Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF
CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 1

Dr. Ibrahim Radwan

DISTINCTIVE BY DESIGN

ACKNOWLEDGEMENT OF COUNTRY

- I would like to acknowledge the Ngunnawal people who are the Traditional Custodians of the Land that we are on.
- I would also like to pay respect to the Elders both past and present of the Ngunnawal Nation and extend that respect to other Indigenous Australians who are present.

UNIT OVERVIEW

- This course is about **understanding** different concepts of Data Science
- Also, you will be **acquiring** practical skills of how to perform data analysis and modelling
- We will use **R language** in our practical work
- There is no specific Prerequisites or Co-requisites of this unit
- This is neither a machine learning unit nor a programming unit. However, both will be considered and touched, while going through the unit

WHAT WILL BE COVERED IN IDS/IDS G?

- What Data Science is
- The life cycle of a typical Data Science project
- Data Preparation and Wrangling
- Data Exploration and Visualization
- Basics of the statistical modelling
- Basics of supervised and unsupervised modelling
- How to present and document your final findings
- All the above with practices in R language, so basics of R language will be covered too

WHO IS YOUR LECTURER?

UNIVERSITY OF CANBERRA

- Assistant Professor of Machine Learning and Artificial Intelligence
- Between 2017 – 2019, I had worked as a research fellow with the Australian National University.
- Between 2014 – 2016, I had worked as a computer vision and machine learning researcher in the Algorithm group, at Seeing Machines Ltd.
- I have done my PhD (2011 - 2014) at University of Canberra in Computer Vision
- Before that I got a Master and a Bachelor of Computer Science at Zagazig University, in Egypt
- I am a father of three kids

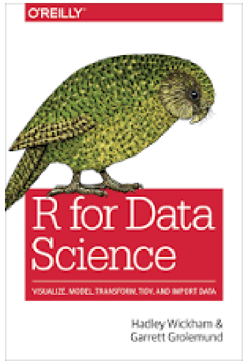
UNIT OUTLINE

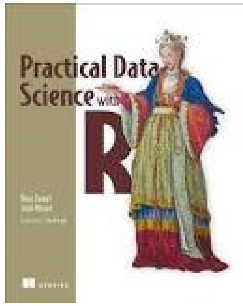
UNIVERSITY OF CANBERRA

- Very important document
- Represents the “contract” between you and us
- Describes what is being covered in this unit (*i.e. unit contents*)
- Describes how you will be assessed
  - Assignments
  - Quizzes
  - Assessment criteria
  - Minimum requirements to pass the unit
- Available on the unit website on UC Learn
  - https://unicanberra.instructure.com/courses/9040
  - Please read it carefully and let us know if anything is not clear

LECTURES AND MATERIALS

- Lecture notes and the recording of the lectures will be made available on the IDS UC Learn website (aka Canvas).
  - The lecture notes and slides will be posted before the lecture every week
- Additional reading
  - Recommended:
    - "R for Data Science" by *Garrett Grolemund* and *Hadley Wickham* (https://r4ds.had.co.nz/)
    - "Practical Data Science with R Second Edition" by Nina Zumel and John Mount (available as a hard copy at the UC library)
  - Supplementary:
    - "Modern Data Science with R" by Benjamin S. Baumer and Daniel T. Kaplan
    - "Introduction to Data Science, Data Analysis and Prediction Algorithms with R" by *Rafael A. Irizarry,* https://leanpub.com/datasciencebook





LECTURES AND MATERIALS (2)

- **USB Drives**: You are recommended to have two USB sticks / thumb drives / disks to store and to backup your work.

HANDOUTS

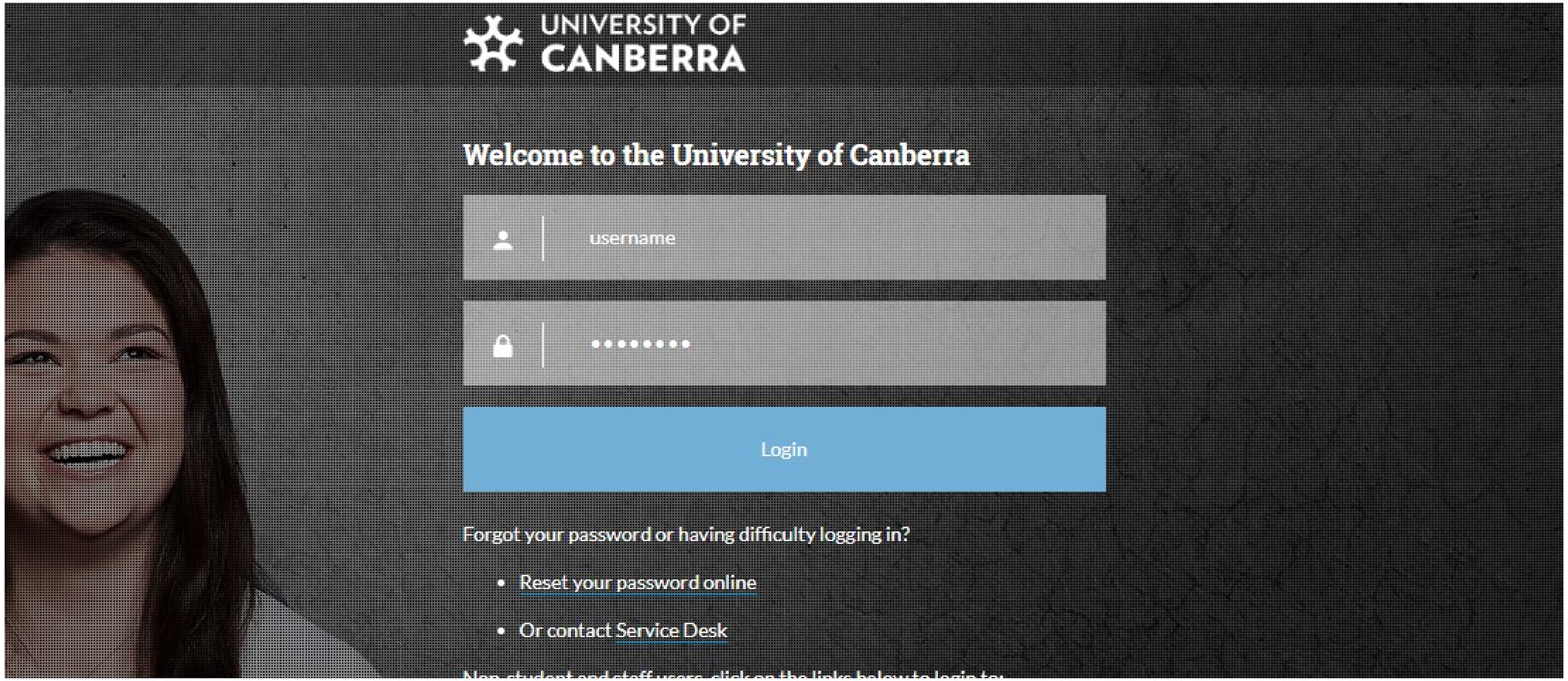
- There will be **no** handouts in this unit, only the lecture slides and the lab notes.
- Let’s save some trees, everything will pretty much be available as a PDF on the UC Canvas site
- Download and print a copy from the UC Learn website, if you really need a hardcopy

