Digital Investigation 11 (2014) 273–294

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Impacts of increasing volume of digital forensic data: A survey and future research challenges

Darren Quick*, Kim-Kwang Raymond Choo

Information Assurance Research Group, Advanced Computing Research Centre, University of South Australia, Mawson Lakes Campus, Mawson Lakes Boulevard, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 1 September 2014
Accepted 4 September 2014
Available online 1 October 2014

Keywords:
Data mining
Data volume
Digital forensics
Evidence discovery
Forensic computer analysis
Intelligence analysis
Knowledge management

ABSTRACT

A major challenge to digital forensic analysis is the ongoing growth in the volume of data seized and presented for analysis. This is a result of the continuing development of storage technology, including increased storage capacity in consumer devices and cloud storage services, and an increase in the number of devices seized per case. Consequently, this has led to increasing backlogs of evidence awaiting analysis, often many months to years, affecting even the largest digital forensic laboratories. Over the preceding years, there has been a variety of research undertaken in relation to the volume challenge. Solutions posed range from data mining, data reduction, increased processing power, distributed processing, artificial intelligence, and other innovative methods. This paper surveys the published research and the proposed solutions. It is concluded that there remains a need for further research with a focus on real world applicability of a method or methods to address the digital forensic data volume challenge.

Introduction

The increase in the number and volume of digital devices seized and lodged with digital forensic laboratories for analysis has been an issue raised over many years. This growth has contributed to lengthy backlogs of work (Gogolin, 2010; Parsonage, 2009). A significant growth in the size of storage media combined with the popularity of digital devices and the decrease in the price of these devices and storage media has led to a major issue affecting the timely process of justice. There is a growing volume of data seized and presented for analysis, often now consisting of many terabytes of data for individual investigations. This has resulted from;

(a) An increase in the number of devices seized per case.

(b) The number of cases with digital evidence is increasing (anecdotal information indicates the last case observed without digital evidence was at least 3 years old).

(c) The size of data on each individual item is increasing.

The increasing number of cases and devices seized is further compounded with the growing size of storage devices (Garfinkel, 2010). Existing forensic software solutions have evolved from the first generation of tools and are now beginning to address scalability issues. However, a gap remains in relation to analysis of large and disparate datasets. Every year the volume of data is increasing faster than the capability of processors and forensic tools can manage (Roussev et al., 2013).

Processing times are increasing with the increase in the amount of data required to be analysed. In the last decade, there have been many calls for research to focus on the timely analysis of large datasets (Garfinkel, 2010; Richard

* Corresponding author. Tel.: +61 8 8172 5074.
E-mail addresses: Darren.Quick@gmail.com, darren.quick@mymail.unisa.edu.au, darren_q@hotmail.com (D. Quick), raymond.choo@unisa.edu.au (K.-K.R. Choo).

http://dx.doi.org/10.1016/j.diin.2014.09.002
1742-2876/© 2014 Elsevier Ltd. All rights reserved.

and Roussev, 2006a; Wiles et al., 2007) including the application of data mining techniques to digital forensic data in an endeavour to address the issue of the growing volume of information (Beebe and Clark, 2005; Palmer, 2001).

Serious implications relating to increasing backlogs include; reduced sentences for convicted defendants due to the length of time waiting for results of digital forensic analysis, suspects committing suicide whilst waiting for analysis, and suspects denied access to family and children whilst waiting for analysis (Shaw and Browne, 2013). In addition, employment can be affected for suspects under investigation for lengthy periods of time, and ongoing difficulties can be experienced by suspects and innocent persons when computers and other devices are seized, for example; the child of a suspect may have school assignments saved on a seized computer, or the partner of a suspect may have all their taxation or business information saved on a laptop.

In this paper we study literature examining the digital forensic data volume issue, including the volume of data, the growth of media, and research challenges. We review publications focussing on data mining, data reduction, triage, intelligence analysis, and other proposed methodologies. We then summarise the findings, and future directions for research are outlined in the conclusion.

We located material published in the last 15 years (i.e. 1/1/1999–14/6/2014) by searching various academic databases, including IEEE Xplore, ACM Digital Library, Google Scholar, and ScienceDirect using keywords such as; “Digital Forensic Data Volume”, “Computer Forensic Volume Problem”, “Forensic Data Mining”, “Digital Forensic Triage”, “Forensic Data Reduction”, “Digital Intelligence”, “Digital Forensic Growth”, and “Digital Forensic Challenges”. In addition, we browsed all papers published in Digital Investigation: The International Journal of Digital Forensics & Incident Response, and The Journal of Digital Forensics, Security and Law. A summary table of key papers and topics is listed in Table 4 (see Discussion section).

Survey

Volume of data (1999–2009)

Digital forensics plays a crucial role in society across justice, security and privacy (Casey, 2014). Concerns regarding the increasing volume of data to be analysed in a digital forensic examinations have been raised for many years. McKemmish (1999) stated that the rapid increase in the size of storage media is probably the single greatest challenge to digital forensic analysis. In 2001, Palmer published the results of the first Digital Forensic Research Workshop (DFRWS), which included a section from Dr Eugene Spafford discussing various challenges posed to computer forensics and stated, ‘Digital technology continues to change rapidly. Terabyte disks and decreasing time to market are but two symptoms that cause investigators difficulty in applying currently available analytical tools’ (Palmer, 2001).

Sommer (2004) outlined the issues with the increasing data size and number of devices in a legal environment, which is slow to understand the resources and procedures involved, is resulting in methods which do not scale to cope with the increases. Roussev and Richard (2004) stated that the vast amounts of disk storage in use by ordinary computer users would soon overwhelm digital forensic investigators. Ferraro and Russell (2004) discuss the increase ubiquitousness of computers, coupled with a notion of a forensic scientist conducting examinations in every computer related crime, leading to demand for forensic science services which outstrips the resources available, and that alternative methods will be required. Ferraro and Russell (2004) also outline the average time digital evidence is retained, stated to be between three and five years or more, and that orders from courts which can mandate impossible or time consuming procedures in evidence handling, can impede timely processing of evidence. Rogers (2004) reported on a study relating to the needs of digital forensic practitioners, and listed the top issues from a survey conducted of 60 respondents indicating that education, training and certification was the most reported issue, and a lack of funding was the least reported issue, with ‘technology’ and ‘data acquisition’ in the top four concerns raised by the respondents.

Brown et al. (2005) stated the challenge in digital forensics is locating relevant information in large datasets, analogous to finding ‘needles in haystacks’, or in some instances ‘bits of needles in bits of haystacks’. Beebe and Clark (2005) state that ‘the sheer volume and “noisiness” of … data is absolutely overwhelming and incompatible with manual data analysis techniques.’ The unique requirements that make the field of forensic analysis different from traditional pattern analysis include; data that is both related and unrelated, ‘interesting’ data may be low frequency rather than repetitive, data sources are large and can include multiple sources, differing data types, and that the cost of missing relevant data is large (Brown et al., 2005). Sheldon (2005) stated that due to the increasing capacity of storage devices and the increase in time to undertake analysis, it is ‘not feasible to continue performing forensic analysis using the accepted approaches that we use today’.

Alink et al. (2006) state that the volume of data in typical investigations is huge, with modern systems containing hundreds of gigabytes, and large investigations often consist of multiple systems totalling terabytes of data, and in addition, the diversity of data can be overwhelming. Richard and Roussev (2006a) made the observation that most current digital forensic tools are unable to deal with the ever growing size of media, and that new analysis techniques will be required, such as automatic categorisation of pictures. Adelstein (2006) states that the nature of a digital forensic investigation has changed, and the larger disk sizes has resulted in an increase in the time required for collecting a full disk image and then conduct analysis. Furthermore, the nature of digital forensic investigations calls for ongoing technology developments to provide significantly better tools for practitioners (Richard and Roussev, 2006b). As an example, Alink et al. (2006) describe their prototype system, which can display timestamp information merged from different tools, highlighting that current tools, such as EnCase, display time ordered views of file-system metadata only.

Wiles et al. (2007) when discussing the challenge of exponential growth in data, specifically the volume and cost to analyse, alluded to forensic practitioners attempting to locate needles in haystacks which becoming larger and more compact. Khan et al. (2007) highlight that the increasingly larger data volumes with a varying level of diversity result in a need for additional resources and greater cost, and that much of the data consists of text, hence text mining methodologies have been proposed to look for patterns.

Case et al. (2008) stated that the 'leading challenge for digital forensic investigations is that of scale.' As the complexity of forensic data increases, forensic tools must adapt to have a broader focus. Rather than concentrating on first-order information, which is merely presented by the current tools in volume to an analyst who then has to understand the information, forensic tools should correlate information from multiple sources, bringing together information generated by a variety of processes (Case et al., 2008). Nance et al. (2009) outline the results of a brainstorming session from the 2008 Colloquium for Information Systems Security Education addressing the development of a research and education agenda for Digital Forensics, which included the challenge of 'Data Volume', stating it is 'common for digital forensic investigations to be overwhelmed with massive volumes of data'. Riley et al. (2008) examine the time involved in imaging common (for the time) hard drives, and state that the trend of increasing sized drives presents a challenge to forensic investigators, and with the combination of increasing crime and shortage of examiners there is a large backlog of processing and analysis of evidence.

Casey et al. (2009) highlight that the growing size of storage, the variety of devices, increasing caseloads, and limited resources, 'are combining to create a crisis for Digital Forensics Laboratories' with many labs having 6–12 month backlogs. Beebe (2009) lists 'Volume and Scalability' as an important strategic direction for research, as the size of data is ever increasing. Ayers (2009) states that existing forensic software and tools are becoming inadequate due to complexity and increasingly large volumes of data. Cloud computing is discussed by Biggs and Vidalis (2009) who highlight that it will likely become a major issue in relation to creation, storage, processing, and distribution of illicit material. Garfinkel et al. (2009) state that there is a need for forensic tools which can reconstruct, analyse, cluster, mine, and make sense of the increasing variety and scale of data.

Turnbull et al. (2009) undertook analysis of the figures from the South Australia Police Electronic Crime Section for the period from 1999 to 2008, which showed an average increase of 20% per year growth for the number of requests, and the number of property items also increasing each year, quantifying the generally known growth trend. A white paper from Access Data Corporation discussed the issue of the FBI cybercrime labs reported to having a large backlog of cases which delayed investigations, and quoted FBI Executive Assistant Director Stephen Tidwell as saying '[t]he pervasiveness of the Internet has resulted in the dramatic growth of online sexual exploitation of children, resulting in a 2,000 percent increase in the number of cases opened since 1996' (AccessDataCorporation, 2010). It is further stated that 'it's not only the number of delayed cases that make this an urgent matter. It is the nature of most of these cases that dramatically increases the pressure on computer forensics labs to implement more efficient policies and practices to overcome this issue.' Biggs and Vidalis (2009) also discuss paedophile criminal activity, stating that this crime type accounts for between 70 and 80% of an investigators workload, and that 'the cloud could prove to be a haven that the paedophile may wish to exploit to fulfil his heinous needs. If data content is not monitored by cloud vendors, then this type of remote storage and relative anonymity of the cloud account holder, may further stretch law enforcement resources beyond breaking point.'

Volume of data (2010–2014)

Casey (2010) discusses the pervasiveness of electronic devices in society, in particular the major challenge to keep pace with new developments. Gogolin (2010) surveyed law enforcement agencies in Michigan, USA, and was reported that many agencies reported that 50% of cases have a digital component, and many digital forensic labs had backlogs approaching or exceeding two (2) years, with average caseloads far exceeding average case analysis timeframes.

Garfinkel (2010) discusses the potential challenges of the next decade, and alludes to the growing size of storage devices as 'the coming digital forensic crisis', where due to the growing size of storage devices there will be difficulties in imaging, processing, and assembling terabytes of data into concise reports. The focus of forensic software has been on thoroughness, and minimal development has focussed on performing rapid analysis, i.e. within 5-min (Garfinkel, 2010). However, many commercial forensic software companies have taken steps to address the changing nature of digital evidence and the challenges of the growing volume of data by radically changing the method of storing case files and processing data (AccessDataCorporation, 2010; DFI_News, 2011). However, this does not necessarily alleviate or address the volume data issue.

Garfinkel (2012b) highlights the totality of information which makes digital forensic software development distinct from other software development due to; the diversity of data, the volume of data, the need to be using the latest software and operating systems, the pressure on a small group of human practitioners, the time to train new practitioners and programmers (~2 years), and unreasonable expectations of clients and judiciary, the so-called "CSI effect." Bhoedjang et al. (2012) state that individual cases often consist of terabytes of data, which could be reviewed by investigators with few technical skills, freeing up technical specialists for higher level tasks, and that '[d]ata volumes and processing times have increased to the point that desktop processing is no longer cost-effective.'

The volume of data is exacerbated by the nature of electronic evidence, in that, top-of-the-line systems are seized and presented for analysis, hence forensic labs need to be equipped with top of the line systems to undertake analysis in a few hours on what has been in use for months or years, concluding that 'we will never get ahead of the performance curve' (Garfinkel, 2012b). Jones et al. (2012)

outline the increasing demand placed on the Australian New South Wales Police Force State Electronic Evidence Branch for digital forensic support, requiring the development of a sampling process to reduce the time spent on analysis of Child Exploitation Material investigations.

