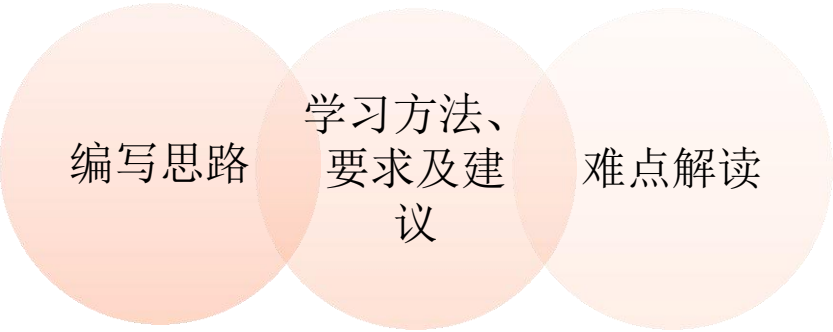


第5章 数据处理



编写思路 学习方法、
要求及建
议 难点解读

1.本章定位与内容简介



5.1 数据科学与数据加工

5.2 探索型数据分析

5.3 数据大小及标准化

5.4 缺失数据及其处理方法

5.5 噪声数据及其处理

5.6 数据维度及其降维处理

5.7 数据脱敏及其处理

5.8 数据形态及其规整化方法

5.9 Python 编程实践

5.10 继续学习本章知识

习题

2.本章学习提示及要求

了解

- 数据加工在数据科学中的重要地位
- 大数据环境下的数据加工的新含义和新要求

理解

- 探索型数据分析方法
- 规整数据的概念及基本原则

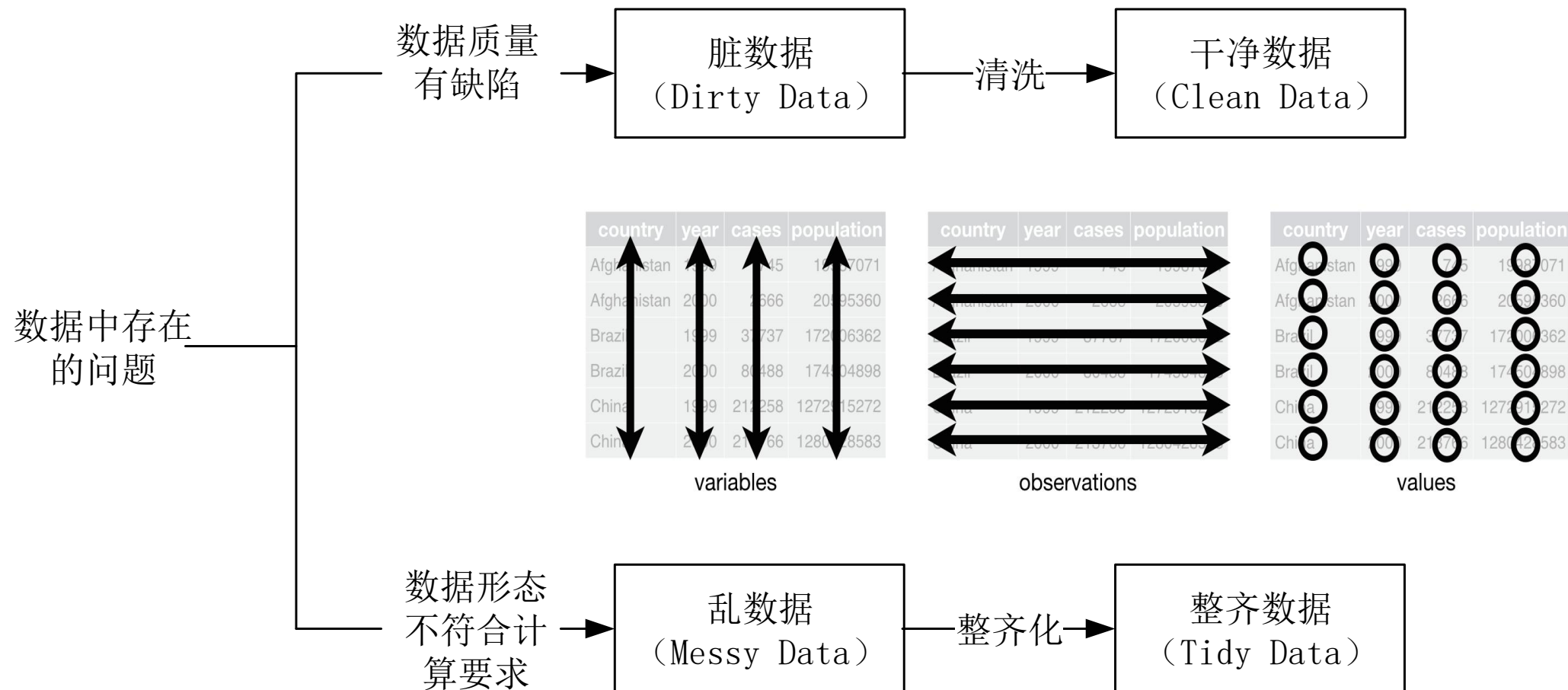
掌握

- 数据大小及其标准化
- 缺失数据及其处理方法
- 噪声数据及其处理方法
- 数据降维及其处理方法
- 数据脱敏及其处理方法
- 数据形态及其规整化方法

熟练掌握

- 基于Python的数据加工方法

数据加工 (wrangling or munging)



3.规整数据 (Tidy Data)

■ Data tidying

- structuring datasets to facilitate analysis.

■ Tidy Data的三个基本原则

- (1) 每个观察占且仅占一行。
- (2) 每个变量占且仅占一列。
- (3) 每一类观察单元构成一个关系（表）。

country	year	cases	population
Afghanistan	1999	1845	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	17206362
Brazil	2000	80488	174604898
China	1999	212258	1272015272
China	2000	213766	1280423583

variables

country	year	cases	population
Afghanistan	1999	1845	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	17206362
Brazil	2000	80488	174604898
China	1999	212258	1272015272
China	2000	213766	1280423583

observations

country	year	cases	population
Afghanistan	99	1845	19987071
Afghanistan	00	2666	20095360
Brazil	99	31737	17206362
Brazil	00	80488	174604898
China	99	212258	1272015272
China	00	213766	1280423583

values



Journal of Statistical Software

August 2014, Volume 59, Issue 10.

<http://www.jstatsoft.org/>

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible.

This paper introduces the tidy data framework, which provides a simple and effective way to clean data.

The tidy data framework is based on the principle that data should be organized in a way that is easy to understand and use.

The tidy data framework is based on the principle that data should be organized in a way that is easy to understand and use.

The tidy data framework is based on the principle that data should be organized in a way that is easy to understand and use.

