Catching the Banksters: The Use of Big Data Analytics in Billion Dollar Regulatory Investigations

Daniel Gozman
Henley Business School
daniel.gozman@gmail.com

Wendy Currie
Audencia Business School
w.currie@audencia.com

Jonathan Seddon
Audencia Business School
j.seddon@audencia.com

Abstract

Following the financial crisis, emboldened regulators have increased the magnitude of fines levied for financial malfeasance. The automation of the data discovery process underpins the rise in internal investigations, which financial organizations are obliged to conduct on the behest of regulators, keen to reduce information asymmetries and bolster transparency. Yet little research exists into the technologies which underpin post-crisis regulatory agendas. Our study focuses on big data technologies (eDiscovery tools) which facilitate investigations, where rare yet serious breaches have occurred. We focus on the micro/data level (volume, veracity, variety and velocity) to understand how these tools are influencing regulatory outcomes. The findings illustrate the need for financial organizations to adopt robust information governance policies to ease future investigatory efforts. We identify various practices which may help compliance managers better respond to regulatory investigations faster and more easily to ease the burden of post-crisis regulation.

1. Introduction

This study addresses how financial firms are facing burdensome demands to meet regulatory mandates using analytics (e-Discovery tools[1]) and how regulators are increasingly requiring organizations to conduct vast searches of their organizational data (structured and unstructured) to avoid sanctions or, instead, disclose levels of malpractice. The paper illustrates how the use of analytics is now part of a wider compliance regime in financial institutions where the risk of sanctions and reputational damage are ever present if malpractice is uncovered. Through our analysis, we provide guidance for practitioners seeking to navigate the complex world of regulatory compliance in the post-crisis world. The study also illustrates broader issues regarding the automation of professional services (paralegal work), in the era of machine learning and big data.

1.1 Problematizing big data in financial services

Despite the extensive use of mathematical models within capital markets which give an aura of impartiality and reliability, finance is not physics and to a large degree operates on trust and faith ultimately underpinned by transparency and the availability and accuracy of underlying data. The UK Regulator's Risk Outlook for 2014 [1], which outlined the major risks the industry was facing, highlights lack of transparency and asymmetric information as an ongoing risk: 'Information asymmetries – when one party in a transaction has more or better information than the other party – are common in most retail and wholesale financial markets' transactions. They potentially affect outcomes along the distribution chain, causing mis-selling and reduced trust and can affect market integrity if used to benefit the firm at the expense of one or more conflicted clients'. Thus, at a time where volumes of digital data are increasing exponentially, the use of technology and analytics to provide transparency into employees' conduct and culture is becoming increasingly pivotal and so deserves scrutiny and the attention of researchers.

Prior to the use of eDiscovery tools, transparency was facilitated by organizations in partnership with their legal teams, by reviewing and disclosing paper documents and print outs, of a relatively small number of electronic documents, to courts or regulators. As data intricacy increases, new challenges in meeting disclosure obligations emerge. Related eDiscovery

[1] Electronic discovery (also called e-discovery or eDiscovery) refers to any process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a regulatory, civil or criminal investigations.

URI: http://hdl.handle.net/10125/50459
ISBN: 978-0-9981331-1-9

projects present an increasing cost for regulated financial firms - not least as regulators' demands for firms to evidence compliance though data disclosure become more onerous and regular. Consequently, firms are being driven to revisit and improve data governance practices that underpin eDiscovery projects so that they can more easily and quickly respond to the demands of regulators. In summary, this study examines how global financial institutions are using big data compliance analytics[2] to support their governance operations and manage regulatory obligations. Thus, we are guided by the following high-level research question:

- How can big data tools intervene, when serious regulatory breaches occur, to automate the identification, collection, analysis and disclosure of structured and unstructured data?

2. Contextualizing regulatory risk

Following the 2008 financial crisis, operational failures and related malpractices have increased demands for more transparency and regulatory scrutiny of management practices [2]. Firms are now faced with a 'new normal' of higher operational costs, derived from the need to meet a 'tsunami' of new regulatory rules, with short deadlines for implementation, whilst being subject to heightened levels of supervision [3]. Figure 1 highlights the fines and penalties levied by the UK financial services regulator since the financial crisis. In Figure 1, the sharp increase in fines[3] in 2014 reflects large penalties levied against financial organizations for rigging the FX and LIBOR inter-banking benchmark rates, see Table 1. Often the levying of fines are precipitated by a regulatory investigations.

