Glossary

Agency
The capacity of individuals or entities to act or exert influence on their surroundings. In historical and social contexts, agency is often discussed in terms of an individual's relationship to power structures or social dynamics that enable or constrain one's ability to act. Agency can be analyzed at the level of grammar as it is represented through sentence structure—specifically, who is positioned as the subject (the one performing the action) and who is the object (the one receiving the action). This linguistic framing can shape perceptions of power, responsibility, and control in discourse.

Archives
Collections of historical records, documents, or data that are preserved for reasons such as cultural heritage or research. Archives can be physical (e.g., government records, letters, manuscripts) or digital (e.g., databases, repositories). Notable digital archives for historical research include HathiTrust, Internet Archive, Project Gutenberg, and Chronicling America. These repositories provide access to historical texts, newspapers, and various multimedia.

Argument, Function
In logic and programming, an argument refers to a value or variable that is passed to a function. A function is a reusable block of code designed to perform a specific task. Analysts typically write functions to keep from rewriting the same code repetitively, as the function name can be used instead of rewriting code line-by-line. A function typically consists of three key parts: a name, a list of input arguments (also called parameters), and a body—the set of code instructions or operations that define what the function does. When the function is called or "invoked," the arguments are passed in, and the body code is executed using those values. The function may return a result, often referred to as the output.

Britain / British Empire
In the context of the nineteenth century, Britain refers to the United Kingdom, particularly England, Scotland, Wales, and Ireland. The British Empire (c. 16th–20th century) was a global colonial and imperial entity, exerting political, economic, and military control over vast territories across Africa, Asia, the Americas, and Oceania. Parliamentary debates during this period—including discussions of imperial policy and colonized regions—are recorded in Hansard, the official transcript of proceedings in the British Parliament. These records give insight into the language and ideology underpinning British imperial rule.

Cleaning
The process of preparing and refining data for analysis by identifying and removing inconsistencies, errors, or irrelevant data that might detract from the analysis. While cleaning may be essential in most computational research, it is not a neutral process. Decisions about what to remove or what to retain involves the analyst's judgment and can significantly shape the outcomes of analysis. Cleaning can improve the quality of data, but it also risks introducing distortions, biases, or reinforcing assumptions that may obscure the complexity of the original source.

Code
A structured set of instructions written in a programming language to perform a specific task or set of tasks. Code serves as a way for humans to communicate with computers, translating logical steps into operations a machine can execute. It can be used for many purposes including data analysis and automation of repetitive tasks. Code is written in languages such as Python, R, JavaScript, or C++, each with its own syntax, strengths, and disadvantages. Well-written code is often reusable and accompanied by comments to improve readability and collaboration. The structure of code typically includes variables, functions, conditionals, loops.

Coding Environment
A software platform or integrated development environment (IDE) where analysts write, test, and run code. Coding environments often provide helpful functionality for writing code such as syntax highlighting and debugging. They may be language-specific (e.g., RStudio for R) or support multiple languages. Just a few more examples of coding environments include: Visual Studio Code, Jupyter Notebook, and Google Colab. Google Colab also supports collaboration in that it allows multiple people to access a single notebook at a given time.

Comments (in Code)
Annotations written within code to explain specific sections for readers. Though they do not affect how the program runs, comments are useful for documenting reasoning, outlining structure, and making notes to oneself. They can capture the analyst's assumptions or interpretive choices, providing a layer of meaning that exists parallel to the programming logic.

Concordance
A structured list of all occurrences of a specific word or phrase within a text or corpus, typically displayed with a snippet of surrounding context (known as "key word in context," or KWIC). Concordances allow researchers to examine patterns of language usage and variation across different sources or time periods. They are foundation tools in both linguistic, literary, and historical analysis, helping researchers contextualize words and trace shifts in meaning or the framing in which they are presented. In historical text analysis, concordances can offer an entry point into the interpretive labor of distant reading, or an intermediary transition point between a distant reading and can also act as bridges to more focused close readings of full documents.

Corpus vs. Dataset
A corpus is a collection of texts assembled for analysis, often used in linguistic, literary, or historical research. These texts are usually selected from a specific genre, time period, or discourse domain. For example, NovelTM (Underwood et al.) includes thousands of English-language novels for computational literary analysis, while the Hansard Corpus contains transcriptions of British parliamentary debates from the 19th and 20th centuries for political and historical research. Digital corpora used in computational research are often processed using natural language processing techniques. This may include tokenization, part-of-speech tagging, syntactic parsing, or the addition of structured metadata such as author, date, and publication context. They are typically curated with attention to genre or historical period.A dataset, by contrast, is a broader term that refers to any organized collection of data—textual, numerical, categorical, or otherwise—structured in a way that supports computational analysis. While all corpora are datasets, not all datasets are corpora. The distinction highlights differing priorities: corpora emphasize language and interpretive context, whereas datasets are defined more generally by format and function, and may prioritize measurement, classification, or predictive modeling over textual interpretation.

Corpus Linguistics
The study of language based on large, structured collections of texts—or, corpora—analyzed using computational and statistical methods. These methods enable researchers to identify patterns such as word frequency, syntactic constructions, and semantic shifts over time. For example, scholars like Marc Alexander and Jessie Demmen have used the Hansard Corpus—over 1.6 billion words of British parliamentary debate stretching back to 1803—to examine how language reflects shifting attitudes toward national identity, race, and empire. Their work demonstrates how computational approaches can uncover the subtle rhetorical strategies and evolving ideological frameworks that characterize political discourse across centuries. Corpora are typically annotated with metadata (e.g., date, genre, speaker) and may undergo natural language processing to include linguistic features such as part-of-speech tags or syntactic parses to support quantitative analysis. Rather than making claims based on isolated examples, corpus-based research asks researchers to ground their interpretations in patterns that emerge from broader language use. In historical and cultural research, corpus linguistics is especially valuable for tracing diachronic shifts in meaning. For example, researchers might use corpora to examine how the lexical field surrounding terms like liberty, race, or gender changes across centuries, or how political speech evolves in response to social upheaval. In this way, corpus-based methods illuminate how language both shapes and reflects broader cultural, political, and epistemological transformations.

