Welcome to PHYS591000
Hands-on AI for Physics
物理與人工智慧(AI)實作導論

Spring 2022

National Tsing Hua University

清華

- Time : 10:10-13:00 Wed
- Location : General Building II (綜二館) R521
- Web : https://nthu-phys591000.github.io/AIPHYS2022/

Announcement

- This course is offered in English.
- No additional sign-ups. (不加簽). If you wish to try you may keep adding the class through the University system during the course shopping period, in case somebody withdraws.
- Access to Gen. Bldg II 5F (綜二5F門禁開放) during the class will be granted after the shopping period.

Course Description

- The course “Hands-on Artificial Intelligence (AI) for Physics” is a project-based upper division course aimed to provide practical skills to perform research with neural networks.

- Each class is divided into a lecture part and a ‘lab’ part.

Class Format

- For in-class exercise and labs: work in groups of two assigned by instructors; rotate every week.

- Experience-based learning:
  - Work out a set of in-class exercise during lectures

- Post-lecture ‘Lab’:
  - Carry out hands-on assignments (homework)

Class Format

- More about the ‘Lab’ part:
  - Submit one homework (lab) for two people
  - Lab due at noon on Friday of the week
    - The Lab part gives you a chance to work with your teammate with TA’s around.
    - You can submit the lab by the end of the class.
    - You don’t need to stay till 1PM; Be sure to have a plan with your teammate to finish the lab before Friday noon before you leave class.

Class Format

- Final project:
  - Apply machine learning techniques on data provided by instructors (Details will be provided later)
  - Find your own teammate (2 people/team).
    - At least two teams will be working on the same project
    - A kaggle competition will be setup for each project
    - An oral presentation (in English) for each team. Both members have to speak.

Teaching Staff

Instructor

Pai-hsien Hsu
徐百嫻 (Jennifer)

Teaching Assistants

Yi-Lun Chung
鍾沂倫 (Alan)

Chiau Jou Li
李巧柔

Prerequisite

- Python  (assume you have heard/are familiar with Numpy, Pandas, Matplotlib, etc.)

- Will give you an idea of the level needed in today's exercise

Grading Policy

- Class Participation - 10%
  - Headcount (出席) + participation in discussion
- Lab/Homework - 60%
  - 10-12 sets of assignments
  - Due: **Noon on Fridays** (No late submissions)
- Final Project - 30%
  - Oral presentation: 15%
  - Final project results: 10%
  - Kaggle competition: 5%

Grading Policy

- Note: 10 points (10% of final grades) off for absence without permission for the following events

  - Guest lecture (Week 15 05/25; TBC)
  - Final project discussions (Week 16 06/01)
  - Final project oral presentations (Week 17 06/08)

- Do **not** attend the class if you have respiratory symptoms (呼吸道症狀). Just email us.

Tips for ‘enjoying’ this course

- Do not expect that everything will be covered in lectures
  - Ask Google and discuss with you teammate
- It is very important that you *learn how to* work together
  - In real world you usually work with others for a project
- Take a minute to know your teammate for today!
  - E.g. ‘Do you have experience working with AI?’

