Canadian Journal of Law and Technology

Volume 9
Number 1 *1 & 2*

Article 2

6-1-2011

Overcoming the Digital Tsunami in e-Discovery: is Visual Analysis the Answer?

Victoria L. Lemieux

Jason R. Baron

Overcoming the Digital Tsunami in e-Discovery: is Visual Analysis the Answer?*

*Victoria L. Lemieux and Jason R. Baron***

ABSTRACT

New technologies are generating potentially discoverable evidence in electronic form in ever increasing volumes. As a result, traditional techniques of document search and retrieval in pursuit of electronic discovery in litigation are becoming less viable. One potential new technological solution to the e-discovery search and retrieval challenge is Visual Analysis (VA). VA is a technology that combines the computational power of the computer with graphical representations of large datasets to enable interactive analytic capabilities. This article provides an overview of VA technology and how it is being applied in the analysis of e-mail and other electronic documents in the field of e-discovery, as well as discussing several challenges and limitations of the technology. The article concludes that VA has the potential to overcome some of the limitations of current search and retrieval techniques, but that addressing the digital tsunami is more likely to be achieved by using VA in combination with other search and retrieval technologies in the context of creating an effective data governance program.

I. INTRODUCTION: THE E-DISCOVERY PROBLEM

At the beginning of the second decade of the 21st century, the legal profession continues to confront exponentially increasing volumes of data across a spectrum of old and new varieties of electronic applications and formats, ranging from e-mail and traditional word processing, to web pages and even newer forms of web 2.0 social media. It is also not unusual for lawyers to confront the task of having to sift through millions of such files contained on electronic media of all types, from databases and online networked systems, websites, and disaster recovery backup tapes, all for the purpose of performing their searches for relevant evidence.

At the same time, in the midst of this growing technological thicket, the responsibility of the lawyer in civil litigation remains unchanged, namely: to use due diligence and reasonable means to ferret out what constitutes relevant evidence of importance to his or her client. This is in addition to responding to requests for documents made by opposing parties in the lawsuit and preparing the lawsuit for trial (or, as is increasingly the case, for ensuring a fair settlement). In the over-

---

* A version of this paper was originally prepared for a session at the Society of American Archivists' Annual Meeting, Washington, D.C., 10–14 August, 2010.

** Victoria L. Lemieux, School of Library, Archival and Information Studies, University of British Columbia, vlemieux@mail.ubc.ca. Jason R. Baron, Office of the General Counsel, National Archives and Records Administration, jason.baron@nara.gov.

whelming majority of cases, legal discovery demands still consist of formal requests that in the main expressly ask for some variation of the phrasing, "*provide any and all*" documents relevant to the specific issue(s) being litigated. But faced with the prospect of reviewing terabytes of information and beyond,[1] do lawyers use optimal means for actually performing their e-discovery searches in response to such requests? Put another way, do lawyers understand the limitations of present-day search methods or what alternatives may exist that would be useful in supplementing their existing practice?

To be sure, in various jurisdictions new rules of civil procedure have been adopted ensuring that lawyers and judges are on notice of their obligations to produce documentary evidence in electronic form. For example, as of January 1, 2010, Ontario amended its Civil Rules of Procedure to provide for "Principles re Electronic Discovery," in the form of Rule 29.1.03(4) which states:

> (4) In preparing the discovery plan, the parties shall consult and have regard to the document titled "The Sedona Canada Principles Addressing Electronic Discovery" developed by and available from The Sedona Conference.[2]

Previously, in December 2006 the U.S. Federal Rules of Civil Procedure were amended to expressly incorporate the term "electronically stored information." and to require that lawyers representing all parties to proceedings meet and confer early on in litigation to hammer out how electronic evidence will be preserved, formatted, and accessed. An increasing number of judges are on record as expecting that as part of these early on discussions counsel will discuss if not begin negotiating over what constitutes an adequate "search protocol" for each side to adhere to as part of the discovery process.[3] For a discussion of evolving Canadian e-discovery law, see (Force 2010).[4]

In light of all of these developments, there is growing recognition in many quarters of the legal profession that a need exists to ask the questions posed here, namely, about the limitations of present day searches and alternatives to the status quo ante in how the legal obligations are met. Given the avalanche of electronic data that lawyers confront on a daily basis, the profession is ripe for considering

---

1 See generally George L. Paul & Jason R. Baron, "Information Inflation: Can the Legal System Adapt?" (2007) 13 Rich JL & Tech 10, online: <http://law.richmond.edu/jolt/v13i3/article10.pdf>; Jason R. Baron & Ralph C. Losey, "e-Discovery: Did you know?" (2010), Multimedia Presentation, online: YouTube <http://www.youtube.com/watch?v=bWbJWcsPp1M>.

2 *Rules of Civil Procedure*, RRO 1990, Reg. 194, s. 29.1.03(4), online: <http://www.e-laws.gov.on.ca/html/regs/english/elaws_regs_900194_e.htm#ys29p1p04>. The Sedona Conference is a not for profit legal think tank that has published numerous commentaries on e-discovery, including the referenced paper that can be found online: <www.thesedonaconference.org/publications>.

3 See generally Jason R. Baron & Edward C. Wolfe, "A Nutshell on Negotiating E-Discovery Search Protocols" (2010) 11 The Sedona Conference Journal 229 [Baron, "Nutshell"].

4 Donald Force, "From Peruvian Guano to Electronic Records: Canadian E-Discovery and Records Professionals" (2010) 69 Archivaria 49, online: <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13261>.

new ways of interpreting and making sense out of vast quantities of data on a scale never previously confronted.