Critical Evaluation (of Code)
A systematic assessment of code's suitability and accuracy in relation to the interpretive and methodological aims behind a given project. This process includes analyzing the underlying logic, clarity, and comprehensiveness of the code, as well as validating its outputs against expected or theoretically grounded results. In digital humanities research, such evaluation extends beyond functionality to include questions of epistemological rigor—whether the code reflects sound interpretive reasoning, aligns with the objectives of the analysis, and engages responsibly with the data or cultural texts under examination.

Critical Search
Using critical thinking to engage with data by putting them into conversation with primary and secondary sources and theory. To critically search is to engage in an iterative process of asking questions with data, rather than to execute a single algorithmic procedure and to treat the output as an intervention. Critical search may be conceived of as movement between seeding (i.e. using the known literature to propose a series of questions that a dataset can help to answer and how), broad winnowing (i.e. successive passes through the data, which iteratively narrow the body of work to be consulted), and guided reading (where the analyst turns to carefully read primary sources, guided by findings in the dataset).

Critical Thinking (about History)
In the discipline of History, critical thinking enters the process at almost every stage of explaining what happened in the past and analyzing change over time. The analyst must ask of each primary and secondary source: what is the bias that this source carried? Questions such as these afford the opportunity to ask who participated in, experienced, or caused different episodes of change over time, and whether the event was experienced the same way by each of them. Each event, in turn, is subject to the same questioning process: when did it begin? When did it end? Do we understand the event differently if we look over a longer or shorter period of time? Critical thinking about the past is impossible without an appreciation of the primary textual sources that lead to interpretation as well as an understanding of how other scholars have seen the event in the past, and these pressures put real constraints on any form of digital analysis divorced from the interpretive concerns of historians.

Critical Theory
An interdisciplinary, philosophical framework and evolving intellectual tradition that encompasses a wide range of approaches for analyzing the construction of knowledge, and the role of culture in shaping social life and cultural artifacts like text. While often associated with the critique of power—examining how systems of domination and inequality are embedded in language, institutions, and ideology—critical theory also engages with questions of aesthetics, ethics, epistemology, and identity. It includes diverse strands such as Marxist theory, feminist theory, postcolonial theory, critical race theory, queer theory, and deconstruction, each offering different methods for interpreting texts and understanding the social world. When applied to historical corpora such as the Hansard debates, critical theory enables researchers to explore how parliamentary language did not merely describe policy but helped constitute political realities—reinforcing imperial hierarchies, gender norms, or liberal ideals of citizenship. A Marxist reading might trace how economic interests and class ideologies shaped debates about labor or land; a postcolonial perspective might focus on how the discourse of civilization was mobilized to justify colonial violence; while a feminist analysis could reveal how women were rendered invisible in legislative speech. Rooted in the work of thinkers such as Karl Marx, Theodor Adorno, Michel Foucault, bell hooks, and Judith Butler, critical theory continues to inform inquiry across the humanities. Its strength lies in its capacity to open up texts to deeper questions about meaning and the conditions of knowledge production.

Cultural Analytics
Approaches to the study of culture through computational methods, data analysis, and visualization tools. Cultural analytics bridges the concerns of the humanities and data science, and frequently requires inspecting data about cultural artifacts – such as text, images, music, film, or social media – at scale.

Data
Across both the sciences and the humanities, data provides the evidentiary basis for analysis. It refers to a collection of observations, values, or representations that can be systematically processed or analyzed, encompassing a wide range of forms—including numerical values that can be measured or counted, or qualitative, capturing descriptive, contextual, or interpretive content such as interview transcripts, ethnographic

notes, or visual artifacts. In digital humanities research, the term "data" also invites critical scrutiny—not only of how information is collected and encoded, but of the epistemological and ethical assumptions that underlie its classification and use. This is because data collection is not a neutral task–it is shapped by human choices on what is collected, how it is categorized, and what is excluded. In New Digital Worlds (2018), Roopika Risam argues that data in the digital humanities must be understood as a product of cultural and historical processes. As Risam notes, the design of data curation reflects power relations and value systems, and thus requires ongoing critical reflection—particularly in projects that engage marginalized histories and cultural memory.

Historical Data
Information drawn from records, documents, artifacts, and other sources that provide insight into the past. Historical data is foundational to fields such as history, digital humanities, archival studies, and cultural analysis—but its relevance extends well beyond academia. Industries such as finance, insurance, energy, and urban planning rely on historical datasets to model risk or track trends. For example, shipping logs and climate records inform supply chain logistics and infrastructure resilience; census data underpins market research and demographic forecasting; and historical patent filings shape technological assessments. This data may be structured, like tax records or parliamentary transcripts, or unstructured, such as oral histories or marginalia. Researchers working with historical data face persistent challenges, including gaps in the archival record, biases in what was preserved, and distortions introduced through digitization. Nevertheless, historical data is vital for modeling the past and for understanding how institutions evolve and how memory and identity are encoded in language and practice. Working with historical data also demands a critical awareness of the ethical and epistemological implications of digitization, particularly as historical silences and exclusions of people are reproduced in digital corpora.

Data Science vs. Computer Science vs. Information Science
* Data Science is an interdisciplinary field focused on extracting knowledge and actionable insights from data. It employs computational modeling techniques such as statistical analysis or machine learning to identify patterns or generate predictions. Data science is both methodological and applied, emphasizing not only the development of algorithms but also their interpretation in context. Its scope spans a variety of domains—including public health, climate science, economics, and the humanities—where complex, often unstructured data must be transformed into forms amenable to analysis. As a practice, it draws from computer science for its infrastructure and tools, and from information science for concerns around data provenance, quality, and stewardship. * Computer Science is a foundational discipline concerned with the principles and technologies that enable computation. It encompasses both theoretical and practical dimensions: the former includes areas such as automata theory, formal languages, and computational complexity, while the latter includes software engineering or and human-computer interaction. Computer science provides the algorithms, data structures, and processing systems upon which fields like data science depend. It is concerned less with the content of data and more with the logic, structure, and efficiency of computational processes themselves. * Information Science examines the life cycle of information—its creation, organization, dissemination, retrieval, and use—within human and technological systems. It is inherently socio-technical, concerned not only with information infrastructures like databases and digital archives but also with the social dimensions of information access and management. Information science engages with questions of classification, metadata standards, privacy, and information equity. It has deep historical ties to library and archival studies but has evolved to include the study of digital platforms, user experience, and information policy in both public and private sectors.

