PIVOT AND JOIN

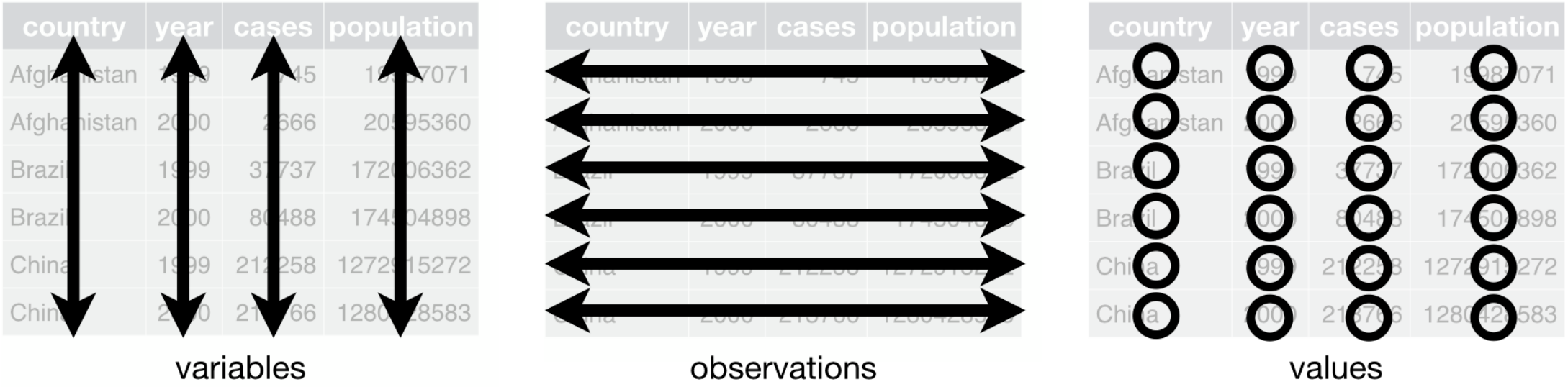
Part 1: tidy data for functional programming and iteration

John Little
Duke University Libraries
Center for Data & Visualization Sciences

2023-09-22

TIDY DATA

TIDY DATA[1]



1. A robust discussion of *tidy data* can be found in *R for Data Science* (Wickham, Cetinkaya-Rundel, and Grolemund 2023): *https://r4ds.had.co.nz/tidy-data.html*

