ZSOLT ZŐDI *

Law and Legal Science in the Age of Big Data

Intersections. EEJSP
3(2): 69-87.
DOI: 10.17356/ieejsp.v3i2.324
http://intersections.tk.mta.hu

* [zodi.zsolt@tk.mta.hu] (Hungarian Academy of Sciences Centre for Social Sciences Institute for Legal Studies)

This paper has been prepared as the part of OTKA 116979 research project, *Regulatory Issues of the Internet Intermediaries.*

Abstract

The connection between Big Data (BD) and law can be thematised in several ways. This article aims to contribute to the understanding of the different levels of interplay between Big Data, law and legal science. The paper firstly considers Big Data as the subject of legal regulation. Accordingly, it overviews the moral questions surrounding Big Data, BD's predictive potential as well as the impacts of it on legal framework rules regarding privacy, data protection, competition and business regulation. The next section understands Big Data as a tool in the regulator's and the lawyer's hand. It discusses the new ways of 'Big Data-based social engineering' as well as the creation of predictive tools and inferencing techniques based on Big Data in policing, law enforcement and litigation. Then the paper investigates the use of BD in legal science, thus the fourth section considers Big Data as a research tool. It seeks to explore the use of legal data-sets and textual corpuses as BD. In addition, it sheds some light on the wider impacts of statistical analysis, natural language processing, content analysis, machine learning and behavioural prediction on legal science. Finally, the paper gives some insight into the relationship between traditional doctrinal scholarship and the new types of BD-based research.

Keywords: Big Data, Legal Regulation, Legal Science, Law-making, Prediction in Law, Data-driven Lawyering.

1. Introduction

The phrase 'Big Data' (hereinafter: BD or BD phenomenon) was born in the IT sector, and it was first described by IT (e.g. Ahlberg, 2011) and business journals (see e.g. Economist, 2010), and later in the seminal book by Mayer-Scönberger and Cukier (Mayer-Scönberger and Cukier, 2013). It later became a buzzword in the business sciences, sociology and public policy. Though it has influenced law and legal science, and the Mayer-Schönberger – Cukier book itself also devotes a whole chapter to the risks (practically regulatory aspects) of BD, still, the number of reflections in law is significantly lower than in the domains mentioned above.[1] This paper aims to contribute to the understanding of the connection of law and legal science, on the one hand, and the BD phenomenon, on the other.

The term Big Data is used in a rather loose way in the literature. Most of the sources are using, or at least are mentioning the 'magic' 3-4-5, 'V'-s (Volume, Velocity, Variety, Veracity and Value)[2] as a definition. However this can hardly be considered a classical definition: it only gives the *differenta specifica* describing how Big Data differs from other 'things', (in the 3-4-5 'V'-s) but does not provide the *genus proximum,* (the family of phenomena to which the Big Data belongs). Does BD herald a new period in history? Or only a new era in the development of the information society? Or is it a new *driving force* that is changing our society? A new *mindset*? A set of *attitudes*? And if it is, what is the scope of this mindset? My definition in this article is a narrow and simple one: I use the term Big Data for the new (technical) ways, solutions and methods of producing, collecting, processing and using data, which together, as a driving force, might ultimately change the mindset, the attitudes, and all of society, – including the law. Big Data therefore initiates social changes, but it is not the social change, nor the new historical period *in itself*.

A vast amount of data is generated on the internet every second by people and by machines (sensors). These data mainly provide information about people. (Data on natural phenomena, such as the data from the Large Hadron Collider or weather data, are sometimes also considered to be 'Big', but this aspect has no relevance here). It includes a whole range of data from the cell information, location and call (meta)data of mobile phones, the search strings typed into search engines, and click information on websites, as well as the data generated by sensors, smart meters, and online cash machines, or the millions of pages written and published by government officials, also known as 'open data' (Mayer-Schönberger and Cukier, 2013: 116–118). It no longer represents an IT or a data storage problem, but a social phenomenon, since this amount of data not only requires different storage methods, handling and interpretation techniques, but can be and already is being used in totally different ways than 'normal' data. These new data-usage practices will have a severe social impact. Some aspects of this impact are already detectable, but some are still to come.

[1] At Wiley alone, 30 books were recently available analysing BD in business, while there is no monograph in the legal field, although recently, HeinOnline contained some 180 legal articles. The White House is also very active in producing policy papers, reports, and other documents (see White House, 2012; White House, 2012c; White House, 2014; White House, 2015a; White House, 2016; NFS, 2015).

[2] For the three V-s: (Eaton et al., 2012: 5), for the four and five: (IBM Data Hub, 2016), but there are already sources who mention seven: (DeVan, 2016)

The aim of this article is to collect and systematise the arguments concerning the impact of BD on law (on legal regulation and lawyering) and its potential effects on legal science. My ambition here thus is not more than to collect and systematise the arguments, and insights from the available legal literature, and other influential resources – mainly from policy papers - that has a legal (regulatory) relevance, and partly to frame these issues and arguments. Hence, this writing is a collection of ideas and arguments, sometimes predictions and framings of leading authorities in the field, completed with my own observations. As with all predictions of this kind, those provided here are subject to mistakes. Finally, some of my observations and thoughts are purely descriptive, while others may have prescriptive elements, (like the 'ethics of BD' section).

BD as a term appeared before 2010, but it became a popular subject of analysis only at the beginning of the present decade. This was the time during which most scholars realised that the mass production of data was a sign of deeper changes beneath the surface of society. Since then, hundreds of articles have been published in leading legal journals on different aspects of law and BD. Some of the papers have analysed BD in general, addressing questions such as the ethical problems raised by BD or its impact on data protection. Some other works have analysed specific problems arising from concrete well-known cases.

