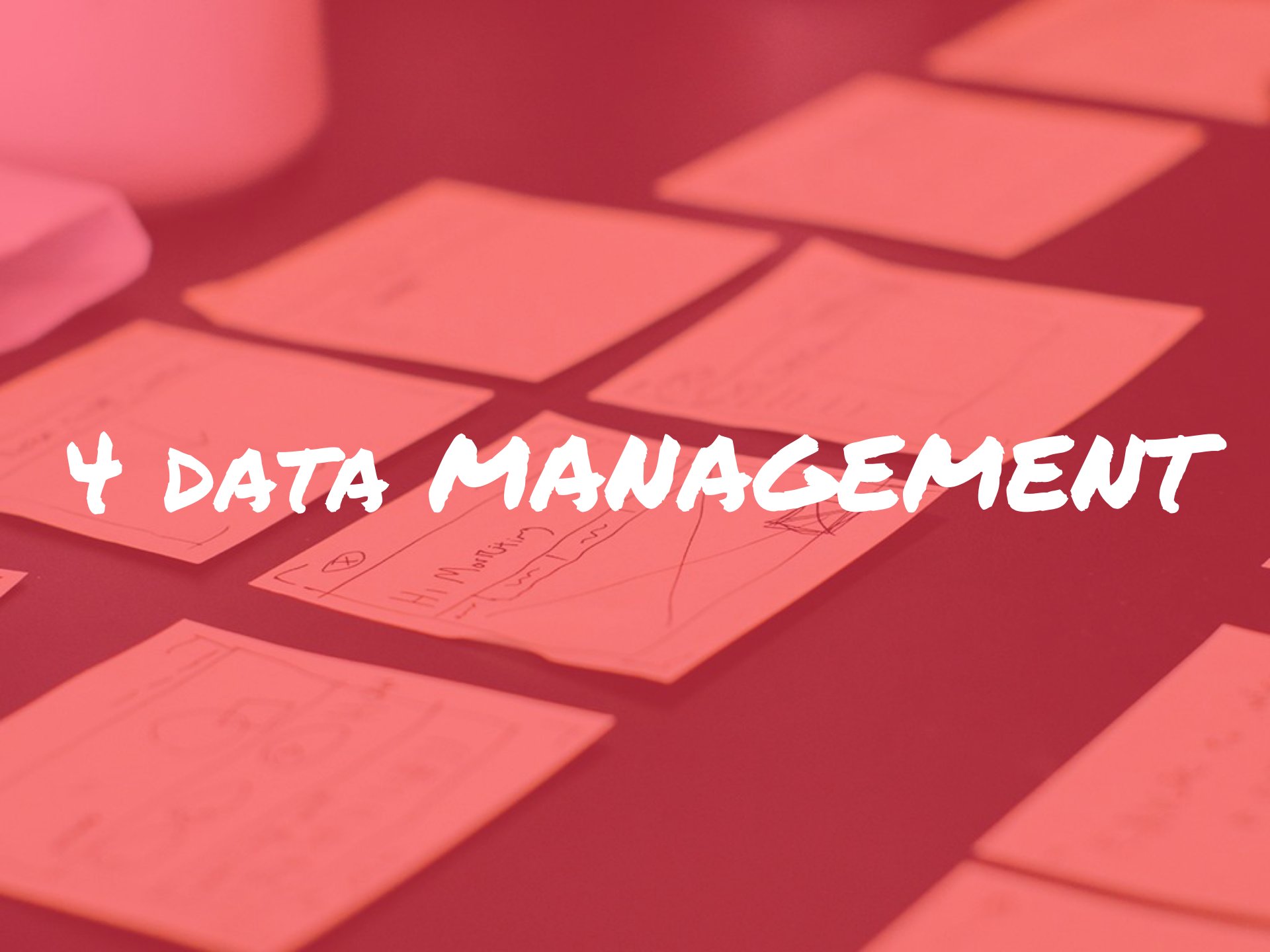

4 DATA MANAGEMENT

數據清洗 迷思

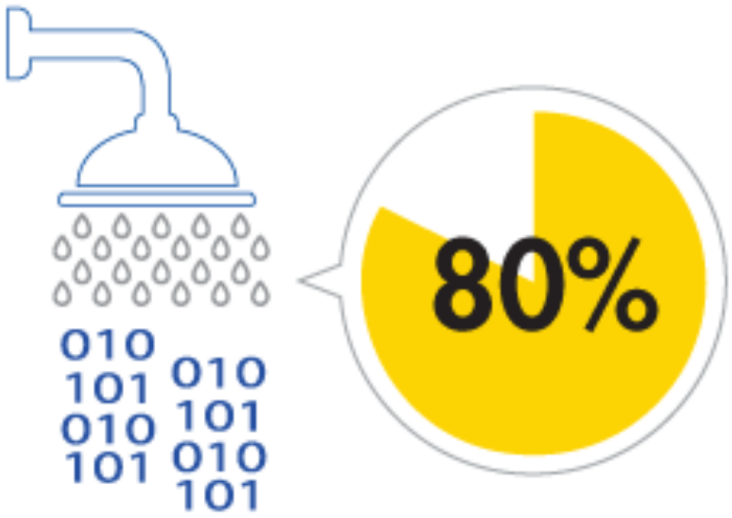
數據工程師的事

很無聊

只是寫code

數據清洗

指刪除、更正數據中錯誤、不完整、格式有誤或多餘的數據



瑣碎又讓人頭痛的過程

非常* ∞重要的過程
是項目品質的一个重要步驟

面對處理過完整乾淨的數據
才可在數據分析中玩出許多有趣的東西

Feature Engineering

- Feature Engineering -> Machine Learning
- Feature Learning -> Deep Learning (images/videos/audio/text)
  Train deep neural networks to extract features.

表格資料

Tabular data are in the form of a table,
feature columns of numeric / categorical / string type

1. Int / float: directly use or bin to unique int values

2. Categorical data: one-hot encoding
   - Map rare categories into "Unknown"

fish	cat	mouse	dog	others
[0,	1,	0,	0,	0]
[0,	0,	0,	1,	0]

3. Date-time: a feature list such as
   - [year, month, day, day_of_year, week_of_year, day_of_week]

