Forthcoming in *Handbook of New Institutional Economics*, edited by Claude Ménard and Mary M. Shirley, Springer 2024.

Combining Machine Learning and Econometrics to Examine the Historical Roots of Institutions and Cultures

PETER GRAJZL and PETER MURRELL

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

Murrell: Department of Economics, University of Maryland, College Park, MD 20742, USA.pmurrell@umd.edu

Abstract: Machine learning (ML) and associated computational advances have opened entirely new avenues for the processing and analysis of large data sets, especially those containing text. In this chapter, we show how ML can extend the scope of historical institutional and cultural analysis. We first provide an overview of some of the scattered existing literature using ML methods to study historical institutions and culture. We then use our own work on pre-19th century English caselaw and print culture to illustrate the possibilities and the challenges in using ML as a tool for systematic quantitative inquiry into the origins, change, and impact of institutions and culture. We highlight the power of ML for distilling core facts from large corpora and generating datasets amenable to analysis using conventional econometric analysis. We demonstrate how our work allowed us to explore the deep institutional roots of specific legal and cultural ideas, analyze the coevolution of ideas within caselaw and culture, examine the impact of caselaw on economic development both before and during the Industrial Revolution, and discern critical junctures in England's legal and cultural development. Focusing on historical institutions and culture, the chapter illuminates the types of lessons that can be learned from the application of ML in new institutional economics. It also suggests a pathway that researchers applying ML to history can follow when trying to find a practical, implementable set of methods among the proliferation of new techniques that is usual when an area of research is in its infancy.

Keywords: Machine learning, econometrics, England, topic modeling, word embeddings, VAR, texts, law, culture

1. INTRODUCTION

We provide an overview of our own experience in combining machine learning (ML) and more traditional econometrics to investigate the historical development of institutions and culture. This experience was gained while building a portfolio of research projects on early-modern England, but the lessons we have learned could apply in any historical context. Our intended audience is NIE scholars who are currently interested in developments in this area of research, but are, like we have been, primarily focused on the application of tried-and-tested techniques. We hope to convey two sets of insights in tandem: first, what these new methodological tools can add to the more traditional literature on the origins, evolution, and effects of institutions and culture; second, the insights into English institutional and cultural development that have arisen in our research.

An overarching element of our research follows McCloskey (2016: xiii): "Our riches did not come from piling brick on brick…but from piling idea on idea". Institutional change is thoroughly linked to that process of accumulation of ideas, which are part of culture. In most contexts, the raw data on institutions and culture appears as text, not numbers. This dominance of text is multiplied if, as in our work, the primary focus is on developments before 1800. Then, it is entirely possible that a complete empirical project will have to proceed without the input of any pre-existing numerical data.

The challenges abound. One is the size of text data sets, both in terms of the number of observations and the dimensions of the core element of each observation, a legal opinion or a sermon, perhaps. For English at least, language usage became much more uniform only during the 19th century. In the English of the 17th century, one might find many different spellings of exactly the same word. Any given word in a 17th-century document cannot be relied upon to have the same meaning as it would have today. Untranslated Latin words appear often. The past is a foreign country where Google Translate does not reach.

Gradually over the last few decades, these challenges have receded in difficulty. Desktop computers can now handle datasets of the size likely to be available for early-modern history. The bespoke software and dictionaries needed to standardize spelling and translate Latin words can be created with user-friendly programming languages. Powerful off-the-shelf ML tools can reduce the dimensionality of the individual observations. These same tools can represent the text data in numerical form, making the resultant datasets analyzable using existing econometric methods. By summarizing documents as mixtures of ideas, the methods facilitate an understanding of the estimated ideas that reflects the specific context, even though those ideas are expressed in a 17th-century English that is very different from our own. By leveraging word-use context, individual words can be represented as multi-dimensional numerical vectors, thereby allowing one to identify synonyms by finding those word-vectors that are most similar to the vector representing a word of interest. Then, study of a word's close synonyms gives insights into the way that the word was used in context, for example in 17th century English legal opinions. These are truly tools to produce "verbal thermometers" (McCloskey 2016: 236). None of this is to say that the computer is a substitute for the close reading of texts. Rather, it complements reading as the researcher tries to

figure out history's puzzles. And in sharp contrast to human analysts, these ML tools can synthesize and summarize the content of tens of thousands of texts without importing preconceptions.

Any researcher who has discovered a treasure trove of texts and wants to apply ML will be immediately struck by the large variety of tools that are available and the lack of a clear guide as to which ones will be most useful. This is a characteristic of any powerful new methodology in its infancy. In addition to the proliferation of different tools, many papers contribute ad hoc methods of summarizing the content of the texts. There is nothing that corresponds to a core econometrics textbook that leads one easily into the methodological literature. There are not readily identifiable schools of thought to orient one's thinking, such as the clear divide between structural and reduced-form methods in econometrics. To be sure, there are very good reviews, but the differences in perspective between these attest to the fact that this area of research has not reached a level of maturity where it is obvious where to begin a rigorous piece of applied work that is focused purely on generating new substantive insights. And these reviews are certainly not focused on the trials and tribulations of dealing with pre-19th century data. For examples of such works, see Burkov (2019), Athey and Imbens (2019), Gentzkow et al. (2019), Prüfer and Prüfer (2020), James et al. (2023), and Grimmer et al. (2021).

We have no ambition of bringing order to the inchoate sets of methodologies. This is not a work of synthesis. Rather, by showing the reader what we have done, we hope to convey one viable route to producing new sets of quantitative insights from texts. By emphasizing that the combination of ML methods and econometrics is the cornerstone upon which a research project is built, we highlight the types of fundamental substantive results that one could hope to generate. Indeed, some of our findings recast elements of conventional wisdom that have been in place for centuries.

To set the stage, we begin in Section 2 with a brief overview of the scattered literature that uses ML to advance the understanding of the historical origins and evolution of institutions and culture. Sections 3 to 6 chronicle the processes and the deliberations involved in our application of ML. We describe text preprocessing and then the key ML methods relevant to institutional economists primarily interested in learning about the past. We discuss the methodological paradigm that the use of ML entails. We also clarify that the real power of ML appears when it is used to develop a comprehensive empirical overview of institutional and cultural development. We illustrate this point with examples from our own research. Section 7 concludes with broader reflections on the use of ML and empirical research in economics.

II. MACHINE LEARNING APPLIED TO INSTITUTIONS AND CULTURE IN HISTORICAL SETTINGS: A SCATTERED LITERATURE

With institutions and culture now occupying a central place in theories and narratives of comparative economic development, scholars have turned to empirically examining the historical origins and evolution of institutions and culture. In this pursuit, new ML techniques have emerged as a valuable complement to traditional econometrics. Here, we point to a sampling of studies demonstrating the power of the new methods either as a means of addressing preexisting empirical

challenges or, more interestingly, charting new research avenues. We focus primarily on scholarship in the era before the 19th century. Because this earlier setting inherently comes with an abundance of data challenges, it facilitates discussion of the productive application of the new methodological tools as well as provides insights into their limits.

As we clarify below, the research examined in this section addresses institutions and cultures scattered across a wide range of subperiods and geographical contexts. It is the existing or potential application of ML that is the primary cross-cutting theme. This section then provides a useful backdrop to the discussion of our research, which has endeavored to combine ML and econometrics to examine one specific historical setting, early-modern England.

Addressing Challenges in Conventional Historical Research: Data Collection

A first challenge in quantitative historical research on institutions and culture is to create the datasets themselves. The creation of new datasets has dramatically increased at the same time as the application of ML, often using the pre-processing tools that have gained widespread use in the service of creating text-based-data for ML. For example, Pagé-Perron (2018) employs text processing and network analysis on a cuneiform corpus, providing insights into how ML converts unwieldy corpora into a more usable form that will facilitate future research into Mesopotamian culture and institutions. This is one aspect of ML that we emphasize in future sections. In old texts, tokenization, the separation of a document into lexical units, and lemmatization (removal of inflections to focus on the same basic words) will often require the creation of bespoke processing software. But after the first creation, methods and database will be available for future researchers. Many other studies have made available databases constructed using ML methods. Moreover, datasets are also being constructed for more traditional empirical exercises without the direct application of ML algorithms, but with the use of the types of computational resources for big data that have become available only in recent decades. Table 1, part A, lists examples of relevant contributions.

Addressing Challenges in Conventional Historical Research: Data Analysis

One common empirical challenge that can arise in historical research on institutions and culture involves addressing the curse of dimensionality. In pre-19th century historical settings, where cross-sectional samples tend to be small and the frequency of observations over time is low, the number of observations may barely exceed, or even fall short of, the number of variables. ML can then be productively employed in choosing predictors and allowing for flexible model specifications.

In standard settings, techniques such as ridge regression and lasso explicitly select a subset of pre-specified variables to avoid overfitting. For example, Michalopoulos and Xu (2021) use lasso to analyze age-old folklore concepts as predictors of present-day cultural values and beliefs as measured in surveys. Notably, the authors also discover that the set of models selected by ML might exhibit inherent instability across various specifications, particularly when predictors are highly correlated. Therefore, a supplementary human classification exercise was essential to increase precision of the results.