Textbook

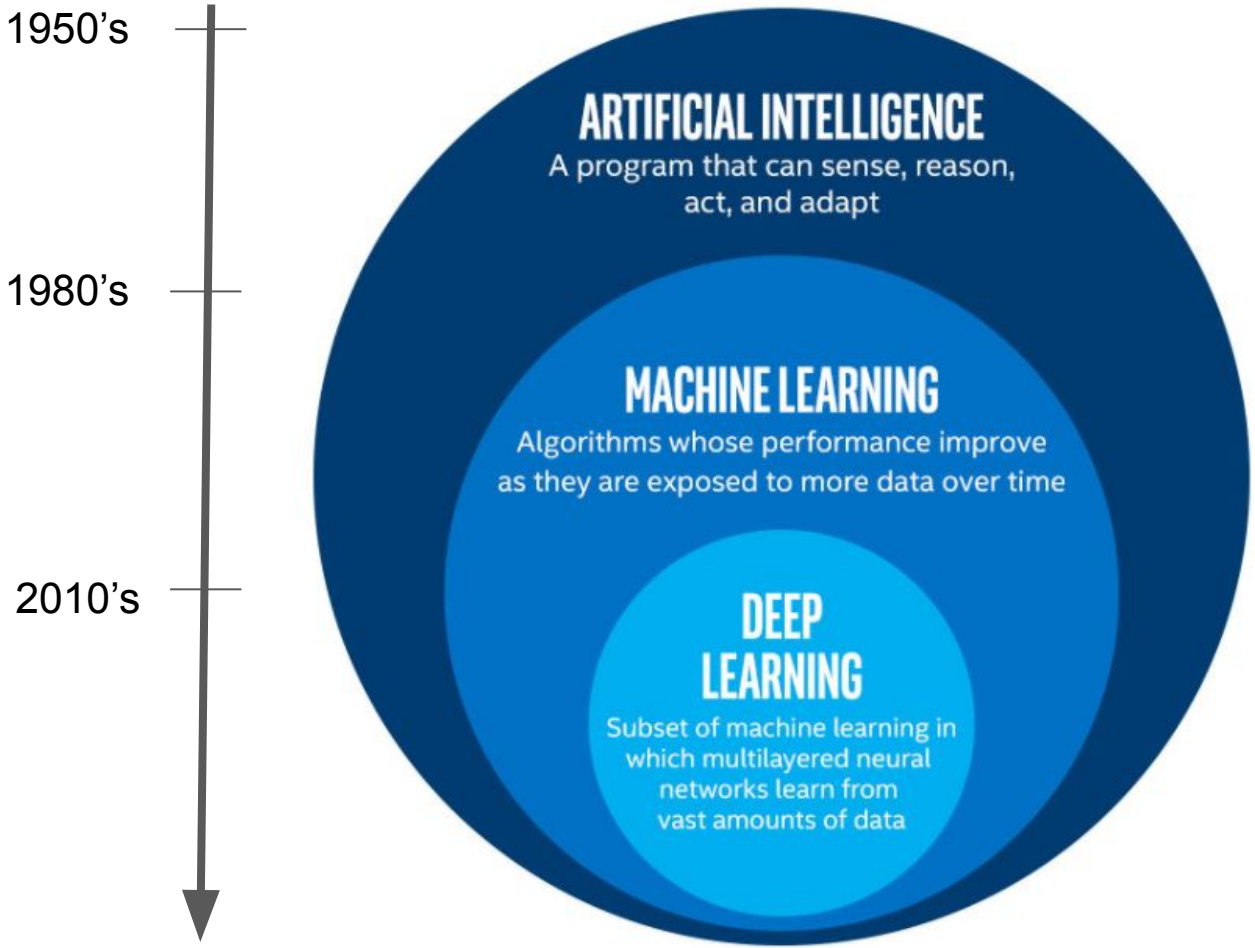
- No official textbook
- Useful on-line references
  - https://nthu-phys591000.github.io/AIPHYS2022/resource.html
  - Deep learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
  - Machine Learning Yearning by Andrew Ng (practical concepts; available online)
  - A Course in Machine Learning by Hal Daume III (Introduction; available online)
  - Deep Learning with Python by Francois Chollet (learning through examples; Keras)
  - Kaggle Courses https://www.kaggle.com/learn/overview

AI?

Machine Learning?

Overview: What is AI?

Deep Learning?



