Using Topic-Modeling in Legal History, with an Application to Pre-Industrial English Caselaw on Finance[*]

Peter Grajzl[†] and Peter Murrell[‡]

February 8, 2022

Abstract

We argue that topic-modeling, an unsupervised machine-learning technique for analysis of large corpora, can be a powerful tool for legal-historical research. We provide a non-technical introduction to topic-modeling driven by the presentation of an example of how researchers can use the data that topic-modeling produces. The context of the example is pre-industrial English caselaw on finance. We generate new insights on the timing of pertinent legal developments, the linkages of law on finance to other areas of law, and the relative importance of common-law and equity in the emergence of law and legal ideas relevant to finance. We argue that topic-modeling has the potential to bridge traditional legal history and economics, increasing the influence of the former on the latter, which is overdue. The output of topic-modeling includes the data required to generate a quantitative macroscopic overview of the flow of legal history. These data can be used in many ways in subsequent legal-historical research. Epistemologically, topic-modeling offers an escape from the temptations of Whig history and opens up new avenues for inductive analysis characteristic of traditional historical research.

Keywords: Topic-modeling, legal history, pre-industrial England, finance, caselaw, common-law, equity

JEL Classifications: C8, G20, K00, O17, N2, P10

[*]We thank two referees for their insightful comments and especially the editor, Gautham Rao, for his encouragement and insights. We are also grateful to Jonathan Gong for research assistance and participants at the AI4Law workshop in February 2021 at the University of Oxford for helpful comments.

[†]Department of Economics, The Williams School of Commerce, Economics, and Politics, Washington and Lee University, Lexington, VA 24450, USA and CESifo, Munich. Email: grajzlp@wlu.edu

[‡]Department of Economics, University of Maryland, College Park, MD 20742, USA. Email: pmurrell@umd.edu

I. Introduction

The last few decades have seen the ever-increasing importance of quantitative empirical methods in historical studies in general, and economic history in particular. However, these methods have made few inroads into pre-twentieth-century, and especially pre-industrial, legal history, despite the central place of law in the history of world economic development.[1] No doubt the relative absence of such quantitative legal history is because the legal record is mostly in words, the processing of which requires computational power that is orders of magnitude beyond that needed for numbers. However, with huge increases in computer power in recent years and the associated development of desktop text-analyzing software, the menu of research methods and results available to all legal historians is now rapidly changing. Text can be processed and analyzed as quickly and easily as numbers were two decades ago.[2] Libraries of readily usable computational packages are available for the statistical analysis of texts. We now have the possibility of using the text of centuries ago as data.[3]

The objective of the present paper is to convey to traditional legal historians the role that these new computational techniques can play in legal-historical research. We do so by presenting an example of the types of results that can be produced with these new tools.[4] As we present the example, we outline the steps that must be taken in the computational-statistical process. But our presentation does not require readers to be conversant with the intricacies of such methods. We provide verbal, intuitive descriptions of the methods used and the tasks that must be accomplished.

[1] S. Robertson, "Searching for Anglo-American Digital Legal History," *Law and History Review* 34 (2016): 1047-69, noting that "as the fields of digital humanities and digital history have grown in scale and visibility since the 1990s, legal history has largely remained on the margins of those fields." There are some important very recent examples for the nineteenth century, such as K. Funk and L.A. Mullen, "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice," *American Historical Review* 123 (2018): 132-64. In recent years, a number of empirical papers are appearing that use data from the eighteenth century made available by the *Old Bailey Proceedings* project. See T. Hitchcock, R. Shoemaker, C. Emsley, S. Howard and J.McLaughlin, "The Proceedings of the Old Bailey, 1674-1913", (www.oldbaileyonline.org). Both of these works rely on the types of computational advances that we highlight in this paper and that we feel will lead to a quiet revolution in legal historical studies.

Existing, more traditional studies on the period before the nineteenth century usually contain very small samples or few variables, implying that there is a limited ability to apply the types of empirical methods that are now commonplace in economic history. E. Cavell, "The Measure of Her Actions: A Quantitative Assessment of Anglo-Jewish Women's Litigation at the Exchequer of the Jews, 1219-81," *Law and History Review* 39 (2021): 135-72 provides a recent example of a very interesting exercise in early legal history that is, understandably, limited by a small sample with few variables. D. Klerman, "Settlement and the Decline of Private Prosecution in Thirteenth-Century England," *Law and History Review* 19 (2001): 1-65 is notable in providing a very early example of pre-industrial legal history that is exceptional for the centrality of empirical methods in its contribution.

[2] The general problem is usefully captured as "How do you write a national history that was the product of lawmaking in 50 separate jurisdictions?", as posited by E. Nystrom and D. Tanenhaus, "The Future of Digital Legal History: No Magic, No Silver Bullets," *American Journal of Legal History* 56 (2016): 150-67. This problem is multiplied in caselaw where one is studying hundreds of years and thousands of cases. The methods we describe in this paper almost completely remove the sample-size and limited-observations constraint referred to in the previous footnote. Notably, the general project that includes the current paper did not rely on any extramural funding, emphasizing that the techniques we describe are within the reach of all scholars.

[3] See, for example, J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21 (2013) 267-97; M. Gentzkow, B. Kelly, and M. Taddy, "Text as Data," *Journal of Economic Literature* 57 (2019): 535-74; and M.A. Livermore and D. N. Rockmore (ed.), *Law as Data: Computation, Text, and the Future of Legal Analysis* (Santa Fe: SFI Press, 2019).

[4] In legal history, two early examples of the use of the new sets of computational tools are provided by D. Tanenhaus and E. Nystrom, "Let's Change the Law: Arkansas and the Puzzle of Juvenile Justice Reform in the 1990s," *Law and History Review* 34 (2016): 957-97 and C. Romney, "Using Vector Space Models to Understand the Circulation of Habeas Corpus in Hawai'i, 1852-92," *Law and History Review* 34 (2016): 999-1026. In contrast to the exercise reported in this paper, these two examples do not use the computational methods to drive an empirical exercise but rather use these methods as search procedures to find those legal materials on which a more traditional analysis should be focused.

Our view is that what we offer in this paper will be instructive for scholars currently using traditional legal-historical approaches: If the past two decades of methodological developments in the humanities, the social sciences, and law teach us anything, it is more or less inevitable that the new computational methods will become a part of the toolkit of legal history.

Importantly, we do not argue that the new computational approach will replace existing methods. In fact, we do the opposite. As we present our example, we detail many instances where our use of the existing work of traditional legal historians has played an absolutely vital role in our ability to produce any novel insights from our application of the new tools. Thus, through the use of example, we hope to show how traditional and computational legal history can complement each other as the field of legal history moves into the new age in which the use of computational methods will become standard. In doing so, we are also able to pinpoint where each of the traditional and computational approaches to legal history has their comparative advantage.

To make this paper more accessible to those unfamiliar with any of these new methods, we focus on only one, topic-modeling, which indeed is one of the most popular machine-learning techniques that has been applied in history, law, and social science.[5] Concentrating on one method allows us to focus on the essential characteristics of machine-learning and to discuss them in intuitive, non-technical ways, addressing our exposition not to those who want to learn the details of the computational analysis but rather to those who want to understand the types of insights that computational-text-analysis can bring to substantive domain-specific research. As Grimmer, Roberts, and Stewart argue, "machine learning is as much a culture defined by a distinct set of values and tools as it is a set of algorithms."[6]

This is a paper written by economists. One additional impetus underlying the writing of this paper followed from our observation that the mainstream economics literature has tended to ignore the insights of the historians of caselaw, while traditional caselaw historians hardly refer to the methods and findings of economists.[7] Perhaps this is because economists are more moved by

[5] On the popularity of topic-modeling, see Gentzkow et al., "Text as Data" and J. Guldi and B. Williams "Synthesis and Large-Scale Textual Corpora: A Nested Topic Model of Britain's Debates over Landed Property in the Nineteenth Century," *Current Research in Digital History* 1 (2018).

Several other machine-learning and related computational approaches have been utilized to investigate law-as-data. Machine-learning methods have been used, for example, to predict court outcomes; see e.g., D.M. Katz, M.J. Bommarito, and J. Blackman, "A General Approach for Predicting the Behavior of the Supreme Court of the United States," *PLoS ONE* 12 (2017): e0174698. Word and document embedding models represent words and documents as numerical scores for a long list of variables, thereby helping to quantify the meaning of words and documents on the basis of their proximity to other words and documents in the corpus; see e.g. E. Ash and D.L. Chen, "Case Vectors: Spatial Representations of the Law Using Document Embeddings," in *Law as Data*, ed. M.A. Livermore and D.N. Rockmore (Santa Fe: SFI Press, 2019), 313-37. Embedding approaches have been employed, for example, to investigate the presence of racial bias in judicial opinions; see e.g. D. Rice, J.H. Rhodes, and T. Nteta, "Racial Bias in Legal Language," *Research & Politics* April-June (2019). For an overview of the use of machine-learning and computational methods in the emerging research field of computational analysis of law-as-data, see J. Frankenreiter and M.A. Livermore, "Computational Methods in Legal Analysis," *Annual Review of Law and Social Science* 16 (2020): 39-57. For innovative applications of computational methods to legal-historical themes, but not focusing on English caselaw, see e.g. S. Klingenstein, T. Hitchcock, and S. DeDeo, "The Civilizing Process in London's Old Bailey," *Proceedings of the National Academy of Sciences* 111 (2014): 9419-24 and Funk and Mullen, "The Spine of American Law".

[6] J. Grimmer, M. E. Roberts, and B. Stewart, "Machine Learning for Social Science: An Agnostic Approach", *Annual Review of Political Science* 24 (2021): 395-419.

[7] R. Harris, "The Encounters of Economic History and Legal History," *Law and History Review* 21 (2003): 297-346 identified this separation of these fields and his conclusions still seem to apply today.

quantitative evidence, which is not easily found in the history of caselaw. This paper is an attempt to straddle the two fields, to show that there can be strong complementarities between them.

To illustrate the power of topic-modeling for legal history, we provide new quantitative information on developments in English caselaw and legal ideas from the mid-16th century to the mid-18th century. Thus, central to our approach in this paper is showing the usefulness of the computational methods by providing an example of their application to ongoing debates in legal history. In contrast to much existing work in digital history, we do not argue for the productiveness of the computational methods by focusing on the methods themselves. Rather, we endeavor to make the case by providing an example of the contribution of the methods to an understanding of the past that is directly relevant to the disciplines of legal history and economics.[8]

This era of English law has been of particular interest to both legal historians and economists, for related reasons: for the former because much law relevant to the modern era was created then; for the latter because of the possible connection between legal developments and the rise of Britain as the first industrial power. In particular, the progress of the financial sector in the decades preceding the industrial revolution has received much attention in economics. However, the work of economists on pre-industrial finance has placed little emphasis on caselaw, which for many is the defining characteristic of the English legal family. We show how topic-modeling can use the caselaw record to cast new light on the patterns and sources of finance-related legal developments in England from the middle of the 16th century to the industrial revolution. In doing so, we find invaluable the accumulated insights of legal historians, echoing the views of users of topic models in other fields who emphasize how the traditional 'close' reading of texts must be used alongside the 'distant' reading provided by machine-learning.[9] The outputs generated on the basis of the new methods are the complements of traditional legal-historical research. The results from topic-modeling are not replacements for the detailed, and immensely valuable, contextual analysis of

[8] On these points more generally, see S. Robertson and L. Mullen, "Arguing with Digital History: Patterns of Historical Interpretation," *Journal of Social History* 54 (2021): 1005-22, who argue that "Digital history has only rarely contributed interpretative or argumentative scholarship that contributes to disciplinary understandings of the past", largely because of its focus on the methodological. Beyond the example appearing here, the use of the methods outlined in this paper and of the dataset discussed here are provided in several additional papers that contribute to disciplinary understandings of the past: See P. Grajzl and P. Murrell, "A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution II: Applications," *Journal of Institutional Economics* 17 (2021): 201-16; P. Grajzl and P. Murrell, "A Macrohistory of Legal Evolution and Coevolution: Property, Procedure, and Contract in Pre-Industrial English Caselaw" (https://dx.doi.org/10.2139/ssrn.4005612); and P. Grajzl and P. Murrell "Of Families and Inheritance: Law and Development in Pre-Industrial England" (https://dx.doi.org/10.2139/ssrn.3975015).

