LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Machine learning in particle physics

LIP Internship Program
Summer 2021

Miguel Crispim Romão
mcromao@lip.pt

Pheno Group

What is Machine Learning?

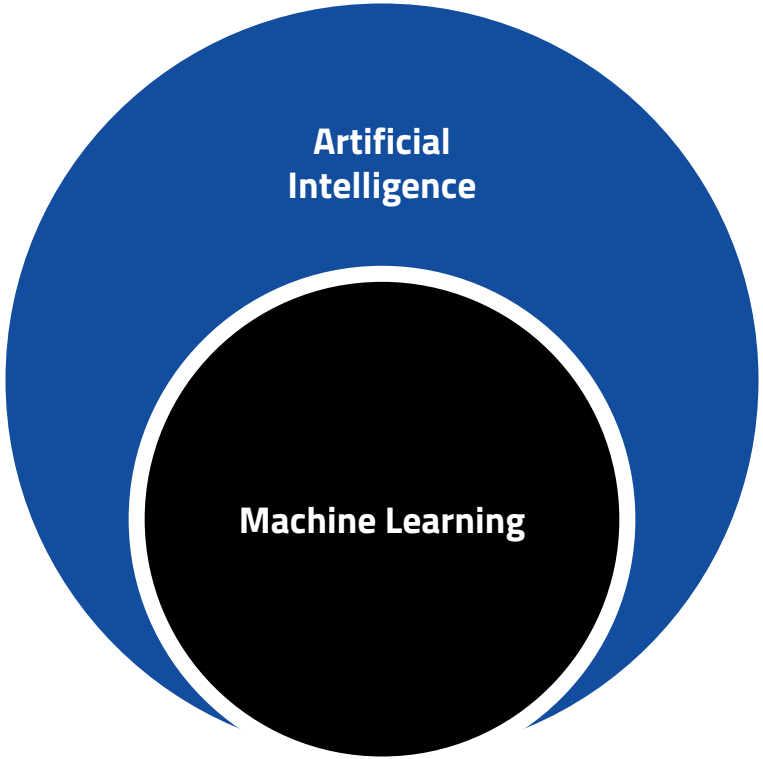
From an Artificial Intelligence Perspective

“Artificial Intelligence is the quest of creating machines that think and act intelligently

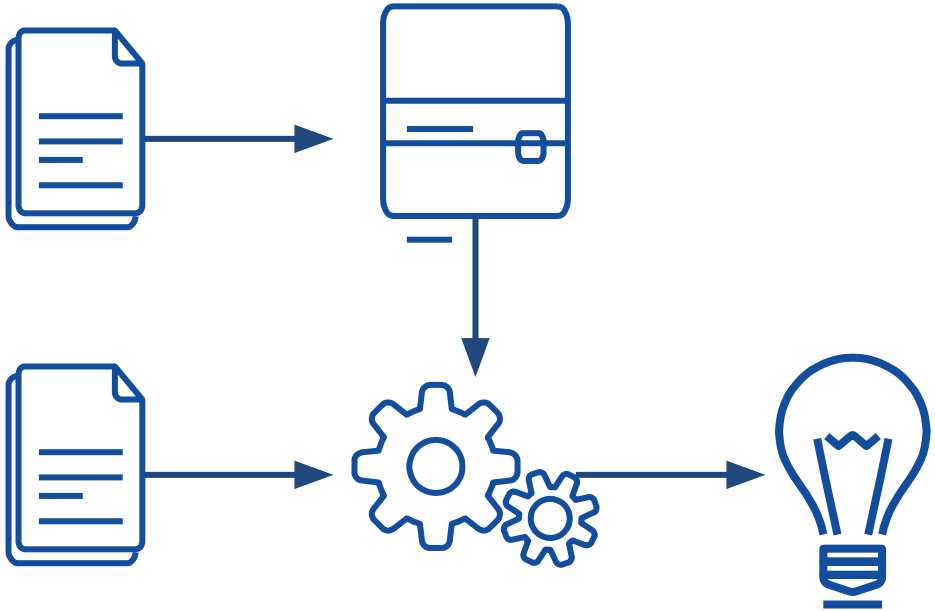
Artificial Intelligence is a big topic
and covers many problems

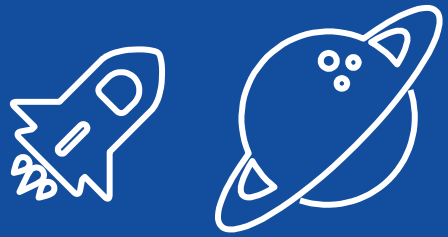
- Reasoning and Problem-solving
- Knowledge Representation
- Planning
- **Learning**
- Natural Language Processing
- Perception
- Motion and Manipulation
- Social Intelligence
- “General Intelligence”

Machine Learning is the subfield of AI that concerns how a machine can learn to perform tasks



A machine learns how to perform a task by creating a model that will act intelligently on new data

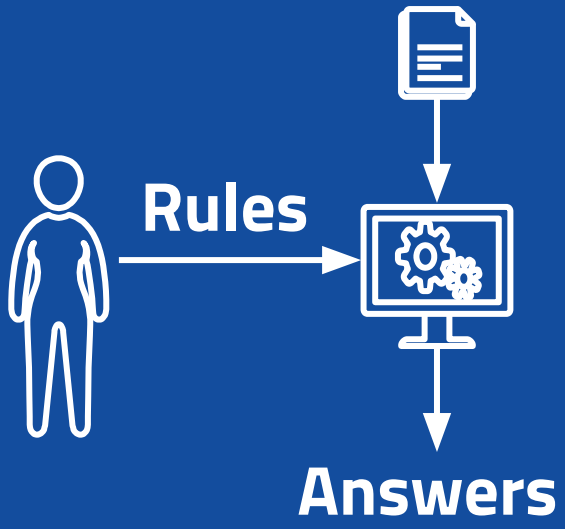




Self-Taught Code

Machine Learning is a different paradigm of computing: a program that learns what it has to do

