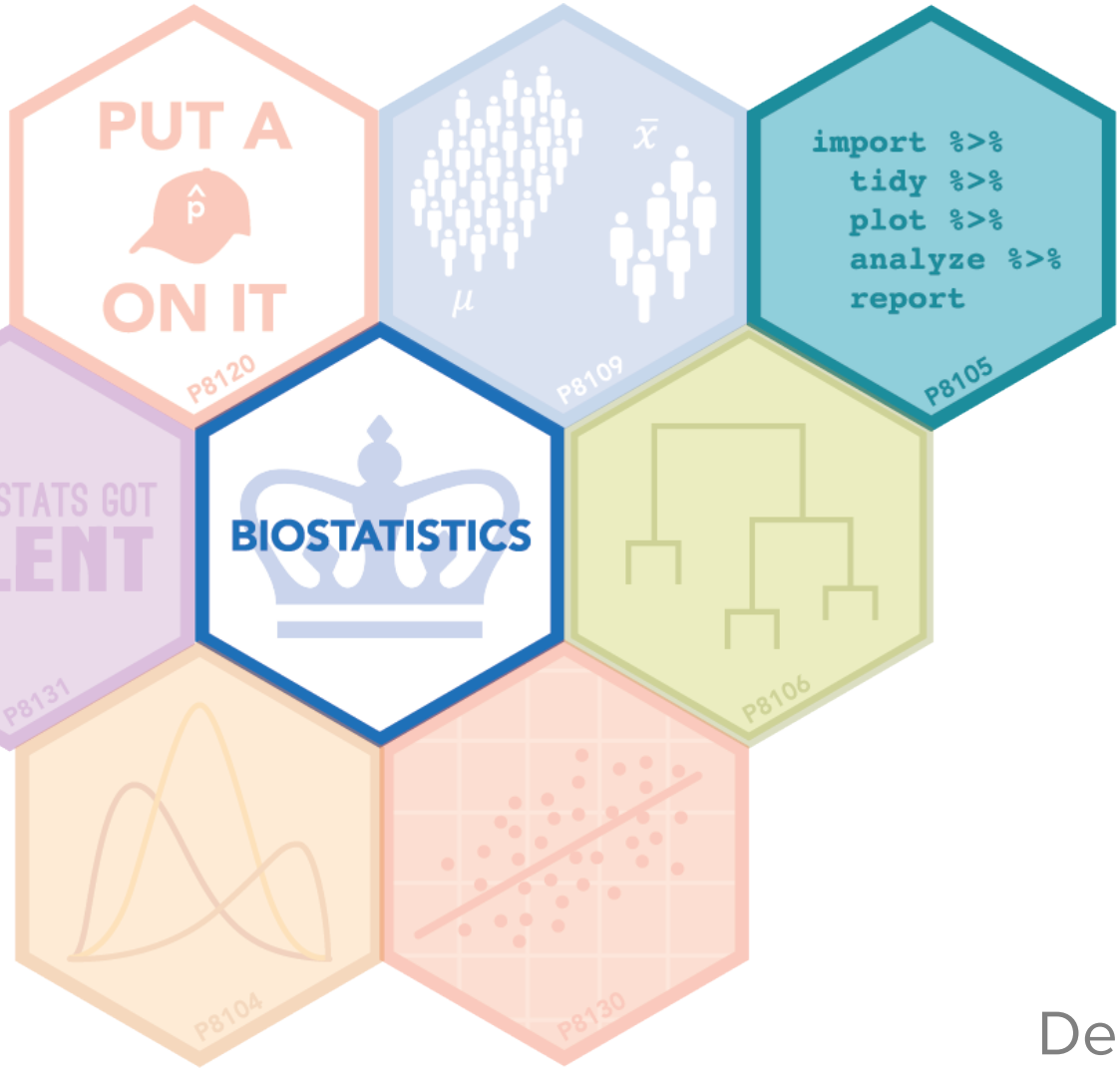
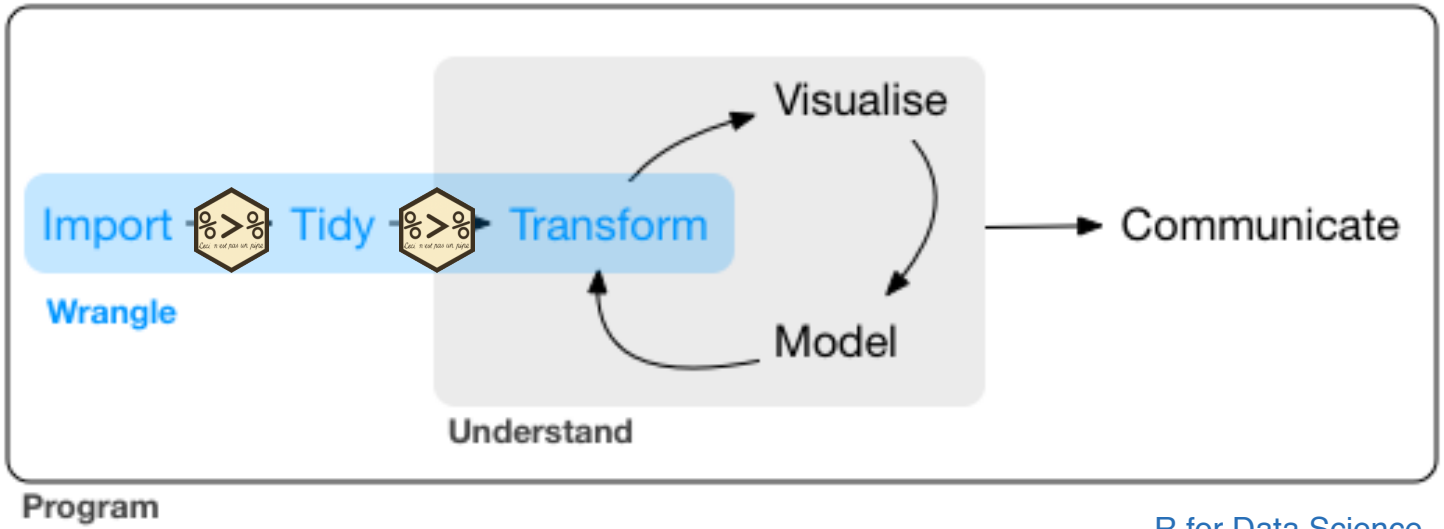


