Introduction of machine learning for single-cell data analysis

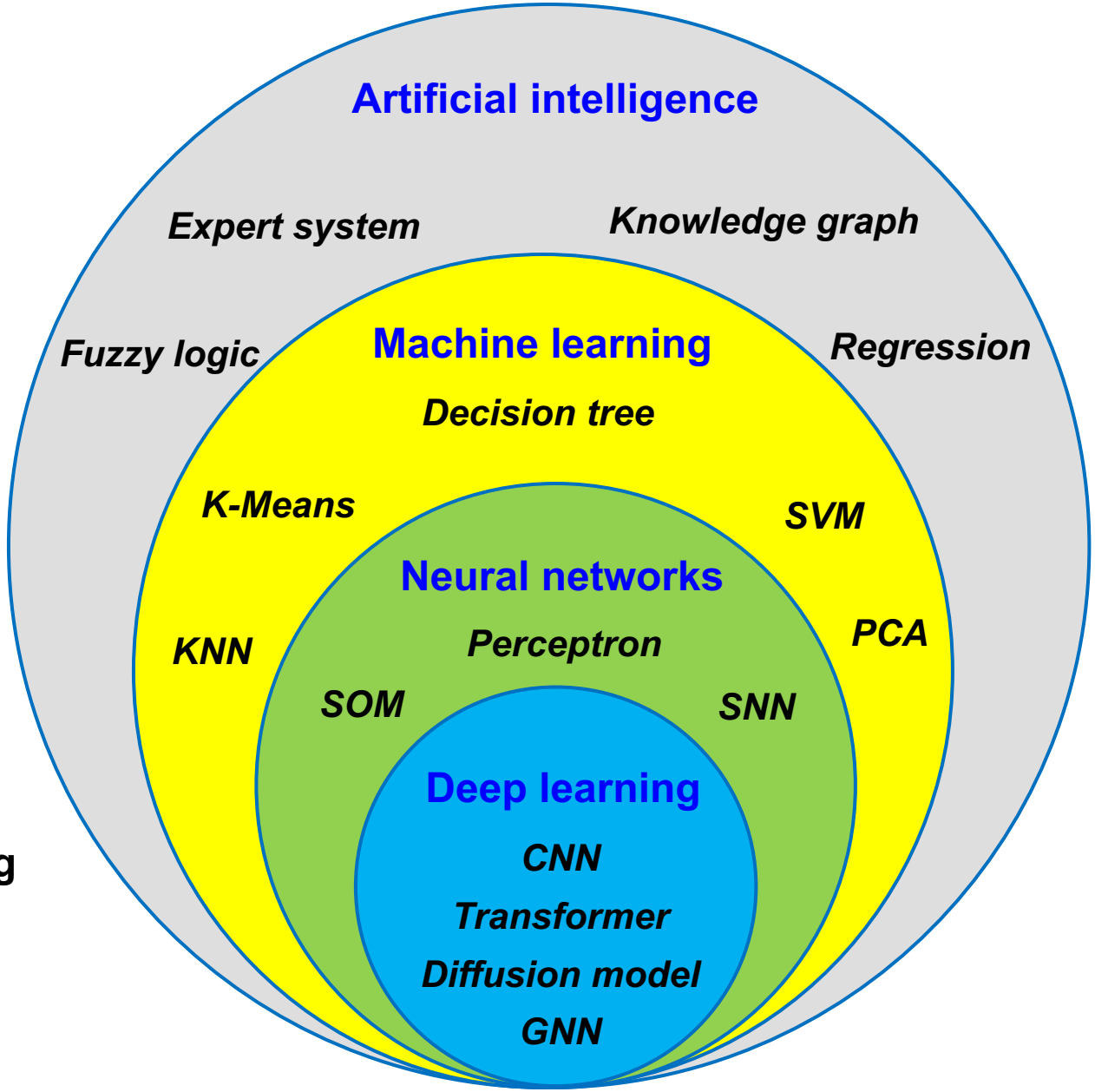
Dong Xu

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



- Unsupervised learning
- Supervised learning
- Self-supervised learning
- Reinforcement learning

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

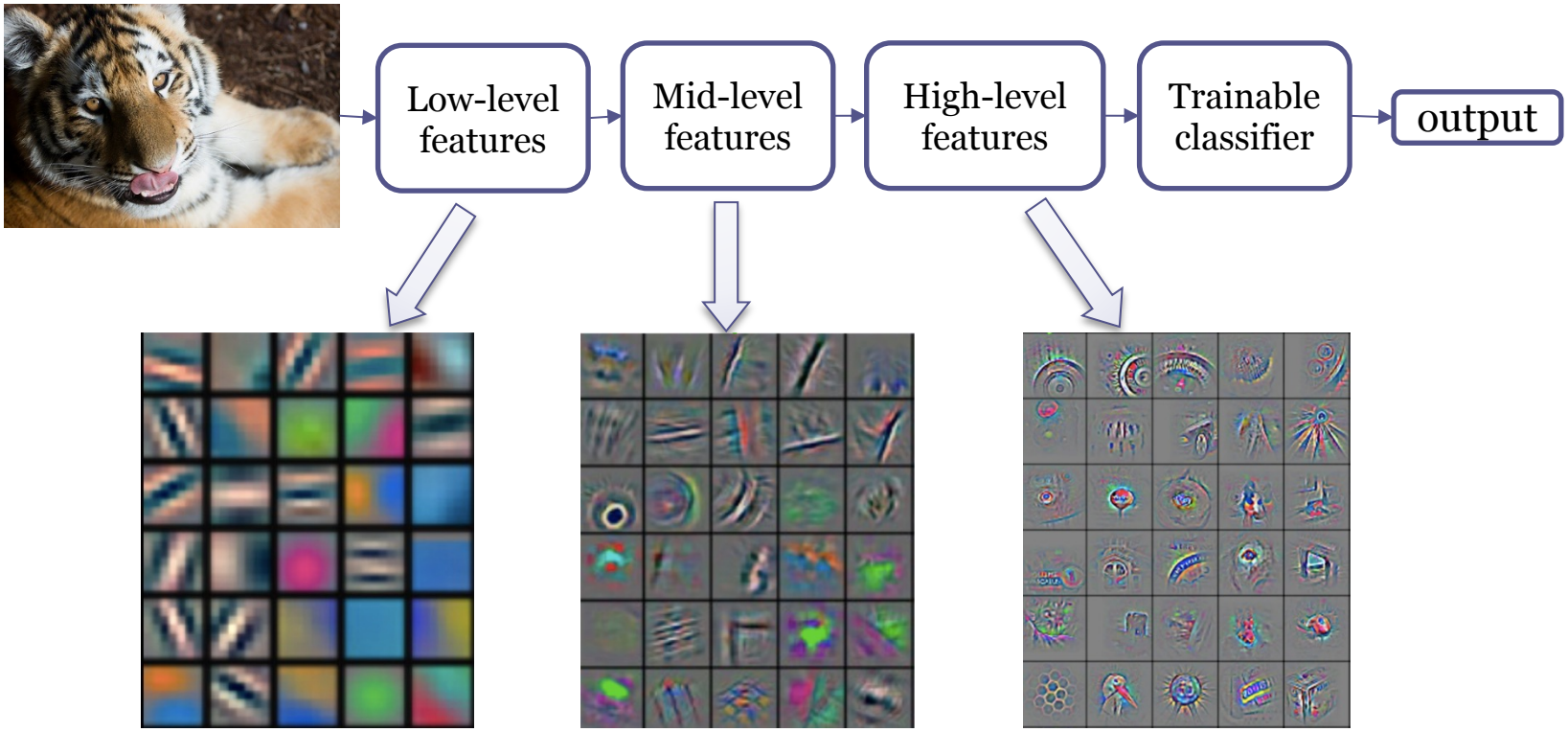
Generalization

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



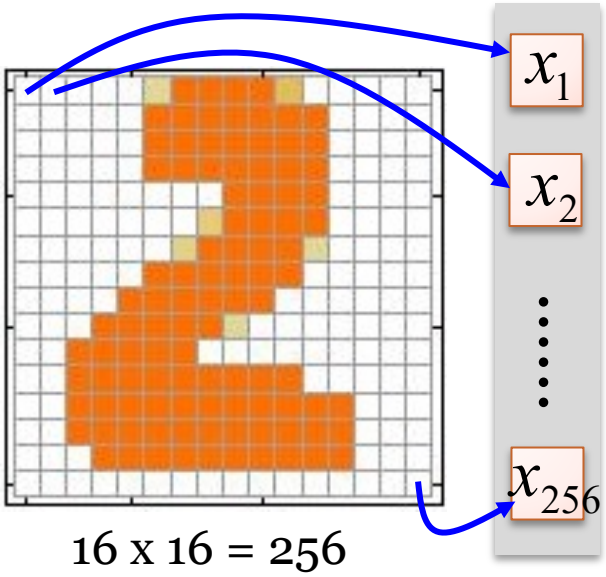
Hierarchy of representations with increasing level of abstraction