II. IS TECHNOLOGY THE SOLUTION?

As stated in a report by PwC UK entitled *e-Disclosure 2020*, far from providing a panacea, technology seems to create problems faster than it solves them, or at a minimum, the solutions seem destined to lag behind. Easy information creation and inexpensive storage appears long before tools to organise or catalogue that information.[5] Indeed, we know that current e-discovery search methods are not sufficient to overcome the digital tsunami: the most common methods currently used in e-discovery — keyword searching and linear review — are increasingly ineffective for the massive volumes of data that must be sifted through for each case. There have been a number of studies highlighting the limitations of existing search and retrieval techniques. In one study lawyers overestimated the effectiveness of their keyword-based search strategies by as much as 55%.[6] Dabney (1986), Bing (1987) and Schweighofer (1999) all provide in-depth reviews of the limitations of full text searching for legal documentation.[7] More recently, a multi-year study evaluating the efficacy of various search methods known as the "TREC Legal Track" demonstrated that traditional Boolean search methods failed to find up to 78% of relevant documents that other automated search methods accounted for (Tomlinson et al, 2008).[8]

In 2007, The Sedona Conference®, a leading legal non-profit think tank, issued a Commentary where the limitations of keyword searching were comprehensively described. In relevant part:

> Keyword searches work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all

---

5 PricewaterhouseCoopers, *E-Disclosure 2020 — Creating a Strategic Framework for the Future*London, PwC, 2010 [PwC, "E-Disclosure 2020"], online: http://www.pwcwebcast.co.uk/pwc_uk_edisclosure_2020.pdf.

6 David C. Blair & M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document retrieval system" (1985) 28:3 Communications of the ACM 289.

7 Daniel P. Dabney, "The Curse of Thamus: Full Text Legal Document Retrieval" (1986) 78:5 Law Libr J 5; Jon Bing, "Performance of Legal Text Retrieval Systems: the Curse of Boole" (1987) 79 Law Libr J 187 [Bing, "Curse of Boole"]; Erich Schweighofer, "The Revolution in Legal Information Retrieval or: The Empire Strikes Back" (1999) JILT, online:

<http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/1999_1/schweighofer>.

8 S. Tomlinson et al, "Overview of the TREC 2007 Legal Track", in *The Sixteenth Text Retrieval Conference (TREC 2007) Proceedings*, online:

<http://trec.nist.gov/pubs/trec16/t16_proceedings.html>.

> other languages).

(The Sedona Conference 2007).[9]

All of these prior reports and studies are in line with results of an online survey of legal and technical professionals in the UK and two roundtable discussions on e-discovery conducted by PwC indicating that keyword searching is increasingly untenable.[10] Panellists noted the difficulties of choosing key words, reporting that "[e]ven if you have a brilliant, absolutely focussed search, you are still going to end up with too many documents to review and within those there will still be a very large proportion of irrelevant material."[11] Data volumes are quickly becoming such that even with the best keyword search terms and an army of reviewers, it could still take months or years to sift through all the data and there would still be no guarantee of satisfactory results.[12] New approaches are therefore very much needed.

Some leading thinkers in the legal profession have recognized that alternatives to traditional keyword searching do exist, and, in spite of the dearth of reported opinions on the subject, many practitioners are incorporating a variety of new search techniques in their everyday practice. In the one reported decision that does go into alternative search techniques at some length, *Victor Stanley v. Creative Pipe*,[13] Judge Grimm, citing to (The Sedona Conference 2007), recognized that in addition to keyword searches, "other search and information retrieval methodologies" exist such as

> probabilistic search models, including "Bayesian classifiers" (which searches by creating a formula based on values assigned to particular words based on their interrelationships, proximity, and frequency to establish a relevancy ranking that is applied to each document searched); "Fuzzy Search Models" (which attempt to refine a search beyond specific words, recognizing that words can have multiple forms. By identifying the "core" for a word the fuzzy search can retrieve documents containing all forms of the target word); "Clustering" searches (searches of documents by grouping them by similarity of content, for example, the presence of a series of same or similar words that are found in multiple documents); and "Concept and Categorization Tools" (search systems that rely on a thesaurus to capture documents which use alternative ways to express the same thought).[14]

---

9 "The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery" (2007) 8 The Sedona Conference Journal 189.

10 PwC, "E-Disclosure 2020", *supra* note 5.

11 PwC, "E-Disclosure 2020", *supra* note 5 at 23.

12 See Examiner's Report by Anton R. ValukasLehman Brothers Holdings Inc Chapter 11 ProceedingsVol. 7, Appx. 5 (reporting that out of a universe of 350 billion pages — three petabytes — the Examiner narrowed the collection by selecting key custodians and using dozens of separate Boolean searches to collect for review in excess of five million documents, representing 40 million pages, for two further levels of manual review by 70 contract attorneys supplemented by others), online: <http://lehmanreport.jenner.com/>.

13 250 FRD 251 (D Md 2008).

14 *Ibid* at 259 n9.

In line with *Victor Stanley's* recognition of concept and categorization methods, it has been reported that the "use of auto-categorization systems can potentially reduce document request times from over four months to as little as thirty days for even the largest datasets" (Oot et al, 2010).[15]

These and other candidate technological solutions (including near de-duplication, e-mail threading, and predictive coding,[16]) will be deployed on a more widespread basis in litigation in the future, as lawyers increasingly become convinced that efficiencies in cost and scale are achieved through their greater use.