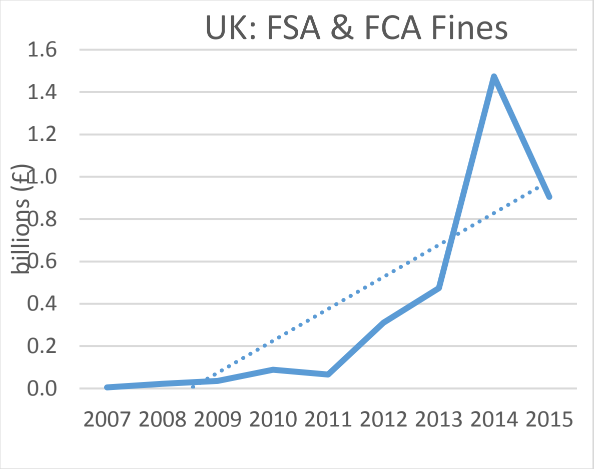

Figure 1: Post crisis financial penalties in the UK. Source: FSA and FCA[4]

	LIBOR Benchmark Rigging	**Foreign Exchange (FX) Benchmark Rigging**
Summary	In 2012, an investigation into the London Interbank Offered Rate, or Libor, which underpins over $300 trillion worth of loans worldwide, revealed collusion across multiple banks to manipulate interest rates for their own profit from 2003.	Similar to the LIBOR scandal in 2013 an investigation by UK, USA, and Swiss regulators, assisted by authorities in Hong Kong, revealed they were scrutinizing 15 banks for manipulating a benchmark for setting the price of major currencies from 2006. This market is the world's largest where turnover is over $5 trillion a day.
Penalties	Fines have been levied across multiple regulatory bodies in the UK, USA and the EU, currently more than $9 billion for rigging Libor. From 2015 investigations are continuing with other institutions expected to be implicated and related fines and civil lawsuits likely to ensue.	Multiple banks have paid a total of $5.6 billion. The FBI has described the scandal as involving criminality on a massive scale. Further regulatory investigations and law suits are expected as are criminal charges.

[2] Compliance analytics or just analytics hereafter refers to calculative functions for meeting regulatory obligations which utilise algorithms and draw upon data sets with volume, variety velocity and veracity. Visualization software (e.g. dashboards) may then be required to present the outputs in a way where it is easily understandable to humans.

[3] Fines have been levied across multiple regulatory bodies in the UK, USA and the EU, more than $9 billion for rigging Libor and $5.9 for FX. For example, The Financial Conduct Authority (FCA), the UK's sole financial services regulator, has imposed fines totalling £1,114,918,000 ($1.7 billion) on five banks for failing to control business practices in their G10 spot foreign exchange (FX) trading operations: Citibank N.A. £225,575,000 ($358 million), HSBC Bank Plc £216,363,000 ($343 million), JPMorgan Chase Bank N.A. £222,166,000 ($352 million), The Royal Bank of Scotland Plc £217,000,000 ($344 million) and UBS AG £233,814,000 ($371 million).

[4] The FSA operated between 2001-2013. After which, the FCA replaced it along with the PRA (Prudential Regulatory Authority). Total fines for 2013 include fines levied by both the FSA and FCA.

Table 1: Summary of LIBOR and FX rate rigging scandals

2.1 Regulatory investigations

Regulatory investigations may often incorporate 'dawn raids.' Such raids are defined as searches of individuals and businesses offices, often carried out in the early hours, by the FCA (UK financial services regulator) under warrant and in the presence of a police officer. The FCA undertakes these raids in order to prevent the removal of laptops, desktops, PDAs and mobile devices and the destruction of electronic documents and paper files. From 2012 to 2013, the number of dawn raids conducted by the FCA almost doubled from 11 to 20 raids.

Regulatory investigations may not always take the form of dawn raids. Regulators also have the power to require financial organizations to conduct internal investigations and report back. Where regulators suspect that misconduct may have occurred or want to clarify that it has not, the regulator may instruct financial organizations to conduct an investigation and submit relevant data and commentary to them in a prescribed format. An example is when the UK regulator wished to enlarge the scope of its investigations into rate rigging (see Table 1), and so instructed more financial organizations to conduct investigations into employee misconduct. Where such malpractice is thought to be widespread, the regulator may require firms to prove they have not been involved through the disclosure of unstructured data such as including emails or chat room data. Such investigations may be costly as the regulator may come back to the firm and ask them to widen the scope by including more individuals, more data types or lengthen the time periods reviewed. Often the timeframes for reporting back are tight. In such cases, financial organizations often look to their general council who, in turn, may look to external legal firms and eDiscovery consultancies for additional resource and expertise.

In the wake of the financial crisis, the UK regulator faced strong critiques for adopting a light touch principles based regulation of financial organizations [4]. Consequently, the regulator introduced more severe practices. However, intensified monitoring and sanctioning of financial organizations has not been without controversy. The dismissal of the head of the FCA (Martin Wheatley) by Britain's Chancellor of the Enqueuer in 2016 was interpreted by many as a reaction to criticism levied by banks insurers who complained that the regulator had adopted a "guilty until proven innocent" attitude to regulation. With Wheatly famously being quoted as saying he would, "Shoot first and ask questions later". While others suggested that the regulator, under Wheatly, foremost interest was in healing its reputation and so was 'obsessed' with media management [5].

