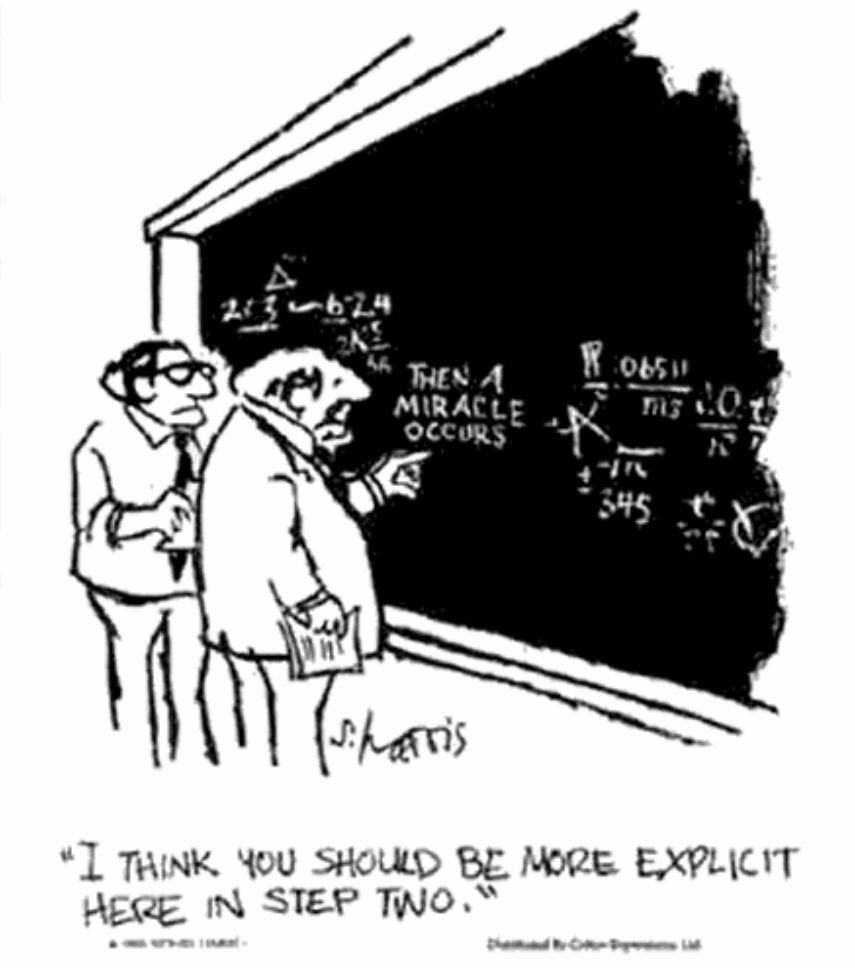
Reproducibility and Dynamic Documents

Roxanne Connelly
roxanne.connelly@york.ac.uk
@RoxanneConnelly

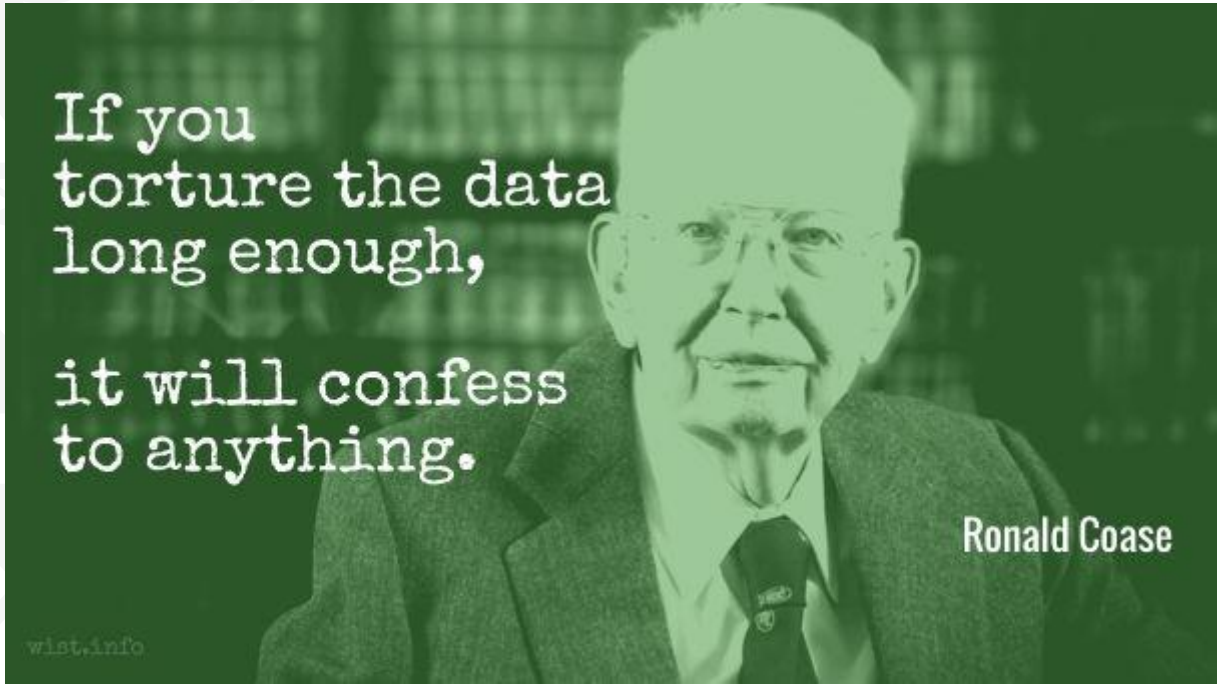
AQMEN DATA SCIENCE TRAINING AND CAPACITY BUILDING WORKSHOP
InTuition House, London
27th March 2019

Problem

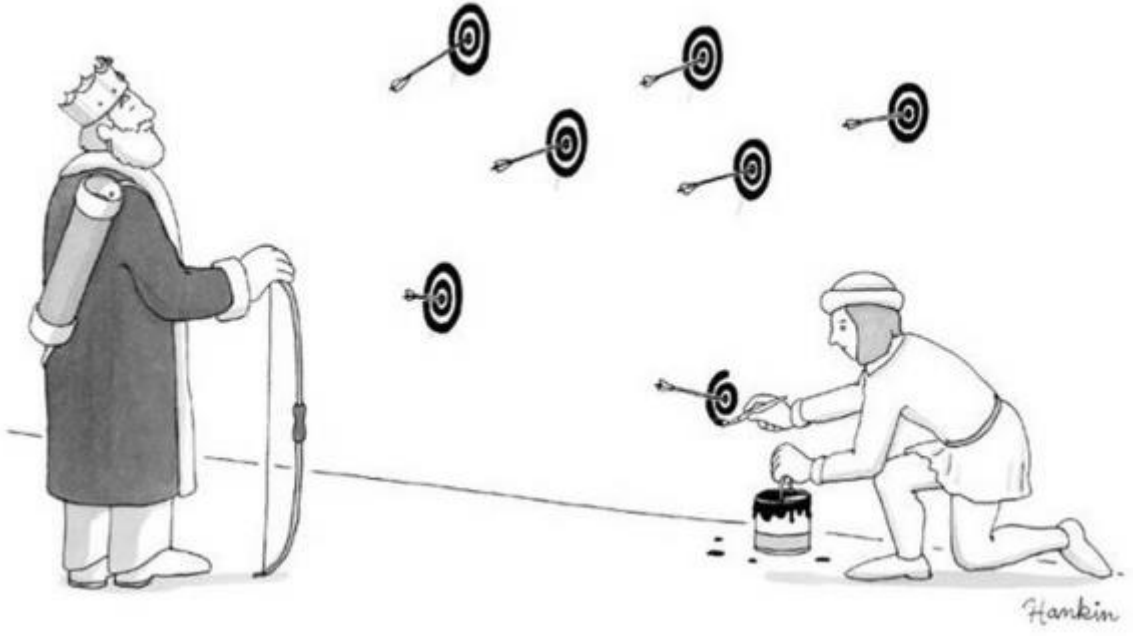

THEN A MIRACLE OCCURS
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."


NULLIUS
IN
VERBA

P Hacking


If you
torture the data
long enough,
it will confess
to anything.
Ronald Coase
wist.info

HARKing
Hypothesizing After the Results are Known



Example: Screen Time

nature human behaviour

UNIVERSITY of York

Article | Published: 14 January 2019

The association between adolescent well-being and digital technology use

Amy Orben & Andrew K. Przybylski

Nature Human Behaviour **3**, 173–182 (2019) | Download Citation

Abstract

The widespread use of digital technologies by young people has spurred speculation that their regular use negatively impacts psychological well-being. Current empirical evidence supporting this idea is largely based on secondary analyses of large-scale social datasets. Though these datasets provide a valuable resource for highly powered

UNIVERSITY of York

TECHNOLOGY

Have Smartphones Destroyed a Generation?

More comfortable online than out partying, post-Millennials are safer, ally, than adolescents have ever But they're on the brink of a mental- crisis.

TWENGE SEPTEM

Are smartphones causing more teen suicides?

