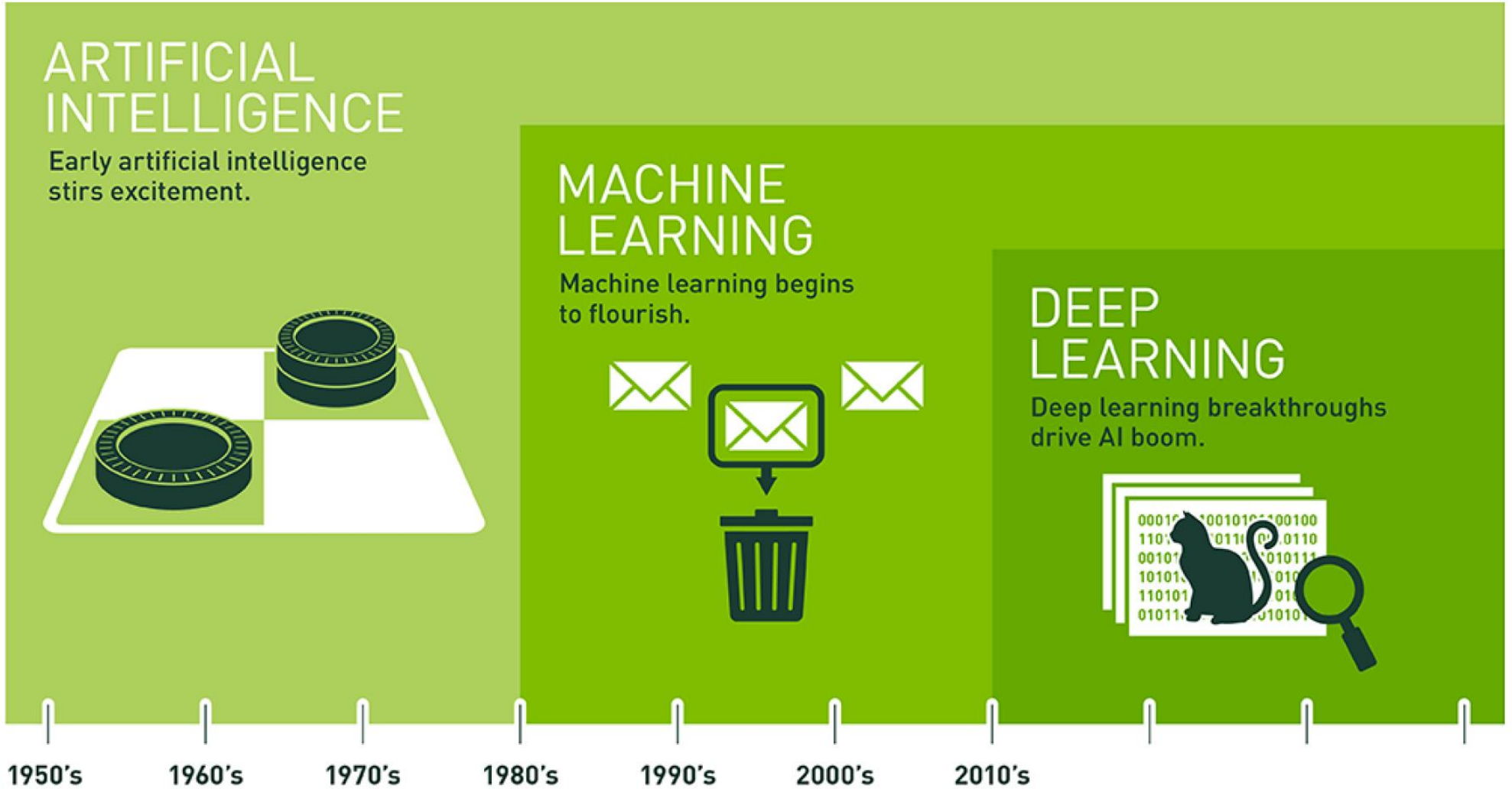
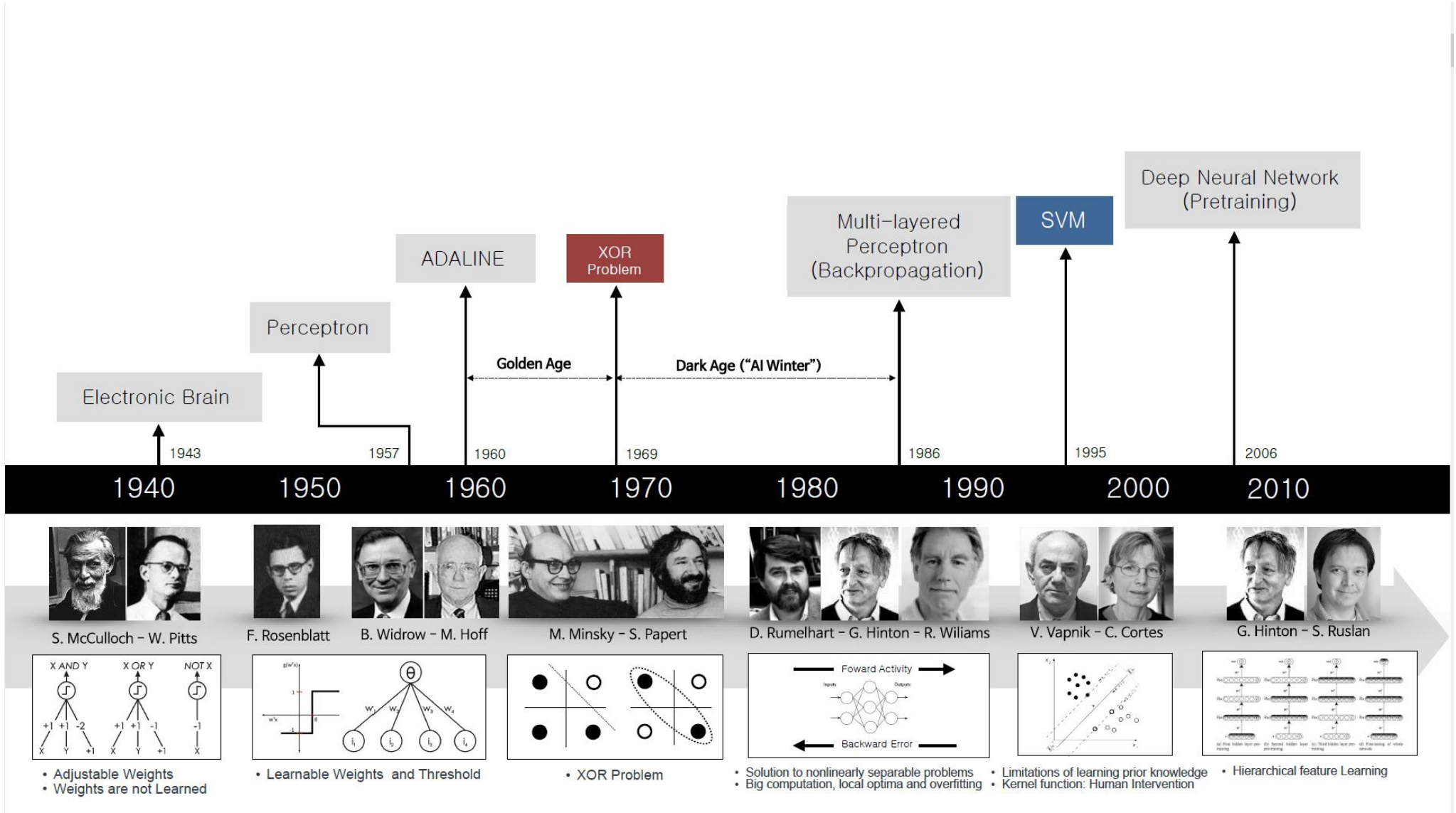
第一章 生物大模型简介

人工智能的发展历程



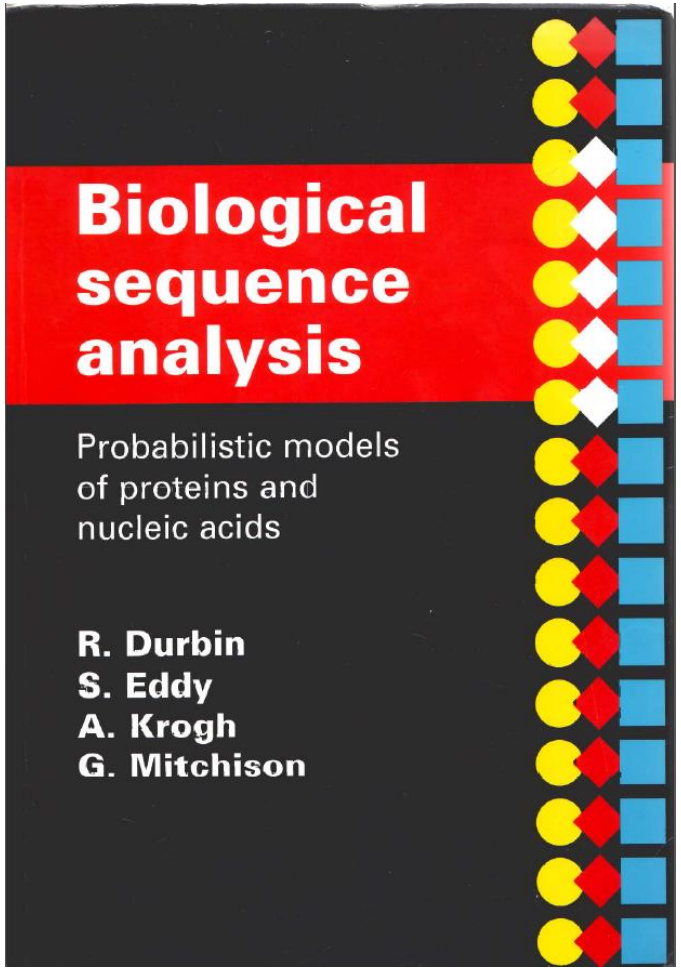
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

人工智能发展的里程碑事件



《人工智能简史》，尼克 著

经典生物信息学/生物信息学 1.0



At a Snowbird conference on neural nets in 1992, David Haussler and his colleagues at UC Santa Cruz (including one of us, AK) described preliminary results on modelling protein sequence multiple alignments with probabilistic models called 'hidden Markov models' (HMMs). Copies of their technical report were widely circulated. Some of them found their way to the MRC Laboratory of Molecular Biology in Cambridge, where RD and GJM were just switching research interests from neural modelling to computational genome sequence analysis, and where SRE had arrived as a new postdoctoral student with a background in experimental molecular genetics and an interest in computational analysis. AK later also came to Cambridge for a year.

