AN INTRODUCTION TO R FOR POLICY ANALYSIS

COURSE SCHEDULE:

Webinar each Monday @ 10am (AEST) ~ 20 July to 17 August

- ~~Week 1 ~ An introduction to programming, the R language and Rstudio~~
- ~~Week 2 ~ Doing stuff with data - Importing, Exploring and Summarizing Data~~
- **Week 3 ~ Making pixels pretty (1) - Data Cleaning, Merging and Basic Visualization**
- Week 4 ~ Making pixels pretty (2) - Data Cleaning and Visualization
- Week 5 ~ Bringing together the R programming pipeline

WORKSHOP 3 OUTLINE:

- R in the Wild: Anya Cushnie Mills – M & E Advisor
- Why we viz – example
- Reviewing the basics
- Stacking, binding, merging and reshaping data
- Data viz in R – an applied introduction to ggplot2
- Policy Scenario : Unemployment and Poverty in the US
- Workshop evaluation

POLL: COMPLETED SWIRL EXERCISES?

- Subsetting Vectors
- Matrices and Data Frames
- Looking at Data
- Dates and Times

Why Use R?

Anya Cushnie Mills
PhD Student, Epidemiology and Public Health
Tampere University, Finland
Anya.Cushnie@gmail.com

- ✓Used R exclusively for 5 years
- ✓Open access – no license required
- ✓All numerical analysis possible
- ✓Huge online community
- ✓Never have to memorize a code
- ✓Writing code is just like writing a sentence
- ✓Several packages to simplify tasks

- **Assessing HIV program outcomes for Jamaica before and after "Treat All": A retrospective study using the national treatment services database.**

Anya V. Cushnie[§1], Ralf Reintjes[1, 2¶], Susanna Lehtinen-Jacks[1¶], J. Peter Figueroa[3¶]

- 1 Unit of Health Sciences, Faculty of Social Sciences, Tampere University, Tampere, Finland
- 2 Department of Health Sciences, Hamburg University of Applied Sciences, Hamburg, Germany:
- 3 Department of Community Health and Psychiatry, University of the West Indies, Mona, Jamaica

Analysis was done using *R*, version 3.5.3 and the finalfit package was used to generate regression results and plots.

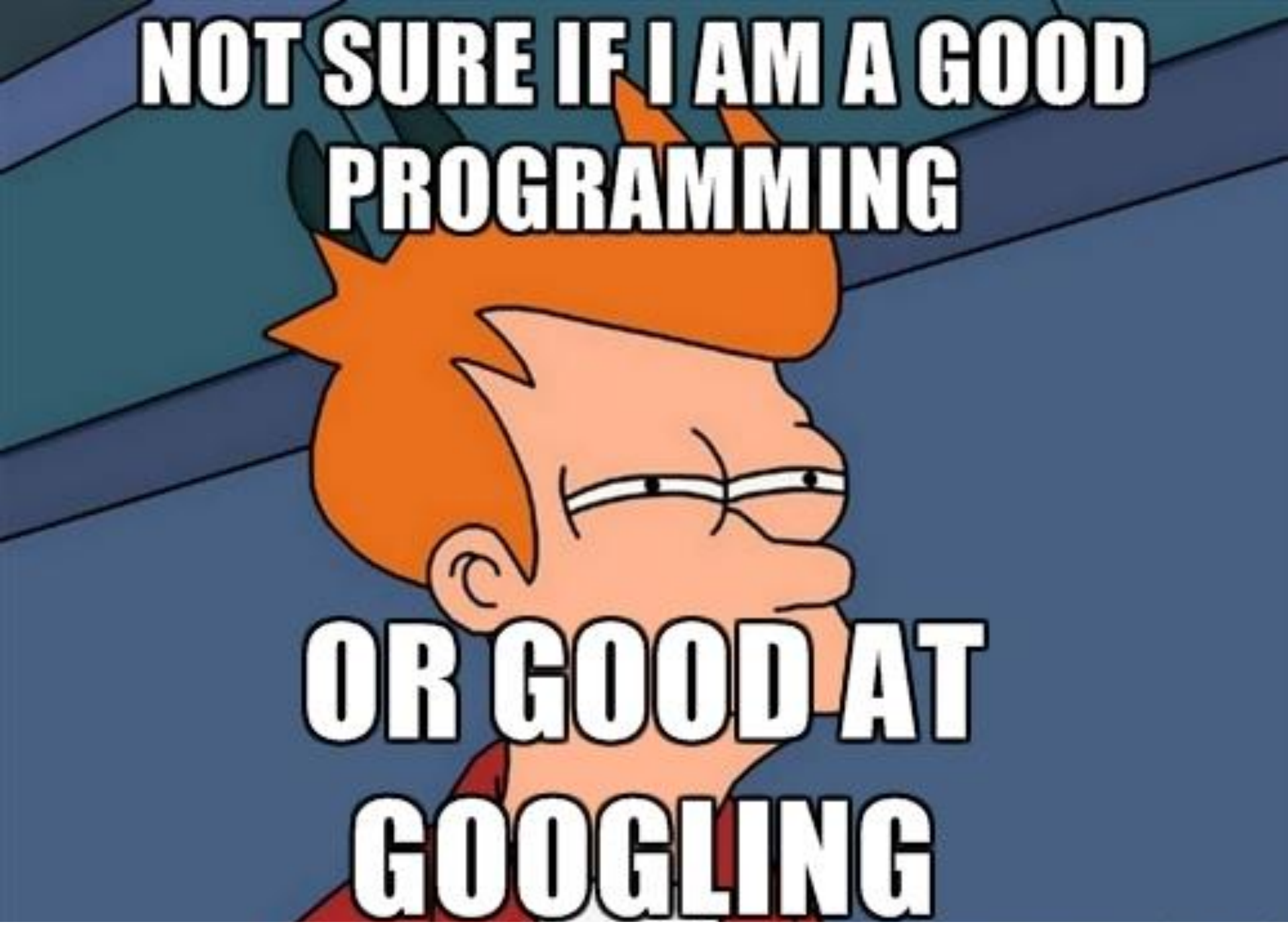
LEARNING R

During the Course:

- Complete the swirl courses
- Think through and attempt the scenario before
- Review the slides before each session
- Ask questions (either now or via slack)
- Tell me what worked/didn't in the evaluation

After the Course:

- Focus on learning what interests you about R
- Start using R alongside other software
- Commit to using R for a specific project



https://ardalis.com/working-through-roadblocks-a-guide-for-new-programmers

MOTIVATING EXAMPLE

EXAMPLE: SOME DATASET

Being a jerk, your colleague has gone on holiday and left you with his uncompleted project. The project requires understanding the relationship between x and y for different groups. However, it's not clear what the groups are as your colleague didn't label anything :/

- How can we start to understand the data?
- How can we tell if the groups are similar or not?

Open up the 'some analysis' script in the 'Some data' folder in week 3 and let's have a look.

WHY WE VIZ

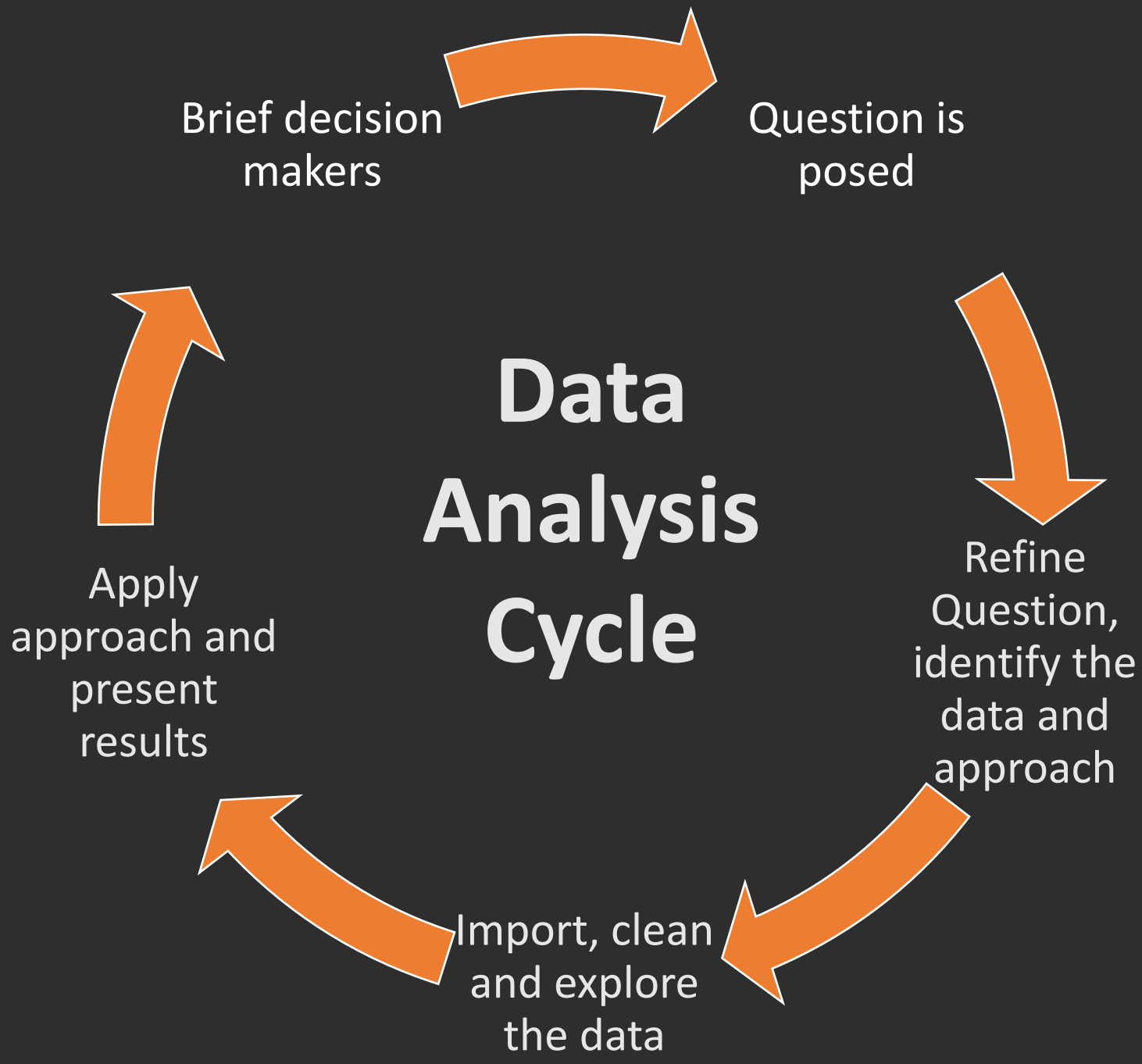
Summary statistics alone can hide complexity