III. WHAT ABOUT VISUAL ANALYSIS?

A technology emerging as one possible solution to the e-discovery problem is visual analysis (VA). VA is a relatively new technology that combines analytical reasoning facilitated by interactive visual interfaces. It has been described as:

- The science of analytical reasoning facilitated by interactive visual interfaces.[17]
- More than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis.[18]
- The formation of abstract visual metaphors in combination with a human information discourse (usually some form of interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces.[19]

Though VA is a dynamic process not a static one, it has its origin in the static visual representation of data. The field of visualization is, of course, very old: maps, graphs and charts have been in use for many centuries. Computer-assisted visualization is of more recent vintage. There is growing research interest reflecting the many advantages that visualization is said to offer over traditional textual representations: increased cognitive resources, such as by using a visual resource to expand human working memory; reduction of search, such as by representing a large amount of data in a small space; enhanced recognition of patterns, such as when

---

15 P. Oot, A. Kershaw & H.L. Roitblat, "Mandating Reasonableness in a Reasonable Inquiry" (2010) 87 Denv UL Rev 533 at 551. See also Ronni D. Solomon & Jason R. Baron, "Bake Offs, Demos & Kicking the Tires: A Practical Litigator's Brief Guide to Evaluating Early Case Assessment Software & Search and Review Tools" (The Sedona Conference, 2009), online:

<http://www.kslaw.com/Library/publication/BakeOffs_Solomon.pdf>.

16 PwC, "E-Disclosure 2020", *supra* note 5.

17 J.J. Thomas & K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, (Los Alamitos, CA: IEEE Computer Society Press, 2005), online: <http://nvac.pnl.gov/agenda.stm> [Thomas, "Illuminating the Path"].

18 D.A. Keim et al, "Challenges in Visual Data Analysis" in *Proceedings of Information Visualization, 2006. IV 2006. Tenth International Conference on* (2006) at 9, online: IEEE <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1648235>.

19 Kris Cook, Rae Earnshaw & John Stasko, "Guest Editors' Introduction: Discovering the Unexpected, Computer Graphics and Applications" (2007) 27 IEEE Computer Graphics and Applications 15.

data are organized spatially to indicate their temporal relationships; and support for the easy perceptual inference of relationships that are otherwise more difficult to induce.[20]

Over time, visualization research has expanded beyond static visual representations to applications that allow users to interact with visualizations in an exploratory fashion. VA tools apply visualizations such as bar charts, node link diagrams, clusters or time lines to represent the results of computer analysis (see Figure 1).[21] Users are then able to interact with the visual representations to conduct further analysis in an iterative manner. The results of these additional analytical steps also are presented as visual metaphors or representations that can be further analysed if desired. VA tools have been designed specifically to deal with:

- Processing masses of dynamic data
- Answering an array of often ambiguous questions
- Keeping humans in the loop and at the centre of analysis
- Blending computational analysis with interactive visualization of the results of that analysis
- Providing quick answers with on demand improvement of analytical results
- Incorporating presentation linked with analysis
- Exporting easy to understand representations of results in real time[22]

VA has emerged as a multi-disciplinary field (see Figure 1)[23] in which researchers from computer science, visual arts, cognitive psychology and other domain-specific areas collaborate to explore the basic components of the technology: analytical reasoning techniques; visual representation and interactions; data representations and transformations; and techniques to support production, presentation, and dissemination.[24]

VA's original domain of application was in science but it has now moved into other areas, such as biology; business intelligence; fraud detection; and epidemiology.[25] The kinds of questions and issues that have attracted other domains of analysis to this technology are similar to the types of questions and issues faced in e-discovery.

---

20 Thomas, "Illuminating the Path", *supra* note 17; Daniel T. Gilbert, *Stumbling on Happiness* (New York: Knopf, 2006) at 99; Ron Rensink, "Change Detection" (2002) 53 *Annual Review of Psychology* 245.

21 As it was not possible to include colour graphics in the text of this article and colour is an important feature of many visualizations, including those selected to illustrate points in this article, the authors have chosen to make all figures available online at http://www.ciferresearch.org/research/current_project&pid=25>.

22 VisMaster, "Mastering the Information Age", Video production, (France: INRIA, 2010), online: YouTube <http://www.youtube.com/watch?v=5i3xbitEVfs> [VisMaster video].

23 Op. cit. note 21.

24 Thomas, "Illuminating the Path", *supra* note 17.

25 VisMaster video, *supra* note 22.

IV. VISUAL ANALYSIS AS APPLIED TO E-DISCOVERY

VA tools are relatively new to the e-discovery domain. To date, most of the tools have focussed on the analysis of e-mails, though some have incorporated other documentary forms as well. Prior to the emergence of VA, visualization has been used to analyze both personal e-mail collections and public archives of threaded discussions.[26] For example, TimeStore employed a redesign of the e-mail inbox and filing system which automatically arranged e-mail in a two dimensional grid, with one axis being people and the other, time.[27] Sack studies e-mail exchanges between thousands of Usenet users and in a 2001 paper discusses the "Conversation Map" system, a visual system for browsing and navigating very large-scale conversations.[28] Other researchers have applied Treemaps in the Netscan project to visualize Usenet postings.[29] Re-mail provided overviews of correspondents and messages to help spot those with similar attributes.[30] PostHistory was a personal information management tool for using timelines and contact overviews to help generate insights that would be socially relevant to the owner of an e-mail collection.[31] Social-Network Fragments complemented that approach, identifying communication clusters in social networks of authors.[32] There has also been work on showing the structure of threaded e-mail conversations over time.[33]

---

26 Hyunmo Kang et al, "Making Sense of Archived E-mail: Exploring the Enron Collection with NetLens" (2008) 61 *Journal of the American Society for Information Science and Technology* 723 [Kang, "Making Sense"].

27 Yiu, Kelvin S et al, "A Time-based interface for electronic mail and task management" In Design of Computing Systems: Proceedings of HCI International '97. Vol. 2. Elsevier, 19–22.