All of us quickly adopted the ideas of probabilistic modelling. We were per-

Richard Durbin

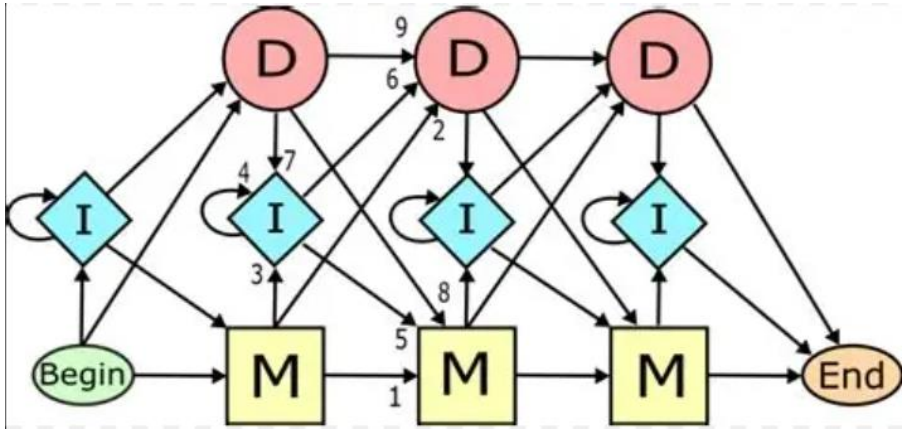
Sean Eddy

Anders Krogh

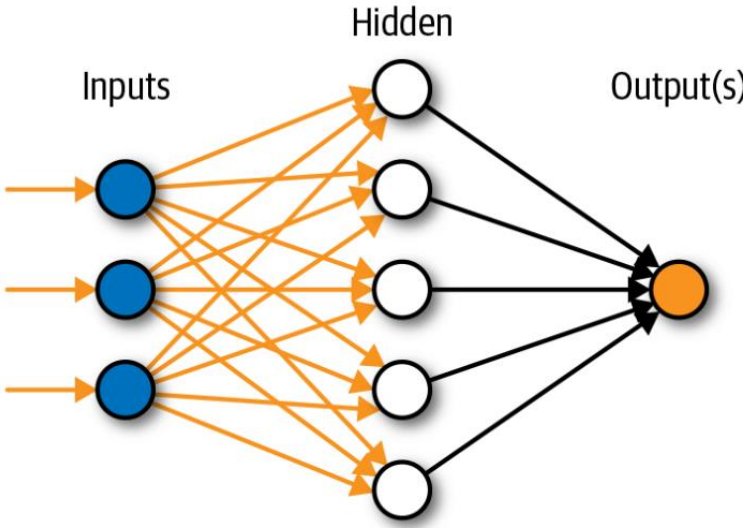
Graeme Mitchison

经典机器学习

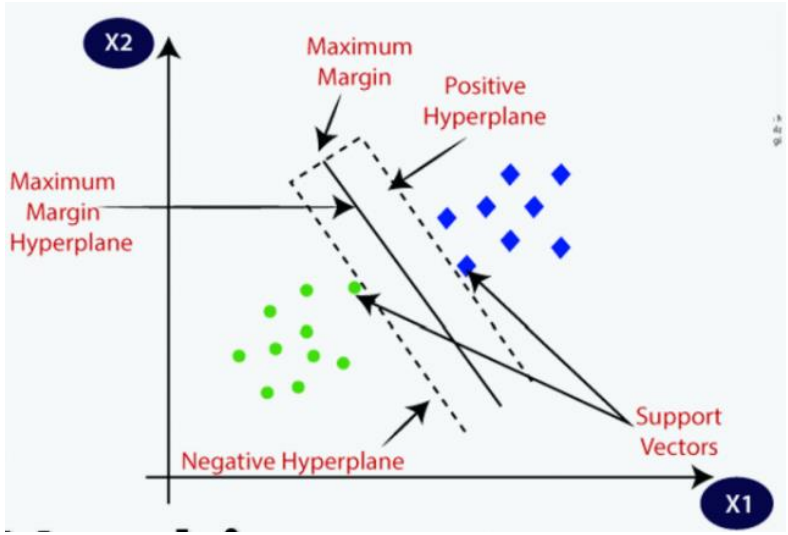
隐马尔科夫模型



人工神经网络

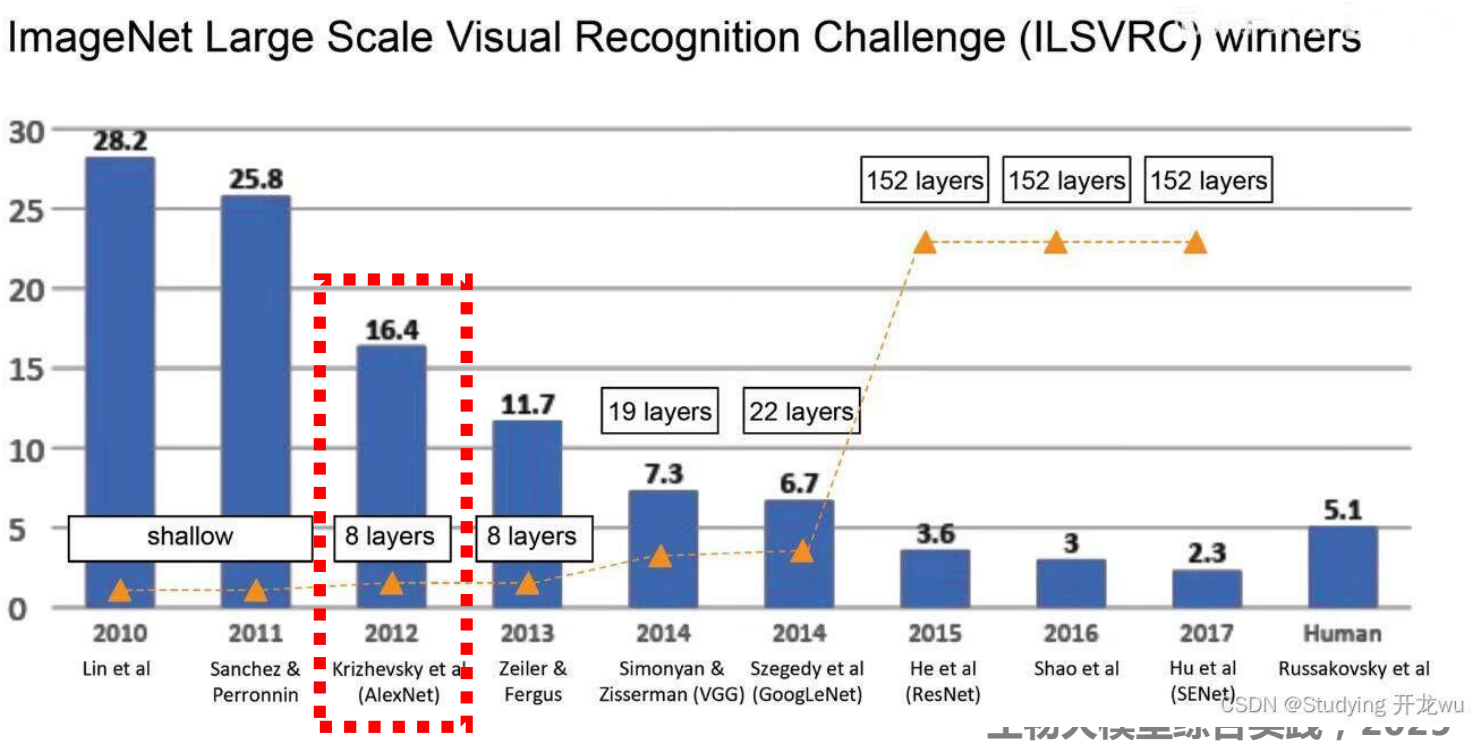


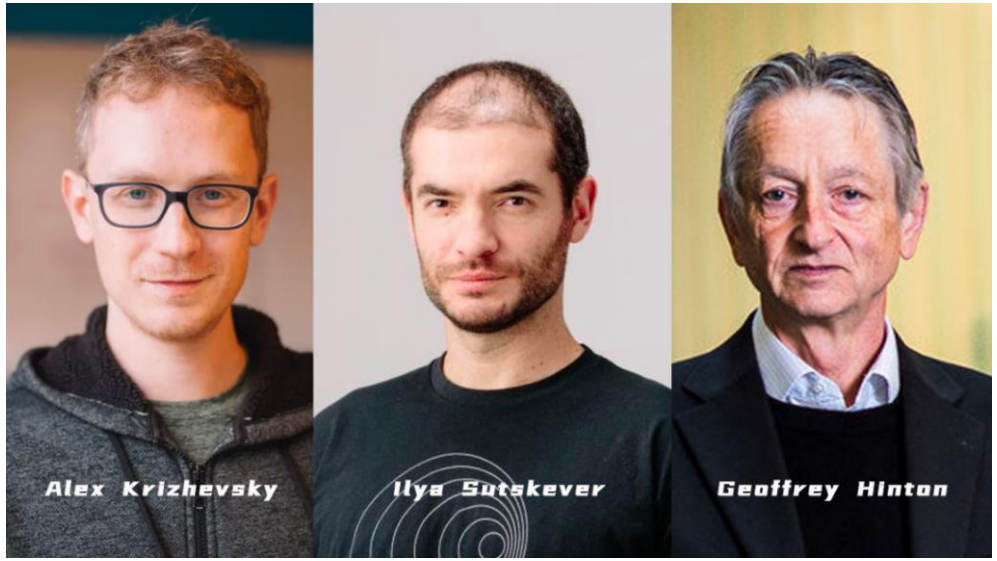
支持向量机



技术突破：卷积神经网络

- Geoffrey Hinton：2001-2014，多伦多大学计算机科学系教授
- 2012 年 12 月，AlexNet，两个天才学生Alex Krizhevsky 、Ilya Sutskever —— “以学生为中心”
- 将图像识别的错误率从>25%首次降低到<20%



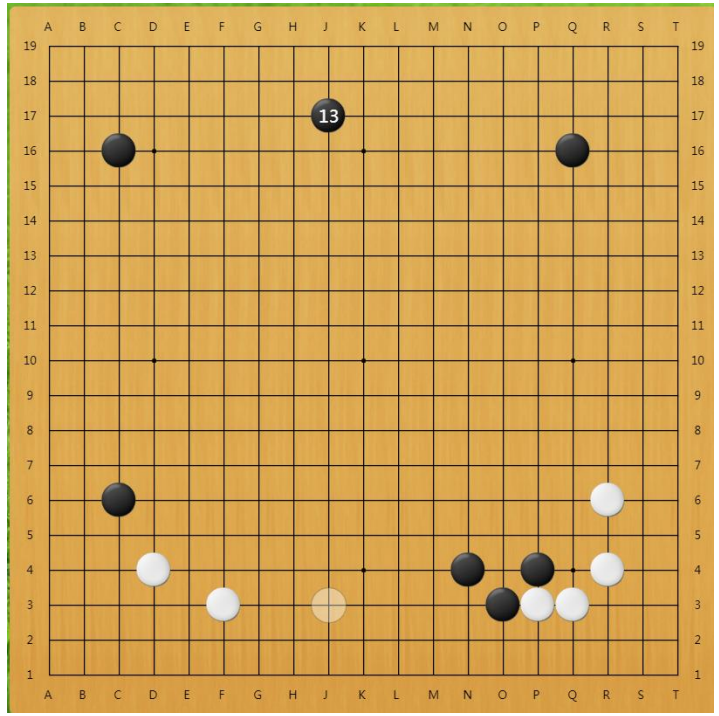


技术突破：围棋 & 蛋白质结构预测

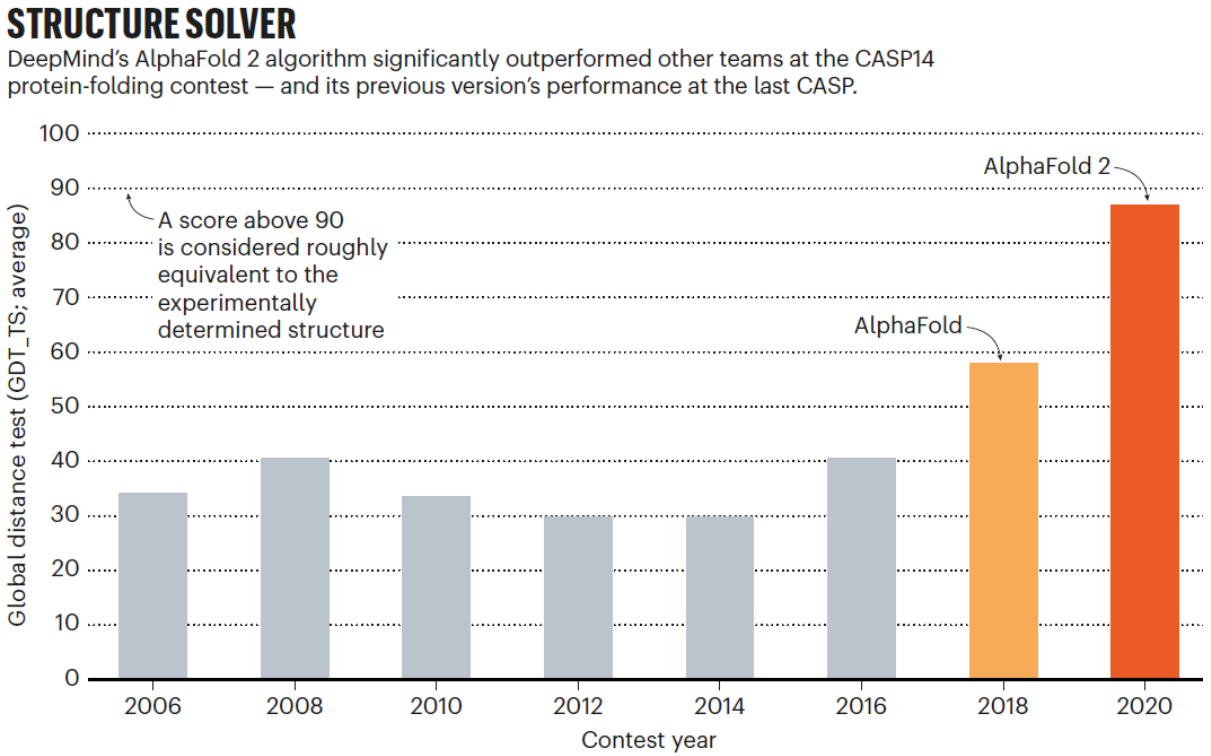
围棋 AlphaGo：战胜人类围棋世界冠军

蛋白质 AlphaFold 2：蛋白质三级结构预测的准确性媲美实验手段

2016.3.9，李世石 *vs.* AlphaGo



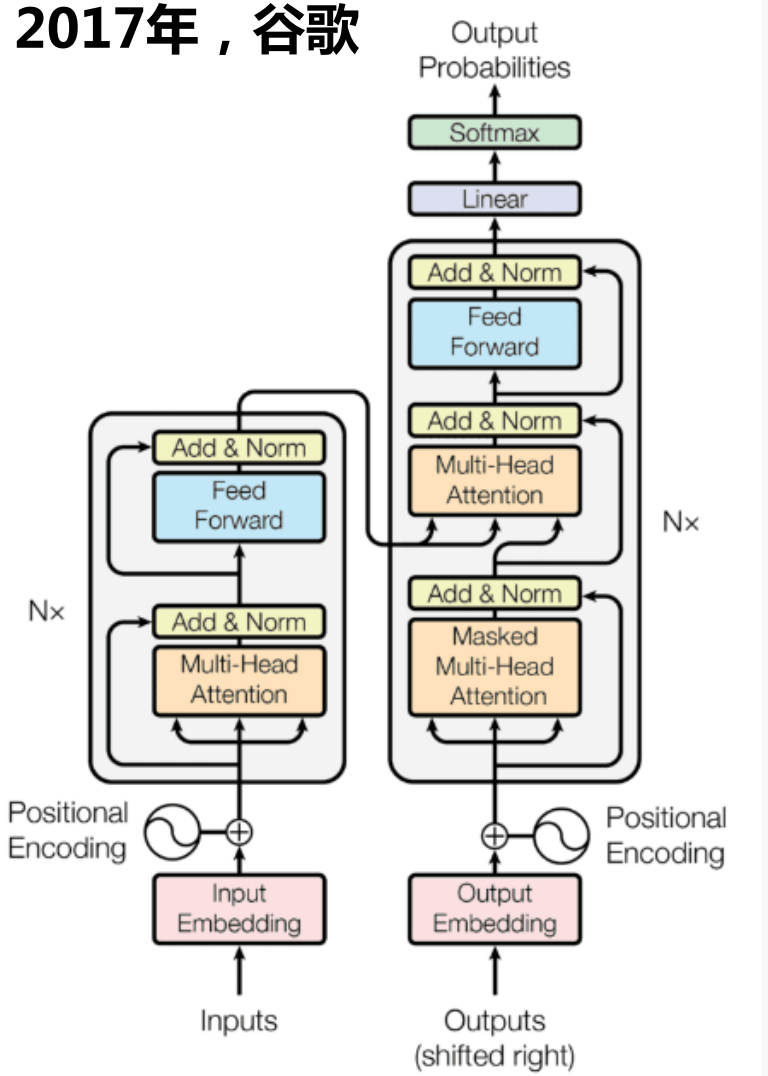
Nature, 2016, 529, 484-9



Nature, 2020, 577, 706-710

Transformer框架

2017年，谷歌

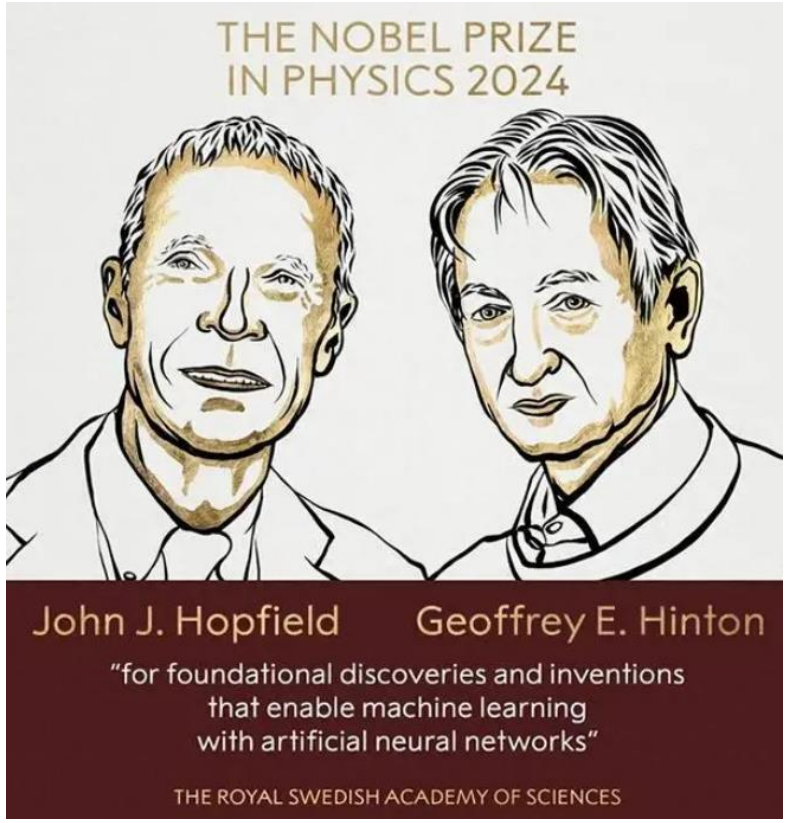


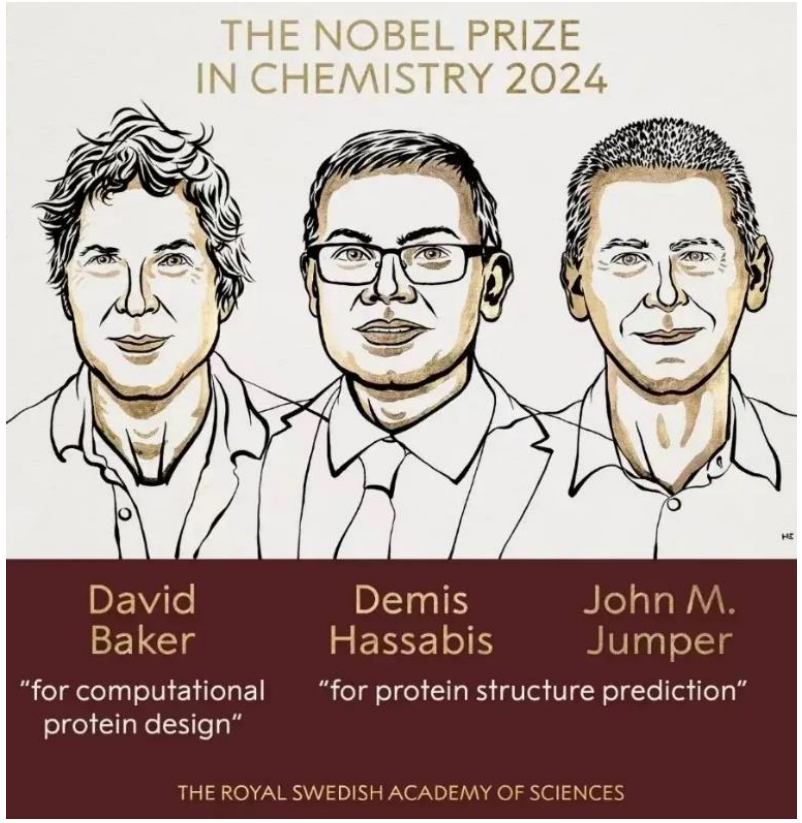
The encoder-decoder structure of the Transformer architecture
Taken from "Attention Is All You Need"

- 大语言模型的哲学理念
  - “压缩即智能”
- 无监督学习
  - 通过压缩来理解数据的结构
  - 压缩越有效，越能找到数据共有的结构
  - 好的压缩器也是好的预测器

Ilya Sutskever

2024年诺贝尔物理学奖 & 化学奖





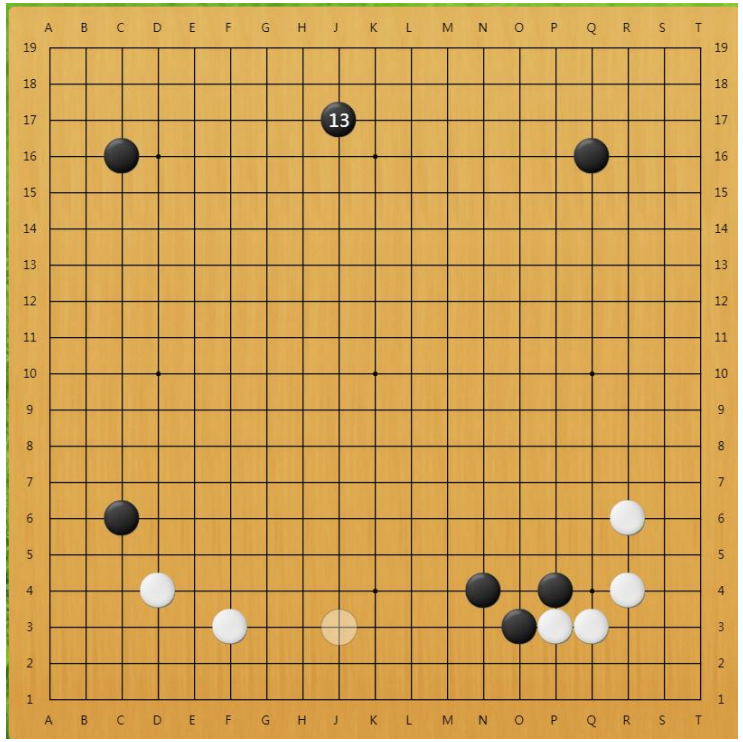
“基于人工神经网络实现机器学习的基础性发现和发明”

