

Python 3 玩儿转机器学习

讲师：liuyubobobo

版权所有 侵权必究
liuyubobobo

慕课网《Python3机器学习》

机器学习基础概念

讲师：liuyubobobo

版权所有，侵权必究

慕课网《Python3机器学习》

关于数据

讲师：liuyubobobo

版权所有，侵权必究

数据

- 著名的鸢尾花数据 https://en.wikipedia.org/wiki/Iris_flower_data_set



Iris setosa



Iris versicolor



Iris virginica

数据

Iris Plants Database

=====

Notes

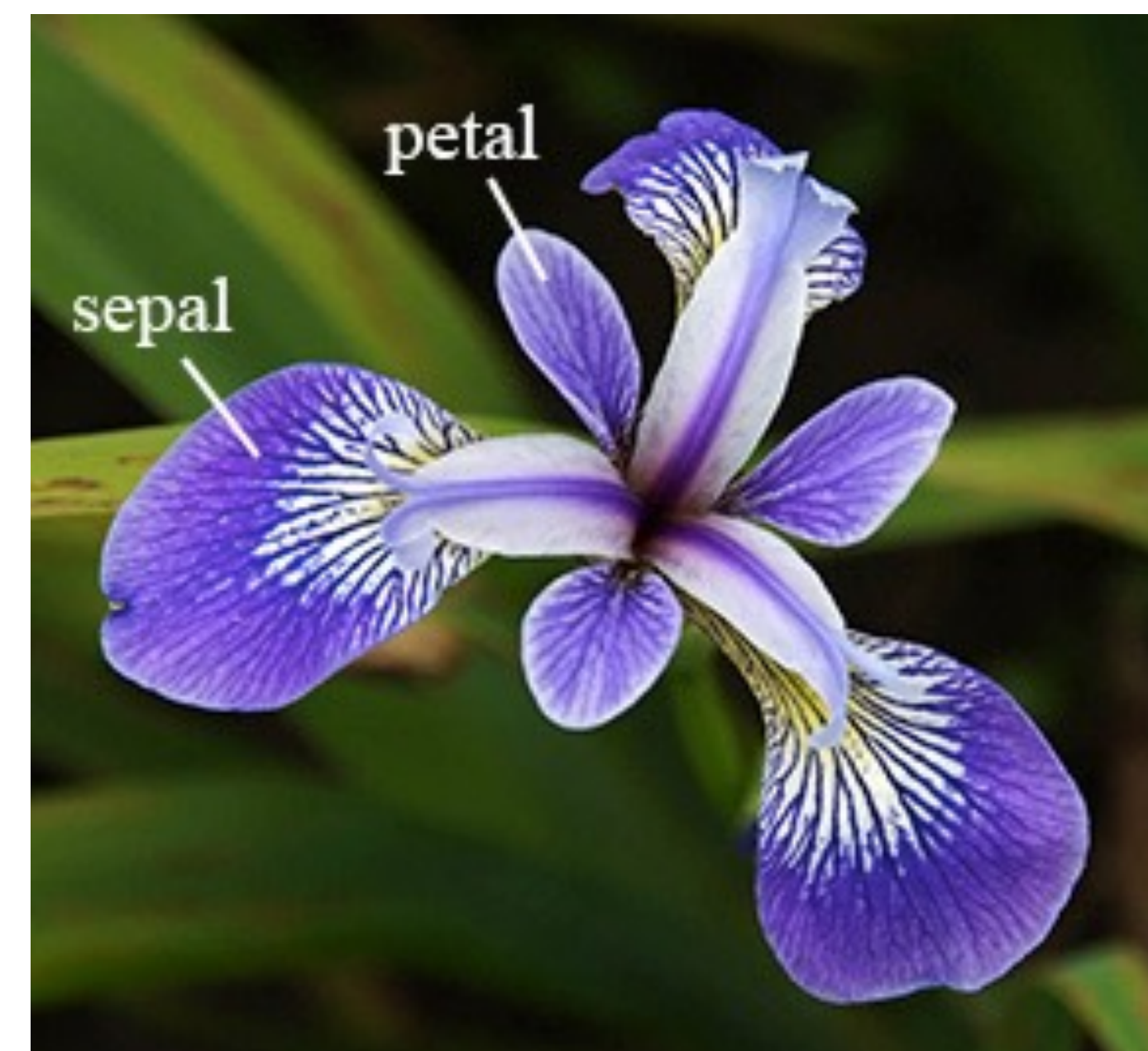
Data Set Characteristics:

:Number of Instances: 150 (50 in each of three classes)

:Number of Attributes: 4 numeric, predictive attributes and the class

:Attribute Information:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:
 - Iris-Setosa
 - Iris-Versicolour
 - Iris-Virginica



数据

萼片长度	萼片宽度	花瓣长度	花瓣宽度	种类
5.1	3.5	1.4	0.2	se (0)
7.0	3.2	4.7	1.4	ve (1)
6.3	3.3	6	2.5	vi (2)

数据

萼片长度	萼片宽度	花瓣长度	花瓣宽度	种类
------	------	------	------	----

5.1	3.5	1.4	0.2	se (0)
-----	-----	-----	-----	--------

7.0	3.2	4.7	1.4	ve (1)
-----	-----	-----	-----	--------

6.3	3.3	6	2.5	vi (2)
-----	-----	---	-----	--------

- 数据整体叫数据集 (data set)
- 每一行数据称为一个样本(sample)
- 除最后一列，每一列表达样本的一个特征(feature)
- 最后一列，称为标记(label)

