

Introduction

Introductory Data Science for Innovation (995N1)

Week 1 – 27 September 2021

Frédérique Bone & Daniele Rotolo



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT

Did you get to install RStudio or connect to RStudio cloud?



Make sure you also downloaded the R script available in Canvas.

What do you expect to learn in this module?

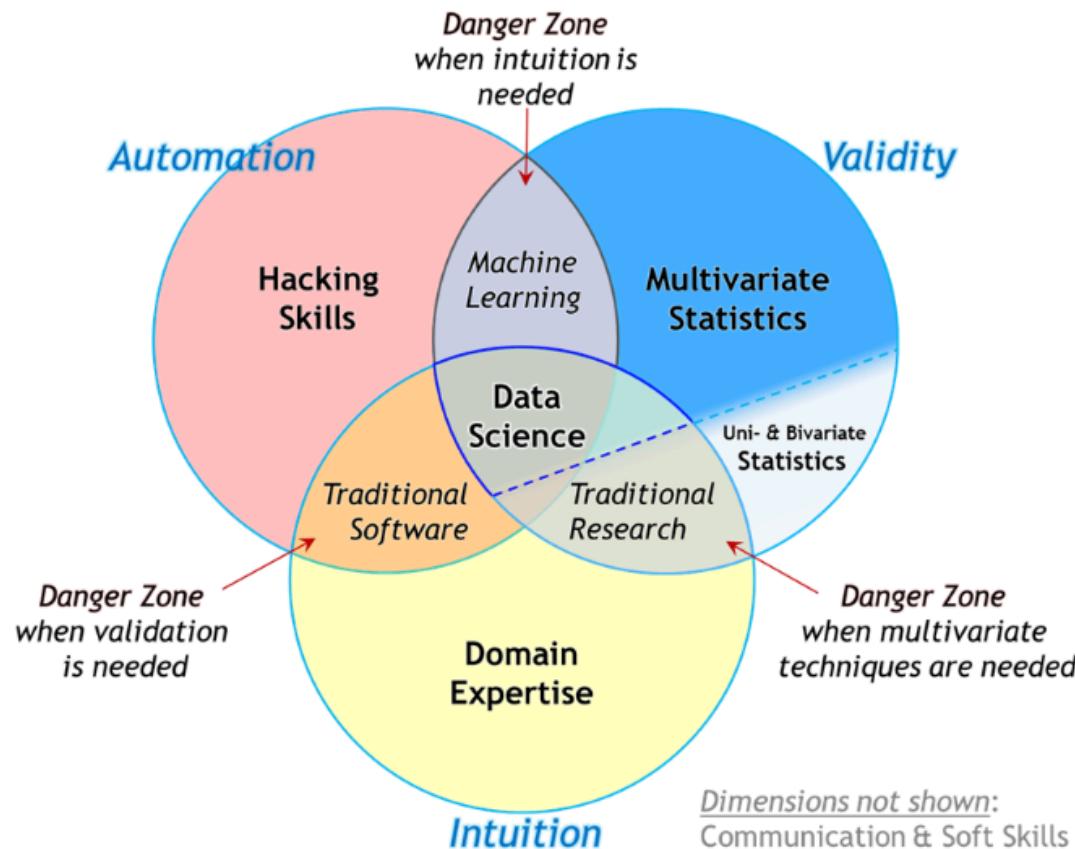
What is Data Science?



BUSINESS
SCHOOL

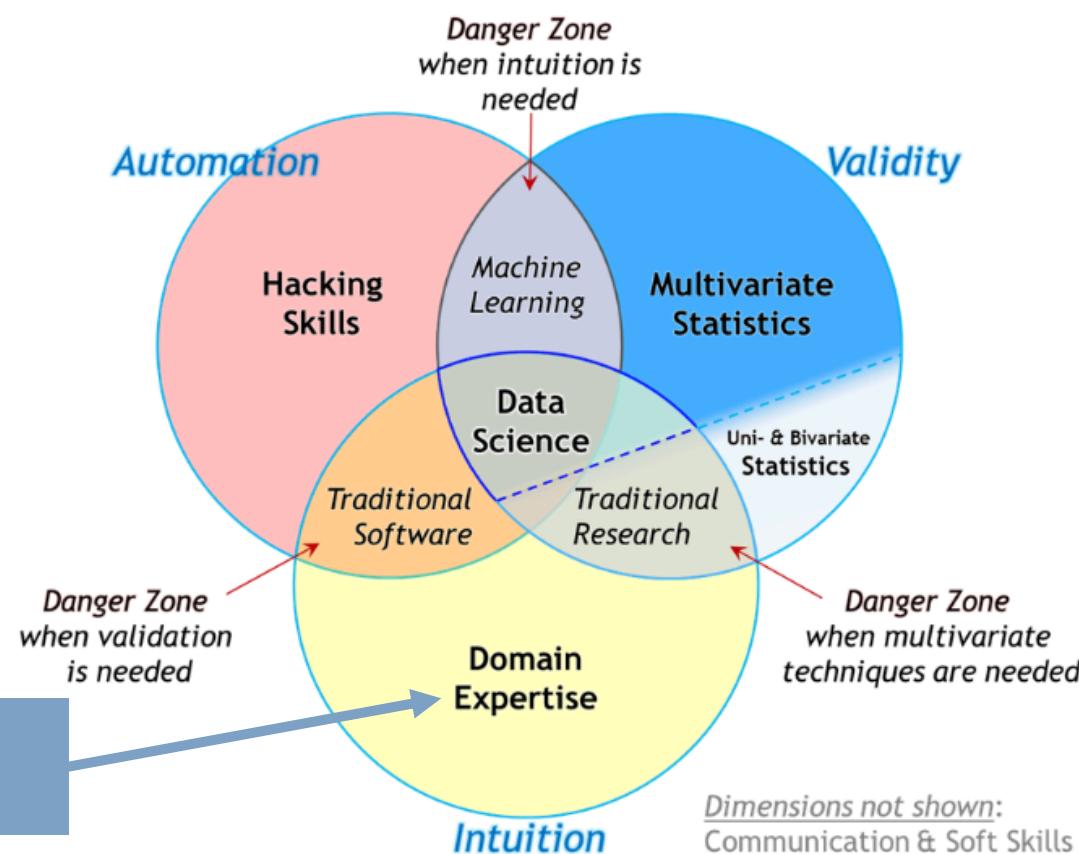
SCIENCE POLICY
RESEARCH UNIT

What is data science?



