Datasets from unstructured text in document collections (corpora)

CONTEXT RULE ASSISTED MACHINE LEARNING (CRAML)

Peter Norlander, Ph.D.,
Quinlan School of Business,
Loyola University Chicago
August 4, 2023

Software is at
github.com/sjmeis/CRAML_beta/

This research has been generously supported by the Russell Sage Foundation, and the Economic Security Project's Antimonopoly Fund. Outstanding software development and programming by Stephen Meisenbacher.

FOR INDUCTIVE OR DEDUCTIVE RESEARCH

THE PROBLEM

RESEARCH QUESTIONS

How many franchise documents have no poach and noncompete clauses?

How many job ads advertise:

- hiring workers on visas / restrictions on workers on visas?
- work from home? Independent contractors? Union jobs?

THE PROBLEM – MISSING DATA

Information buried in unstructured text

Lack of public information

Single-source bias from surveys

– one respondent does not represent firm behavior

Proprietary Data Brokers

- Trade secret methods for producing data

- Inconsistent with some core values (equity, transparency, replicability)

Sworn

Single-Source

Sold

Secret

More problems - Replication, black box, inter-operability, access, resources

IS AI THE ANSWER?

ML and AI tools necessary and helpful

Which tools?

Whose data?

How do we know?

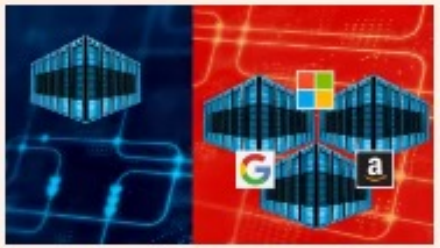
How to fix errors?

Opinion **Media**

Hollywood's fight is your fight

There will be more battles over how to divide the intellectual property pie in all industries

RANA FOROOHAR + Add to myFT



Inside Business.

Risk of 'industrial capture' looms over AI revolution

Madhumita Murgia

FOR INDUCTIVE OR DEDUCTIVE RESEARCH

A SOLUTION

Context Rule Assisted Machine Learning

A VERSATILE SUITE OF TOOLS FOR ANY CONCEPT OR CORPUS

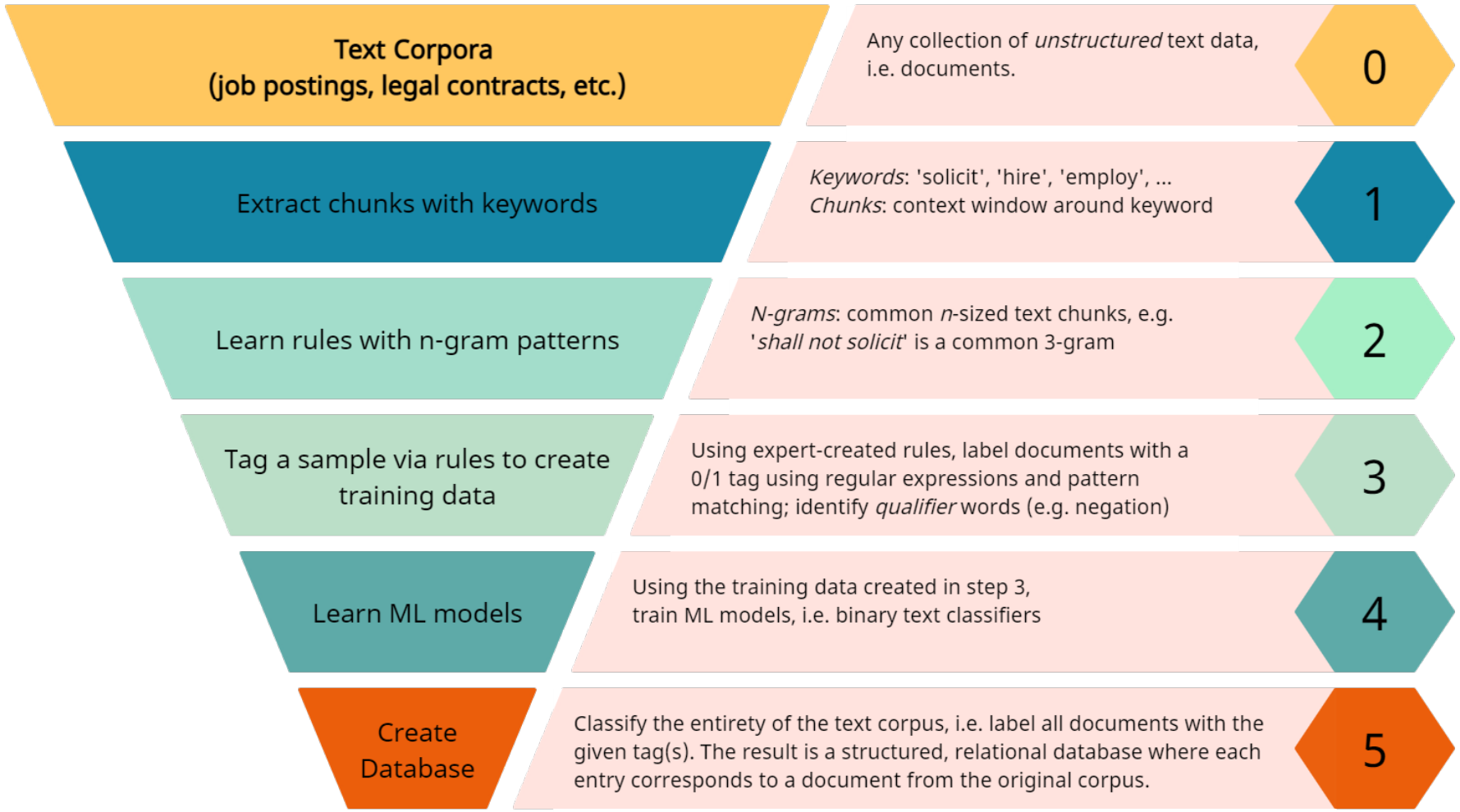
- For qualitative researchers - develop and scale original constructs over massive text corpora
- For quantitative researchers – output in rows and binary feature columns
- Ordinary desktop or laptop computer equipment required
  - Sample your corpus, set context window, and focus only on the text you care about
- Little or no code solution