[9] The useful distinction between close and distant reading arose among scholars of literature in what has become known as the digital humanities, where debates about the usefulness of computational methods, particularly topic-modeling, were both early and very spirited. For the digital humanities, see, for example, A. Goldstone and T. Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 5 (2014): 359-84. For history, see the very early study by S. Block "Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources," *Commonplace* 6 (2006); and more recently J. Guldi, "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora," *Journal of Cultural Analytics* 3 (2018). For the same emphasis in political science, see Grimmer and Stewart, "Text as Data" and, in a joint product of a sociologist and two computer scientists, P. DiMaggio, M. Nag, D. Blei, "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics* 41 (2013): 570-606. For legal history, see Robertson, "Searching".

traditional legal historians, but instead simply offer a different sort of lens for studying legal-historical phenomena.

The focus is on the use of the quantitative output produced by one existing topic-modeling exercise, that of Grajzl and Murrell, henceforth referred to as GM.[10,11] By building on an existing implementation of topic-modeling, this paper can omit descriptions of the technical nuances and the details of data construction, making it accessible to a wider audience. We do, however, provide an intuitive description of the methods used to generate the raw quantitative output of the topic model, a description that is intended to be accessible to those not versed in the details of computational-statistical modeling.

We then use that intuitive description of topic modeling to describe its data outputs. Importantly, the outputs of a topic model are not the endpoint of such an exercise. Rather, they constitute data that can be productively employed as an input into subsequent analyses. Thus, we turn to examples of the substantive insights that can be generated from the dataset produced by the machine learning. We highlight the types of information that can be generated and made readily available to other scholars. That information can be easily used by those who have no intention of implementing the methods themselves but rather are interested in the types of substantive results that can be generated by the data that is the output of a topic model.

Section II presents the informal overview of topic-modeling. It begins with a brief history of how this tool has been used in the humanities, law, and the social sciences, showing that the particular exercise that this paper presents is a natural outgrowth of two decades of development and application of topic-modeling. This short history argues that topic-modeling should not be regarded as immediately alien to legal history in view of the fact that it has been applied in fields whose objects of study share many features with the history of the law.

Then, we proceed with a non-technical discussion of the assumptions, methods, and outputs of topic-modeling. This informal overview has the advantage that it lays bare the types of assumptions about texts that machine-learning uses, so that the weaknesses of the new approaches can be clearly seen.

The raw data used for the topic model discussed here are virtually all reports on cases heard before 1765 that appear in the *English Reports*, a corpus comprising 52,949 reports.[12] Topic-

[10] See P. Grajzl and P. Murrell, "A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution I: Generating and Interpreting the Estimates," *Journal of Institutional Economics* 17 (2021): 1-19; and P. Grajzl and P. Murrell, "A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution II: Applications," *Journal of Institutional Economics* 17 (2021): 201-16.

[11] The work on English case reports is part of a much larger project on using computational and statistical techniques to understand English history. The earliest products of this project combined legal history and intellectual history, with two papers addressed to understanding more general sets of ideas, focused on Francis Bacon and Edward Coke. See P. Grajzl and P. Murrell, "Toward Understanding 17th Century English Culture: A Structural Topic Model of Francis Bacon's Ideas," *Journal of Comparative Economics* 47 (2019) 111-35; and P. Grajzl and P. Murrell, "Characterizing a Legal-Intellectual Culture: Bacon, Coke, and Seventeenth-Century England," *Cliometrica* 15 (2021): 43-88.

[12] The digitized copies of the *English Reports* were purchased from a publishing company domiciled in South Africa. It is beyond the scope of this paper to provide the many details of the initial processing of these digital copies, and the cleaning of them. Suffice it to say that a very large proportion of GM's labor time devoted to the pertinent research projects was consumed in all of these

modeling produces parsimonious summaries of this enormous amount of text information, which comprises 31,057,596 words. Because topic-modeling is an unsupervised machine-learning technique, the shape of the summaries themselves is not produced in order to answer a particular question or to test a particular hypothesis. Rather, the text-data themselves shape their own synopsis.

The summaries are in the form of 100 'topics', as if the computational methods had produced a new digest of English law, divided into 100 sections. This is the 'dimensionality reduction' aspect of machine-learning, producing an organized summary of an enormous amount of text that no human being could possibly hope to read (or at least retain and organize in memory).[13] In this respect, topic-modeling dovetails with one of the central concerns of historians, to provide compelling narratives. The computer is essential to the production of the narrative because so much information is captured and condensed. This is especially the case when the attempt is to capture ebbs and flows over centuries: Guldi and Armitage emphasize the potential in big data to return historical studies to the longue durée.[14]

In the case of topic-modeling the computer output itself is only the beginning and much interpretation is needed. The sections of the digest come without names—one just knows which case reports feature a particular digest section most prominently and which vocabulary that section most favors. The detailed work of legal historians over the centuries then provides the background for analysis of this information, enabling the researcher to understand which areas of law a particular section of the digest contains, thereby driving the crucial step of topic naming. Close reading undertaken in the context of traditional legal history is essential to interpret the output of the computer's distant reading.

Importantly, the naming of topics can be done by any researcher who has obtained the data produced by the topic model: it is undertaken quite separately from the computational analysis. This possibility for sharing of the generated data is one of the most important contributions that the new methods can offer to legal history and legal-historical research: the results that are the output of the topic-modeling exercise can be made available and used as inputs by all researchers.

Once the underlying nature of each topic is understood, the researcher can then proceed to analyze the vast amount of quantitative information produced by the computational methods. This

tasks. For more details, see GM and M. Schmidt, *Institutional Persistence and Change in England's Common Law: 1700-1865* (Ph.D. dissertation, University of Maryland, 2015).

Because a central objective of GM was to include as many reports as possible, which necessarily implied computational processing of all reports, there was a need to exclude a small percentage of reports with too many words that did not have a counterpart in either modern English or standard Latin. Chiefly, this had the effect of excluding reports in Law French. There is no doubt that this is a blemish on the application of the computational methods. Initially, there were 60,249 pre-1765 reports in the data set, but 6,917 were dropped because they were in Law French and a further 383 were removed because they contained too many unrecognizable words. This left 52,949 reports.

[13] The summary does not rely on any existing classifications: we return to this point in the conclusion.

[14] See J. Guldi and D. Armitage, *The History Manifesto* (Cambridge: Cambridge University Press, 2014), emphasizing that "Over the last decade, the emergence of the digital humanities as a field has meant that a range of tools are within the grasp of anyone, scholar or citizen, who wants to try their hand at making sense of long stretches of time. Topic modeling software can machine read through millions of government or scientific reports and give back some basic facts about how our interest in ideas have changed over decades and centuries."

paper provides an example of how such information can be used: the present paper's input data is the output data of GM and we use those data to present new results and provide insights that any readers could have produced had they availed themselves of the same data.

Each of Sections IV, V, and VI is built around just one just evocative figure intended to provide an interpretation of the development of caselaw and legal ideas relevant to finance in pre-industrial England. The origins of our interest in these developments lies in our background as economists. In Section III, we review the debates that have made the history of English law on finance important in that discipline, and explain how the combination of machine-learning methods and the prior insights of legal historians offers new information pertinent to those debates. We ask and answer the following questions: Which time periods evidence the most intense development of that area of law and legal ideas? Which pre-existing elements of law, such as property or contract, were most important as inputs into this development, and when? What were the relative roles of common-law and equity in spurring these developments?

Section IV introduces the 15 of the 100 GM-estimated topics that are most relevant to finance, the most salient sections of the machine-produced digest. These 15 topics were identified by the authors on the basis of topic content, and therefore the overall category of finance is not an entity produced by the topic modeling itself. This is just one of the many examples we provide in this paper of the fact that the modeling of the topics themselves is not the endpoint of the analysis, but rather provides the data that the researcher uses in combination with existing information and judgment to proceed to real substance.

The periods of the most intense development of the relevant caselaw become evident by examining timelines that show when these 15 topics are most prevalent in the *English Reports*. For example, the timelines show what will be very familiar to legal historians, that ideas on assumpsit developed in the early 17th century. But the timelines can add to these insights by demonstrating that attention to assumpsit peaked around 1630, while the development of ideas on the validity of contracts, for example, was largely a product of the 1690's and later. Cumulatively, our timelines of the finance-related areas of law suggest that the 17th century witnessed many advances in caselaw that became relevant to 18th-century finance.

Section V considers connections between the developments that take place in differing areas of law. Any case report usually incorporates ideas from varied legal domains even if one specific issue is central to the case: a single case is indexed within many sections of the digest. The topic-modeling produces data on the proportion of each of the 100 estimated topics that is present in each of the 52,949 reports of cases. Thus, one can find, for example, whether a case that is very much centered on trusts tends to emphasize contract issues or property considerations. By examining such connections in general, one can make conclusions about the legal ideas in one domain that were relevant to, and possibly fed into, the legal ideas in another domain. In Section V, we identify the links among the 15 topics (the digest sections) identified with finance, as well as links between these 15 topics and ones not classified within finance. We find, for example, that early-17th-century developments in the caselaw of contracts had significant effects on later developments in caselaw relevant to finance.

Section VI examines the relative importance of common-law and equity in producing law relevant to finance. Although case reports are unambiguously assignable to courts and although specific legal notions were often the particular province of either common-law or equity, each type of court absorbed ideas from the other. For example, a case on trusts in Chancery (an equity court) could well use ideas on contract developed in Common Pleas or King's Bench (common-law courts). The development of ideas in a given legal domain can then be ultimately viewed as reflecting debates in both common-law and equity. We examine the relative importance of law and equity for each topic related to finance. Interestingly, our evidence shows that many of the critical areas of law on finance were a product of equity, and not of the common-law. To state the implied conclusion in its most contentious form, Britain might never have been economically powerful enough to spread its common-law around the world had it relied solely on the common-law at the time that it began spreading its system of law around the word.

Section VII concludes, providing reflections on both the promise of computational text analysis for legal history and its pitfalls. We comment on what topic-modeling can and cannot do. We emphasize that topic-modeling can provide new sources of data for other researchers: once a massive volume of texts are summarized, the quantitative summaries themselves can provide inputs into further research. Peering into the future, one can detect signs that unsupervised machine-learning might be gradually changing the research perspectives of social science, with descriptive analyses now becoming more acceptable. The almost exclusive emphasis on the hypothetico-deductive method is waning (very slightly at the moment) and exercises in the inductive spirit are gaining credibility. This change would naturally lead to much more complementarity between traditional legal historians and those who favor the use of computational and statistical methods in the social sciences.

II. An introduction to topic-modeling

The techniques that we describe here are descendants of the seminal paper by Blei, Ng, and Jordan,[15] particularly the structural topic model by Roberts, Stewart, and Airoldi, which is the

[15] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993-1022. One measure of the prominence of this contribution is that this is the 7th most cited article in computer science that was produced this millennium (https://citeseerx.ist.psu.edu/stats/articles). To be sure, there were a number of similar algorithms developed before the Blei et al. contribution, but early excitement about these methods seems to have focused on Blei et al., perhaps because of the accessible software developed for implementation. See A.K. McCallum, "MALLET: A Machine Learning for Language Toolkit," (http://mallet.cs.umass.edu). In their note explaining this software, S. Graham, S. Weingart, and I. Milligan, "Getting Started with Topic Modeling and MALLET," (https://programminghistorian.org/en/lessons/topic-modeling-and-mallet) state: "You will sometimes come across the term 'LDA' when looking into the bibliography of topic modeling. LDA and Topic Model are often used synonymously, but the LDA technique is actually a special case of topic modeling created by David Blei and friends…. It was not the first technique now considered topic modeling, but it is by far the most popular…They all work in much the same way." One such earlier algorithm was used in study by Newman and Block in the first history publication to use topic-modeling. See D. J. Newman and S. Block, "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper," *Journal of the American Society for Information Science and Technology* 57 (2006): 753-67.

version of topic-modeling used by GM to produce their results.[16] Topic-modeling originated in computer science, in pursuit of using computational methods to summarize large amounts of text information. Within the social sciences and humanities, the field in which topic-modeling first flourished was the digital humanities, particularly literature, obviously a field for which text is central. In that discipline, the rise in popularity was probably fueled by the rhetoric of the assertions of the advantages of distant reading over traditional close reading, and ensuing debates. Text, rather than numbers, providing much of the core data in politics, political science was the next major discipline to see the advantages of the new machine-learning approaches, particularly topic-modeling. It is much more difficult to find applications in political theory, which is perhaps the closest analog in political science to caselaw.[17] Political science was naturally followed by law, also presumably due to the fact that much of its data are texts, but legal history has been slow to follow. Digital humanities, political science, and law seem to be the three major non-computational-science disciplines where applications using topic-modeling, and related techniques, appear regularly in the top journals and are cited regularly within the mainstream of the field.

