An essay prepared for the Third Edition of the *Handbook of Cliometrics* published by Springer Nature, edited by Claude Diebolt and Michael Haupert.

How Machine Learning Will Change Cliometrics

Peter Grajzl, Peter Murrell

July 2023

Abstract

Machine learning (ML), it is sometimes claimed, will change the world. In this chapter, we argue that ML does offer great potential for advancing the field of cliometrics. We first attempt to demystify ML by describing its most widely-used methods and showing how these are often natural extensions of traditional approaches reshaped to take advantage of increases in computer power. We then proceed to discuss applications of ML in existing cliometric research. These fall into three categories. First, ML provides a tool for improving solutions to preexisting empirical problems, especially in causal identification. Second, ML can produce new representations of data that facilitate a fresh way of looking at the world with conventional cliometric techniques. Third, those new representations naturally lead to a quantitative approach to the inductive generation of new facts and theories. We suggest that this third category offers the most likely route by which ML will change the way historical research is done. To buttress this conclusion, we reflect on our own experience when employing ML methods in the study of English legal and cultural history. We discuss how ML allowed us to reimagine the way in which we conduct our research. We conjecture that ML will not only expand both the scope and breadth of cliometric research but also realign its orientation with what has always been one of the goals of historical research—to describe the ebb and flow of history.

Keywords: machine learning, cliometrics, causality, induction, England, legal history, culture

* We thank Norbert Hornstein and William Idsardi for very helpful comments over the years about the characteristics of ML when applied to language.

Introduction

Google Translate is arguably one of the great success stories of machine learning (ML). In late 2021, when having our first thoughts on this paper, we fed 'machine learning' into Translate and obtained the Latin 'apparatus doctrina'. Feeding that into Chinese gave '仪器学说', and then the Mongolian 'хэрэгсэл', and back into English, we obtained 'tools'. Was that a failure of Google Translate? No, it uncovered a deeper, unsought truth: ML is just a set of tools, often delivering insights that the user has not envisaged. As we will argue below, one of ML's most productive uses in cliometrics will be to let the data speak for itself, and to deliver new unexpected findings that uncover the flow of history. (We wrote this essay before ChatGPT became widely available. Extensive experience with that amazing tool has served only to reinforce this conclusion.)

Machine learning is in vogue. Athey (2019) argues that ML "will have a dramatic impact on the field of economics within a short time frame". But in what ways do ML tools differ from those employed in conventional cliometric research? Can ML augment the methodological toolkit currently available to cliometricians? Economists tend to emphasize the potential of ML to offer a more powerful means of conducting microeconomic causal inference, the holy grail of much empirical economics. But can ML also deliver new ways of generating quantitative insights, perhaps extending beyond a focus on causality?

These are the questions we address in the present chapter. We first give an overview of what we think ML is, demystifying it by arguing that it is a set of disparate methods that emerged from existing techniques in an age when increasing computer power changed the way that researchers developed new methods. We clarify the general ideas underlying ML and its relation to more traditional methods, defining specific terms and giving a bird's eye view of its most common methods. Throughout the discussion, our focus is on the intuitive and applied, drawing analogies with conventional cliometric techniques rather than discussing technical nuances.

We then provide an overview of the cliometric uses of ML aimed at solving empirical problems that were clearly understood before the advent of ML but whose solution really gained from the deployment of the extra computer power that ML leverages. Drawing on a series of cliometric studies that have employed ML in this vein, we argue that in these specific studies ML adds value by augmenting and finetuning empirical implementation. But the use of ML is hardly ever at the heart of the project's conception: the underlying research could have been conducted even without the use of novel ML techniques. For studies in this category, then, ML primarily replaces an army of research assistants (Athey 2019).

The next section, the fourth, explores what we view as the most exciting and promising set of developments in the application of ML to cliometrics. ML offers an entirely new lens for illuminating the flow of history. Unlike much of conventional cliometrics, grounded in the hypothetico-deductive reasoning that has come to dominate economics and related social sciences, a subset of ML tools favors the more inductive approach that has been used by historians as far back as Thucydides. One reason for this change in the tenor of quantitative research is that

increased computer power has opened up whole new vistas in the analysis of texts, a source of data that increasingly dominates the historical record as one recedes into the past. ML facilitates the quantification of large amounts of text, enabling the discovery of facts that would have otherwise remained buried in the ocean of the 'great unread'. Unexpected substantive insights are generated, which can in turn stimulate new theoretical conjectures. We suggest that ML's facilitation of an inductive approach will lead to ML changing historical research by as much as it will change research in any other discipline.

To illuminate this central point as well as to suggest a pathway into the use of ML for other researchers, we describe our own experience with applying ML to the study of English legal and cultural history. We highlight how and why we saw ML as providing a powerful tool for this endeavor, showing how it allowed us to reimagine our research. We provide examples of specific substantive questions, both descriptive and causal, that we have been able to address as a result of applying the new methods. We emphasize the tremendous potential of ML to generate novel representations of datasets that can be replicated, re-investigated, and further examined using conventional cliometric approaches. We illustrate the power of ML by briefly describing some of the central substantive findings that we generated on this journey. And we integrate broader reflections on the ways in which ML can aid cliometric research.

What Is Machine Learning?

One of our goals here is to demystify the concept of ML, to make it clear how the advent of ML reflects continuing progress in methodological development rather than a rupture that sets data analysis on a wholly new trajectory. The need to demystify is obvious when one observes the use of 'machine learning' in the titles and abstracts of papers intended to convey sophistication, allied with the intimidating thought that machines can really out-learn humans, a view promoted most notably by the artificial intelligence industry.[1] In fact, mention of the use of ML in describing the techniques to be used in research conveys about as much information as mention of 'econometrics' would do: very little. To proceed, we have to accomplish three tasks. First, describe what we think ML is. Second, define a set of terms that have come into common use with the advent of ML. Third, give a bird's-eye view of some of the most important methods of ML, focusing on the types of problems they solve.

General

Given the focus of this Handbook, it is befitting to begin with econometrics, which we view as the application of the insights of statistics for the purpose of estimating relationships between a set of dependent variables and some, manageable, number of variables that have been theorized to

[1] Consider this statement, probably aimed at methodologically naïve corporate executives: "The development of neural networks has been key to teaching computers to think and understand the world in the way we do, while retaining the innate advantages they hold over us such as speed, accuracy and lack of bias." What Is The Difference Between Artificial Intelligence And Machine Learning? Bernard Marr Dec 6, 2016, Forbes. accessed 11/4/2022. https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/?sh=54b420212742

provide underlying causes. Within econometrics in general, and this is going to be important later, latent variable techniques have been primarily used as a tool to justify a statistical framework for estimation: they are usually not viewed as being of independent interest themselves (e.g., in probit). What has driven econometrics over time is not the power of computation, but the power of the human mind to conceptualize complicated decision frameworks within simpler models that facilitate the estimation of core parameters of interest using limited amounts of data.

Around the turn of the century, the huge increase in computer power began to have significant effects. This happened primarily within computer science, outside social science and even statistics departments. Thus machine learning grew most prominently as a way of using the power of the computer to solve practical problems that often, although not necessarily, involved large data sets. A large number of new algorithms to solve these practical problems were constructed. The underlying conception of these algorithms did not incorporate much of the primary concern of economists, models that could tease out causality, nor, for the most part, on the primary concern of statisticians, inference. It is not surprising therefore that economics fell far behind many other disciplines in adopting these techniques. A notable contrast is with the study of literature, not a place that economists would immediately look to for new methodological developments.

The coining of the term machine learning is conventionally attributed to Samuel (1959) who showed that "a computer can be programmed so that it will learn to play a better game of checkers [draughts in English] than can be played by the person who wrote the program." A definition of machine learning that has taken hold and for which Samuel is often cited, but never wrote, is "the field of study that gives computers the ability to learn without explicitly being programmed". For those emphasizing that particular definition, what seems to be the essential feature of ML is that the program itself can generate information that will allow it to derive an improved solution to a problem. But the same description could be made of any numerical routine that operates in iterative fashion (e.g., Newton-Raphson). Moreover, the computer is explicitly programmed and there is much hands-on fine tuning that is necessary to get the computer to produce anything that would be deemed a satisfactory result. What is perhaps qualitatively different about ML is that less explicit theoretical structure needs to be embodied in the algorithm, primarily as a result of greater computer power and more data. Seen that way, the move from traditional econometrics to ML is analogous to the move in statistics from parameterized regression to non-parametric regression.