DataFrame
A two-dimensional, tabular data structure widely used in digital historical research for organizing and analyzing structured data. Commonly implemented in programming environments such as Python's pandas library or the R language, a dataframe arranges information into rows and columns, where each row might correspond to a discrete historical observation—such as a census record, parliamentary speech, notarial act, or newspaper article—and each column represents a specific variable, such as date, location, author, or thematic tag. This format supports systematic analysis of large-scale historical datasets, facilitates comparisons across time periods or sources, and supports reproducible research through code-based data processing. Researchers can easily apply filters to dataframes or create visualizations as part of computational historical inquiry.

Data Type
The classification of data based on its nature and the kinds of operations that can be performed on it. Common data types include numerical data, such as integers and floating-point numbers, which are used for calculations and measurements. Categorical data represents discrete groups or labels and can be either nominal (without order, like country names) or ordinal (with a meaningful order, like rankings). Boolean data captures binary values such as true or false, often used in logical operations. Text data, or strings, consists of characters and is used to represent words, sentences, or symbols. Date and time data captures temporal information, including timestamps and durations, allowing for chronological analysis and time-based comparisons.

Digital Humanities
An interdisciplinary domain that integrates computational methodologies with humanistic inquiry as part of the interpretive process. Drawing on techniques such as text mining, data visualization, spatial analysis, and network analysis, digital humanities (DH) enables scholars to explore cultural, historical, or literary questions at scale, while also attending to the complexities of humanities interpretation. The field has its roots in the mid-20th century tradition of "humanities computing," exemplified by Father Roberto Busa's collaboration with IBM to produce a digital concordance of Thomas Aquinas's works. In the early 21st century, DH expanded significantly, shaped by institutions such as the Stanford Literary Lab, co-founded by Franco Moretti. Moretti's concept of "distant reading" advocated for the use of quantitative methods to analyze large corpora of literature. This approach, and others like it, challenged traditional methodologies in literary studies by treating literature as a dataset to be modeled. Today, digital humanities encompasses a wide range of practices—from building digital archives to designing interactive platforms and a space for reflecting on the epistemological consequences of computational research in the humanities.

Distinctiveness (e.g., TF-IDF)
A statistical principle used to evaluate the significance of a word in a specific document relative to a larger collection of texts, or corpus. It combines term frequency (TF), which counts how often a word appears in a document, with inverse document frequency (IDF), which reduces the weight of words that appear commonly across many documents. TF-IDF was originally developed in the 1970s by information retrieval researcher Karen Spärck Jones as part of her pioneering work in information retrieval—the science of designing systems (like search engines) that help users find relevant documents in large collections of text. At the time, one of the central problems in information retrieval was ranking which documents were most relevant to a user's query. Simply counting how often a term appeared in a document (term frequency) did not work well because some words–like the, and, or even common topic words–appeared frequently across many documents and thus provided little discriminative power. Spärck Jones introduced inverse document frequency (IDF) to address this issue. Her insight was that the informativeness of a word is inversely related to how frequently it appears across a corpus. Words that are rare across documents but frequent within a single document are more likely to be meaningful for distinguishing that document's content. In computational historical analysis, TF-IDF can be used to help researchers identify which terms were especially prominent or meaningful in particular speeches or debates relative to broader discourse trends over time.

Distance (e.g., JSD)
A mathematical measure for quantifying the difference or dissimilarity between two sets of language data. Divergence metrics are widely used in computational text analysis to track how word usage shifts across corpora, time periods, or social contexts. One prominent example is Jensen–Shannon Divergence (JSD), a method that measures how two distributions—such as word frequencies—differ while remaining interpretable. In digital humanities research, JSD has proven especially powerful for studying long-term changes in discourse. For example, Sara Klingenstein, Tim Hitchcock, and Simon DeDeo used JSD in their study "The Civilizing Process in London's Old Bailey" to examine how courtroom language changed over more than a century of trial transcripts. By comparing distributions of words used in earlier versus later trials, they demonstrated a gradual shift from violent and physical descriptions of crime to more abstract and moralizing language—evidence of what they interpret as a "civilizing" transformation in legal discourse. Building on such methods, Partial JSD extends the analysis by identifying which specific terms contribute most to the divergence, allowing researchers to analyze the lexical drivers of historical or cultural change over time.

Digitization

The process of converting physical records—such as manuscripts, printed books, or maps—into digital formats that can be stored, accessed, and analyzed with a computer. Digitization ranges from basic scanning, which generates static image files, to techniques such as optical character recognition (OCR), which transforms printed or handwritten text into machine-readable data. In digital humanities research, digitization is a critical step that enables large-scale text mining of historical materials. Advances in vision-language models (VLMs), have introduced new possibilities for transcription accuracy, especially for handwritten and multilingual documents as they are able to interpret more complex layouts or faded and irregular handwriting with greater precision than traditional OCR tools. For instance, projects such as Transkribus have begun to incorporate deep learning and VLM-based pipelines for the transcription of early modern handwriting.

Embeddings
Embeddings are a method for translating words or documents into numerical representations that reflect patterns of meaning and usage. In these representations, words that appear in similar contexts are positioned closer together in a multi-dimensional space. Words that are considered as more dissimilar to one-another appear further away. This allows researchers to analyze the embedding and identify relationships between words based on how they are used in context, rather than relying on exact matches. For example, within a word embedding representation of a corpus, an analyst might recognize that "labor" and "workforce" may be used similarly even if the exact word doesn't appear in both texts. In analyzing change over time, word embeddings can reveal how associations between words shift across different historical periods. For example, they might show that in the United States Congressional Record, the word man became increasingly associated with the term gay in later decades—indicating a change in how certain concepts were linked in political discourse (Buongiorno).

Event
A discrete unit within a dataset that signifies a notable action or occurrence. In historical research, an event typically refers to a specific, documented moment—such as a political speech, a singular vote, the passage of legislation, a protest, a battle, or the publication of an influential text—that signals a transformation in social or cultural dynamics. In computational analysis, events may be identified from either structured data (e.g., metadata with dates and locations) or unstructured sources (e.g., patterns in language use, the clustering of key terms, or temporal spikes in document mentions). Detecting and analyzing events in this way enables researchers to identify historical patterns and situate individual occurrences within larger historical processes.

