Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 2

Dr. Ibrahim Radwan

DISTINCTIVE BY DESIGN

OUTLINE

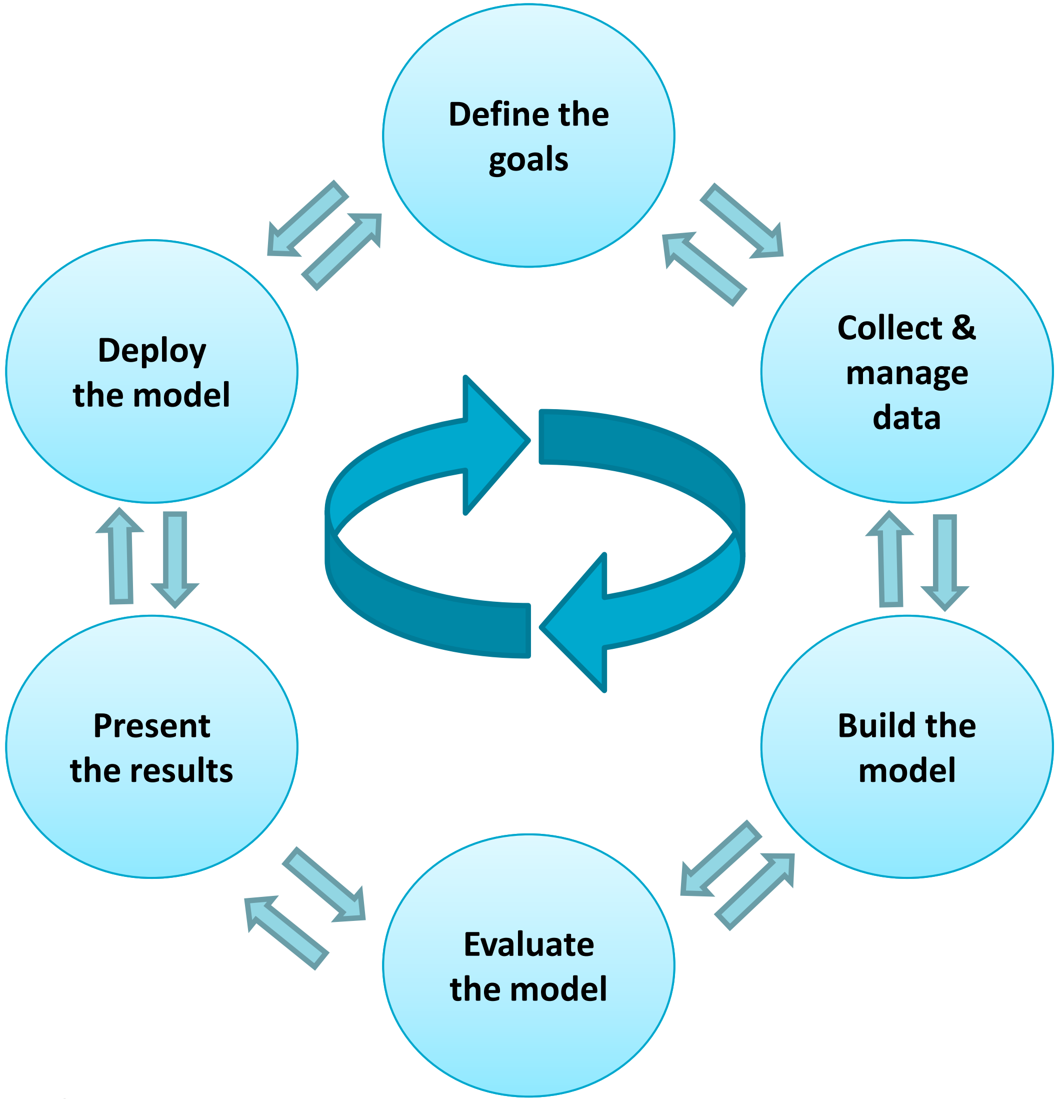
- Stages of a Data Science Project
- Professional Ethics of Data Science
- Programming in R - Basics:
  - Why R?
  - Variable assignments
  - Basic data types
  - Mathematical and logical operations in R

STAGES OF A DATA SCIENCE PROJECT

UNIVERSITY OF CANBERRA

It is a project life cycle, so loops inside loops

You may move back and forth from any stage to any stage



The boundaries are fluid, so the activities are overlapping

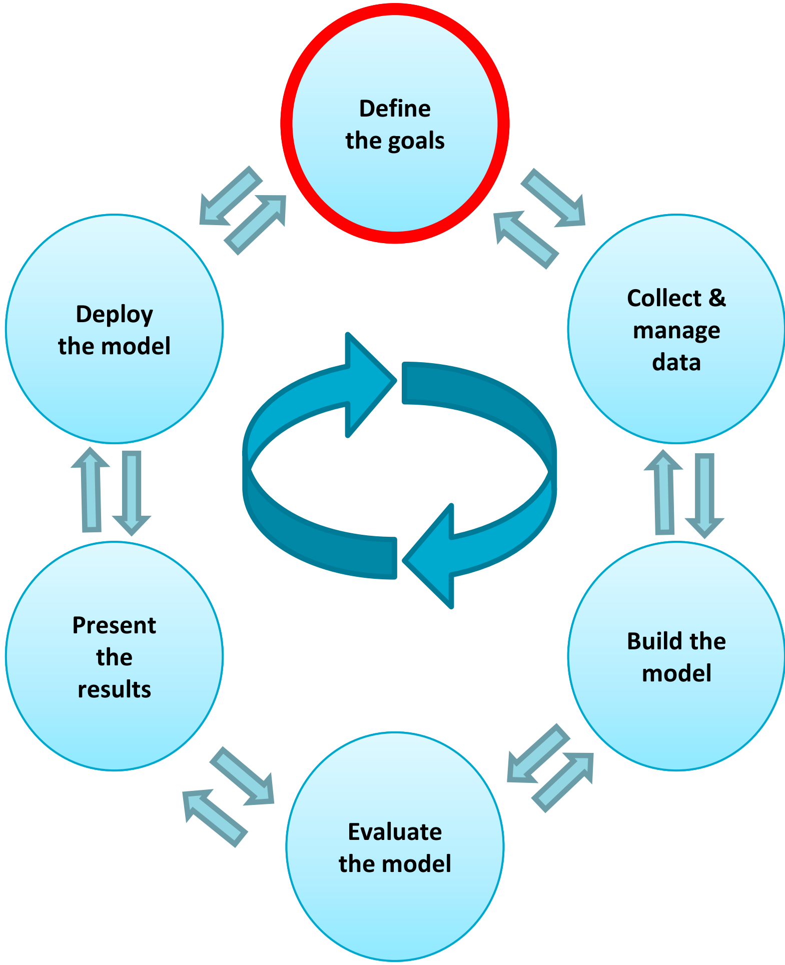
Mostly, the project will be ended with a follow-up one