Casey et al. (2013) discuss the increase in the number of cases and the increase in the amount of data for each examination, highlighting the need for more efficient tools and processes. Overill et al. (2013) highlight the growing gap between the demand for services and the capability in law enforcement digital forensic units, and with public sector budget cuts in the current economic climate there ‘is no realistic possibility of increasing the level of resourcing to match the ever-increasing demand.’ This is supported by Shaw and Browne (2013), who also state that forensic labs globally are failing to keep pace with demand for services, and that forensic triage is suggested as a method to deal with the “big data” problem. Roussev et al. (2013) examine the process of triage, and highlight that storage capacity is increasing which will continue to put pressure on forensic tool developers, with a call to focus on real-time processing at the time of imaging, preferably processing is undertaken at the same time as imaging, so the bottleneck is I/O speed, but conclude that ‘with current software, keeping up with a commodity SATA HDD at 120 MB/s requires 120–200 cores.’

Quick and Choo (2013) discuss the increasing use and availability of cloud storage, which can, and is, being used by criminals and criminal organisations, further adding to the complexity of the growth in data, and impacting on timely analysis. Alzaabi et al. (2013) discuss the growth in storage capacity and decreasing cost of devices, and whilst there are tools and techniques to assist an investigator, the time and effort to undertake analysis remains a serious challenge. Raghavan (2013) states that an ‘exponential growth of technology has also brought with it some serious challenges for digital forensic research’, and the ‘volume problem’ is the ‘single largest challenge to conquer’.

van Baar et al. (2014) discuss the increasing number of devices and caseload, and outline the shortcomings of many current forensic lab processes, such as having forensic practitioners involved in low level tasks such as network management, system administration, imaging, and other tasks, not necessarily utilising their higher level skill sets for analysis tasks. Breitinger and Roussev (2014) state that ‘one of the biggest challenges facing a digital forensic investigation is coping with the huge number of files that need to be processed’ and that *known file filtering* or *hashing* is difficult to maintain as the underlying data within files is altered on a regular basis resulting in reference databases becoming obsolete. Breitinger et al. (2014) reinforce this, stating the time to compare known hash values using current processes is time consuming, and that improved processes can achieve the same result in much faster timeframes.

Vidas et al. (2014) highlight that the increasing volume of data is leading to backlogs of cases within forensic labs, and analysis times of days or weeks, which is counterintuitive to the need for fast analysis in many cases. Noel and Peterson (2014) discuss the ‘big data problem’ which is complicating digital investigations, resulting in poor decision making, lost opportunities, failing to discover evidence, and potentially loss of life.

Growth of media

In Survey section we outlined the many digital forensic papers discuss the problem of increasing volumes of data and devices, in this section we will focus on the growth of media. Moore's Law is the observation of an average doubling of the number of transistors on an integrated circuit every 18–24 months, which assists in predicting development of computer technology (Wiles et al., 2007). Kryder (as quoted by Walter, 2005) made the observation that in the space of 15 years, the storage density of hard disks had increased 1000 fold, from 100 million bits per square inch in 1990, to 2005 when 110 gigabit drives were released by Seagate. Kryders' Law equates to storage density of hard drives doubling every 12 months, holding true since 1995 (Wiles et al., 2007). The doubling of hard drive size is about twice the pace of Moore's Law (Coughlin, 2001). This highlights that storage capacity is doubling roughly every year and the capacity to process data is doubling every 18–24 months, leading to a potential growing gap in the capability to process data. Roussev and Richard (2004) discuss this in relation to the growth in CPU speeds in comparison with I/O transfer speeds, and observe that I/O speeds have not kept up with CPU speeds, and if they had, ‘it would take about the same amount of time to image a 20 GB drive as it would to image a 200 GB drive. Anybody who has tried a similar experiment knows that this is absolutely not the case.’ Roussev et al. (2013) undertook further study in relation to real-time processing of hard drive data, and concluded that the volume of data is increasing exponentially, and the amount of processing required is increasing faster than the capability of workstations and many forensic tools.

In 1999, a 10 GB hard drive was considered a large amount of data (McKemmish, 1999). Culley (2003) stated that it was not uncommon for computer evidence to consist of a terabyte or more in 2003. Alink et al. (2006) describe testing of their prototype system on forensic images ranging from 40 to 240 GB. Wiles et al. (2007) state that as companies and government agencies are now storing petabytes of data, it is possible that large digital forensic cases may approach this volume of data. Pringle and Sutherland (2008) stated that in 2007 it took approximately 24 h to image a 500 GB hard drive. When Riley et al. (2008) reviewed the imaging times for common hard drives, they used 80 GB, 120 GB, and 250 GB hard disk drives, with imaging times ranging from 30 min to nearly 2 h Zimmerman (2013) also conducted testing of imaging times, using a 1 TB hard drive for a range of testing, with time ranging from 2.5 h to 5.5 h.

Craiger et al. (2005) report on the volume of data examined by the United States Federal Bureau of Investigation (FBI) Computer Analysis Response Team (CART) for the Years 1999–2003, as presented to the 14th INTERPOL Forensic Science Symposium, and calculate the number of cases increasing threefold, and the volume of data increasing by forty-six times, stated to be many times the volume of data in the largest library on Earth, the United

States Library of Congress. Averaging the number of cases with the data burden (Table 1) shows an increase in average per-case size from 8 GB in FY1999 to 119 GB in FY2003. In 2002 it was stated that fifty percent of the cases opened by the FBI involved a computer (Peisert et al., 2008). In 2010, fifty percent (50%) of cases in the Michigan, USA area involved digital evidence (Gogolin, 2010).

In a further effort to assess the growth in digital forensic data volume, information from the FBI Regional Computer Forensic Laboratory (RCFL) Annual Reports from fiscal year (FY) 2003–2012 was reviewed. The data and figures in the reports were compiled and are summarised in Table 2 (FBI_RCFL, 2003–2012). The figures show growth in the total volume of data, from 82 Terabytes (TB) in 2003 to 5986 TB (5.8 Petabytes) in 2012, an overall increase of an average of 67% per annum. In FY2003, the average size for a case was approximately 83 GB, which has grown to approximately 699 GB in FY 2012.

Roussev et al. (2013) outline the acquisition rates based on maximum sustained throughput of hard drives to approximate that, in 2003 a 200 Gb hard drive should take about 1 h to image at 58 MB/s, and in 2013 a 3 TB hard drive to take almost 7 h at 123 MB/s, but they then state that in reality a 3 TB hard drive actually took over 11 h to image. They demonstrate that imaging and processing times are increasing, and that this will continue as I/O improvements do not keep pace with volume increases, and conclude that current workstation hardware does not offer enough processing power to keep up with a SATA hard drive.

The International Data Corporation has stated that the world's volume of data doubles every 18 months, and the cost of storage has fallen dramatically; for example, a 2 terabyte (TB) hard drive in 2010 costing the same as an 18 gigabyte (GB) hard drive in 1998 (Wong, 2010). A single megabyte (MB) of hard drive storage in 1956 was US$10,000, which dropped to US$300 by 1983, US$0.07 in 1998, and US$0.0005 in 2010 (with a 250 GB drive costing $125) (Growchowski, 1998; Wong, 2010). According to PC World, hard disk drives, which were introduced in 1956, took 35 years to reach one gigabyte, 14 more years to reach 500 GB, and only two more years to double to one terabyte (Reyes et al., 2007).

Sommer (2004) wrote that a typical retail PC at that time contained an 80 GB hard disk drive, which was an increase from 20 GB the previous year. In the same year, Roussev and Richard (2004) state that a 200 GB hard disk cost approximately $165, and that a terabyte-class storage system cost under $1000. When Turner (2005) outlined a method of selective imaging to deal with the so-called ‘imminent doomsday’ relating to the volume of data, a single hard disk drive at that time was stated to be ‘in excess of 350 Gb’. In 2008, 2TB of storage was available for under $500 (Riley et al., 2008).

Table 1
FBI CART examinations (INTERPOL, 2004).

US FY	1999	2000	2001	2002	2003
Caseload	2084	3591	5166	5924	6546
Data burden (TB)	17	39	119	358	782
Average case size (GB)	8	11	23	60	119

In 2014, 4TB hard drives are now available for less than US$150 (Walmart, 2014). Casey states that 200 GB is capable of storing about 4,000,000 pictures. Extrapolating this, 4TB would be capable of storing 80 million pictures, an enormous volume of pictures for an examiner to view and classify. In addition, it is possible to store 120,000 min of compressed video on a 4 TB hard drive (Bell, 2013). To put that into more comprehensible numbers, this equates to 2000 h, or 83 days' worth of video, which for an examiner (working 9 am to 5 pm Monday to Friday) would take 50 weeks to play every video file in full.

Data from the South Australia Police (SAPOL) Electronic Crime Section (ECS) for 2013 indicates cases with hard drive sizes of 1–2 TB for single drives, with some up to 4 TB, and some computers containing half a dozen terabyte size hard drives in a RAID configuration. The majority of cases have multiple computers, storage media, and portable devices per investigation, often comprising many terabytes of data in total (Authors' compilation). This raises questions regarding hard drive and storage media sizes over the last 10 years in comparison with the sizes of hard drives seized for digital forensic analysis, and the average number of devices per investigation over this period of time. It is, perhaps, timely to revisit the work of Turnbull et al. (2009), with updated data, to determine the data volume trend since 2009.

Table 3 compiles the information discussed in this section in a timeline format, highlighting the average sizes of hard disk drives, the average memory (RAM) in consumer computers, milestones for technology such as USB, IDE, SATA, Windows operating systems, Intel processors, and includes the FBI CART and RCFL average case volumes (from Tables 1 and 2). Fig. 1 charts the general trend in the volume of Hard Drives with the growth in RAM of Apple Mac Pro computers and the average case size for the FBI CART and RCFL from Tables 1–3. A logarithmic scale base 2 is used as the increase in hard drive size is greater than the RAM and FBI sizes, skewing the chart when a standard scale is used.

In relation to processing the imaged forensic data, Roussev and Richard (2004) outline processing and indexing timeframes using FTK 1.43a compared with a prototype distributed system over a six gigabyte (6 GB) forensic image (which would now be considered quite small) and discussed timeframes of about 2 h when using FTK 1.43a, and over four days for an 80 GB hard disk drive, using a ‘high-end’ Windows XP machine consisting of a 3 GHz Pentium 4 processor with 2 GB of RAM. Even at the time, when extrapolating these timeframes, the processing time becomes a major bottleneck to investigations. Riley et al. (2008) report on research conducted in regard to indexing an 80 GB hard drive, stated at the time to take more than four days, and conclude that the time to index 2 TB of data would be over two months, although they state the system resources would fail before it could complete. At the time of writing (June 2014), processing and indexing this volume of data is possible with current forensic tools, and, whilst time consuming, is not too onerous. However, there is a lack of current or historical research examining processing times of common amounts of data using common processing solutions. SAPOL ECS figures (and the authors' personal experience) indicate

Table 2
FBI RCFL annual reports 2003–2012 (FBI_RCFL, 2003–2012).

US fiscal year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Service requests received	1444	1548	3434	4214	4567	5057	5616	5985	6318	5060	43,243
Examinations conducted	987	1304	2977	3633	4634	4524	6016	6564	7629	8566	46,834
TB processed	82	229	457	916	1288	1756	2334	3086	4263	5986	20,397
Average case size (GB)	83	176	154	252	278	388	388	470	559	699	

Table 3
Timeline of developments of computers and hard drives.[a]

Year	HDD	FBI	FS & USB	Common operating systems, processors, and PCs		
1956	RAMAC 1 MB cost $10k					
~	~					
1977						Apple][(4 KB RAM) CBM PET
1980	IBM 1 GB HDD for $40k		FAT12			
1981	5 MB Seagate HDD			MSDOS	IBM Acorn	VIC20 (5 KB RAM)
1982					286	C64 (64 KB RAM)
1983						
1984						Apple MacIntosh (128 KB RAM)
1985					386	Amiga 1000 (256 KB RAM)
1986			IDE HD	MS Windows		
1987	20 MB/40 MB		FAT16			Amiga 500 (512 KB RAM)
1988						
1989					486	
1990	40 MB 3500 RPM			Windows 3.1		Amiga 3000 (2 MB RAM)
1991	1 GB					Apple Powerbook (2 MB RAM)
1992	2.1 GB		FAT32			Amiga 4000 (2 MB RAM)
1993					Pentium	
1994			NTFS			
1995	3.2 GB 5400 RPM			Windows 95		Apple PowerMac (8 MB RAM)
1996						Apple PowerBook (16 MB RAM)
1997	16.8 GB		USB1.1		Pentium II	
1998	18 GB HDD	(GB)		Windows 98		Apple iMac (32 MB RAM)
1999		8			Pentium III	Apple iBook (32 MB RAM)
2000		11	USB2		Pentium 4	
2001		23		Windows XP	Xeon	Apple OSX
2002	40 GB 7200 RPM	60				
2003		119		AMD 64 bit … Pentium M		
2004		176	SATA HD			Apple iBook G4 (512 MB RAM)
2005	500 GB HDD	154				Apple iMac G5 (256 MB RAM)
2006	750 GB	252		Windows Vista		Apple Macbook Pro (1 GB RAM)
2007	1 TB	278				Apple Mac Pro (Intel) (2 GB RAM)
2008	1.5 TB	388	USB3			Apple Mac Pro (3 GB RAM)
2009	2 TB	388	SATA3	MS Windows 7		
2010		470			2nd Gen Intel	Apple iPad
2011	4 TB	559				
2012		699		Windows 8	3rd Gen Intel	Apple Mac Pro (6 GB RAM)
2013	5 TB		USB3.1			
2014	6 TB					Apple Mac Pro (12 GB RAM)

[a] Sources: http://support.apple.com/kb/SP24, http://support.apple.com/kb/sp43, http://support.apple.com/kb/sp169, http://support.apple.com/kb/SP506, http://support.apple.com/kb/SP69, http://www.apple.com/au/30-years/, http://www.apple.com/mac-pro/specs/, http://windows.microsoft.com/en- AU/windows/history, http://www.commodore.ca/history/company/chronology_portcommodore.htm, http://www.computerhistory.org/timeline/?category=cmptr, http://www.intel.com/content/dam/www/public/us/en/documents/corporate-information/history-intel-chips-timeline-poster.pdf, http://www.livescience.com/20718-computer-history.html, http://www.pcworld.com/article/127105/article.html, http://www.seagate.com/au/en/about/company-information/?navtab=company-history, http://www.tomshardware.com/reviews/15-years-of-hard-drive-history,1368-2.html, Growchowski (1998), McKemmish (1999), Reyes et al. (2007), Sommer (2004), Wong (2010).

examiners often have many terabytes of data for a single investigation, which is spread over a multitude of devices, with various file and operating systems (Authors compilation). The total time to image and process the media in these cases is lengthy. The capability of a human examiner to understand the massive volume of data is not able to keep pace with the ability to gather and store the data (Fayyad et al., 1996a). Hence, new techniques are required in relation to undertaking digital forensic copying and digital forensic analysis.

