Lists in R

Data structure: a structure, usually with a specific class, that holds data. eg vector, matrix, array, list, tibble

We are already familiar somewhat with vectors

Vectors are a type of datastructure.

```
fib <- c(1, 2, 3, 5, 8, 13)
```

* all values in the list have the same class

class(fib) is ?

```
she_loves_me_she_loves_me_not <-
    c(TRUE, FALSE, TRUE, FALSE, TRUE)
```

```
word_on_street <- c("I", "am", "the", "king")
```

Vectors are very important in R.

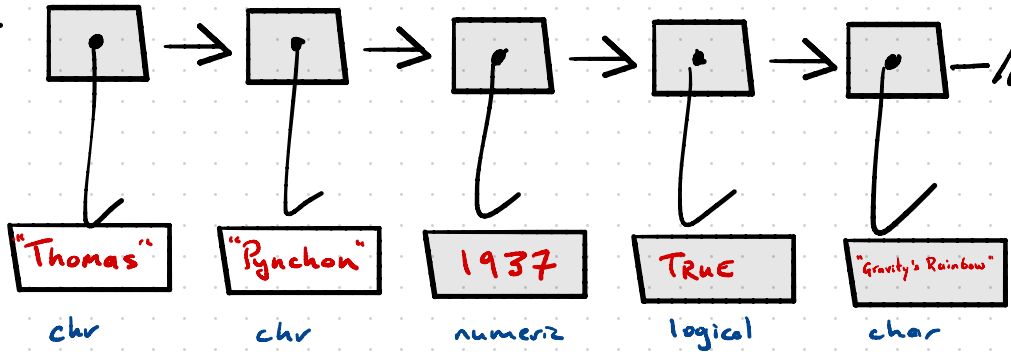
Think of one column (variable) in a tibble.

Lists are more general, more flexible.

lists are another type of data structure.

```
L ← list("Thomas", "Pynchon", 1937,
         TRUE, "Gravity's Rainbow")
```

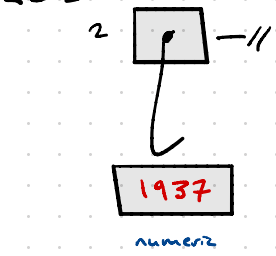
L2



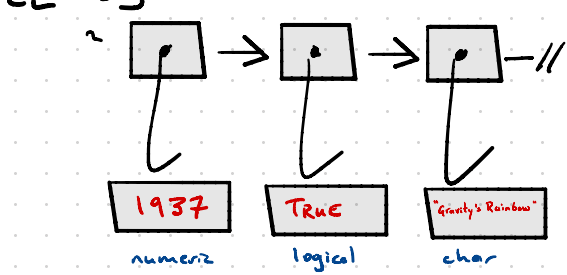

```
class(L); length(L); names(L)
```

list 5 NULL

L[3]

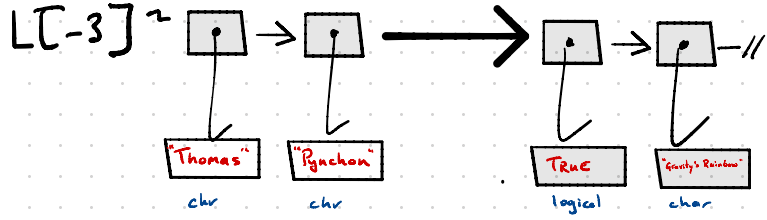


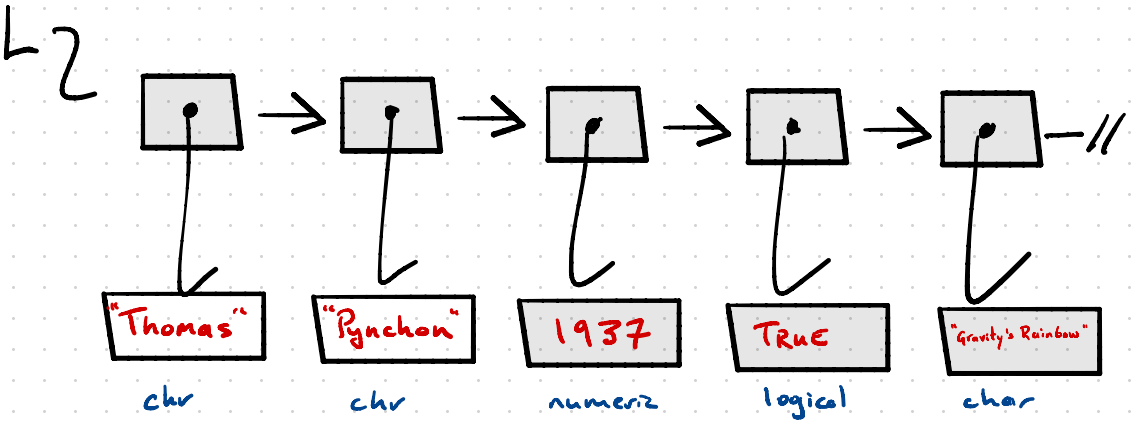
L[3:5]

