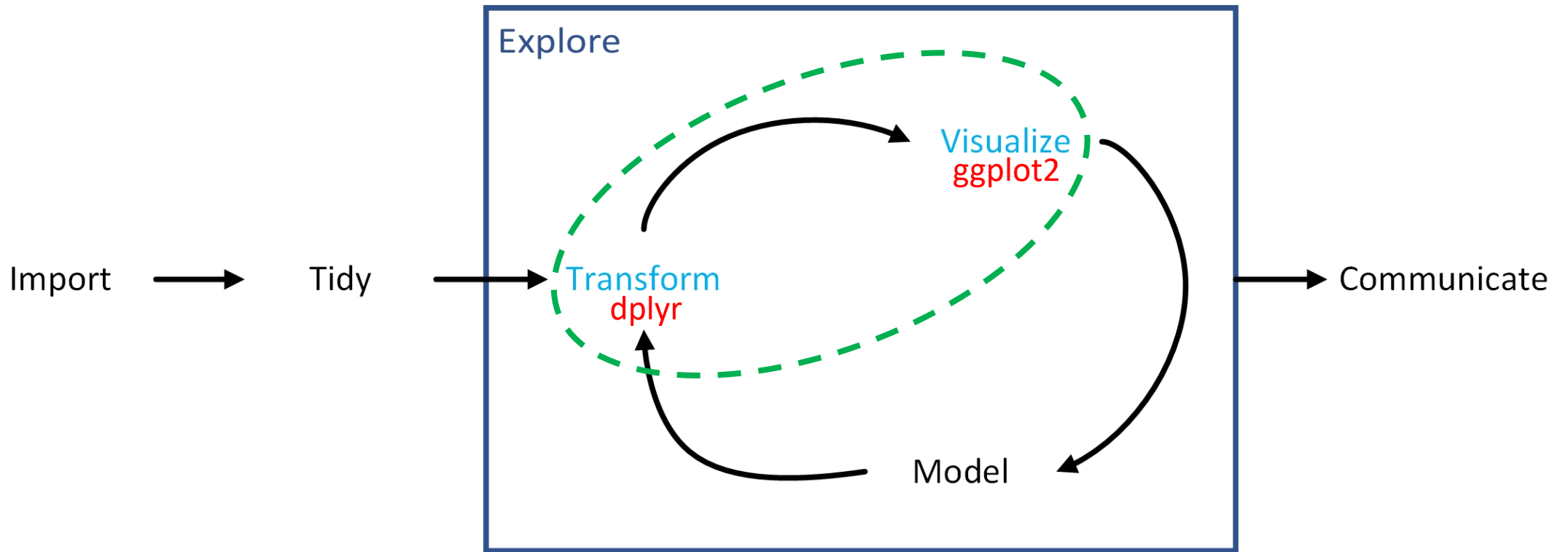# Data Transformation with dplyr & stringr + ggplot2
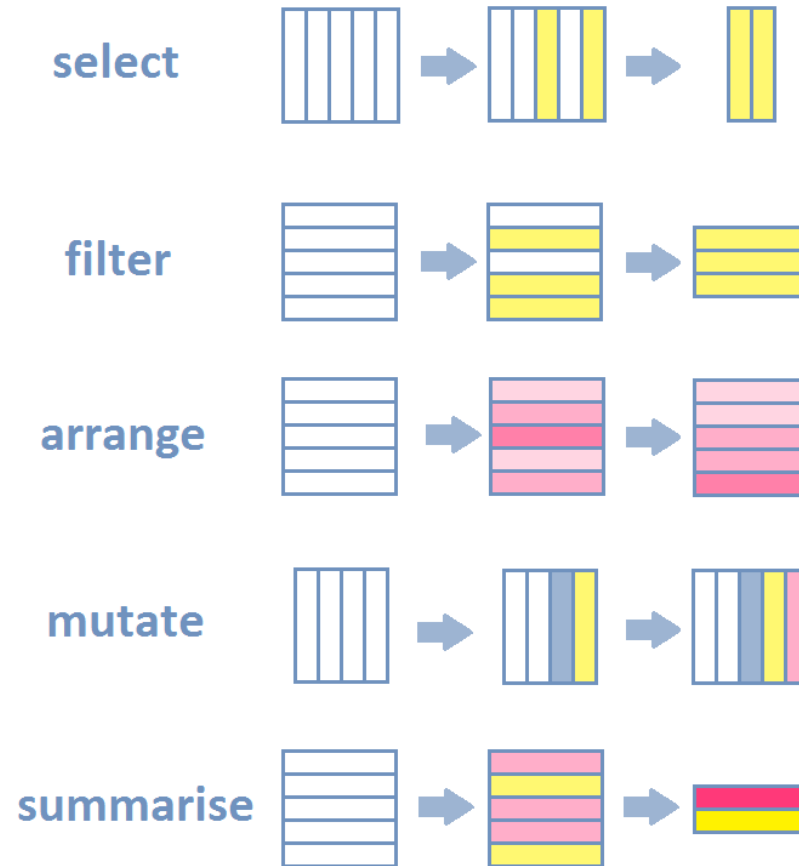
CMB 710 Tidybiology

