

Adventures in Open & Reproducible Research

Charlie Rahal

LCDS, Department of Sociology and Nuffield College

University of Oxford, 16th May, 2023



In the spirit of ‘Open Science’:

crahal.github.io/files/dspi_talk.pdf

and *all* of the research you see here today is at:

github.com/crahal

Lots of examples: hopefully something for everyone!

Everything here is made with free, and ‘Open’ tools:

- Linux (operating system)
- Python (data wrangling and statistical software)
- $\text{\LaTeX}$ (typesetting)

Everything here is made with free, and ‘Open’ tools:

- Linux (operating system)
- Python (data wrangling and statistical software)
- $\text{\LaTeX}$ (typesetting)

Why mention this? We’re talking about ‘Open’ Science!

What is ‘Open Science’?

‘The movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional’

Pragmatically:

1. Allows creation of value and advancement.
2. Allows verification of the scientific record.
3. Allows public disclosure and engagement.
4. Allows science to become accessible/transparent.

What motivates a talk about ‘Open Science’?

- Three anecdotes from my time as a doctoral student:

What motivates a talk about ‘Open Science’?

- Three anecdotes from my time as a doctoral student:
 1. Make sure your coauthors have code.

What motivates a talk about ‘Open Science’?

- Three anecdotes from my time as a doctoral student:
 1. Make sure your coauthors have code.
 2. Critically analyze everything.

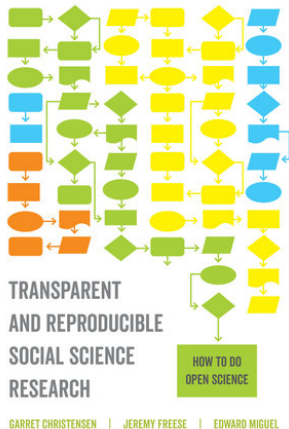
What motivates a talk about ‘Open Science’?

- Three anecdotes from my time as a doctoral student:
 1. Make sure your coauthors have code.
 2. Critically analyze everything.
 3. Consider using open-source software.

What motivates a talk about ‘Open Science’?

- Three anecdotes from my time as a doctoral student:
 1. Make sure your coauthors have code.
 2. Critically analyze everything.
 3. Consider using open-source software.
- I’m presently the LNL Co-ordinator for the UKRN.

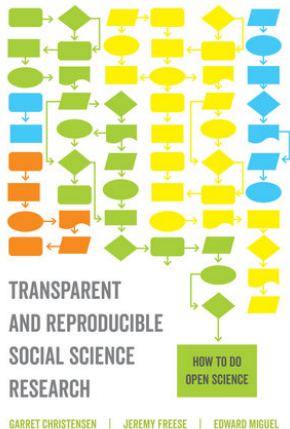
More traditional motivation, just in case you aren't sold...:



This book...

Is most definitely worth
reading!*

More traditional motivation, just in case you aren't sold...:



This book...

Is most definitely worth
reading!*

*Not just because one of
the authors is a coauthor of
a paper included later!

It does an **excellent** job of
motivating replications.

One of the most important parts of the book is this table:

Compared to Original	Original Procedures	Key Differences
Same Data	Verification	Reanalysis
Different Data	Direct Replication	Extension

- I previously taught a 'Replication Project' on the MPhil in Sociology and Demography.
- Increasing push towards teaching concepts of 'Replication'.
- This is one of the reasons why Sociology has moved towards.

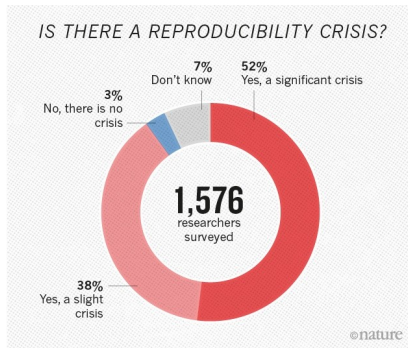
How it started:

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • <https://doi.org/10.1371/journal.pmed.0020124>

How it's going:



What does ‘Open Science’ look like?

1. Preprints and ‘Open Access’.
2. Publicly available data.
3. Open source code



What does ‘Open Science’ look like?

1. Preprints and ‘Open Access’.
2. Publicly available data.
3. Open source code


HARVARD
 Dataverse

Add Data ▾
 Search ▾
 About
 User Guide
 Support
 Sign Up
 Log In

Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

Add a dataset ➕

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

Add a dataverse ➕

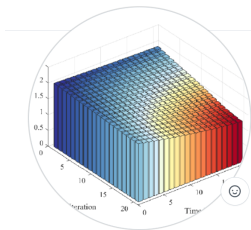
Publishing your data is easy on Harvard Dataverse!

Learn about getting started creating your own dataverse repository here.

Getting started ↗

What does 'Open Science' look like?

1. Preprints and 'Open Access'.
2. Publicly available data.
3. Open source code



Pinned

Customize your pins

[GWASReview](#)

Public ::

A Scientometric Review of Genome Wide Association Studies

HTML ☆ 15 🗑️ 7

[NHSSpend](#)

Public ::

A home for the NHSSpend Library

HTML ☆ 1 🗑️ 1

[InterModel_Vigorish](#)

Public ::

Introducing the InterModel Vigorish (IMV)

Jupyter Notebook ☆ 1

[OxfordDemSci/gwasdiversitymonitor](#)

Public ::

Source code for the GWAS Diversity Monitor

CSS ☆ 12 🗑️ 1

These slides are made with an open source typesetting language!

I_make_this_slide.tex

```
\begin{frame}[]
  These slides are made with an open source typesetting language!
  \begin{center}
    \begin{scriptsize}
      \begin{minipage}{.95\linewidth}
        \begin{myenv}{guess$_\_what$__this$__does.py}
          \begin{minted}{latex}
            # Hello friends :)
          \end{minted}
        \end{myenv}
      \end{minipage}
    \end{scriptsize}
  \end{center}
\end{frame}
```

Open source statistical code can come in the form of Python (or R):

guess_what_this_does.py

```
import json, io, imageio, requests
from PIL import Image

def get_cat_bounded(url):
    api_ret = requests.get(url)
    if api_ret.status_code==200:
        text = json.loads(api_ret.text)
        while (text[0]['width']<=400 or text[0]['width']>=600) or \
            (text[0]['height']<=400 or text[0]['height']>=600):
            api_ret = requests.get(url)
            if api_ret.status_code==200:
                text = json.loads(api_ret.text)
            r = requests.get(api_text[0]['url'])
            return Image.open(io.BytesIO(r.content))

url = 'https://api.thecatapi.com/v1/images/search'
images = []
for _ in range(0, 50):
    images.append(get_cat_bounded(url))
imageio.mimsave('CATS.gif', images, fps=1)
```

Introduction and Motivation
oooooooooooooooo●ooo

Scientometrics
ooooooo

Civic Technology
ooooooooo

Inequality
oooo

Model Uncertainty
ooooo

PRNGs
ooooo

Conclusion
oo

Thats right: Cats!

