nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-019-0620-8

Predicting history

Joseph Risi[1,4], Amit Sharma[2,4], Rohan Shah[2], Matthew Connelly[3*] and Duncan J. Watts[1*]

Can events be accurately described as historic at the time they are happening? Claims of this sort are in effect predictions about the evaluations of future historians; that is, that they will regard the events in question as significant. Here we provide empirical evidence in support of earlier philosophical arguments[1] that such claims are likely to be spurious and that, conversely, many events that will one day be viewed as historic attract little attention at the time. We introduce a conceptual and methodological framework for applying machine learning prediction models to large corpora of digitized historical archives. We find that although such models can correctly identify some historically important documents, they tend to overpredict historical significance while also failing to identify many documents that will later be deemed important, where both types of error increase monotonically with the number of documents under consideration. On balance, we conclude that historical significance is extremely difficult to predict, consistent with other recent work on intrinsic limits to predictability in complex social systems[2,3]. However, the results also indicate the feasibility of developing 'artificial archivists' to identify potentially historic documents in very large digital corpora.

Almost two centuries ago, Hegel[4] observed that "the owl of Minerva spreads its wings only with the falling of the dusk", meaning that philosophical understanding of an era is possible only as it is ending. Arthur Danto, an analytical philosopher, developed this intuition into a theoretical argument for why the judgement of history is intrinsically unpredictable[1]. Danto imagined a hypothetical entity—the ideal chronicler—that possesses complete information about the state of the world, along with its entire history up to that point, and unlimited ability to integrate and analyse that information. Danto then argued that the ideal chronicler would still be unable to anticipate the significance that will be assigned to contemporaneous events by future historians because their judgements are inevitably informed by events that have not yet taken place. Evaluating the historical significance of a single event would in principle require the ideal chronicler to accurately predict not only its impact on subsequent events, but also how those events would impact broader social structures and practices[5]. For example, to have anticipated the historical significance of the storming of the Bastille on 14 July 1789, the ideal chronicler would have had to predict not only the whole sequence of events leading to the French Revolution, but also their transformation of political and philosophical notions of sovereignty—a profound structural shift that has arguably shaped all subsequent revolutionary movements[6].

Danto's argument is compelling in theory, but as a practical matter it may still be the case that some, or even most, historically significant events could be identified at the time they are happening or shortly thereafter. For example, while the full significance of the Bastille took many years to develop, contemporaneous accounts show that within days it was perceived as a legitimate act of popular revolution, the implications of which might well have already seemed historic. One can also easily think of other events, such as Einstein's formulation of special relativity, the moon landing, the fall of the Berlin Wall and the discovery of the double-helical structure of DNA, that were immediately hailed as historic and continue to be viewed that way decades later. Certainly, in our current era, politicians, journalists and experts alike do not experience any difficulty in deciding what future historians will judge to be important: a simple web search for 'historic moment' reveals more than 100 usages per month. So, who is right?

Conceptually, one way to answer such a question would be with the aid of a hypothetical dataset that satisfies three key properties. First, it would constitute a complete census of events. Second, the events would be well defined and separable, allowing unambiguous evaluation. Third, each event would be labelled with two importance scores (one assigned by contemporary observers and one by future historians). Imagine, for example, a team of experts who develop a codebook for defining events and rating their historical significance. These experts would train until they achieved high intercoder reliability, and then set about evaluating all candidate events as they occur, including many that may seem trivial. Decades or even a century or more later, another team would use the same coding criteria and training regimen to evaluate the same data. One could then simply compute the classification accuracy of the contemporaneous experts using standard metrics such as precision and recall.

By definition, a dataset of this sort would take at least decades to construct. It also presumes that historians treat events as discrete and uniform units of analysis when, in fact, events in the historical literature are extremely heterogeneous. For example, events such as the assassination of Archduke Ferdinand and the negotiations leading to the signing of the Treaty of Versailles occurred on two very different time scales and could even be conjoined into a single event on yet a third time scale. World War I and World War II, in turn, could be described as distinct or as parts of something bigger: a Thirty Years War, a Forty Years War or an even longer "Crisis of the Middle Powers".[7] Debates about how to categorize and periodize such complex phenomena in fact lie at the very core of historical scholarship and are rarely resolved to universal satisfaction. Confronted with this potentially insoluble conceptual ambiguity, we proceed in much the same way as historians do when attempting to identify and define events: by turning to the written historical record; that is, to collections of documents that have been preserved and curated by archivists. Given limited time and storage capacity, both archivists and historians must continually make judgements about which documents will be most valuable in representing, illuminating and explaining events and trends. In practice, therefore, when historians debate the relative importance of different events, they typically cite different historical documents.

[1]Microsoft Research New York, New York, NY, USA. [2]Microsoft Research India, Bangalore, India. [3]Department of History, Columbia University, New York, NY, USA. [4]These authors contributed equally: Joseph Risi, Amit Sharma. *e-mail: mjc96@columbia.edu; duncan@microsoft.com; djwatts@seas.upenn.edu

Thus motivated, we now outline a more tractable version of the hypothetical exercise above—one that combines digital archives of the sort that are commonly used by historians with machine learning methods capable of labelling millions of documents automatically. Specifically, we study a collection of 1,952,029 US Department of State (DOS) cables from 1973 (when the State Department began storing them as electronic records) to 1979 (the last year for which these records have been declassified and released to researchers; http://history-lab.org/cables). Addressing the first desired property, cables were the dominant mode for communicating important information between the DOS in Washington and embassies and consulates around the world. While our dataset clearly represents only a small portion of all contemporaneous reports of events taking place within the 1973–1979 period, it is highly likely that any matter of importance in US foreign affairs will be mentioned in this collection. Equally important for our purpose, the vast majority of cables describe routine and often unimportant matters (for example, travel schedules of various officials and embassy social functions) associated with the routine business between the United States and other states; thus, consistent with our idealized design, our collection includes many events that would have seemed trivial at the time for each one that seemed important.