COURSE LOGISTICS AND TEAM

- Unit Convener
  - Dr. Ibrahim Radwan
  - Rm: 6C45, Tel: (02) 6201 5538
  - ibrahim.radwan@canberra.edu.au
- Moderator
  - Dr. Shuangzhe Liu,
  - Rm: 6C34, Tel: (02) 6201 2513
  - Shuangzhe.Liu@canberra.edu.au
- Tutors
  - Dr. Karam Sallam
- Unit website:
  - https://unicanberra.instructure.com/courses/9811

UC LEARN ONLINE (AKA CANVAS)

- https://unicanberra.instructure.com/courses/9811
- All material –lecture notes, lab notes and any other material –is available on the unit web site on UC Learn (via the MyUClogin)



Welcome to the University of Canberra

Forgot your password or having difficulty logging in?

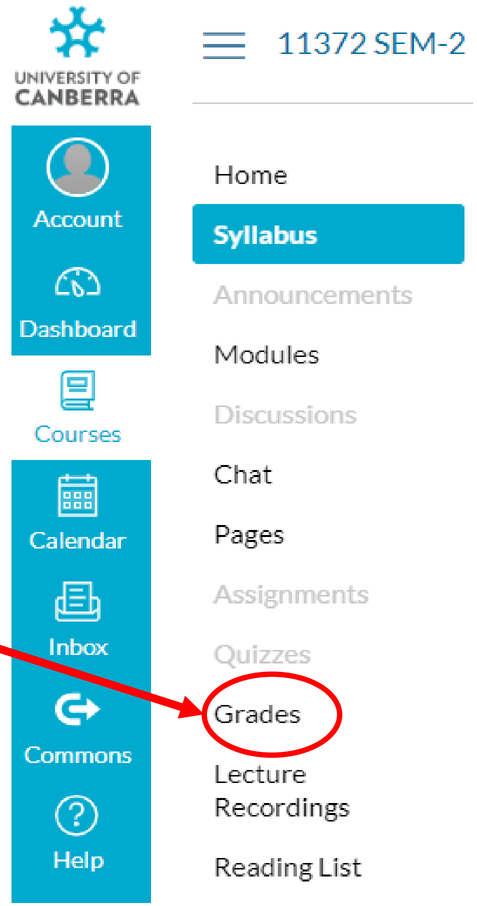
- Reset your password online
- Or contact Service Desk

Non-student and staff users, click on the links below to login to:

CHECKING YOUR PROGRESS

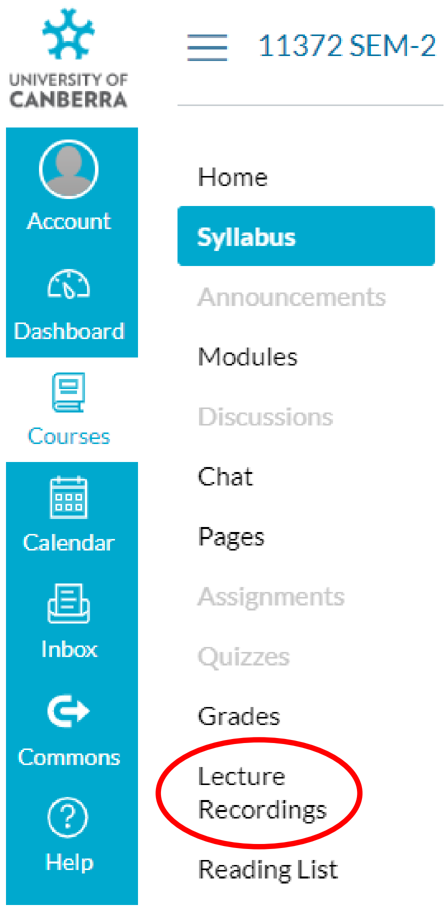
- As you complete assignments and tests, your test marks and assignment status are recorded on the IDS website on UC Learn. This is called ***Grades*** in Canvas.

- All of the information about your performance is recorded there, so please if you think that any of the information is incorrect, discuss it with:
  - Your tutor, in the first instance, or
  - Your lecturer, if the tutor was unable to help, as soon as possible.



