A Primer on Deep Learning

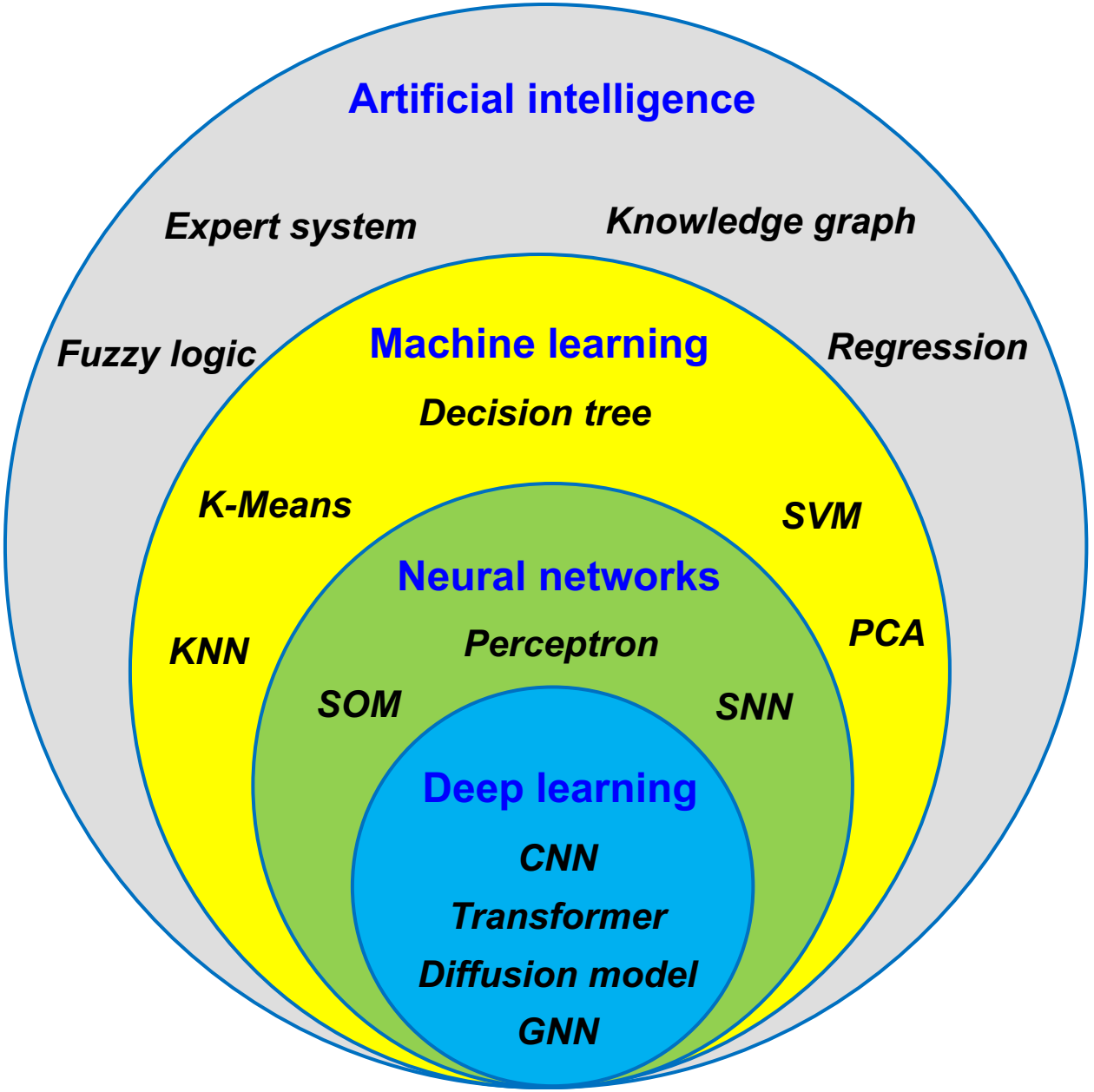
Dong Xu

EECS Department
C. S. Bond Life Sciences Center
University of Missouri, Columbia
http://digbio.missouri.edu

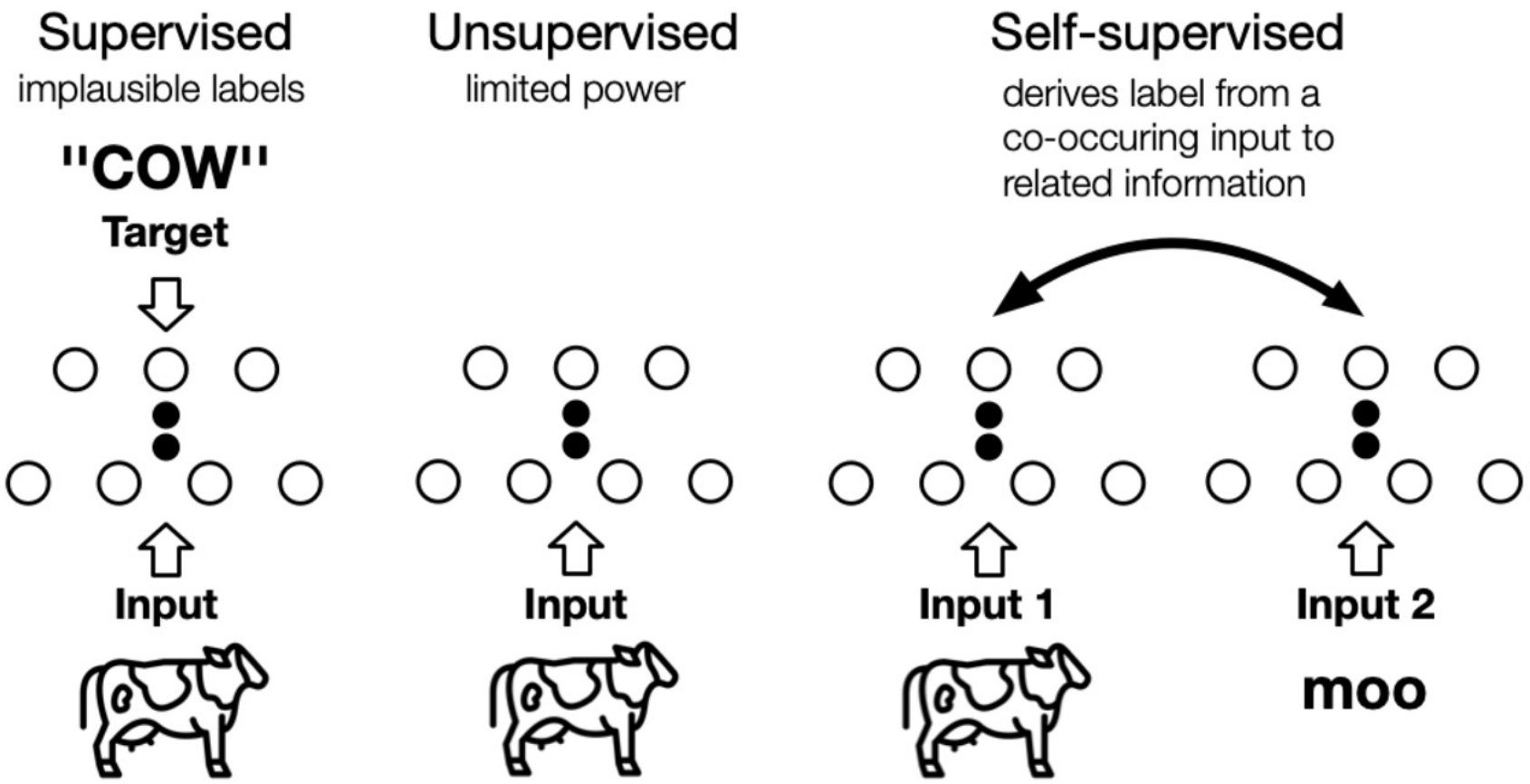
Introduction

- Primer for those who are unfamiliar with deep learning methods
- A high-level view of deep learning strategy and methods
- Deep learning has been successfully applied in single-cell data analyses