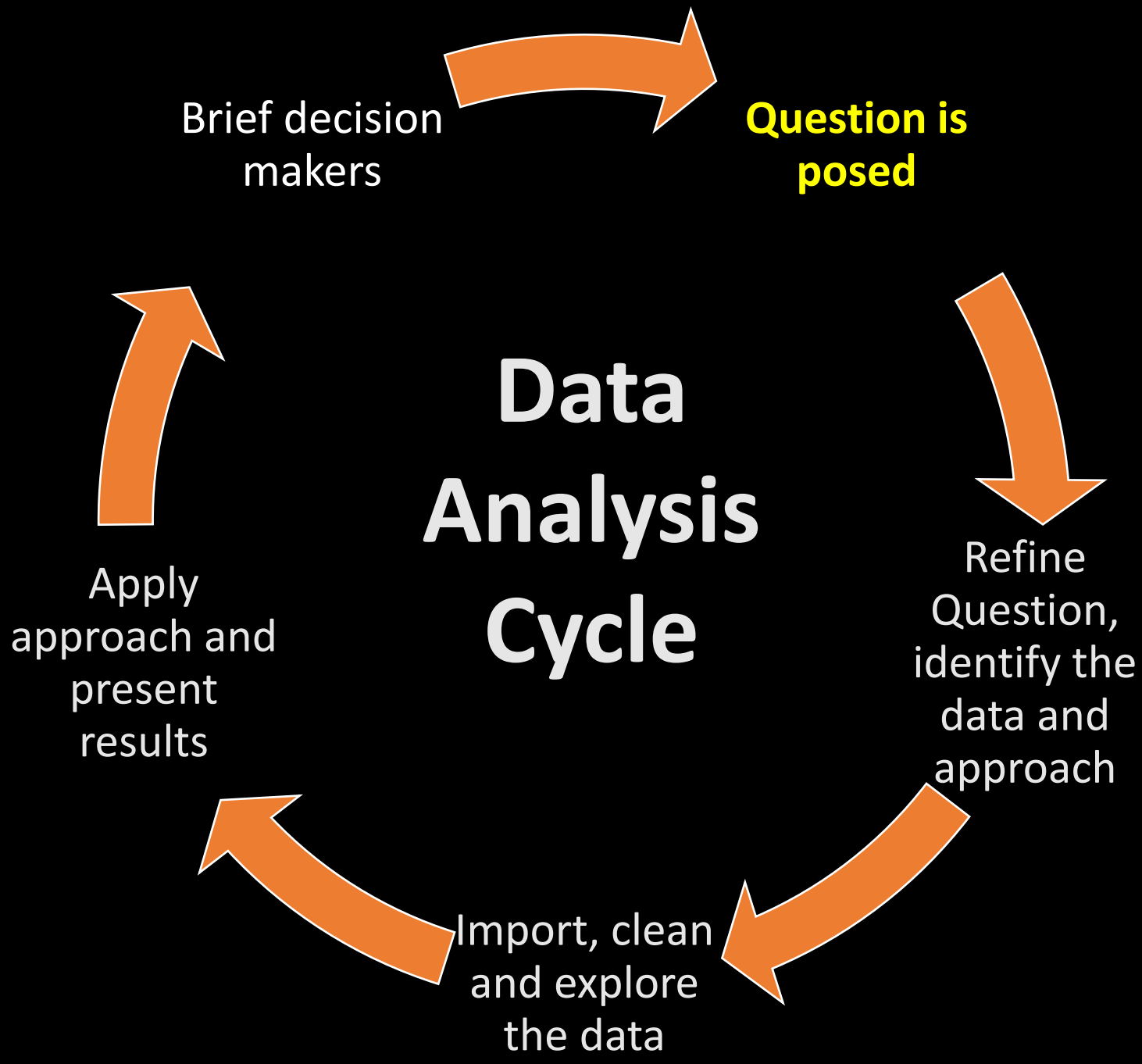
Good data viz also:

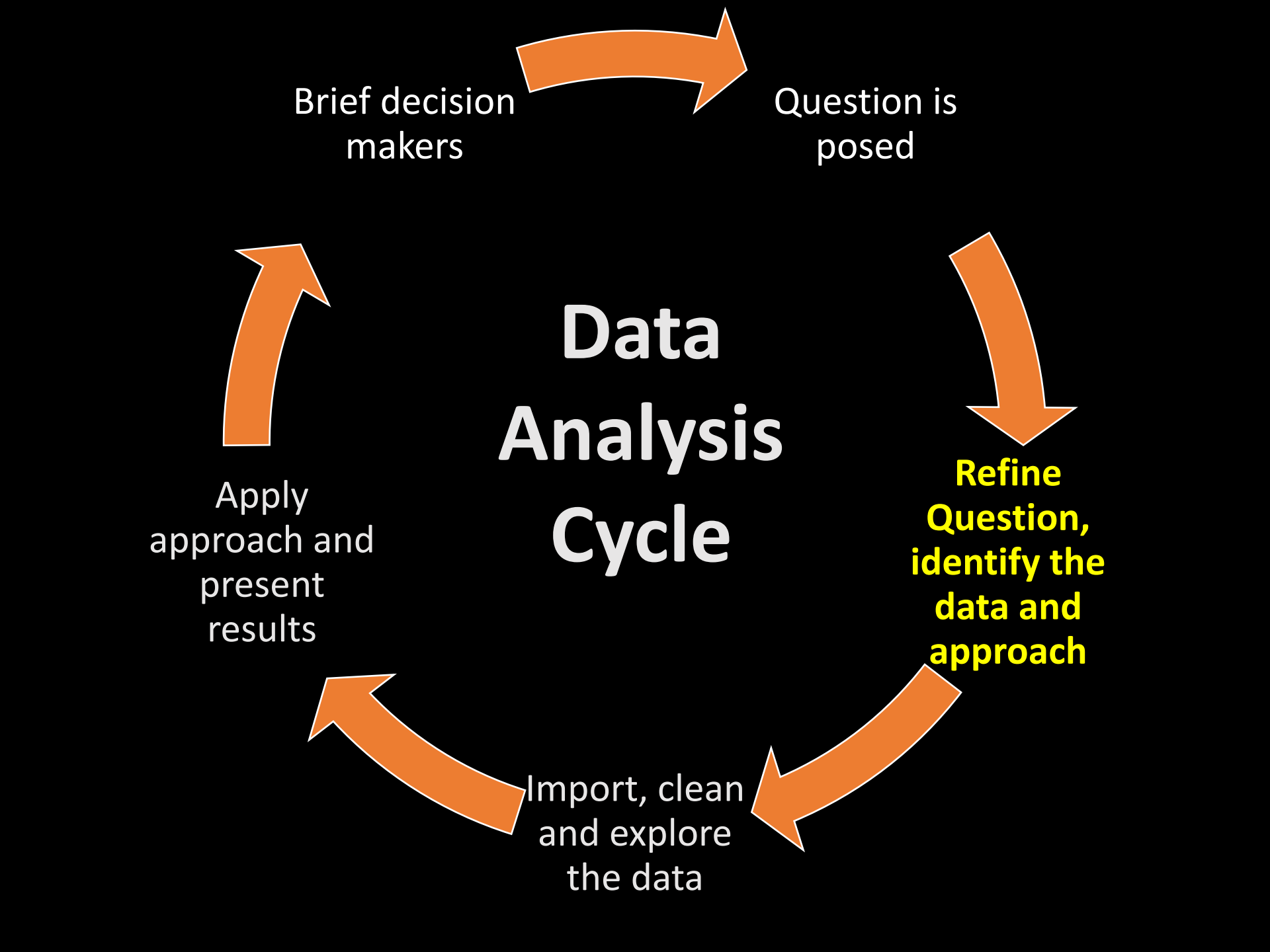
- Helps information be more quickly interpreted
- Highlights relationships, trends and patterns
- Allows us to focus on either the forest or the trees
- Helps us guide good policy by telling a 'data story'

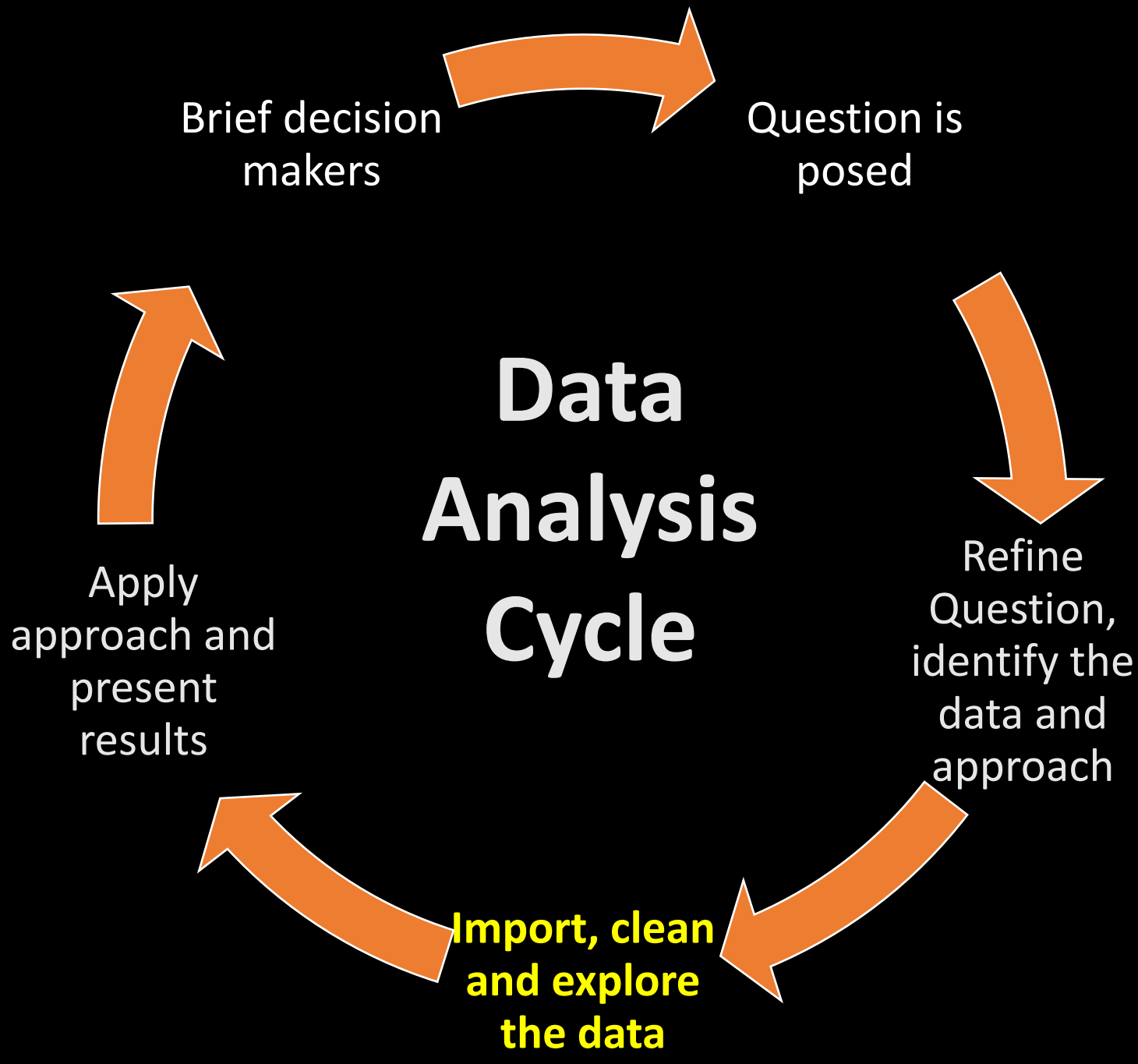
R has a reputation for being able to generate an unlimited range of high quality graphics