Economics and history, particularly legal history, the disciplines reflected in this paper, are ones where the application of topic models has lagged. In economics, this is readily explained by the enormous influence of the hypothetico-deductive paradigm, with its emphasis on the testing of hypotheses concerning isolated causal facts rather than an interest in broad narrative.[18] The uses of topic-modeling in economics most usually focus on new measurements of highly specific phenomena, to fit into a particular implementation of that paradigm.[19] Our use of topic-modeling is therefore rather different from the few applications in the mainstream of our field: our objective is to provide a broad narrative of finance-related English caselaw over two centuries. To the extent that we match our data to specific hypotheses, it is because we came to realize after the construction of our narrative how our narrative naturally reflected on these hypotheses, not because we aimed originally to test them.

The reason for history's lag in applying machine-learning in general, and topic-modeling in particular, is less clear, to us at least.[20] As already mentioned, topic-modeling leads naturally to a

[16] See M.E. Roberts, B.M. Stewart, and E.M. Airoldi, "A Model of Text for Experimentation in the Social Sciences," *Journal of the American Statistical Association* 111 (2016): 988-1003, whose general approach is very similar to that of Blei et al. but has an emphasis on incorporating document meta-information (such as date of publication) directly into the analysis. Small details would have changed had we used LDA, but we are sure the overall picture would have remained the same. For copious detail on the structural-topic-model approach to topic-modeling, including how to get started on implementation, see https://www.structuraltopicmodel.com/.

[17] On this point, see H. Bonin, "From Antagonist to Protagonist: 'Democracy' and 'people' in British Parliamentary Debates, 1775-1885," *Digital Scholarship in the Humanities* 35 (2020): 759-775. One example using a topic-model-type method is L. Blaydes, J. Grimmer, and A. McQueen, "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds," *Journal of Politics* 80 (2018): 1150-67.

[18] Applications in sociology have also lagged, perhaps because the hypothetico-deductive method has had increasing sway in that field as well. For the lag in sociology, see N.C. Lindstedt, "Structural Topic Modeling for Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005-2017," *Social Currents* 6 (2019): 307-18.

[19] For example, S. Hansen and M. McMahon, "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication," *Journal of International Economics* 99 (2016): S114-S133.

[20] Stephen Robertson emphasizes that text-analysis in history has been held back simply by the availability of a large stock of digital texts. S. Robertson, "The Differences between Digital Humanities and Digital History," *Debates in Digital Humanities*

historical narrative. But Guldi and Armitage, who emphasize this point also, argue that research in history has turned away from exercises that examine long time periods and expansive subjects, exactly the areas in which machine-learning can contribute. The high degree of technical complexity in existing applications of topic-modeling to history might also have discouraged some researchers.[21]

However, as we hope to show in this paper, researchers interested in using the output of topic models do not themselves have to engage in all the complexities of producing topic model estimates. If that output is freely available to all, as is the case with GM, it is enough for subsequent researchers to understand how to interpret that output when using it as a source of data as a basis for further exploration. An analogy is helpful here. Economic historians using estimates of national income are not required to produce those estimates themselves, nor even to grasp all the complexities of data gathering and index number construction. As we show below, by example, the output data of topic-modeling can be used as input data for further exercises in an exactly analogous way.

A. The Topic Model

The algorithms producing topic-modeling estimates begin with a conceptualization of the process of document (in our context, case report) generation that is extremely crude but lends itself to formalization in a statistical model. It is the explicitness of the conceptualization that facilitates interpretation of the results of the analysis, producing the insights that legal historians might appreciate. Such an interpretation is often not possible with the results of other machine-learning techniques, such as neural networks, in which the focus is on prediction or problem-solving, rather than description. But the relative ease of interpretation comes with a cost: the simple conceptualization will surely foster a general skepticism.[22] We give an unvarnished view here to emphasize limitations, and why they arise.

The process of generating case reports envisaged by topic-modeling may be summarized as follows. An author (in our context, a legal reporter) is viewed as beginning with a fixed number of topics, essentially lodged in his or her brain and available for use when writing. Topics might be well-identified legal concepts, such as assumpsit or habeas corpus, or ideas that cut across many domains of law, such as revocation, or even a particular reporting style.[23] When a particular topic

(2016). This constraint is rapidly being relaxed. Indeed, one of the contributions of GM is to make machine readable, cleaned versions of the English Reports available for scholars in general. See GM and the concluding section of this article for more details.

[21] For interesting articles of this kind, see A. Barron, J. Huanga, R. Spang, and S. DeDeo, "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences* 115 (2018): 4607-12; and A. Rule, J. Cointet, and P. Bearman, "Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790-2014," *Proceedings of the National Academy of Sciences* 112 (2015): 10837-44.

[22] As counterpoint to this apology for simplification, see S. Robertson, "Digital Humanities" in *The Oxford Handbook of Law and Humanities* ed. S. Stern, M. Del Mar, and B. Meyler, (Oxford: Oxford University Press, 2019), emphasizing that "If humanities scholars chafe at such simplification, it is worth noting that narrative, the favored representational model of humanities scholars, is a deliberately simplified account that is illuminating because of, not despite, its simplification."

[23] The productive use of machine-learning to detect style was emphasized by Matthew L. Jockers, one of the most forceful advocates of machine-learning in the digital humanities; see M.L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013).

is used, the author simply has a greater preference for the vocabulary more closely associated with that topic than for other words. For example, when the author refers to the topic assumpsit, the author will have a greater likelihood of using the word promise; similarly mention of bail will be frequent when using the topic habeas corpus. The production of a document, a case report, then entails the author choosing to emphasize some topics less and some more, depending on the general context of that report. A document will be a mixture of topics. Thus, a particular case report might tend to emphasize, for example, both assumpsit and habeas corpus because the defendant was in debtor's prison as a result of a case involving non-payment of a contractual debt. The words promise and bail would then appear prominently in this case report, but words such as daughter or wife would hardly appear because they are associated with topics that emphasize estates or wills, which are of no relevance for these particular types of cases.

Thus, a topic model will view a document as one where the author has chosen to emphasize certain topics, which in turn emphasize their own characteristic vocabularies. Consistently, a document is fed into the statistical analysis as a bag of words that has been stripped of all syntactic and sentence structure. However, each word choice is based on the emphasized topics and the vocabulary emphasized by these topics. This conceptualization then views semantic content as becoming embedded in a report because word choices will be highly correlated across reports. For example, contract and promise will appear frequently together, but their presence will be negatively correlated with the appearance of daughter and will, which in turn frequently co-occur. The topic-modeling algorithm produces results that reflect semantic content because it leverages these patterns of correlations across documents. In the phrasing of Mohr and Bogdanov, "relationality trumps syntax."[24] Similarly, topic-modeling is able to 'see' through polysemy because meanings are embodied in combinations of word usage not in single words.[25] The model will reflect the sense of 'extent' in 'he is not bound to prove the whole extent of a debt' very differently from the one in 'the Crown may not proceed against its debtor either by extent or scire facias' because of the repetition of the accompanying words across many cases.

The bag-of-words assumption is obviously a stylization that does no justice to the process of writing. One should note, however, that this assumption is partially a consequence of current limitations in computational power. With expected increases in computational power, much more acceptable characterizations of the process of authoring a document will become available when using techniques that are descendants of the ones described here.[26]

One final step in the conceptualization of the document-writing process is to acknowledge that different authors of case reports have different characteristics, and indeed the same author will be influenced by circumstances such as the timing of the case and the court adjudicating it. This can be explicitly incorporated into the estimation process when using the structural topic model.

[24] J. W. Mohr and P. Bogdanov, "Introduction–Topic Models: What They Are and Why They Matter," *Poetics* 41 (2013): 545-69.

[25] DiMaggio et al., "Exploiting Affinities".

[26] One could instead view documents as collections of two- or three-word chunks, or even larger phrases. But the required processing-power increases proportionately with the number of distinct phrases, which increases exponentially with the number of words allowed to be in a phrase.

In the application reported in this paper, the author of a specific case report is viewed as influenced by the year in which the report is written and the court in which the case is heard.

At this stage, we would imagine that readers unversed in topic-modeling, and machine-learning methods more broadly, are immensely skeptical. We were too when we began using such techniques. But after several years poring over results, comparing those results to existing ideas in the literature, and seeing the value added of insights that were not possible to reach when utilizing conventional approaches to analysis of texts, we saw how topic-modeling can provide a powerful complement to the traditional work of historians. Moreover, "legal history is better positioned for a digital turn than most historical fields when it comes to the amenability of legal sources to computational analysis" because reporters followed consistent forms of presentation using specialized vocabulary, where the correspondence between words and meaning remained much more constant both over time and between individuals than would have been the case for ordinary language.[27]

B. Estimation

Estimation begins with the observations that are available to the researcher: the documents (case reports) and the information that characterize authors. The researcher must decide on the number of topics, that is, the number of sections of the digest. Using statistical criteria and a more subjective evaluation of the coherence and meaning of the produced topics, GM judged that 100 different topics adequately captured the salient emphases in the reports on pre-1765 cases.[28] This element of human judgment is part of the process of validating the overall topic model: "Researchers must also interpret the topic model output, probably iteratively, so that a best fit can be found between the number of topics and an overall level of interpretability."[29]

Given the estimated topics, machine-learning provides a measure of the importance of each vocabulary word to each topic. In the current example, this is the proportion of each of the 41,174 distinct vocabulary words in each of the 100 topics. The estimation also predicts the proportion of any given one of the 52,949 documents that can be attributed to the use of each topic. And given that each document is labeled as reporting on a case heard at a specific time in a specific court, the estimates provide information on how the use of various topics varies with those characteristics, year or court.

C. What Are the Topics?

Topic-modeling is an unsupervised machine-learning exercise: the estimation of the topics is not guided by any objective to match topics to pre-existing ideas about what is in the law. Thus,

[27] See Robertson, "Searching". On this point, see also P. Grajzl and P. Murrell, "Lasting Legal Legacies: Early English Legal Ideas and Later Caselaw Development During the Industrial Revolution," *Review of Law & Economics* (2022) (in press).

[28] As evidenced by the large, related literature, computational scientists and statisticians usually emphasize rule-based criteria for model choice, relying solely on numerical information derived from the estimating process or the output data. In contrast, practitioners emphasize the element of subjective judgment, which would take into account the perceived quality of the topics reflecting both the uses to which they are to be put and the nature of the text data that is used in estimation. See, for example, Gentzkow et al., "Text as Data"; DiMaggio et al., " Exploiting Affinities"; and Mohr and Bogdanov, "Introduction–Topic Models".

[29] See Mohr and Bogdanov, "Introduction–Topic Models", 560.

the produced objects, the topics, come unlabeled: the researcher must provide titles for the sections of the machine-produced digest. This requires applying insights from existing legal-historical research. The information described in the previous paragraph is matched against those insights: one examines closely the vocabulary most used by a topic and one closely reads those case reports in which the topic is most prominent. This is an extremely laborious task, but GM found that it was not conceptually difficult to identify the idea or ideas underlying each and every one of their 100 topics.

One important part of the general methodology to emphasize here is that the identification of what a topic refers to cannot rely solely on a perusal of the vocabulary words that a topic most uses, even those words that a topic most uses relative to other topics. It is absolutely essential to read the documents in which a topic is most prominent. The labeling of a topic must make sense in relation to the content of all the other estimated topics because the specific emphasis in one topic might only be clear when contrasting that topic to a closely related one with a different emphasis. The reason to highlight this point is that many, probably a large majority, of the papers that have used topic-modeling to date base the interpretation of topics only on perusal of the words that a topic most uses. A reading of the documents requires domain-specific knowledge, and in the case of pre-industrial English history, it certainly requires struggling with a very different form of English. That is one reason why we emphasize that modern machine-learning and traditional doctrinal text analyses are complements.

This painstaking naming process is an essential ingredient of the validation of a topic model exercise—simply making sure that its results provide a coherent whole, both within topics and across topics.[30] The relative ease, in the conceptual sense, of topic naming in GM does suggest validity to the whole topic-modeling exercise. If many topics were simply mysterious, then one would harbor doubts that the specific features of the machine-learning process were not well suited to the texts being analyzed.

