ON THE PRIVILEGE OF TURNING OUR WORLD INTO DATA

YOUNG IRISH STATISTICIANS
ROHAN ALEXANDER, 16 DECEMBER 2021.

ROHAN ALEXANDER

UNIVERSITY OF TORONTO

- **Assistant Professor, Information and Statistical Sciences**
- **Assistant Director, CANSSI Ontario**
- **Senior fellow, Massey College**
- **Faculty affiliate, Schwartz Reisman Institute**

- ***Telling Stories With Data***
- ***Multilevel Regression and Poststratification: A Practical Guide and New Developments***

- **Co-organizer, Toronto Data Workshop**

rohanalexander.com @RohanAlexander rohan.alexander@utoronto.ca


GUINNESS
GUINNESS


RELAND

OUTLINE

1. Origins of data science.
2. What I see data science as.
3. Some of our roles and responsibilities.
4. Various applications:
    1. understanding the effect of elections;
    2. hate speech detection; and
    3. the reproducibility of COVID-19 pre-prints.
5. A few open issues.

WHAT IS DATA SCIENCE?

"...AN EXCITING DISCIPLINE THAT ALLOWS YOU TO TURN RAW DATA INTO UNDERSTANDING, INSIGHT, AND KNOWLEDGE."

"...THE SCIENCE OF EXTRACTING MEANINGFUL INFORMATION FROM DATA."

"...THE PROCESS OF FORMULATING A QUANTITATIVE QUESTION THAT CAN BE ANSWERED WITH DATA, COLLECTING AND CLEANING THE DATA, ANALYZING THE DATA, AND COMMUNICATING THE ANSWER TO THE QUESTION TO A RELEVANT AUDIENCE."

"DATA SCIENCE AS THE PROCESS OF GENERATING INSIGHT FROM DATA THROUGH REPRODUCIBLE AND AUDITABLE PROCESSES"

WHAT IS DATA SCIENCE?

"HUMANS MEASURING STUFF, TYPICALLY RELATED TO OTHER HUMANS, AND USING SOPHISTICATED AVERAGING TO EXPLAIN AND PREDICT"

WHAT WE CAN LEARN FROM A HISTORICAL DUBLIN CENSUS?

“AN ESSAY ON THE POPULATION OF DUBLIN. BEING THE RESULT OF AN ACTUAL SURVEY TAKEN IN 1798, WITH GREAT CARE AND PRECISION, AND ARRANGED IN A MANNER ENTIRELY NEW”

AN

ESSAY

ON THE

POPULATION OF DUBLIN.

BEING THE

RESULT OF AN ACTUAL SURVEY

TAKEN IN 1798, WITH GREAT CARE AND PRECISION,

AND ARRANGED IN A MANNER ENTIRELY NEW.

BY THE

REV. JAMES WHITELAW, M.R.I.A.

VICAR OF ST. CATHARINE'S.

TO WHICH IS ADDED,

The General Return of the District Committee in 1804,

WITH

A COMPARATIVE STATEMENT OF THE TWO SURVEYS,

ALSO,

SEVERAL OBSERVATIONS

ON THE

PRESENT STATE

OF THE

Poorer Parts of the City of Dublin.

DUBLIN:

PRINTED FOR THE AUTHOR,

BY GRAISBERRY AND CAMPBELL, NO. 10, BACK-LANE.

1805.

YORK-STREET.

{East end, 44, 0 Feet wide. / West end, 44, 6 Feet wide.}

Number of Houses.	Number on Door.	State of Repair.	Stories High.	UPPER AND MIDDLE CLASS.			SERVANTS OF DITTO.			LOWER CLASS.			Total Males.	Total Females.	Grand Total.	NAMES AND OCCUPATIONS OF PROPRIETORS, &c.
				Males.	Females.	Total.	Males.	Females.	Total.	Males.	Females.	Total.				
				AUNGIER-STREET.												
1	1	m	4	0	0	0	0	0	0	8	10	18	8	10	18	Elizabeth Nowlan, Haberdasher, L. H.
2	1	m	4	0	0	0	0	0	0	5	6	11	5	6	11	Lau. Birmingham, Porter-house, L. H.
3	2	n	4	1	4	5	0	1	1	0	0	0	1	5	6	George Shee.
4	3	m	2	0	0	0	0	0	0	3	4	7	3	4	7	I. Woffington, Hair-dresser.
5	4	m	2	0	0	0	0	0	0	5	7	12	5	7	12	{John Maxwell, Law Scrivener. / Thomas Irwin, Taylor, shop only.
6	5	n	4	1	0	1	1	1	2	0	0	0	2	1	3	P. Marsh.
7	5	n	4	3	0	3	0	1	1	0	0	0	3	1	4	Samuel Montgomery.
8	6	n	4	2	2	4	1	1	2	0	0	0	3	3	6	Mr. Cary.
9	7	n	4	4	3	7	1	2	3	0	0	0	5	5	10	George Lyndon.
10	8	n	3	1	0	1	0	2	2	0	0	0	1	2	3	Charles Fleetwood, Attorney.
11	9	n	4	1	2	3	1	1	2	0	0	0	2	3	5	Mrs. Robnet.
12	10	n	4	3	3	6	3	2	5	0	0	0	6	5	11	William Glascock, Attorney.
13	11	n	4	1	1	2	3	3	6	0	0	0	4	4	8	James Glascock, Attorney.
14	12	n	4	3	3	6	1	2	3	0	0	0	4	5	9	W. Bourne, Attorney.