----- [Table 1 about here] -----

In settings where non-linearities are particularly important, decision-tree methods can be employed effectively to cast light on the determinants and importance of institutions and cultures. Such methods, often implemented in the form of random forests (ensembles of individual trees), select non-linearities and variable-interactions in an iterative fashion, with little prior model specification by the researcher. While random forests allow the researcher to ascertain the relative importance of the different variables in predicting an outcome, this information is usually uninformative of the direction of a variable's influence. Quite often, therefore, it is valuable to complement the random-forest results with structural regression estimates, where the model facilitates easy interpretation of the sign of an effect. Table 1, part B, offer examples of pertinent contributions.

Leveraging Text-as-Data

Several studies from widely different historical contexts use ML methods for analysis of text as data to illuminate previously unknown aspects of institutional and cultural evolution. For instance, what were the themes emphasized by newspapers in colonial and early independent America? In, to our knowledge, the first application in a historical-cultural setting of topic modeling, a technique discussed later in the paper, Newman and Block (2006) focuses on the Pennsylvania Gazette. Their findings indicate, for example, that the presence of ideas involving a national government experienced a surge from the 1760s to the 1790s. But perhaps even more importantly, the authors demonstrate the potential of ML to "reveal cultural…histories", to enable an investigation devoid of "fallible human indexing or their own preconceived identification of topics", and to utilize "orders-of-magnitude more documents than a person can reasonably view".

Also drawing on information encapsulated in or pertaining to texts, studies of networks using ML methods have become common across a wide variety of fields. The idea is to take a set of interactions (articles and their citations are a modern example) and use the computational methods to find patterns in interactions when the core dataset might contain millions of potential interactions. For example, Van Vugt (2022) employs a network-based approach to investigate the exchange of letters between Antonio Magliabechi, an eminent Florentine librarian, and his fellow scholars. The analysis involves examining data on Magliabechi's correspondents, the timing of the correspondence, and the connections of those correspondents beyond direct interactions with Magliabechi. The findings provide a nuanced depiction of the dynamic and evolving relationships among a group of interacting social agents, that is, on the spread of culture. Table 1, part C, lists examples of a number of recent contributions leveraging text as data.

3. THE CORPORA UTILIZED IN OUR ANALYSIS

Our analysis utilizes two major text corpora: the English Reports (Renton 1900-1932; hereafter ER) and the Early English Books Online-Text Creation Partnership (2022; hereafter EEBO-TCP).

In the following, we describe each of the two corpora and the pre-processing challenges that they entailed.

The English Reports (ER)

The ER compiles the definitive set of 129,042 reports on decisions rendered in the superior English courts of law between the early 13th century and the mid-19th century. Coverage is sparse for the early years, but the number of reports becomes substantial by the mid-16th century.

As is familiar in historical research, the ER constitutes neither the population nor a random sample of cases adjudicated in the English courts. Rather, the ER includes a subset of the cases heard by the superior courts. Reporters were especially eager to provide a record of cases that elucidated novel or unsettled aspects of law, that is, cases giving rise to legal development (Grajzl and Murrell 2021b). However, there is no precise record detailing the criteria used to select cases for incorporation into the compilations of reports that were subsequently consolidated in the ER.

The ER became the de facto record of court cases that the English legal profession used as its authoritative source for legal precedent. Thus, the ER contains a record of nearly all cases that came to influence the law. In this sense, the ER provide unique insight into the nature of English caselaw development between the mid-16th and late 18th centuries. There is no alternative machine-readable legal-historical corpus of comparable depth and breadth available for this era. Any exploration of English legal history has to use the cases in the ER.

The Early English Books Online-Text Creation Partnership (EEBO-TCP)

The EEBO-TCP is a machine-readable corpus of 60,331 texts, which captures core works representative of pre-1700 English print culture. (Gavin (2017) provides a history of the EEBO-TCP project, with interpretation.) The texts were prepared by the Text Creation Partnership (TCP) and included in Early English Books Online (EEBO). The starting point of the creation of the texts were catalogs that "trace the history of English thought from the first book printed in English in 1475 through to 1700" (TCP 2022). The EEBO-TCP corpus comprises about one half of the texts listed in a master catalog of works known to still exist.

Texts were included in the EEBO-TCP in an attempt to create a representative corpus. The underlying vision was "to key as many different works—as much different text—as possible" (TCP 2022), that is, to build a corpus that approximated a random sample of available works. Yet, it would be inappropriate to consider the EEBO-TCP corpus as representing a random sample of English culture. The texts mirror the culture embedded in printed materials rather than representing broader popular culture. The texts included are those that have survived, thus reflecting the extent to which subsequent generations valued them. And the EEBO-TCP project prioritized first editions, making TCP oriented to the creation of culture rather than the consumption of culture.

Nevertheless, the EEBO-TCP corpus is unique as a resource for understanding pre-18th-century English print culture. No alternative machine-readable corpus has been assembled that is both capable of supporting a quantitative exploration of English culture before 1700 and has such a comprehensive range of texts. The EEBO-TCP is an invaluable collection for those interested in understanding English history from the advent of the printing press to the time when peaceful

accession of new governments was assured. It is especially valuable for those who want to use computational methods.

Pre-Processing the Corpora

Computational analysis of historical texts requires much arduous, time-consuming work pre-processing texts. The challenges are numerous. We escaped the step of parsing the inscrutable historical fonts because we were able to access two corpora that had already been made machine-readable. But there are many steps between machine-readable and ML friendly. English orthography was not standardized until at least eighteenth century. A machine does not know that thynkynge and thinking are the same word. Modern natural language programs will not know archaic inflections, and therefore that endeth indicates the presence of the verb-stem end. And given that the learned in the 17th century and before were fond of using Latin, the computer will not understand, unless one instructs it, that equus can be rendered as horse. Nor will it know that equo is the dative of equus, a problem that we solved by programming a Latin stemmer, that is, an algorithm to convert words to their Latin roots.

Thus, the texts were replete with words that could not be found in extremely comprehensive dictionaries of modern English words. Translation was required for all these words. This entailed the construction of bespoke project-specific dictionaries to do translations by machine, since any process relying on humans would be infeasible without a large budget. At the same time, all production-inserted formatting symbols were removed, as well as non-recognizable symbols that had strayed into the files. Then we dropped documents that contained either an especially small number of words or an uncharacteristically high share of words that could not be matched to any word in the modern English dictionary even after the processing. Some early English legal cases were written in Law French, an idiosyncratic blend of French, Latin, and English. Unfortunately, we had to drop these cases, a lingering regret.

We then assigned to each document metadata variables such as the year of publication, the number of words, and, in the case of the ER, the adjudicating court and reporter name. The resultant corpora were additionally processed using the steps recommended for the variety of ML that we first used, a structural topic model (Roberts et al. 2014, 2016). We thereby converted all words to lower case and applied a standard stemming algorithm that converted all inflected words to their root form. Finally, we removed standard English stop words (e.g., the, an, etc.), numbers, words with fewer than three characters, words appearing in only one document, and punctuation.

4. THREE EXAMPLE TECHNIQUES

So, one has a data set that comprises a corpus that has been cleaned. What does one do with it? Here, we do not discuss the technicalities of the methods. Understanding the basics of what is being estimated does not require knowledge of the technicalities. Moreover, we only discuss the tools that we think will provide easy entry points for institutional economists whose primary goal is to learn about the past using tools readily accessible to the non-specialist. The selection of methods to discuss is based on our own experience.

Topic Modeling

In the ML literature on text-analysis, by far the most widely used tool is topic modeling. Topic modeling takes a set of documents and asks what subjects or ideas (i.e., topics) can be found in these documents. Like all data analysis it is about finding useful (aggregate, predictive, evocative, causal) patterns in the data, patterns that can be interpreted as reflecting a subset of the ideas that one could plausibly expect to find in the data.

Topic modeling is an unsupervised ML method, meaning that the dataset does not contain any labeled observations that are examples of the pattern that one is seeking. Thus, if one takes the works of Francis Bacon (English philosopher, statesman, and scientist) and Edward Coke (the most celebrated common-law lawyer), as we did in Grajzl and Murrell (2021a), one might expect that the corpus would contain ideas about epistemology and constitutional law. But these ideas appear in greater or lesser amounts in different documents, and in some documents they might be only fleetingly employed. Thus, there are no preexisting document labels, saying, for example, 'this document is primarily about epistemology and not about constitutional law'. One simply has patterns in the correlations of word usage across documents. Topic modeling uses these patterns to identify ideas emphasized in the corpus: these are the topics (e.g., epistemology and constitutional law). Of course, we do not directly observe the ideas themselves and cannot test how well we have characterized the ideas.

