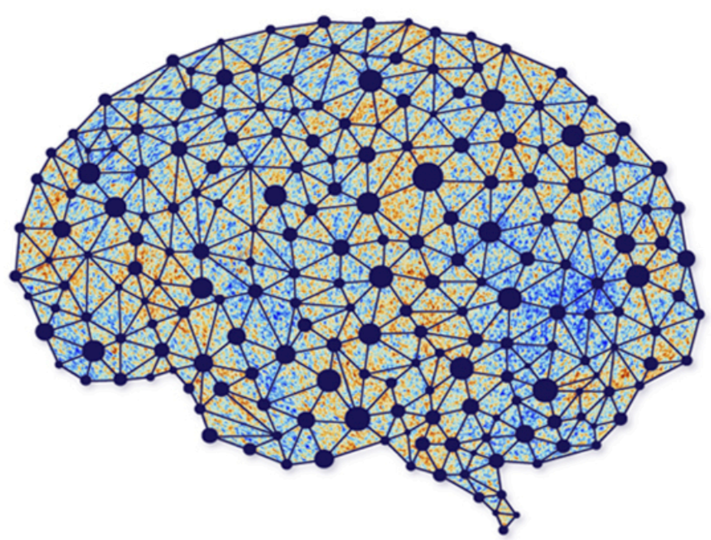
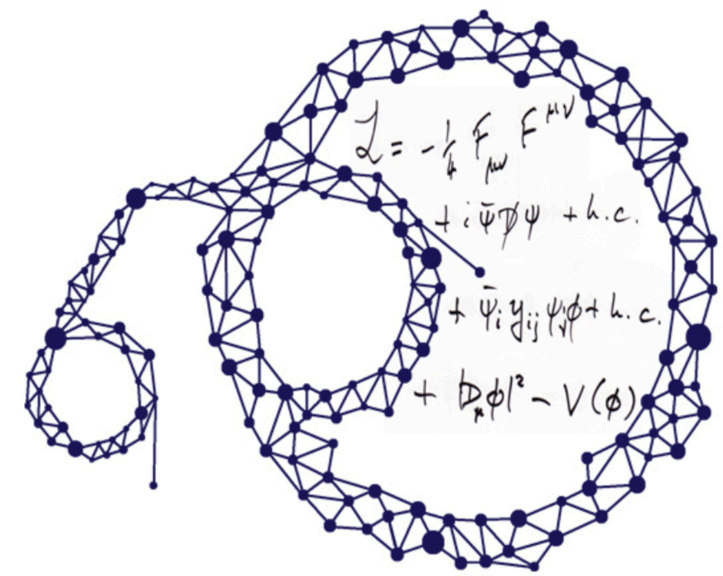
Physics 361 - Machine Learning in Physics

Lecture 1 – Introduction

Jan. 21st 2025



AI
∩
Universe



Moritz Münchmeyer

Instructors

Lecturer: Prof. Moritz Münchmeyer

- Computational and theoretical Cosmologist
- My research: http://munchmeyer.physics.wisc.edu/
- Office: 6205
- Email: muenchmeyer@wisc.edu

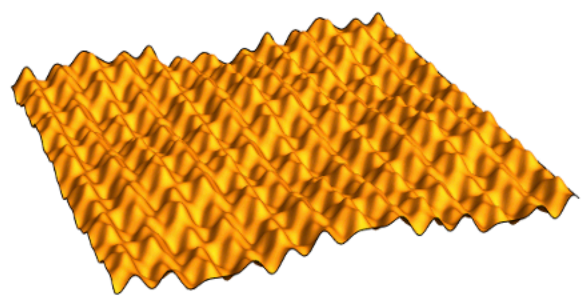
TA: Yurii Kvasiuk

- Computational and theoretical Cosmologist

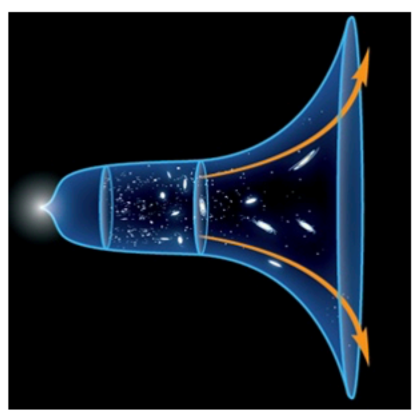
My area of physics: Cosmology

Fundamental physical quantities

“Big bang”: Primordial particles and interactions (Quantum mechanics, General Relativity)

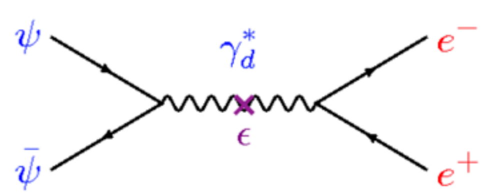


Expansion of the universe (General Relativity)

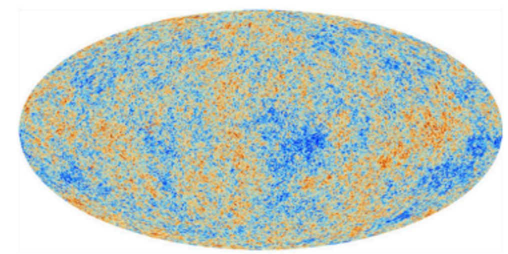


Discover magazine

Properties of matter and radiation (Special Relativity, Quantum Mechanics)