The connection of BD and law can be thematised in several ways. I distinguish between the effect *to law, and legal science*. First I discuss the interplay between law (legal regulation) and BD. Here, a further distinction is made, whereby BD can be the *subject* of legal regulation, but it can also be a *tool* for better, 'predictive' law making and application of law and policing. After this, in the third section I will discuss the potential impact of BD on *legal science*. The following table illustrates the three domains of interaction between law, legal science and BD.

Table 1. The role of BD in law and legal science.

	BD as a subject	BD as a tool
Law	❶ How should law frame, define and regulate the BD phenomenon? How will BD change existing privacy, data protection, competition, business regulatory, etc. rules? What will the new rules regulating BD look like? Methodological and theoretical (including ethical) questions about BD regulation, methodological and theoretical questions about using BD methods in law making and law enforcement. Moral dilemmas of prediction.	❷ How can we exploit the new possibilities provided by BD in law making, policy creation and the application of law? How can we design new ways of 'BD-based social engineering'? How can we create predictive tools and inferencing techniques based on BD in policing, law enforcement and litigation
Legal science	❸ BD as a new research tool in legal science. The use of big data-sets and textual corpuses as BD. How will these 'super-empirical' research methods change legal scholarship? What is the relationship between traditional doctrinal scholarship and the new types of BD-based research? How can we use statistical analysis, natural language processing, content analysis, machine learning, behavioural prediction, etc. in legal science?	

2. BD as the subject of legal regulation

2.1 Risks

The first point of connection is that the BD phenomena visibly raises a whole range of risks that eventually must be handled in some way – also by the law. And BD – as Mayer-Schönberg and Cukier put it, are not only increased risks of the past, but the BD 'changes the character' of the risks, (Mayer-Schönberger and Cukier, 2013: 153). Regulators are facing a serious dilemma: BD offers new possibilities in business and government, but implies dangers that are not clearly foreseeable. The dilemma is present in policy papers and discussions, (White House, 2012; White House, 2016; FTC Conference, 2014) and within the literature (e.g. Tane and Polonetzky, 2012: 63–69.) As usual, even the need for any, and especially any new regulation is questioned sometimes (Big Data and the Law Blog, 2014).

As for the risks, the literature mentions the following interrelated dangers:

1. Even in the case of data collected with the consent of the data-subject, because of the quantity of the data and its connection with other data elements, power is created on the part of the data-owner which is far beyond the normal data-protection scenario. The classic example here is the Target case, in which a BD-based algorithm that processes the buying habits of customers figured out the pregnancy of a 16-year-old girl (Duhigg, 2012; Mayer-Schönberger and Cukier, 2012: 152–153). An aspect of this danger is, that within big databases, de-anonymisation can be relatively easily based on metadata and the use of certain algorithms (Ohm, 2010: 1718; Mayer-Schönberger and Cukier, 2012: 154). On a wider horizon, as Tane and Polonetzky (2012) and Crawford and Schultz (2014: 94) have pointed out, nearly all categories of 'traditional' data protection are being questioned, and especially 'notice and consent', data minimisation' and 'principles of purpose' elements. For example, traditional data protection regulation is based on the consent of the person. But in the age of the BD, so much data is generated by a person that simply no one can control it. 'Can you imagine Google trying to contact hundreds of millions of users for approval to use their old search queries?' – ask Mayer Scönberger and Cukier. Or what 'legitimate purpose' of the data processing means, when 'the most innovative secondary uses haven't been imagined when the data is first collected' (Mayer-Schönberger and Cukier, 2012: 153).

2. There is a new phenomenon: the 'predictive power' of BD (McGregor et al., 2013; Siegel, 2013; Simon, 2014; White House, 2014; Ferguson, 2015; Jeon and Jeong, 2016). A company can look much more deeply than before into its customers' habits with the help of BD, and based on that, it can exercise, for example, discriminatory practices. The problem has been discussed extensively in policy papers (White House, 2015b), as well as in a conference organised by the Federal Trade Commission in 2014 (FTC Conference, 2014). The FTC also detected another new BD-based risk, namely discriminatory pricing. Further, everyone is familiar with (and sometimes, does not like) the surge pricing of UBER, for example, which is also based on BD algorithms, and many people do not like it (Dholakia, 2015).

It is quite normal, that when a new social phenomenon is forming, there is neither consensus about a definition among scientists and experts, nor any agreement

among regulators and stakeholders whether to regulate it or not, and if so, how. But it is also a common experience that eventually, the definition *of the law* will finally be much simpler – sometimes surprisingly so – than the definitions given by the experts or social scientists. It is likely that this will be the case with the BD phenomenon. I think, that from the legal, regulatory, 'risk-centred' point of view, the important aspect of BD is not the 'four Vs' or that it has been collected without explicit consent, but that there are huge data sets that have been collected with *one* particular purpose (or even generated spontaneously, without any aim), which are used *for another purpose*. The other peculiarity of BD is that, under certain conditions, one can make relatively accurate *predictions* based on it.

BD therefore creates an information 'super-power' on the part of the data owner. A further problem is that these predictions are made by algorithms that are not transparent to the average citizen, and their inferences cannot be understood by 'common sense knowledge'. Most of the recommendations appear to be aimed at mitigating or compensating for this superpower, and they urge transparency concerning the algorithms and the decisions generated by these algorithms (Mattioli, 2014: 537; EU Regulation, 2016: 13.2.f, 14.2.g).

Therefore, this article opines that BD is primarily *not a data protection problem*. Traditional data protection regulations can be applied to the BD world, but they would deprive BD of its value-creating characteristics. The paradox is that the risks of BD are identical to its most significant value-creating power.

