大数据时代的算法设计:
学习增强的数据流算法

董寅灏
计算机科学与技术学院

中国科学技术大学第一届“德·创”学科交叉研究生学术论坛 · 学科交叉中心分论坛
2026 年 1 月 15 日

什么是算法 (Algorithm)?

什么是算法 (Algorithm)?

- 解决问题的方法、步骤

什么是算法 (Algorithm)?

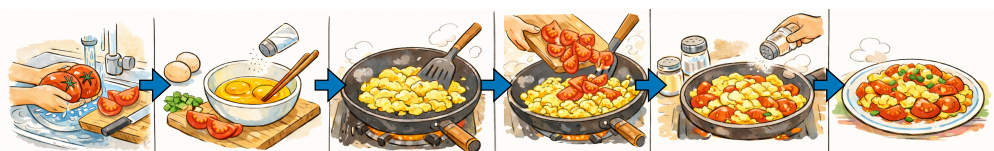
- 解决问题的方法、步骤
- **问题：**如何做一盘西红柿炒鸡蛋?

什么是算法 (Algorithm)?

- 解决问题的方法、步骤
- **问题：**如何做一盘西红柿炒鸡蛋?
  - **输入：** 
  - **输出：** 

什么是算法 (Algorithm)?

- 解决问题的方法、步骤
- **问题：**如何做一盘西红柿炒鸡蛋?
  - **输入：** 
  - **输出：** 
  - **算法：**



算法无处不在

路径规划

视频推荐

人脸识别

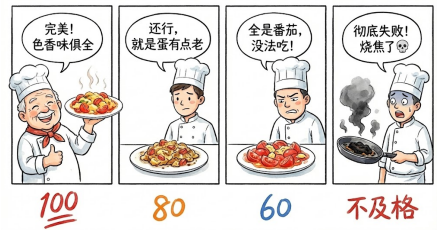
什么是“好”的算法?

什么是“好”的算法?

- **质量** (越高越好)

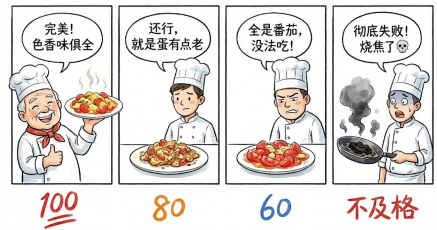
什么是“好”的算法?

- **质量** (越高越好)
  - 目标函数 f(外观, 口感, 健康)



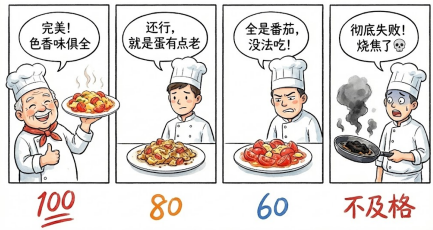
什么是“好”的算法?

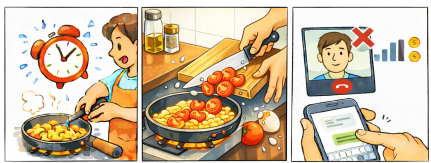
- **质量** (越高越好)
  - 目标函数 f(外观, 口感, 健康)
- **资源开销** (越少越好)



什么是“好”的算法?

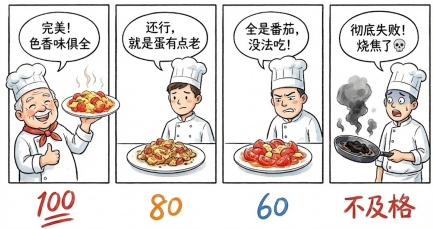
- **质量** (越高越好)
  - 目标函数 f(外观, 口感, 健康)
- **资源开销** (越少越好)
  - **时间**：午休很短，必须尽快做完
  - **空间**：台面很小，放不下所有食材，必须边切边下锅
  - **通信**：需要远程请教，但流量很贵，不能一直视频

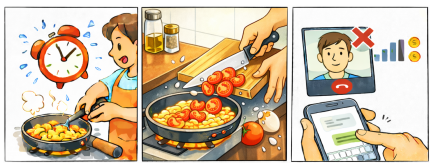




什么是“好”的算法?

