新媒体大数据分析

New Media Big Data Analysis

专题：自然语言处理与大模型

黄振亚，朱孟潇，张凯

Email: huangzhy@ustc.edu.cn，mxzhu@ustc.edu.cn

课程主页：

http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html

助教：齐畅，朱家骏

bigdata_2025@163.com

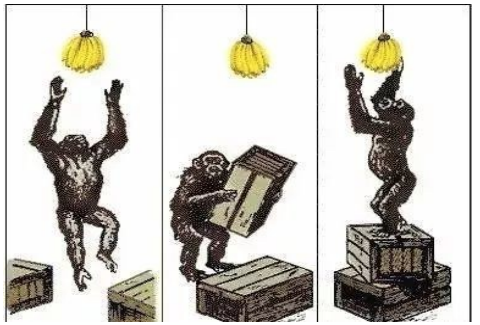
Outline

- 人工智能
- 自然语言处理发展
- 大模型的原理与技术
- 大模型应用
- 未来展望

什么是智能

- 系统通过获取和加工信息而取得的一种能力，从而实现简单到复杂的演化
  - 系统，可以是有机的生物体系统，也可以是计算机系统

- 以生物体为载体称为生物智能，以机器为载体的称为机器智能。
- 两者可以遵循相同或者相似的规律。

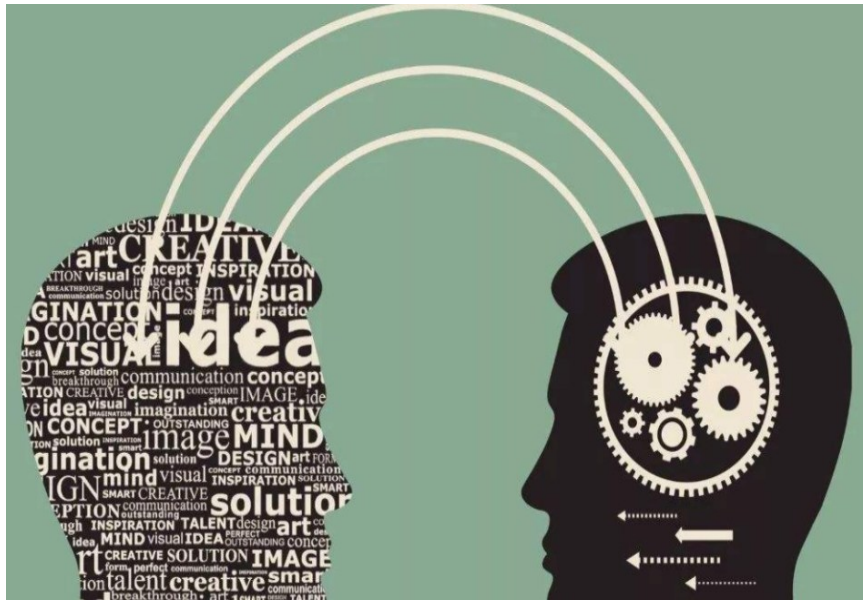


学习能力　　理解能力　　逻辑思维能力　　解决问题能力

什么是人工智能

人工智能
(Artificial Intelligence)

能够和人一样进行感知、认知、决策、执行的人工程序或系统



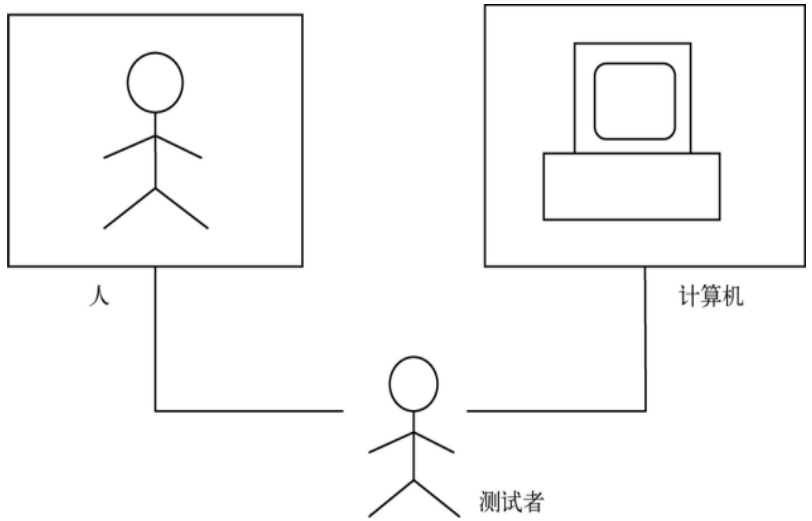
- **人工智能(Artificial Intelligence, AI)**，是研究、开发用于模拟、延伸和扩展**人类智能**的理论、方法、技术及应用的一门新的技术科学。
- 人工智能是对人的意识、思维的信息过程的模拟。人工智能不是人的智能，但能像人那样思考、也可能超过人的智能。
- **ANI (Artificial Narrow Intelligence)**
  --> **AGI (Artificial General Intelligence)**
  --> **ASI (Artificial Super Intelligence)**

图灵测试　　Can machines think?

1950年，**“计算机之父”和“人工智能之父”艾伦·图灵（Alan M. Turing）**发表了论文《计算机器与智能》，这篇论文被誉为人工智能科学的开山之作。在论文的开篇，图灵提出了一个引人深思的问题：“机器能思考吗？”。这个问题激发了人们无尽的想象，同时也奠定了人工智能的基本概念和雏形。

艾伦·图灵 (Alan Turing, 1950)

这位英国数学家、计算机科学之父，在他划时代的论文《计算机器与智能》中，提出了一个振聋发聩的问题。为了回答这个问题，他设计了著名的“模仿游戏”，即后人所称的“图灵测试”，为衡量机器智能提供了一个可操作的哲学和科学标准。