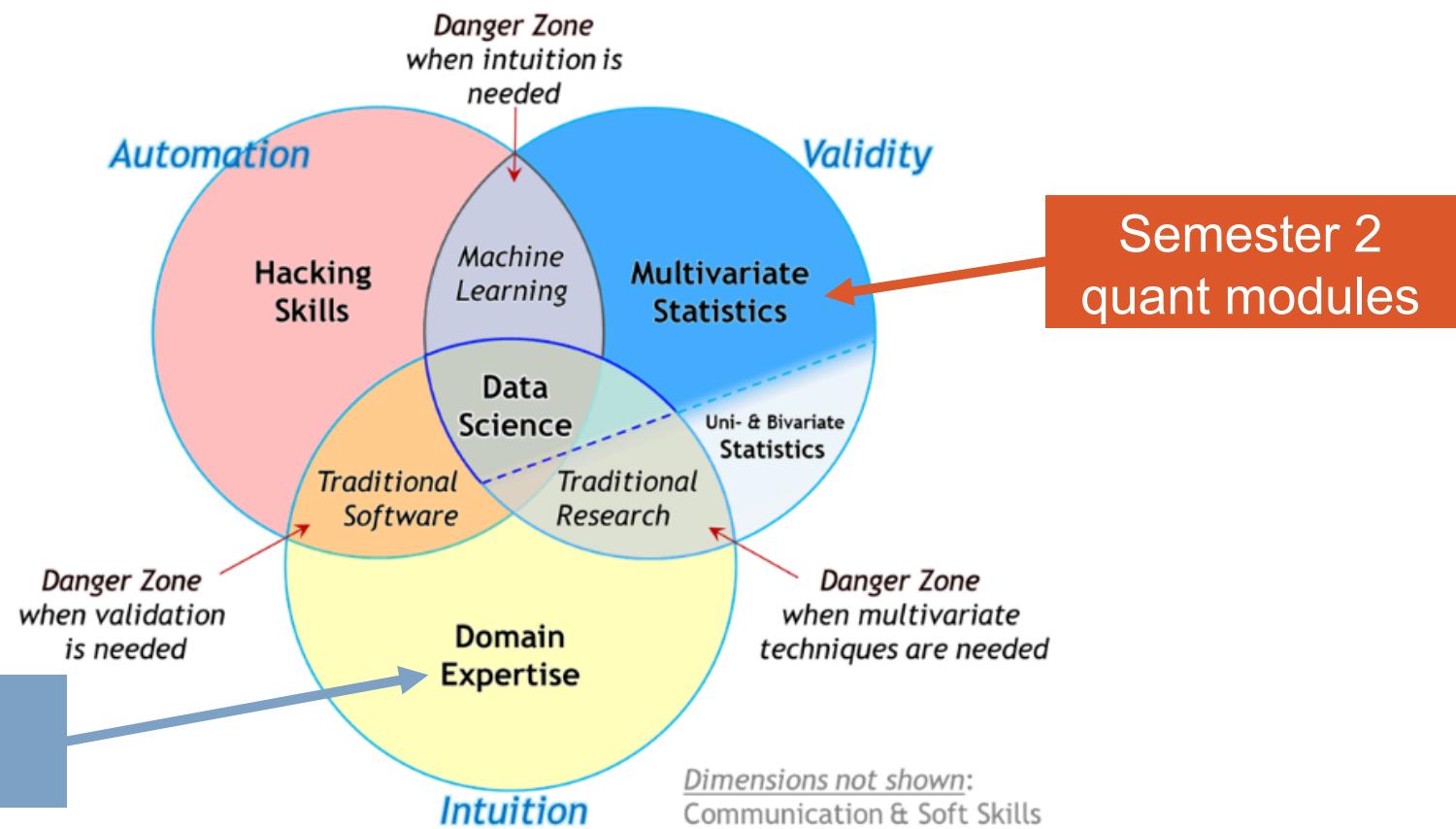
Source: Andrew Silver, <https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40>

What is data science?



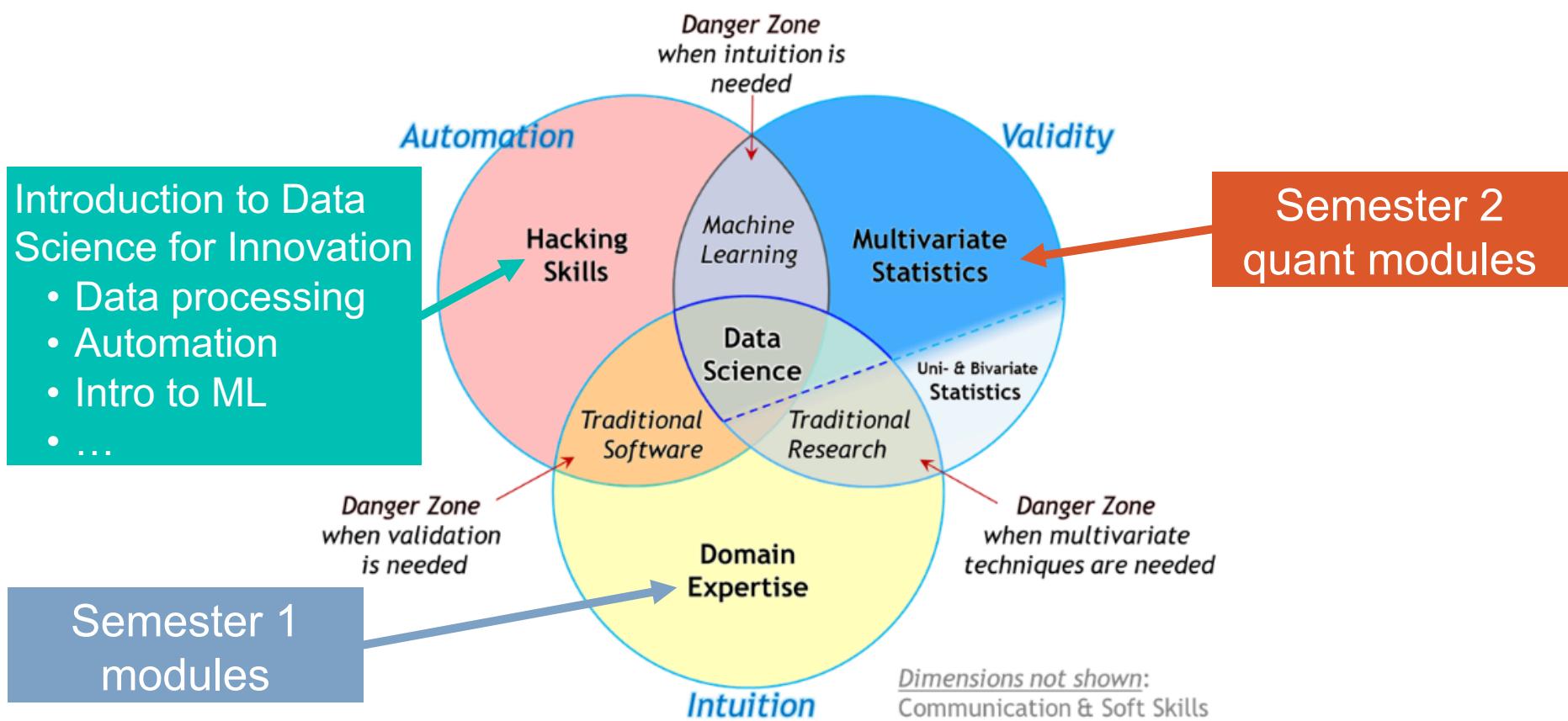
Source: Andrew Silver, <https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40>

What is data science?



Source: Andrew Silver, <https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40>

What is data science?



Source: Andrew Silver, <https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40>

What will you learn in this module?

