Reproducible Data Science with R

(Slightly) modified from slides by: David Smith
R Community Lead, Microsoft
@revodavid

What is reproducibility?

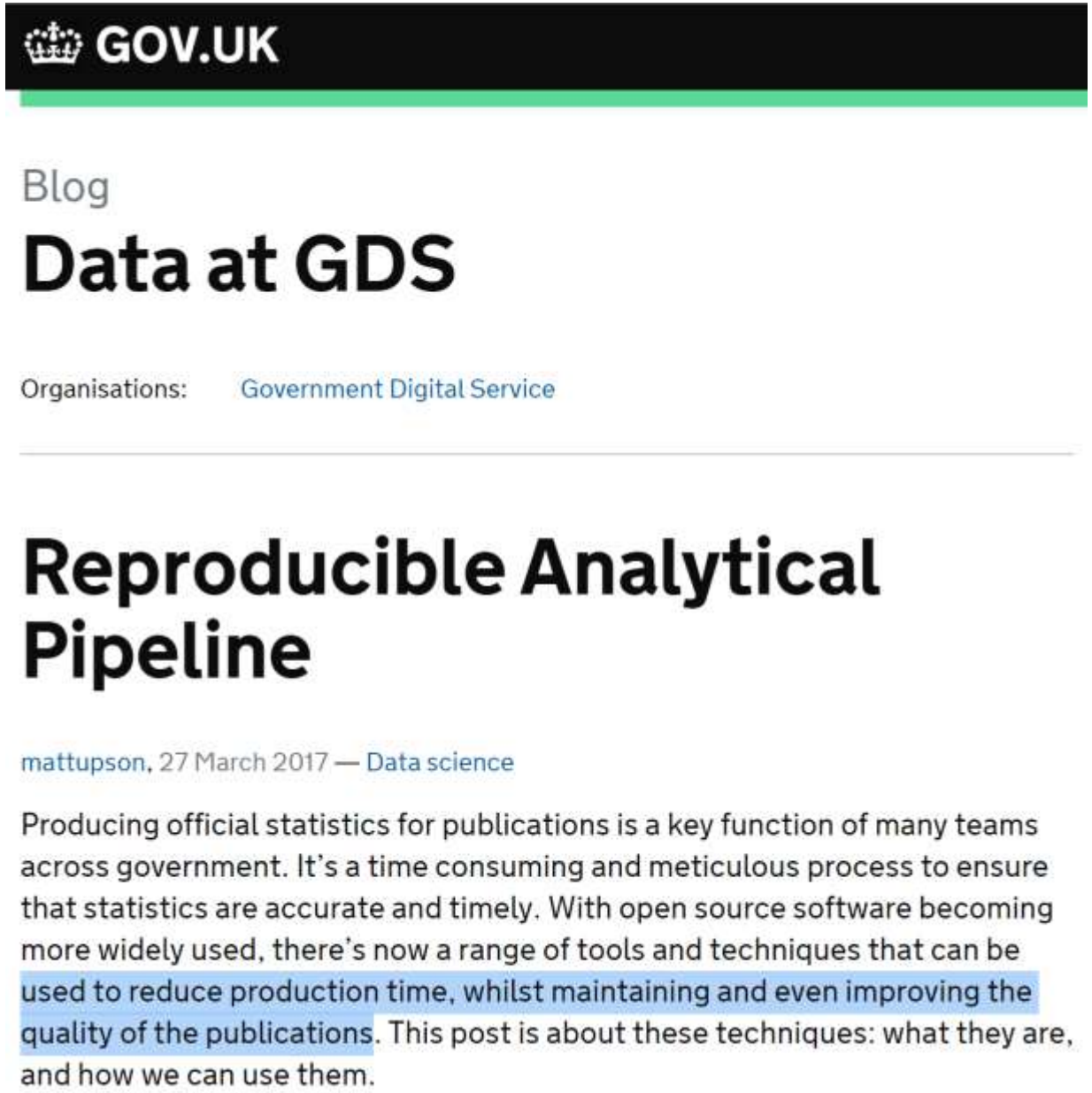
"Two honest researchers would get the same result"

– John Mount

- Transparent data sourcing and availability
- Fully automated analysis pipeline (code provided)
- Traceability from published results back to data

Why Reproducibility?