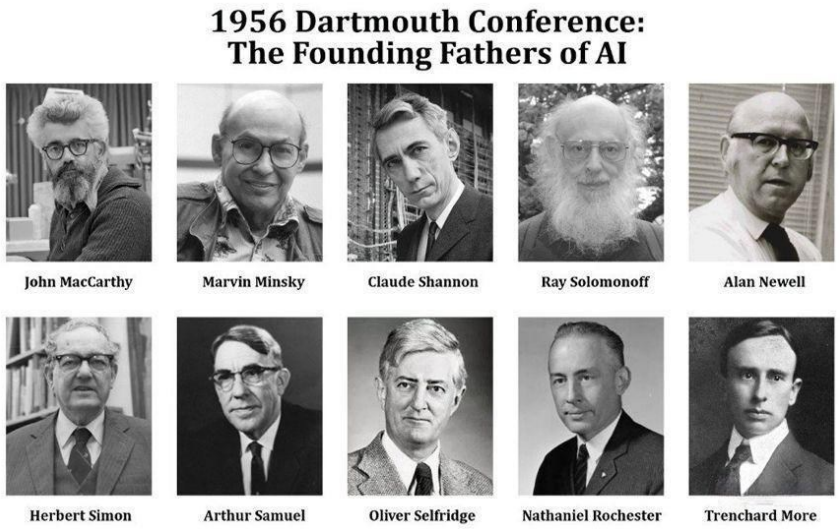
测试内容

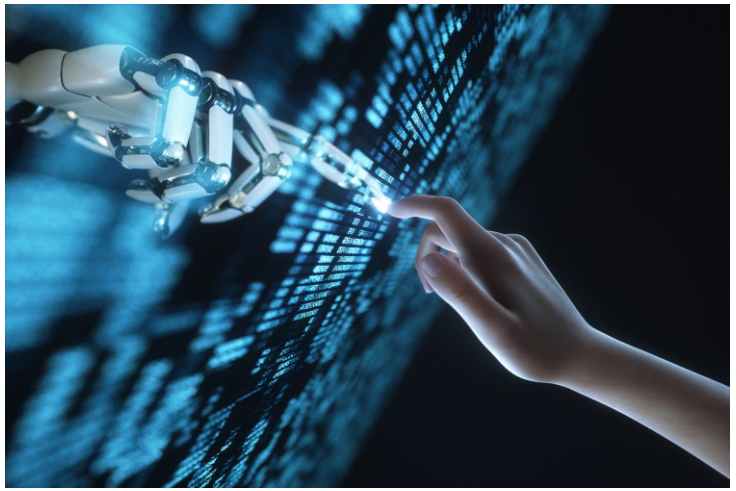
若机器在对话中能让人无法分辨其机器身份

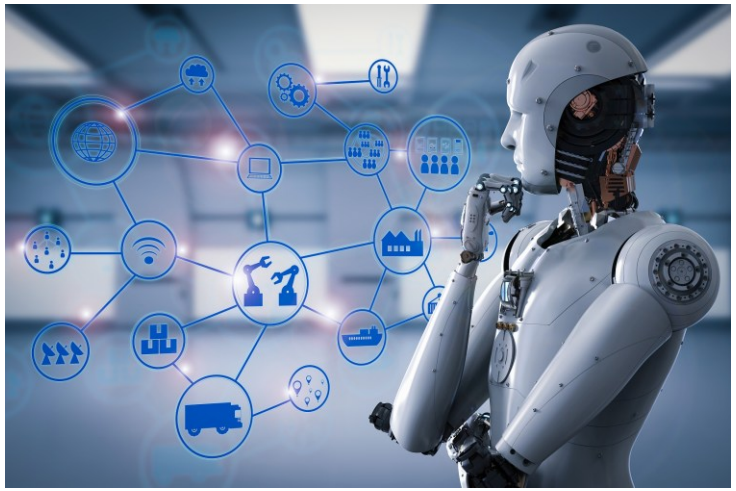
则视为具备智能

人工智能的诞生

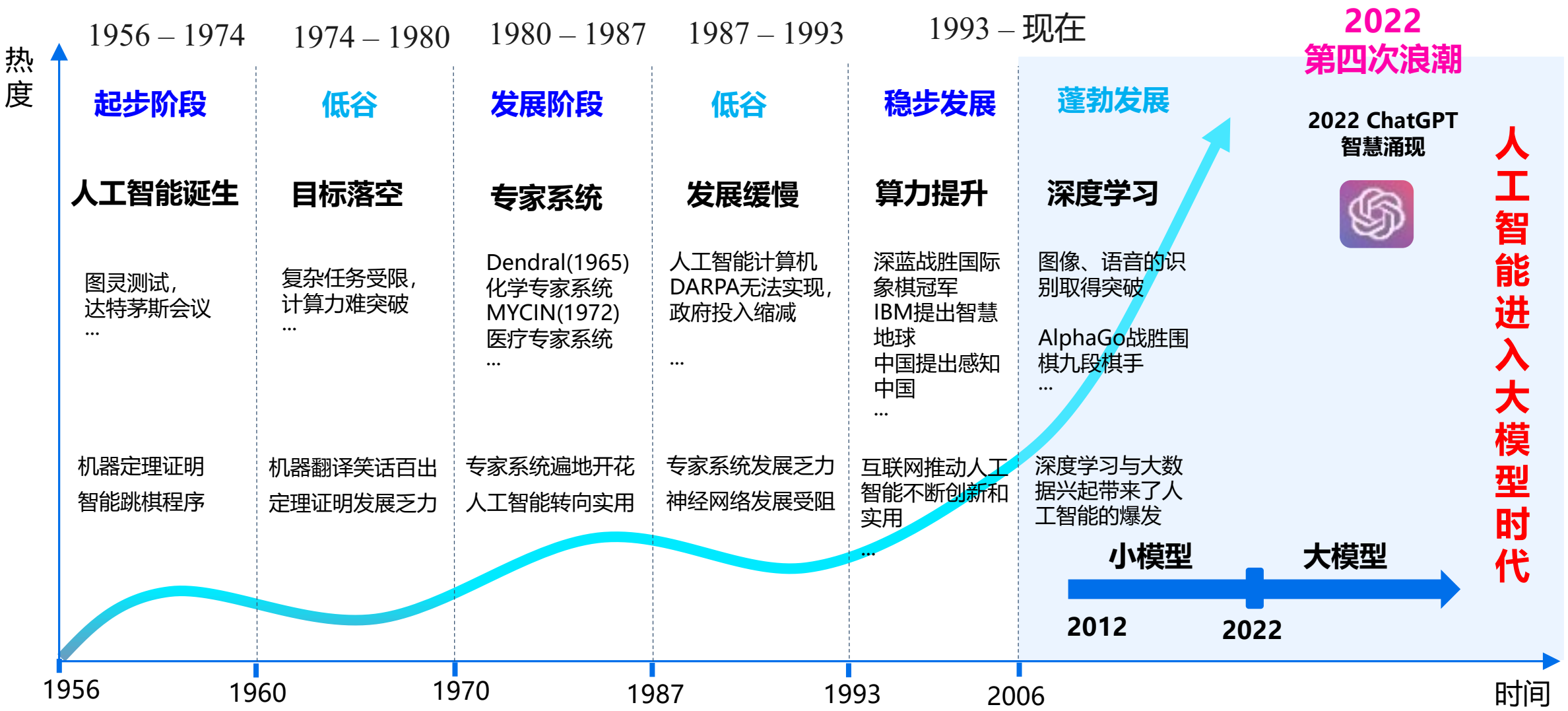
- **1956年8月，在美国达特茅斯学院举办的人工智能夏季研讨会，是人工智能领域具有里程碑意义的一次重要会议**。这次会议汇聚了众多杰出的科学家和工程师，他们共同探讨和研究人工智能的发展和应用前景。
- 这次会议的主题围绕着**人工智能的定义、研究方法和应用场景**展开。与会者们深入探讨了人工智能的基本概念、算法和技术，以及其在各个领域的应用潜力。他们共同认识到，人工智能的研究和发展将为人类带来巨大的变革和进步。
- “人工智能”这个词汇被约翰.麦卡锡（John McCarthy）首次提出



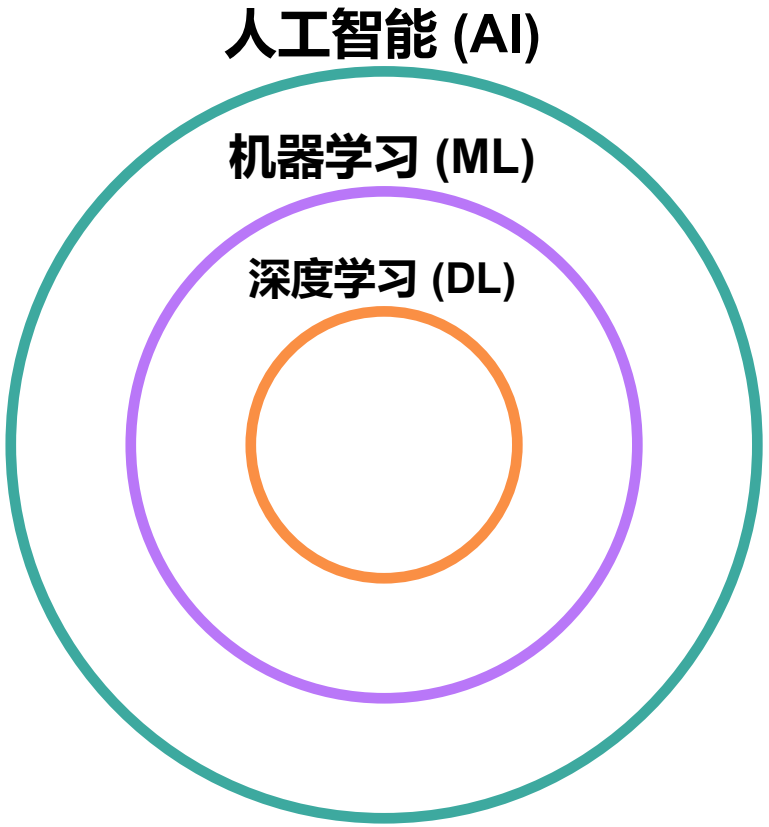




人工智能发展过程


热度
1956 – 1974
起步阶段
人工智能诞生
图灵测试，
达特茅斯会议
…
机器定理证明
智能跳棋程序
1974 – 1980
低谷
目标落空
复杂任务受限，
计算力难突破
…
机器翻译笑话百出
定理证明发展乏力
1980 – 1987
发展阶段
专家系统
Dendral(1965)
化学专家系统
MYCIN(1972)
医疗专家系统
…
专家系统遍地开花
人工智能转向实用
1987 – 1993
低谷
发展缓慢
人工智能计算机
DARPA无法实现，
政府投入缩减
…
专家系统发展乏力
神经网络发展受阻
1993 – 现在
稳步发展
算力提升
深蓝战胜国际
象棋冠军
IBM提出智慧
地球
中国提出感知
中国
…
互联网推动人工
智能不断创新和
实用
…
蓬勃发展
深度学习
图像、语音的识
别取得突破
AlphaGo战胜围
棋九段棋手
…
深度学习与大数
据兴起带来了人
工智能的爆发
2022
第四次浪潮
2022 ChatGPT
智慧涌现
人工智能进入大模型时代
小模型
大模型
2012
2022
1956
1960
1970
1987
1993
2006
时间

人工智能并非单一技术，而是一个技术体系



人工智能 (AI)

定义: 一个广阔的科学领域，旨在构建能够模仿人类智能的 机器，执行看、听、说、分析、决策等认知任务。

通俗理解: 这是我们的终极目标——让机器像人一样“思考”。

机器学习 (ML)

定义: AI的核心子集，让机器从大量数据中自动“学习” 规律和模式，以完成特定任务，而无需编写固定规则。

通俗理解: 给它看大量案例，让它自己总结方法。

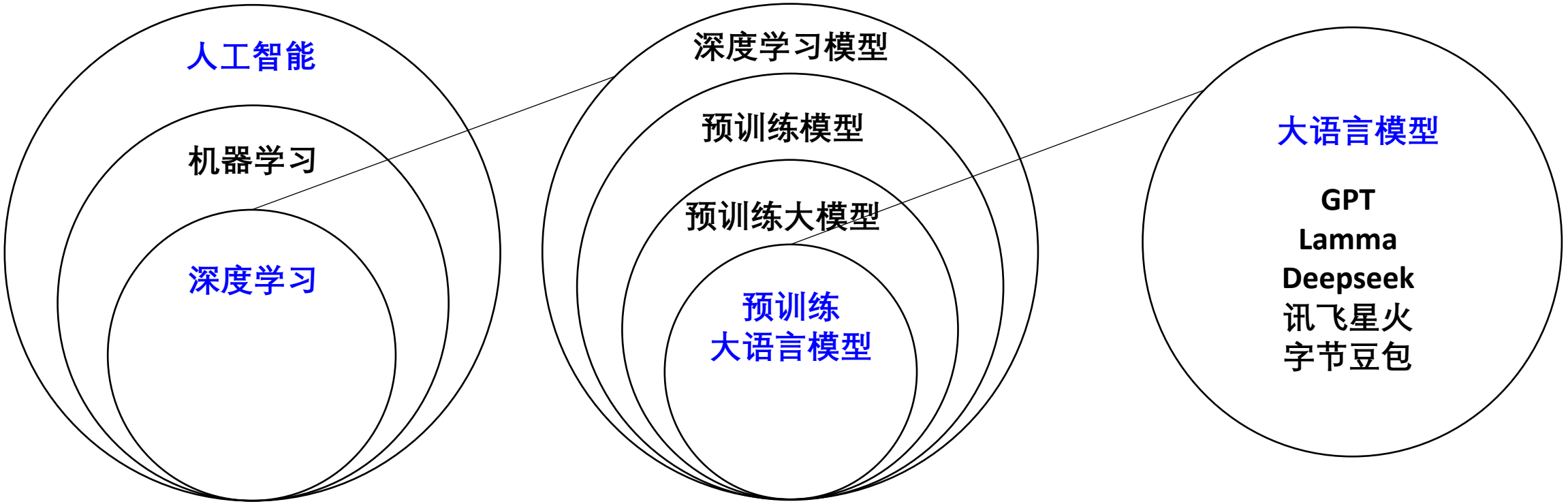
深度学习 (DL)

