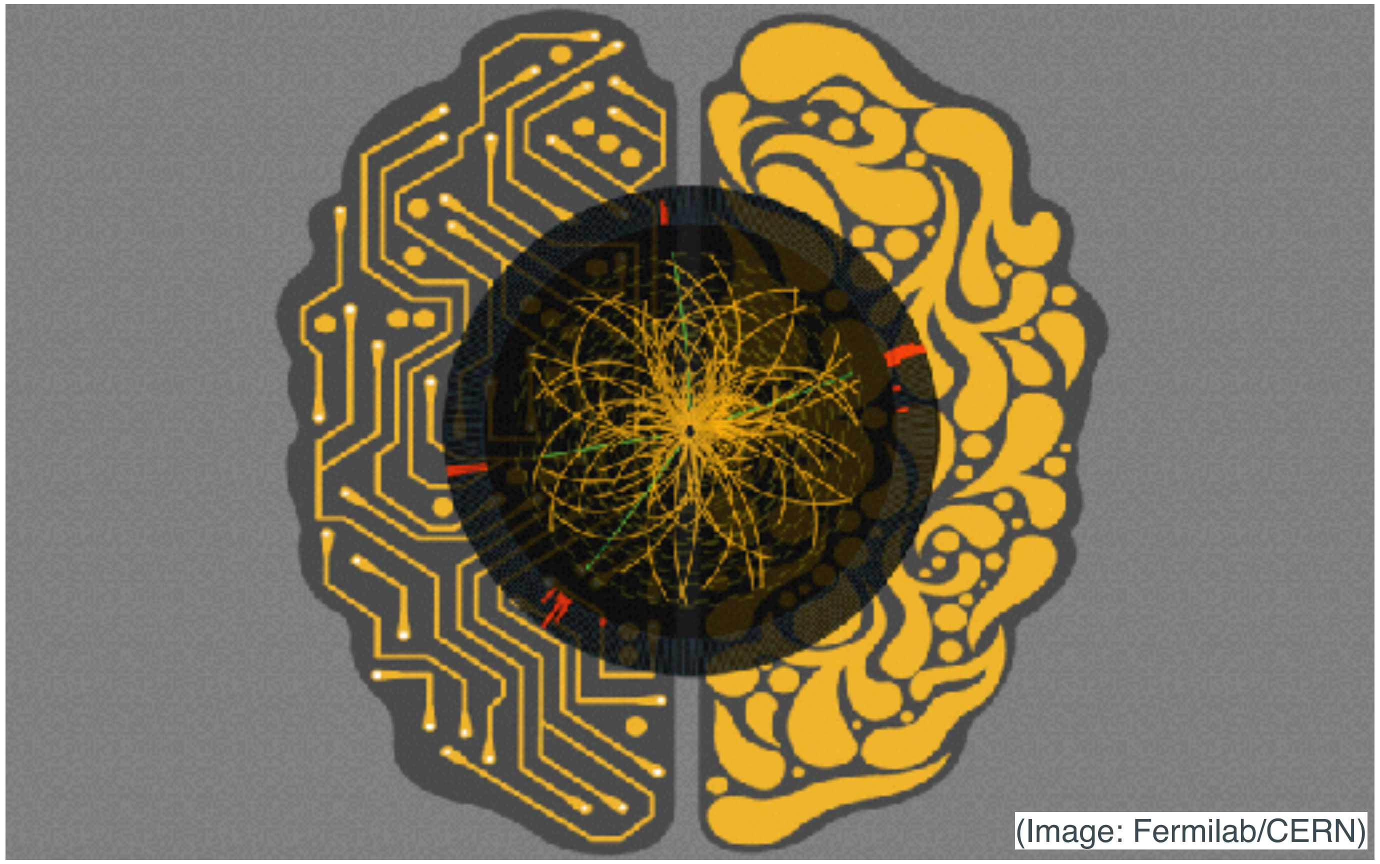
PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



