Digital Forensics, AKA eDiscovery, in a Cyber Security Context

by Thibault Dambrine & Laura Smith

Introduction

When a cyber breach occurs, it is almost always accompanied by data exfiltration and threats to disseminate or sell stolen data on the "dark web". Among all data, the most immediately valuable to hackers is (but not limited to) Personally Identifiable Information (PII), and Personal Health Information (PHI).

Companies and organizations who hold confidential customer data are entrusted to keep it safe. When that trust is broken for any reason, even if by no specific or intentional fault of their own, those same companies and organizations may be the target of lawsuits, based on Privacy Laws. Figuring out what information is at risk in such circumstances is where digital forensics, often described as "eDiscovery" crosses the path of cyber security incident remediation.

In this article, we will describe:

- eDiscovery, origins, evolution and disambiguation for the use of the term
- Privacy Laws
- How digital forensics, also known as "eDiscovery", are linked to Cyber Security Breach Resolution
- Perspectives on the unique challenges at the intersection of eDiscovery and Cyber Breach Resolution

eDiscovery Overview

"Discovery" is the root of "eDiscovery". In legal context, "Discovery" is the term used for the initial phase of litigation. This is where parties in a legal dispute must provide relevant information and records, along with other evidence related to the case. The key difference between "Discovery" and "eDiscovery" is that eDiscovery is related specifically to electronic document formats. Discovery on the other hand, encompasses both data stored on paper format and data stored in digital format.

Over time, the market for legal eDiscovery software tools has grown to become a multi-billion-dollar industry. With the emergence of cybercrime, eDiscovery software companies traditionally operating in the legal market have expanded into the post-cyberbreach territory. With this, the usage of the word "eDiscovery" has grown to describe both legal and port-breach processes. For disambiguation, lets first differentiate the two versions of "eDiscovery".

1) "eDiscovery" In the original, legal context involves a team of licensed attorneys combing through electronic documents for specific elements. They will then use their legal knowledge, experience and applicable framework to decide whether or not specific documents or elements harvested may be relevant and useful to the case.

2) "eDiscovery" In a post-cyberbreach, digital forensics context is a specialized form of eDiscovery. Its aim is to find specific elements of PII data contained within data sets that may have been compromised, copied, or downloaded during a cyber breach, for subsequent illegal distribution. In this scenario, key functions are:
    - Understanding what type of personal identifiable information (PII) data may have been contained in a compromised data set
    - Linking each PII element to its respective owner (defined as the affected person or company)

    This type of "eDiscovery work is typically performed by individuals who have data breach investigation experience.
    Other names used to refer to this specific sub-discipline are "Cyber Discovery", Post Breach eDiscovery" and/or Post Breach Data Mining.

The two "eDiscovery" types, legal and post-cyberbreach, are similar in process. Both involve identifying, collecting and preserving data. The key differences between the two is the purpose of the exercise, and who is analyzing the data. Going forward, this article will focus on the post data breach, "Digital Forensics", version of "eDiscovery".

When thinking of compromised data in context of post-cyberbreach eDiscovery, sources can be almost unlimited, but a short list will typically include at least some of the following:

On-premises or Cloud-based computing and storage devices	Social Media	Third parties who may have relevant documents
○ Back-up systems Archival systems ○ Email servers, including Webmail accounts e.g., Gmail, Hotmail etc. ○ Network Servers ○ Document management systems ○ Mobile phones, tablets, and other handheld devices Social Insurance numbers (note that these appear also in the "Financial Data" section)	○ Chat accounts ○ Social media accounts	○ Agents, consultants, or advisers ○ Suppliers

eDiscovery is a mature discipline. It is governed by a standardized framework established in 2005 by George Socha and Tom Gelbmann, to address the [then] lack of standards in the eDiscovery market. The Electronic Discovery Reference Model (EDRM) breaks down the eDiscovery process in 9 stages:

1) Information Governance – Getting your electronic house in order to mitigate risk & expenses, should e-discovery become an issue, from initial creation of ESI (electronically stored information) through its final disposition.
2) Identification – Locating potential sources of ESI & determining its scope, breadth & depth.
3) Preservation – Ensuring that ESI is protected against inappropriate alteration or destruction.
4) Collection – Gathering ESI for further use in the eDiscovery process (processing, review, etc.).
5) Processing – Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review & analysis.
6) Review – Evaluating ESI for relevance & privilege.
7) Analysis – Evaluating ESI for content & context, including key patterns, topics, people & discussion.
8) Production – Delivering ESI to others in appropriate forms & using appropriate delivery mechanisms.
9) Presentation – Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native & near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

Two points on the structure above:

a) Each of these stages is a building block in a methodology designed to yield solid, un-altered electronic evidence. In the case of a post-cyberbreach eDiscovery exercise, the first step in the process would be number (2) in the list above: "Identification".

