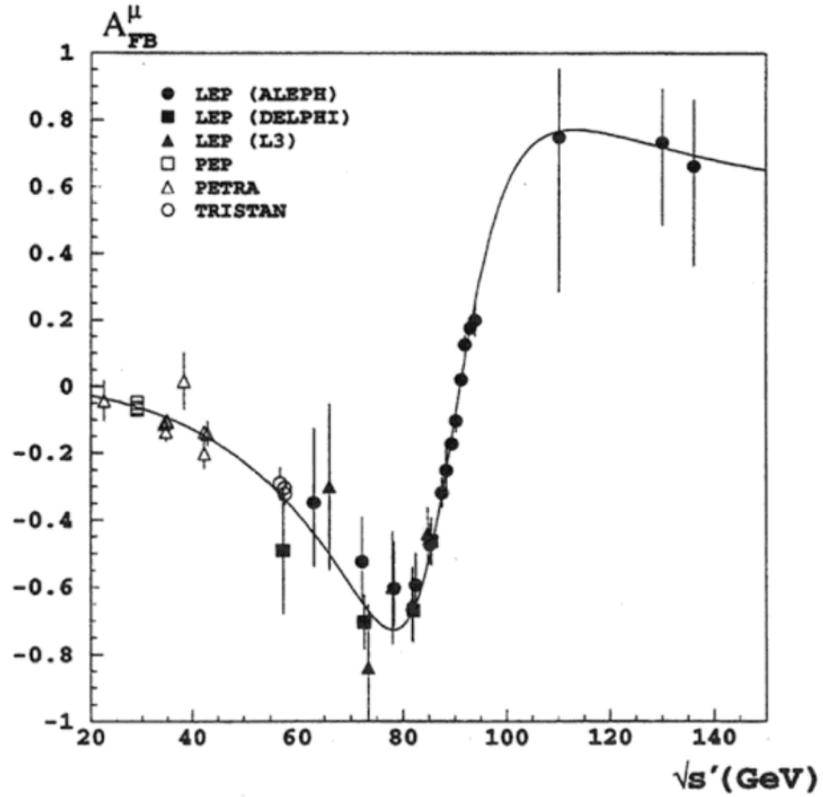
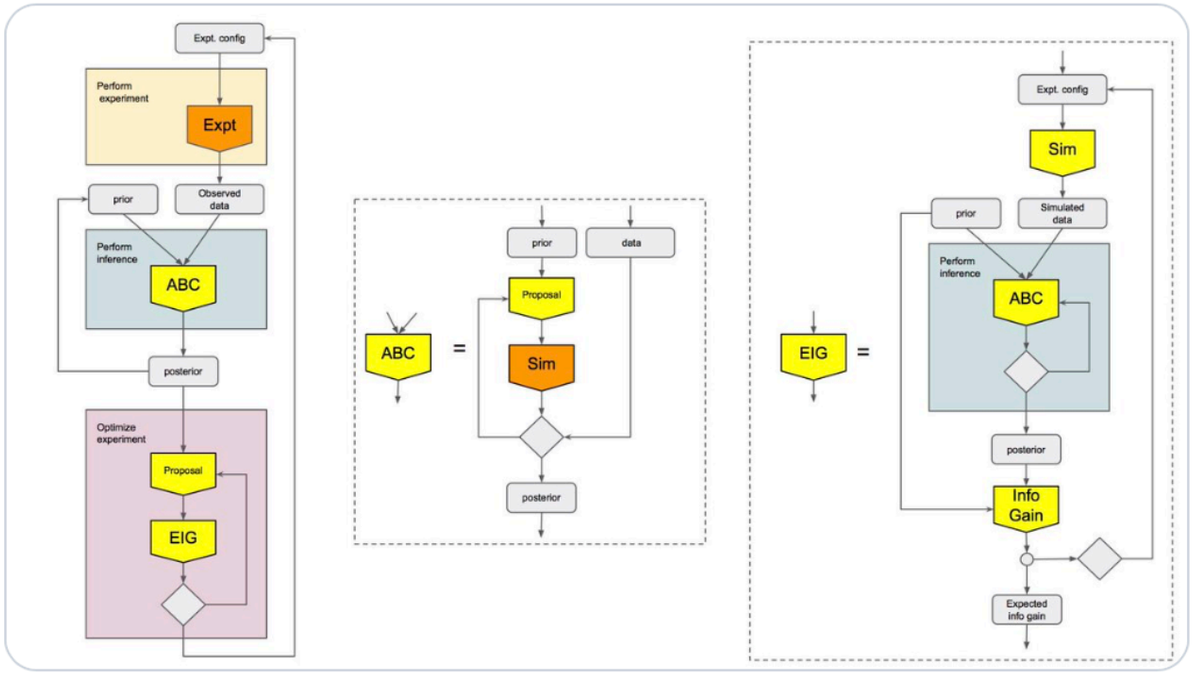


Figure 2: Measured forward-backward asymmetries of muon-pair production compared with the model independent fit results.



Kyle Cranmer @KyleCranmer · Jun 11, 2017

Demo for YComb research
active learning + workflows + implicit models = #ActiveSciencing
@lukasheinrich_ @glouppe
github.com/cranmer/active...



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

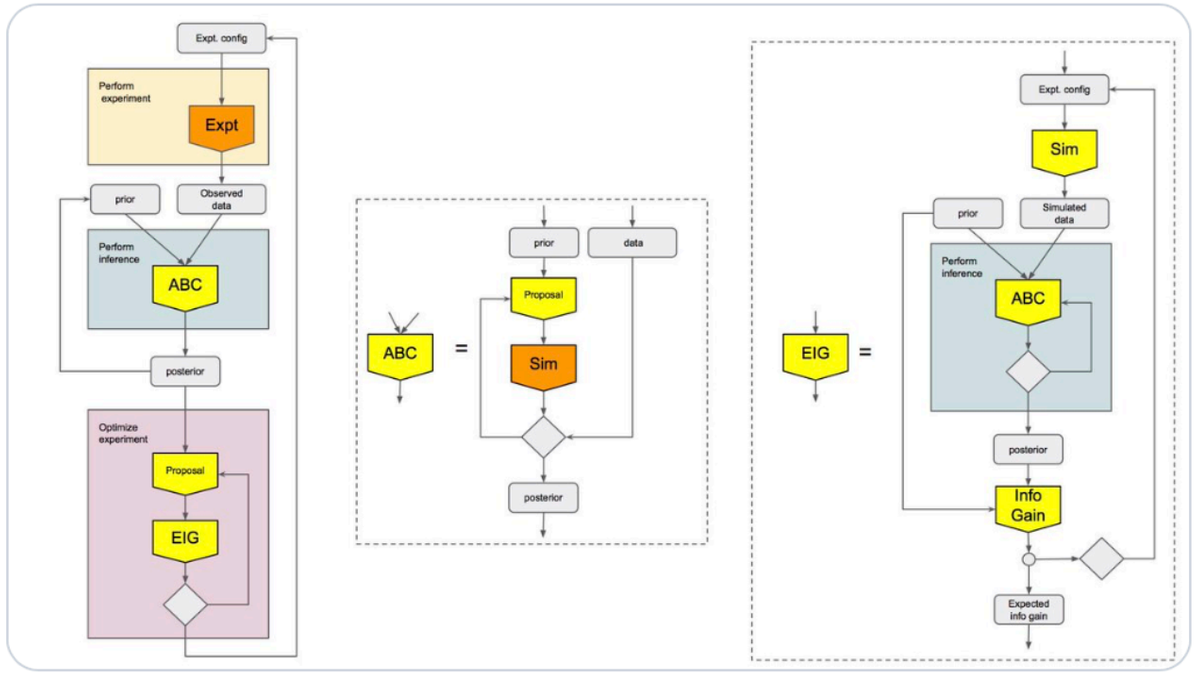
Replying to @KyleCranmer @lukasheinrich_ and @glouppe

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...



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...

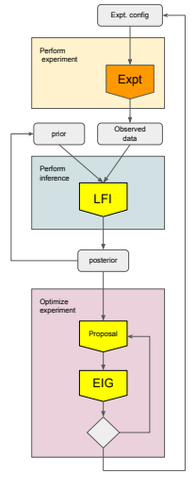
Keep in mind that
- the simulator model was specified
- the space of experimental configurations was well specified

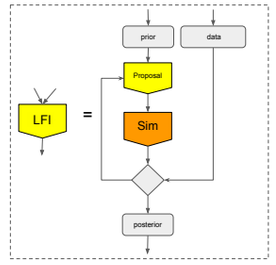
Still it was hard enough!

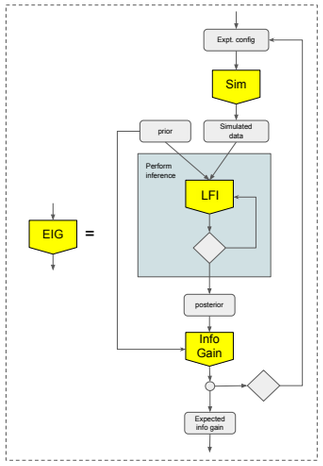
Going to open world of experimental configurations and potential models much harder.

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:

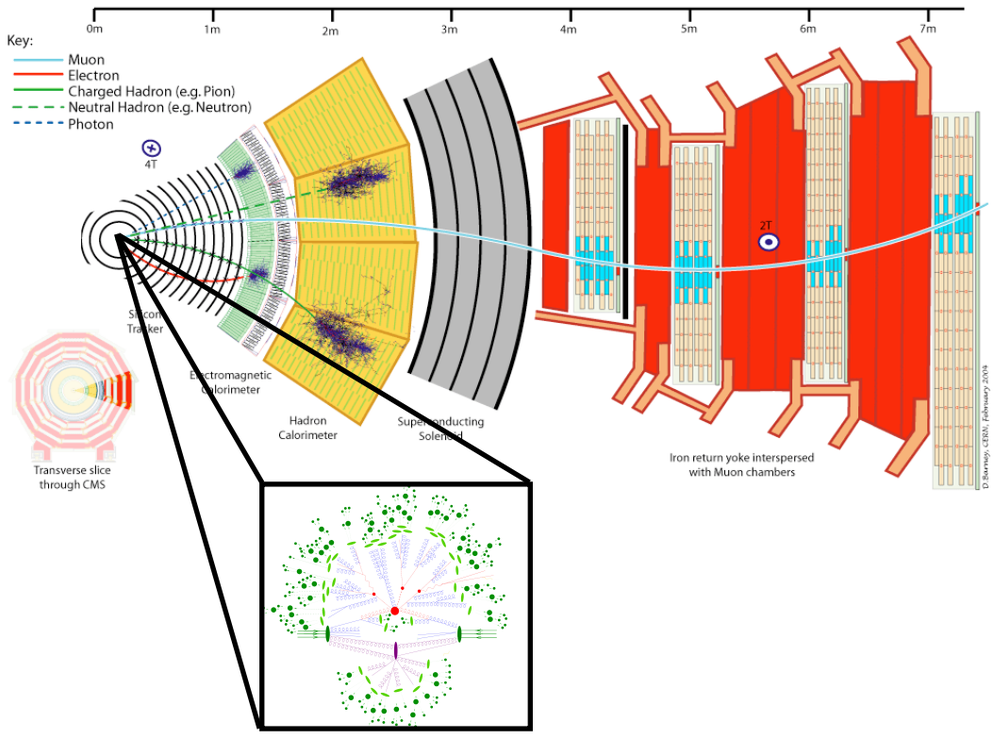
  - likelihood, posterior, utility (EIG) surface, acquisition, …

https://github.com/cranmer/active_sciencing

APPROACHES TO LIKELIHOOD-FREE INFERENCE

Use simulator

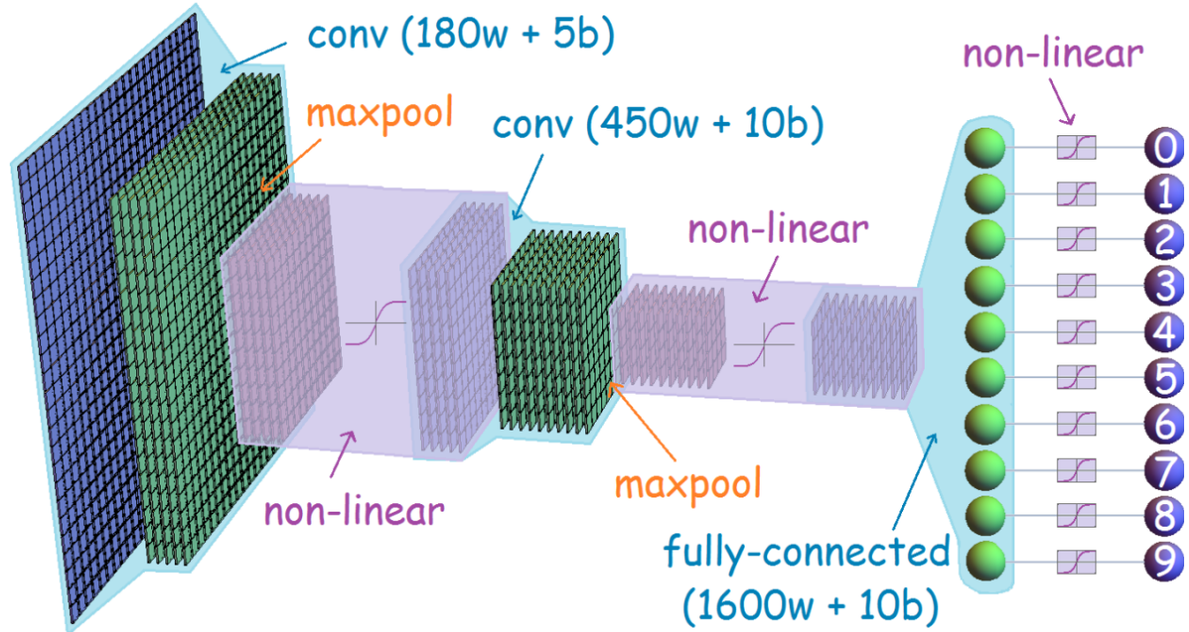
(much more efficiently)



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

Learn simulator

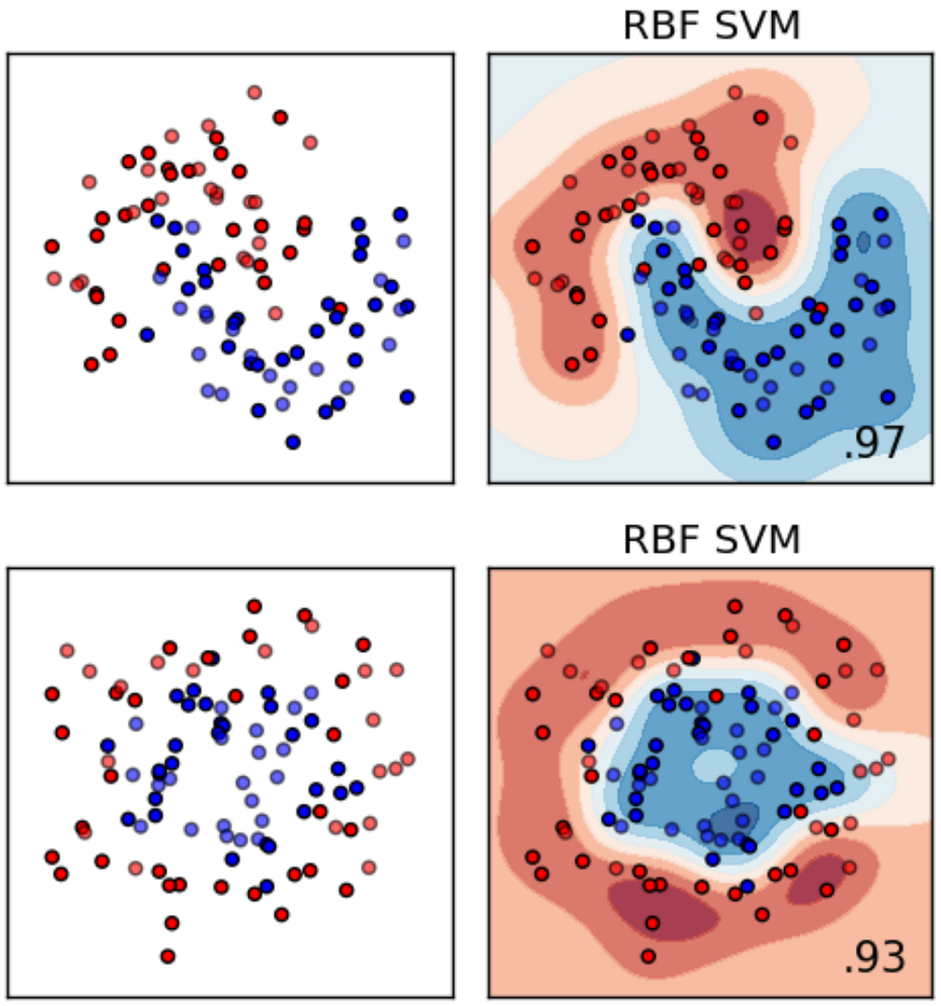
(with deep learning)

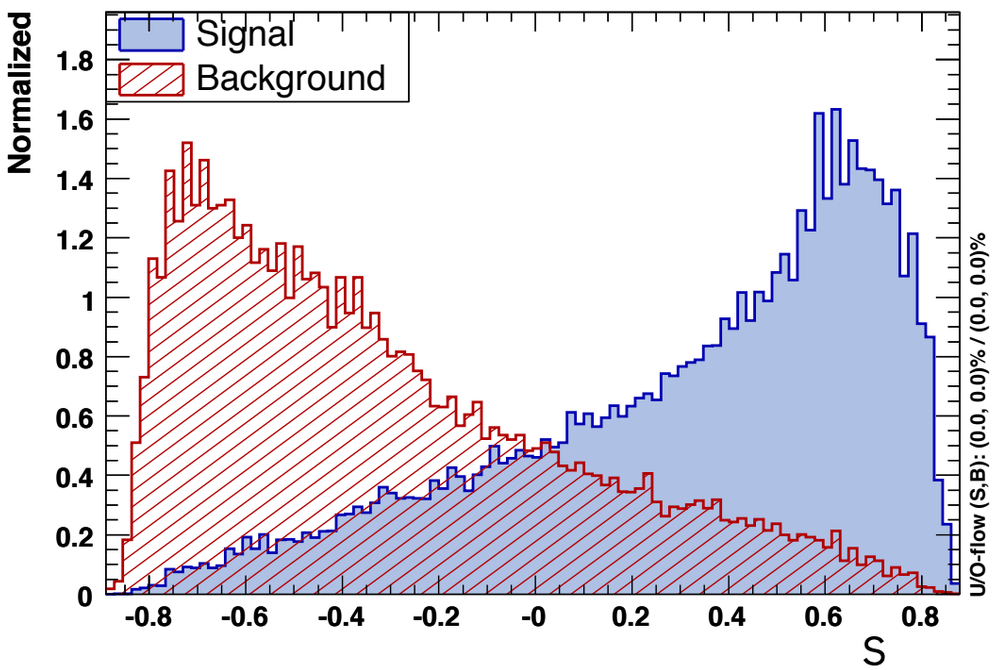


- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autoregressive models, Normalizing Flows

[image credit: A.P. Goucher]

LIKELIHOOD RATIO TRICK





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

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

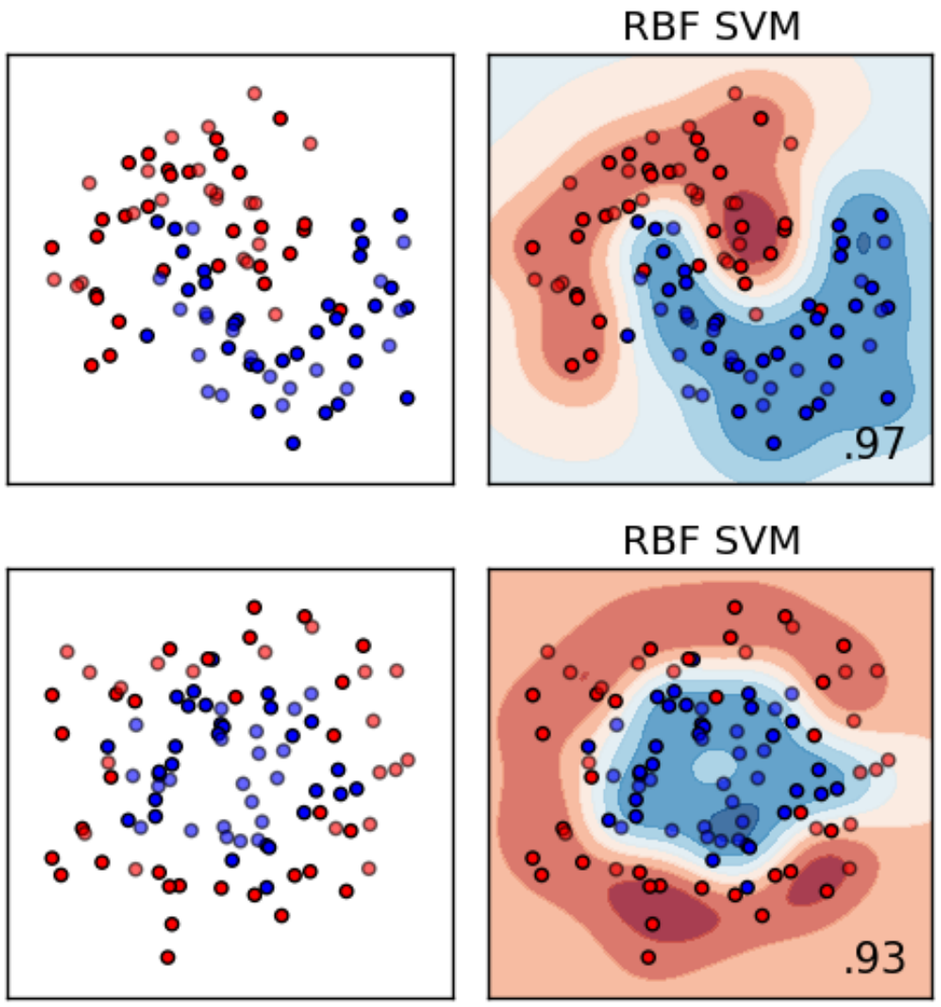
- i.e. approximate the Bayes optimal classifier

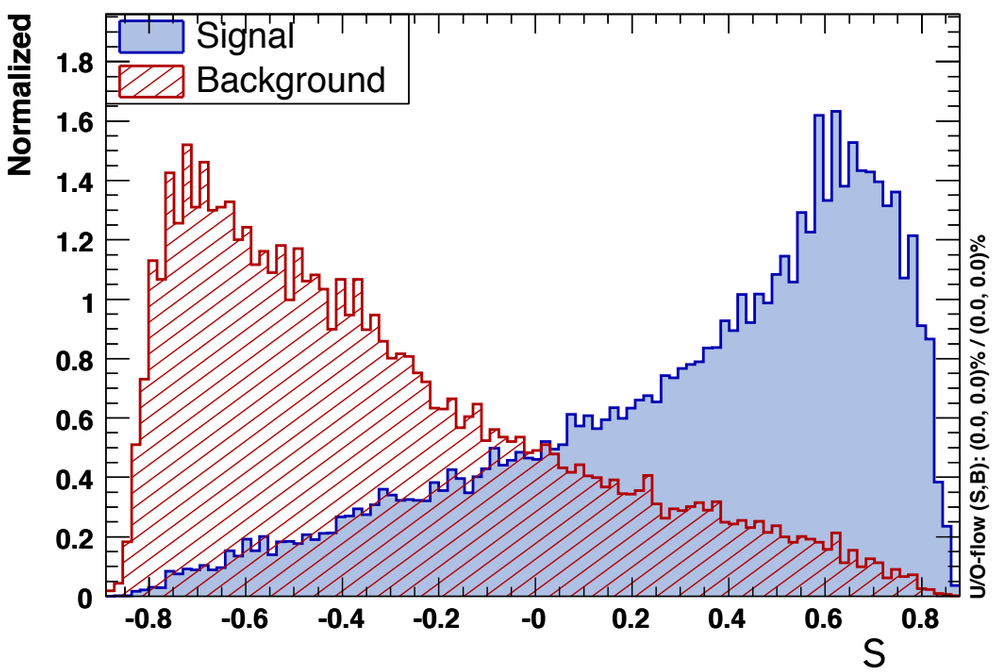
$$s(x) = \frac{p(x|H_1)}{p(x|H_0) + p(x|H_1)}$$

- 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

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

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

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

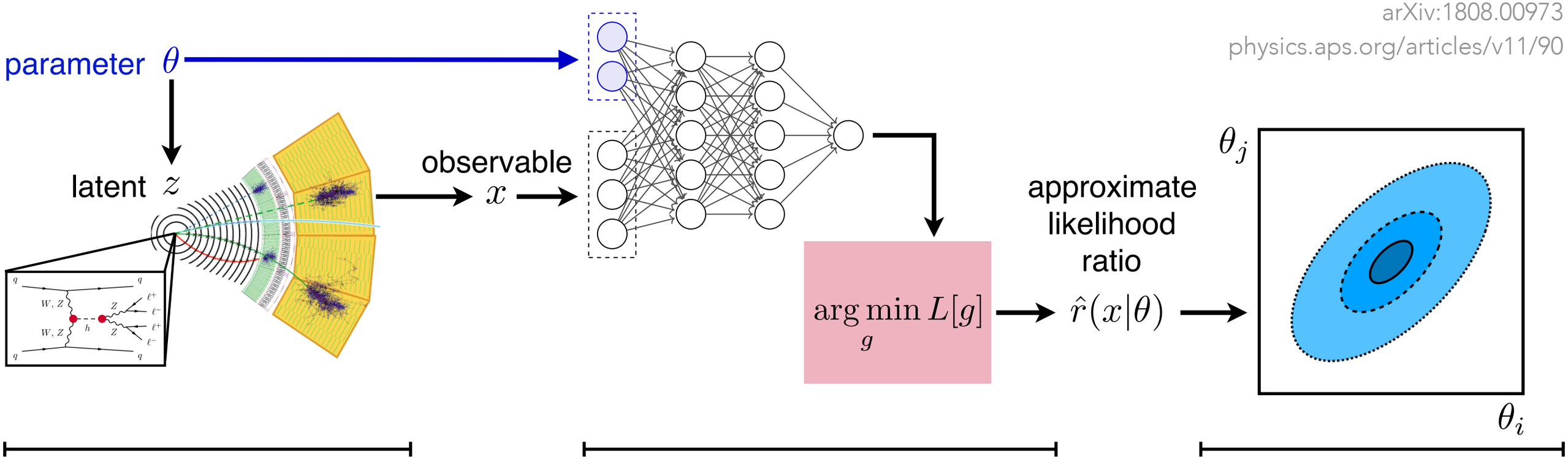
Can do the same thing for any two points θ_0 & θ_1 in parameter space Θ. I call this a **parametrized classifier**

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

K.C., G. Louppe, J. Pavez: Approximating Likelihood Ratios with Calibrated Discriminative Classifiers [arXiv:1506.02169]