Some of GM's topics fit snuggly within existing concepts in the legal, historical, and traditional text-analysis literature. For example, Assumpsit, Bankruptcy, and Uses resonate closely with legal concepts and instruments covered at length in textbooks on the history of English law.[31] Other sets of topics split a single broad subject into several constituent areas (e.g. Implementing Ambiguous Wills, Contingency in Wills, Validity of Wills). Yet further types of topics encompass substantive issues that cut across many substantive areas of law (e.g., Revocation, Determining Damages & Costs) or refer to general legal ideas and modes of reasoning about cases as opposed to specific domains of application (e.g. Coke Reporting).[32] This an example of topic-modeling as

[30] See Grimmer et al., "Machine Learning for Social Science", stating: "Rather than place our trust fully in models and fit statistics, we argue that human feedback is essential for judging the quality of model results used for discovery."

[31] In order to distinguish our topic names clearly in the remainder of this paper, we capitalize them.

[32] An additional type of topic is identified by DiMaggio et al., "Exploiting Affinities", who argue that "Topic models often shunt noisy data into uninterpretable topics in ways that strengthen the coherence of topics that remain." In fact, our experience is not that the topics are uninterpretable, per se, but rather the interpretation means that the topic tells one nothing about the substantive inquiry in question. For example, GM find a topic that they call Non-Translated Latin. Sixteenth and seventeenth century lawyers

an exercise in discovery, rather than an exercise in prediction or hypothesis testing, which would instead be focused on a search for anticipated patterns in caselaw or legal ideas.[33]

When economists name topics in such an analysis, there is undoubtedly a tendency to focus on the functional domain to which the law is applied. A legal scholar would probably focus more on the legal doctrines captured in a topic and the historical origins of those doctrines. We are therefore sure that legal historians would have chosen a slightly different set of names than GM did for at least a subset of the 100 topics, probably finding labels that resonate more with internal characteristics of the legal system and less with outward effects on economic agents.[34]

The fact that the list of topic names, the titles of the sections of the digest of pre-1765 English caselaw and associated legal ideas, only partially matches the chapter and section headings of a legal history textbook can be viewed as either a vice or a virtue, depending on the reader's perspective. It might be troubling for some readers to look at a topic like Geographic Jurisdiction of Laws and realize that this topic is prominent in case reports that deal with such divergent areas of law as the relations between parishes and the legal status of individual citizens of belligerent nations. This topic appears prominently in case reports from the whole time period covered by our data and it appears in cases heard in all of the major courts. Thus, some areas of emphasis suggested by topic-modeling do not fit comfortably within existing classifications based on more traditional techniques. But this, in fact, shows the power of these machine-learning methods, highlighting how legal ideas can appear in many different types of cases. By covering the gamut of case reports in a particular time period, topic-modeling is an exercise in discovery, unearthing substantive patterns and connections between seemingly disparate notions that would likely remain unnoticed with the use of traditional methods restricted by the limits of human memory and reason. We return to this point in the conclusion, where we comment on how machine-learning is changing the research practices in several fields, lessening the hold of the hypothetico-deductive method and opening up possibilities for inductive exercises.

III. The Elements of English Legal History Emphasized in Mainstream Economics

In this section, we explain why we, as economists, chose to focus on the law relevant to finance in articulating the properties, promise, and pitfalls of topic-modeling. Ideas about the history of the law have made a difference in economics. Some of the conventional wisdom that drives important areas of mainstream economics reflects on subjects that are of great interest to those legal historians studying developments before the 20th century. However, there appears to

not only had their own version of English, their Latin was also highly idiosyncratic. The text preparation procedures were able to handle idiosyncratic English and standard Latin, but not idiosyncratic Latin.

[33] This point is much emphasized in the literature, something to which we return more fully in the conclusion. See, for example, DiMaggio et al., "Exploiting Affinities"; Mohr and Bogdanov, "Introduction–Topic Models"; A. Goldberg, "In Defense of Forensic Social Science," *Big Data & Society* (2015); L.K. Nelson, "Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South," *Poetics* 88 (2021): 101539.

[34] Since the topic-modeling algorithm produces unlabeled topics and since the data output from that algorithm can easily be transmitted, other researchers could easily produce their own set of names for the 100 topics. Indeed, doing so could inspire much more research that benefits from topic models. One research team produces the output data from the topic model, which can then easily be the input data to the work of other researchers.

have been little cross-fertilization between the literatures of the two fields, certainly as far as those literatures focus on the caselaw of the pre-industrial era.[35] Hence, the specific ideas embraced by mainstream economists do not always match the legal history that has been developed by those researchers whose primary audience is legal scholars and who approach the study of legal history with traditional text-analysis methods.

Finance is not normally a category or an immediate domain of interest within the legal-history literature of the pre-industrial era.[36] However, this area of law is vital to economic history because England's financial revolution preceded, and was perhaps a key input into, the industrial revolution. Ideas about English legal history have been influential in areas of economics as diverse as the regulation of modern financial markets, protection of investor rights, and the relief of poverty in the poorest countries. This is no doubt due largely to the global influence of Britain from the 18th century on, the importance of the British financial and industrial revolutions, and the spread of the common-law around the globe. It is also certainly due to the fact that understanding the sources of economic development is often considered the most important question of economics, and Britain led the world in political and economic development for more than two centuries.

Our focus here is on the two most influential strains of thought that are driven by interpretations of English legal history and that have had wide currency in mainstream economics. Given this focus, we unfortunately cannot do justice to the many authors, particularly those studying institutional and economic history, who challenge these views, and offer nuanced caveats.[37] The two legal-history-based paradigms are those following the seminal papers by North and Weingast (henceforth NW)[38] and La Porta, Lopez-de-Silanes, Shleifer, and Vishny (henceforth LLSV).[39] Both paradigms focus on high-level, even constitutional, elements of the legal system rather than on the information that occupies most of English legal history and which

[35] Some economic historians have been very aware of detailed developments in the legal sphere, but it seems to be the case that such economic historians have had little effect on the perspectives on English legal history that are dominant in the mainstream of economic analysis, as exemplified in the works to be discussed in the ensuing paragraphs. R. Harris, "The Encounters", was early in making a case for productive exchange between legal history and economics, stressing that legal historians did not pay sufficient attention to the economic history literature. We are more concerned here with the lack of interchange in the reverse direction.

[36] The word 'finance' appears only twice in J.H. Baker, *An Introduction to English Legal History*, Fifth edition (Oxford: Oxford University Press, 2019) and the pertinent issues are in separate discussions, included under property and contract.

[37] Some salient critiques are N. Sussman and Y. Yafeh, "Institutional Reforms, Financial Development and Sovereign Debt: Britain 1690-1790," *Journal of Economic History* 66 (2006): 906-35; P. Murrell, "Design and Evolution in Institutional Development: The Insignificance of the English Bill of Rights," *Journal of Comparative Economics* 45 (2017) 36-55; L. Neal, "How It All Began: The Monetary and Financial Architecture of Europe During the First Global Capital Markets, 1648-1815," *Financial History Review* 7 (2000): 117-40; P. O'Brien, "The Nature and Historical Evolution of an Exceptional Fiscal State and Its Possible Significance for the Precocious Commercialization and Industrialization of the British Economy from Cromwell to Nelson," *Economic History Review* 64 (2011): 408-46; S. Ogilvie and A.W. Carus, "Institutions and Economic Growth in Historical Perspective" in *Handbook of Economic Growth*, ed. P. Aghion and S.N. Durlauf (Amsterdam: Elsevier, 2014): 403-513; D. Coffman, A. Leonard, and L. Neal (ed.), *Questioning Credible Commitment: Perspectives on the Rise of Financial Capitalism* (Cambridge: Cambridge University Press, 2013); and G.M. Hodgson, "1688 and All That: Property Rights, the Glorious Revolution and the Rise of British Capitalism," *Journal of Institutional Economics* 13 (2017): 79-107.

[38] D.C. North and B. Weingast, "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England," *Journal of Economic History* 49 (1989) 803-32.

[39] R. La Porta, F. Lopez-De-Silanes, A. Shleifer, A., and R.W. Vishny, "Legal Determinants of External Finance," *Journal of Finance* 52 (1997): 1131-50; and R. La Porta, F. Lopez-De-Silanes, A. Shleifer, A., and R.W. Vishny, "Law and Finance," *Journal of Political Economy* 106 (1998): 1113-1155.

provides the data for this paper, that is, the vast collection of reports on the deliberations within the courts. Both sets of works have had enormous influence in economics, in areas far removed from their original domain of application.[40]

The approach of NW is that "the institutional changes of the Glorious Revolution permitted the drive toward British hegemony and dominance of the world".[41] In emphasizing the effects of constitutional measures, particularly the Bill of Rights and the Act of Settlement, NW are followed by the influential works of Acemoglu and Robinson[42] and North, Wallis, and Weingast.[43]

LLSV also emphasize overarching features of the legal system.[44] Their focus is on the overall characteristics of law-making and legal adjudication and how these produce different types of legal processes in common-law and civil-law countries. In many works following on the original papers, summarized by La Porta, Lopez-de-Silanes, and Shleifer[45], the authors, and others, bring enormous amounts of modern data at a very detailed level to bear on their work. But to the extent that they engage with legal history it is at the level of the approaches to law that were developed in England and France and their effect on system-wide characteristics such as judicial independence, the use of juries, organization of the legal system, and the sources of law.

The reader will notice from the above summary that the two influential legal-history paradigms that have had a broad influence across a swathe of economics do not rest on detailed examinations of the vast number of routine developments in the law that is the stuff of the history emphasized by traditional legal historians. These paradigms do not invoke characterizations of the development of English law in the period 1550-1750 that are based on the records of the courts and apply to domains that are crucially important for a capitalist economy—contract, property, and tort. They do not reflect the painstakingly slow developments occurring in procedures, precedent, and forms of legal action, which affected how the courts functioned and how litigants could use the law. In short, within the two institutional narratives that have been most successful in using English legal history to influence the way economists think about the world, the work of scholars within traditional legal-history is largely missing.

[40] A computational search of JSTOR reveals how unusual these two works are in their spread across the whole of economics. North and Weingast, "Constitutions and Commitment", appears in the *Journal of Economic History* and is referred to in JSTOR thirteen times as much as the typical article published at the same time in that journal. The references to North and Weingast, "Constitutions and Commitment", are twice as common in the journals outside economic history as in economic history journals, while for the typical article published in the same journal at the same time, the ratio is 0.6. Similarly, La Porta et al., "Legal Determinants", appears in the *Journal of Finance* and is referred to in JSTOR twenty times as much as the typical article published at the same time in the same journal. The references to La Porta et al., "Legal Determinants", are twice as common in the journals outside finance as in finance journals, while for the typical article at the same time in the same journal, the ratio is 0.24.

[41] North and Weingast, "Constitutions and Commitment", 830.

[42] D. Acemoglu and J.A. Robinson, *Why Nations Fail: The Origins of Power, Prosperity and Poverty* (New York: Crown Business, 2012). See, for example, p. 102 reiterating that "The Glorious Revolution limited the power of the king and the executive, and relocated to Parliament the power to determine economic institutions …The Glorious Revolution was the foundation for creating a pluralistic society…The government…steadfastly enforced property rights… Historically unprecedented was the application of English law to all citizens. Arbitrary taxation ceased, and monopolies were abolished almost completely…"

[43] D.C. North, J.J. Wallis, and B.R. Weingast, *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History* (Cambridge: Cambridge University Press, 2009).

[44] See La Porta et al. "Legal Determinants" and La Porta et al., "Law and Finance".

[45] R. La Porta, F. Lopez-de-Silanes, and A. Shleifer, "The Economic Consequences of Legal Origins," *Journal of Economic Literature* 46 (2008): 285-332.

A machine-learning history of English caselaw offers the chance to bridge the fields of economics and legal history. By using as input the reports on tens of thousands of historical cases, it absorbs, imperfectly for sure, the most important information used by legal historians, the micro-level case-report data that is far removed from the macro-level constitutional and legal-system arrangements emphasized by NW and LLSV. By interpreting the results of the analysis using centuries of insights developed by scholars who have focused on caselaw, a machine-learning approach incorporates elements of traditional legal-historical research and complements existing exegeses on legal development. At the same time, a machine-learning history offers the type of broad narrative about caselaw that is so difficult for the outsider to the field of legal history to gain, even with the use of such a superb textbook as that by Baker.[46] In the ensuing sections, we illustrate the power of topic-modeling in the context of the developments and features of caselaw and legal ideas pertinent to finance.

