Data Wrangling

INFO 201

d3.unconf()

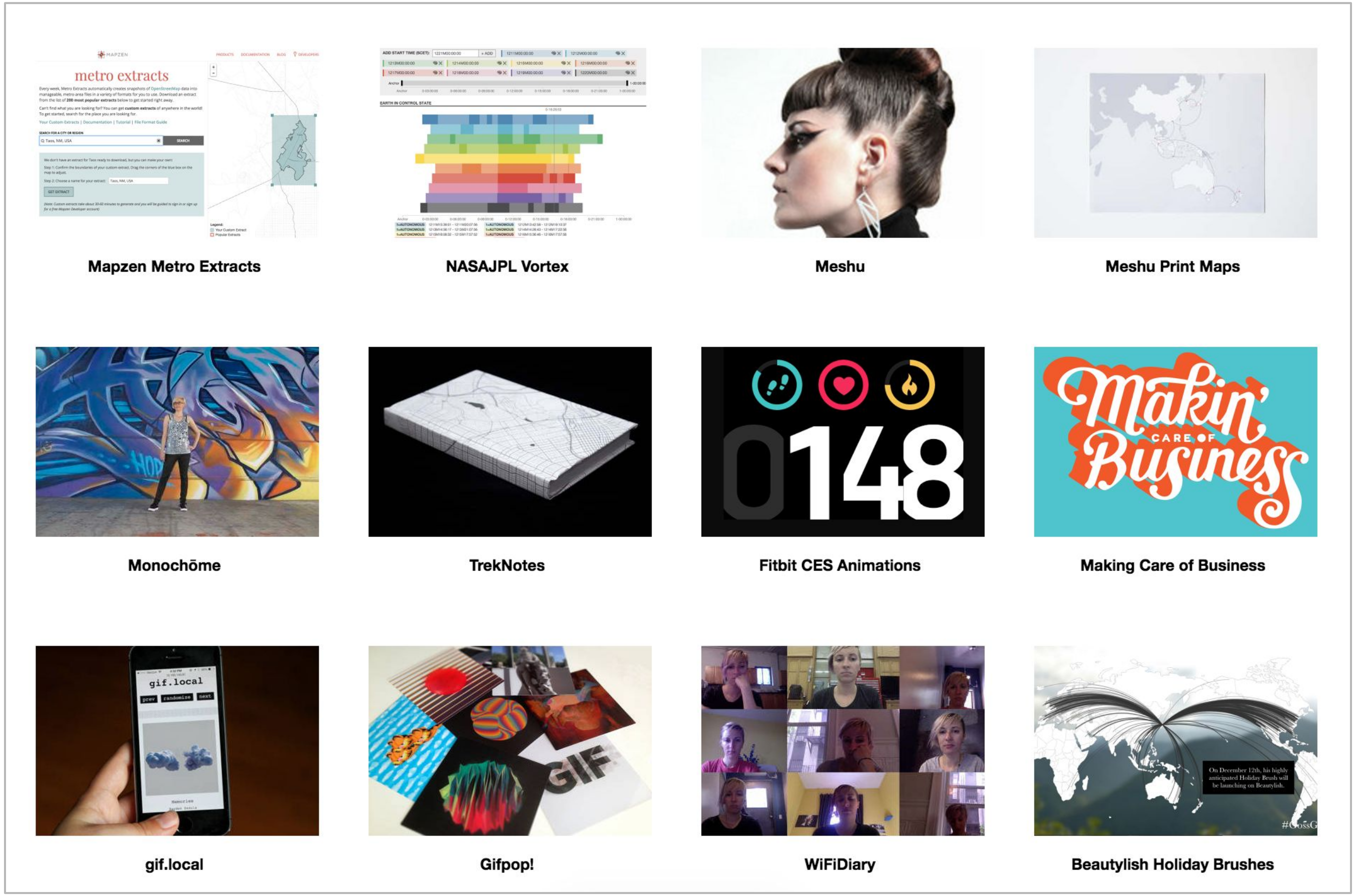
nE&s

2016

data sketch|es

a Nadieh & Shirley collaboration

Nadieh Bremmer + Shirley Wu



Mapzen Metro Extracts

NASAJPL Vortex

Meshu

Meshu Print Maps

Monochōme

TrekNotes

Fitbit CES Animations

Making Care of Business

gif.local

Gifpop!