定义: 机器学习的一种特殊形式，使用复杂的“人工神经网 络”结构，擅长从图像、文本等非结构化数据中发现模式。

通俗理解: 模仿大脑处理信息的方式，能看懂图片、听懂语音。

人工智能的技术关系

人工智能包含了机器学习，机器学习包含了深度学习，深度学习可以采用不同的模型，其中一种模型是预训练模型，**预训练模型包含了预训练大模型（可以简称为“大模型”），预训练大模型包含了预训练大语言模型（可以简称为“大语言模型”）**，预训练大语言模型的典型代表包括OpenAI的GPT和百度的文心ERNIE，ChatGPT是基于GPT开发的大模型产品，文心一言是基于文心ERNIE开发的大模型产品。

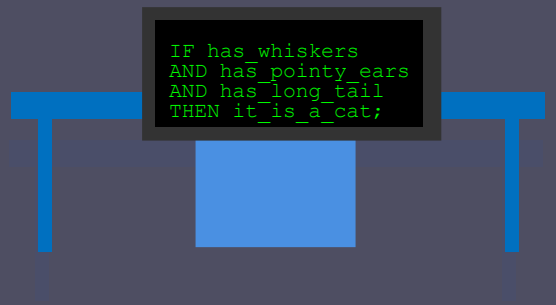


人工智能技术原理—机器学习

机器学习如何工作：一个简单的类比

机器学习的范式转变——从“授人以鱼”（给出规则）到“授人以渔”（学会学习）

传统编程



方法: 程序员试图写下所有关于“猫”的规则。

"IF 有胡须 AND 有尖耳朵 AND 有长尾巴 THEN 是猫"

问题: 方法非常脆弱。规则永远写不完，
也无法覆盖所有情况（例如：无尾猫、折耳猫）。

机器学习

方法: 我们不再编写规则，而是给计算机“喂”入成千上万张已标记的图片（例如，猫 vs 非猫）。

过程: 算法自动分析数据，自己学习和总结“猫”的视觉特征，这些特征可能非常复杂和抽象。

结果: 模型学会了识别猫，并具有很好的泛化能力。

人工智能技术原理—深度学习

生物学启发

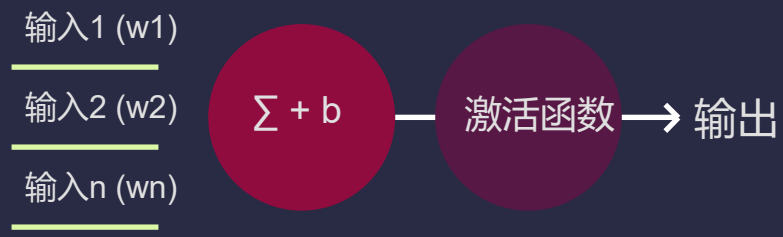
人工神经网络（Artificial Neural Networks, ANNs）的设计灵感，来源于人脑的基本构成单位——生物神经元网络。

基本单元对比

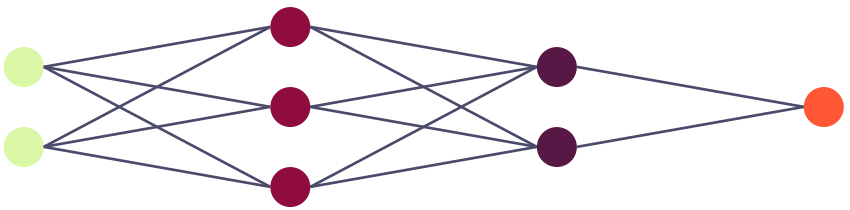
生物神经元

通过树突接收来自其他神经元的信号，在细胞体内进行处理，当信号强度超过某个阈值时，通过轴突将信号传递给下一个神经元。

人工神经元 (节点)



网络结构



成千上万个人工神经元连接在一起，分层排列，就构成了人工神经网络

能够自动发现数据中从微观到宏观的复杂结构和模式

人工智能技术原理—深度学习

深度学习的应用技术

计算机视觉

Computer Vision

“让机器「看懂」世界”

应用领域：

- 人脸识别与安防监控
- 自动驾驶中的障碍物检测
- 医疗影像分析（如识别肿瘤）
- 工业产品缺陷的自动质检等

核心技术:

深度卷积神经网络(CNN)是该领域的关 键模型，它能高效地处理图像数据。

自然语言处理

Natural Language Processing

“让机器「理解」语言”

应用领域：

- 智能搜索引擎、机器翻译
- 情感分析（判断用户评论的情绪）
- 智能客服与问答系统
- 文本自动生成与摘要等

核心技术:

循环神经网络(RNN)、长短期记忆网络(LSTM)以及Transformer是处理序列数据(如文本) 的关键模型

智能语音

Speech Recognition

“让机器「听懂」人类”

应用领域：

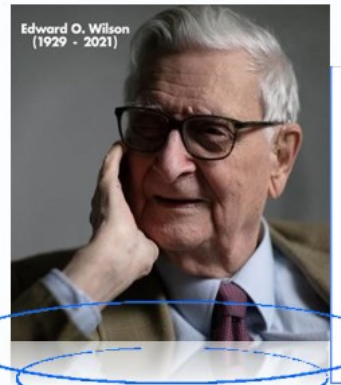
- 智能手机的语音助手（如Siri） -
智能音箱、会议录音实时转文字
- 语音指令控制等

核心技术:

深度学习模型能够有效地将连续的声波信号转换为文本，准确率远超传统方法

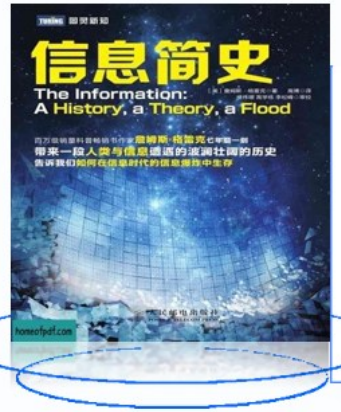
自然语言处理

语言 是人类交流思想、表达情感最自然、最深刻、最方便的工具



“语言是继真核细胞之后最伟大的进化成就”

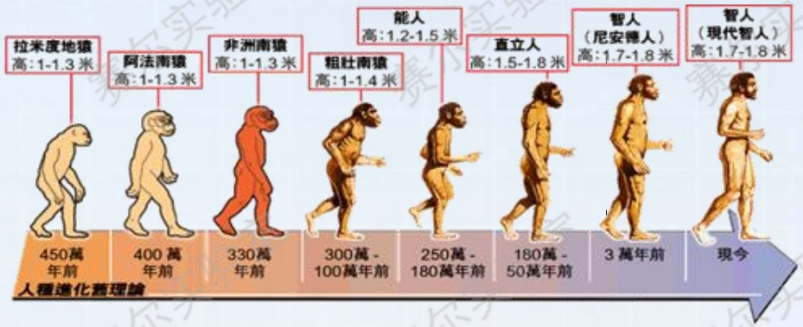
—— 社会生物学之父爱德华·威尔逊



“语言本身就是人类有史以来最大的技术发明”

—— 詹姆斯·格雷克《信息简史》

人类历史上大部分知识是以语言文字形式记载和流传的



自然语言处理

- 认知智能是人工智能的重要阶段，让机器能理解，会思考
- 自然语言理解（NLU）是认知智能的重要内容，通往强人工智能的必经之路

自然语言处理

自然语言处理成为 制约人工智能取得更大突破和更广泛应用的瓶颈

“深度学习的下一个大的进展应该是让神经网络真正理解文档的内容”

——诺贝尔奖得主、图灵奖得主、深度学习之父Geoffrey Hinton

“深度学习的下一个前沿课题是自然语言理解”

——图灵奖得主、Meta AI负责人 Yann LeCun

“如果给我10亿美金，我会建造一个NASA级别的自然语言处理研究项目”

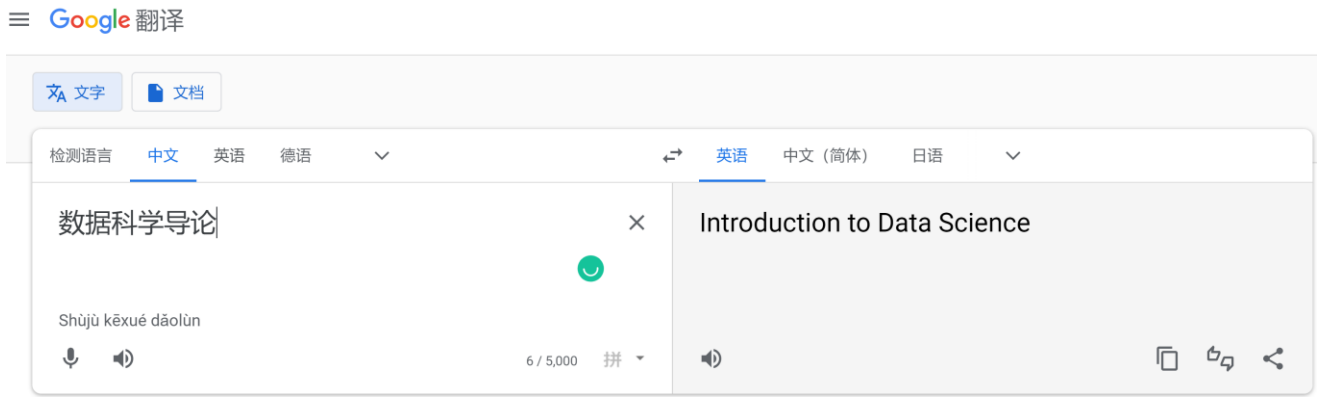
——美国双院院士、世界知名机器学习专家 Michael I. Jordan

“下一个十年，懂语言者得天下”

——美国工程院士、微软前全球执行副总裁沈向洋

自然语言处理

- 自然语言理解——NLU



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
5 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
6 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967

SQuAD2.0
The Stanford Question Answering Dataset

自然语言处理

- 自然语言生成—NLG

哪首诗是人写的?