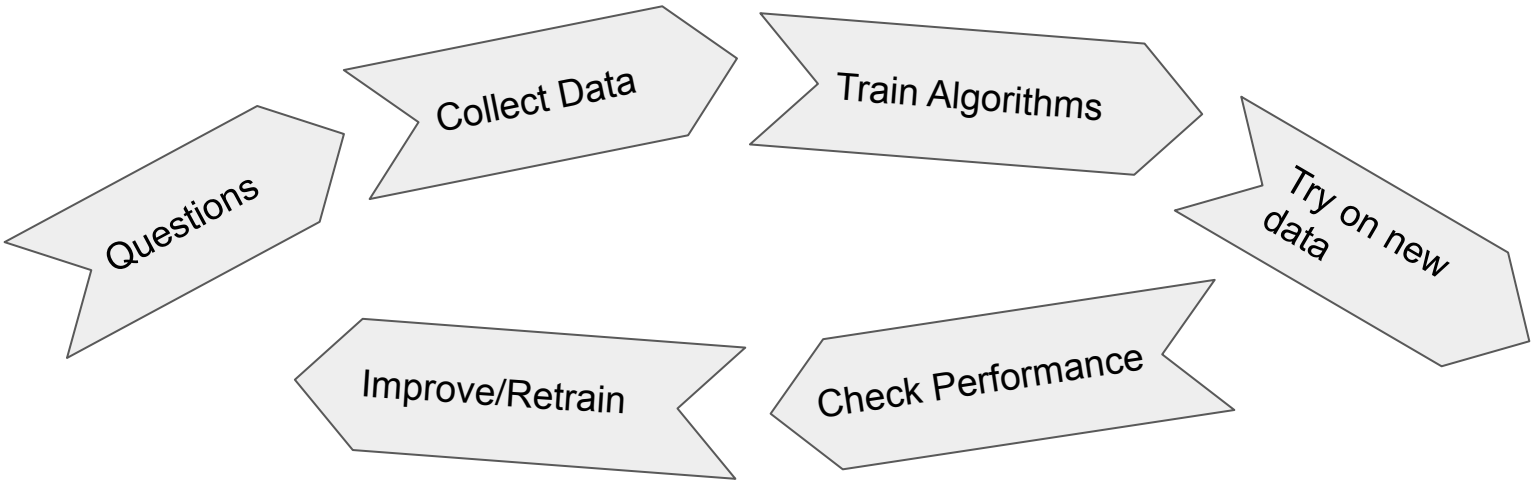
techplusmedia

Artificial Intelligence (AI)

- The science of training machines to perform human tasks
  - Understand human speech
  - Pattern/image recognition
  - Play strategic games (e.g. GO, chess)
  - Drive cars
- An ‘old’ idea since 1950’s