WiFiDiary

Beautylish Holiday Brushes

Rachel Binx

Today’s Objectives

Consider how to map from analytical steps to programming tasks

Understand how use DPLYR's ***data manipulation verbs*** to wrangle data

Practice chaining methods together by using the ***pipe operator***

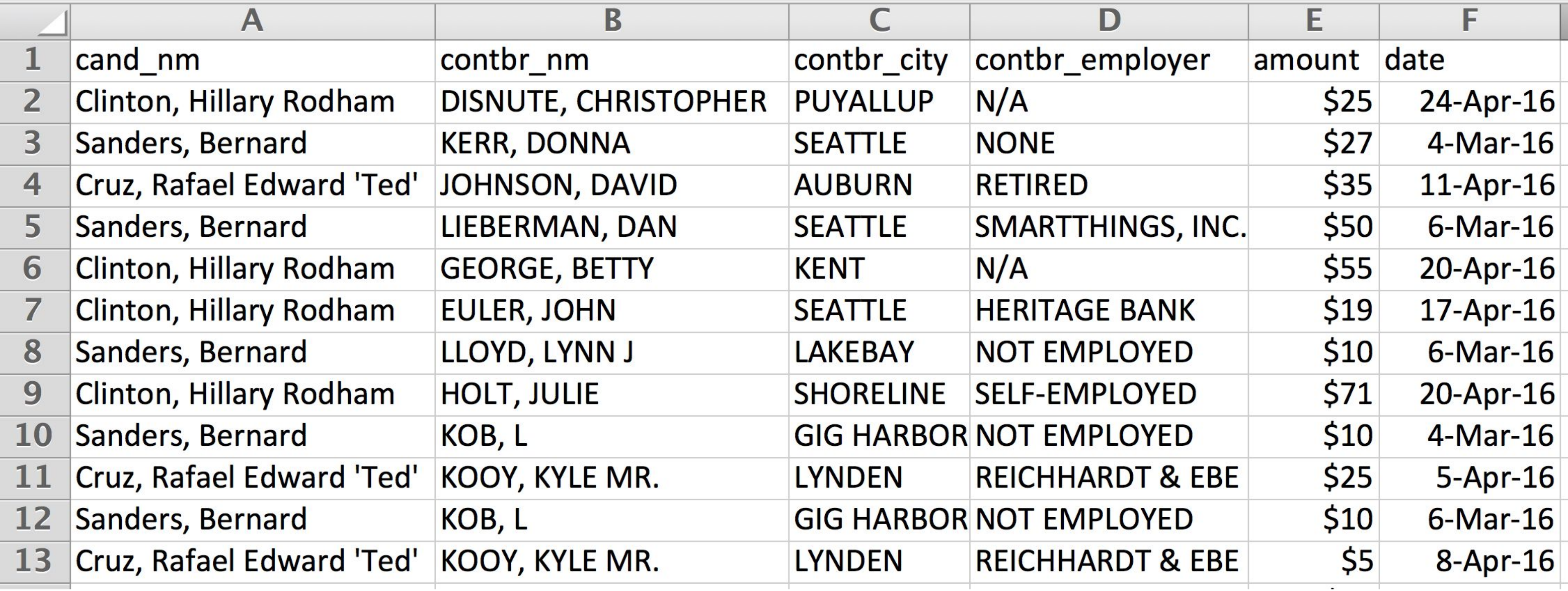
Analytical steps

Steps for Data Analysis

Articulate a research question of interest

Translate your questions into code

Execute your program



	A	B	C	D	E	F
1	cand_nm	contbr_nm	contbr_city	contbr_employer	amount	date
2	Clinton, Hillary Rodham	DISNUTE, CHRISTOPHER	PUYALLUP	N/A	$25	24-Apr-16
3	Sanders, Bernard	KERR, DONNA	SEATTLE	NONE	$27	4-Mar-16
4	Cruz, Rafael Edward 'Ted'	JOHNSON, DAVID	AUBURN	RETIRED	$35	11-Apr-16
5	Sanders, Bernard	LIEBERMAN, DAN	SEATTLE	SMARTTHINGS, INC.	$50	6-Mar-16
6	Clinton, Hillary Rodham	GEORGE, BETTY	KENT	N/A	$55	20-Apr-16
7	Clinton, Hillary Rodham	EULER, JOHN	SEATTLE	HERITAGE BANK	$19	17-Apr-16
8	Sanders, Bernard	LLOYD, LYNN J	LAKEBAY	NOT EMPLOYED	$10	6-Mar-16
9	Clinton, Hillary Rodham	HOLT, JULIE	SHORELINE	SELF-EMPLOYED	$71	20-Apr-16
10	Sanders, Bernard	KOB, L	GIG HARBOR	NOT EMPLOYED	$10	4-Mar-16
11	Cruz, Rafael Edward 'Ted'	KOOY, KYLE MR.	LYNDEN	REICHHARDT & EBE	$25	5-Apr-16
12	Sanders, Bernard	KOB, L	GIG HARBOR	NOT EMPLOYED	$10	6-Mar-16
13	Cruz, Rafael Edward 'Ted'	KOOY, KYLE MR.	LYNDEN	REICHHARDT & EBE	$5	8-Apr-16

What are 5 questions that you have about this dataset?

Example data

downloaded here

Sample Questions

Who donated the most money?

Which city did the largest donation come from?

When was the smallest donation made?

Sample Questions

Who donated the most money?

Which city did the largest donation come from?

When was the smallest donation made?

Select a **column** of interest

Sample Questions

Who donated the **most money**?

Which city did the **largest donation** come from?

When was the **smallest donation** made?

Filter down to a specific **row**

Grammar of Data Manipulation

Select particular columns

Filter down to specific rows

Arrange (sort) your dataset by values

Mutate your dataframe to add a column

Summarise your dataframe (calculate summary info, mean)

module 9 exercise-1

DPLYR

DPLYR

"A grammar for data manipulation"