X

y

第*i*个样本行写作 $X^{(i)}$ 第*i*个样本第*j*个特征值 $X_j^{(i)}$ 第*i*个样本的标记写作 $y^{(i)}$

数据

萼片长度	萼片宽度	花瓣长度	花瓣宽度
------	------	------	------

5.1	3.5	1.4	0.2
7.0	3.2	4.7	1.4

6.3	3.3	6	2.5
-----	-----	---	-----

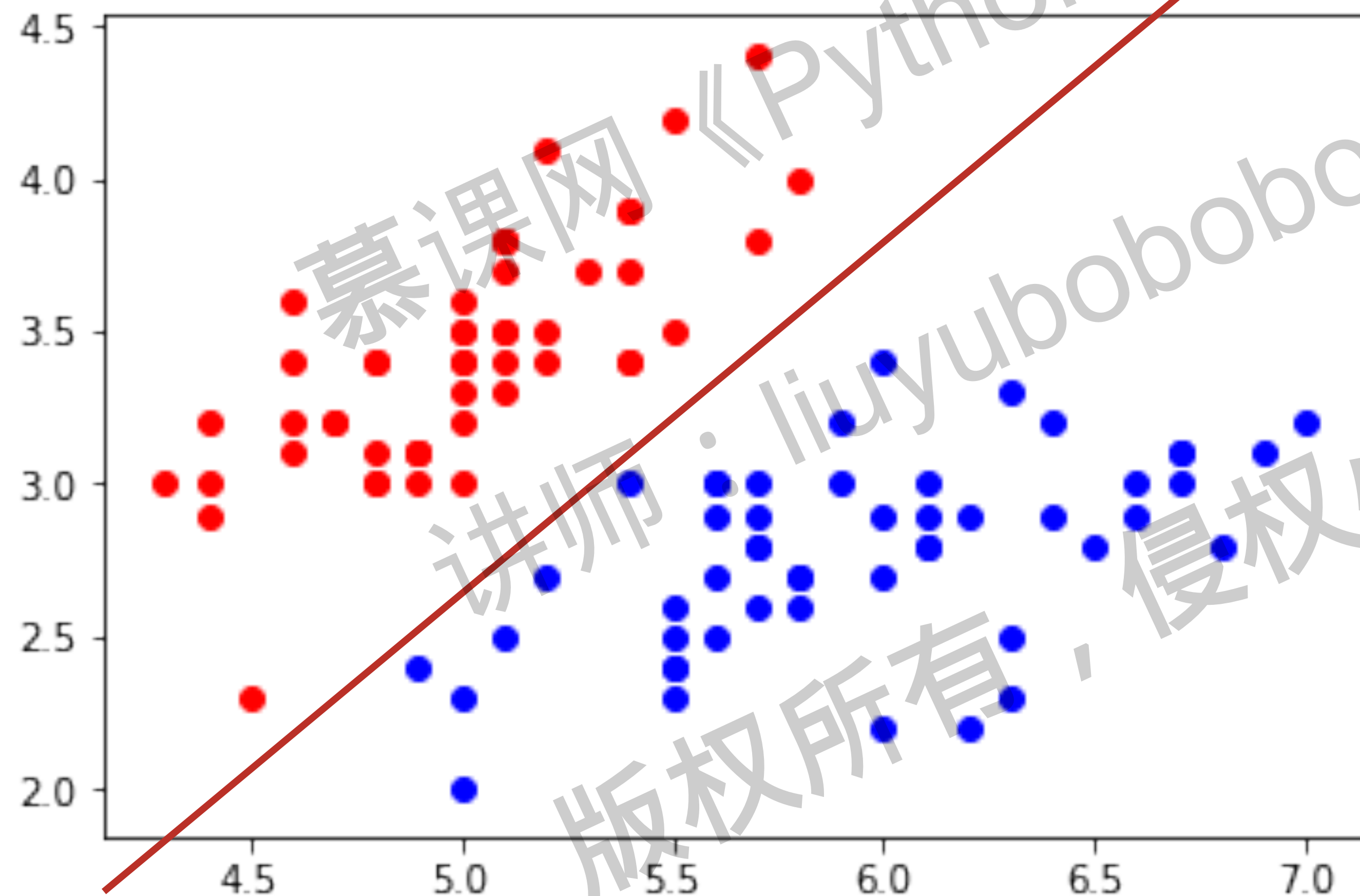
→ 特征

→ 特征向量 $X^{(i)}$

$$\begin{pmatrix} (X^{(1)})^T \\ (X^{(2)})^T \\ (X^{(3)})^T \\ \dots \end{pmatrix}$$

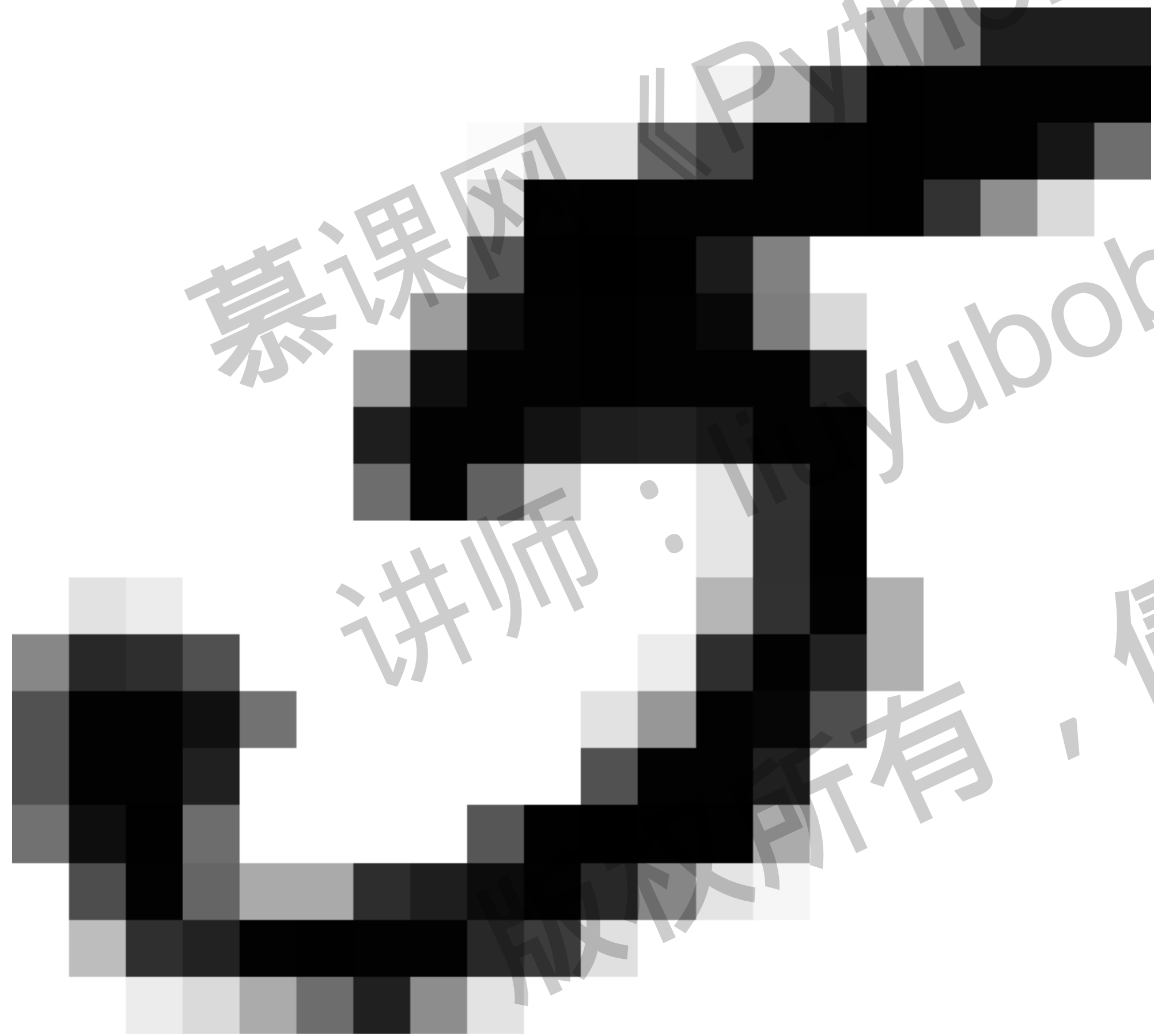
$$\begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}$$

数据



- 特征空间 (feature space)
- 分类任务本质就是在特征空间切分
- 在高维空间同理

特征可以很抽象



- 图像，每一个像素点都是特征
- 28×28 的图像有 $28 \times 28 = 784$ 个特征
- 如果是彩色图像特征更多

慕课网《Python3机器学习》

机器学习的基本任务

讲师：liuyubobobo

版权所有，侵权必究

机器学习的基本任务

- 分类

- 回归

分类任务



分类任务



分类任务

- 二分类



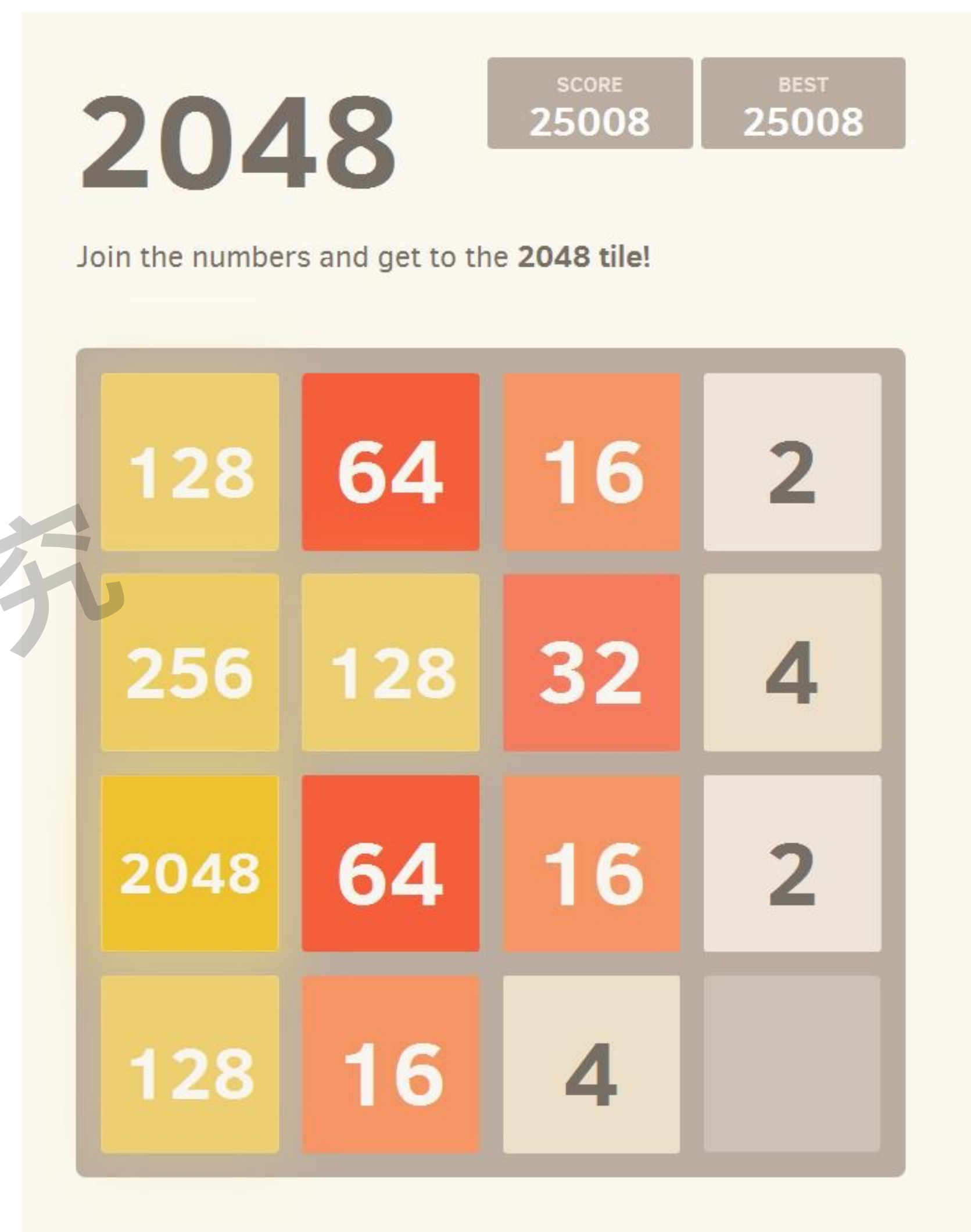
- 判断邮件是垃圾邮件； 不是垃圾邮件
- 判断发放给客户信用卡有风险； 没有风险
- 判断病患良性肿瘤； 恶性肿瘤
- 判断某支股票涨； 跌