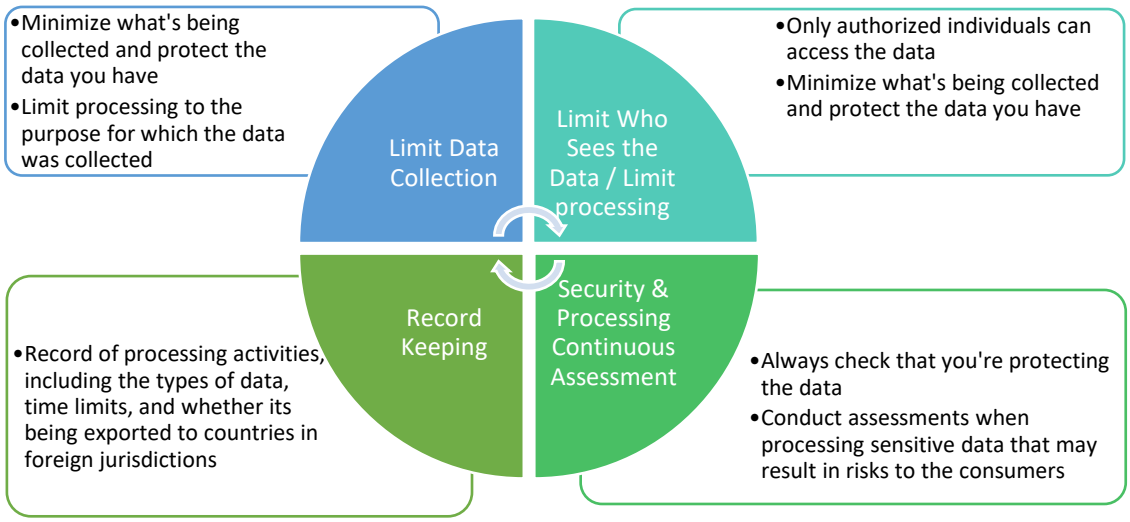
b) Point number (1) in the list above, "Information Governance", is not part of the process in a post-cyberbreach eDiscovery assignment. It is however an often overlooked and less prioritized task. Within Information Governance, the task of "purging obsolete data" is one of those that seldom gets all the attention it deserves.
Many organizations only discover too late (e.g., in post cyberbreach eDiscovery exercises), the value purging obsolete PII data. Like un-exploded ordinance, un-purged private data, from former employees for example, may create unforeseen liabilities if those elements are hacked or stolen.

Privacy Laws

According to Wikipedia, Privacy law is "*the body of law that deals with the regulating, storing, and using of personally identifiable information, personal healthcare information, and financial information of individuals, which can be collected by governments, public or private organisations, or other individuals. It also applies in the commercial sector to things like trade secrets and the liability that directors, officers, and employees have when handing sensitive information.*"
Examples of such laws are:

- Canada: Personal Information Protection and Electronic Documents Act (PIPEDA)
- European Union: General Data Protection Regulation (GDPR)
- UK: The "Data Protection, Privacy and Electronic Communications", the UK post-Brexit GDPR-equivalent legislation
- U.S. : Family Educational Rights and Privacy Act (FERPA)
- U.S. : The Health Insurance Portability and Accountability Act (HIPPA)

These laws are meant to ensure organizations and companies actively protect individual PII, PHI data against misuse, including ones caused by cyber breaches.



Privacy Law Principles

eDiscovery, in a post-cyberbreach context

After a cyber breach, one of the key steps in remediation is to understand if data was stolen. If this was the case, was the stolen data subject to one or more of the privacy laws described above. Under privacy laws, companies and organizations may face significant legal risks and/or fines if they fail to notify the affected parties within prescribed delays.

Enabling damage control with a quantified understanding of the extent of the breach is a key driver to limit these risks.
The eDiscovery deliverable in this context, must be as accurate as possible, identifying relevant PII elements involved in the data breach as fast as possible. The eDiscovery report will enable:

- Notification to outside stakeholders (whose data may have been stolen)
- Compliance with applicable privacy laws
- Initiation of damage control to reduce reputation impact as soon as possible
- Readiness with accurate and well researched facts, in case of possible class action lawsuits

Key eDiscovery Challenges

Data thefts are particularly pernicious, as they can be used to put customers as well as employee identities and financial situations at risk. Most corporate systems have no shortage of sensitive and valuable data to take, resell or exploit. Each of which carry their own sets of risks. Examples of those PII identifiers are:

Civic identifiers	**Financial identifiers**	**Medical Data**
Passport numbers	Bank account numbers	Medicare numbers
Citizenship Numbers	Swift Bank Codes	Treating physician
Driver's license numbers	Credit Card numbers	Treating facility
Physical address and telephone numbers	Social Security numbers	Medications
Social Insurance numbers (note that these appear also in the "Financial Data" section)		Diagnosis

In post-breach eDiscovery investigations, over and above the complexities of network considerations and data storage, there is complexity induced by the file systems themselves. When sifting through large amounts of data, in search of specific elements, one may encounter a variety of file formats. For example,

- Application data, which may be stored in database systems
- Individual documents, which may be found in
  - Searchable, such as MS Word or PDF form documents
  - Non-searchable, such as a picture of a passport attached to an email message
  - Buried and compressed with a mass of other documents, grouped inside a Zipped, compressed archive
  - Attachments to email messages stored in mail systems

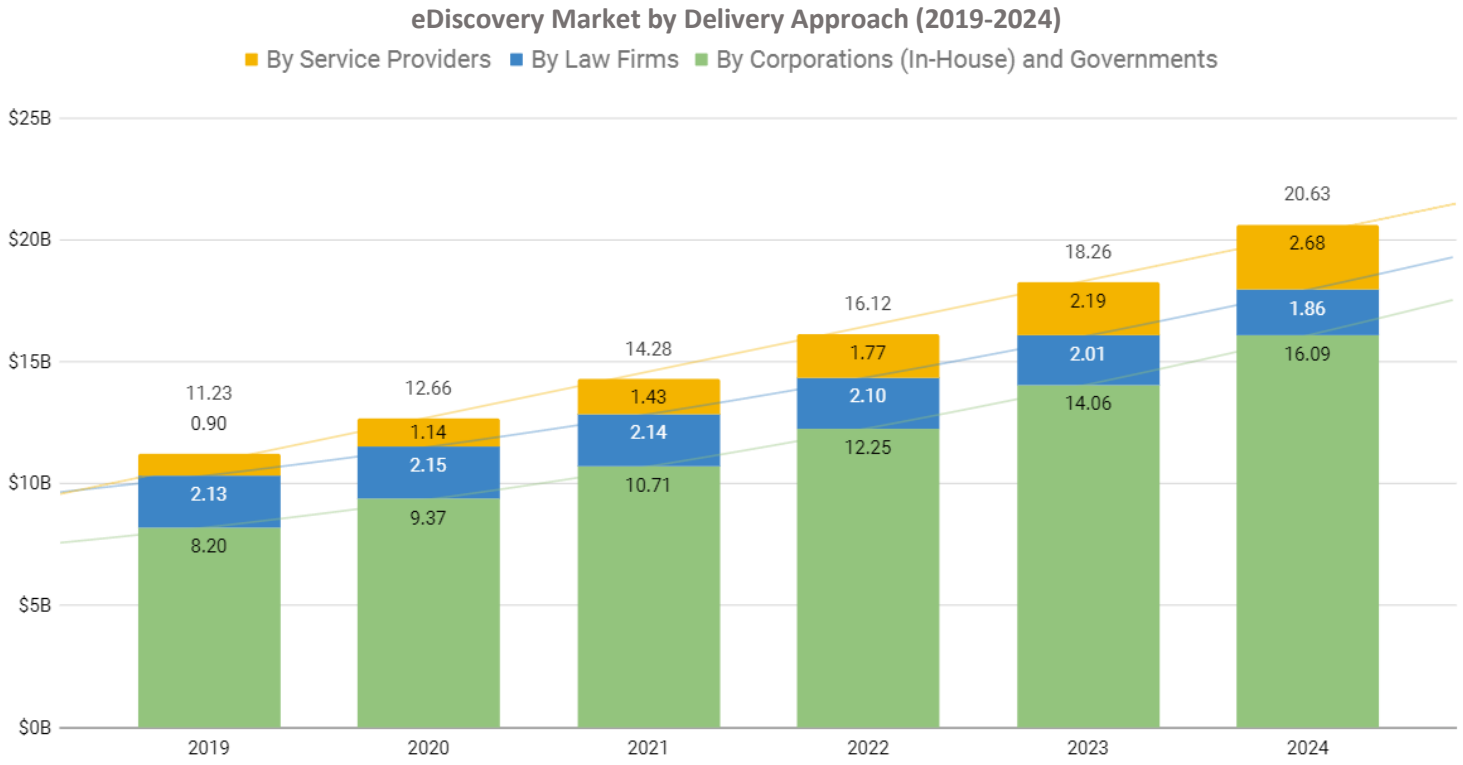
eDiscovery Efficiency: Points of Reference

In a data breach situation, anxiety levels are high, as are the risks to the organization. Budgets are limited and time is short. Every dollar, every minute counts. In this context, eDiscovery functions are typically most effectively performed by experienced professionals with software tools, understanding of the best options and established procedures.