Junqi Lu

ggplot2 Works with dplyr for Data Communication
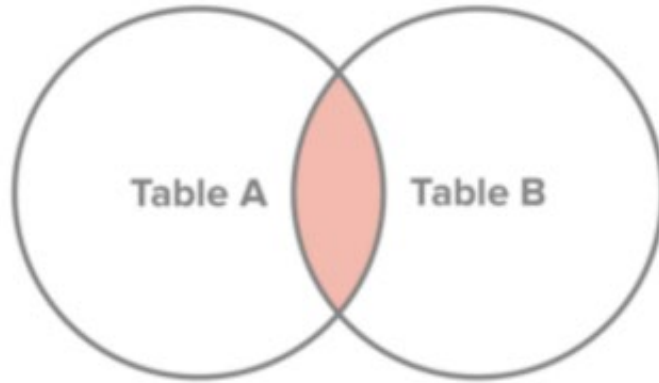
Quick Review of Basic dplyr
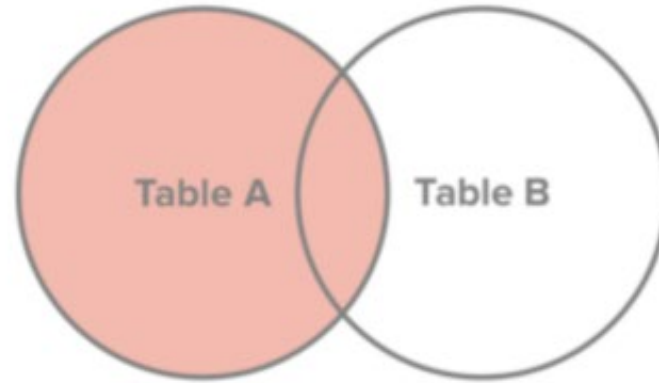

