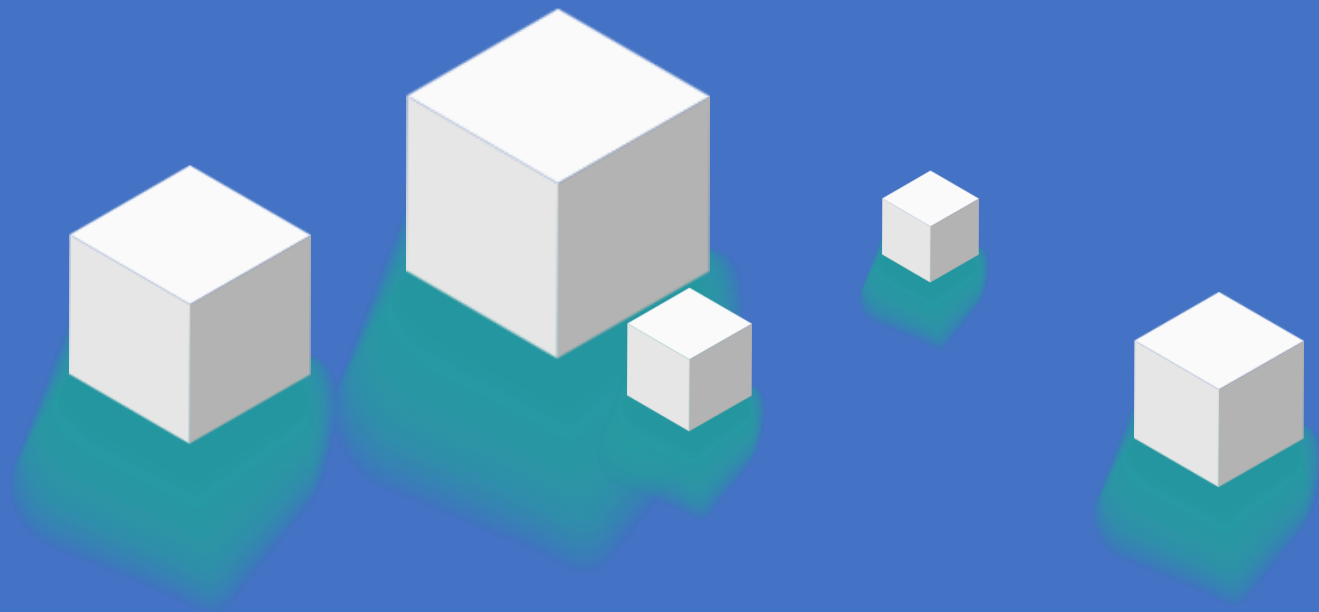
机器学习七步法



Thinking： 什么是人工智能?

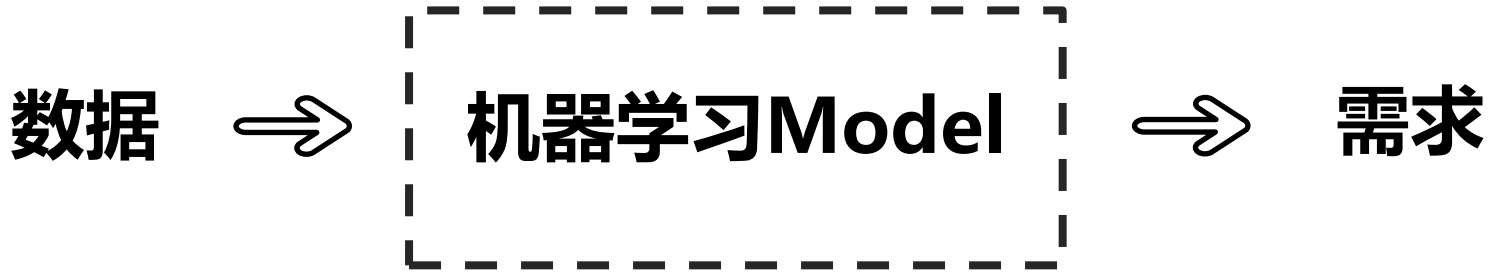
The theory and development of computer systems able to perform tasks normally requiring human intelligence.

