Acta Univ. Sapientiae, Legal Studies, **10**, 2 (2021) 287–302

DOI: 10.47745/AUSLEG.2021.10.2.09

Big-Data-Based Legal Analytics Programs. What Will Data-Driven Law Look Like?

Zsolt ZŐDI
PhD, Senior Research Fellow
Institute of the Information Society
University of Public Service, Budapest (Hungary)
e-mail: zodi.zsolt@uni-nke.hu

Abstract. Big-data-based legal analytics programs (LAP) appeared in the US in the early 2010s. They work by collecting large amounts of textual data from public databases, usually on websites, refining this data, linking it to other data, and then analysing and presenting it with special software. In this study, I first present the history of LAPs, their different types, key features, and their content and technology fundamentals. In a highlighted example, I also describe their uses through the 'Judge Analyzer' module. I will write later in this section about the upheaval that a judge analyser service has caused in France and the arguments and counterarguments that have been voiced in this debate. In the second part, the study describes the legal uses of LAPs and the related concerns. The study divides the concerns into two groups. The first type of general concern points to the possibility of a limited use of 'data-based law' and 'prediction' in law. The second type of counterargument focuses on the difference between common law and continental legal systems. Finally, the study briefly describes the future of LAPs.

Keywords: big data, data-based legal analytics programs, judge analyser, prediction of case law, France

1. What Are Big-Data-Based Legal Analytics Programs?

The term 'big data' spread in the early 2010s.[1] It marks the process by which more and more data, mainly from the open Internet and the Internet of things, makes it possible to develop new types of analysis, to identify contexts, to draw conclusions, and to make predictions in different areas of life.

The big data phenomenon can best be seen as an approach and narrative that seeks to apply long-established statistical methods to new areas where insufficient

[1] Schönberger–Cukier 2013.

data has been available so far, such as to predict customer behaviour in business or the occurrence of certain social phenomena. Big data has also revolutionized the use of artificial intelligence, perhaps its most significant effect. This is because the large amount of data collected in different areas of life also functions as teaching data for applications based on machine learning. In the second half of the 2010s, this brought real breakthroughs in many areas such as machine translation or medical diagnostics based on imaging procedures.

The big data approach soon appeared in the world of law.[2] Many have praised the potential it offers to help legislation (e.g. through 'experimentation' with real-time data and immediate feedback) and jurisprudence (e.g. text mining analysis of court judgments), whether in court, in law offices, or at authorities.[3] The following article analyses a special type of software (mainly operating in the form of an online service) that has been developed mainly to help the work of lawyers in law offices and legal departments. These legal analytics programs (LAPs) can retrieve, display, analyse, and visualize all kind of legal data, extracted from legal documents and databases.

1.1. The History of Legal Analytics Software

In the United States, pilot projects were launched in the mid-2000s to provide a new type of tool for lawyers through innovative aggregation and processing of available open data generated at courts, the patent office, and at other authorities.

The first such analytics program was *Lex Machina*, whose immediate predecessor was developed in collaboration with Stanford Law School and the Stanford Computer Science Department back in 2006.[4] The project was backed by another initiative, the IP (Intellectual Property) Litigation Clearinghouse, which sought to set up a search system for disputes over intellectual property (primarily patents). Intellectual property law is a particularly good area for experimentation because there is a large capital movement around the field, litigation is usually very costly, and cases involve a relatively narrow subject area (domain), so the range of arguments used in them can be well delineated. (It is recalled here that one of the first successful legal expert systems, HYPO, operated in the area of trade secrets, which has similar characteristics.)[5] Lex Machina continued to gather investors and customers until 2015, when it was acquired by LexisNexis.[6]

At the same time, other service providers appeared. Currently, according to the 360Quadrants website,[7] 18 companies offer such services, although their

[2] Ződi 2017; Rapoport–Tiano 2019.

[3] Katz 2013.

[4] Hoge 2013.

[5] Ashley 1991.

[6] *Lex Machina*. 2015.

[7] *360Quadrants: Best Legal Analytics Software*. 2021.

functionality is diverse. In addition to ‘general’ analysis, there are LAPs that help the process of e-discovery, or legal due diligence. The functionality of these programs is very similar (search, aggregation/visualization, prediction), but instead of processing *public legal documents,* they analyse the client’s own documents, though sometimes the data extracted from these are compared with the data in public databases.

