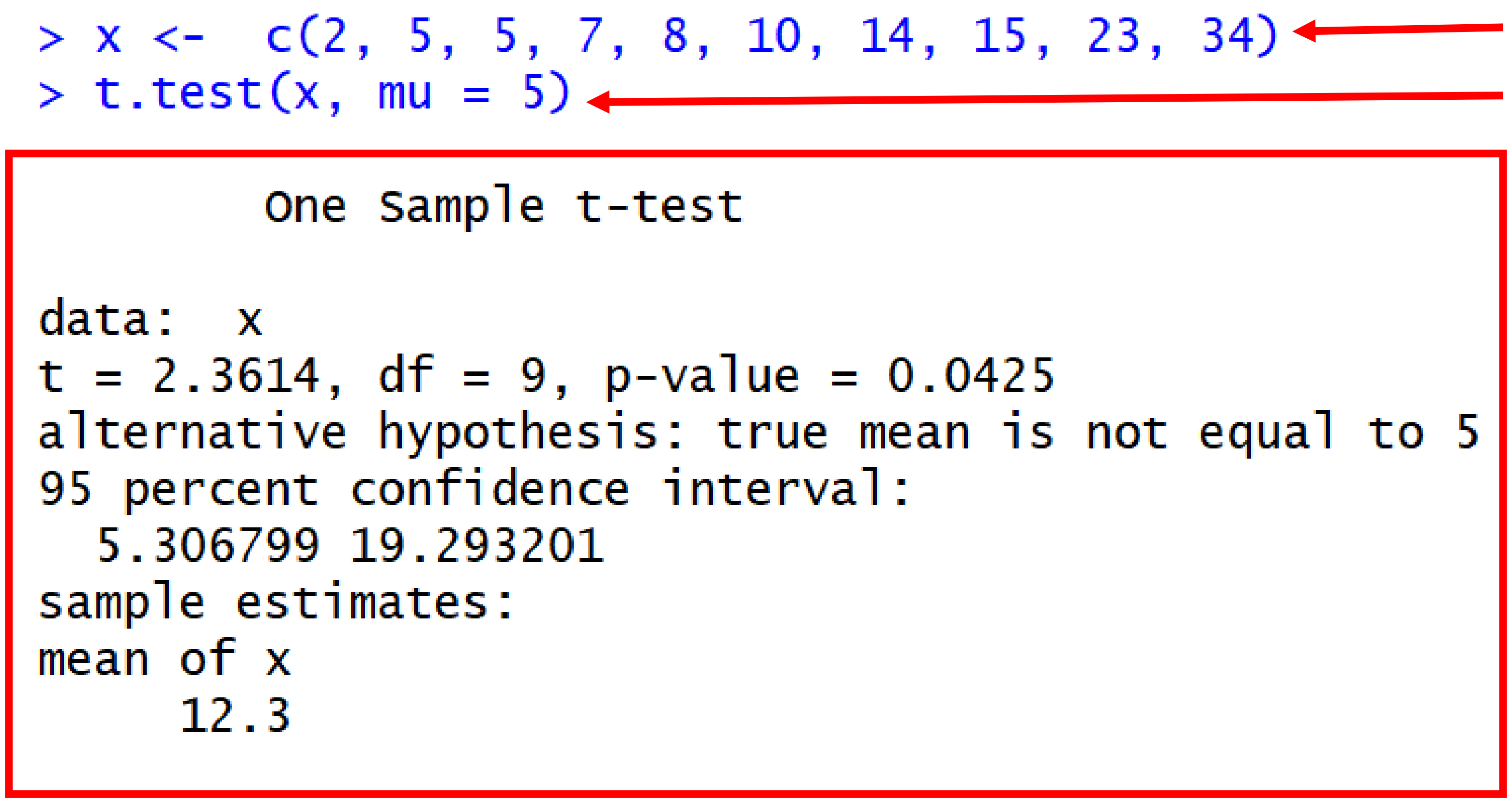
Workshop 6
The tidyverse and beyond

- Don't look back in anger

Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

R is a statistical programming language

- While the input data to models needs to be tidy, unfortunately the models outputs are less than neat



```
> x <-  c(2, 5, 5, 7, 8, 10, 14, 15, 23, 34)
> t.test(x, mu = 5)

        One Sample t-test

data:  x
t = 2.3614, df = 9, p-value = 0.0425
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
  5.306799 19.293201
sample estimates:
mean of x
     12.3
```