2.2 High impact low frequency breaches

The Basel Committee on Banking Supervision [6] defines Operational Risk as, 'the risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events.' While a related category of risk, termed 'Compliance Risk', addresses, 'the risk of legal or regulatory sanctions, material financial loss, or loss to reputation a bank may suffer as a result of its failure to comply with laws, regulations, and rules.' Often, firms organise their compliance function within their operational risk function as there is a close relationship between the two. A third relevant risk category is termed 'Regulatory Risk', which refers to the risk that a change in regulatory rules and laws may impact a business [7]. These definitions provide us with a useful point of departure from which to consider the use of big data technologies for managing compliance and investigating breaches.

In a paper for the International Monetary Fund [8] Jobst suggests, 'the typical loss profile from operational risk contains occasional extreme losses among frequent events of low loss severity. Hence, banks categorize operational risk losses into expected losses (EL), which are absorbed by net profit and unexpected losses (UL), which are covered by risk reserves through core capital and/or hedging.' The LIBOR and FX rate rigging scandals and rogue trader malpractice are examples of rare operational risk events leading to considerable fines and reputational damage [5]. We build on Jobst's representation of operational risk in order to frame our study, see Figure 2. This study addresses low probability breaches which occur much more rarely and are often distinguished by huge fines and substantial changes and refinements to regulatory frameworks.

3. Related literature

Previous research has focused on the strategic implications of big data, but not much research has considered how these technologies are implicated in regulatory investigations which may yield fines amounting to billions of dollars. Economists have studied the LIBOR and FX rigging scandals in relation to operational risk [5], arbitrage, market- making and the transfer of financial risk [4], the origins of the scandals [9] and the ethical implications for managing the risk culture of financial organizations [10, 11].

However, there remains an absence of work which addresses the tools big implicated in investigating low frequency yet high impact regulatory breaches which may result in controversial billion dollar fines for banking institutions.

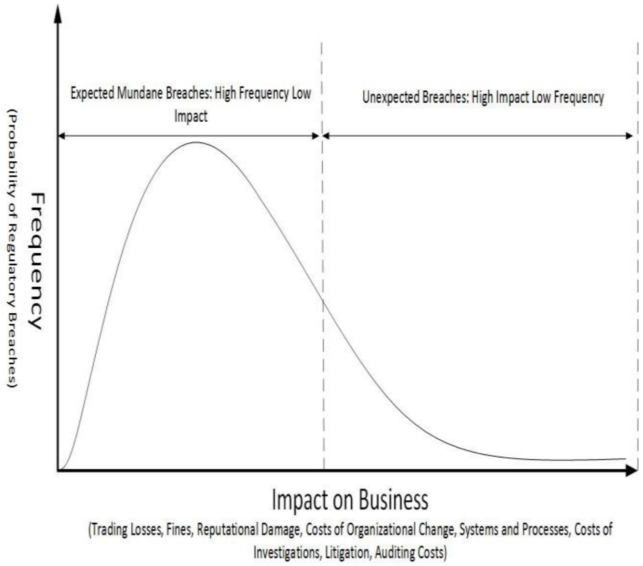

Figure 2: Frequency and impact of regulatory breaches

While the practice of managing large data has been a perennial topic for information systems for decades [9], few studies are situated within financial services which link important topics of regulation, compliance, technology and the professional practices of individuals, such as lawyers, compliance managers, fund managers and traders. Prior work on managing technology in financial services has widely addressed data and information issues around trading [10, 11] and more recently, on analytics and inter-organizational standards in the mortgage industry [12]. The move from manual based to electronic trading following the 'Big Bang' in 1986 has generated interesting studies about the use of technology and data in organizational change [13]. A study on regulation and IT following the financial crisis observed the scope of the credit crisis and resultant great recession (marked by the collapse of Lehman Bros and actions required to save Northern Rock) extended well beyond the corporate failures of the dot.com era [3]. However, there are relatively few studies from the information systems' community that focus on the wider policy issues relating to financial regulation, technology and data.

3.1 Theorizing big data (4Vs)

More and more specialist tools, such as eDiscovery tools, are being utilized to traverse large volumes of structured and unstructured data held within organizations but across borders to help evaluate compliance breaches and assist with litigation. Business analysts suggest, '*Big data has been a reality for eDiscovery for longer than it has in most other application areas. The volume of information collected in response to legal and regulatory challenges has grown from thousands, to hundreds of thousands, to millions of documents over the last few years.'* [14]. As volumes of data have increased, correspondingly academic and practitioner interest in big data has grown. One definition states, 'big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time' [15]. A common theorization of big data, the four Vs, has focused on what differentiates big data from common analytics. The 4Vs framework provides underpinning concepts which differentiate big data and facilitate related analysis. The **volume** of data sets, the speed of data creation and availability (**velocity**), the **variety** of data types (e.g. social media, emails, videos, GPS signals) and the trustworthiness, integrity and accuracy of the data (**veracity**) collectively define this phenomenon [16]. Furthermore, machine learning technologies review and learn from data sets (with 4V properties) to make predictions and recognize patterns that can allow firms to better identify misconduct and risks.

