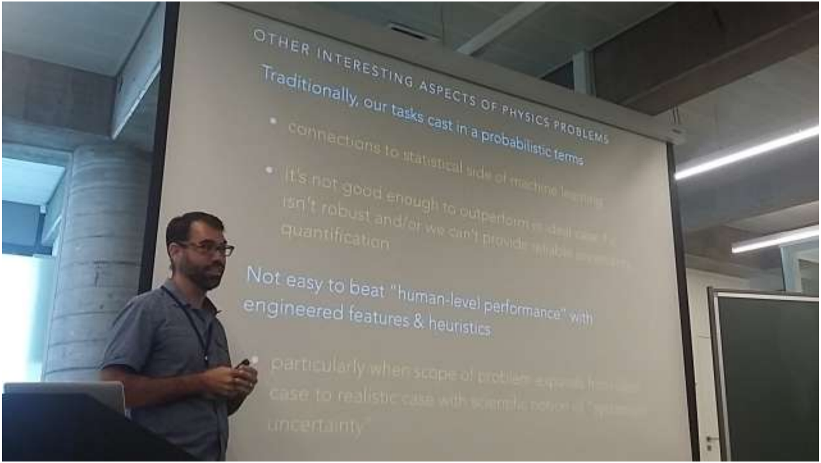
Teaching machines to discover particles

Gilles Louppe



OTHER INTERESTING ASPECTS OF PHYSICS PROBLEMS

Traditionally, our tasks cast in a probabilistic terms

- connections to statistical side of machine learning
- It's not good enough to outperform in ideal case if it isn't robust and/or we can't provide reliable uncertainty quantification

Not easy to beat "human-level performance" with engineered features & heuristics

- particularly when scope of problem expands from ideal case to realistic case with scientific notion of "systematic uncertainty"

Catch-up session for Kyle's yersterday talk

Background

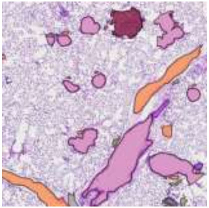
A few words about myself

Background:

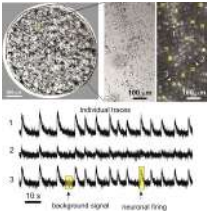
- Training in computer science
- PhD in machine learning
  - Contributions to random forests (interpretation, randomness, scalability, etc)

Machine learning for Science:

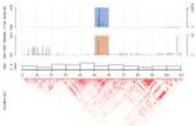
- As a PhD, I grew an interest for scientific applications of ML.



Recognition algorithms for biomedical images



Connectome reconstruction algorithms



Genome-wide associations studies with ML

Postdoc'ing at CERN + NYU

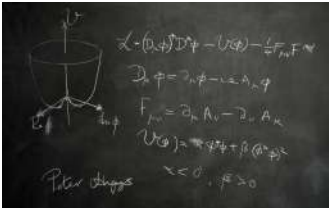
- Joined CERN, and then NYU, as a postdoc with the goal of applying ML to particle physics data.

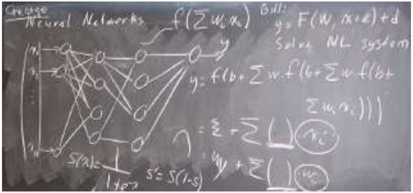
- Switched gears in terms of research:
  - Contributions in likelihood-free inference, adversarial learning, domain adaptation, ...
  - Driven by particle physics applications.

- Team work with physicists and researchers in ML.

Physics jargon vs. ML lingo

Physicists and machine learning researchers
do not speak the same language.


$\mathcal{L} = (D_\mu\phi)^\dagger D^\mu\phi - V(\phi) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}$
$D_\mu\phi = \partial_\mu\phi - ieA_\mu\phi$
$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$
$V(\phi) = \alpha\phi^*\phi + \beta(\phi^*\phi)^2$
$\alpha < 0, \beta > 0$
Peter Higgs

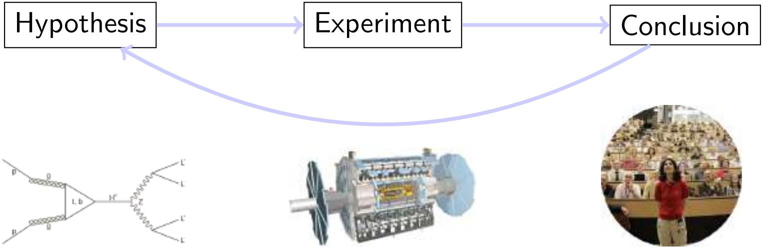


- Due/thanks to its large collaborations, particle physics has often siloed itself and (re-)developed its own tools.
- This results in a barrier between physicists and outsiders, despite sometime using the same underlying concepts.