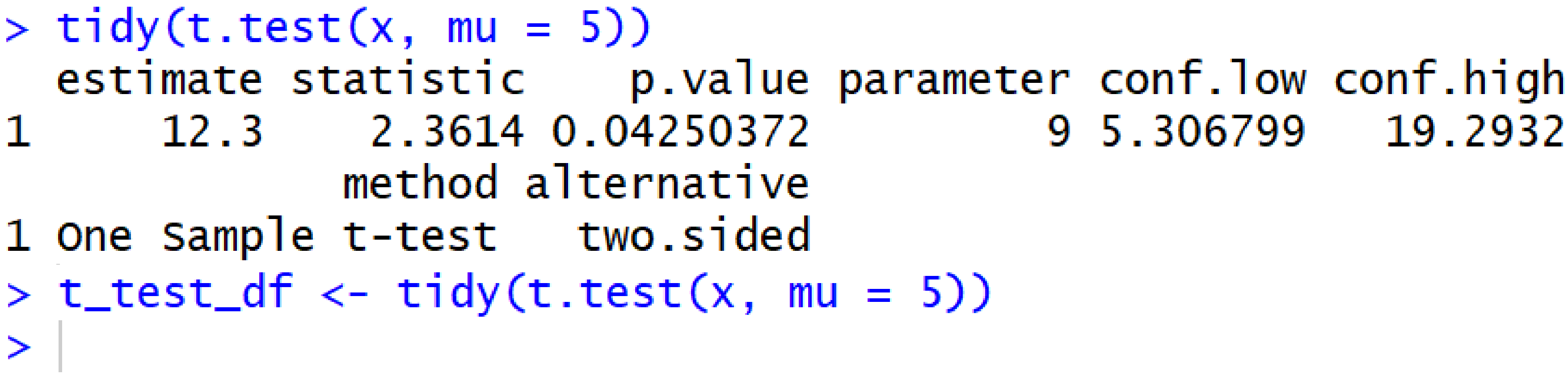
Input data

Function call

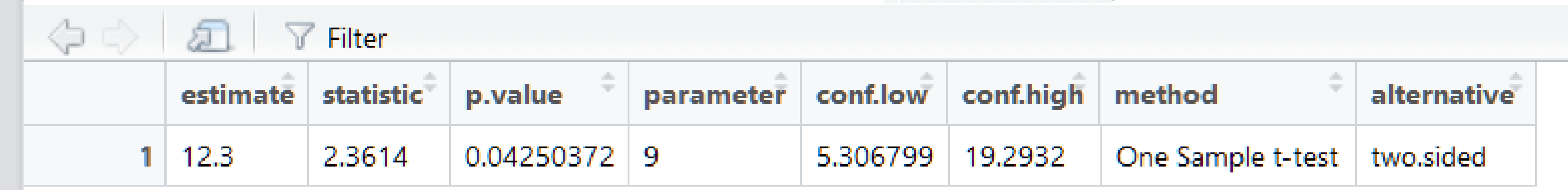
Output

The broom package tidies up the output

- The broom package will take the outputs of your test and place them in an easy to access table



```
> tidy(t.test(x, mu = 5))
  estimate statistic    p.value parameter conf.low conf.high
1     12.3    2.3614 0.04250372         9 5.306799   19.2932
             method alternative
1 One Sample t-test   two.sided
> t_test_df <- tidy(t.test(x, mu = 5))
> 
```



	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1	12.3	2.3614	0.04250372	9	5.306799	19.2932	One Sample t-test	two.sided

- The modelr package combines base R modelling with %>%
- To gain more of an insight into model building, I recommend working through Part IV of “R for Data Science”

Worksheets

ws6_script1_stats_basics.R

ws6_script2_model_outputs.R

Inconsistent function names

- R is a very versatile language
  - The main drawback of this versatility for beginners is the variety of ways to do the same task
  - Often a painful learning curve

```
names, colnames
row.names, rownames
rowSums, rowsum
browseURL, contrib.url, fixup.package.URLs
package.contents, packageStatus
getMethod, getS3method
read.csv and write.csv, load and save, readRDS and saveRDS
Sys.time, system.time
```

Variable selection

```
summary(starwars$name)

summary(starwars$"name")

summary(starwars["name"])

summary(starwars[,"name"])

summary(starwars[["name"]])

summary(starwars[1])

summary(starwars[,1])

summary(starwars[[1]])
```

Worksheet

Open ws6_script3_too_much_choice.R

Writing R scripts to make them reusable

Consistency

- Throughout this course we've discussed how the tidyverse is more human readable

- Also, the functions are designed to do one task well

- The underlying syntax is simplified and consistent

- This does not mean that the choice has disappeared

```
summary(select(starwars, names))
starwars %>% with(summary(names))
starwars %>% summary(.$names)
starwars %>% summary(names)
```

