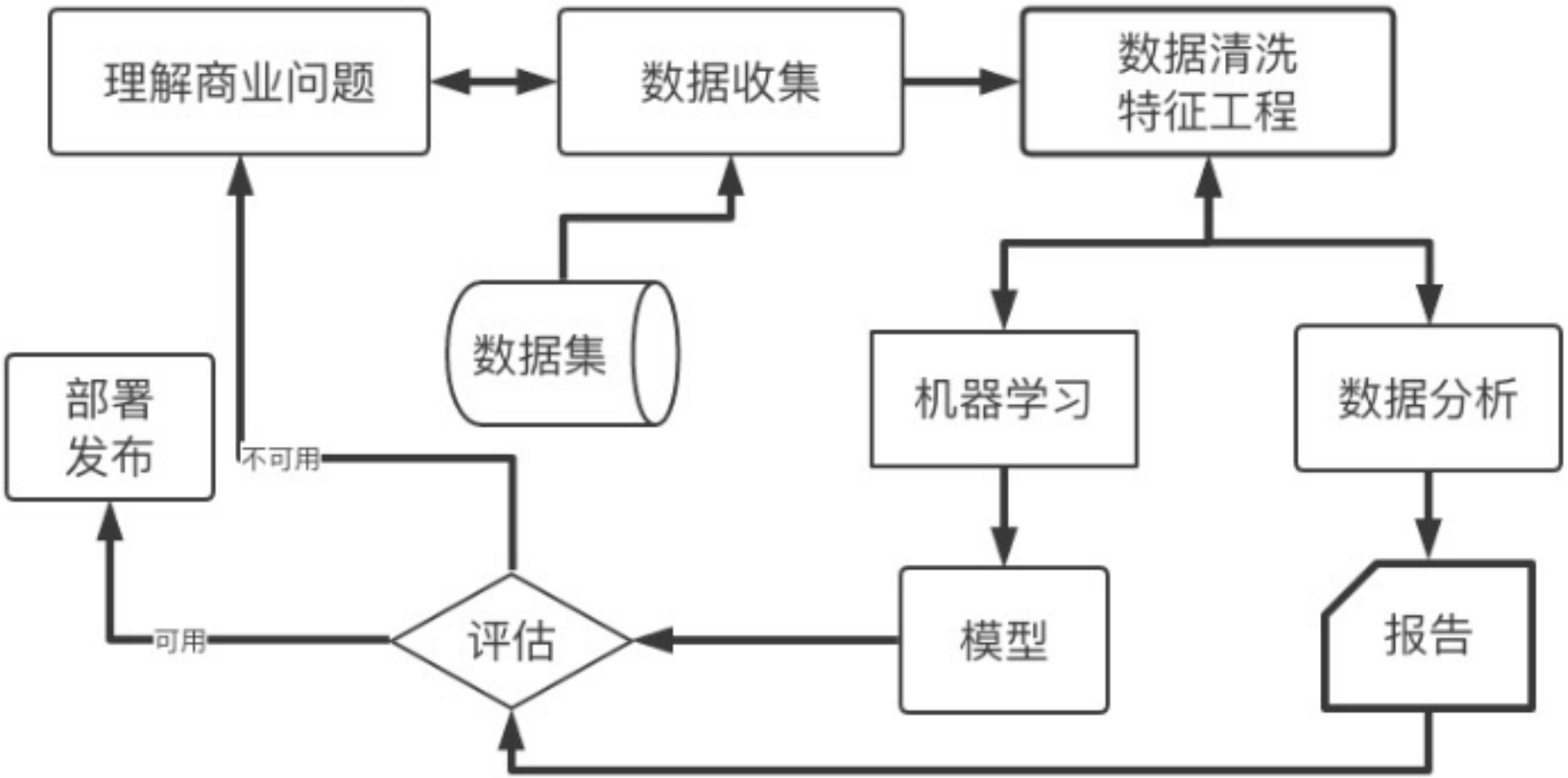
数据工程流程



理解商业问题

从业者——数据工程师，必须对相应的业务有所了解，这也是数据工程师特有的市场价值之一。

理解商业问题，并非是成为业务高手，而是要能够从业务中梳理出与数据工程项目有关的环节，特别是将业务中某些问题转化为数据问题。

数据收集

数据收集和前述理解商业问题，两者之间是一个互动关系。研究收集数据的方法，也是对商业问题的再度理解。
此外，数据收集还包含着从某个数据集中获得数据的含义。这里所说的数据集，包括但不限于：

- 数据库，包括关系型和非关系型
- 数据接口（API）
- 保存数据的文件，比如 Excel、CSV 文档等

以上这些是常用的数据集。如何从这些数据集中读取到数据？需要的技能应该是：
熟练使用 SQL
熟练使用某种编程语言（本课程使用的是 Python 语言）

数据清洗和特征工程

假设已经通过某种合法的方式得到了某些数据，接下来要做的是了解这些数据，主要通过以下两种方式：

- 对数据进行简单的描述性统计
- 对数据实行可视化，直观地了解数据概况

这里就用到了“数据可视化”的技能。

然后就是“数据清洗”和“特征工程”，这是另外两个重要工作（本课程也会涉及）

两个分支

有了数据之后，根据商业问题的目标，可以从事两个方面的具体工作。

（1）数据分析

应用各种数据分析的方法，最终得到一份分析报告。

分析结果，除了用数字表达之外，可视化是不可避免的。

（2）机器学习

机器学习是另外一个专门领域，目前正火热中。

通过机器学习算法，实现对数据的分类、预测和聚类等操作。

评估

不论是机器学习，还是数据分析，其结果都要进行评估。

对于机器学习而言，有专门的模型评估方式。即便如此，用可视化的方式把结果表达出来，也是一种重要的手段。

根据评估结果，确定是否采用机器学习所获得的模型，亦或数据分析的报告是否被采纳。

以上是数据工程项目的基本流程，从中可知，**“数据可视化”并不是流程中的一个独立环节，它是几个环节中必不可少的实现手段。**

数据报告

一个完整的数据报告，应至少包含以下六块内容：

1.报告背景

2.报告目的

3.数据来源、数量等基本情况

4.分页图表内容及本页结论

5.各部分小结及最终总结

6.下一步策略或对趋势的预测