Legally Speaking: Of Mindfields and Minefields: Legal Issues in Text and Data Mining

William M. Hannay
Schiff Hardin, LLP, whannay@schiffhardin.com

Bruce Strauch
The Citadel, strauchb@citadel.edu

Bryan M. Carson
bryan.carson@wku.edu

Jack Montgomery
Western Kentucky University, jack.montgomery@wku.edu

Follow this and additional works at: http://docs.lib.purdue.edu/atg
Part of the Library and Information Science Commons

Recommended Citation
DOI: https://doi.org/10.7771/2380-176X.6663

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
I have been asked to address the topic of opportunities and challenges for text and data mining in academia, with a special emphasis on the legal issues involved in these research techniques. To do so, I will begin by taking you on a sightseeing trip through some mining regions outside the ivy-covered walls. Only by seeing the broader contexts in which text and data mining is used can one appreciate the legal issues in which academics must operate.

**Background**

What has come to be called “data mining” is the process of extracting hidden knowledge from large amounts of raw data. Data mining goes beyond traditional searches of databases. In traditional database searches, information is returned in response to a direct query. Think historically of searching through a card catalogue based on the Dewey Decimal System or, much more recently, doing an Internet search for books written by that eminent legal scholar, Bill Hannay. For the purpose of this speech, for example, I did a traditional search in the Lexis-Nexis legal database for court cases in the United States mentioning text mining or data mining. (The answer was 116 cases, but a number of them were false hits.)

By contrast, in “real” data mining, what is retrieved is not explicitly in the database. Rather, the desired information implicitly emerges from patterns and relationships. The process of discovering (and then analyzing) such patterns and relationships has come to be termed “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.”

The term became popularized in the early 1990s coextensively with the term “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.”

Data mining finds these sorts of patterns and relationships using data analysis tools and techniques to build models. It combines tools from the discipline of statistics and the field of artificial intelligence (“AI”) with database management techniques. A data mining project can look either for trends or for anomalies. In the latter regard, one of the earliest commercial successes was in credit card fraud detection.

By contrast, in “real” data mining, what is retrieved is not explicitly in the database. Rather, the desired information implicitly emerges from patterns and relationships. The process of discovering (and then analyzing) such patterns and relationships has come to be termed “data mining.” The term became popularized in the early 1990s coextensively with the term “data mining.”

For example, the insurance industry compiles hundreds of thousands of claims reports from different insurance carriers into a common database and uses data mining techniques to identify potential claim fraud schemes. Where individuals have unusually frequent accidents and the same groups of doctors, lawyers, and chiropractors repeatedly show up in connection with “rear-ender” collisions or “slip and falls,” it may indicate patterns which deserve further investigation.

In the wake of September 11, 2001, for example, a data mining company presented the Department of Defense with a data pattern analysis proposal geared toward improving the security of military installations in the United States and possibly abroad. It suggested that a rigorous analysis of personal characteristics of persons who sought access to military installations might be used to predict which individuals pose a risk to the security of those installations.

At the same time, the pervasiveness of massive databases — especially ones containing personally identifiable information or matters of personal sensitivity — has led to increased concern about the government or private industry making dangerous intrusions into the private lives of millions of Americans.

In June 2013, U.S. and U.K. newspapers reported that the FBI and the National Security Agency (“NSA”) had for a number of years been obtaining access to vast amounts of telephone and Internet data on millions of American citizens, including those who only make calls to other U.S. numbers, for data mining purposes unrelated to any specific target or investigation. The news stories resulted from the leak of classified and unclassified documents to The Guardian newspaper by former NSA consultant Edward Snowden. The furor over the NSA data mining has intensified in the past few months and led to both law suits and proposed legislation.

The pharmaceutical industry has also been making substantial use of data mining, particularly in connection with marketing to the doctors that prescribe medications. Drug stores and pharmacies, as a matter of business routine and federal law, receive prescriber-identifying information when processing prescriptions. Many pharmacies sell this information to IMS America and other data-collating firms that analyze prescriber-identifying information and produce reports on prescriber behavior. “Data miners” lease these reports to pharmaceutical manufacturers subject to nondisclosure agreements. In turn, “detailers” — salesmen who represent the drug manufacturers — then use these reports to refine their marketing tactics aimed at doctors in order to increase sales.