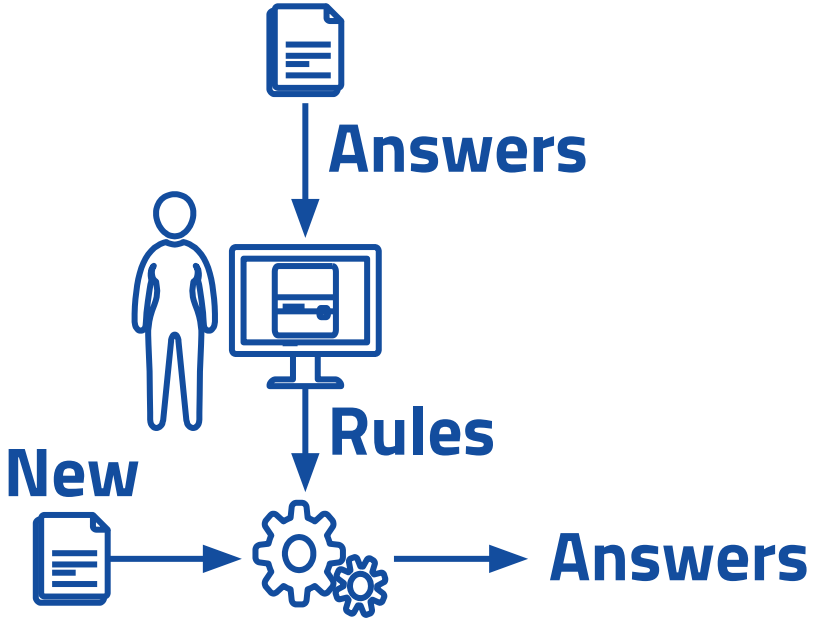
Classical Programming



Machine Learning

= Decision Function

= Data



Machine Learning Taxonomy

What is out there and what tasks can we solve?

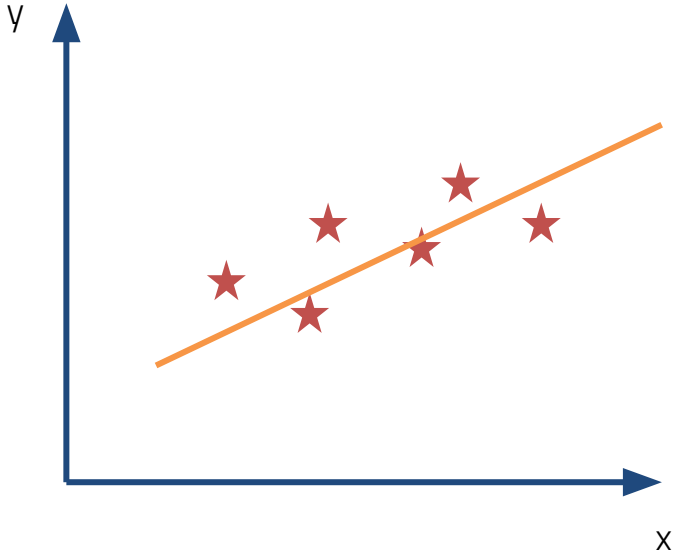
Machine Learning
Taxonomy: Types of Learning

The main differentiator is the type of learning, i.e. by **task**

- Supervised
  - Data includes the answers
- Unsupervised
  - Algorithm embodies the answers
- Other types
  - Semi-supervised
  - Self-supervised
  - Reinforcement