2.2 Ethics of BD

Traditional doctrinal scholarship can do little or nothing regarding the BD phenomenon, because there is not yet any regulation or jurisprudence in the field. So, in the case of BD and legal science, one must pursue other avenues, such as Richards and King (2014) who aim to establish the foundations for future regulations addressing 'Big Data ethics'. As they stated: 'We have some privacy rules to govern existing flows of personal information, but we lack rules to govern new flows, new uses, and new decisions derived from that data' (Richards and King, 2014: 408). They lay down four high level principles. First, privacy will not be dead in the era of Big Data, but it should rather be perceived as 'information rules' than as 'information we can keep secret or unknown'. They stated:

> Privacy should not be thought of merely as how much is secret, but rather about what rules are in place (legal, social, or otherwise) to govern the use of information as well as its disclosure (Richards and King, 2014: 411).

Second, in the BD era, we must rethink our attitudes towards sharing personal information. Shared private information can still remain confidential, and that is what counts. Information always exists in intermediate states between completely public and completely private. We often share information with trust, expecting that it will remain confidential. The third ethical standard in BD ethics is transparency. Transparency, as the authors stated, 'fosters trust by being able to hold others accountable'. According to the authors, BD practices should be as transparent as

possible, though they also admit that this will create a problem they call 'a transparency paradox':

> Transparency of sensitive corporate or government secrets could harm important interests, such as trade secrets or national security. Too little transparency can lead to unexpected outcomes and a lack of trust. Transparency also carries the risk that inadvertent disclosures will cause unexpected outcomes that harm privacy and breach confidentiality (Richards and King, 2014: 420).

Finally, the fourth standard is the standard of identity. In the BD era, based on inference and predictive algorithms, governments, companies, and organisations can create a profile on us, and practically decide who we are, under which categories we belong, before we make up our own minds. There should be rules that empower us to define ourselves against the machine-made identity.

There are some insights beyond these ethical standards. The first is that BD's predictive and inferential power will enable the machines to make decisions that cannot be explained by our traditional narratives or justified by our 'traditional' justification techniques. And this problem is not solved by the rule that 'meaningful information about the logic involved' should be provided by the controller (EU Regulation, 2016: 14.2.g). Imagine that a machine makes a prediction that a certain group of men with a definite skin and hair colour, height, social status, and shoe size (just to be even more absurd) will commit violent crimes with a 90 per cent probability. Will the authorities stay idle? Or will they at least place these people under surveillance? And if they do, how this will be justified? How can any measurement be justified that is based on attributions that are not under the control of the person? How can the inference of a machine be justified that is not based on our 'normal' moral narratives and 'causal explanations', but on some hidden interrelationships based on a huge amount of data? Let us just consider the terrorist dilemma (Brugger, 2000), assuming that the machine pinpoints a person who will, with 99 per cent probability, commit a terrorist attack. Will we do anything, and if we do, what will be the underlying reasoning?

2.3 The dilemma

If we have so many risks and fears, why should we not just put a ban on BD practices? First, because it is impossible, second because it is very inexpedient. Nearly all scholars agree that BD has an enormous value-creating potential. Byers (2015) pointed out five areas in business for which BD can create value.

1. 'Creating transparency to big data often exposes variability in performance and results, leading to changed behavior for more economic impact' (Byers, 2015: 758). This means that BD encourages economic performance.
2. BD enables experimentation and gives direct feedback for different solutions, business models, and product-types. For example, Tane and Polonetzky (2012) discussed BD-based web-analytics as follows:

> (W)eb analytics - the measurement, collection, analysis, and reporting of internet data for purposes of understanding and optimizing web usage - creates rich value by ensuring that products and services can be improved to better serve consumers (67).
> 3. BD enables companies or organisations to segment populations in a very sophisticated way. BD can not only be a tool for marketers but also an excellent new basis for improved risk-management.
> 4. In certain fields, human decisions (and human errors) can be replaced by BD-based decision-making.
> 5. Big data enables innovations in business models or pricing. UBER's surge pricing is an excellent example.

BD as the most important driving force of the future economy is also present in policy papers both in the EU and in the USA. A recent EU document states: 'Big data technology and services are expected to grow worldwide to USD 16.9 billion in 2015 at a compound annual growth rate of 40 per cent – about seven times that of the information and communications technology (ICT) market overall.' (Communication WP, 2014:2) The document mentions the smart grid, health, transport, environment, retail, manufacturing and financial services as BD areas (Communication from the Commission, 2014: 2). The White House also shares this optimism concerning BD: 'big data technologies continue to hold enormous promise, as the report identified—to streamline public services, to advance health care and education, and to combat fraud and complex crimes like human trafficking' (White House, 2014).

However, the value creation potential of BD prevails only if BD sets are disclosed. Mattioli (2014) argued that disclosure of the BD sets should be encouraged:

> Much of the rhetoric describing big data's potential for innovation assumes that data can be easily and meaningfully reused and recombined in order to examine new questions [...] Most significantly, big data's producers tend to infuse their products with subjective judgments that, when left undisclosed, limit the data's potential for future reuse. [...] These conclusions point toward the need for new policies designed to encourage the disclosure of big data practices (544, 549, 570).

An important contradiction is apparent here. On the one hand, BD and BD's predictive power create a dangerous imbalance and increasing vulnerability among customers. On the other hand, the BD on which these predictive algorithms are based represent huge potential and value-creating power. Some assert that BD sets should be disclosed in order to increase their value-creating ability. But disclosure will increase the vulnerability of private persons, and since some of the BD sets collected by private companies are some of their most valuable assets, they are not eager to share them with anyone else. One of the most serious issues in the coming years will be to find an equilibrium in regulation between the values of transparency and limited usage, and value-creating freedom of use.