28 Warren Sack, "Conversation Map: An Interface for Very Large-Scale Conversations" (Winter 2000/01) 17 *Journal of Management Information Systems* 73 [Sack, "Conversation Map"].

29 Marc Smith & Andrew Fiore, "Visualization components for persistent conversations" in *Proceedings of the ACM CHI '01 Human Factors in Computing Systems Conference* (New York: ACM Press, 2001) 136.

30 S.L. Rohall et al, "Re-mail: A reinvented e-mail prototype (demonstration)" In Proceedings of the SIGCHI conference on Human Factors in Computing Systems: CHI '04 Extended Abstracts on Human Factors in Computing System (New York: ACM Press, 2004) 791-792.

31 F. Viegas et al, "Digital Artifacts for Remembering and Storytelling: *PostHistory* and *Social Network Fragments*" In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences* (Washington, DC: IEEE, 2004) 109.

32 *Ibid*.

33 Bernard Kerr, "Thread arcs: An Email Thread Visualization" In *2003 IEEE Symposium on Information Visualization*: 27, online:

<http://www.computer.org/portal/web/csdl/doi/10.1109/INFVIS.2003.1249028>; Gina Venolia & Carmen Neustaedter, "Understanding sequence and reply relationships within e-mail conversations: A mixed-model visualization" In Gilbert Cockton and Panu Korhonen, eds, *Proceedings of the ACM CHI 2003 Human Factors in Computing Systems Conference April 5–10, 2003, Ft. Lauderdale, Florida, USA* (New York: ACM Press, 2003) 361.

E-mail visualization has recently started to receive some attention from researchers seeking to support retrospective analysis as well. For example, Perer, Shneiderman and Oard used timelines to study the rhythms in e-mail use over time with a case study based on 20 years of a single individual's e-mail.[34] Perer and Smith looked at three simple visualizations of e-mails to construct portraits of e-mail practices, and collected feedback from eight users looking at their own e-mail store.[35] The eArchivarius project showed clusters based on content or co-addressing, along with timelines and biographies to explore a small set of government e-mails that had been released in response to *Freedom of Information Act* requests.[36]

Visual analytics goes one step beyond these efforts by incorporating capabilities that permit user directed exploration. An interactive graph-representation tool was developed by Heer to explore the social network of correspondents for the half-million-document Enron collection for example.[37]

A review of the research literature indicates that visual analysis functionality (and associated tools) as applied in the domain of e-mail analysis can be classed broadly into two categories: 1) Tools that represent communications patterns and 2) tools that represent content. Of the tools that visualize communications patterns, there are two sub-categories: 1) those that represent time lines of communications and 2) those that visualize social networks.[38] Tools that focus on social networks analysis often visually represent analytic results using node link or network diagrams as these types of diagrams lend themselves well to representing social actors and relationships between them. Typically, though not exclusively, these types of visualizations are generated on data drawn from metadata such as the header of an e-mail or, in some cases, user supplied data.[39] Tools that focus on content, on the other hand, can use traditional keyword search techniques, for example, NetLens E-mail or vector space clustering algorithms that facilitate clustering of data into "gal-

---

34 Adam Perer & Ben Shneiderman, "Beyond threads: Identifying discussions in Email archives" *2005 IEEE Information Visualization*, 41-42; Adam Perer, Ben Shneiderman & Douglas W. Oard, "Using rhythms of relationships to understand e-mail archives" (2006) 57 *Journal of the American Society of Information Science and Technology* 1936 [Perer, "Using rhythms"]; Adam Perer & M. Smith, "Contrasting Portraits of E-mail Practice: Visual Approaches to Reflection and Analysis" In *Proceedings of Conference on Advanced Visual Interfaces (AVI)* (Venezia, Italy: ACM Press, 2006) 389–395.

35 *Ibid*.

36 A. Leuski, D. Oard, & R. Bhagat, R., "eArchivarius: Accessing collections of electronic mail" [Demonstration] In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 468). (New York: ACM Press, 2003) at 468. The e-Archivarius Project website is at: http://people.ict.usc.edu/~leuski/projects/earchivarius/.

37 Jeffrey Heer, "Exploring Enron: Visual Data Mining of E-mail" (2005), online: <http://jheer.org/enron>.

38 For a good overview of approaches to visualization of e-mail, see Judith Donath, "Visualizing E-mail Archives", online:

<http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf>.

39 Sack, "Conversation Map", *supra* note 28 is an example of social network analysis that draws on semantic analysis of content.

axies", "spires", and "bubbles" conceptualizing the topics and entities that exist in the documents of interest.[40] In other words, automatically generating an ontology based on an analysis of the content of a particular data set.[41] Some tools, such as NetLens E-mail, combine elements of all these forms of visualization. While VA research and development in the domain of e-discovery has focused primarily on the problem of searching for and retrieving e-mail documents, some studies incorporate a wider range of material. Kang et al describe incorporating transcriptions of voice messages into their analysis of Enron documentation and discuss the possibility of incorporating content drawn from external sources found on the Web such as biographies of persons of interest.[42] Esteva et al have conducted research using a collection of unstructured Spanish documents varying in length and size that approximates the type of documentation that might be discoverable in litigation.[43]

A number of technology vendors offer visual analytics solutions. Without any intent to endorse particular products or services,[44] we see these as being grouped into three broad categories: 1) those that are legal discovery "pure play" technology companies (e.g., MetaLINCS and Attenex), 2) those that offer more general search or data management software and services (e.g., Autonomy) and have incorporated visualization into their offerings, and 3) "pure play" visual analytics tools that can be used to conduct discovery but also used for other purposes (e.g., Starlite).[45] These tools range from those that offer visualization as an "add on" mode of displaying search results, which in itself can be extremely helpful to reviewers, to those that offer "true" visual analytic capabilities — the ability to have a human analyst interact with a visualization in order to really delve into data to structure, organise, and, ultimately, understand them.