In general, digital evidence must be relevant and credible, and the collection, fusion and correlation of

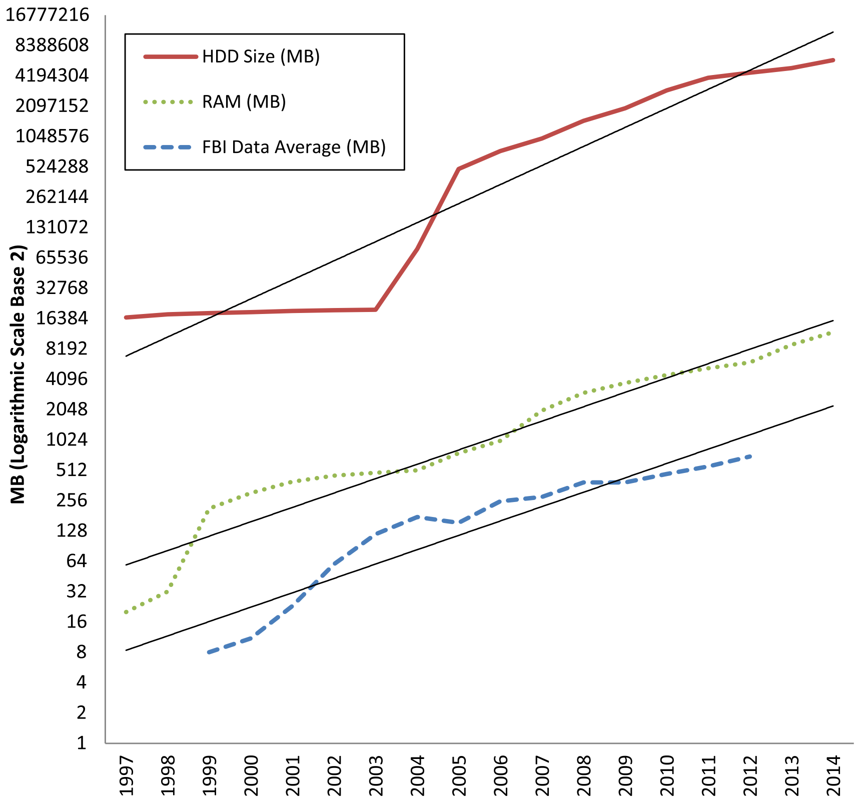

Fig. 1. General trend in HDD (MB), RAM (MB) and FBI case size (MB) between 1997 and 2014.

disparate data is vital to digital investigations (Palmer, 2002). The challenges do not only relate to processing the volume of data in a timely manner, but also copying and storing greater and greater volumes of information (McKemmish, 1999). Mobile portable devices, such as mobile phones and tablets, also complicate forensic examinations. Mobile phones are a particular problem due to the large variety of manufacturers, models, and operating systems, making it difficult to gather a full physical copy of data, and relying on logical access to collect data (Reyes et al., 2007). The pervasiveness of these mobile devices results in many criminal and civil investigations having electronic evidence.

McKemmish (1999) stated that as the size of hard drives increases, so too does the amount of data stored, including multimedia, and the demand for greater sizes is also fuelled by the Internet and access to media content to the average user. Palmer (2002) discusses challenges to digital forensic analysis includes the trend towards disparate information, with networked systems and devices hosting data, potentially anywhere in the world. Wiles et al. (2007) observed that as the size of media increases, users are storing more, including; every email, document, spreadsheet, picture and video they have, and as storage is easily purchased and inexpensive, there is little need to organise their data. Cloud stored data poses difficulties to forensic analysis, as there can be large volumes of data hosted in disparate locations (Quick et al., 2014).

Proposed solutions

Data mining

Spafford (as cited in Palmer, 2001) listed data mining as a field of specialty which may assist in digital forensic analysis. Beebe (2009) also stated that the use of data mining techniques may be another solution to the volume challenge, and also that data mining has the potential to locate trends and information that may otherwise be undetected by human observation. Beebe also raised a list of topics for further research, including;

- a method for implementing subset collection,
- how data mining research can be extended to digital forensics,
- what adaptions are required to apply data mining to forensic datasets, and
- whether link analysis techniques be applied to digital forensic data (Beebe, 2009).

Data mining is the process of 'extracting useful information from large datasets or databases' and utilises various fields of information processing, including; statistics, machine learning, data management, pattern recognition, and artificial intelligence. Data mining has arisen from the interest in tapping the growing volume of large databases, in diverse business areas, such as retail transaction data, credit card use records, telephone call data, and government statistics' (Hand et al., 2001). Data mining methodologies have been applied to security and intrusion detection, an associated field to digital forensics, to discover patterns of user behaviour to recognise anomalies and intrusions (Lee and Stolfo, 2000). Data mining has been used in a variety of areas with large datasets, is useful to summarise incomplete data, and can be used to potentially identify user behaviour to develop user profiles (Abraham, 2006). There is potential to apply data mining techniques to digital forensic data to assist with the data volume challenge, and for knowledge and intelligence discovery purposes.

Brown et al. (2005) developed a process of mining forensic data, with the focus on picture analysis and detecting and filtering partially clothed persons from other pictures using colour space and vector composition. In conjunction with other filtering systems, this process can greatly reduce the amount of information a practitioner must sift through. Huang et al. (2010) outline a framework using ontology matching and machine learning to gather knowledge from large volumes of digital evidence by matching conceptual models to enable data mining and knowledge discovery. They propose the use of a Bayesian Networks approach, but state that whilst rule based algorithms are fast, there is a potential to miss information (Huang et al., 2010). Both offer solutions, but apply only to small aspects of the overall digital forensic analysis process, and do not offer comprehensive methodology which can be applied across the whole range of information and evidence required to be analysed.

Data mining has the potential to benefit digital forensics in relation to reducing the processing time, improving the information quality, reducing the cost of analysis, and improving the ability to discover patterns which may otherwise remain unknown (Beebe and Clark, 2005). However, there are limitations, as raised by Shannon (2004) in relation to missing important information. Beebe and Clark (2005) also state that additional limitations relate to the untested nature of applying data mining techniques to forensic data, and the general lack of understanding of data mining techniques in the field of digital forensics. To address the limitations, there is a need to increase the awareness of data mining techniques in the digital forensic community, train examiners in data mining techniques, and to create a framework for using data mining techniques in examinations, calling for active research to extend data mining to digital forensics and investigations (Beebe and Clark, 2005).

The process of understanding data is known by a variety of names, such as; data mining, 'knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing' (Fayyad et al., 1996a, 1996b). Data mining is a step in the process of Knowledge Discovery in Databases (KDD) which is a process of extracting useful information from the growing volume of digital data, which is used in a variety of fields, such as; business, manufacturing, scientific, and personal information (Fayyad et al., 1996a). KDD is a process of understanding data, and the methods and techniques to do this, which are addressed by mapping low-level data which is too large to easily understand, using specific data mining methods relating to pattern discovery and extraction. The nine-step process of KDD, namely: (1) learning the domain, (2) creating a target dataset, (3) data cleaning and preprocessing, (4) data reduction and projection, (5) choosing the function of data mining, (6) choosing the data mining algorithm, (7) data mining, (8) interpretation, and (9) using the knowledge (Fayyad et al., 1996b), is visualised in Fig. 2;

The knowledge discovery (KD) process is aimed at identifying potentially useful patterns from large collections of data, and includes data mining as a step (Cios and Kurgan, 2005). KD is the overall process of preparing, mining, verifying, and applying the discovered knowledge from data. There are a variety of data mining algorithms with different techniques suited to different problems, and a variety of methods to determine the appropriate algorithm to solve problems (Fayyad et al., 1996b). The general approach of data mining can be for predictive and/or descriptive reasons, or a combination of both (Fayyad et al., 1996b). In addition, the technique of 'Content Retrieval' deals with extracting information from complex, semi-structured, or unstructured data (Hand et al., 2001). Content retrieval techniques are usually applied to the domain of textual data, multimedia data, Internet data, spatial data, time-series or sequential data, and complex data (Beebe and Clark, 2005). This has, perhaps, the most promise in relation to digital forensic data.

Hearst (1999) discusses the process of extracting meaning from large text collections, which are difficult to decipher, and outlines the application of information access and corpus based computational linguistics to enable text data mining. Digital forensic evidence can consist of structured and unstructured data, noisy and clean, and from a variety of sources (Beebe and Clark, 2005). Hence, a variety of data mining techniques may be required. A text mining process across an information repository also has potential to assist examiners (Weiser et al., 2006).

Hand et al. (2001) discuss the various methods of predictive, descriptive, and content retrieval data mining. All three methods have potential application in relation to the data mining of digital forensic data holdings, and content retrieval is, perhaps, the most promising due to the variety of structured and unstructured data common to digital forensic evidence. Beebe and Clark (2005) state that descriptive data modelling is likely to have limited application in digital

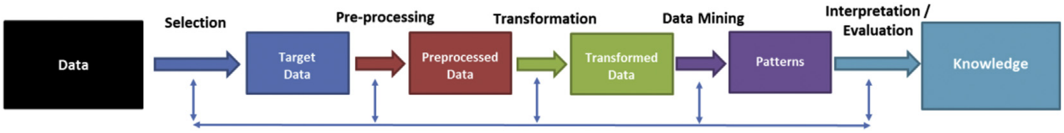

Fig. 2. Overview of the steps constituting the KDD process. Adapted from Fayyad et al., 1996b.

forensic analysis due to issues around data loss, but that the process may have application in relation to internal investigations and military investigations.

Content retrieval, also known as text mining or information retrieval, is well researched, due to the recent increase in demand for business intelligence, and the increase in data availability (Beebe and Clark, 2005). Shannon (2004) outlines a content mining technique called Forensic Relative Strength Scoring (FRSS) in which ASCII proportionality and entropy calculation is used to create a value for text and data, and this is used to filter the data to locate information which can be of benefit to an examiner. However, Shannon (2004) states that there is a chance that the FRSS process will miss important information, and is intended to be a guide, and not to be depended on to locate information. Hence, there are potential issues in relation to using this technique when applied in a legal environment. Noel and Peterson (2014) propose the use of natural language processing method of Latent Dirichlet Allocation (LDA) to process user data within forensic data holdings. However, in testing they conclude that the LDA process is much slower than a regular expression (regex) search, with an example of a regex search taking 1 min in comparison with the LDA process taking over 8 h. In a situation where a life is at risk, the large difference in time to process may be hazardous.

Business analytics is a process of designing the needs of data analysis into business systems from the outset, specifically including and addressing the areas of; data collection, data generation, data storage, and integration from multiple sources (Kohavi et al., 2002). This field provides methods to implement data analysis within entire business systems, and aspects of this may have application in the overall setup of a digital forensic analysis environment. However, as digital forensic analysis has differing requirements to the usual fields that these techniques are applied to, there is a need for specific research to address the unique issues applicable to forensic analysis. The steps of the business analytics process align to the processes described for KD, KDD, Intelligence Analysis, and Digital Forensics. The process of business analytics is visualised in Fig. 3;

Data mining methodologies can assist with analysis of digital forensic data, and the sub-phases of data surveying, extraction, and examination (Beebe and Clark, 2005). Descriptive modelling can be used to profile use and activity during the initial data survey phase. Classification techniques can be used to reduce the volume of data to be analysed, and entity extraction techniques, content retrieval, and link analysis follow this (Beebe and Clark, 2005). Link analysis can be used to visualise associations amongst entities identified in the content retrieval stage. The crime data analysis phase can be broken up into sub-phases of transformation, detection, and visualisation. During the transformation stage, data from the various disparate sources is converted to a common format to develop an understanding of the data. Next, predictive mining techniques can be used to identify clusters or groups of entities, associations, and networks, using visualisation techniques to display and analyse the information (Beebe and Clark, 2005).

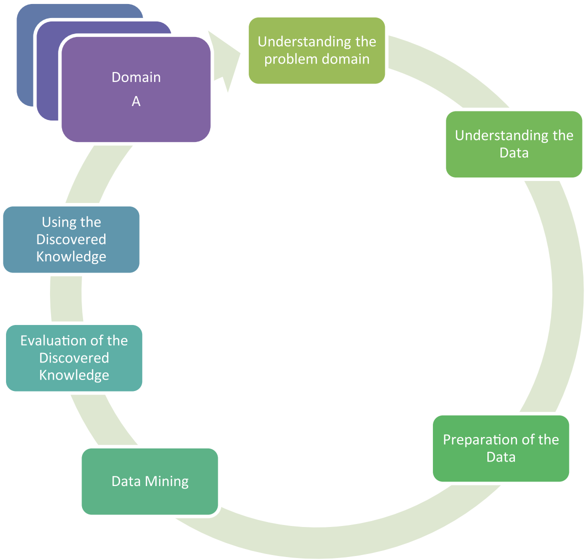
Domain
A
Understanding the problem domain
Understanding the Data
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
Using the Discovered Knowledge

Fig. 3. Business analytics. Adapted from Kohavi et al., 2002.

