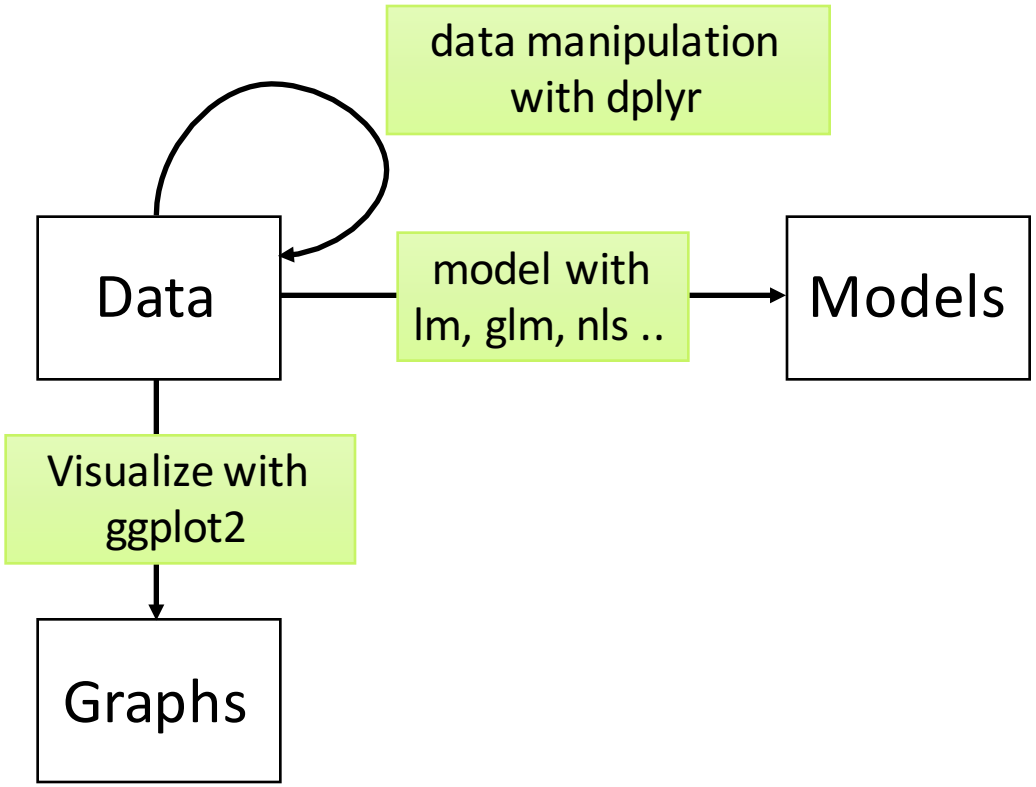
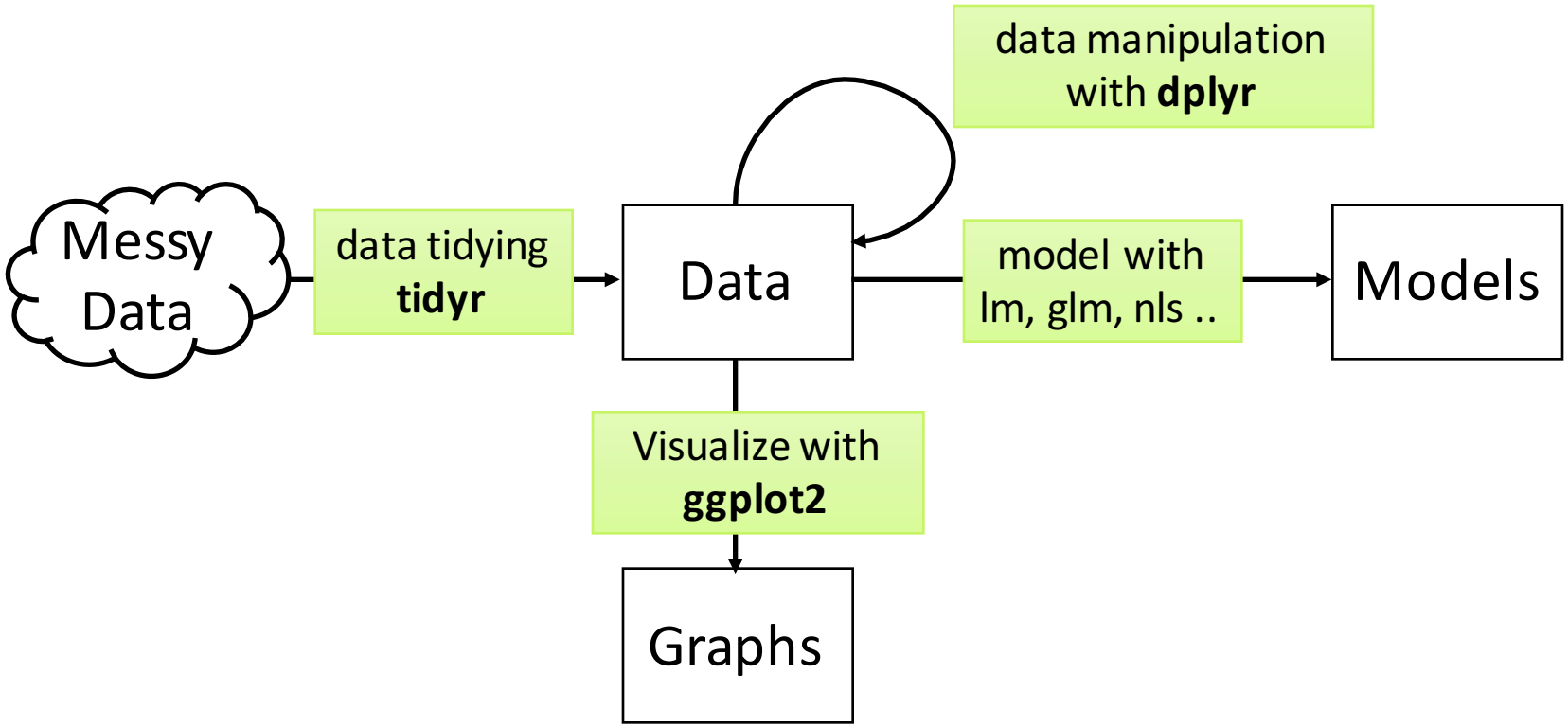
Tidy your data and results with tidyr & broom