---

40 Kang, "Making Sense", *supra* note 26. See also Starlite, online:<http://www.futurepointsystems.com> and IN-SPIRE, online: <http://inspire.pnl.gov>. Both use clustering algorithms to group the content of documents by semantic similarities.

41 For an overview on the subject of ontologies, see Tom Gruber, *"Ontology"* in Ling Liu & M. Tamer Özsu, eds, *The Encyclopedia of Database Systems* (New York: Springer-Verlag, 2008) 1963, online: <http://tomgruber.org/writing/ontology-definition-2007.htm.

42 Kang, "Making Sense", *supra* note 26.

43 Esteva, Maria et al. "Finding narratives of activities through archival bond in electronically stored information (ESI)" (Paper delivered at the 2009 Society of American Archivists' Research Forum, Austin, Texas, August 2009) [Esteva, "Finding narratives"].

44 Examples given of goods and services in the e-discovery marketplace are simply for illustrative purposes only, and no endorsement by the authors is either expressed or implied.

45 For information about MetaLINCS see "Enron Emails NowAvailable for Public Analysis: MetaLINCS E-Discovery Software Allows Anyone to Explore Enron's Email Secrets Ahead of Landmark Trial", online: The Free Library <http://www.thefreelibrary.com/Enron+E-mails+Now+Available+for+Public+Analysis%3B+MetaLINCS+E-Discovery. . .-a0140996677>. For Attenex see online: <http://www.ftitechnology.com/products/attenex_patterns.aspx>. For Autonomy see online: <http://www.zantaz.com/products/electronic-discovery/index.htm>. For Starlite, see *supra* note 40.

There are an increasing number of companies who offer visual analysis services as well, so organisations who want to experiment with visual analytics can choose whether to "buy" or "let" this technology. The decision to invest in acquiring the technology in-house or whether to buy it as a service will rest, in part, on the scale of legal discovery requirements: large corporate organisations carrying out discovery on a regular basis will benefit most from acquiring the technology in-house and integrating it with their existing litigation and discovery support tools, while those involved in discovery less often might be better offer availing themselves of a service-oriented offering.

V. WHY COULD VISUAL ANALYSIS BE AN EFFECTIVE E-DISCOVERY TOOL?

In 2006, Forrester Research predicted that visual analytics was going to be "the next big thing" in e-discovery.[46] Why? One reason is that with traditional tools of analysis, reviewers face a long process of culling and de-duplicating large volumes of electronic datasets to produce relevant documents or e-mails, with associated metadata, stored in their native form, or in some normalized form such as EDRM XML, tiff, or pdf. The documents are then indexed and made available for search using standard search query types such as: term queries; phrase queries; near queries; range queries; wildcard queries; fuzzy queries and Boolean queries. One limitation of this approach is that the reviewer may not be certain at the start of a case of the important topics, key players, significant dates, or specific vocabulary in use. She may think she has an idea of what might be important and so try out a few simple queries to see what she can find, but her results may return zero finds. She may give up faced with the impossibility of searching for "unknown unknowns." Belkin et al. recognized this problem in information seeking, stating that "in general the user is unable to specify precisely what is needed."[47]

Attfield, De Gabrielle and Blandford have recently written about the ineffectiveness of traditional information retrieval approaches in the e-discovery domain. Drawing on data from case-studies of e-discovery, these researchers found that document reviewers would benefit from support in drawing together emergent document classes — groups of related irrelevant documents and groups of related relevant documents — which the reviewer becomes aware of during the review task. They note that many traditional review systems fail to assist the reviewer in this and so adversely affect cognitive momentum and the efficiency and effectiveness of the task. They conclude that interactive information visualisations provide new opportunities to move beyond such limitations.[48] VA tools, in providing a visual

---

46 Barry Murphy, "Believe it — eDiscovery Technology Spending to Top $4.8 Billion by 2011" (Cambrdge, MA: Forrester Research, Inc., 2006).

47 N.J. Belkin, R.N. Oddy & H.M. Brooks, "ASK for Information Retrieval: Part 1. Background and Theory" (1982) 38:2 *Journal of Documentation* 61.

48 Simon Attfield, Stephen De Gabrielle & Ann Blandford, "The Loneliness of the Long-Distance Document Reviewer: E-Discovery and Cognitive Ergonomics. *DESI III Global E-Discover/E-Disclosure Workshop: 12th International Conference on Artificial Intelligence and Law*. Barcelona, Spain. Available online at <http://www.law.pitt.edu/DESI3_Workshop/DESI_III_papers.htm>.

representation of concepts and their interrelationships in a domain — have been shown to be extremely helpful to people who need to learn about a domain.[49] Some researchers go beyond even this claim to suggest that interactive visualization tools can prompt reviewers to think more creatively.[50]

The School of Library, Archival and Information Studies (SLAIS) at the University of British Columbia (UBC) has been experimenting with VA tools specifically designed to overcome this cognitive barrier by applying the computational power of computers to analyze large datasets in order to "see the unseen." Put simply, with the kinds of tools being used at UBC-SLAIS the reviewer does not need to know in advance what he is searching for. These VA tools use vector space clustering algorithms to analyze a dataset and present the entire "universe" of data found within the chosen dataset in the form of a visual representation (e.g., nodes, galaxies, spires).[51] The capability of these tools surmounts a problem that often occurs with keyword searching where documents can fall through the cracks of a search. These types of VA tools see all documents and cluster them together, so no document is missed out accidentally. A further advantage is that the reviewer is then able to engage interactively with the visualization to discover more about the cluster of documents its represents in order to test hypotheses and develop new ones.

