Legal Text as a New Source of Economic Evidence

Mona Birjandi,[*] Steven Singer [**] , Tahmid Ahmed [***]

[*] Mona Birjandi, Dr. Mona Birjandi is the Principal Economist and Director of Data Analytics at Outten & Golden LLP. She leads the firm's economics and data analytics team, supporting employment litigation through statistical analysis, economic modeling, and data-driven insights. Before joining the firm, Dr. Birjandi worked in both economic consulting and corporate analytics. She has advised on matters involving antitrust, intellectual property, and labor and employment, and previously served as Principal Economist at Petco Health and Wellness. She holds a Ph.D. and M.A. in Economics from Yale University and an M.A. in Law and Diplomacy from Tufts University. She may be reached at mbirjandi@outtengolden.com.

[**] Steven Singer, Steven Singer is a Data Scientist at Outten & Golden LLP, where his research lies at the intersection of data analytics, optimization, and social justice. Before joining Outten & Golden, Steven served as a Senior Analyst at an economic consulting firm, where he conducted economic and statistical analyses in antitrust litigation and supported expert witnesses in complex cases. He previously worked as a Research Assistant at the Kellogg School of Management, examining how alignment between firm ownership and leadership affects IPO outcomes. Steven holds a B.A. in Mathematics from Northwestern University and can be reached at ssinger@outtengolden.com.

[***] Tahmid Ahmed is a Data Analyst at Outten & Golden, LLP where he has developed proprietary tools for economic analysis and supports lawyers through data analytics and statistical analysis. Before joining the firm, Tahmid worked in economic consulting and data science at Analysis Group where he conducted analysis for antitrust, labor and employment and healthcare matters. He previously served as a research assistant in the Government Department at Harvard and holds an A.B. in Government and Economics at Harvard University. He can be reached at tahmed@outtengolden.com.

I. Abstract

Legal systems produce an enormous volume of textual records that encode economically relevant information (in U.S. federal courts alone, PACER contains over one billion documents across tens of millions of cases). Yet this trove of legal text has remained largely underutilized in empirical economics, due to the difficulty of processing lengthy, unstructured legal information at scale. Recent advances in AI, particularly in natural language processing and large language models, make it feasible to analyze these documents in bulk and convert qualitative legal information into structured, quantitative data that can be tabulated and processed systematically. We highlight how this enables new research applications across labor markets, technology, macroeconomics, and competition. For example, mining expert witness filings reveal details about a high-skill labor market for expert services that was previously invisible. Analyzing patent litigation records shows how enforcement risks may deter entry and shape innovation incentives. Similarly, surges in breach-of-contract suits during crises can serve as real-time indicators of economic stress, while class action and wage-and-hour lawsuits yield data on product failures and workplace wage violations. By treating court documents as data, economists can observe market behavior, enforcement actions, and institutional processes at a depth not captured by traditional datasets. We also examine the methodological, technical, and institutional challenges of using legal text at scale, such as converting document formats, handling varied legal jargon and procedures, dealing with incomplete data coverage, and ensuring strong quality control, all of which requires significant effort to fully realize the potential of extracting data sources from court documents.

II. Introduction

Courts around the world generate an immense volume of textual data each day, including case filings, motions, judicial opinions, expert reports, and trial transcripts. In the United States federal court system alone, the Public Access to Court Electronic Records (PACER) system houses over one billion documents across tens of millions of cases,[i] and state courts add volumes more.[ii] These legal documents encompass rich information that is highly relevant to economists, policymakers, and legal professionals, such as case metadata (e.g., filing dates, number of parties, court level, length of trial), monetary amounts (e.g., damages claimed, fines, attorney fees, settlement amounts, expert witness fees), party and representation data (e.g., attorneys, victims), and case outcome metrics, among others. However, historically, this information has been severely underutilized due to difficulty of accessibility and production of lengthy, structured data at scale manually.

Recent advances in Large Language Models (LLMs) promise to change this. LLMs can read and summarize complex language, identify patterns, and even extract structured data from free-form text. This article explores how AI can transform legal documents into valuable datasets, the types of insights we can extract, and the potential applications in economic research. We also discuss the benefits of this approach and the challenges or risks. The goal is to provide a conceptual, yet comprehensive, overview of this emerging interdisciplinary frontier for researchers, economists, legal professionals, policymakers, and technologists alike.

III. Legal Text as An Untapped Economic Resource

1) Scale and Diversity of Legal Text

Modern court systems generate an extraordinary volume of text as a byproduct of routine adjudication. Dockets accumulate complaints, answers, motions, briefs, judicial

opinions, expert reports, deposition designations, hearing and trial transcripts, sentencing memoranda, settlement agreements, and remedial orders. These materials capture the actions of individuals, firms, and government agencies in legal contexts where incentives and strategic behavior are often clearly visible. As a result, court text contains information that is difficult to observe in other data sources such as allegations and defenses about business practices, the timing and content of enforcement activity, contract terms and disputed interpretations, the economic logic offered by experts, and the remedies ultimately imposed or negotiated.

In the United States, the sheer scale alone places court text in the same "big data" class as administrative microdata or large commercial datasets. PACER provides instantaneous access to more than 1 billion documents filed at more than 200 federal courts.[iii] The federal case management databases encompass around 65 million cases and approximately 1.7 billion docket entries as of mid-2025.[iv] This volume is steadily growing as each year the federal courts see 2 million new cases and tens of millions of new documents added to the system.[v] State courts, administrative tribunals, and regulatory bodies contribute substantial additional volume. State courts handle the vast majority of cases (roughly 95% of all U.S. cases), amounting to tens of millions of new cases annually.[vi] Recent data indicate about 66 million cases per year in state courts.[vii] Administrative courts and agencies add further to this collection: for instance, the U.S. immigration court system's backlog reached approximately 3.6 million pending cases by 2024.[viii] Class action settlement administrators, arbitration forums, and regulatory enforcement actions similarly produce large collections of documents (e.g. settlement notices, monitoring reports, hearing transcripts).

Notably, much of this textual output is at least partially accessible to researchers and the public. Some of these are for free, for example systems like PACER (and state equivalents) do allow retrieval of filings, however, a barrier is that certain sources that house

and preprocess court data (i.e., DocketAlarm, Lexis) often require paid fees.[ix,x] Despite the massive scale and detail of court texts, and the emerging infrastructure to process and house these data, they remain a largely untapped resource in economic research.

