数据预处理基础

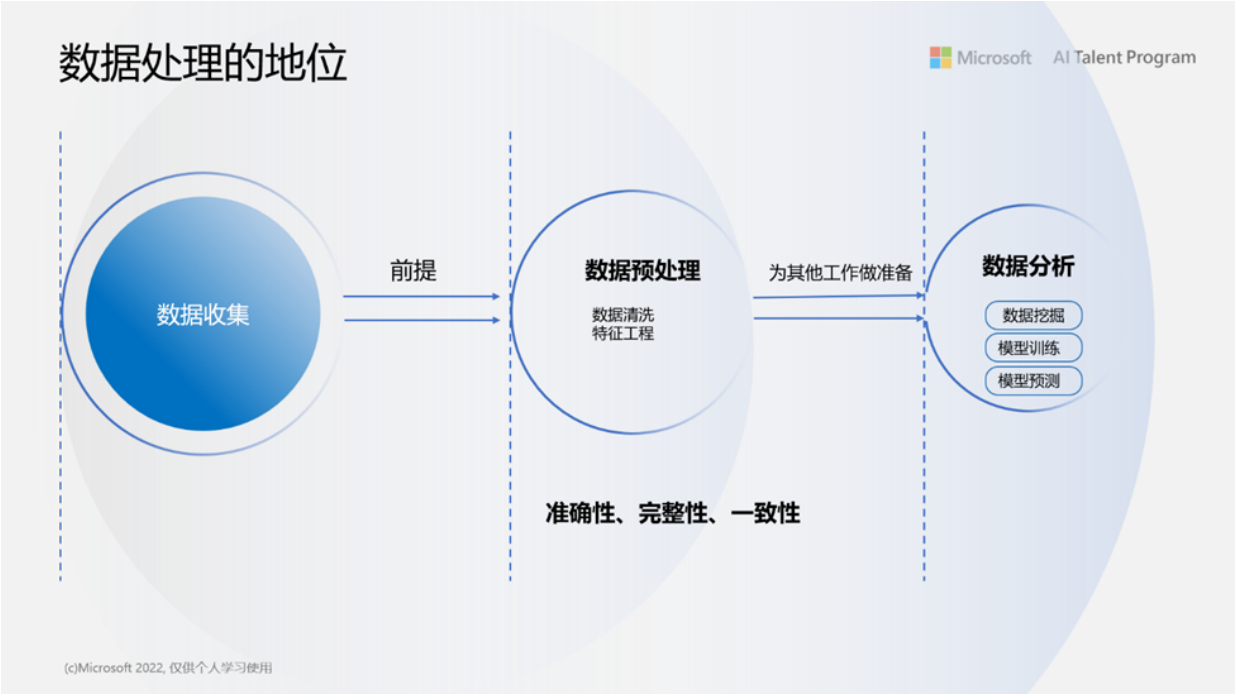
该课程主要为大家讲授如下的内容：

- 数据预处理的地位
- 数据预处理的必要性

1. 数据预处理的地位

1. 什么是数据预处理

数据分析的流程一般分为三个步骤：数据收集、数据预处理和数据分析。数据预处理作为中间环节，承接了前一阶段收集来的原始数据，并向数据分析、模型训练和数据可视化等下游任务提供优质的原材料。数据预处理的目标就是提高整个数据的质量。

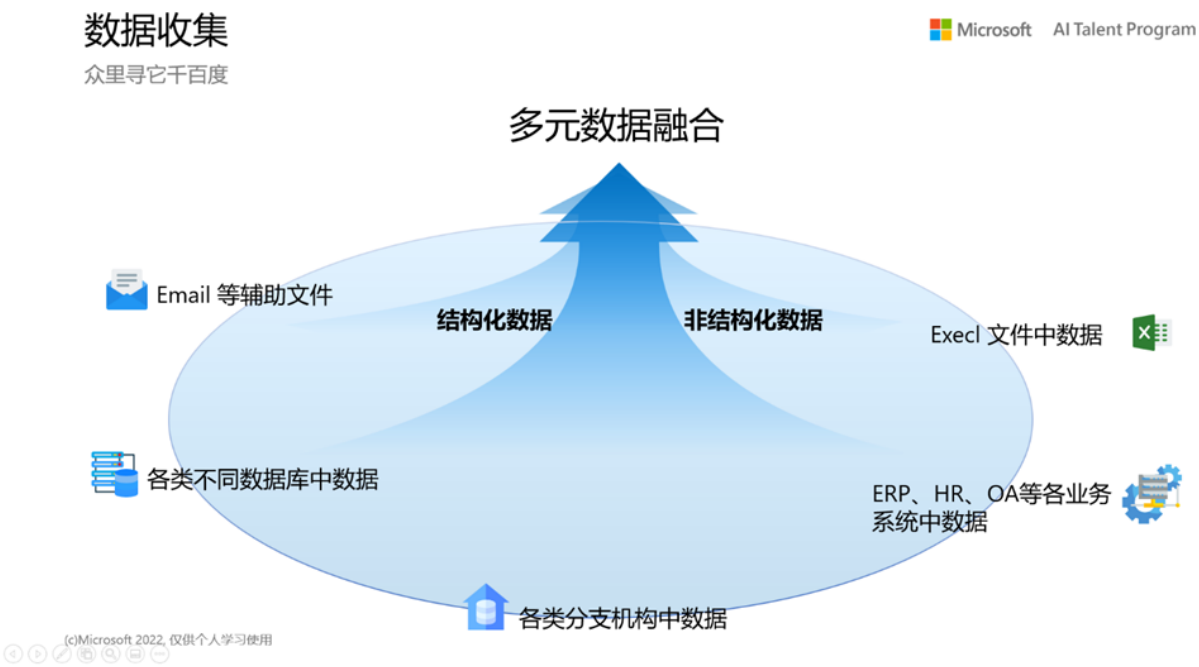


2. 数据质量的衡量标准

- 准确性：数据是否正确，是否包含异常值或错误值；
- 完整性：数据是否存在缺失现象；
- 一致性：数据内部是否服从同样的尺度、标准，数据之间的逻辑是否一致。

2. 数据预处理的必要性

一定要有数据预处理这个步骤才能保证数据的质量吗？直接把收集来的原始数据送去分析不可以吗？我们从数据收集和数据分析两个角度来看待这个问题。



1. 从数据收集角度看

数据收集就是把我们所需的数据都收集并且存储起来，用于后续的分析、挖掘。在这个阶段，原始数据本身就会存在很多问题。

1) 数据收集的渠道是多种多样的。从介质的角度来说，可以是街头发放的问卷、录制的音频、电子系统的log。对于问卷，需要用人工录入或者是OCR识别；对于录音，需要人工转录或者语音识别；对于电子系统，需要额外的前后端的开发工作。

2) 不同来源的数据有不同的存储方式。文本，excel，数据库（sql，nosql），同样带来了各式各样的数据类型和数据结构。

2. 从数据分析角度看

对于数据可视化，或机器学习、深度学习任务来说，数据的质量是尤为关键的。如果数据中包含着异常值、错误值，制作出来的数据图表很可能没有可读性；错误的数据会引入过多的噪声，或者是冗余的特征，导致了模型产生严重偏差。

另外，并不是所有正确、干净的数据都要全部展现在可视化当中。作为模型训练的输入，还需要对数据进行编码、归一化、特征构造和选取等进一步的操作。

由此可见，数据预处理在整个数据分析的流程当中是非常重要的。