Addressing the second criterion, we treat individual cables as proxies for different events. To be sure, some events, such as treaty negotiations, may play out over extended periods and thus be the subject of many cables. Conversely, more than one event could be referenced in a single cable. In practice, however, diplomats were trained to summarize essential information as concisely as possible, both because of the limited time of senior policymakers and the limited carrying capacity of communications and encryption systems. They were also required to identify and classify the subject of each cable, and break up longer, more complex reports and analyses into multiple communications, such as when a meeting dealt with several subjects. Although imperfect, cables can therefore serve as references to both simple and complex phenomena, which is why scholars so often cite and quote from cables to represent historical events.

Finally, addressing the third criterion, we created two sets of labels: one that reflects the perceived importance of each cable in the eyes of contemporaneous observers (in this case, the professional diplomats and other US State Department employees who authored the cables); and the other that reflects the perceived importance of the same cable to professional historians decades later. For the former, we constructed a perceived contemporaneous importance (PCI) score $\rho_i = \sum_j w_{ij}$ for each cable, i, where $-10 \leq w_j \leq 20$ are weights assigned to 22 metadata fields (weights were assigned manually by expert human annotators; see Supplementary Methods for details). Entries in these fields, such as how the cable was to be handled, how it was classified and whether it was designated for high-level attention, can be interpreted as the author's assessment of the importance of the content (see Methods). For the latter, we consulted a second document collection entitled the *Foreign Relations of the United States* (FRUS). By law, the Office of the Historian of the DOS must certify that the documents in FRUS constitute a "thorough, accurate, and reliable" account of US foreign policy[8]. They include diplomatic cables, transcripts of telephone calls, planning memoranda and other official records, compiled decades after the events in question by professional historians with the explicit goal of conveying historically important information (http://history-lab.org/frus). Although the collection dates to 1861, here we are exclusively interested in the volumes spanning 1973–1979 (the period that corresponds to the first electronic State Department cables). Out of the total collection of nearly 2 million cables, only 1,723 (0.09%) were selected for inclusion in FRUS, consistent with our requirement that only a small fraction of records should be considered historically important.

Using these data, we trained and evaluated a series of statistical models to answer three questions that are motivated by Danto's hypothesis. First, to address the question of how well contemporaneous impressions of importance reflect future historical judgements, we evaluated how accurately the PCI score of a given cable can predict its subsequent appearance in FRUS. Second, to address the question of how well any contemporaneous entity, human observer or hypothetical ideal chronicler can anticipate historical importance, we evaluated the predictive accuracy of a machine learning classifier that uses all available information for the same set of cables. Finally, to interpret our findings, we manually inspected a selection of true positive, false positive and false negative cables, to identify features associated with each.

Addressing our first question, we first checked whether the PCI score contains any information relevant to the FRUS versus non-FRUS distinction. Figure 1a shows P_{FRUS}—the probability of a cable being included in FRUS—as a function of the PCI score for four samples of cables: (1) all 1,723 FRUS cables and a randomly sampled subset of the same number of non-FRUS cables (that is, a 1:1 sample); (2) all FRUS cables and a randomly sampled subset of $50 \times 1{,}723 = 86{,}150$ non-FRUS cables (a 50:1 sample); (3) all FRUS cables and a randomly sampled subset of $100 \times 1{,}723 = 172{,}300$ non-FRUS cables (a 100:1 sample); and (4) the complete set of all 1,723 FRUS and 1,950,306 non-FRUS cables (that is, a 1,132:1 ratio). The 1:1 non-FRUS:FRUS sample corresponds to a 'balanced' classification task wherein the classifier is presented with two cables (one FRUS and the other non-FRUS) and asked to identify which is which. The larger sample sizes correspond to increasingly difficult, and more realistic, classification tasks, wherein the classifier must distinguish the FRUS cable from an ever-larger collection of non-FRUS cables.

Figure 1a shows a positive and monotonically increasing dependency of P_{FRUS} on the PCI score, confirming that the PCI score does indeed contain relevant information about P_{FRUS}. For example, of the ten cables with the highest overall PCI scores, eight relate to the most important events in this period, such as the Iranian Revolution, the second Strategic Arms Limitation Treaty, the Vietnamese invasion of Cambodia and the civil war in El Salvador. However, Fig. 1a also shows that the relationship becomes progressively weaker and less informative as the sample size increases. For example, while for the 1:1 sample the difference between the highest and lowest PCI scores can discriminate between FRUS and non-FRUS with near certainty, even the highest scoring cables in the full (that is, 1,132:1) sample are only four percentage points more likely than the lowest scoring cables to belong to FRUS. Figure 1b summarizes the same information, showing the correlation between PCI score and P_{FRUS} as a function of sample size. Consistent with Fig. 1a, PCI score is positively correlated with P_{FRUS} for all four samples, but also declines steadily as the ratio of non-FRUS to FRUS increases, from 0.89 for the 1:1 sample to 0.45 for the full sample.