“计算蛋白质设计”、“蛋白质结构预测”

范式变革：人工智能 + 生物学

- 2019.10.23，张辰宇，人工智能生物学（AIBIO）
- 2022.8.16，中国生物物理学会人工智能生物学分会
- 徐涛院士，生物物理所：**AI就是物理！**

2016.3.9，李世石 *vs.* AlphaGo



中国科学: 生命科学 2022 年 第 52 卷 第 3 期: 291 ~ 300
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
CrossMark click for updates
评 述 中国知名大学及研究院所专栏 南京大学生命科学学院专辑

人工智能生物学——生物学3.0

周祯[1,2], 闫超[1,2], 张辰宇[1,2*]

1. 南京大学生命科学学院, 医药生物技术国家重点实验室, 南京 210023;
2. 南京大学人工智能生物医药技术研究院, 南京 210031
* 联系人, E-mail: cyzhang@nju.edu.cn

收稿日期: 2021-08-10; 接受日期: 2022-01-27; 网络版发表日期: 2022-02-23

摘要 现代生物学一方面取得了重大进展, 另一方面也面临着巨大的发展瓶颈, 而伴随着人工智能(artificial intelligence, AI)理论和技术的快速发展, AI与生物学深度融合的生物学V3.0——人工智能生物学(artificial intelligence biology, AIBIO)已经呼之欲出. 人们将人工智能生物学定义为利用人工智能的原理和手段来研究生命系统基本规律的科学. 其研究特点是: 动态整合多层面与多因素, 从而真正理解生命现象中的分子间相互作用与相互调控的规律, 解决生命科学中的重大基本问题. 作为一个全新的生命科学学科, 人工智能生物学将全面提升生物学研究的高度, 革新生物学研究的现有范式, 拓展生物学研究的范围, 实现生命科学和医学科学关键领域的实质性突破. 及时掌控并引领相关领域的研究, 对推动生命科学领域的基础研究、技术发展, 甚至对整个社会的进步都至关重要.

生物大模型综

生物信息学 *vs.* AI生物学

Primer

https://doi.org/10.1038/s41587-024-02123-4

Designing proteins with language models

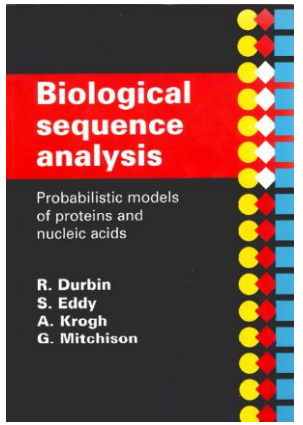
Foundations of protein language models

Fundamentally, protein language models aim to predict how likely we are to observe a particular protein sequence S given all the protein sequence data collected thus far. We denote a protein sequence $S = (s_1, s_2, \ldots s_N)$, where s_i represents the amino acid at position i in the sequence. As a first approximation, we might consider the probability of observing a protein as the joint probability of observing each of its constituent amino acids. Under this model, referred to as unigram, we calculate the probability of a sequence S as

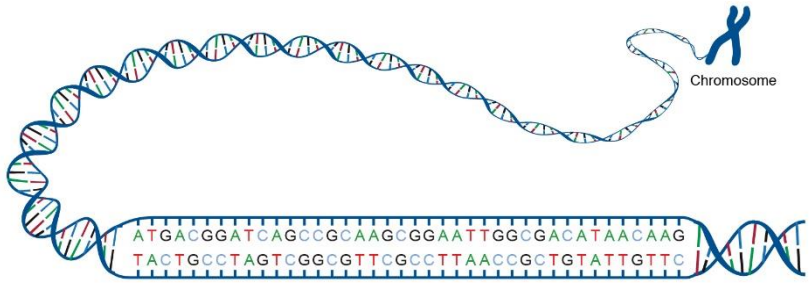
$$P(S) = \prod_i^N P(s_i)$$

In practice, to compute $P(S)$, we simply tabulate the frequency of each amino acid occurring in our sequence database and multiply the probabilities for the specific sequence S. However, proteins are not unordered collections of amino acids. Rather, the specific order in which we observe the amino acids is a critical determinant of structure and function. To capture this order dependency, we can use the preceding residues to inform the probability of the next amino acid. In an n-gram model, we multiply these contextualized probabilities to form the overall probability of the sequence:

$$P(S) = \prod_i^N P(s_i | s_{i-(n-1)}, \ldots, s_{i-1})$$



生物信息学（1.0）
“骰子模型”
线性、指令式、解析式



人工智能生物学（2.0）
“语言模型”
非线性、人机交互式、
生成式

Nat Biotechnol, 2024, 42, 200-202

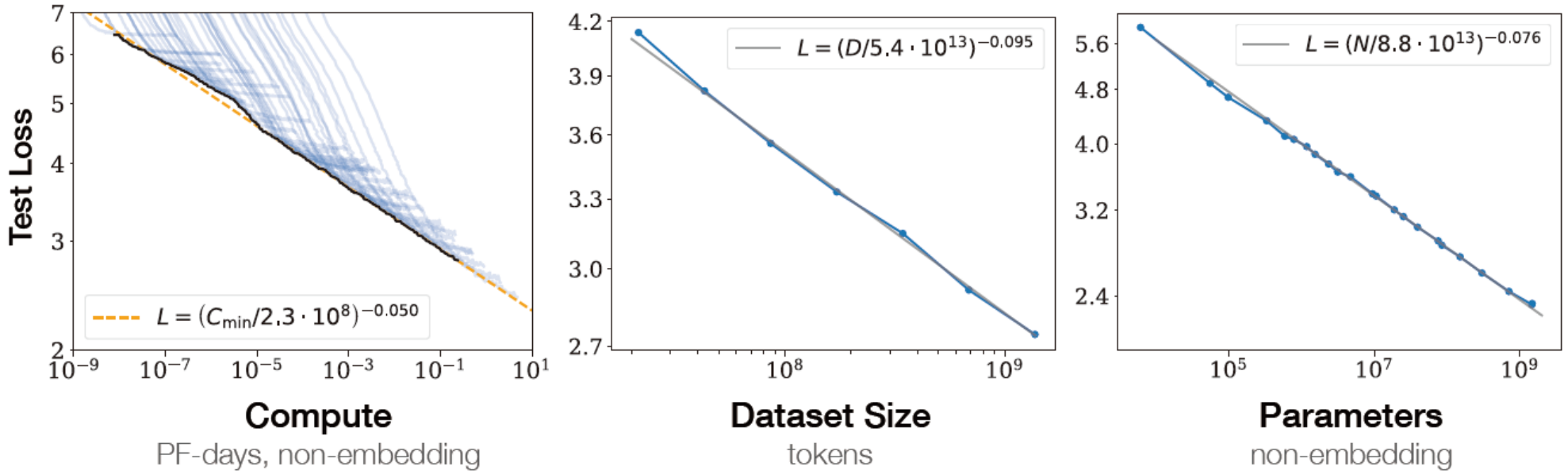
大语言模型的摩尔定律

- 尺度定律：“大力出奇迹”

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com



arXiv:2001.08361, 2020

大语言模型的涌现特征

- 语境学习（In-context learning）
- 小样本学习（Few-shot learning）
- 零样本学习（Zero-shot learning）
- 机器推理：思维链（Chain-of-thought, CoT）
- …

A Comprehensive Overview of Large Language Models

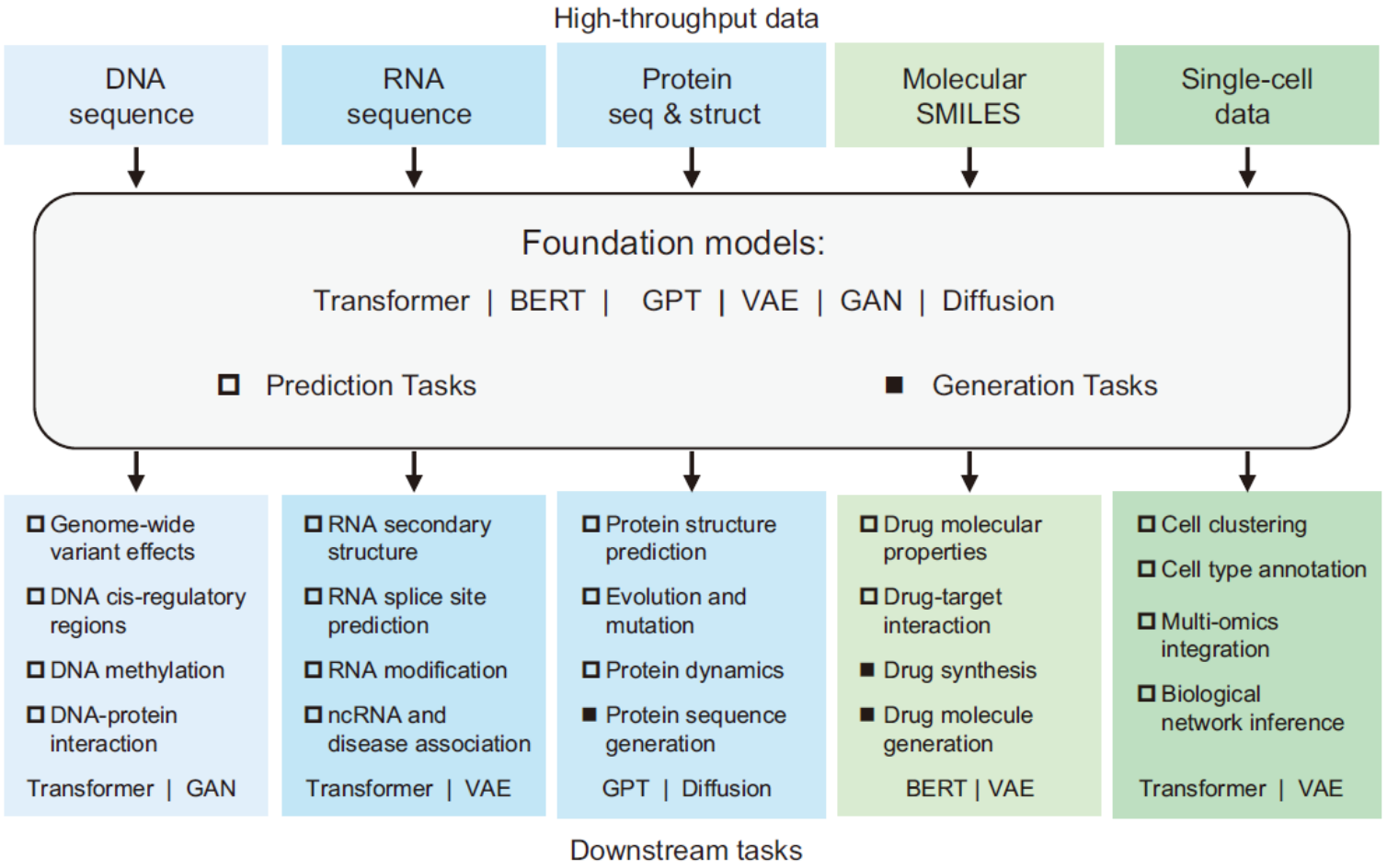
Humza Naveed[a], Asad Ullah Khan[a,*], Shi Qiu[b,*], Muhammad Saqib[c,d,*], Saeed Anwar[e,f], Muhammad Usman[e,f], Naveed Akhtar[g,i], Nick Barnes[h], Ajmal Mian[i]

arXiv:2307.06435, 2023

生物大模型/基础模型

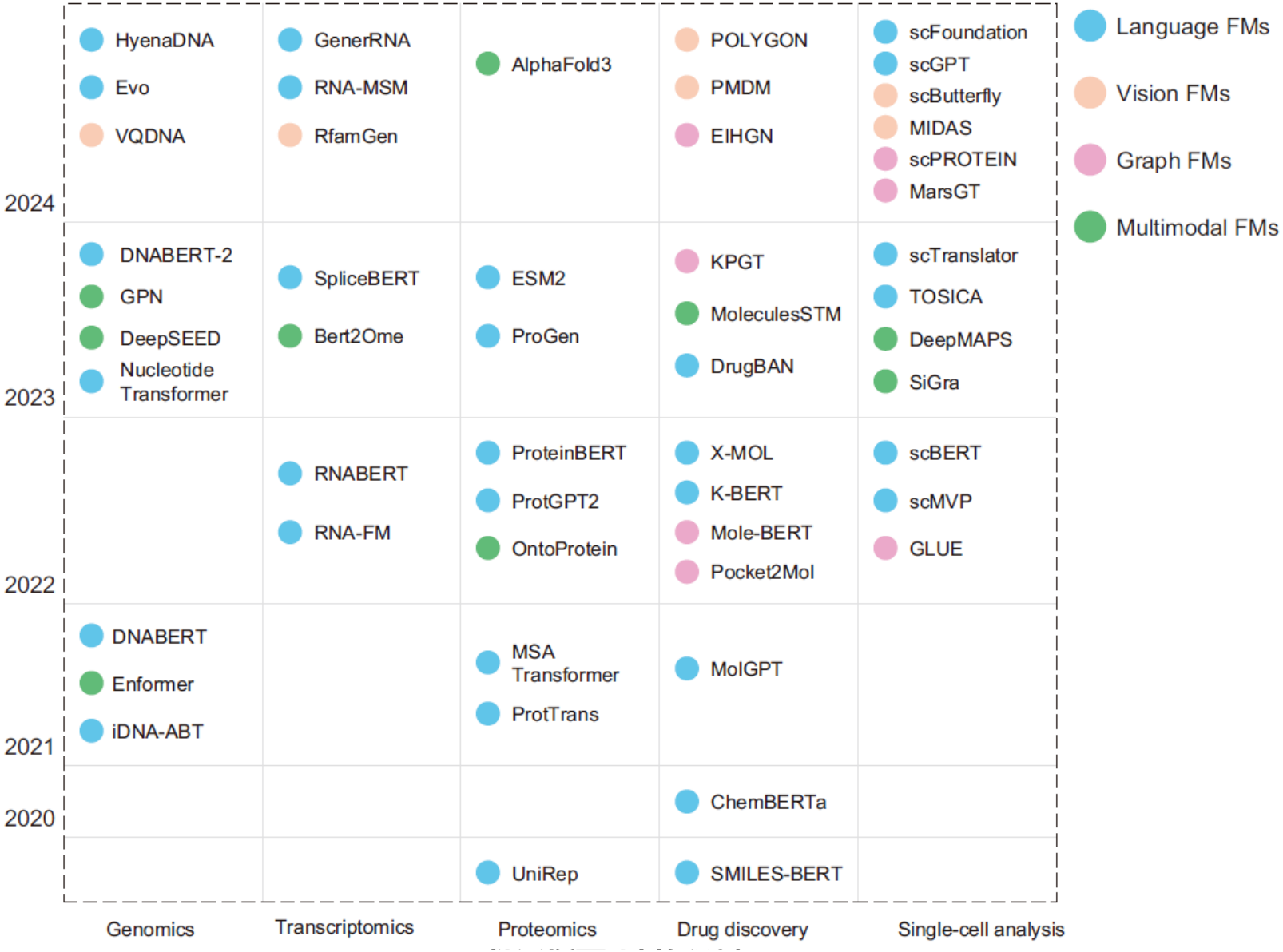
Foundation models in bioinformatics

Fei Guo[1,2], Renchu Guan[3], Yaohang Li[4], Qi Liu[5], Xiaowo Wang[6], Can Yang[7] and Jianxin Wang[1,2,*]



