Content analysis of 150 years of British periodicals

Thomas Lansdall-Welfare[a], Saatviga Sudhahar[a], James Thompson[b], Justin Lewis[c], FindMyPast Newspaper Team[d,1], and Nello Cristianini[a,2]

[a]Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1UB, United Kingdom; [b]Department of History, University of Bristol, Bristol BS8 1TB, United Kingdom; [c]School of Journalism, Media and Cultural Studies, University of Cardiff, Cardiff CF10 3NB, United Kingdom; and [d]FindMyPast Newspaper Archive Limited (www.britishnewspaperarchive.co.uk), Dundee DD2 1TP, Scotland

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved November 30, 2016 (received for review April 21, 2016)

Previous studies have shown that it is possible to detect macroscopic patterns of cultural change over periods of centuries by analyzing large textual time series, specifically digitized books. This method promises to empower scholars with a quantitative and data-driven tool to study culture and society, but its power has been limited by the use of data from books and simple analytics based essentially on word counts. This study addresses these problems by assembling a vast corpus of regional newspapers from the United Kingdom, incorporating very fine-grained geographical and temporal information that is not available for books. The corpus spans 150 years and is formed by millions of articles, representing 14% of all British regional outlets of the period. Simple content analysis of this corpus allowed us to detect specific events, like wars, epidemics, coronations, or conclaves, with high accuracy, whereas the use of more refined techniques from artificial intelligence enabled us to move beyond counting words by detecting references to named entities. These techniques allowed us to observe both a systematic underrepresentation and a steady increase of women in the news during the 20th century and the change of geographic focus for various concepts. We also estimate the dates when electricity overtook steam and trains overtook horses as a means of transportation, both around the year 1900, along with observing other cultural transitions. We believe that these data-driven approaches can complement the traditional method of close reading in detecting trends of continuity and change in historical corpora.

artificial intelligence | digital humanities | computational history | data science | Culturomics

The idea of exploiting large textual corpora to detect macroscopic and long-term cultural trends has been discussed for many years (1, 2), promising to empower historians and other humanities scholars with a tool for the study of culture and society. Many studies have been published over the past few years (3–6), some going as far as to propose a quantitative and data-driven approach to the study of cultural change and continuity, owing as much to the methods of modern genomics as to those of the humanities.

A seminal study of 5 million English-language books published over the arc of 200 years (1) showed the potential of this approach, generating a debate about the possible advantages and drawbacks of this new methodology. The study made various claims about both the evolution of language and that of culture (for example, measuring the time required by various technologies to become established or the duration of celebrity for various categories of people as well as studying changes in English grammar). However, one of the key criticisms was that it was based almost entirely on counting words, ignoring both semantics and context (7). Additional criticism was that it did not cover periodicals (8) and that the data sample might have been biased, representing only those books found in the libraries (9).

A later study (10) discussed the possible benefits of mining corpora of digitized newspapers and proposed the use of "distant reading" techniques (11) in this domain, but it was severely constrained by the tools that it used, which only allowed for the querying of individual words. It concluded by advocating for the use of big data methods for newspaper analysis and proposing specific criteria for the design of such experiments.

Although the "Culturomics" study (1) was based on the idea of introducing quantitative and measurable aspects to the study of cultural change, using high-throughput methods for data acquisition and analysis, additional developments in the field of Natural Language Processing (NLP) now allow for more sophisticated information to be extracted from text, allowing previous criticisms to be overcome in many ways (12, 13).

In this study, following on from a series of articles pioneering the use of high-throughput data for the study of culture (1, 4–6, 14, 15) and drawing on the debate that followed their publication (7–9), we assembled a massive dataset of newspapers and periodicals aimed at verifying or contextualizing some of the findings of the study on books (1) using unique and more refined methods and incorporating into the interpretation of results various valuable lessons learned from the subsequent debate.

We first present *n*-gram trends as used in the Culturomics paper before moving beyond simple word counting methods to incorporate more semantic information about named entities and their properties. The corpus that we assembled is formed by 28.6 billion words from 120 regional or local news outlets contained in 35.9 million articles that were published in the United Kingdom between 1800 and 1950. This sample represents approximately 14% of all regional newspapers published over that period in the United Kingdom and covers

Significance

The use of large datasets has revolutionized the natural sciences and is widely believed to have the potential to do so with the social and human sciences. Many digitization efforts are underway, but the high-throughput methods of data production have not yet led to a comparable output in analysis. A notable exception has been the previous statistical analysis of the content of historical books, which started a debate about the limitations of using big data in this context. This study moves the debate forward using a large corpus of historical British newspapers and tools from artificial intelligence to extract macroscopic trends in history and culture, including gender bias, geographical focus, technology, and politics, along with accurate dates for specific events.

Author contributions: T.L.-W., J.T., and N.C. designed research; T.L.-W., S.S., and N.C. performed research; T.L.-W., S.S., J.T., J.L., and N.C. analyzed data; F.N.T. supplied the data and the description of its generation; and T.L.-W., J.T., J.L., and N.C. wrote the paper.

Conflict of interest statement: F.N.T. is a team within the company FindMyPast. Its main role has been to provide part of the data and relative text.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper are available online at data.bris.ac.uk/data/dataset/dobuvuu00mh51q773bo8ybkdz.

[1]The FindMyPast Newspaper Team: Amy Gregor, Boon Low, Toby Atkin-Wright, Malcolm Dobson, and Richard Callison.

[2]To whom correspondence should be addressed. Email: nello.cristianini@bristol.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.1606380114

newspapers obtained from all of the main geographical regions in the United Kingdom. We made various efforts to ensure that the data sample is as representative as possible of United Kingdom local newspapers, covering all main regions, time periods, and key outlets.

To keep this study focused on the trends that we extract and not on the engineering techniques that were used, we have only made use of methods that have already been deployed in other published studies and can be considered stable. Drawing on the subject expertise of the multidisciplinary research team, knowledge of the historical, media, and sociological context was used to inform each stage of the study design: from the careful selection of newspapers and the selection of keywords to the interrogation and interpretation of the results. Where appropriate, the data queries were sampled and read closely to address potential noise in the optical character recognition (OCR) text or ensure that concepts were being accurately tracked.

The study is intentionally wide-ranging, enabling a broad assessment of the potential of the approach. Given space constraints, the discussion of historical context is rendered necessarily concise. Contextual awareness was, however, central to making sense of the findings. To give an example, analyzing the term “suffragette”—a word popularized by a specific segment of the media as a politicized exercise in “catchword” creation—can only be understood in relation to both the history of media and the history of the struggle for voting rights of women in Britain.

Our hope is to concentrate the attention of the reader on the main important point that we are trying to make: it is possible today to detect long-term trends and changes in history by means of big data analysis of the vast corpora that are becoming available. These findings can include studies about politics, technology, the economy, values, gender, and much more. These trends and changes, which might otherwise go unnoticed, can be discovered by machine, enabling a complementary approach with closer investigation by traditional scholars.

Results

Differences Between Books and Newspapers. A starting point for our study was to compare some results for our corpus with those for the Google books corpus (1), showing the similarities and differences between using a corpus of books and one of newspapers and highlighting that we can find the same trends in our corpus but also, that an analysis of newspapers may be more sensitive to certain cultural shifts—notably because of their closer relationship to current events—than books.

