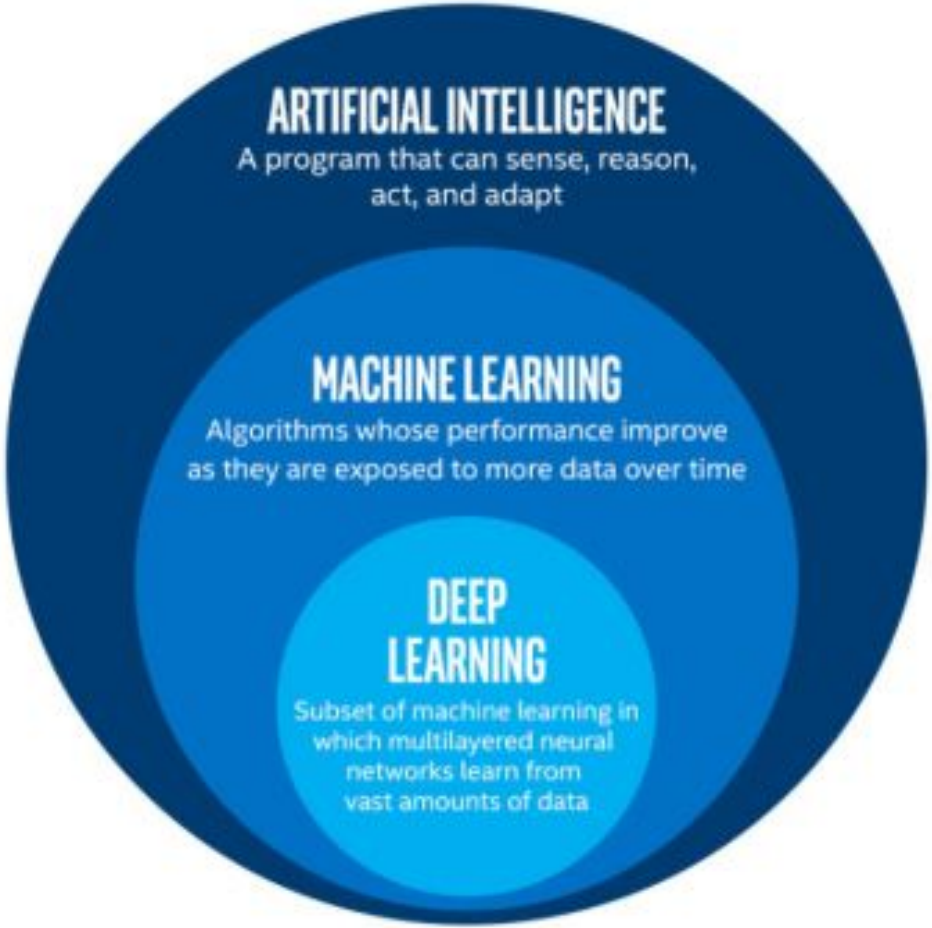
Machine Learning and Data Representation