These practices took a dark turn when drug manufacturers began using the mined data to encourage physicians to prescribe drugs for “off-label” uses. Off-label use is the use of pharmaceutical drugs for an unapproved indication or in an unapproved age group, unapproved dosage, or unapproved form of administration. While it is legal for a physician to independently decide to prescribe a drug off-label, it is illegal for the drug company to promote off-label uses to prescribers. Under the Food, Drug, and Cosmetic Act, manufacturers are prohibited from directly marketing a drug for a use other than the FDA approved indication.

The stakes for drug makers can be high. In 2009, Pfizer agreed to pay $2.3 billion in fines and penalties to resolve charges brought by the U.S. Department of Justice, arising from the illegal “off-label” marketing of prescription drugs by the company. Using sophisticated “prescription data mining” and “influence mapping” analyses, Pfizer had targeted specific physicians for visits by Pfizer sales representatives to promote off-label uses of Pfizer drugs. Litigation has also been brought based on the....
Legally Speaking

from page 52

sale of pharmacy customer prescription information to data mining companies who sell that information to drug manufacturers for marketing purposes.11

In an effort to ban this sort of data mining, various state legislatures enacted laws prohibiting the use of prescription information for marketing purposes. For example, in 2007, the state of Vermont enacted the Prescription Confidentiality Law (Act 80), one component of which is the following:

A health insurer, a self-insured employer, an electronic transmission intermediary, a pharmacy, or other similar entity shall not sell, license, or exchange for value regulated records containing prescriber-identifiable information, nor permit the use of regulated records containing prescriber-identifiable information for marketing or promoting a prescription drug, unless the prescriber consents…. Pharmaceutical manufacturers and pharmaceutical marketers shall not use prescriber-identifiable information for marketing or promoting a prescription drug unless the prescriber consents…. [18 Vt. Stats. Ann. § 4631(d).]

Drug manufacturers fought back and brought litigation in federal court alleging that their constitutional free speech rights had been violated. Though they lost at the trial court level, the manufacturers won a victory in a 6-3-3 vote in the U.S. Supreme Court.

Writing for the majority, Supreme Court Justice Kennedy held in Sorrell v. IMS Health12 that the manufacturers’ First Amendment rights had been violated. Sorrell struck down the Section 4631(f) of the Vermont law forbidding the sale of prescriber-specific information by pharmacies to “pharmaceutical manufacturers and pharmaceutical marketers.” The Supreme Court rejected Vermont’s explanation that the law was intended to protect public health and keep health care costs in check, saying that the law could not withstand “heightened scrutiny.” The Supreme Court found that heightened scrutiny was the appropriate standard of review because the Vermont law was a constitutional free speech rights had been violated. Though they lost at the trial court level, the manufacturers won a victory in a 6-3-3 vote in the U.S. Supreme Court.

Writing for the majority, Supreme Court Justice Kennedy held in Sorrell v. IMS Health12 that the manufacturers’ First Amendment rights had been violated. Sorrell struck down the Section 4631(f) of the Vermont law forbidding the sale of prescriber-specific information by pharmacies to “pharmaceutical manufacturers and pharmaceutical marketers.” The Supreme Court rejected Vermont’s explanation that the law was intended to protect public health and keep health care costs in check, saying that the law could not withstand “heightened scrutiny.” The Supreme Court found that heightened scrutiny was the appropriate standard of review because the Vermont law was a content-based — only forbidding the marketing of drugs — and a speaker-based — only silencing pharmaceutical marketers and manufacturers — prohibition on speech. The Supreme Court held that the speech’s commercial nature did not negate the need for heightened scrutiny because “[w]hile the burdened speech results from an economic motive, so too does a great deal of vital expression.”13

In addition to these large scale commercial mining operations, a number of cases have arisen in connection with individuals or companies attempting to engage in data mining on the Internet. These cases often involve efforts to collect customer information off of the Internet through the use of so-called “Web crawlers” or “bots” (short for robots), meaning programs that search through hundreds or thousands of Websites to collect information on product pricing or availability. The same techniques may be used in an effort to collect personal information off of social networking sites, such as Facebook. Such efforts are often prohibited by courts because data mining is contractually prohibited by the Website.14 (Remember the “contract” you clicked okay to when you signed up for the service or entered the Website?)

