Deep learning for Regulatory genomics

Mahboobeh (Mariya) Golchinpour leili

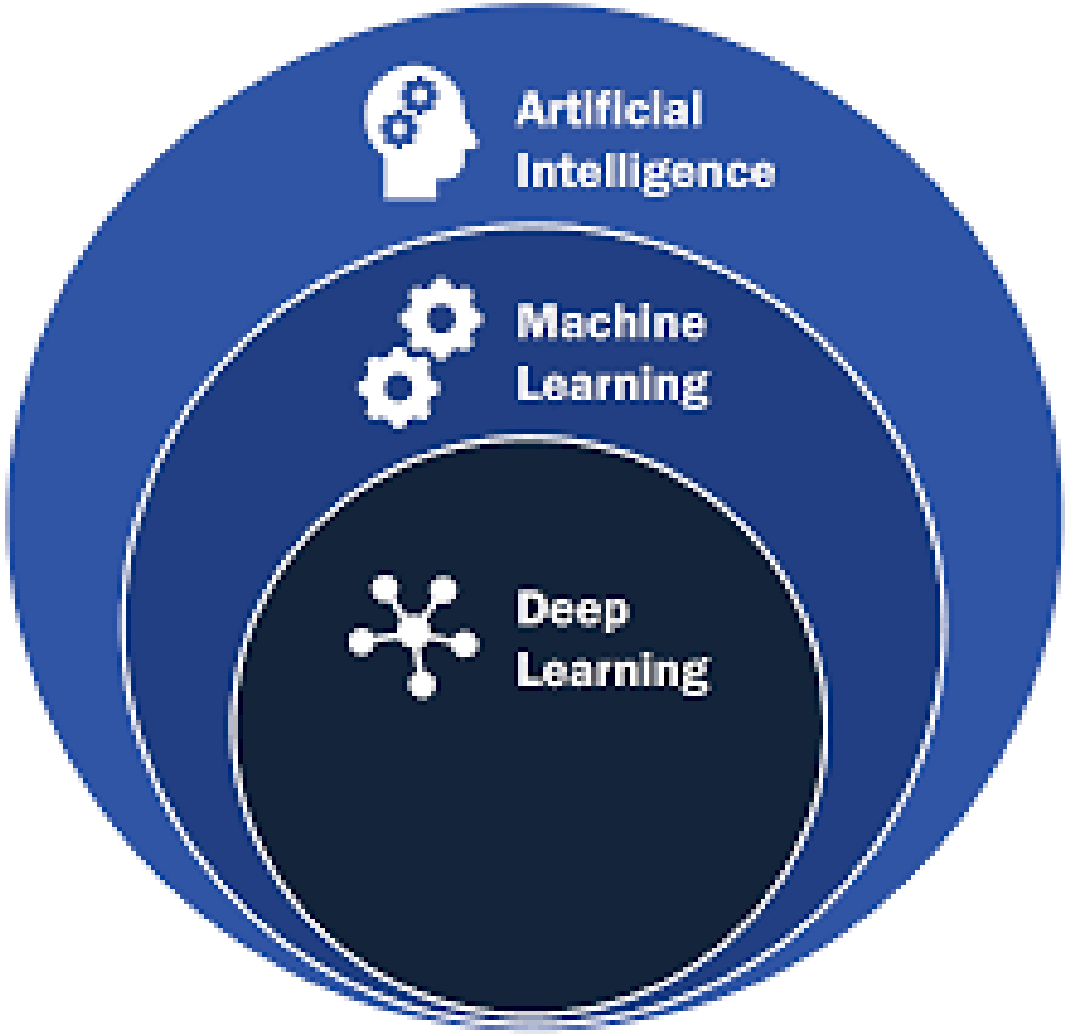
PHD student

Department of Bioinformatics, I.B.B (Institute of Biochemistry & Biophysics)

University of Tehran

Outline

- What is Artificial intelligence?
  What is Machine Learning?
- Deep neural network
- Tools
- Deep neural network for gene regulation

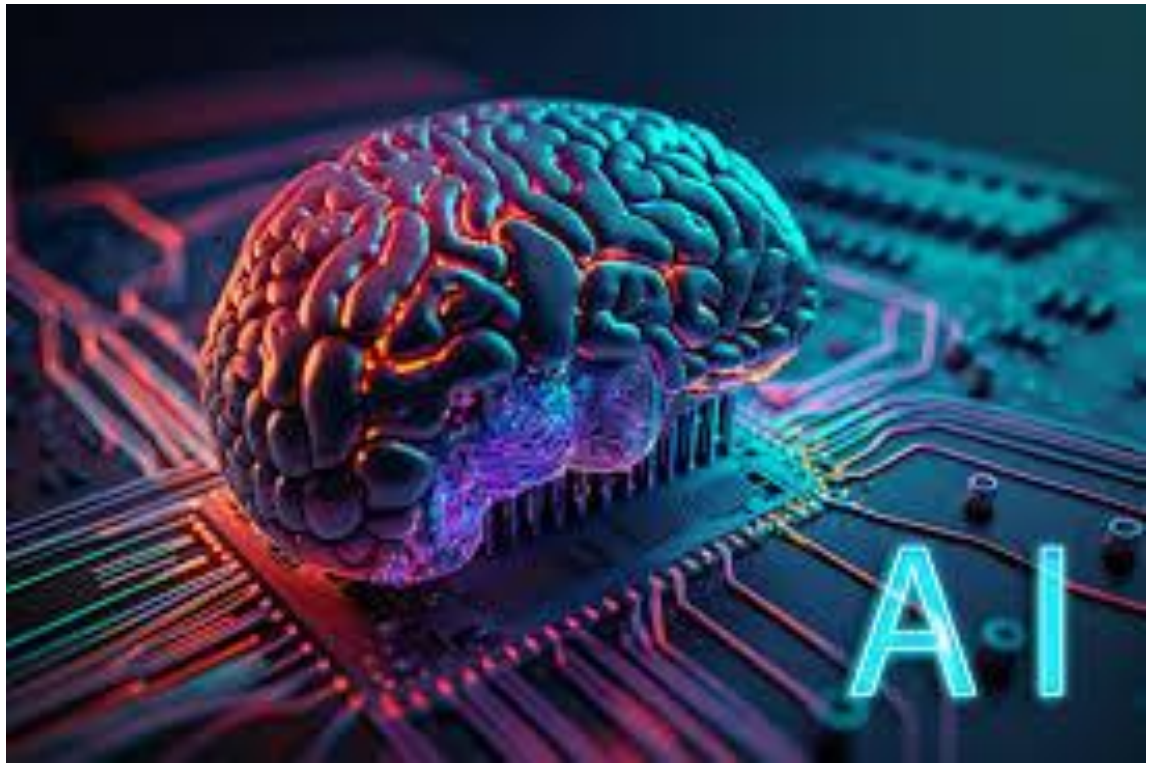


Artificial intelligence
&
Machine Learning

Artificial intelligence

Artificial Intelligence (AI): The concept of artificial intelligence dates back to the 1950s when Alan Turing and others introduced early ideas about the ability of machines to learn.

The theory and development of computer systems capable of performing tasks that historically required human intelligence, such as recognizing speech, making decisions, and identifying patterns.

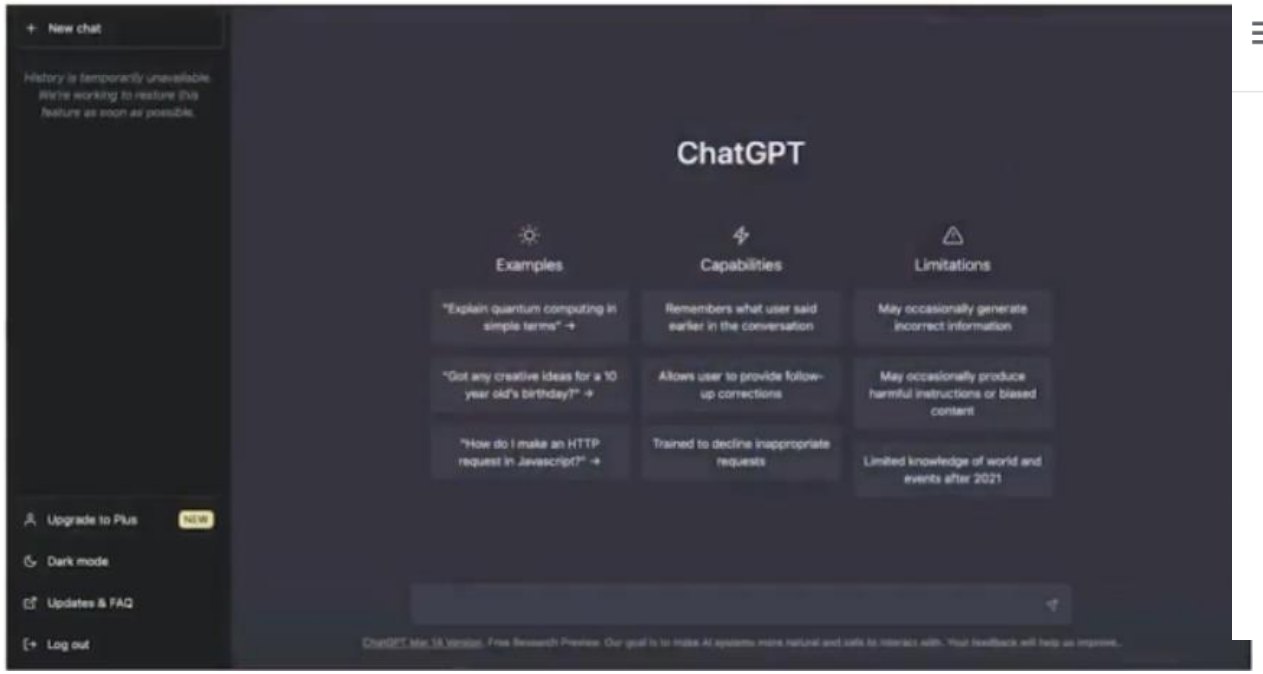


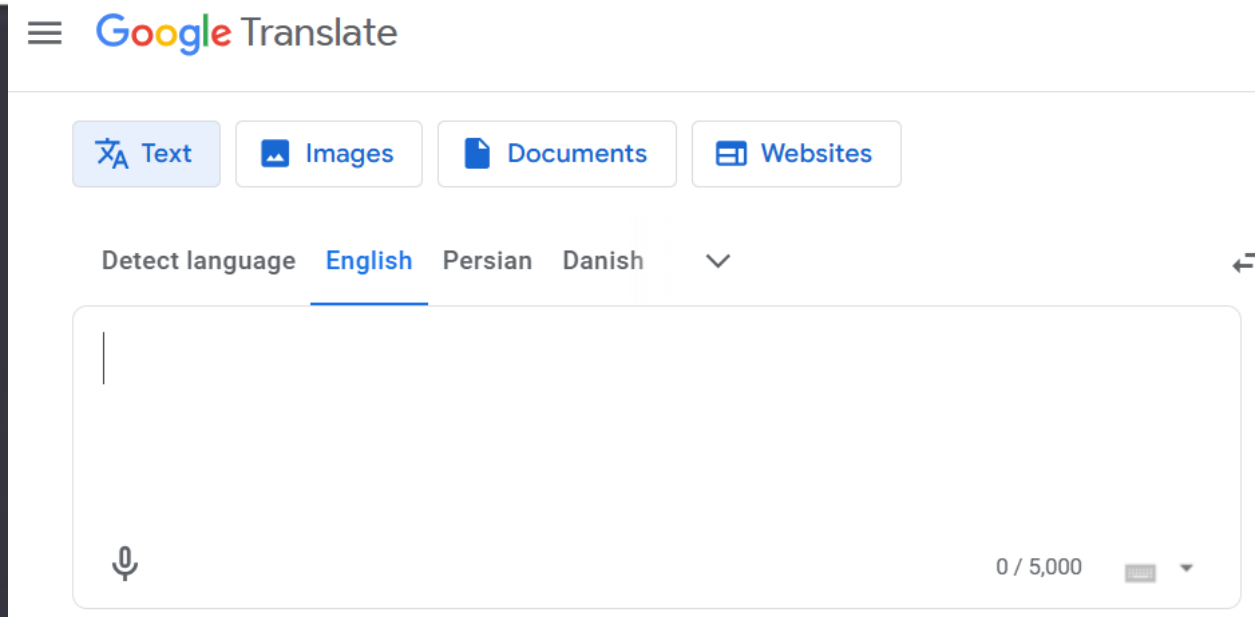
Common examples of AI

Natural language processing(NLP)

ChatGPT: Uses large language models (LLMs) to understand and generate text in response to questions. Large language models, also known as LLMs, are very large deep learning models that are pre-trained on vast amounts of data.

Google Translate: Uses deep learning algorithms to translate text from one language to another.





Computer Vision

Enable machines to see and interpret images

What is Machine Learning?

Machine Learning: It advanced during the 1980s and 1990s with the development of statistical algorithms and self-learning systems.

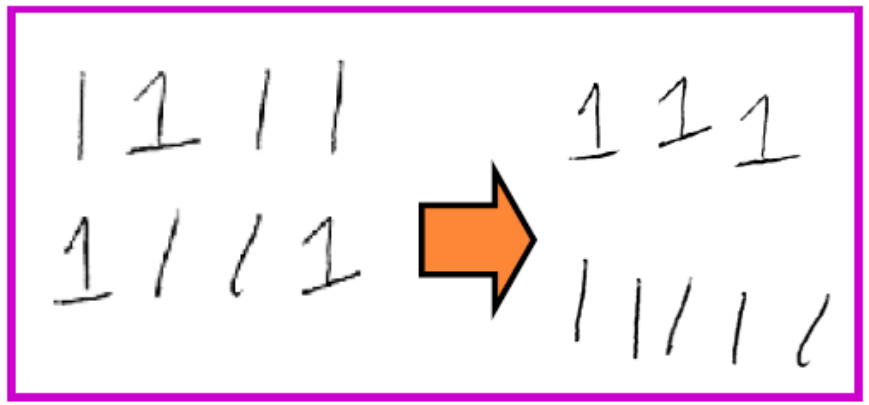
Tom M. Mitchell's definition of machine learning:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Defining the Learning Task

Improve on task T, with respect to
performance metric P, based on experience E

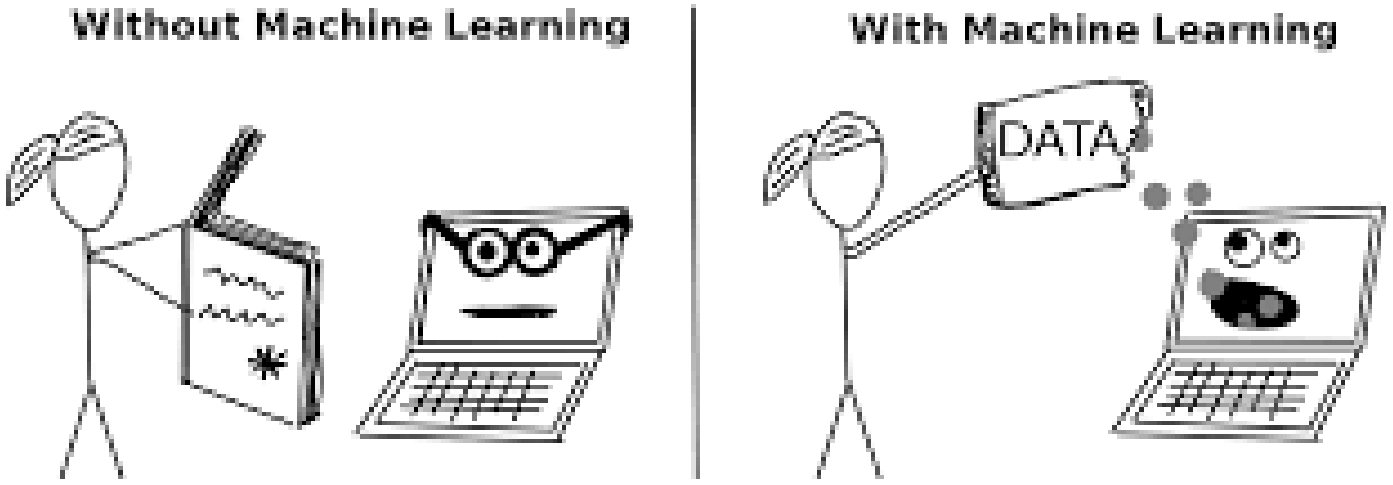
T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words



T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

What is Machine Learning?