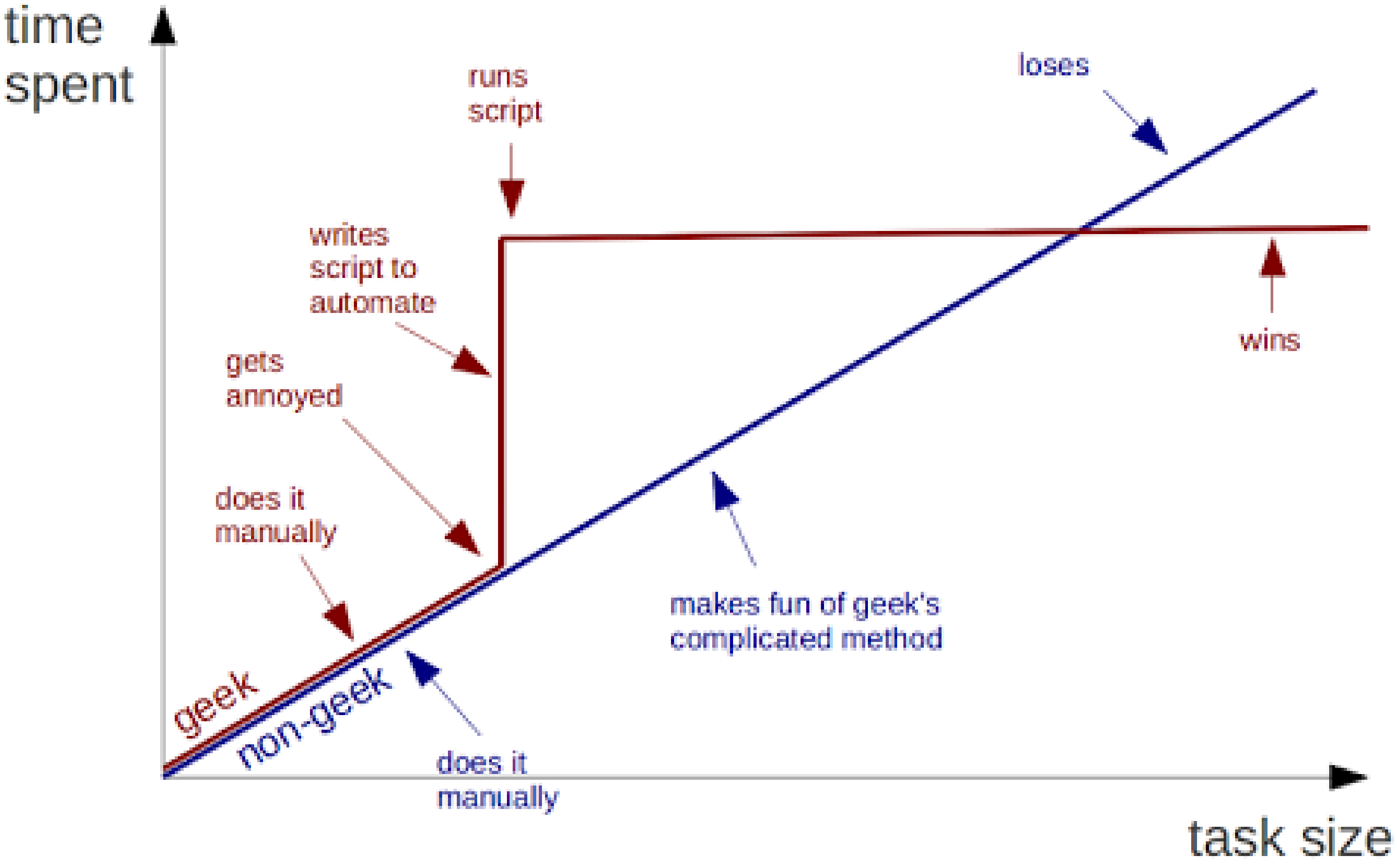
Style guide

http://style.tidyverse.org/syntax.html

Worksheet

Open ws6_script4_good_habits.R

Why bother?

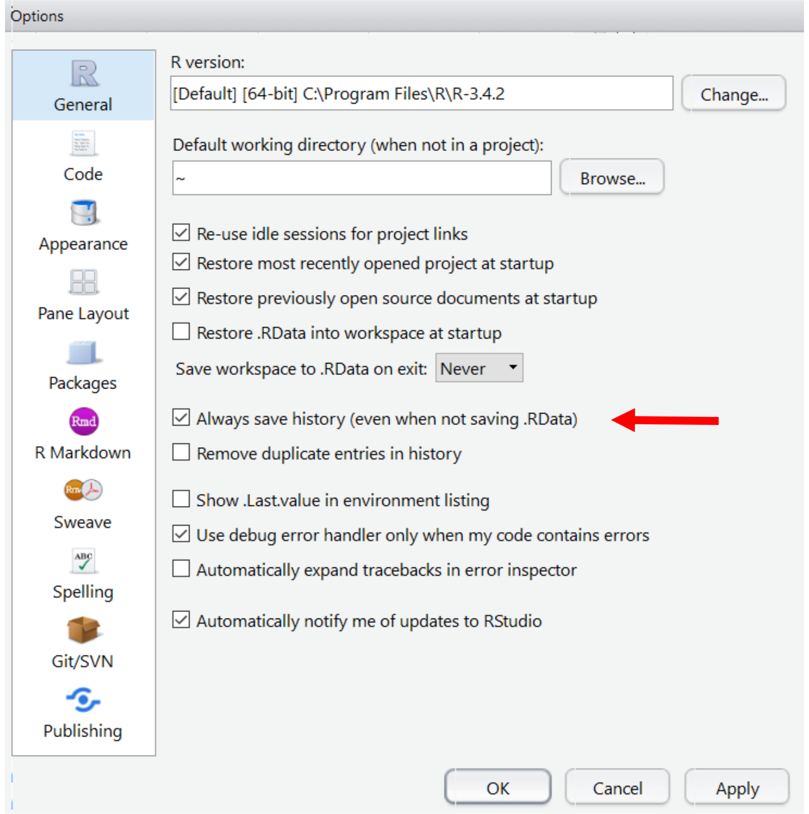


https://nicercode.github.io/blog/2013-04-05-projects/

Other points to note

- You might consider your environment as “real”

- If you continue to use R, it is better for you to consider your R scripts as “real”, as these should recreate the environment



R version:

[Default] [64-bit] C:\Program Files\R\R-3.4.2

Default working directory (when not in a project):

~

☑ Re-use idle sessions for project links
☑ Restore most recently opened project at startup
☑ Restore previously open source documents at startup
☐ Restore .RData into workspace at startup