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

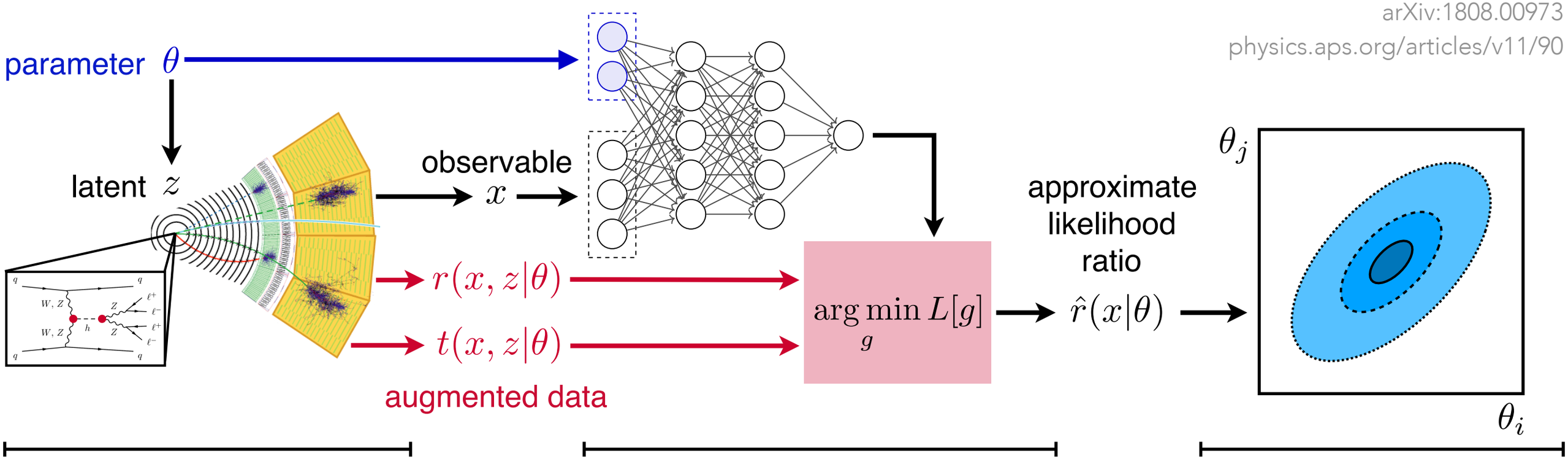
Currently a 2-stage process:

1. learning surrogate
2. Inference on parameters of simulator

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

LEARNING THE LIKELIHOOD RATIO

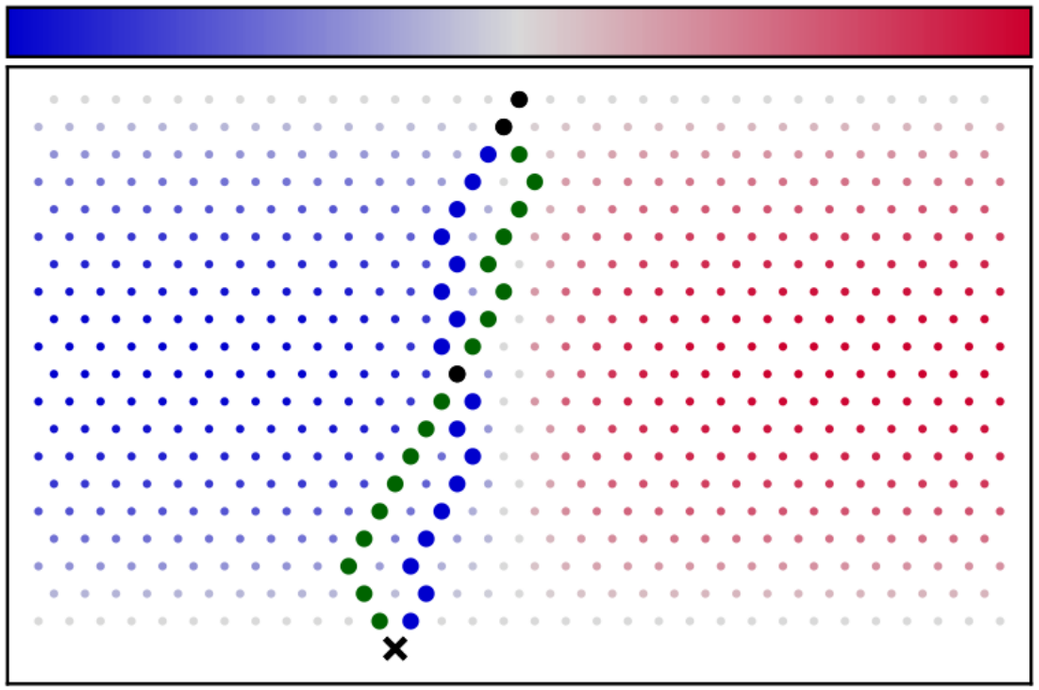
arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
arXiv:1808.00973
physics.aps.org/articles/v11/90

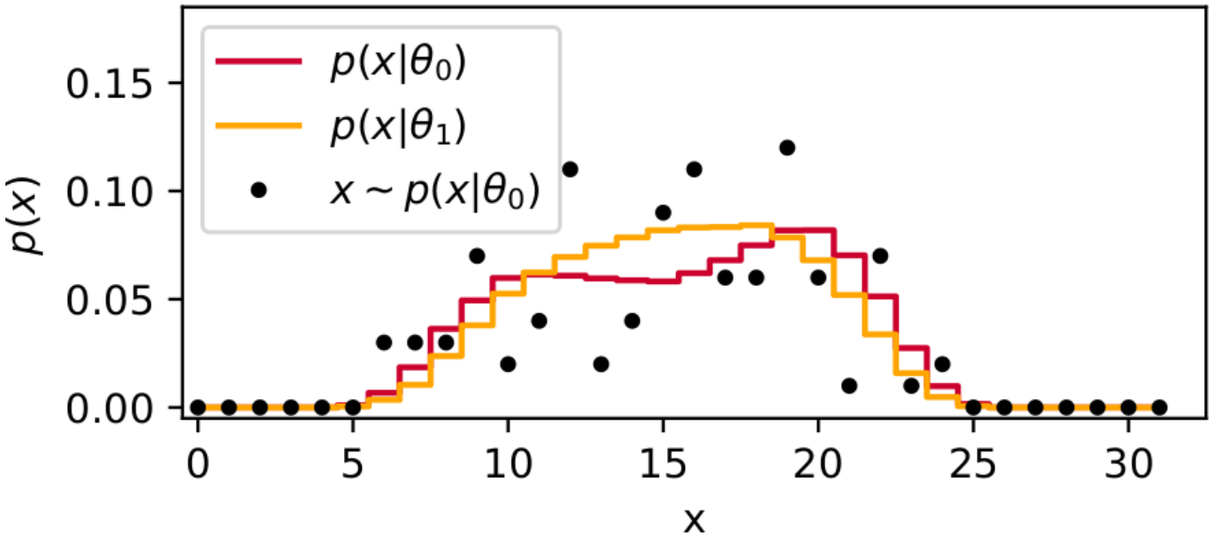


Recently, we realized we can **extract more from the simulator**.
We can use **augmented data** to improve training



MINING GOLD





While implicit density is intractable

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

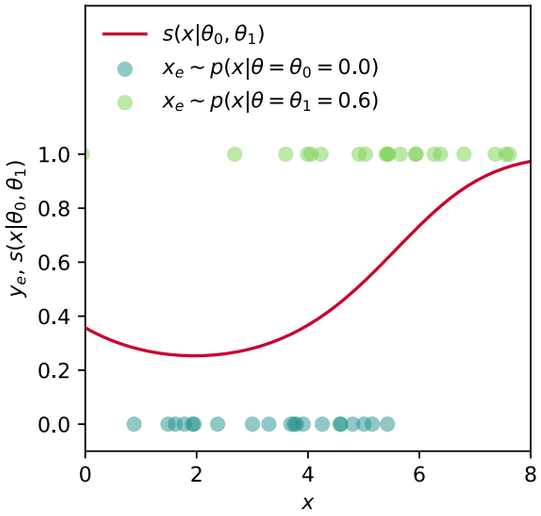
Some quantities conditioned on latent z are tractable:

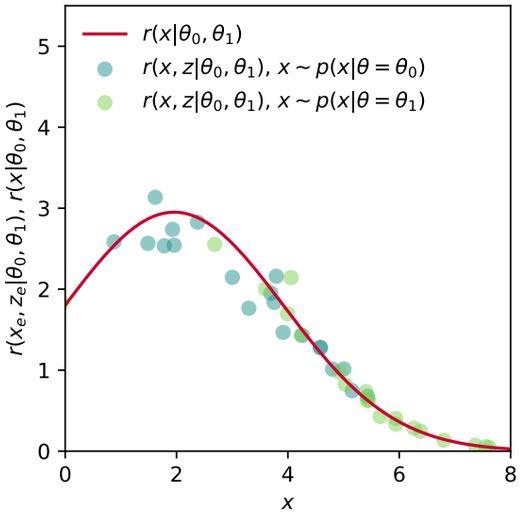
$$r(x, z|\theta_0, \theta_1) = \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)}$$

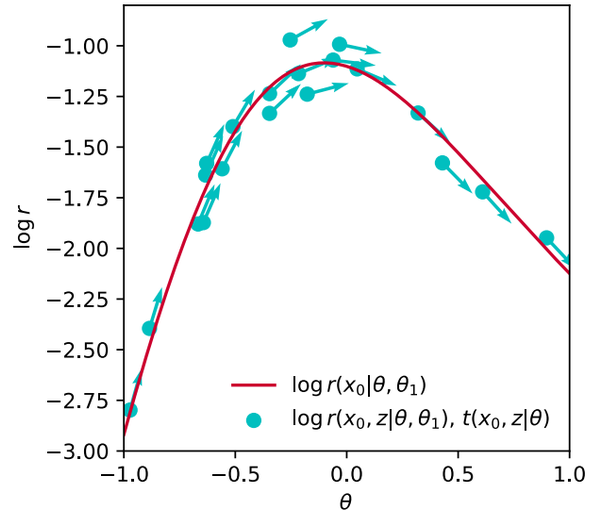
and similar to REINFORCE policy gradient

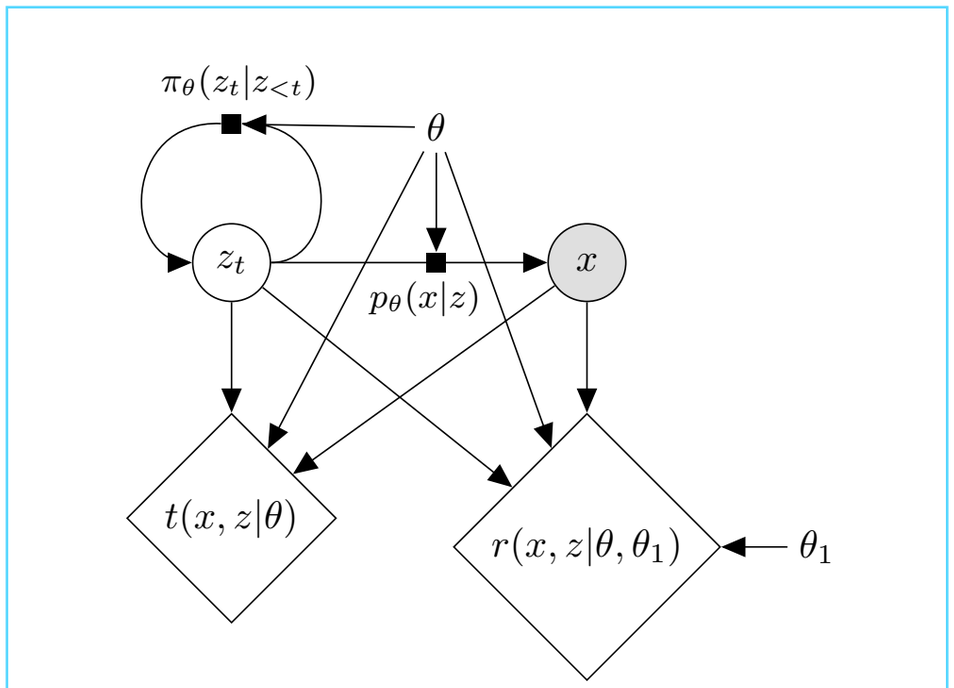
$$t(x, z|\theta_0) = \frac{\nabla_\theta p(x, z|\theta)|_{\theta_0}}{p(x, z|\theta_0)} = \nabla_\theta \log p(x, z|\theta)|_{\theta_0}$$

PUTTING IT ALL TOGETHER

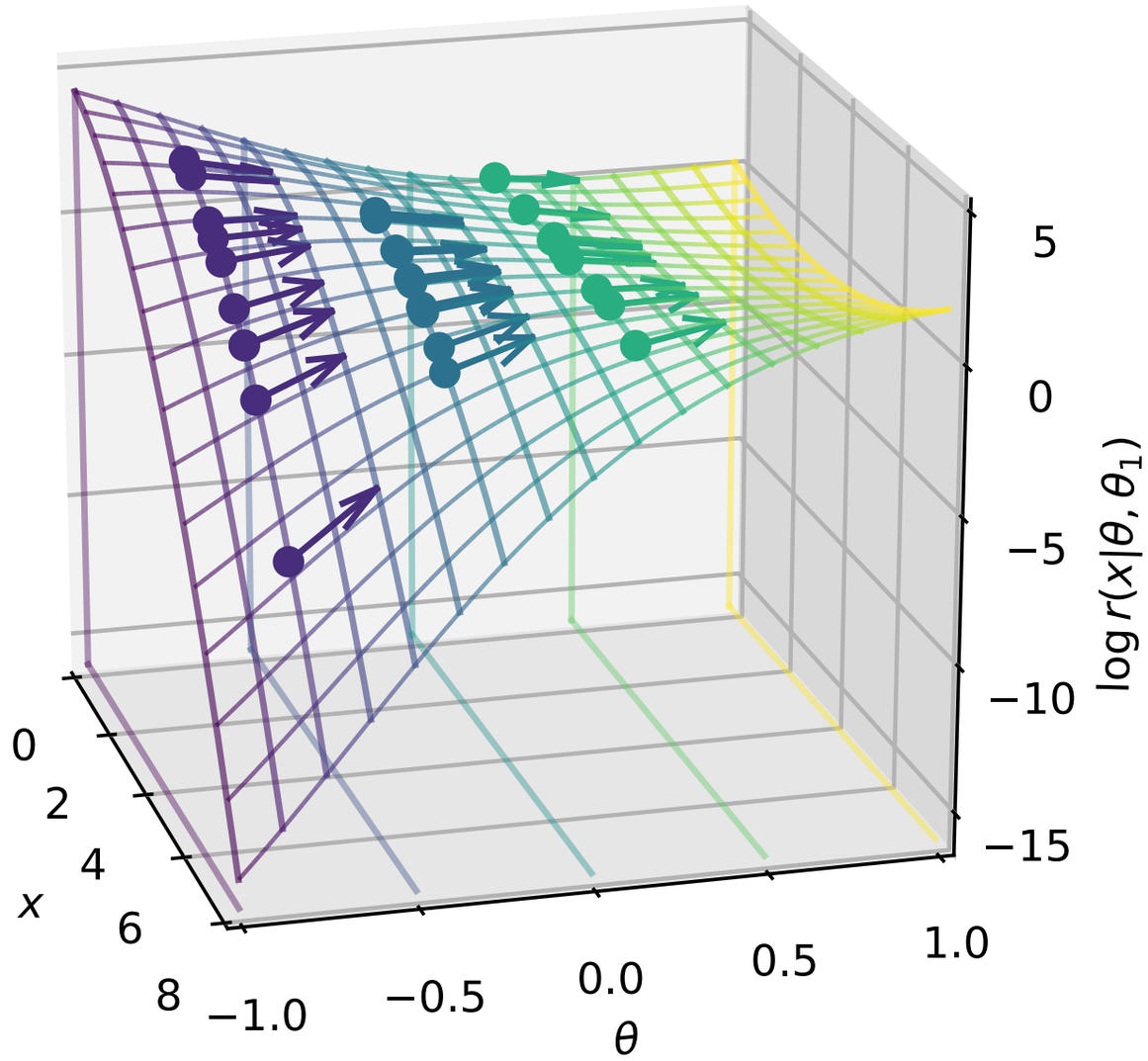






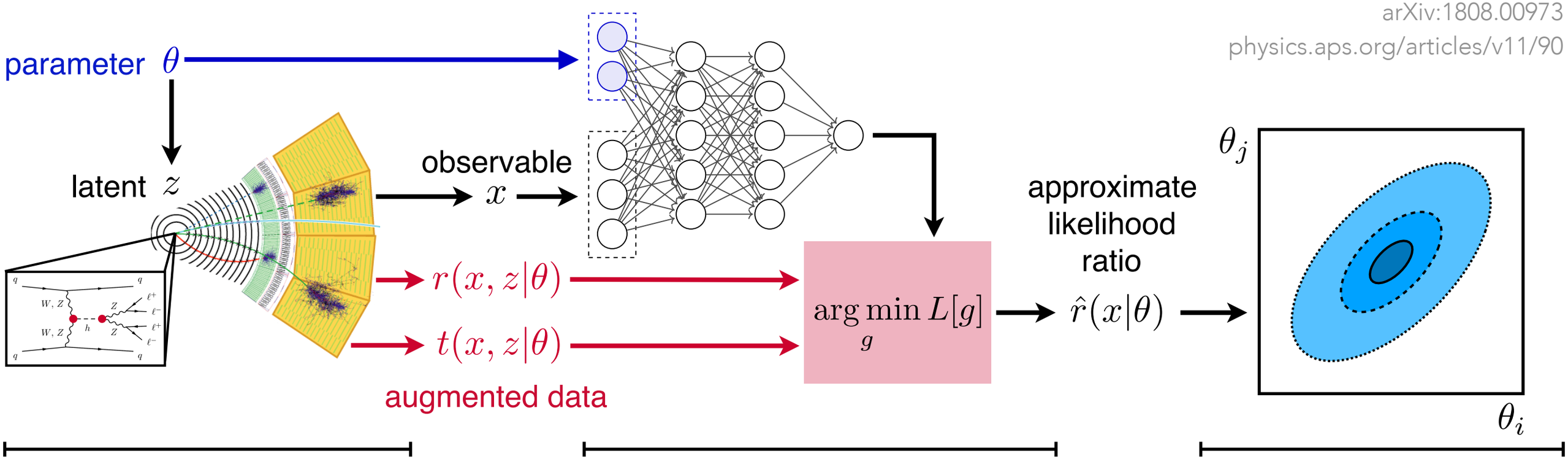


can think of simulator as policy πθ in language of reinforcement learning



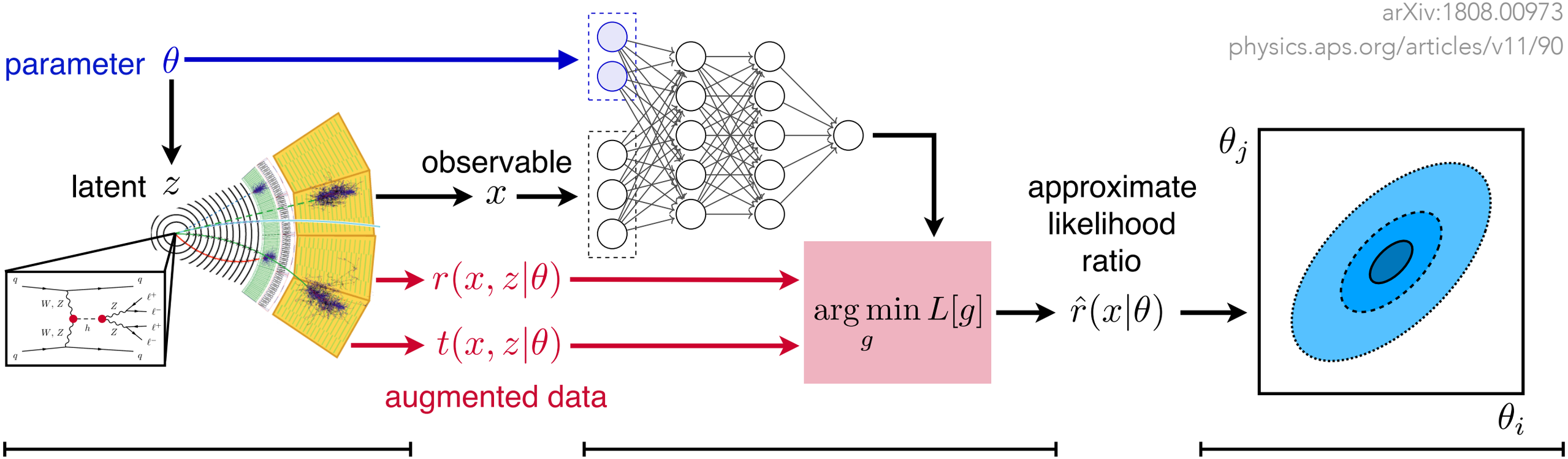
LEARNING THE LIKELIHOOD RATIO

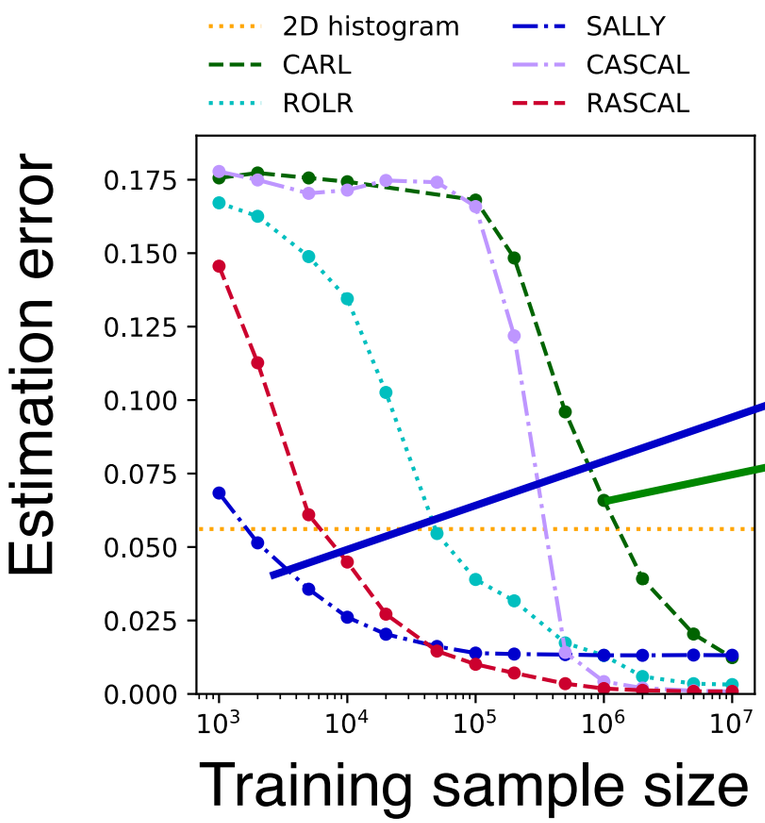
arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
arXiv:1808.00973
physics.aps.org/articles/v11/90



LEARNING THE LIKELIHOOD RATIO

arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
arXiv:1808.00973
physics.aps.org/articles/v11/90





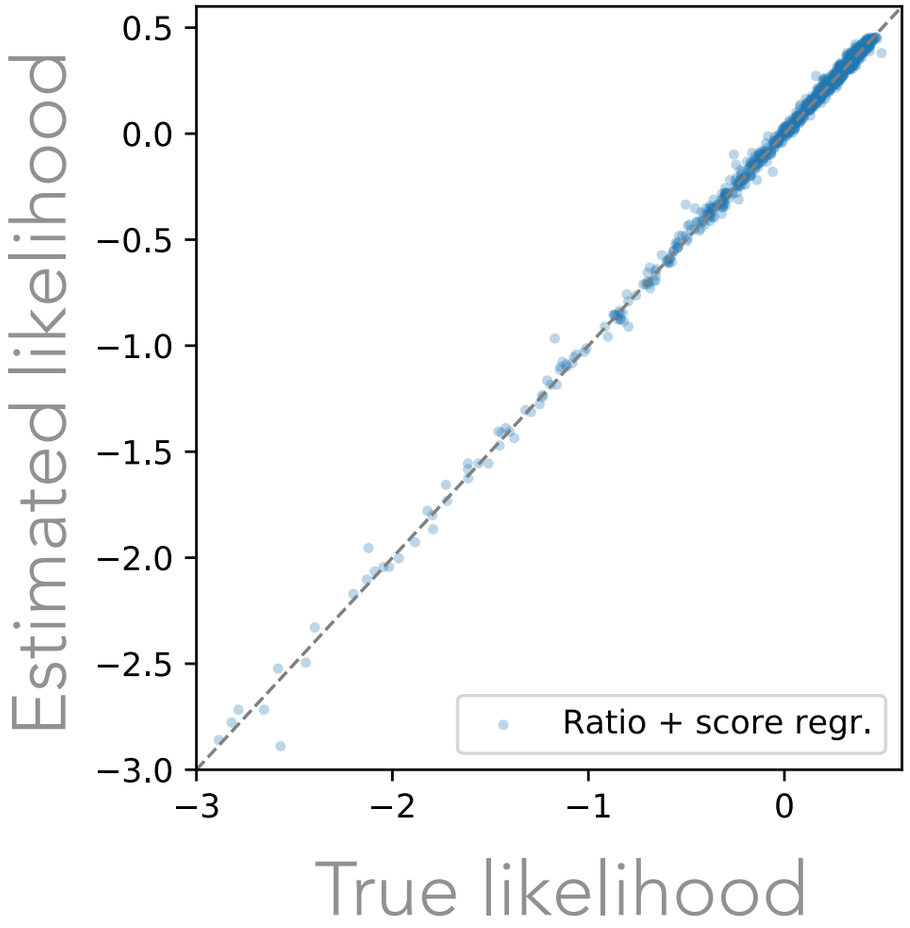
New techniques
require less data than
without augmented data

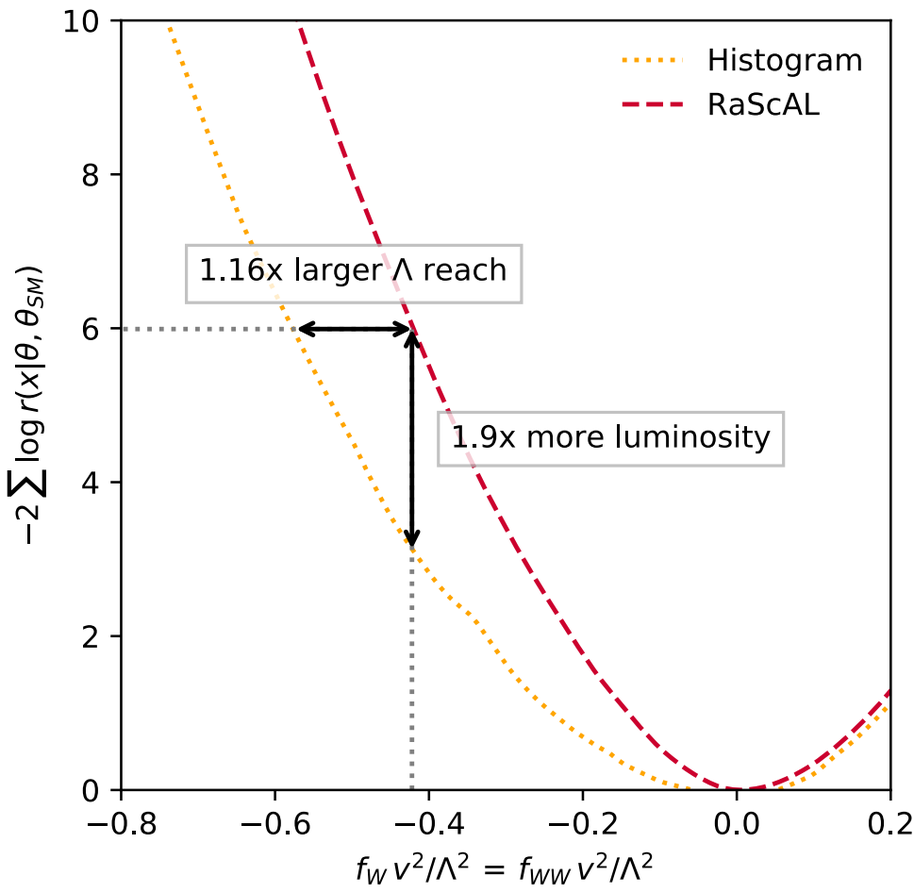
Traditional Approach no NN

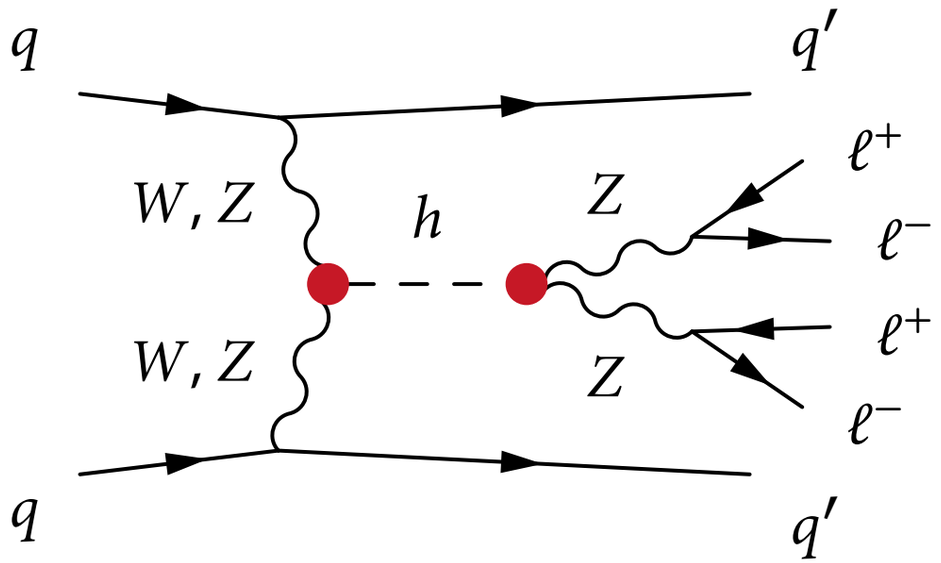
Two Examples

IMPACT ON STUDIES OF THE HIGGS BOSON

(based on a 42-Dim observation ***x***)