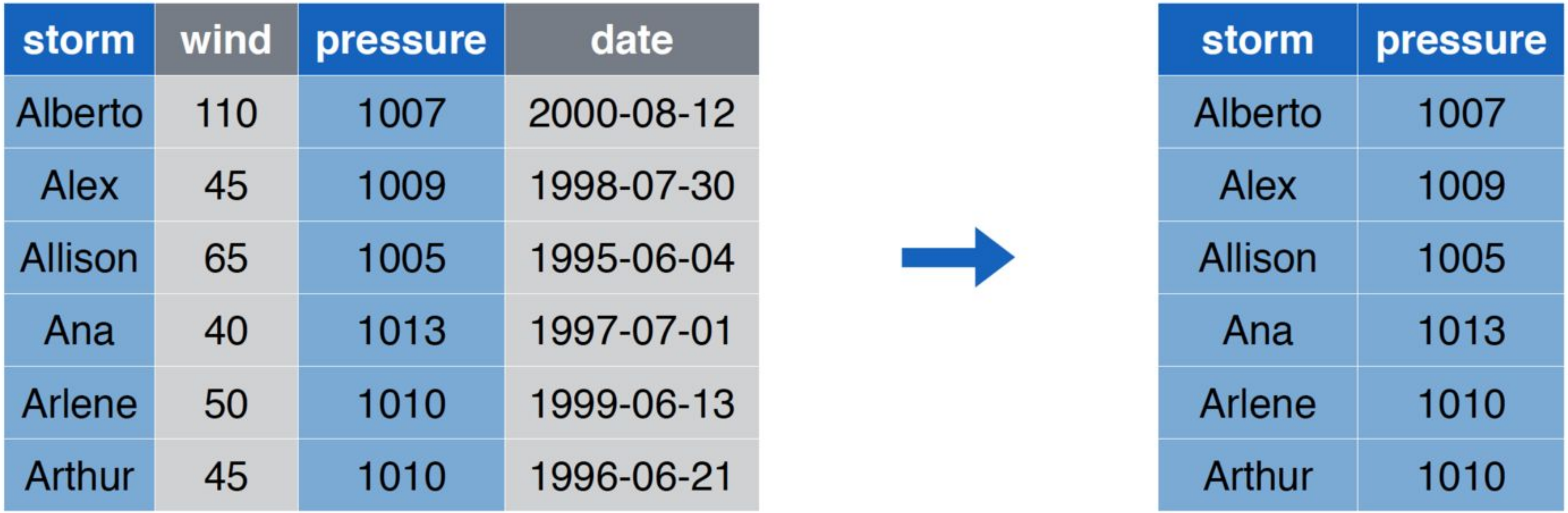
Provides verbs for common tasks

Make your code easier to write and read

Written by Hadley Wickham

select()

storms

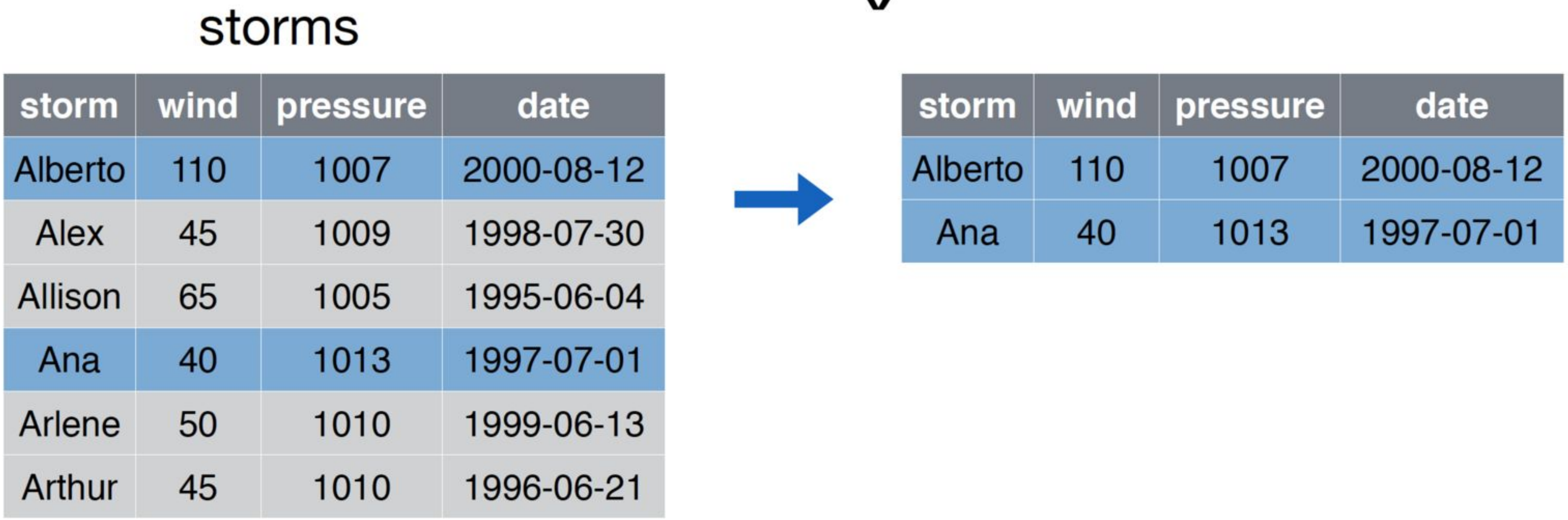


storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	pressure
Alberto	1007
Alex	1009
Allison	1005
Ana	1013
Arlene	1010
Arthur	1010

```
storms <- select(storms, storm, pressure)
```

filter()