Data mining processes have potential application to assist forensic examiners. However, an overall method to apply this to forensic examinations remains a gap. Okolica et al. (2007) discuss the use of the Author Topic model to identify the interests of a person from email text in an effort to identify insider threats. Iqbal et al. (2008) outline a method for applying data mining techniques to email text to identify authors in an endeavour to locate fraudulent messages. Iqbal et al. (2010) build on this work to identify writing styles in an email collection.

Casey (2014) comments that as the volume of data increases, more attention is placed on research using big data solutions and visualisation techniques. Garfinkel (2012b) advocates that the difficulty in applying “big data” solutions to digital forensic data is due to the diverse nature of forensic data, such that solutions for other problems often consist of large amounts of data which can be reduced and locally processed, and that state based agencies do not have the budget or personnel to utilise the data processing methods of large world-renowned physics labs. Hence data mining alone may not provide a complete solution to the growing volume of digital forensic data, and other methods may need to be combined for an overall solution.

Data reduction and subsets

Data reduction is a step prior to data mining, and in this era of vast data volumes seized for analysis, the application of this could have potential in digital forensic examinations. Spafford (as cited in Palmer, 2001) stated that there is a need to understand what information needs to be collected to provide an accurate analysis for a particular circumstance. This alludes to a process of not collecting everything, but to focus collection of that which will provide an accurate analysis. ACPO (2006) also support this, advising that ‘[p]artial or selective file copying may be considered as an alternative in certain circumstances e.g. when the amount of data to be imaged makes this impracticable. However, investigators should be careful to ensure that all relevant evidence is captured if this approach is adopted.'

Culley (2003) raised the issue of the efficiency of imaging entire hard drives, but stated that full imaging remained the desirable and thorough option at that time. Now that larger datasets are becoming the norm (and often referred to as 'Big Data'), and logical imaging is implemented in commercial forensic analysis software, it is timely to consider the process of collecting targeted data within the larger volume of evidence and the potential time and resource savings this can provide. Pollitt (2013) mentions "sufficiency of examination", stated to be coined by Mike Phelan, late Director of DEA, alluding to doing enough examination to answer the required questions, and no more. Quick and Choo (2014) outline a Data Reduction Framework which runs before, in parallel with, or subsequent to, full forensic analysis, and serves to collect a reduced subset of data for the purpose of a fast review, intelligence analysis, archiving, and future enquiries – see Fig. 4 in Growth of media section. Collecting a data subset is potentially easily implemented, with minimal changes to current systems and procedures, and can be undertaken in a forensically sound manner when abiding with common forensic principles (e.g., ACPO, 2006 or NIJ, 2004).

With the volume of data increasing as discussed, one method to address this is to only image and index selected files, and exclude files and data which does not assist an examiner. There is a potential to reduce time and resource requirements by making decisions prior to imaging to only examine that which may be relevant. Ferraro and Russell (2004) state that, in relation to child exploitation image investigations, it should be sufficient to locate the necessary evidence to support a criminal investigation and prosecution, without the need for a full forensic examination for all cases. Turner (2005, 2006) discusses selective imaging and "Digital Evidence Bags" as a method to store a variety of digital evidence whilst retaining information relating to the source and location of the data subset. Schatz and Clark (2006) in building on this, proposed the concept of a "Sealed Digital Evidence Bag", providing for referencing between evidence bags to build a corpus of evidence, using the Resource Description Framework (RDF) to annotate information contained within datasets. Alzaabi et al. (2013) describe the use of RDF when applied to data obtained from Android devices and the use of an ontology based system to interpret data traces for analysts. Commercial forensic software such as Guidance Software EnCase and AccessData Forensic Tool Kit now provide the capability of selectively imaging files to support the collection of subset data into logical evidence files.

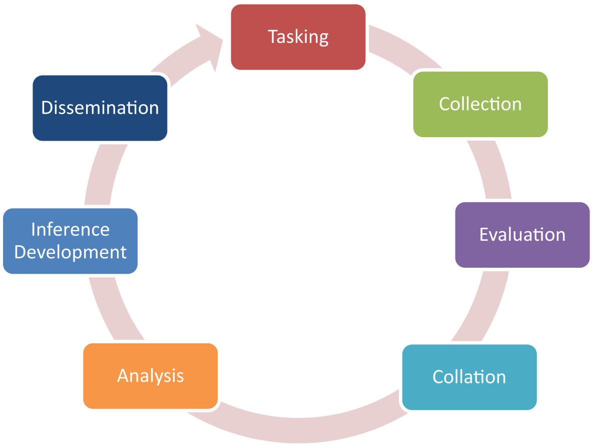

Fig. 4. The intelligence cycle. Adapted from UNODC, 2011.

Garfinkel (2012a) discusses Digital Forensics XML (DFXML) as a method to store common forensic information as a generic data store for a range of tools, enabling parsed data to be stored in a common format to enable greater analysis across a variety of information types. XIRAF (XML Information Retrieval Approach to digital forensics) uses XML as a format to store the output of a range of forensic tools (Alink et al., 2006). It is also stated that data reduction is a key approach with XIRAF, along with caching, parallel processing, and close interaction with users (Bhoedjang et al., 2012).

The process of selecting which files to image can be a challenge (Beebe, 2009). Many child exploitation investigations require additional and thorough analysis to determine what actions a user has undertaken with specific files. In these cases, a variety of other files become important, such as Link files, Prefetch files, Windows Registry files, and other structured and unstructured data. Some of this information may be in unallocated space and, hence, access to the full image may still be necessary. An examiner could be criticised if potential evidence has been discarded due to the perceived cost in time and resources of making a full forensic image, and conducting thorough analysis. There is still a great opportunity for the process of data subsets and selective imaging (Quick and Choo, 2014). Whilst in some instances this does not entirely replace the need for full analysis, a process of data reduction can serve to enable a rapid initial analysis.

Child exploitation investigations are one of the most investigated categories for digital forensic analysis (Turnbull et al., 2009). These investigations have a general focus towards pictures, videos, internet chat, and browser history. With these investigation types, a question is raised whereby; "Is it necessary to image entire hard drives, especially when these drives are in the multi terabytes and can often contain known data of little relevance to the investigation (e.g. movie files, television shows, and music)?". Richard and Roussev (2006b) discuss a process of selective imaging by removing known files at the imaging stage, and also only collecting pictures when dealing with a child exploitation investigation. However, there are additional files that have potential evidence for these investigation types, which should be collected and analysed. In some investigations an examiner may need to undertake analysis of a full forensic image if evidence is not located in the subset collected, for example; picture files within an encrypted or compressed file may be missed in a subset collection methodology.

Kenneally and Brown (2005) discuss a 'Risk Sensitive Collection Methodology' whereby selected artefacts are extracted during the initial collection stage, and can be applied to the collection stage when examining so-called 'dead' systems, and also live system extractions. This approach is designed to reduce costs in relation to time and resources, and also evidence storage requirements (Kenneally and Brown, 2005). In addition, Kenneally and Brown (2005) further explain a process of selective imaging to address risks associated with collecting full forensic images from large hard drives, which are stated to be primarily the cost in time and resources. This is done by selecting data to image at the collection stage. Legal standards of reasonableness and relevance are raised to address concerns in not undertaking analysis of a full forensic image. However, it could be argued that if the difference relates to hours or days, in a criminal or civil investigation (which can potentially take many months or years) it could be deemed reasonable to take a full bit-for-bit image, and conduct analysis with all available and potentially relevant data where appropriate.

Garfinkel (2006) introduces Forensic Feature Extraction (FFE) and Cross Drive Analysis (CDA) methods. FFE is outlined as a focus on particular information, such as; email addresses, message information, date and time information, cookies, social security and credit card numbers (Garfinkel, 2006). Information from a scan of a drive is stored for analysis and comparison. However, by interpreting the data and not storing the original, there may be instances where new techniques cannot be applied to the original data. There have been developments in recent years whereby new information is able to be extracted from data holdings that were previously unknown. For example, Windows Registry analysis methodologies include new areas for locating information (Carvey, 2011; Mee et al., 2006). Hence, there is value to storing and retaining original files and the contained data within a logical evidence file to allow for subsequent processing in the future, as and when new tools and processes are developed.

Greiner (2009) refers to the searching of entire forensic datasets as akin to 'shotgun forensics', i.e. imaging everything and looking for relevant information, and that a more appropriate approach is 'sniper forensics', whereby a targeted approach focussing on pertinent information is used, with investigators knowing what to look for, and when the information is found to stop looking. However, in some investigations, there is a chance that information in relation to other offending may be discovered whilst undertaking analysis, and crucial information may be missed by focussing on specific information based on an investigation typology. As an example, in a child exploitation investigation, an examiner may locate Internet chat which provides evidence of "making a child amenable". If the focus is only on the possession of pictures or other media, then this potential offending may be missed, and a child may remain at risk. Hence, the process of data reduction should collect all data that may provide information without necessarily basing this on a type of suspected offending.

Beebe (2009) poses a solution to the volume and scalability challenge as 'selective digital forensic acquisition', or identifying subsets of data for imaging rather than imaging entire physical devices. However, Beebe also states that selective acquisition should include allocated and unallocated space, but does not allude to the reasoning behind this statement, stating that research is needed to determine the process of identifying which files are necessary to acquire. Richard and Roussev (2006a) discuss selective imaging, but state that the number of potentially excluded files may be small, and hence not achieve much of a reduction in volume and, hence, other approaches should be examined. However, Quick and Choo (2014) outline a process of collecting a subset of data which results in a vast reduction in total volume required to be imaged, analysed, and archived (0.196% of initial volume).

Casey et al. (2013) propose a processing method whereby data is fed simultaneously to separate extraction operations, whilst a full forensic image is taken, hence providing the ability to image, verify, extract and carve with a 32% increase in time over standard imaging. They also propose to store extracted information in standardised database format such as XML and SQLite. Whilst this potentially serves the need to undertake analysis of the information, there is still a need to undertake a full forensic image, resulting in this method increasing processing time, and also interprets data for central storage, which has the implication that if the original is not available in future, new techniques to process original data will not be able to be applied to historical data. Roussev et al. (2013) examine the process of real-time digital forensics and triage by researching average times for different collection tasks under a triage scenario, within time constraints, with a focus on collecting the richest and most relevant information, which include; file system metadata, Windows Registry, file metadata, and file content. They outline a method of collecting relevant data by utilising Latency-optimised target acquisition (LOTA) whereby the data is collected sequentially, negating the need for a drive to continually seek to locate data.

Jones et al. (2012) outline the implementation of a method of reducing the volume of material examined in CEM cases within the NSWPOL SEEB by automated scripted sampling of digital forensic images to produce a smaller subset of data, which is reflective of the contents of the entire media. This method has merit, but is limited to investigations relating to CEM possession, whereas in other child related investigations full analysis is still required. Also, the process may be difficult to apply in various jurisdictions due to differing legislation, and does not address victim identification as a random sample of files is taken. In addition, there is no method of collection of common files which would be useful in providing intelligence in relation to the user, or overall trends across cases, although collection of important files can be easily added to the SEEB process (such as Quick and Choo, 2014).

Raghavan (2013) discuses a solution to the volume problem is to use data reduction techniques and remove known and irrelevant data prior to analysis (However, we argue it would be better to remove irrelevant data prior to imaging). XIRAF uses a process of eliminating 'uninteresting' information subsequent to imaging and pre-processing (Bhoedjang et al., 2012). The focus of an applicable data reduction methodology could be to only collect identified data which will provide the greatest information return, rather than collecting everything and then discarding known data. By first collecting only important data, the time savings in relation to collection and analysis are vastly reduced, and the storage requirements are potentially vastly less than storing full images (Quick and Choo, 2014).

Breitinger and Roussev (2014) outline a process of using *bytewise similar* hashing rather than *bytewise identical* hashing, to exclude known data and files, and introduce an automated method to evaluate matching results, but they do not outline processing timeframes, nor whether this is undertaken on full forensic images or data subsets. Breitinger et al. (2014) state the time to compare known hash values using current processes is time consuming and that improved processes can achieve the same result in much faster timeframes, highlighting the results of experiments across 4457 files comprising 1.9 GB of data in the order of seconds rather than minutes. Extrapolating this to current hard drives (with data volumes in the terabytes) results in lengthy timeframes for this process.

Greiner (2009) reminds us that there needs to be a focus on the goal of an investigation, knowing what needs to be answered, what is needed to answer it, how to get the data, and what the data tells us and, importantly, the principle of Occam's Razor, where the 'simplest answer is usually the right one, reminding us not to speculate, but let the data speak for itself.' The benefit of quickly analysing subset of data is that if the first round of collection does locate evidence, the practitioner can finalise the case and move on to the next one. If the first round of collection does not locate data of interest, then a forensic practitioner can return to the original media and conduct further analysis.

Triage

Digital forensic triage is 'a process of sorting enquiries into groups based on the need for or likely benefit from examination' (Parsonage, 2009). It is seen as a way to make greater use of limited resources (Pollitt, 2013). Digital triage has potential to assist with the data volume issue by identifying which items potentially contain evidence. In the context of this research study, the focus is on triage processes applied to identify which item/s from a group are likely to contain evidence, such as a rapid examination by an experienced practitioner to determine which item may contain evidence. The focus is the technical triage process, rather than an Administrative Triage which is applied when determining case acceptance or case prioritisation.

