Font-Pomarol et al. *EPJ Data Science* (2023) 12:2
https://doi.org/10.1140/epjds/s13688-022-00376-0

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Socially disruptive periods and topics from information-theoretical analysis of judicial decisions

Lluc Font-Pomarol[1], Angelo Piga[1,2], Rosa Maria Garcia-Teruel[2], Sergio Nasarre-Aznar[2], Marta Sales-Pardo[1] and Roger Guimerà[1,3*]

*Correspondence:
roger.guimera@urv.cat
[1]Dept. Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia
[3]ICREA, 08017 Barcelona, Catalonia
Full list of author information is available at the end of the article

Abstract

Laws and legal decision-making regulate how societies function. Therefore, they evolve and adapt to new social paradigms and reflect changes in culture and social norms, and are a good proxy for the evolution of socially sensitive issues. Here, we use an information-theoretic methodology to quantitatively track trends and shifts in the evolution of large corpora of judicial decisions, and thus to detect periods in which disruptive topics arise. When applied to a large database containing the full text of over 100,000 judicial decisions from Spanish courts, we are able to identify an abrupt change in housing-related decisions around 2016. Because our information-theoretic approach pinpoints the specific content that drives change, we are also able to interpret the results in terms of the role played by legislative changes, landmark decisions, and the influence of social movements.

Keywords: Judicial decisions; Topic model; Information theory

1 Introduction

Technological advances have facilitated the generation and storage of digital documents stemming from human activities, from financial transactions to medical records or drug prescriptions. These digital traces open the door to understanding human behavior in new ways [1]; digital documents allow to analyze the temporal evolution of the interactions between social actors, making it possible to infer the sociological and cultural processes beneath human activities [2]. In this endeavor, computational efforts are critical to automatize the processing and extraction of information from large-scale corpora of documents [2]. Indeed, computational methods enable the quantitative analysis of the content of documents and their evolution; they are powerful tools to understand the underlying social processes by capturing trends and patterns that result from the prevalence, extinction or substitution of specific practices and ideas [2, 3].

One of the last domains to enter the digitization era is that of legal studies. In recent years, despite the fact that there are still some barriers that prevent open access to digital court records [4], there has been a steady increase in the availability of digital documents

Springer

related to court activities (such as judicial decisions or processes, especially in the US and Europe) and legal processes in general [5]. In fact, despite some reluctance to incorporate evidence-based methodologies in the study of legal processes [6–8], some voices have advocated for systematic and quantitative approaches for which the availability of digital legal documents is crucial [4, 8–11]. Recently, systemic, computational approaches have been able to start extracting and analyzing large corpora of judicial decisions [7, 12], which has enabled the study of the use and propagation of precedent through the network of citations between judicial decisions [11, 13–17]. In this sense, while a thorough reading of a decision is the only way to fully comprehend the legal reasoning, computational analysis can uncover large-scale patterns in the legal system as a whole [18, 19].

Besides enabling the systematic study of legal processes, digitized legal documents are also a good proxy for the evolution of sensitive social issues. Indeed, because of their key role in determining how societies function, law and legal decision-making are subject to public opinion [20] and constrained to evolve and adapt to new paradigms [21–23]. Therefore, legal documents reflect changes in culture and social norms [24]. Here, we investigate whether major social events leave measurable footprints in judicial records. We show that, indeed, the evolution of content in a large corpus of tens of thousands of judicial decisions from Spanish courts reveals the emergence and evolution of a socially disruptive issue. In particular, we focus on decisions related to housing in the context of the global financial crisis. Since 2007 and in less than a decade, more than 700,000 home-related foreclosure procedures were started (including those that do not result in a court procedure, as well as those that do, in all instances), which had a devastating effect in a significant fraction of the population in urban areas [25]. We use an information-theoretic methodology to quantify the footprint that such a major social issue left in the judiciary, by tracking the main trends and shifts in the content of decisions, both in terms of the full text of the decisions and of their citations to existing legislation. Specifically, our analysis shows an abrupt change in the content of housing-related decisions culminating in 2016, which is in stark contrast to the smooth evolution of two control corpora related to issues that did not produce social unrest during the same period. Moreover, because the approach we use pinpoints the specific content that drives change, we are able to interpret the results in terms of the role played by legislative changes, landmark decisions, and the influence of social movements.

2 Data and methods

2.1 Judicial decisions data set

We examined three different corpora of judicial decisions corresponding to different areas of law. In the first one, housing case law (H), decisions are related to housing issues such as evictions, foreclosing procedures, and squatting. In the other two corpora, used as a control group, decisions are related to homicides (HO) and condominium (C), respectively. In all three cases, decisions were mainly decided by courts of appeal (Provincial Court, 89%), but some of them were decided by other higher courts (e.g. the Supreme Court at the national level, 8% or regional Superior Courts, less that 1%). We do not analyze decisions ruled by courts of first instance, since those are not digitally published in the national database. We restricted the analysis to the period 2001 to 2018, resulting in a selection of 22,983 decisions in housing, 15,648 in homicides and 59,516 in condominium (for more details see Supplementary Text S1 in Additional file 1 and Additional files 2-4).

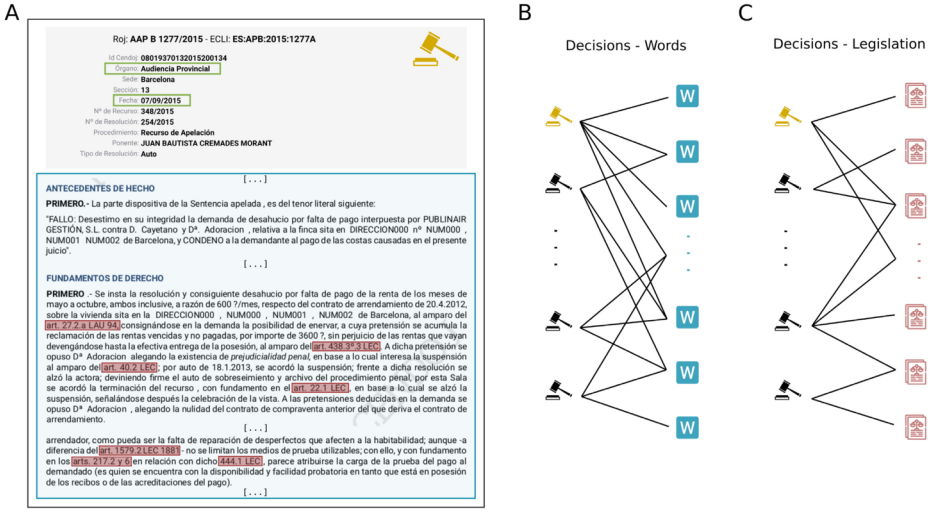

Figure 1 Representing judicial decisions as bipartite networks. (**A**) For each decision in the corpus we consider the words used (blue), the cited legislation (red), and metadata such as the date of the ruling and the court (green). We then create two bipartite networks: (**B**) A network of decisions and words; (**C**) A network of decisions and cited legislation. We select relevant words through a process of n-gram extraction and removal of words with low informational content, whereas we obtain cited legislation directly from the data (Section 2)

The data were provided to us by Tirant Online, one of the largest and most comprehensive databases for judicial decisions in Spain. The corpus provides the full text and a wide range of metadata extracted by Tirant Online for each judicial decision, from which we consider the cited legislation and the date of the ruling (see Fig. 1).

