ETC5523: Communicating with Data

Week 1

Advanced Reproducible Practices

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5523.Clayton-x@monash.edu
4th August 2020

👩🏻‍🏫 ETC5523 Teaching Team

Emi Tanaka

Lecturer Wk 1-6

Chief Examiner

Stuart Lee

Lecturer Wk 7-12

Mitchell O'Hara-Wild

Tutor

ℹ️ ✉ ETC5523.Clayton-x@monash.edu

for confidential matters

🎯 ETC5523 Learning Objectives

1. Effectively communicate data analysis, using a blog, reports and presentation.
2. Learn how to build a web app to provide an interactive data analysis.
3. Learn to construct a data story.

🗣️ Communicating

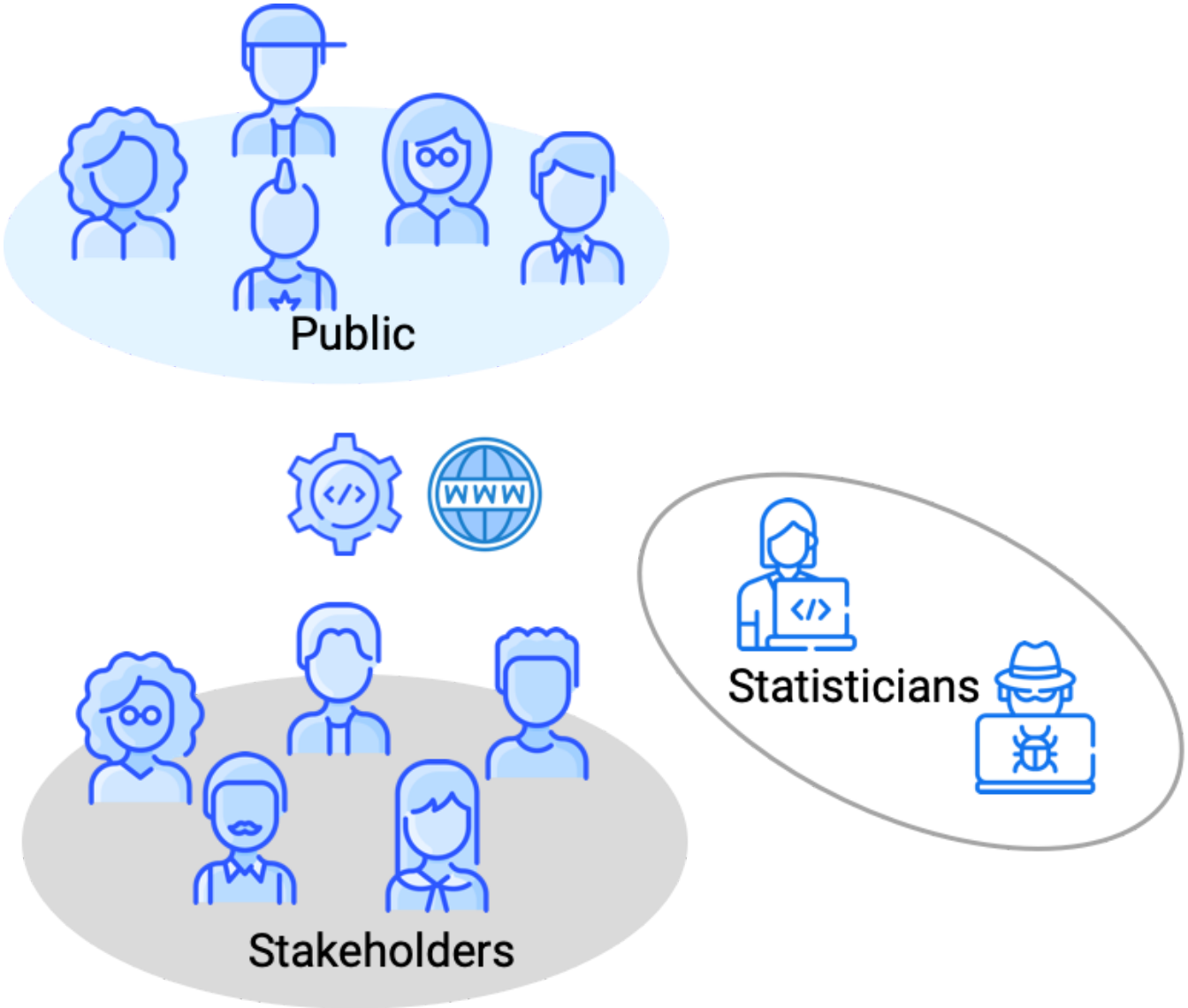
> To effectively communicate, we must realize that we are all different in the way we perceive the world and use this understanding as a guide to our communication with others.
>
> —Anthony Robbins

Your engagement with your peers will be critical to form your understanding and become a better communicator.