Perhaps the reason why Samuel's apocryphal definition has been so eagerly adopted is its most conspicuous aspect—the notion that the computer actually learns. But a computer does not learn in the way humans learn. It produces results, of which it has no understanding. A better way to think of how ML changes estimation is the following. Using traditional approaches, the analyst can easily envisage the set of possible results. With ML, it is much harder to envisage that set, and therefore surprises are more likely. A comparison between two methods long-used by cliometricians is instructive: linear regression and principal components. Solving linear regression leads to exactly what was envisaged, indications of which variables are related to the dependent variable. But the set of outputs that principal components could possibly produce is much larger,

an enormous variety of variable aggregates, together with information on how many aggregates the analyst should take seriously. Principal components is indeed often classified as an ML method, whereas linear regression is usually not.

And so we come to our definition of machine learning. ML refers to a large number of algorithms that leverage the power of modern computation, thereby expanding the set of tools available to researchers to analyze data. Athey (2019) expresses the point pithily: "one way to conceptualize ML algorithms is that they perform like automated research assistants—they work much faster and more effectively than traditional research assistants at exploring modeling choices." However, because of their origins, primarily in computer science, these algorithms have tended to have a very different style than those traditionally used in economics. Computer scientists are more interested in computational efficiency than statistical inference, and more interested in solving real-world problems than in teasing out causality. This has meant that ML algorithms have been primarily focused on two objectives: prediction and data summary.

No doubt these two foci stand at the fore because increases in computer power have inevitably led to much larger data sets. These facilitate the testing of prediction models by allowing the analyst, more or less costlessly, to divide the data set into two parts—a training set used for estimation and a testing set used to evaluate the estimates. If the goal is to obtain reasonable prediction, then we are spared imposing a procrustean structure that facilitates the application of statistical theory. Moreover, the larger a data set the more likely that parsimonious summaries are needed. In history, the largest datasets contain texts. Summarizing these implies converting the texts into numerical variables, with the number of possibilities close to infinity even with a small vocabulary. This has led to the development of sets of algorithms that characterize texts using a manageable number of variables.

Specific Terms

Because of their different goals and their origin outside the social sciences, and at least partially outside statistics, the new algorithms have brought with them a new set of terms that are useful to review before we proceed. ML algorithms have been designed to process numbers, texts, images, and sounds. The analyst always has to choose how to represent these data inputs within the algorithm and to choose between different representations, a process that is usually referred to as the extraction of features. Ultimately, a feature is a variable.

A major distinction is between supervised and unsupervised learning. All data analysis is about finding useful (predictive, evocative, causal) patterns in the data. The difference between supervised and unsupervised methods is that in the former each observation provides an example that approximately fits the pattern that one is aiming to find. In unsupervised learning, the world has not provided any observations that are examples of that pattern, examples that could be used as a litmus test of the pattern-finding exercise. Regression is a supervised technique: one is trying to estimate a function when one has data on both explanatory and explained variables. Then there are easily available methods of assessing how well the data fit the estimated relationship. Contrast an exercise that has been one objective of historians since the dawn of the discipline—trying to

figure out what ideas were in the brain of an author when she was writing her books. One simply has the books, which exhibit patterns in word usage. From one's own knowledge of word usage, one is able to associate those patterns with important ideas. However, we do not observe the ideas themselves and therefore cannot test how well we have fit the data, at least in the very direct way that is done in supervised learning.

Supervised methods, then, can be viewed very broadly as extensions of empirical techniques that economists would conventionally refer to as regressions. The application of these empirical techniques has been made possible by powerful machines, large data sets, and the associated emergence of new computational algorithms, whose precise properties, especially statistical ones, are often unknown. Unsupervised methods can be viewed as extensions of existing latent variable techniques in which the prime interest is in the estimated latent variables themselves. The latent variables might be estimates of ideas, as above, or the unobservable classes into which a certain entity falls, or, as in social network analysis, the closeness of two entities. But, importantly, before beginning the estimation, one has no data directly capturing the ideas, or the classes, or the closeness.

We turn now to a few terms that are less relevant in what follows, but have become so much a part of the hype surrounding computational data analysis that it is necessary to define them to clarify why they are not particularly relevant to our concerns. ML is usually viewed as part of artificial intelligence (AI). What AI adds to ML is active use of the estimates. When your adaptive thermostat learns from its temperature readings today what time to start heating tomorrow, it engages in ML. When it raises the heating setting at the right time tomorrow, it engages in AI—using the learning to do something that otherwise a human would have to do. The part of AI that is not ML is irrelevant here.

Deep learning refers to the use of a software architecture that has been found useful when finding patterns in large data sets. That architecture, usually called a neural network, was thought, probably incorrectly, to mimic the processes of the human brain. Ultimately, one single component ('layer') of the structure roughly amounts to feeding a set of input values into a set of estimated linear functions and then feeding the resultant function values into many estimated non-linear functions to obtain new output values. This is one layer of computation. Then, one can feed the outputs from one layer into the next layer, which does exactly the same thing. The term 'deep' refers to the multiplication of layers. The property of deep learning that is of interest to social scientists is that it, very roughly, provides a general-purpose function-fitting tool. Therefore, as a tool to either predict or to summarize, deep learning is very powerful The property that likely makes it less useful for understanding social processes is that it relies on the estimation of a huge number of parameters whose values usually have no intuitive interpretation.

Methods

Our aim in this subsection is not to explain the details of how ML algorithms work, but indicate what types of problems the algorithms solve. This subsection will be most useful for cliometricians new to ML who want to gain initial insights into which methods they might want to delve into.

For the delving, we recommend Burkov (2019), James et al. (2021), and Grimmer et al. (2021b) as good places to start.

Consider the following problem, which Mosteller and Wallace (1963) solved in what can be viewed as a first cliometric application of ML. One has a set of 100 articles, 50 of which are definitely known to be written by person A and 40 definitely known to be written by person B. One would like to predict who wrote the other 10, knowing that it is either A or B. Count some frequencies of word usage in all the articles, say the proportion of times that each article uses "thus", "therefore", or "hence". Run a linear regression using the 90 observations with a dummy variable for authorship as the predicted variable and the word counts as explanatory variables. Then use the estimated coefficients to predict the dummy variable for the 10 articles for which authorship is unknown. One has filled in gaps in the historical data using a technique almost a century old, linear discriminant analysis. Support vector machines and Bayes classifiers are much more powerful tools for performing classification tasks. The former method replaces the linear function used in the regressions by a general non-linear one and estimates it with non-parametric techniques. The latter uses Bayes theorem to generate the probabilities of being in the different classes.

It is well known that standard regressions perform very poorly on out-of-sample mean squared error when the degrees of freedom of a regression are small. One standard solution to this problem is to modify the least squares objective function by adding a penalty term that causes the estimated coefficients to be closer to zero, hence the term shrinkage methods. When the penalty is the sum of the absolute values of all coefficients, then many coefficient estimates will be zero. Lasso imposes such a penalty and therefore provides a method of automatic variable selection. It is particularly useful in cases where the number of explanatory variables is close to, or even greater, than the number of observations.

In a regression context, suppose that one had a large set of possible predictors and no idea about either which ones could be relevant or the functional form that should be applied to each individual predictor. That large set might include many speculative predictors constructed from interactions of variables. Regression tree methods provide flexible approaches to estimating the regression equation. They rely upon an iterative process which begins by dividing into two subsets the set containing all possible observations on the predictors. The division is made so that a dummy variable capturing membership in the subsets minimizes the residual sum of squares. Then on the second iteration, one of those two subsets is divided into two in pursuit of the same objective, and so on. The result is a regression that chooses a subset of predictors, which each enter the prediction function in a highly non-linear way. Several techniques have been developed to improve the properties of estimated regression-trees, including random forests, bagging, and boosting. For the purposes of this paper, it is sufficient to think of these techniques as regression-tree methods.

All the foregoing are supervised methods. We now turn to unsupervised methods, which we believe are the tools that will provide the most exciting avenues for new directions in cliometrics. Of all such methods, topic modeling is the most used. Topic modeling takes a set of documents

and asks what subjects or ideas (i.e., topics) can be found in these documents. It can be viewed very narrowly, as a data reduction technique, or very broadly, as a way to uncover subjects or deeper ideas that are implicit in many documents, even ideas that are never the explicit focus of any specific document. The algorithms that produce the estimates are based on a crude model of how documents are produced from the authors' ideas.(A much less used topic-modeling method, matrix-completion, is purely data driven and not underpinned by any specific model of document production.)