Deep learning

raw features, deep/sophisticated architectures, more data, more compute

Handwriting Digit Recognition

Input

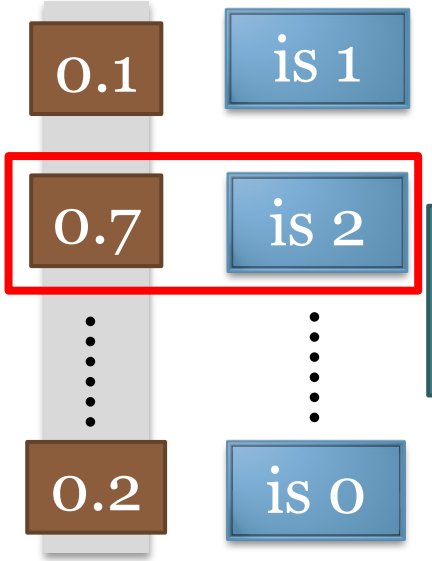


16 x 16 = 256

Ink → 1

No ink → 0

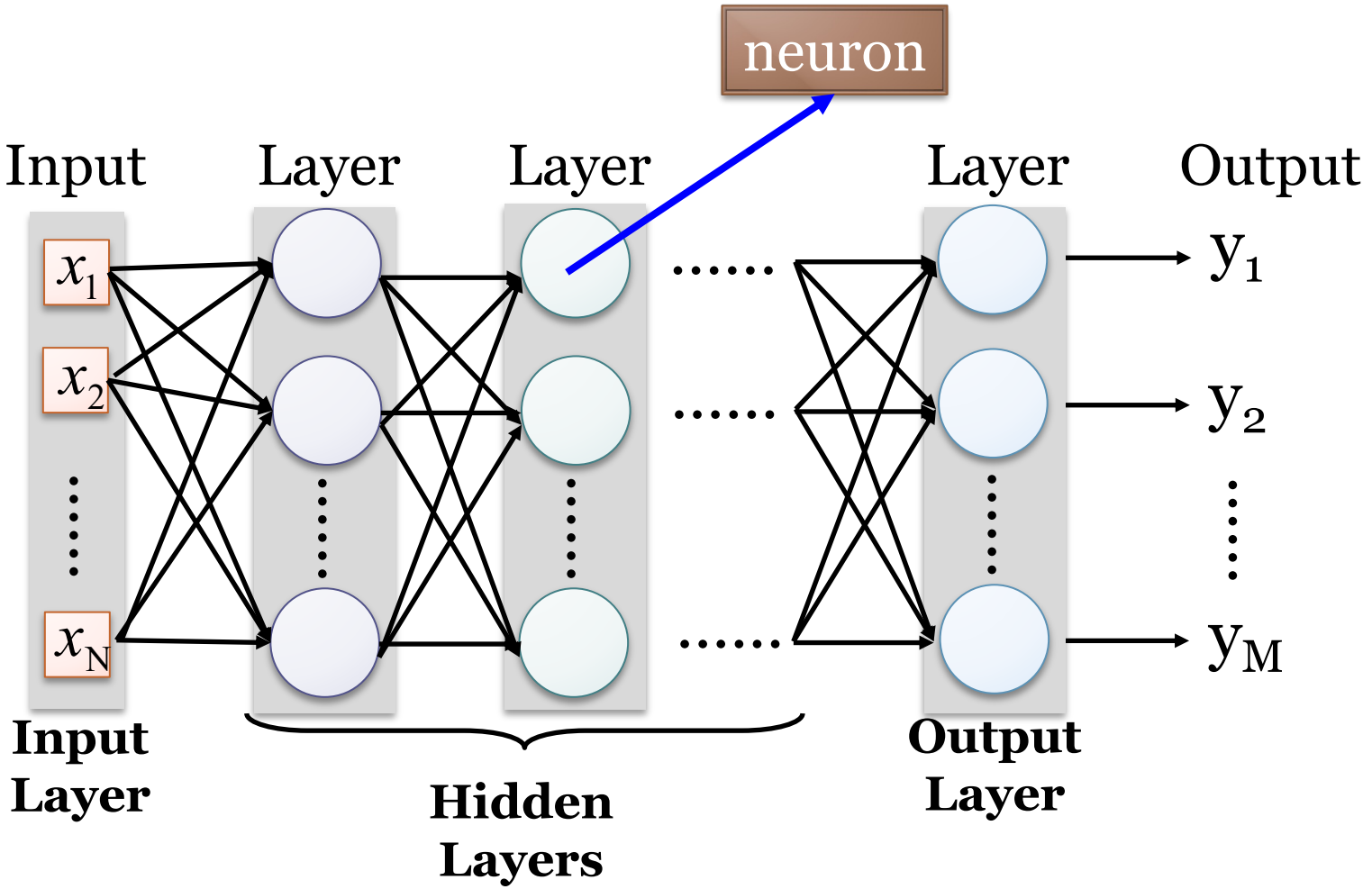
Output



The image is "2"

Each dimension represents the confidence of a digit.

Deep Neural Network (DNN)

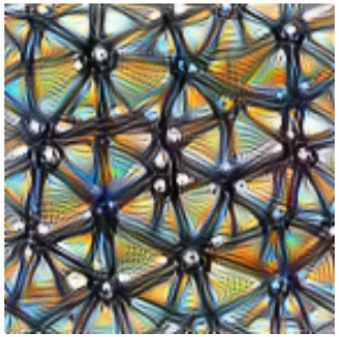


Deep means many hidden layers

Latent Representation

Neuron

`layer`$_n$`[x,y,z]`

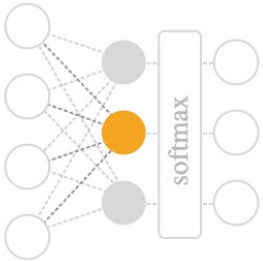


Channel

`layer`$_n$`[:,:,z]`

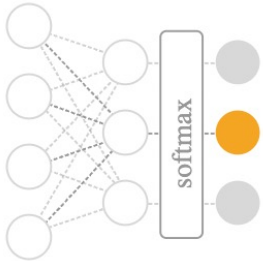
Layer/DeepDream

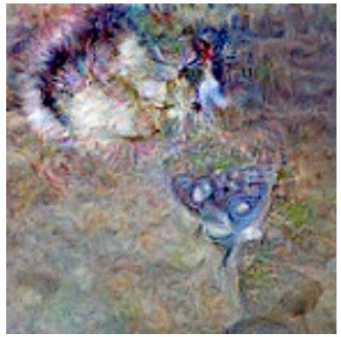
`layer`$_n$`[:,:,:]`^{2}



Class Logits

`pre_softmax[k]`





Class Probability

`softmax[k]`

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

Abstraction and Representation

physical space → latent/embedding space (manifold)

