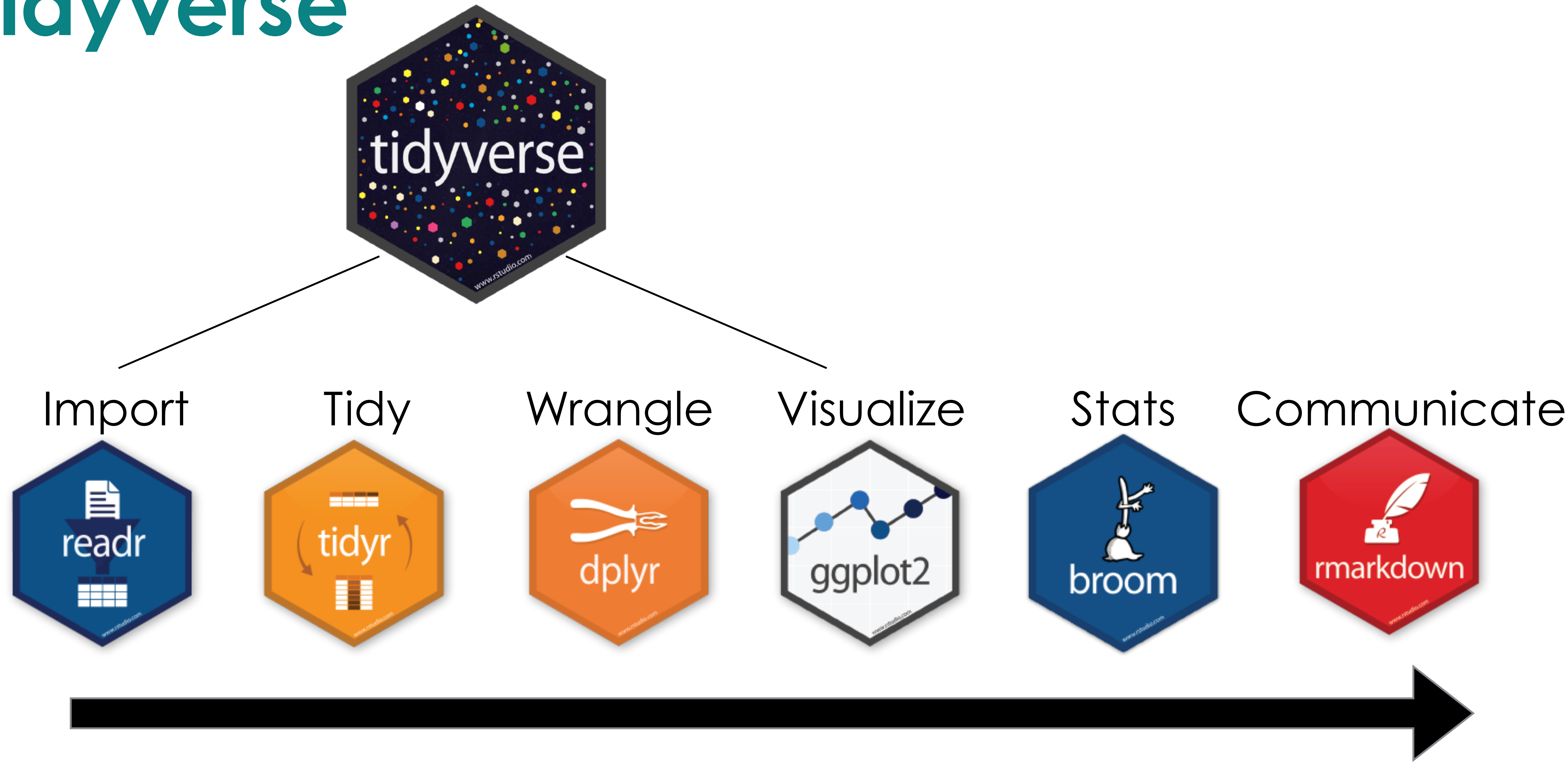
Introduction to R Workshop

Session 2: Importing and Tidying Data w/ R

Session 2: Goals

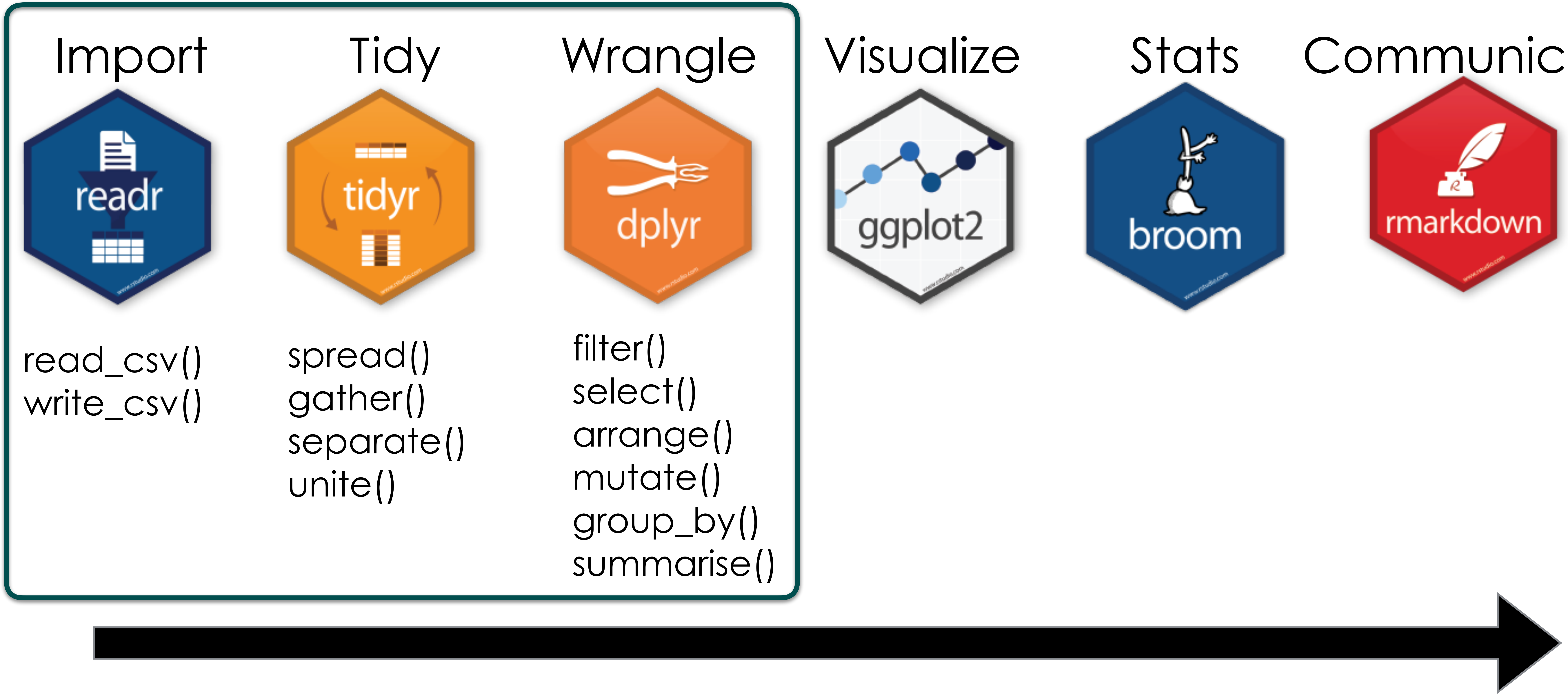
- **Import** data with readr
- **Tidy** a dataset
- **Transform/Wrangle** data

Data Analysis in the Tidyverse



Import Tidy Wrangle Visualize Stats Communicate

Data Analysis in the **Tidyverse**



Import

read_csv()
write_csv()

Tidy

spread()
gather()
separate()
unite()

Wrangle

filter()
select()
arrange()
mutate()
group_by()
summarise()

