# 一小时带你了解数据清洗

## 是什么玩意儿

菊安酱

2019.4.11

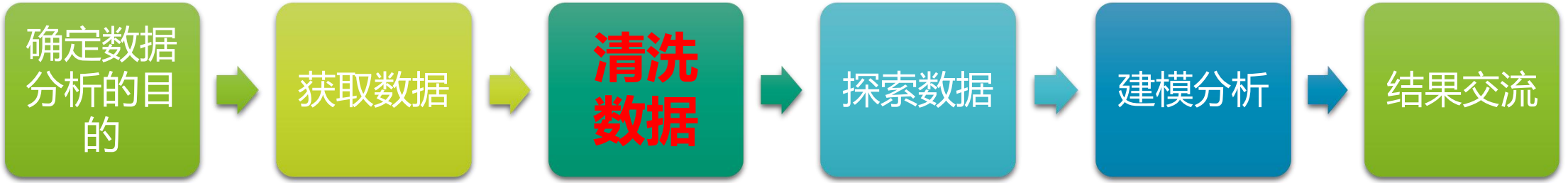
# 目录

- 什么是数据清洗
- 数据清洗的流程有哪些
- 常用的数据清洗方法
- 去哪儿网文本数据清洗案例

# 什么是数据清洗?

在讲解数据清洗前，我们先来看一看数据分析的整个流程

**> 数据分析流程:**




明确的分析目的就像是瞄准了正确的靶心，才能使后续的动作有意义

收集原始数据，数据来源可能是丰富多样的，格式也可能不尽相同

理顺杂乱的原始数据，并修正数据中的错误，这一步比较繁杂，但确是整个分析的基石

进行探索式分析，对整个数据集有个全面的认识，以便后续选择何种分析策略

常常用到机器学习、深度学习等算法

使用报告、图表等形式展现出来，与他人交流

# 通俗的来讲数据清洗~

我们把数据分析整个流程比作一个做菜的过程

确定目标：
麻辣香锅

# 通俗的来讲数据清洗~

我们把数据分析整个流程比作一个做菜的过程

获取原材料

**蔬菜类：** 青菜、白菜、芹菜、菠菜、胡萝卜、青笋、莲藕、西兰花、土豆、红薯
**干货类：** 木耳、腐竹
**海鲜类：** 鱿鱼、虾
**肉类：** 五花肉、午餐肉

然后你就到菜市场去买菜了……

然而这些菜并不能直接入锅呀……

So?









