

Open & Reproducible Research with Semi-structured Data

Charlie Rahal

LCDS, Department of Sociology and Nuffield College

University of Oxford, 22nd November, 2022



Open & Reproducible Research (with mostly Semi-structured Data)

Charlie Rahal

LCDS, Department of Sociology and Nuffield College

University of Oxford, 22nd November, 2022



In the spirit of ‘Open Science’:

crahal.github.io/files/departamental_seminar.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)
- L^AT_EX (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 and transparent.

What motivates a talk about ‘Open Science’?

- Different to the other presentations, I thought I'd tell three anecdotes from my time as a doctoral student.

What motivates a talk about ‘Open Science’?

- Different to the other presentations, I thought I'd tell three anecdotes from my time as a doctoral student.

1. Make sure your coauthors have code.

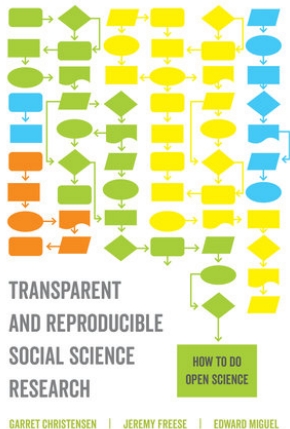
What motivates a talk about ‘Open Science’?

- Different to the other presentations, I thought I'd tell 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’?

- Different to the other presentations, I thought I'd tell 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.

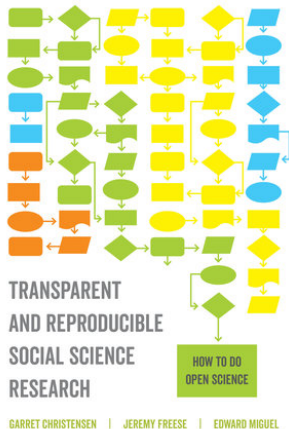
More traditional motivation, just in case you aren't sold...:



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

Michael mentioned in this series (2019?) that replications should be a quintessential part of graduate school: I totally agree!

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](#) +

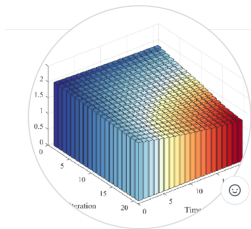
Publishing your data is easy on Harvard Dataverse!

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)
```

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.

Perhaps controversial, but most classes here are now taught in R.

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.

Perhaps controversial, but most classes here are now taught in R.



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...

What does a return from these API's typically look like?