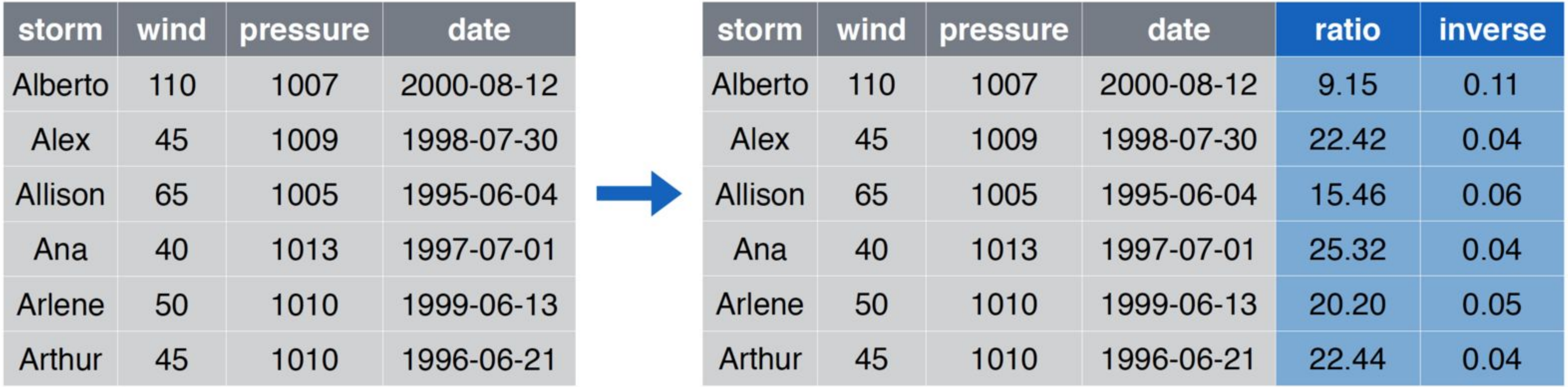
storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Ana	40	1013	1997-07-01

```
storms <- filter(storms,storm %in% c('Ana', 'Alberto'))
```

mutate()