Regression Example

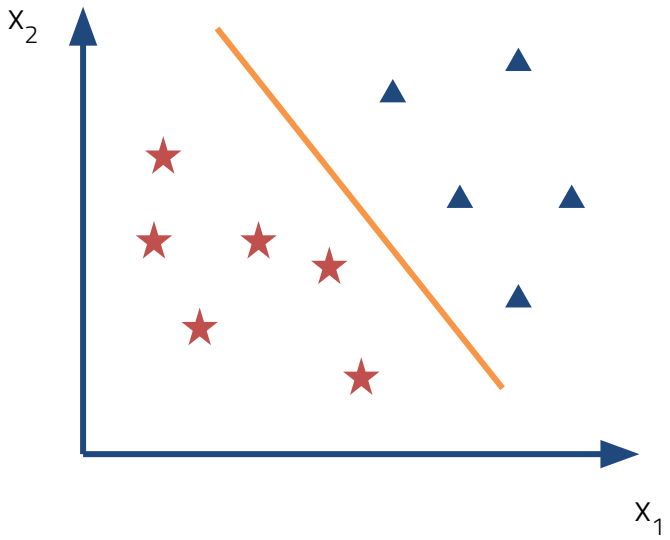
Linear Regression



$$y = wx + b$$

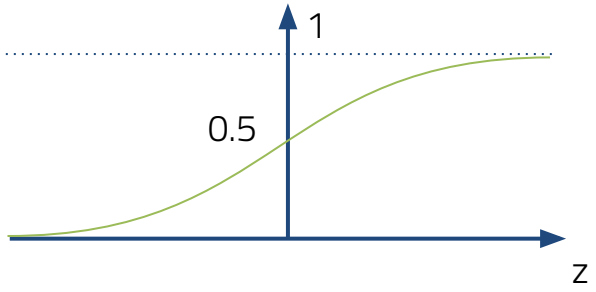
Classification Example

Logistic Regression: Parametric Example



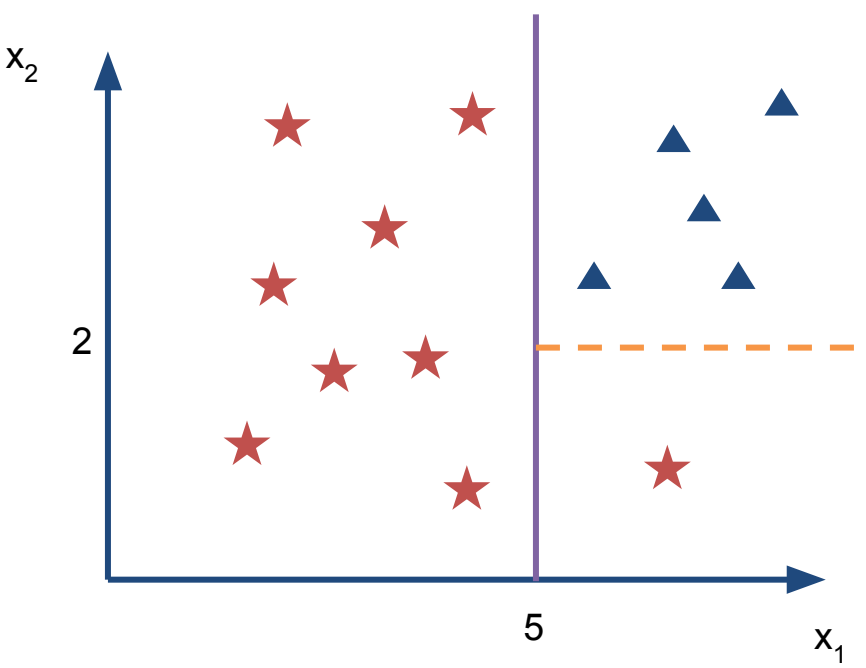
$$\sigma(x) = \frac{1}{1 + e^{-z}}$$

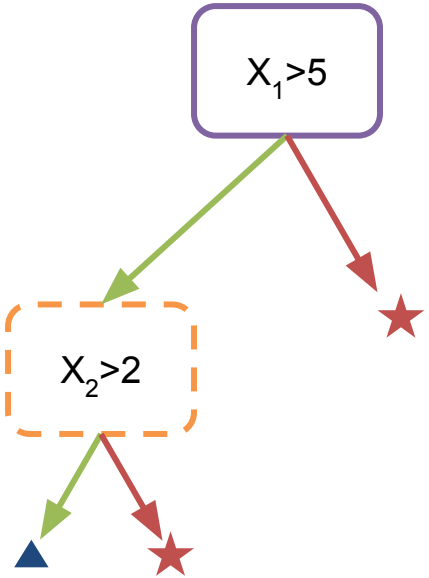
$$z = \vec{w} \cdot \vec{x} + b$$



Machine Learning

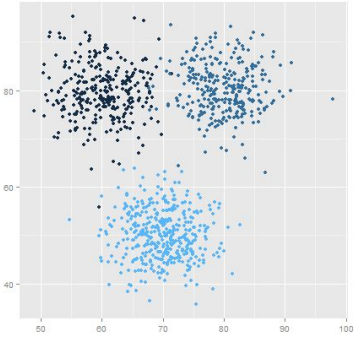
Decision Tree: Non-parametric example

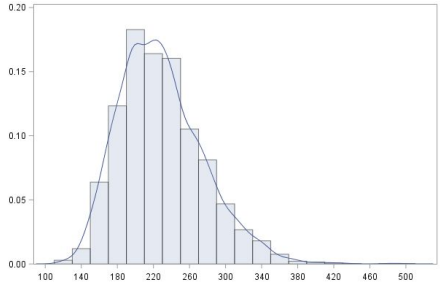


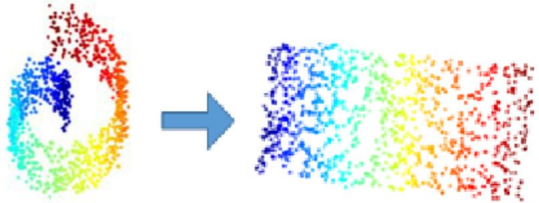


Machine Learning
Taxonomy: Unsupervised Learning

- The training data does not include the answer we want to reproduce
- The answer is embodied in the Learning Algorithm (i.e. provided by a human)
- The model will learn how to map the X to the answers
- Answers define the type of model
  - Clustering
  - Density Estimation
  - Dimensional Reduction