N. Zumel and J. Mount, "Practical Data Science with R"

STAGES OF A DATA SCIENCE PROJECT(2)

Business understanding

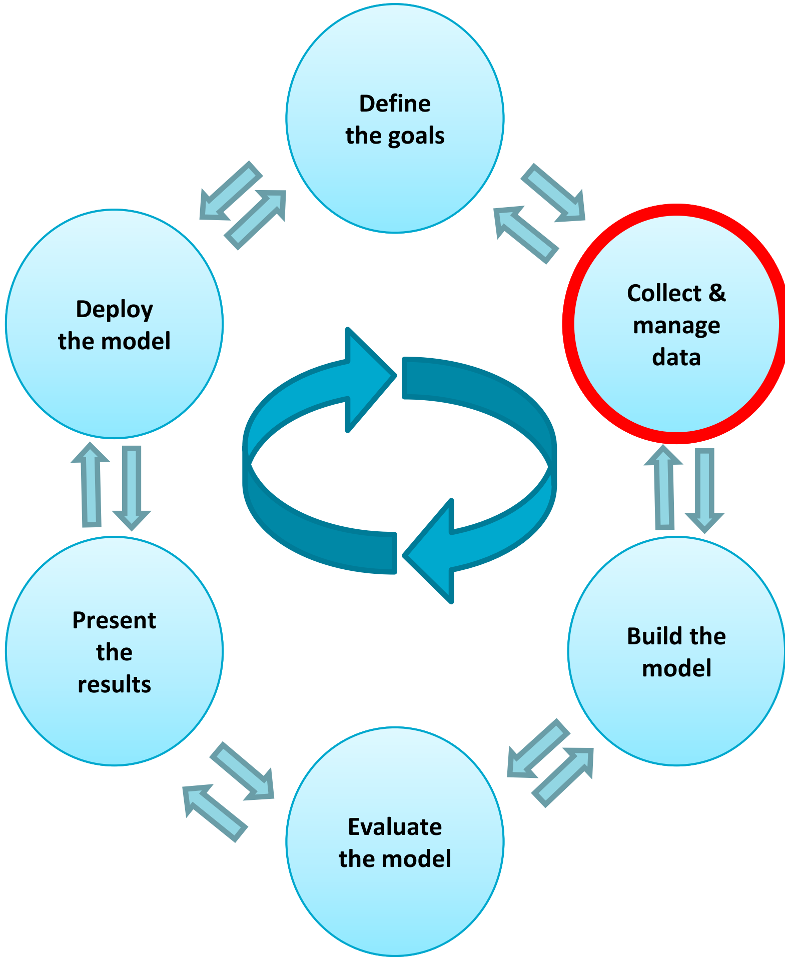
- To define the goals for your project, you need to find answers to these questions:
  - What is the problem that I am solving?
  - What do the project sponsors lack, and what do they need?
  - What are they doing to solve the problem now, and why is not it good enough?
  - What resources (e.g. data) do I need?
  - Will I have domain experts to collaborate with?
  - What are the thoughts of the project sponsors to deploy the results?
  - Do I have a "*Definition of Done*" for this project?



Measurable and quantifiable goals are the key toward achieving the objectives

STAGES OF A DATA SCIENCE PROJECT(3)

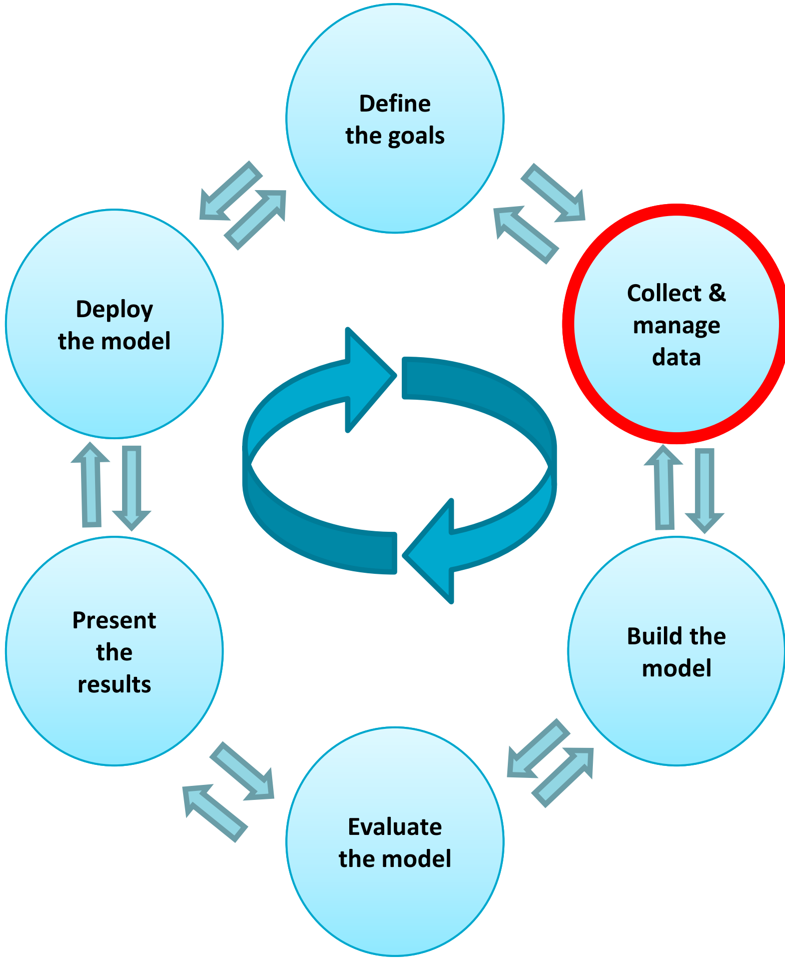
- Start with these questions to understand what data you have:
  - What data is available to me?
  - Will it help to solve the problem? Is it enough?
  - What quality is the data in?
- Data Acquisition:
  - If needed, you search for other data sources or collect some.
  - You may need to understand the data sources: on-premises vs. cloud databases vs. Files
  - You may need to understand the rates of the Data: Streaming vs. Batch or Low vs. High-quality



