# Introduction to R and Data Science

Bhavesh Shah, Scientist (Environment), V&A

And student on MSc Data Science Course at UCL

**b.shah@vam.ac.uk**

**#IconMMN**

#NOMATERIALTOOMODERN

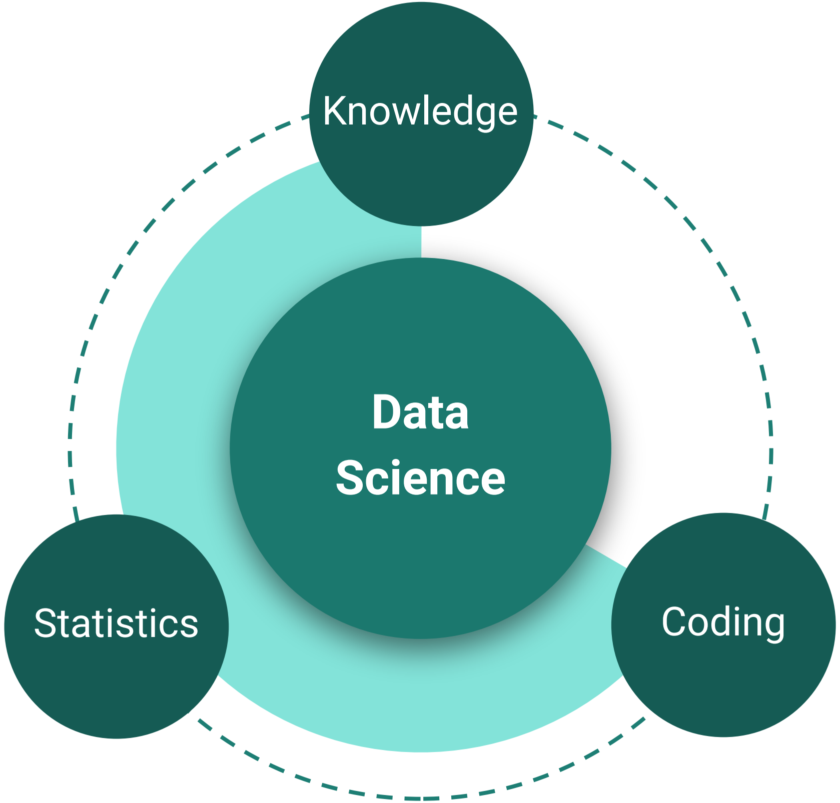
# Data science

Breakthroughs in technology, computing and algorithms

And lots and lots of data...

New kid on the block?

Or just research using a computer?