LECTURES

- Will cover the theory and code examples using R language
- Lectures will be recorded, you can access the recordings via ‘Lecture Recordings’ on the UC Learn site (after the lecture’s time)
  - Tue (15:30–17:20) in (virtual room)
  - Start in Week 1, 9th of February
- If you foresee any problem in being able to watch the lectures and/or having any problem to access the recordings, please send me an email
  - The later you leave it, the less likely it is that a satisfactory solution can be found, so in your own interest, discuss it early!



TUTORIALS AND COMPUTER LABS

- Practical work, hands-on training
- Working on problems using R language
- A combined 2 hours (1h tutorial & computer lab + 1h unsupervised activities)
- Start in week 2
- UG groups
  - Wed (15:30 - 16:30) in (1C33)
  - Fri (10:30 - 11:30) in (virtual room)
- G groups:
  - Wed (11:30 - 12:30) in (6B4)
  - Fri (15:30 - 16:30) in (virtual room)

PRACTICAL WORK

- To be performed using R-Studio as a user interface for R language on your own PC/laptop
  - R is a programing language that can be used to data analytics and statistical computing.
  - R-studio is an open-source and a cross-platform tool
  - R-Studio is an IDE that can be used for writing and running R programs
  - R and R-Studio can be installed on Windows, Linux and Mac-OS
  - You may use alternative IDE on your laptop, such as PyCharm or Spyder

PRACTICAL WORK (2)

UNIVERSITY OF CANBERRA

- Students are encouraged to use USBs to backup their work
- You will use your student ID to access any computer in the university’s campus
- Access to the buildings after 6.30 pm may require swipe card access

PLAGIARISM (CHEATING)

UNIVERSITY OF CANBERRA

- Handing in work that is not your own, or very similar to another student's work
- Treated very seriously by university (*Student Conduct), i.e.* can result in Fail
  - https://www.canberra.edu.au/current-students/queensland-students/student-conduct
- Won't help as won't prepare you for final assessment
- Lecturer / Tutor can orally question students on their submitted work
- Avoid this by starting early and seeking help
- Seek help from tutor, student resource centre, lecturer, etc.

ASSESSMENT DETAILS

- **Unit Readiness Test**:
  - This test is an early assessment item, which you will be required to complete several associated tasks and a short quiz.
  - **Due Date:** 23:59 Sunday, Week 4
  - **Weighting Percentage:** 10%
- **Assignment 1**:
  - In this assignment you will be developing and changing programs in R language toward the learned skills of data collection, data wrangling and data exploration.
  - You will be given a template for the assignment and you need to submit a working version of this template as well as output of the visualization part/s as a single ZIP file
  - **Due Date:** 23:59, Sunday, Week 10
  - **Weighting Percentage:** 30%

ASSESSMENT DETAILS (2)

- **Week 9 online test**
  - 1-hour test via UC learn online site
  - **Due Date:** 23:55, Sunday, Week 9
  - **Weighting Percentage:** 20%

- **Final Assignment** :
  - In this assignment you will be developing and changing programs in R language toward the learned skills through the whole unit.
  - **Due Date:** Week 14-15
  - **Weighting Percentage:** 40%

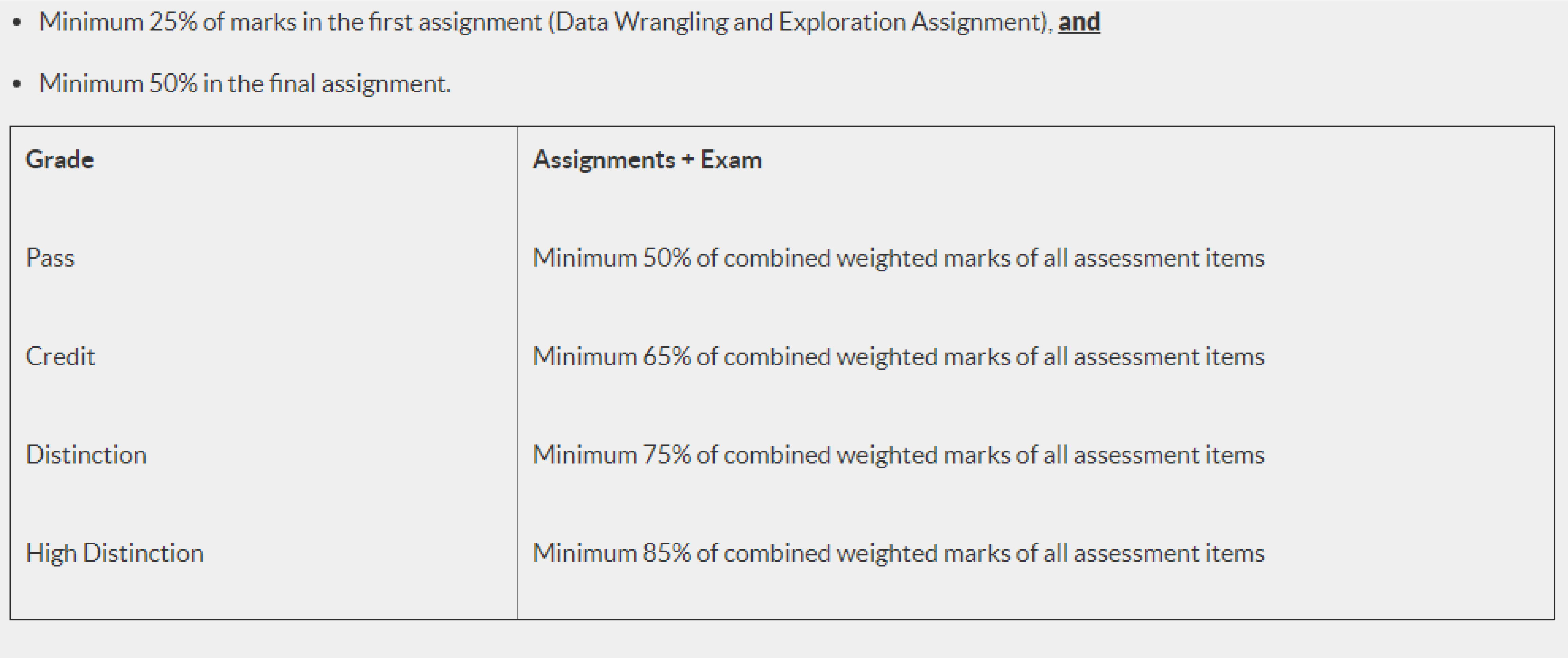
ASSESSMENT DETAILS (3)