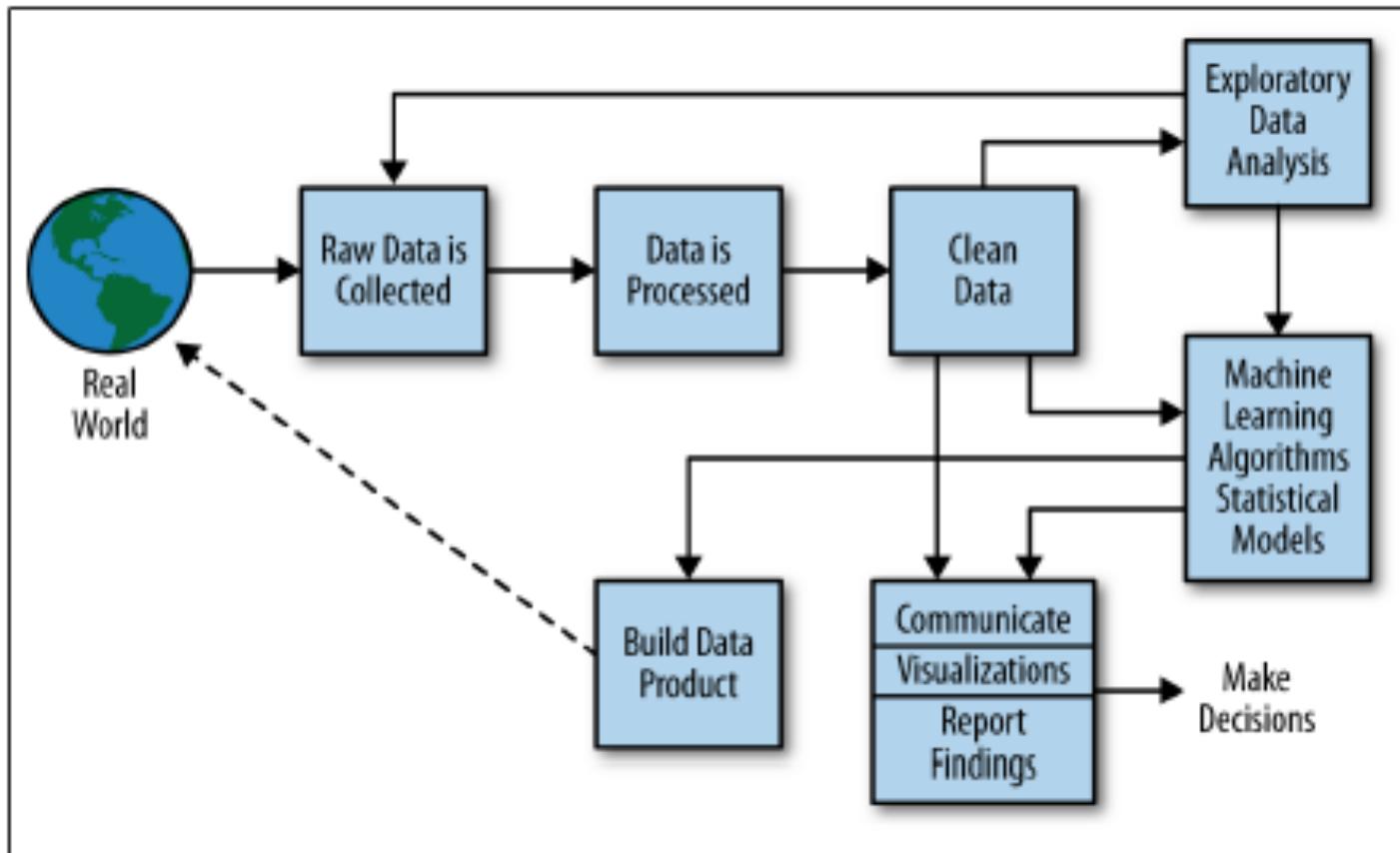
What you will learn on other Semester 1 modules?

You will develop an understanding of science and innovation policy with associated concepts and literature: together with your practical knowledge of science and technology policy, this provides you with '**domain expertise**'.

This course aims to provide you with practical skills:

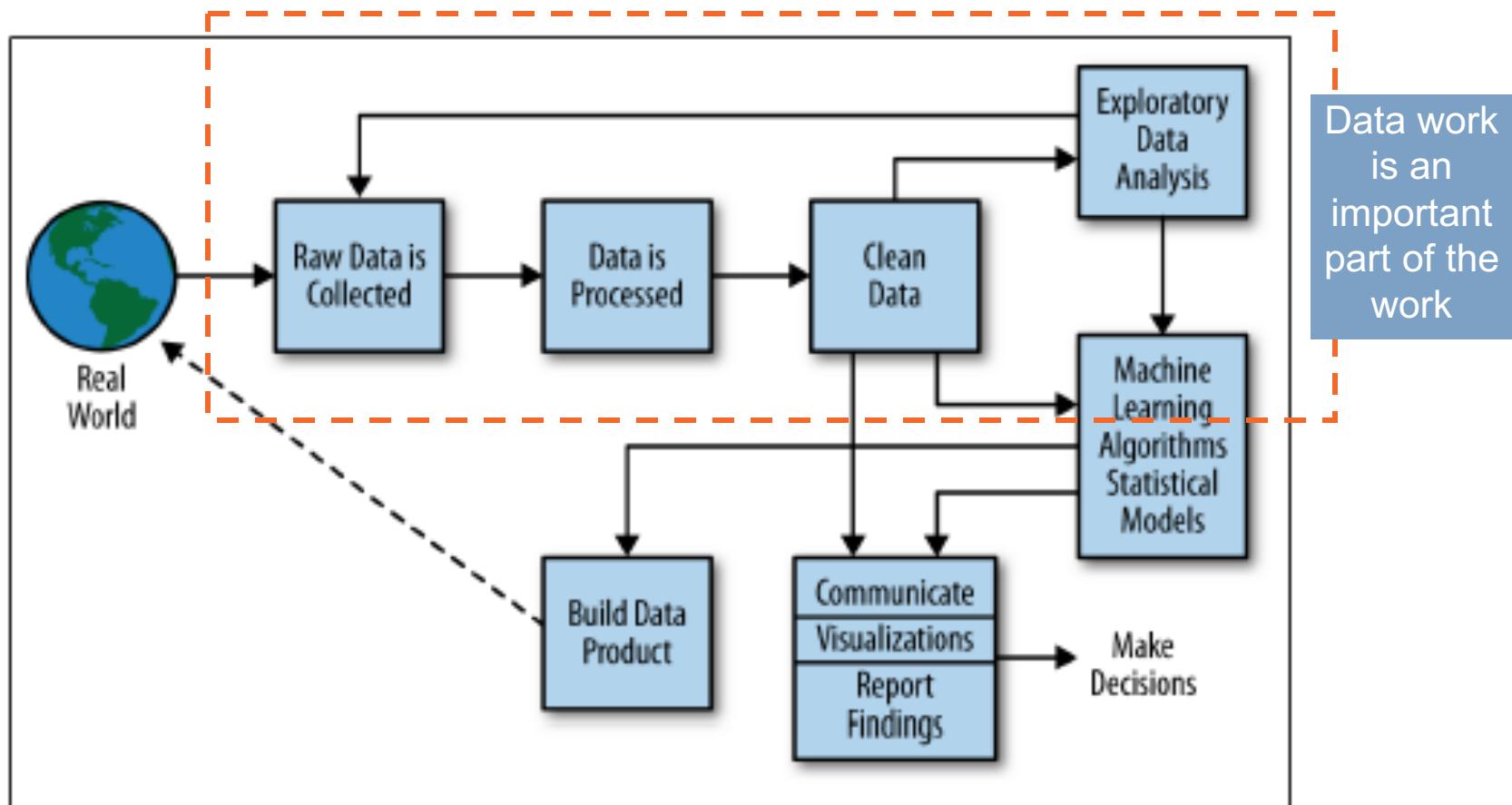
- To work with innovation related data ('into the wild')
- To clean and transform data in an efficient manner
- To generate analysis and visualisation for policy and decision making
- Through the use of computational tools (learning how to program)

What is data science?



Source: The Data Science Process, By Rachel Schutt & Cathy O'Neil, Doing Data Science, P. 41

What is data science?



Source: The Data Science Process, By Rachel Schutt & Cathy O'Neil, Doing Data Science, P. 41

What is data science?

Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

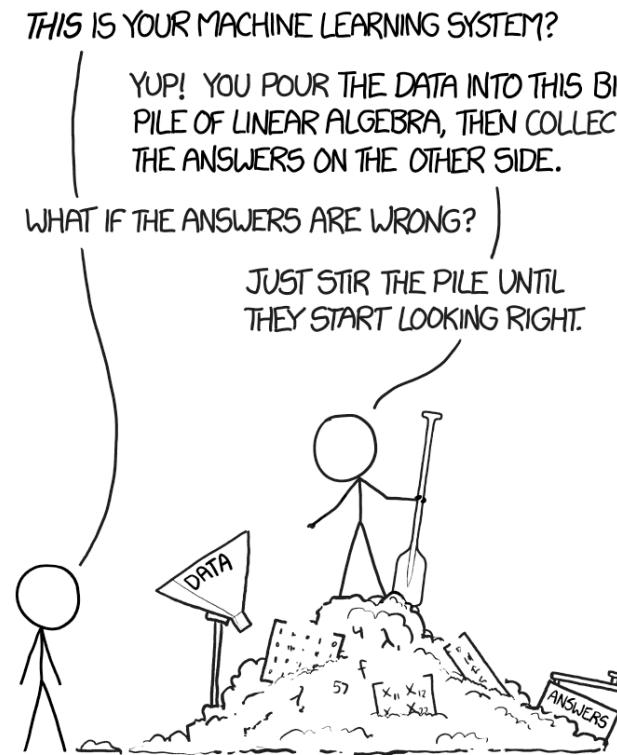
Source: New York Times , 17 Aug. 2014

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

Garbage in → Garbage out

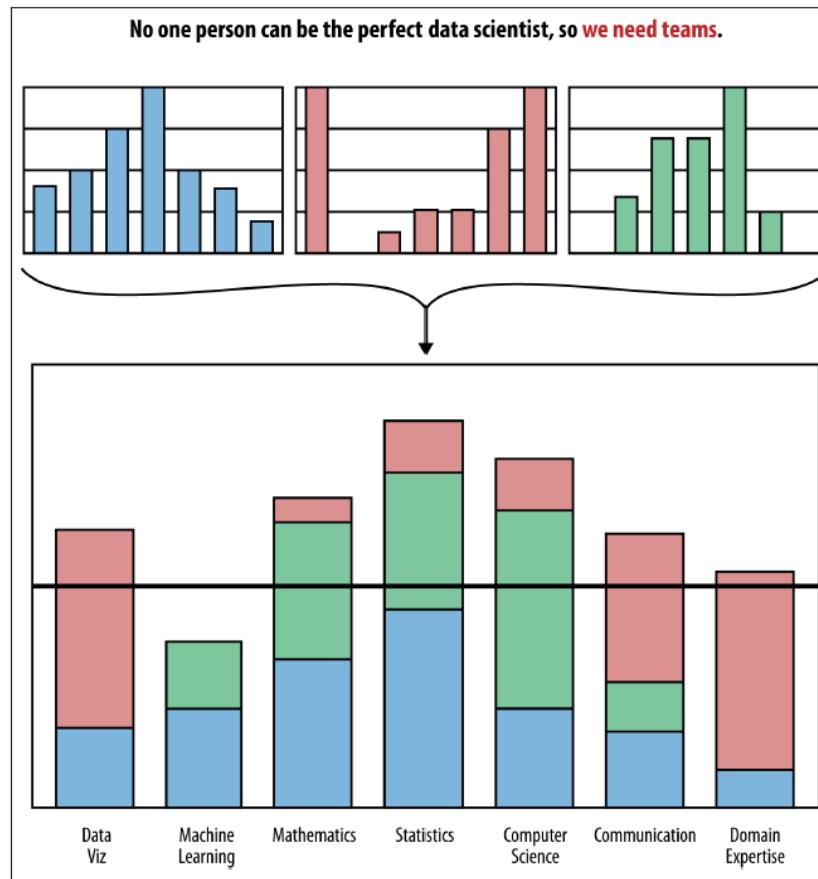
What is data science?

Garbage in → Garbage out



<https://xkcd.com/1838/>

What is data science?



Source: The Team profiles, By Rachel Schutt & Cathy O'Neil, Doing Data Science, P. 12

Introduction to Data Science for Innovation

You won't become computing/machine learning experts.

But you will gain knowledge on the workings of each part of the pipeline with practical cases of science and innovation data.

You will learn how to learn.

It's up to you to take your knowledge to the next level.

Why is it important?



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT

Why learning data science?

Science and Innovation policy is
evidence based

- Example from Rebecca Endean,
Strategy Director of UK Research and Innovation
Presenting at the ResearchFish conference 2018.

(https://5ba0c32f-08ac-4ea4-84b2-3adc5e339103.filesusr.com/ugd/fc7e18_f53daa3096ed4822a6df0783bfbbfa51.pdf)

Why learning data science?

UKRI's approach to assessing impact is based around recognising key differences in its activities and focusing on relevant impacts

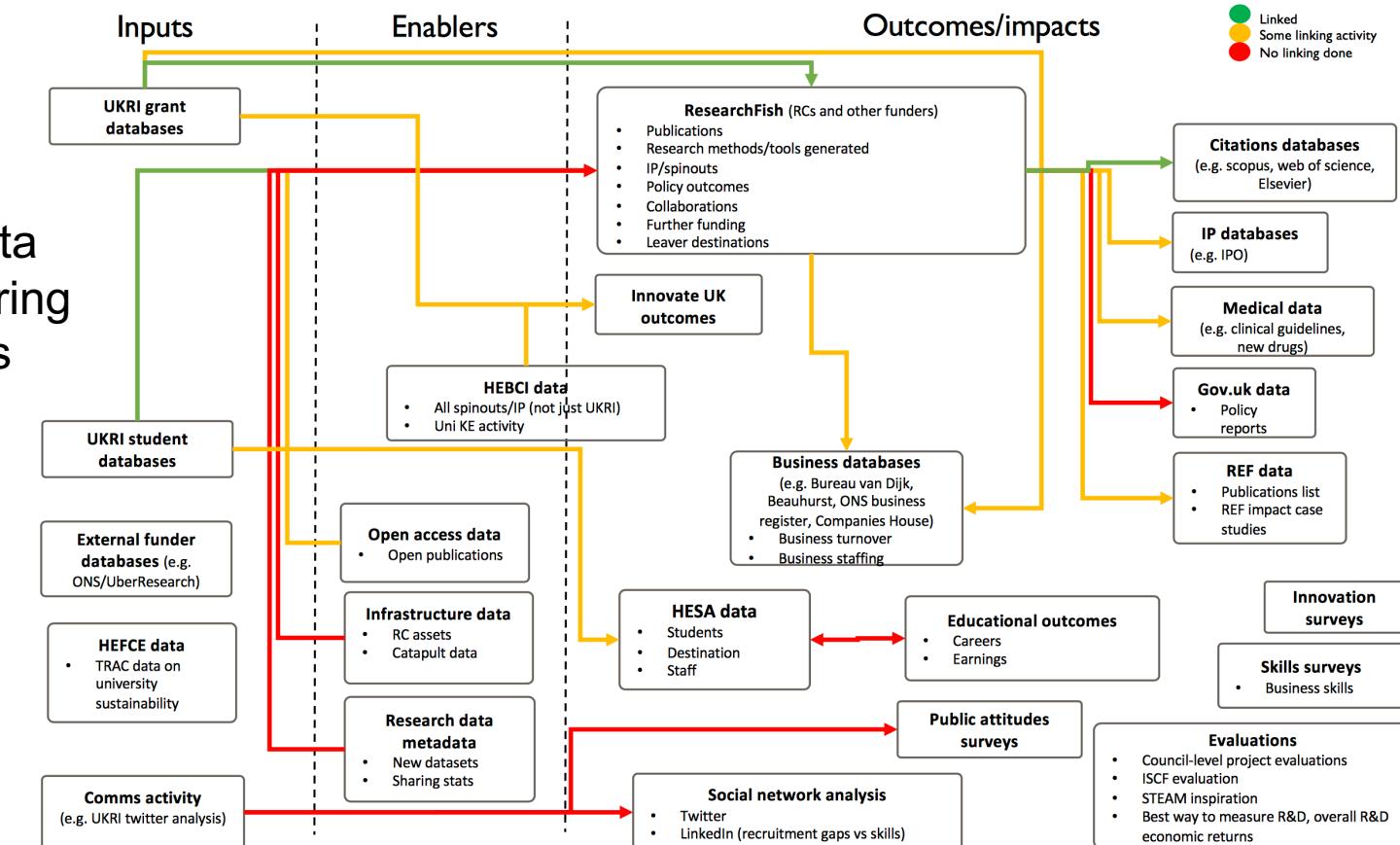
There is no one-size-fits-all approach to evaluating research and innovation programmes. UKRI is developing an overarching evaluation strategy which:

- Is systematic, evidence-based and dynamic;
- Assesses impact beyond counting citations;
- Does not privilege one discipline over another because its outputs are more tangible and/or immediate;
- Understands that there is a lag between funding and realisation of impact;
- Understands that impact may not be in the form of a product or tool, but contributing towards a web of knowledge which allows for further understanding/work to be done; and
- Makes a convincing and robust case for impact, utilising effective tools and methods for assessing impact.