- **Save time**
- Better science
- More authoritative research
- Reduce risk of errors
- Facilitate collaboration



GOV.UK

Blog

Data at GDS

Organisations: Government Digital Service

Reproducible Analytical Pipeline

mattupson, 27 March 2017 — Data science

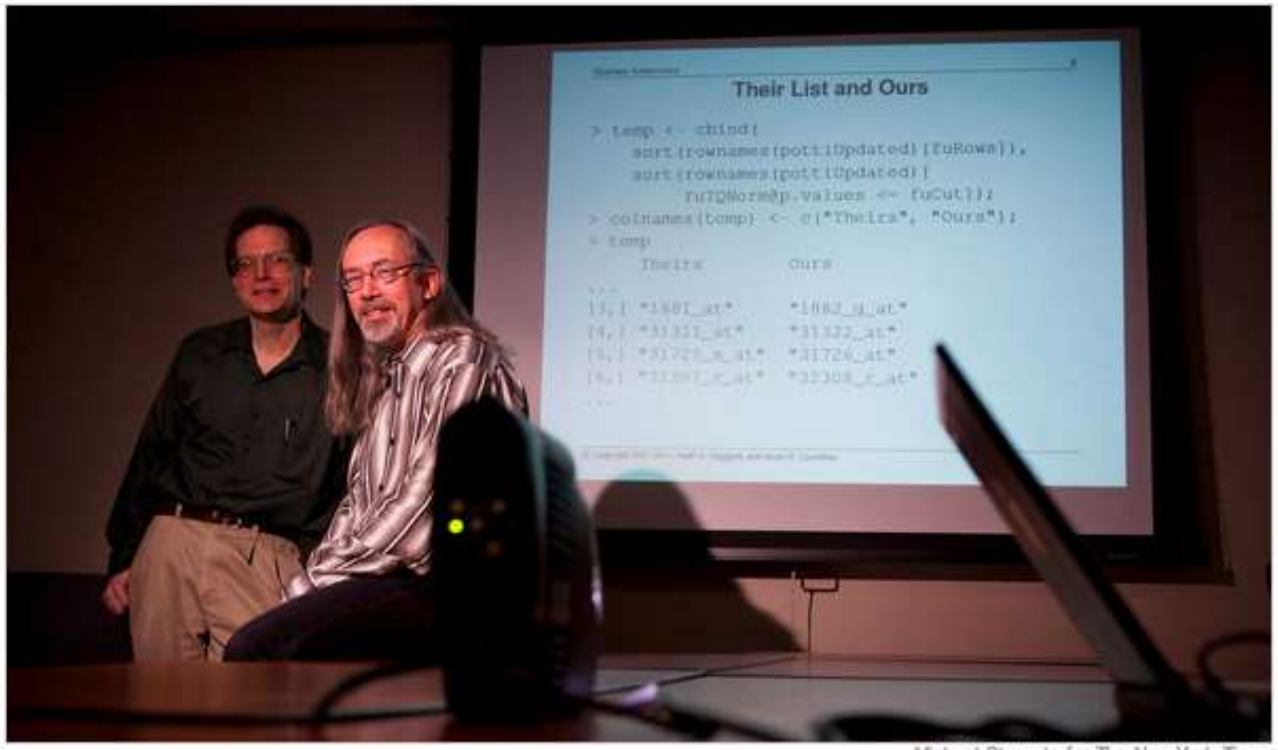
Producing official statistics for publications is a key function of many teams across government. It's a time consuming and meticulous process to ensure that statistics are accurate and timely. With open source software becoming more widely used, there's now a range of tools and techniques that can be used to reduce production time, whilst maintaining and even improving the quality of the publications. This post is about these techniques: what they are, and how we can use them.

Microsoft

Why Reproducibility?

- Save time
- **Better science**
- More authoritative research
- Reduce risk of errors
- Facilitate collaboration

How Bright Promise in Cancer Testing Fell Apart



Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By GINA KOLATA
Published: July 7, 2011

Microsoft

Why Reproducibility?

- Save time
- Better science
- **More authoritative research**
- Reduce risk of errors
- Facilitate collaboration

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Rights & Permissions

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

Brian Nosek's team set out to replicate scores of studies.

