Lecture 1
Introduction

Lin ZHANG, PhD
School of Software Engineering
Tongji University
Fall 2023

Course Info

Contact Information

Room 418L, Jishi Building
TA: Tianjun Zhang, 1911036@tongji.edu.cn

All material can be found at

http://cslinzhang.gitee.io/home/

Materials

- Major materials
  - My slides
- References
  - 《计算机视觉：原理算法与实践》（草稿），张林等
  - 《机器学习》，周志华，2016
  - 《统计学习方法》（第2版），李航，2019
  - Some papers

Examination

- Homework 30%: 3 times, and each time 10%.
- Paper reading and presentation 20%
  - Read a paper related to machine learning and do a presentation
- Final report and presentation 50%
  - Select a problem related to your research direction, try to solve it with machine learning techniques, write an essay and finally do a presentation
- Being absent >=1/3 lectures, you will fail this course

Arrangement of Lectures (temporarily)

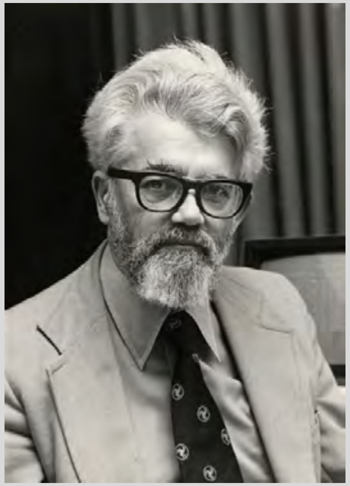
- Basic Concepts and Model Evaluation
- AdaBoost and Cascade Structure
- Principle Component Analysis
- Sparse Representation based Classification
- Linear Model
- Neural Network and CNN
- Applications of CNN
- Least Squares
- Fundamentals of Convex Optimization
- Support Vector Machines
- Other Topics*

A little history about AI

人工智能

1956年，**麦卡锡**召集哈佛大学、麻省理工学院、IBM公司、贝尔实验室的研究人员召开**达特茅斯会议**正式提出“**人工智能**”

2006年达特茅斯会议当事人重聚，左起：**摩尔、麦卡锡、明斯基、赛弗里奇、所罗门诺夫**



John McCarthy

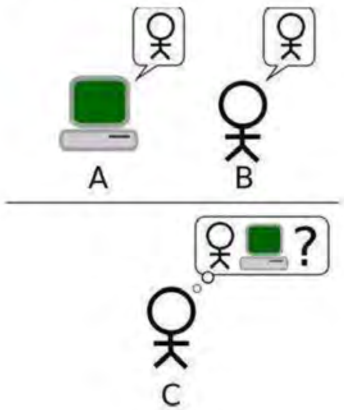
人工智能之父

人工智能是指计算机系统具备的能力，该能力可以履行原本只有依靠人类智慧才能完成的复杂任务

A little history about AI

什么是人工智能?

- 指由人制造出来的机器所表现出来的智能
  - 通常指通过计算机程序来呈现人类智能的技术
- **遗憾的是，“智能”本身难以定义清楚！**
  - 行为定义的智能Behavior defined intelligence
  - 即**图灵测试**定义的智能（不管内涵，只管外延）



A little history about AI

什么是人工智能?

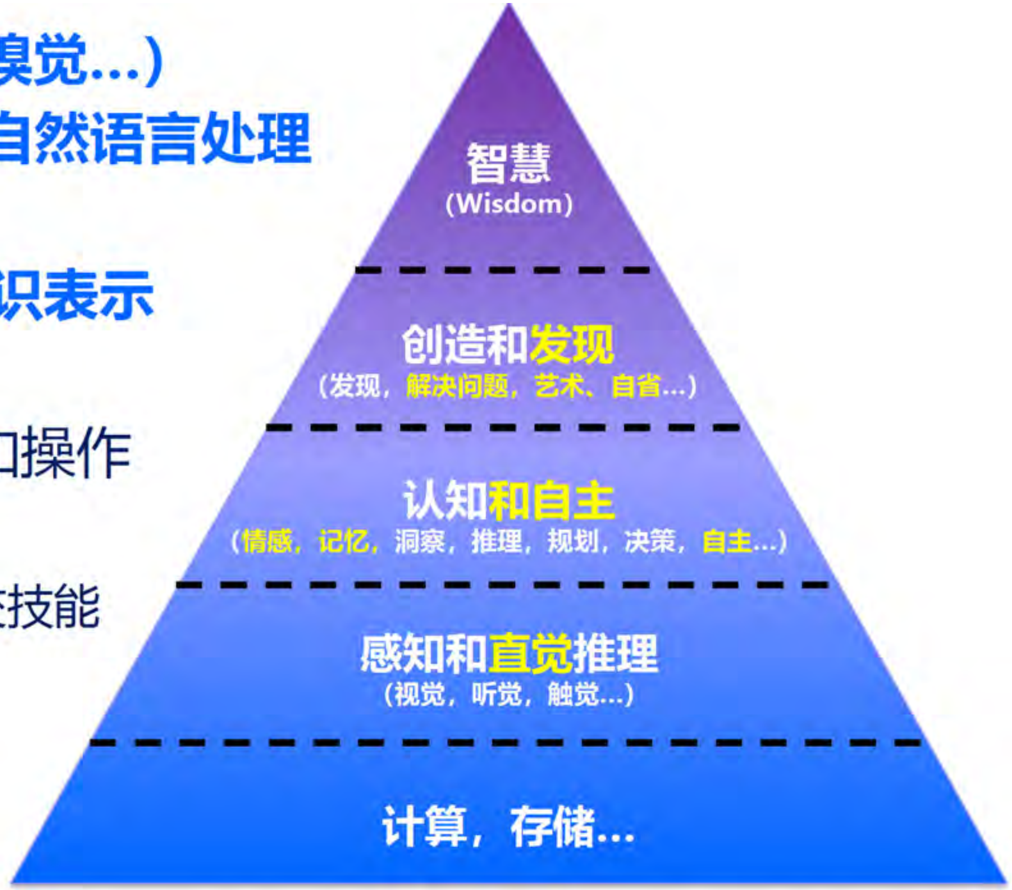
- 行为定义的智能Behavior defined intelligence
  - **系统的表现是智能的**

- 在计算机领域，人工智能是指对“智能代理”的研究
  - 任何可以**感知环境**并**采取行动**以最大可能达成其**特定目标**的任何设备都是智能代理【维基百科】