分类任务

- 多分类
- 数字识别
- 图像识别
- 判断发放给客户信用卡的风险评级

分类任务

- 多分类
- 很多复杂的问题也可以转换成多分类问题



分类任务

- 多分类



分类任务

- 多分类



分类任务

- 多分类
 - 一些算法只支持完成二分类的任务
 - 但是多分类的任务可以转换成二分类的任务
 - 有一些算法天然可以完成多分类任务

分类任务

- 多标签分类



另一类数据

房屋面积 (平方米)	房屋年龄 (年)	卧室数量 (间)	最近地铁站距离 (千米)	价格 (万元)
80	3	1	10	300
120	8	3	5	500
200	5	4	12	700

回归任务

- 结果是一个连续数字的值，而非一个类别
 - 房屋价格
 - 市场分析
 - 学生成绩
 - 股票价格

回归任务

- 结果是一个连续数字的值，而非一个类别
- 有一些算法只能解决回归问题
- 有一些算法只能解决分类问题
- 有一些算法的思路既能解决回归问题，又能解决分类问题

回归任务

- 一些情况下，回归任务可以简化成分类任务

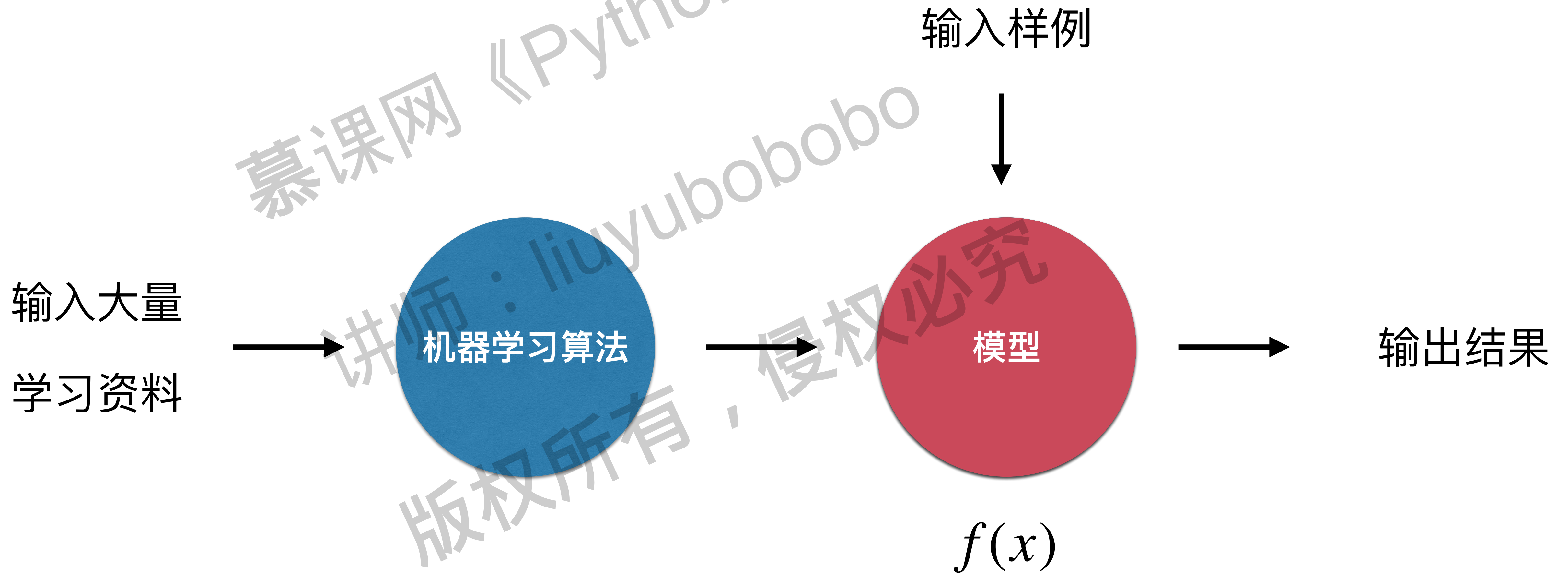


机器学习的基本任务

- 分类

- 回归

什么是机器学习



监督学习

- 分类

- 回归

慕课网《Python3机器学习》

监督学习，非监督学习 半监督学习和增强学习

讲师：nuyubobobo
版权所有，侵权必究

机器学习方法的分类

- 监督学习
- 非监督学习
- 半监督学习
- 增强学习

监督学习

给机器的训练数据拥有“标记”或者“答案”

狗



猫



监督学习



MNIST数据集

监督学习

- 图像已经拥有了标定信息
- 银行已经积累了一定的客户信息和他们信用卡的信用情况
- 医院已经积累了一定的病人信息和他们最终确诊是否患病的情况
- 市场积累了房屋的基本信息和最终成交的金额
-