2) Why This Resource Has Remained Under-Processed

For decades, the rich textual data produced by courts was accessible but difficult to analyze at scale. Early empirical legal studies had to rely on small hand-coded samples or very limited text mining because parsing the nuance of legal documents was beyond the reach of simple algorithms. Several key barriers kept most court filings and transcripts from being quantitatively analyzed:

1. **Unstructured format:** Legal filings are long, narrative documents rather than neatly formatted databases. Important information is buried in prose and footnotes, making it hard for software to extract facts or figures without human guidance. Unlike structured datasets, a court case file might span dozens of pages of free-form text with no uniform template.
2. **Complex language:** Legal documents use specialized jargon and dense syntax that require domain expertise to interpret. The writing is highly formal and designed for precision – even a single comma can alter obligations or rights. They rely on archaic terms and legalese not found in everyday language, making them less accessible to non-specialists and traditional NLP techniques, and documents are filled with cross-references to statutes or prior cases. This linguistic complexity has historically confounded one-size-fits-all text parsers.[1][xi]
3. **Variability:** Every case and jurisdiction can have its own terminology and document organization. A contract dispute filing looks very different from a criminal indictment

[1] https://arxiv.org/pdf/2410.21306.

or a Supreme Court brief. Even within the same genre, lawyers have individual styles. This heterogeneity defeats rule-based extraction systems that expect consistent formats. Moreover, many legal documents (especially older ones) exist only as scanned PDFs or images, which creates additional challenges for machine reading and can render some files completely unreadable.

4. **Volume:** As previously discussed, the sheer scale of court text is overwhelming and constantly growing. Manually processing and coding millions of pages across non-uniform documents to extract structured data can be prohibitively time-consuming and costly, making large-scale quantitative analysis impractical without automation. As a result, vast swaths of court filings remained untapped for quantitative analysis, despite the valuable information they contain about legal processes and outcomes.
5. **Context and semantic understanding:** Extracting meaningful variables from legal text requires interpreting what the document is legally about, not merely searching for keywords. A simple keyword search for "disparate impact," for example, will indeed retrieve Title VII cases, but it will also pull in non–Title VII cases that use the same terminology. In *Smith v. City of Jackson* (2005), the Court discusses "disparate impact," yet the claim arises under the Age Discrimination in Employment Act (ADEA) rather than Title VII.[xii] This illustrates why reliable extraction pipelines must infer the statutory basis and legal theory from context (e.g., identifying the pleaded cause of action and cited statute) rather than treating keyword hits as labels. This level of contextual extraction at scale has only become feasible with the advent of large language models.

As a result, despite being routinely generated and increasingly accessible to policymakers and researchers, court records have rarely been systematically mined for causal or descriptive empirical analysis of policy effectiveness or economic outcomes. The result is

a familiar problem in economics: institutions overflow with data but lack usable information, not because data is unavailable, but because transforming unstructured text into reliable, analyzable measures remains costly and complex.[xiii]

Other professions have begun to overcome similar challenges as text analysis technology improved. In finance, for example, researchers now mine unstructured corporate filings to extract sentiment and predictive indicators,[xiv] while innovation economists use patent language to measure technological novelty and impact.[xv] These examples show how economists are increasingly using computational linguistics to turn unstructured, institutionally generated text into structured data for empirical analysis.[xvi] The success of these methods with SEC filings, patents, and news articles highlights the untapped potential of legal texts which were long overlooked simply because the tools to analyze them weren't ready.

3) AI and LLM Breakthrough

The advent of powerful LLMs changes the equation as AI systems can now read and interpret legal text with an increasingly advanced level of comprehension. Modern language models have demonstrated the ability to summarize judicial opinions, extract key facts from complaints, classify legal issues, and even draft documents.[xvii] Crucially, they can do this at scale and speed. For instance, an AI can be instructed to summarize thousands of pages of contract or case law in minutes, a task that would take a human team weeks.[xviii] It is important to note that large-scale processing may require the use of an API and/or more computational infrastructure for more advanced processing. Regardless, this combination of scalability and understanding means we can finally unlock the economic insights latent in court documents. In short, court text is poised to become the next major untapped economic

dataset, as AI tools make it feasible to analyze legal documents in bulk and transform qualitative legal information into quantitative data.

IV. Examples of Datasets from Legal Documents

Using AI to process legal documents can yield a variety of novel datasets that were previously difficult to compile. Below we outline several types of data that can be harvested from court records, illustrating the range of economic insights they offer:

- **Expert Billing Rates**: Many court filings contain detailed information on testifying experts' compensation, including hourly rates, total billed, and sometimes hours worked by task (e.g., data review, report drafting, deposition prep, testimony). By mining expert reports, declarations, invoices, and related fee submissions, AI tools can convert these scattered references into a structured dataset of expert billing rates and total expert costs across industries, case types, courts, and time. For empirical economists, this enables direct study of the market for testifying expert services which is a high-skill and niche labor market that is otherwise largely unobserved. With a sufficiently large dataset, researchers can (i) estimate pricing gradients by forum (federal vs. state), region, and case complexity; (ii) quantify returns to observable expert attributes (discipline, seniority, publication record, prior testimony frequency); (iii) measure changes over time; and (iv) examine how expert pricing interacts with litigation incentives—such as whether higher expert costs discourage filing, or shift bargaining power in settlement negotiations. In short, extracting expert-fee information from court text converts a largely opaque professional-services market into analyzable microdata that can support research in labor market.

To illustrate what this looks like in practice, below we report results from a limited pilot search of a very small set of case filings (not a comprehensive sample). Even this narrow scrape yields extractable expert-rate observations, including: (i) an expert engagement in an ADA class action reflecting a $300/hour rate (see Figure 1 below);[xix] (ii) an expert declaration discussing the reasonableness of a $585/hour rate (2025);[xx] and (iii) a Texas property appraisal matter reflecting a $200/hour expert rate related to fee reasonableness.[xxi] These examples are intended to demonstrate feasibility and the types of fields that can be extracted (rate, case context, expert role, year, jurisdiction, retaining party), not to characterize the market as a whole.

Figure 1- Example of Expert Billing Rate Information Contained in a Legal Document

> We have been engaged by Nye, Stirling, Hale & Miller, LLP in this matter. PFC is not engaged on a contingency basis for this case. An engagement letter has been attached as Exhibit C. My hourly rate for this engagement is $300 per hour, for both analysis and testimony. Neither my compensation nor my opinions are dependent on the outcome of this litigation.

