2.1 绕不开的房价预测

由回归问题开始了解机器学习

主讲人：李辉楚吴

前情提要——ML和DL与AI的关系

$$AI \subset ML \subset DL$$

人工智能（Artificial Intelligence，AI）

让机器展现出甚至超越人类的智慧

机器学习（Machine Learning，ML）

当前实现人工智能的一种主流方法

深度学习（Deep Learning，DL）

机器学习中的一种“大力出奇迹”方法

前情提要——ML和DL是什么?

ML和DL皆以学习的方式

从数据中寻找一个解决问题的函数

$$f(x) = y$$

前情提要——$f(x)$能解决哪些问题？

感知世界

回归
分类
聚类

创造（虚拟）世界

文本生成
图像生成
视频生成
语音生成

改变（物理）世界

具身智能
人形机器人
无人驾驶

从房价预测开始的机器学习

回归问题是什么？

讨论：准确判断房屋价格应该怎么做呢？

收集信息，信息越丰富越好！

学区房	景观房	商业中心
医疗条件	居民经济条件	交通情况
房型结构	房屋面积	……

找到房屋信息与房价间的联系

$$f\left(\begin{array}{|c|c|c|c|c|c|c|c|c|}\hline \text{学区房} & \text{湖景房} & \text{商业中心} & \text{医疗条件} & \text{经济条件} & \text{交通情况} & \text{房型结构} & \text{房屋面积} & \cdots\cdots \\ \hline \end{array}\right) = \text{房屋价格}$$