```
{
  "creator": ["AbdouMaliq Simone"],
  "datePublished": "2014-01-01",
  "docType": "chapter",
  "id": "ark://27927/pbd16g75hqk",
  "isPartOf": "Jakarta, Drawing the City Near",
  "language": ["eng"],
  "pageCount": 60, "pageEnd": "82",
  "pageStart": "23", "pagination": "23 82",
  "provider": "portico",
  "publicationYear": 2014,
  "publisher": "University of Minnesota Press",
  "sequence": 6,
  "tdmCategory": [{
    "label": "Business - Business operations",
    "category": "Business",
    "sub_category": "Business operations"
  }],
  "title": "1. The Near-South: Between Megablock and Slum",
  "wordCount": 22167
}
```

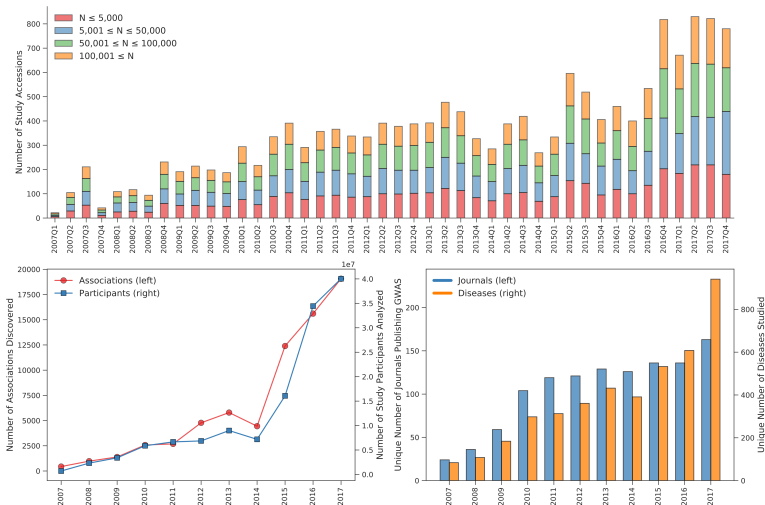

My early scientometric work analyzed GWAS.

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:

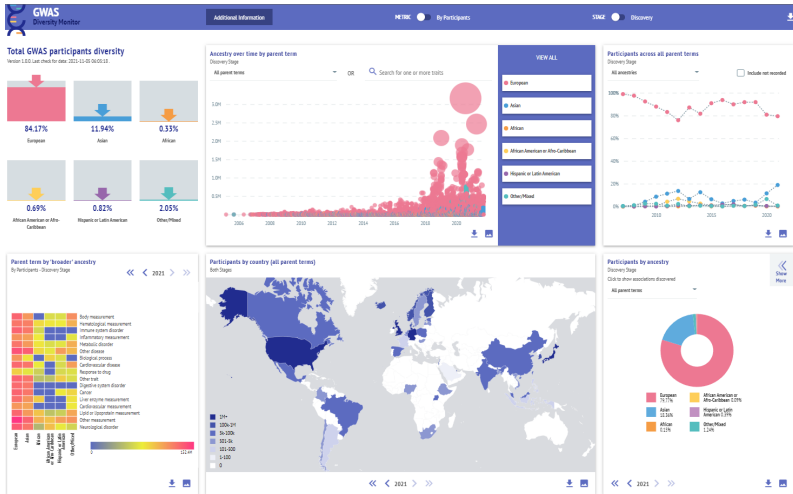
1. Conduct a thorough scientometric review.
2. Launch a substantial interactive dashboard.





'The Scientometrics of Genome Wide Association Studies'

with Melinda C. Mills

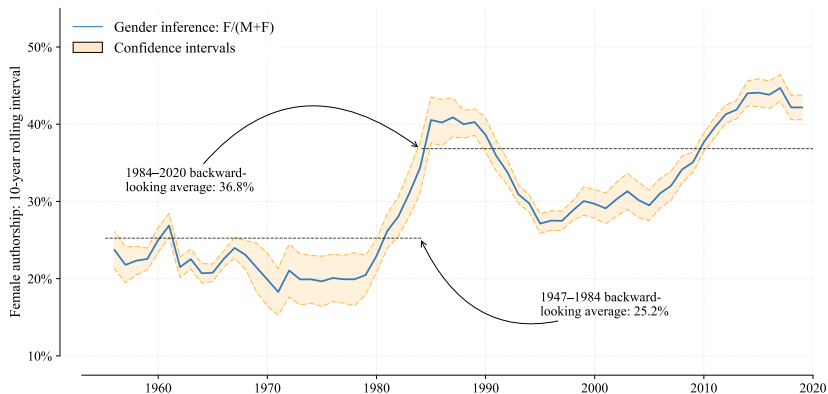


‘The GWAS Diversity Monitor Tracks Diversity in Real Time’
with Melinda C. Mills

More broadly, this approach allows us to consider;

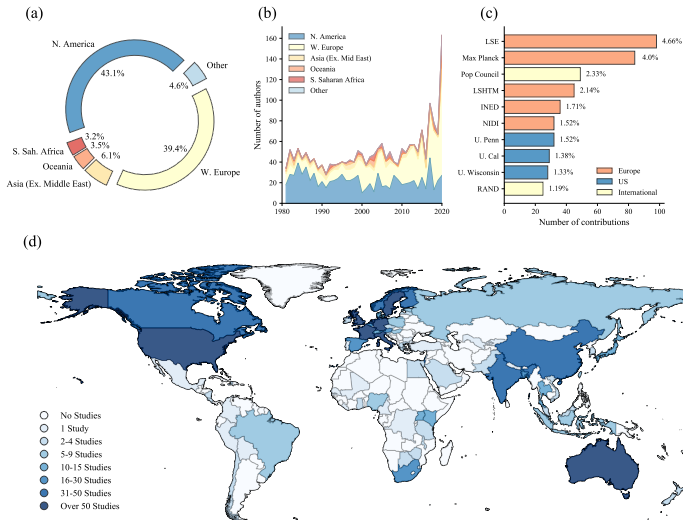
- Entire journals (e.g. ISSN)
- Entire topics (e.g. machine learning)
- Many (, many) other things still in progress.

Lets see a couple of examples of the above now!

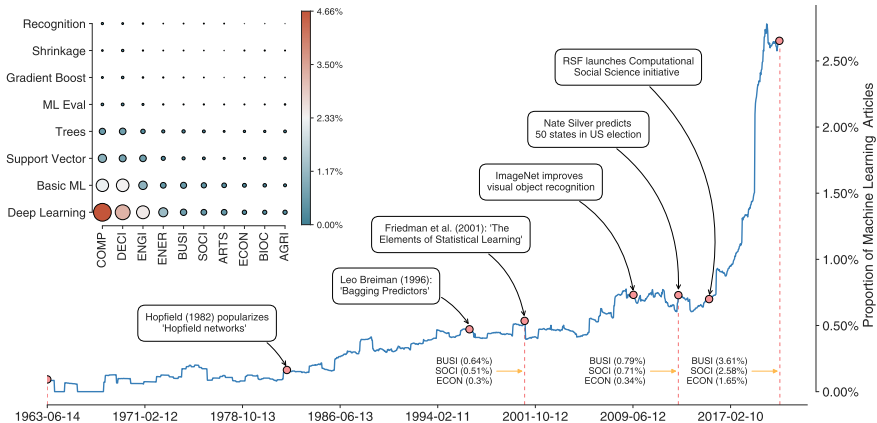


'Population Studies at 75 Years: An empirical review'

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
Open Access Home Other versions **DONATE** About

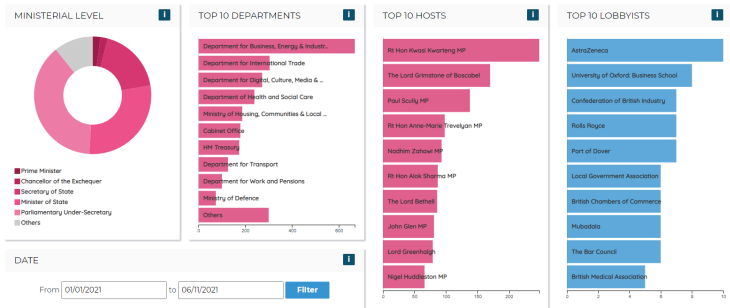
2,623 out of 74,435 meetings

SEARCH BY LOBBYIST, HOST, SUBJECT

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...

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	Agency	City	Agency CCG	Entity	Entity Sector	Program	Program Number	AP Amount	VAT registration number	Postcode	Invoice number
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	COMMUNITY SERVICES	BEXLEY COUNCIL	11599653	868 000.00		6100071865
