# 1.数据分析基本概念

1. 数据分析中的三个主要场景:

商业智能 BI，数据仓库 DW（Data Warehouse），数据挖掘 DM

1. 在数据仓库中，还有一类重要的数据是元数据：

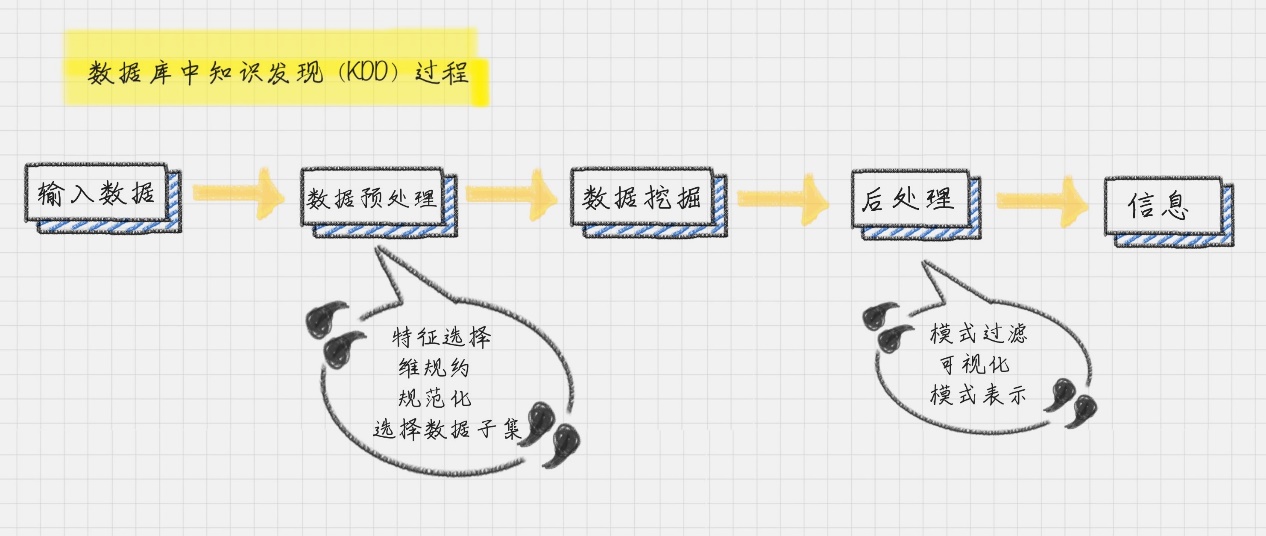
元数据（MetaData）：描述其它数据的数据，也称为“中介数据”。例如一个事物下的所有描述，书：作者，出版社，发行日期，刊号等，这些属性为一套书的元数据。

数据元（Data Element）：就是最小数据单元。

1. 数据挖掘过程：

1）在数据挖掘中，有几个非常重要的任务，就是分类、聚类、预测和关联分析。

2）KDD：Knowledge Discovery in Database, 数据挖掘的一种解释



1. 数据预处理：
2. 数据清洗：主要是为了去除重复数据，去噪声（即干扰数据）以及填充缺失值。
3. 数据集成：是将多个数据源中的数据存放在一个统一的数据存储中。
4. 数据变换：就是将数据转换成适合数据挖掘的形式。比如，通过归一化将属性数据按照比例缩放，这样就可以将数值落入一个特定的区间内，比如 0~1 之间。

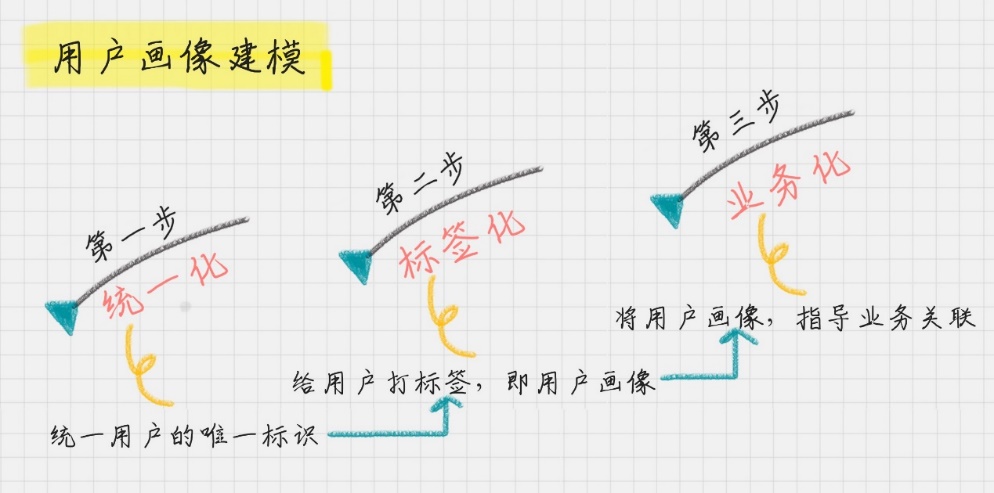
# 2.标签化就是数据的抽象能力

1. 用户画像的准则：

都是谁：需要统一标识用户 ID，方便我们对用户后续行为进行跟踪。

从哪来：我们需要对这些用户进行标签化，方便我们对用户行为进行理解。

要去哪：们要将这些用户画像与我们的业务相关联，提升我们的转化率，或者降低我们的流失率。”