——— *Oxford Dictionary*

Using data to solve problems.

—— *cy*

Using data to solve problems

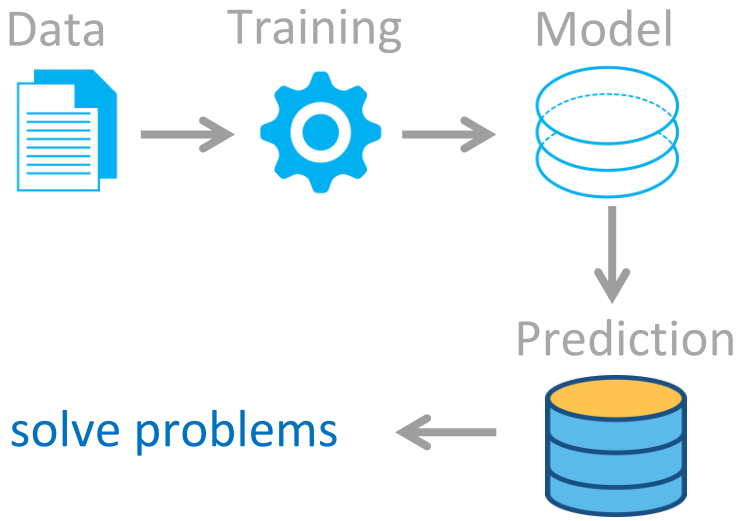


AI的本质

AI就是利用数据，解决问题

Using data to solve problems

Training	Prediction
Using data	solve problems



训练阶段：通过对数据的训练，创建一个预测模型并对其进行微调。