Topics are distributions over words, reflecting the words most used to refer to a particular idea. Documents are distributions over topics, reflecting the ideas that the production of the document relied upon. Thus by far the most interesting output produced by the topic model is the document-topic prevalence matrix showing the proportion of each document devoted to each particular topic. By associating this matrix with data on the characteristics of documents a wealth of opportunities to discover emerges, exploring how ideas change over time, how they differ between authors, etc. Of course, all such explorations rely ultimately on the analyst's interpretation of the content of each topic. This requires examining which words a topic most uses and which documents most use a topic.

Whereas topic models focus on finding the ideas or subjects contained in a set of documents, word embeddings focus on the meanings that have been attached to the words used in those documents. Any specific word is viewed as situated in a high dimensional vector space, 300 being a conventional number. The position of the word in the space is assumed to capture its meaning. Usually, estimating the vector representing any given word relies on algorithms that use data on the neighboring words in a large set of documents. Given the estimates, one can gain insights into the ways in which the words are used by examining the relationships between the vectors for the words. For example, if one found in a set of legal texts that the difference between the vectors for lord and lady was very similar to the difference between the vectors for property and children one might conclude something about gender roles in the society under investigation. Then changes in this degree of similarity over time would tell us something about changing gender roles.

Cluster analysis is an old technique that has gained new life in the ML era. It begins with a set of observations that are effectively points in n-dimensional space. (A document might be an observation if it is summarized using a vector of word counts.) The starting point is that these observations can be divided into a number of non-overlapping subsets, in which members of a subset share some property that non-members do not. Cluster analysis allocates observations to subsets, grouping similar observations on the basis of some metric (e.g., distance between points). The hope is that deeper characteristics of the entity summarized by the observation are reflected in cluster membership. Finding those characteristics, however, depends on understanding the

historical context and interpreting the determinants of cluster membership in light of the meaning of each dimension of the Euclidean space.[2]

Network methods begin with observations on the connections between any pair of entities contained in a set. The entities might be intellectuals or legal cases, for example, and the data on individual connections might be whether the intellectuals have written to each other or whether the legal cases cite each other. In such data sets there are often thousands of entities and millions of connections. The pertinent methods aim to generate parsimonious summaries of the whole network, asking questions such as which entities stand at the center, whether the network is separable into easily distinguishable sub-networks, who influences whom, etc. Although network analysis is a part of ML because of its dependence on computational algorithms, it has developed somewhat separately from mainstream ML given the non-overlap between the methods for networks and for those ML methods mentioned above.

Machine Learning, a New Tool for Solving Well-Understood Empirical Problems

One productive and immediate application of ML in cliometrics has been in the use of the new tools to solve problems that have been long known to exist when analyzing historical data with standard econometric methods. In these contexts, ML aids cliometric research that would have been feasible even before the advent of ML, but ML still clearly adds value. In what follows, we give an overview of the corresponding uses of ML methods together with specific examples, showing how such research fits into a small set of overarching themes. Our coverage of these applications is non-exhaustive, partially because new cliometric research employing ML tools is emerging as we write. Moreover, we concentrate only on highlighting which ML methods are used to solve which types of data or estimation problems, not on the substantive findings of the papers we discuss.

Improving Causal Estimates

ML has been fruitfully employed in addressing well-understood causality problems. Indeed, given the increasing emphasis on causal inference as the touchstone of social-science empirical research, some economists have contended that the use of ML to address causality would be the primary application of ML in cliometrics (Mitchener 2015, Athey and Imbens 2019). This is a view that we contest later in the paper.

ML has facilitated the construction of new variables that address a pre-existing omitted variable problem. For example, in the cliometric literature on wage disparities, one common concern has been unobserved human capital heterogeneity. To investigate black-white wage disparities in the U.S. South of the 1940s, Carruthers and Wanamaker (2017) impute missing scores on an Army test. They use regression trees that predict test scores as a function of a standard

[2] One extension of the basic cluster model (hierarchical clustering) explores which clusters become merged with each other in a process of sequential aggregation. The algorithm begins with the maximum number of clusters and merges them in a series of steps until there are just two clusters left. Another extension (fuzzy clustering) relaxes the assumption that each observation must belong to only one cluster.

set of socioeconomic measures, using the observations where test scores are available. Out-of-sample predictions from this regression are then used as an additional control variable in a racial-wage-gap regression.

In a standard IV-2SLS setting, ML can improve the first stage of the model, a prediction problem. In one application of central interest to cliometricians, Diallo (2021) reexamines Acemoglu et al.'s (2001) (AJR) evidence on the colonial origins of comparative development. Diallo uses a support vector machine to estimate a highly non-linear version of the first stage regression of institutional quality on settler mortality. This approach confirms AJR's substantive conclusions, but suggests there is less of a bias in the OLS estimates than was found by AJR. Thus, the inherent flexibility of ML in estimating a non-linear first stage might offer a way to improve traditional methods.

Poulos and Zeng (2021) use ML to increase the power of synthetic control methods to estimate the effect of homestead policy (the treatment) on state-level public education spending in the 19th-century US. At their core, such methods require a prediction of what would have happened to treated units had those units not been treated. To estimate that counterfactual, one needs aggregate predictions from a set of untreated units that are as similar as possible to the treated unit. Poulos and Zeng (2021) use neural networks to form those aggregate predictions, where the advantage of neural networks is the degree of complexity that can be implicitly embodied in the estimated, predictive equation and in the construction of the aggregate control units.

Addressing the Curse of Dimensionality

When data on many variables are collected and one allows for interactions and non-linearities, the number of variables can grow to be as large as, or larger than, the number of observations. This is the so-called curse of dimensionality, which especially arise in historical settings where cross-sectional samples are small and the frequency of observations over time is low. ML is useful in choosing predictors and their functional forms in such situations. Lasso does this explicitly by selecting variables from a set specified by the analyst. For random forests, the process is more implicit and allows for estimation of very flexible functional forms.

The curse of dimensionality is especially likely to arise when using text data, given the number of important words in any vocabulary and the numbers of ways in which these can be combined to obtain different meanings. Thus, Gentzkow et al. (2019) rely on lasso to address the dimensionality problems arising when estimating partisanship based on the text of Congressional speeches. Similarly, Michalopoulos and Xu (2021) employ lasso when examining the use of folklore concepts (embodied in keywords) as predictors of cultural values that are measured in modern-day surveys. Interestingly, the authors find that the ML-selected models can be inherently unstable across different specifications, especially when predictors are highly correlated. Therefore, a complementary human classification exercise was necessary to improve the results.

Several studies employ random forest methods. Düben and Krause (2023) use these methods to examine the interplay of geographic and institutional factors as predictors of the location of

administrative cities in imperial China. Ma and Li (2022) rely on random forest methods to find the predictors of the highest rank of Chinese officials prior to the 13th century. Kelly and Ó Gráda (2015) use a similar approach to examine the predictors of local population changes before and during Ireland's great famine.

Creating or Completing Datasets

A number of papers have used ML to create or complete datasets, where the objective of the studies has not been to answer a particular substantive research question. Rather the goal has been to employ ML for the purpose of improving datasets and supplying methods and data for other scholars.

A longstanding empirical challenge for cliometricians has been the linking of various historical datasets, for example, to study intergeneration mobility or the long-run effects of formative events or government programs (Abramitzky 2021). This is a difficult task because different datasets often do not have common identifiers, the spelling of names can vary, reports contain errors, and optical-character recognition is imperfect. Supervised ML algorithms (e.g., random forests or support vector machines) can be used to improve record matching. The algorithms involve estimating a match-function using a dataset that contains known good and bad matches and then predicting match quality out of sample. The known good matches might be obtained from manually coded data or from an external source, such as a public wiki-style repository (Goeken et al. 2011, Feigenbaum 2016, Poulos 2019, Price et al. 2021). The known bad matches are easily created artificially.

The record-matching problem can be viewed more generally as a scenario where the same individuals are in different datasets based on different data structures and so the individuals cannot be easily matched across the datasets. In the case of patents, identical ideas might fall into different categories in different patent classification systems. To study the historical development of patenting, it is beneficial to identify those ideas. Billington and Hanna (2021) do this using topic modeling applied to multiple historical patent datasets. Their data are the words of patent titles within existing taxonomies. Their estimated taxonomy could be used to classify all patent data consistently, an important goal given the key role of innovation in economic growth and given the need to conduct growth studies over time and across countries.

Historical studies sometimes have to rely on proxies for variables of interest. Saavedra and Twinam (2020) construct a new proxy for personal income in historical times using modern data. They predict incomes using a large number of predictors reflecting industry, occupation, and demographics. If those predictors, but not income, are available historically, then the income proxy is simply the prediction of income in the historical dataset. ML comes into play because of the sheer number of variables that could be used as predictors. They use lasso to solve the problem of variable selection.