Data Mining in Academia

So now at last we come to the subject of data mining in academia. We have come the long way round in order to illustrate that the legal issues affecting the topic for today are derivative of — and dwarfed by — the far, far larger context of the use of data mining in law enforcement, national security, and commercial marketing operations. Rules developed by courts or legislatures to deal with fraud, terrorism, or the invasion of privacy may be ill-fitting shoes when applied to the efforts of liberal arts or scientific scholars to find some new angle or insight into their disciplines.

What is an example of academic data mining? Two years ago, Folger Shakespeare Library director Michael Witmore gave a speech describing his use of data-mining methods to analyze Shakespeare’s First Folio.15 Witmore processed 767 different thousand-word excerpts of plays from the First Folio through a piece of software called “DocuScope.” The software is based on a database of 40 million English linguistic patterns sorted into more than 100 categories. Filtering Shakespeare’s classics through DocuScope uncovered patterns in Shakespeare’s work that a human scholar, trained in traditional academic reading methods, would never see. Such as the fact that — in purely linguistic terms — Shakespeare’s Othello is a comedy. Surprising, no? Well, I’ll say no more on the merits of this piece of research at this point. I cite it merely as an example of how data mining might be used in the liberal arts. More to the point is to ask whether this research would have encountered any legal issues that needed to be solved.

The first issue that might arise in text mining is whether there is any copyright problem in obtaining and searching the “database” (here, the First Folio). Maybe, but maybe not. Master Will is of course long dead, and his works have been in the public domain for nearly four hundred years.16 Yet there might be a question about copyright. Why? The answer lies in the particular version of the First Folio the researcher wishes to use. All of the numerous printed editions of Shakespeare’s works through the 19th century are out of copyright, but it would likely be true that any modern, electronic, word-searchable edition is in copyright. Shakespeare’s words are not copyrightable, but the electronic version of them might be. If you go to the trouble of converting an old, out-of-copyright print edition of Shakespeare into a Microsoft Word document or some other e-text, your e-text of it may well be copyrightable. Thus, while there are several Websites which offer a free electronic edition of Shakespeare, you should check to make sure that the version you plan to search is not somehow protected.17

We will need to revisit the copyright question later, but let’s go on with our discussion of the Shakespeare data mining project. So there would seem to be no copyright issue as such. And presumably there is no problem in accessing the non-copyrighted edition of the First Folio. By that I mean that the version you are working with has no electronic lock on it that you would have to unlock (or possibly “circumvent”). So, is there anything else to worry about? Yes. We need to know what search engine or data mining software you are planning to use?

Possibly the software is proprietary. You may need permission to use it. DocuScope, for example, is owned by Carnegie Mellon University, from whom you must obtain permission to access and use the software. I have not seen their form of license agreement, but it may include some requirements relating to publication of research using DocuScope.

Well, that’s a pretty easy set of steps to go through in order to data mine the First Folio. Easy, compared to other situations. The difficulty factor would increase if the contents of the database you want to search is copyrighted. Let’s say you want to apply DocuScope to the collected works of William Faulkner and Ernest Hemingway, both of whose works are — so far as I know — still under copyright. First, you would have to find an electronic database containing their novels and short stories. Let’s assume that you can find all of them in the Google Books Project or HathiTrust Digital Library (“HDL”). The issue of whether Google had a right to scan all the books in the world is still an open question before Judge Chin,18 but the District Court decision in the HathiTrust case19 issued by Judge Baer recognizes that such copying constitutes “fair use” of the works under 17 U.S.C. § 107 because the use for scholarship and research was “transformative” by providing superior search capabilities rather than actual access to the copyrighted works and facilitating access for print-disabled persons.