模型生成：预测模型可以从这些数据背后找出答案来，帮我们解决某个问题。

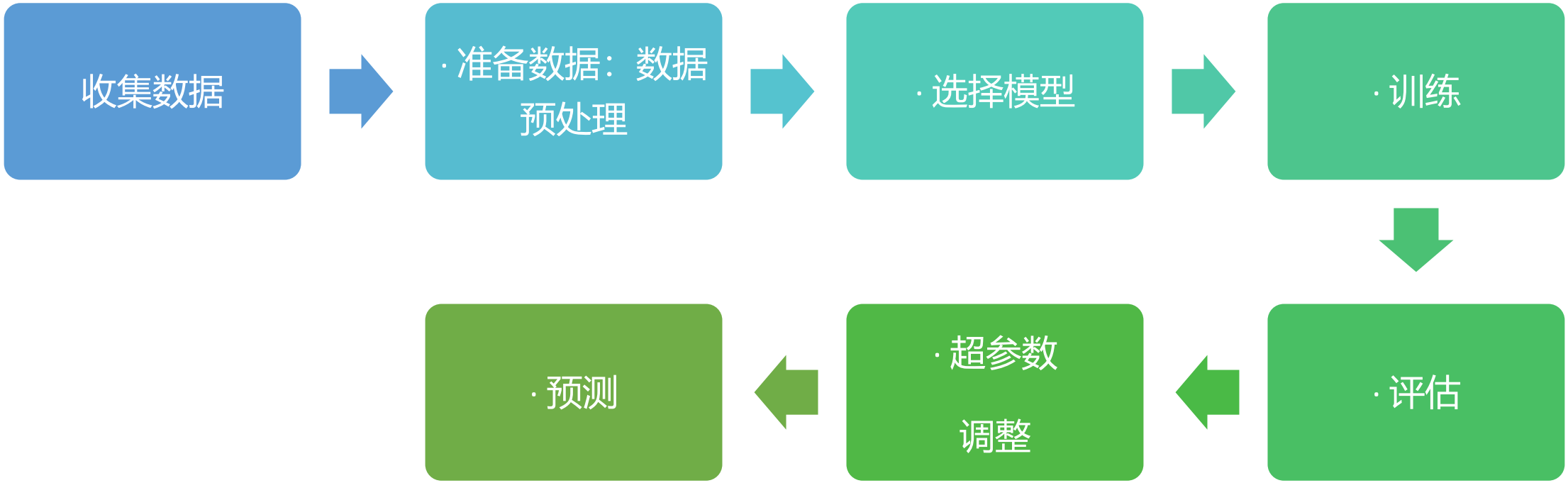
预测阶段：通过测试集完成模型评估，从而了解模型在测试集中的有效性。

过程中，预测模型会被不断改进和使用。

机器学习的步骤

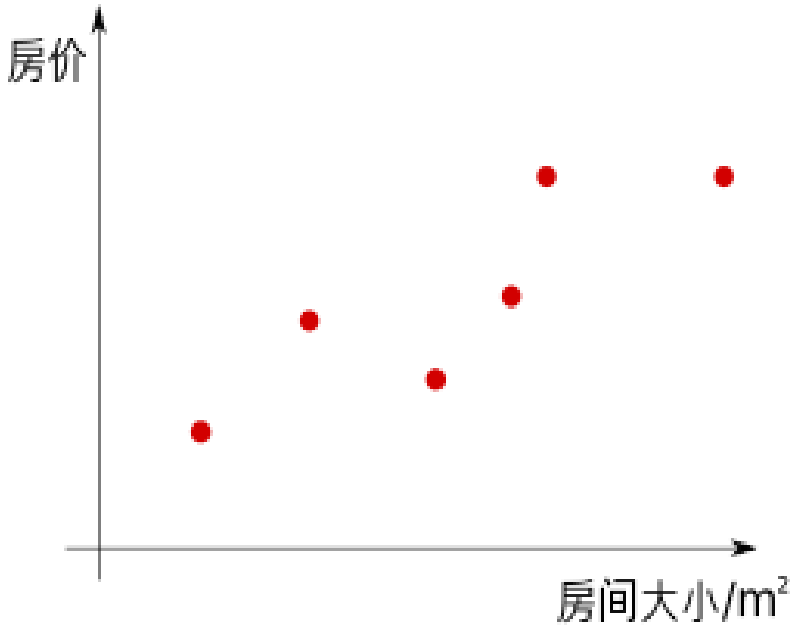
- Thinking：如何预测房价?

机器学习的7个步骤



机器学习的步骤

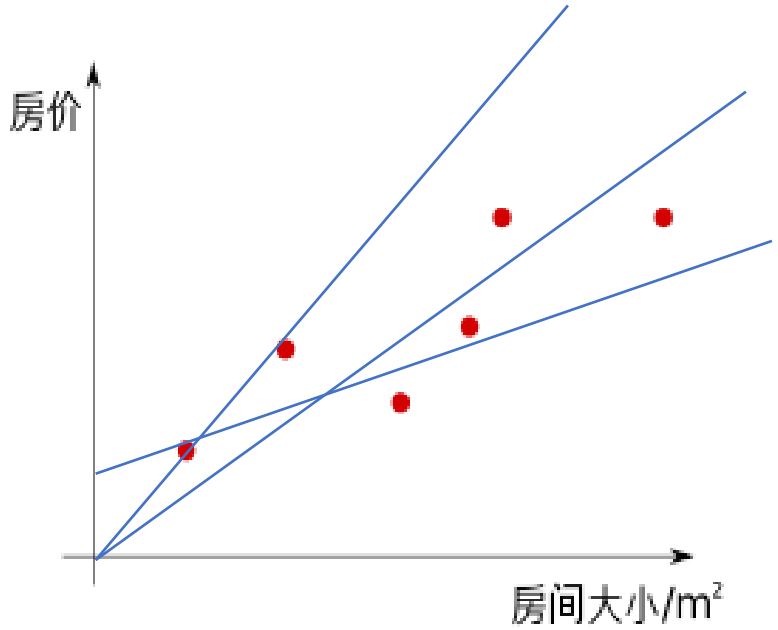
- Thinking：如何预测房价?



