

Extract-Transform-Load

Yishih Chung

National Chiao Tung University

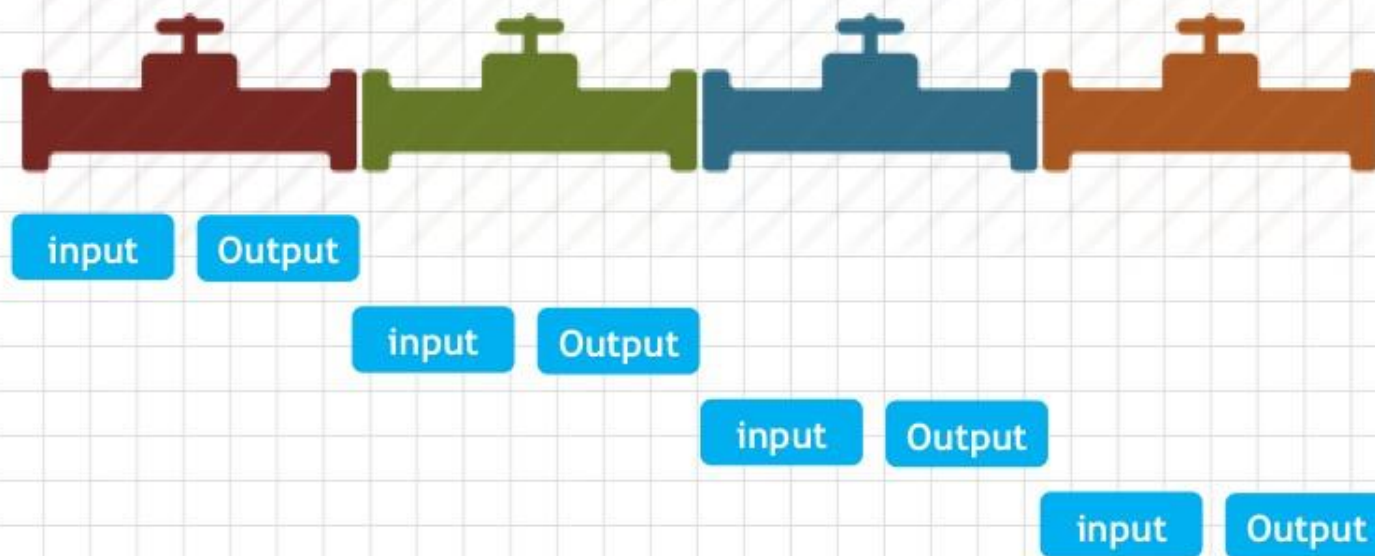
Data wrangling

- The process of transforming and mapping data from one “raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
- 萃取 (extract)、轉置 (transform)、載入 (load)至目的端
- 進行資料分析/科學之前最重要/浪費花時間的工作
- 常用套件
 - reshape2, dplyr, tidyr, data.table
 - 套件安裝：install.packages(c("dplyr", "reshape2"))
 - 套件使用：library(pkg_name)

欄位名稱	代號	欄位名稱	代號
日期	date	車輛數標準差	std.sbi
時間	hour	平均空位數	avg.bemp
場站代號	sno	最大空位數	max.bemp
場站區域	sarea	最小空位數	min.bemp
場站名稱	sna	空位數標準差	std.bemp
經度	lat	氣溫	temp
緯度	lng	最高溫	max.temp
總停車格	tot	最低溫	min.temp
平均車輛數	avg.sbi	溼度	humidity
最大車輛數	max.sbi	氣壓	pressure
最小車輛數	min.sbi		

Pipeline

- Pipeline: input & output chain




$f(x, \dots) \%>\% f(x, \dots) \%>\% f(x, \dots) \%>\% f(x, \dots)$

常用dplyr指令

- `filter` 對列做篩選
- `select` 對欄做篩選
- `mutate` 更改欄或新增欄 `mutate`
- `arrange` 排列 `a->z` or `z->a`
- `group_by` + `summarise` 依照類別處理
- ^{column合併}`bind_cols`, `bind_rows` 單一資料源整併
- `merge`, `left_join` 多資料源合併