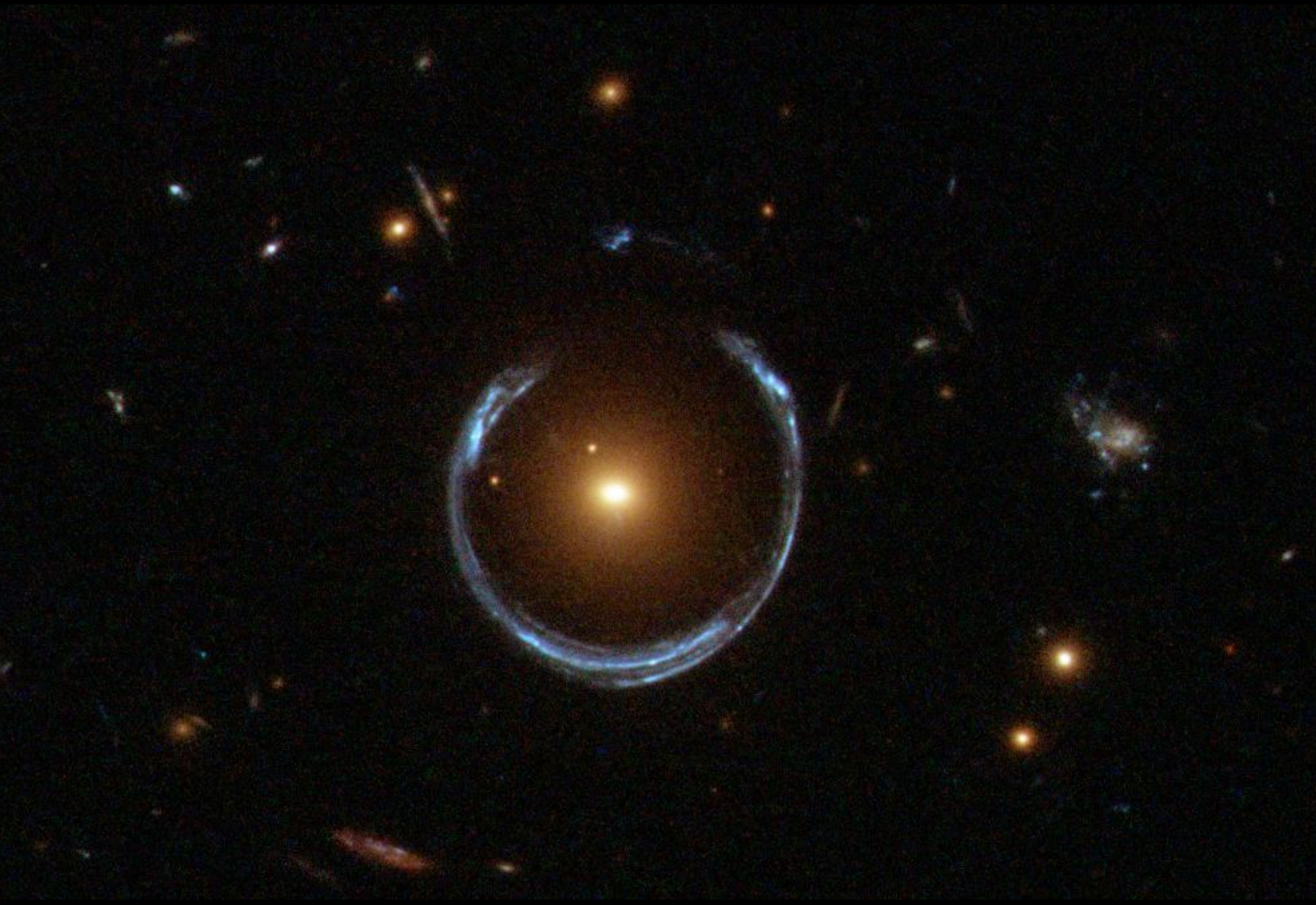


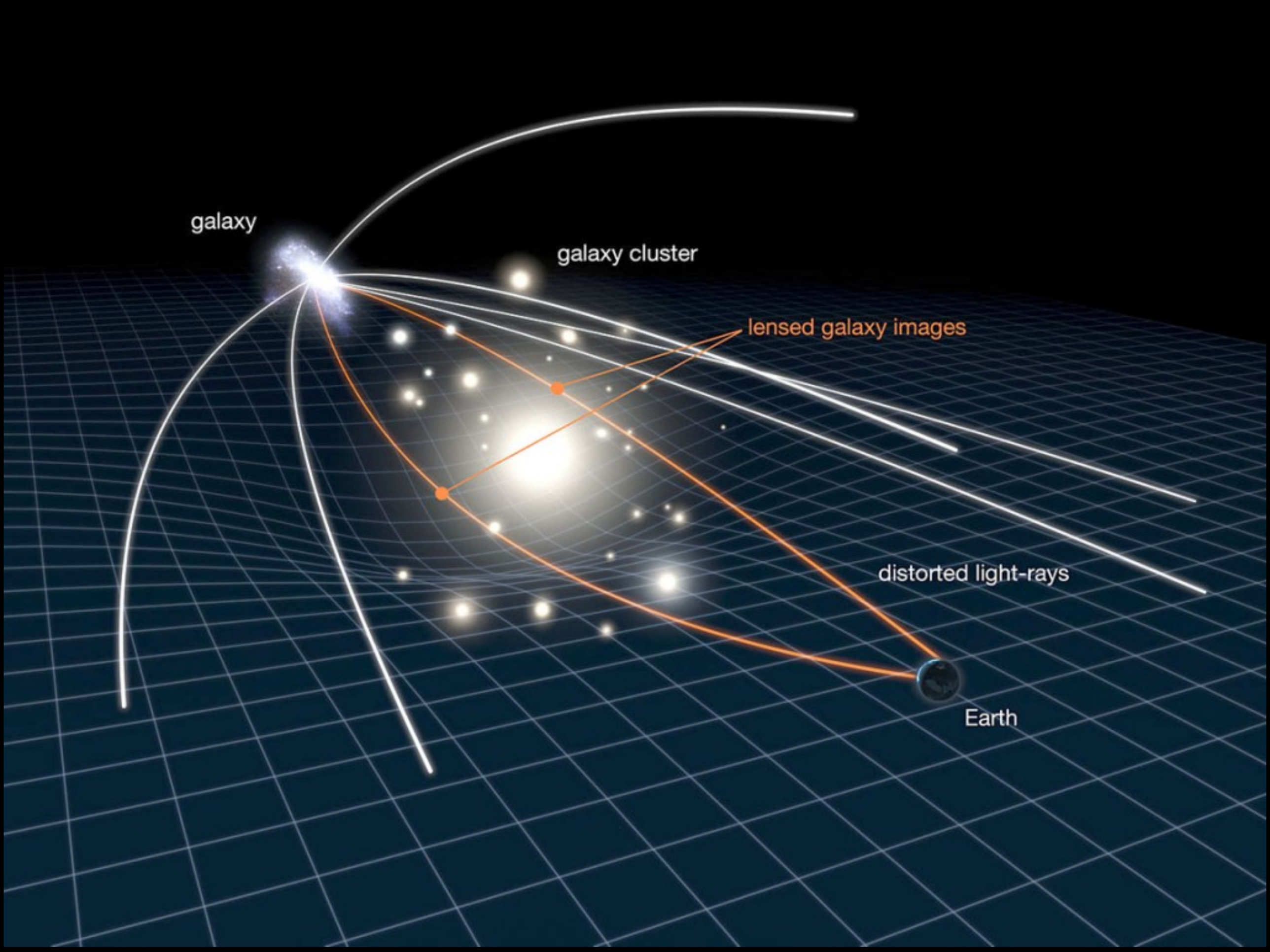






J Brehmer, J Pavez, G Louppe, K.C. PRL & PRD 2018 [arXiv:1805.00013 & arXiv:1805.00020]
"Better Higgs Measurements Through Information Geometry" [arXiv:1612.05261] & CARL [arxiv:1506.02169]



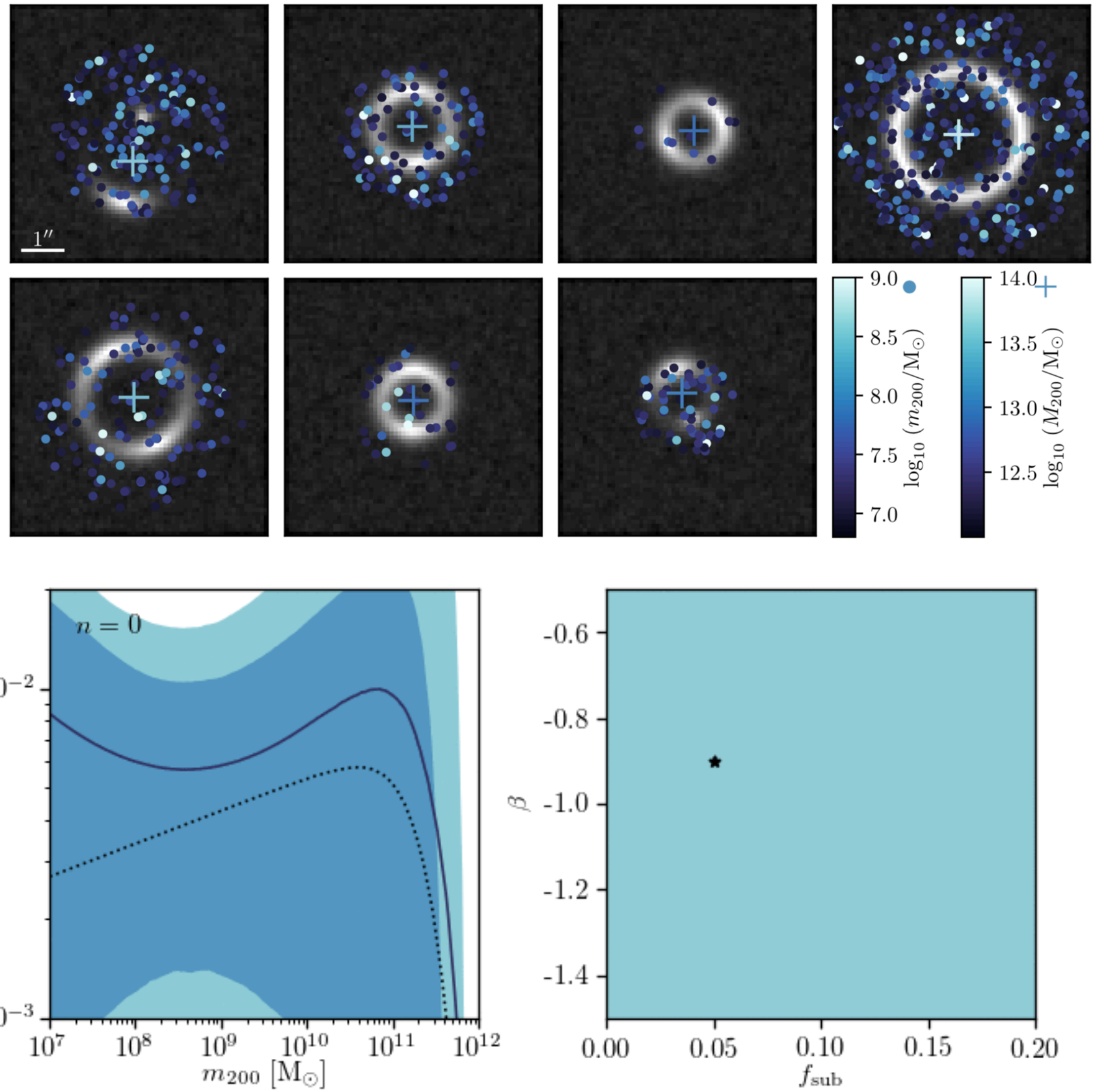

galaxy
galaxy cluster
lensed galaxy images
distorted light-rays
Earth

arXiv:1909.02005

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

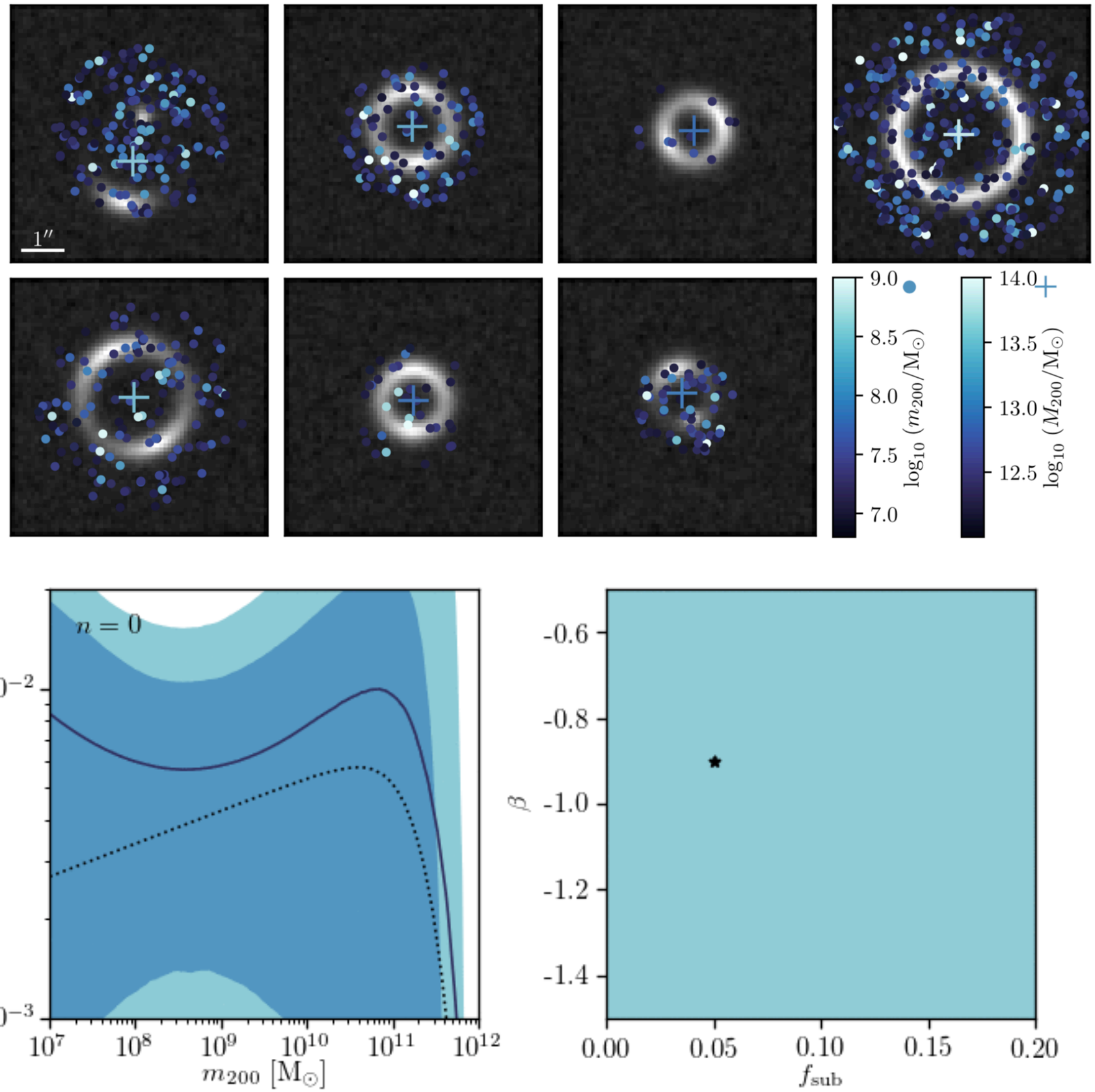


arXiv:1909.02005

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

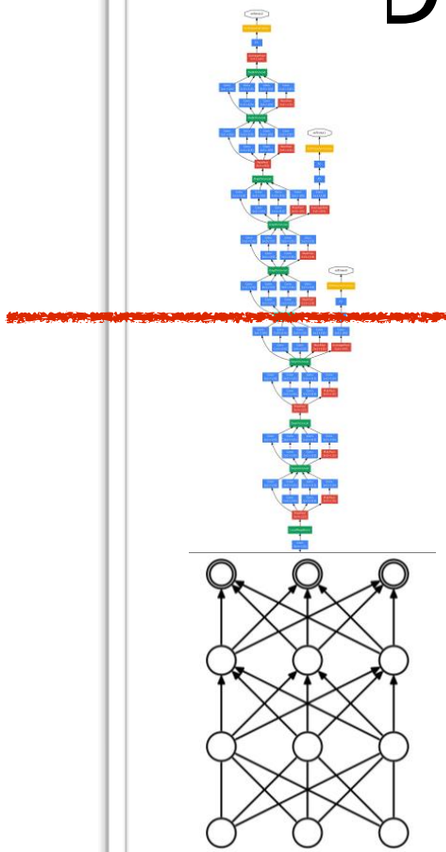


TWO PARADIGMS IN MACHINE LEARNING



Max
Welling

Discriminative or Generative?



-Deep Learning

-Kernel Methods

-Random Forests

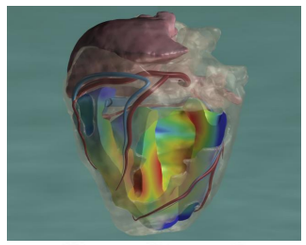
-Boosting

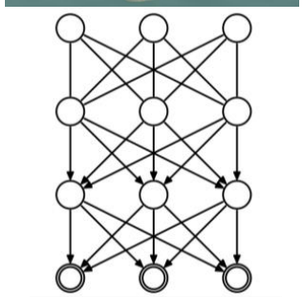
Variational Auto-Encoder

-Bayesian Networks

-Probabilistic Programs

-Simulator Models





- 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

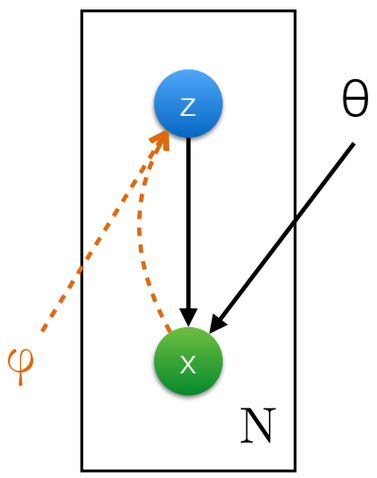
Deep Generative Models

VARIATIONAL AUTO-ENCODER

[Slides from D. Kingma NIPS 2015]

Auto-Encoding Variational Bayes

[Kingma and Welling, 2013/2014]
[Rezende et al, 2014]



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

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



Diederik (Durk) Kingma



Max Welling



Danilo J. Rezende

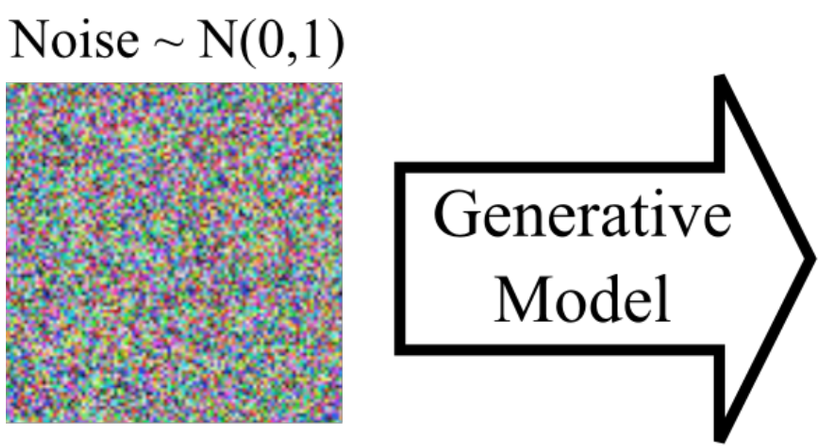
Conv. net as encoder/decoder, trained on faces

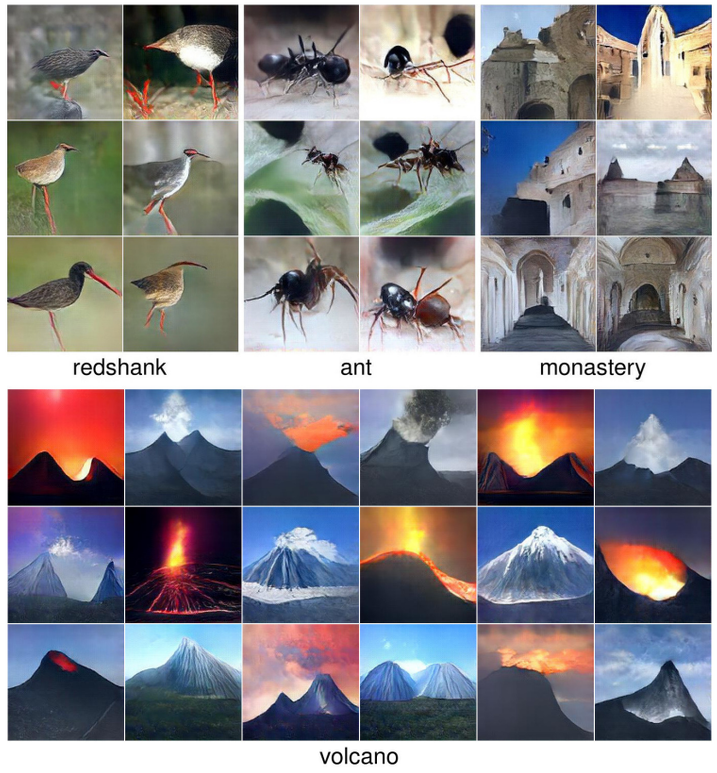


GENERATIVE ADVERSARIAL NETWORKS

Z

X



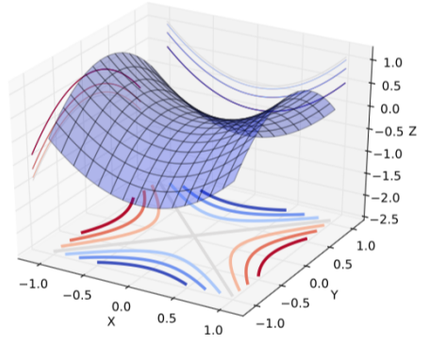




Leo is G Tom is D

- 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

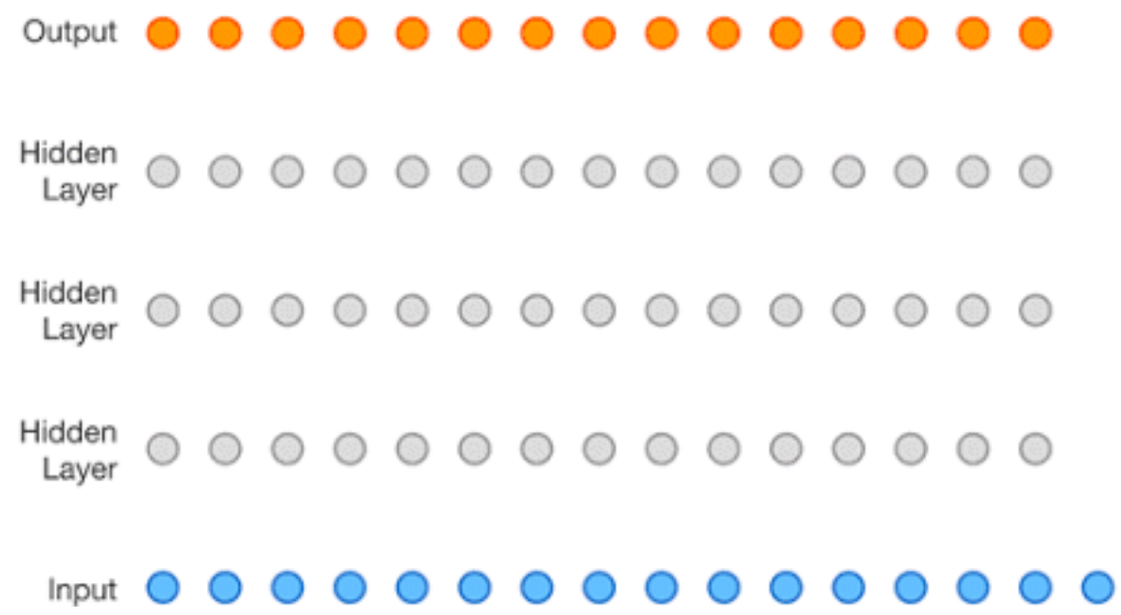
Are there other deep generative models that can help?