http://perso.ens-lyon.fr/lise.vaudor/dplyr/

Data Transformation Cheatsheet: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf
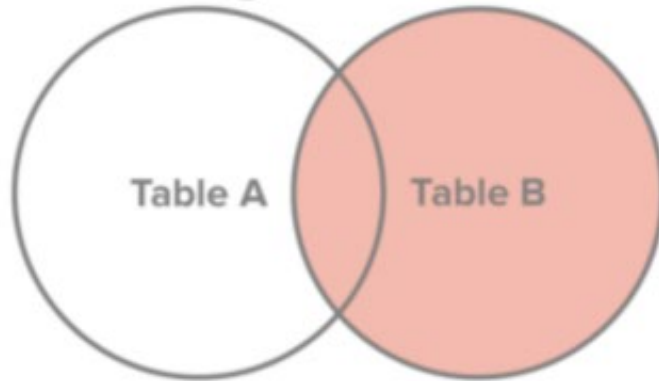
Graphical Review of 4 join()
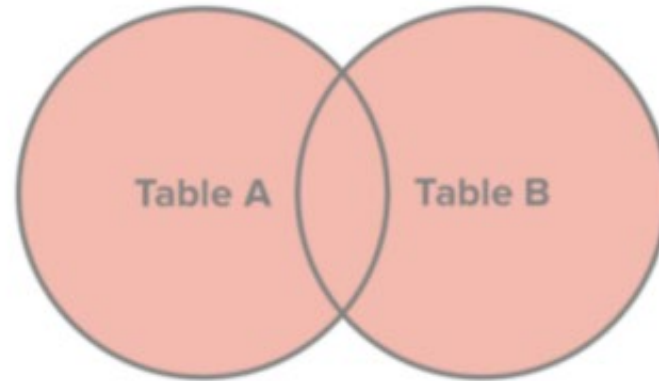
Dummy DataFrames for join() revisit
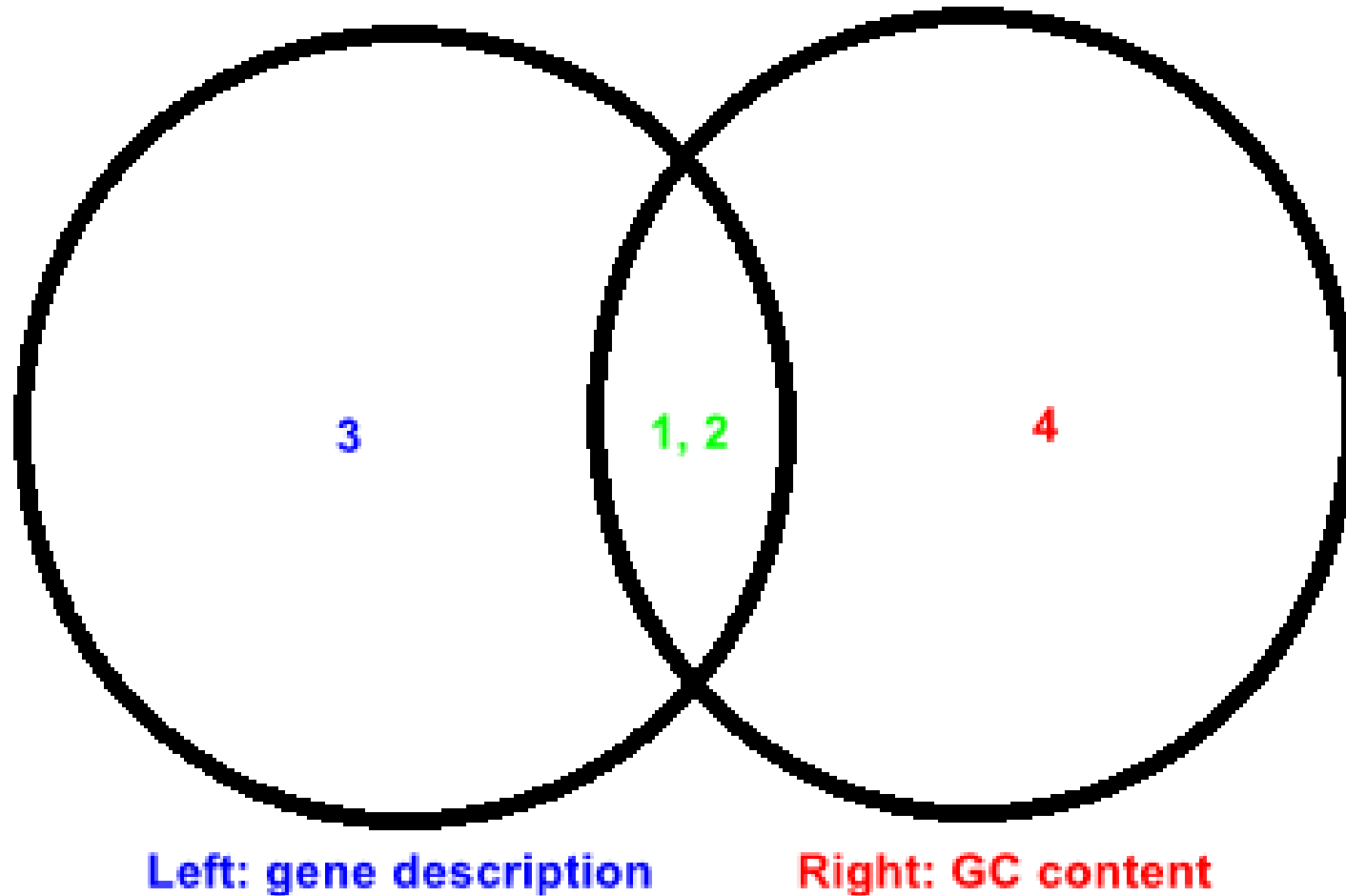
Graphical Review of 4 Mutating join()