4. Methodology and research context

To fulfil our research goal, we selected an eDiscovery and data forensics consultancy based in London (UK) and serving a variety of financial organizations worldwide. The study used semi-structured interviewing techniques with 33 interviews conducted in total. Senior business managers, lawyers, data forensic experts, project managers, compliance officers and eDiscovery consultants were interviewed across financial organizations, law firms and the eDiscovery consultancy. Our inductive (theory-building) approach allowed us to build our analysis initially from a series of 5 pilot interviews to validate and develop the research instrument, with informants from the consultancy. From the outset of this study it was important to develop a working definition of the concept of 'big data' relevant to the financial industry and the technology under investigation. The results of these interviews with business and IT managers showed that big data was characterized in three ways. First, informants discussed big data in terms of increasing volumes where lawyers, compliance managers, fund managers and traders now work with granular data (reported on an item-by-item basis). Second, the velocity of data has grown where data is frequently updated and analyzed. Third, the variability of data has increased where data can be structured or unstructured (i.e. text, video).

To control the scope of our study, our interview schedule situated 'big data' around how consultancy was changing products/services and client requirements for conducting regulatory investigations. Our aim is to impose discipline on our research design by carrying out open-end interviews on a more narrow range of areas and topics in an attempt to avoid some of the methodological pitfalls facing qualitative researchers. A common problem is that qualitative interviews generate numerous amounts of data which is 'messy' and difficult to organize [17]. The result is often an over-scoping of the study, where the phenomenon (in this case, big data) becomes lost in translation as the situations and contexts to which informants refer are not well defined.

Data analysis was conducted through long established interpretive techniques for analyzing data through the recursive identification of patterns, first through categorization and then abstraction [18]. During the process of data analysis, primary and secondary data were closely reviewed to determine points of importance and interest [17]. Common themes were identified and categories assigned for each case independently. Thus, long interviews were simplified through the adoption of simple categories. The analysis adopted a two cycle approach to coding. The first cycle adopted a 'Descriptive Coding' approach for summarizing segments of data. This method is appropriate for inductive studies utilizing semi-structured protocols [19]. This approach requires the application of a content phrase to a segment of data representing a topic of inquiry, and so related to the risks and challenges being faced, for example 'Regulatory Investigations', 'Unstructured Data' or 'Changes in Data Volume.' The second cycle adopted a 'Pattern Coding' approach to identify major themes by searching for causes and explanations from the data. Such an approach builds on the first cycle of analysis and are, 'explanatory or inferential codes, that identify an emergent theme, configuration or explanation. They pull together a lot of material into more meaningful and parsimonious units of analysis. This analysis was guided by existing theorizations of big data. Examples include 'Volume', 'Veracity' and 'Digitization'. Scope, depth and consistency were achieved by discussing key concepts, constructs and terminology with each of the informants and triangulating the findings across primary and secondary data sources [18]. Secondary data included white papers, press releases and speeches, regulatory mandates, marketing materials and commentary from legal and accounting firms. For example, interviewee references to particular areas of regulation were triangulated with the original regulations and industry commentary to ensure key points were fully understood and consistent across sources.

4.1. Case: eDiscovery consultancy

Our case study focuses on a full service eDiscovery firm. Millnet is one of the UK's largest legal data services and document solutions providers, with clients in over 60 countries. The firm was incorporated in 1996 and has evolved from providing traditional legal print services to providing electronic document consulting, processing and review. Millnet's clients' include Legal 500 firms and FTSE 100 companies. Millnet is not a software vendor (it works with a number of vendors) but instead utilizes best in class eDiscovery software to provide consultancy, infrastructure and expertise. The firm supports the investigation and review of structured and unstructured electronic data[5] held within financial organizations, which may relate to serious internal investigations, litigation or regulatory breaches. Millnet recently moved premises and invested £1M in a new facility. This investment allowed them to double their square footage to facilitate growth in personnel, allowed for the integration of purpose built forensic and server rooms, and upgrades to their data network security and biometric entry control systems for quarantined areas.

Within the UK and the USA, the legal profession has been transformed through a combination of technological advancement and related alterations in the legislative landscape. In 2006, the USA's Federal Rules of Civil Procedure (FRCP) and in 2013, the Jackson Reforms were brought into effect in the UK. Both sets of legislation address how technology may be used to support civil cases. A crucial development is that electronically stored information (ESI) has been accepted as being of equal evidentiary weight and value as conventional paper documents. Deloitte [20] suggests that, 'It is often the case that an entire business dispute, regulatory investigation, or

[5] Including, emails, voice recordings, video streaming, chat rooms, spreadsheets and text based documents.

multimillion pound litigation may hinge on identifying when a single piece of data was communicated, generated, altered or deleted, by and to whom and under what circumstances.' Table 2 highlights some examples of how these tools have been used in regulatory investigations and the value they create. The Electronic Discovery Model (EDRM), Figure 3, represents a conceptual presentation of the eDiscovery process. The model should not be interpreted as a literal, linear or waterfall model. Systems and firms may facilitate discrete elements or the whole model, particularly as software vendors begin to consolidate functionality across the EDRM. The process depicted should be viewed as iterative. The same activities may be repeated many times to create an increasingly accurate set of results. It may also be necessary to cycle through earlier steps to define the approach being adopted as investigators obtain a better understanding of the data or the context regarding the investigation