这就是回归问题

T 博士
《灌篮高手》

特点1：任务是预测或者推断

特点2：目标是找出输入和输出之间的函数关系

特点3：输出的是连续数值

f(大气压力 | 温度 | 湿度 | 风速 | 风向 | 降水 | 历史气候 | 云形 | ……) = 天气变化

f(GDP | 通货膨胀 | 财务状况 | 交易情况 | 股价走势 | 股权变动 | 舆论行情 | 行规变动 | ……) = 股价变化

这就是回归问题

T 博士
《灌篮高手》

特点1：任务是预测或者推断

特点2：目标是找出输入和输出之间的函数关系

特点3：输出的是连续数值

用机器学习的方法解决回归问题

案例研究

1978波士顿区域房价预测

波士顿房价预测问题——If…else…方法

步骤1：数据准备

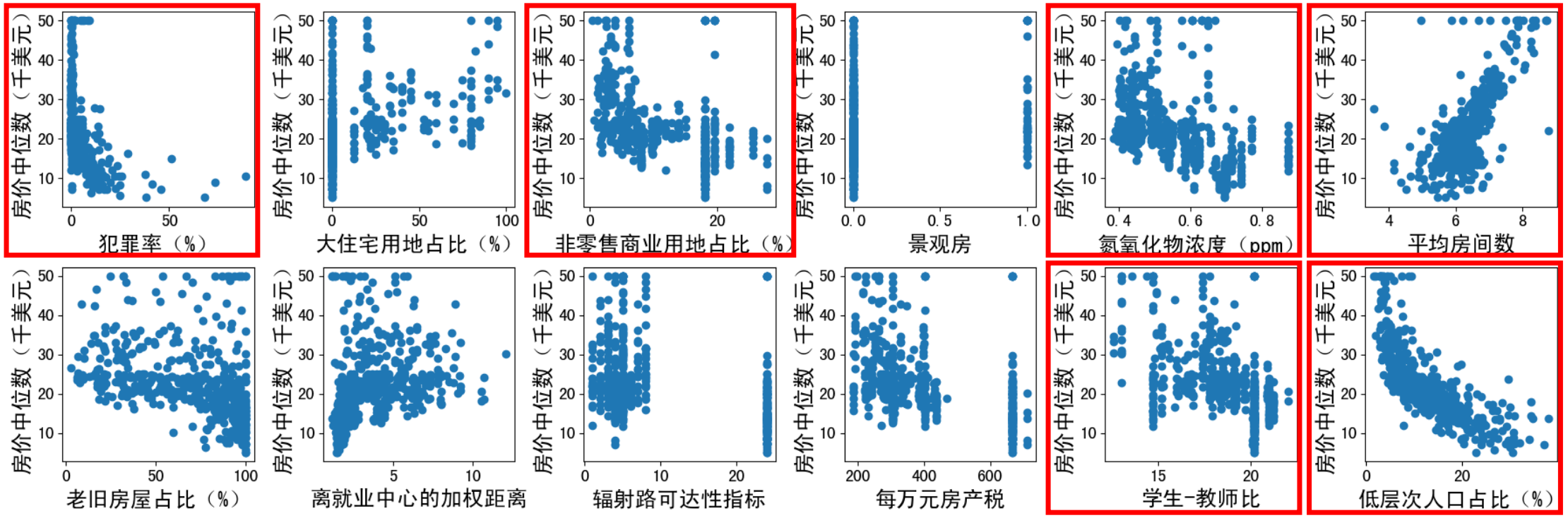
1978年波士顿区域房屋价格表

x_1：犯罪率（%）	0.00632	0.02731	0.02729	0.03237	0.06905
x_2：大住宅用地占比（%）	18.00	0.00	0.00	0.00	0.00
x_3：非零售商业用地占比（%）	2.31	7.07	7.07	2.18	2.18
x_4：景观房（0/1）	0	0	0	0	0
x_5：氮氧化物浓度（ppm）	0.538	0.469	0.469	0.458	0.458
x_6：平均房间数（个）	6.575	6.421	7.185	6.998	7.147
x_7：老旧房屋占比（%）	65.2	78.9	61.1	45.8	54.2
x_8：离就业中心的加权距离	4.09	4.9671	4.9671	6.0622	6.0622
……	……	……	……	……	……
y：房价中位数（千美元）	**24.00**	**21.6**	**34.7**	**33.4**	**36.2**

Harrison, D. and Rubinfeld, D.L. "Hedonic prices and the demand for clean air", J. Environ. Economics & Management, vol.5, 81-102, 1978.

波士顿房价预测问题——If…else…方法

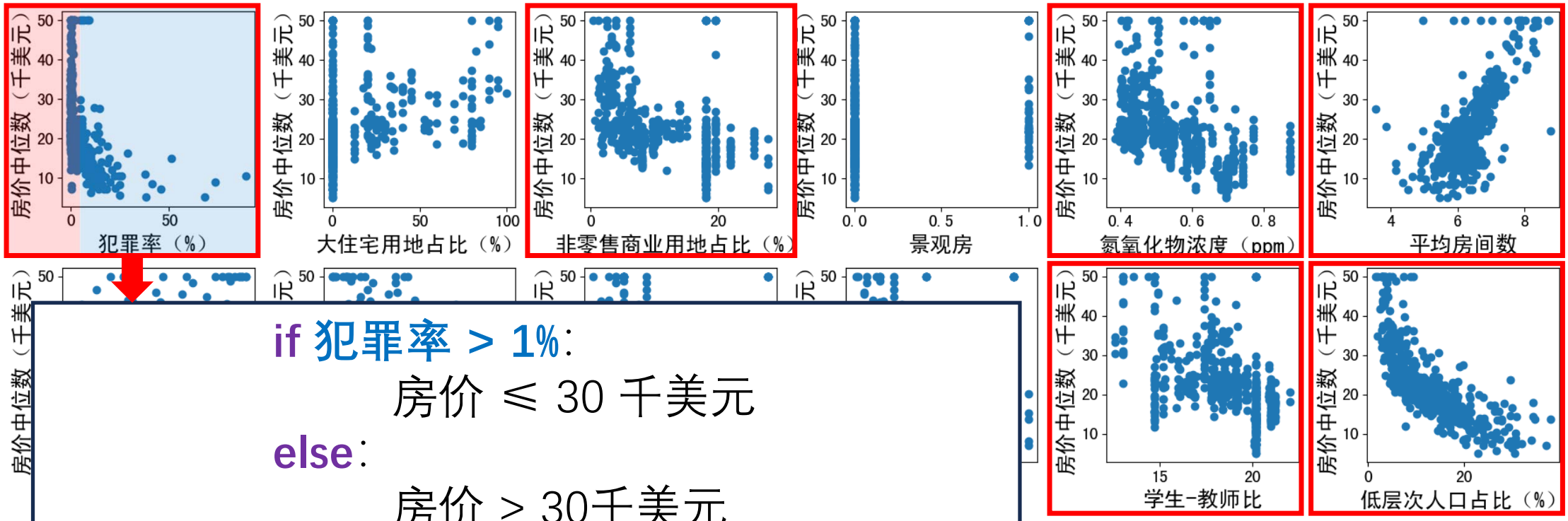
步骤2：分析数据，发现规律，构建规则（人工）



部分特征与房价明显相关，部分特征有一定影响

波士顿房价预测问题——If…else…方法

步骤2：分析数据，发现规律，构建规则（人工）



if 犯罪率 > 1%:

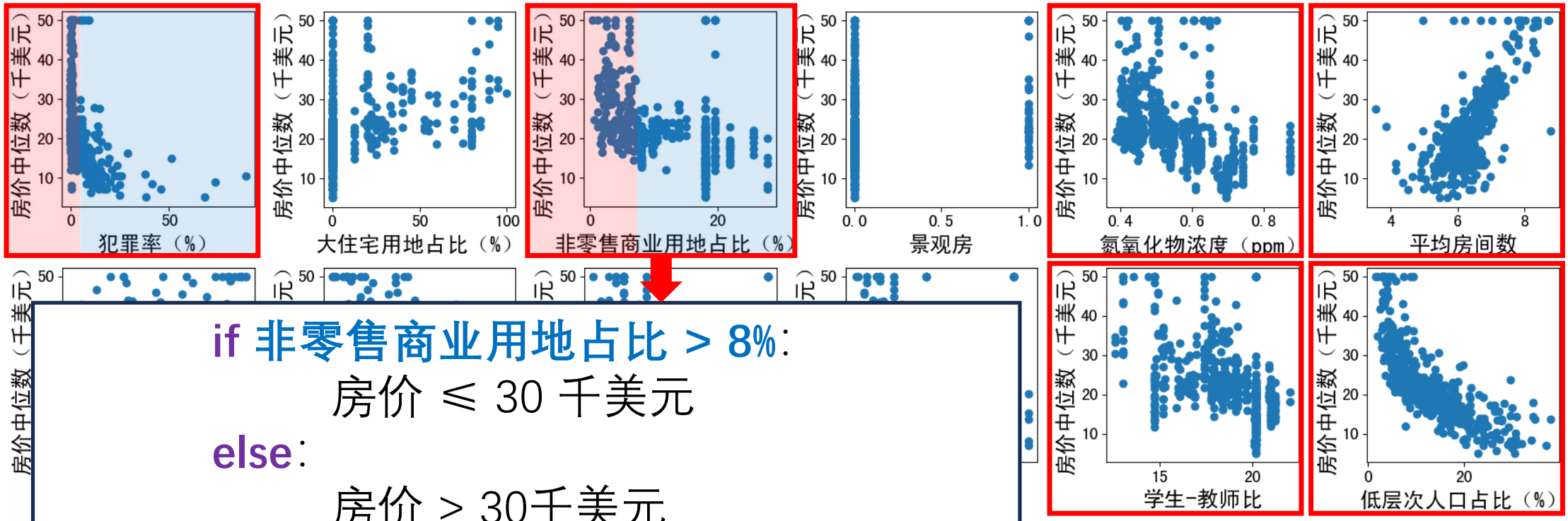
房价 ⩽ 30 千美元

else:

房价 > 30千美元

波士顿房价预测问题——If…else…方法

步骤2：分析数据，发现规律，构建规则（人工）



if 非零售商业用地占比 > 8%:

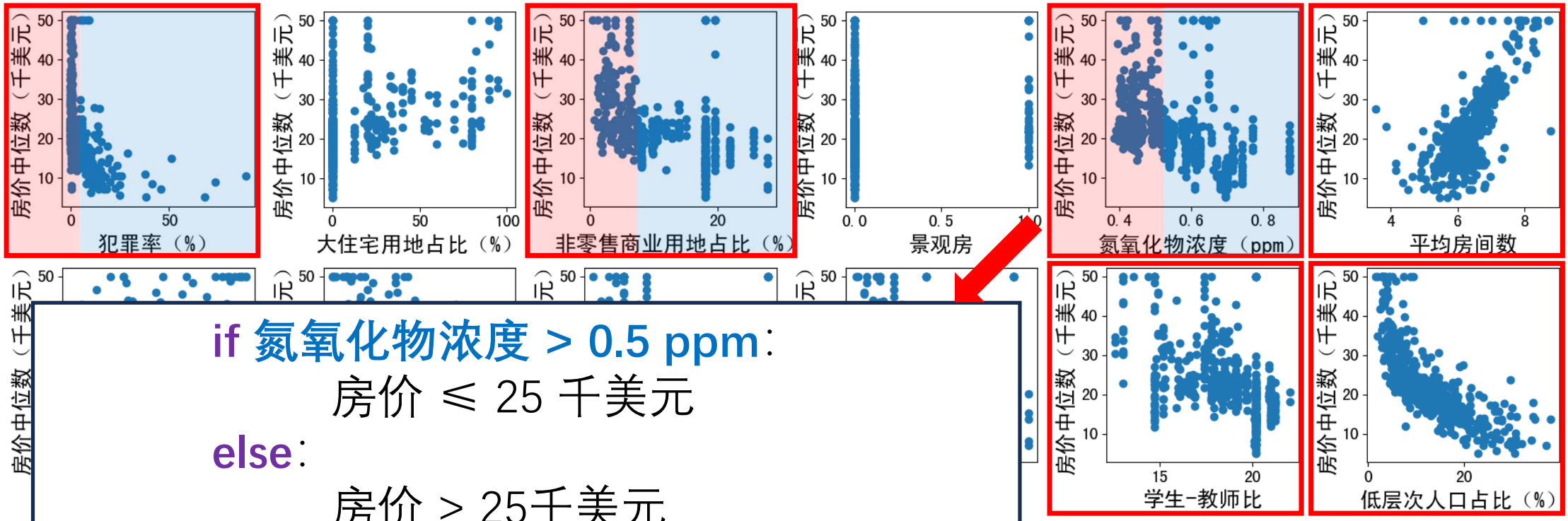
房价 ⩽ 30 千美元

else:

房价 > 30千美元

波士顿房价预测问题——If…else…方法

步骤2：分析数据，发现规律，构建规则（人工）



if 氮氧化物浓度 > 0.5 ppm:

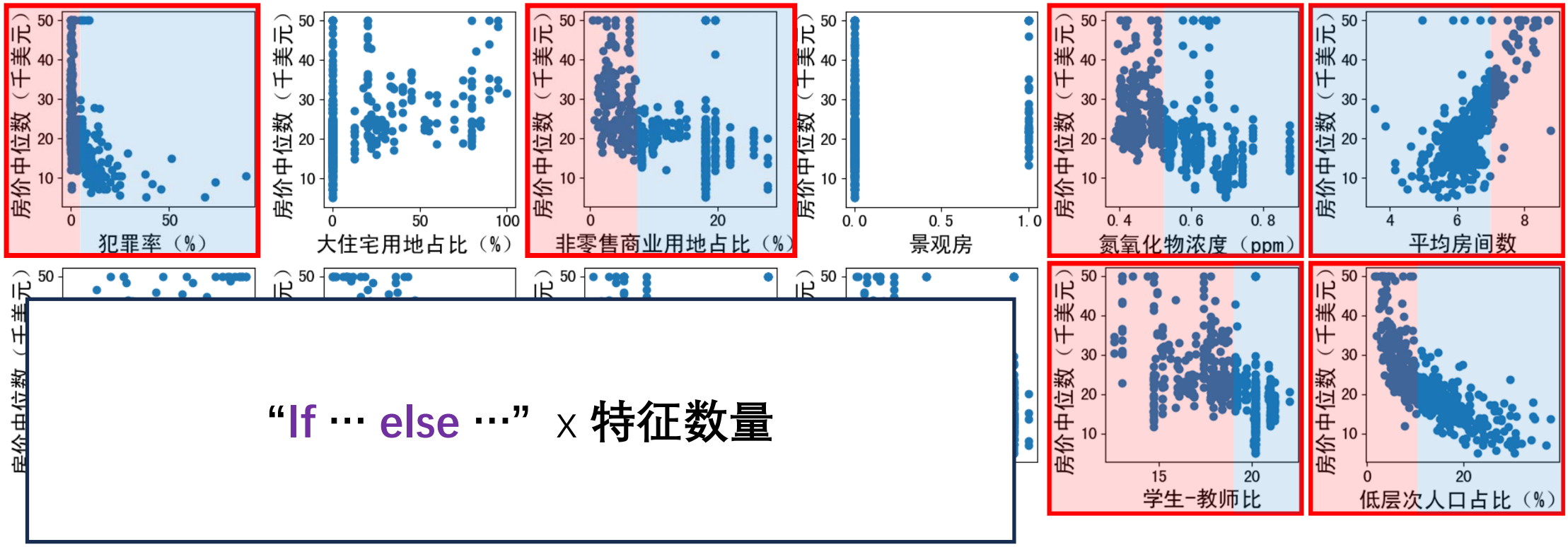
房价 ⩽ 25 千美元

else:

房价 > 25千美元

波士顿房价预测问题——If…else…方法

步骤2：分析数据，发现规律，构建规则（人工）



“If … else …” x 特征数量

波士顿房价预测问题——If…else…方法

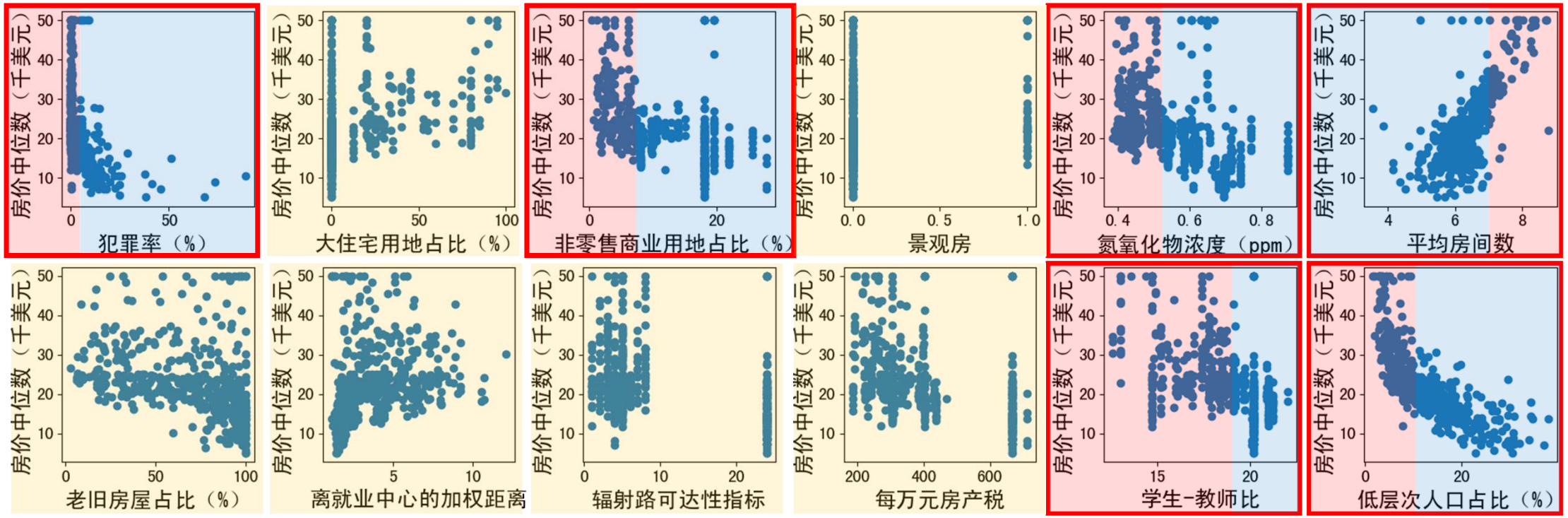
步骤3：根据专家知识，编写推理引擎

if 犯罪率 < 0.05% & 非零售商业用地 < 2% & 氮氧化物浓度 < 0.4 ppm & …:
    房价 > 50 千美元
else if 0.05% < 犯罪率 < 0.1% & 非零售商业用地 < 2% & 氮氧化物浓度 < 0.4 ppm & … :
    50 千美元 ≥ 房价 > 45 千美元
else if 0.05% < 犯罪率 < 0.1% & 2% < 非零售商业用地 < 4% & 0.4 ppm < 氮氧化物浓度 < 0.5 ppm & … :
    45 千美元 ≥ 房价 > 40 千美元
else if 0.1% < 犯罪率 < 0.2% & 4% < 非零售商业用地 < 8% & 0.4 ppm < 氮氧化物浓度 < 0.5 ppm & … :
    40 千美元 ≥ 房价 > 35 千美元
else if 0.1% < 犯罪率 < 0.2% & 4% < 非零售商业用地 < 8% & 0.5 ppm < 氮氧化物浓度 < 0.6 ppm & … :
    35 千美元 ≥ 房价 > 30 千美元
……