The same approach can be scaled systematically to cover experts across different case categories, geographies, time periods, and expert types and seniority levels. With a sufficiently large dataset, economists can use the resulting microdata to study the market for testifying expert services by estimating pricing gradients across forums and regions, quantifying returns to expert attributes measuring trends over time, and examining how expert costs interact with litigation incentives and access to justice.

- **Case Outcomes and Litigation Metrics**: Court dockets, orders, and opinions can be converted into standard outcome variables used in empirical work such as disposition (dismissed/settled/trial/judgment), prevailing party, monetary

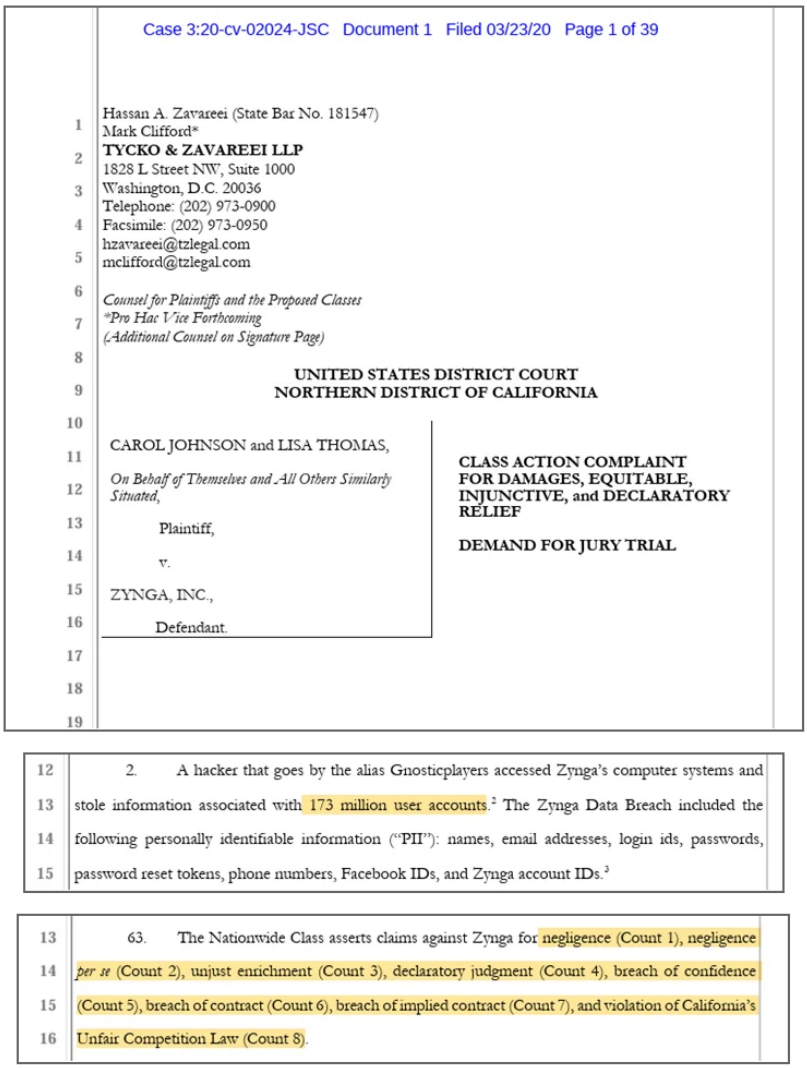
outcomes (damages, fee awards, disclosed settlement amounts), and time-to-resolution (filing-to-termination). Recent NLP/LLM approaches show these measures can be recovered at scale.[xxii] As another example, in securities class actions, by analyzing legal text, a Stanford University report finds that 105 cases settled in 2022, with a $3.8B total settlement value and a $13M median settlement.[xxiii] In patent litigation, Lex Machina reports >$4.3B in damages awarded in 2024, across 90+ cases with damages awards.[xxiv] These types of statistics ultimately originate in court records and related legal documents, which are then compiled into structured datasets. For economists, these AI-extracted outcomes make it possible to take litigation seriously as an economic institution—one that shapes incentives, bargaining, and market structure.[2]

- **Allegations, Legal Issues, and Fact Patterns**: The text of complaints, briefs, and transcripts can be transformed into datasets about what is being litigated and how. For example, AI could read thousands of complaints to categorize the types of allegations companies are facing and the factual scenarios behind them. We could construct time-series data on the frequency of certain claims, for example, tracking a rise in "privacy and data breach" lawsuits in the last five years, or mapping which industries are seeing more antitrust enforcement. As an illustrative example of how a single complaint becomes structured data, is the complaint for *Johnson & Thomas v. Zynga, Inc.* (N.D. Cal., Case No. 3:20-cv-02024-JSC, filed 03/23/2020), which alleges a major data breach affecting 173 million user accounts and identifies the compromised data elements (e.g., names, email addresses, login IDs, passwords, and password

[2] Using this approach, the authors are developing a comprehensive dataset of Title VII disparate-impact cases in the United States, including standardized case-level metrics. If completed in time, the dataset and findings from the data will be shared with the audience at the presentation of this research.

reset tokens), along with asserted legal theories (including negligence and consumer protection claims). A corresponding dataset row could include fields such as: (i) case_number: 3:20-cv-02024-JSC, (ii) filing_date: 2020-03-23, (iii) allegation_category: privacy / data breach, (iv) accounts_affected (alleged): 173,000,000, (v) plaintiffs: Carol Johnson; Lisa Thomas, (vi) defendant: Zynga, Inc., (vii) negligence; negligence per se; unjust enrichment; declaratory judgment; breach of confidence; breach of contract; breach of implied contract, and violation of California's Unfair Competition Law.[xxv] (see Figure 2 below for excerpts from the complaint)

Figure 2- Excerpts from Johnson & Thomas v. Zynga, Inc. Complaint



Case 3:20-cv-02024-JSC Document 1 Filed 03/23/20 Page 1 of 39

Hassan A. Zavareei (State Bar No. 181547)
Mark Clifford*
TYCKO & ZAVAREEI LLP
1828 L Street NW, Suite 1000
Washington, D.C. 20036
Telephone: (202) 973-0900
Facsimile: (202) 973-0950
hzavareei@tzlegal.com
mclifford@tzlegal.com

Counsel for Plaintiffs and the Proposed Classes
**Pro Hac Vice Forthcoming*
(Additional Counsel on Signature Page)

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

CAROL JOHNSON and LISA THOMAS,

On Behalf of Themselves and All Others Similarly Situated,

Plaintiff,

v.

ZYNGA, INC.,

Defendant.