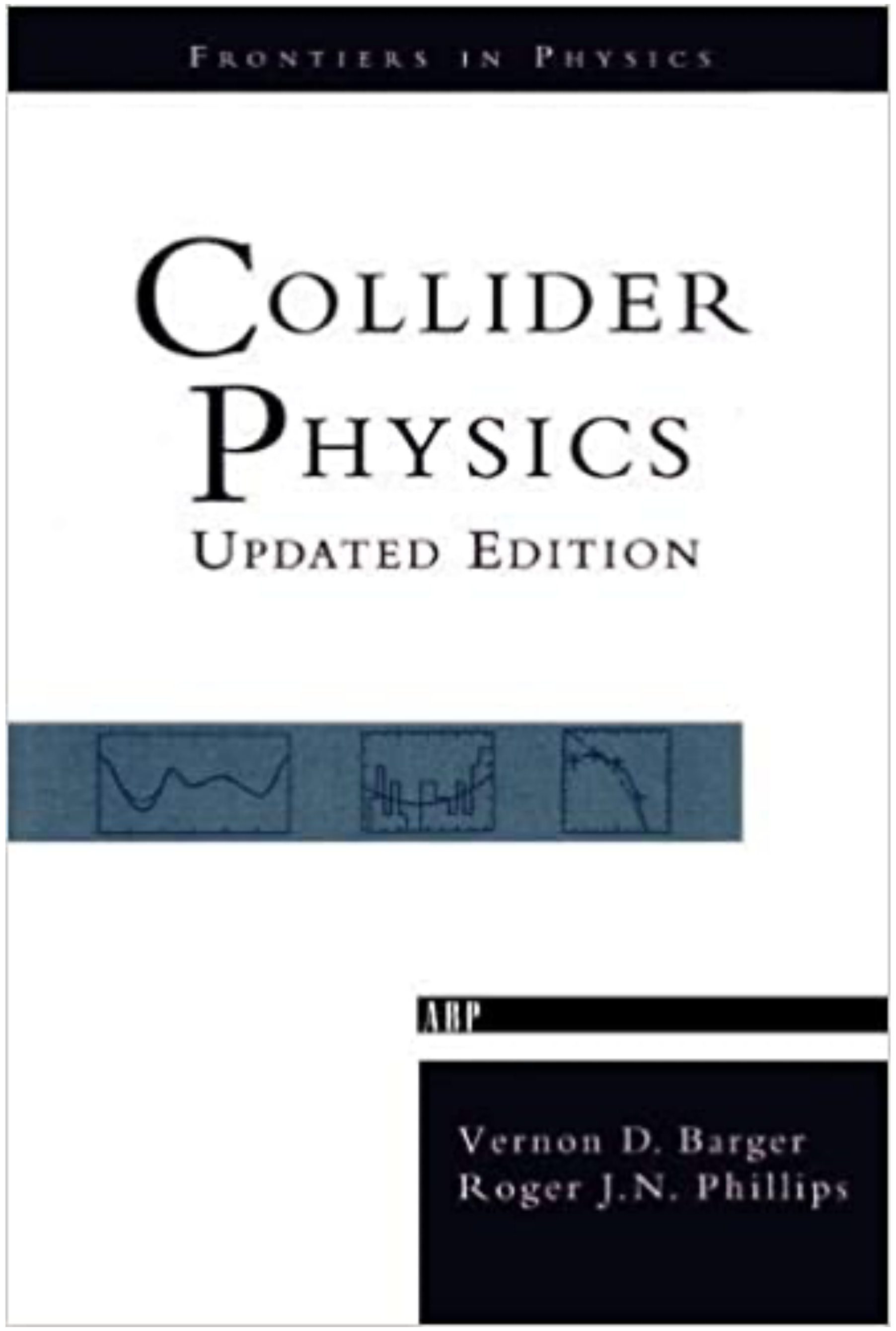
Lecture 1: Overview

Collider Physics

- The goal is to understand the fundamental laws of nature from the high energy scatterings of particles in a complex collider environment.

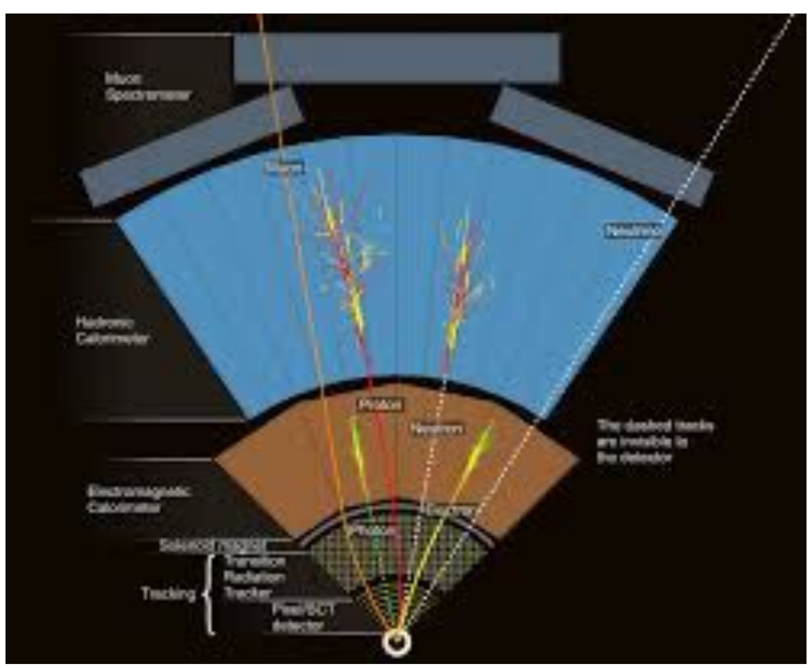
- A multifaceted program: model building, cross-section calculations, kinematic treatment, developing software packages for simulations (including detector effects) and analysis of data.

- An essential bridge between theory and experiment.