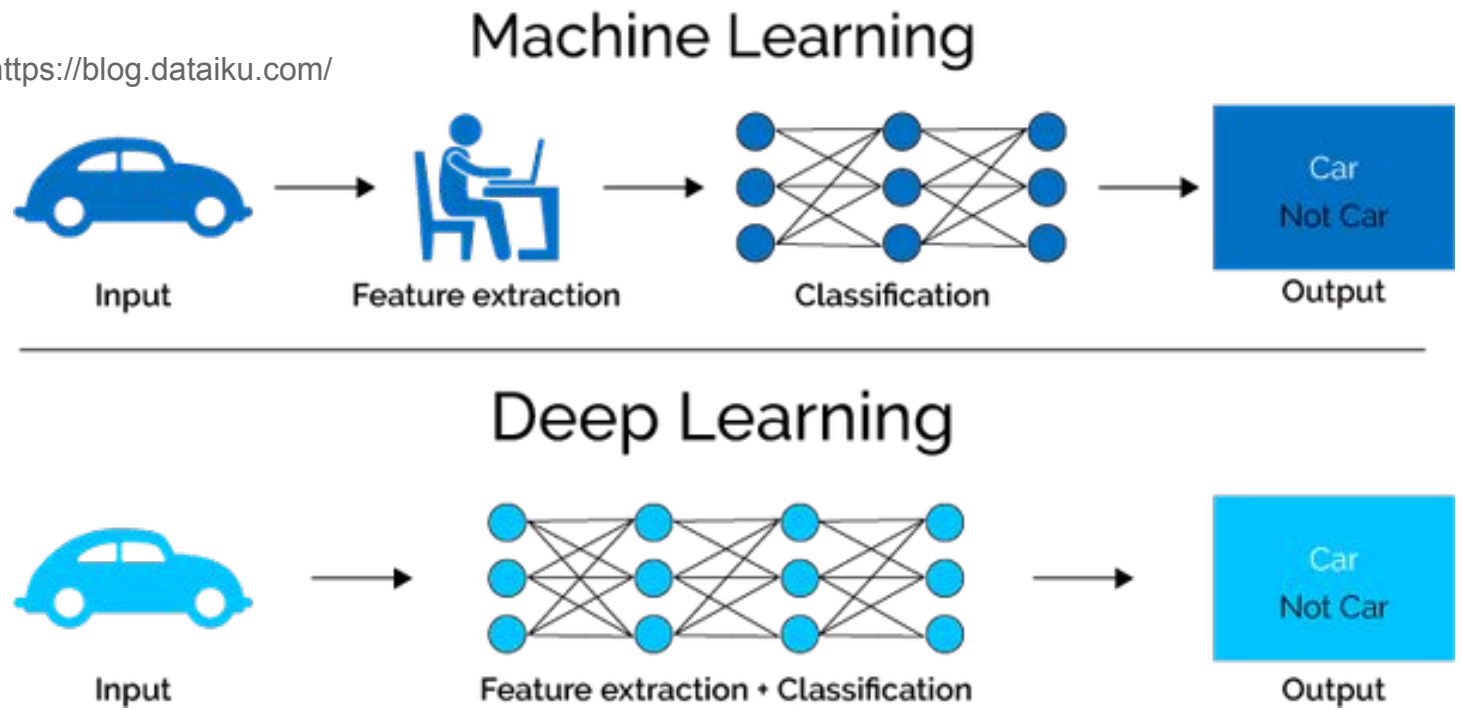
Machine Learning

- A subfield of AI which trains machine how to learn from data *without being explicitly programmed*