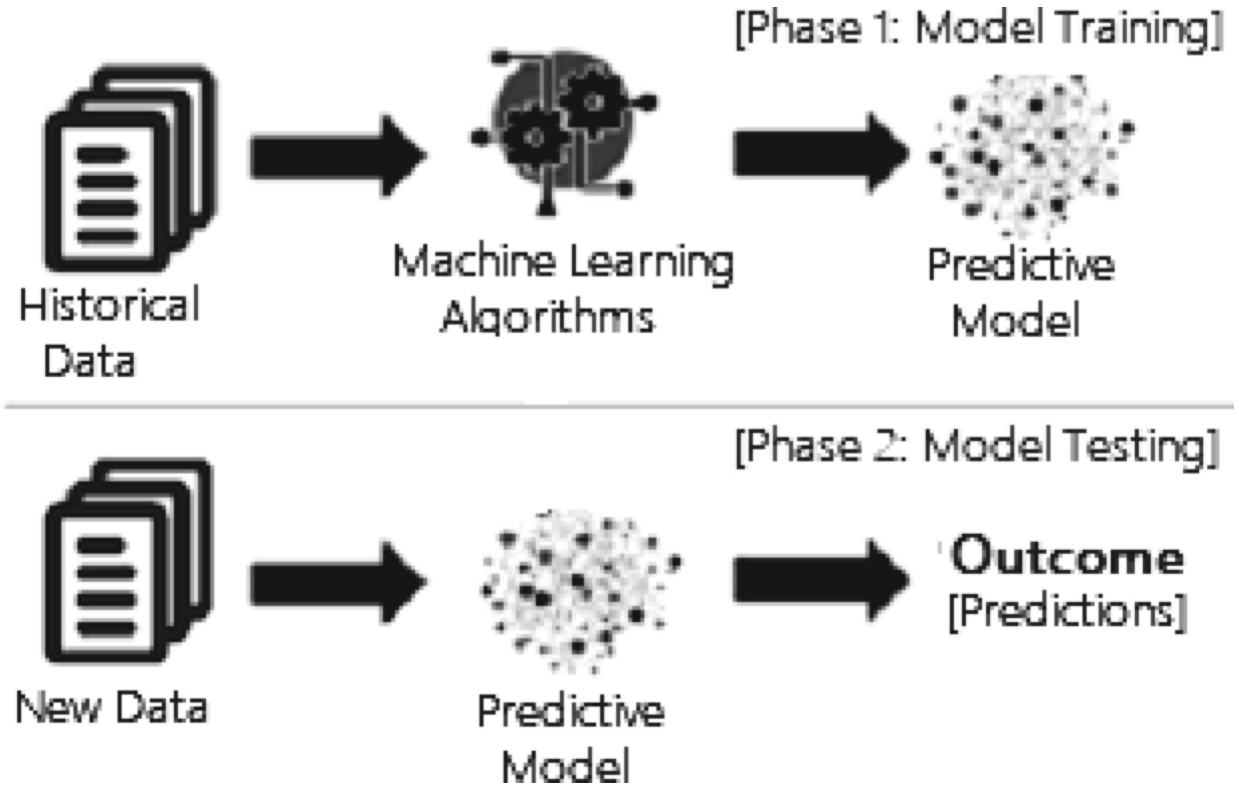
"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." -Arthur Samuel (1959)



What is Machine Learning?

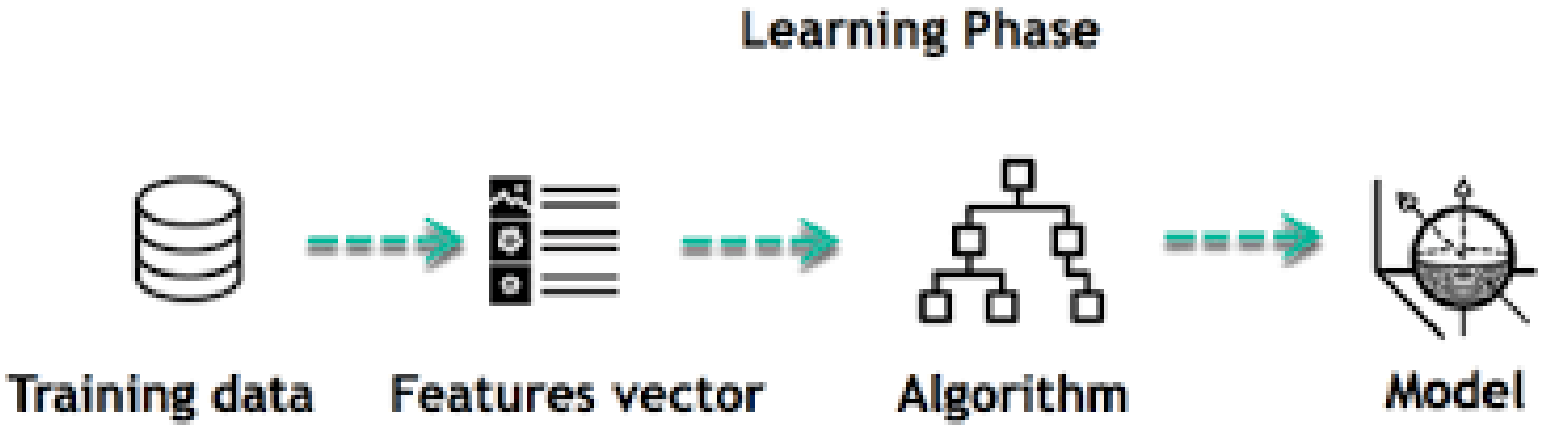
Machine learning constructs algorithms that can generalize patterns from data.

These algorithms analyze input data and identify underlying rules that can be applied to new, unseen data.



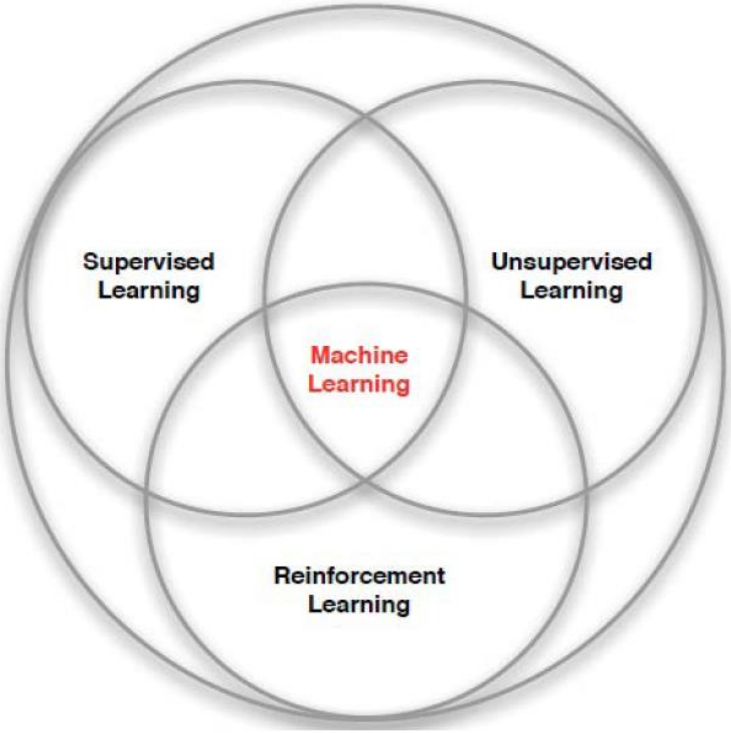
Examples

- Classification: Categorizing data into predefined groups, like spam email detection.
- Uncovering hidden structures: Finding clusters in data or relationships



Types of Learning

- **Supervised (inductive) learning (Regression, Classification)**
  - Given: training data + desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Reinforcement learning**
  - Rewards from sequence of actions



Supervised Learning

- **Data have Label**
- **Label represent data**
- Learn to predict outcomes using **Labeled data**

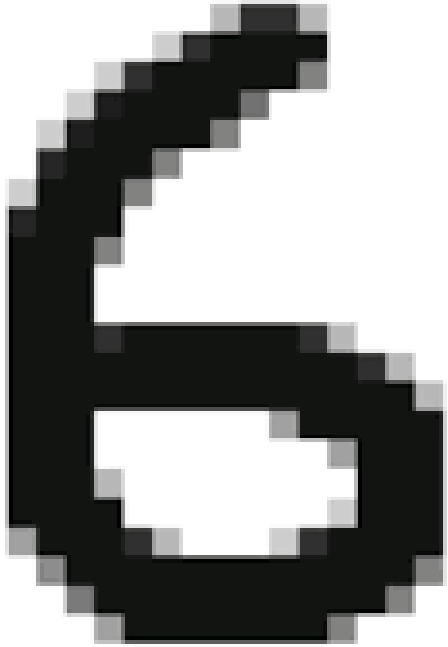
Label: cat

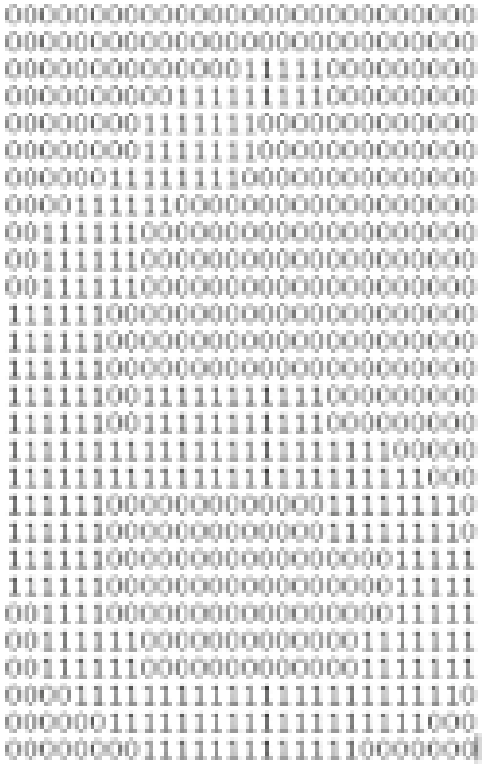
label: dog

Supervised Learning

- **Definition**: A form of machine learning where the model learns from labeled data $\{(x_i, y_i)\}$ to predict an output y given an input x.
  - Input: X1,X2,X3,…Xn
  - Output: y
  - Goal:

$$y = f(x) + \epsilon$$



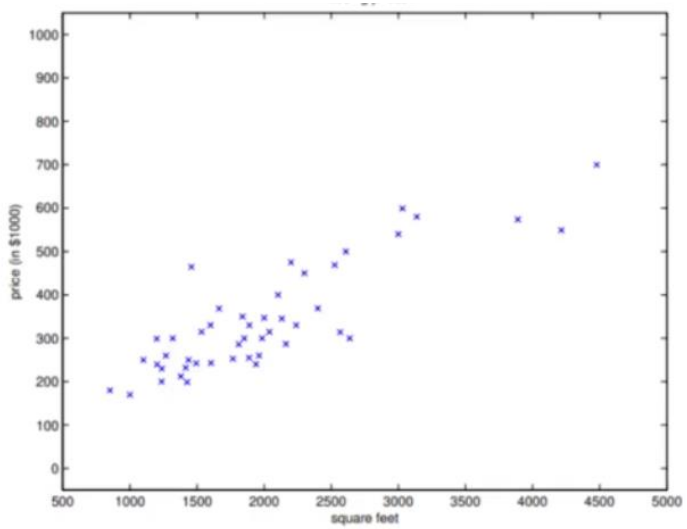


Example: Predict House prices by Linear Regression

Predict House prices from Data

Features like, size, location, number of rooms

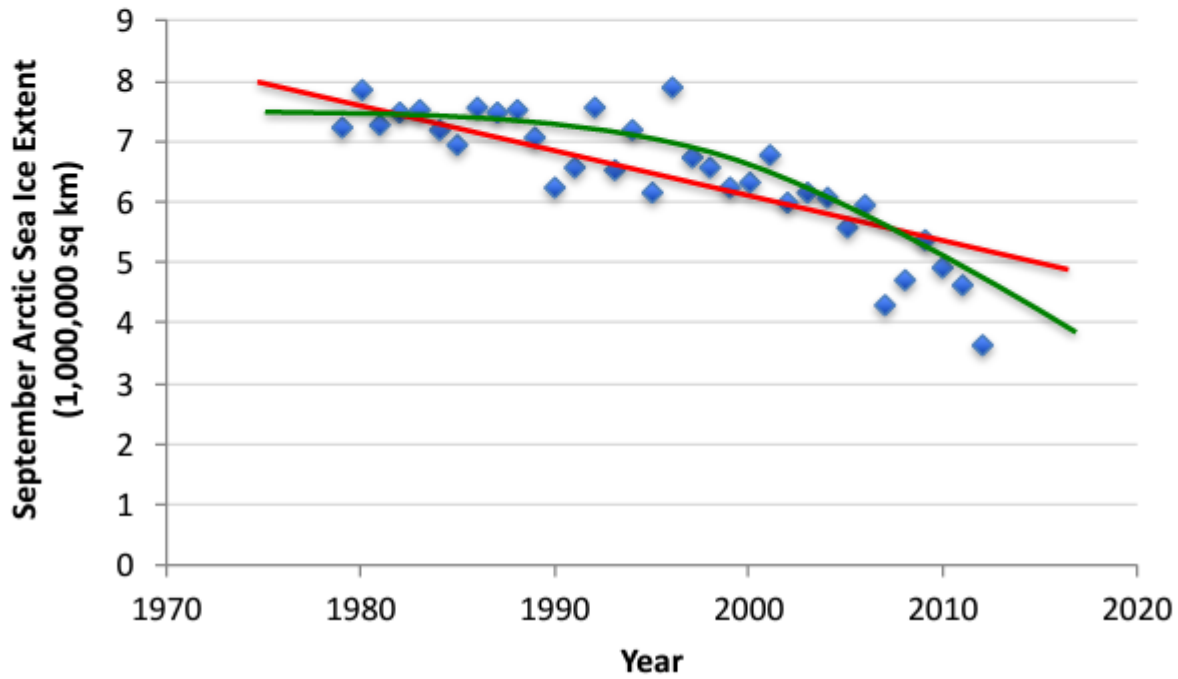
- Input: X1,X2,X3,…Xn
- Output: y
- Goal: $y = f(x) + \epsilon$



Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

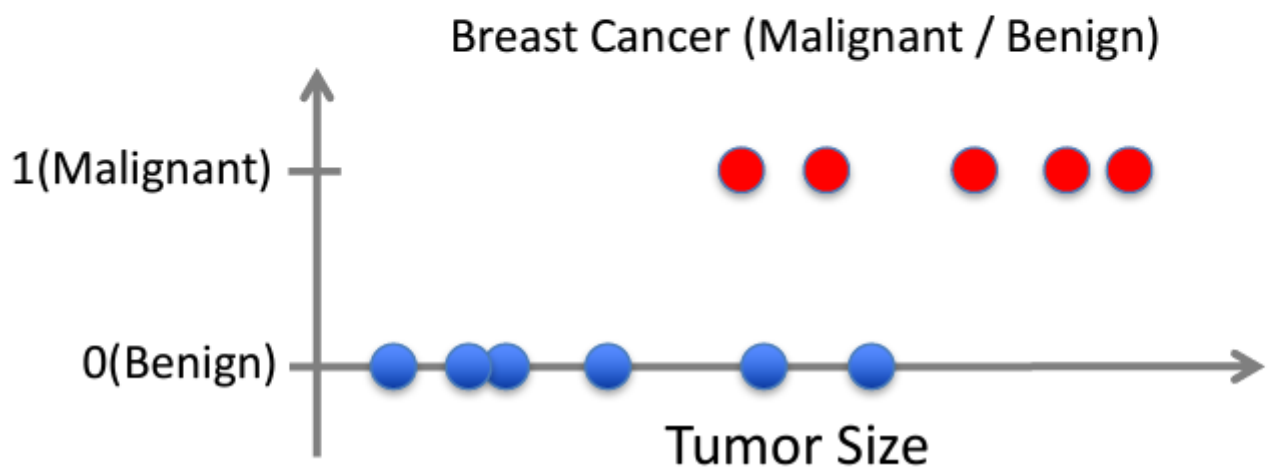
Supervised Learning: Regression

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Supervised Learning: Classification

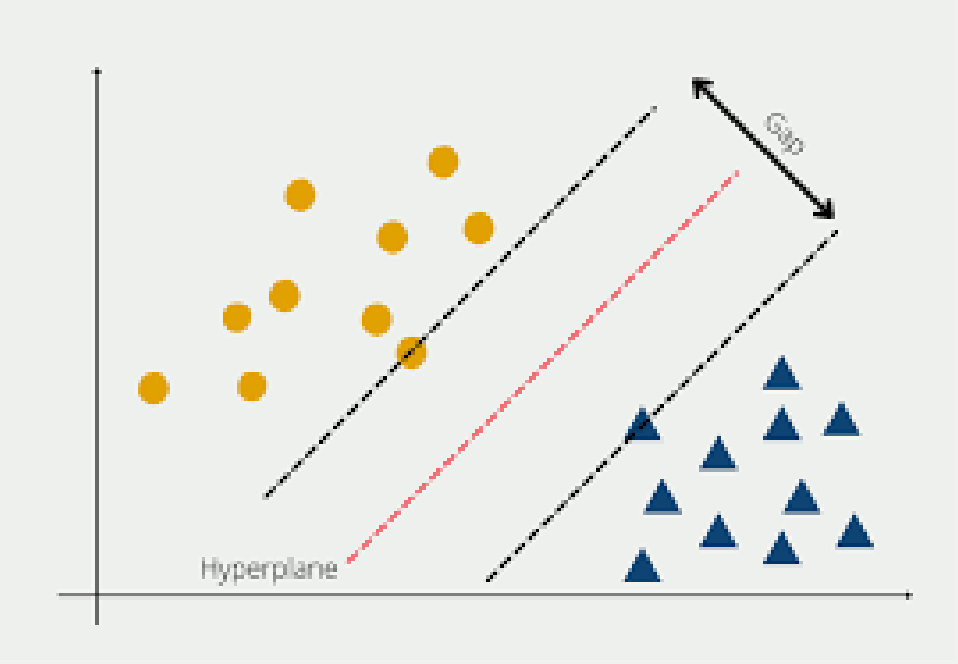
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Support Vector Machine (SVM)

Machine learning algorithm for classification.