A little history about AI

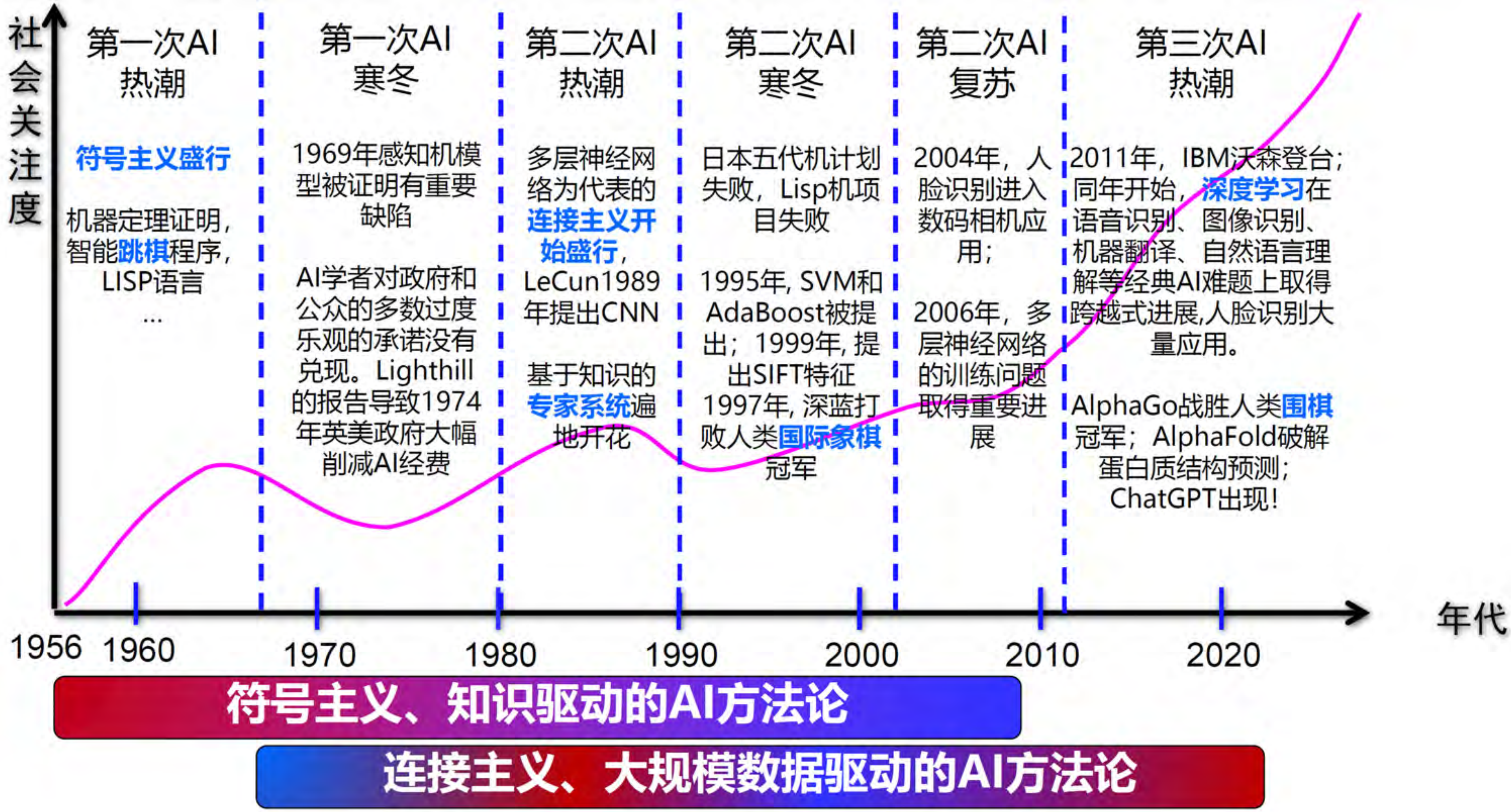
人工智能的内涵和任务

- **Perception感知（视觉，听觉，嗅觉...）**
- **Natural language processing自然语言处理**
- **Learning学习**
- **Knowledge representation 知识表示**
- Planning规划
- Motion and manipulation运动和操作
- Social intelligence社会智能
  - Affective Computing情感计算/社交技能
- Reasoning, problem solving
- Creativity创造力
- General intelligence通用智能



A little history about AI

人工智能发展的历史沿革



A little history about AI

三次人工智能浪潮背后的方法论

- **第一次浪潮：基于符号的推理与搜索**
  - 模拟人的符号推理方式
  - 搜索树：解决迷宫问题
- 符号主义
  - 符号主义的实现基础是纽威尔和西蒙提出的物理符号系统假设
    - 人类**认知和思维的基本单元是符号**，而**认知过程就是在符号表示上的一种运算**。人是一个物理符号系统，计算机也是一个物理符号系统，故可用计算机来模拟人的智能行为，即用计算机的符号操作来模拟人的认知过程
    - 实质就是模拟人的左脑抽象逻辑思维，通过**研究人类认知系统的功能机理，用符号之间的逻辑关系来描述人类的认知过程**，并把这种符号输入到能处理符号的计算机中，就可以模拟人类的认知过程，从而实现人工智能

三次人工智能浪潮背后的方法论

- **第一次浪潮：基于符号的推理与搜索**
  - 模拟人的符号推理方式
  - 搜索树：解决迷宫问题
- 符号主义
  - 太乐观了【 承诺太多，最终做不到 】！
    - 解决不了更复杂的现实问题
      搜索空间太大了：如围棋
    - **大量问题难以转化为符号推理问题**
      例如：人脸识别、语音识别等模式识别问题（非结构化数据的结构化）

三次人工智能浪潮背后的方法论

- **第二次浪潮：依赖人类符号化知识的专家系统**
  - **以依赖符号化知识库和符号推理的专家系统为主**
  - 知识表示是人工智能的核心难题
  - 人工智能研究早期主流的知识表示方法——**符号主义的知识观**
    - 从符号主义的观点来看，**认知就是符号的处理过程**，是智能的基础
    - **符号化的知识表示、推理、运用**是人工智能的核心
    - 知识表示：采用符号表示（实体、关系等）所有知识
    - 知识推理：推理是采用启发式知识及启发式搜索对**问题求解的过程**，其过程可以用某种**形式化的语言**来描述，因而有可能建立起基于符号化知识的人类智能和机器智能的**同一理论体系**

A little history about AI

三次人工智能浪潮背后的方法论

- **第二次浪潮：依赖人类符号化知识的专家系统**
  - 多数AI系统建立在符号基础上的知识表示（知识库、知识图谱）
    - •例：医疗辅诊系统，症状->疾病
  - 巅峰之作：IBM 的沃森自动问答系统
    - 2011年，IBM 沃森在问答竞赛 《 危险边缘 》（Jeopardy）上击败人类
    - 类似问题：地球上最北端的机场是哪个?
    - 背后的技术

      自然语言处理、消息检索、知识表示、自动推理、机器学习等开放式问答技术

A little history about AI

三次人工智能浪潮背后的方法论

- **第二次浪潮：依赖人类符号化知识的专家系统**
  - 还是太乐观了
    - 雄心勃勃的日本**第五代计算机**计划失败
    - 美国Cyc常识知识库项目陷入困境（至少不算成功）

      1984年启动，Douglas Lenat教授领衔，以手工建立知识库为主，包含了320万条人类定义的断言，涉及30万个概念，15000个谓词
    - **挑战性问题：常识是否可穷尽枚举?**
    - 难以解决复杂的现实问题

      知识表示困境：文字识别尚可，人脸识别用什么知识表示?
  - 方法论层面
    - 方法论上的悄然变迁，基于**专家知识人工设计特征**，采用**统计模式识别**和**机器学习**（包括神经网络）工具，学习较小规模数据之间统计关系，成为主流方法

三次人工智能浪潮背后的方法论

- **第二次浪潮：依赖人类符号化知识的专家系统**
  - 第二次浪潮末期：数据驱动的机器学习方法的崛起
  - 基本原理—基于函数拟合的预测问题

- 用[较大量的]成对的(x_i, y_i)数据，拟合一个带有θ**参数**的函数f
  - 本质：学习x和y的相关性；类比：学生学习过程，x_i是**考题**，y_i是**答案**
- 函数f经常是人工设计的，例如：线性函数$y = f(x) = Ax$
- **参数**θ量相对较少（但也经常在数十万甚至数百万量级）