Vijay Ivaturi
Center for Translational Medicine
University of Maryland, School of Pharmacy


data manipulation
with dplyr
Data
model with
lm, glm, nls ..
Models
Visualize with
ggplot2
Graphs


data manipulation
with **dplyr**
Messy
Data
data tidying
tidyr
Data
model with
lm, glm, nls ..
Models
Visualize with
ggplot2
Graphs

tidyr

A package that ***reshapes*** the layout of tables

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

```
df <- data.frame(  id = 1,
date = paste0(seq(as.Date("2015/10/20"), by = "day", length.out =
3),"T","9:00:00"),
parent = c(round(sort(10 - rnorm(2,1,1),decreasing = TRUE),2),NA),
met1 = c(round(sort(8 - rnorm(2,1,1),decreasing = TRUE),2),NA),
wt = c(70,rep(NA,2)),
age = c(50,rep(".",2)))
```

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

gather()

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

(former column names)

key

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

(former cells)

key values

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

gather()

Collapses multiple columns into two columns:

1. a key column that contains the former column names
2. a value column that contains the former column cells

```
gather(df, analyte,DV,parent:met1)
```

gather()

Collapses multiple columns into two columns:

1. a key column that contains the former column names
2. a value column that contains the former column cells

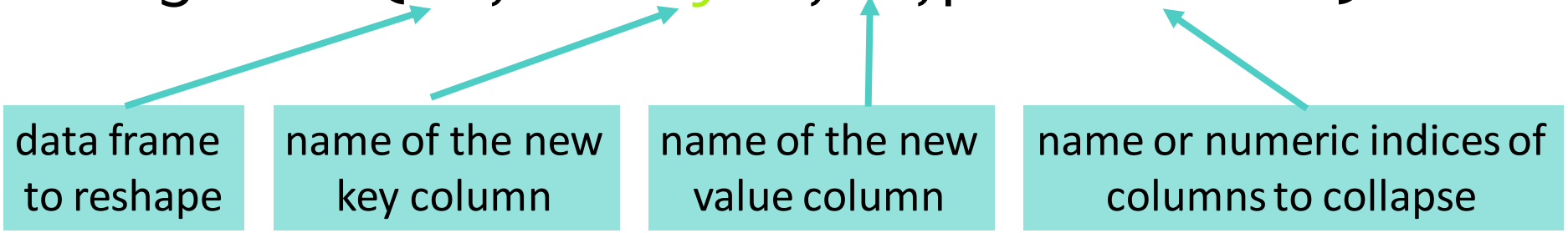
```
gather(df, analyte,DV,parent:met1)
```

gather()

Collapses multiple columns into two columns:

1. a key column that contains the former column names
2. a value column that contains the former column cells

```
gather(df, analyte,DV,parent:met1)
```



df %>% gather(analyte,DV,parent:met1)

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

Active Moiety Concentration
= Parent + Met1

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

mutate

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	
1	2015-10-21T9:00:00	NA	.	7.58	6.41	
1	2015-10-22T9:00:00	NA	.	NA	NA	

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

spread()

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

key (new column names)

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

key **value** (new cells)