The goal of SVM is to find a "maximum margin" that separates the classes.



Evaluate method

How to measure error

- **Linear Regression Hypothesis:**

$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

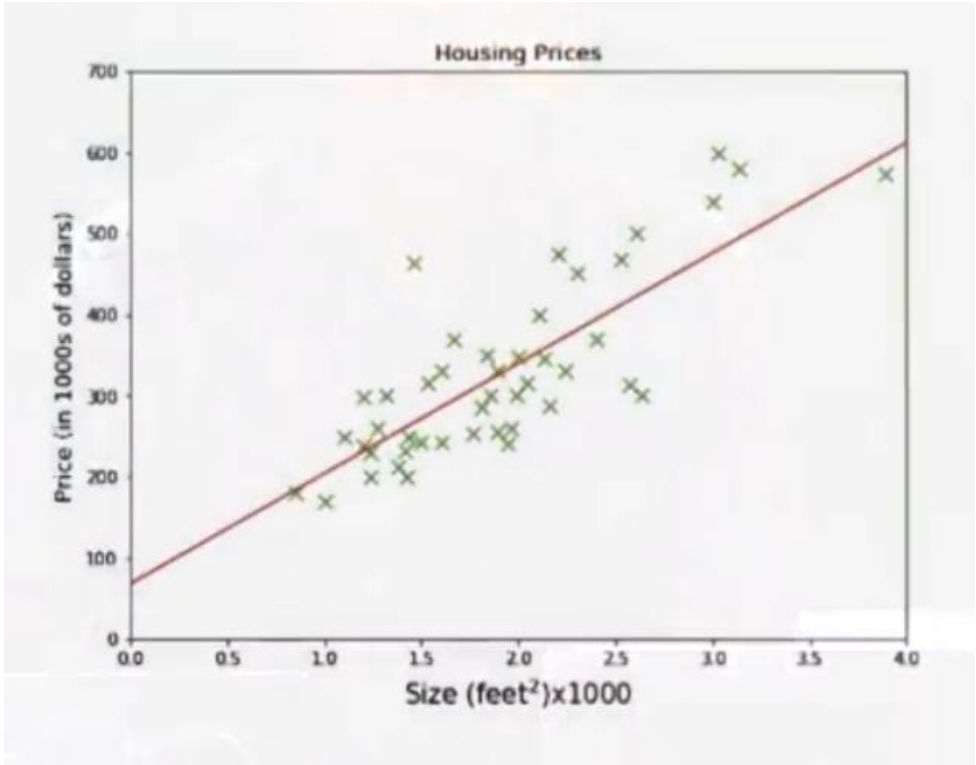
- **Input Vector x:**

$$\mathbf{x} = [x_0 = 1, x_1, x_2, \ldots, x_D]$$

- **Parameter Vector w:**

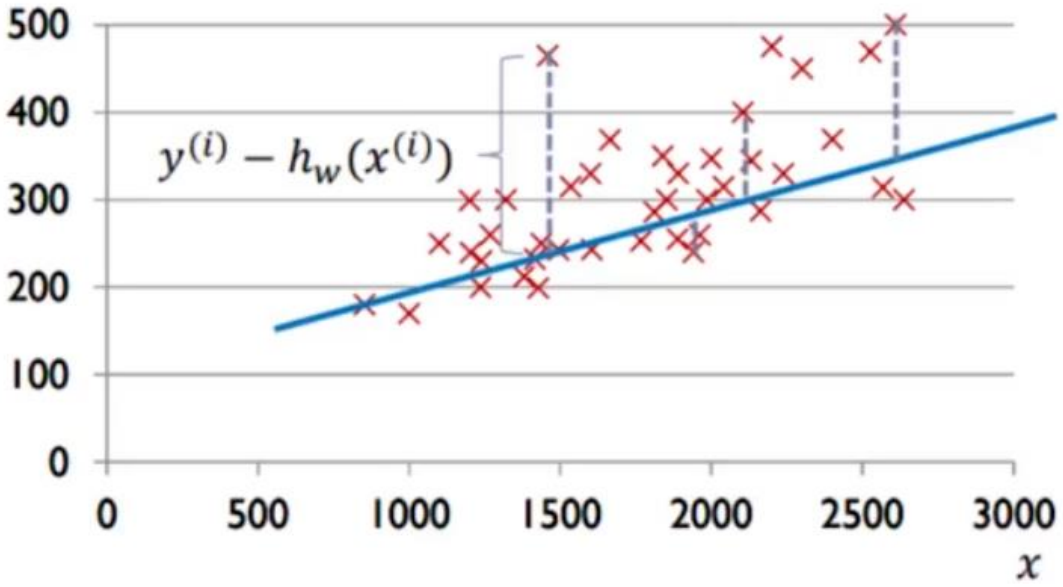
$$\mathbf{w} = [w_0, w_1, w_2, \ldots, w_D]$$

$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + \sum_{i=1}^{D} w_i x_i$$



Mean Squared Error -cost function

Mean Squared Error is the sum of the squared differences between the prediction and true value.



$$J(w) = \sum_{i=1}^{n} \left(y^{(i)} - h_w(x^{(i)}) \right)^2$$

$$= \sum_{i=1}^{n} \left(y^{(i)} - w_0 - w_1 x^{(i)} \right)^2$$

Unsupervised Learning

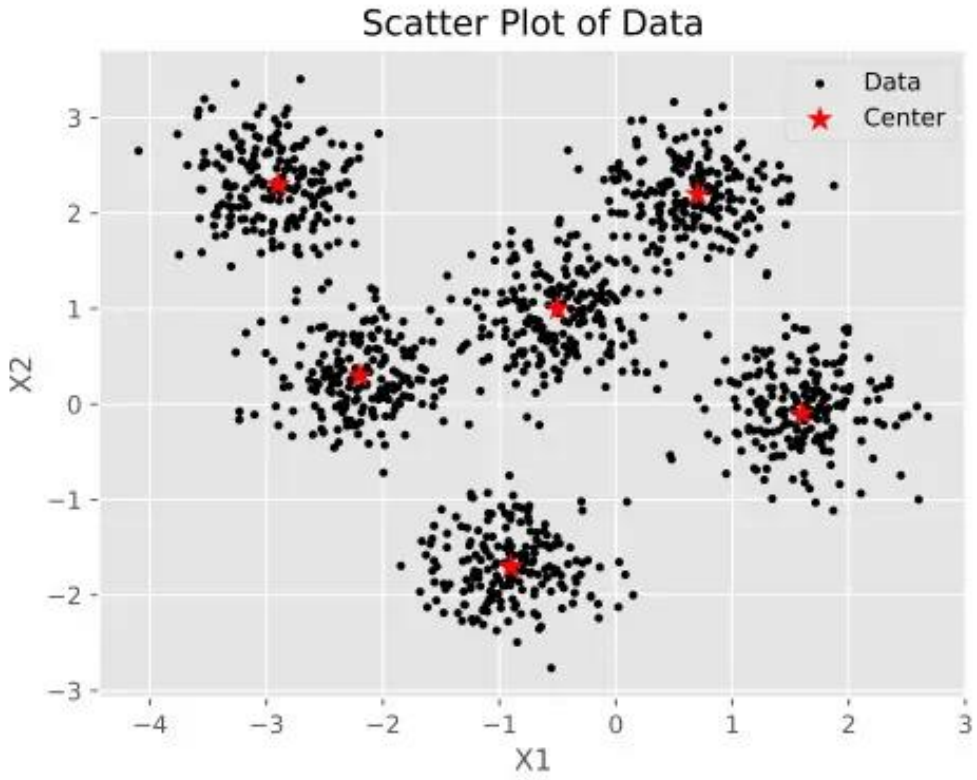
Learn to predict outcomes using **unlabeled data**

Clustering

Example: Customer segmentation, news clustering.

K-means

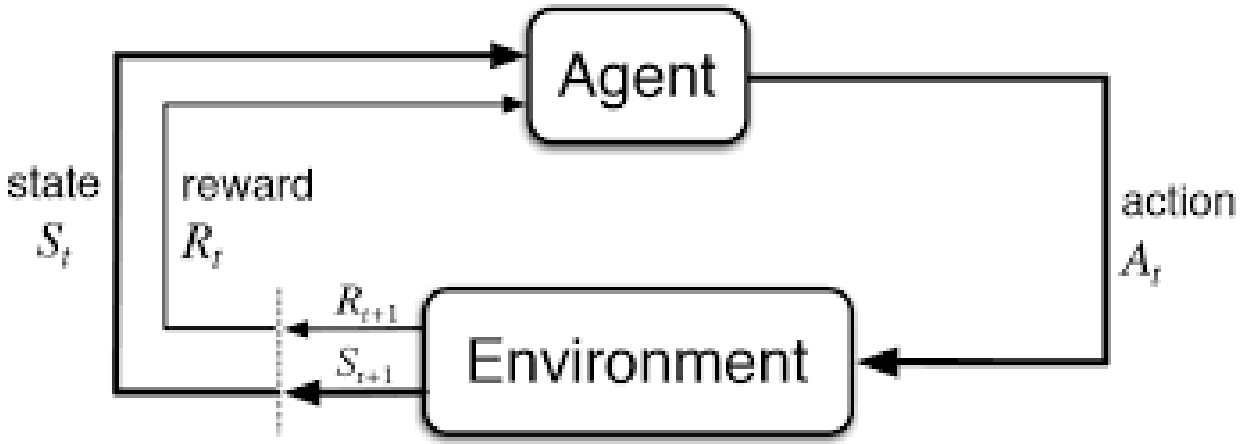
- This algorithm divides data into k clusters (k is determined by the user).
- The process involves randomly selecting k points as the initial cluster centers.
- Then, data points are assigned to clusters based on the shortest distance to these centers.
- In subsequent steps, the cluster centers are updated based on the average of the data points within each cluster.
- This process continues until the centers no longer change.



Reinforcement Learning

- Reinforcement learning (RL) is a ML technique that trains software to make decisions to achieve the most optimal results.

- A type of ML where an agent learns to behave in an environment by performing actions and seeing the results.

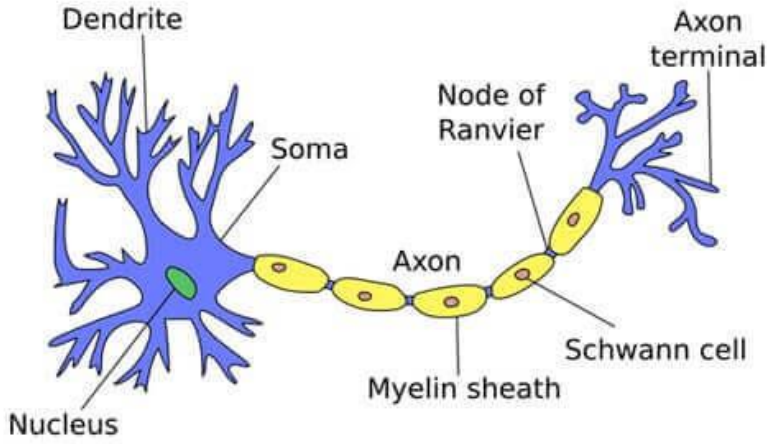


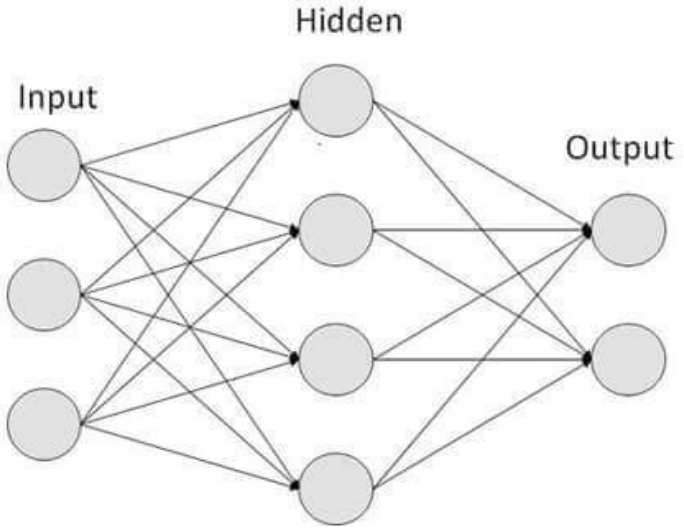
Neural Network

Mimic human brain to solve complex tasks

A **neural network** is a computational model inspired by the structure and function of the human brain. It consists of layers of interconnected nodes (or "neurons") that process data by simulating the way biological neurons transmit signals.

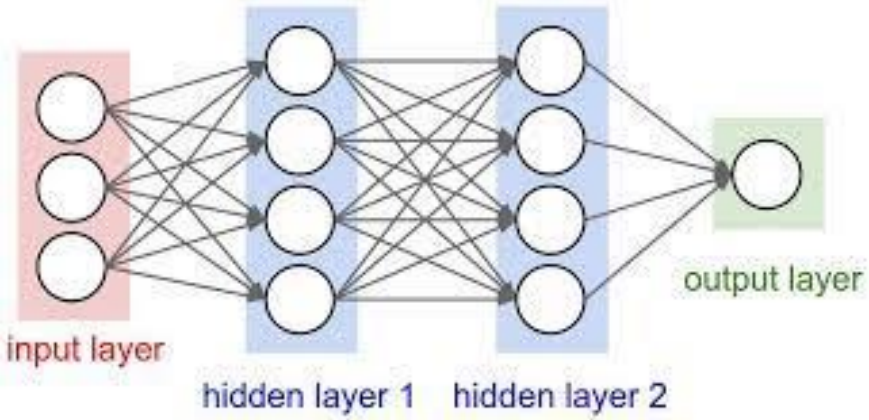
- Input Layer-Hidden Layers- Output Layer



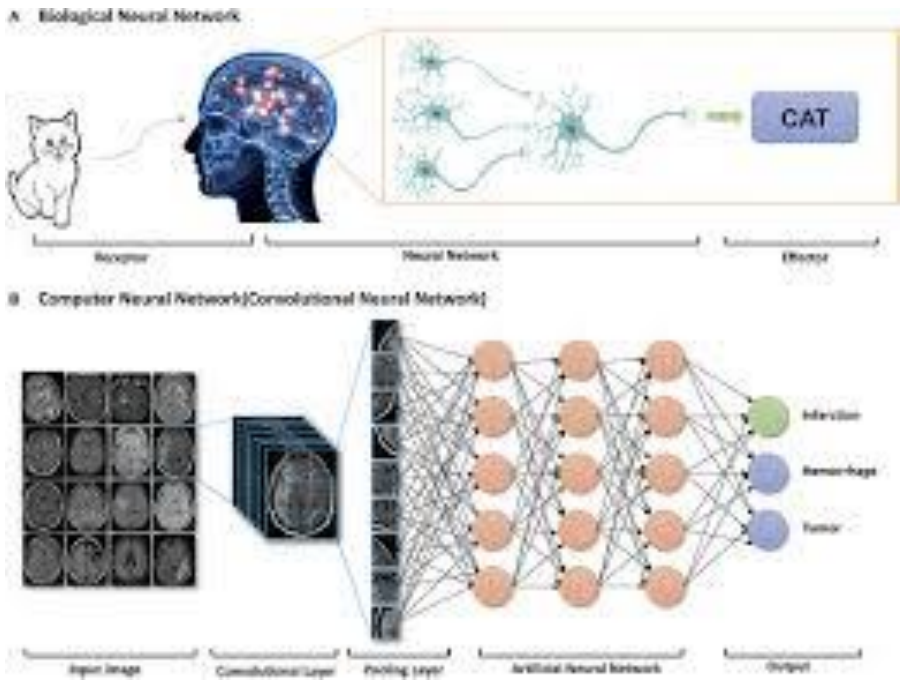


Deep-learning

Deep Learning: The progress of deep learning began in the early 2010s, when neural networks achieved high accuracy in solving complex problems such as facial recognition and image analysis.