监督学习

- 分类

- 回归

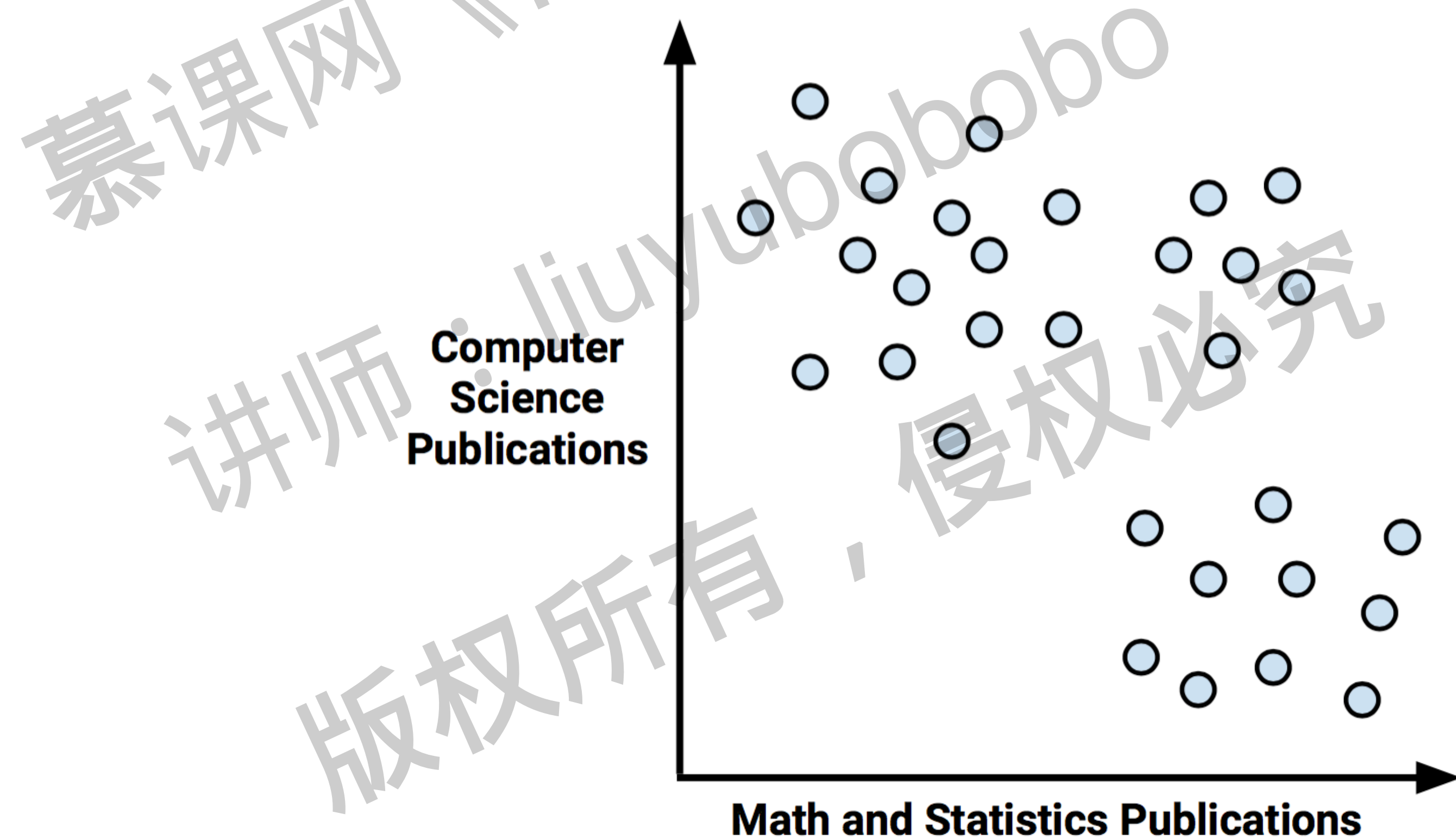
监督学习

我们在这个课程中学习的大部分算法，属于监督学习算法

- k近邻
- 线性回归和多项式回归
- 逻辑回归
- SVM
- 决策树和随机森林

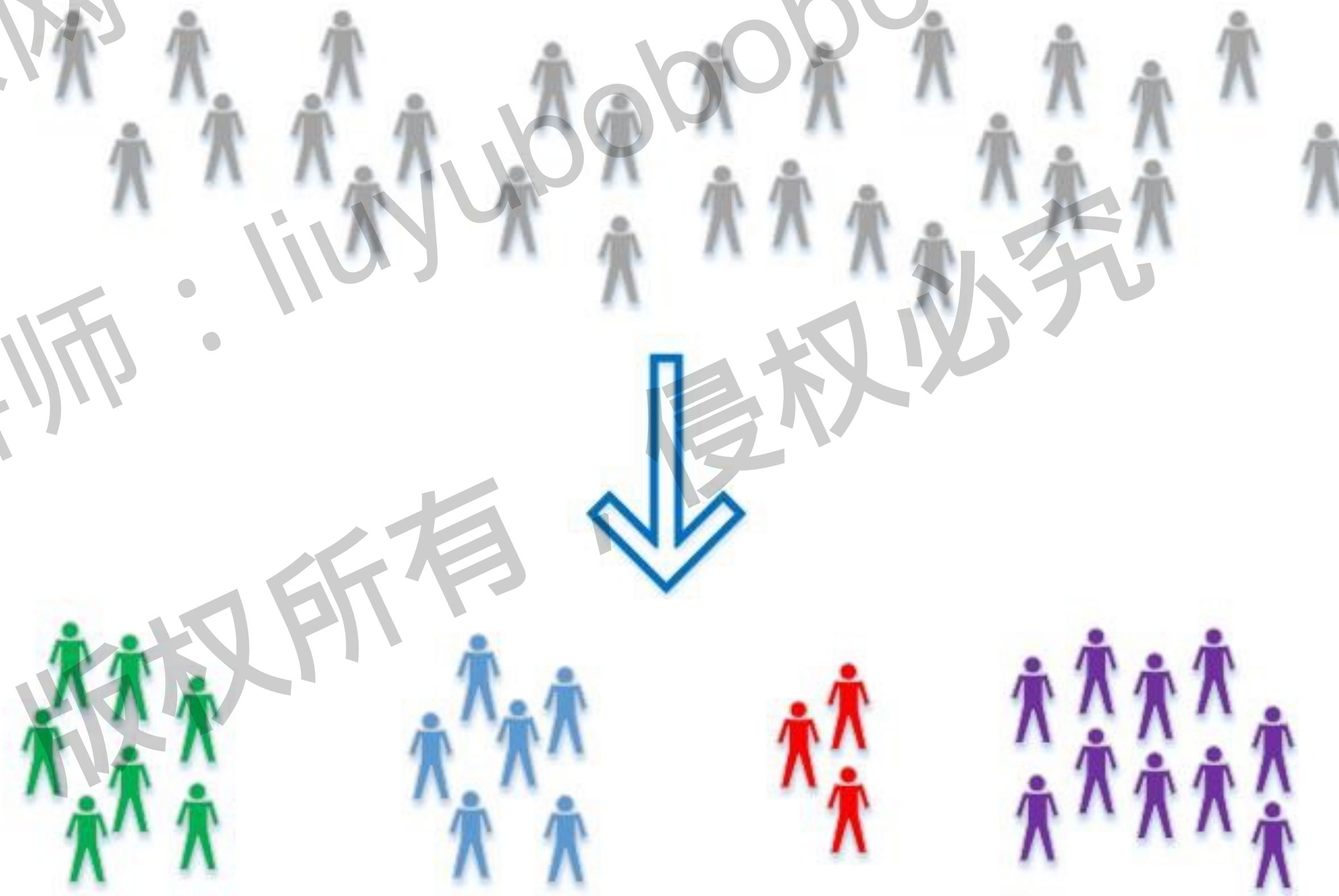
非监督学习

给机器的训练数据没有任何“标记”或者“答案”