IV. Characterizing temporal change: The development of caselaw and legal ideas relevant to finance

Within the sections of the 100-topic machine-produced digest of pre-1765 English caselaw and associated legal ideas, we identified 15 topics as pertinent to finance. Topic-modeling does not tell us which topics to designate as relevant to finance: this is our judgment based on an understanding of the content of all topics estimated by GM. The topics we designated as finance ones are: Arbitration & Umpires, Assumpsit, Bankruptcy, Bonds, Claims from Financial Instruments, Contract Interpretation & Validity, Executable Purchase Agreements, Execution & Administration of Estates, Identifying Contractual Breach, Implementing Trusts, Mortgages, Negotiable Bills & Notes, Pleadings on Debt, Prioritizing Claims, and Repaying Debt. Table 1 contains a brief description of these topics, focusing on select key words (or rather their stems) and the top case-reports identified by topic-modeling.[47]

< [Insert Table 1 here] >

Figure 1 presents timelines for these 15 topics over the years 1550-1750. To interpret these figures, it is best to focus on a particular example: we will use Assumpsit. Taking a particular year, say 1600, the figure indicates that the topic Assumpsit occupied roughly 3% of the attention in the case-reports heard in that year.[48] These timelines offer a feature of topic-modeling that has been much emphasized in the literature. They capture the changing amount of attention in English courts in a very long time period reflecting thousands of cases, focusing not on landmark rulings, but rather overall trends reflecting data that might be only a tiny part of each individual case. As Goldstone and Underwood found for the digital humanities, "Quantitative methods may be

[46] See Baker, *An Introduction*.

[47] Many more details on these topics can be found in GM and the corresponding appendices. See note 10 above.

[48] Of course, despite the large number of reports used to produce the data, any given year might have only a few cases. Therefore the figures are moving averages, producing smoothness, especially removing prominent idiosyncrasies arising in years when the data are sparse. Additionally, such figures are usually accompanied by confidence intervals that indicate how imprecise the estimate of the timeline is in any given year. In our applications, those intervals are very narrow for all the timelines. Thus, it is sufficient to focus only on the averages that appear in the diagram.

especially useful for characterizing long, gradual changes, because change of that sort is otherwise difficult to grasp."[49]

< [Insert Figure 1 here] >

There is a crucial question of how to interpret the meaning of that amount of attention, which is a central concern of GM. It is natural to think that the height of a timeline reflects how much a certain area of law is used or not, and this is exactly the assumption in the oft-used word-frequency analysis. The fallacy of such an approach becomes evident on examining our example topic, Assumpsit. Its timeline exhibits an inverted-U, with attention to the topic almost vanishing from case reports during the 18th century. But we know from the careful work of legal historians that the idea of assumpsit was thoroughly embodied in law by that time. So the height of the timeline does not show how much litigants and judges actually depend on a particular idea at a given point in time.[50] We know in fact from the detailed legal history that assumpsit was more and more accepted in the late 16th century, became authoritative early in the 17th century, and was elaborated in many cases in subsequent decades. Therefore, the height of the timeline in a particular year is informative of the rate of development of doctrines in that year rather than the use of the doctrine. This is a reflection of the obvious: litigants do not waste time litigating elements of the law that are accepted by all; judges emphasize the matters that are in dispute; and writers of case reports attract readers by telling them something new, rather than rehashing settled matters.

To explore this logic, GM build a simple evolutionary model of the production of case reports. Here the logic can be easily explained using a simple analogy with a subject that is painfully familiar to us all. The spread of an idea is like the spread of a virus. The inverted-U is like the pandemic curve that we all want to see flattened. Case reports will show a lot of attention to an idea when it is relatively new and becoming more important, just as the count of positive tests for the presence of a virus will rise when the pandemic is becoming very serious. Once an idea is old, it will not show up in the body of case reports, just as there will no longer be many new infections when herd immunity arrives.

Thus the timelines provide a very simple answer to the question of when various aspects of legal development occurred. They are crude, missing out many nuances of the legal record, but that is the cost of trying to summarize masses of data in a parsimonious way. That element of simplicity is present in all statistical work endeavoring to extract simple core facts from masses of data. A non-machine-learning approach to answering this question would necessarily involve deeper investigation into how the language was being used in individual reports and how those reports resonated with the wider context. As in many instances in this paper, we emphasize that

[49] Goldstone and Underwood, "The Quiet Transformations", 379. This resonates with comments in Guldi and Armitage, *The History Manifesto*, and in Flanders on what computers can do: J. Flanders, "Detailism, Digital Texts, and the Problem of Pedantry," *TEXT Technology* 2 (2005): 41-70.

[50] Note that it is entirely possible that the topic Assumpsit vanished from cases in the early 18th century while the word assumpsit was used in a considerable number of case reports from that era. This is possible because topics reflect the co-occurrence of related words rather than only the frequency of single words. When the word assumpsit is used in later cases it might be invoked very briefly to reference a huge area of law without being accompanied by many words that were necessary to use in earlier cases, before the notion of assumpsit became readily accepted.

the two different approaches, our statistical distant reading and the more traditional close reading, are complements. The former is much more likely to reflect the development of ideas within a broad swathe of all cases, including lesser ones. The latter would naturally reflect a narrower set of cases found to be especially influential.

Given that we can view the height of the timeline at any moment as capturing the incremental rate of development of legal ideas, there is an even simpler way to summarize the cumulative development of the law. This will be especially useful in the interpreting the information that appears in the next two sections. Given the evolutionary logic, for any given topic, one can calculate the year that marks the passing of the halfway mark of all legal development that did occur during 1550-1750. (Think of the virus analogy when a vaccine is not available: we could find the precise year in which the proportion of the population that had been infected passed 50%.) We have made this calculation for all 15 topics included in Figure 1, and the relevant years are marked on that figure with vertical lines. To take the example of Assumpsit again, the vertical line is placed at 1631, indicating that half of the legal development pertinent to Assumpsit that would occur during 1550-1750 actually had occurred by 1631. Even though the landmark decision, in Slade's case, was rendered in 1602, our data summary suggests that much development of related law still occurred after that decision. This is not surprising: landmark cases establish a principle that needs to be fully articulated in a variety of settings.

One of the findings that is immediate from a quick perusal of the timelines and dates in Figure 1 is that several pertinent areas of law were substantially settled well before 1688, the period typically given short shrift in the study of English financial arrangements in the economics literature. Significantly, even late developers such as Implementing Trusts and Negotiable Bills & Notes show spikes in attention during the third quarter of the 17th century. Well before the Glorious Revolution, there was broad acceptance by the legal profession of many of the ideas relevant to modern finance. The financial revolution in England was occurring throughout the 17th century, at least as far as the development of pertinent legal ideas was concerned. This is decidedly not the picture that emerges from the main strands of the relevant literature in economics. At the same time, it would be difficult to make this precise conclusion from the traditional legal-historical literature alone: we are not aware of any scholar who has stated this conclusion, let alone documented it in as precise a manner as our use of topic-modeling data does.

Examining the early and late developing topics in Figure 1, it is clear that the areas of law that developed early are rather broad, in the sense that they are not about specific financial instruments per se, but rather about more general areas of law, where progress is perhaps a pre-condition for the use of specific financial instruments. The earliest developing areas are Assumpsit, Bonds, Identifying Contractual Breach, and Pleadings on Debt, all of which are relevant to a wide spectrum of economic activity. In contrast, the areas of law that developed later pertain to much more specific financial arrangements such as Bankruptcy, Mortgages, and Negotiable Bills & Notes.

More generally, for the reader interested in areas of law beyond finance, recall that Figure 1 focuses on just 15 of the 100 topics. Many different lessons on the development of various areas of law could be extracted from the complete set of timelines presented in GM.

V. Uncovering interconnections: The links between finance and other areas of law

We know that a report of a case will normally refer to many different legal ideas, even though the decision in a particular case usually hinges on one particular aspect of law.[51] Detailed rules on Repaying Debt are formulated in the context of earlier developments in Assumpsit and Bonds, for example. Therefore important insights about legal development can be obtained by examining whether case reports emphasizing one particular topic also emphasize other specific topics. Co-occurrence of two topics at the case-report level is evidence of complementarity in the use of legal ideas. It shows that the corresponding topics aid each other in expressing a specific set of ideas, indicating a shared conceptual foundation.

This is (positive) topic correlation, a measure of the degree to which a pair of topics tend to be mentioned in the same case reports. Finding those topic pairs with the largest positive correlations is a first step in detecting associations between different areas of legal development. If one finds that topics X and Y are highly correlated and, furthermore, that X developed earlier than Y, then that is suggestive of causality, with X an input into Y rather than vice versa. For example, the development of law relevant to Bonds is more likely to have provided input into the development of law on Repaying Debt than vice versa, given that these topics are strongly positively correlated and given the information on their timing in Figure 1.

To illustrate these considerations, consider the justifiably uncelebrated case of *Alcock v Blowfield*, heard by the King's Bench in the third year of the reign of Charles I.[52] The case report is an unusual one because one topic dominates: Assumpsit accounts for 69% of the case according to the GM topic-model estimates. Procedural Rulings on Actions accounts for a further 5% of the case report. If this pattern were repeated over a sufficient number of case reports, then one would find that these two topics, one contract and the other procedural, would be correlated with each other. This is indeed the case, with these two topics exhibiting a correlation of 0.25, a rather high level of inter-relationship. However, since the corresponding areas of law were developing at the same time (see Figure 1), we have no strong indication of the direction of causality for this particular topic pair.

It is worth emphasizing that *Alcock v Blowfield* is just one of the 52,949 case reports in the data. The type of information given in the previous paragraph is available for all cases. Because the GM topic model produces data on the proportion of the 100 topics that occupies each of the reports, it is then trivial to find correlations between reports in topic usage. By providing

[51] DiMaggio et al., "Exploiting Affinities", 582, point out, in a rather different context, that topic-modeling's assumption of many ideas mixed in a single text provides a significant advantage: "[A] virtue of topic modeling is its deep affinity to the central insight in the sociology of culture that texts do not necessarily reflect a single perspective but are often characterized by *heteroglossia*, the co-presence of competing 'voices'--perspectives or styles of expression--within a single text."

[52] Alcock v Blowfield (1627) 95 E.R. 74, 1061.

information about the connection between apparently disparate cases, the statistical analysis offers clues that might ultimately be helpful to the more traditional type of analysis usually undertaken by legal historians. Moreover, if the correlations are based on subtle connections between topics that appear in many cases, their existence might be very difficult to detect without quantitative tools: the computer is "a device that extends the range of our perceptions to phenomena too minutely disseminated for our ordinary reading. The computer is…being asked to help the researcher perceive patterns at a finer-than-human level of granularity."[53]

A. The Criteria for Displaying Connections and the Resultant Network of Legal Ideas

With 100 topics, there are 4,950 distinct correlations and therefore a need to focus on the most important. We consider only correlations that are greater than 0.15, of which there are only 85: these are the strongest 2% of the correlations. We are interested primarily in the 15 finance-related topics. Nevertheless, in examining the development of law related to finance it is important to focus not only on these 15 topics, but also on any topics that are related to them, since law outside finance can surely influence the development of finance-related law. In examining correlations, we therefore include all topics related to a finance topic via at most two steps: a non-finance topic is included if it has a correlation greater than 0.15 with any topic that has a correlation of greater than 0.15 with a finance-related topic. This leaves us with 57 links to study, half of which are direct links to the finance topics themselves. From this fact alone, an interesting observation arises. Two-thirds of the most important links in our data, 57 of 85, connect to finance, and one third are directly connected to finance topics. In contrast, finance topics are only 15% of all topics. This is evidence that the development of law related to finance is at the center of English legal developments in the period under study.

Focusing on the top 2% of correlations is a very stringent criterion, forced upon us by a combination of two factors. First, parsimony is essential to extract lessons from overwhelming amounts of data. Second, we are examining an area of law that seems to have many connections with other areas of law. However, if a reader were interested into burrowing down into an area of law that was much less broadly connected with other areas, a weaker criterion for the size of the correlation could be used: the narrowness of the area of the law would provide its own parsimony.[54]

Even 57 correlations are hard to parse if one solely focuses on a list of topics and their associated correlations. In this case a picture is certainly worth a thousand words. We present our findings with the aid of Figure 2. All relevant topics and connections, given the above criteria, appear in the diagram: there are 15 finance topics, 24 non-finance topics that are related to the 15 finance topics, and 57 connections, indicated by dashed lines. The names of the 15 finance topics

[53] Flanders, "Detailism", 57.

