AI + Bio

Lan XUE

2021.4.10

NJU-AIA

Overview

- I . The methodology and technology in biological research
- II. A Primer on Machine Learning
- III. Bio+AI: two demos

Overview

- I . The methodology and technology in biological research
- II. A Primer on Machine Learning
- III. Bio+AI: two demos

We now better understand how species are related

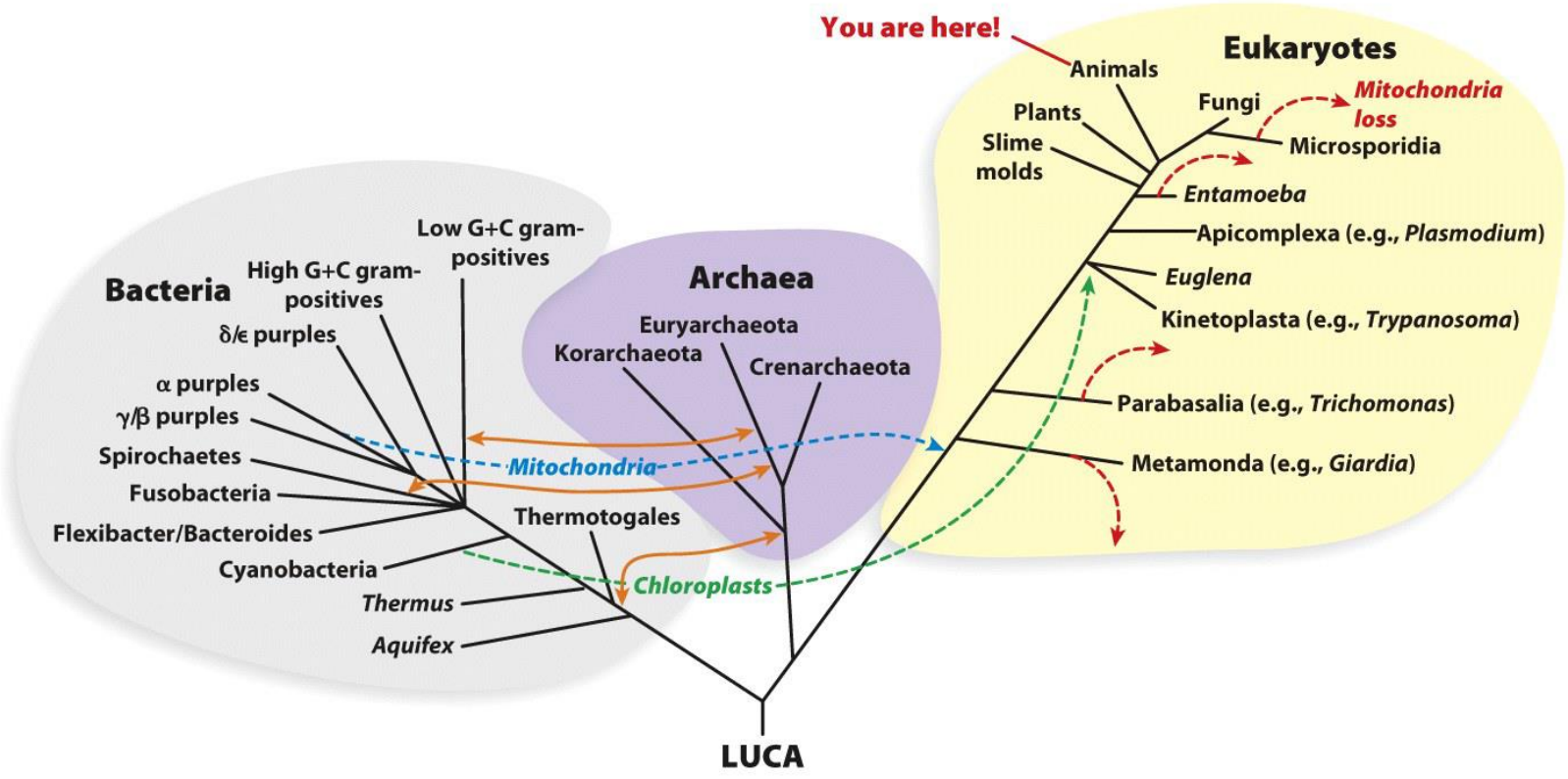
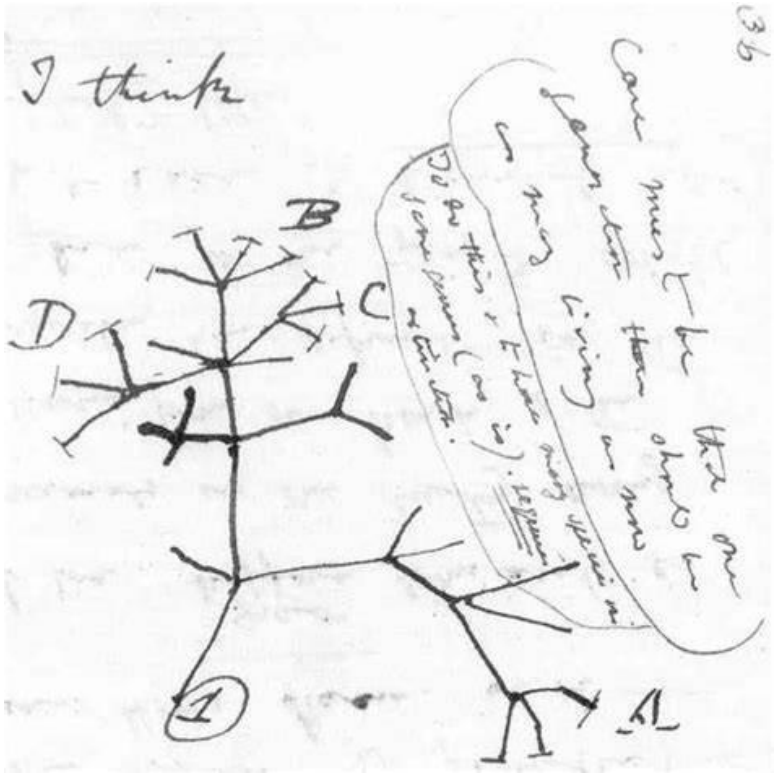


Figure 8-17a
Molecular Biology: Principles and Practice
© 2012 W. H. Freeman and Company



Trees depicting evolution.

The first known **evolutionary tree** drawn by Darwin in 1837 (right)

Hypothesis-Driven Research

Hypothesis

Experiment design

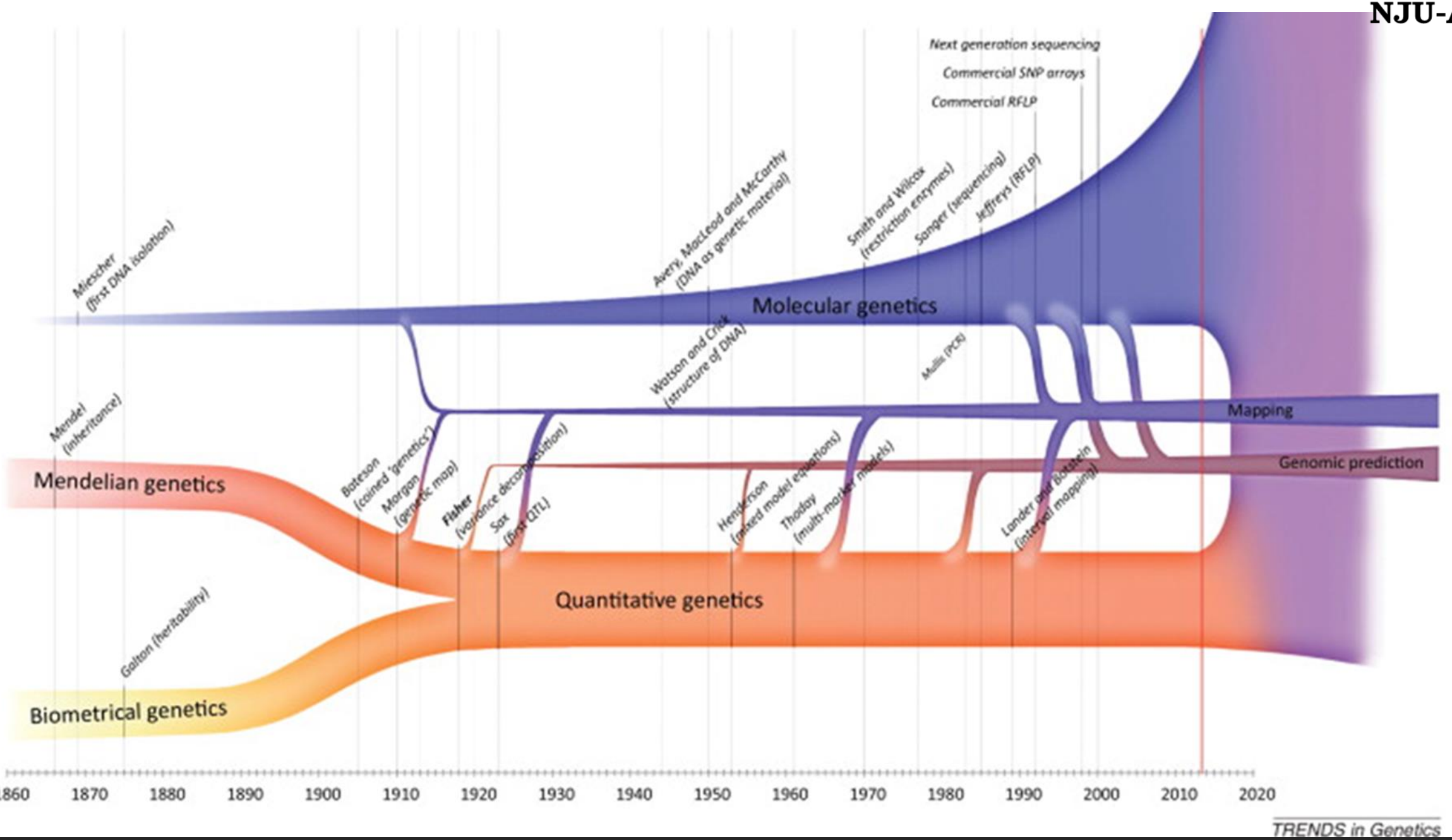
Proven

New theory

Falsity

New hypothesis

© Long W


Next generation sequencing
Commercial SNP arrays
Commercial RFLP
Miescher (first DNA isolation)
Avery, MacLeod and McCarthy (DNA as genetic material)
Smith and Wilcox (restriction enzymes)
Sanger (sequencing)
Jeffreys (RFLP)
Molecular genetics
Watson and Crick (structure of DNA)
Mullis (PCR)
Mapping
Mendel (inheritance)
Genomic prediction
Mendelian genetics
Bateson (coined 'genetics')
Morgan (genetic map)
Fisher (variance decomposition)
Sax (first QTL)
Henderson (mixed model equations)
Thoday (multi-marker models)
Lander and Botstein (interval mapping)
Quantitative genetics
Galton (heritability)
Biometrical genetics
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
TRENDS in Genetics

You can now have your genome fully sequenced for less than$999

myGenome is a whole genome sequencing and interpretation service to help you and your physician improve your health, longevity, and much more.

https://www.veritasgenetics.com/mygenome

BGI华大

首页 科学研究 科技服务 医学服务 仪器平台 华大概况 国家基因库 火眼实验室 集团站群 EN

临床检测

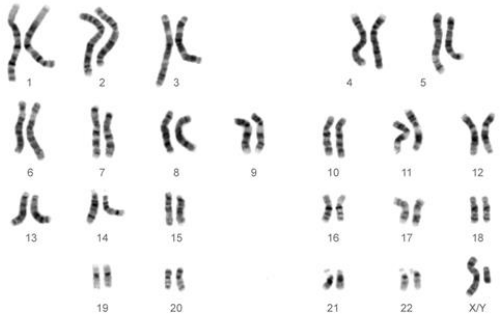
- 生育健康
  - 孕前检测
  - 产前检测
  - 新生儿检测
  - 单基因病检测
  - 女性健康检测
- 遗传咨询
  - 遗传咨询
- 肿瘤防治
  - 遗传性肿瘤基因检测
  - 肿瘤用药基因检测
  - 临床营养代谢检测
- 传染感染
  - PMseq®病原微生物高通量基因检测

体外诊断

- 检测服务
  - 肿瘤标志物检测
  - 营养干预及用药安全检测
  - 宫颈癌筛查
  - 优生优育检测
  - 内分泌检测
  - 生化类检测
- 试剂产品
  - 联合检测试剂
  - 质谱检测试剂
  - 化学发光检测试剂
  - 核酸检测试剂
  - 酶联免疫检测试剂
  - 胶体金检测试剂

鉴定服务

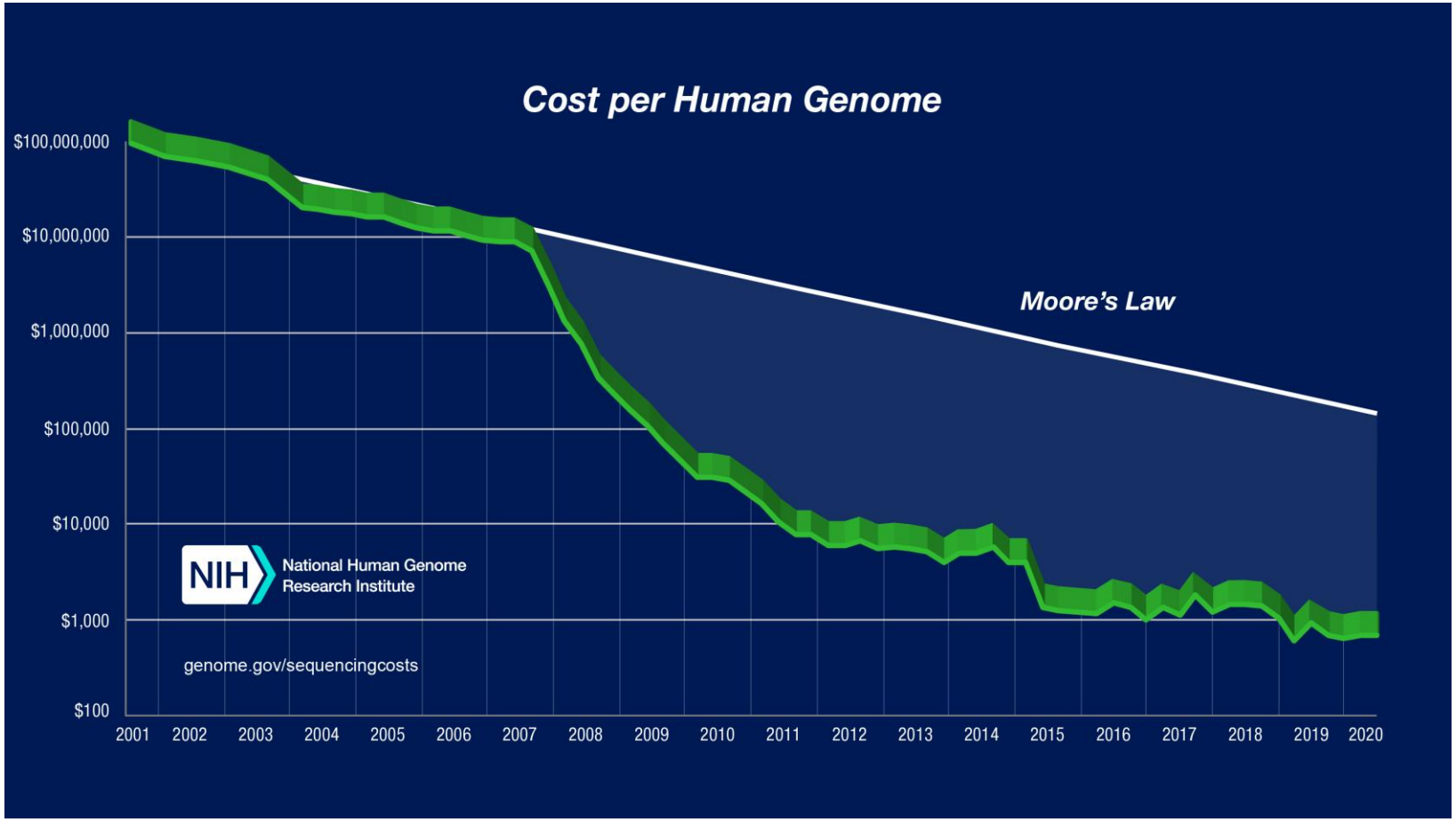
- 检测服务
  - DNA亲缘检测
  - 动植物检测
  - 毒物检测
- 试剂服务
  - 试剂盒



https://www.genomics.cn/

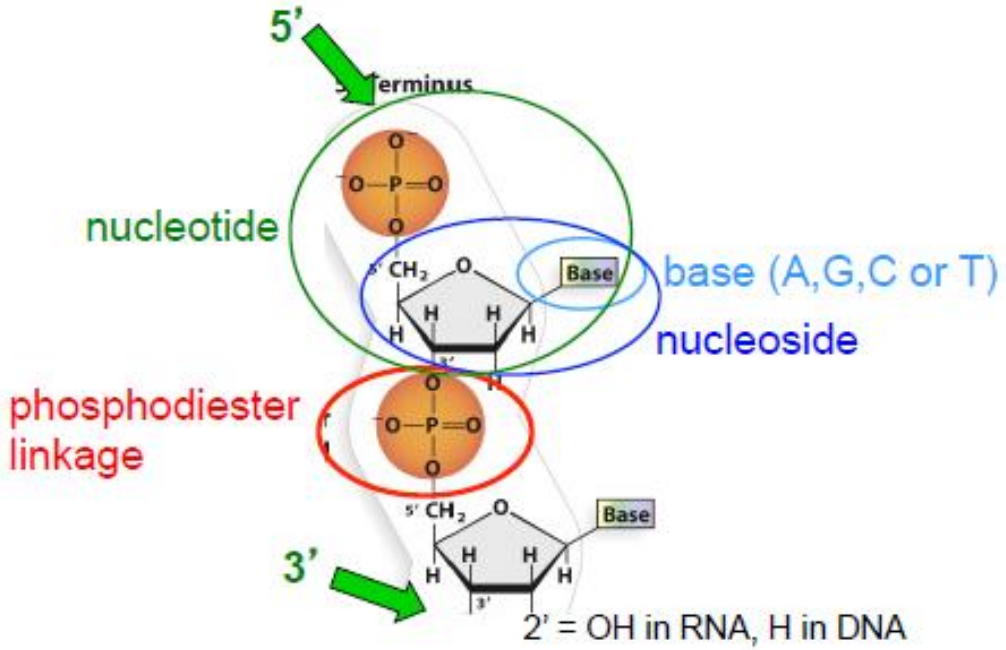
The cost of sequencing started decreasing precipitously in 2007

Why?