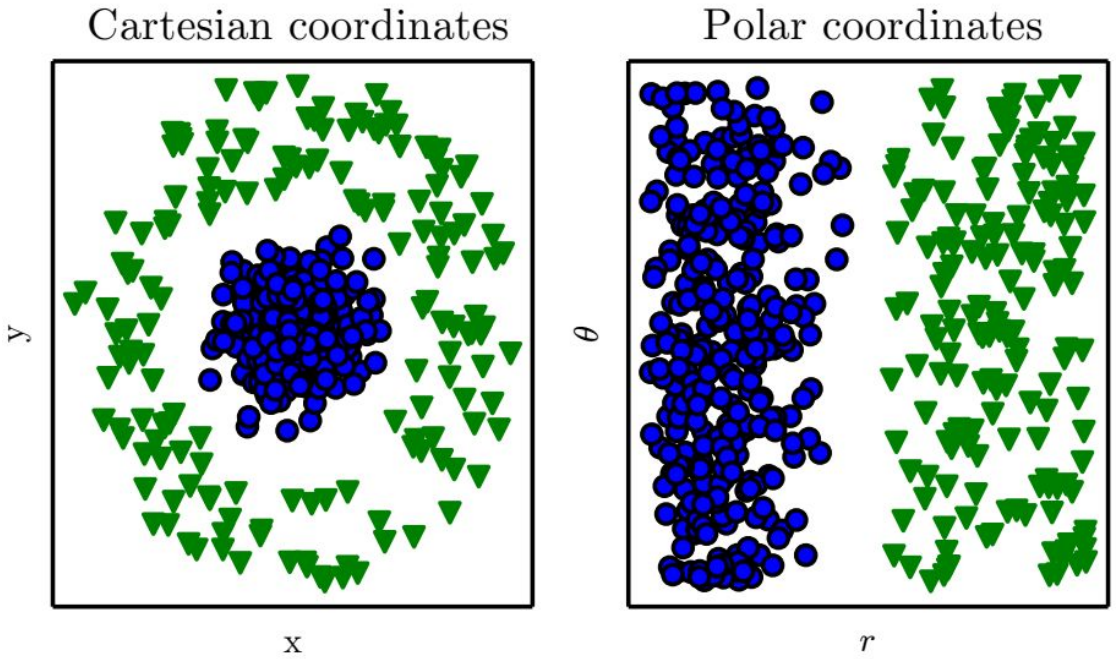
Week 02
PHYS591000 Spring 2021

Inspired by “Deep Learning” Ian Goodfellow, Yoshua Bengio, Aaron Courville



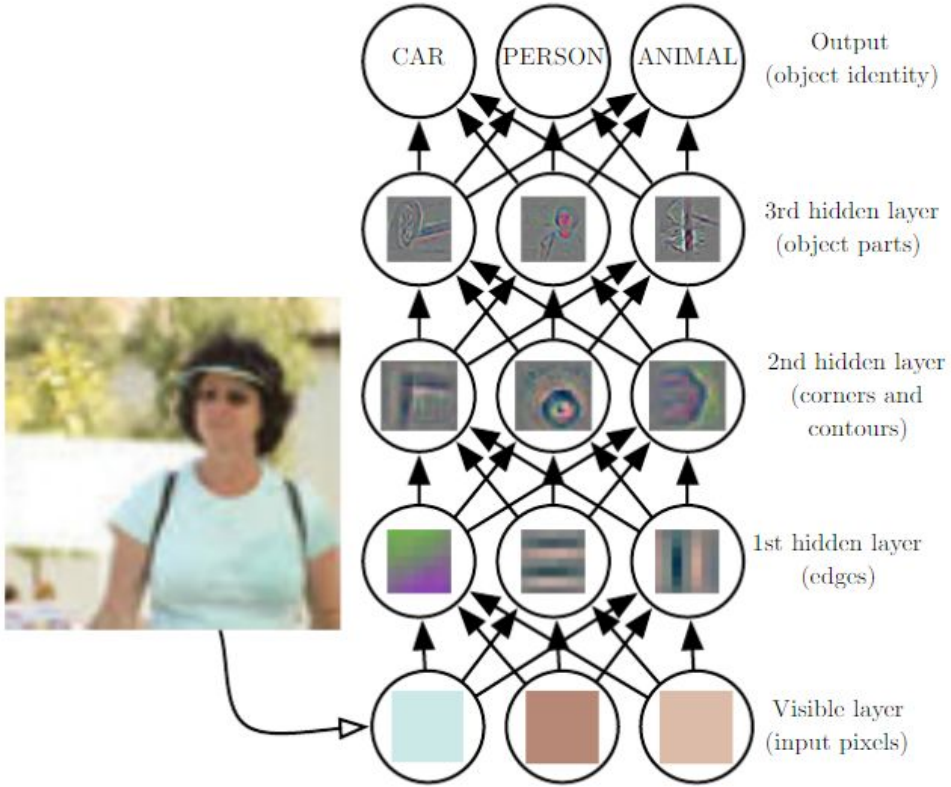
