Vanderbilt University Law School

Legal Studies Research Paper Series
19-36

Exploring the Interfaces Between Big Data and Intellectual Property Law

10 (2019) Journal of Intellectual Property, Information Technology and Electronic Commerce Law (JIPITEC) 22

Professor Daniel J. Gervais
Vanderbilt University School of Law

Exploring the Interfaces Between Big Data and Intellectual Property Law

by **Daniel Gervais***

Abstract: This article reviews the application of several IP rights (copyright, patent, sui generis database right, data exclusivity and trade secret) to Big Data. Beyond the protection of software used to collect and process Big Data corpora, copyright's traditional role is challenged by the relatively unstructured nature of the non-relational (noSQL) databases typical of Big Data corpora. This also impacts the application of the EU sui generis right in databases. Misappropriation (tort-based) or anti-parasitic behaviour protection might apply, where available, to data generated by AI systems that has high but short-lived value. Copyright in material contained in Big Data corpora must also be considered. Exceptions for Text and Data Mining (TDM) are already in place in a number of legal systems and likely to emerge to allow the creation and use of corpora of literary and artistic works, such as texts and images. In the patent field, AI systems using Big Data corpora of patents and scientific literature can be used to expand patent applications. They can also be used to "guess" and disclose future incremental innovation. These developments pose serious doctrinal and normative challenges to the patent system and the incentives it creates in a number of areas, though data exclusivity regimes can fill certain gaps in patent protection for pharmaceutical and chemical products. Finally, trade secret law, in combination with contracts and technological protection measures, can protect data corpora and sets of correlations and insights generated by AI systems.

Keywords: Copyright; patent; data exclusivity; artificial intelligence; big data; trade secret

Recommended citation: Daniel Gervais, Exploring the Interfaces Between Big Data and Intellectual Property Law, 10 (2019) JIPITEC 22 para 1

A. Introduction

1 The interfaces between "Big Data" (as the term is defined below) and IP matters both because of the impact of Intellectual Property (IP) rights *in* Big Data, and because IP rights might interfere with the generation, analysis and use *of* Big Data. This Article looks at both sides of the interface coin, focusing on several IP rights, namely copyright, patent, data exclusivity and trade secret/confidential information.[1] The paper does not discuss trade marks in any detail, although the potential role of Artificial Intelligence (AI), using Big Data corpora,[2] in designing and selecting trade marks certainly seems a topic worthy of further discussion.[3]

B. Defining Big Data

2 The term "Big Data" can be defined in a number of ways. A common way to define it is to enumerate its three essential features, a fourth that, though not essential, is increasingly typical, and a fifth that is derived from the other three (or four). Those features are: volume, veracity, velocity, variety, and value.[4] "Volume" or size is, as the term Big Data suggests, the first characteristic that distinguishes Big Data from other ("small data") datasets. Because Big Data corpora are often generated automatically, the question of the quality or trustworthiness of the data ("veracity") is crucial. "Velocity" refers to "the speed at which corpora of data are being generated, collected and analyzed".[5] The term "variety" denotes the many types of data and data sources from which data can be collected, including Internet browsers, social media sites and apps, cameras, cars, and a host of other data-collection tools.[6] Finally, if all previous features are present, a Big Data corpus likely has significant "value".

3 The way in which "Big Data" is generated and used can be separated into two phases.[7]

4 First, the creation of a Big Data corpus requires processes to collect data from sources such as those mentioned in the previous paragraph. Second, the corpus is analysed, a process that may involve Text and Data Mining (TDM).[8] TDM is a process that uses an Artificial Intelligence (AI) algorithm. It allows the machine to learn from the corpus—hence the term "machine learning" (ML) is sometimes used as a synonym of AI in the press.[9] As it analyses a Big Data corpus, the machine *learns and gets better at what it does.* This process often requires human input to assist the machine in correcting errors or faulty correlations derived from, or decisions based on, the data.[10] This processing of corpora of Big Data is done to find correlations and generate predictions or other valuable analytical outcomes. These correlations and

[*] Dr. Gervais is Professor of Information Law at the University of Amsterdam and the Milton R. Underwood Chair in Law at Vanderbilt University. The author is grateful to Drs. Balász Bodó, João Quintais, and to Svetlana Yakovleva of the Institute for Information Law (IvIR), to participants at the University of Lucerne conference on Big Data and Trade Law (November 2018), to Ole-Andreas Rognstad and other participants at the Data as a Commodity workshop at the University of Oslo (December 2018), and to the anonymous reviewers at JIPITEC for most useful comments on earlier versions of this Article.

[1] The Article considers IP rights applied by all or almost all countries, namely those contained in the Agreement on Trade-related Aspects of Intellectual Property Rights, Annex 1C of the Agreement Establishing the World Trade Organization, 15 April 1994. As of January 2019, it applied to the 164 members of the WTO, including all EU member States and the EU itself.

[2] This use of the term "corpus" in this context is an extension of its original meaning as either a "body or complete collection of writings or the like; the whole body of literature on any subject", or the "body of written or spoken material upon which a linguistic analysis is based". Oxford English Dictionary Online (accessed 21 December 2018).

There is a debate about the proper form of the plural. Both Oxford and Merriam-Webster indicate that "corpora" is the proper form, although the author has encountered the form "corpuses" in the literature discussing Big Data. See e.g., the 2014 White House report to the President from the President's Council of Advisors on Science and Technology titled "Big Data and Privacy: A Technological Perspective", at x. "Corpora" is the form chosen here, although the predicable future is that the perhaps more intuitive form "corpuses" will win this linguistic tug-of-war.

[3] For example, AI systems can create correlations between trademark features (look, sound etc.) and their appeal, thus allowing the creation and selection of "better" marks.

[4] Jenn Cano, 'The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity', XSNet (March 11, 2014), <https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity> (accessed 10 December 2018).

[5] Ibid.

[6] The list includes "cars" as cars as personal vehicles are one of the main sources of (personal) data—up to 25 Gigabytes per hour of driving. The data are fed back to the manufacturer. See Uwe Rattay, 'Untersuchung an vier Fahrzeugen - Welche Daten erzeugt ein modernes Auto?', ADAC, <https://www.adac.de/infotestrat/technik-und-zubehoer/fahrerassistenzsysteme/daten_im_auto/default.aspx> (accessed 11 December 2018).

[7] The two components are not necessarily sequential. They can and often do proceed in parallel.

[8] See Maria Lillà Montagnani, 'Il text and data mining e il diritto d'autore' (2017) 26 AIDA 376.

[9] Cassie Kozyrkov, 'Are you using the term 'AI' incorrectly?', Hackernoon (26 May 2018), <https://hackernoon.com/are-you-using-the-term-ai-incorrectly-911ac23ab4f5>.

[10] How IP will apply to the work involved in the human training function of machine learning is one of the interesting questions at the interface of Big Data and IP. The term "training data" is used in this context to suggest that the machine training is supervised (by humans). See Brian D Ripley, *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press, 1996) 354.

insights can be used for multiple purposes, including advertising targeting and surveillance, though an almost endless array of other applications is possible. To take just one different example of a lesser known application, a law firm might process hundreds or thousands of documents in a given field, couple ML with human expertise, and produce insights about how they and other firms operate, for example in negotiating a certain type of transaction or settling (or not) cases.

5 A subset of machine learning known as *deep learning* (DL) uses neural network, a computer system modelled on the human brain.[11] This implies that any human contribution to the output of deep learning systems is "second degree". When considering the possible IP protection of outputs of such systems, this separation between humans and the output challenges core notions of IP law, especially authorship in copyright law and inventorship in patent law.

C. Framing the issues

6 ML and DL can produce high value outputs. Such outputs can take the form of analyses, insights, correlations, and may lead to automated (machine) decision-making. It can be expected that those who generate this value will try to capture and protect it, using IP law, technological measures and contracts. One can also expect competitors and the public to try to access those outputs for the same reason, namely their value.

7 How far should IP go to protect value generated by ML? The old adage that "if it is worth copying it is worth protecting" has long been discarded.[12] A more nuanced question to ask might be, do entities that collect, process and use Big Data *need IP incentives* or *deserve additional rewards* to do what they do. Is protecting Big Data corpora and their processing outputs comparable to providing an incentive for trees to grow leaves in the spring? Specifically, does the creation of incentives help generate *new or better* data corpora, analyses, and thus produce welfare increases, taking account of welfare losses that rights *in* Big Data might cause, such as increased transaction and licensing costs?

8 In many cases, Big Data corpora are protected by secrecy, a form of protection that relies on trade secret law combined with technological protection from hacking, and contracts. Deciding which IP rights may apply should thus distinguish Big Data corpora that are not publicly accessible (say the Google databases powering its search engine and adverting) and those that are. A secret corpus is often de facto protected against competitors due to its secrecy, meaning that competitors may need to generate a competitive corpus to capture market share.[13] A publicly available corpus, in contrast, must rely on erga omnes IP protection—if it deserves protection to begin with. Copyright protects collections of data; the sui generis database right (in the EU) might apply; and data exclusivity rights in clinical trial data may be relevant. All three are topics explored below.