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	COMMUNITY SERVICES	BEXLEY COUNCIL	11599653	100 000.00		6100071861
Department of health	NHS	Bexley	CCG	10042601	Other Gen Supplies & Inv	COMMUNITY SERVICES	BEXLEY COUNCIL	11599653	100 000.00		6100071869
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	WYLLER RESOURCES	BEXLEY COUNCIL	11599652	64 000.00		6100071867
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11607291	68 226.00	604483687	610008044
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING - NON ACUTE	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11607291	68 226.00		610008044
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	COMMUNITY SERVICES	BEXLEY COUNCIL	11615604	30 000.00		6100071863
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Independent Sector	ACUTE COMMISSIONING	IN-HEALTH LTD	11615704	38 528.24	787424655	HL00056682
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	NEWWAY NHS FOUNDATION TRUST	11615704	48 147.4		6100071862
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	HOSPICES	GREENWICH & BEXLEY COMMUNITY HOSPICE	11623377	218 000.00		871
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	LEWISHAM & GREENWICH NHS TRUST	11636201	4 440 140.00		6448267
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	COMMUNITY SERVICES	NHS PROPERTY SERVICES LTD	11641867	213 124.11		6100071868
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	BARROFFORD & BEXLEY COMMUNITY HOSPICE	11642864	25 672.53		48413715
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	CARTFORD AND GRAVESEND NHS TRUST	11642923	-30 000.00		200004184
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	CARTFORD AND GRAVESEND NHS TRUST	11642923	200 000.00		200004139
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11644232	-3 626 119.34		CM00218970
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	AMBUANCE SERVICES	LONDON AMBUANCE SERVICE NHS TRUST	11663375	587 757.29	604811392	10011114