Field (in Academic Study)
An organized area of scholarly research defined by a common set of questions, methods, theories, and subject matter. Fields can range from broad disciplines like history or computer science to more specialized areas like digital humanities or forensic linguistics. Fields provide a framework and template for how research is conducted, such as what tools and approaches are prioritized. They also influence how scholarly work is received by an audience and evaluated. Fields also structure academic communities by forming networks of journals and conferences. Over time, fields may shift or expand as new topics gain attention, interdisciplinary work emerges, or social and technological changes influence research priorities.

Field (in a Dataset)
A column or attribute in a structured dataset that represents a specific category of information. For example, in a dataset of historical speeches, fields might include "Date," "Speaker," "Location," and "Debate Title" Each field contains a particular type of data—such as text, numbers, or dates—and helps organize the dataset in a consistent, analyzable format. Fields are used for filtering and sorting data for quantitative or qualitative analysis.

Filter (in a Dataset)
A process used to refine or subset a dataset by applying specific conditions to one or more fields. By using filters, researchers can extract only the records that meet certain criteria—such as selecting speeches delivered by a particular politician, documents from a specific decade, or entries that mention a key term. Filtering narrows down large datasets to focus on relevant subsets and for conducting targeted analysis in historical research and allowing for more precise comparisons.

GitHub

A web-based platform for version control and collaborative software development, built on the Git system, which tracks changes in code and text files over time. In the context of libraries and the digital humanities, GitHub serves as an infrastructure for managing research workflows, sharing tools, and ensuring transparency and reproducibility. Many R packages relevant to historical and legal data analysis—such as hansardr (for British parliamentary debates), usdoj (for U.S. Department of Justice data), and noaa (for climate and weather datasets)—are hosted on GitHub, making them openly accessible to the public. Along with providing access to data and code, GitHub also fosters collaboration across disciplines by supporting documentation. In this way, GitHub advances the principles of open scholarship and aligns with the FAIR data guidelines (Findable, Accessible, Interoperable, and Reusable), helping research communities build sustainable, extensible tools for digital inquiry (Wilkinson et. al 2016).

Grammar
The system of rules that governs the structure and use of language, encompassing elements such as syntax (sentence structure), morphology (word formation), and semantics (meaning). In linguistic and historical analysis, grammar can provide insight into interpreting how agency, authority, and responsibility are constructed in discourse. Subject-object relationships, verb tense, and voice (e.g., active vs. passive) shape how actions are attributed and events are framed. In this way, grammar offers insights into power dynamics and ideological positions embedded in historical texts.

Hansard
The official transcript of parliamentary debates from Great Britain, Hansard documents the speeches and decisions made in the UK Parliament and serves as a key resource for historical and political research. Multiple versions exist for different scholarly purposes. The Historic Hansard (1803–2005) is publicly available through the UK Parliament, while curated versions like those by Egger and Spirling or the hansardr R package offer cleaned and structured data for computational analysis. Proprietary versions, such as the version of Hansard hosted by ProQuest, provide additional metadata and search functionality but often require institutional access.

hansardr
hansardr is an R data package designed to make access to 19th-century UK Hansard parliamentary records straightforward within the R programming environment. It can be installed from GitHub and provides preprocessed, analysis-ready versions of the Hansard debates, organized by decade to support efficient loading and processing. The package allows researchers to query speeches, retrieve speaker and session metadata, and conduct large-scale analyses of legislative discourse using R tools. By default, hansardr includes a sample dataset containing 10 rows per decade to enable rapid testing and exploration. To access the full corpus, users can run the function `download_hansard()`, which replaces the samples with the complete records for each decade of the 19th century. The Hansard corpus is structured by decade and includes four interlinked datasets per period, all keyed by sentence_id: `hansard_YYYY` (debate text), `debate_metadata_YYYY` (speech date and title), `speaker_metadata_YYYY` (original and disambiguated speaker names), and `file_metadata_YYYY` (speech IDs and source file metadata). This structured approach enables joined queries across text, speakers, and metadata, supporting both close and distant reading methodologies in digital historical research.

Historiography
Historiography is the study of how history has been written about over time. Historians rarely agree about which events are important, who participated in those events, what caused them or what their most important consequences were. Reviewing the tensions between different perspectives in the work of historiography. In the professional discipline of history, reviewing historiographical debates is often understood as a necessary preliminary step before studying new historical material, primary sources, or data – because only after considering a perspective on how other scholars have interpreted the past, agreed or disagreed in the past, can a scholar be confident in a new interpretation.

History
History is the study of the past, especially through the examination of change over time as marked by "events," where culture, society, ideas, politics, or economics shifted in a way noticeable to participants. History also encompasses the study of periods of time, which are stretches of time punctuated by events. Periods and events may both be long or short. Widespread consensus about the major episodes in collective history has long been understood as a necessary precondition for a society to accept a shared form of

governance, especially democracy. At the same time, dissent about history is typically considered productive of integrating new points of view into a perspective on the past, so long as the dissent in history is based on facts – that is, forms of evidence that survive from the past, whether in the form of documents, landscapes, data, or folklore – whose origin in the past is recognized as valid by most careful observers, whether or not those observers share other interpretations of the facts' consequence.

Iteration
A repeated process of refinement and revision. Iterative research processes are central to both computational research and historical analysis. In data science, iteration often involves adjusting algorithms, models, or visualizations based on insights from previous results. In historical research, iteration may take the form of moving between different layers of evidence—starting with a visualization or summary of data, then returning to the original dataset, consulting relevant secondary literature, and ultimately reexamining primary sources. This iterative process allows analysts to deepen their interpretations, test their assumptions, and build more nuanced arguments by continually revisiting and refining their understanding of the material and the method through which they arrived and their understanding.

Joins
A data processing operation that combines data from multiple tables based on a shared field (or key). Joins are a key part of data manipulation in R, particularly when using packages like `dplyr` (from Hadley Wickham's `tidyverse`). Common join types in R include **inner_join()**, which returns only the rows with matching keys in both datasets, and **left_join()**, which preserves all rows from the first dataset while merging in any matching data from the second.

Keywords
Significant words or phrases used to index, search, or analyze texts. Keywords can guide digital historical research. Their notion stems from cultural theorist Raymond Williams who famously explored the layered and contested meanings of socially significant words in his book **Keywords: A Vocabulary of Culture and Society (1976)**, arguing that certain terms serve as windows into broader ideological, cultural, and historical shifts. Building on this foundation, keywords are used by historians as conceptual anchors to trace changes in discourse, examine the evolution of political or cultural values, and understand how language reflects and shapes social life. In computational research, keywords can be manually assigned—such as metadata tags in digital archives—or automatically extracted using methods like TF-IDF (Term Frequency-Inverse Document Frequency), which identifies terms that are especially distinctive within a given document or corpus.