Topic modeling then has two fundamental objectives. The more prosaic is simply a reduction of dimensionality in the data, for example, reducing to 105 topics the 142 million words in 67,455 reports on court cases heard in England's high courts between 1765 and 1865 (Grajzl and Murrell 2023c). In this sense, topic modeling can be viewed simply as an extension of existing latent variable techniques for numerical data, such as principal components. When pursuing dimensionality-reduction, the prime consideration is to finish up with a set of variables—often authors focus on only one—that can be used in some other statistical exercise, such as a differences-in-differences regression.

But, for us, the most exciting aspect of topic modeling follows from a second objective—using the results to interpret and use a large set of latent variables that summarize the texts completely. It is a way to uncover deeper ideas that are implicit in many documents, even ideas that are never the explicit focus of any specific document. In Grajzl and Murrell (2019), where we focused on Bacon alone, the topic epistemology was, unsurprisingly, found to be one of the most important, but so was a less expected topic, religious law, reflecting the emphases of a man bathed in Puritanism from an early age and later to become Lord Chancellor. Even less expected, our later investigation of the outputs of the topic model showed that these two topics were deeply connected, suggesting the origins of Bacon's epistemology in his religious and legal background.

Topic-modeling algorithms can be motivated by a crude conceptualization of document generation that lends itself to formalization. The model presumes that the author of a specific document being examined begins with a fixed number of ideas, or topics, which are available for use when writing. Use of a particular topic implies that the author has greater affinity for the vocabulary more closely associated with that topic. Moreover, the author's emphasis will vary

across documents. Then, even with each document viewed simply as a frequency distribution of word usage, semantic content is embedded in the corpus via correlations between word distributions across documents.

What in particular does a topic model estimate? The researcher specifies ex ante the number of topics (T), or sets of ideas, that are in the dataset, a corpus of documents. The algorithm estimates T separate probability distributions over vocabularies. Those distributions are the topics. This is intuitive. When authors write about transubstantiation, as in the corpus examined in Grajzl and Murrell (2023d), they will place a high probability on the use of words such as sacrament, body, and bread. In contrast, when they are producing their lusty entertainments they would tend to favor a different set of words, such as cuckold, fop, and whore. Thus, the words most used by a topic are one key to interpreting the ideas captured by the topic. A majority of papers use this information, and only this information, to give the topic a name.

We think it unfortunate that, in naming the topics, most papers rely only on the words most associated with each topic, without additionally considering which documents feature each topic most prominently. In our work, we found it equally important to read those top documents, sometimes 20, but more usually as many as 40 or more per each topic. This was crucial: topics can only be understood in relation to all other topics—nuances appear in the differences between closely related topics. This reading requires domain-specific knowledge: ML is never purely computational. For example, in Grajzl and Murrell (2021b), the topics Contract Interpretation & Validity and Identifying Contractual Breach had many characteristics in common and it was only by comparing the two that the underlying ideas behind each topic could be readily isolated. Our experience has been the opposite of Rule et al.'s (2015) who view estimated topics as being inherently hard to identify. Indeed, we have found that scrutiny of the estimated topics helps ease one into study of a historical period whose language and context might, at first, feel very foreign.

By far the most useful output of the topic model is the document-topic matrix (DTM). With N documents in a corpus and T topics estimated, then the DTM is an N×T matrix where element ij indicates the proportion of document i that arises from the author's use of topic j. Connecting this matrix to data on the characteristics of documents provides a wealth of opportunities to explore how ideas change over time, how they differ between authors, etc. For example, if one is interested in running a regression with documents as observations, then one might include just one column of the matrix as an explanatory variable. If one is interested in changing emphases over time and one knows the year of each document, then it is trivial to aggregate rows of this matrix to produce a year-by-topic matrix showing how the attention to each idea, or topic, changed over time.

Casting this abstract description into a concrete setting might be helpful to readers. In Grajzl and Murrell (2023d) we aim to use ML to further the understanding of English culture in the 16th and 17th centuries. After pre-processing, the final corpus comprised 57,863 documents and contained 83,337,912 words. After a series of exercises using both objective statistical measures and subjective evaluation of the topic-modeling output, we chose 110 topics. We then examined the words most highly associated with each topic and the documents prominently featuring the

topic. If the topic model is being used to present an integrated picture of relevant history, then this very laborious step is necessary to confidently draw conclusions, or state new facts.

For example, the top word-stems for one topic emphasized logical connectives, such as yet, though, and thus. The top documents often contained the word 'experiment' and most focused on religion. One top document stated that "If a man were but well read in the story and various passages of his life, he might be able to make an experimentall divinitie of his own. He that is observant of Gods former dealings and dispensations towards him, may be thence furnished with a rich treasury of experience..." We named the topic Baconian Theology, indicating theological ideas that used the epistemology that is characteristic of Bacon. Then, a timeline of the use of this topic was produced. This topic appeared in the corpus before the contributions of Bacon, suggesting that Bacon used what was present in English culture before he himself made his mark. This fundamental observation about the evolution and origins of ideas is readily, almost trivially, extracted from the relevant data, once the data have been generated.

Importantly, the topic-naming process can lead to new findings. Therefore, it is important not to place this naming exercise within the straitjacket of clearly identified categories already appearing in the literature. If such matches do occur, then it should be because the topic-modeling estimates dictated them. If they did not occur, then one enters discovery mode, which was exactly the situation arising in the detection of Baconian Theology, a specific set of ideas that none of the preexisting literature prepared us for. Therefore, it is important to identify the ideas underlying all estimated topics.[1] Thereby, one is forced to go beyond one's presuppositions and identify the unexpected. Being able to name all topics with terms that are logically compelling is a verification of the quality of the topic modeling.

Above we have suggested two different levels of analysis—use a very small number of topic-related variables in a standard econometric analysis or develop a broad picture using all topics, often 100 or more, to provide an overview of an era of institutional or cultural history. But there is an intermediate level: aggregating the topics (the columns of the DTM) into a smaller number of themes, maintaining a comprehensive overview but fewer variables, a number that might be more manageable in a traditional econometric analysis. Thus, Grajzl and Murrell (2023c) characterize caselaw development during 100 years of English industrialization (1765-1865) with 105 topics aggregated into 13 readily inferred broader themes. The topics are grouped into themes using researcher judgment. This fits with our general philosophy of using ML where necessary (creating a 67,455-by-105 DTM) and using judgment and existing knowledge where possible (aggregating 105 columns of that matrix into 13).

The above description reflects elements of the methodologies of a number of papers, covering several different subjects and historical eras. The methodologies had many common elements. The findings are disparate and cannot be combined into a simple story. But perhaps one element

[1] Sometimes topics are interpretable but say nothing about substance. For example, Grajzl and Murrell (2021b) found the topic non-translated Latin, arising because pre-processing worked imperfectly on the highly idiosyncratic Latin of sixteenth and seventeenth century lawyers.

running through all the findings can be emphasized. We find that specific ideas appeared earlier than is generally assumed in the existing literature. Thus, our analysis of early English law suggests that major developments in financial markets were already ongoing in the first part of the 17th century (Grajzl and Murrell 2021c), earlier than mooted in most previous studies. In addition to the findings on Baconian Theology noted above, in Grajzl and Murrell (2023b) we locate the stirrings of the English scientific revolution in 1558 to 1610, earlier than is conventional (Wootton 2015). Given the central place of the financial and scientific revolutions in any story of England's rise, these are fundamental findings.

We believe that there are two prime reasons for this earlier dating. The first is familiar in modern academia. The synthetic work that summarizes new developments in an organized and evocative way gets more attention than the often-inchoate publications that were the most innovative (McMahan and McFarland 2021). Indeed, in our ongoing work, we show that this very process might account in no small part for an over-estimation of the novelty of the ideas of Edward Coke, often characterized as the greatest lawyer in the period 1550-1750. The second is due to the difference between ML and traditional text analysis. Fragments of ideas spread across many documents might be as important as the same ideas dominating a single work. The latter is much more likely to appear later, when ideas have coalesced, than the former, a result of incomplete ideas percolating. The latter will normally be dominant in traditional text analysis, but not in ML.

The above has only referred to the descriptive findings of our topic modeling analysis, findings that come almost transparently from the topic-model estimates. But much more can be done with the DTM, as we will discuss in Section 6 on the integration of topic modeling and standard econometrics. In the next two subsections, we turn to alternative ways to encapsulate and investigate large text datasets.

Measuring Novelty Using Words

Wootton (2015: 18) provides a capsule summary of English history that has a large effect on our work: "…let us take for a moment a typical well-educated European in 1600 – we will take someone from England, but it would make no significant difference if it were someone from any other European country as, in 1600, they all share the same intellectual culture. …. But now let us jump far ahead. Let us take an educated Englishman a century and a quarter later…[Then] what was true of an educated Englishman in 1733 would not be true of a Frenchman, an Italian, a German or even a Dutchman." Although science is the subject of Wootton's book, this observation should not be restricted to science, but to all manner of English ideas. So something remarkable happened in England during 1550-1750, with many debates about what that was and when that was. Our topic modeling results, as we have mentioned, do reflect on that, but alternative methods are available.