Visualize

Stats

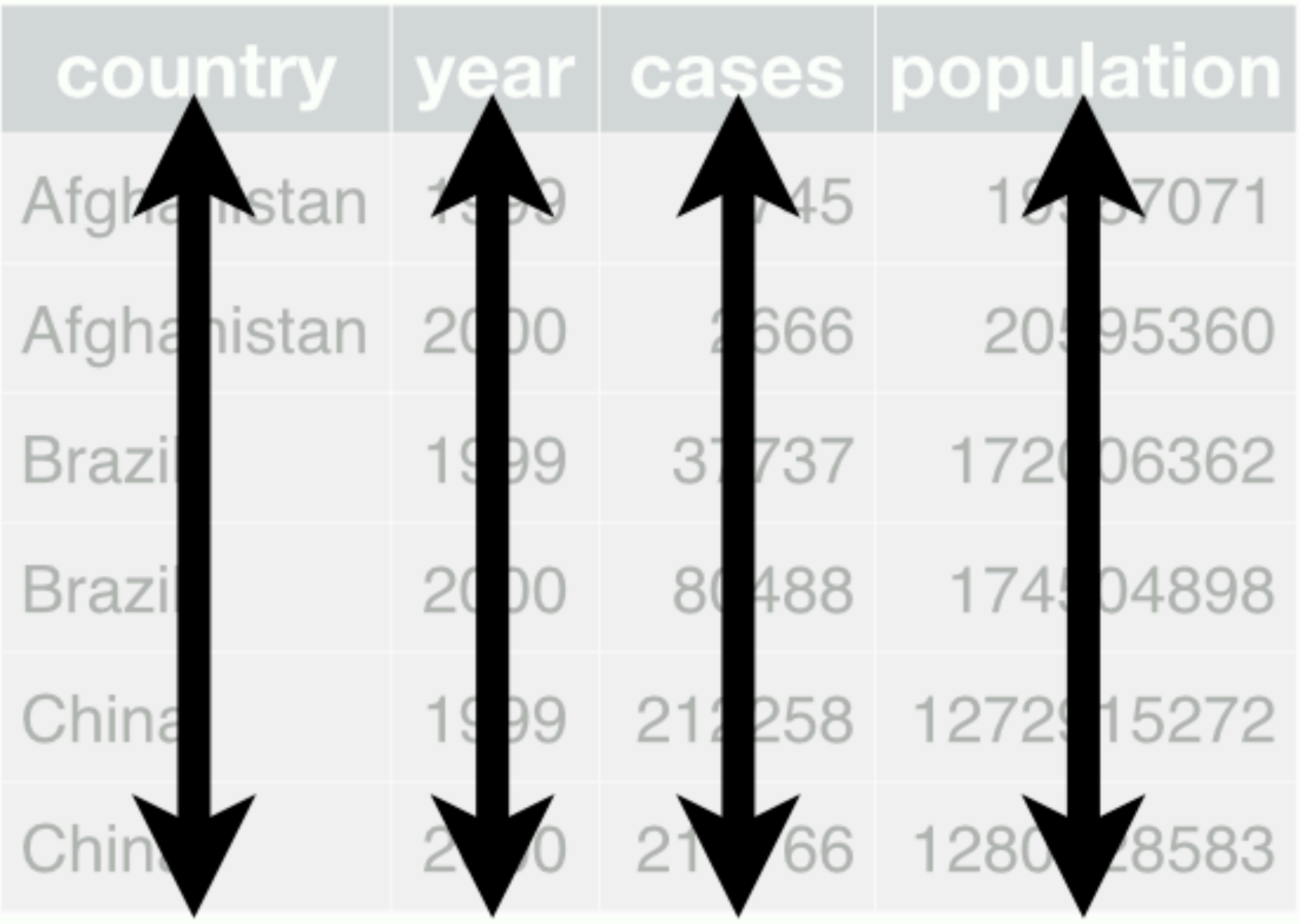
Communicate

readr

read_csv() - import .csv file

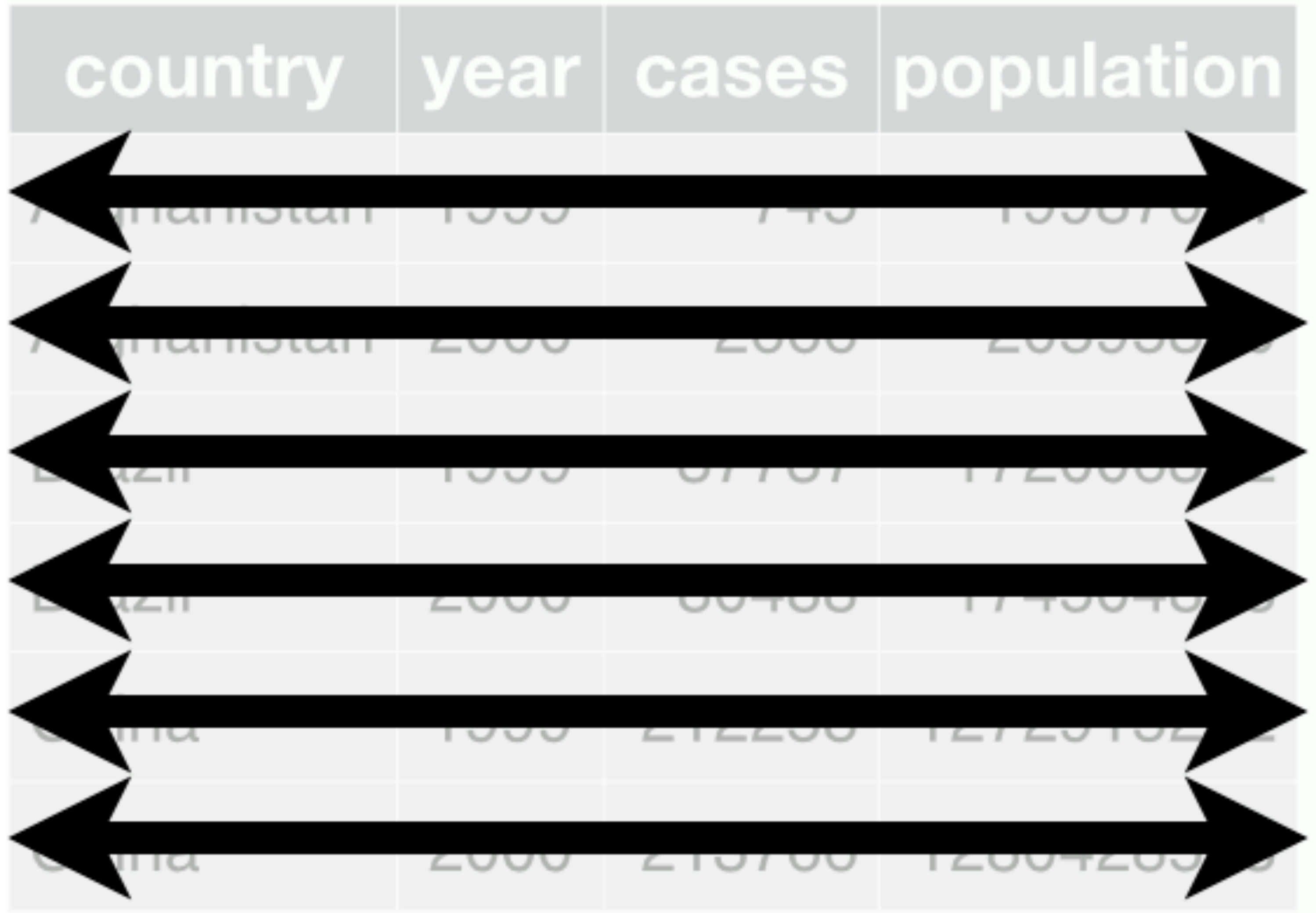
write_csv() - export .csv file

Tidy data

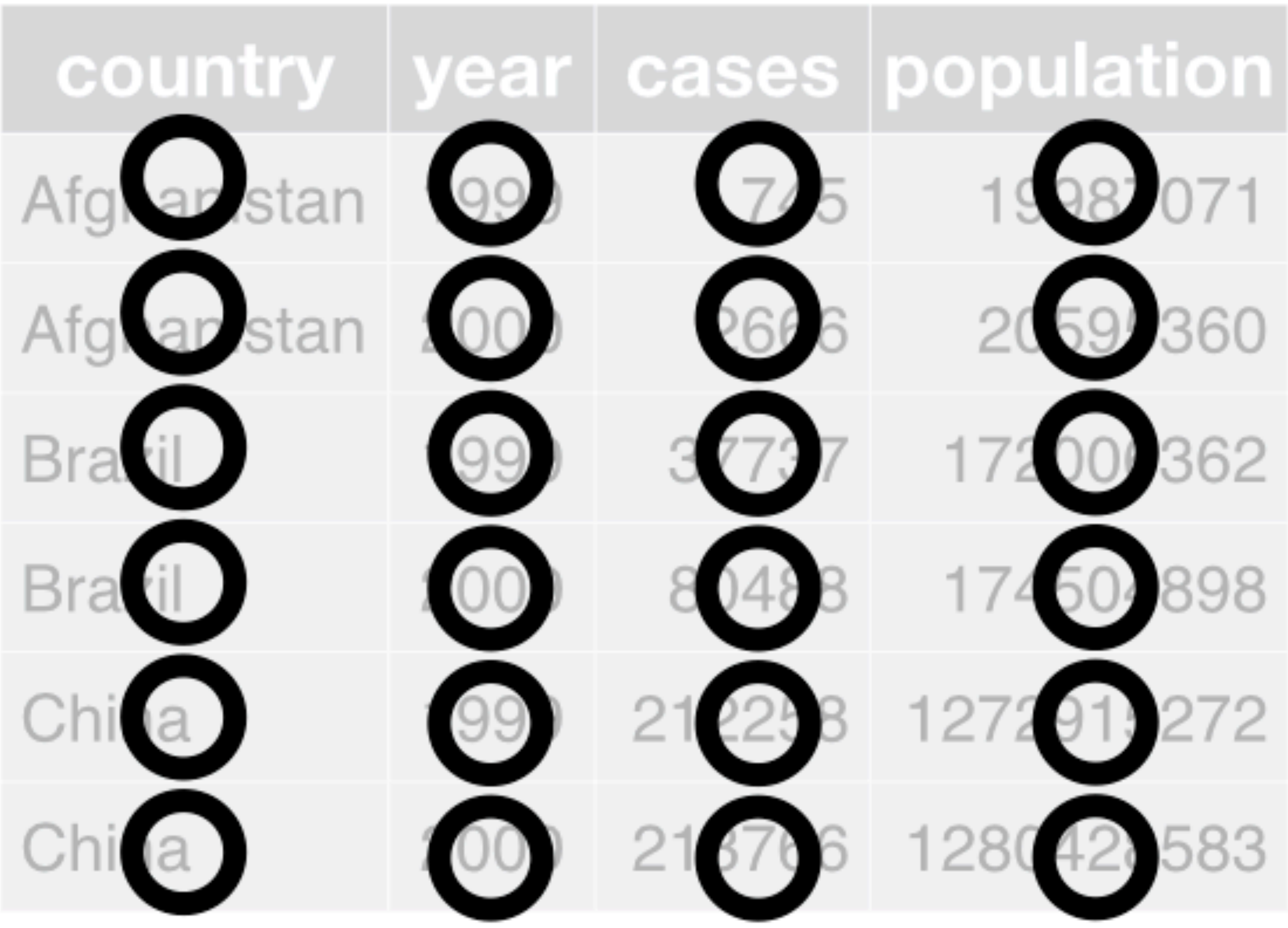


country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

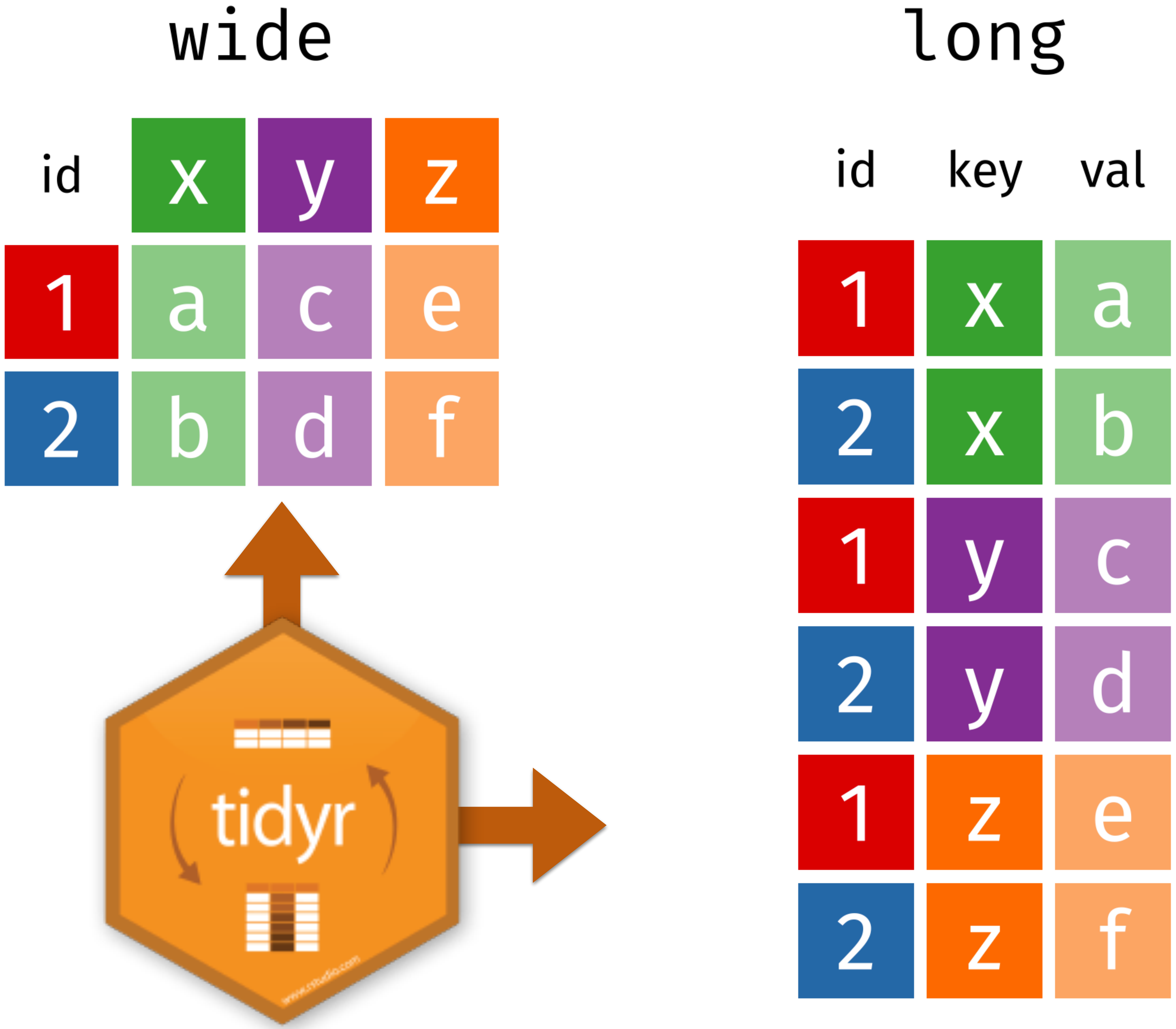


observations