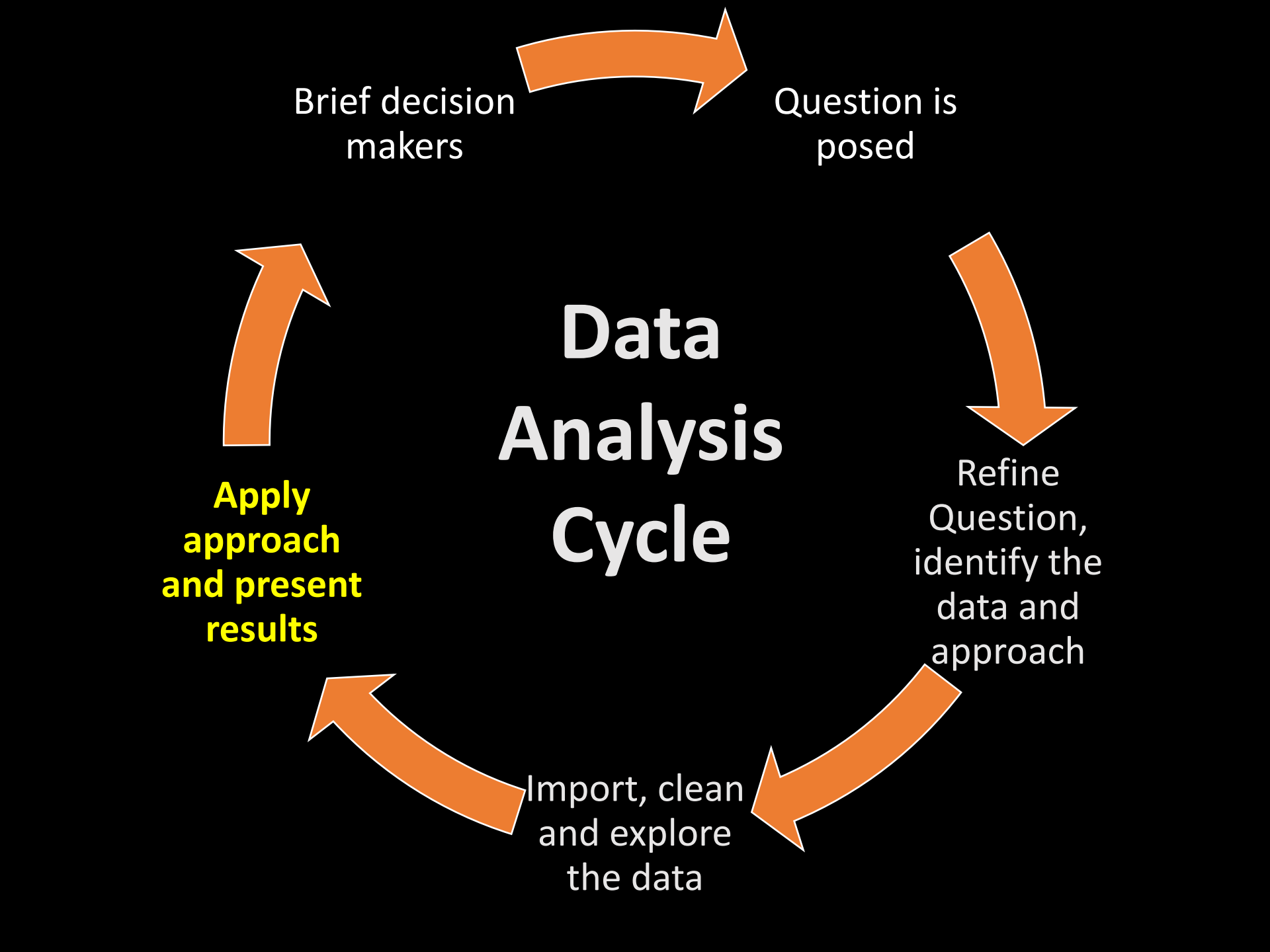
REFRESHER : THE BASICS

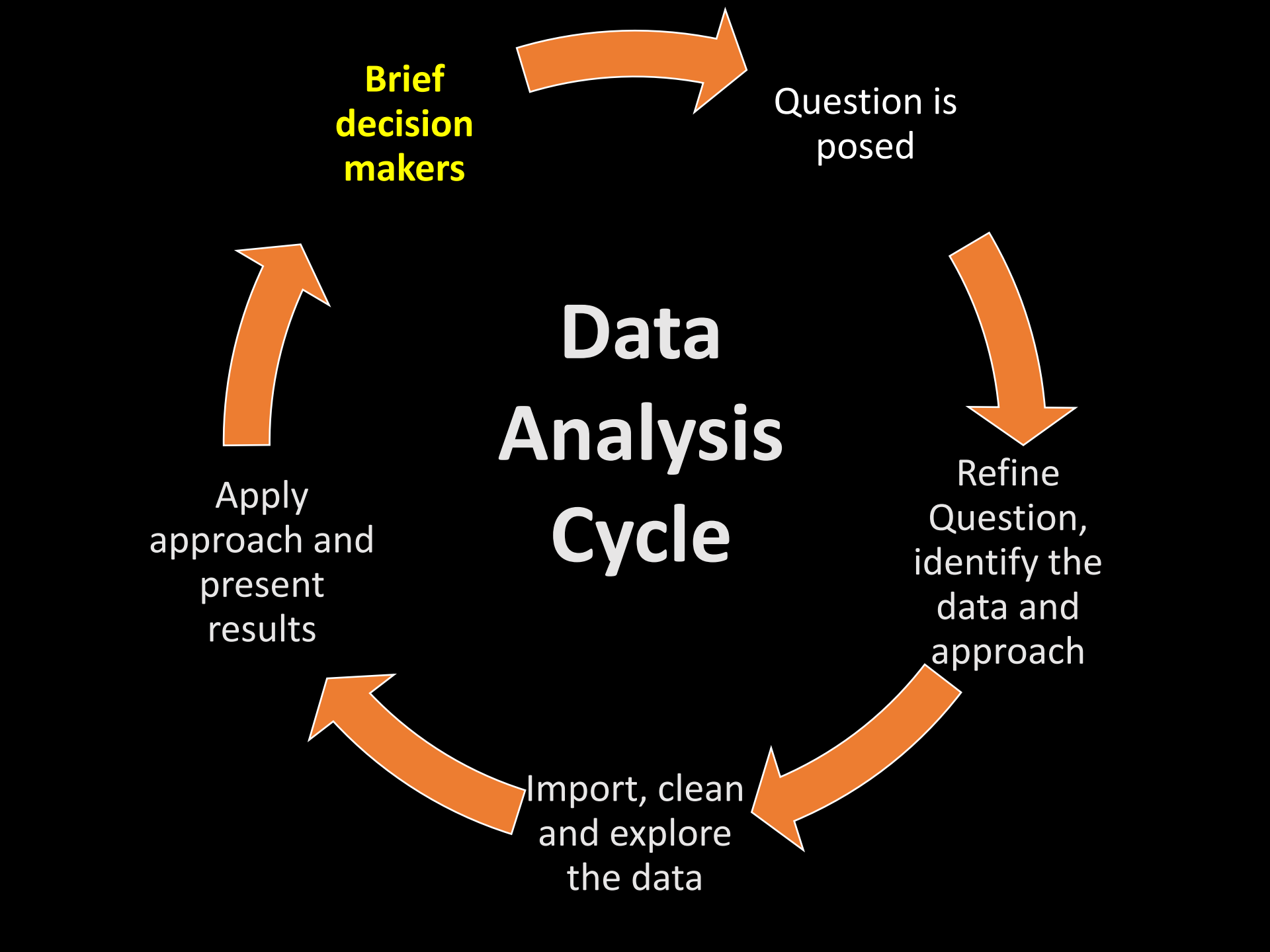

Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle

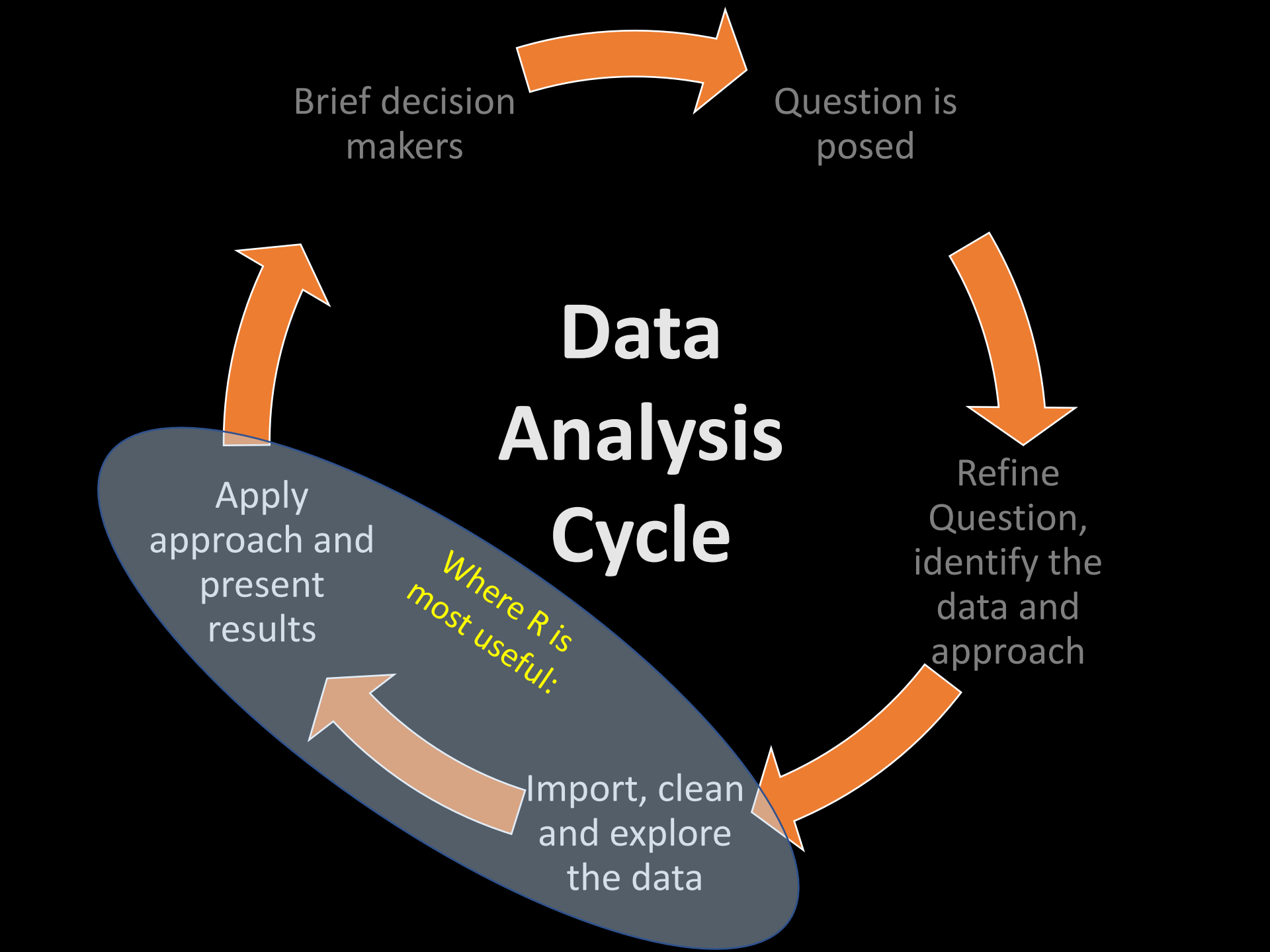

Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle


Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle


Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle


Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle


Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Apply approach and present results
Data Analysis Cycle


Brief decision makers
Question is posed
Refine Question, identify the data and approach
Import, clean and explore the data
Where R is most useful:
Apply approach and present results
Data Analysis Cycle

TYPICAL STEPS IN R ANALYSIS:

1. #**Set-up** - the working directory ----
2. #**Load** the necessary **packages** ----
3. #**Import**, **Explore + Clean** the data ----
4. #**Explain** - analyse, understand and visualize ----
5. #**Save -** the results ----

Reminder:

'#' tells R not to run the line

'----' code sections in RStudio

TESTING CONDITIONS WITH LOGIC

- **Less than:** <
- **Greater than:** >
- **Equal to:** == or **Not equal to** !=
- **And:** &
- **Or:** |
- **Not:** ! (eg, not equal !=, not greater than !> and not less than !<)

These are useful for extracting and summarizing data based on logical conditions. The result will be a logical vector:

For instance: *1:5* > *2* results in:

FALSE FALSE TRUE TRUE TRUE

SUBSETTING DATA + LOGIC

Vectors (use x[row])

33rd item: **vector_object[33]**

Data frames :

x[row,columns] OR x$column[rows]

Combine with logic to select data based on condition:

```
rand_numbers<-rnorm(1000)
#select numbers >1
rand_numbers[rand_numbers >1 ]
```

COMMON ERRORS

Unmatched parenthesis: ie not including a close bracket to tell R the arguments passed to a function have been supplied and to apply the function – will tell you with '+'

Misplaced commas: eg not separating arguments by commas when applying a function

Applying functions to an incompatible element or data object type – eg trying to multiply text by 2

"cannot open" – attempts to read something that doesn't exist or isn't accessible (eg because the name or directory is wrong or R can't access it due to it being locked/open in another program).

"could not find function" – not loading the library needed for a function or misspelling the name of the function (eg Plot instead of plot).

"Error in eval" – cause by references to something that doesn't exist (eg misspelling an object name)