秋夕湖上	秋夕湖上
一夜秋凉雨湿衣， 西窗独坐对夕晖。 湖波荡漾千山色， 山鸟徘徊万籁微。	荻花风里桂花浮， 恨竹生云翠欲流。 谁拂半湖新镜面， 飞来烟雨暮天愁。
机器	**宋代诗人葛绍体**

Zhe Wang, Haifeng Wang and Enhong Chen, Chinese Poetry Generation with Planning based Neural Network, COLING'2016

自然语言处理

- 自然语言理解在学术界和产业界中都有大量研究和应用
  - 学术界刷新各榜单
  - 产业界落地各类应用带来显著效果

学术榜单

CoQA
A Conversational Question Answering Challenge

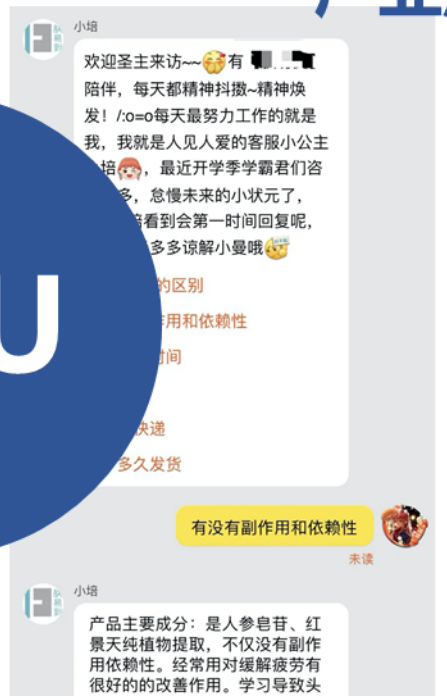
SQuAD2.0
The Stanford Question Answering Dataset