FLEXIBILITY AND AGNOSTICISM

"Any concept, any corpus"

- Build Dictionaries –
  - run a dictionary on the corpus
- Build Rules - aids rule discovery, validation, and implementation
  - Sample the corpus
  - Validate the sample
  - Extrapolate rules with exact string and REGEX
- ML / AI
  - Supports similarity search with concept embeddings
  - Supports construction of niche ML models by building training data from rules
- Databases and metadata
  - Works with SQL or spreadsheet
- Input formats
  - Txt, csv, json

The CRAML Pipeline, Simplified

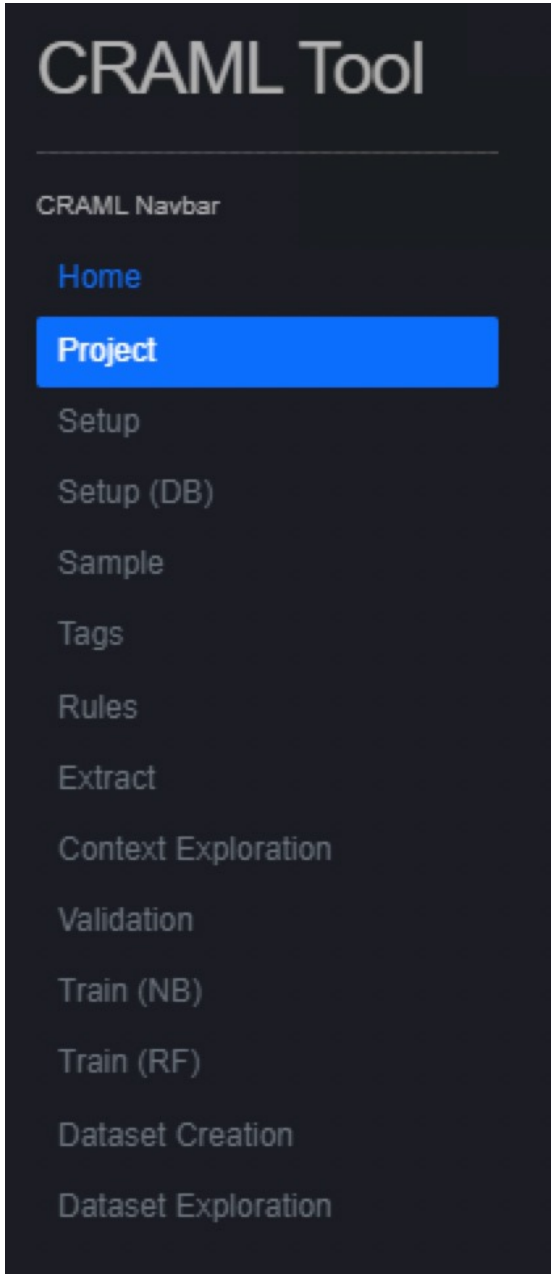


HTTPS://GITHUB.COM/SJMEIS/CRAML_BETA/

GUI INTERFACE – OPENS IN WEB BROWSER

Requires some Python knowledge

Setup your data and get started



FOR INDUCTIVE OR DEDUCTIVE RESEARCH

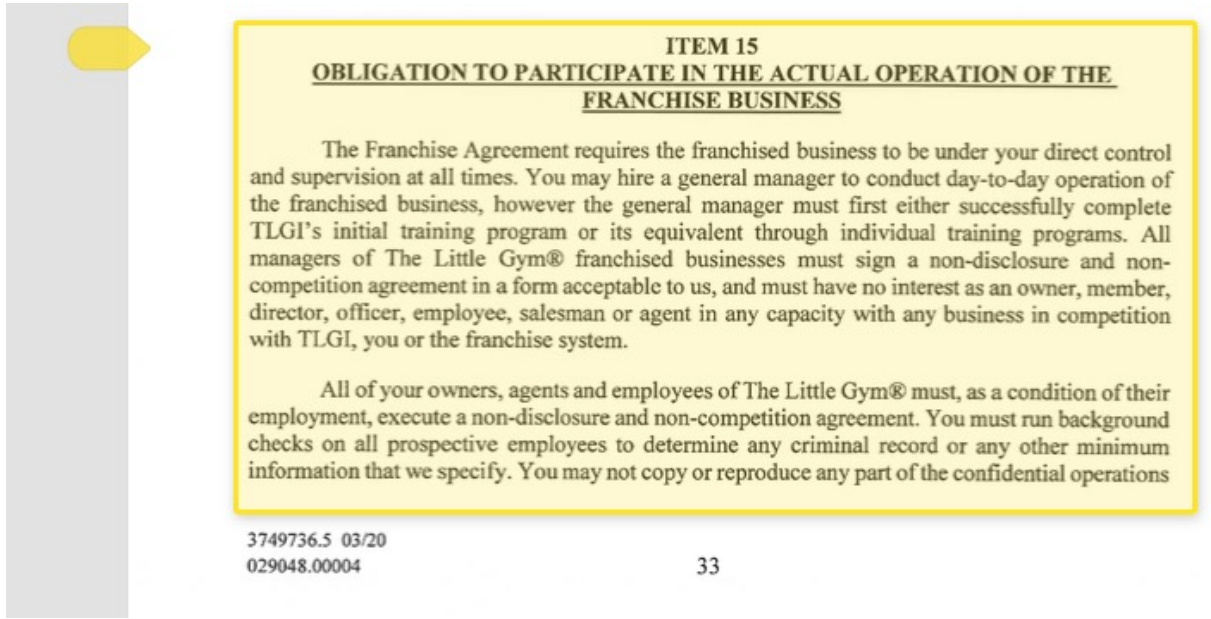
USE CASES

Context Rule Assisted Machine Learning

AS A DISCOVERY TOOL WITH FULL PIPELINE AND ML: ANTICOMPETITIVE CLAUSES IN FRANCHISE DOCUMENTS

Transparent
Open-Source
Replicable
Machine Learning
Expert-curated training data

Toward rows and columns!



ITEM 15
OBLIGATION TO PARTICIPATE IN THE ACTUAL OPERATION OF THE FRANCHISE BUSINESS

The Franchise Agreement requires the franchised business to be under your direct control and supervision at all times. You may hire a general manager to conduct day-to-day operation of the franchised business, however the general manager must first either successfully complete TLGI's initial training program or its equivalent through individual training programs. All managers of The Little Gym® franchised businesses must sign a non-disclosure and non-competition agreement in a form acceptable to us, and must have no interest as an owner, member, director, officer, employee, salesman or agent in any capacity with any business in competition with TLGI, you or the franchise system.

All of your owners, agents and employees of The Little Gym® must, as a condition of their employment, execute a non-disclosure and non-competition agreement. You must run background checks on all prospective employees to determine any criminal record or any other minimum information that we specify. You may not copy or reproduce any part of the confidential operations