Why show you this? Well, Python is great at lots of useful things:

1. Scientific Computing (SciPy, numpy)
2. Data visualisation (Matplotlib, seaborn)
3. Web scraping (requests)
4. Data wrangling (Pandas)

A natural choice for un/semi-structured data. Python also excels at NLP/ML, but less with classical statistics.

Python is **great** at building new datasets. But why bother?

‘We have been using the same data sets over and over again, and since we wanted new answers, we have been developing new econometric techniques to try to transform the data, and get more meaningful information out of them. Just having a fresh, new data set brings a whole new perspective, and I think people are starting to realize that, and gradually people are becoming more interested in data collection itself’

A. Cavallo, 15 November 2012 in Taylor et al. (2014)

Python is **great** at building new datasets. But why bother?

‘We have been using the same data sets over and over again, and since we wanted new answers, we have been developing new econometric techniques to try to transform the data, and get more meaningful information out of them. Just having a fresh, new data set brings a whole new perspective, and I think people are starting to realize that, and gradually people are becoming more interested in data collection itself’

A. Cavallo, 15 November 2012 in Taylor et al. (2014)

New datasets allow fresh **insight** and are typically **open**!

But what is **semi-structured** data?

“A form of structured data that does not obey the tabular structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements”

Different to **tabular** surveys classical social scientists are used to.

But, often, our goal is to parse it and make it tabular all the same.

A lot of my semi-structured work is ‘**scientometric**’ and ‘**bibliometric**’ in nature:

- **Scientometrics** is the field of study which concerns itself with measuring and analysing science.
- **Bibliometrics** is the use of statistical methods to analyze quantitative information pertaining to books, articles and other publications.

These fields are naturally related to ‘**Open Science**’.

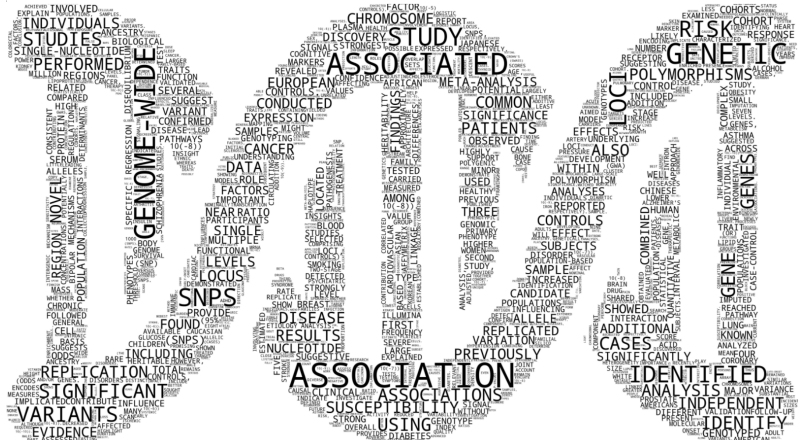
Data comes from various providers (usually APIs).

Lets use a 'Traffic Light' system to rank by openness:

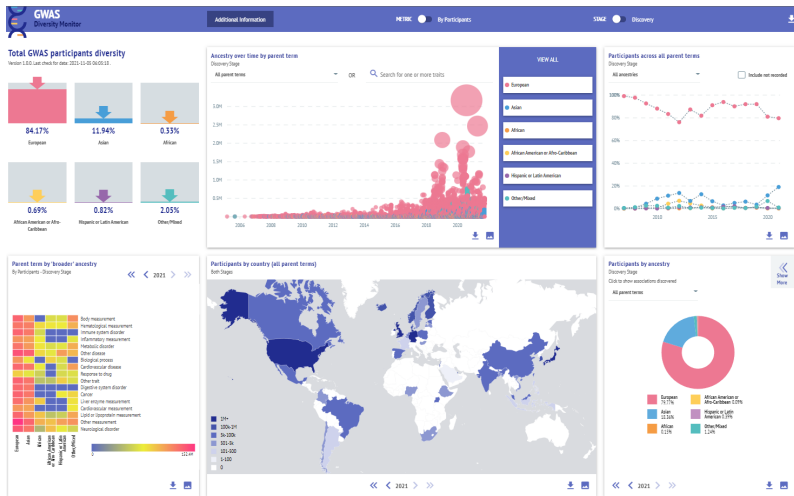
- Web of Science
- Google Scholar
- Scopus
- PubMed
- JSTOR/Constellate
- CrossRef
- The General Index
- Dimensions

This list gets longer every day...

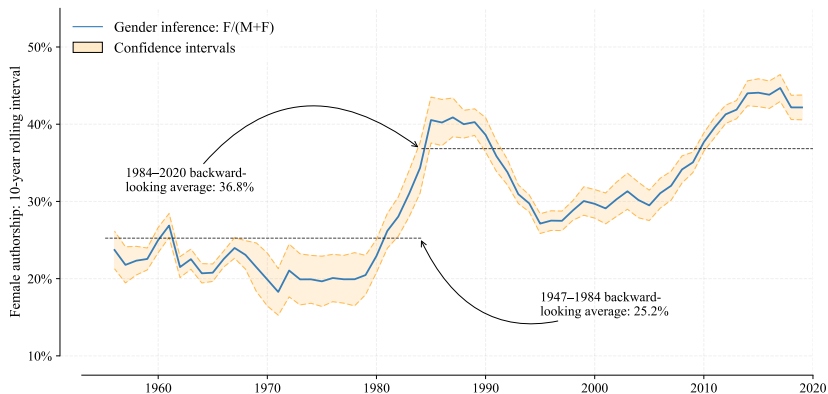
- My early scientometric work analyzed GWAS.
- It took curated lists of catalogued information from the the European Bioinformatics Institute. It then called **semi-structured** data from PubMed.
- We published two papers which:
 1. Conduct a thorough scientometric review.
 2. Launch a substantial interactive dashboard.