One such method that we are pursuing in research that is still forthcoming is to implement the method of Kelly et al. (2021) on our legal and cultural databases. The essence of that method is as follows. Any text document can be characterized by a frequency distribution over the vocabulary it uses. Because of the computational demands resulting from the use of large matrices, pre-processing steps beyond those normally applied have to be used to reduce the number of items in

the vocabulary, generating a computationally manageable representation of the corpus. Standard packages are available to do this. For the 1765-1865 ER corpus, for example, this distribution is summarized by a vector of length 10,791 for each document. Take any two documents and visualize their word distributions as vectors in n-dimensional space. Then, the cosine of the angle between the vectors is a measure of how similar the documents are. A document is more novel the smaller is the sum of its cosine-similarities with all the documents produced in previous years.

Thus, one can calculate a novelty score for every single document in the corpus. It is then trivial to calculate timelines of aggregate yearly novelty. And using data characterizing documents one can generate disaggregated timelines. For example, using the document-theme matrix from a previous topic modeling exercise on the development of caselaw, one can find which of procedure caselaw (subsuming topics such as procedural bills, correct pleas, and evidence gathering and admissibility) or real property caselaw (encompassing topics such as manorial tenures, common-land disputes, and uses) was most novel, and when. Our early results suggest the former after 1640 and the latter before 1640.

Or one can find out which particular courts were producing the most novel decisions and when, contrasting, perhaps, the contributions of equity (especially in Chancery) and common-law courts. (Equity was a distinct system of legal principles and remedies that supplemented the common law. Equity relied on laxer pleading rules, used an inquisitorial procedure without juries, and provided relief in the form of decrees and injunctions.) Our early results suggest the common-law was dominant in innovation until the 1660's but after that equity courts were hugely important.

Or one might be interested in, for example, which cultural outputs were emphasized to a greater or lesser degree at different periods of history. Take literature or institutions for example. Our early results suggest that, not surprisingly, 1640 was a great dividing line, with institutions emphasized much more thereafter, as the production of literature became relatively less important.

Both of the methods we have discussed above disregard word order when analyzing a text. But methods that leverage order, albeit in a limited way, also exist. The next subsection describes one that was first proposed in 2013.

Word Embeddings to Characterize Change in Ideas

Word embeddings focus on capturing the semantic relationships between individual words. Any specific word is characterized as a vector in a low dimensional space (e.g., 300 dimensions has been a common choice). The position of the word in the vector space reflects the word's meaning. Estimates of the numerical vector for each word are obtained by leveraging the words that in a text appear close to the target word. Utilizing these word-embeddings estimates, one can gain insights into the usage of the words by examining the relationships between the vectors for the words. For example, if one found in a given corpus that the difference between the vectors for man and woman was very similar to the difference between the vectors for career and family, then this would be an indication of stereotyping of gender roles (see, e.g., Ash et al. 2024).

One can readily see how word-embeddings models can be productively employed to cast light on the historical evolution of institutions and culture. For instance, Gennaro and Ash (2022) investigate the text of speeches given in the U.S. Congress from 1858 onwards in order to

characterize the relative emphasis of reason and emotionality in political discourse. The authors first estimate a word-embedding vector for each word in the corpus, and then, by averaging the vectors for each word in a document, obtain a vector for each document. At the same time, they use standard external dictionaries such as those included in the LIWC (2022) to define reason- and emotionality-related word lists. A vector for each word list can be obtained as the average of vectors for individual words on the list. Gennaro and Ash (2022) then characterize every speech along an emotion-reason continuum by computing measures of vector similarity between the document vector and vectors for the emotionality and reason word lists. An examination of the results indicates, for instance, that emotional expression in the U.S. Congress remained comparatively low and stable until the mid-20th century but experienced a notable increase from the late 1970s onwards.

In our ongoing work, we are using a similar approach to offer the very first quantitative assessment of one the most famous hypotheses on socio-legal evolution—Henry Maine's 1861 dictum that the evolution of progressive societies follows a pattern "from Status to Contract". In premodern societies, legal rights and obligations, according to Maine, were rooted in 'status.' This status was defined by intricate relational and hierarchical networks within the family and broader societal structures. A person's ascribed position in society determined the range of choices available to them in, for example, commerce and marriage. In contrast, in what Maine perceived as progressive societies, such as Victorian England, there emerged a shift toward relationships based on 'contract'. In this context, the law explicitly acknowledges the capacity of individuals to freely assume powers, responsibilities, and authority.

To assess the validity of Maine's conjecture, we are using a word-embeddings approach to generate document-level measures of the relative emphasis on status and contract concepts in early-modern English caselaw and culture. One could, of course, simply follow the approach by Gennaro and Ash (2022) and rely on predetermined external dictionaries to find the words related to these concepts. However, the use of word-embeddings ultimately requires that the dictionaries be specified sufficiently reliably, a task that is especially challenging in historical contexts such as ours.

For instance, in our setting, it would be difficult even for a well-trained modern-day historian to specify ex ante an encompassing set of terms that adequately reflect the notion of tenure, the essence of status-based relations in a feudal society. By inspecting the set of words that each have a word-embedding highly correlated with that for tenure, one generates a list of synonyms that are highly pertinent for the 17th century. This reveals the importance of, e.g., knight-service, heriot-service, and frankalmoign (tenure in free alms), which can be added to the word list of tenure-related terms. The procedure might also reveal words that must be dropped from an initial word list because their early-modern usage was very different from usage today. Thus, our procedure uses the word embeddings not only to characterize concepts and documents, as in Gennaro and Ash (2022), but also to finetune the external dictionary word lists by identifying a whole set of relevant synonyms for any particular word. Importantly, the synonyms implied by word embeddings reflect word usage in exactly the historical period and context that one is studying.

The corresponding iterative, word-embeddings-aided approach to dictionary construction (see, e.g., Rice and Zorn 2021) is therefore especially productive in the context of specialized corpora, such as legal documents, and generally for corpora from times when word usage was very different from our own. In our setting, the word embeddings effectively teach the researcher 17th century English. For example, in early-modern caselaw, drawer is most closely associated with endorsor and acceptor. In early-modern print culture, fellow is most similar to rascal and rogue.

Our preliminary results indicate that the sociolegal evolution of English society was not nearly as linear as Maine would have predicted. Moreover, there were considerable differences between caselaw and print culture in the temporal evolution of the emphasis on status versus contract.

5. A NECESSARY DIGRESSION: A DIFFERENT METHODOLOGICAL PERSPECTIVE?

The previous sections have focused primarily on the new methodologies. The following section will discuss our combining of these methodologies with standard econometrics and the substantive findings that resulted. But before we move to that discussion, it is necessary to consider the overarching methodological paradigm that has been the natural outcome of our work.

The reader will notice that, so far, we have not talked about standard errors, hypothesis testing, and suchlike. We have been more concerned with how to expand the set of reliable and, hopefully, broadly accepted, historical facts. There are several reasons for this.

First, the processing of texts brings a whole new range of information into easy view. After all, where, for example, could one ever hope to obtain a timeline depicting the way in which lusty entertainments replaced historical romance within the production of English culture (Grajzl and Murrell 2023d)? The questions of what and when come before the question of why.

Second, the hypothetic-deductive methodological straitjacket—theory, hypothesis, and one statistical test—fits awkwardly with an exercise in which data is being analyzed for the first time, where the most pressing task is to find out what could be in the data. At that stage, it is not clear what one should be theorizing upon. The isolation of interpretable facts from raw text data is itself a creative enterprise in which discovery leads to new ideas that in turn could be fashioned into a new theory. Induction seems much more apposite in providing a methodological underpinning to this exercise than the hypothetico-deductive method.

Third, the very methods of ML have not been shaped by those who prioritize the hypothetic-deductive methodology. Instead, the primary objectives for developers of ML tools have been prediction and summarization. Therefore, the set of easily implementable routines—for those of us who are not computer scientists, with limited computer power, and without an army of research assistants—does not usually focus on standard errors of output. Perhaps this will change, but we do not think that will be very soon. Instead, for the type of work that we do, the original output of an ML exercise—say, the document-topic matrix—is treated as if it were a set of solid facts.

Notably, this is exactly the way national income numbers are used in an enormous number of econometric exercises.[2]

Fourth, the hypothetic-deductive methodology is impossible to implement when one is trying to make sense of new data types. We confronted this in our study of English law. A particularly clear example was the presence of the idea of assumpsit in English law, essentially the circumstances under which a debt could be assumed to arise after the non-fulfillment of a contractual promise. The usual narrative has centered on a long struggle between courts, then a decisive case very early in the 17th century, and acceptance thereafter. But we found a timeline of assumpsit's appearance in English law as an inverted-U with a peak at 1635, indicating a significant rise before 1600, followed by a fall to insignificance as the 17th century proceeded. Yet we knew that the idea of assumpsit was universally accepted as the 17th century drew to a close. We then had to build a model to understand why the topic-model output could look this way. The theory followed from insights provided by the data, rather than preceding the empirical analysis. This is the very essence of induction. Our model incorporated the evident notion that ideas are most discussed when they are in contention or being spread to new situations. When they are broadly accepted, they do not appear in discussion.