- **质量** (越高越好)
  - 目标函数 f(外观, 口感, 健康)
- **资源开销** (越少越好)
  - **时间**：午休很短，必须尽快做完
  - **空间**：台面很小，放不下所有食材，必须边切边下锅
  - **通信**：需要远程请教，但流量很贵，不能一直视频
- **算法设计的目标**：在质量与资源开销之间权衡





我的研究领域：理论计算机科学

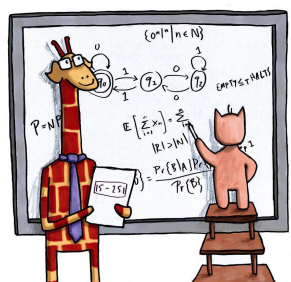
目标：设计有**理论保证**的算法

我的研究领域：理论计算机科学

目标：设计有**理论保证**的算法

理论保证：

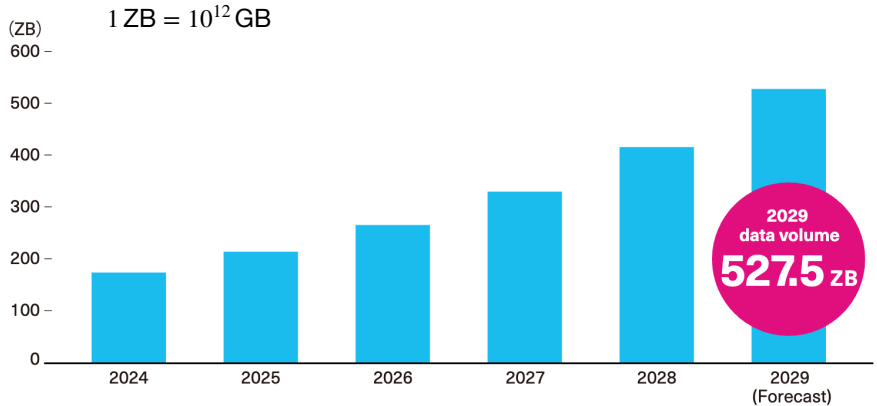
对于任意输入数据 (不管多“刁钻”)，算法的输出**质量**和**资源开销**都有**可证明**的保证 (而不只在某些数据上好)



图片来源：CMU CS251课程主页

我们正处在大数据时代

Fig. 1: Global Data Generation (ZB)[1]



图片来源：Kioxia, Integrated Report 2025; 数据来源：IDC, Worldwide IDC Global DataSphere Forecast, 2025-2029, #US53363625

我们正处在大数据时代

大数据时代给算法设计带来了挑战和机遇

图片来源：Kioxia, Integrated Report 2025; 数据来源：IDC, Worldwide IDC Global DataSphere Forecast, 2025-2029, #US53363625

大数据时代算法设计的挑战

大数据时代算法设计的挑战

- 传统计算模型的默认假设：输入数据可以全部放进内存，并反复读取

大数据时代算法设计的挑战

- 传统计算模型的默认假设：输入数据可以全部放进内存，并反复读取
- 这需要线性空间 (与输入数据规模成正比)，在大数据时代不切实际！

大数据时代算法设计的挑战

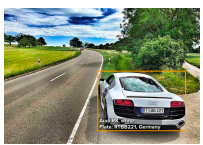
- 传统计算模型的默认假设：输入数据可以全部放进内存，并反复读取
- 这需要线性空间 (与输入数据规模成正比)，在大数据时代不切实际！

- 需要新的计算模型：在**内存空间远小于输入数据规模**时仍能(近似)求解

大数据时代算法设计的机遇

大数据时代算法设计的机遇

- 人工智能、机器学习技术蓬勃发展
  - 人工智能 (Artificial Intelligence, AI)：让机器模仿人类行为的技术总称
  - 机器学习 (Machine Learning, ML)：让机器**从数据中**学习规律的技术