Finally, a key problem in historical records is insufficient and fragmentary evidence. Both supervised and unsupervised ML algorithms can help fill gaps in records. Thus, there are studies

that have provided predictions to complete ancient inscriptions (Roueché 2022), that identify family names in primary sources (Liu and Hearne 2022), that classify archeological objects into appropriate historical periods (Ünlü 2019), and that disambiguate authorship (De Gussem 2017, Mosteller and Wallace 1963).

Generating a Crucial New Variable for Use in Conventional Regression Frameworks

In this section, we have focused on studies where ML has aided research that would have been easily conceptualized even before the advent of ML, and in most cases could have been accomplished without ML (but with a huge amount of human labor or using less nuanced methods). In many cases, ML has simply improved the quality of research that would have been feasible before the development of ML. However, in some cases, the additional element supplied by ML, say the construction of a new variable, is crucial in carrying the research to fruition. This is particularly the case when the new variable reflects text and therefore captures a phenomenon not discernible from standard numerical data available before ML.

Bi and Traum (2019) examine how newspaper reporting impacted U.S. government bond prices in the 1840s. The authors apply clustering and topic modeling to the corpus of newspaper articles for each state, allowing them to construct a state-specific, time-varying index of state legislative activities and fiscal actions. They then use this variable as a determinant of bond prices. The project would have taken on a very different hue if the authors had relied on traditional data collection rather than using unsupervised ML to extract the key variable of interest from a large amount of (text) data.

McCannon and Porreca (2021) also use topic modeling to construct a variable that could be used in an otherwise conventional regression setting. Drawing on the Old Bailey records on criminal trials in 19th-century London, the authors estimate the attention paid to topics that reflect particular crimes. They then use relative topic attention 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, a feature more important for some crimes than others. Again, it would have been very challenging to bring this project to fruition without the unsupervised ML.

Machine Learning, a New Lens for Illuminating the Flow of History

The previous section has highlighted ML as providing a new set of quantitative tools for tackling preexisting empirical problems, often focused on the economist's El Dorado, consistent causal estimates. This has been the predominant use of ML in cliometric research. We argue, however, that this is only one, and quite possibly not the most productive, use of ML for the study of history. ML also provides a new set of tools to offer wholly new quantitative insights into what has happened in the past. This is consistent with another striking difference between the research surveyed in the previous section and that reviewed in this section. In the former, ML primarily added to existing approaches. Here, the research projects are entirely dependent on the use of ML: the ML tools form the bedrock upon which the projects are built. These two aspects of the difference between the two sections are complementary. New insights are much more likely to

flow from analyses that view the data from a wholly new perspective. When the focus of research is a new way of describing the past, then inductive generalizations are bound to follow, providing one route for the emergence of new theories, as opposed to testing existing ones. ML thereby provides a new lens for investigating the flow of history.

The application of this new descriptive approach emanating from ML is not common within cliometrics. However, scholars in many other disciplines have pointed out the promise of ML in aiding description, induction, and theory generation. (See Grajzl and Murrell (2022b) and Grimmer et al. (2021a) for many references.) Kahneman (2019) 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". Ferguson-Cradler (2023) emphasizes "the need for what might be described as a 'narrative turn' in the discipline" and highlights the promise of computational text-analysis as a means to discover macro-level patterns in economic and social history (Mongin 2019). And recent work in the digital humanities points to the promise of macro-analytical, exploratory computational approaches in generating novel literary insights (see, e.g., Siewert and Reiter 2018).

ML has coevolved with the advent of big data and often its power rests on the existence of large datasets. When researching the, sometimes distant, past, where might big data be found to generate the new quantitative insights? The answer is in texts, and also images, the media primarily responsible for driving the development of historical ideas for millennia. The sheer volume of texts dwarfs the amount of quantitative data that has traditionally driven cliometric analyses. Before the advent of ML, the avenues for quantitative analysis of texts were limited. Now, as emphasized in the second section of this paper, ML offers an ever-expanding array of tools for quantitative analysis of large text corpora. The application of these tools makes possible a systematic inquiry into the 'great unread', the body of historical texts that are too many, and often too mundane, to have been absorbed into the analyses of scholars restricted by library stacks, time, and human memory.

In this section, we examine the types of questions that scholars have been able to address when applying ML in this new vein of research, providing examples of successful contributions. In the next section, we review our own contributions to English legal and cultural history, the development of which led us to develop the ideas on the place of ML in cliometrics that is the theme of this essay.

Developing the Broad Picture

In what is, to our knowledge, the first application of topic modeling in a historical context, Newman and Block (2006) describe the themes present in early American newspapers, focusing on the Pennsylvania Gazette. The paper is primarily a proof of concept: Newman and Block clearly illuminate the potential of unsupervised ML to "reveal cultural…histories", to facilitate an analysis devoid of "fallible human indexing or their own preconceived identification of topics", and incorporate "orders-of-magnitude more documents than a person can reasonably view".

Enlightening, specific facts are also generated: for example, the presence of ideas involving a national government nearly triples from the 1760s to the 1790s.

One frequent area of the application of topic modeling has been to describe the historical development of specific disciplines, for example, Peirson et al. (2017) on the history of biology and Bohr and Dunlap (2018) on the history of environmental sociology. One noteworthy example particularly relevant to present concerns is Wehrheim (2019), which applies topic modeling to articles published in the *Journal of Economic History* since 1941. This facilitates the identification of facts about the methodological evolution of the field, for example, the finding that the pertinent research took a distinctly cliometric turn in the 1960s.

Distilling Fascinating Facts from the Broad Picture

If one were to know the real story of the genesis of journal articles using unsupervised ML, we suspect that many began with the identification of a fascinating corpus and a glimmer of an idea about what might be found there, without the formulation of a concrete hypothesis. An initial analysis developed the broad picture. Then, some findings stood out as really remarkable, perhaps related to that initial idea, perhaps surprising. The publication process being what it is, the final product focused more on the remarkable finding than the overall picture. This is not to denigrate this approach—we actually think this is the genesis of much scholarly work, with induction its core and hypothesis testing its veneer. Indeed, we suspect that this is the real process of research underlying many papers that are ultimately framed in terms of the hypothetico-deductive method.

The importance of the above observation in the present context is that the reader should remember that development of a broad picture using ML underlies most of the work we discuss, even though in the following discussion we highlight only the main substantive findings, due to space constraints. The much larger number of interesting substantive findings in each surveyed contribution underscores the general point we stress: the new methods of ML provide an immensely powerful tool to describe past events and to make inductive generalizations.

Blaydes et al. (2018) draw on Muslim and Christian political-advice texts from the medieval era, providing insight into when European modes of political thought diverged from those emphasized in other areas of the world. They estimate a topic model, identifying topics common to both Muslim and Christian polities and then examine how the emphasis on the estimated topics changes over time. In Muslim texts, there is an increasing emphasis on the duties, obligations, and skills of kings between the eleventh and thirteenth centuries. For Christian texts, there is a decline in the prevalence of religious topics from the Middle Ages onwards, a distinct contrast with the Muslim texts.

Miller (2013) estimates a topic model on the whole corpus of surviving administrative records from late imperial China. The interest is in separating episodes of rebellious violence from more mundane criminal violence, given that a common vocabulary is used for both. Rather than imposing a particular view of word usage in analyzing the texts, Miller lets the texts speak for themselves. Six distinct topics relate to different facets of violent behavior. By examining the

words and documents most associated with the topics, he is able to characterize which topics capture the different types of violence. Then it is straightforward to show how the incidence of each type of violence waxed and waned.

Gennaro and Ash (2022) examine the text of six million speeches given in the U.S. Congress from 1858 onwards, focusing on the position of their content on an emotionality-reason spectrum. One could of course apply traditional dictionary methods in this endeavor, simply counting the number of words signifying emotions compared to those indicating reasoning. But ML facilitates the creation of an improved document-level measure. Applying word embedding, one can estimate a vector characterizing every word in the corpus. From these embeddings, one finds the average vectors for emotion words and for reason words, where these sets of words are obtained from standard dictionaries. Then, using measures of vector closeness, one can characterize every word and speech on the emotion-reason continuum. The analysis reveals, for example, that emotionality was relatively low and stable until the mid-20th century but increased significantly from the late 1970s.

Pozen et al. (2019) investigate a related issue, polarization, using a dataset comprising the text of all remarks made on the U.S. House and Senate floors from 1873 to 2016. To measure constitutional polarization, they examine how hard or easy it is for a supervised ML-based classification algorithm to predict a speaker's party affiliation using only the words in their remarks. When the algorithm makes precise predictions, then polarization is high. The analysis demonstrates that U.S. political discourse has become increasingly polarized over the past four decades and polarization has grown faster in constitutional discourse than in non-constitutional discourse.