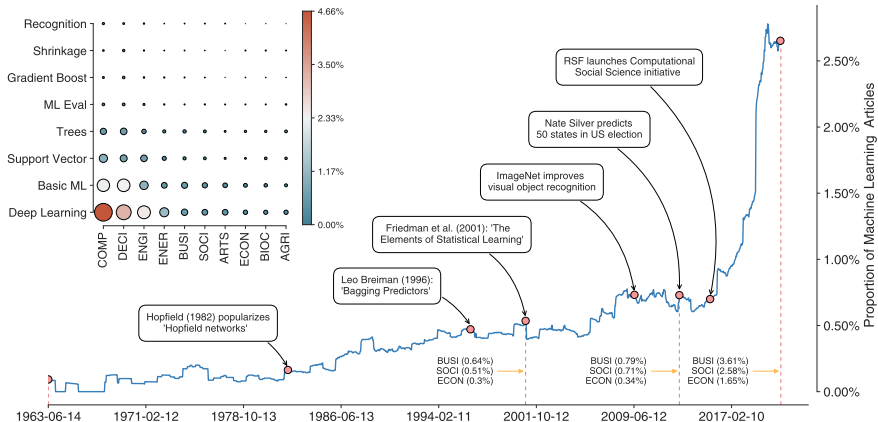
'The Scientometrics of Genome Wide Association Studies'
With Melinda C. Mills



‘The GWAS Diversity Monitor Tracks Diversity in Real Time’
with Melinda C. Mills



‘Population Studies at 75 Years: An empirical review’
 With Melinda C. Mills



'The Rise of Machine Learning in the Academic Social Sciences'

With Dave Kirk and Mark Verhagen

Administrative Data is an invaluable resource for Good.

ADR UK (2018)

- UK leads the world in terms of 'Openness' of government data.
- The value of public government data is increasingly recognised.
 - Increasingly recognised by the ESRC (e.g. ADRNs).
- For better or worse:

I'm more interested in 'hard to use', semi-structured data.

Administrative Data is an invaluable resource for Good.

ADR UK (2018)

- UK leads the world in terms of 'Openness' of government data.
- The value of public government data is increasingly recognised.
 - Increasingly recognised by the ESRC (e.g. ADRNs).
- For better or worse:

I'm more interested in 'hard to use', semi-structured data.

Using curated data from the ONS SRS would be too easy.

TRANSPARENCY INTERNATIONAL UK fighting corruption worldwide Open Access Home Other versions **DONATE** About

2,623 out of 74,435 meetings

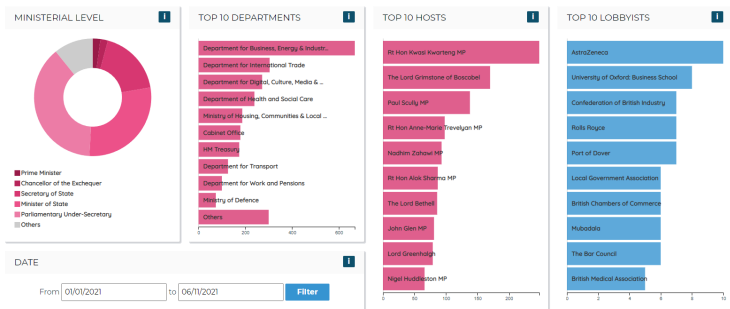
SEARCH BY LOBBYIST, HOST, SUBJECT

Reset

Reset

UK GOVERNMENT

LOBBYIST



openaccess.transparency.org.uk

With Ian Knowles, Paulo Serôdio, David Stuckler and TI UK

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

If only there was a **Government Data Service** to curate it...

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

If only there was a **Government Data Service** to curate it...

However, I am currently working with the ONS to improve collection!

Lets list a few of the problems with procurement data. Some are common to semi-structured data, some are not:

1. Some organisations don't believe in their legal obligation.
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations on the internet.)
3. Data commonly converted from a tabular file...

Lets list a few of the problems with procurement data. Some are common to semi-structured data, some are not:

1. Some organisations don't believe in their legal obligation.
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations on the internet.)
3. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema.)
4. Naturally, of course, there are no supplier UIDs.

Just one quick example, followed by what we can do...