Thus, to us at the moment, the most productive use of ML in historical institutional and cultural research lies in adding to, or modifying, the set of facts that are readily accepted as characterizing a historical period. So just as historians are no longer prone to debating whether Henry VIII had five or seven wives, with the accepted answer even being the cornerstone of a Broadway musical, we would hope that new implementations of ML establish similar facts with regard to institutional and cultural history—facts that, eventually, do not need further questioning. For example, in our work, we show a change towards a less combative tone in culture during the 16th and 17th centuries: English religious and political discourse became less antagonistic and more scholarly in tone, while authority relationships became less important, literature more playful, and economic topics more prominent. Similarly, we establish that the chief legal ingredients of a financial revolution were present in caselaw well before the Glorious Revolution, the political event in 1688 that institutional scholars previously long perceived as critical to subsequent financial development.

Is this a radically new way of thinking about historical research? We think not. Rather, it just describes candidly the way things have always been done. We believe that much research that is framed in hypothetico-deductive terms is really inductivist in disguise. Consider digging only slightly below the surface of virtually any applied econometrics exercise demonstrating the presence of a hypothesized effect and produced by any researcher that you truly admire and whom you believe approaches research with the highest levels of integrity. You would invariably encounter a process in which the hypothesis has been shaped by perusals of the data, in which various approaches have been followed and discarded, and in which the results that have been

[2] And indeed, it is exactly in this way that researchers treat a variable they have derived from a ML process and fitted into a regression. There is little acknowledged recognition that a value of this variable would be best represented by a distribution, rather than a single number.

chosen for presentation are those viewed most representative of the many produced. It is induction, or perhaps more properly abduction, under the cloak of hypothetico-deductive.

But ardent hypothetico-deductivists might object that their method is always about testing general rules, ones that will inform about something more than the events in the particular time and place characterized by the data. But this is not usually the sphere of historical research, where data are scarce and have to be constructed using myriad assumptions. Historical research is primarily about time and place where omitted variables and data-selection issues are legion. In that case, the hypothetico-deductivist method in history boils down to isolating specific facts, rather than discovering general broadly applicable rules.

There is another sense in which we believe that ML historical research is very similar to the ways things have always been done. Humans and machines do the same thing: topic modeling. Informal topic-modeling exercises have been carried out for centuries by learned scholars reading many texts, finding commonalities between those texts, and interpreting the underlying ideas. That is after all how McCloskey (2016) discovered a Bourgeois Civilization in Western Europe; how Skinner (1965) found the idea of a Norman Non-Conquest in 16th- to 18th-century England; how Hirschman (1977) saw the Interests taming the Passions at the same time; and how Zweigert and Kötz (1992) were able to find empiricism in common law and abstraction in continental law.

We have stated these points bluntly, because we feel they need to be stated so. There seem to be vast terrains of economics where these points are not recognized, at least openly. But as ML has become more common in the social sciences there is more acknowledgement of these points. Kahneman (2019: 609) notes that when applying ML to large datasets, "you will find out much more than your theory is designed to explain", which in turn allows for the possibility that "machine learning can be a source of hypotheses". 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" (Grimmer et al. 2021: 396). 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…" (DiMaggio et al. 2015: 4). 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" (Nelson 2021: 2). And even in economics, "In many applications of topic models the goal is to provide an intuitive description of text rather than inference…Interpretation or story building…tends to be a major focus for topic models and other unsupervised generative models" (Gentzkow et al. 2019: 556).

6. FROM MACHINE LEARNING TO ECONOMETRICS

As we have commented above, ML analysis of texts reduces data dimensionality. For example, our application of topic modeling to a pre-1765 ER corpus summarizes 31 million words with a

52,949-by-100 document-topic matrix. The resulting estimates can then be productively employed in subsequent econometric analysis.

The existing literature using ML for text-as-data in historical-institutional contexts has, by and large, focused on extracting from the ML estimates a single variable to solve a specific data-econometric problem. For example, Bi and Traum (2019) explore how U.S. government bond prices in the 1840s were affected by newspaper reporting. Upon applying clustering and topic modeling to the corpus of newspaper articles for each state, the authors construct a state-specific, time-varying index of state legislative activities and fiscal actions. The resulting variable is then used as a key determinant of bond prices. Similarly, McCannon and Porreca (2023) apply topic modeling to the Old Bailey records on criminal trials in 19th-century London to estimate the attention paid to a topic that is interpreted as reflecting the professionalization of the court. They then use the attention to the pertinent topic as an outcome variable in a difference-in-differences framework where the treatment of interest is a change in law that introduced the right to counsel.

The focus on a single variable, as in the above studies, is usually driven by the hypothetico-deductive method. It has a laser-like focus on constructing a new variable rather than leveraging the whole set of dimensionality-reduced data produced via ML. In these types of studies, the ML methods are handmaids to existing approaches: they do not form the bedrock upon which the projects are built.

In contrast, when the goal is to paint a broader empirical picture of the historical roots, flow, and repercussions of ideas about institutions and culture, the use of ML tools will, by necessity, form the very foundation on which a project is built. In our view, it is this conception that offers the opportunity for an especially productive combination of ML and statistical-econometric methods. In what follows, we offer examples from our own research.

Example 1: Uncovering the Legal Legacies of Early English Caselaw

The path-dependence of caselaw development is broadly acknowledged by legal historians. But which early-modern legal ideas were especially relevant for later caselaw development during the Industrial Revolution? We addressed this core question about the evolution of English law in Grajzl and Murrell (2022a). In conducting the analysis, our unit of observation was a pre-1765 document (a case). Our dependent variable was post-1764 citations to pre-1765 cases (Murrell 2021). A citation reflects the precedential relevance of the cited case to the citing case, and thus citations to case reports are a quantitative measure of the influence of those case reports on the development of the law.

We used a standard econometric model for count data, explaining post-1764 citations to pre-1765 cases with the characteristics of the pre-1765 cases. Under the reasonable assumption that all aspects of a prior case could be an input into future deliberations, and thus decisions, the number of explanatory variables was determined by our ML exercise on the corpus of pre-1765 cases, as many as 100 given that exercise. We therefore encountered a methodological challenge for which the use of ML has been strongly advocated: perhaps we, too, should apply lasso to reduce the number of predictors (see, e.g., Michalopoulos and Xue 2021). But, in fact, our number of observations (documents in a particular post-1764 period) was large enough that we could easily

estimate all of the desired parameters. Using lasso would have unduly restricted the substantive scope of our analysis, thereby failing to provide a comprehensive overview of the early-modern legal origins of later legal development.

In substantive terms, we found that the ideas (i.e., topics) having the strongest effects were precedent-based thought and legal reporting in the style of Edward Coke, one of the preeminent makers of English law. The primary legacy of early English caselaw, therefore, lay in bestowing modes of reasoning. The other side of the coin was that a readily identified subset of preexisting legal ideas played no discernible role in caselaw development during the Industrial Revolution. For example, questions of jurisdiction—which courts had which authority—exhibited the weakest effect of all the sets of legal ideas.

We then investigated why a subset of preexisting legal ideas played no discernible role in subsequent caselaw development during the Industrial Revolution. Is it that those legal ideas became so widely accepted within the legal profession by the late 18th century that, even though they were relevant, legal professionals using them in the later era no longer felt the need to cite specific cases (hypothesis A)? Or is it that those legal ideas were generally less applicable in the late 18th and 19th century (hypothesis B)? The essence of our approach in distinguishing between these two possibilities rests on using a measure of the degree to which the corresponding legal ideas had been settled in pre-1765 caselaw. If hypothesis A is true, we would expect this pre-1765 measure to be negatively related with whether the ideas exert a detectable effect post-1764. In contrast, if hypothesis B is true, we would not expect to observe a systematic relationship between this pre-1765 measure and an indicator for whether the applicable ideas exert a detectable effect post-1764. Our evidence is consistent with hypothesis B. Thus, the reason why certain preexisting legal ideas do not exert a detectable effect is that those ideas were generally no longer key to post-1764 legal disputes.

As we emphasize in Grajzl and Murrell (2022a), our estimates of the later effects of earlier legal ideas provide a picture of the relevance of early English law for caselaw development in the Industrial Revolution that has never been provided before in such a systematic quantitative, comprehensive fashion. Our approach to investigating legal development could be applied in many other contexts.