The studies they took on ranged from whether expressing insecurities perpetuates them to differences in how children and adults respond to fear stimuli, to effective ways to teach arithmetic.

Microsoft

Why Reproducibility?

- Save time
- Better science
- More authoritative research
- **Reduce risk of errors**
- Facilitate collaboration

After the London Whale trade blew up, the Model Review Group discovered that the model had not been automated and found several other errors. Most spectacularly,

> *"After subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the modeler had intended. This error likely had the effect of muting volatility by a factor of two and of lowering the VaR . . ."*

JPMorgan Embarrassed Over $2 Billion in Losses [Update]

By **Adam Pasick**

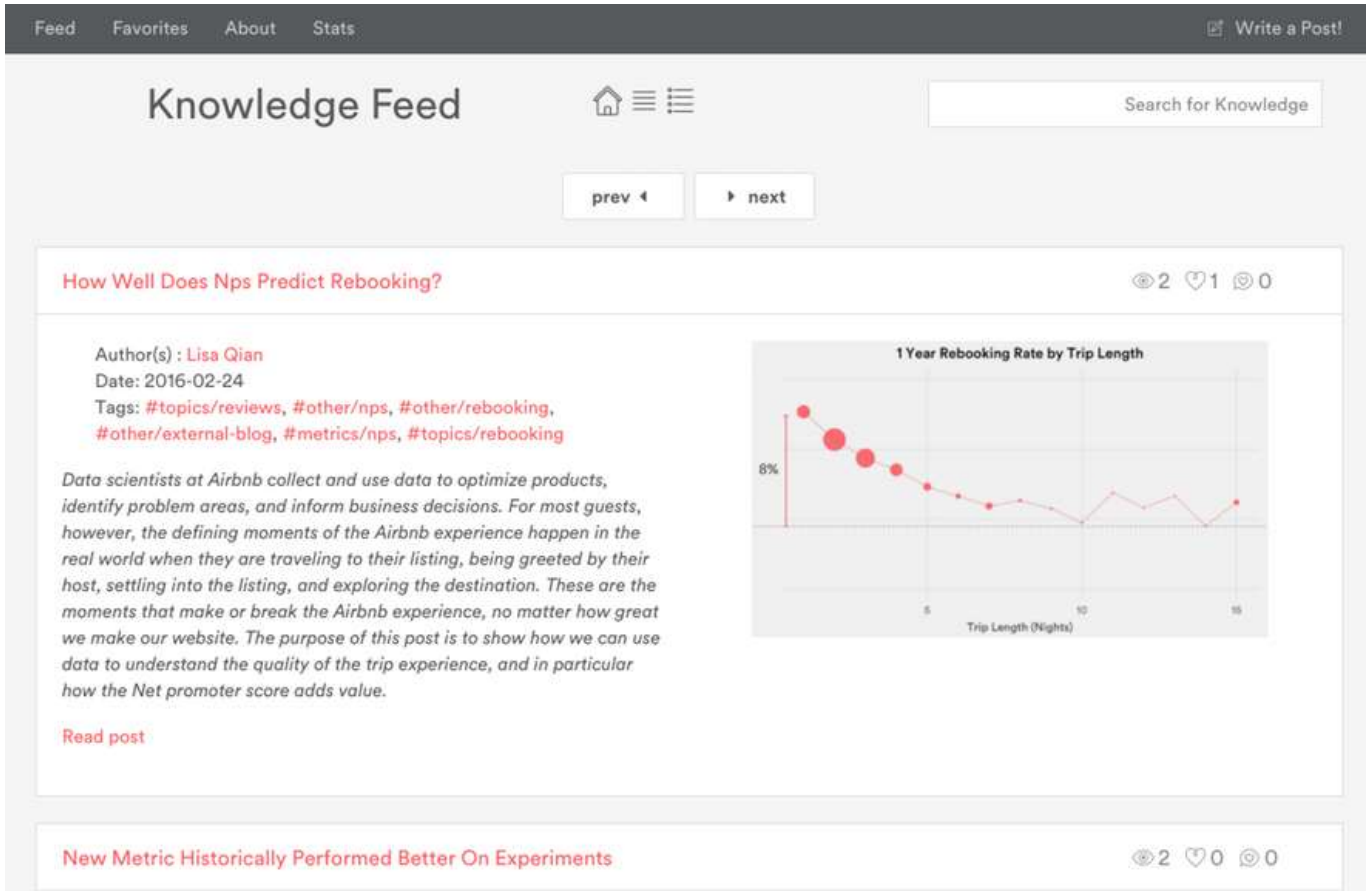
Photo: Mario Tama/2012 Getty Images

JPMorgan Chase CEO Jamie Dimon made a shocking disclosure Thursday night that some of the company's bets on credit markets have gone bad to the "significant" tune of $2 billion. The losses stemmed from a hedging

Microsoft

Why Reproducibility?

- Save time
- Better science
- More authoritative research
- Reduce risk of errors
- **Facilitate collaboration**



Accessing Data

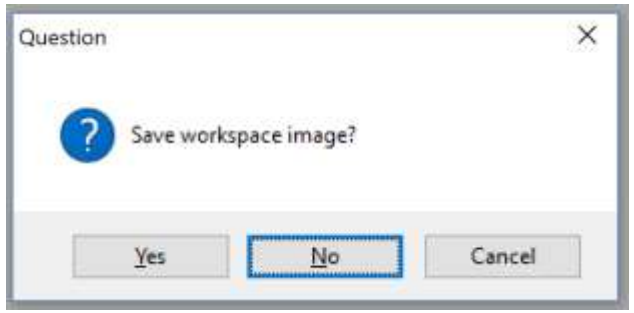
- Data sources
  - databases, sensor logs, spreadsheets, file download, APIs, …
  - Remember, these are *sources*: **don't modify them**!
  - Create a dedicated directory for input data
- Snapshot data into local static files
  - You will likely include some ETL steps here
  - Record a timestamp of when data was extracted: `Sys.time()`
  - Document how this was all done, **preferably** with code (source file)!
- Import text files using R functions
  - e.g. `read.table`

Analysis process

- Interactively explore data & develop analyses as usual
- **Capture** the entire process in an R script
  - `library(tidyverse)` is helpful for cleaning, feature generation, etc.
  - Re-usable components can be shared as (private) packages
- Generate artifacts using scripts
  - Graphics (please, no JPGs!)
  - Tables
  - Documents
- Develop a project directory, use local paths
  - data/, code/, figures/, output/