dplyr:: select

X1	X2	X3	X4	X5
V		V	V	
V		V	V	
V		V	V	
V		V	V	
V		V	V	
V		V	V	



X1	X3	X4
V	V	V
V	V	V
V	V	V
V	V	V
V	V	V
V	V	V

dplyr::filter

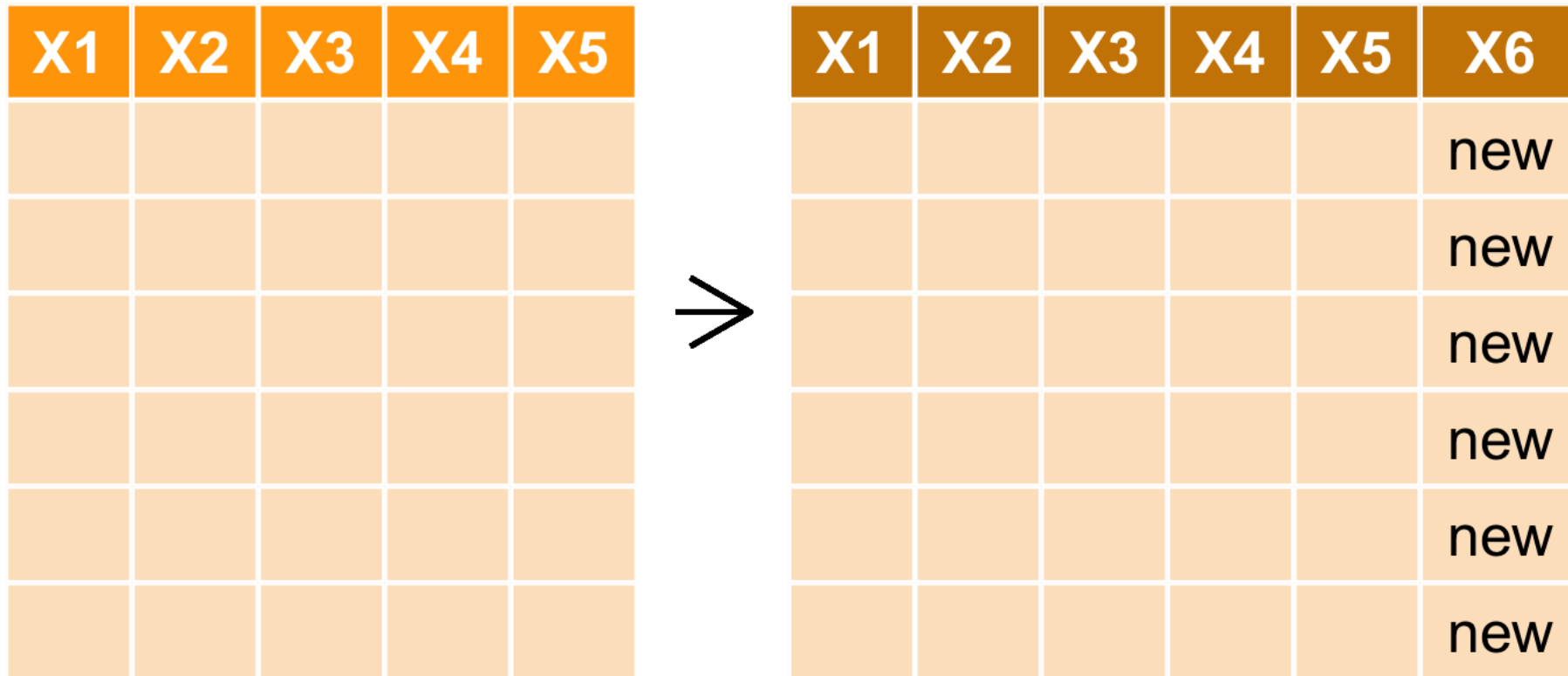
X1	X2	X3	X4	X5



文字

X1	X2	X3	X4	X5

dplyr::mutate



dplyr::group_by

資料分成三群

X1	X2	X3	X4	X5



X1	X2	X3	X4	X5



Dplyr::group_by + summarise

行政 dock數
東區 100
東區 80
東區 20
西區 50
西區 40
西區 10

行政 avg dock數
東區 100
西區 50

X1	X2	X3	X4	X5



X1	X2	X3	X4	X5

temp%>%group_by(行政)
%>%summarise(x1=mean(dock))