LAPs can thus be clearly linked to the massive amount of data that appears on the open Internet and to the Anglo-Saxon legal culture, where litigation (and every legal process, even a pre-contractual negotiation) is perceived as a cognitive, argumentative battle in which the party with more information and insight wins.

1.2. Technology and Functions

LAPs use legal documents as raw data. Documents, as I mentioned above, can be public documents from public databases or the user’s own documents. Lex Machina, for example, is currently using four data sources: PACER (Public Access to Court Electronic Records), EDIS (Electronic Document Information System of the International Trade Commission), US PTO (Patent Office – database), and dockets of some county courts.[8] The former are collected by most service providers by web scraping technologies[9] (bulk downloads), but direct data transfer by individual data providers is also not excluded. Several techniques are then used in the texts to extract data and information from them. The most frequently used of these information extraction technologies is the so-called RegEx (Regular Expression) and Named Entity Recognition.[10] The former is essentially data recognition of certain character patterns (strings) described by rules. The system finds the strings defined in the rules (e.g. legal references in a legal document) and then indexes them. NER can be understood as a subset of the RegEx method, where, from unstructured or partially structured texts, *names* (in a broad sense), i.e. ‘rigid markers’ (personal names, institution names, place names, etc.), and numerical data (date, amount of money, or other entities) are captured and stored. The ultimate goal of both methods is to extract structured information from partly or entirely unstructured corpuses of texts. The services and functions of LAPs are then based on this data.

The functions of LAPs can be divided into three groups: search, aggregation/visualization, and predictive functions.

Firstly, all systems have ‘traditional’ search features. Due to the rich set of metadata (the wide variety of extracted data), these systems usually allow much more accurate, complex searches than systems that search simply in the full text or

[8] *Lex Machina – How It Works?* 2021.

[9] Mitchell 2018.

[10] Zhang et al. 2018, Csányi et al. 2021.

based on a few simple parameters. Another feature is that, unlike general Internet search engines, they allow a wide range of pre-sets for various metadata/parameters (so-called parametric search) when searching. In general web search engines, there is usually only one search box in the main interface, and the search engine tries to 'figure out' what the searcher's intention is, because it can be very different for the same search string. In this respect, the handbook for Google search testers is interesting and instructive because it takes into account the user's personal information and search history. For example, if a user enters the word 'apple', the search engine takes Apple Inc. as the 'dominant interpretation' and treats the 'fruit' result only as a 'common' search intent. However, professional search engines work differently because they only consider the search string and cannot measure the user intent. Different metadata are meant to 'express' user intent. In this sense, LAPs' search engines are so-called 'vertical or topical search engines'.

Another function of the systems is the aggregation/statistics function. This provides much more than the result list of traditional search engines because it does not return information items (e.g. legislation or court judgments, i.e. documents) as a result of a search but data items extracted from the records, and it can display them in a predefined format. The most spectacular of these aggregation functions, and, of course, the most controversial one, is the one that summarizes the data in a system related to a judge, which I will discuss in detail below. But aggregation can, of course, be made for any other data. For example, in the case of a database containing court decisions, the data of a law firm, a court, or a client (company) can be aggregated and presented. Thus, the success rate of the law firm, the list of lawyers acting on behalf of the law firm, the list of clients of the law firm, the 'relationship' between the firm and a judge, and so on, can be displayed. Similarly, it is possible to put a particular court at the centre of the inquiry: for example, the frequency of a particular type of case, the average length of cases, and the length of each stage of the proceedings (pre-litigation, first instance, second instance, etc.) can be also scrutinized.

In the same way, data collection and display can be performed for a group of documents, e.g. a specific narrower area of law; for example, for matters relating to tenancies, this group might include: who acts in these cases, how long the procedure lasts, with what result the procedure typically ends (whether the tenant or the landlord typically wins), and much more. In essence, all the data that can be extracted can be shown together with all the other data. This in itself is not very innovative, as, for example, the essence of relational databases is that any data can be queried in conjunction with any other data – what is innovative about these systems is that they extract and organize data that was previously only available in an unstructured form, so that it can be queried. Of course, not only can the system present data in the form of simple lists and tables, but it can also visualize it in a variety of ways. This can be a simple diagram, but it can also take other spectacular forms: a Gantt-chart-like form, a table, a network graph, a heatmap, and so on.

