机器学习

yasaka

大数据炙手可热

- 大数据公司主要有四类：
- 1,数据拥有者，数据源，PB级数据的包子铺
- 2,大数据咨询公司，Cloudera--CDH
- 3,大数据工具公司，Databricks--Apache Shark
- 4,整合应用型，结合机器学习来解决更多实际的痛点

机器学习有什么重要性

Baidu 百度

Google

facebook

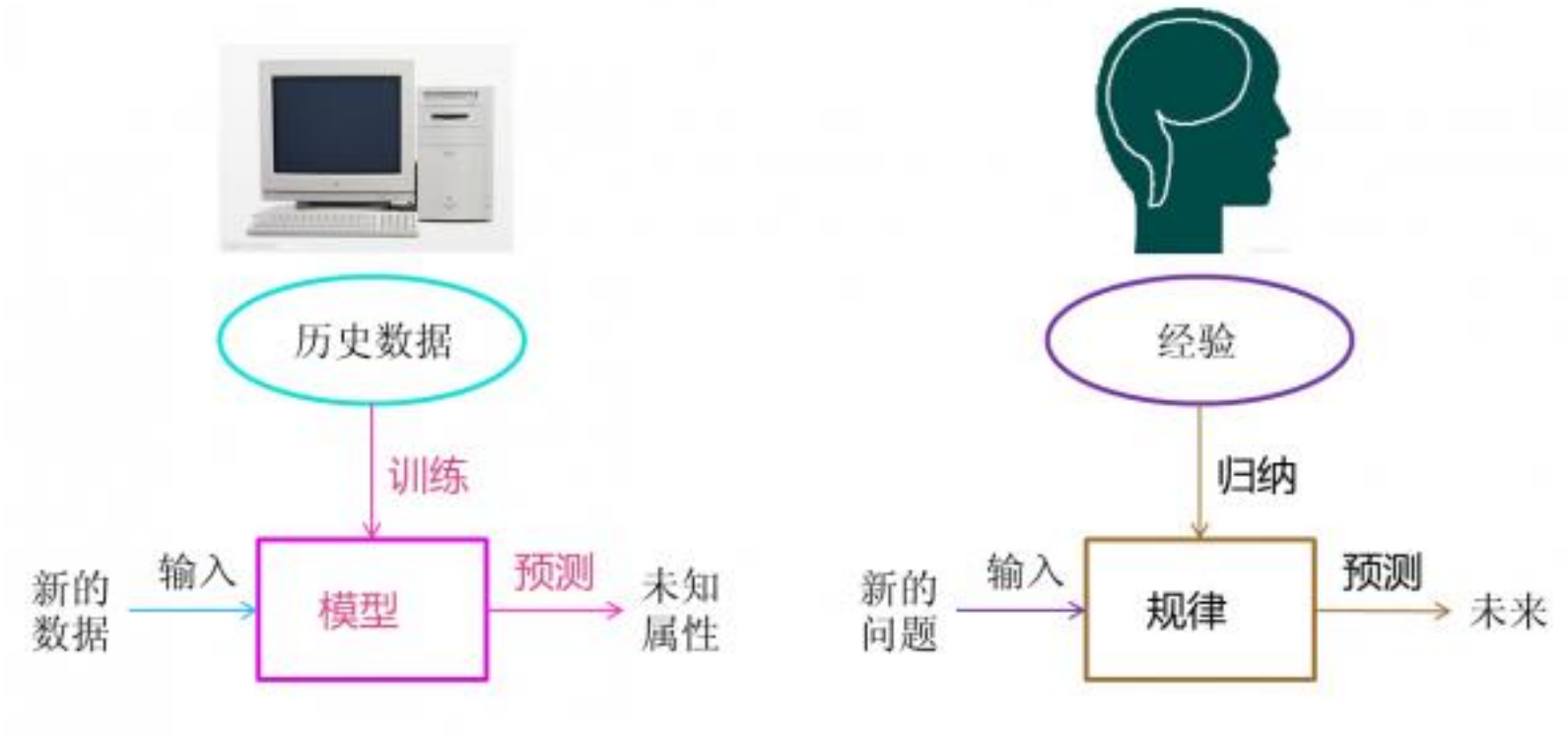
Siri, Cortana, Now

机器学习是什么

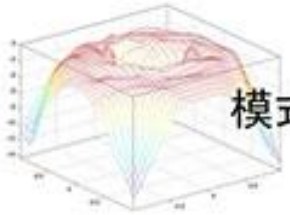
- 已有的数据(经验)
- 某种模型(迟到的规律)
- 利用此模型预测未来(是否迟到)
- 机器学习界“数据为王”思想