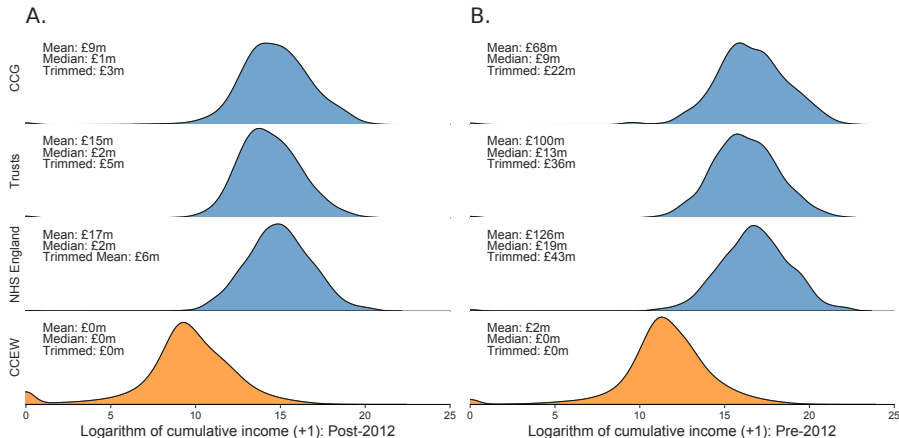
A3131

Expenditure Over Threshold

Department/Entity	City	Entity	Entity Sector	Subject	Transaction number	AP Amount	VAT registration number	Postcode/Invoice number
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	COMMUNITY SERVICES	11599633	606 000.00	6100071663
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	COMMUNITY SERVICES	11599635	400 000.00	6100071661
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	COMMUNITY SERVICES	11599637	15 000.00	6100071669
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	WYLLIE RESTORATION	11599642	64 000.00	6100071687
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11600729	68 226.00	6054836607
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING - NON ACUTE	11600730	68 226.00	6100071685
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	COMMUNITY SERVICES	11615604	30 000.00	6100071689
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	ACUTE COMMISSIONING	11615610	38 526.24	787424655
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	ACUTE COMMISSIONING	11615612	58 127.4	HL00056682
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	HOSPICES	11623337	218 000.00	871
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11636201	444 000.00	60482627
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11636202	218 126.1	60482627
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11642864	25 672.53	48413715
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11642923	-30 000.00	2000004184
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11642927	200 000.00	200041339
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11642932	-3 662.119 34	CM00218970
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	AMBULANCE SERVICES	11663375	587 757.29	100111174
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11666893	768 636.00	227 489
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11666895	760 530.34	233561
Department of health	Benoni	Benoni CCG	03042601	Cont Care Adult 100% Fully Funded	FUNDED NURSING CARE	11666922	26 512.00	2389731641
Department of health	Benoni	Benoni CCG	03042601	Charges From CSU	RECHARGES	11666995	85 249.00	ALM3036302CALM303658
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11673003	278 826.00	1206094
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11680115	82 568.00	CM00254903
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	COMMUNITY SERVICES	11701702	228 969.00	278586
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	COMMUNITY SERVICES	11701707	141 917.00	37932
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	COMMUNITY SERVICES	11701710	807 726.00	37892
Department of health	Benoni	Benoni CCG	03042601	Income Mx < 11%	BALANCE SHEET	11720068	47 733.1	6044623331
Department of health	Benoni	Benoni CCG	03042601	Income Mx < 11% - Student Loans	H/MRC	11720069	738.00	1206100555770 MAR 18
Department of health	Benoni	Benoni CCG	03042601	National Insurance - 1 vNED-RMS	BALANCE SHEET	11720065	29 114.84	6044623331
Department of health	Benoni	Benoni CCG	03042601	National Insurance - 1 vNED-RMS	H/MRC	11720068	19 284.18	1206100555770 MAR 18
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11724081	213 245.00	2029126
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11724085	268 141.00	2295894
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11724086	27 120.00	2029126
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11724099	266 424.00	231364
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11724114	224 739.00	231364
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	RECEIPTS FROM	11734031	440.00	CM00292676
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	11738262	-30 056.34	1365
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	11738269	55 540.00	CM0029686
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	COMMUNITY SERVICES	11748440	246 505.00	175451
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11748386	47 300.00	2010009982
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11771723	10 816.00	263322
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11771733	2 691 512.00	1000044300
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	MENTAL HEALTH CONTRACTS	11777627	25 202.00	800027387
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	COMMUNITY SERVICES - NON ACUTE	11777630	10 000.00	CM00015080
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	MENTAL HEALTH CONTRACTS	11782324	1 725 240.00	37897
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	MENTAL HEALTH CONTRACTS	11801176	36 000.00	37896
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	PLANNED CARE	11801435	10 000.00	484138351
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	11818020	258 781.00	37805
Department of health	Benoni	Benoni CCG	03042601	Hibiscus Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	11818040	55 506.00	37804
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	PLANNED SUPPORT SERVICES	11818100	1 416.00	6000019263
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	DAY THERAPIES	11818190	276 741.00	589
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	DAY THERAPIES	11818192	25 000.00	6000019263
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-voluntary Sector	DAY THERAPIES	11818193	25 000.00	6000019263
Department of health	Benoni	Benoni CCG	03042601	Clinical/Medical-Clinic Public Sector	LEARNING DIFFICULTIES	11854467	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11854468	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	11854469	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854470	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854471	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854472	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854473	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854474	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854475	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854476	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854477	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854478	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854479	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854480	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854481	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854482	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854483	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854484	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854485	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854486	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854487	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854488	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854489	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854490	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854491	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854492	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854493	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854494	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854495	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854496	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854497	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854498	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854499	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854500	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854501	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854502	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854503	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854504	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854505	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854506	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854507	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854508	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854509	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854510	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854511	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854512	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854513	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854514	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854515	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854516	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854517	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854518	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854519	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854520	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854521	250 000.00	6100071247
Department of health	Benoni	Benoni CCG	03042601	Health Srv Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	11854522	250 000.00	

Conclusion

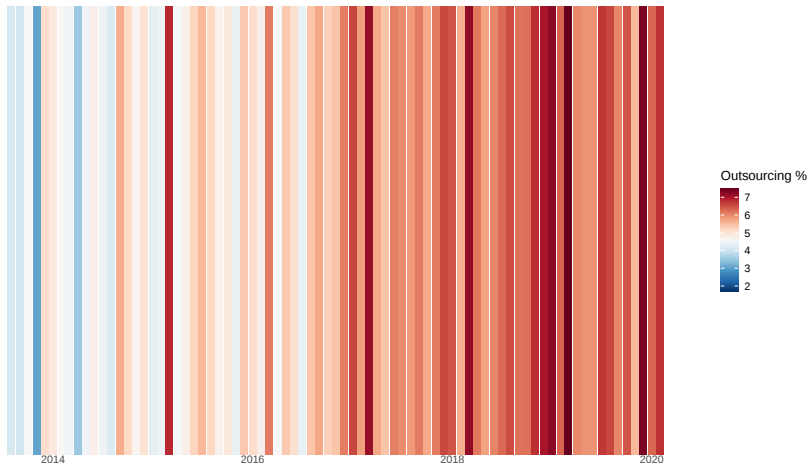




‘The Role of Non-Profits in Public Health Service Provision’

With John Mohan

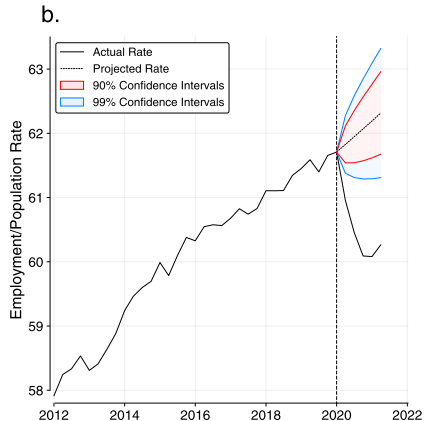
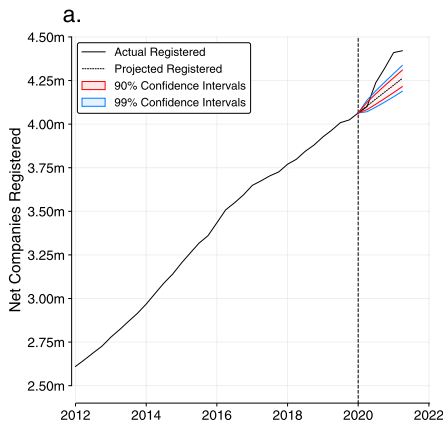
Creeping Privatisation in England's NHS



Source: Raw CCG payments approximately matched to Companies House and the Charity Commission, then manually verified

‘Creeping Commercialization in the NHS’

By Ben Goodair and Aaron Reeves

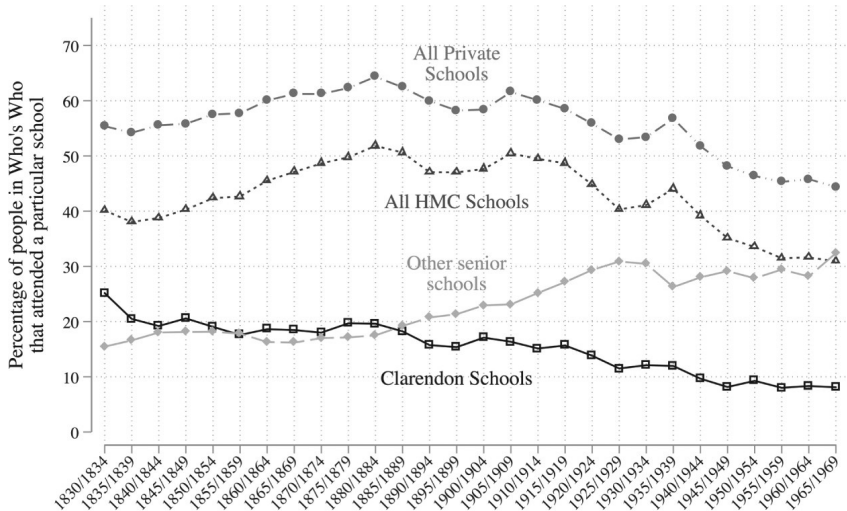


‘Modelling Labour Market Inequalities and the Covid-19 Pandemic’