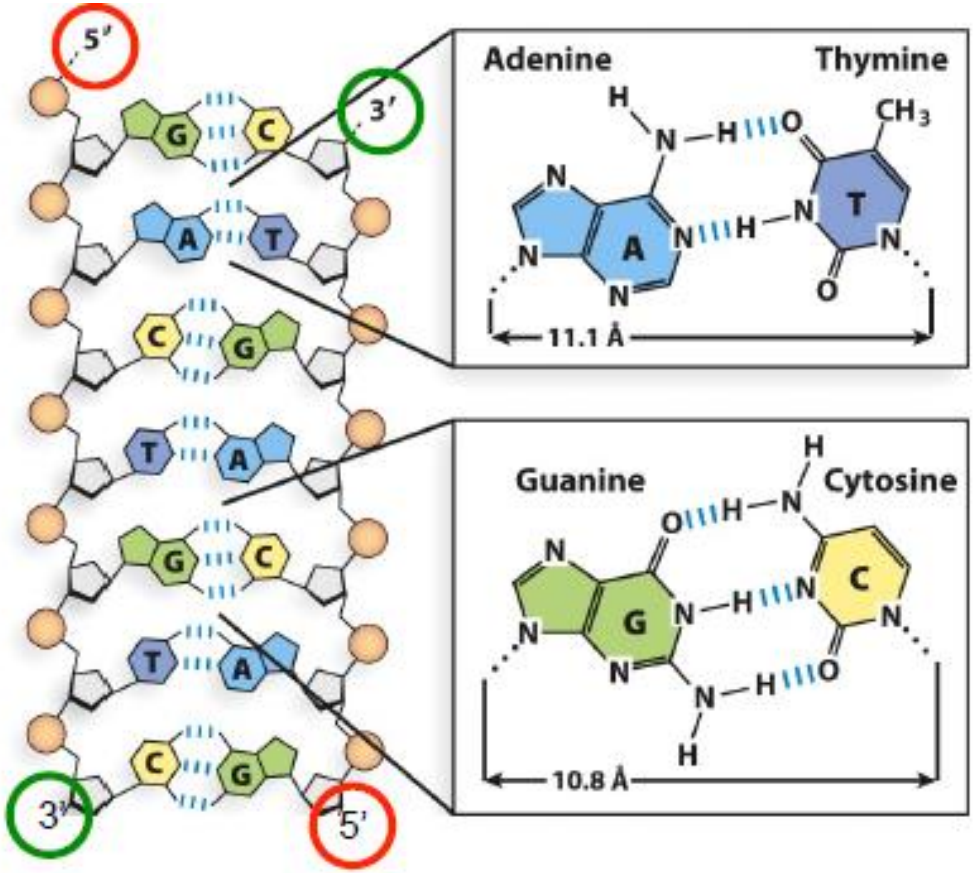
http://www.genome.gov/sequencingcosts/

DNA structure



Oligonucleotides are
connected
by phosphodiester bonds

Double stranded
DNA is antiparallel



Sanger sequencing (First generation sequencing)

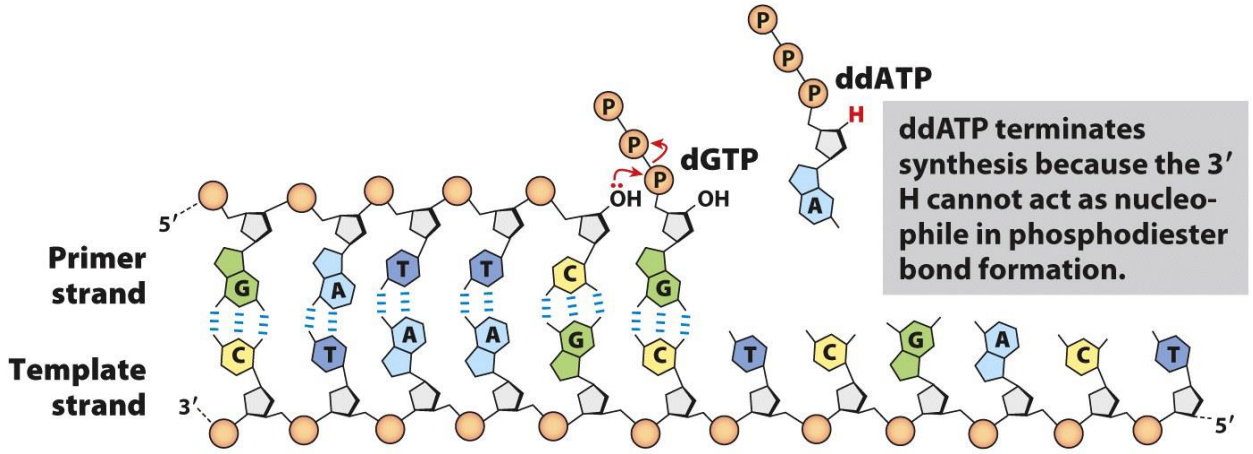
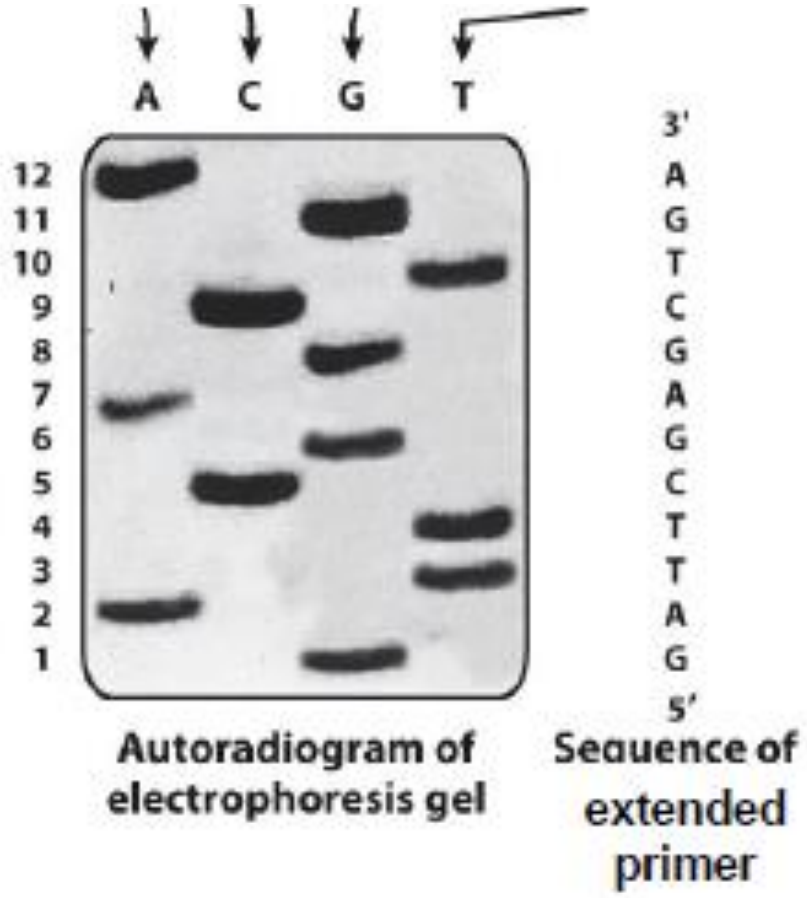


Figure 7-11a
Molecular Biology: Principles and Practice
© 2012 W. H. Freeman and Company



"Next generation" sequencing costs much less per base than Sanger sequencing because millions of sequences are read out at the same time

Illumina sequencing

[Source: (c) Courtesy Michael Cox]

Fluorescently labeled, reversible terminator nucleotides

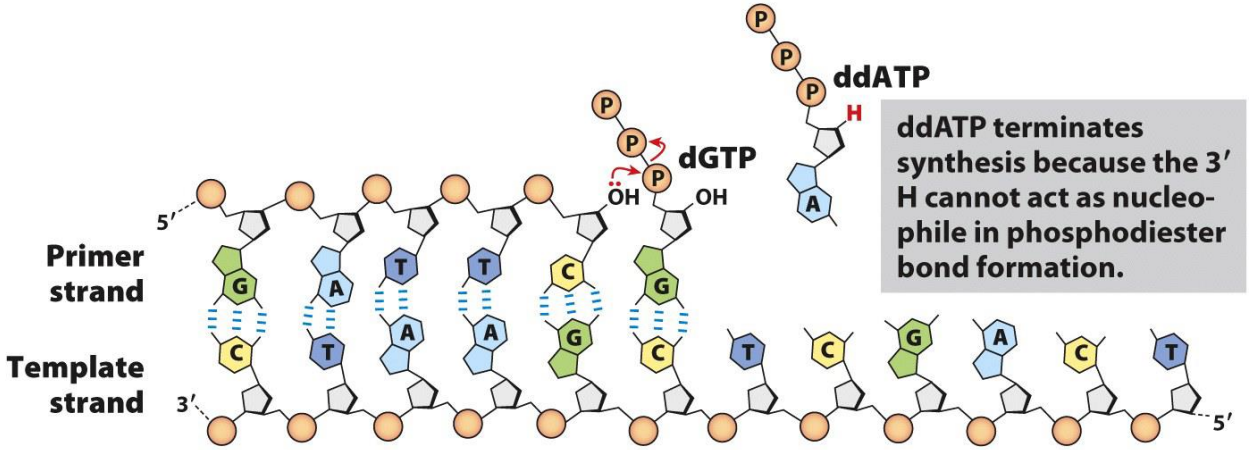
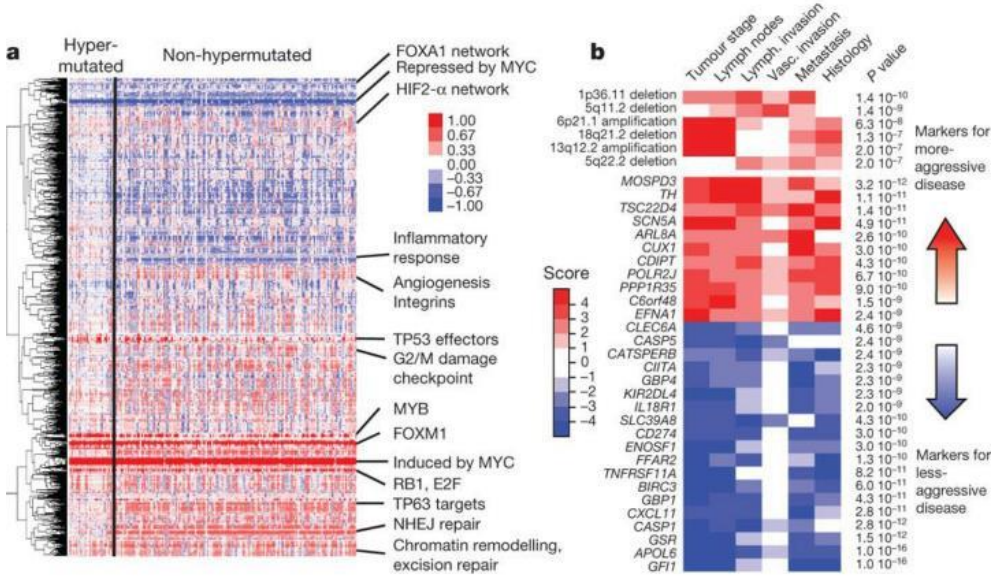


Figure 7-11a
Molecular Biology: Principles and Practice
© 2012 W. H. Freeman and Company

Dideoxy nucleotides (ddATP, ddGTP, etc.) are used as chain terminators in **Sanger sequencing**

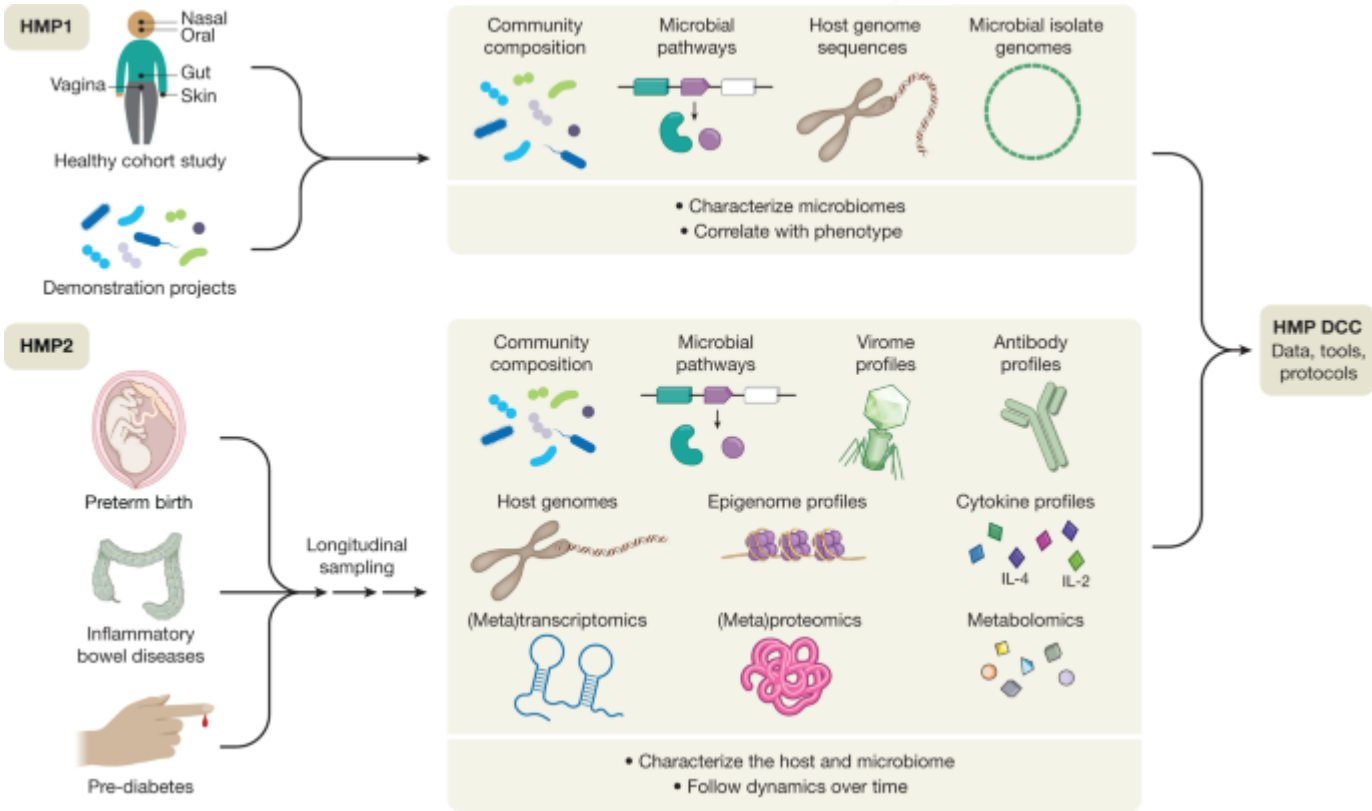
Community efforts across research disciplines are regularly generating **petabytes of data**.

Genomics: Cancer Genome Atlas



Cancer Genome Atlas Network. *Nature*. 2012;487:330-337.

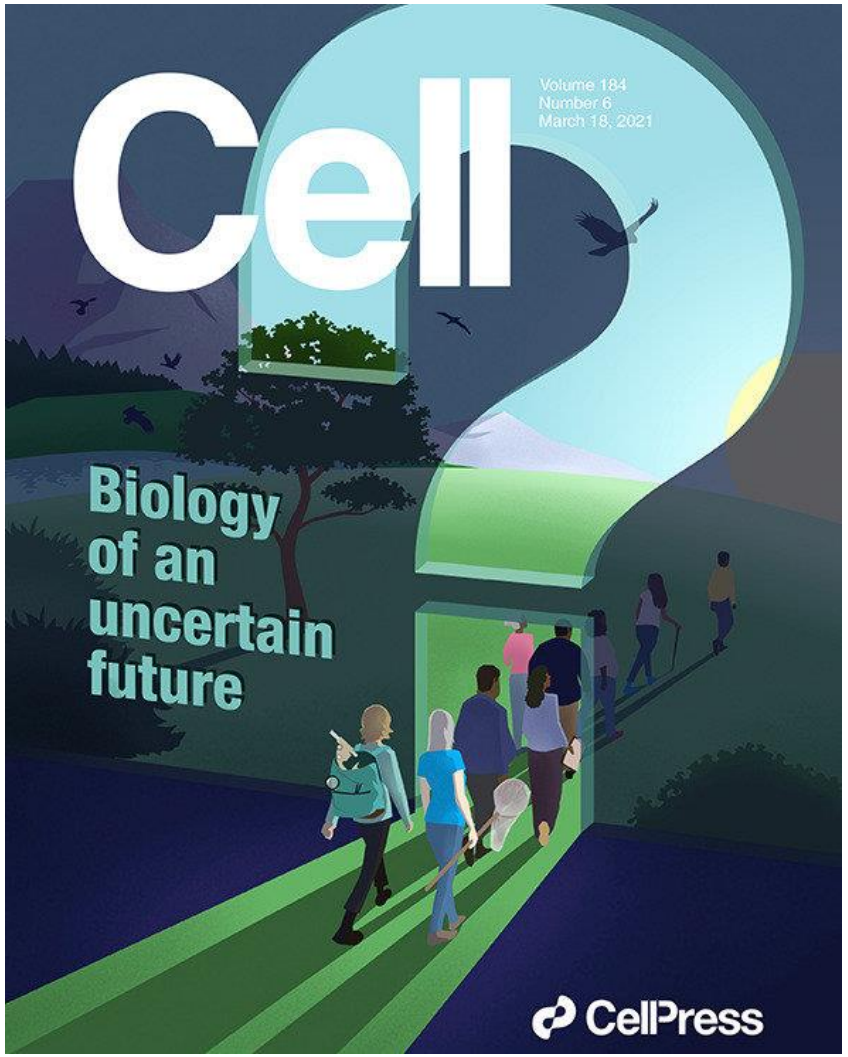
HUMAN MICROBIOME PROJECT



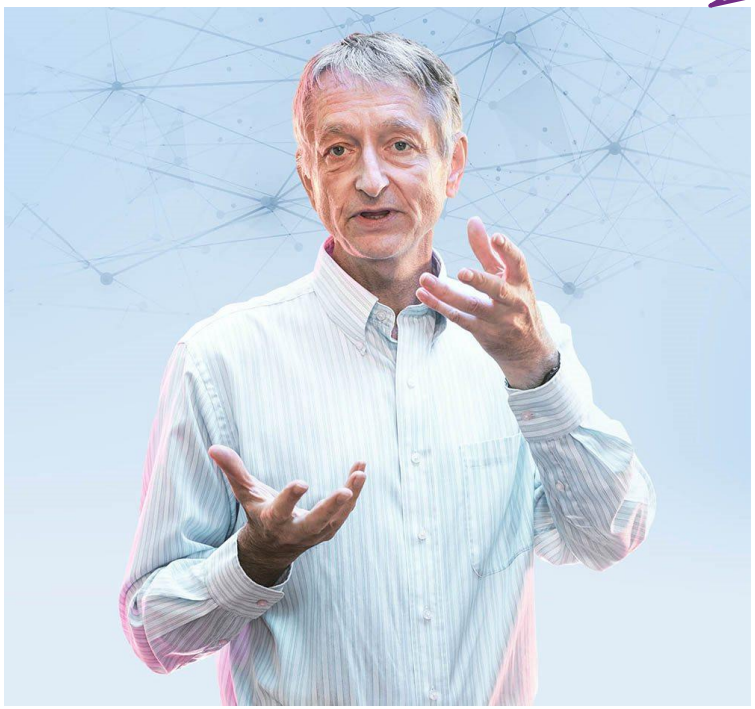
Biology of an uncertain future?

- The cost of sequencing started decreasing precipitously in 2007.
- Highly complex datasets has been being generated from biological experiments.
- While entering the second year of a pandemic., it is urgent to unlock basic science problems.

Can next generation machine learning be
a game changer?



We have about 10^14 synapses and only live for about 10^9 seconds. Experience are expensive, synapses are cheap.



© Geoffery E.Hinton

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Overview

- I . The methodology and technology in biological research
- II. A Primer on Machine Learning
- III. Bio+AI: two demos

What enables machine learning?