4. Feature combination: Cartesian product of two feature groups
   - [cat, dog] x [male, female] -> [(cat, male), (cat, female), (dog, male), (dog, female)]

1 創建新變量

算術運算符号

運算符號	描述
+	加
-	減
*	乘
/	除
^或**	求冪
x%%y	求餘。5%%2結果為1

2 重新編碼

重編碼涉及根據同一個變量或從其他變量的現有值創建新值的過程

① 將一個連續型變量修改為一組類別值
② 將誤編碼的值替換為正確值
③ 基於一組分數線創建一個表示及格/不及格的變量

邏輯預算符號

運算符號	描述
<	小於
<=	小於或等於
>	大於
>=	大於或等於
==	等於
!=	不等於
!x	非x
x | y	x或y
X & y	X和y

2-1 Example

男性和女性在領導方式上的不同

典型的問題如下

- 處於管理崗位的男性和女性在聽從上級的程度上是否有所不同？

- 這種情況是否依國家的不同而有所不同，或者說這些由性別導致的不同是否普遍存在？

解答這些問題的一種方法是讓多個國家的經理人的上司對其服從程度打分，使用的問題類似於

這名經理再做出決策之前，會詢問我的意見				
1	2	3	4	5
非常不同意	不同意	不同意 也不反對	同意	非常同意

各行數據代表某個經理人的上司對他的評分

表 領導行為的性別差異

	manager	date	country	gender	age	q1	q2	q3	q4	q5
1	1	02/24/18	US	M	32	5	4	5	5	5
2	2	03/28/18	US	F	45	3	5	2	5	5
3	3	01/1/18	UK	F	25	3	5	5	5	2
4	4	03/12/18	UK	M	39	3	3	4	NA	NA
5	5	01/01/18	UK	F	99	2	2	1	2	2

