Institute
and Faculty
of Actuaries

Reproducible data science techniques in actuarial work

What can actuaries learn from open science and other professions?

Philip Darke FIA, Mercer and Newcastle University
Dr Matthew Forshaw, Newcastle University

February 2019

Reproducible data science techniques for actuaries

We often focus on the glamourous side of data science: big data sets, artificial intelligence and the replacement of humans by machine.

This presentation considers data science from a different perspective – how do data scientists work?

We make the case that actuaries can learn from the working methods used by data scientists and apply them to existing actuarial work, as well as when adopting data science techniques.

Content

- Case studies from UK Government and the BBC News website
- How reproducibility can help address the challenges of working in data intensive fields
- Why this is important for actuaries
- A short introduction to some of the tools available

This presentation is accompanied by exercises covering some of the techniques and tools introduced.

Institute
and Faculty
of Actuaries

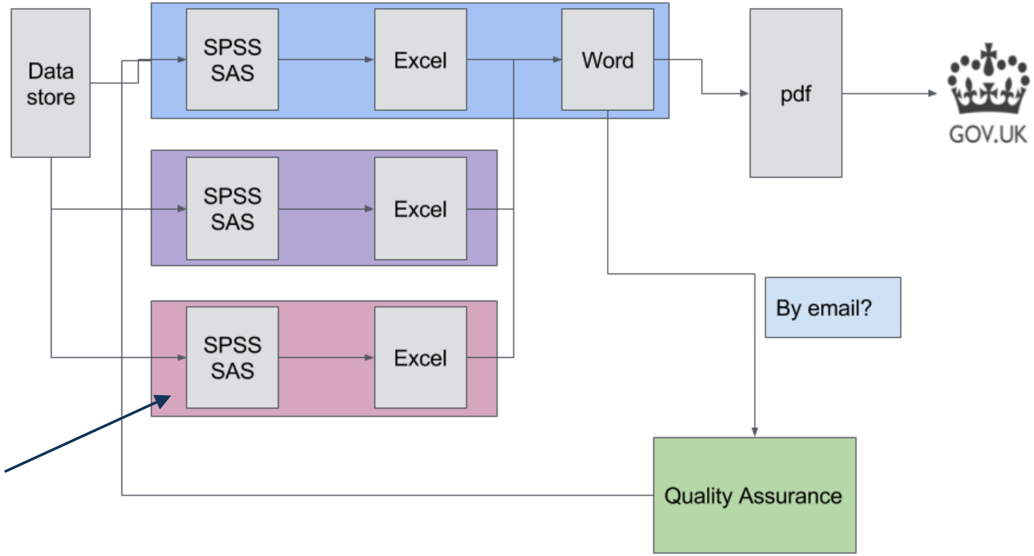
Case study

Producing official statistics in UK Government

Producing official statistics in government

Data scientists at the UK Government Digital Service have worked with a number of government departments to transform the production of official statistics.

In general, analytical processes in government involve a mixture of manual and semi-manual processes looking something like this:



Proprietary software

https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process

Government statistics – the existing process

This process varies across government departments and projects but it typically involves:

- Extracting data from databases, spreadsheets and other sources
- Processing the data and carrying out analysis in proprietary software
- Manually copying results to Excel for further analysis and the production of tables/graphs
- Copying and pasting output to a Word document for reporting and formatting
- Converting the report to a PDF suitable for publication
- Along the way, documents might be emailed between colleagues resulting in multiple versions of files

What do your existing analytical processes look like? Does the above look familiar?

Government statistics – problems with this approach

The Government Digital Service recognised a number of issues with this approach:

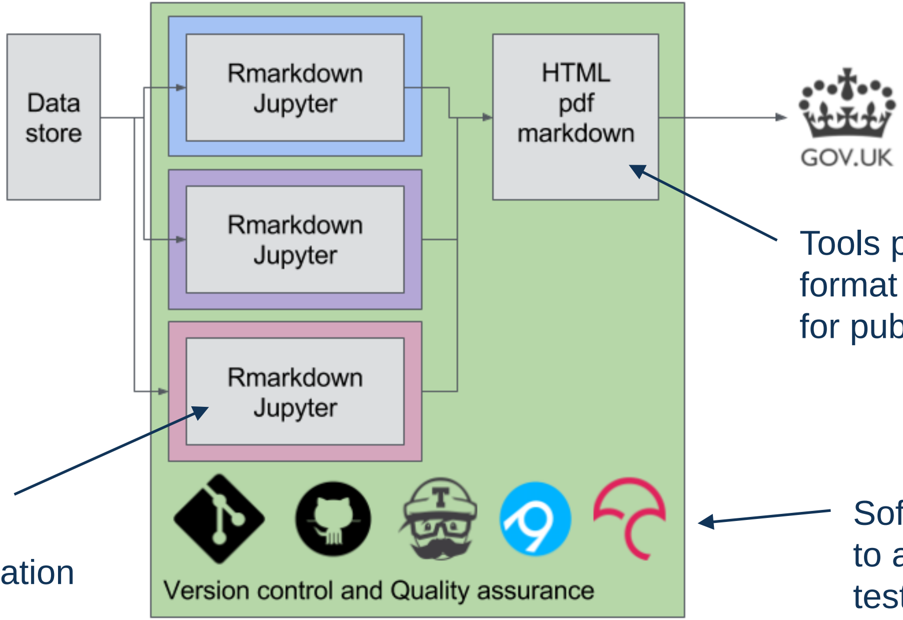
- Heavy reliance on manual processes is time consuming
- Errors in spreadsheets are common due to human error and are often difficult to catch
- Further errors can be introduced when copying data/output between software tools
- It is difficult to keep an audit trail of the process when working in a team (i.e. who did what)
- Checking and testing often happens at the end of the process, rather than embedded throughout
- Any changes made following checking/peer review require the manual process to be repeated – at the risk of introducing further errors

They also identified the difficulty in reproducing the results of previous reports.

Are you confident you could easily reproduce your last board paper/client report/statistical analysis?

Government statistics – their solution

To try and solve these problems, they took ideas from software development and academia, using open source tools to devise a pipeline that reduced production time whilst arguably improving the quality of the output. It looks like this:



Tools produce and automatically format reports and web pages for publication

Open source software that combines analysis, reporting and documentation

Software development tools to automate the audit trail and testing process

https://ukgovdatascience.github.io/rap_companion/why.html#desired-reproducible-analytical-pipeline

Reproducible analytical pipelines

The Government Digital Service call this approach *Reproducible Analytical Pipelines* (RAP):

- Data is loaded into software once (no cutting and pasting!)
- One tool is used for all data processing, analysis, production of tables/graphs and report writing
- Tools and techniques from software development are used to automate testing and maintain an audit trail

> “With RAP, analytical teams can automate time-consuming processes of data assembly, verification and integration, generate charts and tables, and set up and populate statistical reports. **The potential time savings for analysts are enormous, freeing them up to focus on the interpretation of the results.** The other huge benefit comes from building a process that is fully transparent, auditable and verifiable – **reducing risk and improving quality.**
>
> Matt Upson and Mat Gregory, Government Digital Service ”