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

spread()

Generates multiple columns from two columns:

1. each unique value in the key column becomes a column name
2. each value in the **value** column becomes a cell in the new columns

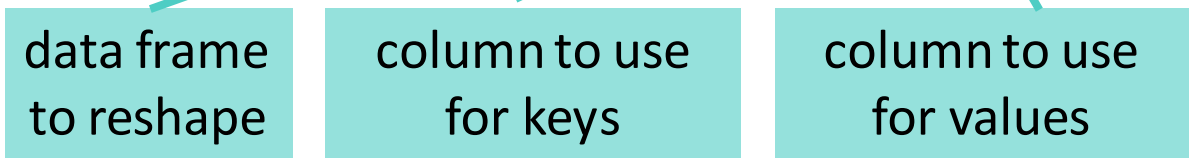
```
spread(df, analyte,DV)
```

spread()

Generates multiple columns from two columns:

1. each unique value in the key column becomes a column name
2. each value in the **value** column becomes a cell in the new columns

```
spread(df, analyte,DV)
```



```
df %>% gather(analyte,DV,parent:met1)
```

id	date	wt	age	parent	met1
1	2015-10-20T9:00:00	70	50	9.37	8.05
1	2015-10-21T9:00:00	NA	.	7.58	6.41
1	2015-10-22T9:00:00	NA	.	NA	NA

```
df %>%
spread(analyte,DV)
```

id	date	wt	age	analyte	DV
1	2015-10-20T9:00:00	70	50	parent	9.37
1	2015-10-21T9:00:00	NA	.	parent	7.58
1	2015-10-22T9:00:00	NA	.	parent	NA
1	2015-10-20T9:00:00	NA	50	met1	8.05
1	2015-10-21T9:00:00	NA	.	met1	6.41
1	2015-10-22T9:00:00	NA	.	met1	NA

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	NA	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	NA	.	NA	NA	NA

BQL	unknown	-99

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	NA	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	NA	.	BQL	unknown	-99

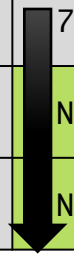
replace_na()

- replaces missing values with specific value of choice
- replaces many missing values at once to any “type”

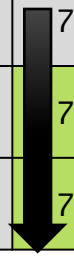
```
df %>% replace_na(list(parent="BQL",met1="unknown",AM=-99))
```

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	NA	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	NA	.	BQL	unknown	-99



id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	70	.	BQL	unknown	-99



fill()

- fills missing values with the most recent non-empty cell

- fill multiple columns at once by specifying bare variable names, e.g. wt:age or exclude certain variables using -age

```
df %>% fill(wt)
```

What is the outcome of fill()'ing age?

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	70	.	BQL	unknown	-99

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	70	.	BQL	unknown	-99

id	date	time	wt	age	parent	met1	AM

id	date	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	70	.	BQL	unknown	-99

```
df %>% separate(date,into=c("date","time"),sep="T")
```

id	date	time	wt	age	parent	met1	AM
1	2015-10-20	9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21	9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22	9:00:00	70	.	BQL	unknown	-99

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

id	date	time	wt	age	parent	met1	AM
1	2015-10-20	9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21	9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22	9:00:00	70	.	BQL	unknown	-99

```
df %>% unite(datetime,date,time,sep="T")
```

id	datetime	wt	age	parent	met1	AM
1	2015-10-20T9:00:00	70	50	9.37	8.05	17.42
1	2015-10-21T9:00:00	70	.	7.58	6.41	14.26
1	2015-10-22T9:00:00	70	.	BQL	unknown	-99

Verb	Usage
gather	collapses multiple columns in key-value pairs
spread	spreads a key-value pair across multiple columns
replace_na	replace missing values
fill	fills missing values by using previous entry
separate	turns a single character column into multiple columns
extract	turns each regex capture group into a new column
unite	paste together multiple columns into one
complete	explicitly completes missing data combinations

```
df_comp <- dplyr::data_frame(id = c(1,1,1,2,2),
                             time = c(0,1,2,0,1))
```

id	time
1	0
1	1
1	2
2	0
2	1

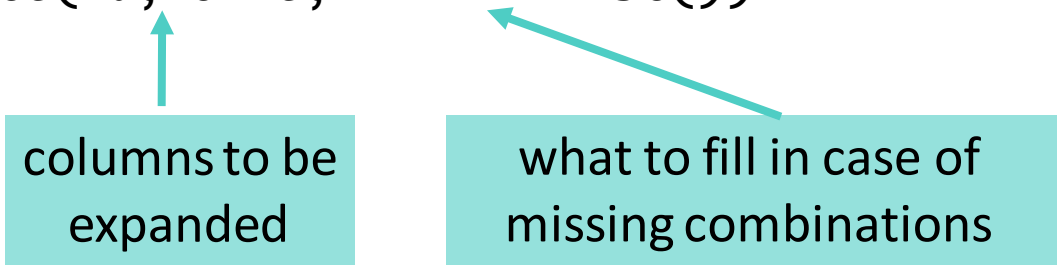
```
df_comp %>% complete(  id, time)
```

id	time
1	0
1	1
1	2
2	0
2	1
2	2

complete()