Example 2: Assessing the Effects of Caselaw on Economic Performance

England's Industrial Revolution was an epoch-defining event. But just how important was the law emanating from English courts in facilitating industrialization? We tackled this pivotal, and previously unanswered, question in Grajzl and Murrell (2023c). Methodologically, we combined ML with conventional time-series econometrics. In particular, given the inherently intertwined nature of legal and economic development, we used vector autoregression (VAR). In modern macroeconomics, VARs are used extensively to characterize the dynamics of multiple time series and to conduct inference in settings that are fraught with endogeneity but lack natural experiments of the kind usually exploited by empirical microeconomists. As such, VARs have often been used in economic history to examine the interaction among socioeconomic variables.

We do not view the application of VAR as in any way superior to analyses employing natural experiments. Natural experiments provide incisive lessons when a suitable truly exogenous variable is available. But, as we argue below, in earlier historical periods, and particularly when examining broad institutional and cultural data, the required truly exogenous variable to proxy an appropriate shock to the system is usually unavailable. In such cases, VARs can still provide valuable insights.

In one such exercise, examining the ER cases from 1765 to 1865, we first estimated a 105-topic model. With our intended unit of analysis being England in a particular year, it would have not been feasible to employ as many as 105 variables when estimating a VAR. Therefore, we used the intermediate level of analysis discussed in Section 4. Using the estimated document-topic prevalence matrix and the year of each case, we aggregated the topics to produce annual time series of attention to 13 distinct themes. Examples of such themes are procedure and reasoning, contract, debt and finance, and inheritance. Thus, a 67,455-by-105 document-topic proportions matrix was replaced in the analysis by a 67,455-by-13 document-theme-proportions matrix.

We then aggregated the document-level theme-proportions at the yearly level, generating 13 caselaw time series of per-capita annual attention of the ER to the pertinent themes. Crucially, and as indicated in Section 4, a simple model of cultural diffusion allows one to conclude that the attention paid to any theme in a given year reflects the amount of change in adherence to the corresponding legal ideas in that year. Therefore, our time series also measure the intensity of development in the pertinent areas of law. Finally, we augmented the 13 caselaw series with a real per-capita GDP series.

We were focused on examining the interaction over time of our caselaw variables and real per-capita GDP. These variables coevolve, each potentially affecting each other, most likely with a lag, but possibly contemporaneously. Those conditioned by standard microeconometrics will immediately think of unaccounted-for factors. These would be of three broad types: proximate causes of per-capita GDP, secular trends, and exogenous shocks. Proximate causes, for example capital accumulation, can be omitted if one is clear that the analysis then focuses on caselaw as a fundamental legal-institutional driver of economic performance. Secular trends are included via a linear time trend. Idiosyncratic (exogenous) shocks, unrelated to both past developments in caselaw and secular trends, are subsumed in error terms. Given this view of the world, one has a structure that exactly corresponds to a standard VAR.

However, as is well known, extracting the maximum amount of information from such a structure requires incorporating information to aid identification. Our goal was to offer a comparative account of the importance of different caselaw domains—13 in total—for England's economic performance. There was no possibility of finding a source of exogenous variation in each of these areas of law. Ours is therefore a very different empirical context than those encountered in a subset of recent studies that also employ historical time-series data but are interested in identifying the effects of one specific exogenous variable (e.g., Jordà et al. 2022, Palma 2022). We thereby followed the standard approach employed by macroeconomists, using a recursive identification scheme naturally implied by our legal-institutional setting. This

identification method effectively implies making assumptions on which variables cannot affect which other variables simultaneously.

Our estimates reveal that caselaw developments were a crucial determinant of economic fluctuations. Caselaw shocks together accounted for more of the variability in per-capita GDP around its long-term trend than did shocks directly to per-capita GDP. The direction of the response of per-capita GDP to caselaw innovations, however, critically depended on the legal domain. Developments in caselaw on intellectual property, organizations, debt and finance, and inheritance boosted economic performance. For instance, during this era, court decisions were critical in stimulating patenting, facilitating incorporation, resolving insolvency, and disambiguating wills, all developments that would have stimulated the supply and reallocation of capital. In contrast, developments in property and ecclesiastical caselaw exerted a negative effect on per-capita GDP. For example, 18th-century developments in rules on tithes, which remained in the domain of ecclesiastical courts, disincentivized capital accumulation.

Interestingly, our analysis identified an era when the legal system evidenced reduced attention to output-fostering areas of law and increased attention to output-hindering. We named this the 'bleak-law era', depicted in Dickens' *Bleak House*. This was also the time when Bentham was making his thoroughgoing critiques of English legal processes.

Similarly, in Grajzl and Murrell (2023a) we investigated if caselaw developments impacted English economic development in the preindustrial era. Our dataset used variables standard in the Malthusian framework (real per-capita income and vital rates) alongside three yearly time-series of legal themes, that is, aggregates of legal topics that were estimated using the pre-1765 ER corpus. These three naturally fit into the Malthusian context: caselaw on land, inheritance, and families. Using a similar methodology to that in Grajzl and Murrell (2023c) described above, we estimated a VAR on the resulting dataset. The results show that preindustrial economic development was profoundly shaped by caselaw developments, with caselaw on families and inheritance being especially impactful. In the preindustrial era, developments in inheritance and family caselaw, for example, facilitated greater access to capital and enhanced financial security for wives, widows, younger sons, and daughters, reduced transaction costs within the family, and facilitated productive matching in the marriage market.

Example 3: Estimating the Coevolution of Ideas

Our exploration of the ER and EEBO-TCP corpora using the combination of ML and econometrics has allowed us to generate a rich set of entirely novel insights about the historical evolution of institutions and cultures. Among these, we have especially focused on eliciting patterns in the coevolution of ideas. In the legal context, for example, scholars have long been pondering the question of the degree of autonomy of legal processes (Tomlins 2007). Similarly, cultural historians have been debating the influence of religion on scientific thought (see, e.g., Merton 1938, Wootton 2015). Our application of topic modeling and time-series methods casts systematic quantitative light on these issues in the context of early-modern England.

In Grajzl and Murrell (2022c), we explored the coevolution of legal ideas on property, contract, and procedure, the three core elements of English caselaw. Once more we used topic-modeling on

the pre-industrial part of the ER corpus (Grajzl and Murrell 2021b, 2021c) to construct yearly measures of attention to the three legal domains. Our VAR approach then directly implies two distinct conceptualizations of legal autonomy: vis-à-vis the rest of the legal system and vis-à-vis societal factors fully external to the legal system. The main insight generated by our empirical analysis is that the internal dynamics of the legal system explain much less of the development of caselaw on contract than that on property and procedure. Contract caselaw was more buffeted by unusual events occurring outside the normal functioning of the system of litigation (e.g., civil conflict, passage of important legislation, ascent of an especially innovative judge). Thus, in comparison with property and procedure, contract development was relatively more autonomous from the internal dynamics of the legal system, but relatively less autonomous from society.

In Grajzl and Murrell (2023d) we turned our attention to culture. Upon estimating a 110-topic model using the TCP corpus, we used multivariate time-series methods analogous to those in Grajzl and Murrell (2022c, 2023a, 2023c), casting light on the coevolution of ideas within three broad themes: religion, science, and institutions. The question of how developments in these three areas affected each other is a staple of historical inquiry. Our analysis reveals, for example, that innovations in religious ideas induced strong responses in both science and institutions, especially at times when Puritanism was prominent in religious discourse. Thus, the religion-to-science link was strong already in the second half of the 16th century, much earlier than stressed by previous scholars (Merton 1938: 414-416; Mokyr 2016). Again, as emphasized in Section 4, our analysis seems to identify significant events taking place earlier than is the prevailing consensus in the literature.

Finally, two episodes that have received great attention from historians and social scientists interested in institutional change, the Civil War and the Glorious Revolution, did not lead to an unusually large amount of attention in print culture to matters pertaining to institutional development. Rather, cultural debates about institutions preceded, and thus perhaps helped spur, each of those revolutions.

Example 4: Finding Critical Junctures

As we have commented in Section 4, something remarkable happened in England during 1550-1750, and one of the enduring puzzles of history is what that was and when it happened. Grajzl and Murrell (2023b) addresses this puzzle directly. In the age of two major political revolutions, this paper looks for quiet revolutions—periods of fundamental institutional and cultural shifts that do not occur via force nor have clear beginning and ending dates. To identify quiet revolutions, we apply the econometrics of unknown structural breaks to each of the time-series of 100 caselaw topics and 110 culture topics discussed above. The insight that drove this particular choice of econometric tool is that attention to a topic rises steeply (i.e., features a structural break) at the beginning of a conceptual revolution, as proponents of new ideas preach their gospels to the unconverted. But one such upturn does not identify a quiet revolution; rather, a quiet revolution is characterized by a cluster of upturns for similar topics in a short time period.

The results can be conveyed in two evocative figures, showing when England experienced change and in which areas of ideas. The visual exposition of the results is a crucial part of the

analysis; without it, the reader would be left at the half-way stage of verbal summary of 210 facts isolated from each other. The results show that early-modern England featured several profound conceptual revolutions. The financial revolution began by 1660. Significant changes in land law began during the Interregnum. Caselaw relating to families began to change significantly in the late 17th century. Elizabethan times saw dramatic change in the emphasis on basic skills. At the same time, but perhaps not coincidently, a Puritan revolution affected both theology and political ideas. An upturn in dissent preceded the mid-17th century revolution and Civil War.