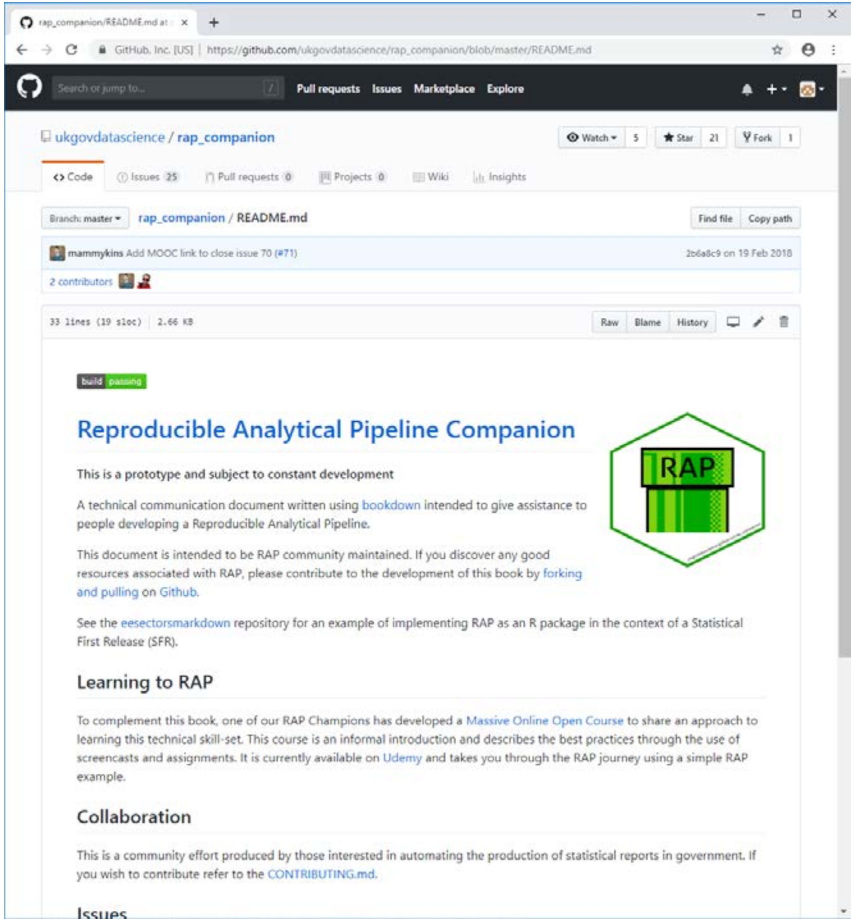
Government statistics – further reading

- Blog posts here and here introduce the RAP project

- A free online book, the RAP Companion, is available here that explains the process in detail

- In addition, the data scientists involved in this project created a free online course at Udemy

- An example statistical release using RAP can be found on GitHub here



rap_companion / README.md

mammykins Add MOOC link to close issue 70 (#71) 2b6a8c9 on 19 Feb 2018

2 contributors

33 lines (19 sloc) | 2.66 KB

Reproducible Analytical Pipeline Companion

This is a prototype and subject to constant development

A technical communication document written using bookdown intended to give assistance to people developing a Reproducible Analytical Pipeline.

This document is intended to be RAP community maintained. If you discover any good resources associated with RAP, please contribute to the development of this book by forking and pulling on Github.

See the eesectorsmarkdown repository for an example of implementing RAP as an R package in the context of a Statistical First Release (SFR).

Learning to RAP

To complement this book, one of our RAP Champions has developed a Massive Online Open Course to share an approach to learning this technical skill-set. This course is an informal introduction and describes the best practices through the use of screencasts and assignments. It is currently available on Udemy and takes you through the RAP journey using a simple RAP example.

Collaboration

This is a community effort produced by those interested in automating the production of statistical reports in government. If you wish to contribute refer to the CONTRIBUTING.md.

Issues

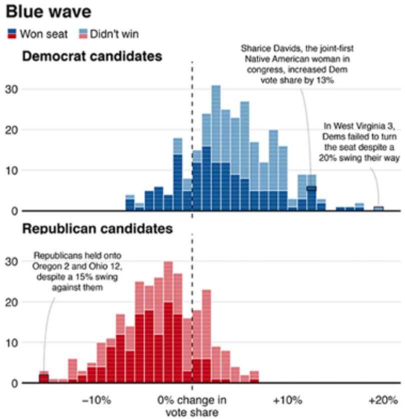
Institute
and Faculty
of Actuaries

Case study

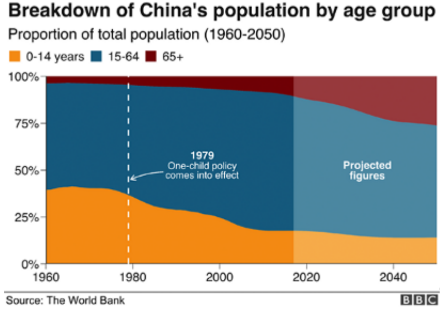
Producing graphics for the BBC News website

BBC News graphics

The BBC visual and data journalism team have used a similar approach to fundamentally change how they produce graphics for the BBC News website.



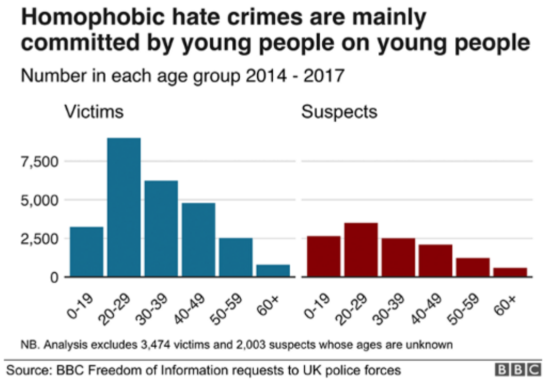
Blue wave



Breakdown of China's population by age group

Proportion of total population (1960-2050)