WIDE DATA

- Code

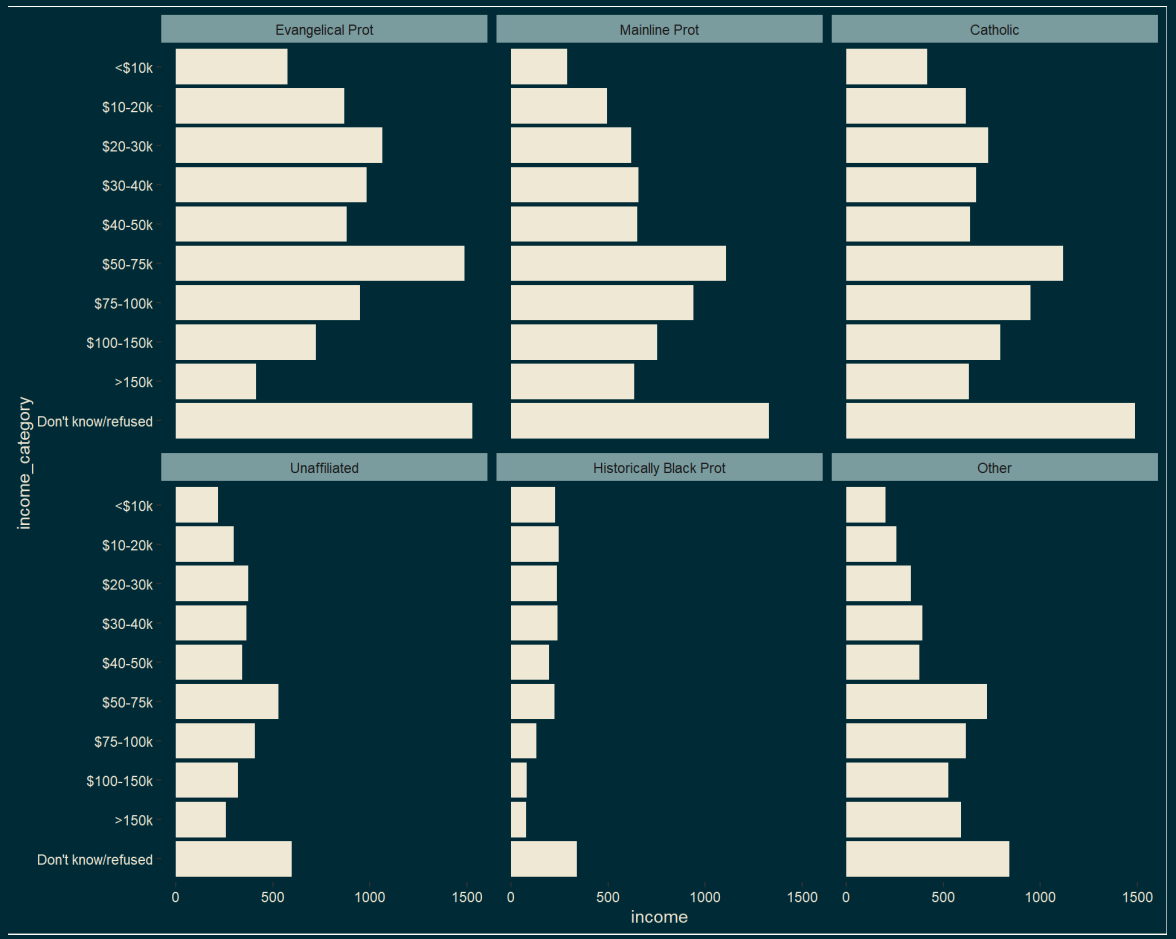
	RELIGION	<$10K	$10-20K	$20-30K	$30-40K	$40-50K	$50-75K	$75-100K	$100-150K	>150K	DON'T KNOW/REFUSED
1	Agnostic	27	34	60	81	76	137	122	109	84	96
2	Atheist	12	27	37	52	35	70	73	59	74	76
3	Buddhist	27	21	30	34	33	58	62	39	53	54
4	Catholic	418	617	732	670	638	1116	949	792	633	1489
5	Don’t know/refused	15	14	15	11	10	35	21	17	18	116
6..17											
18	Unaffiliated	217	299	374	365	341	528	407	321	258	597

TALL DATA

▶ Code

	RELIGION	INCOME_CATEGORY	INCOME
1	Agnostic	<$10k	27
2	Agnostic	$10-20k	34
3	Agnostic	$20-30k	60
4	Agnostic	$30-40k	81
5	Agnostic	$40-50k	76
6..179			
180	Unaffiliated	Don't know/refused	597

▶ Code



CODE

```
relig_income |>
  pivot_longer(cols = -religion, names_to = "income_category") |>
  ggplot(aes(value, income_category)) +
  geom_col() +
  facet_wrap(vars(religion))
```