The EU's approach to the BD phenomenon is apparently also controversial. As we know, the EU initiated the revision of the data protection directive (EU

Directive 1995/46/EC) in 2012 (European Commission Press Release, 2012); at that time, Big Data was simply not an issue or at least not in the recent narrative framework. The narrative of the EU during the revision was that the level of data protection should be *increased* and should be brought to the same level across Europe. According to the reasoning underlying the Regulation, a higher data protection standard would leverage trust, because '(b)uilding trust in the online environment is key to economic development' (Commission WP, 2012: 7). I have doubts whether this argument is so simple. Creating higher standards can result in a higher level of trust, but at the same time it increases administrative burden, or can even create obstacles for the enterprises that want to exploit the power of BD.

The Regulation contains only a few amendments with a connection to Big Data, and these amendments clearly show that this regulatory environment is not aimed primarily towards a 'data-driven economy'. Let us take one of the most important fields of BD: regulation of automated decision making and profiling. There are two main rules in this field. First, '(t)he data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her'(EU Regulation, 2016: 22.1), and second, the data subject has the right to obtain information about 'the existence of automated decision-making [...] and meaningful information about the logic involved' (EU Regulation 2016: 15.1.h)) He/she is also entitled to ask for 'human intervention' (EU Regulation, 2016: 22.3) at any time. I have doubts if these rule can be aligned, and it is also unclear whether this will facilitate a 'thriving data-driven economy'. Later, the case of two American legal information service companies are expounded, that are building their services on the liberal publication policy of American court documents. These services, in their existing form, simply *would be legally impossible to build in Europe*. But it is not only about data-protection rules. I already mentioned the connection between open data and the data economy. Despite all efforts, open data initiatives are developing slowly across Europe (Nicol et al., 2013; Open Data Maturity in Europe, 2016).

Because of all these controversies, the contents of the future regulation cannot yet be ascertained. Will it simply amount to some new rules within the existing data protection regulation and for e-commerce, or will it change the whole regulatory landscape? Traditional data protection rules in the 1980s and '90s protected ordinary citizens against governments, and later, up to the present, against large companies. In this respect, BD has created a new situation. It is collected and used in a non-transparent way, and it enables the data owners to make predictions and thereby to 'control the future'. But there is still too little evidence about this predictive power. It is unclear whether famous oft-cited cases (like the Target-case) are really the forerunners of incidents that will occur very frequently or if such cases are only accidental and isolated stories. We do not yet know whether the BD collected about us and our fellow-citizens/customers/parties will enable these organisations to really know even more about us than we know about ourselves. The other side of the coin is that it has also not been established that BD will bring about a new era characterised by a 'data-driven economy' (whatever this means). As long as the answers to these questions are unknown, we must be very cautious about designing new rules.

3. BD as a tool in the regulator's and the lawyer's hands

Law is not just regulating, but can also *rely on* BD. As in many field of the business, healthcare, education and other sectors, it can also use it as a tool. In the field of law, however, we can differentiate between two levels of reliance: BD, on the one hand, can support *law making* and policy design, but on the other, it can be a tool in the hand of officials, lawyers, and judges in law application (lawyering), law enforcement and litigation.

3.1 BD in law making

In the field of law making and policy design, it is likely that better rules and regulations can be created with the help of BD: The EU also appears to adhere to this idea (H2020 Call, 2016). BD can open new perspectives in the preparation and design of rules, but also in the measurement of the effects of the amended rules. There is already a quite simple requirement that regulation should be based upon facts and data. (In Hungary for example this is also a legal requirement, since § 17 of the Act CXXX of 2010 on Law-making rules, that the law-maker should prepare an impact assessment on economic, social, budgetary, environment, safety fields – Act CXXX of 2010) For example, in planning VAT revenue it is important to know something about the gross retail turnover. If corporation tax is raised in the hope that it will bring in extra revenue, the number of companies and their profits should be studied. However, the possibilities offered by BD go far beyond this. In BD sets the data representing a certain social aspect is complete and available in real time. For example, the data generated by the online cash registers recently introduced in Hungary are the comprehensive and real-time set of data for the retail sector, and are not retrospectively collected or representative sample-based. The traffic information recorded by highway cameras registers every vehicle. The cell information from mobile phones shows the real movement of citizens, which is not distorted by the memory of the person who recalls it. Communications via social media show real-time human interactions.

Therefore, BD can support law-making in several ways. First, the effect of a policy decision can be measured by data outputs, which show changes at the micro-level. For example, a policy (law) change aimed at companies can be measured by the company register and the balance sheets or P&L statements published by companies. Second, in the era of BD, the initial data on which a policy decision is based are available in a complete and real-time format. The cases of the cash registers or the traffic information provided by highway cameras were mentioned above. At present, these data – if considered at all during law making – have been available only in an incomplete form, showing the past, not the present. Third, BD enables law-makers to experiment and to simulate certain policy decisions in smaller populations and to immediately measure the consequences of these decisions on certain outputs (for experiments with BD, see Byers, 2015). The acceptance of a decision or a policy change can be immediately monitored via social media, for example, or the increase or decrease of crimes via the information provided by CCTV cameras or information systems operated by the police.

3.2 BD in lawyering

The application of the law (law enforcement, litigation, decision making, drafting of documents, etc.) can be supported by BD as well. This is a field in which there are existing examples. Lex Machina and Ravel are two functioning applications, both of which are based on BD.

Lex Machina, which was recently acquired by Reed Elsevier, offers 'legal analytics' in three fields and argues that its software represents the 'third leg to the law practice stool' next to traditional legal research and legal reasoning.[3] The product captures the litigation data and documents published in PACER,[4] UPSTO,[5] and EDIS,[6] – all open data – then mines and analyses the data with the help of artificial intelligence software. This means that it extracts data from these documents (players, asserted properties, findings, and outcomes, including damages awarded, etc).

The logic behind Lex Machina's competitor, Ravel,[7] is nearly the same. It extracts information from litigation documents with the help of natural language processing algorithms, which, at the same time, have the ability to engage in machine learning and visualise the results in a very spectacular form.