Source: The World Bank

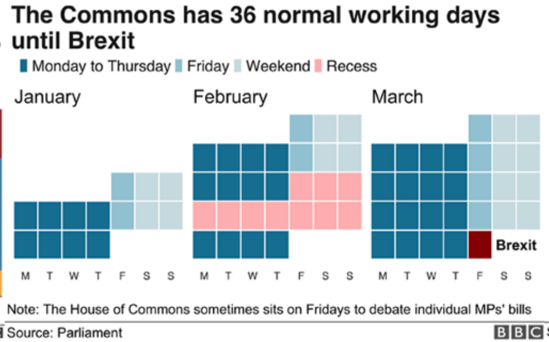


Homophobic hate crimes are mainly committed by young people on young people

Number in each age group 2014 - 2017

NB. Analysis excludes 3,474 victims and 2,003 suspects whose ages are unknown

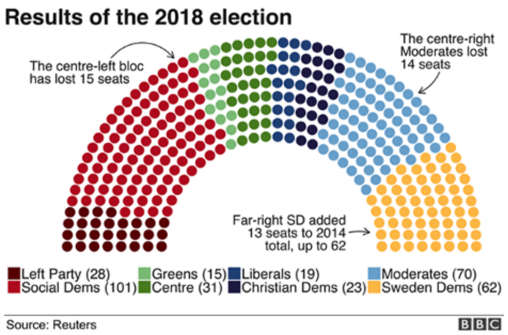
Source: BBC Freedom of Information requests to UK police forces



The Commons has 36 normal working days until Brexit

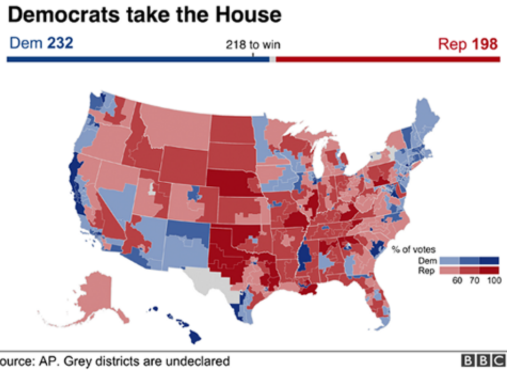
Note: The House of Commons sometimes sits on Fridays to debate individual MPs' bills

Source: Parliament



Results of the 2018 election

Source: Reuters



Democrats take the House

Source: AP. Grey districts are undeclared

BBC News graphics

The statistical programming language R has been used by the BBC for data analysis for some time, however website graphics were previously produced by an in-house design team.

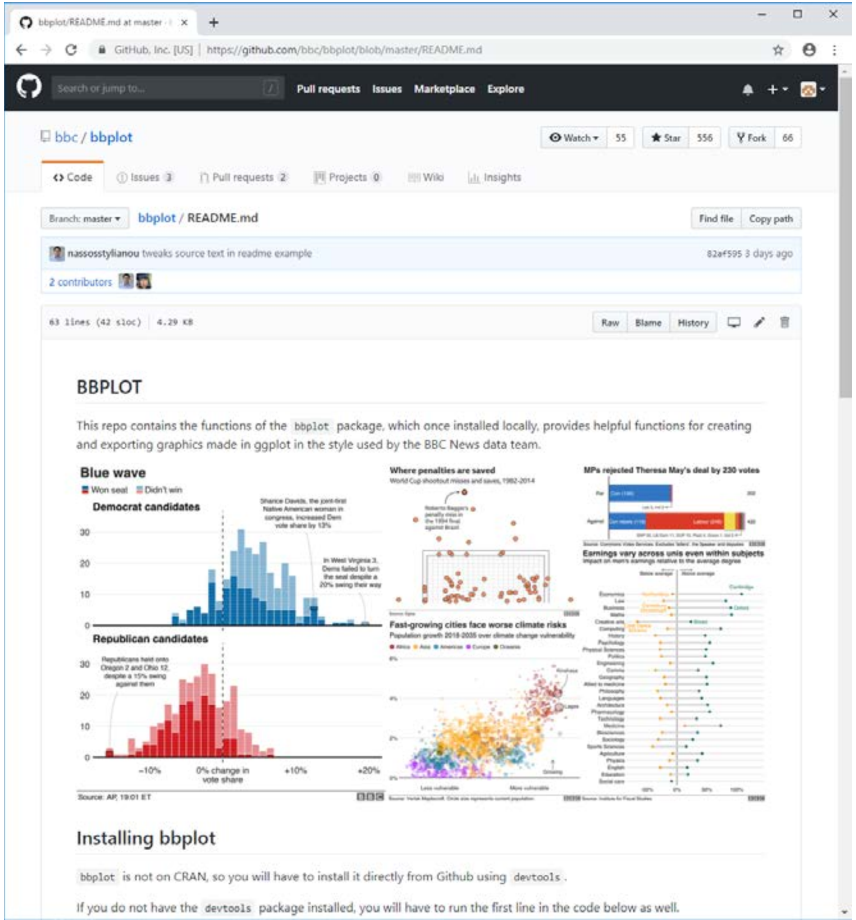
At the start of 2018, the BBC visual and data journalism team looked at whether R could be used for the whole process – in effect automating the design element of the process. In March 2018, they published the first chart that was produced from start to finish in R using the `ggplot2` package.

