复习课 by Nag1

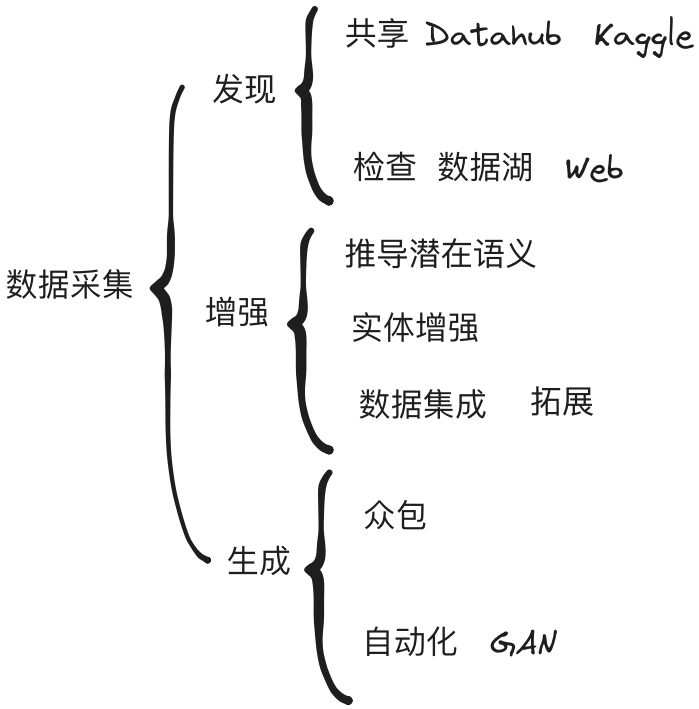
15选择x2
10判断x1
（前五章）
3简答x10（主要流程，包含哪几类预处理；K均值主要流程）
3计算x10（PPT中例题）

1. 绪论

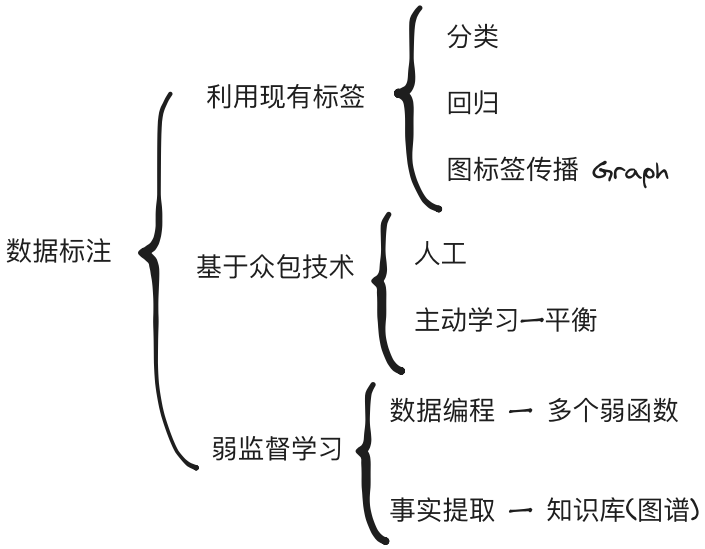
- 数据 定量、定性的属性
- 大数据 规模大 种类多 速度快 计算能力 存储需求大
- 数据挖掘 : 数据 $\xrightarrow{\text{发现}}$ 知识