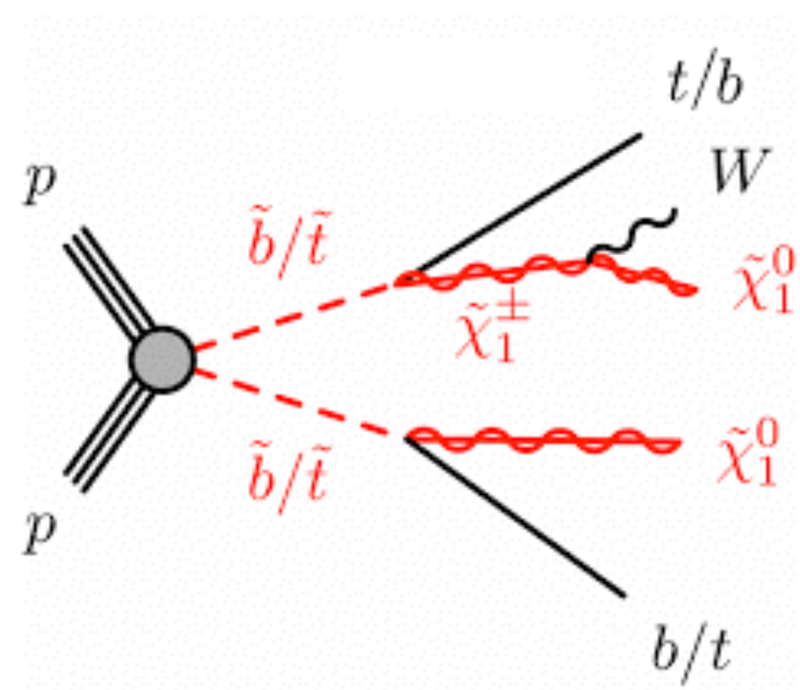


Collider Physics

- It is a living subject, and constantly evolving. In recent years, a great deal of effort is developing ML tools for collider physics.
- Particles collide in the Large Hadron Collider (LHC) detectors (with ~ 10^8 sensors) approximately 1 billion times per second, generating about one petabyte of collision data per second.
- How do we parse this huge amount of data to infer the underlying theory?