- All assessment items will receive a numerical mark. The final grade will be a weighted average of the individual assessment items

- You need to achieve a minimum 25% of the marks in first assignment and a minimum 50% in the final assignment to pass this unit
- A delay in submitting any of the assessment items will be penalised with 5% every for 24 hours, until 7 days, unless you have an approved extension request.

ASSESSMENT DETAILS (4)

- **To be awarded a particular grade in IDS/IDS G, you must meet both the assignment requirements as in the table below:**



- Minimum 25% of marks in the first assignment (Data Wrangling and Exploration Assignment), **and**
- Minimum 50% in the final assignment.

Grade	**Assignments + Exam**
Pass	Minimum 50% of combined weighted marks of all assessment items
Credit	Minimum 65% of combined weighted marks of all assessment items
Distinction	Minimum 75% of combined weighted marks of all assessment items
High Distinction	Minimum 85% of combined weighted marks of all assessment items

OPPORTUNITIES FOR LEARNING

- **Lectures**
  - Pay attention
  - Annotate your notes
  - Ask questions

- **Tutorials and Labs**
  - Come prepared – read the tutorial work before the tutorial and attempt any practical work. The tutorial and lab classes are for discussion and to resolve difficulties. Actively participate!
  - Do the R exercises in the tutorial and computer lab notes
  - If you expect to magically pick up the 'crumbs of wisdom' from just attending the classes, be warned that this will not be sufficient to pass the unit

OPPORTUNITIES FOR LEARNING (2)

- **Assignments**
  - Start early
  - Do your own work
  - Seek help and do so early

- **Textbooks**
  - We have <u>recommended</u> readings every lecture

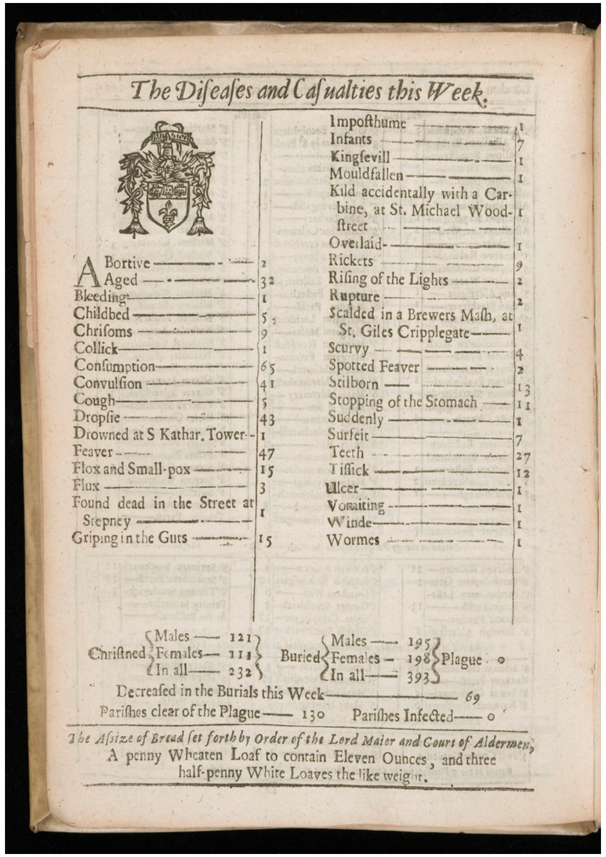
CONSULTATION TIME

- Tue (12:30 – 13:30) in (6C45 and/or remotely through virtual room)
- Other times by prior appointment only!
  - Best, arrange a time with me by email
  - Ibrahim.Radwan@Canberra.edu.au
- Make use of the **online discussion forum** on the UC Canvas site!
  - It's a much faster way of getting in contact and also has the benefit of other students, who might have a similar questions, being able to see the answer as well

WHAT IS DATA SCIENCE?

UNIVERSITY OF CANBERRA

- Is Data Science new?
  - No, people used to use different methods to extract information from data, (e.g. “bills of mortality” in 1592). These bills were weekly statistical summary of the mortalities in London.
- … so, is Data Science when you deal with large amount of data “*Big Data*”?
  - No, Data Science can deal with small and large data.
- Does Data Science only exist coupled with Machine Learning?
  - No, there is also statistical and mathematical modelling.
- … so, what is Data science?



The Diſeaſes and Caſualties this Week.

