1. 问题描述：
   1. 交互推荐：

用户曾经对某商品看过，收藏过，加入过购物车或者购买过，其中，**交互且购买**的为正样本，**交互未购买**的为负样本。属于监督学习。

* 1. 未交互推荐：

用户过去并为对于该商品有任何操作，但可以根据与用户行为相似的其他用户或者用户过去的购买行为进行推断。属于无监督学习。

1. 基本过程：
   1. 数据清理：
      1. 清除爬虫用户
      2. 清除异常用户
      3. 降低特殊事件段的影响（双12）
      4. 对于缺失的部分数据，是否考虑进行平滑？
   2. 数据转换：
      1. 日期字符串 -> 可以比较大小的整数
      2. 地理数据字符串 –>其实同样可以转化为可以比较大小的字符串，甚至可以考虑分成7个部分来存，每个部分用ASCII码顺序。
      3. 归一化
   3. 数据分析：
      1. 我们需要对于数据中的正负样本的量进行初步估计，对于具有相似特征的用户与商品进行一定的分析；
      2. 正负样本的抽样，遴选，是否要下采样与过采样；
      3. 选择Offline Validation Dataset
      4. 关于特征分析这一步分，我们可能要对于样本进行充分的可视化以及统计特性处理，分析统计显著性。
   4. 特征选择：
      1. 核心思想：所有特征构建围绕：品牌热门程度，用户购买能力，品牌购买周期，用户购买周期，用户重复性行为，用户行为偏好，用户交互价值，品牌价值，用户价值，行为贡献这几个总体方向去考虑
      2. 选择比较有代表性的“清水湾沙滩青年流浪者”的方案进行分析：给定<user\_id, item\_id>，分析以下属性：
         1. 当前用户的总体特征，比如其购买习惯，购买行为，交互行为
         2. 当前item的特征，比如被多少人买过之类
         3. Cross特征1： item与其他item之间的关联特征；
         4. Cross特征2：item与user之间的关联特征；
         5. Cross特征3：对于我们这次竞赛，由于引入了地理信息，是否要分析与该用户地理位置相近的用户行为，等等，即user-user关联特征
         6. Cross特征其他：由于还涉及了item\_category，是否要考虑同类商品之类的信息？
         7. 总结：由于本次数据字段大大扩充，因此Cross特征的分析是我们的重要任务。
      3. 待分析内容：
         1. 各种行为的次数，人数，订单数，
         2. 相隔日期，
         3. 各种数据之间的比例
      4. 粒度选择：
         1. 日期：近期密，远则稀疏，但是注意交叠
         2. 次数：同样要进行分箱操作
   5. 模型学习：
      1. 总体趋势：多个model结果的融合：
         1. 不同类型的model：如Logistic Regression， GBRT， RF，SVM等
         2. 不同参数，同model：GBRT可选不同迭代次数，不同深度
         3. 同参数，同模型，不同训练集：对于训练数据进行时间维度上的分割
         4. 同参数，同模型，同训练集：由于算法的不稳定性，可能每次训练得到的结果不太一样
      2. 融合策略：
         1. Top K输出合并
         2. 加权
         3. 排序融合
         4. 多层次融合：即基于上面不同的model类型区分方法，建立层次

以上都是从之前的答辩slides中进行的相关总结，但是对于S1是否真的合适？根据之前的几个采访，S1在小数据量的情况下，并没有采用比较高的维度特征（90D及其以下），其次，优先考虑的也是规则的方式进行推荐。目前有些不确定的是，规则的方式一般是如何进行的？这个可能要杰哥给我们介绍一下。

之所以在S2中大规模使用LR,GBRT等直接融合，我个人猜测是因为平台限制，参赛者也没有太多时间花在算法的创新实现上，这也是与我之前参加数学建模比赛时的经验是一致的，我们没有必要在这类比赛中尝试过多的创新，“**灵活地基于现有算法+人工数据干预**”可能是比较好的方案。

关于现有的算法的实现，考虑采用平台：Python

数据库框架：MySQL和Python的链接：MySQLdb，在windows下安装：不会。。主要是64位windows没有直接的binary包，编译我这儿一直悲剧TT，要不就都在linux下用吧。

机器学习框架：采用scikit-learn：

Windows和Linux下都可以，可以直接用pip来安装,pip install scikit-learn

需要注意的是，scikit-learn需要numpy和scipy支撑，二者在linux下安装可以直接用sudo apt-get install python-numpy python-scipy完成。Windows下感觉好麻烦。。因为numpy还依赖lapack之类的矩阵库。。要不写代码一律linux吧。。好慌。。我在windows下装了几次numpy都悲剧了。