Given these data, we encode each decision as two unsorted lists, one of words and one of law articles. For words, we removed numbers and non-word characters, grouping and substituting statistically significant n-grams ($n = 2, 3$) (that is, groups of n words that appear together very often), and filtering stop-words using an information theoretic approach [26] that allows us to reduce considerably the length of the word list, while still keeping the informative terms (see Supplementary Text S2 in Additional file 1 and Additional files 5 and 6). In the case of legislation, we took the full list of cited law articles directly from the metadata.

2.2 Network-based topic models to characterize the content of judicial decisions

We quantify the content of decisions by modeling the topics for each corpus separately. Topic modeling is an approach used to classify large textual corpora and quantify content differences between documents, among other applications, by breaking down the content of each document into latent topics, which are groups of words used similarly in documents [27]. Because of the limitations of some of the most commonly used topic model approaches [28], we use the network-based method proposed by Gerlach et al. [29], which infers a hierarchical stochastic block model (SBM) from the bipartite network formed by documents and their words [29] (Fig. 1B and Fig. 2). In probabilistic terms, the inferred SBM is the most plausible one given the data; in information-theoretical terms, it has the shortest description length [30], that is, it is the model that best compresses the observed data. Unlike other approaches, the number of topics and levels in the hierarchy is inferred from the data rather than chosen manually.

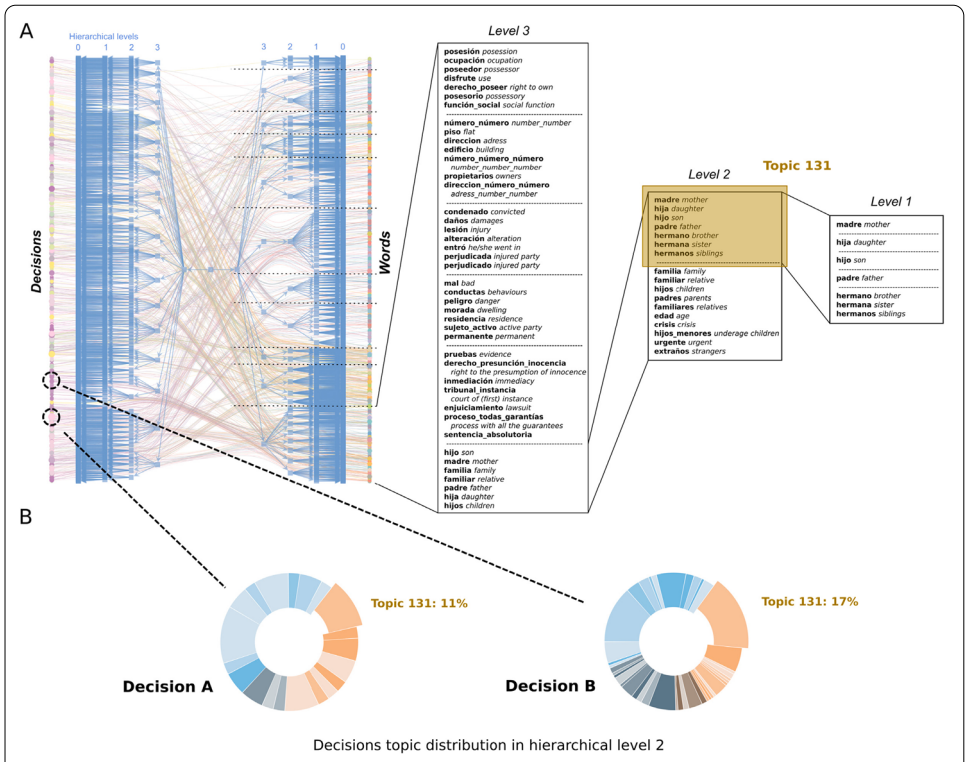

Figure 2 Hierarchical topic models and decisions as distributions over topics. (**A**) Using a network-based topic modeling approach that uses hierarchical stochastic block models [29], we infer hierarchical partitions in the bipartite network of judicial decisions and words. We show word topics in the housing decisions corpus. The hierarchical structure of the model is illustrated by expanding a particular topic: from left to right, we take a topic at the highest level of the hierarchy (level 3) and expand specific sub-topics at successively lower levels. (At level 3, only the top words in each subtopic are shown.) Topics contain words that are used in similar contexts, and this similarity increases as we go down in the hierarchy. (**B**) Given the membership of each word in a given topic, decisions can be characterized as distributions over topics by measuring the number of times each word appears in a document (Section 2). For example, topic 131 accounts for 11% of the words in Decision A, and for 17% of the words in Decision B. We follow the same procedure for the bipartite network of documents and legislation (Fig. 1 and Fig. S12 in Additional file 1). Likewise, we infer the word topics and the legislation topics for the other two corpora: homicides and condominium. Tables S1-S4 in Additional file 1 in provide the number of topics at each hierarchical level for each topic model

Similarly, by building a bipartite network of documents and cited legislation, we also obtain hierarchically-nested legislation topics (Fig. 1C and Fig. S12 in Additional file 1). Thus, while typically topic models are used to coarse-grain the words in texts, we extend this practice by concurrently modeling the use of legislation, obtaining two different topic models for each corpus.

Because topics are organized hierarchically, documents can be modeled with different degrees of coarse graining. On the one hand, lower levels in the hierarchy tend to describe very specific concepts (in the case of word topics) or to group together law articles of the same law (in the case of legislation topics, more than expected by chance, Supplementary Text S4 and Fig. S2 in Additional file 1). On the other, we find very general topics at higher levels, containing words that are only vaguely related or law articles of related laws. For details regarding the number of topics provided by the models at each level, see Supplementary Text S3 and Additional files 7-12. For simplicity, all explanations in the rest of

this section will refer to word topic models of a specific corpus, but they are equivalent for legislation topic models.