1.3. The Predictive Function of LAPs

The third and most controversial function of LAPs is the prediction function, where it is possible to select which case we want to make a 'prediction' for on the basis of several parameters. Incidentally, in many cases, the prediction is functionally nothing more than the classification of documents (the system classifies the document into a kind of decision type), but sometimes a prediction takes the form of a probability of winning/losing a case in percentage format. To be able to imagine prediction based on document grouping, it is worth recalling a famous experiment that is still one of the most cited in the subject.

Here, Aletras et al.[11] processed the judgments of the European Court of Human Rights, including Article 3 (prohibition of torture and inhuman and degrading treatment), 6 (right to a fair trial), and 8 (a right to respect for one's 'private and family life, his/her home, and his/her correspondence'), as interpreted by a large number of judgments. (250, 80, and 254 judgments, respectively, were available for these three articles at the time of writing.) All the judgments were then downloaded and cut up along the subheadings of the judgment, i.e. the texts were separated from each other.

The problem of predicting the decisions of the ECtHR was defined in the paper as a binary classification task. An equal number of cases ruling a violation of the Convention and not finding a violation were then selected. Information referring to the outcome of the sentence (violated / did not violate) was removed from each judgment, and the text was 'stripped down' using RegEx and highlighting stop words (waste words such as adjectives, conjunctions, etc.).

From all parts of the judgments thus stripped (remember, judgments by subheadings, e.g. facts, litigation history, etc.), information was obtained by two natural language processing methods. On the one hand, the so-called bag of words model (BOW) was adopted, i.e. word co-occurrence frequencies were searched and stored without any grammatical and syntactic features. 2,000 so-called n-grams were found and counted, each for each part of the judgment (n-gram is a specific word group, expression, or phrase). On the other hand, the so-called topic model was used, formed by grouping n-grams.

The actual prediction was made by SVM (Support Vector Machine). The essence of SVM is that from certain linguistic features (in this case, the previously formed n-grams and topics) they form a vector in a multidimensional space, thus representing the meaning of the text. In order for a machine to predict based on a 'similar meaning' (the distance of vectors relative to each other), a set of

[11] Aletras et al. 2016.

judgments must be used to teach what the final outcome will be. For this, the machine is given a set of judgments in which the final result is also told. (This is called training data.) Then, the machine was asked about another set of data to make a prediction, that is, to classify the judgment deprived of its end result in one of the groups, based on the use of words. The end result of the experiment was that the machine was able to predict the result with roughly 80% accuracy, relying mainly on the text of the factual description and legal argument.

The further results of the experiment are not relevant here, I have only described it in order to get an idea of how a prediction works in a given application. As we can see, the prediction can be of two types, percentage or binary. Percentage prediction differs from the above method in that it places a given text on a scale of 0 to 1, thus expressing the chances of a lawsuit or a motion being successful. Systems also differ in the way they predict. In some systems, only the parameters can be set (e.g. 'I am a defendant in a trademark lawsuit, before Y court, Q judge, and I want to file an X-type of motion, what is the chance of being accepted?' – such as the *Motion analyser* functions in the *Premonition* LAP[12]). Here the system does not actually make a prediction but shows a percentage distribution based on the parameter set. ('For this type of application in this type of lawsuit, the success rate before this judge is 65%.') The other type is when they show a text pattern ('a fact') to the machine. In practical applications, this is usually part of e-discovery software, and it is not the main function of the system but the exploration of all documents related to a given case on the client's servers, correspondence, etc. In this case, the software performs the same sequence of actions that we saw in Aletras et al. This thus allows the quite surreal function in these systems to be able to make an estimate of the chances of litigation based on the set of documents found.

1.4. The 'Judge Analyser' and Its Public Reception

Judge analysers do nothing more than show the data collection extracted from the documents for a judge, that is, using the judge as a query. It is easy to see that the greater the granularity of the documents, i.e. the more data they can extract from them, the larger the 'data sheet' they can compile from a judge.