9 The *outputs* of the processing of Big Data corpora may contain or consist of subject matter that facially could be protected by copyright or patent law. Big Data technology can be—and in fact is—used to create and invent. For example, a Big Data corpus of all recent pop music can find correlations and identify what may be causing a song to be popular. It can use the correlations to write its own music.[14]

10 The creation of (potentially massive amounts of) new literary and artistic material without direct human input will challenge human-created works in the marketplace. This is already happening with machine-written news reports.[15] Deciding whether machine-created material should be protected by copyright could thus have a profound impact on the market for creative works. If machine created material is copyright-free, machines will produce free goods that compete with paid ones, that is, those created by humans expecting a financial return. If the material produced by machines is protected by copyright and its use potentially subject to payment, this might level the commercial playing field between human and machine, but then who (which natural or legal person) *should* be paid for the computer's work? Then there will be border definition issues. Some works will be created by human and machine working together. Can we apply the notion of joint authorship? Or should we consider the machine-produced portion (if separable) copyright-free, thus limiting the protection to identifiably human-authored portions?

[11] With "deep learning model, the algorithms can determine on their own if a prediction is accurate or not… through its own method of computing – its own 'brain', if you will" Brett Grossfeld, 'A simple way to understand machine learning vs deep learning', ZenDesk (18 July 2017), online: <https://www.zendesk.com/blog/machine-learning-and-deep-learning/>.

[12] *University of London Press v University of London Tutorial Press*, [1916] 2 Ch. 601 at 610.

[13] Thanks to Prof. Bernt Hugenholtz (Univ, of Amsterdam) for discussing this insight with me.

[14] See Gaëtan Hadjeres & François Pachet, 'DeepBach: A steerable model for Bach chorales generation' (3 December 2016) 1, online: <https://arxiv.org/pdf/1612.01010v1.pdf>.

[15] See Corinna Underwood, 'Automated Journalism – AI Applications at New York Times, Reuters, and other mediants', eMerj (22 June 2018, updated 29 November 2018), online: <https://bit.ly/2Q84BTV>.

11 If such major doctrinal challenges—each with embedded layers of normative inquiries—emerge in the field of copyright, Big Data poses existential threats in the case of patents. AI tools can be used to process thousands of published patents and patent applications and used to *expand the scope of claims in patent applications*. This poses normative challenges that parallel those enunciated above: who is the inventor? Is there a justification to grant an exclusive right to a machine-made invention? To whom? Then there are doctrinal ones as well. For example, is the machine-generated "invention" disclosed in such a way that would warrant the issuance of a patent?

12 It gets more complicated, however. If AI machines using patent-related Big Data can broaden claim scope or add claims in patent applications, then within a short horizon they could be able to *predict the next incremental steps in a given field of activity* by analysing innovation trajectories. For example, they might look at the path of development of a specific item (car brakes, toothbrushes) and "predict" or define a broad array or what *could* come next. Doctrinally, this raises questions about inventive step: If a future development is obvious to a machine, is it obvious for purposes of patent law? Answering this question poses an epistemological as well as a doctrinal challenge for patent offices. The related normative inquiry is the one mentioned above, namely whether machine-made inventions (even inventions the scope [claims] of which were merely "stretched" using Big Data and AI) "deserve" a patent despite their obviousness (to the machine).

13 This use of patent and technological Big Data could lead to a future where machines pre-disclose incremental innovations (and their use) in such a way that they constitute publicly available prior art and thus make obtaining patents impossible on a significant part of the current patentability universe. Perhaps even the best AI system using a Big Data corpus of all published patents and technical literature will not be able to predict the next pioneer invention, but very few patents are granted on ground-breaking advances. AI systems that can predict *most* currently patented inventions, (which tend to be only incrementally different from the prior art) would wreak havoc with the patent-based incentive system.[16]

14 Let us take an example. It is possible that deep-learning algorithms could parse thousands of new molecules based on those recently patented or disclosed (in applications) and even predict their medical efficacy. If such data (new molecules and predicted efficacy) were available and published, it would significantly hamper the patentability of those new molecules due to lack of novelty.

15 The unavailability of patents would dramatically increase the role of data exclusivity rights (the right to prevent reliance in clinical data submitted to obtain marketing approval) in the pharmaceutical field.[17] If this prediction of future inventions by AI became an established practice in fields where this (separate) protection by data exclusivity is unavailable, the very existence of the incentive system based on patents could be in jeopardy.

16 In the pages that follow, the Article takes a deeper look at each of these challenges and draws the contours of possible answers.

D. Copyright

17 Let us get an easy point out of the analytical picture at the outset: the human-written (AI) software used to collect (including search and social media apps), store and analyse Big Data corpora is considered a literary work eligible for copyright protection, subject to possible exclusions and limitations.[18] The analysis that follows focuses on the harder question of the protection of the *Big Data corpora* and of the *outputs* generated from the processing of such corpora.

18 Before we delve more deeply into the interface between Big Data and copyright, it is necessary briefly to review briefly a fundamental element of copyright law, namely originality.

I. The Key Role of Originality

19 The main international instrument in the field of copyright is the Berne Convention, to which 176 countries were party as of January 2019.[19] That

[16] See Shlomit Yanisky Ravid & Xiaoqiong (Jackie) Liu, 'When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3a Era' (2018) 39 Cardozo L Rev 2215, 2254; and Ted Baker, 'Pioneers in Technology: A Proposed System for Classifying and Rewarding Extraordinary Inventions' (2003) 45 Arizona L Rev 445.

[17] See Daniel Gervais, 'The Patent Option' (*forthcoming*, North Carolina J. L. & Tech), available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266580> (accessed 15 December 2018).

[18] This is recognized for example in Article 10.1 of the TRIPS Agreement (note 1 above), which provides that "[c]omputer programs, whether in source or object code, shall be protected as literary works under the Berne Convention (1971)".

[19] Berne Convention for the Protection of Literary and Artistic Works, of 9 September 1886, last revised at Paris on 24 July 1971, and amended on September 28, 1979 [hereinafter Berne Convention]. On membership of the Berne Union (countries party to the Convention), see <http://www.wipo.

Convention protects "literary and artistic works", a term that the Convention only defines by providing a list of categories of "productions" (another undefined term) that fit into the literary and artistic categories.[20]

20 There is more to this story, however. A Committee of Experts meeting under the auspices of the World Intellectual Property Organization (WIPO), which administers the Berne Convention, concluded that, although this is not specified expressly in the text of the Convention, the only mandatory requirement for a literary or artistic work to be protected by the Convention is that it must be "original". To arrive at this conclusion, the Committee considered both the Convention's drafting history and the use of the expression "intellectual creation" in the Convention as a functional synonym of the term "work".[21] This also means that *no mandatory formality* may be required to obtain copyright protection.[22] The same statement, namely that the only applicable criterion is originality, can be made about EU law.[23]

21 The Convention contains important hints as to what constitutes an "original" work. In its Article 2, when discussing the protection of "collections", it states that "[c]ollections of literary or artistic works such as encyclopaedias and anthologies which, by reason of the *selection and arrangement of their contents*, *constitute intellectual creations* shall be protected *as such*, without prejudice to the copyright in each of the works forming part of such collections."[24] Selection and arrangement are exemplars of what copyright scholars refer to as "creative choices".[25] Creative choices need not be artistic or aesthetic in nature, but it seems they do have to be human.[26] Relevant choices are reflected in the particular way an author describes, explains, illustrates, or embodies his or her creative contribution. In contrast, choices that are merely routine (e.g. the choice to organize a directory in alphabetical order) or significantly constrained by external factors such as the function a work is intended to serve (e.g. providing accurate driving directions), the tools used to produce it (e.g. a sculptor's marble and chisel), and the practices or conventions standard to a particular type of work (e.g. the structure of a sonnet) are not creative for the purpose of determining the existence of a sufficient degree of originality.

22 When the Berne Convention text was last revised on substance in 1967,[27] neither publicly available "electronic" databases nor any mass-market database software was available. The "collections" referred to in the Convention are thus of the type mentioned by the Convention drafters: (paper-based) anthologies and encyclopaedias. The negotiators' objective was to create a separate copyright for the maker (or "arranger") of a collection, knowing that most if not all of the entries in the collection (say, an encyclopaedia) were written by third parties, each an expert in her or his own field and each entitled to his or her own copyright in the entry. In a collection of this type, there are thus two layers of copyright; first, a right in each entry, and in each illustration or photograph, which is either transferred or licensed to the maker or publisher of the collection; and, second, a copyright in what one might call the "organizational layer", granted to the maker of the collection based on the "selection or arrangement" of the individual entries, photographs and illustrations. The second layer—the collection such as encyclopaedia—is generally treated as a collective work.[28]

int/treaties/en/ShowResults.jsp?lang=en&treaty_id=15>.

[20] Ibid art. 2. The term "production" seems to refer to the fact that a work must be objectified to be protected, that is, a work is not a work if it only exists in the mind of an author. See Ivan Cherpillod, *L'Objet du Droit d'Auteur* (Centre du Droit de l'Entreprise de l'Université de Lausanne, 1985) 35-41.