With Aaron Reeves and Naomi Muggleton

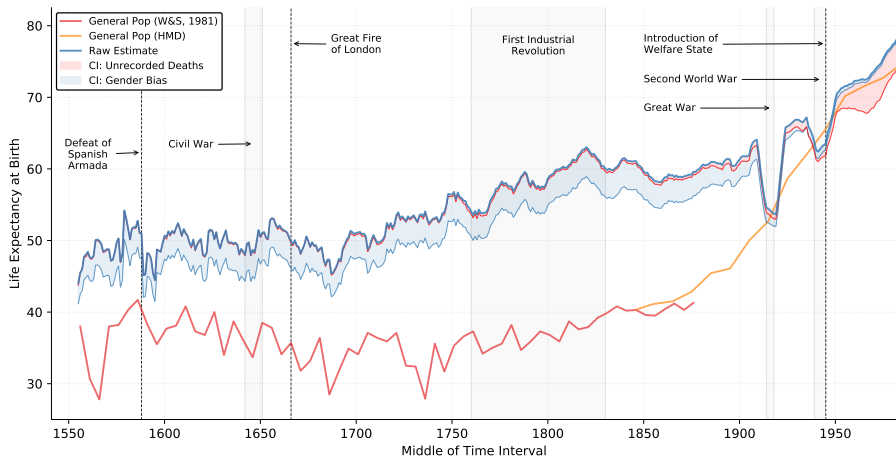
Semi-Structured Data for Inequality Research

- Data doesn't always come from APIs.
- Sometimes, it can come from raw `.html` itself.
- One example of this is when we scraped Who's Who to get information on education background of 'elites'.
- Another is when we scraped `thepeerage.com` to get data on dates of birth and death of British Peers.
- We might want to (there are competing views):
 - Check ToS and `robots.txt`.
 - Contact the webmaster first.



‘The decline and persistence of the old boy’

With Aaron Reeves, Sam Friedman. and Magne Flemmen



‘The Legacy of Longevity: Inequalities in UK Life Expectancy 1500-2016’ with Aaron Reeves, Felix Tropsch and Darryl Lundy

Ongoing research (With Aaron Reeves and Naomi Muggleton) uses and probate records from 1858-1996.

The idea is to:

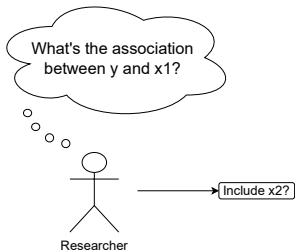
1. Scrape 800k .pdfs containing 24m probates: **Done!**
2. Preprocess images ready for OCR: **Done!**
3. Run OCR on HPC in parallel: **Done!**
4. Build lifetables to calculate e_0 by wealth at death...: **Pending!**