values

Tidy data



wide

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

long

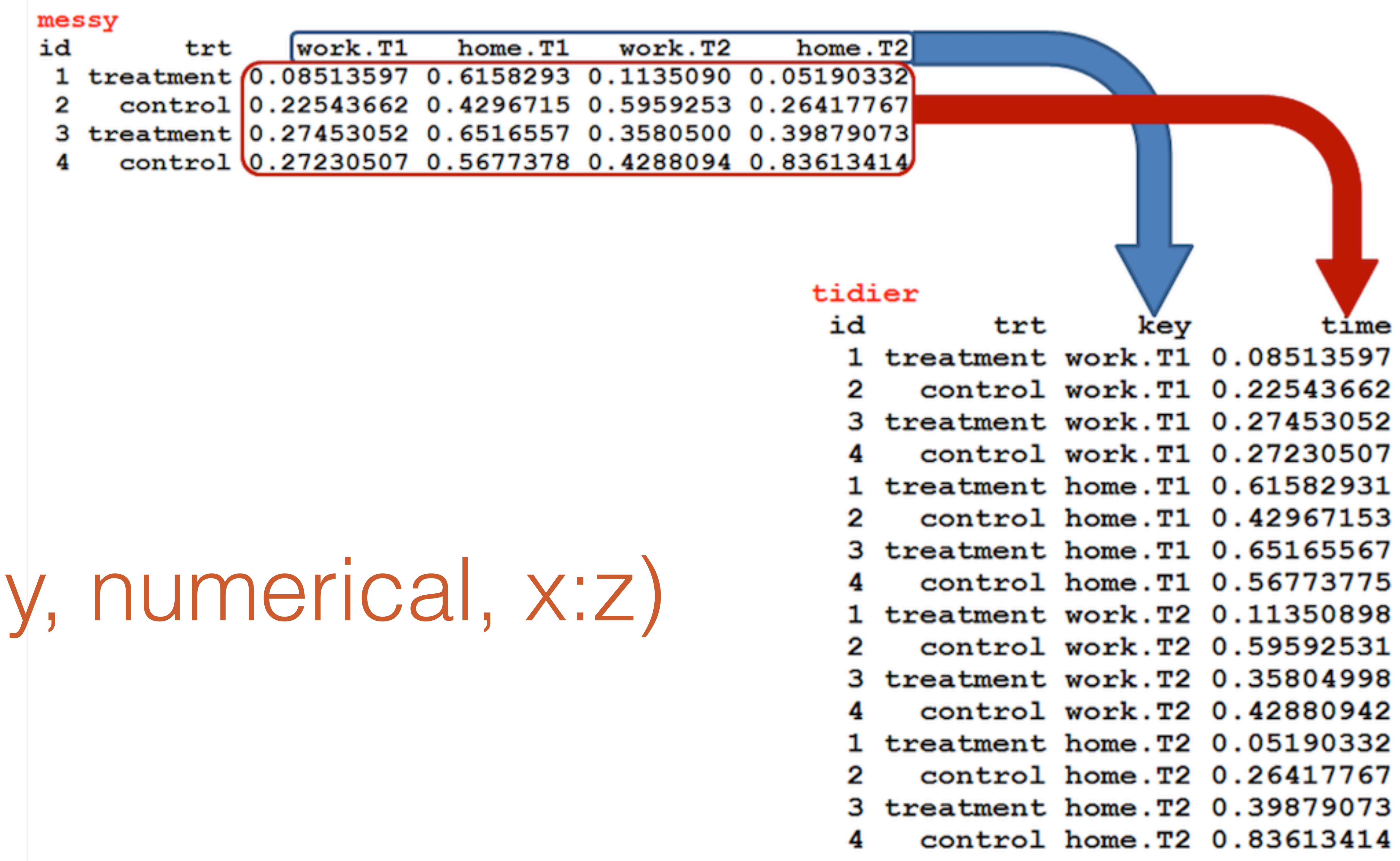
id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

tidyr - tidy up a dataset

- **gather()**
- **spread()**
- **separate()**
- **unite()**

gather() - reshapes data from ‘wide’ to ‘long

gather(key, time, 3:6)