Kozlowski et al. (2019) use a neural network to estimate a word embeddings model that reflects the Google Books Ngrams (Michel et al. 2016) corpus for the 20th century. By examining the similarity between the word-embedding vectors for two words, one captures the extent to which the two words are associated within general cultural discourse. For example, Kozlowski et al. find that the association between affluence and education grows over the 20th century while the association between affluence and morality declines. Remarkably, a comparison of texts from each decade of the 20th century indicates that the cultural dimensions of class comprise a rather stable semantic structure.

Giorcelli et al. (2022) aim at a more pinpointed finding when they investigate whether Darwin's 1859 publication of *On the Origin of Species* affected the broader cultural discourse. To assess the extent of cultural change induced by Darwin's work, they focus on core Darwinian notions (e.g., evolution and selection) and use preexisting word embeddings trained on Google Books Ngrams. The analysis reveals how Darwinian ideas gradually spread within the broader cultural discourse. For example, the notion of evolution, which was initially related most closely to concepts from chemistry and physics became semantically much closer to concepts in human society and biology.

Predicting History?

Risi et al. (2022) ask whether it is possible to accurately characterize events as historically prominent at the time when they are occurring. This is a fascinating question that has long intrigued philosophers, but it has been difficult to even conceive in quantitative terms until the emergence of ML. They examine a corpus of U.S. diplomatic cables from 1973 and 1979. The value of this dataset is that a subset of these cables were identified in hindsight by government historians as "conveying historically important information" relevant to U.S. foreign policy. To examine whether it is conceivable that observers at the time knew which cables would be later recognized by historians, Risi et al. model an ideal chronicler who is able to draw on all information available at the time of an original cable. The question is whether this ideal chronicler would be able to predict which of the original cables would be included in the selective corpus produced by historians.

This is a classification task, for which ML is ideally suited. For predictors, all information on any specific cable is used (e.g., the secrecy level) as well as indicators of the content of the cable. To this end, a topic model with 500 topics is estimated, providing a set of 500 topic proportions. When predicting which cables would come to be viewed as referring to significant events, the ideal chronicler is able to perform better than an index of how humans felt about cable importance at the time the cables were sent. But Risi et al. are ultimately forced to conclude that their ideal chronicler does not do particularly well in spotting events of truly historical importance. The fog of diplomacy is not lifted by ML.

Unearthing Buried Interconnections

Understanding socioeconomic interconnectivity and commonality has been one important focus of historical research. Cluster analysis and network approaches, both unsupervised ML methods, are tools for estimating such phenomena in practical settings, thereby facilitating new interpretations and new theorizing about historical events.

Van Vugt (2022), for example, applies a network approach to examine correspondence between Antonio Magliabechi, a prominent Florentine librarian, and his peers within a community of scholars. The analysis uses data on whom Magliabechi corresponded with and when the correspondence happened, plus whom those correspondents corresponded with. Then network algorithms can identify who are the brokers (those who connect disparate parts of the network), the degree of closure (the extent to which all participants in a network are in touch with each other), and when ruptures occur in the social network. The results paint a nuanced picture of the ebb and flow of relations between a group of interacting social agents.

Pagé-Perron (2018) similarly applies network analysis to a large administrative cuneiform corpus to identify previously unknown patterns of social organization within ancient Mesopotamia, together with the dynamics of those patterns. And Franzosi et al. (2012) use network analysis to elucidate the relationships between actors (African-Americans, whites) and the nature of actions (violence, coercion) described in a corpus of newspaper stories of lynchings in Georgia between 1875 and 1930.

In contrast, Perrin (2022) is primarily interested in investigating the nature of county-level heterogeneity in 19th-century France in order to shed light on France's early demographic transition to low fertility. Perrin's approach therefore uses hierarchical cluster analysis to uncover a typology of pre-19th-century French counties based on their socioeconomic and demographic characteristics. The use of ML thereby provides quantitative insight into the sets of features that are associated with the early spread of fertility control in some French counties, but not in others.

Seeing the Past

ML holds the promise of improving the understanding of historical actors and societies by dissecting their visual and physical creations. Recent applications of computer vision and augmented reality, surveyed in Kee and Compeau (2019), offer persuasive examples. Sheratt and Bagnall (2019), for instance, employ face-detection techniques to examine thousands of photographs of 19th and 20th century Australian immigrants. The analysis sheds new light on the discriminatory practices directed toward non-European immigrants under the 'White Australia' agenda. More generally, the Kee and Compeau (2019) edited volume demonstrates that, in historical research, "discovery comes from speculation and a playful approach to historical questions and problems".

A Machine-Learning Odyssey through English History

We now offer a personal perspective on the use of ML based on our own research. Once more, our goal is not to elaborate on the nuances of techniques. Rather, we explain how we came to view ML as an immensely productive set of tools for pursuing a cliometric agenda as we grappled with its particular problems and as ML led us to change our methodological perspective. We have two aims: first, to provide cliometricians interested in using ML techniques with practical insights and possibilities for analyzing historical texts and data; second, to argue that releasing the full power of ML requires a change in methodological style. Grimmer et al. (2021a: 397) argue that "machine learning is as much a culture defined by a distinct set of values and tools as it is a set of algorithms".

Given our extensive use of topic modeling, much of the discussion of ML is oriented around application of that technique. But the insights apply more generally, especially on methodological issues and on routine yet vital matters such as the preprocessing of historical texts.

Our substantive application is focused on England, primarily in the 17th century. Data availability greatly improves as one moves to the 20th century. For historians looking for existing applications of ML on which to model their work, a big problem is that most applications use episodes where data paucity is not a concern, making them less relevant to cliometric-specific difficulties. This is one reason why the present chapter has different points of emphases than those found in the more general ML literature. A large majority of historical settings will face the same data challenges, estimation conundrums, interpretational puzzles, and research opportunities that we faced.

From Interest in English History to Curiosity about ML

Our early joint work focused on the circumstances under which a powerful legal profession would facilitate or hinder institutional development (Grajzl and Murrell 2006). 17th-century England was a natural case study because of the acknowledged importance of the common law. Insights from this early work led Murrell (2017) to re-examine the prominent (North and Weingast 1989; Acemoglu and Robinson 2012) that new institutions designed and implemented at the highest-level provided the key step in turning the chaos of the 17th-century into the powerful economy and society of the 18th century. In contrast, the whole perspective of caselaw is of an edifice built slowly from the bottom up, via the accretion of past decisions.

Given this backdrop, a next step was to better understand how early-modern English lawyers and jurists thought about law and institutional change. We studied the writings of lawyer-scholars of the 16th and 17th centuries. We saw that the lawyers held a conceptualization of institutional development that foreshadowed many elements of Darwinism (Grajzl and Murrell 2016). We also noted the intimate connection between English law and the broader culture.

The methodology of this early exercise conformed to what digital humanities scholars would call a 'close reading' of historical texts, not a methodology that is common within economics, nor one consistent with our favored approach, the application of quantitative empirical tools. The close reading of traditional legal history has another major deficiency. Each contribution relies on only a limited number of texts: a close reading of the entire body of law is infeasible and the body of law that could be recalled when writing is even smaller than the one that could be read. We wondered how one might quantify and measure the sea of information embodied in various text corpora. Earlier work (Murrell and Schmidt 2011) had made clear to us that large text corpora were needed. Given that unsupervised ML had not yet entered economics, we needed to turn to other disciplines for lessons on how to analyze large corpora.

For economists, a journal called *Poetics* would not be the first choice to consult. However, the 2013 special issue on 'Topic Models and the Cultural Sciences' was especially informative and inspiring. The issue featured contributions by several key contributors to the emerging field of computational analysis of texts (e.g., David Blei, John Mohr, Justin Grimmer, David Mimno). The issue focused on the application of unsupervised approaches. The contributions illustrated the power of computational text analysis while stressing the complementarity between traditional (close reading) approaches and ML. There were articles that fit squarely into cliometrics, for example Miller (2013), which was described earlier in this essay. The 'distant' reading of 17th-century law and culture became an obvious route for us to take. Nevertheless, we were still very skeptical about whether ML could produce genuine substantive insights.

Testing the ML Waters

Before 2016, our only experience with programmable software was with the use of canned statistical packages. We typically prefer doing our own computation rather than relying on research assistants: we feel we really understand our data only when we wrestle with those data ourselves.