techplusmedia

Representation Matters

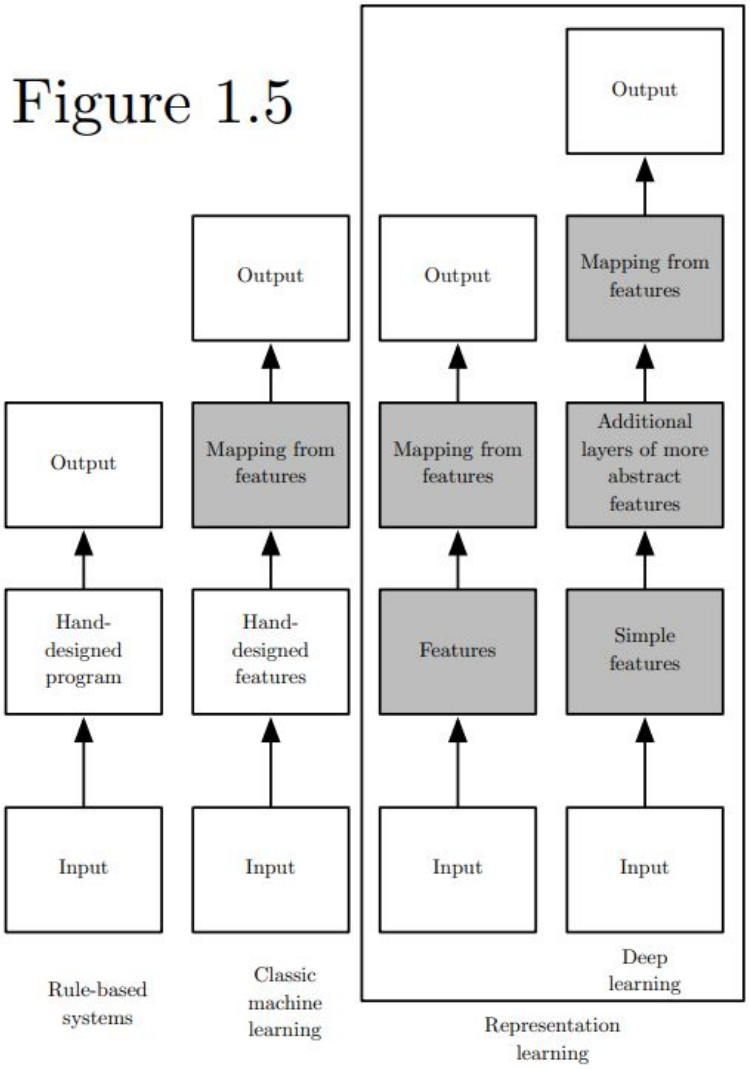


Each piece of information included in the representation of data is known as a *feature*