Inner join

Outer join

https://r4ds.had.co.nz/relational-data.html#relational-data
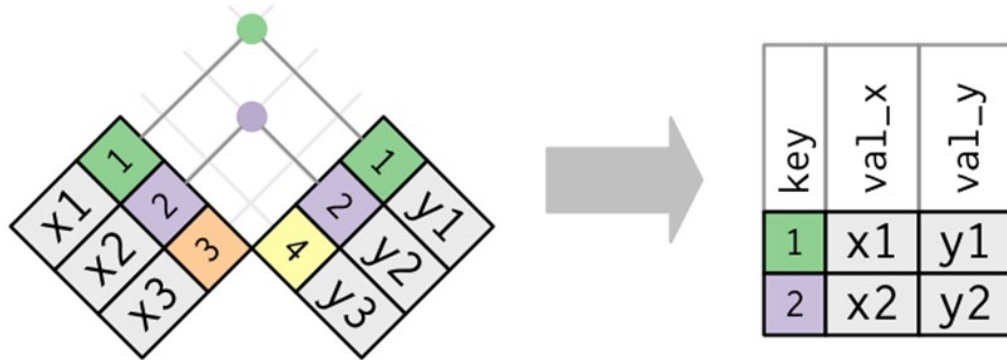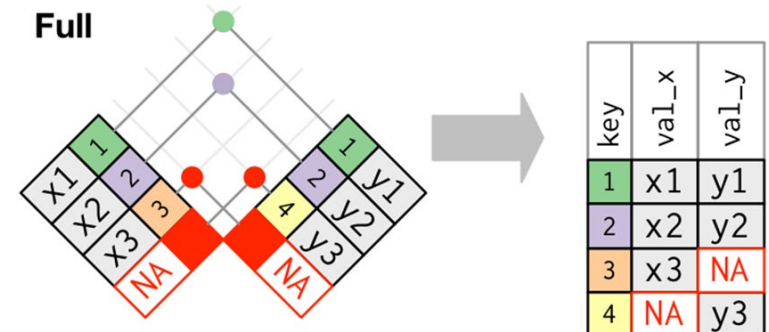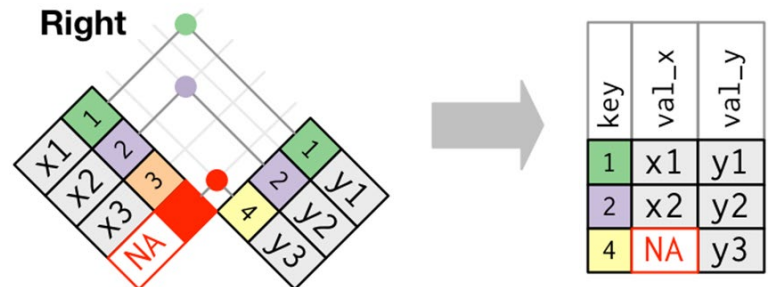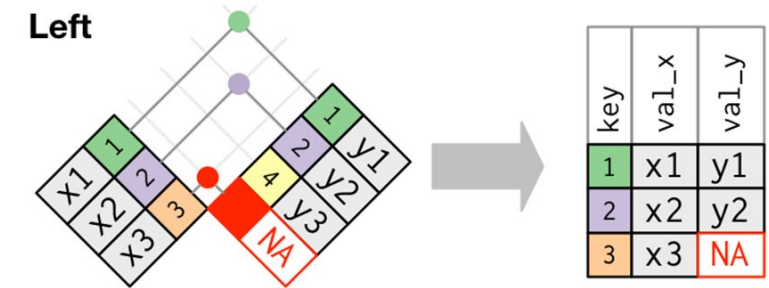
2 More Filtering join()

semi_join(x, y) keeps all observations in x that have a match in y

anti_join(x, y) drops all observations in x that have a match in y.

DataFrames binding/assembly

bind_rows() faster than rbind()

bind_cols() faster than cbind()