3749736.5 03/20
029048.00004

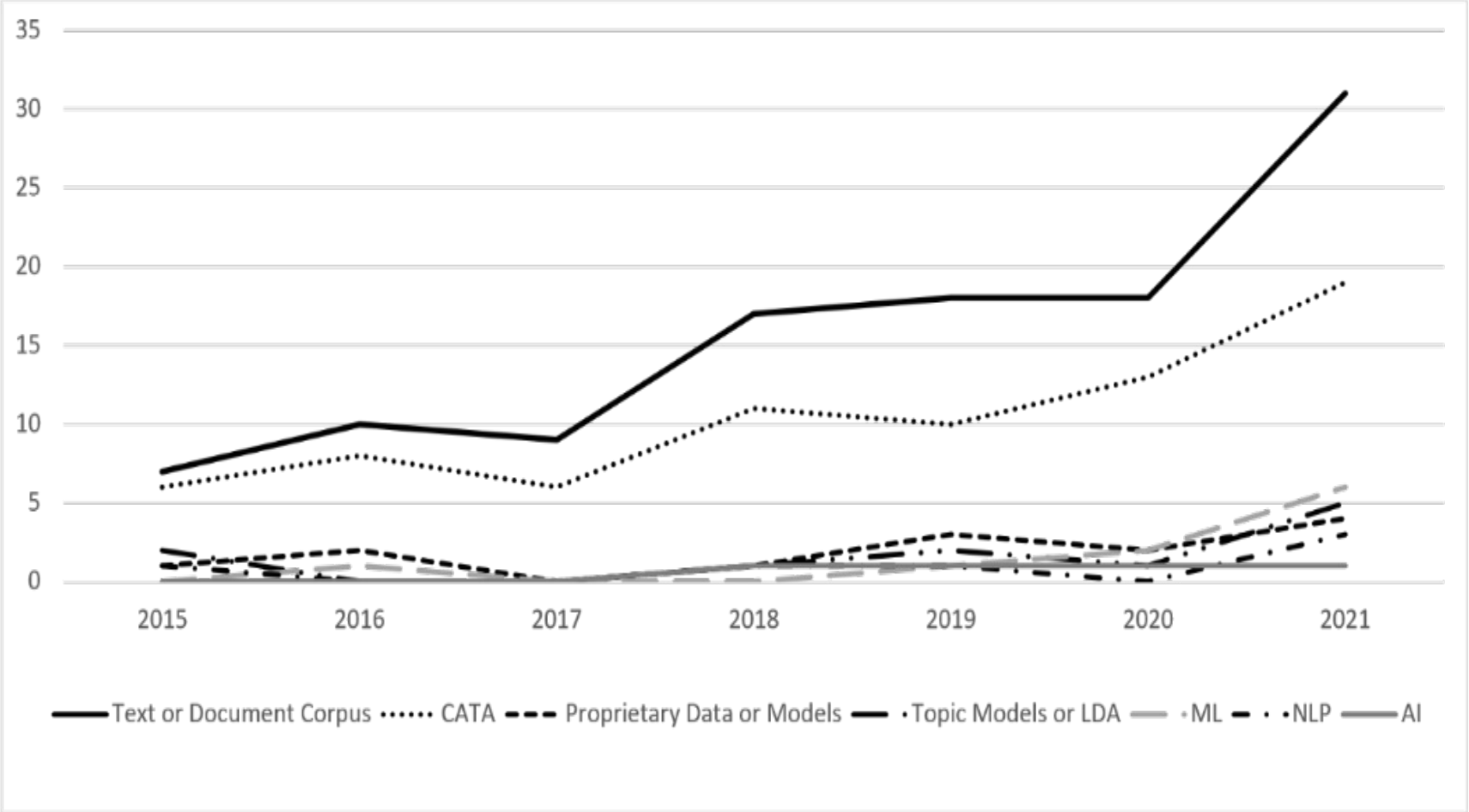
33

Working paper is up - franchise no poach, non-compete and other clauses.
Documents are up on DocumentCloud
Replication materials are up.

Norlander, P. "New Evidence on Employee Noncompete, No Poach, and No Hire Agreements in the Franchise Sector." Working Paper. Washington Center for Equitable Growth.

AS A DICTIONARY

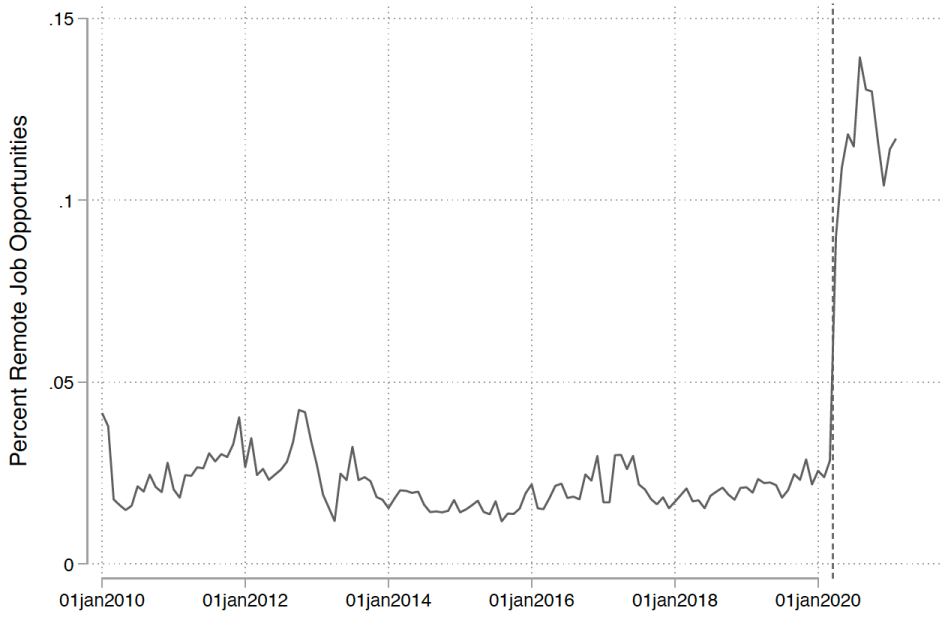
- How many papers in management literature refer to text or document corpora?
- Of those which specific text analysis methods are used?