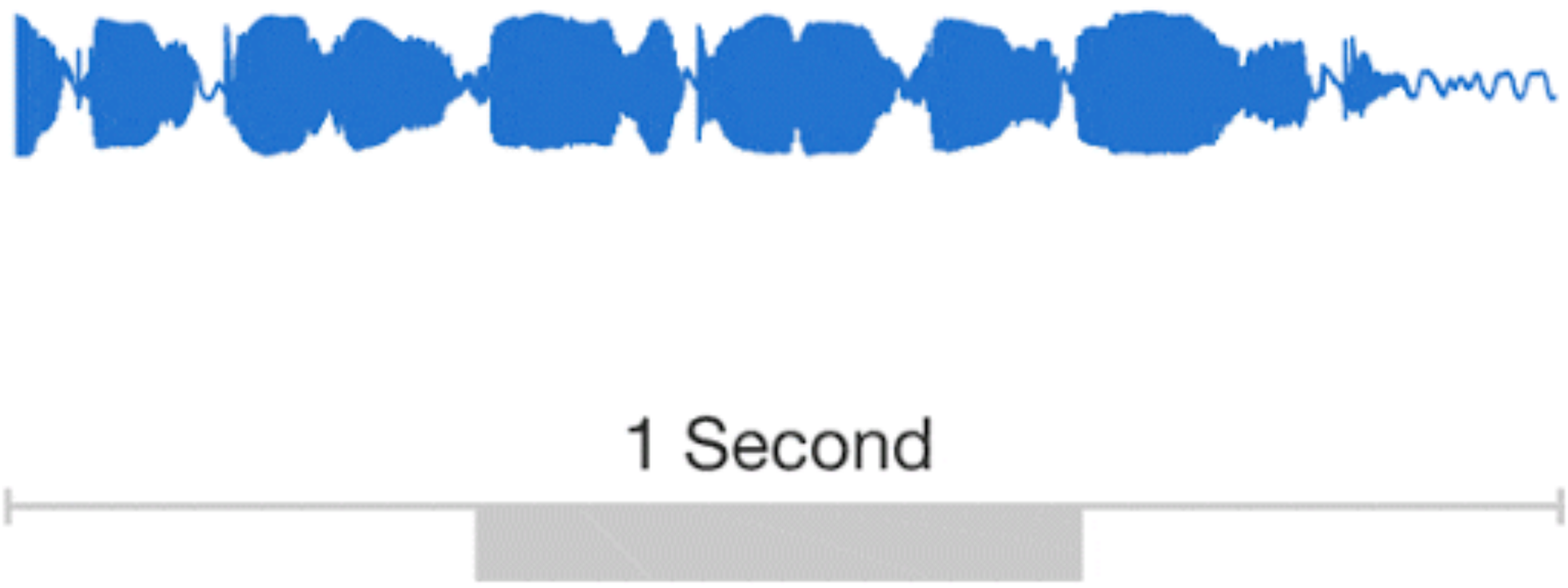
WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

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

have a tractable density. Train via maximum likelihood



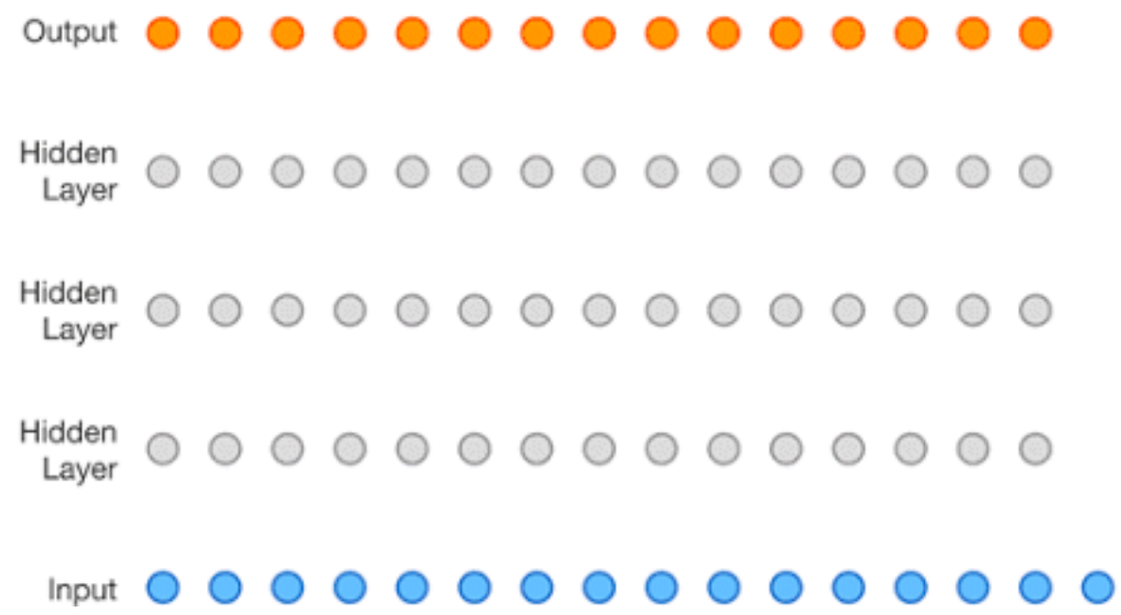


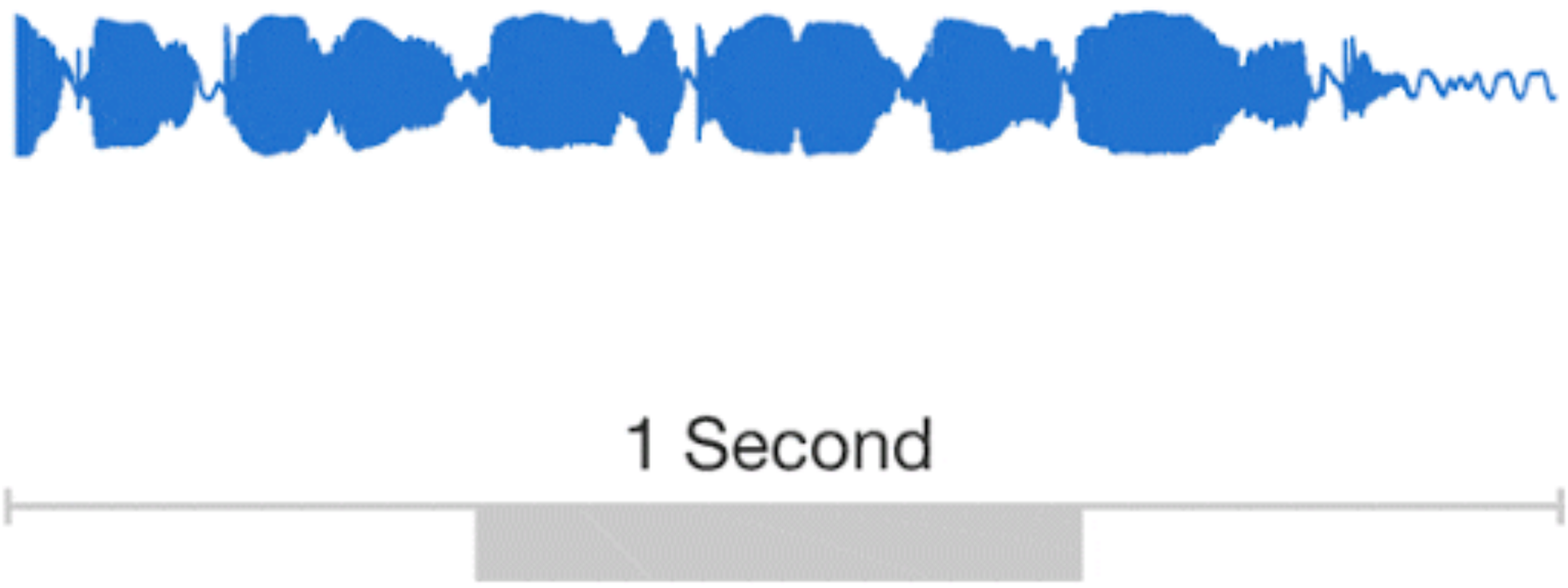
WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

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

have a tractable density. Train via maximum likelihood



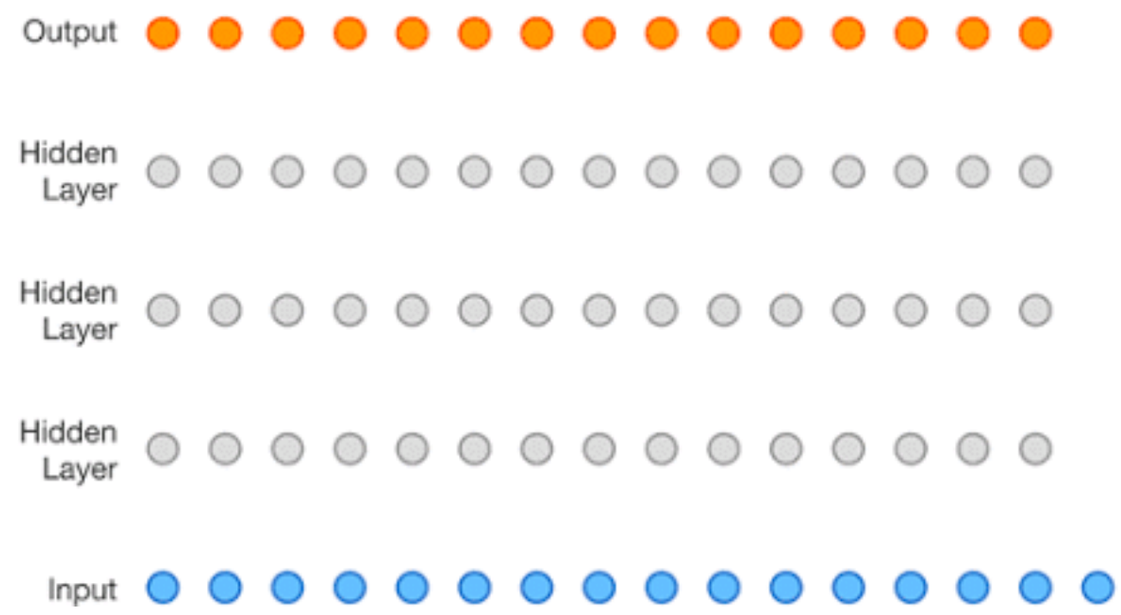


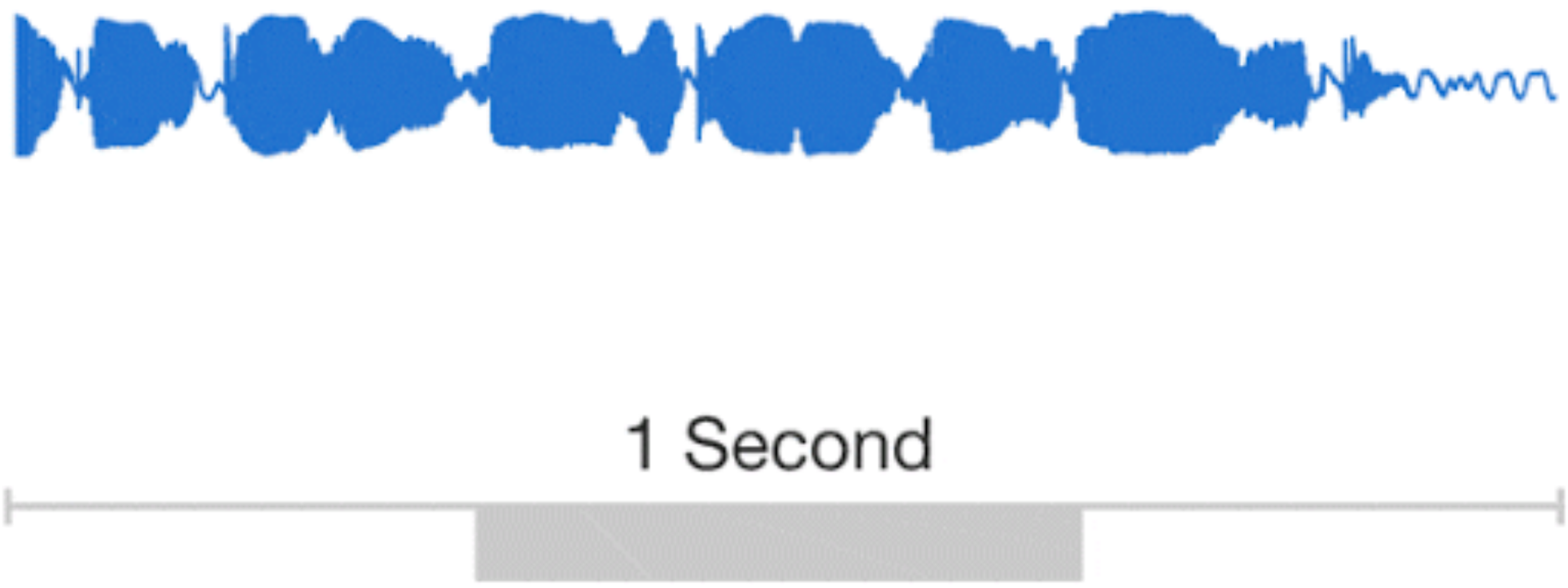
WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Autoregressive models defined by

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

have a tractable density. Train via maximum likelihood





Approximations using Change-of-variables

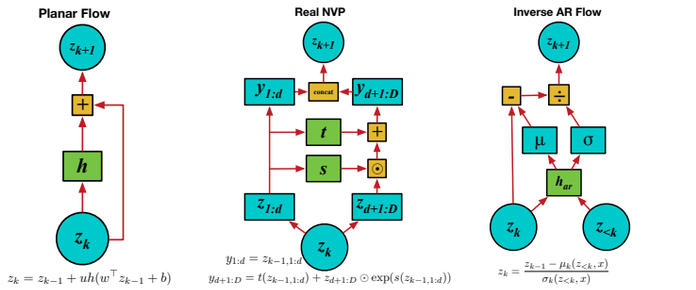
Exploit the rule for change of variables for random variables:

- 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

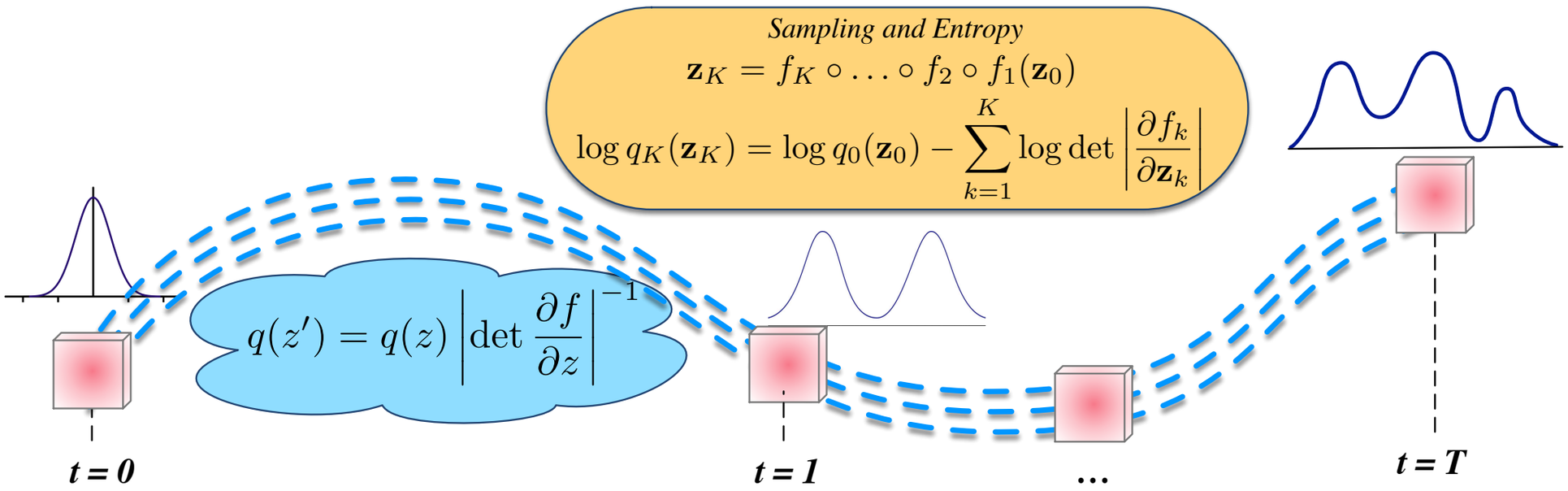
$$\mathcal{L} = \mathbb{E}_{q_0(\mathbf{z}_0)}[\log p(\mathbf{x}, \mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z}_0)}[\log q_0(\mathbf{z}_0)] - \mathbb{E}_{q_0(\mathbf{z}_0)}\left[\sum_{k=1}^{K} \log \det \left|\frac{\partial f_k}{\partial \mathbf{z}_k}\right|\right]$$

- 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

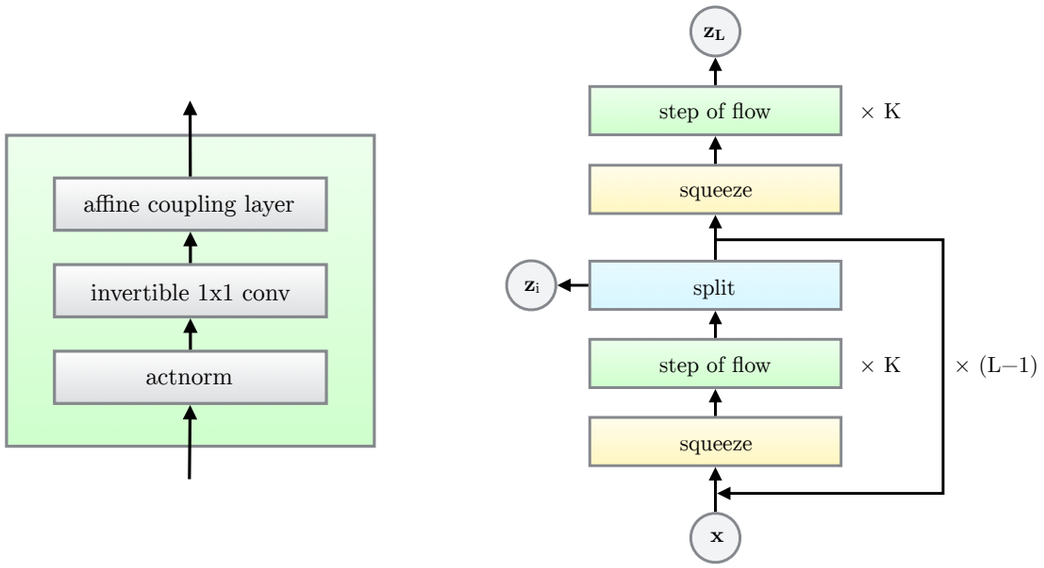
[Rezende and Mohamed, 2015]

FLOWS WITH INVERTIBLE CONVOLUTIONS

https://arxiv.org/abs/1807.03039

Glow: Generative Flow with Invertible 1×1 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 Hoogeboom [1 2] **Rianne van den Berg** [1] **Max Welling** [1 3]

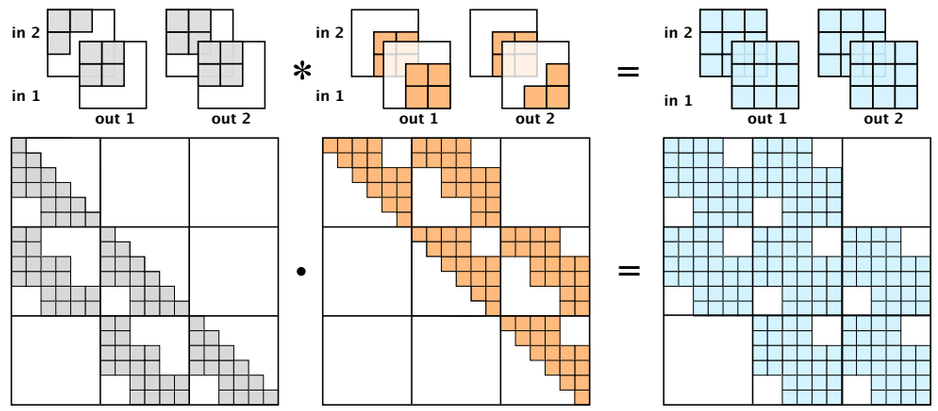


Table 3. Performance of Emerging convolutions on CIFAR10, ImageNet 32x32 and ImageNet 64x64 in bits per dimension (negative $\log_2$-likelihood), and ± reports standard deviation.

	CIFAR10	ImageNet 32x32	ImageNet 64x64
Real NVP	3.51	4.28	3.98
Glow	3.36 ±0.002	**4.09**	**3.81**
Emerging	**3.34** ±0.002	**4.09**	**3.81**

FLOWS WITH CONTINUOUS TIME

FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models

Will Grathwohl[*†‡], **Ricky T. Q. Chen**[*†], **Jesse Bettencourt**[†], **Ilya Sutskever**[‡], **David Duvenaud**[†]

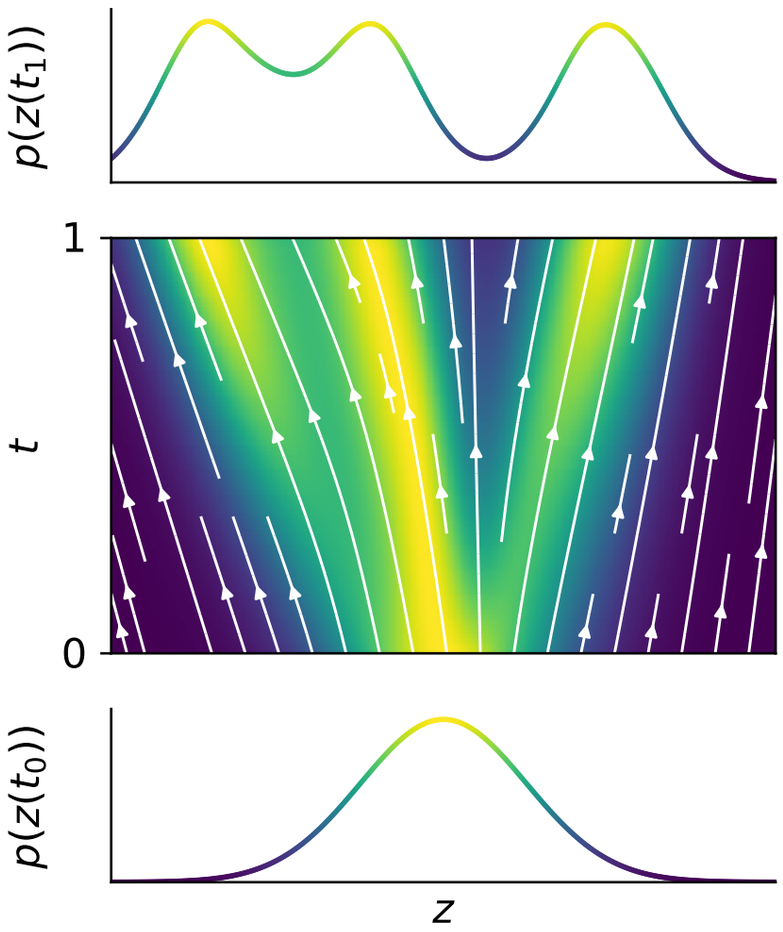
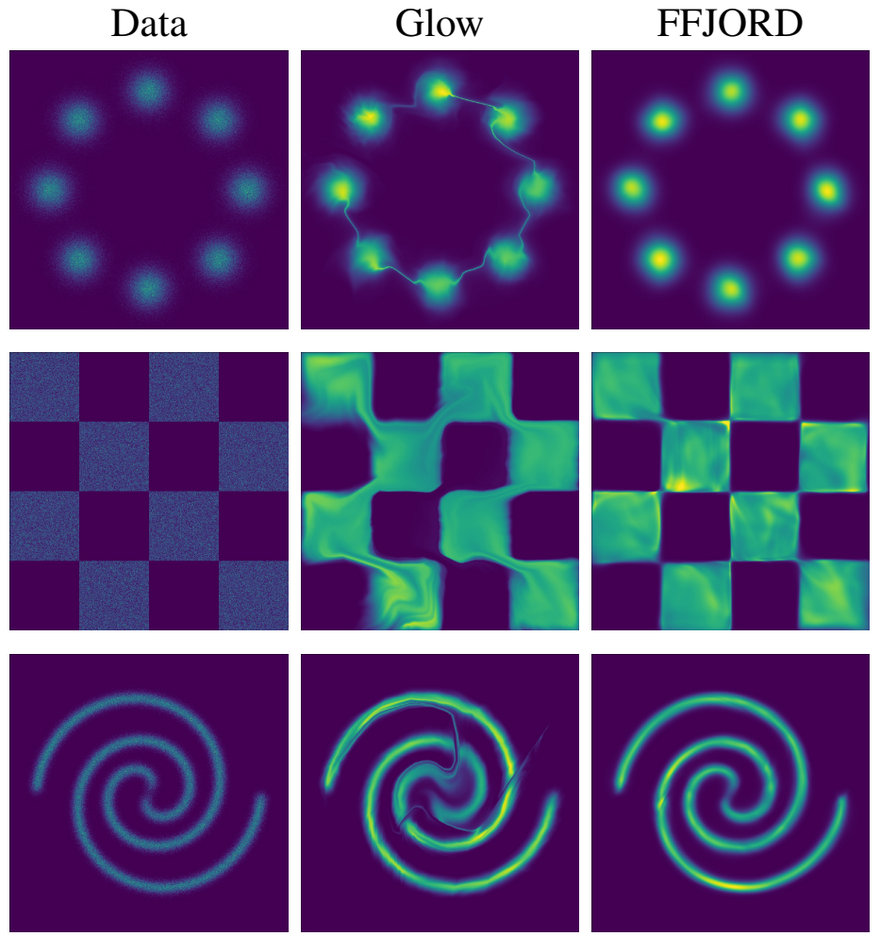


Figure 1: FFJORD transforms a simple base distribution at t_0 into the target distribution at t_1 by integrating over learned continuous dynamics.



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

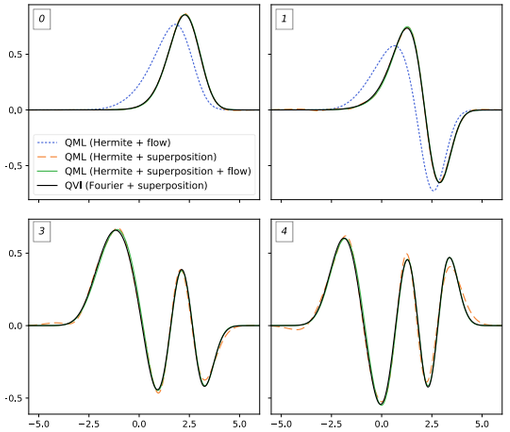
Table 1: A comparison of the abilities of generative modeling approaches.