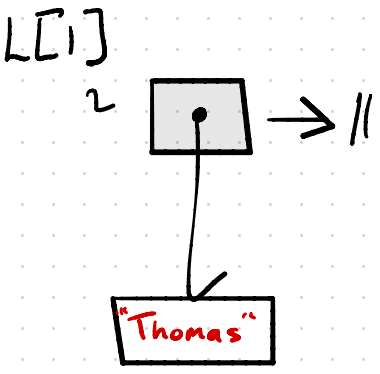


class(L[-3]) = ???

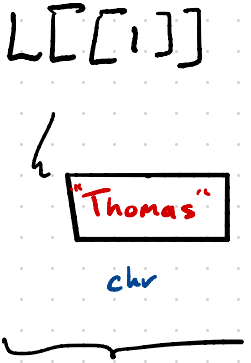
L[-3]







THIS IS STILL A LIST (WITH JUST ONE ITEM)



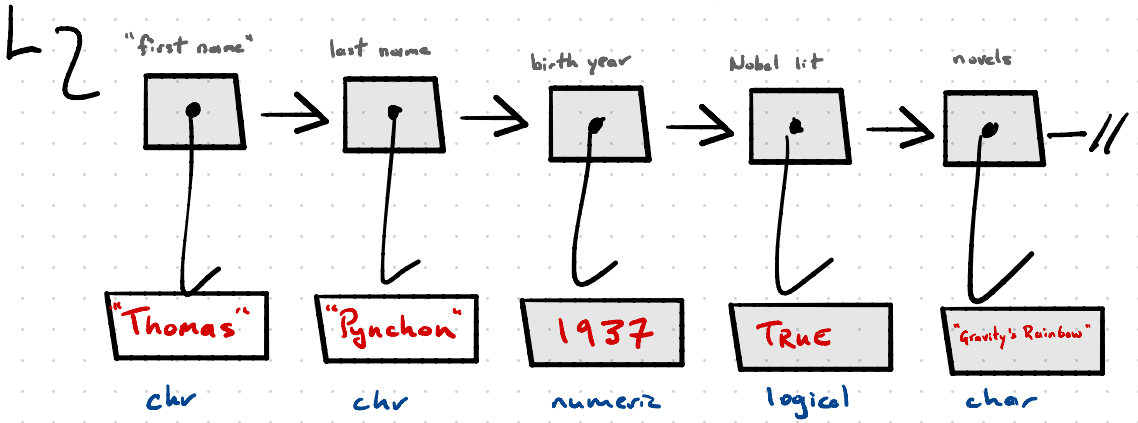
THIS IS THE VALUE IN THE LIST
"Thomas" of class chr.

L[[1:3]] ?

L[[1]]
L[[2]]
L[[3]]

Naming Elements of Lists

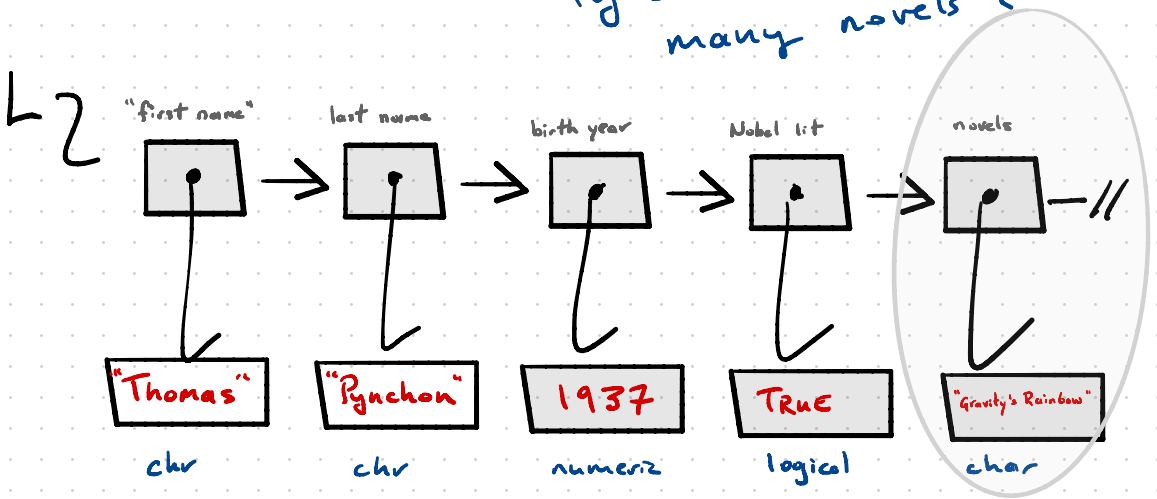
names(L) ← c("first name", "last name", "birth year", "Nobel Lit", "novels")