Abortive	2
Aged	32
Bleeding	1
Childbed	5
Chriſoms	9
Collick	1
Conſumption	65
Convulſion	41
Cough	5
Dropſie	43
Drowned at S Kathar. Tower	1
Feaver	47
Flox and Small-pox	15
Flux	3
Found dead in the Street at Stepney	1
Griping in the Guts	15

Impoſthume	1
Infants	7
Kingſevill	1
Mouldfallen	1
Kild accidentally with a Carbine, at St. Michael Woodſtreet	1
Overlaid	1
Rickets	9
Riſing of the Lights	2
Rupture	2
Scalded in a Brewers Maſh, at St. Giles Cripplegate	1
Scurvy	4
Spotted Feaver	2
Stilborn	13
Stopping of the Stomach	11
Suddenly	1
Surfeit	7
Teeth	27
Tiſſick	12
Ulcer	1
Vomiting	1
Winde	1
Wormes	1

Chriſtned	Males	121	Buried	Males	195	Plague 0
	Females	111		Females	198	
	In all	232		In all	393	

Decreaſed in the Burials this Week — 69

Pariſhes clear of the Plague — 130 Pariſhes Infected — 0

The Aſsize of Bread ſet forth by Order of the Lord Maier and Court of Aldermen,
A penny Wheaten Loaf to contain Eleven Ounces, and three
half-penny White Loaves the like weight.

Bills of Mortality Feb 21 - 28 1664.
Credit: Wellcome Collection.

WHAT IS DATA SCIENCE? (2)

- Before we jump into the definition of the Data Science, let us understand the difference between:
- "**Data**": is a raw, unstructured pieces or records that are needed to be processed to have a meaning,
- and, "**Information**" is where these data are processed and presented in a given context so as to make it useful.

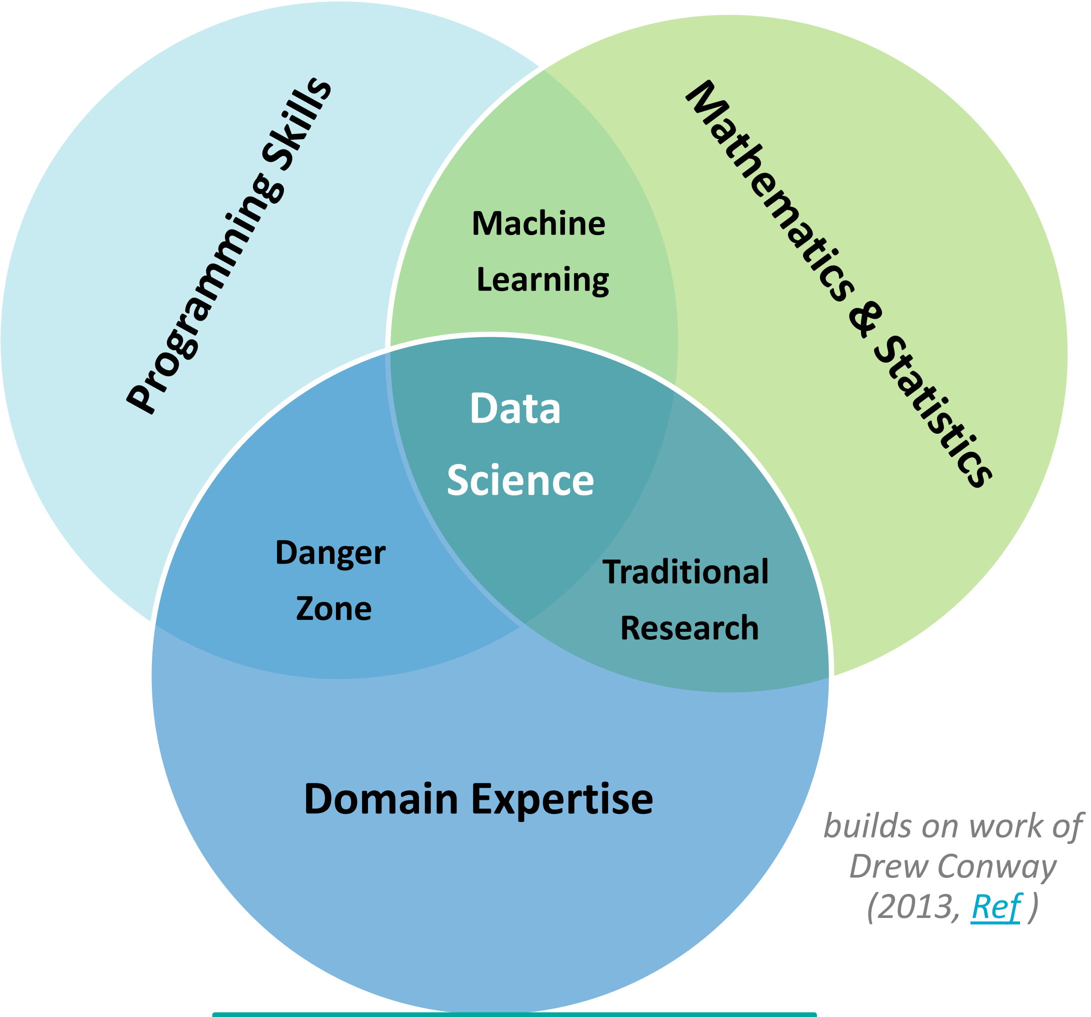
- Data Science is the science that encapsulates set of components that can be used to turn raw data into actionable insights.

WHAT IS DATA SCIENCE? (3)

- What is the Data Science?
- Data Science Venn diagram
- Why is Data Science so important?

Data

Information



builds on work of Drew Conway (2013, Ref)

Data Science Venn diagram

WHAT IS DATA SCIENCE? (4)

Computer science basics, such as algorithms, programming, etc.