Experiment



Theory

Why Machine Learning?

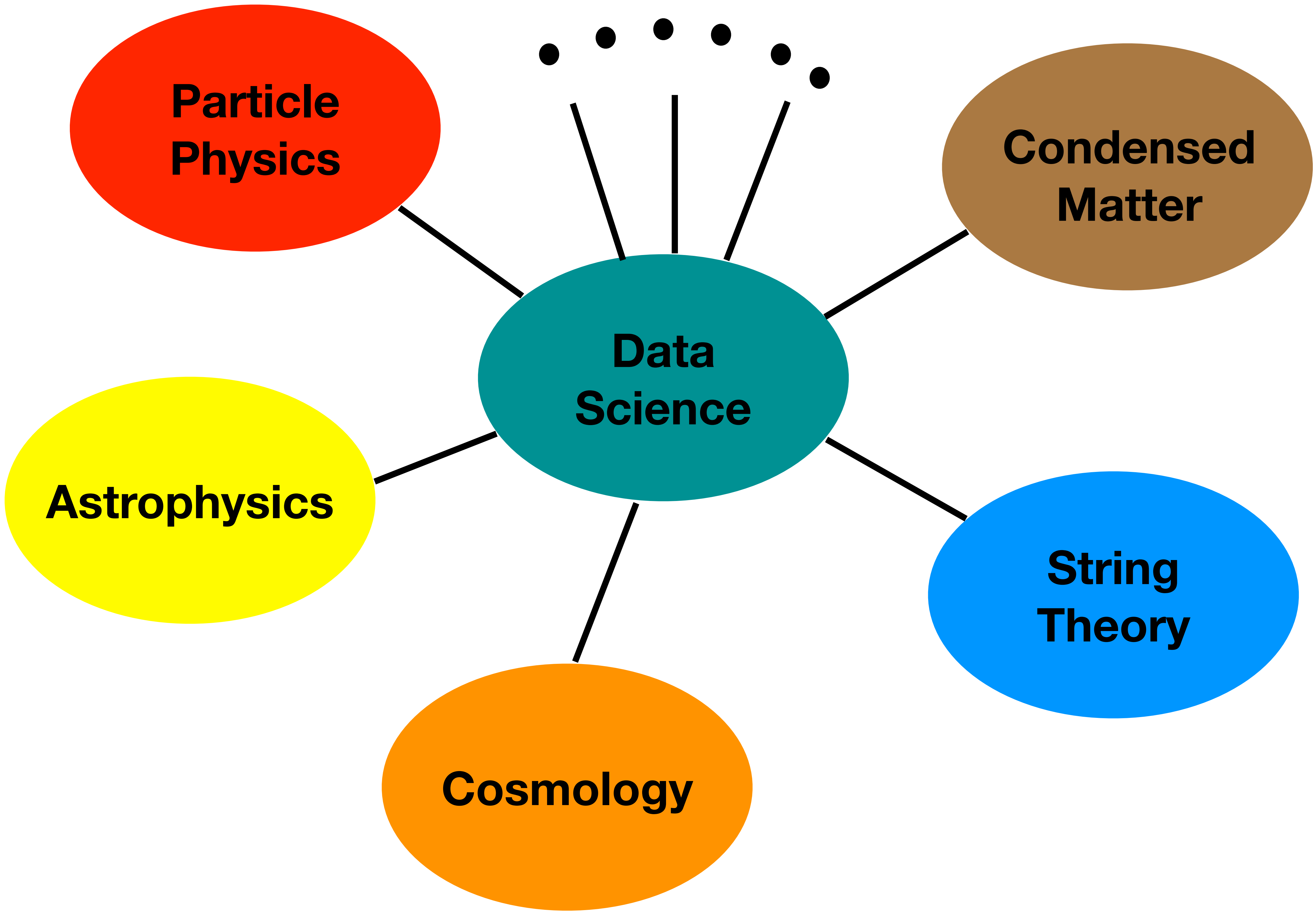
- Analyses of data such as classification, hypothesis testing, regression, and goodness-of-fit testing are based a statistical model $p(x|\theta)$ describing the probability of observing x given the parameters of a theory θ.

- **High dimensionality** and **large volume** of particle physics data make these computationally formidable.

- Traditionally, raw sensor data are processed into low-level objects e.g. calorimeter clusters & tracks. From these low-level components, we use algorithms to estimate the energy, momentum, & identity of particles. Event-level summaries are obtained from these reconstructed objects.

- A central role of machine learning in collider physics is to improve this **data reduction**, reducing the relevant information contained in the low-level, high-dim. data into a **higher-level, smaller-dim. space**.