Increases in depression and suicide appeared among teens in 2012 – the same time smartphone ownership became the norm

INDEPENDENT

JOIN US? SUBSCRIBE REGISTER

NEWS POLITICS VOICES FINAL SAY SPORT CULTURE VIDEO INDY/LIFE INDYBEST LONG READS INDY100 VOUCHERS

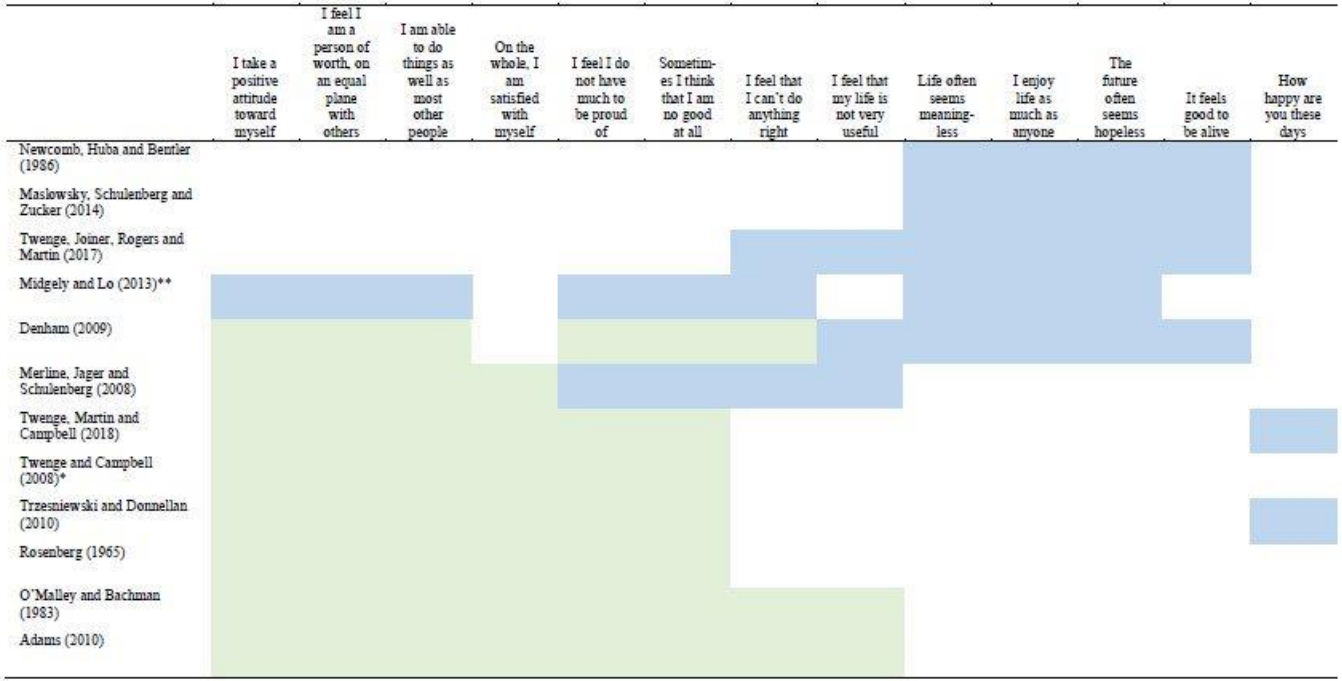
SMARTPHONES 'CAUSING MENTAL HEALTH PROBLEMS IN TWO-YEAR-OLDS'

Neutral	Perceived weight	×1.02	-	-
Factors	Potatoes	×0.86	-	-
	Asthma	×1.34	-	-
	Milk	×0.28*	-	-
	Going to movies	-	×11.51*	-
	Religion	-	×16.29*	-
	Music	-	×32.68	-
	Homework	-	×3.57*	-
	Cycling	-	-	×1.88*
	Height	-	-	×1.53*
	Glasses	-	-	×1.45
	Handedness	-	-	×0.10

Table 1 | Possible specifications (analytical decisions) used to test a simple linear regression between technology use and adolescent well-being in the datasets YRBS, MTF and MCS

Decision	YRBS	MTF	MCS
Operationalizing adolescent well-being	Mean of any possible combination of five items concerning mental health and suicidal ideation	Mean of any possible combination of 13 items concerning depression, happiness and self-esteem	Mean of any possible combination of 24 questions concerning well-being, self-esteem and feelings (cohort members), or mean of any possible combination of 25 questions from the Strengths and Difficulties Questionnaire (caregivers)
Operationalizing technology use	Two questions concerning electronic device use and TV use, or the mean of these questions	Eleven technology use measures concerning the Internet, electronic games, mobile phone use, social media use and computer use, or the mean of these questions	Five questions concerning TV use, electronic games, social media use, owning a computer and using the Internet at home, or the mean of these questions
Which co-variates to include	Either include co-variates or not	Either include co-variates or not	Either include co-variates or not
Other specifications	Either take mean of dichotomous well-being measures, or code all cohort members who answered 'yes' to one or more as 1 and all others as 0		Use well-being measures declared by cohort members or those declared by their caregivers

Table showing how research papers have used different combinations of MTF measures to define depressive symptoms (blue) and self-esteem (green). This illustrates the abundance of analytical flexibility in this area. We also include Newcomb, Huba and Bentler (1986) and Rosenberg (1965) who originally devised parts of the scales.

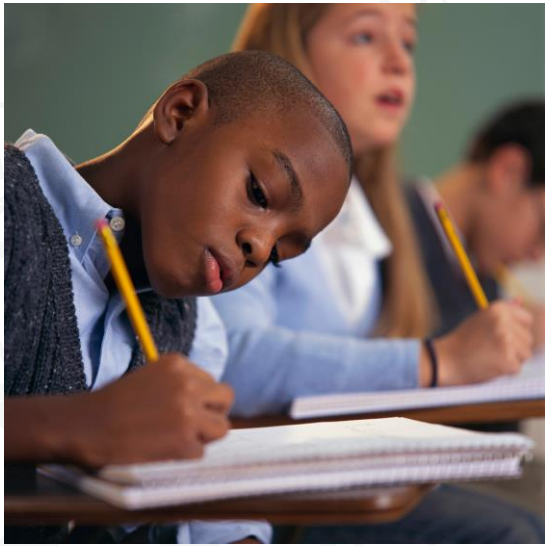


Solutions

“It’s a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of **utter honesty** - a kind of leaning over backwards. For example, if you’re doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it…”

(Feynman 1974)

Feynman, Richard P. "Cargo cult science." Engineering and Science 37, no. 7 (1974): 10-13.



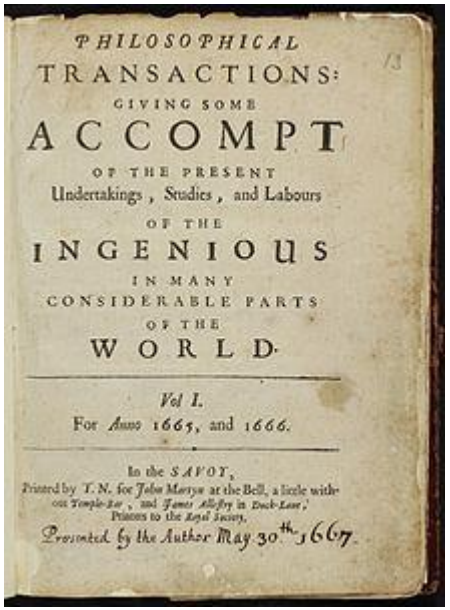
Remember to show your working out as well as your answers.

6X8=48

$$\begin{array}{r} {}^{6}\cancel{2}{}^{1}\cancel{7}3 \\ -\ 47 \\ \hline 226 \end{array}$$

First Journal Publications

- In 1665 the first peer reviewed journal articles were published (Journal des Sçavans and Philosophical Transactions of the Royal Society of London).


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
Presented by the Author May 30th 1667.

Modern Journal Publications

“The paper is just a palimpsest”
(Gayle 2016)

Documentation

- You should document the whole research process (including the data management):
  - ‘Methods’ sections of publications do not allow sufficient detail to be presented.

We need to share details of the whole research process!

The TOP Guidelines

- The TOP Guidelines were created by journal, funders and societies to align scientific ideals with practices.

- Published in 2015, the Transparency and Openness Promotion Guidelines include eight modular standards, each with three levels of increasing stringency.

- Over 5,000 journals and organizations have already become signatories of the TOP Guidelines.

https://cos.io/our-services/top-guidelines/

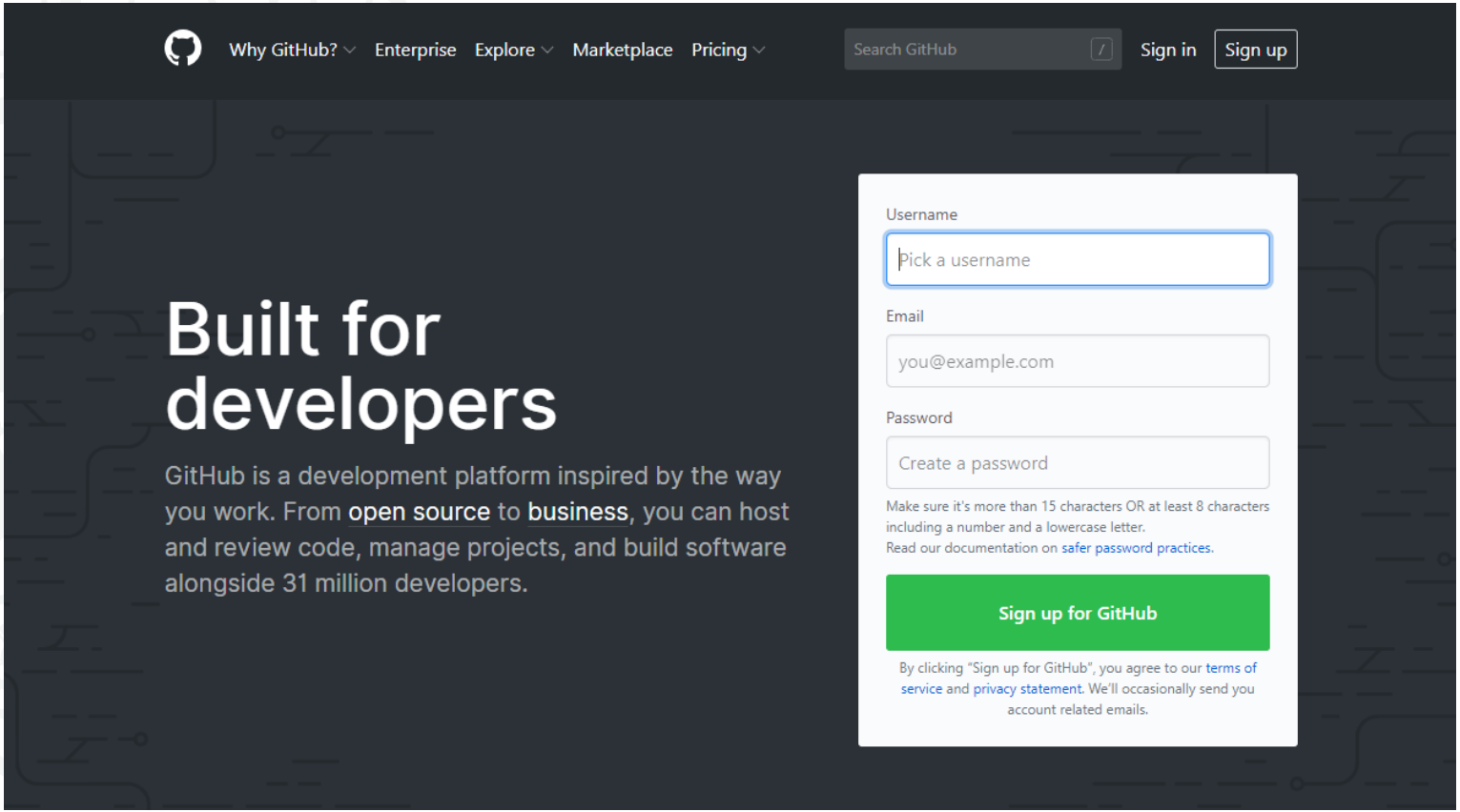
American Journal of Political Science

Replication Files and Article Verification Process

The author of an accepted manuscript must provide replication materials that are sufficient to enable interested researchers to reproduce all of the analytic results that are reported in the text and supporting materials. The document titled "*American Journal of Political Science* Guidelines for Preparing Replication Files" provides useful information about what information is needed and how it should be organized. All replication files must be stored in a Dataset within the *AJPS* Dataverse, located on the Harvard Dataverse Network. The document titled "*American Journal of Political Science* Quick Reference for Uploading Replication Files" provides information about creating a Dataset on the *AJPS* Dataverse and depositing materials into it. Note that authors also can make their replication files available elsewhere (e.g., their personal website, other data repositories, etc.) as long as all of the necessary files are included in the Dataset on the *AJPS* Dataverse.