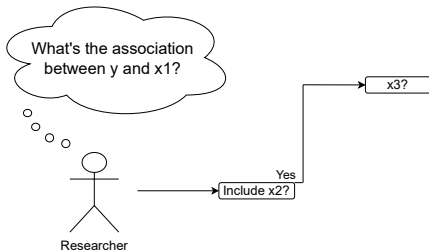
689	ABBOTT Henry Bernard of 22 Isley Road, London died 18 April 1907 at 80. Probate London 19 May to Louisa Harriette Abbott widow. Effects £21 10s. 6d.	ABBOTT Sarah Catherine of 32 Heywood-road Seven Kings (wife of Benjamin Harrier Abbott) died 18 Dec 1897. Probate London 19 February to the ss Benjamin Harrier Abbott father's saleman. Effects £209 15s.
	ABBOTT Isaac of Thorpston Northamptonshire bootmaker died 16 October 1908 Administration Peterborough 13 December to Sarah Hannah Abbott widow. Effects £221 10s.	ABBOTT Susan Georgina of Barchin House George-st Eastbourne (wife of Henry Susan Abbott) died 11 Mar 1908 Probate London 31 July to the ssd Henry Susan Abbott Lecturer-Cathedral His Majesty's Indian Affairs. Effects £680 1s. 3d.
	ABBOTT James of The Elms Road Gorse-road Ramsgate licensed victualler died 14 February 1908 Probate London 19 April to Elizabeth Abbott widow. Effects £284 14s. 2d.	ABBOTT Thomas of Duncroft gentleman died 17 Novemb 1907 Probate Wakefield 3 March to William Johnes Butler and Henry David Body auctioneers. Effects £768 12s. 4d. Rescued £7109 14s. 2d.
	ABBOTT John of 13 Leman-road Leythorpe Essex died 29 January 1908 Administration London 3 March to John Benjamin Abbott contractor. Effects £449 17s. 2d.	ABBOTT Thomas of 103 Abbe-street Rochdale-road Manchester died 29 March 1908 Probate Manchester 29 Apr to Thomas Edward Mellor and William Sedg compositions. Effects £108 11s.
	ABBOTT John of Betchford Lincolnshire died 20 February 1908 Administration London 24 February to Thomas Abbott farmer. Effects £293 4s.	ABBOTT Thomas of 30 Marshall-street Middlesbrough died 2 June 1908 Administration London 9 July to Effie Abbott widow. Effects £668 8s. 2d.
128	ABBOTT John of Mayfield House Northfield House Yorkshire died 28 June 1908 Probate London 10 September to Mary Abbott widow and William Abbott railway clerk. Effects £1916 12s. 4d.	ABBOTT Walter of Fairwood near Witham Essex died 2 November 1907 Probate London 29 July to Walter Perc Abbott and Osborne Hedley Abbott gentlemen. Effects £569 10s. 12d.
86	ABBOTT John White of 2 Westbury-gate Hanger Lane Ealing Middlesex died 21 April 1908 Probate London 27 June to Gertrude Agnes Abbott widow. Effects £2139 15s. 2d.	ABBOTT William of Kiverton-road Rawmarsh Butcher Yorkshire colliery-undert-manager died 1 November 1907 Probate Wakefield 15 February to Ellen Abbott widow Effects £134 10s. 8d.
	ABBOTT Joseph of 8 Progression-croft Clough Field Rawtestall Lincolnshire organist died 7 November 1908 Administration London 19 December to Maria Jane Abbott widow. Effects £432 14s. 2d.	ABBOTT William of Horncroft Lincolnshire died 23 January 1908 Probate Lincoln 23 March to John Elmy Chatterton clerk to the Governor. Effects £131.
3	ABBOTT Lucia Augusta of Banisterie Lowland Islands (widow of 8 June 1908 Administration (with Will) London 15 January to August Augustus Wade natus clerk attorney of 10 Abbe-st Abbott and Anne Margaret Stammers Abbott spinners. Effects £717 3s. 3d.	ABBOTT William of 68 Post-street Fitchburg Wigan townsmaster died 14 October 1908 Administration Liverpool 10 November to Sarah Ann Lowe (wife of Edwin Lowe Effects £57 3s.
165	ABBOTT Mary of Colsterworth Lincolnshire widow died 13 January 1908 Probate London 14 December to Charles Alfred Ernest Abbott bank clerk and George Edward Abbott bank manager. Effects £1042 15s. 4d.	ABBOTT William Maynes of 37 York-road Southend-on-Sea died 19 November 1908 Administration (with Will) London 3 July to George Abbott commercial traveller. Effects £30 10s. 2d.
	ABBOTT Martha of Heathcote Court Ayrton Hill End Northamptonshire (widow of 18 May 1908 Probate London 3 December to Elizabeth Abbott widow. Effects £149 18s. 10d.	ABBOTT Henry of The Cranhill Cottage near Woodrove Northamptonshire (widow of 18 May 1908 Probate London 3 December to Elizabeth Abbott widow. Effects £203 7s. 1d.
	ABBOTT Percy Willard of 33 Ashdale Park Villa Ravenscourt Park Middlesex died 5 May 1908 at sea. Administration London 1 August to 10 Arthur Abbott chemist's manager. Effects £443 15s. 3d.	ABBOTT Jane of West Beckham Norfolk widow died 27 Apr 1908 Probate Norwich 10 June to James Kemp maid Effects £109.
	ABBOTT Samuel of Cornwell Yorkshire died 21 June 1908 Administration (with Will) York 8 December to George Albert Teale Butcher and Christopher Arthur Baskin millers. Effects £43.	ABBOTT Robert of Roston Norfolk retired farmer died 4 October 1908 Probate Norwich 17 October to Robert James Abbott (Sharnagh) and William Abbott infirmans.
1693	ABBOTT Samuel Mitchell of Elshfold Falskote-road Binsley Common Bramley Kent died August 1908 Administration (with Will) London 21 December to Elizabeth Abbott widow. Effects £152.	ABBOTT Daphne Mary Thelma of 28a Rue de Malmaison English born (wife of Sir William Neville Noel Bernard) died 12 September 1908 Administration Leeds 29 October to George Herbert Caine commercial clerk Effects £250.
52	ABBOTT Sarah of Dauntsey Somersetshire spinster died 10 December 1908 Administration (with Will) London 23 April to Emily Gowler widow. Effects £55 3s. 10d.	ABBOTT Anne of Upsettine Stasley-road Sutton Senary widow died 21 November 1908 Probate London 15 December to Magdalen Anne Mary Miller (wife of Ernest Bos Miller) Effects £68 10s. 1d.
697	ABBOTT Sarah of 11 Springstead Road Portsmouth widow died 28 April 1908 Probate London 6 May to Eliza Taylor (wife of Alfred Taylor). Effects £157.	ABBOTT Anna Maria of Trench Norfolk widow died 12 Mar 1908 Probate Norwich 6 June to the executor Willa Frederick Kerns clerk. Effects £467 10s. 8d.
1015	ABBOTT Sarah of The Square Bedford Nottinghamshire (wife of Thomas Abbott) died 17 June 1908 Probate London 29 July to Thomas Abbott engineer. Effects £92 8s. 6d.	ABBOTT Ezra of 11 Denestable-road Luton Bedfordshire gent died 18 September 1908 Probate Northampton 14 November to Earn South-engineer and Alfred William Spink baker. Effects £7152 16s. 3d.

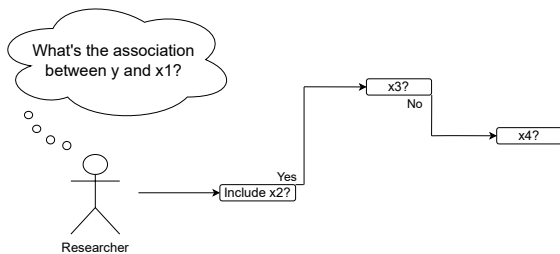
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By opening research, we can examine the certainty of statistical results.
- Does β_1 really equal 7?

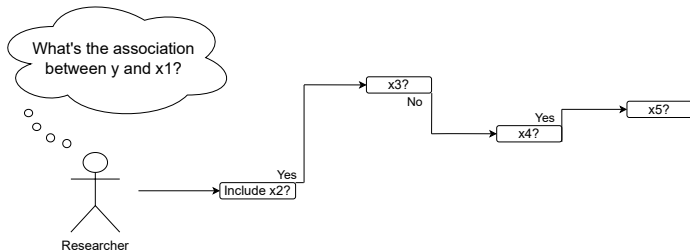
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By opening research, we can examine the certainty of statistical results.
- Does β_1 really equal 7?
- How can we "take the ‘con’ out of econometrics"?

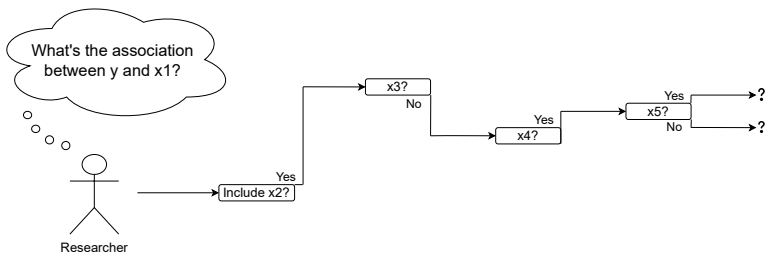
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By opening research, we can examine the certainty of statistical results.
- Does β_1 really equal 7?
- How can we "take the ‘con’ out of econometrics"?
- Transparency and more computational procedures!