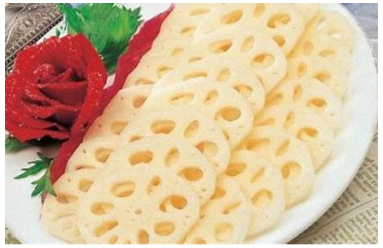



入锅之前最重要且最繁琐的一道工序来啦！！！

择菜、洗菜、切菜三部曲

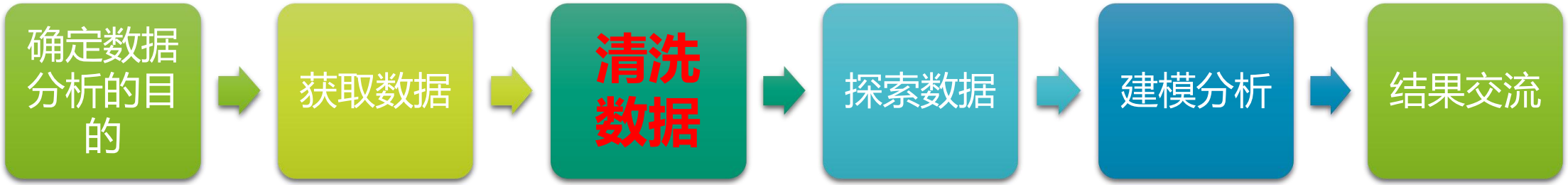





接下来，你就可以准备入锅炒菜啦~

# 数据分析 vs 做饭

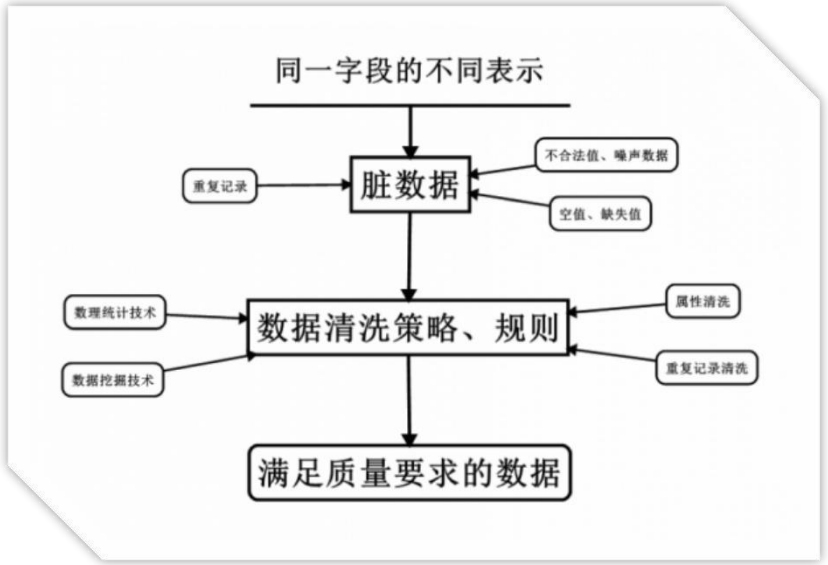

确定数据分析的目的
获取数据
清洗数据
探索数据
建模分析
结果交流


确定做麻辣香锅
菜市场购买食材
择菜、洗菜、切菜
炒菜
拍照、发朋友圈

# 那.....什么是数据清洗?

- 维基百科定义：数据清洗是从记录表、表格、数据库中检测、纠正或删除损坏或不准确记录的过程。

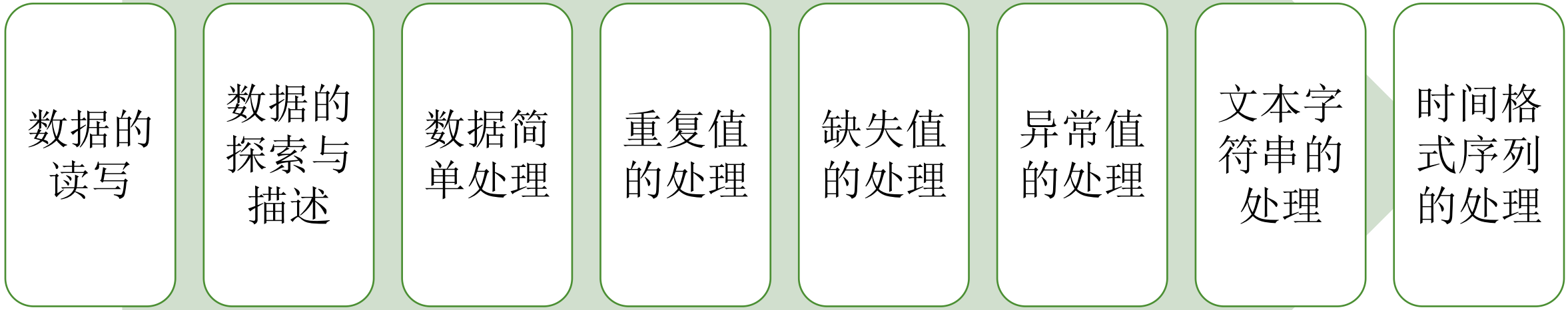
- 简单来说，数据清洗就是把“脏数据”变为“干净的数据”。




脏数据：
残缺数据、错误数据、重复数据、不符合规则的数据.....

干净的数据：
可以直接带入模型的数据

# 数据清洗流程




可根据实际情况调整顺序

# 数据清洗常用方法

**数据的读写**
- pd.read_csv('文件路径')
- pd.read_excel('文件路径')

**数据的探索与描述**
- df.info()
- df.describe()

**数据简单处理**
- 去除数据间的空格
- 英文字母大小写的转换

**重复值的处理**
- duplicated()
- drop_duplicates()

**缺失值的处理**
- 删除缺失值
- 均值填补法
- 向前填充/向后填充
- 模型填补法，如随机森林

**异常值的处理**
- 删除异常值的记录
- 作为缺失值处理
- 平均值修正、盖帽法修正
- 不处理：业务分析挖掘价值

**文本字符串的处理**
- 去除前后空格处理
- 处理中间有，（）之类的数据：replace(',','')
- 正则表达式提取所需数据

**时间格式序列的处理**
- 将系统时间格式化
- 系统时间和时间戳相互转换
- 年月日的提取

# 文本清洗案例

# Thank you~

菊安酱
2019.4.11