R 語言
tidyverse: dplyr & tidyr

王貿

國立臺灣大學行為與資料科學研究中心助理研究員

國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

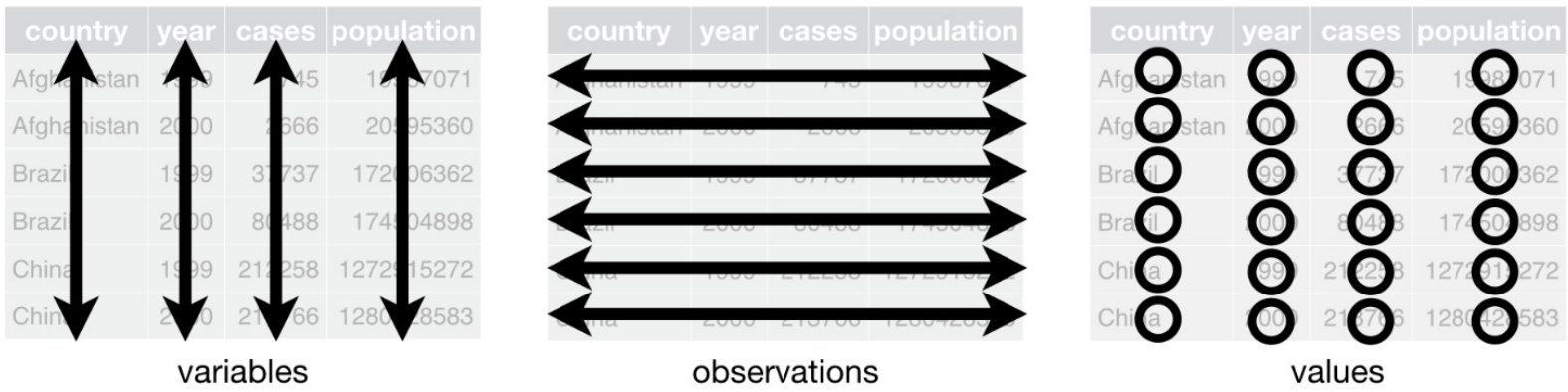
課程主題重點

- **dplyr**
  select, filter, group_by, summarize, mutate, arrange

- **dplyr**
  join
- **tidyr**
  gather (pivot_longer),
  spread (pivot_wider)

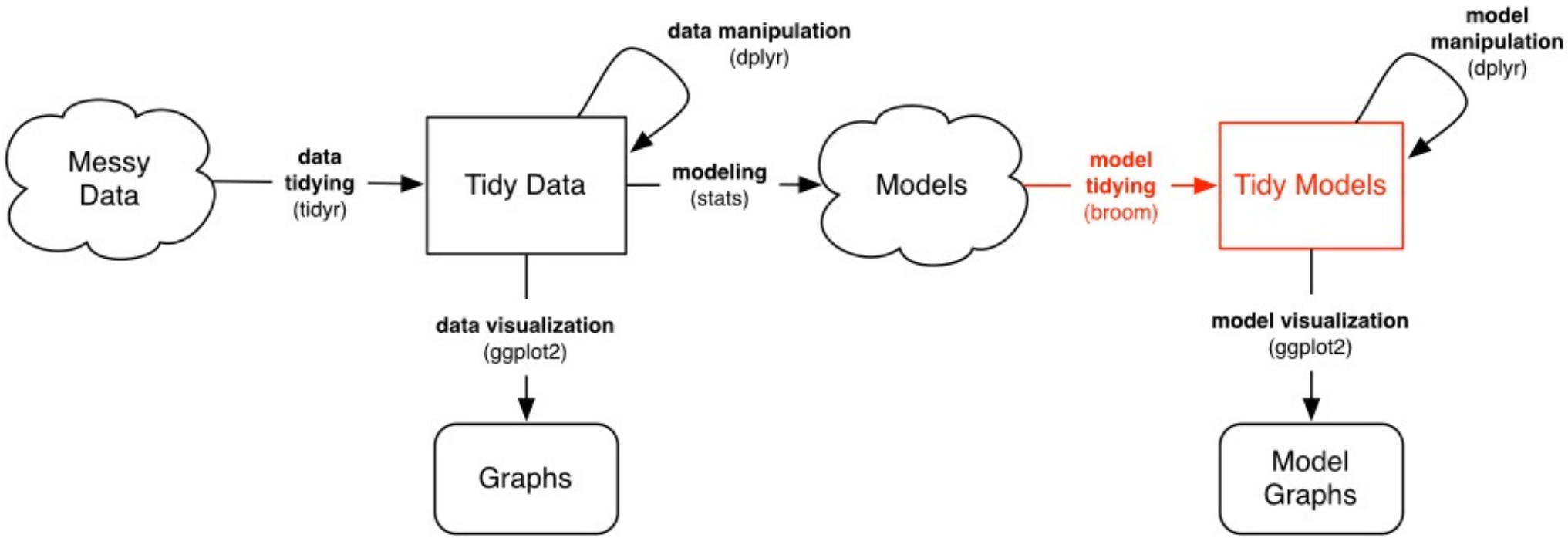
Tidy Data

1. Each **variable** must have its own **column**.
2. Each **observation** must have its own **row**.
3. Each **value** must have its own **cell**.