TIDY DATA

Jeff Goldsmith, PhD
Department of Biostatistics

Tidy data

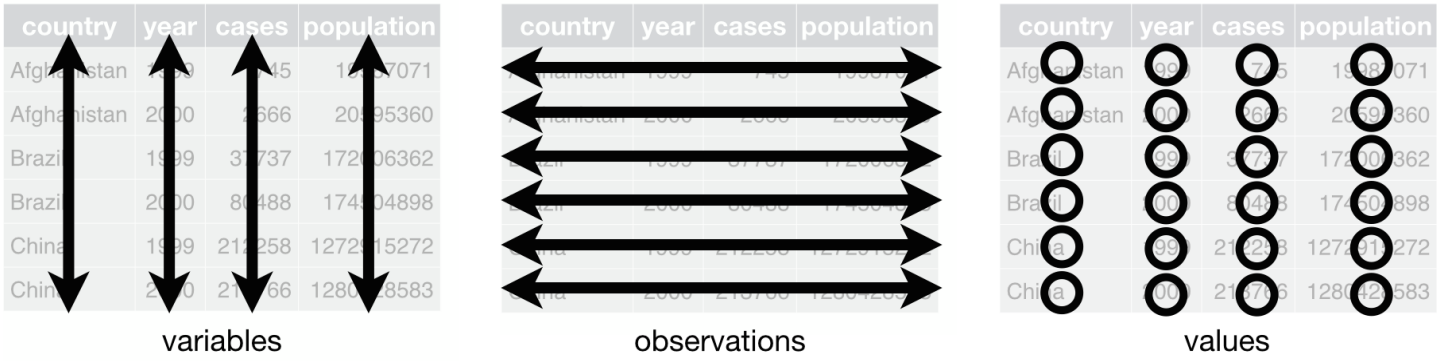
- "Middle" step in the wrangling process



R for Data Science

Rules for tidy data

- Data tables have an implied structure which the “tidy data” framework makes explicit
  - Columns are variables
  - Rows are observations
  - Every value has a cell



R for Data Science

Why tidy your data?