“THERE IS THE FAMOUS STORY BY EDDINGTON ABOUT SOME PEOPLE WHO WENT FISHING IN THE SEA WITH A NET. UPON EXAMINING THE SIZE OF THE FISH THEY HAD CAUGHT, THEY DECIDED THERE WAS A MINIMUM SIZE TO THE FISH IN THE SEA! THEIR CONCLUSION AROSE FROM THE TOOL USED AND NOT FROM REALITY.”

HAMMING (1996, 177)

MATERNAL MORTALITY IS THE NUMBER OF WOMEN WHO DIE WHILE PREGNANT, OR SOON AFTER A TERMINATION, FROM A CAUSE RELATED TO THE PREGNANCY OR ITS MANAGEMENT

WHO (2019)

ON BALANCE BETWEEN DATA AND SCIENCE


oo
nn
mm
ll
kk
ii
hh
gg
ff
ee
dd
cc
bb
LITERARIA INSTRUXI
ANNO

[J]UST PRIOR TO THE TIME WHEN MY OWN CAREER IN STATISTICS WAS COMMENCING, AN IMPORTANT DEVELOPMENT TOOK PLACE IN WHICH I WAS THEREFORE ABLE TO PARTICIPATE STRAIGHT AWAY, AND WHICH HAS SINCE HAD A GREAT EFFECT ON THE DEVELOPMENT OF THE SUBJECT. I MEAN OF COURSE THE INVENTION OF THE COMPUTER.

FOSTER


“THE BEST DATA SCIENCE
ALWAYS STARTS WITH THE
SCIENCE, NOT THE DATA”
JEFF LEEK AND ROGER D. PENG 'ADVANCED DATA SCIENCE 2020'

EXAMPLES OF MY WORK

Statistics > Applications

[Submitted on 17 Nov 2021]

The Increased Effect of Elections and Changing Prime Ministers on Topics Discussed in the Australian Federal Parliament between 1901 and 2018

Rohan Alexander, Monica Alexander

Politics and discussion in parliament is likely to be influenced by the party in power and associated election cycles. However, little is known about the extent to which these events affect discussion and how this has changed over time. We systematically analyse how discussion in the Australian Federal Parliament changes in response to two types of political events: elections and changed prime ministers. We use a newly constructed dataset of what was said in the Australian Federal Parliament from 1901 through to 2018 based on extracting and cleaning available public records. We reduce the dimensionality of discussion in this dataset by using a correlated topic model to obtain a set of comparable topics over time. We then relate those topics to the Comparative Agendas Project, and then analyse the effect of these two types of events using a Bayesian hierarchical Dirichlet model. We find that: changes in prime minister tend to be associated with topic changes even when the party in power does not change; and the effect of elections has been increasing since the 1980s, regardless of whether the election results in a change of prime minister.

Comments: 50 pages, 20 figures, 6 tables
Subjects: **Applications (stat.AP)**
Cite as: arXiv:2111.09299 **[stat.AP]**
(or arXiv:2111.09299v1 **[stat.AP]** for this version)

LIVE CODE

[Submitted on 22 Jul 2021 (v1), last revised 8 Dec 2021 (this version, v2)]

Reproducibility of COVID-19 pre-prints

Annie Collins, Rohan Alexander

To examine the reproducibility of COVID-19 research, we create a dataset of pre-prints posted to arXiv, bioRxiv, medRxiv, and SocArXiv between 28 January 2020 and 30 June 2021 that are related to COVID-19. We extract the text from these pre-prints and parse them looking for keyword markers signalling the availability of the data and code underpinning the pre-print. For the pre-prints that are in our sample, we are unable to find markers of either open data or open code for 75 per cent of those on arXiv, 67 per cent of those on bioRxiv, 79 per cent of those on medRxiv, and 85 per cent of those on SocArXiv. We conclude that there may be value in having authors categorize the degree of openness of their pre-print as part of the pre-print submissions process, and more broadly, there is a need to better integrate open science training into a wide range of fields.

Comments: 14 pages, 6 tables, 4 figures 2021-12-08 replacement fixes a few incorrect references and adds reference to some additional papers

OPEN QUESTIONS

SOME OPEN AREAS

Demonstrate:

1. **How do we write unit tests for data science?**
2. **What happened to the revolution?**
3. **How do we think about power?**

THANK YOU

rohanalexander.com

@RohanAlexander

rohan.alexander@utoronto.ca