Unity of Physics



Data is BIG

Cosmology is marching into a big data era:

Experimental Data	2013	2020	2030+
Storage	1PB	6PB	100-1500PB
Cores	10^3	70K	300+K
CPU hours	3x10^6 hrs	2×10^8 hrs	$\sim 10^9$ hrs
Simulations	2013	2020	2030+
Storage	1-10 PB	10-100PB	$>$ 100PB - 1EB
Cores	0.1-1M	10-100M	$>$ 1G
CPU hours	200M	>20G	$>$ 100G

Table taken from 1311.2841

	data volume	schedule
SDSS	40 TB	2000-2020
DESI	2 PB	2019-2027
LSST	> 60 PB	2020-2030
Euclid	>10 PB	2020-2027
WFIRST	>2 PB	2023-2030
CMB-S4	$\mathcal{O}(1)$ (PB)	2020-2027(?)
SKA	4.6 EB	2019-2030(?)

Data is BIG

Cosmology is marching into a big data era:

Experimental Data	2013	2020	2030+
Storage	1PB	6PB	100-1500PB
Cores	10^3	70K	300+K
CPU hours	3x10^6 hrs	2×10^8 hrs	$\sim 10^9$ hrs
Simulations	2013	2020	2030+
Storage	1-10 PB	10-100PB	> 100PB - 1EB
Cores	0.1-1M	10-100M	> 1G
CPU hours	200M	>20G	> 100G

Table taken from 1311.2841

	data volume	schedule
SDSS	40 TB	2000-2020
DESI	2 PB	2019-2027
LSST	> 60 PB	2020-2030
Euclid	>10 PB	2020-2027
WFIRST	>2 PB	2023-2030
CMB-S4	$\mathcal{O}(1)$ (PB)	2020-2027(?)
SKA	4.6 EB	2019-2030(?)

~ 200PB of ***archived data*** in the first 7 years of the **LHC.**

Data is BIG

Cosmology is marching into a big data era:

Experimental Data	2013	2020	2030+
Storage	1PB	6PB	100-1500PB
Cores	10^3	70K	300+K
CPU hours	3x10^6 hrs	2×10^8 hrs	$\sim 10^9$ hrs
Simulations	2013	2020	2030+
Storage	1-10 PB	10-100PB	$>$ 100PB - 1EB
Cores	0.1-1M	10-100M	$>$ 1G
CPU hours	200M	$>$20G	$>$ 100G

Table taken from 1311.2841

	data volume	schedule
SDSS	40 TB	2000-2020
DESI	2 PB	2019-2027
LSST	> 60 PB	2020-2030
Euclid	>10 PB	2020-2027
WFIRST	>2 PB	2023-2030
CMB-S4	$\mathcal{O}(1)$ (PB)	2020-2027(?)
SKA	4.6 EB	2019-2030(?)

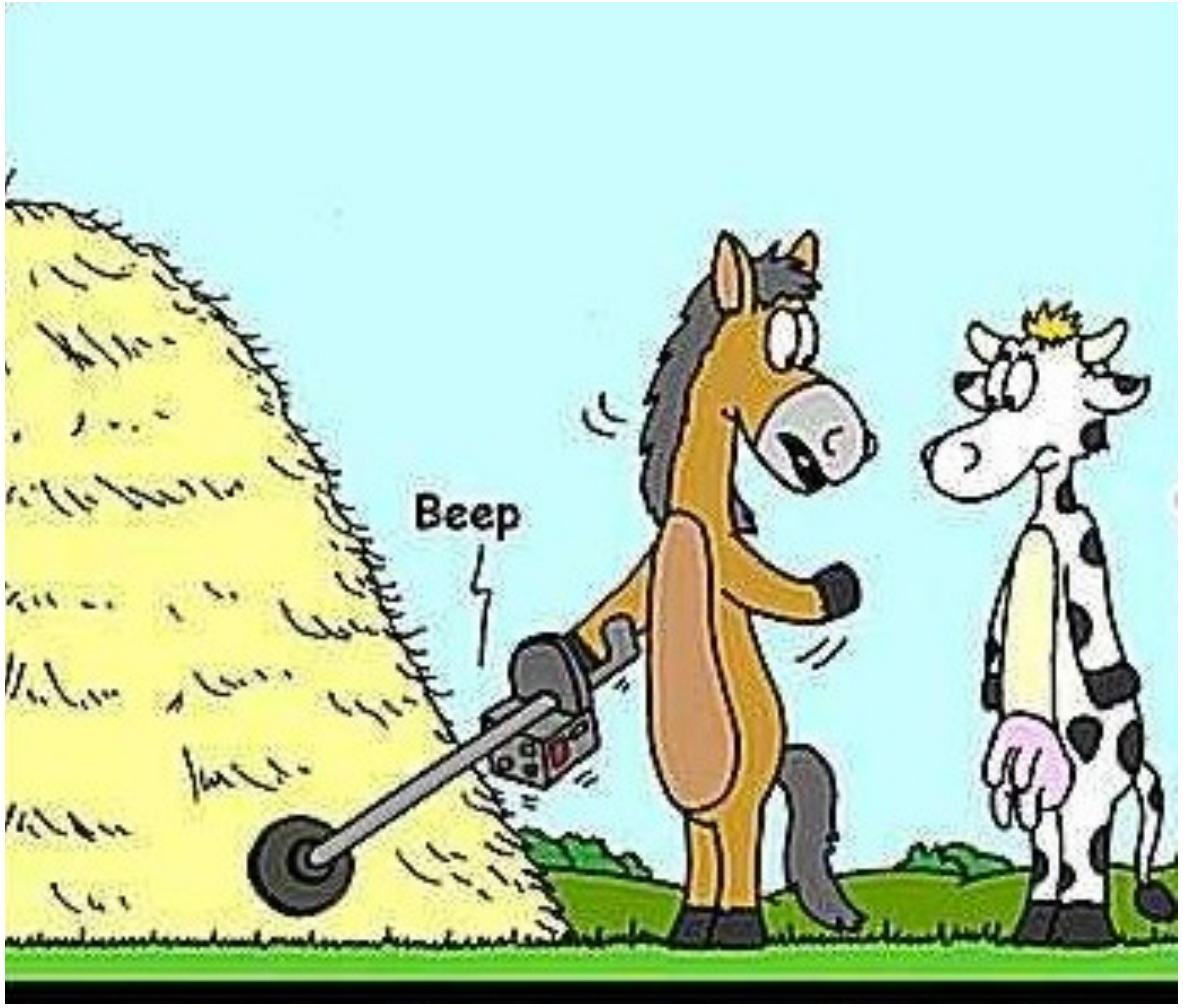
~ 200PB of ***archived data*** in the first 7 years of the **LHC.**