STAGES OF A DATA SCIENCE PROJECT(4)

UNIVERSITY OF CANBERRA

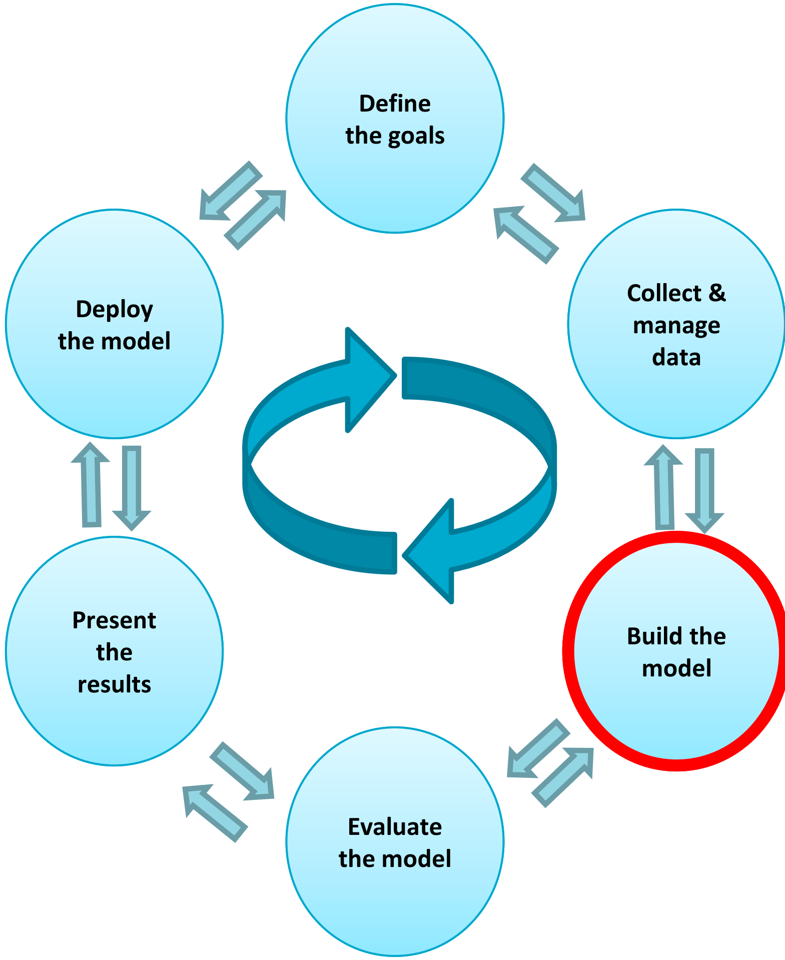
- Data Management:
  - Wrangling
    - Transforming the data into proper format for the sake of analysis
    - Same as data munging
  - Exploration
    - Summarise the statistics of different variables of the data
    - Use visualisation tools to visualise the variable distributions to understand the characteristics of these data
  - Cleaning
    - Remove duplicates and non-relevant records
    - Remove NAs, or replace them with specific values



STAGES OF A DATA SCIENCE PROJECT (5)

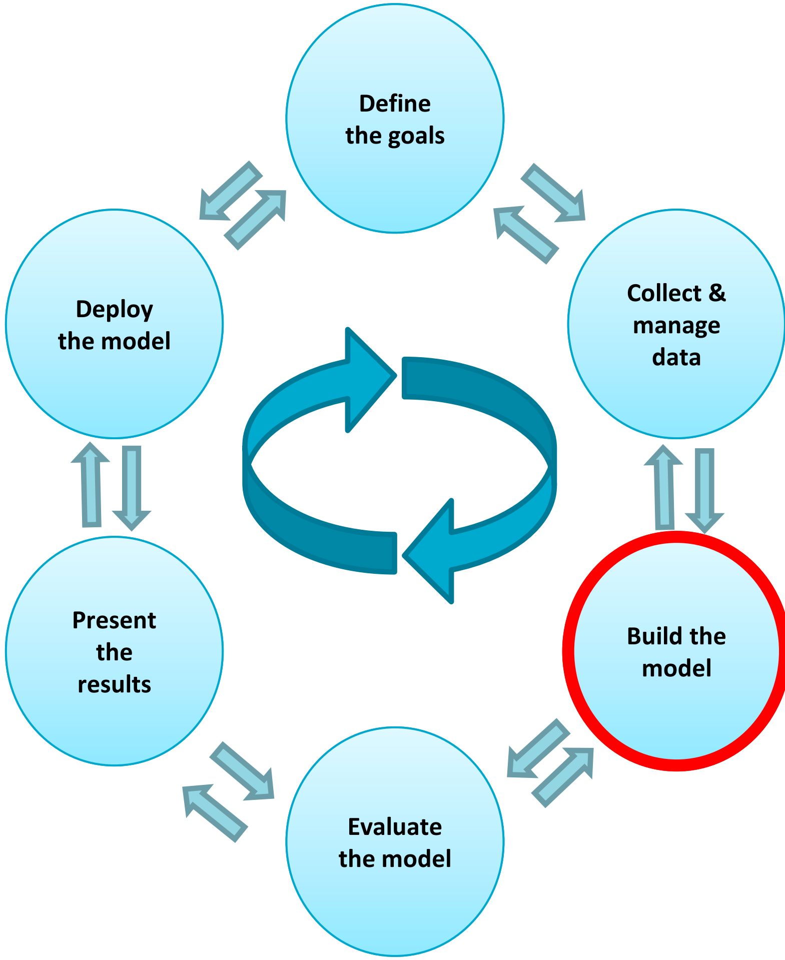
- Data Modelling makes assumptions about the data distributions & relationships
  - , therefore, we may go back and forth between this step and the data management step

- Also, using Statistical and machine-learning based approaches to build models that are useful to extract insights from the data



STAGES OF A DATA SCIENCE PROJECT (6)