The model gives the membership of each word to a given topic T_i^ℓ at a given level ℓ in the hierarchy. Then, by counting the number of times each topic appears in a decision, we obtain the distribution over topics of each decision. Analogously, we obtain the yearly distribution of topics by considering all the words in all the decisions ruled in a given year—given a non-informative prior for the parameters of a topic distribution (a uniform Dirichlet distribution), the posterior yearly distribution of topics is

$$P\left(T_i^\ell|y\right) = \frac{n_y(T_i^\ell) + 1}{N_y + K^\ell}, \tag{1}$$

where $n_y(T_i^\ell)$ is the number of times that a word belonging to topic T_i^ℓ appears in the decisions ruled in year y, $N_y = \sum_i^{K^\ell} n_y(T_i^\ell)$, and K^ℓ is the number of topics in a given hierarchical level ℓ.

2.3 Time evolution of topics and Kullback-Leibler surprise

We analyze the time evolution of each topic T_i^ℓ by calculating its importance in a given year relative to its maximum importance across years

$$E_{T_i^\ell}(y) = \frac{1}{\max_y\{P(T_i^\ell|y)\}} P\left(T_i^\ell|y\right). \tag{2}$$

To quantify the extent to which the content of decisions changes, we compute the Kullback-Leibler (KL) surprise $\mathcal{S}_{-\tau}^\ell(y)$ between the yearly topic distribution (Eq. (1)) at year y and that at another year in the past $y - \tau$:

$$\mathcal{S}_{-\tau}^\ell(y) := D_{\mathrm{KL}}\left(P\left(\boldsymbol{T}^\ell|y\right)|P\left(\boldsymbol{T}^\ell|y-\tau\right)\right) = \sum_i \mathcal{S}_{-\tau}^\ell\left(y; T_i^\ell\right), \tag{3}$$

$$\mathcal{S}_{-\tau}^\ell\left(y; T_i^\ell\right) := P\left(T_i^\ell|y\right) \log \frac{P(T_i^\ell|y)}{P(T_i^\ell|y-\tau)}. \tag{4}$$

Here, $D_{\mathrm{KL}}(\boldsymbol{p}|\boldsymbol{q}) = \sum_i p_i \log \frac{p_i}{q_i}$ is the KL divergence [31, 32]. This metric is the relative entropy of a distribution $\boldsymbol{p}$ when expecting another distribution $\boldsymbol{q}$, and measures the average log-likelihood of observations being distributed as $\boldsymbol{p}$ when actually they were generated by $\boldsymbol{q}$ [32].

The KL divergence has been shown to describe cognitive surprise [33] and has been used to measure dissimilarity between speeches or texts using topic models [22, 34–36], word frequency models [37] or other low dimensional representations of textual content [38]. Here, we adapt some of these ideas to measure the extent to which the content of judicial decisions in one year differs from those in previous years (Eq. (3)).

To account for local temporal changes, we calculate the KL divergence between a topic distribution and the same topic distribution the year before ($\tau = 1$ in Eq. (3)). To analyze long-term patterns we compare to the distribution at all previous years in the past.

We evaluate the robustness of our results by comparing them with alternative ones corresponding to higher description length models. This analysis shows that there is very little qualitative variation of our results (see Supplementary Text S5 in Additional file 1).

We also conducted tests of human-annotated coherence of the topics (see Supplementary Text S5 in Additional file 1) [39].

2.4 Sampling factor

When computing information-theoretic measures from discrete distributions that have been learned from data, important biases appear when the size of the sample is far from the limit $N \gg K$, where N is the number of counts (words or law articles in our case) and K the number of bins in the discrete distribution (topics in our case). Although several approaches have been proposed to avoid such biases in entropy measures and other related quantities [40, 41], to our knowledge none have been proposed to address the bias when measuring the KL divergence. However, the bias depends on the so-called sampling factor N/K; therefore, we control the sampling factor to make legitimate comparisons between different KL measures.

In particular, when computing a given surprise $S_{-\tau}(y)$, we use the same sampling factor for all the years and for each corpora, so that

$$\frac{N_H}{K_H} = \frac{N_{HO}}{K_{HO}} = \frac{N_C}{K_C}, \tag{5}$$

where N is the number of words/citations used to estimate the distribution and K the number of topics, for each of the three corpora. All measures reported in the manuscript are averages over sub-samples of the corpus obtained with a fixed sampling factor.

3 Results

We hypothesized that the major social unrest that followed the collapse of the housing market in Spain should have left measurable footprints in legal documents, and particularly judicial decisions. Therefore, we analyze the corpus of housing-related decisions looking for these measurable footprints in terms of changes both at the level of the corpus and at the level of the topics responsible for those global changes. To that end, we use a combination of network-inference and information-theoretic approaches (see Section 2). To fully calibrate the changes observed in the housing corpus, we compare them with those found in the other two corpora, homicides and condominium.

3.1 Word and legislation topics give a global view of the evolution of decision contents

By representing the yearly topic distribution for both word and legislation topics, we obtain a global view of the time evolution of the content of the corpora (see Fig. 3; Section 2, Eq. (1)). By analyzing yearly changes in the prominence of topics, we observe that for housing-related decisions some topics that were prominent in the first years were later replaced by others (Fig. 3**A**-**B**). For instance, a word topic associated to leases loses importance, whereas one associated to abusive mortgage clauses gains prominence.

In Fig. 3**C**-**H**, we show the overall evolution of the relative importance of topics, for the housing corpus and our two control corpora, and for both word and legislation topics. Qualitatively, the importance of word topics in the homicide and condominium corpora remains very stable throughout the whole period 2001-2018. By contrast, in the housing