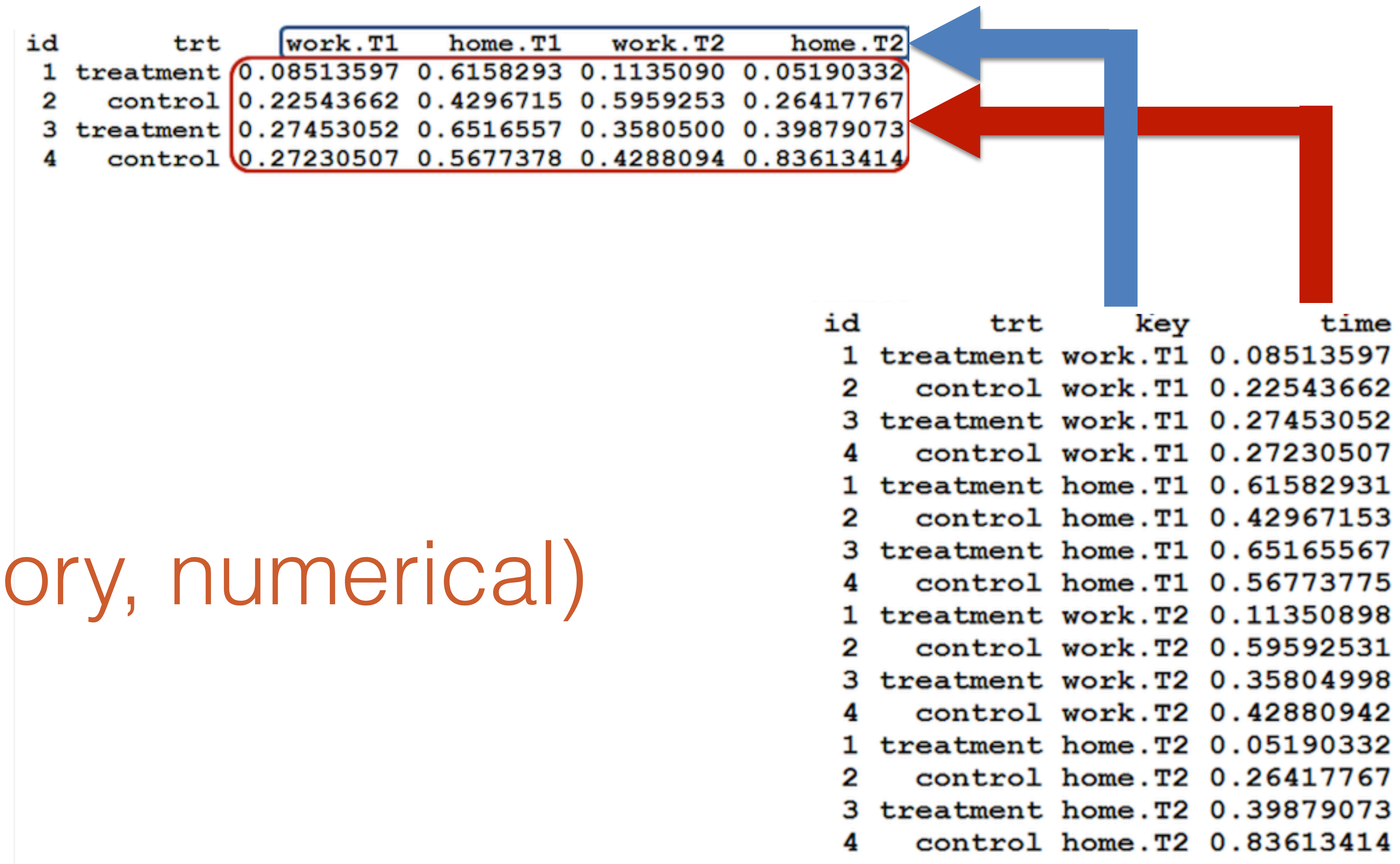
```
messy
id       trt    work.T1   home.T1   work.T2    home.T2
 1 treatment 0.08513597 0.6158293 0.1135090 0.05190332
 2   control 0.22543662 0.4296715 0.5959253 0.26417767
 3 treatment 0.27453052 0.6516557 0.3580500 0.39879073
 4   control 0.27230507 0.5677378 0.4288094 0.83613414
```

```
tidier
 id       trt     key       time
  1 treatment work.T1 0.08513597
  2   control work.T1 0.22543662
  3 treatment work.T1 0.27453052
  4   control work.T1 0.27230507
  1 treatment home.T1 0.61582931
  2   control home.T1 0.42967153
  3 treatment home.T1 0.65165567
  4   control home.T1 0.56773775
  1 treatment work.T2 0.11350898
  2   control work.T2 0.59592531
  3 treatment work.T2 0.35804998
  4   control work.T2 0.42880942
  1 treatment home.T2 0.05190332
  2   control home.T2 0.26417767
  3 treatment home.T2 0.39879073
  4   control home.T2 0.83613414
```

Formula: gather(category, numerical, x:z)

spread() - reshapes data from ‘long’ to ‘wide’

spread(key, time)



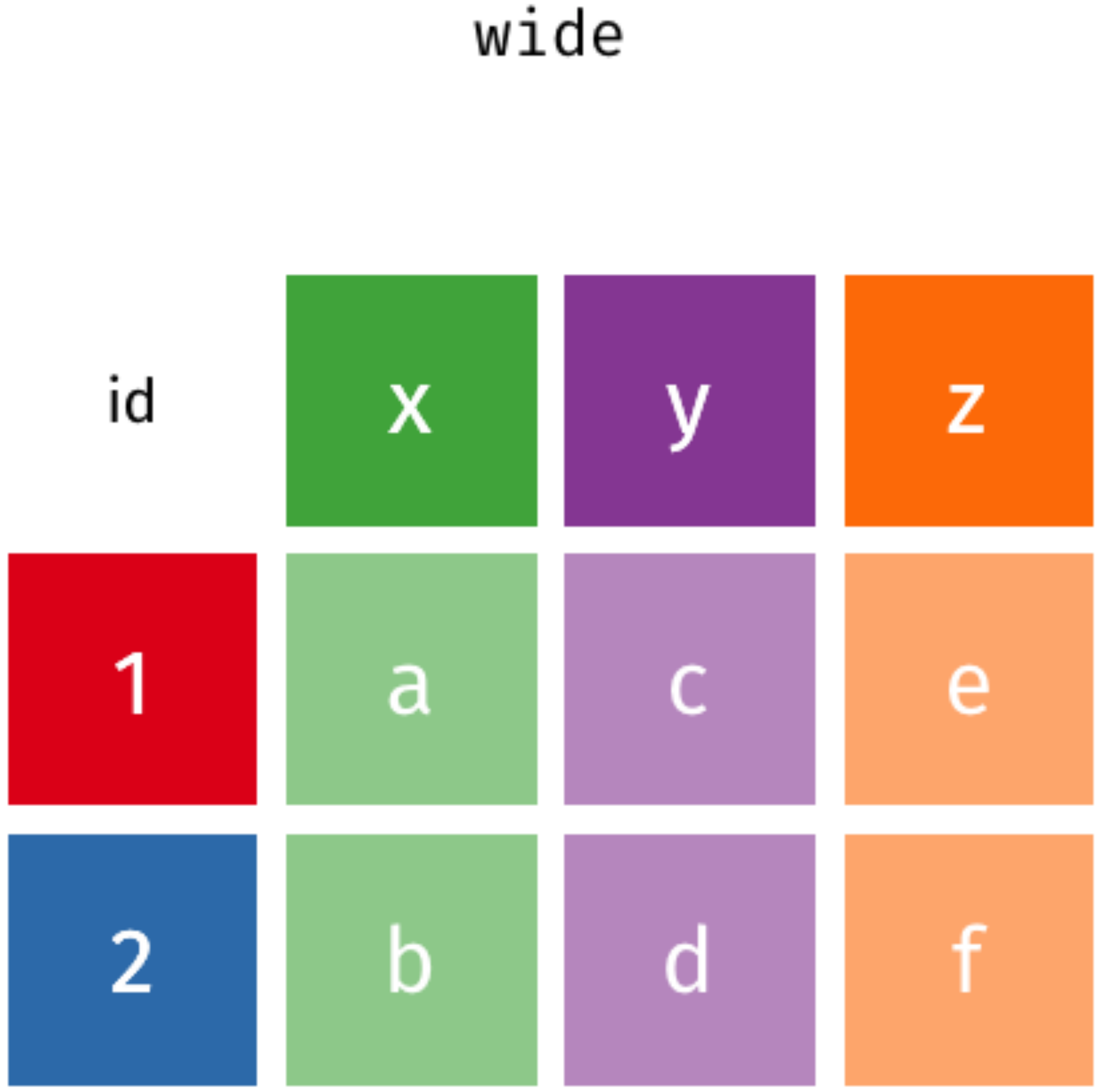
```
id       trt    work.T1   home.T1   work.T2    home.T2
 1 treatment 0.08513597 0.6158293 0.1135090 0.05190332
 2   control 0.22543662 0.4296715 0.5959253 0.26417767
 3 treatment 0.27453052 0.6516557 0.3580500 0.39879073
 4   control 0.27230507 0.5677378 0.4288094 0.83613414
```

```
id       trt     key       time
 1 treatment work.T1 0.08513597
 2   control work.T1 0.22543662
 3 treatment work.T1 0.27453052
 4   control work.T1 0.27230507
 1 treatment home.T1 0.61582931
 2   control home.T1 0.42967153
 3 treatment home.T1 0.65165567
 4   control home.T1 0.56773775
 1 treatment work.T2 0.11350898
 2   control work.T2 0.59592531
 3 treatment work.T2 0.35804998
 4   control work.T2 0.42880942
 1 treatment home.T2 0.05190332
 2   control home.T2 0.26417767
 3 treatment home.T2 0.39879073
 4   control home.T2 0.83613414
```