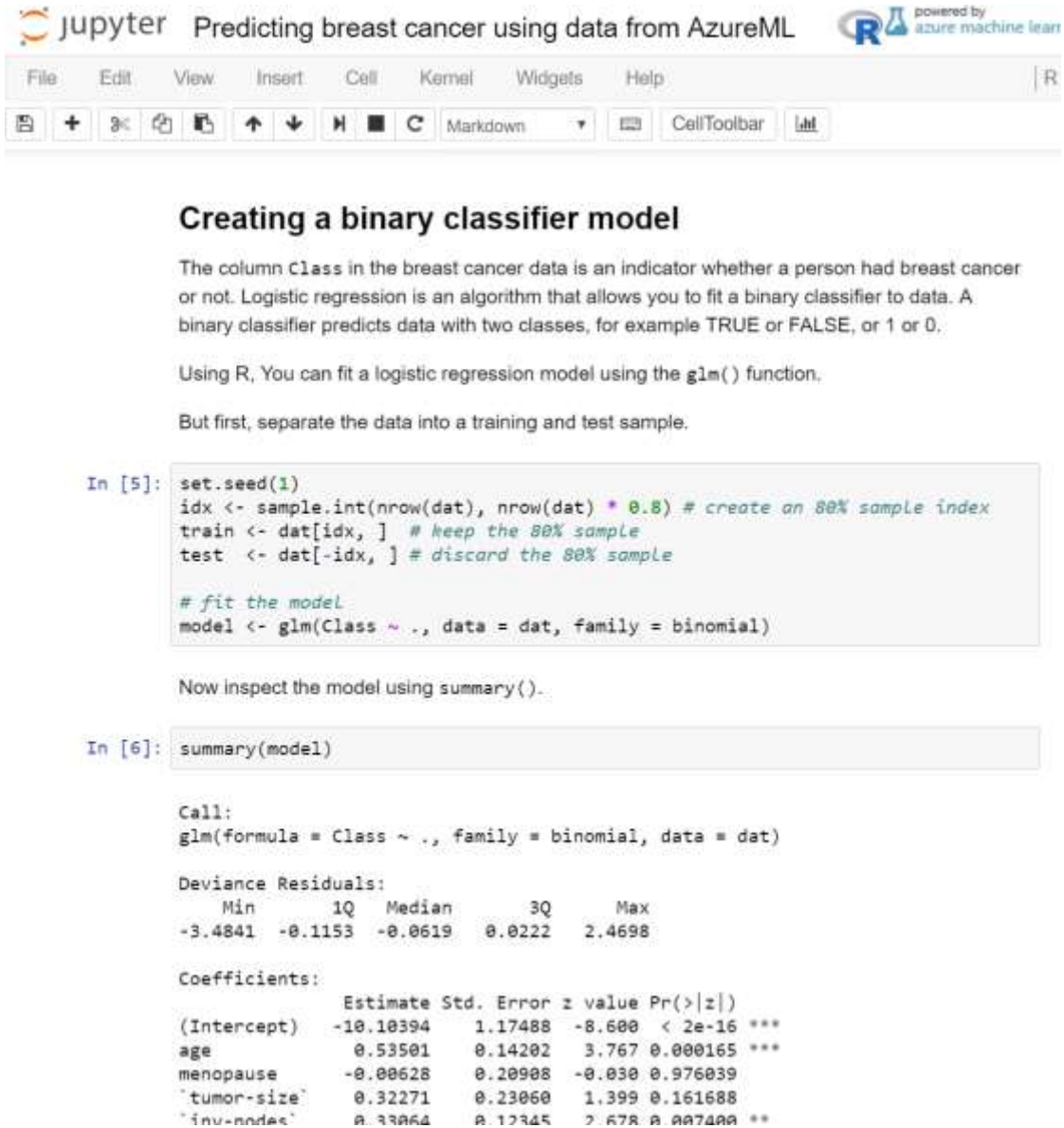
A reproducible analysis environment

- Operating system and R version
  - For most purposes, not the biggest cause of issues
    - But do document your R session info: `sessionInfo()`
- A clean R environment
  - Organize work into independent projects (directories)
  - Use **relative** paths in scripts
  - Avoid use of .Rprofile