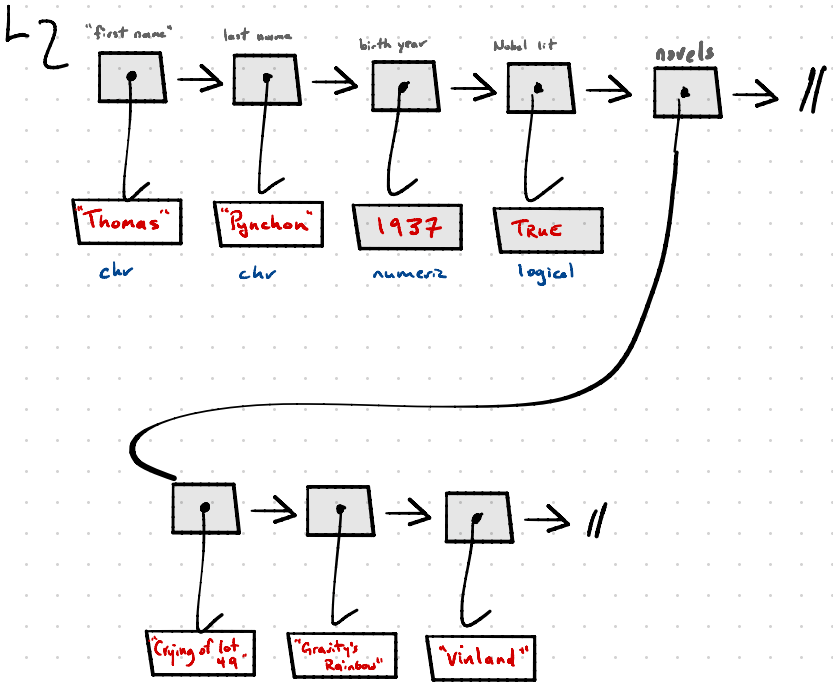
L$"birth year" Same as L[[3]]

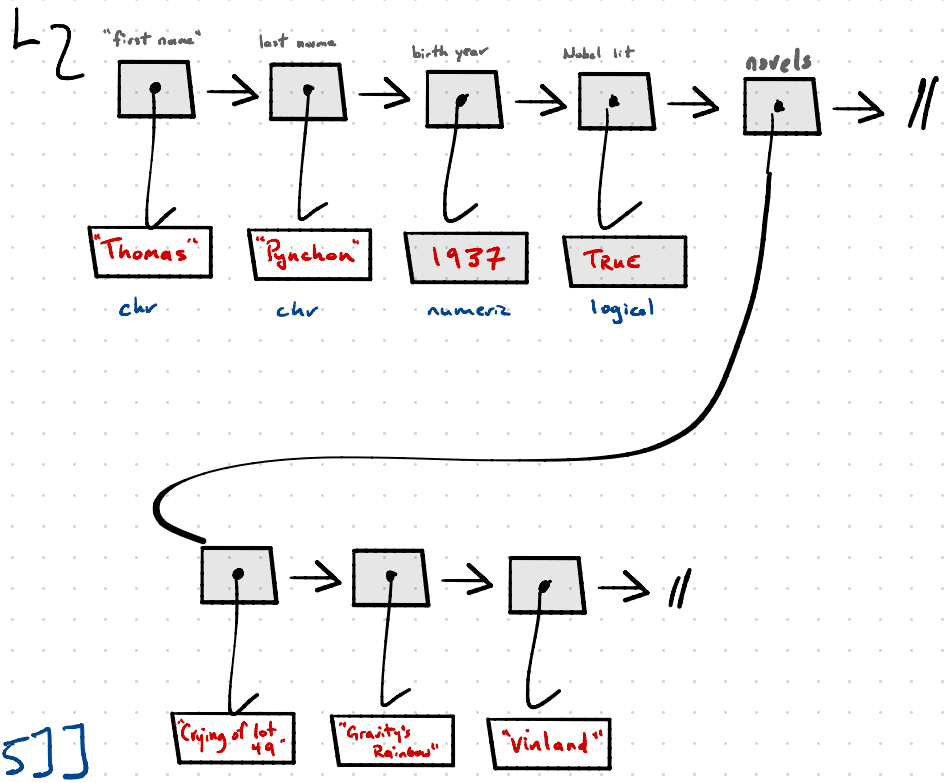
L$"novels" same as L[[5]]

Pynchon has written many novels?