CLASS ACTION COMPLAINT FOR DAMAGES, EQUITABLE, INJUNCTIVE, and DECLARATORY RELIEF

DEMAND FOR JURY TRIAL

2. A hacker that goes by the alias Gnosticplayers accessed Zynga's computer systems and stole information associated with 173 million user accounts.[2] The Zynga Data Breach included the following personally identifiable information ("PII"): names, email addresses, login ids, passwords, password reset tokens, phone numbers, Facebook IDs, and Zynga account IDs.[3]

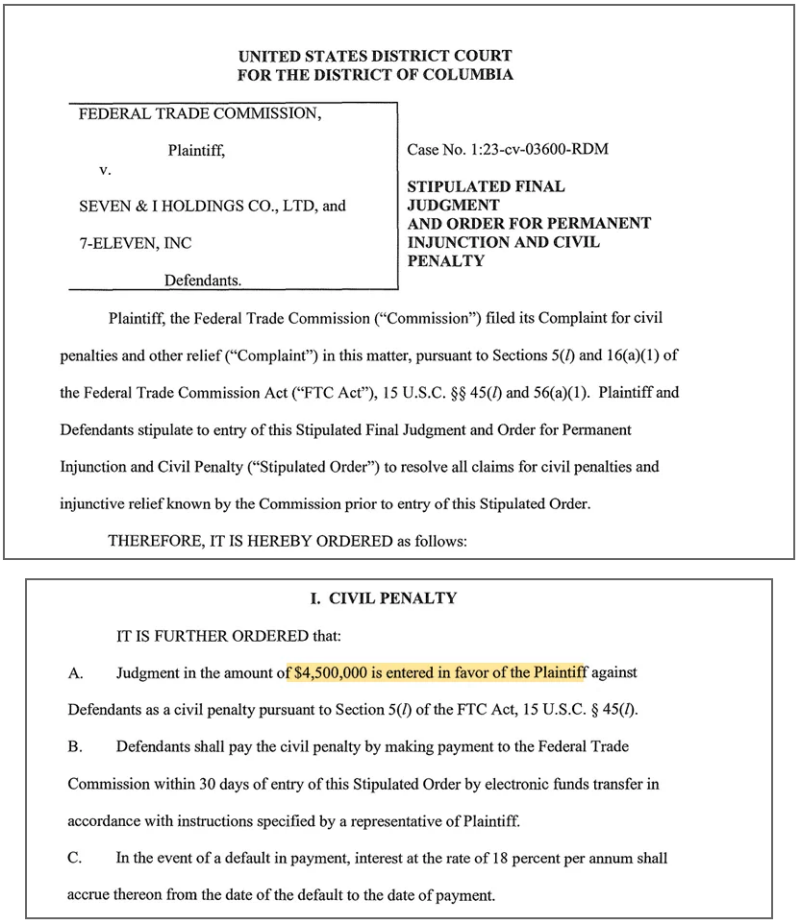
63. The Nationwide Class asserts claims against Zynga for negligence (Count 1), negligence *per se* (Count 2), unjust enrichment (Count 3), declaratory judgment (Count 4), breach of confidence (Count 5), breach of contract (Count 6), breach of implied contract (Count 7), and violation of California's Unfair Competition Law (Count 8).

Similarly, mining trial transcripts or deposition texts could produce datasets on witness testimony—like how often certain keywords appear in cases involving mergers. These kinds of datasets let researchers measure trends in corporate behavior and regulatory focus through the lens of litigation. They also enable social scientists to detect patterns like the prevalence of certain evidence or arguments (for instance, how often is "email evidence" cited in employment cases, or how frequently do defendants use a particular defense strategy in toxic tort cases).

- **Regulatory Enforcement and Fines**: Many regulatory actions end up in either court filings or administrative proceeding records that can be parsed. AI can extract data on enforcement trends, such as the size of fines or penalties imposed in securities cases, or the frequency of enforcement actions by agencies like the SEC, EPA, or FTC. For example, by reading consent decrees or settlement documents, one might build a dataset of corporate penalties (fines, required remedies) and see how they correlate with the targeted company's stock performance or industry sector. These data help evaluate the economic impact of regulation and the deterrence effect of enforcement, and how regulatory focus shifts over time.

  As an example, a recent enforcement case demonstrates this feasibility: 7-Eleven, Inc. agreed to pay a $4.5 million civil penalty to settle an FTC lawsuit alleging violation of a prior antitrust consent order. The FTC's stipulated judgment noted that this $4.5 million fine was the largest ever collected for a notice-order violation.[xxvi] By processing documents like consent decrees and settlement orders, an AI system could systematically capture fields such as the (i) penalty amount ($4.5M), (ii) the violation (failing to give required merger notice), (iii) the enforcing agency (FTC), and (iv) the date (2025) – building a dataset of enforcement actions and fines for economic analysis of regulatory policy. (*see* Figure 3 below for excerpts from the complaint)

Figure 3- Excerpts from FTC v. Seven & I Holding Co. LTD Complaint



UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

FEDERAL TRADE COMMISSION,

Plaintiff,

v.

SEVEN & I HOLDINGS CO., LTD, and

7-ELEVEN, INC

Defendants.

Case No. 1:23-cv-03600-RDM

STIPULATED FINAL JUDGMENT AND ORDER FOR PERMANENT INJUNCTION AND CIVIL PENALTY

Plaintiff, the Federal Trade Commission ("Commission") filed its Complaint for civil penalties and other relief ("Complaint") in this matter, pursuant to Sections 5(*l*) and 16(a)(1) of the Federal Trade Commission Act ("FTC Act"), 15 U.S.C. §§ 45(*l*) and 56(a)(1). Plaintiff and Defendants stipulate to entry of this Stipulated Final Judgment and Order for Permanent Injunction and Civil Penalty ("Stipulated Order") to resolve all claims for civil penalties and injunctive relief known by the Commission prior to entry of this Stipulated Order.

THEREFORE, IT IS HEREBY ORDERED as follows:

I. CIVIL PENALTY

IT IS FURTHER ORDERED that:

A. Judgment in the amount of $4,500,000 is entered in favor of the Plaintiff against Defendants as a civil penalty pursuant to Section 5(*l*) of the FTC Act, 15 U.S.C. § 45(*l*).