AI Scope


Artificial intelligence
Expert system
Knowledge graph
Fuzzy logic
Machine learning
Regression
Decision tree
K-Means
SVM
Neural networks
KNN
Perceptron
PCA
SOM
SNN
Deep learning
CNN
Transformer
Diffusion model
GNN

Machine Learning Types



Supervised Machine Learning

- Apply a prediction function to a feature representation of image to get the desired output:

f() = “apple”

f() = “tomato”

f() = “cow”

$$y = f(\mathbf{x})$$

Output (label)

Prediction function

Image feature

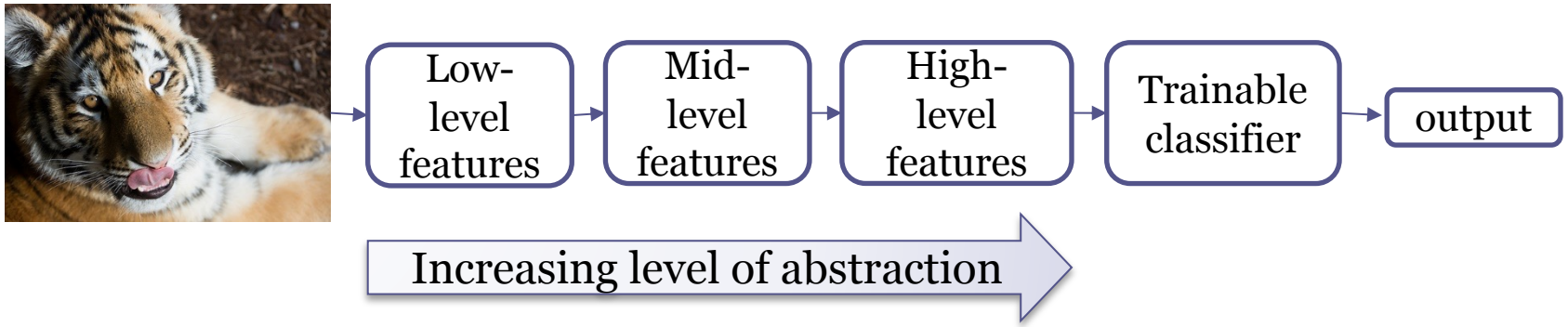
Generalization

Training set (labels known)

Test set (labels unknown)

- How well does a learned model generalize from the data it was trained on to a new test set?

Learning Hierarchical Representations



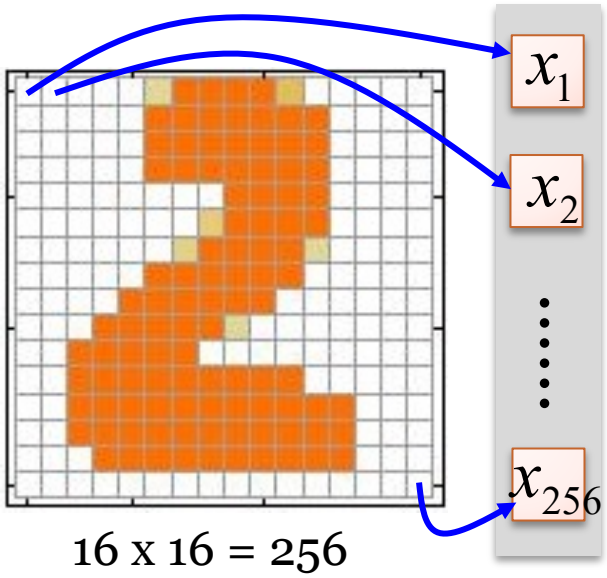
- Hierarchy of representations with increasing level of abstraction. Each stage is a kind of trainable nonlinear feature transform
- Image recognition