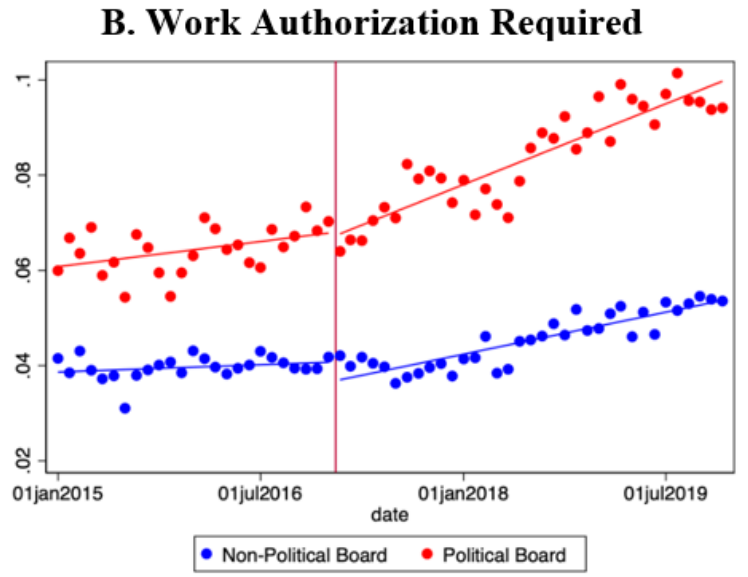
Meisenbacher, Stephen, and Peter Norlander. "Transforming Unstructured Text into Data with Context Rule Assisted Machine Learning (CRAML)." arXiv, January 20, 2023. https://doi.org/10.48550/arXiv.2301.08549.

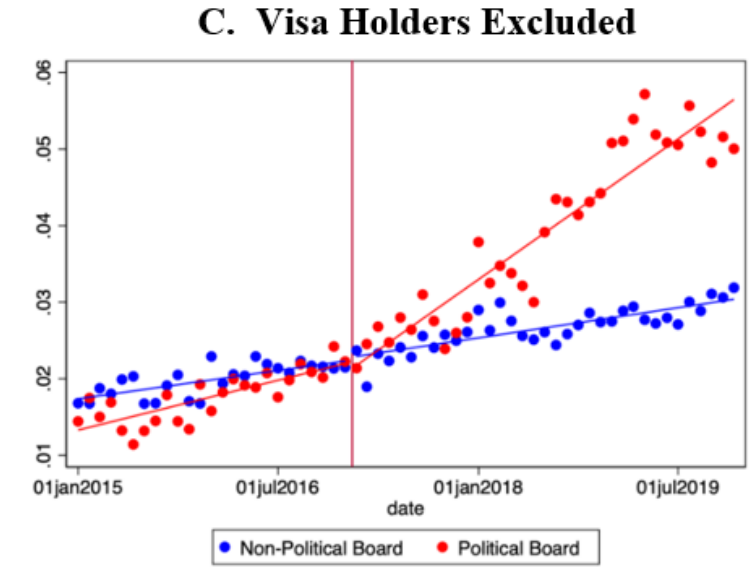
AS AN INPUT TO HYPOTHESIS DRIVEN DEDUCTIVE RESEARCH

FIGURE 1 - TIME SERIES OF JOBS ADVERTISING REMOTE JOB OPPORTUNITIES



Norlander, P and Erickson, C. "The Role of Institutions in Job Teleworkability Before and After the Covid-19 Pandemic."





Sauerwald, S. and Norlander, P. "When Trump Said Jump: Political Directors and the Recruitment of Foreign Workers." Under review

QUESTIONS?