KWIC (Key Word in Context)
A method in text analysis that displays a selected keyword alongside a fixed number of words before and after it, providing a window into the term's immediate context like a concordance. KWIC concordances are especially useful for examining how words are used in practice—revealing patterns in meaning, tone, and grammatical usage across different texts. This approach is widely used in linguistic, historical, and digital humanities research to trace shifts in discourse over time. In R, the `quanteda` package offers a built-in `kwic()` function that allows researchers to generate KWIC displays quickly from large corpora.

LLM (Large Language Model)
Large Language Models (LLMs) are a form of artificial intelligence (AI) trained on massive amounts of text to generate, interpret, and analyze language. Notable examples include OpenAI's GPT-4 and GPT-3.5, Google's Bard, and Anthropic's Claude. In historical research, LLMs can assist with tasks such as brainstorming, code generation, metadata extraction, and exploratory analysis of large corpora. They can be used for generating initial interpretations that scholars can then refine. However, LLMs do not truly reason — their outputs are probabilistic, based on predicting likely word sequences rather than understanding meaning or context. As a result, while LLMs can be valuable tools, their outputs must be critically evaluated, as they may contain inaccuracies such as "hallucinations" (responses that are semantically plausable but factually incorrect), reflect biases, or omit the complexities of historical evidence.

Longue Durée, Big History, Moyenne Durée, Microhistory
Longue-durée history is history of long spans of time – as short as a few decades or as long as multiple centuries, although typically history longer than a few millenia is classified as "big history," because its

analysts must have recourse to sources from evolution and neuroscience, rather than the documentary set of sources upon which most modern historians rely. In theory, the analysis of datasets can help analysts to expand their account of history over decades to a century or more, although in reality, most text-based datasets linked to institutions (such as Hansard) are limited to a few decades or at most two or three centuries. Approaches that aggregate information across time and space at scale are contrasted with microhistory, an approach to social history that developed in Italy and France in the 1960s, and which emphasizes the intensive reading of local archives in order to develop a rich account of the past. Originally, local microhistories made room for longue-duree analysis, although social historians of the 1980s, 1990s, and 2000s tended to gravitate to the examination of local history over timescales as short as a year or several years.

Metadata
Descriptive information about data that provides context and structure and transforms documents into searchable, analyzable datasets. In sum, metadata is data about data. In the context of Hansard, metadata includes details such as the name of the speaker, the date of the debate, the debate title. This metadata is important for organizing the parliamentary record and enables researchers to filter speeches by speaker or trace debates across specific time periods. For historians, metadata in Hansard supports both targeted searches—such as locating all speeches by a particular MP—and large-scale computational analysis, supporting comparisons across decades.

N-grams (Unigram, Bigran, Trigram)
Sequences of n words or characters often used in computational text analysis to examine patterns, co-occurrence, and linguistic structure. N-grams are derived through the process of "tokenization" or breaking text into meaningful units, which can be individual words, phrases, or characters.

- Unigram (1-gram): A single word (e.g., history).
- Bigram (2-gram): A two-word sequence (e.g., digital humanities).
- Trigram (3-gram): A three-word sequence (e.g., machine learning model).

Normalization
Normalization is a data processing technique used to standardize data to ensure consistency and comparability across a dataset. In computational text analysis normalization typically involves transforming text to a uniform format by applying steps such as lowercasing all words, removing diacritics (e.g., converting *résumé* to *resume* or vice-versa), and using lemmatization or stemming to reduce words to their base or root form (e.g., if lemmatizing, run, running, and ran are each transformed into run). The goal may be to minimize variation of word forms that could otherwise distort frequency counts or pattern recognition. In historical research, normalization allows scholars to compare terms across different time periods, spelling conventions, or textual sources, making it a common early step for an analysis.

Operating System (OS)
The software that manages a computer's hardware and software, enabling users and programs to interact. Common operating systems include Windows, macOS, and Linux. Each OS offers different capabilities and interfaces. In computational historical analysis or data science, the choice of operating system can significantly affect programming environments, software installation, and computational workflows. For example, certain R or Python packages may require Unix-based systems (like macOS and Linux) to install correctly, and command-line tools often function differently across OS platforms. Understanding OS compatibility is therefore important for reproducibility and managing software dependencies.

Parliament
A legislative body responsible for debating, shaping, and enacting laws within a government. It serves as a central institution in many democratic systems, often comprising elected representatives who engage in public debate and decision-making. Parliamentary records, such as the UK Hansard, are primary sources for historical and political research. These records provide detailed accounts of legislative proceedings, including speeches, debates, and policy discussions, offering insights into how laws were formed and how political ideologies evolved to influence public opinion and governance over time.

Pedagogy
The theory and practice of teaching, pedagogy encompasses the methods, design, and underlying philosophy

that guide how knowledge is structured and delivered to learners. In fields such as digital humanities and data science, pedagogical approaches are often aligned with the values and practices of the discipline itself. For example, teaching in the digital humanities may emphasize interpretation of digital texts and critical engagement with technology, while data science pedagogy often focuses on problem-based learning or reproducible research practices. In both cases, pedagogy is not just about content delivery but about shaping how students participate in the intellectual and methodological frameworks of the field.

Periodization
The practice of dividing history into distinct time periods based on shared social, cultural, political, or economic characteristics. These divisions—-such as the Renaissance, Enlightenment, or Industrial Revolution-—help historians organize historical developments and make sense of change over time. However, periodization is not a neutral or purely objective process; it often reflects the priorities and interpretive frameworks of a historiographical tradition. While useful for structuring narratives and comparative analysis, period labels can obscure continuities or impose artifical boundaries that marginalize human experiences. As such, periodization is both a tool for analysis and a subject of critique.

Plots
Visual representations of data that make it possible to glean patterns, trends, and relationships that might be difficult to detect in raw numerical or textual data. They are useful tools in both exploratory analysis and the communication of results, allowing researchers to summarize complex information in a visual fashion In historical and computational research, plots can be used to track change over time—for example, a line plot might show how the frequency of a political term changes across decades, while a bar chart might compare the number of speeches given by different political parties in a given period. Categorical data, which refers to variables that represent distinct groups or labels (such as speaker names, political affiliations, or geographic regions), is often visualized through bar charts or grouped scatter plots. Other plots, like histograms, can illustrate the distribution of values, such as word counts or sentence lengths in a corpus.

