

# 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

# 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

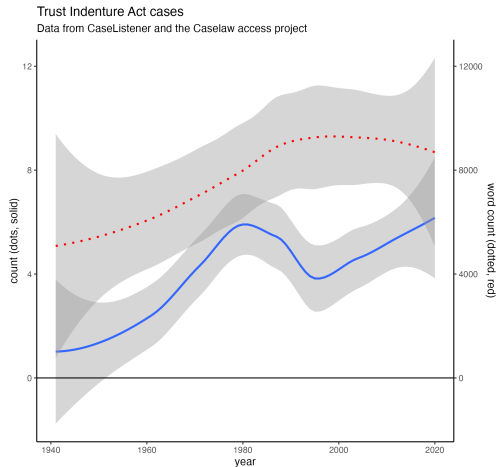
## Tracking Regulatory Evolution

- 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 (ooooooooh, 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)
- Insights into litigation strategies and case outcomes from docket reports.

# 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 Scandinavian 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 in econometrics and NLP.
- Enhancing workflow through automation and scalability.
- Advancing investor protection and societal goals.

Questions?