GLUE

SuperGLUE

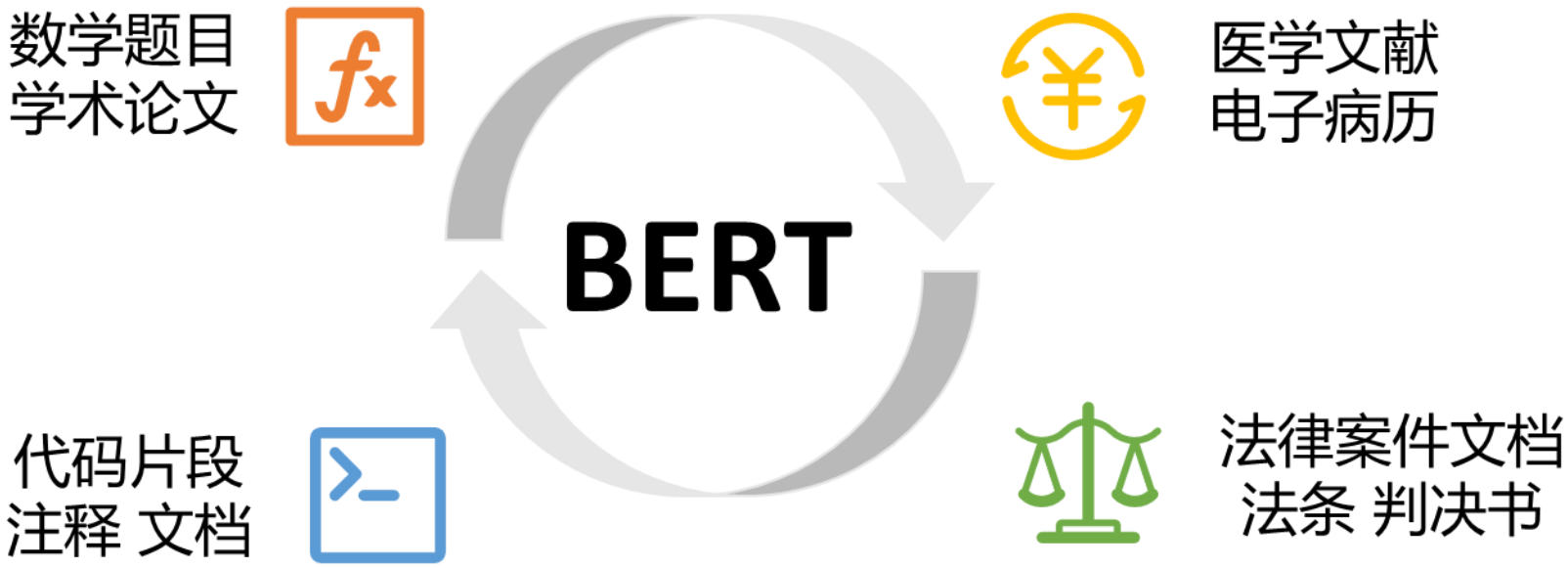
NLU

产业落地

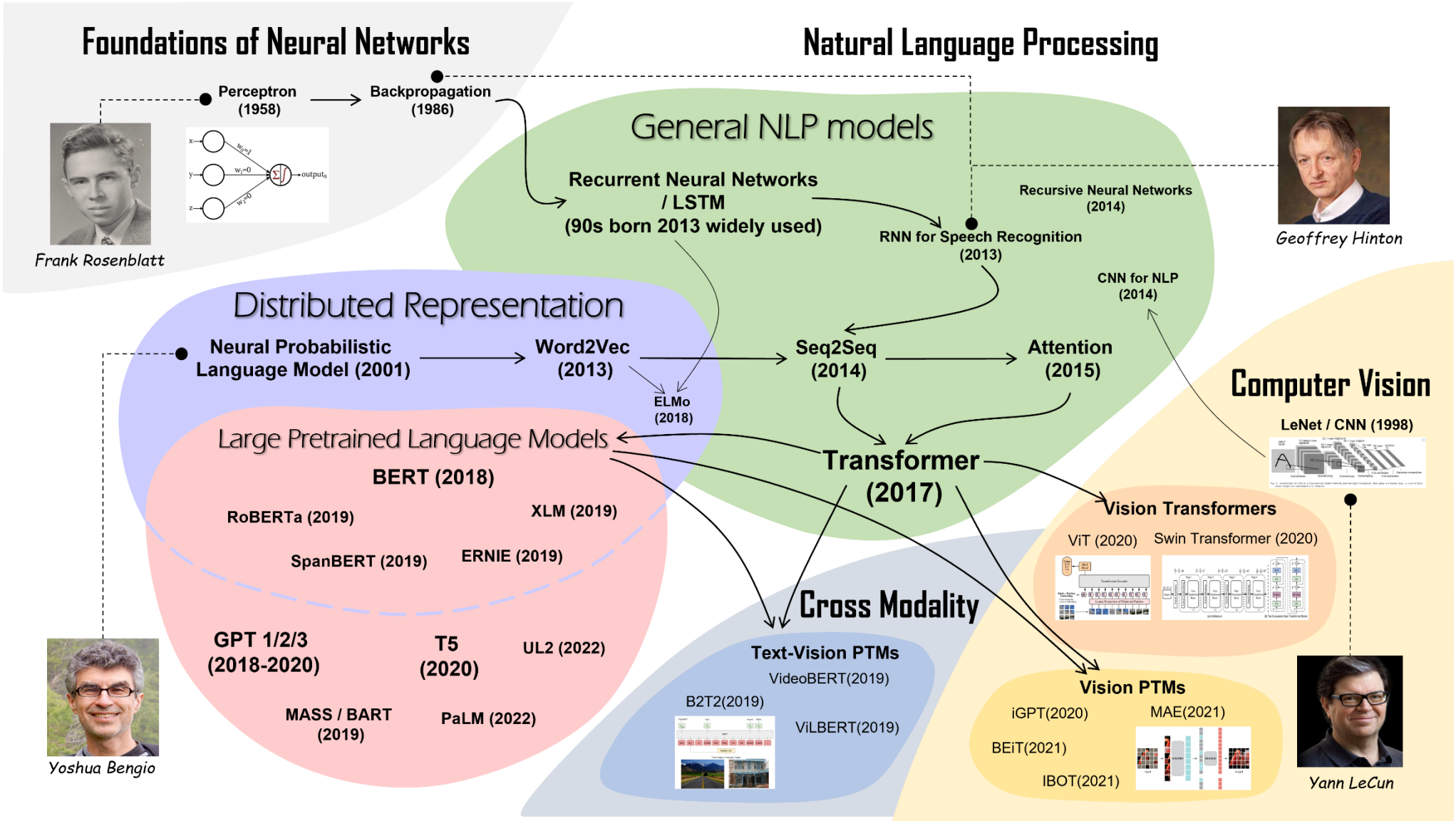


自然语言处理

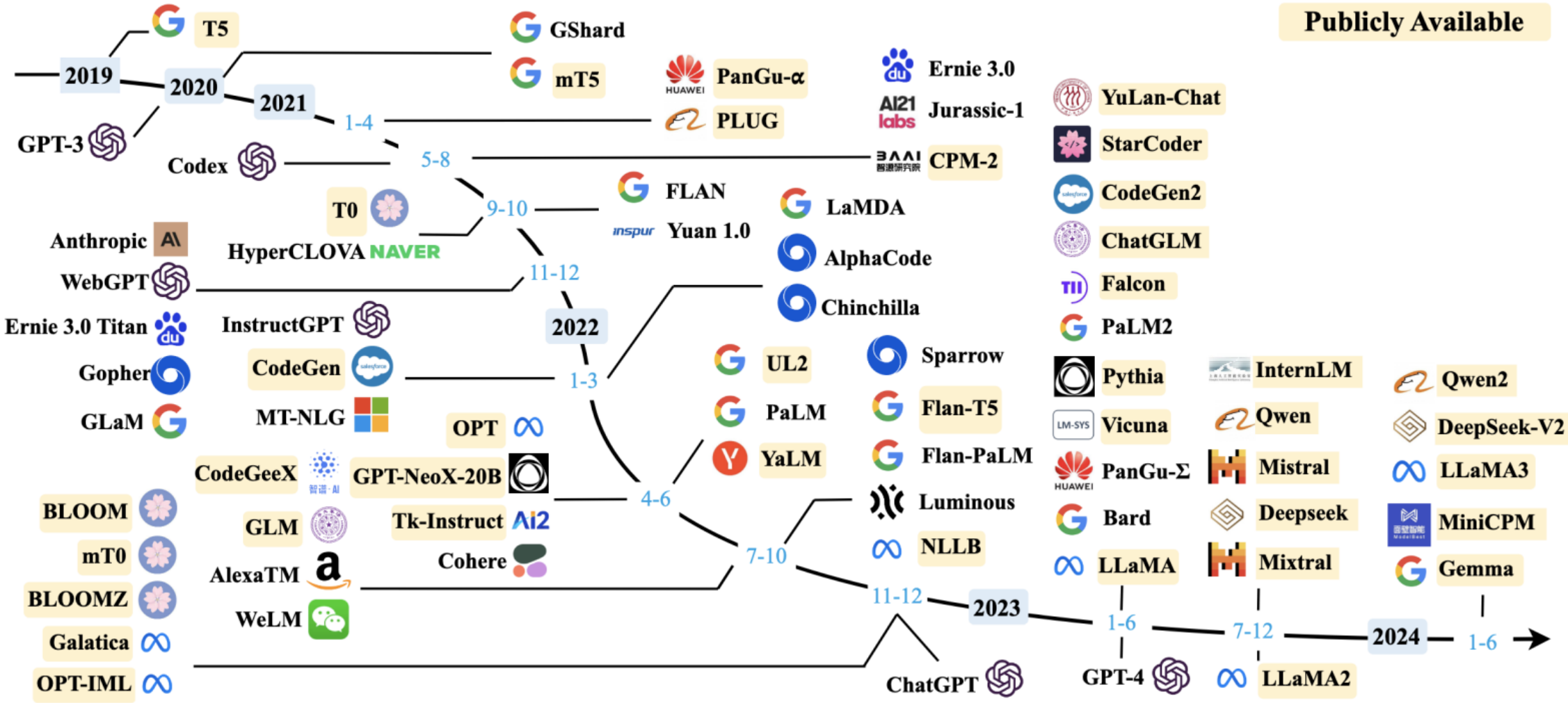
- 自然语言理解被应用于越来越多的领域
  - 法律、数学、代码、医学 ……
- 针对不同领域的**语料和特点**，提出特定的方法



自然语言处理

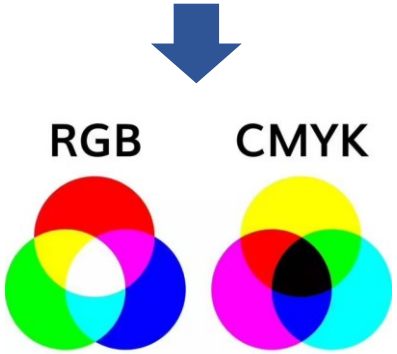

Foundations of Neural Networks
Natural Language Processing
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
General NLP models
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Computer Vision
LeNet / CNN (1998)
Large Pretrained Language Models
BERT (2018)
RoBERTa (2019)
XLM (2019)
SpanBERT (2019)
ERNIE (2019)
GPT 1/2/3 (2018-2020)
T5 (2020)
UL2 (2022)
MASS / BART (2019)
PaLM (2022)
Transformer (2017)
Vision Transformers
ViT (2020)
Swin Transformer (2020)
Cross Modality
Text-Vision PTMs
VideoBERT(2019)
B2T2(2019)
ViLBERT(2019)
Vision PTMs
iGPT(2020)
MAE(2021)
BEiT(2021)
IBOT(2021)
Yoshua Bengio
Yann LeCun

自然语言处理


Publicly Available
2019
2020
2021
2022
2023
2024
T5
GPT-3
GShard
mT5
PanGu-α
PLUG
Ernie 3.0
Jurassic-1
CPM-2
YuLan-Chat
StarCoder
CodeGen2
ChatGLM
Falcon
PaLM2
Codex
T0
HyperCLOVA
Anthropic
WebGPT
Ernie 3.0 Titan
Gopher
GLaM
FLAN
Yuan 1.0
LaMDA
AlphaCode
Chinchilla
InstructGPT
CodeGen
MT-NLG
OPT
CodeGeeX
GPT-NeoX-20B
Tk-Instruct
Cohere
GLM
AlexaTM
WeLM
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
UL2
PaLM
YaLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
ChatGPT
Pythia
Vicuna
PanGu-Σ
Bard
LLaMA
InternLM
Qwen
Mistral
Deepseek
Mixtral
LLaMA2
GPT-4
Qwen2
DeepSeek-V2
LLaMA3
MiniCPM
Gemma
1-4
5-8
9-10
11-12
1-3
4-6
7-10
11-12
1-6
7-12
1-6

自然语言处理

- 自然语言理解难在哪里
  - 语言是**高度抽象的产物**，基本组成单位不是明确的物理量



Outline

- 人工智能
- 自然语言处理发展
- 大模型的原理与技术
- 大模型应用
- 未来展望

人工智能技术原理—自然语言处理

- 语言模型的本质是对一段自然语言的文本预测概率的大小

S1: 语言模型的本质是对一段自然语言的文本进行预测概率的大小
S2: 语言模型的本质是对自然一段语言的文本进行预测概率的大小
S3: 语言模型的本质是对自然语言一段的文本进行预测概率的大小

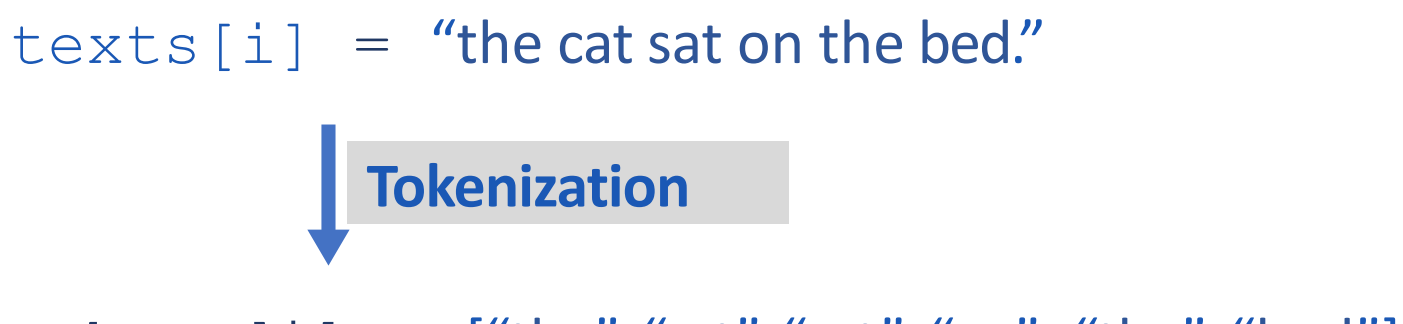
如何判断哪个句子更像一个合理的句子，如何量化评估

$$P(S) = P(w_1, w_2, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1, w_2, \dots, w_{n-1})$$

语言模型 $L = \sum_{w \in C} log\, P(w|context(w))$

自然语言处理

- Step1: 词条化（Tokenization）
  - 将给定的字符序列拆分成一系列子序列的过程
  - 其中，每个子序列被称为一个词条（Token）
  - 词条化的主要任务就是确定正确的词条，并避免标点等因素干扰



自然语言处理

- 英文分词
  - 词与词组的切分
    - To be or not to be…
  - 标点符号的影响
    - 连字符：Self-motivation
    - 引号：大鲨鱼奥尼尔（O’Neal）
  - 专有名词的拆分
    - New York University or New / York University?