三次人工智能浪潮背后的方法论

■ **第二次浪潮：依赖人类符号化知识的专家系统**

■ 方法论层面——**基于知识的特征设计**

□ **两篇文章**的相似度计算

三次人工智能浪潮背后的方法论

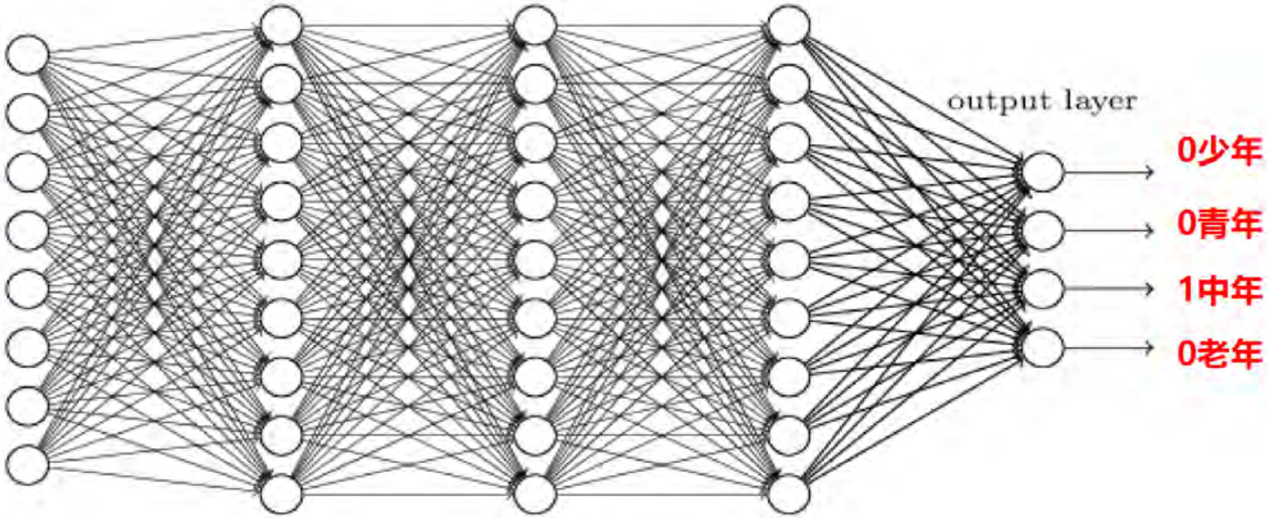
- **第二次浪潮：依赖人类符号化知识的专家系统**

- 方法论层面——**基于知识的特征设计**
  - **两张人脸**的相似度计算
    - **步骤1：** 图像中若干个点形成的**微模式类型**
    - **步骤2：** 统计人脸上**不同微模式的出现频次**作为不同人脸的**特征表示**

三次人工智能浪潮背后的方法论

- **第三次浪潮：依赖大量数据的深度学习方法**
  - 用神经网络作为映射函数f**直接学习**从输入x预测输出y
    - **较少依赖人工设计**
  - **题海战术（动辄百万，千万量级）**

x



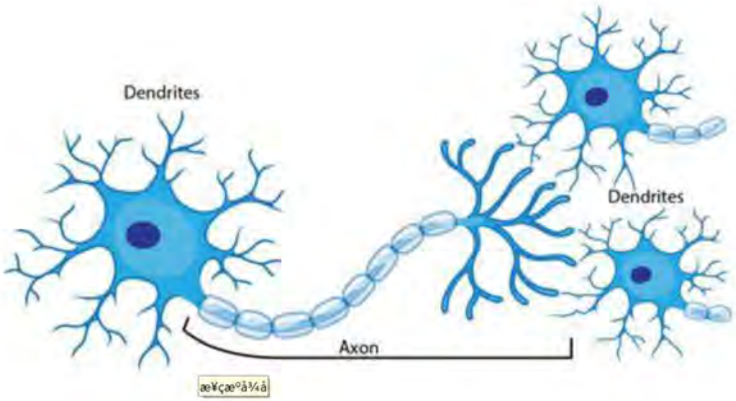
A little history about AI

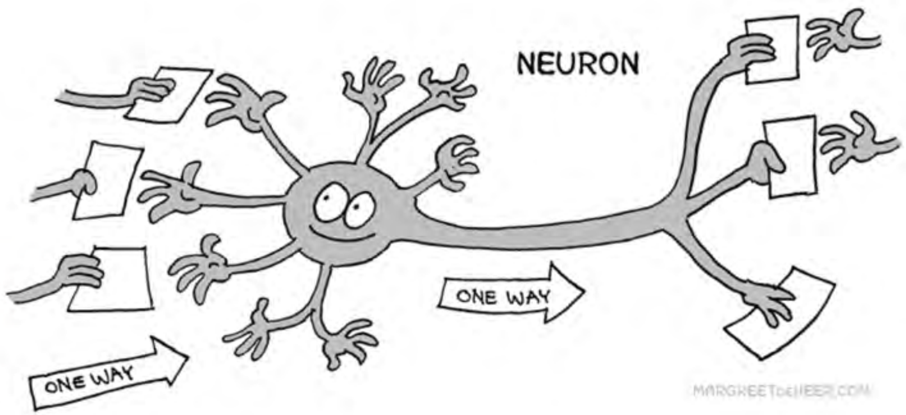
三次人工智能浪潮背后的方法论

- 第三次浪潮：依赖大量数据的深度学习方法

深度学习(深度卷积神经网络)的缘起

- 生物脑中的神经网络，单个神经元的**功能**
  - 接收前面神经元的输入，汇总➔决策➔传递





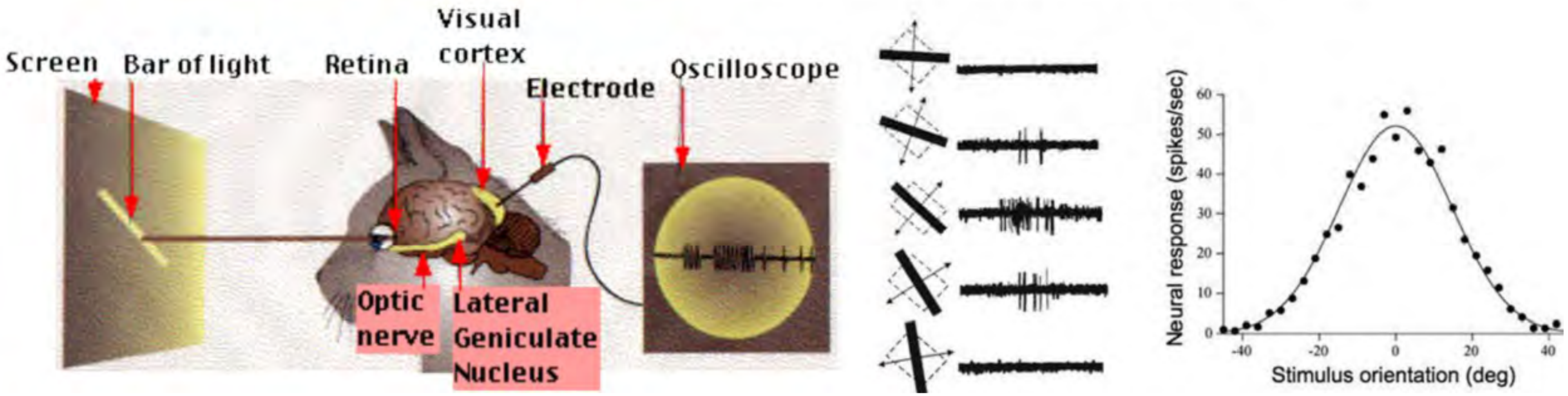
A little history about AI

三次人工智能浪潮背后的方法论

■ **第三次浪潮：依赖大量数据的深度学习方法**

深度学习(深度卷积神经网络)的缘起

- 生物脑中的神经网络，单个神经元的**功能**
  - 接收前面神经元的输入，汇总→决策→传递
  - 初级视觉皮层(V1)区简单细胞
    - **功能是检测不同朝向的线段** Hubel & Wiesel, 1959, 1962, ...