National Science Review, 2025, 12, nwaf028, https://doi.org/10.1093/nsr/nwaf028

生物大模型的演化


2024
HyenaDNA
Evo
VQDNA
GenerRNA
RNA-MSM
RfamGen
AlphaFold3
POLYGON
PMDM
EIHGN
scFoundation
scGPT
scButterfly
MIDAS
scPROTEIN
MarsGT
2023
DNABERT-2
GPN
DeepSEED
Nucleotide Transformer
SpliceBERT
Bert2Ome
ESM2
ProGen
KPGT
MoleculesSTM
DrugBAN
scTranslator
TOSICA
DeepMAPS
SiGra
2022
RNABERT
RNA-FM
ProteinBERT
ProtGPT2
OntoProtein
X-MOL
K-BERT
Mole-BERT
Pocket2Mol
scBERT
scMVP
GLUE
2021
DNABERT
Enformer
iDNA-ABT
MSA Transformer
ProtTrans
MolGPT
2020
ChemBERTa
UniRep
SMILES-BERT
Genomics
Transcriptomics
Proteomics
Drug discovery
Single-cell analysis
Language FMs
Vision FMs
Graph FMs
Multimodal FMs

AI生物学的研究范式

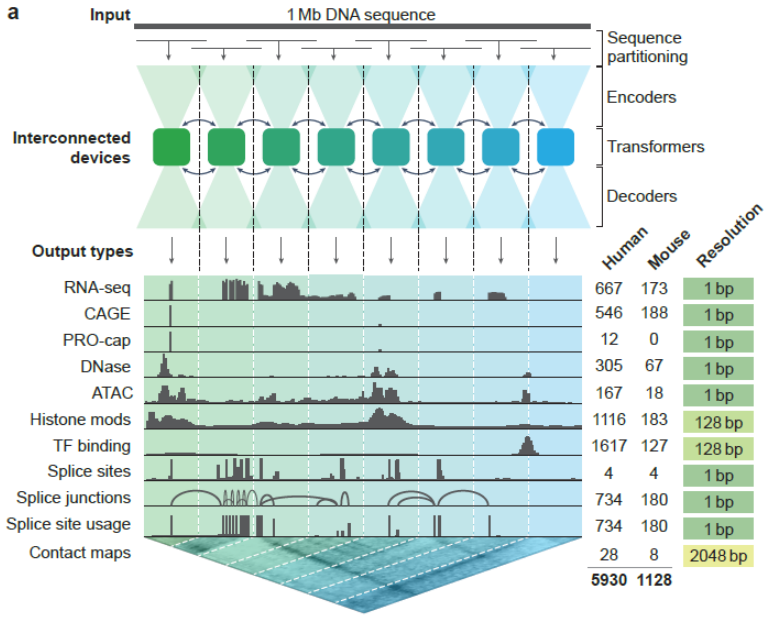
- **大模型、生成式、多模态、小样本**

- **预训练 + 微调**
  - X 自行构建预训练模型
  - √ **公共生物大模型 + 特定问题相关的数据资源**

大模型

AlphaGenome：变异-效应预测

AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model



上下文序列：1 mb

预测精度：1bp bins

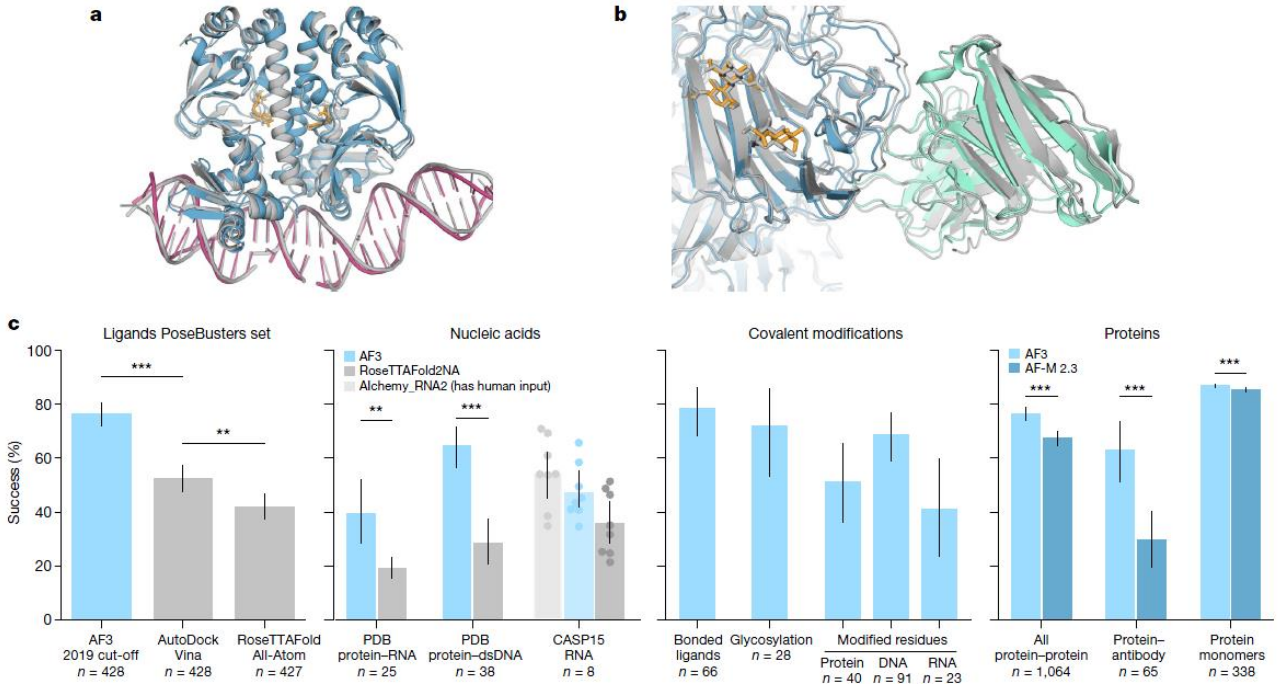
DeepMind

doi: https://doi.org/10.1101/2025.06.25.661532

AlphaFold 3：生物大分子相互作用预测

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

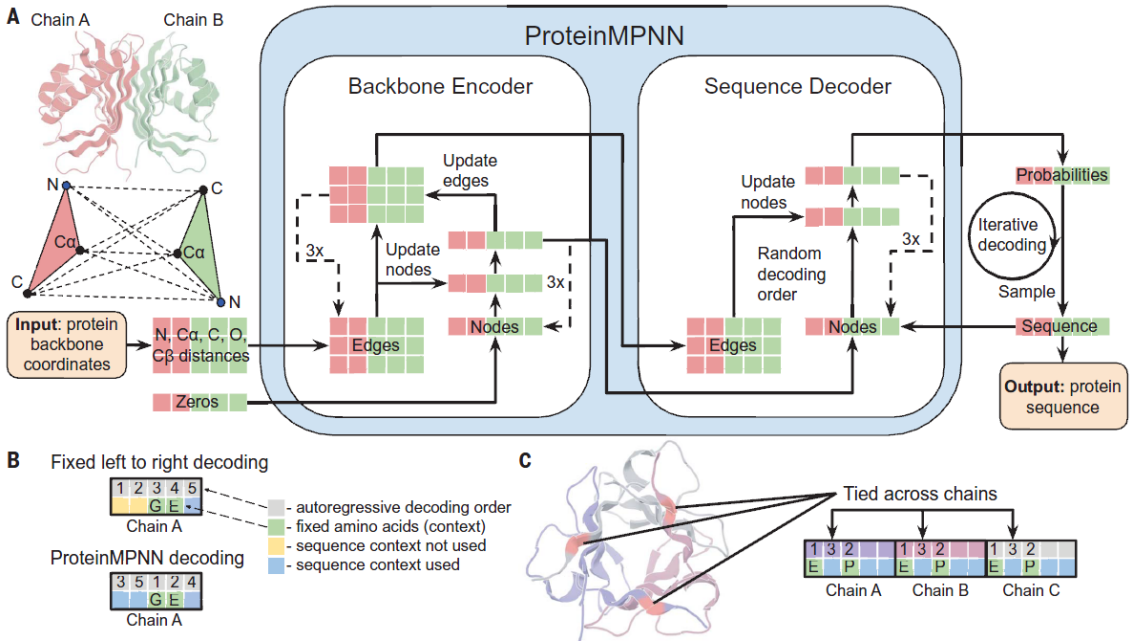


Nature, 2024, 630, 493-500

生成式

PROTEIN DESIGN

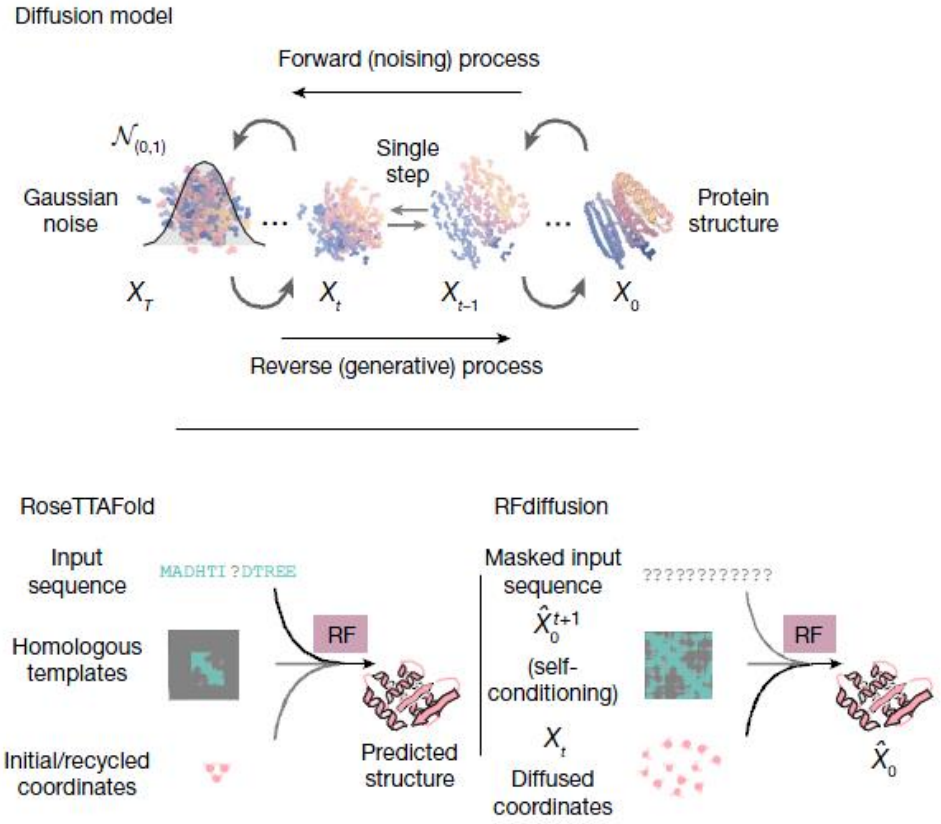
Robust deep learning–based protein sequence design using ProteinMPNN



Science, 2022, 378, 49-56

Article

De novo design of protein structure and function with RFdiffusion

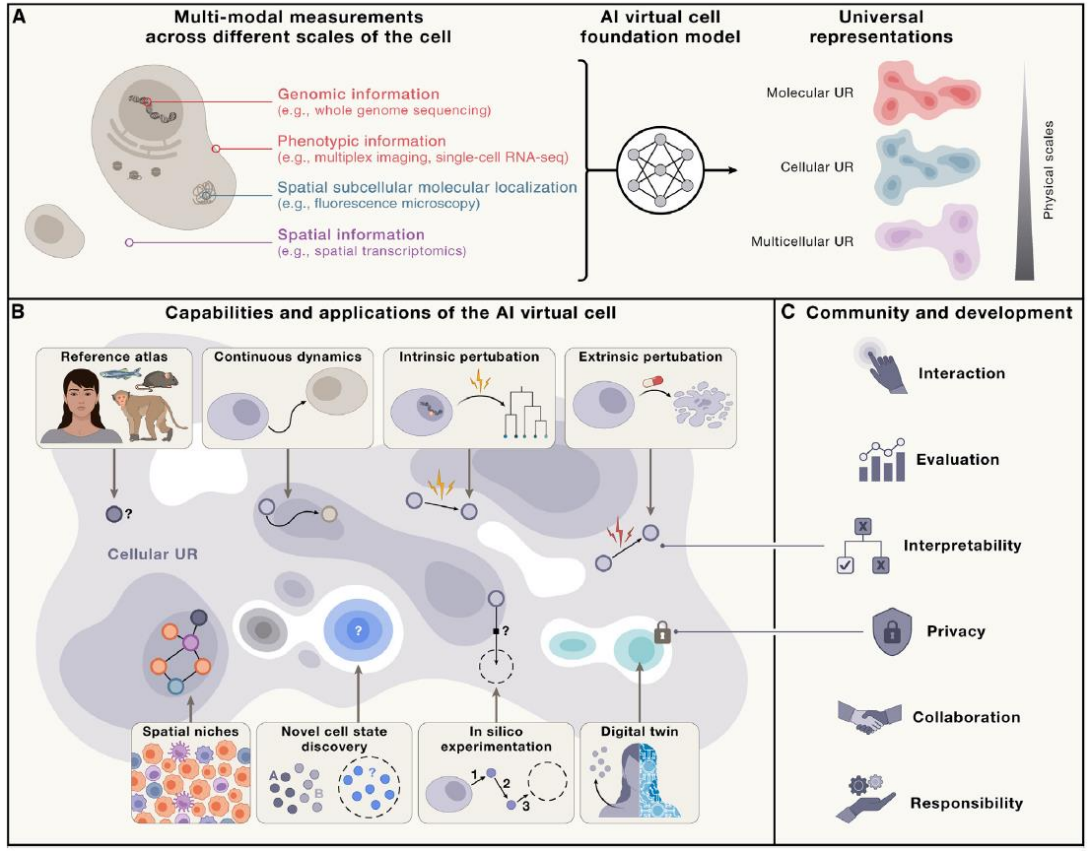


Nature, 2023, 620, 1089-1100

多模态

Perspective

How to build the virtual cell with artificial intelligence: Priorities and opportunities



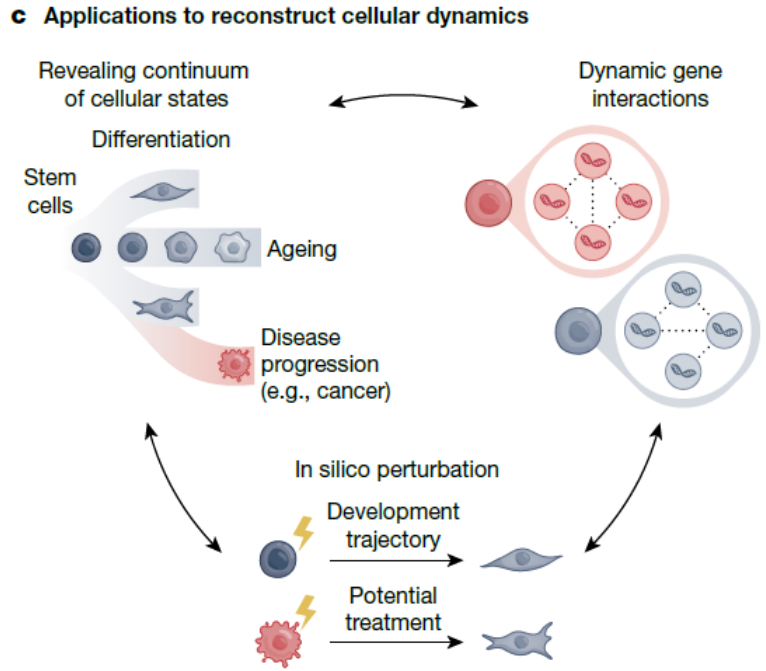
Cell, 2024, 187, 7045-7063

Perspective

Towards multimodal foundation models in molecular cell biology

https://doi.org/10.1038/s41586-025-08710-y
Received: 17 October 2023

Haotian Cui[1,2,3], Alejandro Tejada-Lapuerta[4,5], Maria Brbić[6,7,8], Julio Saez-Rodriguez[9,10], Simona Cristea[11,12], Hani Goodarzi[13,14], Mohammad Lotfollahi[15,16], Fabian J. Theis[4,5,17] & Bo Wang[1,2,3,18]



新细胞类型识别，生物标志物发现，
基因调控推断，扰动分析

Nature, 2025, 640, 623-633

小样本

PanPep

nature machine intelligence

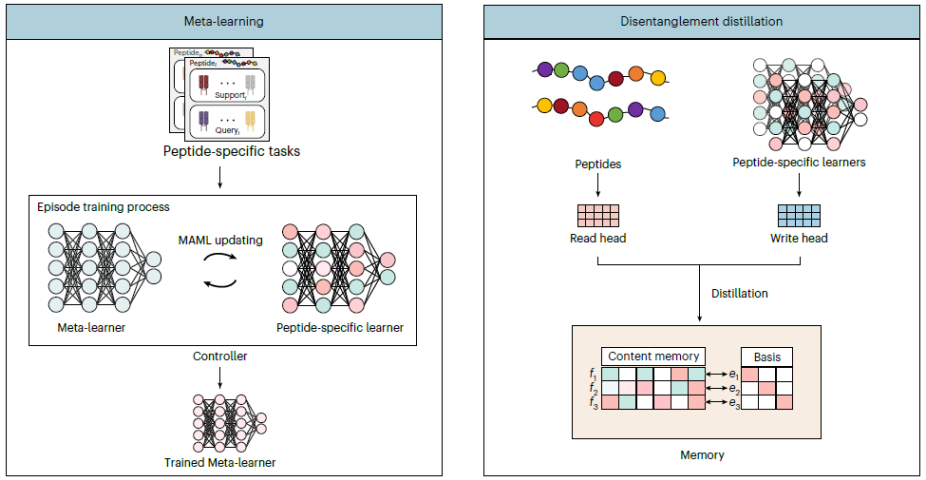
Article

https://doi.org/10.1038/s42256-023-00619-3

Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition

Received: 8 October 2022

Yicheng Gao[1,2,5], Yuli Gao[1,2,5], Yuxiao Fan[1,2], Chengyu Zhu[1,2], Zhiting Wei[1,2], Chi Zhou[1,2], Guohui Chuai[1,2], Qinchang Chen[3], He Zhang[2] & Qi Liu[1,2,3,4]