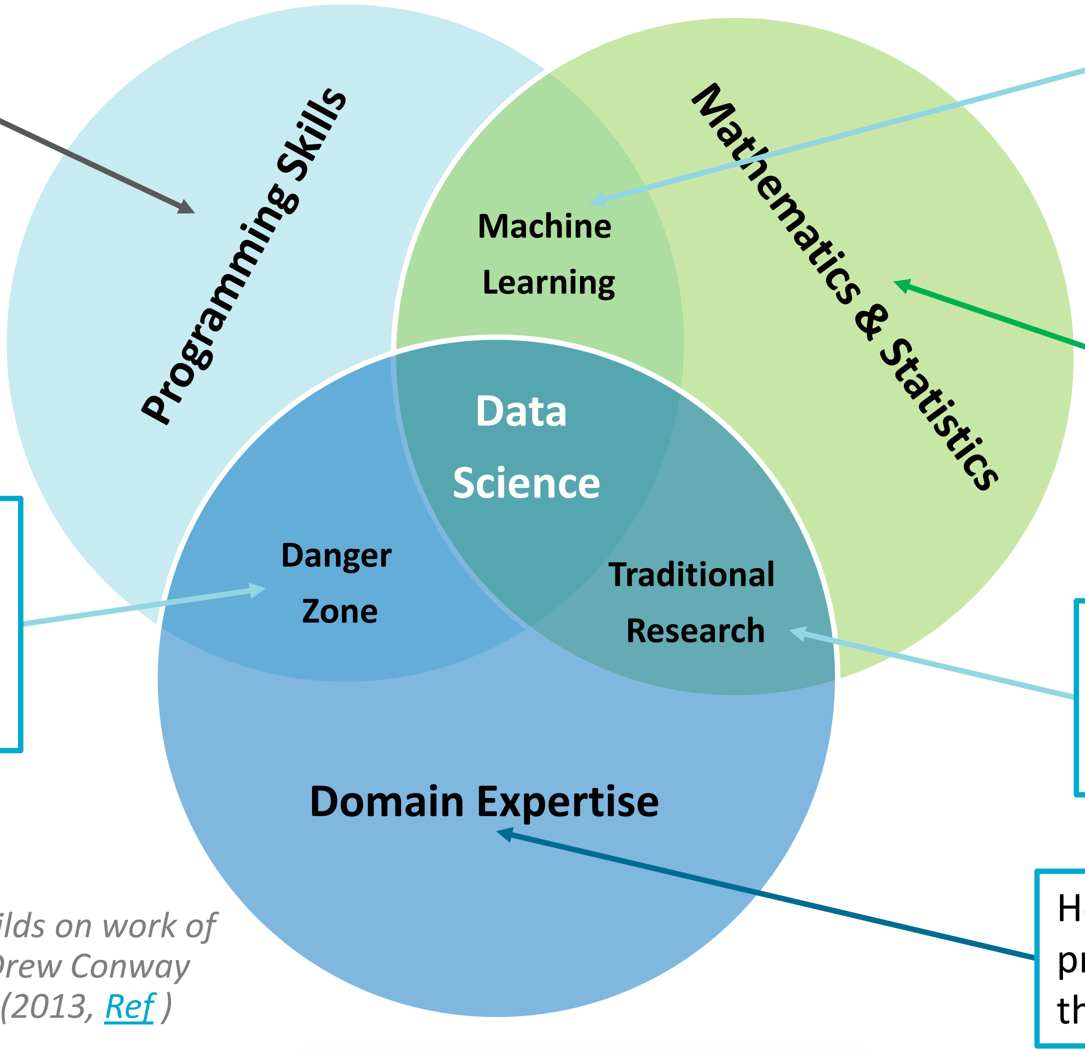
Build models to predict, score or recommend output some values

Distributions of the variables, different modelling strategies, etc.

When we don't know why it works or why it is not. Lack of mathematical and statistical interpretation

Understanding the data records and the problem domain

Helps you to understand the problem and the meaning of the data records



builds on work of Drew Conway (2013, Ref)

Data Science Venn diagram

WHY DATA SCIENCE?

- A simple answer could be:
  - We have lots of data and what is really needed is the information,
  - because there is more data available than ever before, more data than we could understand.
  - Therefore, we need data science to give us the ability to extract the insights and the important information from these unstructured data
- Is not the statistics or the machine learning enough?
  - Statistics or machine learning are quite useful in building models on the data, but firstly, you need to understand that data
  - Combining the statistics and machine learning with understanding of the problem and data results in Data Science
- Video about what and why data science