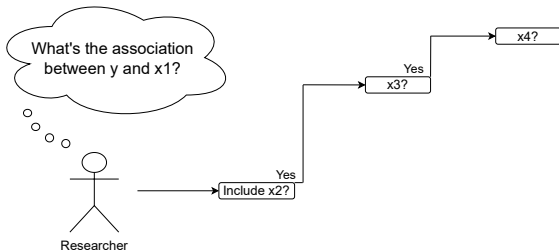


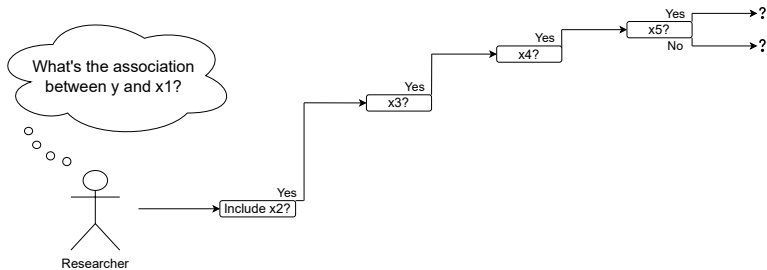


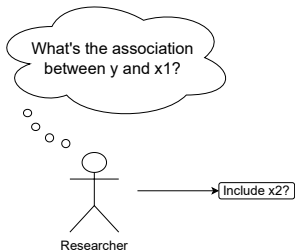


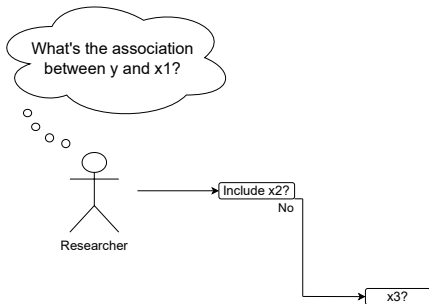


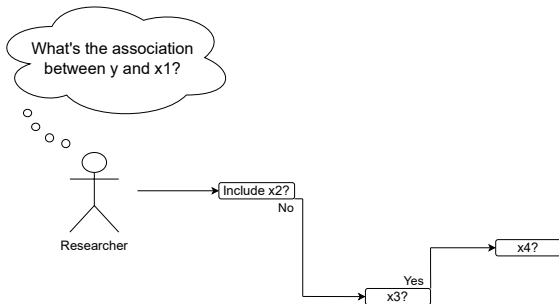


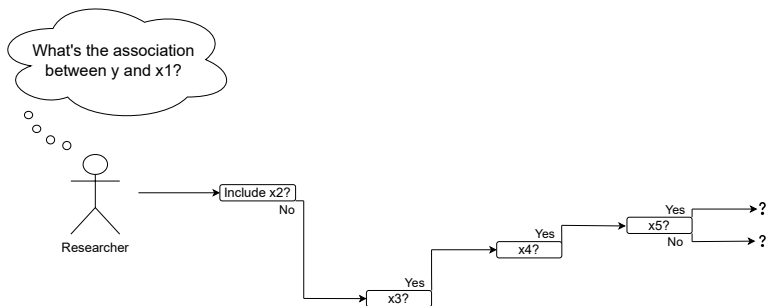


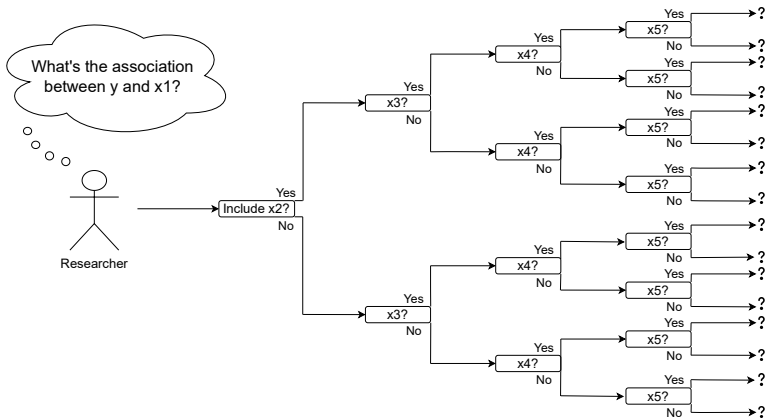




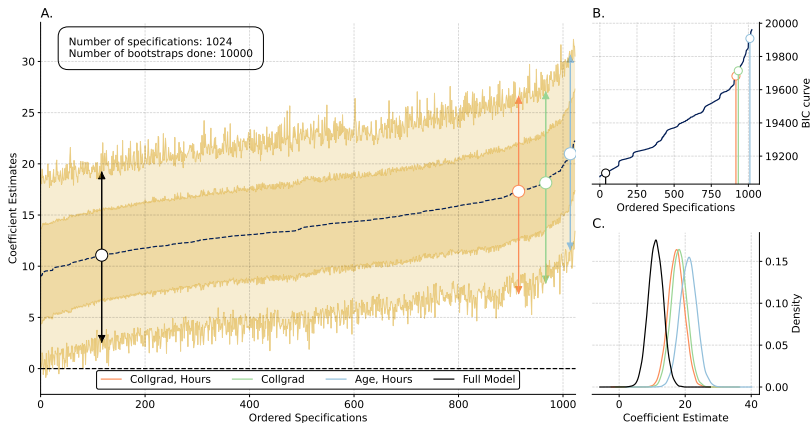




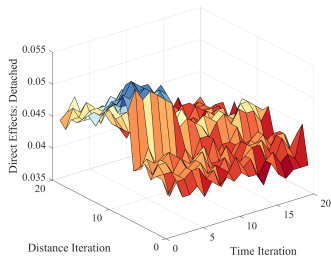
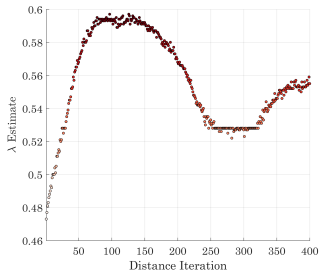
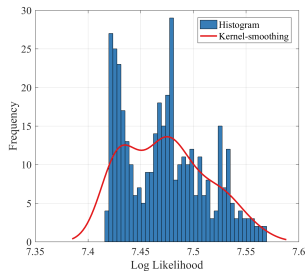
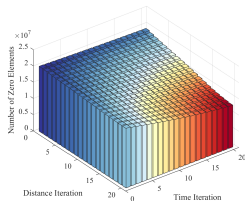




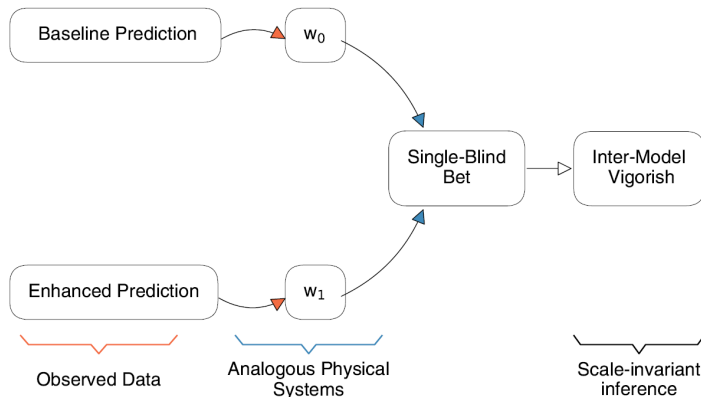
- Our work tackling this type of uncertainty is still in progress!