Even in legal systems where only judgments are available as public documents, a wealth of data can be extracted. This data can be divided into three groups:

– *procedural and technical* details of the case (case number, dates, related judgments and appeals, identity of the trial court, co-judges, members of the panel, other dates and names – names of the parties and their representatives, possibly other parties involved in the case);

[12] *Motion Analyzer by Premoniton.ai.* 2021.

– *content-oriented (meta-)data* such as the subject of the case or the larger area of law, the direction of the decision according to the litigation roles (plaintiff–defendant) or according to the substantive legal roles in the decision (e.g. landlord–tenant), references to legislation and case law, including quasi-normative documents, such as judgments of the domestic Constitutional Court or the European Court of Human Rights, the direction of the judgment on appeal – upholding, annulling, altering, etc.);

– *data generated by other, more sophisticated NLP methods* (BOW, or other data – see above – highlighted data, word-to-word occurrences, etc., from which text style, unique features, or other things can be inferred).

From this data, a profile can then be compiled that provides quite a lot of information about the judge even without further inferences. For example, the judge's previous 'portfolio', what cases s/he has heard, in what courts s/he has worked, other professionals s/he has 'relations' with, and what cases s/he has tried can all be seen. Further, the success rate of her cases in higher courts, the most frequently cited sources of law, including precedents, and what roles s/he prefers in her judgements, e.g. whether s/he favours landlords or tenants, banks or debtors, could be revealed. Of course, all this can be generated for individual judges, or as an average for the whole court.

The judge analyser does not enjoy universal popularity even in those Anglo-Saxon countries where the publicity of court hearings and judgments is much more accepted, and the analysis of judges' judgments, subsequent citations,[13] or even their political leaning[14] is part of the legal culture. No wonder, then, that when the first such service appeared in France, it almost immediately caused a storm.

In July 2019, several legal technology portals dramatically reported that in France the 'profiling of judges' had been banned, and 'data scientists had been fired from the courts',[15] thereby significantly reducing the transparency of courts.[16] The owner of the best-known French company for data-based legal analysis (Predictice) called the decision a 'shame for democracy'.[17]

The legislation, passed by the French parliament back in the spring,[18] was enacted as part of judicial reforms and, of course, was not just about banning, but about the publication of, essentially all the French court judgments in an anonymized form on the World Wide Web. However, at the same time the decision was published, a rule was introduced that data contained in published documents may not be used to 'evaluate, analyse, and compare' the work of

[13] Kosma 1998.

[14] Lim 2000.

[15] Taylor 2019.

[16] McGinnis 2019.

[17] Tashea 2019.

[18] LOI n° 2019-222 du 23 mars 2019. 2019.

judges and court staff. Violators of the rule would be punished for the crime of 'misuse of personal data' of the French Criminal Code.[19]

An article published on the Vervassungsblog[20] suggests a previous case lay behind the stricter approach: in 2016, a lawyer-engineer named Michaël Benesty published an article on the asylum practice of French courts based on data analysis and machine learning. It turned out that there were no justifiable differences between the judgments of individual courts and judges.[21] The French courts reacted quickly, partly by denying the charges and partly by questioning Benesty's methodology.

Of course, there were also defenders of the new legislation.[22] A law professor at the University of Lorraine, for example, argued that the reaction of the Anglo-Saxon press was a simple misunderstanding and stemmed from the difference between common law and French law. In France, judges do not pass judgment in their own name but in the name of the republic, and to this day they cling to the fiction of logical syllogism (a necessary conclusion). In addition, judges in many cases sit in chambers, within which we cannot see the position of individual judges. Therefore, the profile of judges is difficult to establish and can be very distorting. Finally, as the law only prohibited the analysis, comparison, and ranking of judges and court staff by name, other data-based analytics are still possible. Judgments, groups of judgments from different perspectives, courts, and even chambers can still be analysed and compared. Predictions can still be made about the possible future outcomes of cases.

2. Opportunities and Concerns about Big Data Analytics in Law. The Future of Analytics

The French case illustrates the dilemmas of data-based legal analysis: these have been on the agenda almost constantly since the big data narrative appeared in law. Of these, there are concerns that are general and independent of the legal system and some that explicitly imply that this software type is tied to the common law legal culture. I will examine these two sets of arguments below. However, before turning to the concerns, I will briefly recapitulate the arguments of proponents of data-based law, relying on Daniel Martin Katz's much-cited article.[23]

[19] Code pénal Article 226-18.