Rebecca Endean, UKRI, 2018

Why learning data science?

UKRI
From data
To measuring
impacts



Rebecca Endean, UKRI, 2018



So what will we learn in
the rest of the course?

Outline of the course

Outline

Introduction to programming and data manipulation:

- Week 1: Introduction
- Week 2: Data structure
- Week 3: Data Collection and Transformation

Lecture outline

Data in the wild:

- Week 4: Data sources and innovation indicators
innovation 1
- Week 5: Data sources and innovation indicators
innovation 2 (Guest lecture – ResearchFish
TBC)

Lecture outline

Analysing innovation data:

- Week 6: Textual data and text-mining (P1)
- Week 7: Textual data and text-mining (P2)
- Week 8: Data visualisation
- Week 9: Introduction to machine learning

Lecture outline

Wrapping up:

- Week 10: Revisions
- Week 11: Data Science, Policy and Practice
(Guest lecture –TBC)

Assessments



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT

Assessments

Learning outcome	Assessment mode
• Identify data sources to address research questions and the main ethical concerns associated with data usage	REP
• Identify the most appropriate approaches to read, process and transform data	REP, GPN
• Describe and apply major text-mining techniques to generate intelligence for decision making	GPN
• Employ data analysis techniques to produce effective visualisations	REP, GPN

Assessments

Coursework, (GPN):

Group presentation (30% of the final mark)

A small text mining project.

Timeline:

Week 1-3: creation of groups (2-3 students)

Week 3-10: project development (including supervision)

Week 11: project submission

Week 11: group presentation

Submit your presentation to f.bone@sussex.ac.uk and d.rotolo@sussex.ac.uk by Week 11.

Assessments

Assessment (REP):

Report – 3000 words (70% of the final mark)

- The report will need to examine a topic or a phenomenon related to Science & Technology Policy using the analytical and visualisation techniques introduced in the module
- The analysis should rely on a working script in R to be included as part of the report.
- The report should draw on the content of other modules that you have attended to produce a critical interpretation of the results of your analysis.
- Further details will communicated.

Learning to program



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT

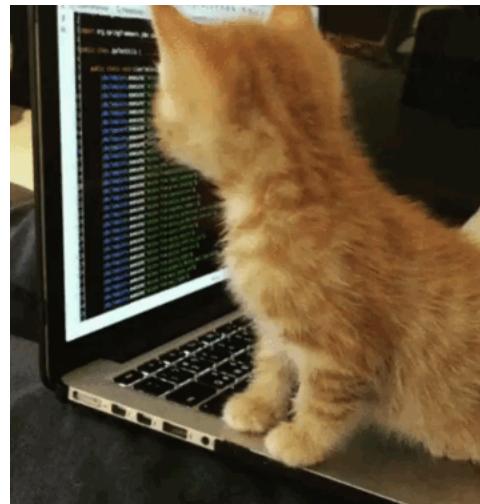
Learning to program

Do you have any programming experience?



Learning to program

Do you have any programming experience?



Don't worry we'll teach you the basics ...

Learning to program

What language should I learn?

We will learn R!

- R is an easy to use language
- It started as a language for statistics
- But becomes increasingly popular for data science projects

It includes many packages for:

- importing data
- data wrangling
- data visualisation

Comes with a good environment/user interface



Learning to program

The **workshops** will give you hands-on exercises to learn how to use R

- Introduce the RStudio environment.
- Getting data and handling them in R
- Using programming logic to solve problems
- Working with text data
- Introduce machine learning
- Visualise the data to take decisions



Learning to program

There is a lot of useful functions in all the packages available (maybe too many to remember):

Use Cheat Sheets



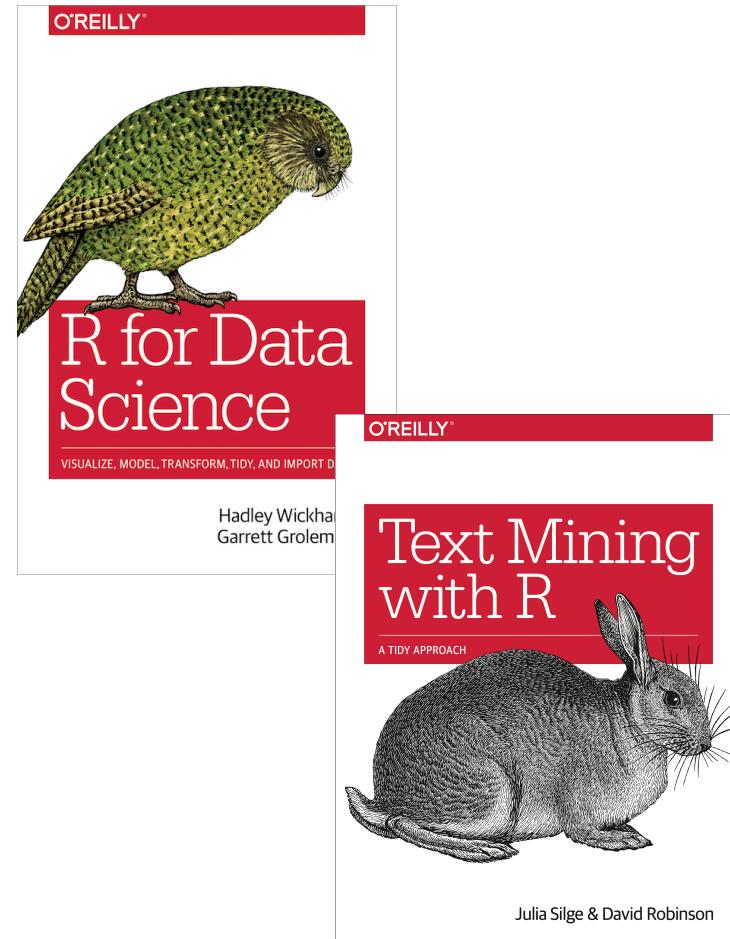
<https://rstudio.com/resources/cheatsheets/>

Learning to program

Many suggested readings will include chapters from the ‘R for Data Science’ book.
(<https://r4ds.had.co.nz/index.html>)

For text mining, the book ‘Text Mining with R’ is also a reference.
(<https://www.tidytextmining.com>)

They are both available online for free.



Learning to program

The other main language for data science is 'Python'.

Python is a language used by engineers, computer scientists.

It is more likely that the latest machine learning algorithms are first available on python.

Both R & Python have their strength for data science (see this [blog](#)).

If you ever wanted to start using python and still working with R, you can use reticulate.





Let's practice with R!



Overview of RStudio



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT

Objectives

1) Guided tour of RStudio

2) Basic operations in R

Through 4 small exercises:

1. Do simple computations in R
2. Use variables in cumulative steps
3. Glimpse a dataset
4. Work with R packages (to use specific functions developed by others)

To do the exercises, download the R sheet on canvas.

Rstudio

We are going to work with the RStudio environment.

Use either your own copy of Rstudio (see canvas for installation guide)



Or go to the Sussex server (on campus only):

<http://rstudio.uscs.susx.ac.uk/>

View of RStudio

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Shows a script named "09.3.dynamic_analysis_kws_draft.R" containing R code. The code includes comments explaining simple arithmetic operations and variable assignment.
- Environment Browser:** Shows the "Titanic" dataset loaded into the Global Environment. It displays 32 observations and 5 variables, with a value of 3 assigned to variable "a".
- Plots:** A histogram titled "Histogram of Titanic\$Freq" showing the frequency distribution of the "Freq" variable. The x-axis ranges from 0 to 700, and the y-axis ranges from 0 to 25. The distribution is highly right-skewed, with the highest frequency occurring between 0 and 100.