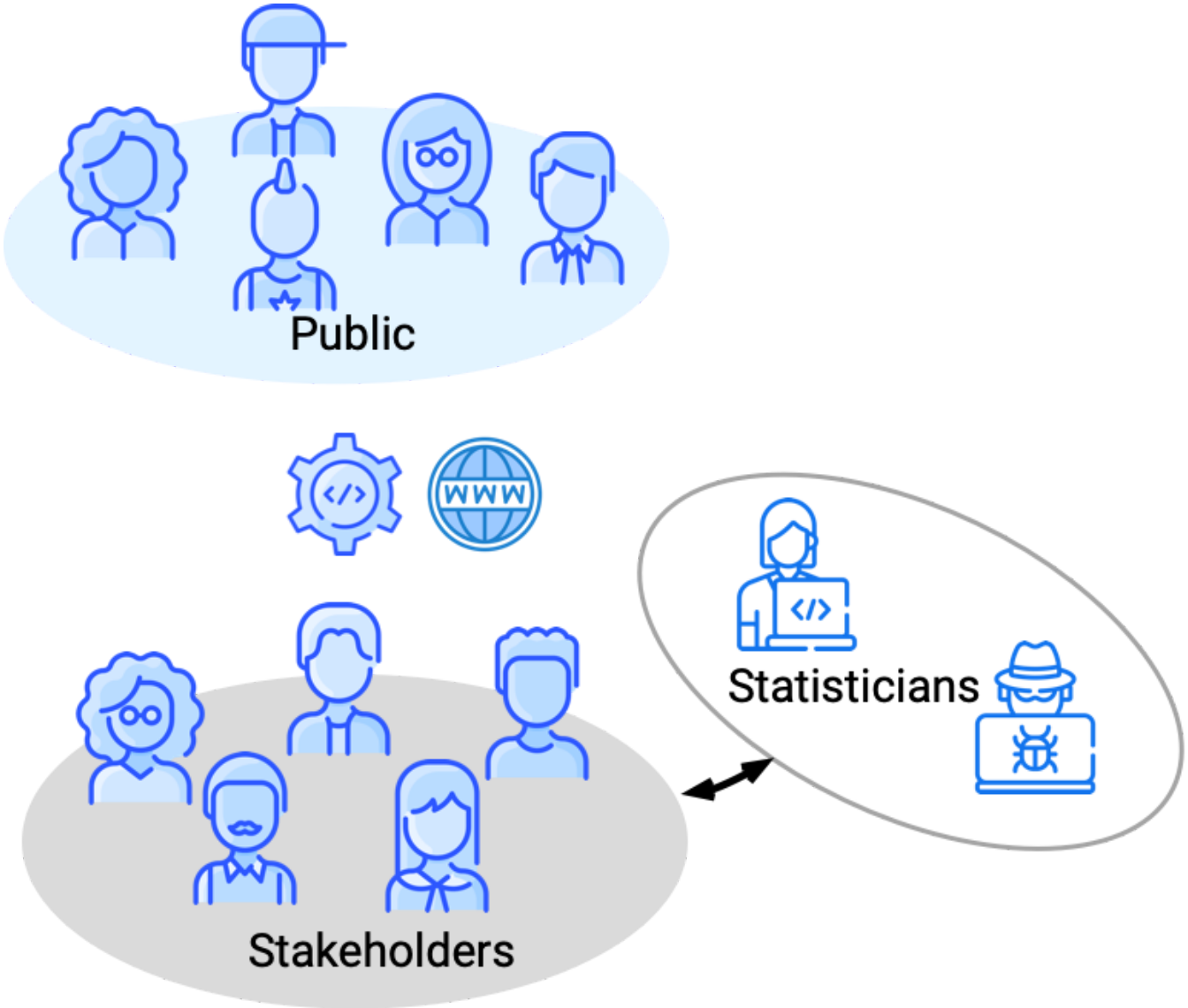
📈 Communicating *with Data*

> The two words 'information' and 'communication' are often used interchangeably, but they signify quite different things. Information is giving out; communication is getting through.
>
> —Sydney J. Harris

In this course, you will construct narratives from your number-crunching to tell a compelling data story.