This streamlined the production of graphics and enabled more people at the BBC to easily produce graphics in their house style.

> “[This approach] saves a huge amount of time and effort, in particular when working with data that needs updating regularly, with **reproducibility a key requirement of our workflow**. In short, it was a game changer…
>
> BBC Visual and Data Journalism team ”

BBC News graphics – further reading

- Blog post here explains what the BBC did and what they learnt along the way
- The cookbook here is used by the BBC to share tips and tricks, and solutions to common problems
- The package that creates graphics in the BBC style can be found on GitHub here



bbplot/README.md at master

GitHub, Inc. [US] | https://github.com/bbc/bbplot/blob/master/README.md

bbc / bbplot

Branch: master ▾ bbplot / README.md

nassosstylianou tweaks source text in readme example 82af595 3 days ago

2 contributors

63 lines (42 sloc) | 4.29 KB

BBPLOT

This repo contains the functions of the `bbplot` package, which once installed locally, provides helpful functions for creating and exporting graphics made in ggplot in the style used by the BBC News data team.

Search or jump to...
Pull requests
Issues
Marketplace
Explore
Watch ▾ 55
Star 556
Fork 66
Code
Issues 3
Pull requests 2
Projects 0
Wiki
Insights
Find file
Copy path
Raw
Blame
History
Blue wave
Won seat
Didn't win
Democrat candidates
Sharice Davids, the joint-first Native American woman in congress, increased Dem vote share by 13%
In West Virginia 3, Dems failed to turn the seat despite a 20% swing their way
Republican candidates
Republicans held onto Oregon 2 and Ohio 12, despite a 15% swing against them
30
20
10
0
-10%
0% change in vote share
+10%
+20%
Source: AP, 19:01 ET
BBC
Where penalties are saved
World Cup shootout misses and saves, 1982-2014
Roberto Baggio's penalty miss in the 1994 final against Brazil
Fast-growing cities face worse climate risks
Population growth 2018-2035 over climate change vulnerability
Africa
Asia
Americas
Europe
Oceania
Less vulnerable
More vulnerable
MPs rejected Theresa May's deal by 230 votes
For
Against
Earnings vary across unis even within subjects
Economics
Law
Business
Maths
Physics
English
Education
Social care