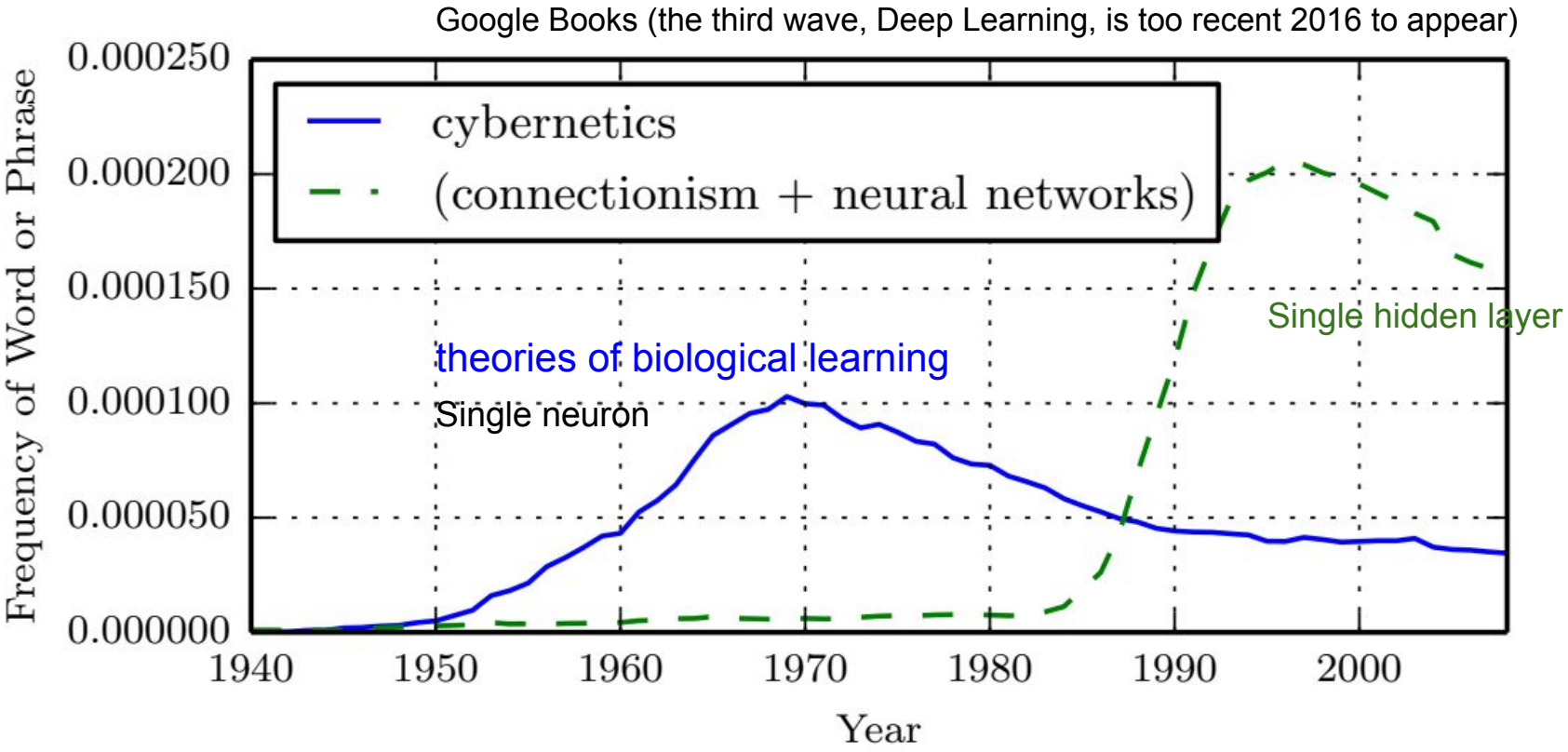
Illustration of Deep Learning



Learning multiple component



Historical Waves



Work in Machine Learning

- Applied Machine Learning
  - Collect data, build models, train models, analyze performance, evaluate errors
- Machine Learning developers
  - Implement Machine Learning algorithms and infrastructure
- Machine Learning research
  - Design and analyze models and algorithms

In this course, we will focus on “Applied Machine Learning”

Questions relevant in machine learning

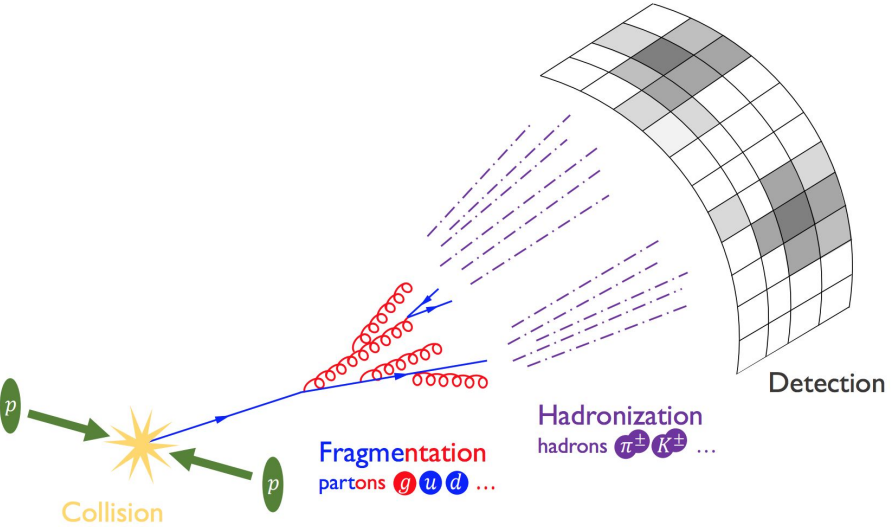
- What is the core problems?
- What features are available?
- Is there enough information to learn relationship between input and output?
- What are model assumptions?

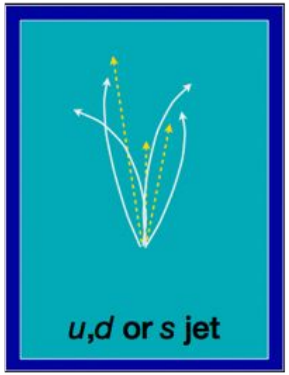
Feature Engineering

- Filtering
  - Multiple filters, keep only events of interest
  - Example: “events with one electron or muon with pT > 23 GeV”
- Prepare “low level features”
  - Every event is associated to a matrix of particles and features
  - Example: Feature = [“energy1”, “px1”, “py1”, “pz1”, “energy2”,”px2”,”py2”,”pz2”]
- Compute “High Level features”
  - Computed from low level particle features based on domain-specific knowledge
  - Example: M_{event} = invariant mass of electron-muon pair

How to solve problems without Machine Learning?