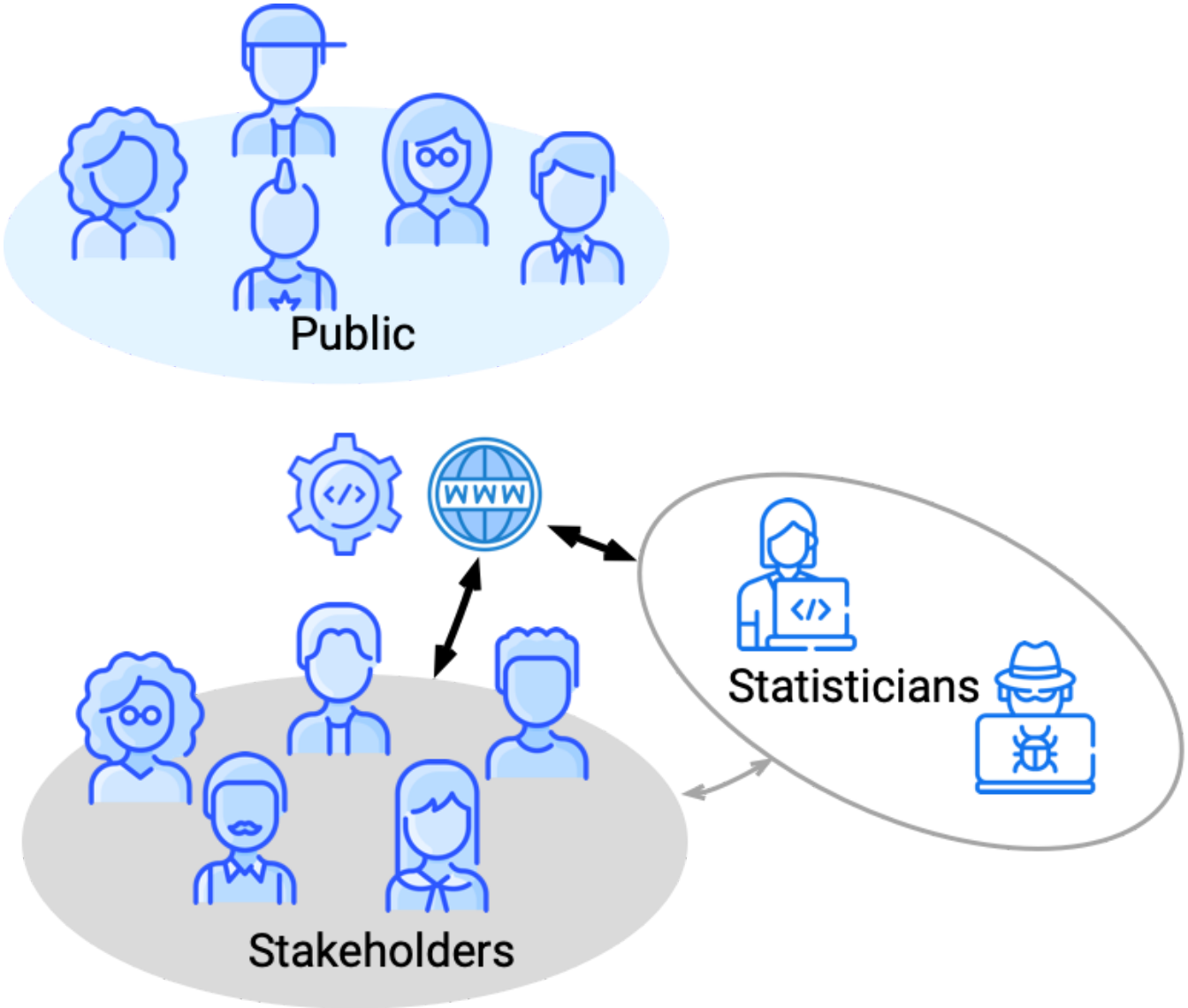


Who do we communicate with as analysts, statisticians, or data scientists?



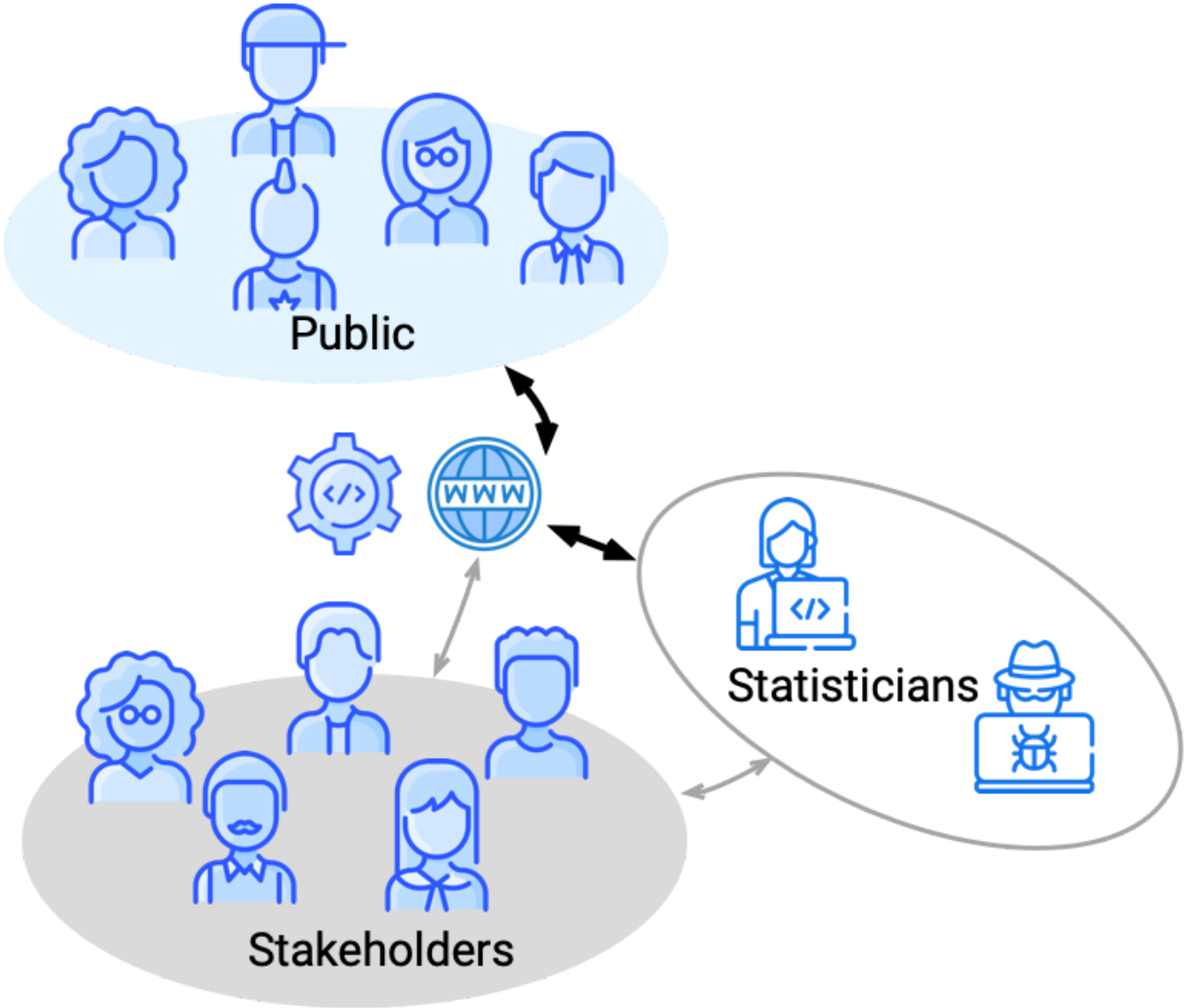
Statisticians
↕
Stakeholders

- Stakeholders can have a span from zero to proficient statistical knowledge.
- They may be technicians, domain-knowledge experts, managers, decision-makers and so on.



Statisticians ↕ Stakeholders

- You may not directly interact with the stakeholders but rather through product (e.g. reports or tools) that you provide the stakeholders.