Installing bbplot

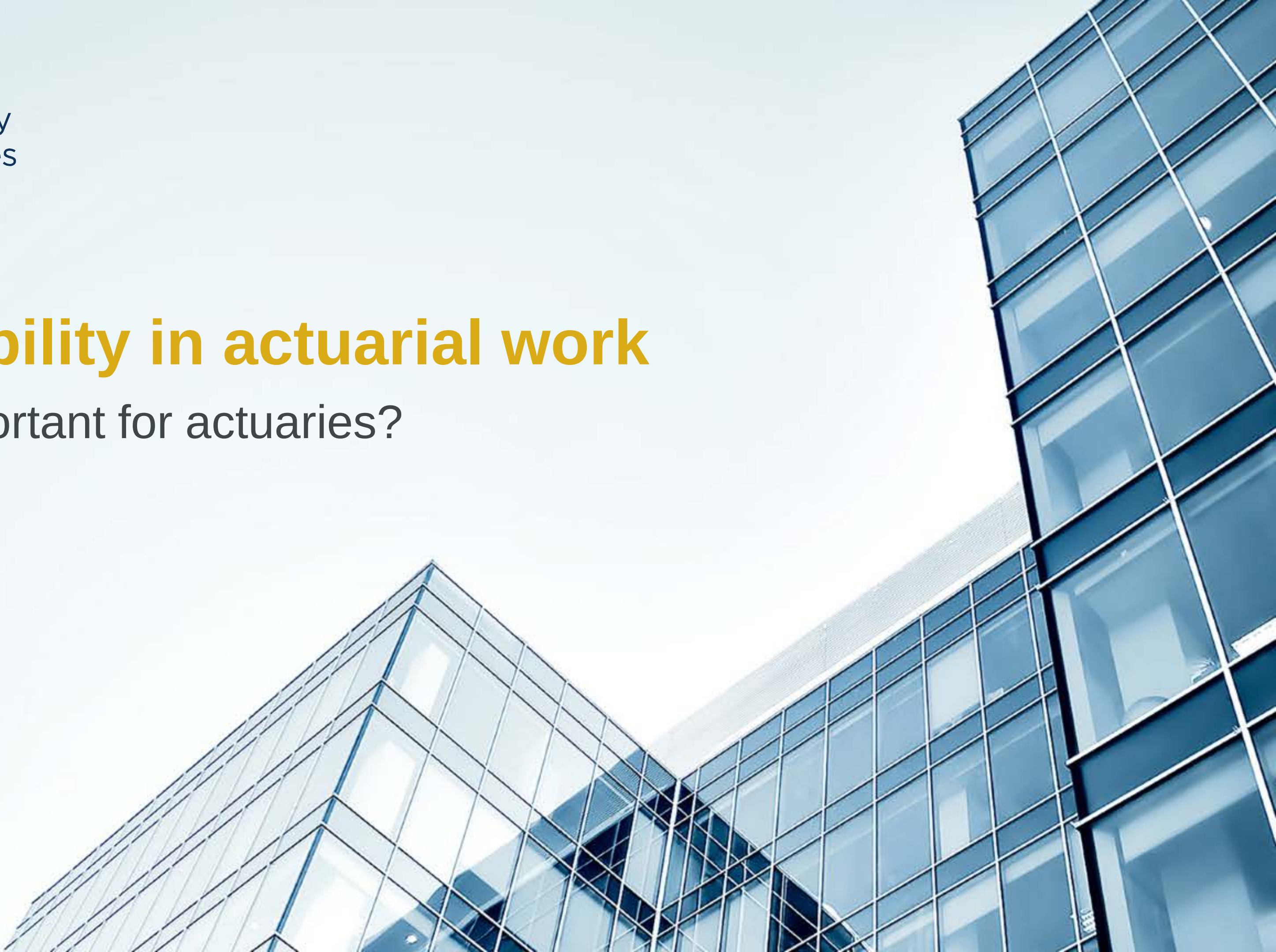
`bbplot` is not on CRAN, so you will have to install it directly from Github using `devtools`.

If you do not have the `devtools` package installed, you will have to run the first line in the code below as well.

Institute
and Faculty
of Actuaries

Reproducibility in actuarial work

Why is this important for actuaries?

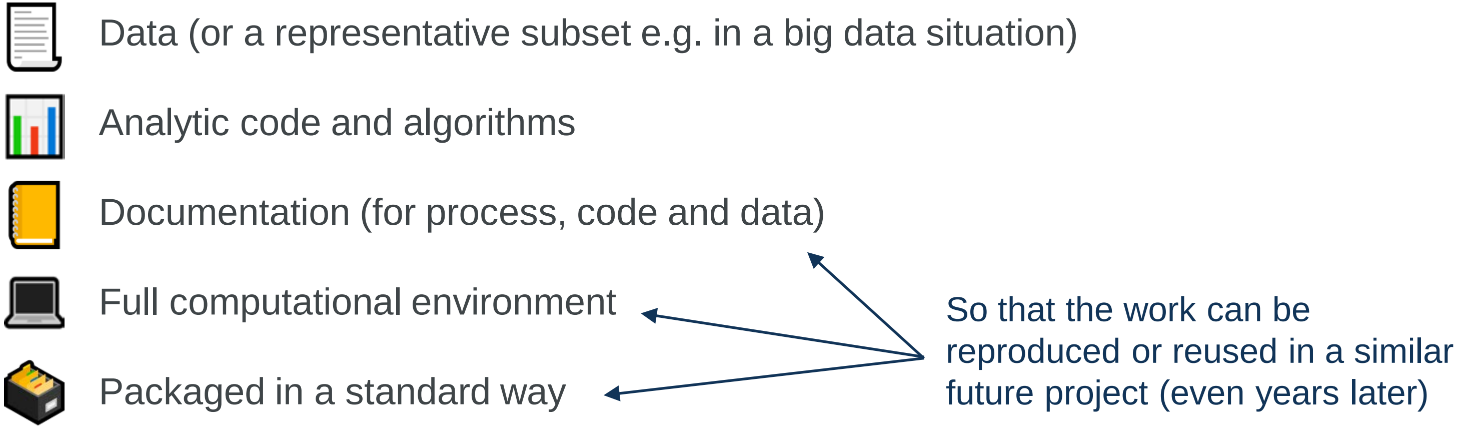


What is reproducibility?