QUANTUM FLOWS

https://arxiv.org/abs/1904.05903

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

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

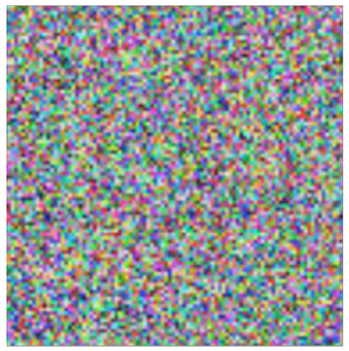
Correlation ≠ Causation

DEEP GENERATIVE MODEL VS. SIMULATION

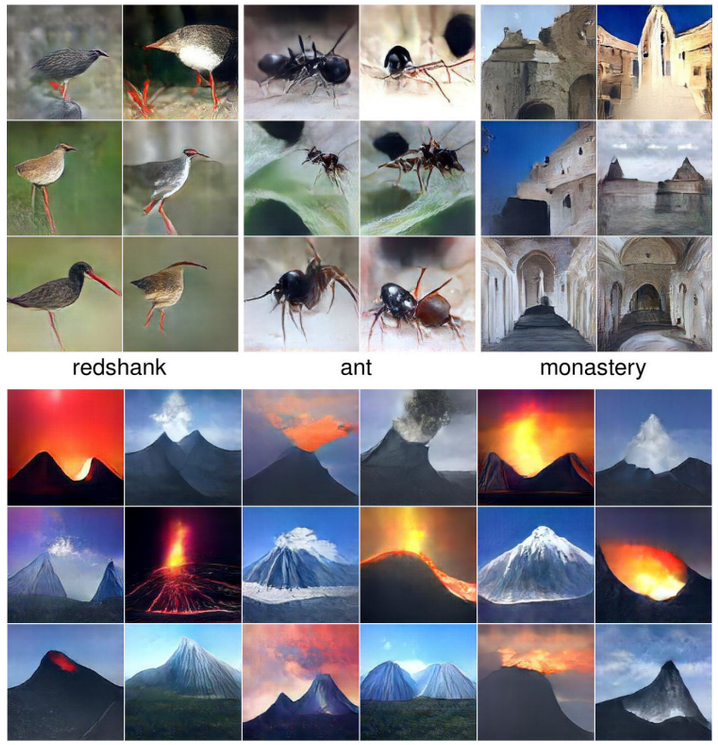
Z

X

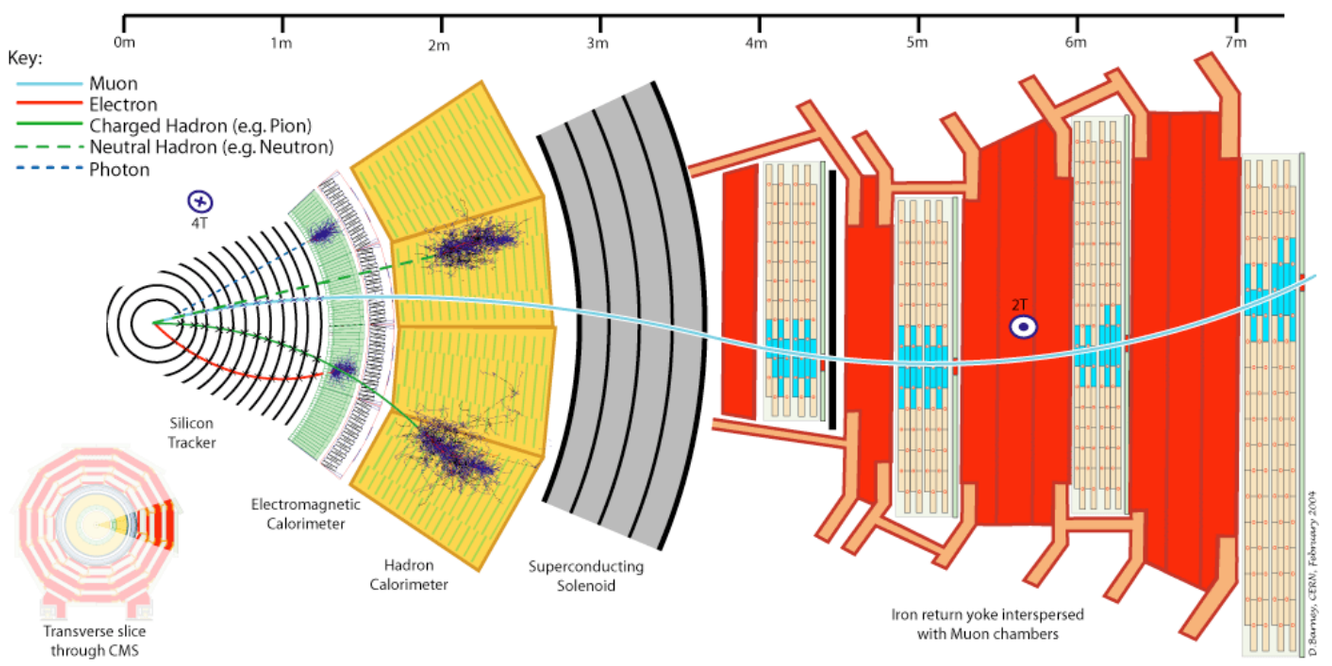
Noise ~ N(0,1)



Generative Model







http://torch.ch/blog/2015/11/13/gan.html

Paganini, et. al. Phys.Rev.Lett. 120 (2018) no.4, 042003 ; & ATL-SOFT-PUB-2018-001

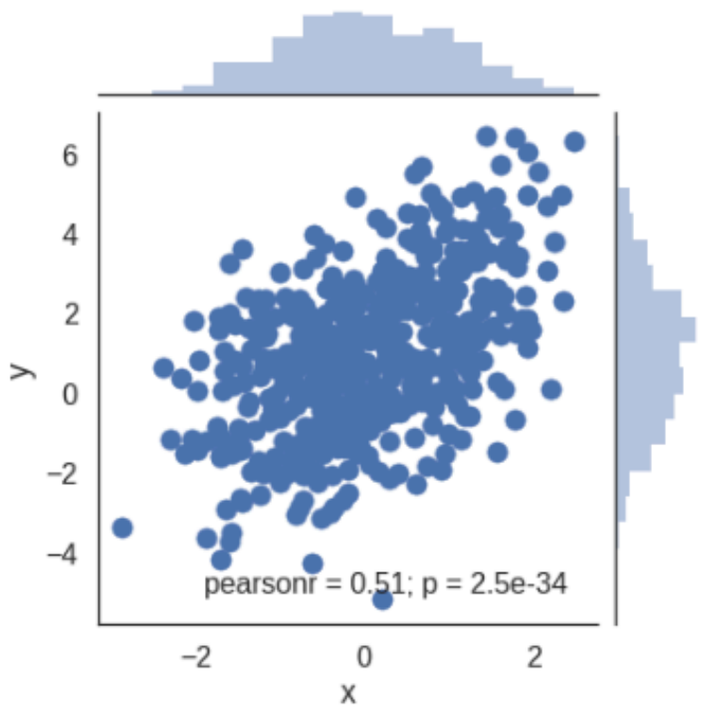
Judea Pearl

Figure 1. The causal hierarchy. Questions at level 1 can be answered only if information from level i or higher is available.

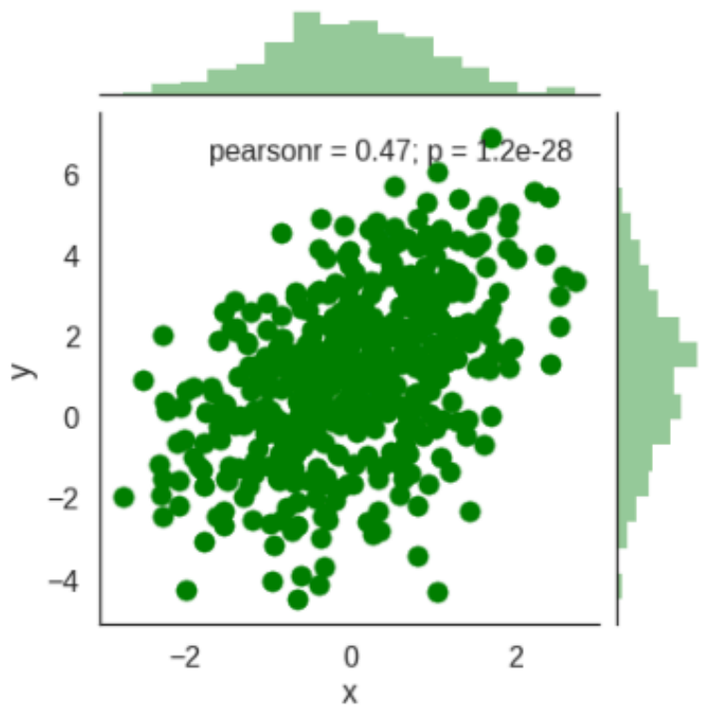
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief inY?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

SAME JOINT DISTRIBUTION, DIFFERENT CAUSAL MODEL

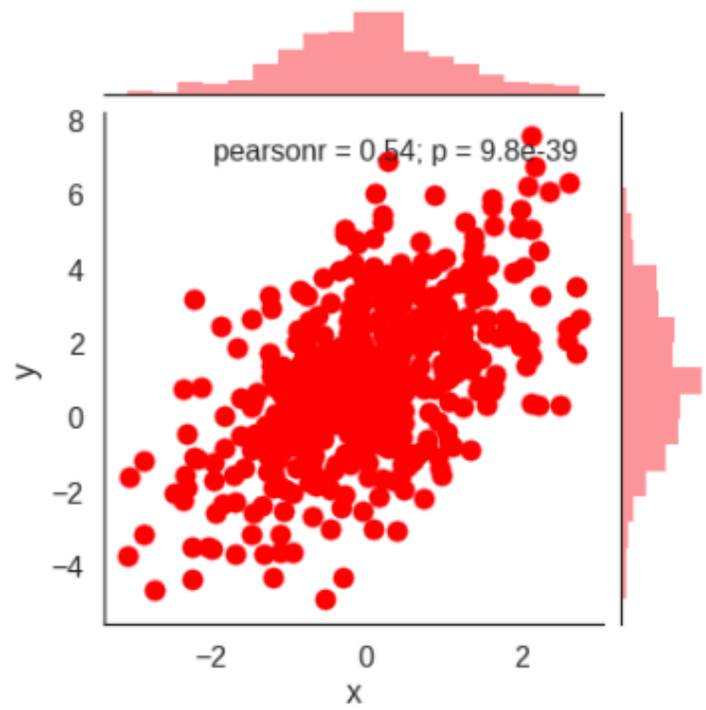
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```




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



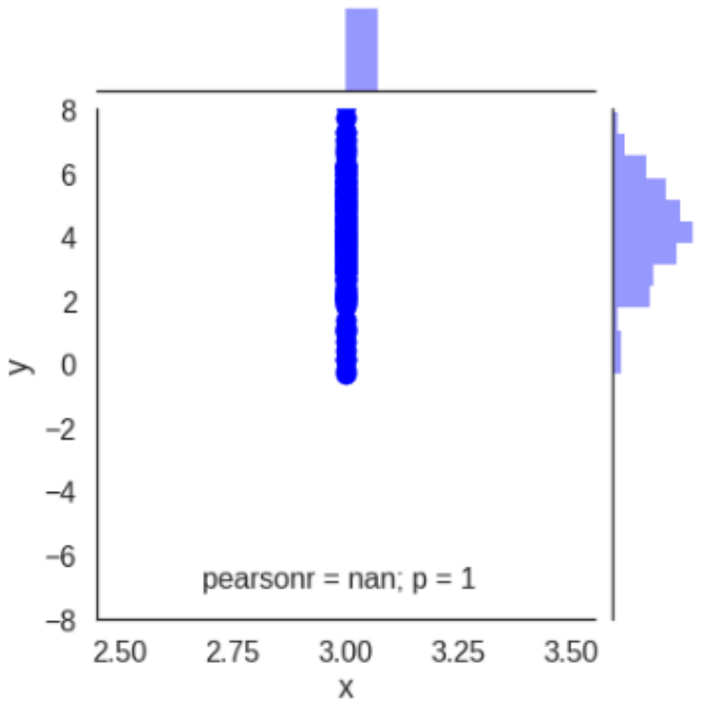

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



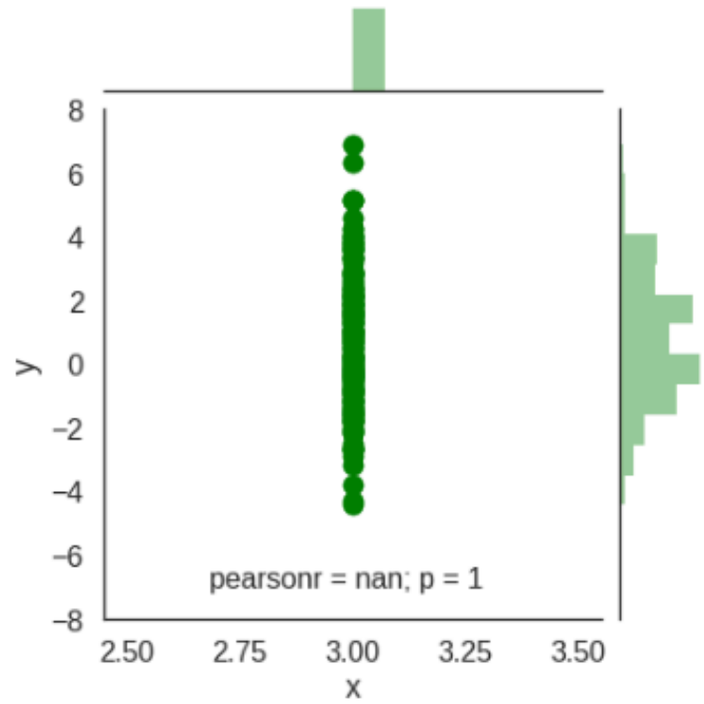
https://colab.research.google.com/drive/1rjjjA7teiZVHJCMTVD8KlZNu3EjS7Dmu#scrollTo=TlzzvcGOZdvW
https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/

CAUSATION > CORRELATION

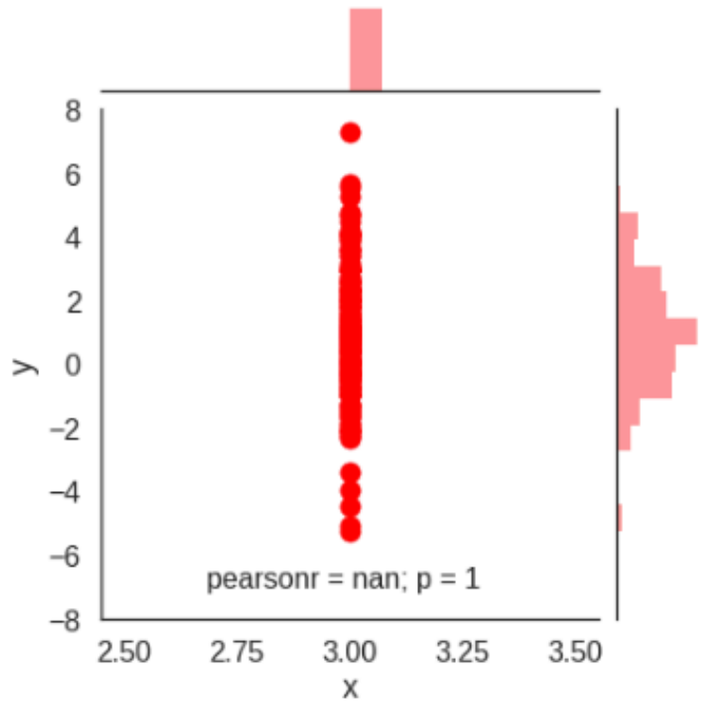
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3



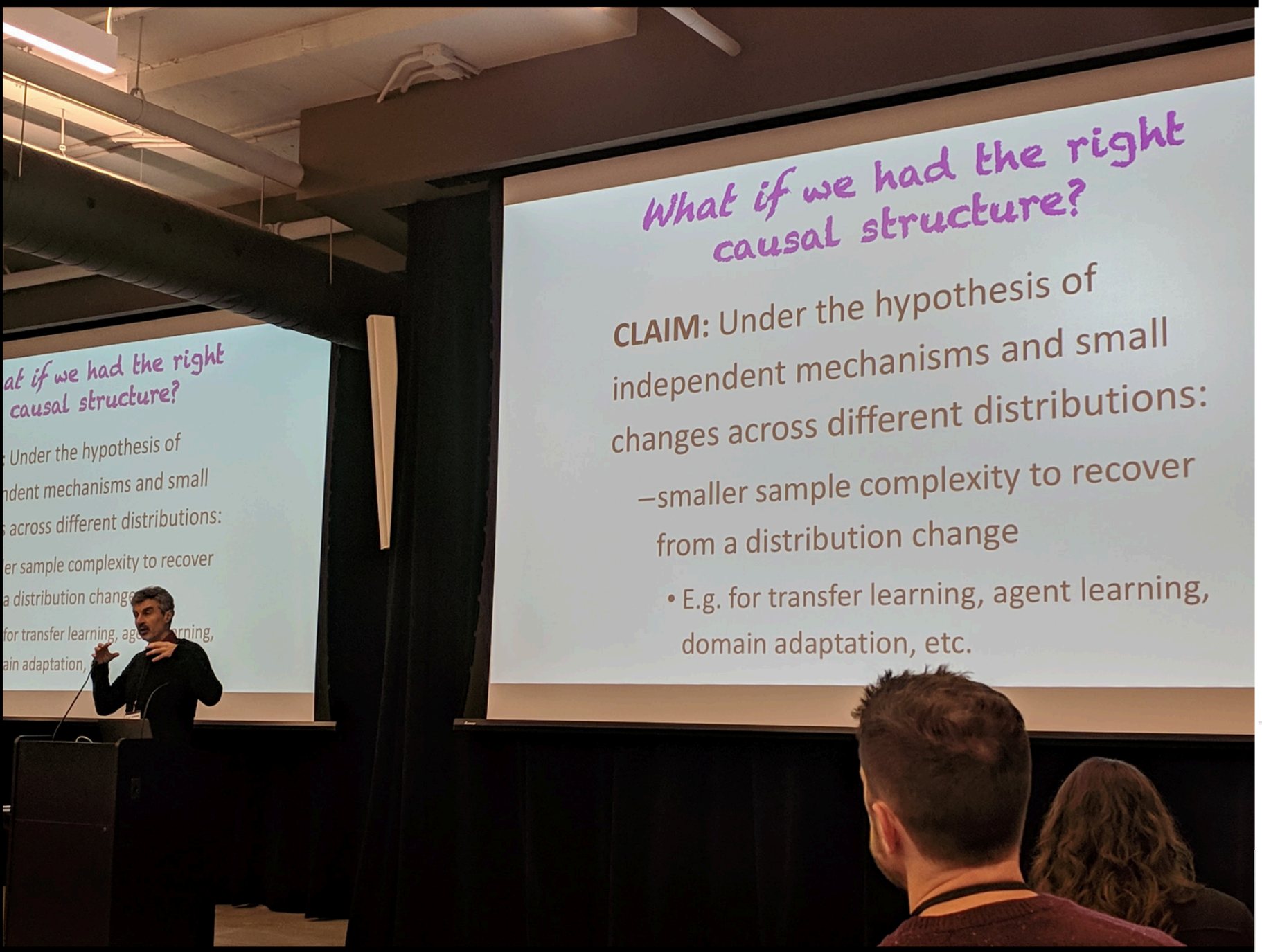
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3



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



https://colab.research.google.com/drive/1rjjjA7teiZVHJCMTVD8KlZNu3EjS7Dmu#scrollTo=TlzzvcGOZdvW
https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/



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

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 😉

Inductive Bias

Compositionality

Symmetry

Causality



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.

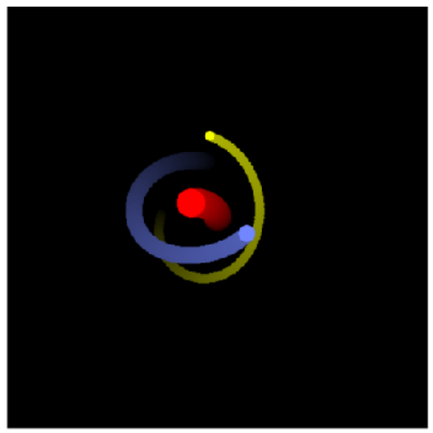
- Objects and relations reflect *decisions* made by evolution, experience, and task demands about how to represent the world in an *efficient and useful way*
- Intelligence is about *model-building*, beyond just recognizing patterns (Tenenbaum)
- *Combinatorial generalization* via abstraction and compositionality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

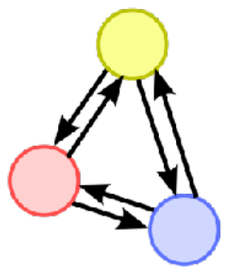
Peter Battaglia

Insight of data generating process informs inductive bias on architecture

Physical systems as graphs

n-body

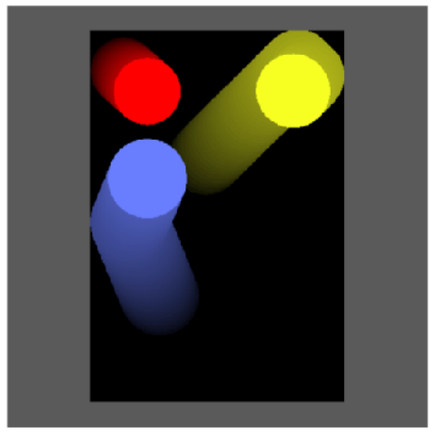


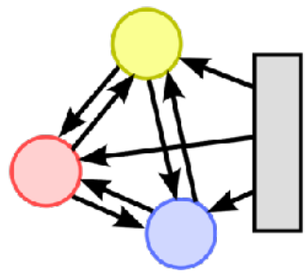


Nodes: bodies

Edges: gravitational forces

Balls

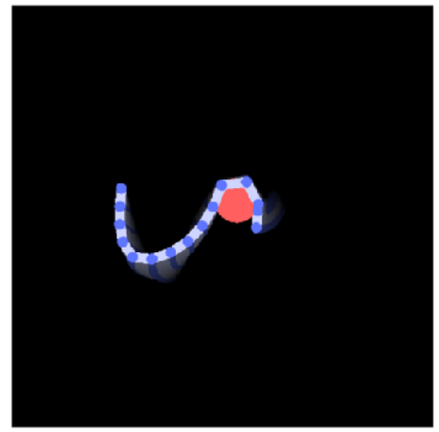


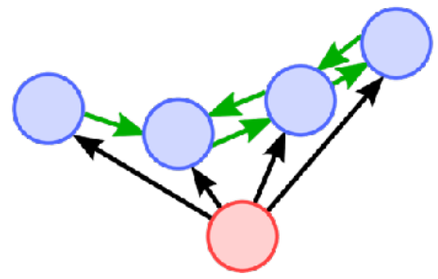


Nodes: balls

Edges: rigid collisions between balls, and walls

String





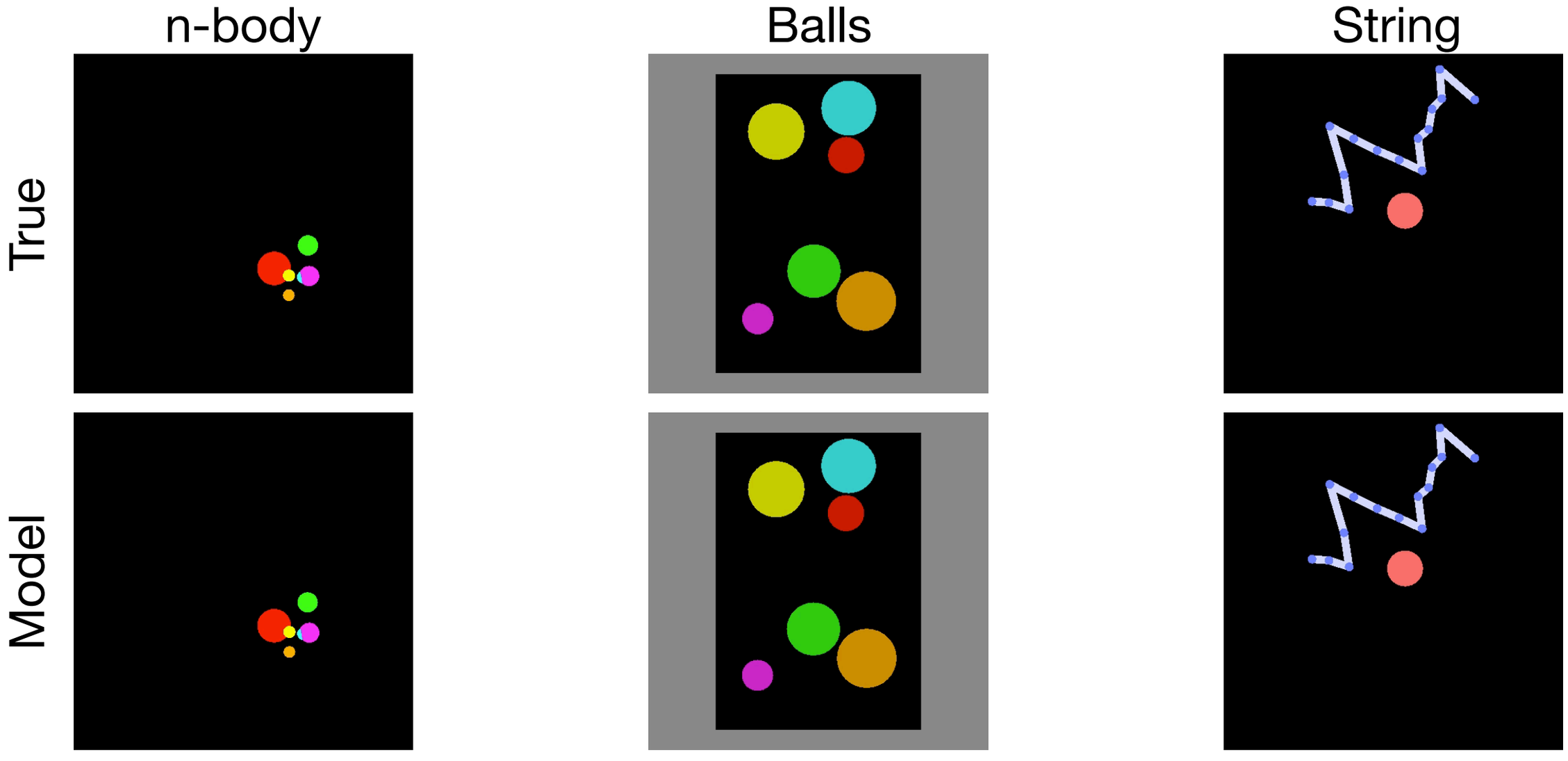
Nodes: masses

Edges: springs and rigid collisions

Battaglia et al., 2016, NeurIPS

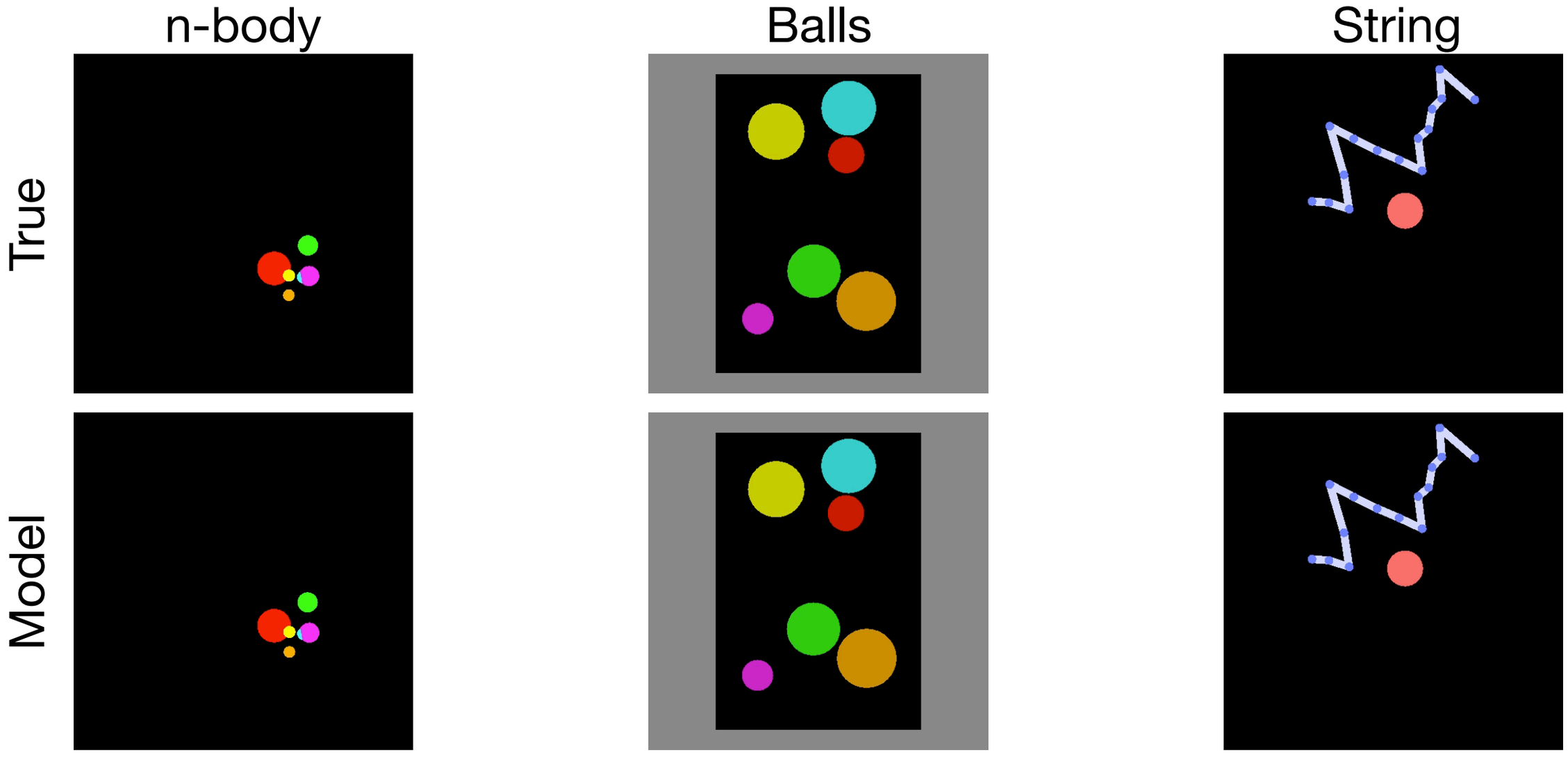
Slides borrowed with permission from Peter Battaglia based on Battaglia et al., 2016, NeurIPS