Next, Fig. 1c shows the precision–recall curves for a single sample of each of the 1:1, 50:1, 100:1 and full sample sizes, along with the corresponding maximum value of the F1 score (see Methods). For the 1:1, 50:1 and 100:1 samples, we repeated this exercise for 10 randomly drawn samples of non-FRUS cables and computed the average best F1 score and standard errors. Figure 1d shows the F1 score and standard error for each sample size, along with corresponding scores for a naive or 'zero information' model that predicts that every cable will be in FRUS. Although extremely simple, the naive model is useful as a baseline against which we can compare our model performance. For example, the naive model has perfect recall (by construction), and hence does reasonably well for the balanced sample, but its precision (and hence F1 score) gets progressively worse as the number of non-FRUS cables in our sample increases. Similarly, the PCI model performs well for the 1:1 sample ($F1 = 0.86$) and its performance also decreases as the sample size

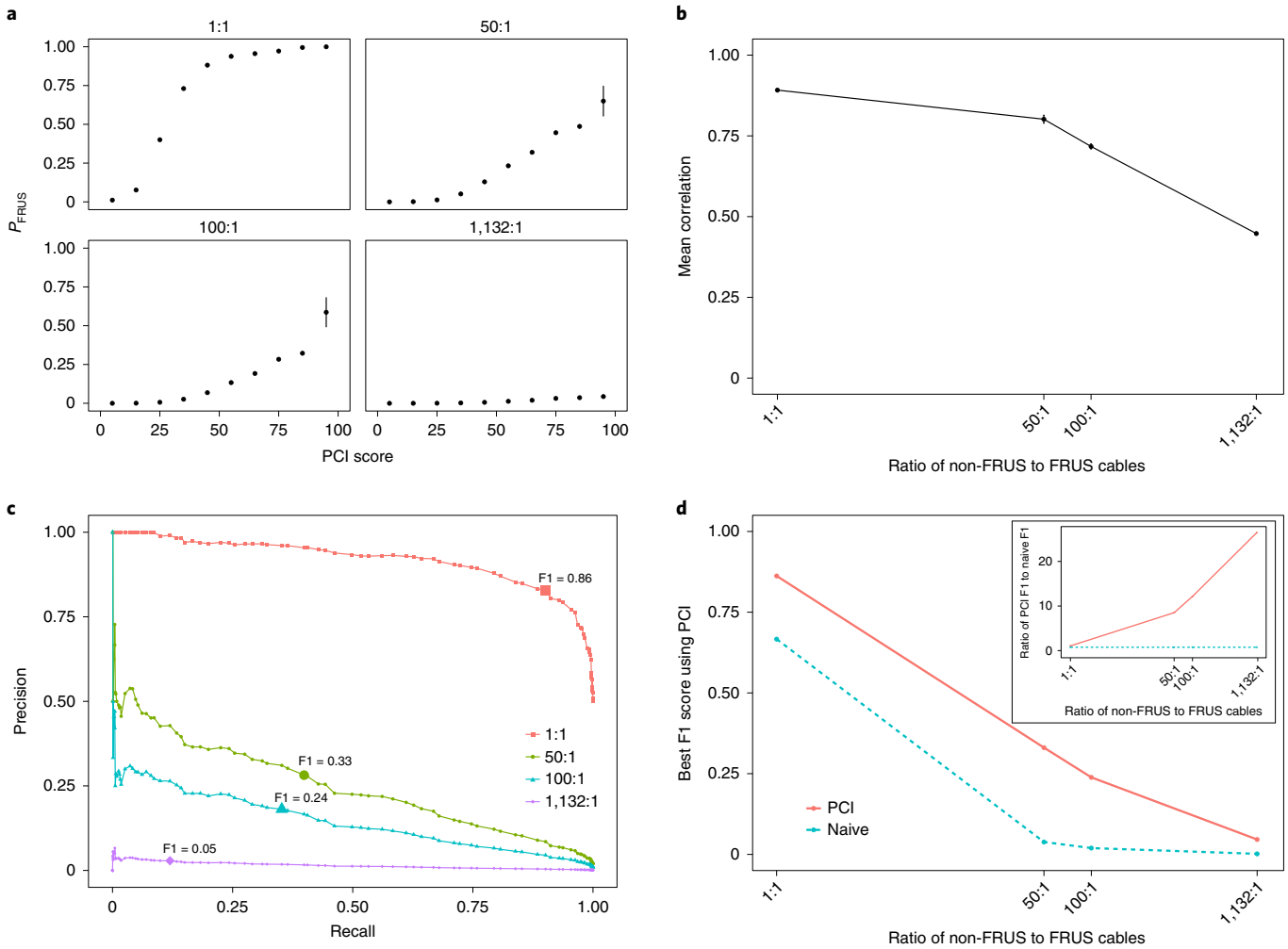

Fig. 1 | Performance of the PCI model. a, P_{FRUS} versus the PCI score. **b**, Correlation of the PCI score with P_{FRUS} as a function of the sample size. **c**, Precision–recall curve for the PCI predictor (maximum F1 score indicated). **d**, Maximum F1 score versus the sample size. Inset: ratio of the maximum F1 score for the PCI model to the naive model F1 score. Error bars represent s.e. and are, in most cases, smaller than the symbols.

increases. However, compared with the naive model, the performance of the PCI degrades more slowly with increasing sample size, hence its relative performance increases correspondingly (Fig. 1d, inset).

These results provide mixed support for Danto's hypothesis. On the one hand, for the smallest sample, the PCI model is highly reliable in absolute terms, and dramatically outperforms the naive model even for large samples. Both of these results suggest that what contemporary observers judged to be important contained some signal regarding their cables' future historical significance. On the other hand, the diminishing F1 scores for larger sample sizes show that contemporaneous labels become rapidly less indicative of historical significance as the level of 'noise' (that is, non-significant cables) increases. In a world where the number of historically insignificant events vastly outnumbers significant events—which almost certainly is the world in which we live—Danto's scepticism about the ability of contemporary observers to estimate historical significance seems warranted.