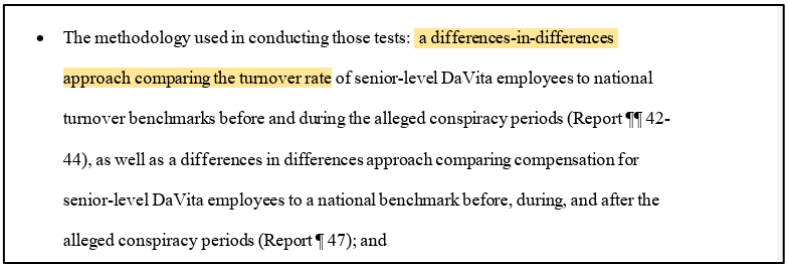
B. Defendants shall pay the civil penalty by making payment to the Federal Trade Commission within 30 days of entry of this Stipulated Order by electronic funds transfer in accordance with instructions specified by a representative of Plaintiff.

C. In the event of a default in payment, interest at the rate of 18 percent per annum shall accrue thereon from the date of the default to the date of payment.

- **<u>Economic Methods and Modeling Choices in Expert Reports</u>**: A large and economically rich category of court filings consists of expert reports, many of which contain sophisticated applied microeconomics and econometrics (e.g., regression and matching designs in discrimination cases, damages and pass-through models in antitrust, hedonic and conjoint analyses in product markets, event studies in securities, royalty and bargaining models in IP, and survey-based valuation in consumer cases). By systematically processing these reports, AI can build datasets that code (i) the methods used (e.g., DiD, fixed effects, hazard models, event studies, conjoint, nested logit, structural demand, bargaining/royalty frameworks); (ii) key design choices (unit of observation, control groups, time windows, functional forms, clustering choices); (iii)

reported effect sizes and uncertainty (coefficients, standard errors, confidence intervals); and (iv) indicators of judicial reception and admissibility by linking reports to Daubert rulings and evidentiary orders. This opens the door to new empirical research on the market for economic evidence, such as identifying which modeling choices are more likely to be admitted in court, whether certain methods tend to draw greater judicial skepticism, and how the availability of court-accepted economic analysis shifts in response to legal standards or precedent. In this context, expert reports can be viewed as a body of applied economics, offering a unique lens through which to study how analytic practices evolve under adversarial conditions and how evidentiary rules influence the development and credibility of empirical work. As an example, in an antitrust no-poach case, the defense expert's report applied a difference-in-differences econometric model to assess the impact of alleged agreements on employee turnover. Specifically, the expert compared the firm's employee turnover rates before vs. during the alleged conspiracy period to national benchmarks, employing a differences-in-differences benchmark methodology to isolate the effect of the no-poach agreement.[xxvii] This example shows that complex economic methods (here a DiD analysis) are indeed found in court filings. An AI-driven review of many such expert reports could tabulate (i) which methods are used in which cases, (ii) the context, and even (iii) link to outcomes like whether the court deemed the analysis credible or admissible. (*see* Figure 4 below for excerpts from the expert report)

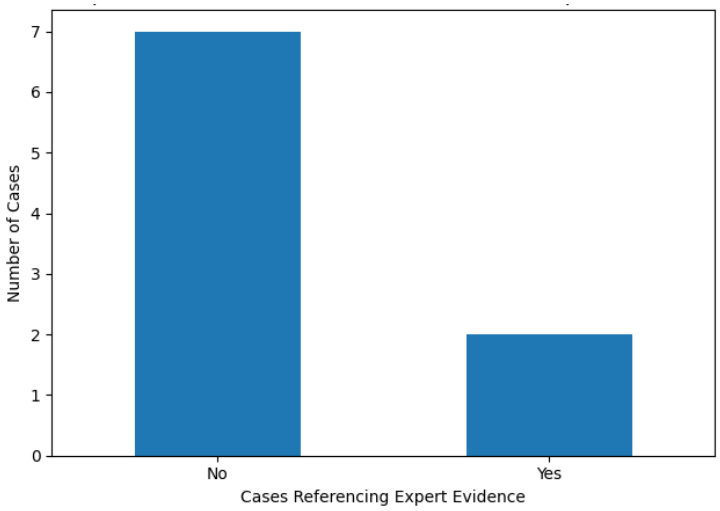
Figure 4-Excerpts from Legal Text Related to Case No. 21-cr-0229-RBJ



- The methodology used in conducting those tests: a differences-in-differences approach comparing the turnover rate of senior-level DaVita employees to national turnover benchmarks before and during the alleged conspiracy periods (Report ¶¶ 42-44), as well as a differences in differences approach comparing compensation for senior-level DaVita employees to a national benchmark before, during, and after the alleged conspiracy periods (Report ¶ 47); and

As another simple illustrative exercise using publicly available Supreme Court opinions (e.g., via Justia),[xxviii] we identified all nine antitrust decisions issued by the U.S. Supreme Court between 1980 and 1989 and searched the full text of each opinion for references to "expert". Only two cases—*Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, 472 U.S. 585 (1985),[xxix], [xxx] and *Cargill, Inc. v. Monfort of Colorado, Inc.*, 479 U.S. 104 (1986)[xxxi]—explicitly reference economic expert testimony or expert witnesses in the Court's discussion. In *Aspen Skiing*, the Court cites testimony from an expert marketing witness regarding consumer demand, while in *Cargill* it references an expert's estimates of post-merger market capacity. Thus, only 2 of 9 antitrust decisions (~22%) from the 1980s explicitly reference expert economic evidence. While purely descriptive, this example illustrates how structured analysis of legal text can reveal when and how economic expertise enters judicial reasoning, even using simple methods and fully public data sources.

Figure 5- Supreme Court Antitrust Decisions of the 1980s: Expert Evidence



In summary, the examples above are illustrative rather than exhaustive: AI-enabled processing can convert unstructured legal texts into structured datasets spanning expert fees, litigation costs, case outcomes and durations, the substantive content of disputes, and the evolution of legal rules. Each such dataset can unlock new empirical insights. The broader point is that nearly any fact, choice, or event recorded in court documents can become a measurable variable once extraction is reliable and scalable. The next section discusses how researchers and analysts can use these data in economic and policy research.

V. Legal Data in Economic Research

Below we develop several examples illustrating how litigation records can answer canonical economic questions across various disciplines.

- ***Innovation/IP:*** *Does patent-enforcement risk deter entry and redirect R&D—i.e., do incumbents use suits to raise rivals' costs?*

  Traditional measures like patent counts, citations, and R&D spending reflect innovation activity and impact, but they overlook an important dimension: the hidden cost of innovating in certain technological areas, including the risk of litigation, the extent of contested claims, and patterns of strategic enforcement.