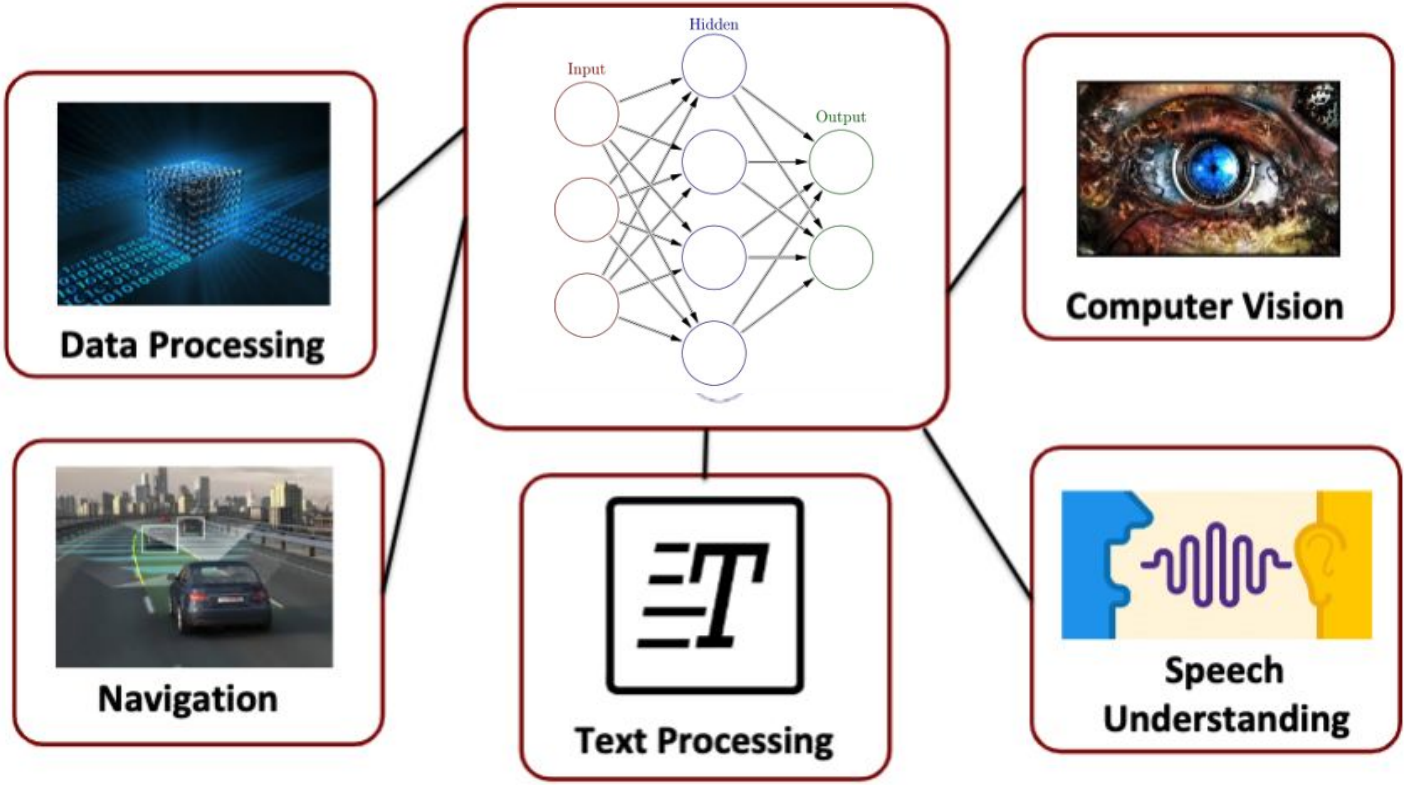


Classical Machine Learning vs Deep Learning

Source: https://blog.dataiku.com/



Artificial Neural Network



Course schedule

- Before Spring break (Week 8; no class): ‘Classical’ machine learning and introduction to neural networks
  - Goal: Prepare you for your final project

- After Spring break: Variations of neural networks
  - Goal: Tools/ideas for your final projects

- Guest lecture + final project discussions and presentations

AI applications in Physics

- In this course we focus on questions in physics. For example,
  - How to tell a signal event from a background event in particle experiments?
  - What is the true energy of a particle given the measurement made?
  - Which stars/objects belong to the same galaxy?

AI applications in Physics

- In this course we focus on questions in physics. For example,
  - How to tell a signal event from a background event in particle experiments?

    -> **Classification**
  - What is the true energy of a particle given the measurement made?

    -> **Regression**
  - Which stars/objects belong to the same galaxy?

    -> **Clustering**

AI applications in Physics

- In this course we focus on questions in physics. For example,
  - How to tell a signal event from a background event in particle experiments?

    -> **Classification** **Supervised learning**
  - What is the true energy of a particle given the measurement made?

    -> **Regression** **Supervised learning**
  - Which stars/objects belong to the same galaxy?

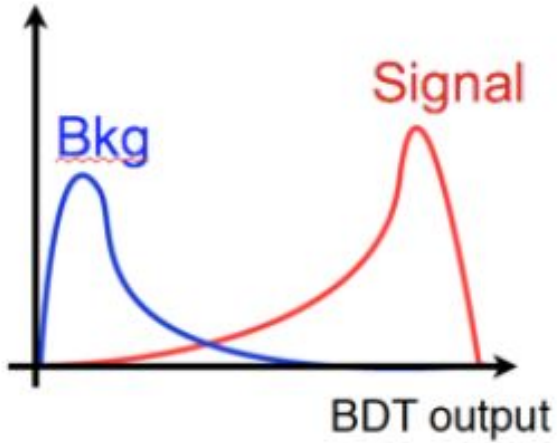
    -> **Clustering** **Unsupervised learning**

Supervised Learning

- The right answer is given in the training data (‘labeled’)

  E.g. Train a boosted decision tree (BDT) with data labeled as signal and background

  Classification: to make *discrete* predictions (True/False, Signal/Background, Type I/II/III Supernovae, etc.)