- 每位經理人的上司根據與服從權威相關的五項陳述（q1到q5）對經理人進行評分
- 日期一欄記錄了進行評分的時間

- 例如
① 經理人1是一位在美國工作的32歲男性，上司對他的評價是慣於順從
② 經理人5是一位在英國工作的，年齡未知（99可能代表缺失）的女性，服從程度評分較低

POINT

一個數據集中可能含有幾十個變量和成千上萬的觀測，但為了簡化示例，我們僅選取了5行10列的數據。另外，已將關於經理人服從行為的問題數量限制為5。

在現實的研究中，很可能會使用10到20個類似的問題來提高結果的可靠性和有效性。

5-2 QUIZ

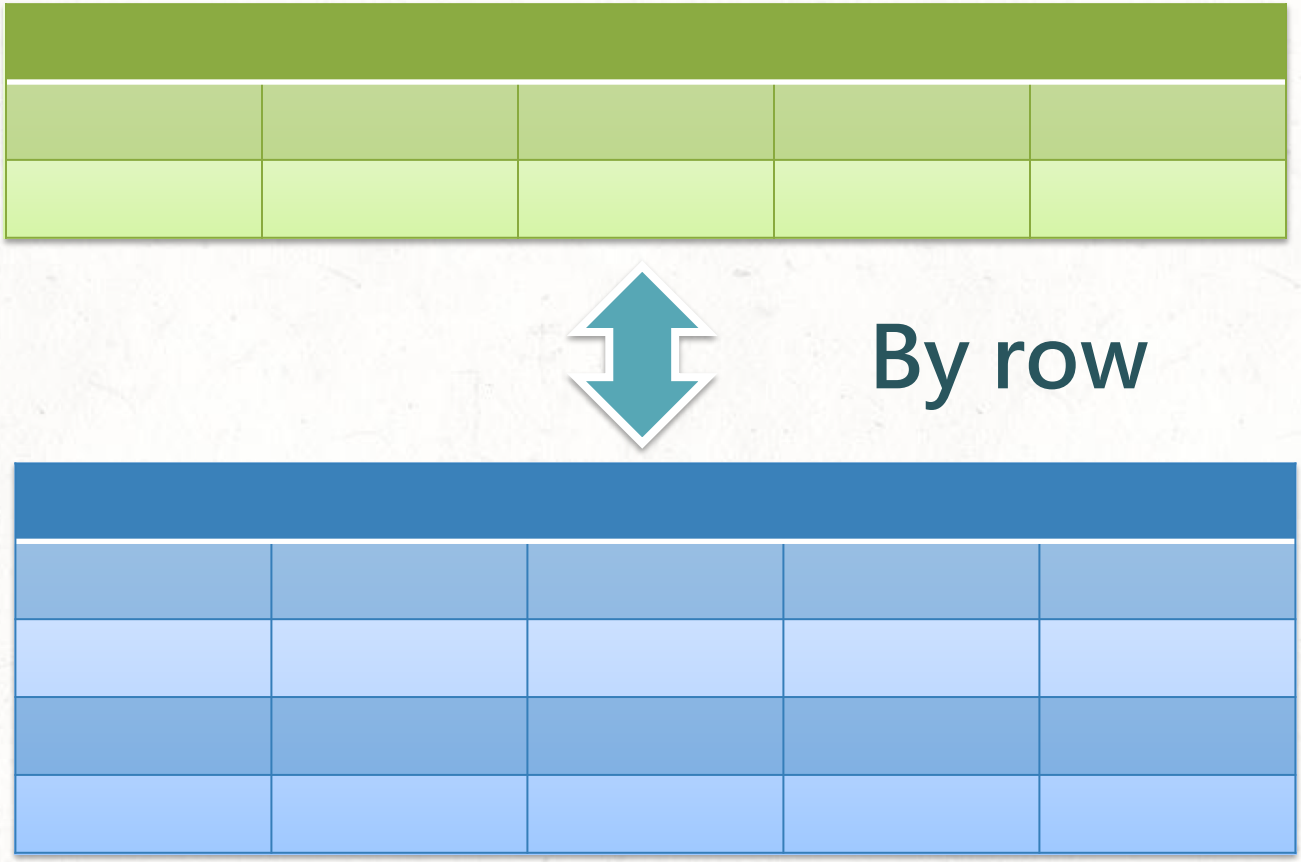
"hsb" contains 200 observations from a sample of high school students with demographic information about the students