自然语言处理

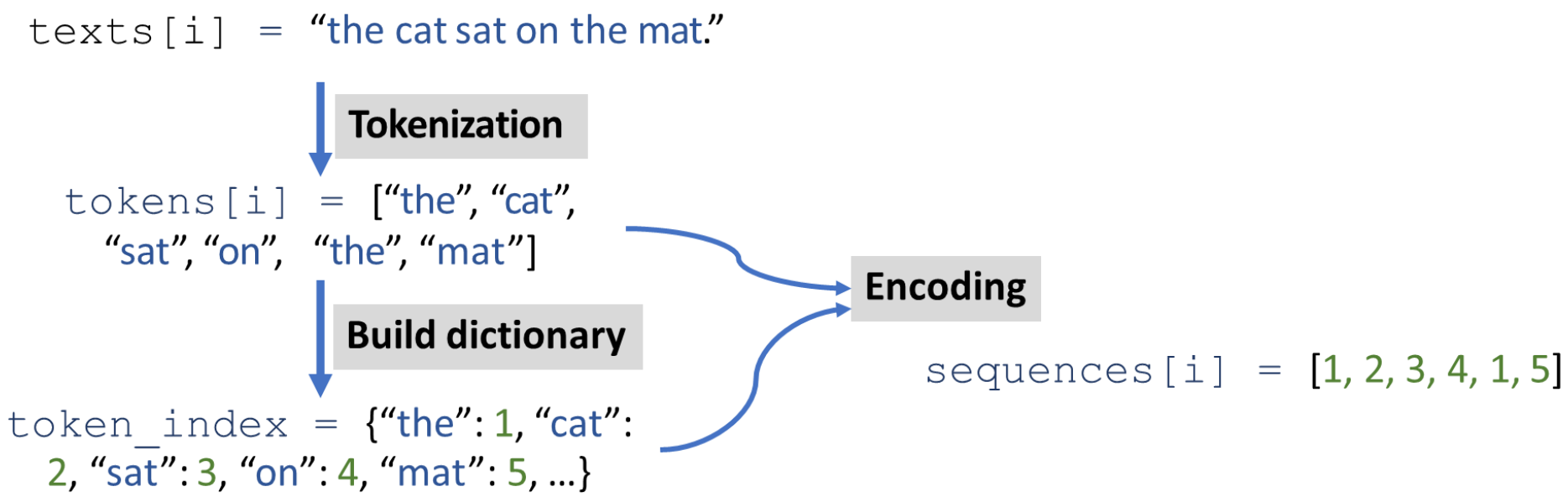
- 中文分词
  - 语素 是最小的语音语义结合体，是最小的语言单位。
  - **“字”**：简单高效，表示能力较差，不能独立地完整地表达语义信息。
    - 国家标准GB2312-80 中定义的常用汉字为6763个。
  - **“词”**：具有固定的语音形式，可以独立运用的最小的语言单位。
    - 词的表示能力较强，但汉语词的个数在10万个以上，面临复杂分词问题
  - 最大的挑战：没有显式分隔符（如空格）
    - 英语可视作词的集合，而汉语则是字的集合
    - 无显式分隔符使分辨不同组合方式更加困难
  - 中文对虚词的运用：不单独表意，但影响句意
    - 古虚词：之乎者也，现代虚词：的、了、吧…
  - 分词歧义、未登录词等

自然语言处理

- 分词可能带来的隐患
  - 分词带来的大量低频词，导致严重的数据稀疏。越来越多的OOV（Out of Vocabulary）词。
  - 分词中难免的错漏将导致额外的噪声
- 去停用词
  - 停用词（Stopwords）指文档中频繁出现或对实际语义影响不大的词语。
    - 例如，英文中的The、of，中文中的“的”、“是”等。
    - 数字、副词等与语义关系不大的词常作为停用词被处理。
  - 为什么要去除停用词?
- 停用词类型与识别
  - 停用词的设置与语料库的性质有关
    - 例如，URL中的www，Wikipedia中的wiki
  - 常用的停用词识别方法
    - 较为成熟的停用词识别方法有: 文本频率、词频统计、熵计算等。
  - 常用的停用词表：哈工大停用词表、百度停用词表、 NLTK停用词表等

自然语言处理

- Step 2: 建立词表（dictionary）
  - 每个词都可以用一个正整数表示



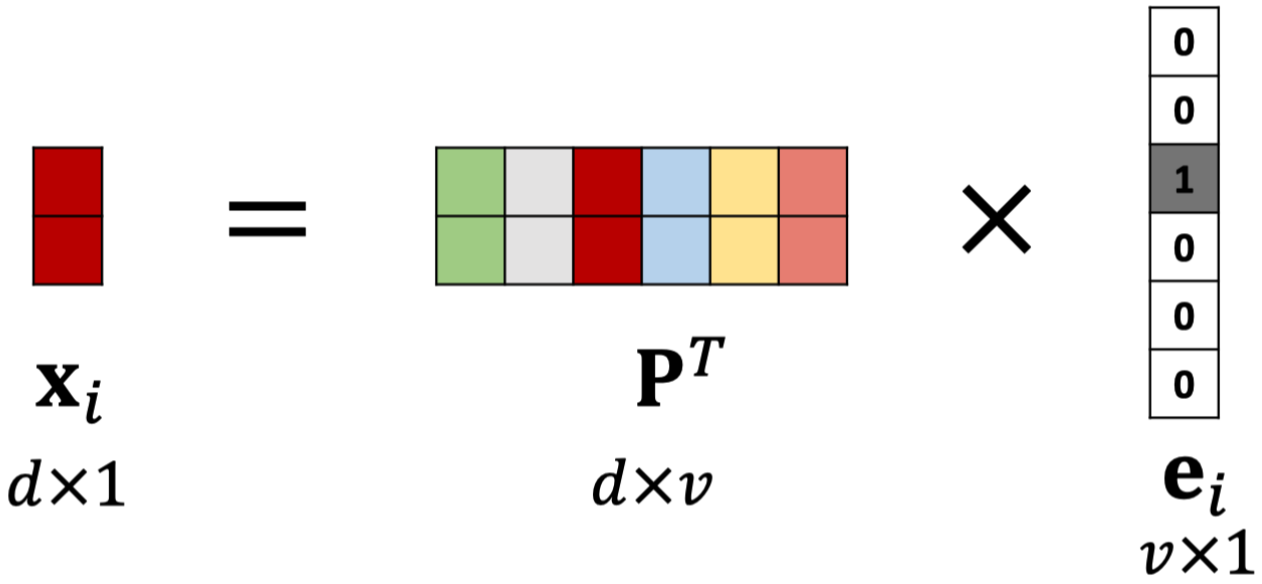
自然语言处理

- Step3: 词嵌入（Word Embedding）
  - 独热编码（One-hot encoding）
    - 假设字典里的单词数量为v
    - 每个单词使用一个v维向量表示，一个位置为1，其余为0
    - 问题1：维度太大，向量的维度等于词表大小
    - 问题2：任意两个向量都是正交的，独热编码不存在相似的概念

Word	Index	One-hot encoding
“movie”	1	$\mathbf{e}_1 = [1, 0, 0, 0, 0, \cdots, 0]$
“good”	2	$\mathbf{e}_2 = [0, 1, 0, 0, 0, \cdots, 0]$
“fun”	3	$\mathbf{e}_3 = [0, 0, 1, 0, 0, \cdots, 0]$
“boring”	4	$\mathbf{e}_4 = [0, 0, 0, 1, 0, \cdots, 0]$
…	…	…

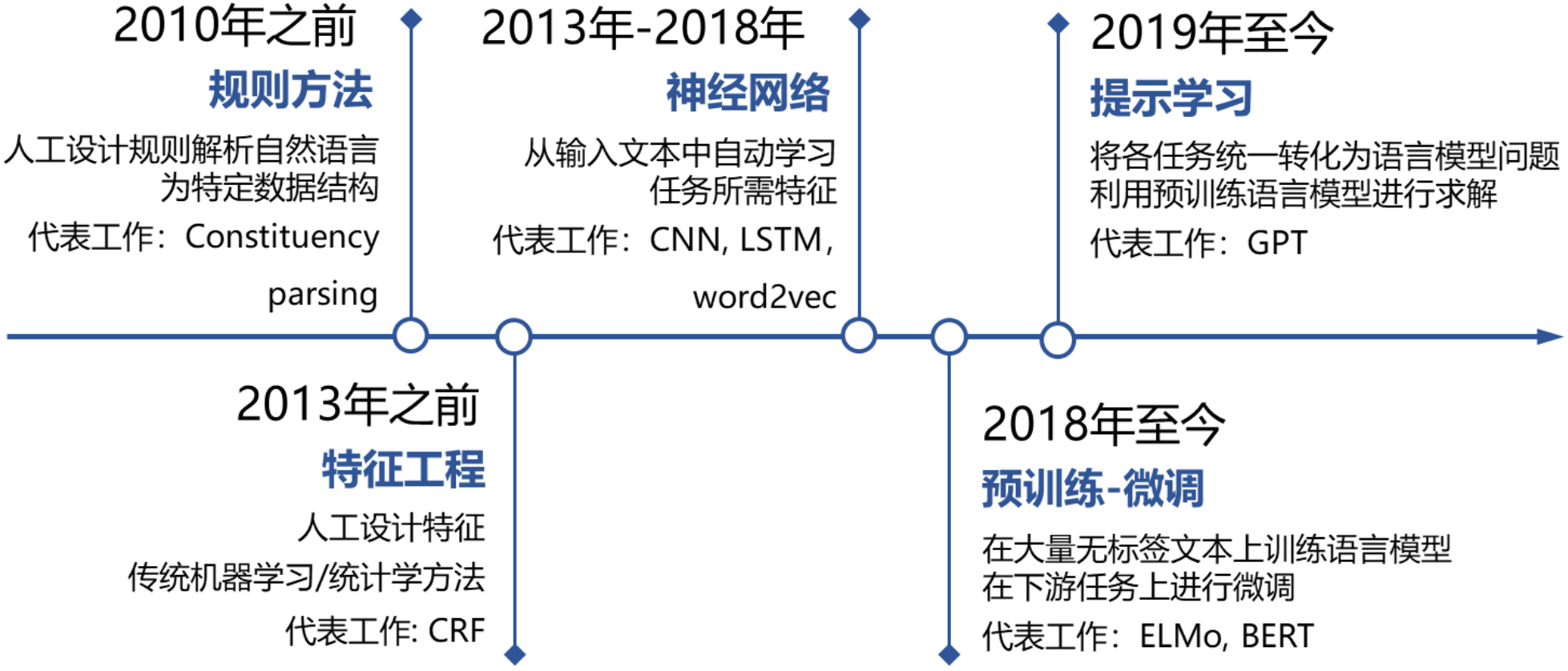
自然语言处理

- 词嵌入（Word Embedding）
  - 为了解决上述问题，词嵌入方法将one-hot向量映射成低维向量
  - D为词向量的维度，由用户自己决定，d的大小会影响模型的表现
  - 参数矩阵P是模型在训练过程中学习得到的