Prediction
A computational process that estimates future values, classifications, or outcomes based on patterns in existing data. It plays a central role in both machine learning and historical modeling, where algorithms are trained to recognize trends and make inferences. Methods of prediction include statistical forecasting, such as projecting economic indicators over time; text classification, such as identifying the likely sentiment or political leaning of a speech; and topic modeling, where prediction refers to estimating the probability that certain latent themes are present in a document based on word co-occurrence patterns. Predictive methods must be used with caution in historical contexts, as language is shaped by cultural and temporal factors that do not necessarily follow consistent or repeatable patterns. Words in a historical corpus do not "predict" future usage in a statistical sense, and applying predictive models without context risks oversimplifying complex historical phenomena. As such, prediction in historical research is best understood as a tool for exploring patterns and generating interpretive questions, rather than as a method for drawing definitive conclusions about the future based on the words used in the past.

Primary Sources
Original documents or artifacts that offer firsthand accounts of historical events, created by individuals who experienced or observed them directly. These sources include manuscripts, letters, diaries, newspapers, government records, oral histories, photographs, and material culture. In the context of political and institutional history, **Hansard**, the official transcript of UK parliamentary debates, serves as a primary source that captures the language of legislators. Unlike secondary sources, which analyze or interpret historical evidence, primary sources provide direct access to the voices and contexts of the past.

Programming Language
A formal system of symbols used to instruct computers to perform specific tasks, ranging from data processing to app development. Different languages are suited to different purposes based on their syntax, functionality, and ecosystem–that is, the larger community that develops and contributes to the programming language. In data science and the digital humanities, languages like Python and R are widely used for tasks such as text analysis, data cleaning, and visualization. SQL may be used for querying and managing structured data in databases, making it valuable for working with archival metadata or large corpora. For building web-based projects or interactive visualizations, languages such as JavaScript, HTML, and CSS are commonly

used. The choice of programming language shapes both the technical capabilities of a project and the ways researchers engage with their data.

R
A programming language and open-source software environment originally developed in the early 1990s by statisticians Ross Ihaka and Robert Gentleman to support statistical computing and data analysis. It has since become a widely used tool in the digital humanities for analyzing large textual corpora, modeling patterns in historical data, and producing reproducible research. R's extensive ecosystem of packages—especially those in the tidyverse—allows researchers to clean, structure, and visualize data with syntactic and stylistic consistency. Tools like quanteda, tidytext, and tm support text mining, keyword analysis, topic modeling, and other forms of computational textual interpretation. Its community and emphasis on transparency and openness make R especially valuable for open science projects that combine humanistic inquiry with data science methods.

RStudio
An integrated development environment (IDE) for R that provides a comprehensive interface for writing, debugging, and running R code. RStudio uses multiple panels to organize the workspace for analysts. The Source panel allows users to write and manage R scripts, while the Console panel is where code is executed interactively. The Environment/History panel tracks the objects in the current session and maintains a history of commands. The `Files/Plots/Packages/Help` panel offers access to file directories, visualizations, installed packages, and R documentation. These panels work together to provide a cohesive environment to support data analysis. Additionally, RStudio supports interactive data visualization for exploring data through charts and plots.

Sampling (vs. Subsetting or Filtering)

- Sampling: The process of selecting an often representative subset of data from a larger dataset, often used in analysis to infer trends from a smaller portion of data. Common methods include random sampling (selecting data points at random) and stratified sampling (dividing the population into subgroups and sampling from each to ensure representation).
- Subsetting: The process of selecting a portion of a dataset based on specific criteria, but without necessarily being representative of the whole. For example, a researcher might subset data to focus on all records from a particular year or region.
- Filtering: A specific form of subsetting where only rows meeting a defined condition are retained. For example, filtering a dataset to include only texts written by a particular author or only data points with a value above a certain threshold.

Secondary Sources
In historical research, secondary sources are scholarly works that interpret or analyze primary evidence to construct arguments about the past. Rather than offering firsthand accounts, these sources draw upon archival materials, official records, newspapers, and other primary documents to provide context, critique, and narrative. Historians rely on secondary sources to engage with existing interpretations and to situate their work within broader scholarly debates. Such sources are commonly found in academic monographs published by university presses, articles in peer-reviewed journals indexed in databases like JSTOR, Project MUSE, or ProQuest Historical Newspapers, and dissertations cataloged through platforms such as the MLA International Bibliography or WorldCat.

Snippets
Short extracts or sections of text, code, or data that are used for quick reference or demonstration. Snippets allow for quick exploration of concepts or examples without needing to engage with an entire dataset or program. In text analysis, snippets may provide contextual examples of how keywords or phrases appear within larger bodies of text, helping researchers understand their usage and significance. In programming, code snippets illustrate small, functional segments of code, often shared to demonstrate specific techniques or solve particular problems.

Social Sciences
The study of human behavior, societies, and institutions using empirical research methods. Key disciplines within the social sciences include sociology, anthropology, political science, economics, psychology, and communication studies. These fields use both qualitative and quantitative methods to analyze social phenomena. The insights from these fields provide insights that inform policy and theory. History is sometimes viewed as a social science because it uses similar empirical methods to study societal behavior and institutions. When historians analyze past social, political, or economic trends using data science approaches or statistical tools, history aligns with the social sciences by seeking to understand patterns and causes of historical events.

Software Package / Library
A collection of prewritten code designed to perform specific actions. Packages may be designed by professionals or members of a community to extend the capabilities of a programming language without needing to write new code from scratch. A library is often a specific type of package that provides reusable functions or classes, typically organized around a particular domain or task, such as natural language processing or data visualization. Examples of packages include `tidyverse` (R), which is used for data manipulation and visualization, and `spaCy` (Python), a library focused on natural language processing. `ggplot2` (R) is another example, a package that provides functionality for data visualization, particularly in the context of the `tidyverse`.