Variable	Position	Label	Value	Label
id	1			
female	2		.00 1.00	Male Female
race	3		1.00 2.00 3.00 4.00	Hispanic Asian african-amer White
ses	4		1.00 2.00 3.00	Low Middle High
schtyp	5	type of school	1.00 2.00	Public private
prog	6	type of program	1.00 2.00 3.00	general academic vocation
read	7	reading score		
write	8	writing score		
math	9	math score		
science	10	science score		
socst	11	social studies score		

such as their gender (female), socio-economic status (ses) and ethnic background (race). It also contains a number of scores on standardized tests, including tests of reading (read), writing (write), mathematics (math) and social studies (socst).

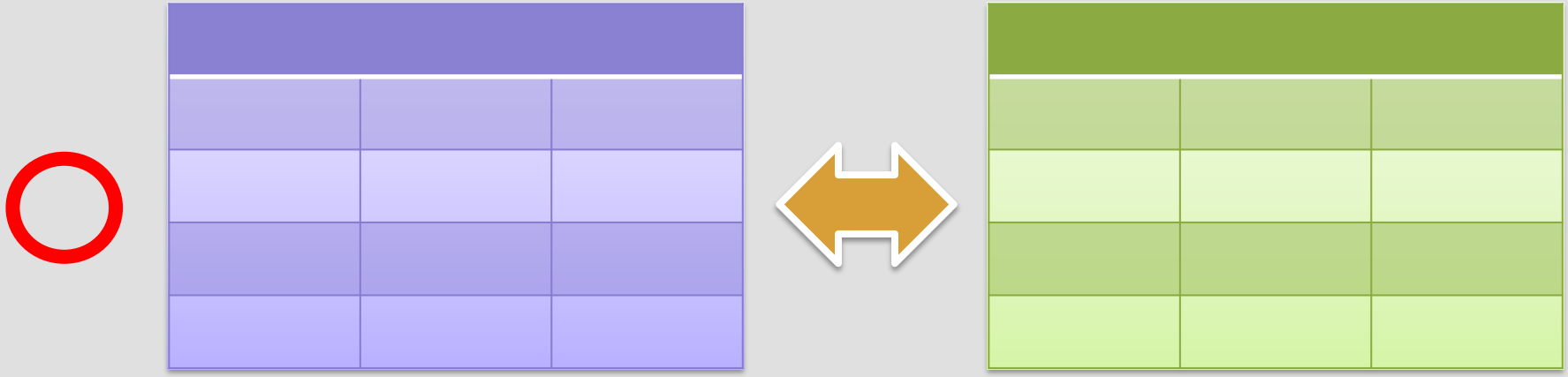
6. Merging Data

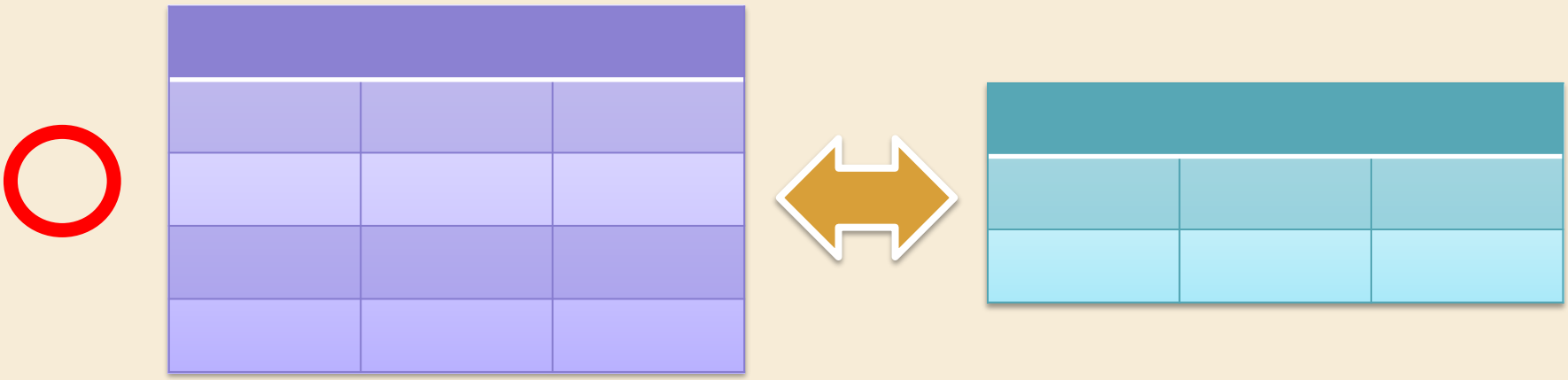
6-1 rbind()



- ✓Make sure the number of columns match.
- ✓The names and classes of values being joined must match.

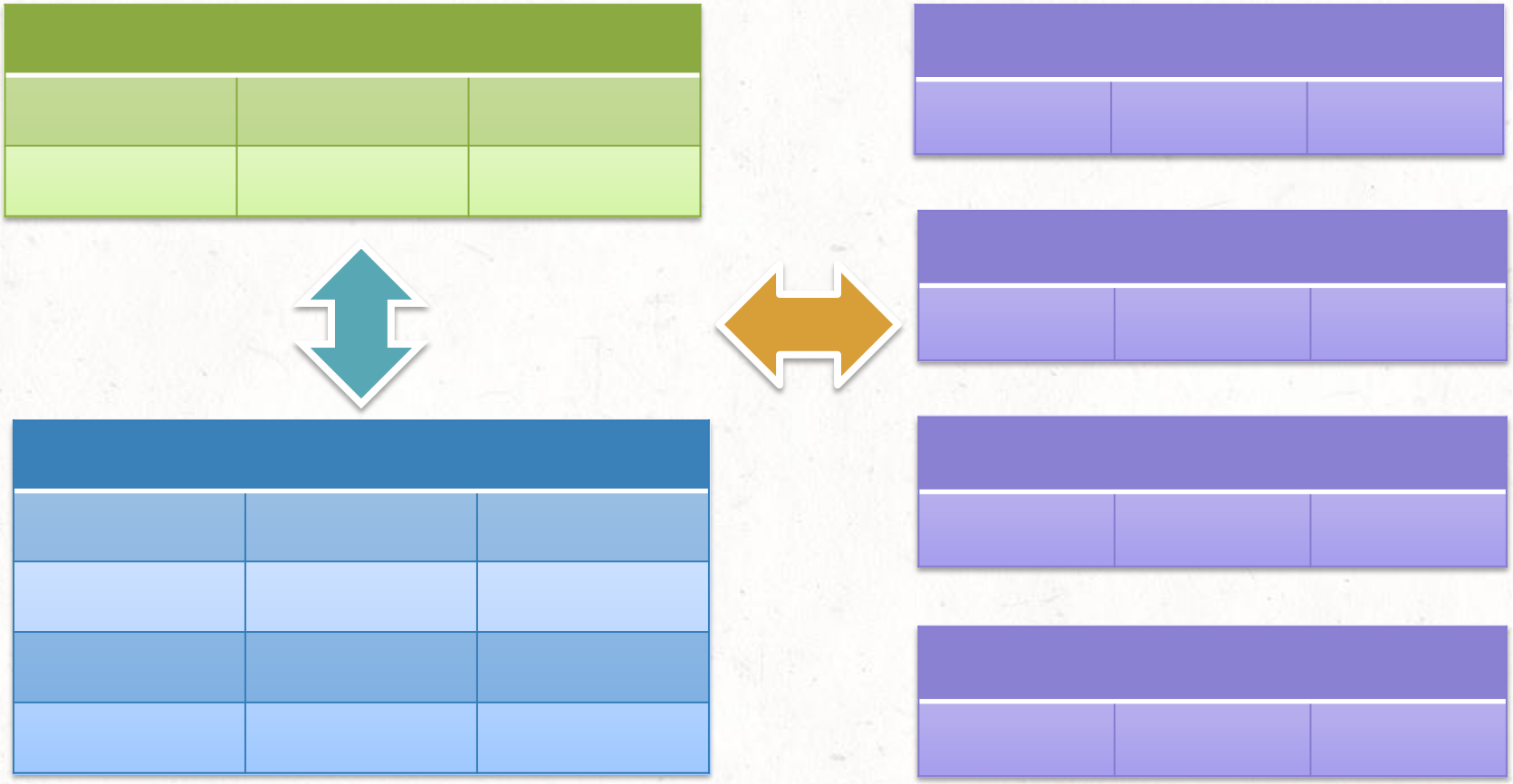
6-2 cbind()