  Pixel → edge → texton → motif → part → object
- Text

  Character → word → word group → clause → sentence → story

Handwriting Digit Recognition

Input

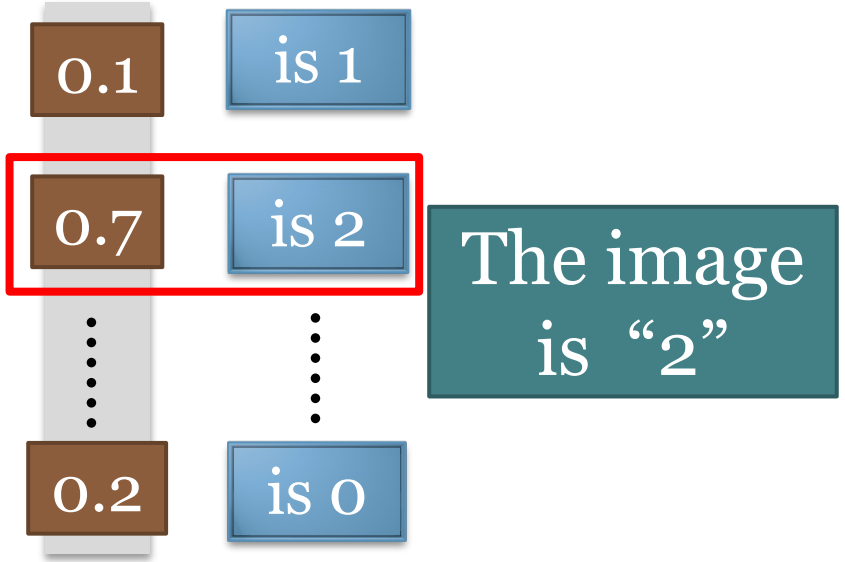


16 x 16 = 256

Ink → 1

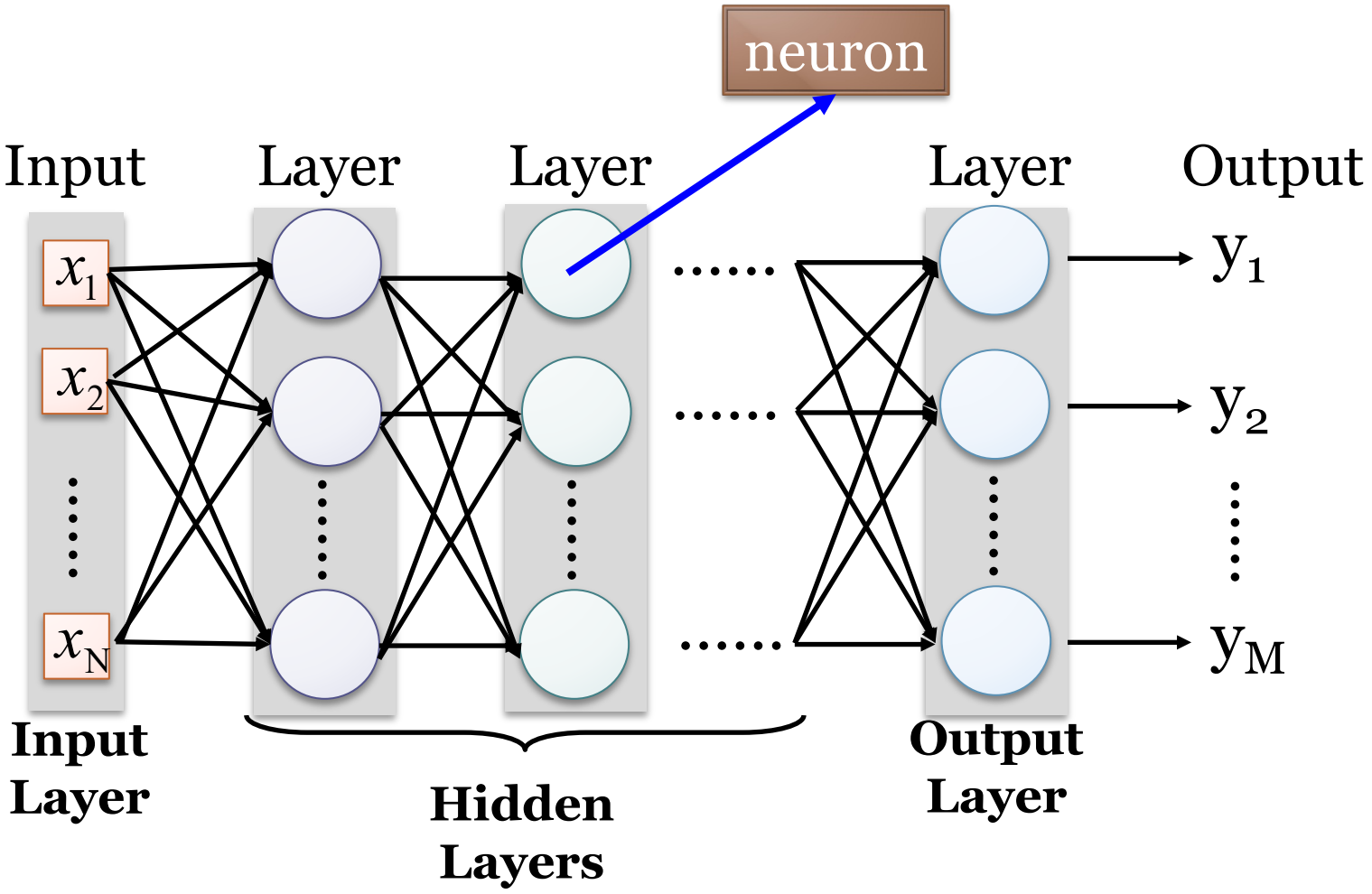
No ink → 0

Output



Each dimension represents the confidence of a digit.

Deep Neural Network (DNN)



Deep means many hidden layers

Abstraction and Representation

physical space → latent/embedding space (manifold)