Disclaimer. Ask if things are unclear!

Particle Physics 101

The scientific method



The Higgs boson exists

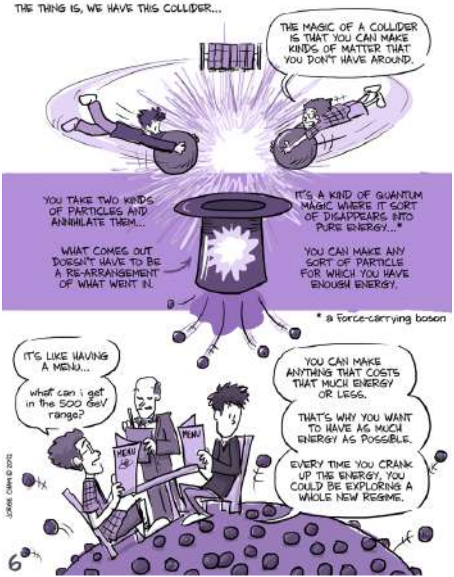
LHC+ATLAS+CMS

Discovery!

- The scientific method = recurrence over the sequence "hypothesis, experiment and conclusion".
- Conclusions are routinely automated through statistical inference, in which machine learning methods are embedded.
- Hypothesis and experiments are usually left for the scientists to decide.

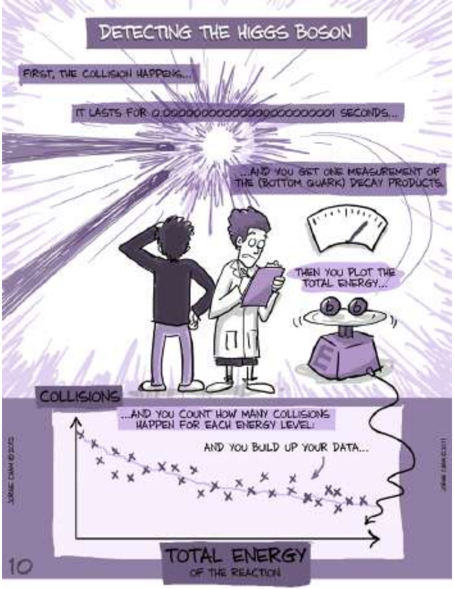
Credits: Jorge Cham, PHD comics

Testing for new physics


THE THING IS, WE HAVE THIS COLLIDER...
THE MAGIC OF A COLLIDER IS THAT YOU CAN MAKE KINDS OF MATTER THAT YOU DON'T HAVE AROUND.
YOU TAKE TWO KINDS OF PARTICLES AND ANNIHILATE THEM...
IT'S A KIND OF QUANTUM MAGIC WHERE IT SORT OF DISAPPEARS INTO PURE ENERGY...*
WHAT COMES OUT DOESN'T HAVE TO BE A RE-ARRANGEMENT OF WHAT WENT IN.
YOU CAN MAKE ANY SORT OF PARTICLE FOR WHICH YOU HAVE ENOUGH ENERGY.
* a force-carrying boson
IT'S LIKE HAVING A MENU...
what can i get in the 500 GeV range?
MENU
MENU
YOU CAN MAKE ANYTHING THAT COSTS THAT MUCH ENERGY OR LESS.
THAT'S WHY YOU WANT TO HAVE AS MUCH ENERGY AS POSSIBLE.
EVERY TIME YOU CRANK UP THE ENERGY, YOU COULD BE EXPLORING A WHOLE NEW REGIME.
JORGE CHAM © 2012
6