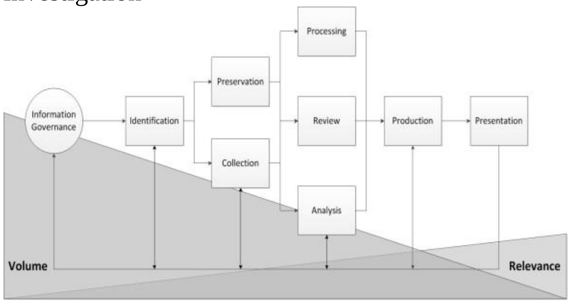

Figure 3. Electronic Discovery Reference Model v3.0 Source: edrm.net.

More recently, eDiscovery vendors have sought to incorporate more automation in order to assist with the increasing data complexities. Where key word searches are unable to deal with the variety and volumes of data being considered, predictive coding is increasingly used when there is a need to investigate large volumes of varied structured and unstructured data in a cost effective manner.

Predictive coding involves using sophisticated machine learning algorithms to determine the relevance of documents based on feedback from a human. Instead of junior staff reviewing large volumes of data, the senior partners will review and code a 'seed' set of documents. As this process continues, the system learns more about the coding approach and begins to predict the reviewers' coding. At the point where the reviewers and systems coding are sufficiently similar, the system is deemed to have learned enough to make confident predictions regarding the remaining documents.

	Potential contentious matter	Financial Services Market Abuse Investigation
Investigation	Email accounts belonging to 23 potentially relevant people in three jurisdictions covering an 18 month period This returned over 3.6 million documents. By locating the documents relied upon by the senior individual and documents highlighted by witnesses and applying the "email threading" functionality, we quickly identified 1,198 highly relevant documents. A number of complicated, targeted, custodian and key word searches (in English, German and French), refined by specific deduplication searches to overcome the challenge of email address fields not always being identical when processed, reduced the dataset to just over 7,400 documents which required human review.	Original collection of 20,000 documents (T1) – review completed. Second collection of 300,000 documents (T2) with less than 4 weeks to review. Using the relevant documents from T1 + 'good example/key documents' and subject matter experts to train Relativity Assisted Review on what makes a document relevant vs not. Relativity Assisted Review provides a complete audit trail on every decision the computer makes based on what is deemed to be a seed document as reviewed by the subject matter expert.
Value created	It cost £145,000 to review this dataset. Over 350 man days saved - approximately £263,000 (60%) cost savings compared to a traditional keyword driven document review.	85,000 documents were reviewed at first pass and 9,000 reviewed at the second pass. Total cost of review to production exercise was £175,000 compared with traditional document review at a document by document level would have cost £465k and taken 4662 review hours to complete first pass review alone.

Table 2: Summary of LIBOR and FX Rate Rigging Scandals. Source: Simmons and Simmons

5. Findings: Managing the 4Vs

The veracity, variety, velocity, and volume of the data integral to regulatory investigations pose specific challenges. As Millet's website states, 'Banking matters tend to involve vast amounts of information and can often include unusual file types such as Bloomberg messaging and audio files.' A key challenge for those conducting regulatory investigations is reviewing a vast ‘universe’ of structured and unstructured data and then narrowing down the amount of files which are actually passed on to be reviewed by expensive legally trained individuals, whose time should be maximized.

A Partner at law firm Simmons and Simmons and client of Millnet commented on how the **volume** and **velocity** of files has grown in recent years at a rapid pace along with the technology employed, *‘About six, seven years ago now, electronic data was becoming more of a challenge previously when it was all hard copy lots of paper files came in and we had to deal with it manually. We could just print the emails. Over the last two, three years the volumes of data have gone through the roof. You're no longer dealing with data sets that tend to bulk out to about 20 to 30,000, you're talking about millions. So from a lawyer's perspective, they are going from, ‘I got a box of files or maybe on a bad day I got ten boxes of files, to, I've suddenly now got a warehouse full. A conceptual warehouse full and you're obviously not going to print them all out. So big data for us, or what counts as big, is things in the millions. And actually to be honest, things in the hundreds of thousands, anything where you're not going to be able to have a whole bunch of humans looking at it. The last two years have seen developments in the infrastructure, both the software and the hardware that enable people to do a lot more a lot quicker. We’re talking days for* hundreds *of gigabytes, days rather than weeks.’*