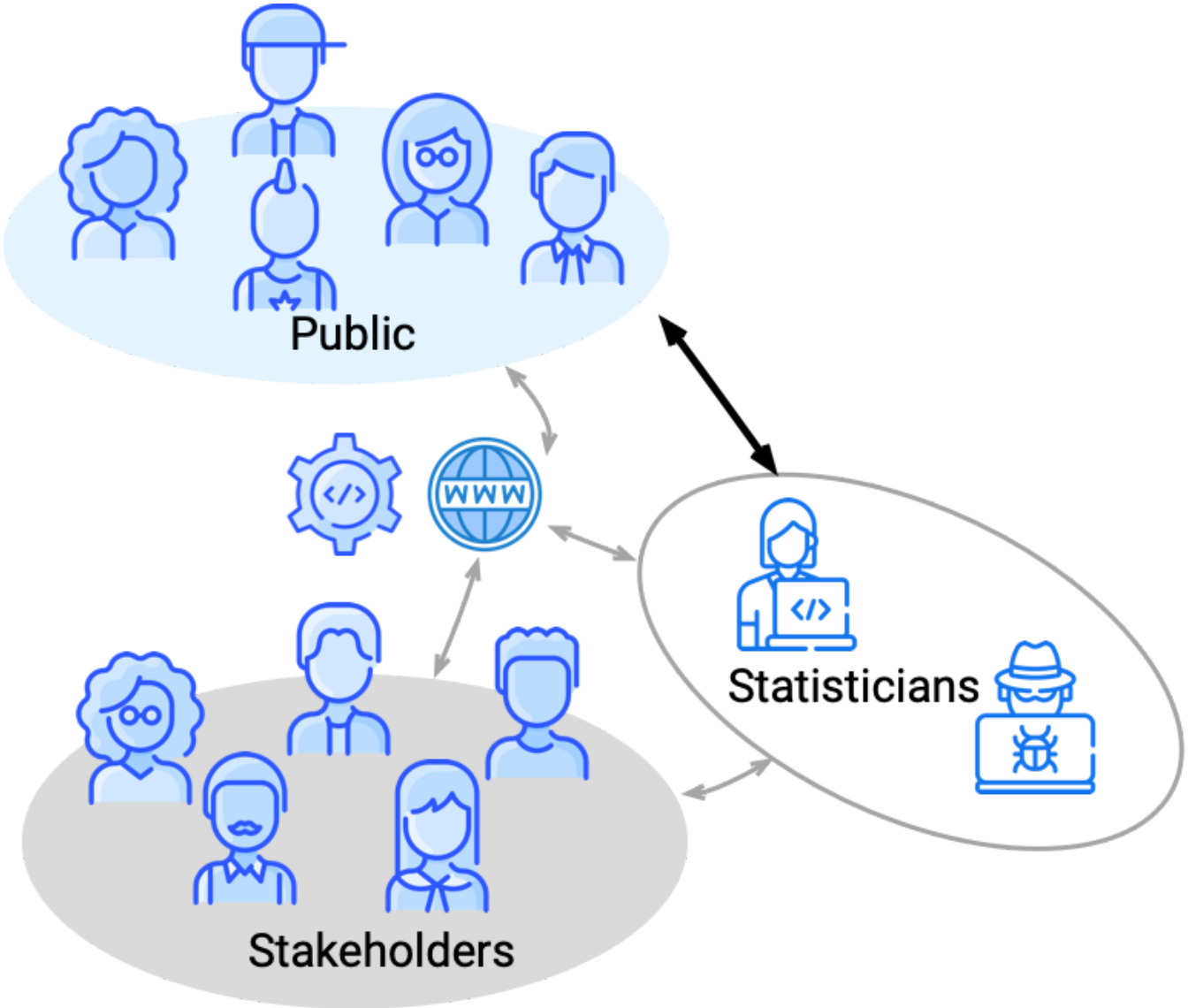


Statisticians

↕

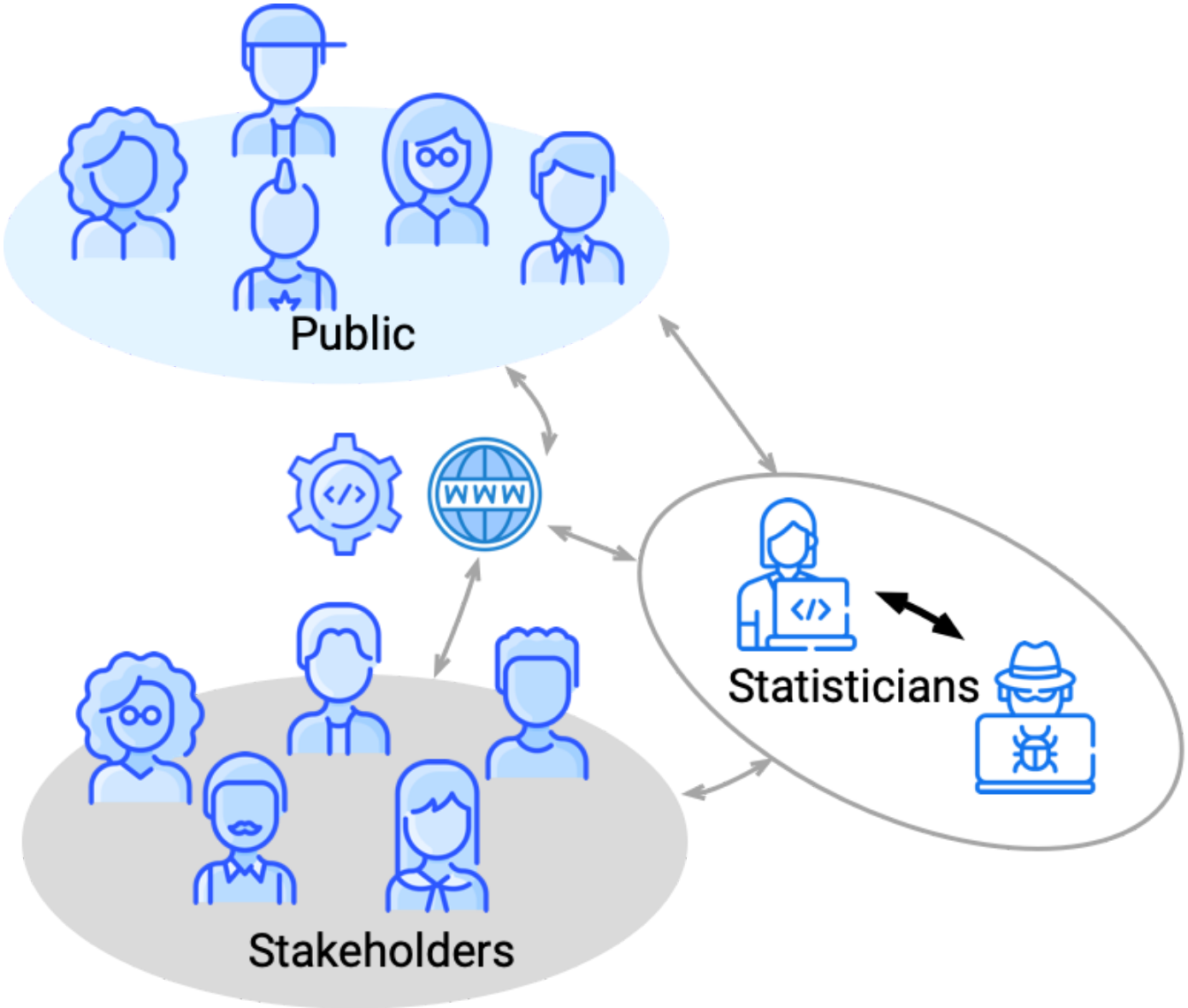
General Public

- Reports, blogs, social media posts, tools and so on may also be consumed by the general public.