- The idea is to advance the **specification curve literature**.
- In particular: selection, **averaging**, and **influence analysis**.
- See github.com/centre-for-care/nrobust for updates!



‘A Grid Based Approach to Analysing Spatial Weighting Matrix Specification’



‘InterModel Vigorish (IMV): A novel approach for quantifying predictive accuracy when outcomes are binary’

With Ben Domingue, Alex Rigos, Jeremy Frees *et al.*

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.
- **Q:** Has anyone ran a program twice, with different results?

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.
- **Q:** Has anyone ran a program twice, with different results?
- **Q:** Has anyone here ever set a 'seed'?

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.
- **Q:** Has anyone ran a program twice, with different results?
- **Q:** Has anyone here ever set a 'seed'?
- **Q:** How did you decide on what seed to use?

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.
- **Q:** Has anyone ran a program twice, with different results?
- **Q:** Has anyone here ever set a 'seed'?
- **Q:** How did you decide on what seed to use?
- Maybe to eliminate variation in algorithms with PRNGs?
 - This seems to be the current 'best practice' advice.

Pseudo-Random Number Generation

- I'd like to end with 5 slides on what I'm working on **right now**.
- **Q:** Has anyone ran a program twice, with different results?
- **Q:** Has anyone here ever set a 'seed'?
- **Q:** How did you decide on what seed to use?
- Maybe to eliminate variation in algorithms with PRNGs?
 - This seems to be the current 'best practice' advice.
- We argue this is the **opposite of what we want!**
 - We propose you pre-specify **multiple** (complex) seeds.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.

Pseudo-Random Number Generation (Cont.)

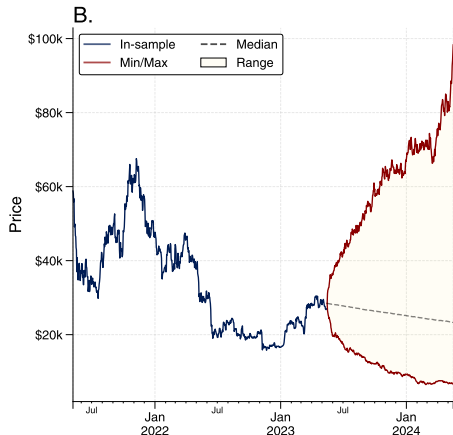
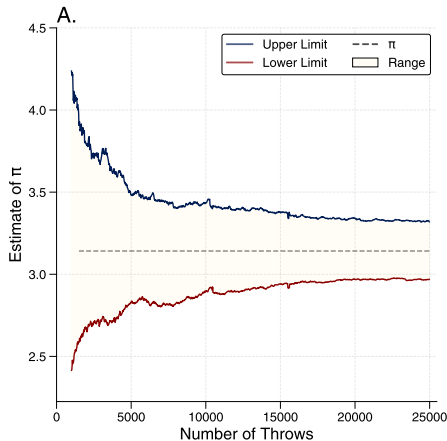
- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**
- The variation in estimand can be **huge**.

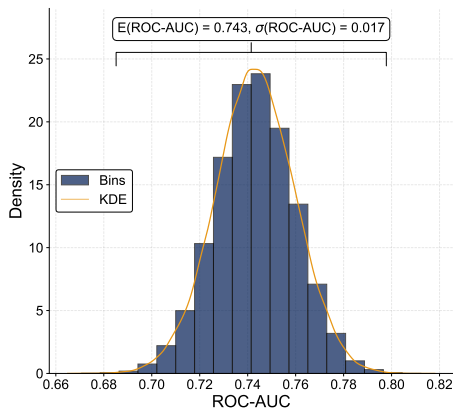
Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**
- The variation in estimand can be **huge**.
- We bring attention to this through multiple types of replications.
 - Simulations, machine learning, and inferential research.

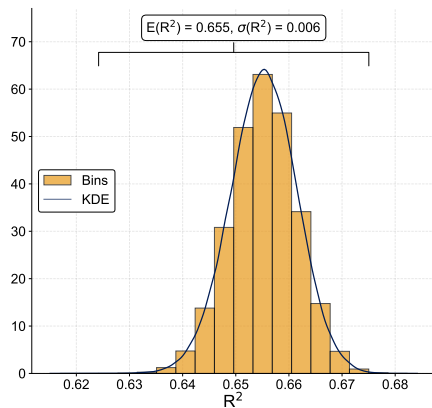


‘From a Seed of Doubt Grows a Forest of Uncertainty’
With Arun Frey, Jiani Yan and Mark Verhagen

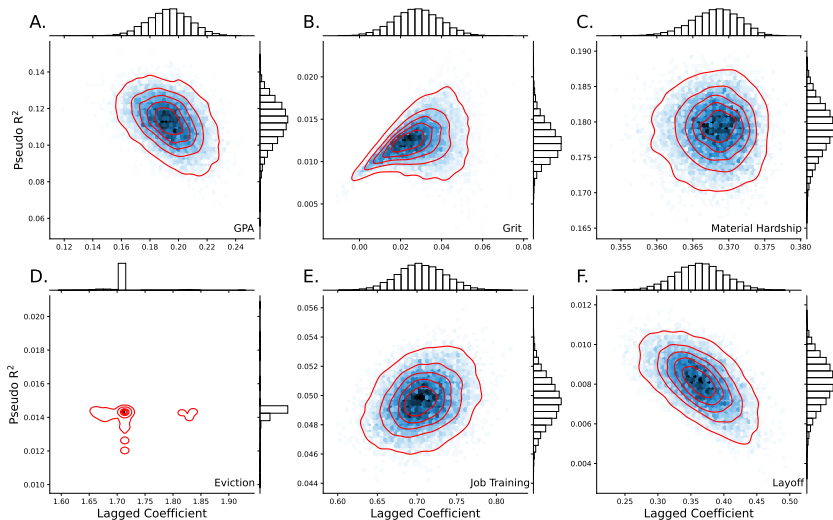
A.



B.



‘From a Seed of Doubt Grows a Forest of Uncertainty’
 With Arun Frey, Jiani Yan and Mark Verhagen



‘From a Seed of Doubt Grows a Forest of Uncertainty’
 With Arun Frey, Jiani Yan and Mark Verhagen

I hope I've been able confer my excitement, enthusiasm and inspiration for:

- 'Assumption free' social science.
- 'Semi-structured' data found 'in the wild'.
- Openness and Transparency in research.

As we argue for in a recent preprint (with Dave and Mark), the future of Open CSS is (hopefully) bright!

Thank you!

Don't hesitate to get in touch at:

charles.rahall@sociology.ox.ac.uk

and

github.com/crahal!