Finding elements of personally identifiable information within large amounts of files stored in diverse formats takes method and effort. While human review is the gold standard for precision, it is comparatively slow and expensive. Scalability is often constricted by the availability and the cost of experienced reviewers.

When deciding on an eDiscovery vendor, organizations and companies have choices. Fortune Business Insights pegs the value of the 2021 cyber security market at USD 155.83 billion. This is big enough to make room for fragmentation and specialization. With this, eDiscovery companies have grown to focus specifically on individual components and sub-specialties. Post-Breach, digital forensics eDiscovery is one of those. To offer more competitive prices for specific document review, some eDiscovery companies differentiate themselves is by offering lower human review cost. This is done by exporting human document review tasks to countries in foreign jurisdictions where staff are paid at a lower rate. The trade-off in this scenario, is that the data to be reviewed will be seen out of country and many times, out of continent.

In an eDiscovery exercise, when scanning large amounts of documents in search of sensitive data, every attempt will typically be made to reduce the time and cost. Software, more than humans, can efficiently eliminate documents that are obviously not containing any trace of sensitive information. Once that is done, the remaining documents will go through a second level of scrutiny and if necessary, human reviews will be done, but on a smaller scale.



Total Worldwide Market for eDiscovery by Delivery Approach - Estimated 12.93% Compound Annual Growth Rate

Source: https://complexdiscovery.com/an-ediscovery-market-size-mashup-2019-2024-worldwide-software-and-services-overview/

Challenges

Post-cyberbreach eDiscovery challenges vary with each case. Significant parameters which may influence the methods, efforts and costs are:

- The amount of data to process, the type of storage and the variety of file formats, e.g., searchable, graphic, archive files, as well as the availability of accompanying metadata will all influence the amount of processing, effort, and cost. Typically, the first cut aims to use programs to rapidly eliminate documents which would contain no data of interest and thus reduce the cost of more deeply scrutinizing the remaining documents.
- Understanding the risks and potential liability associated with the compromised and/or stolen data. Privacy law regulations tend to be implemented for specific data types. Understanding what the applicable penalties are may be a decision factor for eDiscovery budgets. Sometimes however, the type of data stolen is only known after the eDiscovery process is complete.

- The value of the data will drive the need for precision. Higher value will increase the need for accuracy, the use of human review and cost
- On Day-One of a cyber breach, time is short and an "eDiscovery budget" may not be readily available.

Conclusions

Companies and organizations store customer and/or employee data as part of running their businesses. Based on Privacy Laws, they have custodial obligations to their stakeholders to secure their data.

Privacy laws now exist in most countries. They are designed to ensure these duties are fulfilled and set penalties for failing to secure stakeholder data against misuse, including failure to protect those data elements against cyberattacks.

The purpose of post-breach eDiscovery (as opposed to legal eDiscovery) is to:

- Find elements of personally identifiable information (PII) within compromised data sets.
- To associate PII data with individual stakeholders, who can then be notified after cyber breach, in compliance with Privacy Laws.
- Quantify the extent of the breach and the amount of PII data put at risk.

After a cyber breach, time becomes particularly expensive. In this context, using an experienced eDiscovery team can make a significant difference in quickly understanding the extent of the data at risk and how to handle subsequent steps.

Acronyms & Definitions

EDRM: Electronic Discovery Reference Model
ESI: Electronically stored information
HIPPA: Health Insurance Portability and Accountability Act
FERPA: Family Educational Rights and Privacy Act
PHI: Personal Health Information
PII: Personally identifiable information
PIPEDA: Personal Information Protection and Electronic Documents Act
GDPR : General Data Protection Regulation
The Data Protection, Privacy and Electronic Communications – Explanatory Memorandum

Dark Web Points of Reference

The Dark Web (often referred to): https://en.wikipedia.org/wiki/Dark_web
Overlay Network (part of the Dark Web structure) https://en.wikipedia.org/wiki/Overlay_network
The Tor Network (often used when ransomware is involved) https://en.wikipedia.org/wiki/Tor_(network)
The Tor browser (often used when ransomware is involved) https://en.wikipedia.org/wiki/Tor_(network)#Tor_Browser

---

Thibault Dambrine is an IT consultant with Keyera Corp. (keyera.com) in Calgary, Alberta.
At the time of writing, he was working for CyberClan (http://www.cyberclan.com).
Thibault can be reached at dambrine@gmail.com.

Laura Smith is the Head of eDiscovery and Project Management for CyberClan, based in the United Kingdom.
She can be reached at mlaura.smith@cyberclan.com.

Thanks and gratitude also goes to each and every reviewer who helped us make this article what it has become.