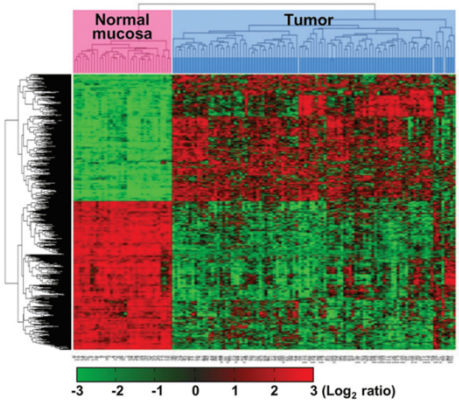
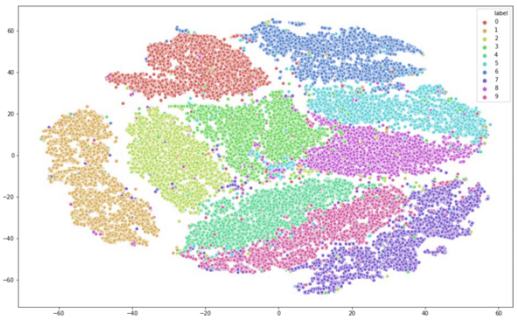
Lecture 20 Deep Learning

- Recap from last lecture
  - What is meant by finding the RNA motif
    - ✓ False: the RNA binding site
    - ✓ True: Binding pattern
  - The gene expression matrix
    - ✓ Biggest difference between single & bulk: Specificity
  - Challenges in single-cell data analytics (& corresponding solutions)
    - ✓ Noise (denoise, normalization)
    - ✓ Doublet (KNN etc.)
    - ✓ Dropout (treat the missing values, e.g take average)
    - ✓ Batch Effect (normalization, advanced technologies)
  - Visualize gene expression data in 2D





    - ✓ In huge dimension, we can't see whether they are close to each other, so we realize the data in 2D to find out ant similarities/differences
  - The process of t-SNE
    - ✓ Time consuming but powerful
  - Protein binding has preference (very specific)
  - From aligned sequences to motif
    - ✓ Realize motif

- In this module, we focus on personal health, which is a more daily life topic, we will apply the models we learned in Module 2 Genomics to handle these kinds of data.

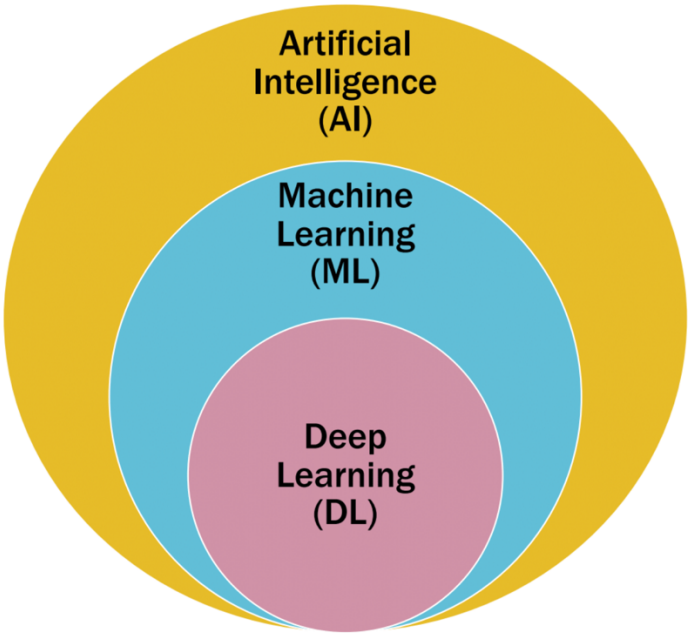
- Why do we care about health data?
  - Doctors/clinicians: diagnosis, treatment
    - ✓ They can only diagnosis from the data/information (symptoms & lab test)
    - ✓ For example, we apply a machine to take a photo of your body (maybe a specific part), and then we realize it into visualization information
    - ✓ 望聞問切 (Biomedical imaging, symptom, disease history, lab tests)
    - ✓ Without data, doctors cannot diagnose precisely

- For example, if one was bite by the snake, you need to figure out its species to decide is it poisonous/non-poisonous? If it is poisonous, what is the type of the poison (influence on blood/neural cells etc.)