COMPUTING HARDWARE

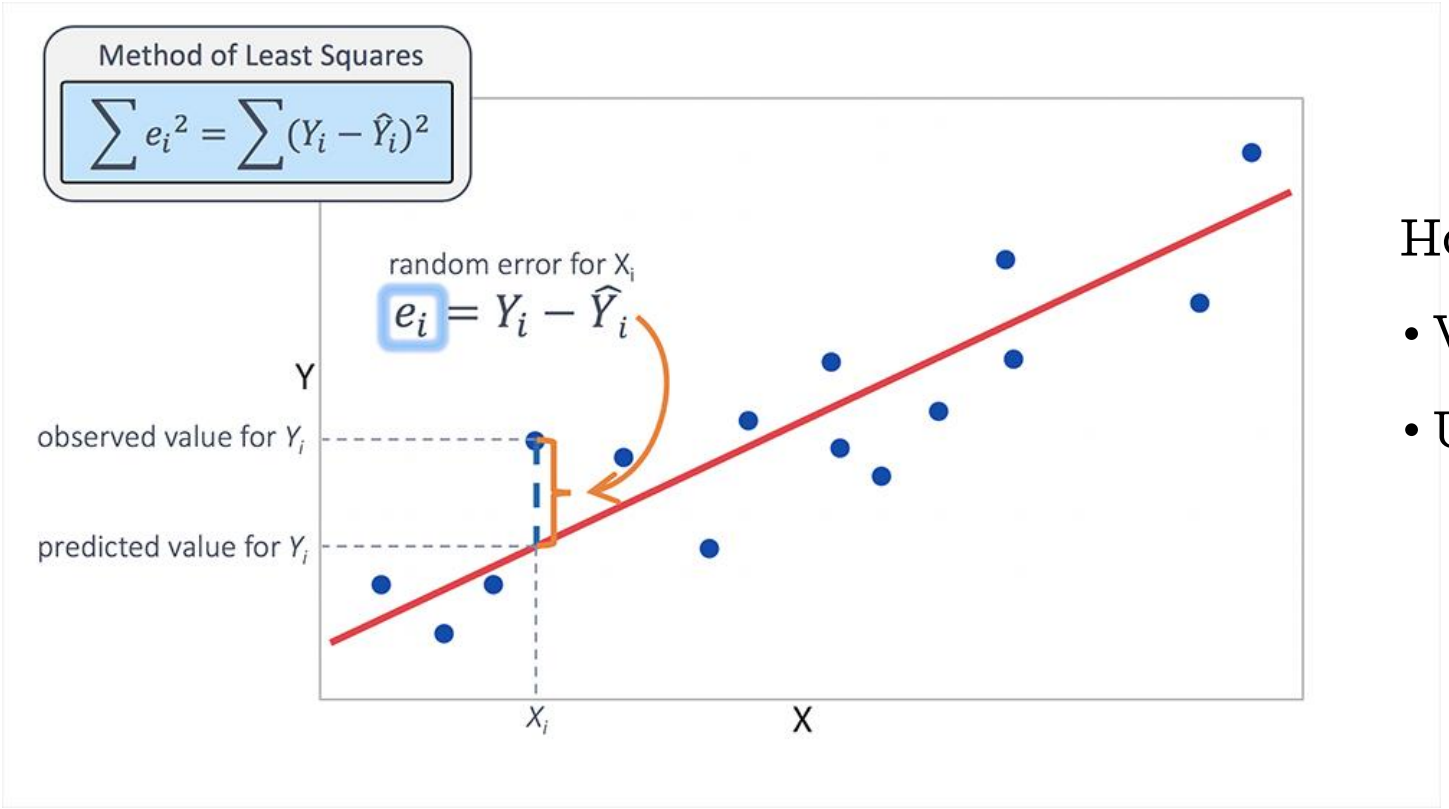
ALGORITHM

BIG DATA

DOMAIN

Regression and Classification

Regression

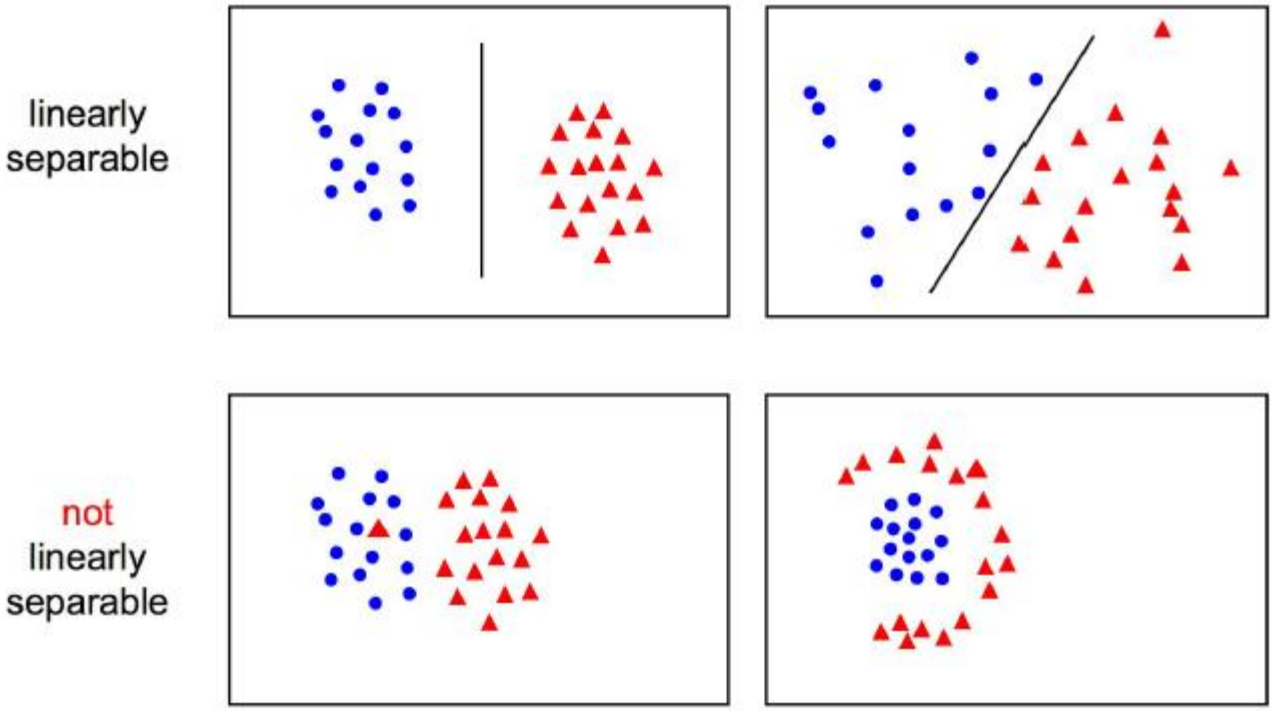


How to get this line:

- We have: Y = aX + b
- Use **Least Square Method** to get a and b

Regression and Classification

Classification



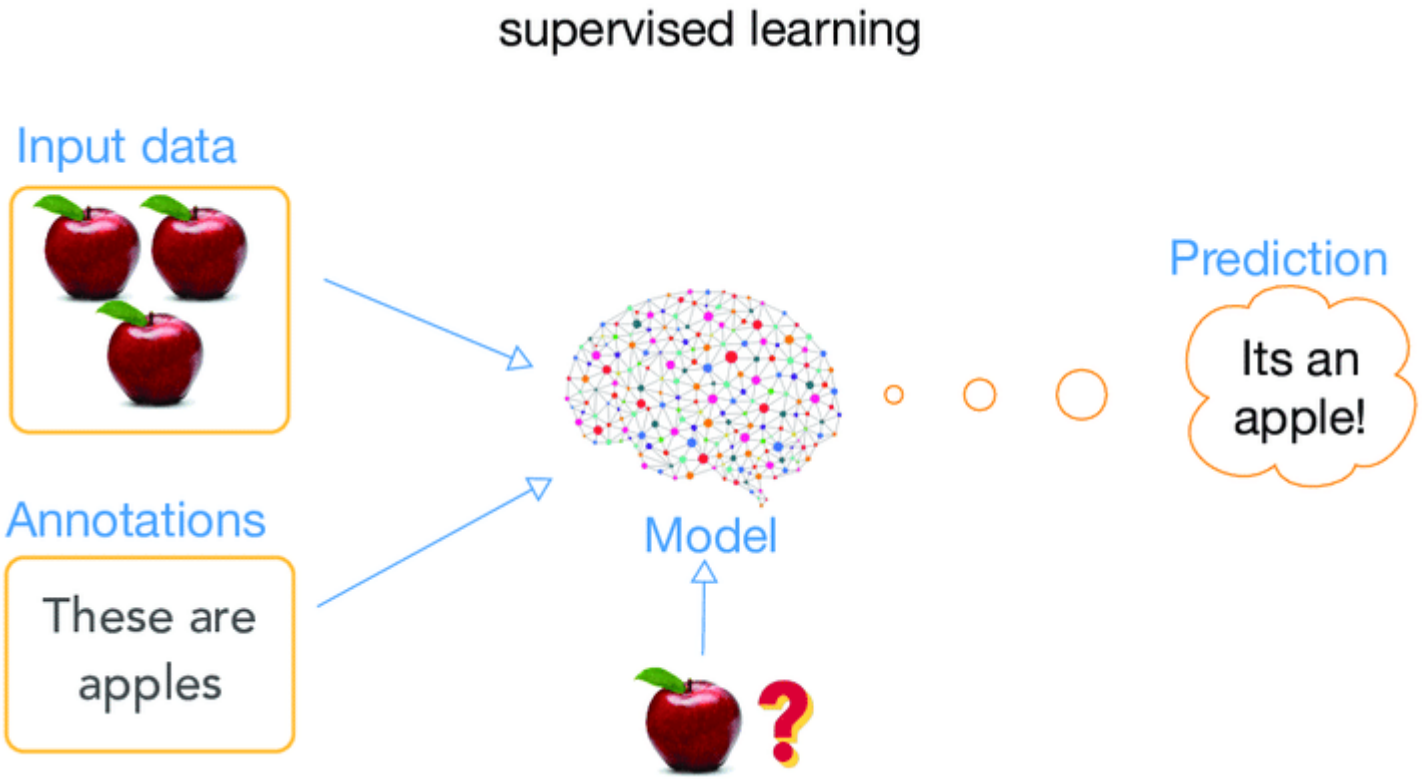
How to get this line:

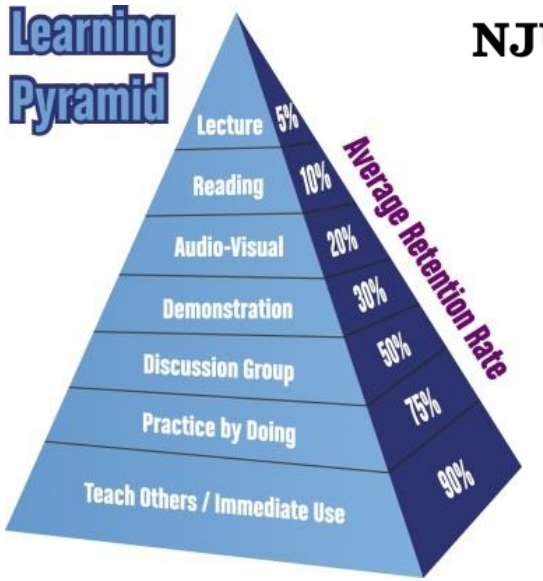
- We have: $Y = aX + b$
- Use **Linear Classification Method**

to get a and b

Supervised Learning

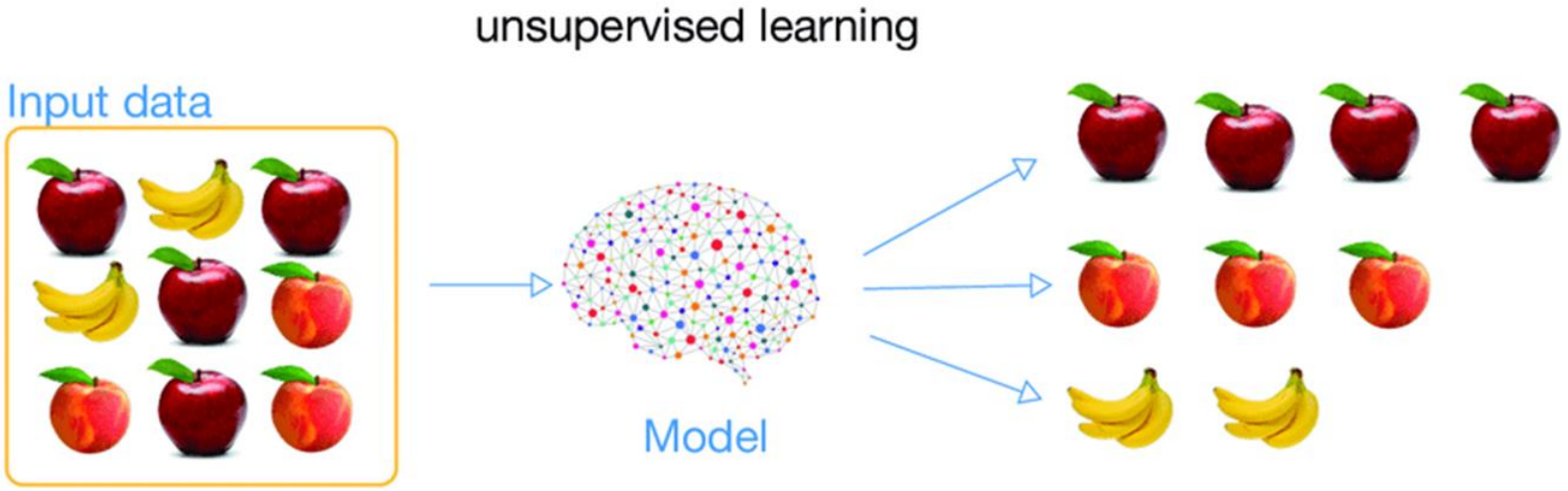
Two concepts: data, target/label





Unsupervised Learning

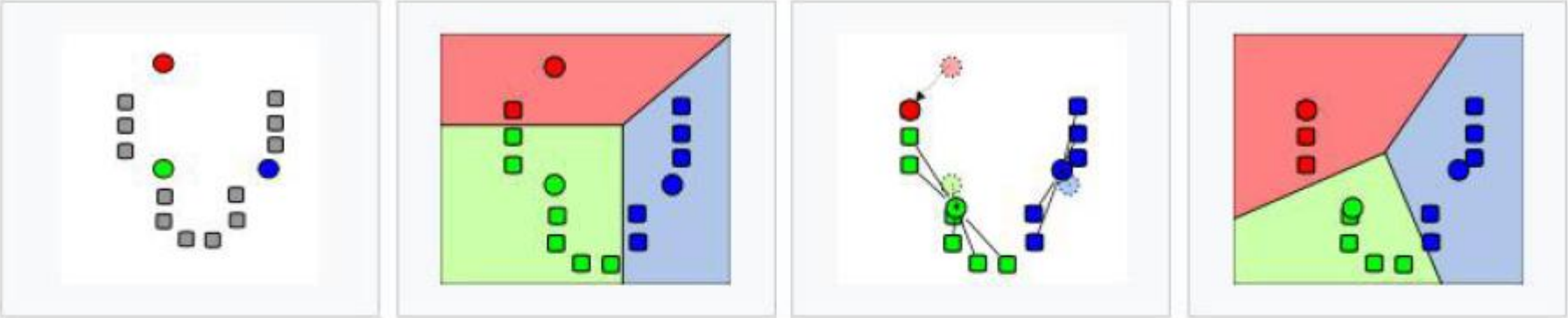
Only data



Unsupervised Learning

K means

Clustering



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

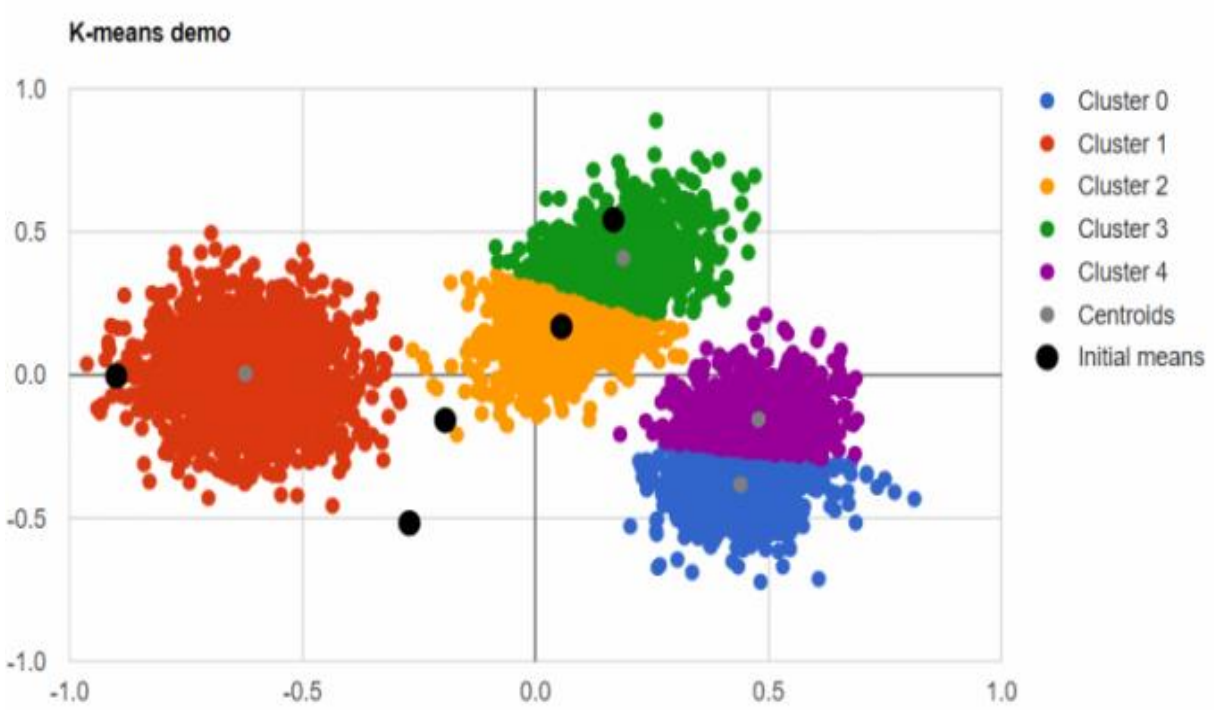
3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