Both companies can analyse individual judgements, areas of law, judges, and courts, and in certain areas, they can also do predictions on the outcome of a certain type of case, offer a certain type of language that has proven to be preferred by a particular judge, or plan a litigation strategy.

So, the use of BD in this area is already a reality. But beyond its predictive possibilities, which is also a key element here, these services throw light on all the methodological questions *on the interpretation* of data as well. It is common knowledge that the interpretation of statistical data has raised methodological issues and can be the source of huge errors, even if the data collection is carefully planned (see e.g. the 'McNamara sin' mentioned by Mayer-Schönberger and Cukier, 2013: 163–165). But BD has created many new problems, since the data *collection is done from data sets which were not originally designed for that particular purpose*. In the case of judgements and other free text documents, natural language further increases the possibility of misinterpretations and false conclusions. As Kris Hammond (2015) stated:

[3] https://lexmachina.com/law-firms/

[4] Public Access to Court Electronic Records (PACER) is an electronic public access service that allows users to obtain case and docket information online from federal appellate, district, and bankruptcy courts: https://www.pacer.gov/

[5] The website and database of the U.S. Patent and Trademark Office: https://www.uspto.gov/

[6] The website and database of the U.S. International Trade Commission (USITC): https://www.usitc.gov/

[7] http://ravellaw.com/

> Decision makers don't want data. They want to understand what's happening in the world. Data for the sake of data is a waste of time and money. Spreadsheets, visualizations, and dashboards fail because they may express the data, but they don't communicate facts and the events in the world that gave rise to them. [...] Likewise, the data associated with us as individuals, including the wealth of data from the emerging Internet of Things will be transformed into reports that real people will be able to read and understand. Rather than seeing data, they will see stories of their own lives mapped out for them based on artificial intelligence language systems looking at their data and explaining it to them (Hammond, 2015).

If we just consider the 'Judge analyzer' service of Ravel, whose system promises to uncover 'the rules and specific language your judge favors and commonly cites' and to 'pinpoint distinctions that set your judge apart', the hidden narrative behind it is that using the language, the distinctions, the arguments, the concepts and sources a particular judge prefers, can help to win the case. An even further and deeper narrative behind this is that there is a connection between the quality of the reasoning and winning the case. This narrative is certainly not self-evident for continental legal systems, where the quality of the reasoning is often not determinative in legal proceedings – and sometimes of course this is not the case in the Common Law systems either.

In the world of BD, it can sometimes turn out that our narratives – the big tales and the common interpretational frames – fail. What kind of narrative can be attached to the fact that a positive relationship exists between disgust sensitivity and political conservatism? (Inbar et al., 2011). How many such hidden interrelationships will be discovered that do not fit our existing narratives? Will BD be the next field for which we need to adjust our traditional narratives as we did after the development of quantum physics?

3.3 BD-based application of rules

If the predictive power of BD analytics is so powerful, is it not better to use these algorithms, which are based on real time, complete and detailed data, for example, to establish sentences in criminal cases to eliminate proven sentencing disparities? (Kunz and Majairan, 2016; Volkov, 2016; Windergren et al., 2016). Or is it not possible to use this power in civil law cases in which judges must interpret discretionary categories, such as a 'reasonable time' or 'fair compensation'. Would it not be better to use BD-based algorithms to actually consider every detail?

Outside the realm of the law, these BD-based decisions are already quite common. Just think about the scoring process used by banks when they decide whether to grant a loan, which is, in great part, based on data and algorithms. The process eventually ends with a number. The same applies to insurance companies' risk assessment process. They have one thing in common: Applicants do not receive a justification after the decision. This is partly because the decision is based on personal characteristics that the applicant cannot change, and partly because the decision is made based on data-based algorithms, which either cannot be explained using plain

words, or it is not in the interest of the decision-maker to disclose the underlying rules. The European General Data Protection Regulation contains rules on automated decision making, but as I hinted above, this rule can be counterproductive and can hinder the EU from achieving a 'thriving data-driven economy'.

If this brave new world arrives, there will be several consequences. Just to mention two: first, imagine if sentencing was BD based. When creating the algorithms, the value judgements that until then, were hidden, would be explicit.[8] In the case of sentencing, the goal of 'special prevention' (that the punishment should prevent the perpetrator from committing a crime again) and the goal of 'general prevention' (that the punishment should deter others from committing crimes) should be represented in the algorithm, and therefore, should be transformed into variables, which are made explicit.

Another consequence could be even more interesting. The unity (or uniformity) of the decisions within a legal system is an important constitutional principle. We tend to think that BD-based algorithms will produce more uniform decisions, because the same algorithm can be used across the whole legal system. But the opposite could occur. BD-based algorithms, simply because they are able to examine and process far more considerations parameters and circumstances than a human, can make more *diverse* decisions. This brings us back us to problem #1: in these cases the two sides (two sub-principles) of the same principle, – the justice – are conflicting. The first says, that since there are no two similar cases, every case must be treated differently. But the other principle of justice says that 'like cases should be treated alike'. There is a certain point, where the decisions of complex algorithms simply cannot be explained by plain human words, because they do not fit into our everyday narratives. In these cases it will be for us to decide to use these algorithms, and create a justification, or to ignore them and take back control over the decision process.

4. BD and legal science

4.1 BD as a 'super-empirical' method

In many respects, BD-based research projects are not different from 'simple' statistical research projects, which are also based on great volumes of data, performed by computers, and use statistical and mathematical algorithms to process data. Nevertheless, BD has changed the landscape of the social sciences, and there have been extensive debates about how it will affect the methodology of social science research (Williford and Henry, 2012). It is far beyond the scope of this article to elaborate on the differences between the 'old' methods and the 'BD' methods. This article merely seeks to draw attention to some spectacular ones.

