The grammar of graphics is a framework for constructing statistical graphics in a principled way.

The grammar defines an explicit relationship you have between the variables (a quantity, quality, or property that you can measure) in your data and the graphic or plot you wish to represent. It was first defined by Lee Wilkinson and extended by Hadley Wickham in the R package `ggplot2` .

Throughout this step, you will uncover how the charts you know by name can be created via `ggplot2` , learn how to construct a `ggplot2` graphic and then build a plot up, layer-by-layer.

Why use the grammar of graphics?

The grammar of graphics allows you to **define** the mapping between variables in the data, with elements of the plot. It allows you to see and understand **how** plots are similar or different.

The grammar also helps you see how variations in the definition create variations in the plot. Using named plots, for example, a pie chart, bar chart, scatterplot, in some ways is like seeing animals in the zoo.

World Health Organisation (WHO) tuberculosis case notifications data

To showcase `ggplot2` you will build graphics using a data set provided by the World Health Organisation (WHO) (https://www.who.int/). This data is notifications of recent tuberculosis (TB) cases aggregated by **country**, **sex** and **age** group. We have tidied up this data for you and saved it in the R data storage format.

Before making your way through this step, we strongly recommend that you download the data set (https://github.com/datascienceprogram/ids_course_data/blob/master/tb_tidy.rds) then place it in a subfolder of your project folder on your computer. For example, if your project folder name is called **first_project**, then the TB data may have the directory: **first_project/data/tb_tidy.rds**.

About the data

The data consists of six columns:

- **country:** the country where the TB case(s) occurred
- **iso3:** a three letter standardised country code
- **year:** the year when the TB case(s) occurred
- **sex:** whether the TB case(s) belonged to males or females.
- **age_group:** whether the TB case(s) belonged to someone aged between 15 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 or 65 and above
- **count:** the total number of TB cases for a given country, year, sex and age group.

Load the data in RStudio on your computer

On your computer, open RStudio and then load the TB data in with the `read_rds()` function (contained in a package in the tidyverse)

```
library(tidyverse)
tb <- read_rds("data/tb_tidy.rds")
tb
```

```
## # A tibble: 47,866 x 6
##    country     iso3   year count sex   age_group
##    <chr>       <chr> <dbl> <dbl> <fct> <fct>
##  1 Afghanistan AFG    1997    10 M     15-24
##  2 Afghanistan AFG    1998   129 M     15-24
##  3 Afghanistan AFG    1999    55 M     15-24
##  4 Afghanistan AFG    2000   228 M     15-24
##  5 Afghanistan AFG    2001   379 M     15-24
##  6 Afghanistan AFG    2002   476 M     15-24
##  7 Afghanistan AFG    2003   511 M     15-24
##  8 Afghanistan AFG    2004   537 M     15-24
##  9 Afghanistan AFG    2005   606 M     15-24
## 10 Afghanistan AFG    2006   837 M     15-24
## # ... with 47,856 more rows
```

For the examples throughout Week 2, you will use a subset of this data corresponding to TB cases in Australia.

```
tb_au <- filter(tb, country == "Australia")
tb_au
```

```
## # A tibble: 224 x 6
##    country   iso3   year count sex   age_group
##    <chr>     <chr> <dbl> <dbl> <fct> <fct>
##  1 Australia AUS    1997     8 M     15-24
##  2 Australia AUS    1998    11 M     15-24
##  3 Australia AUS    1999    13 M     15-24
##  4 Australia AUS    2000    16 M     15-24
##  5 Australia AUS    2001    23 M     15-24
##  6 Australia AUS    2002    15 M     15-24
##  7 Australia AUS    2003    14 M     15-24
##  8 Australia AUS    2004    18 M     15-24
##  9 Australia AUS    2005    32 M     15-24
## 10 Australia AUS    2006    33 M     15-24
## # ... with 214 more rows
```

Don’t forget to save your work!

When you’re ready, save your work as `tb_au` in your project folder - you’ll continue to build on this throughout the week.

That’s it for now, but over the next steps of the course you will create charts from the TB data to find out if there are differences between age groups and sex for TB cases in Australia.