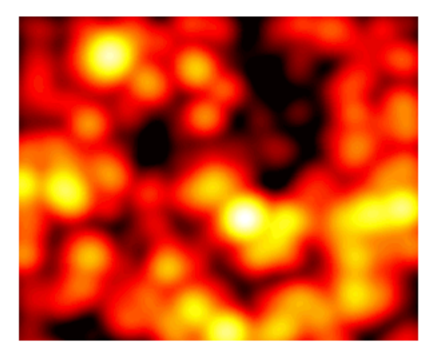
Observations



Cosmic Microwave background

NASA – Hubble deep field

Galaxy surveys



Kovetz et. al. 2017

Light emission in different wave lengths

inference

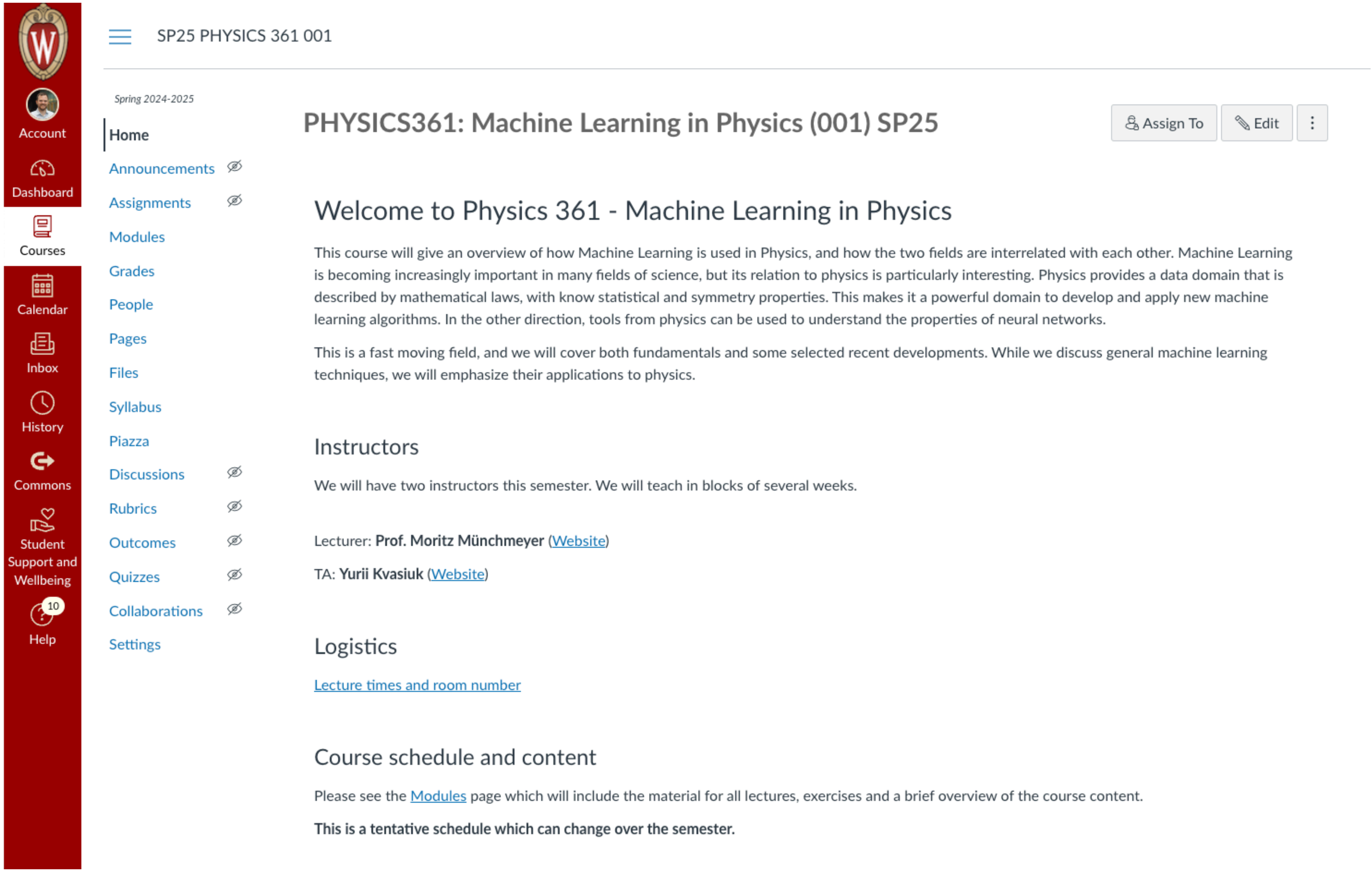
Tell us about your experience

- We have a very mixed class in terms of preparation of students.
- We want to adjust the level of the course to make it appropriate for the majority of students.
- Let’s gauge your your experience with:
    - Programming (python)
    - Machine learning
    - Physics
    - Math

Course logistics

Course Canvas Page

The schedule and all lectures and exercises will be uploaded on the course canvas page https://canvas.wisc.edu/courses/450587



If you are comfortable with it **please upload a profile picture of you to canvas** to help us learn your names.

Grading

- **Problem sets**
  - There will be approximately 8 to 10 problem sets. Download and upload on Canvas. Late submission requires prior permission.
- **Final project:**
  - Apply some of the techniques we learn to a concrete project of your choice. Write a course paper.
  - Final grade will be made out of problem sets and final project. Tentative weighting: 50/50 problem sets and final project.

Text books and Reviews

- We won't follow a single textbook closely. However these references can be used to supplement the course notes:
  - **Classics on General ML**
    - **Bishop: Pattern Recognition and Machine Learning** (classic)
    - **MacKay: Information Theory, Inference, and Learning Algorithms**, CUP (free online version)
    - **Goodfellow, Bengio, Courville: Deep Learning**, MIT Press: deeplearningbook.com
  - **Recent popular textbooks on General ML**:
    - https://udlbook.github.io/udlbook/ **Understanding Deep Learning**
    - https://d2l.ai/ **Dive into Deep Learning** (fully online and with code)
  - **ML in Physics**
    - Mehta, Bukov, Wang, Day, Richardson, Fisher, Schwab: **A high-bias, low-variance introduction to Machine Learning for physicists** (1803.08823)
    - Carleo, Cirac, Cranmer, Daudet, Schuld, Tishby, Vogt-Maranto, Zdeborova: **Machine Learning and the Physical Sciences** 1903.10563
    - Kaplan: **Notes on Contemporary Machine Learning for Physicists**
    - Acquaviva: **Machine Learning for Physics and Astronomy**

Learning tools

- We will be using **Piazza for questions and answers**.
    - You are encouraged to ask your questions on Piazza, rather than e.g. by email to your instructors, so that everybody can discuss, answer, and learn from the answers.
    - You can post questions and answers anonymously if you prefer.
- **Computational tools**:
    - We will use **Jupiter or Google Colab.**
    - The programming language will be **Python.**
    - I will upload a basic tutorial for python which you can work through. But if you have not worked with python you will need to do some self study. This is well invested time.