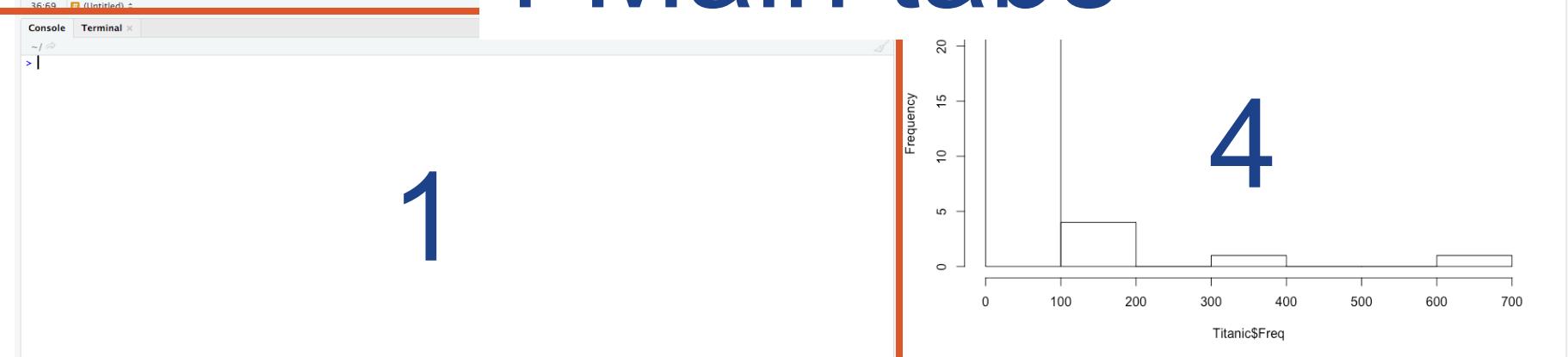
View of RStudio

The screenshot displays the RStudio interface with the following components:

- Script Editor (Left Panel):** Shows an R script titled "09.3.dynamic_analysis_kws_draft.R". The script contains comments and code snippets related to Data Science for Innovation, Seminar 1, and Frédérique Bone. It includes sections for simple computations, variable assignment, and a check for the expected value of a variable.
- Environment Browser (Top Right Panel):** Shows the global environment with the "Titanic" dataset loaded, containing 32 observations and 5 variables. A variable "a" is assigned the value 3.
- Plots (Bottom Right Panel):** Displays a histogram titled "Histogram of Titanic\$Freq" showing the frequency distribution of the "Freq" variable. The x-axis is labeled "Titanic\$Freq" and ranges from 0 to 700. The y-axis is labeled "Frequency" and ranges from 0 to 25. The histogram has three bars: one between 0 and 100 with a frequency of approximately 25, one between 100 and 200 with a frequency of approximately 5, and one between 300 and 400 with a frequency of approximately 1.
- Console (Bottom Left Panel):** Shows the command prompt with the current working directory as "/".

View of RStudio

1

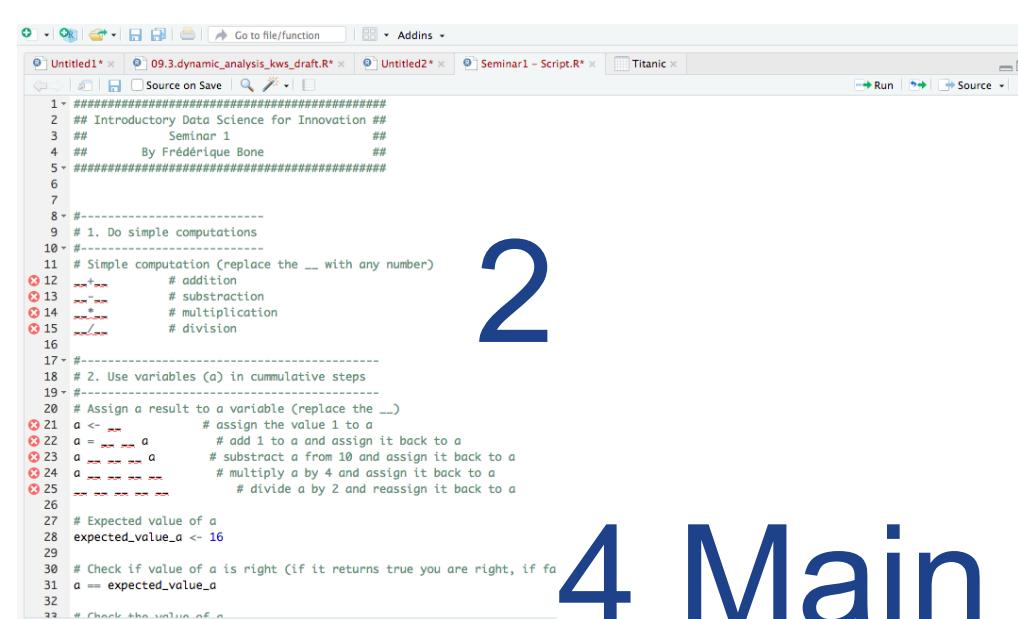


```

Console Terminal x
~/ ...
> |

```

2

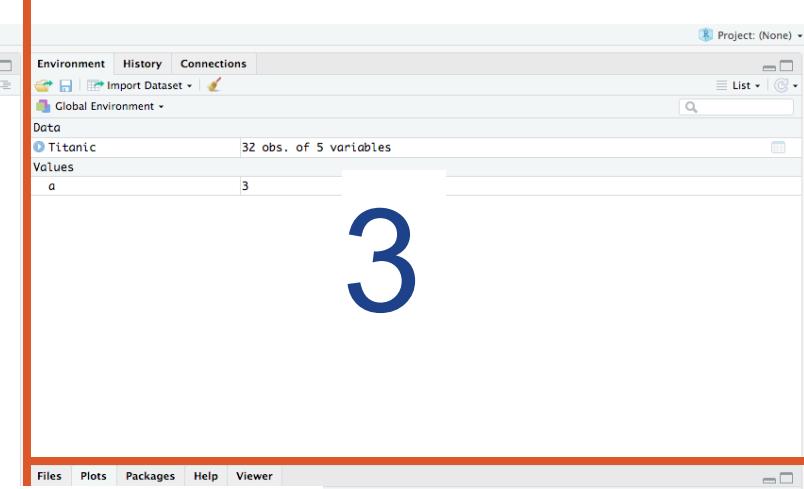


```

1 #####
2 ## Introductory Data Science for Innovation ##
3 ## Seminar 1 ##
4 ## By Frédérique Bone ##
5 #####
6
7
8 #-----
9 # 1. Do simple computations
10 #-----
11 # Simple computation (replace the ___ with any number)
12 ___+___ # addition
13 ___-___ # subtraction
14 ___*___ # multiplication
15 ___/____ # division
16
17 #-----
18 # 2. Use variables (a) in cumulative steps
19 #-----
20 # Assign a result to a variable (replace the ___)
21 a <- ___ # assign the value 1 to a
22 a = ___ + a # add 1 to a and assign it back to a
23 a = ___ - a # subtract a from 10 and assign it back to a
24 a = ___ * a # multiply a by 4 and assign it back to a
25 a = ___ / a # divide a by 2 and reassign it back to a
26
27 # Expected value of a
28 expected_value_a <- 16
29
30 # Check if value of a is right (if it returns true you are right, if fa
31 a == expected_value_a
32
33 # Checks when certain ref.n
36:69: 0 (Untitled.R)