Judge Baer stated:

Transformative uses are likely to satisfy the first factor. Campbell, 510 U.S. at 575 (“The central purpose of this investigation is to see . . . whether the new work merely superseded[s] the objects of the original creation . . . or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message; it asks, in other words, whether and to what extent the new work is ‘transformative.’”) (internal citations and quotation marks omitted). A transformative use may be one that actually changes the original work. However, a transformative use may also be one that serves an entirely different purpose. Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 609 (2d Cir. 2006) (affirming district court’s conclusion that the use of entire copyrighted concert posters in a book to “document and represent the actual occurrence” of the concerts was different from the “dual purposes of artistic expression and promotion of the original use”). The use to which the works in the HDL are
put is transformative because the copies serve an entirely different purpose than the original works: the purpose is superior search capabilities rather than actual access to copyrighted material. The search capabilities of the HDL have already given rise to new methods of academic inquiry such as text mining.20

In a footnote, the court made the following observation:

22/ Mass digitization allows new areas of non-expressive computational and statistical research, often called “text mining.” One example of text mining is research that compares the frequency with which authors used “is” to refer to the United States rather than “are” over time. See Digital Humanities Amicus Br. 7 (“[i]t was only in the latter half of the Nineteenth Century that the conception of the United States as a single, indivisible entity was reflected in the way a majority of writers referred to the nation.”).21

Thus, if Judge Baer’s opinion is upheld on appeal,22 there is a good argument that a researcher would not be engaging in copyright infringement by mining works that are still in copyright. The harder question is whether the researcher can obtain access to the copyrighted works.

Logically, a researcher only needs — or would want — to engage in text or data mining if, first, there is a large-enough collection of books or articles to make the effort meaningful and, second, the collection of works is machine-readable ... preferably in a common format. Thus, a biochemical researcher might be interested in data mining the past 50 years’ worth of scientific journals but only if they are electronically searchable in the right kind of way. But access is no easy matter. There’s the rub.

To obtain access to all of Elsevier’s past and current journals, for example, may require special permissions and payment of fees that are not inherent in the university’s current subscriptions. Elsevier’s current policy on data mining23 includes the following:

- We wish to understand our customers’ text mining requirements and as practically every content mining request has a different goal there is not a common solution to provide this. Consequently we request that customers looking to mine our content should speak to their Elsevier Account Manager or should contact us directly at <universal.access@elsevier.com>.

- We will then discuss the mining request, access to the content (see below), licensing, and (where applicable) pricing for the project.

- Mining requests are often content specific. Customers can choose to mine our full-text content, abstracts, data, and other materials. A charge may be applicable dependent on the request.

Hold up on the idea of obtaining permissions for a moment, and let’s talk about whether the researcher has the right and ability to obtain access without going down the permission route.

If you already have access to the copyrighted work via Google Books or HathiTrust’s HDL, you may choose to skip seeking permission to data mine. Of course you better check with your university’s lawyers first, but conceptually you might be able to run the search.

But if there are access problems because, for example, “technological protection measures” (TPMs) have been installed on the database to control or limit access, there is more of an issue.24 Assuming you had access to some sort of device or software that could cancel or circumvent the TPM, can you legally use it to do so? The answer to this question is on the outer edge of legal certainty.

The Digital Millennium Copyright Act (“DMCA”) of 1998 includes TPM provisions that ban both acts of circumventing TPMs used by copyright owners to control access to their works, as well as devices, services, or technology that is primarily designed or useful for circumvention. However, the DMCA does include seven limited (and generally inadequate) exceptions including for library acquisitions, security testing, reverse engineering of software, encryption research, and law enforcement. Arguably, the use of TPMs as a practical matter nullifies the ability to make “fair use” of protect ed digital works. The DMCA bans consumers from circumventing TPMs to make fair use of a protected digital work, such as making a back-up copy of a copy-protected CD or DVD that they have purchased. A possible crack in this wall has opened up in the last few years as a result of two decisions by the U.S. Court of Appeals for the Federal Circuit.