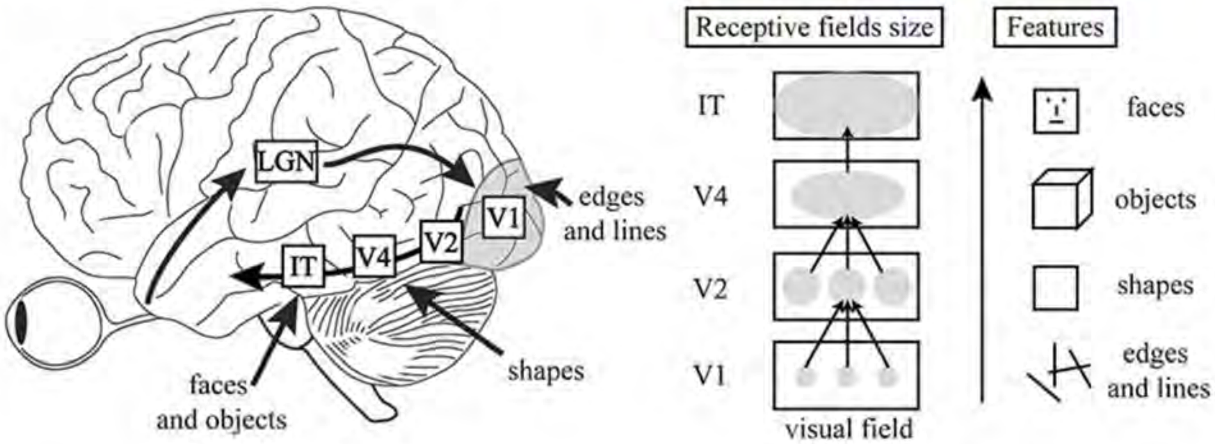


三次人工智能浪潮背后的方法论

■ **第三次浪潮：依赖大量数据的深度学习方法**

深度学习(深度卷积神经网络)的缘起

- 生物脑中的神经网络，**大量神经元互联**
  - 视觉通路神经细胞层级感受野假设
    - 响应越来越复杂的模式→祖母细胞理论
    - 可见越来越大的(视网膜)感受野：**类比从普通士兵到总司令**



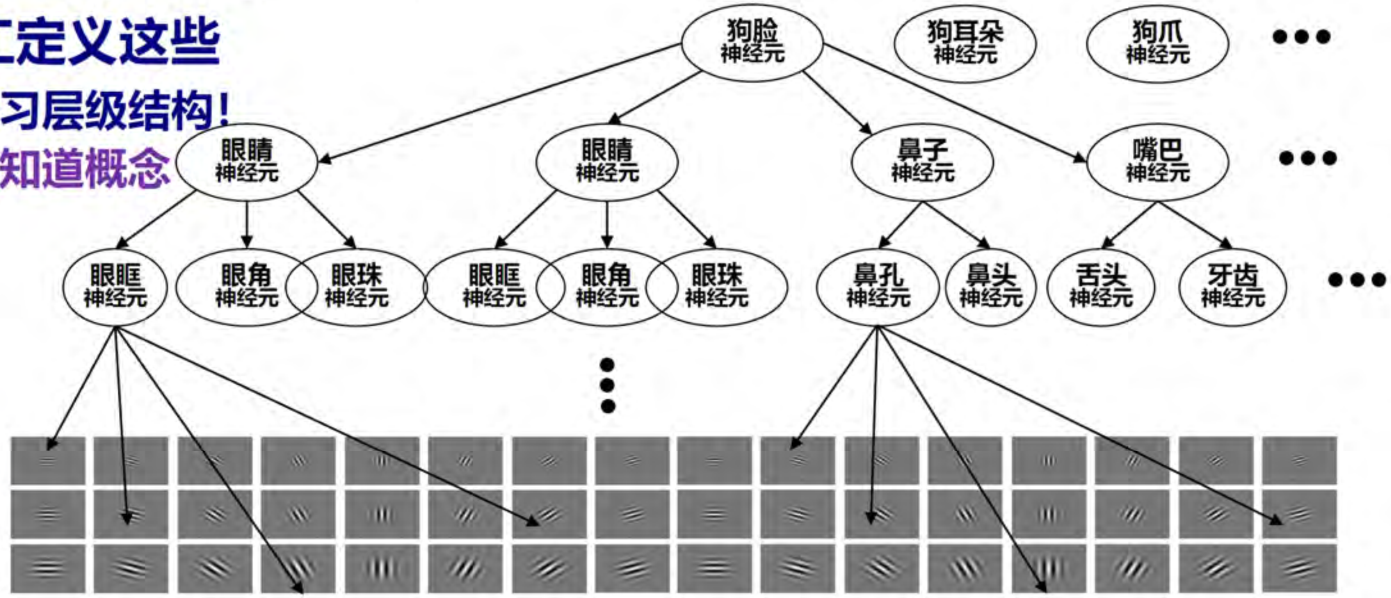
A little history about AI

三次人工智能浪潮背后的方法论

- **第三次浪潮：依赖大量数据的深度学习方法**

□ **所谓深度学习主要是指多层神经网络**

- **一个例子：多层神经网络怎么找到狗?**
- **但不需要人工定义这些**
  - □ **算法自动学习层级结构!**
  - □ **算法不需要知道概念**