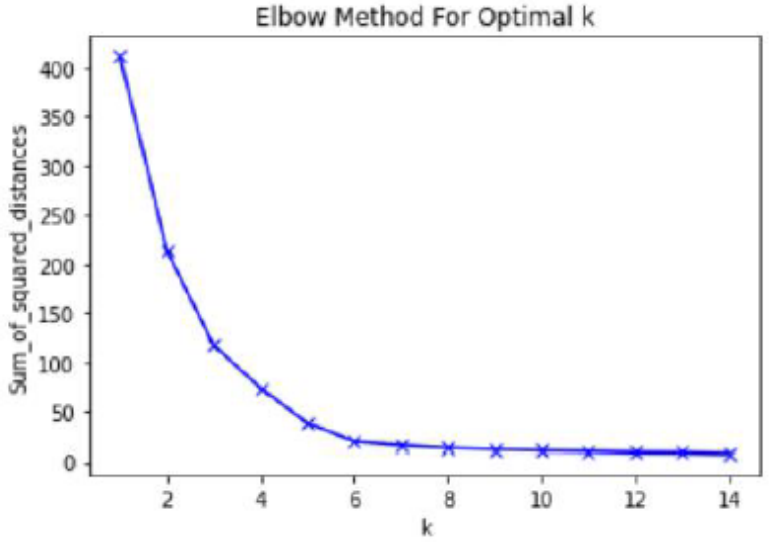
Unsupervised Learning

K means

Clustering

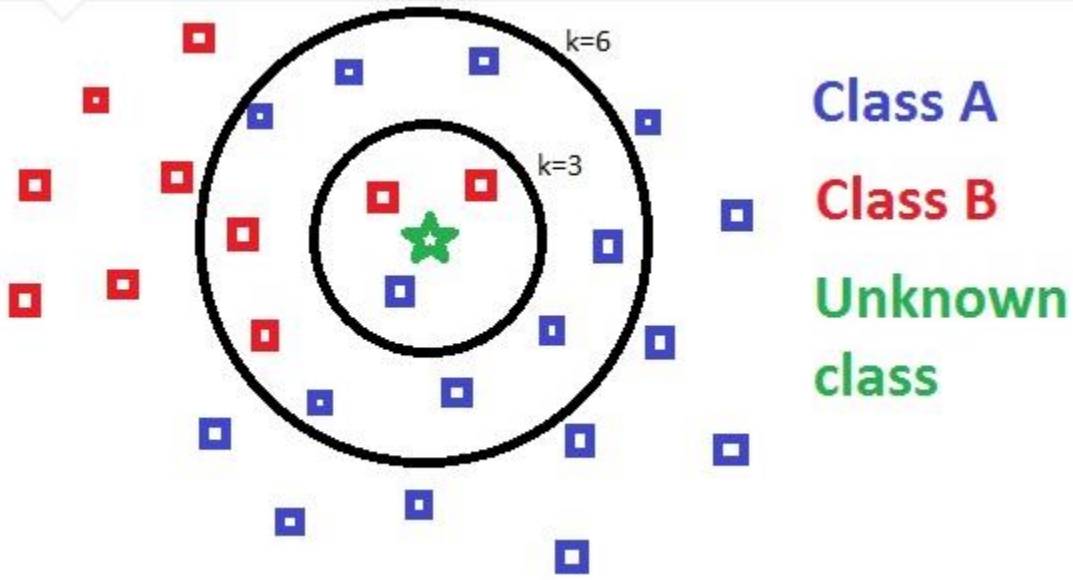


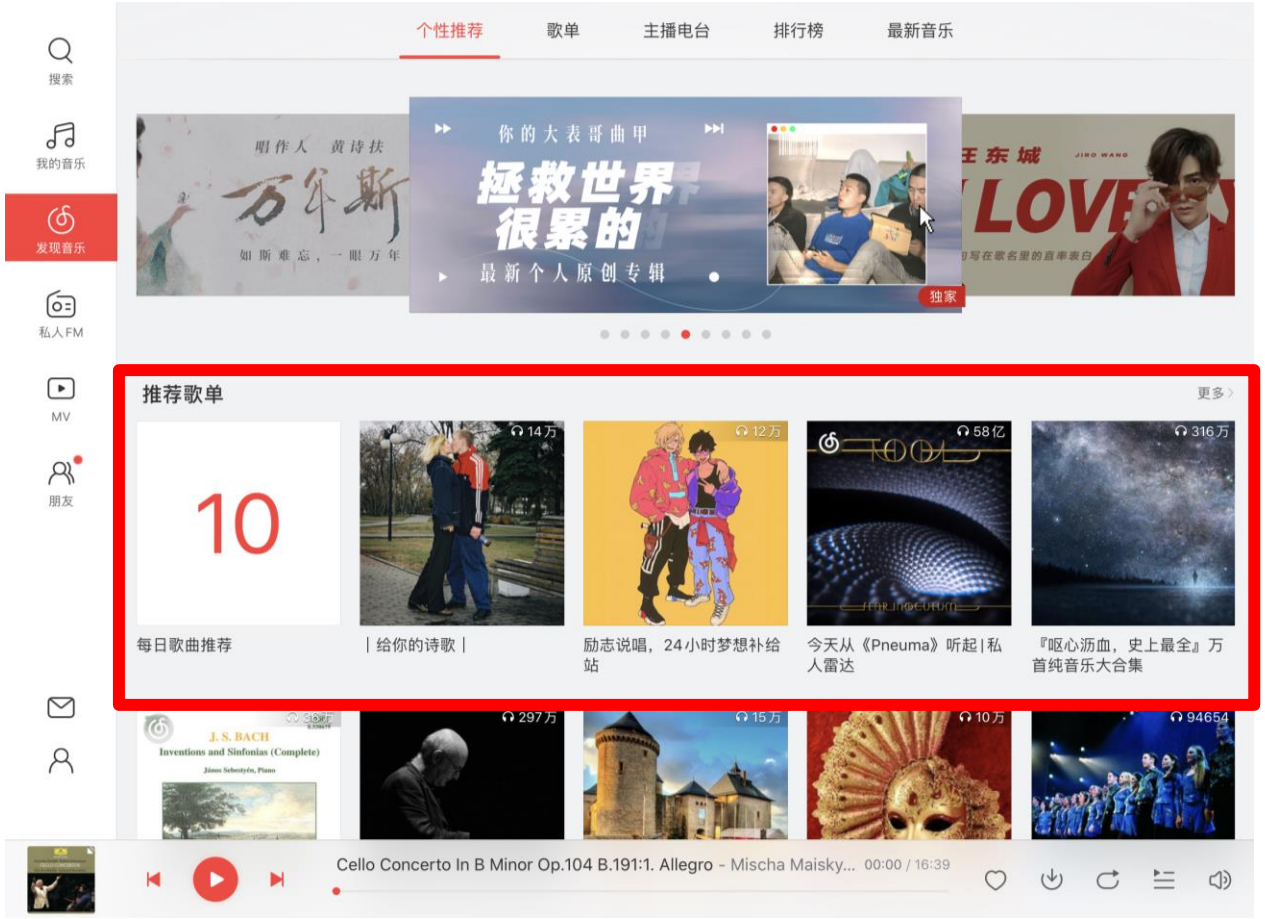
- How to decide the number of centroids ?
- These centroids start from where ?
- When to stop ?



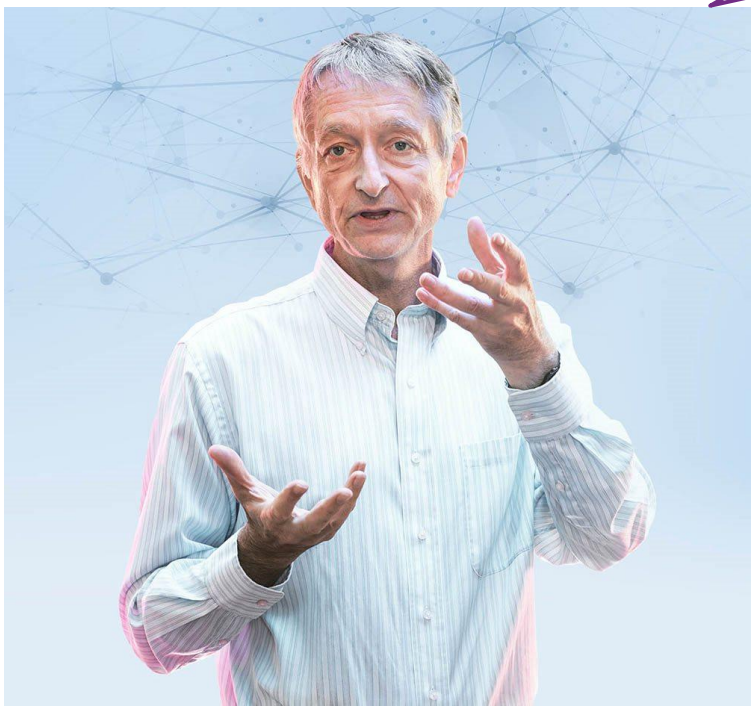
Supervised Learning

K-nearest neighbor (KNN)





We have about 10^14 synapses and only live for about 10^9 seconds. Experience are expensive, synapses are cheap.



© Geoffery E.Hinton

REVIEW

doi:10.1038/nature14539

Deep learning

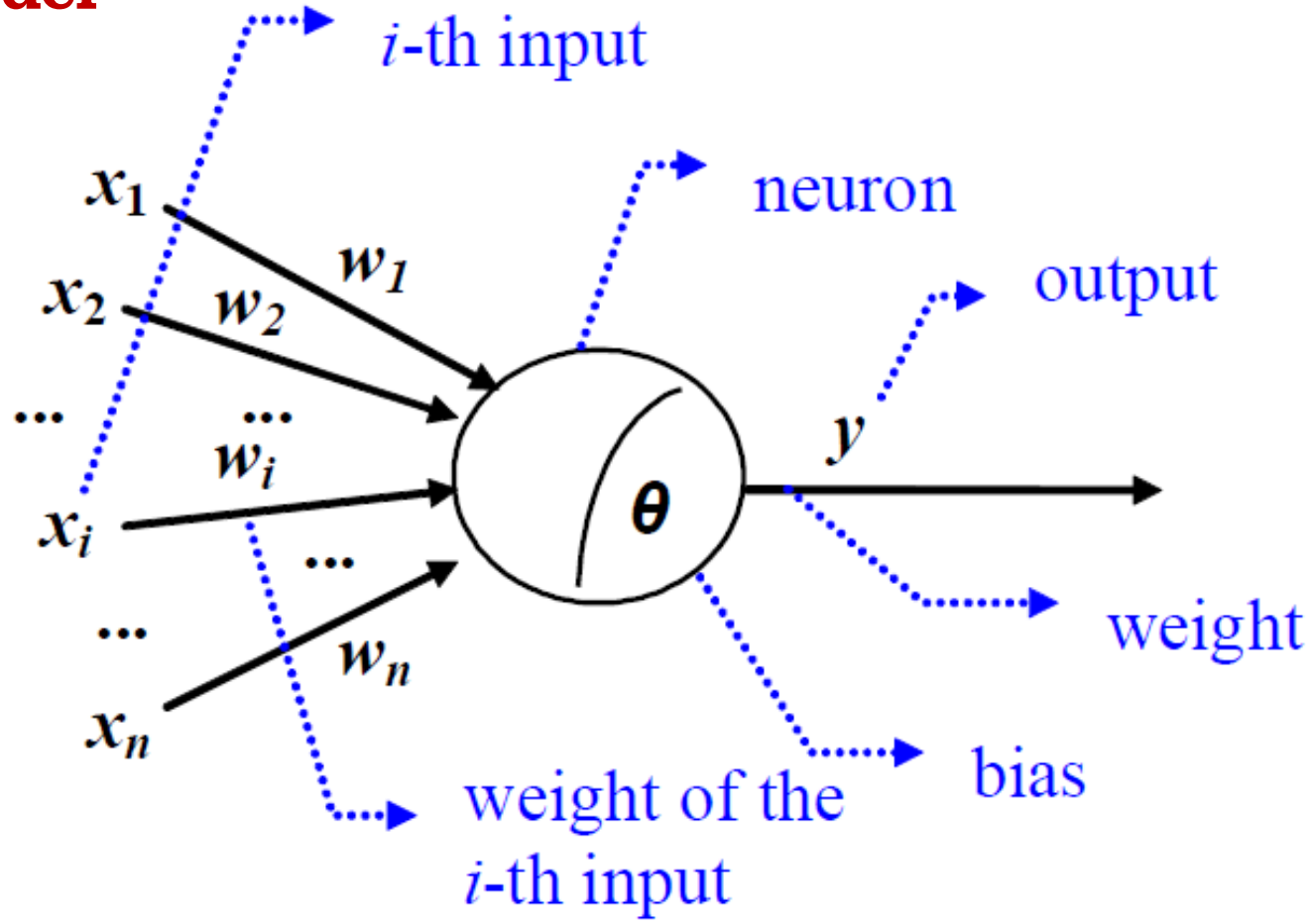
Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

What is a neural network?

Neural networks are massively parallel interconnected networks of simple (usually adaptive) elements and their hierarchical organizations which are intended to interact with the objects of the real world in the same way as biological nervous systems do [T. Kohonen, 1988]