➢ AI + Health data: Diagnosis & treatment

- Deep learning has been applied in disease screening

- Artificial Intelligence VS Machine Learning VS Deep Learning



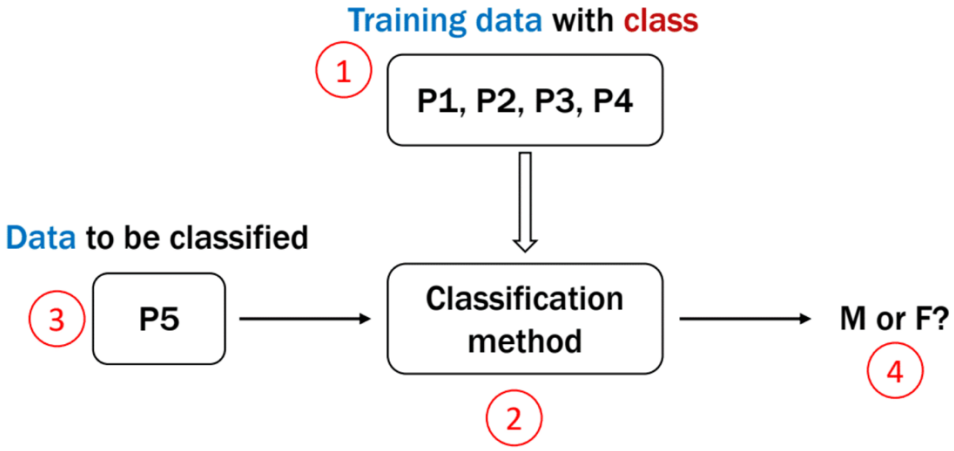
➢ AI: Any techniques which enable computers to mimic human behavior
  ✓ Very huge scope
➢ ML: A subset of AI, which effectively perform a specific task without using explicit instructions, relying on patterns and inference from the data (even without instruction)
➢ Any example of it is AI but not ML
  ✓ For example, traditional car(driver will give a clear instruction) and autonomous car (Based on environment to perform specific operations)
➢ DL: A subset algorithms of ML, which takes advantage of multi-layer neural networks.
  ✓ Give instruction VS Based on environment

- There are 309 keywords in AAAI, machine learning is just a specific direction

- Machine learning tasks
  ➢ Unsupervised leaning (Without predefined label): Clustering, Dimension Reduction
  ➢ Supervised learning (Has predefined label): Regression, Classification
  ➢ Reinforcement learning: system interact with the environment, learning adjust itself

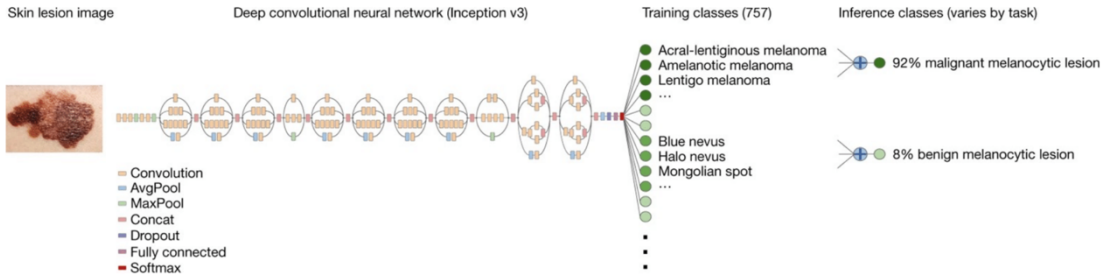
- Deep Learning (just for reference)
  - Large language models (Human/Protein/RNA)

- Deep learning for disease screening
  - Simplified model we learned in Module 2



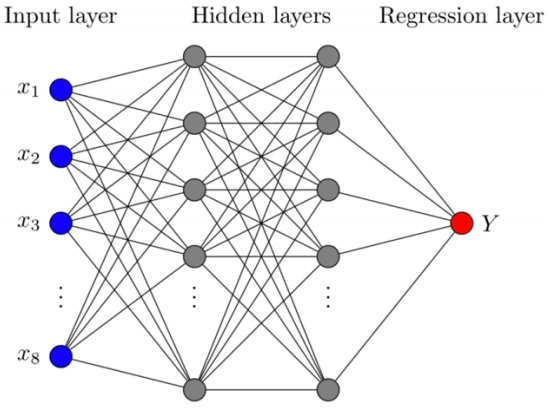
  - Real life example: Patient's data→Cancer? →If yes, which stage?

- Build models for real-life healthcare problems

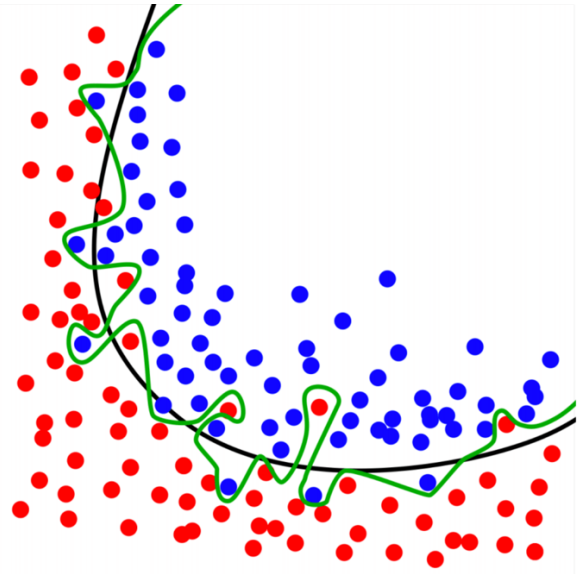


  - Complicated

- We have introduced fully connected networks
  - To resolve complicated problems
    - ✓ Increase the number of nodes and layers
    - ✓ Add non-linear function



- Problems?
  - ✓ Suppose we have an image with size (256*256*3), and 3 layers for binary classification. The internal layer has 1000 nodes. How many parameters will we have?
    - (256*256*3+1)*1000 + (1000 + 1)*1 = 196610001, which is super complex
  - ✓ Overfitting issue: the model is too complicated that it may fit the noise in the data



  - ✓ Other Problems besides overfitting:
    - How to determine the number of nodes and layers?
    - Storage
    - Running time (Embedded systems)
    - Hard to train
    - Prior knowledge is ignored (Images do have knowledge/spatial information, for example, 2 pixels near each other will have similar pixel value)

- Images are different from data matrix
  - Because after shuffling, the spatial information will change significantly in image