目标检测

语音识别

AlphaGo 围棋机器人

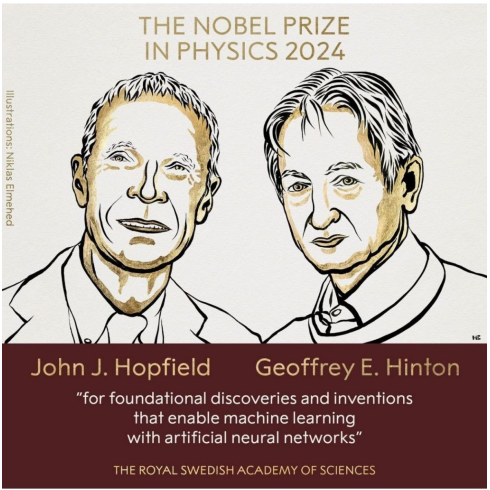
图像生成

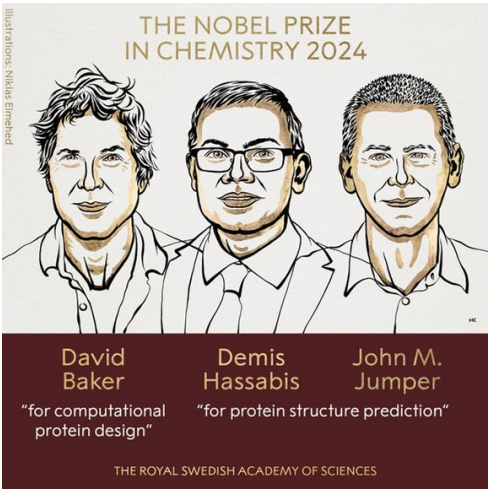
视频生成

自动驾驶

聊天机器人

2024年诺贝尔物理学奖、诺贝尔化学奖





我的研究目标

空间受限场景、机器学习辅助

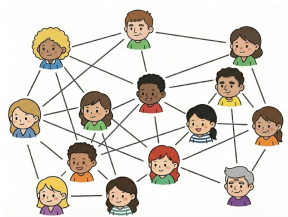
设计有理论保证的算法

优化“质量—空间”权衡

以一个具体问题为例：社区发现

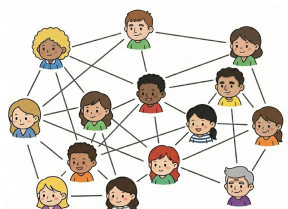
以一个具体问题为例：社区发现

- **输入**：社交网络 (计算机科学中称为“图”)
  - 把每个人看成一个点，两个人有联系/认识就连一条线 (图论中叫“边”)



以一个具体问题为例：社区发现

- **输入**：社交网络 (计算机科学中称为“图”)
  - 把每个人看成一个点，两个人有联系/认识就连一条线 (图论中叫“边”)
- **输出**：把所有人分成若干个社区 (每个人属于一个社区)
  - 社区内联系多，社区之间联系少



以一个具体问题为例：社区发现

- **输入**：社交网络 (计算机科学中称为“图”)
  - 把每个人看成一个点，两个人有联系/认识就连一条线 (图论中叫“边”)
- **输出**：把所有人分成若干个社区 (每个人属于一个社区)
  - 社区内联系多，社区之间联系少

以一个具体问题为例：社区发现

- **输入**：社交网络 (计算机科学中称为“图”)
  - 把每个人看成一个点，两个人有联系/认识就连一条线 (图论中叫“边”)
- **输出**：把所有人分成若干个社区 (每个人属于一个社区)
  - 社区内联系多，社区之间联系少
- **实际应用**：内容推荐 (社交媒体中“你可能感兴趣的内容”)

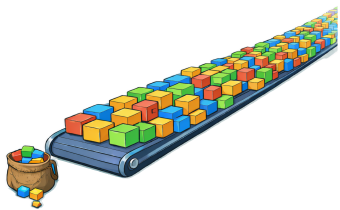
以一个具体问题为例：社区发现

以抖音为例，月活跃用户数已超 9 亿

- **输入**：社交网络 (计算机科学中称为“图”)
  - 把每个人看成一个点，两个人有联系/认识就连一条线 (图论中叫“边”)
- **输出**：把所有人分成若干个社区 (每个人属于一个社区)
  - 社区内联系多，社区之间联系少
- **实际应用**：内容推荐 (社交媒体中“你可能感兴趣的内容”)

大数据时代算法设计的理论模型

- **模型1：空间受限场景的建模**
  - 数据流模型
- **模型2：利用机器学习辅助算法设计**
  - 学习增强模型



空间受限场景建模：数据流模型

(Streaming Model)

空间受限场景建模：数据流模型

(Streaming Model)

- 1996 年提出，**获 2005 年哥德尔奖** (理论计算机科学领域最高奖)

空间受限场景建模：数据流模型

(Streaming Model)

- 1996 年提出，**获 2005 年哥德尔奖** (理论计算机科学领域最高奖)
- 输入数据不是完整给出，而是以**元素序列 (数据流)** 的形式到达

空间受限场景建模：数据流模型

(Streaming Model)

- 1996 年提出，**获 2005 年哥德尔奖** (理论计算机科学领域最高奖)
- 输入数据不是完整给出，而是以**元素序列 (数据流)** 的形式到达

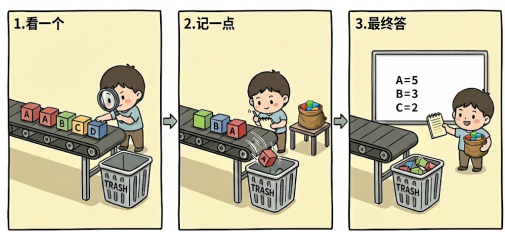
社交网络中，元素就是一条条连线/“边”

空间受限场景建模：数据流模型

(Streaming Model)

- 1996 年提出，**获 2005 年哥德尔奖** (理论计算机科学领域最高奖)
- 输入数据不是完整给出，而是以**元素序列 (数据流)** 的形式到达
- **算法框架：看一个、记一点、最终答**

社交网络中，元素就是一条条连线/“边”



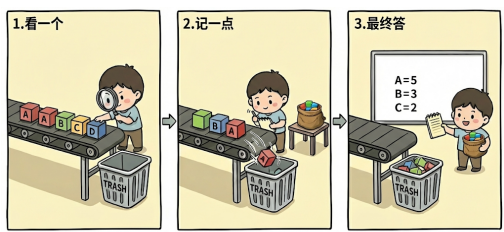
空间受限场景建模：数据流模型

(Streaming Model)

- 1996 年提出，**获 2005 年哥德尔奖** (理论计算机科学领域最高奖)
- 输入数据不是完整给出，而是以**元素序列 (数据流)** 的形式到达

社交网络中，元素就是一条条连线/“边”

- **算法框架：看一个、记一点、最终答**
  - 依次扫描数据流中的元素
  - 对于每个元素，进行一些操作
    (如保存当前元素或更新内存中信息)
  - 扫描结束，根据内存中的信息输出问题的解



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

机器学习辅助算法设计：学习增强模型

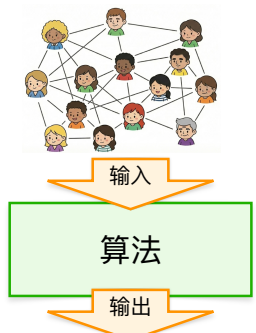
(Learning-Augmented Model)

- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”

机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

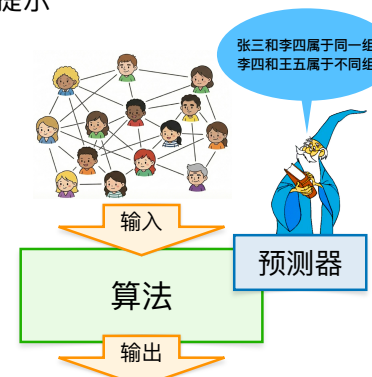
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

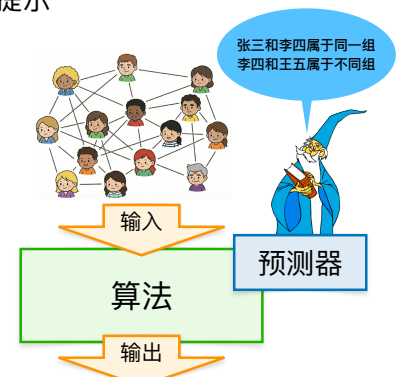
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

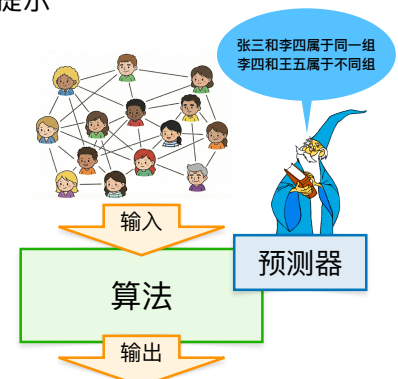
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”
  - 预测的形式是算法设计的一部分



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

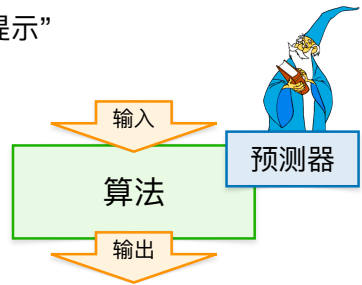
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”
  - 预测的形式是算法设计的一部分
  - 对算法设计者而言，预测器是黑盒
    - 只使用它的输出，不关心如何得到
    - 不知道预测质量，预测可能不准确 (甚至很差)



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

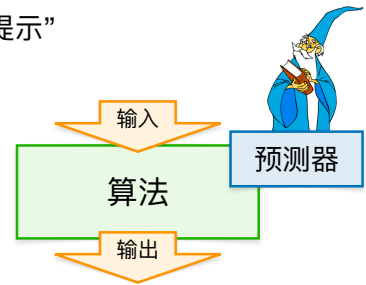
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”
  - 预测的形式是算法设计的一部分
  - 对算法设计者而言，预测器是黑盒
    - 只使用它的输出，不关心如何得到
    - 不知道预测质量，预测可能不准确 (甚至很差)
- **目标：“扬长避短”**



机器学习辅助算法设计：学习增强模型

(Learning-Augmented Model)

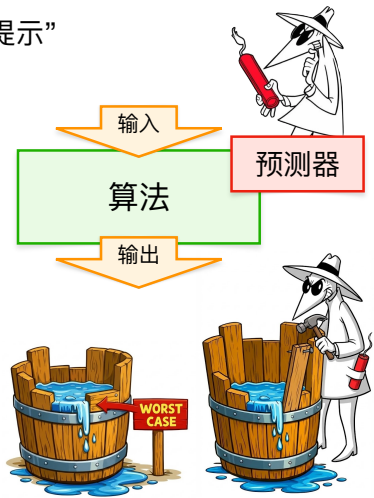
- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”
  - 预测的形式是算法设计的一部分
  - 对算法设计者而言，预测器是黑盒
    - 只使用它的输出，不关心如何得到
    - 不知道预测质量，预测可能不准确 (甚至很差)
- **目标：“扬长避短”**
  - 预测好时： 显著优于不使用预测的经典算法



机器学习辅助算法设计：学习增强模型

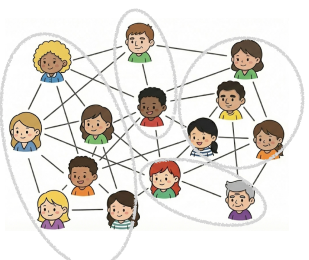
(Learning-Augmented Model)

- 算法可以访问一个**预测器**，获得关于输入数据的某种“提示”
  - 预测的形式是算法设计的一部分
  - 对算法设计者而言，预测器是黑盒
    - 只使用它的输出，不关心如何得到
    - 不知道预测质量，预测可能不准确 (甚至很差)
- **目标：“扬长避短”**
  - 预测好时： 显著优于不使用预测的经典算法
  - 预测差时： 仍不劣于不使用预测的经典算法



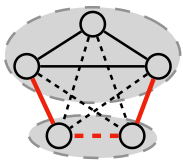
我的研究：学习增强的数据流算法

关注问题：社区发现 (图聚类/划分)



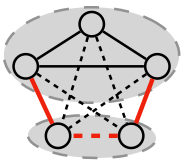
研究成果 (1)

- **问题1：关联聚类**
  - 把点分成若干组，使得“分错”的边数尽可能少
  - 应用背景：社区发现、图像分割、自动标注



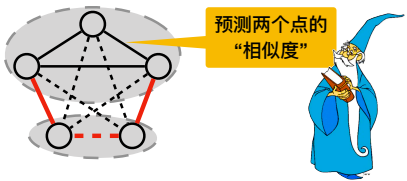
研究成果 (1)

- **问题1：关联聚类**
  - 把点分成若干组，使得“分错”的边数尽可能少
  - 应用背景：社区发现、图像分割、自动标注
  - 经典数据流算法(不用预测)的理论保证 [CKLPU24]
    - 对于任意输入图，$\frac{\text{算法解的值}}{\text{最优解的值}} \leq 3$，空间 $o(n^2)$，其中 n 为图中点数



研究成果 (1)

- **问题1：关联聚类**
  - 把点分成若干组，使得“分错”的边数尽可能少
  - 应用背景：社区发现、图像分割、自动标注
  - 经典数据流算法(不用预测)的理论保证 [CKLPU24]
    - 对于任意输入图，$\frac{\text{算法解的值}}{\text{最优解的值}} \le 3$，空间 $o(n^2)$，其中 n 为图中点数
  - 我们算法(使用预测)的理论保证 [DJLP25]

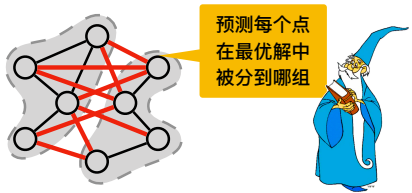


定理： 对于任意输入图，$\frac{\text{算法解的值}}{\text{最优解的值}} \le \min\{2.06\beta, 3\}$，空间 $o(n^2)$，其中 $\beta \ge 1$ 为预测质量
(β 越小, 预测越好)

[DJLP25] **Yinhao Dong**, Shan Jiang, Shi Li, Pan Peng. *Learning-Augmented Streaming Algorithms for Correlation Clustering*. **NeurIPS 2025**. (机器学习领域顶级国际会议, 中国计算机学会推荐A类会议)

研究成果 (2)

- **问题2：最大割** (只估计最优解的值)
  - 把点分成两组，使得两组之间的边数尽可能多
  - 应用背景：电路设计、统计物理
  - 经典数据流算法(不用预测)的理论保证
    - 平凡算法：对于任意输入图，$\frac{\text{最优解的值}}{\text{算法解的值}} = 2$，空间 $O(\log n)$，其中 n 为图中点数
    - 可以证明 [KK19]：要达到 $\frac{\text{最优解的值}}{\text{算法解的值}} \leq 2 - \epsilon$，空间至少为 $\Omega(n)$
  - 我们算法(使用预测)的理论保证 [DPV25]



定理： 对于任意输入图，$\frac{\text{最优解的值}}{\text{算法解的值}} \leq 2 - \epsilon$，空间 $O(\log n)$

[DPV25] **Yinhao Dong**, Pan Peng, Ali Vakilian. *Learning-Augmented Streaming Algorithms for Approximating MAX-CUT*. **ITCS 2025**. (理论计算机科学领域重要国际会议, 清华大学计算机学科群/中国科大信智学部推荐B类会议, 中国科大学生首篇)

总结

- 大数据时代给算法设计带来了挑战和机遇
  - 挑战：输入数据规模远超内存空间 → 怎么办? → **数据流算法**
  - 机遇：机器学习蓬勃发展 → 能帮什么? → **学习增强的算法**
- 目前以理论研究为主，期待推动实际应用
  - “*The best theory is inspired by practice. The best practice is inspired by theory.*” — Donald Knuth (高德纳，计算机科学家，1974年图灵奖得主)
- 学科交叉的可能?

致谢：本PPT中部分图片由ChatGPT、Gemini生成

总结

- 大数据时代给算法设计带来了挑战和机遇
  - 挑战：输入数据规模远超内存空间 → 怎么办? → **数据流算法**
  - 机遇：机器学习蓬勃发展 → 能帮什么? → **学习增强的算法**
- 目前以理论研究为主，期待推动实际应用
  - “*The best theory is inspired by practice. The best practice is inspired by theory.*” — Donald Knuth (高德纳，计算机科学家，1974年图灵奖得主)
- 学科交叉的可能?

谢谢大家！欢迎交流与合作😊

致谢：本PPT中部分图片由ChatGPT、Gemini生成