photo

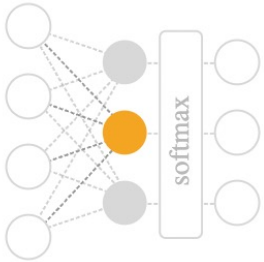
impressionism

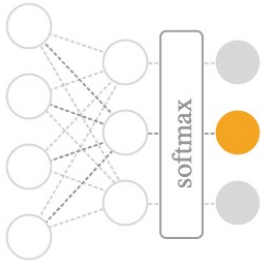
expressionism

cubism

abstract expressionism

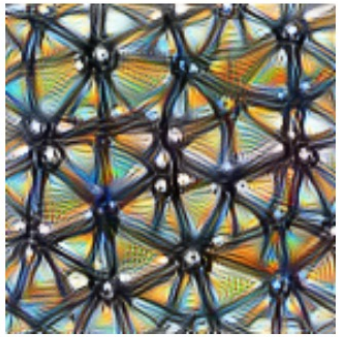
Latent Representation





Neuron

$\text{layer}_n[x,y,z]$



Channel

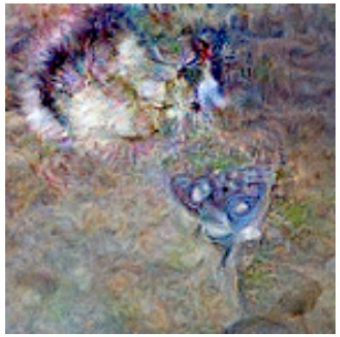
$\text{layer}_n[:,:,z]$

Layer/DeepDream

$\text{layer}_n[:,:,:]^2$

Class Logits

$\text{pre\_softmax}[k]$

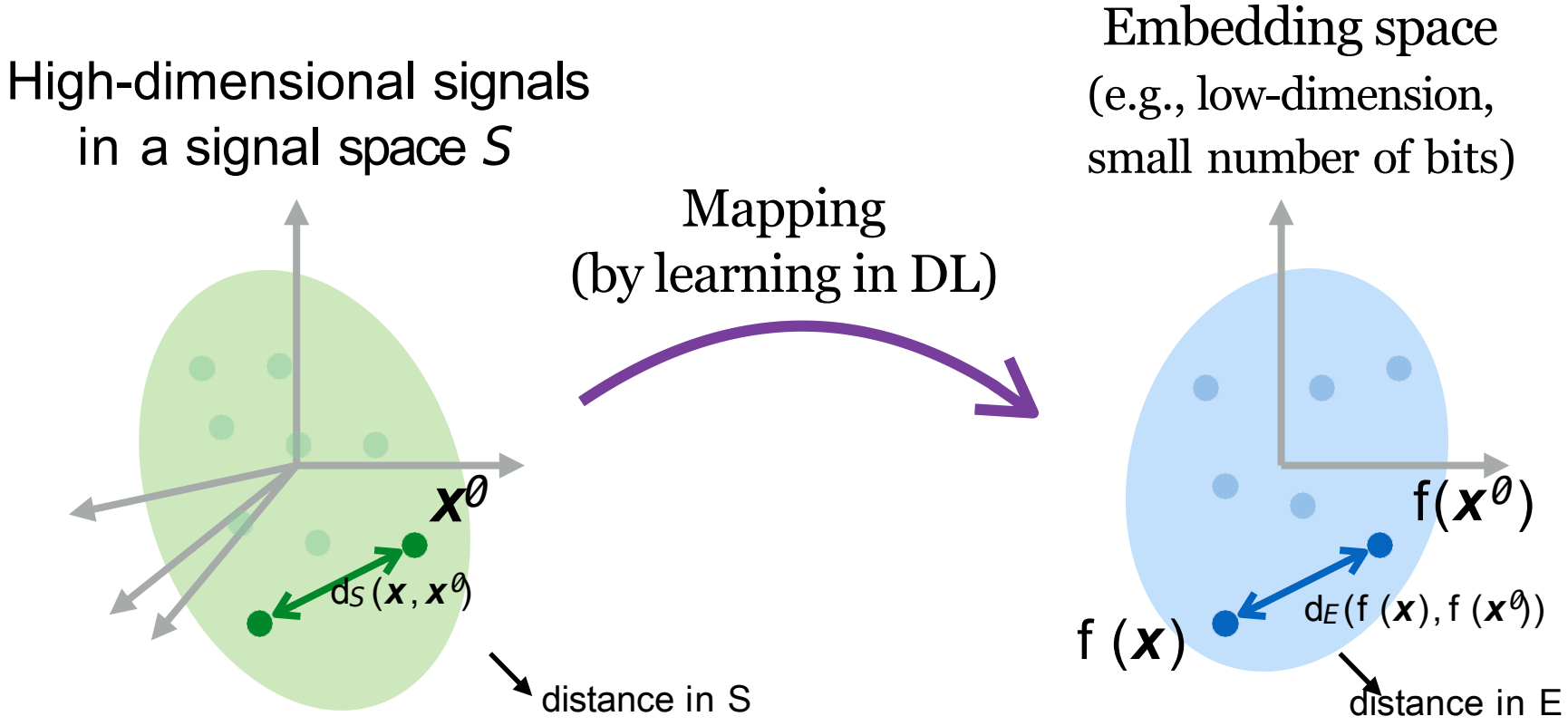


Class Probability

$\text{softmax}[k]$

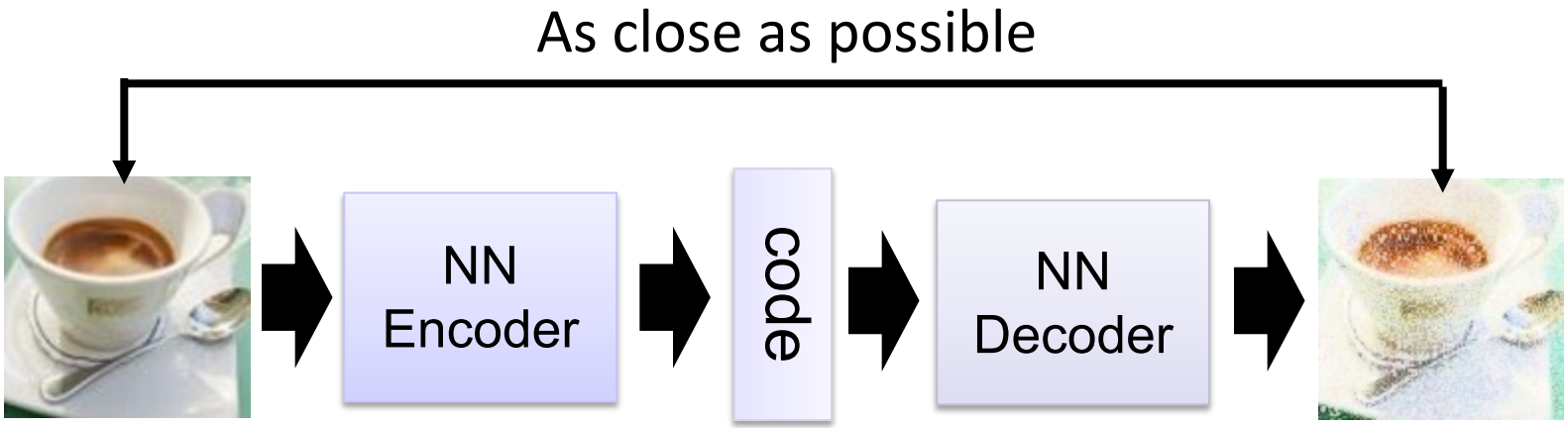
https://distill.pub/2017/feature-visualization/

Unsupervised Learning & Embedding



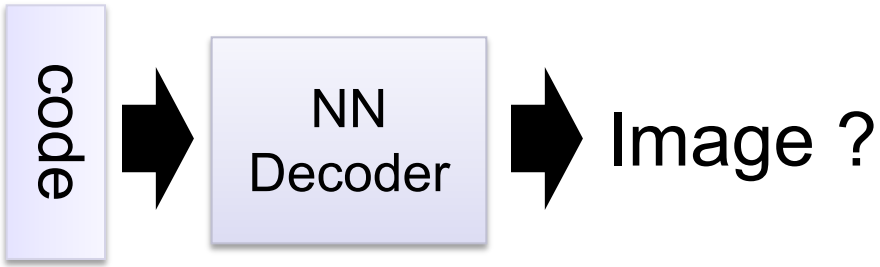
An embedding is a function from an original space to an embedding space that preserves aspects of the geometry of the original space

Autoencoder

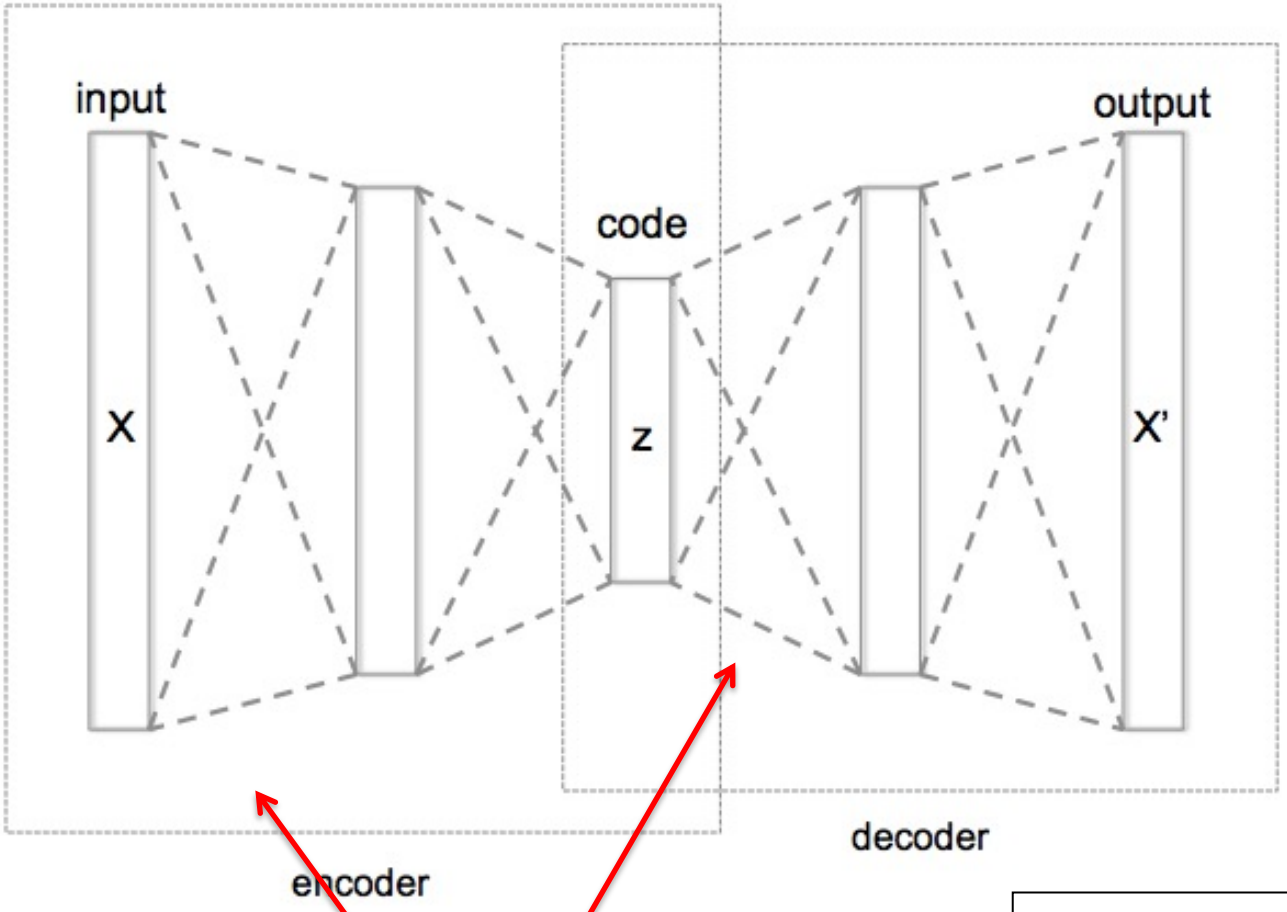


A type of unsupervised learning which discovers generic features of the data (**learn data patterns**)