Planned schedule of topics

- We will cover the following units, each with application to physics:
  - **Probability theory and Information theory background**
  - **Basics of Machine Learning**
    - **Optimization and Regularization**
    - **Basic Architectures**
  - **Generative models: Diffusion, Normalizing Flows, Flow Matching**
  - **Simulation-based inference & Uncertainty Quantification**
  - **Learning on graphs and other data structures**
  - **Transformers, LLMs, Foundation Models, Reasoning**
  - **Solving Inverse Problems and PDEs with NN**
- These are a lot of topics so we will cover them somewhat briefly.
- Since this is a new class, the list of topics may evolve over the semester.

Why a class on Machine Learning in Physics?

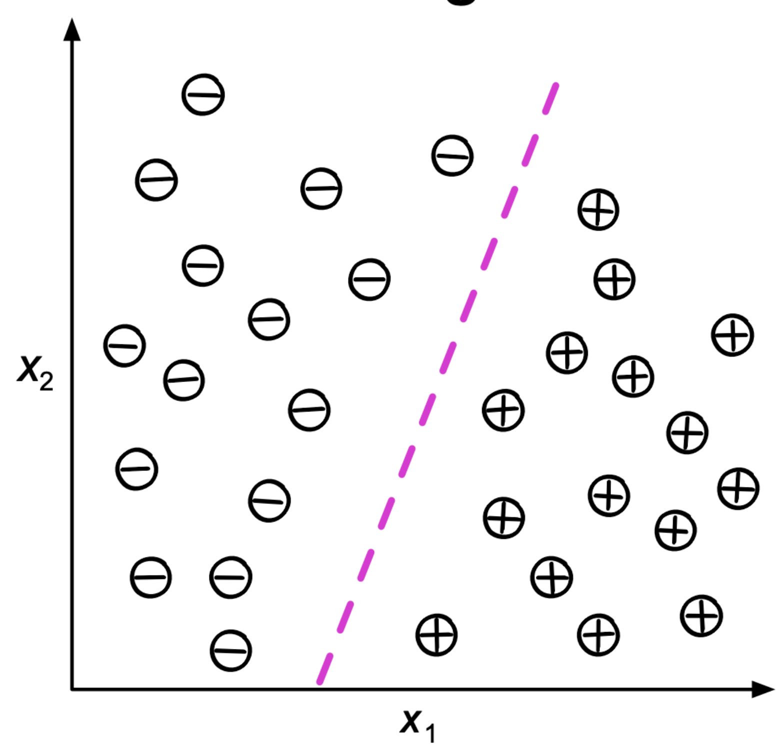
- At the very least, Machine Learning is a tool, like Likelihoods and MCMC, of which **every physicist needs to know the basics** now.
- However “Black box” applications of Machine Learning in Physics without insights from domain experts usually don’t work. We need to **understand how to use domain knowledge**.
- Physics provides a data domain that is described by mathematical laws, with know statistical and symmetry properties. This means that we can often **combine analytic methods with machine learning**.
- There are many ingenious uses of machine learning in physics, that go far beyond training standard ML models on physics simulations.
- Sometimes it is possible to **solve previously intractable real world problems**.
- There are also theoretical connections between physics and machine learning.
- In the future, we may have “AI physicists” that can do science independently, but we still seem far away from that. This course: Understand the state of the art.

Broad Overview of Machine Learning

Categories of Machine Learning

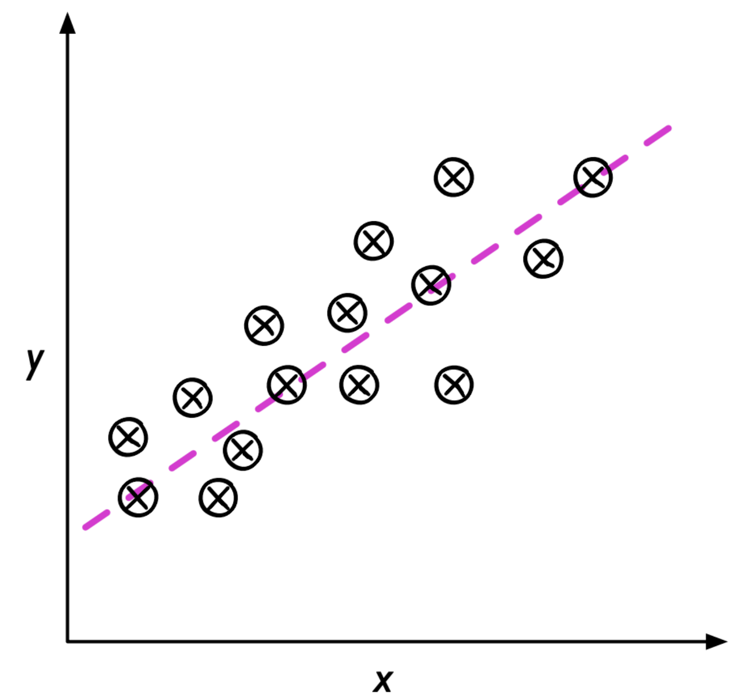
- **Supervised Machine Learning**
  - Labelled data
  - Direct feedback
  - Predict label (regression or classification)
- **Unsupervised learning**
  - No labels/targets
  - Find hidden structure of the data
  - Includes clustering and generative models
- **Reinforcement learning**
  - Decide on actions to take
  - Rewards and environment change with time
- The lines between these can be blurred. For example there is **weakly supervised learning** (with partial or noisy labels).