photo

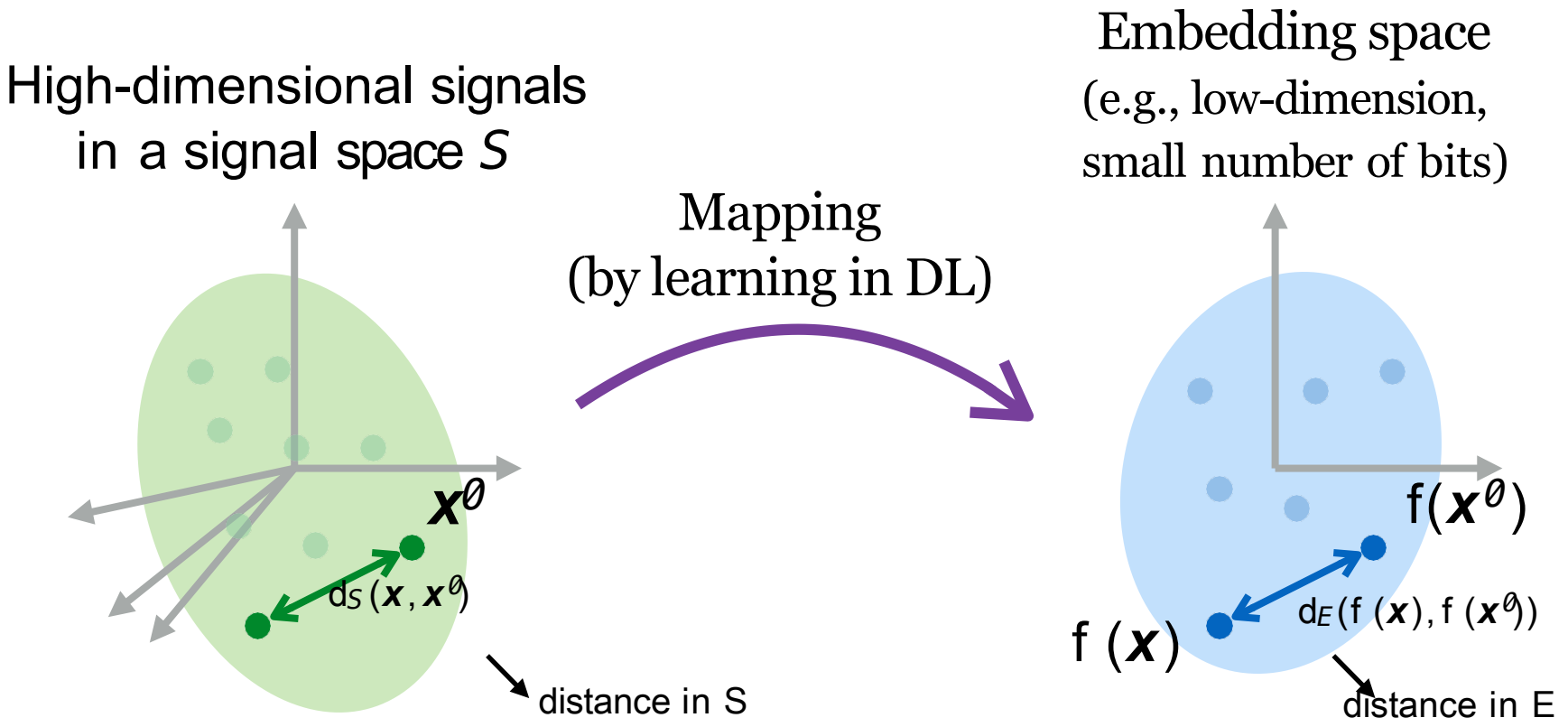
impressionism

expressionism

cubism

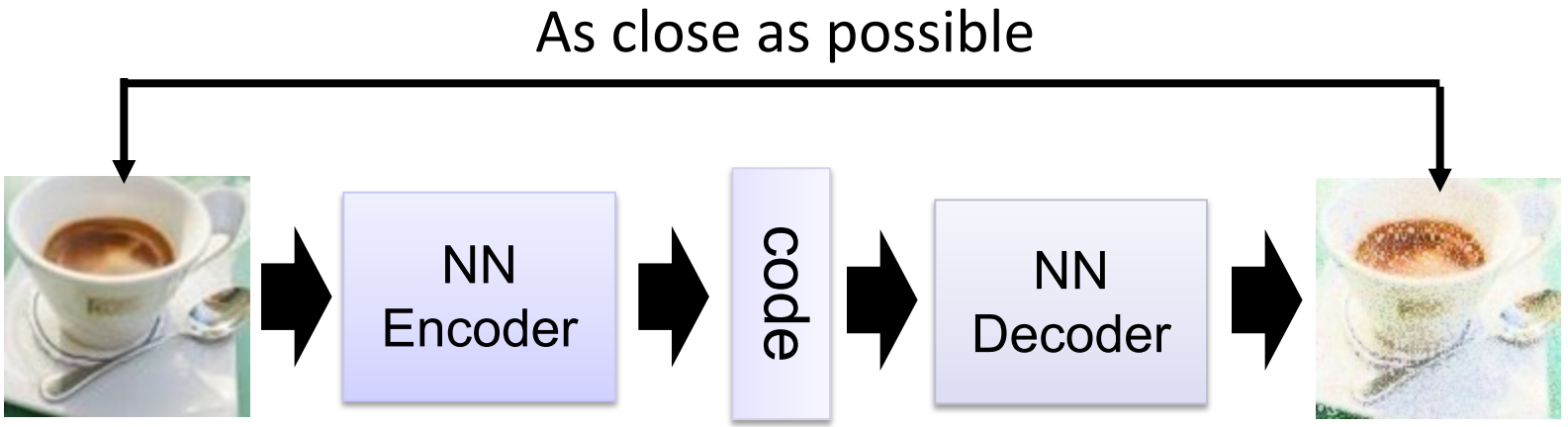
abstract expressionism

Unsupervised Learning & Embedding



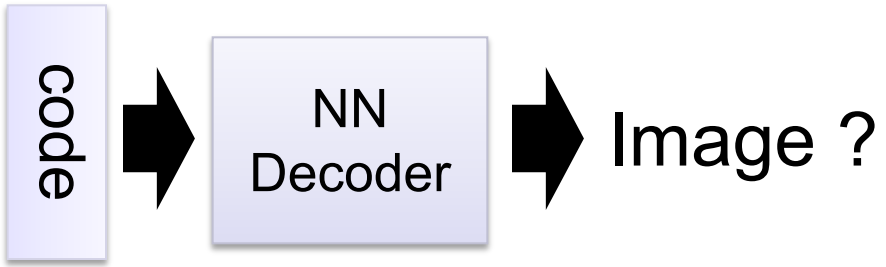
An embedding is a function from an original space to an embedding space that preserves aspects of the geometry of the original space

Autoencoder

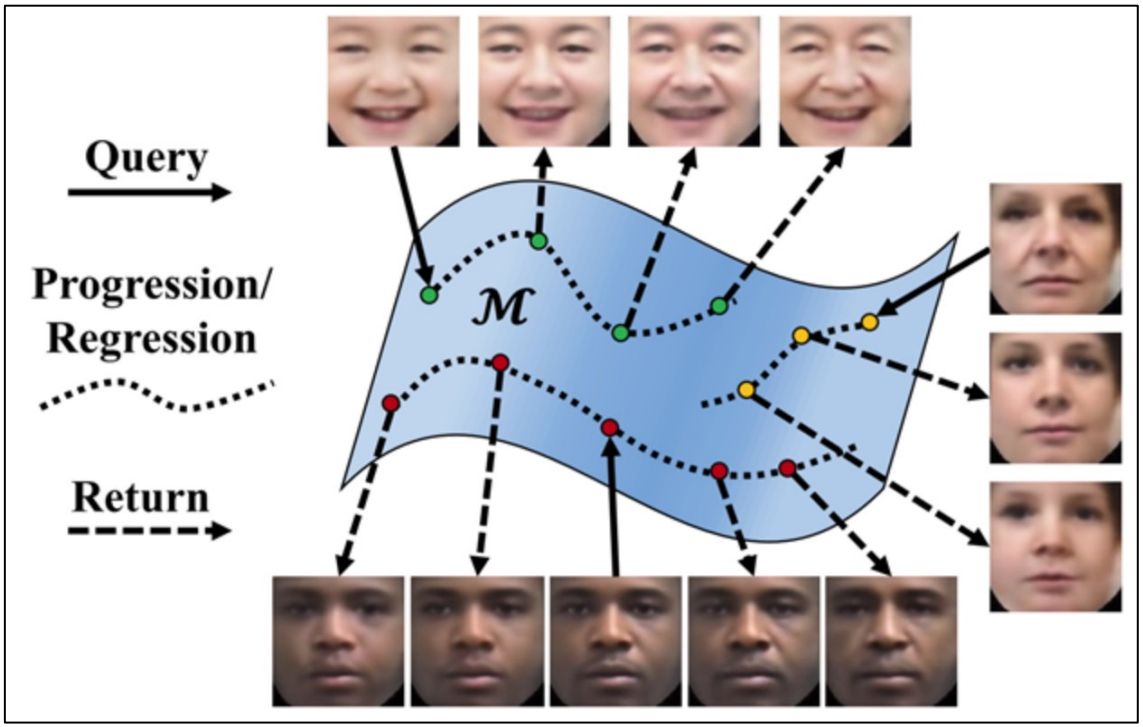


A type of unsupervised learning which discovers generic features of the data (**learn data patterns**)