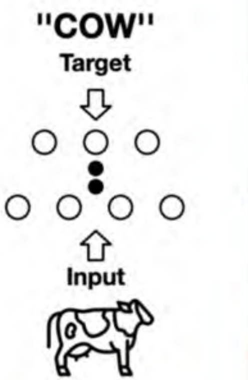


A little history about AI

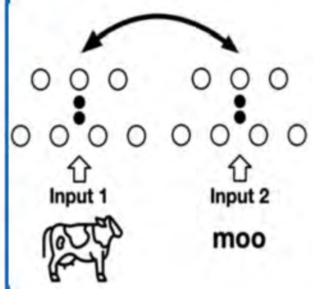
预训练大模型有望突破通用人工智能瓶颈

开启了基于自监督学习的“大数据+大模型”新范式，从大规模的无标注数据中挖掘隐含的监督信息进行通用知识学习，成为迈向通用人工智能的重要途径

1 从有监督到自监督



有限人工标注训练样本



大规模无标注训练样本

2 从专用小模型到通用大模型

目标检测

人脸识别

高度定制

大模型

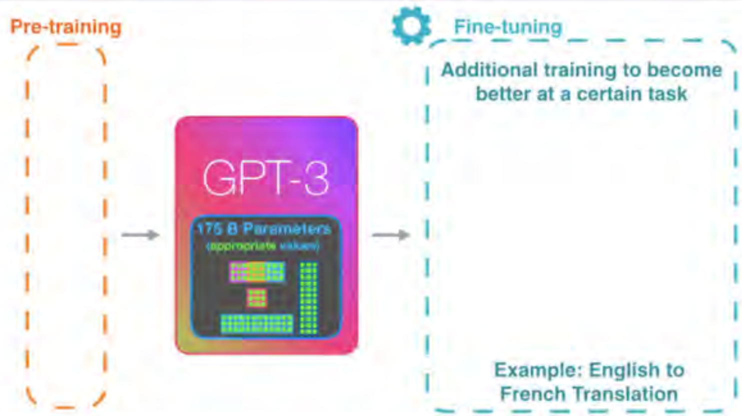
多专多能

A little history about AI

预训练大模型有望突破通用人工智能瓶颈

自然语言理解领域的大杀器——GPT-3

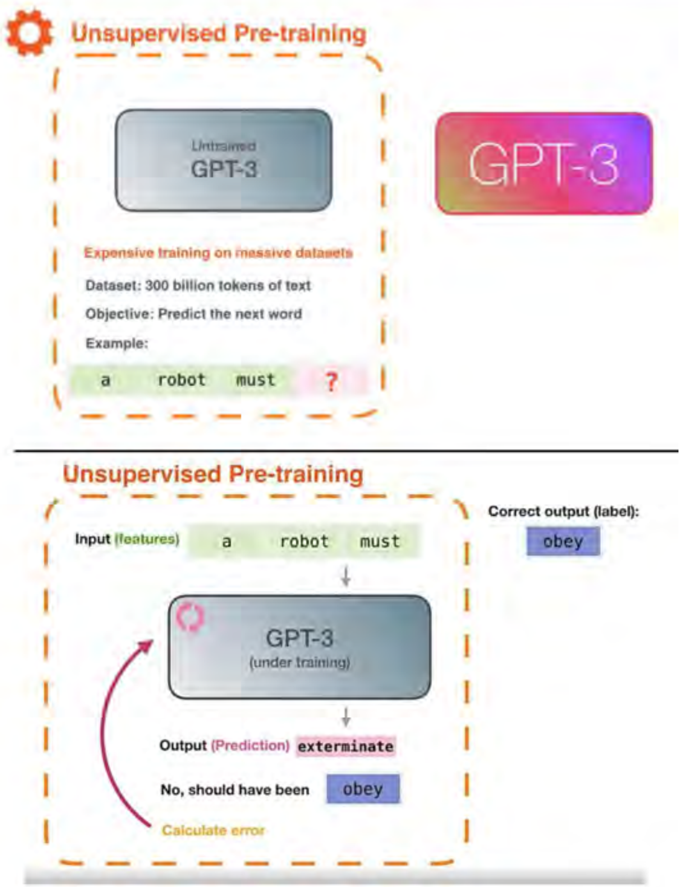
- **2020年6月，OpenAI GPT-3**
  - □1750亿参数，比其前身多100倍
    - 比之前最大NLP模型要多10倍
    - 花费460万美元进行训练
  - □大力出奇迹：见过巨量的人类语言
    - 训练语料：**3000亿单词（tokens）**
      - □ 60%：C4语料库（爬虫项目 Common Crawl 在 2019 年 4 月全网部分文本快照）
      - □ 22%：WebText2 （OpenAI自己收集的，未全部开放）
      - □ 16%：Books
      - □ **3%：Wikipedia**
    - 整个英语维基百科（约600万个词条）仅占其训练数据的3%



预训练大模型有望突破通用人工智能瓶颈

自然语言理解领域的大杀器—GPT-3预训练

- GPT-3 预训练
  - □无监督学习（自监督学习）
  - □Language Modeling：从前述词预测下一个词
    - 如左图中通过 “a robot must” 来预测下一个词 “obey”
  - □通过学习语言，同时学到了以自然语言中表达的大量 “知识”



来源：How GPT3 Works - Visualizations and Animations, Jay Alammar

A little history about AI

预训练大模型有望突破通用人工智能瓶颈

GPT, GPT-2, GPT-3

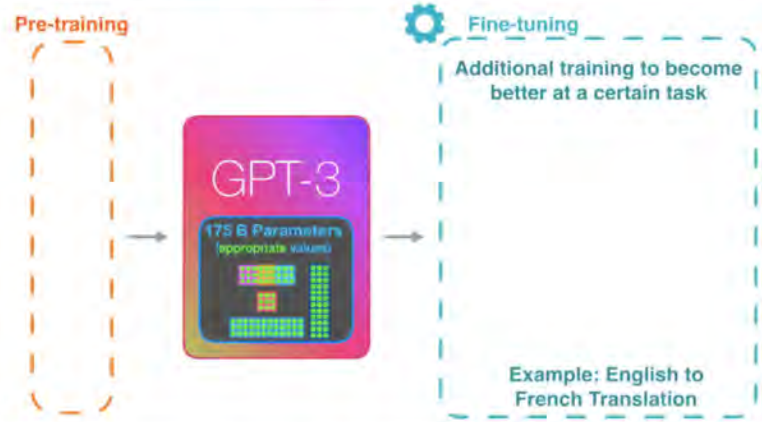
	GPT	GPT-2	GPT-3
数据集	5GB: BookCorpus	40GB: WebText	45TB: Common Crawl, WebText2, Books1, Books2, Wikipedia
参数量	117M	1.5B	175B
训练方法	Unsupervised pre-training, fine-tuning on each task	Unsupervised multitask pre-training via meta-learning, zero shot	Unsupervised multitask pre-training via meta-learning, zero/one/few shot
模型结构	Decoder (layer=12, dim=768, head=12)	Decoder (layer=48, dim=1600)	Decoder (layer=96, dim=12888, head=96)

A little history about AI

预训练大模型有望突破通用人工智能瓶颈

自然语言理解领域的大杀器——GPT-3

- **2020年6月，OpenAI GPT-3**
  - □1750亿参数，比其前身多100倍
    - 比之前最大NLP模型要多10倍
    - 花费460万美元进行训练
  - □大力出奇迹：见过巨量的人类语言
    - 整个英语维基百科（约600万个词条）仅占其训练数据的3%



- 可以做什么?
  - □回答问题，基于问题的搜索引擎，聊天机器人，机器翻译，续写文章…

预训练大模型有望突破通用人工智能瓶颈

ChatGPT是什么?

- ChatGPT基于大规模语言模型GPT3.5，通过**人类反馈学习微调**而来的对话生成大模型。不再是传统意义的人机对话系统，是以自然语言为交互的通用语言处理平台
- 超出预期的交互体验
  - 通用的意图理解能力
  - 强大的连续对话能力
  - 智能的交互修正能力
  - 较强的逻辑推理能力