X1	X2	X3	X4	X5

分群後summarise 他

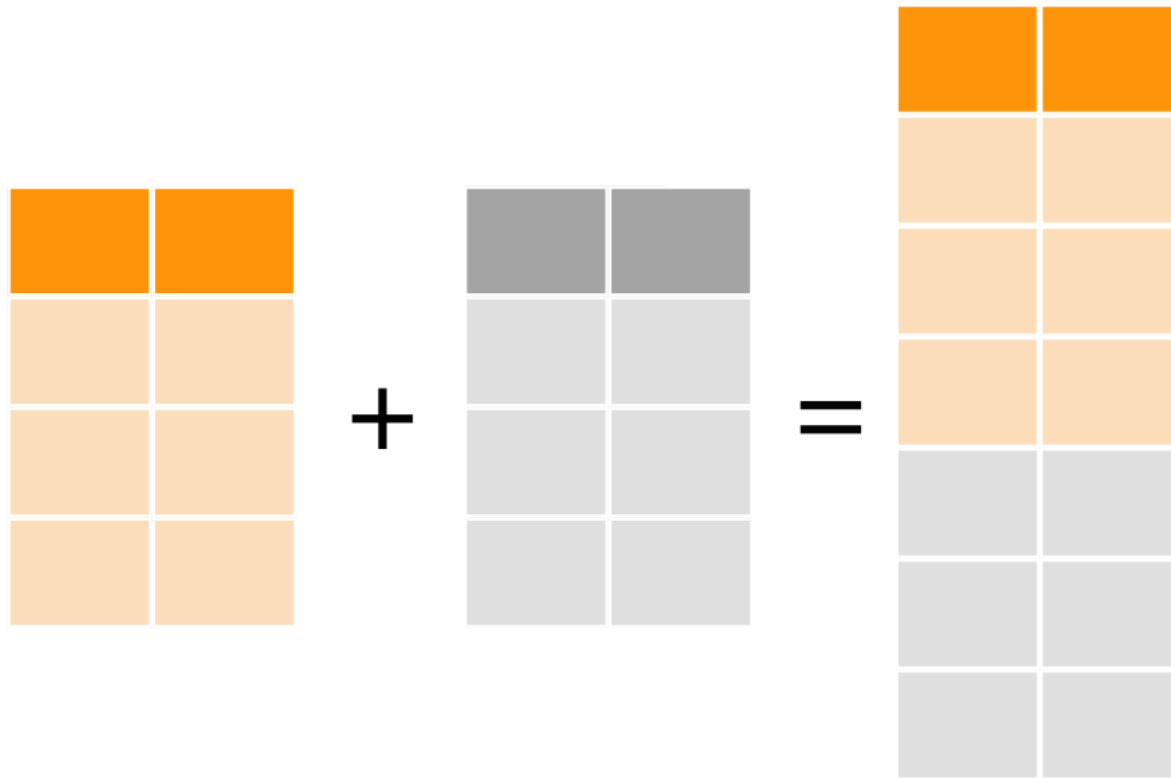
dplyr::arrange

X1	X2
3	Ben
4	Johnson
1	Rafe
2	Ning



X1	X2
1	Rafe
2	Ning
3	Ben
4	Johnson

`dplyr::bind_rows(a,b)`



`dplyr::bind_cols(a,b)`



library(reshape2)

- Long & Wide Data
- melt
 - wide format -> long format
- cast
 - long format -> wide format
 - dcast for data.frame
 - acast for vector, matrix and array

dcast formula `dcast(aql, month + day ~ variable, value.var = "value")`

ID variables (left side of formula)	Variable to swing into column names (right side of formula)	Values (value.var)
--	---	-----------------------

Long-format data

month	day	variable	value
5	1	ozone	41
5	2	ozone	36
5	3	ozone	12
5	4	ozone	18
5	5	ozone	NA
5	6	ozone	28

Wide-format data

month	day	ozone	solar.r	wind	temp
5	1	41	190	7.4	67
5	2	36	118	8.0	72
5	3	12	149	12.6	74
5	4	18	313	11.5	62
5	5	NA	NA	14.3	56
5	6	28	NA	14.9	66

In-class practice

1. 請幫忙找出信義區各車站八點車子最多的站
2. 在中和找一個下午三點風比較小的地點
3. 信義區的腳踏車站晴天和雨天的使用率有何差別：
 - filter、mutate、select、group_by、summarise
 - dcast
 - arrange