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



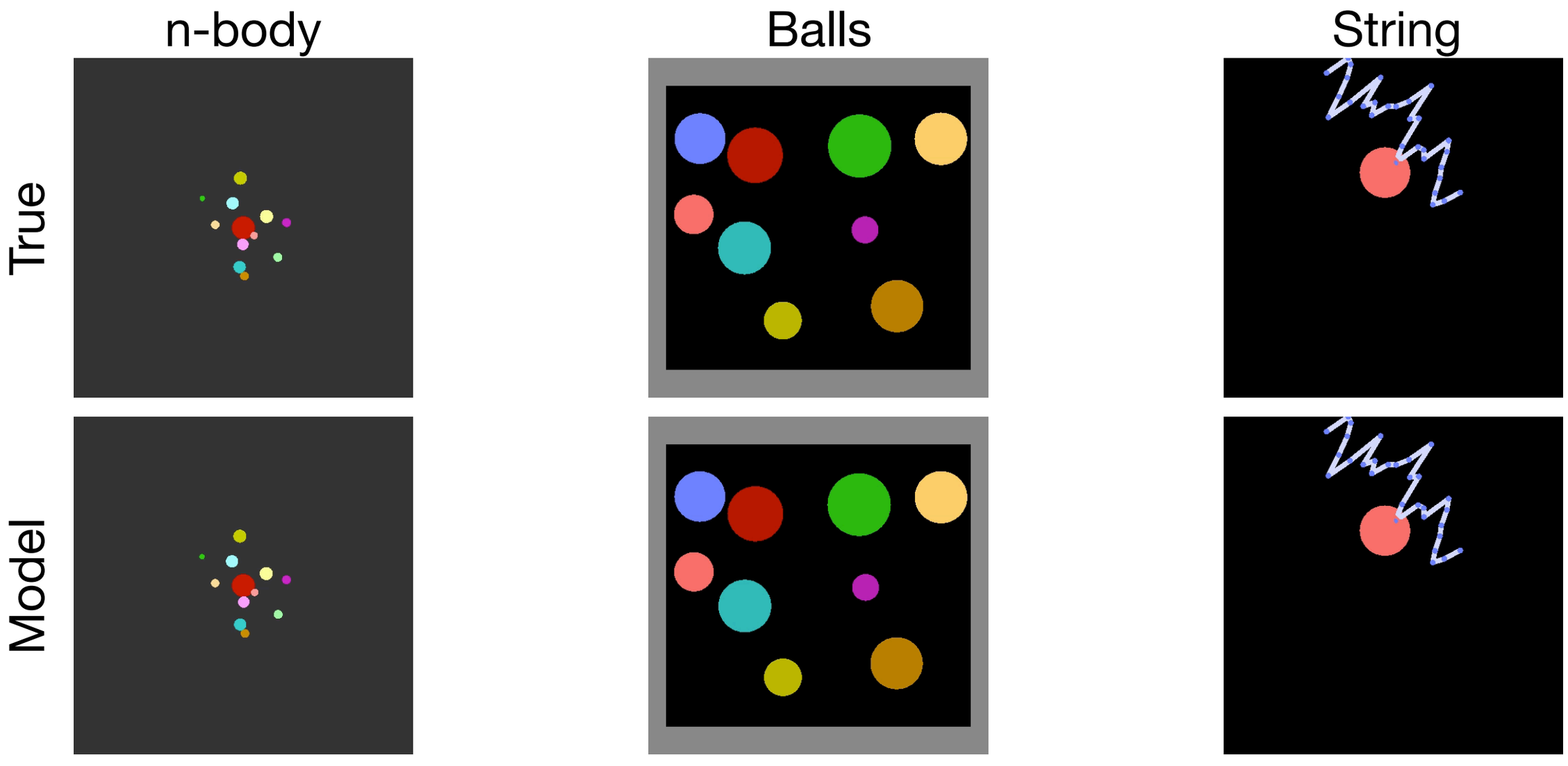
Slides borrowed with permission from Peter Battaglia based on Battaglia et al., 2016, NeurIPS

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



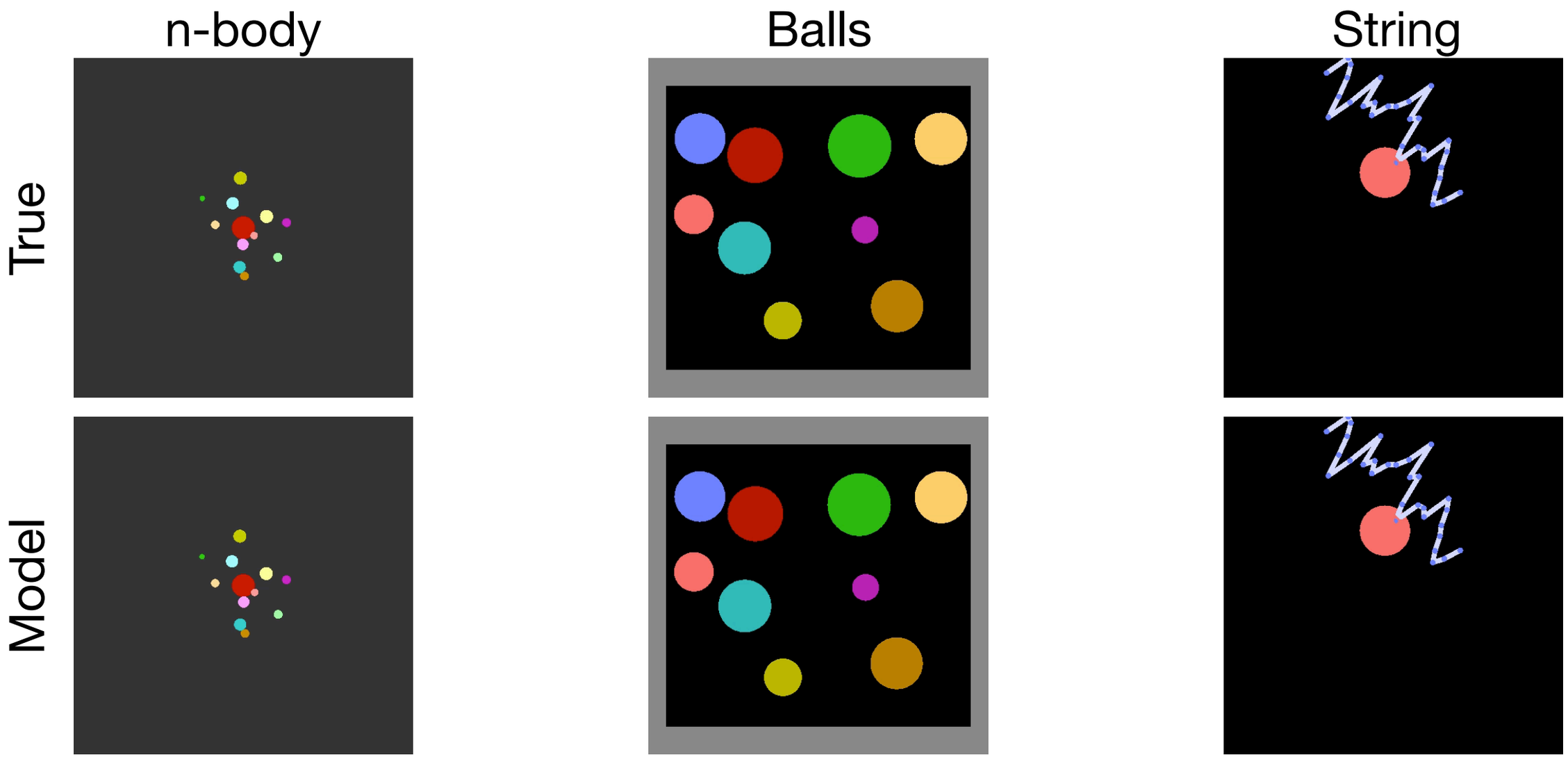
Slides borrowed with permission from Peter Battaglia based on Battaglia et al., 2016, NeurIPS

Zero shot generalisation to larger systems



Slides borrowed with permission from Peter Battaglia based on Battaglia et al., 2016, NeurIPS

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.

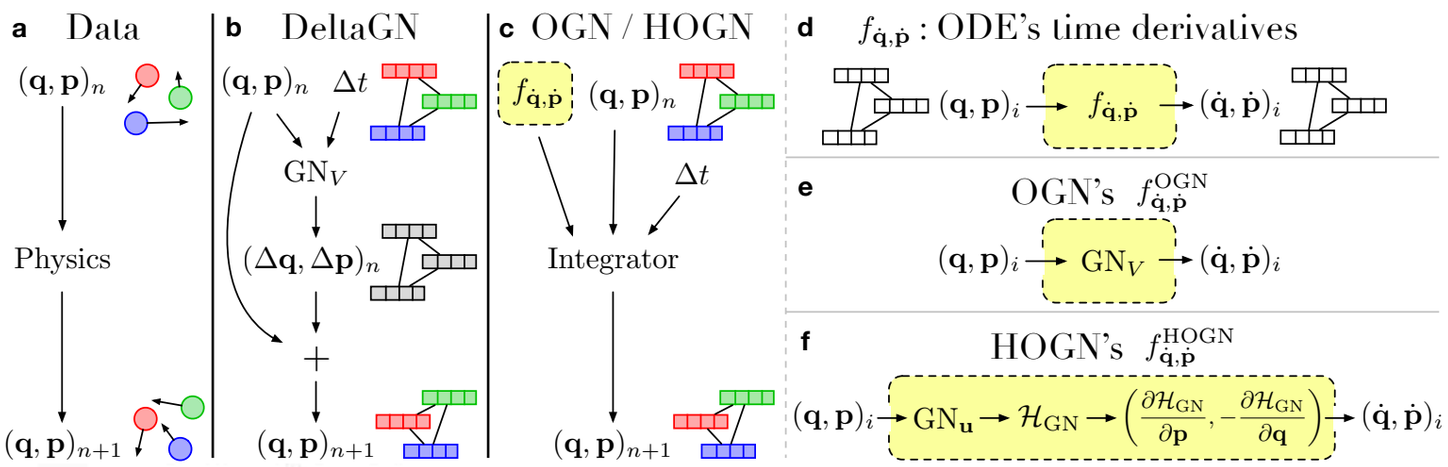
Hamiltonian Graph Networks with ODE Integrators

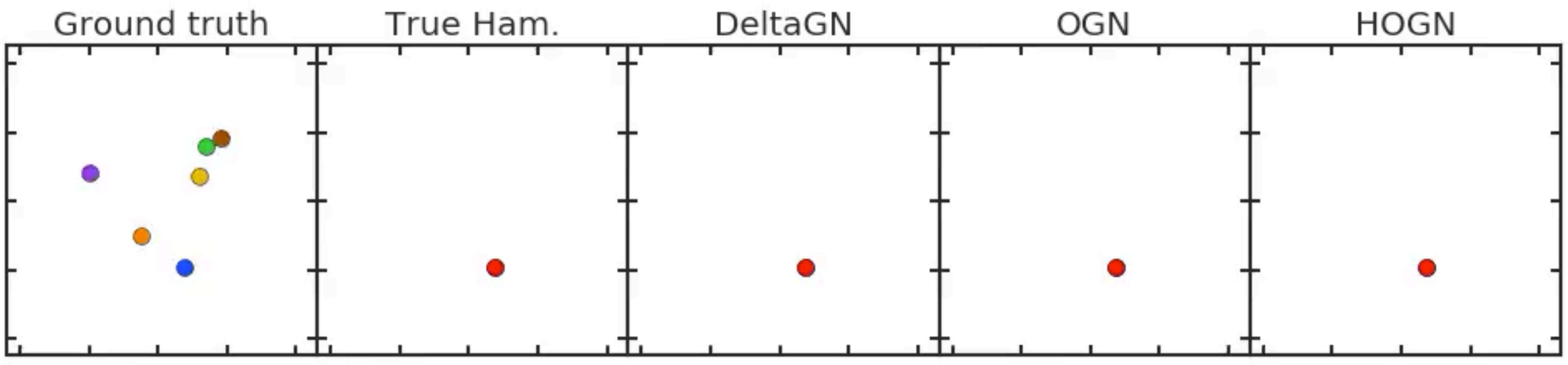
Alvaro Sanchez-Gonzalez
DeepMind
London, UK
alvarosg@google.com

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.

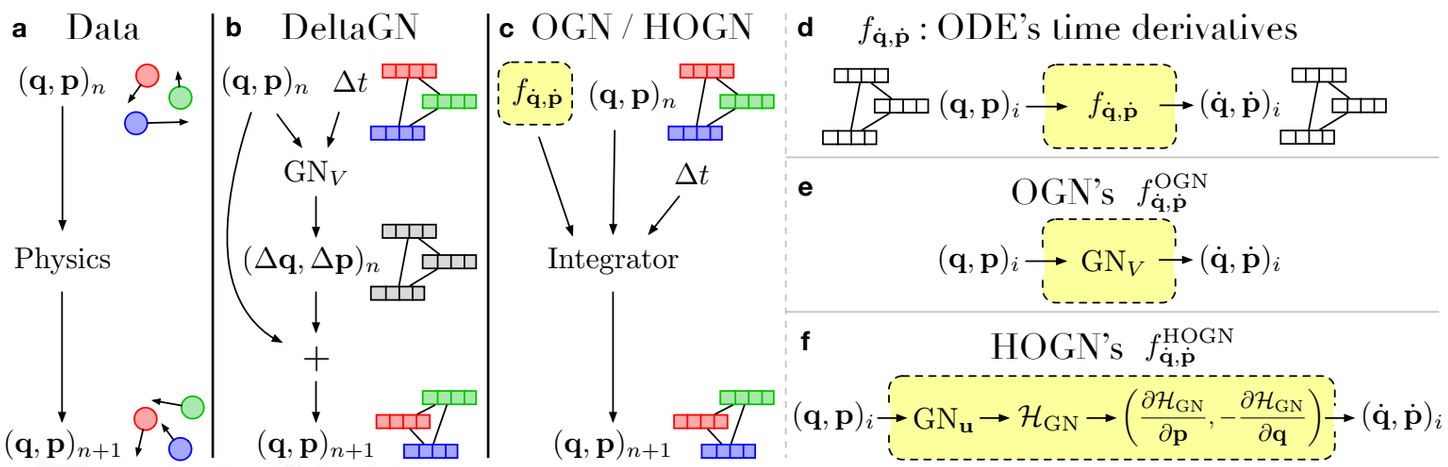
Hamiltonian Graph Networks with ODE Integrators

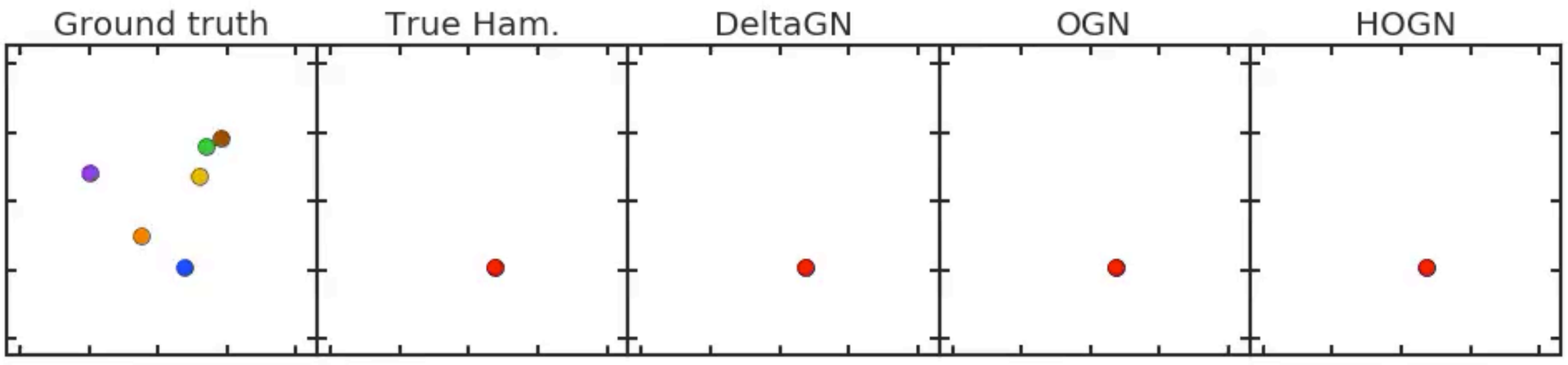
Alvaro Sanchez-Gonzalez
DeepMind
London, UK
alvarosg@google.com

Victor Bapst
DeepMind
London, UK
vbapst@google.com

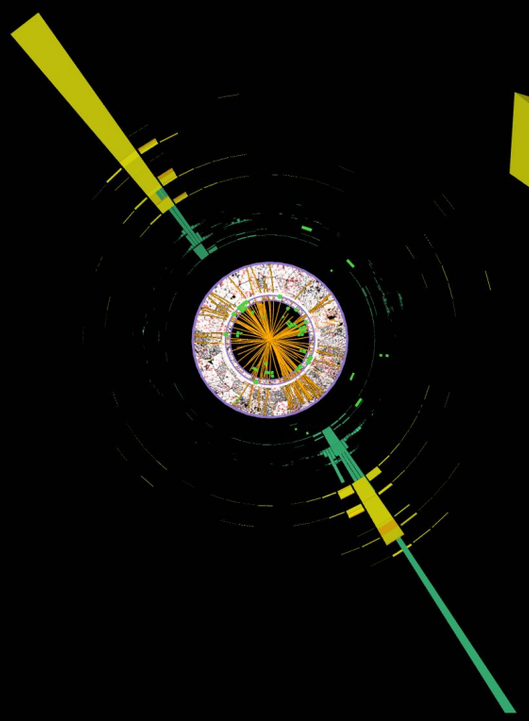
Kyle Cranmer
NYU
New York, USA
kc90@nyu.edu