Formula: spread(category, numerical)

gather() - reshapes data from ‘wide’ to ‘long

spread() - reshapes data from ‘long’ to ‘wide’



wide

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

separate() - split single column to many

```
separate(key, into=c(“location”, ”when”), sep = “.”)
```

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998
4	control	work	T2	0.42880942
1	treatment	home	T2	0.05190332
2	control	home	T2	0.26417767
3	treatment	home	T2	0.39879073
4	control	home	T2	0.83613414

unite() - combine multiple columns

unite(key, location, when, sep = “.”)

id	trt	key	time
1	treatment	work.T1	0.08513597
2	control	work.T1	0.22543662
3	treatment	work.T1	0.27453052
4	control	work.T1	0.27230507
1	treatment	home.T1	0.61582931
2	control	home.T1	0.42967153
3	treatment	home.T1	0.65165567
4	control	home.T1	0.56773775
1	treatment	work.T2	0.11350898
2	control	work.T2	0.59592531
3	treatment	work.T2	0.35804998
4	control	work.T2	0.42880942
1	treatment	home.T2	0.05190332
2	control	home.T2	0.26417767
3	treatment	home.T2	0.39879073
4	control	home.T2	0.83613414

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998
4	control	work	T2	0.42880942
1	treatment	home	T2	0.05190332
2	control	home	T2	0.26417767
3	treatment	home	T2	0.39879073
4	control	home	T2	0.83613414

tidyr - tidy up a dataset

- **gather()** - ‘wide’ to ‘long’
- **spread()** - ‘long’ to ‘wide’
- **separate()** - split up a column
- **unite()** - merge multiple columns

Demo!

dplyr verbs:

filter()

select()

rename()

arrange()

mutate()

group_by()

summarise/summarize()

filter()- picks rows based on values

Fruit	Count
Apple	34
Raspberry	67
Pear	35
Plum	27
Peach	5
Strawberry	2
Melon	97
Mango	5

```
filter(Fruit == “Raspberry”)
```

Fruit	Count
Raspberry	67

```
filter(Count < 10)
```

Fruit	Count
Peach	5
Strawberry	2
Mango	5

filter()- picks **rows** based on values

```
filter(column == “value”)

filter(year <= 1995)

filter(column %in% c(“Mary”, ”Mari”))

filter(column %in% c(“Mary”, ”Mari”) & year >1940)

filter(!name == “Dave”)  - filters out/omits
```

Try to isolate all :

Flights on May 9th

Flights in January and February

Flights to LAX and SFO

Flights delayed by >60min

Flights that departed between 12am and 6am

select() - pick specific columns

select(2:49)

select(Day, Month, Year)

select(-xlkjgtklj) - removes “xlkjgtklj”

select(starts_with(delay): names starts with delay)

rename() - change column names

Formula: rename(new_column = old_column)

```
rename( patient_ID = id,
        hours = time)
```

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

patient_ID	trt	location	when	hours
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

arrange()- changes row order

Fruit	Count
Apple	34
Raspberry	67
Pear	35
Plum	27
Peach	5
Strawberry	2
Melon	97
Mango	5

```
arrange(desc(Count))
```

Fruit	Count
Melon	97
Raspberry	67
Pear	35
Apple	34
Mango	5
Peach	5

mutate() - create **new column** from **existing data**

Formula: **mutate**(new_column = columnA - columnB)
mutate(new_column = columnA * columnB)
mutate(new_column = log2(columnA) / columnB)

mutate(minutes = time * 60)

id	trt	location	when	time
1	treatment	work	T1	0.08513597
2	control	work	T1	0.22543662
3	treatment	work	T1	0.27453052
4	control	work	T1	0.27230507
1	treatment	home	T1	0.61582931
2	control	home	T1	0.42967153
3	treatment	home	T1	0.65165567
4	control	home	T1	0.56773775
1	treatment	work	T2	0.11350898
2	control	work	T2	0.59592531
3	treatment	work	T2	0.35804998

id	trt	location	when	time	minutes
1	treatment	work	T1	0.08513597	5.1081582
2	control	work	T1	0.22543662	13.5261972
3	treatment	work	T1	0.27453052	16.4718312
4	control	work	T1	0.27230507	16.3383042
1	treatment	home	T1	0.61582931	36.9497586
2	control	home	T1	0.42967153	25.7802918
3	treatment	home	T1	0.65165567	39.0993402
4	control	home	T1	0.56773775	34.064265
1	treatment	work	T2	0.11350898	6.8105388
2	control	work	T2	0.59592531	35.7555186
3	treatment	work	T2	0.35804998	21.4829988

Demo!

- **group_by()**- ‘**lock-in**’ by certain criteria
- **summarize()** - **reduce** multiple values to a **single value**

Cat	Fruit	Count
1	Apple	34
1	Raspberry	67
1	Pear	35
1	Plum	27
2	Peach	5
2	Strawberry	2
2	Melon	97
2	Mango	5

```
data %>%
     group_by(Cat) %>%
     summarize( Total = sum(Count))
```

Cat	Total
1	163
2	109

Try to:

Compute speed in mph from (time) and distance (miles)

- Which flight flew fastest?

What was the longest flight delay in JFK in November?