"no applicable method" – trying to apply a function to an object/element type that doesn't support it

"**subscript out of bounds"** – telling R to access data that doesn't exist/ is out of bounds (eg asking for row 34 in a 20 row vector)

When data/elements are the wrong type they can be converted using:

as.data.frame(), as_tibble(), as_vector(), as_character(), as_numeric(), etc

DPLYR VERBS

Provides a grammar for basic data manipulation:

- **select** – pick the columns you want
- **filter** – choose the relevant observations
- **mutate** – create new variables
- **group_by** – define groups you're interested in for summarise
- **summarise** – create summary statistics
- **arrange** – sort your data

enter '?summarise' to see help and a list of summary statistics

TIDY DATA

TIDY DATA:

Vehicle	Efficiency	Power	Year
Mazda RX4	21	110	2001
Mazda RX4 Wag	21	110	2015
Merc 230	22.8	95	2015
Merc 280	19.2	123	1983
Datsun 710	22.8	93	2017

- ✓ Each variable is a column.
- ✓ Each observation a row.
- ✓ Each experiment/survey a separate table

TIDY DATA:

Because R operations are done by column, repetitive tasks can be greatly simplified if we get everything to a single column first

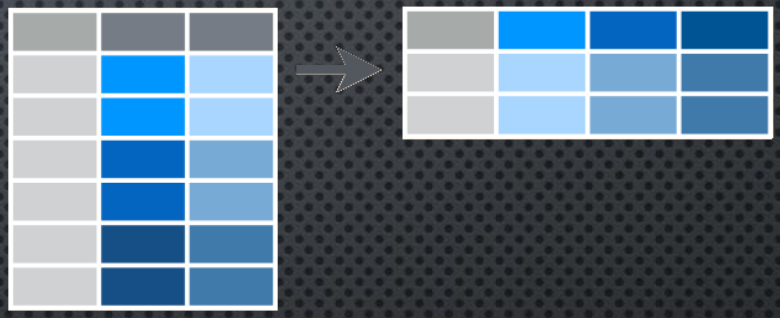
Eg: If columns specify both year and measurement we can:

1. Make the data long;
2. Split the time and measurement into two columns; then
3. Widen the data based on the variable type.

‘TIDYR’ CAN HELP YOU GET DATA IN THE RIGHT SHAPE

pivot_longer()

pivot_wider()



separate()

unite()

(This can also be useful for making data useful for pivot tables in excel)

MAKING DATA LONGER

pivot_longer(data, cols)

cols = columns to reshape/pivot from wide to long (eg a variable/measurement type).

eg: to aid time series analysis when years stored by column

MAKING DATA WIDER

pivot_wider(data, id_cols = NULL, names_from ='name', values_from = 'value')

id_cols = columns that uniquely identifies observations

names_from = the name of the column(s) to make wide

values_from = the column where values are stored

eg to make variables 'long' so they're easier to access

SPLITTING COLUMNS

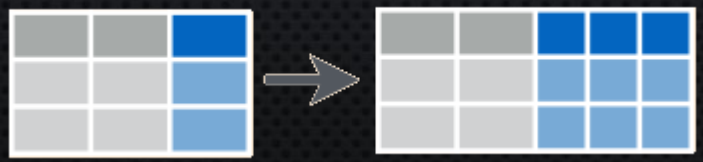
separate(data,col,into,sep)

col: column where data should be split

sep: character where data should be split

into: names of columns pieces to be split into

eg: to split first and last names into separate columns



MERGING MULTIPLE COLUMNS

unite(data, col, ..., sep="_")

col = name of the new column with merged data

... = columns to merge into one

sep = separator to use between values (defaults to "_")