[20] Langford–Madsen 2019.

[21] Benesty 2016.

[22] G'Sell 2019.

[23] Katz 2013.

2.1. Benefits of Big-Data-Driven Law

Katz's starting point is that the legal professions, especially the work of attorneys, have a wealth of predictive elements. Do we have a chance in this matter? How will the court judge? How much will this cost me? What are the risks of this or that contractual provision being left out of the contract? A lawyer should be able to answer all of these questions on a daily basis, but s/he can only rely on his or her own past experience and intuition. However, there are no such limitations to data-driven analysis. It can analyse almost every relevant judgment in a field of law and analyse an amount of data that a person is incapable of. It is unbiased and tireless.[24]

Katz first examines the costs of litigation. He gives an example of the TyMetrix system,[25] where the data on lawsuits in each area of law can be used to query what legal costs can be expected. Another example is the possibility to compare the effectiveness of lawyers.[26] This can be obtained from the litigation data: it is clear who the attorney was and whether or not s/he won the lawsuit. And, finally, of course, the most important thing is that the outcome of the cases can be predicted with considerable accuracy, Katz says. It analyses predictive systems in three areas: the U.S. Supreme Court, the Patent Court, and securities fraud class action on capital market fraud.

It is also interesting to take a closer look at an experiment conducted on the judgements of the U.S. Supreme Court in 2002.[27] Here the authors predicted the outcome of all of the 2002 Supreme Court judgments automatically from a variety of case-derived variables. A group of expert lawyers was used as a control group. The machine achieved 75% accuracy, while the expert control group achieved 59%. Machine prediction projected variables extracted from past data (such as the subject matter of the case and the political orientation of the judge) into the present judgments and made predictions based on them. What is astonishing is that the 'subject matter', the facts, and the legal arguments of the case did not play a role in this model. The machine predicted much better than experts based on six simple and not 'legal-professional' variables, such as circuit of origin, issue area of the case, type of petitioner (e.g. the United States, an employer, etc.), type of respondent, ideological direction (liberal or conservative) of the lower court ruling, and whether the petitioner argued that a law or practice is unconstitutional.[28]

[24] Katz 2013. 928.

[25] TyMetrix is a SaaS (Software as a Service) e-invoicing and business management solution that has a prediction module. The system is currently owned by Wolters Kluwer. https://www.wolterskluwer.com/en/solutions/enterprise-legal-management/tymetrix-360/modules.

[26] Premonition has a module with this function.

[27] Ruger et al. 2004.

[28] Ruger et al. 2004. 1163.

2.2. General Concerns

According to Devins et al.,[29] three arguments can be made in the field of law regarding big-data-based analyses.

The first argument is that no dataset and data analysis or conclusion that seems to be the most objective can be completely 'objective' and 'value-neutral' as such services often claim to be. On the one hand, all observations are 'theory-laden', especially in complex systems such as law. So, it already requires a preliminary theory of what data we collect at all. On the other hand, the interpretation of the data requires further theory, framing, value choices, and thus a series of moments from which it loses its objectivity. To this argument, we can add that the theories are laden with values, and it is especially characteristic that values and their legal precipitations and principles are contradictory. In practice, this means for these big data applications that the same language pattern or dataset, due to both its theoretical framing and its embedding in some value context, can lead to completely different results that may not be perceived by the machine.

The second argument against big-data-based law is that its ultimate virtue, its predictive power is simply illusory and virtually useless. Law is not a deterministic system and, moreover, is constantly subject to change. There are always actors who take unexpected steps, or apply unexpected reasoning that can lead to success. And the percentage probability says nothing about the prospects for a particular case, as it is not based on causation. And in law, causal relationships play a major role.

This argument actually contains two arguments as follows:

The first half of the argument is that legal proceedings are cognitive struggles, linguistic 'battles' that are not limited to specific data in the battle, legal references, and so on, but they also involve the convincing power of all this, i.e. it is the quality of the legal argument that counts. This is true for two types of legal arguments: arguments connected to the interpretation of the law on the one hand and those related to the 'construction' of the facts on the other. Both areas leave a lot of room for skills in reasoning. This is also the reason why the explanation of an algorithmic output and the explanation of a legal decision, that is, legal reasoning, are so dramatically different.[30] The proper narration and presentation of facts, including the embedding of these in our ordinary narratives, play a very important role in legal reasoning, as does the interpretation of the 'open texture' of legislation.[31] In fact, the legal decision is the proper mixture of these two operations (storytelling and interpretation) *in relation to each other*. However, machine analyses are based on hard facts and hard rules. Facts are not narratives

[29] Devins et al. 2017.