A class of neural network models that uses multiple layers to extract complex patterns from data.



Deep learning requires large amounts of data and high computational power and is currently highly effective in applications such as face recognition, image detection, and natural language processing.

Deep-learning architectures

- Multi-layer perceptron
- Convolutional neural network
- Recurrent neural network
- Bi-LSTM
- Transformer-Attention
- Autoencoder

Tools

Scikit-Learn: free and open-source machine learning library

classification, regression and clustering algorithms including support-vector machines, k-means

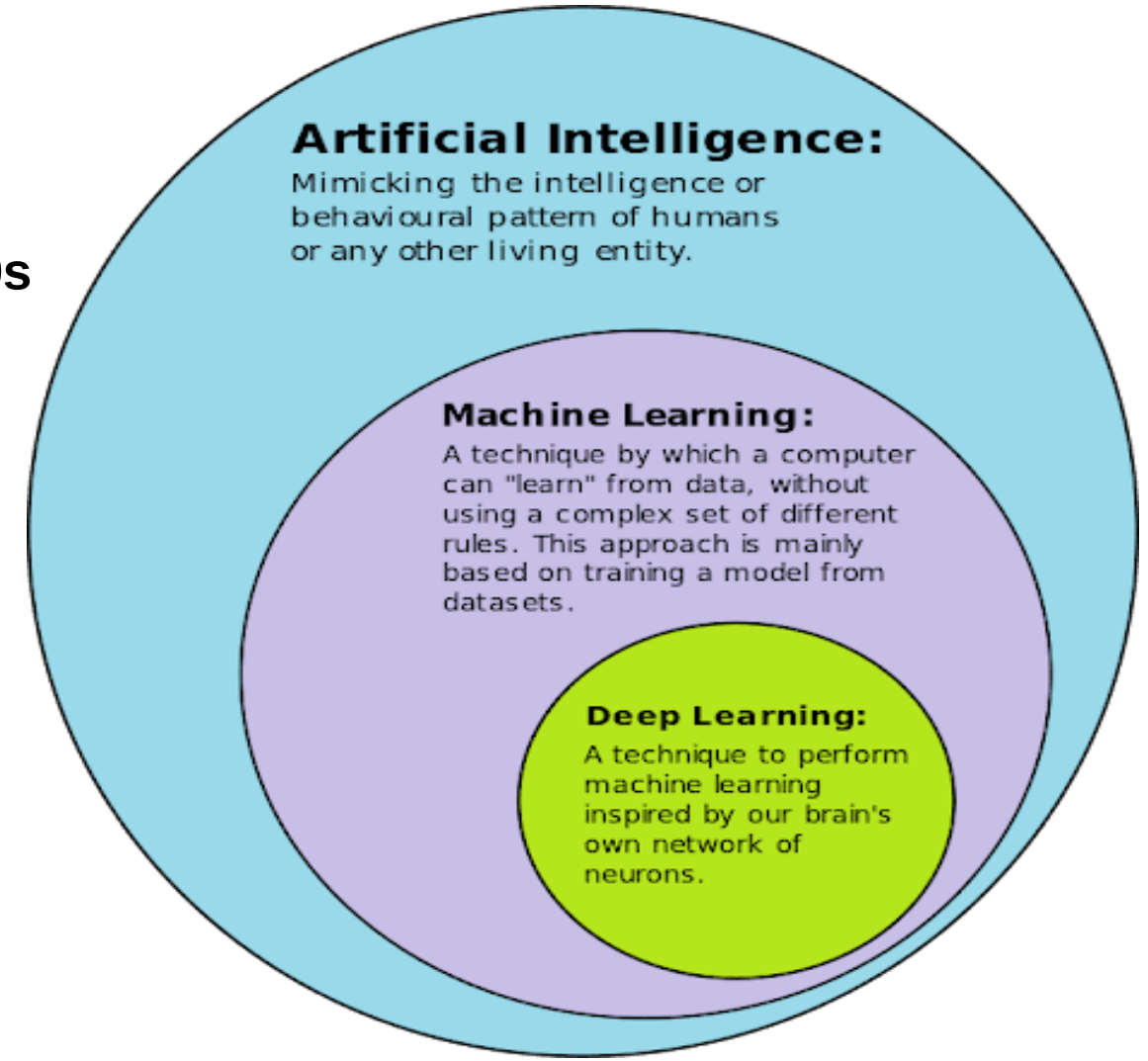
TensorFlow and PyTorch: Deep learning library for construct and train Deep Neural networks

What Is Differences between Artificial intelligence, Machine Learning & Deep Learning?

Artificial Intelligence: 1950s

Machine Learning: Developed in the 1980s and 1990s

Deep Learning: Advances in deep learning began in the early 2010s



Limitations

Need for Large Data: Deep learning requires vast and diverse datasets to deliver accurate results.

Model Complexity: Deep learning models are complex, requiring powerful computational resources and advanced technical skills.

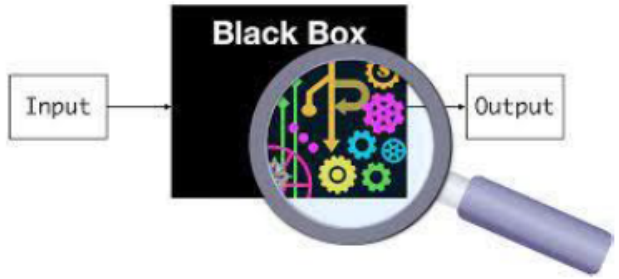
Limitations in Specific Conditions: AI may perform poorly in unusual scenarios or with incorrect data, which can be risky in critical fields like medicine.

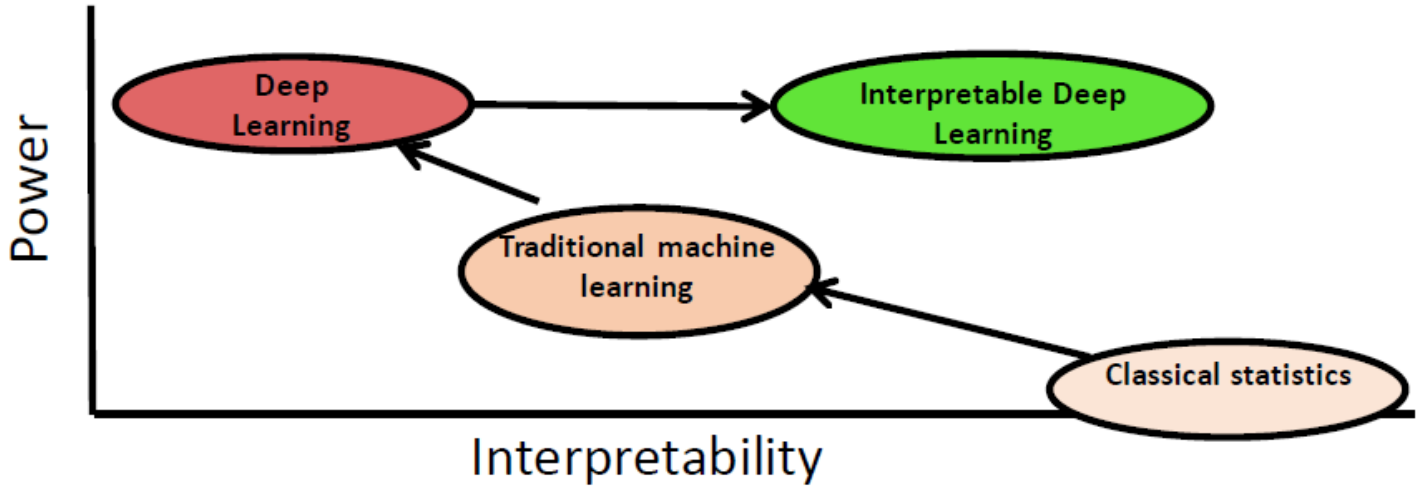
Limited Interpretability: Some deep learning models, such as neural networks, are difficult to interpret, making it challenging to understand their decision-making process.

Deep Learning & Poor interpretability

Black box of AI

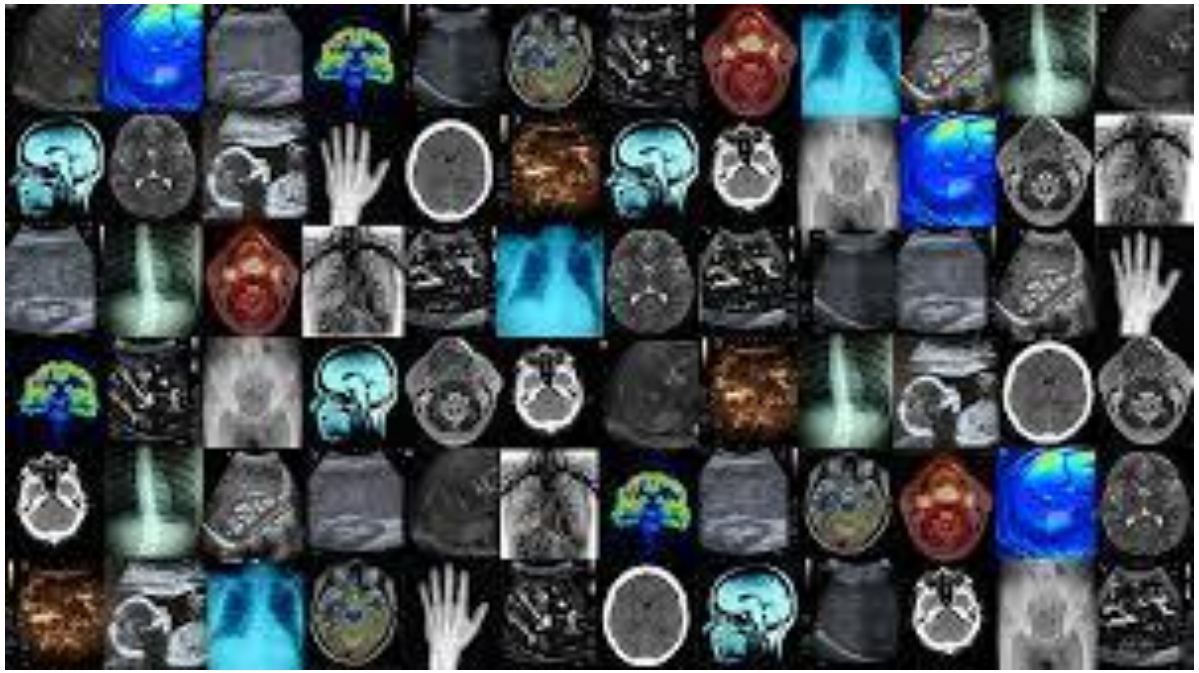
- Need to understand how machines learn.





Applications in Medicine

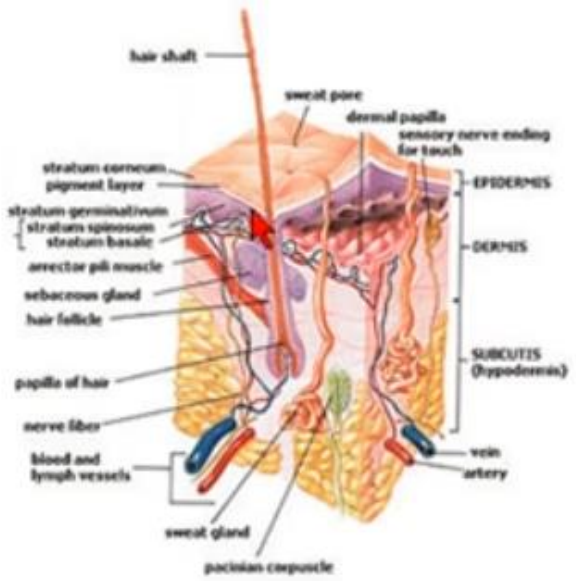
Disease Diagnosis: Machine learning and deep learning algorithms are used for diagnosing diseases such as cancer, diabetes, and heart conditions, helping doctors identify them more quickly.

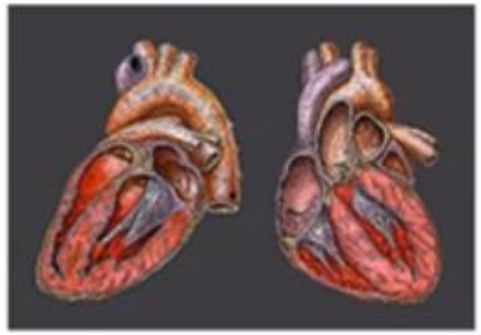


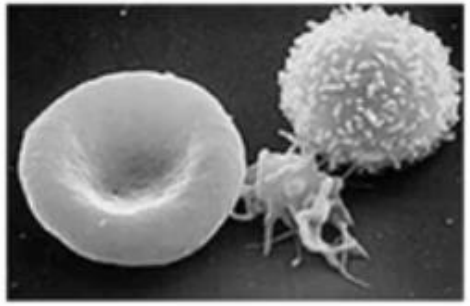
Deep learning for Regulatory genomics

One Genome –Many Cell Types

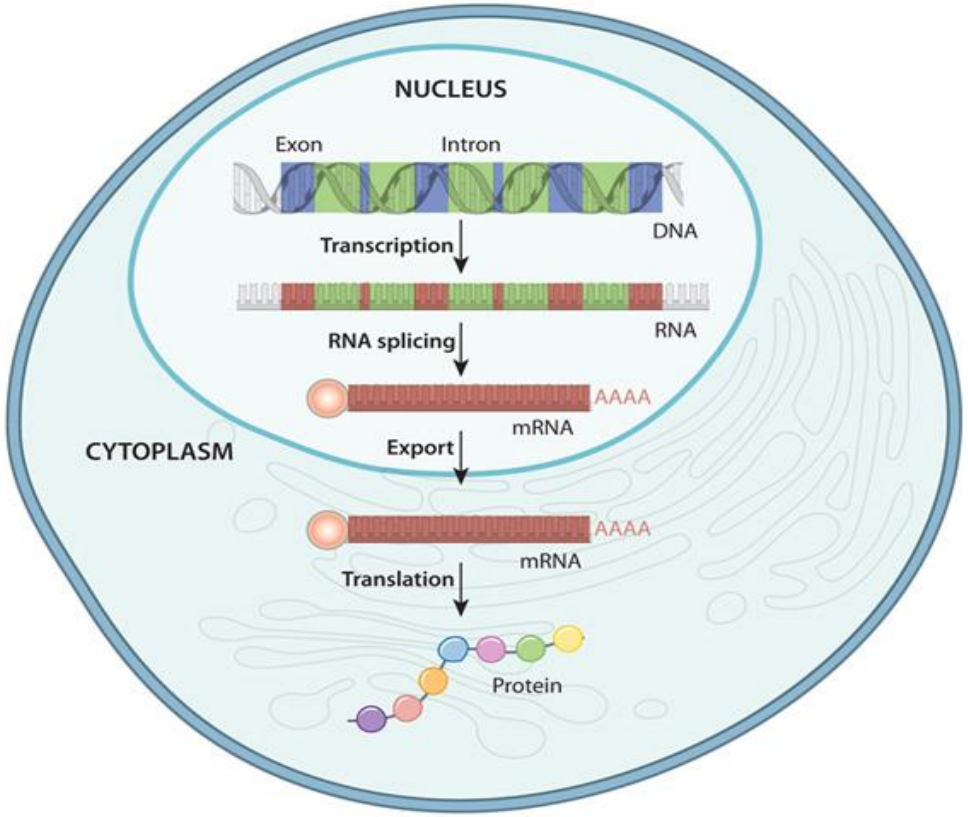
ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