https://ajps.org/guidelines-for-accepted-articles/

GitHub (github.com)



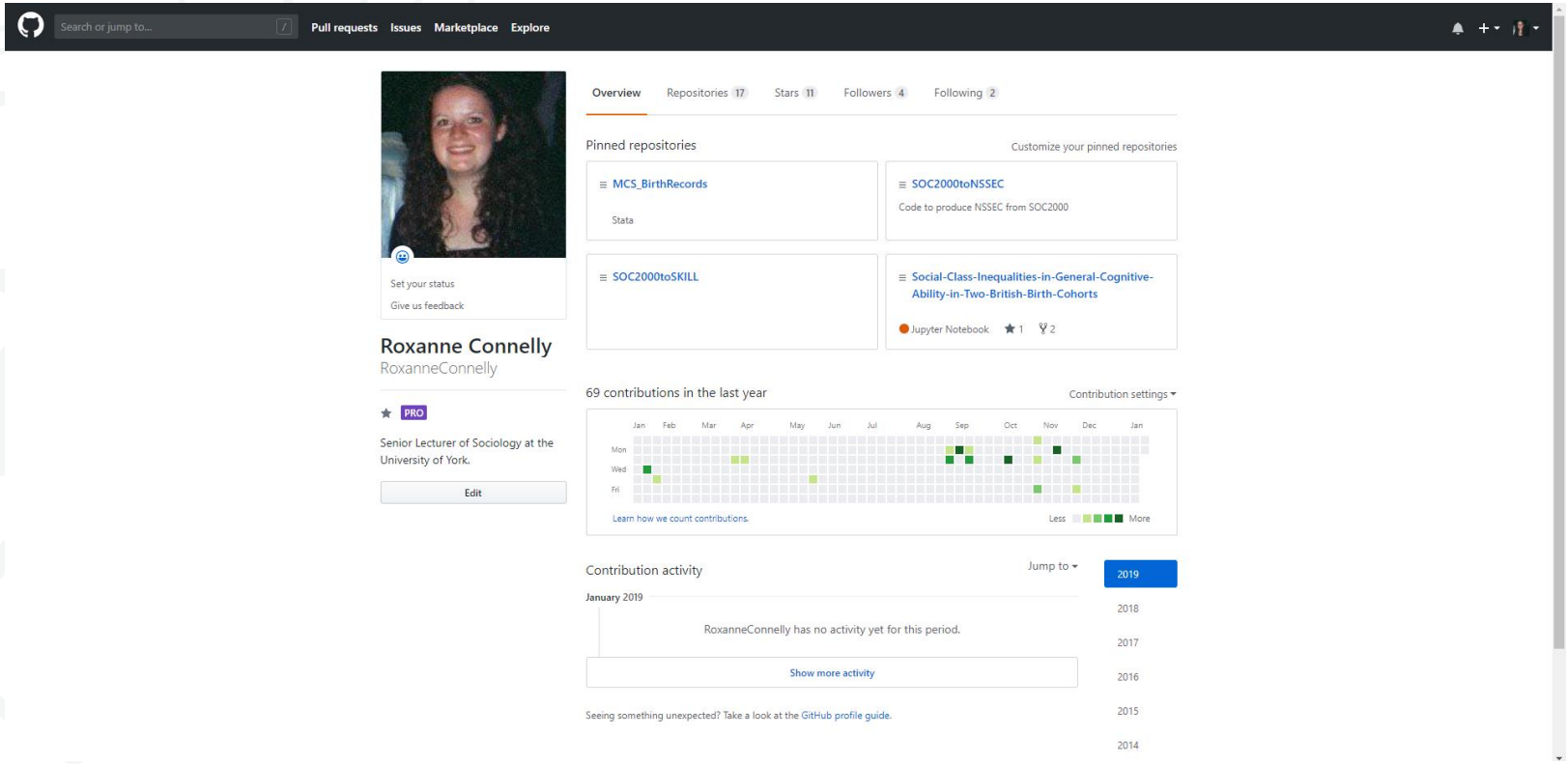
GitHub

- GitHub is a web-based system to keep multiple revisions of your files and to share code easily with collaborators.

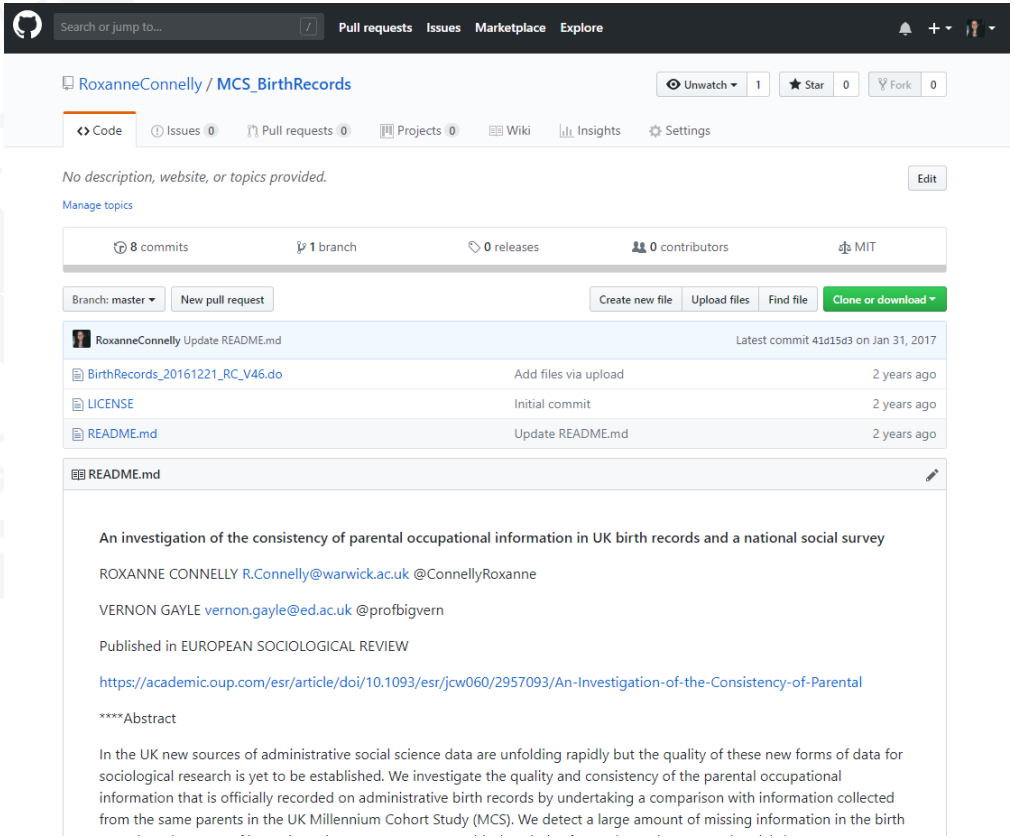
Version Control Software

- Version control systems keep track of every change to a file over time so early versions can be restored.