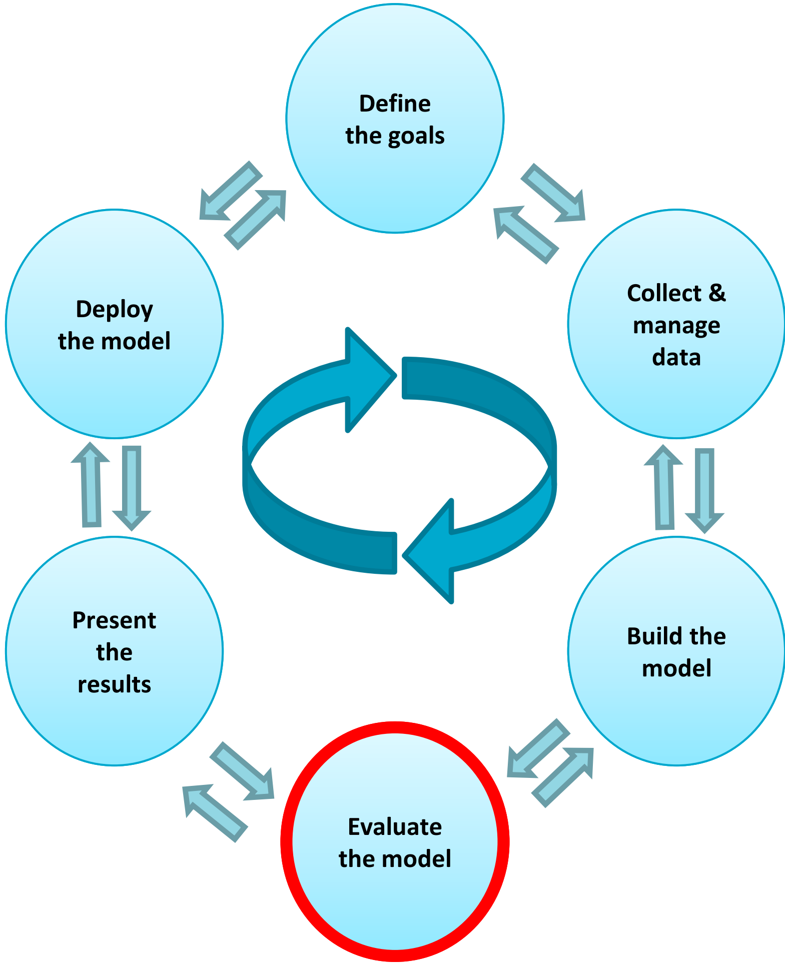
- Common Data Modelling tasks include:
  - Classification
  - Regression
  - Clustering
  - Ranking
  - Scoring
- Data Modelling can be classified into:
  - Supervised Learning,
  - Unsupervised Learning and
  - Reinforcement Learning



STAGES OF A DATA SCIENCE PROJECT (7)

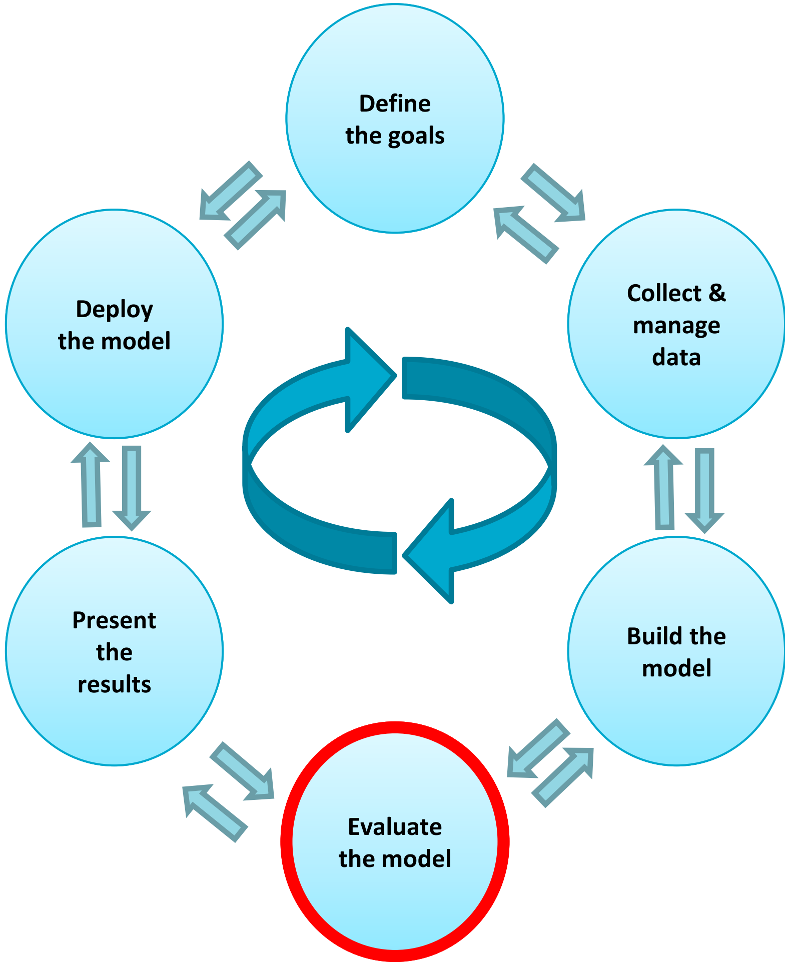
- Model Evaluation
  - you need to determine if the built model meets your goals and is accurate enough
- Is it better than the obvious guess?
- Do the results make sense in the context of problem-domain?

- If "No" to any of the previous questions, you may need to loop back to the modelling step or even to the data selection and management process



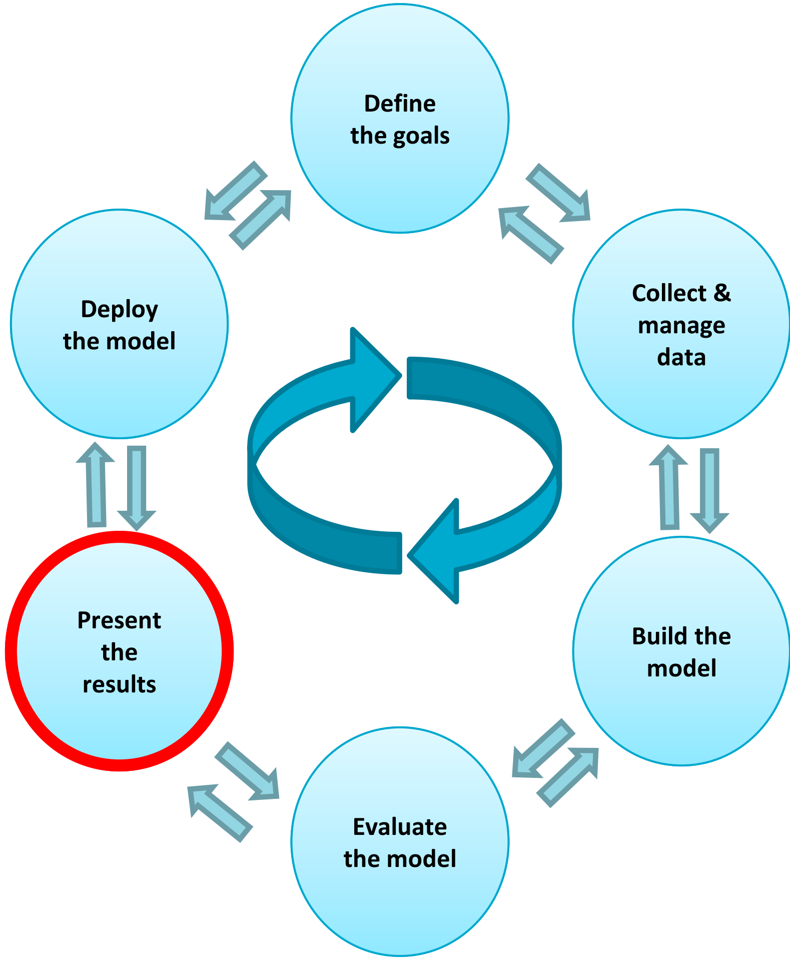
STAGES OF A DATA SCIENCE PROJECT (8)