[21] See WIPO Committee of Experts on Model Provisions for Legislation in the Field of Copyright, First Session, document CE/MPC/I/3, of March 3, 1989, at 16; and Memorandum prepared for the WIPO Committee of Experts on Model Provisions for Legislation in the Field of Copyright, document CE/MPC/I/2-III of Oct. 20, 1988, at 10.

[22] See Jane C. Ginsburg, '"With Untired Spirits and Formal Constancy": Berne Compatibility of Formal Declaratory Measures to Enhance Copyright Title-Searching' (2013) 28:3 Berkeley Technology LJ 1584-1622. Countries are allowed to impose a second requirement, namely fixation. Berne Convention, art. 2(2).

[23] Football Dataco, CJEU 1 March 2012, C-604/10, para. 40. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [hereinafter "Database Directive"]. Recital 16 of the Database Directive, for example, notes "no criterion other than originality in the sense of the author's intellectual creation should be applied to determine the eligibility of the database for copyright protection, and in particular no aesthetic or qualitative criteria should be applied". See also Daniel Gervais and Estelle Derclaye, 'The Scope of Computer Program Protection after SAS: Are We Closer to Answers?' (2012) 34:8 EIPR 565

[24] Berne Convention (n 11) art. 2(5) (emphasis added).

[25] See Daniel Gervais and Elizabeth Judge, 'Of Silos and Constellations: Comparing Notions of Originality in Copyright Law' (2009) 27:2 Cardozo Arts & Entertainment LJ 375.

[26] Deciding whether Big Data corpora are protectable in the absence of an identifiable human author would be the subject of a separate analysis, well beyond the scope of this paper. Suffice it to say that views differ. Contrast s. 9(3) and 178 of the CDPA with this statement from the United States Copyright Office: "Examples of situations where the Office will refuse to register a claim include: [...] The work lacks human authorship". United States Copyright Office, Compendium of U.S. Copyright Office Practices, (3rd edition, 2017) 22.

[27] An Appendix for developing countries was added in Paris in 1971 but it did not modify the definition of "work".

[28] For example, section 101 of the US Copyright Act (Title 17

II. Application to Big Data

23 When "electronic" databases started to emerge in the 1990s, data generally had to be indexed and re-indexed regularly to be useable. The TRIPS Agreement (signed in 1994 but essentially drafted in the late 1980s up to December 1990), is a reflection of this development.[29] Using language meant to parallel art. 2(5) of the Berne Convention, it states that:

> *Compilations of data or other material, whether in machine-readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.*[30]

24 The data in typical (relational or "SQL") databases in existence in the 1990s generally was "structured" in some way, for example via an index, and that structure might qualify the database for (thin) copyright protection in the database's organizational layer. Older databases also contained more limited datasets ("small data").

25 Facebook, Google, and Amazon, to name just those three, found out early on that relational databases were not a good solution for the volumes and types of data that they were dealing with. This inadequacy explains the development of open source software (OSS) for Big Data: the Hadoop file system, the MapReduce programming language, and associated non-relational ("noSQL") databases such as Apache's Cassandra.[31] These tools and the corpora they helped create and use may not qualify for protection as "databases" under the SQL-derived criteria mentioned above. This does not mean that no work or knowhow is required to create the corpus, but that the type of structure of the dataset may not qualify. As the CJEU explained in *Football Dataco*,

> *[S]ignificant labour and skill of its author, as mentioned in section (c) of that same question, cannot as such justify the protection of it by copyright under Directive 96/9, if that labour and that skill do not express any originality in the selection or arrangement of that data.*[32]

26 Indeed, Big Data is sometimes defined in *direct contrast* to the notion of SQL database and reflected in the TRIPS Agreement (and the EU database directive discussed in the next section). A McKinsey report, for example, notes that "Big Data" referred to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse."[33] Those data are often generated automatically but at times less so, as when Google scanned millions of books for its massive book scanning project.[34] This was a most ambitious project but copyright "got in the way", especially for access to the corpus outside the United States:

> *Google's idea was to digitize as many published works as possible in as many languages as possible for the purpose of creating a universal digital library made up all printed books from every culture. The problem is that books are intellectual property, and intellectual property laws, cultures, and practices are not uniform around the world.*[35]

27 Big Data software is unlikely to "select or arrange" the data in a way that would meet the originality criterion and trigger copyright protection. In the *Google Books* case, the database basically consists of word-searchable scans of the books. From a copyright standpoint, therefore, it is doubtful whether a Big Data corpus of this sort, or a "dump" of, say, personal data scraped from online search engines or social media sites, would benefit from copyright protection. Hacking and other methods of unauthorised access to such corpora might be better handled via computer crimes and torts.

28 An argument has been made that tables or other outputs (such as analysis results generated by a TDM system) can be protected by copyright. An example

of the United States Code) defines "collective work" as "a work, such as a periodical issue, anthology, or encyclopedia, in which a number of contributions, constituting separate and independent works in themselves, are assembled into a collective whole".

[29] For a longer description of the negotiating history, see Daniel Gervais, *The TRIPS Agreement: Drafting History and Analysis* (4th ed) (Sweet & Maxwell, 2013).

[30] TRIPS Agreement (n 1) art. 10.2 (emphasis added). A difference between Berne and TRIPS that need not be belaboured here but is worth noticing is the different conjunction used between "selection" and "arrangement". Emphasis added. See also s. 3A of the Copyright, Designs and Patents Act 1988 (CDPA).

[31] See April Reeve, 'Big Data and NoSQL: The Problem with Relational Databases' (7 September 2012), available at <https://infocus.dellemc.com/april_reeve/big-data-and-nosql-the-problem-with-relational-databases/> (accessed 18 November 2018). It is worth noting that it is because code is protected by copyright (see TRIPS Agreement, art. 10(1)), that owners of code can license it and impose open source terms.

[32] Football Dataco (n 15) para 42.

[33] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, Big Data: The next frontier for innovation, competition, and productivity, at 1, (McKinsey, 2011), available at <https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx>.

[34] See books.google.com. See also Daniel Gervais, 'The Google Book Settlement and the TRIPS Agreement' [2011] Stanford Tech LR 1.

[35] Lyombe Eko, Anup Kumar, Qingjiang Yao, 'To Google or Not to Google: The Google Digital Books Initiative and the Exceptionalist Intellectual Property Law Regimes of the United States and France' (2012) 15 J Internet L 12, 13–14.

often mentioned in this context is the controversial car valuation database case concerning the catalogue of used car prices known as the *RedBook* in the United States.[36] The US Court of Appeals for the Second Circuit found that a collection of prices of used cars based on an algorithm factoring in age, mileage, model, etc. could benefit from protection.[37] The court's opinion "seems quite artificial and not directed to preserving the creativity and ingenuity inherent in any view of creative authorship."[38] It obscures the principle that ideas are not protected by copyright, an internationally recognized principle.[39] Moreover, even if that case is still good law, the question whether machine-created productions can qualify as copyright works is either still open, or resolved in favour of a need for human authorship.[40]

29 An interesting argument has been put forward by Harvard law professor Ruth Okediji for a different role for copyright in this context. She asserts that governments could claim protection of data-driven innovation to allow them to "develop appropriate conditions that ensure that more members of the public have access to any new works created."[41] The purpose would be to ensure that "free or heavily subsidized access to Big Data is available to the broader public at marginal cost or not much more."[42] This idea resembles the General Public License (GPL) model, which uses copyright licenses to maintain the "open" nature of computer code based on previous open source software.[43] Indeed, OSS has been critical in shaping the technology that supports Big Data.[44]

30 Finally, it is worth noting that, in some jurisdictions, even absent copyright protection for Big Data, other IP-like remedies might be relevant, such as the tort of misappropriation applicable to "hot news" in US law, or the protection against parasitic behaviour available in a number of European systems.[45] This might apply to information generated by AI-based TDM systems that have initially high but fast declining value, such as financial information relevant to stock market transactions, as data "has a limited lifespan--old data is not nearly as valuable as new data--and the value of data lessens considerably over time".[46]

III. The Sui Generis Database Right

31 In EU law, there is also a *sui generis* right in databases.[47] This right is not subject to the originality requirement.[48] The Directive refers to the database maker's investment in "obtaining, verification or presentation of the contents" and then provides a right "to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database."[49] The directive also mentions in its recitals that a database includes "collections of independent works, data or other materials which are *systematically or methodically arranged* and can be individually accessed."[50] This, according to Professor Bernt Hugenholtz, "squarely rules out protection –

[36] See eg Peter DiCola et al., 'Legal Problems in Data Management: IT & Privacy at the Forefront: "Big Data": Ownership, Copyright, and Protection' [2015] John Marshall J. Information Technology & Privacy L, 565, at 576.

[37] CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, 44 F.3d 61 (2d Cir.1994).

[38] C.D. Freedman, 'Should Canada Enact A New Sui Generis Database Right?' (2002) 13 Fordham Intell. Prop. Media & Ent. LJ 35, 85.

[39] See TRIPS Agreement, art. 9(2).