# Why code?

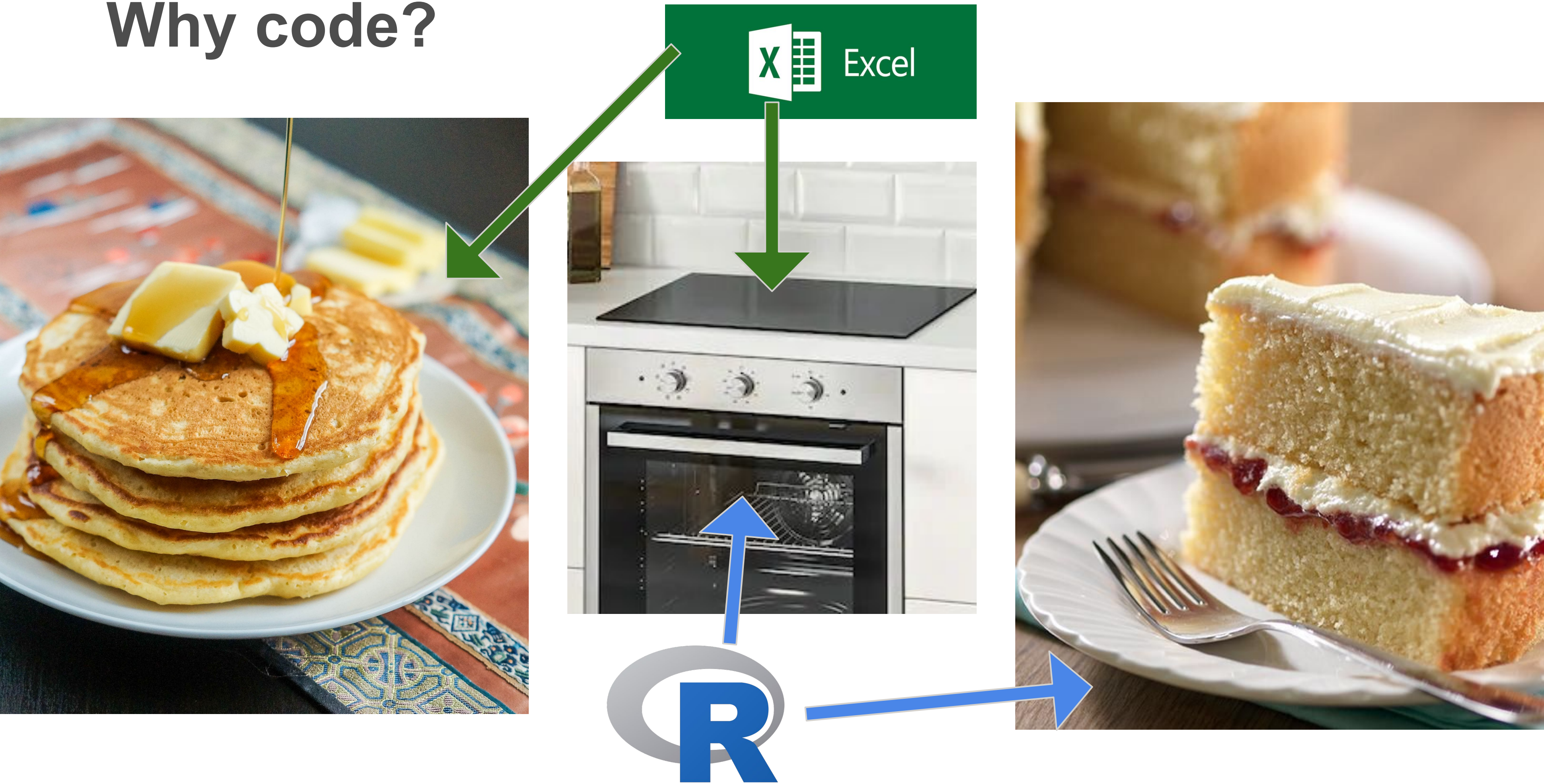
Let’s try to answer a different question:

If you have (managed to find) flour and eggs, would you rather have...