[54] For example, if one were interested in the workings of the poor laws one might want to examine topics related to Geographic Settlement of Children. Then one would be led to examine a narrow but interesting set of topics: Reviewing Local Orders, Employment of Apprentices & Servants, Decisions After Criminal Conviction, and Clarifying Legislative Acts.

are capitalized to distinguish them.[55] The topic names are accompanied by the estimate of the mid-year of topic development discussed in the previous section.

< [Insert Figure 2 here] >

B. Insights from the Network of Topics Related to Finance

What such a diagram has the potential to offer is the easy detection of patterns that indicate broad lessons in the development of the law. These patterns are readily found in Figure 2 and they are not difficult to interpret. The core finance-related topics are in a block in the lower left of the diagram with many interconnections between them. To the right of these are a set of topics whose development was concentrated in the first half of the 17th century. These topics are related most closely to contract law and to procedural developments relevant to litigants pursuing contract cases in court. The fact that Assumpsit, an early topic, is connected with procedural topics suggests that the procedural rigor of early common-law was of key importance in addressing matters of debt. Above these topics are a small block of very early developing areas of law connected to transfer of ownership of property or transfer of the right to use the property, for example on leases. The reason for the connection between these and the broader elements of contract law is transparent.

The largest contrast is between the topics in the lower-right of the diagram, connected to contract, and those in the upper-left of the diagram. The latter group focuses on inheritance and wills. Those are topics whose development came much later in the 17th century than the contract-related topics discussed in the previous paragraph. The topics in the upper-left focus on inheritance and wills and mainly concern property issues, as is inevitable given the importance of land as the basis of family relationships at that time. And given the importance of trusts in dealing with these complicated family-inheritance relationships it is not surprising that the topic Implementing Trusts should be intimately connected to this block of topics.

If one wanted to tell an overarching story of development of caselaw and legal ideas relevant to finance that is evoked by this figure but removed from the nuances of specifics, it would be the following. Early stirrings of an agricultural revolution and the growth of the rural textile industry stimulated a market in the transfer of land-use rights. This led to cases concerning disputes on leases and rentals, which in turn spurred refinements in contract law. Such refinements were closely associated with the development of court procedures that channeled contract disputes as they entered the court system. These developments naturally fed into the law relevant to the exchange of financial property and to the debts that arose as a result. But a separate relationship was with the law relevant to both property and the family because the types of arrangements that are so important for finance, trusts and mortgages, for example, were intimately connected with the way in which English families were trying to structure their inheritance arrangements. Given

[55] For an understanding of what the finance topic names signify, the reader is directed to Table 1. For reasons of brevity, similar discussions of the topic names for non-finance topics are omitted, with the reader referred to the relevant elements of GM. After the publication of GM, one topic name that appears in Figure 2 was reconsidered and changed. Interacting in Court has been changed to Decisional Logic, with the renaming prompted by a further reading of the case-reports that most use this topic.

the timing of events, it seems that the two areas of law, finance and family-inheritance, were developed in tandem, rather than one obviously being the precursor of the other.

For the reader interested in examining interconnections between different areas of law, we must emphasize that we have only provided one example of many that could be carried out using as data the correlations derived from the topic-modeling exercise. As far as we are aware, there exists no network analysis on any subject in the pre-industrial legal history literature that is similar to the one explored in Figure 2, even though some aspects of the connections appearing in that figure have certainly been known to legal historians. Where topic-modeling goes beyond what already exists in the legal history literature is that it is a tool to tell a broader story, leveraging a comprehensive set of cases, picking up patterns that might be reflected only in the repetition of thousands of minute sections of text, introducing easily-understood quantifications, and facilitating the use of visualizations that aid the genesis of fresh legal-historical insights.

VI. LAW VERSUS EQUITY IN CASELAW AND LEGAL IDEAS ON FINANCE

In examining the development of legal doctrines, legal historians are very careful to differentiate between law and equity, between the activities of the common-law courts and those outside this system, particularly the Court of Chancery.[56] Nevertheless, this distinction is not made as clearly as it should be in the related economics literature, especially when interpreting the development of law on finance and understanding the strengths of the English legal system. Our methods can clarify which topics—which sections of the digest—are primarily common-law ones and which are equity ones, producing quantitative observations on which types of courts—common-law or equity—were most important in legal developments connected to finance.

The legal record assigns case reports unambiguously to courts. Of the case reports in our data, 23% are from Chancery, and 75% are from the three principal common-law courts, King's Bench, Common Pleas, and Exchequer.[57] Therefore, for the most parsimonious presentation of our results on common-law versus equity, we can simply contrast the presence of finance topics in Chancery reports relative to the presence of the same topics in the reports of all other courts.

Figure 3 contains the pertinent information for the 15 finance topics. Each topic name is accompanied by the estimate of the mid-year of topic development, discussed in Section IV. The topics are ordered vertically by estimated mid-year, with the earliest developing topics at the top

[56] Our interest in examining law versus equity was stimulated by J. Morley, "The Common Law Corporation: The Power of the Trust in Anglo-American Business History," *Columbia Law Review* 116 (2016): 2145-98. Morley describes the crucial role of equity in the development of trusts and argues that trusts afforded many of the legal properties now associated almost uniquely with the modern, legislated corporate form.

[57] In fact, Exchequer considered both common-law and equity cases. See W. H. Bryson, *The Equity Side of The Exchequer: Its Jurisdiction, Administration, Procedures and Records* (Cambridge: Cambridge University Press, 1975). However, the equity side accounted for many fewer cases than the common-law side. For example, in the mid-17th-century Exchequer reports of Hardres, fewer than one-quarter of the cases are equity cases. Moreover, Exchequer reports as a whole are small in number compared to the number of reports from the other three major courts. Lastly, the decision to not take into account the mixed set of cases in Exchequer actually biases our results against finding the conclusions that we reach in this section; that is, a more refined treatment of the division between common-law and equity Exchequer reports would actually strengthen our conclusions.

of the figure. The bars show the proportion of the total attention to a topic due to Chancery. A bar that ends at 0.5 indicates that the associated topic is as likely to appear in equity reports as in common-law reports.[58]

< [Insert Figure 3 here] >

As in the previous two sections, the figure that we use to convey the information in this section contains many more details than we will choose to comment upon. What stands out starkly in this figure are two prominent patterns. First, in the earlier years of the 17th century, the legal developments related to finance were concentrated in the common-law courts. Second, as the 17th century proceeded, Chancery played more and more of a role. Later developments in caselaw and legal ideas on finance are, by and large, concentrated in Chancery.[59]

No doubt these empirical patterns can be explained in different ways.[60] Our favored interpretation is as follows. A workable financial system depended upon improved contract law and related procedures, which were primarily the province of the common-law courts. Without progress in these areas, disputes about debts resultant on financial contracts would have been much harder to resolve, slowing the growth of finance itself. Later, new institutional arrangements were intimately connected to a more detailed articulation of property rights in land connected to inheritance. These changes in property rights depended upon the clear separation between the formal title to the land and the beneficial ownership of the land, the first being a common-law right and the second an equitable one. With that distinction, there was scope for new arrangements that could become the basis of modern finance, such as trusts and mortgages.

VII. CONCLUDING REFLECTIONS

A. What Topic-Modeling Can (and Cannot) Do

We hope that we have been able to convey to legal historians of all stripes the methods and the promise, but also the limitations, of topic-modeling, the most popular method of unsupervised machine-learning. Topic-modeling obtains its power from integrating information from large numbers of documents that could not possibly be summarized in a lifetime pursuing traditional forms of text analysis. In this paper, it leads to the production of data that can quantify the overall timing of specific developments in the law (Figure 1), uncover connections between apparently disparate elements of the law such as Claims from Financial Instruments and Marriage Settlement (Figure 2), and reveal in which courts specific developments principally occurred (Figure 3), showing, for example, the importance of equity to finance. It is this ability to provide the

[58] Usually such figures are accompanied by confidence intervals that indicate the imprecision of the estimates. Yet in the present context (of abundance of data), those intervals are very small, so focusing on the sizes of the bars alone is sufficient.

[59] The reader might be tempted to conclude that the results in Figure 3 are simply due to the fact that Chancery reports became relatively more numerous as the 17th century proceeded. But the pertinent statistical procedures control for year. Therefore the results summarized in Figure 3 do not reflect the relationship between the time period in which a particular issue is prominent and the relative number of all the case reports emanating from the different courts during that time period.

[60] For example, one reviewer of an earlier version of this paper suggested that the patterns in Figure 3 are also consistent with the rise of early corporate law and the increasing use of equity for purposes of resolution of especially complex creditor-debtor arrangements.

information necessary to produce a broad, quantitative macrohistorical overview that is one of the major contributions that can be made by a machine-learning analysis.

Many of the limitations of machine-learning are simply the other side of the coin. Our topic-modeling produced a satellite image rather than a land registry plot. It cannot reveal the crucial moment at which an imaginative leap moved legal thought into new territory, nor the exact source of that new idea. Our application of topic-modeling did not incorporate information from other detailed sources found outside the case reports, for example, from the legislative record or biographies, which might provide clues as to why and when a specific decision was made in the particular circumstances of one case. Despite advances in artificial intelligence, the human reader will, at this point, still move much more easily than the algorithm from the language of case reports to the language of memoir and legislation. Similarly, the human reader will effortlessly detect the sentiments underlying a particular case report: identifying the acceptance or rejection of a particular doctrine in a case report is trivial for human readers, but topic-modeling usually does not distinguish hostility or receptiveness.[61] Topic-modeling detects categories of debate rather than subtleties of position.[62]

Despite current limitations, a careful application of topic-modeling, and indeed machine-learning in general, can now make a large contribution to legal-historical research. We have endeavored to demonstrate this by referring to elements of the economics literature that has used selective elements of English legal history as central assumptions to drive influential theories. Overall developments in English caselaw from 1550 to 1750 bear on these assumptions and a macroscopic legal history can show in a compact way the ebb and flow of caselaw over such a long time period.

In this macro legal history, we were able to show that areas of finance-related caselaw developed much earlier than is assumed within the conventional wisdom in economics, in which the great constitutional and political changes consequent on the Glorious Revolution are seen as defining events. We also showed that a considerable part of the development of the law on finance occurred outside the common-law system itself, in equity. These observations lead to two further correctives to the economics literature discussed in Section III. First, much of the law on finance was created in a court whose officials did not have the type of legally established judicial independence that is often emphasized as being crucially important and that was legislated for common-law judges in the Act of Settlement of 1701. This is additional support for the conclusion

[61] Sentiment analysis is another vibrant area of machine-learning research; see Frankenreiter and Livermore, "Computational Methods". Sentiment analysis, which focuses on specific types of sentiments, was not suitable for our analysis given our objective in the present exercise—to obtain a broad historical overview from a large corpus.

[62] For spirited criticisms of topic-modeling and related computational techniques, see N. Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45 (2019): 601-39; and G. Brookes and T. McEnery, "The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation," *Discourse Studies* 21 (2019): 3-21. Da primarily focuses on the problems of machine learning techniques when applied to literature. Brookes and McEnery criticize a particular implementation of topic-modeling from the perspective of corpus linguistics. We view these contributions as interesting in detecting pitfalls in applications of topic-modeling but not dispositive on its value in general.

that the great constitutional measures following the Glorious Revolution do not seem to have been crucial to the development of the law on finance.[63]

Second, our evidence indicates that a large part of the creation of English law on finance was a product of equity, whose formal legal procedures were perhaps closer to those of the civil-law system than to those of the common-law courts that shared the same building. Thus, juries, adversarial processes, and common-law procedural rules, which are so often emphasized as central to the English legal heritage, were not a critical element in the development of financial law.[64]

B. Topic-Modeling Output as a New Source of Data

A central goal of this paper was to highlight the type of data output produced by topic-modeling. In some sense, our topic-modeling was done by the end of Section II. The remainder of the paper focused on how the output of one topic-modeling exercise can provide the input into separate, self-contained empirical exercises. Unsupervised machine-learning can therefore play a role analogous to many other data-producing exercises that rely on innovative methods to combine large amounts of micro data to construct a macro dataset that many other researchers could use.[65] Topic-modeling, then, is rather like the construction of measures of national income, the production of which might lead to insights in itself, but the objective of which is often the creation of a dataset that can be used as an input into further research (for example, to investigate the determinants of long-run development).