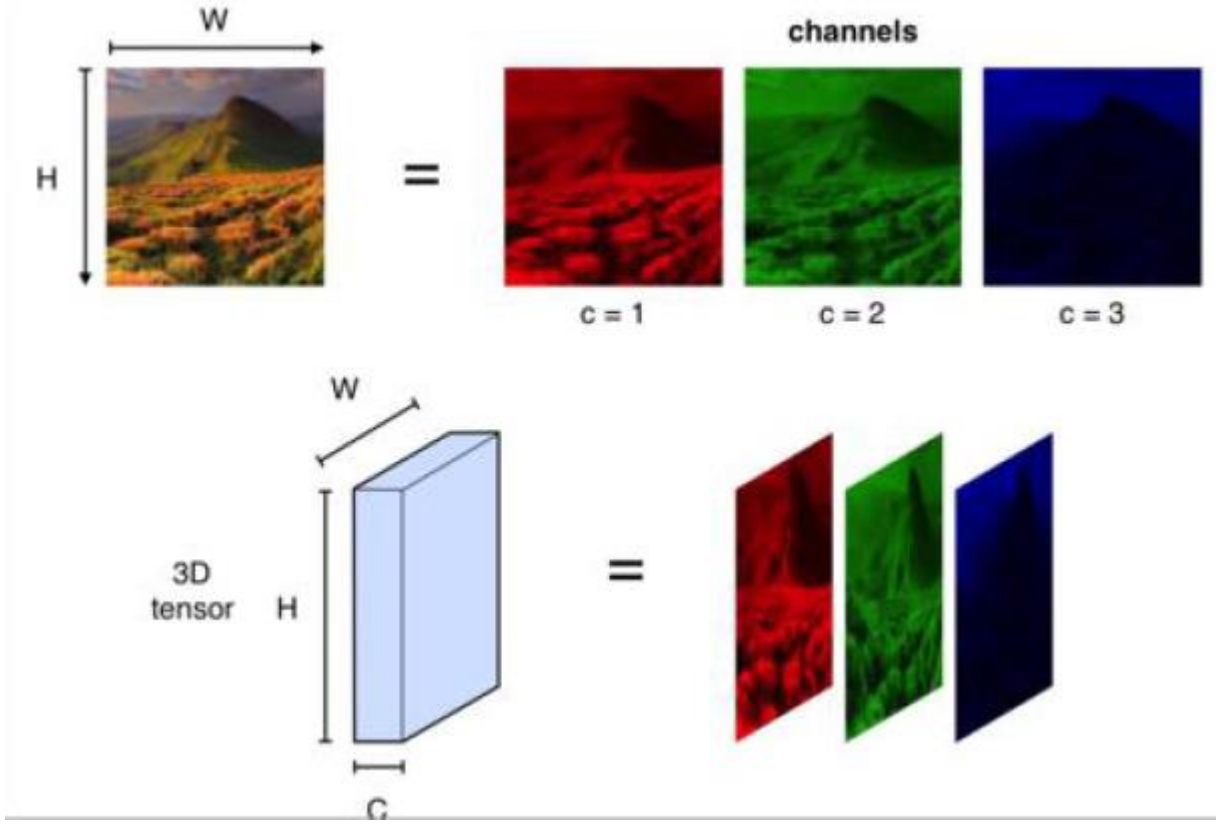
M-P model

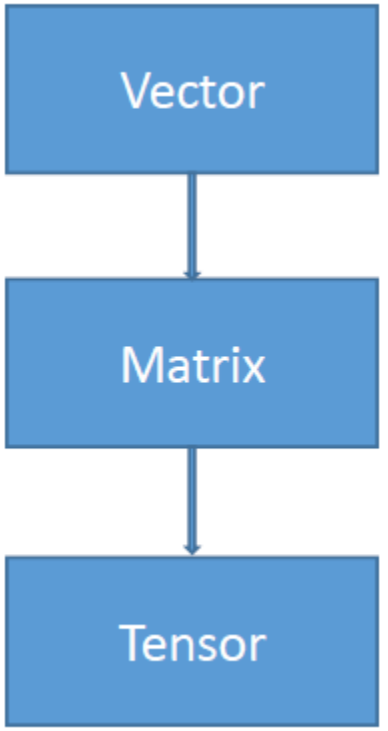


neurons are connected by weights

Data

e.g. RGB Images

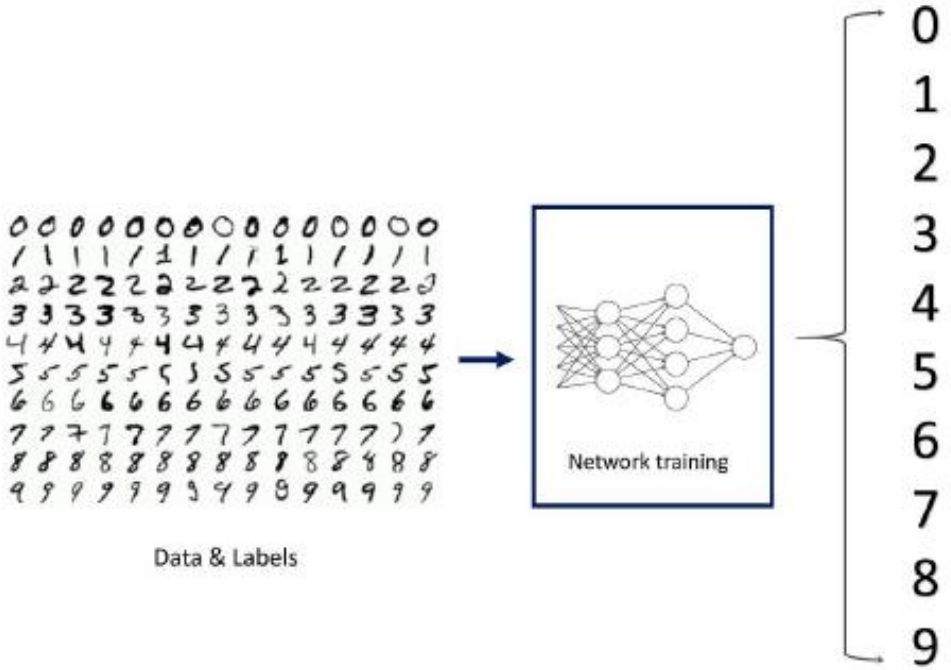


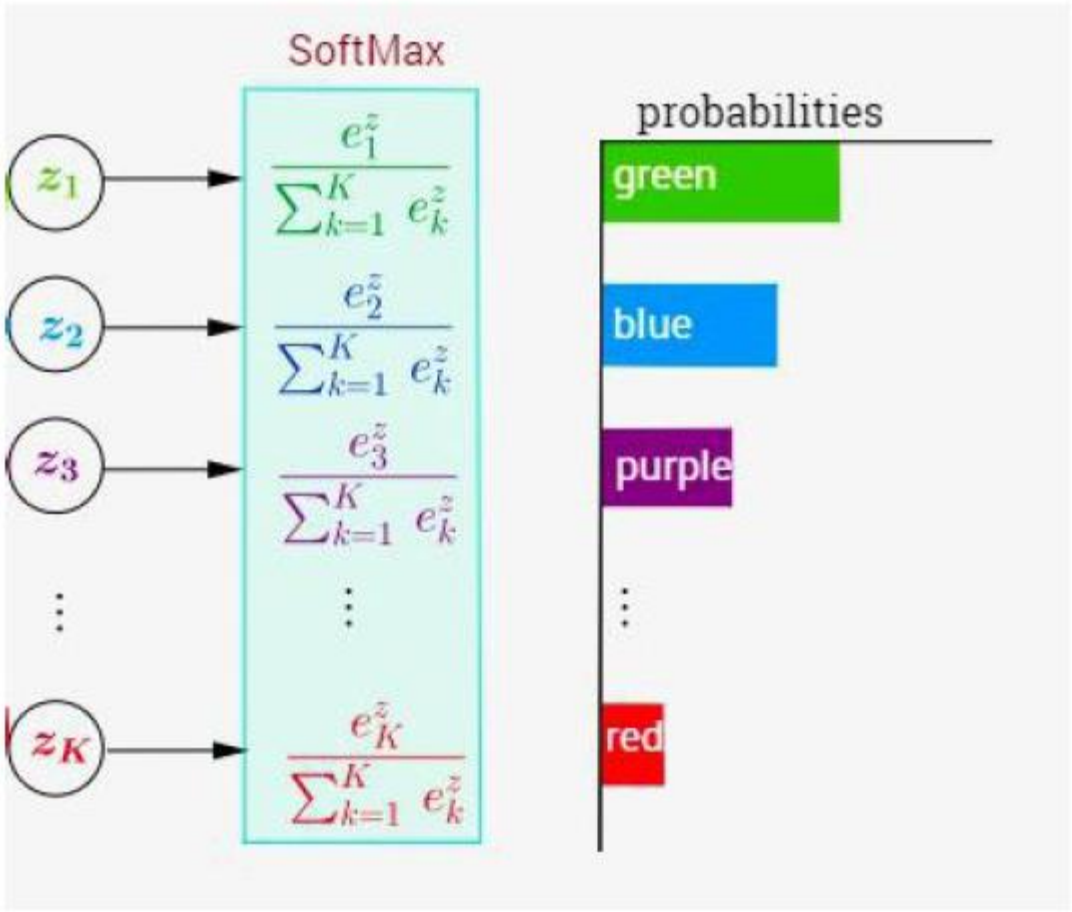


Softmax

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

Mapping vector into probability





Softmax

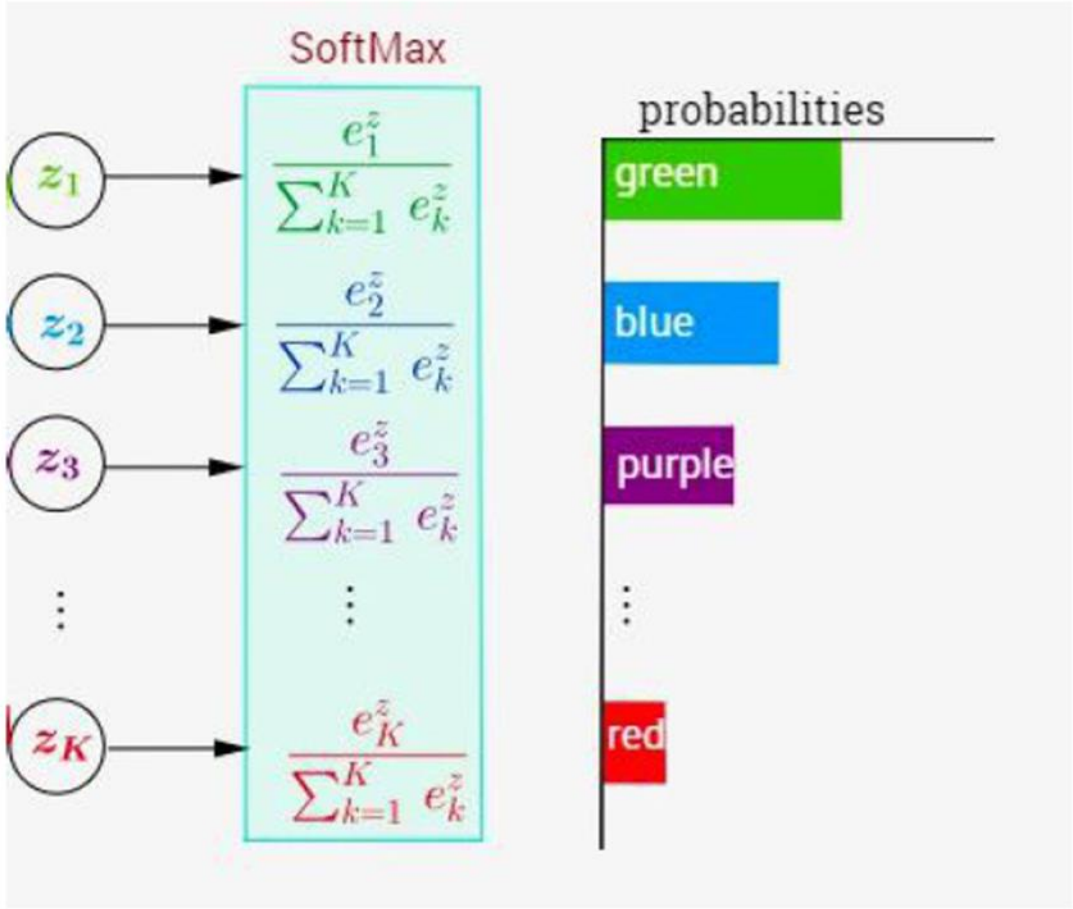
Mapping vector into probability

2
4
1
-2
-3
1
-3
2.01
1

Softmax

0.09492586938614
0.70141257413392
0.034921275782485
0.0017386279448756
6.396054767659E-4
0.034921275782485
6.396054767659E-4
0.095879890234077
0.034921275782485

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$



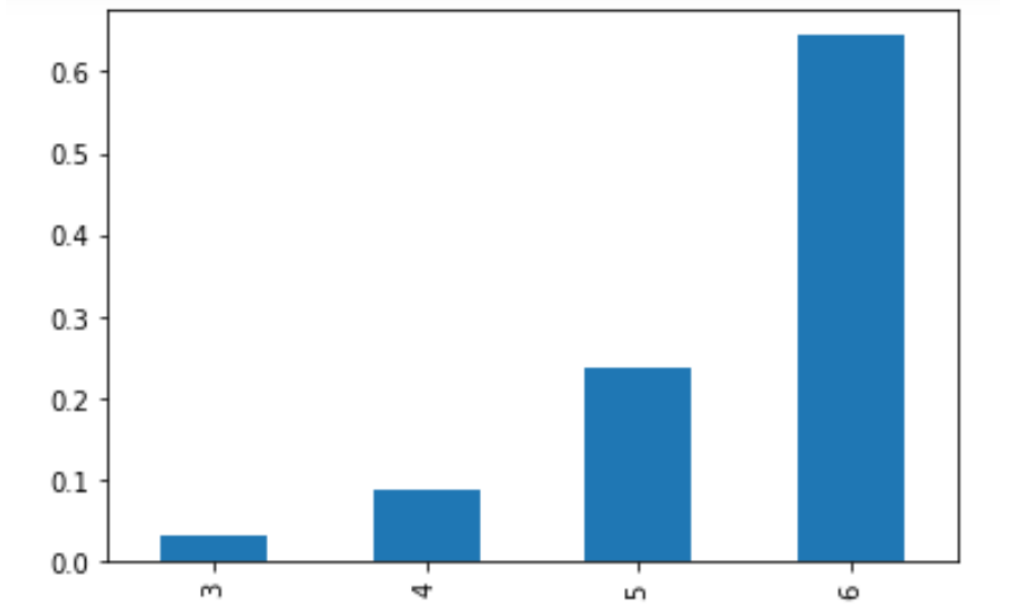
Hands-on

Implement Softmax

1. [1,2,3,4] → [3,4,5,6]
2. [101,102,103,104]

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

def softmax(x):
    sum_exp = np.sum(np.exp(x))
    softmax = np.exp(x)/sum_exp
    return softmax


i = ['1','2','3','4']
a = [1,2,3,4]
asm = softmax(a)
pasm = pd.Series(asm,index = i)
pasm
```

```
In [4]: b = [3,4,5,6]
        i = ['3','4','5','6']
        bsm = softmax(b)
        pbsm = pd.Series(bsm,index = i)
        pbsm

Out[4]: 3    0.032059
        4    0.087144
        5    0.236883
        6    0.643914
        dtype: float64

In [5]: pbsm.plot(kind = 'bar')

Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x2336a32a470>
```



Loss function

Compare output with label

Cross Entropy

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

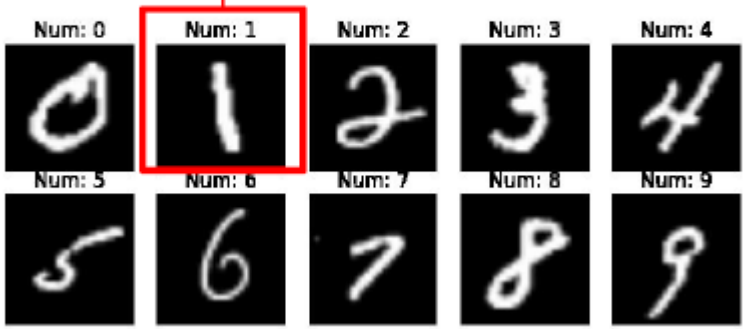
Question: can we change p and q?

2	0.094925869386l4	0
4	0.701412574133 92	1
1	0.034921275782485	0
-2	0.0017386279448756	0
-3	6.396054767659E-4	0
1	0.034921275782485	0
-3	6.396054767659E-4	0
2.01	0.095879890234077	0
1	0.034921275782485	0

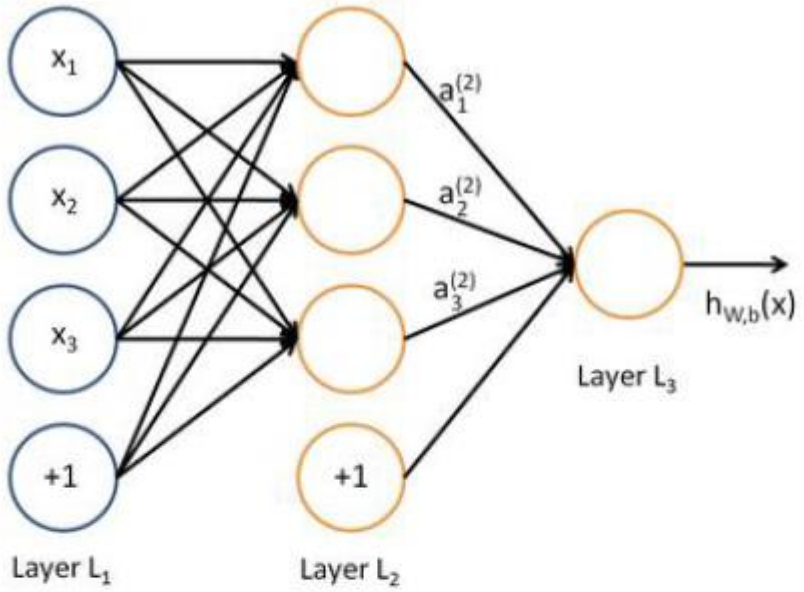
Softmax

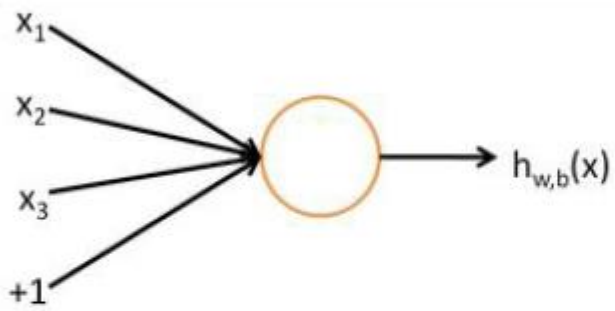
$H(p,q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$

Compare



Linear Layer



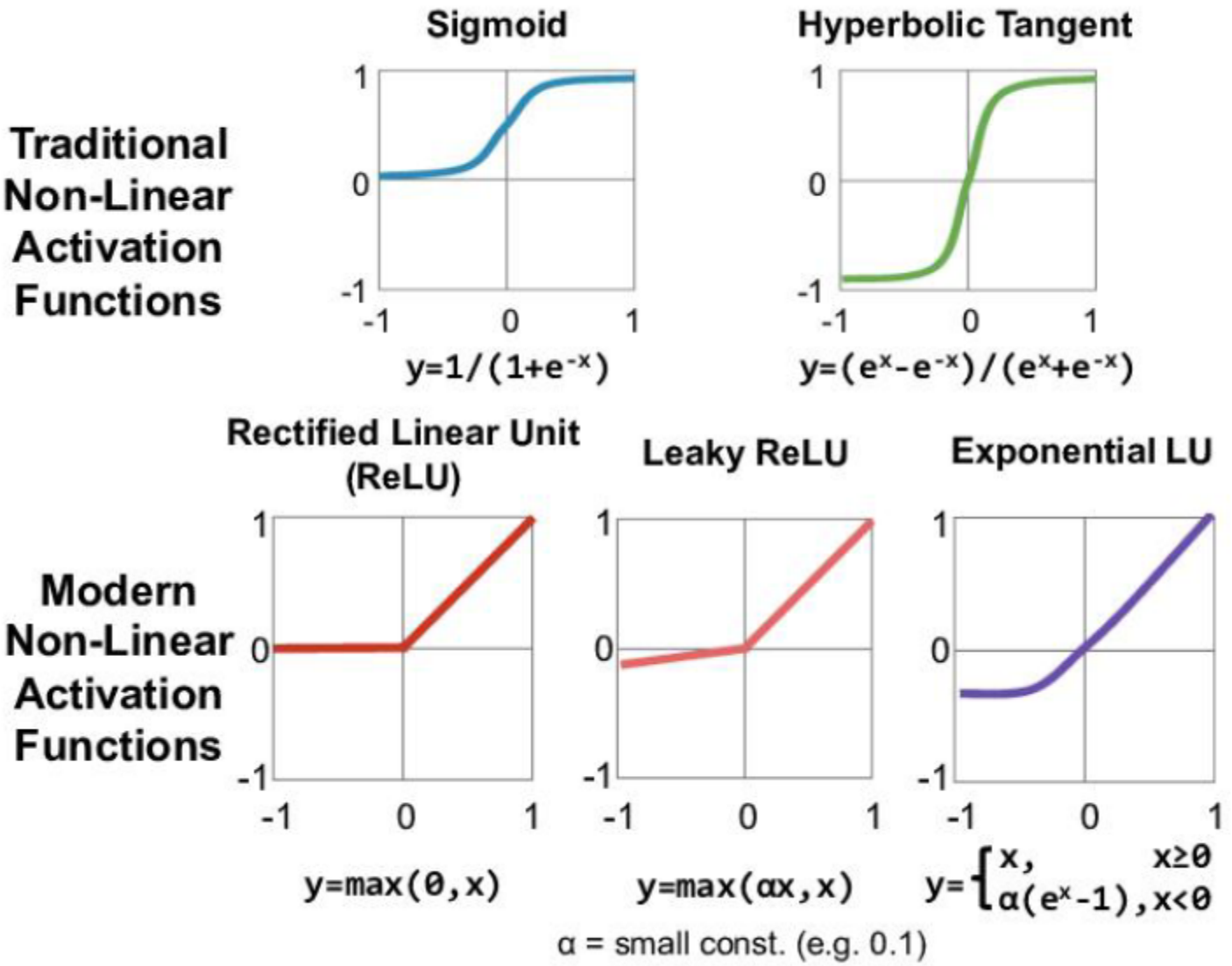


$$
\begin{aligned}
a_1^{(2)} &= f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}) \\
a_2^{(2)} &= f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}) \\
a_3^{(2)} &= f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}) \\
h_{W,b}(x) &= a_1^{(3)} = f(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)})
\end{aligned}
$$

$$h_{W,b}(x) = f(W^T x) = f(\textstyle\sum_{i=1}^{3} W_i x_i + b)$$

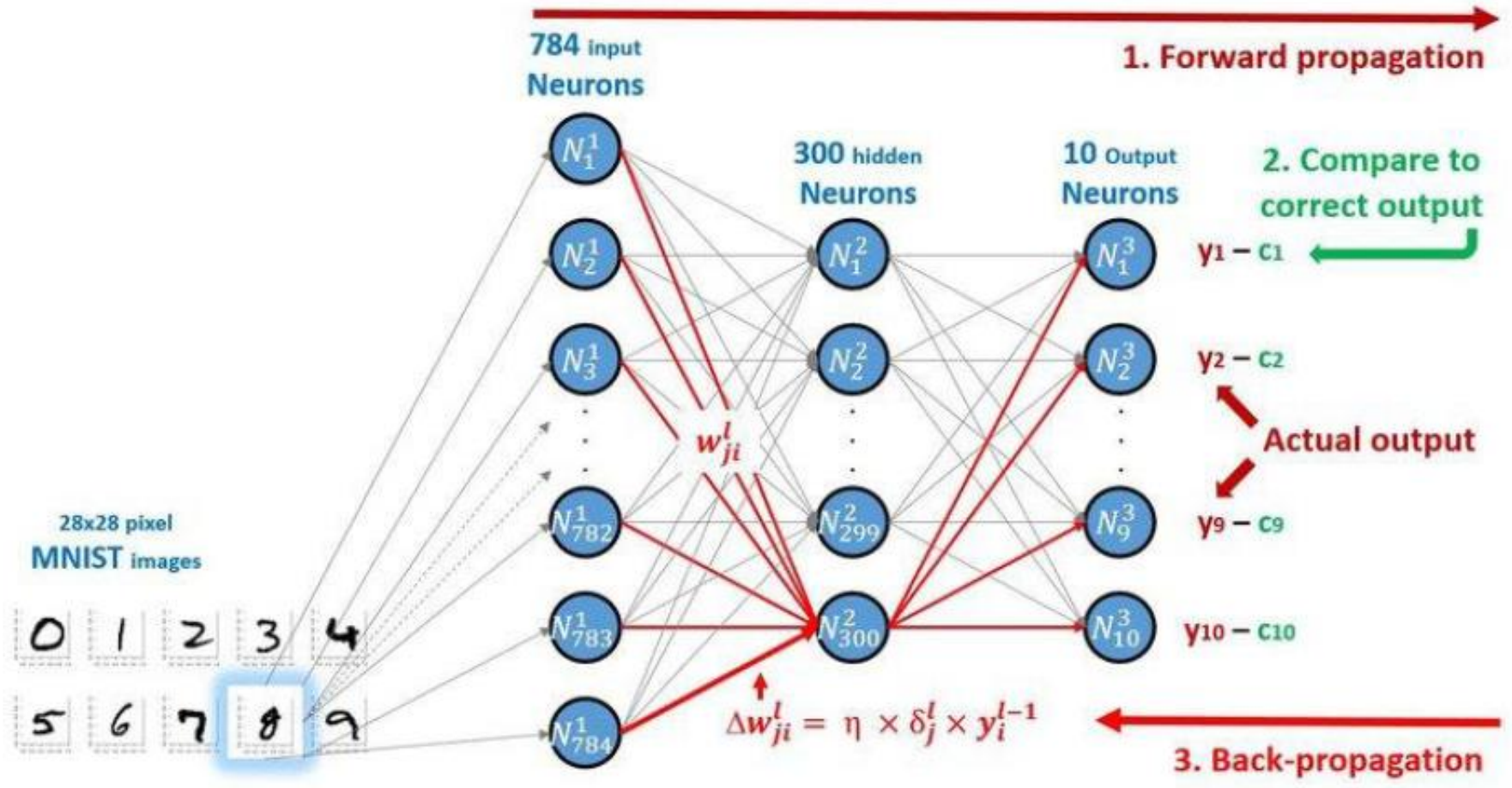
$$f(z) = \frac{1}{1+\exp(-z)}. \quad f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}},$$

Activation Layer

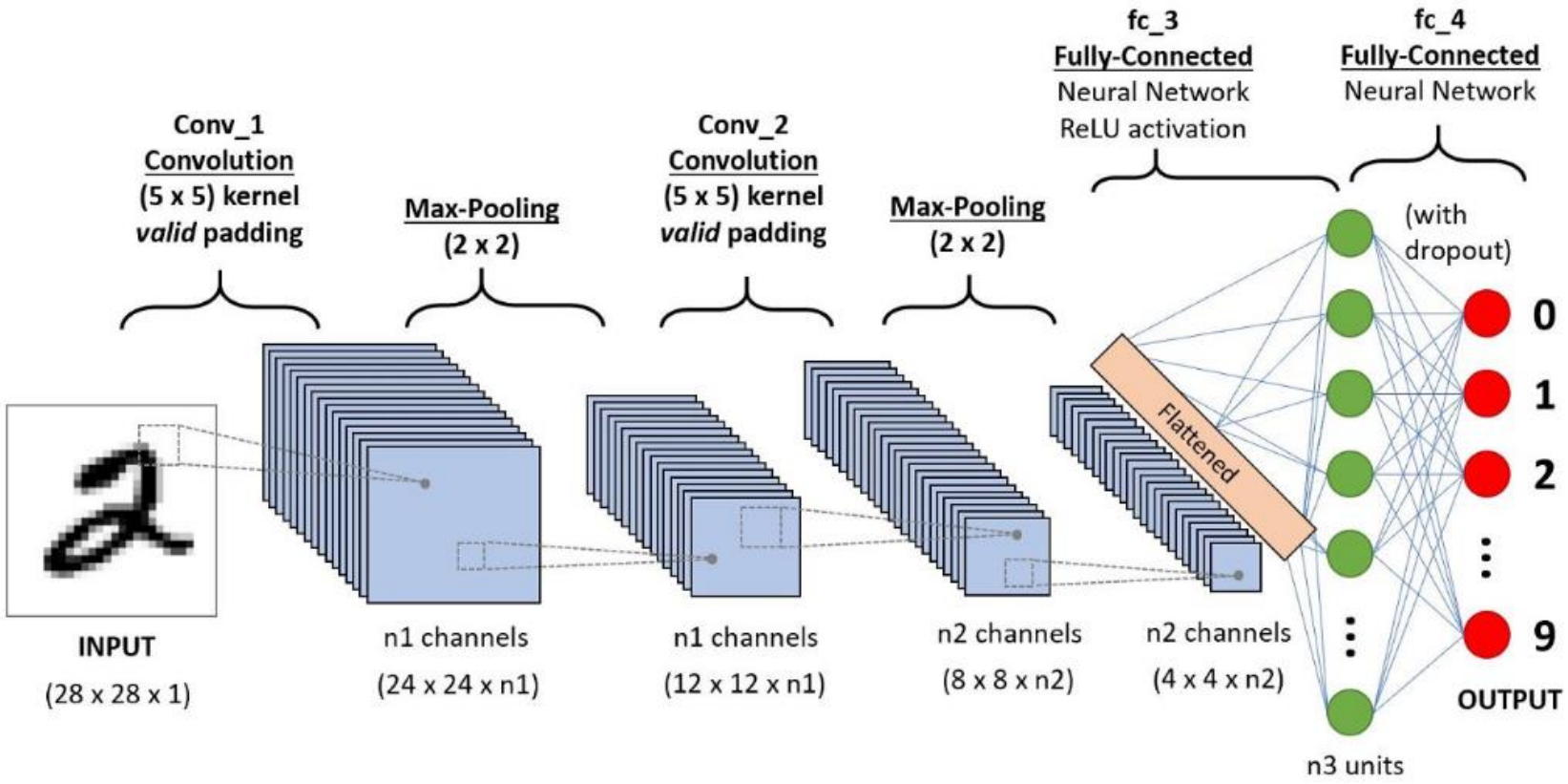


Back Propagation

Forward Pass and Backward Pass



The structure of a convolutional neural network(CNN)



BP (backpropagation)

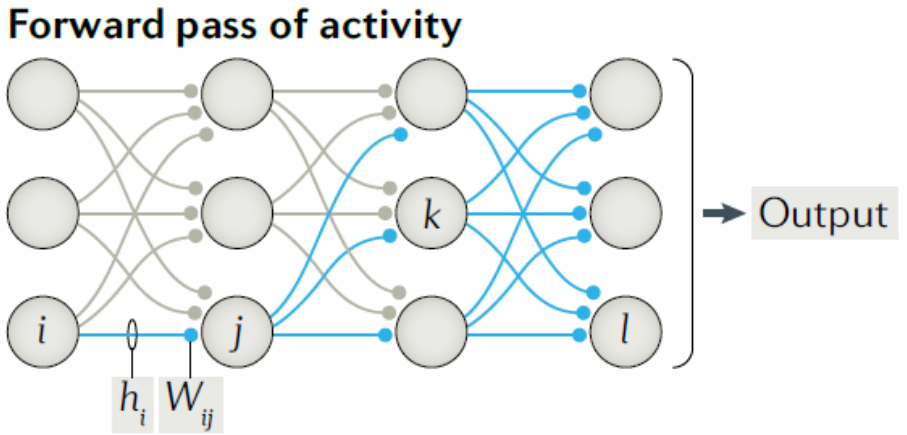
Step 1: feedforward input from input layer to hidden layer to output layer

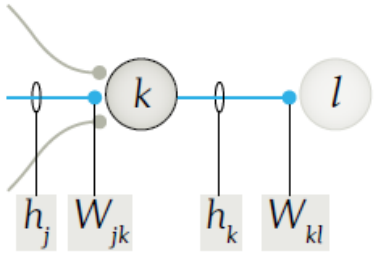
Step 2: compute the error of the output layer

Step 3: backpropagate the error from output layer to hidden layer

Step 4: adjust weights and biases

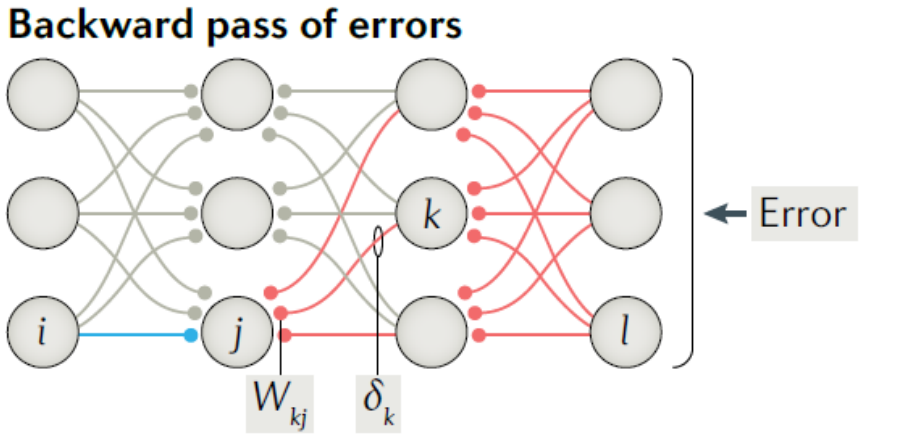
Step 5: if termination criterion is satisfied, stop. Otherwise go to step1

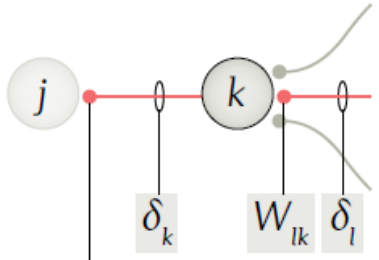




$$h_k = f(a_k)$$

$$a_k = \sum_j h_j W_{jk}$$





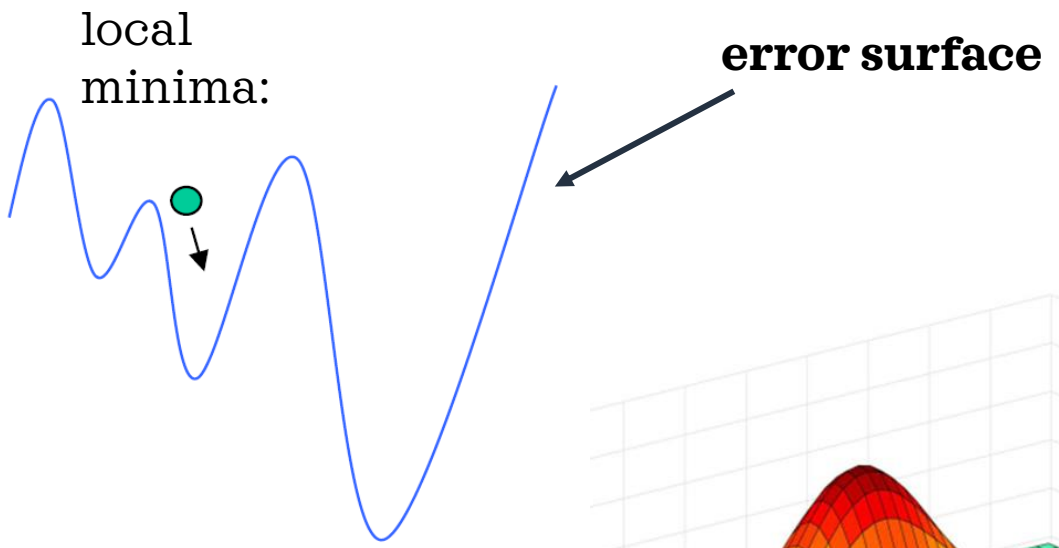
$$W_{kj} = W_{jk}$$

$$\delta_k = e_k f'(a_k)$$

$$e_k = \sum_l \delta_l W_{lk}$$

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta h_i \delta_j \text{ where } \delta_j = e_j f'(a_j) = \left(\sum_k \delta_k W_{jk} \right) f'(a_j)$$

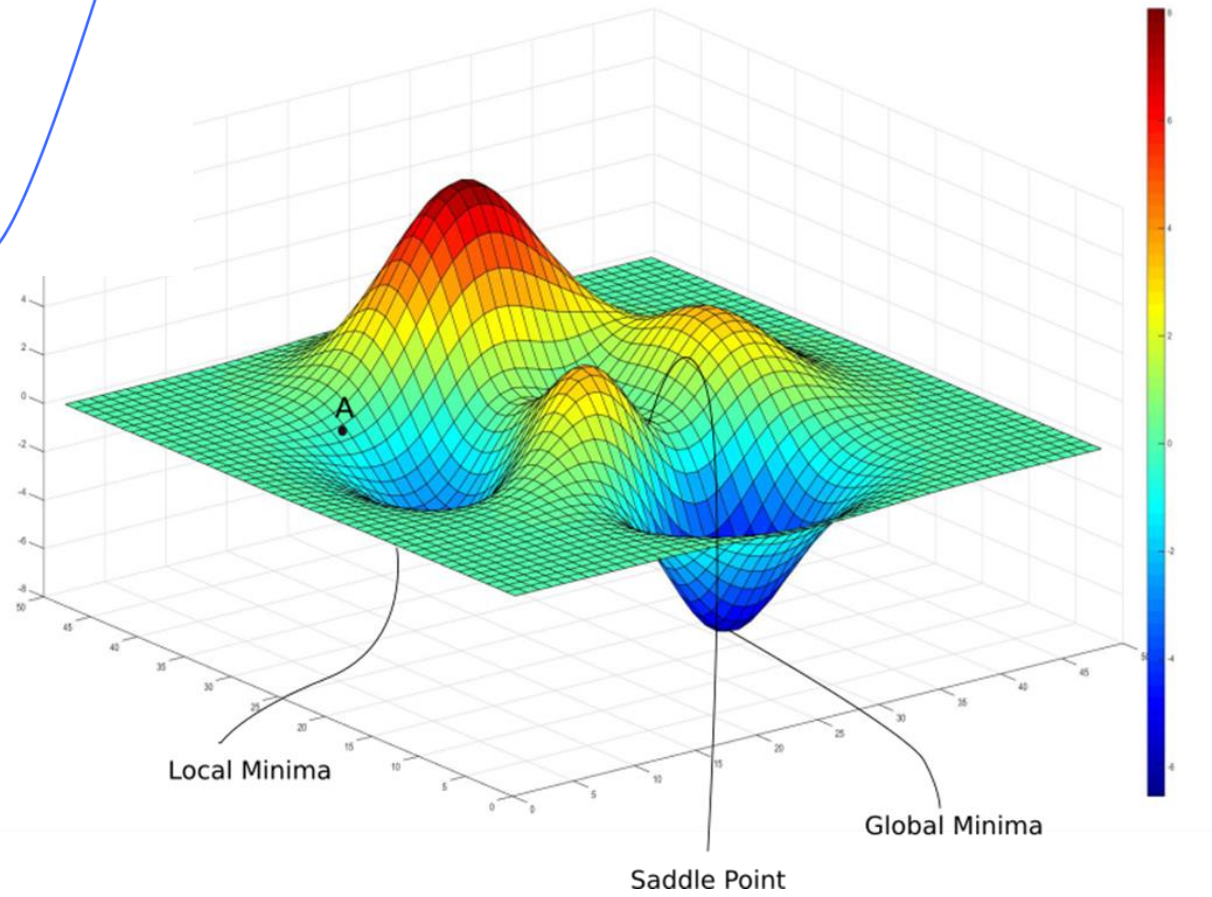
Remarks



Solution of BP:

- Guaranteed local minima
- Random initiation and stochastic gradient descent make it less likely

to get stuck in the local minima



Pytorch

1. Concept of tensor. https://pytorch.org/docs/stable/tensors.html

2. Cross Entropy Loss in Pytorch: Be Careful
https://pytorch.org/docs/stable/nn.functional.html?highlight=cross%20entropy#torch.nn.functional.cross_entropy

E.g.

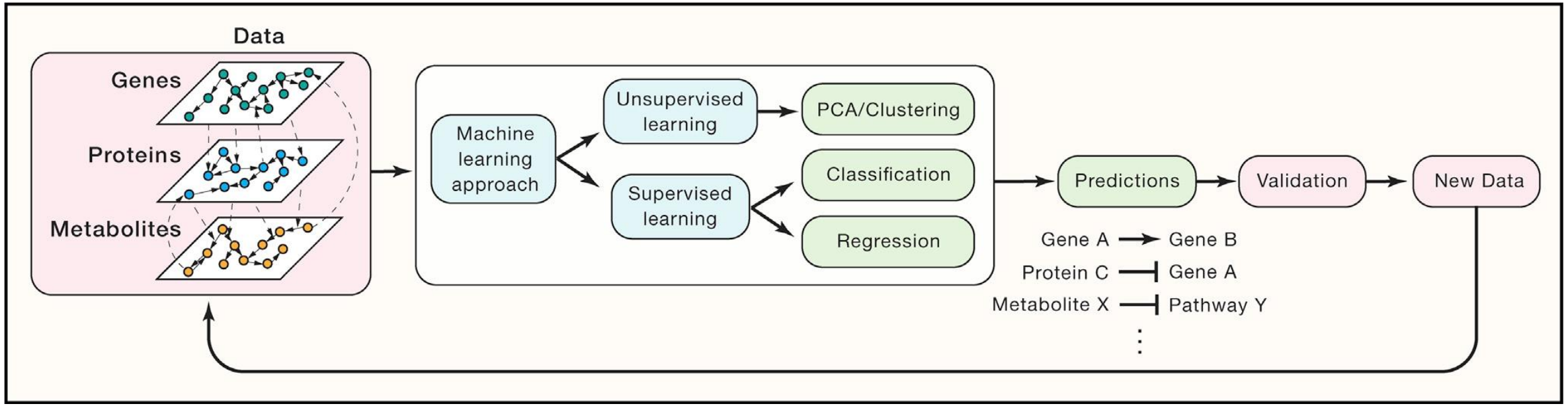
```
torch.nn.functional.cross_entropy(torch.tensor([[1.+2,2.+2,3.+2]]), torch.tensor([1]))
```