eg: create a date from separate day/month/year columns

STACKING, BINDING AND JOINING DATA

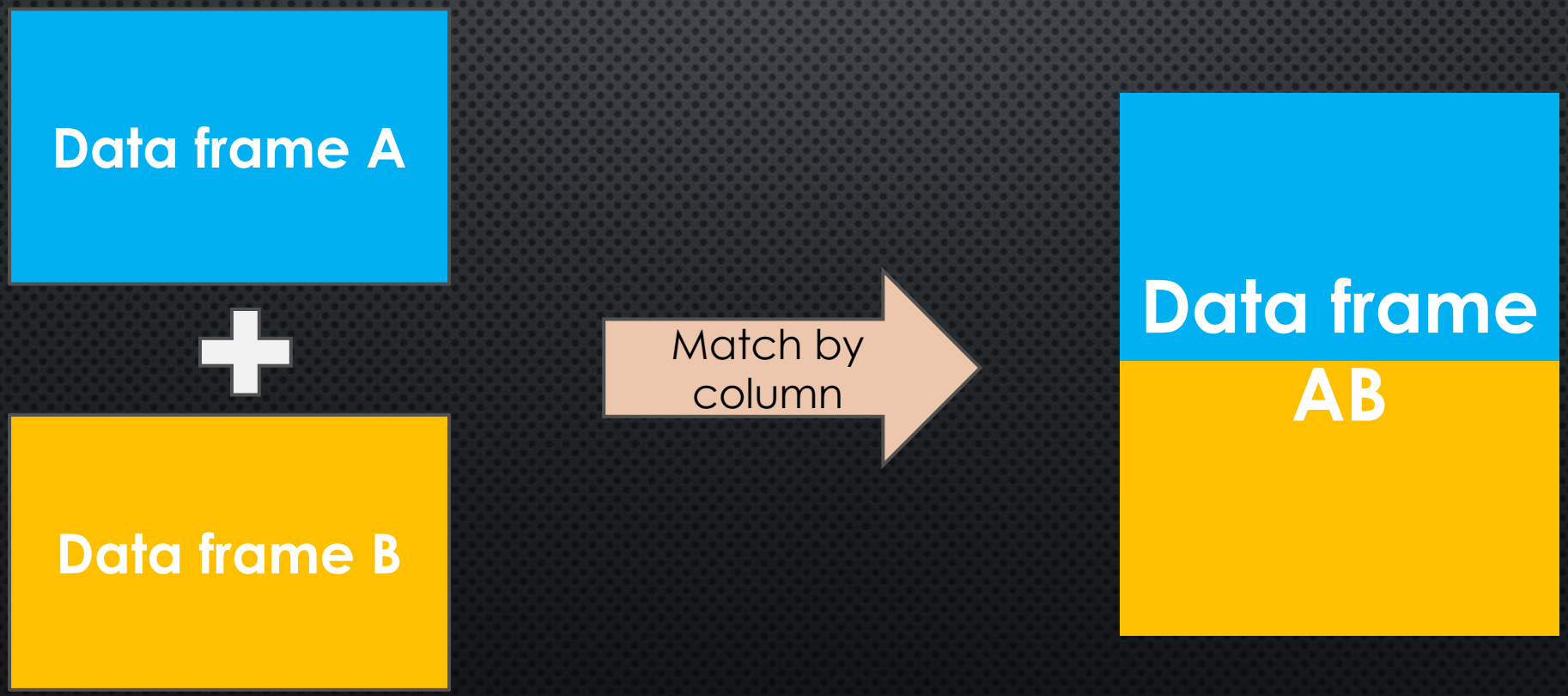
STACKING, BINDING AND JOINING DATA

- **Stack** - data using columns
- **Bind**– data based on rows
- **Join** – data using a common key

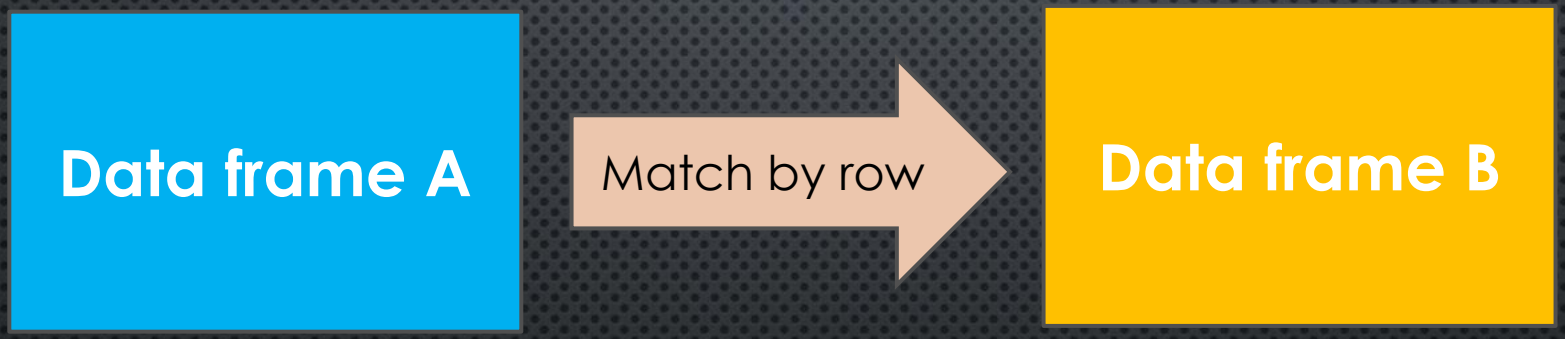
Data frame A

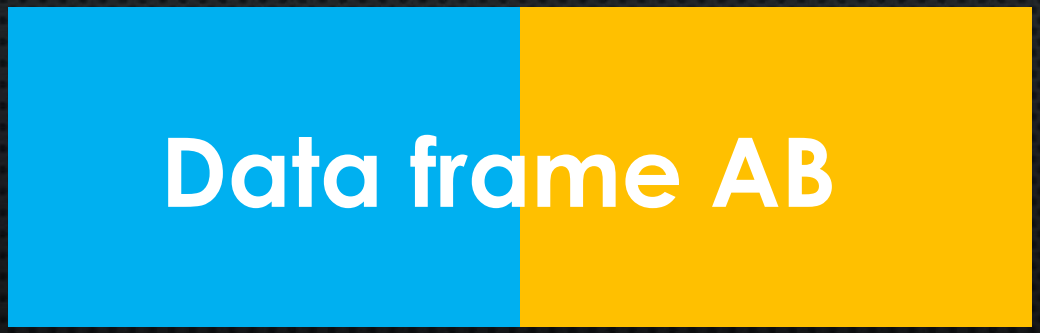
Data frame B

STACKING - ADDING NEW ROWS/OBSERVATIONS

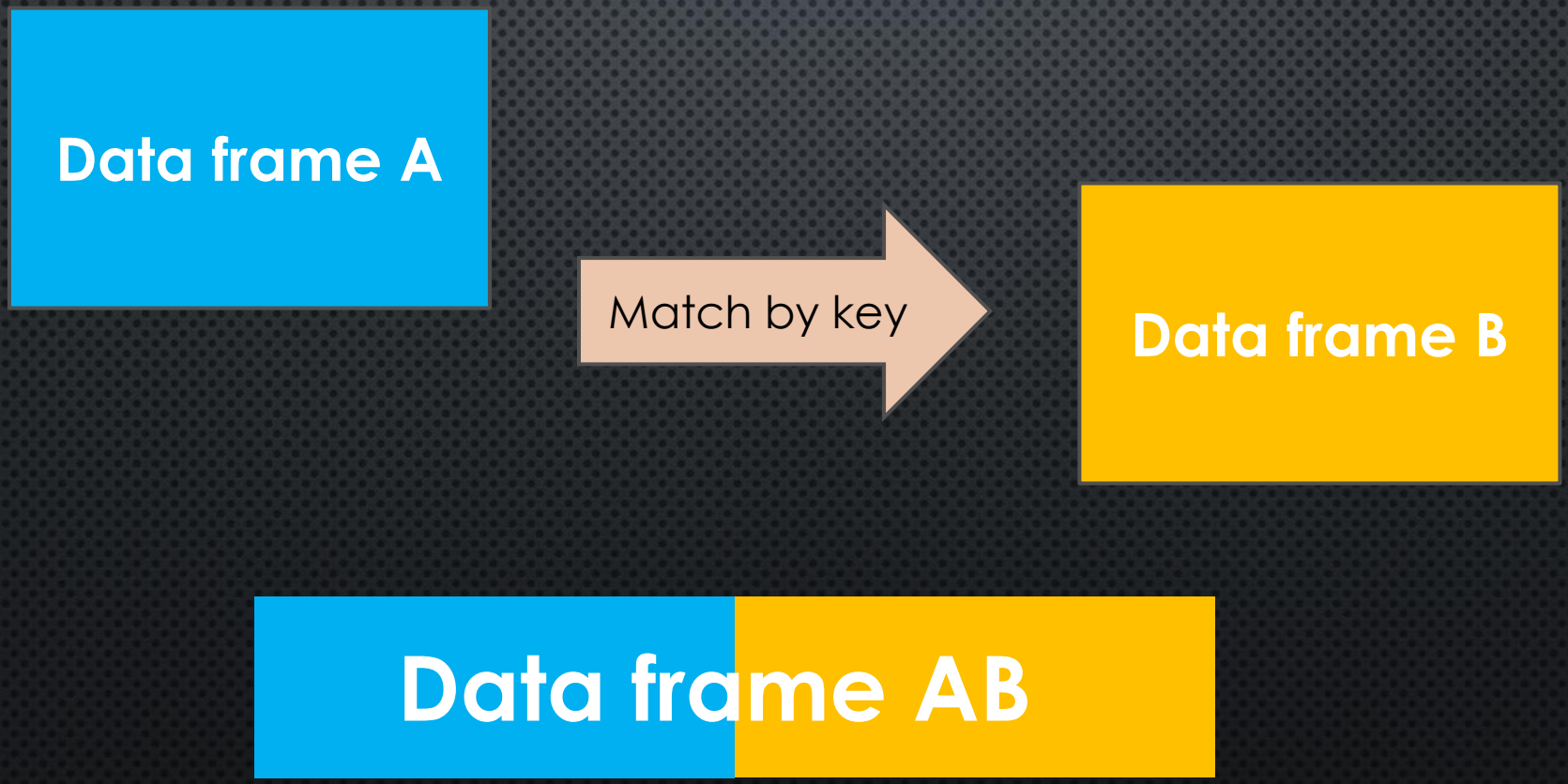


BINDING – ADDING NEW COLUMNS/VARIABLES





JOINING – JOING USING A COMMON KEY



STACKING, BINDING AND JOINING DATA

Base:

- rbind(),
- cbind()
- merge()

- **Dplyr:**
  - bind_rows()
  - bind_cols()
  - left_join(), right_join() etc

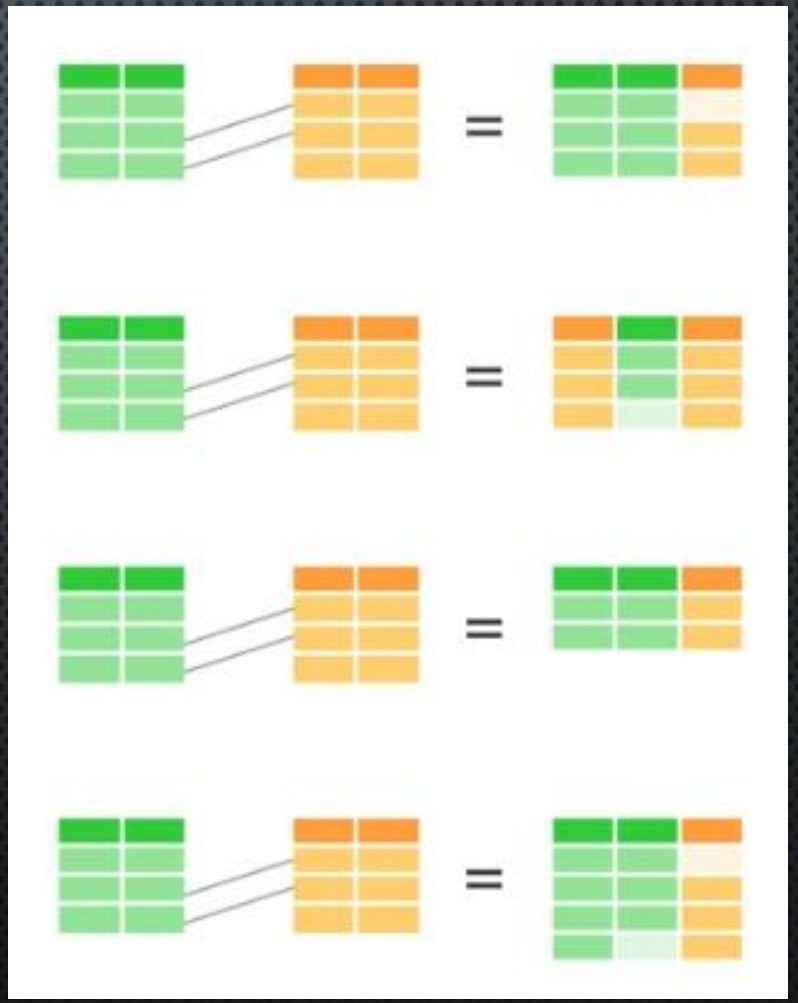
DPLYR JOINS:

left_join(x,y)

right_join(x,y)

inner_join(x,y)

full_join(x,y)