This is a crucial point to emphasize. Because it produces machine-readable output, topic-modeling makes a standard part of the social-science tool-kit a natural part of legal-historical studies. This is the creation and sharing of datasets.[66] Using GM's posted dataset, any researcher could re-estimate the topic model with a different number of topics, rename the topics, and replicate or challenge the exercises carried out in the latter half of this paper. Such researchers would be able to by-pass the really laborious tasks of corpus preparation and proceed to substance immediately, producing new and different results.

[63] P. Murrell, "Did the Independence of Judges Reduce Legal Development in England, 1600-1800?" *Journal of Law and Economics* 64 (2021) 539-65 extends this conclusion more generally to areas outside finance using citation analysis to show that granting independence to judges in England might have had deleterious effects on the development of caselaw in the 1600-1800 time period.

[64] In the papers that are seminal for economics (see note 39 above), LLSV do not mention or allude to the distinction between common-law and equity in the context of the English legal tradition; the authors only emphasize the importance of overall 'legal style', a construct perhaps intended to encompass more than only the common-law system itself. The empirical work in economics that follows LLSV has focused on easily-discernible, legal-system-wide attributes, likely because it is easier to put those in numerical form than to make data out of texts. This often results in the use of data that reflects the operations of common-law courts and not those of equity courts. For example, the economics-oriented empirical research following LLSV (see note 45 above) emphasizes judge-made law, adaptability, and judicial independence as institutional sources of superiority of common-law legal systems over civil-law legal systems. Yet it was Lord Chancellors who made law in instances when the common-law judges, bound by strict procedural rules, could not: equity was far more flexible than the common-law. And the Lord Chancellor was a government official.

[65] For one outstanding example in British economic history, see S. Broadberry, B.M.S. Campbell, A. Klein, M. Overton, and B. van Leeuwen, *British Economic Growth, 1270-1870* (Cambridge: Cambridge University Press, 2015)

[66] GM's dataset is freely available at http://www.econweb.umd.edu/~murrell/Data/ER/ER.html. For any help needed to process this dataset, please contact the authors directly.

The most important direct output of GM is one data set, a 52,949 by 100 table that shows the proportion of each topic in each case report.[67] Ancillary matching datasets provide information helpful in interpreting the data in this matrix, linking each case report to more details of the report, such as the case name, reporter volume, and year. A host of statistical, descriptive, and case-study analyses could be carried out using these data alone: we have presented a sampling above. Beyond this, users could link elements of this data table to their own datasets, matching on years, characteristics of cases, or topics, to produce their own analyses of outcomes that might be far removed from law.

C. Topic-Modeling as an Escape from Whig History

Whig history is one particular instantiation of a more general phenomenon—interpretation of data using a perspective that is narrowed by assumptions particularly relevant to the current age. Goldberg evocatively summarizes the more general problem for his own field: "…Sociologists often round up the usual suspects. They enter metaphorical crime scenes every day, armed with strong and well theorized hypotheses about who the murderer should, or at least plausibly, might be. It is not unlikely that many perpetrators are still walking free as a consequence. Consider them the sociological fugitives of hypothesis testing."[68] And, as he comments, hypothesis testing dominates the social sciences. For economists and for the economic historians who identify as economists, the hypothetico-deductive method has become almost a straightjacket, an enforced avoidance of the types of narratives that do not focus on a particular hypothesis, and that would so resonate with traditional history.

By using and summarizing masses of data in an unsupervised manner, topic-modeling naturally avoids the selective interpretations that can often result when the focus is on a few prominent historical Acts of Parliament or a particular conflict, whether in the form of politics, war, or revolution.[69] Moreover, with an unsupervised approach, the unexpected will surface.[70] In examining the results from a topic model of developments in their own discipline, Goldstone and Underwood provide a justification that could equally apply to the current paper, mutatis mutandis: "…Our model adds nuance to accounts that emphasize a few individual actors in conflict. It shows the emergence and subsequent naturalization of the discourse of criticism over the whole course of the twentieth century, reminding us that the very idea of the discipline of literary study as criticism is the product of a historical development…Topic modeling thus challenges presentist assumptions in methodological debate…allowing literary historians to dramatize changes that may be too gradual, too distributed, or too unconscious to condense into a case study."[71] The methods

[67] This dataset is the easiest to one to convey (in a spreadsheet) to those who have no intention of using the topic-modeling software itself. Use of that software implies that more data are immediately available.

[68] Goldberg, "In Defense", 1.

[69] See J.W.F. Allison, "History to Understand, and History to Reform, English Public Law," *Cambridge Law Journal* 72 (2013) 526-57. Allison emphasizes the selective invocations to which some uses of legal history have fallen prey, and suggests as one remedy the widening of sources.

[70] Topic-modeling can help scholars circumvent the limitations of existing theories and look at the data anew; see R.S. Buurma, "The Fictionality of Topic Modeling: Machine Reading Anthony Trollope's Barsetshire Series," *Big Data & Society* (2015).

[71] Goldstone and Underwood, "The Quiet Transformations", 370.

we have used therefore provide a way to escape from the ever-present temptations of Whig history and its cousins, and to obtain a summary that is as far removed from the hold of past interpretations and theoretical predispositions as one could hope.[72]

One particular consequence in this paper is that the names of the chapters of our machine-produced digest do not correspond to the traditional chapter headings of a conventional study in law. We regard this as a positive outcome, providing the possibility of new insights.[73] But we also recognize this might be regarded by some as a concern, as moving one step too far from existing structures of analysis.

D. The Increasing Importance of the Descriptive and the Inductive

As machine-learning has become more common in the social sciences (it is still very much a minority pursuit), a sense has arisen that the tenor of research is changing. To quote Goldberg again: "The problem with hypothesis testing is not its epistemological foundations, or its ontological validity; rather that, as a practice it has become entirely taken for granted."[74] Descriptive analyses are shunted aside. Unsupervised machine learning forces us to realize that this is happening and it is a problem. In political science, "…The introduction of machine learning methods also invites us to reevaluate the typical model of social science…the current abundance of data allows us to break free from the deductive mindset that was so previously necessitated by data scarcity."[75] In sociology, "Engagement with computational text analysis entails more than adapting new methods to social science research questions. It also requires social scientists to relax some of their own disciplinary biases, such as a preoccupation with causality…".[76] In the digital humanities "…the mathematical assumptions of machine learning—both unsupervised and supervised approaches—are…better equipped for use in the type of inductive, exploratory, and contextual research traditionally conducted using qualitative methods."[77] And even in economics, "In many applications of topic models the goal is to provide an intuitive description of text rather than inference…Real research often involves an iterative tuning process with repeated rounds of estimation, validation, and adjustment…Interpretation or story building…tends to be a major focus for topic models and other unsupervised generative models."[78]

Thus, there are signs that machine-learning, especially of the unsupervised variety, is gradually moving social science research in a direction that historians might find more complementary to their own methods. The inductive will become more acceptable as machine-learning makes clearer the constraints of a tight focus on hypothesis testing. The historical method has always been more inductive in its approach: historians explore the archives, or the yearbooks, or the case reports not to search for the hypothetical needle in a haystack, but rather in the hope

[72] Allison, "History to Understand", makes clear that traditional legal history is not immune to such problems.

[73] Robertson, "Digital Humanities", also emphasizes the positive: finding legal ideas where they are not expected by working outside traditional legal classifications.

[74] Goldberg, "In Defense", 1.

[75] Grimmer et al., "Machine Learning ", 2.

[76] P. DiMaggio, "Adapting Computational Text Analysis to Social Science (and Vice Versa)," *Big Data & Society* (2015).

[77] Nelson, "Leveraging the Alignment", 2.

[78] Gentzkow et al., "Text as Data", 549, 555, 556.

that by becoming immersed in new information, new explanations will arise. The application of topic-modeling reported here is similar in approach: it aimed at providing a broad quantitative narrative of English caselaw on finance over the period of two centuries, and then let insights on the development of law emerge from the data.

There is potential for traditional legal history and empirical economics to become much more complementarity than they have ever been. Indeed, the extensive use of input from traditional legal-historical analysis has been absolutely essential in providing the background for the interpretation of the output of the topic model discussed here. As the digital humanities came to realize very quickly, close and distant reading must be combined. The machine can organize a phenomenal amount of facts, but does not interpret them: when topic-modeling legal history, interpretation is a process requiring the use of the accumulated wisdom of legal historians working for centuries on individual texts. The older methods and the newer ones are complements, not substitutes.