Peter Battaglia
DeepMind
London, UK
peterbattaglia@google.com





JETS

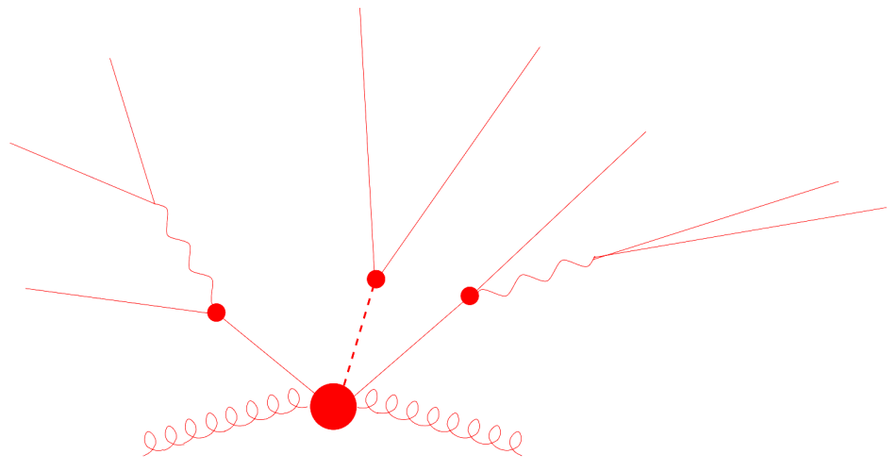


Run: 329716
Event: 857582452
2017-07-14 10:48:51 CEST

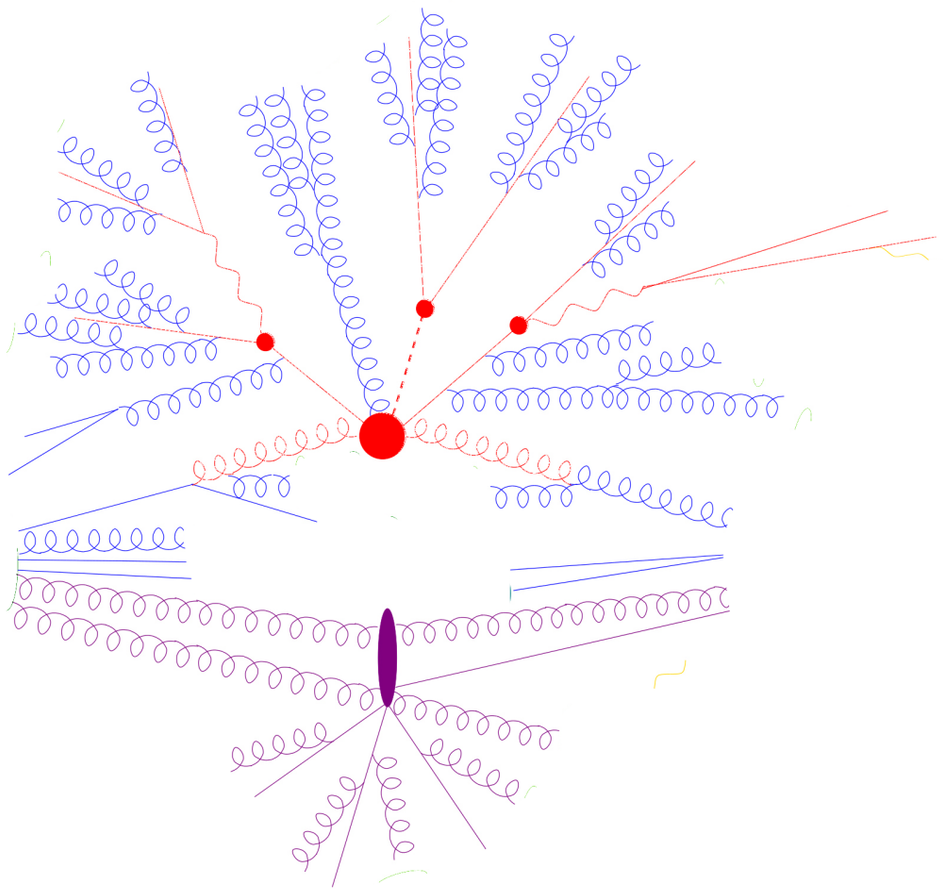
ATLAS
EXPERIMENT

CAUSAL, GENERATIVE MODEL FOR JETS

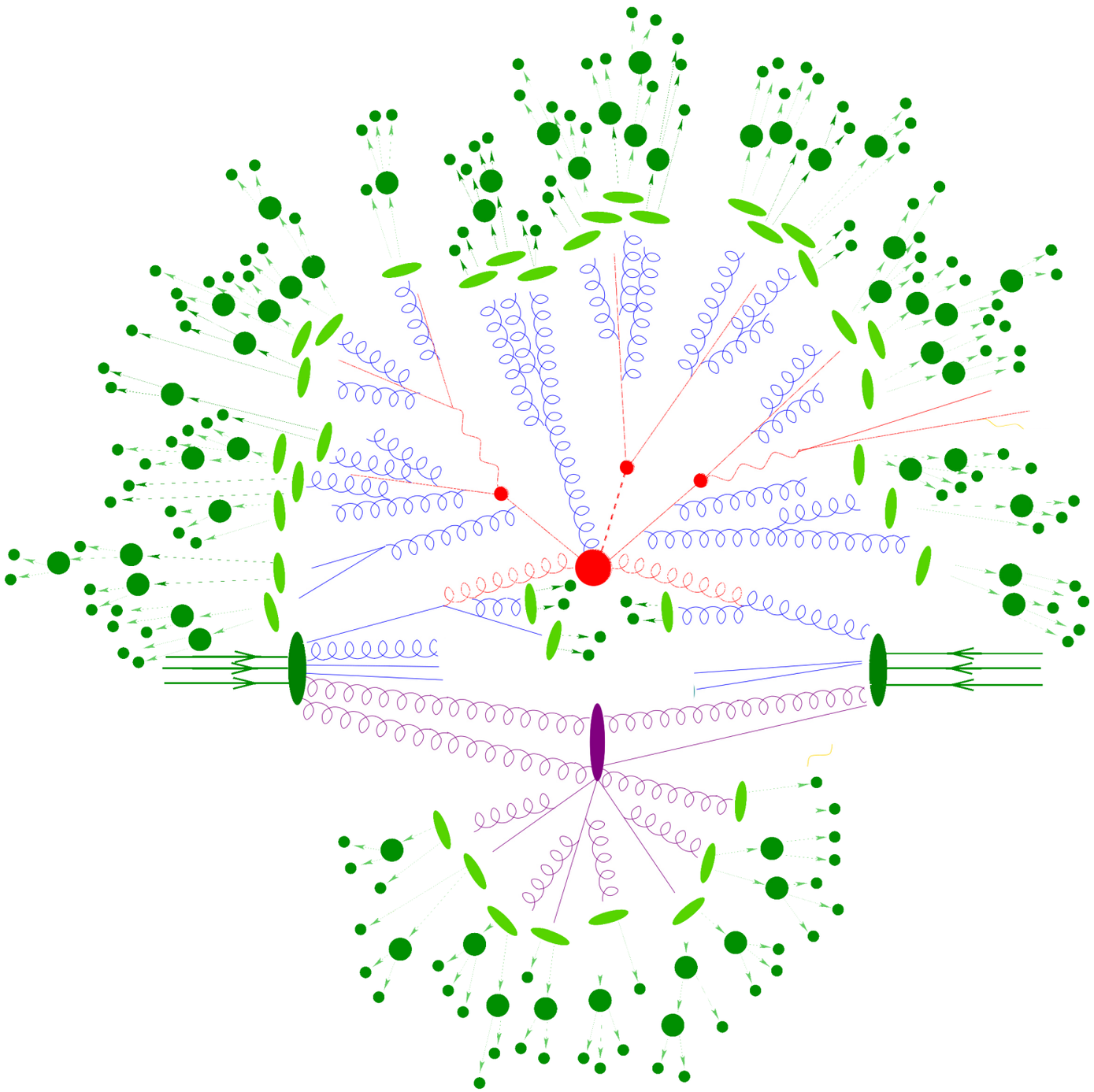
CAUSAL, GENERATIVE MODEL FOR JETS



CAUSAL, GENERATIVE MODEL FOR JETS

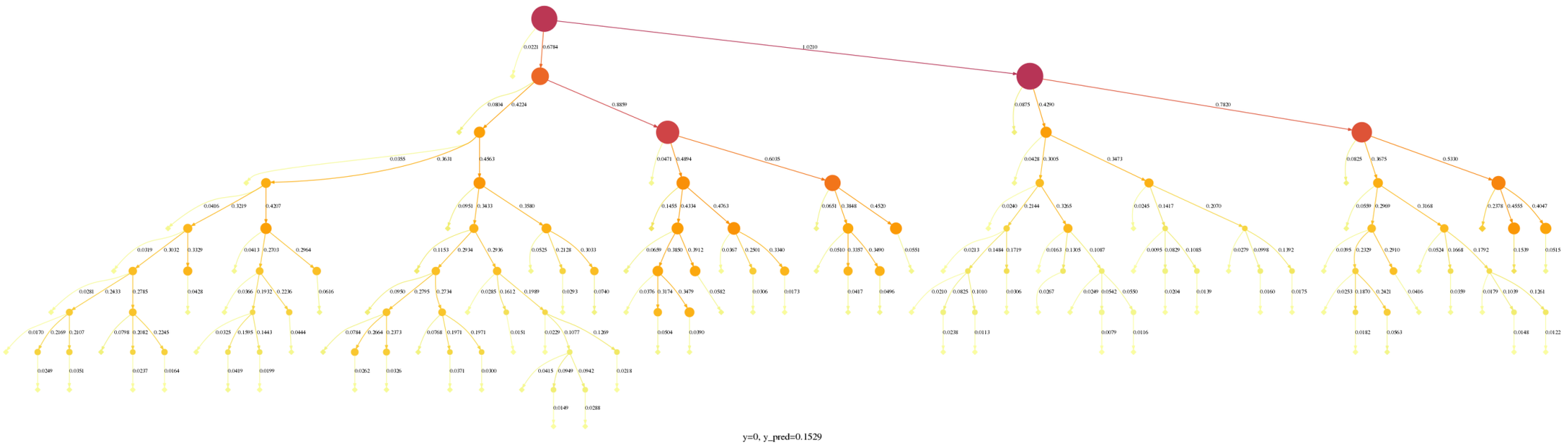


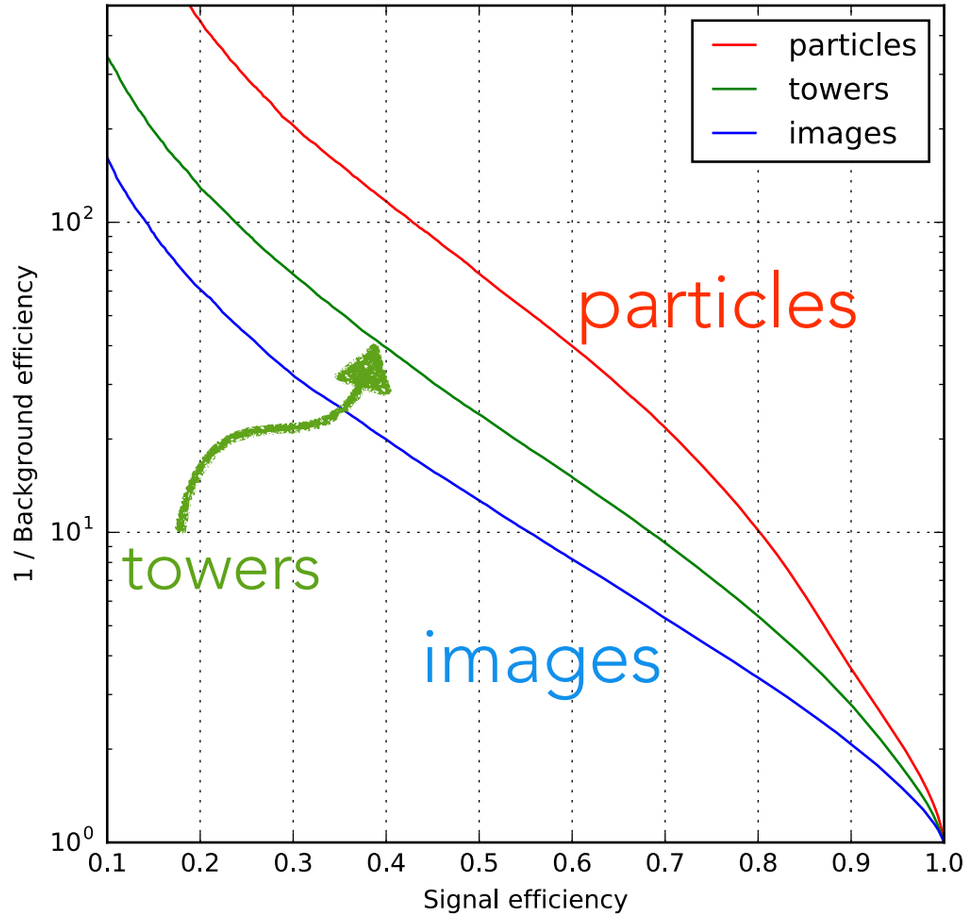
CAUSAL, GENERATIVE MODEL FOR JETS



QCD-INSPIRED RECURSIVE NEURAL NETWORKS

Insight of data generating process informs inductive bias on architecture





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

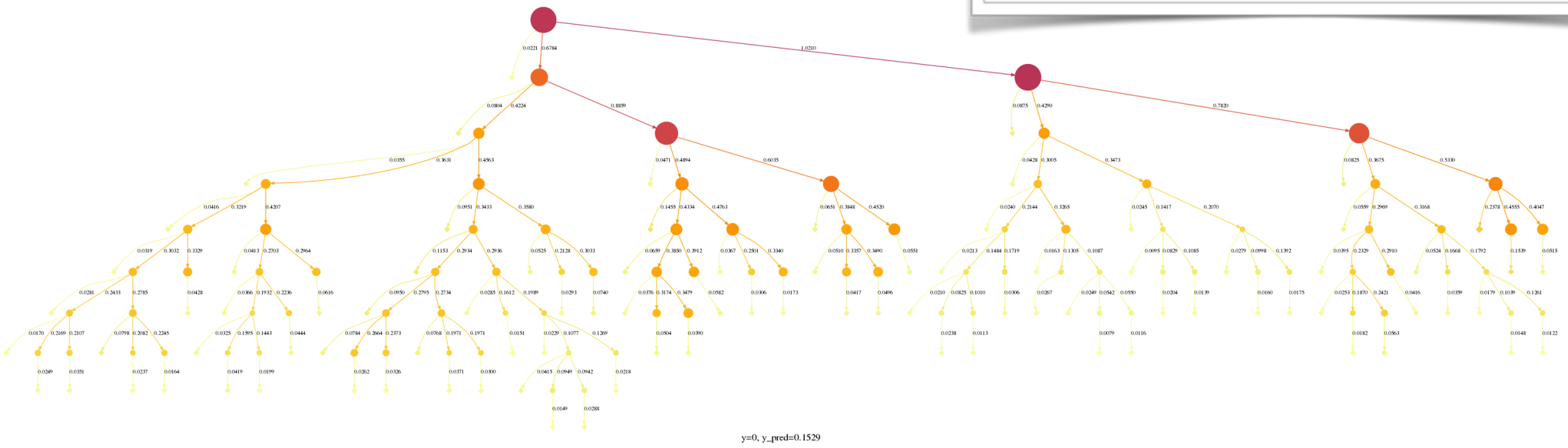
Louppe, Cho, Becot, KC [arXiv:1702.00748]

INDUCTIVE BIAS

arXiv:1902.09914

The Machine Learning Landscape of Top Taggers

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

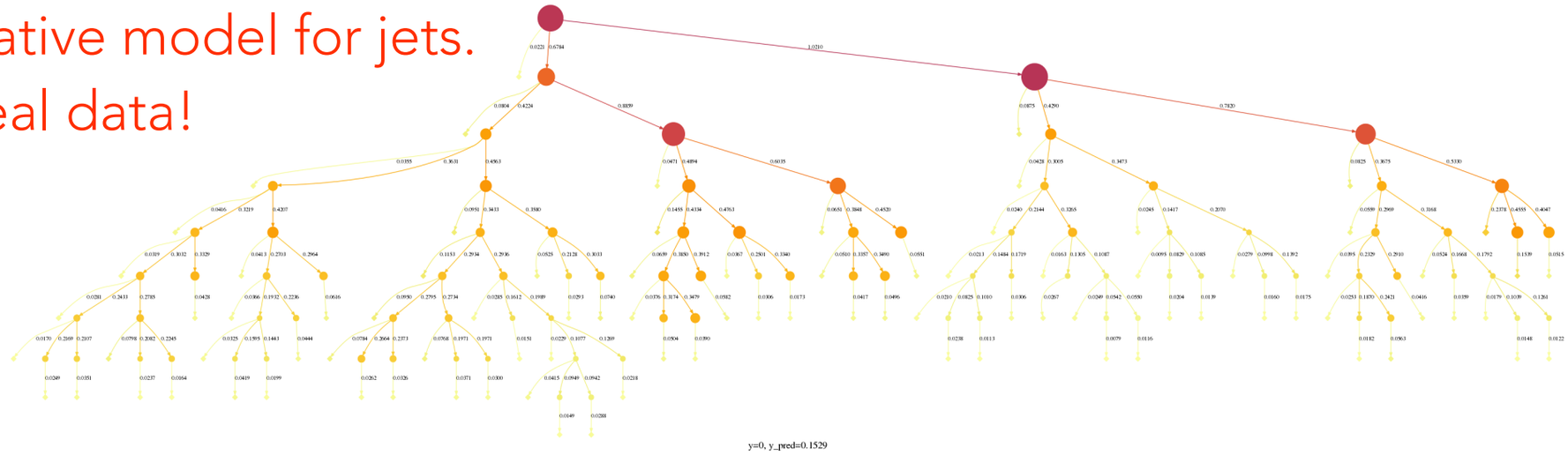


	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	966±18	610k
ResNeXt [30]	0.984	0.936	1122±47	1246±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	378± 5	391 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	802±12	783±13	57k
Multi-body N-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.

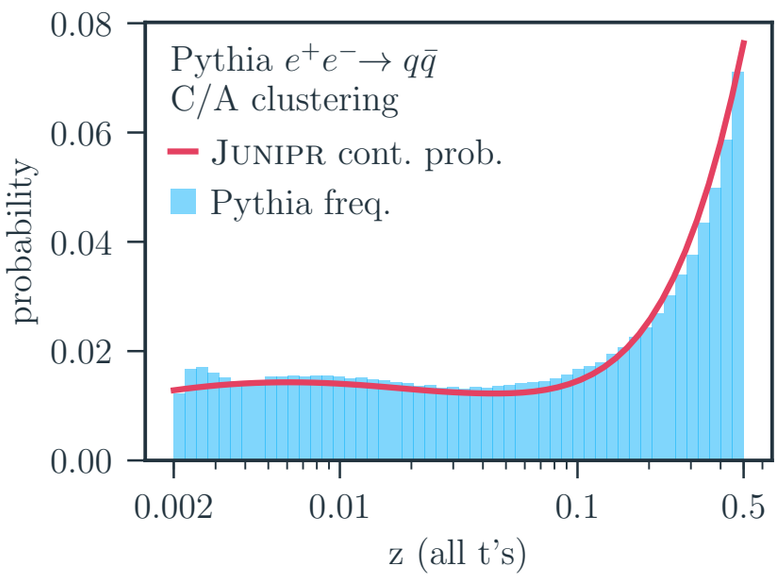
Can train on real data!

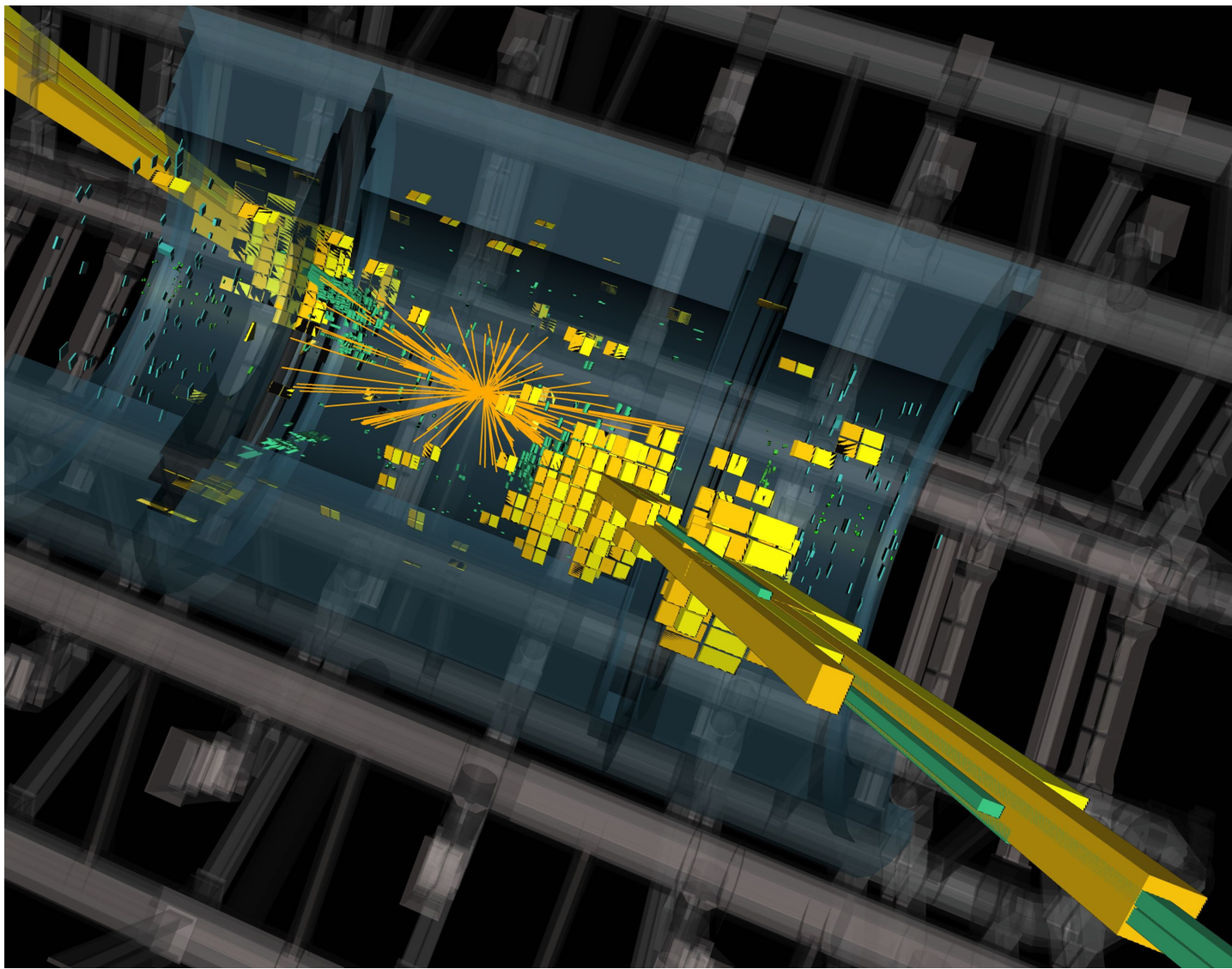


tractable likelihood

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

… and it is interpretable



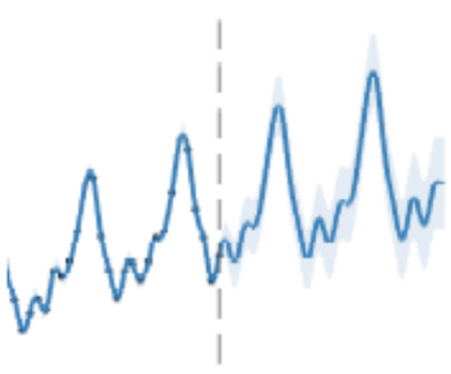


Andreassen, Feige, Frye, Schwartz arXiv:1804.09720

COMPOSITIONALITY

Vocabulary of kernels + grammar for composition

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

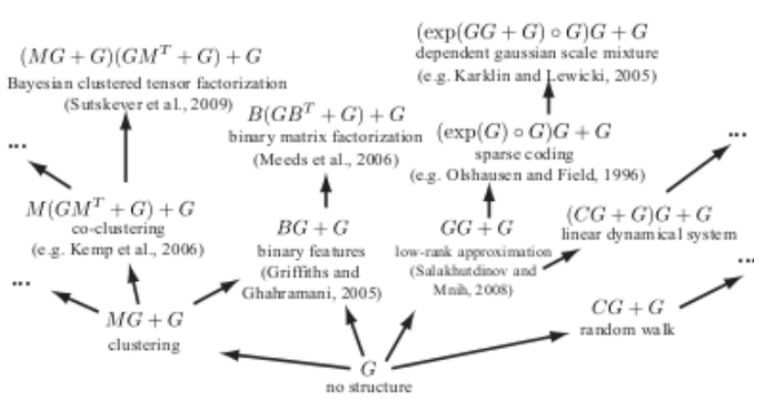


Structure Discovery in Nonparametric Regression through Compositional Kernel Search

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, Zoubin Ghahramani

International Conference on Machine Learning, 2013

pdf | code | poster | bibtex

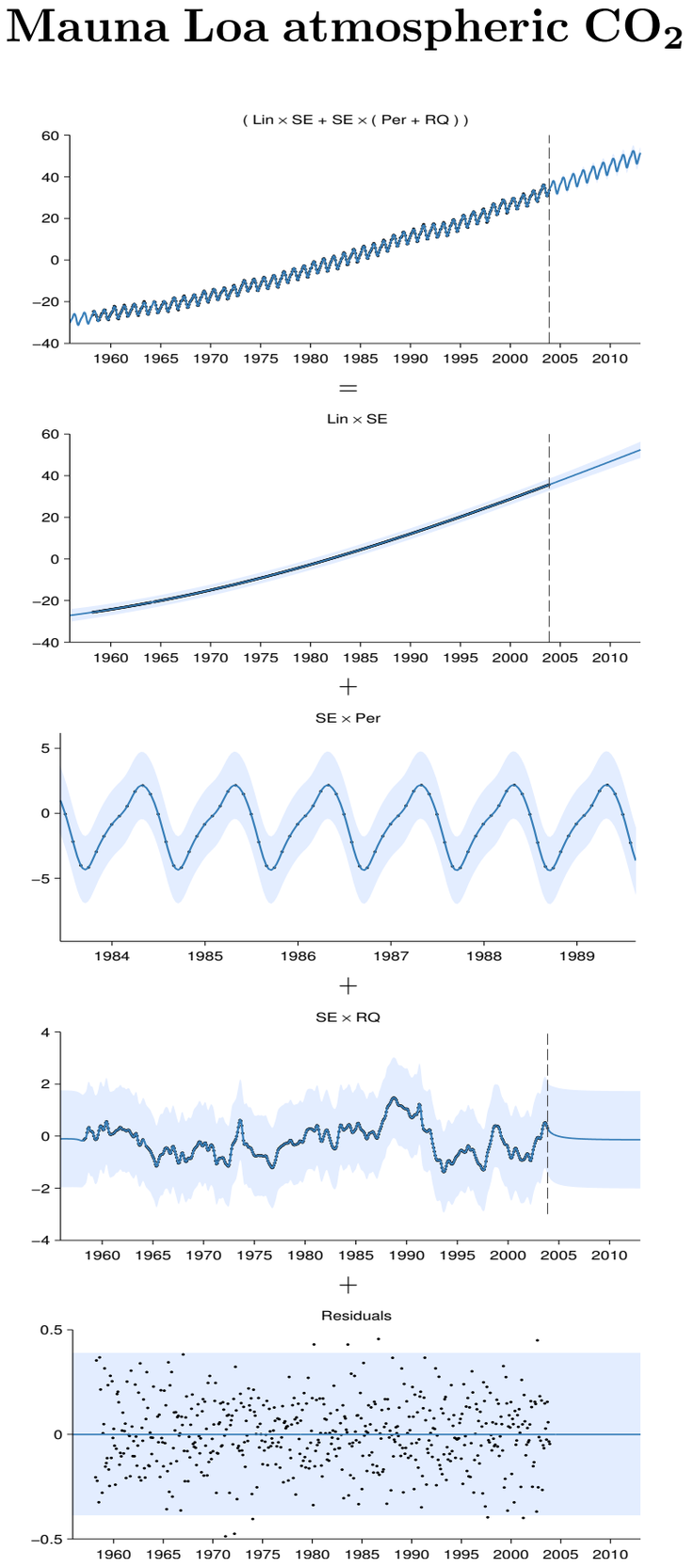


Exploiting compositionality to explore a large space of model structures

Roger Grosse, Ruslan Salakhutdinov, William T. Freeman, Joshua B. Tenenbaum

Conference on Uncertainty in Artificial Intelligence, 2012

pdf | code | bibtex

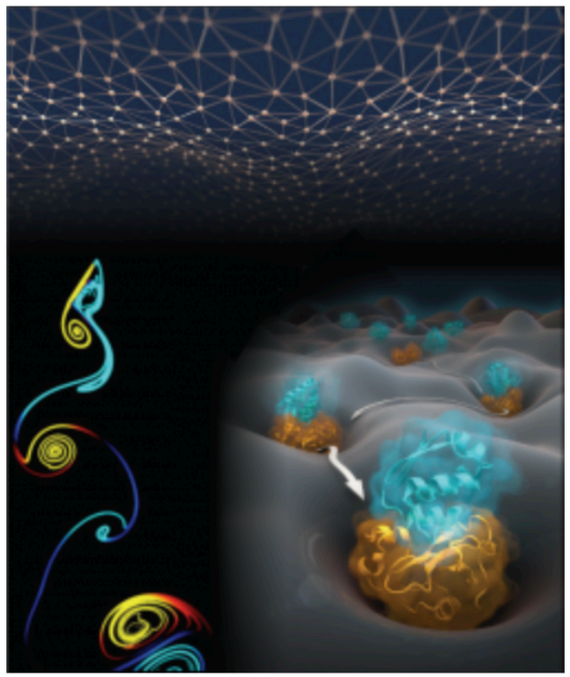


Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019

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.

Teacher → Causal, Generative Model (Simulator)

Richer set of problems can be investigated.

Insight of data generating process informs inductive bias on architecture



Backup

HOW DO YOU USE THE GOLD?

arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
arXiv:1808.00973
physics.aps.org/articles/v11/90

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

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

One can show it is minimized by

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

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

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

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

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

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

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

One can show it is minimized by

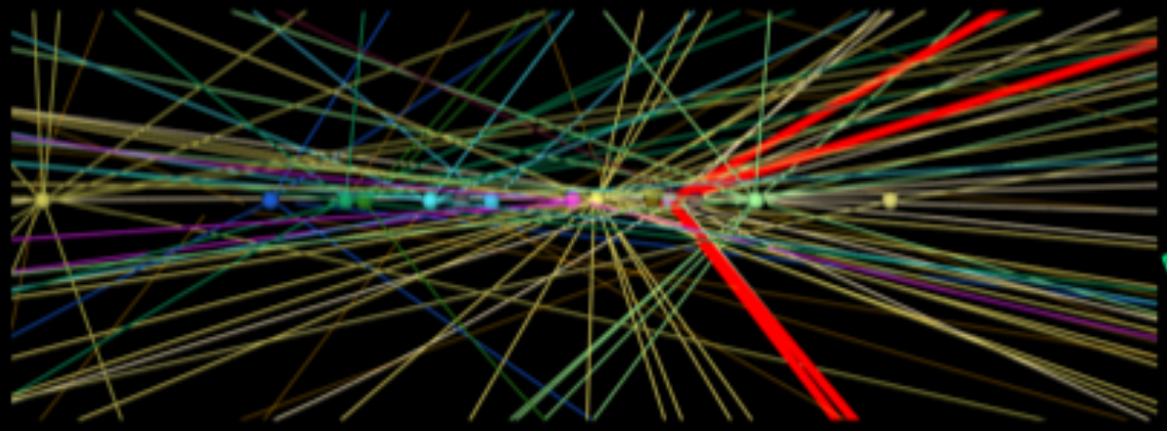
$$t(x|\theta_0) = \arg\min_{\hat{t}(x|\theta_0)} L_t[\hat{t}(x|\theta_0)] \; .$$

Again, we implement this minimization through machine learning

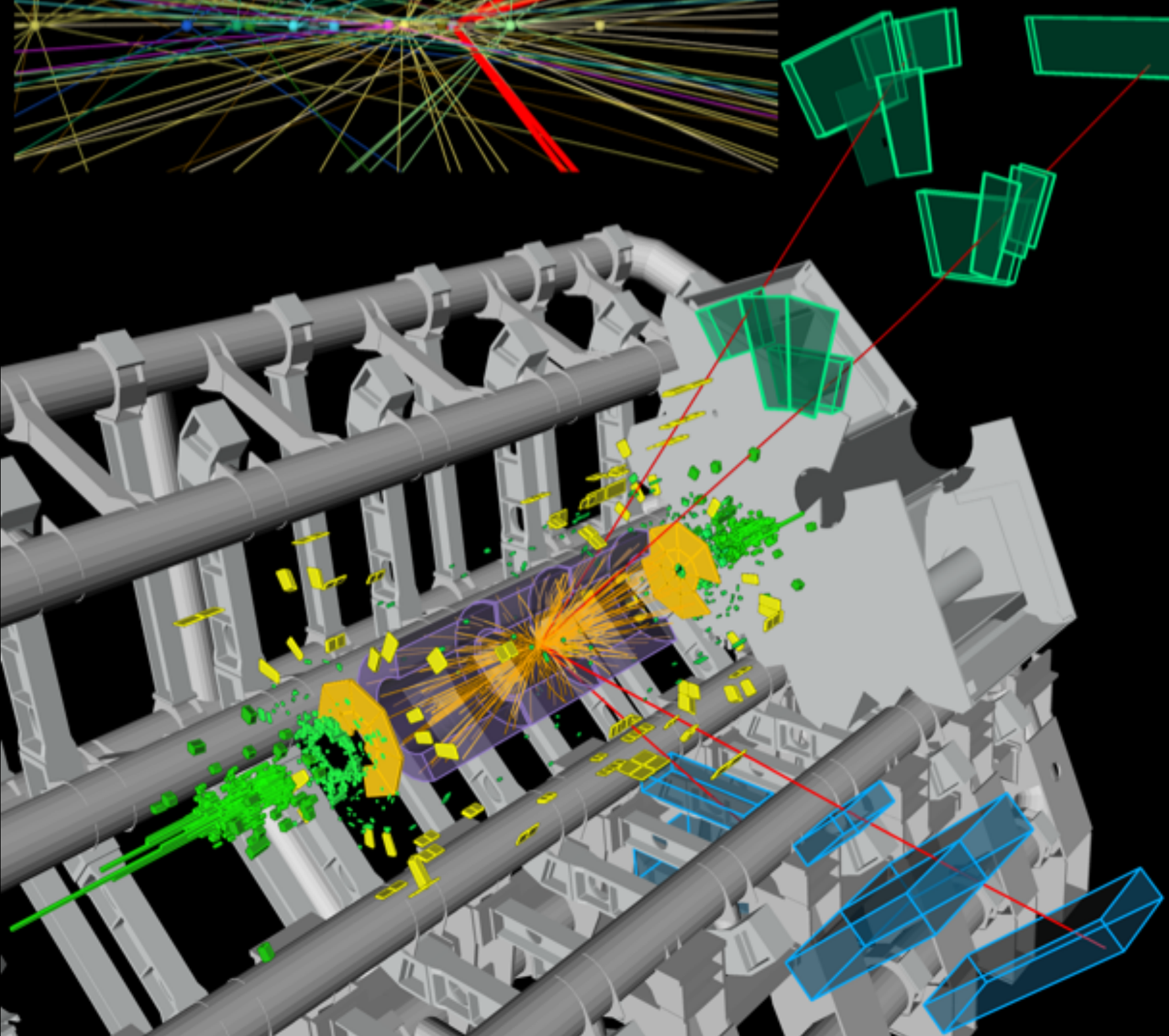
We want score

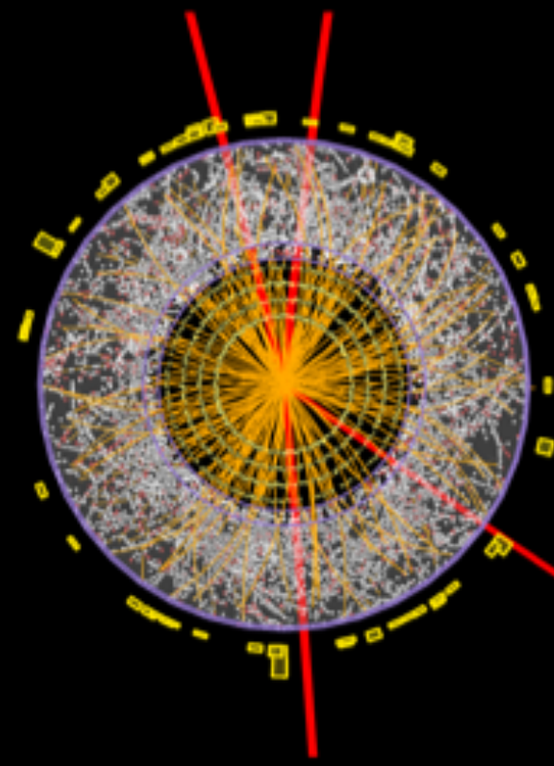
$$t(x|\theta_0) \equiv \nabla_\theta \log p(x|\theta)\Big|_{\theta_0}$$