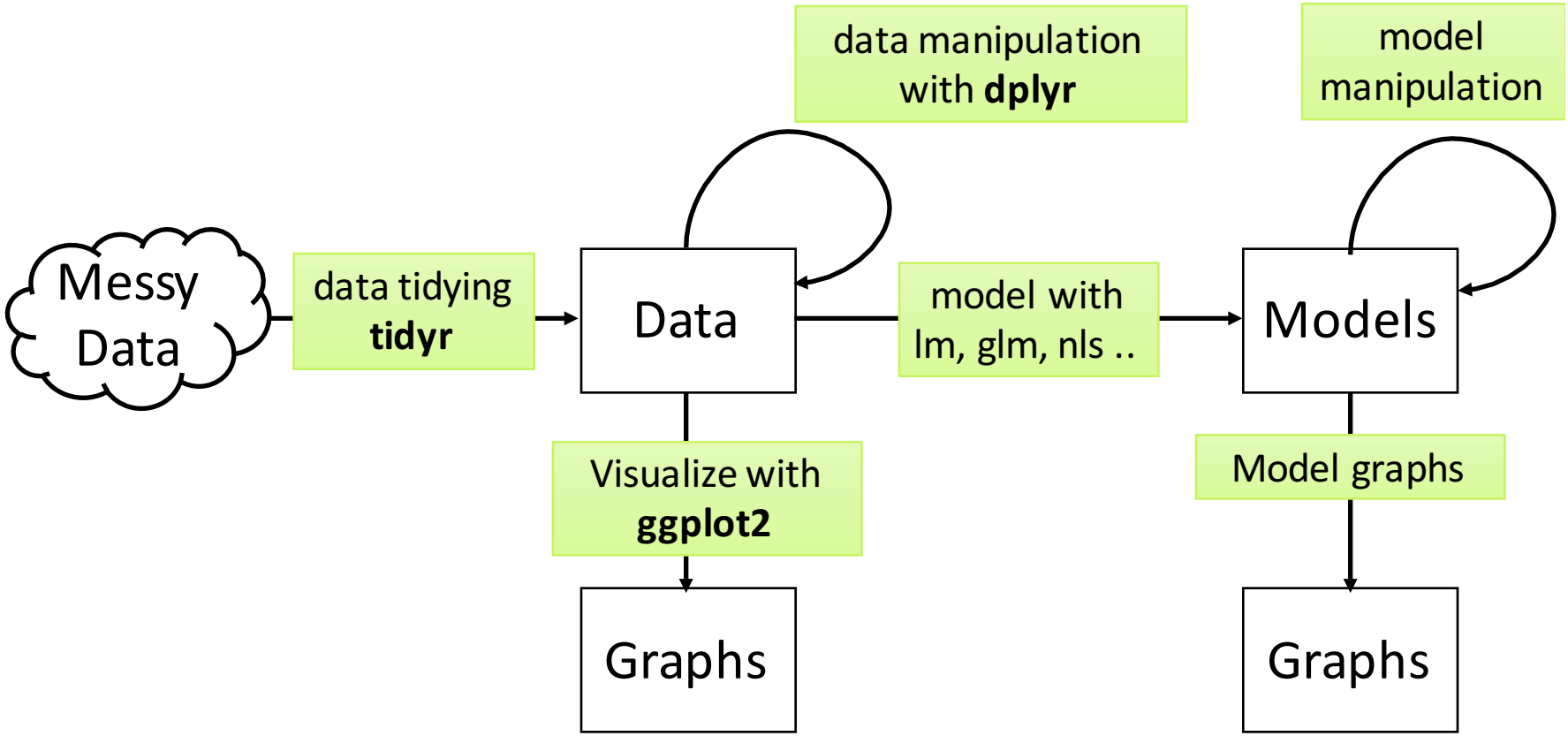
- explicitly completes missing data combinations
- an advanced version of expand.grid

```
df_comp %>% complete(id, time, fill = list())
```



broom

An R Package to Convert Statistical Models into Tidy Data Frames


Messy Data
data tidying
tidyr
Data
data manipulation
with dplyr
model with
lm, glm, nls ..
Models
model
manipulation
Visualize with
ggplot2
Graphs
Model graphs
Graphs

Verb	Usage
tidy	Component/fit level statistics
augment	Observation level statistics
glance	Model level statistics