- The general public will have a variety of statistical background.



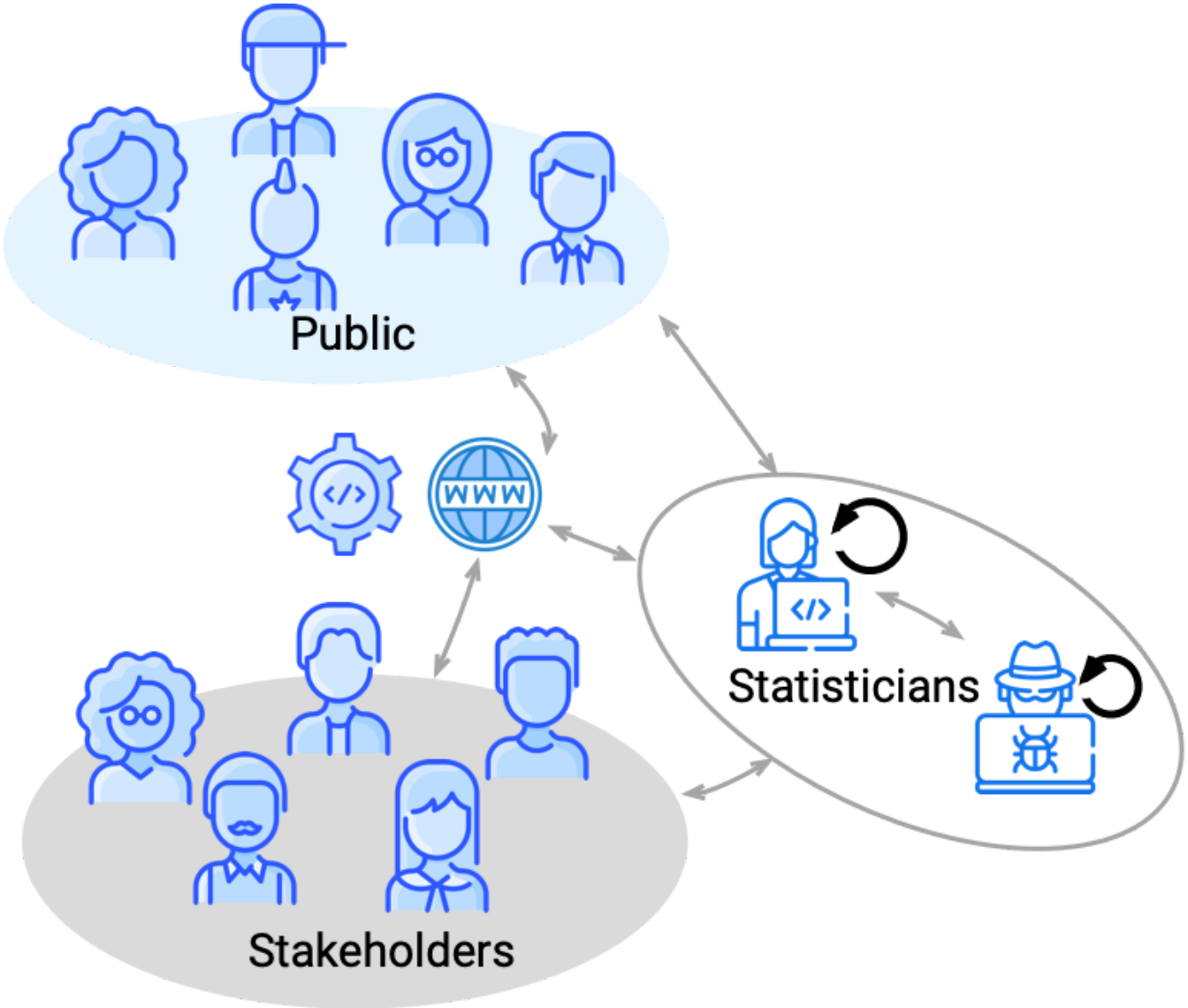
Statisticians
↕
General Public

- You may directly communicate with public as well (presentations, workshops, outreach and so on)



