数据科学基础 I
(Matlab)

东北大学

中国大学MOOC

数据分析流程

东北大学

数据科学的工作流程

确定分析目标　　数据加工与规整

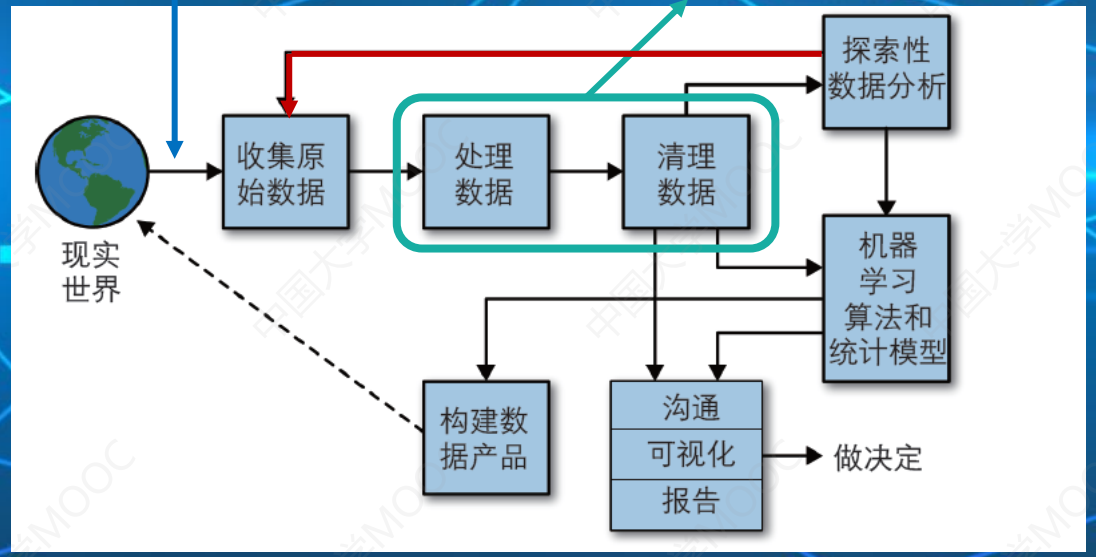

Image:《数据科学实战》

收集数据

数字化

定制化

数据加工与规整

脏数据

数据清洗

干净数据

乱数据

数据规整

规整数据

质量要求

计算要求

脏与乱的原因

- 重复（Duplicate Data）
- 不完备（Incomplete Data）
- 过时（Outdated Data）
- 不安全（Insecure Data）
- 错误（Inaccurate/Incorrect Data）
- 不一致（Inconsistent Data）
- 冗余（Too Much Data）

清洗/规整的方案

- 分析脏乱的原因
- 评估脏乱的影响

中国大学MOOC

探索性数据分析

Exploratory Data Analysis（EDA）是指对已有的数据在尽量少的先验假定下进行探索，并通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

Reference：Hoaglin,D.C. 1982. Exploratory data analysis. In encyclopedia of statistical sciences, Volume 2

模型选择与算法设计

- 以对消费者的建模为例：
  - 划分消费者群体：聚类，分类；
  - 购物篮分析：相关，聚类；
  - 购买额预测：回归，时间序列；
  - 满意度调查：回归，聚类，分类；

中国大学MOOC

数据驱动的成功案例

一级标题

什么是信息？

二级标题

数据科学

三级标题

文字：思源宋体 CN Heavy
思源黑体 CN Medium
思源黑体 CN Normal

主要用色

SOURCEHANSANSCN-MEDIUM.OTF

SOURCEHANSANSCN-NORMAL.OTF

SOURCEHANSERIFCN-HEAVY(1)(1).OTF

中国大学MOOC