```
tensor(1.4076)
```

Overview

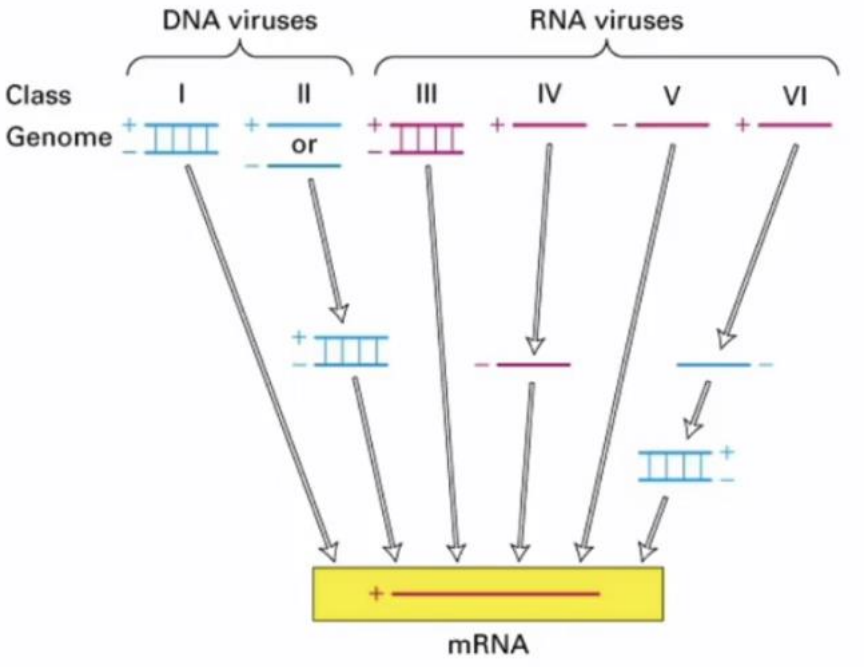
- I . The methodology and technology in biological research
- II. A Primer on Machine Learning
- III. Bio+AI: two demos

Machine-Learning Applications Build Models to Interpret and Analyze Datasets

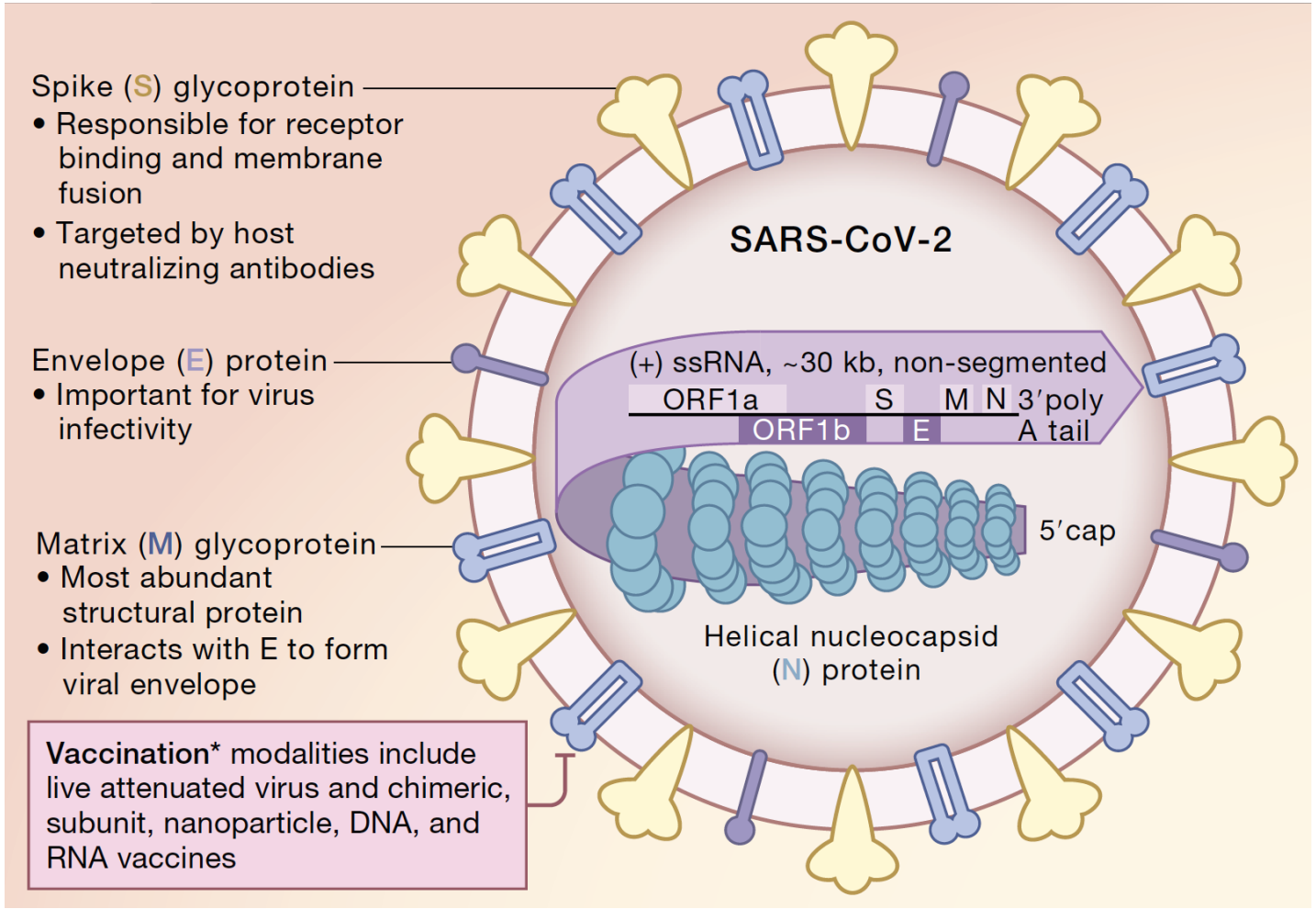


Data consist of features measured over many samples, including quantification of genes, proteins, metabolites, and edges within networks.

SARS-CoV-2 is an approximately 30-kb, single-stranded, positive-sense RNA



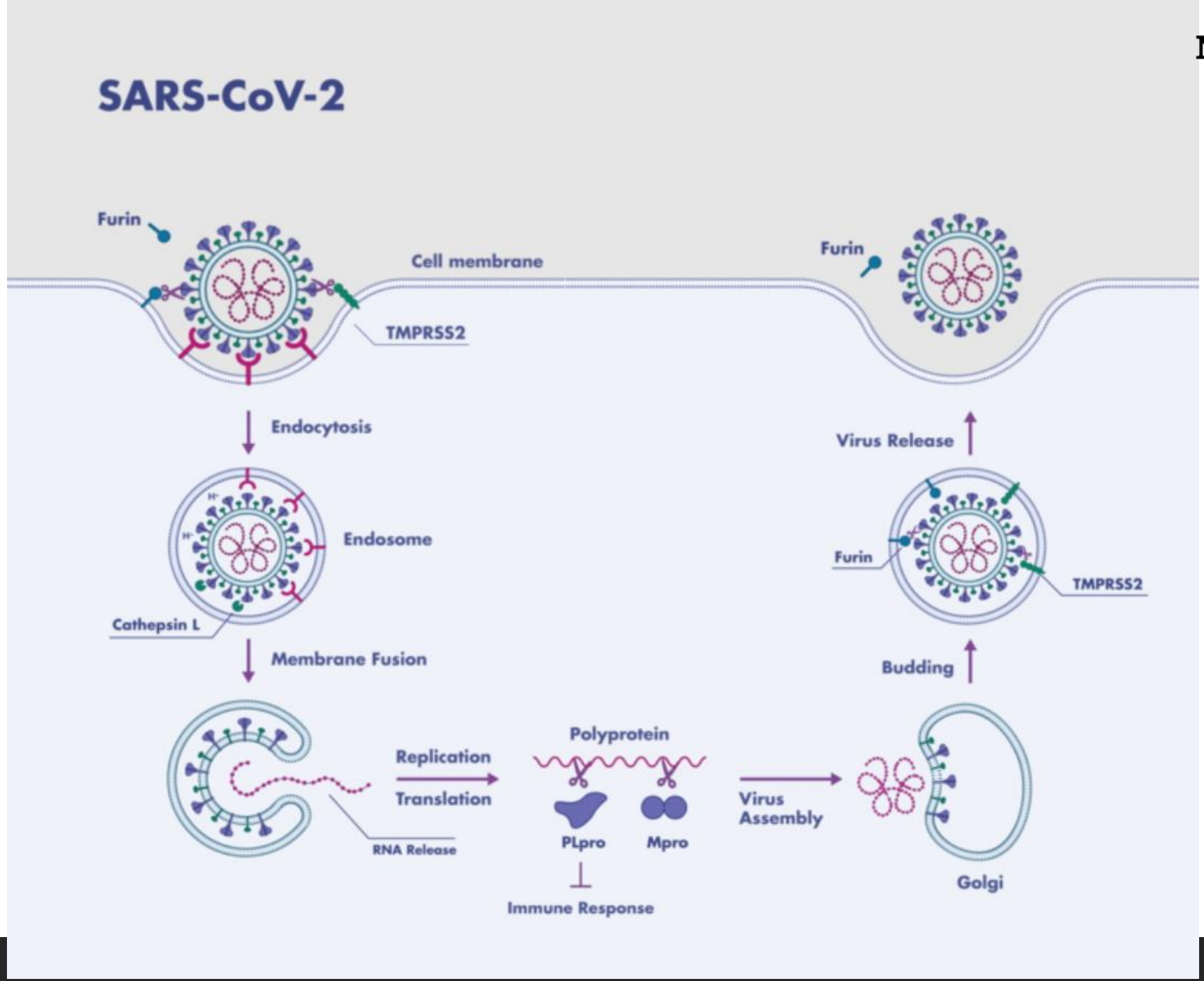
29(16+4+9) proteins

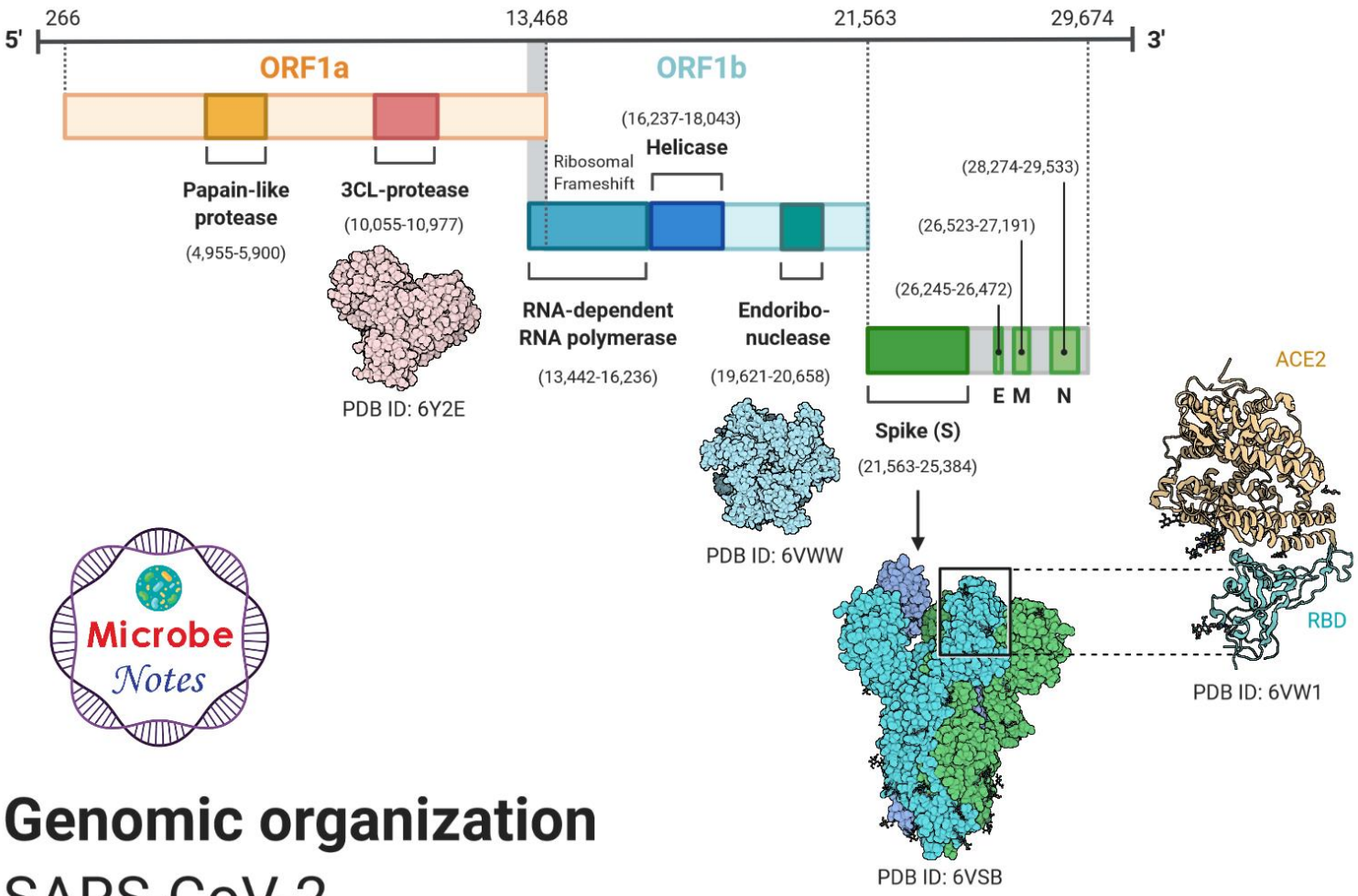


1. Entry

2.Transcription

3. Assembly

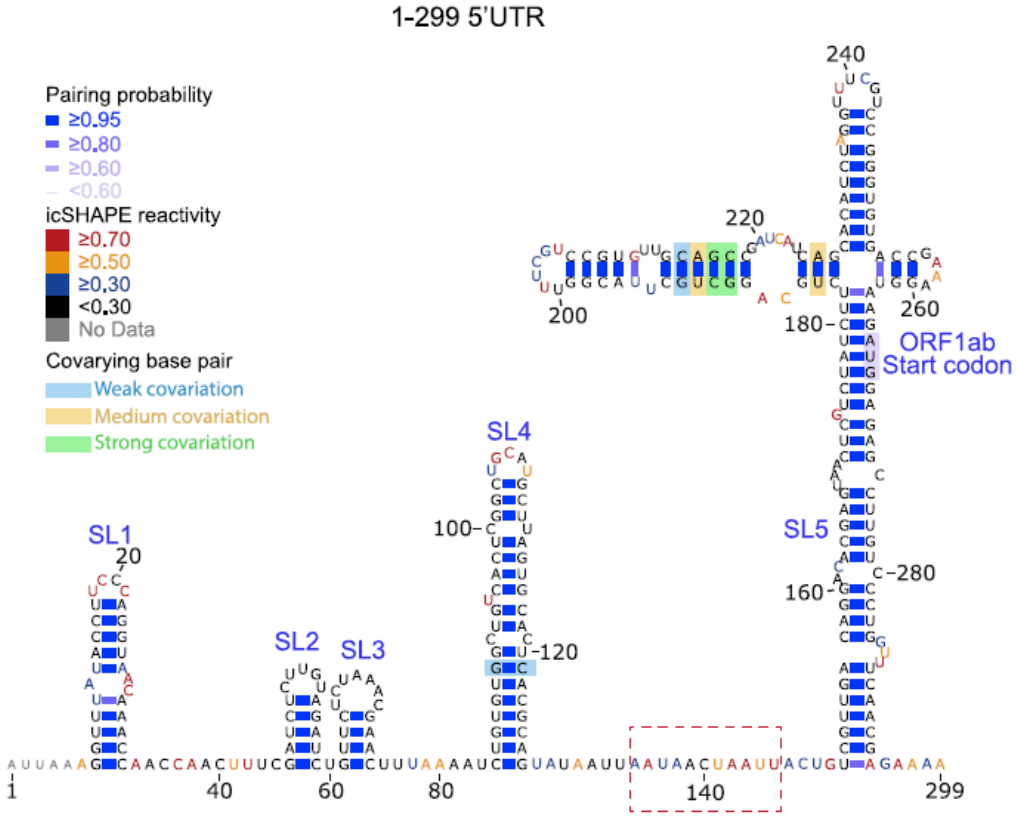


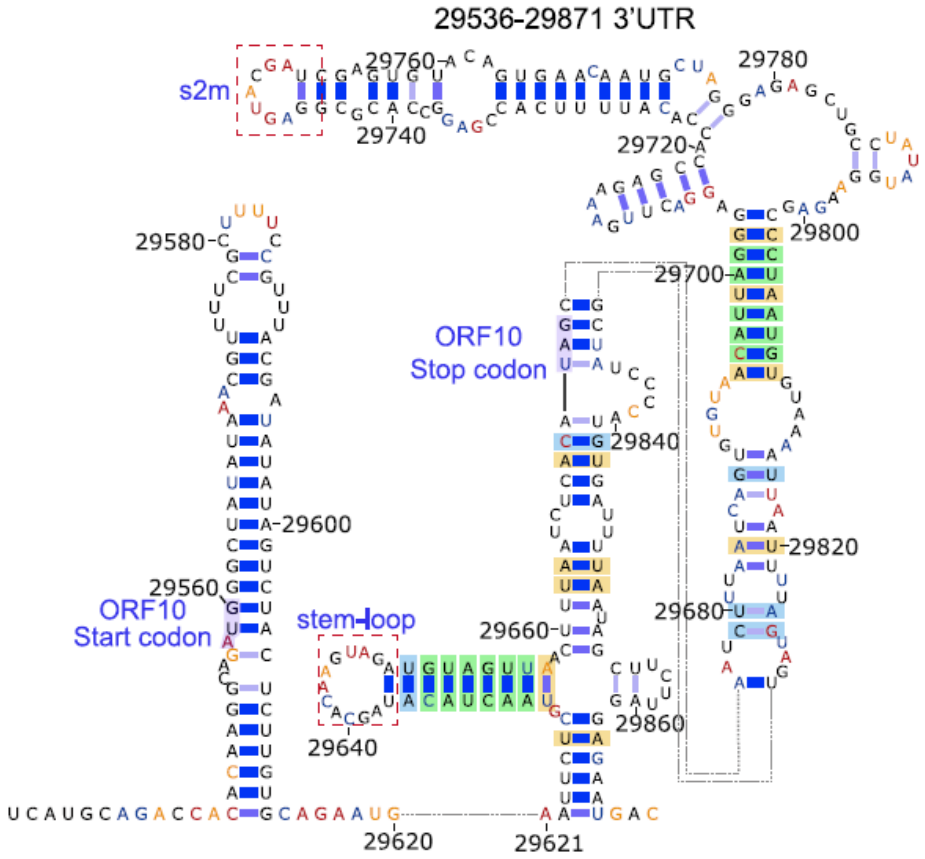


Genomic organization

SARS-CoV-2

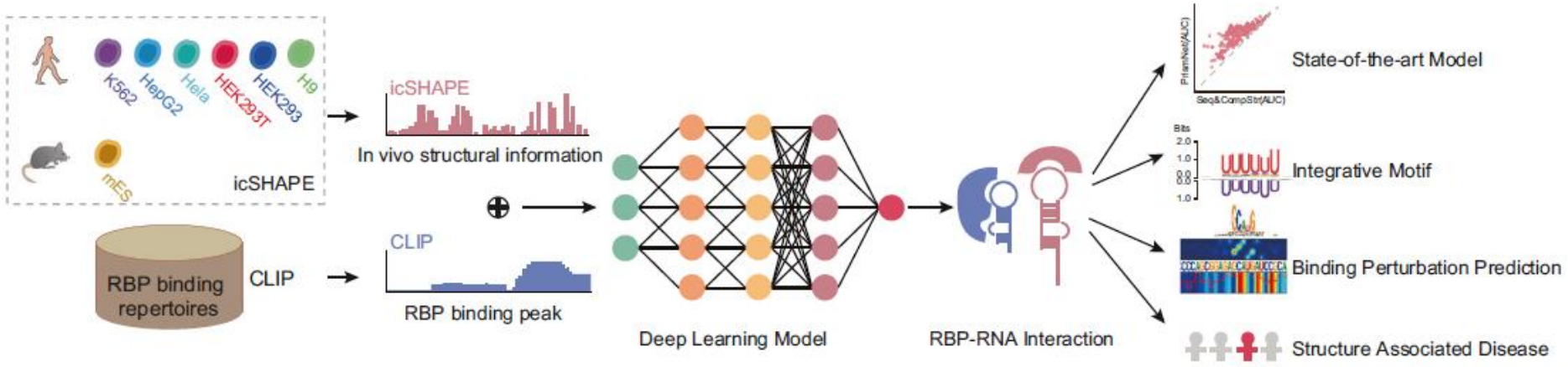
Demo I A neural network model that integrates in vivo RNA Bing Protein binding and RNA features in matched cells

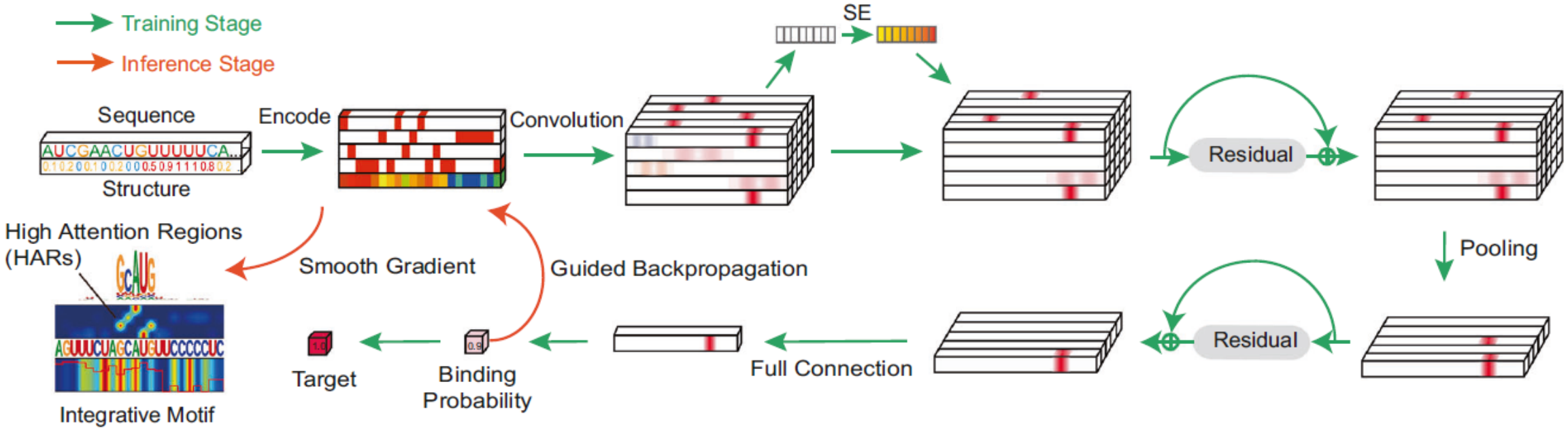




High attention regions (HARs) are predicted to be the exact locations of RBP binding nucleotides

NJU-AIA



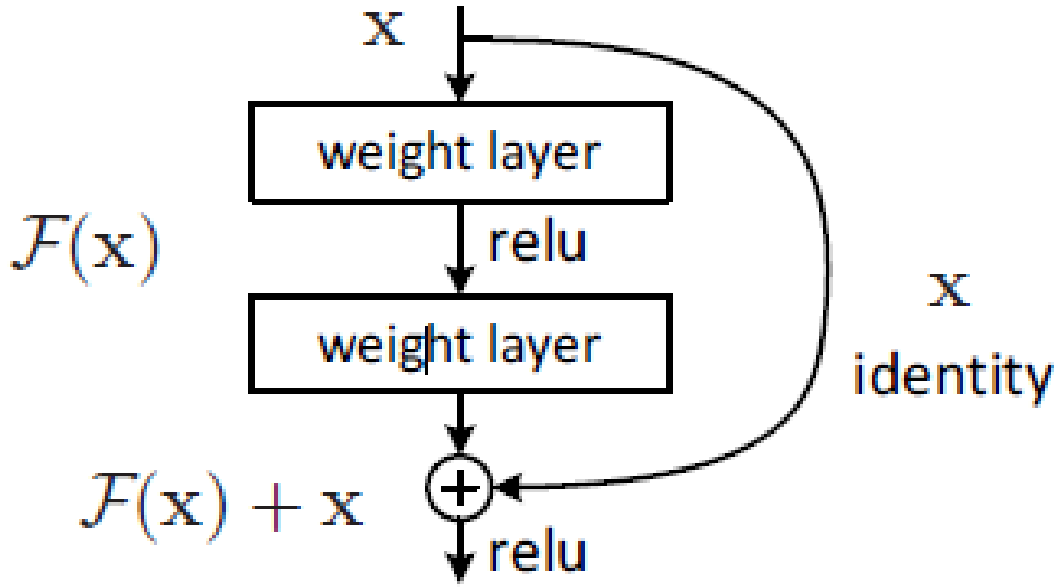


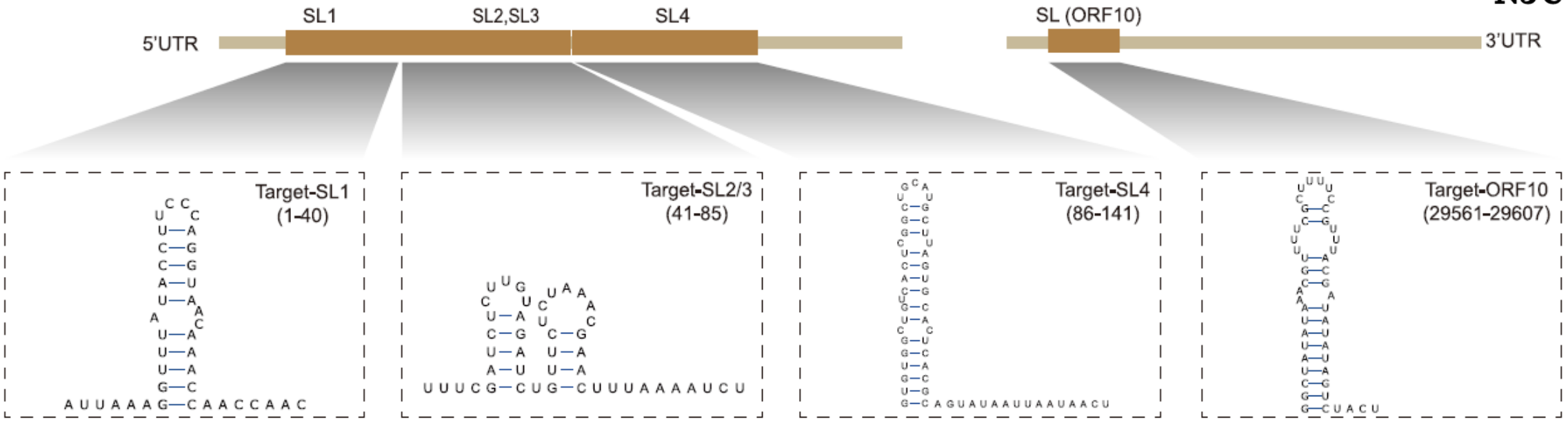
ResNet

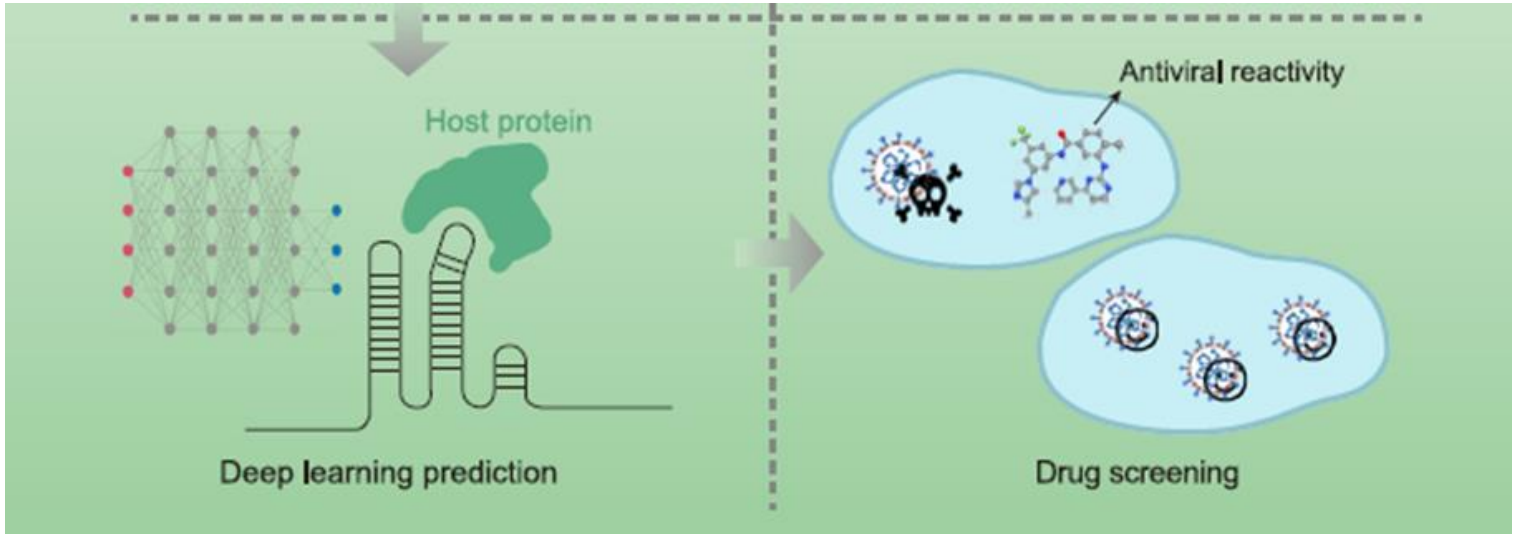
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

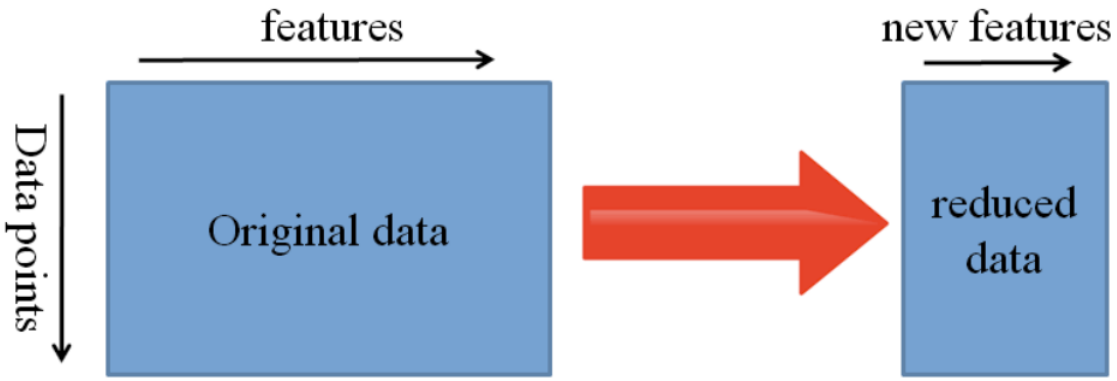


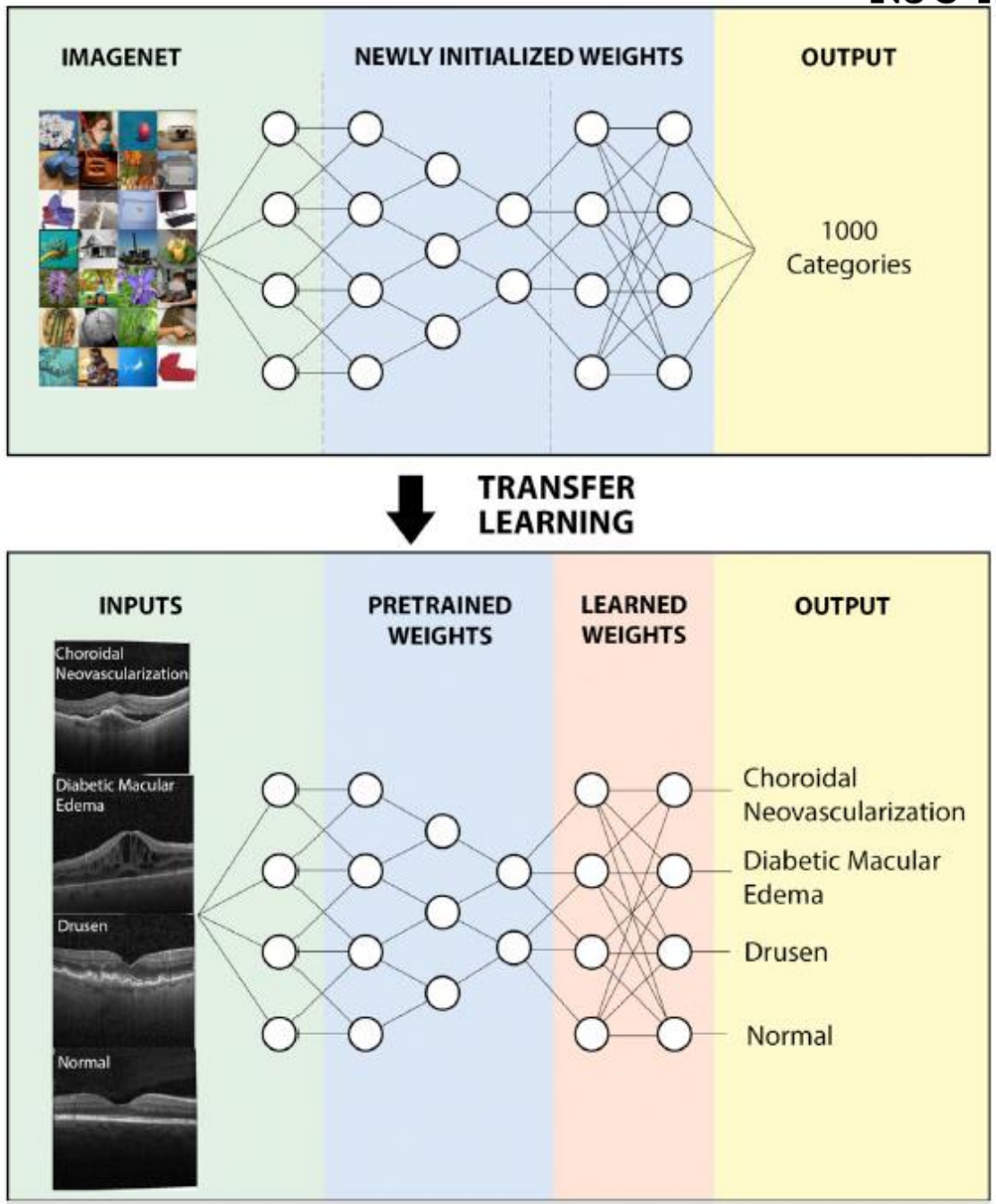

SL1
SL2,SL3
SL4
SL (ORF10)
5'UTR
3'UTR
Target-SL1
(1-40)
Target-SL2/3
(41-85)
Target-SL4
(86-141)
Target-ORF10
(29561-29607)

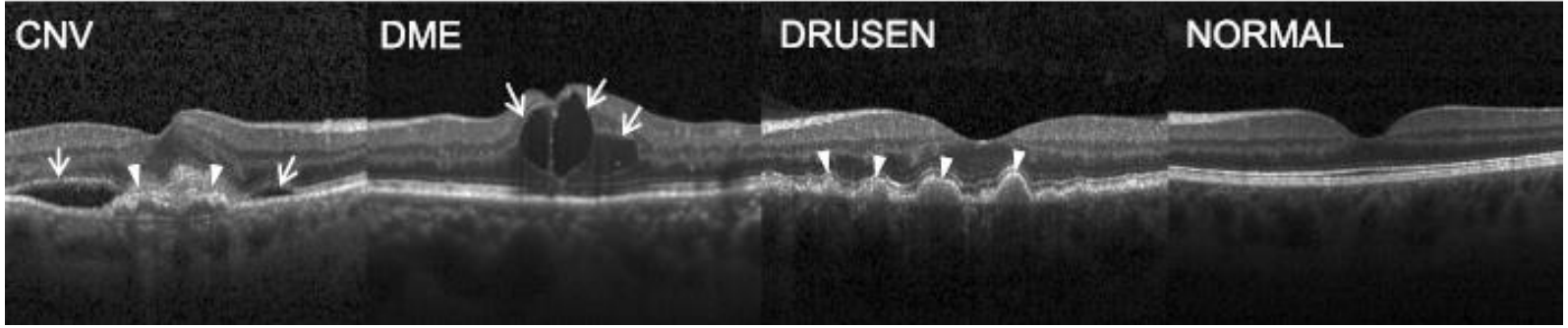

Host protein
Deep learning prediction
Antiviral reactivity
Drug screening

Demo II Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning

Transfer Learning







NJU-AIA

B

True Labels \ Predicted Labels	NORMAL	DRUSEN	CNV	DME
NORMAL	246	5	0	3
DRUSEN	4	236	3	1
CNV	0	7	242	9
DME	0	2	5	237

Total image numbers