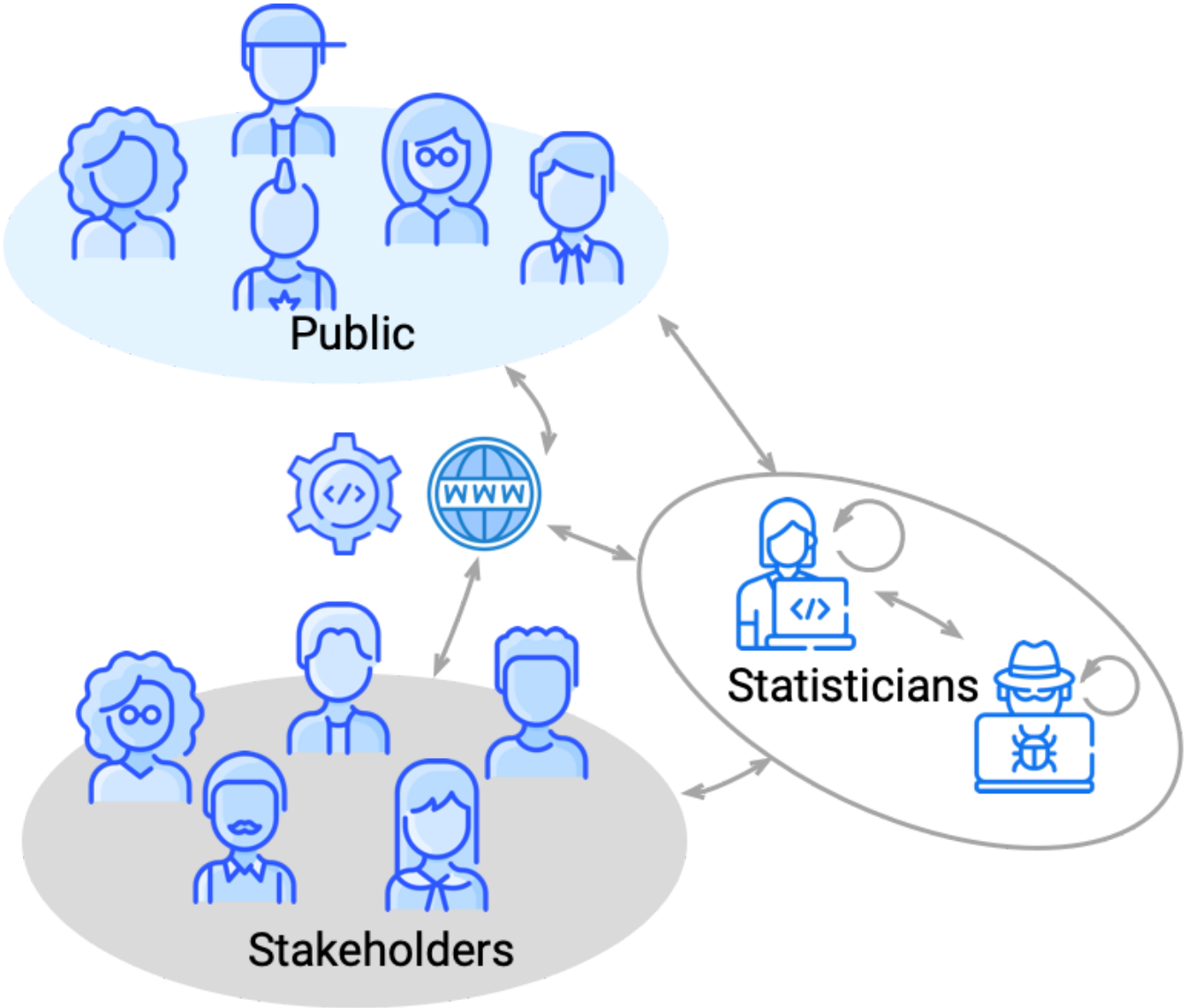
Statistician
↕
Statistician

- You also talk to your peers (directly or indirectly)



You
↓
Future You

- Future you also needs to understand what past you did!



You
↕
Another Party

- All communications are represented as two-way here

  (with one exception, as you can't travel back in time to talk to past self)

- Ideally you want the communication to be two-way, but in practice, communication may only be possible one-way.

Communication isn't just between humans

Computer systems communicate with each other.

You'll learn about web and data technologies next week.

🤝 Key Ingredient of a Successful Communication

TRUST

Trust in the **data collection process**

Trust in the **data entry**

Trust in the **data processing**

Trust in the **number crunching**

Trust in **your skills**

Trust **assumptions are reasonable**

Trust in the **software**

Trust in **interpretations**

👍 Building Trust

Minimise errors

- We are all **prone to errors** no matter how skilled we are.
- Adopting **reproducible practices** (e.g. git + R Markdown) helps to **minimise this error**. (note: it does not completely eradicate it)

Be transparent

- Share code and data (where you can). This not only builds trust in the results, this gives opportunities for others to correct any errors. See also about open science, open-source software and open-data.
- Give credit and acknowledgement where due.

Interrogate your data, tools and sources

- Use trusted software (preferably with proper testing).
- Sanity check your data, tools and sources.

What statistical software do you use and why?

03:00

Unit Structure

2 hour lectures

Tues 4.00-6.00PM

1.5 hour tutorials

you can only attend the one that you are allocated in

Weds 4.00-5.30PM OR Thu 5.00-6.30PM

Attendances are recorded.

Resources

Course homepage

this is where you find the course materials
(lecture slides, tutorials and tutorial solutions)

https://cwd.numbat.space/

Moodle

this is where you find discussion forum, zoom links, assignments and marks

https://lms.monash.edu/course/view.php?id=81235

Engagement (total 10%)

- Lecture and tutorial participation will contribute to this.
- If you have some clashes for lecture or tutorial, there are other opportunities to show your engagement, including:
  - posting in the Moodle discussion forum,
  - github commits (more on this later in the course),
  - peer reviewing (more on this later in the course), and
  - online submissions (see Tutorial Q1C for example).
- There should be **at least 3 measurable and meaningful engagement activities each week** (contributing to 1% of your final mark capped to 10% over the whole semester; in other words, you can miss 2 weeks, or 4 weeks counting mid-semester break although harder to gain engagement marks over that period, without impacting your mark).