7. CONCLUSION

The machine learning of history is not normal science in the sense articulated by Kuhn. When implementing the core methods, the beginning researcher cannot fall back on decades of scholarship, which usually guides decisions on the choice of many assumptions in contemporaneous research. There are not standard tests produced by tried and true methods, which usually dictate the acceptability of a piece of research in the relevant scientific community. The machine learning of history will therefore involve invoking new types of assumptions or combining old ones in unusual recombinations. Inevitably, these new ways of approaching research will collide with the normal-science paradigm expected of contributors to a particular discipline. We ourselves have received bemusing criticisms of the type of research that we are currently undertaking, some with phrasing that seemed to indicate we had breached some unwritten code of conduct. Our answers to these critics appear in the previous pages, an integrated body of findings, developed using rigorous, transparent, and replicable methods, adding fundamental insights about English legal and cultural history.

When powerful new tools drive radical changes in methods, the research community enters the early stages of a multi-strategy, multi-period, coordination game, moving hesitantly toward an equilibrium set of procedures that are widely spread and broadly understood. Eventually, a workable equilibrium will be found, establishing a whole new set of standard assumptions and practices that will guide a large part of the research. The equilibrium is Kuhn's normal science. But costs are inevitable before that coordination equilibrium is found.

The process of reaching an equilibrium of a coordination game involves competitive elements: researchers have an incentive to develop approaches that they hope will one day become standard. In our reading of the literature, we have found that this seems to provide too many incentives to do things differently from other researchers. When there is no set of broadly accepted norms it is too tempting to construct ad hoc embellishments of techniques. These make it harder to interpret results across studies on similar subjects. These ad hoc elements of new studies also make it more difficult for a beginning researcher to understand which are the most productive ways of proceeding in a new research effort.

These observations have guided the way that we have pursued the research summarized in the previous pages. We are not interested in technique, per se, but have been motivated by two simple facts: texts provide a large proportion of historical data; and computational power and algorithmic development has proceeded far enough to make the processing and summarization of those texts

possible for all researchers. There is now no need for the researcher interested primarily in the substance of history to engage in ad hoc development of new methods. This straightforward approach is exemplified in our own work on novelty, where we rely on standard measures of the similarity of texts.

However, the text processing and summarization does present distinctive challenges. A beginning one is that the production of text-data suitable for machine learning will rest on bespoke databases that depend highly on the historical context. In our case, this involved understanding the nuances of 17th-century English. Nevertheless, as we showed in discussing our use of word embeddings, machine learning itself can provide tools to refine the accuracy of those databases.

Next is the problem of the sheer number of variables produced by the summarization of a large corpus of texts. After all, a useful cataloging of the fundamental ideas in English 17th-century law, for example, can be expected to comprise many dozens of variables. This, above all, suggests that descriptive analyses will provide a core output of any machine-learning exercise. Those analyses will require a thorough understanding of context, as in our discussion of developments in English law and culture. Moreover, in order to evocatively convey the lessons of, perhaps, hundreds of stark facts, the researcher will have to develop new types of figures (as we did in studying quiet revolutions) or find ways of conveying the results of many estimated parameters (as we did in examining the legal legacies of early caselaw).

The fact that standard econometric methods are most useful when the focus is on a limited set of variables presents another problem. This means that the researcher will usually need to further aggregate the results of the summarization (as when we produced themes from topics for both our legal and cultural databases). But then a further challenge often arises: the text summarization might produce a set of variables that are all endogenous, with no obvious exogenous explanatory variables to aid identification. This will require the use of techniques such as VAR together with an approach to identification that relies on broad theorizing rather than the adding of data that provides exogenous information (as when we explored the role of law in the Malthusian framework or studied the interaction of religion and science). But this use of macro empirical tools that does not rely on natural experiments leveraging exogenous variation can fit uncomfortably with the perspectives of those scholars who are only willing to embrace the prevailing treatment-effect paradigm that nowadays dominates microeconomic empirical research.

In conclusion, machine learning applied to historical texts provides a whole new analytical toolkit facilitating insight into the historical roots of institutions (e.g., caselaw) and culture (e.g., theology), the study of which has previously relied predominantly on the reading of a limited number of texts. As we have made clear, the potential is enormous. But the beginning researcher considering using these tools needs to reflect on the fact that they will be working outside the framework of Kuhn's normal science. Then, pre-conceived ideas about what constitutes publishable research can affect evaluations in ways that are beyond the examination of rigor and quality. We have provided one methodological model for such research by describing the processes and products of our own work on English history. It is work that combines new machine learning methods with econometric techniques that have already been tried and tested, with the primary

objective being the production of new substantive insights. We believe that following this model offers the opportunity to expand knowledge dramatically in a broad swathe of economic, legal, cultural, and intellectual history.

REFERENCES

Almelhem, Ali, Murat Iyigun, Austin Kennedy, and Jared Rubin. 2023. "Enlightenment Ideals and Belief in Science in the Run-up to the Industrial Revolution: A Textual Analysis." IZA Discussion Paper No. 16674.

Ash, Elliott, Daniel L. Chen, and Arianna Ornaghi. 2024. "Gender Attitudes in the Judiciary: Evidence from US Circuit Courts." *American Economic Journal: Applied Economics* 16(1): 314-350.

Athey, Susan and Guido W. Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics* 11: 685-725.

Barber, Luke, Michael Jetter, and Tim Krieger. 2023. "Foreshadowing Mars: Religiosity and Pre-Enlightenment Warfare." CESifo Working Paper No. 10806

Barron, Alexander T.J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences* 115(18): 4607-4612.

Bi, Huixin and Nora Traum. 2019. "Sovereign Risk and Fiscal Information: A Look at The U.S. State Default of the 1840s." Federal Reserve Bank of Kansas City working paper no. 19-04.

Blaydes, Lisa, Justin Grimmer, and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80(4): 1150-1167.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley E.B. Wilbert. 2023. "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project." NBER Working Paper No. 31671.

Burkov, Andriy. 2019. *The Hundred-Page Machine Learning Book*. Quebec City: Andriy Burkov.

Clark, Gregory. 2023. "The Inheritance of Social Status: England, 1600 to 2022." *Proceedings of the National Academy of Sciences* 120(27): e2300926120.

DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* 2(2): 1-5.

Dittmar, Jeremiah and Skipper Seabold. 2019. "New Media and Competition: Printing and Europe's Transformation After Gutenberg." CEP Discussion Paper No 1600.

Düben, Christian and Melanie Krause. 2023. "The Emperor's Geography – City Locations, Nature, and Institutional Optimisation." *Economic Journal* 133(651): 1067-1105.

Funk, Kellen and Lincoln A. Mullen. 2018. "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice." *The American Historical Review* 123(1): 132-164.

Gavin, Michael. 2016. "Historical Text Networks: The Sociology of Early English Criticism." *Eighteenth-Century Studies* 50(1): 53-80.

Gavin, Michael. 2017. “How to Think about EEBO.” *Textual Cultures* 11(1/2): 70-105.

Gennaro, Gloria and Elliott Ash. 2022. “Emotion and Reason in Political Language.” *Economic Journal* 132(643): 1037-1059.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature* 57(3): 535-74.

Grajzl, Peter and Peter Murrell. 2019. “Toward Understanding 17th Century English Culture: A Structural Topic Model of Francis Bacon’s Ideas.” *Journal of Comparative Economics* 47(1): 111-135.

Grajzl, Peter and Peter Murrell. 2021a. “Characterizing a Legal-Intellectual Culture: Bacon, Coke, and Seventeenth-Century England.” *Cliometrica* 15(1): 43-88.

Grajzl, Peter and Peter Murrell. 2021b. “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(1): 1-19.

Grajzl, Peter and Peter Murrell. 2021c. “A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution II: Applications.” *Journal of Institutional Economics* 17(2): 201-216.

Grajzl, Peter and Peter Murrell. 2022a. “Lasting Legal Legacies: Early English Legal Ideas and Later Caselaw Development During the Industrial Revolution.” *Review of Law & Economics* 18(1):85-141.

Grajzl, Peter and Peter Murrell. 2022b. “Using Topic-Modeling in Legal History, with an Application to Pre-Industrial English Caselaw on Finance.” *Law and History Review* 40(2): 189-228.

Grajzl, Peter and Peter Murrell. 2022c. “A Macrohistory of Legal Evolution and Coevolution: Property, Procedure, and Contract in Early-Modern English Caselaw.” *International Review of Law and Economics* 73: 106113.

Grajzl, Peter and Peter Murrell. 2023a. “Of Families and Inheritance: Law and Development in England Before the Industrial Revolution.” *Cliometrica* 17(3): 387-432.

Grajzl, Peter and Peter Murrell. 2023b. “Quiet Revolutions in Early-Modern England.” *Public Choice*, forthcoming.