[40] The US Copyright Office, for example takes that view, See United States Copyright Office. Compendium of U.S. Copyright Office Practices, 3rd edition (2017) at 3-4. See Amir H. Khoury, 'Intellectual Property Rights for "Hubots": On the Legal Implications of Human-Like Robots as Innovators and Creators' (2017) 35 Cardozo Arts & Ent. LJ 635, 665. For an older but potentially still relevant article on the same topic, see Daniel Gervais, 'The Protection Under International Copyright Law of Works Created with or by Computers', (1991) 5 IIC Int'l Rev. Ind'l Prop. and Copyright Law, 629, 644-45. For a critique, see Shlomit Yanisky-Ravid, Luis Antonio Velez-Hernandez, 'Copyrightability of Artworks Produced by Creative Robots and Originality: The Formality-Objective Model' (2018) 19 Minn. J L Science & Tech. 1. A recent proposal suggests applying the work-made-for-hire doctrine for AI works so that the human operating the AI system would be the author under US law. See Shlomit Yanisky Ravid and Samuel Moorhead, 'Generating Rembrandt: Artificial Intelligence, Accountability and Copyright -The Human-Like Workers are Already Here - A New Model' [2017] Michigan State LR 659.

[41] Ruth L. Okediji, 'Government as Owner of Intellectual Property? Considerations for Public Welfare in the Era of Big Data' (2016) 18 Vanderbilt J Entertainment and Technology L 331, 361.

[42] Ibid.

[43] "The distributor of a GPL-licensed work, for example, must make the source code of that work available under the terms of the GPL". Eli Greenbaum, 'Open Source Semiconductor Core Licensing' (2011) 25 Harvard J L & Tech. 131, 139.

[44] David J. Kappos, 'Open Source Software and Standards Development Organizations: Symbiotic Functions in the Innovation Equation' (2017) 18 Columbia Science & Technology LR 259, 261. Mr, Kappos is the former head of the United States Patent and Trademark Office.

[45] See Victoria Smith Ekstrand and Christopher Roush, 'From "Hot News" to "Hot Data": The Rise of "FinTech", the Ownership of Big Data, and the Future of the Hot News Doctrine' (2017) 35 Cardozo Arts & Entertainment LJ 303.

[46] D. Daniel Sokol & Roisin Comerford, 'Antitrust and Regulating Big Data' (2016) 23 George Mason LR. 1129, 1138.

[47] Database Directive (n 19). See also Daniel Gervais, 'The Protection of Databases' (2007) 82:3 Chicago-Kent LR 1101.

[48] See P Bernt Hugenholtz, 'Intellectual Property and Information Law' in Jan J.C. Kabel and Gerard J.H.M. Mom (eds.), *Essays in Honour of Herman Cohen Jehoram* (The Hague/London/Boston: Kluwer Law International 1998) 183-200.

[49] Directive (n 22), art 7(1).

[50] Ibid, recital 7.

whether by copyright or by the sui generis right – of (collections of) raw machine-generated data."[51] The use of noSQL technologies may mean that Big Data corpora are not protected by the sui generis right. It also seems fair to say that the machine produced outputs (such as new data corpora) based on analyses of Big Data are neither "obtained" nor "collected"; they are generated by the machine. This would seem to leave them unprotected by the sui generis right.

32 The Database Directive also mentions, however, that if there is an *investment* in obtaining the data, that investment may be sufficient for the corpus to qualify as a database.[52] "Recitals 10-12 preceding the Directive illustrate that the principal reason for introducing the sui generis right was to promote investment in the (then emerging) European database sector".[53] If the directive were applied to Big Data corpora, then crawling through the data might constitute prohibited "extraction" unless it was minimal.[54]

33 While this matter cannot be fully investigated here, there are serious doubts about the power of this argument to justify the application of the directive to Big Data corpora. The Court of Justice of the European Union defined "investment" in obtaining the data as "resources used to seek out existing materials and collect them in the database but does not cover the resources used for the creation of materials which make up the contents of a database."[55] Professor Hugenholtz explained that "the main argument for this distinction, as is transparent from the decision, is that the Database Directive's economic rationale is to promote and reward investment in database production, not in generating new data".[56] This casts doubt on whether the notion of investment is sufficient to warrant sui generis protection of Big Data corpora, though Matthias Leistner suggested caution in opining that the "the sweeping conclusion that all sensor- or other machine-generated data will typically not be covered by the sui generis right is not warranted".[57]

34 Arguably, indirect confirmation that "Big Data" corpora are protected neither by copyright nor by the sui generis right in database may be found in a Commission staff document accompanying a 2017 Communication from the Commission in which the idea was floated to create a data producer's right.[58] The Staff document noted that

> *"[T]he Database Directive did not intend to create a new right in the data. The CJEU thus held that neither the copyright protection provided for by the Directive nor the sui generis right aim at protecting the content of databases. Furthermore, the ECJ has specified that the investment in the creation of data should not be taken into account when deciding whether a database can receive protection under the sui generis right".*[59]

35 The idea of creating a new exclusive right in data was conspicuously absent in an April 2018 document on the creation of a "European data space".[60]

IV. Exceptions and Limitations for Big Data TDM

1. The need for exceptions and limitations

36 TDM software used to process corpora of Big Data might infringe rights in databases that are protected either by copyright or the EU sui generis right, thus creating a barrier to TDM.[61] The rule that *copyright works* reproduced in a Big Data corpus retain independent copyright protection has not been altered. This means that images, texts, musical works and other copyright subject-matter

[51] P. Bernt Hugenholtz, 'Data Property: Unwelcome Guest in the House of IP', [2018] Kritika. Essays on Intellectual Property, vol. III. See also Estelle Derclaye 'The Database Directive', in Irini Stamatoudi and Paul Torremans (eds), *EU Copyright Law* (E. Elgar, 2014) 302-303.

[52] Database Directive (n 12) art. 7(1).

[53] Mark J. Davison & P. Bernt Hugenholtz, 'Football fixtures, horse races and spin-offs: The ECJ domesticates the database right' (2005) 27:3 EIPR 113, 116.

[54] See Irini A. Stamatoudi, 'Text and Data Mining', in Irini A. Stamatoudi (ed.), *New Developments in EU and International Copyright Law* (Wolters Kluwer, 2016) 266.

[55] *Fixtures Marketing Ltd v Oy Veikkaus Ab*, ECJ 9 November 2004, case C-46/02, ECR [2004] I-10396; *British Horseracing Board v William Hill Organization*, ECJ 9 November 2004, case C-203/02, ECR [2004] I-10415; *Fixtures Marketing Ltd v Svenska Spel AB*, ECJ 9 November 2004, case C-338/02, ECR [2004] I-10497; *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)*, ECJ 9 November 2004, case C-444/02, ECR [2004] I-10549.

[56] Hugenholtz (n 27).

[57] Matthias Leistner, 'Big Data and the EU Database Directive 96/9/EC: Current Law and Potential for Reform' (September 7, 2018). Available at SSRN: <https://ssrn.com/abstract=3245937>.

[58] European Commission, 'Staff Working Document on the free flow of data and emerging issues of the European data economy', Brussels, 10 January 2017, SWD(2017) 2 final, 33-38. See also European Commission, 'Building A European Data Economy', Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 10 January 2017, COM(2017) 9 final, 13.

[59] Ibid. p. 20.

[60] See Communication from the Commission to the European Parliament, The Council, the European Economic and Social Committee and the Committee of the Regions, "Towards a common European data space", COM(2018) 232 final, 25 April 2018.

[61] See Daniel L. Rubinfeld & Michal S. Gal, 'Access Barriers to Big Data' (2017) 59 Arizona Law Review 339, 368.

contained in a Big Data corpus are still subject to copyright protection until the expiry of the term of protection. This second point is by far the one that has attracted the largest amount of attention. Geiger et al. opined that "[o]nly TDM tools involving minimal copying of a few words or crawling through data and processing each item separately could be operated without running into a potential liability for copyright infringement."[62] This might explain why several jurisdictions have introduced TDM limitations and exceptions.

37 Four examples should suffice to illustrate the point. The German Copyright Act contains an exception for the "automatic analysis of large numbers of works (source material) for scientific research" for non-commercial purposes.[63] A corpus may be made available to "a specifically limited circle of persons for their joint scientific research, as well as to individual third persons for the purpose of monitoring the quality of scientific research."[64] The corpus must also be deleted once the research has been completed.[65]

38 France introduced an exception in 2016 allowing reproduction, storage and communication of "files created in the course of TDM research activities."[66] The reproduction must be from lawful sources.[67]

39 The UK statute provides for a right to make a copy of a work "for computational analysis of anything recorded in the work," but prohibits, however, dealing with the copy in other ways and makes contracts that would prevent or restrict the making of a copy for the purpose stated above unenforceable.[68]

40 Finally, the Japanese statute contains an exception for the reproduction or adaptation of a work to the extent deemed necessary "the purpose of information analysis ('information analysis' means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information."[69]

2. Designing Big Data/TDM exceptions

41 The examples in the previous paragraphs demonstrate a similar normative underpinning, namely a policy designed to allow TDM of data contained in copyright works. They disagree on the implementation of the policy, however. Based on those examples, the questions that policy-makers considering enacting an explicit TDM exception or limitation should include:

- Whether the exception applies to only one (reproduction) or all rights (including adaptation/derivation);
- Whether contractual overrides are possible;
- Whether the material used should be from a lawful source;
- What dissemination of the data, if any, is possible; and
- Whether the purpose of TDM is non-commercial.