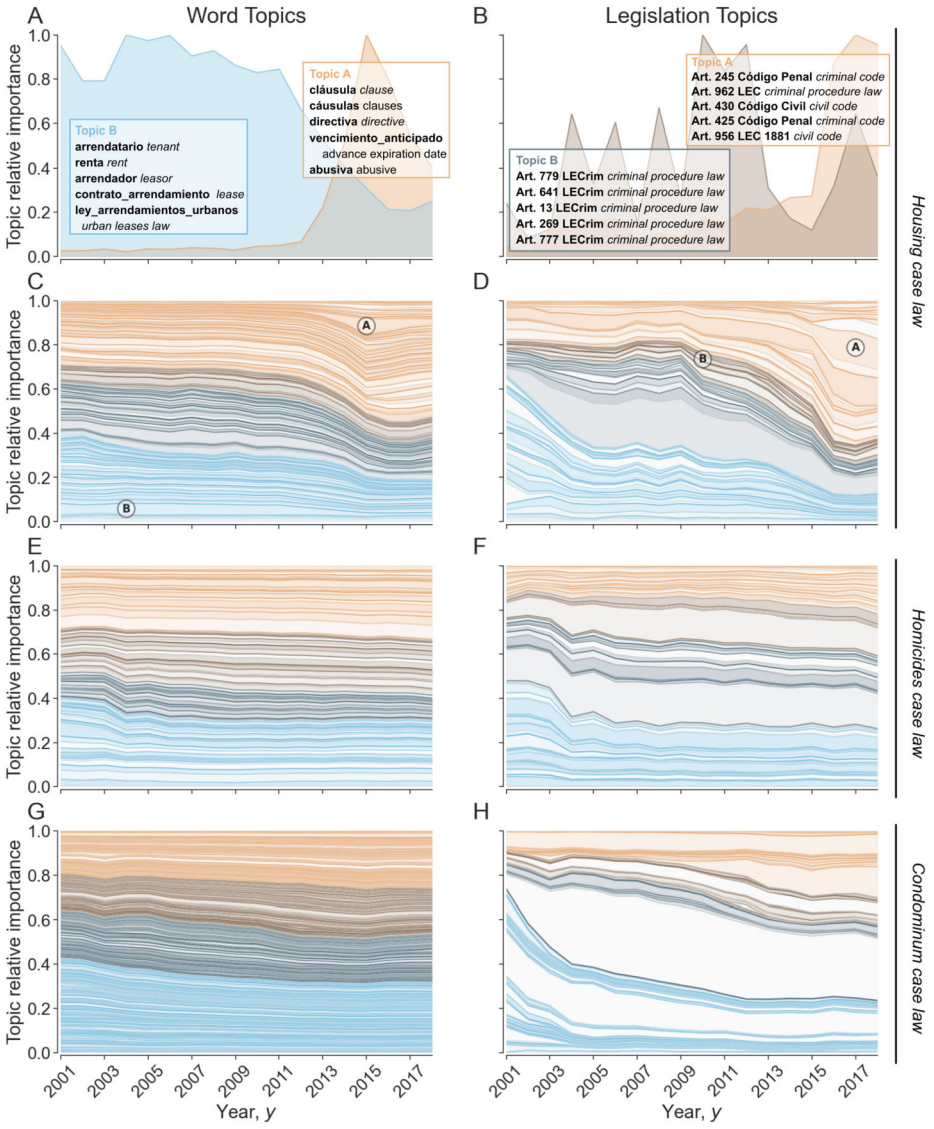

Figure 3 Time evolution of word topics (left column) and legislation topics (right column). For each year y, we compute the topic distribution $P(\boldsymbol{T}|y)$ (Section 2). (**A-B**) Time evolution of the importance of some illustrative topics in the housing corpus, normalized by their maximum importance across years (Section 2). While some topics decrease in the period 2012-2018, others show the opposite behavior. Topics depicted in blue reach their maximum importance in early years, whereas orange topics reach their maximum importance in the final years. Topics that reach their maximum importance in intermediate years are depicted in gray. (**C-H**) Evolution of all topics for the three corpora: (**C-D**) housing; (**E-F**) homicides, and (**G-H**) and condominium. Each layer corresponds to a topic and their thickness corresponds to their relative importance in a given year. Topics shown in panels (**A**) and (**B**) are identified in (**C**) and (**D**), respectively. Hierarchical levels: C, L = 2; D, L = 1; E, L = 2; F, L = 0; G, L = 1; H, L = 0; we show the analogous results for other hierarchical levels in the models in Figs. S3 and S4 in Additional file 1

corpus we observe a major, systemic shift in topic importance in the period 2013-2016: Some word topics that accounted for 30% of the words in the decisions in 2011 end up accounting for over 50% only five years later. Similarly, the evolution of legislation topics is similar for the three corpora in the period 2001-2012; then a major shift occurs in housing-related decisions in the years 2013-2016, which we do not observe in the homicide and

condominium corpora. Indeed, legislation topics that accounted for 20% of the cited law articles in housing related-decisions prior to 2010 end up accounting for over 60% in 2016. All these results are robust when using other hierarchical levels in the topic model (Fig. S3-S4 in Additional file 1). Therefore, we observe a genuine shift in the language and the legislation used by judges in judicial decisions related to housing in the period 2012-2016, which we do not observe in other areas of the law.

3.2 Bayesian surprise reveals disruptive periods and topics

As discussed above, we quantify the significance of the changes we observe in Fig. 3C-D by means of an information-theoretic measure of surprise, the Kullback-Leibler (KL) divergence [31] (see Section 2). When we compare topic distributions corresponding to pairs of consecutive years (Fig. 4), the KL surprise quantifies the changes that are qualitatively apparent in Fig. 3. Until 2012, the textual content and the legislation cited in judicial decisions changes, from one year to the next, at rates that are similar among the three corpora. After 2012, the surprise between consecutive years is considerably higher in housing case law than in the two control corpora. Remarkably, both word and legislation topics display a pronounced peak in 2016, which we observe for all hierarchical levels of topics (Figs.