The need for eDiscovery systems to deal with a **variety** of file types has also become increasingly important. The need to investigate chat data has become common in regulatory investigations, particularly those involving multiple organizations. Yet, several of the study’s participants highlighted chat data as providing particular challenges. A Millnet eDiscovery consultant commented, ‘*We are seeing more of chat room data because people are not just using emails, they’re using chats, they’re using their internal chat programmes, and they’re using the Reuters and Bloomberg chatrooms. Chat data are big, long streams of text, maybe 800 pages. It comes out in long transcript and is not pretty on the eye and is not easy to review. More often than not it’s got hundreds of hits and somebody just has to sit there and go through it. Also, you see a lot of noise so everybody coming in and out you see everyone's email, everyone's company disclaimers, and you've got to wade through all of this and within that there may be something dreadful going on. But how, as a human being, you're going to find it? The other challenge for chat rooms, it's the phraseology people use. So it's not text searchable easily because people don't say, ‘I'm going to go and manipulate x.’* Our participants highlighted how technologies allow for the reduction of ‘noise’ essential to allowing human reviews. An experiment conducted by a UK law firm using two individuals to review the same set of five documents revealed that chat data could be reviewed 40% quicker using an eDiscovery platform which removed the ‘noise’.

In addition to unstructured data, structured data (data held within relational databases) also presents challenges. Financial organizations often have large numbers of bespoke, vendor and legacy systems containing vast amounts of structured data. Examples include customer relationship management tools, accounting tools and trading and risk platforms. Data schemas inherent in such systems allow the data held to be accessed quickly and easily to facilitate business as usual processes. The foundation of eDiscovery tools is the ability to turn unstructured data into structured data. That is, to identify, analyze, search and present vast quantities of unstructured data. In order to do so the system creates a database of structured data populated by unstructured data. Thus, eDiscovery tools ensure that the data held within the database is searchable and can be presented in a format which is easy for humans to understand. Consequently, it may be assumed that taking data which is already structured and importing it into an eDiscovery tool might be easier. An eDiscovery Project Manager commented, *‘Structured data is a strange one because it feels like it should be the Holy Grail. All of eDiscovery is about taking unstructured data and turning it into structured data, that’s what the damn process is all about. And the data is already structured, it should be easy. You should be able to run your queries and find all your relevant events or client log activities or whatever it is. And my experience is that you almost never can.’*

There are several reasons why analyzing structured data present additional challenges. Often, the information systems implemented by financial organizations contain structured data not designed for eDiscovery purposes but are instead designed for people to conduct their day-to-day work, for example, systems which maintain customer data. This often creates **veracity** problems when conducting eDiscovery searches, where the data schema of the database is not designed to facilitate related queries and may return inaccurate data. Another reason cited was that it is often not easy to mine the data from the system. Software vendors may not include functionality to allow the extraction of the data as it is not usually necessary and the inclusion of such functionality may provide opportunities for data theft. These challenges are eased where organizations use well known systems from vendors such as Microsoft or Oracle. However, further challenges occur, where the eDiscovery team may not have access to the vendors' license and their data schema or related design documentation, or where the system in question was bespoke, and the design is not obvious or is a legacy system no longer supported by the vendor. An eDiscovery Consultant commented, '*So extraction doesn't exist to a huge degree, which is really bizarre and it means that, on the occasions that we do end up doing structured data in a huge way, it ends up being treated much more like forensics because you are having to piece together a system, quite often from its back end without its interface, which you normally don't have a license for, or perhaps an installer for, or just perhaps an environment in which you can install them. So you're picking to bits a database which, it's much, much worse than unstructured data because the unstructured data is basically a load of formats that we deal with every day. Yeah, the data schemas are difficult to recreate. But decoding these structures, if it's noisy or not obvious how to recreate something that's useful can be difficult.'*

A common challenge across both structured and unstructured data types includes the need to understand what constitutes duplication and so to remove irrelevances in the data to further ensure the accuracy (**veracity)** of the data. For example, email trails are often duplicated where individuals forward or reply to existing email trails. Duplication complexity is increased where emails are held in different formats across numerous devices, including the exchange server folder, local inboxes on desktops and laptops and mails stored on mobile devices. Furthermore, while each email may look similar to a human, each mail's meta-data relating to author, recipient, date and time will also differ. An eDiscovery consultant provided an example of the problems meta-data can cause, *'I can give you a real world example, which is if you created some documents in 2012 and today you copy and paste them onto a USB stick, actually what you'll do in doing that is you will reset the creation date of the copy documents to today's date. Now you'll get some people that will do a collection where they say, right, we want all documents, I don't know, related to mis-selling between 2009 and 2011. If the IT department has gone at some stage and copied the documents from one system onto another they have basically reset the creation dates, so there'll be great chunks of documents there that actually aren't within the search'* Other complexities occur in defining and applying keyword searches which run the risk of being, 'both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English'. Simple keyword searches when used in isolation may only reveal 20% of relevant evidence in a large, complex dataset, such as an email collection. Instead, search terms should be thoroughly tested for efficacy and accuracy, part of which would include sampling to ensure that categories are neither over nor under inclusive and that there exists an iterative feedback loop to ensure that terms are refined appropriately.