By using AI to process patent complaints, claim charts, expert reports, rulings, verdicts, and settlement or licensing disclosures, researchers can construct a litigation-risk map of the innovation landscape. This allows for the analysis of which technology clusters face frequent legal challenges, the dynamics of who sues whom, how closely asserted patents align with the products they target, and the timing and stakes of enforcement efforts. These new measures make it possible to study how legal risk affects real economic behavior. Researchers can examine whether it deters market entry, influences R&D decisions after litigation, or reflects strategic targeting around product launches or financing events. In turn, these insights help link legal conflict to patterns of investment, competition, and technological progress.

- ***Macroeconomics:*** *How do large shocks (pandemics, financial crises, disasters, major regulatory changes) propagate through the economy, and which sectors/regions show strain first?*

  Traditional macro data (such as GDP, employment, defaults, surveys) are essential but largely capture outcomes and are released on scheduled timetables that can lag the underlying events.[3] [xxxii] Court filings, by contrast, enter the public record essentially as disputes arise and can be accessed quickly. A concrete illustration is the pandemic spike in federal civil litigation: civil case filings in U.S. district courts increased from 297,877 (FY2019) to 470,581 (FY2020)—a ~58% jump reported by the U.S. Courts.[4] [xxxiii] Using AI on breach-of-contract suits, insurance disputes, and bankruptcy petitions, researchers can construct a legal stress index: counts of shock-related cases,

[3] https://www.bea.gov/news/schedule.
[4] https://www.uscourts.gov/data-news/reports/statistical-reports/judicial-business-united-states/judicial-business-2021/us-district-courts-judicial-business-2021

shares invoking "force majeure/impossibility," and topic mixes by industry and geography. These measures allow macro and regional analysis of adjustment mechanisms which identify where contract failures cluster, and how legal institutions shape the allocation of losses and the speed of recovery.

- ***Industrial Organization:*** *Where do products and industries fail technically and what are the economic consequences?*

  Large-scale consumer class action filings provide a high-frequency record of alleged harm and its resolution. In 2023 alone, federal courts saw 9,705 class action filings, including 4,460 related to consumer protection, according to Lex Machina's Consumer Protection Litigation Report.[xxxiv] These cases often surface technical failures that have crossed a threshold of public and legal concern, such as defects in batteries, brakes, software algorithms, or contaminated products. With tools like AI, researchers can now analyze the underlying complaints, expert reports, class certification briefs, and settlement agreements to build detailed datasets by industry and time. These datasets can track the type of failure, the affected product lines, when and where incidents occurred, and how the issues were ultimately resolved—through recalls, repair programs, settlements, or other remedies. This kind of analysis opens the door to new economic research. Scholars can examine whether competitive pressure or cost-cutting leads to more defects, how quickly problems spread across firms and suppliers, and how markets react once failures become public. It also allows for estimates of consumer harm and assessments of how well compensation mechanisms address losses like downtime, safety risks, or out-of-pocket costs. Importantly, it sheds light on how legal exposure

influences firms' decisions about pricing, warranties, product design, and investments in quality control

- ***Labor/Employment:*** *Where do unequal treatment and wage violations concentrate, and do legal institutions change employer behavior?*

  Employment lawsuits provide a detailed and time-stamped view of workplace disputes and how they are resolved. With the help of AI, researchers can analyze a wide range of court documents, such as complaints, class certification briefs, expert reports, settlement agreements, consent decrees, and attorney fee petitions, to build datasets that reflect both worker and firm characteristics. These datasets can track the nature of the legal claims, including distinctions between disparate treatment and disparate impact, or misclassification and unpaid overtime. They can also reveal information about the protected class involved, the job context, the outcome of the case, and the types of remedies awarded, such as back pay, policy changes, or injunctive relief. Over time, these records help show how long cases take to resolve, how often they recur, and how outcomes vary across industries, firms, and regions. This structured view of employment litigation supports several lines of economic research. It allows analysts to examine how firms respond to legal exposure, such as whether they change hiring, promotion, or pay practices after being sued. It provides insight into wage theft by using recovered back pay as a proxy for underpayment, helping researchers understand where underpayment is most common and how enforcement affects wages and employment. It also sheds light on barriers to legal access, showing how outcomes differ between represented and unrepresented workers, and highlighting the role of legal aid in shifting bargaining power. By treating

employment litigation as a source of empirical data, researchers can better measure labor market harms and the legal institutions that influence how those harms are addressed.

These examples are intended to be illustrative rather than exhaustive. They show how litigation records, once processed, can be used to study core economic mechanisms. More broadly, our analysis shows that almost any piece of economically relevant information found in court records can be turned into data, giving researchers new tools to study markets and institutions beyond what's possible with traditional government or commercial datasets.

VI. Lessons and Pitfalls

As researchers increasingly use AI to extract data from court documents, it is essential to understand the method's capabilities, limitations, and downstream consequences. This section draws on applied experience to highlight key lessons and pitfalls for producing reliable, responsible, and reproducible datasets from legal text.

1) Legal Document Format and Conversion

Legal documents are typically created and stored as PDFs, but PDFs are not reliably machine-readable. They often include scanned pages, complex layouts, tables, and footnotes that disrupt straightforward text extraction. To make them usable for AI processing, significant preprocessing is usually required (OCR, layout recovery, and cleaning). However, these steps can introduce systematic errors, like missing sections or scrambled text, and often slow down the entire pipeline.

One important lesson from real-world AI applications is that file format matters. For large-scale analysis, it is best to convert source documents into open, structured formats that preserve both the text and its context. JSON, for example, is best used for storing legal

information because it handles hierarchical structure and metadata well.[5][xxxv] CSV/Parquet support tabular outputs for statistical analysis; and SQL databases facilitate querying at scale. XML is another format sometimes used in legal data exchange because it allows definition of a schema; initiatives like the National Open Court Data Standards (NODS) project are developing technical court data standards to support the creation, sharing and integration of court records across jurisdictions.[6][xxxvi]

In general, plain text files (txt) are simple and easy to process with AI tools, but they lack metadata and structural information, which can limit their usefulness. In short, choosing the right format is essential. Structured, open formats like JSON, CSV, and XML make it much easier to analyze, share, and scale legal data projects. In contrast, working directly with unstructured or proprietary formats, like PDFs, Word documents, or image scans, often leads to delays, errors, and added complexity.