Randomly generate a vector as code



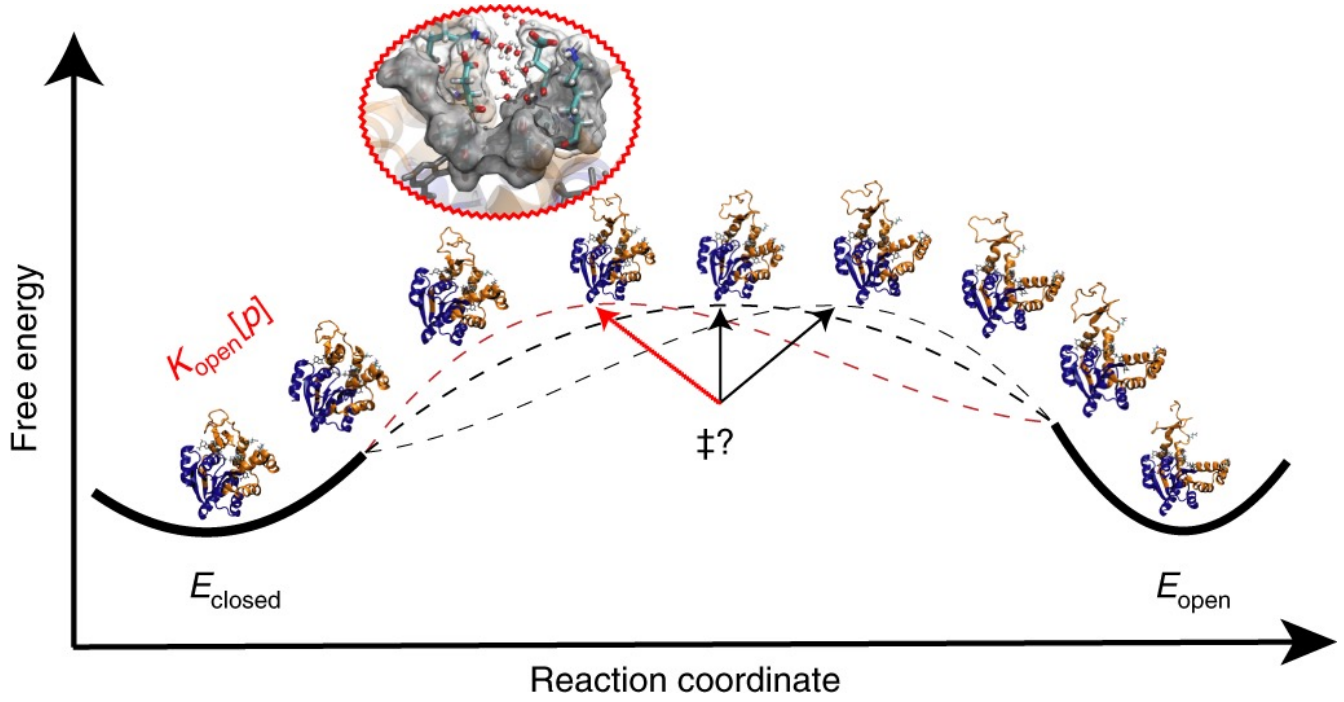
Manifold



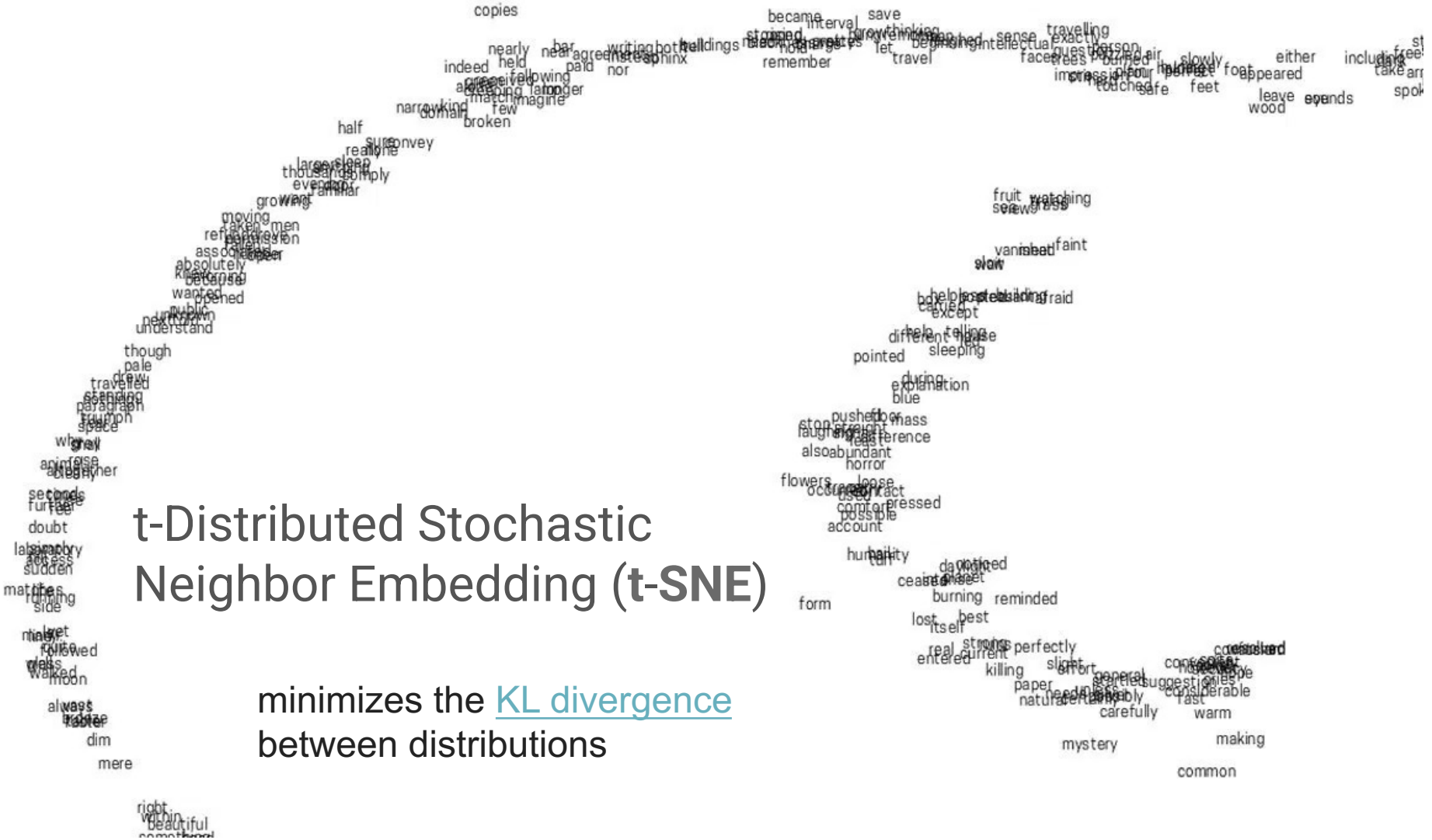
- A Manifold is a topological space that locally resembles Euclidean space near each point
- n-dim manifold → topological space M, every point x ∈ M has a neighbor homeomorphic (isomorphismic) to Euclidean space Rn