The first involves the population being studied. In a narrow social domain, BD shows *the whole picture*, (Mayer-Schönberger and Cukier, 2013: 22–31) and not a picture of a random or a representative sample. Normally, the domain under study is

[8] An interesting example of making hidden value preferences explicit for the algorithm of an autonomous car is the 'Moral Machine' project by MIT: (http://moralmachine.mit.edu/).

smaller compared to traditional empirical research, and there are different distortions, compared to the previous research. If research is conducted on people's movements based on the location data of their cell phones, this will use real-time and undistorted data, but will not show, for example, the aim of the movement, which would be asked in a survey.

The second difference is that a statistical analysis is always preceded by a conscious and planned data collection process. The data collection is preceded by the data collection design (e.g. the drafting of questionnaires), and this is preceded by a hypothesis based on a narrative or a paradigm. Therefore 'normal' empirical research can never change the starting paradigm or narrative. It can falsify the hypotheses, and start the whole process from scratch, but it cannot change the paradigm itself. In case of BD research, *the birth of the data precedes the research*, and the researcher must somehow process and interpret the data after it has been created. Therefore, in BD research, if the hypothesis is falsified, one must rethink the paradigm as well. If the data does not fit into the existing narrative, one must change it or find another one (Deardorff, 2016).

Third, it seems that because of the volume and complexity of BD, the visualisation and interpretative tools are far more important in presenting the results of BD-based research than in any other field. It is quite natural for statistical results to be presented in diagrams and not only in tables. But in the case of BD, the visualisation is not simply a way to *better present* the data, but sometimes the only way to present it. Normally, BD simply cannot be presented in its raw form, like, for example, a report on statistical research normally presents the survey questions and the dispersion of answers for every question.

4.2 'Doctrinal' and 'empirical' legal science

How does all of this affect legal science?

In the past few years, – as happened more than once in the last century of legal science – it has become one of the leitmotifs of the methodological writings concerning legal science that traditional 'doctrinal' scholarship seems to be in a crisis (Bodig, 2015; Dyevre, 2016); one of the escape routes could be empirical, data-based research, through which legal science could become a 'real' social science.

To understand the problem, we should first clarify the relationship between 'traditional' legal science and 'empirical' science, which is considered to be 'real' social science, and the further difference between the 'old' empirical methods and the new method offered by BD in the legal domain.

Legal science's traditional role, which is sometimes called 'doctrinal', or in the German-speaking parts of Europe, 'dogmatical', 'ranges between straightforward descriptions of (new) laws, with some incidental interpretative comments, on the one hand, and innovative theory building (systematisation) of the other' (Hoecke, 2011: vi). Regardless of how innovative it is, doctrinal science always analyses *texts*, namely some *important* texts in the framework of normative concepts, which is partly established by the text of the laws, partly by judicial practice, and partly by legal scholars.

Empirical research, on the other hand, is centred on social, and sometimes, psychological 'facts'. Further, research can be empirical without using data or bigger samples, so there are additional (narrower) categories of research methods, namely those which are 'data-based' and 'statistical'. The former uses numbers and variables to describe certain social phenomena, and the latter relies on representative samples and statistical methods such as dispersion and correlation, and standard ways of segmenting a population.

Therefore, the different types of research in the legal domain are the following:

Table 2. Different types of legal research.

	Methodology and conceptual network	Observed object	Observed population	Reliability of prediction
Doctrinal	Desktop, using an existing normative, legal conceptual framework	Manifestations of the normative object – texts	Some 'important' texts	–
Empirical	Based on sociological methods, using social science methods, sometimes based on 'numbers' using concepts of social science and mathematical methods	Objectivations of the social phenomena and/or texts, or data taken on social phenomena and/or data taken on texts	Accidental selection of social facts.	Low
Statistical	Based on representative data, using concepts of social science, mathematics, and statistics	Data taken on social phenomena and/or data taken on texts	Representative sample	High
BD based	Mathematical methods, narratives and conceptual framework employed retrospectively	Data sets, in most of the cases, huge text corpuses processed as data	The total population/data-set	Very high

For the same research question (for example, 'how has medical malpractice litigation changed in the last five years, and what are the future trends?'), there are five possibilities to elaborate the topic. The doctrinal research will comprise the reading of the most important higher court decisions and the analysis of the conceptual framework within these documents. An empirical study can include a questionnaire completed by counsels and judges active in the malpractice field. Data-based research can complete this with data connected to medical malpractice, for example, the length of the court procedures or the damages paid by hospitals. A statistical enquiry would conduct all of them using representative samples. Finally, the BD-based research could include complete data-sets, such as the whole aggregation of hospital and litigation documentation, which can 'say' anything about malpractice litigation.

Why is the situation of the legal domain special? Empirical *legal* research is always in a very strange, in-between situation for two reasons. First, the social facts it observes (such as the 'behaviour of judges' or the 'medical malpractice' itself) are based on normative constructions (the 'judge', the 'malpractice'). Any empirical research in the field of law requires – beyond the general social distinctions such as gender and age – these sets of normative concepts. While it seems, that these concepts are much more stable than social ones, as they defined in legal sources, this is not entirely true. They are, very frequently also fiercely debated, – like many constitutional concepts recently in Hungary: including the concept of 'rule of law' itself

– and they are also subject to regulatory and interpretational changes too. This result, that any BD research in law will be the subject of not only general interpretational doubts, well-known from social science, but will be debated by the 'traditional' doctrinal scholars as well, using 'traditional' conceptual arguments. Second, empirical legal research should very often (though not always) rely on, process, control and interpret *texts*. This is true for 'normal' empirical research (such as research on the attitudes of judges), but it deeply pervades empirical research projects that are based on text analysis. These 'text-empirical research projects' can be subdivided into two types: projects that are based on human reading and coding (See for example, the famous Supreme Court (Spaeth) database,[9] or the European Conreason project[10] and projects based on machine processing (see e.g. Fowler et al., 2007; Ződi, 2015). In the case of the manually coded research, it is quite clear that coding sometimes requires interpretation and partially arbitrary decisions, but even in the case of a machine-made analysis during the construction of the text-analysing algorithms, one must make certain decisions which can distort the *data collection itself*.