Our progress since then shows that ML should be feasible for any researcher willing to try. But when we began in 2016, we sought a manageable and relatively easy-to-process corpus to begin with, a decision made in part because self-help manuals on methods and machine-readable historical texts were so much scarcer then than they are today. However, even now, we would recommend this approach to cliometricians diving into ML for the first time: gaining a nuanced understanding of a new methodology is much easier if one does not have to grapple with data and programming challenges at the same time.

Of course, we simultaneously had substantive aims: to generate novel insights into early-modern English law and culture. We thus assembled a corpus of those works of Francis Bacon and Edward Coke that were already available in a machine-readable format. Simultaneously, we familiarized ourselves with Python (for the text processing); discovered, in our view, the most productive topic-modeling approach, structural topic modeling (henceforth STM; Roberts et al. (2014, 2016); and then learned enough R to implement STM.

Recognizing the Potential of ML

Our initial cliometric investigation of Bacon's and Coke's works (Grajzl and Murrell 2019, 2021a) was very much exploratory, not driven by the objective of testing a hypothesis. We hoped that the data generated via topic modelling would lead to new insights, some of which would perhaps bear on claims made in the existing literature. For example, a handful of Bacon scholars had alluded to the possibility that Bacon's path-breaking epistemological ideas could have sprung from his experience with the common law. If Bacon was indeed primarily an influential mouthpiece for a preexisting set of ideas, rather than a lone genius who provided a new one, then such a finding would have deep implications for the understanding of England's early economic rise.

How did the ML estimates help us understand the origins of Bacon's epistemology? Positive correlations of topic proportions across documents show that the topics share a conceptual foundation. Bacon's legal topics and his methodological topics were correlated. Timelines of topic usage showed clearly that Bacon's legal ideas developed well before his epistemological insights. Further insights were gained from the data produced by the topic modeling on the use of specific types of vocabulary. Following a difference-in-differences mode of reasoning, we examined whether the extent of overlap in inductive or fact-finding vocabulary between legal and epistemological topics was greater than the extent of overlap in the usage of such vocabulary between science and epistemological topics. Law and epistemology in Bacon had more overlap than law and science, indicating that Bacon moved from law to science.

For us, the most important insight from this early work was that ML could open new avenues of inquiry into areas of controversy that have existed for centuries. We do feel that Grajzl and Murrell (2019, 2021a) contains some of the most compelling evidence to date showing the origin of Bacon's epistemology in the common-law culture, a view that is decidedly a minority one among traditional cultural historians. But ML does not directly lead to the smoking gun—a statistically significant parameter estimate, sought by advocates of the hypothetico-deductive method. Instead,

it provides a patchwork quilt of evidence, which can be used to cast light on existing hypotheses or to generate new ones.

We also learned much about the practical side of applying ML. At the estimation stage, not all decisions can be made using statistical criteria; judgment has to be used, for example, in deciding how to pre-process the texts and in choosing the number of topics. At the interpretation stage, the reading of the documents that featured a given topic most prominently was essential to grasping the substantive meaning of the estimated topics. Close reading and distant reading are complements.

Most importantly, this early work gave us increased confidence in the power of ML. We were initially skeptical: the generative model underlying topic modeling assumes that any text can be completely characterized by the frequency distribution of the vocabulary in that text. Word order does not matter. This is preposterous. But despite this we found that topic modeling produces results that were intuitive, highly relevant, and even captured important semantic nuances.

On to Caselaw

The next step was to proceed to the work that had always been a prime objective: could we use ML to quantify and understand the historical evolution of English caselaw? Our objective was not to focus on the landmark cases, as traditional legal historians would, but rather tap into the 'great unread' to understand the macro-evolution of caselaw occurring within the thousands and thousands of often mundane cases that form the body of law.

We had to assemble the appropriate corpora. Luck played a role, as it often does in cliometric research. Schmidt (2016) found DVDs of a machine-readable version of the English Reports (Renton 1900-2932; ER, in short). The ER comprises a selection of the cases heard by the superior courts from the mid-15th century to the mid-19th century. The selection was by reporters who were especially focused on contentious or novel aspects of law: the ER thus certainly reflects broad developments in English caselaw. But the selection of the texts that are available for use will always be problematic in cliometric research: with details of selection processes lost to history, thinking carefully about the conditionality behind one's findings is the only solution.

Aided by earlier hands-on experience with topic modeling, we understood that careful cleaning of the texts would be immensely important. But the existing literature on computational analysis of texts, given its focus on contemporary settings, provided little guidance. There is the chaotic orthography and varying inflections of centuries of the changing English language. (To a computer, lysteth and lists have no connection.) There is the frequent use of Latin.[3] (To a computer, king and rex have no connection.)

[3] Also, a small subset of texts was in Law French, a language once used by all judges, which reads like a French essay of an English schoolboy who also knows Latin and whose essay exhibits a constant bemusement about how people across the channel manage to communicate with each other. Following the schoolboy's teacher, we gave up on trying to read these texts.

We addressed these immense challenges in pragmatic ways. Latin was translated word by word, making use of online Latin dictionaries. We tackled the problem of many variants of spelling in old English by translating it into modern-day English using online dictionaries supplemented by many thousands of bespoke additions (e.g., righteousenesse and righteousness). We were forced to drop a small number of documents that featured an uncharacteristically high share of words that could not be matched to any word in the English dictionary. The pragmatism means that our ER corpus is far from perfect. Nevertheless, we are sure that it is very suitable for deploying ML in cliometric research. Some scholars have used Google Ngrams to provide them with data input into their ML exercises. But a quick perusal would be enough for any reader to be convinced that this resource does not solve the evident problems we have referred to in this paragraph, especially for texts before the mid-19th century. We also chose to split the corpus into two parts, a pre-1765 corpus, focusing on the pre-industrial era, and a post-1764 corpus. Most of our research to date has focused on the former.

A large proportion of ML algorithms require choosing values of some parameters before estimation begins. In most cases, the literature offers no universally agreed-upon approach for these decisions. In the case of topic modeling, one must choose the number of topics. Our own experience is that although there are many statistical measures of the quality of the estimates, one's own judgment in interpreting topics goes a long way in providing an informed and defensible answer. For our 16th to 18th century corpus of 52,949 documents containing 31 million words, we chose 100 topics. The core output of ML was a 52,949-by-100 matrix showing the proportion of each document due to each topic, but other outputs might be useful such as the proportion of each topic due to each vocabulary word.

Interpreting the Estimates

Then, much laborious and challenging work follows. The 100 topics need to be evocatively labeled. A simple and quick route taken by many papers is to examine the words that are most associated with each topic. This is not sufficient and can often be misleading. Crucial information is also contained in the documents in which a topic is most featured. A reading of those will require knowledge of the broader institutional setting. Close reading of many documents and the study of (legal) history was therefore vital to interpreting the output: ML-based distant reading and the more traditional close reading are complements, not substitutes. Final decisions on topic names will require both extensive insight into a topic itself and comparisons between closely related topics. Quite often, the meaning of a topic only becomes clear when one understands which ideas are present in other topics. We also learned that there can be residual topics that have no clear substantive interpretation, which was rather disturbing when we first encountered one. However, topic models "shunt noisy data into uninterpretable topics in ways that strengthen the coherence of topics that remain" (Di Maggio et al. 2013).

Interpreting the Flow of Legal History

The document-topic prevalence matrix together with the metadata on the year of each document can be utilized to produce timelines that show the attention to specific topics over time.

An important challenge arose in interpreting the timelines. It is tempting to think that the proportion of a corpus accounted for by a topic in any given year is a measure of the importance of that topic's substance within the society in that year. This assumption is almost universally made by scholars engaged in the analogous pursuit of examining variations in word frequencies. However, we quickly realized this was not correct in our legal-institutional context. An inverted-U is characteristic of the temporal pattern in the amount of attention paid to topics reflecting legal ideas that had gradually become accepted as settled law.

Resolving this apparent paradox required some simple formal modeling. Using an evolutionary game-theoretic framework, we showed that the attention paid to a topic is a function of the amount of change in adherence to the corresponding legal ideas (Grajzl and Murrell 2021b). Hence, the timelines capture the intensity of yearly development of an area of caselaw, viewed from the perspective of those writing the documents, in our case legal professionals (judges and lawyers).

One example of the insights that flowed from the timelines is the dating of the development of legal ideas relevant to the financial revolution (Grajzl and Murrell 2021c). Of the 100 estimated topics, 11 refer to aspects of the caselaw relevant to the financial revolution. Examining timelines for these 11 topics, one finds that several pertinent areas of caselaw were settled well before the time typically associated with the financial revolution (the era after the Glorious Revolution of 1688).