However, addressing our second question, we note that Danto's argument applies not only to the authors of the cables themselves, but also to any conceivable observer, including his hypothetical ideal chronicler who can make use of any information that could, in principle, have been knowable at the time. The ideal chronicler, in other words, can build up an arbitrarily complex and sophisticated predictive model that is not limited by the intuition of contemporaneous humans about what is and is not likely to be important. Moreover, the ideal chronicler can learn the sorts of events that turn out to be historically important based on its observations of which events have previously been labelled as important by historians. We operationalize this idea by training a type of machine learning classifier known as a gradient-boosted decision tree[9,10]. To avoid overfitting[11], we fitted the model on a subset of the data comprising all FRUS cables from the years 1973–1978, along with four matching sets of non-FRUS cables, corresponding to the same ratios as above (1:1, 50:1, 100:1 and 1,132:1), finding the set of weights that minimizes the classification error in the predictions over the training set (see Supplementary Methods for more details on the model). We then evaluated the model—now with fixed weights—on a separate 'held-out' dataset comprising just FRUS cables from 1979 and the same multiples of non-FRUS cables as used in the training stage.

This simulated ideal chronicler (SIC) goes beyond the PCI model in two key respects. First, whereas the PCI score incorporates a small number of metadata fields into a single feature, the SIC model uses many thousands of features, including additional metadata as well as the cable's content (that is, the message text). Specifically, we characterized the content of each cable using a topic model[12]—a type of unsupervised machine learning model that treats each cable (or 'document') as a mixture of 'topics', which are in turn treated as

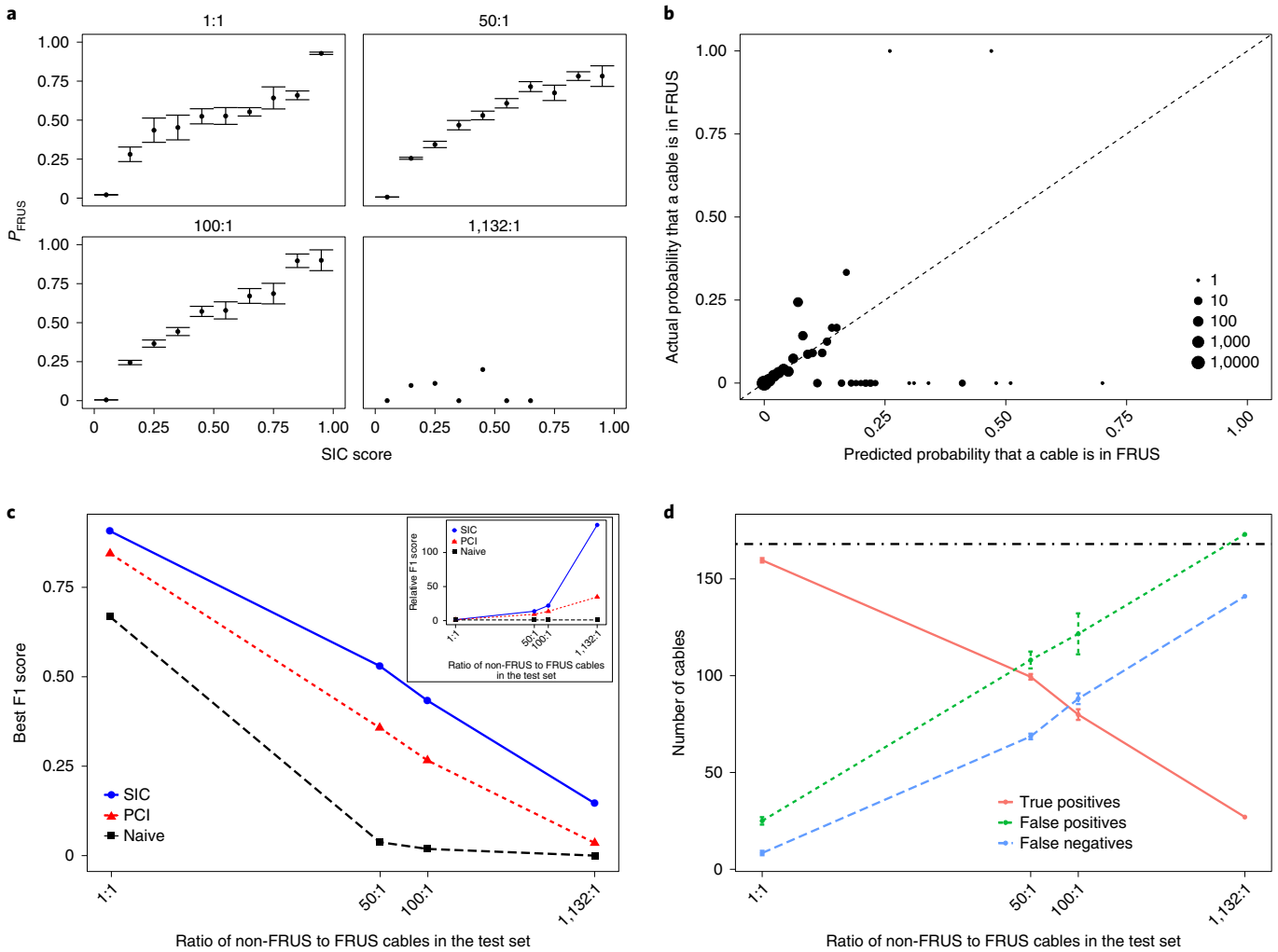

Fig. 2 | Performance of the SIC model. a, P_{FRUS} versus SIC score. **b**, Calibration of the SIC score for the 1,132:1 sample. **c**, Maximum F1 score versus the sample size for the SIC model (blue line), PCI model (red line) and naive (zero information) model (black line). Inset: ratio of SIC and PCI model maximum F1 scores to the naive model F1 score. **d**, Number of true positives (red line), false positives (green line) and false negatives (blue line) as a function of sample size. The dot-dashed line shows the total number of FRUS documents ($n = 168$). In all panels, error bars represent s.e. and are, in most cases, smaller than the symbols.