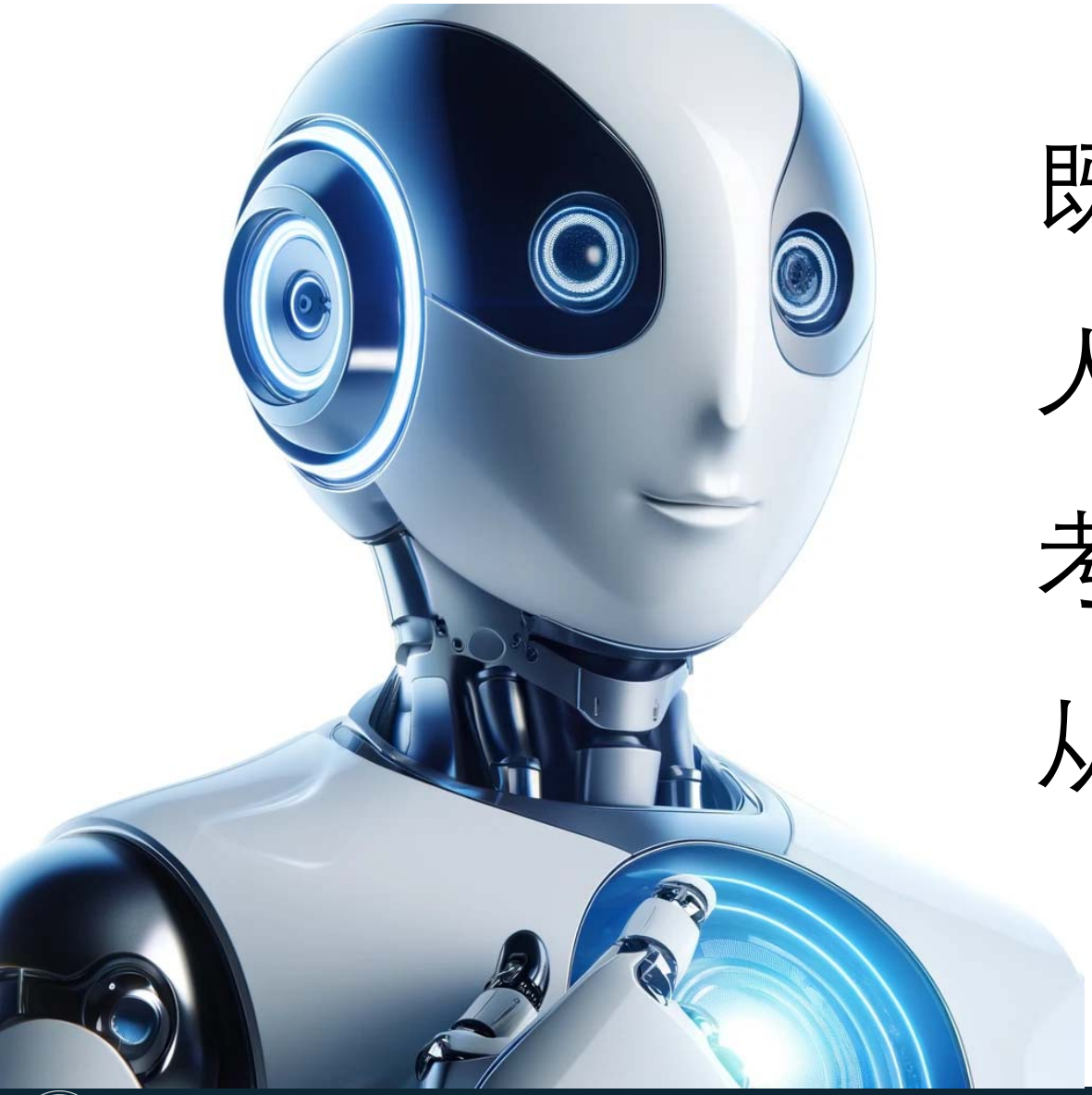
容易理解，但是极其复杂，难以用于其它的情况（难泛化）

波士顿房价预测问题——If…else…方法

步骤2：分析数据，发现规律，构建规则（人工）



容易忽视其它特征的影响



既然这么繁琐复杂，

人类，要不……

考虑一下让机器

从数据中学习规律

波士顿房价预测问题——机器学习方法

一大堆数据

- 输入：房价特征
- 输出：房价（标签）
- 定义问题形式
- 阐述正确的结果

x：平均房间数（个）	y：房价中位数（千美元）
6.575	24.00
6.421	21.60
7.185	34.70
…	…

波士顿房价预测问题——机器学习方法

一大堆数据

- 输入：房价特征
- 输出：房价（标签）
- 定义问题形式
- 阐述正确的结果

一堆模型
$f_1, f_2, \cdots, f_n$

- $y = \Sigma\alpha x + \beta$，$x$是特征值
- f由参数α和β定义

  $f_1 \rightarrow y = 3.5x + 2$

  $f_2 \rightarrow y = 1.5x + 3$
- f有无数的可能

波士顿房价预测问题——机器学习方法

- 输入：房价特征
- 输出：房价（标签）
- 定义问题形式
- 阐述正确的结果

平均房间数	6.575	6.421	7.185
房价（千美元）	**24**	**21.6**	**34.7**

✓	25.0	24.5	27.1
✗	12.9	13.6	13.8

- $y = \Sigma\alpha x + \beta$，$x$是特征值
- f由参数α和β定义

$f_1 \rightarrow y = 3.5x + 2$

$f_2 \rightarrow y = 1.5x + 3$

- f有无数的可能

波士顿房价预测问题——机器学习方法

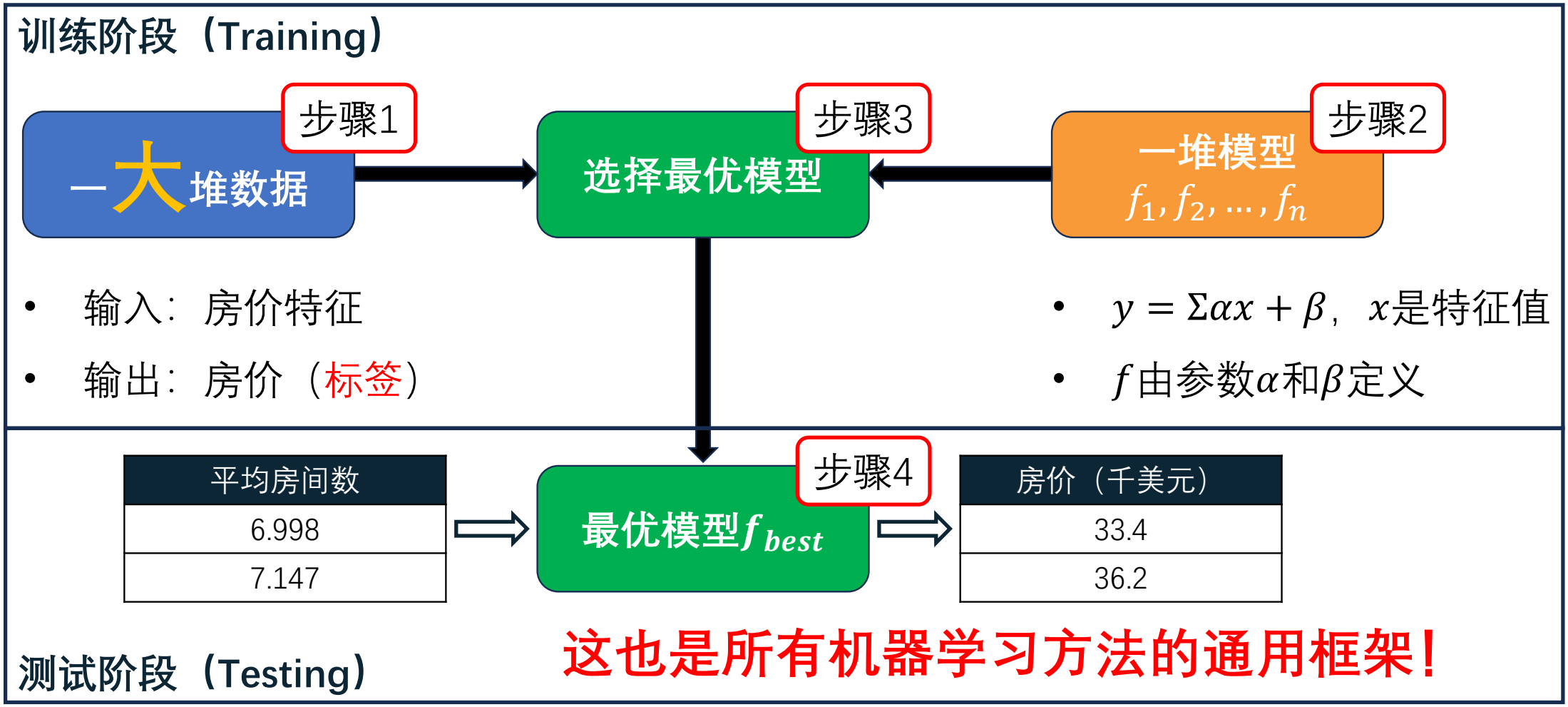
- 输入：房价特征
- 输出：房价（标签）

- $y = \Sigma\alpha x + \beta$，$x$是特征值
- f由参数α和β定义

平均房间数
6.998
7.147

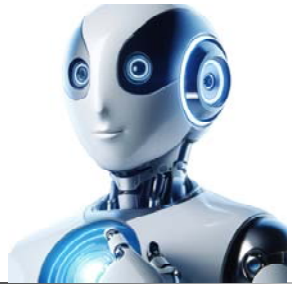
房价（千美元）
33.4
36.2

波士顿房价预测问题——机器学习方法



(专家系统)

IF-ELSE vs 机器学习

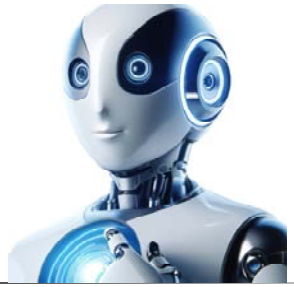


	(专家系统) IF-ELSE	机器学习
工作原理	由人类专家手工 编写规则	由算法自动学习 模式和关系
数据需求	依赖专家知识 数据需求低	依赖数据质量 数据需求高
优　势	可解释性强 高度定制化	专业知识需求低 泛化能力强
劣　势	知识获取难 规则维护难	对数据要求高 可解释性差

应该选哪一种方法呢?