Rogers et al. (2006) outline a triage process model developed in the field on actual cases, court decisions, and prosecutor directions, designed to rapidly locate evidence and identify persons at risk by focussing on user profiles, timelines, internet activity, and case specific information, defining triage as '[a] process in which things are ranked in terms of importance or priority.' LaVelle and Konrad (2007) outline a method for using the Microsoft Windows Robocopy command line tool via a GUI to preserve a subset of data, mainly aimed at large computer servers and storage systems. Casey (2009) highlights the specialised expertise required when undertaking on-site examinations in a corporate or enterprise environment.

Reyes et al. (2007) introduce 'fast forensics' which is the processes undertaken within the first few hours to locate information to use during an interview with a suspect. These usually relate to on-site or field examination to locate evidence and intelligence to assist investigators to use in interview or other searches, but can also be applied in lab situations. Casey et al. (2009) discuss a triage process at a variety of levels, from the administrative decisions about thresholds to accepting cases, to a process of triage consisting of; survey/triage ⇛ preliminary examination ⇛ in-depth examination, whereby cases are assessed and when accepted, decisions made to progress through the levels of examination based on each case and individual requirements. Garfinkel (2010) discusses a method of prioritisation of analysis, in which a triage process is implemented to enable the practitioner to be presented with important information rapidly, and mentions a commercial system that can process media and display contents on a hand-held touch-screen user interface (IDEAL STRIKE) but this system is not widely marketed.

Roussev and Quates (2012) propose a content triage process which is undertaken on forensic images, subsequent to the process of acquiring images of media. Using similarity digests and correlation hashing methodologies, with the aim of building a picture of the contents of forensic images, they apply the process to the M57 case study (Garfinkel et al., 2009) to demonstrate the application of the proposed method across a range of forensic images. We suggest that the process of imaging larger and larger drives places an undue burden on the time to undertake a rapid triage, and hence it would be beneficial to first undertake a data reduction process to rapidly improve the triage timeframe (as per Quick and Choo, 2014).

Overill et al. (2013) outline a process for managing triage within a digital forensic lab with investigations involving multiple devices which is based on the underlying crime type of an investigation and the role of the digital evidence, ie whether it is central (specific), or auxiliary to the investigation. The process utilises automated triage software tools, and that the triage process will stop when a value criteria for a specific investigation type is reached. Shaw and Browne (2013) focus on *technical triage* rather than *administrative triage*, highlighting the risk of inadequately trained personnel reviewing artefacts which require a high degree of knowledge and experience in digital forensic analysis to interpret. The technical triage process they outline involves booting a suspect device with a Linux OS and then conducting a thorough scan of the entire media, parsing file systems and recovering artefacts from unallocated, interpreting data and storing the results in report format. Marturana and Tacconi (2013) propose an automated method which gathers file and data statistical information, which can be applied during *live* or *post-mortem* situations, and provide an examiner information summarising the file types contained on media, such as a hard drive or mobile device.

O'Connor (2004) discusses the use of PXE boot in a corporate environment involving a large number of systems in an endeavour to review systems and identify those potentially containing evidence. Koopmans and James (2013) discuss an automatic triage process using PXE boot as applied to networked computers to locate specific hosts which contain information matching key terms setup from a central server, and highlight specific investigations where the process has been successfully utilised. Vidas et al. (2014) outline the use of OpenLV (previously LiveView) and its application as a triage tool, and whilst this tool enables a forensic image or physical device to be booted in a virtual environment without changing the original data, it does not perform any analysis, relying on an examiner to traverse a computer as a user would, potentially missing data which is not usually presented easily to a user, such as registry data, or information held in other databases (Internet History, Chat History, Deleted Emails, etc). Shiaeles et al. (2013) compare three triage tools (TriageIR, TR3Secure, and Kludge) and report on their suitability, concluding that not one tool is able to fulfil the needs of every situation.

A triage process has potential to quickly identify which evidence items are likely to contain evidential data and, hence, once the identified data is examined and reported, there may be an opportunity to close the investigation and move on to the next one. Software is available to conduct initial triage of digital media, such as ADF Triage, EnCase Portable, and AccessData Triage. Many of these tools interpret the data and present the findings in a report, without opportunity to collect the original files for archival storage, or collect data in proprietary format which doesn't provide for subsequent processing, data mining, or analysis with alternative tools. There is an opportunity to undertake research to address the triage process with a method of collecting and reviewing a subset of data quickly, which serves the purpose of a triage to quickly identify media with evidential data, and also serve to support other forensic analysis purposes; data mining, intelligence analysis, knowledge discovery, archival, and retrieval.

Intelligence analysis and digital intelligence

The discussion thus far has focussed on using digital forensic data for evidence. However, there is a potential to utilise digital forensic data for intelligence and knowledge discovery, or 'digital intelligence'.

Evidence is data which is used to establish proof, whereas intelligence is information which is processed into knowledge designed for action (UNODC, 2011). 'Criminal intelligence is the creation of an intelligence knowledge product that supports decision making in the areas of law enforcement, crime reduction, and crime prevention.' (UNODC, 2011). Intelligence-led policing is a model where crime intelligence and data analysis provide valuable input to a decision making framework in an effort to reduce, disrupt, and prevent crime through strategies and management (Ratcliffe, 2007). A common misconception is that criminal intelligence is surveillance and other covert activities. However, a wide variety of information can be combined with covert information to provide a broader picture to be used by decision makers, and is then criminal intelligence (Ratcliffe, 2007). There is a potential for digital forensic data to provide valuable information as part of the intelligence-led policing process, and criminal intelligence, which can also assist digital forensic examiners.

Three types of criminal intelligence are; Tactical, Operational, and Strategic (UNODC, 2011). Tactical Intelligence supports front line staff and investigators, and is often tied to an investigation leading to an arrest or gathering evidence. Tactical intelligence is mainly short term, arrest-focussed activity directed to front line operational officers. Operational Intelligence sits at a broader organisational level to support area commanders and regional managers in crime reduction activity. This is a mid-level focus to assist with tackling organised crime groups, and for decision makers to determine the priorities for limited resource allocation. Strategic Intelligence aims to provide an understanding into patterns of criminal behaviour and the criminal environment, with a focus on future activities (Ratcliffe, 2007). Strategic analysis is aimed at higher level decision makers, with a longer term focus (UNODC, 2011).

The intelligence analysis process is a cycle of Tasking, Collection, Evaluation, Collation, Analysis, Inference Development, and Dissemination, visualised in Fig. 4 (UNODC, 2011). As outlined in the Criminal Intelligence Manual for Analysts (UNODC, 2011), the tasking phase is designed to outline the scope and requirements. The Collection phase is to identify and collect the data required to achieve the task. Data can be from open source, closed source, and classified information. The Evaluation phase is a process of assessing the source and quality of the information, often by rating information using a scale. Collation is the process of organising the data into a format to allow for retrieval and analysis. Data integration and analysis is a careful examination of the information to discover meaning. Various techniques are used, such as link charts, event charts, activity charts, financial profiling, and data correlation. Then the information is interpreted to determine relevance, and to develop inferences in relation to the key pieces of information. An inference can be as a hypothesis, prediction, estimation, or conclusion, and is then tested before acceptance. Dissemination is then the communication of the intelligence, as a formal reports, formal oral briefing, weekly bulletin, or an ad-hoc briefing (UNODC, 2011).

The application of intelligence analysis techniques can potentially assist digital forensic examiners to process the vast information contained within common digital forensic investigations. Al-Zaidy et al. (2012) apply criminal network analysis techniques to text documents to discover direct and indirect relationships between persons, addresses, and other entities. Garfinkel (2010) discusses the process of 'cross-drive analysis', and states that the perception of casting doubt on the admissibility of evidence has prevented the adoption of this technique. Cross drive analysis is the use of statistical techniques to correlate information within a single disk image and across multiple disk images, which has potential use for intelligence analysis.

Ribaux et al. (2006) discuss how traditional forensic case data can have valuable input to the crime intelligence analysis process, the benefits that can arise from this, and state that forensic science should participate and provide input to crime intelligence holdings. Their discussion relates to traditional forensic information, such as DNA, hair/fibre, blood, and ballistic analysis, and it is further argued that there is a lack of recognition of intelligence needs within the general forensic science community (Ribaux et al., 2010). There is also a great potential to utilise valuable information stored within the multitude of digital devices and data seized for digital forensic analysis, widening the scope of analysis to include criminal intelligence beyond only examining the data for evidential purposes. Weiser et al. (2006) propose a National Repository of Digital Forensic Intelligence, comprising four aspects; Information Knowledge Base, Best Practice, Tools Index, and a Case Index, with the aim of sharing ideas and methodologies leading to efficiency gains, however, impediments to large scale adoption of this include; control of data, confidentiality and classification, task load, and discovery issues. Initial establishment of such an intelligence capability is encouraged within and across agencies, as a first step towards a national data resource (see Australian Criminal Intelligence Database and Australian Law Enforcement Intelligence Network (ACC, 2013)).

Whilst Garfinkel (2010) raises the issue of lost opportunity for data correlation, it is discussed in relation to single investigations involving multiple drives, and the potential for examiners, who are working on one drive at a time, to miss data linkages across multiple drives. Whilst there is a great benefit to investigations to correlate data from multiple drives, there is also an opportunity to use stored subsets of data from a range of cases with intelligence analysis techniques to discover trends across a wide variety of investigations. This intelligence can also be used to provide management level information to determine the appropriate deployment of resources. There is potentially less hesitation to undertake cross-drive analysis if it can be communicated to examiners that forensic principles are not discarded, and traditional principles are observed to ensure the original media is not altered in any way. In addition, the use of data subsets facilitates the process of cross-drive analysis for many reasons, such as to research trends over time, gain tactical intelligence, and locate potential evidence. The data reduction process outlined in Quick and Choo (2014) serves to collect a reduced subset of data for the purpose of a fast review, intelligence analysis, archiving, and undertaking future enquiries.

Raghavan et al. (2009) introduce the concept of an open Forensic Integration Architecture (FIA) to enable the merging of different evidence items and outlines a framework for formalising the analysis of evidence from multiple sources simultaneously. The FIA identifies information from multiple sources to build theories relating to the investigation. However, the case study used demonstrates the concept applied for a single investigation, and is not discussed in terms of undertaking analysis across a variety of disparate investigations. Alink et al. (2006) discuss the integration of mobile phone information into the XIRAF system to match information from other sources, such as forensic hard drive images. Case et al. (2008) propose a framework and software termed ‘Forensics Automated Correlation Engine’ (FACE) which is intended to be an automated correlation of disparate evidence. Garfinkel (2010) outlines a method of Forensic Feature Extraction (FFE) which can be used to locate information within email communications, and potentially identify the primary user of a computer. This technique can be implemented across a variety of data subsets to locate intelligence in relation to communications across a variety of investigations, and can be utilised for tactical or operational intelligence purposes, along with other techniques applied to other data types.

Producing a profile of a suspect is another facet of criminal intelligence. Nykodym et al. (2005) discuss the application of psychological profiling techniques to create profiles in relation to cyber criminals and cyber-crimes. Abraham (2006) discusses the application of investigative or criminal profiling to identify the personality characteristics of a user, such as email authorship analysis. Shaw (2006) outlines the application of psychological profiling with insider threat cases using text analysis to identify suspects based on behaviour. Garfinkel (2010) lists examples of areas to focus methods for representing and analysing information, which includes signature metrics, metadata representation, system information, application profiles, communication information, and user profiles, based on what applications the user runs, when and why. Criminal profiling is therefore an intelligence process which can benefit from the inclusion of digital information, such as digital forensic data. An opportunity exists to research methods of rapidly building a psychological profile of a user, based on the information observed within digital media. This could include information from seized media, such as; websites visited and times, Internet searches, Internet chat, information extracted from documents, spreadsheets and other user created files, the video files the user watches and stores, the music the user listens to, and other such information.

There is also a potential benefit to a digital forensic examiner in liaising with investigators and reviewing intelligence holdings for a suspect or the user of computers and electronic devices. An examiner may be able to speed up analysis with prior knowledge of user, case, and investigation details, and provide greater information to investigators, legal counsel, and other potential beneficiaries.

Other proposed solutions to the data volume challenge

There have been other proposed solutions to the issue of increasing volume of data seized for analysis, including distributed and parallel processing, visualisation, digital forensic as a service (DFaaS), and the use of artificial intelligence techniques.

Parallel and distributed processing offers a potential to speed up analysis of forensic data. Roussev and Richard (2004) examined the application of distributed processing in an effort to deal with the exponential growth in data presented for analysis, outlining a method for dividing the tasks of indexing and searching across multiple workstations to undertake these tasks in a timely manner. Lee et al. (2008) outline the application of a Tarari content processor[1] in relation to searching of data, concluding that there are positive benefits, but that further research is necessary. Nance et al. (2009) state that whilst there have been some developments in relation to implementing parallel

[1] Tarari content processors are designed to distribute processing of data across multiple threads to speed up regular expression search times, and other functions, and are often used in intrusion detection applications (LSI, 2011).

processing of forensic data (Access Data Distributed Network Attack for password recovery), there are other areas which could benefit from parallel processing, such as data carving and generating timeline reports. They also mention that the process of imaging could benefit from parallel processing. However, the process of imaging relies on the speed of the source media, and attempting to have multiple processes calling for access to different sections of a hard drive could confuse the hardware controller and potentially lead to longer read/seek times, rather than speeding up the process.