[30] Ződi 2022.

[31] Hart 1961. 124.

but only data, and rules are not human texts written in ordinary language – which can be interpreted in multiple ways – but algorithms.

The second half of the argument also contains a philosophical (though not epistemological but ontological) argument, namely that statistical relationships (e.g. correlation) that work with probabilistic relationships do not work well in law because law always deals with individual cases and is not interested in statistical correlations. How did it happen in this particular case? – asks the lawyer. And the fact that the parties lose 90% of a particular case type says nothing about whether or not I will be among the 10%.

Finally, the third argument Devins and co-authors make against data-based law is that its application in certain areas can be particularly dangerous. Here we read the well-known concerns about discrimination 'encoded' in algorithms and data and that big-data-based thinking would undermine the main source of innovation as data-based law always relies only on the past.

2.3. Concerns over Differences between Common Law and Civil Law

LAPs were invented in the Anglo-Saxon area. It is much more difficult to set up and run such services on the continent not only because the underlying legal culture is different but also because there is not as much data (documents, records) available as in the US. As I have already indicated in the description of the Lex Machina system, it processes the files in the PACER system. This means (except in cases where a sealing order has been requested or the documents are not public due to the nature of the case) that the entire file is open: with the submissions of the parties, minutes of negotiations, expert opinions, and so on. It is easy to see how much more information can be extracted from these records than from anonymized judgments, which is typically the only publicly available data source in civil law legal systems. It should be added that in a number of countries it is even not the complete set of judgments that is published, nor even a representative one, but a small fraction of it – as is the case in Hungary.

The system of data protection is also completely different in the two cultures. Data protection standards are much higher in Europe. If we look only at the simple fact that in common law systems the case is named using the names of the parties, one can immediately see the difference between the two legal cultures. In addition, existing and legitimate data protection barriers are sometimes hampered by almost incomprehensible additional barriers such as the rule that in business litigation company names should also be removed from judgments (as in Hungary), although companies are clearly not protected by the data protection provisions.

Perhaps more importantly, one of the main uses of U.S. systems is forum shopping, meaning that the parties decide where to initiate proceedings in the light of statistics and data. This is not possible in most continental jurisdictions due to strict rules regarding the distribution of cases at territorial and institutional levels. At the same time, as we read in the introduction to this volume, there is a competition between the world's legal systems, independent of forum shopping, and data-driven analysis also sheds new light on the issue of forum choice.

But the cultural difference is even more significant. In the French judge analyser debate, it was an important argument that the judge is delivering an impersonal decision, based on logic. This idea has deep roots in French legal culture. In contrast, in common law systems, it has always been part of the legal culture that the law is fundamentally a rhetorical performance, which excludes the syllogistic model.

2.4. Quo Vadis LAPs?

The reception of the French ban, and the public upheaval around it, has shown that even a partial restriction is not a viable option: if there is data, these services, including judicial profiling, will continue to be with us, at most 'at home, between the four walls' of those who hope to gain an advantage in a proceeding. Of course, this can be handled sceptically or with sad indulgence, but it is much better to look more at the benefits and what these services can be used for and what they should not be used for.

Most importantly, opponents of the services seem to be confusing two things: big-data-based legal decision-making and big-data-based decision support. From the fact that a judge or a system cannot decide on a statistical basis, why should a lawyer not look at the statistics of a judge, or a field of law, or collect and analyse the main arguments and the directions of the judgment? This first more general conclusion can therefore be formulated in such a way that big-data-based LAPs can indeed have a raison d'être as an element of the work of a lawyer.

Judges can also benefit from LAPs, as they can gather information about themselves that would otherwise come from public databases, and this may help them to identify their own bad instincts, previously unrecognized habits, stereotypical vocabulary, and so on, and improve on them.