Lists can even contain lists.

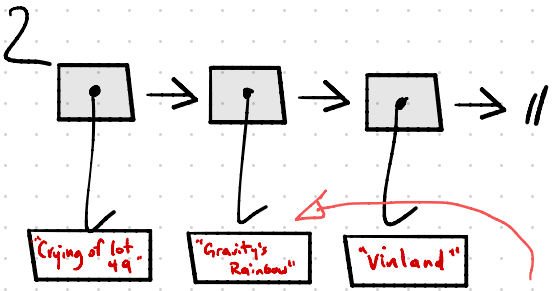




L[[5]]

L$novels

L$"novels" ?

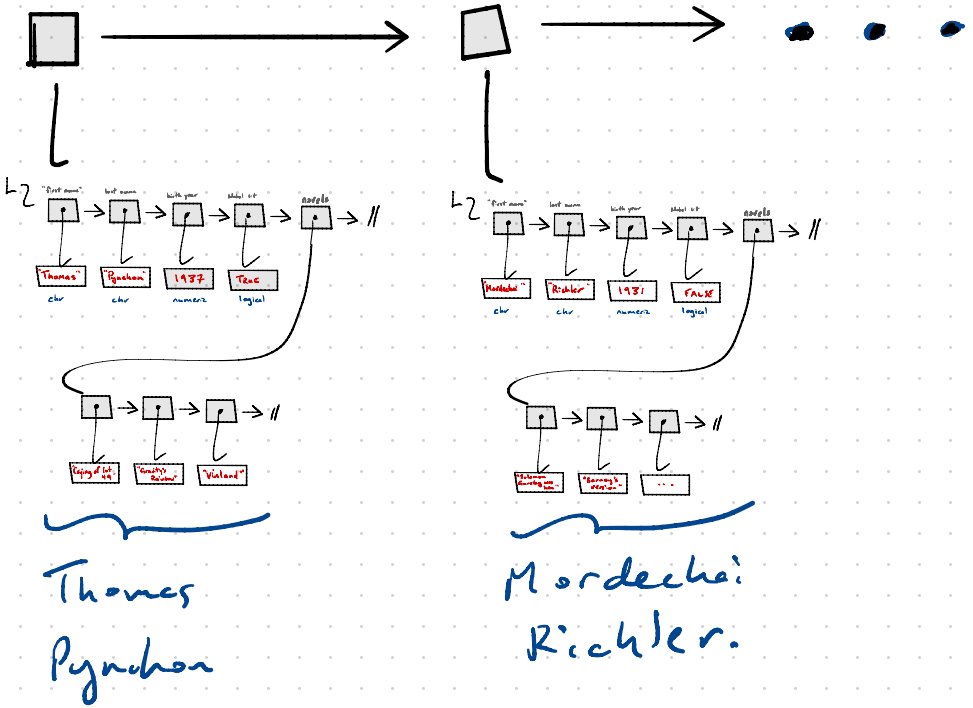


L$novels[2:3] ?

L$novels[[2]] ?

Literature

2



Thomas Pynchon

Mordecai Richler.

DAVID SANKOFF

UdeM

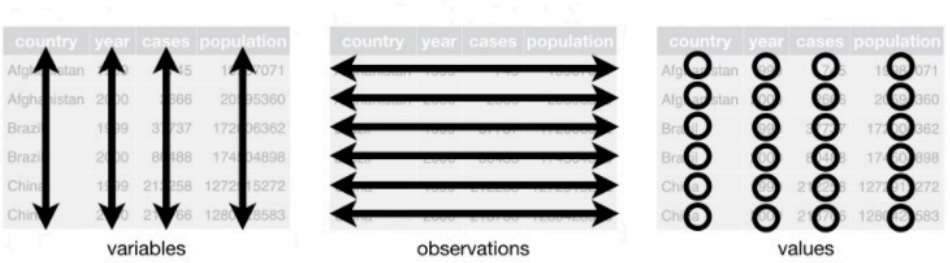
U de Ottawa.