Manifold Hypothesis

- **Deep Learning Central Hypothesis**: Data concentrates around a low-dim manifold (**relevant dimension**)
- Mimic human learning



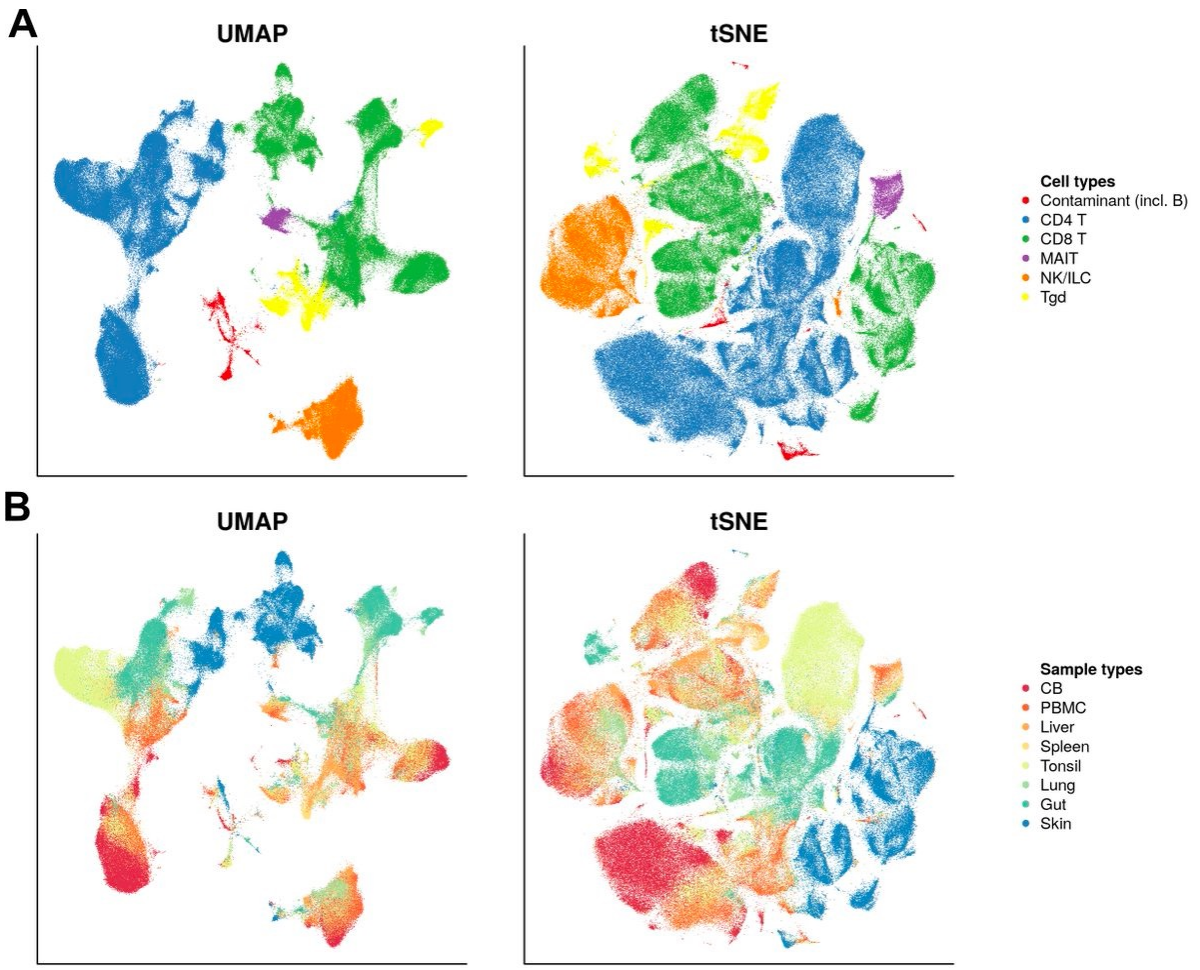
T-SNE Visualization



U-Map

Uniform Manifold Approximation and Projection (UMAP)

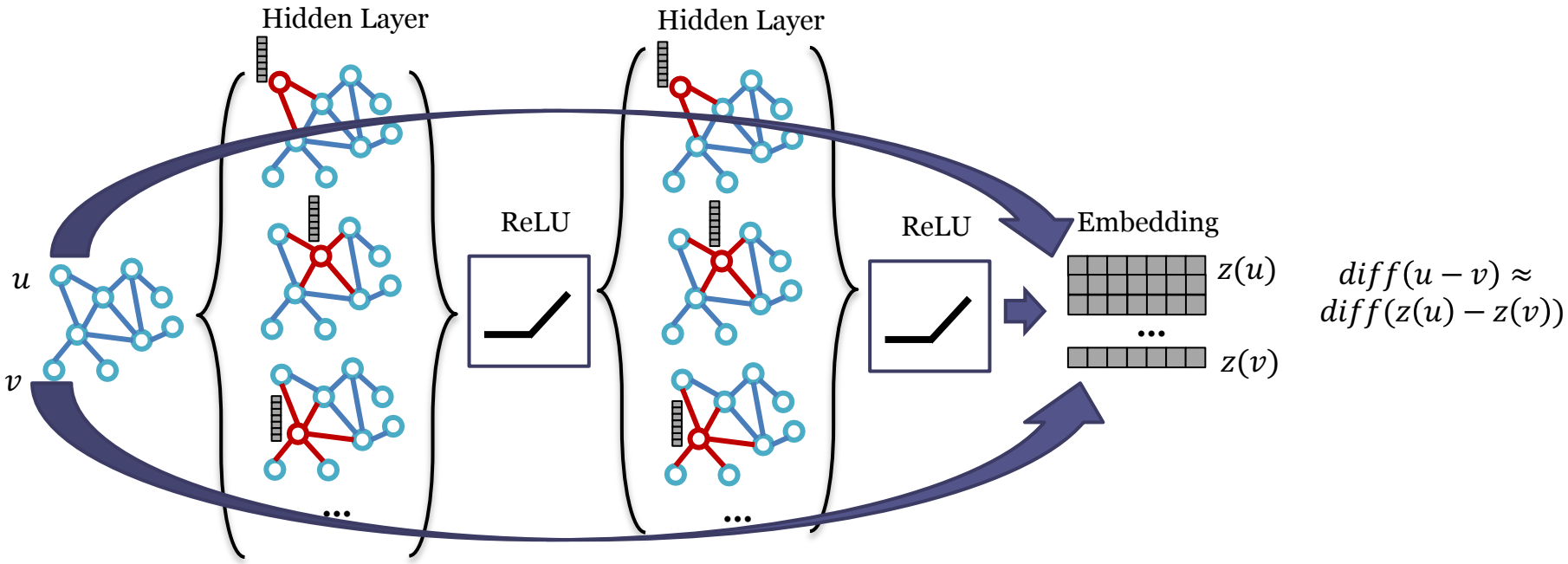
find a topological representation of the data in a lower dimensional space through manifold learning technique



https://github.com/lmcinnes/umap

Graph Neural Network (GNN)