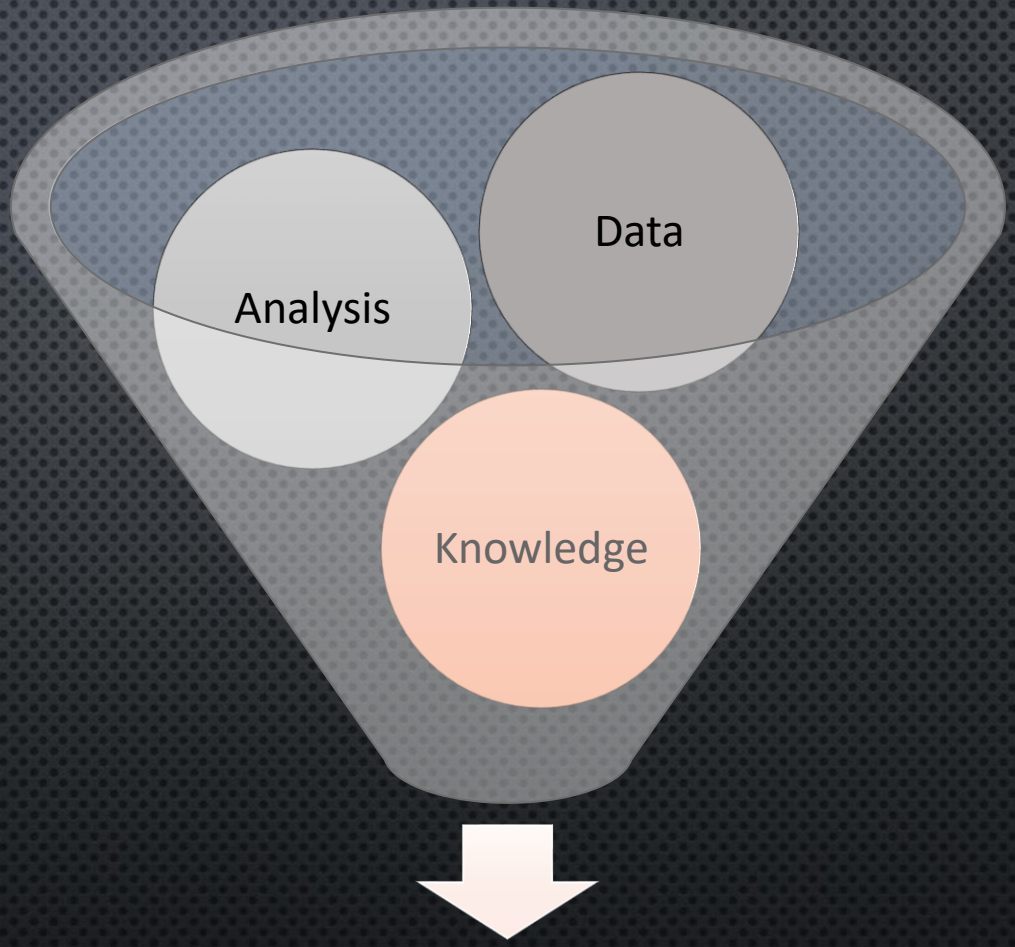
See: https://rpubs.com/williamsurles/293454

DATA VIZ IN R

DATA VISUALISATION

- Graphs are about making sense of data
- People are better at interpreting visual information than numbers
- Graphs are also useful for you to explore data quickly – seeing trends etc
- A good graph can help you make an argument more convincing and interesting (particularly if part of a ‘data story’)

EXPLORATORY VS. EXPLANATORY ANALYSIS



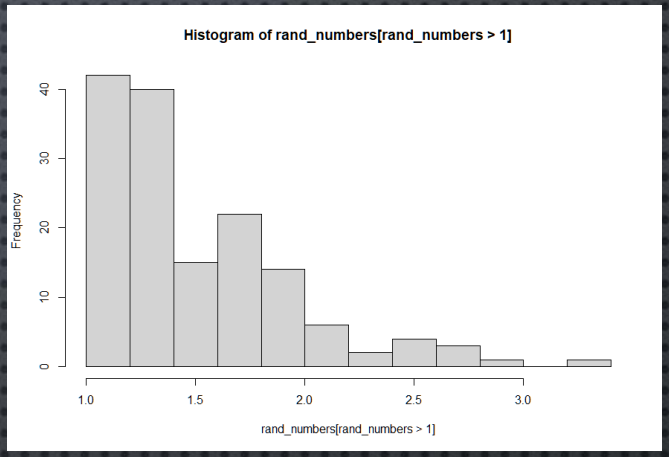
Exploring →

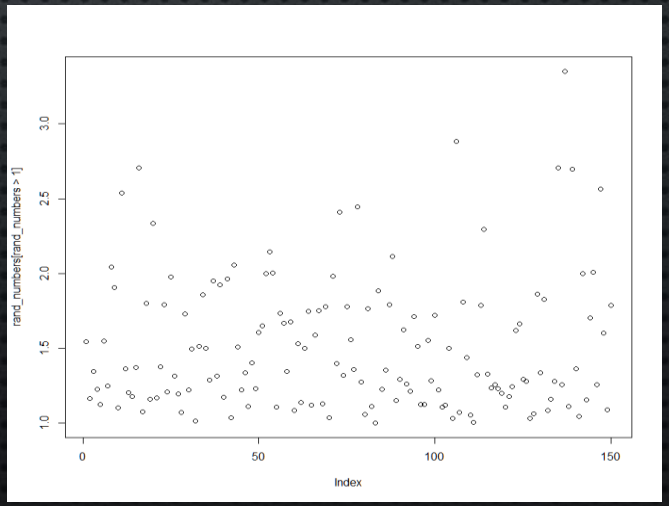
Explaining → Insights/Results

BASE PLOTS

- Are built into R
- Are fast and useful
- Don't have to be ugly
- Uses commands, rather than using a 'grammar'
  - Eg plot(), barplot(), hist(), plot.ts() etc

(Great for **exploratory analysis**)

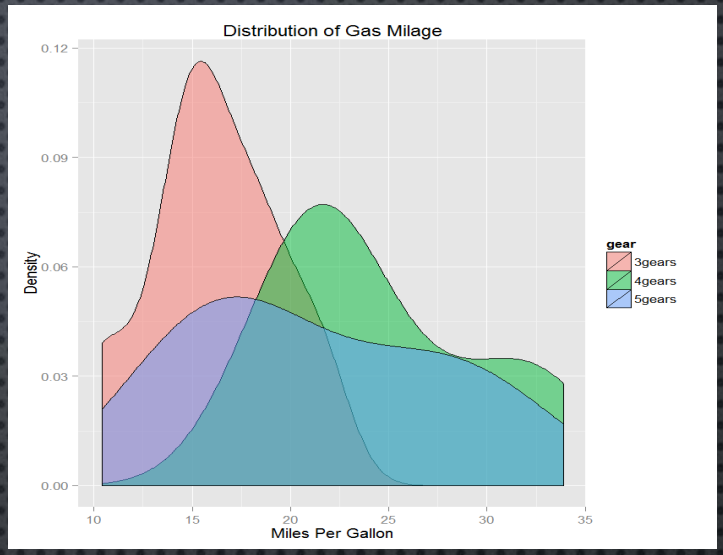


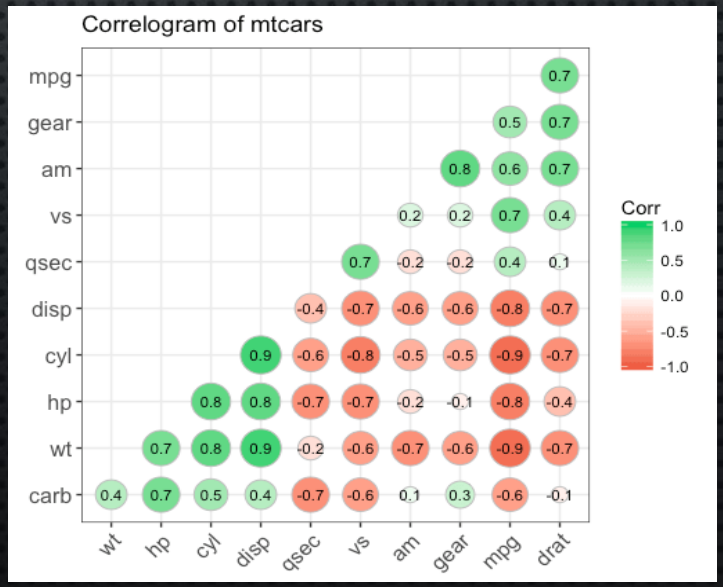


GGPLOT:

- Requires the ggplot2 package
- Can be slower
- Tries to be prettier by default
- Has a consistent logic and 'grammar'

(Great for **explanatory analysis**)





GGPLOT2 – KEY GRAMMAR

GGLOT2 is based on the 'grammar of graphics', the idea that every graph is comprised of the same basic components

- Data;
- A geometry: **'geom'** and coordinate system.
- Aesthetics: **'aes()'**.