distributions over 'words'. Topics generated in this manner comprise lists of words with weights that correspond to the frequency with which each word appears in each topic. Topic models have been used previously to identify historical trends in corpora of newspapers[13,14], scientific papers[15–17] and transcripts of political debates[18]. Here, we treat the resulting topics as features that predict the historical significance of the corresponding document. For our purpose, we estimated a topic model with $K = 500$ topics, effectively adding up to 500 new features to each cable (see Supplementary Methods for details). A second important difference between the SIC and PCI models is that, whereas the weights assigned to features in the PCI model are predetermined, the SIC learns the optimal weights from the training data (as with the PCI model, it is evaluated on the held-out test data; see Methods). In other words, the SIC is given as many advantages as can be allowed without knowing specifically which historical records in the test set (that is, 1979) will be judged important enough to be included in FRUS.

Figure 2 shows that the SIC model performs substantially better than the PCI model. Figure 2a shows P_{FRUS} versus the SIC model score for all four sample sizes. Compared with Fig. 1a, the SIC scores are better calibrated (that is, closer to the diagonal) than the PCI scores. Figure 2b shows an expanded view of the 1,132:1 sample, with the circles sized by the number of observations, indicating that the mass of the distribution lies in the lower left corner. As with the PCI model earlier, next we considered the maximum F1 score attained using a threshold on the SIC score. Figure 2c shows the corresponding F1 scores for both models (the naive model is again included for comparison), showing that although performance diminishes with sample size for all models, the relative difference in performance increases: for the full sample, the maximum F1 score for the SIC model is four times better than the model using PCI alone. To understand the relative improvement, we note that the F1 score can increase either because the number of true positives goes up, thereby increasing both precision and recall, or because the number of false positives goes down, thereby increasing precision. We also note that the total number of positive cases is necessarily the sum of true positives and false negatives, and it is a constant (Fig. 2d); thus, false negatives cannot vary independent of true positives. As shown in Table 1, the improved F1 score is driven both by a decrease in false positives (173 versus 753) and by an increase in true positives (27 versus 17), but the former is considerably larger than the latter.

Put another way, whereas for the PCI model the by-far larger source of error is false positives, for the SIC model, false positives

Table 1 | Confusion matrices for the PCI model and SIC model

	FRUS	Non-FRUS
PCI model		
Predicted FRUS	True positive = 17	False positive = 753
Predicted non-FRUS	False negative = 151	True negative = 320,733
SIC model		
Predicted FRUS	True positive = 27	False positive = 173
Predicted non-FRUS	False negative = 141	True negative = 321,313

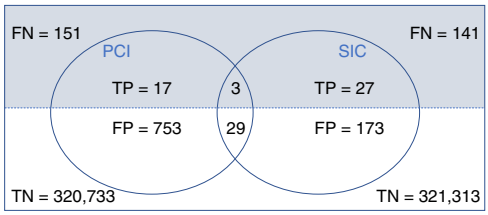

Fig. 3 | Venn diagram of true and false positives and negatives for the PCI and SIC models. FN; false negatives; FP, false positives; TN, true negatives; TP, true positives.

and false negatives contribute roughly equally (173 and 141, respectively). The high false positive rate for the PCI model is consistent with our opening observation about the frequency with which contemporary observers proclaim events to be historic, and also suggests that it might be possible to avoid some of these mistakes by paying close attention to the features of events that do ultimately register as historic. However, as shown in Fig. 3, the SIC model does not outperform the PCI model simply by making fewer of the same mistakes, but rather by generating largely distinct sets of true positives and false positives. For example, of the 17 FRUS cables correctly predicted by the PCI model, only 3 are also predicted by the SIC model, even though the SIC model correctly predicted 27 FRUS cables in total. Although this result seems puzzling at first, it follows naturally from the way in which machine learning models are evaluated. To elaborate, both PCI and SIC models produce scores that can be interpreted as P_{FRUS}. To classify the cable as FRUS or non-FRUS, we also require a threshold τ such that if $P_{\text{FRUS}} > \tau$, we label it a positive prediction; otherwise, we label it a negative prediction. Naturally, increasing τ will result in fewer false positives (higher precision) but also more false negatives (lower recall); thus, depending on the precise distribution of model scores and the current value of τ, F1 could be increased either by increasing or lowering τ. As both the distribution and the optimal value of τ will be different for different models, there is no guarantee that the true positive set for one model should be a subset of the other.

These results suggest two tentative conclusions: (1) more complex models outperform simpler ones, albeit with sharply diminishing marginal returns (see Supplementary Methods for more models and discussion); and (2) interpreting the differences between models is non-trivial and requires additional evidence. To address the second observation, we examined the 25 highest-ranking false positives for the PCI model (that is, cables that the models were most confident would be in FRUS but were not). These cables were almost always classified 'secret', and frequently had high-level 'Cherokee' and/or 'flash' urgency designations. The top two PCI false positives were typical in that they reported on highly sensitive contacts with foreign leaders (see http://history-lab.org/documents/1979STATE040191 and http://history-lab.org/documents/1979STATE266494). The SIC model correctly identified both as true negatives, in large part because it incorporated the topics of communications. For example, it favoured cables involving more consequential political negotiations and military strategy. The very highest-ranked true positive concerned ministerial-level negotiations over Namibian independence, including the military implications (see http://history-lab.org/documents/1979SECTO03013). The SIC model also ranked, in the top 0.03% of all cables, one titled "*Passing the Torch: Saddam Is Solidly In Charge*". It had no urgency or handling restrictions and received a wide distribution. However, in the SIC model, the topic, with terms such as 'Iraq', 'Baghdad' and 'Saddam', was a more important feature than a secrecy classification (see http://history-lab.org/documents/1979BAGHDA01528).