42 The answers to all five questions can be grounded in a normative approach, but they should be set against the backdrop of the three-step test, which, as explained below, is likely to apply to any copyright exception or limitation.

43 Before taking a look at the five points in greater detail, it is worth recalling that there are other types of exceptions that might allow TDM in specific instances, such as general exceptions for scientific research and fair use.[70]

44 As to the first question, if allowing TDM is seen as a normatively desirable goal, then the right holder should not be able to use one right fragment in the bundle of copyright rights to prevent it. In an analysis of rights involved, Irini Stamatoudi came

[62] Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data' (2018) 49:7 EIPR 814, 818.

[63] Copyright Act of 9 September 1965 (Federal Law Gazette I, p. 1273), as last amended by Article 1 of the Act of 1 September 2017 (Federal Law Gazette I p. 3346), art. 60d. Available at <https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html>.

[64] Ibid.

[65] Ibid.

[66] Geiger et al. (n 51) 830.

[67] Law No. 2016-1231§ for a Digital Republic and art. L122-5 of the Intellectual Property Code.

[68] Added by the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, 2014 No. 1372. Online <https://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made>.

[69] Copyright Law of Japan, art. 47*septies*, translated by Yukifusa Oyama et al., available at <http://www.cric.or.jp/english/clj/doc/20161018_October,2016_Copyright_Law_of_Japan.pdf>.

[70] An example of the former may be found in arts. 5(3)(a) and 6(2)(b) of the Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, O.J. L 167, 22/06/2001 P. 0010 – 0019('InfoSoc Directive').

to the conclusion that right fragments beyond reproduction and adaptation were much less relevant.[71] Still, it would seem safer to formulate the exception or limitation as a non-infringing *use*, as in section 107 (fair use) of the US Copyright Act for example.[72]

45 Second, for the same reason, contractual overrides should not be allowed. One can hardly see how they can be effective unless perhaps there was only one provider of TDM for a certain type of work. Even if a provision against contractual overrides was absent from the text of the statute, the restriction could be found inapplicable based on principles of contract law.[73]

46 Third, the lawful source element contained in French law is facially compelling. It seems difficult to oppose a requirement that the source of the data be legitimate. There are difficulties in its application, however. First, it is not always clear to a *human* user whether a source is legal or not; the situation may be even less clear for a machine. Second, and relatedly, if the source is foreign, a determination of its legality may require an analysis of the law of the country of origin, as copyright infringement is determined based on the *lex loci delicti*—and this presupposes a determination of its origin (and foreignness) to begin with. Perhaps a requirement targeting sources that the user *knows or would have been grossly negligent in not knowing* were illegal might be more appropriate.[74]

47 The last two questions on the list above are somewhat harder. Dissemination of the data, if such data includes copyright works, could be necessary among the people interested in the work. German law makes an exception for a “limited circle of persons for their joint scientific research”, and “third persons for the purpose of monitoring the quality of scientific research.”[75] This is a reflection of a scientific basis of the exception, which includes project-based work by a limited number of scientists and monitoring by peer reviewers. This would not allow the use of TDM to scan libraries of books and make snippets available to the general public, as Google Books does, for example. An interpretation of the scope of the exception might depend on whether the use is commercial, which in turn might vary according to the definitional approach taken: is it the commercial nature of the *entity* performing the TDM that matters, or the specific use of the TDM data concerned (i.e., is that specific use monetized)?

48 As of early 2019, the EU was considering a new, mandatory TDM exception as part of its digital copyright reform efforts.[76] Article 3, which contains the proposed TDM exception, has been the focus of intense debates. The September 2018 (Parliament) version of the proposed TDM exception maintains the TDM exception for scientific research proposed by the Commission but adds an optional exception applicable to the private sector, not just for the benefit of public institutions and research organisations.[77] Members of the academic community have criticised the narrow scope of the Commission’s proposed exception, which the Parliament’s amendments ameliorated.[78] The European Copyright Society opined that “data mining should be permitted for non-commercial research purposes, for research conducted in a commercial context, for purposes of journalism and for any other purpose”.[79]

49 One should note, finally, that when a technological protection measure or “lock” such as those protected by art. 11 of the 1996 WIPO Copyright Treaty, is in place preventing the use of data contained in copyright works for TDM purposes, the question is whether a TDM exception provides a “right” to perform TDM and thus potentially a right to circumvent the TDM or obtain redress against measures designed to restrict it.[80] This might apply to traffic management (e.g., throttling) measured used to slow the process down. Those questions are worth pondering, but they are difficult to answer, especially at the international level.[81]

[71] Stamatoudi (n 32) 262.

[72] The US Copyright Act (17 USC s 107) reads in part as follows: “the fair use of a copyrighted work ... is not an infringement of copyright”.

[73] See for example Lucie Guibault’s detailed analysis of the possible application of the German *Sozialbinding* principle in this context. Lucie M.C.R. Guibault, *Copyright Limitations and Contracts: An Analysis of the Contractual Overridability of Limitations on Copyright* (Kluwer Law International, 2002) 224-225.

[74] This language echoes footnote 10 (to art. 39(2)) of the TRIPS Agreement (n 1).

[75] See n 54.

[76] Geiger at al. (n 51) 832-33.

[77] The Parliamentary version and the Commission’s proposal are compared in amendments 64 and 65 of the document ‘Amendments adopted by the European Parliament on 12 September 2018 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market’ (COM(2016)0593 – C8-0383/2016 – 2016/0280 (COD)) (1), online: <https://bit.ly/2SS3HYA>.

[78] See e.g. Martin Senftleben, ‘EU Copyright Reform and Startups – Shedding Light on Potential Threats in the Political Black Box’ (undated), at p. 9. Online: <https://bit.ly/2kiJgFq>.

[79] European Copyright Society, General Opinion on the EU Copyright Reform Package, 24 January 2017. Online: <https://bit.ly/2k2k3jD>.

[80] WIPO Copyright Treaty, 20 Dec. 1996.

[81] For a brief discussion, see Geiger at al. (n 51) 836-838.

3. Application of the Three-Step Test

50 The three-step test sets boundaries for exceptions and limitations to copyright rights.

51 The original three-step test is contained in art. 9(2) of the Berne Convention. Its purpose is to allow countries party to the Convention to make exceptions to the right of reproduction (1) "in certain special cases", (2) "provided that such reproduction does not conflict with a normal exploitation of the work", and (3) "does not unreasonably prejudice the legitimate interests of the author".[82] The test was extended to all copyright rights by the TRIPS Agreement, with the difference that the term "author" at the end was replaced with the term "right holder".[83]

52 The test was interpreted in two panel reports adopted by the World Trade Organization's Dispute-Settlement Body.

53 The first step ("certain special cases") was interpreted to mean that "an exception or limitation must be limited in its field of application or exceptional in its scope. In other words, an exception or limitation should be narrow in quantitative as well as a qualitative sense".[84] The Study Group discussed the possible inclusion of the test in the Berne Convention before the 1967 (Stockholm) revision had opined that the test should require that any exception to the right of reproduction be "for clearly specified purposes".[85]

54 The normative grounding to justify a TDM exception is fairly clear. Indeed, exceptions and limitations have already been introduced in major jurisdictions. A well-justified exception or limitation with reasonable limits and a clear purpose is likely to pass the first step.

55 The second step (interference with normal exploitation) was defined as follows. First, exploitation was defined as any use of the work by which the copyright holder tries to extract/maximize the value of her right. "Normal" is more troublesome. Does it refer to what is simply "common", or does it refer to a normative standard? The question is particularly relevant for new forms and emerging business models that have not, thus far, been common or "normal" in an empirical sense. At the revision of the Berne Convention in Stockholm in 1967, the concept was used to refer to "all forms of exploiting a work, which have, or are likely to acquire, considerable economic or practical importance".[86] In other words, if the exception is used to limit a commercially significant market or, a fortiori, to enter into competition with the copyright holder, the exception is prohibited.[87]

56 Could a TDM exception be used to justify scanning and making available entire libraries of books still under active commercial exploitation? The answer is negative, as this would interfere with commercial exploitation. For books still protected by copyright *but no longer easily available on a commercial basis*, the absence of active commercial exploitation would likely limit the impact of the second step, however, subject to a caveat. Some forms of exploitation are typically done by a third party under license and do not need any active exploitation *by the right holder*. For example, a film studio might want the right to make a film out of a novel no longer commercially exploited. That may in turn generate new demand for the book. This is still normal exploitation. One must be careful in extending this reasoning too far, for example, and assume that every novel will be turned into a movie.

57 TDM is quite comparable to the not adaptation of a novel to the big screen. Its purpose is *not* to convey the same or similar expressive creativity via a different medium. TDM is looking, if anything, for *ideas* embedded in copyright works. Because Big Data corpora used for TDM are necessarily composed of large numbers of works and other data, the TDM function cannot be performed if licensing work by work is required. This is also differs in the case of a film adaptation, a scenario in which it seems reasonable to expect that the author (or her representative) and the film producer might negotiate a license.