Using a similar approach, we computed the use frequency of 1-grams and n-grams over time, where a 1-gram is a string of characters uninterrupted by a space that includes words (“adventurous” and “puppy”), numbers (“1.215”), and typographical errors (“wrods”), whereas an n-gram is an n-length sequence of 1-grams, such as the phrases “United Kingdom” (a 2-gram) and “in the past” (a 3-gram). The use frequency for an n-gram was computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus that year. We restricted n to three and limited our study to n-grams that occur at least 10 times within the corpus.

We found that the impact of key events, such as coronations, conclaves, wars, and epidemics, was much more obvious in our corpus, with peaks allowing us to identify specific years in which events occurred. For the books corpus, the impact of key events was much less clear (Fig. 1), highlighting that regional newspapers are much closer than books to the events covered in both time and space. Fig. 1 helps to show the differences between the two types of written medium, with newspapers offering a closer representation of historical shifts, whereas books are more reflective in nature and less time-bound (for example, a book’s narrative might be set in the past).

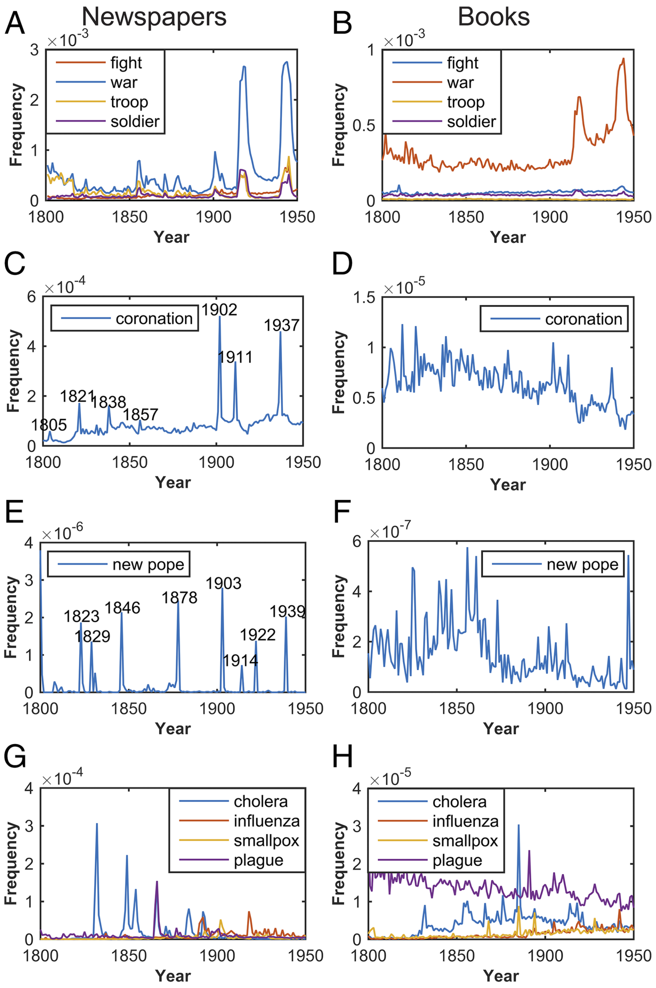

Fig. 1. Comparison between (*A*, *C*, *E*, and *G*) our corpus of British periodicals and (*B*, *D*, *F*, and *H*) the Google books corpus (1) using n-gram trends identifying (*A* and *B*) major wars, (*C* and *D*) coronations, (*E* and *F*) conclaves, and (*G* and *H*) epidemics between 1800 and 1950 in the United Kingdom. Events are clearly identifiable in the periodical corpus, whereas it is more difficult to distinguish exact years of events in the books corpus.

Open-Ended Measurements. We then looked at more open-ended questions, which included measurements of more general and less well-established relations. We divide our analysis into the following spheres: values and beliefs, United Kingdom politics, technology, economy, social change, and popular culture. Again, we selected topics and keywords in a way to avoid ambiguities and performed close reading of some of the articles identified by our analysis to ensure that the keywords represented the intended topic.

In values and beliefs, we test the hypothesis put forward by Gibbs and Cohen (3) of a decline in so-called “Victorian values” during the period under investigation. We find that mentions of certain key Victorian values (3) are in overall decline, although terms like “duty,” “courage,” and “endurance” find new impetus in times of war, whereas other key terms, notably “thrift” and “patience,” do not exhibit a downward trend, qualifying straightforward accounts of the supposed demise of Victorian values (Fig. 2*A* and *B*).

In United Kingdom politics, Gladstone and Disraeli are often seen as the key political figures of the 19th century; however, our findings suggest that Gladstone was significantly more newsworthy during the 19th century itself than Disraeli (Fig. 2*C*). This finding could be partly because of Gladstone’s greater political

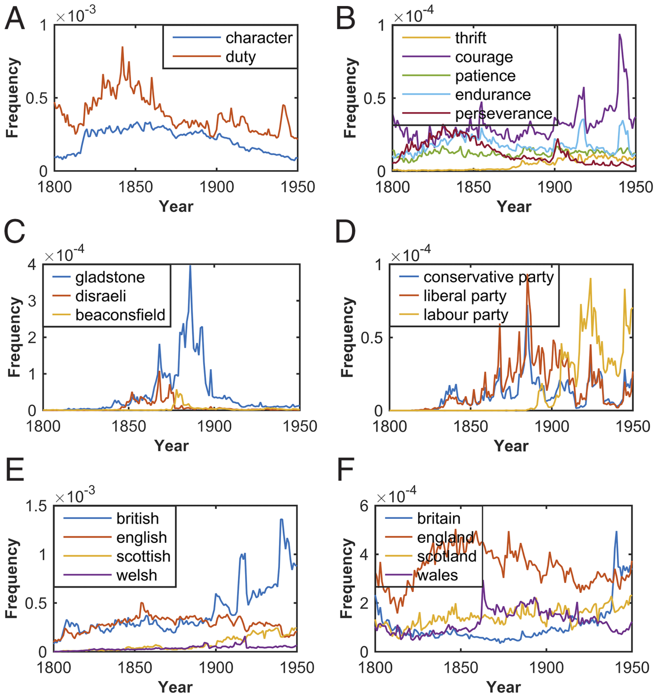

Fig. 2. Values, beliefs, and United Kingdom politics. *n*-Gram trends showing (*A* and *B*) a decline in Victorian values as put forward by Gibbs and Cohen (3), (*C*) that Gladstone was much more newsworthy than Disraeli, (*D*) that liberals are more mentioned than conservatives until the 1930s, and (*E* and *F*) that reference to British identity takes off in the 20th century.

longevity, although it is notable that Gladstone received more coverage even during Disraeli's years as Prime Minister and was a towering figure in press coverage of the period in a way that Disraeli was not.

Overall, the Conservative and Liberal Parties received broadly similar levels of coverage during the 19th century, although they are both eclipsed from the 1920s onward by the Labour Party (Fig. 2*D*). This change cannot, of course, be assumed to reflect levels of political support, but it does suggest that the emergence and growth of the Labour Party was setting the agenda for the regional and local press from 1920 to 1950 (notably after the first Labour Party government in 1924).

Our findings also suggest a very clear timeline in the emergence of "Britishness" as a popular idea, with the term "British" overtaking the term "English" at the end of the 19th century (Fig. 2 *E* and *F*). Thereafter, we see a significant increase in the use of the term British in the first half of the 20th century, with dramatic increases during both world wars. The term English declined during the same period (and indeed, suffers small dips during World War 1 and World War 2)—to such an extent that the term "Scottish" overtakes it in the late 1940s, suggesting that British replaced English as a default national identifier. Although scholarship suggests that the development of Britishness predates this rise (16), these data suggest that the dominance of Britishness in the popular imagination is a 20th century phenomenon.