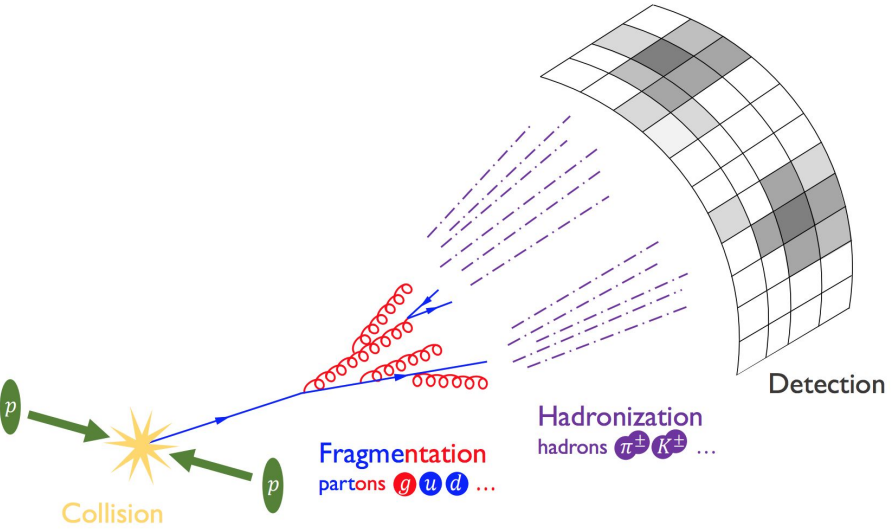
- Recruit a “physics expert” to teach you about different physics features (e.g. mass peak, color, angular distribution
- Recognize these features in a given distribution, and then come up with a best guess of signal process

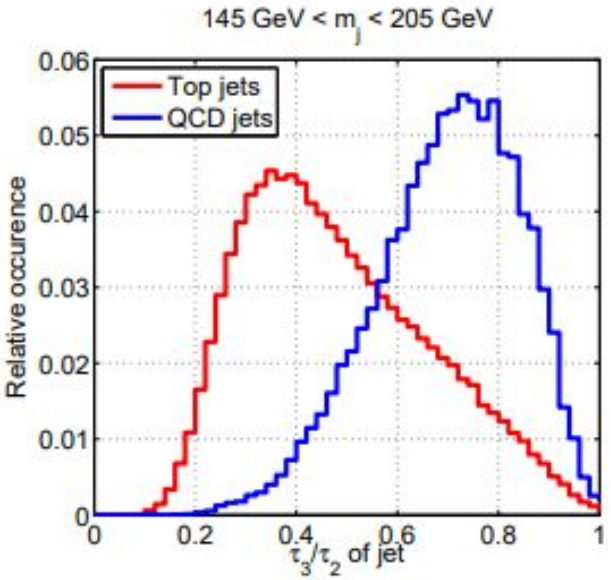




High Level Features - domain experts

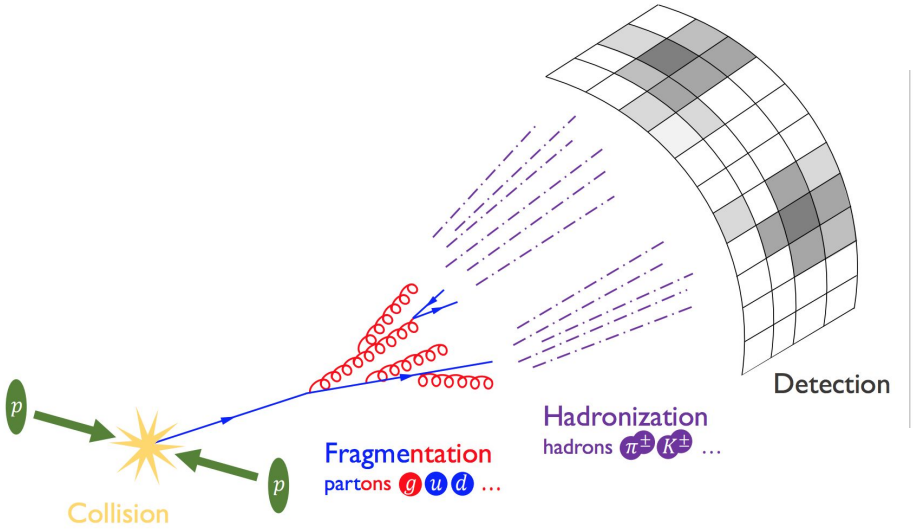
- Recruit a “physics expert” to teach you about different physics features (e.g. mass peak, color, angular distribution
- Recognize these features in a given distribution, and then come up with a best guess of signal process