Supervised Learning: Classification



Find what class a data vector belongs to, e.g. classify pictures into cat, dog, car etc.

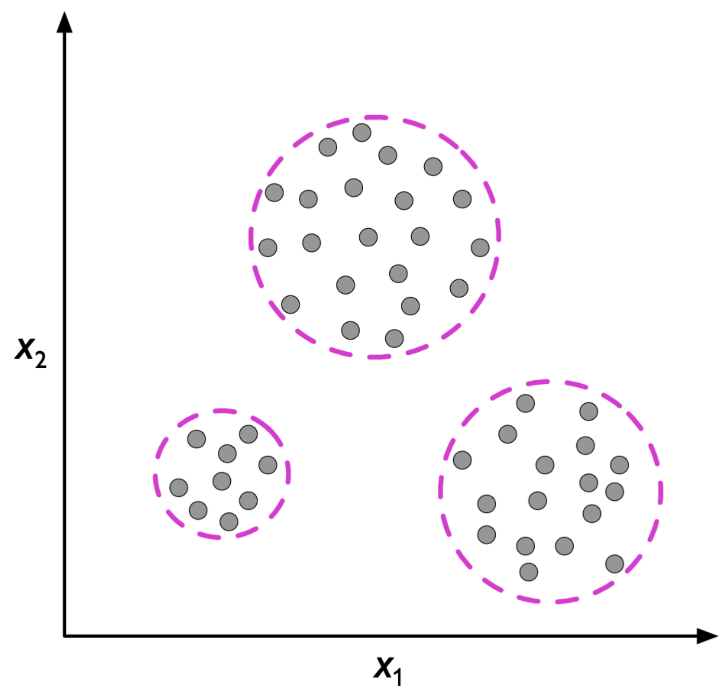
Supervised Learning: Regression



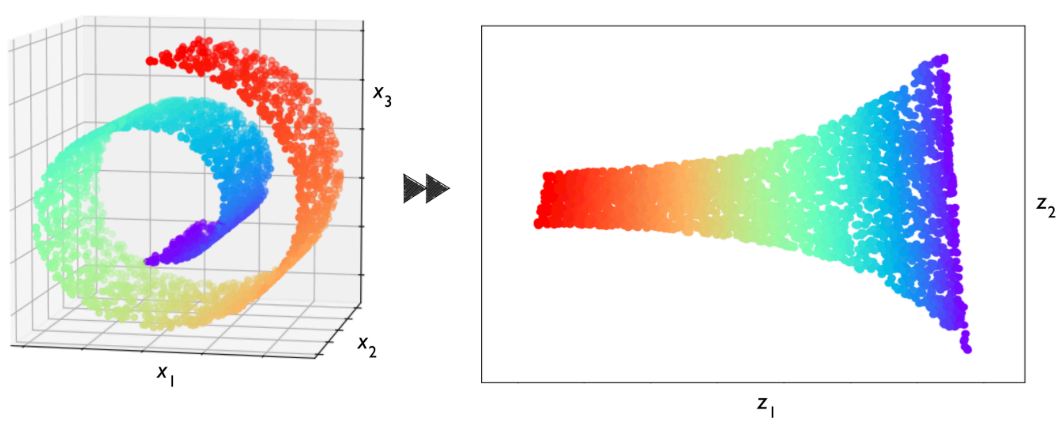
Measure an (often real valued) parameter from the data. E.g. given a picture of a car estimate its value.

Plots: sebastianraschka.com

Unsupervised Learning -- Clustering

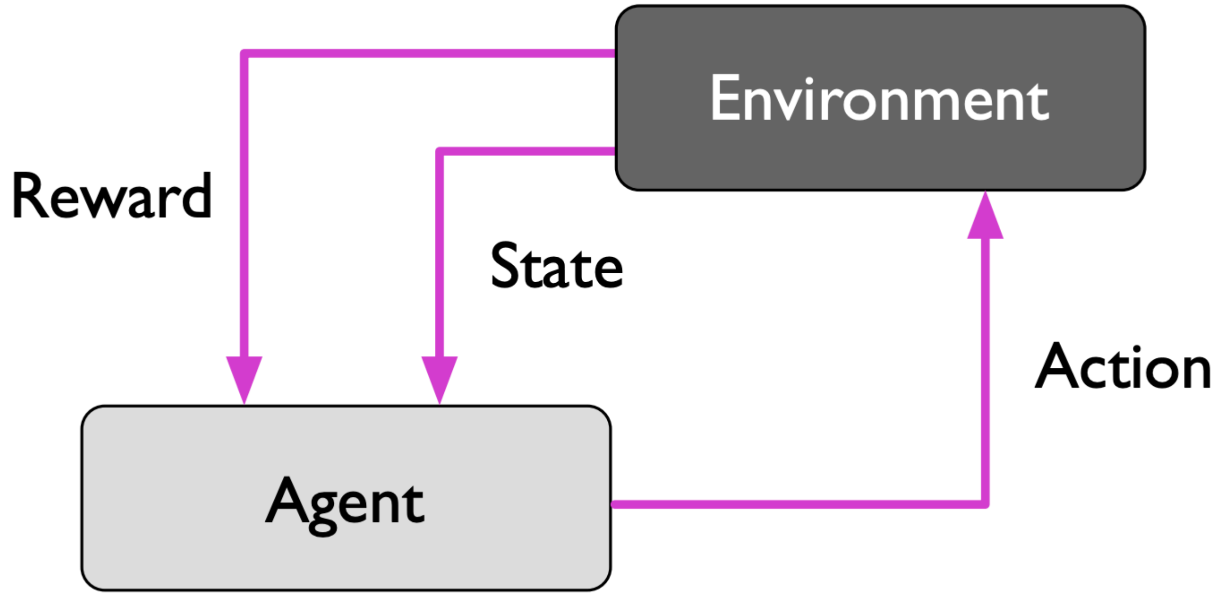


Unsupervised Learning -- Dimensionality Reduction



Unsupervised learning — Generative modelling

Reinforcement Learning



Generative vs discriminative models

Informally:

- **Generative models** can generate new data instances.
- **Discriminative models** discriminate between different kinds of data instances.
- A generative model could generate new photos of animals that look like real animals, while a discriminative model could tell a dog from a cat.

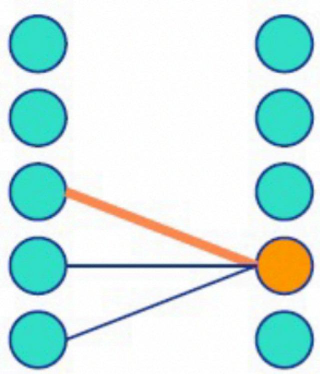
More formally, given a set of data instances X and a set of labels Y:

- **Generative models** capture the joint probability $p(X, Y)$, or just $p(X)$ if there are no labels.
- **Discriminative models** capture the conditional probability $p(Y \mid X)$.

There are many types of generative models, including **GANs, VAE, diffusion models and normalizing flows**.

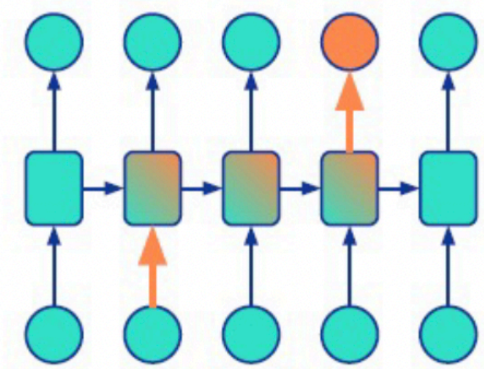
https://developers.google.com/machine-learning/gan/generative

Examples of neural network architectures



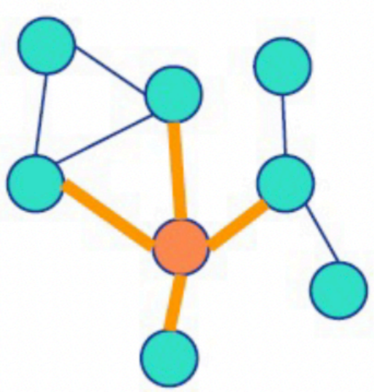
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours



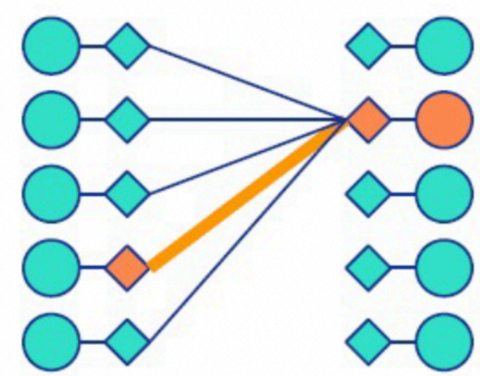
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges



Attention Module (e.g. language)

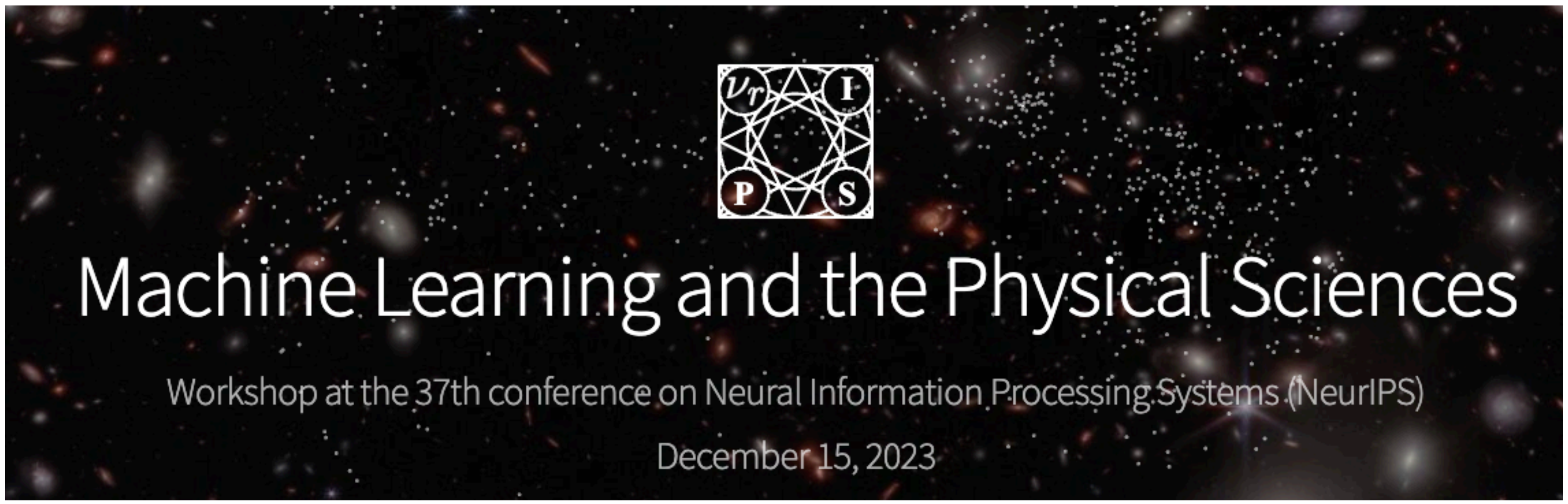
- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

We will meet all of these architectures later in the course.

Some Examples of ML in the Physical Sciences

Where to see recent work?

- To get an overview, check out for example this yearly NeurIPS physics workshop:
  - https://ml4physicalsciences.github.io/2023/ (yearly workshop)



  - At the end of the semester I hope you will be able to understand what many of these papers are about, at at superficial level.
- Other workshops that are relevant:
  - https://ai4sciencecommunity.github.io/neurips23.html AI for Science: from Theory to Practice
  - A few others: https://neurips.cc/virtual/2023/events/workshop
- I'll mention other conferences and specific papers later on.

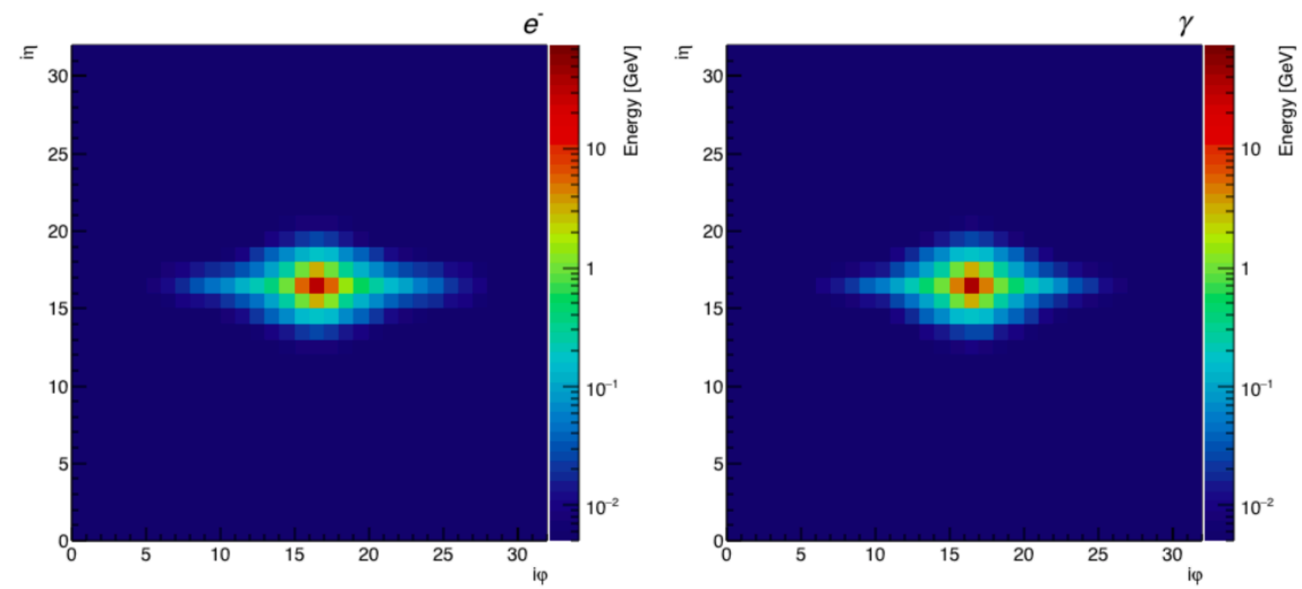
Classifying Events and Objects

- Examples:

  - LHC particle collisions. ML has a long history in particle physics, reaching back several decades.

Arxiv: 1807.11916
End-to-End Physics Event Classification with CMS Open Data

(Here and below I select papers somewhat at random, there are MANY other good papers in each domain)



(a) Average electron shower (b) Average photon shower

Figure 1: e/γ showers averaged over 50k showers. The e shower is slightly more spread out in ϕ-in addition to being slightly asymmetric-due to bremsstrahlung effects.

  - Ice cube particle shower classification. E.g. 2209.03042

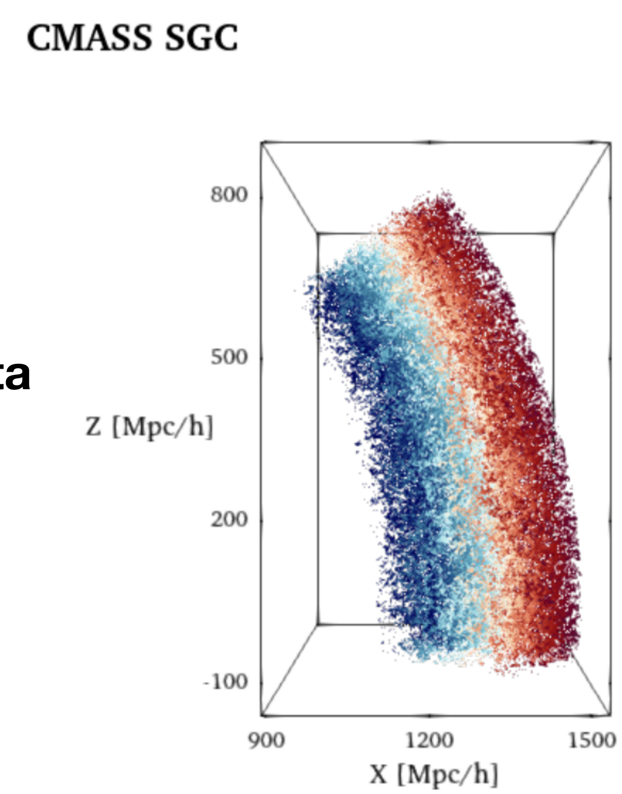
  - Galaxy type classification. In the past, different galaxy types were classified by researchers by eye. Not possible with millions of galaxies.

Measuring physical parameters

- It is often not clear how to measure a parameter from a collection of data.
- If we have reliable simulations, we can train a neural network to perform the measurement, using supervised learning.
- Example: Measuring cosmological parameters (age of the universe, amount of dark matter etc.) from a galaxy survey

SimBig project 2211.00723

Galaxy data



→ **CNN** → **Parameter Measurements**

- Main challenge: Reliability of training data.