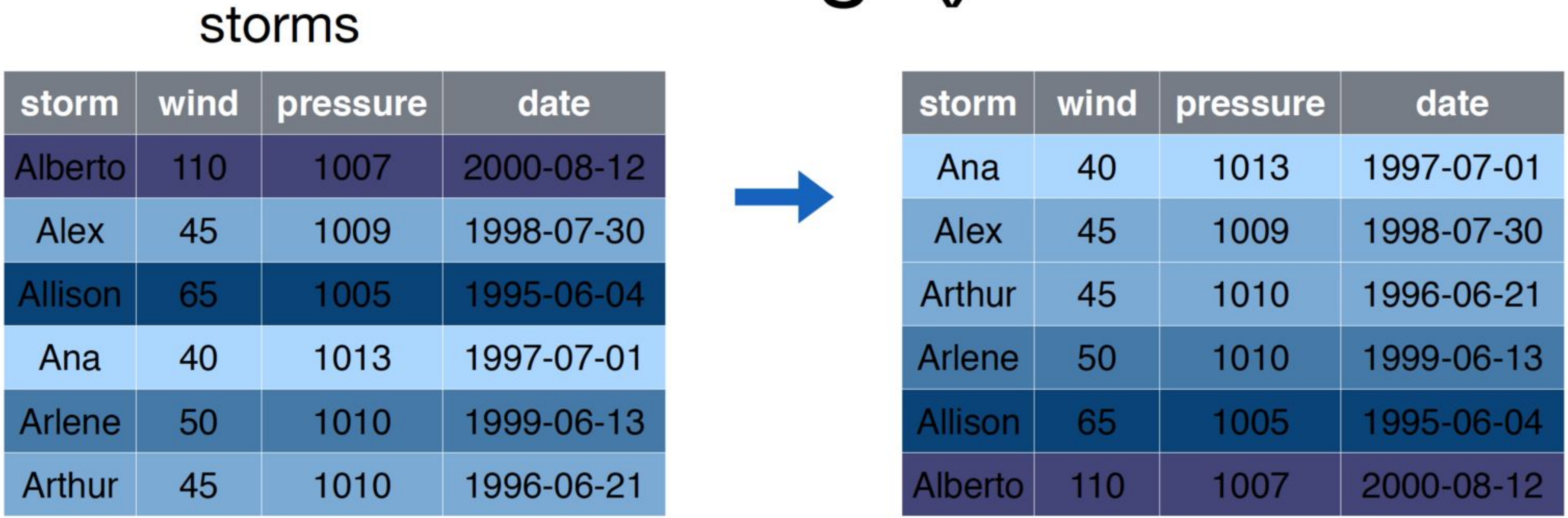


storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	wind	pressure	date	ratio	inverse
Alberto	110	1007	2000-08-12	9.15	0.11
Alex	45	1009	1998-07-30	22.42	0.04
Allison	65	1005	1995-06-04	15.46	0.06
Ana	40	1013	1997-07-01	25.32	0.04