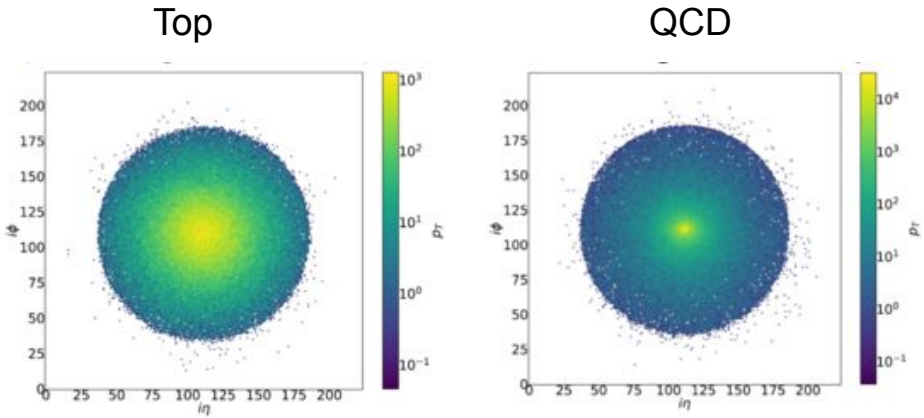




Low Level Features using Deep Learning

- Use simple representatation to treat data as pixelated image





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.

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*.

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

Week 2 class discussion

In-class Quiz and Lab session: https://www.kaggle.com/t/f9f694670e674ee2af94cd8ff506ed96

In-class Whiteboard [URL]

Seat Assignment		
13	14	15
10	11	12
7	8	9
4	5	6
1	2	3
	White board	

No.	StudentA	StudentB
1	106022116 江瑞軒	109022421 陳星宏
2	106022119 沈昀臻	109022502 許為吉
3	106022201 陳立豪	109022504 陳品翰
4	106022215 謝明儒	109022512 翁昊丞
5	106022216 蕭科洋	109022515 林慧琪
6	106022219 鄔玟潔	109022516 陳沅彰
7	106022231 吳芯瑀	109022521 岑美茵
8	107022101 胡英祈	109022531 黃佳敏
9	107022107 沈奕廷	109022533 王榆惠
10	107022108 陳正力	109022554 葉百祥
11	107022120 王 晨	109022701 謝沛軒
12	107022135 周晏霆	105035423 孫睿狄
13	107022207 林宗彣	107022523 郭俊賢
14	108022516 吳宣霈	108022901 張宸瑜
15	108022550 張藝薰	108022553 何銘鴻

Select correct statements

- [ ] Deep Learning is a subset of Machine Learning
- [ ] Artificial Intelligence is a subset of Machine Learning
- [ ] Deep Learning input is usually a simpler representation
- [ ] Machine Learning heavily relies on the representation

Select correct statements

Checkboxes

- [x] Deep Learning is a subset of Machine Learning
- [ ] Artificial Intelligence is a subset of Machine Learning
- [x] Deep Learning input is usually a simpler representation
- [x] Machine Learning heavily relies on the representation

Select correct statements

- [ ] There are two historical waves of the AI development
- [ ] We are currently in the second wave of the AI development
- [ ] The capability of solving hidden layer neuron is the achievement in the 1st wave
- [ ] None of the above

Select correct statements

Checkboxes

- [ ] There are two historical waves of the AI development
- [ ] We are currently in the second wave of the AI development
- [ ] The capability of solving hidden layer neuron is the achievement in the 1st wave
- [ ] None of the above

Select correct statements

- [ ] Deep Learning uses low level features
- [ ] Machine Learning uses high level features
- [ ] high level feature is built from low level features
- [ ] domain expert knowledge is required to do analysis without machine learning

Select correct statements

Checkboxes

- [x] Deep Learning uses low level features
- [x] Machine Learning uses high level features
- [x] high level feature is built from low level features
- [x] domain expert knowledge is required to do analysis without machine learning

Select correct statements

- [ ] NumPy is built on top of Pandas
- [ ] NumPy is more efficient for 500K rows or more
- [ ] Pandas is more efficient for 50K rows or less
- [ ] Pandas is for tabular data

Select correct statements

Checkboxes

- [ ] NumPy is built on top of Pandas
- [ ] NumPy is more efficient for 500K rows or more
- [ ] Pandas is more efficient for 50K rows or less
- [ ] Pandas is for tabular data