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11666893	786 836.00		227 489
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11666893	760 530.34		233561
Department of health	NHS	Bexley	CCG	10042601	Cont Care Adult 100% Fully Funded	FUNDED NURSING CARE	BURA CARE HOMES	11669882	26 512.00	2387911641	ALM3036320YALM303658
Department of health	NHS	Bexley	CCG	10042601	Charges From CCG	RECHARGES	NHS SOUTH EAST CCG	11669995	85 249.00	GB054442045	731160994
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11701003	213 828.00		1200694
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	MOORFIELD EYE HOSPITAL NHS FOUNDATION TRUST	11650115	82 568.00		CC00254903
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701702	228 988.00		278586
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701702	141 917.00		37932
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11701710	807 726.00		37892
Department of health	NHS	Bexley	CCG	10042601	Income Mx < 11%	BALANCE SHEET	HMC	11720565	47 733.11	604483331	1009100565770 MAR 18
Department of health	NHS	Bexley	CCG	10042601	Income Mx < 11% Student Loans	BALANCE SHEET	HMC	11720565	738.00	604483331	1009100565770 MAR 18
Department of health	NHS	Bexley	CCG	10042601	National Insurance < 1 YN-EMS	BALANCE SHEET	HMC	11720565	29 114.84	604483331	1009100565770 MAR 18
Department of health	NHS	Bexley	CCG	10042601	National Insurance < 1 YN-EMS	BALANCE SHEET	HMC	11720565	19 284.18	604483331	1009100565770 MAR 18
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724081	213 245.00		200106
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724081	288 141.00		229584
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724081	27 100.00		200106
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11722029	266 424.00		231364
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11724114	224 793.00		200106
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	RETAIL REVENUE	RETAIL REVENUE	11724114	440.00		200106