Tidy Data

There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Figure 12.1 shows the rules visually.



This means for most real analyses, you'll need to do some tidying. The first step is always to figure out what the variables and observations are. Sometimes this is easy; other times you'll need to consult with the people who originally generated the data. The second step is to resolve one of two common problems:

1. One variable might be spread across multiple columns.
2. One observation might be scattered across multiple rows.

Action Item:

Find examples in the Tara Oceans or TCGA data.

ESR1 is one variable but it has 2 columns.

LONGER

① Tm

participant	at diagnosis ESR1	after treatment ESR1
AINF	628	527
⋮	⋮	⋮

```
T %>%
pivot_longer(c("at diagnosis ESR1", "after treatment ESR1"),
             names_to = "measurement point",
             values_to = "expression")
```

tibble ~

participant	at diagnosis ESR1	after treatment ESR1
AINF	628	527
⋮	⋮	⋮

part.	measurement pt.	expressio-
AINF	at diagnosis ESR	628
AINF	after treatment	527

(2)

WIDER — One obs. across many rows.

T2

participant	gene	count
AINF	ESR1	628
AINF	ANLN	527

2 rows

AINF IS ONE OBSERVATION (FOR WHICH WE MEASURE MANY VARIABLES).

```
T %>%
  pivot_wider( names_from = gene, values_from = count )
```

parti	ESR1	ANLN
AINF	628	527.

- separate - when a single variable has >1 pieces of information (values).

ıta published at	Station identifier [TARA_station#]	Date/Time [yyyy-mm-ddThh:mm]	Latitude [degrees North]	Lon [deg East
h?All&q=TARA_X00000...	TARA_004	2009-09-15 18:00:00	36.5533	
h?All&q=TARA_Y20000...	TARA_004	2009-09-15 11:30:00	36.5533	
h?All&q=TARA_A20000...	TARA_007	2009-09-23 16:08:00	37.0541	
h?All&q=TARA_A20000...	TARA_007	2009-09-23 12:50:00	37.0510	
h?All&q=TARA_X00000...	TARA_009	2009-09-28 16:59:00	39.0609	
h?All&q=TARA_X00000...	TARA_009	2009-09-28 12:18:00	39.1633	

Showing 1 to 7 of 243 entries, 17 total columns

Console Terminal Jobs

/cloud/project/

```
#   www.marineregions.com]` <chr>
>
> View(T2)
> T2$`Date/Time [yyyy-mm-ddThh:mm]`
  [1] "2009-09-15 18:00:00 UTC" "2009-09-15 11:30:00 UTC"
  [3] "2009-09-23 16:08:00 UTC" "2009-09-23 12:50:00 UTC"
  [5] "2009-09-28 16:59:00 UTC" "2009-09-28 12:18:00 UTC"
  [7] "2009-11-02 14:07:00 UTC" "2009-11-02 14:07:00 UTC"
  [9] "2009-11-02 08:13:00 UTC" "2009-11-02 08:13:00 UTC"
```

T1 %>%

separate(`Date/Time...`, into = c("date", "time"), sep = " ")

space

not optimal

...	date <chr>	time <chr>	...
	"2009-09-15"	"18:00:00"	