- Examples of evaluation criteria to evaluate a model:
  - Confusion Matrix,
  - Precision-Recall,
  - Area under Curve
  - F1 Score
  - Mean Absolute Error
  - Mean Squared Error



STAGES OF A DATA SCIENCE PROJECT (9)

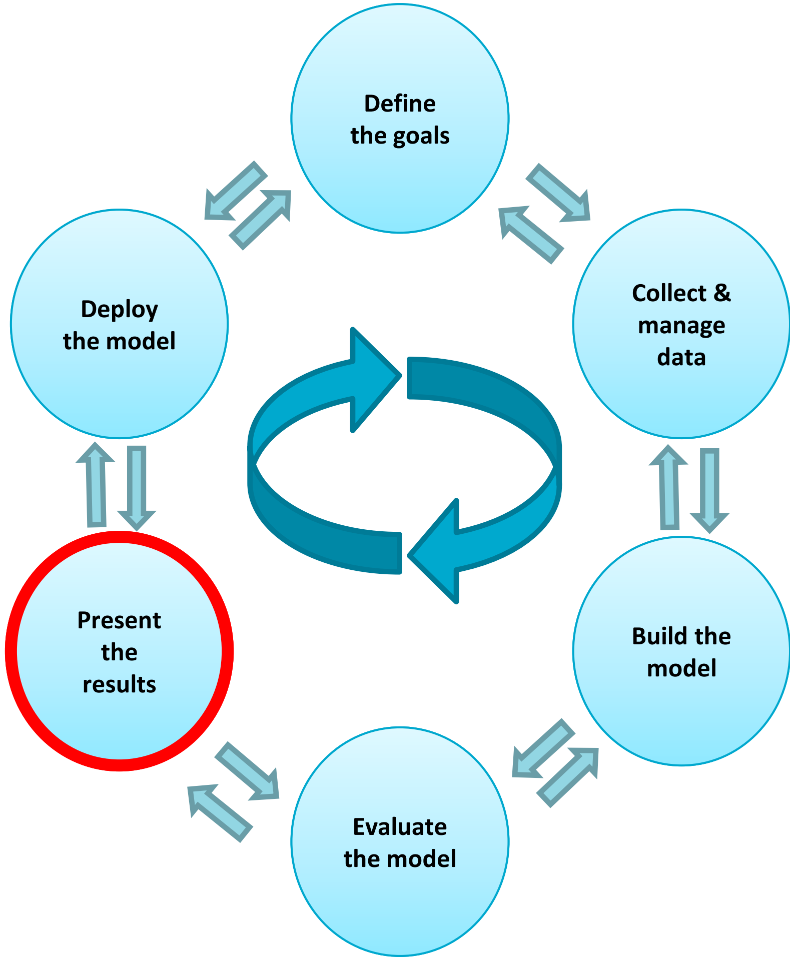
- Once you have a model that meets your selection criteria:
  - You need to present the results to the project sponsors and the different stakeholders with a language that is understandable by them
  - Different audiences require different kind of information
  - Focus on the most interesting findings or recommendations



STAGES OF A DATA SCIENCE PROJECT (10)

UNIVERSITY OF CANBERRA

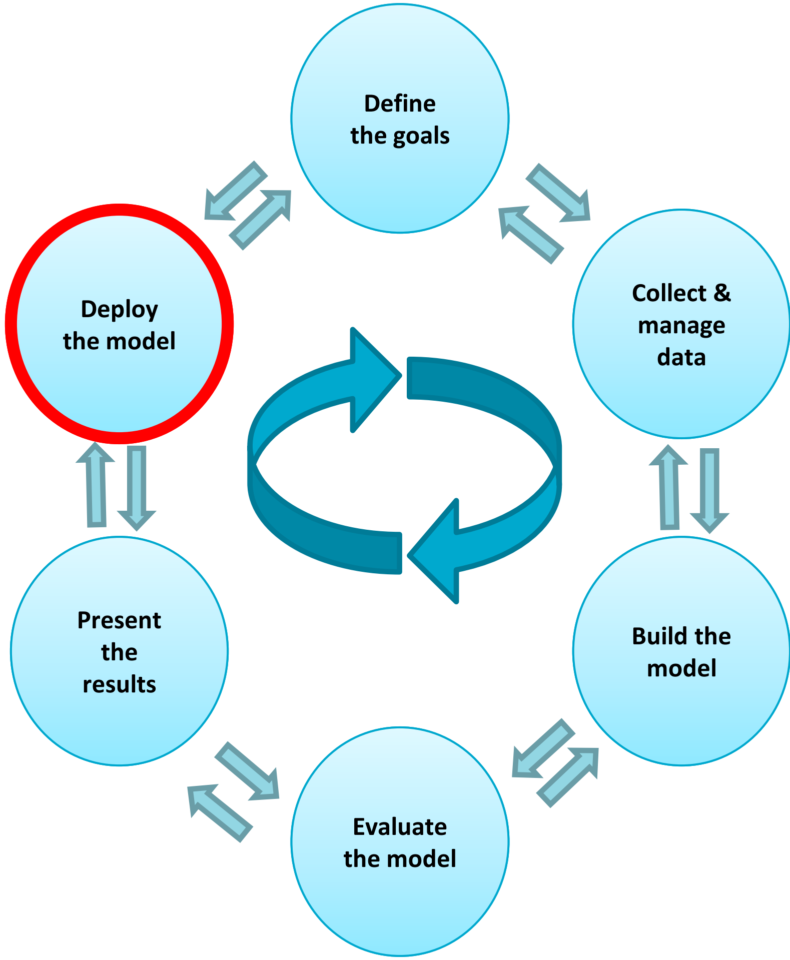
- While presenting and documenting your results, you need to consider the following:
  - How should they interpret the model?
  - What does the model output look like?
  - If the model uses a specific evaluation criteria, such as F1 score, how should they use this criteria?
  - What are the impacts of your model on the resources, which the operation staff are responsible for?