In terms of sheer volume, nothing trumps the volume of ***theoretical data of string vacua***. A rough estimate gives:

$$10^{500} \text{ (Type IIB flux vacua)} \qquad 10^{272,000} \text{ (F theory flux vacua)}$$

Big Dataset

- LHC (raw data/event ~ 1MB), $6x10^8$ events/second.
- GAIA: $1.1x10^9$ stars
- LSST: 10 billion galaxies.
- Searching in large datasets is key. How to find needle in a haystack.
- Automation is much needed to enable analysis of dataset (~getting self driving cars to work).



Astrophysics

- **Galaxy classification:** given an astrophysical observation, which galaxy type do we see?
- Done by human for a long time (e.g., Galaxy Zoo).
- Greatly enhanced by ML: using technology from image classification.

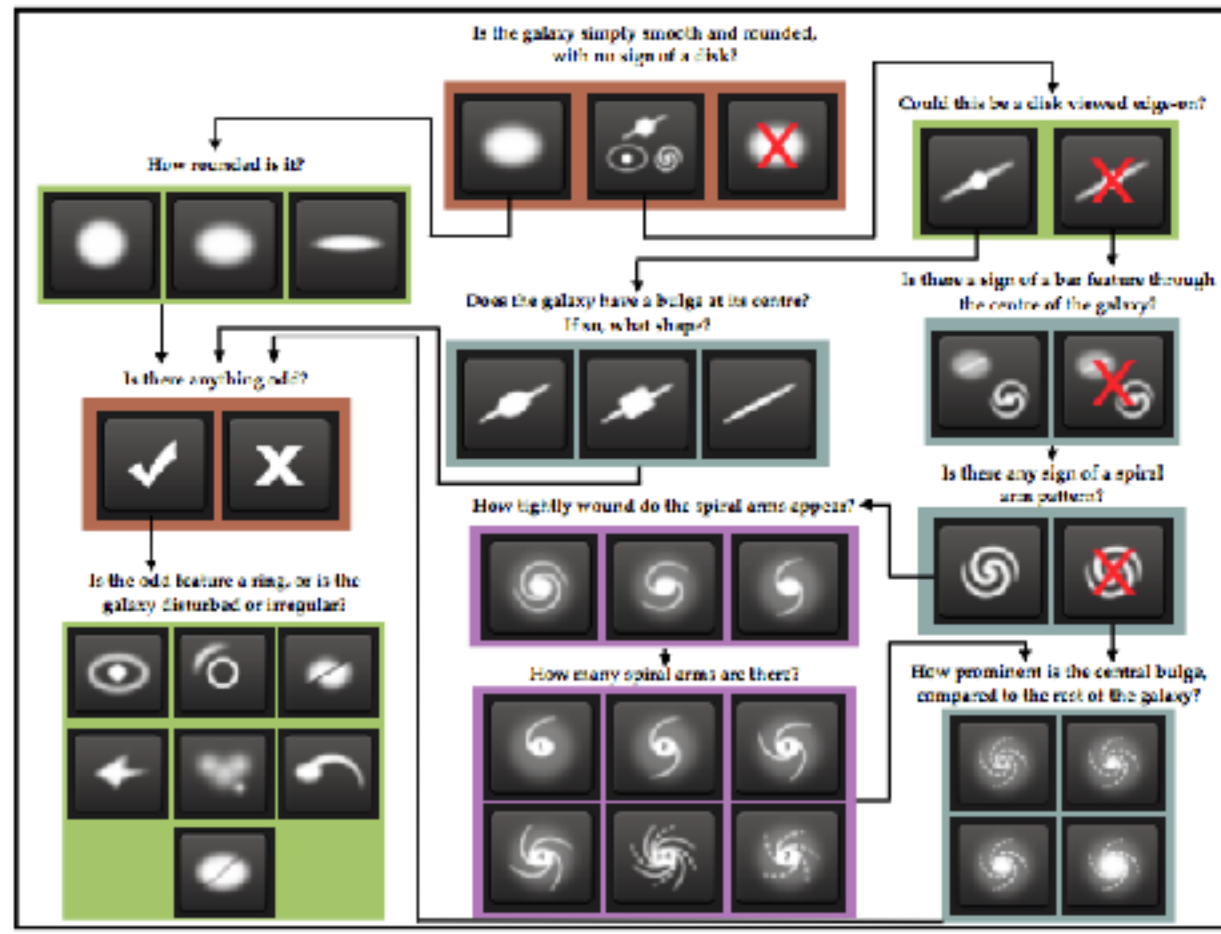


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/overview/description

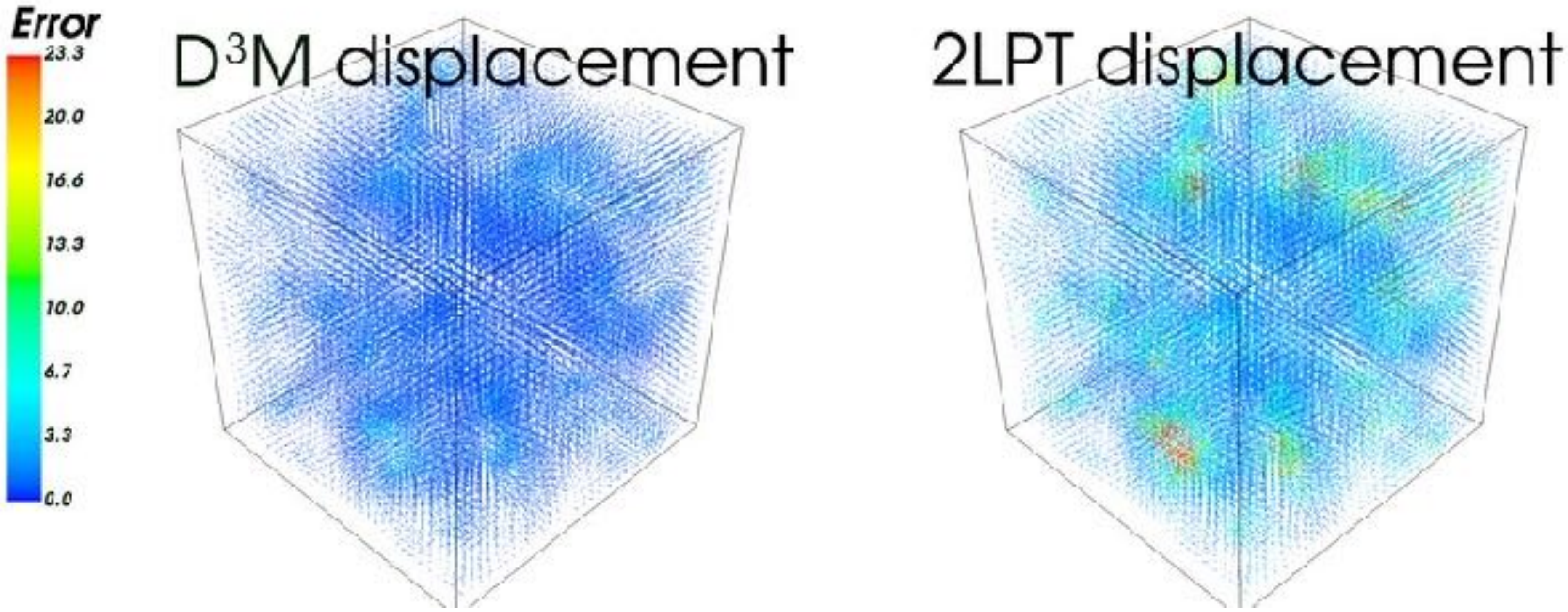
http://benanne.github.io/2014/04/05/galaxy-zoo.html

Accelerating Simulations

- Problem: Generating samples from high-dimensional probability distributions (e.g. to understand structure formation in the Early Universe or expected number of events at the LHC).

- ML offers shortcuts to standard Monte Carlo techniques.

- Relating to image generation, image translation (medical physics)