# Why code?

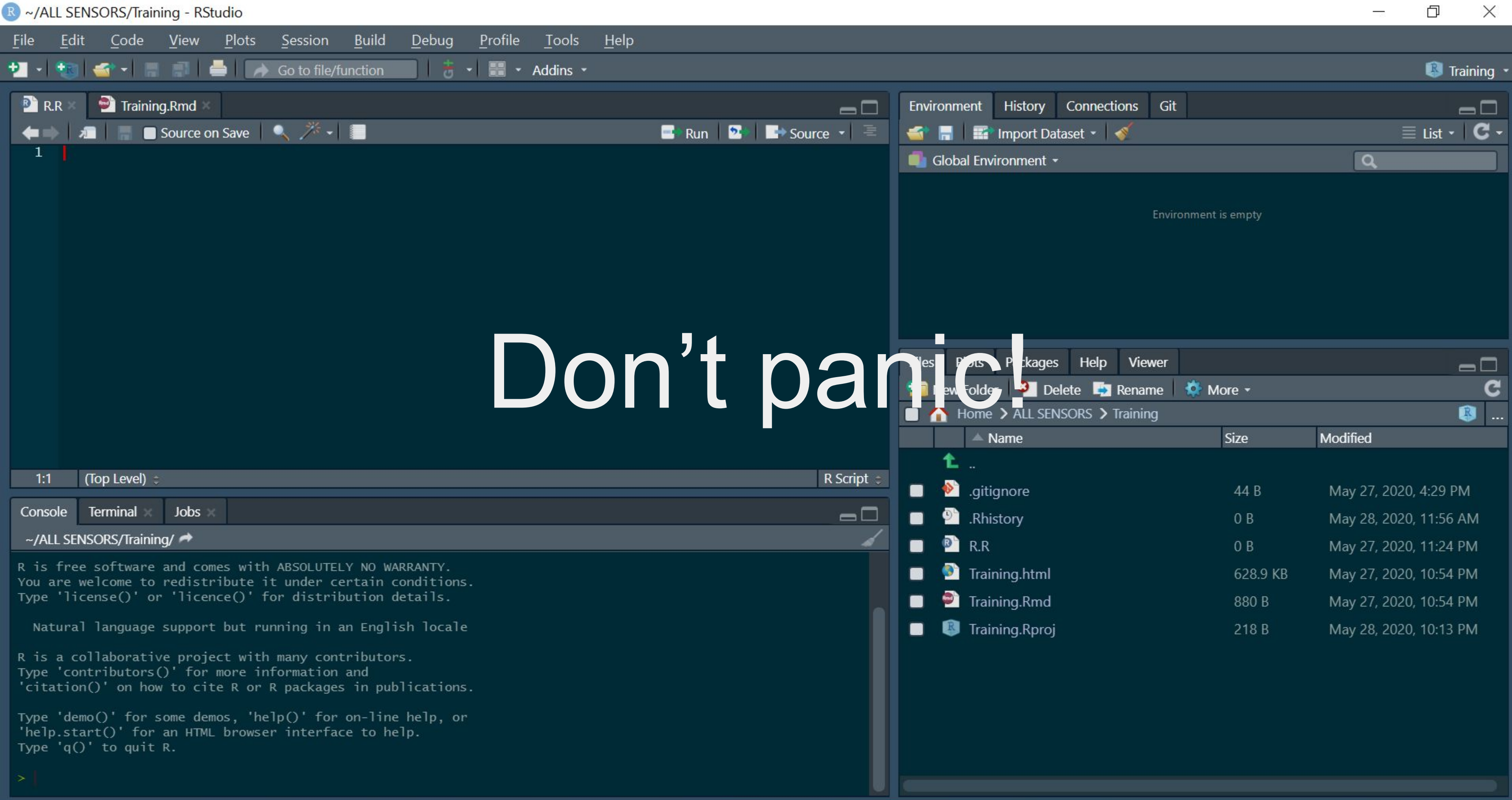
Pancake
or
cake?