In technology, we track the spread of innovations in energy, transportation, and communications. In the first area, we observe the steady decline of steam and the constant increase of electricity, with a crossing point in 1898 (Fig. 3*A*). In the area of transportation, we observe how trains overtook horses in popularity in 1902, well after the dawn of the railway age that began in the 1840s, showing the cultural significance of horsepower throughout the 19th century (Fig. 3*B*).

In the area of communications, we examine the rate of adoption of the telegraph, telephone, radio, and television, supporting previous findings (1) that observed an ever-increasing rate of uptake of new technologies that culminated with the rapid rise of television (Fig. 3*C*).

In economy, we find that discussions of the economy as a distinct concept and field began in late Victorian times. The decline in reference to political economy and the growth of reference to the economy manifest the emergence of a sharper idea of the economy as a distinct knowable entity with its own features and rhythms, separable from those of politics (Fig. 3*D*). It is important, however, to note that, on closer reading, reference to the economy seems to be about the need for savings, which is apparent in 1922 and 1932. It is the secular trend evident comparing the economy with political economy that is suggestive.

We also find that it is striking that the term panic emerges as corresponding to volatile downward financial markets without needing to involve concerns about morality or crime, linking clearly when inspected under closer reading to banking crises with pronounced peaks in 1826, 1847, 1857, and 1866 (Fig. 3*E*). This conjecture can be further explained and examined collectively with 19th century press and financial history but would be difficult to express without this complementary, distant-reading approach. More speculatively, it is notable that sampling regional newspapers and thus, mitigating "London-centric" bias, nonetheless, reveal the centrality of financial markets in the City of London to discussions of panic.

In social change, we observe sharp temporal boundaries in phenomena, such as the suffragette movement and the period of anarchist activity; we observe the peaks of unrest that correspond with well-known periods of strike action in 1912 and 1919, whereas the expression revolt corresponds with tension in

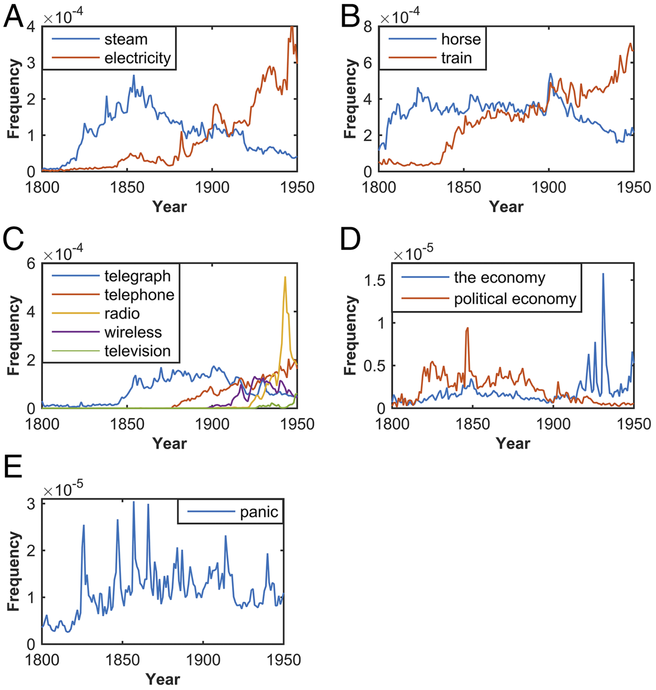

Fig. 3. Technology and economy. *n*-Gram trends showing (*A*) the steady decline of steam and the rise of electricity, (*B*) the waning popularity of horses and the increase in trains, (*C*) the rate of uptake for different communication technologies, (*D*) "the economy" as a concept beginning in late Victorian times after a decline in "political economy," and (*E*) that the four largest peaks for "panic" correspond with negative market movements linked to banking crises in 1826, 1847, 1857, and 1866.

British colonies, notably the Lower Canadian rebellion of 1837–1838 and the "Indian mutiny" of 1857 (Fig. 4*A*).

The frequency of suffragette has a clearly delimited time interval (1906–1918) (Fig. 4*B*), which corresponds with the period from the popularizing of the term in response to the disruption of public meetings to the achievement of suffrage for many, although not all, adult women in 1918. Despite the many years of political campaigning that preceded it, we see a sharp rise in coverage of the suffragettes (and suffragists) following the dramatic death of Emily Wilding Davidson, who was trampled to death by the King's horse at Ascot. This sharp rise in coverage is, perhaps, an early 20th century example of the importance of a "media event" to a political campaign and its ability to capture the journalistic imagination.

The time interval for anarchism is mostly present in the interval from 1882 to 1920, corresponding to the heyday of concern over anarchist direct action before the rise of fascism and bolshevism, whereas slavery includes the movement for abolitionism and the American Civil War (Fig. 4*C*).

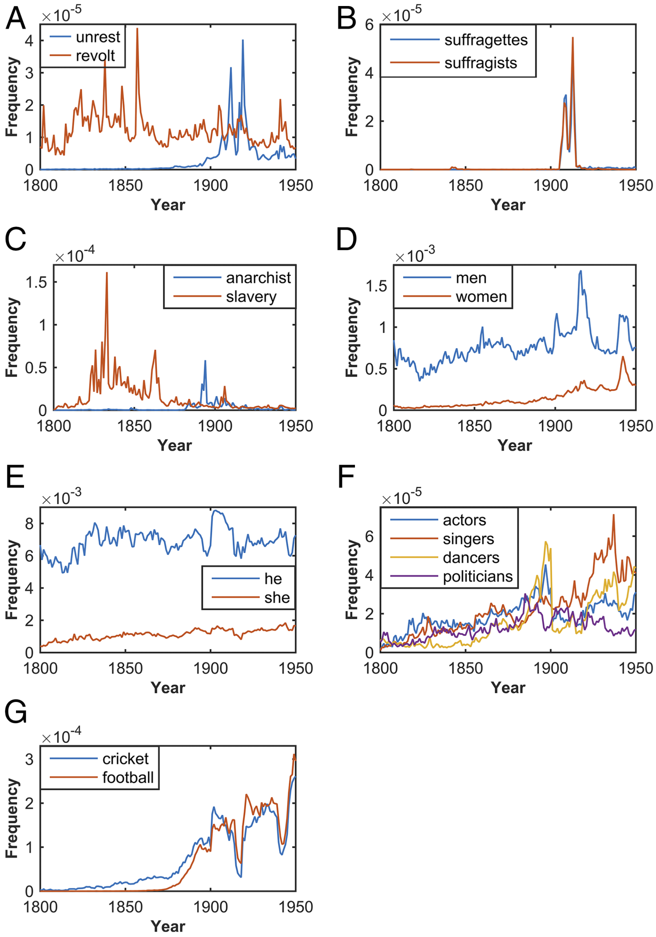

Fig. 4. Social change and popular culture. *n*-Gram trends showing that (*A*) "unrest" corresponds with well-known periods of social tension, whereas "revolt" corresponds with tension in British colonies; (*B*) the suffragette movement falls within a delimited time interval; (*C*) "slavery" includes the movement for abolitionism and the American Civil War, whereas "anarchist" corresponds to the heyday of concern over anarchist direct action before the rise of fascism and bolshevism; (*D*) the gender gap in mentions of men and women is closing, with women making advances during the two wars; (*E*) the gender gap is also closing when measured using the pronouns he and she; (*F*) actors, singers, and dancers begin to increase in the 1890s, rising significantly thereafter, whereas references to politicians, by contrast, gradually decline from the early 20th century; and (*G*) football is more prominent than cricket from 1909 on.

As we might expect, the *n*-gram "men" is mentioned more often than "women," and the same is true for the *n*-gram "he" compared with "she," indicating that we are accessing information about the actual number of men and women in the news. It is interesting to note that the relative proportion of men and women is not very different in today's news (17). Additional analysis with more sophisticated methods is reported below, supporting this conclusion. We can also see a slow increase in the mentions of women and she over the course of 150 years (Fig. 4 *D* and *E*), suggesting a steady increase in the role of women in public life over the whole period, with a more dramatic rise in the consciousness of women as a group in the 20th century during the two world wars. In both cases, we measured the slope of the line of best fit for the time series representing the ratio between the relative frequencies of the *n*-grams women and men as well as that for the *n*-grams she and he, finding both to be positive.

In popular culture, media scholars have documented the growth in human interest news (and the proportionate decline in public affairs), with these data suggesting a clear timeline for the increasing importance of popular culture in news coverage. For example, we see references to "actors," "singers," and "dancers" begin to increase in the 1890s, rising significantly thereafter, whereas references to "politicians," by contrast, gradually decline from the early 20th century (Fig. 4*F*). We see the same pattern in the increasing coverage of *n*-grams "football" and "cricket," with football more prominent than cricket from as early as 1909 (Fig. 4*G*).

Beyond Counting Words. Techniques from NLP allow us to move beyond simply counting word frequencies and focus instead on the frequency with which given entities are mentioned in the text. Named entities include people, locations, and organizations, and references to them can be formed by sets of *n*-grams: generally, multiple references can be used for the same entity. It is possible to automatically resolve these coreferences, therefore creating an automated way to generate multiple *n*-grams related to a given entity.

This step moves us closer to the level of concepts and semantics and also allows us to bypass many of the risks associated with the selection of keywords (*Materials and Methods*). It is further possible to automatically link named entities with existing databases of entities that have recently become available that offer an authoritative list of people, locations, and organizations. These open-source lists include Yago (18) and DBpedia (19), and they allow us to automate the inclusion of external information about different entities that is not present in the corpus itself, such as the gender and occupation of a person or the coordinates of a location. Parsing the text in this way resulted in the extraction of 263,813,326 mentions to 1,009,848 different entities in the corpus.

Discovering every time that a person mentioned within the corpus is also present in DBpedia (19) or another knowledge base often enables us to map them to an occupation type. This procedure allows us to automate the study (1) of fame for people in different careers over their lifetime (Fig. 5*A*).

Among other things, we confirm their finding that politicians and writers are most likely to achieve notoriety within their lifetimes, whereas scientists and mathematicians are less likely to achieve fame; however, we also observe a decline for politicians and writers in news that was not observed in books, whereas time seems to be kinder to scientists and mathematicians. This method has enormous potential for media content analysis, allowing researchers to do widespread and detailed analysis of the sources used in news and explore, for example, the predominant political and ideological affiliation of the sources used in news reporting.

We also extract every single mention of a person in the corpus (regardless of whether they are present in external resources)

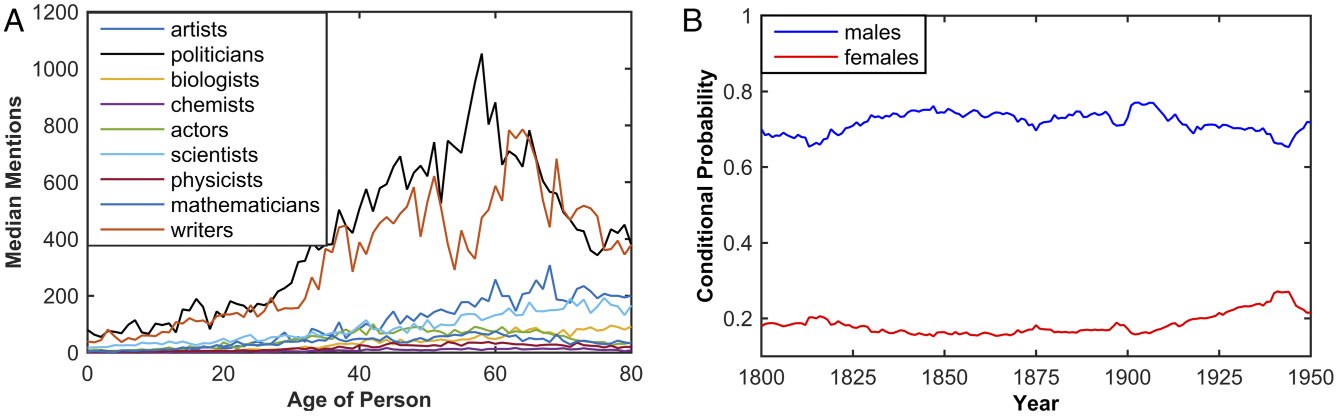

Fig. 5. People in history. (*A*) Replicating the study (1) on famous personalities by occupation using all extracted entities associated with a Wikipedia entry, we found that politicians and writers are most likely to achieve notoriety within their lifetimes, whereas scientists and mathematicians are less likely to achieve fame but decline less sharply. (*B*) We computed the probability that a given reference to a person is to a male or a female person. We find that, although males are more present than females during the entire period under investigation, there is a slow but steady increase of the presence of women after 1900, although it is difficult to attribute this to a single factor at the time.

and infer gender using the ANNIE plugin of GATE, a standard tool for NLP (20). This process gave us over 35 million references to people with a resolved gender, allowing us to calculate the overall probability that a person mentioned in the news is male (or female) and finally, study how this probability changes over time (Fig. 5*B*).

This result confirms—with higher sophistication—the results obtained using the *n*-grams trends (Fig. 4 *E* and *F*), showing that women are consistently represented less than men during the entire period under investigation, and it allows us to explore the nuances and character of various assumptions made about gender. This more refined approach also shows a slow but steady increase of the presence of women after 1900. These results can be read in combination with analogous ones for modern news (17), showing that gender bias within the media does not seem to have changed very much, with approximately three times as many males as females in modern newspapers.

Furthermore, revisiting the concepts that we explored with *n*-gram trends, we compiled geographical maps for the United Kingdom for each of the terms displaying a gradual increase or decline (rather than a spike of activity) (Fig. 6). We extracted all locations found in the articles that mention one of the concepts, disambiguated them again using DBpedia (19), and retrieved their geographical coordinates.

We observe that the terms British and English were reasonably widespread in use across most of the United Kingdom in 1854. By 1940, the use of English had dwindled, with British becoming the default national identifier (Fig. 6*A*).

During 1885, we can see scattered mentions of the Liberal Party around the United Kingdom, with a focus on London, whereas there is very little mention of the yet to be formed Labour Party. However, by 1924, this situation had changed, when the Labour Party achieved its first minority government and replaced the Liberal Party as the party mentioned across the country, again with a geographical focus around London (Fig. 6*B*).

The geographical focuses of technological advances over time were also observed, which we show for the transition from steam to electricity (Fig. 6*C*) and from horses to trains (Fig. 6*D*). For steam, we can see that mentions during its highest use year in 1854 are widespread, with concentrations focused around major ports. However, the adoption of electricity replaces steam by 1947, with electricity being mentioned particularly in reference to London, Leeds, and areas of the Southwest (Fig. 6*C*). During the earliest peak of attention to horse in 1823, we see that mentions are mainly diffused across the country without a distinctive pattern, indicative of their use in rural communities, and there is only the odd mention of train, which on closer reading, was revealed to be generally in a different context (referring to animal training or processions). By 1948, the decline of horse has clearly taken effect, all but disappearing from that map, whereas train is heavily mentioned, particularly around major cities, displaying a similar pattern to that of electricity.

Discussion

The key aim of this study was to show an approach to understanding continuity and change in history based on the distant reading of vast news corpora, which is complementary to the traditional close reading by historians. We showed that changes and continuities detected in newspaper content can reflect properties of culture, biases in representation, or actual real-world events.

With this approach, historians can explore the complex relationship between public discourse and lived experience by detecting trends in statistical signals extracted from large-scale textual corpora. The method is intended to be used in combination with traditional approaches, which are needed for both the design of the study and the interpretation of the findings. Nevertheless, it provides conjectures and answers that would be very difficult to formulate by using close reading alone.

In particular, we showed that computational approaches can establish a meaningful relationship between a given signal in large-scale textual corpora and verifiable historical moments, which was shown in the trends for coronations and epidemics displayed in Fig. 1, and that newspapers provide increased clarity to the analysis of these events that may not be possible in other cultural forms, such as books. We further showed that the approach can reveal or confirm ways in which news media represent particular people or issues over time, as evidenced by the existence of a gender bias that is still present in the media today (17), and that historical trends in public discourse can be made accessible through the same means.

Importantly, this complementary approach provides a layer of cultural understanding that can be used to augment established historical perspectives, evidenced in this study by the temporal and geographical patterns in the uptake of various technologies and concepts shown in Fig 6, which can provide benefit to traditional economic and technical histories of the period.

In this study, special care was devoted to the choice of events that were used for analysis and the keywords chosen to represent them, because we should all be aware of the risk of detecting

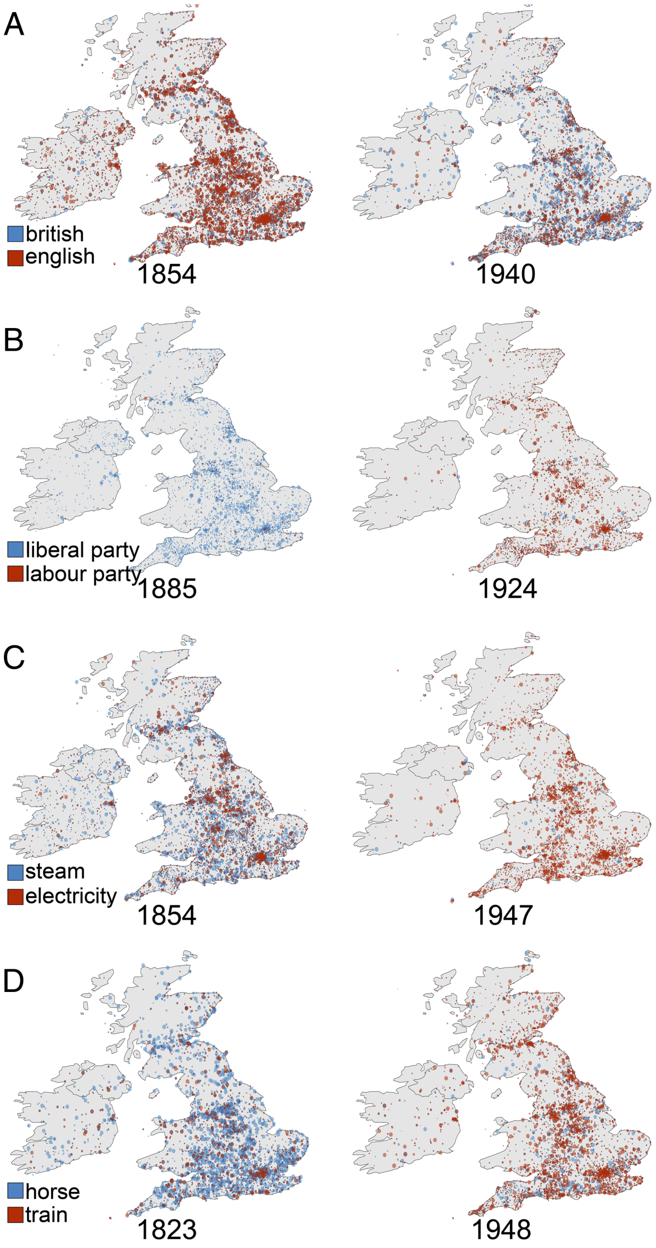

Fig. 6. Changes in geography over time. Maps of the United Kingdom showing the changes in geographical focus of locations extracted from articles containing the terms (*A*) British and English, (*B*) Liberal Party and Labour Party, (*C*) steam and electricity, and (*D*) horse and train for the years in which each concept received its peak attention.

spurious signals in large datasets. As recommended by Nicholson (10), one should try to choose words that have high sensitivity and specificity for the concept being investigated and at the same time, are not too susceptible to semantic shifts and errors in the OCR process. Each of these steps could, in principle, be formalized and automated to some extent: for example, the use of Automatic Query Expansion (21) could likely return a viable set of words that pertains to a specific concept [this set approximating the lexical field (22) of that concept]. However, we feel that it is ultimately the role of the historian to use her judgment and cultural knowledge in the choice of these keywords and the subsequent interpretation of the results. In this way, we focused our analysis on events and keywords that are not ambiguous: coronations, wars, technological artifacts, etc., all of which are events and trends that can be represented by a small set of specific words and often had clear dates attached to them.

Various authors have voiced concern that digital humanities might be just a colonization of the humanities by the sciences, but doing so is not the purpose of this study. On the contrary, we feel that the practice of close reading cannot be replaced by algorithmic means. Indeed, our methods can only detect increased or decreased attention toward a given topic or idea over the decades, offering a complementary approach to close reading, but they cannot explain the reasons behind those changes, which are best understood by other means. We believe, however, that other criticisms are less warranted: the inability of computational methods to introduce contextual knowledge, access semantic information, or work in the presence of OCR noise or the issues related to bias in the original corpus selection are probably all issues that can be solved or accounted for over time.

Future work will indeed include denoising of the data, linking of the data with other corpora or data sources, better disambiguation of entities, and more refined information extraction within a context. Additionally, the evaluation of suitable keywords could be partially automated by including information about OCR noise to help guide the analyst, with recent developments also offering the promise of capturing the extent to which the meaning of a specific word has changed over time (23). These directions are part of engineering work already underway.

What cannot be automated is the understanding of the implications of these findings for people, which will always be the realm of the humanities and social sciences and will never be that of machines.

Materials and Methods

Data Source Background. The British Library's newspaper collections are among the finest in the world, containing most of the runs of newspapers published in the United Kingdom since 1800. The scale of the newspaper publishing industry from the early 19th century onward was enormous, with many cities and towns publishing several newspapers simultaneously and other newspapers that aimed for a wider county circulation providing an unrivalled picture of provincial life spanning the whole of the 19th century and half of the 20th century (24).

In May of 2010, FindMyPast began a partnership with the British Library to digitize millions of pages of these historic newspapers and make them available for the public to search online at www.britishnewspaperarchive.co.uk.

New pages are being scanned all of the time as part of the 10-year project, which once finished, will contain over 40 million newspaper pages from the British Library's newspaper collection. To date, FindMyPast has made available over 12 million pages from 535 different newspaper titles published between 1710 and 1959, adding over 8,000 new pages each day.

The newspaper collection is further supplemented with digitized newspaper records provided by the Joint Information Systems Committee (JISC) that cover the same geographic regions and time period. The digitization of these newspapers was funded by the JISC to provide a representative sample of United Kingdom newspapers spanning all geographical regions, making them suitable for a large-scale automated content analysis of Britain during the 19th and 20th centuries. The data from the JISC form approximately 20% of the resulting corpus.

In this study, we selected a subset of the entire corpus that had been scanned at the time, aiming to assemble a corpus for the study of Britain between 1800 and 1950. To do so, we undertook several significant steps relating to the selection of news outlets to provide a balanced representation in terms of geographic region, time period and quality of the texts, the digitization process and extraction of the associated metadata, and the extraction of information from the raw text of the corpus.

The corpus is accessible under a subscription model at www.britishnewspapersarchive.co.uk, whereas enquires about bulk access to raw data should be directed to FindMyPast. The exact list of articles and newspaper outlets from FindMyPast along with secondary data produced during this study are openly available, including time series of the million most frequent *n*-grams and the 100,000 most frequent named entities extracted by AIDA (25), which are available at data.bris.ac.uk/data/dataset/dobuvuu00mh51q773bo8ybkdz (26).

Data Selection. Newspaper issues were selected from those that had been digitized by FindMyPast from the British Library archives with an eye to form a representative sample of newspapers. The selection was performed by committee, and the criteria for inclusion were (*i*) the completeness of the runs of issues, (*ii*) the number of years that an issue covers, (*iii*) the geographical region that the issue is from, (*iv*) the quality of the OCR output for the issue, and (*v*) the political bias of issues.

Our principal aim was to cover all geographical regions and time intervals as fairly as allowed by the available data. Issues were first separated into the registered geographical region for the publisher, and within each region, newspaper issues were ranked by a combination of the number of years covered (favoring issues with continuous coverage of many years), their average OCR quality, and the total size of data available for the issue. Issues were then selected from the ranking until the region had good coverage. Using domain knowledge, consideration was also taken to ensure that the selection of newspaper issues represents the balance of political opinion in the regional press at the time.

In total, the corpus includes 120 titles selected to best cover the following 12 geographical regions of the United Kingdom using the above criteria for inclusion: East, East Midlands, Northern Ireland, London, Northwest, Northeast, Scotland, Southeast, Southwest, Wales, West Midlands, and Yorkshire.

We estimated the number of regional newspapers within the United Kingdom during the period from 1800 to 1950 using statistics on the number of newspapers in circulation (27). Specifically, we take the average number of papers in circulation from newspaper directories for 1847, 1877, and 1907. This calculation gave us an estimate of 835 newspaper titles in existence through the period. Our corpus contains 120 newspaper titles, giving us an estimate that the corpus covers approximately 14% of the regional papers for the United Kingdom during that time.

Data and Associated Metadata.

Digitization process. The original newspapers were provided by the British Library to the FindMyPast Newspaper Team as either microfilm or the original bound newspapers. Original bound newspapers were scanned using Zeutschel A0 Sheet-Bed Scanners, creating high-quality digital copies of the newspapers as TIFF images at 400 dots per inch (dpi) in 24-bit color before being converted to JPEG2000 format images for archiving. Images created from digitizing the microfilm resulted in grayscale images at 300 dpi.

The raw images created during digitization were digitally cropped, cleaned, and contrast enhanced before being segmented into classified areas corresponding to the type of content, such as news articles, advertisements, and obituaries; structural information, such as page numbers, headers, and footers; and title information, such as issue title, date of publication, and price.

The images were then passed through an OCR process to identify the text used in each section of the page, whereas the associated metadata for each issue were passed through a quality assurance check to correct any mistakes in the structural extraction step.

Data provided via the JISC collection were digitized using a similar workflow through an external supplier (28).

Structure extraction. The raw images from the scans of the original bound newspaper or the microfilm after cropping, cleaning, and contrast correction were next processed in a step to segment each page into classified areas. This process was performed by FindMyPast in two different ways.

The majority of the corpus (78%) was manually segmented into different classified areas relating to the content of the page, structural information, or title information. It was found that this manual process was prohibitively expensive after a certain point within the project, and therefore, the remaining corpus was processed using an automated method using the CCS docWorks software (29).

OCR. OCR was performed on the digital images within the CCS docWorks software (29) by the FindMyPast team. This process outputs the recognized text in the image along with associated information (such as the location and layout on the page) and percentage word accuracies for each word in the standard Metadata Object Description Schema (MODS), Metadata Encoding and Transmission Standard (METS), and Analyzed Layout and Text Object (ALTO) formats (30).

The percentage word accuracy for OCR is calculated automatically by the OCR software and used as a measure of how confident the software is that the characters making up the words were interpreted correctly. Each individual character in each word is assigned a character score between zero and nine (with nine being 100% confidence) for how confident the software is that the character has been read correctly. The overall score for the word was then calculated by taking the average confidence score for the letters that compose it. More widely, the word accuracy score is averaged over a set of words to assign a quality score to how well the text in the image has been recognized at the article level.

Because errors in OCR can be affected by a number of systematic factors, including but not limited to font, size, and physical condition of the original paper copy, the error rate varies across titles; therefore, we aimed to select those titles that had the best possible OCR quality for inclusion in our corpus so as not to detrimentally affect the analysis by introducing low-quality texts. Of course, certain typographical considerations must also be considered when analyzing the data, such as the common use of the long s, which can often be mistaken for an f. Additional work can certainly be done to account for these types of mistakes and will improve in the future.

Overall, in the corpus selected for analysis, the average percentage word accuracy was estimated to be 77.0%, with an SD of 5.78, taking the average score assigned to each newspaper outlet per year by CCS docWorks (29) weighted by volume across all articles in the corpus.

Metadata. The process that was undertaken by the British Library and the FindMyPast Newspaper Team to annotate the data was again managed through the process management pipeline based on CCS docWorks (29). Metadata relative to each outlet were manually entered at the time of the digitization based on the British Library newspaper catalog. The location assigned to each outlet was identified based on the location of original publication. The date was extracted from each newspaper issue by human operators and then, validated during quality control checks. The page segmentation and headline OCR for material processed early in the project were manually corrected by operators; later in the project, these steps were performed without human intervention. A human editor was used to run quality control checks on the structural data extracted by the software, and the workflow software identified systematic issues that were then manually corrected by operators. This process included verifying that the outlet name is correct, the date of the issue is correct, and the pages have been segmented into correctly identified types along with any other quality assurance steps taken.

Automated Content Analysis. After the digitization process had been completed, the FindMyPast team provided the Bristol team with a collection of documents containing the textual content from the newspaper articles along with associated metadata relating to the title of the article, the date of publication, the title that published the article, the location for the publisher, and so forth. Documents were converted from the METS, MODS, and ALTO formats into JavaScript Object Notation (31) documents and stored with their associated metadata in a MongoDB NoSQL collection (https://www.mongodb.com/).

Each document in the database was then subjected to an information extraction procedure (described below), which aimed to allow us to generate time series of any *n*-gram, extract references to entities within the text and resolve the entities, and link the entities to external databases where possible to enrich the information contained within each document.

***n*-Grams.** *n*-Grams were extracted from the main textual content of each document, beginning with tokenizing the text, counting the frequency of each *n*-gram across the entire corpus, and then, filtering the *n*-grams so as to only keep those that occur a minimum number of times (in this case, at least 10 times).

Generation. Raw text data are stored as a string of characters, with no explicit word information. Tokenization splits the string of characters up into a string of words, also referred to as tokens, terms, or 1-grams, for which we can then compute the frequency. Tokenization was performed using the assumption that contiguous sequences of alphabetic characters form a single 1-gram in the vocabulary, which is separated by whitespace characters. Numeric characters that form contiguous sequences are also considered a 1-gram, whereas special characters, such as punctuation, are treated in different ways depending on the specific character. For our purposes, the alphabet used is the Unicode Transformation Format 8 (UTF-8) character set.

The tokenization was implemented using the Word Break rules from the Unicode Text Segmentation algorithm following the specification in the Unicode Standard Annex 29 (unicode.org/reports/tr29/). *n*-Grams were further processed to remove possessives (trailing "'s" at the end of words), lowercased, and stemmed using the Porter stemmer algorithm (32). Tokenization was performed using the Lucene analyzer library available at https://lucene.apache.org/core/4_0_0/analyzers-common/overview-summary.html.

Frequency of n-grams. We calculated the frequency of each *n*-gram (up to a length of three) in the corpus by first counting how often each *n*-gram occurs within each document with a publication date of the same year in the corpus and then, dividing this number by the total volume of terms (1-grams) occurring within the documents published in the same year. This calculation gives a relative importance to each *n*-gram at the time of its use. This calculation was computed within the Hadoop map-reduce framework available at hadoop.apache.org/, allowing us to distribute the computation and work on the large corpus used in this study.

When estimating the relative frequency of an *n*-gram for each year, we also calculated a confidence interval for that estimate using the Yate's score interval (33). The resulting confidence bars were not discernible when plotted, as they were very small because of the very large size of the data used to calculate the time series. As an example, comparing steam with electricity, the size of the change between 1800 and 1950 is at least two orders of magnitude larger than the mean confidence intervals relative to those time points.

Entities. We used standard text engineering tools to extract named entities from the text, linking them with external sources of information where possible. Entities were extracted using AIDA (https://github.com/yago-naga/aida), a framework for entity detection and disambiguation (25) including both person and location types. Additionally, we extracted references to people, including those not necessarily present in any external databases, using the ANNIE plugin of the General Architecture for Text Engineering (https://sourceforge.net/projects/gate/) (20).

Although we note that both of these tools do not have 100% accuracy, in detecting the entity from the text or linking it with the correct information from external sources, it is important to also note that we mitigate the risk by removing those entities for which high confidence could not be achieved as explained below. Although performance of the tools cannot be assessed on the historical corpus (for lack of a "ground truth"), each tool does, however, achieve high performance on benchmarking tasks, with AIDA reporting a mean average precision of 89.05% on the Association for Computational Linguistics' Special Interest Group on Natural Language Learning Conference on Computational Natural Language Learning (CoNLL) 2003 dataset (25), and our entity extraction tool based on the ANNIE plugin for GATE achieved an accuracy of 97.1% on news media from the web (17). Furthermore, for each of the tools, we developed quality control checks and filters to ensure that we only keep the predictions for which we have a high level of confidence, because it should be noted that these tools are not specifically trained on digitized historical newspapers.

For the entities extracted using AIDA (25), although only more prominent people or locations are identified (because they must first appear in an external database, such as Wikipedia), it was sufficient for our purpose of identifying different personalities by their specific occupations (e.g., scientists, writers, politicians, etc.). In doing so, we are able to replicate the Google books study (1) using all personalities that we extracted from the corpus rather than limiting ourselves to the top 25. Overall, there were 263,813,326 mentions of 1,009,848 different entities mentioned in Wikipedia.

This study was performed by grouping all personalities by their occupation types in DBpedia (available for download from wiki.dbpedia.org) (19) as extracted by AIDA before resolving hyponyms to their hypernym occupation type using the WordNet ontology (34) (available at https://wordnet.princeton.edu/wordnet/download/). Personalities were filtered to remove spuriously extracted entities, where entities were identified as spurious by first retrieving the birth date for each entity from DBpedia and then, removing those for which the majority of their mentions occur before the entity is born. In this way, we reduce the number of personalities that has been erroneously linked to a specific Wikipedia entry.

For studying gender balance over the course of history within the corpus, we wanted to avoid any systematic effects caused by our gender detection procedure. A method based on linking entities to DBpedia (which is based on Wikipedia) would likely suffer from the same gender imbalance discovered in Wikipedia (35). Therefore, we used instead the ANNIE plugin of GATE to extract every reference to a person within the corpus and classify their gender into male, female, or unknown using contextual information (such as titles, first names, pronouns, etc.). We considered only those references for which we could obtain an unambiguous gender, discarding more ambiguous entities where we received more than one distinct gender label for the entity. In doing so, we are able to show the number of references to males and females over the course of 150 years in United Kingdom newspapers. In total, there were 25,896,979 unambiguous references to males, 10,198,490 unambiguous references to females, and only 309,098 ambiguous references (assigned to both genders by the tool) found within the corpus, showing that we can unambiguously find the gender of an entity for 99.15% of the entities in the historical newspaper corpus.

We additionally compared our findings with those coming from the Google Books *n*-gram corpus (1) along with our own results using the independent *n*-gram method. This combined use of large numbers of references and the comparison with independent sources of information gives us confidence that we can separate the signal from the noise.

Locations were also extracted using AIDA (25), disambiguating each mention of a location with its Wikipedia page. Geographical coordinates were retrieved from DBpedia for each location or parsed from the live Wikipedia page when no coordinates were resolved from DBpedia.

Geographical focus maps of different concepts, such as British or train as displayed in Fig. 6, were generated by visualizing all locations that were present in any article containing the concept *n*-gram and that occur a minimum of three times in any article containing the concept *n*-gram in the same year. This threshold was used to both filter very low-frequency locations and obtain a more readable map. Location markers were sized according to a combination of the natural log of their total mentions in the corpus (more mentioned locations are given greater weight) and the probability that, within a given year, a location is mentioned in the same article as the concept *n*-gram (the size of the intersection between a concept and a location with a year is a measure of how related they are at that time).

Statistical Robustness of Methods. When working in a data-driven high-throughput way, which is the case in this distant reading project, it is necessary to automate most steps, and this automation does create the problem that each step might introduce errors: OCR will corrupt some characters, named entity recognition might fail to recognize a location, and disambiguation steps might link an entity to the wrong entry in external resources. However, the size of the dataset and careful design can be used to mitigate this risk.

Our focus is on detecting large statistical patterns, and therefore, we can tolerate less than perfect performance at each stage of the analysis and still extract a reliable signal—if we carefully design the analysis. In this way, we are not different from previous Culturomics studies (1), and as Nicholson (10) observed, "this is the price one must pay to explore the 'vast terra incognita' of 19th century print culture."

Sanity checks performed at the end of the pipeline show that indeed—for all of the errors that may be introduced—we can still reliably detect historical events, such as coronations, wars and epidemics. Among the many design choices involved in the study, we compare relative frequencies of a given word (e.g., train vs. horse) or relative changes in the ratio between male and female entities, ensuring that we are comparing signals that are affected by the same type of noise. Additional sanity checks, by comparing time series generated by words, such as he and she or men and women, with those generated by the overall mention of male and female entities, show that any noise found in the processing pipeline does not cancel the signal.

On the Selection of Keywords and Other Signals. As pointed out by Nicholson (10), one of the key design choices in these studies is the selection of keywords. There are various risks involved in this step: a word might not represent well the concept under investigation, perhaps because it is ambiguous, or it might not be semantically stable during the period under study; perhaps that word might not be robust under OCR noise. Indeed, we might want to look at several words to represent a concept (as we did for Victorian values) or sometimes, entire lexical fields (22). Our approach has been to use judgment based on historical knowledge for the assessment of the relevance and stability of each word, make use of carefully selected lists of words already used in previous relevant studies, assess keywords by close reading some of the articles matching them, and use automated means to go beyond counting words and therefore, bypass the risks associated with selecting keywords entirely.