  - Set explicit random seeds
  - Do **not** save R workspace



Presenting results

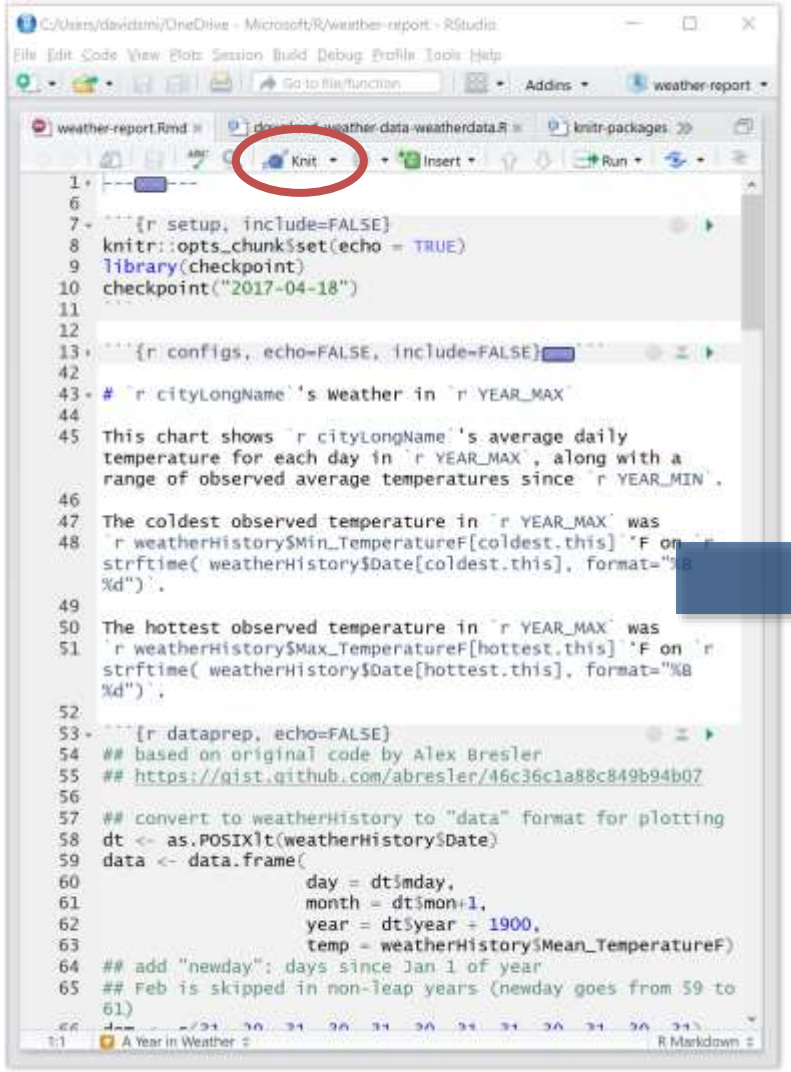
- Eliminate manual processes (as far as possible)
  - Annotations (graphs / tables)
  - Cut-and-paste into documents
- Notebooks
  - Combines code, output and narrative
  - Good for collaboration with other researchers
- Document Generation
  - Best for automating reports