Randomly generate a vector as code



Autoencoder with 3 Layers

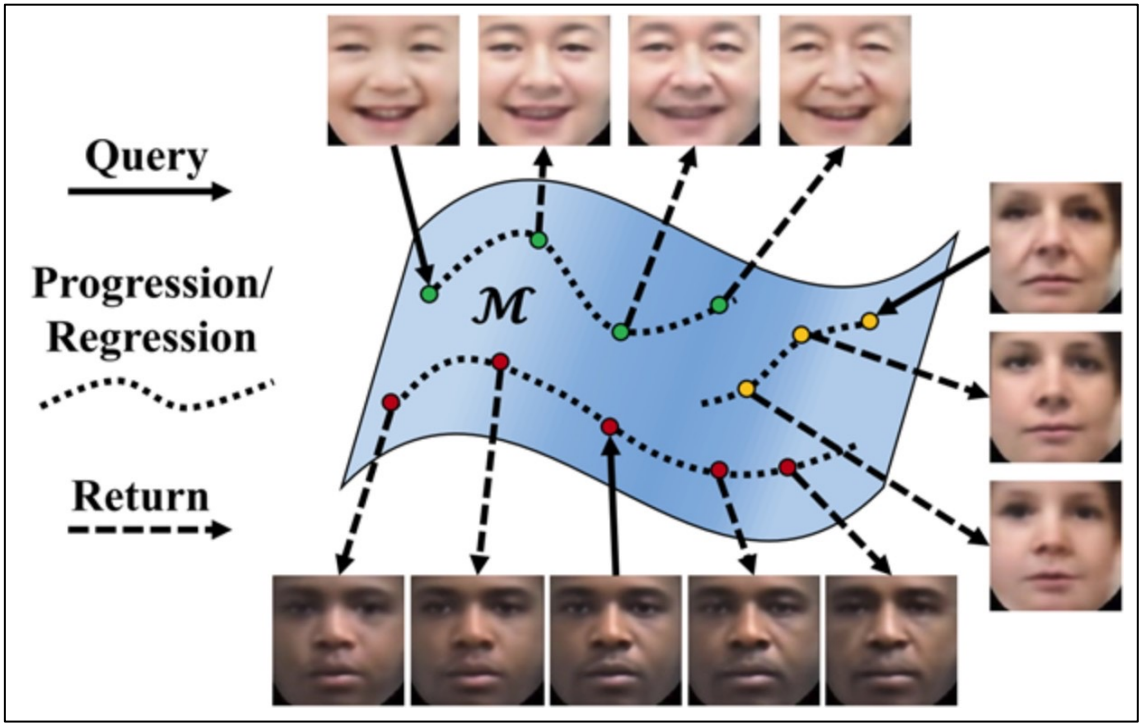


Large → small, learn to compress

Z: Latent variable; embedding

Training: model.fit(X,X)
Cost function: $\Sigma_{k=1..N} (x_k - x'_k)^2$

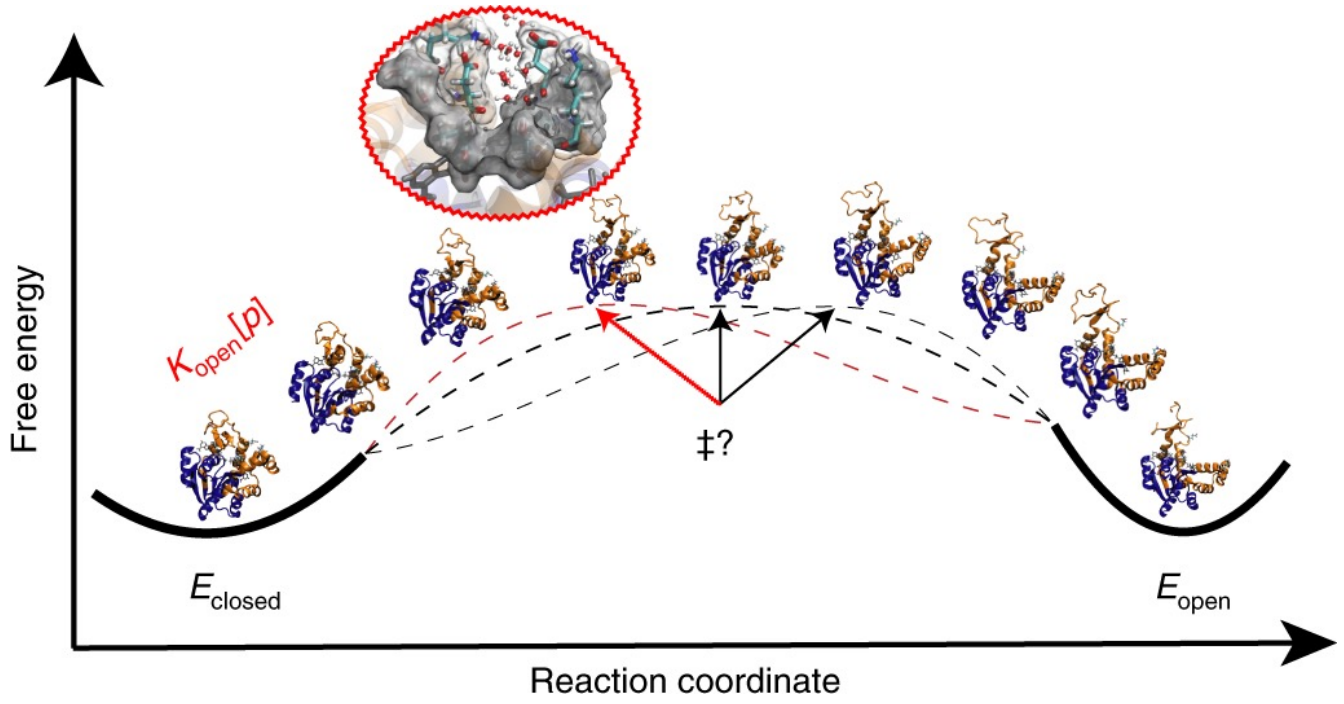
Manifold



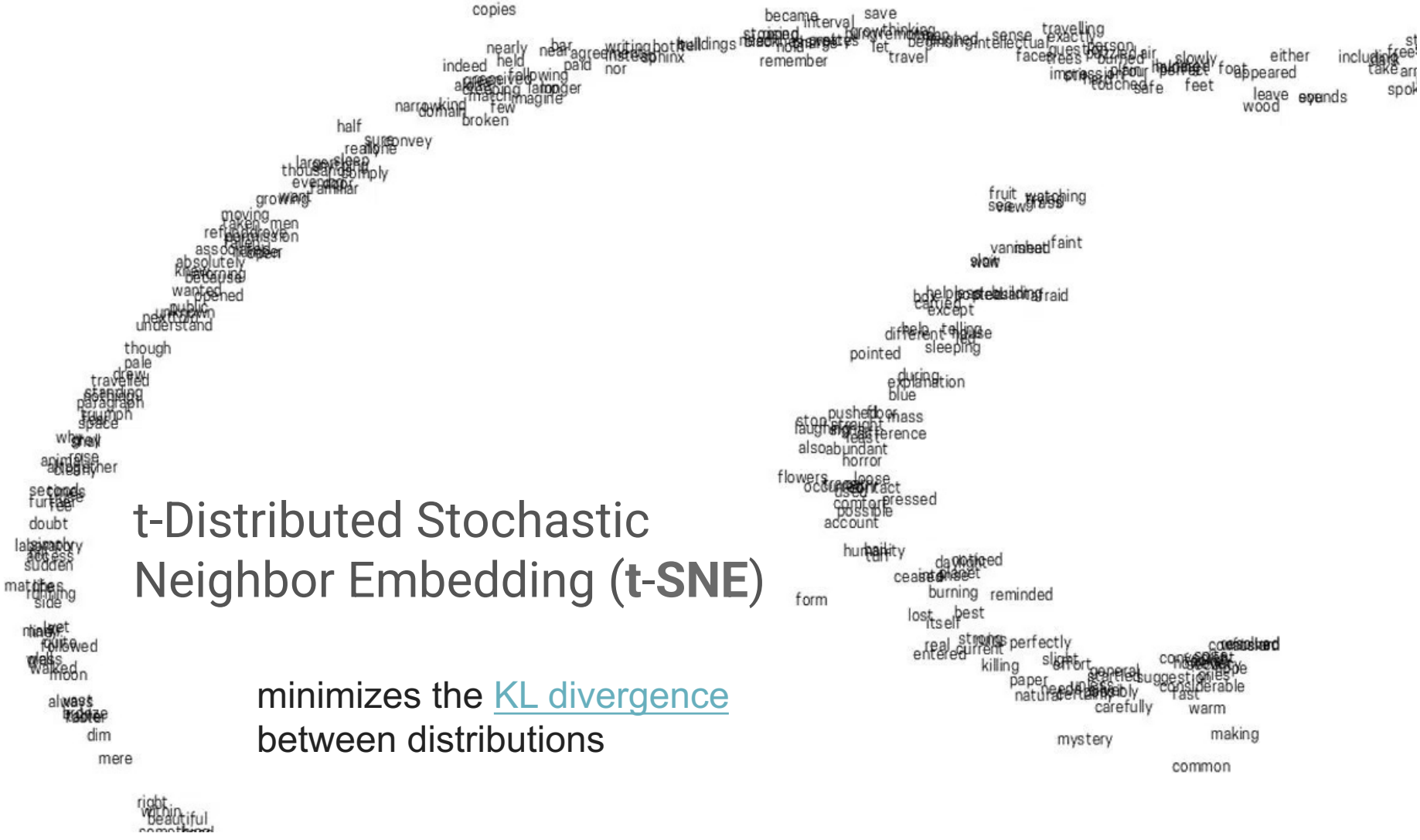
- A Manifold is a topological space that locally resembles Euclidean space near each point
- n-dim manifold → topological space M, every point x ∈ M has a neighbor homeomorphic (isomorphismic) to Euclidean space Rn

