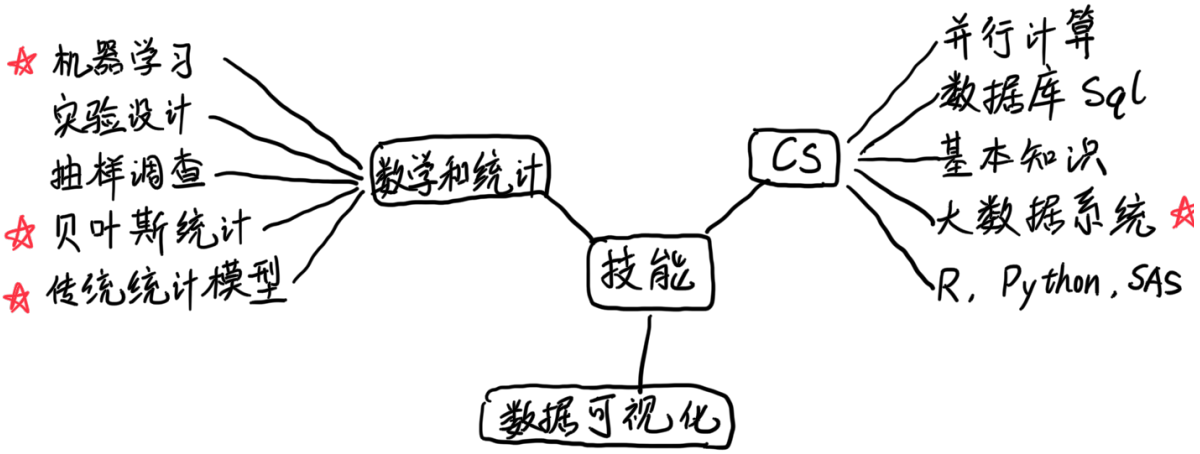
数据科学家必备技能



如何成为数据科学家：

学习资源

在线资源、
教材、文献

→

动手项目

在线课程
kaggle/算法比赛

→

分析工具

数据加工和描述统计
Sql, Excel

菜单/拖拽式分析
SPSS, SAS EM

编程：Python, R

算法拆解：Spark

岗位：

数据科学家：

⇒ 数据建模，主导复杂实验

⇒ 统计/机器学习

数据工程师：

⇒ 构建数据管线，提供平台和工具给 DS 和 DE

⇒ CS，工程背景

数据分析师：

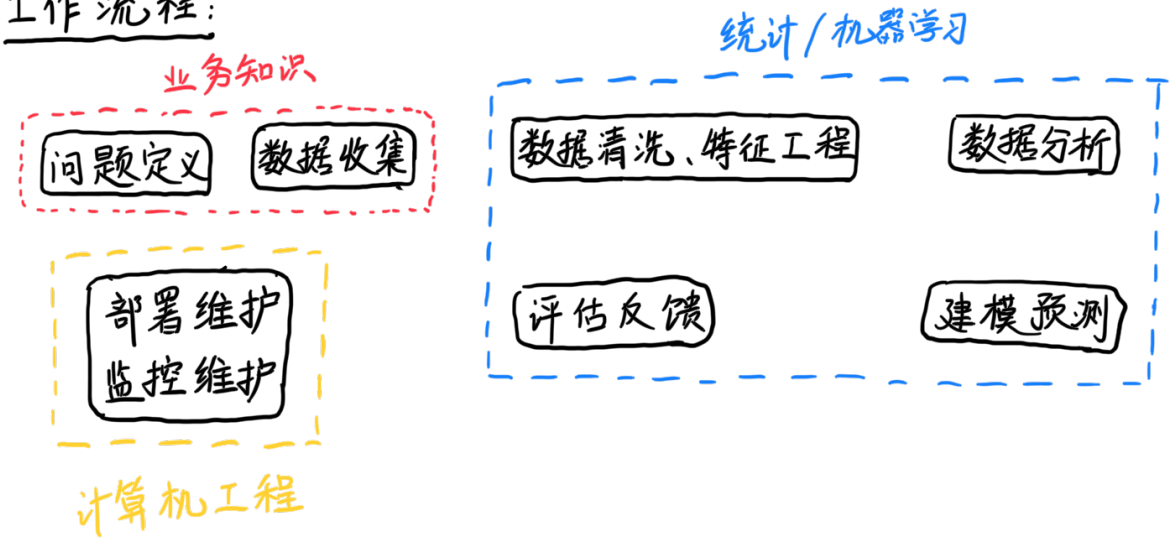
⇒ 构建数据体系，定义跟踪 metrics

⇒ ETL 抓取，提供 data insight (extract, transform , load)