LAPs also help to reveal the inconsistencies in judicial practice. Many argue that no two cases are the same, and 'there is certainly a reason' if a judge decides otherwise in a seemingly similar case. At the same time, it should be seen that such inequalities, which may be apparent but potentially real, also become apparent more easily through machine analysis. And then what we do with the data brought to the surface is up to us.

Foreseeing the future and making better decisions based on it is, of course, the ultimate goal, but it can also be a closer goal to simply better understand certain phenomena based on 'hard facts'. If we see all the information about a judge that can be read from his or her judgments, we get a much more nuanced picture of them, even if we do not want to make predictions about a particular case. An overview and understanding of the judicial practice in a field of law is greatly aided by seeing the length of lawsuits, which judges are the most influential in the field, the proportion of first-instance judgments upheld in the second instance, and so on. Data-based analysis can be especially helpful for doctrinal scholarship, being a control tool of doctrinal theories.

Most providers (including Premonition, for example) argue that anyone who uses such systems will gain an advantage in legal cognitive combat. However, this benefit only lasts until the other party subscribes to the same service. That is why it is also worth wasting a few words on what the legal system and legal practice will look like when this data is available to everyone. It is worth dividing this question into two parts; on the one hand, the data that show an aggregate picture of a bulk of documents (e.g. a group of judgements or a court), and the conclusions drawn from them, and, on the other hand, the data that show the performance or professional background of *people* (judges, attorneys, or parties involved in litigation).

The first type – non-personal data – will, in my view, work primarily to promote out-of-court settlements by predicting the prospects for litigation. There are already indications of this type of usage: due to the predicted legal costs, large companies already use the forecasts of their legal advisers or lawyers, often expressed as a percentage (e.g. to create a provision of the potential cost of losing a lawsuit or to calculate the amount offered in an out-of-court settlement). Arguments presented systematically in an aggregated form can be a useful aid to a lawyer in relatively simple cases to develop a tactic or, if they represent a statistically 'losing' side, to develop new arguments. For this reason, LAPs do not necessarily stifle legal innovation as an 'aggregated past' because by systematically presenting case law, they can even help improve legal innovation.

The situation is different with the analysis of legal representatives and judges. Calculating and showing a 'win rate' can have a very serious prestige-harming effect on a lawyer, so in this area some kind of restriction or regulation is desirable. (As is the case for credit ratings.) In my view, uniform methods will be developed soon for collecting and processing information about legal professionals, and these methods will be increasingly transparent. I do not even rule out the possibility of prescribing at the legislative level how to calculate a lawyer's win rate, what other information should be provided in this regard, and in which situations these indicators cannot be used. (For example, if there is not enough data.) Judges will be aware of their own profile, and if it is not to their liking, they will act against it, by consciously moving against their own data-based image. (If, for example, the assessment is that 'this judge always decides in favour of the banks', then s/he can pay more attention to making this evaluation more balanced.) What is certain, however, is that the charm of the novelty of these services will soon fade and will blend into the everyday life of traditional law, as has been the case with computer-aided legal research, e-discovery tools, or the other technical novelties of recent years.

References

ALETRAS, N.–TSARAPATSANIS, D.–PREOȚIUC-PIETRO, D.–LAMPOS, V. 2016. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science* 2: e93. https://doi.org/10.7717/peerj-cs.93 (accessed on: 22.10.2021).

ASHLEY, K. D. 1991. Reasoning with Cases and Hypotheticals in HYPO. *International Journal of Man–Machine Studies* 34: 753–796.

BENESTY, M. 2016. L'Impartialité de certains juges mise á mal par l'intelligence artificielle. *village-justice.com* – https://www.village-justice.com/articles/impartialite-certains-juges-mise,21760.html (accessed on: 22.10.2021).

CSÁNYI, G. M.–NAGY, D.–VÁGI, R.–VADÁSZ, J. P.–OROSZ, T. 2021. Challenges and Open Problems of Legal Document Anonymization. *Symmetry* 13: 1490. https://doi.org/10.3390/sym13081490 (accessed on: 22.10.2021).

DEVINS, C.–FELIN, T.–KAUFMANN, S.–KOPPL, R. 2017. The Law and Big Data. *Cornell Journal of Law and Public Policy* 27: 357–414.

G'SELL, F. 2019 Predicting Courts' Decisions Is Lawful in France and Will Remain So. *actualitesdudroit.fr.* – https://www.actualitesdudroit.fr/browse/tech-droit/donnees/22630/predicting-courts-decisions-is-lawful-in-france-and-will-remain-so (accessed on: 22.10.2021).