How do VA tools such as those employed in the UBC-SLAIS experiments work? The reviewer begins by preparing the dataset for input into the VA tool. This usually involves some amount of pre-processing, such as de-duplication and XML tagging. Once the dataset is loaded, analysis can begin. In the vector space model of analysis, documents are represented as vectors of the descriptors that are employed.[52] A vector corresponds to the number of the words in the body of text. Every attribute can be weighted according to its importance. In the simplest case, the attribute receives value 1 if the descriptor occurs and 0, if this is not the case. The similarity between two vectors is usually computed as a function of the number of qualities which are common to both objects. A ranking of documents is inherent to the vector space model. Information representation is done according to similarity to the search vector. Similar documents can be combined into clusters through matching the search vectors with the central vector.

Once the initial clusters have formed, a process of iterative visual analysis of the visualizations to validate hypotheses or discover emergent, unexpected patterns is possible. The reviewer focuses on clusters of interest by selecting a cluster and

---

49 Ozgur Turetken & Ramesh Sharda, "Visualization of Web Spaces: State of the Art and Future Directions" (2007) 38:3 *Database for Advances in Information Systems* 51; Richard E. Mayer, *Multimedia Learning* (New York: University Press, 2001).

50 Seyour Papert, *The Children's Machine: Rethinking School in the Age of the Computer* (New York: Basic Books, 1993).

51 Two tools have been used in research conducted by UBC-SLAIS to date: 1) Starlite <http://www.futurepointsystems.com> and 2) IN-SPIRE; see *supra* note 40. Both use clustering algorithms to group the content of documents by semantic similarities.

52 Gerard Salton, *The SMART Retrieval System — Experiments in Automatic Document Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1971); Bing, "Curse of Boole", *supra* note 7 at 165.

dropping everything else out of the calculations (e.g., making some words temporarily into outliers) and then evaluating and corroborating (or modifying by labelling) the key topics of the cluster, but also finding peculiar words for the given cluster. For example, a number of iterations of the visualizations might be created by eliminating terms from the visual clusters that were either too specific (e.g., "ISBN") or two general (e.g., "criminal") to produce meaningful clusters of conceptually related data. Using this technology with an iterative analytic technique, the reviewer is able to hone in quickly on the key topics in a large data set, rather than having to guess what might be covered and generate key words that may or may not reveal relevant items. The reviewer also is able to interrogate the results of the initial search to further analyze the data set by, for example, testing out competing hypotheses about the words referred to in the records.

VI. THE CHALLENGES OF RELYING ON VISUAL ANALYSIS

As promising a technology as visual analysis is, there are still many barriers to full adoption and reliance on it to overcome the e-discovery challenge. The following sections discuss a number of these challenges.

(a) Ease of Use

SLAIS researchers have had the opportunity to work with a number of VA tools and have found that the technical skills needed to learn and operate the software effectively are still significantly more than one could expect of a casual user. In other words, it takes training and regular use of the tools to yield effective results. It would be difficult for someone — an attorney or records professional — who did not use VA frequently to use the technology now and again with good results. On the other hand, good results are difficult to achieve if the reviewer only has technical knowledge of how to use the software and is not a domain expert because of the iterative analytical process involved in visual analysis. The exploratory path is not predetermined but emerges as the reviewer discovers new insights from each visual representation — insights that a domain expert is more capable of developing. This is in contrast to a study done by Efthimiadis and Hotchkiss on more traditional search techniques.[53] In this study the researchers ran a test to discover whether domain expertise is necessary in legal researchers. They found that a group of non-experts outperformed a group of experts.

53 Efthimis Efthimiadis & Mary A. Hotchkiss, "Legal Discovery: Does Domain Expertise Matter?" In proceedings of the American Society for Information Science and Technology 45,1 (2008): 1-2 [Efthimiadis, "Legal Discovery"].

To address the need for domain expertise in the use of VA, UBC-SLAIS researchers have found a process called "pair analysis" very effective.[54] In pair analysis a technical expert and a domain specialist work together with the dataset to conduct the visual analysis. It may be that in future, improved user interfaces will make it unnecessary to employ pair analysis, but for the time being this method can be used to improve the effectiveness of the process.

(b) Analytical Reasoning Techniques

One of the challenges of using visual analysis tools is that there is no one predefined pathway or protocol for the process of analytical reasoning to be followed to interact with a dataset. For example, Kang et al describe significant differences in the process of sense-making between an experienced analyst and an archivist in interacting with VA tools and data.[55] It is entirely up to the reviewer to follow hunches and lines of reasoning. It can also be quite difficult to know when there is nothing more to be gleaned by further interrogating or tinkering with a visual representation. As Perer and Schneiderman observe, interactive techniques can yield valuable discoveries, but current data analysis tools typically support only opportunistic exploration that may be inefficient and incomplete. These two researchers propose a methodology of interaction with visualizations that they label "Systematic Yet Flexible Discovery."[56] Other methods, such as Analysis of Competing Hypothesis (ACH) may also be used. Each method is likely to yield quite varied results.[57] Where the analytical process is unique, if not idiosyncratic, questions may arise about whether two reviewers working with the same dataset could produce the same results. Such questions potentially could add further delays or expense to the e-discovery process if opposing parties challenge VA results.

(c) Data Input

The task of preparing data for analysis using VA software is not an insignificant one at this point either. Datasets come in many formats and documents have diverse structures and lengths; nevertheless, they must be rendered in a form that can be read by a VA tool. It can still take a significant amount of effort, as the UBC-SLAIS research team has discovered, to de-duplicate, cleanse, and structure large quantities of data in order to prepare them for analysis. So, better tools and techniques will be needed to import data in diverse formats into VA tools.