Which flights departed from LGA arrived to DTW early?

dplyr verbs:

filter() - pick specific **rows**

select() - pick specific **columns**

rename() - **change** column names

arrange() - **sort** by row values

mutate() - add **new column** from existing data

group_by() - ‘**lock-in**’ by variables

summarise/summarize() - **reduce** multiple values to a **single value**

Try to determine:

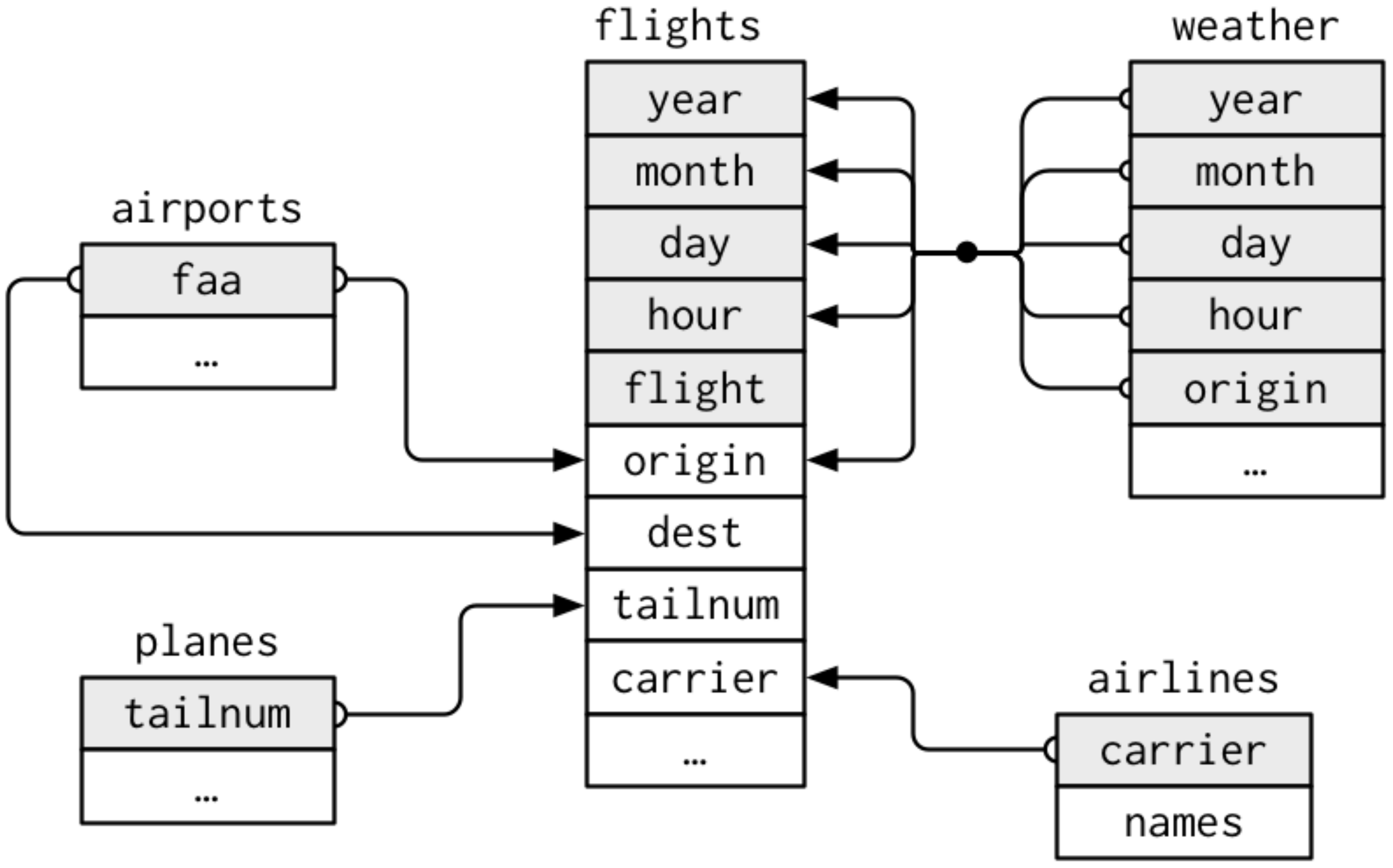
Which airport had the most flights in December?

Which NYC airport has the most airlines?

How many United Airlines flights depart from JFK to ORD?

Joins combine datasets based off of set values

nycflights13 package



flights

year

month

day

hour

flight

origin

dest

tailnum

carrier

…

weather

year

month

day

hour

origin

…

airports

faa

…

planes

tailnum

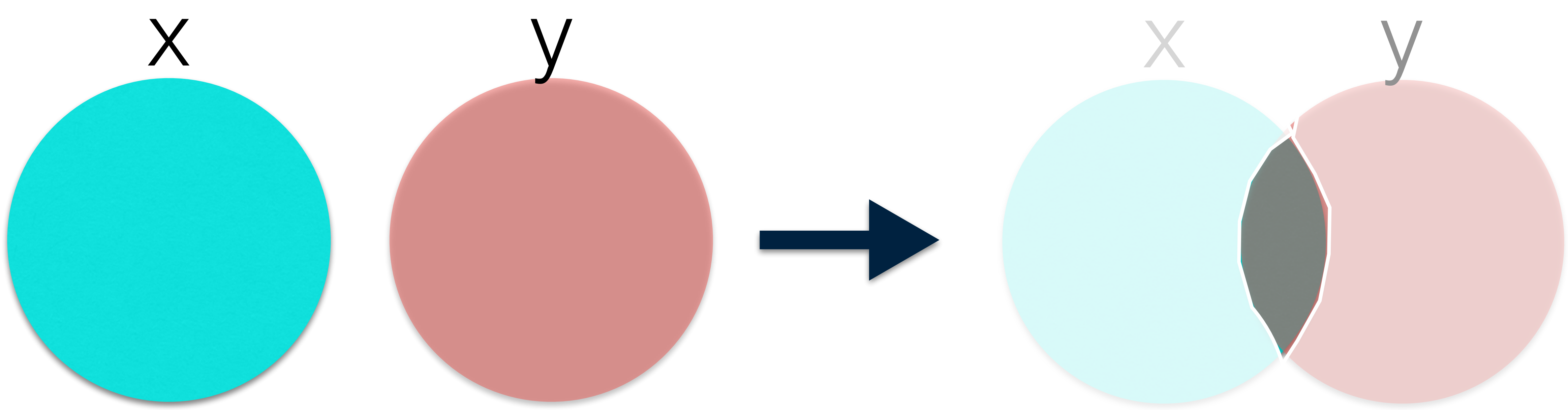
…

airlines

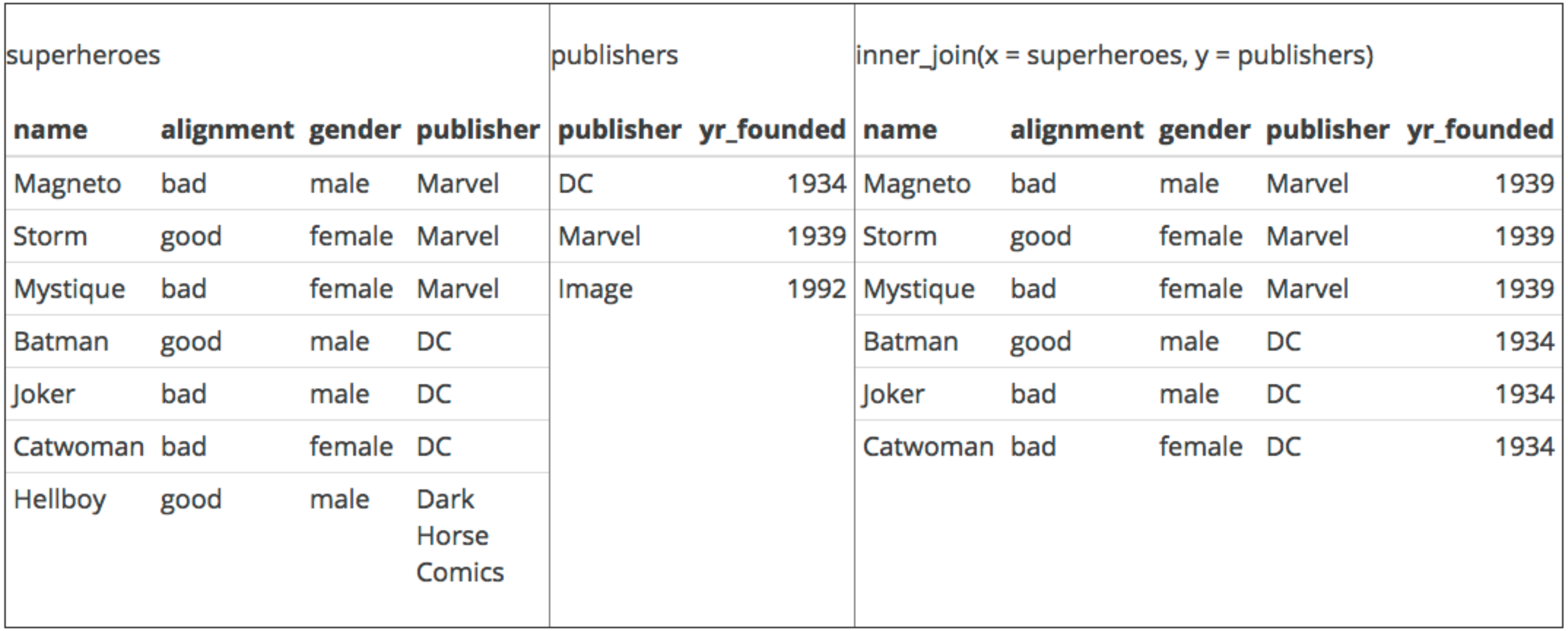
carrier

names

inner_join(x, y)