Manifold Hypothesis

- **DL Central Hypothesis**: Data concentrates around a low-dim manifold (**relevant dimension**)
- Mimic human learning
- Not all embeddings produce manifold



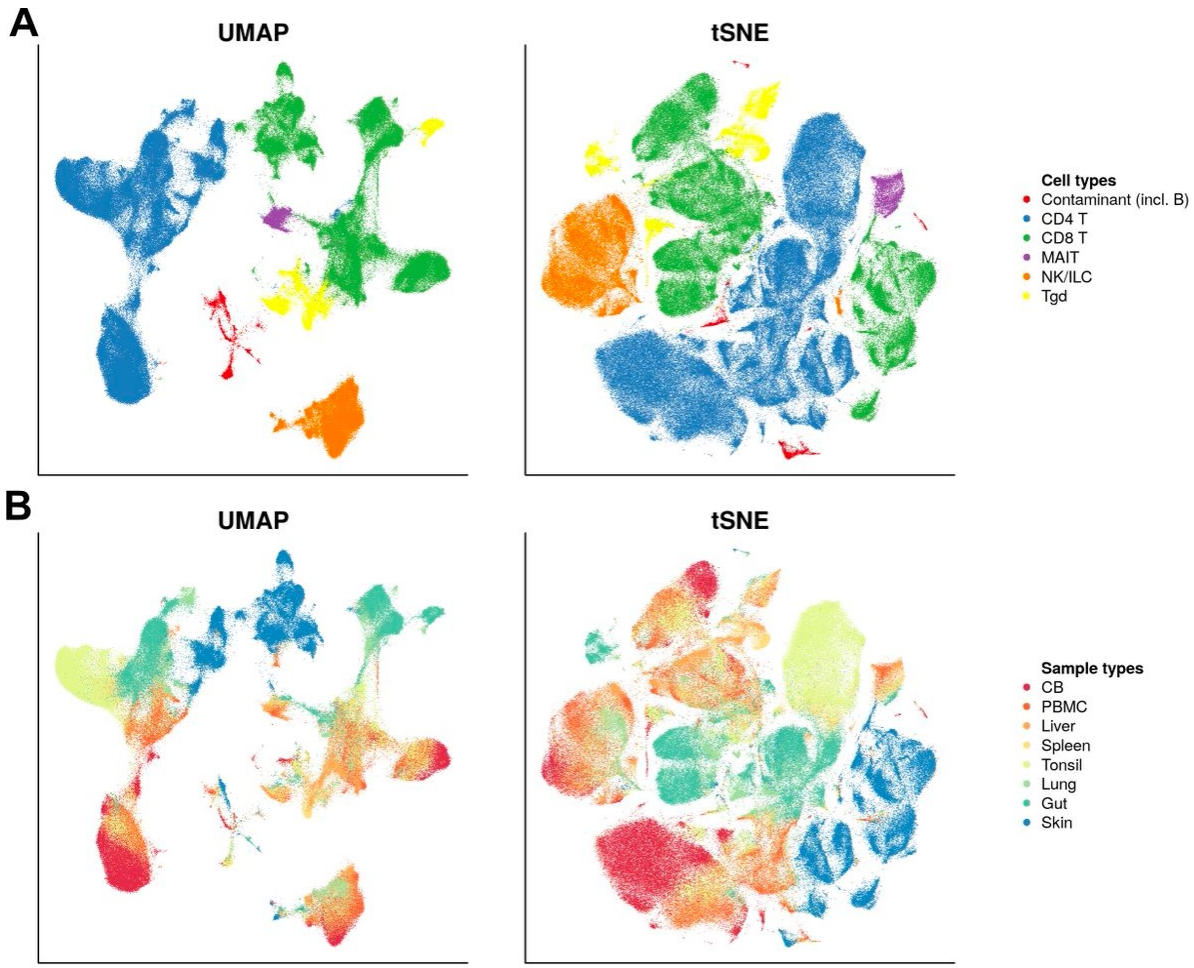
T-SNE Visualization



U-Map

Uniform Manifold Approximation and Projection (UMAP)

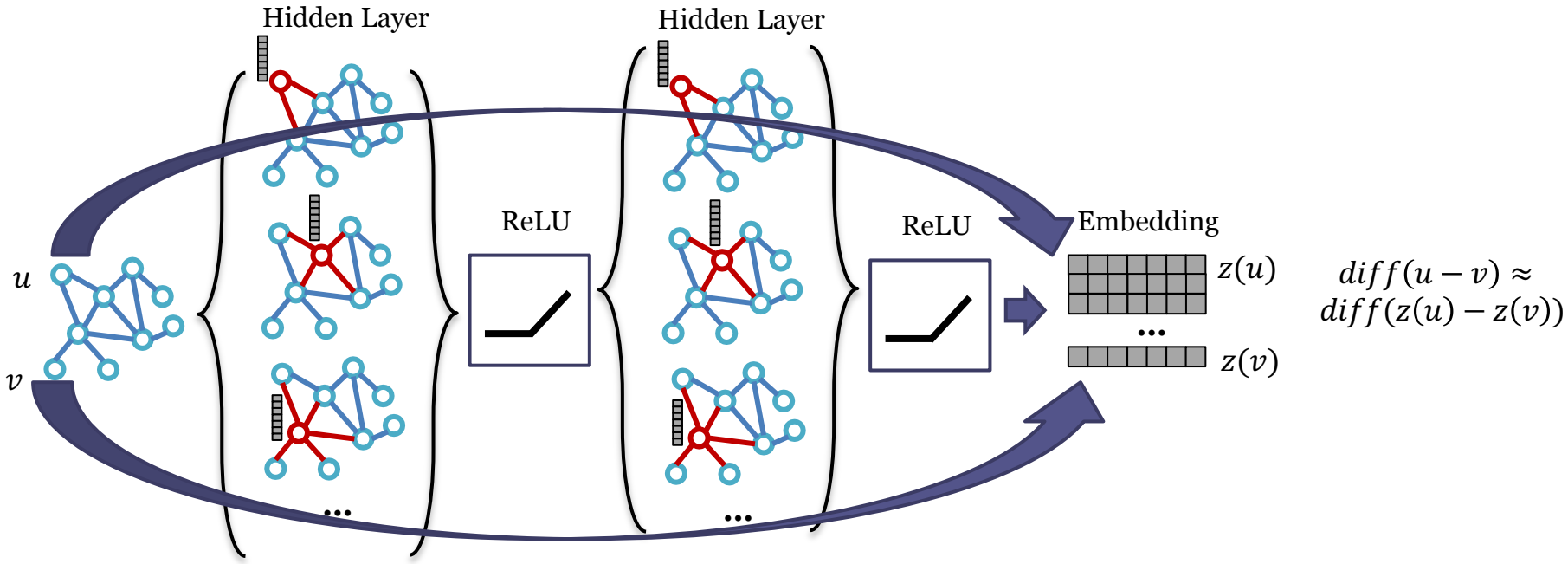
find a topological representation of the data in a lower dimensional space through manifold learning technique



https://github.com/lmcinnes/umap

Graph Neural Network (GNN)