n = 207,130 OCT images

Label/grade images

Create model using transfer learning on TensorFlow platform

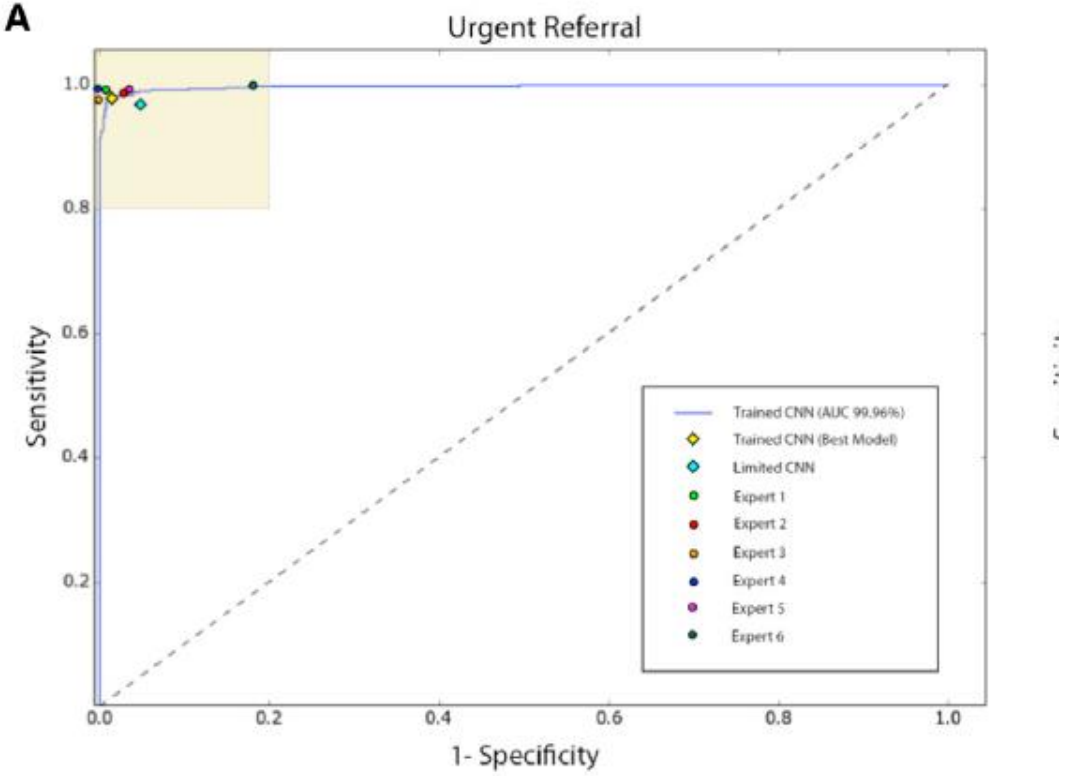
Training dataset
n = 108,312
(37,206 CNV, 11,349 DME, 8,617 Drusen, 51,140 Normal)
4,686 patients

Validation dataset
n = 1,000
(250 CNV, 250 DME, 250 drusen, 250 normal)

Outcome evaluations
(AUC, errors, accuracy)

Occlusion test

A



井

All knowledge is, in final analysis, history.

All sciences are, in the abstract, mathematics.

All judgements are, in their rationale, statistics.

(在终极的分析中，一切知识都是历史；

在抽象的意义下，一切科学都是数学；

在理性的世界里，所有的判断都是统计学。)

-- C. R. Rao.

References

1. LeCun, Yann, et al. “Deep Learning.” Nature, vol. 521, no. 7553, 2015, pp. 436–444.
2. Wu, Xindong, et al. “Top 10 Algorithms in Data Mining.” Knowledge and Information Systems, vol. 14, no. 1, 2007, pp. 1–37.
3. Zeiler, Matthew D., and Rob Fergus. “Visualizing and Understanding Convolutional Networks.” 13th European Conference on Computer Vision, ECCV 2014, 2014, pp. 818–833.
4. Neural Network Structure pictures credit to Simon Xin Dong.
5. Kermany, Daniel S., et al. “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.” Cell, vol. 172, no. 5, 2018, pp. 1122–1131.
6. Camacho, Diogo M., et al. “Next-Generation Machine Learning for Biological Networks.” Cell, vol. 173, no. 7, 2018, pp. 1581–1592.
7. Finkel, Yaara, et al. “The Coding Capacity of SARS-CoV-2.” BioRxiv, 2020.
8. Sun, Lei, et al. “Predicting Dynamic Cellular Protein–RNA Interactions by Deep Learning Using in Vivo RNA Structures.” Cell Research, 2021, pp. 1–22.
9. Sun, Lei, et al. “In Vivo Structural Characterization of the SARS-CoV-2 RNA Genome Identifies Host Proteins Vulnerable to Repurposed Drugs.” Cell, 2021.
10. He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

Thanks for your attention!

Being provident.

欢迎加入NJU AIA！

南京大学人工智能协会**公众号**

网站: njuaia.github.io