- **combine things in common between x and y**



superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

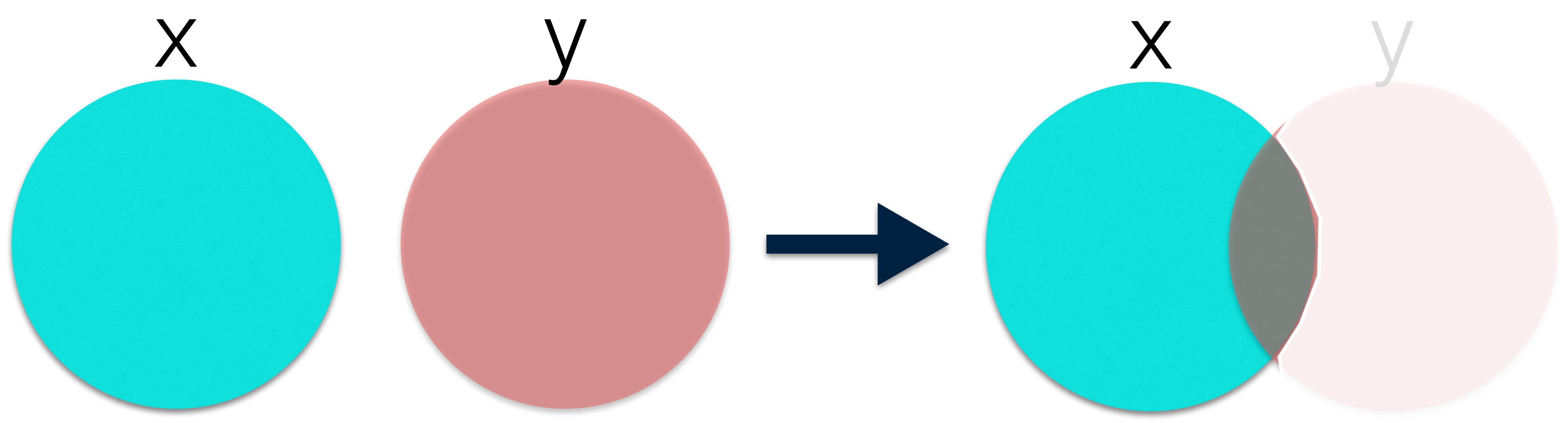
publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

inner_join(x = superheroes, y = publishers)

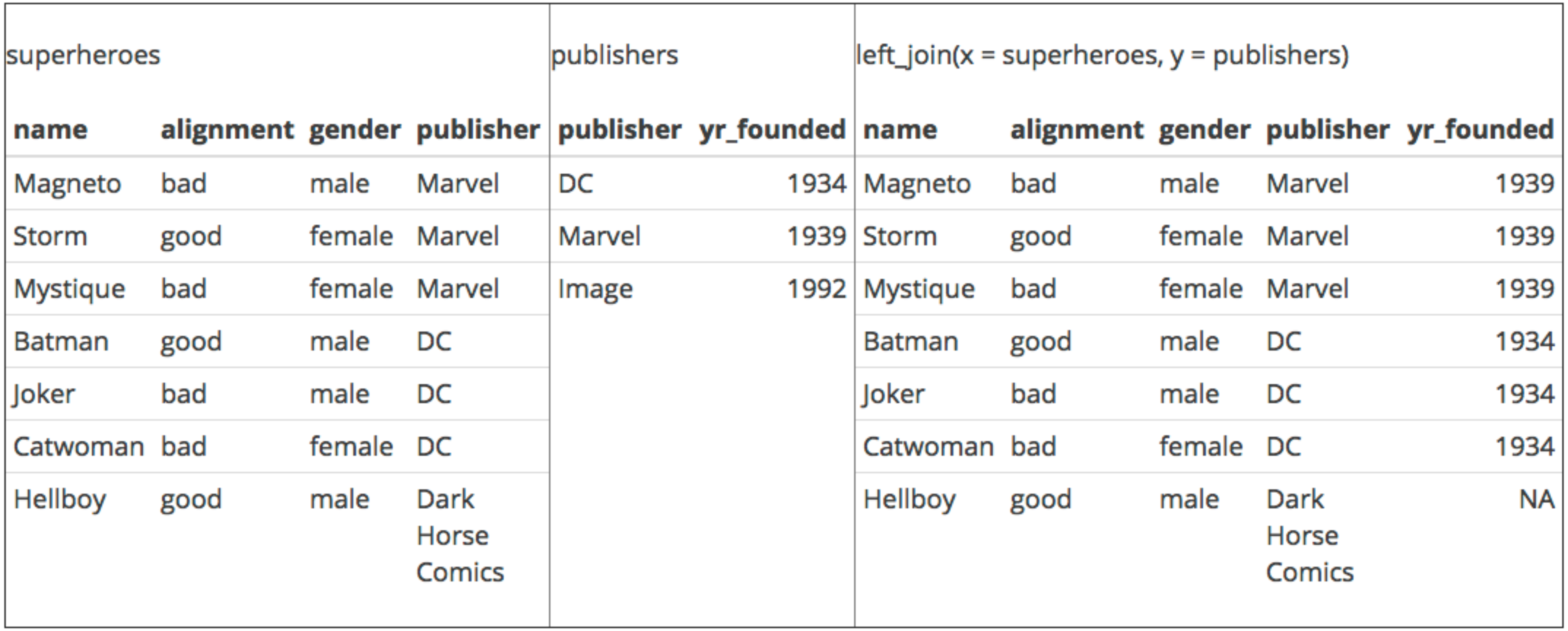
name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934

(source:Jenny Bryan - Stat545)

left_join(x, y)



return all rows of x and all columns from x and y



superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

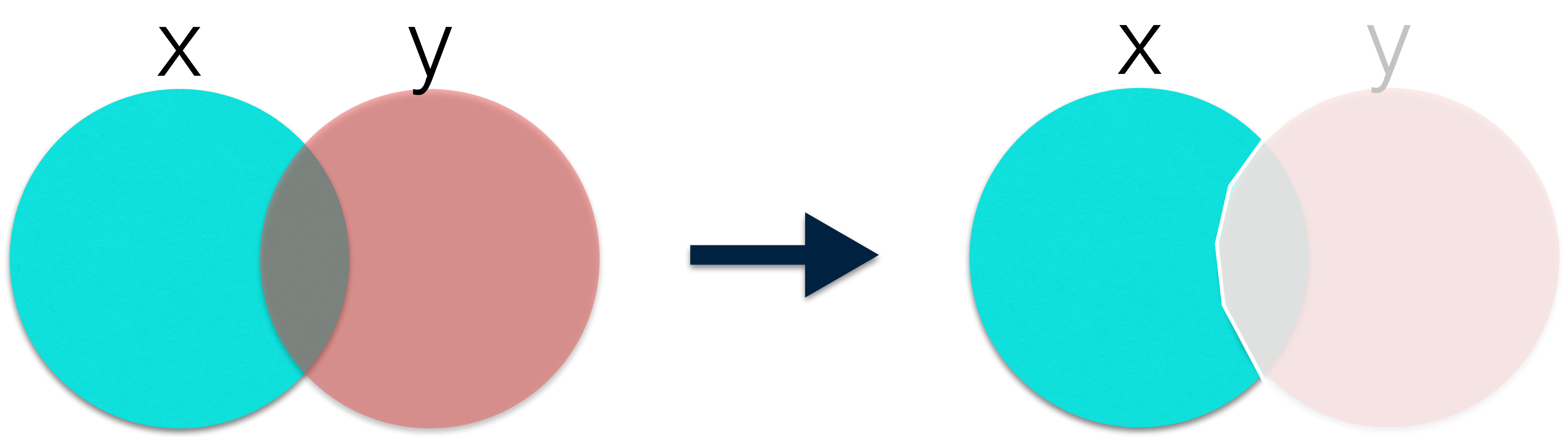
left_join(x = superheroes, y = publishers)