58 One way to pass the second step would be for a TDM exception to allow limited uses that do not demonstrably interfere with commercial exploitation, such as those allowed under the German

[82] Berne Convention (n 11) art. 9(2).

[83] TRIPS Agreement, art. 13. The test is now used as the model for exceptions to *all copyright rights* in TRIPS (art. 13); Articles 10(1) and (2) of the WIPO Copyright Treaty (20 December 1996); Article 16(2) of the WIPO Performances and Phonograms Treaty (also adopted on 20 December 1996); Article 13(2) of the Beijing Treaty on Audiovisual Performances (24 June 2012); and Article 11 of the Marrakesh Treaty to Facilitate Access to Published Works for Persons who are Blind, Visually Impaired or Otherwise Print Disabled (27 June 2013). Interestingly, in TRIPS, it is also the test for exceptions to industrial design protection (art. 26(2)) and patent rights (art. 30).

[84] WTO Report of the Panel WT/DS160/R of 15 June 2000 on United States – Section 110(5) of the US Copyright Act, para 6.109 (emphasis added and citations omitted). [hereinafter "panel report"]. The second case led to the following panel report: WT/DS114/R of 17 March 2000 on Canada – Patent Protection of Pharmaceutical Products.

[85] Records of the Intellectual Property Conference of Stockholm: June 11 to July 14, 1967 (WIPO, 1971) 112.

[86] Ibid, at 112.

[87] Paul Goldstein, *International Copyright: Principles, Law, and Practice* (OUP 1998) 295.

statute. Another example is the use of "snippets" from books scanned by Google for its Google Books project, which was found to be a fair use by the US Court of Appeals for the Second Circuit. This matters not just as a matter of US (state) practice but because at least the fourth fair use factor ("the effect of the use upon the potential market for or value of the copyrighted work") is a market-based assessment of the impact of the use resembling the three-step test's second step.[88] The Second Circuit noted that this did not mean that the Google Books project would have *no* impact, but rather that the impact would not be meaningful or significant.[89] It also noted that the type of loss of sale created by TDM "will generally occur in relation to interests that are not protected by the copyright. A snippet's capacity to satisfy a searcher's need for access to a copyrighted book will at times be because the snippet conveys a historical fact that the searcher needs to ascertain."[90] In the same vein, one could argue that the level of interference required to violate the second step of the test must be significant and should be a use that is relevant from the point of view of commercial exploitation.

59 The third step (no unreasonable prejudice to legitimate interests) is perhaps the most difficult to interpret. What is an "unreasonable prejudice", and what are "legitimate interests"? Let us start with the latter. "Legitimate" can mean sanctioned or authorized by law or principle. Alternatively, it can just as well be used to denote something that is "normal" or "regular". The WTO dispute-settlement panel report concluded that the combination of the notion of "prejudice" with that of "interests" pointed clearly towards a legal-normative approach. In other words, "legitimate interests" are those that are protected by law.[91]

60 Then, what is an "unreasonable" prejudice? The presence of the word "unreasonable" indicates that *some level or degree* of prejudice is justifiable. Hence, while a country might exempt the making of a small number of private copies entirely, it may be required to impose a compensation scheme, such as a levy, when the prejudice level becomes unjustified.[92] The WTO panel concluded that "prejudice to the legitimate interests of right holders reaches an unreasonable level if an exception or limitation causes or has the potential to cause an unreasonable loss of income to the copyright holder".[93]

61 Whether a TDM exception is liable to cause an unreasonable loss of income to copyright holders is analytically similar to the second step of the test as interpreted by the WTO panels. It is not, however, identical: The owner of rights in a work no longer commercially exploited may have a harder case on the second step. It is not unreasonable, however, for a copyright holder, to expect some compensation for use of a protected work even if it is not commercially exploited. For example, the owner of rights in a novel may expect compensation for the republication by a third party or translation of the book. The major difference between the second and third step in this regard is that the third step condition may be met by compensating right holders. This would allow the imposition of a compulsory license for specific TDM uses that overstep the boundary of free use, for example to make available significant portions of, or even entire, protected works that are no longer commercially exploited. For example a TDM engine could find all works that fit a user's criteria (say, 20th century novels, in any language, where a murder by poison takes place and both Pontius Pilate and a cat play a prominent part in the plot).[94] Then the system could (a) make the text or part thereof available, against adequate compensation, especially if no e-book database existed; or (b) generate a translation or summary if the book, especially if no linguistic version of use to the searcher was available.[95]

E. Patents

I. The role of Big Data in patent disclosures

62 The interface between patents and Big Data is interesting on several levels.

63 First, TDM might be used in enhancing the use of patent information.[96] The "patent bargain" is basically a fair disclosure of an invention in exchange for a limited monopoly on its use,

[88] The fourth fair use factor contained in the US Copyright Act (17 USC s 101) reads as follows: "the effect of the use upon the potential market for or value of the copyrighted work."

[89] *The Authors Guild v. Google, Inc.* 804 F.3d 202 (2d Cir, 2015), cert. denied 136 S.Ct. 1658.

[90] Ibid.

[91] Panel Report, paras 6.223–6.229. In para. 6.224 the Panel tried to reconcile the two approaches: "[T]he term relates to lawfulness from a legal positivist perspective, but it has also the connotation of legitimacy from a more normative perspective, in the context of calling for the protection of interests that are justifiable in the light of the objectives that underlie the protection of exclusive rights".

[92] Records (n 72) 1145–46.

[93] Panel Report (n 71) para. 6.229.

[94] The reader will have recognized the unlikely plot of Mikhail Boulgakov's masterpiece, *Master and Margarita*.

[95] The application of both the Berne Convention Appendix (for developing countries) and the Marrakesh VIP Treaty might also be considered in this context.

[96] See Dario Mastrelia, 'Patent information and technology transfer in the information society era: From the current scenario to new business ideas' (2018) 40:7 EIPR 460.

especially on a commercial basis.[97] Unfortunately, patent information is often mired in a difficult language known as "patentese", which obscures the informational function of published patents.[98] An AI-capable TDM system might be able not just to find but also to *interpret* useful information and facilitate technology transfers.[99] Relatedly, AI and patent information could be combined not just to interpret patent claims but also to determine their validity.[100]

64 AI applications in this field already go further, however, and the trajectory of their development leads to some potentially remarkable conclusions. First, existing AI-based systems using Big Data (e.g. databases of published patents and technical literature) allow patent applicants to maximize the exclusivity claimed in their patent applications by identifying material analogous to the invention that can also be claimed—essentially variations on the theme of the invention—thus potentially broadening its scope *beyond what the applicant actually invented.*[101]

II. Big Data and the future of innovation

65 This section is admittedly at the border between current technology and the future. Part of it is thus speculation based on how current AI systems using patent corpora and AI are likely to evolve. Various options are considered. Hopefully, the reader will find some of it useful.

66 The kind of claim-broadening system described above can be used for a different purpose, namely to *disclose* (*without* claiming patent rights) incremental variations on claims of existing patents, thus potentially preventing patenting of improvements and even derivative and incremental inventions in the future.[102] Are AI-generated disclosures of variations on existing inventions or incremental innovations sufficient to defeat novelty?[103] If massive disclosures through AI-systems of incremental variations on existing patents become common, patent courts and offices might be tempted—for both institutional and normative reasons—to limit the patent-defeating power of such disclosures, for example by insisting that they do not sufficiently enable or describe the invention, which would remain patentable, therefore, when an application is filed by a (human) person providing a more complete disclosure. More neutral outcomes might be obtained in higher courts.

67 The discussion of the role of Big Data-based AI systems in innovation disclosures can be taken up a level. As Yanisky Ravid and Liu note:

> *AI systems create a wide range of innovative, new, and non-obvious products and services, such as medical devices, drug synthesizers, weapons, kitchen appliances, and machines, and will soon produce many others that, had they been generated by humans, might be patentable inventions under current patent law.*[104]

68 There is little doubt that Big Data-based AI systems will innovate, that is, they will produce what one might call "inventions". Indeed, Google's AI system, known as DeepMind, already thinks it does and it has filed patent applications.[105] The first question to ask in this context is whether such inventions are patentable. The second is, what will the broader impact on innovation be?

69 As noted in the introductory part, Big Data-based AI systems are more likely to generate incremental innovations than pioneer inventions. They could so, however, at a pace of innovation that could eclipse any previous period in human history, causing an exponential increase over the (already very fast) pace of current technological change.

[97] The obligation to disclose is reflected in art. 29.1 of the TRIPS Agreement. See also Katherine J. Strandburg, 'What Does the Public Get? Experimental Use and the Patent Bargain' [2004] Wisconsin. LR 81, 111-17.

[98] Sean B. Seymore, 'The Teaching Function of Patents' (2010) 85 Notre Dame LR 621, 633–34.