STAGES OF A DATA SCIENCE PROJECT (11)

UNIVERSITY OF CANBERRA

- Final step is the model deployment and maintenance
- The model is now into operation
- You need to ensure that the model runs smoothly and will not make disastrous unsupervised decisions
- Usually, the deployment starts with small pilot version to test the performance in real scenarios
  - this may bring out issues that you did not anticipate,
  - then the model need to be adjusted accordingly



PROFESSIONAL ETHICS

- Some of the other professions have their own Oath or code of conduct
  - The most known one is the “*Hippocratic Oath*” for physicians
  - Some principles of that oath still hold for the nowadays data science work such as:
    - *“First, do no harm.”*
    - *“I will not be ashamed to say "I know not"*
    - *“I will respect the privacy of my patients”.*
    - *“I will remember that I remain a member of the society”*
  - Some other professions such as lawyers, journalists have their own code of conduct
  - We need a similar code for data scientists

PROFESSIONAL ETHICS (2)

- Regulations are not enough?
  - As the technology advances quickly and the regulations move slowly, it is hard to rely on the regulations. So we need to stand for our community as data scientists and agree up on our code of ethics.
    - e.g. automated vehicle

PROFESSIONAL ETHICS (3)

- The main value between the data scientist and the client is the **trust**
- The professional ethics should represent the special responsibilities not to take unfair advantage of this trust
- Example of breaking this trust is "*Truthful falsehoods*", which is about misinterpreting the results.