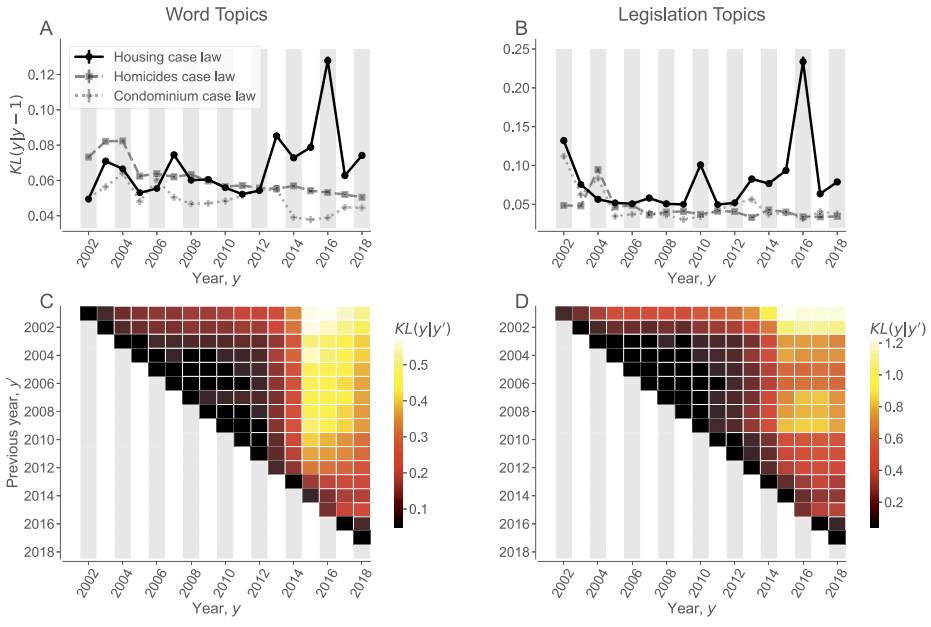

Figure 4 Kullback-Leibler (KL) divergence to measure surprise in the time evolution of topics. We use the KL divergence to measure the change in the topic distribution on year y with respect to the distribution on: (**A**-**B**) the year before; (**C**-**D**) a specific year y' in the past (for the housing corpus only). Although the number of decisions per year and the number of topics in the model is different for each corpus, we ensure the comparability of the results by using the same sampling factor (that is, the same ratio between the number of words used to estimate a topic distribution and the number of topics in the model; see Section 2). Error bars, often smaller than the symbols, correspond to the standard deviation of the mean over several sub-samples using a fixed sampling factor. We compute the KL divergence from the topic distributions at each hierarchical level within each topic model (see Fig. 2 and Section 2). Here, we show results for the following levels L and the corresponding number of topics K. (**A**, **C**) Word topics for housing: L = 1 and K = 564 word topics; homicides: L = 1 and K = 688 word topics; condominium: L = 1 and K = 525 word topics. (**B**, **D**) Legislation topics for housing: L = 2 and K = 30 legislation topics; homicides: L = 1 and K = 17 legislation topics; condominium: L = 1 and K = 27 legislation topics. Results obtained at other hierarchical levels are qualitatively very similar (Figs. S5-S7 in Additional file 1)

S5-S6 in Additional file 1). Note that measuring surprise at the lowest levels in the hierarchy reveals changes occurring in the most specific topics, while doing so at the highest level reveals changes in the most general ones. Therefore, a consistent peak throughout levels in the hierarchy implies that the changes occurred around this date stem from a deep reorganization of topics, rather than a shallow reorganization of sub-topics.

To further characterize the changes occurred in the topic landscape, we also calculate the KL surprise between the topic distribution in one year and all years in the past. Within the corpus of housing decisions, this analysis highlights the discontinuity between decisions ruled before and after 2012 and, especially, before 2010 and after 2014 (Fig. 4A-B, Fig. S7-S8 in Additional file 1). Additionally, this analysis reveals the scope of the shift occurred in 2016—distributions after this date show high KL divergence with respect to years earlier than 2012, but relatively low KL divergence with respect to 2016, which indicates that the shift occurred in 2016 persists in the following years.

3.3 Legal interpretation links disruptive topics to landmark decisions and law modifications

Finally, we investigate the contribution of individual topics to the KL surprise, which allows us to go beyond identifying periods of rapid change in judicial decisions, and to actually pinpoint the specific topics that most contributed to the disruption in the years around 2016. In particular, we focus on the topics that most contribute to the changes occurred in 2016: Word topics 108 and 14, and legislation topics 41 and 42 (Fig. 5A and 5B respectively, see also Fig. S9-S10 in Additional file 1). Each topic legal interpretation follows the words/legislation found in the topic, some of which we include in the following (in italics; for more details on the content of each topic, see Supplementary Text S3 in Additional file 1).

When interpreting the topics that disrupt the most, we choose the hierarchical level that balances internal coherence and level of detail of topics. Indeed, these two qualities behave in opposite ways when going from the lowest (finest) level of the topic hierarchy to the highest (broadest). For instance, if we consider word topic 14 in level 2 (see Fig. S13 in Additional file 1), we observe that some sub-topics present very specific coherence but very insufficient detail. For example, in the scope of housing, the terms *denunciado, denunciante* (plaintiff, respondent) only frame the legal analysis in the area of criminal law. It is only by considering sub-topics (that is, going up in the hierarchy) that we learn more details that narrow housing criminal law to the light criminal offense of squatting. Unfortunately, going too high in the hierarchy makes interpretation more difficult because words in a topic can become specific to a set of decisions, which can distort the interpretation.

Word topic 108 Words in this topic are typically related to mortgage loan contracts, a hotly debated subject in the years after the global financial crisis of 2007, when mortgage enforcement skyrocketed. The increase in the use of such terms around 2016 stems from the following events. In the well-known case Aziz 2013 [42], the European Court of Justice (ECJ) established a doctrine involving the Directive 93/13/EEC on unfair terms in consumer contracts [43]. The directive stated that, in order to protect consumers (*consumidores*, in Spanish), the validity of possible unfair terms (*cláusulas abusivas*) included in mortgage loan contracts (*contratos de préstamo hipotecario*) could be discussed during

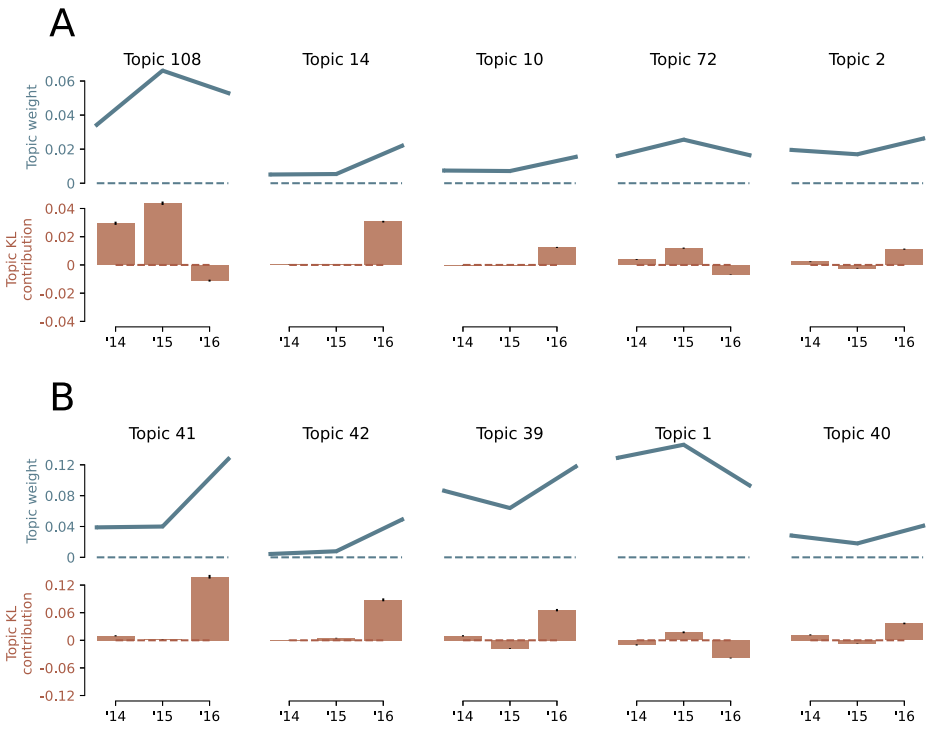

Figure 5 Topics in housing-related decisions that change abruptly and contribute to disruption in 2016. We show the weights of each topic (blue line) and their contribution to the Kullback-Leibler (KL) divergence (red bars) over the years preceding 2016, where there is an important disruption in the content of decisions (Fig. 4). We show the top-5 topics with the highest contribution to the KL divergence in the selected period for both word (**A**) and legislation topics (**B**). Topic weights correspond to topic distributions at hierarchical level 2 for word topics, and level 1 for legislation topics. While the weight shows the relative importance a topic has in a given year, the KL contribution shows the extent to which the importance of a topic is different from the year before. We show the evolution for all topics in the same hierarchical level in Figs. S9-10 in Additional file 1. The words and law articles associated to each of these topics can be found in Additional file 7 and 8, respectively