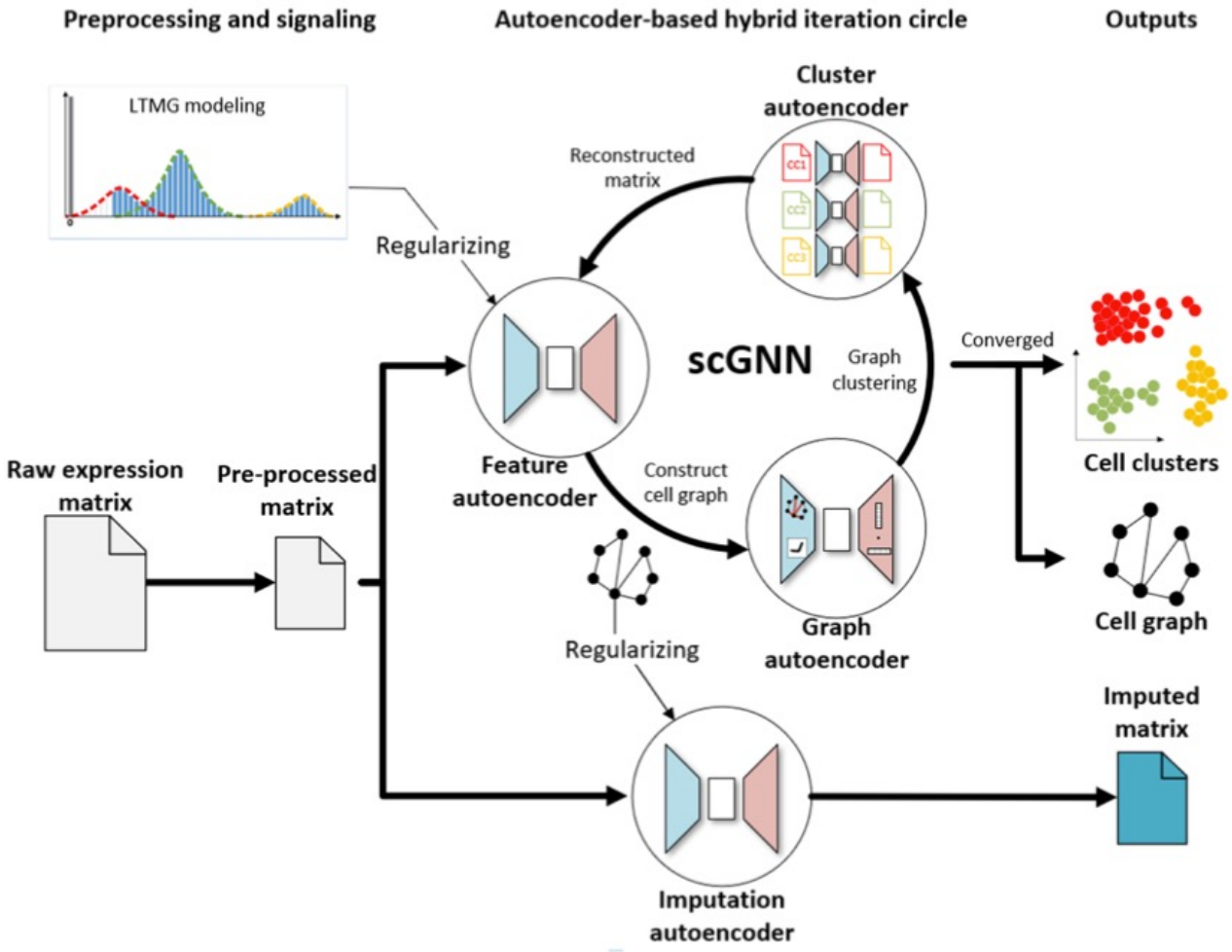
- GNN deconvolutes node relationships through neighbor information propagation in a deep learning architecture.
- Generate node embeddings based on local neighborhoods



Application of Graph Neural Network in Single-cell Analysis

Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, Wang C, Fu H, Ma Q, Xu D. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nature Communications 12, 1-11, 2021.

Cited >500 times



Foundation Model Era

- Machine learning paradigms

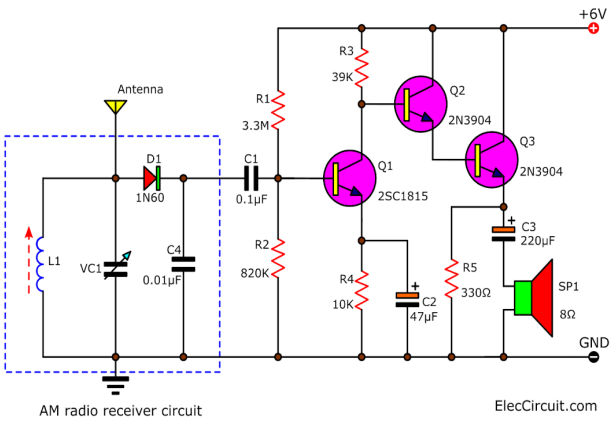
  Feature engineering: manual feature extraction (SVM, LightGBM, XGBoost)

  Architecture engineering: raw features, design deep network (CNN, LSTM)

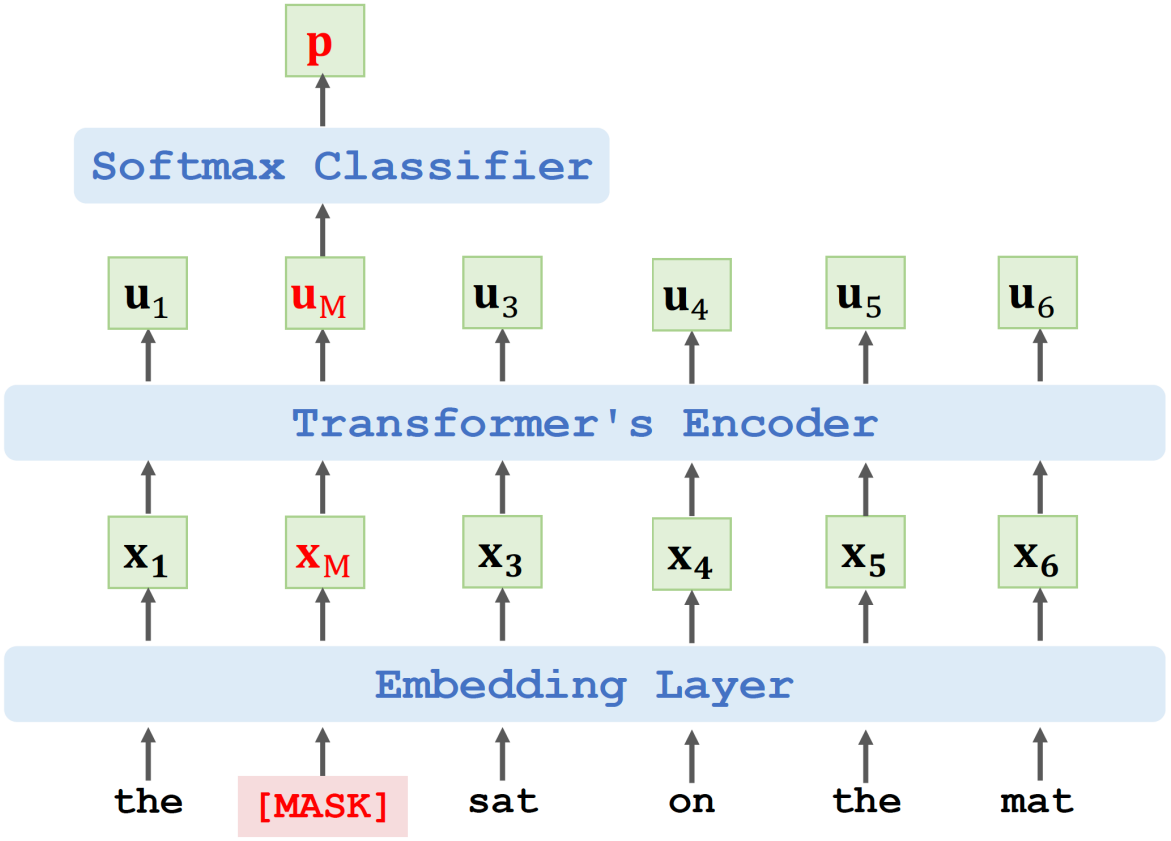
  Objective engineering: pre-train large model and fine-tune it (ResNet50, Bert)