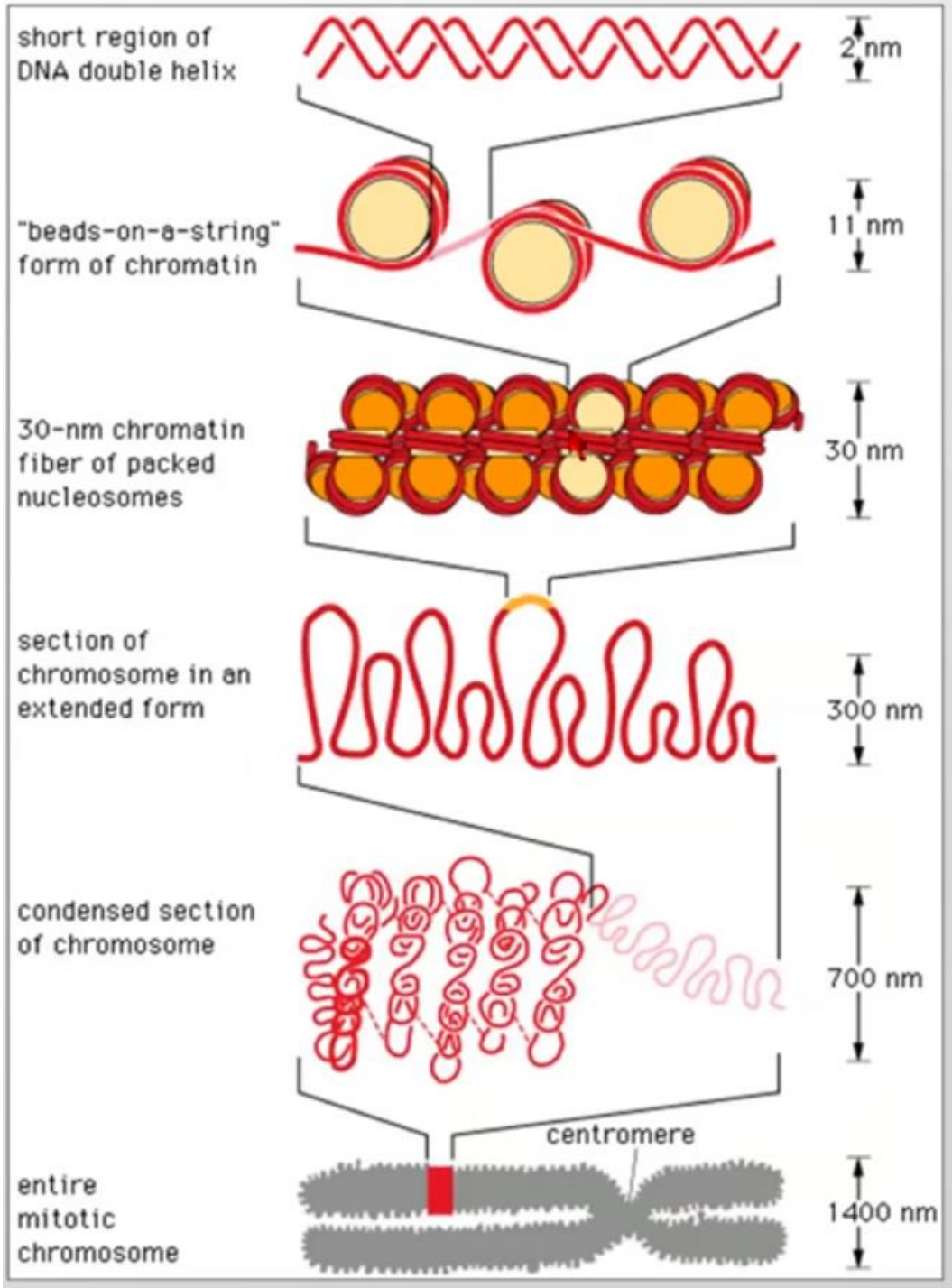






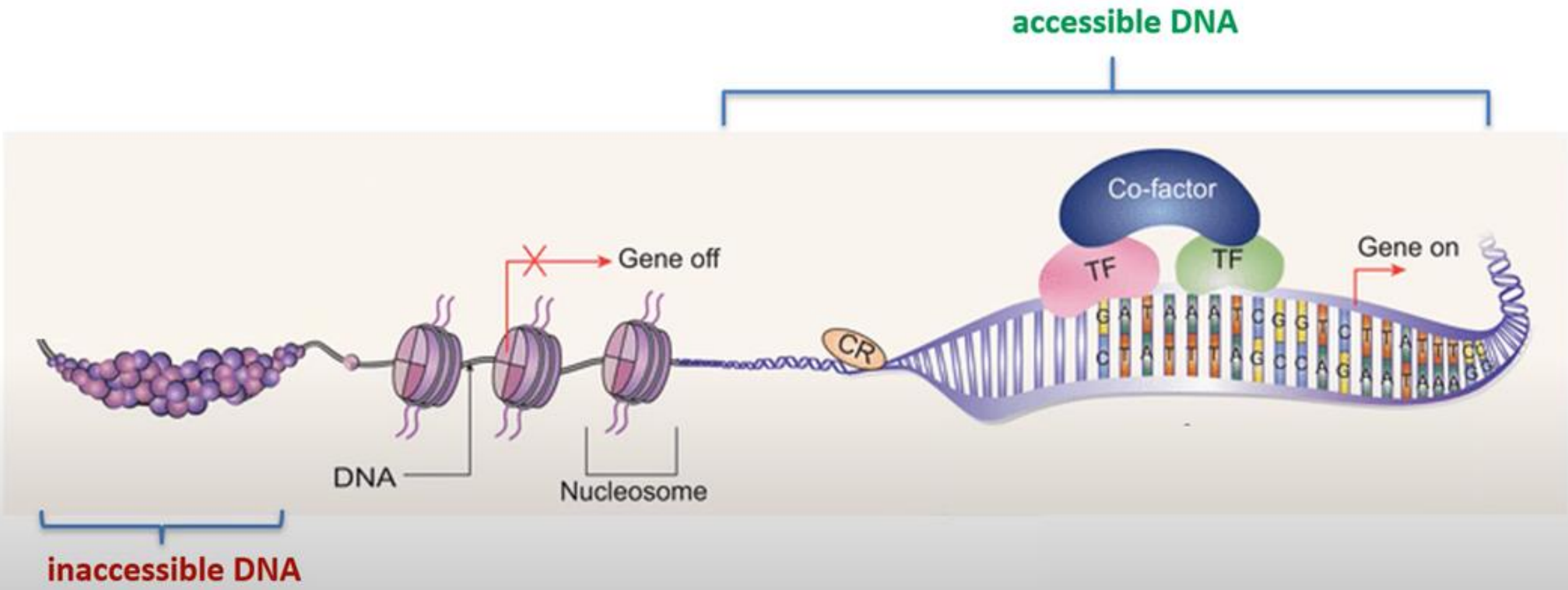
DNA packaging



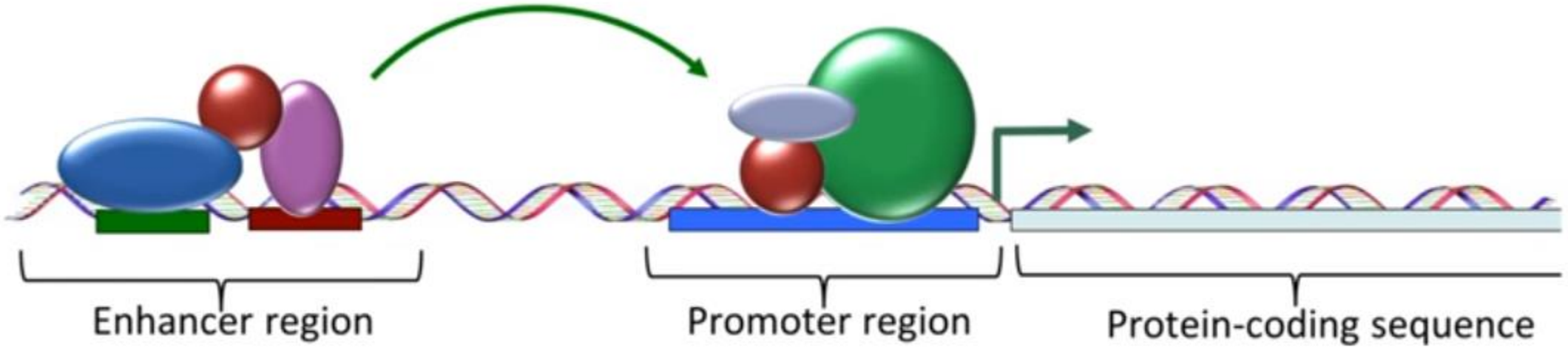


Accessible chromatin and Transcription factor (TF) binding

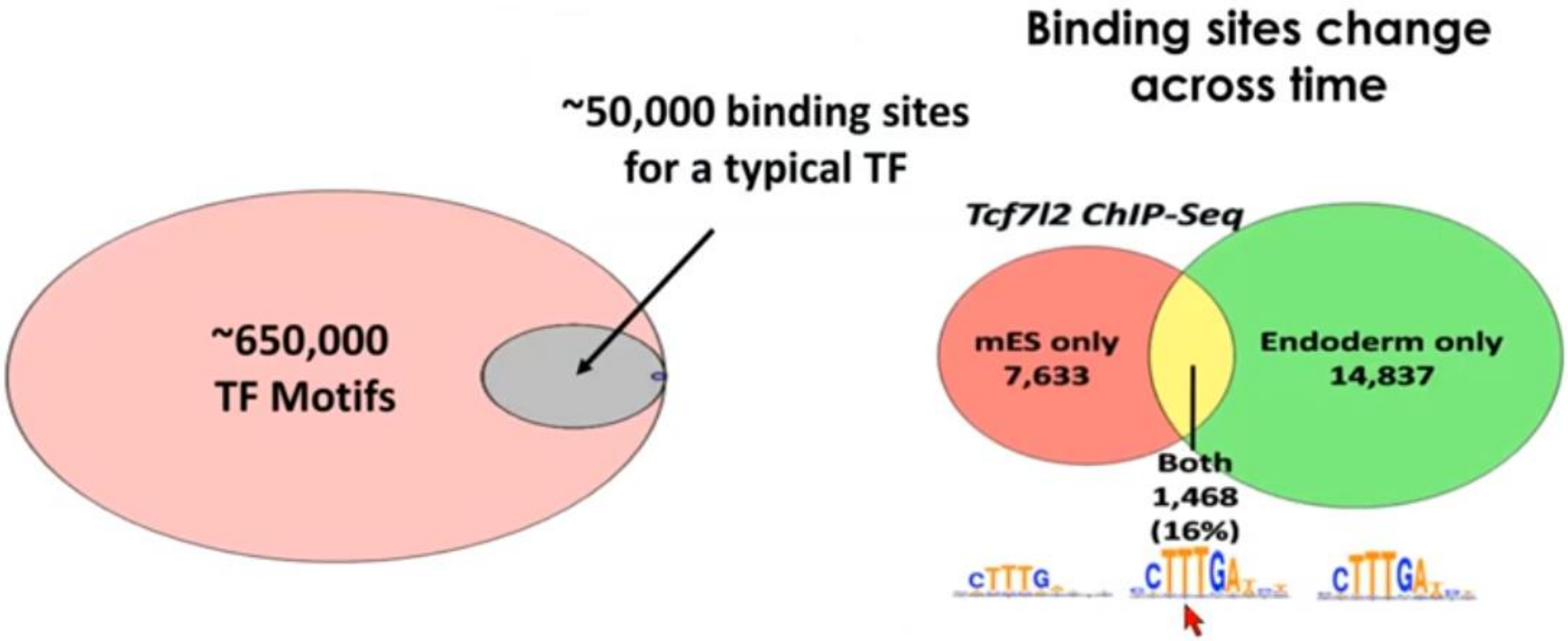
- TFS binds to DNA at transcription factors binding sites (TFBSs)



Transcription factors control activation of cell-type –specific promoters and enhancers

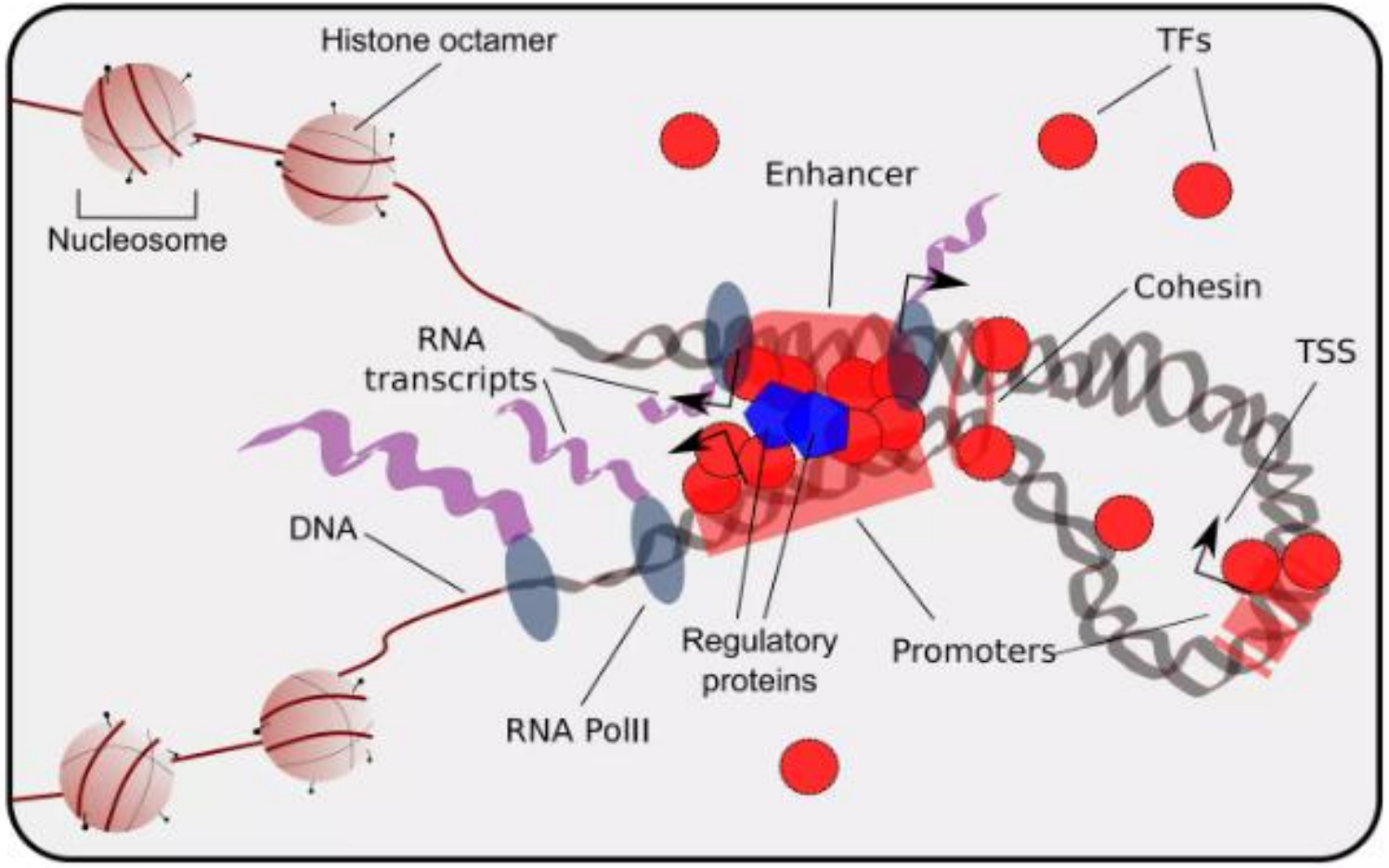


Motifs can predict TF binding



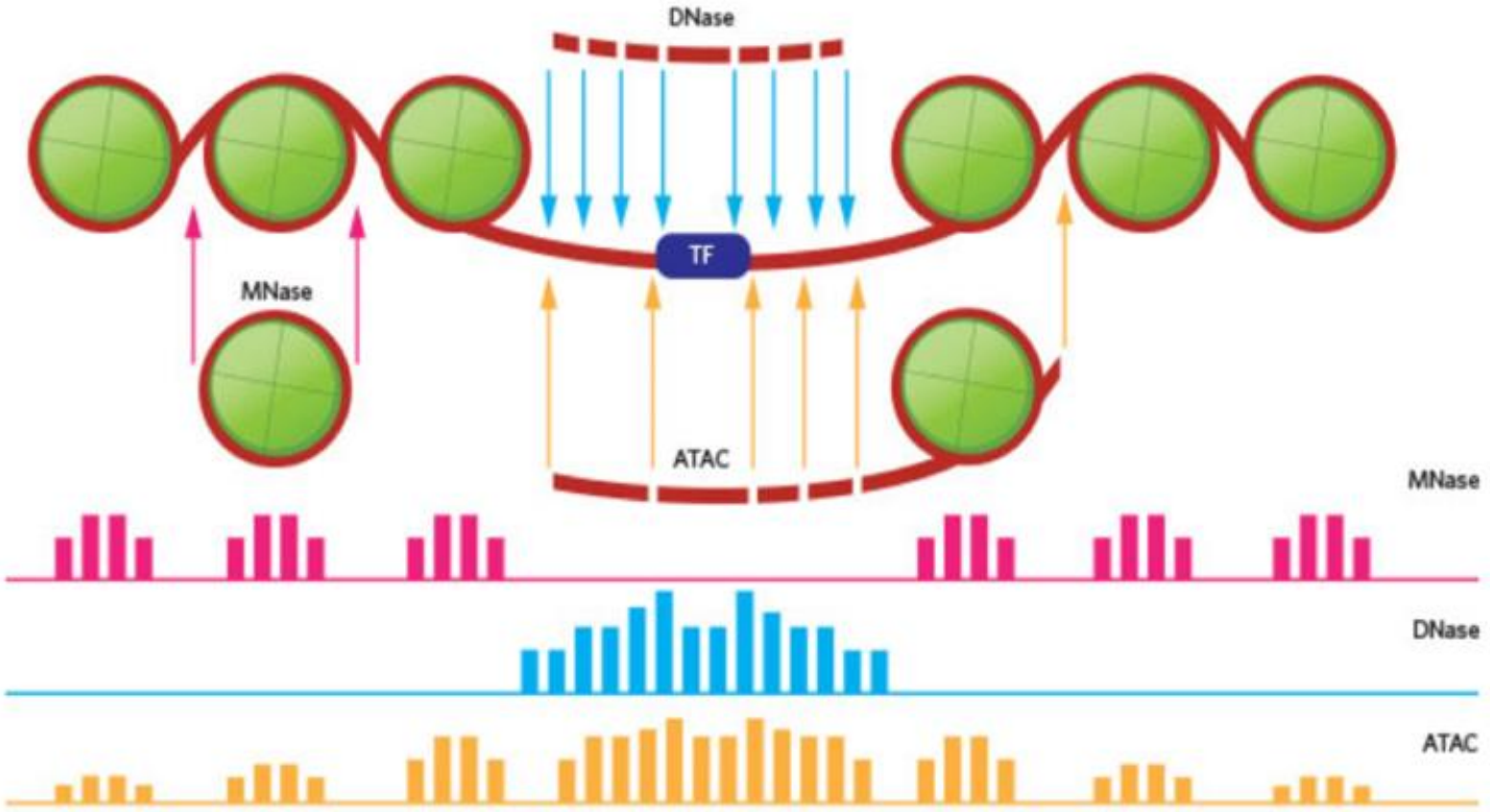
Transcription factor Binding Site

- TFBS are often located in: Gene promoters ,Distal regulatory elements, such as: enhancers, silencers, insulators.



