A Research Agenda for Big Data and Securities Regulation

Presentation for “AI, Securities Regulation, and Using Big Data to Explore Securities Regulation Issues” AALS 2025

James F. Tierney

Chicago-Kent College of Law

January 9, 2025

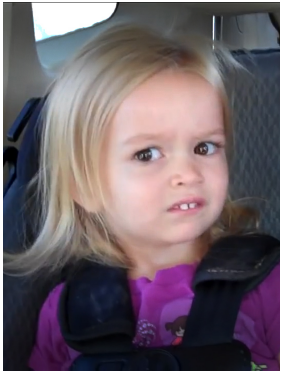
Introduction and Context

- There's too much content (data) in the world relevant to our space; how do we get a handle on it?
- Some of us also have "close reading" problems (e.g., ADHD, lack of time to read ourselves, lack of funding for RAs to read for us)
- One answer is "distant reading," or NLP and text-as-data methods (a.k.a. "forest for the trees" or "birds eye view")
- Compute is cheaper, tech is simpler: Opportunities from increased data availability and computational advancements open up new roles for regulatory and scholarly debates

My pitch: securities regulation scholars should lean into building new scholarship off of big data and legal analytics

This panel? Really??

- Never in my dreams did I think there'd be a sec reg panel on this topic or that I'd be here
- Former tween computer nerd, no PhD
- Wife banned me from making sourdough during early pandemic so, knowing I was going into academia shortly, I "learned to code" as a Covid hobby



Big data methods

One view of the cathedral:

- **Text** (e.g., disclosure) is the bread and butter of our field
- Disclosure policy shapes research questions in public markets.
- Required disclosures (e.g., financial statements, MD&A reports) as well as voluntary disclosures (e.g., investor call transcripts) are rich data sources.
- Can we derive insight from these textual sources—and do so cheaply and easily through computational methods?

Analyzing Securities-Related Corpora

Securities Data Sources

- Textual corpora include 10-K disclosures, investor call transcripts, and analyst reports.
- Combination with econometric data enriches corporate governance studies.
- Adjacent fields leverage these datasets for insights into EDGAR filings (e.g., Bodnaruk, Loughran, and McDonald 2015), transcripts of investor calls (Cai and Yung 2022), and analyst reports (Li et al. 2024)

Court Decisions and ”Distant Reading”

Legal Text Analysis

- Court decisions provide historical insights into securities law evolution.
- “Close reading” is what law professors are used to; e.g., Pritchard & Thompson 2023, gleaning meaning of Supreme Court sec reg cases
- ”Distant reading” enables pattern discovery in judicial reasoning (Jockers and Thaklen 2020)
- Example: Caselaw Access Project spans centuries of federal and state court decisions. What if we wanted to know how courts are applying a federal securities law over time?

Regulatory Texts and Trends

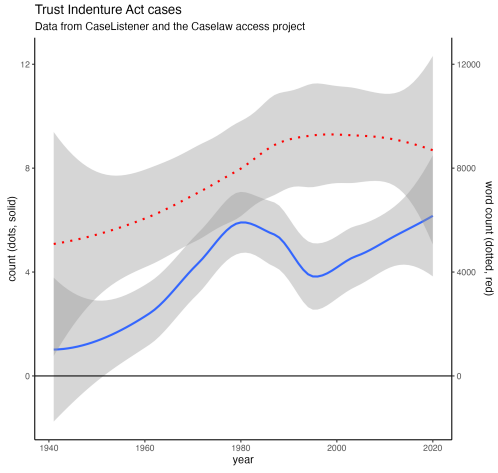
Tracking Regulatory Evolution

- Federal Register filings reveal shifts in SEC focus and methods.
- Analysis includes responses to crises, tech advancements, and political shifts.
- Example studies: Choi (2021) on statutory interpretation; Tierney (2025) on SRO rulemaking.

Applications of NLP in Legal Analysis

NLP in Securities Regulation

- Methods can run the range from inductive and descriptive (e.g., counts, dictionary matches) to deductive and testing (e.g., topic modeling, word embeddings)
- Suppose we want to know how federal courts are applying the Trust Indenture Act (oooooh, exciting)



283 cases involving the TIA from the CAP, took about an hour

Querying Large Data Sources

Automating Data Analysis

- Programmatic queries reduce manual effort in data extraction.
- Examples: FINRA BrokerCheck API for broker disclosures (e.g., Honigsberg and Jacob 2021; Alexander and Iannarone 2021b; Tierney 2024 (see next slides))
- Insights into litigation strategies and case outcomes from docket reports.