THE HIGGS BOSON



ATLAS EXPERIMENT
http://atlas.ch





Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

THE CAUSAL, GENERATIVE MODEL

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

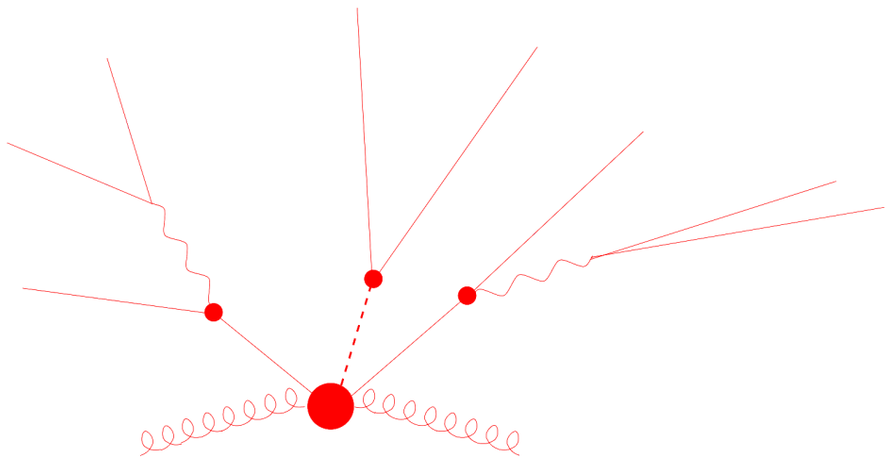
THE CAUSAL, GENERATIVE MODEL

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L \;+\; \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



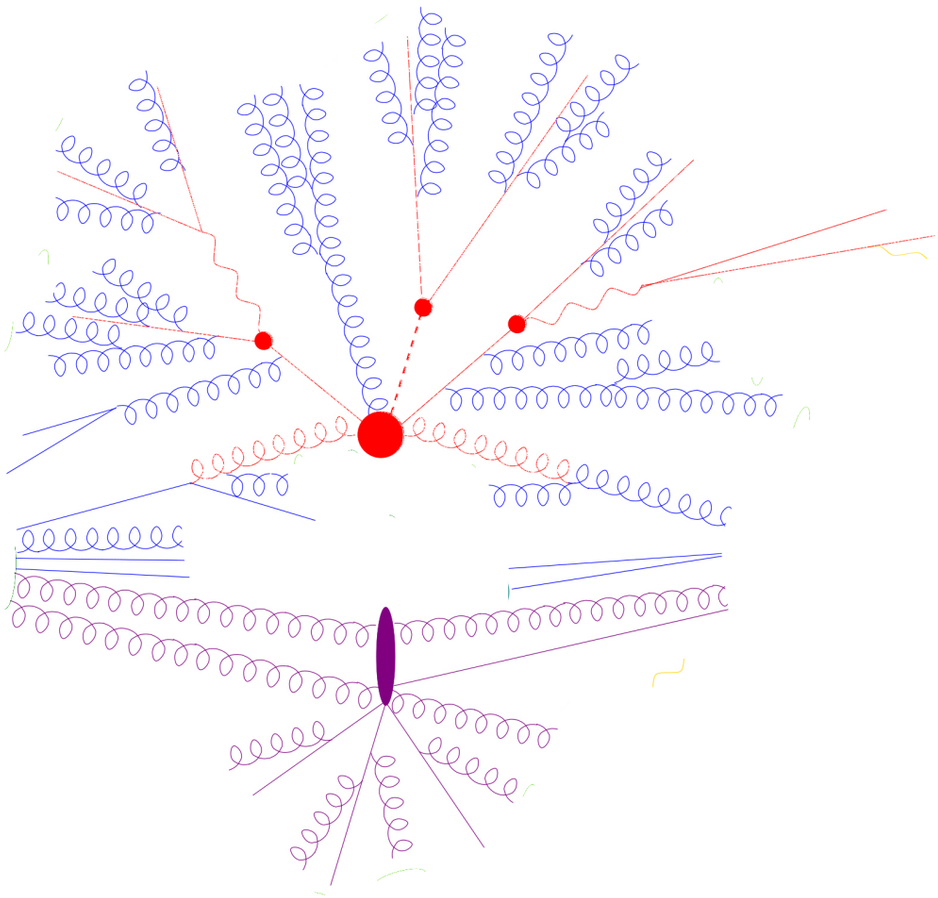
THE CAUSAL, GENERATIVE MODEL

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

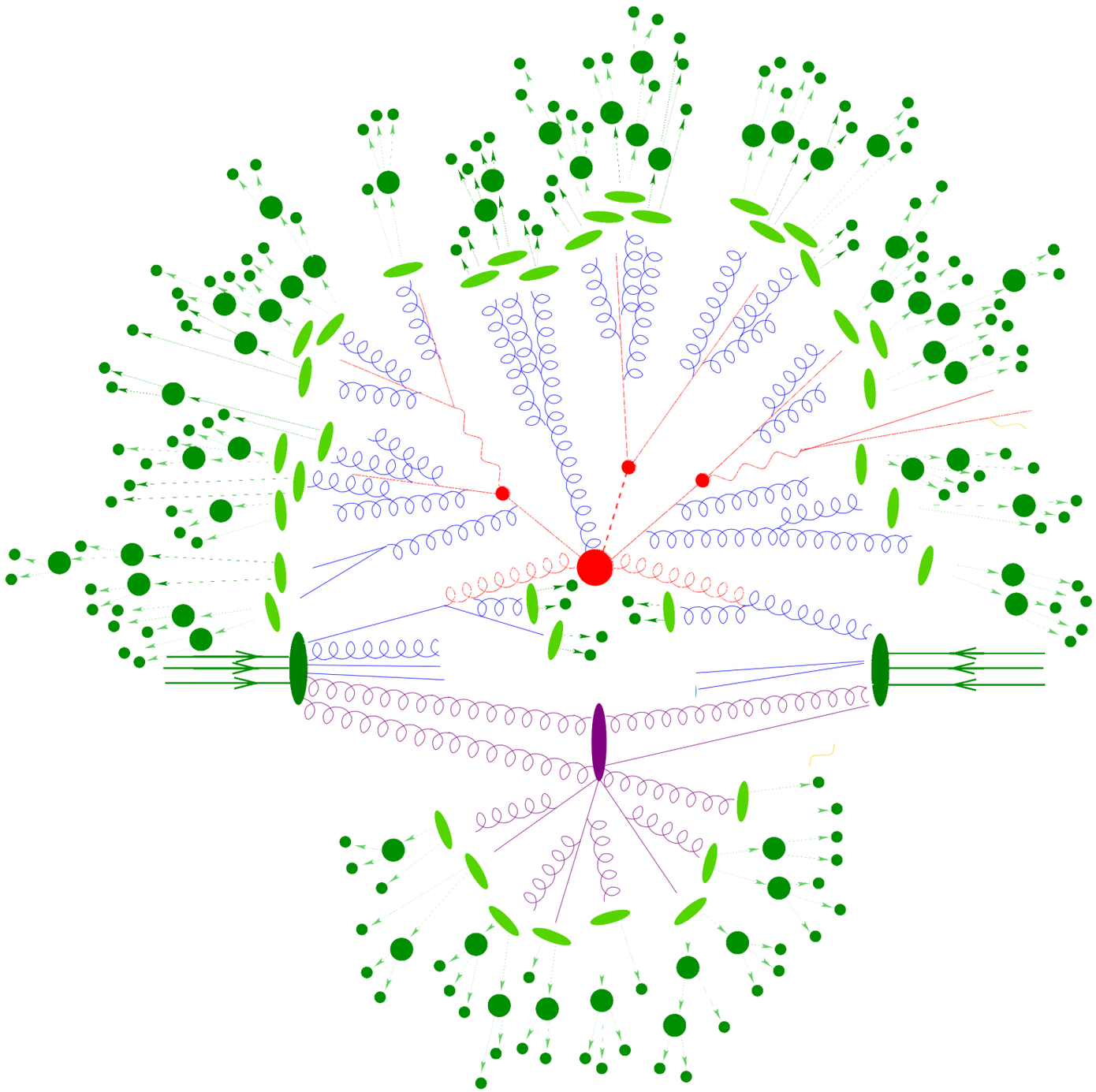
$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



THE CAUSAL, GENERATIVE MODEL



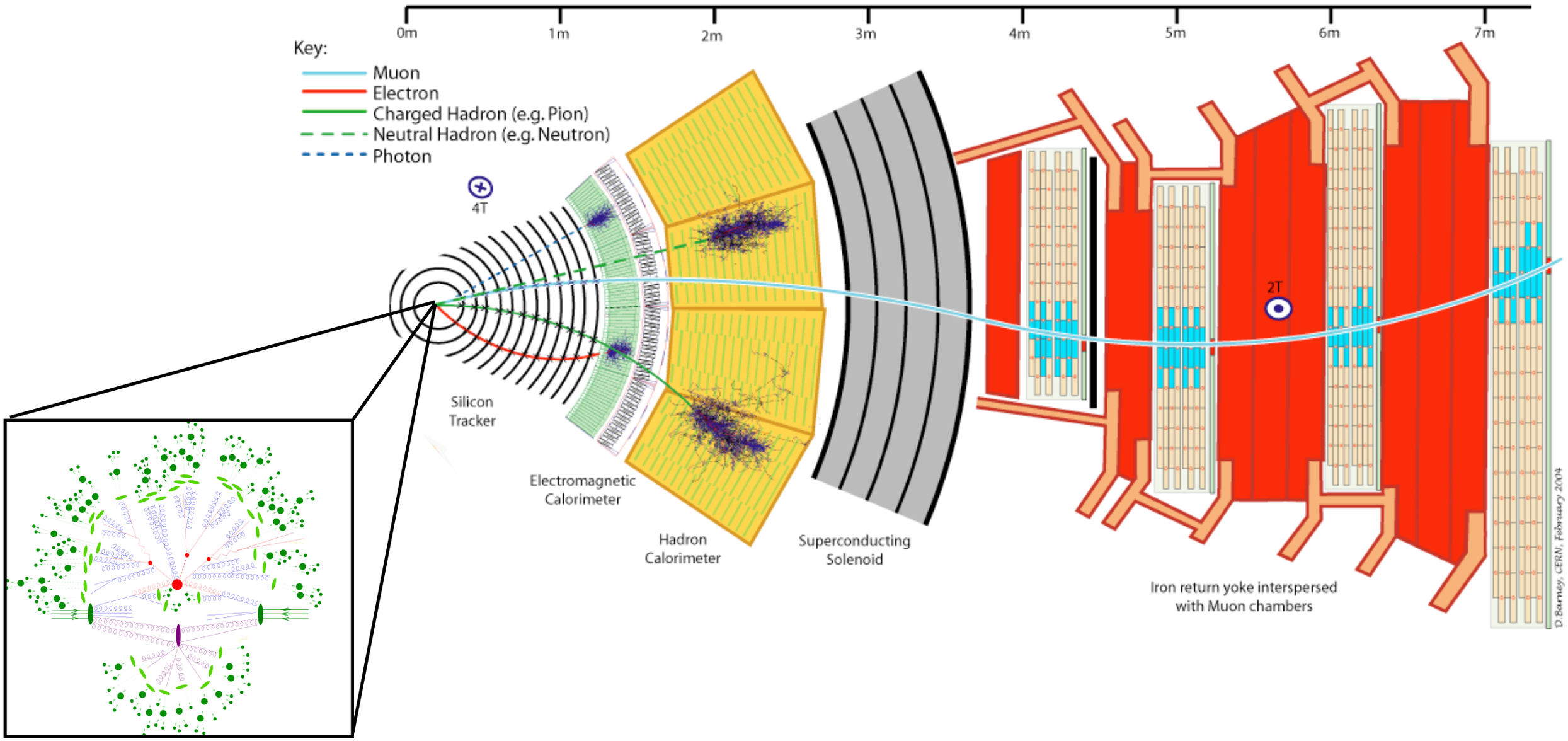
$$
\begin{aligned}
\mathcal{L}_{SM} = & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu}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

THE CAUSAL, GENERATIVE MODEL

Conceptually: Prob(detector response | particles)

Implementation: Monte Carlo integration over micro-physics

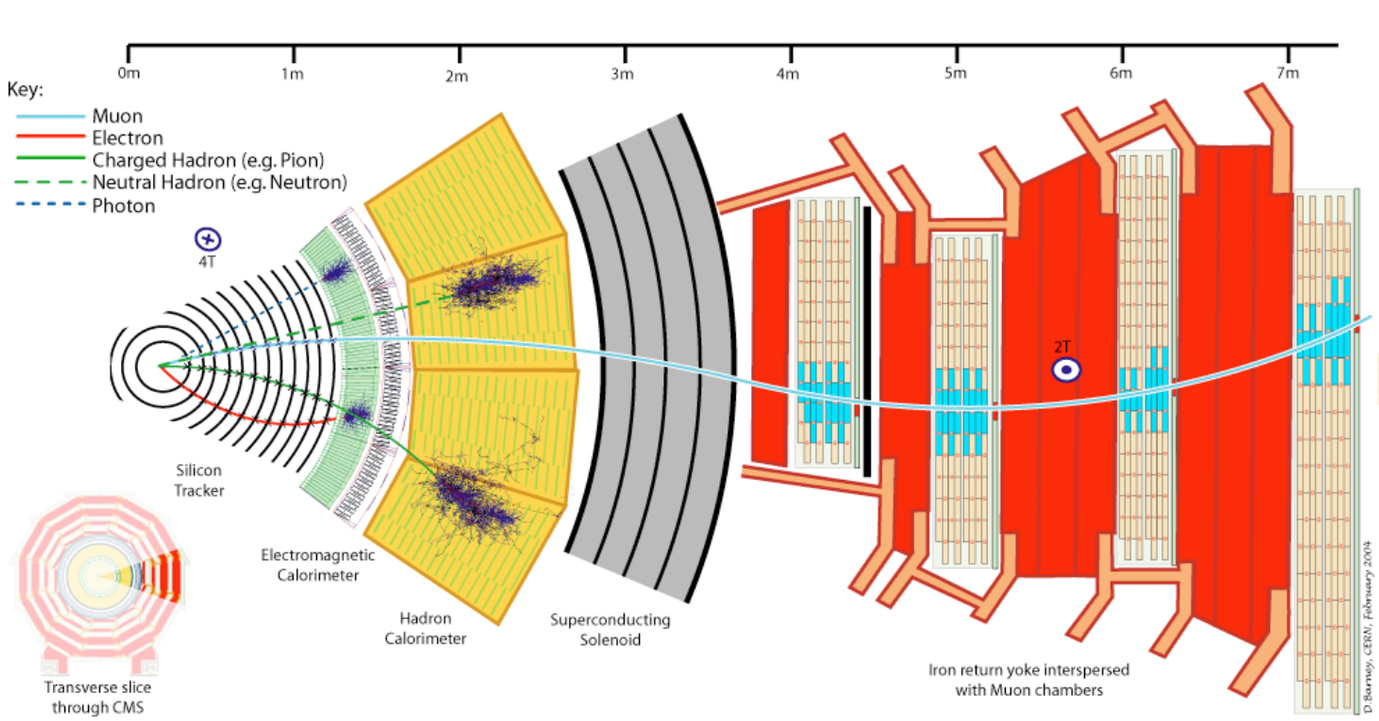
Consequence: evaluation of the likelihood is intractable

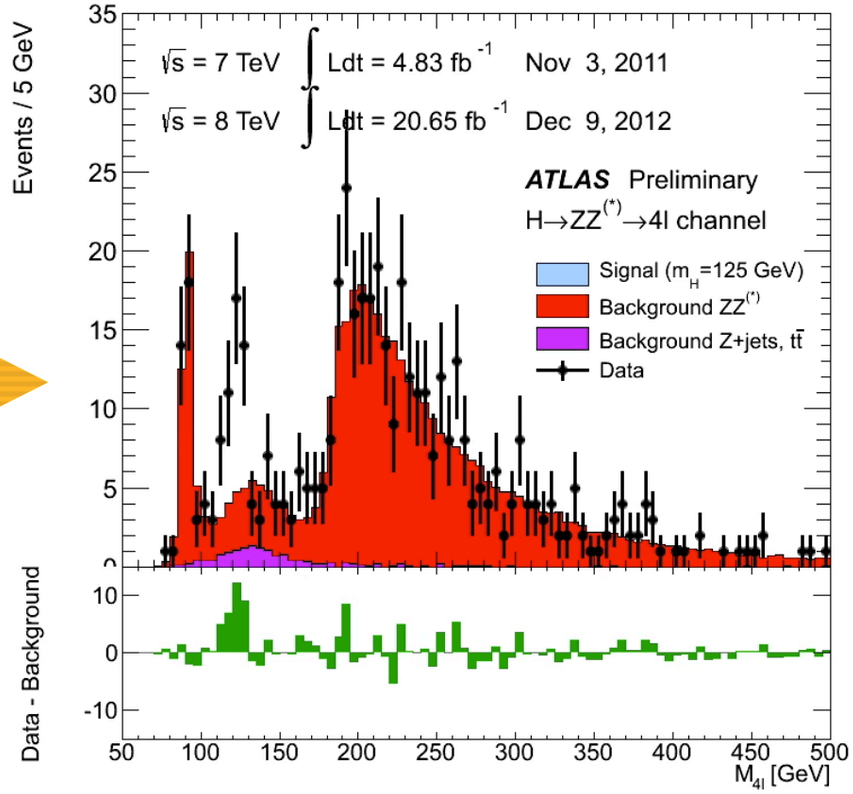


10^8 SENSORS → 1 REAL-VALUED QUANTITY

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)

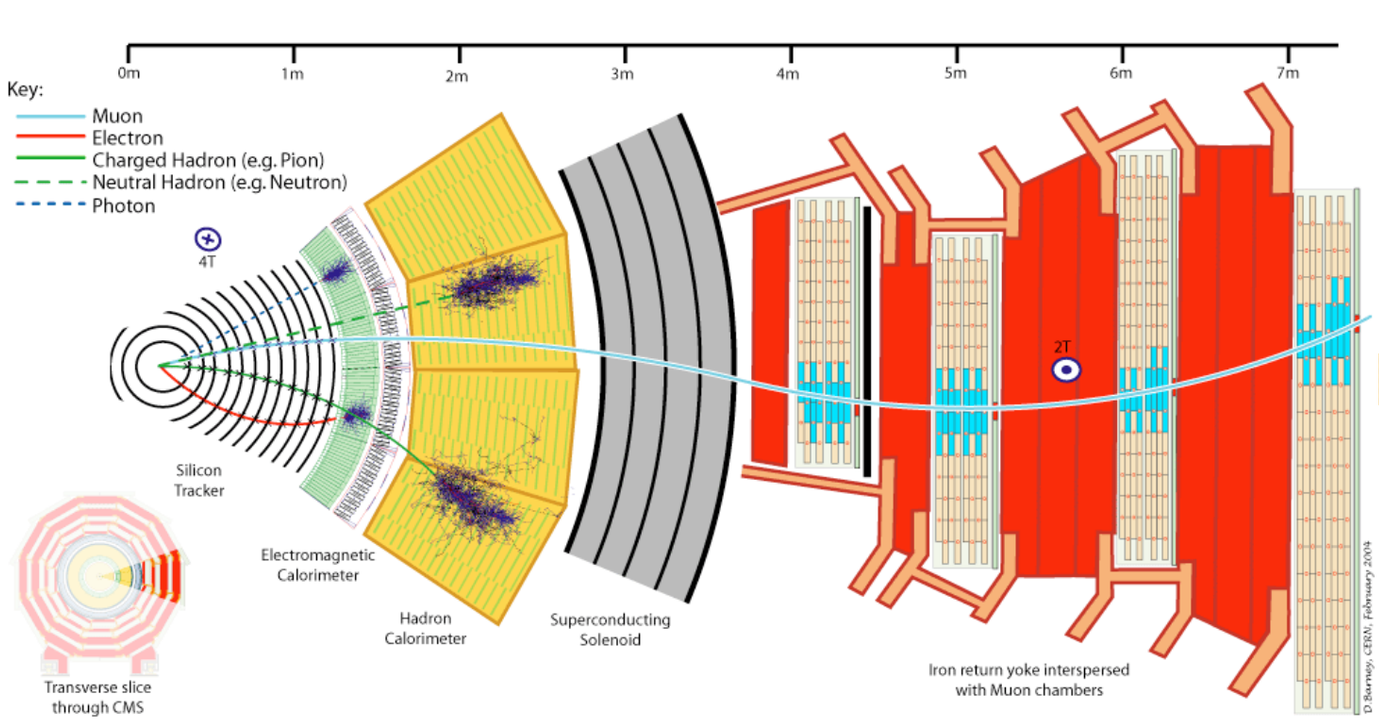


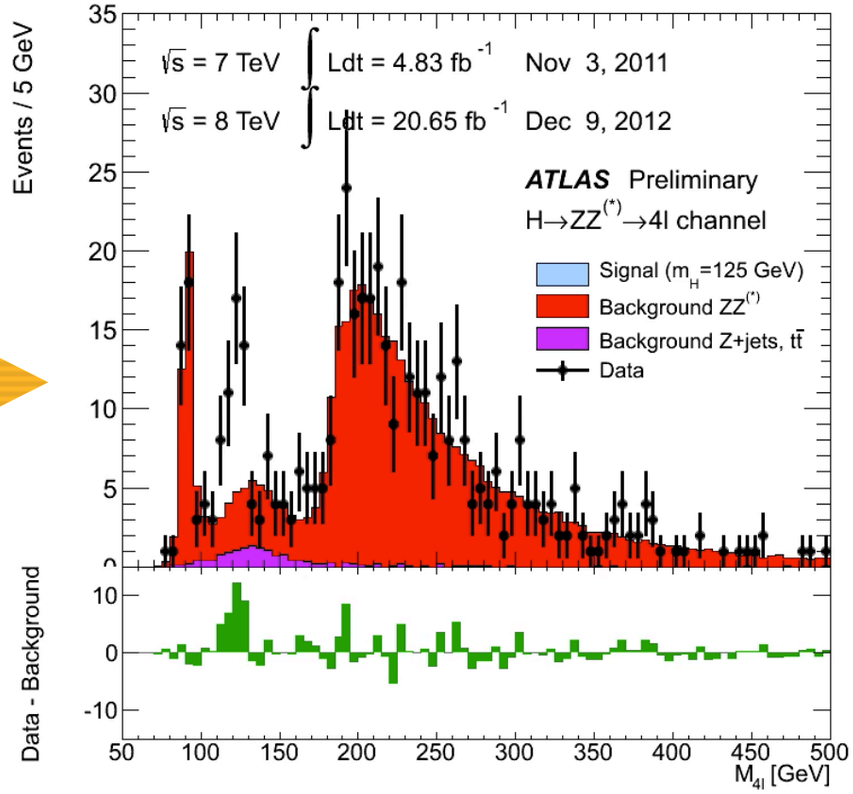


10^8 SENSORS → 1 REAL-VALUED QUANTITY

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

observed

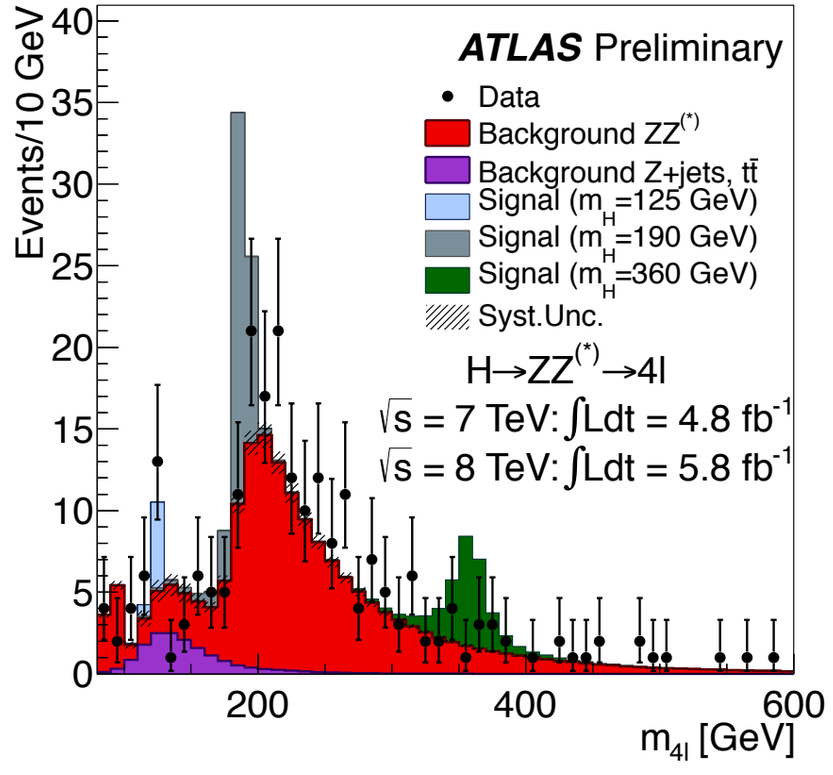
MC Sampling

simulation

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

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

histogram approximation



THE CRUX, AN INTRACTABLE INTEGRAL

observed

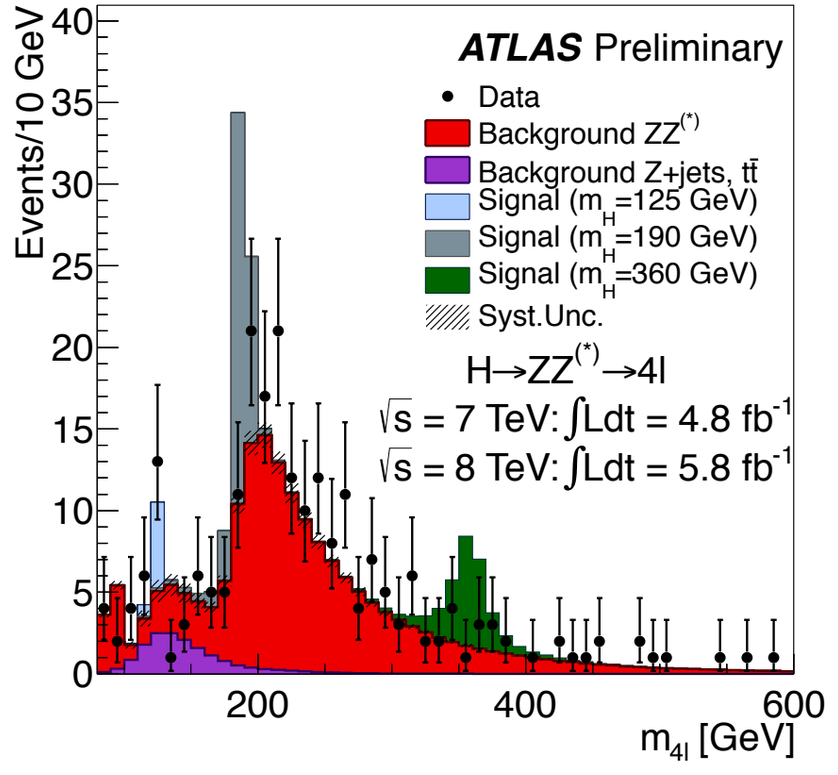
MC Sampling

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 **x** is high-dimensional