PIVOT

{ TIDYR }

`tidyr::pivot_longer()` `tidyr::pivot_wider()`

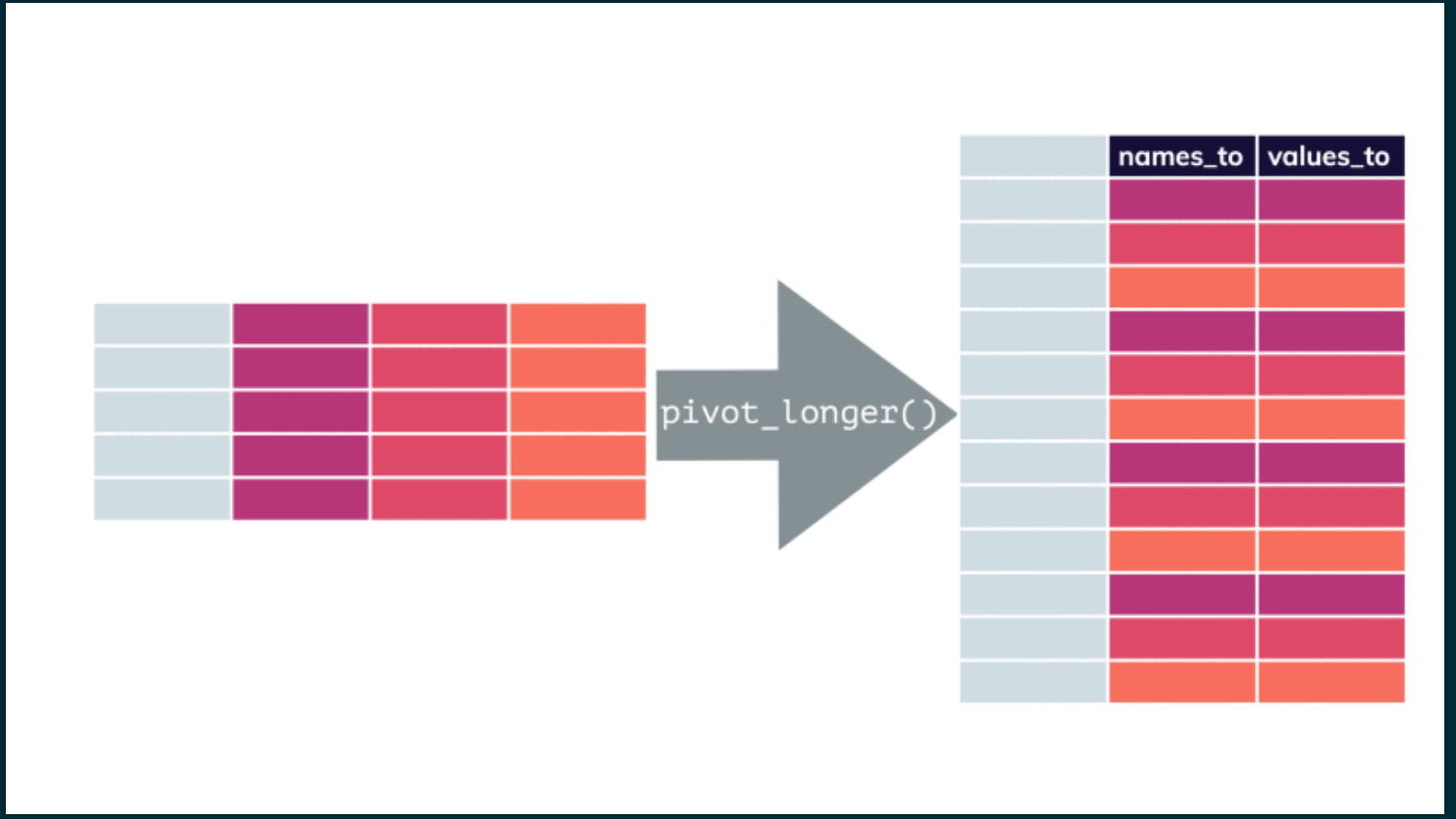
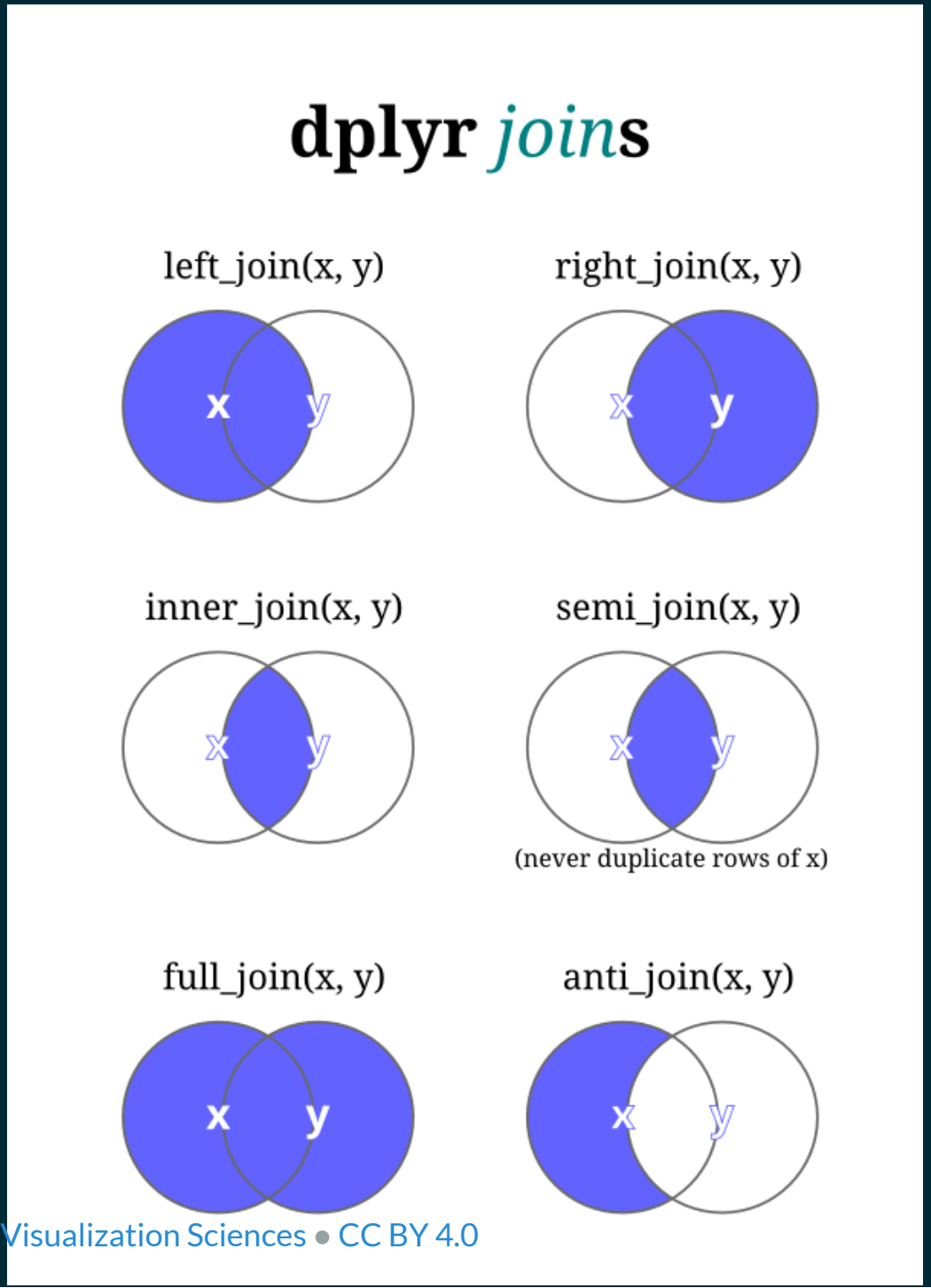