4.基于Python的数据加工常用方法

Pivot

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t



```
df.pivot(index='foo',  
          columns='bar',  
          values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

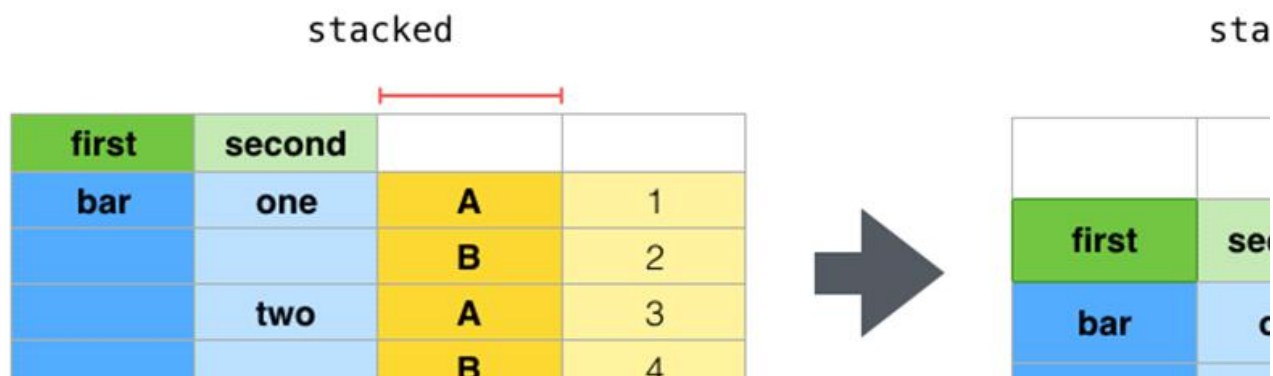
4.基于Python的数据加工常用方法



多级索引

多级索引

4.基于Python的数据加工常用方法



多级索引

多级索引

4.基于Python的数据加工常用方法

id类列+variable列+value列
未加入标识列的所有列名放在variable下

Melt

df3

	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150



df3.melt(id_vars=['first', 'last'])

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

5.如何继续学习本章知识

数据加工的动机

- 数据质量要求
- 数据计算要求

数据加工的应用

- 多种方法的综合运用
- 不同方法并非正交

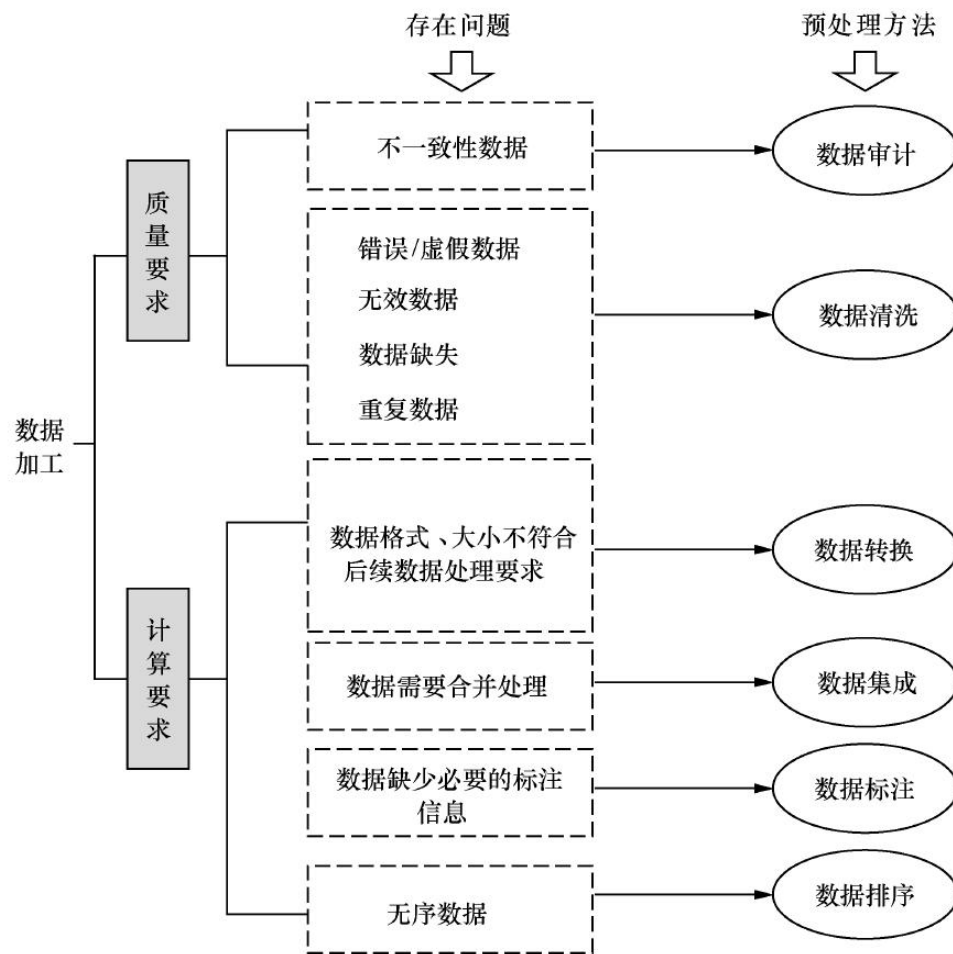


图 5-14 数据预处理方法



小结



1.本章定位与内容简介

2.本章学习提示及要求

3.规整数据（Tidy Data）

4.基于Python的数据加工常用方法

5.如何继续学习本章知识