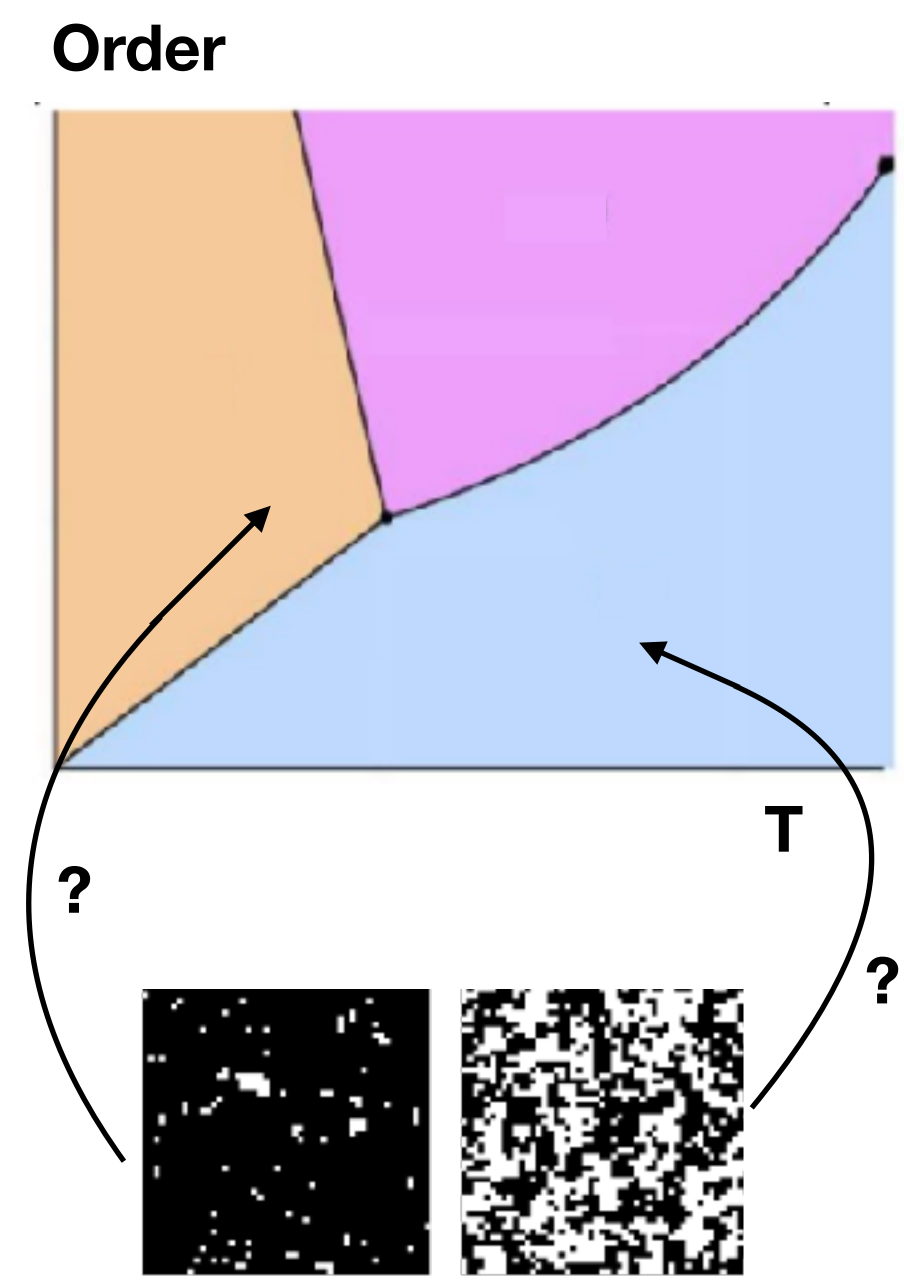


https://www.simonsfoundation.org/2019/06/26/ai-universe-simulation/

How about in Theoretical Physics?

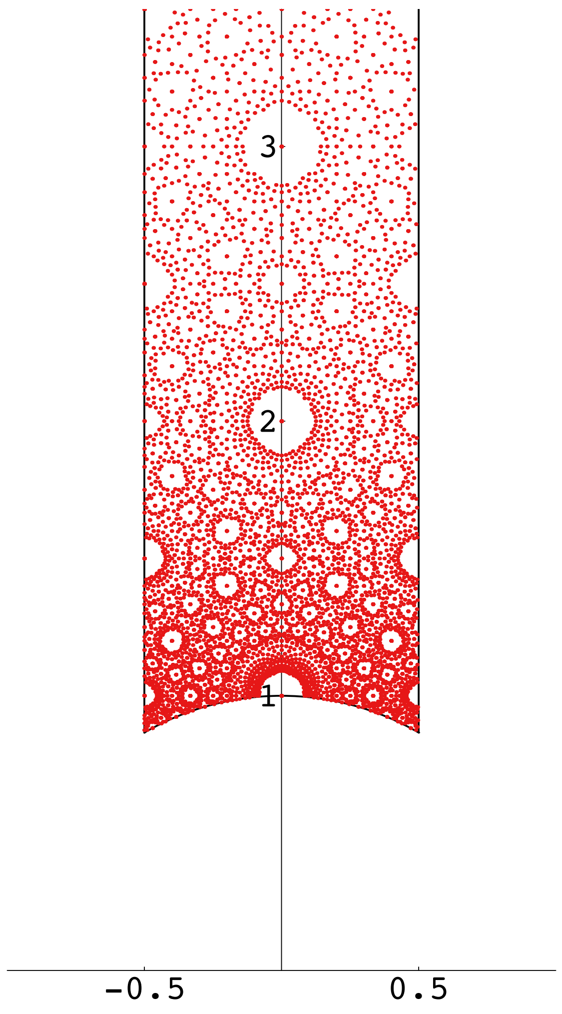
Condensed Matter Physics

- Machine Learning and Physics share deep connections, in particular with statistical/many body physics (Boltzmann machines, softmax, etc).
- Classification of phases of matter: Finding boundaries in the phase diagram.
- Done by humans (e.g., 2D Ising model)
- ML techniques have been developed.
- Again using technology from (image) classification for physical dataset.



String Theory and Mathematical Physics

- Detecting features in string theory solutions.
- Found mostly by analyzing simple examples.
- Can more features be found by ML?
- Finding “good”/relevant features without domain knowledge can be done with “unsupervised” learning (e.g. dimensionality reduction, topological data analysis, ….).
- Large mathematical datasets: Calabi-Yau manifolds (extra dimensions in string theory), …

