

2023-2024学年第一学期本科生课程

《神经网络与深度学习》

第一节：课程导入

主讲人：戴金成(副教授，博士生导师)

daijincheng@bupt.edu.cn

神经网络与深度学习课程组



北京郵電大學

Beijing University of Posts and Telecommunications

内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



深度学习涉及的基本概念



实验平台介绍(助教)

内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



深度学习涉及的基本概念



实验平台介绍(助教)

为什么选择本课程？



海绵·有志青年·宝宝



章鱼·怨种打工人·哥

为了给今后的学术科研和个人成长打下好基础，为我校双一流学科建设做出..... (以下省略100字)

算法岗，好搞钱。现在学好深度学习，毕业进大厂搞算法发财！

别问我，我就为了蹭个学分.....

不知道，妈妈没教.....



蟹·赚钱爱好者·老板

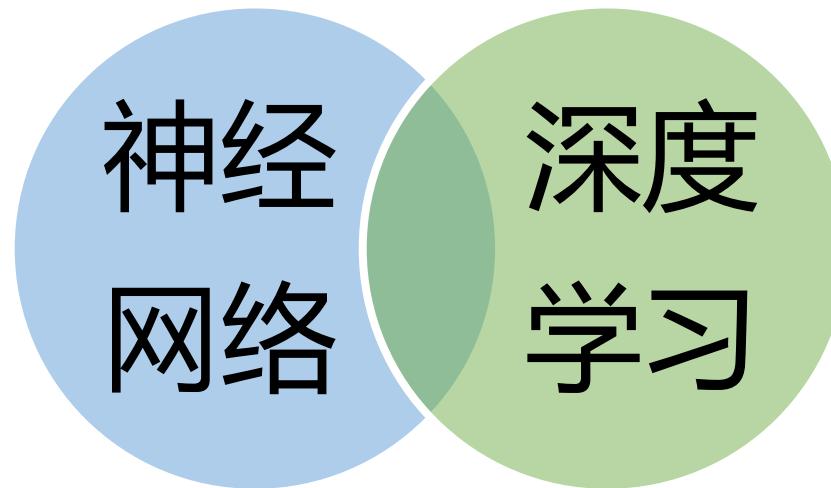


派大·能活一天是一天·星

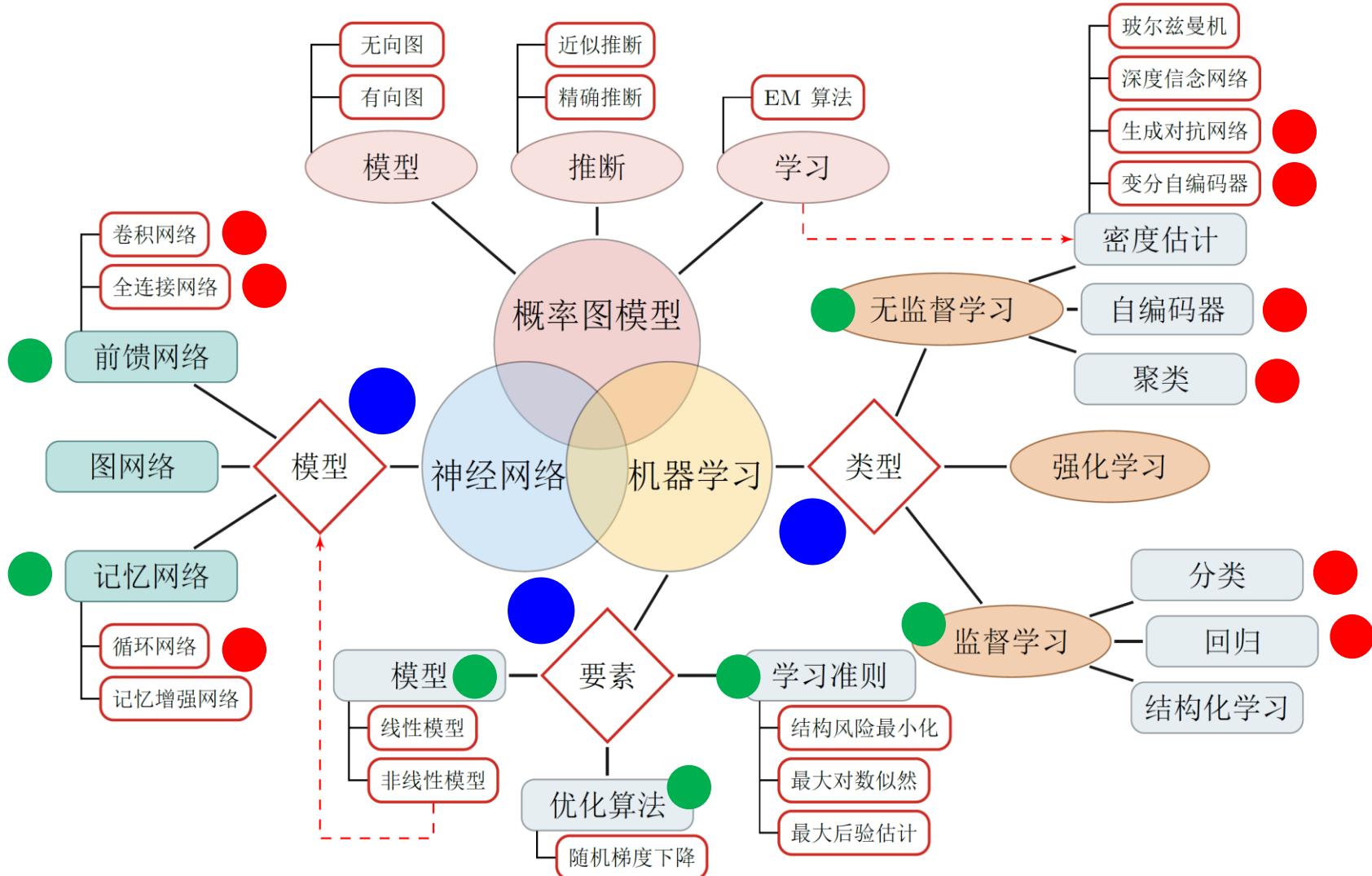
关于本课程

人工智能的一个重要子领域

- 神经网络：一种以（人工）神经元为基本单元的模型
- 深度学习：一类机器学习问题，主要解决贡献度分配问题。



更详细的课程概括



教学安排

教学周	大 纲	内 容
3-7 (国庆不上课)	深度学习基础 (实验作业1)	课程导入, 深度学习概述
		机器学习, 线性回归
		线性与非线性分类
		多层感知机
		常用激活函数和损失函数
8-9	计算机视觉基础 (实验作业2)	卷积神经网络(CNN)
		经典卷积神经网络模型
10-11	模型优化	网络参数优化算法
		网络正则化方法
12	无监督学习基础 (实验作业3)	无监督学习基本概念
		自编码器(AE)

教学安排

教学周	大 纲	内 容
13-14/15	自然语言理解基础 (实验作业4)	循环神经网络(RNN), LSTM
		序列到序列模型(Seq2Seq)
		注意力机制(Attention), 自注意力, 多头注意力, 交叉注意力
		词嵌入, Transformer, 逐层归一化
		BERT, GPT
14/15-17	生成模型 (实验作业5)	生成问题概述
		变分自编码器(VAE)
		生成对抗网络(GAN), GAN的改进
18	总复习	总复习

推荐教材

□ 主要教材：邱锡鹏，神经网络与深度学习

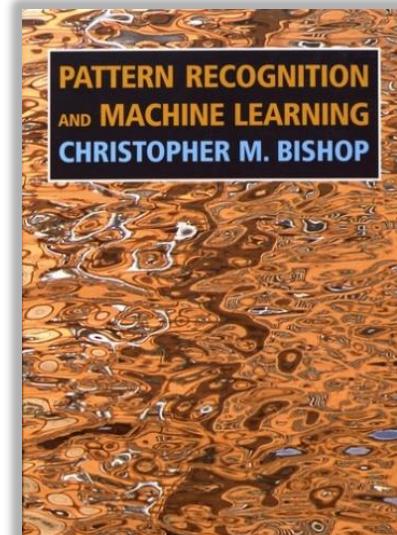
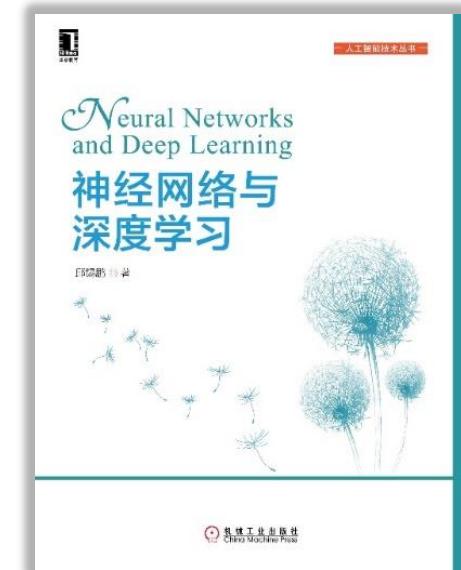
，机械工业出版社, 2020, ISBN
9787111649687

➤ 电子版教材同步至北邮教学云平台

□ 其他参考教材：Bishop, C.M. (2006).

Pattern recognition and Machine
Learning. Springer. ISBN
9780387310732.

➤ 网上有中文版 by 马春鹏



推荐网上课程

□ 李宏毅机器学习系列课程

➤ <https://www.bilibili.com>

□ 斯坦福大学CS224n: Deep Learning for Natural Language Processing

➤ <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/>

➤ Chris Manning 主要讲解自然语言处理领域的各种深度学习模型

□ 斯坦福大学CS231n: Convolutional Neural Networks for Visual Recognition

➤ <http://cs231n.stanford.edu/>

➤ Fei-Fei Li Andrej Karpathy 主要讲解CNN、RNN在图像领域的应用

□ 加州大学伯克利分校 CS 294: Deep Reinforcement Learning

➤ <http://rail.eecs.berkeley.edu/deeprlcourse/>

课外阅读：AI顶会论文

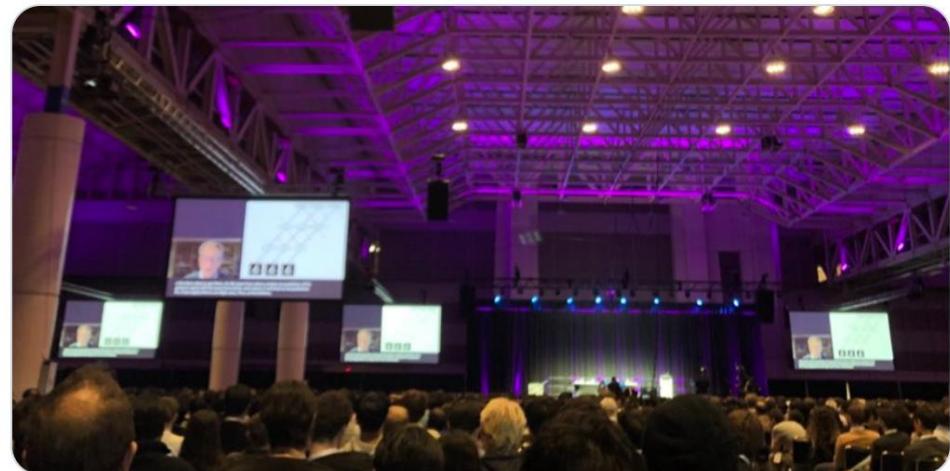
- NeurIPS、ICLR、ICML
- ACL、EMNLP
- CVPR、ICCV、ECCV
- AAAI、IJCAI
- ...



NeurIPS Conference @NeurIPSConf · 6小时

...
@geoffreyhinton is a real Rockstar! He packs a 10k room without being in it!

He presents an insightful technical keynote on the next big thing!
Go check it out



助教与课程群

口课程助教



吕守晔

口课程微信群

联系方式：

lyushouye@bupt.edu.cn

若需答疑或其他事宜，请在周一至
周五9:00-17:00工作时间进行



该二维码7天内(9月16日前)有效，重新进入将更新

内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



深度学习涉及的基本概念



实验平台介绍(助教)

课程考核

□ 平时实验成绩(60%)

- 5次实验作业，若干不同分值的候选题目，需要从中选够满分 ≥ 100 分的题量(**可以选题总分超过100，但平时成绩满分不超过100分×60%**)
- 每次作业提交代码和实验报告，编程语言：Python，深度学习框架：PyTorch/TensorFlow
- 可以使用ChatGPT，但实验报告必须标明哪些地方用到了ChatGPT帮助
- 切勿抄袭，抄袭作业or不标明ChatGPT辅助的作业将会被判为0分

□ 期末考试(40%)