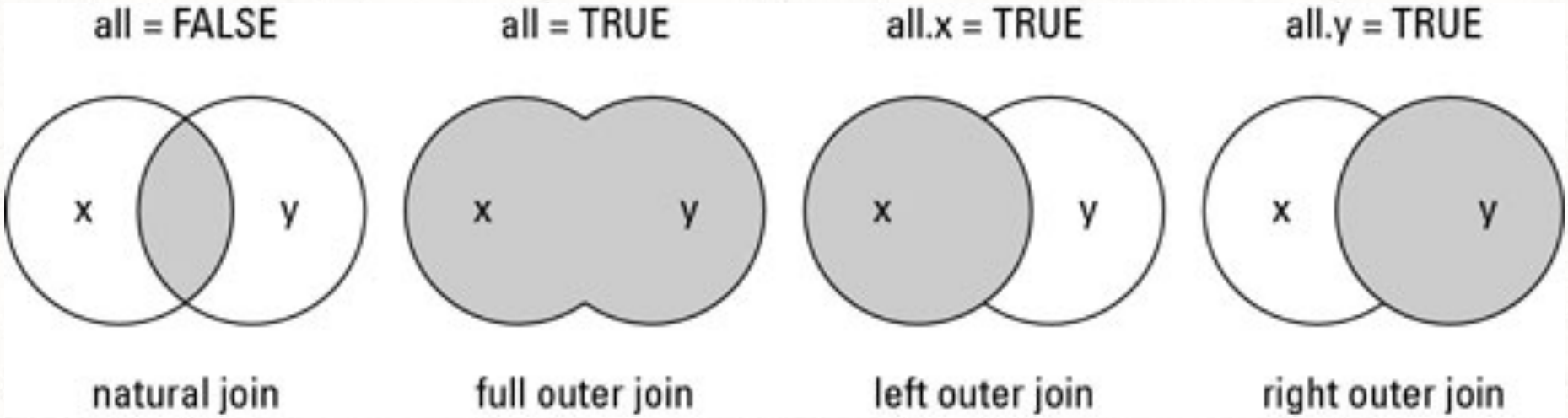
✓cbind does not require matching heights; if one data frame is shorter

6-3 rbind() & cbind()



WARNING: cbind does not require matching heights; if one data frame is shorter, it will recycle it.

6-4 merge()


all = FALSE
x
y
natural join
all = TRUE
x
y
full outer join
all.x = TRUE
x
y
left outer join
all.y = TRUE
x
y
right outer join


Thanks!