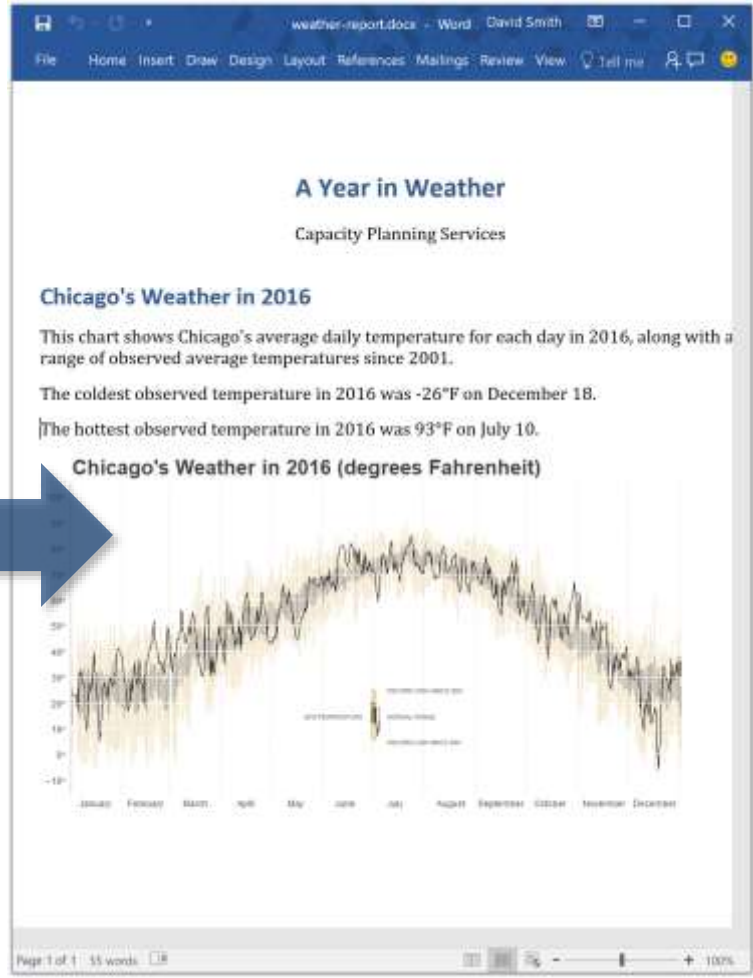


Microsoft

knitr / Rmarkdown

- Generate HTML, Word, or PDF reports
  - Or books, blogs, slides, …
- Combine narrative and R code in a single document
- Human-readable, easy to edit
- Single-click update!