元学习预测T细胞受体-抗原结合

Nat Mach Intell, 2023, 5, 236-249

pFunK

nature metabolism

Article

https://doi.org/10.1038/s42255-024-01093-w

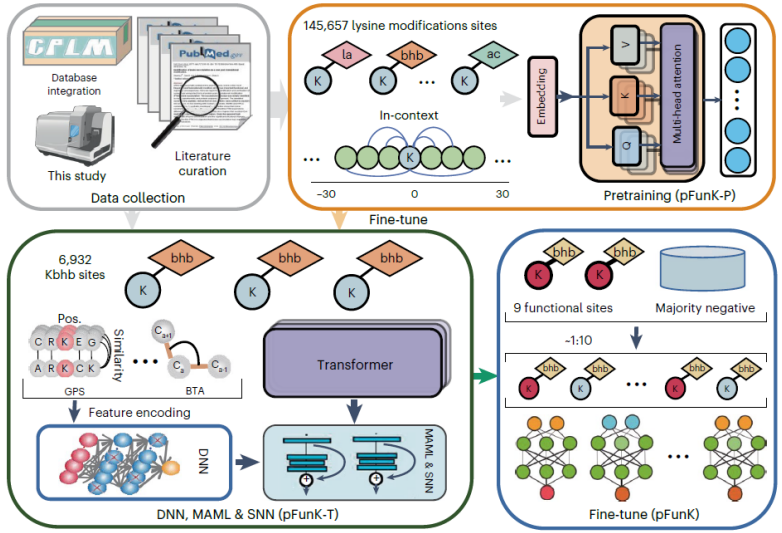
Ketogenic diet reshapes cancer metabolism through lysine β-hydroxybutyrylation

Received: 1 August 2023

Accepted: 2 July 2024

Published online: 12 August 2024

Junhong Qin[1,7], Xinhe Huang[2,7], Shengsong Gou[1,7], Sitao Zhang[1,7], Yujie Gou[2], Qian Zhang[1], Hongyu Chen[1], Lin Sun[3], Miaomiao Chen[2], Dan Liu[2], Cheng Han[2], Min Tang[1], Zihao Feng[2], Shenghui Niu[1], Lin Zhao[1], Yingfeng Tu[1], Zexian Liu[4], Weimin Xuan[3], Lunzhi Dai[5], Da Jia[1] & Yu Xue[2,6]

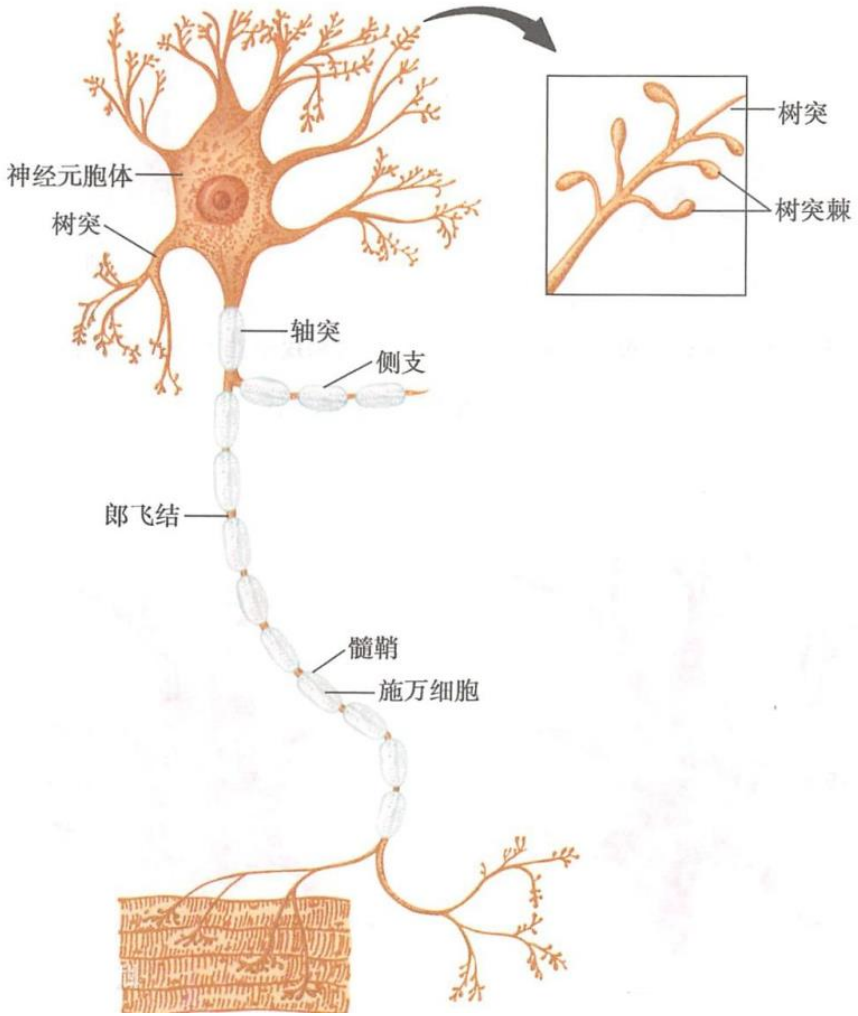


蛋白质修饰功能预测的小样本学习

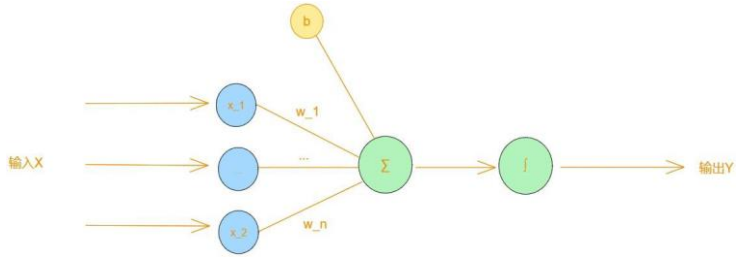
Nat Metab, 2024, 6, 1505-1528

神经网络与神经元

生物界中的神经元结构

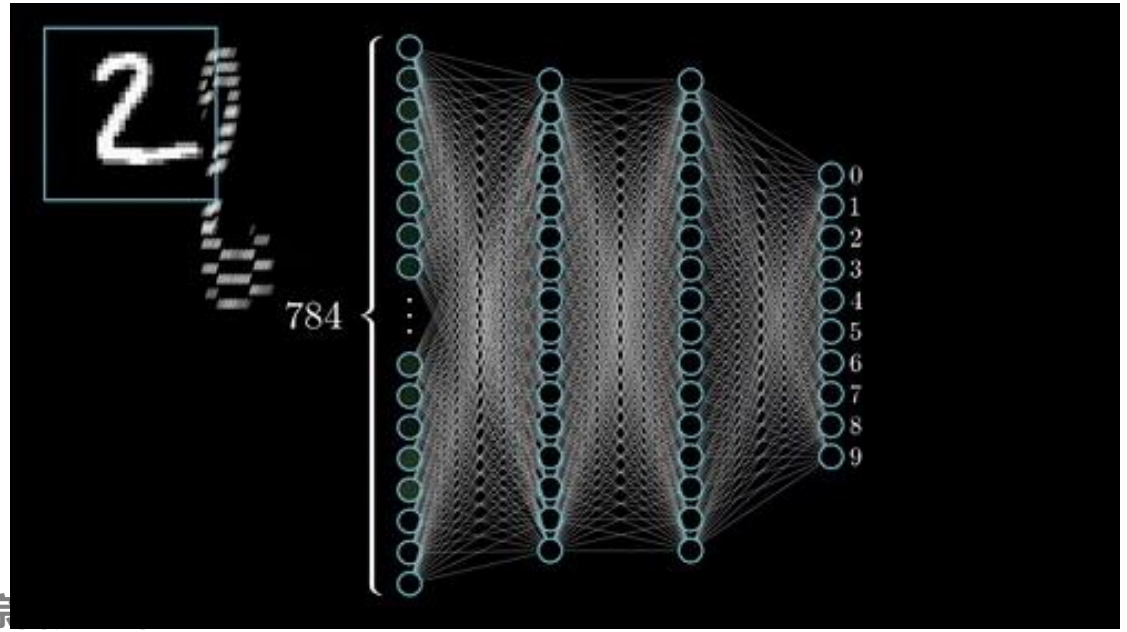


计算机中的神经元模型

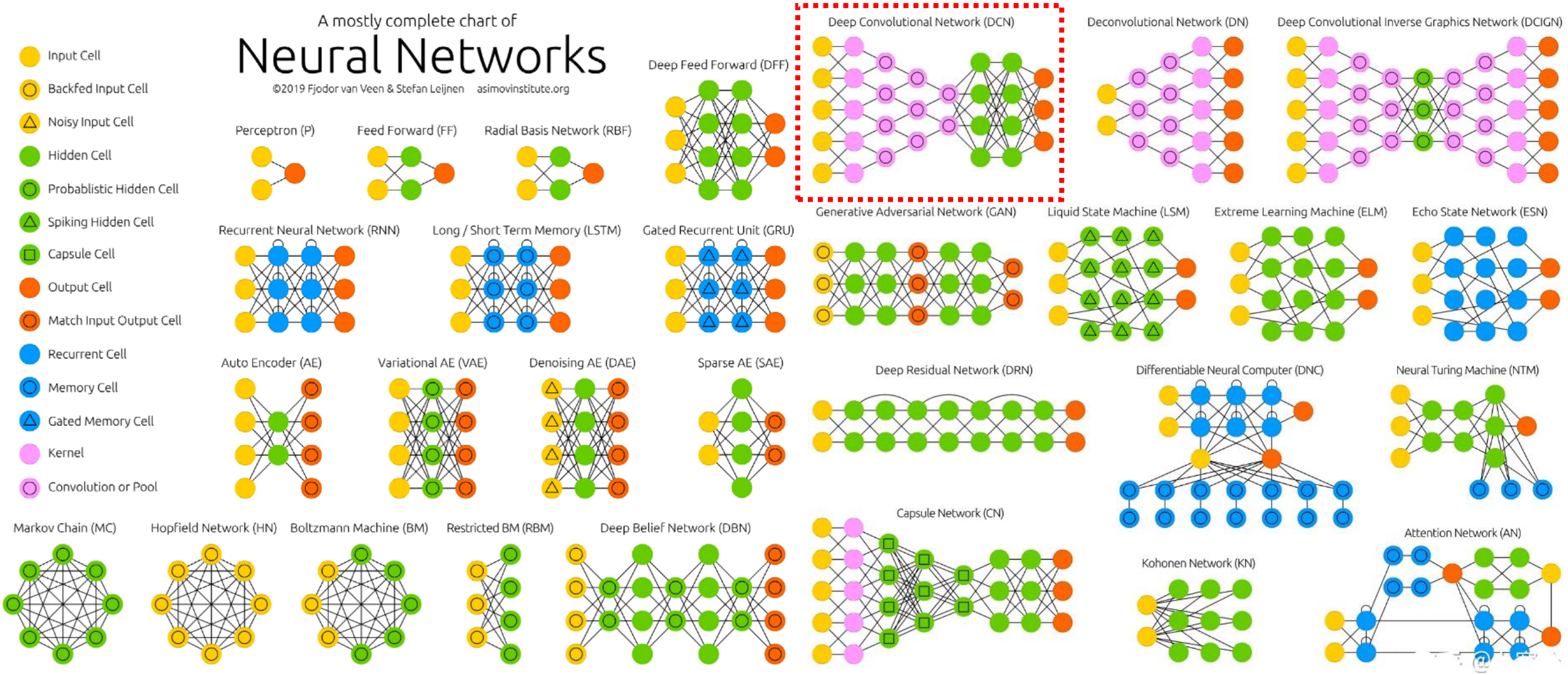


$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) = f\left(\vec{W} \cdot \vec{X} + b\right)$$