数据获取* → 预处理* → 数据仓库 → 数据挖掘* → 模式评估 → 可视化 → 决策支

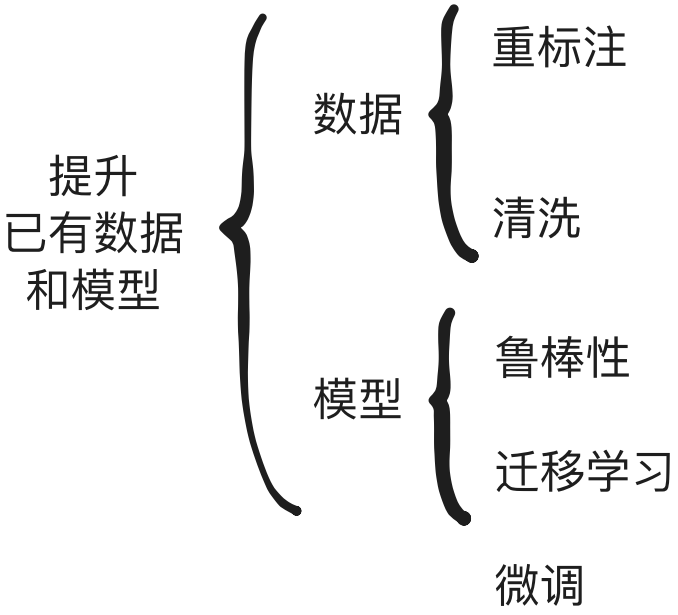
2. 数据获取



> 这里的数据湖是企业用的，和数据仓库不一样，扔的都是原始数据，了解即可。



> 图标签传播的重点是**自动**



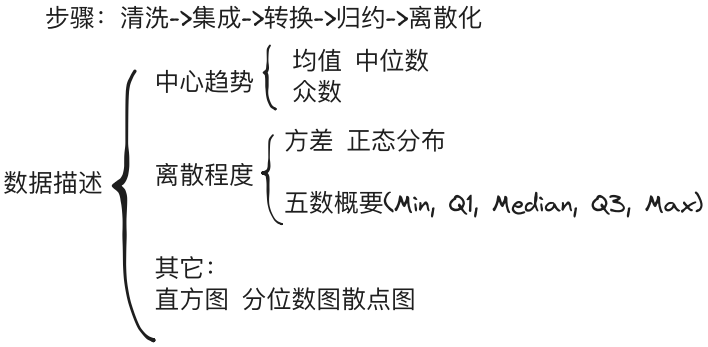
3. 数据预处理

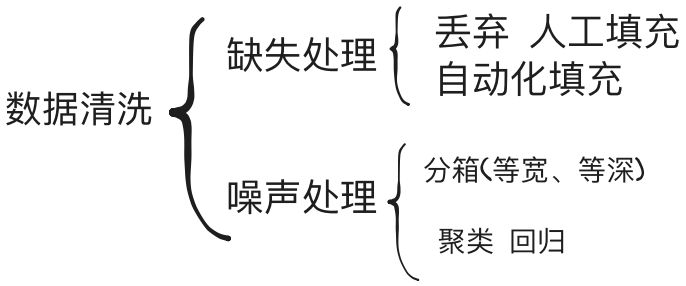
为什么预处理?

- 噪声
- 离群点（可以是合法的，可以很有价值）

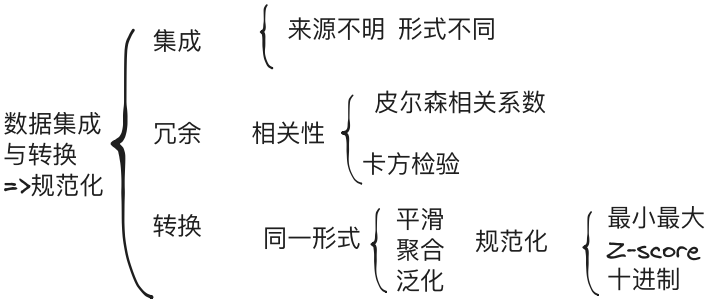
真实数据质量无法满足数据挖掘的要求：

- 不完整：没有关心的属性（比如温度差）
- 不准确：噪声、人工错误
- 不一致：冲突（比如一部分是华氏度，一部分摄氏度）





- 人工填充：根据属性推断，受人工主观因素、知识本精、人工成本限制。
- 自动化填充：
  全局统一赋值（0、"NaN"）、属性**均值**、预测缺失值（线性回归、决策树）



数据集成

- 数据的冗余性可以通过对数据进行**相关性分析**检测到
- 当相关系数$r_{A,B} > 0$，表明A与B呈正相关
- 当相关系数$r_{A,B} = 0$，表明A与B无关
- 当相关系数$r_{A,B} < 0$，表明A与B呈负相关
- 相关系数的**取值范围**是$[-1, 1]$

数据转换

- 数据聚合：OLAP（在线分析处理）和 数据立方体
- 最小最大：

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

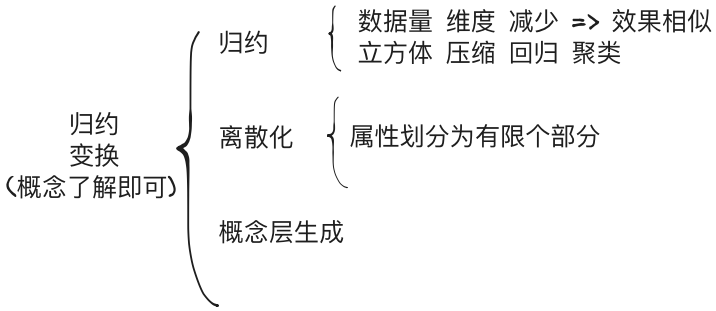
- z-score：

$$x' = \frac{x - \mu}{\sigma}$$

- 十进制：

```c
[564,46,-234,-19]
转化为
[0.564,0.046,-0.234,-0.019]
```

06 数据归约



04 数据仓库(背下描述就行)

后面的内容：K-means流程、决策树等算法流程过一遍，把PPT上计算题例题看一遍就就没问题了，老师很厉害很好但是这课本身太水。