2) Processing and Maintaining Legal Databases

Maintaining AI-created legal databases is an ongoing, resource-intensive process that blends technical, legal, and logistical challenges. Given the scale of legal records, which can span millions of dockets, filings, and opinions, any serious attempt to structure and analyze this information requires robust backend infrastructure. As discussed previously, legal documents can be quite lengthy and unstructured, often in formats like PDF that demand preprocessing such as OCR, classification, and parsing before they are fed into AI models. This in turn requires significant computation resources (which also depends on the AI model used), including scalable cloud storage, high-throughput data pipelines, and GPU-accelerated environments for running large language models efficiently. Costs associated with

[5] File Formats
[6] Interoperable Legal AI for Access to Justice | Yale Law Journal

commercial APIs or model fine-tuning can quickly escalate as the volume of data increases, making infrastructure planning and budgeting a critical part of legal data projects.

3) Coverage Gap and Keeping Data Current

A key limitation is that many legal documents are not publicly accessible. Filings in state courts or sensitive cases are often paywalled, redacted, or never released. As a result, AI-based datasets may offer only a partial view of the litigation landscape. Addressing these gaps often requires additional resources to locate and obtain restricted documents. In addition, legal cases are constantly updating: dockets evolve, new evidence is filed, decisions are amended, and appeals reshape outcomes. A database that captures only the initial complaint or a single snapshot in time risks becoming outdated or misleading. Effective systems must therefore be designed to refresh data continuously, monitoring dockets for new entries and automatically integrating subsequent filings into existing records. Without these updates, even well-structured databases will degrade in reliability over time.

4) Other Processing Challenges

There are several additional challenges in creating clean, reliable datasets from legal text:

- **Label ambiguity and drift:** Many target variables are not stated in a single, unambiguous location (e.g., “who won,” “settlement amount,” “reason for dismissal”). They may be implicit, contested, or described differently across documents. Even when a label is well defined, its legal meaning can vary across jurisdictions and change over time, creating inconsistent data extraction. A more advance model or API may solve this problem but there could be significant issues at scale.

- **Jurisdictional heterogeneity:** Terminology, docketing practices, document types, and procedural rules vary widely across federal, state, and county courts. A model or prompt set that performs well in one jurisdiction may fail in another, generating systematic cross-jurisdiction inconsistencies.
- **Cross references and versioning:** Legal text is dense with references to exhibits, prior filings, record citations, and sealed appendices. If referenced material is unavailable, extracted variables may be incomplete. Amendments and superseded filings add further pitfalls: the operative complaint, controlling order, or final version is not always obvious without careful tracking and inspection.
- **Nonrandom missingness from sealing and redaction:** Redactions and sealing are not random as sensitive matters are more likely to be withheld or heavily redacted. This creates selection bias in outcomes researchers often care about (e.g., high-dollar settlements or confidential terms).
- **Entity resolution errors:** Names in legal records are often inconsistent. A single name might refer to different entities, or the same entity may appear under multiple variations due to mergers, rebranding, or formatting differences. Without careful entity resolution, researchers risk misattributing outcomes, combining unrelated records, or splitting one actor into several, which can distort results and undermine analysis.
- **Monetary amount misclassification:** Court text contains multiple "amounts" that can be easily confused (e.g., requested versus awarded versus paid, gross settlement fund versus net-to-class). LLMs can extract the wrong number with high confidence unless the pipeline enforces clear context rules.

- **Reproducibility under model and API changes:** With proprietary APIs or frequently updated models, the same prompt may yield different outputs over time. Without strict version control (model version, prompt templates, temperature, retrieval context, and document snapshot), results may be difficult to replicate.
- **Hallucination:** Legal settings are sensitive to fabricated or inferred content. Yet many extraction tasks do not come with explicit "guardrail metrics" that automatically validate correctness (e.g., whether a cited statute appears in the record, whether a settlement amount is explicitly stated, or whether the extracted disposition matches the docket entry). As a result, LLM outputs can appear coherent while being wrong—misstating a holding, inventing a number, or attributing a legal theory that is not pleaded. Small hallucinations can materially change an extracted variable or interpretation. This means extraction pipelines must be prompt- and schema-disciplined (clear definitions, constrained output formats, and document-grounded quoting or span references) and paired with validation checks and human review.

These challenges show that AI extraction isn't fully automatic. Building reliable legal datasets require consistent quality checks, clear documentation, and transparency about data sources and assumptions. Most importantly, human review is essential as auditing outputs and interpreting legal concepts that models alone can't resolve.

VII. Concluding Remarks

This study has demonstrated that legal text is abundant in nature but rarely utilized in economic research. With advances in AI and natural language processing, information embedded in dense legal filings can now be transformed into structured data, opening new opportunities for empirical analysis. What was once difficult to observe, such as the details of

disputes, enforcement actions, or contractual breakdowns, can now be studied at scale. Examples like patent litigation, crisis-related contract disputes, consumer class actions, and employment cases highlight the range of economic questions that can be explored using court records. More broadly, these documents offer a rich source of behavioral and institutional data, comparable in scope and utility to traditional administrative datasets.

However, fully realizing this potential requires overcoming significant challenges. Technical hurdles include converting and parsing millions of diverse, often non-machine-readable documents at scale: many legal records are scanned or complex PDFs that demand OCR and extensive preprocessing, as well as substantial computing infrastructure to handle the volume. Methodological challenges arise from the heterogeneity of legal language and procedure across jurisdictions and over time, which complicates consistent data extraction and variable definition. Institutional challenges involve incomplete or biased coverage of the legal text archive: many filings (especially in state courts or sensitive cases) are paywalled, sealed, or redacted, leading to non-random gaps in the data. Additionally, legal datasets must be continuously updated, since dockets evolve with new filings and outcomes.

Consequently, transparency in the data-generation process, rigorous error checking, and human expert oversight remain essential to ensure that legal text analyses are reliable. Economists venturing into this area must implement systematic quality controls and clearly document how raw texts are transformed into features of a dataset

Despite these challenges, the ability to turn court records into data represents a major opportunity for empirical research. Court documents comprise a rich but currently untapped reservoir of economic information. With the aid of modern AI tools and careful validation, this resource can be harnessed to significantly broaden the scope of economic analysis, shedding new light on how legal processes and disputes shape economic outcomes at large.