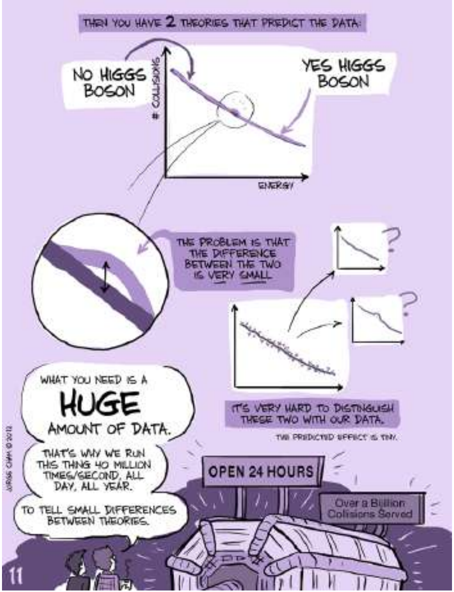
Credits: Jorge Cham, PHD comics

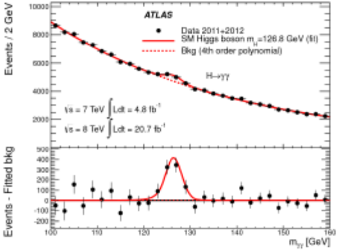
Testing for new physics


DETECTING THE HIGGS BOSON
FIRST, THE COLLISION HAPPENS...
IT LASTS FOR 0.000000000000000000000001 SECONDS...
...AND YOU GET ONE MEASUREMENT OF THE (BOTTOM QUARK) DECAY PRODUCTS.
THEN YOU PLOT THE TOTAL ENERGY...
b
b
E
COLLISIONS
...AND YOU COUNT HOW MANY COLLISIONS HAPPEN FOR EACH ENERGY LEVEL.
AND YOU BUILD UP YOUR DATA...
TOTAL ENERGY
OF THE REACTION
10

Credits: Jorge Cham, PHD comics

Testing for new physics


THEN YOU HAVE 2 THEORIES THAT PREDICT THE DATA:
NO HIGGS BOSON
YES HIGGS BOSON
# COLLISIONS
ENERGY
THE PROBLEM IS THAT THE DIFFERENCE BETWEEN THE TWO IS VERY SMALL.
IT'S VERY HARD TO DISTINGUISH THESE TWO WITH OUR DATA.
THE PREDICTED EFFECT IS TINY.
WHAT YOU NEED IS A HUGE AMOUNT OF DATA.
THAT'S WHY WE RUN THIS THING 40 MILLION TIMES/SECOND, ALL DAY, ALL YEAR.
TO TELL SMALL DIFFERENCES BETWEEN THEORIES.
OPEN 24 HOURS
Over a Billion Collisions Served
JORGE CHAM © 2012
11



Hypothesis test based on the likelihood ratio

$$\frac{p(\mathbf{x}|\text{background})}{p(\mathbf{x}|\text{background}+\text{signal})}$$

Credits: Kyle Cranmer

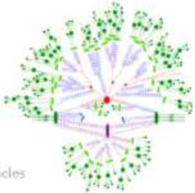
The Standard Model

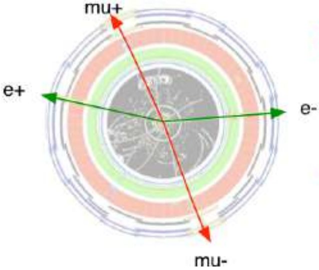
$$
\begin{aligned}
\mathcal{L}_{SM} = \; & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
+ \; & \underbrace{\bar{L}\gamma^\mu(i\partial_\mu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

1) We begin with Quantum Field Theory

2) Theory gives detailed prediction for high-energy collisions

hierarchical: 2 → O(10) → O(100) particles





3) The interaction of outgoing particles with the detector is simulated.

>100 million sensors

4) Finally, we run particle identification and feature extraction algorithms on the simulated data as if they were from real collisions.

~10-30 features describe interesting part

The **uniqueness** of particle physics lies in its highly precise and compact model.

Machine learning $\cap$ Particle physics

The players

$\boldsymbol{\theta} := (\boldsymbol{\mu}, \boldsymbol{\nu})$

Parameters

$\boldsymbol{\mu}$

Parameters of interest

$\boldsymbol{\nu}$

Nuisance parameters

$\mathbf{z}$

Latent variables

Forward modeling
Generation
Simulation

Prediction

$$p(\mathbf{x}|\boldsymbol{\theta})$$

Inference

Inverse problem
Unfolding
Measurement
Parameter search

$\mathbf{x} \sim p_r(\mathbf{x})$

Observations drawn from Nature

$\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$

Simulated data (a lot!)