Tidy data should all be in the same tibble.

```
> T1
# A tibble: 243 x 18
   sample_label ac_sample ac_run eng… pangea_id pange_data station date       time  latitude longitude depth feature
   <chr>        <chr>     <chr>  <chr> <chr>     <chr>      <chr>   <date>     <tim>    <dbl>     <dbl> <dbl> <chr>
 1 TARA_004_DC… ERS487936 ERR59… http… TARA_X00… http://ww… TARA_0… 2009-09-15 18:00     36.6     -6.57    40 (DCM) …
 2 TARA_004_SR… ERS487899 ERR59… http… TARA_Y20… http://ww… TARA_0… 2009-09-15 11:30     36.6     -6.57     5 (SRF) …
 3 TARA_007_DC… ERS477953 ERR31… http… TARA_A20… http://ww… TARA_0… 2009-09-23 16:08     37.1      1.95    42 (DCM) …
 4 TARA_007_SR… ERS477931 ERR31… http… TARA_A20… http://ww… TARA_0… 2009-09-23 12:50     37.1      1.94     5 (SRF) …
 5 TARA_009_DC… ERS488147 ERR59… http… TARA_X00… http://ww… TARA_0… 2009-09-28 16:59     39.1      5.94    55 (DCM) …
 6 TARA_009_SR… ERS488119 ERR59… http… TARA_X00… http://ww… TARA_0… 2009-09-28 12:18     39.2      5.92     5 (SRF) …
 7 TARA_018_DC… ERS488346 ERR59… http… TARA_S20… http://ww… TARA_0… 2009-11-02 14:07     35.8     14.3     60 (DCM) …
 8 TARA_018_DC… ERS488354 ERR59… http… TARA_A10… http://ww… TARA_0… 2009-11-02 14:07     35.8     14.3     60 (DCM) …
 9 TARA_018_SR… ERS488330 ERR59… http… TARA_A10… http://ww… TARA_0… 2009-11-02 08:13     35.8     14.3      5 (SRF) …
10 TARA_018_SR… ERS488340 ERR59… http… TARA_A10… http://ww… TARA_0… 2009-11-02 08:13     35.8     14.3      5 (SRF) …
# … with 233 more rows, and 5 more variables: size_lower <chr>, size_upper <dbl>, pelagic_biomes <chr>, regions <chr>,
#   `pelagic_biomes MRGID` <chr>
> T5
# A tibble: 141 x 5
   pangea_id       `16S_miTAGs` richess chao1 shannon_diversity
   <chr>                  <dbl>   <dbl> <dbl>             <dbl>
 1 TARA_B100000965       103040   3053. 4404.              6.83
 2 TARA_B100000959       111240   3172. 4805.              6.68
 3 TARA_B100000963       122482   2412. 3504.              6.60
 4 TARA_B100000902       112336   2931. 4241.              6.78
 5 TARA_B100000953        56449   3108. 4723.              6.68
 6 TARA_B100000900        74287   2293. 3198.              6.62
 7 TARA_B100000927        77265   2501. 3567.              6.68
 8 TARA_B100000929       110154   3372. 4999.              6.83
 9 TARA_B100000925       128823   2356. 3372.              6.65
10 TARA_B100001113       128677   2874. 4207.              6.80
# … with 131 more rows
> T8
# A tibble: 245 x 37
   pangea_id data                latitude longtitude  depth `temp[C]` `salinity[PSU]` `oxygen[umol/kg… `Mean_Nitrates[…
   <chr>     <dttm>                 <dbl>      <dbl>  <dbl>     <dbl>           <dbl>            <dbl>             <dbl>
 1 TARA_B10… 2011-04-15 13:10:00   -13.0       -96.0  57.6      20.6            35.5            217.               1.50
 2 TARA_B10… 2011-04-16 16:09:00   -12.9       -96.1 175.       13.0            34.8              0.708           21.1
 3 TARA_B10… 2011-04-15 13:10:00   -13.0       -96.0   5.48     25.3            35.8            200.               4.59
 4 TARA_B10… 2011-04-22 19:51:00    -5.27      -85.2  45.7      19.6            34.9            104.              20.3
 5 TARA_B10… 2011-04-22 14:16:00    -5.27      -85.2 476.        9.20           34.7              4.43            40.0
 6 TARA_B10… 2011-04-21 20:16:00    -5.25      -85.2   5.48     24.9            34.7            206.              11.7
 7 TARA_R10… 2011-05-13 00:00:00     2.08      -84.5  29.8      26.5            34.3            203.               3.89
 8 TARA_B10… 2011-05-13 00:00:00     2.08      -84.5  29.8      26.5            34.3            203.               3.89
 9 TARA_B10… 2011-05-13 19:44:00     2.07      -84.5 376.       11.3            34.8              2.52            33.9
10 TARA_R10… 2011-05-12 13:38:00     1.99      -84.6   5.40     27.6            33.4            199.               1.21
# … with 235 more rows, and 28 more variables: `NO2 [umol/L]**` <dbl>, `PO4 [umol/L]**` <dbl>, `NO2NO3 [umol/L]**` <dbl>,
#   `SI [umol/L]**` <dbl>, `AMODIS:PAR8d,Einsteins/m-2/d-1` <dbl>, `Okubo-Weiss` <dbl>, Lyapunov_exp. <dbl>,
#   grad_SST_adv <dbl>, retention <dbl>, `Mean Depth MLD Sigma [m]*` <dbl>, `Mean Depth Max Fluo [m]*` <dbl>, `Mean Depth
#   Max N2 [m]*` <dbl>, `Mean Depth Max O2 [m]*` <dbl>, `Mean Depth Min O2 [m]*` <dbl>, `Mean Depth Nitracline
#   [m]*` <dbl>, miTAG.SILVA.Taxo.Richness <dbl>, miTAG.SILVA.Phylo.Diversity <dbl>, miTAG.SILVA.Chao <dbl>,
#   miTAG.SILVA.ace <dbl>, miTAG.SILVA.Shannon <dbl>, OG.Shannon <dbl>, OG.Richness <dbl>, OG.Evenness <dbl>, `FC -
#   heterotrophs [cells/mL]` <dbl>, `FC - autotrophs [cells/mL]` <dbl>, `FC - bacteria [cells/mL]` <dbl>, `FC -
#   picoeukaryotes [cells/mL]` <dbl>, `minimum generation time [h]` <dbl>
```