Save workspace to .RData on exit: Never

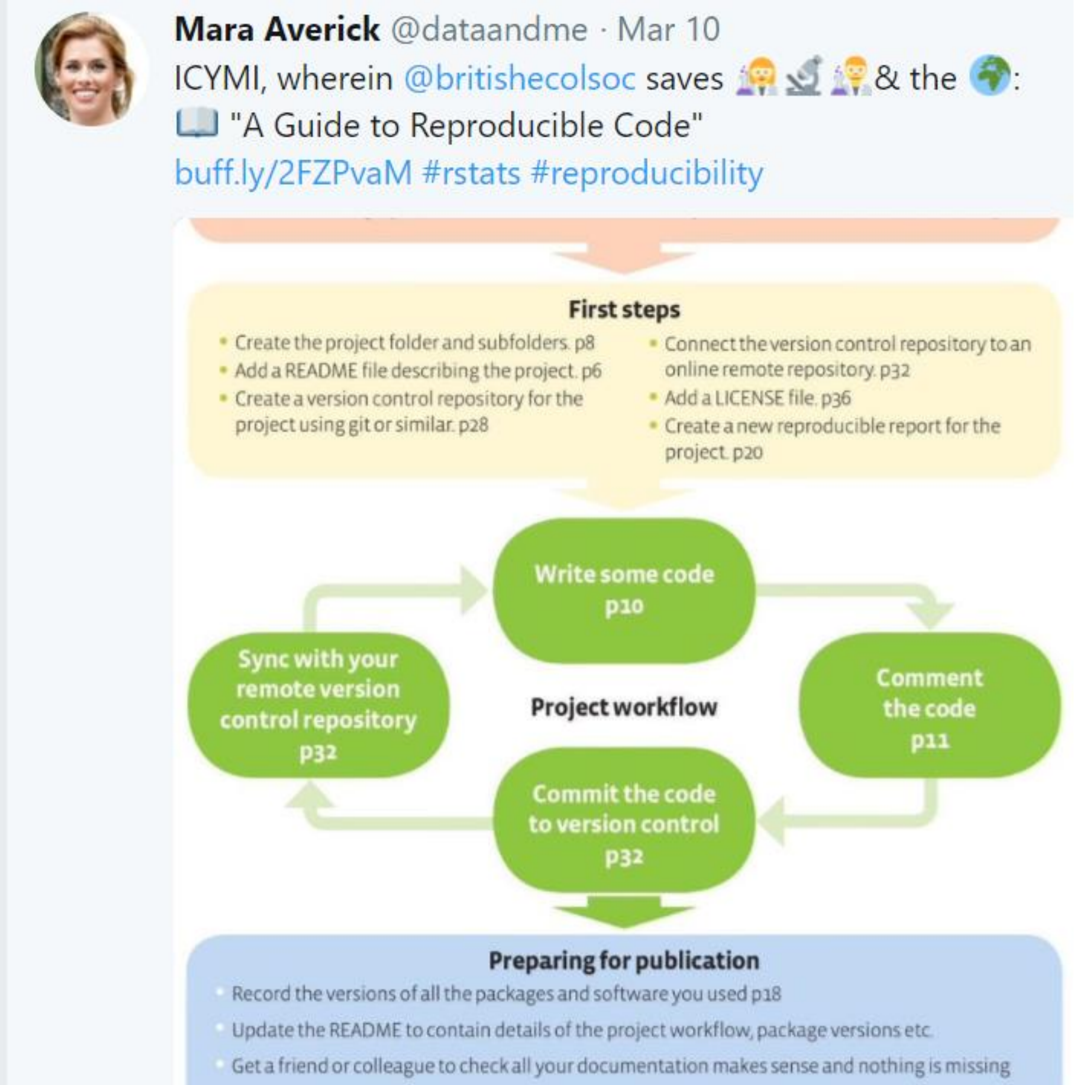
☑ Always save history (even when not saving .RData)
☐ Remove duplicate entries in history

☐ Show .Last.value in environment listing
☑ Use debug error handler only when my code contains errors
☐ Automatically expand tracebacks in error inspector

☑ Automatically notify me of updates to RStudio

- You may suffer short term pain

- This will prevent long term agony



Mara Averick @dataandme · Mar 10

ICYMI, wherein @britishecolsoc saves 👩‍🔬🔬👨‍🔬 & the 🌍:
📖 "A Guide to Reproducible Code"
buff.ly/2FZPvaM #rstats #reproducibility

First steps

- Create the project folder and subfolders. p8
- Add a README file describing the project. p6
- Create a version control repository for the project using git or similar. p28
- Connect the version control repository to an online remote repository. p32
- Add a LICENSE file. p36
- Create a new reproducible report for the project. p20

Project workflow

Write some code p10

Comment the code p11

Commit the code to version control p32

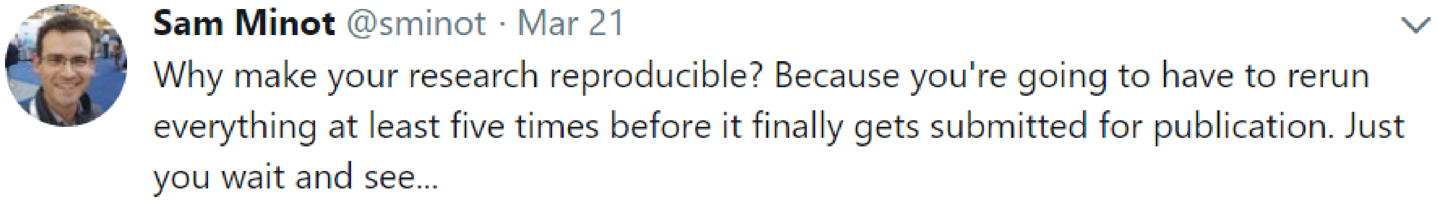
Sync with your remote version control repository p32