房间大小x	房价y
50	82
80	118
100	172
200	302
……	……

机器学习的训练过程

- 训练是机器学习的主要步骤
- 针对预测房价这个例子，我们可以
  用简单的线性模型：y=w*x + b



机器学习的训练过程

- 在机器学习中，我们有很多特征，基于这些特征，我们需要训练在Model中的权重w
- 这些特征值构成的矩阵，称之为权重矩阵 weights
- 同时，还存在偏差，称之为 biases

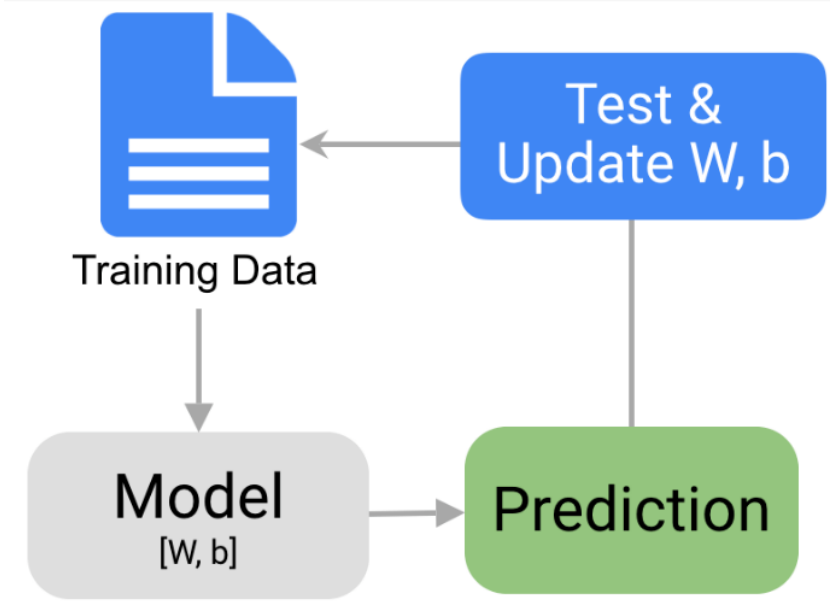
房间大小	区域	周围绿化	周边配套	房型	房价y
50	海淀	A	A	style1	82
80	通州	B	A	style1	118
100	朝阳	C	B	style2	172
200	海淀	C	C	style3	302
……	……	……	……	……	……

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

$$f(x) = \mathbf{w}^T \mathbf{x} + b$$

机器学习的训练过程

- 机器学习的过程，就是在搜索空间中对W和B进行搜索的过程，使得模型的准确率达到某个标准
- 一个训练步骤(training step)，称之为一次迭代。目的在于更新权重和变量
- 通过多次迭代，模型中的参数不断进行更新。就好像是在数据中进行线性拟合
- 当完成训练时，可以使用模型对房价进行预测



机器学习的模型选择

- Thinking：什么是回归问题，什么是分类问题?
- Thinking：什么是线性回归，什么是逻辑回归?

机器学习的模型选择

- 判断一个问题是分类，还是回归：

输出的数据类型：离散 or 连续

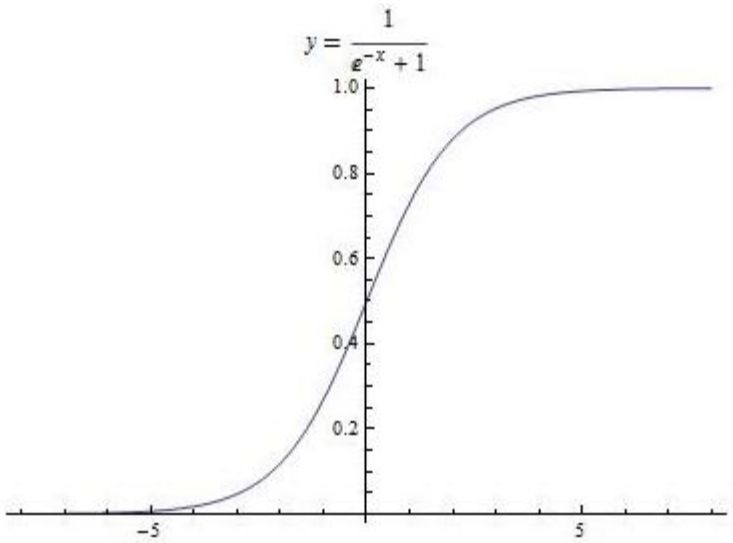
- 线性回归：

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

$$f(x) = \mathbf{w}^T \mathbf{x} + b$$

- 逻辑回归：

使用sigmod函数，实际上是分类算法



$$y = \frac{1}{1 + e^{-(w^T x + b)}}$$

机器学习的模型选择

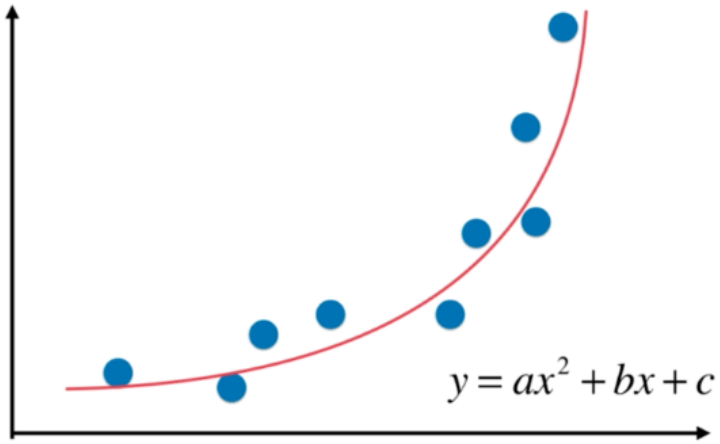
- Thinking：如何判断杯子里盛的是水，还是饮料？

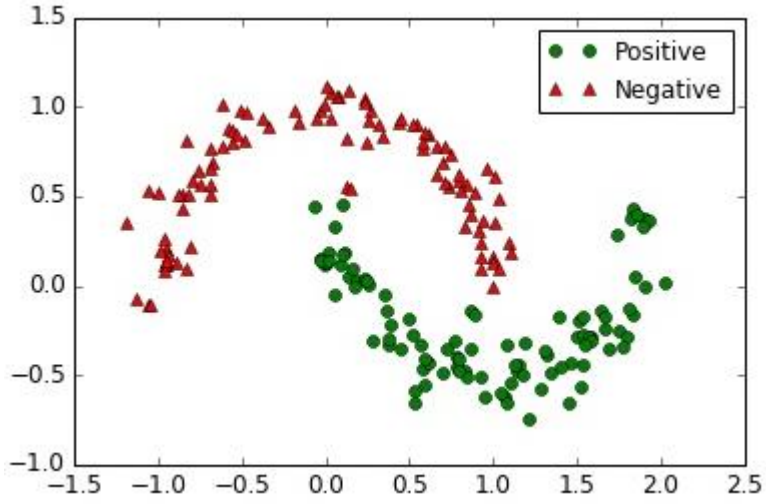
机器学习的模型选择

Color	Sugar	Classification
252	0.1%	water
210	4%	beverage
150	8%	beverage
250	0.5%	water
……	……	……

可见光的波长为400～760nm，白色是包含光谱中所有颜色的集合
因此采用Color这里采用颜色值

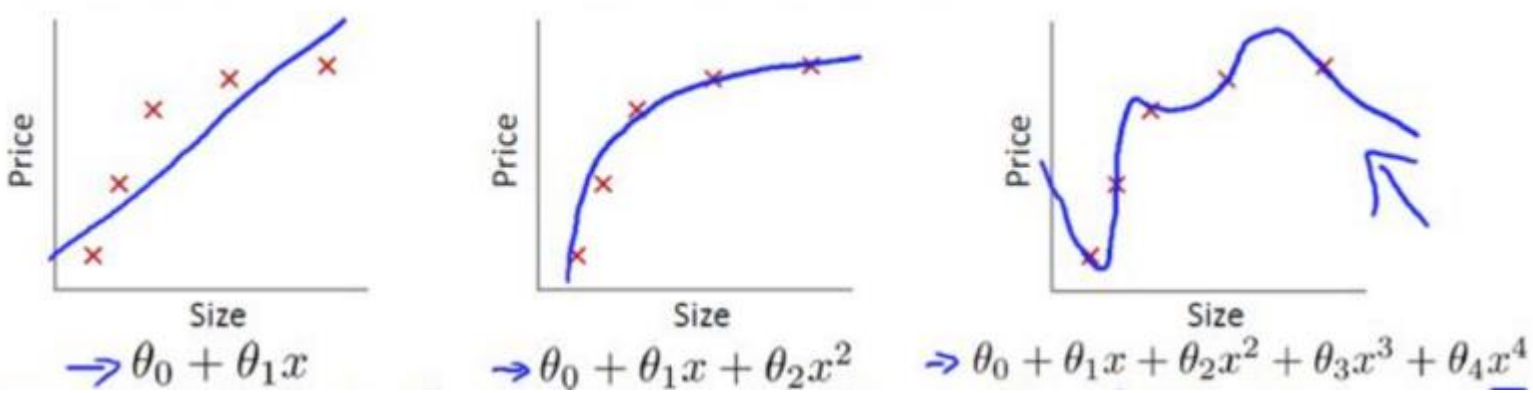
机器学习的模型选择


$y = ax^2 + bx + c$


Positive
Negative
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

机器学习的特征构造

如何用线性回归模型拟合非线性关系



机器学习的评估

- 对数据的评估有多种方式：
- 我们通过会选择一部分数据作为测试集，比如20%或者10%

超参数调整

- 我们还可以对模型中的参数进行调整，比如epoch的次数，学习率等
- 这些参数通常被称为超参数。调整超参数的过程比起科学更像是艺术。这是实验性的过程，并很大程度上取决于具体的数据集、模型和训练过程

数据分析模型

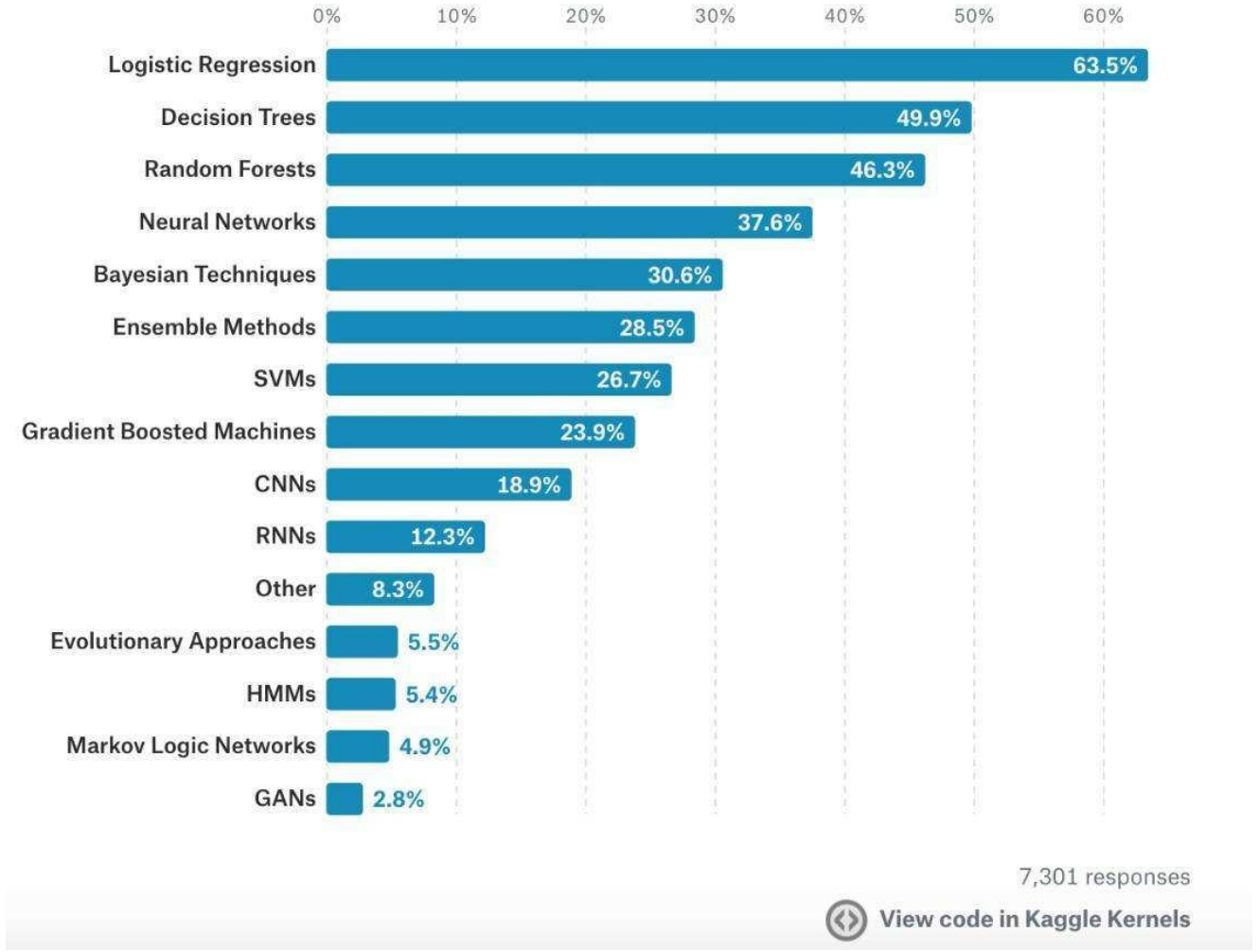
——10大经典模型

- 分类算法：C4.5，朴素贝叶斯（Naive Bayes），SVM，KNN，Adaboost，CART
- 聚类算法：K-Means，EM
- 关联分析：Apriori
- 连接分析：PageRank

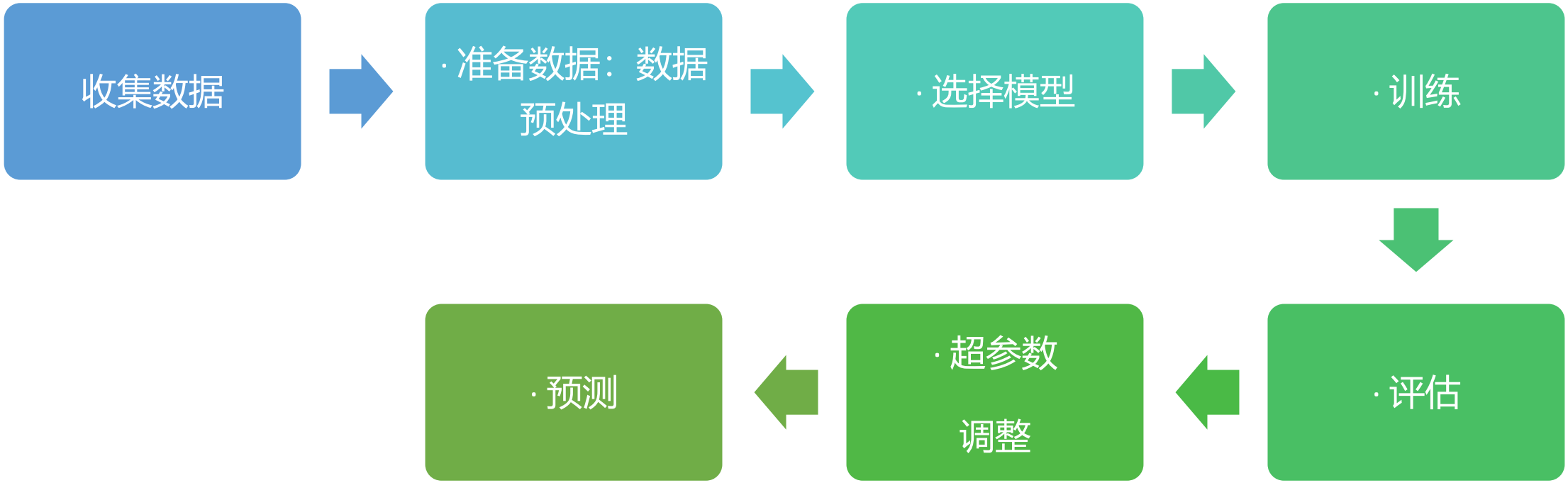
数据分析模型

主流模型

- Logistic Regression, Decision Trees, Random Forests在业界依然是主流



Summary



Thinking & Action

Thinking1：假设你是某商业银行的战略经理，需对现有业务线进行分类，并提出战略建议。

现有业务数据：

- 零售银行储蓄账户：市场增长率3%，本行市场份额2.5%
- 数字钱包/移动支付：市场增长率18%，本行市场份额1.2%
- 小微企业贷款：市场增长率8%，本行市场份额3.0%
- 传统柜台汇款服务：市场增长率-5%，本行市场份额1.0%
- 绿色债券承销：市场增长率35%，本行市场份额0.5%

以上业务如何分类？针对每类业务提出1-2条战略建议；

Thinking & Action

Action1：银行客户定期存款预测

https://tianchi.aliyun.com/competition/entrance/531993/introduction

葡萄牙银行，开展营销活动吸引更多人认购定期存款，通过营销活动沉淀的数据，想要预测未来这些人是否会进行认购，方便进行精准营销

训练集：train.csv 4459条

测试集：test.csv 49342条

字段	说明	
age	年龄	客户基本信息
job	职业：admin, unknown, unemployed, management…	
marital	婚姻：married, divorced, single	
default	信用卡是否有违约: yes or no	
housing	是否有房贷: yes, no, unknown	
Loan	是否有个贷：yes, no, unknown	
Contact	联系方式：unknown, telephone, cellular	本次活动的联系情况
Month	上一次联系的月份：jan, feb, mar, …	
day_of_week	上一次联系的星期几：mon, tue, wed, thu, fri	
Duration	上一次联系的时长（秒）	
Campaign	活动期间联系客户的次数	
Pdays	上一次与客户联系后的间隔天数	
Previous	在本次营销活动前，与客户联系的次数	
Poutcome	之前营销活动的结果：unknown, other, failure, success	
emp_var_rate	就业变动率（季度指标）	市场经济特征
cons_price_index	消费价格指数（月度指标）	
cons_conf_index	消费者信心指数（月度指标）	
lending_rate3m	银行同业拆借率 3个月利率（每日指标）	
nr_employed	雇员人数（季度指标）	
subscribe	客户是否进行存款：1 或 0	预测结果


Thank You
Using data to solve problems