Arlene	50	1010	1999-06-13	20.20	0.05
Arthur	45	1010	1996-06-21	22.44	0.04

```
storms <- mutate(storms, ratio = pressure/wind, inverse = 1/ratio)
```

arrange()



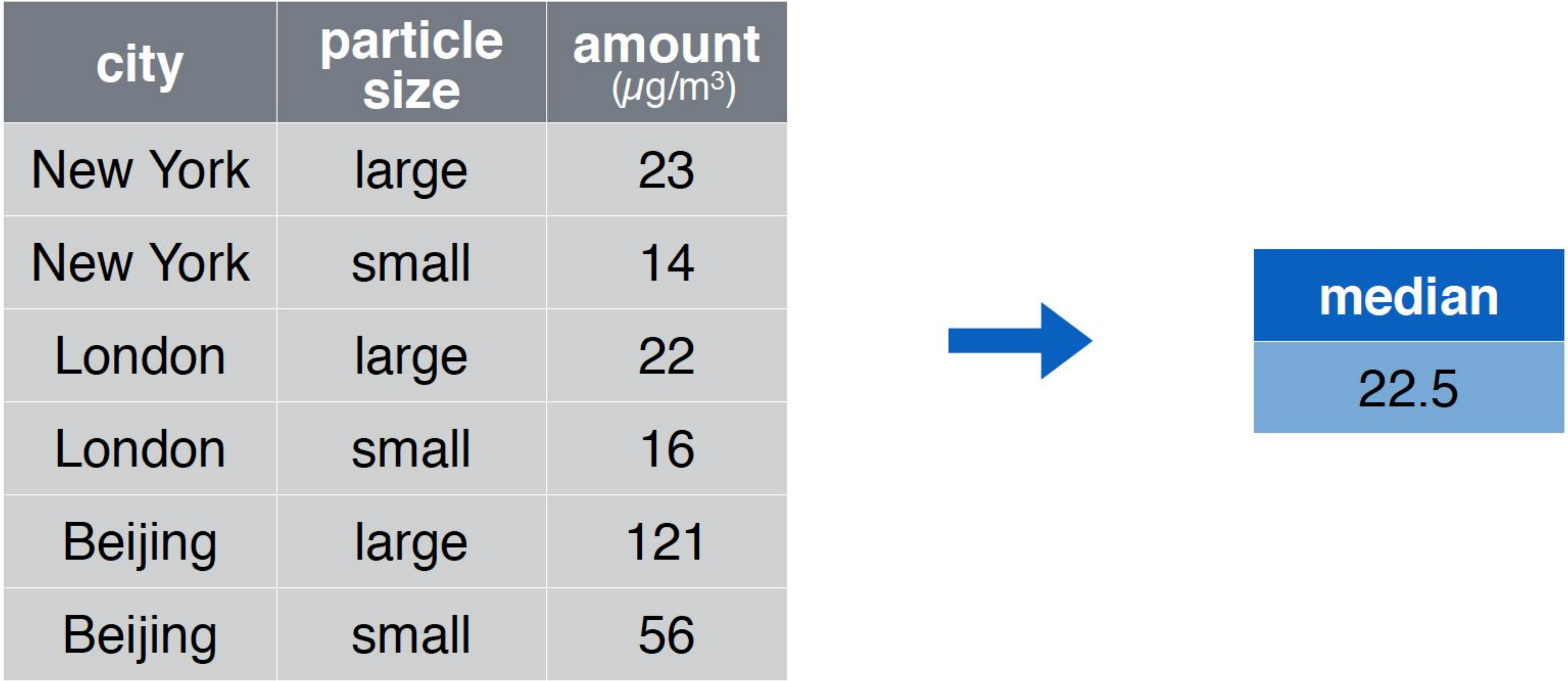
storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