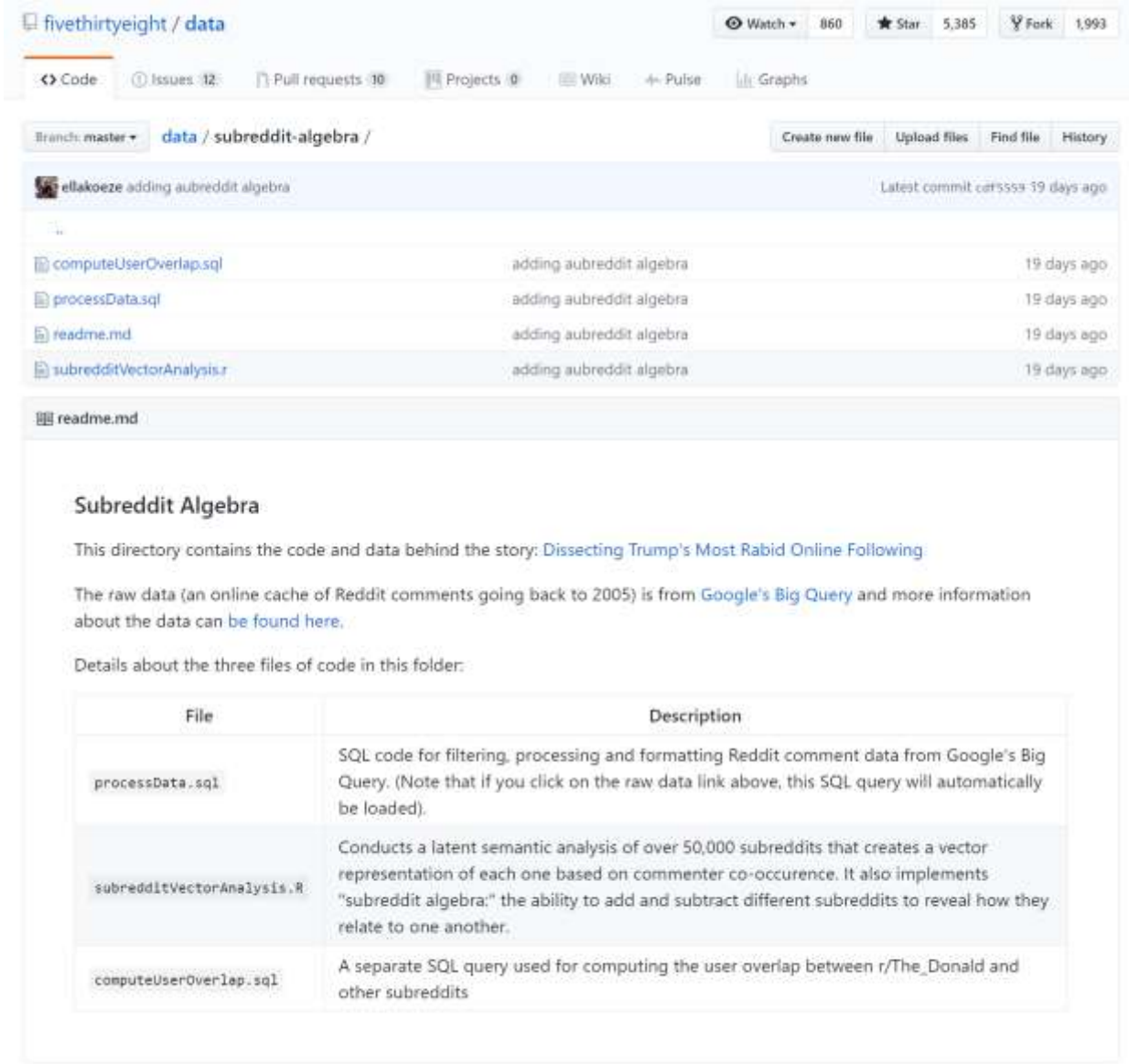




Microsoft

Collaboration and sharing

- Just share R project folder
- Publish on Github
  - **Happy Git and GitHub for the useR**, Jenny Bryan http://happygitwithr.com/
  - Version retention and tracking
  - Collaboration (code and comments)



Take-Aways

Reproducibility is Beneficial

- Saves time
- Produces better science
- More trusted research
- Reduced risk of errors
- Encourages collaboration

Reproducibility is Simple

- Document and automate processes with R scripts
- Read and clean data with **tidyverse**
- Use **checkpoint** to manage package versions
- Generate documents with **knitr**
- Share reproducible projects with Github