BEING A DATA SCIENTIST

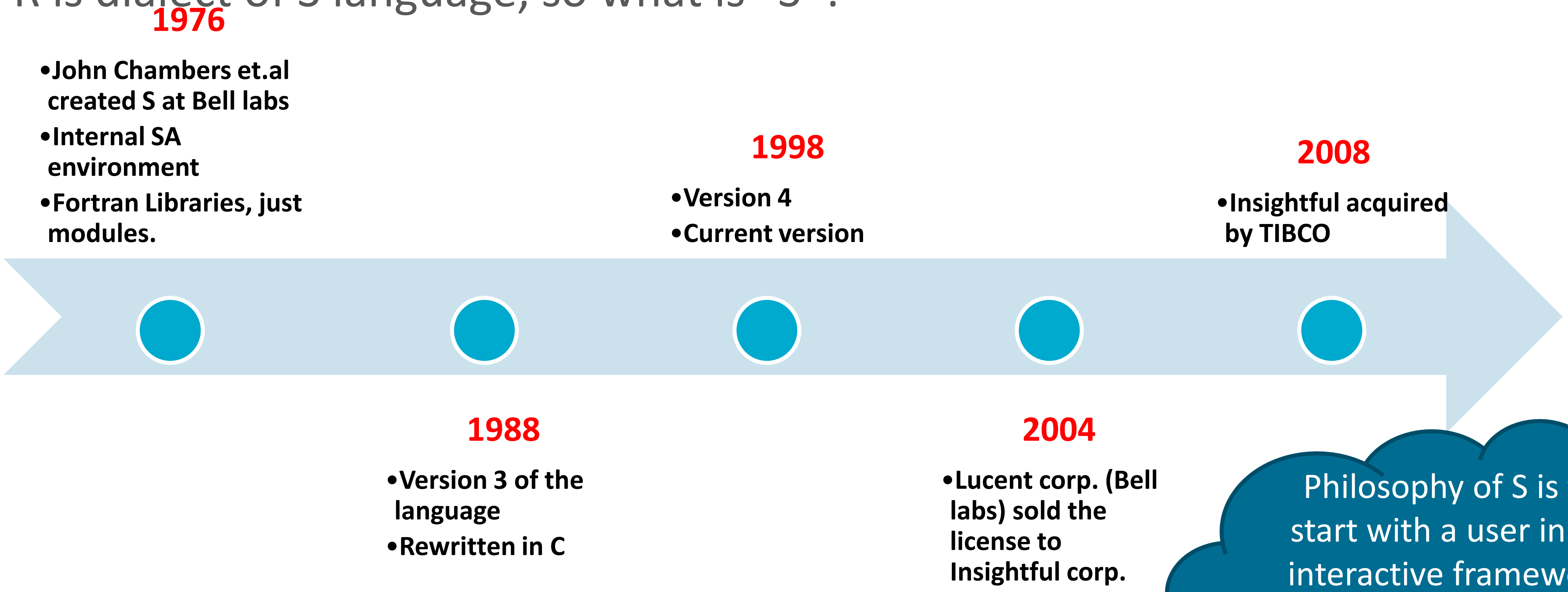
- What does it mean to be a data scientist?
  - For every new problem, you may go into a new field to try to understand how it works, to massage the data until you understand them, to try to acquire all the knowledge of the field without being a specialist. (a detective job? ☺)
  - Be updated all the time with the new tools, methods and workflows, as it is rapidly evolving, you do not need to fall behind
  - As a data scientist, you need to try different ideas using scientific methods to achieve critical tasks
  - Being ready to understand, analyse and process messy data structures and build efficient models out of these data to predict valuable information, is not it exciting? ☺

PROGRAMMING IN R

- R is a programming language and a free software environment for statistical computing and graphics.
- R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS
- https://www.r-project.org/
- A bit about the history of it and the connection with S

OVERVIEW & HISTORY OF R

- R is dialect of S language, so what is “S”?



1976

- **John Chambers et.al created S at Bell labs**
- **Internal SA environment**
- **Fortran Libraries, just modules.**

1988

- **Version 3 of the language**
- **Rewritten in C**

1998

- **Version 4**
- **Current version**

2004

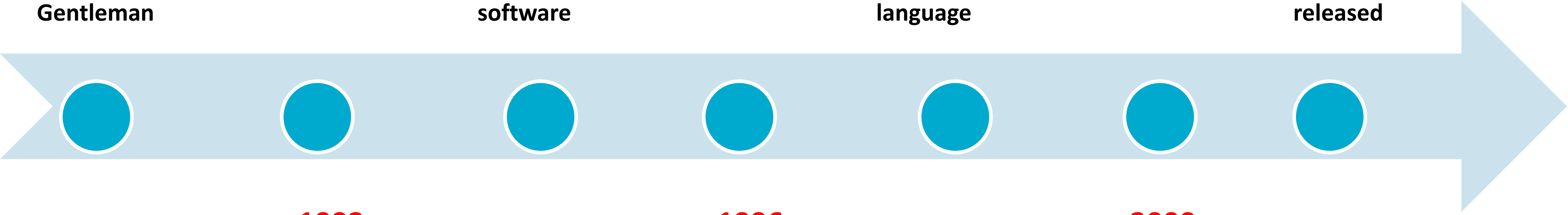
- **Lucent corp. (Bell labs) sold the license to Insightful corp.**

2008

- **Insightful acquired by TIBCO**

Philosophy of S is to start with a user in an interactive framework and slide him gradually to the programming space

OVERVIEW & HISTORY OF R (2)