Preparing for publication

- Record the versions of all the packages and software you used p18
- Update the README to contain details of the project workflow, package versions etc.
- Get a friend or colleague to check all your documentation makes sense and nothing is missing

Matt Motyl @MattMotyl · Mar 23

I'm writing my first research report using RMarkdown and it is figuratively blowing my mind. I can't believe how much time I've wasted in my life transcribing data/analyses into word documents. Game changer. #awe



Sam Minot @sminot · Mar 21

Why make your research reproducible? Because you're going to have to rerun everything at least five times before it finally gets submitted for publication. Just you wait and see...

R Markdown

- R Markdown combines the code you wrote, the output produced and your own comments

- You can view it as a digital lab notebook, where you are both recording what you're doing, and what you were thinking while you were doing it!

- R Markdown outputs can take many forms
    - Word documents, PDFs, slideshows etc.

- Once created the .Rmd file get sent to knitr, which executes the chunks of code and creates a new markdown document (.md)
    - this is then processed by pandoc which creates the finished file
        - knitr and pandoc are external websites

R Markdown

YAML header

````
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
````

Chunks of code

~~~
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
~~~

Plain text with integrated outputs from R

````
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
````

Chunks of code

~~~
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
~~~

Worksheet
ws6_script5_Rmarkdown_example.R

Open http://rpubs.com/bpalmer/337383

Follow the instructions at the bottom of the webpage link and have ago at creating your own R Markdown document

Worksheet

Open ws6_script6_writing_scripts.R

- This script outlines the various steps to work on at your own pace

- Open a blank script and populate that with your code

- Try to do each step independently

- Once you've succeeded, attempt to pull them all together using "%>%" where feasible

- Include some informative sentences to help make the code more understandable should you need it in the future
    - i.e. tricky steps that required workarounds
    - details about the data and the steps needed to process it

WHO dataset

- contains tuberculosis (TB) cases by year, country, age etc.

- Typical real life messy data set

Tips:

- country, iso2, iso3 redundantly specify the country
- You’ll need to gather together all the columns from “new_ep_f014” to “new_sp_m65”
- These columns are likely to be values and not variables
- Examine your data frame as you go
- Once tidied, decide on elements you’d like to examine
  - e.g. data by country, by age etc.
  - group the data, summarise it
- produce some graphical
- add layers, titles, legends etc. to your graphs

Worksheet

Open ws6_script7_sample_analysis.R

Where to next?

- Keep using R
    - The more you practise/use it, the easier it becomes

- If you haven't already, join the **Cork R Meet-Up group**
    - If you want another workshop around a topic specific for your work, we can help organise that!
    - https://www.meetup.com/Cork-Ireland-R-Users-Group/

- Find the course that meets your needs and do it at your own pace
    - then another
        - then another

- Follow the R community online and engage with it:
    - Twitter
    - https://community.rstudio.com/tags/teaching

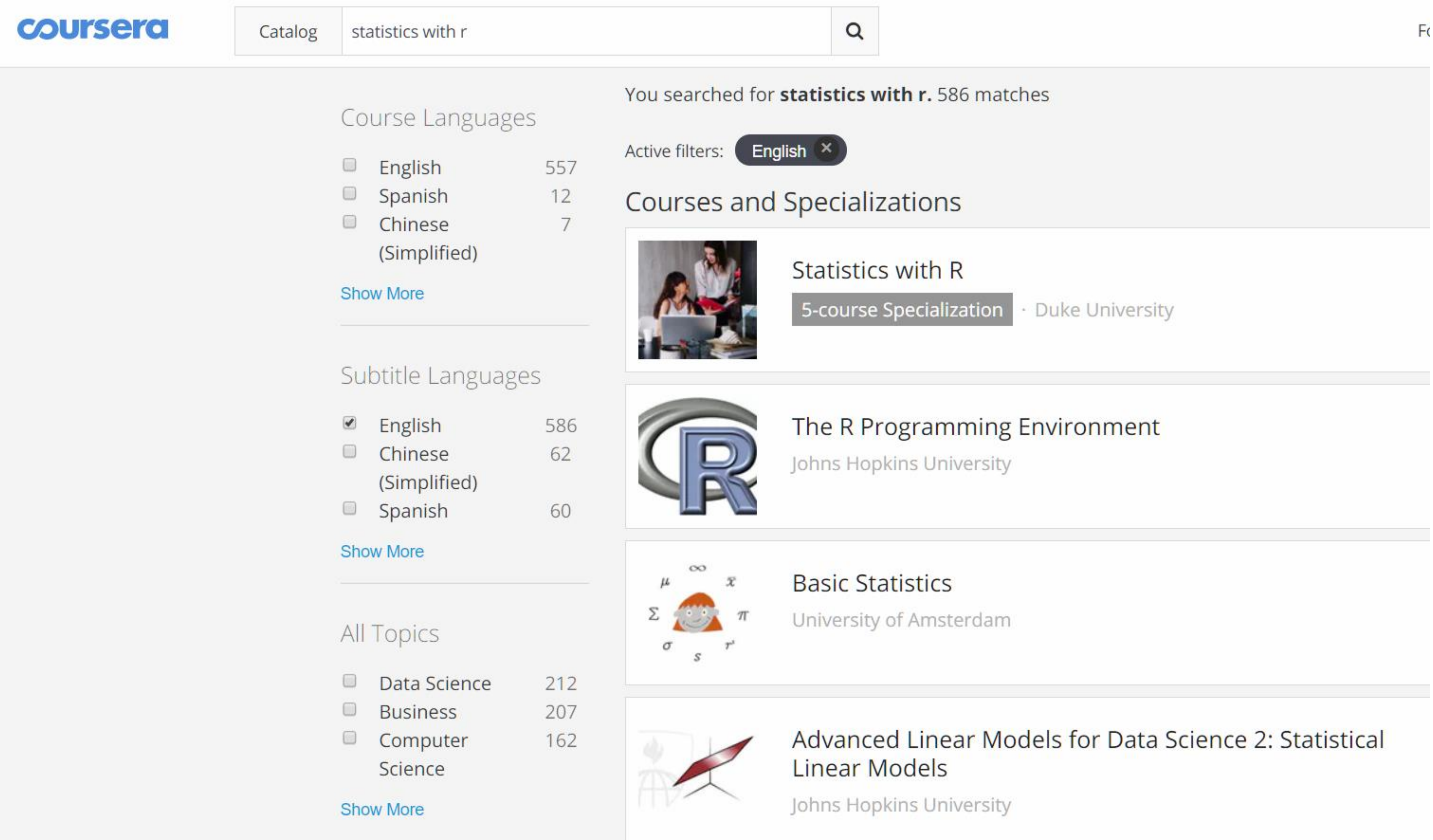
Where to next?