[99] See Mastrelia (n 83) 465. It may also be useful to recall that art. 7 of the TRIPS Agreement mentions that "the transfer and dissemination of technology, to the mutual advantage of producers and users of technological knowledge and in a manner conducive to social and economic welfare, and to a balance of rights and obligations".

[100] See Ben Dugan, 'Mechanizing Alice: Automating the Subject Matter Eligibility Test of Alice v. CLS Bank' [2018] U Illinois J L Tech & Policy 33.

[101] See Ben Hattenbach, Joshua Glucoft, 'Patents in an Era of Infinite Monkeys and Artificial Intelligence' (2015) 19 Stanford Technology LR 32, 35, describing a company called CLOEM using "brute-force computing to mechanically compose text for thousands of patent claims covering potentially novel inventions and also to generate defensive publications to prevent others from obtaining patent protection in the same field".

[102] See Ryan Abbott, 'I Think, Therefore I Invent: Creative Computers and the Future of Patent Law' (2016) 57 Boston College LR 1079, describing "projects such as "All Prior Art" and "All the Claims" which attempt to use machines to create and publish vast amounts of information to prevent other parties from obtaining patents".

[103] Though there is no formal international test, typically this would require that the disclosure provide enough information for a person skilled in the art to make or practice the invention. For a discussion (under US law) see Jennifer L. Kisko and Mark Bosse, 'Enablement and Anticipation' (2007) 89 J Patent & Trademark Office Society 144, 151.

[104] Yanisky Ravid and Liu (n 9), 2219-2220.

[105] Mike James, 'Google's DeepMind Files AI Patents', i-programmer (11 June 2018) <https://bit.ly/2ATh5or>.

One company active in the field markets itself as creating "commercially relevant inventions at high speed and with great diversity" and notes that "[h]undreds of patents based on *our inventions* have been filed by some of the best-known technology companies worldwide".[106] If this type of technology continues to grow, as it surely will, we could reach a *singularity of innovation*.[107] The notion of "singularity" became well-known after the publication of Ray Kurtzweil's famous 2006 book on the topic.[108] The singularity, according to Kurzweil, will be a reality when computers become more "intelligent" than humans.[109]

70 An innovation singularity would compel a fundamental rethink of the innovation incentive system. From a first to disclose (and patent) system, one might need to consider a "first to develop" system. Such a system would lead to a series of both doctrinal and normative questions, including: whether any period of exclusivity is essential and then how long; who can apply; what period of time do they have to actually develop; and then develop what (proof of concept, actually marketable product, etc.); to which territory does it apply, and the list goes on.

71 The future might not take a public domain path (through massive disclosures) and opt for a proprietary route instead. Big Data based "inventions" reflecting the deep learning ability of AI systems might deserve protection by patents *even if no discernible human contribution to the inventive process has taken place*. The forces that might restrict the scope of novelty-destroying disclosures mentioned in the previous paragraphs might push back against a public domain trajectory and help grant patents even if the broader scope of claims in applications is the product of claim-broadening algorithms. This would mean that claims added or broadened by a Big Data based AI system to a patent application (and possibly entire new applications) might have to be granted to a person (natural or legal) for inventions that the applicant does not actually possess and is very possibly unable to exploit. Whether this occurs, in turn, might depend on the ability of the AI system to explain its invention.[110]

72 The impact of such a scenario might depend on how the market would react. If owners of patent rights in inventions they cannot exploit license them to companies that can exploit them, then private ordering might solve the otherwise massive blocking effect. The blocking effect could become a patent troll's dream, however, allowing the capture of vast areas of incremental innovation and thus exponentially expanding the reach of trolls in this space.[111]

73 As with copyright "authorship", one might fairly ask whether there must be human inventorship for a patent to be granted. No definitive answer can be given under current law, and a full analysis is beyond the scope of this Article. Divergences of views have emerged.[112] One might add that this

[106] <http://www.iprova.com/about-us/> (accessed 21 January 2019).

[107] See Ryan Abbott, 'I Think, Therefore I Invent: Creative Computers and the Future of Patent Law' (2016) 57 Boston Coll LR 1079, 1079–80 ("A creative singularity in which computers overtake human inventors as the primary source of new discoveries is foreseeable".)

[108] Ray Kurzweil, *The Singularity Is Near* (Viking, 2006). It seems, however, that the notion originated earlier. For example, it can be found Vernor Vinge, 'The Coming Technological Singularity: How to Survive in the Post-Human Era' (Winter 1993) *Whole Earth Review* (online <https://edoras.sdsu.edu/~vinge/misc/singularity.html>.

[109] See ibid. Vinge also discussed the idea that those computers might somehow become "aware".

[110] Explanation in this context is sometimes referred to as dumbing it down for humans to understand the machine's "thinking", or explaining "to a lay audience in such a way that they can make use of such explanations." Sandra Wachter et. al., 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 Harvard J L & Tech. 841, 851.

The problem is that the best AI insights may be the ones that the machine is least able to explain. For example, if a Big Data based AI system was excellent at diagnosing a certain disease, explanation might not be possible, but then I suspect that in such a case the value of excellent diagnostic capabilities would outweigh the need for an explanation.

[111] A patent is blocking if "if circumventing it (1) is not commercially practicable, or (2) will not produce a commercially viable product". Ian Simmons, Patrick Lynch, Theodore H. Frank, '"I Know It When I See It": Defining and Demonstrating "Blocking Patents"' (2002) 16 Antitrust 48, at 49.

As professor Robert Merges noted, "patent law's property rule, which requires a voluntary patentee-infringer bargain or an injunction against infringement, assumes that if a bargain would benefit both parties, they will reach one". Robert Merges, 'Intellectual Property Rights and Bargaining Breakdown: The Case of Blocking Patents' (1994) 62 Tennessee LR 75, 78. That assumption is questionable. However, the problem that AI might cause may also be solved (in part) by AI by facilitating contacts between potential licensor and licensee (Thanks to Florent Thouvenin (University of Zurich) for this insight).

[112] In the United States, though the law seems to require human inventive activity, the Patent Office (USPTO) has reportedly granted 'several patents with nonhuman inventors, albeit not explicitly and not necessarily with their knowledge". Russ Pearlman, 'Recognizing Artificial Intelligence (Ai) As Authors and Inventors Under U.S. Intellectual Property Law' (2018) 24 Richmond J.L. & Technology 2, 23. Normatively, "[t]he concept of an inventor does not fit neatly into scenarios in which the invention emerges from random interactions between existing computer programs, repeated computer simulations using all possible scenarios, or other forms of data mining, perhaps with little or no direction or forethought on the part of the human operator". Liza Vertinsky & Todd M. Rice, 'Thinking About Thinking Machines: Implications of Machine Inventors for

presupposes that one actually *knows* whether a human or a machine is the "inventor". If the patent applicant does not need to provide proof of human invention, perhaps courts will require it later on in infringement proceedings and invalidate patents for lack of (human) inventorship.

74 The last question in this section is whether there can be patents on AI systems themselves. International patentability criteria are contained in art. 27 of the TRIPS Agreement. This provision leaves World Trade Organization (WTO) members a fair degree of flexibility in determining what constitutes an "invention", and then whether such invention is new, involves an inventive step (or is non-obvious) and is industrially applicable (or useful).[113] The European Patent Office (EPO) issued new Examination Guidelines (in force November 2018) noting that "[a]rtificial intelligence and machine learning are based on computational models and algorithms for classification, clustering, regression and dimensionality reduction, such as neural networks, genetic algorithms, support vector machines, k-means, kernel regression and discriminant analysis", and that ["s]uch computational models and algorithms are *per se* of an abstract mathematical nature, irrespective of whether they can be 'trained' based on training data".[114] In the United States, algorithms are also essentially unpatentable since the US Supreme Court's decision in *Alice v. CLS Bank,* which imposed a two-part test that most computer programs are unlikely to pass.[115] The focus is now on the machine: "If the novel feature is the use of a computer, the patent will likely be invalid, while if the novel feature is a better computer, the patent will likely be valid."[116] The role of patents in protecting algorithms thus seems fairly narrow going forward.

III. Localization and working requirements

75 There is a final point, arguably tangential but nonetheless potentially relevant, to be made in connection with patents and Big Data. In 1995, when the TRIPS Agreement entered into force, rules were meant to limit or eliminate the so-called working requirements in patent law, which were legal under previous international rules.[117] This requirement was seen, in a number of (mostly developing) countries as a part of the patent bargain.[118] A patent, as defined in TRIPS, is a right to exclude not conditioned on either availability or manufacture or other use of the patented invention in the territories where a patent is in force.[119] Prior to TRIPS, certain countries imposed a (local) working requirement to make sure that patented inventions would be available (and the technology used) in the country. The TRIPS rationale is, in short, that companies should be able to produce patented inventions wherever they believe it is more efficient and export to other territories. Local working requirements parallels the current clash between personal data protection and (free) trade.

76 This is relevant to Big Data because a common form of personal data protection is *data localization*.[120] Is the assumption that free trade is a desirable normative goal applicable here? Cross-border data flow limits seem to be a pushing back against free trade.[121] This indirectly imposes a local "working requirement" on AI corpora containing personal data. If IP law is prologue, free trade (i.e. free cross-border data flows) will win that debate.