a mortgage enforcement procedure (which, by definition should be fast and efficient, as all its terms should already had been reviewed by a notary public when the mortgage was arranged). Since, until then, such a discussion was not formally allowed according to Spanish civil procedure law (LEC, articles 552, 557 and 561), that case forced a law reform in Spain through Act 1/2013, which was still unclear, insufficient and led to further problems and cases: since 2013, whether the amount of default interest rate (*intereses de demora*), the debt acceleration clause (*cláusulas de vencimiento anticipado*) or the unilateral liquidation of the debt by the creditor (all of them possible unfair terms) were in accordance with European law, was left to the courts to determine on a case-by-case basis during a mortgage enforcement. This process ran in parallel with the discussion about the validity of the floor clauses (*cláusulas suelo*) in mortgage loan contracts, which was not cleared up until the decision 9-5-2013 by the Spanish Supreme Court (and, at a European level, until the ECJ decision 21-12-2016).

The events of 2013 (the Aziz case with the subsequent Spanish law reform, Act 1/2013, and the doctrine of the Supreme Court on floor clauses) explain the increase in the number of decisions related to topic 108, with a peak in 2015 (see Fig. 5A) which is consistent with the 2/3-year delay expected in decisions ruled by courts of appeal with respect to the first instance case. Additionally, as a counterpart to the increase in the number of first instance

cases related to this topic (which 2/3 years later arrived to the courts of appeal), after 2013 the number of mortgage enforcements slowed down progressively (from 38,961 in 2013 to 24,555 in 2016) due to non-judicial agreements between mortgagees and mortgagors [44].

Word topic 14 Most of the words in this topic are related to a criminal offense. For instance, we find the words plaintiff and respondent (*denunciante* and *denunciado*), which appear in a criminal procedure to refer to the parties involved (instead of claimant and defendant that we would find in a civil procedure). Indeed, this topic refers to a specific type of criminal offense (minor criminal offence, *delito leve*, was introduced in Organic Act 1/2015 in substitution for misdemeanors, *faltas*; articles 13.3 and 4 and 33.4 Criminal Code): non violent (against people) squatting (article 245.2 of the Spanish Criminal Code). This is because this topic includes references to both what is protected through this minor criminal offense, which is possession of a property (disturbance of possession, *perturbación de la posesión*; possession endangerment, *riesgo de la posesión*); and to the role that criminal law should have in this kind of situations, namely, that criminal law should only intervene as a last resort (principle of minimal intervention, *principio de intervención mínima*; penal intervention, *intervención penal*) as far as most situations are, at least theoretically, protected through civil law using possession claims (article 250 Spanish civil procedure Law). However, those mechanisms were not effective in protecting owners, many of whom resorted to filing a criminal lawsuit, which most times was dismissed by judges according to the aforementioned rule of minimal intervention. This situation contributed to the enormous increase of squatters (see legal interpretation of legislation topics below) and to a reform of the Spanish civil procedure Law to facilitate the civil way, which did not arrive until 2018 (by Act 5/2018).

Legislation topic 41 This topic has 96% of the weight in article 245 of the Spanish criminal code, whose Sect. 2 includes the aforementioned minor criminal offense of squatting (*usurpación*) of a usually uninhabited dwelling (see word topic 14). The sharp increase in the use of this article in housing-related decisions in 2016 is also illustrative. During the first 5 years since the start of the global financial crisis of 2007 there were no legal dispositions to stop the crisis or to palliate its consequences, while the ones since 2012 had been very feeble and non-structural [45]. This fact, coupled with social movements that supported squatting as a solution to mitigate the housing problem, lead to an increase in both criminal and civil (forced dispossessions) squatting cases. According to the Spanish General Council of the Judiciary [46], convictions for squatting-related crimes went from 420 in 2007 to more than 6000 annually between 2016 and 2018; between 2008 and 2018 convictions for squatting grew more than ten-fold.

Legislation topic 42 The most important law article in topic 42 is *Article 82 of the Spanish Organic Law of the Judicial Power (LOPJ)*, which deals with the functions of the courts of appeal in the field of criminal law. In 2015, it was modified by Organic Law 13/2015 and the word misdemeanors (*faltas*) was substituted by minor criminal offenses (*delitos leves*), in the field of criminal offenses related to housing. The second most import law article in the topic is *Article 876 of the Criminal Procedural Law*, which deals with the process of notification of the judicial decision. This Article was also modified to substitute the same words as in *Article 82*. Since these two articles account for 80% of the weight in the

topic (see Supplementary Text S3 in Additional file 1), we can say that the topic is related principally to the mentioned modification of different criminal law norms.

The aforementioned law articles address very general aspects in the scope of criminal procedural law, which hinders the task of interpreting the raise in the number of judicial decisions citing them that we observe. However, knowing that they have been modified in relation to words that appear in word topic 14, and knowing that the importance of both topics evolved in parallel (see Fig. 5B, 5G), we can say that the observed raise in the use of these articles is linked to these modifications. All in all, this is an example of the interplay between word topics and legislation topics.

4 Discussion

For centuries, humans have left traces of their stories, activities and values in books, newspapers, and transcribed speeches, among others. Legal documents, such as codes, laws and, especially, judicial decisions, are particularly useful because they reflect changes in society and the evolution of the most socially sensitive issues and debates, which are the ones that end up in court. Our study presents and validates a methodology to exploit the information contained in digitized judicial decisions and to detect, quantify, and explain social disruptions from them.

One particularity of our approach is that it can be used to analyze, simultaneously and coherently, the textual content of decisions and the use of existing legislation by judges. While words shape the discourse and the arguments, citation to existing legislation summarizes the mechanisms by which judges fit their ideas into the applicable framework at each point in time; thus the importance of considering both elements. Moreover, while analyzing word topics only requires linking words to legal concepts, analyzing legislation topics is more challenging because law articles can be extensive and address very general aspects. Then, using words in word topics can leverage the interpretation of law articles present in legislation topics and vice-versa, as shown in our legal analysis. Remarkably, our results show that social disruption leads to abrupt changes at both levels simultaneously, that is, judges change their discourse at the same time that they change the legal mechanisms by which they justify their decisions.