Assays to study Transcription factor binding sites (TFBSs)

- Protein binding microarray (PBM)
- Chip-seq
- ChIP-exo
- DNase-seq
- FAIRE-seq
- MNase-seq
- ATAC-seq



Representing TFBS: Position Weight matrix(PWM)

- PFM is a 4×L matrix.
- Wαm is the probability of seeing nucleotide α at position m.
- PWMs assuming **nucleotide independence** within TFBSs.

Motifs	T	C	G	G	G	G	g	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C
SCORE(*Motifs*)	3	+ 4	+ 0	+ 0	+ 1	+ 1	+ 1	+ 5	+ 2	+ 3	+ 6	+ 4 = 30
COUNT(*Motifs*)												
A:	2	2	0	0	0	0	9	1	1	1	3	0
C:	1	6	0	0	0	0	0	4	1	2	4	6
G:	0	0	10	10	9	9	1	0	0	0	0	0
T:	7	2	0	0	1	1	0	5	8	7	3	4
PROFILE(*Motifs*)												
A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4
CONSENSUS(*Motifs*)	T	C	G	G	G	G	A	T	T	T	C	C

From motif matrix to count matrix to profile matrix to consensus string to motif logo

K-mer

- ***K*-mer** to refer to a substring of length ***k*** in a string
- Define **COUNT(*Text*, *Pattern*)** as the number of times that a ***k-mer*** *Pattern* appears as a substring of ***Text***

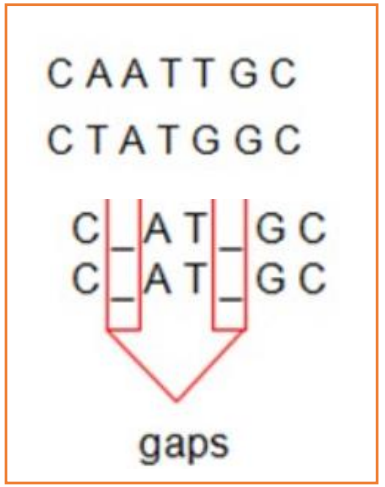
$$\text{COUNT}(\text{ACA}\mathbf{\text{ACTAT}}\text{GCAT}\mathbf{\text{ACTAT}}\text{CGGGA}\mathbf{\text{ACTAT}}\text{CCT}, \mathbf{\text{ACTAT}}) = 3$$

Different parameters for k-mers

Length	Window	Tokenized
3	3	ATC GCG TAC GAT CCG
4	4	ATCG CGTA CGAT
5	5	ATCGC GTACG ATCCG
4	2	ATCG CGCG CGTA TACG CGAT ATCC
4	3	ATCG GCGT TACG GATC

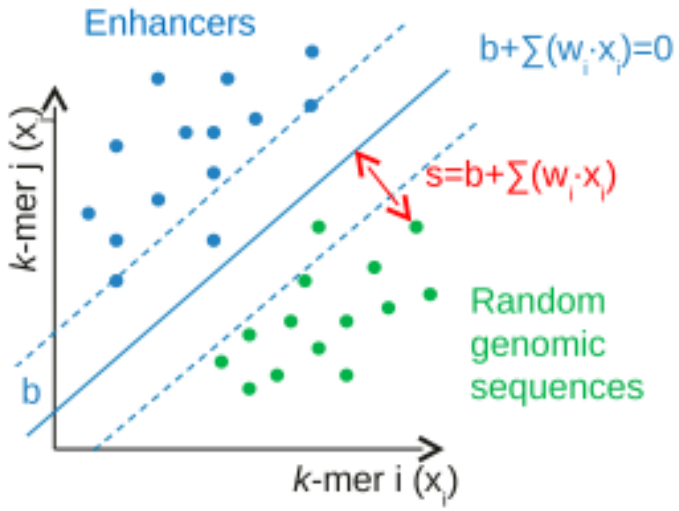
SVM based (kmer-SVM) framework for Enhancer prediction

- The EP300 gene provides instructions for making a protein called p300 (turning on transcription).
- Using the SVM to distinguishes (enhancer) and negative (random genomic) sequence sets.
- Gapped k-mers allow for gaps, providing a more flexible representation of sequence motifs.



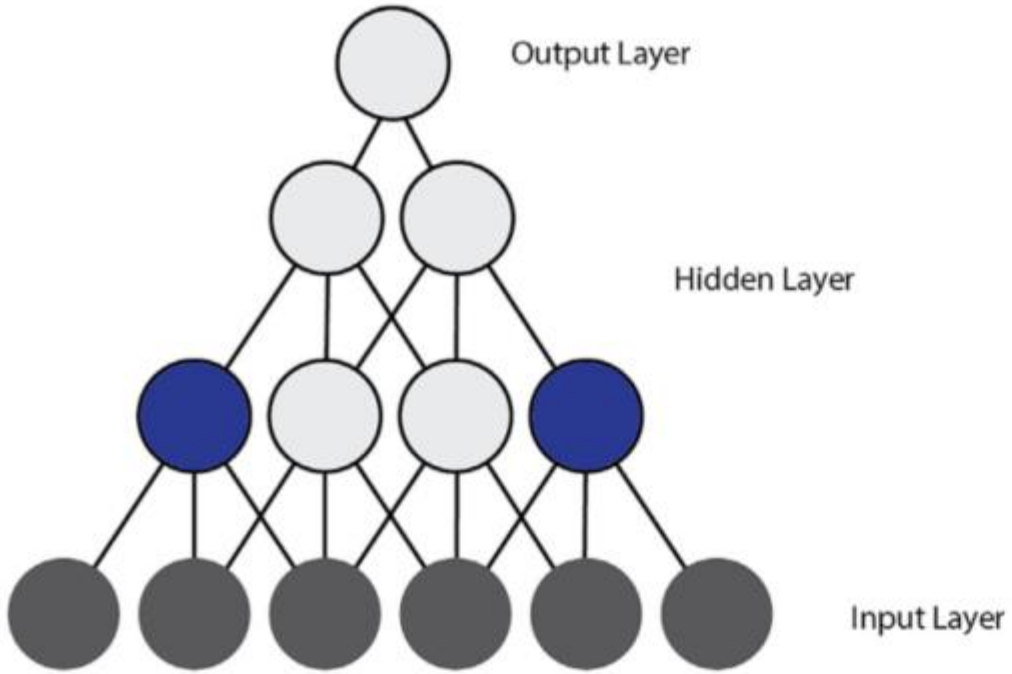
Sequences → *k*-mer frequencies → SVM training → Predictive feature analysis

k-mer	counts
5′-AAAAAA-3′ 3′-TTTTTT-5′	x_1
5′-AAAAAC-3′ 3′-TTTTTG-5′	x_2
5′-AAAAAG-3′ 3′-TTTTTC-5′	x_3
...	...
5′-TTTAAA-3′ 3′-AAATTT-5′	x_n

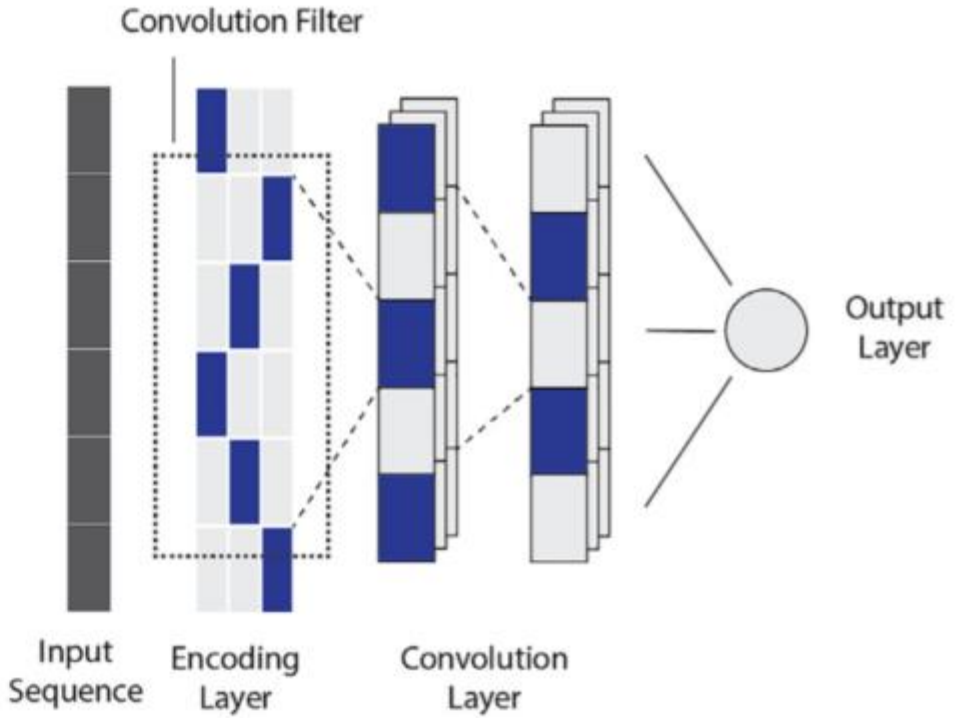


k-mer	weights (w_i)
5′-AATGAG-3′ 3′-TTACTC-5′	+3.94
5′-AATTAG-3′ 3′-TTAATC-5′	+3.84
5′-AGCTGC-3′ 3′-TCGACG-5′	+3.65
...	...
5′-CAGGTA-3′ 3′-GTCCAT-5′	-2.06

MLP & CNN



Multi-layer Perceptron

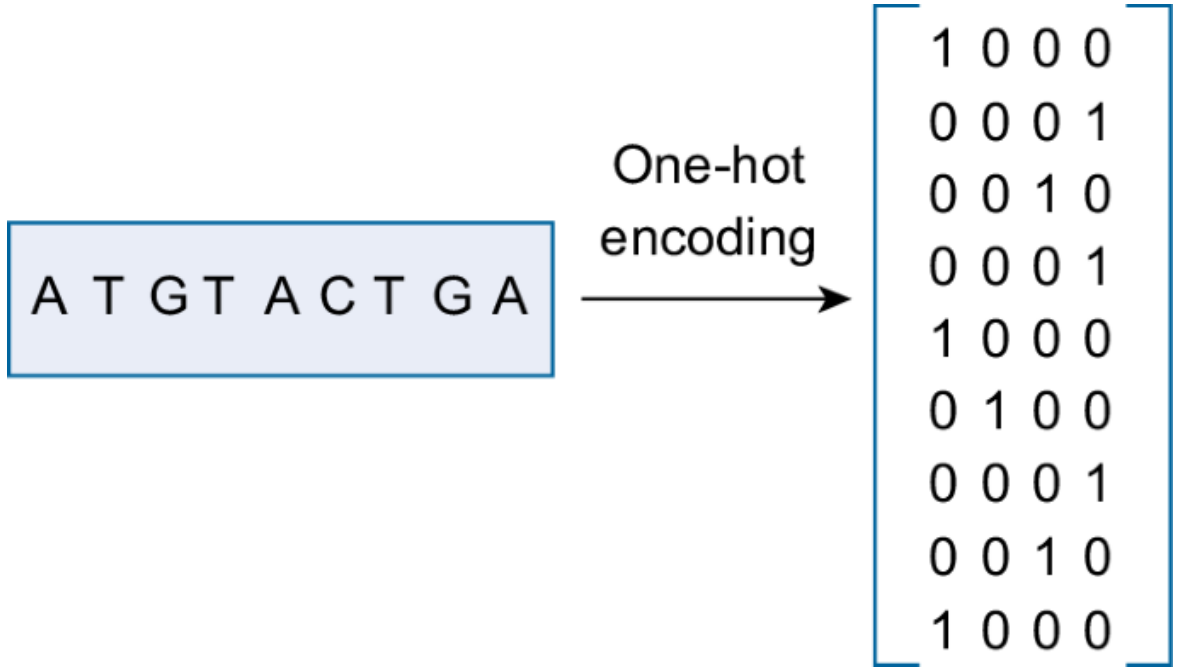


Convolutional Neural Network

One hot-encoding for DNA sequence

- Each nucleotide is represented as a one-hot vector
- RNA sequences can also be encoded similarly by simply changing T to U

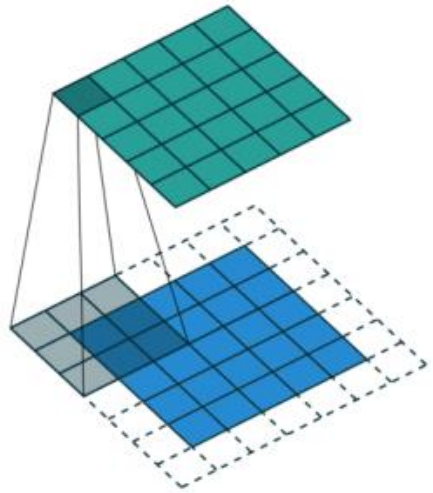
A = (1,0,0,0)
G = (0,1,0,0)
C = (0,0,1,0)
T = (0,0,0,1)



Convolution over one hot-encoding matrix

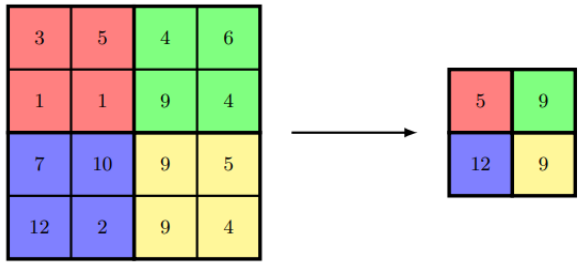
- CNNs represent genomic sequences as 1D or 2D images with four associated channels (A, C, G, T)