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	QUEEN VICTORIA HOSPITAL NHS FOUNDATION TRUST	11738362	-30 056.34		1365
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Over Under Performance	ACUTE COMMISSIONING	WICKSTEAD EYE HOSPITAL NHS FOUNDATION TRUST	11738362	53 540.00		CM004848
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	COMMUNITY SERVICES	OXLEAS NHS FOUNDATION TRUST	11748440	246 505.00		17541
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	QUEEN VICTORIA HOSPITAL NHS FOUNDATION TRUST	11748368	47 300.00		201000982
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	WICKSTEAD EYE HOSPITAL NHS FOUNDATION TRUST	11771723	10 826.00		26332
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	CARTFORD AND GRAVESEND NHS TRUST	11771723	2 691 512.00		1000041308
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	MENTAL HEALTH CONTRACTS	SOUTH LONDON AND MAUDSLAY NHS FOUNDATION TRUST	11777727	25 202.00		80007387
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Independent Sector	COMMUNITY SERVICES - NON ACUTE	BEXLEY COUNCIL	11801116	46 300.00		6100071860
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	MENTAL HEALTH CONTRACTS	OXLEAS NHS FOUNDATION TRUST	11782582	1 728 324.00		37897
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	MENTAL HEALTH CONTRACTS	OXLEAS NHS FOUNDATION TRUST	11801116	36 000.00		37898
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	RETAILS AND FACILITIES	NHS PROPERTY SERVICES LTD	11801116	51 209.19		484138351
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	OXLEAS NHS FOUNDATION TRUST	11801820	258 781.00		37895
Department of health	NHS	Bexley	CCG	10042601	Hithou Foundation Trusts	CHILD AND ADOLESCENT MENTAL HEALTH	OXLEAS NHS FOUNDATION TRUST	11813430	55 506.00		37894
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	LEARNING DIFFICULTIES	BEXLEY COUNCIL	11813430	20 000.00		6000718263
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	HOSPICES	GREENWICH & BEXLEY COMMUNITY HOSPICE	11815190	276 741.00		589