HART, H. L. A. 1961. *The Concept of Law*. Oxford.

HOGE, P. 2013. Lex Machina: 'Moneyball' Meets Patent Lawsuits. *San Francisco Business Times* 13.09.2013. https://www.bizjournals.com/sanfrancisco/blog/2013/09/lex-machina-mines-lawsuit-data.html?page=all (accessed on: 22.10.2021).

KATZ, D. M. 2013. Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry. *Emory Law Journal* 62: 909–966.

KOSMA, M. N. 1998. Measuring the Influence of Supreme Court Justices. *Journal of Legal Studies* 27: 333–372.

LANGFORD, M.–MADSEN, M. 2019. France Criminalises Research on Judges. *vervassungsblog.de* 12.06.2019. https://verfassungsblog.de/france-criminalises-research-on-judges/ (accessed on: 22.10.2021).

LIM, Y. 2000. An Empirical Analysis of Supreme Court Justices' Decision Making. *Journal of Legal Studies* 29: 721–752.

MCGINNIS, J. O. 2019. Transparency and the Law in France. *lawliberty.org* https://www.lawliberty.org/2019/06/20/transparency-and-the-law-in-france/ (accessed on: 22.10.2021).

MITCHELL, R. 2018. *Web Scraping with Python*. Sebastpol (USA).

RAPOPORT, N. B.–TIANO, T. R. Jr. 2019. Legal Analytics, Social Science, and Legal Fees: Reimagining Legal Spend Decisions in an Evolving Industry. *Georgia State University Law Review* 35: 1269–1304.

RUGER, T. W.–KIM, P. T.–MARTIN, A. D.–QUINN, K. M. 2004. The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision Making. *Columbia Law Review* 104: 1150.

SCHÖNBERGER, M. V.–CUKIER, K. 2013. *Big Data. The Essential Guide to Work, Life, Learning in the Age of Insight*. London.

TASHEA, J. 2019. France Bans Publishing of Judicial Analytics and Prompts Criminal Penalty. *abajornal.com* – http://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics (accessed on: 22.10.2021).

TAYLOR, S. 2019. French Data Analytics Law Won't Stop Analytics, Attorneys Say. *law.com* 07.06.2019. https://www.law.com/legaltechnews/2019/06/07/french-data-analytics-law-wont-stop-analytics-397-21251/?slreturn=20191006132927 (accessed on: 22.10.2021).

ZHANG, S.–HE, L.–VUCETIC, S.–DRAGUT, E. C. 2018. Regular Expression Guided Entity Mention Mining from Noisy Web Data. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels): 1991–2000.

ZŐDI, ZS. 2017. Law and Legal Science in the Age of Big Data. *Intersections. East European Journal of Society and Politics* 3: 69–87.

2022. *Algorithmic Explainability and Legal Reasoning. The Theory and Practice of Legislation*. https://doi.org/10.1080/20508840.2022.2033945 (accessed on: 05.01.2022).

*** *360Quadrants: Best Legal Analytics Software*. 2021. https://www.360quadrants.com/software/legal-analytics-market (accessed on: 22.10.2021).

*** *Code pénal Article 226-18*. https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000006417968&cidTexte=LEGITEXT000006070719 (accessed on: 22.10.2021).

*** *Lex Machina*. 2015. LexisNexis Acquires Premier Legal Analytics Provider. 11.23.2015. https://lexmachina.com/media/press/lexisnexis-acquires-lex-machina/ (accessed on: 22.10.2021).

*** *Lex Machina – How It Works?* 2021. https://lexmachina.com/how-it-works/ (accessed on: 22.10.2021).

*** *LOI n° 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice Article 33*. 2019. https://www.legifrance.gouv.fr/eli/loi/2019/3/23/2019-222/jo/article_33 (accessed on: 22.10.2021).

*** *Motion Analyzer by Premoniton.ai*. 2021. https://premonition.ai/legal_analytics/#motions (accessed on: 22.10.2021).

*** *TyMetrix*. https://www.wolterskluwer.com/en/solutions/enterprise-legal-management/tymetrix-360/modules (accessed on: 22.10.2021).