- 开卷考试
- 只允许携带纸质材料，不允许携带电子设备

内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



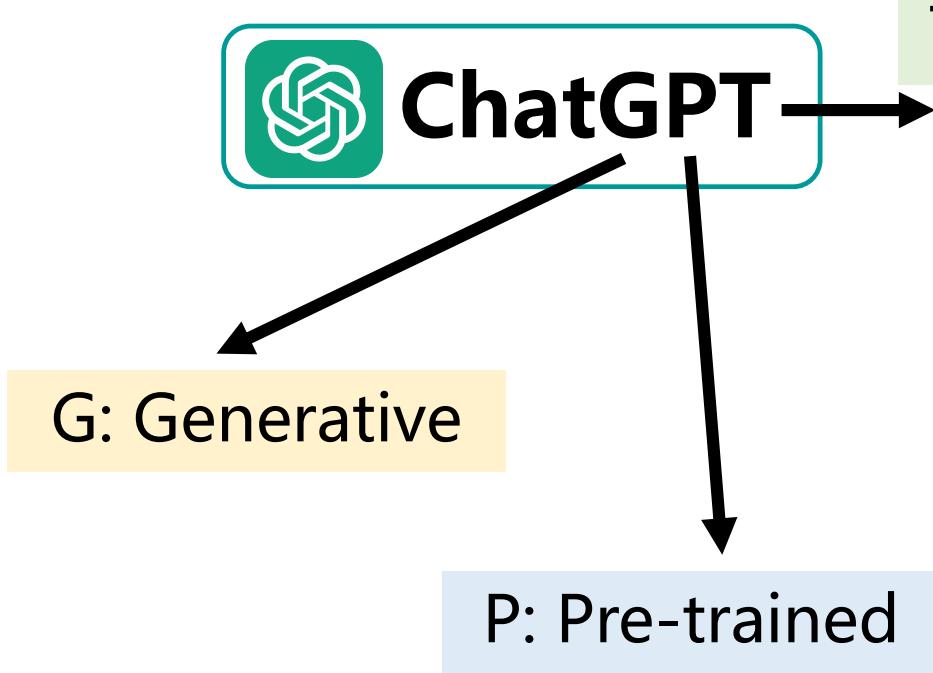
深度学习涉及的基本概念



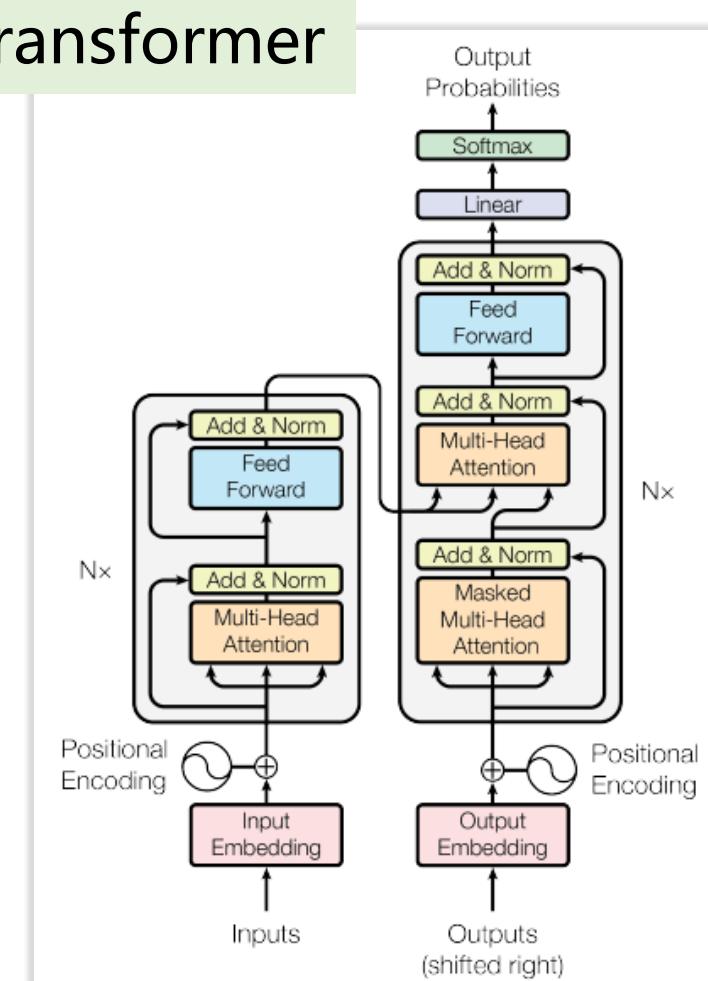
实验平台介绍(助教)

深度学习关键技术：从ChatGPT谈起

口 GPT的全称是什么？从中你能发现哪些关键技术？



关键技术：注意力机制，生成模型，预训练技术，强化学习...



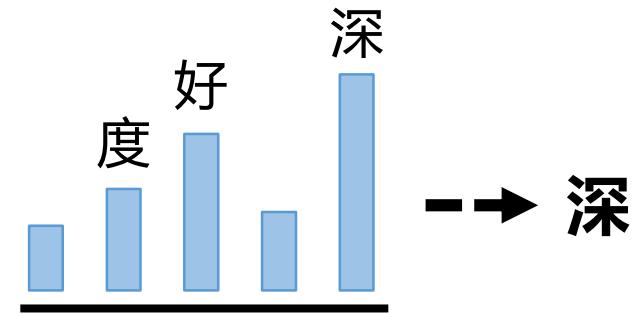
ChatGPT真正在做的事：文字接龙

一个函数(function)

什么是深度学习？



f

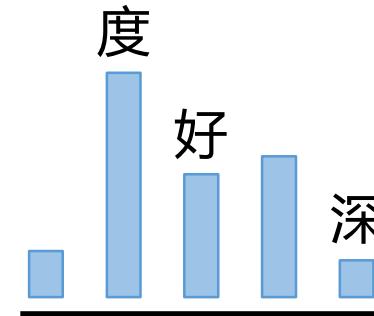


→ 深

什么是深度学习？深



f



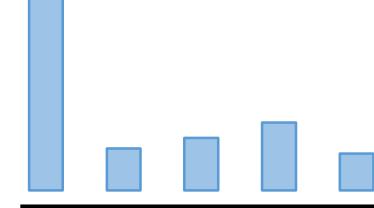
→ 度

什么是深度学习？深度学习是机器学习的一个子领域，.....。



f

END



→ END

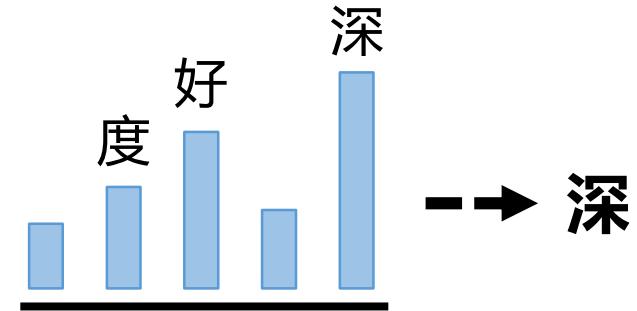
ChatGPT真正在做的事：文字接龙

一个函数(function)

什么是深度学习？



f



→ 深

SE 什么是深度学习？



深度学习是机器学习的一个子领域，它模拟人类大脑中神经元的工作方式，以尝试自动从数据中学习和提取特征。以下是一些关键概念和点：



- 神经网络：**深度学习的核心是人工神经网络，特别是那些有很多层的神经网络，因此得名“深度”。每一层都从其前一层的输出中学习一些特征。
- 数据驱动：**与传统的编程方法相比，深度学习是数据驱动的。这意味着模型的性能通常与提供给它的数据量成正比。
- 自动特征提取：**传统的机器学习方法通常需要手动选择和提取特征。而在深度学习中，模型能够自动从原始数据中学习和提取合适的特征。
- 应用广泛：**深度学习已经在许多领域取得了突出的成绩，包括图像识别、自然语言处理、语音识别、游戏（如围棋）等。
- 训练需要的资源：**深度学习模型通常需要大量的数据和计算资源来训练，特别是对于非常复杂的任务。这是近年来由于大数据的兴起和GPU计算能力的提高，深度学习才得以实现的主要原因。
- 网络结构的多样性：**除了常见的前馈神经网络，还有很多其他类型的网络结构，如卷积神经网络。

Regenerate

Send a message



SE 什么是深度学习？



深度学习是机器学习的一个子领域，它使用了特定的算法结构——深度神经网络，来模拟人脑中神经元的结构和功能，从而从数据中自动学习和提取特征。以下是关于深度学习的一些关键点：



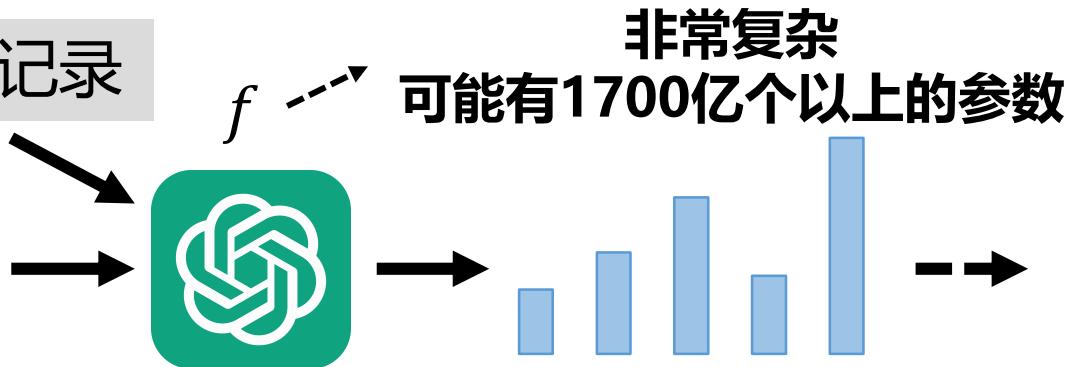
- 神经网络和深度：**深度学习最常见的模型是人工神经网络，特别是深度神经网络。“深度”一词意味着网络有多个连续层，这些层用于从输入数据中提取不同级别的特征。
- 自动特征提取：**与传统的机器学习方法不同，深度学习能够自动从数据中提取和学习特征，而无需手动设计。
- 大数据和计算：**深度学习模型通常需要大量的标记数据和强大的计算能力来训练，尤其是当处理复杂任务时。随着数据量的增加和硬件（如GPU）的进步，深度学习在各种任务上都取得了显著的成功。
- 多种应用：**深度学习在许多应用中都已证明其效果，包括图像识别、语音识别、自然语言处理和强化学习等。
- 常见的网络类型：**虽然存在许多种类型的深度学习结构，但一些比较常见的包括卷积神经网络。

Regenerate

ChatGPT：大语言模型

同一段对话过去的历史记录

请给我一些建议



SE 请想象你是我的朋友，我会对你抱怨，希望你可以用中文提供安慰，并试图跟我聊聊，在对话过程中请展现出同情心，现在我们开始。



当然可以，我会尽量为你提供安慰。



你：今天真的很糟糕，感觉一切都不顺心。

我：哎，听起来你今天真的很不容易。发生了什么事情让你这么难过吗？



最近工作很累，事情堆积如山！



我：哎，真是太不容易了。现代的工作压力确实很大，特别是当任务堆积的时候。住，你不是一个人在战斗，我会尽我所能陪伴和支持你。有时候，当我们深陷困境人分享你的感受可能会让心情好一些。你想详细地说说你的工作中遇到的难题吗？聊会让你感觉轻松些。



下周就要上课了，所以这周末也不能干别的，就要好好备课了

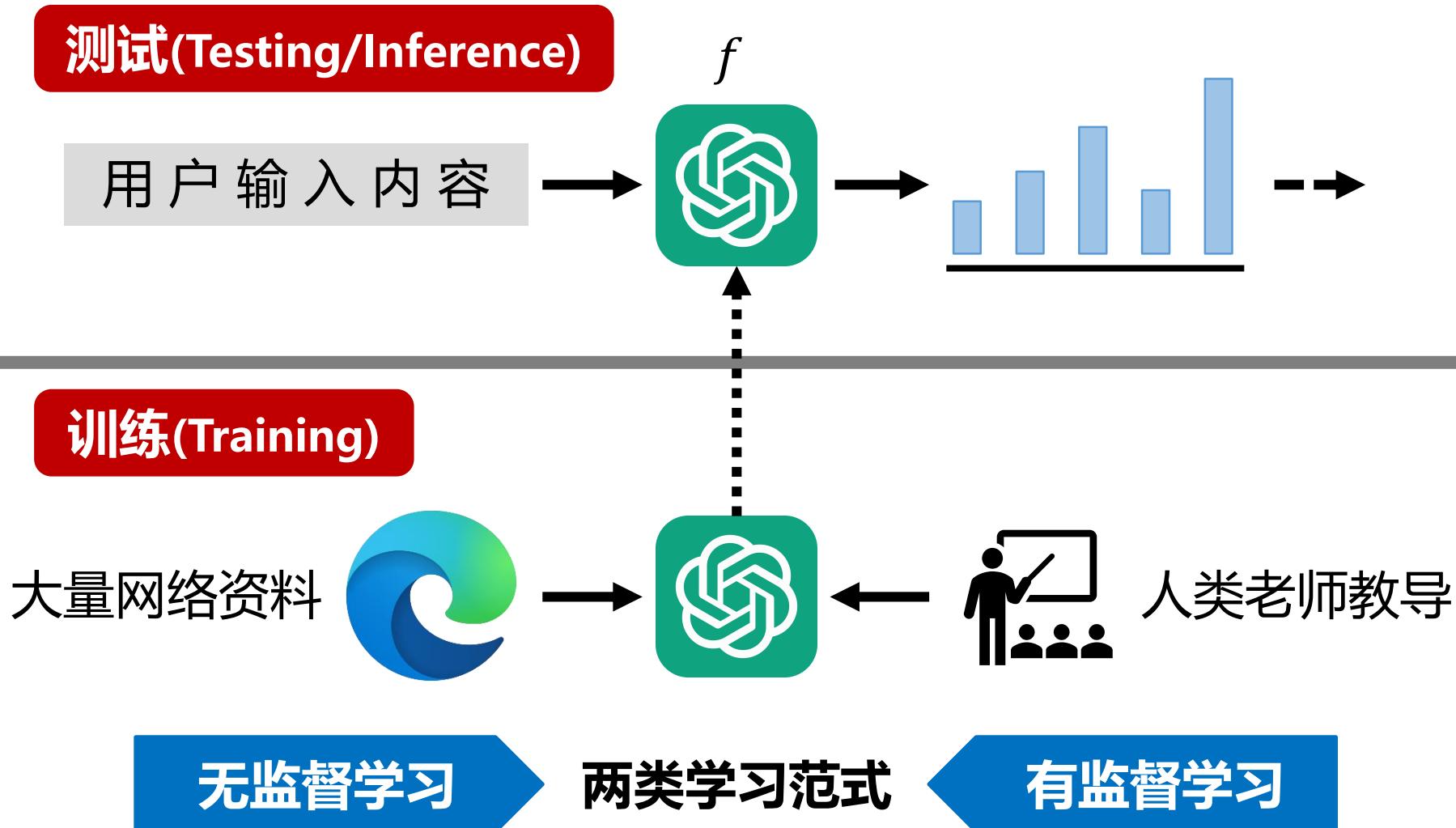


我：明白你的感受，备课确实是一项既重要又费时的任务，特别是当你希望为学生提供最好的教育体验时。但请记住，你的努力和付出都是值得的，因为你正在为学生的未来做出积极的贡献。



不过，在备课的同时，也请不要忘记照顾自己。即使是短暂的休息和放松，也会帮助你提高工作效率。你有没有考虑过定一些短暂的休息时段，或者做一些你喜欢的事情来放松一下呢？

ChatGPT是如何得到的?

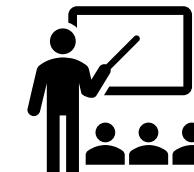


机器是如何学习的？

以英文/中文翻译为例

有监督学习

I eat an apple \longleftrightarrow 我吃苹果

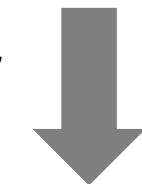


人为收集
中英对照
标签数据

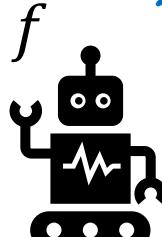
You eat an orange \longleftrightarrow 你吃橘子

⋮

机器自己寻找函数 f



I = 我, you = 你, apple
= 苹果, orange = 橘子



You eat an apple \xrightarrow{f} 你吃苹果

机器是如何学习的?

口有监督学习套用到ChatGPT上

输入：中国首都是哪? → 输出：北京

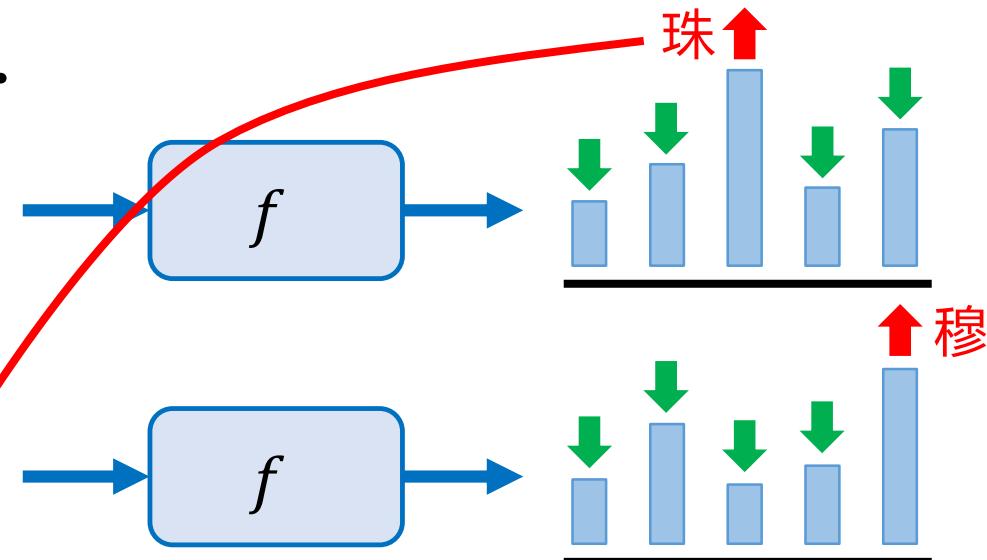


输入：帮我修改这段文字... → 输出：好的...

输入：教我做坏事... → 输出：这是不对的...

机器自己寻找函数 f 使得...

世界第一高峰是哪一座?



世界第一高峰是哪一座? 珠

机器是如何学习的?

口有监督学习套用到ChatGPT上

输入：中国首都是哪? → 输出：北京



输入：帮我修改这段文字... → 输出：好的...

输入：教我做坏事... → 输出：这是不对的...

机器自己寻找函数 f 使得...

世界第一高峰是哪一座? → f → ???

- 假设机器真的可以根据老师的教导找出函数 f , 他的能力可能还非常有限, 因为**人类可以提供的成对数据量(有标签数据量)非常有限**

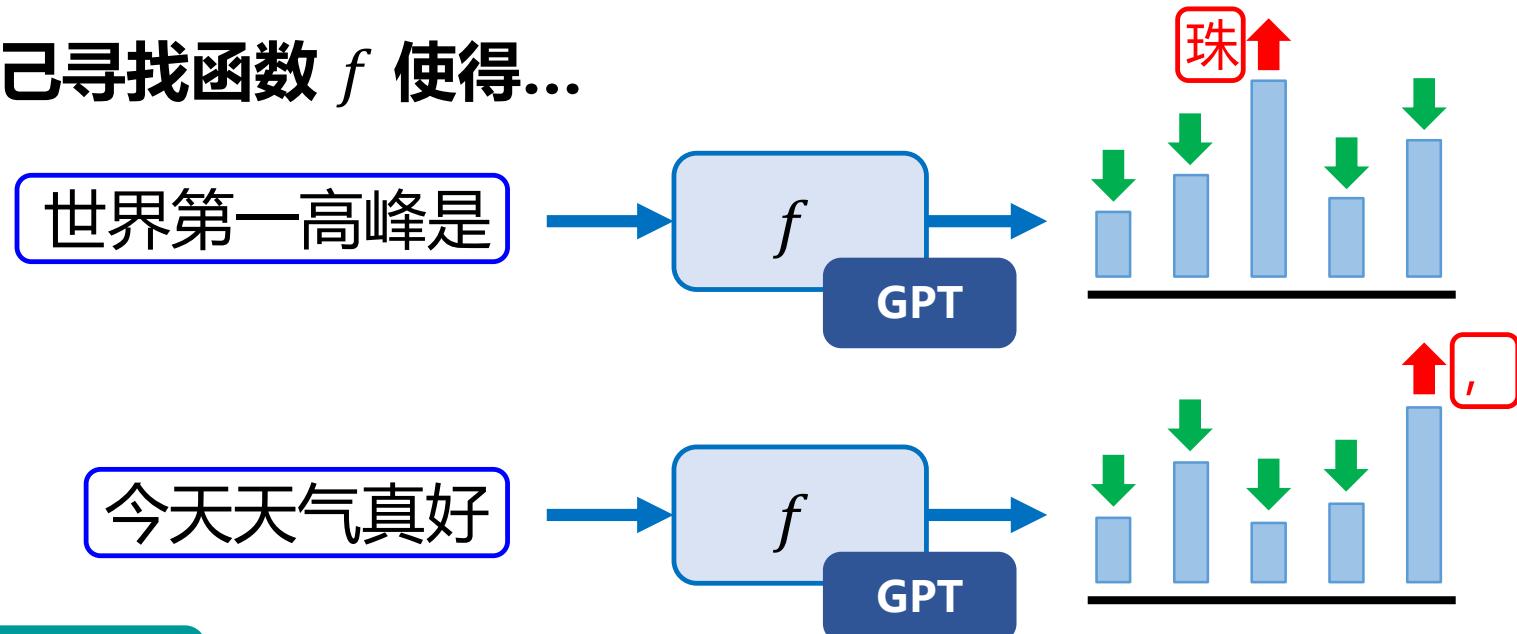
如何制造大量成对数据?

机器是如何学习的？

口 网络上每一段文字都可以教机器做文字接龙



机器自己寻找函数 f 使得...



自监督学习

ChatGPT有多大?

□ 在ChatGPT之前的GPT系列



GPT-2
(2019)

数据:
40GB

ChatGPT有多大?

口在ChatGPT之前的GPT系列

**ChatGPT
(GPT-3.5?)**

<https://platform.openai.com/docs/model-index-for-researchers>

模型: 175B

数据: 570GB
(45T before filtering)

数据:
40GB
模型:
1542M

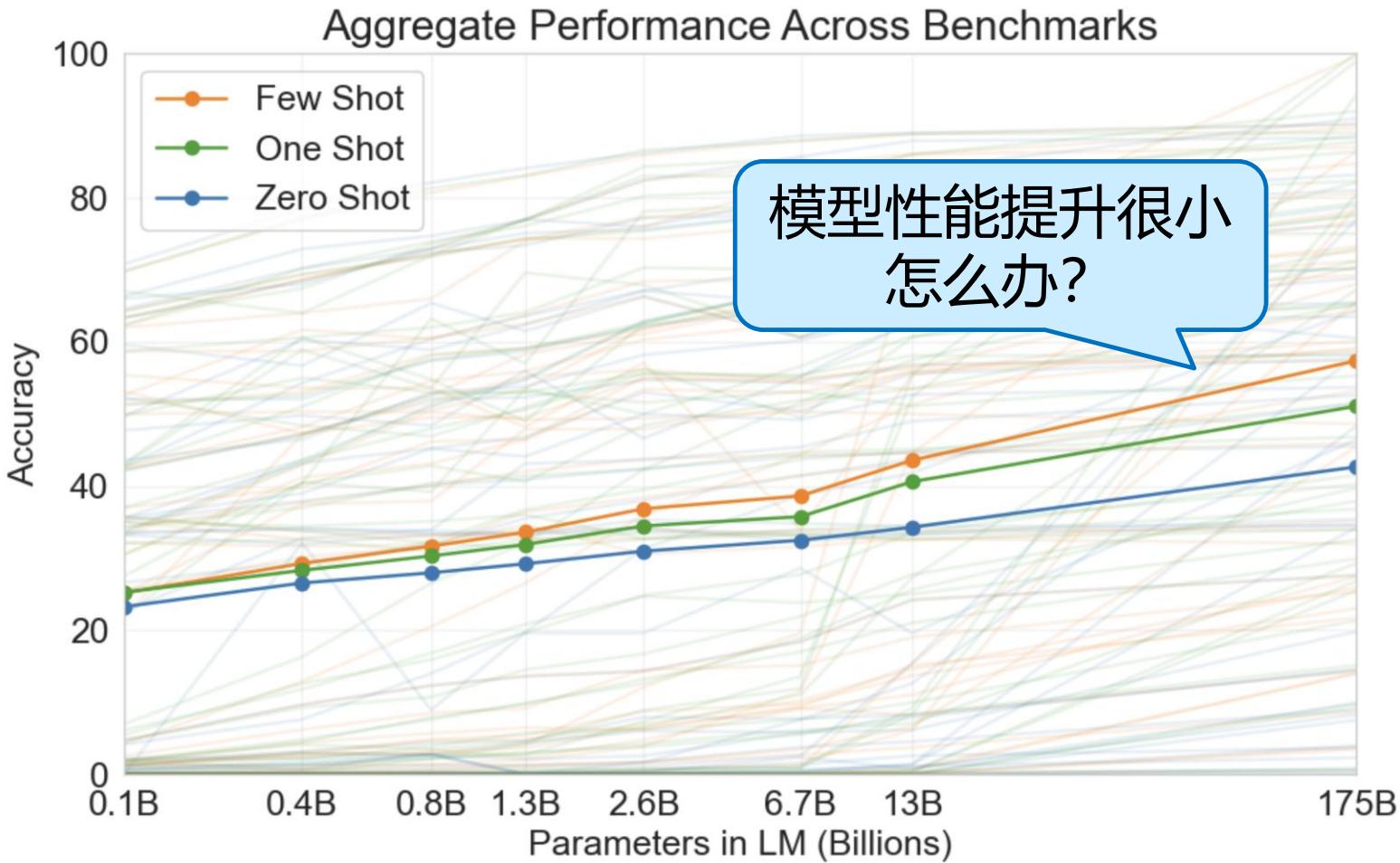


**GPT-2
(2019)**



**GPT-3
(2020)**

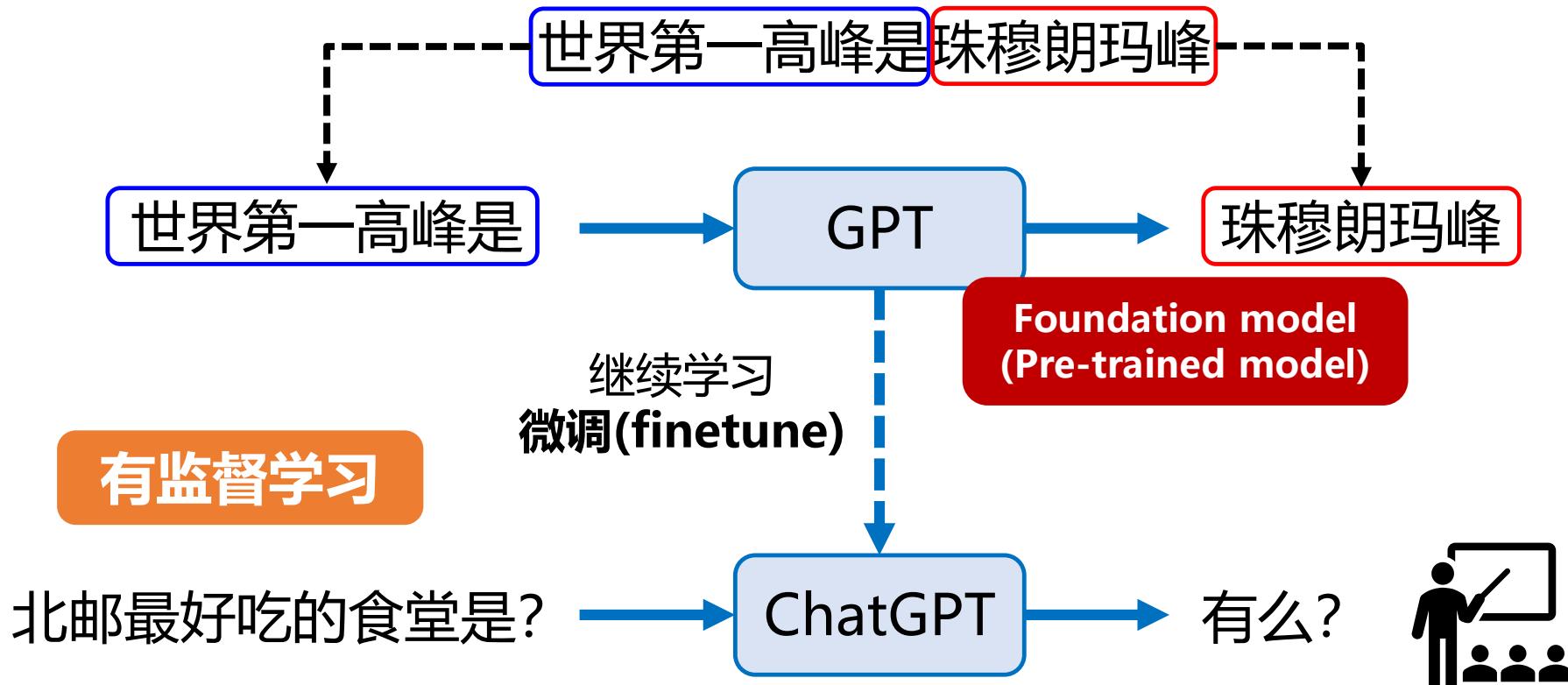
GPT-3的局限



Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901.

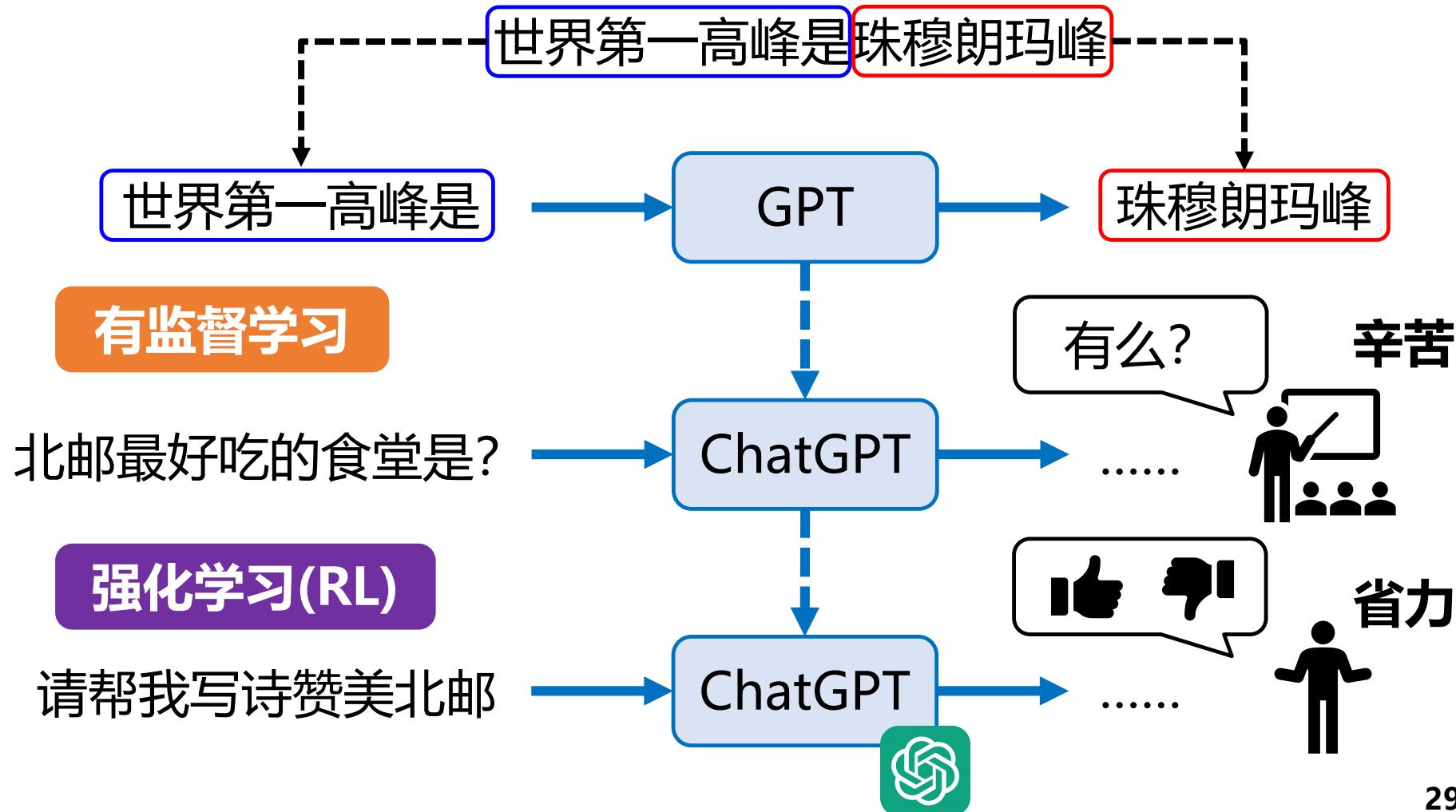
预训练技术

预训练(自监督学习)



强化学习

预训练(自监督学习)



内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



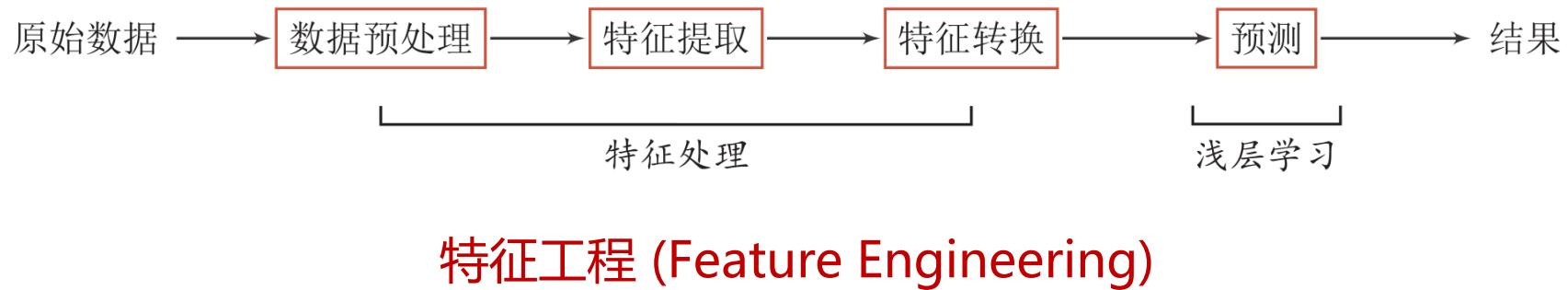
深度学习涉及的基本概念



实验平台介绍(助教)

机器学习

当我们用机器学习来解决一些模式识别任务时，一般的流程包含以下几个步骤：



浅层学习(shallow learning): 不涉及特征学习，其特征主要靠人工经验或特征转换方法来抽取。

语义鸿沟：人工智能的挑战之一

底层特征 VS 高层语义

- 人们对文本、图像的理解无法从字符串或者图像的底层特征直接获得



床前明月光，
疑是地上霜。
举头望明月，
低头思故乡。

表示学习(Representation Learning)

口数据表示是机器学习的核心问题

- 特征工程：需要借助人类智能

口表示学习

- 如何自动从数据中学习好的表示

口难点

- 没有明确的目标

什么是一个好的数据表示?

- “好的表示” 是一个非常主观的概念，没有一个明确的标准。
- 但一般而言，一个好的表示具有以下几个优点：
 - 应该具有很强的表示能力(紧致特点)
 - 应该使后续的学习任务变得简单
 - 应该具有一般性，是任务或领域独立的

语义表示

如何在计算机中表示语义？

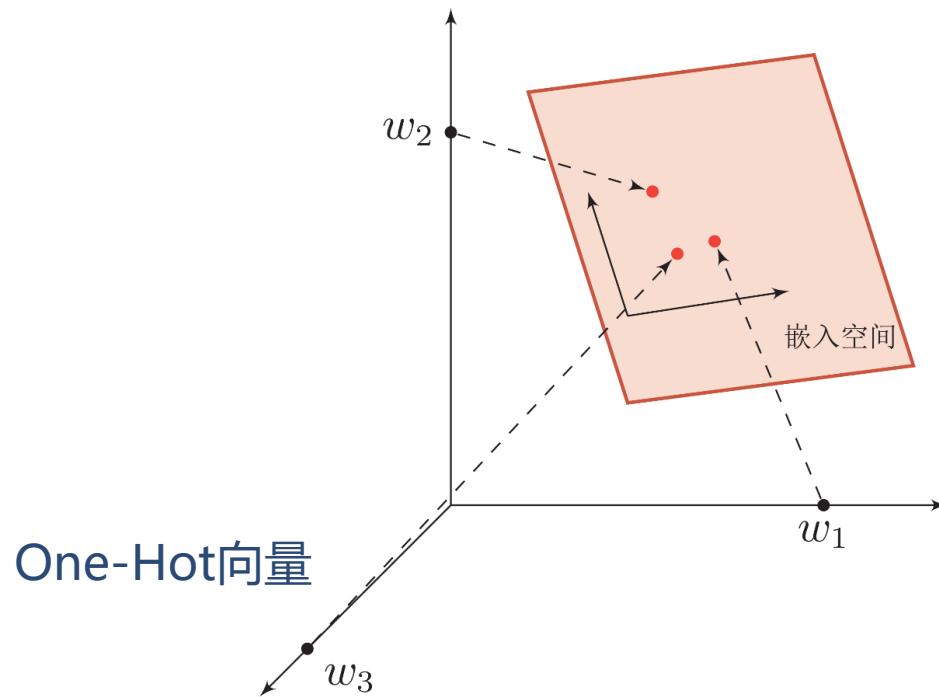
局部（符号）表示

分布式表示

知识库、规则

嵌入：压缩、低维、稠密向量

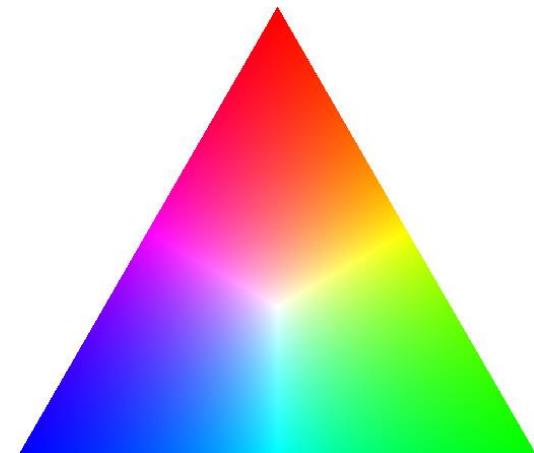
Embedding



语义表示

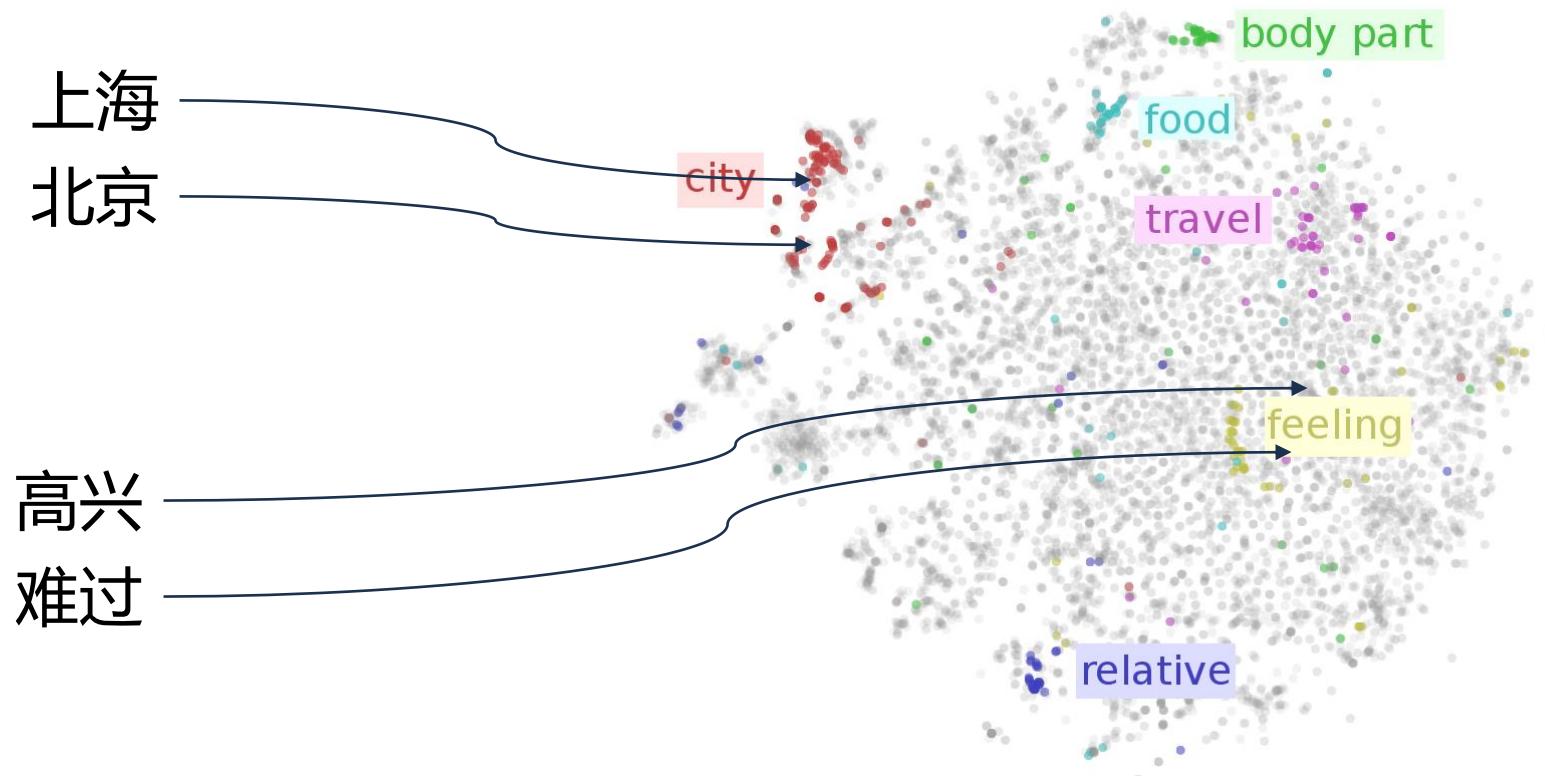
一个生活中的例子：颜色RGB编码

颜色	局部表示	分布式表示
琥珀色	$[1, 0, 0, 0]^T$	$[1.00, 0.75, 0.00]^T$
天蓝色	$[0, 1, 0, 0]^T$	$[0.00, 0.5, 1.00]^T$
中国红	$[0, 0, 1, 0]^T$	$[0.67, 0.22, 0.12]^T$
咖啡色	$[0, 0, 0, 1]^T$	$[0.44, 0.31, 0.22]^T$



语义表示

词语嵌入 (Word Embeddings)

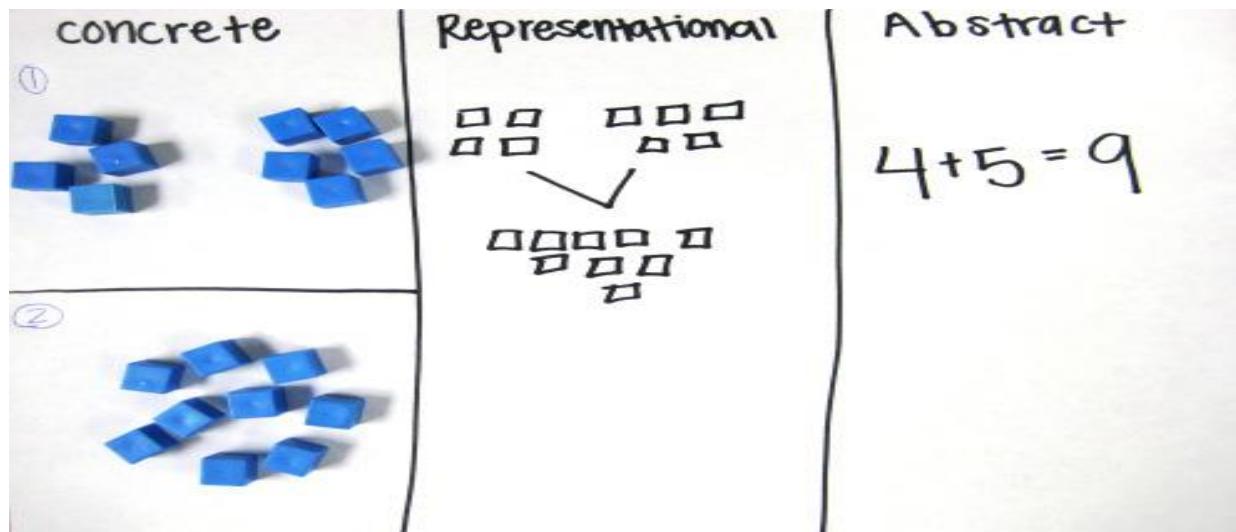


<https://indico.io/blog/visualizing-with-t-sne/>

表示学习与深度学习

一个好的表示学习策略必须具备一定的深度

- 特征重用
 - 指数级的表示能力
- 抽象表示与不变性
 - 抽象表示需要多步的构造



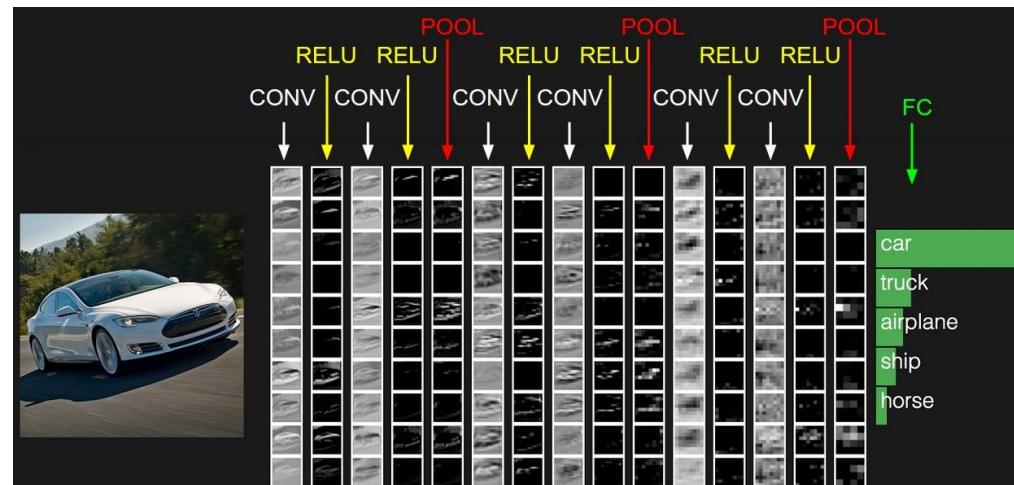
<https://mathteachingstrategies.wordpress.com/2008/11/24/concrete-and-abstract-representations-using-mathematical-tools/>

深度学习

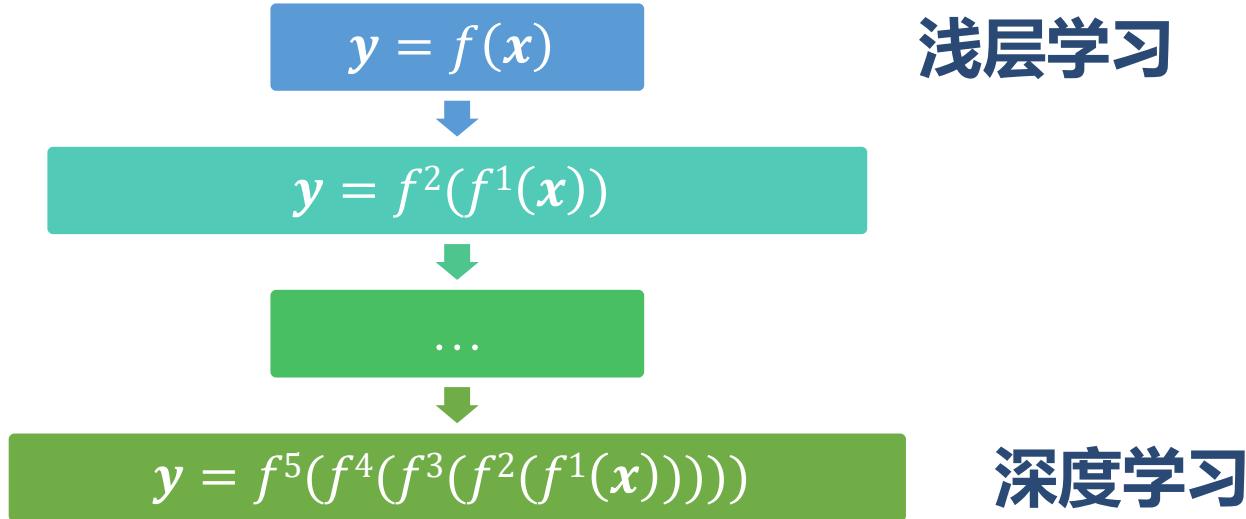
通过构建具有一定“深度”的模型，可以让模型来自动生成好的特征表示（从底层特征，到中层特征，再到高层特征），从而最终提升预测或识别的准确性。

原始数据 → 底层特征 → 中层特征 → 高层特征 → 预测 → 结果

表示学习
深度学习



深度学习的数学描述



$f^i(x)$ 为非线性函数，不一定连续

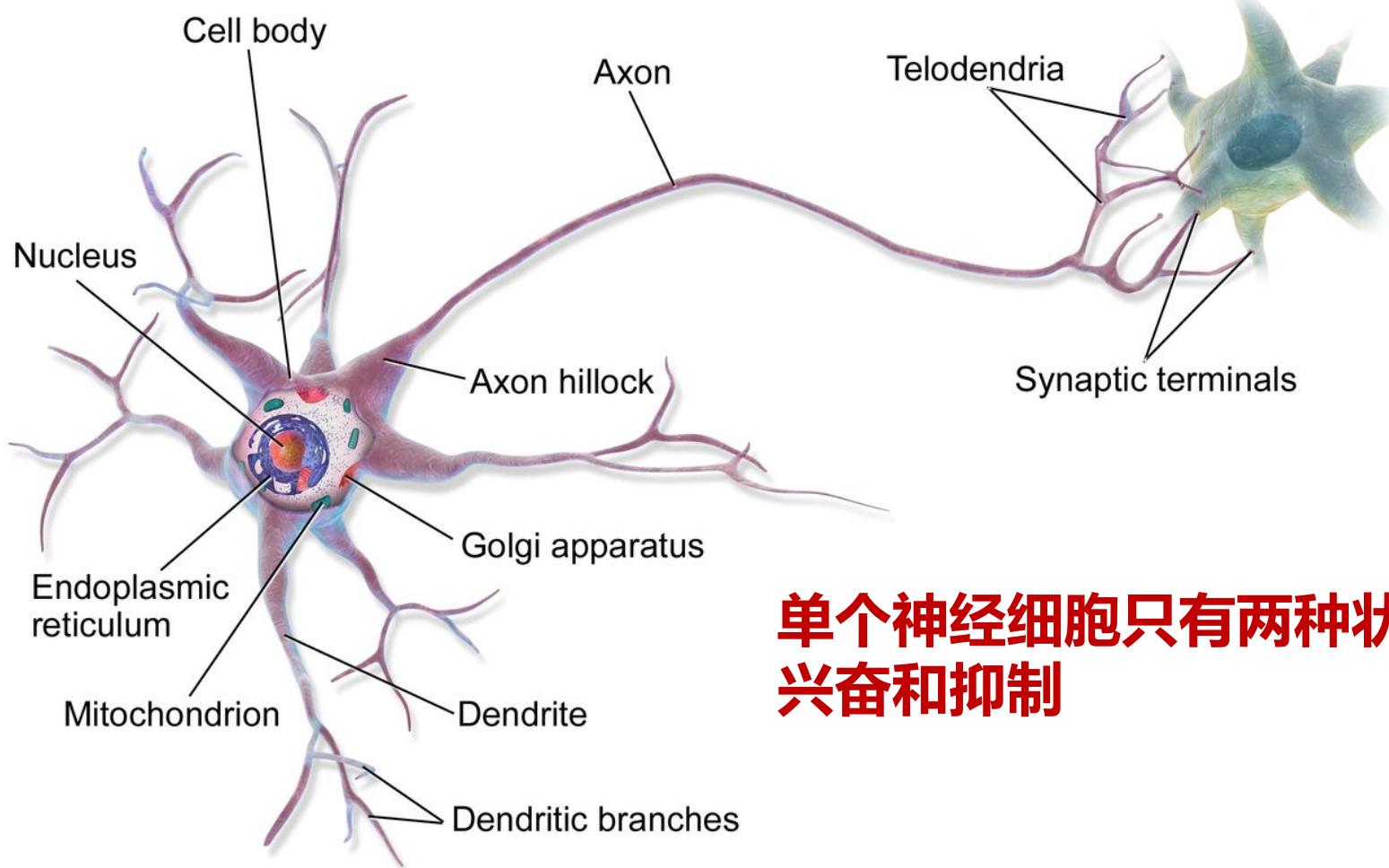
当 $f^i(x)$ 连续时，这个复合函数称为**神经网络**

神经网络

口生物神经元

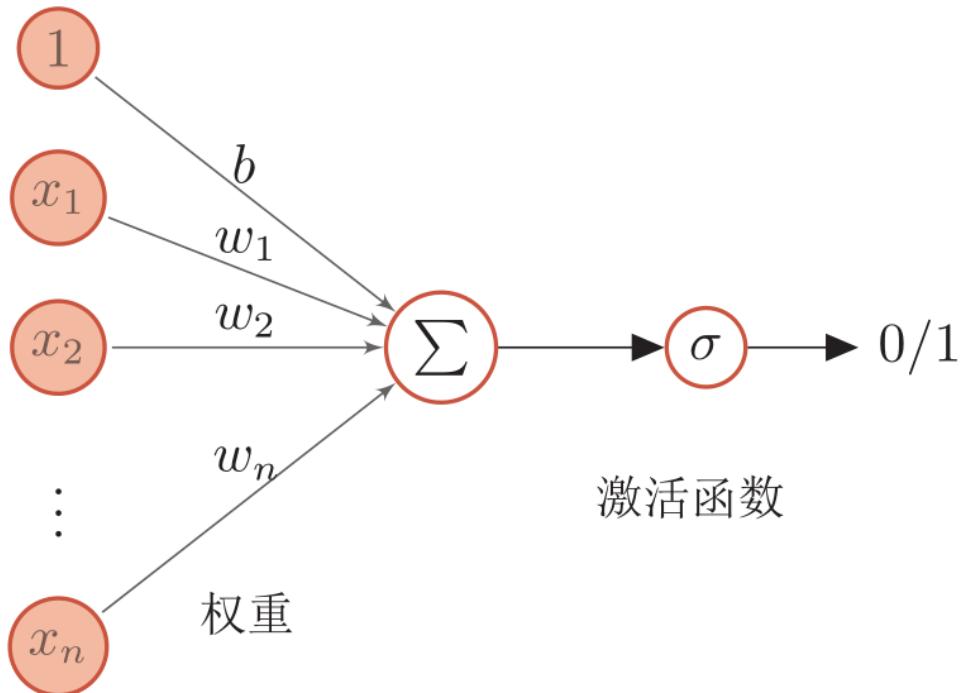
人脑有860亿个神经元

[video: structure of brain](#)



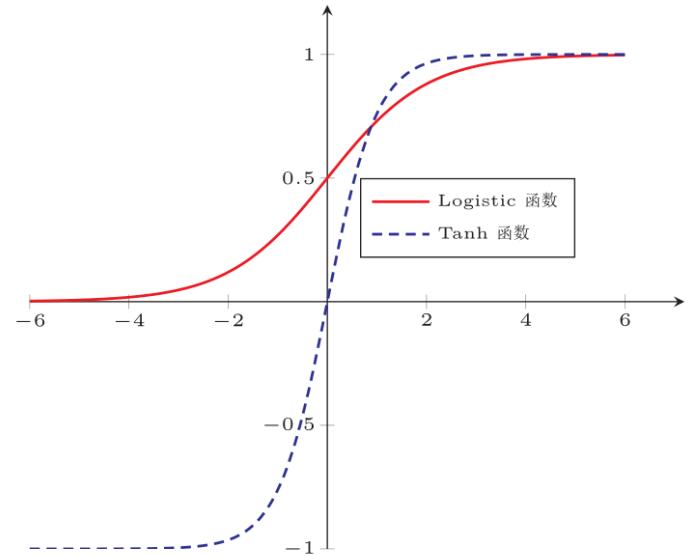
**单个神经细胞只有两种状态：
兴奋和抑制**

人工神经元



激活函数

权重



人工神经网络

口人工神经网络主要由大量的神经元以及它们之间的有向连接构成。因此考虑三方面：

口神经元的激活规则

- 主要是指神经元输入到输出之间的映射关系，一般为非线性函数

口网络的拓扑结构

- 不同神经元之间的连接关系

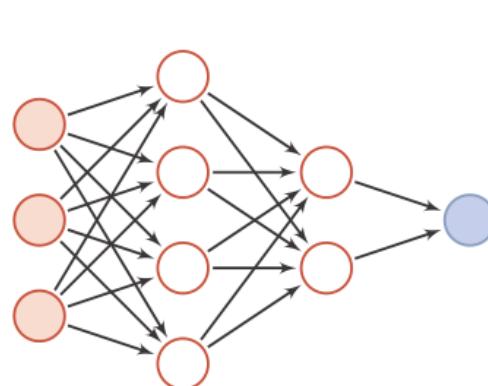
口学习算法

- 通过训练数据来学习神经网络的参数

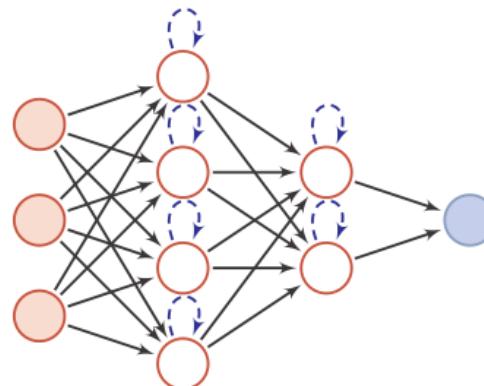
人工神经网络

口人工神经网络由神经元模型构成，这种由许多神经元组成的信息处理网络具有并行分布结构

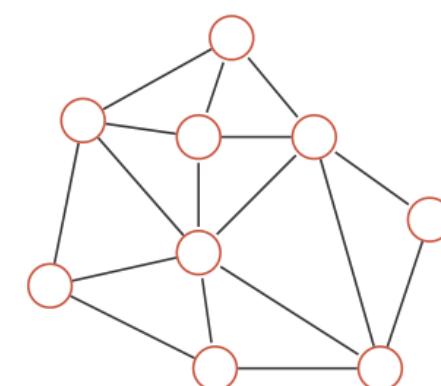
- 虽然这里将 神经网络结构大体上分为三种类型，但是大多数网络都是复合型结构，即一个神经网络中包括多种网络结构。



(a) 前馈网络

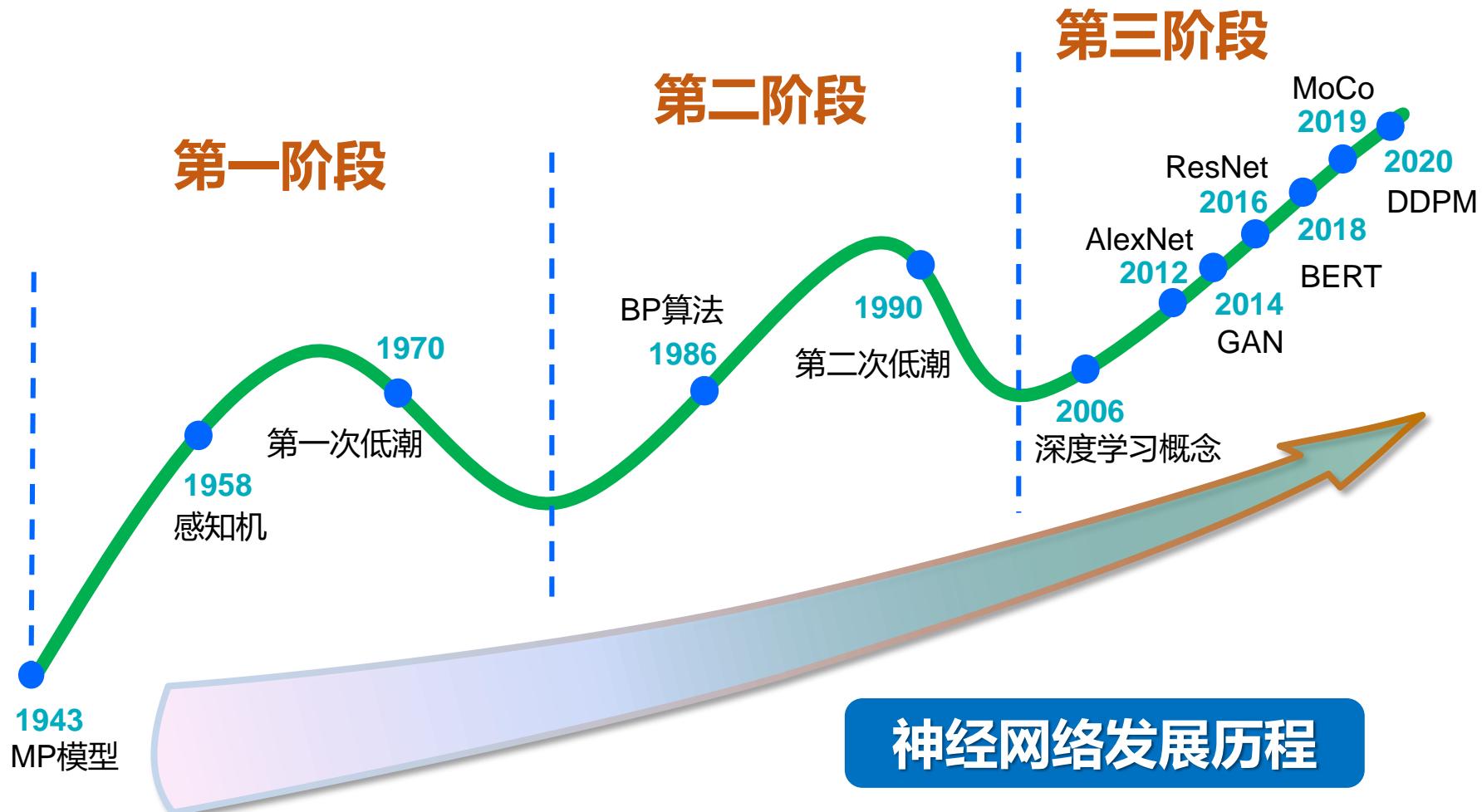


(b) 记忆网络



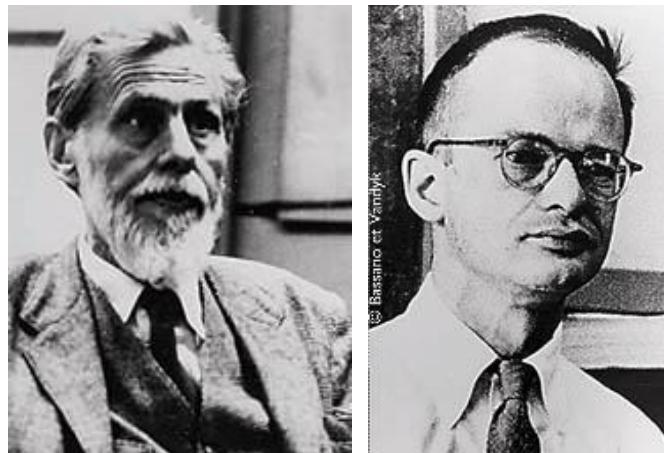
(c) 图网络

深度学习发展历程



深度学习发展第一阶段

- 1943 年，神经生物学家 Warren McCulloch 和数学家 Walter Pitts 对生物神经元进行建模，首次提出一种神经元模型，即广为认知的 M-P 模型
- 1958年，Roseblatt 最早提出可以模拟人类感知能力的神经网络模型，并称之为感知机
- 1970年，计算能力无法满足复杂任务，AI 第一次低潮到来



Warren McCulloch Walter Pitts

深度学习发展第二阶段

- 1986年，Geoffrey Hinton 等人将反向传播算法引入到多层感知机，完美的解决了非线性分类问题
- 1990年，日本“第五代工程”宣告失败，迎来 AI 第二次低潮
- 20世纪 90 年代中期，统计学习理论和以支持向量机为代表的机器学习模型开始兴起，神经网络由于理论基础不清晰、优化困难、可解释性差等问题陷入低潮



Geoffrey Hinton

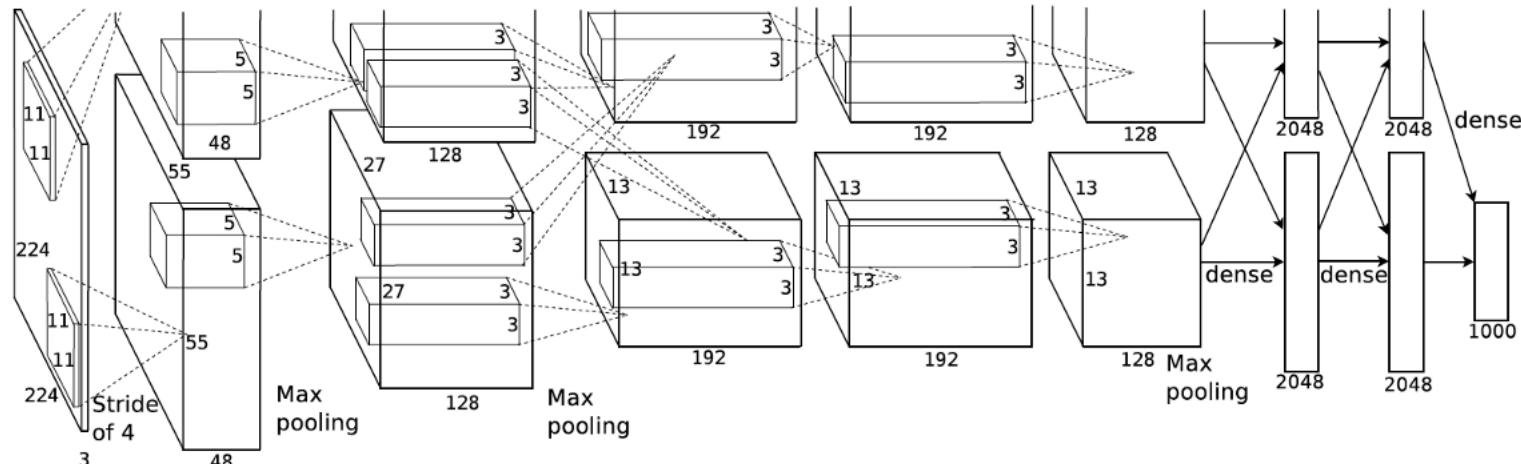
深度学习发展第三阶段

- 2006 年， Geoffrey Hinton 及其团队在 Science 上发表了关于神经网络理念突破性的文章，首次提出了深度学习的概念，被称为“深度学习觉醒”的一年



深度学习发展第三阶段

- 2012年，Hinton 领导的小组采用深度学习模型 AlexNet，在 ImageNet 数据大赛一举夺冠，碾压第二名 SVM 算法的效果
- AlexNet 是第一个现代深度卷积网络模型，是深度学习技术在图像分类上取得真正突破的开端



AlexNet

Refer: Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems. 2012: 1097-1105.

内容导览



课程教学内容概览



课程考核方式



深度学习关键技术：从ChatGPT谈起



深度学习涉及的基本概念



实验平台介绍(助教)

实验平台

选择一：天池 Notebook

天池notebook以机器学习PAI DSW(DataScienceWorkshop)作为天池实验室底座，为大家提供完备的IDE以及丰富的计算资源，单次使用时长8小时

配置：Nvidia A10

使用：总共使用时间不限制CPU和GPU单次链接8小时，每次链接可以运行多个Notebook

地址：

<https://tianchi.aliyun.com/notebook-ai/>

 阿里云 | TIANCHI 天池

选择二：Kaggle Notebook

Kaggle自带的NoteBook部分，基本操作同jupyter notebook一致，运行时默认认为CPU运行，可以支持在线加载数据集

配置：Nvidia Tesla P100-PCIE-16GB 1.3285GHz

使用：CPU时间不限，GPU每周35小时可以同时运行多个CPU Notebook，只能运行一个GPU Notebook

地址：

<https://www.kaggle.com/code>



天池Notebook使用介绍

□ 注册并登录，新建Notebook

阿里云 | TIANCHI 天池

数据可视化分析学习赛

首页 天池学习 天池大赛 数据集 天池实验室 在线编程 技术圈 其他

aliyun8195960196 LV1 | 退出 中

我的剩余资源

GPU剩余用量: 59.9小时/60小时
有效期至2023-12-31 23:59:59
积分达200加赠30小时, 点此get升

CPU剩余用量: 不限

天池Notebook

天池notebook集成机器学习PAI DSW (DataScienceWorkshop) 探索者版, 成为天池实验室底座, 为大家提供完备的IDE以及丰富的计算资源

新建Notebook → 我的实验室 →

所有项目 我的Star 我的项目

搜索关键字 排序: 时间 语言: 所有语言 大赛: 所有赛事及其他

淘宝用户购物行为数据可视化分析
aliyun1234567 2023-03-02 10:27:44

☆ 73 ￥ 221 🎁 46 ⚙ 11622

免费试用

天池Notebook使用介绍

□ 新建Notebook界面如下，每个单元格可以逐个运行，方便调试

Draft 2023-09-09 11:53:37

关闭实例 发布代码 | 课程 ▾ 数据集 ▾ Notebook ▾ | GPU独享型 ▾ 07:59:37 aliyun8195960196 ▾

The screenshot shows the Tianchi Notebook interface. On the left is a file browser with a search bar. The main area contains a code editor with several cells. The top cell lists installed packages. The second cell installs a package. The third cell provides an example of using `sklearn2pmml`. The bottom cell shows a complex matplotlib plot generation script. A context menu is open over the third cell, with 'GPU独享型' (GPU Exclusive) selected. The right side of the interface includes a memory monitor showing 229 / 4096 MB, a sidebar with '免费试用' (Free Trial), '云开发' (Cloud Development), '挑战赛' (Challenge Cup), and a status bar at the bottom.

```
[ ]: # 查看当前kernel下已安装的包 List packages
!pip list --format=columns

[ ]: # 安装拓展包
pip install some_package --user

[ ]: # 举例：安装sklearn某依赖库
!pip install sklearn2pmml --user

[ ]: # 举例：绘图案例 an example of matplotlib
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
from scipy.special import jn
from IPython.display import display, clear_output
import time
x = np.linspace(0,5)
f, ax = plt.subplots()
ax.set_title("Bessel functions")

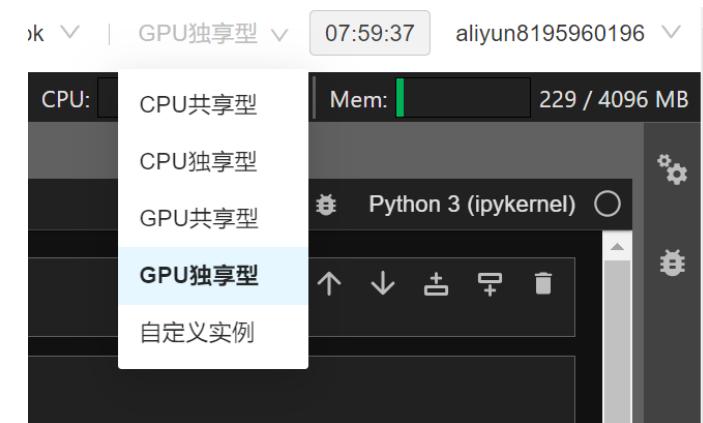
for n in range(1,10):
    time.sleep(1)
    ax.plot(x, jn(x,n))
    clear_output(wait=True)
    display(f)

# close the figure at the end, so we don't get a duplicate
# of the last plot
plt.close()
```

天池Notebook使用介绍

□ 各种服务器类型

- CPU共享型： 2C4G 存在多个人共用一个容器的情况，运行环境通过kernel隔离，notebook文件编辑直接保存到天池实验室,防止多人同时在容器编辑相互影响，该方式下由于多个人共用一个容器，因此请注意隐私
- CPU独享型： 2C4G 每个人单独一个容器，安全性相比共享型更高
- GPU共享型： 2C6G T4(显存2G) 与CPU共享型使用方式相同，使用时请注意隐私
- GPU独享型： 2C6G T4(显存2G) 每个人单独一个容器 与CPU独享型使用方式相同
- 自定义类型：由于天池实验室的资源有限，如有更高资源需求可通过领取PAI-DSW免费试用方式启动PAI-DSW，关联到实验室使用。领取及使用方法参见：<https://tianchi.aliyun.com/notebook/511513>



天池Notebook使用介绍

以【Python入门系列】用Pandas揭秘美国选民的总统喜好赛事baseline教程为例

Fork 教程内容到自己的天池实验室：进入到教程页面，点击右上角的Fork按钮，即可将教程内容复制到自己的天池实验室。

The screenshot shows a page from the Tianchi Notebook section. At the top, there's a navigation bar with links for Home, Tianchi Learning, Tianchi Competition, Tianchi Laboratory, Online Programming (NEW), Tianchi No. 7 Hall, Technology Circle, and Others. On the right side of the header, there are user icons for 193 notifications, a profile picture, Tianchi Little T, and LVS, along with a Logout link.

The main content area has a breadcrumb trail: Tianchi Laboratory > Tianchi notebook > 正文. Below it is a title: "数据分析入门活动赛事，利用Pandas分析美国选民总统喜好..." and a timestamp: 2020-10-22 21:57:33. There are also stats: 5808 views, 588 stars, and 2798 forks. A large red arrow points to the "Fork" button, which is highlighted with a red border.

The page features a tab navigation with "内容" (Content) selected, followed by "评论(3)" (Comments 3), "Star(588)", and "fork记录(2798)". To the right, there's a dropdown for "版本 12".

The main content starts with a section titled "1、赛前准备" (1. Preparations). It includes a sub-section "1.1 前言" (1.1 Preface) and a detailed description of the competition: "本次赛事由开源学习组织Datawhale主办，主要带领学习者利用Python进行数据分析以及数据可视化，包含数据集的处理、数据探索与清晰、数据分析、数据可视化四部分，利用pandas、matplotlib、wordcloud等第三方库带大家玩转数据分析~还有丰富礼品等你来领取哦~". Below this is a link: "学习赛事地址: <https://tianchi.aliyun.com/competition/entrance/531837/introduction>".

On the right side, there's a "目录" (Table of Contents) sidebar with a hierarchical list:

- 1、赛前准备
 - 1.1 前言
 - 1.2 数据集来源介绍
 - 1.3 需要提前安装的包
 - 1.4 需要提前下载好数据集
- 2、数据处理
 - 2.1 将委员会和候选人一一对应...
 - 2.2 将候选人和捐赠人一一对应...

天池Notebook使用介绍

□ 进入DSW进行代码编辑、运行、学习

The screenshot shows the Tianchi Notebook interface. At the top, there is a navigation bar with links: 阿里云 | TIANCHI 天池, 首页, 天池学习, 天池大赛, 数据集, 天池实验室 (highlighted with a red box), 在线编程, 技术圈, 其他, and a search bar. To the right of the search bar is a user profile icon with a '328' badge and a 'LV4' level indicator. Below the navigation bar, there is a large orange button labeled '我的剩余资源'. On the left, there is a section titled '天池Notebook' with a sub-section titled '我的实验室' (highlighted with a red box). In the center, there is a resource summary: GPU剩余用量: -4.6小时/60小时 (Valid until 2023-12-31 23:59:59) and CPU剩余用量: 不限. The background features a stylized illustration of a modern city skyline.

天池Notebook使用介绍

□ 进入“我的实验室”后，点击“编辑”按钮进入代码编辑环境

The screenshot shows the Tianchi Notebook interface. At the top, there is a navigation bar with links for 阿里云 | TIANCHI 天池, 首页, 天池学习, 天池大赛, 数据集, 天池实验室, 在线编程, 技术圈, 其他, and a search bar. On the right side of the header, there is a user profile with 328 contributions, the username itranner33, a level 4 badge (LV4), and a log out link. Below the header, the main content area displays a notebook titled "fork from 数据分析入门活动赛事, 利用Pandas分析美国...". The notebook details include: forked from "数据分析入门活动赛事, 利用Pandas分析美国...", created by itranner33 on 2023-08-30, tagged with PYTHON3, and part of the "【Python入门系列...】" series. There are 0 comments, 0 stars, and options to attach data, edit/run code, or publish. Below the title, there are tabs for 内容 (Content), 已挂载数据 (Attached Data), 评论(0) (Comments 0), Star(0) (Stars 0), and fork记录(0) (Fork History 0). The content section starts with a heading 1. 赛前准备 (Preparation) and a sub-section 1.1 前言 (Foreword). It describes the competition as being organized by Datawhale, involving Python data analysis and visualization, and includes sections on data processing, data exploration, and data analysis. A link to the competition introduction is provided: <https://tianchi.aliyun.com/competition/entrance/531837/introduction>. On the right side of the content area, there is a sidebar with a "版本 0" (Version 0) label and a "目录" (Table of Contents) section. The table of contents includes: 1、赛前准备 (1. Preparation), 1.1 前言 (1.1 Foreword), 1.2 数据集来源介绍 (1.2 Introduction to Dataset Sources), 1.3 需要提前安装的包 (1.3 Packages to be installed in advance), 1.4 需要提前下载好数据集 (1.4 Pre-download data sets), 2、数据处理 (2. Data Processing), 2.1 将委员会和候选人一一对应... (2.1 Match committees and candidates one-to-one...), 2.2 将候选人和捐赠人一一对应... (2.2 Match candidates and donors one-to-one...), 3、数据探索与清洗 (3. Data Exploration and Cleaning), and 4、数据分析 (4. Data Analysis). At the bottom left, there are buttons for "新建" (New) and "上传" (Upload). A red box highlights the "编辑" (Edit) button in the toolbar above the content area.

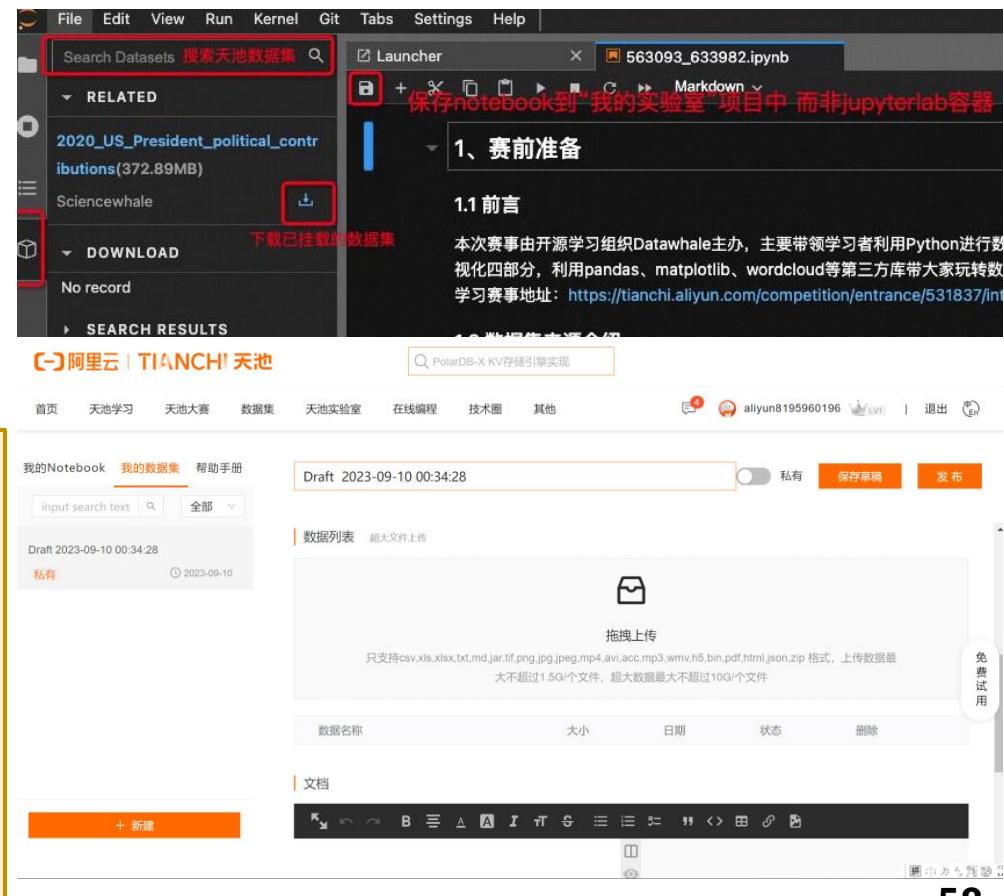
天池Notebook使用介绍

□ 下载并解压对应数据集

- 使用天池已经开放或自己在实验室上传的数据集，可通过jupyterlab页面左侧的数据集插件进行搜索和下载
- 自己上传数据集的方式
如下右图
- 使用外部其他数据，可直接通过wget方式下载
数据到容器中使用

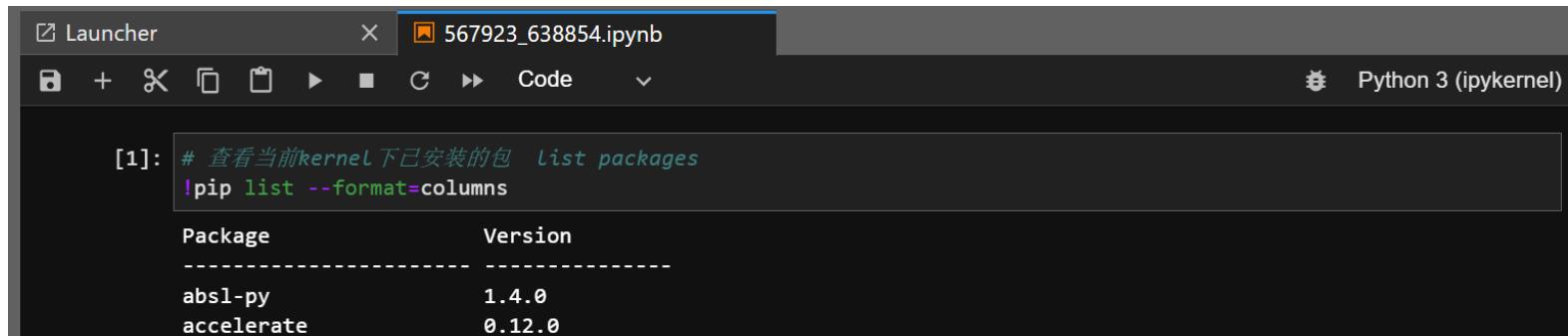
```
# 使用wget命令从网络下载  
一个文件并保存在当前目录  
wget http://cn.wordpress.or  
g/wordpress-3.1-zh_CN.zip
```

```
# 解压方法，（压缩包名字  
需要带后缀，如：data.zip）  
!unzip -o -q 压缩包名字
```



天池Notebook使用介绍

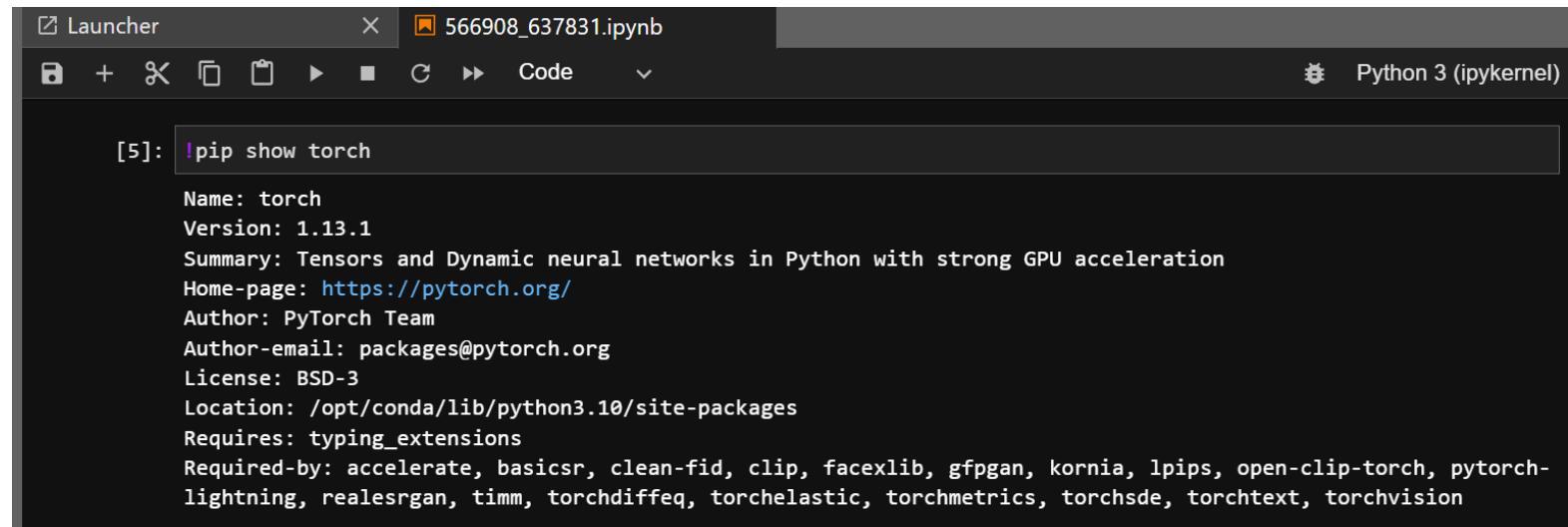
□ pip list可以查看所安装的包



```
[1]: # 查看当前kernel下已安装的包 List packages
!pip list --format=columns

Package           Version
-----
absl-py          1.4.0
accelerate       0.12.0
```

□ pip show xxx可以查看包的版本信息



```
[5]: !pip show torch

Name: torch
Version: 1.13.1
Summary: Tensors and Dynamic neural networks in Python with strong GPU acceleration
Home-page: https://pytorch.org/
Author: PyTorch Team
Author-email: packages@pytorch.org
License: BSD-3
Location: /opt/conda/lib/python3.10/site-packages
Requires: typing_extensions
Required-by: accelerate, basicsr, clean-fid, clip, faceexlib, gfpgan, kornia, lpips, open-clip-torch, pytorch-lightning, realesrgan, timm, torchediffeq, torchelastic, torchmetrics, torchsde, torchtext, torchvision
```

天池Notebook使用介绍

- nvidia-smi可以查看GPU信息

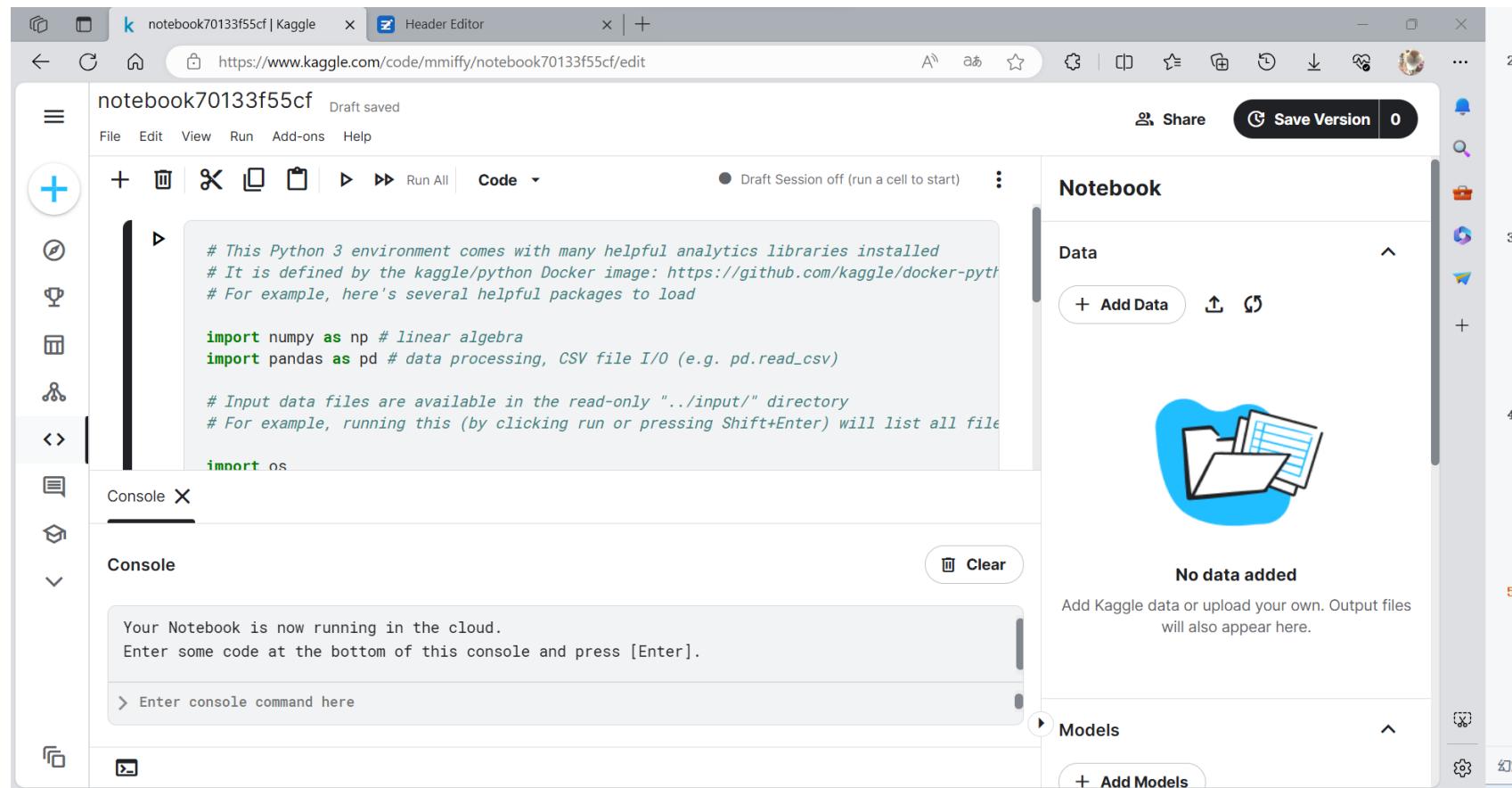
```
[9]: !nvidia-smi
```

```
Sat Sep  9 17:14:48 2023
+
| NVIDIA-SMI 470.103.01    Driver Version: 470.103.01    CUDA Version: 11.4    |
+-----+-----+-----+-----+-----+-----+-----+-----+
| GPU  Name      Persistence-M| Bus-Id      Disp.A  | Volatile Uncorr. ECC | | | |
| Fan  Temp     Perf  Pwr:Usage/Cap| Memory-Usage | GPU-Util  Compute M.  |
|          |          |             |              |          |          MIG M. |
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0  NVIDIA A10        On  | 00000000:00:08.0 Off |          0 | | | |
| 0%   31C     P8    20W / 150W |      0MiB / 22731MiB |     0%       Default |
|          |          |             |              |          |          N/A |
+-----+-----+-----+-----+-----+-----+-----+-----+
+
| Processes:
| GPU  GI  CI      PID  Type  Process name          GPU Memory |
| ID   ID              ID           Usage            |
+-----+-----+-----+-----+-----+-----+
| No running processes found
+-----+-----+-----+-----+-----+-----+
```

Kaggle Notebook使用介绍

□ 注册并登录

- 不翻墙的情况下注册可能会遇到人机验证无法进行的问题，解决方法可以参考[Kaggle注册过程中的人机验证问题解决方案](#)



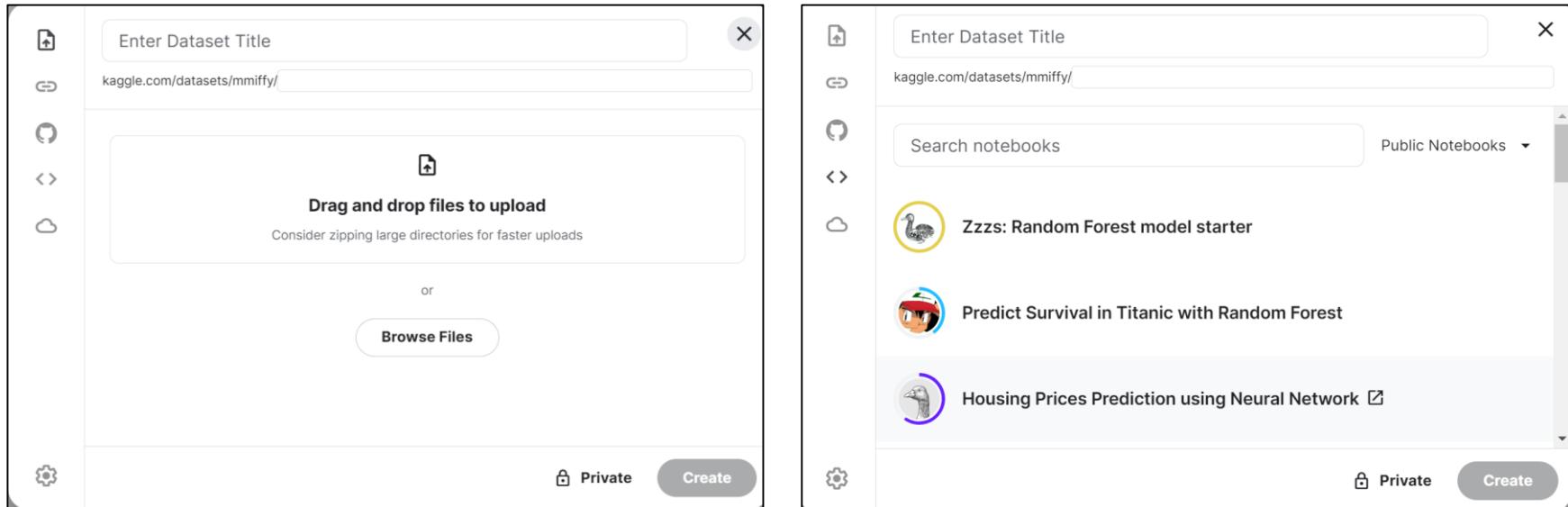
Kaggle Notebook使用介绍

□ 上传自己的数据集或者搜索kaggle中已有的数据集

The screenshot shows the Kaggle Notebook interface. On the left, a sidebar menu includes options like Create, Home, Competitions, Datasets (which is highlighted with a red box), Models, Code, Discussions, Learn, More, Your Work, and EDITED. Below these are two entries: notebook0512f5b78e. The main content area is titled "Datasets" and features a search bar, a "New Dataset" button, and a "Your Work" section. It also includes a search for datasets, filters, and categories such as All datasets, Computer Science, Education, Classification, Computer Vision, NLP, and Data Visualization. A "Trending Datasets" section displays four thumbnail images: basketball players, a pizza, a person working at a desk, and a landscape.

Kaggle Notebook使用介绍

- 上单次只能上传一个文件，因此最好将数据集打成压缩包进行上传



以Kaggle中的Housing Prices Prediction using Neural Network项目为例

Kaggle Notebook使用介绍

- 进入notebook后，可以看到右边Data下有input和output的相关信息

The screenshot shows a Kaggle Notebook titled "Housing Prices Prediction using Neural Netw...". The notebook interface includes a sidebar with various icons, a toolbar with cell creation and execution buttons, and a main area where Python code is written in a code cell. The code imports several libraries like numpy, pandas, matplotlib, seaborn, sklearn, and tensorflow. To the right of the notebook area, there is a sidebar titled "Notebook" which contains a "Data" section. This "Data" section is highlighted with a red box and contains two subsections: "Input" and "Output". The "Input" section lists several files: "home-data-for-ml-course" (with subfiles "data_description.txt", "sample_submission.csv", "sample_submission.csv.gz", "test.csv", "test.csv.gz", "train.csv", and "train.csv.gz"). The "Output" section shows a folder named "/kaggle/working" with a "load more" button. At the bottom of the sidebar, there is a "Models" section.

```
import numpy as np
import pandas as pd

pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import matplotlib as mpl

from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer

from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

from sklearn.model_selection import train_test_split

import tensorflow as tf
```

Kaggle Notebook使用介绍

□ 运行单元格可以逐步得到运行结果，但需要按顺序运行

Housing Prices Prediction using Neural Netw...

The screenshot shows a Kaggle Notebook interface for "Housing Prices Prediction using Neural Netw...". The top navigation bar includes File, Edit, View, Run, Add-ons, Help, Share, Save Version (0), and session status (Draft Session (1m)). The toolbar features icons for New, Open, Save, Run Cell, Run All, and Code dropdown.

The main area displays a code cell with a red box around the play button icon:

```
train_data = pd.read_csv("/kaggle/input/home-data-for-ml-course/train.csv")
train_data.head()
```

Output of the cell shows the first 5 rows of the dataset:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	Pool	
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Reg	Lvl	AllPub	...	0	Ni
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Reg	Lvl	AllPub	...	0	Ni
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Reg	Lvl	AllPub	...	0	Ni
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Reg	Lvl	AllPub	...	0	Ni
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Reg	Lvl	AllPub	...	0	Ni

Text below the table: 5 rows x 81 columns

Bottom of the cell: + Code, + Markdown

Another code cell below:

```
test_data = pd.read_csv("/kaggle/input/home-data-for-ml-course/test.csv")
test_data.head()
```

The right sidebar contains sections: Notebook, Data (+ Add Data), Input (home-data-for-ml-course), Output (56KB / 19.5GB), Models, Submit to competition, Notebook options, and Schedule a notebook to run.

Kaggle Notebook使用介绍

□ 点击Draft Session可以看到session运行时的各项指标

The screenshot shows a Kaggle Notebook interface for a "Housing Prices Prediction using Neural Netw..." project. The top navigation bar includes File, Edit, View, Run, Add-ons, Help, Share, and Save Version (0). A red box highlights the "Draft Session (6m)" entry in the session list, which is connected by a red arrow to a detailed modal window. The modal displays session statistics: GPU P100 On, Session 6m (12 hours), Disk 4GB Max 73.1GB, CPU usage 0.00%, and RAM 705.7MB Max 30GB. The main notebook area shows code for loading data and a preview of the "train_data" DataFrame.

Housing Prices Prediction using Neural Netw...

File Edit View Run Add-ons Help

Share Save Version 0

Draft saved

+ Markdown

Draft Session (6m)

Session 6m (12 hours)

Disk 4GB Max 73.1GB

CPU 0.00%

RAM 705.7MB Max 30GB

Loading Data

[4]:

```
train_data = pd.read_csv("/kaggle/input/home-data-for-ml-course")
train_data.head()
```

[4]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour
0	1	60	RL	65.0	8450	Pave	NaN	Reg	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	

5 rows x 81 columns

Notebook

Data

+ Add Data

Input

- home-data-for-ml-course

Output (56KB / 19.5GB)

- /kaggle/working

Models

Submit to competition

Notebook options

LANGUAGE

Python