If one possesses any type of variable that can characterize each document, then one can use that variable in combination with the document-topic prevalence matrix to make further characterizations of legal development. For example, one might ask: what are the actual legal origins of the caselaw and the legal ideas relevant to finance? As we demonstrate in Grajzl and Murrell (2022b), early 17th-century legal developments related to finance were concentrated in the common-law courts. Subsequently, however, equity (a non-common-law court) played an increasingly important role. As we conclude in that paper, had Britain relied solely on the common law when it began spreading its system of law around the world, Britain might never have been economically powerful enough to spread its common law around the world.

Topic modeling allied with simple tools to process its output can therefore lead to insights that were impossible to draw before the advent of ML. Perhaps, some legal historians had earlier conjectured the conclusions that we highlight in the two previous paragraphs. But we have not seen such conclusions stated explicitly, nor in quantitative terms, nor directly reflecting such a broad swathe of legal decisions. ML has much to add in simply facilitating different ways of describing the world, which can then provide substance for inductively generating new theories.

Beyond Descriptive

Description is not enough for social science, especially economics, with its fixation on the hypothetico-deductive method. Topic modeling, and more generally unsupervised ML, does not readily resonate with that approach. But topic modeling does generate a dataset that lends itself to

the application of standard econometric tools and hypothesis testing. Interested researchers can access our datasets at http://www.econweb.umd.edu/~murrell/.

For example, one might ask to what extent were early-modern legal ideas relevant for caselaw development during the Industrial Revolution. In Grajzl and Murrell (2022a), our unit of observation was a pre-1765 document (a case). We used a conventional (negative binomial) regression framework. Our dependent variable was post-1764 citations to pre-1765 cases (Murrell 2021). Our explanatory variables were the estimated topic prevalences, 100 for each document. We therefore had a problem for which the use of ML has been strongly advocated: should we use lasso to reduce the number of explanatory variables? In fact, our number of observations (documents) 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. In substantive terms, we found, inter alia, that one of the most important early-modern topics reflected Coke-style thought, a conclusion built on evidence much stronger than had previously been martialed to understand the legacies of one the greatest lawyers in history.

When one has time series of related phenomena, one can examine how these coevolved. Using the estimated document-topic prevalence matrix and the date of each case, in Grajzl and Murrell (2022c) we aggregated topics to produce annual time series of attention to three core elements of English caselaw: property, contract, and procedure. With three variables that are plausibly quite interdependent and no obvious and available exogenous measures to explain them, a vector autoregression (VAR) is the appropriate approach. In fact, exogenous events ('shocks') are estimated as a by-product of VAR when one uses the legal-historical setting to justify identification assumptions. One can thus separate the normal coevolutionary operations of the legal system from the effect of external shocks. Our analysis shows, for example, that caselaw on procedure and property closely coevolved, while the development of contract law was relatively autonomous. One of the largest shocks appeared in procedure from 1605 to 1619, exactly the time when Coke was exerting his outsize influence on legal development.

The next step was to ask whether caselaw development had any repercussions for economic development. For the pre-1765 analysis, analysis of Malthusian mechanisms using time series data was already well established in economic history. In Grajzl and Murrell (2022d), our dataset used variables standard in the Malthusian framework (real per-capita income and vital rates) alongside three aggregates of our legal topics that fit into that framework, caselaw on land, inheritance, and families. Following the methodology of Grajzl and Murrell (2022c) described above, we estimated a VAR on this expanded dataset. The results show that preindustrial caselaw development profoundly influenced economic development, with caselaw on families and inheritance being especially prominent. We know of no other paper that even toys with the idea of delving into this type of effect of caselaw: Grajzl and Murrell (2022d) contains results that are fundamentally new to the literature. The methodology of this paper could not have been conceived before ML made possible the quantifying of legal ideas embodied in law reports.

We then turned our attention to an analysis of the data produced from a topic model of the reports on post-1764 cases (Grajzl and Murrell 2022e). The analysis followed, mutatis mutandis, that in Grajzl and Murrell (2022c, 2022d). We supplemented 13 caselaw time series with a real per-capita GDP series and estimated a VAR. Our evidence shows that caselaw developments were a key driver of economic fluctuations. But precise effects depend on the legal domain. Developments in caselaw on intellectual property, organizations, debt and finance, and inheritance increased per-capita GDP, while developments in property and ecclesiastical caselaw reduced it. Again, we know of no other paper that even toys with the idea of contrasting effects of different areas of caselaw: obtaining such results would have been unimaginable before ML made the quantifying of legal texts possible.

We feel that there are two ways of viewing this extensive set of VAR results. One is to follow the standard view in macroeconomics, that the estimated relationship capture important aspects of causality. A second is to view these results as providing valuable descriptive evidence derived from a compelling approach to analyzing patterns in the data, for example, VAR-derived estimates of exogenous shocks. We are agnostic on this choice, even disinterested. Our views follow from a perspective that sees the question of the effect of English caselaw on English development as tremendously important and terribly neglected. Any rigorously developed evidence on this question constitutes a significant contribution to knowledge. Of course, this evidence would not satisfy those who believe that the only way to address such questions is via natural or quasi-natural experiments. But we view rough answers to questions of huge importance as at least as valuable as very precise answers to questions of lesser importance.

Culture

As had been our objective from the beginning, we also directed our efforts to studying culture (Grajzl and Murrell 2022f). Drawing on the Text Creation Partnership (2022; https://textcreationpartnership.org/) corpus, which covers the 16th and 17th centuries, we followed the steps outlined above for the ER corpus and estimated a topic model. Our 110 cultural topics synthesize the content of 57,863 texts comprising more than 83 million words.

Just a simple perusal of the topics and their timelines results in many observations that add to the stock of knowledge on English cultural development. For example, we name one topic Baconian Theology because it uses Baconian-style epistemology to analyze Biblical episodes and personal moral experiences in an effort to know God and interpret one's place in the world. Notably, we find attention to this topic increasing when Bacon was only a young man, well before his epistemological contributions. Could our topic model have uncovered an existing cultural source that stimulated Bacon's ideas?

Using analogous multivariate time-series methods as in Grajzl and Murrell (2022c, 2022d, 2022e), we examined 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. We show, for example, that innovations in religious ideas spurred strong

responses in the two other areas. Could innovations in theology have systematically influenced inquiry in secular domains?

We have ended the two previous paragraphs with questions because our work on culture is at an early stage. But even now, this work has shown us that our methods developed while analyzing Bacon, Coke, and then caselaw are eminently applicable in the context of culture. ML-based methods can address profound questions in a different way than ever before; they can be a tool for discovering facts, adding new items to the catalog of the information on which history is built; they can paint a picture that can spur hypothesis generation; they can generate data that facilitate hypothesis testing.

Not a Conclusion

This is the wrong time to draw a conclusion about the place of ML in cliometric research. Our conjectures are contained implicitly in what appears above. But ML techniques are changing fast. Large cliometric databases are accumulating and becoming more easily accessible. ML will be used in as many cliometric studies in the next five years as it has in all past years.

We have provided a snapshot of a fast-moving train that will be in a very different place ten years from now. We are sure that it is the time for economic historians, and historians more generally, to buy their tickets.

References

Abramitzky R, Boustan L, Eriksson K, Feigenbaum J, Pérez S (2021) Automated linking of historical data. Journal of Economic Literature 59(3): 865-918.

Acemoglu D, Johnson S, Robinson J (2001) The colonial origins of comparative development: an empirical investigation. American Economic Review 91(5):1369-1401.

Acemoglu D, Robinson JA (2012) Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown, New York.

Athey S (2019) The impact of machine learning on economics. In: Agrawal A, Gans J, Goldfarb A (eds) The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, Chicago, pp 507-547.

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annual Review of Economics 11:685-725.

Awad M, Khanna R (2015) Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress, New York.

Bi H, Traum N (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

Billington SD, Hanna AJ (2021) That's classified! Inventing a new patent taxonomy. Industrial and Corporate Change 30(3): 678-705.

Blaydes L, Grimmer J, McQueen A (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.

Bohr J, Dunlap RE (2018) Key Topics in environmental sociology, 1990–2014: results from a computational text analysis. Environmental Sociology 4(2): 1-15.

Burkov A (2019) The Hundred-Page Machine Learning Book. Andriy Burkov, Quebec City.

Carruthers CK, Wanamaker MH (2017) Separate and unequal in the labor market: human capital and the Jim Crow wage gap. Journal of Labor Economics 35(3): 655-696.

De Gussem J (2017) Bernard of Clairvaux and Nicholas of Montiéramey: tracing the secretarial trail with computational stylistics. Speculum 92(S1): S190-S225.