In fact, the performance of the SIC model may have been better than Fig. 2 indicates. The reason is that, as noted earlier, it is likely that events that seemed important at the time would have been the subject of more than one cable (for example, negotiations between Egypt and Israel, South Africa and Namibia, and Saddam Hussein's increasingly autocratic rule). However, in such cases, it is unlikely that FRUS historians would have included all of the relevant cables, either because they did not personally vet all 1.95 million cables or because they elected to include only a representative sample of those they did vet. If true, it follows that at least some cables that are currently scored as false positives might be more appropriately classified as true positives. To check the robustness of our findings to this possibility, we recomputed the precision of the SIC model under increasingly relaxed requirements for a positively classified cable to 'match' the text of a FRUS cable, finding that precision was essentially invariant (see Supplementary Methods for details). We also manually inspected all 200 cables the SIC model identified as historic, finding that of the 173 false positives 83 were associated with the same subjects as the true positives. Assuming that all of these matches are really true positives, the true F1 score for the SIC model for the full sample would increase from 0.14–0.24. Although substantial, even this increase would not qualitatively alter the shape of the performance curve in Fig. 2a, especially as we would expect much smaller effects for the 1:1, 50:1 and 100:1 samples (as they are much less likely to contain redundancies). However, it would mean that false negatives would become the dominant source of error for the SIC model. In contrast with the PCI model, in other words, the problem for the SIC model is not so much that events that it thinks will be historic turn out not to be (although that is still a problem), but rather that it fails to identify which seemingly unimportant events will later be recognized as significant. To illustrate, the FRUS cable with the second lowest SIC score is part of a review of a departmental policy that precluded hiring gay people (https://history.state.gov/historicaldocuments/frus1969-76v30/d88). It is the only one like it in all of the FRUS volumes from 1973–1979, so the term 'homosexuality' does not rank high in any of the 500 topics; also, the cable is identified as administrative and is not classified[8]. Looking back, however, scholars can now see this brief message as reflecting an important moment in the history of lesbian, gay, bisexual and transgender rights.

To conclude, how do these results help us to evaluate Danto's provocative claim that the historical significance of events cannot be known at the time? On the one hand, the direction and magnitude of our results clearly contradict the claim that nothing whatsoever can be predicted about history. Although the PCI model tended to badly overstate the importance of contemporary events—consistent with anecdotal evidence from news reports—the SIC model performed much better and may have performed even better than our analysis allows due to the redundant mapping of cables to events. On the other hand, our results also show that historical significance is extremely hard to predict as the number of candidate events grows large. Figure 2 makes this point most directly, showing that as the ratio of non-FRUS to FRUS cables increases, even the best-performing model performs poorly. Also, whereas PCI errors are

more likely to be false positives than false negatives, for the SIC model the main problem is likely to be false negatives. As the true number of events is almost certainly vastly larger than even our very large corpus allows, the difficulty faced by Danto's ideal chronicler of simultaneously minimizing false positives and false negatives appears overwhelming.

Therefore, on balance, our results suggest that Danto was substantively correct. As the number of events being evaluated grows, successful predictions will be increasingly outnumbered by events that seem insignificant at the time, but which come to be viewed as important by future historians in part because of events that have not yet taken place. More generally, our results provide further evidence for the observation that the combination of nonlinearity, stochasticity and competition for scarce attention that is inherent to human systems poses serious difficulties for ex ante predictions[2,3,19]—a pattern that has previously been noted in outcomes such as political events[20], success in cultural markets[21], the scientific impact of publications[22] and the diffusion of information in social networks[23]. Given that historical significance is typically evaluated on longer time scales than these other examples, it is especially vulnerable to unintended consequences[24], sensitivity to small fluctuations[25] and reinterpretation of previous information in light of new discoveries or societal concerns. A further complication is that historical significance, even when it can be meaningfully assigned, is specific to observers whose evaluation may depend on their own idiosyncratic interests and priorities. Although we speak of history as a single entity, in reality there may be many histories, within each of which the same set of events may be recalled and evaluated differently[5].

Despite these difficulties, we close by noting that research using algorithmic predictions may nonetheless prove useful to historians in practice. As we have already discussed, some false positives of the SIC may in fact be considered true positives that were omitted by the FRUS historians, or simply overlooked among millions of other cables. This result suggests that our SIC could serve as an artificial archivist: a machine learning model that could identify a set of potentially significant documents for human consideration that is much smaller than the full corpus but contains a large portion of the true positives. For example, a classifier resembling our SIC model, but optimized for recall rather than F1 score, might correctly identify more than 80% of FRUS cables. Although such a model would also have a higher rate of false positives than the model presented here, and hence worse overall performance, it could still enable historians to focus on the small percentage of records likely to include those they would have rated as historic without assistance, along with many more they might otherwise have missed. Testing this concept would also address a more basic question: to what extent do scholars agree about which records are historic? Alternatively, are they more likely to agree with a machine learning algorithm than they are with one another? Regardless, given the ever-increasing volume of historical data (including billions of emails, text messages and social media posts), we predict that even less-than-ideal chroniclers will become increasingly important for the future of historical research.