Simulation-based Inference

- When we measure parameters, we also need error bars (or better the full posterior).
- **Simulation-based inference is the process of finding parameters of a simulator from observations, probabilistically**.
- In “traditional” data analysis in physics we often make analytical assumptions of the statistics of an observable, most commonly that it is Gaussian distributed.
- With machine learning one can **learn the probability distribution of observables from simulations.** In a Bayesian analysis, the likelihood or the posterior can be learned from simulations.
- This is usually done using a **Neural Density Estimator, such as a Normalizing flow**.
- See e.g. arxiv:1911.01429 The frontier of simulation-based inference

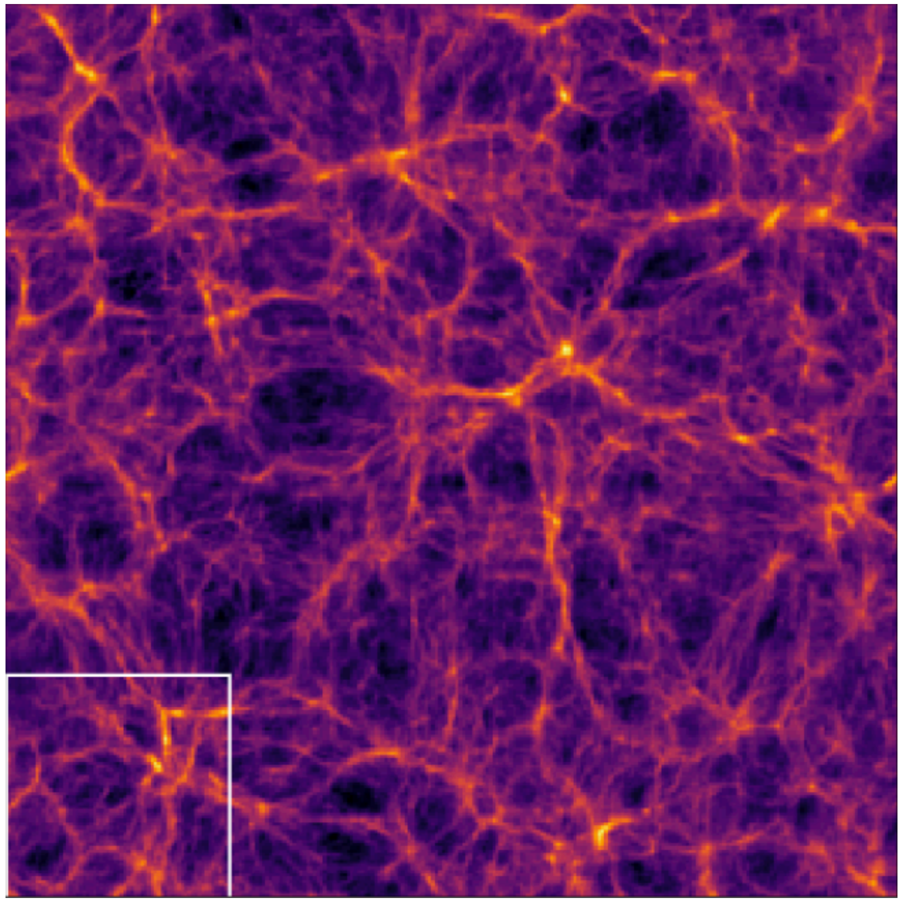
Generating Simulations / Emulators

- Neural networks can be used as **surrogate models to replace computationally expensive simulations.** These are often called Emulators.

- Once trained on data or simulations, an emulator can make new “simulations” much faster.

  - Example from my own research:

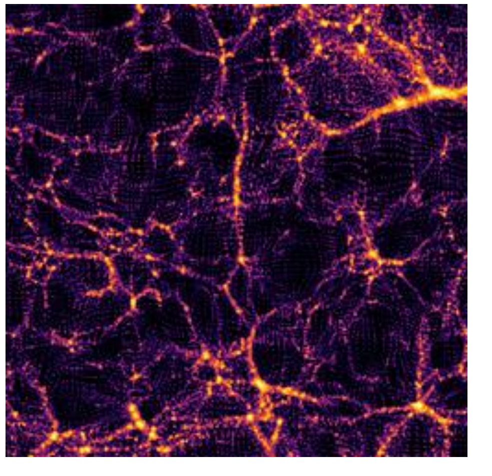
    - Generating 3d simulations of the matter distribution of the universe using a diffusion model. (Arxiv: 2311.05217)



- Machine Learning is often used to **speed up classical methods**.

Auto-differentiation without ML

- To train neural networks, computational techniques were developed that can train models with **billions of free parameters.** This is done with auto-differentiation libraries such as
  - **PyTorch**
  - **JAX**
  - **Tensorflow**
- This software is useful in physics **even if you don't use any machine learning**.
- Physicists **re-write their codes in auto differentiable language**, which allows efficient optimization with respect to any parameters. Some examples from my field:
  - CosmoJax, a differentiable cosmology library
  - Differentiable cosmology simulations, e.g. pmwd