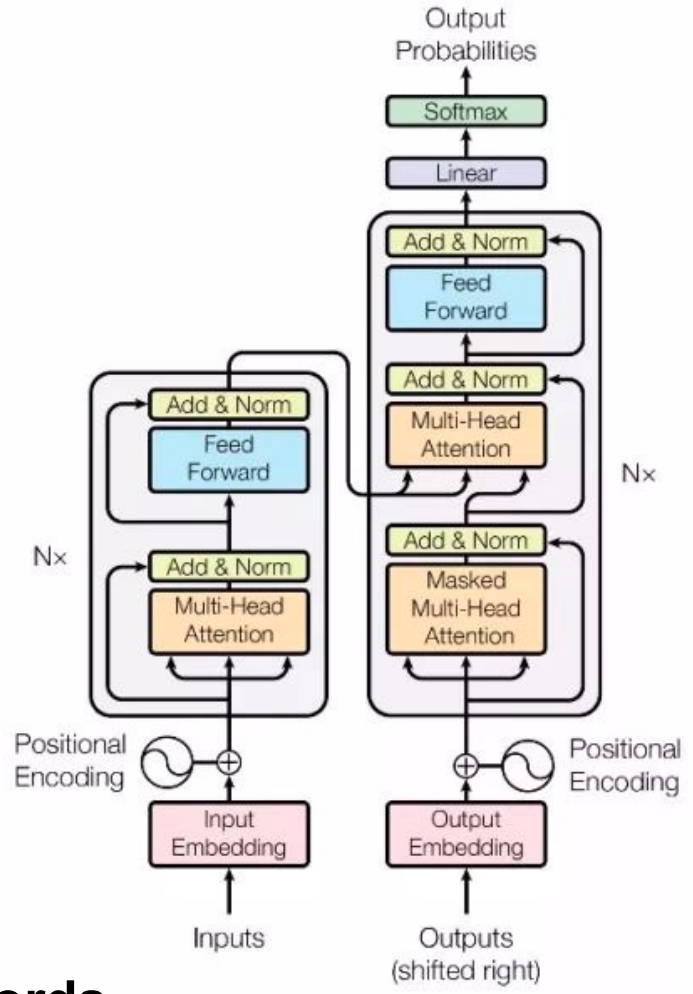
  Prompt engineering: prompt **foundation model** in zero/few shots

- Industrial era of artificial intelligence



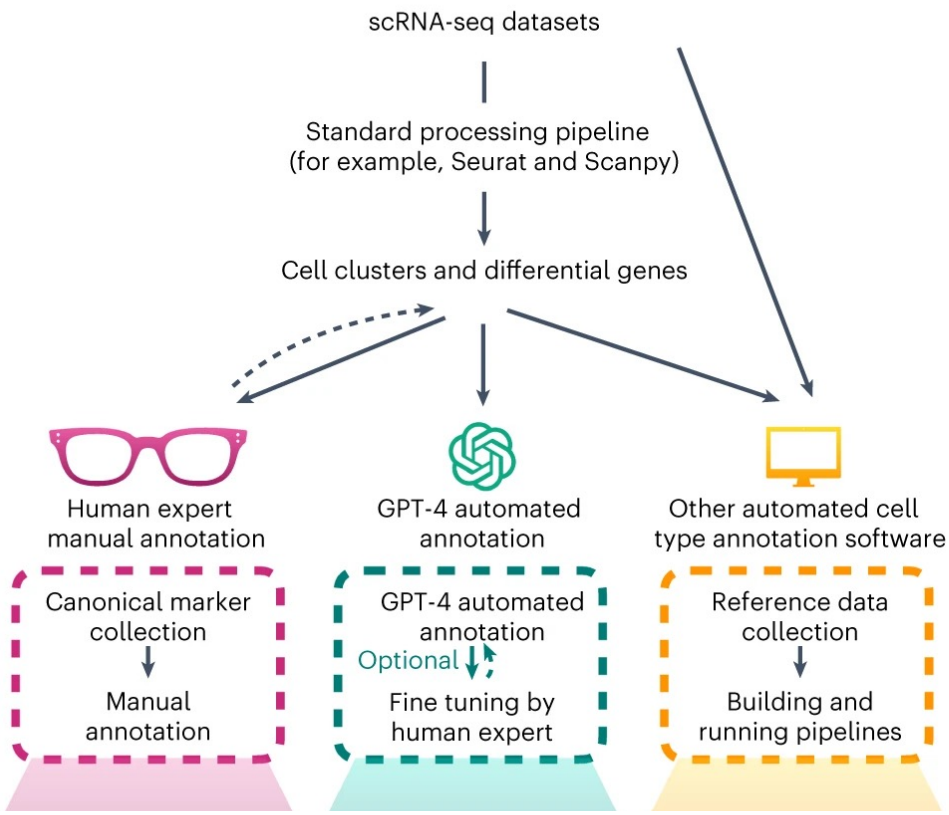
From Transformer to GPT





Self-supervised learning to predict missing words

GPT-4's Cell Type Annotation



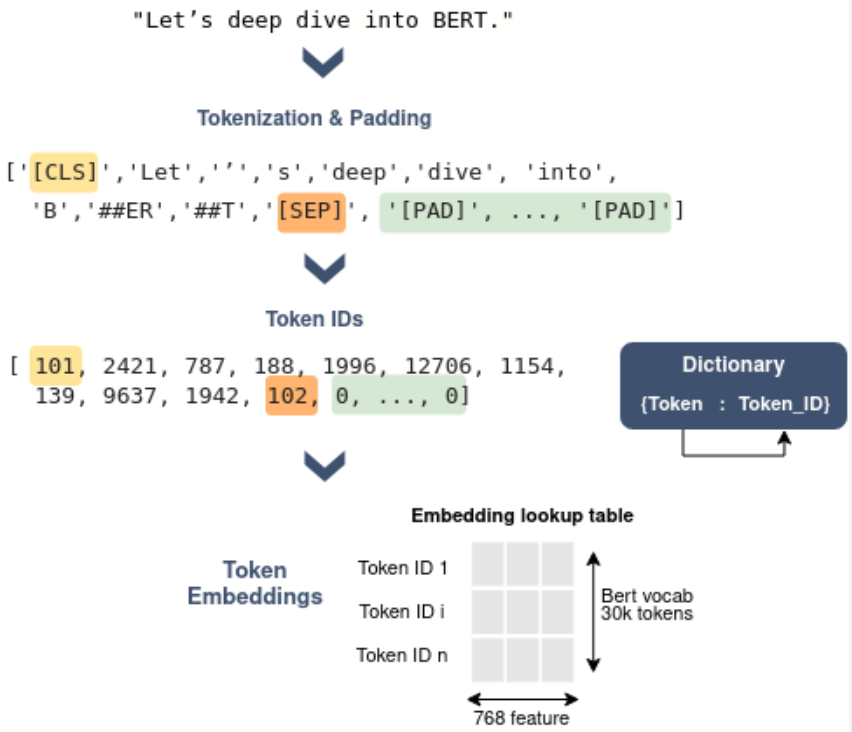
Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name.

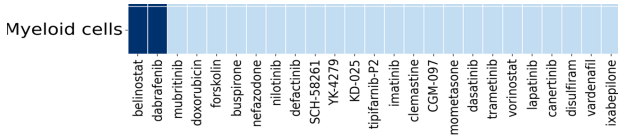
CR2, CD24, FAS, CXCR3, CD1c
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROXl, LYVE1
TPSAB1, FCER1A, TPSB2, KIT, CD69, HDC
ACTA2, MY01B, ACTA2, ANPEP, DES, MCAM, PDGFRB, CSPG4

1. Dendritic cells
2. Luminal epithelial cells
3. Lymphatic endothelial cells
4. Mast cells
5. Pericytes

Hou, W., Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat Methods 21, 1462–1465 (2024).

