7.3. 数据科学的项目管理

▲7.2 数据产品及开发

▼7.4 数据能力

数据科学的项目管理

图 7-3 项目管理的主要内容

数据科学项目应遵循一般**项目管理的原则和方法**，涉及范围、时间、成本、质量、 风险、人力资源、沟通、采购及系统管理等 9 个方面的管理，如图 7-3 所示。

1 数据科学项目中的主要角色

表 7-1　数据科学项目中的主要角色及其任务

角色	描述
项目发起人（Project Sponsor）	项目的投资者，代表的是项目最终利益与目的
项目经理 （Project Manager）	项目的实际管理者，包括项目范围、时间、成本、质量、风险、人力资源、沟通、采购及系统的管理
客户（Client）	项目的最终用户，代表的是项目的用户需求。同时，客户往往是数据科学项目中扮演领域专家的角色
数据科学家（Data Scientist）	负责项目发起人、经理、客户、数据工程师之间的有效沟通；负责数据管理 策略以及数据处理方法与技术方案的选择；负责数据产品的研发，如数据处理结果的可视化等
数据工程师（Data Engineer）	负责在具体的软/硬件上部署和实施数据科学家提出的方法与技术方案
操作员（Operations）	负责管理软硬件系统和基础设施（如云平台等）。例如，系统管理员、硬件维护人员等

2 数据科学项目中的主要活动

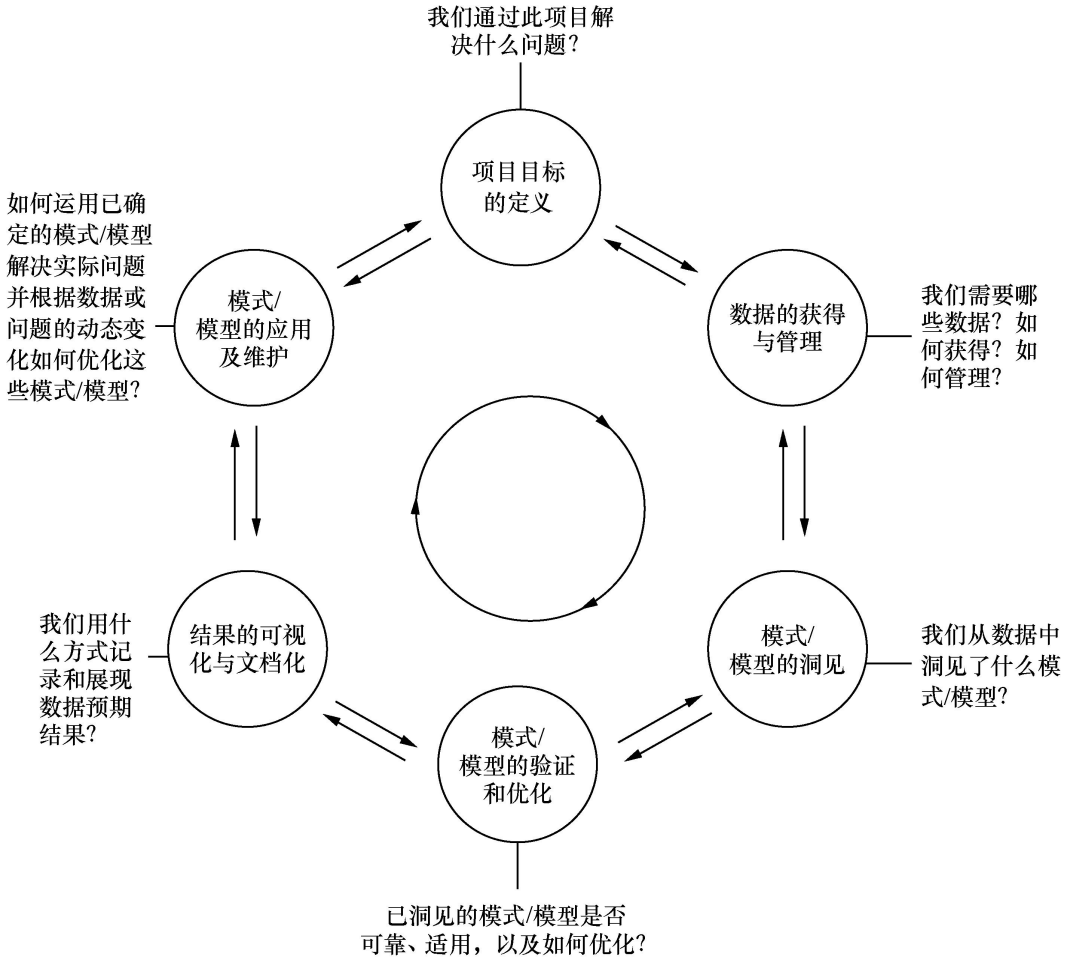

图 7-4 数据科学项目的基本流程

从图 7-4 可以看出，数据科学项目是由“项目目标的定义”到“模式/模型的应用及维护”的一系列**双向互联的互动链条组成的循序渐进**的过程，主要涉及的活动如下。

（1）项目目标的定义

- 主要回答的问题是“我们通过此项目解决什么问题”。
- 项目目标的定义应符合 SMART 原则的要求，即具体（Specific）、可测量（Measurable）、可实现（Achievable）、 相关（Relevant）和可跟踪（Traceable）。
- 定义目标的前提是调查项目需求—问题域、研究假设与项目边界，尤其是项目干系人（Stakeholders）最关心的核心问题。
- **项目干系人“最关心的问题”不一定是数据科学项目要解决的“最核心问题”。**

（2）数据的获得与管理

- 主要回答的问题是“我们需要哪些数据？如何获得？如何管理”。
- 在定义项目目标的基础上，进一步分析项目所需的数据及其属性，并判断其“可获得性”。
- 如果“可获得”，需要“自己收集”还是“利用已有数据”？
- 还需要考虑是否需要进行 数据加工、数据计算所需的平台，以及数据管理技术。

（3）模式/模型的洞见

- 主要回答的问题是“我们从数据中洞见了什么模式/模型”。
- 采用数据统计和机器学习的知识对数据进行分析与处理，**挖掘数据中隐藏的有用的**“信息”或（和）“知识”，为项目目的的实现提供“可能的解决方案”。

（4）模式/模型的验证和优化

- 主要回答的问题是“已洞见的模式/模型是否可靠、可用，以及如何优化”。
- 在洞见可能的解决方案—数据中隐藏的模式/模型之后，需要对其进行可靠性验证和可用性分析，分析我们已发现的模式/模型的信度和效度，并判断是否可用于解决项目的研究问题。
- 可以以已发现的模式/模型为基础，利用历史数据或新增数据，进一步优化模式/模型。

（5）结果的可视化与文档化

- 主要回答的问题是“我们用什么方式记录和展现数据结果”。
- **结果的可视化和文档化**分别代表的是数据项目结果的可视化表达和文档化记录（包括**故事化描述**）。
- 可视化和文档化方式的选择对于数据科学项目的成功，尤其是项目干系人的正确理解具有重要意义。

（6）模式/模型的应用及维护

- 主要回答的问题是“如何运用已确定的模式/模型解决实际问题，并根据数据或问题的动态变化优化这些模式/模型”。
- 在完成模型的验证和优化以及结果的预期表达方式的选择基础上，我们需要运用模型来解决现实世界的问题—项目干系人最关心的核心问题。