非监督学习的意义

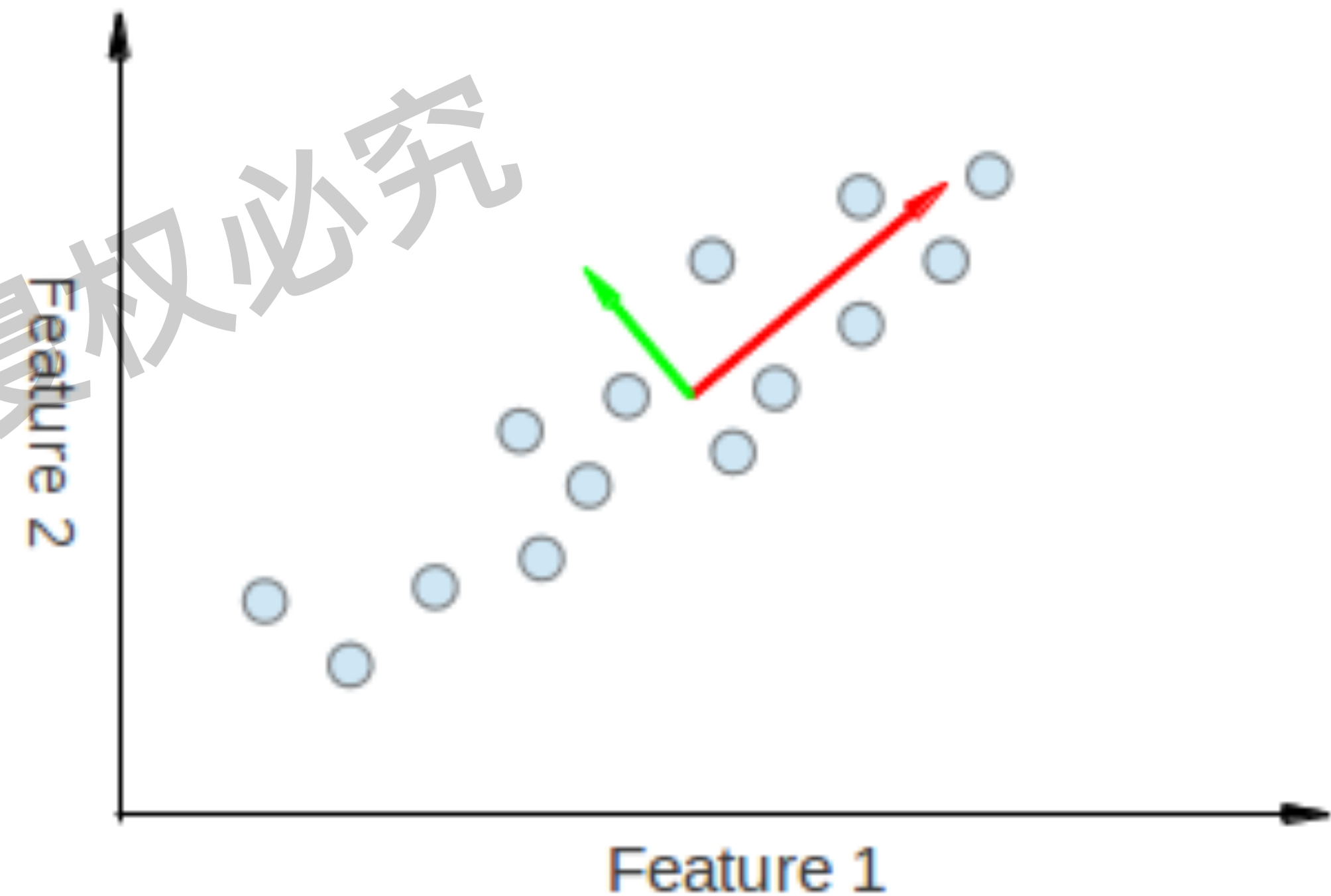
对没有“标记”的数据进行分类 - 聚类分析



非监督学习的意义

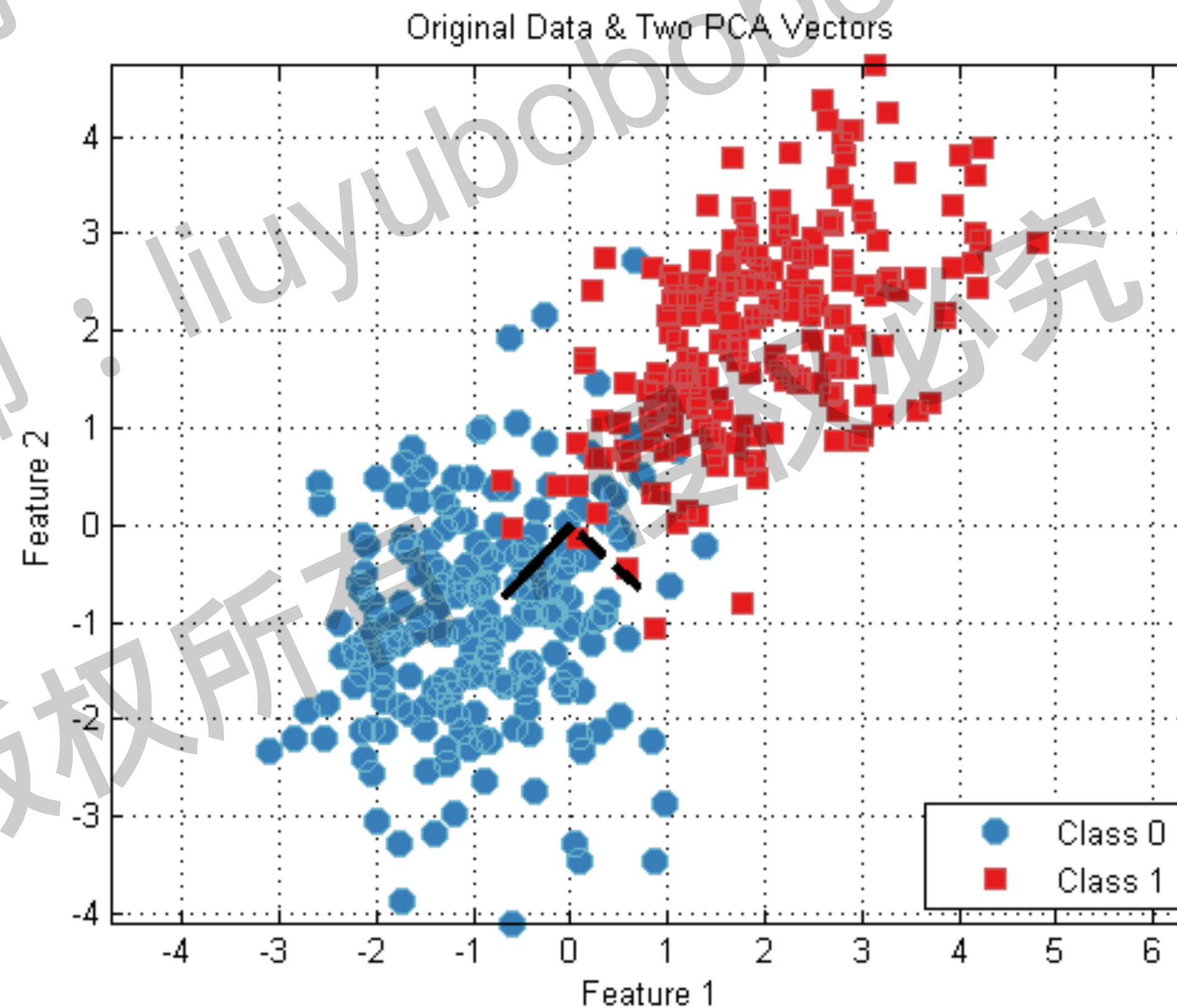
对数据进行降维处理

- 特征提取：信用卡的信用评级和人的胖瘦无关？
- 特征压缩：PCA



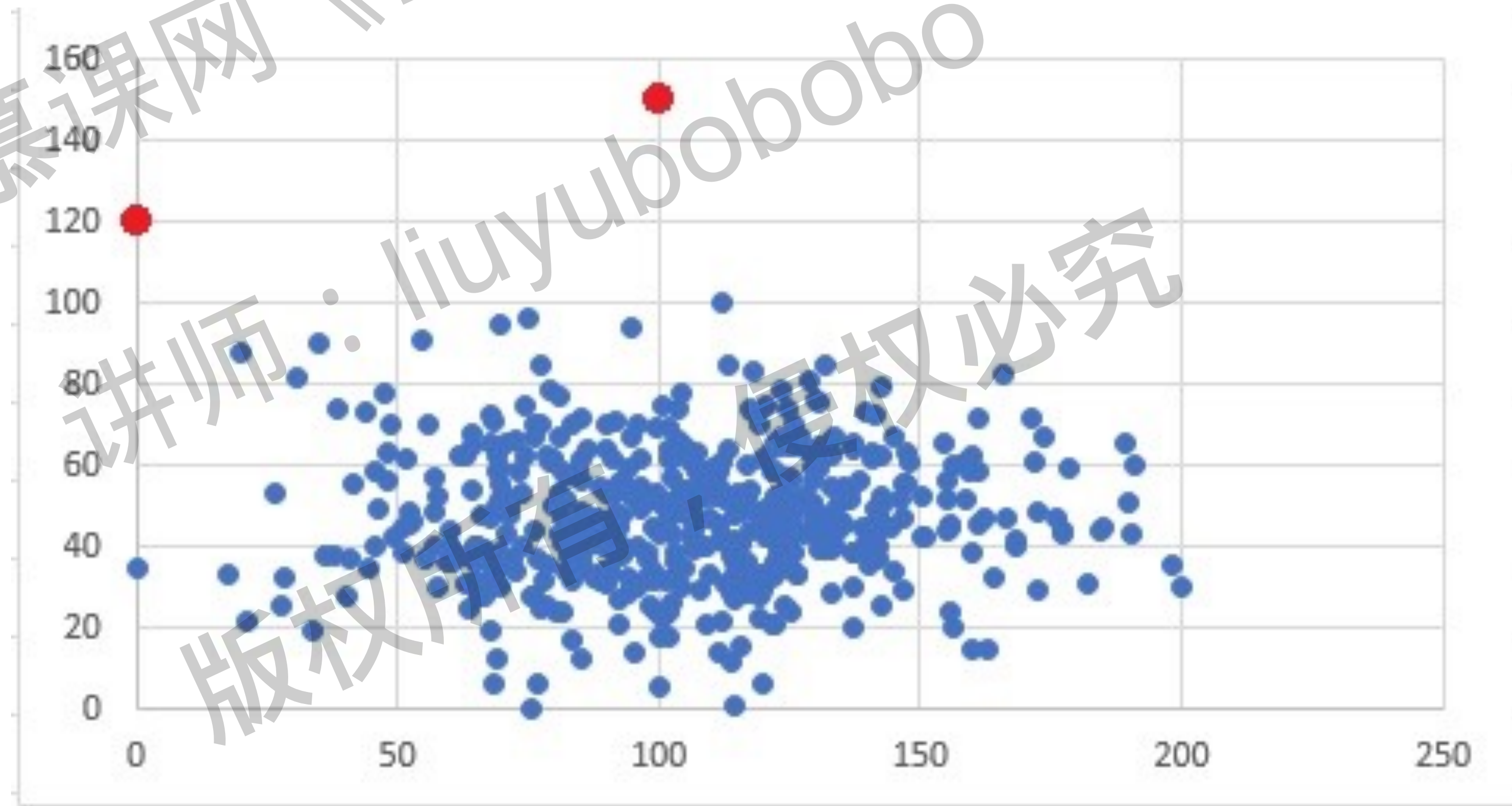
非监督学习的意义

降维处理的意义：方便可视化



非监督学习的意义

异常检测



半监督学习

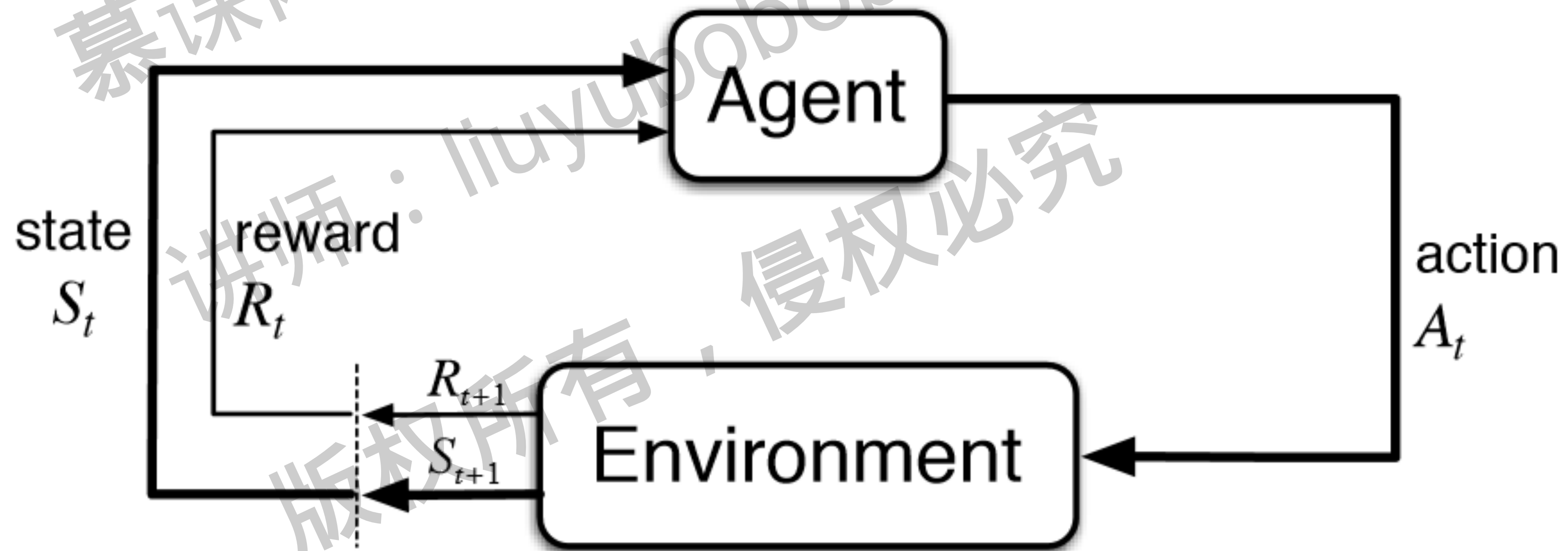
一部分数据有“标记”或者“答案”，另一部分数据没有

更常见：各种原因产生的标记缺失

通常都先使用无监督学习手段对数据做处理，之后使用监督学习手段做模型的训练和预测

增强学习

根据周围环境的情况，采取行动，根据采取行动的结果，学习行动方式。



增强学习

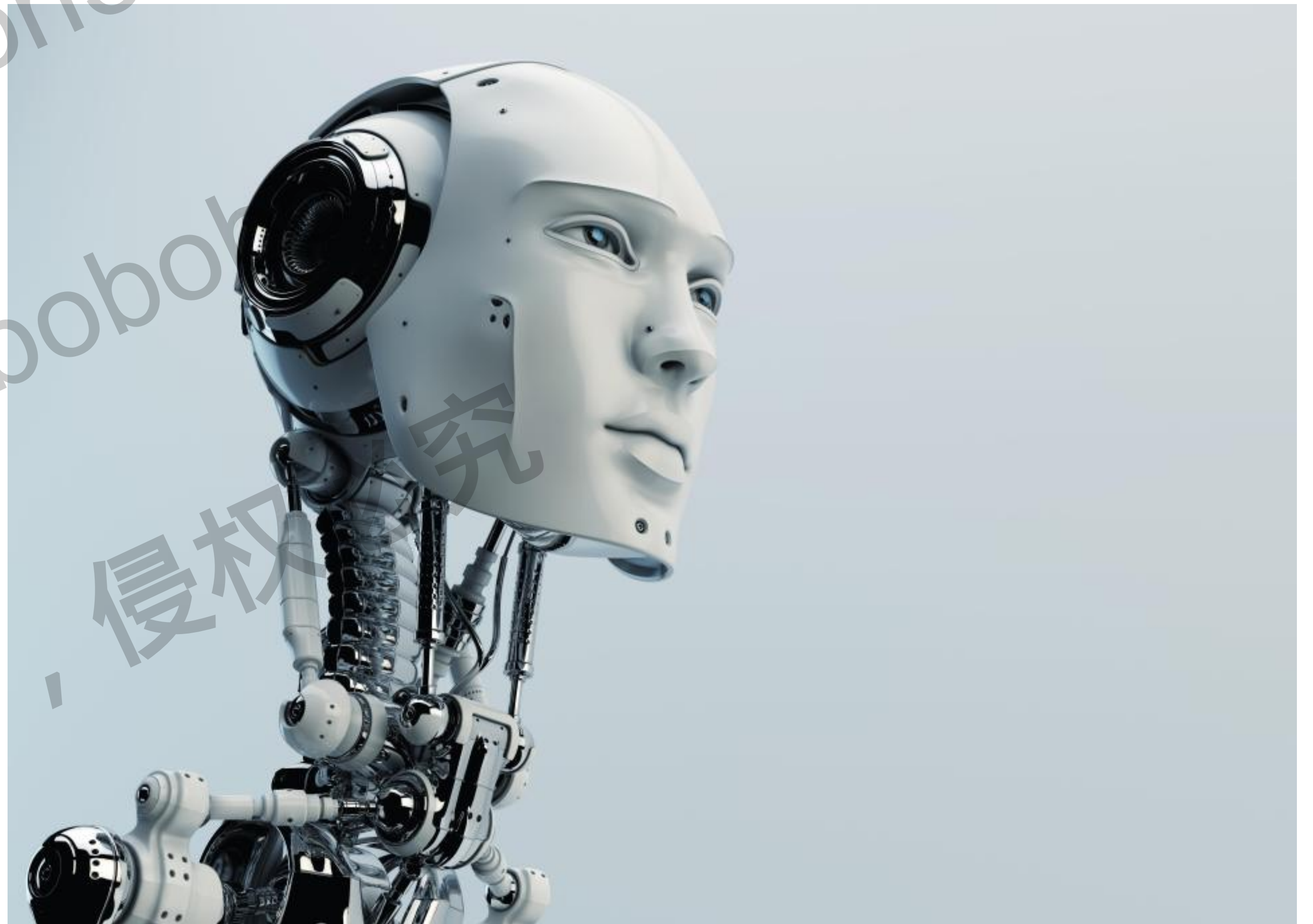


增强学习

无人驾驶

机器人

监督学习和半监督学习是基础



机器学习方法的分类