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	HOSPICES	GREENWICH & BEXLEY COMMUNITY HOSPICE	11815190	276 741.00		589
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11824800	1 869 517.00	882426016	6100071867
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Other Public Sector	LEARNING DIFFICULTIES	BEXLEY COUNCIL	11844467	250 000.00		6100071847
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11844467	11 126.00		6100071862
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Independent Sector	COMMUNITY SERVICES	BELLGROVE ROAD SURGERY	11863191	65 592.39		BA-0320156
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	NHS 111	NHS BROMLEY CCG	11923323	71 147.88		1002102149
Department of health	NHS	Bexley	CCG	10042601	Clinical/Medical-Independent Sector	COMMUNITY SERVICES	CARTFORD AND GRAVESEND NHS TRUST	11933529	26 300.00		6100071868
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11933529	1 137 848.98		234487
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	GUY'S & ST THOMAS HOSPITAL NHS FOUNDATION TRUST	11931340	1 455 963.00	604923417	2106559
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	GUY'S & ST THOMAS HOSPITAL NHS FOUNDATION TRUST	11933529	8 133 540.00	604923417	9046655
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	LEWISHAM & GREENWICH NHS TRUST	11933529	-129 548.00		8466605
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951210	1 187 569.00		232736
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951210	-308 156.00		CM00218959
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951210	229 254.29		233240
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	PLANNED CARE	KING'S COLLEGE HOSPITAL NHS FOUNDATION TRUST	11951210	-411 500.00		CM00218959
Department of health	NHS	Bexley	CCG	10042601	Cont Care Adult 100% Fully Funded	CHC AD FULL FUND PERKS HLTH RD	BEXLEY COUNCIL	11961590	269 534.97		6100071866