- Understanding basic statistical concepts
www.khanacademy.org

- Collection of YouTube videos describing statistics through R
http://rafalab.github.io/pages/harvardx.html

- You know what you want to do, but don't know how to do it
https://stats.stackexchange.com/

Structured training…….



You searched for **statistics with r.** 586 matches

Active filters: English ✕

Course Languages

☐	English	557
☐	Spanish	12
☐	Chinese (Simplified)	7

Show More

Subtitle Languages

☑	English	586
☐	Chinese (Simplified)	62
☐	Spanish	60

Show More

All Topics

☐	Data Science	212
☐	Business	207
☐	Computer Science	162

Show More

Courses and Specializations

Statistics with R

5-course Specialization · Duke University

The R Programming Environment

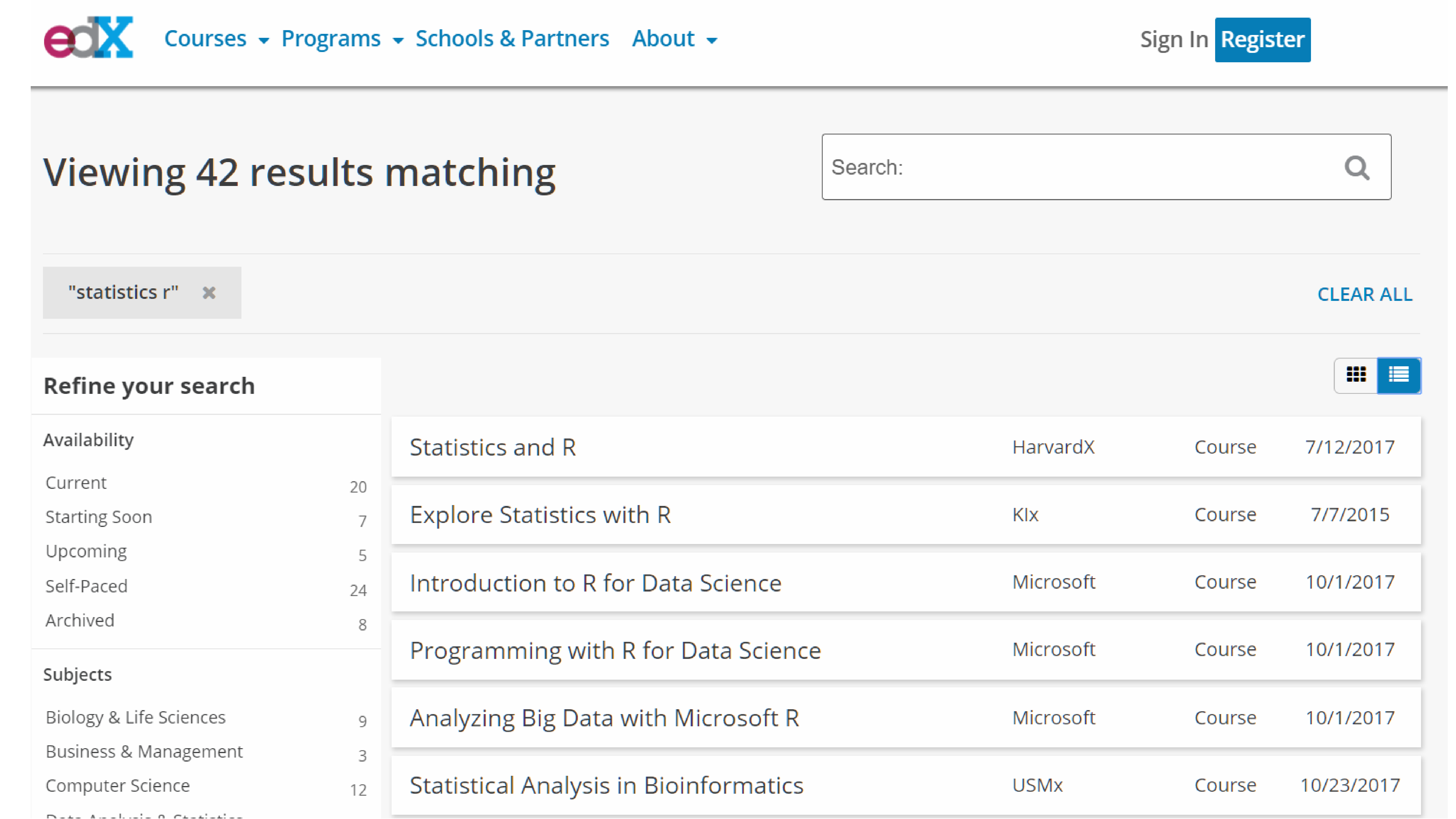
Johns Hopkins University

Basic Statistics

University of Amsterdam

Advanced Linear Models for Data Science 2: Statistical Linear Models

Johns Hopkins University



Viewing 42 results matching

Refine your search

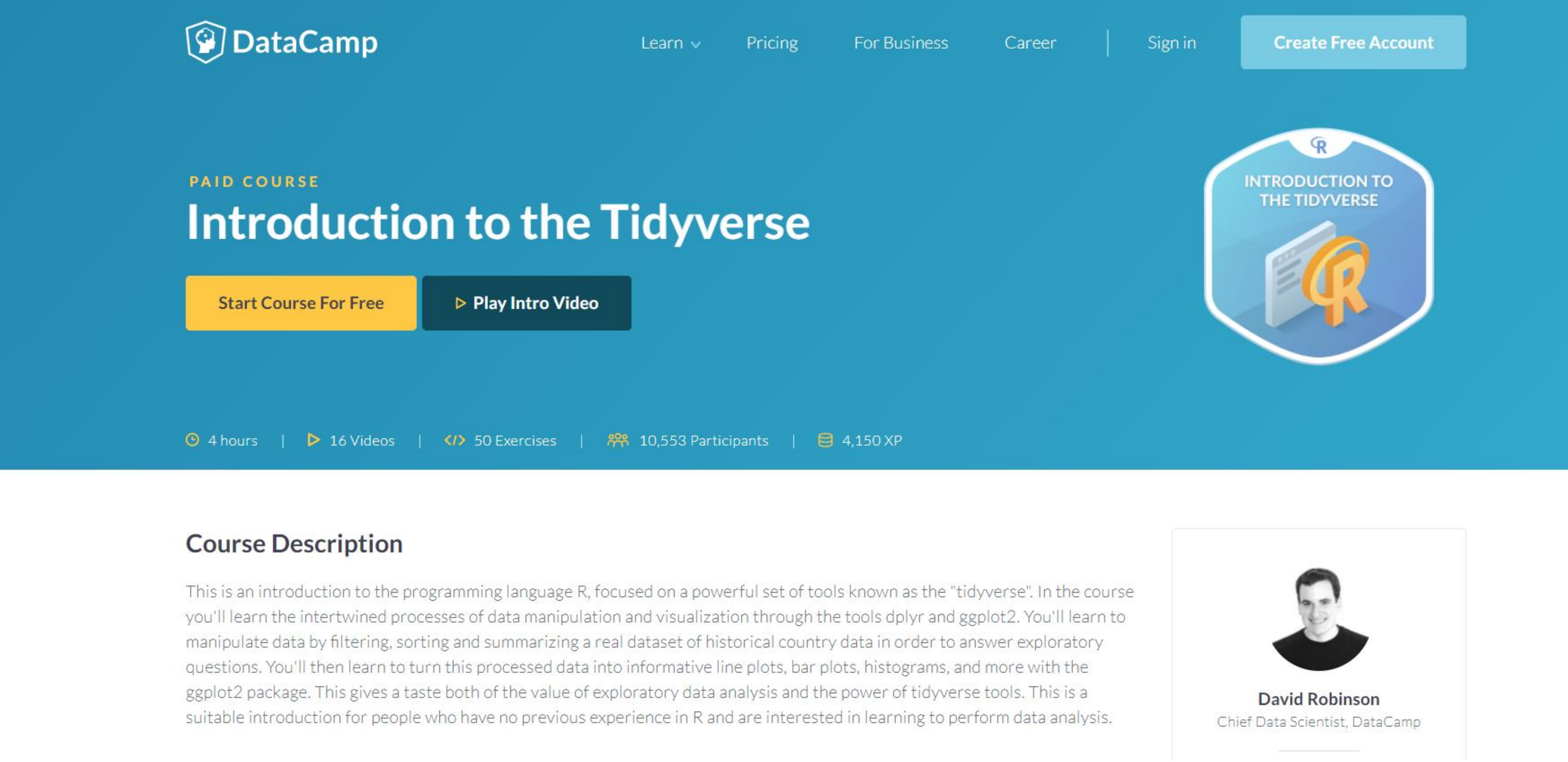
Availability

Current	20
Starting Soon	7
Upcoming	5
Self-Paced	24
Archived	8

Subjects

Biology & Life Sciences	9
Business & Management	3
Computer Science	12
Data Analysis & Statistics	