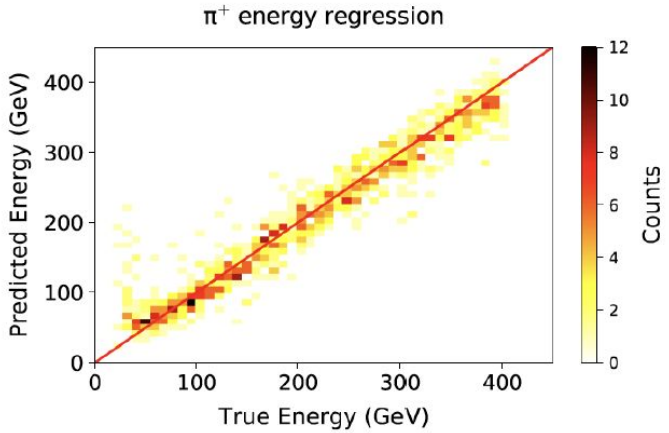


Supervised Learning

- The right answer is given in the training data (‘labeled’)

  E.g. Given the performance of the calorimeter, what is the true energy of a particle corresponding to a certain measured value.

  Regression: to make *continuous* estimation



Unsupervised Learning

- The right answer is **not** given in the training data (‘unlabeled’)

  E.g. Given the observation, divide the stars into different groups

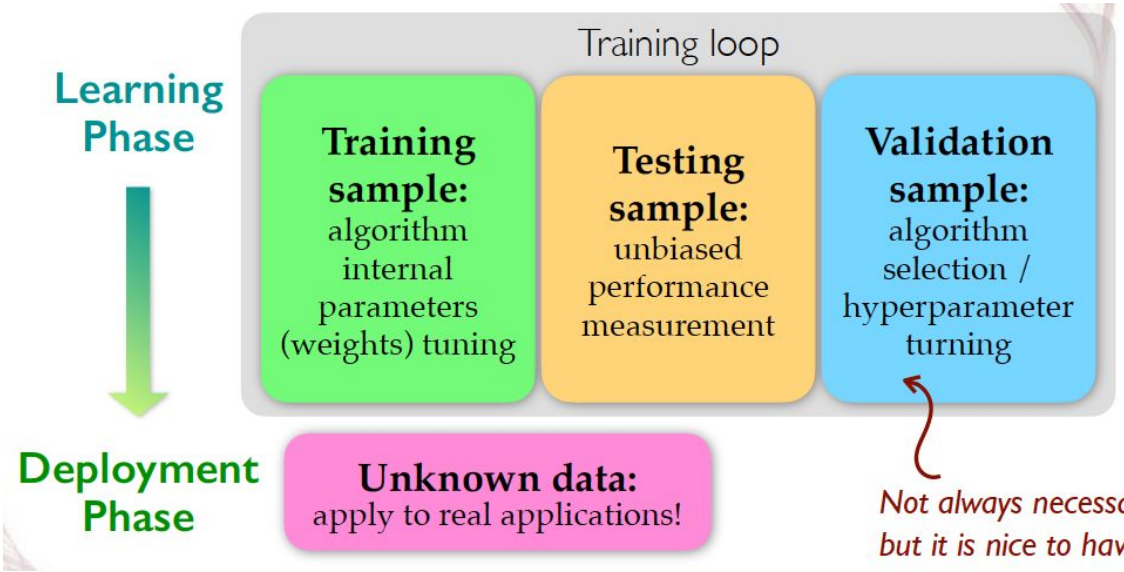
  Clustering: The groups are not known beforehand

Machine Learning

- Key ideas: identify a problem you want to solve, and learn from data
  - **Domain knowledge**: Understand what the problem is, and what information are needed to solve the problem
  - In this course we assume you are ‘domain experts’ in physics! – Be able to learn from google for jargons in physics