- 监督学习
- 非监督学习
- 半监督学习
- 增强学习

慕课网《Python3机器学习》

机器学习的其他分类

讲师：liuyuboboo

在线学习和批量学习（离线学习）

参数学习和非参数学习

版权所有，侵权必究

批量学习和在线学习

- 批量学习 Batch Learning
- 在线学习 Online Learning

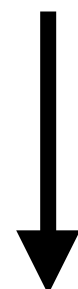
批量学习

输入大量
学习资料



输出结果

输入样例



批量学习

- 优点：简单

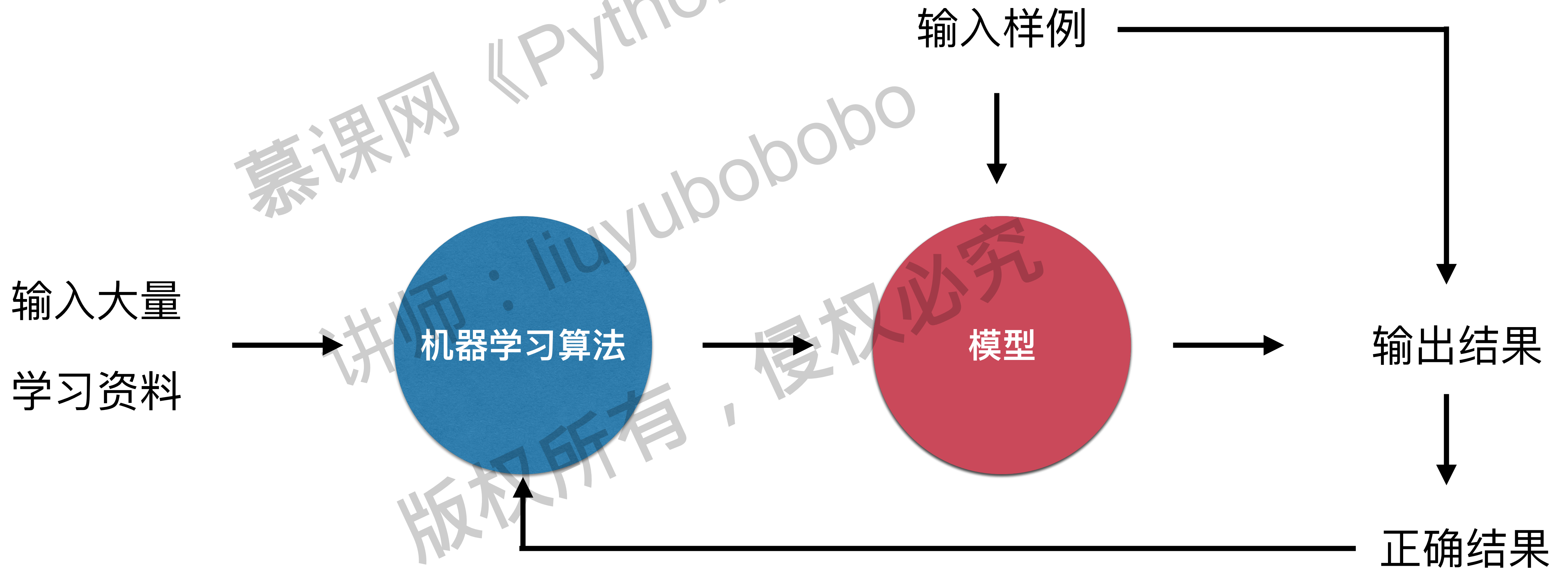
- 问题：如何适应环境变化？

解决方案：定时重新批量学习

- 缺点：每次重新批量学习，运算量巨大；

在某些环境变化非常快的情况下，甚至不可能的。

在线学习



在线学习

- 优点：及时反映新的环境变化

- 问题：新的数据带来不好的变化？

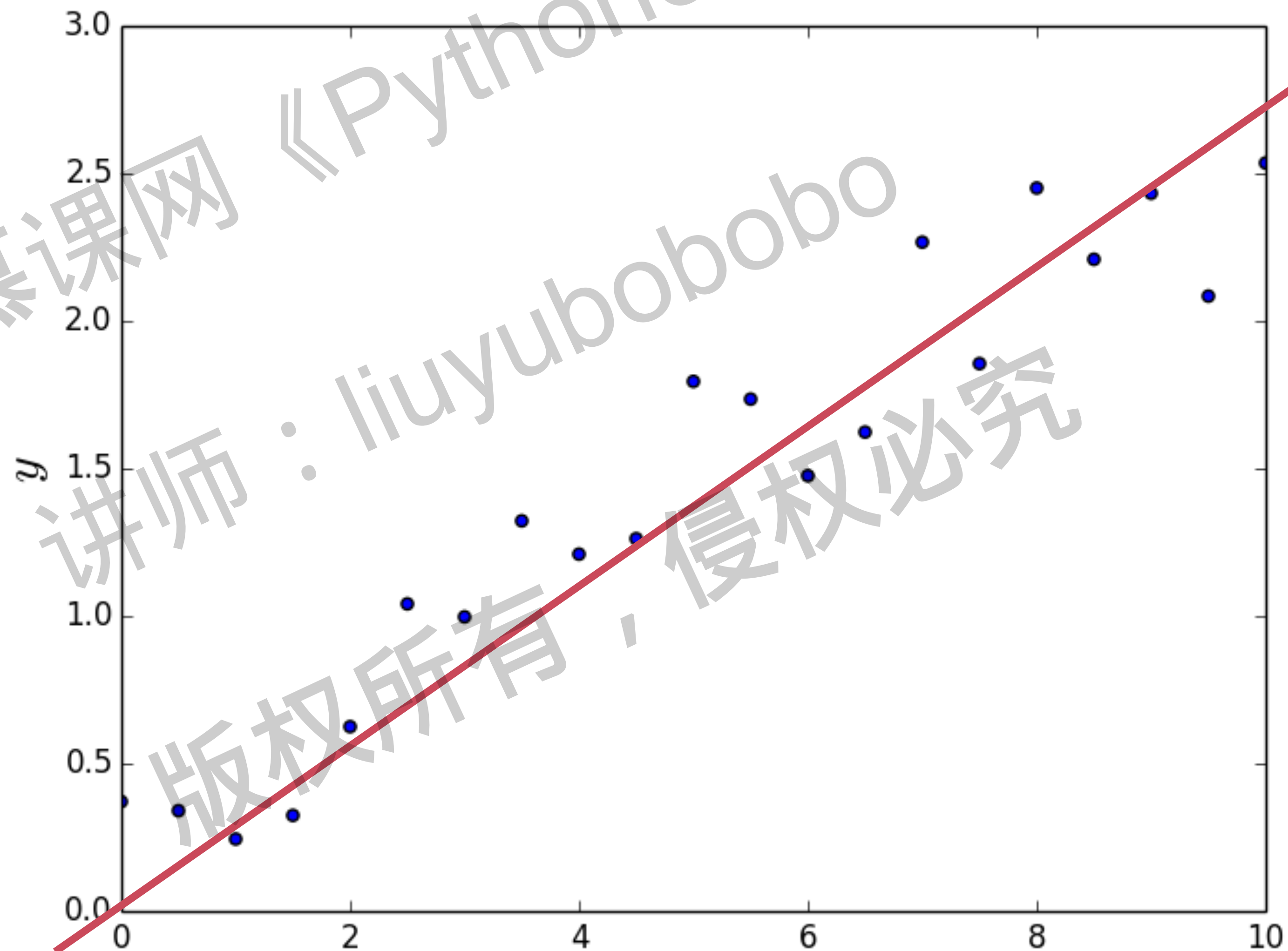
解决方案：需要加强对数据进行监控

- 其他：也适用于数据量巨大，完全无法批量学习的环境。

参数学习和非参数学习