---

54 R. Arias-Hernandez et al, "Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics," *Proceedings of Hawai'I International Conference on System Sciences 44*, January 2011, Koloa, Hawaii (2011).

55 Kang, "Making Sense", *supra* note 26.

56 Adam Perer & Ben Shneiderman, "Systematic Yet Flexible Discovery: Guiding Domain Experts Through Exploratory Data Analysis" in *Proceedings of the International Conference on Intelligent User Interfaces*, 2008: 1–10.

57 ACH is an eight-step procedure grounded in basic insights from cognitive psychology, decision analysis, and the scientific *method* and is a tool to aid judgment on issues requiring weighing of alternative explanations or conclusions. See Charles Gettys et al, *Hypothesis Generation: A Final Report on Three Years of Research*, (University of Oklahoma, Decision Processes Laboratory, 1980).

An additional related challenge is that an increasing amount of the data that might be of interest in an investigation or litigation is non-textual — voice-mail and video clips, for example. Speech recognition software exists to translate speech into text to enable analysis, but tools to render images are still at the development stage.[58]

(d) Visual Representation and Interactions

A number of VA tools use vector space algorithms to cluster terms into groups according to the relative occurrence or co-occurrence of key terms. There are algorithms that cluster according to a "bag of words" approach and those that cluster according to a "phrase-based" approach.[59] Each Algorithm performs differently, has an underlying logic and may have weaknesses or limitations that it is important to understand. For example, some algorithms are better than others at maintaining clusters when new documents are incrementally added to datasets.[60] While it is important to understand how different algorithms function, it may not be easily done when a commercially-available VA tool is used: such algorithms tend to be trade secrets and vendors may be reluctant to be too transparent about their structure and how they function.

VA tools still do not deal with the issue of context well. As Esteva et al. point out in their study on discovering trails of documents relating to the same business function in a large collection of unstructured heterogeneous documents, it may be very challenging "to find documents related to an activity that may encompass different document types and writing styles, and include various sub-topics. These documents may differ in length and therefore contain varied keyword frequencies particular to the topic of interest. Moreover, keywords related to the topic may vary, depending on the writing style of the different authors involved. In addition,

---

58 *Ibid*.

59 In the "bag of words" approach to text retrieval a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order. A common use of this approach is in spam filtering where one "bag" contains words identified with spam and the other "bag" contains legitimate e-mails. Using bayesian statistical analysis, the spam filter determines which bag an e-mail is likely to fit into. In a phrase-based approach, a retrieval system will use phrases to index, search and retrieve documents. The system would look at whether the use of a phrase is statistically significant and how often certain phrases appear in use with other phrases. Returning to the spam filtering example, phrases are identified that predict the presence of other phrases in documents. Documents are then indexed according to their included phrases. A spam document is identified based on the number of related phrases included in a document. For more information, see Bill Slawsky, "Phrased Based Information Retrieval and Spam Detection", online: SEO by the Sea <http://www.seobythesea.com/?p=413>.and David Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval" *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998) 4–15.

60 Niall Rooney, et al, "An Investigation into the Stability of Statistical Document Clustering" (2008) 59:2 *Journal of the American Society of Information Science and Technology* 256.

records that treat similar topics but are not directly associated with the activity of interest may introduce noise in the trail.[61]" Depending on the algorithm, terms referring to a project code-named "Orpheus" may cluster with documents referencing Greek mythology and the story of Orpheus. Without explicit tagging, documents generated in the course of the same business function or project but which discuss a variety of unrelated matters are unlikely to cluster together. The problems become compounded as the size of the dataset increases.[62] VA software tools that UBC-SLAIS researchers have been using in the lab alleviate this problem somewhat by clustering groups of co-occurring terms (phrase-based clustering), thereby providing more context than would be achieved by clustering according to a single word at a time. Reviewers also are able to choose to view one, two, three or more terms representing each cluster to gain a better sense of the significance of the cluster. However, there is no question that the issue of context and the ability to discover linkages between documents related to particular functions generates many outliers in a cluster and remains a challenge.

Finally, visual representations represent data differently. For example, a "galaxy view" in the UBC lab (see Figure 2)[63] indicates relative relationship between documents in a dataset, but does not provide much information about the density of a cluster (i.e., how many documents fall within a particular grouping). In contrast, another tool offers a "Spire" visual (see Figure 3)[64] that answers both questions. Visual representations and can invite divergent interpretations of the same dataset.[65] The effect of algorithms used to optimize the visual layouts may lead to misleading interpretations of the data.[66] Just as with an optical illusion, it is possible to see different things and extract diverging meanings from an image. In some cases, where the visualization is very dense (e.g. visualizations of networks that look like "hairballs"), it may be impossible to draw inferences from the diagram at all.[67] VA is a scientific technique, but not an exact or uncomplicated one; as such, it still involves a fair degree of art in the analysis and interpretation of visual representations of data. Further, as with any image, visual representations can be altered or enhanced using software tools such as Photoshop to emphasize or de-emphasize aspects of the results depending on the objectives. This has obvious implications if the intention is to use the results of a visual analysis in legal proceedings.[68]

---

61 Esteva, "Finding narratives", *supra* note 43.

62 *Ibid*.

63 Op. Cit. note 21.

64 Op. Cit. note 21.

65 Edward R. Tufte has written extensively on the subject of visual perception and system design. See e.g. Edward R. Tufte, *Envisioning Information* (Cheshire, CT: Graphics Press, 1990).