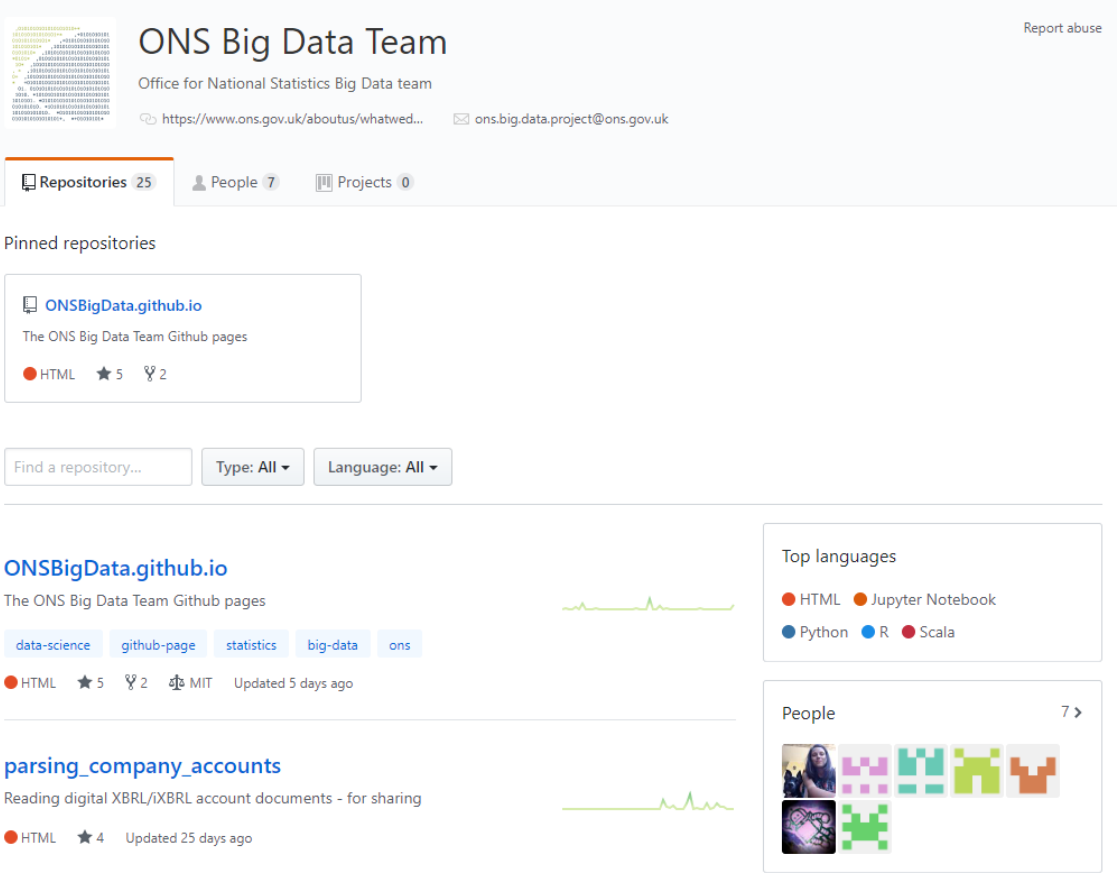
GitHub


Search or jump to...
Pull requests
Issues
Marketplace
Explore
Overview
Repositories 17
Stars 11
Followers 4
Following 2
Pinned repositories
Customize your pinned repositories
MCS_BirthRecords
Stata
SOC2000toNSSEC
Code to produce NSSEC from SOC2000
SOC2000toSKILL
Social-Class-Inequalities-in-General-Cognitive-Ability-in-Two-British-Birth-Cohorts
Jupyter Notebook
Set your status
Give us feedback
Roxanne Connelly
RoxanneConnelly
PRO
Senior Lecturer of Sociology at the University of York.
Edit
69 contributions in the last year
Contribution settings
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan
Mon
Wed
Fri
Learn how we count contributions.
Less
More
Contribution activity
Jump to
2019
2018
2017
2016
2015
2014
January 2019
RoxanneConnelly has no activity yet for this period.
Show more activity
Seeing something unexpected? Take a look at the GitHub profile guide.

GitHub


Search or jump to...
Pull requests
Issues
Marketplace
Explore
RoxanneConnelly / MCS_BirthRecords
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Insights
Settings
No description, website, or topics provided.
Edit
Manage topics
8 commits
1 branch
0 releases
0 contributors
MIT
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
RoxanneConnelly Update README.md
Latest commit 41d15d3 on Jan 31, 2017
BirthRecords_20161221_RC_V46.do
Add files via upload
2 years ago
LICENSE
Initial commit
2 years ago
README.md
Update README.md
2 years ago
README.md
An investigation of the consistency of parental occupational information in UK birth records and a national social survey
ROXANNE CONNELLY R.Connelly@warwick.ac.uk @ConnellyRoxanne
VERNON GAYLE vernon.gayle@ed.ac.uk @profbigvern
Published in EUROPEAN SOCIOLOGICAL REVIEW
https://academic.oup.com/esr/article/doi/10.1093/esr/jcw060/2957093/An-Investigation-of-the-Consistency-of-Parental
****Abstract
In the UK new sources of administrative social science data are unfolding rapidly but the quality of these new forms of data for sociological research is yet to be established. We investigate the quality and consistency of the parental occupational information that is officially recorded on administrative birth records by undertaking a comparison with information collected from the same parents in the UK Millennium Cohort Study (MCS). We detect a large amount of missing information in the birth

GitHub



ONS Big Data Team

Office for National Statistics Big Data team

https://www.ons.gov.uk/aboutus/whatwed... ons.big.data.project@ons.gov.uk

Report abuse

Repositories 25 People 7 Projects 0

Pinned repositories

ONSBigData.github.io

The ONS Big Data Team Github pages

HTML 5 2

Find a repository... Type: All Language: All

ONSBigData.github.io

The ONS Big Data Team Github pages

data-science github-page statistics big-data ons

HTML 5 2 MIT Updated 5 days ago

parsing_company_accounts

Reading digital XBRL/iXBRL account documents - for sharing

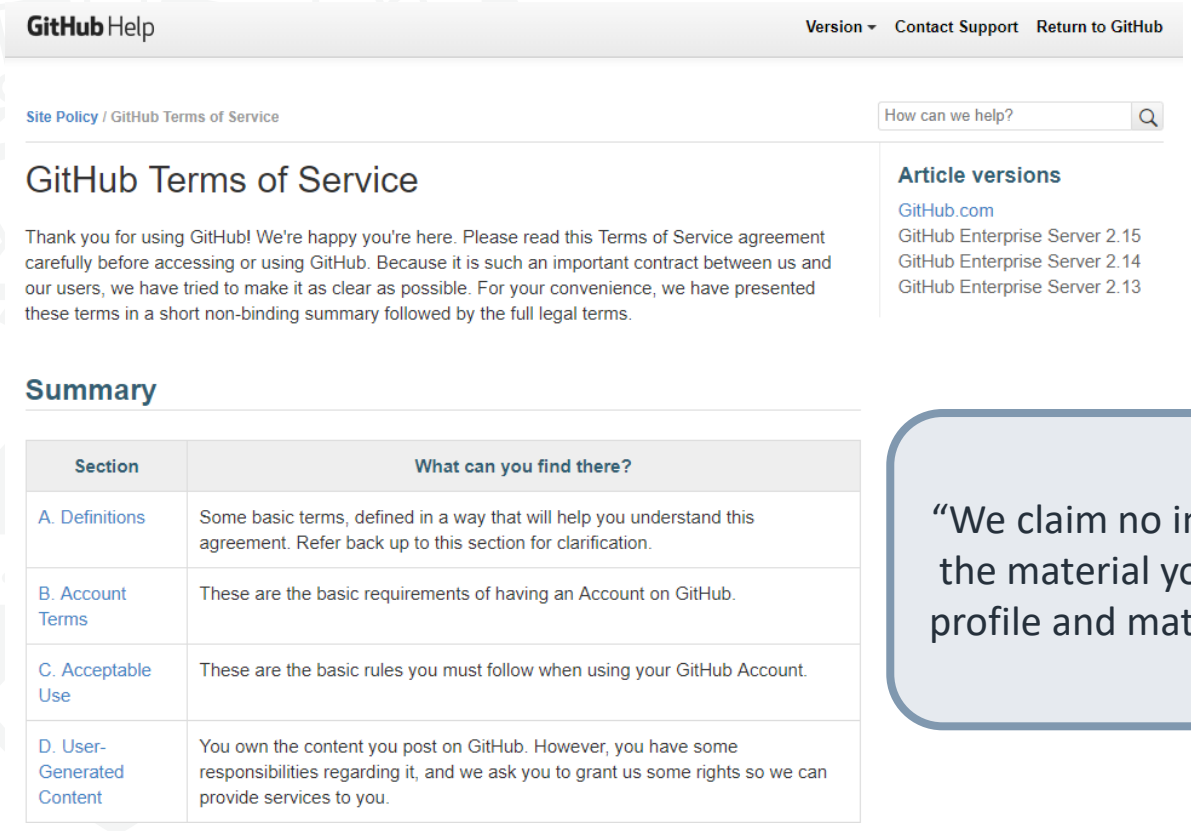
HTML 4 Updated 25 days ago

Top languages

HTML Jupyter Notebook Python R Scala

People 7

GitHub



GitHub Help

Version ▾ Contact Support Return to GitHub

Site Policy / GitHub Terms of Service

How can we help?

GitHub Terms of Service

Thank you for using GitHub! We're happy you're here. Please read this Terms of Service agreement carefully before accessing or using GitHub. Because it is such an important contract between us and our users, we have tried to make it as clear as possible. For your convenience, we have presented these terms in a short non-binding summary followed by the full legal terms.

Article versions

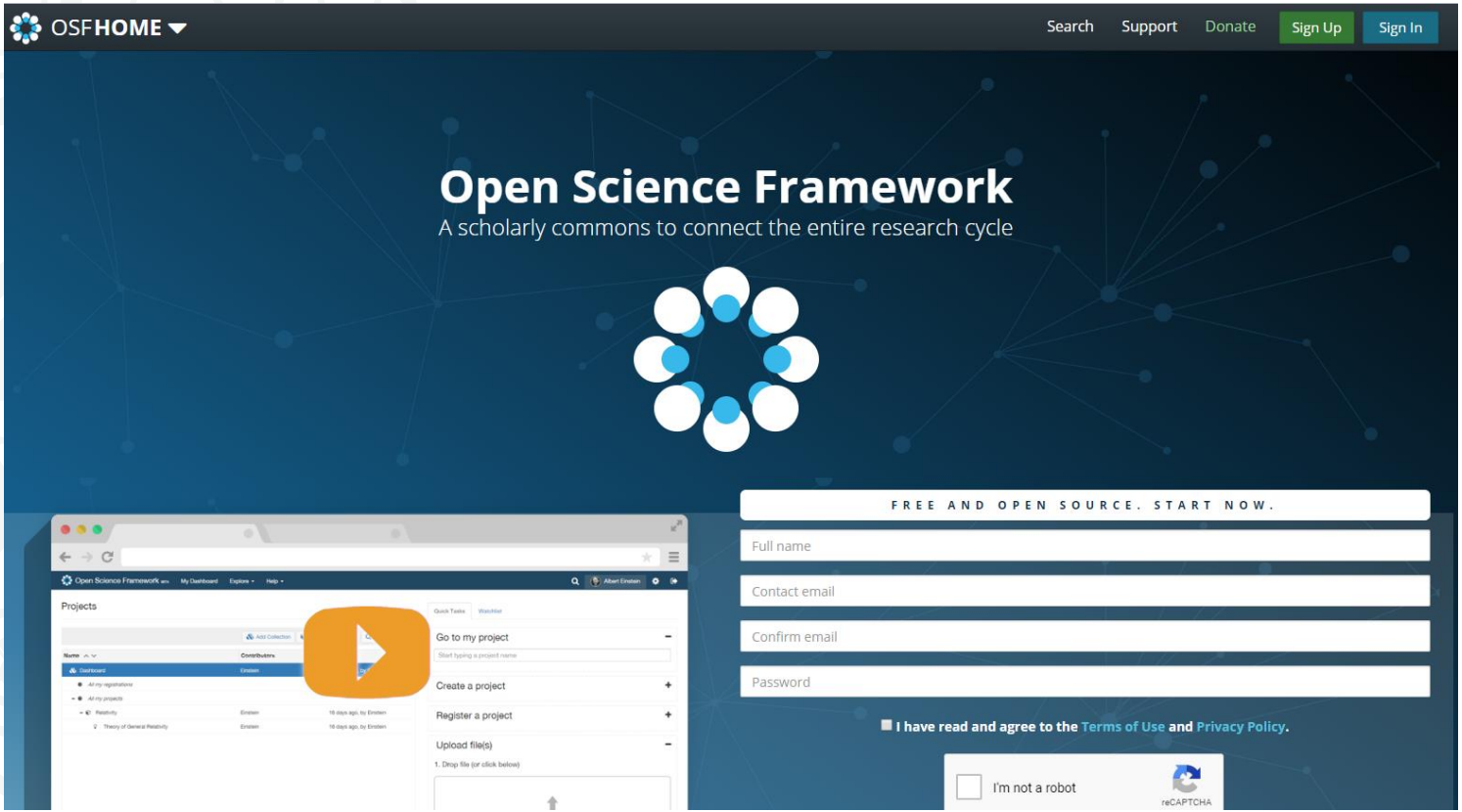
GitHub.com
GitHub Enterprise Server 2.15
GitHub Enterprise Server 2.14
GitHub Enterprise Server 2.13