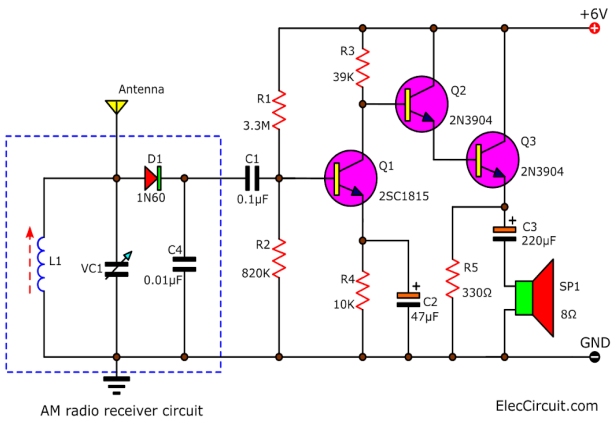
- GNN learns a task-independent representation of a graph by deconvoluting node relationships through neighbor information propagation in a deep learning architecture.
- Generate node embeddings based on local neighborhoods



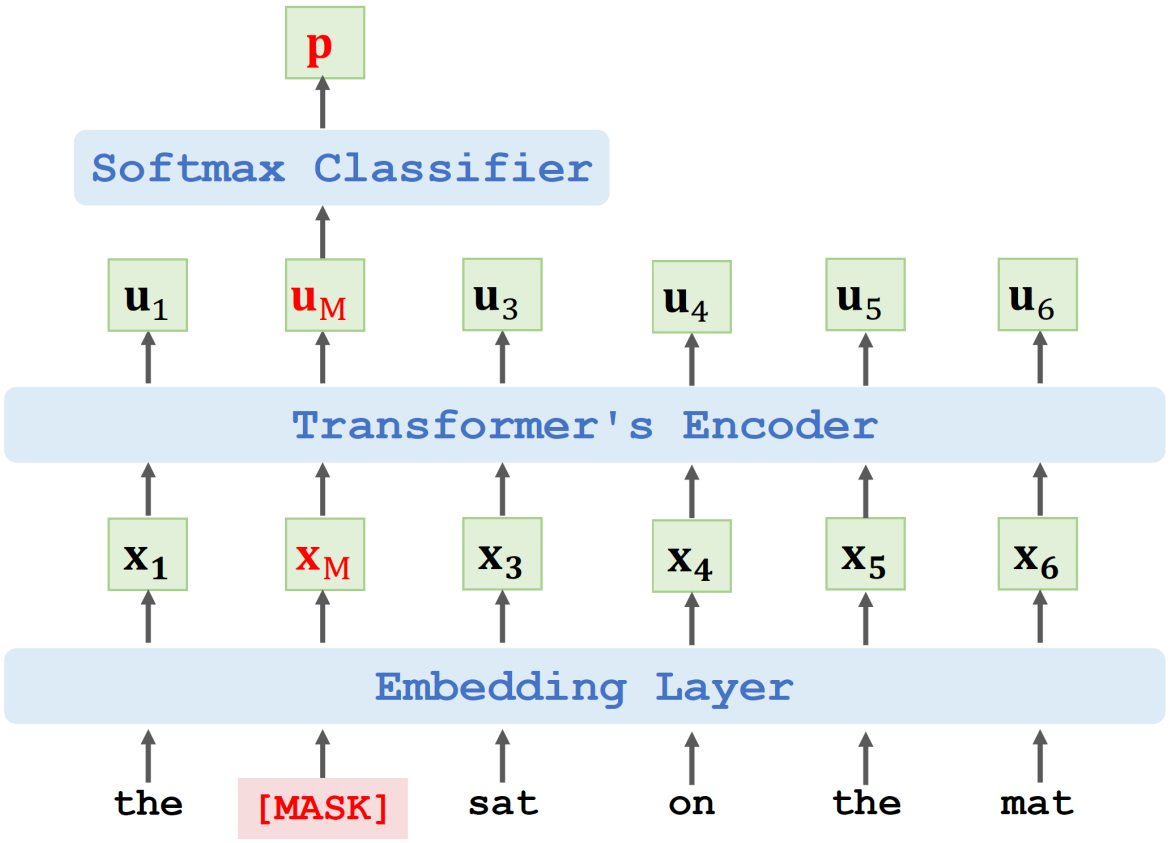
Foundation Model Era

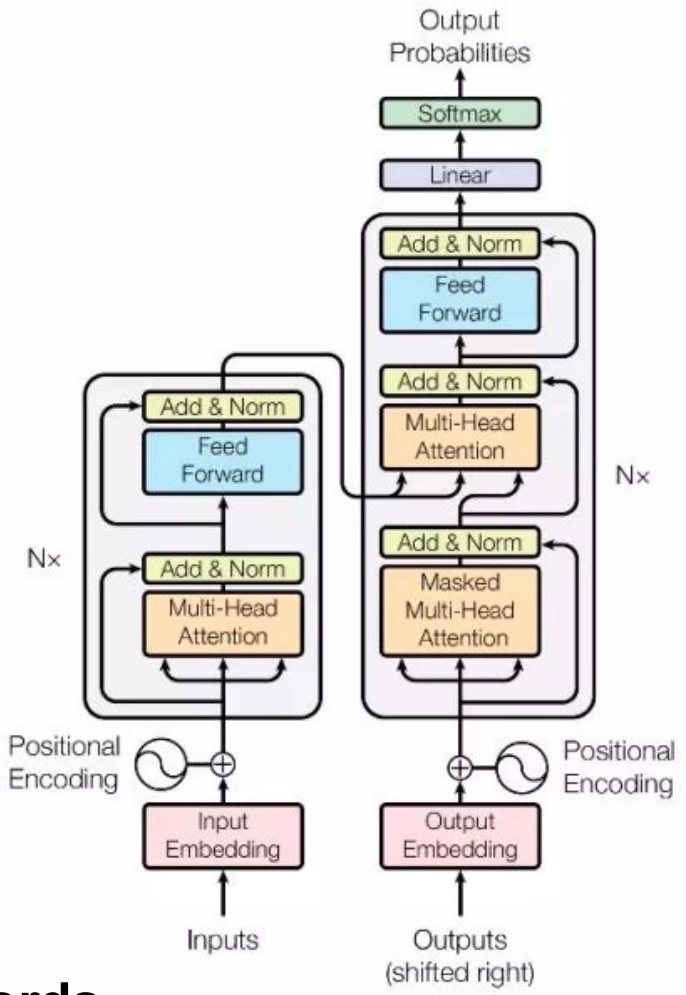
- Machine learning paradigms
  - **Feature engineering**: manual feature extraction (SVM, LightGBM, XGBoost)
  - **Architecture engineering**: raw features, design deep network (CNN, LSTM)
  - **Objective engineering**: pre-train large model and fine-tune it (ResNet50, Bert)
  - **Prompt engineering**: prompt **foundation model** in zero/few shots