Likelihood-free assumptions

Operationally,

$$\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta}) \equiv \mathbf{z} \sim p(\mathbf{z}|\boldsymbol{\theta}), \mathbf{x} = g(\mathbf{z}; \boldsymbol{\theta})$$

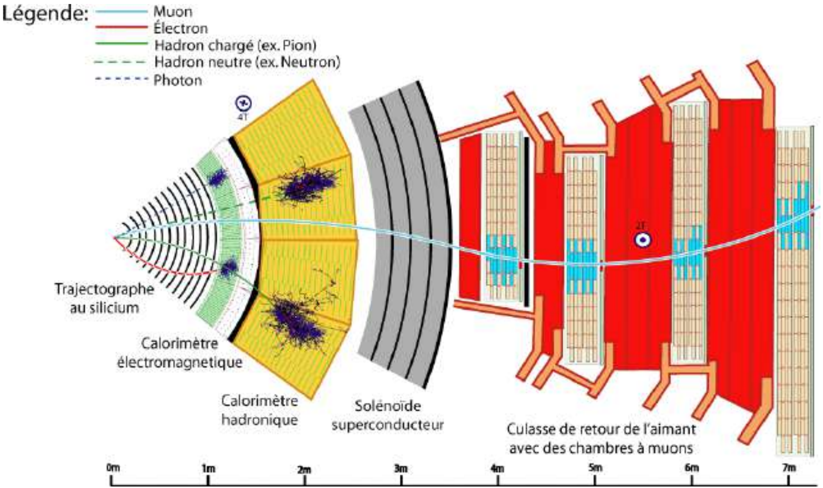
where

- $\mathbf{z}$ provides a source of randomness;
- g is a non-differentiable deterministic function (e.g. a computer program).

Accordingly, the density $p(\mathbf{x}|\boldsymbol{\theta})$ can be written as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int_{\{\mathbf{z}:g(\mathbf{z};\theta)=\mathbf{x}\}} p(\mathbf{z}|\boldsymbol{\theta})\mu(d\mathbf{z})$$

Evaluating the integral is often intractable.


Légende:
Muon
Électron
Hadron chargé (ex. Pion)
Hadron neutre (ex. Neutron)
Photon
4T
2T
Trajectographe au silicium
Calorimètre électromagnetique
Calorimètre hadronique
Solénoïde superconducteur
Culasse de retour de l'aimant avec des chambres à muons
0m
1m
2m
3m
4m
5m
6m
7m

Determining and evaluating all possible execution paths and all $\mathbf{z}$ that lead to the observation $\mathbf{x}$ is not tractable.

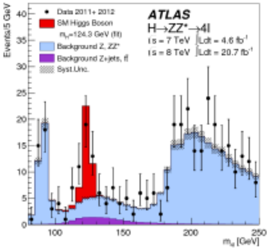
(And even less, normalizing that thing!)

Testing hypothesis (Inference)

Formally, physicists usually test a null $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ by constructing the likelihood ratio test statistic

$$\Lambda(\mathcal{D}; \boldsymbol{\theta}_0) = \prod_{\mathbf{x} \in \mathcal{D}} \frac{p(\mathbf{x}|\boldsymbol{\theta}_0)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta})}$$

- Most measurements and searches for new particles are based on the distribution of a single variable $\mathbf{x} \in \mathbb{R}$.
- The likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ is approximated using 1D histograms. (Physicists love histograms!)
- Choosing a good variable $\mathbf{x}$ tailored for the goal of the experiment is the physicist's job.



See Glen's talk today!

Credits: 1012.3589

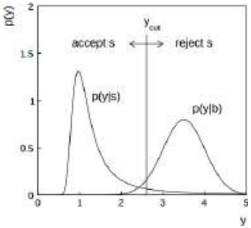
Supervised learning (Inference)

Setup:

- Training data $\{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} | \mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\mu} = y_i)\}_{i=1}^N$
- Learn a function $f : \mathcal{X} \to \mathcal{Y}$.

In particle physics:

- Part of a larger analysis
- To recognize signal from background events and build a test statistic in the region of acceptance (e.g., "cut-and-count" analysis).
- To compress the data into a 1D value (more later).

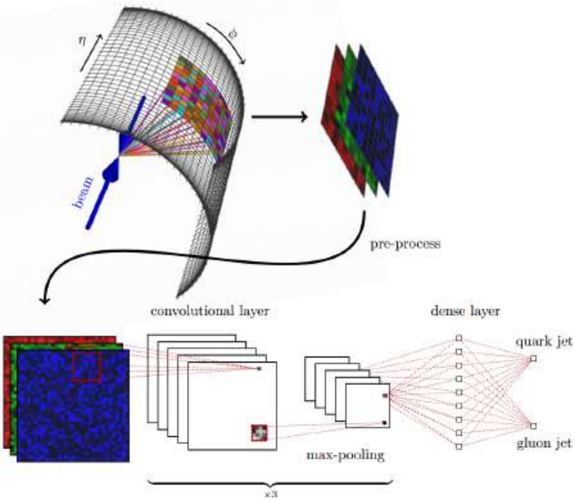


- Domain knowledge is traditionally incorporated as engineered features.

- **New paradigm:** Recent successes with deep learning models built on raw data is tickling physicists' curiosity.
    - How to recast physics problems into well-studied ML problems?
    - How to incorporate domain knowledge?
    - Can we learn what these models have learned? (*See Daniel's talk tomorrow*)

Images: 1612.01551

Particle physics detector as a camera

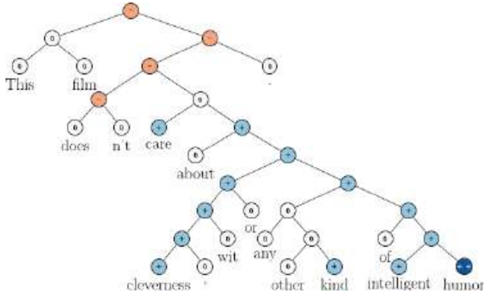


Challenges:

- 3D volume of pixels
- Non-uniform geometry
- Mostly sparse

Credits: 1702.00748

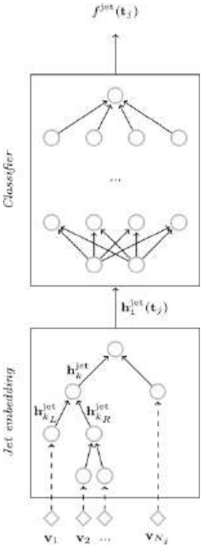
Collision events as text paragraphs



Analogy:

- word → particle
- sentence → jet
- parsing → jet algorithm
- paragraph → event

Domain knowledge is used to template the structure of the network, on a per-event basis.


$f^{\text{jet}}(\mathbf{t}_j)$
Classifier
...
$\mathbf{h}_1^{\text{jet}}(\mathbf{t}_j)$
Jet embedding
$\mathbf{h}_k^{\text{jet}}$
$\mathbf{h}_{k_L}^{\text{jet}}$
$\mathbf{h}_{k_R}^{\text{jet}}$
$\mathbf{v}_1$
$\mathbf{v}_2$
...
$\mathbf{v}_{N_j}$

QCD-aware recursive networks

Domain adaptation, Transfer learning (Inference)

Setup:

- Test data $\{\mathbf{x}_i \sim p_r(\mathbf{x})\}_{i=1}^N$
- $p_r(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$

In particle physics:

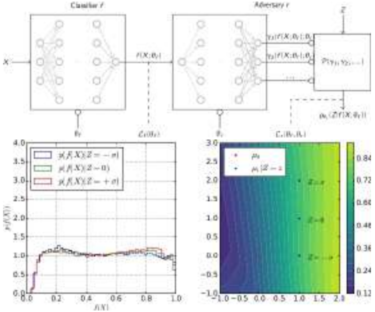
- How does one build a model from simulated data that transfers well-enough to the true data distribution?
- How does one ensure the model does not exploit simulation artefacts?

Attend Michael's talk on Saturday!

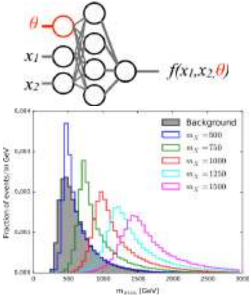
Credits: 1611.01046, 1601.07913

Learning under uncertainty (Inference)

- Despite the precision of the SM, we still have to deal with:
  - statistical uncertainties (inherent fluctuations)
  - systematic uncertainties (the known unknowns of the model)
- Uncertainty is usually formulated as nuisance parameters $\boldsymbol{\nu}$.



With adversarial training, force the model to be independent of $\boldsymbol{\nu}$.



Add $\boldsymbol{\nu}$ as an input to the model and profile it out later.

When to use one strategy over the other?

Likelihood-free inference (Inference)

Given observations $\mathbf{x} \sim p_r(\mathbf{x})$, we seek:

$$\boldsymbol{\theta}^* = \arg\max_{\theta} p(\mathbf{x}|\boldsymbol{\theta})$$

- Histogramming $p(\mathbf{x}|\boldsymbol{\theta})$ does not scale to high dimensions.
- Can we automate or bypass the physicist's job of thinking about a good and compact representation for $\mathbf{x}$, without losing information?
- Hint: We do not need to know $p(\mathbf{x}|\boldsymbol{\theta})$ to find $\boldsymbol{\theta}^*$.

Credits: 1506.02169

Approximating likelihood ratios with classifiers

The likelihood ratio $r(\mathbf{x})$ is invariant under the change of variable $\boldsymbol{u} = s(\mathbf{x})$, provided $s(\mathbf{x})$ is monotonic with $r(\mathbf{x})$:

$$r(\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}_0)}{p(\mathbf{x}|\boldsymbol{\theta}_1)} = \frac{p(s(\mathbf{x})|\boldsymbol{\theta}_0)}{p(s(\mathbf{x})|\boldsymbol{\theta}_1)}$$

A classifier s trained to distinguish $\mathbf{x} \sim p(\mathbf{x}|\theta_0)$ from $\mathbf{x} \sim p(\mathbf{x}|\theta_1)$ satisfies the condition above.

This gives an automatic procedure for learning a good and compact representation for $\mathbf{x}$!

Credits: 1506.02169

Therefore,

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_\theta p(\mathbf{x}|\boldsymbol{\theta}) \\ &= \arg\max_\theta \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x}|\boldsymbol{\theta}_1)} \\ &= \arg\max_\theta \frac{p(s(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\theta}_1)|\boldsymbol{\theta})}{p(s(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\theta}_1)|\boldsymbol{\theta}_1)}\end{aligned}$$

where $\boldsymbol{\theta}_1$ is fixed and $s(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\theta}_1)$ is a family of classifiers trained to distinguish between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$.

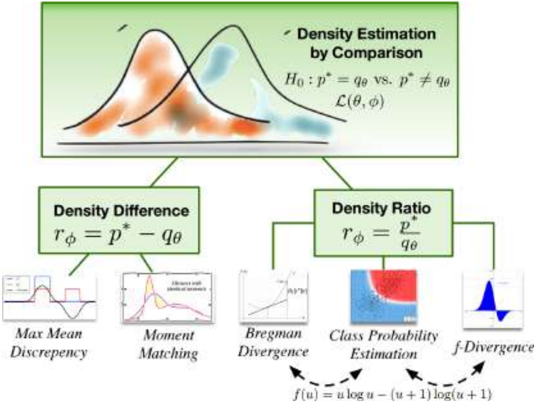
Credits: 1610.03483

Learning in implicit generative models

Likelihood-free inference can be cast into the framework of "implicit generative models".

This framework ties together:

- Approximate Bayesian computation
- Density estimation-by-comparison algorithms (two sample testing, density ratio, density difference estimation)
- Generative adversarial networks
- Variational inference



Density Estimation by Comparison

$H_0 : p^* = q_\theta$ vs. $p^* \neq q_\theta$

$\mathcal{L}(\theta, \phi)$

Density Difference

$r_\phi = p^* - q_\theta$

Density Ratio

$r_\phi = \frac{p^*}{q_\theta}$

Max Mean Discrepency

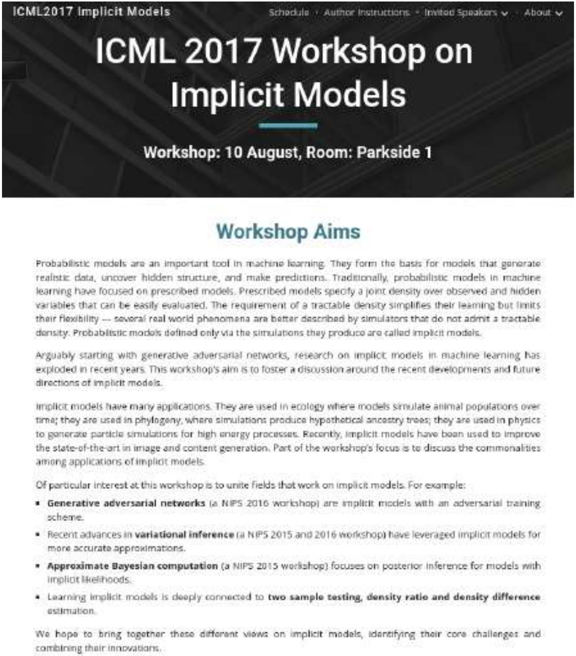
Moment Matching

Bregman Divergence

Class Probability Estimation

f-Divergence

$f(u) = u \log u - (u+1) \log(u+1)$



ICML 2017 Workshop on Implicit Models

Workshop: 10 August, Room: Parkside 1

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.

Hot topic in machine learning!