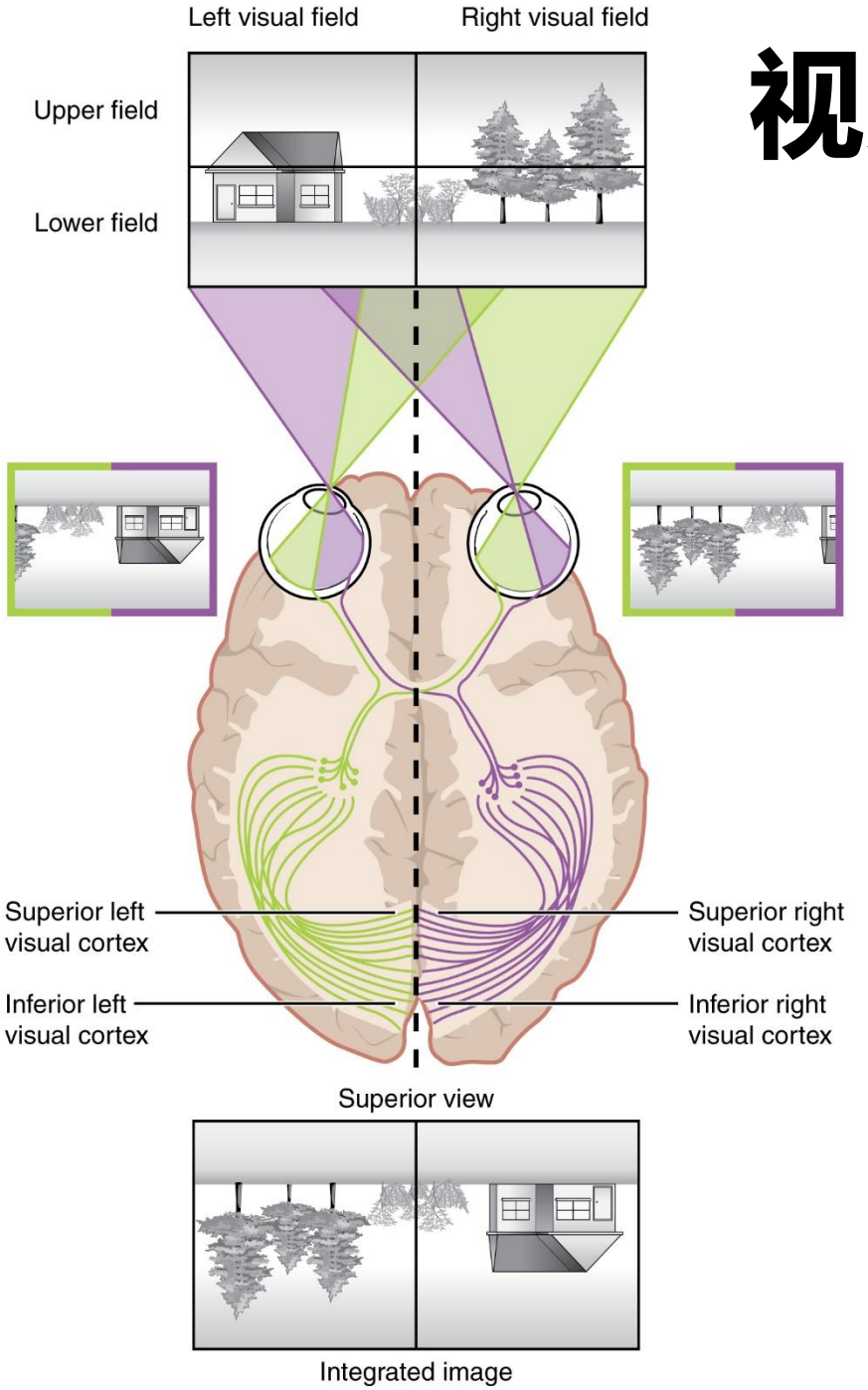
深度神经网络预测数字图像

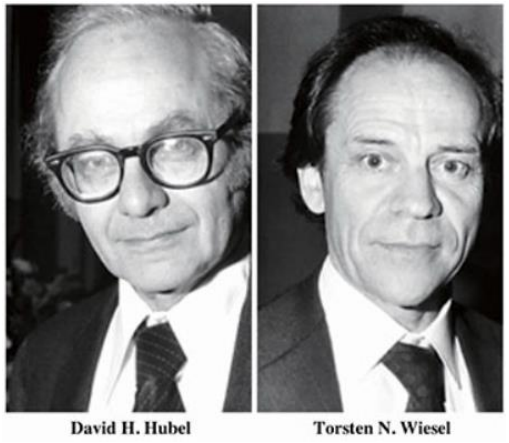


神经网络类型


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

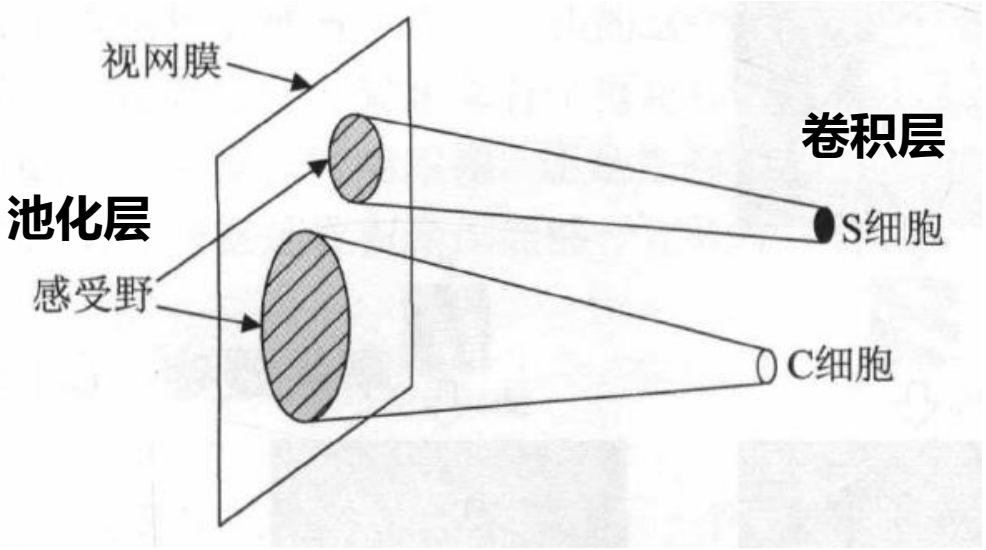
视觉系统的信息处理



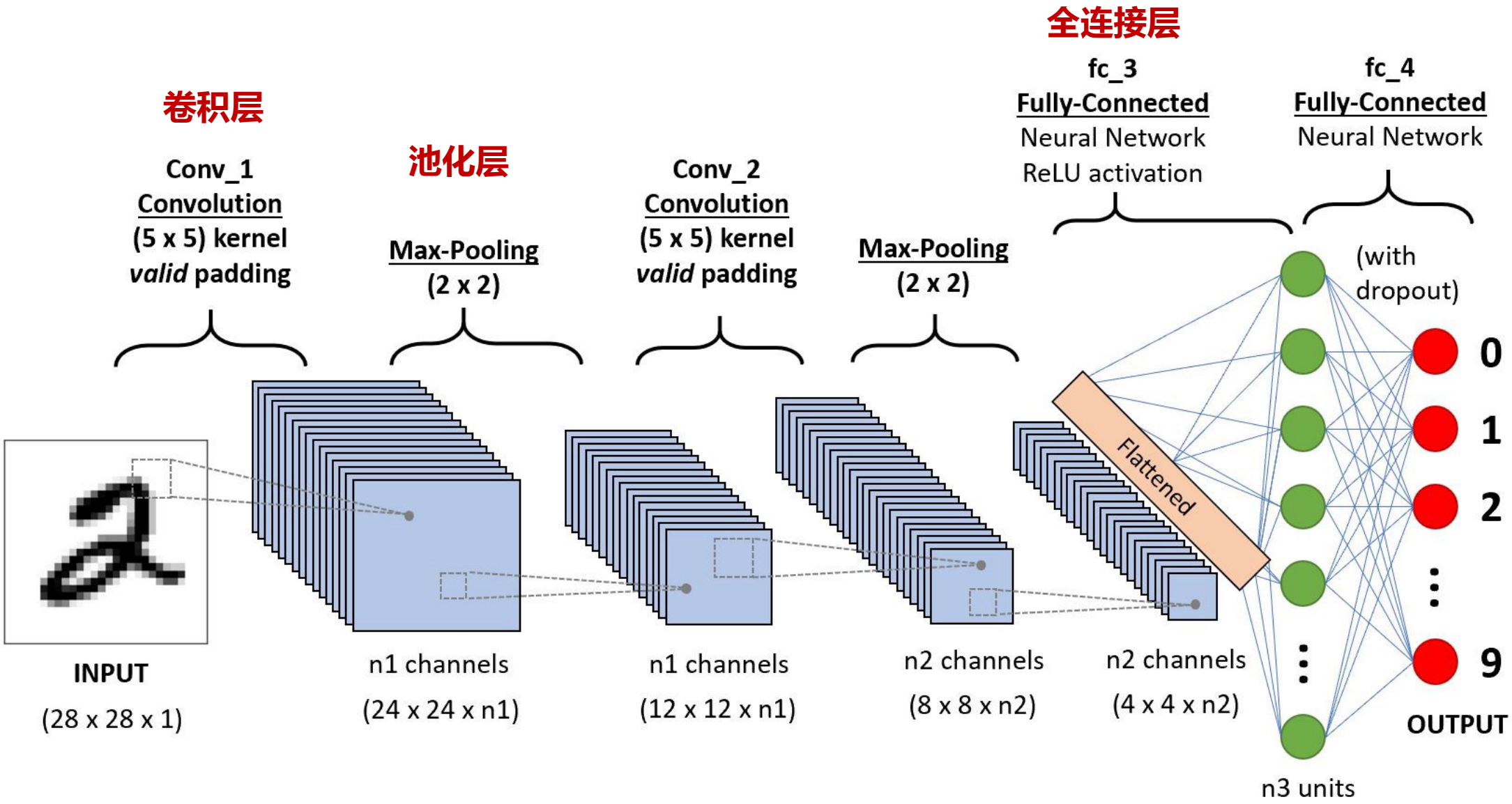


1981年诺贝尔生理学或医学奖

福岛邦彦，1980年提出
Neocognitron



卷积神经网络架构


全连接层
卷积层
池化层
fc_3
Fully-Connected
Neural Network
ReLU activation
fc_4
Fully-Connected
Neural Network
Conv_1
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
Conv_2
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
(with
dropout)
Flattened
0
1
2
9
INPUT
(28 x 28 x 1)
n1 channels
(24 x 24 x n1)
n1 channels
(12 x 12 x n1)
n2 channels
(8 x 8 x n2)
n2 channels
(4 x 4 x n2)
n3 units
OUTPUT

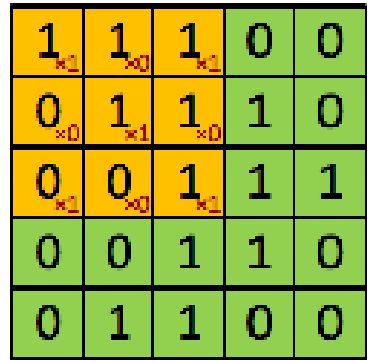
卷积运算与二维卷积

- **离散形式的卷积：**

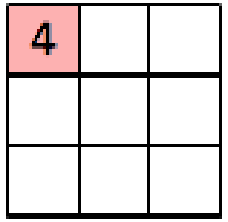
$$s(t) = \sum_{\alpha=-\infty}^{+\infty} x(\alpha)w(t-\alpha)$$

- **二维卷积：**

$$S(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n).$$



Image



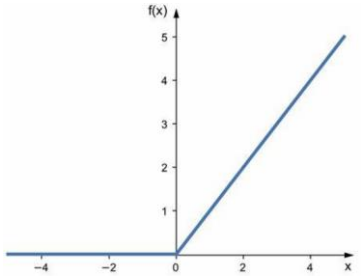
Convolved Feature

CSDN博客《卷积神经网络》

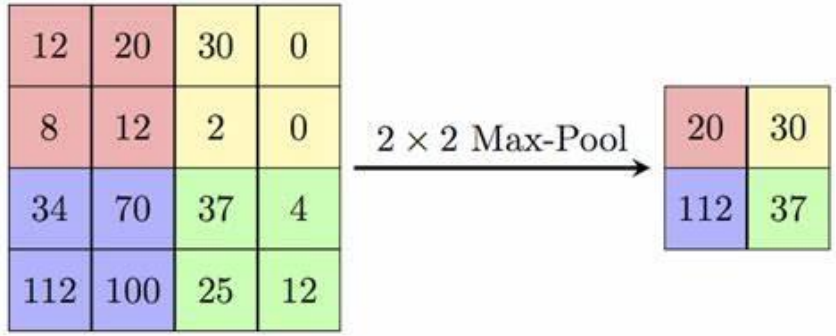
激活函数、卷积层、池化层和输出层

- 神经元激活函数

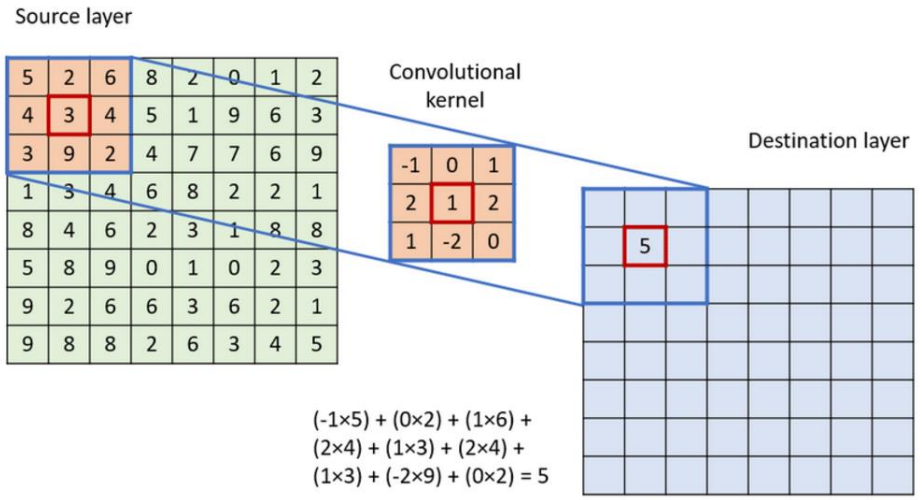
$$\mathrm{ReLU}(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases}$$