1991

- Created @ Auckland university by Ross Ihaka and Robert Gentleman

1993

- First Appear to the public as alternative to S

1995

- Agreed to be released under GNU public license as free software

1996

- Public mailing list (r-help & r-devel) are created

1997

- R core group has been created to look after the source of the R language

2000

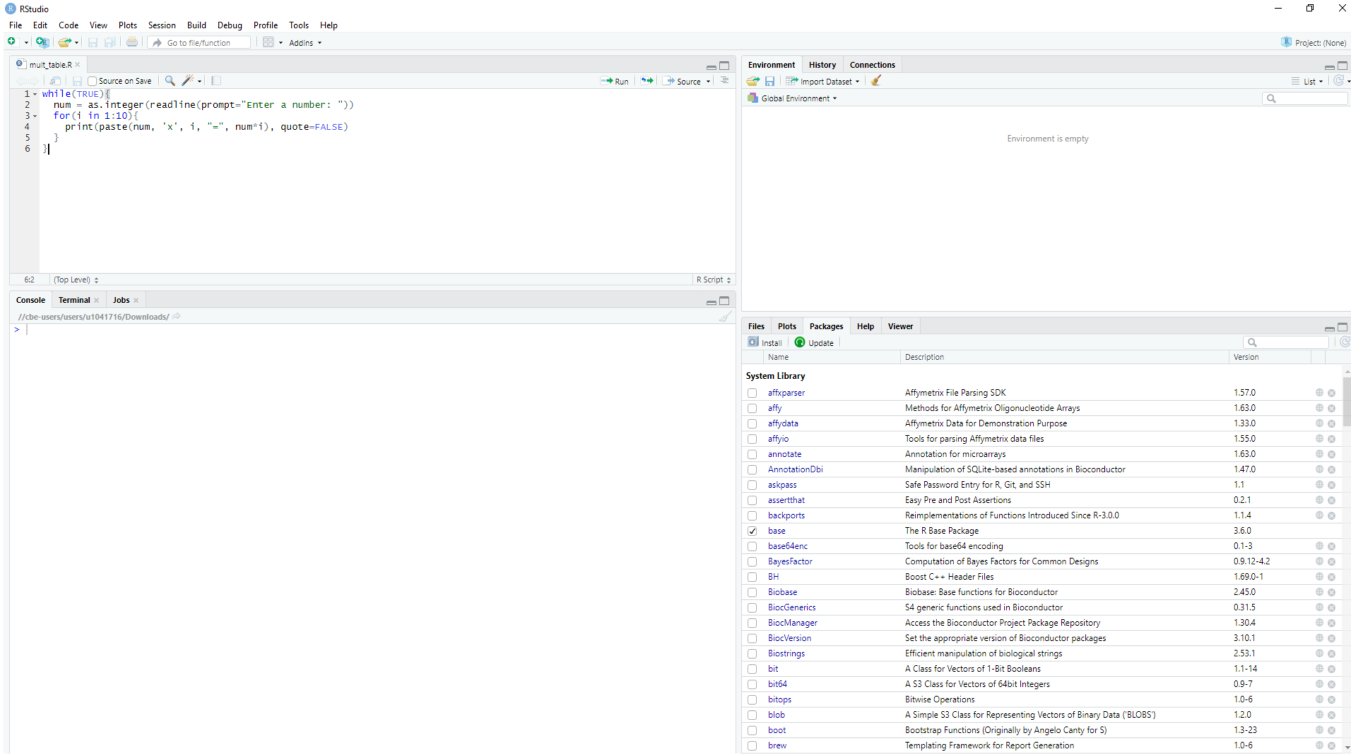
- Version 1.0.0 of R is released

2013

- Version 3.0.2 is released

SETUP YOUR ENVIRONMENT

- Two main components
  - R (https://cran.r-project.org/mirrors.html)
  - Graphical user interface, GUI, such as R-Studio (https://www.rstudio.com/)



```
while(TRUE){
  num = as.integer(readline(prompt="Enter a number: "))
  for(i in 1:10){
    print(paste(num, 'x', i, "=", num*i), quote=FALSE)
  }
}
```

```
//cbe-users/users/u1041716/Downloads/
>
```

Environment is empty

System Library

	Name	Description	Version
□	affxparser	Affymetrix File Parsing SDK	1.57.0
□	affy	Methods for Affymetrix Oligonucleotide Arrays	1.63.0
□	affydata	Affymetrix Data for Demonstration Purpose	1.33.0
□	affyio	Tools for parsing Affymetrix data files	1.55.0
□	annotate	Annotation for microarrays	1.63.0
□	AnnotationDbi	Manipulation of SQLite-based annotations in Bioconductor	1.47.0
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.4
☑	base	The R Base Package	3.6.0
□	base64enc	Tools for base64 encoding	0.1-3
□	BayesFactor	Computation of Bayes Factors for Common Designs	0.9.12-4.2
□	BH	Boost C++ Header Files	1.69.0-1
□	Biobase	Biobase: Base functions for Bioconductor	2.45.0
□	BiocGenerics	S4 generic functions used in Bioconductor	0.31.5
□	BiocManager	Access the Bioconductor Project Package Repository	1.30.4
□	BiocVersion	Set the appropriate version of Bioconductor packages	3.10.1
□	Biostrings	Efficient manipulation of biological strings	2.53.1
□	bit	A Class for Vectors of 1-Bit Booleans	1.1-14
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
□	bitops	Bitwise Operations	1.0-6
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.0
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-23
□	brew	Templating Framework for Report Generation	1.0-6

SETUP YOUR ENVIRONMENT (2)

- **To Install R**
  - Open an internet browser and search for CRAN (**C**omprehensive **R A**rchive **N**etwork) and go to the https://cran.r-project.org
  - Select the Download link that is suitable for your operating system
  - Click on the file containing the latest version of R
  - Save the .pkg file (MAC OS) or .exe file (Windows OS)
  - Double-click it to open, and follow the installation instructions
  - Now that R is installed, you need to download and install RStudio

SETUP YOUR ENVIRONMENT (3)

- **To Install RStudio**
  - Go to www.rstudio.com and click on the "Download RStudio" button.
  - Click on "Download RStudio Desktop"
  - Click on the version recommended for your operating system, or the latest for Mac OS or Windows OS,
  - For MAC OS,
    - save the .dmg file on your computer, double-click it to open, and then drag and drop it to your applications folder.
  - For Windows OS,
    - save the executable file. Run the .exe file and follow the installation instructions.
    - leave all default settings in the installation options.

KEY TAKEAWAYS

UNIVERSITY OF CANBERRA

- Data Science encapsulates different discipline to extract insights from messy data
- A data science project involves going back and forth between different stages until reaching successful results
- Data science is not just statistics, it is more about understanding the interests and the needs of the stakeholders
- As a data scientist you need to be equipped with the passion to understand different problems from different domains
- Before jumping into solving a problem or an issue, you need to ensure that you have a clear and quantifiable goals
- Data science is quite exciting if you are passionate about discovering root causes of the issues and ready to solve them

RECOMMENDED READINGS

UNIVERSITY OF CANBERRA

- Chapter 1 from “Practical Data Science with R Second Edition” by Nina Zumel and John Mount. A hard copy of the book exists in the library
  - https://livebook.manning.com/book/practical-data-science-with-r-second-edition/chapter-1/

- Warm-up with R and R-Studio:
  - How to install R and R-Studio?
  - Navigating the R-Studio Environment?