- Consistent data structures will simplify your thought process
  - Especially true if you use tools designed for tidy data
  - Sounds like something the “tidyverse” would help with…
- Data written for computers is easier to work with

Not all data are tidy

- Columns are values, not variable names
- Single columns contain multiple variables
- Data are stored in multiple tables

- Non-tidiness is sometimes (if only rarely) intentional
- Data written for humans is generally not tidy
  - Human readability is important, but should be a deliberate choice

- Some data aren't really amenable to tidiness
  - Genomics; neuroimaging

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

VS

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

"Tidy Data", H. Wickham, JSS

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

VS

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

"Tidy Data", H. Wickham, JSS

The Fellowship Of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

The Two Towers

Race	Female	Male
Elf	331	513
Hobbit	0	2463
Man	401	3589

The Return Of The King

Race	Female	Male
Elf	183	510
Hobbit	2	2673
Man	268	2459

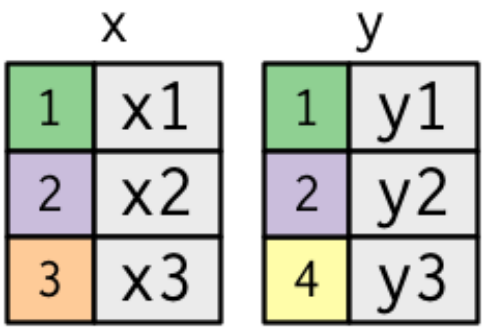
VS

Film	Gender	Race	Words
The Fellowship Of The Ring	Female	Elf	1229
The Fellowship Of The Ring	Male	Elf	971
The Fellowship Of The Ring	Female	Hobbit	14
The Fellowship Of The Ring	Male	Hobbit	3644
The Fellowship Of The Ring	Female	Man	0
The Fellowship Of The Ring	Male	Man	1995
The Two Towers	Female	Elf	331
The Two Towers	Male	Elf	513

https://github.com/jennybc/lotr-tidy/blob/master/01-intro.md

Relational data

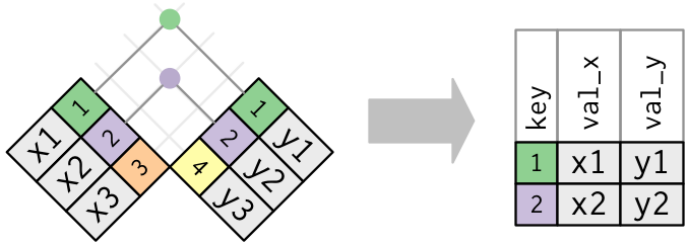
- Data spread across tables with defined relations
- Variables used to define these relations are keys
- Tables are combined by joins



Join types

- Joining datasets x and y

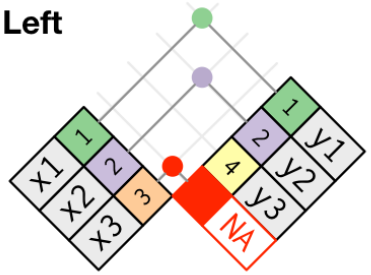
Inner joins



key	val_x	val_y
1	x1	y1
2	x2	y2

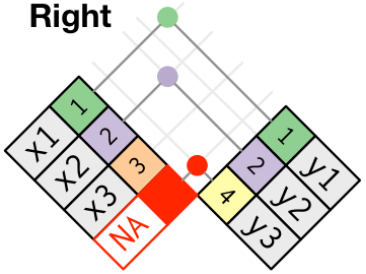
Outer joins

Left



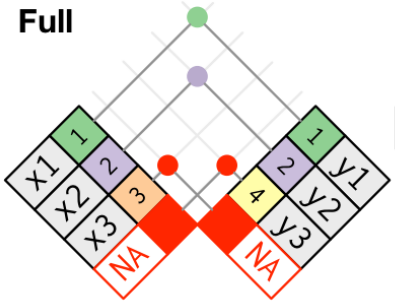
key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Right



key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Full



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

R for Data Science

Key functions

- For tidying single tables
  - `gather`
  - `separate`
- For untidying single tables
  - `spread`
- For combining multiple tables
  - `bind_rows`
  - `*_join`