- Industrial era of artificial intelligence



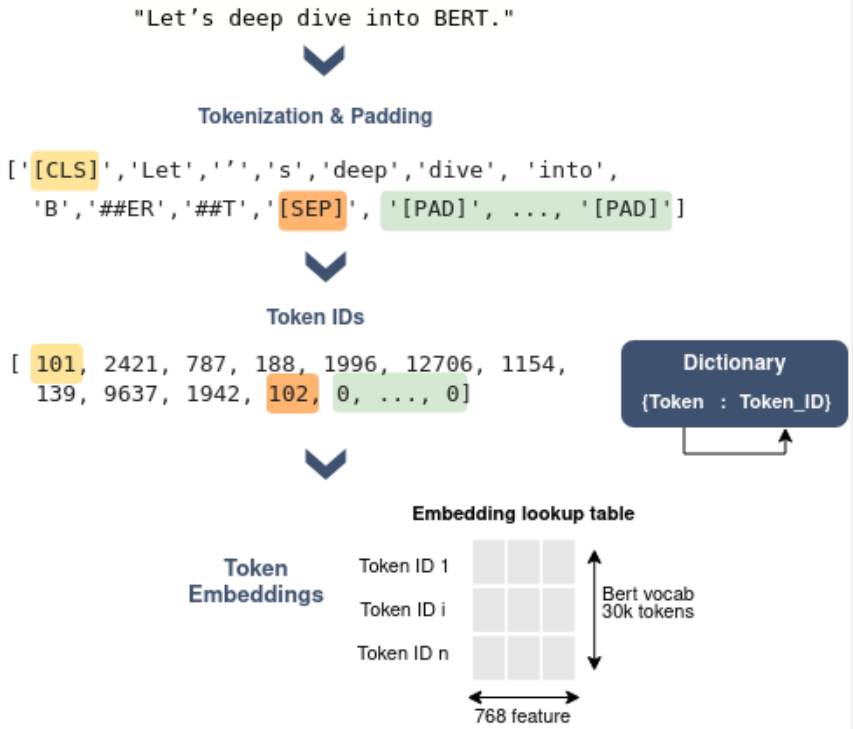
From Transformer to GPT

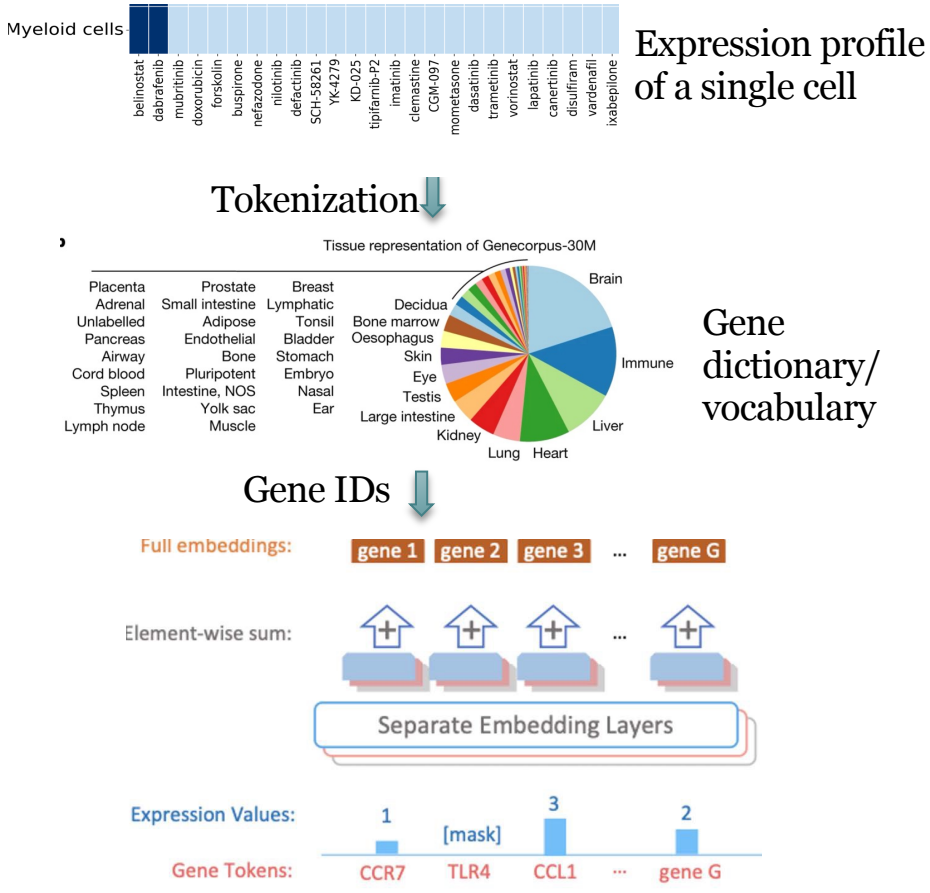




Self-supervised learning to predict missing words

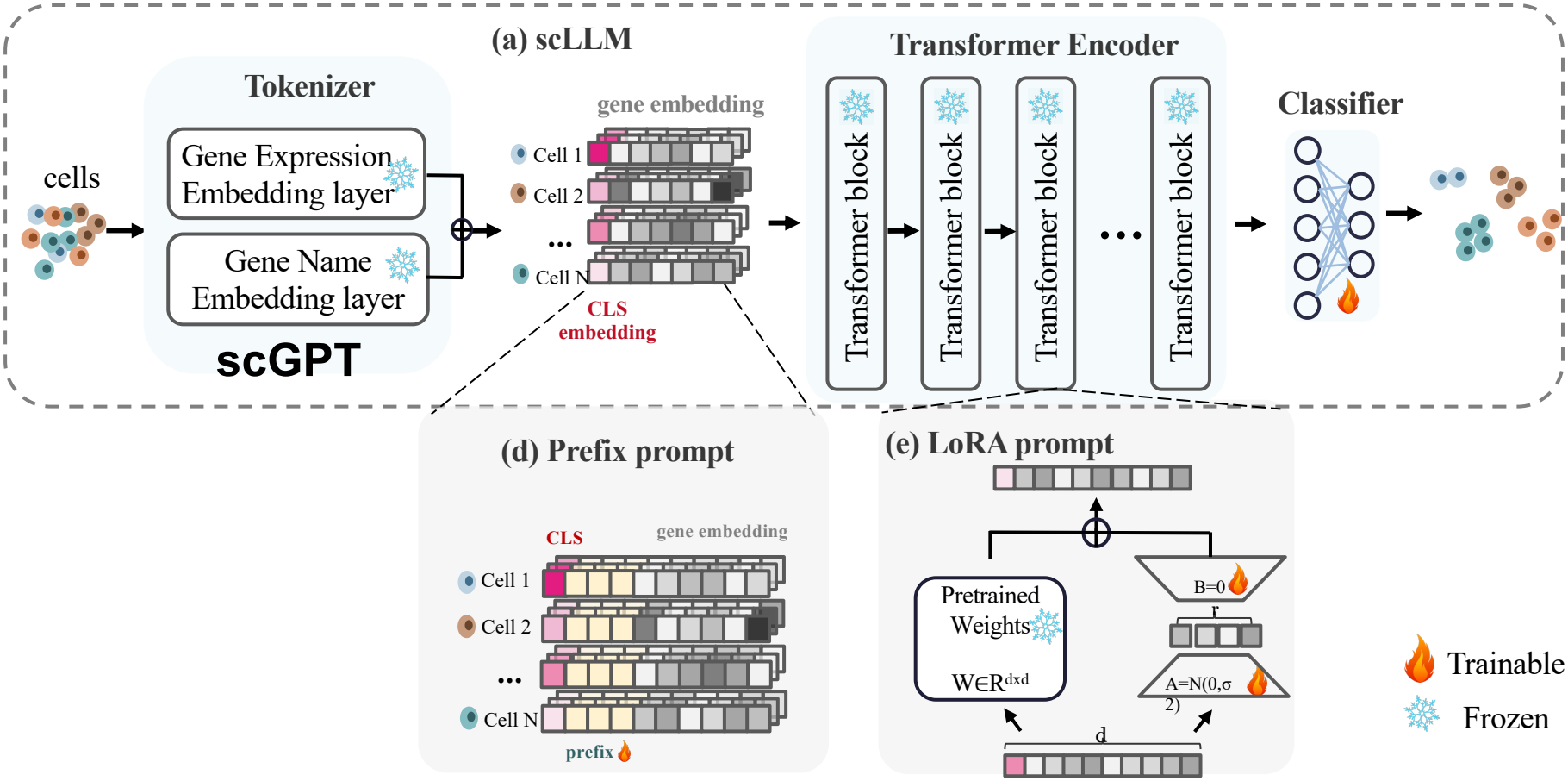
From LLM to Single-cell LLM





Embed gene expression values or gene expression order in each cell

Prompt-based Learning on scLLMs



Add small adapter to scLLM and train the adapter using small data

Summary

- Deep learning methods are evolving fast
- Deep learning add values for single-cell data analyses
- New opportunities to apply deep learning to extract more valuable insights from single-cell data
- Deep learning is not hard to learn and apply for practical purposes

Acknowledgments

This file is for the educational purpose only. Some materials (including pictures and text) were taken from the Internet at the public domain.