Summary

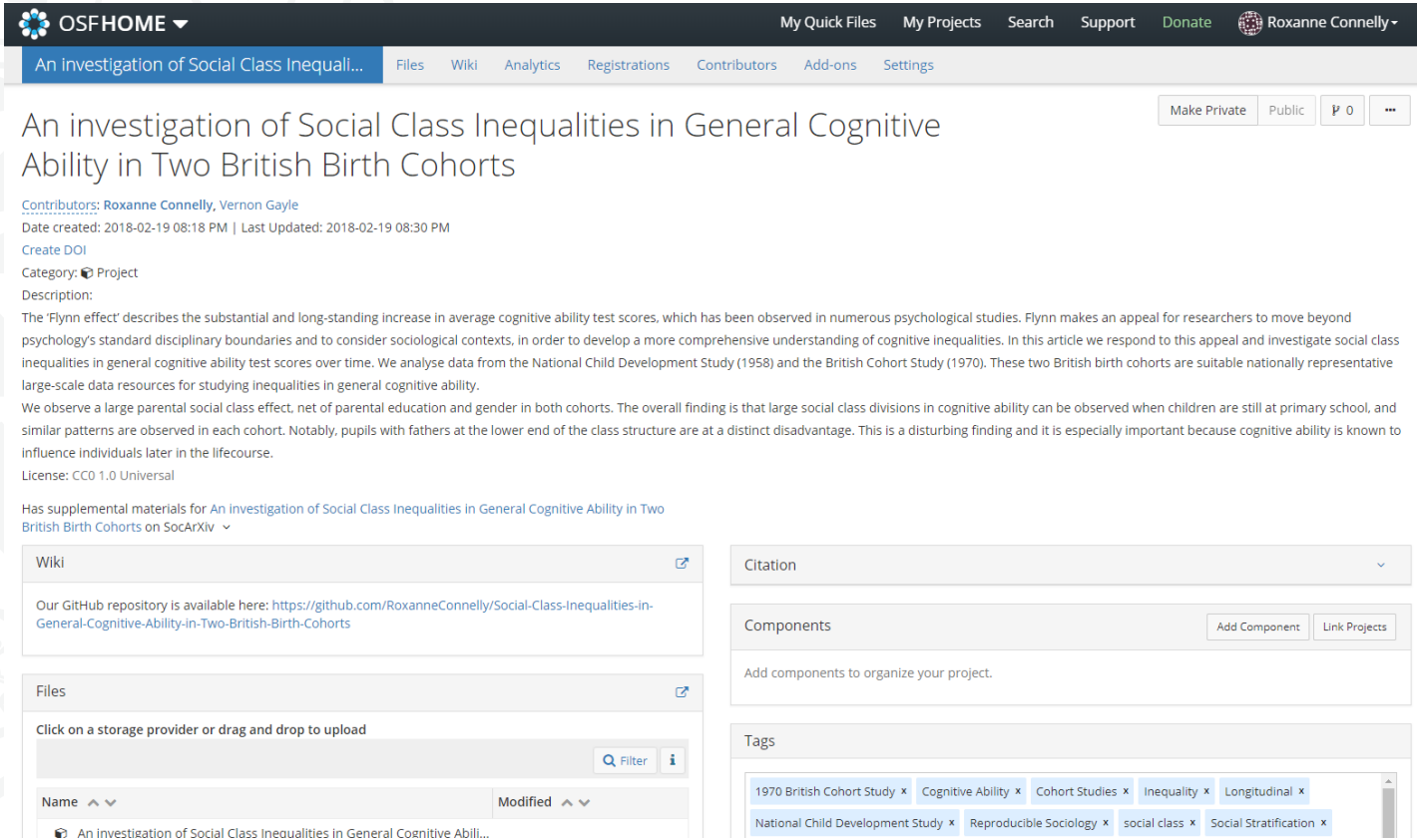
Section	What can you find there?
A. Definitions	Some basic terms, defined in a way that will help you understand this agreement. Refer back up to this section for clarification.
B. Account Terms	These are the basic requirements of having an Account on GitHub.
C. Acceptable Use	These are the basic rules you must follow when using your GitHub Account.
D. User-Generated Content	You own the content you post on GitHub. However, you have some responsibilities regarding it, and we ask you to grant us some rights so we can provide services to you.

"We claim no intellectual property rights over the material you provide to the Service. Your profile and materials uploaded remain yours."

Open Science Framework (https://osf.io/)



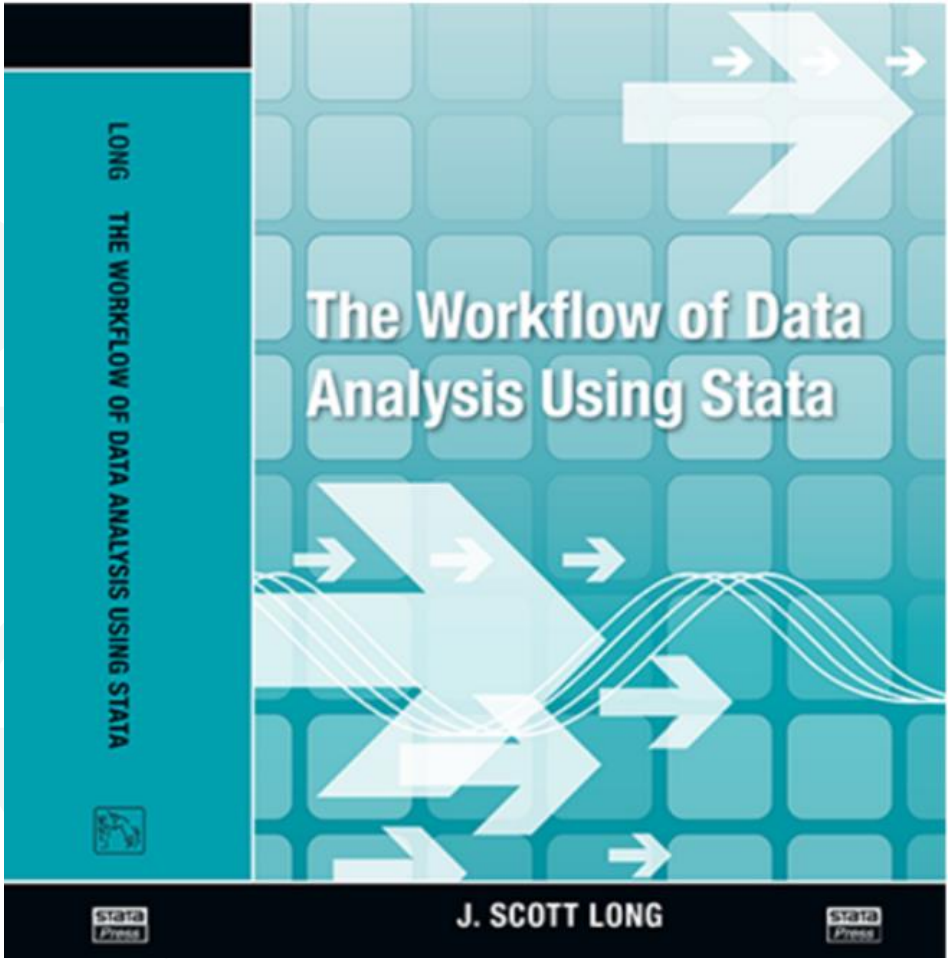
Open Science Framework (https://osf.io/)



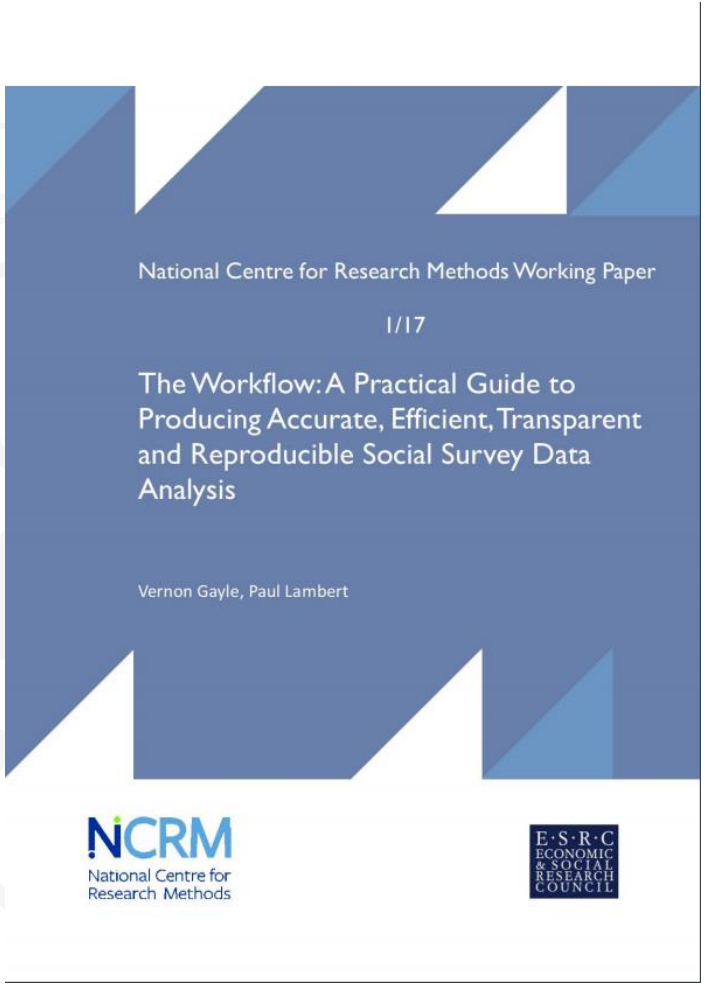

VERY POOR
POOR
FAIR
GOOD
EXCELLENT

No Drop Down Menus

- GUIs will leave you in a sticky mess!