https://r4ds.had.co.nz/tidy-data.html#tidy-data-1

什麼時候資料會需要tidy？

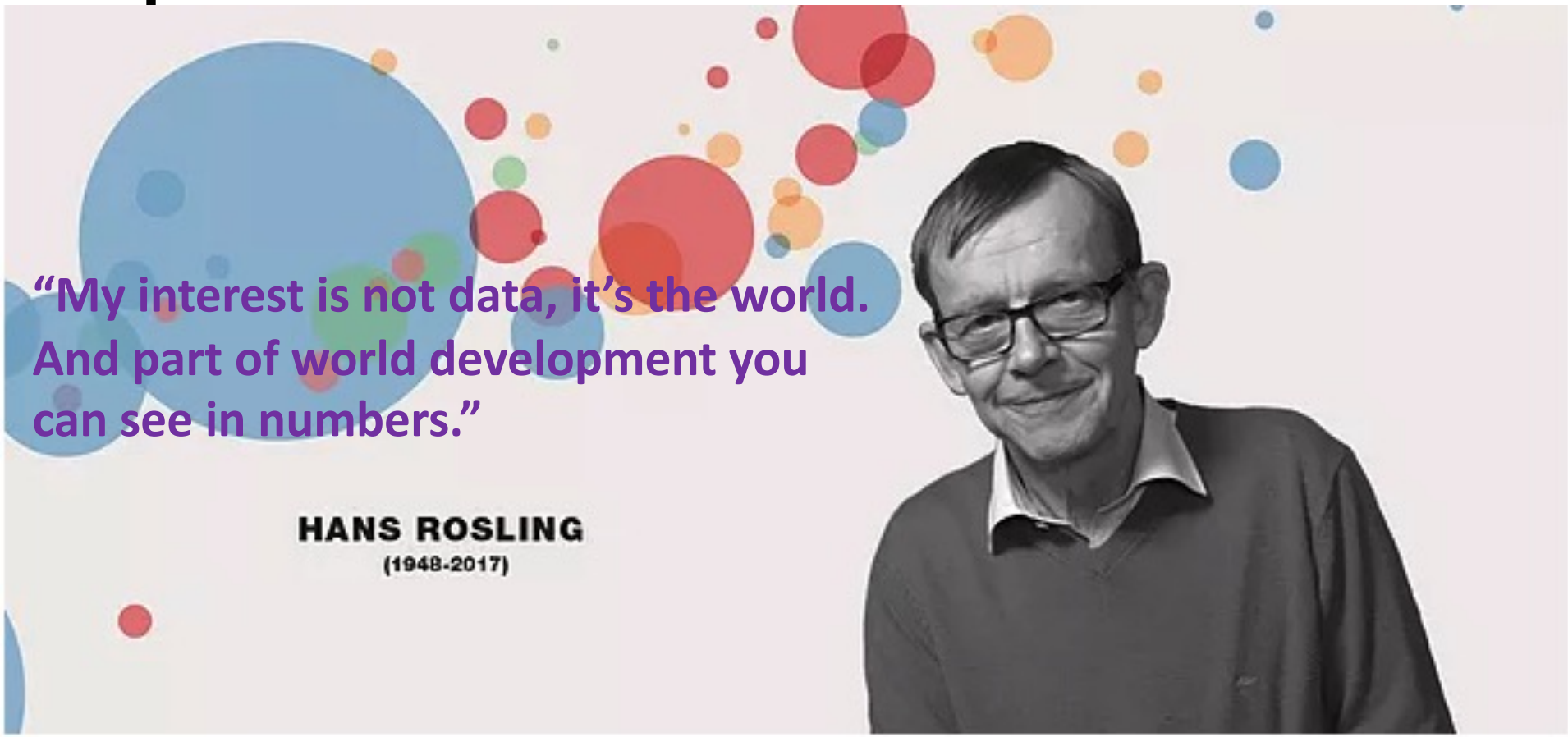


http://varianceexplained.org/r/tidy-genomics-biobroom/

Tidy function 的基本邏輯

- 英文**動詞**做為 function 名稱。
- 資料（data）都放在 function 的**第一個 argument 的位置**。
- 神奇的 **%>%**（pipe）**連接器**。
- 大量借用 SQL（資料庫語言）。