Ayers (2009) defines a series of requirements for second generation forensic tools, including parallel processing, data storage, and accuracy. Pringle and Sutherland (2008) examine the process of developing a high capacity forensic computer system using grid computing systems. Pringle and Burgess (2014) discuss previous research and highlight the inherent issues with current distributed tools, for example the use of central storage, and propose a cluster based file storage system (FClusterfs) and methodology in an endeavour to balance and distribute processing, with research and testing is still underway. Garfinkel (2012b) states that efforts to utilise multi-threading and high performance computing have been problematic, although some successful results have been achieved. Parallel processing has potential to assist with processing greater volumes of data, but may not scale as rapidly as the rate of data growth.

Marziale et al. (2007) discuss the use of Graphic Processing Units (GPU) in experiments to evaluate offloading processing to a GPU in comparison with a multicore Central Processing Unit (CPU). They conclude that the use of GPUs can provide positive benefits to processing times, especially in relation to binary string searches, and that future research is worthwhile. To date major commercial forensic software companies have changed the underlying methods of storing and processing information, and have introduced some parallel and distributed processing options for their forensic analysis software (AccessDataCorporation, 2010; DFI_News, 2011). As the size of digital forensic investigations continues to increase, the current tools still take considerable time to process the increasing volume of data (Marziale et al., 2007). Hence multiple strategies may be required to address the current data volume challenges, without relying on one single potential solution.

Beebe (2009) discusses ‘intelligent analytical approaches’ and the use of analytical algorithms to reduce the time it takes to locate information, allowing investigators to get to relevant data quickly and reduce the superfluous information in a typical case. Sheldon (2005) discusses concepts for the future of digital forensics, stating that the pace of development in technology affects digital forensics, and that in an ideal world, future digital forensic systems will be able to utilise artificial intelligence to assimilate the contents of digital media, and use inference rules to produce information that can guide an analyst when conducting analysis (at that time, hoped to be possible in 2015). Rules can be established to interpret differing data structures, combined with an ability to learn from each case it is used on, and linkage to a global network of information, to assist examiners (Sheldon, 2005). Hoelz et al. (2009) outline the potential application of Artificial Intelligence techniques to digital forensic data through a multi agent system and case-based reasoning to analyse and correlate data, presenting the results to an examiner, and included the application of a distributed platform in the experiments.

Alink et al. (2006) developed an approach to the volume of data problem which consists of the following; feature extraction and analysis, XML based output from various tools, and XML database storage and query. They developed a prototype system called XIRAF (XML Information Retrieval Approach to digital forensics) which extracts information from forensic images and stores the information in a database, accessed via a web interface. They outline times to process data, including; hashing files can take several hours, extracting EXIF data takes several hours, and parsing unallocated space also takes several hours. Bearing in mind these tasks are undertaken consecutively, this adds up to considerable time to process data, at that time stated to be tested on forensic images ranging from 40 GB to 240 GB. With today's hard drive sizes in the many terabytes, the processing times would be considerably larger if new techniques are not implemented.

Bhoedjang et al. (2012) update the progress of developing XIRAF (six years later) to consume larger data volumes. Search functionality has been implemented, using indexed search methods. Remote access is now available using a web interface, allowing a wider range of clients to access forensic data, including experts and non-experts; analysts, lawyers, and detectives. User management is addressed with the web interface, which also includes project management. The XIRAF system automates many tasks, freeing up specialists for higher level tasks, and by presents information in a common format, removing the need for an examiner to understand the technical details of different forensic software, although they state the amount of data can still be overwhelming. Pre-processing utilises commodity hardware, and is reported to take a day to run a single pass over input data, even with previous tools being merged, and the implementation of tools being run in parallel. However, processing times are stated to be acceptable, and although distributed processing of single images is thought to significantly complicate the system, it is stated to be not clear if this would justify the cost of implementation (Bhoedjang et al., 2012).

van Baar et al. (2014) further discuss the application of XIRAF as a whole of process model, applied in the Netherlands, with implementation as Digital Forensics as a Service (DFaaS). They highlight that traditional forensic labs generally make specialists responsible for a variety of administrative tasks, and hence have less time to perform specialised analysis. In the DFaaS implementation they have a team of support personnel, responsible for administration of; applications, databases, storage, infrastructure, and other systems. This frees up specialists to perform specialised tasks. In relation to DFaaS being applied in a real world situation, they state that it would be beneficial for digital information to be available within the first few days to investigators, but in traditional systems, this is not the case.

Teelink and Erbacher (2006) outline the application of non-hierarchical and hierarchical visualisation techniques to folder tree structures of hard disk drives, and conducted experiments comparing visualisation with text searches, concluding that visualisation methods have potential benefits to forensic analysis. Furthermore in relation to visualisation, Olsson and Boldt (2009) discuss timestamp information and how this is a common denominator for the large variety of structured and unstructured data common to digital forensic holdings, and can be used to create visualisations of digital forensic data. Alink et al. (2006) use the timestamp information from multiple forensic tools to produce timeline visualisations which includes data from multiple tools and evidence in their XIRAF prototype. Stevens (2004) examines time information in relation to correlation from different sources, synchronisation, and a model to simulate clock behaviour to address errors to enable the unification of clock information into single timelines.

Khan et al. (2007) outline a proposed method of generating timeline activity from digital evidence by applying post-event reconstruction using neural network techniques. Raghavan (2013) discusses the challenges of generated a unified timeline across multiple sources of evidence, which include time zone interpretation, clock skew, and syntax. Schatz et al. (2006) discuss time and date issues in relation to forensic analysis, studying clock skew and corroboration. Buchholz and Tjaden (2007) undertook a study in relation to the clocks of hosts connected to the Internet within the scope of forensic investigations. Marrington et al. (2011) produced prototype software which detects inconsistencies in computer activity timelines. As time and date issues can be crucial in a forensic examination, confirming the accuracy of data and the subsequent forming of subsequent conclusions is necessary to ensure accuracy, and acceptance of evidence in court (Boyd, 2004). Visualisation techniques are an important aspect to be considered, along with a comprehensive approach, without relying on one method or technique to address the volume challenge.

Beebe (2009) observes that forensic research has followed the digital forensic process: Prepare → Respond → Collect → Analyse → Present → Complete. The initial focus of research has been on response and collection (hence hardware write blockers and live response procedures have been developed), and the focus of research is now swinging towards analysis and presentation, with the analysis phase having the greater questions to be answered (Beebe, 2009).

Discussion

As outlined, there has been much discussion regarding the data volume challenge, and many calls for research into the application of data mining and other techniques to address the problem. Nevertheless, there has been very little published work in relation to a method or framework to apply data mining techniques, or other methods, to reduce or analyse the large volume of real-world data. In addition, the value of extracting or using intelligence from digital forensic data has had minimal discussion.

There are a variety of research fields which have potential to impact the volume data challenge, including Knowledge Discovery, Knowledge Management, Data Mining, and Criminal Intelligence Analysis. Knowledge Discovery and Knowledge Management are overall processes to extract valuable knowledge from data (Cios and Kurgan, 2005; and see Survey section). Data mining is a step of the knowledge discovery process and may offer a way to comprehend the large volume of seized data (Fayyad et al., 1996a). Whilst there have been many calls for research in relation to the large volume issue, which include research into whether data mining methodologies can be applied to digital forensic data (Beebe and Clark, 2005; Palmer, 2001), there is very little published research which progresses this. This is, perhaps, because digital forensic research is relatively new compared to other forensic science disciplines or the information security discipline. For example, *Digital Investigation*, the only journal dedicated to Digital Forensics with an impact factor, which ranks six (and is the top digital forensic publication) in Google Scholar's (general) Forensic Science category,[2] is only in its 11th year in 2014.

Significant gaps remain in relation to applying data mining methodology to digital forensic data, including a methodology which can be applied to real world data, the benefits which may be observed, and the most appropriate methodology to achieve the desired results including; a reduction in analysis time, a method of archiving and retrieving data, a rapid triage process, and a methodology to gain knowledge from seized data. Whilst data mining alone may not be an overall solution to the many issues raised, it will perhaps have best application in relation to intelligence and understanding of disparate case information to provide an overall increase in knowledge, and perhaps should be researched with that in mind.

It was observed that there is minimal published information relating to the average processing times of average evidence amounts per year, such as applying common forensic analysis tasks to common data. There is an opportunity to use a standard corpus of data and apply common processing techniques to this, and record the timeframe for processing, extrapolating this to hard drive sizes of the time. This would serve to highlight whether there is a growing gap in processing times, and when undertaken for a period of time, serve to show whether processing techniques are improving analysis timeframes.

There is great potential to develop a data reduction methodology which serves to collect a subset of important information for the purposes of triage, rapid analysis, data mining, intelligence analysis, and archiving, such as the Digital Forensic Data Reduction Framework – Fig. 5 (Quick and Choo, 2014). Data reduction techniques have perhaps the greatest opportunity to influence the various stages of forensic analysis, and also to provide benefits to other areas not generally discussed (i.e. intelligence, archiving, and knowledge of trends).

[2] http://scholar.google.com.au/citations?view_op=top_venues&hl=en&vq=soc_forensicscience; last accessed 26 May 2014.

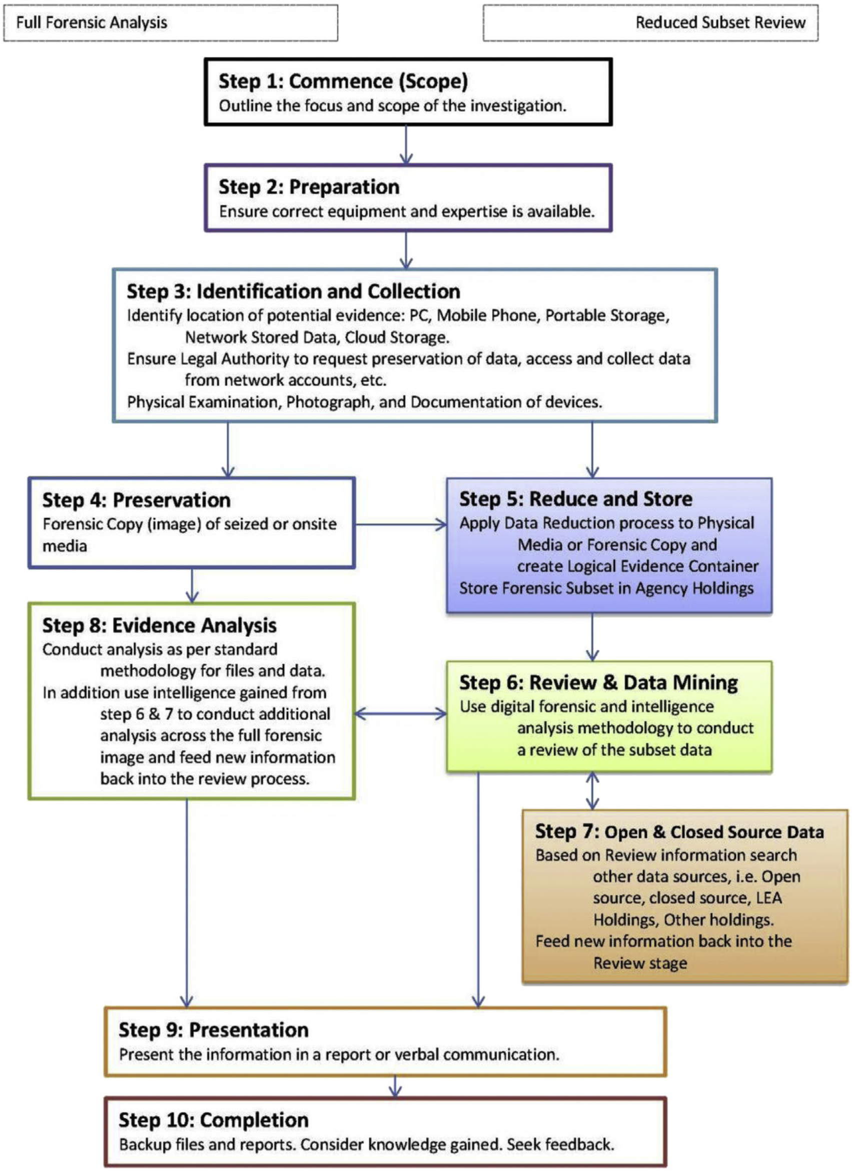
Full Forensic Analysis
Reduced Subset Review
Step 1: Commence (Scope)
Outline the focus and scope of the investigation.
Step 2: Preparation
Ensure correct equipment and expertise is available.
Step 3: Identification and Collection
Identify location of potential evidence: PC, Mobile Phone, Portable Storage, Network Stored Data, Cloud Storage.
Ensure Legal Authority to request preservation of data, access and collect data from network accounts, etc.
Physical Examination, Photograph, and Documentation of devices.
Step 4: Preservation
Forensic Copy (image) of seized or onsite media
Step 5: Reduce and Store
Apply Data Reduction process to Physical Media or Forensic Copy and create Logical Evidence Container
Store Forensic Subset in Agency Holdings
Step 8: Evidence Analysis
Conduct analysis as per standard methodology for files and data.
In addition use intelligence gained from step 6 & 7 to conduct additional analysis across the full forensic image and feed new information back into the review process.
Step 6: Review & Data Mining
Use digital forensic and intelligence analysis methodology to conduct a review of the subset data
Step 7: Open & Closed Source Data
Based on Review information search other data sources, i.e. Open source, closed source, LEA Holdings, Other holdings.
Feed new information back into the Review stage
Step 9: Presentation
Present the information in a report or verbal communication.
Step 10: Completion
Backup files and reports. Consider knowledge gained. Seek feedback.