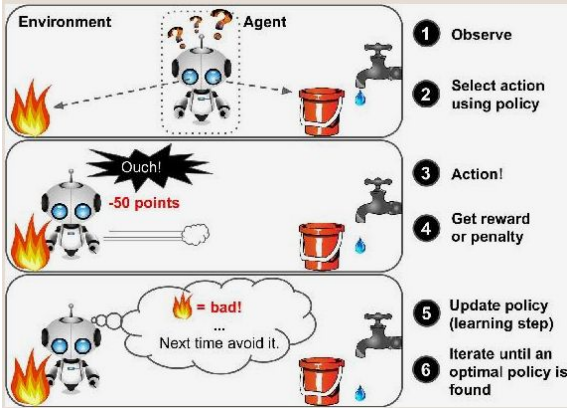




Machine Learning

Taxonomy: Other types of learning

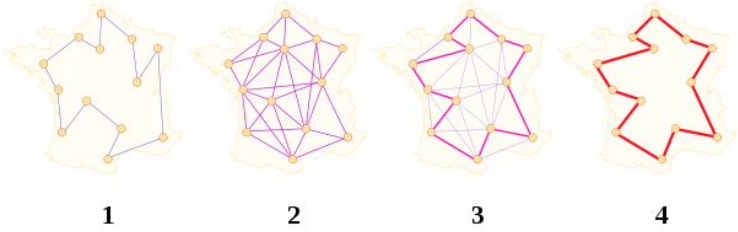
- Reinforcement learning:
  - An agent interacting with environment
- Self-supervised:
  - Representation learning
  - Generative models



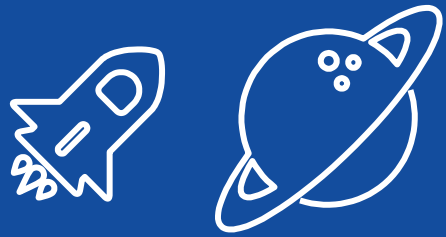
https://thispersondoesnotexist.com/

Machine Learning
Taxonomy: Other AI approaches

- Search
  - Travel salesman problem
- Optimisation
  - Bayesian optimisation
  - Genetic algorithms



Main Point



AI/ML can provide alternative approaches for any task that is either **data** or **computationally** intensive

Why is High-Energy Physics Ideal for AI/ML?
A match made in heaven…

- To Mario's disappointment, current and future collider experiments are data heavy (c.f. Michele's talk)
- Data generated are inherently probabilistic due to quantum mechanics: ML loves a good distribution
- Data simulation and calibration tasks are computationally heavy

278

petabytes of data

In the last decade, LHC experiments collected almost 280 petabytes of data, which scientists recorded on tape. You would need to stream Netflix 24/7 for more than 15,000 years to eventually use that much data! But from another perspective, platforms like Facebook (which has 2.5 billion users) collect that much data in 70 days!

7.5 billion

Worldwide LHC Computing Grid requests

Physicists need a huge amount of computing power to do their research—much more than a standard laptop can support. Every day several thousand physicists submit a total of about 2 million "jobs" to the WLCG. Each "job" is an important brick in the growing body of scientific work.

https://www.symmetrymagazine.org/article/10-years-of-lhc-physics-in-numbers

Machine Learning in the Wild

High-Energy Physics Applications

Many many applications nowadays
Won't cover all

- HEP community has progressed significantly on AI/ML applications in the past few years
- Exhaustive review is impossible
- The community has put together a living review
  - https://iml-wg.github.io/HEPML-LivingReview/

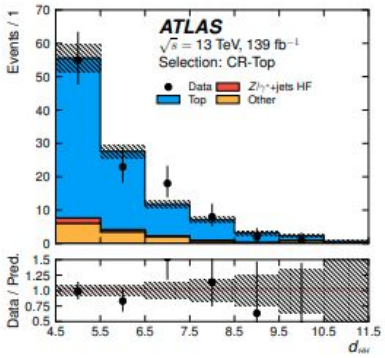
I will focus on broad areas of application and examples of what we do at LIP with AI/ML

Data Intensive Tasks

Classification: Looking for something specific

Better New Physics Analysis

- We start with many (tens) of different variables: which is the best to find the events of interest?
  - Use a supervised classifier (trained on simulation) to combine them all into a single discriminant (c.f. Ana Peixoto's talk)
  - We will be seeing this tomorrow in the tutorial session
  - By isolating the signal, we increase the **statistical efficiency** of our analysis (c.f. this afternoon's tutorials)
  - **Better efficiency = better exclusions or discovery**



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Phys. Lett. B 801 (2020) 135145
DOI: 10.1016/j.physletb.2019.135145

CERN-EP-2019-143
7th February 2020

Search for non-resonant Higgs boson pair production in the $bb\ell\nu\ell\nu$ final state with the ATLAS detector in pp collisions at $\sqrt{s}$ = 13 TeV

Classification: Looking for something specific

Better Event Tagging

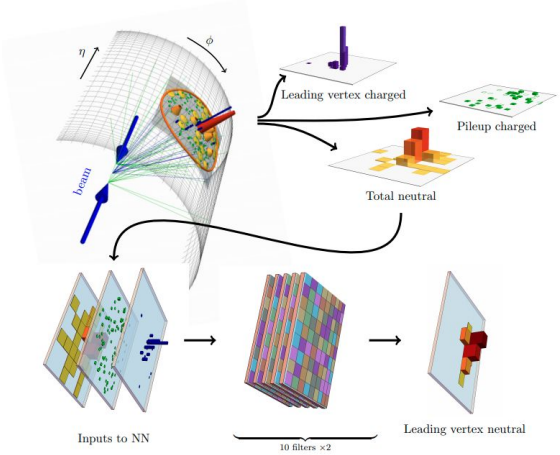
- Correctly identify known SM processes
  - Top, Strange, B quarks
  - Tau
  - Higgs
  - Z/W
  - Quark vs Gluon Jets
- Or rare phenomena
  - Quark-Gluon Plasma modified jets (c.f. Liliana's talk)
- **Better tagging = Better Physics studies**