Stop Words
Stop words are common words like "the," "and," "of," or "is" that are often removed in text analysis because they carry little meaning on their own and appear frequently in many corpora. These words are often removed because they can obscure the more meaningful parts of the text. For example, without removing stop words, a word frequency visualization might be dominated by common words like "the," making it difficult to highlight the more relevant or interesting terms. However, the decision to remove stop words is not without its challenges. What qualifies as a stop word can be subjective, depending on the context and the goals of the analysis. Removing stop words can sometimes distort the data by omitting important context that might affect the text's meaning. For instance, many popular natural language processing tools, such as Stanford's NLP toolkit, remove gender pronouns (e.g., "he," "she") by default as part of their stop word lists. While this can be useful for certain types of analysis, it may be misleading when analyzing texts where gendered language is critical to understanding the content. Therefore, researchers must be critical and selective in deciding which stop words to remove, as this decision can significantly influence the results and interpretations of the analysis.

Subsetting
The process of extracting a portion of a dataset based on specified conditions. Unlike sampling, which often is used to create a representative selection, subsetting focuses on isolating specific rows, columns, or values that meet certain criteria for further analysis. This method is useful when researchers want to examine particular partitions of data, such as records from a specific time period, a certain category, or values within a defined range.

Syntax (in Code)
The set of rules that define the structure and format of a programming language. It governs how commands, functions, and expressions must be written in order for the code to execute properly. For example, in R, the correct syntax would be, `my_data <- read.csv("data.csv")`, while the following would be incorrect due to improper syntax, `my data = read csv("data.csv")`. "my data" has a gap between text, which would result in an error. Instead of using the `<-` assignment operator we used `=`, which technically works but is stylistically incorrect for this context. Instead `=` is typically used in function calls. Incorrect syntax can lead to errors, preventing the code from running or producing unintended results. Unconventional syntax can lead to miscommunications between groups working on shared code.

Syntax (in Text)
The arrangement of words and phrases to form grammatically correct sentences in a human language. It dictates how different elements of a sentence—such as subjects, verbs, and objects—are structured to convey meaning clearly and coherently. Syntax plays a crucial role in natural language processing because it influences how meaning is represented within text. In computational text analysis, software like `spacy` are commonly used to analyze sentence structure and perform part-of-speech tagging, which identifies the

syntactic roles of words (e.g., noun, verb, adjective). In **Text Mining for Historical Analysis** syntactic analysis is used to understand the relationships between parts-of-speech in historical texts and demonstrate how analyzing parts-of-speech can provide insight into gender and power dynamics.

Text Mining vs. Generative AI
Involves extracting insights, patterns, or structured data from (often large) text corpora using computational techniques such as word extraction, TF-IDF, named entity recognition, and topic modeling. Text mining is primarily used to analyze existing text data to engage with language features within the data. On the other hand, Generative AI refers to AI systems that generate new text, images, or data based on patterns learned from training data, such as GPT-4 for text generation or Stable Diffusion for image creation. In this way, Generative AI can support creative processes like brainstorming. Understanding the differences between typical text mining approaches and generative AI is important to understand, as generative AI can return incorrect assumptions about history using a confident sounding tone as the generated responses are not bound by the original data's structure, which can affect their accuracy and reliability.

Tidy Data
A structured data format designed to simplify analysis and visualization by making data more human-readable. In tidy data, each column represents a variable, which is any characteristic or feature that can vary across each row For example, in historical analysis, variables could include the date of a speech, the speaker's name, or the location of an event. Each row represents an observation, which is a single instance or record within the dataset—such as a specific speech given by a politician or a historical event on a given day. Each cell contains a value, which is the actual data point or measurement for a given variable and observation, such as the length of the speech or the word frequency of key terms in the speech. This structure makes data easier to process, read, and visualize, and it is particularly useful when using tools like the tidyverse in R, which is designed to work with tidy data formats.

Tidyverse
A collection of R packages developed primarily by Hadley Wickham that are designed to support data science through a unified framework for data processing. The Tidyverse has significantly influenced contemporary data science practice, promoting a philosophy in which data structure is treated as foundational to analytical rigor. Central to its design is the concept of "tidy data," a set of structuring principles articulated in Wickham's 2014 paper "Tidy Data." These principles assert that each variable should form its own column, each observation its own row, and each type of observational unit its own table. This structure facilitates clarity and interoperability across packages. Core Tidyverse tools—including (but not limited to) dplyr for data manipulation, tidyr for reshaping data, and ggplot2 for visualization—operate most effectively when data conforms to this tidy format. Together, these packages have reshaped how researchers conceptualize and engage with data.

Vector
A data structure in programming and data analysis, representing an ordered list of values. Vectors can be of different types, such as numeric (e.g., `c(1, 2, 3)`), character (e.g., `c("word1", "word2")`), or boolean (e.g., `c(TRUE, FALSE, TRUE)`). In machine learning, vector representations, such as word embeddings, map text data into numerical spaces for analysis. These vectors can be visualized and analyzed by historians. For example, in scatter plots, each point in the plot can represent a vector. This allows researchers to identify patterns or trends in the data that may not be easily discernible from raw values. In text analysis, visualizing word embeddings as vectors in two-dimensional space can help uncover semantic relationships, such as words that are closer together in the plot being more contextually related.

Validation
An iterative process that involves assessing the accuracy, reliability, and relevance of claims derived from historical data or models. It requires not only evaluating the results of an analysis but also returning to the original sources from which the data was drawn. This process often involves cross-referencing claims with primary sources, such as archival materials, speeches, or official documents, to ensure that the conclusions align with the historical record. Historians often validate their findings by consulting secondary sources, such as scholarly articles and books, which provide context and alternative interpretations. This cyclical process of validation-—comparing computational results with traditional methods—-contexutally grounds

claims about history, helping to safeguard against misinterpretation or the overgeneralization of data and historical events.

Visualization
The graphical representation of data to facilitate analysis and communication. Visualization transforms complex datasets into interpretable insights. Common visualization techniques include bar charts, which are used for comparisons across categories; scatter plots, which help reveal relationships between two variables; heatmaps, which highlight patterns or correlations within a dataset; and word clouds, which represent the frequency of terms in a text, making it easier to identify dominant themes. In historical analysis, visualization is a key component of storytelling and exploratory data analysis, allowing historians to view trends, compare historical events, and communicate findings.

Vocabulary
The set of unique words or terms that characterize a corpus. They play an important role in guiding historical analysis. Historians can, for example, focus their analysis by examining the natural vocabulary embedded within texts or by curating their own set of terms to steer the output of algorithms. In computational historical analysis, historians often define a set of keywords or topics to track across a corpus, concentrating on terms that represent important historical concepts, such as "democracy," "empire," or "class." This vocabulary can be used to analyze trends over time, compare term usage across regions or political groups, or identify shifts in how key ideas are discussed.