# Why code?



# Coding

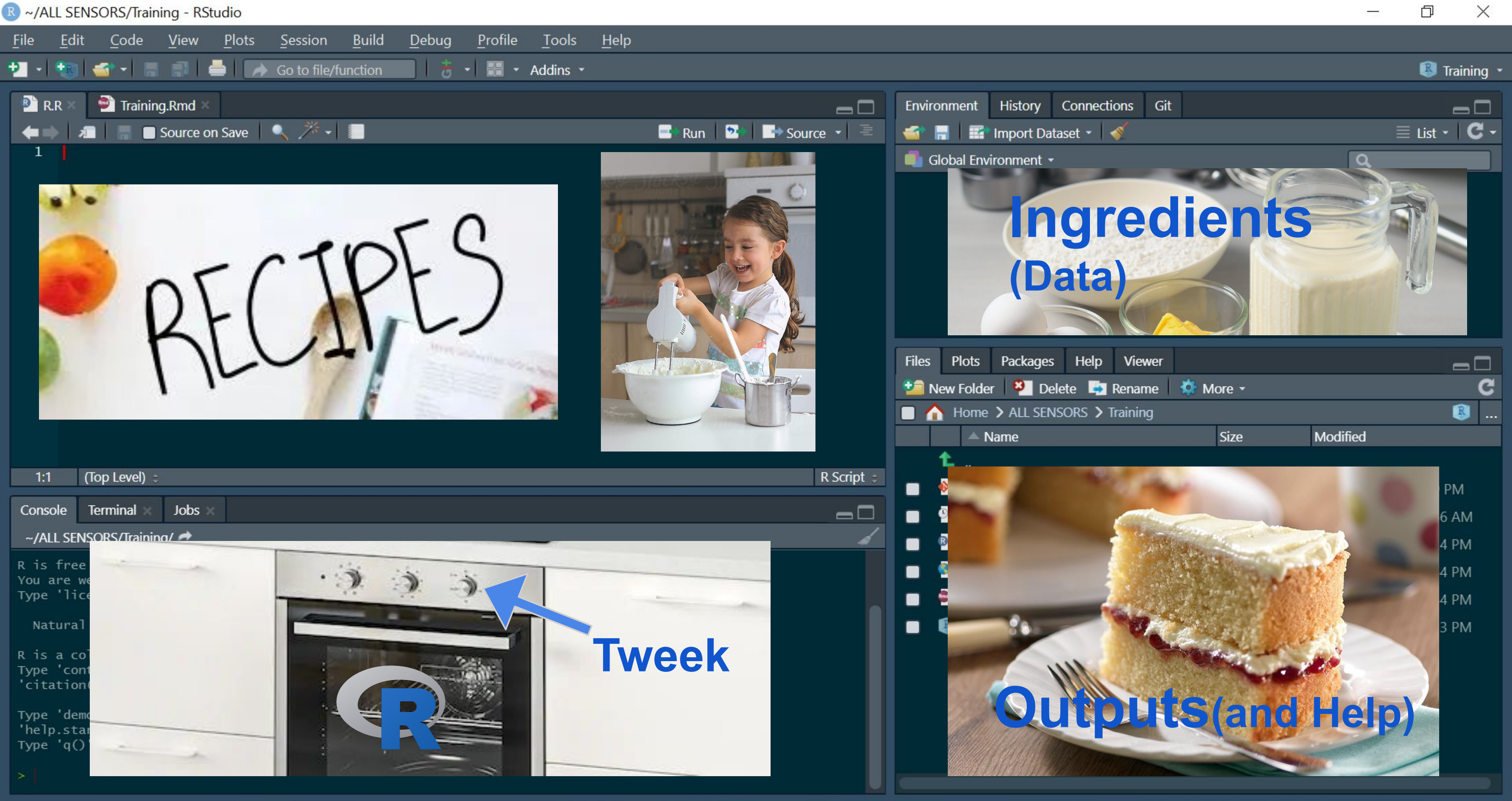
1) You write the recipes (code) and prepare the ingredients (data)
2) R does the baking - you cross your fingers
3) Add the finishing touches - and serve!



Don’t panic!

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

| Name | Size | Modified |
|---|---|---|
| .. | | |
| .gitignore | 44 B | May 27, 2020, 4:29 PM |
| .Rhistory | 0 B | May 28, 2020, 11:56 AM |
| R.R | 0 B | May 27, 2020, 11:24 PM |
| Training.html | 628.9 KB | May 27, 2020, 10:54 PM |
| Training.Rmd | 880 B | May 27, 2020, 10:54 PM |
| Training.Rproj | 218 B | May 28, 2020, 10:13 PM |