Gapminder



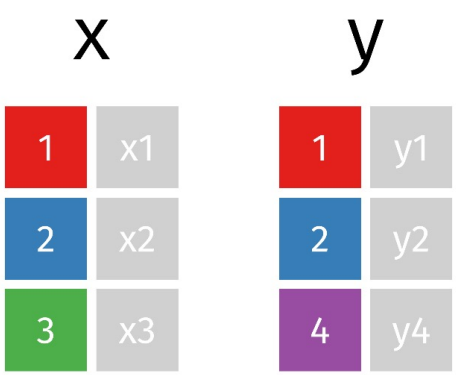
https://www.gapminder.org/tools/#$chart-type=bubbles

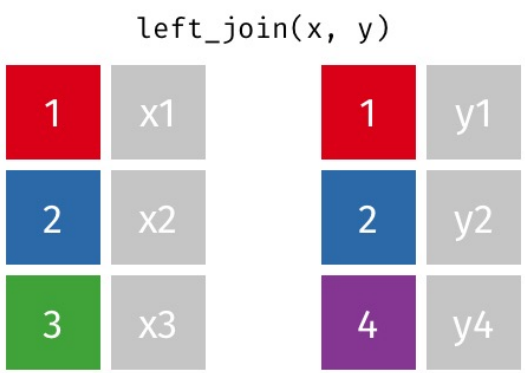
Tidy Animated Verbs

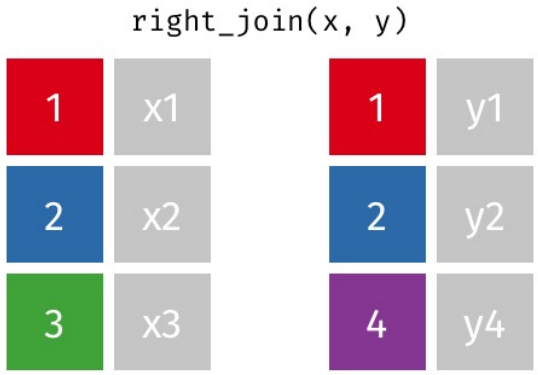
- https://github.com/gadenbuie/tidyexplain

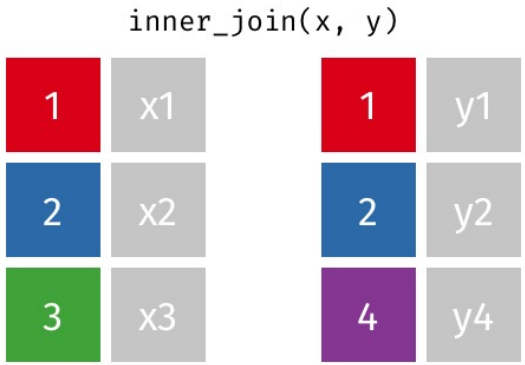
- **Mutating Joins**
- **Filtering Joins**
- Set Operations