Deep learning
Novel Approaches to Old Problems

- The current Machine Learning interest is very motivated by Deep Learning
  - A class of Machine Learning models that are very versatile
    - Can intake data in any formats (even very low-level without any human pre-processing)
      - Images, Text, Audio, Video, etc
      - **Go beyond tabular data**
    - Can tackle any problem which can be framed through a differential loss
      - Generative models, Deep Reinforcement Learning, etc
      - **Go beyond traditional discrimination tasks**

Deep learning
Novel Approaches to Old Problems: Jet Images

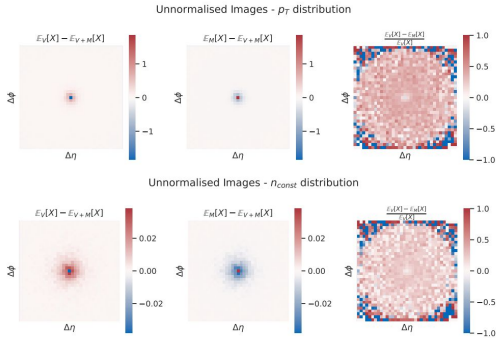
- As you have learnt from Michele's lecture, at colliders we only have two things
  - Tracks of charged particles
  - Jets of from energy deposits in calorimeters
- As you have learnt from Agostinhos' lecture
  - Calorimeters are composed of cells forming a grid. Each grid works as an "eye", or better yet: a **pixel**
    - **Can represent the jets as images**

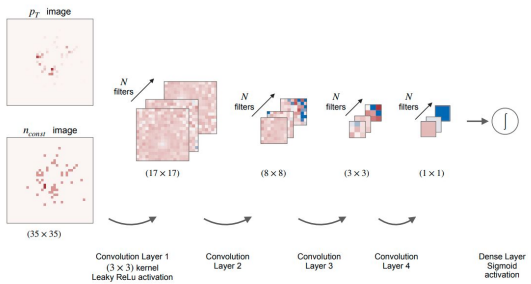


https://arxiv.org/abs/1707.08600

Deep learning
Novel Approaches to Old Problems: Jet Images

- Fresh out of the press: Deep Learning for the Classification of Quenched Jets [MCR, L. Apolinario, N. F. Castro, J. G. Milhano, R. Pedro, F. C. R. Peres] https://arxiv.org/abs/2106.08869
  - Differentiate jets that only lived in vacuum from those that might have interacted with the Quark-Gluon Plasma
  - Used Jet Images and Lund plane paths (Physics input)

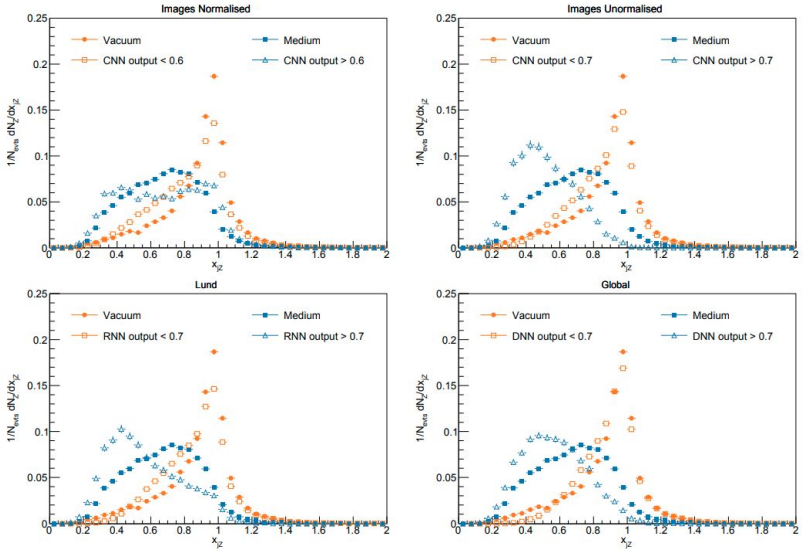




Deep learning

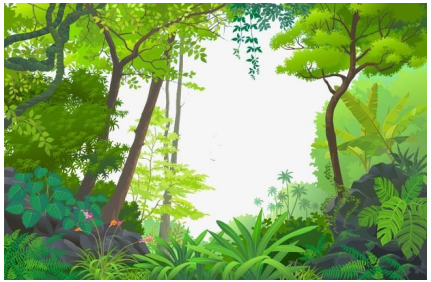
Novel Approaches to Old Problems: Jet Images

- With no high-level features, the networks performed better than the customary variables
- Despite the complexity of the problem, vacuum-like jets were consistently identified
  - Allow to purify samples of modified jet to further study the Quark-Gluon Plasma: ML enhanced Physics!