Fig. 5. Digital forensic data reduction framework (Quick and Choo, 2014).

Another major gap relates to the use of intelligence gained during forensic analysis, which has potentially a large benefit to policing agencies, and yet remains unaddressed. The current focus is on attending to urgent jobs, with no time to consider other matters which may provide valuable input to current investigations. Historically, there has very little discussion of a methodology to utilise the intelligence gained from one investigation to assist with other investigations, nor to build a body of knowledge inherent in historical cases. In addition, the use of open and closed source intelligence for investigations could provide a manner of improving the time and value of analysis of digital forensic investigations. For example, information stored on a phone seized for one investigation may provide information to other seemingly unrelated investigations. Without an intelligence or knowledge management process, this information or linkage remains undiscovered. Using knowledge management, intelligence analysis, and data mining methodology, it is envisioned that a large volume of information could be aggregated into common data holdings, and include the capability for rapid searches to link associated information and assist in investigations.

Whilst many of the problems posed by the increasing volume of data are addressed in part by new developments in technology, another challenge is that the law is somewhat slower to address issues relating to digital forensic analysis (McKemmish, 1999). Hence, any new methodologies must abide to the legal environment to which they are being applied, as the pace of change in law makes it difficult to implement unique solutions to challenges. Dr Eugene

Table 4
Summary of digital forensic data volume papers and topics.

Year	Author	Data mining	Data reduction	DFaaS	Distributed parallel	Intelligence	Machine learning	Triage	Visualisation	Volume
1999	McKemmish									✓
2001	Palmer	✓	✓							✓
2002	Palmer									✓
2003	Culley		✓							✓
2004	Ferraro and Russell		✓							✓
2004	O'Connor							✓		
2004	Rogers									✓
2004	Roussev and Richard				✓					✓
2004	Shannon	✓								
2004	Sommer									✓
2004	Stevens								✓	
2005	Beebe and Clark	✓								✓
2005	Brown et al.	✓								✓
2005	Craiger et al.									✓
2005	Kenneally and Brown		✓							
2005	Nykodym et al.					✓				
2005	Sheldon						✓			✓
2005	Turner		✓							✓
2006	Abraham	✓				✓				
2006	Adelstein									✓
2006	Alink et al.		✓	✓		✓			✓	✓
2006	Garfinkel		✓							
2006	Ribaux et al.					✓				
2006	Richard and Roussev (2006a,b)		✓							✓
2006	Rogers et al.							✓		
2006	Schatz and Clark		✓						✓	
2006	Shaw					✓				
2006	Teelink and Erbacher								✓	
2006	Turner		✓							
2006	Weiser et al.					✓				
2007	Buchholz and Tjaden								✓	
2007	Khan et al.								✓	✓
2007	LaVelle and Konrad							✓		
2007	Marziale et al.				✓					
2007	Okolica et al.	✓								
2007	Reyes et al.							✓		✓
2007	Wiles et al.									✓
2008	Case et al.					✓				✓
2008	Iqbal et al.	✓								
2008	Lee et al.				✓					
2008	Peisert et al.									✓
2008	Pringle and Sutherland				✓					✓
2008	Riley et al.									✓
2009	Ayers				✓					✓
2009	Beebe	✓	✓				✓			✓
2009	Biggs and Vidalis									✓
2009	Casey et al.							✓		✓
2009	Garfinkel et al.							✓		✓
2009	Greiner		✓							
2009	Hoelz et al.						✓			
2009	Nance et al.				✓					✓
2009	Olsson and Boldt								✓	
2009	Parsonage							✓		
2009	Raghavan et al.					✓				
2009	Turnbull et al.									✓
2010	Casey									✓
2010	Garfinkel					✓		✓		✓
2010	Gogolin									✓
2010	Huang et al.	✓								
2010	Iqbal et al.	✓								

Table 4 (*continued*)

Year	Author	Data mining	Data reduction	DFaaS	Distributed parallel	Intelligence	Machine learning	Triage	Visualisation	Volume
2010	Iqbal et al.	✓								
2010	Ribaux et al.					✓				
2011	Marrington								✓	
2012	Al-Zaidy et al.					✓				
2012	Bhoedjang et al.		✓	✓						✓
2012	Garfinkel, 2012a		✓							✓
2012	Garfinkel, 2012b	✓			✓					
2012	Jones et al.		✓							✓
2012	Roussev and Quates							✓		
2013	Alzaabi et al.		✓							✓
2013	Casey et al.		✓							✓
2013	Koopmans and James							✓		
2013	Marturana and Tacconi							✓		
2013	Overill et al.							✓		✓
2013	Pollitt		✓					✓		
2013	Quick and Choo									✓
2013	Raghavan		✓						✓	✓
2013	Roussev et al.		✓							✓
2013	Shaw and Browne							✓		✓
2013	Shiaeles et al.							✓		
2013	Zimmerman									✓
2014	Breitinger and Roussev		✓							✓
2014	Breitinger et al.		✓							✓
2014	Casey	✓								
2014	Noel and Peterson	✓								✓
2014	Pringle and Burgess				✓					
2014	Quick and Choo	✓	✓			✓		✓		✓
2014	Quick et al.									✓
2014	van Baar et al.			✓						✓
2014	Vidas et al.							✓		✓

Spafford (as cited in Palmer, 2001) states that a comprehensive approach to academic research is required to address the range of challenges, not just the technical ones, and include social, legal, and procedural. Procedural guidelines and practices which focus on collecting every piece of data in an investigation can lead to a requirement to thoroughly examine large volumes of data. Social issues relate to the storage of data for long periods of time, which use resources that could be used to address other issues as opposed to storage concerns. Legal issues relate to the ability of law to keep pace with the changes in technology.

There is an opportunity to develop a framework for digital forensic practitioners to apply data mining and intelligence analysis techniques to digital forensic data to reduce the time to collect and gain an understanding of the data seized. Future research opportunities exist in relation to determining appropriate data reduction and data mining techniques to apply to the volume of structured and unstructured data atypical of seized evidential data. There is a need for digital forensic data mining and intelligence analysis techniques to be developed and tested against test data and real world (anonymised) data to determine an appropriate methodology or methodologies which can be applied in real world situations in an effort to address the digital forensic volume data issue.

Table 4 summarises the papers discussed in Survey section, highlighting the general topic or topics covered in each paper from 1999 to 2014.

The process of triage has been widely discussed, including special publications devoted to the topic, and many methods have been proposed. A gap in relation to research into the application of a methodology, with supportive figures would serve to highlight to impact of triage methods, and which are successful and applicable to an organisation.

There is much discussion about ‘growing backlogs’ of cases awaiting analysis, yet no real-world figures are listed to enable accurate understanding of the problem. It would be appropriate for actual figures to be available, although it is understood that many (if not all) agencies would not allow this information to be released for public discussion. The information in the FBI RCFL Annual Reports is of great value to enable an understanding of the size of data and number of cases each year over a period of time. Further information relating to device volume, number of devices, timeframes, backlogs, number of cases with data in cloud storage, imaging times, processing times, analysis times, review times, report volumes, and successful presentations in a legal context, would all be relevant to enable researchers to understand where to focus their work.

Whilst many concepts are put forward to improve triage, imaging, processing, and analysis, there is little discussion about scaling these to the volume of current case requirements. Some methods are applied to very small data volumes, and when extrapolated to current case sizes would result in many hours or days of processing required.

It would be appropriate to scale these methods to apply to real-world data volumes, and also make greater use of real-world data corpus (such as that of Garfinkel, 2012a,b). It would also be beneficial to undertake common imaging and processing tasks using a range of common forensic tools across a common forensic corpus to compare processing times over a period of years, to determine if there have been improvements in technology and tools.

There are also research gaps in relation to the effectiveness of the various approaches in relation to which are applicable in practicable terms. Future research could examine the effectiveness of the various proposed solutions in terms of evidence identification in a large corpus, and in terms of the admissibility of the data into court. For example, a review of the acceptance of triaged evidence or evidence derived from data reduction in a legal environment, and include a review of any concerns raised in relation to whether a data reduction process is potentially missing exculpatory evidence.

The legal issues raised by Spafford (as cited in Palmer, 2001) relate to ensuring a compliance with law and legal procedures, as there is little point in undertaking research and deploying advanced technology if it doesn't comply with the law and rules in the environment which it is deployed. RFC 3227 lists the legal considerations in relation to computer evidence, including the need to ensure it is; Admissible, Authentic, Complete, Reliable, and Believable (Brezinski and Killalea, 2002). Where 'Complete' is listed, it is in relation to telling 'the whole story, and not just a particular perspective', and does not necessitate a need to image everything. As long as legal issues are considered when determining new methods to copy, store, and process data, it will be far easier to gain acceptance from forensic practitioners within a legal environment. Future efforts of research should also abide with legal, social, and procedural guidelines, such that legal considerations are compliant, timeliness of processes ensures social acceptance, and procedural issues in relation to adoption of techniques are discussed.

Conclusion

Our literature survey identified that research gaps still remain in relation to the digital forensic data volume challenge. For example, there remains a need for research to be undertaken into data reduction techniques, data mining, intelligence analysis, and the use of open and closed source information. This should include research into the application of processes in a real world environment, and its acceptance in Courts and other tribunals.

Data Mining offers a potential solution to understanding the increasing volume of data, but may be more suited as an intelligence and knowledge tool, rather than an evidence focussed tool. Understanding the intelligence value of digital forensic data is a gap, and making use of this knowledge from the volume of data is an untapped resource which agencies should consider. Merging disparate data into a common body of knowledge may provide linkages which are currently unknown. A survey of current processing and analysis tools using a standard corpus would serve to provide an understanding of the current time for these tasks, and ongoing comparisons would serve to measure the progress of forensic tools.

Triage processes are well discussed, and perhaps seen as a solution to the growing volume of devices and data. Further research would be appropriate to assess the influence of various triage processes on real world devices and data to determine the most applicable methodology to deploy, which also provides for future needs, and a review of the acceptance of triaged evidence in a legal environment, including whether the process is potentially missing exculpatory evidence. A review of backlog of cases within agencies is also needed to quantify the actual extent of the problem, as this information is currently anecdotal, confidentiality issues notwithstanding.

A data reduction process holds potential to influence a range of digital forensic stages; such as collection, processing and analysis, and provide for intelligence, knowledge, and future needs. Any methodology developed should follow a comprehensive framework to address; data reduction, review, analysis, data mining, intelligence analysis, data storage, archiving, and retrieval. Building on our previous work (Quick and Choo, 2014), we are researching the appropriate files and data types to include in a Data Reduction process to explore the effectiveness of this approach.

Acknowledgements

The views and opinions expressed in this article are those of the authors alone and not the organisations with whom the authors are or have been associated or supported.

References

Abraham T. Event sequence mining to develop profiles for computer forensic investigation purposes. In: ACSW frontiers '06: proceedings of the 2006 Australasian workshops on grid computing and e-research; 2006. p. 145–53.
ACC. In: aCC, editor. Board of the Australian Crime Commission, chair annual report 2012–13. Commonwealth of Australia; 2013.
AccessDataCorporation. Divide & conquer: overcoming computer forensic backlog through distributed processing and division of labor white paper. 2010.
ACPO. Good practice guidelines for computer based evidence v4.0. Association of Chief Police Officers; 2006. www.7safe.com/electronic_evidence [viewed 05.03.14].
Adelstein F. Live forensics: diagnosing your system without killing it first. Commun ACM 2006;49:63–6.
Al-Zaidy R, Fung BCM, Youssef AM, Fortin F. Mining criminal networks from unstructured text documents. Digit Investig 2012;8:147–60.
Alink W, Bhoedjang RAF, Boncz PA, de Vries AP. XIRAF – XML-based indexing and querying for digital forensics. Digit Investig 2006;3(Suppl.):50–8.
Alzaabi M, Jones A, Martin TA. An ontology-based forensic analysis tool. J Digit Forensics, Secur Law 2013;(2013 Conference Suppl.):121–35.
Ayers D. A second generation computer forensic analysis system. Digit Investig 2009;6:S34–42.
Beebe N. Digital forensic research: the good, the bad and the unaddressed. Springer; 2009. p. 17–36.
Beebe N, Clark J. Dealing with terabyte data sets in digital investigations. Adv Digit Forensics 2005:3–16.
Bell L. Seagate launches 4TB hard disk engineered for video content. The Inquirer 2013. http://www.theinquirer.net/inquirer/news/2269518/seagate-launches-4tb-hard-disk-engineered-for-video-content [viewed 22 June].
Bhoedjang RAF, van Ballegooij AR, van Beek HMA, van Schie JC, Dillema FW, van Baar RB, et al. Engineering an online computer forensic service. Digit Investig 2012;9:96–108.

Biggs S, Vidalis S. Cloud computing: the impact on digital forensic investigations. In: IEEE international conference for internet technology and secured transactions (ICITST 2009). IEEE; 2009. p. 1–6.

Boyd C. Time and date issues in forensic computing – a case study. Digit Investig 2004;1:18–23.

Breitinger F, Baier H, White D. On the database lookup problem of approximate matching. Digit Investig 2014;11(Suppl. 1):S1–9.

Breitinger F, Roussev V. Automated evaluation of approximate matching algorithms on real data. Digit Investig 2014;11(Suppl. 1):S10–7.

Brezinski D, Killalea TRFC. 3227-Guidelines for evidence collection and archiving. 2002.

Brown R, Pham B, de Vel O. Design of a digital forensics image mining system. Knowl-Based Intell Inf Eng Syst 2005:395–404.