Using the same topic modeling approach to the cited legislation has the additional advantage of facilitating explanation. Indeed, law articles, like words, are semantically related with each other at different levels. In order to produce interpretable results it is crucial to detect these relationships and reduce the dimensionality of the legislation space; that is, to go from thousands of articles to hundreds or even tens of legislation topics, as we have defined them. In other contexts, where legal precedents are more important than in the Spanish system (for example, in common law countries such as the United States), it may be appropriate and useful to extend the topic modeling approach to study precedent. More broadly, using topic models for content elements other than words could also be useful in different digitized documents such as academic papers, where one could use it to coarse-grain the list of references by clustering them into topics.

Going back to the use of judicial decisions to analyze social change, we have shown that our approach provides an accurate and robust description of the shift in the content of decisions in the case of housing-related decisions in Spain. Our approach reveals that this shift is not a shallow reorganization of subtopics, but rather a deep change affecting all levels of the topic hierarchy for both words and cited law. Finally, we have shown that

the topics responsible for the sharp changes we observe in the content of housing-related decisions in 2016 can be unambiguously interpreted in terms of the legal and social context at that time.

Legal documents are, together with religious documents, among the oldest written information sources that have been preserved through history. Our work shows that it is possible to extract and interpret information from such documents, and to reveal disruptive societal events. Thus, although we analyzed very recent documents, we hope that our methodology will be useful to precisely localize and interpret historical events, even in the distant past.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-022-00376-0.

Additional file 1. Supplementary information (PDF 6.3 MB)
Additional file 2. Supplementary information (CSV 628 kB)
Additional file 3. Supplementary information (CSV 403 kB)
Additional file 4. Supplementary information (CSV 1.6 MB)
Additional file 5. Supplementary information (TXT 5 kB)
Additional file 6. Supplementary information (CSV 1 kB)
Additional file 7. Supplementary information (CSV 590 kB)
Additional file 8. Supplementary information (CSV 1.2 MB)
Additional file 9. Supplementary information (CSV 398 kB)
Additional file 10. Supplementary information (CSV 525 kB)
Additional file 11. Supplementary information (CSV 335 kB)
Additional file 12. Supplementary information (CSV 619 kB)
Additional file 13. Supplementary information (CSV 1 kB)

Acknowledgements
We thank Tirant Online for providing us with the digitized and parsed corpora of legal decisions, and for their help with data processing.

Funding
This research was funded by the Social Observatory of the "la Caixa" Foundation as part of the project LCF/PR/SR19/52540009, by MCIN/AEI/10.13039/501100011033 (Project No. PID2019–106811GB-C31) and by the Government of Catalonia (Project No. 2017SGR-896).

Abbreviations
H, Housing case law; HO, Homicides case law; C, Condominium case law; SBM, Stochastic block model; KL, Kullback Leibler; ECJ, European Court of Justice; LEC, *Ley de Enjuiciamiento Civil* (Spanish Civil procedure law); LOPJ, *Ley Orgánica del Poder Judicial* (Spanish Organic Law of the Judicial Power).

Availability of data and materials
We provide the full list of IDs corresponding to all the judicial decisions studied in this work, see Supplementary Text S1. Using these IDs, the information from each decision can be retrieved downloading the opinions from the following public database https://www.poderjudicial.es/search/indexAN.jsp.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contribution
MS-P and RG designed the research. LF-P, MS-P and RG carried out computational experiments. All authors discussed and analysed the results and wrote and edited the manuscript. All authors read and approved the final manuscript.

Author details
[1]Dept. Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia. [2]Dept. Private Law, Universitat Rovira i Virgili, 43003 Tarragona, Catalonia. [3]ICREA, 08017 Barcelona, Catalonia.

Publisher's Note

Received: 1 July 2022 Accepted: 20 December 2022 Published online: 03 February 2023