Reproducibility is the process of making code and data available so that others can easily replicate your analysis.

Data scientists achieve this by using automation and standard tools/processes that reduce the number of manual steps required in a piece of analysis. This allows them to work with larger datasets and more complex workflows.

A reproducible process is one that brings together the:



- Data (or a representative subset e.g. in a big data situation)
- Analytic code and algorithms
- Documentation (for process, code and data)
- Full computational environment
- Packaged in a standard way

So that the work can be reproduced or reused in a similar future project (even years later)

What are the key points from the case studies?

Reproducible workflows improve productivity

- Reproducible workflows are more efficient, freeing up analysts to focus on interpreting results, innovating and solving problems – not wrestling with software or formatting documents.

Reproducible workflows help improve the quality of work

- Reproducible work is more easily checked and audited. Having fewer manual processes limits the potential for human error. It allows work to be revisited, updated and improved down the line.

Use of open source software

- Organisations are increasingly moving towards the use of open source software – making it possible to introduce reproducible workflows to the workplace.

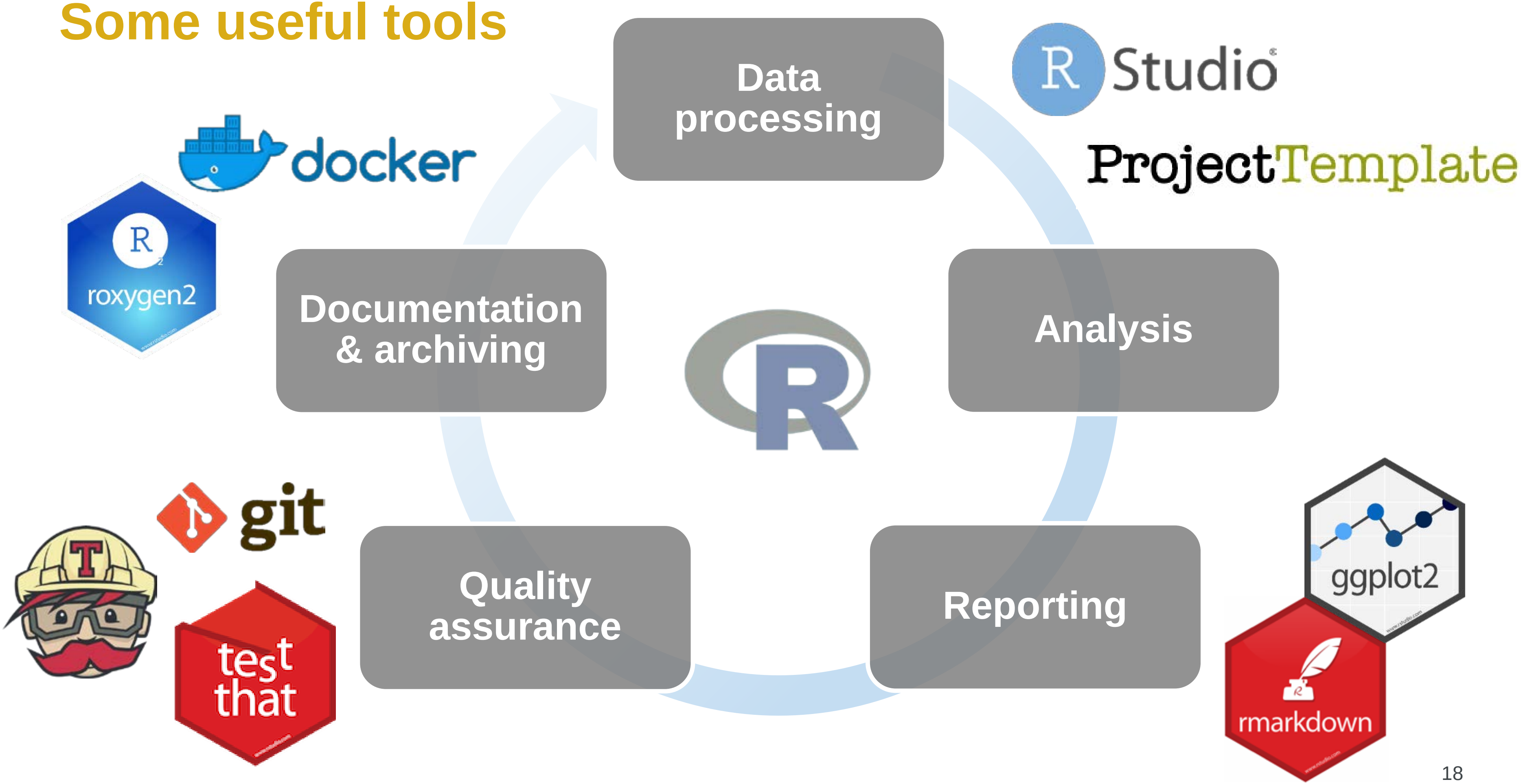
Sharing best practice throughout the organisation