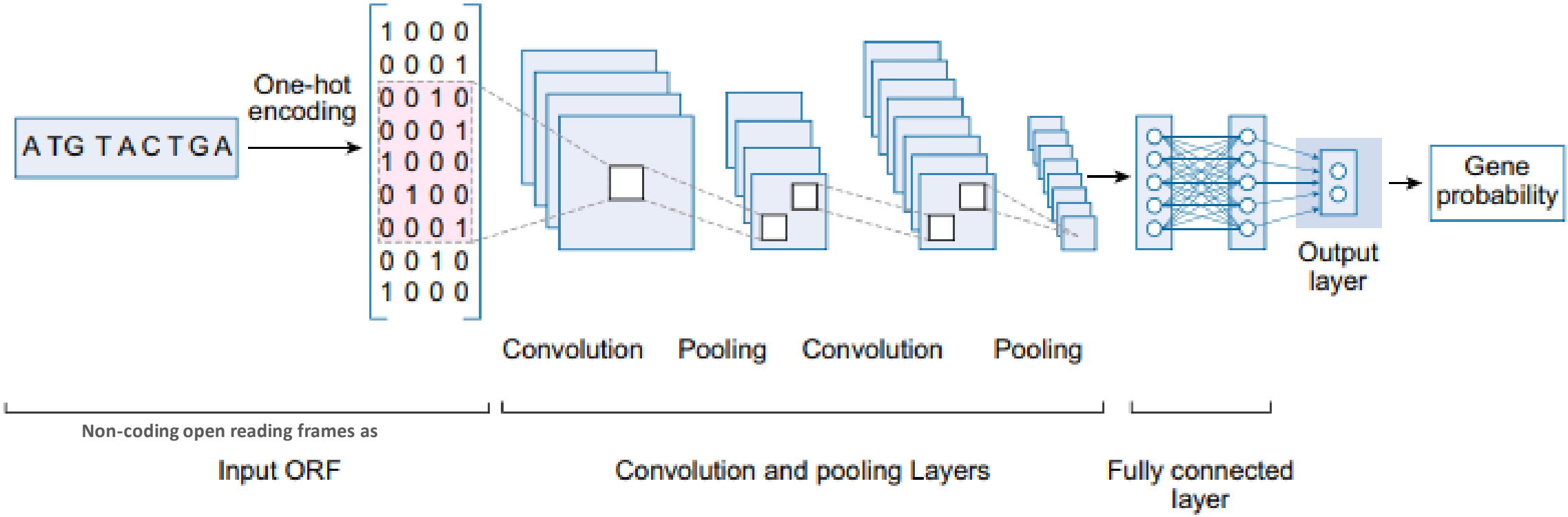
	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1



Convolution Neural Network for DNA sequence

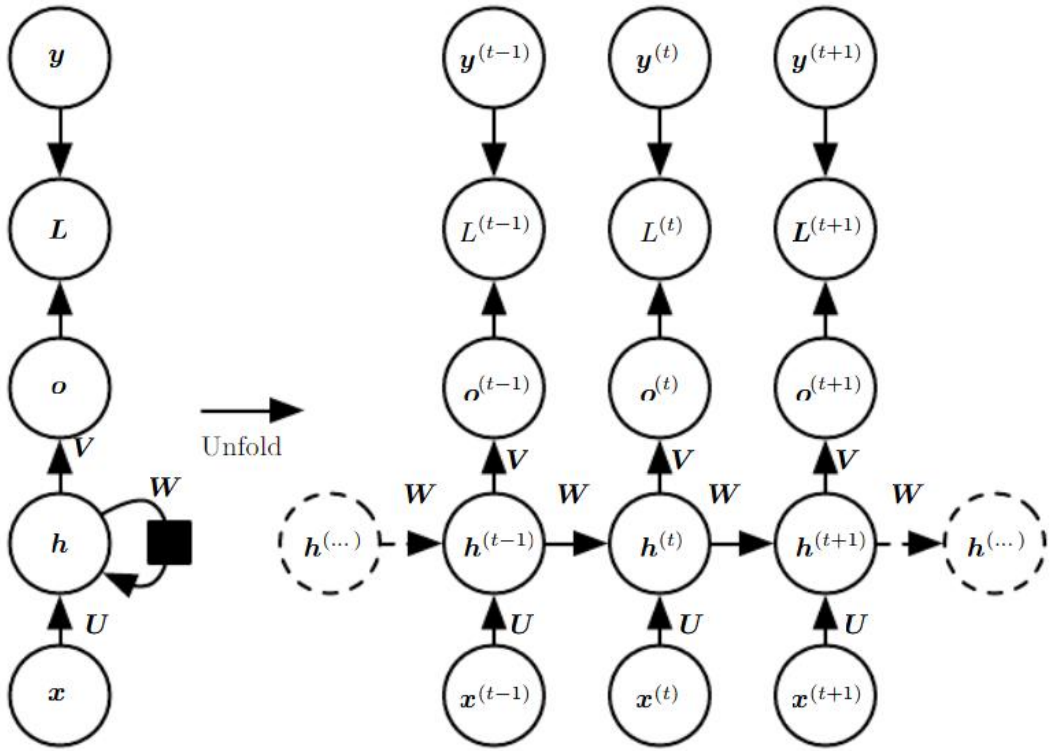
- CNN architectures designed for motif discovery and classification consist of one or more sets of four layers.
  - Convolutional layer
  - Pooling layer(Max/average)
  - Fully connected NN layer
  - Output layer





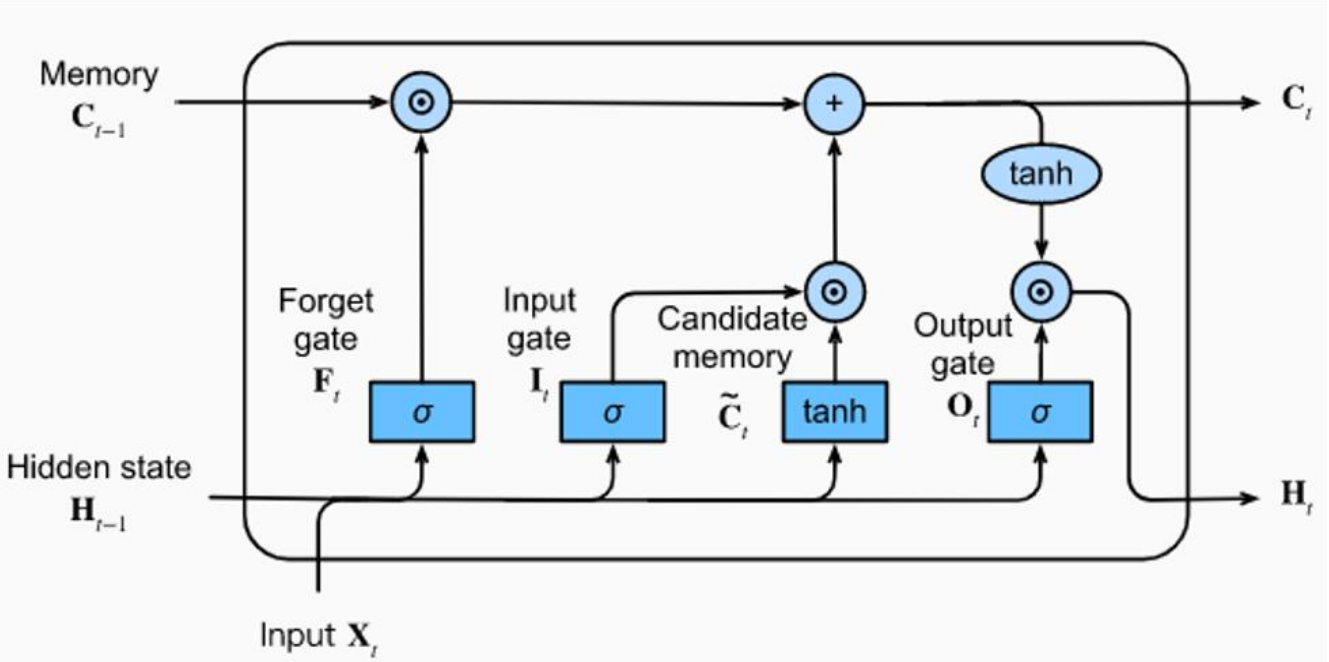
Recurrent neural network (RNN)

- Recurrent networks have recurrent connections between hidden units
  - Allow information to be passed from one step of the sequence to the next.
  - This recurrence allows the network to capture dependencies and patterns in the data that involve temporal relationships.
- RNN architecture is incapable of learning long-term dependencies.

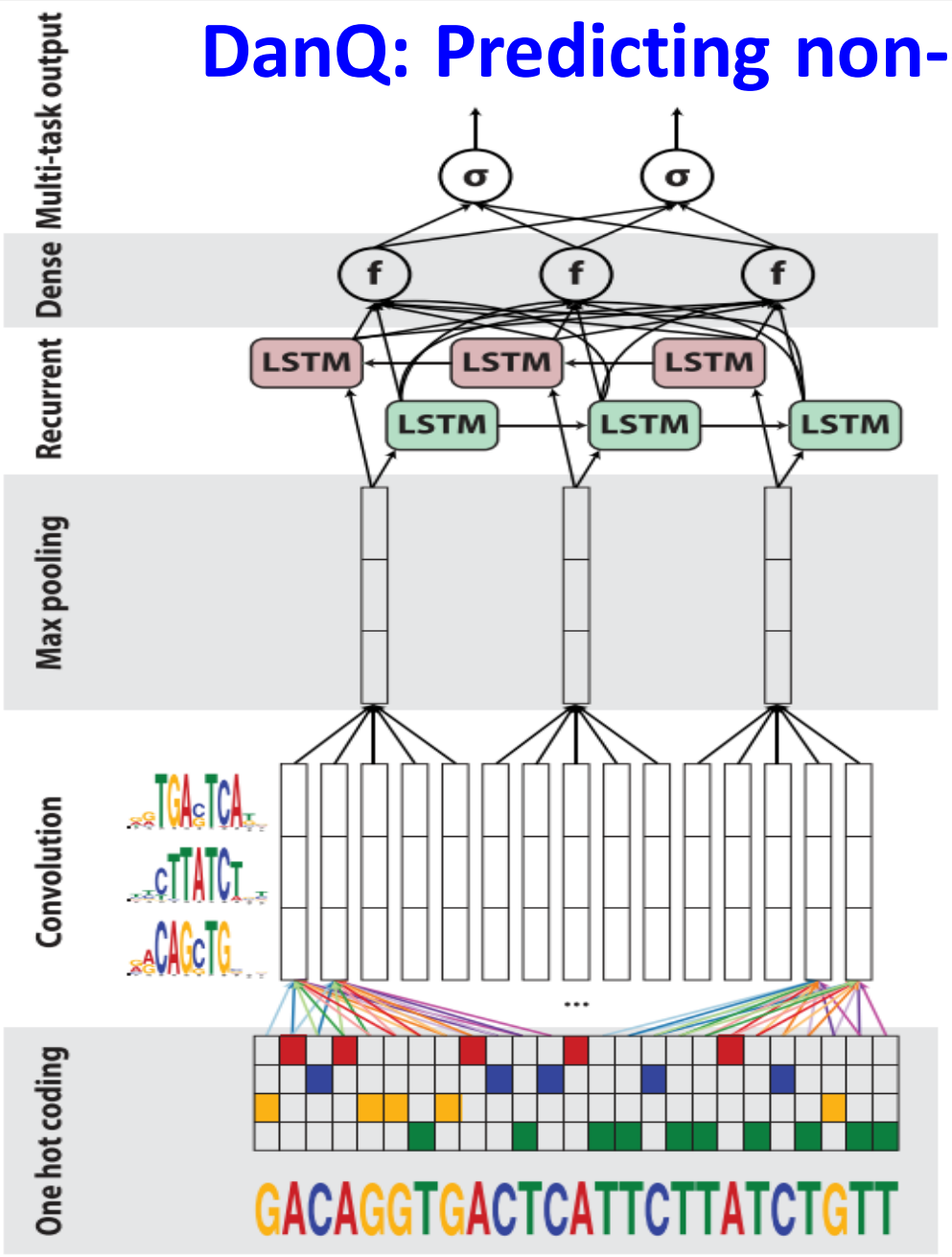


Long Short-Term Memory (LSTM) & Bi-LSTM

- To address issue of RNN, introduced Long Short-Term Memory (LSTM).
- LSTM consists of an input gate, output gate, and forget gate
- forget gate that allows the model to either reflect or forget the impact of input data at each time step.



DanQ: Predicting non-coding function de novo from sequence

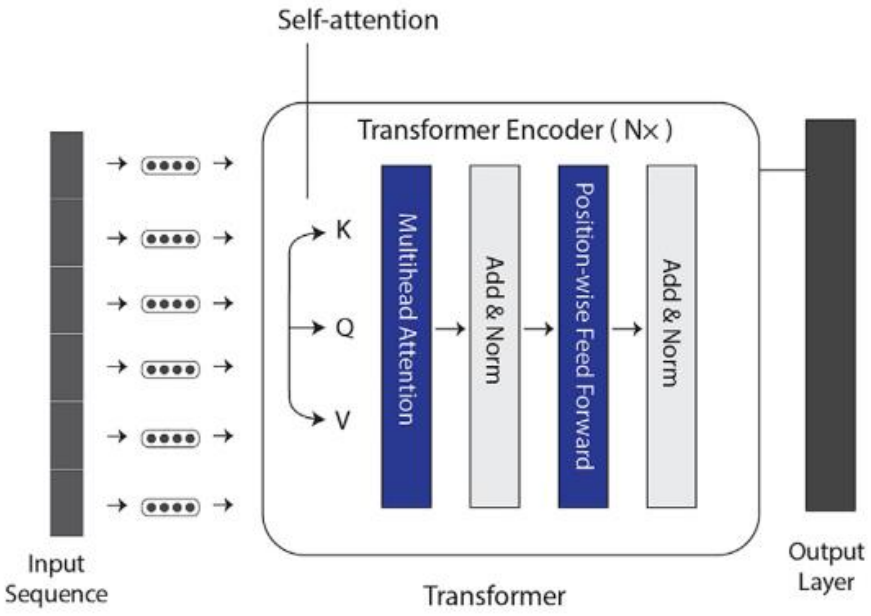


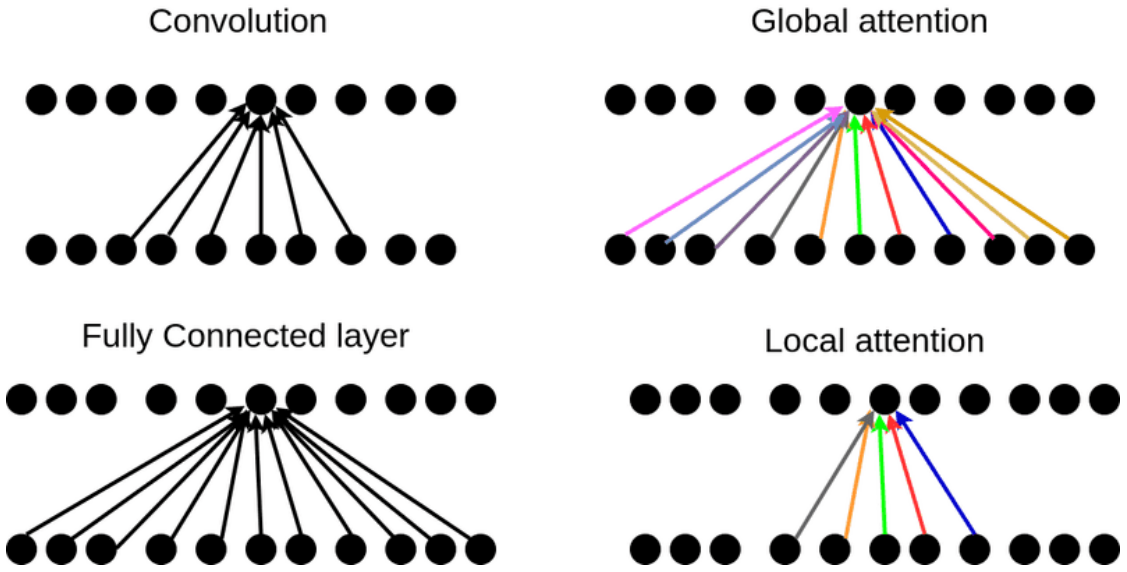
CNN & RNN(Bi-LSTM)

- hybrid convolutional : captures **regulatory motif**
- Bi-LSTM framework: captures **long-term dependencies between the motifs** in order to **learn a regulatory 'grammar'** to improve predictions.