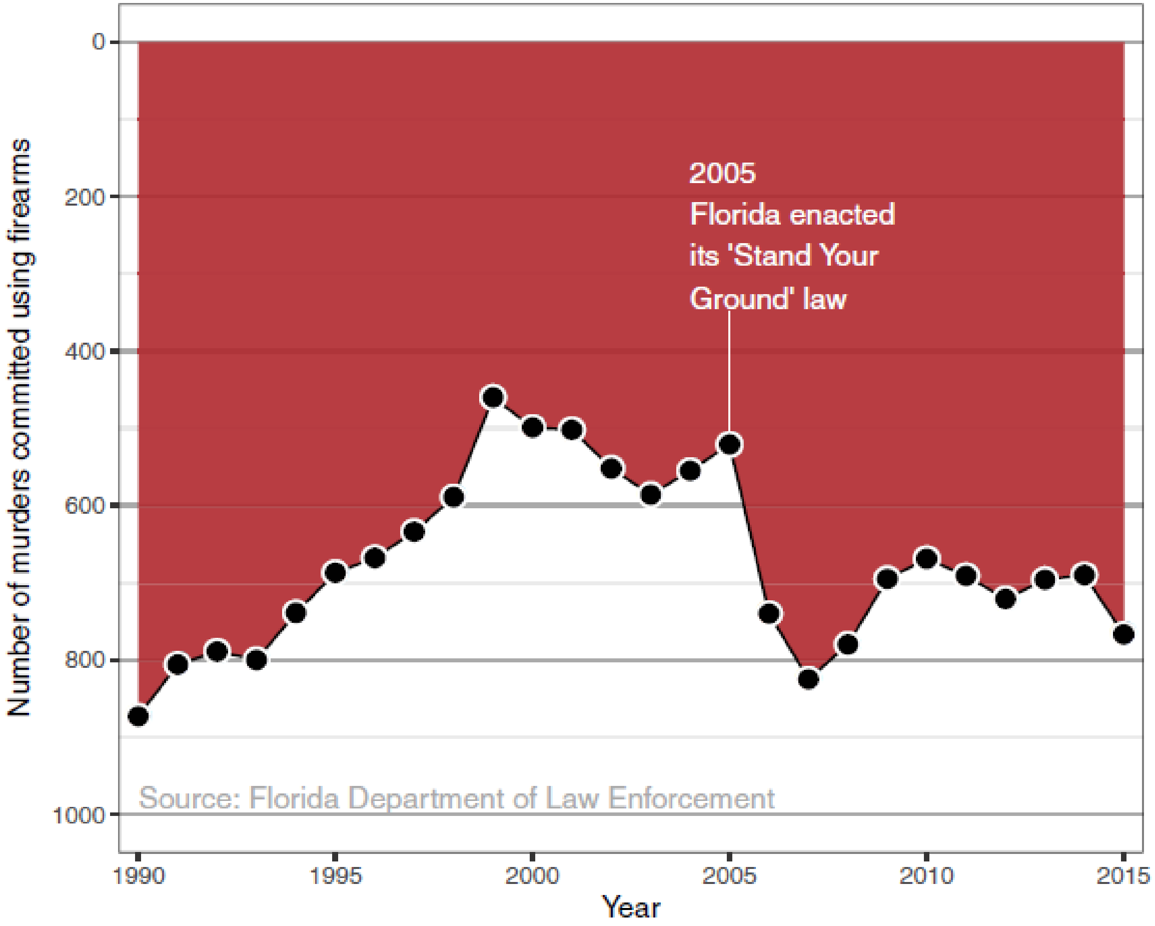


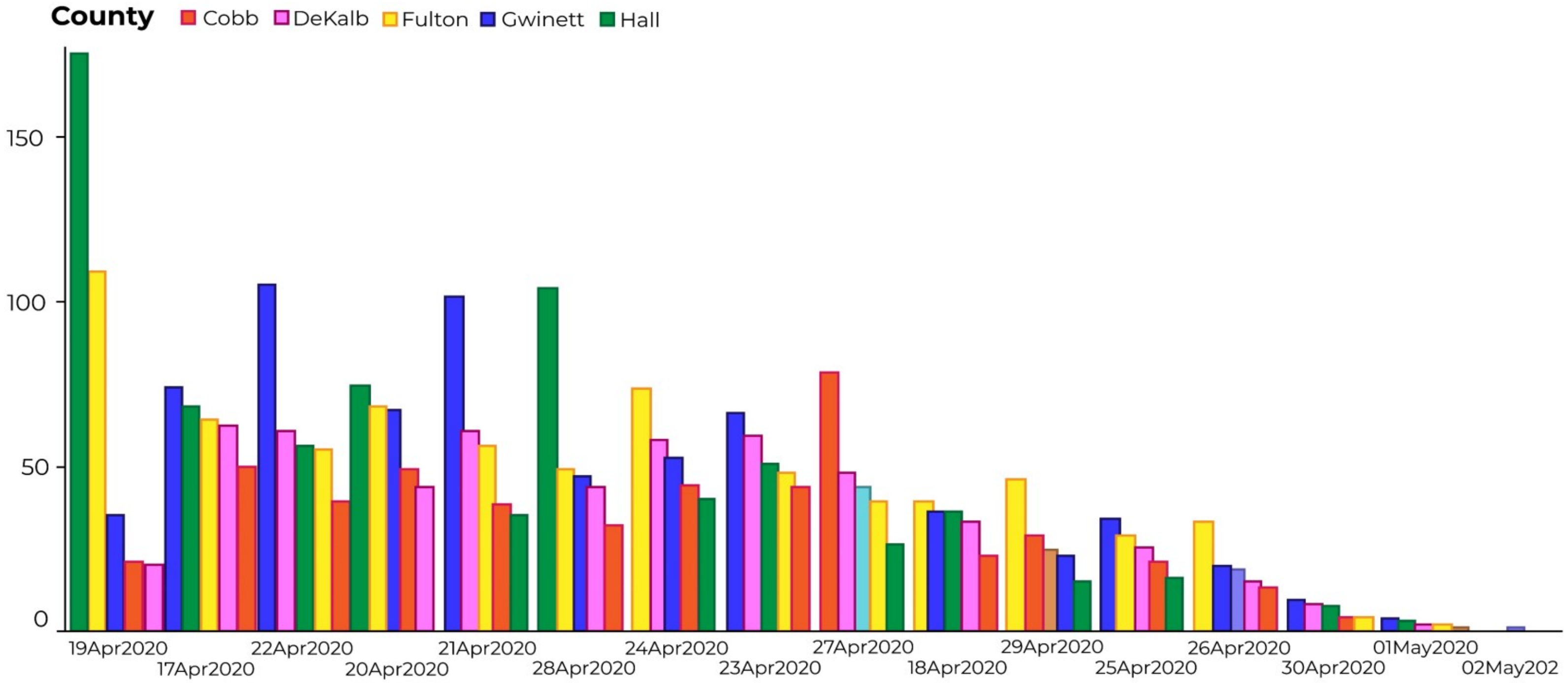
Figure 6.1: Reproduction of a data graphic reporting the number of gun deaths in Florida over time. The original image was published by Reuters.

Chapter 8, Modern Data Science 2nd Edition

PROFESSIONAL ETHICS (4)

Top 5 Counties with the Greatest Number of Confirmed COVID-19 ***(Reproduction of Figure)***

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



Chapter 8, Modern Data Science 2nd Edition

CODE OF ETHICS FOR DATA SCIENTISTS

- Data are everywhere and large volumes of data are collected every day
- Data science is meant to extract good things from the data, so it is all about doing what is good for the people and what is better for lives
  - Prediction of floods by analysing the satellite images
  - Preventing Suicide by understanding the causes from previous data and build programs to intervene when necessary
  - Different studies to look after the animals and the different species in our planet
- Data is an incredible tool for change, so we need to make sure that this change is what we all want to see

CODE OF ETHICS FOR DATA SCIENTISTS (2)

UNIVERSITY OF CANBERRA

- How ?
  - The code of ethics should represent the principles, values, and standards that govern our behaviour and actions
  - Community effort
  - We need to join in a global conversation about what the standards of dealing with the data should be.
    - https://youtu.be/i_bUa0BUg8Y
  - In the level of your organisation, get your team of data scientists at a meetup every now and then and start talking about what a Code of Ethics would look like.
  - We need to make our own values and standards for data collection and analysis.

CODE OF ETHICS FOR DATA SCIENTISTS (3)

- The following is an example of a "*code of conduct*", which has been put, after many discussions between data scientists, within the Data Science Association:
  - https://www.datascienceassn.org/code-of-conduct.html
- Mostly, organisations will be interested only in its minimum accountability to the law and to the regulations.
- Make your own standards and values when dealing with the clients (professional standards) and with the data you are collecting or analysing

CODE OF ETHICS FOR DATA SCIENTISTS (4)

- Examples of these values could be:
  - Produce truthful, interpreted results:
    - Is the data analysis valid?
    - Is the interpretation of the results fair and making sense?
    - What are the social consequences of your outcome?
  - Respect privacy of the data,
    - *“Don't do data mining on someone else's data if you don't want it done on your data.”*
    - Anonymise the data when possible
    - Don’t use the data beyond the goals of the project or for any personal use
  - What else?

WHY R FOR DATA SCIENTIST?

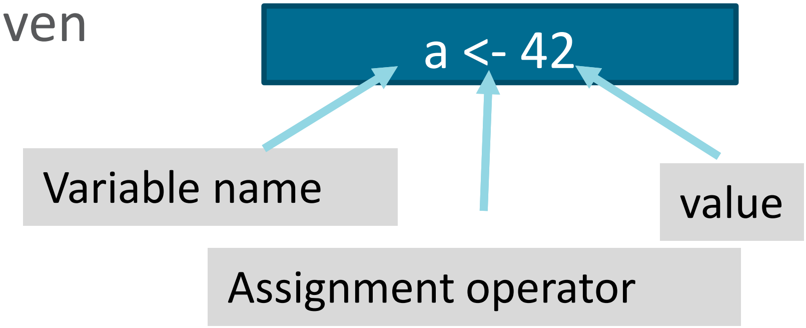
- Statistically comprehensive
  - R is the most comprehensive statistical analysis package as new technology and ideas often appear first in R
- Vector-based language
  - , so you can represent everything as vectors, which leads to accelerating the operations on the data
- It is a complete programming language,
  - so that you can build programs by using functions, objects, packages, etc.,
  - Also, you are not only using the installed packages and functions, but also can create your own scripts

WHY R FOR DATA SCIENTIST? (2)

- It is good for business
  - It is free and open source software, can be used without the need to pay for a licence
  - It is great in creating impressive visualisations
  - R is a cross-platform which runs on many operating systems, so can run anywhere at any time
- Easy Environment
  - R is open source for long time now so, you’re likely to be able to find support for any statistical analysis you need to perform
  - Easy to learn and use
    - With few lines, you can create complex statistical analysis and build impressive charts