机器学习与人类思考的类比

- 历史往往不一样，但历史总是惊人的相似



交叉学科



模式识别

计算机视觉

数据挖掘

机器学习

语音识别

统计学习

自然语言处理

关系

- 模式识别=机器学习
- 数据挖掘=机器学习+数据库
- 统计学习近似等于机器学习
- 计算机视觉=图像处理+机器学习
- 语音识别=语音处理+机器学习
- 自然语言处理=文本处理+机器学习

利用大数据预测H1N1在美国某小镇的爆发

百度预测2014年世界杯

机器学习子类深度学习

Microsoft

Baidu IDL 百度深度学习实验室
Relentless for Ultimate Intelligence

MIT

人工智能是机器学习的父类

智慧的最佳体现是什么？

计算：云计算　　推理：专家系统

灵敏：事件驱动　　**智慧：机器学习**　　知识：数据仓库

检索：搜索引擎

疑犯追踪


PRIMARY THREAT
STATUS: MISSING
NAME: [REDACTED]
但它却看到了一切
but it sees everything...

root

大数据炙手可热

- 大数据公司主要有四类：
- 1,数据拥有者，数据源，PB级数据的包子铺
- 2,大数据咨询公司，Cloudera
- 3,大数据工具公司，Databricks
- 4,整合应用型，结合AI来解决更多实际的痛点

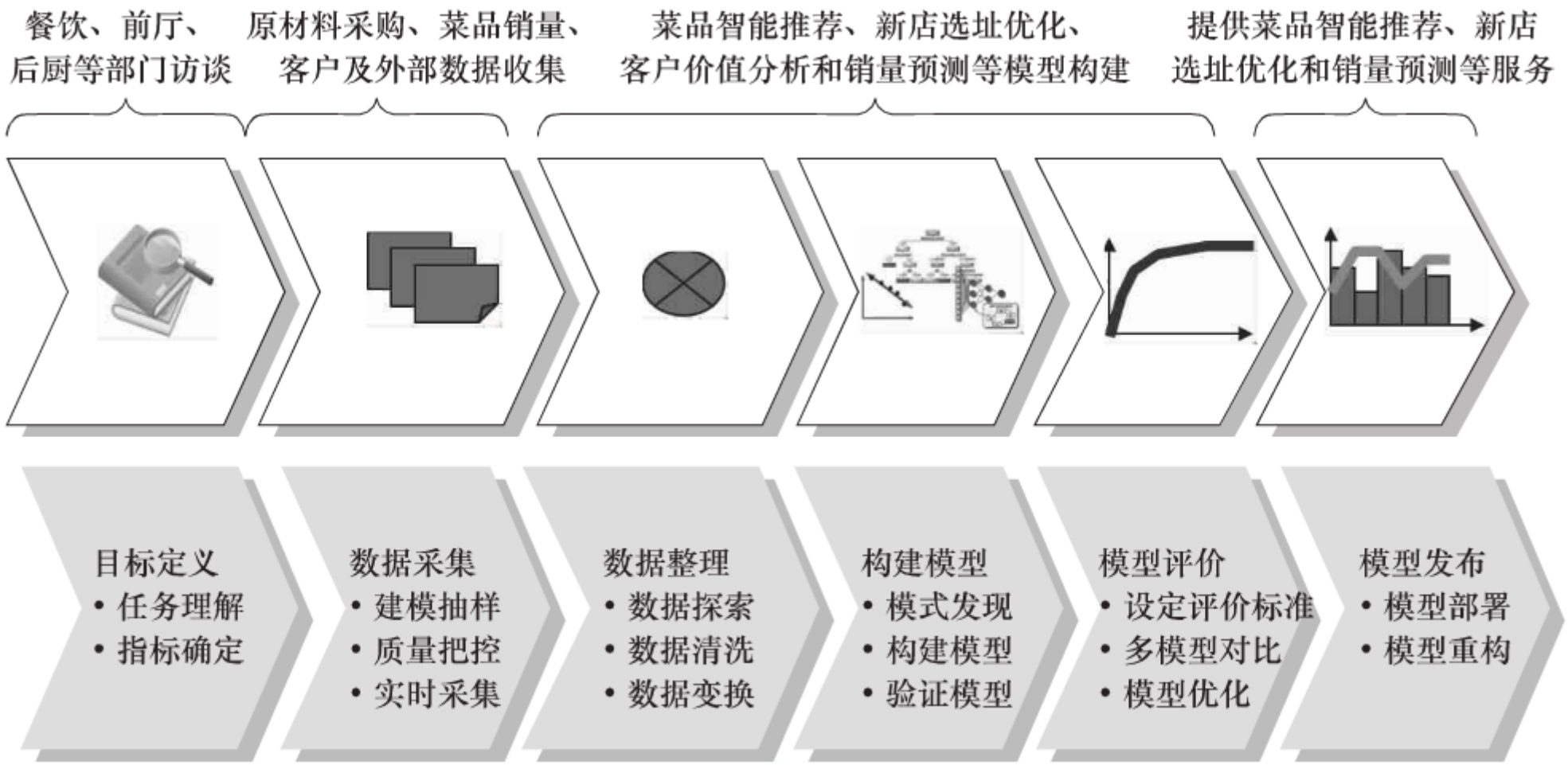
- 数据挖掘的基本任务包括利用分类与预测、聚类分析、关联规则、时序模式、偏差检测、智能推荐等方法，帮助企业提取数据中蕴含的商业价值，提高企业的竞争力。