Department of health	NHS	Bexley	CCG	10042601	Health Svc Res P8th Trust-Contract Baseline	ACUTE COMMISSIONING	UNIVERSITY COLLEGE LONDON HOSPITALS NHS FOUNDATION TRUST	11962539	68 226.00	604483687	610008044

Expenditure Over Threshold

32 / 50

Expenditure Over Threshold

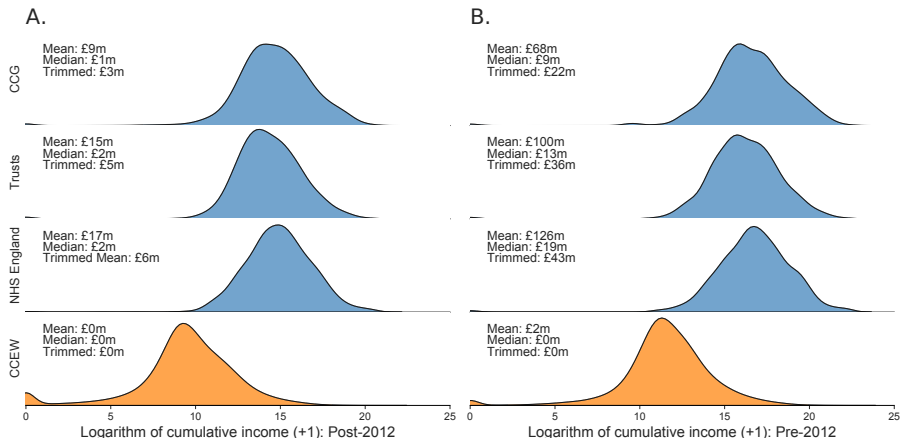
32 / 50

Expenditure Over Threshold

32 / 50

Conclusion

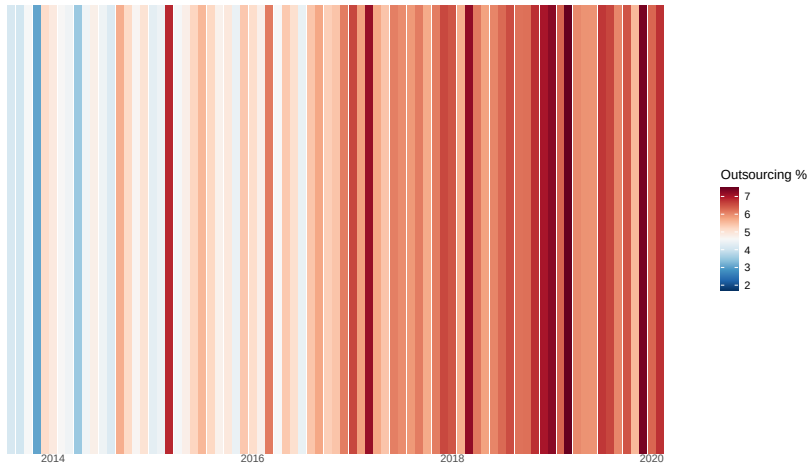




‘The Role of Non-Profits in Public Health Service Provision’

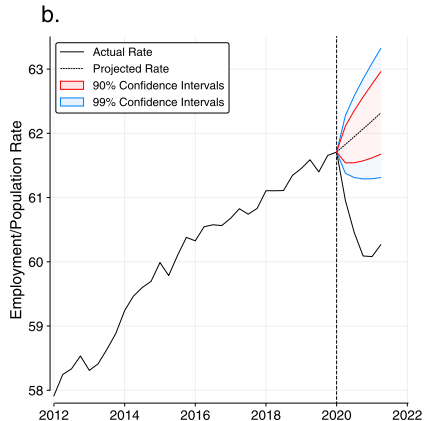
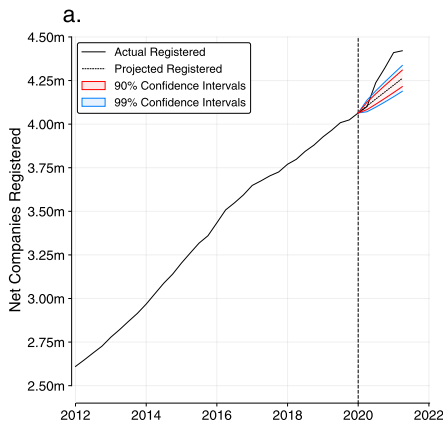
With John Mohan

Creeping Privatisation in England's NHS



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

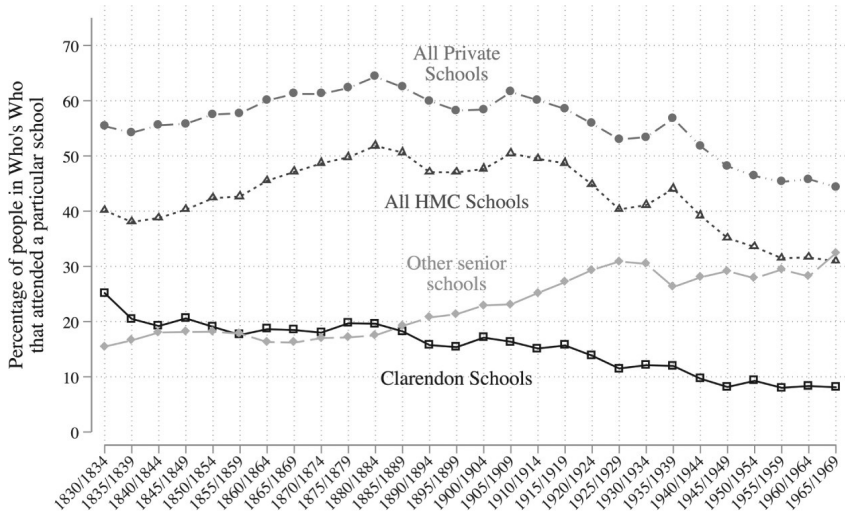
'Creeping Commercialization in the NHS'
With Ben Goodair and Aaron Reeves



‘Modelling Labour Market Inequalities and the Covid-19 Pandemic’
 With Aaron Reeves and Naomi Muggleton

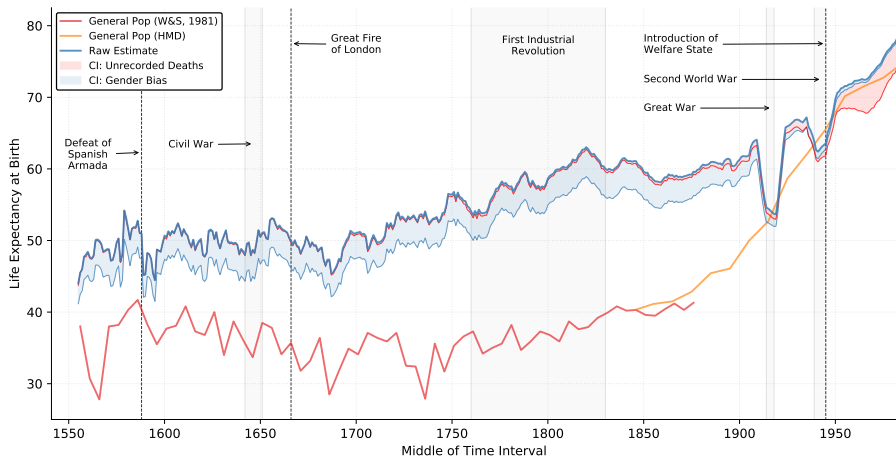
Semi-Structured Data for Inequality Research

- 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.
- When doing this, we might choose 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!**

There are close links between ‘Open Science’ and traditional survey based modelling 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 survey based modelling 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"?

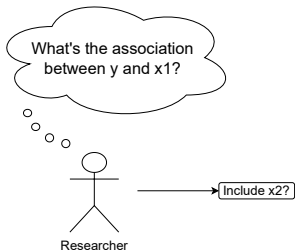
There are close links between ‘Open Science’ and traditional survey based modelling paradigms!

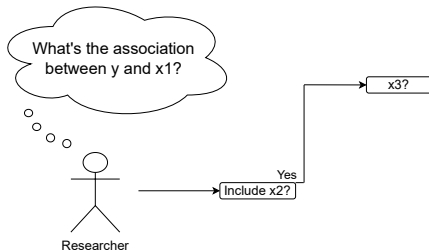
By opening research, we can examine the certainty of statistical results.

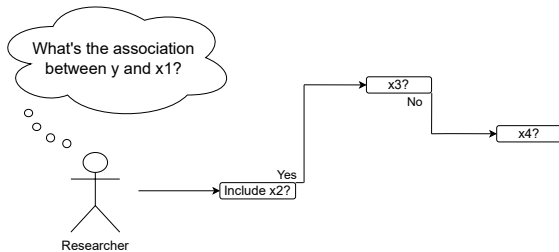
Does β_1 really equal 7?

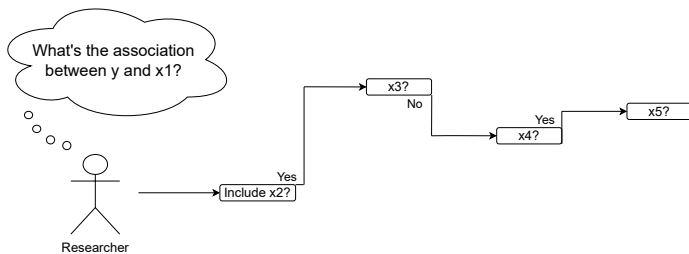
How can we "take the ‘con’ out of econometrics"?

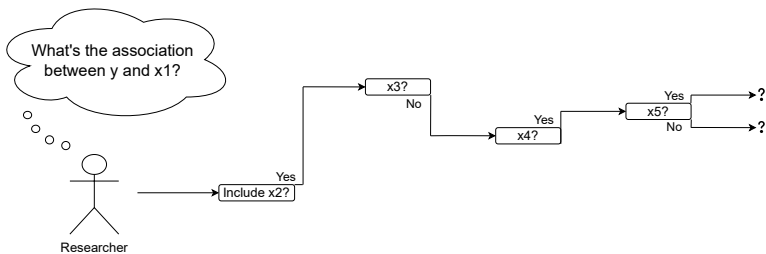
Transparency and slightly more computational procedures!

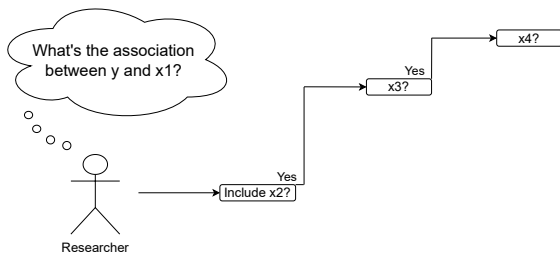


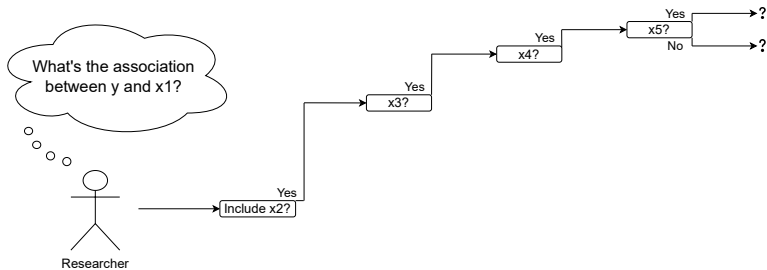


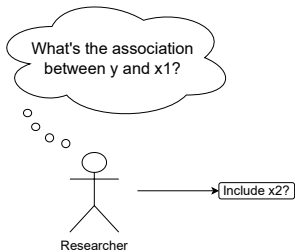


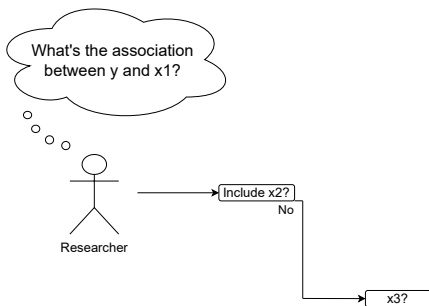