There is a second risk involved in the selection of keywords: when mining vast corpora, there is always the risk of finding a spurious signal (for example, a time series that has accidental resemblance with some historical trend). The risk is higher when using high-throughput methods because of the statistical phenomenon of "multiple testing": even if each keyword has a very low chance of showing accidental correlations, when we can analyze tens of thousands of keywords, this risk is multiplied accordingly. The problem is further increased by the inherent ambiguity of the tasks described in this study: the lexical field (22) relative to an event or cultural phenomenon is not well-defined a priori, and therefore, there is significant freedom for the analyst to—involuntarily—select words that confirm a hypothesis.

These risks can be mitigated by various technical and statistical approaches. For example, making use of precompiled lists of keywords from previous studies, such as the Victorian values, is a standard statistical method to account for multiple testing by reducing the space of possible testing being conducted, whereas it should also be possible to generate a list of keywords that relate to a specific concept by using techniques from the field of Automatic Query Expansion (21), therefore approximating its lexical field. However, ultimately, it will be the job of the analyst to make careful judgments and use the findings with the necessary care. We have made every effort to select nonambiguous terms and events to avoid the risk of generating a spurious signal, ensuring that we generate the keywords for analysis in a way that is independent of their temporal behavior in the corpus.

ACKNOWLEDGMENTS. We thank Ilias Flaounas for his help in the initial stages of the study and Patrick Fleming and the JISC for making their corpus available. Part of the work was carried out using the computational facilities of the Advanced Computing Research Center, University of Bristol. This study was made possible by FindMyPast Newspapers Archive Ltd (www.britishnewspaperarchive.co.uk), which shared the data. Most University of Bristol members of the team (T.L.-W., S.S., and N.C.) were supported by European Research Council Advanced Grant ThinkBIG awarded to N.C.

1. Michel JB, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.
2. Reddy R, StClair G (2001) *The Million Book Digital Library Project*. (Carnegie Mellon University, Piittsburgh). Available at www.rr.cs.cmu.edu/mbdl.htm. Accessed December 19, 2016.
3. Gibbs FW, Cohen DJ (2011) A conversation with data: Prospecting Victorian words and ideas. *Vic Stud* 54(1):69–77.
4. Mauch M, MacCallum RM, Levy M, Leroi AM (2015) The evolution of popular music: USA 1960–2010. *R Soc Open Sci* 2 (5):150081.
5. Leetaru K (2011) Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday* 16(9).
6. Flaounas I, et al. (2013) Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital Journalism* 1(1):102–116.
7. Gooding P (2013) Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Lit Ling Comput* 28(3):425–431.
8. Morse-Gagné EE (2011) Culturomics: Statistical traps muddy the data. *Science* 332(6025):35.
9. Schwartz T (2011) Culturomics: Periodicals gauge culture's pulse. *Science* 332(6025):35–36.
10. Nicholson B (2012) Counting culture; or, how to read Victorian newspapers from a distance. *J Vic Cult* 17(2):238–246.
11. Moretti F (2013) Distant Reading (Verso Books, London).
12. Borin L, et al. (2013) Mining semantics for culturomics: Towards a knowledge-based approach. *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, eds Liu X, Chen M, Ding Y, Song M (ACM, New York), pp 3–10.
13. Suchanek FM, Preda N (2014) Semantic culturomics. *Proc VLDB Endowment* 7(12):1215–1218.
14. Lansdall-Welfare T, Sudhahar S, Veltri GA, Cristianini N (2014) On the coverage of science in the media: A big data study on the impact of the Fukushima disaster. *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data)*, eds Lin J, Pei J, Lin TY (IEEE, New York), pp 60–66.
15. Flaounas I, et al. (2010) The structure of the the EU mediasphere. *PLoS One* 5(12):e14243.
16. Colley L (2005) *Britons: Forging the Nation, 1707–1837* (Yale Univ Press, New Haven, CT).
17. Jia S, Lansdall-Welfare T, Sudhahar S, Carter C, Cristianini N (2016) Women are seen more than heard in online newspapers. *PLoS One* 11(2):e0148434.
18. Suchanek FM, Kasneci G, Weikum G (2007) Yago: A core of semantic knowledge. *Proceedings of the 16th International Conference on World Wide Web*, eds Williamson C, Zurko ME, Patel-Schneider P, Shenoy P (ACM, New York), pp 697–706.
19. Lehmann J, et al. (2015) DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semant Web*, eds Williamson C, Zurko ME, Patel-Schneider P, Shenoy P 6(2):167–195.
20. Cunningham H, Maynard D, Bontcheva K, Tablan V (2002) GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ed Isabelle P (ACL, Stroudsburg, PA), pp 168–175.
21. Carpineto C, Romano G (2012) A survey of automatic query expansion in information retrieval. *ACM Comput Surv* 44(1):1–50.
22. Öhmann E, Trier J (1931) *Der Deutsche Wortschatz im Sinnbezirk des Verstandes* (C. Winter, Heidelberg).
23. Weston J, Ratle F, Mobahi H, Collobert R (2012) Deep learning via semi-supervised embedding. *Neural Networks: Tricks of the Trade*, eds Montavon G, Orr GB, Mller K-R (Springer, Berlin), pp 639–655.
24. Findmypast Newspaper Archive Limited (2016) About the British Newspaper Archive. Available at www.britishnewspaperarchive.co.uk/help/about. Accessed September 26, 2016.
25. Hoffart J, et al. (2011) Robust disambiguation of named entities in text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, eds Merlo P, Barzilay R, Johnson M (Association for Computational Linguistics Stroudsburg, PA), pp 782–792.
26. Lansdall-Welfare T, et al. (2016) FindMyPast Yearly n-Grams and Entities Dataset. Available at data.bris.ac.uk/data/dataset/dobuvuu00mh51q773bo8ybkdz. Accessed December 19, 2016.
27. Walker A (2006) The development of the provincial press in England c. 1780–1914: An overview. *Journal Stud* 7(3):373–386.
28. Shaw J (2009) British Library Newspapers Digitisation Report. Available at www.webarchive.org.uk/wayback/archive/20140614080134/www.jisc.ac.uk/media/documents/programmes/digitisation/blfinal.pdf. Accessed September 26, 2016.
29. CCS (2016) Content Conversion Specialists - Digitization Services. Available at https://content-conversion.com/#digitization-services. Accessed September 26, 2016.
30. Impact Centre of Competence in Digitisation (2016) Recommendations on Formats and Standards Useful in Digitisation. Available at www.digitisation.eu/training/recommendations-for-digitisation-projects/recommendations-formats-standards-recommendations/. Accessed September 26, 2016.
31. Bray T (2014) *The JavaScript Object Notation (JSON) Data Interchange Format* (IETF, Fremont, CA).
32. Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3):130–137.
33. Wallis S (2013) Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *J Quant Linguist* 20(3):178–208.
34. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to wordnet: An on-line lexical database. *Int J Lexicogr* 3(4):235–244.
35. Wagner C, Garcia D, Jadidi M, Strohmaier M (2015) It's a man's wikipedia? Assessing gender inequality in an online encyclopedia. *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, eds Quercia D, Cha M, Mascolo C, Sandvig C (AAAI Press, Palo Alto, CA), pp 454–463.