推出**2个月**即达到**1亿**活跃用户
历史上**增长最快**的消费者应用程序

将对文字编辑、程序编译、智能问答等行业带来巨大冲击

A little history about AI

预训练大模型有望突破通用人工智能瓶颈

ChatGPT基础数据：文本与代码

该页slide来自中科院自动化所刘静研究员

2020年OpenAI利用 **45T** 文本数据，通过自监督训练获得基础大模型GPT-3，实现**流畅性、知识性**

专业书籍
维基百科
互联网文本
……

丰富的语言知识
多样的语言表达

GPT-3
能说会道

更多更新数据

C++
Java
Python
……

全面的逻辑实现
详细的代码注释

CodeX
逻辑编程

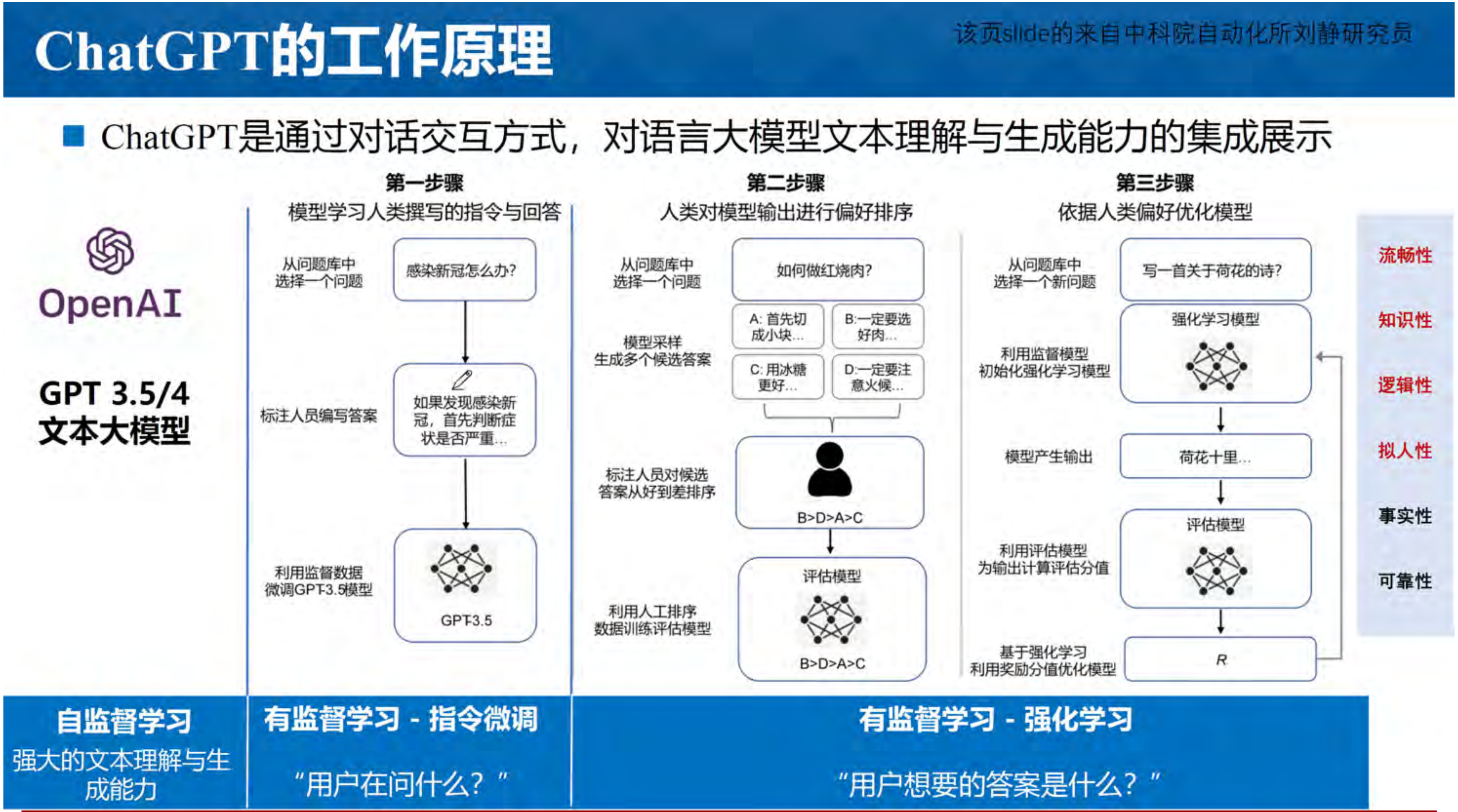
更多更新代码

2022年OpenAI利用更多更新文本数据和代码数据的混合学习，得到更强的基础大模型GPT-3.5，成为ChatGPT的基础模型，实现了**流畅性、知识性和逻辑性（推理能力）**

2021年OpenAI在GPT-3基础上利用 **179G** 代码数据，通过自监督训练获得**逻辑**编程模型Codex

A little history about AI

预训练大模型有望突破通用人工智能瓶颈

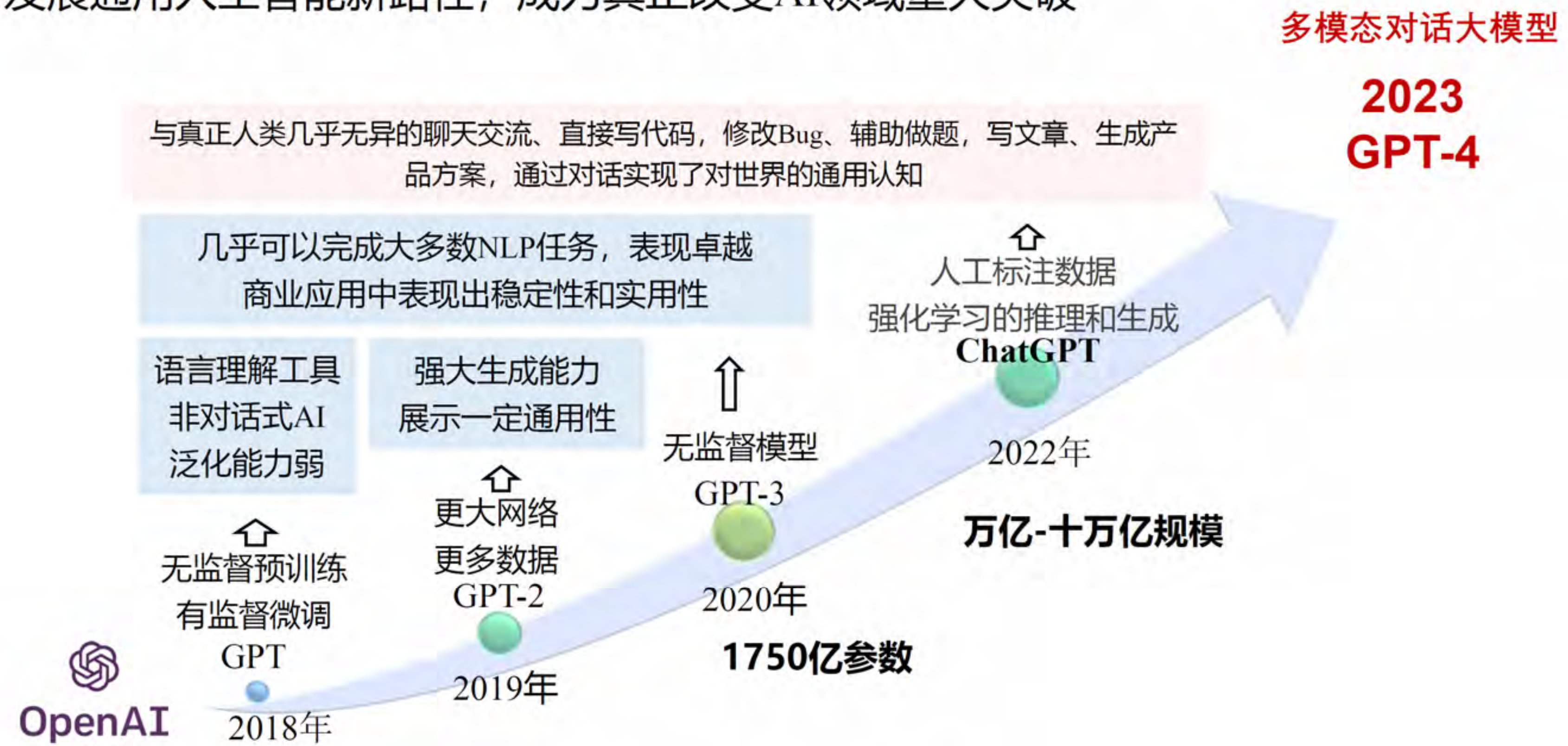


A little history about AI

预训练大模型有望突破通用人工智能瓶颈

ChatGPT以产品为导向，众多技术与成果的集大成者

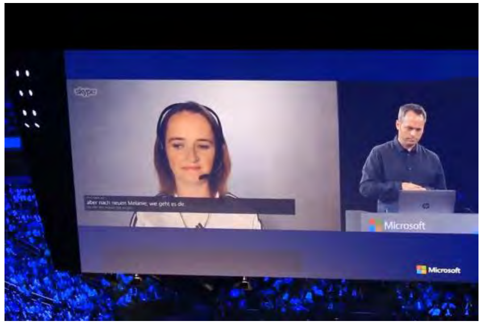
- **大模型技术**与**人类反馈强化学习**融合，实现知识逻辑涌现和人类价值观模拟，探索出了发展通用人工智能新路径，成为真正改变AI领域重大突破



A little history about AI

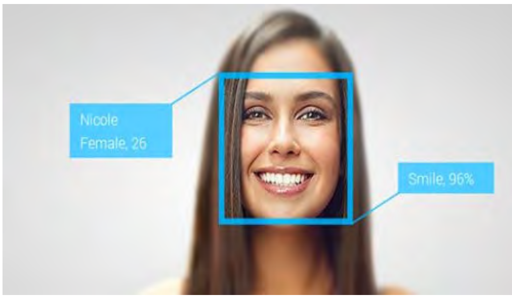
人工智能产业发展加速明显

自然语言处理（NLP）：
微软Skype Translator实现同声传译



计算机视觉（CV）：
格林深瞳的视频监控可智能识别犯罪

计算机视觉（CV）：
Face++的人脸识别云服务



感知、规划和决策：
Google无人驾驶汽车

A little history about AI

人工智能成为世界焦点

人工智能目前已经成为世界各国关注的焦点。2017年7月，中国政府发布了“新一代人工智能发展规划”