6. Discussion

Our study shows how the complexity and heterogeneity of underlying data and related analytics provides a further layer of technical complexity to banking matters and so adds further opacity to understanding controls, behaviors and misdeeds. For example, one must understand the nature of eDiscovery search capabilities and related data issues to run effective searches. Predictive coding affords the automation of operational practices for discovery and so shapes this process iteratively as the system initially learns from human input and eventually takes over (velocity). Data accuracy (veracity) may also act to unduly influence outcomes. This underscores the need to study big data analytics at the level of micro practice and from the bottom up.

As the use of big data analytics within financial firms becomes further embedded and institutionalized, the ability of firms to facilitate analytics and reduce related costs and overheads through information governance will become increasingly important. Yet, our study shows that proactively structuring and managing data is of a low priority for many managers as the volume and variety of regulatory rules increases along with related costs and overheads. A further contribution is made in reviewing the complexities of dealing with different data types and how paper documents may still present challenges to those conducting regulatory investigations. Many discussants of big data overlook the fact that large volumes of important documents (e.g. financial records, health records) are often still held in paper form and that transferring these to searchable electronic documents may not be as straight forward as assumed.

6.1 Managerial and Policy Implications

As Constantiou and Kallinikos [21] succinctly note, *'it makes a great deal of difference whether data is gathered through a carefully laid out cognitive (semantic) architecture or, by contrast, is captured and stored without such a plan and on the assumption* *that it may be variously used a posteriori.'* The purpose of eDiscovery tools is to manage heterogeneous data created in haphazard fashion and to apply and impose a clear structure upon it so that it can be searched and analyzed. Where new data types, such as chat room data, become relevant to regulatory investigations such systems must be flexible enough to incorporate such variety. An important function of such systems is to create structured data out of unstructured data. eDiscovery systems classify and assemble data which has been generated as part of everyday working practices and communications and stored at the point of creation with little view as to how such data may be structured to support future regulatory investigations and litigation. Building on this perspective **we suggest that organizations may seek to apply order across haphazard data and thereby reduce related complexities by implementing proactive data and information governance practices**. Respondents felt that future compliance pressures and risks could be significantly mitigated through proactive categorization and management of data by financial organizations, yet often information and data governance policies within financial organizations was felt to be not well implemented and not a current priority. This is perhaps unsurprising in the post financial crisis environment where operations' budgets are often consumed with meeting new compliance practices and where there exists little residual appetite or resource for implementing proactive measures aimed at improving or gold platting existing compliance measures. However, **we suggest that firms which proactively organize and manage their data will find the pain of compliance and managing breaches easier in the years to come.** As regulatory investigations and related litigations becomes increasingly common, financial organizations which are likely to have to undertake future eDiscovery projects may use information governance techniques to reduce the need to rely on costly eternal resources. Where information can be found quickly and easily, organizations can react more quickly. Our respondents suggest that one of the key challenges in responding to regulatory investigations was the tight timeframes set by the regulatory bodies. Tight deadlines for responses may create further challenges where financial organizations see eDiscovery searches as simplistic and so do not appreciate the intricacies involved at the micro/data level, including reducing 'noise', accessing and managing structured data, preserving metadata and approaches for scanning, analyzing and indexing paper documents. Consequently, they may leave interacting with eDiscovery experts too close to the deadline. The eDiscovery consultants interviewed felt that was often because, initially, the scope and complexity of the investigation was misunderstood or that the ability of technology to automate work and reveal in the early stages the impact of the investigation was underestimated. Consequently, **we would advise financial services practitioners conducting eDiscovery projects to engage with technical experts early on who understand the issues at the micro/data level**. Firms which understand the impact of regulatory investigations may formulate appropriate strategies. In regulatory investigations **early determination of whether the firm is likely to be subject to fines and further litigation allows organizations to segregate funds appropriately and put strategies in place to mitigate reputational damage**. Furthermore, regulators have previously reduced fines for organizations which have been the first to come forward and highlight a problem. Harnessing the power of analytics to better understand organizational operations may have many additional benefits beyond compliance. Through better understanding and control of the data their organization holds, firms will be much better placed to reap the benefits of big data analytics. For example, analytics may help firms identify areas where duplication of effort and systems are occurring and so improve processes. Improved understanding of operational risks may also allow firms to reduce their requirements to hold higher levels of regulatory capital. Furthermore, analytics may help organizations better understand how individuals in the firm interact with one another and thereby act to improve lines of communication. Analytics may also assist organizations in vital strategic decision making and related efforts to recruit and retain necessary staff. As a consequence, **firms which embrace information governance techniques are better placed to exploit big data analytics and related future innovations**. To conclude, firms which are able to become masters of their own data and conquer challenges related to volume, velocity, veracity and variety will be able to draw a competitive advantage through enhanced strategic decision making and increased operational efficiency.

7. Concluding comments

Symbolized by the four V's (volume, velocity, variety and veracity) there is no 'one-size-fits-all' template for all organizations and institutions in managing regulators' demands for disclosure. A common challenge for global firms, is the need for each company to keep pace with the ongoing legal and regulatory landscape, where new directives, laws and rules are coercively applied often by regulatory bodies based in different countries.