Take Home Assessment (15%)

- Will be made available at Thu 20 Aug 7PM.
- You will have 24 hours (i.e. **due Fri 21 Aug 7PM**) to complete the assessment.
- You will be given an document that you have to reproduce using R Markdown.

Blog Entries (20%)

- You will have 4 blog entries featuring a data story **due Fri 4 Sep 2020**, **Fri 18 Sep 2020**, **Fri 16 Oct 2020** and **Fri 30 Oct 2020**.
- You will write these blogs using R Markdown.

Assignment 1: Making a Shiny web app (20%)

- You will make an interactive web application to communicate a data story using `shiny`
- The app will be **due Fri 9 Oct**

Assignment 2: Making an R-package (25%)

- You will make an R-package that has clear documentation, vignettes, consistent coding style, and robust functions (as demonstrated by testing).
- The R-package will be **due Fri 23 Oct**

Group Presentation (10%)

- Group of 3-4 to present 5 minutes at **Tue 3 Nov 4PM**

Expectations Part 1/3

- Lectures are recorded but you are expected to have either attended the lecture, or watched the recordings fully, prior to the tutorial for the week.
- Tutorials may not be recorded, and attendance is expected.
- Questions related to the course should be raised at moodle discussion forum.
- For personal or private administrative issues, the email contact is: ETC5523.Clayton-x@monash.edu
- If you register after the start of the semester or if you miss a lecture/tutorial, it is your responsibility to catch up with missed material, learn about due dates for material to be turned in, and getting assigned to a group for team work, as necessary.
- All times are given in AEST (Melbourne time).

🔶 Expectations Part 2/3

- The computer software R and RStudio IDE will be used for the unit.
- You will need to install latest versions of all necessary software (below is not an exhaustive list for this course).

```
install.packages(c("tidyverse", "rmarkdown", "shiny", "broom",
  "skimr", "blogdown", "distill", "knitr", "kableExtra", "feasts",
  "devtools", "usethis", "testthat", "roxygen2", "knitr", "fable",
  "shinydashboard", "flexdashboard", "DT", "xaringan", "tsibble",
  "stargazer", "htmlwidgets", "bookdown", "pagedown", "magick",
  "crosstalk", "plotly", "colorspace", "pkgdown", "posterdown",
  "glue", "lme4", "agridat", "catdata", "gganimate", "htmltools",
  "leaflet", "linl", "patchwork", "shinytest", "shinyjs",
  "tinytex", "xaringanthemer", "lmerTest", "DBI", "dbplyr",
  "RMySQL", "RPostgreSQL", "RSQLite", "bigrquery", "odbc",
  "rvest", "nycflights13", "palmerpenguins", "eechidna"))
```

🔶 Expectations Part 3/3

- **ETC5513 is prerequisite**. Some of you may have exemption (due to evidence of equivalent knowledge to ETC5513) or gone through the bridging course instead.
- It's expected that you know how to use **git**, **github** and basic **R Markdown** (as taught in ETC5513).
- The priority of the ETC5523 teaching team is to support you in ETC5523 material.
- It's essential in this course that you have a GitHub, RPubs, shinyapps.io and Netlify (connected to your GitHub account) accounts. All are free to register.

GitHub Classroom

We are going to use GitHub Classroom to keep track of your repos and distribute templates.

Engagement opportunity

1. Start with the test assignment by clicking on the link given in Moodle **before Fri 7 Aug 11PM** and make sure you identify yourself by connecting your name to the roster.
2. Once you have accepted it (note: some browsers do not work well with GitHub Classroom so use Chrome or Firefox), you can find your repo here:
   https://github.com/etc5523-2020
3. If you don't recall how to connect this repo locally, check this guide.
4. Make some small changes to your repo and make sure these changed are pushed to GitHub.

Questions?

R Markdown

Case Studies

There are more features in R Markdown than we can show you.

The aim for you should be to independently search and find your solution.

Case study ❶ knitr::opts_chunk$set

- Repeating chunk options can be painful in your workflow.
- If you are say writing `echo = FALSE` for every single chunk, you might as well set the default chunk option to `echo = FALSE`.
- Remember **don't repeat yourself** (DRY) every time.
- In order to change the default chunk options you can use

```
knitr::opts_chunk$set(echo = FALSE)
```

- You can get the current default by using the following command:

```
str(knitr::opts_chunk$get())

## List of 53
##  $ eval          : logi TRUE
##  $ echo          : logi TRUE
##  $ results       : chr "markup"
##  $ tidy          : logi FALSE
##  $ tidy.opts     : NULL
##  $ collapse      : logi FALSE
##  $ prompt        : logi FALSE
##  $ comment       : chr "##"
##  $ highlight     : logi TRUE
##  $ strip.white   : logi TRUE
##  $ size          : chr "normalsize"
##  $ background    : chr "#F7F7F7"
```

⚙️ Setting Chunk Options

- You can always overwrite the default for each chunk.

```
```{r, message = FALSE, warning = FALSE}
library(tidyverse)
```
```

- For chunk options, check out https://yihui.name/knitr/options/.

🧹 Organising your Rmd file

- ⚙️ It is good practice to set default chunk options at the beginning of your Rmd file just as it is good practice to load all packages needed at the beginning.
- 👁️ This makes it easier to quickly see what the expected behaviour of the chunks for anyone (including yourself) looking at the file.