⇒ 数学/统计

⇒ 业务/产品理解能力

工作流程：



定义问题

⇔ 把业务目标翻译为技术目标：

- 数据科学项目怎么帮助解决业务问题
- 技术选型

目标	技术
预测数值	回归
预测类别	分类
预测偏好	推荐系统
发现	聚类
发现异常数据	异常值检验

- 如何评估

数据探索性分析

数据探索性分析

1. 熟悉数据
   - 是否有时间趋势：周期性，突变峰值或低谷
   - 相关场景和人群：符合业务目标
   - 采样是否有代表性

2. 数据分析
   - 是否可读
   - 数据分布
   - 是否有异常值、缺失值

3. 数据清洗：
   - 唯一性检验：一个字段或多个字段的组合在整个数据集中必须唯一
   - 一致性检验：保证数据在多个数据源中表达的意义相同
   - 完整性检验：检查数据的缺失值、空值，NULL值
   - 有效性检验：在分析的时间节点是有效的
   - 准确性检验：不符合规范的值，错误值

缺失原因：MCAR Missing Completely at Random

MAR: Missing at Random

NMAR: Not missing at Random

解决方法：

① 去掉不用：MAR 或者数据足够多

② Imputation: Ad-hoc：均值，众数，0，last observation carry forward

模型预测：KNN

③ 加入新特征：是否缺失

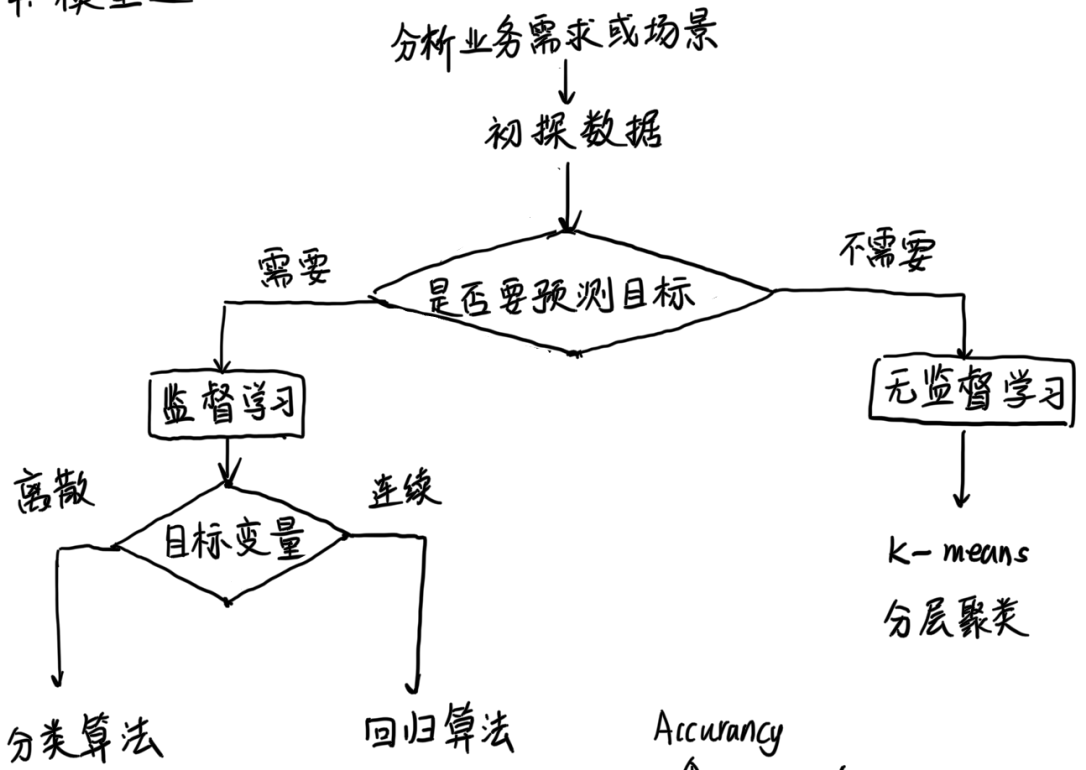
④ 改用缺失不敏感模型

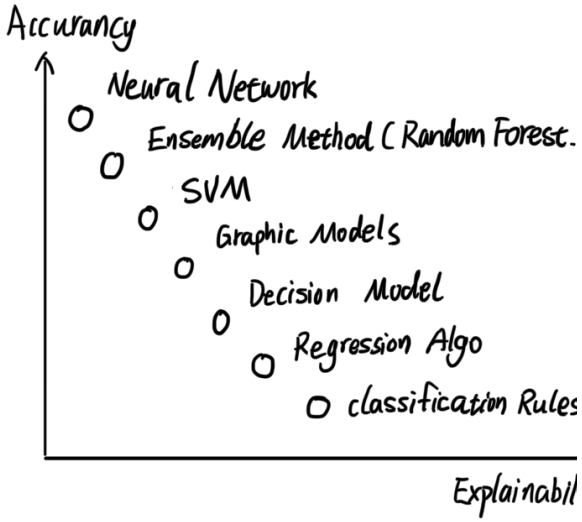
$$f(Q \mid Y_{obs}) = \int f(Q \mid Y_{obs}, Y_{mis}) f(Y_{mis} \mid Y_{obs}) dY_{mis}$$

what you get ← ∫ ← outcome model using ← Impute based on

"Integrate" over imputed dataset　complete data　missing data model

4. 模型选取





分类算法:

- Logistic Regression
- SVM
- Random Forest
- XGBoost