Deep learning
Novel Approaches to Old Problems: New Physics

- Supervised classifiers are great to search for something specific
- In the end of the day, we don't really know what new physics can look like (c.f. Ana Peixoto's talk)
- What if we want to search for **anything new**?
  - We know what we know: The Standard Model
  - We don't know what we don't know: New Physics



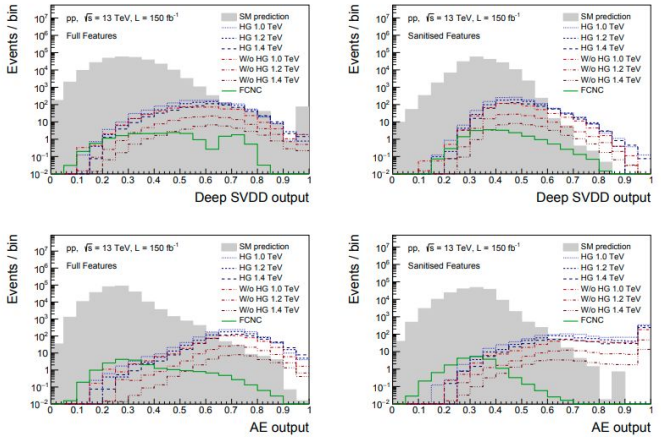

Deep learning
Novel Approaches to Old Problems: New Physics

- Since we know what we know, the rest has to be an **anomaly**
- Use novel Deep Learning methods of **anomaly detection**
  - Auto-Encoders
  - Deep-SVDD

https://arxiv.org/abs/2006.05432

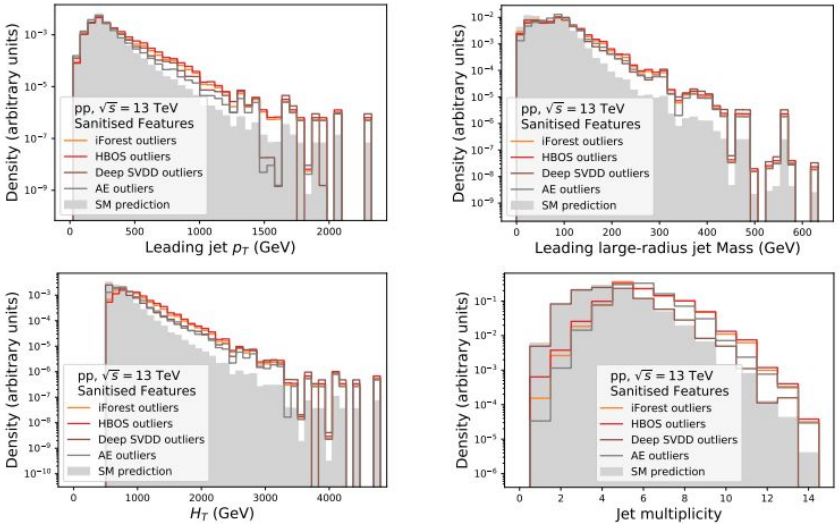
Finding new physics without learning about it: anomaly detection as a tool for searches at colliders

M. Crispim Romão[1], N. F. Castro[1,2], and R. Pedro[1]



Deep learning
Novel Approaches to Old Problems: New Physics

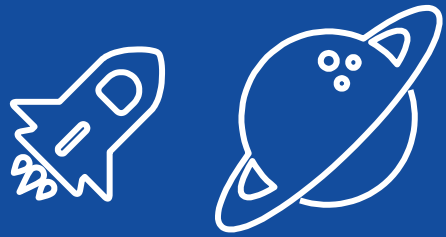
- Can also see how different Anomaly Detection methods detect anomalies
- Likely no “one model to find them all”
- Lot of work to be done down the road…
- **Novel approach to search for New Physics**, complementary to the supervised way, but with the potential of wider coverage



Other Data Intensive Tasks

One slide to cite them all

- Track reconstruction
- Pileup Mitigation
- Calibration
- Applications in Neutrino Physics and Experiments
  - c.f. Paulo Bras talk and their work with unsupervised methods to signal clustering
- Cosmology, Astro Particle, and Cosmic Ray physics
  - c.f. Ruben talk



Disclaimer

This does not, by any means nor extent, cover everything. I’m aware of this.

Computationally Intensive Tasks

Computationally Intensive Tasks

Not all data seems like data

- So far we have seen applications where we have a lot of (real or simulated) data of events.
- Albeit this is the straightforward way to use AI/ML, many other tasks in HEP are computationally intensive and can benefit AI/ML
- Many of these actually do involve a lot of intermediate data, which is the reason for the computational overhead
- And remember: **data is what ML craves**

Computationally Intensive Tasks

Not all data seems like data: Monte Carlo Generators

- One of the main computationally intensive tasks is to simulate experiment data using Monte Carlo generators (c.f. Bernardo talk just now)
  - Simulated data is used to prepare analyses, calibrate setups, and even test models for Quark-Gluon Plasma for example
- Particle Physics processes are **non deterministic** due to their **quantum mechanical nature**
- In order to simulate events at experiments, one needs to simulate a lot of possible events in order to have a good **statistical description** of the process
- (I'm spending a lot of time in this because it'll appear again in the tutorial)

Computationally Intensive Tasks

Not all data seems like data: Monte Carlo Generators

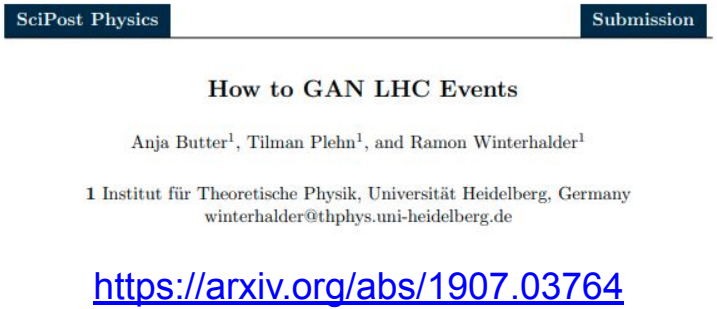
- The generation requires extensive sampling (data!) from unknown distributions. This sampling is **expensive** if one wants to cover the whole underlying (quantum mechanical) distribution

- Solution: Generative methods!