BrokerCheck by FINRA

INDIVIDUAL | FIRM

By clicking the **SEARCH** button or otherwise using BrokerCheck, I agree to **BrokerCheck Terms of Use**

john | at | Firm Name or CRD/SEC# (optional) | in | City, State or ZIP (optional) | SEARCH

Refine Results

We found **75,383** results

1 of 750 pages

List View | Sort By Relevance

KEVIN KURIAN JOHN
(KEVIN **JOHN**...)
CRD#: 6168869

Currently Not Registered

PR Previously Registered Broker

Disclosures	No
Years of Experience	5

MORE DETAILS >

Nixon Michael John
(NIXON **JOHN**...)
CRD#: 6300011

CITIGROUP GLOBAL MARKETS INC. | CRD#: 7059
NEW YORK, NY 10013 +1

B Broker *Regulated by* ***FINRA***

Disclosures	No
Years of Experience	7

MORE DETAILS >

AJITH K JOHN
(**JOHN**...)
CRD#: 4783282

Currently Not Registered

PR Previously Registered Broker

Disclosures	Yes
Years of Experience	9

MORE DETAILS >

ROBERT HENRY JOHN
(R J **JOHN**)
CRD#: 2378311

WELLS FARGO ADVISORS FINANCIAL NETWORK, LLC | CRD#: 11025
HAZLETON, PA 18201 +1

B Broker *Regulated by* ***FINRA***

IA Investment Adviser

Disclosures	No
Years of Experience	31

MORE DETAILS >

```r
session <- bow(url)

bars_list <- session %>%
  scrape()

table_data <- bars_list %>%
  html_nodes("table")

list_of_bars <- table_data %>%
  html_table() %>%
  .[[1]] %>%
  filter(!is.na(CRD)) %>%
  mutate(links = table_data %>%
           html_nodes('tr') %>%
           html_nodes('a') %>%
           html_attr("href"))

scrapable_list_of_bars <- list_of_bars %>%
  mutate(crd = case_when(str_detect(links, "brokercheck.finra.org") ~ TRUE,
                         TRUE ~ FALSE))

scrapable_list_of_bars <- scrapable_list_of_bars %>%
  group_by(CRD) %>%
  unique() %>%
  tibble::rowid_to_column("index")

saveRDS(scrapable_list_of_bars, "list_of_bars.RDS")
```

```r
vector_of_CRDs_to_scrape <- scrapable_list_of_bars %>%
  filter(crd == TRUE) %>%
  .$CRD

user_agent = 'James Tierney law professor jtierney1@kentlaw.iit.edu; polite R package bot'

session <- "https://api.brokercheck.finra.org/" %>%
  bow()

params3 = list(
  `hl` = 'true',
  `includePrevious` = 'true',
  `nrows` = '12',
  `query` = 'john',
  `r` = '25',
  `sort` = I('bc_lastname_sort+asc,bc_firstname_sort+asc,bc_middlename_sort+asc,score+desc'),
  `wt` = 'json'
)

bc_working_tibble <- NA

# the following is the scraper function

extract_data <- function(url){
  Sys.sleep(5)

  url_scraped <- nod(session, paste0("https://api.brokercheck.finra.org/search/individual/", url)) %>%
    scrape(query = params3)

  if (url_scraped$hits$total > 0) {
    json <- gsub("^angular\\.callbacks._1\\((.*)\\);?$", "\\1", url_scraped$hits$hits[[1]]$`_source`) %>%
      fromJSON()
```

```r
  bc_working_tibble <- tibble(
    individualId = json$basicInformation$individualId,
    firstName = json$basicInformation$firstName,
    middleName = json$basicInformation$middleName,
    lastName = json$basicInformation$lastName,
    otherNames = list(json$basicInformation$otherNames),
    sanctions = list(json$basicInformation$sanctions),
    bcScope = json$basicInformation$bcScope,
    iaScope = json$basicInformation$iaScope,
    daysInIndustry = json$basicInformation$daysInIndustry,
    currentEmployments = list(json$currentEmployments),
    currentIAEmployments = list(json$currentIAEmployments),
    previousEmployments = list(json$previousEmployments),
    previousIAEmployments = list(json$previousIAEmployments),
    disclosureFlag = list(json$disclosureFlag),
    iaDisclosureFlag = list(json$iaDisclosureFlag),
    disclosures = list(json$disclosures),
    examsCount = list(json$examsCount),
    stateExamCategory = list(json$stateExamCategory),
    principalExamCategory = list(json$principalExamCategory),
    productExamCategory = list(json$productExamCategory),
    registrationCount = list(json$registrationCount),
    registeredStates = list(json$registeredStates),
    registeredSROs = list(json$registeredSROs),
    brokerDetails = list(json$brokerDetails),
    match = TRUE
  )

} else {

  bc_working_tibble <- tibble(
    match = FALSE
  )
```