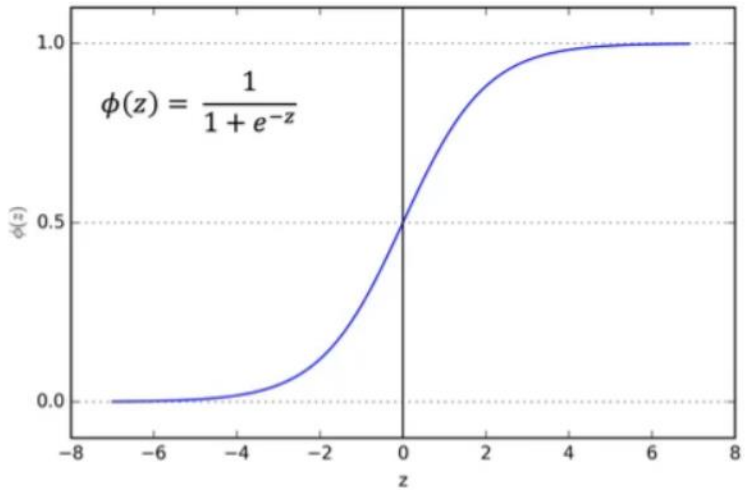
- 最大池化（Max pooling）



- 卷积层



- 输出：sigmoid函数



学习率与误差逆传播

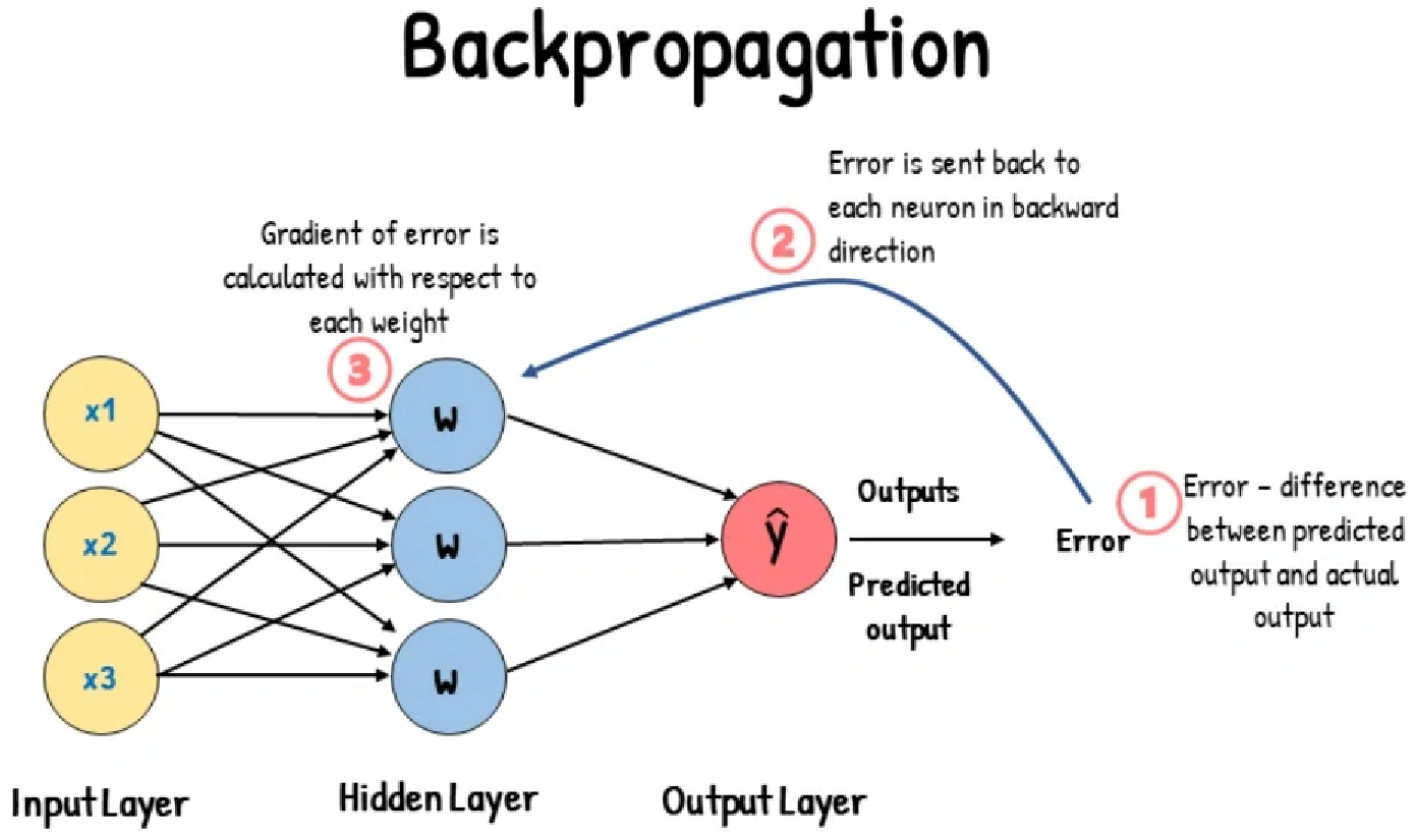
Old weight

Derivative of Error with respect to weight

$$^{*}W_x = W_x - a\left(\frac{\partial Error}{\partial W_x}\right)$$

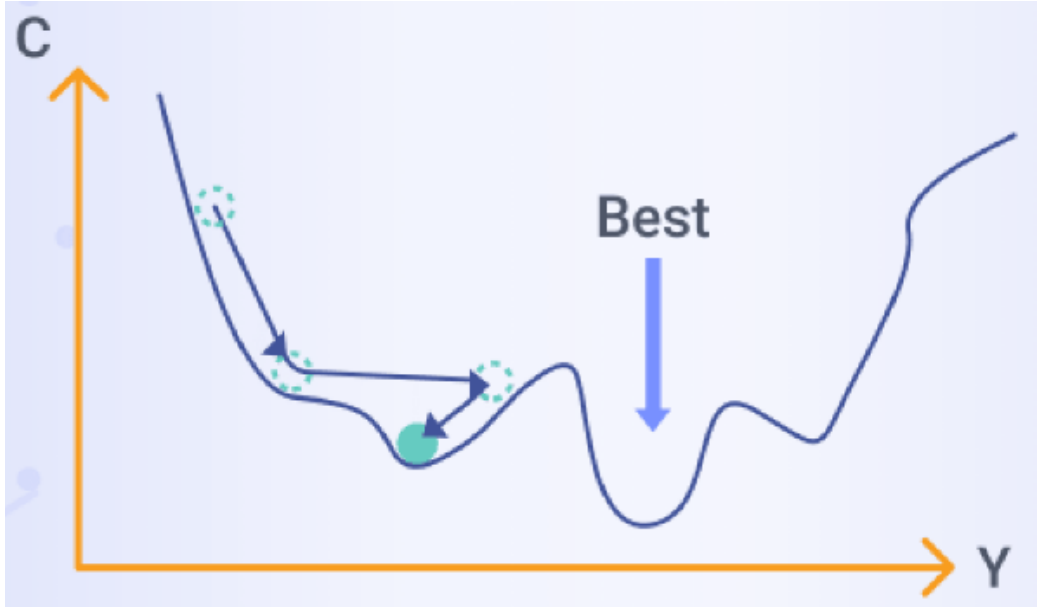
New weight

Learning rate

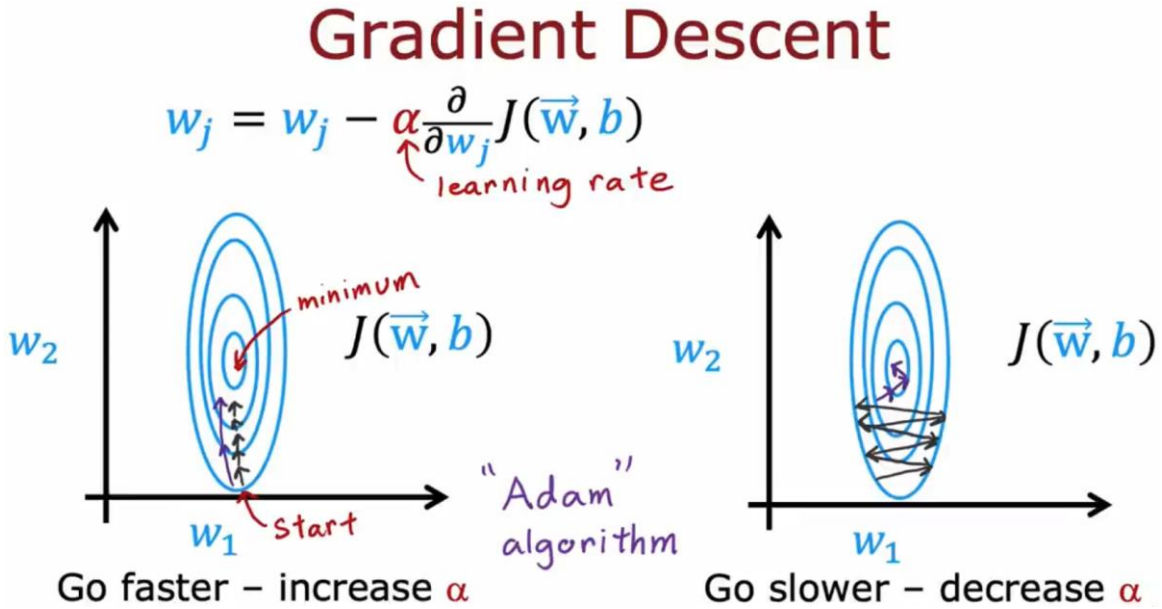


模型优化

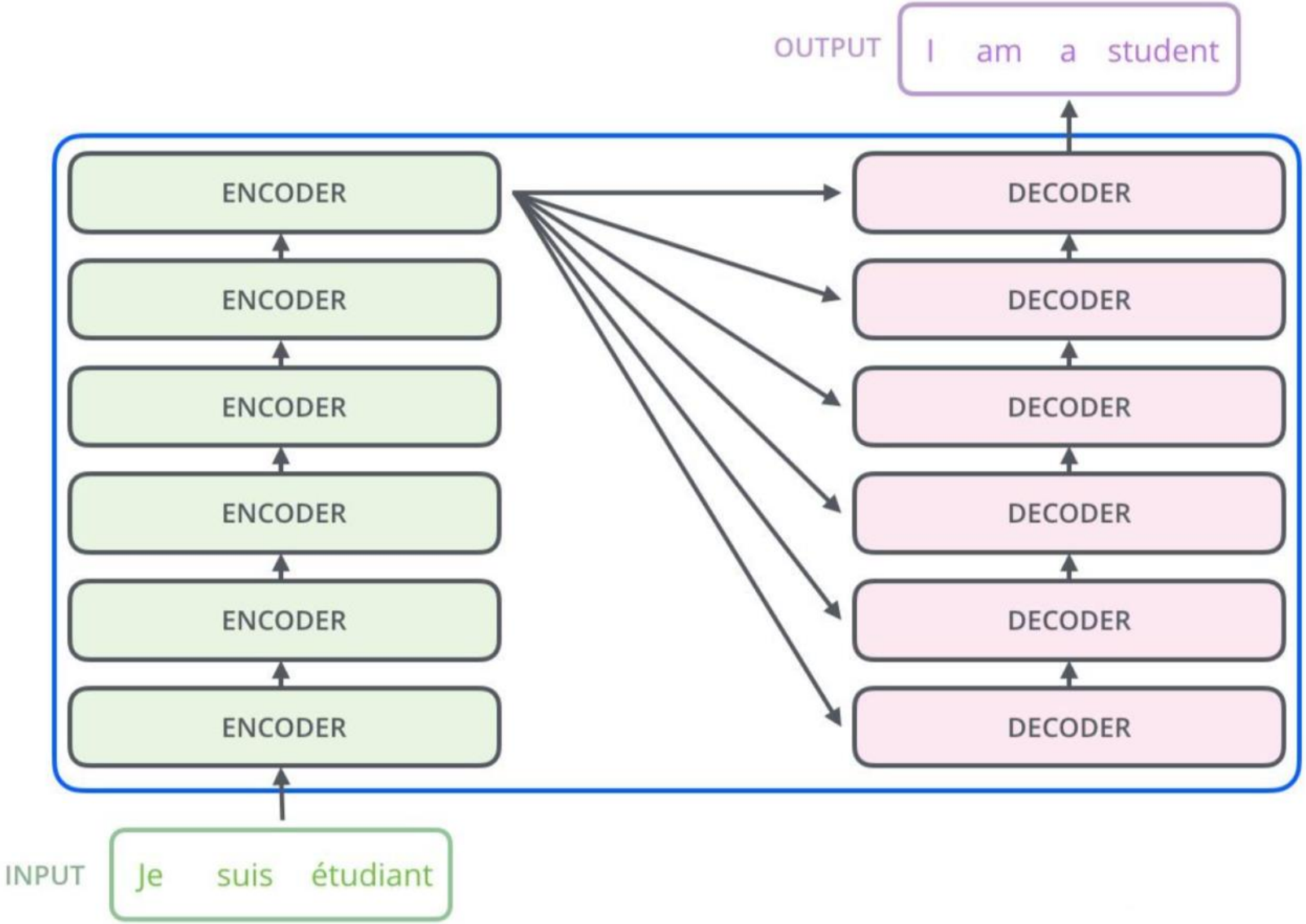
- **随机梯度下降SGD**
- **Stochastic Gradient Descent**



- **自适应矩估计Adam**
- **Adaptive Moment Estimation**



Transformer：编码器与解码器



词向量嵌入

- 1. 词表构建
  - { “I” : 0, “love” : 1, “NLP” : 2, “<PAD>” : 3} # PAD为填充符号
  - 句子“I love NLP”的索引序列为：[0, 1, 2]
- 2. 嵌入层
  - 假设嵌入维度 d_model = 4（实际中通常为512或768）

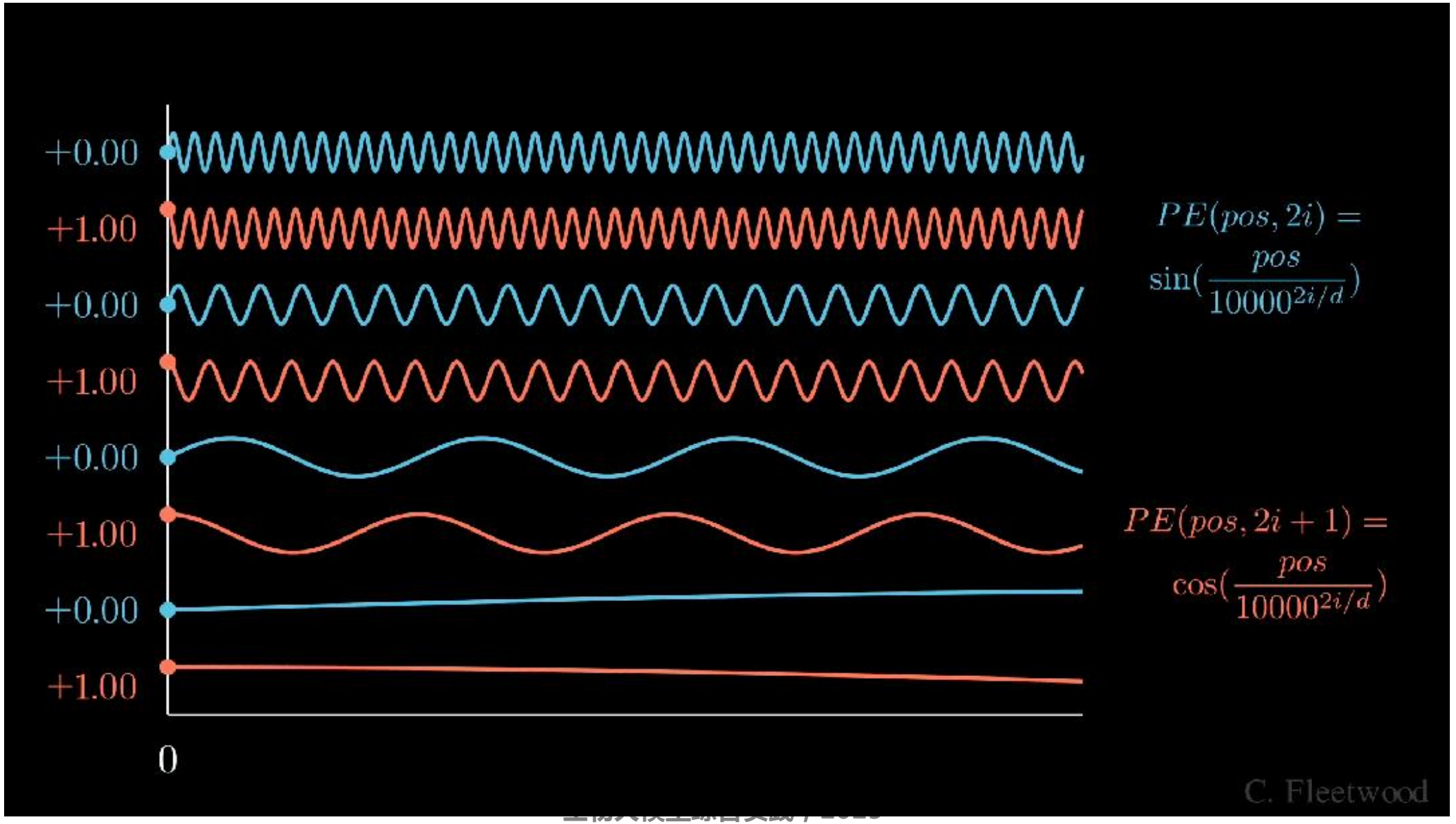
嵌入矩阵 `E` 的形状为 `[词表大小, d_model]`，例如：

```
E = [
    [0.1, 0.2, 0.3, 0.4],  # "I"的向量
    [0.5, 0.6, 0.7, 0.8],  # "love"的向量
    [0.9, 1.0, 1.1, 1.2],  # "NLP"的向量
    [0.0, 0.0, 0.0, 0.0]   # "<PAD>"的向量（全零）
]
```

通过索引查找得到词向量序列：

```
Embedded_words = [
    [0.1, 0.2, 0.3, 0.4],  # "I"
    [0.5, 0.6, 0.7, 0.8],  # "love"
    [0.9, 1.0, 1.1, 1.2]   # "NLP"
]
```

正余弦位置编码


+0.00
+1.00
+0.00
+1.00
+0.00
+1.00
+0.00
+1.00
0
$PE(pos, 2i) = \sin(\frac{pos}{10000^{2i/d}})$
$PE(pos, 2i+1) = \cos(\frac{pos}{10000^{2i/d}})$
C. Fleetwood

正余弦位置编码

$$PE_{pos,2i} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{pos,2i+1} = cos(pos/10000^{2i/d_{model}})$$

- ***pos***：单词在句子中的位置（从0开始）
- d_{model}：嵌入维度（此处为4）
- ***i***：维度索引（0到3）

- 对于位置0（第一个词“I”），位置编码向量是：

PE(0, 0) = sin(0 / 10000^(0/4)) = 0
PE(0, 1) = cos(0 / 10000^(0/4)) = 1
PE(0, 2) = sin(0 / 10000^(2/4)) = 0
PE(0, 3) = cos(0 / 10000^(2/4)) = 1

- 对于位置1（第二个词“love”），位置编码向量是：

PE(1, 0) = sin(1 / 10000^(0/4)) = 0.0001
PE(1, 1) = cos(1 / 10000^(0/4)) = 1.0
PE(1, 2) = sin(1 / 10000^(2/4)) = 0.0087
PE(1, 3) = cos(1 / 10000^(2/4)) = 0.9999

- 位置2（“NLP”）：

类似计算，结果：[0.0002, 1.0, 0.0175, 0.9998]

完整位置编码矩阵：

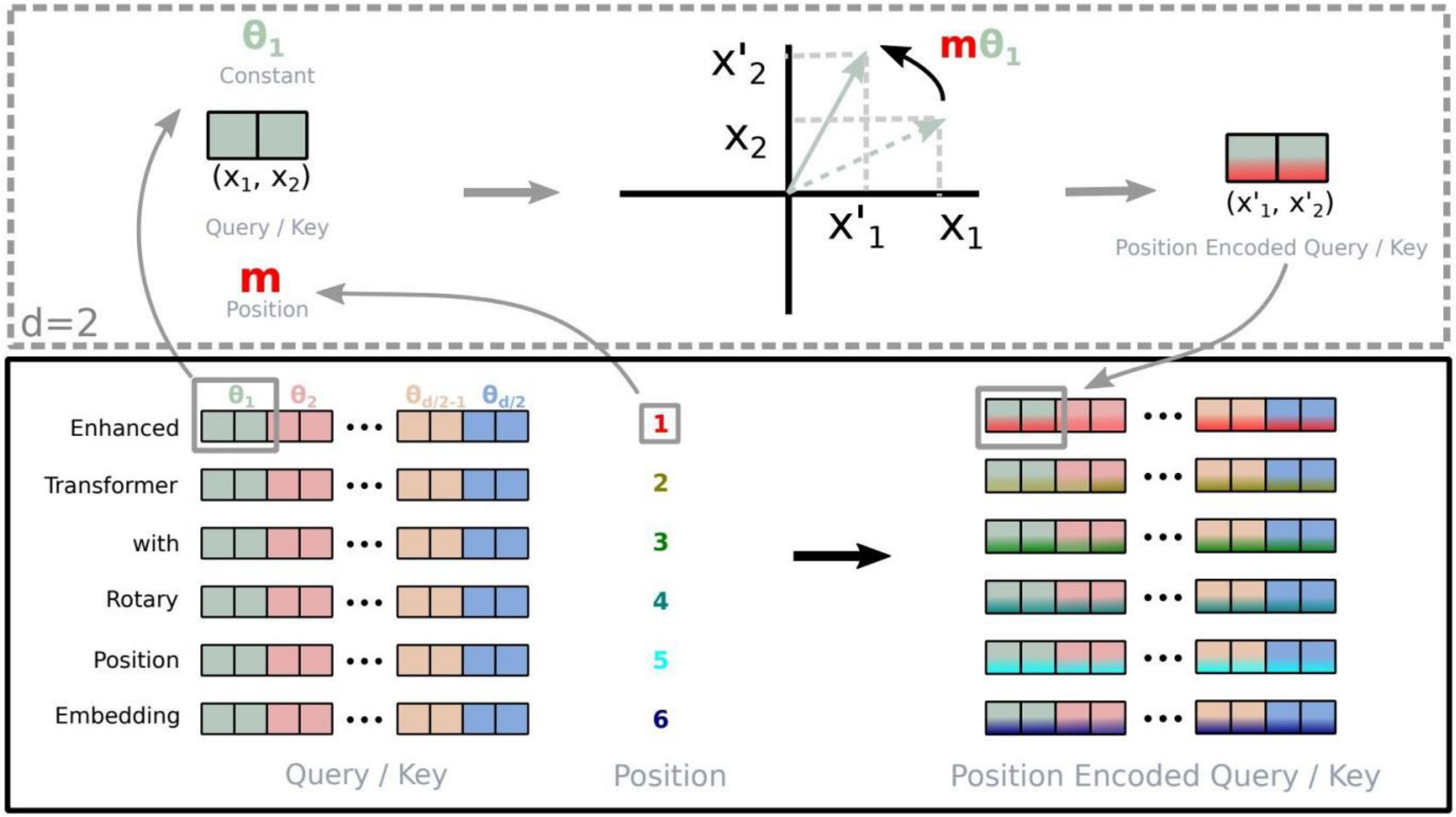
```
PE = [
    [0, 1, 0, 1],                # 位置0
    [0.0001, 1.0, 0.0087, 0.9999],  # 位置1
    [0.0002, 1.0, 0.0175, 0.9998]   # 位置2
]
```