VARIABLES

- Variable:
  - a placeholder, in a computer programs, which refers to a location in memory that stores a value. This value can be numbers (real and complex), words, matrices, and even tables
- R is Case sensitive
- Naming convention:
  - It contains letters, numbers, and only dot or underscore characters
  - Can’t start with a number (e.g.: 2norm)
  - Can’t start with a dot followed by a number (e.g.: .2norm)
  - Can’t start with an underscore (e.g.: _name)
  - Can’t be a reserved keyword



a <- 42

Variable name

Assignment operator

value

VARIABLES (2)

- Don’t use Reserve keywords as a variable name

for	in	repeat	while	function
if	else	next	break	TRUE
FALSE	NULL	Inf	NaN	NA
NA_integer_	NA_real_	NA_complex_	NA_character_	…

myName
my_Name
my.Name
.myName

✓

2myName
.2myName
function
_myName

✗

VARIABLES (3)

- Valid variables? Which is invild?
  - total, Sum, .fine.with.dot, [✓ 5sum], this_is_acceptable
- Invalid variables? Which is valid
  - tot@l, _fine, TRUE, .0ne, [✓ Number5]

➢ Notes:

- R is a dynamic language, where you don’t need to specify specific types for the variables. It will be defined implicitly
- Naming guide:
  - Common one is Google R style guide
    - https://google.github.io/styleguide/Rguide.xml

VARIABLES (4)

- Assign variables
  - Assignment operator
    - match.score <- 300
  - Assign function
    - assign("match.score", 300)

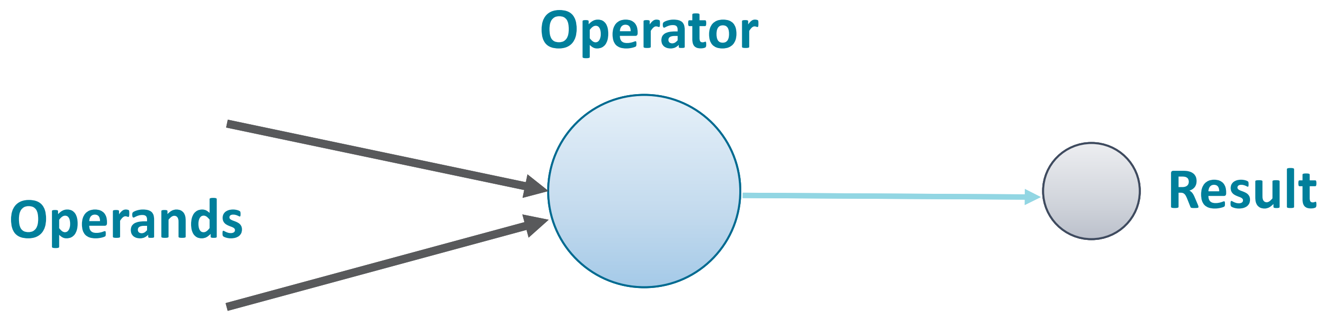
DATA TYPES

- Basic data types:
  - Numeric data types
    - Decimal, x <- 4.5
    - Integers, y <- as.integer(5)
    - By default the type of the natural numbers is “Numeric”
  - Character, with single or double quotes
    - name <- “John”, name <- ‘Peter’
  - Logical, for Boolean values (TRUE, FALSE, NA)
    - Check <- TRUE
  - To get the type of a variable use:
    - typeof(variable_name)
    - class(variable_name)

DATA TYPES (2)

- Data structures:
  - Vectors
  - Lists
  - Matrix
  - Data frame
  - Factors
- These structures will be covered next lecture, so stay tuned!

OPERATORS



Arithmetic operators

Logical operators

MATHEMATICAL OPERATORS

- Addition, +
- Subtraction, -
- Multiplication, *
- Division, /
- Exponentiation, ^ or **
- Modulus, %%
- Integer Division, %/%

```
# Example for R Arithmetic Operators
a <- 10
b <- 5
add <- a + b
print(paste("Addition is:", add))

sub <- a - b
print(paste("Subtracting is:", sub))

multi <- a * b
print(paste("Multiplication is:", multi))

division <- a / b
print(paste("Division is", division))

Integer_Division <- a %/% b
print(paste("Integer Division is:", Integer_Division))

exponent = a ^ b
print(paste("Exponent is:", exponent))
format(exponent, scientific=FALSE)
modulus = a %% b

print(paste("Modulus is:", modulus))
```

MATHEMATICAL OPERATORS (2)

- Built-in mathematical functions, such as:
  - abs(-4) # absolute value → 4
  - log(3) # natural logarithm → 1.098612
  - log(3, base= 10) # change logarithm base to 10 → 0.4771213
  - exp(5) # exponential → 148.4132
  - pi # PI constant value → 3.141593

Why six digits after the decimal point? Can we change it?

SPECIAL NUMBERS

- Inf & -Inf
  - Positive and negative infinity
  - Very big or very small number that the computers can’t handle
- NaN
  - Not a Number
  - No mathematical sense
- NA
  - Not available
  - Missing value
- These special numbers let the program to continue without crashing when an operation results in such values

```
# Infinity
1 / 0 # positive infinity
-1 / 0 # negative infinity
Inf + 5 # operation on Inf
is.finite(1 / 0) # check if finite
is.infinite(1 / 0) # check if infinite

# Undefined
Inf / Inf # NaN (not a number)
is.nan(Inf/Inf) # Check if NaN

# Missing values
NA + 5 # operation on NA
is.na(NA) # check is NA

# NaN is NA, but NA is not NaN
is.na(NaN)
is.nan(NA)
```

LOGICAL OPERATORS

- Results in TRUE or FALSE
  - Greater than, >
  - Greater than or equal to, >=
  - Less than, <
  - Less than or equal to, <=
  - Equal to, ==
  - Not equal to, !=
  - Logical NOT, !
  - Logical OR, |
  - Logical AND, &

```
# Logical operators

5 > 3 # greater than
5 >= 3 # greater than or equal
5 < 3 # less than
5 <= 3 # less than or equal
5 == 3 # equal
5 != 3 # not equal

"B" == "a" # also for characters, depends on coding

!(TRUE) # NOT
TRUE | FALSE # OR
TRUE & FALSE # AND
```

RECOMMENDED READINGS

- Chapter 8 from “**Modern Data Science with R, 2nd Edition”** by Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton”
  - *https://mdsr-book.github.io/mdsr2e/ch-ethics.html*

- *For more about atomic data types in R, please watch this:*
  - https://www.linkedin.com/learning/r-for-data-science-lunchbreak-lessons/r-data-types-basic-types?u=2330002