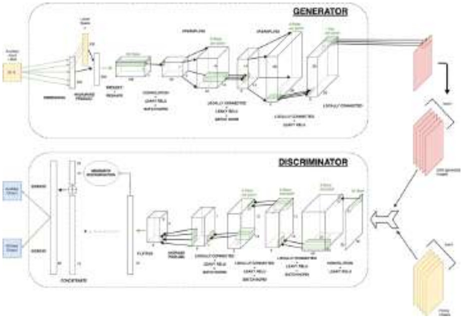
Fast simulation (Prediction)

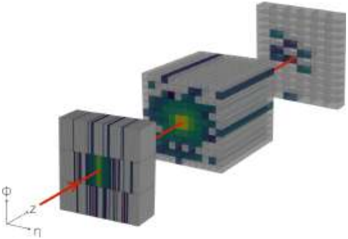
- Half the LHC computing power (300000 cores) is dedicated to producing simulated data.
- Huge savings (in time and $) if simulations can be made faster.
- Hand-made fast simulators are being developed by physicists, trading-off precision for speed.

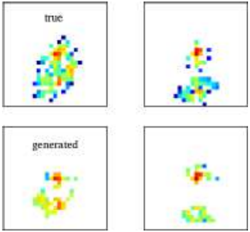
Can we learn to generate data?
(i.e. can we build a fast proxy for $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$?)

Credits: 1701.05927, 1705.02355

Learning generative models (Prediction)





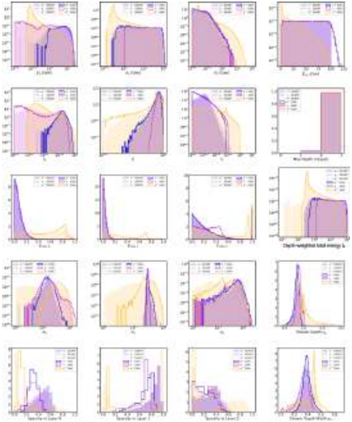


Challenges:

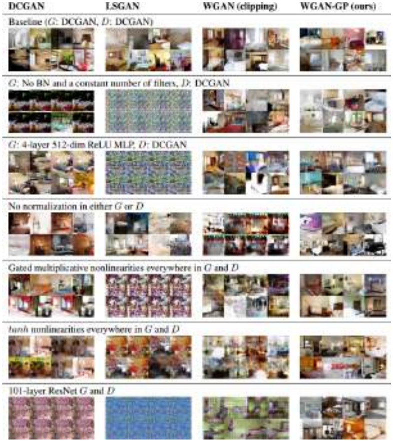
- How to ensure physical properties?
- Non-uniform geometry
- Mostly sparse
- GANs vs. VAE vs. Normalizing Flows?

Credits: 1705.02355, 1704.00028

How to evaluate generative models?



Physics: Evaluate well-known physical variates


DCGAN
LSGAN
WGAN (clipping)
WGAN-GP (ours)
Baseline (G: DCGAN, D: DCGAN)
G: No BN and a constant number of filters, D: DCGAN
G: 4-layer 512-dim ReLU MLP, D: DCGAN
No normalization in either G or D
Gated multiplicative nonlinearities everywhere in G and D
tanh nonlinearities everywhere in G and D
101-layer ResNet G and D

ML: Look at generated images

This is not satisfying.

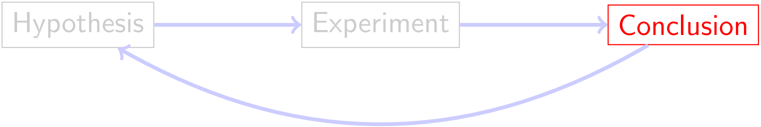
Can't we do better from a methodological standpoint?

(Some first steps at 1511.01844)

Outlooks

(a thought experiment)

Automating the scientific process



Hypothesis

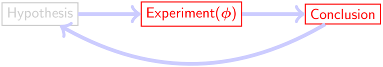
Experiment

Conclusion

Most efforts are focused on automating the analysis of experimental results to draw conclusions, assuming the hypothesis and experiment are fixed.

Can we also automate
the steps of hypothesis and experiments?

Optimal experimental desgin



Hypothesis

Experiment(ϕ)

Conclusion

Parameters $\boldsymbol{\theta}$ of the (standard) model are known with uncertainty $H[\boldsymbol{\theta}]$. *How to best reduce the uncertainty* $H[\boldsymbol{\theta}]$*?*

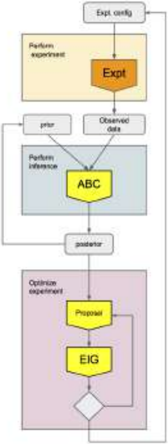
1. Assume an experiment with parameters $\boldsymbol{\phi}$ can be simulated.
2. Simulate the expected improvement $\Delta(\boldsymbol{\phi}) = H[\boldsymbol{\theta}] - \mathbb{E}_{\text{data}|\boldsymbol{\phi}}[H[\boldsymbol{\theta}|\text{data}]]$.
   - This embeds the full likelihood-free inference procedure.
3. Find $\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}} \Delta(\boldsymbol{\phi})$
   - Computationally (super) heavy.