These two peculiarities mean that BD-based research projects in law will not supersede doctrinal efforts; rather, they will rely on them. Doctrinal scholarship will provide the theoretical framework, the concepts and the distinctions that will serve as a basis for the higher narratives on which empirical and BD projects can build. But eventually, there will be a reverse process as well. BD research projects can offer new insights and ideas for which doctrinal scholarship can begin to build new theories and narratives.

5. Conclusions

Big Data already has a severe impact on law, and raises serious dilemmas. While Big Data is often mentioned as the basis for a new economic order, it is increasing risks, (mainly on privacy and anti-discrimination fields), which are different in character than 'normal' privacy issues (Mayer-Scönberger and Cukier, 2013). This new scenario means the 1. unmanageable volume, velocity, and movement of the (personal) data, that we and our devices generate, 2. the predictive power of the data which is resulting in an unbalanced relationship between private persons and those having access to the data and 3. the secondary use of the data, i.e. where the aggregated data is used for purposes that are far from their original ones. Keeping or tightening the existing (data protection and other) standards does not seem to be working, because this deprives society and the economy of Big Data's value-creation power. New rules are needed soon, based on new ethical principles.

BD offers possibilities in law making, lawyering and legal science. Experimental law-making, predictive lawyering and policing, legal enforcement based on data, and in-depth analysis of cases, fields of law, judges and courts will soon become parts of legal practice and will play an increasingly important role in the coming years. This does not mean that the political element in law-making or the moral judgement in legal decision making will disappear. It is rather that the foundation and the reasoning

[9] http://scdb.wustl.edu/
[10] http://www.conreasonproject.com/

structure of law making, lawyering and legal science will start slowly shifting to the direction where arguments based on Big Data will be accepted, and used more and more. Traditional legal science will also stay with us. But it will be controlled and complemented with the insights of (open) data-driven research.

References

Act CXXX of 2010 of Hungary (2010) Act CXXX of 2010 on the Law-making.

Ahlberg, C. (2011) The News Forecast: Can You Predict the Future by Mining Millions of Web Pages for Data? *Wired Magazine*, November 10, http://www.wired.co.uk/article/the-news-forecast Accessed: 20-01-2017.

Big Data and the Law Blog (2014) The Federal Trade Commission Wants In On Big Data Regulation. https://bigdataandthelaw.com/2014/10/02/the-federal-trade-commission-wants-in-on-big-data-regulation/ Accessed: 20-01-2017.

Bodig, M. (2015) Legal Doctrinal Scholarship and Interdisciplinary Engagement. *Erasmus Law Review*, 2: 43–54. DOI: https://doi.org/10.5553/elr.000035

Brugger, W. (2000) May Government Ever Use Torture? Two Responses from German Law. *American Journal of Comparative Law*, 48(4): 661–678. DOI: https://doi.org/10.2307/840910

Byers, A. (2015) Big Data, Big Economic Impact. *I/S: A Journal of Law and Policy for the Information Society*, 10(3): 757–764.

Commission WP (2012) Commission Staff Working Paper. Impact Assessment Accompanying the document [...] General Data Protection Regulation and Directive [...] on the protection of [...] processing of personal data etc. http://ec.europa.eu/justice/data-protection/document/review2012/sec_2012_72_en.pdf Accessed: 20-01-2017.

Communication from the Commission (2014) Towards a thriving data-driven economy (COM/2014/0442 final). http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1404888011738&uri=CELEX:52014DC0442 Accessed: 20-01-2017.

Crawford, K. and J. Schultz (2014) Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, 55(1): 93-128.

Deardorff, J. (2016) Big Data to Transform Social Science Research. Huge Amounts of Data Have the Potential to Change Long-standing Paradigms. *Northwestern University News Center*, May 23, http://www.northwestern.edu/newscenter/archives/special/data-science/day-3.html Accessed: 20-01-2017.

DeVan, A. (2016*)* The 7 V's of Big Data. *Impact Radius Blog*, April 7, https://www.impactradius.com/blog/7-vs-big-data/ Accessed: 28-06-2017.

Dholakia, U. M. (2015) Everyone Hates Uber's Surge Pricing – Here's How to Fix It. *Harvard Business Review*, December 21, https://hbr.org/2015/12/everyone-hates-ubers-surge-pricing-heres-how-to-fix-it Accessed: 20-01-2017.

Duhigg, C. (2012) How Companies Learn Your Secrets. *New York Times,* February 19, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp Accessed: 20-01-2017.

Dyevre, A. (2016) *The Future of Legal Theory and the Law School of the Future.* Cambridge: Intersentia.

Eaton, C., DeRoos, D., Deutsch, T., Lapis, G. and P. Zikopoulos (2012) *Understanding Big Data; Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw Hill.

Economist (2010) *The Data Deluge. The Economist Special Report,* February 27.

EU Directive (1995/46/EC) of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of Such Data.

European Commission Press Release (2012) *Commission Proposes a Comprehensive Reform of Data Protection Rules to Increase Users' Control of their Data and to Cut Costs for Businesses.* January 25, http://europa.eu/rapid/press-release_IP-12-46_en.htm?locale=en Accessed: 20-01-2017.

General Data Protection Regulation (2016) Regulation of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC.