```

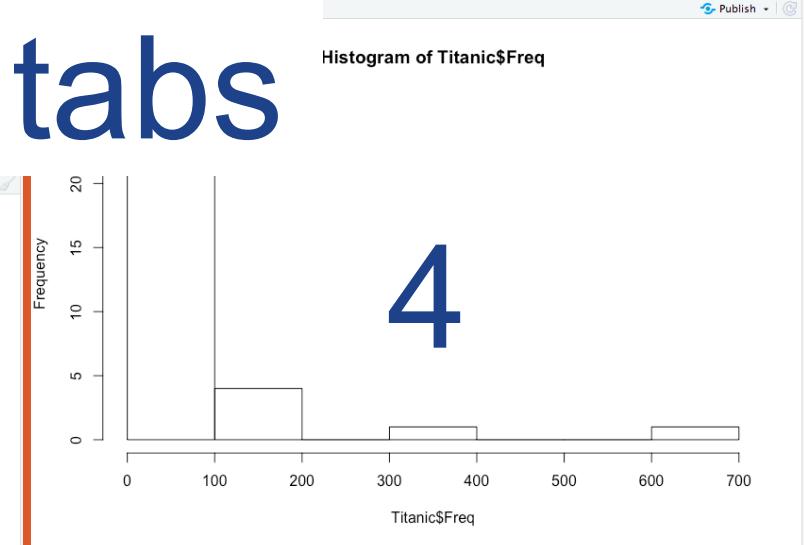
3



	Titanic	32 obs. of 5 variables
Values	a	3

4 Main tabs

4



Histogram of Titanic\$Freq

Frequency

Titanic\$Freq

Bin Range (Titanic\$Freq)	Frequency
0 - 100	0
100 - 200	5
200 - 300	0
300 - 400	1
400 - 500	0
500 - 600	0
600 - 700	1

View of RStudio

The console

```

1
2
3
4

```

1

2

3

4

```

1 #####
2 ## Introductory Data Science for Innovation ##
3 ## Seminar 1 ##
4 ## By Frédérique Bone ##
5 #####
6
7
8 #-----
9 # 1. Do simple computations
10 #-----
11 # Simple computation (replace the ___ with any number)
12 ___+___ # addition
13 ___-___ # subtraction
14 ___*___ # multiplication
15 ___/____ # division
16
17 #-----
18 # 2. Use variables (a) in cumulative steps
19 #-----
20 # Assign a result to a variable (replace the ___)
21 a <- ___ # assign the value 1 to a
22 a = ___ + a # add 1 to a and assign it back to a
23 a = ___ - a # subtract a from 10 and assign it back to a
24 a = ___ * a # multiply a by 4 and assign it back to a
25 a = ___ / a # divide a by 2 and reassign it back to a
26
27 # Expected value of a
28 expected_value_a <- 16
29
30 # Check if value of a is right (if it returns true you are right, if false)
31 a == expected_value_a
32
33 # Checks when certain numbers
34 # (Initiated) a
35
36:9 0 (Untitled.R)

```

ram of Titanic\$Freq

Titanic\$Freq Bin Range	Frequency
0-50	0
50-100	0
100-150	5
150-200	0
200-250	0
250-300	0
300-350	1
350-400	1
400-450	0
450-500	0
500-550	0
550-600	1
600-650	1
650-700	1

The console

The console :
where you give your instructions for some action to be done.

Try typing some basic operations, and press enter

```
> 1+1
[1] 2
> 3-4
[1] -1
> 4*10
[1] 40
> 3/2
[1] 1.5
`-
```

View of RStudio

1

```

1 ##### Introductory Data Science for Innovation #####
2 ## Seminar 1 ##
3 ## By Frédérique Bone ##
4 #####
5 #####
6
7
8 #-----
9 # 1. Do simple computations
10 #-----
11 # Simple computation (replace the ___ with any number)
12 ___+___ # addition
13 ___-___ # subtraction
14 ___*___ # multiplication
15 ___/____ # division
16
17 #-----
18 # 2. Use variables (a) in cumulative steps
19 #-----
20 # Assign a result to a variable (replace the ___)
21 a <- ___ # assign the value 1 to a
22 a = ___ + a # add 1 to a and assign it back to a
23 a = ___ - a # subtract a from 10 and assign it back to a
24 a = ___ * a # multiply a by 4 and assign it back to a
25 a = ___ / a # divide a by 2 and reassign it back to a
26
27 # Expected value of a
28 expected_value_a <- 16
29
30 # Check if value of a is right (if it returns true you are right, if false)
31 a == expected_value_a
32
33 # Checks when certain numbers
34 # (Initiated) a

```

2

3

4

Using scripts

Scripts:

When you want to write something beyond one operation, it may be useful to create a script.

A script enables you to write all your code in a single document.

- Better organisation
- Run the same code at a later stage

Let's upload the exercise script and start section 1

Replace __ by the correct value

Using scripts

Move to exercise 2.

Using the same operators than in exercise 1 use different steps to change the value of a

Using scripts

Now change the first value of a
Rerun that part of the code.

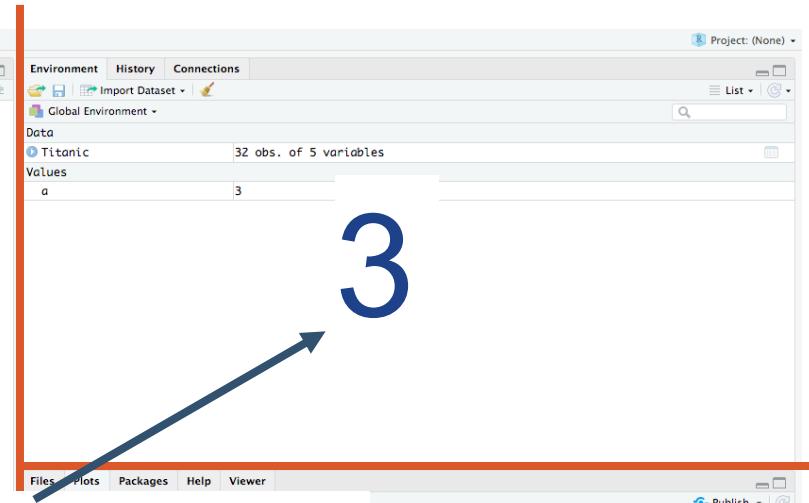
View of RStudio

```

1 #####
2 ## Introductory Data Science for Innovation ##
3 ## Seminar 1 ##
4 ## By Frédérique Bone ##
5 #####
6
7
8 #-----
9 # 1. Do simple computations
10 #-----
11 # Simple computation (replace the ___ with any number)
12 ___+___ # addition
13 ___-___ # subtraction
14 ___*___ # multiplication
15 ___/____ # division
16
17 #-----
18 # 2. Use variables (a) in cumulative steps
19 #-----
20 # Assign a result to a variable (replace the ___)
21 a <- ___ # assign the value 1 to a
22 a = ___ + a # add 1 to a and assign it back to a
23 a = ___ - a # subtract a from 10 and assign it back to a
24 a = ___ * a # multiply a by 4 and assign it back to a
25 a = ___ / a # divide a by 2 and reassign it back to a
26
27 # Expected value of a
28 expected_value_a <- 16
29
30 # Check if value of a is right (if it returns true you are right, if false
31 a == expected_value_a
32
33 # Check other function of a
34 # (Untitled).R

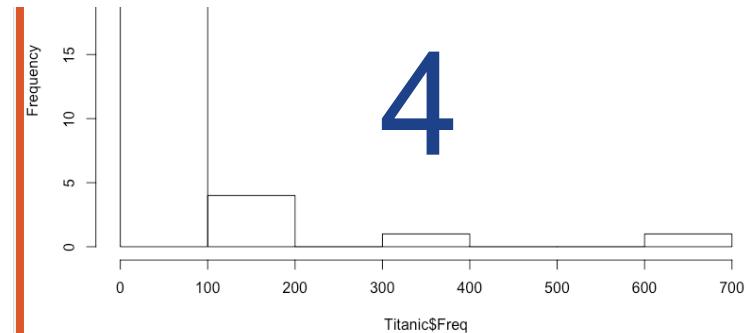
```