Long, J.S., 2009. The workflow of data analysis using Stata. College Station, TX: Stata Press.



UNIVERSITY of York

Gayle, V.J. and Lambert, P.S., 2017. The workflow: A practical guide to producing accurate, efficient, transparent and reproducible social survey data analysis. National Centre for Research Methods.

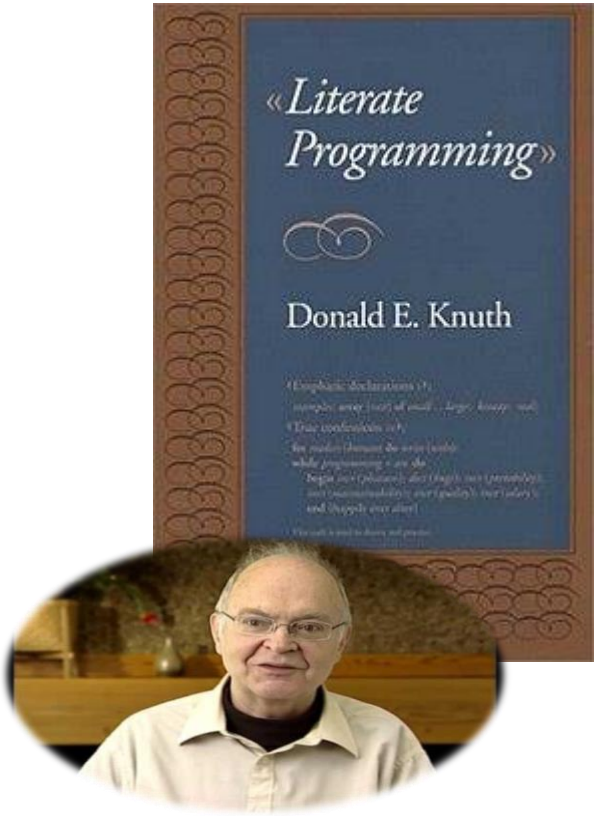
Effective Workflows

1. Accuracy
2. Efficiency
3. Standardization
4. Automation
5. Simplicity
6. Usability
7. Scalability

Literate Programming

- Literate programming is a methodology that combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained.

- The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer.



Research Practice

- **Pair programming** is an agile software development technique in which two programmers work together at one workstation. One, the driver, writes code while the other, the observer or navigator, reviews each line of code as it is typed in.

- **Code peer review** is a software quality assurance activity in which one or several humans check a program mainly by viewing and reading parts of its source code, and they do so after implementation or as an interruption of implementation.

Is this achievable in a plain text document (e.g. a .do file, R script)?

Research Objects

Research Objects (ROs) are ‘semantically rich aggregations of (potentially distributed) resources that provide a layer of structure on top of information delivered as Linked Data… An RO bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation.”

(Bechhofer et al. 2013)

S. Bechhofer et al., “Why linked data is not enough for scientists,” Future Generation Computer Systems, vol. 29, no. 2, pp. 599–611, Feb. 2013. http://dx.doi.org/10.1016/j.future.2011.08.004.

Dynamic Documents

Dynamic Documents

UNIVERSITY of York

Code on Steroids

Dynamic Documents

- Dynamic documents refer to the kind of source documents containing both program code and narratives.

Explorable Multiverse Analyses

- Dragicevic, Jansen, Sarma, Kay, and Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA.

- https://explorablemultiverse.github.io/

The British Journal of Sociology 2019 Volume 70 Issue 1

BJS

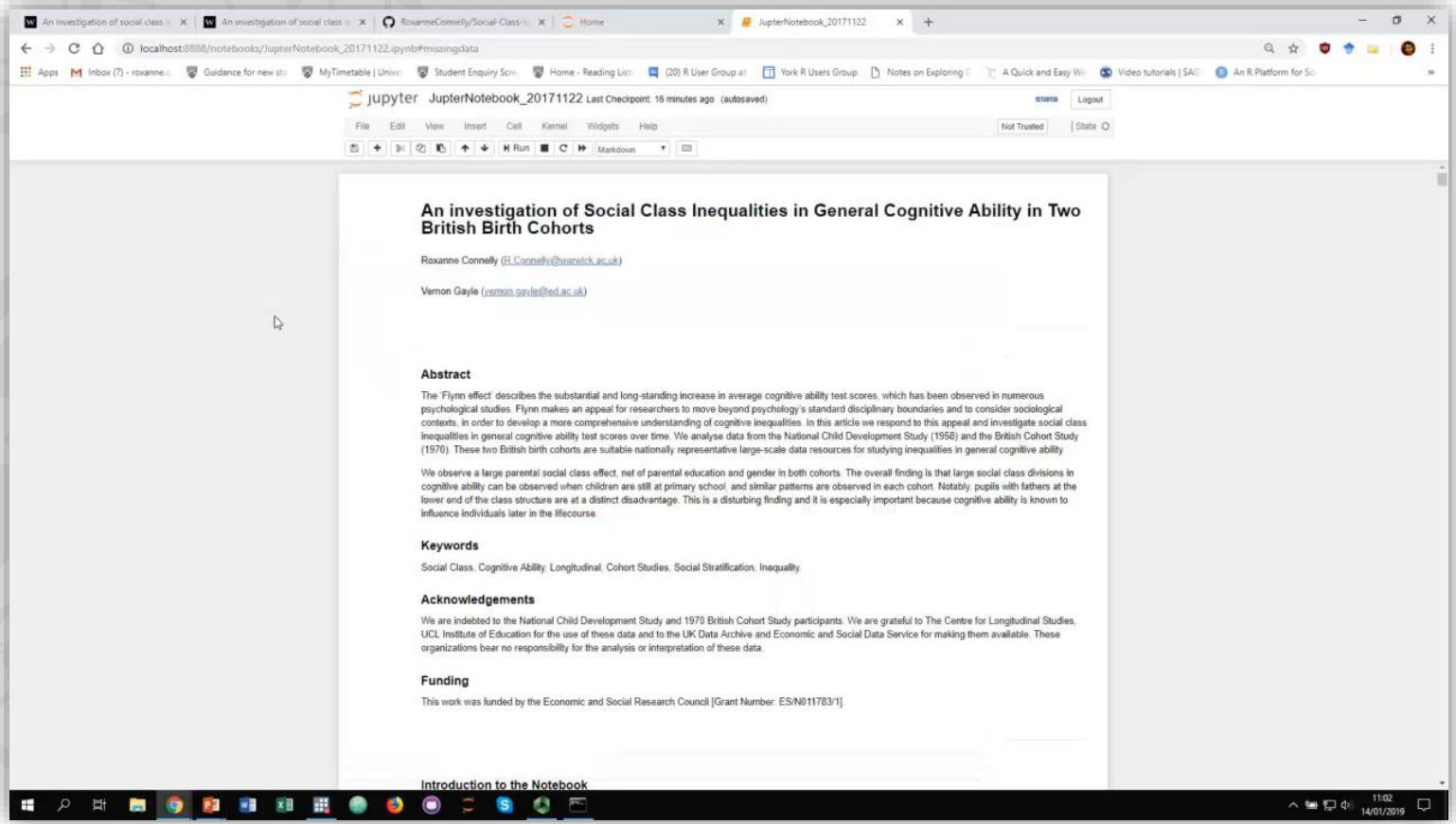
An investigation of social class inequalities in general cognitive ability in two British birth cohorts[1]

Roxanne Connelly and Vernon Gayle

Abstract

The 'Flynn effect' describes the substantial and long-standing increase in average cognitive ability test scores, which has been observed in numerous psychological studies. Flynn makes an appeal for researchers to move beyond psychology's standard disciplinary boundaries and to consider sociological contexts, in order to develop a more comprehensive understanding of cognitive inequalities. In this article we respond to this appeal and investigate social class inequalities in general cognitive ability test scores over time. We analyse data from the National Child Development Study (1958) and the British Cohort Study (1970). These two British birth cohorts are suitable nationally representative large-scale data resources for studying inequalities in general cognitive ability. We observe a large parental social class effect, net of parental education and gender in both cohorts. The overall finding is that large social class divisions in cognitive ability can be observed when children are still at primary school, and similar patterns are observed in each cohort. Notably, pupils with fathers at the lower end of the class structure are at a distinct disadvantage. This is a disturbing finding and it is especially important because cognitive ability is known to influence individuals later in the lifecourse.

Connelly, R. and Gayle, V., 2019. An investigation of social class inequalities in general cognitive ability in two British birth cohorts. The British journal of sociology, 70(1), pp.90-108.


jupyter JupterNotebook_20171122 Last Checkpoint: 16 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Run
Markdown
An investigation of Social Class Inequalities in General Cognitive Ability in Two British Birth Cohorts
Roxanne Connelly (R.Connelly@warwick.ac.uk)
Vernon Gayle (vernon.gayle@ed.ac.uk)
Abstract
The 'Flynn effect' describes the substantial and long-standing increase in average cognitive ability test scores, which has been observed in numerous psychological studies. Flynn makes an appeal for researchers to move beyond psychology's standard disciplinary boundaries and to consider sociological contexts, in order to develop a more comprehensive understanding of cognitive inequalities. In this article we respond to this appeal and investigate social class inequalities in general cognitive ability test scores over time. We analyse data from the National Child Development Study (1958) and the British Cohort Study (1970). These two British birth cohorts are suitable nationally representative large-scale data resources for studying inequalities in general cognitive ability.
We observe a large parental social class effect, net of parental education and gender in both cohorts. The overall finding is that large social class divisions in cognitive ability can be observed when children are still at primary school, and similar patterns are observed in each cohort. Notably, pupils with fathers at the lower end of the class structure are at a distinct disadvantage. This is a disturbing finding and it is especially important because cognitive ability is known to influence individuals later in the lifecourse.
Keywords
Social Class, Cognitive Ability, Longitudinal, Cohort Studies, Social Stratification, Inequality
Acknowledgements
We are indebted to the National Child Development Study and 1970 British Cohort Study participants. We are grateful to The Centre for Longitudinal Studies, UCL Institute of Education for the use of these data and to the UK Data Archive and Economic and Social Data Service for making them available. These organizations bear no responsibility for the analysis or interpretation of these data.
Funding
This work was funded by the Economic and Social Research Council [Grant Number: ES/N011783/1].
Introduction to the Notebook