66 For a discussion of this problem see the work of Martin Kryzwinski at http://www.hiveplot.net/.

67 *Ibid*.

68 Perceptual and legal issues associated with VA artefacts have much in common with similar issues arising from the use of photographs and other visual representations. See Rodney G.S. Carter, "'Ocular Proof': Photographs as Legal Evidence" (2010) 69 Archivaria 23, and also William J. Mitchell, *The Reconfigured Eye*: *Visual Truth in the*

The effective use of an appropriate visual representation requires designers to ensure that reviewers understand how to use, and the implications of using, a particular visual representation. A "one size fits all" approach to visual representations is unlikely to work well. Reviewers will need to understand the relationship between visual representations and interface functionality; determine the best visual representations for particular tasks and platforms; the ways visual representations might be merged effectively to support e-discovery needs; develop guidelines for the application of visual representations and associated knowledge for practical implementation; and consider the ways in which visual representations might be admissible in court (for example, what kind of oral testimony might be needed to attest to the authenticity and integrity of visual representations.).[69]

(e) Scalability

Scalability of different tools and analytic approaches remains a challenge where reviewers must process not just megabytes of data, but terabytes, petabytes, exabytes, and even zettabytes and beyond. Ultimately, we may have to learn to think differently when dealing with data at this order of magnitude.[70]

Aside from these analytical and technical issues, there are also issues arising from the nature and structure of the legal system.

(f) Courts and Judges that are Not Tech-Savvy

Many courts and judges just are not up to speed with technology. Even with respect to keyword searching techniques, a PwC report notes that "The technological black box at the heart of these superior solutions creates uncertainty. While courts, regulators and opposing parties generally understand the outputs from keyword searches-and if not, it can be explained, advanced techniques are less well understood. Those same stakeholders have less assurance and confidence about what they are getting from advanced search technology. Delivering reassurance is a challenge."[71] As one respondent in a PwC legal roundtable expressed it: "I suspect that we will get to the point where judges really understand the concept of search terms and how they work and then we'll say 'Forget search terms, they aren't working we want to use this new software . . . And the judge is going to say 'What! Enough already! You've just got us thinking about search and now you're going to

---

Post-Photographic Era (Cambridge, MA: MIT Press, 2001) (cited *ibid*). Carter's article traces the rules governing the admissibility of photographs into evidence in the courts of law in Canada, the United States and Britain and provides a useful overview of some of the issues associated with and approaches to admissibility of visual evidence that, by analogy, may prove applicable to visualizations resulting from the process of visual analysis.

69 Similar research challenges arise in the use of visualization in the analysis of geospatial data. See, for example, W. Cartwright et al, "Geospatial Information Visualization User Interface Issues" (2001) 28:1 *Cartography and Geographic Information Science* 45.

70 Perer, "Using rhythms", *supra* note 34.

71 Pwc, "E-Disclosure 2020", *supra* note 5 at p. 20.

throw this new thing on us.'"[72]

(g) Adversarial Nature of the Legal System

Litigation is adversarial by nature and e-discovery has become a battleground between opposing counsel that has driven up the costs of litigation and resulted in some spectacular fines.[73] In an effort to reverse this trend, The Sedona Conference has issued a Cooperation Proclamation aimed at facilitating cooperation in legal discovery, with over 100 judges at the time of writing this paper signing on to the document.[74] At present, however, still too many battles are fought in the e-discovery arena. Reliance on a technology such as VA that is seen as a black box because it is still new and not well-understood may exacerbate disputes between counsel over the relevance of documents, and lead to questions going to defensibility of approach. One way of overcoming this issue is to use VA initially to identify subsets of documents of interest (or non-interest, in the case of obviously non-relevant documents). Both parties could then agree on a list of appropriate keywords to be used in searching the dataset, either in an initial round of negotiations or after the owner of the data set has employed some measure of further sampling to produce exemplar documents.[75] In employing sophisticated, iterative search methods, the parties are certainly not limited in fashioning creative approaches aimed at reducing volume while at the same time focussing on what are likely to be the most material documents to a dispute. The fact remains, however, that where there is greater uncertainty, trust and collaboration may be more difficult to reliably establish.

VII. CONCLUSION

It is clear from the above discussion that VA is a promising technology but that it also has a long way to go. Moreover, it is unlikely to overcome the digital Tsunami entirely on its own. E. Efthiamidis argues that "effective information retrieval in today's complex litigation requires a variety of tools and approaches, including a combination of automated searches, sampling of large databases, and a team-based review of these results."[76] Beyond this, search and retrieval is unlikely to be effective where corporate archives are fragmented and in disarray. As recognized in the PwC report on e-discovery, "To successfully address e-[discovery] . . . [organizations] must ask themselves "What information do we have? Why do we have it? How long do we keep it? When do we destroy it? When needed, can we preserve, protect, access, search and produce it? And importantly, "What are the

---

72 *Ibid* at p. 20.

73 *Ibid* at p. 2. See also "Judge Brewster Benchslaps Qualcom Lawyers," *New York Times*, August 8, 2007, online: http://blogs.wsj.com/law/2007/08/08/judge-brewster-benchslaps-qualcomm-lawyers/.

74 "The Sedona Conference Cooperation Proclamation" (Sedona, AZ: Sedona Conference, 2008), online: <http://www.thesedonaconference.org/content/tsc_cooperation_proclamation/proclamation.pdf>.

75 Baron, "Nutshell", *supra* note 3.

76 Efthimiadis, "Legal Discovery", *supra* note 53 at p. 1.

consequences if we cannot?"[77] The bottom line is that, even with the most sophisticated technology, records management fundamentals are still essential.

77 PwC, "E-Disclosure 2020", *supra* note 5 at 7.