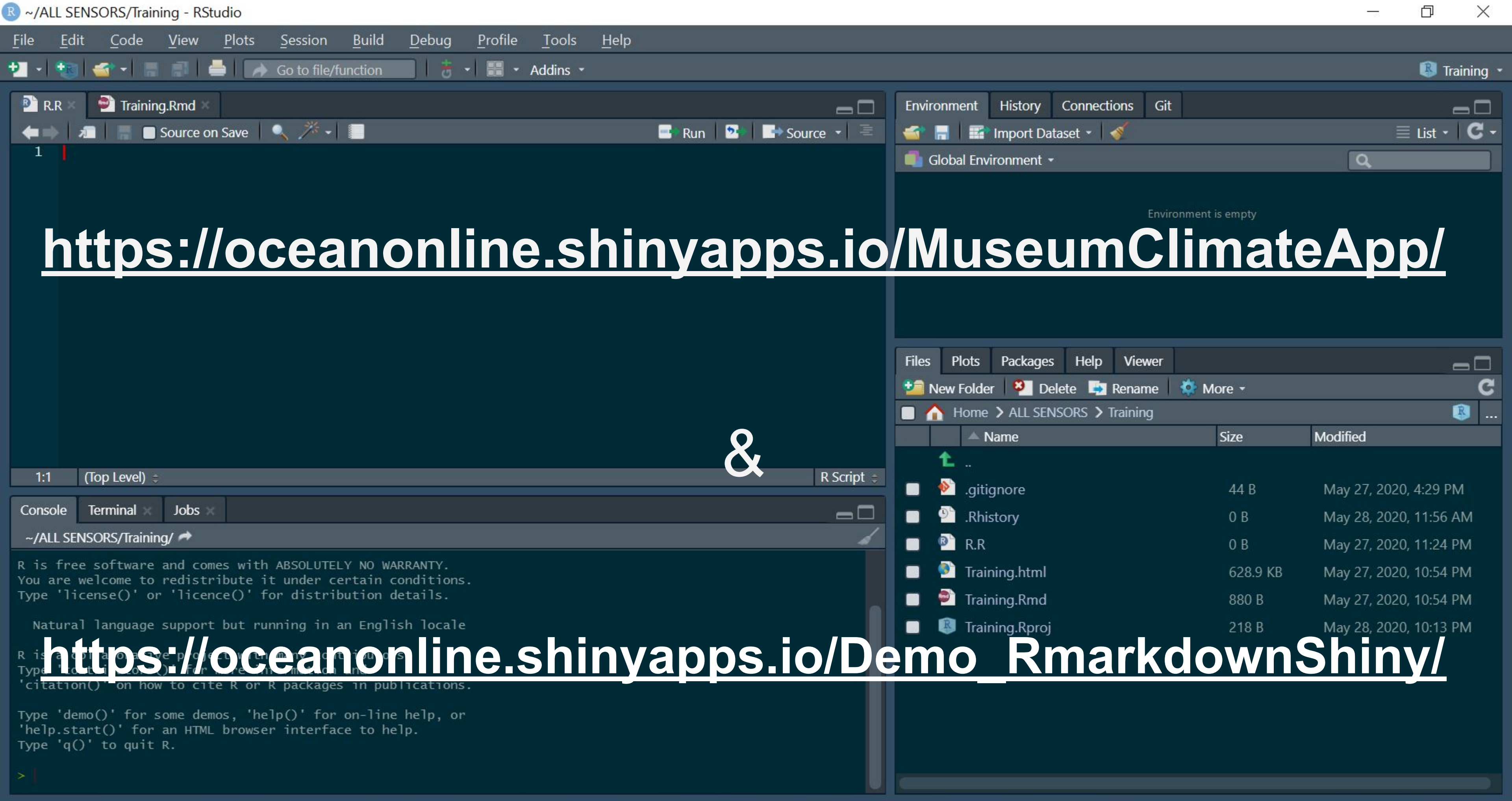


```
R is free
You are we
Type 'lice

  Natural

R is a co
Type 'con
'citation

Type 'dem
'help.sta
Type 'q()
>
```

Ingredients
(Data)

Tweek

Outputs(and Help)



**https://oceanonline.shinyapps.io/MuseumClimateApp/**

&

**https://oceanonline.shinyapps.io/Demo_RmarkdownShiny/**

| | Name | Size | Modified |
|---|---|---|---|
| | .. | | |
| □ | .gitignore | 44 B | May 27, 2020, 4:29 PM |
| □ | .Rhistory | 0 B | May 28, 2020, 11:56 AM |
| □ | R.R | 0 B | May 27, 2020, 11:24 PM |
| □ | Training.html | 628.9 KB | May 27, 2020, 10:54 PM |
| □ | Training.Rmd | 880 B | May 27, 2020, 10:54 PM |
| □ | Training.Rproj | 218 B | May 28, 2020, 10:13 PM |