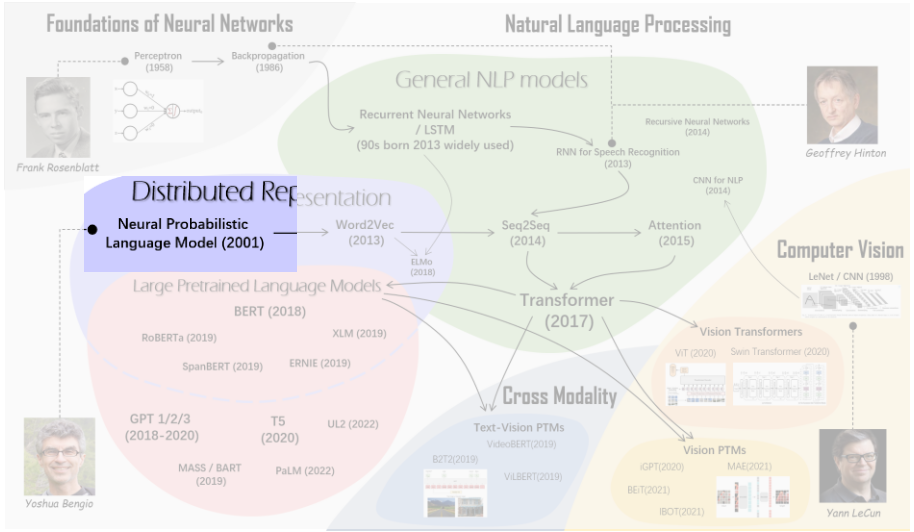
自然语言处理

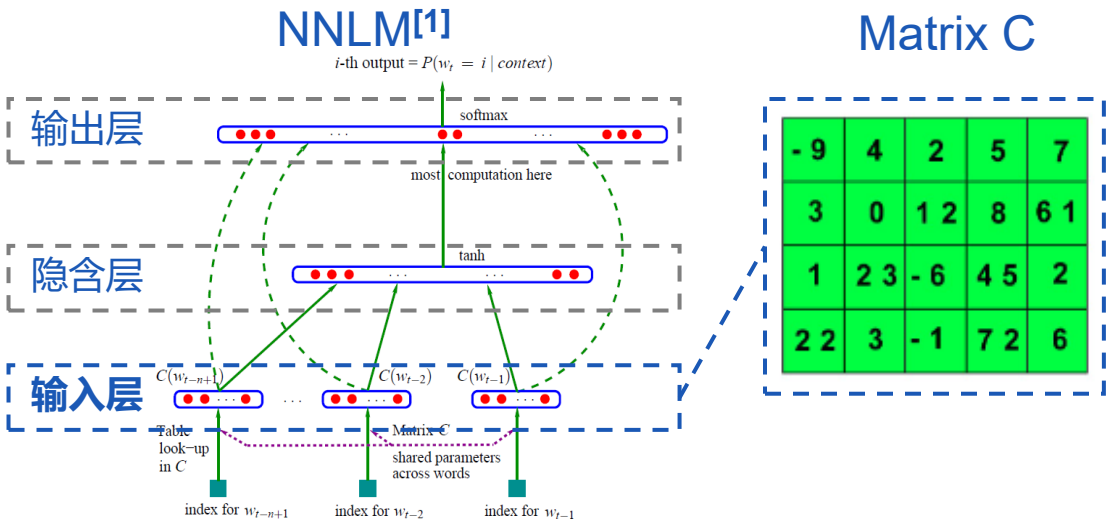
- NLP方法发展趋势



自然语言处理

- NNLM（2001）——使用前馈神经网络搭建的语言模型
  - NNLM包含输入层、隐含层、输出层，其中输入层是一个矩阵C，每来一个单词都相当于从矩阵C中查找代表自己的一行数字，是word embedding的原型