Quite straight forward to link this data.

pangea_id is both a foreign & primary key. (Section 13.3 keys)

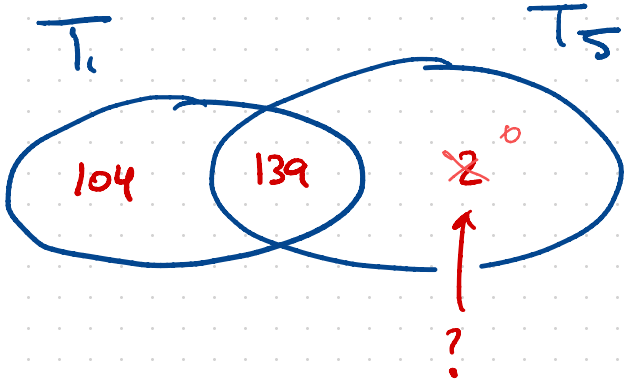
Start with tables T_1 & T_5.

T_1

pangea_id	sample_label	...
TARA_X00...		

T_5

pangea_id	16S_mTAGr	...
TARA_B1...		



In general ...

T_1

key	value	...
T0	5	
T1	7	
T2	3	

T_5

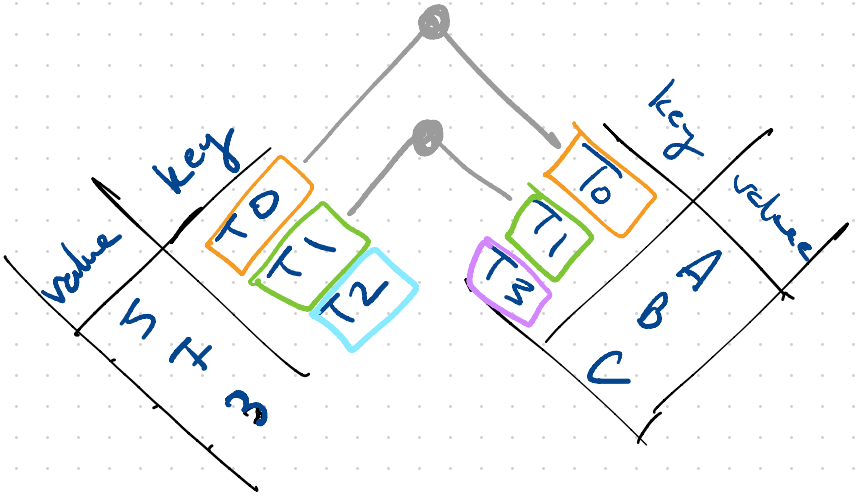
key	value	...
T0	A	
T1	B	
T3	C	

T_1

key	value	...
T0	5	..
T1	7	
T2	3	

T_5

key	value	....
T0	A	...
T1	B	
T3	C	



Inner join:

like an "and" operation

key must be in both tibbles.

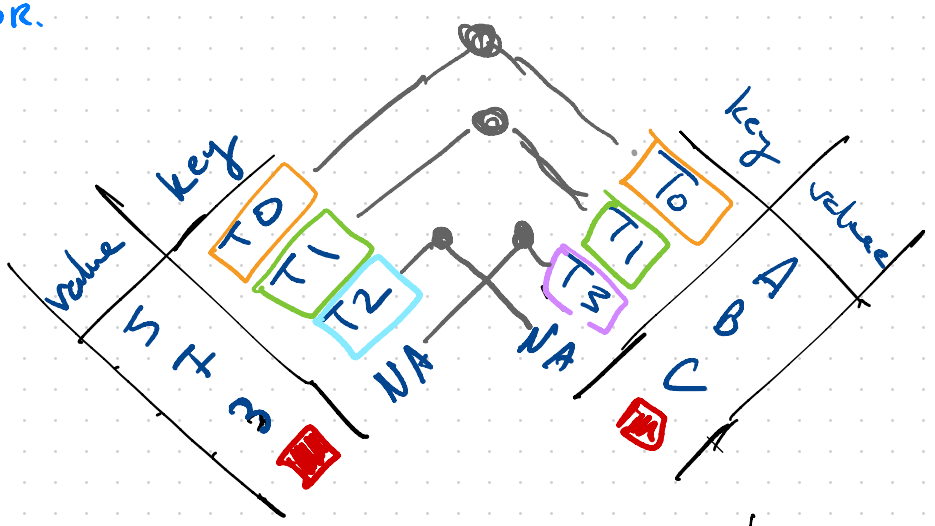
key	$value_{T_1}$	$value_{T_5}$
T0	5	A
T1	7	B