Section 1201(a)(1) of title 17 of the United States Code prohibits any person from “circumvent[ing] a technological measure that effectively controls access to a work protected under this title.” The Federal Circuit confronted the issue in Chamberlain Group, Inc. v. Skylok Technologies, Inc. in 2004.25 There, the court noted that, when Congress enacted the DMCA, it “chose to create new causes of action for circumvention and for trafficking in circumvention devices. Congress did not choose to create new property rights.”26 Accordingly, the court held that section 1201 “prohibits only forms of access that bear a reasonable relationship to the protections that the Copyright Act otherwise affords copyright owners.”27 A copyright owner alleging a violation of section 1201(a) consequently must prove that the circumvention of the technological measure either “infringes or facilitates infringing a right protected by the Copyright Act.”28

The following year, the Federal Circuit again dealt with a claim of circumvention under the DMCA and reaffirmed its ruling in Chamber lain. In Storage Technology Corporation v. Custom Hardware Engineering & Consulting, Inc.,29 plaintiff StorageTek claimed that defendant CHE’s use of certain devices to circumvent a data locking protocol constituted a violation the DMCA. (CHE repairs STK libraries, or “silos,” that are connected to Library Control Units.) The Court of Appeals concluded that it was unlikely that StorageTek could succeed on the merits of its copyright claim and therefore “[t]o the extent that CHE’s activities do not constitute copyright infringement or facilitate copyright infringement, StorageTek is foreclosed from maintaining an action under the DMCA.”30

The court held that, “[t]o the extent that StorageTek’s rights under copyright law are not at risk, the DMCA does not create a new source of liability.”31 The Federal Circuit reasoned that, because the DMCA must be read in the context of the Copyright Act, which balances the rights of the copyright owner against the public’s interest in having appropriate access to the work, “courts generally have found a violation of the DMCA only when the alleged access was intertwined with a right protected by the Copyright Act.”32 This holding was followed in a subsequent case involving CHE.33

Five years later, the Court of Appeals for the Ninth Circuit considered a DMCA circumvention claim and declined to follow the Federal Circuit’s approach or to adopt an infringement nexus requirement.34

Without the Ninth Circuit’s ruling, one might have felt fairly confident that the correct reading of § 1201 of the DMCA is that there cannot be a circumvention violation unless it leads to copyright infringement. Thus, one might have concluded that a circumvention to engage in fair use would not be actionable. However, in light of the Ninth Circuit’s contrary decision, such a conclusion would entail a certain amount of risk. Accordingly, the conservative approach for the moment at least would be to avoid circumventing technological protection measures and to opt to seek permission from the owner of the protected works.

Finally, even without the existence of TPMs barring access to a protected work, there may be contractual or licensing constraints on the ability of an academic researcher to freely mine the works otherwise available to him or her. Contractual provisions in journal or database licensing agreements may place restrictions on the user that either expressly or inferentially bar data mining without permission. Copyright law has not been interpreted to preempt or override such contractual restrictions. The university or library holding the license may be deemed to be in breach of the license agreement if it permits data mining (that is prohibited by the agreement). Even “open access” journals may present a problem if the license is a restricted one (such as an “ND” or No-Derivs one). E.g., the need to provide the attributions required by the CC license may be burdensome when examining books or articles are mined.

With license agreements, as with the other aspects of the data mining project such as copyrightability and TPMs, it is important for the researcher to consult closely with the library staff and with the university counsel before lowering oneself too precipitously down the mine.
Rumors

from page 47

librarians in light of all that is happening in our industry. I was especially struck by the truth of

the “big data” takeaway — “Big data skills will become more important in the years to come. Because of the skill gap, professionals are not investing limited resources in content they do not understand.” (See this issue, p.18.) And we libraries have lots of big data — so much in fact that Dennis Brunning says that “we are headed toward data obesity.” (See p.8, 10.)

Speaking of which, the how-wonderful-that-he-is-talking-to-us Jim O’Donnell will be directing a panel in Charleston this year

continued on page 69