storm	wind	pressure	date
Ana	40	1013	1997-07-01
Alex	45	1009	1998-07-30
Arthur	45	1010	1996-06-21
Arlene	50	1010	1999-06-13
Allison	65	1005	1995-06-04
Alberto	110	1007	2000-08-12

```
storms <- arrange(storms, wind)
```



city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

median
22.5

```
summary <- summarise(pollution, median = median(amount))
```

Summarize

module 9 exercise-2

Chaining Methods

Chaining Methods

What are the steps for answering this question of the mtcars dataset:

> *Which 4-cylinder car gets the best milage per gallon?*

Actually a few steps:

1. **Filter** down the dataset to only 4 cylinder cars
2. Of the 4 cylinder cars, **filter** down to the one with the highest mpg
3. **Select** the car name of the car from step 2.

```r
# Add a column that is the car name
mtcars.named <- mutate(mtcars, car.name = row.names(mtcars))

# Filter down to only four cylinder cars
four.cyl <- filter(mtcars.named, cyl == 4)

# Get the best four cylinder car
best.four.cyl <- filter(four.cyl, mpg == max(mpg))

# Get the name of the car
best.car.name <- select(best.four.cyl, car.name)
```

What we've been doing...

```r
# Add a column that is the car name
mtcars.named <- mutate(mtcars, car.name = row.names(mtcars))

# Write a nested operation to return the best car name

# Select name from the filtered data
best.car.name <- select(
                  # Filter the 4 cylinder data down by MPG
                  filter(
                    # Filter down to 4 cylinders
                    filter(
                      mtcars.named,
                      cyl == 4
                    ),
                    mpg == max(mpg)
                  ), car.name
                )
```

We could also nest…

The Pipe Operator

Takes the ***result from one function*** and passes it in as the ***first argument*** to the next function

Part of the DPLYR package

Written in R as %>% (use the shortcut)

This will completely simplify your code

```r
# Add a column that is the car name
mtcars.named <- mutate(mtcars, car.name = row.names(mtcars))

# Begin your piped operation: filter down to only four cylinder cars
best.car.name <- filter(mtcars.named, cyl == 4) %>%
                       filter(mpg == max(mpg)) %>%
                       select(car.name)
```

Pipe Operator!

module 9 exercise-3

Upcoming...

By Thursday: Be comfortable with **module 9**

Due Tuesday, 10/25 (***before class***): a4-data-wrangling