Methods

Calculation of PCI score. To evaluate the perceived importance of the event or issue, we constructed a PCI score from the rich metadata associated with every cable in our dataset. Cable metadata include: basic metadata (for example, subject, date, from, to, classification and length); tags (a fixed list of 1,749 predefined fields; for example, 'SHUM' for human rights, 'UR' for Soviet Union and 'PINT' for internal political affairs, from which authors can select one or more that apply); and concepts (an open-ended list of 33,357 descriptors; for example, 'surplus weapons disposal', 'negotiations', 'cat-C' and 'religious discrimination') that authors can append to further refine the subject and/or nature of the cable. Some metadata were specifically intended to demarcate the relative importance of different communications. For instance, the 'Cherokee' descriptor in the concept field designated a cable as requiring presidential or secretary of state attention, 'cat-A' was for the attention of other senior officials, and so on. Likewise, 'eyes only' handling instructions and a higher (secret) classification label identified communications with more sensitive information, which is indicative of perceived importance. Conversely, cables also had 'tags' that identified some communications as administrative or routine in nature. On the basis of a close reading of hundreds of cables, we identified 22 pieces of metadata that we judged to be either positively or negatively related to perceived importance. Reflecting these qualitative judgements, we assigned a weight $-10 \leq w_j \leq 10$ to each feature j, where the sign and magnitude of w_j indicates the direction (for example, negative weight lowers importance) and salience of the feature (see Supplementary Table 2 for a complete list of features included in the index, and Supplementary Methods for justification of the corresponding weights). For each cable i, we then computed a PCI score $\rho_i = \sum_j w_{ij}$ defined as the sum of weights w_{ij} for the associated field j. Clearly, the PCI score is an imperfect proxy for what contemporaries thought; although it is based on a close reading of the cables by two historians familiar with the period in question, the choices of features and weights comprising the PCI score were necessarily subjective. In future work, it would be desirable to evaluate alternative scoring rules or to identity corpora that contain more explicit indicators of the author's intent. Nevertheless, a major advantage of this approach is that once the w_j values are fixed, PCI scores can be computed objectively and automatically, meaning that much larger corpora can be consistently coded than would be feasible with human coders.

Calculation of precision–recall curves. To compute the predictive power of the PCI score, we first converted it from an integer value between 0 and 100 to a binary prediction (that is, either FRUS or non-FRUS) by choosing a threshold value y^* such that when $y_i > y^*$, the cable is predicted to be in FRUS, and otherwise not. For any y^*, the model performance can now be quantified in terms of:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

and

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

where a true positive is a cable whose PCI score $y_i > y^*$ and appears in FRUS, a false positive is a cable whose PCI score $y_i > y^*$ but does not appear in FRUS, a true negative is a cable whose PCI score $y_i < y^*$ and does not appear in FRUS, and a false negative is a cable whose PCI score $y_i < y^*$ but appears in FRUS. Setting $y^* = 0$, for example, means that all cables will be classified as belonging to FRUS, resulting in perfect recall but only at the cost of near-zero precision (because of the large number of false positives). Conversely, as $y^* \to 1$, the precision improves (because only cables about which the model is most confident are classified as belonging to FRUS) but recall decreases (because the number of false negatives necessarily increases). The result is a precision–recall curve from which we wish to select a single point that represents the best performance of the model. Because both precision and recall are relevant to our task (that is, we would like to minimize both false positives and false negatives), we choose the value of y^* to maximize the F1 score, defined as the harmonic mean of precision and recall:

$$\text{F1} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

SIC models. We operationalized the idea of an ideal chronicler by training a series of increasingly complex machine learning models on the years 1973–1978, and then testing on data from 1979, where each successive model adds new features, as follows: model 1 = PCI features only; model 2 = PCI features + basic metadata; model 3 = PCI features + basic metadata + tags and concepts; model 4 = PCI features + basic metadata + tags and concepts + topics. For all models, we encoded FRUS cables as the positive class and non-FRUS cables as the negative class. Given cables from 1973–1978, the task is to predict which cables from 1979 will be in FRUS. Model 1 uses the same features as the PCI model above (the PCI score), where the difference is that instead of simply computing ρ_i, we fitted a logistic regression model $y_i = 1/(1 + \exp \beta\rho_i)$ where y_i is the probability that cable i will be selected for inclusion in FRUS, and β is a regression coefficient. For all of the remaining models, we used boosted decision trees[5]—a far more powerful machine learning method than logistic regression. Specifically, the XGBoost algorithm that we implemented is widely used in machine learning contests, where it has been found to perform as well or better than other prediction and classification methods. In addition to the above models, we also included a naive (zero information) model that predicts every cable will be in FRUS. As such, the naive model necessarily has perfect recall, but it also has extremely poor precision, and hence a poor F1 score, which gets progressively worse as the number of non-FRUS cables in our sample increases. Although extremely simple, the naive model is useful as a baseline against which we can compare all of the fitted models.

When tested on the cables from 1979, the models output the probability of membership in FRUS for each cable. As with the PCI model, to evaluate predictive performance, we computed precision and recall at different threshold values y^* of the predicted probability y_i: if the probability is higher than y^*, the model

predicts that the cable is in FRUS. For easy comparison, we again computed the maximum F1 score across all values of y^* for each model. Supplementary Fig. 1 summarizes the predictive performance of all of the models described above for 4 different samples of cables from the year 1979: a 1:1 sample comprising all 1,723 FRUS cables and an equal number of randomly selected non-FRUS cables; a 50:1 sample comprising 50 times as many non-FRUS as FRUS cables; a 100:1 sample comprising 100 times as many non-FRUS as FRUS cables; and the full set of cables.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The original data (that is, cable text and metadata) analysed during the current study are available in the History Lab repository (http://history-lab.org) and US National Archives (https://aad.archives.gov/aad/series-description.jsp?s=4073). Derivative data (for example, model scores) are available at https://osf.io/nhmcd/.

Code availability

All code necessary to reproduce our results is available at https://osf.io/nhmcd/.