F. Data Exclusivity

77 There is a right often closely associated with patents for pharmaceuticals, namely the right of data exclusivity.[122] This is the right to prevent certain

Patent Law' (2002) 8 Boston Univ J Science & Tech L 574, 586.

[113] See Carlos M. Correa, 'Public Health and Patent Legislation in Developing Countries' (2001) 3 Tulane J Technology and Intellectual Prop 1, 8-9.

[114] European Patent Office, 'Guidelines for Examination' (Nov. 2018), sec. 3.3.1. Available at <https://www.epo.org/law-practice/legal-texts/html/guidelines2018/e/g_ii_3_3_1.htm>.

[115] *Alice Corp. Pty. Ltd. v. CLS Bank Int'l* (2014).134 S. Ct. 2347, 2354-55.

[116] Fabio E. Marino & Teri H. P. Nguyen, 'From Alappat to Alice: The Evolution of Software Patents' (2017) 9 Hastings Science & Tech LJ 1, at 28.

[117] TRIPS entered into force on 1 January 1995. The principal set of substantive patent rules before TRIPS were contained in the Paris Convention for the Protection of Industrial Property, of March 20, 1883, last updated in Stockholm (1967), art. 5A.

For a discussion of the working requirement, see Bryan Mercurio & Mitali Tyagi, 'Treaty Interpretation in WTO Dispute Settlement: The Outstanding Question of the Legality of Local Working Requirements' (2010) 19 Minnesota J Intl L. 275, 279-288.

[118] See Katherine J. Strandburg, 'What Does the Public Get? Experimental Use and the Patent Bargain' (2004) Wisconsin LR 81.

[119] TRIPS (n 1), arts. 27(1) and 31.

[120] For a (critical) discussion of national data localization practices, see Bret Cohen, Britanie Hall, Charlie Wood, 'Data Localization Laws and Their Impact on Privacy, Data Security and the Global Economy' (2017) 32 Antitrust 107.

[121] See Svetlana Yakovleva, 'Should Fundamental Rights to Privacy and Data Protection be a Part of the EU's International Trade "Deals"'? (2018) 17:3 World Trade Rev 477.

[122] For a fuller discussion of this interface, see Daniel Gervais, 'The Patent Option' (2019) 20 North Carolina J L & Tech (forthcoming), draft available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266580>.

forms of use of clinical trial data generated to obtain marketing approval for certain pharmaceuticals and chemical products. A basic data exclusivity right is contained in TRIPS.[123] More extensive protection is contained in post-TRIPS (in the so-called "TRIPS-Plus") agreements.[124] There is a concern that such protection might prevent the use of TDM tools, which is seen as a negative development because "it is the collected clinical trial data, and their ability to provide a large and comprehensive dataset, that are highly valuable, not the specific health and safety outcome proven by those data."[125]

78 This right is directly relevant. As discussed in the previous section, patents may become more difficult to obtain due to massive Big Data –based AI disclosures of possibly new incremental innovations. For example, such a system could conceivably disclose new molecules and predict their efficacy. In such a case, it would be near impossible to patent the drug unless patented by the user of the AI "inventor". If it was patented by the AI inventor, then that person's consent could be required to test the new molecule. In both cases the company investing in the testing might not own a patent on the molecule and find it hard to justify the expense of generating clinical test data. The data exclusivity right might fill that void. The right is, however, of limited application beyond the pharmaceutical and agrochemical fields.

G. Trade Secrets and Confidential Information

79 Let us end our *tour d'horizon* with the protection of confidential information, including the subset of confidential information known as trade secrets. Trade secrets and confidential information laws, and contracts, can be used to enable the orderly disclosure of information.[126] That protection is reflected in the TRIPS Agreement.[127] This type of protection of secrets information is compatible with, and often based on, legislation such as the Trade Secrets Directive and a host of national laws.[128]

80 What is the area of application of trade secret law to Big Data? Cristina Sappa analysed the application of trade secret law to data gathered via the Internet of Things (IoT).[129] She suggested three areas which seem to be worthy of further study.

81 First, "within the IoT realm, as in any other business, trade secrets are used to protect information to which access is traditionally limited thanks to (among others) confidentiality clauses or non-disclosure agreements."[130] Thus, trade secret and confidential information law—in this case with the support of contract law—could be used to protect data acquired for purposes of TDM.[131] Trade secret law typically works far better for business information than private data.[132] One might indeed expect the default contracts may not adequately protect the users or consumers—though privacy or consumer protection laws may impose limits on contractual freedoms that include minimum guarantees of confidentiality.[133]

82 Secondly, the protection of confidential information could apply to non-trivial "data coming from a machine-to-machine process".[134] One commentator suggested that "trade secrets, rather than database *sui generis* rights, are the most interesting and flexible property right for coping with the challenge of customer data appropriation in the new, collaborative economy 3.0".[135] For example, if a corpus of Big Data was processed to generate a database of correlations between persons and their preferences (but let us assume that such a database does not or no longer contains the data used to generate the correlations), the new corpus of correlations and insights derived from such correlations may well be protected as a

[123] TRIPS Agreement (n 1) art 39(2).

[124] See Peter K. Yu, 'Data Exclusivities in the Age of Big Data', Texas A&M Univ School of Law Legal Studies Research Paper-Series No. 18-08, at 5-8. Available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3133810>.

[125] Ibid 4.

[126] See Mark Lemley, 'The Surprising Virtues of Treating Trade Secret ad IP Rights' (2008) 61 Stanford LR 311.

[127] TRIPS Agreement (n 1) art. 39.2.

EU law defines a trade secret as "valuable know-how and business information, that is undisclosed and intended to remain confidential" generated by businesses and non-commercial research institutions that "invest in acquiring, developing and applying know-how and information which is the currency of the knowledge economy and provides a competitive advantage". Directive 2016/943 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, recital 1.

[128] Ibid.

[129] Cristiana Sappa, 'What Does Trade Secrecy Have To Do with the Interconnection-Based Paradigm of the Internet of Things?' (2018) 40:8 EIPR 518.

[130] Ibid 521.

[131] TRIPS Agreement (n 1) art. 39.2.

[132] Pamela Samuelson, 'Privacy as Intellectual Property?' (2000) 52 Stanford LR1125, 1151-70.

[133] This would of course include the GDPR (Regulation (EU) 2016/679 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.)

[134] Sappa (n. 113) 523.

[135] Gianclaudio Malgieri, '"Ownership" of Customer (Big) Data in the European Union: Quasi-Property As Comparative Solution?' (2016) 20 J Internet L 3.

trade secret or a database where it exists. Moreover, its use may no longer be limited by the personal data protection that applied to the raw data.

83 Thirdly, Sappa suggests we should consider the "possibilities of welfare gains by third parties, since this regime applying to *knowledge commons* such as the IoT enables spillovers, and therefore its presence may not necessarily be perceived as a bad thing."[136] Excessive restrictions on access to lock-in effects by major data gathering entities might have negative welfare impacts warranting governmental intervention in "data--driven platform markets characterized by strong network and lock--in effects--and in new technological contexts that might otherwise be ripe for competitive innovation."[137]

H. Conclusion

84 This article reviewed the application of IP rights to Big Data. In most cases, AI software is protected by copyright. Copyright's traditional role is otherwise in tension with the creation and use of Big Data corpora, however. The nature of the non-relational (noSQL) databases typical of Big Data corpora implies that such corpora are unlikely to be protected by copyright or by the EU sui generis rights in databases. Misappropriation (tort-based) protection might fill the gap, especially for data generated by AI systems that has high but short-lived value (e.g. in the FinTech sector).[138] Exceptions for Text and Data Mining are probably required to allow TDM using corpora of literary and artistic works, such as texts and images and video. Such exceptions are likely to continue to emerge in more jurisdictions around the world.

85 The questions concerning patents are not easy to answer. AI systems can be used to expand patent applications, but they can also be used to "guess" future incremental innovation and disclose them. Whether that disclosure will be interpreted by patent offices and courts as novelty-defeating is an open question. Whether AI-inventions— with no direct human input—are patentable is a matter under discussion as of this writing.

86 The article also reviewed data exclusivity and trade secrets. The latter might protect correlations and insights generated by AI systems, even if those are based on deep learning including the processing of protected personal data. This might generate tension between personal data protection and IP. The former might fill gaps in patent protection but only in areas where it applies (essentially chemical and pharmaceutical products).

87 In sum, the interfaces between Big Data and IP are about finding ways to adapt IP rights to allow and set proper parameters for the generation, processing and use of Big Data. This includes an analysis of how Big Data may infringe IP rights. There is also an issue of rights *in* Big Data, however. Courts and legislators have years of questions to answer on both constraints in and protection of Big Data.

[136] Ibid.

[137] Kenneth A. Bamberger & Orly Lobel, 'Platform Market Power' (2017) 32 Berkeley Tech. LJ 1051, 1089.

[138] See European Commission, "Consultation document. FinTech: A more competitive and innovative European Financial Sector", 2017, available at <https://ec.europa.eu/info/sites/info/files/2017-fintech-consultation-document_en_0.pdf> (last accessed 15 December 2018).