Grajzl, Peter and Peter Murrell. 2023c. “Caselaw and England’s Economic Performance During the Industrial Revolution: Data and Evidence.” *Journal of Comparative Economics*, forthcoming.

Grajzl, Peter and Peter Murrell. 2023d. “A Macroscope of English Print Culture, 1530-1700, Applied to the Coevolution of Ideas on Religion, Science, and Institutions.” SSRN working paper.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021a. “Machine Learning for Social Science: An Agnostic Approach.” *Annual Review of Political Science* 24: 395-419.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021b. *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton, NJ: Princeton University Press.

Guldi, Jo. 2022. “The Algorithm: Mapping Long-Term Trends and Short-Term Change at Multiple Scales of Time Get access Arrow.” *The American Historical Review* 127(2): 895-911.

Hirschman, Albert O. 1977. *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph*. Princeton, NJ: Princeton University Press.

Hitchcock, Tim, Robert Shoemaker, Clive Emsley, Sharon Howard, and Jamie McLaughlin, et al., The Old Bailey Proceedings Online, 1674-1913. www.oldbaileyonline.org, version 9.0, Autumn 2023.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. *An Introduction to Statistical Learning with Applications in R*. Second edition. New York, NY: Springer.

Jordà, Òscar, Sanjay R. Singh, and Alan M. Taylor. 2022. “Longer-Run Economic Consequences of Pandemics.” *Review of Economics and Statistics* 104(1): 166-175.

Kahneman, Daniel. 2019. “Comment on Artificial Intelligence and Behavioral Economics.” In: Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, Chicago, IL: The University of Chicago Press, pp. 608-610.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. “Measuring Technological Innovation over the Long Run.” *American Economic Review: Insights* 3(3): 303-320.

Law, David S. 2019. “Constitutional Dialects: The Language of Transnational Legal Orders.” In: Shaffer, Gregory, Tom Ginsburg, and Terence C. Halliday (Eds.), *Constitution-Making and Transnational Legal Order*. Cambridge, UK: Cambridge University Press.

LIWC. 2022. https://www.liwc.app/

Ma, Lin and Monica Li. 2020. “What Helped Officials of Song Dynasty in Climbing the Greasy Pole: An Empirical Study.” SSRN working paper.

McCannon, Bryan C. and Zachary Porreca. 2023. “The Right to Counsel: Criminal Prosecution in 19th Century London.” SSRN working paper.

McCloskey, Deirdre N. 2016. *Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World*. Chicago, IL: University of Chicago Press.

McMahan, Peter and Daniel A. McFarland. 2021. “Creative Destruction: The Structural Consequences of Scientific Curation.” *American Sociological Review* 86(2): 341-376.

Merton, Robert K. 1938. “Science, Technology and Society in Seventeenth Century England.” *Osiris* 4: 360-632.

Michalopoulos, Stelios and Melanie M. Xue. 2021. “Folklore.” *Quarterly Journal of Economics* 136(4): 1993-2046.

Miller, Ian M. 2013. “Rebellion, Crime and Violence in Qing China, 1722-1911: A Topic Modeling Approach.” *Poetics* 41(6): 626-649.

Mokyr, Joel. 2016. *A Culture of Growth: The Origins of the Modern Economy*. Princeton, NJ: Princeton University Press.

Murrell, Peter. 2021. “Did the Independence of Judges Reduce Legal Development in England, 1600-1800?” *Journal of Law and Economics* 64(3): 539-565.

Nelson, Laura K. 2021. “Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South.” *Poetics* 88: 101539.

Newman, David J. and Sharon Block. 2006. “Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper.” *Journal of the American Society for Information Science and Technology* 57(6): 753-767.

Pagé-Perron, Émilie. 2018. “Network Analysis for Reproducible Research on Large Administrative Cuneiform Corpora.” In: Bigot Juloux, Vanessa, Amy Rebecca Gansell, and Alessandro di Ludovico (Eds.), *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, pp. 194-223. Brill.

Palma, Nuno. 2022. “The Real Effects of Monetary Expansions: Evidence from a Large-Scale Historical Experiment.” *Review of Economic Studies* 89(3): 1593-1627.

Perrin, Faustine. 2022. “On the Origins of the Demographic Transition: Rethinking the European Marriage Pattern.” *Cliometrica* 16(3): 431-475.

Poulos, Jason. 2019. “Land Lotteries, Long-Term Wealth, and Political Selection.” *Public Choice* 178(1): 217-230.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. “Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set.” *Explorations in Economic History* 80(C): 101391.

Prüfer, Jens and Patricia Prüfer. 2020. “Data Science for Entrepreneurship Research: Studying Demand Dynamics for Entrepreneurial Skills in the Netherlands.” *Small Business Economics* 55(3): 651-672.

Renton, A.W. 1900-1932. *The English Reports. Great Britain. Parliament. House of Lords.* Edinburgh, UK: W. Green & Sons.

Rice, Douglas R. and Christopher Zorn. 2021. “Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies.” *Political Science Research and Methods* 9(1): 20-35.

Rice, Douglas. 2019. “Measuring the Issue Content of Supreme Court Opinions.” *Journal of Law and Courts* 7(1): 107-127.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. “A Model of Text for Experimentation in the Social Sciences.” *Journal of the American Statistical Association* 111(515): 988-1003.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. “Structural Topic Models for Open Ended Survey Responses.” *American Journal of Political Science* 58(4): 1064-1082.

Rockmore, Daniel N., Chen Fang, Nicholas J. Foti, Tom Ginsburg, and David C. Krakauer. 2018. “The Cultural Evolution of National Constitutions.” *Journal of the Association for Information Science and Technology* 69(3): 483-494.

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. 2015. “Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790-2014.” *Proceedings of the National Academy of Sciences* 112(35): 10837-10844.

Skinner, Quentin. 1965. “History and Ideology in the English Revolution.” *The Historical Journal* 8(2): 151-178.

The Text Creation Partnership (TCP). 2022. URL: https://textcreationpartnership.org.

Tomlins, Christopher. 2007. “How Autonomous Is Law?” *Annual Review of Law and Social Science* 3: 45-68.

van Vugt, Ingeborg. 2022. “Networking in the Republic of Letters: Magliabechi and the Dutch Republic.” *Journal of Interdisciplinary History* 53(1): 117-141.

Wootton, David. 2015. *The Invention of Science: A New History of the Scientific Revolution*. New York, NY: Harper. Kindle Edition.

Zweigert, Konrad and Hein Kötz. 1992. *Introduction to Comparative Law*. Second revised edition, translated by Tony Weir. Oxford, UK: Clarendon Press.

Table 1: Examples of articles using ML applied to institutions and culture in historical settings

Part 1: Addressing challenges in conventional historical research: Data collection
- Barber et al. (2023): pre-enlightenment Europe, historical data on religiosity
- Buckles et al. (2023): historical U.S. censuses, record matching
- Clark (2023): England 1600 to present, data on lineage of English people
- Dittmar and Seabold (2019): Europe 1456-1600, data on books
- Hitchcock et al. (2023): English criminal trial records, 1674-1913
- Pagé-Perron (2018): Mesopotamia, cuneiform corpus
- Poulos (2019): 19th-century Georgia, record matching

Part 2: Addressing challenges in conventional historical research: Data analysis
- Düben and Krause (2023): cities in imperial China, random forest
- Ma and Li (2022): government bureaucracy in 13th century China, random forests
- Michalopoulos and Xu (2021): folklore and present-day beliefs, lasso
- Perrin (2022): demographic transition in 18th and 19th century France, clustering

Part 3: Charting new research avenues: Leveraging text as data
- Almelhem et al. (2023): English print texts 1500-1900, topic modeling
- Barron et al. (2018): French Revolution parliamentary assembly speeches, topic modeling
- Bi and Traum (2019) U.S. newspapers 1840s, clustering and topic modeling
- Blaydes et al. (2018): medieval texts, topic modeling
- Funk and Mullen (2018): 19th-century US, similarity measures
- Gavin (2016): early-modern English texts, networks
- Gennaro and Ash (2022): U.S. Congress speeches, 1858 onwards
- Guldi (2022): Britain, similarity measures
- Law (2019): 19th and 20th century national constitutions, topic modeling
- McCannon and Porreca (2023): England, criminal trial records 19th-century, topic modeling
- Miller (2013): late imperial China administrative records, topic modeling
- Newman and Block (2006): 18th-century US newspaper, topic modeling
- Rice (2019): 19th and 20th century US Supreme Court opinions, topic modeling
- Rockmore et al. (2018): 18th to 20th century national constitutions, topic modeling
- Rule et al. (2015): US State of the Union addresses, miscellaneous
- Van Vugt (2022): letters in the Republic of Letters, networks

Notes: The table provides a guide to a sampling of the scattered literature on use of ML applied to institutions and culture in historical settings. Under each part, entries are listed alphabetically. For each entry, we provide brief information on the examined historical context and illustrative use of ML.