机器学习/数据挖掘建模过程

- 定义挖掘目标
- 数据取样
- 数据探索
- 数据预处理
- 挖掘建模
- 模型评价

- 实现菜品智能推荐、促销效果分析、客户价值分析、新店选址优化、热销/滞销菜品分析和销量趋势预测。

机器学习/数据挖掘建模过程



定义挖掘目标

- 实现动态菜品智能推荐，帮助顾客快速发现自己感兴趣的菜品，同时确保推荐给顾客的菜品也是餐饮企业所期望的，实现餐饮消费和和餐饮企业的双赢；
- 对餐饮客户进行细分，了解不同客户的贡献度和消费特征，分析哪些客户是最有价值的，哪些是最需要关注的，对不同价值的客户采取不同的营销策略，将有限的资源投放到最有价值的客户身上，实现精准化营销；
- 基于菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销售进行趋势预测，方便餐饮企业准备原材料；
- 基于餐饮大数据，优化新店选址，并对新店所在位置的潜在顾客口味偏好进行分析，以便及时进行菜式调整。

数据取样

- 根据前面定义的挖掘目标，从客户关系管理系统、前厅管理系统、后厨管理系统、财务管理系统和物资管理系统抽取用于建模和分析的餐饮数据
  - 企业信息：名称、位置、规模、联系方式，部门、人员、角色等
  - 客户信息：姓名、联系方式、消费时间、消费金额等
  - 菜品信息：菜品名称、菜品单价、菜品成本、所属部门等
  - 销售数据：菜品名称、销售日期、销售金额、销售份数
  - 原材料信息：供应商、联系方式、商品名称、客户评价
  - 促销活动数据：促销日期、促销内容、促销描述
  - 外部数据：天气、节假日、竞争对手、周边商业氛围等
- 数据质量：完整性、正确性

数据探索

- 当我们拿到一个样本数据集后
  - 它是否达到我们原来设想的要求(缺失值分析)
  - 其中有没有什么明显的规律和趋势(周期性分析)
  - 有没有出现从未设想过的数据状态(异常值分析)
  - 属性之间有什么相关性(相关性分析)
  - 数据可分为哪些类别等等

数据预处理

- 当采样数据维度过大时，如何进行降维处理
- 数据筛选
- 数据变量转换
- 缺失值处理
- 坏数据处理
- 数据标准化
- 主成分分析
- 属性选择
- 数据归一化

挖掘建模

- 接下来考虑的问题就是判断目标是要做哪类分析？？？
- 选用哪种算法进行模型构建？？？
- 这一步是挖掘工作的核心环节！！！

- 对于举例餐饮行业应用，建模主要包括基于关联规则算法的动态菜品智能推荐、基于聚类算法的餐饮客户价值分析、基于分类与预测算法的菜品销量预测

- 模型说白了就是菜品销量的预测公式，公式可以产生与观察值有相同结构的输出，这就是预测值

模型评价

- 上面建模过程中会得出一系列的分析结果，模型评价的目的之一就是从这些模型中自动找出一个最好的模型，根据业务对模型进行解释和应用
- 分类与预测的模型和聚类分析的模型的评价方法是不同的

常用的机器学习/数据挖掘建模工具

- R
- Python
- SAS
- IBM SPSS
- SQL Server(Analysis Servers)
- MATLAB
- WEKA
- Mahout
- Spark MLlib

总结

- 人类要学会从比特流中解读他人，更要教会机器从比特流中理解人类