Making Dynamic Documents

Dynamic Documents

- Software Specific Dynamic Documents:
  - Stata Dynamic Documents
  - R Markdown

- Language Agnostic Dynamic Documents:
  - Jupyter Notebooks

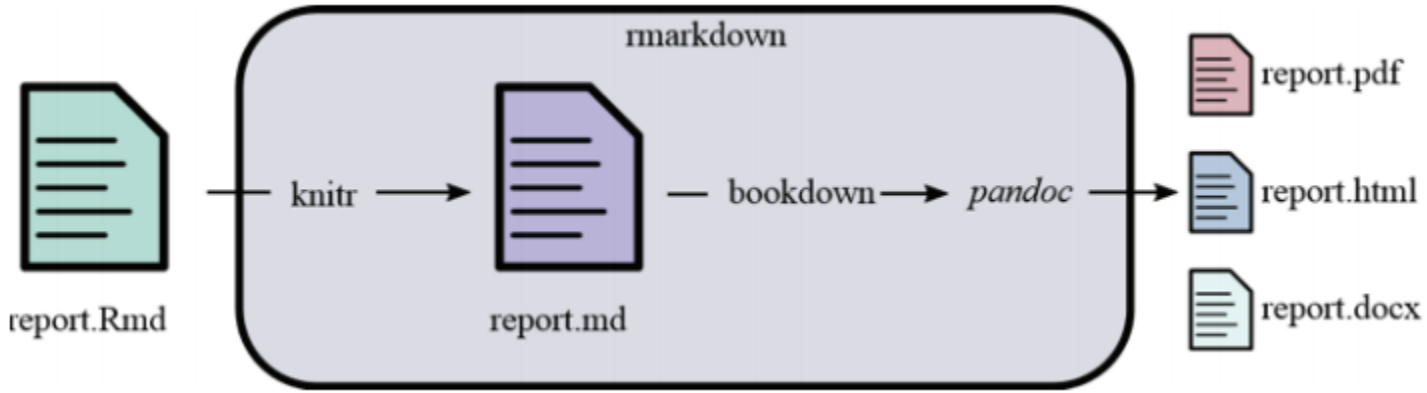
R Markdown

- R Markdown is described as an ‘authoring framework’:
  - You can connect to data and run code:
  - Generate reports of your work,

- An R Markdown file is a simple plain text file with code, text and meta data.

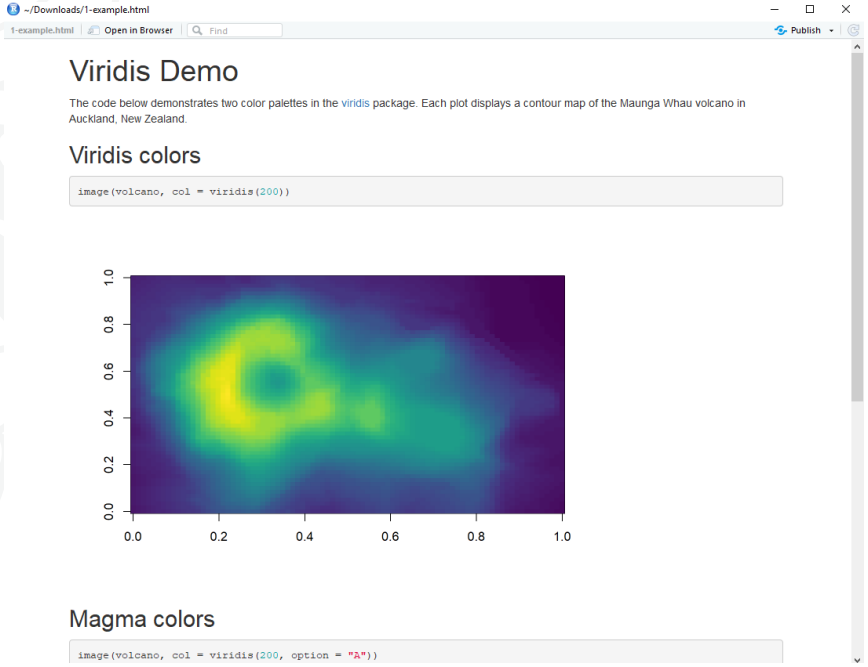
- The finished document will embed the code, narrative and results.

R Markdown



Source: https://www.britishecologicalsociety.org/wp-content/uploads/2017/12/guide-to-reproducible-code.pdf

R Markdown



https://rmarkdown.rstudio.com/lesson-2.html

UNIVERSITY of York

Jupyter Notebooks

Jupyter Notebooks

https://jupyter.org/

The Ipython/Jupyter Notebook

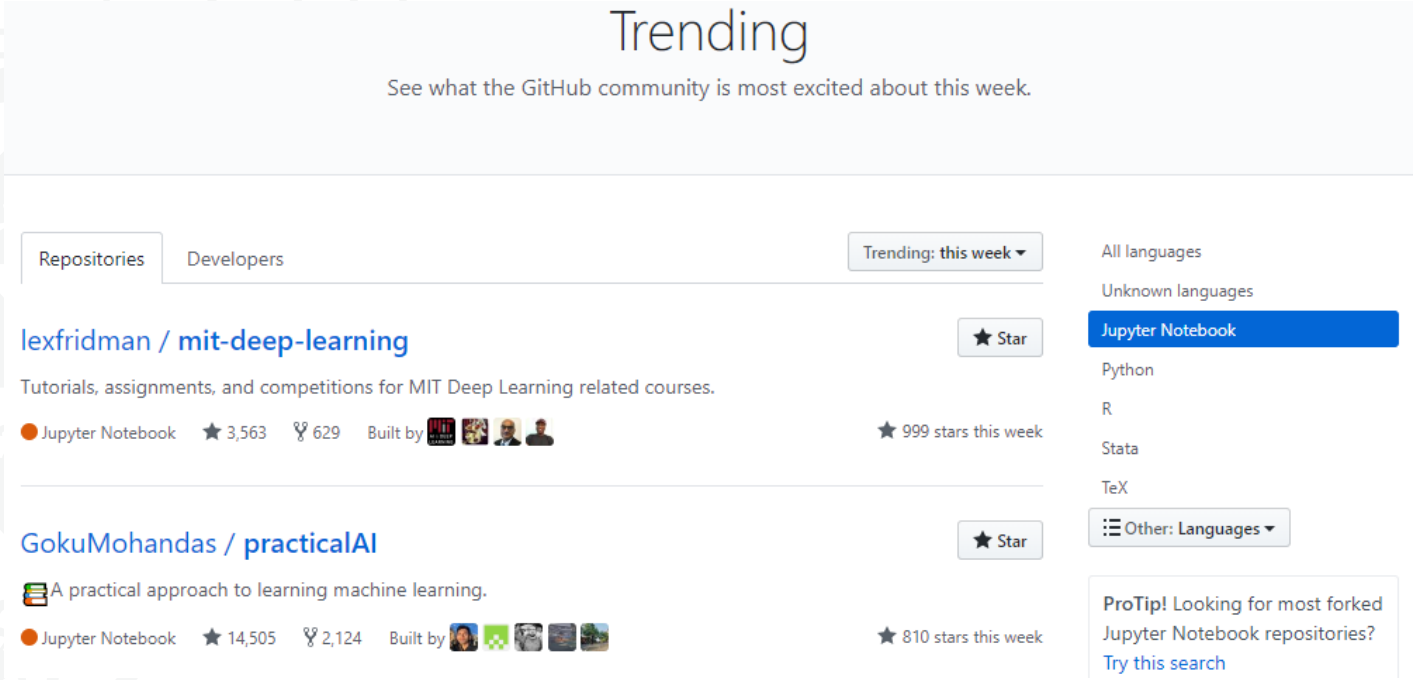
- Text and Maths
- Code
- Code Output
- Results

- Accessible through your web browser

Jupyter Notebooks

https://hub.mybinder.org/user/jupyterlab-jupyterlab-demo-olagq5sj/lab

Jupyter Notebooks



https://github.com/trending/jupyter-notebook?since=weekly

Jupyter Notebooks

UNIVERSITY of York

Gravitational Wave Open Science Center

- Getting Started
- Data
  - Catalogs
  - Bulk Data
- Tutorials
- Software
- Detector Status
- Timelines
- My Sources
- GPS ↔ UTC
- About the detectors
- Projects
- Acknowledge GWOSC

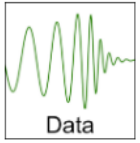
Getting Started

Welcome! The Gravitational Wave Open Science Center (GWOSC) provides access to LIGO and Virgo data, as well as documentation, tutorials, and online tools for finding and viewing data.

What are gravitational-wave interferometers?

For a general introduction, see the LIGO Scientific Collaboration and Virgo Collaboration websites

Where's the data?



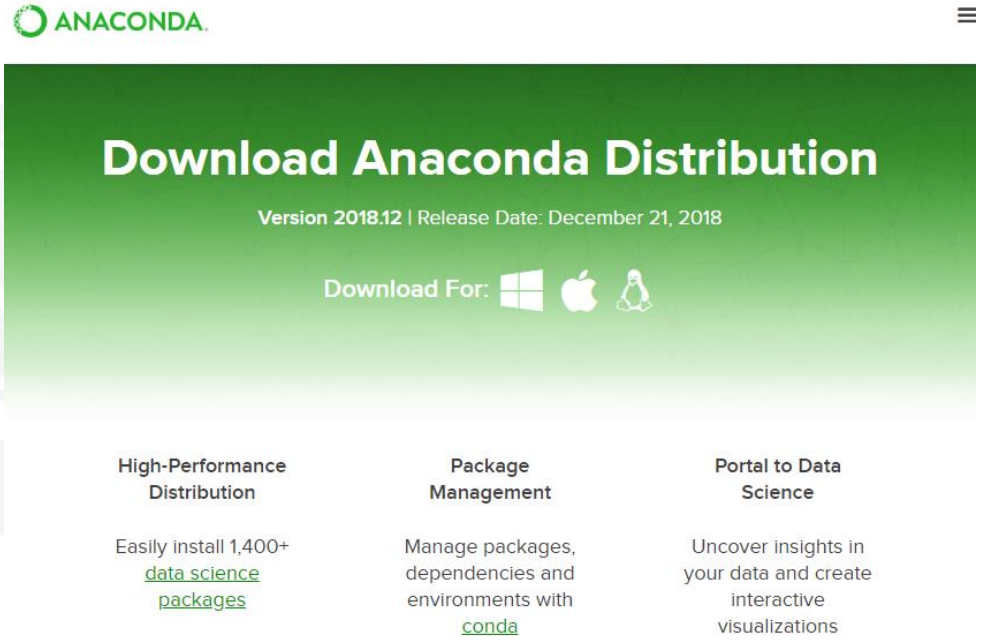
The Data Page allows you to download LIGO and Virgo data.