Diallo B (2022) Machine learning approaches to testing institutional hypotheses: the case of Acemoglu, Johnson, and Robinson (2001). Empirical Economics 62(5):2587-2600.

DiMaggio P, Nag M, Blei D (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics 41(6): 570-606.

Düben C, Krause M (2023) The emperor's geography - city locations, nature, and institutional optimization. Economic Journal 133(651): 1067-1105.

Feigenbaum JJ (2016). Automated census record linking: a machine learning approach. Unpublished manuscript. Available at https://open.bu.edu/handle/2144/27526.

Ferguson-Cradler G (2023) Narrative and computational text analysis in business and economic history. Scandinavian Economic History Review 71(2): 103-127.

Franzosi R, De Fazio G, Vicari S (2012) Ways of measuring agency: an application of quantitative narrative analysis to lynchings in Georgia (1875-1930). Sociological Methodology 42: 1-42.

Gennaro G, Ash E (2022) Emotion and reason in political language: the expression of emotions in 20th century books. Economic Journal, 132(643): 1037-1059.

Gentzkow M, Kelly B, Taddy M (2019) Text as data. Journal of Economic Literature 57(3): 535-574.

Giorcelli M, Lacetera N, Marinoni A (2022) How does scientific progress affect cultural changes? A digital text analysis. NBER Working Paper No. 25429.

Goeken R, Huynh L, Lenius T, Vick R (2011) New methods of census record linking. Historical Methods 44(1): 7-14.

Grajzl P, Murrell P (2006) Lawyers and politicians: the impact of organized legal professions on institutional reforms. Constitutional Political Economy 17(4): 251-276.

Grajzl P, Murrell P (2016) A Darwinian theory of institutional evolution two centuries before Darwin? Journal of Economic Behavior and Organization 131A: 346-372.

Grajzl P, Murrell P (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 P, Murrell P (2021a) Characterizing a legal-intellectual culture: Bacon, Coke, and seventeenth-century England. Cliometrica 15(1): 43-88.

Grajzl P, Murrell P (2021b) 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 P, Murrell P (2021c) 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 P, Murrell P (2022a) Lasting legal legacies: early English legal ideas and later caselaw development during the Industrial Revolution. Review of Law and Economics 18(1): 85-141.

Grajzl P, Murrell P (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 P, Murrell P (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 P, Murrell P (2022d) Of families and inheritance: law and development in England before the Industrial Revolution. Cliometrica, forthcoming.

Grajzl P, Murrell P (2022e) Did caselaw foster England's economic development during the Industrial Revolution? Data and Evidence. SSRN Working Paper.

Grajzl P, Murrell P (2022f) A macroscope of English print culture, 1530-1700, applied to the coevolution of ideas on religion, science, and institutions. Working paper.

Grimmer J, Roberts ME, Stewart BM (2021a) Machine learning for social science: an agnostic approach. Annual Review of Political Science 24: 395-419.

Grimmer J, Roberts ME, Stewart BM (2021b) Text as data: a new framework for machine learning and the social sciences. Princeton University Press, Princeton.

James G, Witten D, Hastie T, Tibshirani R (2021) An Introduction to Statistical Learning with Applications in R. Second Edition. New York, Springer.

Kahneman D (2019) Comment on Camerer CF, Artificial intelligence and behavioral economics. In: Agrawal A, Gans J, Goldfarb A (eds) The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, Chicago, pp 608.

Kee K, Compeau T (eds) (2019). Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History. University of Michigan Press, Ann Arbor.

Kelly M, Ó Gráda C (2015) Why Ireland starved after three decades: the Great Famine in cross-section reconsidered. Irish Economic and Social History 42: 53-61.

Kozlowski AC, Taddy M, Evans JA (2019) The geometry of culture: analyzing the meanings of class through word embeddings. American Sociological Review 84(5): 905-949.

Liu Y, Hearne J (2022) Towards archival reconstruction of Ur III cuneiform tablets. In: Frame G, Jeffers J, Pittman H (eds), Ur in the Twenty-First Century CE, Proceedings of the 62nd Rencontre Assyriologique Internationale at Philadelphia, July 11-15, 2016. Eisenbrauns, University Park, pp 309-314.

Ma L, Li M (2020) What helped officials of Song dynasty in climbing the greasy pole: an empirical study. SSRN Working Paper.

McCannon BC, Porreca Z (2022) The right to counsel: criminal prosecution in 19th century London. SSRN Working Paper.

Michalopoulos S, Xue MM (2021) Folklore. Quarterly Journal of Economics 136(4): 1993-2046.

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinkler S, Nowak MA, Lieberman Aiden E (2010). Quantitative analysis of culture using millions of digitized books. Science, 331(6014): 176-182.

Miller IM (2013) Rebellion, crime and violence in Qing China, 1722-1911: a topic modeling approach. Poetics 41(6): 626-649.

Mitchener KJ (2015) The 4D future of economic history: digitally-driven data design. Journal of Economic History 75(4): 1234-1239.

Mongin P (2019) Analytical Narratives. In: Diebolt C, Haupert M (eds), Handbook of Cliometrics, Second Edition, Springer, Cham, pp 1607-1638.

Mosteller F, Wallace DL (1963) Inference in an authorship problem. Journal of the American Statistical Association 58(302): 275-309.

Murrell P (2017) Design and evolution in institutional development: the insignificance of the English Bill of Rights. Journal of Comparative Economics 45(1):36-55.

Murrell P (2021) Did the independence of judges reduce legal development in England, 1600-1800? Journal of Law and Economics 64(3): 539-565.

Murrell P, Schmidt M (2011) The coevolution of culture and institutions in seventeenth century England. SSRN Working Paper.

Newman DJ, Block S (2006) Probabilistic topic decomposition of an eighteenth-century American newspaper. Journal of the American Society for Information Science and Technology 57(6): 753-767.

North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. Journal of Economic History 49(4): 803-832.

Pagé-Perron É (2018) Network analysis for reproducible research on large administrative cuneiform corpora. In: Bigot Juloux V, Gansell AR, di Ludovico A (eds), CyberResearch on the Ancient Near East and Neighboring Regions. Case Studies on Archaeological Data,

Objects, Texts, and Digital Archiving. Digital Biblical Studies, Vol 2, Brill, Boston, pp 194-223.

Peirson BRE, Bottino E, Damerow JL, Laubichler MD (2017) Quantitative perspectives on fifty years of the Journal of the History of Biology. Journal of the History of Biology 50(4): 695-751.

Perrin F (2022) On the origins of the demographic transition: rethinking the European marriage pattern. Cliometrica 16(3): 431-475

Poulos J (2019) Land lotteries, long-term wealth, and political selection. Public Choice 178(1): 217-230.

Poulos J, Zeng S (2021) RNN-based counterfactual prediction, with an application to homestead policy and public schooling. The Journal of the Royal Statistical Society, Series C (Applied Statistics), 70(4): 1124-1139

Pozen DE, Talley EL, Nyarko J (2019) A computational analysis of constitutional polarization. Cornell Law Review 105: 1-84.

Price J, Buckles K, Van Leeuwen J, Riley I (2021) Combining family history and machine learning to link historical records: the Census Tree data set. Explorations in Economic History 80(C): 101391.

Renton AW (1900-1932) The English Reports. Great Britain. Parliament. House of Lords. W. Green & Sons, Edinburgh.

Risi J, Sharma A, Shah R, Connelly M, Watts DJ (2019) Predicting history. Nature Human Behavor 3: 906-912.

Roberts ME, Stewart BM, Airoldi EM (2016) A model of text for experimentation in the social sciences. Journal of the American Statistical Association 111(515): 988-1003.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG (2014) Structural topic models for open-ended survey responses. American Journal of Political Science 58(4): 1064-1082.

Roueché C (2022) Mind the gap as AI guesses at lost Greek inscriptions. Nature 603(7900): 235-236.

Saavedra M, Twinam T (2020) A machine learning approach to improving occupational income scores. Explorations in Economic History 75(C): 101304.

Schmidt M (2015) Institutional Persistence and Change in England's Common Law: 1700-1865. PhD Dissertation, University of Maryland at College Park.

Sherratt R, Bagnall K (2019) The People Inside. In: Kee K, Compeau T (eds), Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History. Ann Arbor, University of Michigan Press, pp 11-31.

Siewert S, Reiter N (2018) The explorative value of computational methods: rereading the American short story. Amerikastudien/American Studies 63(2): 199-230.

Ünlü R (2019) Classification of historical Anatolian coins with machine learning algorithms. Alphanumeric Journal 7(2): 275-288.

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

Wehrheim L (2019) Economic history goes digital: topic modeling the Journal of Economic History. Cliometrica 13(1): 83-125.