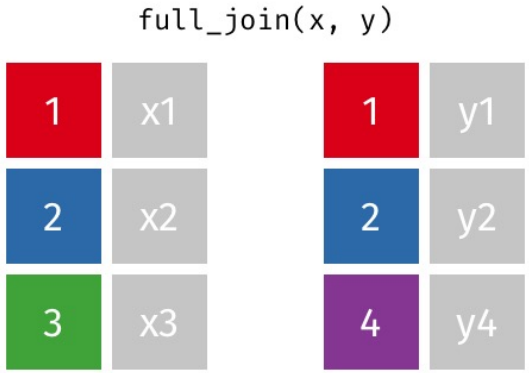
Mutating Joins











Filtering Joins

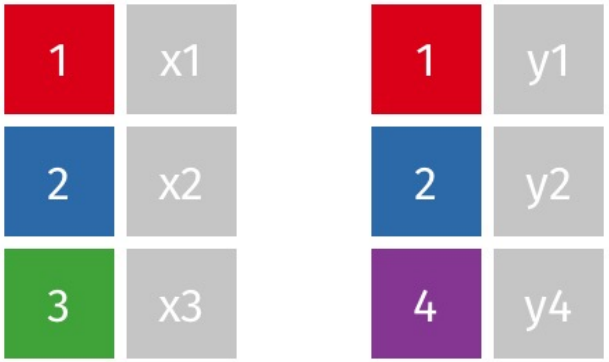
x

1	x1
2	x2
3	x3

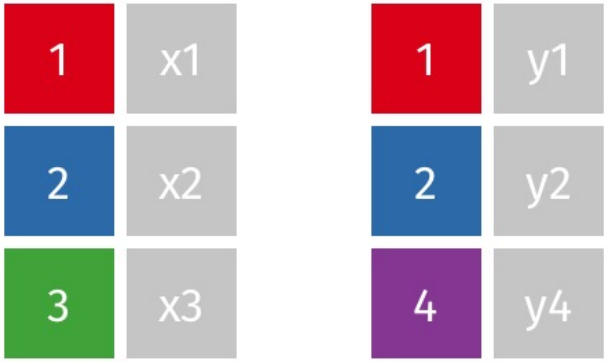
y

1	y1
2	y2
4	y4

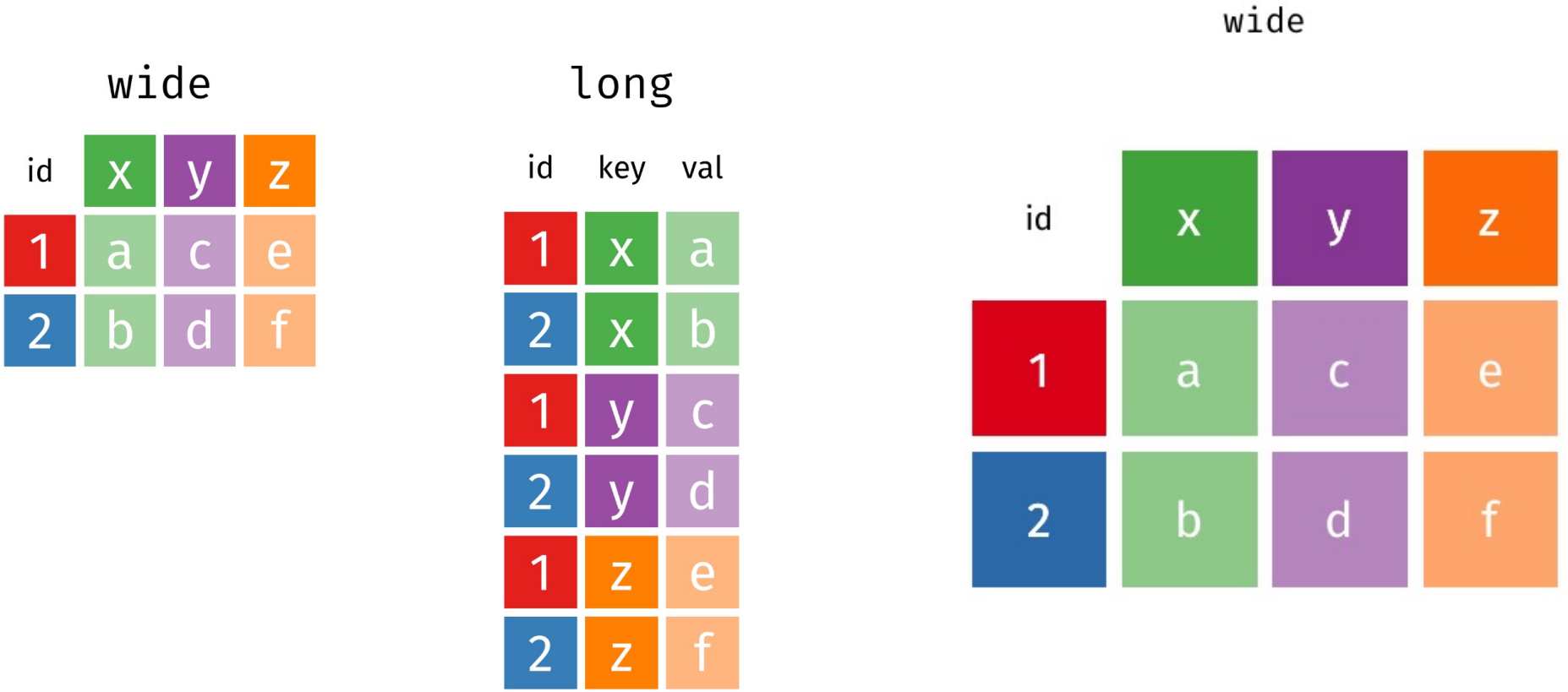
`semi_join(x, y)`



`anti_join(x, y)`



tidyr: gather (longer) & spread (wider)



tidyr: gather & spread

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

Wide format

Sample_ID	Ca	Mg	Na	Cl
P-1	234.3	12.3	4.3	33.5
P-2	432.2	22.3	2.4	12.3

Gather

Spread

Long format

Sample_ID	Key	Value
P-1	Ca	234.3
P-1	Mg	12.3
P-1	Na	4.3
P-1	Cl	33.5
P-2	Ca	432.2
P-2	Mg	22.3
P-2	Na	2.4
P-2	Cl	12.3