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

# The data ingredients

Understanding data is the most important concept:

- Data types
- Data structures
- Data grammar
- How data interacts
- Volume, Veracity, Variety

# Understanding the ingredients

Alphabet - the building blocks

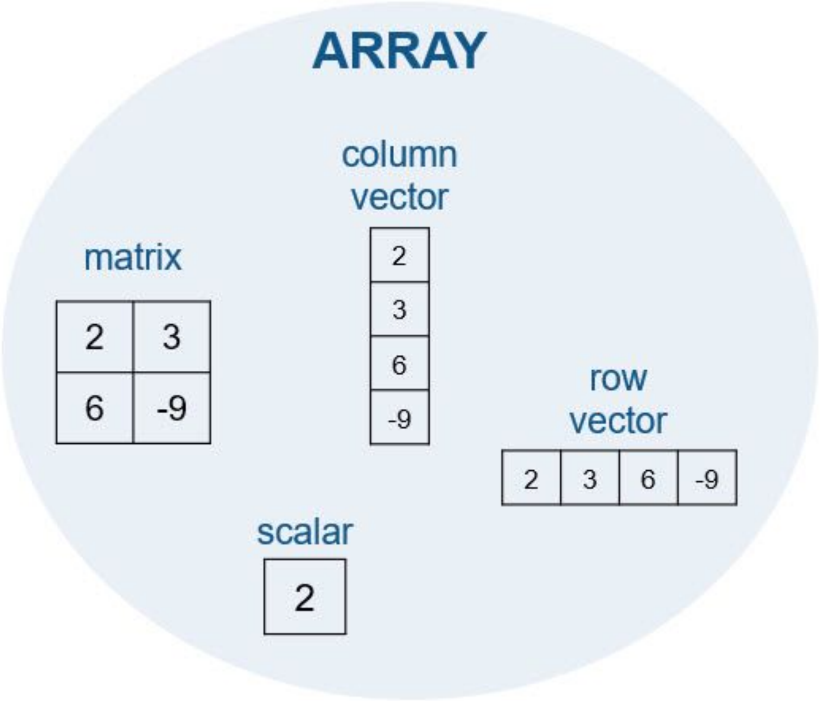
Nouns - also known as ‘objects’

Verbs - the ‘recipes’ and the bits you code

Alphabet - numbers, characters, dates, missing, logical,

Sentences - vectors, matrices, arrays, lists, data frames

Spelling and grammar - are important



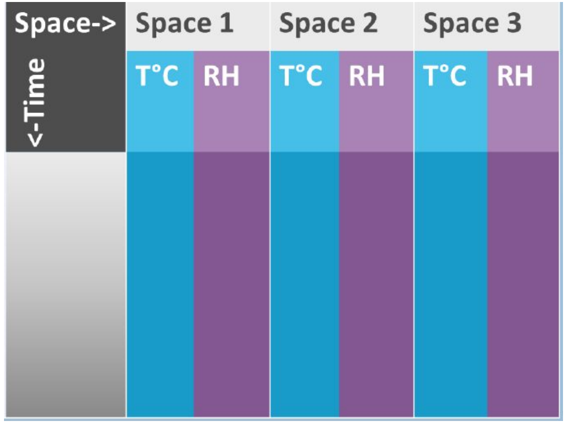
# Preparing the ingredients

“Happy families are all alike; every unhappy family is unhappy in its own way.” – Leo Tolstoy

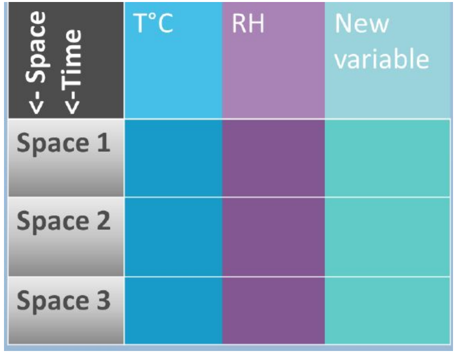
“Tidy datasets are all alike, but every messy dataset is messy in its own way.” – Hadley Wickham