Computationally Intensive Tasks

Not all data seems like data: Monte Carlo Generators

- Generative methods work by **learning** the distribution from where we want to sample. Once learnt, we can sample with almost no computational overhead
- Two approaches:
  - Start with a few examples, learn a distribution from it and hope it interpolates well (works ok)
  - Hybrid method: Monte Carlo sampling on top of progressively learning approximation of the distribution

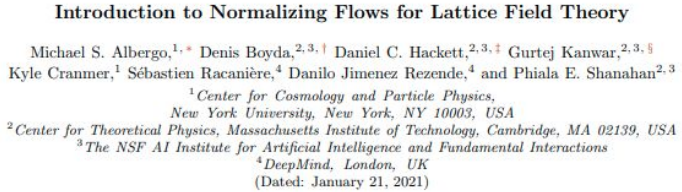

SciPost Physics Submission

How to GAN LHC Events

Anja Butter[1], Tilman Plehn[1], and Ramon Winterhalder[1]

1 Institut für Theoretische Physik, Universität Heidelberg, Germany
winterhalder@thphys.uni-heidelberg.de

https://arxiv.org/abs/1907.03764


Introduction to Normalizing Flows for Lattice Field Theory

Michael S. Albergo,[1,*] Denis Boyda,[2,3,†] Daniel C. Hackett,[2,3,‡] Gurtej Kanwar,[2,3,§]
Kyle Cranmer,[1] Sébastien Racanière,[4] Danilo Jimenez Rezende,[4] and Phiala E. Shanahan[2,3]

[1] *Center for Cosmology and Particle Physics, New York University, New York, NY 10003, USA*
[2] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[3] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[4] *DeepMind, London, UK*

(Dated: January 21, 2021)

https://arxiv.org/abs/2101.08176

Computationally Intensive Tasks

Not all data seems like data: BSM Validation

- Another often overlooked use case is that of validating Beyond the Standard Model models
- Given a model and its parameters, what values for these are still valid against experimental data? How to sample the valid values efficiently?
- AI/ML for the rescue!

Efficient sampling of constrained high-dimensional theoretical spaces with machine learning

Jacob Hollingsworth,[1] Michael Ratz,[1] Philip Tanedo,[2] and Daniel Whiteson[1]

[1] *Department of Physics and Astronomy, University of California, Irvine, CA 92697*

[2] *Department of Physics and Astronomy, University of California, Riverside, California 92521*

(Dated: March 15, 2021)

https://arxiv.org/abs/2103.06957

Computationally Intensive Tasks

Not all data seems like data: Tuning other ML models

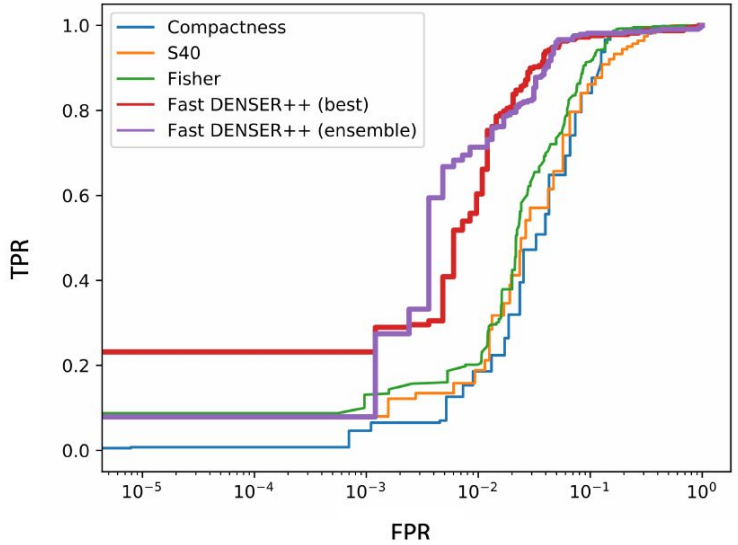
- Even optimising models for data heavy tasks can be done with AI/ML

Automatic Design of Artificial Neural Networks for Gamma-Ray Detection

FILIPE ASSUNÇÃO[1], JOÃO CORREIA[1], RÚBEN CONCEIÇÃO[2], MÁRIO JOÃO MARTINS PIMENTA[2], BERNARDO TOMÉ[2], NUNO LOURENÇO[1], AND PENOUSAL MACHADO[1]

[1]CISUC, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal

[2]LIP/IST, 1600-078 Lisbon, Portugal



The take-home messages

Take-home messages
What Machine Learning is

- ML is a different computing paradigm of self-taught code that learns from previous examples
- A set of solutions for current problems
- A set of novel approaches that opens up new types of research
- A technology that is here to stay and is already embedded in our lives
- An engineering science with little theoretical grounding but huge collection application examples
- Currently profoundly based on statistical learning theory and function approximation/functional analysis
- A skill-set that will be at the same level as coding for your generation