Outer joins: keeps observations (rows) in at least one of the tables.

3 types:

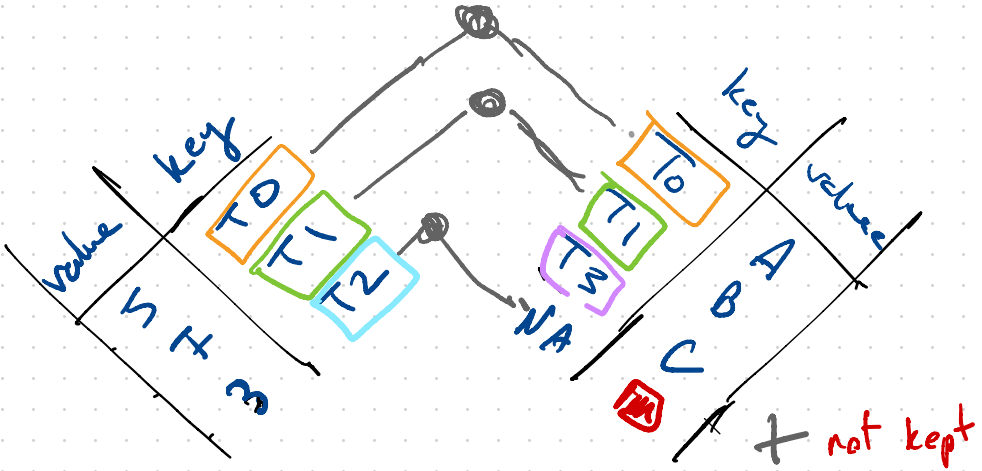
Full join: keeps all observations

↑ sort of like an OR.



key	$value_{T_1}$	value T_5
T_0	5	A
T_1	7	B
T_2	3	NA
T_3	NA	C

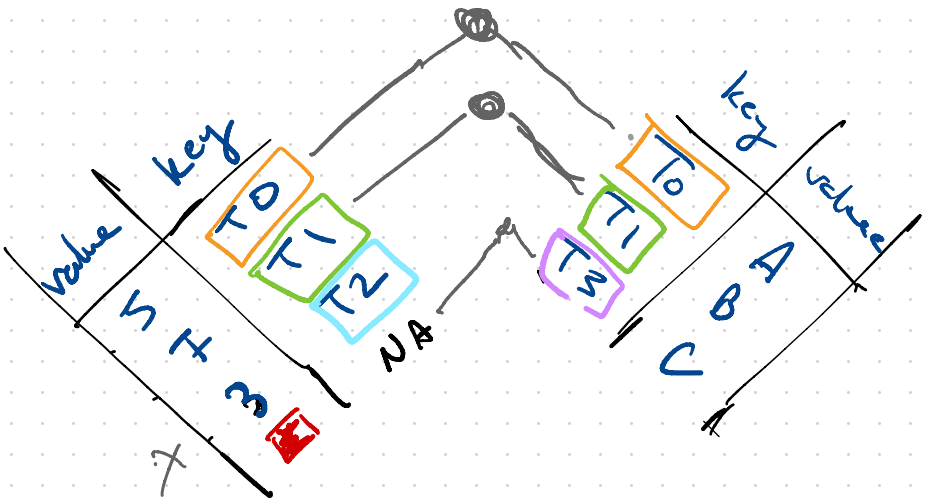
Left join keep everything in first (left) tibble



key	value $_{T_1}$	value $_{T_3}$
T_0	5	A
T_1	7	B
T_2	3	NA

Right join

keep everything in fibble



key	value $_{T_1}$	value $_{T_5}$
T0	5	A
T1	7	B
T3	NA	C

16sRNA miTAG sequences : 500 Mb.

One sample file. What variable in T does this correspond to?

Sample Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- σ^2: variance
- $\sum$: sigma means sum
- n: # of points
- $n-1$: degrees of freedom

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$$