1



3

Looking at variables
in the environment



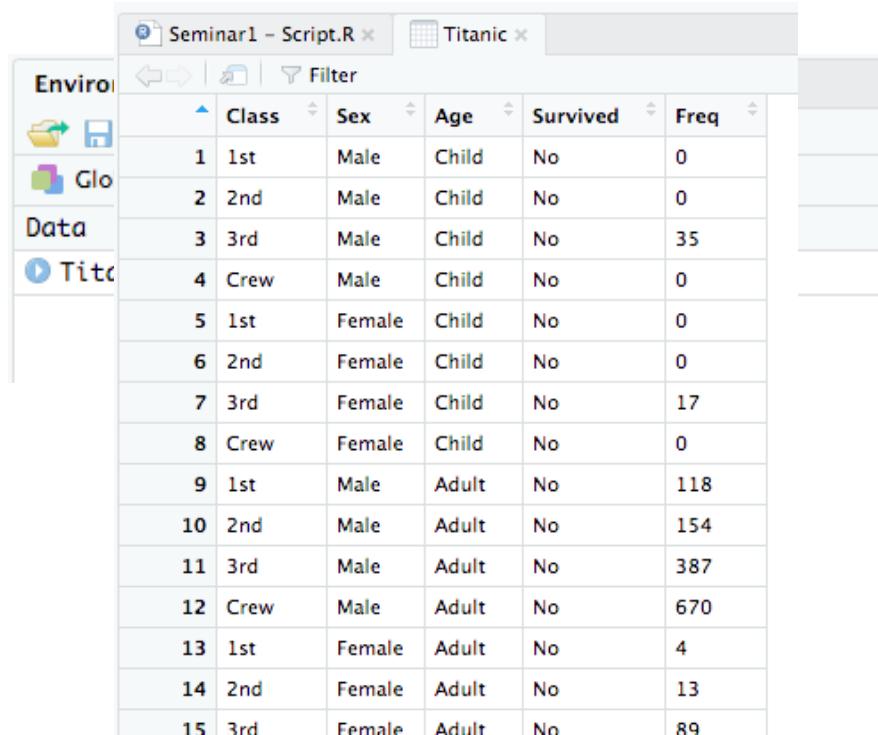
4

Exploring datasets

Move to exercise 3.
We are going to explore a dataset.

Using scripts

What is the size of the dataset?
What are the main variables of the dataset?



The screenshot shows the RStudio interface with the 'Titanic' dataset loaded into a data viewer. The environment sidebar on the left lists 'Environ', 'Global', 'Data', and 'Titanic'. The data viewer window has tabs for 'Seminar1 - Script.R' and 'Titanic'. The 'Titanic' tab is active, displaying a table with 15 rows and 6 columns: Row, Class, Sex, Age, Survived, and Freq. The data shows the distribution of passengers by class, sex, and survival status.

	Class	Sex	Age	Survived	Freq
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0
7	3rd	Female	Child	No	17
8	Crew	Female	Child	No	0
9	1st	Male	Adult	No	118
10	2nd	Male	Adult	No	154
11	3rd	Male	Adult	No	387
12	Crew	Male	Adult	No	670
13	1st	Female	Adult	No	4
14	2nd	Female	Adult	No	13
15	3rd	Female	Adult	No	89

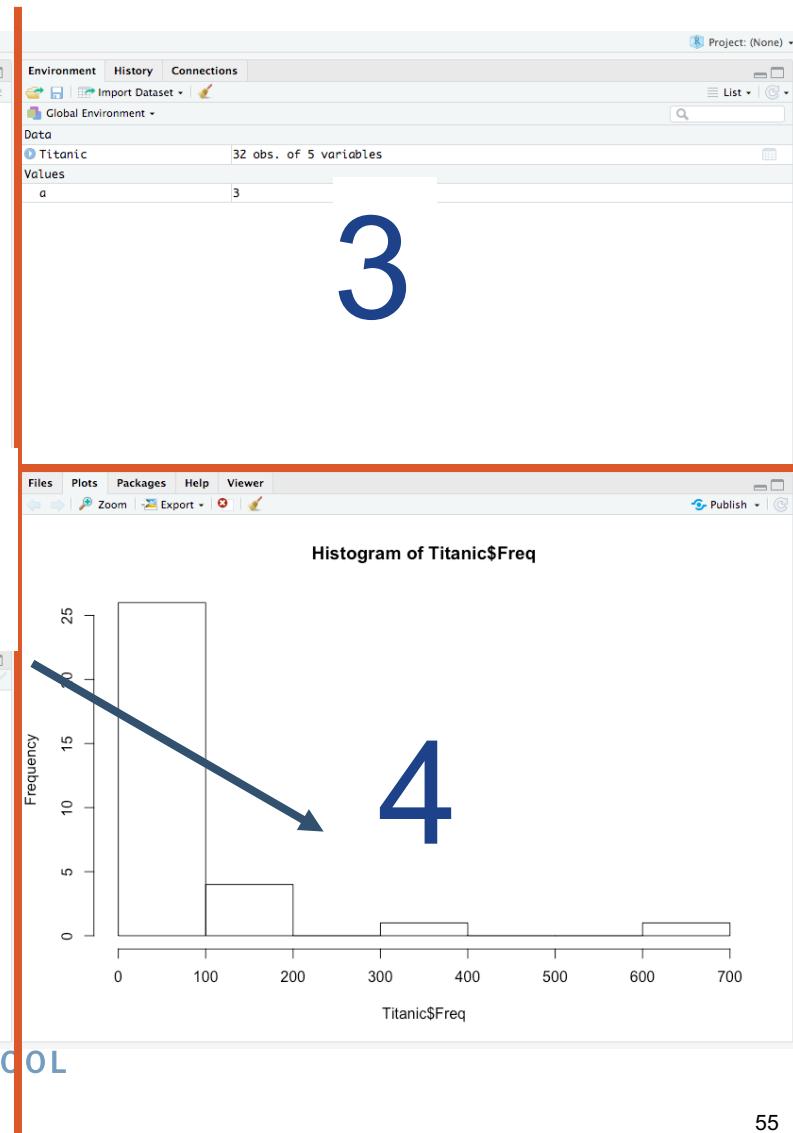
View of RStudio

1
2
Files, Plots,
Packages and Help.

```

1 #####
2 ## Introductory Data Science for Innovation ##
3 ## Seminar 1 ##
4 ## By Frédérique Bone ##
5 #####
6
7
8 #-----
9 # 1. Do simple computations
10 #-----
11 # Simple computation (replace the ___ with any number)
12 ___+___ # addition
13 ___-___ # subtraction
14 ___*___ # multiplication
15 ___/____ # division
16
17 #-----
18 # 2. Use variables (a) in cumulative steps
19 #-----
20 # Assign a result to a variable (replace the ___)
21 a <- ___ # assign the value 1 to a
22 a = ___ + ___ # add 1 to a and assign it back to a
23 a = ___ * ___
24 a = ___ / ___
25 a = ___ - ___
26
27 # Expected value of
28 expected_value_a <-
29
30 # Check if value of
31 a == expected_value_
32
33 # Checks when function is
36:69: 0 (Untitled.R)

```



Installing packages

Move to exercise 4 to install packages.



Thank you.

Contact:

Frédérique Bone
[\(f.bone@sussex.ac.uk\)](mailto:f.bone@sussex.ac.uk)
Daniele Rotolo
[\(d.rotolo@sussex.ac.uk\)](mailto:d.rotolo@sussex.ac.uk)



BUSINESS
SCHOOL

SCIENCE POLICY
RESEARCH UNIT