(These concepts are outlined in the ggplot2 cheat sheet)

GGPLOT2 – KEY GRAMMAR

Geoms	Defining the geometry of the plot
Aesthetics	Defining the aesthetic mappings
Labels	Defines labelling
Themes	Changes how plots look
Facets	Dividing plots based on groups

GGPLOT2 – LAYERS ON A CANVAS

Define the canvas:

```
ggplot(data=titanic_data, aes(x=age)
```

Specify a geometry:

```
ggplot(data=titanic_data, aes(x=age) +
    geom_histogram()
```

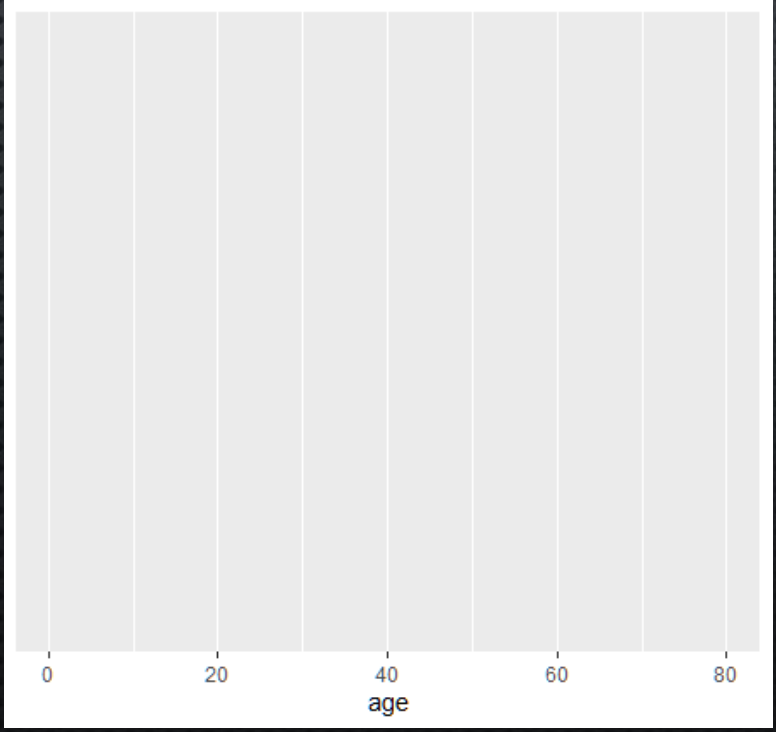
Add layers:

```
ggplot(data=titanic_data, aes(x=age) +
    geom_histogram() +
    labs(y="A Histogram")
```

GGPLOT2 – LAYERS ON A CANVAS

Specify data and aesthetics:

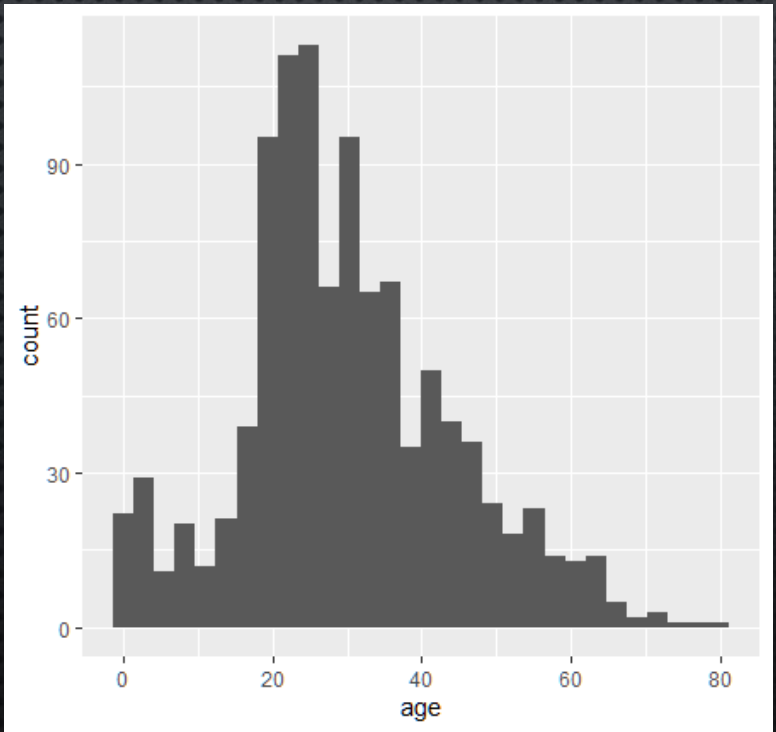
ggplot(data=titanic_data, aes(x=age))



GGPLOT2 – LAYERS ON A CANVAS

Add the geometry:

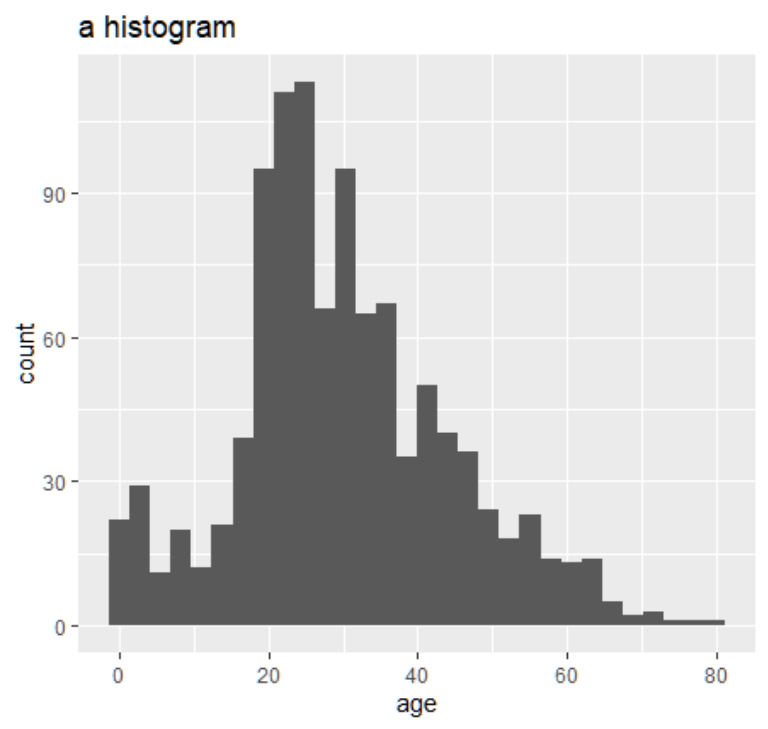
ggplot(data=titanic_data, aes(x=age)) +
geom_histogram()



GGPLOT2 – LAYERS ON A CANVAS

Add labels:

ggplot(data=titanic_data, aes(x=age)) + geom_histogram() + labs(title="a histogram")



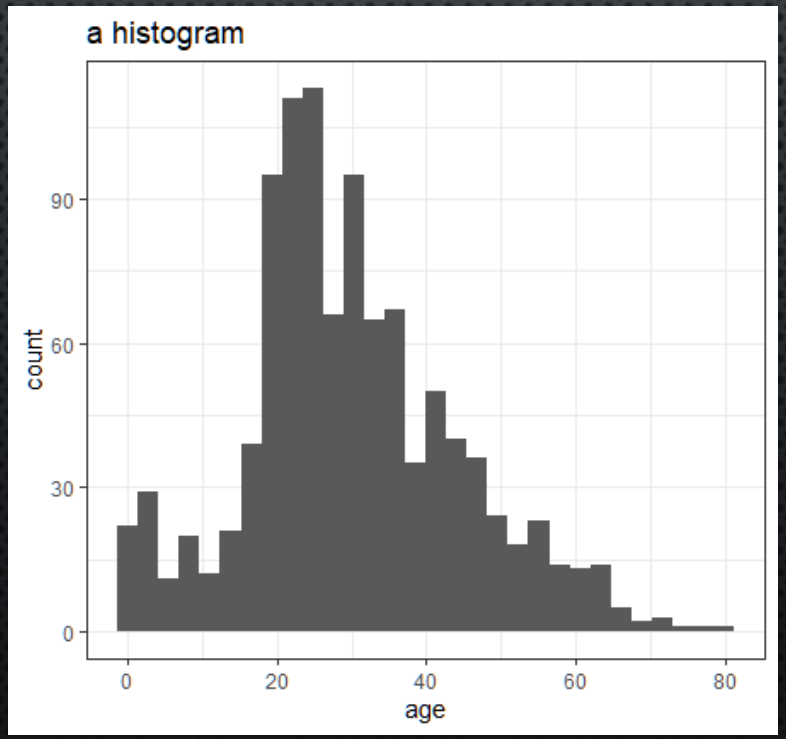
GGPLOT2 – LAYERS ON A CANVAS

Give it a theme:

ggplot(data=titanic_data, aes(x=age)) + geom_histogram() +

labs(title="a histogram") +

theme_bw()



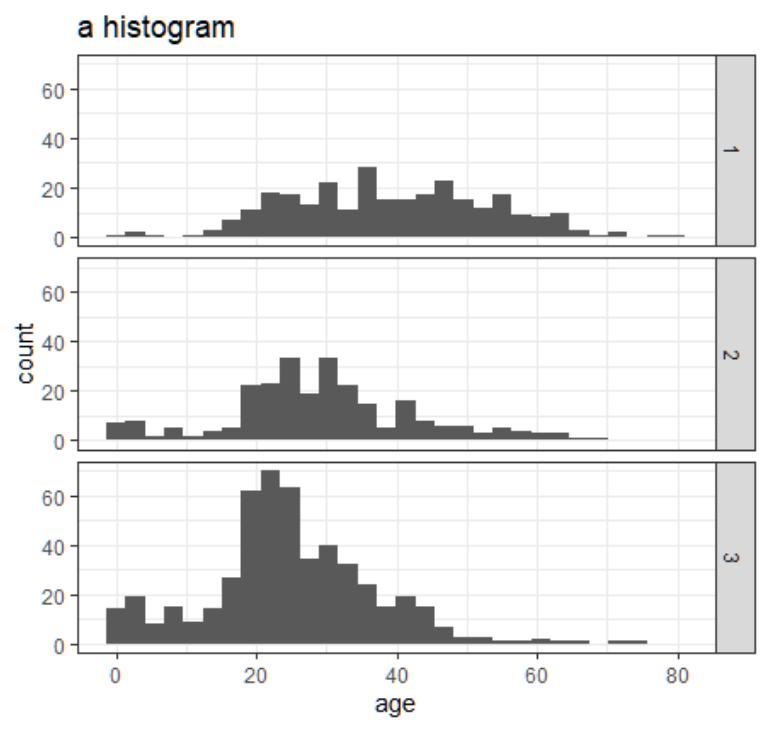
GGPLOT2 – LAYERS ON A CANVAS

Split by facets:

ggplot(data=titanic_data, aes(x=age)) + geom_histogram() +

labs(title="a histogram") + theme_bw() +

facets_grid(



TITANIC ANALYSIS

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

- **Make corrections to the dataset, by:**
  - Changing the boat of passenger number 120 to 122 to "Five" and changing boats listed as "C D" to "C".
  - Chaning the labels for surviving passengers to "Survived" and "Didn't Survived" to make it easier to interpret
- **Plot the most common titles**
  - What if we separate by class?
- **Plot the average fare paid by title**
  - Did this differ by class?
- **Plot the average fair by passenger title, class and sex**
- **Combine the new boat data to see how boat capacity relates to passenger survival**
  - Were people in higher % occupancy boats less likely to survive?