1. 首先，设计唯一标示：用户唯一标识是整个用户画像的核心。设计唯一标识可以从这些项中选择：用户名、注册手机号、联系人手机号、邮箱、设备号、CookieID 等。
2. 其次，给用户打标签：用户行为分析，即用户画像，用户的数学建模，海量数据标签化。

用以下4个维度来进行标签划分：

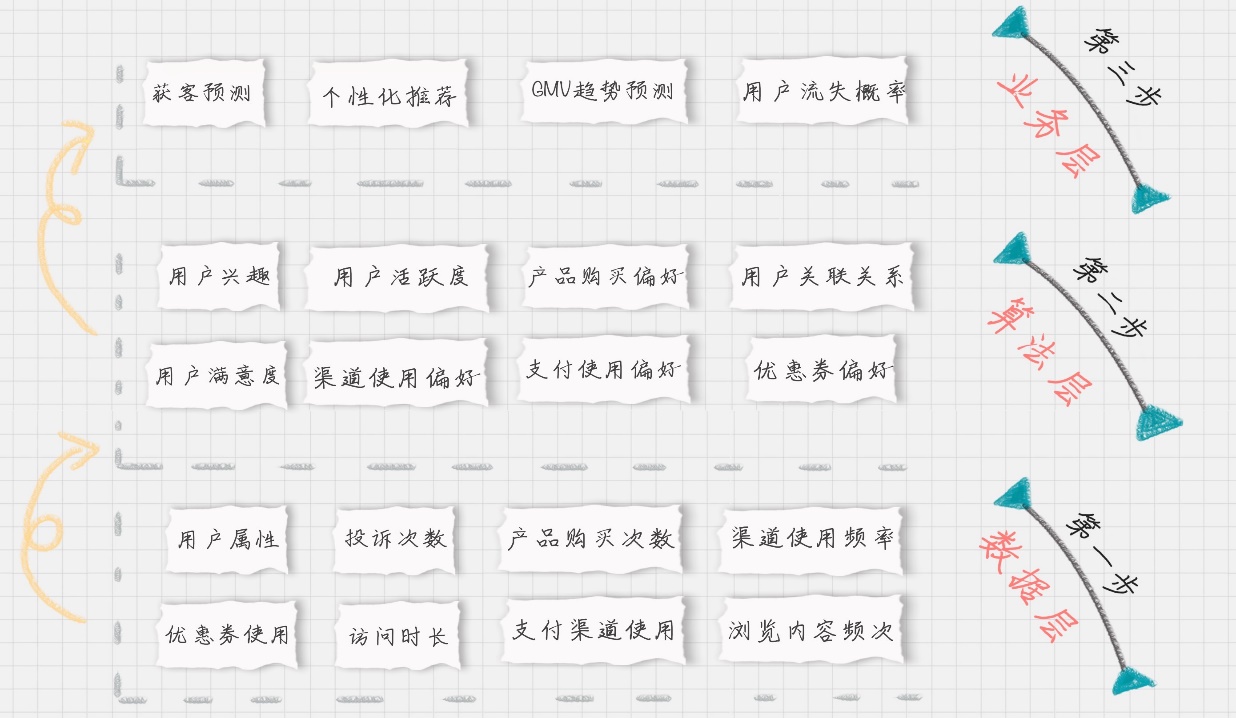
1. 用户标签：包括性别，年龄，收入，职业等，用户的基础属性
2. 消费标签：消费习惯，购买意向，是否对促销敏感，用户的消费习惯
3. 行为标签： 时间段，频次，时长，访问路径，从用户行为得到他们使用APP的习惯
4. 内容分析：用户平时浏览的内容，尤其是对停留时间长，浏览次数多的内容进行分析，分析出用户对那些内容感兴趣，如：金融，科技，娱乐等
5. 标签化的流程：

数据层：用户消费行为里的标签。“事实标签”，作为数据客观的记录。

算法层：透过这些行为算出的用户建模。“模型标签”，作为用户画像的分类标识。

业务层：指的是获客，粘客，留客的手段。“预测标签”，作为业务关联的结果。

所以这个标签化的流程，就是通过数据层的“事实标签”，在算法层进行计算，打上“模型标签”的分类结果，最后指导业务层，得出“预测标签”。



# 3.数据采集：自动化采集数据

一个数据的走势，是由多个维度影响的

我们需要通过多源的数据采集，收集到尽可能多的数据维度，同时保证数据的质量，这样才能得到高质量的数据挖掘结果。

数据源：



如何使用开放数据源：



爬虫抓取：



工具爬虫抓取网页内容：



日志采集工具

日志采集最大的作用就是通过分析用户访问情况，提升系统的性能，从而提高系统承载量。及时发现系统承载瓶颈，也可以方便技术人员基于用户实际的访问情况进行优化。

日志记录了用户访问网站的全过程：哪些人在什么时间，通过什么渠道（比如搜索引擎、网址输入）来过，，都执行了哪些操作；系统是否产生了错误；甚至包括用户的 IP、HTTP 请求的时间，用户代理等。这些日志数据可以被写在一个日志文件中，也可以分成不同的日志文件，比如访问日志、错误日志等。

