**大数据方法论**

1. **理解业务需求**
2. 清晰界定问题，理清提问者的真正目的：如何降低成本？🡪提高效率or提升盈利能力？
3. 分解目标，找出优先关注的问题，围绕该问题进行数据分析
4. 开展分组讨论，有各利益相关者参与，协助确定需求、理清问题

**案例分析：**

* 背景：在公共资金对再次住院费用的给付额度下降的背景下，美国的健康保险公司面临补偿的差额上升而导致向客户收取的保险费率上升（已知不受欢迎）的风险，因而与地方医疗管理机构和IBM数据科学家商讨用如何数据科学解决“如何最优分配有限的医保预算以最大限度用于提供高质量的保障”这个问题。
* 决策路径：
* 确定目的和目标；
* 确定优先讨论“病人再次住院”问题；
* 研究发现：
* 大约30%完成康复治疗的病人在一年内再次住院
* 50%的病人在五年内再次住院
* 充血性心力衰竭患者位于再次入院名单首位
* 进一步确定可应用决策树模型分析上述情形发生的原因；
* 数据科学家发起并召开现场研讨会，关键业务发起人参加，其主要职责包括：
* 设定总体方向（界定问题、项目目标、解决方案需求）
* 持续积极介入并提供指引
* 按需确保必要的支持
* 确定如何建立模型以满足以下四种需求：
* 预测每个充血性心力衰竭患者的再次住院结果（是否再次住院）
* 预测每个病人的再次住院风险
* 理解哪些事件相结合导致了每个病人的预测结果
* 易于理解并应用于新病人以预测其再次住院风险

1. **确定分析方法**
2. 根据提问者提出的问题，选择最合适的分析路径和方法：

* 确定行动的**概率**的问题 🡪 **预测性**模型
* 表示**关系**的问题 🡪 **描述性**方法（着眼于基于事件和偏好的相似活动聚类）
* 需要计数的问题 🡪 统计分析
* 需要**是/否**答案的问题 🡪 预测响应的**分类**模型
* 从数据中辨识关系和趋势 🡪 机器学习
* 学习人类行为 🡪 聚类关联方法

**案例分析：**

* 上述案例中使用决策树模型辨识哪些情况相结合导致了每个病人的结果
* 仔细观察通向树叶的每条路径上的每个节点中通向各自的阈值的变量。这意味着决策树分类器既提供预测结果，又根据各组中占优结果（是或否）的比例提供结果的可能性。本例中，分析师可以获取每个病人的再次住院风险（结果为是的可能性）：{
* 如果占优结果为是 🡪 风险 = 树叶中结果为是的病人比例
* 如果占优结果为否 🡪 风险 = 1 – 树叶中结果为否的病人比例
* 决策树分类模型容易被非数据科学家理解，为新病人的再次住院风险评分。临床医师可以了解什么情况造成病人评分结果为高风险，并且在住院期间不同时点可以构建和应用多重模型。由此形成了病人风险及其如何随着采用不同治疗方法而演变的动态图像。