How to learn from data

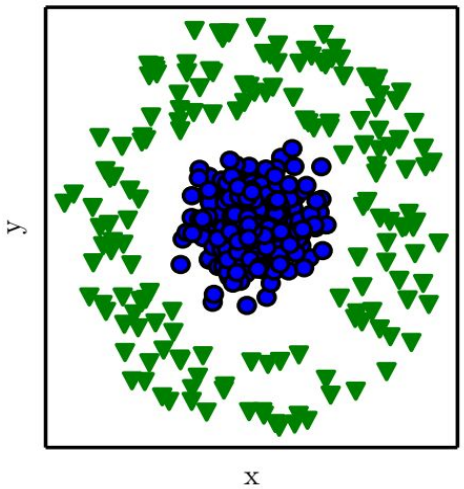
- We need **independent** data samples for training -> testing (-> validation) -> deployment



Courtesy of Prof. Kai-Feng Chen (NTU)

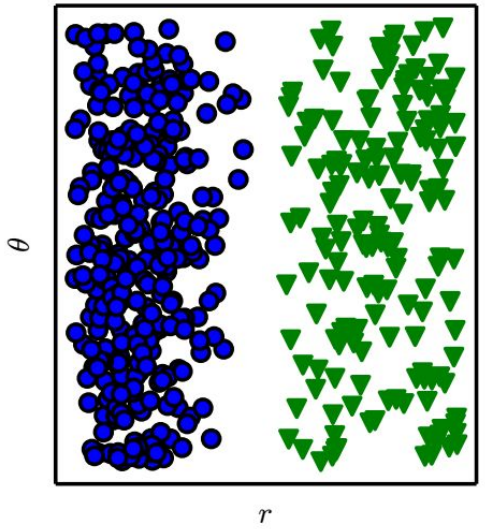
Data structure and visualization

- First need to know what the data looks like
  - **Data structure**: e.g. 1D or 2D arrays (‘shape’ of numpy arrays), label/meaning of each row/column (axis)
    - E.g. positions (x and y) of 100 events



Data structure and visualization

- Find other ways to represent (**visualize**) the data: usually make it easier to separate (classify) or group (cluster) them
  - Compute $r = \sqrt{x^2 + y^2}$ and plot in terms of r and θ



Data structure and visualization

- Information contained in the representation is known as a **feature** of the data

- The data usually come with ‘low-level’ features:
  - E.g. (E, p_x, p_y, p_z) of two particles from a decay

- Compute ‘high-level’ features based on domain knowledge
  - E.g. compute the invariant mass of the mother particle using $m = \sqrt{(E_1+E_2)^2 - (\mathbf{p}_1+\mathbf{p}_2)^2}$ (speed of light c=1)

In-class exercise: week 01

- Let’s remind ourselves with basic python knowledge for checking data structure and visualize the data

- We’ll use the famous MNIST (Modified National Institute of Standards and Technology) database: a database of handwritten digits that is commonly used for training various image processing systems with machine learning.

In-class exercise: week 01

- MNIST data:

MNIST dataset

- Each image has 28x28=784 pixels.

- A number between 0-255 is associated with a pixel, which corresponds to the gray scale:
  - 0=white, 255=black in the original MNIST dataset
  - Intuitively, the amount of ‘ink’ located on each pixel

Working with Kaggle

- For in-class exercise and submission of HW and final project.
- To begin, follow week01 in-class exercise

Computational Platforms

Python

TensorFlow

Jupyter
Notebook

SciKit-Learn

Feature study tools - NumPy and Pandas

https://numpy.org/

- NumPy is the fundamental package for scientific computing in Python. The core is the *ndarray* object.
- Numerical data, n-dimension, less memory consumption, better performance for 50K rows or less

https://pandas.pydata.org/

- Pandas is built on top of NumPy to manipulate tabular data, such as data stored in spreadsheets or databases, called a *DataFrame*.
- Tabular data, upto 3-dimension, more memory consumption, better performance for 500K rows or higher