References

1. Watts D (2007) A twenty-first century science. Nature 445:489
2. Evans JA, Aceves P (2016) Machine translation: mining text for social theory. Annu Rev Sociol 42(1):21–50. https://doi.org/10.1146/annurev-soc-081715-074206
3. García-Gavilanes R, Mollgaard A, Tsvetkova M, Yasseri T (2017) The memory remains: understanding collective memory in the digital age. Sci Adv 3(4):1602368. https://doi.org/10.1126/sciadv.1602368
4. Pah AR, Schwartz DL, Sanga S, Clopton ZD, DiCola P, Mersey RD, Alexander CS, Hammond KJ, Amaral LAN (2020) How to build a more open justice system. Science 369(6500):134–136. https://doi.org/10.1126/science.aba6914
5. Quemy A, Wrembel R (2020) On integrating and classifying legal text documents. In: Hartmann S, Küng J, Kotsis G, Tjoa AM, Khalil I (eds) Database and expert systems applications. Springer, Cham, pp 385–399
6. Hutchinson T, Duncan N (2012) Defining and describing what we do: doctrinal legal research. Deakin Law Review 17:83–119. https://doi.org/10.21153/dlr2012vol17no1art70
7. Panagis Y, Christensen M, Šadl U (2016) On top of topics: leveraging topic modeling to study the dynamic case-law of international courts of law, icourts centre of excellence for international courts. Leg Knowl Inf Syst 294:161–166. https://doi.org/10.3233/978-1-61499-726-9-161
8. Baude W, Chilton AS, Malani A (2017) Making doctrinal work more rigorous: lessons from systematic reviews. Univ Chic Law Rev 84:37–58
9. Hall MA, Wright RF (2008) Systematic content analysis of judicial opinions. Calif Law Rev 96(1):63–122
10. van Gestel R, Micklitz H-W (2014) Why methods matter in European legal scholarship. Eur Law J 20(3):292–316. https://doi.org/10.1111/eulj.12049
11. Šadl U, Olsen HP (2017) Can quantitative methods complement doctrinal legal studies? Using citation network and corpus linguistic analysis to understand international courts. Leiden J Int Law 30(2):327–349. https://doi.org/10.1017/S0922156517000085
12. Medvedeva M, Vols M, Wieling M (2020) Using machine learning to predict decisions of the European court of human rights. Artif Intell Law 28:237–266
13. Mones E, Sapiezynski P, Thordal S, Olsen H, Lehmann S (2021) Emergence of network effects and predictability in the judicial system. Sci Rep 11. https://doi.org/10.1038/s41598-021-82430-x
14. Lupu Y, Voeten E (2012) Precedent in international courts: a network analysis of case citations by the European court of human rights. Br J Polit Sci 42(2):413–439. https://doi.org/10.1017/S0007123411000433
15. Olsen HP, Küçüksu A (2017) Finding hidden patterns in Ecthr's case law: on how citation network analysis can improve our knowledge of Ecthr's article 14 practice. Int J Discrim Law 17(1):4–22. https://doi.org/10.1177/1358229117693715
16. Fowler JH, Johnson TR, Spriggs JF, Jeon S, Wahlbeck PJ (2007) Network analysis and the law: measuring the legal importance of precedents at the U.S. Supreme Court. Polit Anal 15(3):324–346. https://doi.org/10.1093/pan/mpm011
17. Charlotin D (2020) "authorities" in international dispute settlement: a data analysis. (doctoral thesis). PhD thesis, University of Cambridge
18. Guimerà R, Sales-Pardo M (2011) Justice blocks and predictability of U.S. Supreme Court votes. PLoS ONE 6(11):1–8. https://doi.org/10.1371/journal.pone.0027188
19. Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proc Natl Acad Sci 108(17):6889–6892. https://doi.org/10.1073/pnas.1018033108. https://www.pnas.org/content/108/17/6889.full.pdf
20. Sheshadri K, Singh MP (2019) The public and legislative impact of hyperconcentrated topic news. Sci Adv 5(8):8296. https://doi.org/10.1126/sciadv.aat8296
21. Katz D, Coupette C, Beckedorf J, Hartung D (2020) Complex societies and the growth of the law. Sci Rep 10:18737
22. Rockmore DN, Fang C, Foti NJ, Ginsburg T, Krakauer DC (2018) The cultural evolution of national constitutions. J Assoc Inf Sci Technol 69(3):483–494. https://doi.org/10.1002/asi.23971
23. Rutherford A, Lupu Y, Cebrián M, Rahwan I, LeVeck BL, García-Herranz M (2018) Inferring mechanisms for global constitutional progress. Nat Hum Behav 2:592–599
24. Klingenstein S, Hitchcock T, DeDeo S (2014) The civilizing process in London's old Bailey. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1405984111. https://www.pnas.org/content/early/2014/06/11/1405984111.full.pdf
25. Nasarre-Aznar S, Garcia-Teruel RM (2018) Evictions and homelessness in Spain 2010–2017. In: Kenna P, Nasarre-Aznar S, Sparkes P, Schmid CU (eds) Loss of homes and evictions across Europe. Edward Elgar Publishing, USA, pp 292–332
26. Gerlach M, Shi H, Amaral LAN (2019) A universal information theoretic approach to the identification of stopwords. Nat Mach Intell 1(12):606–612. https://doi.org/10.1038/s42256-019-0112-6
27. Blei D, Carin L, Dunson D (2010) Probabilistic topic models. IEEE Signal Process Mag 27(6):55–65. https://doi.org/10.1109/MSP.2010.938079
28. Lancichinetti A, Sirer MI, Wang JX, Acuna D, Körding K, Amaral LAN (2015) High-reproducibility and high-accuracy method for automated topic classification. Phys Rev X 5:011007. https://doi.org/10.1103/PhysRevX.5.011007
29. Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. Science Advances 4(7). https://doi.org/10.1126/sciadv.aaq1360. https://advances.sciencemag.org/content/4/7/eaaq1360.full.pdf
30. Rissanen J (1978) Modelling by shortest data description. Automatica 14:465–471
31. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
32. Shlens J (2014) Notes on Kullback-Leibler divergence and likelihood. CoRR. arXiv:1404.2000
33. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. Vis Res 49(10):1295–1306. https://doi.org/10.1016/j.visres.2008.09.007. Visual Attention: Psychophysics, electrophysiology and neuroimaging
34. Barron ATJ, Huang J, Spang RL, DeDeo S (2018) Individuals, institutions, and innovation in the debates of the French revolution. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1717729115. https://www.pnas.org/content/early/2018/04/16/1717729115.full.pdf
35. Murdock J, Allen C, DeDeo S (2017) Exploration and exploitation of Victorian science in Darwin's reading notebooks. Cognition 159:117–126. https://doi.org/10.1016/j.cognition.2016.11.012
36. Andrei V, Arandjelović O (2016) Complex temporal topic evolution modelling using the Kullback-Leibler divergence and the bhattacharyya distance. EURASIP J Bioinform Syst Biol 16. https://doi.org/10.1186/s13637-016-0050-0

37. Savoy J (2013) Authorship attribution based on a probabilistic topic model. Inf Process Manag 49(1):341–354. https://doi.org/10.1016/j.ipm.2012.06.003
38. Hughes JM, Foti NJ, Krakauer DC, Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. Proc Natl Acad Sci 109(20):7682–7686. https://doi.org/10.1073/pnas.1115407109. https://www.pnas.org/content/109/20/7682.full.pdf
39. Hoyle A, Goel P, Hian-Cheong A, Peskov D, Boyd-Graber JL, Resnik P (2021) Is automated topic model evaluation broken? The incoherence of coherence. In: Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds) Advances in neural information processing systems
40. Nemenman I, Shafee F, Bialek W (2002) In: Entropy and inference, revisited. Dietterich T, Becker S, Ghahramani Z (eds) Advances in neural information processing systems, vol 14. MIT Press, Cambridge. https://proceedings.neurips.cc/paper/2001/file/d46e1fcf4c07ce4a69ee07e4134bcef1-Paper.pdf
41. DeDeo S, Hawkins RXD, Klingenstein S, Hitchcock T (2013) Bootstrap methods for the empirical study of decision-making and information flows in social systems. Entropy 15(6):2246–2276. https://doi.org/10.3390/e15062246
42. ECJ: Aziz judgement, ECLI:EU:C:2013:164. https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=ECLI%3AECLI%3AEU%3AC%3A2013%3A164_1. Last accessed February 2022
43. ECJ: council directive 93/13/EEC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31993L0013. Last accessed February 2022
44. Garcia-Teruel RM, Nasarre-Aznar S (2022) Quince años sin solución para la vivienda. la innovación legal y la ciencia de datos en política de vivienda. Revista Crítica de Derecho Inmobiliario
45. Nasarre-Aznar S (2020) Los Años de la Crisis de la Vivienda. De las Hipotecas Subprime a la Vivienda Colaborativa. Tirant lo Blanch Valencia
46. del Poder Judicial CG. Criminal, civil and labor Data. https://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/Estadistica-por-temas/Datos-penales–civiles-y-laborales/Delitos-y-condenas/Condenados–explotacion-estadistica-del-Registro-Central-de-Penados-/. Last accessed February 2022