（专家系统）

IF-ELSE vs 机器学习



工作原理	由人类专家手工 编写规则	由算法自动学习 模式和关系
数据需求	依赖专家知识 数据需求低	依赖数据质量 数据需求高

没有“一刀切”的方案，具体问题具体分析

抓住问题的关键，选择最合适的方法

课程小结

前情提要 —— 回顾机器学习的概念和典型问题

↓

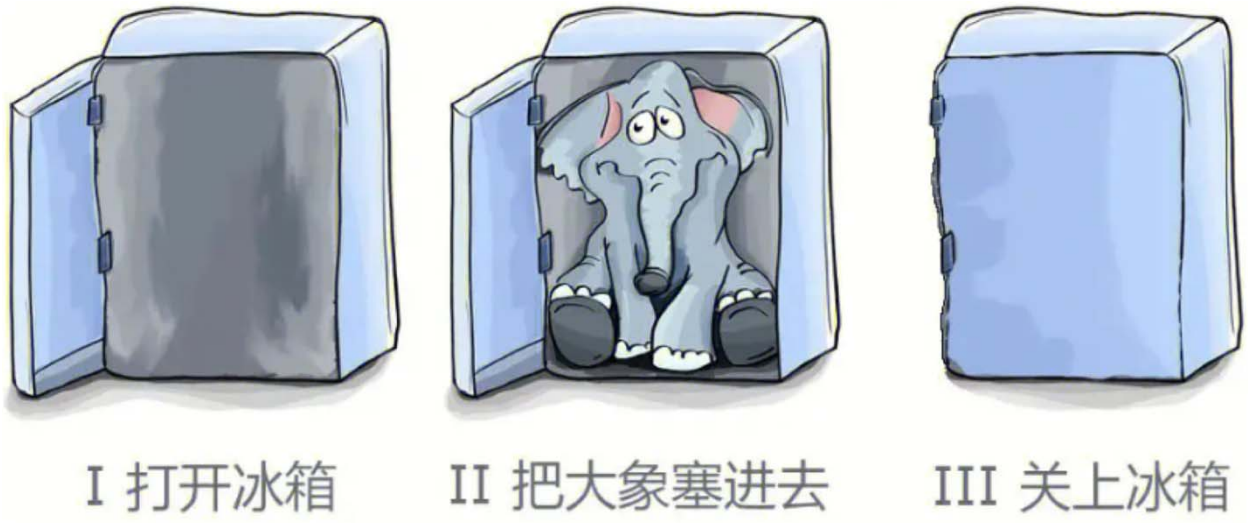
房价预测问题 —— **什么是回归问题?**

↓

波士顿房价案例 —— **如何用机器学习的思想解决回归问题?**

- 专家系统（以If…else…为例）
- 机器学习方法
- 比较两类方法的优劣

《钟点工》2000年春节联欢晚会



成功装箱还需对细节的把握……

思考：仅考虑单个特征，应如何设计预测模型？

如何选择最优模型？

实验2：基于线性回归的房价预测

简介：分析特征与房价的关系，预测房价

数据集：1978波士顿区域房价预测

实验内容应包括：

- 用**散点图**展示房屋特征和房价的关系
- 使用**线性模型**和至少一种**非线性模型**进行房价预测
- 单个特征条件下，使用**蛮力法**选择最优模型
- 所有特征条件下，使用**梯度下降**训练模型
- 分别使用**留出法**、**交叉验证法**和**Boostrapping**方法构建测试集
- 模拟**过拟合**和**欠拟合**的情况，并尝试解决

One more thing

可以尝试使用AI助教

- 回答课程知识点相关的问题
- 提供课程相关的参考材料及资源
- 课程考核材料模板：Notebook和实验报告
- Notebook模板在对应章节结束一周后公布
- 提供实验报告指导，检查内容完整性

机器学习助教巴鲁

https://www.coze.cn/s/i69KaRjd

Balu Machine Learning Teaching Assistant

https://www.coze.com/s/ZmFqEvRLs