词向量与位置编码相加

- 将词向量和位置编码逐元素相加，得到最终输入向量：
  - Final_input = Embedded_words + PE
- 例如，“I”的最终向量为：
  - [0.1+0, 0.2+1, 0.3+0, 0.4+1] = [0.1, 1.2, 0.3, 1.4]
- 最终结果：

```
Final_input = [
    [0.1, 1.2, 0.3, 1.4],  # "I" + 位置0
    [0.5+0.0001, 0.6+1.0, 0.7+0.0087, 0.8+0.9999],  # "love" + 位置1
    [0.9+0.0002, 1.0+1.0, 1.1+0.0175, 1.2+0.9998]   # "NLP" + 位置2
]
```

旋转位置编码RoPE

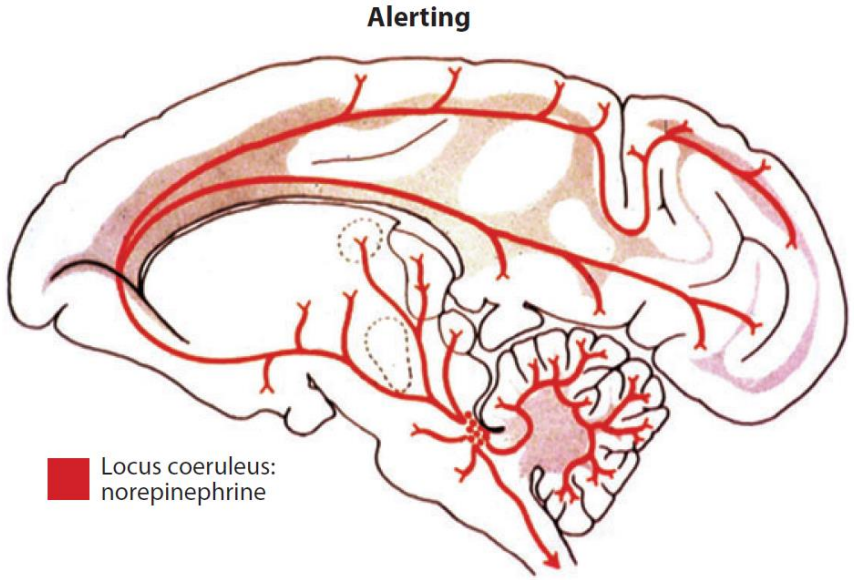

θ1
Constant
(x1, x2)
Query / Key
m
Position
d=2
x'2
x2
x'1
x1
mθ1
(x'1, x'2)
Position Encoded Query / Key
θ1
θ2
θd/2-1
θd/2
Enhanced
Transformer
with
Rotary
Position
Embedding
1
2
3
4
5
6
Query / Key
Position
Position Encoded Query / Key

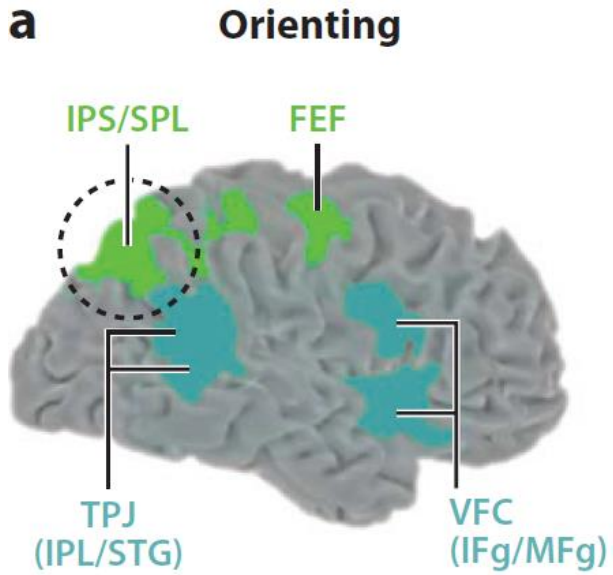
人脑的注意力系统

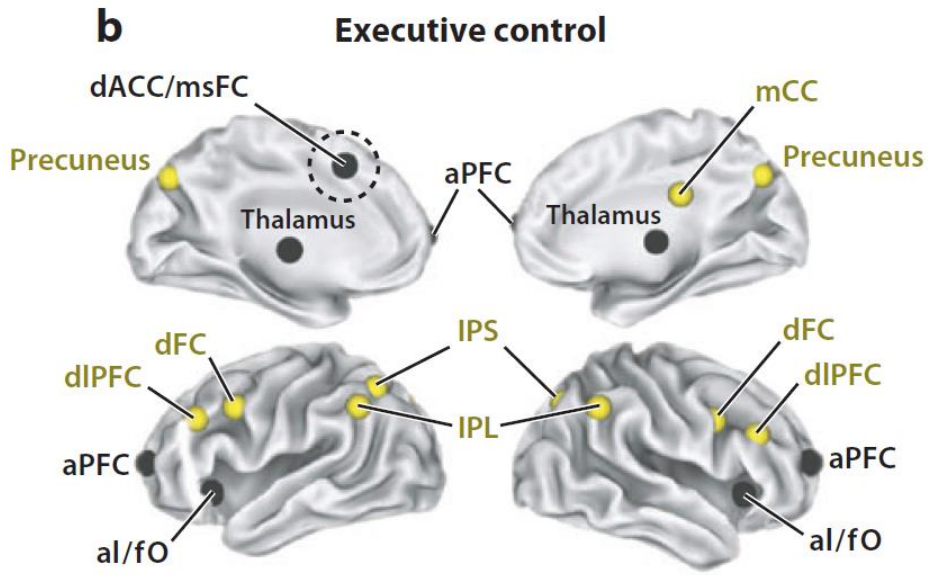
The Attention System of the Human Brain: 20 Years After

Steven E. Petersen[1] and Michael I. Posner[2]

- 警觉系统（Alerting）、定向系统（Orienting）和执行控制系统（Executive Control）

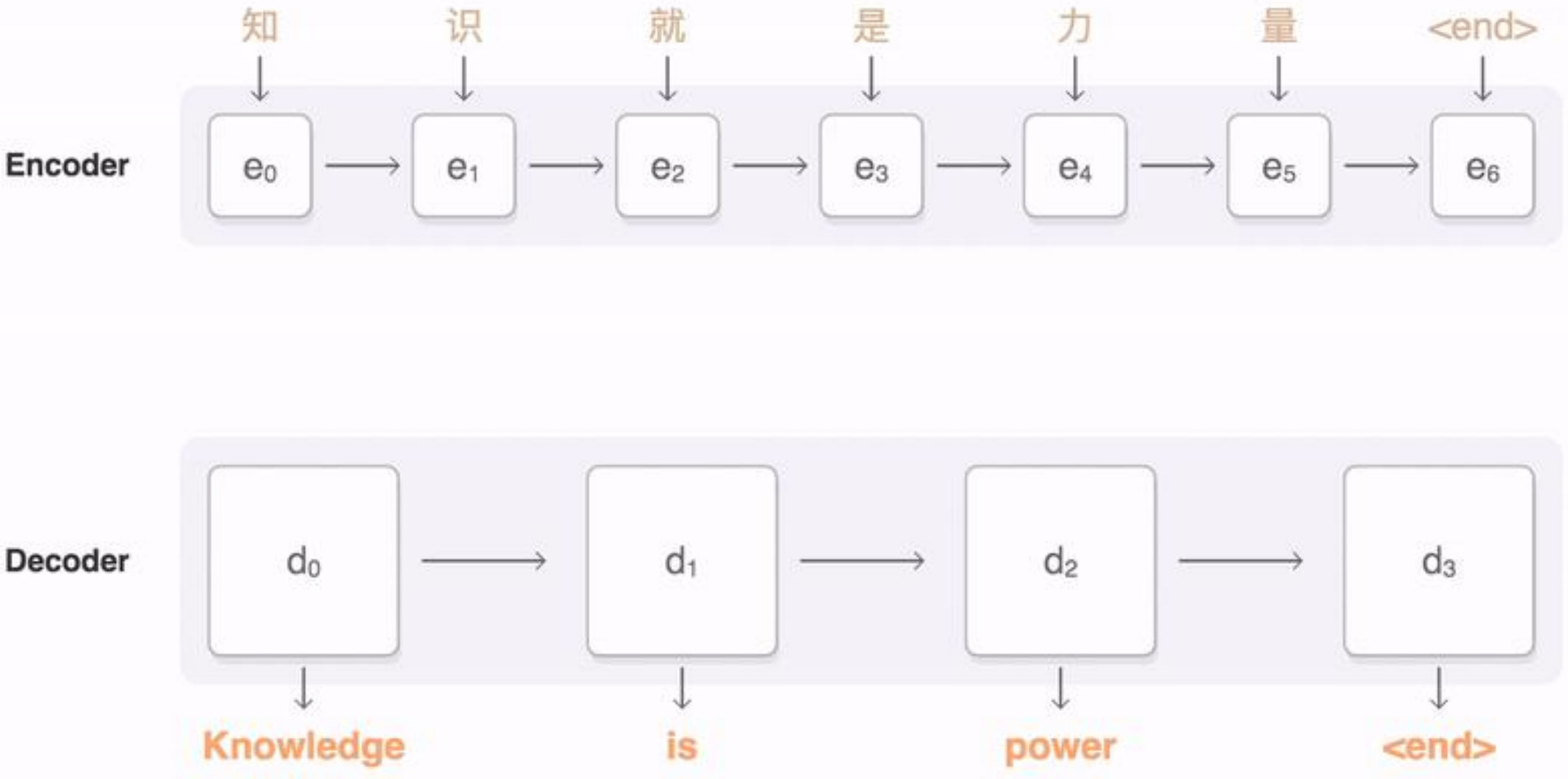




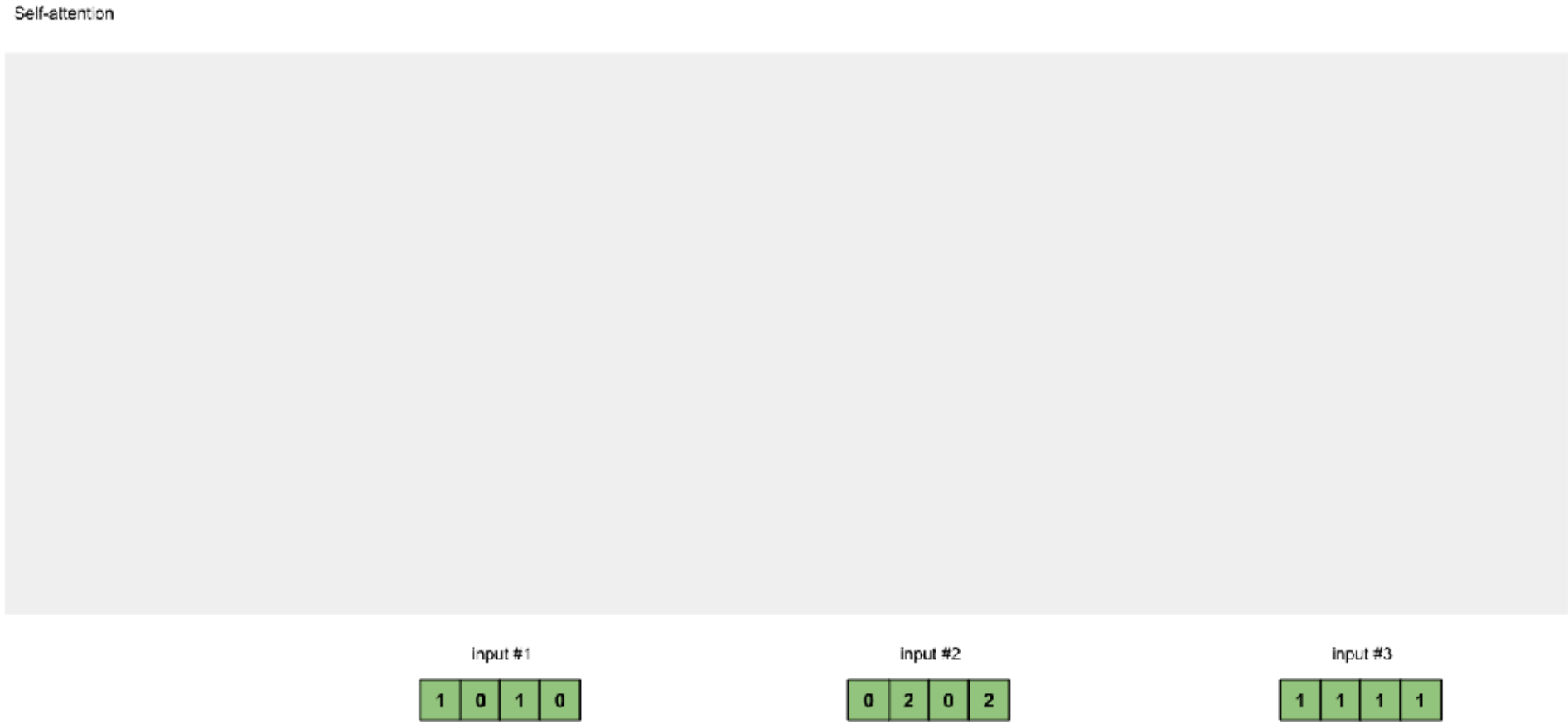


Annu Rev Neurosci, 2012, 35, 73-89

Transformer的注意力机制



注意力机制的计算过程



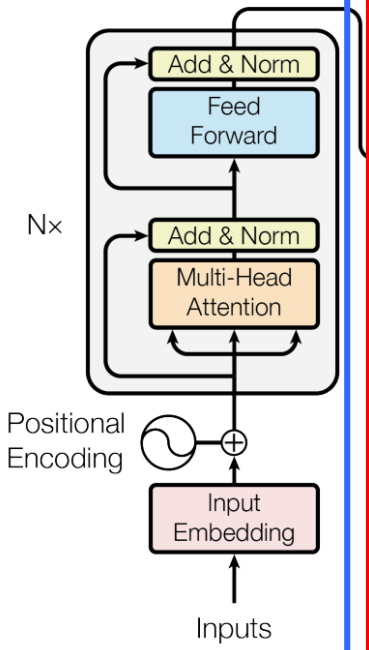
大语言模型的两种主流架构

BERT

编码器

遮蔽词元预测

在训练过程中随机遮蔽一些词元，然后让模型去预测这些被遮蔽的词元

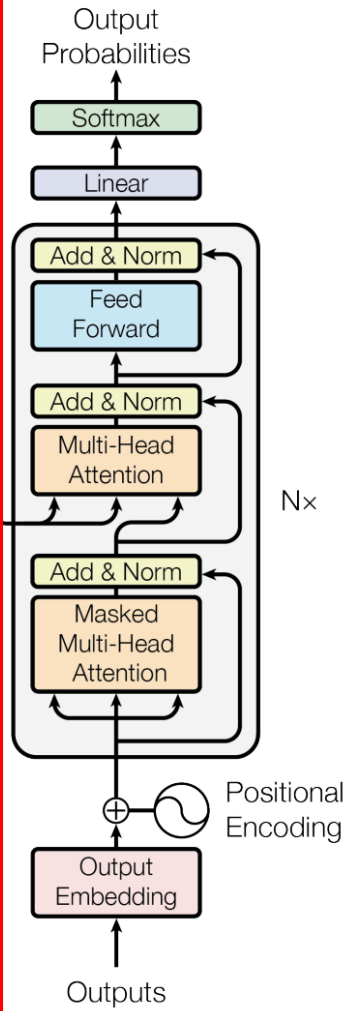


GPT

解码器

下一个词元预测

根据前面的词元来预测下一个词元



Nat Biotechnol, 2024, 42, 200-202

随堂小测 & 课外学习

- 作业：
  - 1. 思考题： 在生物大模型微调中，为何需要使用领域特异性数据集？试举例说明？
  - 2. 实践题：让DeepSeek以句子如“I love HUST”为例，说明transformer是怎么用自注意力机制来进行计算的
- 参考资料：
  - 1. 学术论文：Empowering biomedical discovery with AI agents. Cell, 2024, 187, 6125-6151.
  - 2. 学术论文：The generative era of medical AI. Cell, 2025, 18, 3648-3660.