Take-home messages

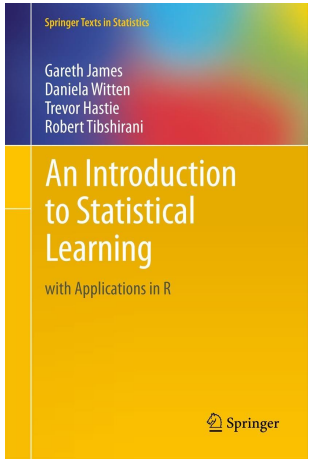
What Machine Learning is not

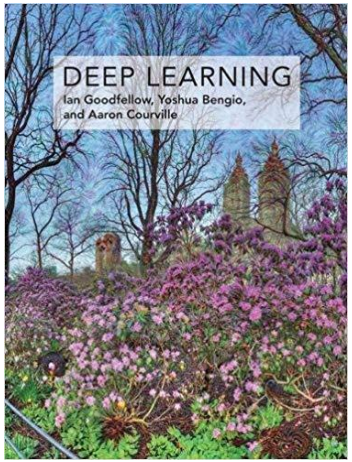
- Capabel to extrapolate and abstract reasoning beyond tasks -> **Ultimately bound to the data where it was trained**
- A solution for every problem -> **Sometimes a nail is just a nail and you only need a hammer**
- A magic framework where everything can be done -> **There are limits to its application**
- A substitute for other computing paradigms -> **Learn how to code**
- An existential threat to humanity -> **Popular culture has created a fantasy idea of AI which has no grounds on the actual technology**

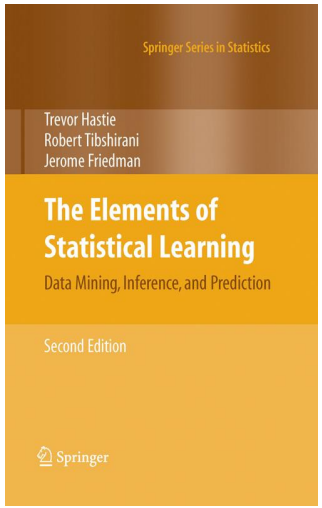
Further resources

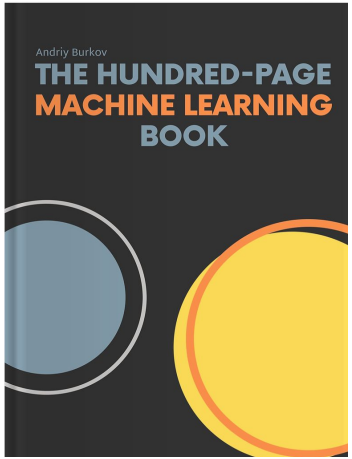
Some of them are free

These are free

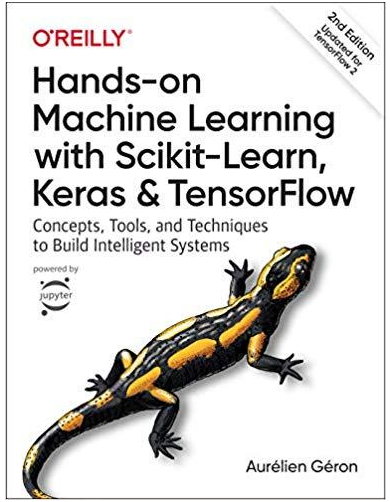


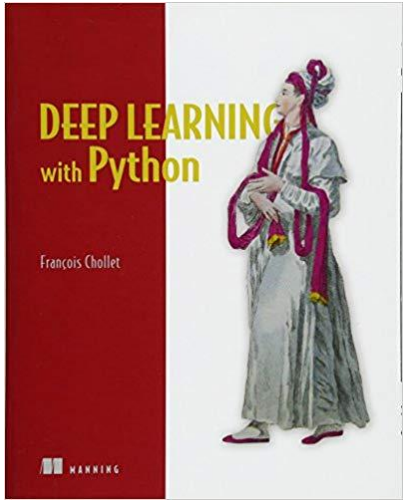


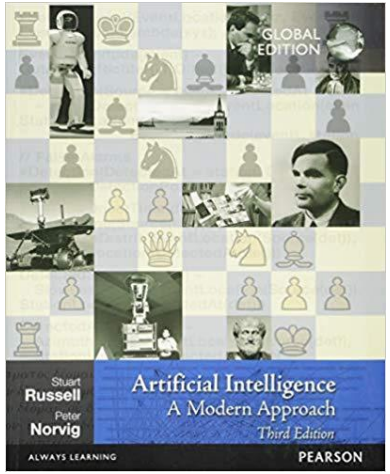


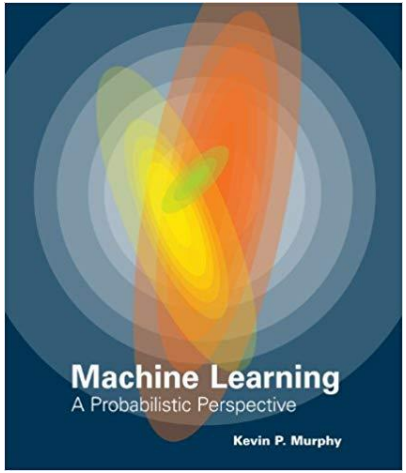


Not free, but very good









For Tomorrow's Tutorial

Tutorial info

- We will be using Google Colab: No need to install anything
  - You are of course more than welcome to set up you own python environment on your computer, but I won't help debugging
- It will be a mix of slides and code-along sessions, followed by breakout rooms with other tutors (Paulo, Rute, Maura, Ceu)
- If you want to prepare read the first two chapters of The Hundred-Page Machine Learning Book
  - https://www.dropbox.com/s/lrhtt1wkffnm4fe/Chapter1.pdf?dl=0
  - https://www.dropbox.com/s/0cprdghmnzpck8h/Chapter2.pdf?dl=0

Thanks!

Any questions?