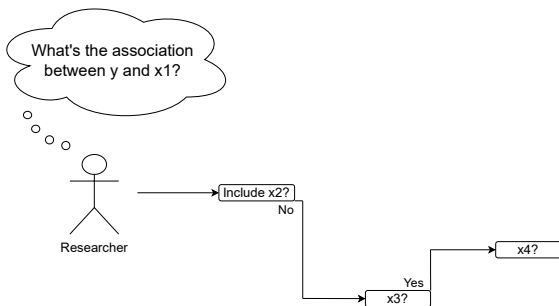


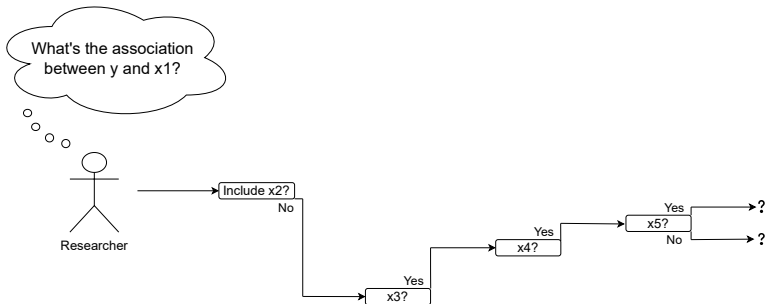


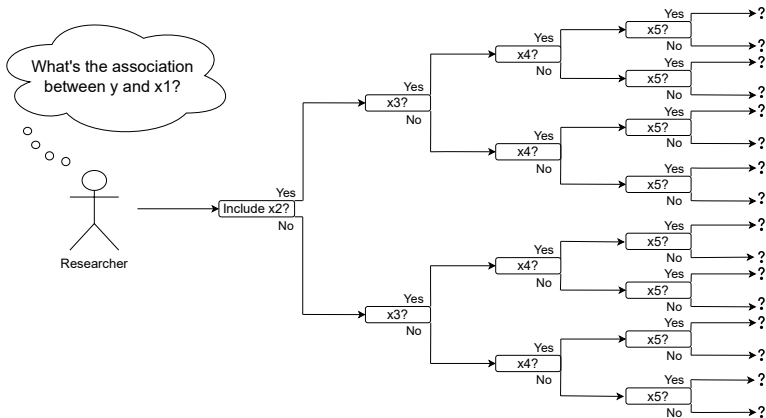












Current research culture around robustness only mandates examination of a couple of choices.

- With current computational power; why not all combinations?
- Is specification choice the only source of variability?

Earlier (unpublished) work involved model boosting, and then...

Moved towards taking one-dimensional estimates (e.g. β_1) and showing visualisations of *feasible* output space of this estimate.

This represents uncertainty around the point estimates.

Our work to tackle this specific type of uncertainty is still in progress!


crahal updated run_union_example and saves out (/w ll) on 28 Dec 2020 19

data	optimized (and correct) bootstrapped specifi...	11 months ago
src/python	updated run_union_example and saves out (/...	11 months ago
.gitignore	added .idea to the gitignore	11 months ago
LICENSE	Initial commit	11 months ago
readme.md	project setup	11 months ago

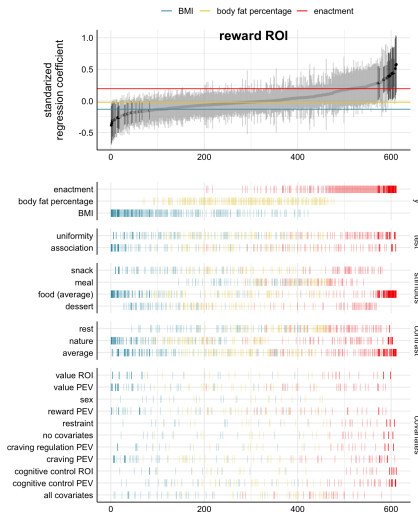
readme.md


nrobust

A set of packages for a robust, certain and stable model space.

I apologise to my coauthors for being so slow!

Our work to tackle this specific type of uncertainty is still in progress!



I apologise to my coauthors for being so slow!

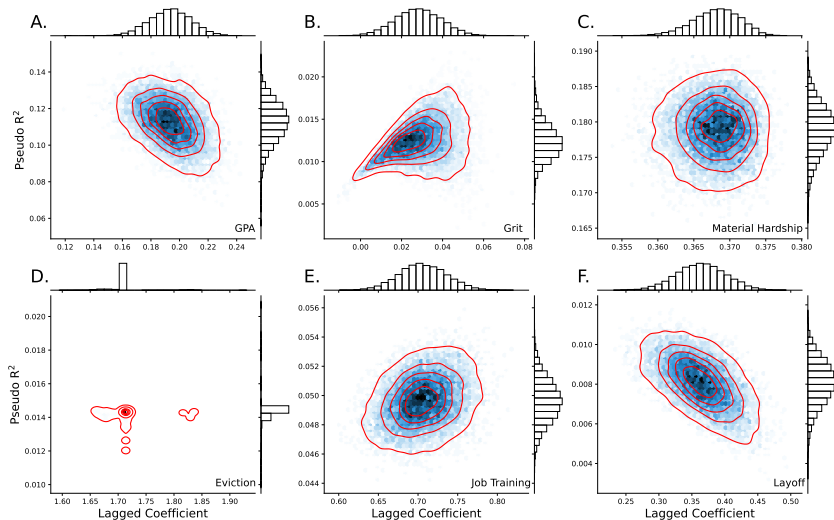
The idea is to advance the **specification curve literature**.

← Disclaimer: Figure not by me!

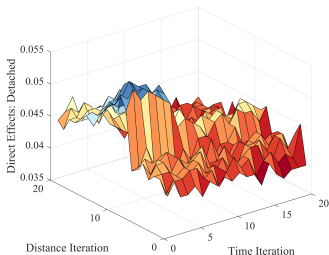
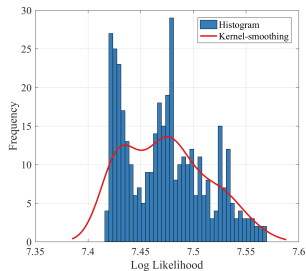
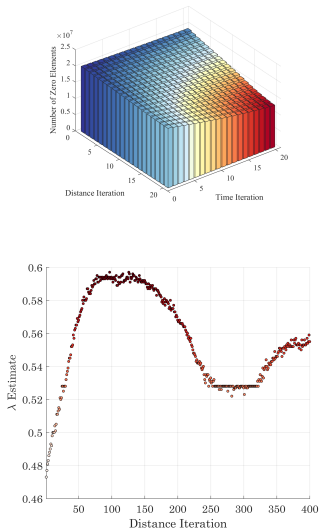
In particular: selection, averaging, influence analysis.

Be sure to watch github.com/crahal/nrobust for updates!

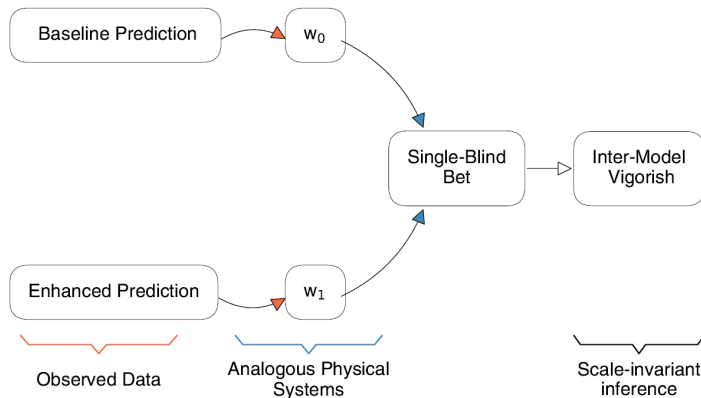
However, there's other types of uncertainty to examine...:



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



‘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.*

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/crahall!