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data Cleaning:

- **Change the boat of passenger number 120 to 122 to "Five" and changing boats listed as "C D" to "C".**
  - titanic_data$boat<-str_replace(titanic_data$boat, "C D","C")
  - titanic_data$boat[120:122]<-"Five"

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data Cleaning:

- **Relabel the survival column to "Survived" and "Didn't Survived"**

```
factor(titanic_data$survived,
levels=c(0,1),
labels=c("Didn't Survive","Survived")))
```

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data Cleaning:

- **Get the title of each passenger**

```
separate(data=titanic_data,
                col=name,
                sep=", | [.]",
                into=c("surname","title","first_last_and_pref_name"),
                remove=FALSE)
```

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data Cleaning:

- **Combine the old data with new data in "titanic boat data.csv"**
  - **Import:** read_csv()
  - **Merge:** left_join(titanic_data,titanic_boat_data)
- **How does boat capacity relates to passenger survival – Were people in higher occupancy boats less likely to survive?**
  - Calculate a measure of boat occupancy as $ of capacity:
    - occupancy_pct = boat_occupancy/ boat_capacity

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data visualization:

- Plot the most common titles
  - What if we separate by class?
- Plot the average fare paid by title
  - Did this differ by class?
- Plot the average fair by passenger title, class and sex
- Were people in higher occupancy boats less likely to survive?

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data visualization:

- **Plot the most common titles**
  - **Two variables (title vs. frequency):** geom _bar()
  - **What if we separate by class?**
    - **Split by class:** facet_grid(class ~ .)

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data visualization:

- **Plot the average fare paid by title**
  - **Two variables (title vs. avg_fare):**

    ggplot(titanic_data,aes(x=title,y=avg_fare)) +

    geom _bar(stat="identity")
  - **Did this differ by class?**

    Split by pclass (rows):

    + facet_grid(pclass ~ .)

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data visualization:

- **Plot the average fair by passenger title, class and sex**
  - **Two variables (title vs. avg_fare):**

    ggplot(titanic_data,aes(x=title,y=avg_fare)) +

    geom _bar(stat="identity")

    Split by pclass and sex (rows ~ columns):

    + facet_grid(pclass ~ sex)

TITANIC ANALYSIS

Despite your work to suggest other movies to the Minister, they're still obsessed with the Titanic and have asked for you to:

Data visualization:

- **Were people in higher % occupancy boats less likely to survive?**
  - **Summarize with Dplyr:**

    Get % of passengers that survived by boat occupancy, sex + pclass
  - **Plot:**

```
ggplot(titanic_pct_summary, aes(x=pct_occupancy, y=pct_survived)) +
geom_point() +
facet_grid(pclass~sex)
```

POLICY SCENARIO: US POVERTY AND UNEMPLOYMENT

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

- A histogram showing the distribution of unemployment
- A line graph showing US Unemployment by state over time
- A series of boxplots showing the distribution of unemployment in the US by county; over time and
- A scatter plot showing county unemployment rates vs. the poverty rate in 2018

US COUNTY UNEMPLOYMENT AND POVERTY

#**Set-up** - the working directory ----

-(set your working directory to where the data is)

#**Load** the necessary **packages** ----

-dlyr, readxl, ggplot2, tidyr + stringr

#**Import**, **Explore + Clean** the data ----

-What do we need to import and how?

-Is the data ready to answer the questions? Why/why not?

#**Explain** - analyse, understand and visualize ----

-What plots do we need to answer the questions?

#**Save** - the results ----

-How are we going to use the results and how should they be saved?

US COUNTY UNEMPLOYMENT AND POVERTY

#**Import, Explore + Clean** the data ----

- **Data doesn't start in the first row**
- **Need to join datasets to compare unemployment and poverty**
- **Not tidy as years and measurements are stored in columns**
  - Reshape to long column names are in one column; and
  - Split into year/measurement
  - Clean the new columns so R knows year are numbers
  - Widen dataset so measurements are columns
- **Check if data is ready for analysis**
  - Sense check based on what *must* be logically true eg:
    - There is only one county observation per variable per year
    - Sum of labor force by county = state labor force totals
    - Unemployment and poverty rates <100%

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

- **A histogram showing the distribution of unemployment**
  - What do we need to filter to make this sensible?
  - How many variables do we want to display?
  - What geom() makes sense?
  - What else do we need to consider?

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

- **A line graph showing US Unemployment by state over time**
  - What do we need to filter to make this sensible?
  - How many variables do we want to display?
  - What geom() makes sense for this?
  - What else do we need to consider?

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

- **A series of boxplots showing the distribution of unemployment in the US by county; over time and**
  - How many variables do we want to display?
  - What geom() makes sense?
  - What else do we need to consider?

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

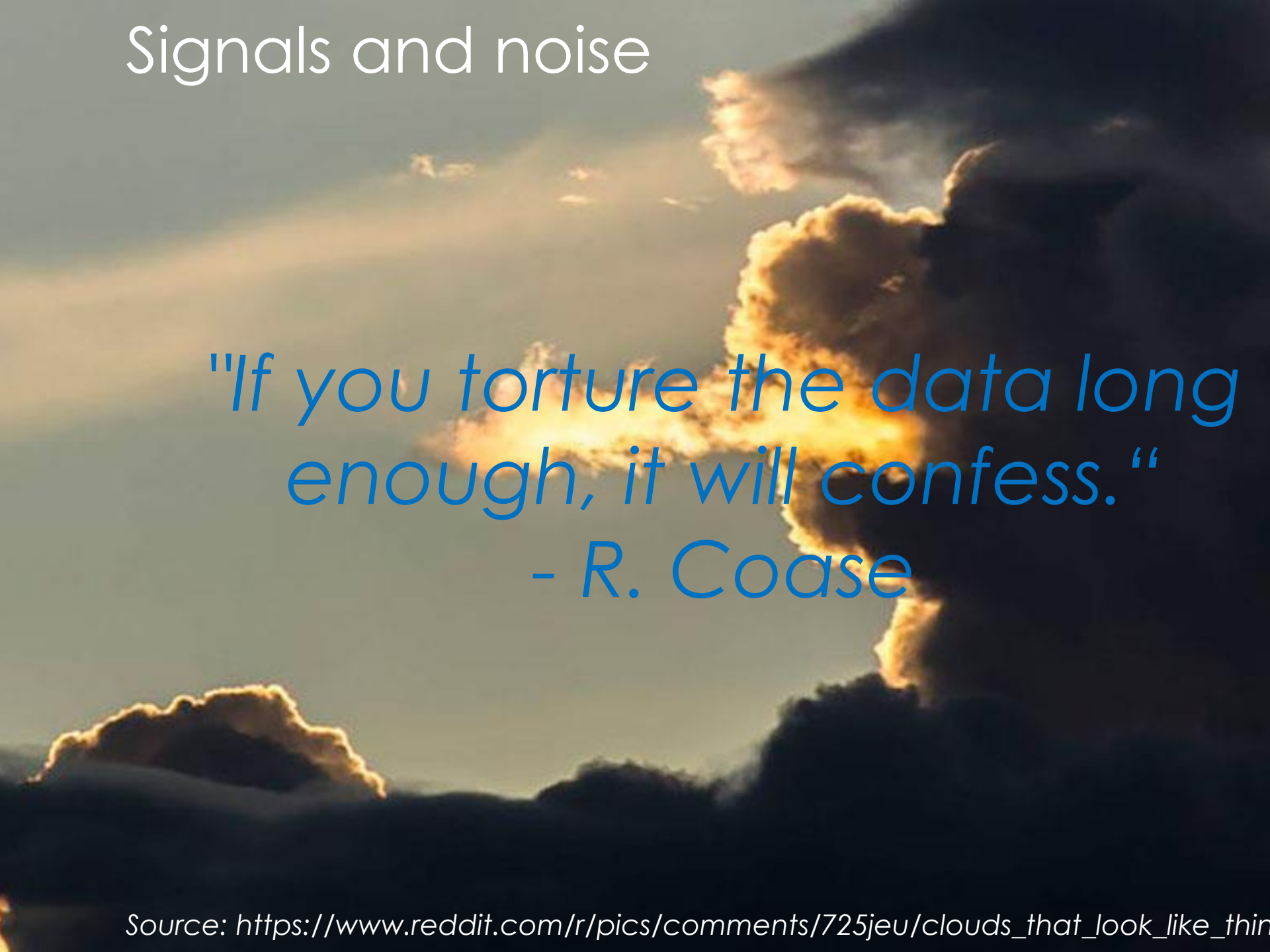
- **A scatter plot showing county unemployment rates vs. the poverty rate in 2018**
  - What do we need to filter to make this sensible?
  - How many variables do we want to display?
  - What geom() makes sense?
  - What else do we need to consider?

US COUNTY UNEMPLOYMENT AND POVERTY

The President has asked for some information to assist their speech on economic growth and poverty in the US. Being an economist they're particularly interested in winning over voters with graphs and would like to present:

- A histogram showing the distribution of unemployment
- A line graph showing US Unemployment by state over time
- A series of boxplots showing the distribution of unemployment in the US by county; over time and
- A scatter plot showing county unemployment rates vs. the poverty rate in 2018

Signals and noise



Source: https://www.reddit.com/r/pics/comments/725jeu/clouds_that_look_like_thin

SUGGESTED SWIRL EXERCISES

- **Before next week complete**
- ~~Subsetting Vectors~~
- ~~Matrices and Data Frames~~
- ~~Looking at Data~~
- ~~Dates and Times~~
- **Base Graphics**
- **Logic**

WEBINAR EVALUATION

(See chat for evaluation link)

AN INTRODUCTION TO R FOR POLICY ANALYSIS