Each **variable** is in its own **column**

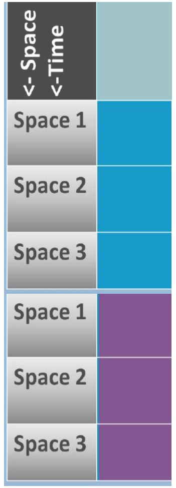
Each **observation**, or **case**, is in its own **row**



| Space-> | Space 1 | | Space 2 | | Space 3 | |
|---|---|---|---|---|---|---|
| <-Time | T°C | RH | T°C | RH | T°C | RH |
| | | | | | | |



| <- Space <-Time | T°C | RH | New variable |
|---|---|---|---|
| Space 1 | | | |
| Space 2 | | | |
| Space 3 | | | |

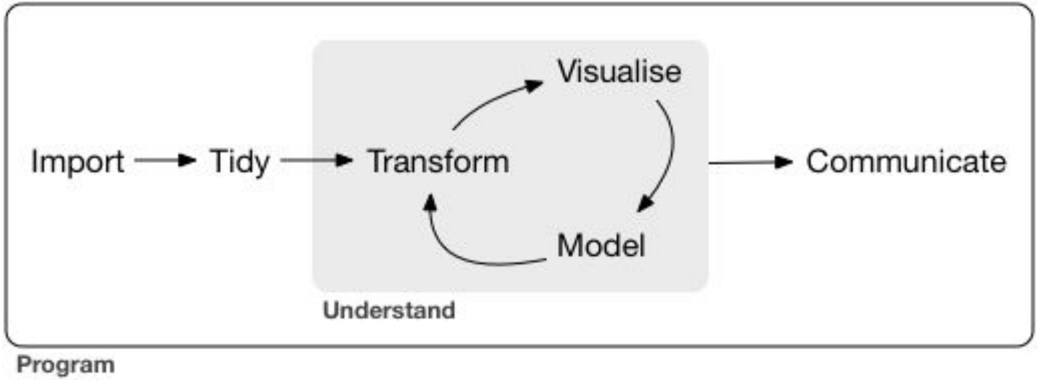


| <- Space <-Time | |
|---|---|
| Space 1 | |
| Space 2 | |
| Space 3 | |
| Space 1 | |
| Space 2 | |
| Space 3 | |

# Writing recipes

Decide what you want to do, who is your audience (your client) and work backwards

Be flexible and adaptive - as you'll likely end up somewhere else



# Icing on the cake

Final pun

Modelling

Machine learning

Graphing

Reporting

Interactives

Online

# Icing on the cake

Final pun

Modelling

Machine learning

Graphing

Reporting

Interactives

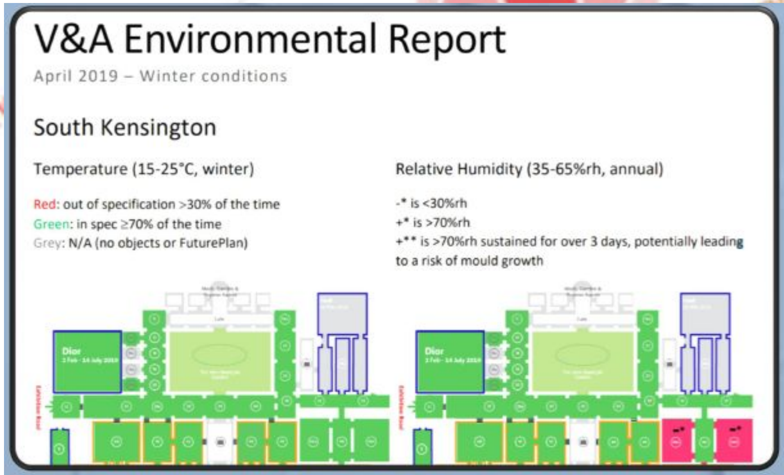
Online

**Statistics**

Known data
Explanatory power

**Machine learning**

Unseen data
Predictive power



## V&A Environmental Report

April 2019 – Winter conditions

South Kensington

Temperature (15-25°C, winter)