By providing empirical examples of how companies operate within their own big data landscape, it is apparent that many of the examples we discuss range from the highly strategic, where each firm has to interpret, develop and implement a data governance strategy, to the very mundane, by considering how each rule or guideline applies to their own operations. While much of the current academic literature looks at the strategic impact of big data, we caution that in many regards, the 'devil is in the detail.' Many of the thorny issues surrounding big data are at the micro-practice level which is less often researched than macro-levels (industry-wide) or meso-levels (across and within companies). We believe that future research which considers big data in the context of financial services and other areas, such as healthcare, may consider multi-level studies which link policy and strategic issues with more granular practices.

The proliferation and reach of big data means that even looking at a single case study, such as a site within a company, poses significant research challenges. This is because the global reach of data now extends well beyond a single site and involves the interventions, decisions, and applications of multiple participants, including regulators, industry professionals, vendor partners, and customers.

In conclusion, the philosophy of reacting to organizational and regulatory failures by increasing the scope and scale of investigations means that regulated activities will become increasingly reliant on analytics. Yet such automation comes at a price by limiting the scope of regulatory structures and analytical processes and does not address deep rooted unethical behavioral practices beyond providing accountability and surveillance after the fact.

Acknowledgments: We are grateful to the SWIFT Institute for funding this research and in particular we would like to thank Peter Ware and Nancy Murphy. We would also like to acknowledge those who participated in our research, as interviewees, including the eDiscovery teams at Millnet and Simmons and Simmons. Finally, we would like to acknowledge the editors and anonymous reviewers for helping us to develop this work.

References

[1] FCA. FCA Risk Outlook. 2014. Retrieved 4th May, 2014 from http://www.fca.org.uk/static/documents/corporate/risk-outlook-2014.pdf

[2] Williams, J.W. Regulatory technologies, risky subjects, and financial boundaries: Governing 'fraud' in the financial markets. *Accounting, Organizations and Society*, 38, 6 (2013), 544-558.

[3] Gozman, D., and Currie, W. The role of Investment Management Systems in regulatory compliance: a Post-Financial Crisis study of displacement mechanisms. *Journal of Information Technology*, 29, 1 (2014), 44-58.

[4] Ashton, P., and Christophers, B. On arbitration, arbitrage and arbitrariness in financial markets and their governance: Unpacking LIBOR and the LIBOR scandal. *Economy and Society*, 44, 2 (2015), 188-217.

[5] McConnell, P.J. Systemic operational risk: the LIBOR manipulation scandal. *Journal of Operational Risk*, 8, 3 (2013), 59-99.

[6] Bank of International Settlements. Basel Committee on Banking Supervision: Operational Risk. 2001. Retrieved 6th June, 2015, from https://www.bis.org/publ/bcbsca07.pdf

[7] Securities Institute. *Operational Risk Official 6th Edition IAQ Workbook* London: Centurion House, 2004.

[8] Jobst, A.A. Operational Risk—The Sting is Still in the Tail but the Poison Depends on the Dose. *IMF Working Paper*, 07, 239 (2007).

[9] Markus, M.L. New games, new rules, new scoreboards: the potential consequences of big data. *Journal of Information Technology*, 30, 1 (2015), 58-59.

[10] Weber, B.W. Next-generation trading in futures markets: A comparison of open outcry and order matching systems. *Journal of Management Information Systems* (1999), 29-45.

[11] Weber, B.W. Adoption of electronic trading at the International Securities Exchange. *Decision Support Systems*, 41, 4 (2006), 728-746.

[12] Markus, M.L., Steinfield, C.W., and Wigand, R.T. Industry-wide information systems standardization as collective action: the case of the US residential mortgage industry. *MIS quarterly* (2006), 439-465.

[13] Clemons, E.K., and Weber, B.W. London's big bang: a case study of information technology, competitive impact, and organizational change. *Journal of Management Information Systems* (1990), 41-60.

[14] Zhang, J., and Landers, G. The State of E-Discovery in 2015 and Beyond. 2015.

[15] Snijders, C., Matzat, U., and Reips, U.-D. Big data: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, 7, 1 (2012), 1-5.

[16] McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., and Barton, D. Big data. *The management revolution. Harvard Bus Rev*, 90, 10 (2012), 61-67.

[17] Silverman, D. *Interpreting Qualitative Data: Methods for Analyzing Talk, Text and Interaction*. London: Sage Publications, 2001.

[18] Miles, M.B., and Huberman, A.M. *Qualitative data analysis: A sourcebook of new methods*. Sage publications, 1984.

[19] Saldana, J. *The Coding Manual for Qualitative Researchers*. Thousand Oaks: Sage, 2009.

[20] Deloitte. Analytic and Forensic Technology. 2015. Retrieved 18th June, 2015, from http://www2.deloitte.com/jp/en/pages/risk/solutions/frs/analytic-and-forensic-technology.html

[21] Constantiou, I.D., and Kallinikos, J. New games, new rules: big data and the changing context of strategy. *Journal of Information Technology*, 30, 1 (2015), 44-57.