[i] *PACER: Federal Court Records*, U.S. COURTS, https://pacer.uscourts.gov/
[ii] Additional platforms to access court documents include*: Case.law*, HARVARD LAW SCHOOL LBRARY, https://case.law/, *Free Law Project*, LEGAL DATA RESOURCES, https://free.law/datasets/, *Stanford Securities Class Action Clearinghouse,* STANFORD LAW SCHOOL, https://securities.stanford.edu/
[iii] About Us, *PACER: Federal Court Records*, U.S. COURTS, https://pacer.uscourts.gov/about-us
[iv] *Courts' Response to Wyden Questions*, CYBERSECURITY & PRIVACY (CCH), https://business.cch.com/CybersecurityPrivacy/courtsresponsetowyden.pdf
[v] *PACER Facts*, FREE LAW PROJECT, https://free.law/pacer-facts/
[vi] *State Courts*, AMERICANS FOR FINANCIAL REFORM, https://afj.org/statecourts/
[vii] *How Many Cases—and What Kind—Do State and Local Courts Handle?*, THE PEW CHARITABLE TRUSTS (Mar. 6, 2025), https://www.pew.org/en/research-and-analysis/data-visualizations/2025/03/how-many-cases-and-what-kind-do-state-and-local-courts-handle.
[viii] *Immigration Courts: Decline in New Cases at the End of FY2024*, CONGRESS.GOV (CRS Product No. IN12463), https://www.congress.gov/crs-product/IN12463
[ix] *Docket Alarm*, https://www.docketalarm.com/
[x] LexisNexis, https://www.lexisnexis.com/en-us/gateway.page
[xi] Farid Ariai, Joel Mackenzie & Gianluca Demartini, *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges*, arXiv:2410.21306 (2024)
[xii] *Smith v. City of Jackson*, 544 U.S. 228 (2005), https://www.law.cornell.edu/supct/pdf/03-1160P.ZO
[xiii] *The Role of Justice in Development: The Data Revolution*, WORLD BANK, https://documents1.worldbank.org/curated/en/423061624976141321/pdf/The-Role-of-Justice-in-Development-The-Data-Revolution.pdf.
[xiv] Nicolás Magner, Pablo A. Henríquez & Aliro Sanhueza, *Decoding Risk Sentiment in 10-K Filings: Predictability for U.S. Stock Indices*, 81 *FIN. RES. LETT.* 107472 (2025)
[xv] Bryan Kelly, Dimitris Papanikolaou, Amit Seru & Matt Taddy, *Measuring Technological Innovation over the Long Run*, 3 *AM. ECON. REV.: INSIGHTS* 303 (2021)
[xvi] Tarek A. Hassan, Stephan Hollander, Aakash Kalyani, Laurence van Lent, Markus Schwedeler & Ahmed Tahoun, *Text as Data in Economic Analysis*, 39 *J. ECON. PERSPS.* 193 (2025).
[xvii] For example, open-source NLP libraries like LexNLP can pull structured facts (dates, dollar amounts, parties) from legal documents. LexNLP, for example, can automatically identify monetary figures, percentages, and dates in unstructured legal text. *lexpredict-lexnlp*, GITHUB, https://github.com/LexPredict/lexpredict-lexnlp
[xviii] Mark J. Davenport, *Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation*, 15 *OPEN J. MOD. LINGUISTICS* 232 (2025).
[xix] Joint Appendix, Lab. Corp. of Am. Holdings v. Davis, No. 22-55873 (9th Cir. filed 2025).
[xx] Declaration, FTC & State of California v. NGL Labs, LLC, No. 2:24-cv-5753 (C.D. Cal. filed 2024).
[xxi] Plaintiff's Original Petition, KISHA LLC v. Galveston Cent. Appraisal Dist., No. 24-CV-1732 (405th Jud. Dist. Ct., Galveston Cnty., Tex. filed Oct. 2, 2024).
[xxii] For example, Vacek et al. (2019) apply NLP to federal dockets and report extracting case outcomes for ~8 million past dockets. Thomas Vacek, Ronald Teo, Dezhao Song, Timothy Nugent, Conner Cowling & Frank Schilder, *Litigation Analytics: Case Outcomes Extracted from US Federal Court Dockets* (2019).
[xxiii] *Stanford Securities Class Action Clearinghouse*, STANFORD SECURITIES CLASS ACTION CLEARINGHOUSE, https://securities.stanford.edu/index.html
[xxiv] *Lex Machina Releases 2025 Patent Litigation Report*, LEXISNEXIS, https://www.lexisnexis.com/community/pressroom/b/news/posts/lex-machina-releases-2025-patent-litigation-report
[xxv] Zynga Inc. v. Harris*, Complaint* (Mar. 2020), COURTHOUSE NEWS SERVICE, https://www.courthousenews.com/wp-content/uploads/2020/03/Zynga.pdf
[xxvi] *Seven-Eleven Stores, Inc. Stipulated Judgment (Redacted)*, FED. TRADE COMM'N, https://www.ftc.gov/system/files/ftc_gov/pdf/c4641sevenelevenstipulatedjudgmentredacted.pdf
[xxvii] *DaVita Inc.'s Opposition to Exclude Expert Testimony* (Mar. 29, 2022), APPLIED ANTITRUST, https://appliedantitrust.com/03_criminal/case_studies/davita2021/davita_dcolo_exclude2_opp2022_03_29.pdf
[xxviii] *Supreme Court of the United States*, JUSTIA, https://supreme.justia.com/

[xxix] *Adarand Constructors, Inc. v. Pena*, JUSTIA, https://supreme.justia.com/cases/federal/us/472/585/
[xxx] *Adarand Constructors, Inc. v. Pena*, 472 U.S. 585 (1985), https://tile.loc.gov/storage-services/service/ll/usrep/usrep472/usrep472585/usrep472585.pdf
[xxxi] *Davis v. United States*, JUSTIA, https://supreme.justia.com/cases/federal/us/479/104/
[xxxii] *News Release and Data Release Schedule*, BUREAU OF ECONOMIC ANALYSIS, https://www.bea.gov/news/schedule
[xxxiii] *Judicial Business 2021: U.S. District Courts*, U.S. COURTS, https://www.uscourts.gov/data-news/reports/statistical-reports/judicial-business-united-states/judicial-business-2021/us-district-courts-judicial-business-2021
[xxxiv] *Lex Machina, 2024 Consumer Protection Litigation Report*, LEX MACHINA, https://lexmachina.com/lex-machina-releases-2024-consumer-protection-litigation-report
[xxxv] *File Formats*, OPEN DATA HANDBOOK, https://opendatahandbook.org/guide/en/appendices/file-formats/
[xxxvi] *Interoperable Legal AI for Access to Justice*, YALE L.J. FORUM, https://yalelawjournal.org/forum/interoperable-legal-ai-for-access-to-justice