- 参数学习 Parametric Learning
- 非参数学习 Nonparametric Learning

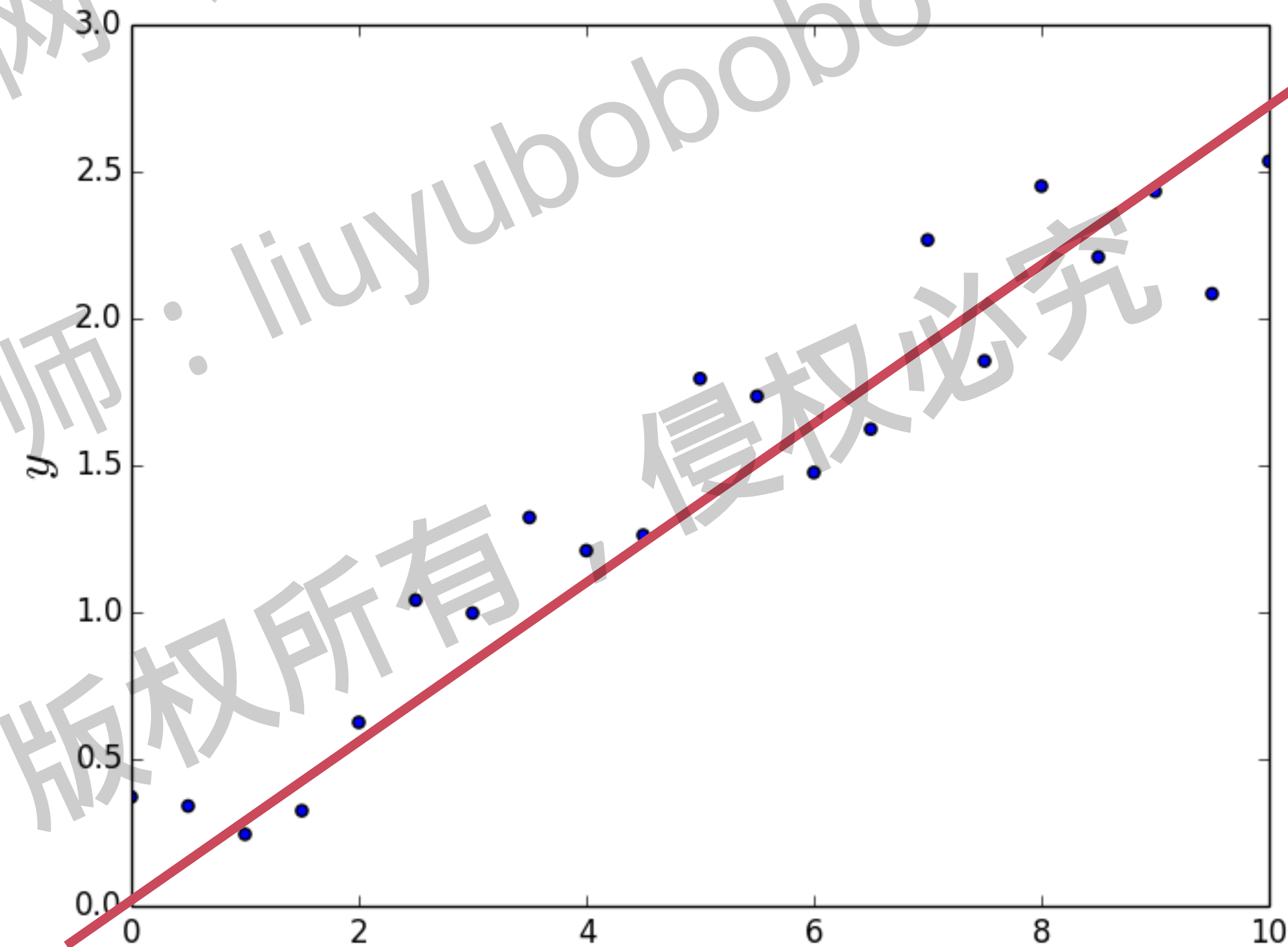
参数学习



$$f(x) = a * x + b$$

参数学习的特点

一旦学到了参数，就不再需要原有的数据集



非参数学习

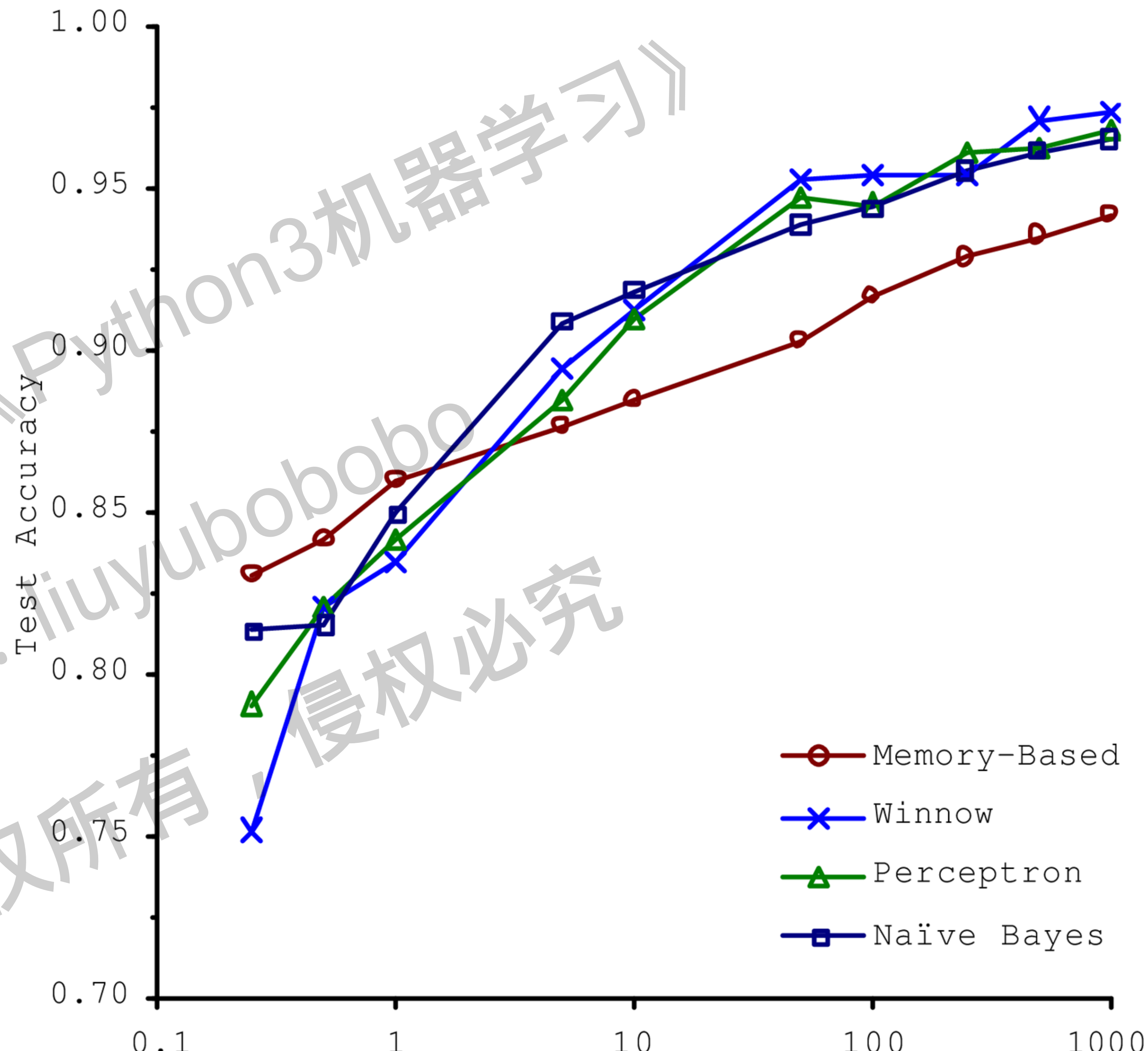
- 不对模型进行过多假设
- 非参数不等于没参数！

和机器学习相关的“哲学”思考

讲师：liuyubobobo

版权所有，侵权必究

2001年，
微软的论文



慕课网《Python3机器学习》

数据即算法？

讲师：liuyubobobo

版权所有，侵权必究

数据即算法?

- 数据确实非常重要
- 数据驱动
- 收集更多的数据
- 提高数据质量
- 提高数据的代表性
- 研究更重要的特征

算法为王?

AlphaGo Zero
Starting from scratch



如何选择机器学习算法?

kNN

线性回归

多项式回归

逻辑回归

模型正则化

PCA

SVM

决策树

随机森林

集成学习

模型选择

模型调试

奥卡姆的剃刀

- 简单的就是好的
- 到底在机器学习领域，什么叫“简单”？

没有免费的午餐定理

- 可以严格地数学推导出：任意两个算法，他们的期望性能是相同的！
- 具体到某个特定问题，有些算法可能更好
- 但没有一种算法，绝对比另一种算法好

没有免费的午餐定理

- 脱离具体问题，谈那个算法好是没有意义的。
- 在面对一个具体问题的时候，尝试使用多种算法进行对比试验，是必要的。

其他思考

面对不确定的世界，怎么看待使用机器学习进行预测的结果？

慕课网《Python3机器学习》

课程环境搭建

讲师：liuyubob6699

版权所有，侵权必究

慕课网《Python3机器学习》



ANACONDA®

讲师: liuyubobobo
版权所有, 侵权必究



课程github

github.com/liuyubobobo/Play-with-Machine-Learning-Algorithms

其他

请大家善于使用慕课网的课程问答区

其他

欢迎大家关注我的个人公众号：是不是很酷



Python 3 玩儿转机器学习

讲师：liuyubobobo

版权所有 侵权必究
liuyubobobo