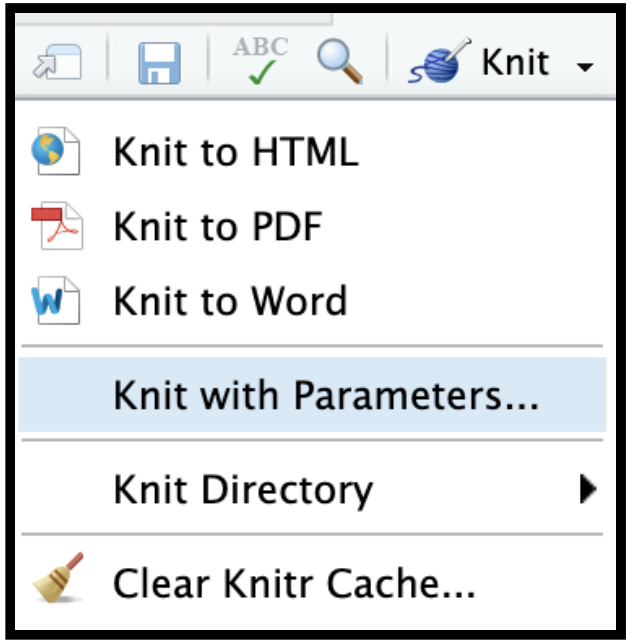
Case Study ❷ Parameterized R Markdown Reports Part 1/2

````
---
title: "Parameterized Report"
params:
   year: 2019
output: html_document
---

```{r, message = FALSE, warning = FALSE}
library(dplyr)
library(eechidna) # get the tcp data
df <- get(paste0("tcp", substr(params$year, 3, 4)))
df %>%
  filter(Elected == "Y") %>%
  group_by(PartyNm) %>%
  tally() %>%
  arrange(desc(n)) %>%
  knitr::kable(caption = paste("Election", params$year))
```
````

Case Study ❷ Parameterized R Markdown Reports Part 2/2

```
---
title: "Parameterized Report"
params:
  year:
    label: "Year"
    value: 2019
    input: radio
    choices: [2001, 2004, 2007, 2010, 2013, 2016, 2019]
output: html_document
---
```



See more input type here.

Case Study ❸ Render Rmd via Command Line

lecture-01suppA.Rmd

````
---
title: "Parameterized Report"
params:
   year: 2019
output: html_document
---
```{r, message = FALSE, warning = FALSE}
library(dplyr)
library(eechidna) # get the tcp data
df <- get(paste0("tcp", substr(params$year, 3, 4)))
df %>%
  filter(Elected == "Y") %>%
  group_by(PartyNm) %>%
  tally() %>%
  arrange(desc(n)) %>%
  knitr::kable(caption = paste("Election", params$year))
```
````

You can knit this file via R command by using the render function:

```
library(rmarkdown)
render("lecture-01suppA.Rmd")
```

You can overwrite the YAML values by supplying arguments to render:

```
library(rmarkdown)
render("lecture-01suppA.Rmd",
  output_format = "pdf_document",
  params = list(year = 2016)))
```

Case Study 4 theme for html_document

You can change the look of the html document by specifying themes:

- default
- cerulean
- journal
- flatly
- darkly
- readable
- spacelab
- united
- cosmo
- lumen
- paper
- sandstone
- simplex
- yeti
- NULL

```
output:
  html_document:
    theme: cerulean
```

These bootswatch themes attach the whole bootstrap library which makes your html file size larger.

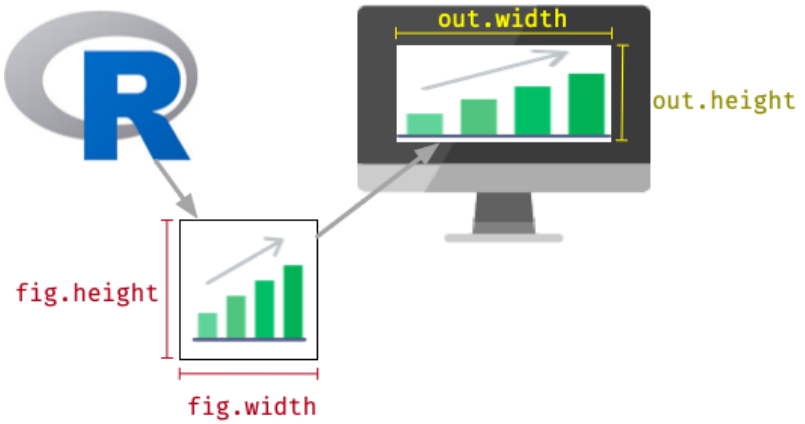
Figures in R

In R, we can classify pictures into two types:

1. Figures constructed using R with result communicated to a specific graphical device.
2. Images already existing as a filie elsewhere with extensions such as `bmp`, `eps`, `gif`, `jpeg`, `pdf`, `png`, `psd`, `svg`, `tex`, `tiff`, or `wmf`.

Case Study 5 Figures in R Markdown

- Some figure chunk options are only for plots created by R, e.g. ⚙ `fig.width`, `fig.height`, and `fig.process`.



- Some are for all, e.g. ⚙ `out.width`, `out.height`, and `fig.link`.
- Some are output specific, e.g. ⚙ **LaTeX**: `fig.env`, `fig.scap`, `fig.pos`, `fig.ncol`, `fig.sep`, **tikz**: `resize.width`, `resize.height`, `external`, `sanitize`, **HTML**: `fig.retina`.

Case Study 6 fig.process

lecture-01suppC.Rmd

lecture-01suppC.html

````
---
title: "Australian Federal Election"
output: html_document
---


```{r setup, include = FALSE}
library(tidyverse)
library(eechidna) # get the tcp data
library(magick)
process_tcp <- function(df) {
  df %>%
  filter(Elected == "Y") %>%
  group_by(PartyNm) %>%
  tally() %>%
  mutate(PartyNm = fct_reorder(PartyNm, -n))
````
````

Case Study 7 Code Snippets

Live Demo

Case Study 8 Sharing HTML reports

Live Demo

That's it!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5523.Clayton-x@monash.edu