The main data are time series sampled at 16384 Hz or 4096 Hz.

Data are calibrated so that gravitational wave signals have units of dimensionless strain (ΔL / L).

https://www.gw-openscience.org/start/

Download Jupyter Notebooks



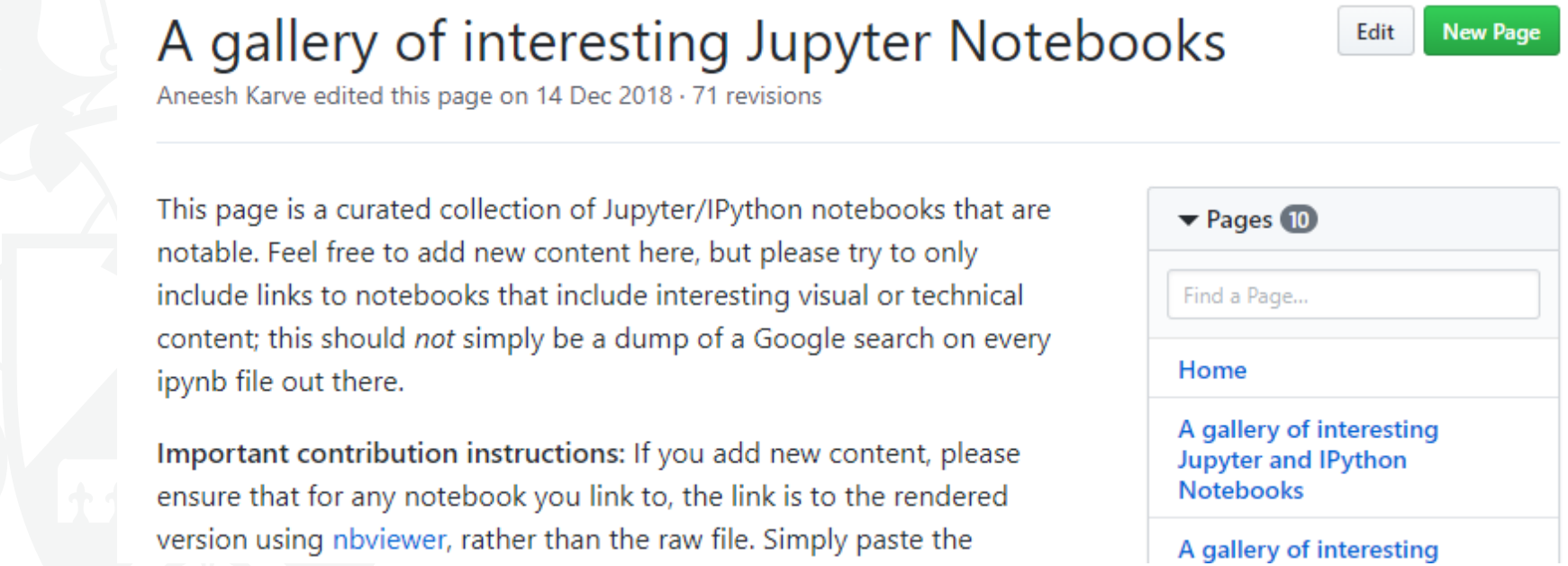
https://www.anaconda.com/download/

Jupyter Kernels

- R Kernel (IRKernel): https://github.com/IRkernel/IRkernel
- Stata Kernel (stata_kernel) https://github.com/kylebarron/stata_kernel

Jupyter: Quick Demo

Jupyter Notebooks



A gallery of interesting Jupyter Notebooks

Edit | New Page

Aneesh Karve edited this page on 14 Dec 2018 · 71 revisions

This page is a curated collection of Jupyter/IPython notebooks that are notable. Feel free to add new content here, but please try to only include links to notebooks that include interesting visual or technical content; this should *not* simply be a dump of a Google search on every ipynb file out there.

Important contribution instructions: If you add new content, please ensure that for any notebook you link to, the link is to the rendered version using nbviewer, rather than the raw file. Simply paste the

Pages 10

Find a Page...

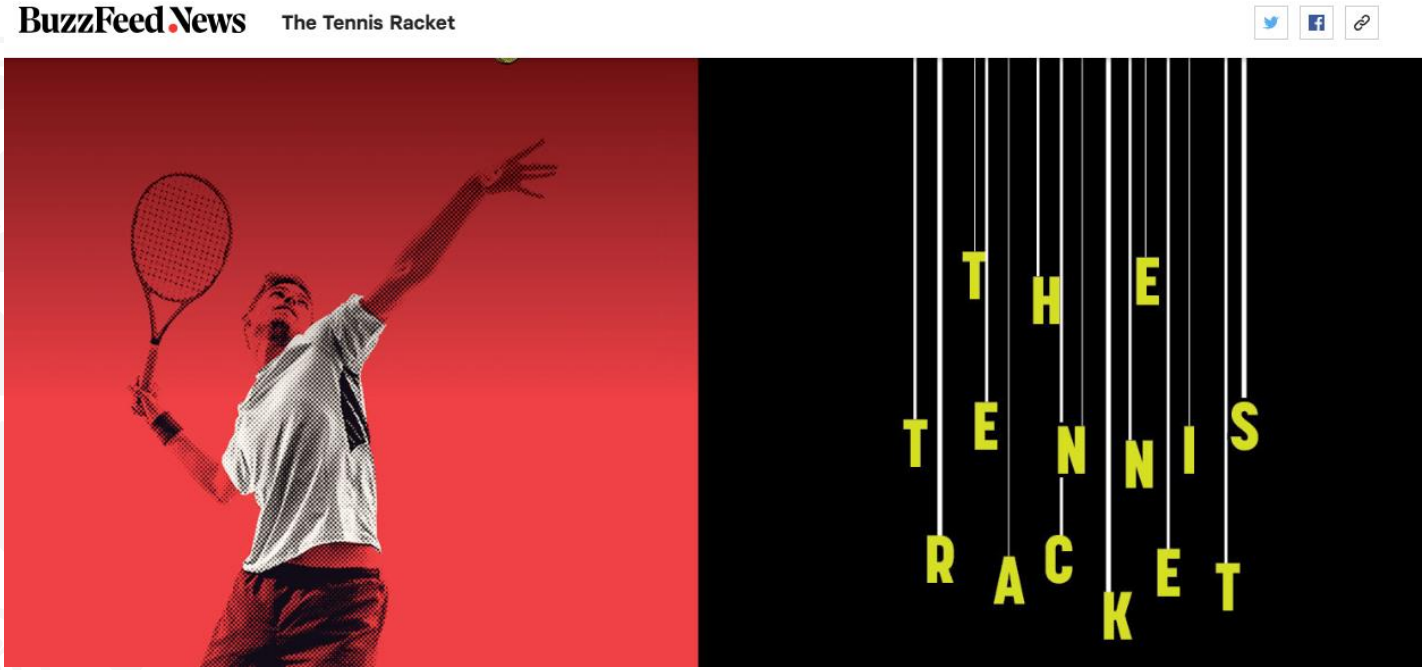
Home

A gallery of interesting Jupyter and IPython Notebooks

A gallery of interesting

https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#introductory-tutorials

Jupyter Notebooks



https://www.buzzfeednews.com/article/heidiblake/the-tennis-racket
https://github.com/BuzzFeedNews/2016-01-tennis-betting-analysis

Summary

Reflections

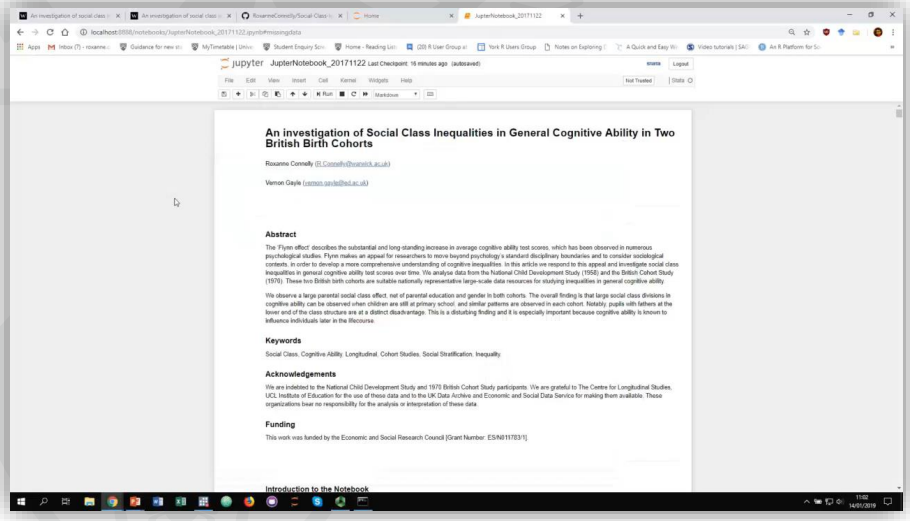
- Dynamic documents are a tool not a panacea.

- The researcher needs to change the way they work:
  - I feel that undertaking research within a dynamic document encourages me to work in a more ‘literate’ manner.

Reflections

- We need to think about how these tools work with social science data that cannot be shared openly.
  - Could we link Jupyter Notebooks up directly with the Data Archives (with researcher log-in required)?
  - How should we work with secure data (e.g. administrative data)?
- We need to continue to think about what should be included in a dynamic document (so it can constitute a research object):
  - Do we need discipline specific guidelines and templates?

Open Social Science



The Unknown

Roxanne Connelly
roxanne.connelly@york.ac.uk
@ConnellyRoxanne

These materials draw heavily on collaborative work developed with **Professor Vernon Gayle, University of Edinburgh**. His contribution is gratefully acknowledged.