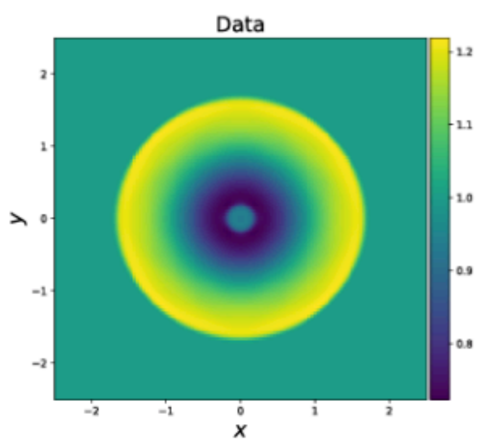
Clustering and Anomaly detection

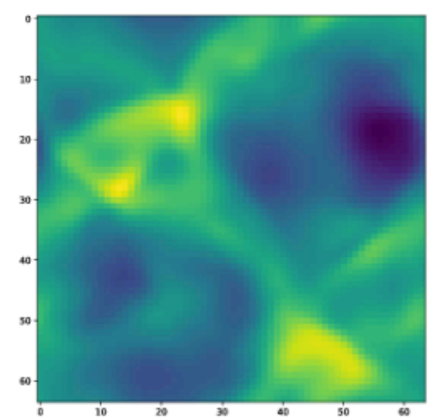
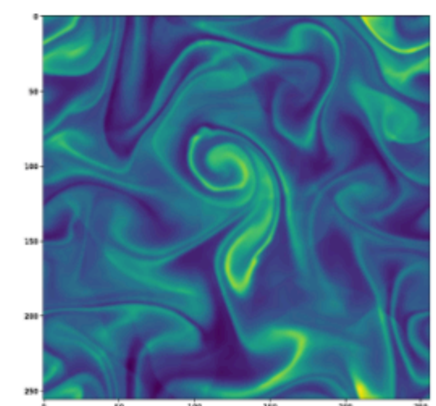
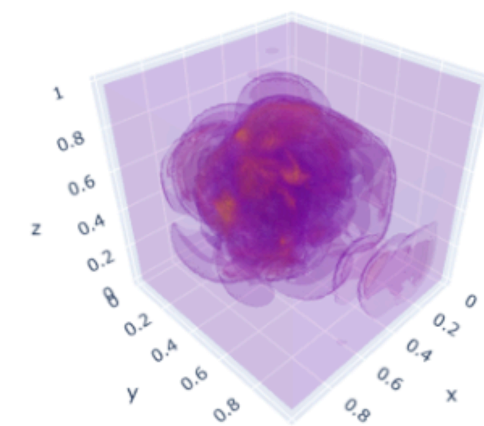
- **How can we organize a large data set of events or objects into classes of similar objects? Clustering and dimensionality reduction algorithms.**
  - Classic k-means is still very useful! E.g. stellar populations.
  - Clustering can also happen in the “latent space” of a generative model.
  - Data visualization, e.g. t-SNE
- **How can we find something “new” without knowing what to look for? Anomaly detection!**
  - Humans are pretty good at anomaly detection by eye, but data sets are too large to be inspected that way and the anomaly may only be visible in the right data representation.
  - Anomalies have been found in archival data, long after the data was taken (example: Fast Radio Bursts). Perhaps there is something exciting hidden in existing data.
  - Unsupervised learning can be used to classify existing events or objects. If an object is not close to any known class, it is flagged as an anomaly.

Solving PDEs and Inverse Problems

- Many problems in physics amount to **solving a complicated set of partial differential equations (PDE)**. There are various ways to use NN for that.

  - Examples (from the PDEBench data set):



- In **Inverse Problems**, one wants to find the input data that produced a specific output. That can mean removing noise or undoing a non-linear evolution. Often they are ill-conditioned and need to be regularized.

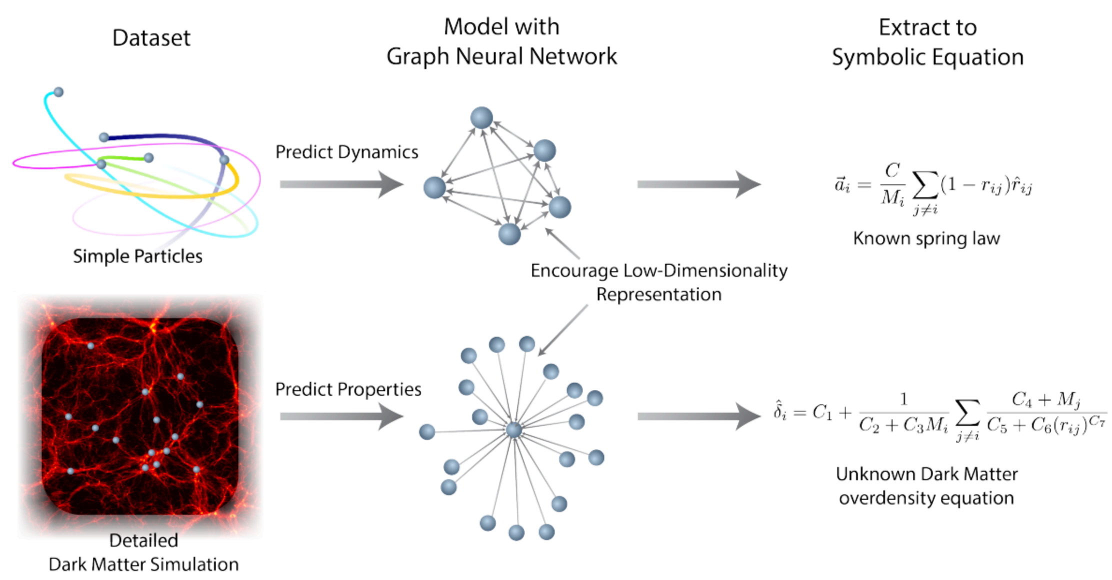
- Neural Networks are being trained to solve such problems approximately.

Symbolic methods

- Theoretical insight in physics come in the form of symbolic expressions. Naturally, combining machine learning and symbolic expressions is an exciting direction.

- Machine learning can be used to improve **symbolic regression**, the process of finding mathematical expressions that describe data.

    - Example: 2006.11287



- Machine learning can come up with **novel proofs and novel solutions**. A large-language model can make “educated guesses” (proposed solutions) that are then verified with a systematic evaluator. e.g. https://www.nature.com/articles/s41586-023-06924-6

Course logistics

- **Reading:**
  - Familiarize yourself with the course canvas page.
  - Check out some of the textbooks and reviews on slide 7.

- **Problem set**: No problem set in the first week