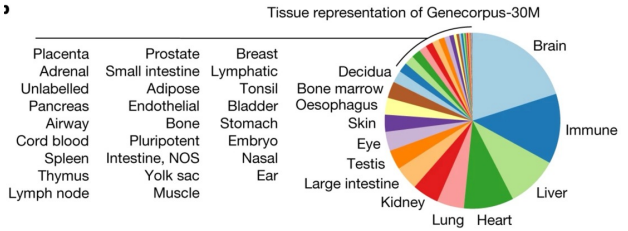
From LLM to Single-cell LLM





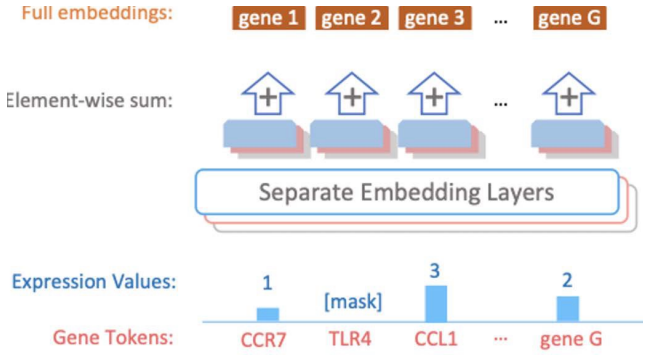
Expression profile of a single cell

Tokenization



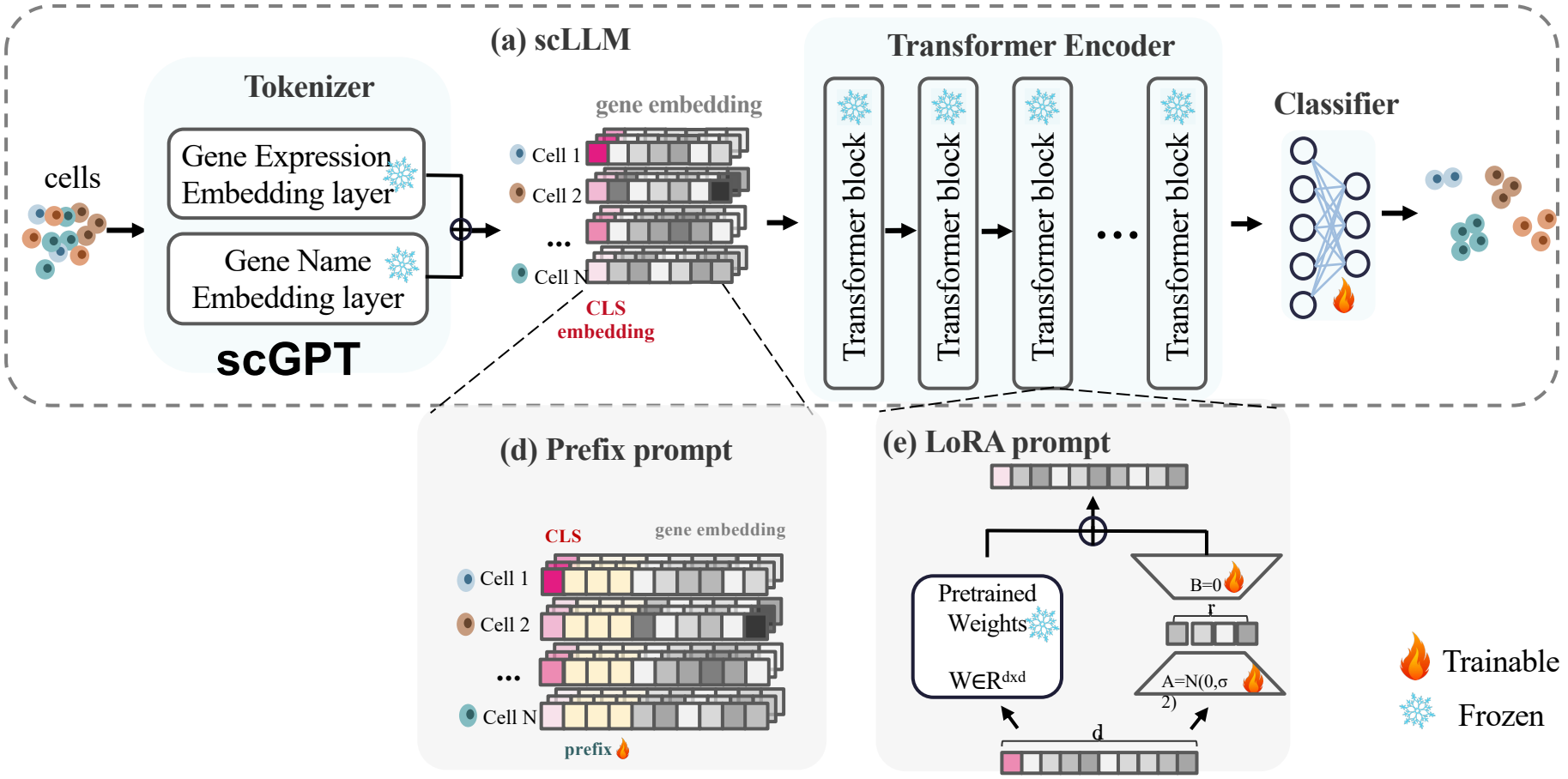
Gene dictionary/ vocabulary

Gene IDs



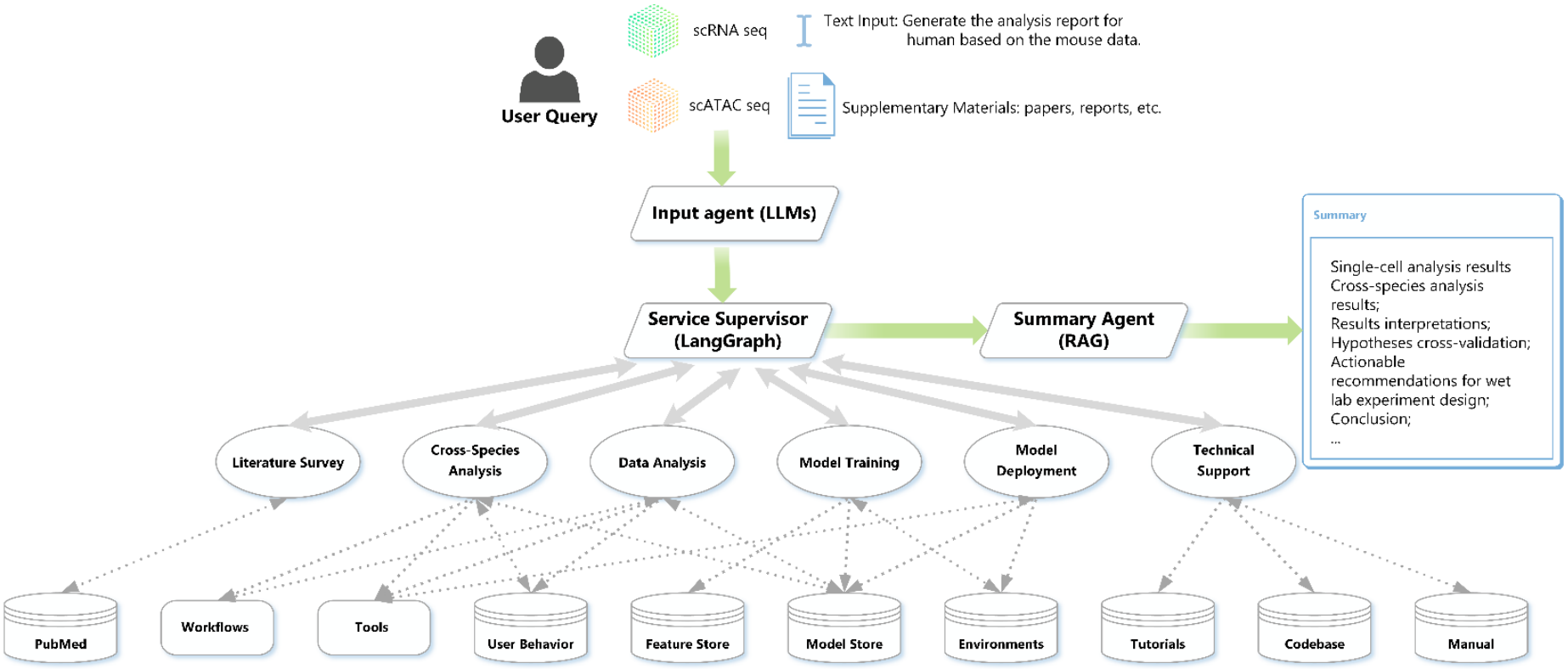
Embed gene expression values or gene expression order in each cell

Prompt-based Learning on scLLMs



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

AI Agent for Single-Cell Analysis



AI agent: autonomous intelligent system performing specific tasks without human intervention

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.