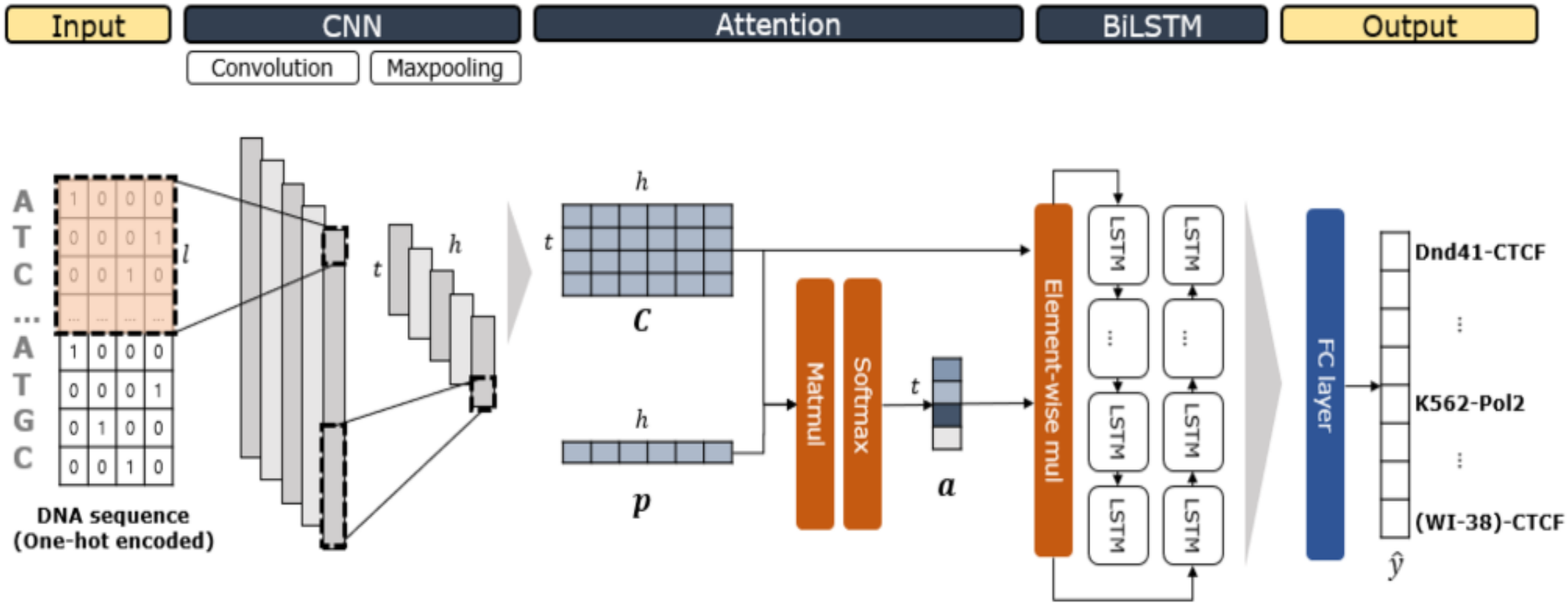
Transformer (Attention)

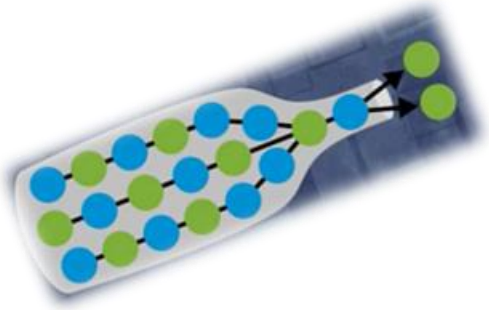
- Attention mechanism can assign different weight scores to each fragment of an input sequence to focus on more important fragments when generating outputs.



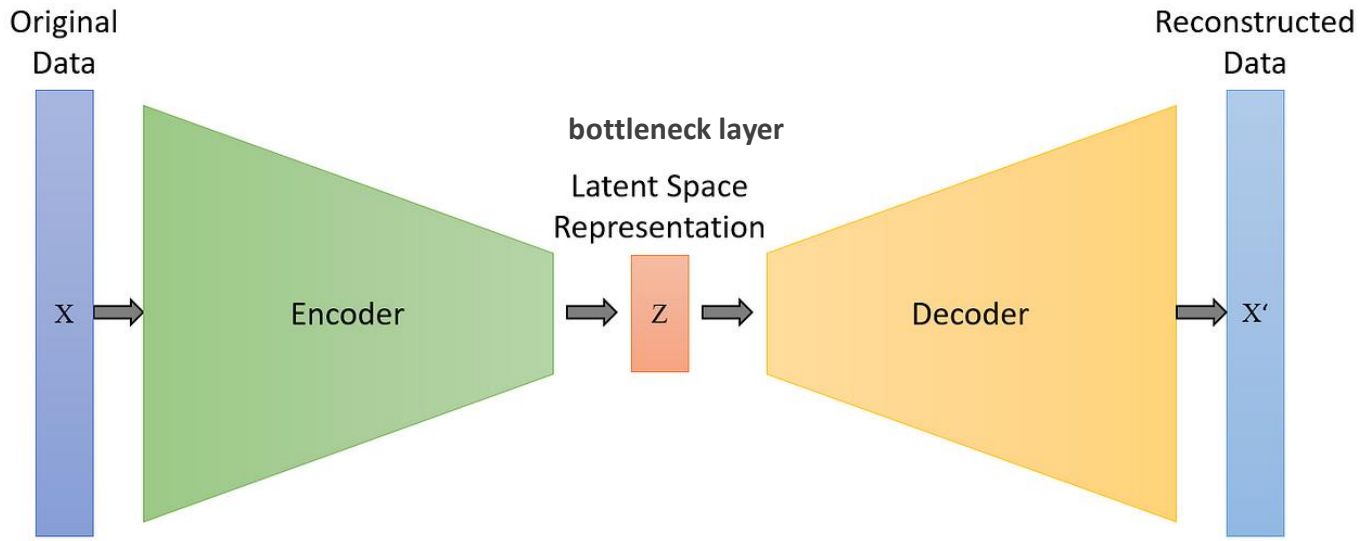


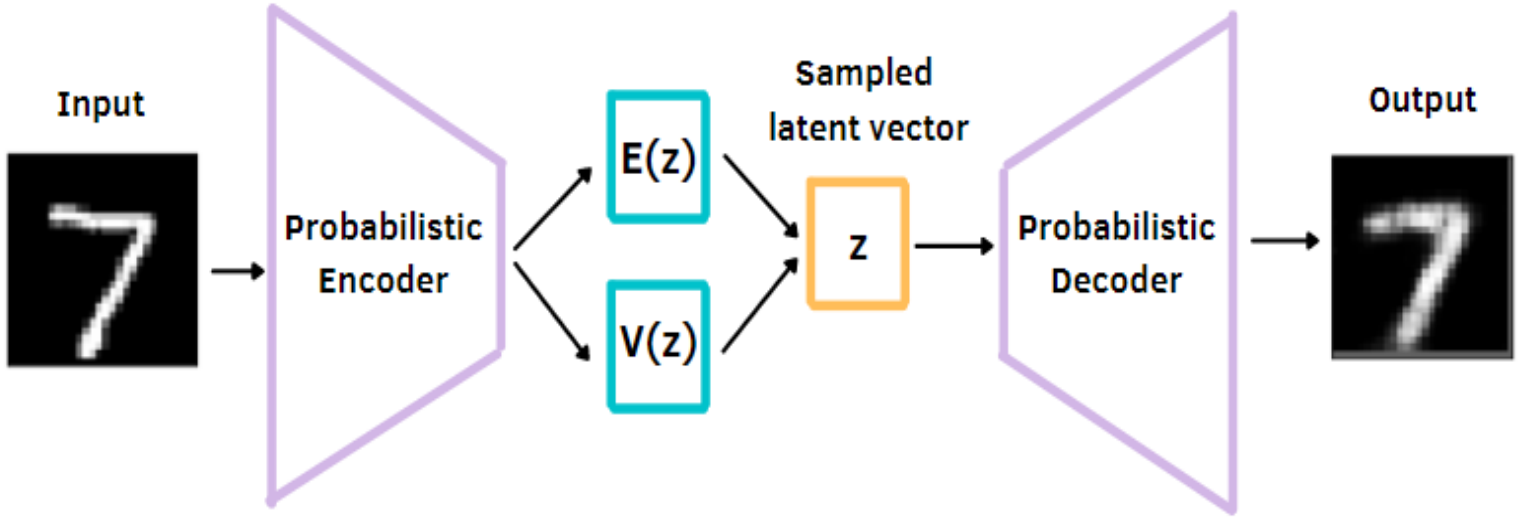
TBiNet
Enhancing the interpretability of transcription factor binding site prediction using attention mechanism





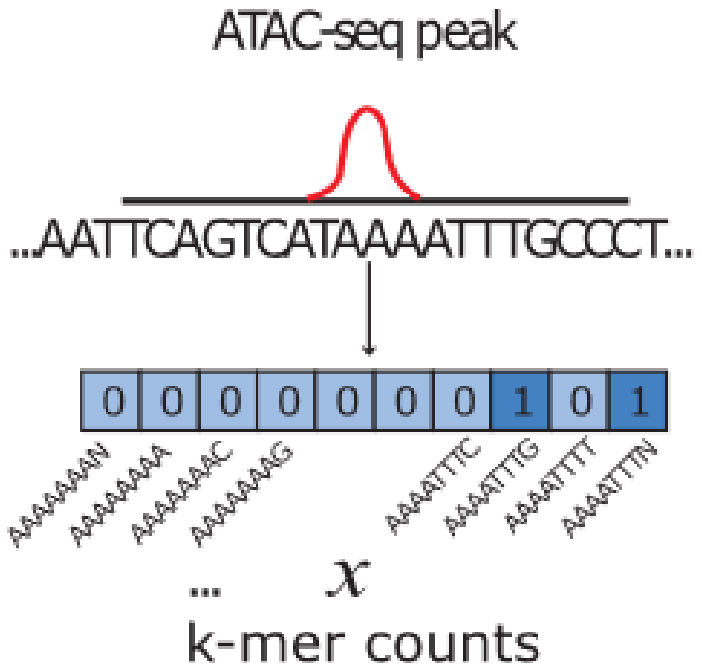
Autoencoder & Variational autoencoder





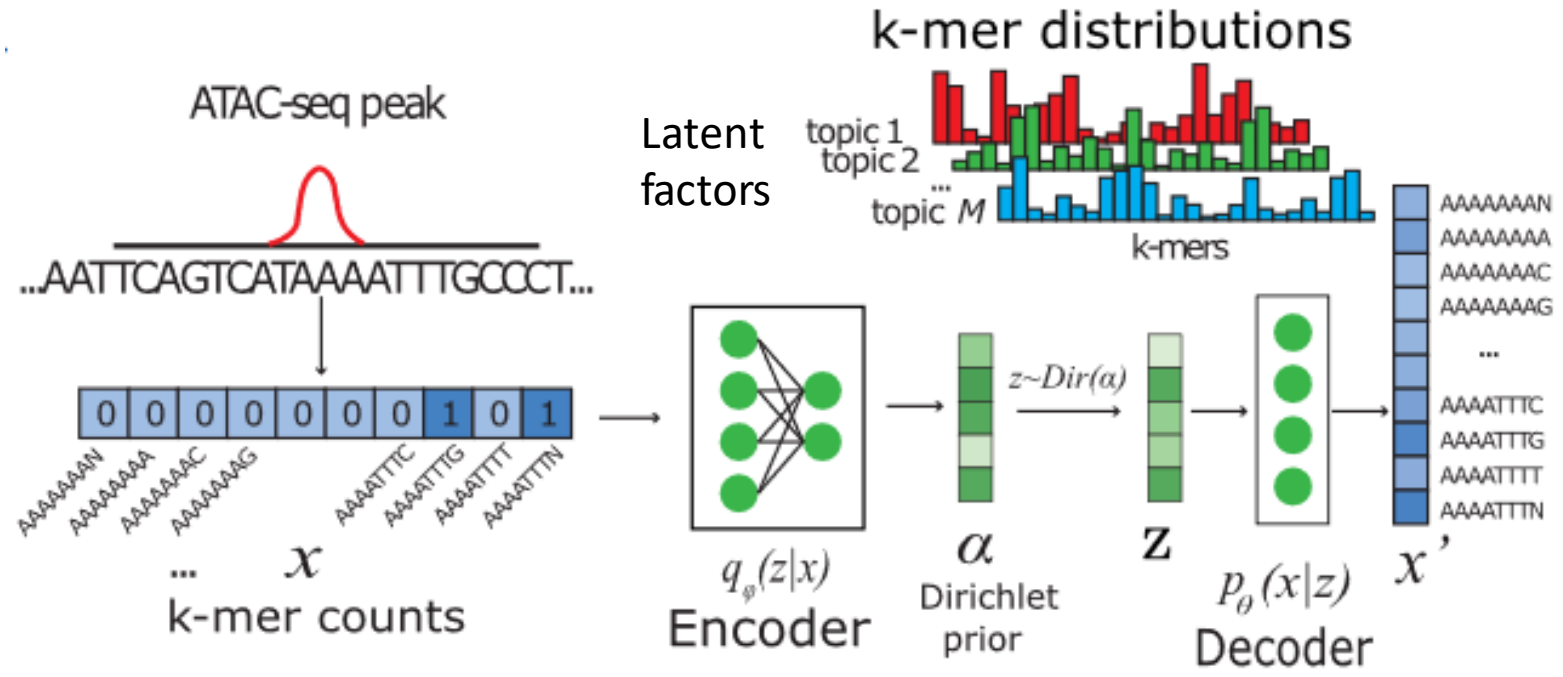
BindVAE
Variational autoencoders for de novo motif discovery from accessible chromatin

- Input: peaks from a cell type (~100k peaks), 200bp length DNA sequences



BindVAE: Variational autoencoders for de novo motif discovery from accessible chromatin

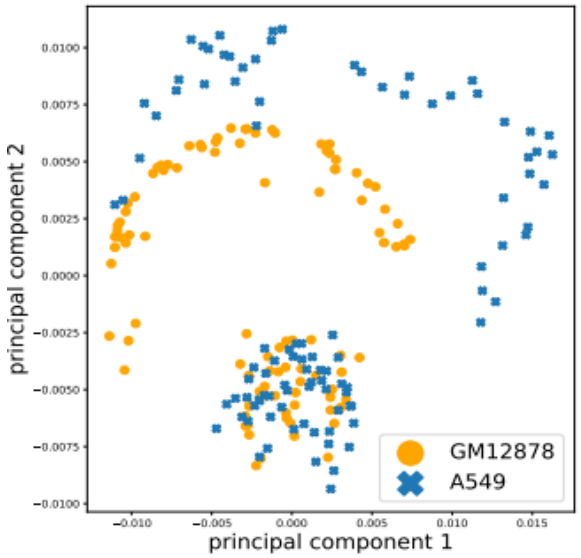
- VAE achieves compression in a probabilistic manner
- Encoder transforms the input x into parameters describing a probability distribution
- The decoder then reconstructs the input from the latent representation z



BindVAE

TF-binding motifs, cell-type specific

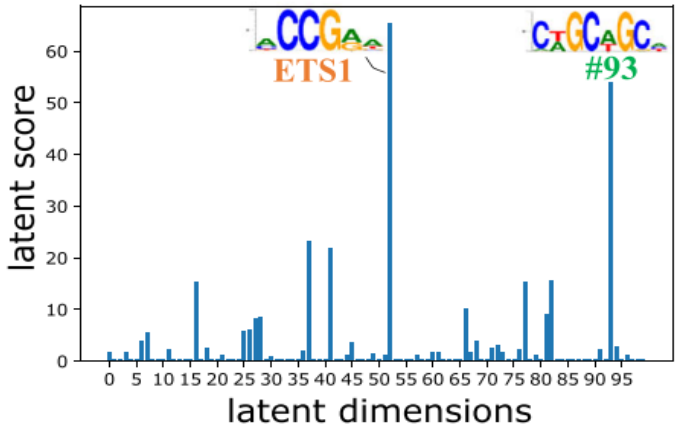
TF	GM12878 motif	A549 motif	CIS-BP motif
HNF4A			
NFIA			
SRY			
ELF5			
OLIG3			



Disentangled output from BindVAE

chr9: 69065430 – 69065460

TTCGGCCCTCTGCAGCCGCCATAGCTCCCCAGCAGAAAC CCGGAAGTGGA



FOXJ3-TBX21 cooperative binding motifs

CAP-SELEX motif
(generated by meme
from enriched probes)

BindVAE motif

CAP-SELEX motifs
from Jolma et al. (2015)

References

- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.
- Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." *Nature methods* 12.10 (2015): 931-934.
- Kshirsagar, Meghana, et al. "BindVAE: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin." *Genome Biology* 23.1 (2022): 174.
- Yuan, Han, et al. "BindSpace decodes transcription factor binding signals by large-scale sequence embedding." *Nature methods* 16.9 (2019): 858-861.
- Park, Sungjoon, et al. "Enhancing the interpretability of transcription factor binding site prediction using attention mechanism." *Scientific reports* 10.1 (2020): 13413.
- Cazares, Tareian A., et al. "maxATAC: Genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks." *PLOS Computational Biology* 19.1 (2023): e1010863.
- Wang, Meng, et al. "DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants." *Nucleic acids research* 46.11 (2018): e69-e69.
- Setty, Manu, and Christina S. Leslie. "SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps." *PLoS computational biology* 11.5 (2015): e1004271.
- Korhonen, Janne H., et al. "Fast motif matching revisited: high-order PWMs, SNPs and indels." Bioinformatics 33.4 (2017): 514-521.

Thanks