Image Credit: apreshill | CC BY 4.0 | https://github.com/apreshill/teachthat/blob/master/pivot/pivot_longer_smaller.gif]

JOIN

{ DPLYR }

```
1  fav_ratings  |>
2    left_join(starwars,
3              by = join_by(name))
```



TWO DATASETS

538.com

	LH join key	
	NAME	FAV_RATING
1	Han Solo	610
2	Luke Skywalker	552
3	Princess Leia Organa	547
4	Anakin Skywalker	245
5	Obi Wan Kenobi	591
6	Emperor Palpatine	110
7	Darth Vader	310
8	Lando Calrissian	142
9..13		
14	Yoda	605

{ dplyr::starwars }

	RH join key				
	NAME	HEIGHT	MASS	GENDER	HOMEWORLD
1	Luke Skywalker	172	77	masculine	Tatooine
2	C-3PO	167	75	masculine	Tatooine
3	R2-D2	96	32	masculine	Naboo
4	Darth Vader	202	136	masculine	Tatooine
5	Leia Organa	150	49	feminine	Alderaan
6	Owen Lars	178	120	masculine	Tatooine
7	Beru Whitesun lars	165	75	feminine	Tatooine
8	R5-D4	97	32	masculine	Tatooine
9..86					
87	Padmé Amidala	165	45	feminine	Naboo

Data gathered then transformed from
https://github.com/fivethirtyeight/data/tree/master/star-wars-survey

JOINED DATA

One data frame

```
fav_ratings  |>
  left_join(starwars,
            by = join_by(name))
```

Join key	Source df	Target data-frame			
NAME	FAV_RATING	HEIGHT	MASS	GENDER	HOMEWORLD
Han Solo	610	180	80.0	masculine	Corellia
Luke Skywalker	552	172	77.0	masculine	Tatooine
Princess Leia Organa	547	NA	NA	NA	NA
Anakin Skywalker	245	188	84.0	masculine	Tatooine
Obi Wan Kenobi	591	NA	NA	NA	NA
Emperor Palpatine	110	NA	NA	NA	NA
Darth Vader	310	202	136.0	masculine	Tatooine
Lando Calrissian	142	177	79.0	masculine	Socorro
Boba Fett	138	183	78.2	masculine	Kamino
C-3P0	474	NA	NA	NA	NA
R2 D2	562	NA	NA	NA	NA
Jar Jar Binks	112	196	66.0	masculine	Naboo
Padme Amidala	168	NA	NA	NA	NA
Yoda	605	66	17.0	masculine	NA

anti_join()

Identifying what does not match

Source df	Target df
FAV_RATING_KEY	STARWARS_KEY
C-3P0	C-3PO
Padme Amidala	Padmé Amidala
R2 D2	R2-D2
Obi Wan Kenobi	Obi-Wan Kenobi
Princess Leia Organa	Leia Organa
Emperor Palpatine	Palpatine

anti_join()

Mutate and regex to modify left-hand side

Source df		Target df
FAV_RATING_KEY	REGEX_MODIFIED_KEY_LH	STARWARS_KEY
C-3P0	c3p0	C-3PO
Padme Amidala	padmeamidala	Padmé Amidala
R2 D2	r2d2	R2-D2
Obi Wan Kenobi	obiwankenobi	Obi-Wan Kenobi
Princess Leia Organa	princessleiaorgana	Leia Organa
Emperor Palpatine	emperorpalpatine	Palpatine

anti_join()

Mutate and regex to modify right-hand side

Source df		Target df	
FAV_RATING_KEY	REGEX_MODIFIED_KEY_LH	REGEX_MODIFIED_KEY_RH	STARWARS_KEY
C-3P0	c3p0	c3po	C-3PO
Padme Amidala	padmeamidala	padméamidala	Padmé Amidala
R2 D2	r2d2	r2d2	R2-D2
Obi Wan Kenobi	obiwankenobi	obiwankenobi	Obi-Wan Kenobi
Princess Leia Organa	princessleiaorgana	leiaorgana	Leia Organa
Emperor Palpatine	emperorpalpatine	palpatine	Palpatine

anti_join()

Incremental improvements

Source df			Target df	
FAV_RATING_KEY	REGEX_MODIFIED_KEY_LH	KEY_MATCH	REGEX_MODIFIED_KEY_RH	STARWARS_KEY
C-3P0	c3p0	FALSE	c3po	C-3PO
Padme Amidala	padmeamidala	FALSE	padméamidala	Padmé Amidala
R2 D2	r2d2	TRUE	r2d2	R2-D2
Obi Wan Kenobi	obiwankenobi	TRUE	obiwankenobi	Obi-Wan Kenobi
Princess Leia Organa	princessleiaorgana	FALSE	leiaorgana	Leia Organa
Emperor Palpatine	emperorpalpatine	FALSE	palpatine	Palpatine

HINTS ON KEYS AND REGEX

- Numeric keys are the best match keys
- Natural language keys are not good match keys
- Regular expressions can normalize keys to some extent
- {fuzzyjoin} package may help