- Government data scientists have shared their findings with over 600 colleagues across government, written a freely available book and a online course. BBC data journalists created a cookbook and a 6 week internal course to upskill colleagues. Their tools are open source and freely available to all (not just the data scientists).

Why is this important for actuaries?

- Actuaries carry out complex analytical work in data intensive fields. The work will continue to get more complex! Other professions are facing the same challenges – actuaries can learn from work done elsewhere.
- Having the right workflow and tools in place make it easier for actuaries to adopt other techniques from data science and to innovate in future.
- Adopting a reproducible workflow is arguably the first step in automating business as usual work.
- A reproducible workflow provides a solid base from which to demonstrate auditability and TAS compliance.
- Proprietary software remains important but actuaries experimenting with data science are beginning to use open source tools. The 2019 curriculum means the next generation of actuaries will be familiar with R.

Some useful tools



Data processing

Analysis

Reporting

Quality assurance

Documentation & archiving

Useful tools for building a reproducible workflow

RStudio is a free and widely used development environment for R that integrates with the tools below.

ProjectTemplate automates the menial parts of statistical analysis and provides a standard way of working in R.

R Markdown is a notebook interface that allows code to sit alongside narrative text and can be used for reporting as part of a reproducible framework with ggplot2 for creating charts and visualisations.

Git is a version control system for managing code and audit trails – it can be used privately in an organisation or with a web-based service such as GitHub.

testthat is a formal automated testing (“unit testing”) package for R.

TravisCI integrates with GitHub to automatically run your tests when code is updated.

roxygen2 automates the production of documentation for your code in R.

Docker packages dependencies inside a container which can run consistently on any infrastructure (see checkpoint/packrat for a lighter touch solution).

Challenges of building a reproducible workflow

- A reproducible approach isn't the solution to all of the complications of analytical work but it's a good first step.
- Developing a reproducible pipeline can be difficult and time-consuming – it's worth exploring whether other solutions are available before developing one from scratch.
- Solutions are generally built around open source software which can be a culture change for some organisations – although this is changing.
- Training requirements for new tools and processes.

Interested in applying these techniques?

Exercises

We have prepared exercises at https://philipdarke.com/reproducible-actuarial-work/ to illustrate how a reproducible workflow can be set up and how some of the tools introduced above are used. The exercises use R to process a data set, carry out some basic analysis and produce a report.

Introducing a reproducible process in your organisation

- Take an existing process that is repeated often
- Develop a minimal viable solution using the tools and techniques introduced (see the exercises)
- Pilot it and let others contribute
- Share what you learn within your organisation
- Consider what you can share more widely e.g. with the actuarial community

Get in touch

We are interested in this! Please get in touch with questions, comments and suggestions.

Dr Matthew Forshaw is a Lecturer in Data Science at Newcastle University, and Data Skills Policy Leader at The Alan Turing Institute working on the Data Skills Taskforce. He is the Programme Director of Newcastle’s Industrial MSc in Data Science.
mattforshaw.com

Philip Darke is an actuary with over 10 years’ consulting experience at Mercer. Philip believes data science will play a significant role in the future of the actuarial profession, and is currently studying at the EPSRC Centre for Doctoral Training in Cloud Computing for Big Data at Newcastle University.
philipdarke.com

Questions

Comments

The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the authors.