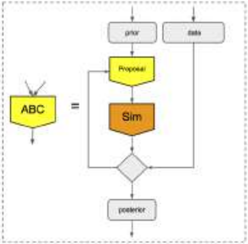
Connections to:

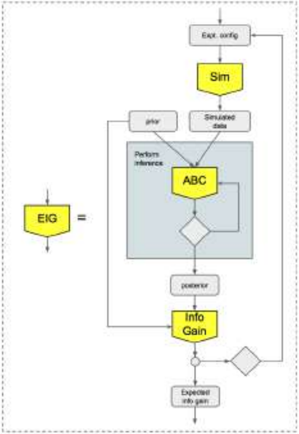
- Bayesian optimization
- Optimal experimental design
- Reinforcement learning (for a sequence of experiments)

Credits: cranmer/active_sciencing

Active sciencing

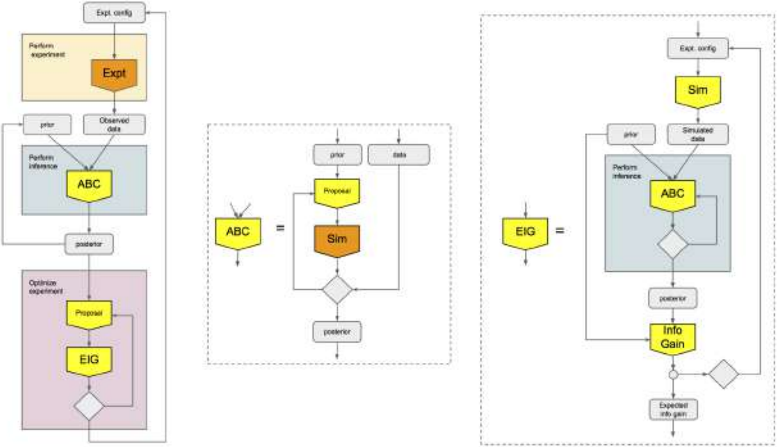


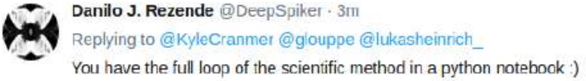




Credits: cranmer/active_sciencing

Active sciencing



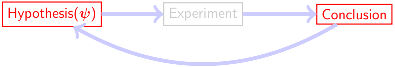


Danilo J. Rezende @DeepSpiker · 3m

Replying to @KyleCranmer @glouppe @lukasheinrich_

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

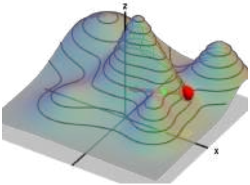
Exploring the theory space

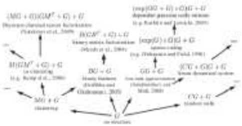


The Standard model admits several extensions.

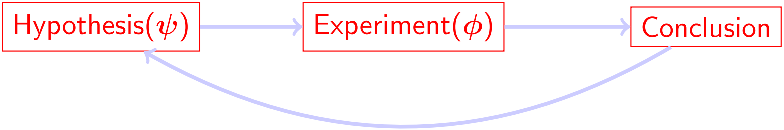
Can we explore the space of theories and find the envelope that agree with the data?

- Assume a generative model of theories, indexed by $\boldsymbol{\psi}$.
- Assume the experiment design $\boldsymbol{\phi}$ is fixed.
- Find $\{\boldsymbol{\psi} | \rho(p_r(\mathbf{x}|\boldsymbol{\phi}), p(\mathbf{x}|\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\theta}^*)) < \epsilon\}$.





AI recipe for understanding Nature



Hypothesis(ψ)

Experiment(ϕ)

Conclusion

Find $\{\boldsymbol{\psi} | \rho(p_r(\mathbf{x}|\boldsymbol{\phi}), p(\mathbf{x}|\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\theta}^*)) < \epsilon, \forall \boldsymbol{\phi}\}$

Summary

Why collaborating with physicists?

- Contribute to the understanding of the Universe.
- Open methodological challenges.
- Test bed for developing ambitious ML/AI methods, as enabled by the precise mechanistic understanding of physical processes.
- Core problems in particle physics transfer to other fields of science (likelihood-free inference, domain adaptation, optimization, etc).