Understanding Large Data Sources

Method challenges

- Most data is **unstructured**, so needs to be made tidy
- Textual analysis of data is **high-dimensional**, making compute costly and the math hard, so dimension-reduction ("make the data simpler, stupid") is an important first-order methodological goal

Neither this token sequence nor the word "blurst" appear in almost any other English language text, so how do we represent it mathematically?

NLP

Pre-processing

- Removing HTML, XML, and other extraneous characters (like the line break \n in plain-text)
- Removing stopwords (the, be, to, of, and, *etc.*)
- Tokenizing — splitting sentences into words
- Stemming/lemmatizing (should we count “sell,” “sold,” and “selling” as the same?)

Note: collecting and cleaning data can easily take as much (if not more) time than writing the paper itself

NLP: Bag of Words Models

Simple Text Representation

- Treats text as a collection of words without considering order.
- Rank by term frequency / inverse document frequency (TF-IDF) so we weight the most importantly unique words for the document
- Benefits: Easy to implement and analyze.
- Limitation: Loses context and syntax.
- Application:
  - → Create a **custom dictionary** of words to find in your corpus for frequency analysis of disclosures or filings
  - → How similar are sets of documents?

Representing the corpus

Suppose we want a corpus C made of m docs (d_1 to d_m) the whole of which include n unique words (w_1 to w_n):

$$C = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ \dots \\ d_m \end{pmatrix} = \begin{pmatrix} [w_1, w_2, w_3, w_4, \dots\ w_n] d_1 \\ [w_1, w_2, w_3, w_4, \dots\ w_n] d_2 \\ [w_1, w_2, w_3, w_4, \dots\ w_n] d_3 \\ [w_1, w_2, w_3, w_4, \dots\ w_n] d_4 \\ \dots \\ [w_1, w_2, w_3, w_4, \dots\ w_n] d_m \end{pmatrix}$$

Document term matrix

	w_1	w_2	w_3	w_4	…	w_n
d_1	4			1		
d_2		1	8	2		3
d_3		5		9		1
d_4	3		7			4
…						
d_m	1			4		2

Reducing the high dimensionality of this sparse data, to something with a few dimensions, might let us do analysis without losing the most important information...

From which we can begin to develop measures of textual similarity (cosine, Euclidean, etc)

NLP: Topic Modeling

Discovering Themes in Text

- Identifies latent topics within large textual datasets.
- Methods: latent Dirichlet allocation (LDA), structural topic model, other kinds of fancy math.
- Application: Exploring trends in regulatory focus over time.

Applications

- Private equity and venture capital: overcoming transparency challenges by mining private fund filings, secondary market transactions, investor communications, etc.
- Retail investors: insights from social media, trading platform data, and portfolios.
- Administrative data: The gold standard is Scandanavian administrative data about every person's microtransaction. To get that here, you need to be friends with the BD's general counsel.
  - → Seriously, administrative data of any quality in the US is hard to come by, making this a matter of relationships as anything else

Key Challenges

- “Learn to code” is not just a playground taunt
- Data-driven policymaking and scholarship aren’t going away, so let’s lean in
  - → OTOH, do we want to be economists or law professors?
- If you’re like me, you don’t have funding for RAs, so we need to find ways to automate and simplify our empirical workflows, balancing methodological rigor with accessibility.

Opportunities

- Methodological improvements from “law as data” and NLP
- Enhancing workflow through automation and scalability
- Advancing investor protection and societal goals, maybe?

Where to begin?

We're in the Bay area, so "learn to code":

- Grole & Wickham, *R for Data Science*
- Imai & Williams, *Quantitative social science: An introduction in tidyverse*
- Slige, *Text mining in R: A tidy approach*

Law as data:

- Livermore, ed., *Law as Data: Computation, text, and the future of legal analysis*
- Stoltz & Taylor, *Mapping Texts: Computational text analysis for the social sciences*
- Stewart, Grimmer & Roberts, *Text as Data: A new framework for machine learning and the social sciences*

Questions?