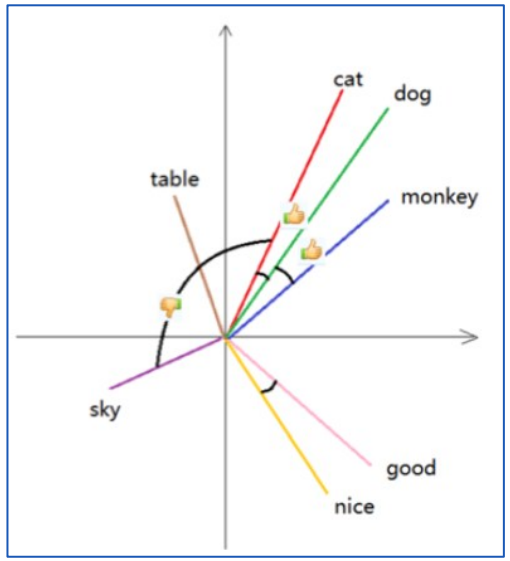




自然语言处理

- Word2Vec词向量能表示词语的语义信息
  - 语义相似性

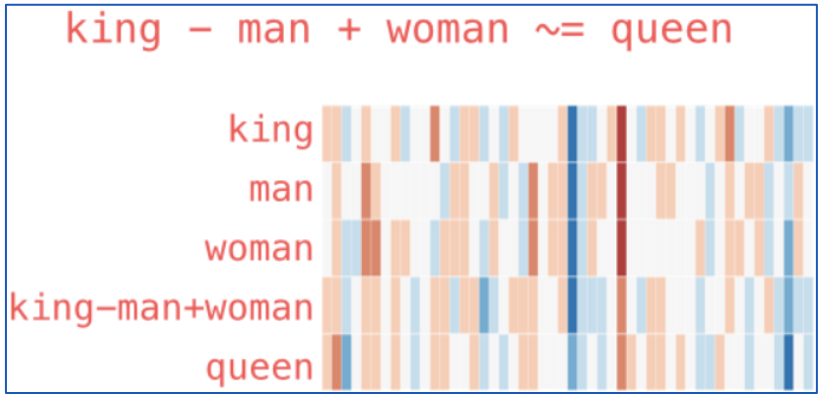
语义相似的词在向量空间的距离更近



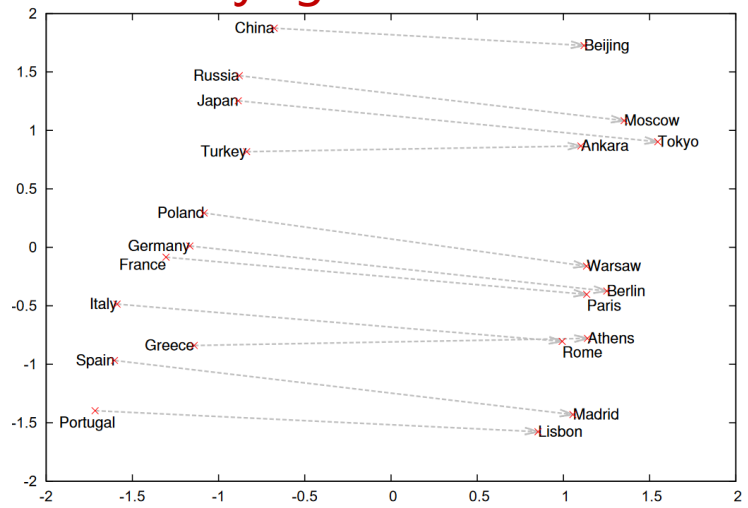
  - 语义类比

语义类比关系表示为词向量运算关系

king – man ≈ queen - woman

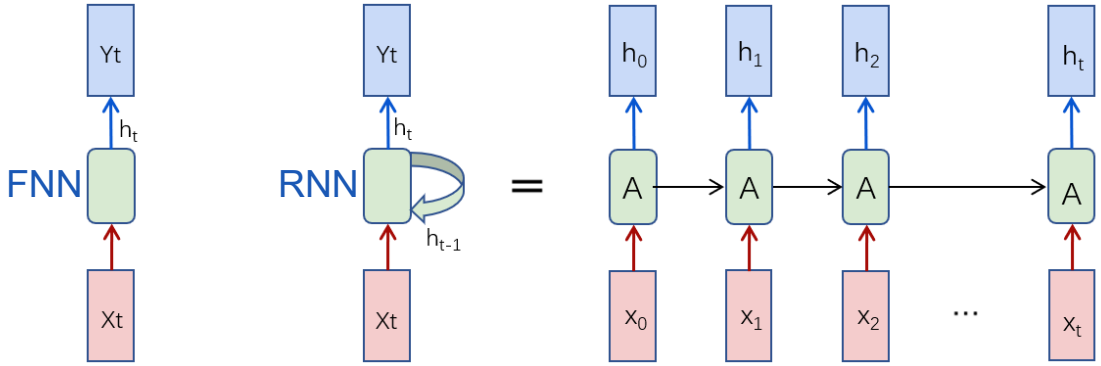


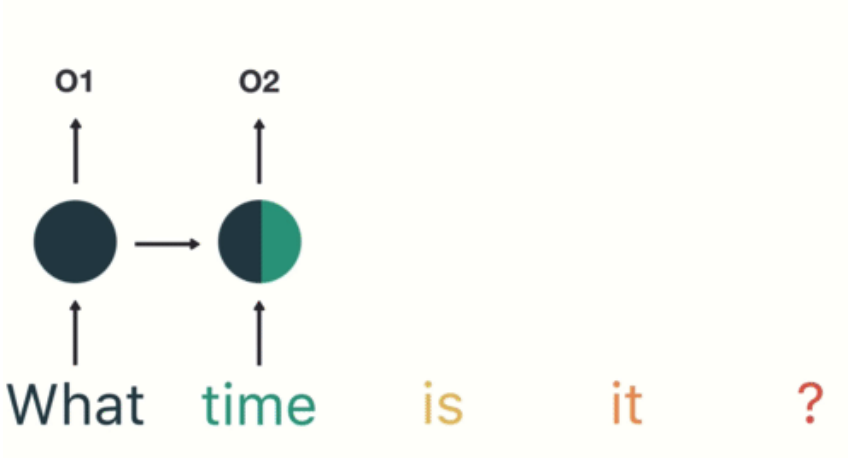
China – Beijing ≈ Russia - Moscow



自然语言处理

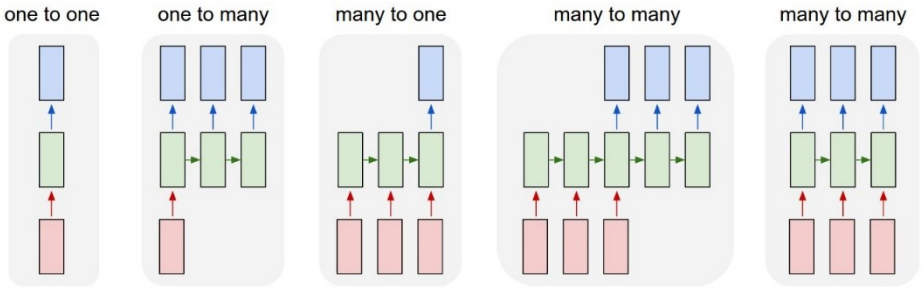
- RNN / LSTM（2013）——从时间的角度对序列建模





one2one：输入输出完全固定，**RNN无需循环**，如图像分类任务

one2many：输入是固定大小，输出是变长序列，**RNN根据输出需要多次循环**，如图像标题生成



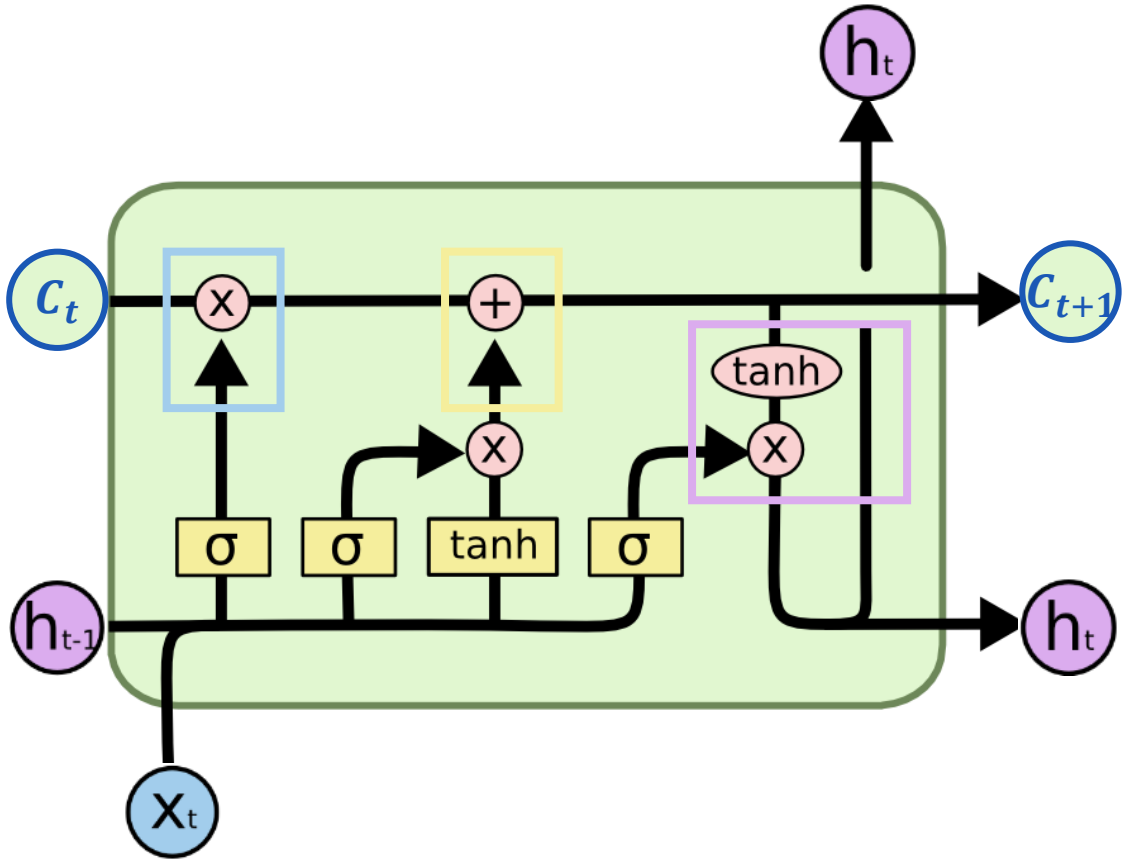
many2one：输入是变长序列，输出是固定大小，**RNN根据输入需要多次循环**，如情感分类

many2many：输入输出是同步变长序列，**RNN根据输入需要多次循环**，如序列标注

many2many：输入输出是异步变长序列，**RNN分别根据输入输出的需求多次循环**，如机器翻译

自然语言处理

- **RNN / LSTM（2013）——从时间的角度对序列建模**
  - 保证长时记忆不遗忘



遗忘门

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

输入门

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

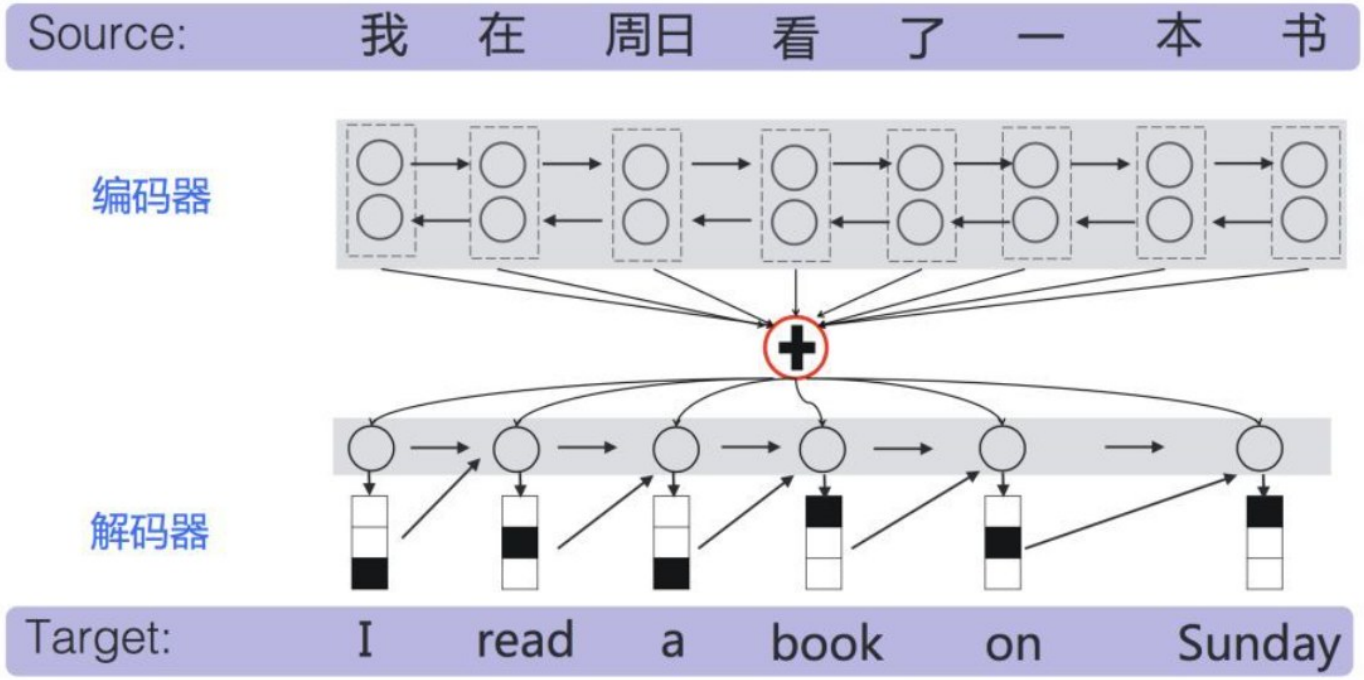
输出门

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

人工智能技术原理—自然语言处理

- **标志技术：Seq2Seq（2014）——序列到序列的神经网络框架**
  - 将序列到序列的过程解耦成序列的表示和生成，Encoder和Decoder分别对应上述功能



谷歌翻译于2016年宣布开始使用神经机器翻译模型，该模型正是应用了Seq2Seq框架