Red: out of specification >30% of the time
Green: in spec ≥70% of the time
Grey: N/A (no objects or FuturePlan)

Relative Humidity (35-65%rh, annual)

-* is <30%rh
+* is >70%rh
+** is >70%rh sustained for over 3 days, potentially leading to a risk of mould growth

# Pros and cons

Free

Community support

10,000 packages

Development

Transferable and high demand skill (and it pays)

The future

Hard

Need to keep learning

Coding ‘rage’

# Tips

Give yourself plenty of time

Think of your future self - Comment as you go along

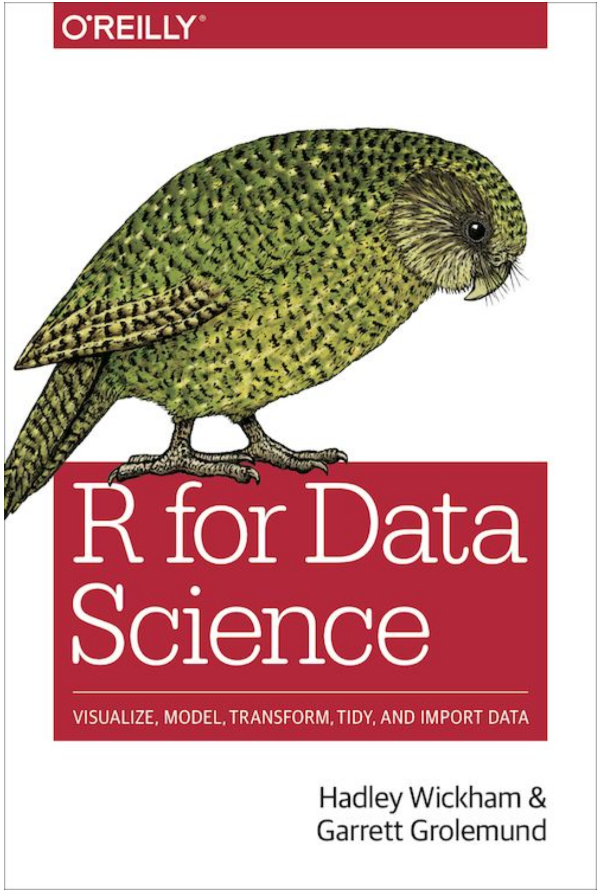
Seek help before code rage!

StackOverflow, Blogs, Twitter, Meetups, Cheat Sheets, YouTube, Kids coding books

-these are your friends

Good place to start: **https://r4ds.had.co.nz/**

Google: **“R for data science Hadley”**



Follow instructions at RStudio for how to download and get started:

https://www.vam.ac.uk/blog/author/bhavesh-shah

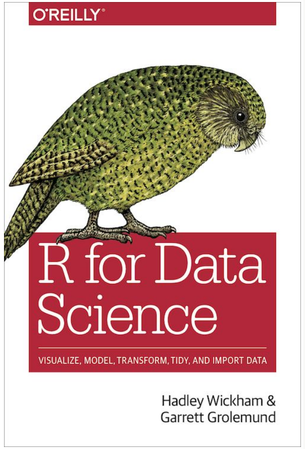
RStudio®

UCL

codecademy

LEARN TO CODE
WITH THESE
COURSERA
PROGRAMMING COURSES

LEARNTOCODEWITH.ME



O'REILLY®

# R for Data Science

VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

Hadley Wickham &
Garrett Grolemund

RaspberryPi

DataCamp

Udemy

# Thankyou

https://github.com/BhavShah01/Training

b.shah@vam.ac.uk

@bhav_shah