Buchholz F, Tjaden B. A brief study of time. Digit Investig 2007;4:31–42.

Carvey H. Windows Registry forensics: advanced digital forensic analysis of the Windows Registry. Elsevier; 2011.

Case A, Cristina A, Marziale L, Richard GG, Roussev VFACE. Automated digital evidence discovery and correlation. Digit Investig 2008; 5(Suppl.):S65–75.

Casey E. "Dawn raids" bring a new form in incident response. Digit Investig 2009;5:73–4.

Casey E. Digital dust: evidence in every nook and cranny. Digit Investig 2010;6:93–4.

Casey E. Growing societal impact of digital forensics and incident response. Digit Investig 2014;11:1–2.

Casey E, Ferraro M, Nguyen L. Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence. J Forensic Sci 2009;54:1353–64.

Casey E, Katz G, Lewthwaite J. Honing digital forensic processes. Digit Investig 2013;10:138–47.

Cios K, Kurgan L. Trends in data mining and knowledge discovery. Adv Tech Knowl Discov Data Min 2005:1–26.

Coughlin T. High density hard disk drive trends in the USA. J Magn Soc Jpn 2001;25:111–20.

Craiger J, Pollitt M, Swauger J. Law enforcement and digital evidence. Handbook of information security, vol. 2; 2005. p. 739–77.

Culley A. Computer forensics: past, present and future. Inf Secur Tech Rep 2003;8:32–6.

DFI_News. Guidance software announces EnCase forensic version 7. 2011. http://www.dfinews.com/product-releases/2011/06/guidance-software-announces-encase-forensic-version-7#.UuR2hLRe6Uk [viewed 26 January].

Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996a;17:37.

Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996b;39:27–34.

FBI_RCFL. In: RCFL, editor. FBI regional computer forensic laboratory annual reports 2003–2012. Quantico: FBI; 2003–2012.

Ferraro MM, Russell A. Current issues confronting well-established computer-assisted child exploitation and computer crime task forces. Digit Investig 2004;1:7–15.

Garfinkel S. Forensic feature extraction and cross-drive analysis. Digit Investig 2006;3:71–81.

Garfinkel S. Digital forensics research: the next 10 years. Digit Investig 2010;7(Suppl.):S64–73.

Garfinkel S. Digital forensics XML and the DFXML toolset. Digit Investig 2012a;8:161–74.

Garfinkel S. Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus. Digit Investig 2012b;9(Suppl.):S80–9.

Garfinkel S, Farrell, Roussev, Dinolt. Bringing science to digital forensics with standardized forensic corpora. In: DFRWS 2009; 2009. Montreal, Canada, http://digitalcorpora.org/corpora/disk-images [viewed 9 September].

Gogolin G. The digital crime tsunami. Digit Investig 2010;7:3–8.

Greiner L. Sniper forensics. netWorker 2009;13:8–10.

Growchowski E. Emerging trends in data storage on magnetic hard disk drives. Datatech. ICG Publishing; September 1998. p. 11–6.

Hand DJ, Mannila H, Smyth P. Principles of data mining. MIT Press; 2001.

Hearst MA. Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on computational linguistics. Association for Computational Linguistics; 1999. p. 3–10.

Hoelz B, Ralha C, Geeverghese R. Artificial intelligence applied to computer forensics. In: SAC '09: proceedings of the 2009 ACM symposium on applied computing. Honolulu, Hawaii: ACM; 2009. p. 883–8.

Huang J, Yasinsac A, Hayes PJ. Knowledge sharing and reuse in digital forensics. In: Systematic approaches to digital forensic engineering (SADFE), 2010 fifth IEEE international workshop on. IEEE; 2010. p. 73–8.

INTERPOL. In: Proceedings of the 14th INTERPOL forensic science symposium; 2004.

Iqbal F, Binsalleeh H, Fung B, Debbabi M. Mining writeprints from anonymous e-mails for forensic investigation. Digit Investig 2010;7:56–64.

Iqbal F, Hadjidj R, Fung BCM, Debbabi M. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digit Investig 2008;5(Suppl.):S42–51.

Jones B, Pleno S, Wilkinson M. The use of random sampling in investigations involving child abuse material. Digit Investig 2012; 9(Suppl.):S99–107.

Kenneally E, Brown C. Risk sensitive digital evidence collection. Digit Investig 2005;2:101–19.

Khan M, Chatwin C, Young R. A framework for post-event timeline reconstruction using neural networks. Digit Investig 2007;4:146–57.

Kohavi R, Rothleder N, Simoudis E. Emerging trends in business analytics. Commun ACM 2002;45:45–8.

Koopmans MB, James JI. Automated network triage. Digit Investig 2013; 10:129–37.

LaVelle C, Konrad A. FriendlyRoboCopy: a GUI to RoboCopy for computer forensic investigators. Digit Investig 2007;4:16–23.

Lee J, Un S, Hong D. High-speed search using Tarari content processor in digital forensics. Digit Investig 2008;5:S91–5.

Lee W, Stolfo SJ. Data mining approaches for intrusion detection. Defense Technical Information Center; 2000.

LSI, LSI Tarari Content Processor Family Enhanced with High-Performance, *Low-Latency Solution*, http://www.lsi.com/about/newsroom/Pages/20100426apr.aspx, viewed 26.01.11.

Marrington A, Baggili I, Mohay G, Clark A. CAT detect (computer activity timeline detection): a tool for detecting inconsistency in computer activity timelines. Digit Investig 2011;8(Suppl.):S52–61.

Marturana F, Tacconi S. A machine learning-based triage methodology for automated categorization of digital media. Digit Investig 2013;10:193–204.

Marziale L, Richard G, Roussev V. Massive threading: using GPUs to increase the performance of digital forensics tools. Digit Investig 2007;4:73–81.

McKemmish R. What is forensic computing? Trends and issues in crime and criminal justice. Australian Institute of Criminology; 1999. p. 1–6.

Mee V, Tryfonas T, Sutherland I. The Windows Registry as a forensic artefact: illustrating evidence collection for internet usage. Digit Investig 2006;3:166–73.

Nance K, Hay B, Bishop M. Digital forensics: defining a research agenda. In: System sciences, 2009 HICSS'09 42nd Hawaii international conference on. IEEE; 2009. p. 1–6.

NIJ. Forensic examination of digital evidence: a guide for law enforcement. 2004. http://nij.gov/nij/pubs-sum/199408.htm.

Noel GE, Peterson GL. Applicability of latent Dirichlet allocation to multi-disk search. Digit Investig 2014;11:43–56.

Nykodym N, Taylor R, Vilela J. Criminal profiling and insider cyber crime. Digit Investig 2005;2:261–7.

O'Connor O. Deploying forensic tools via PXE. Digit Investig 2004;1: 173–6.

Okolica JS, Peterson GL, Mills RF. Using author topic to detect insider threats from email traffic. Digit Investig 2007;4:158–64.

Olsson J, Boldt M. Computer forensic timeline visualization tool. Digit Investig 2009;6:S78–87.

Overill RE, Silomon JAM, Roscoe KA. Triage template pipelines in digital forensic investigations. Digit Investig 2013;10:168–74.

Palmer GA. Road map for digital forensic research. Report from the First Digital Forensic Research Workshop (DFRWS). 2001.

Palmer G. Forensic analysis in the digital world. Int J Digit Evid 2002;1:1–6.

Parsonage H. Computer forensics case assessment and triage – some ideas for discussion. http://computerforensics.parsonage.co.uk/triage/triage.htm; 2009 [viewed 4 August].

Peisert S, Bishop M, Marzullo K. Computer forensics *in forensis*. SIGOPS Oper Syst Rev 2008;42:112–22.

Pollitt MM. Triage: a practical solution or admission of failure. Digit Investig 2013;10:87–8.

Pringle N, Burgess M. Information assurance in a distributed forensic cluster. Digit Investig 2014;11(Suppl. 1):S36–44.

Pringle N, Sutherland I. Is a computational grid a suitable platform for high performance digital forensics?. In: Proceedings of the 7th European conference on information warfare and security. Academic Conferences Limited; 2008. p. 175.

Quick D, Choo K-KR. Dropbox analysis: data remnants on user machines. Digit Investig 2013;10:3–18.

Quick D, Choo K-KR. Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive. Trends Issues Crime Crim Justice September 17, 2014;480:1–11.

Quick D, Martini B, Choo K-KR. Cloud storage forensics. Syngress: An Imprint of Elsevier; 2014.

Raghavan S. Digital forensic research: current state of the art. CSI Trans ICT 2013;1:91–114.

Raghavan S, Clark A, Mohay G. FIA: an open forensic integration architecture for composing digital evidence. In: Forensics in telecommunications, information and multimedia. Springer; 2009. p. 83–94.

Ratcliffe J. Integrated intelligence and crime analysis: enhanced information management for law enforcement leaders. 2nd ed. Washington, DC: Police Foundation; 2007.

Reyes A, Oshea K, Steele J, Hansen J, Jean B, Ralph T. Digital forensics and analyzing data. Cyber crime investigations. Elsevier; 2007. p. 219–59.

Ribaux O, Baylon A, Roux C, Delémont O, Lock E, Zingg C, et al. Intelligence-led crime scene processing. Part I: forensic intelligence. Forensic Sci Int 2010;195:10–6.

Ribaux O, Walsh S, Margot P. The contribution of forensic science to crime analysis and investigation: forensic intelligence. Forensic Sci Int 2006;156:171–81.

Richard G, Roussev V. Digital forensics tools: the next generation. Digital crime and forensic science in cyberspace. 2006. p. 75.

Richard G, Roussev V. Next-generation digital forensics. Commun ACM 2006b;49:76–80.

Riley JW, Dampier DA, Vaughn R. A comparison of forensic hard drive imagers: a time analysis comparison between the ICS image MASSter-Solo III and the Logicube Talon. J Digit Forensic Pract 2008;2:74–82.

Rogers MK. The future of computer forensics: a needs analysis survey. Comput Secur 2004;23:12–6.

Rogers MK, Goldman J, Mislan R, Wedge T, Debrota S. Computer forensics field triage process model. J Digit Forensics, Secur Law 2006;1:19–37.

Roussev V, Quates C. Content triage with similarity digests: the M57 case study. Digit Investig 2012;9(Suppl.):S60–8.

Roussev V, Quates C, Martell R. Real-time digital forensics and triage. Digit Investig 2013;10:158–67.

Roussev V, Richard G. Breaking the performance wall: the case for distributed digital forensics. In: Proceedings of the 2004 digital forensics research workshop; 2004.

Schatz B, Clark AJ. An open architecture for digital evidence integration. In: AusCERT Asia Pacific information technology security conference; 2006.

Schatz B, Mohay G, Clark A. A correlation method for establishing provenance of timestamps in digital evidence. Digit Investig 2006;3(Suppl.):98–107.

Shannon M. Forensic relative strength scoring: ASCII and entropy scoring. Int J Digit Evid 2004;2:151–69.

Shaw A, Browne A. A practical and robust approach to coping with large volumes of data submitted for digital forensic examination. Digit Investig 2013;10:116–28.

Shaw E. The role of behavioral research and profiling in malicious cyber insider investigations. Digit Investig 2006;3:20–31.

Sheldon A. The future of forensic computing. Digit Investig 2005;2:31–5.

Shiaeles S, Chryssanthou A, Katos V. On-scene triage open source forensic tool chests: are they effective? Digit Investig 2013;10:99–115.

Sommer P. The challenges of large computer evidence cases. Digit Investig 2004;1:16–7.

Stevens MW. Unification of relative time frames for digital forensics. Digit Investig 2004;1:225–39.

Teelink S, Erbacher R. Improving the computer forensic analysis process through visualization. Commun ACM 2006;49:71–5.

Turnbull B, Taylor R, Blundell B. The anatomy of electronic evidence; quantitative analysis of police e-crime data. In: ARES '09 international conference on availability, reliability and security; 2009. p. 143–9.

Turner P. Unification of digital evidence from disparate sources (digital evidence bags). Digit Investig 2005;2:223–8.

Turner P. Selective and intelligent imaging using digital evidence bags. Digit Investig 2006;3(Suppl.):59–64.

UNODC. United Nations Office on Drugs and Crime – criminal intelligence manual for analysts. Vienna, Austria: United Nations; 2011. New York.

van Baar RB, van Beek HMA, van Eijk EJ. Digital forensics as a service: a game changer. Digit Investig 2014;11(Suppl. 1):S54–62.

Vidas T, Kaplan B, Geiger M. OpenLV: empowering investigators and first-responders in the digital forensics process. Digit Investig 2014;11(Suppl. 1):S45–53.

Walmart. Western Digital Green 4TB desktop internal hard-drive. http://www.walmart.com/ip/WD-Green-4TB-Desktop-Internal-Hard-Drive/30579528; 2014 [viewed 21 June].

Walter C. Kryder's law. Sci Am 2005;293(2):32–3.

Weiser M, Biros DP, Mosier G. In: Dardick Glenn S, editor. Development of a national repository of digital forensic intelligence. USA: Longwood University Virginia; 2006. p. 5.

Wiles J, Alexander T, Ashlock S, Ballou S, Depew L, Dominguez G, et al. Forensic examination in a terabyte world. Techno security's guide to e-discovery and digital forensics. Elsevier; 2007. p. 129–46.

Wong A. Explosion of data envelops man in the street. The Australian 2010. Australia.

Zimmerman E. Imaging test results. https://docs.google.com/spreadsheet/lv?key=0Al7os14ND-cFdGp1NDR2WGwyakR2TkJtNUFXa29pNXc&type=view&gid=0&f=true&sortcolid=11&sortasc=true&rowsperpage=250; 2013 [viewed 18 June].