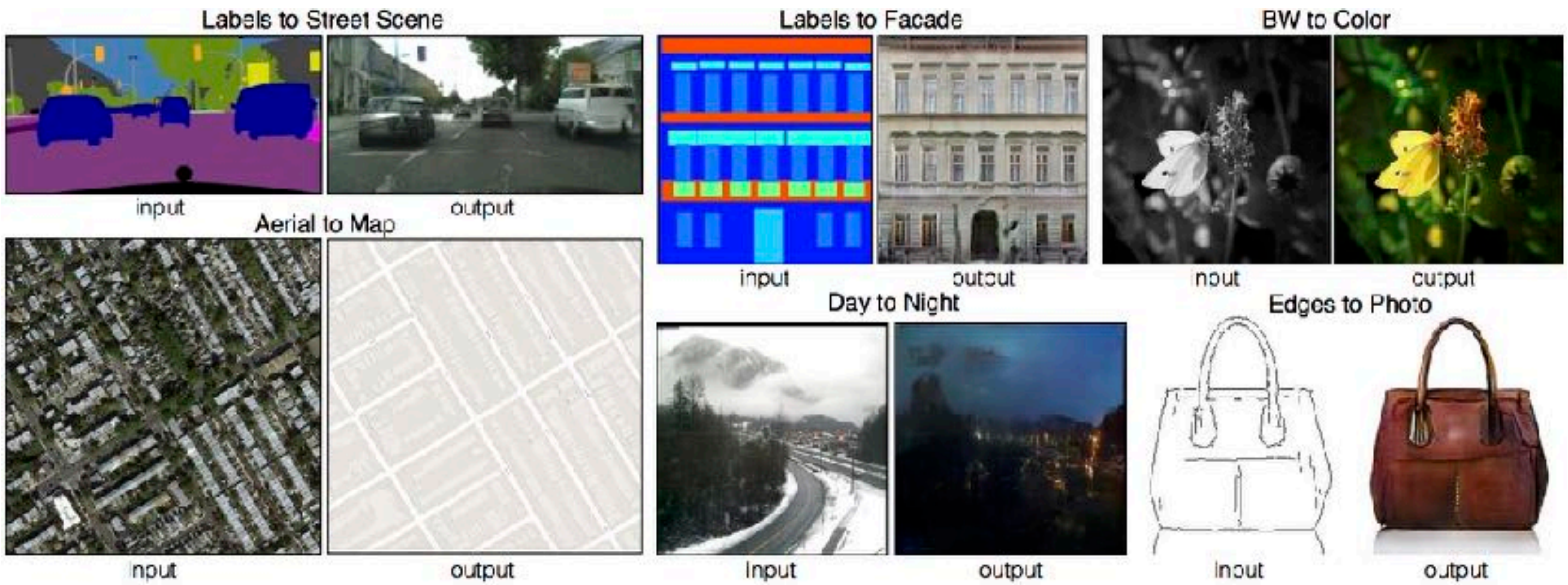


Active area of research with devoted conference series. See e.g. https://indico.cern.ch/event/958074/ for a recent meeting.

Simulations for Theory

- Problem: Generating samples from high-dimensional probability distributions is a ubiquitous problem for any strongly coupled system (condensed matter or QCD).

- Another such unknown distribution are string theory vacua.



https://phillipi.github.io/pix2pix/

What are the goals of this course?

Goals of this Course

- Introduce standard ML tools: you should be able to perform standard ML tasks after this course.
- Programming background is not assumed, only willingness to code. The rest you can pick up from examples...
- This course is not about the fastest implementation of algorithm X, the emphasis is on the concepts rather than efficiencies.
- Discuss examples of physics problems which can be addressed using ML. Hopefully prepare you for research in this direction.

Outline of the Course

- Basic of Machine Learning
- Optimizers
- Regression
- Logistic/Multi-class classification
- A survey of classifiers
- Neural Networks
- Unsupervised learning
- Variational Methods
- Generative Adversarial Networks
- Normalizing Flows
- Reinforcement Learning
- Applications in Physics

References

- Collider Physics (Updated Edition), by Vernon D. Barger and Roger J.N. Phillips
- Deep Learning, by Ian Goodfellow, Joshua Bengio, Aaron Courville
- Information Theory, Inference and Learning Algorithms, by David J.C, MacKay
- A high-bias, low-variance introduction to Machine Learning for physicists, Phys. Rept. 810 (2019): 1-124, by Panjaj Mehta et al.
- Data science applications to string theory, Phys. Rept. 839 (2020), 1-117, by Fabian Ruehle.
- Machine learning and the physical sciences, Rev. Mod. Phys. 91 (2019) no.4, 045002, [arXiv:1903.10563 [physics.comp-ph]], by Giuseppe Carleo et al.

Resources

- ML is a subject that you learn by experimenting – think of this course as a **theory lab** for you to try out various computational, statistical and mathematical methods.
- **Hands on experience** is more valuable than book knowledge. You learn mostly from practical examples.
- Get familiarize with **Python** (mostly python3) and **Jupyter**. Your first assignment is to get to know some commonly used packages.
- Google is your friend. Usually any problem you encounter, somebody else has encountered beforehand. Search for answers!
- Physics ∩ ML is a biweekly seminar series. Please sign up for the mailing list at www.physicsmeetsml.org for zoom links.

Physics ∩ ML

a virtual hub at the interface of theoretical physics and deep learning.

Exercises

- Your grade is based on your participation in the exercises (no exams!)
- The purpose of the exercises is to get you familiarized with the methods/algorithms introduced in lectures and packages installed.
- You are encouraged to discuss with your classmates but you should submit your own solutions (this is how you learn).
- You will be asked to grade each other’s solutions submitted in .ipynb format (so we can test run your code).
- Text-based answers can be included in markdown in these notebooks.
- Many exercises involve plots and it is much more convenient to see them directly in a notebook (think of this as your theory “lab book”).
- Your participation = your solution + your grading of your classmate.

Paper/Presentation

- **Only for those who signed up for 3 credits:** You can give an oral presentation or write up a term paper on a topic related to *Collider Physics and Machine Learning*. I am happy to suggest possible topics.
- The Physics ∩ ML seminar series has many nice talks that would make a good topic for your oral presentation or term paper, e.g.,

Physics ∩ ML

a virtual hub at the interface of theoretical physics and deep learning.

---

13 Jan 2021 — Quantum Machine Learning in High Energy Physics

Sofia Vallecorsa, CERN, 12:00 EDT

20 May 2020 — Building symmetries into generative flow models

Phiala Shanahan, MIT, 12:00 EDT

Summary

- Be able to tell a friend examples of problems where ML can be used in collider physics (and physics in general).
- Where can ML be useful in Theoretical Physics?
- How can a physics problem be related to identifying cats and dots on images
- Remember to start installing the software packages (Exercise 1) and get familiarized with them.