Ferguson, A. G. (2015) Big Data and Predictive Reasonable Suspicion. *University of Pennsylvania Law Review,* 163(2): 327–410. DOI: https://doi.org/10.2139/ssrn.2394683

Fowler, J. H., Johnson, T. R., Spriggs II, J. F., Jeon, S. and P. J. Wahlbeck (2007) Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis,* 15(3): 324–346. DOI: https://doi.org/10.1093/pan/mpm011

FTC Conference (2014) Event Description. *Big Data: A Tool for Inclusion or Exclusion?* September 15, Washington, D.C. https://www.ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion Accessed: 30-01-2017.

H2020 Call (2016) *Policy-development in the Age of Big Data: Data-driven Policy-making, Policy-modelling and Policy-implementation.* https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/co-creation-06-2017.html Accessed: 20-01-2017.

Hammond, K. (2015) *The End of Big Data: AI and the Rise of the Narrative. DataInformed.com*, March 5, http://data-informed.com/the-end-of-big-data-ai-and-the-rise-of-the-narrative/ Accessed: 28-06-2017.

Hoecke, M. (2011) Which Method(s) for What Kind of Discipline? In Hoecke, M. (ed.) *Methodologies of Legal Research.* Oxford: Hart. 1–18. DOI: http://dx.doi.org/10.5040/9781472560896.ch-001

IBM Data Hub (2016) The Four V's of Big Data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data Accessed: 28-06-2017.

Inbar, Y., Pizarro, D., Iyer, R. and J. Haidt (2011) Disgust Sensitivity, Political Conservatism, and Voting. *Social Psychological and Personality Science,* 3(5): 537–544. DOI: https://doi.org/10.1177/1948550611429024

Jeon, J-H and S-R Jeong (2016) Designing a Crime-Prevention System by Converging Big Data and IoT. *Journal of Internet Computing and Services,* 17(3): 115–128. DOI: https://doi.org/10.7472/jksii.2016.17.3.115

Kunz, J. R. and M. P. Majairan (2016) Racial Disparities in Sentencing. *Delaware Lawyer,* 34(1): 18–21.

Mattioli, M. (2014) Disclosing Big Data. *Minnesota Law Review*, 99(2): 535–584.

Mayer-schönberger, V. and K. Cukier (2013) *Big Data; A Revolution that will Transform How We Live, Work and Think*. Boston: Houghton Mifflin Harcourt.

McGregor, V., S. H. Calderon and R. Tonelli (2013) Big Data and Consumer Financial Information. *Business Law Today*. 2013/11. http://www.americanbar.org/publications/blt/2013/11/04_mcgregor.html Accessed: 30-01-2017.

NFS (2015) *Big Data Hubs NFS Program Solicitation.* https://www.nsf.gov/pubs/2015/nsf15562/nsf15562.htm Accessed: 28-06-2017.

Nicol, A., Caruso, J. and É. Archambault (2013) Open Data Access Policies and Strategies in the European Research Area and Beyond. *Science Metrix*, 2013 August, http://www.science-metrix.com/pdf/SM_EC_OA_Data.pdf Accessed: 28-06-2017.

Ohm, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review,* 57(6): 1701–1778.

Open Data Maturity in Europe (2016) European Commission Directorate General for Communications Networks, Content and Technology Unit G.1 Data Policy and Innovation. https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n2_2016.pdf Accessed: 28-06-2017.

Richards, N. M. and J. H. King (2014) Big Data Ethics. *Wake Forest Law Review*, 49(2): 393–432.

Siegel, E. (2013) *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* New Jersey: Wiley. DOI: https://doi.org/10.1002/9781119172536

Simon, P. (2014) Big Data Lessons from Netflix. *Wired Magazine,* 2014 March, https://www.wired.com/insights/2014/03/big-data-lessons-netflix/ Accessed: 30-01-2017.

Tane, O. and J. Polonetzky (2012) Privacy in the Age of Big Data: A Time for Big Decisions. *Stanford Law Review Online,* 64: 63-69.

Volkov, V. (2016) Legal and Extralegal Origins of Sentencing Disparities: Evidence from Russia's Criminal Courts. *Journal of Empirical Legal Studies,* 13(4): 637–665. DOI: https://doi.org/10.1111/jels.12128

White House (2012) *Big Data Research and Development Initiative - Big Data is a Big Deal.* https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal Accessed: 19-06-2017.

White House (2014) *Big Data: Seizing Opportunities, Preserving Values.* https://obamawhitehouse.archives.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf Accessed: 19-06-2017.

White House (2015a) *Big Data Hubs White House Blogpost.* https://obamawhitehouse.archives.gov/blog/2015/11/04/big-announcements-big-data Accessed: 19-06-2017.

White House (2015b) *Big Data and Differential Pricing.* The White House Council of Economic Advisers. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf Accessed: 19-06-2017.

White House (2016) *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.* Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf Accessed: 19-06-2017.

Williford, C. and C. Henry (2012) *One Culture: A Report on the Experiences of First Respondents to the Digging into Data Challenge Computationally Intensive Research in the Humanities and Social Sciences.* Council on Library and Information Resources, Washington, D.C. https://www.clir.org/pubs/reports/pub151/pub151.pdf Accessed: 20-01-2017

Wingerden, S., van Wilsem, J. and B. D. Johnson (2016) Offender's Personal Circumstances and Punishment: Toward a More Refined Model for the Explanation of Sentencing Disparities. *Justice Quarterly,* 33(1): 100–133. DOI: https://doi.org/10.1080/07418825.2014.902091

Ződi, Z. (2015) Citations of Previous Decisions, and the Quality of Judicial Reasoning. *Acta Juridica Hungarica: Hungarian Journal of Legal Studies,* 56(2–3): 129–148. DOI: https://doi.org/10.1556/026.2015.56.2-3.3