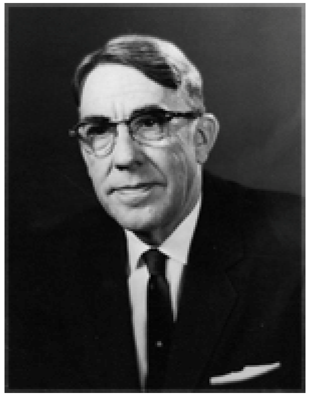
人工智能是开启未来智能世界的秘钥，是未来科技发展的战略制高点；谁掌握人工智能，谁就将成为未来核心技术的掌控者

傍晚，小街路面上沁出微雨后的湿润，和煦的细风吹来，抬头看看天边的晚霞，嗯，明天又是一个好天气。走到水果摊旁，挑了个根蒂蜷缩、敲起来声音浊响的青绿西瓜，一边满心期待着皮薄肉厚瓤甜的爽落感，一边愉快地想着：这学期狠下了功夫，基础概念弄得清清楚楚，算法作业也是信手拈来，这门课成绩一定差不了！

摘自《机器学习》（周志华著，2016）

What is machine learning?

- Gives "computers the ability to learn without being explicitly programmed" (Arthur Samuel in 1959)



Arthur Lee Samuel
(December 5, 1901 – July 29, 1990)

- It explores the study and construction of algorithms that can learn from and make predictions on data
- It is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible

[1] Samuel, Arthur L., Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, 1959

Supervised VS Unsupervised

- Supervised learning
  - It will infer a function from labeled training data; the training data consists of a set of training examples; each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal)
- Unsupervised learning
  - Trying to find hidden structure in unlabeled data; since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution; such as PCA, K-means (a clustering algorithm), AutoEncoder
- Semi-supervised learning
- Reinforcement learning

About sample

- Attribute (feature), attribute value, label, and example

色泽，根蒂，敲声 → {好瓜，坏瓜}

features　　labels

{青绿，蜷缩，浊响：好瓜}

feature values　　label value

one example

Training, testing, and validation

- Training sample and training set

A training set comprising m training samples,

$$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),...,(\boldsymbol{x}_m,y_m)\}$$

where $\boldsymbol{x}_i=(x_{i1},x_{i2},...,x_{id})\in\chi$ is the feature vector of ith sample and $y_i\in\zeta$ is its label

By training, our aim is to find a mapping,

$$f:\chi\mapsto\zeta$$

based on D

If ζ comprises discrete values, such a prediction task is called "**classification**"; if it comprises real numbers, such a prediction task is called "**regression**"

Training, testing, and validation

- Training sample and training set
- Test set
  - A test set is a set of data that is independent of the training data, but that follows the same probability distribution as the training data
  - Used only to assess the performance of a fully specified classifier

Training, testing, and validation

- Training sample and training set
- Test set
- Validation set
  - In order to avoid overfitting, when any classification parameter needs to be adjusted, it is necessary to have a validation set; it is used for model selection
  - The training set is used to train the candidate algorithms, while the validation set is used to compare their performances and decide which one to take

Overfitting, Generalization, and Capacity

- Overfitting
  - It occurs when a statistical model describes random error or noise instead of the underlying relationship
  - It generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations
  - A model that has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data

Overfitting, Generalization, and Capacity

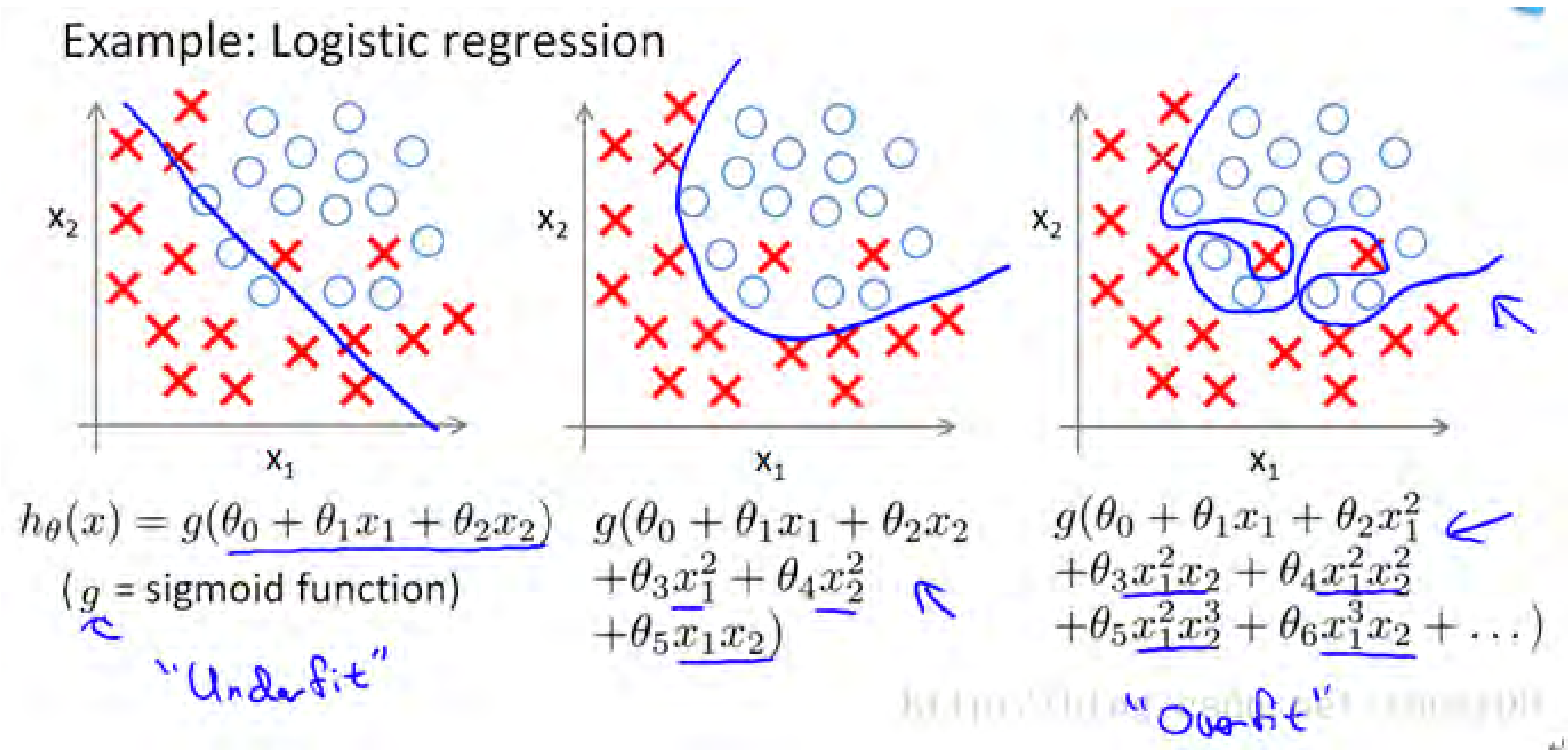
- Overfitting
- Generalization
  - Refers to the performance of the learned model on new, previously unseen examples, such as the test set