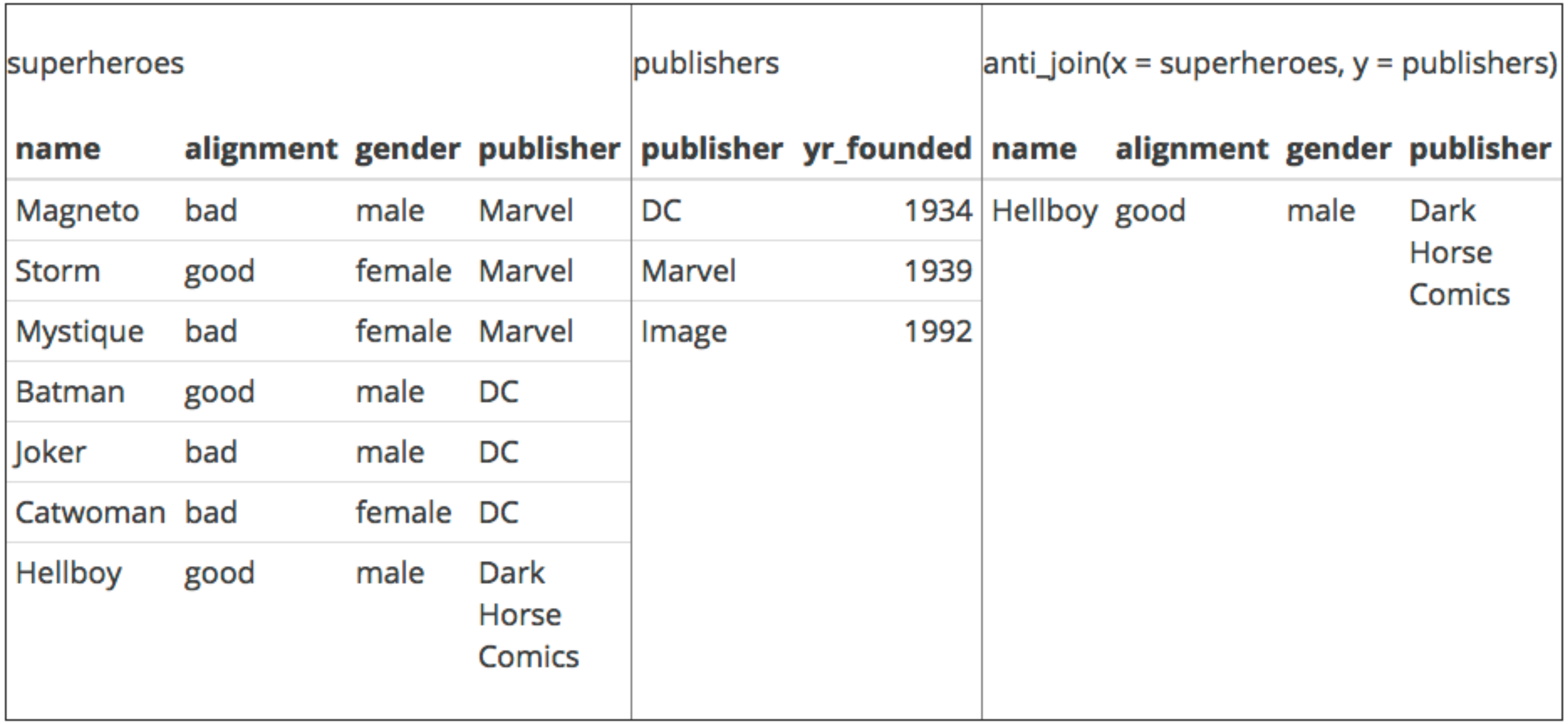
name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA

(source:Jenny Bryan - Stat545)

anti_join(x, y)

keep what is distinct in x only





superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

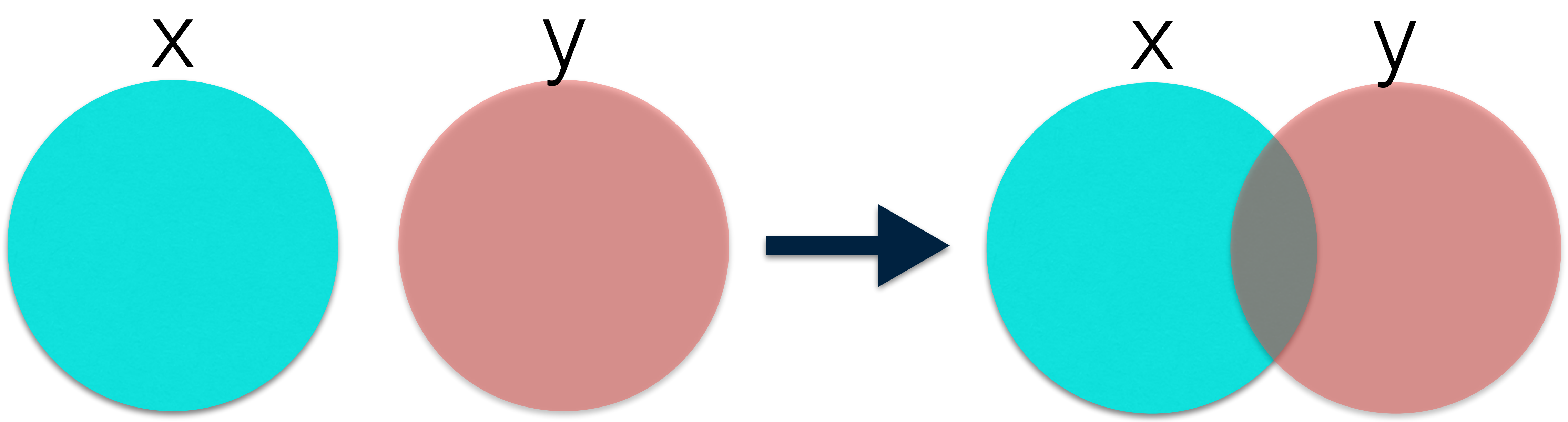
publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

anti_join(x = superheroes, y = publishers)

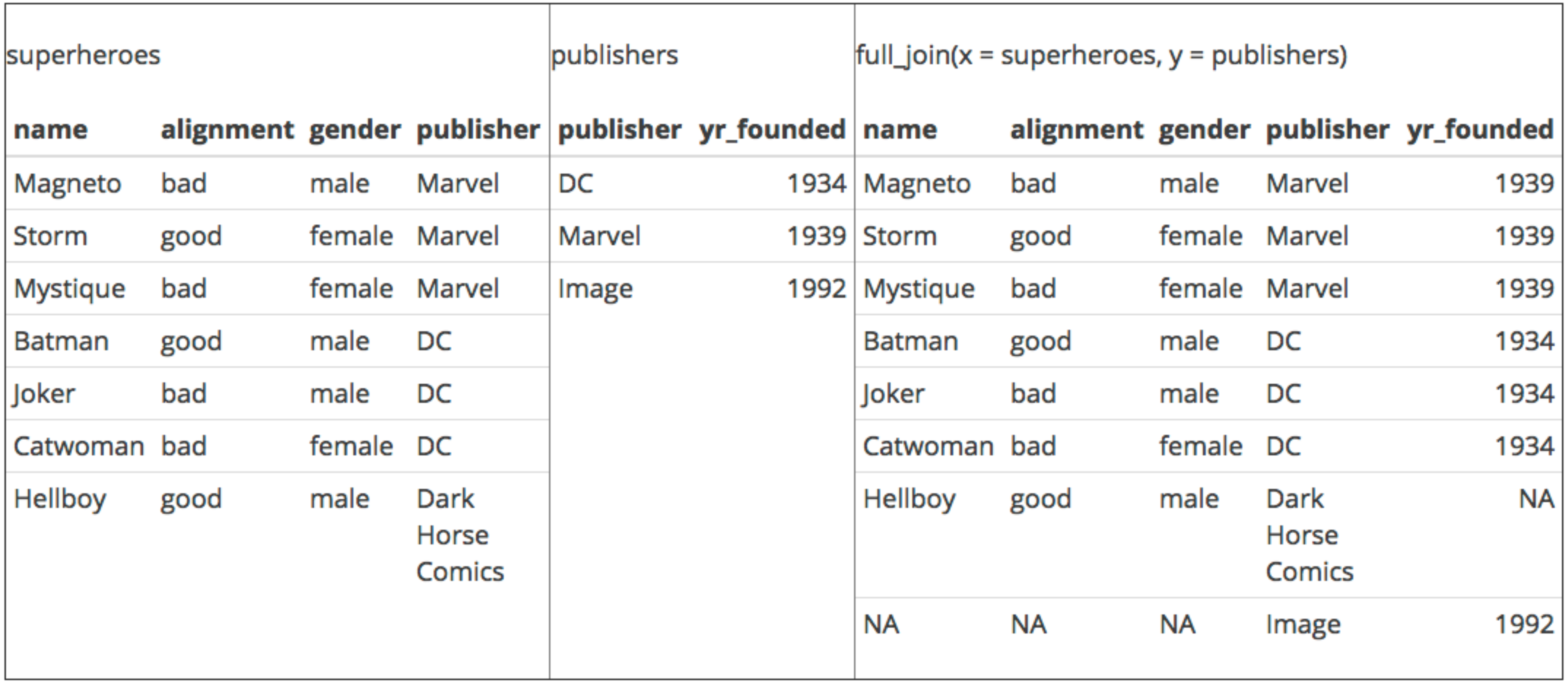
name	alignment	gender	publisher
Hellboy	good	male	Dark Horse Comics

(source:Jenny Bryan - Stat545)

full_join(x,y)



combine x and y, will introduce NAs



superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

full_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA
NA	NA	NA	Image	1992

(source:Jenny Bryan - Stat545)

Demo!