Data transformation on the run for ggplot2

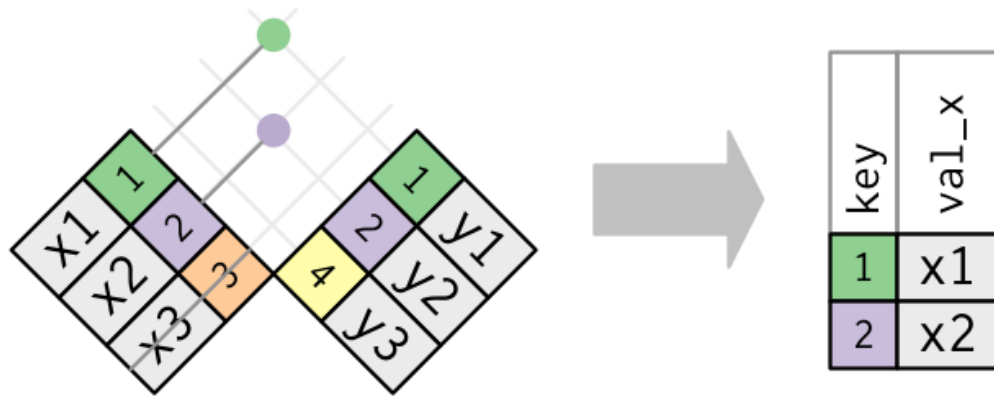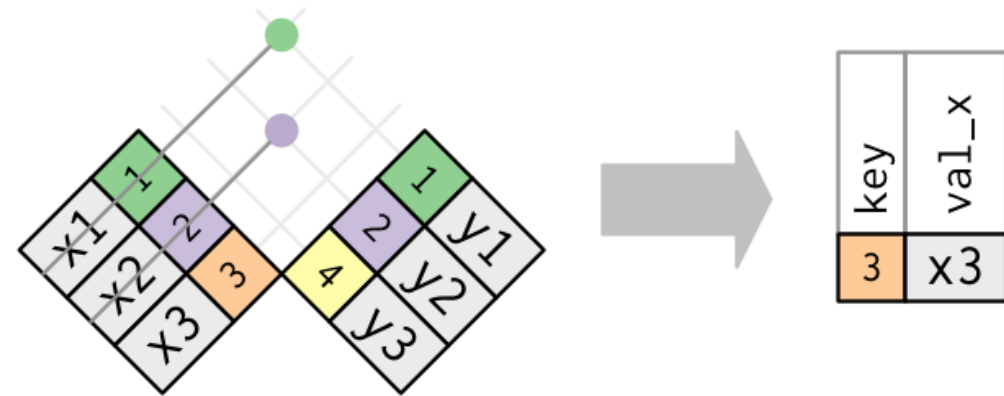- Data filtering by df subseting
  - Syntax: ggplot(df[JUDGEMENT(df$col_name),])+
- Conditional styling by ifelse()
  - Syntax: …styling_parameter=ifelse(JUDGEMENT(df$col_name), ValueIfTrue, ValueIfFalse)…
- If used in combination, need to tell ifelse() that df has changed
  - Method 1
    - NewDf <- df[JUDGEMENT(df$col_name),]
    - ggplot(NewDf)+
      - …styling_parameter=ifelse(JUDGEMENT(NewDf$col_name), ValueIfTrue, ValueIfFalse)…
  - Method 2
    - ggplot(df[JUDGEMENT(df$col_name),])+
    - …styling_parameter=ifelse(JUDGEMENT(df[JUDGEMENT(df$col_name),]$col_name), ValueIfTrue, ValueIfFalse)…

String processing: Stringr ≈ Shape Sorter

Pattern (regex)

str_view()
str_view_all()

String

String

String

Stringr pattern design

Meaning	Syntax	Example pattern	Example result
Or	$C_1|C_2$	str_view_all('abcdefg','bc|f')	abcdefg
One of the collection	$[C_1C_2...]$	str_view_all('abcdefg','[bdf]')	abcdefg
Anything but	$[^C_1C_2...]$	str_view_all('abcdefg','[^bdf]')	abcdefg
Range	$[C_1-C_2]$	str_view_all('abcdefg','[b-f]')	abcdefg
Start of	^C	str_view_all(c('abc','def'),'^a')	abc def
End of	C$	str_view_all(c('abc','def'),'f$')	abc def
0 or 1	C?	str_view_all('loooloolol','o?')	_loooloolol_
0 or more	C*	str_view_all('loooloolol','o*')	_loooloolol_
1 or more	C+	str_view_all('loooloolol','o+')	loooloolol
N or more	C{N,}	str_view_all('loooloolol','o{2,}')	loooloolol

Stringr pattern tester

- Patter explanation with diagram: https://extendsclass.com/regex-tester.html#python

Regular expression to test

/^lo{2,}|.[bd]e$/gm

Flags +

String to test   **Generate a string from RegEx (Beta)**

```
1  loooolool
2  lool
3  abcde
4  abbe
5  abce
6
```

Explanation

RegExp: /^lo{2,}|.[bd]e$/gm

LineBegin!   l   o   2 or more times

One of:

AnyCharExceptNewLine   bd   e   LineEnd!

- Pattern explanation with texts: https://regex101.com/

Realize Ideas to Graphs with dplyr & ggplot2

- Example: join and filter a dataset and make a graph

- Interpret the questions → cheat sheets / Stack Overflow / Google → ggplot2 explorer → finalize details

- Filtering a dataset on-the-fly in ggplot2 by subset()