Overfitting, Generalization, and Capacity

- Overfitting
- Generalization

Overfitting, Generalization, and Capacity

- Overfitting
- Generalization



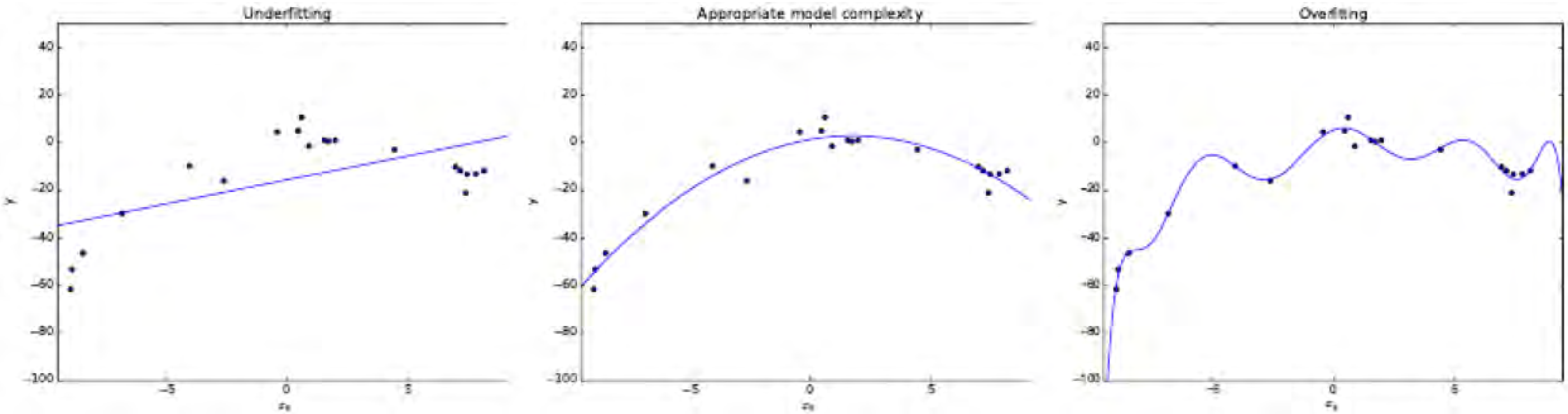
Overfitting, Generalization, and Capacity

- Overfitting
- Generalization
- Capacity
  - Measures the complexity, expressive power, richness, or flexibility of a classification algorithm
  - Ex, DCNN (deep convolutional neural networks) is powerful since its capacity is very large

$$y^* = b + \omega x, \quad y^* = b + \omega_1 x_1 + \omega_2 x_2, \quad y^* = b + \sum_{i=1}^{10} \omega_i x_i$$

$\xrightarrow{\hspace{4em}}$ higher capacity

Overfitting, Generalization, and Capacity



higher capacity

Performance Evaluation

Given a sample set (training, validation, or test)

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

To assess the performance of the learner f, we need to compare the prediction $f(\boldsymbol{x})$ and its ground-truth label y

For regression task, the most common performance measure is MSE (mean squared error),

$$E(f; D) = \frac{1}{m}\sum_{i=1}^{m}\left(f(\boldsymbol{x}_i) - y_i\right)^2$$

Performance Evaluation (for classification)

- Error rate
  - The ratio of the number of misclassified samples to the total number of samples

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(f(\boldsymbol{x}_i)\neq y_i\right)$$

- Accuracy
  - It is derived from the error rate

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(f(\boldsymbol{x}_i)=y_i\right)=1-E(f;D)$$

Performance Evaluation (for classification)

- Precision and Recall

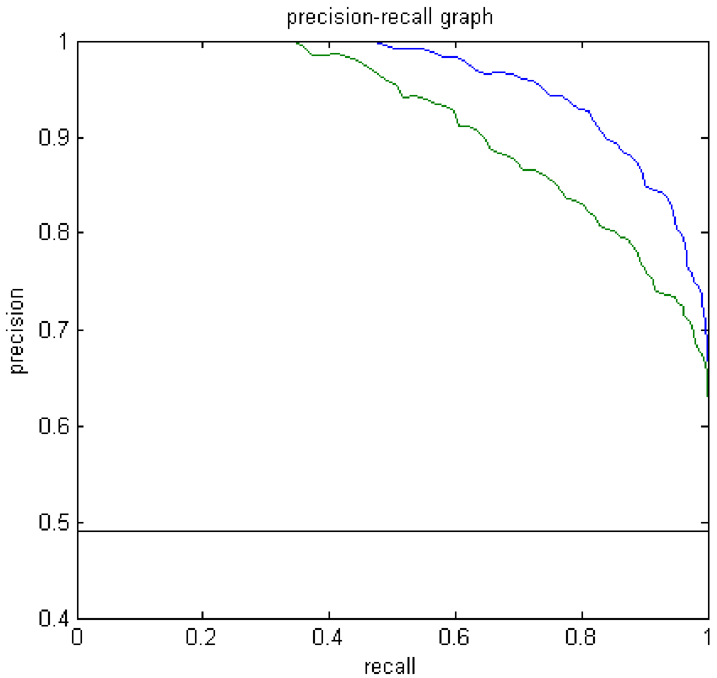
Ground truth	Prediction	
	positive	negative
positive	True Positive (TP)	False Negative (FN)
negative	False Positive (FP)	True Negative (TN)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Performance Evaluation (for classification)

- Precision and Recall
  - Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other
  - Usually, PR-curve is not monotonic



Performance Evaluation (for classification)

- Precision-recall should be used together; it is meaningless to use only one of them
- However, in many cases, people want to know explicitly which algorithm is better; we can use F-measure

$$F_{\beta} = \frac{(1+\beta^2)\times P \times R}{(\beta^2 \times P) + R}$$

Performance Evaluation (for classification)

- To derive a single performance measure

Varying threshold, we can have a series of (P, R) pairs,

$$(P_1, R_1), (P_2, R_2), \ldots, (P_n, R_n)$$

Then,

$$P_{macro} = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad R_{macro} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$F_{\beta-macro} = \frac{\left(1+\beta^2\right) \times P_{macro} \times R_{macro}}{\left(\beta^2 \times P_{macro}\right) + R_{macro}}$$

Model selection—Cross validation

- Simple cross validation
  - Split the dataset at hand into a training set and a validation set
  - Training the models on the training set, and selecting the best model based on the evaluation on the validation set
- S-fold cross validation
  - Randomly split the dataset at hand into S equal-sized subsets; any two subsets do not overlap with each other
  - For one learning model, train it on S-1 subsets and evaluate its performance on the remaining one subset; repeat such a training-evaluating procedure S times, each time using a different subset for evaluation; averaging the obtained S evaluation errors as the performance of this learning model
- Leave-one-out cross validation
  - It can be regarded as a special case of the S-fold cross validation strategy, i.e., $S=m$, m is the number of training samples

Class-imbalance Issue

- Problem definition
  - It is the problem in machine learning where the total number of a class of data is far less than the total number of another class of data
  - This problem is extremely common in practice
- Why is it a problem?
  - Most machine learning algorithms work best when the number of instances of each classes are roughly equal
  - When the number of instances of one class far exceeds the other, problems arise

Class-imbalance Issue

- How to deal with this issue?
  - Modify the cost function
  - Under-sampling, throwing out samples from majority classes
  - Oversampling, creating new virtual samples for minority classes
    - » Just duplicating the minority classes could lead the classifier to overfitting to a few examples
    - » Instead, use some algorithm for oversampling, such as SMOTE (synthetic minority over-sampling techniqe)[1]

[1] N.V. Chawla *et al*., SMOTE: Synthetic Minority Over-sampling Technique, J. Artificial Intelligence Research 16: 321-357, 2002

Class-imbalance Issue

- Minority oversampling by SMOTE[2]

Add new minority class instances by:

- For each minority class instance c
  - neighbours = Get KNN(5)
  - n = Random pick one from neighbours
  - Create a new minority class r instance using c's feature vector and the feature vector's difference of n and c multiplied by a random number
    - » i.e. r.feats = c.feats + (n.feats - c.feats) * rand(0,1)

[2] N.V. Chawla *et al.*, SMOTE: Synthetic Minority Over-sampling Technique, J. Artificial Intelligence Research 16: 321-357, 2002


Q+A