Statistics and R	HarvardX	Course	7/12/2017
Explore Statistics with R	KIx	Course	7/7/2015
Introduction to R for Data Science	Microsoft	Course	10/1/2017
Programming with R for Data Science	Microsoft	Course	10/1/2017
Analyzing Big Data with Microsoft R	Microsoft	Course	10/1/2017
Statistical Analysis in Bioinformatics	USMx	Course	10/23/2017



PAID COURSE

Introduction to the Tidyverse

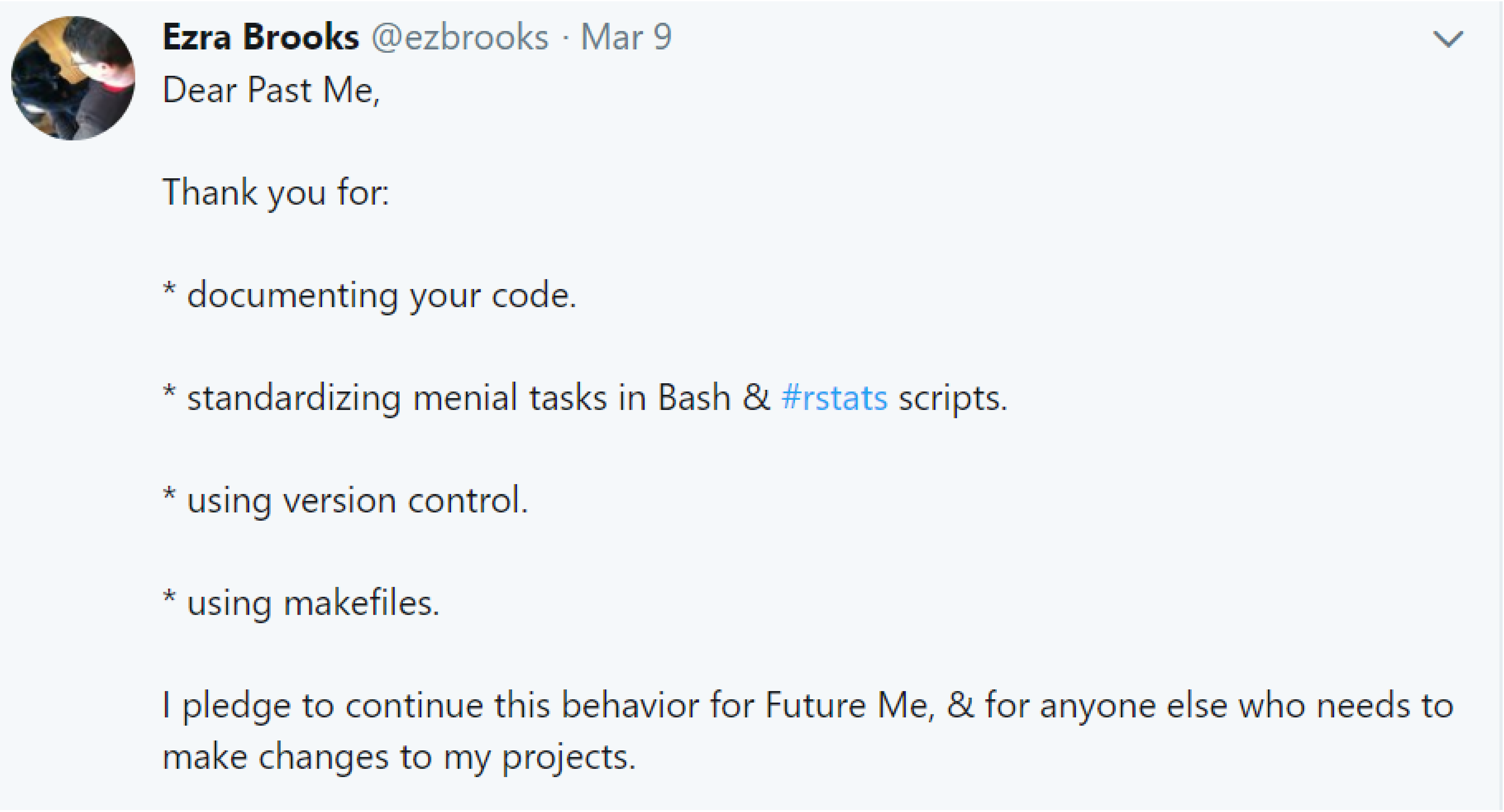
4 hours | 16 Videos | 50 Exercises | 10,553 Participants | 4,150 XP

Course Description

This is an introduction to the programming language R, focused on a powerful set of tools known as the "tidyverse". In the course you'll learn the intertwined processes of data manipulation and visualization through the tools dplyr and ggplot2. You'll learn to manipulate data by filtering, sorting and summarizing a real dataset of historical country data in order to answer exploratory questions. You'll then learn to turn this processed data into informative line plots, bar plots, histograms, and more with the ggplot2 package. This gives a taste both of the value of exploratory data analysis and the power of tidyverse tools. This is a suitable introduction for people who have no previous experience in R and are interested in learning to perform data analysis.

David Robinson
Chief Data Scientist, DataCamp

In conclusion



Ezra Brooks @ezbrooks · Mar 9

Dear Past Me,

Thank you for:

* documenting your code.

* standardizing menial tasks in Bash & #rstats scripts.

* using version control.

* using makefiles.

I pledge to continue this behavior for Future Me, & for anyone else who needs to make changes to my projects.