FIGURE 1: FINANCE TOPICS OVER TIME

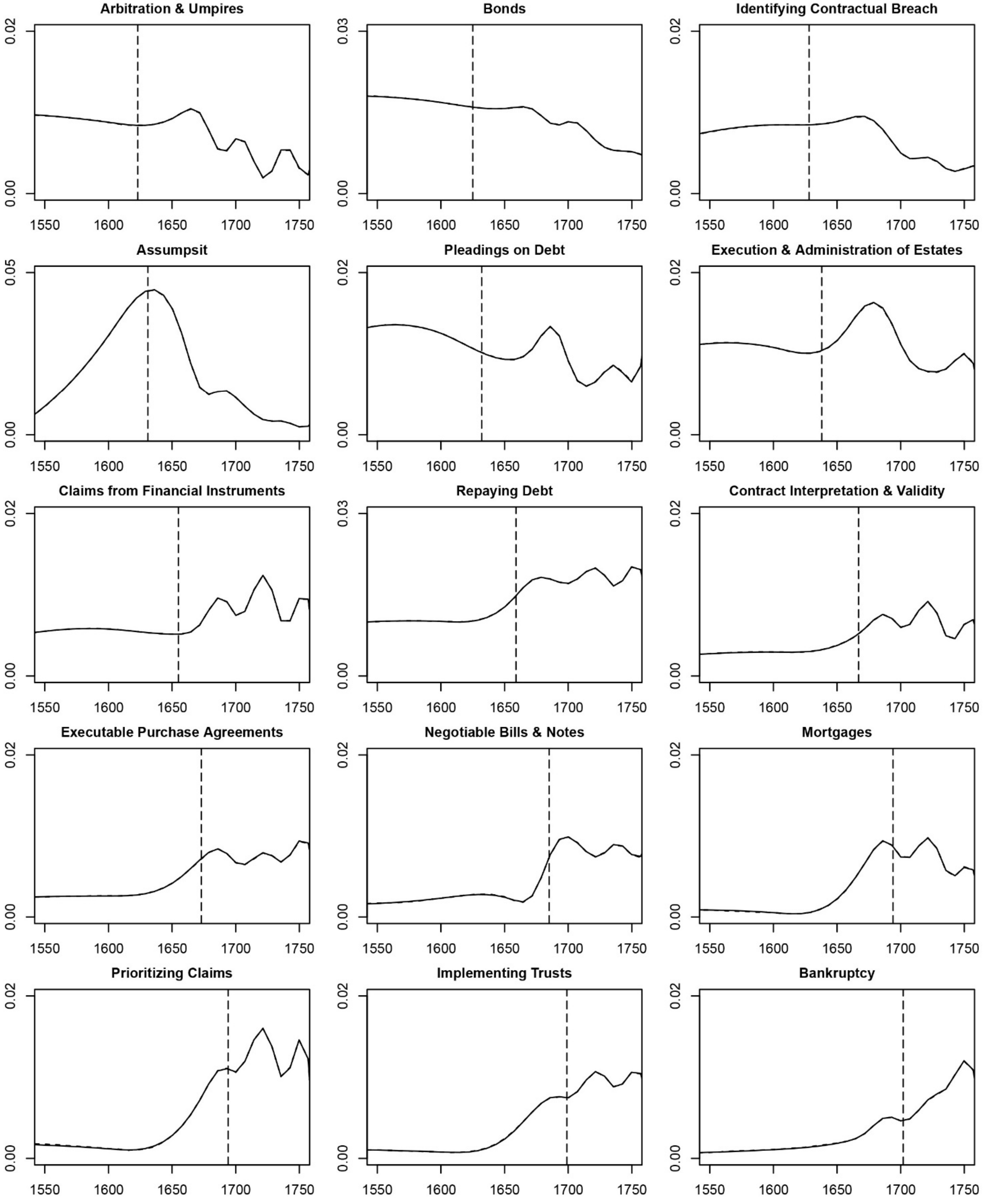

FIGURE 2: INTERCONNECTIONS OF FINANCE TOPICS

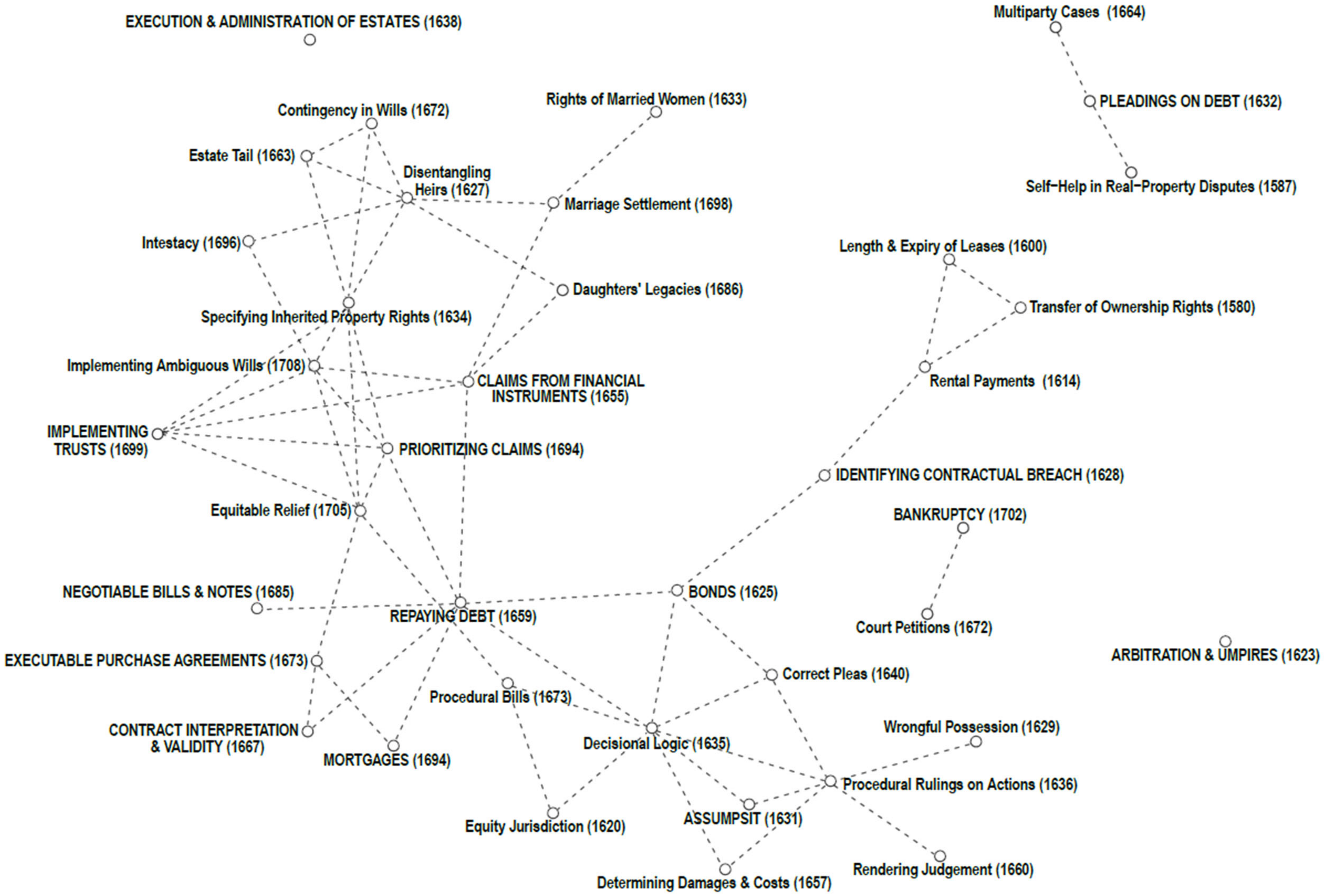
Multiparty Cases (1664)
EXECUTION & ADMINISTRATION OF ESTATES (1638)
Rights of Married Women (1633)
PLEADINGS ON DEBT (1632)
Contingency in Wills (1672)
Estate Tail (1663)
Disentangling Heirs (1627)
Self−Help in Real−Property Disputes (1587)
Marriage Settlement (1698)
Intestacy (1696)
Length & Expiry of Leases (1600)
Daughters' Legacies (1686)
Transfer of Ownership Rights (1580)
Specifying Inherited Property Rights (1634)
Implementing Ambiguous Wills (1708)
CLAIMS FROM FINANCIAL INSTRUMENTS (1655)
Rental Payments (1614)
IMPLEMENTING TRUSTS (1699)
PRIORITIZING CLAIMS (1694)
IDENTIFYING CONTRACTUAL BREACH (1628)
Equitable Relief (1705)
BANKRUPTCY (1702)
NEGOTIABLE BILLS & NOTES (1685)
BONDS (1625)
REPAYING DEBT (1659)
Court Petitions (1672)
ARBITRATION & UMPIRES (1623)
EXECUTABLE PURCHASE AGREEMENTS (1673)
Correct Pleas (1640)
Procedural Bills (1673)
Wrongful Possession (1629)
CONTRACT INTERPRETATION & VALIDITY (1667)
Decisional Logic (1635)
MORTGAGES (1694)
Procedural Rulings on Actions (1636)
ASSUMPSIT (1631)
Equity Jurisdiction (1620)
Rendering Judgement (1660)
Determining Damages & Costs (1657)

FIGURE 3: FINANCE TOPICS IN CHANCERY

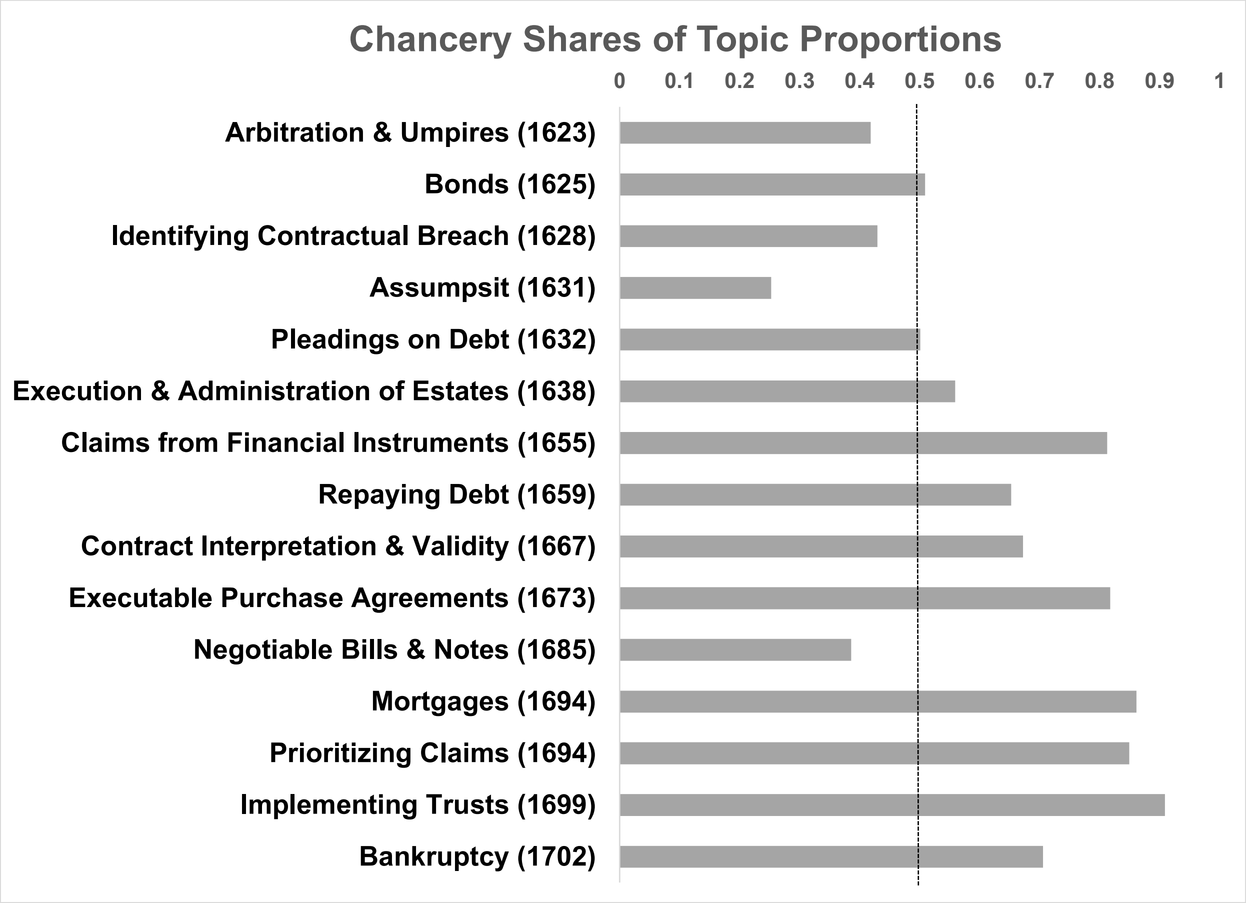

TABLE 1: THE FIFTEEN FINANCE TOPICS BRIEFLY DESCRIBED

Arbitration & Umpires: 0.67% Key word-stems include 'award', 'arbitr', 'umpir', 'arbitra', 'attach', 'releas', 'perform'. Top reports revolve around whether the arbitrators made timely decisions and had chosen an umpire.

Assumpsit: 1.51% Key word-stems include 'assumpsit', 'promis', 'indebitatus', 'consider', 'forbear', 'indebt', 'debt'. Top reports focus on if an assumpsit had taken place and whether an action of assumpsit is allowed.

Bankruptcy: 0.48% Key word-stems include 'bankrupt', 'creditor', 'assigne', 'debt', 'bankruptci', 'assign', 'commiss'. Top reports focus on the assignment of the bankrupt's estate.

Bonds: 1.29% Key word-stems include 'bind', 'condit', 'oblig', 'debt', 'perform', 'void', 'sureti'. Top reports concern bonds, focusing the obligations of the bonds and whether they were satisfied.

Claims from Financial Instruments: 0.75% Key word-stems include 'annuiti', 'cent', 'annum', 'southsea', 'ayear', 'stock', 'dividend'. Top reports describe instances of resolving monetary claims concerning bonds, stocks, dividends, mortgages, annuities.

Contract Interpretation & Validity: 0.56% Key word-stems include 'agreement', 'contract', 'bargain', 'write', 'agre', 'specif', 'sign'. Top reports revolve around interpretation of the meaning of a contract in a given setting.

Executable Purchase Agreements: 0.72% Key word-stems include 'purchas', 'sell', 'convey', 'fraud', 'deed', 'conceal', 'reliev'. Top reports concern contractual transfers of property rights and what renders the contract executable.

Execution & Administration of Estates: 1.12% Key word-stems include 'executor', 'administr', 'testat', 'asset', 'executrix', 'administratrix', 'probat'. Top reports involve the actions of administrators or executors of estates.

Identifying Contractual Breach: 0.64% Key word-stems include 'breach', 'coven', 'perform', 'nonpay', 'evict', 'break', 'refus'. Top reports are about ascertaining and clarifying whether breach of contract has occurred in a given situation.

Implementing Trusts: 0.62% Key word-stems include: 'trust', 'estat', 'chariti', 'profit', 'decre', 'convey', 'beneficiari'. Top reports concern implementation trusts, and rules to determine what is permissible in implementation.

Mortgages: 0.56% Key word-stems include 'mortgag', 'mortgagor', 'redempt', 'equiti', 'encumbranc', 'interest', 'foreclos'. Top reports depict disputes pertaining to rights and obligations of mortgagors, mortgagees, and impacted parties.

Negotiable Bills and Notes: 0.59% Key word-stems include 'bill', 'note', 'accept', 'endorse', 'promissory', 'merchant', 'exchange'. Top reports describe use bills of exchange and promissory notes, focusing on their negotiability.

Pleadings on Debt: 0.99% Key word-stems include 'plea', 'obligatori', 'behalf', 'aforesaid, 'premis', 'verifi', 'attorney'. Top reports focus on the various pleadings to which creditor and debtor have access.

Prioritizing Claims: 0.83% Key word-stems include 'estat', 'debt', 'person', 'shall', 'payment', 'creditor', 'asset'. Top reports focus on who should be paid when claims exceed available funds.

Repaying Debt: 1.59% Key word-stems include 'payment', 'interest', 'due', 'repay', 'discharge', 'indebt', 'lend'. Top reports lay out the details of paying back a sum of money that is owed, often with a focus on interest and often via complex transactions.

Note: The percentage figures are the proportions of the topic in the whole corpus. The mean topic proportion in the whole corpus of reports is 1.0% and the median is 0.81%. The mean topic proportion of the 15 finance topics is 0.85%, the median is 0.67%, and their sum is 12.8%.