Word Count (Raw)
A simple text metric that indicates the total number of words in a document or corpus, without adjustments for stop words, stemming, or lemmatization. This metric provides a basic measure of the length or size of a text and is often used as a foundation for more complex text analyses. While raw word count can offer useful insights, it does not account for semantic meaning, meaning that it can be less informative for deeper textual analysis. However, it is helpful in providing a preliminary overview of the text's volume, which can guide further analytic processes.

X-Axis, Y-Axis
The X-axis and Y-axis are the two perpendicular axes in a Cartesian coordinate system used for plotting data. The X-axis (horizontal) typically represents the independent variable, such as time, categories, or other variables that are controlled or manipulated. The Y-axis (vertical) represents the dependent variable, which is the outcome being measured, such as frequency, magnitude, or quantity. For example, in a scatter plot of top speakers, the X-axis might represent years, while the Y-axis represents word occurrences.

Terms Not Used in This Work

Text Mining for Historical Analysis operates within a specific methodological framework that prioritizes nuanced, contextually grounded insights from historical data. While this book engages with computational and data science approaches to analysis, certain terms are not used because they refer to distinct methodologies or intellectual traditions with different aims. These terms arise from specific academic frameworks or research practices that focus on analysis that do not necessarily align with the more interpretive, context-driven approach taken in **Text Mining for Historical Analysis**.

For example, "culturomics" and "big history" represent two distinct intellectual traditions that emphasize large-scale, often quantitative analyses of culture and history but do not prioritize the nuances of historical interpretation or the specific, localized details that this work focuses on. Similarly, while quantitative history applies statistical and computational techniques to historical data, **Text Mining for Historical Analysis** emphasizes the interpretive and critical dimensions of engaging historical data, such as by forming research processes guided by both close reading and quantiative modeling, rather than relying solely on statistical modeling to drive conclusions.

Here we define these terms:

Culturomics
A quantitative approach to studying cultural trends through large-scale textual analysis, particularly using

resources like the Google Ngram Viewer, which analyzes the frequency of words or phrases in a vast corpus of digitized texts over time. Practitioners of culturomics aim to uncover statistical patterns in language and culture, such as how certain words or ideas gain or lose prominence. This methodology focuses on identifying long-term trends and correlations in language use, often across broad societal shifts or major historical events. While this approach sounds similar to computational text analysis or corpus linguistics, it tends to emphasize quantitative measurements—such as frequency counts—without delving into the deeper, interpretive or contextual aspects of why these changes occur or what this change might mean for human history. It does not employ the qualitative, humanistic insights that come from understanding the underlying social forces shaping language and history. As such, this book does not adopt culturomics' approach, as it emphasizes a more context-based analysis of historical data situated in close reading rather than focusing solely on statistical patterns.

Big History
A multidisciplinary framework that examines history on an expansive scale and in a way that integrates insights from cosmology, geology, biology, and the social sciences to explore long-term patterns of change across the universe, Earth, and human history. It seeks to provide a grand narrative that spans billions of years, from the origins of the universe to the present day, by identifying broad, overarching trends that shape the course of history. While this work engages with historical patterns, it does not adopt the totalizing scope of Big History, which often emphasizes global, universal processes over specific historical contexts or localized events. Instead, this book focuses on more focused, context-sensitive historical analysis, emphasizing the complexities and specificity of human history rather than attempting to unify all of history into one large-scale narrative. While quantitative readings inherently reduce the complexity of history into data points, we believe a narrative of this scale can be overly reductive.

Quantitative History
A methodology that applies statistical and computational techniques to historical data, often associated with economic history. It involves using data science methods to analyze trends and patterns in historical events, frequently focusing on numerical data such as population growth, economic indicators, or social movements. While this work incorporates quantitative methods, the focus of computational historical thinking is of interpretive and critical analysis rather than purely statistical modeling. While economic forces are immensely important, our aim is not just to quantify historical phenomena through understanding the marketplace, but to understand the underlying social, cultural, and political contexts that shape these trends. This approach enables a deeper engagement with the complexities of history by fostering a holistic understanding of historical narratives—not only those driven by economic factors, but also those shaped by power struggles and sociopolitical dynamics that may initially appear less quantifiable, yet can still be meaningfully analyzed using quantitative methods and may not be explained through economic drivers alone.

Cultural Evolution
A theoretical framework that applies evolutionary principles—-such as variation, selection, and inheritance—-to the process of cultural change. It draws parallels between biological evolution and the way cultures adapt and pass on ideas, practices, and technologies over time. While data science approaches can model cultural shifts and identify patterns in how culture changes, this work does not frame these changes in evolutionary terms. Framing cultural change strictly in evolutionary terms can be problematic because it risks oversimplifying complex social and historical processes, reducing cultural shifts to a linear progression. It also overlooks the role of power dynamics, perhaps fostering the false impression that past cultures were inherently "simpler"—and that this supposed simplicity explains their disappearance—while implicitly portraying the present as more advanced or complex by comparison. Such an angle to thinking about history ignores the diverse range of human experiences that influence cultural change.

Metanarrative
A term from critical theory, particularly associated with Jean-François Lyotard, referring to overarching, totalizing explanations of history or knowledge that attempt to provide a single, unified narrative for all events or experiences. These grand narratives often simplify complex histories by imposing broad, universal frameworks that can obscure the diversity of perspectives and historical nuances. This work instead proposes a paradigm through which the application of multiple algorithms and metrics allow for the analysis of historical data in a multi-dimensional way. By employing diverse computational techniques—-such as various text

analysis algorithms and models-—this approach supports analysts exploration of different aspects of historical data, suggesting that analyzing trends and nuance go hand-in-hand, ensuring that multiple perspectives are considered in the analysis of historical events.

By clarifying these distinctions, we acknowledge the breadth of computational and data science approaches to the humanities and social sciences. These different approaches encompass a wide range of methodologies, theories, and intellectual traditions. While these diverse approaches each contribute valuable insights, we argue that it is important to situate computational historical analyses within a specific methodological framework that emphasizes critical analysis, contextual interpretation, and the nuanced understanding of historical data. By doing so, it seeks to balance the abstraction approaches of data science with the richness of historical context so that digital methods serve as a tool for deeper understanding rather than a replacement for interpretive scholarship.