Received: 30 October 2018; Accepted: 25 April 2019;
Published online: 03 June 2019

References

1. Danto, A. C. *Analytical Philosophy of History* (Cambridge Univ. Press, 1965).
2. Martin, T., Hofman, J. M., Sharma, A., Anderson, A. & Watts, D. J. Exploring limits to prediction in complex social systems. In *Proc. 25th International Conference on World Wide Web* 683–694 (International World Wide Web Conference Committee, 2016).
3. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
4. Hegel, G. W. F. *Hegel's Philosophy of Right* (Clarendon Press, 1942).
5. Bearman, P. S., Faris, R. & Moody, J. Blocking the future: new solutions for old problems in historical social science. *Soc. Sci. Hist.* **23**, 501–533 (1999).
6. Sewell, W. H. Historical events as transformations of structures: inventing revolution at the Bastille. *Theory Soc.* **25**, 841–881 (1996).
7. Kennedy, P. M. *The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500 to 2000* (Random House, 1987).
8. McAllister, W. B., Botts, J., Cozzens, P. & Marrs, A. W. *Toward "Thorough, Accurate, and Reliable": A History of the Foreign Relations of the United States Series* (US Department of State, 2015).
9. Schapire, R. E. in *Nonlinear Estimation and Classification* 149–171 (Springer, 2003).
10. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
11. Provost, F. & Fawcett, T. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media, 2013).
12. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
13. Newman, D. J. & Block, S. Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inf. Sci. Technol.* **57**, 753–767 (2006).
14. Yang, T.-I., Torget, A. & Mihalcea, R. Topic modeling on historical newspapers. In *Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 96–104 (Association for Computational Linguistics, 2011).
15. Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proc. Natl Acad. Sci. USA* **101**, 5228–5235 (2004).
16. Blei, D. M. & Lafferty, J. D. A correlated topic model of science. *Ann. Appl. Stat.* **1**, 17–35 (2007).
17. Hall, D., Jurafsky, D. & Manning, C. D. Studying the history of ideas using topic models. In *Proc. Conference on Empirical Methods in Natural Language Processing* 363–371 (Association for Computational Linguistics, 2008).
18. Barron, A. T., Huang, J., Spang, R. L. & DeDeo, S. Individuals, institutions, and innovation in the debates of the French Revolution. *Proc. Natl Acad. Sci. USA* **115**, 4607–4612 (2018).
19. Watts, D. J. Common sense and sociological explanations. *Am. J. Sociol.* **120**, 313–351 (2014).
20. Tetlock, P. E. *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton Univ. Press, 2005).
21. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
22. Clauset, A., Larremore, D. B. & Sinatra, R. Data-driven predictions in the science of science. *Science* **355**, 477–480 (2017).
23. Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. Everyone's an influencer: quantifying influence on Twitter. In *Proc. 4th ACM International Conference on Web Search and Data Mining* 65–74 (Association for Computing Machinery, 2011).
24. González-Bailón, S. *Decoding the Social World: Data Science and the Unintended Consequences of Communication* (MIT Press, 2017).
25. Ferguson, N. *Virtual History: Alternatives and Counterfactuals* (Hachette UK, 2008).

Acknowledgements

The authors are grateful for support from the team at Columbia University's History Lab. This work was supported in part by the National Science Foundation (SBE-1637108). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.J.W. and M.C. conceived and designed the experiments. J.R., A.S. and R.S. performed the experiments and analysed the data. M.C. and R.S. contributed materials. D.J.W. and M.C. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0620-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C. or D.J.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Duncan Watts

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We collected data on State Department Cables using backend MySQL database access provided by history-lab.org, which maintains a digital collection of declassified cables obtained from the U.S. National Archives. We wrote queries in SQL to download the data and extracted comma-separated (CSV) files for analysis.
Data analysis	We wrote custom scripts in R (version 3.x) to do descriptive and predictive analysis. In particular, we used open-source MALLET library (v2.0, http://mallet.cs.umass.edu/) for topic modeling, and open-source xgboost library (v0.7, https://cran.r-project.org/web/packages/xgboost/) for training machine learning models. We also used standard libraries in R for analyzing and plotting data such as dplyr and ggplot2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data (i.e. cable text and metadata) analysed during the current study are available in the History Lab repository, http://history-lab.org and also the US National Archives https://aad.archives.gov/aad/series-description.jsp?s=4073. Derivative data (e.g. model scores) will be made publicly available on OSF.io upon publication.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We apply machine learning models to classify DOS cables from 1973-1979 as belonging/not belonging to FRUS
Research sample	1) A collection of 1,952,029 U.S. Department of State Cables (DOSC) from 1973 – when the State Department began storing them as electronic records – to 1979, the last year in which these records have been declassified and released to researchers. 2) 1723 of the same cables contain in a document collection titled the Foreign Relations of the United States (FRUS).
Sampling strategy	We analyze four samples of cables: (a) all 1723 FRUS cables and a randomly sampled subset of the same number of non-FRUS cables (i.e. a 1:1 sample); (b) all FRUS cables and a sampled subset of 50x1723=86,150 non-FRUS cables (a 50:1 sample); (c) all FRUS cables and a randomly sampled subset of 100x1723=172,300 non-FRUS cables (a 100:1 sample); and (d) the complete set of all 1723 FRUS and 1,950,306 non-FRUS cables (i.e. a 1132:1 ratio).
Data collection	We collected data on State Department Cables using backend MySQL database access provided by history-lab.org, which maintains a digital collection of declassified cables obtained from the U.S. National Archives. We wrote queries in SQL to download the data and extracted comma-separated (CSV) files for analysis.
Timing	1973-1979
Data exclusions	N/A
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging