大數據鑑識：

作者：陳君儀博士(Chun-I P. Chen, Ph.D.) / 美國加州州立大學富勒頓分校教授
編譯：王朝煌 江冠穎 王敦賢 / 中央警察大學資管系

摘要

有些網路觀察家認為某些國家會贊助網路活動進行資料盜取、擾亂網路秩序或更改資訊等行為，顯然大多數的網路犯罪是由一群具專業及高超技術的駭客，協助犯罪組織劫持信用卡和銀行資訊，甚至駭入企業資料庫和竄改紀錄，以收取高額的傭金的犯罪行為。現今的企業不但已面臨與詐欺和貪汙相關的巨大挑戰，在數位時代更曝露於新型威脅，包括身分竊盜、隱私資料竊盜、和智慧財產權竊盜等問題。隨著大數據活動和事件逐漸成為我們日常生活中的一部分，網路犯罪正隨時隨地潛入隱藏在各種數位裝置之中。為了因應網路犯罪嶄新的挑戰，大數據鑑識和鑑識工具的應用，變成為極其重要的課題。本文介紹大數據鑑識基本的概念和原理，涵蓋大數據、資料探勘和機器學習，以及大數據分析技術與大數據鑑識等等。期能透過闡述大數據鑑識相關的概念和原理，提供對此課題有興趣的人士，繼續鑽研或研究的參考。

關鍵詞 **知識階層、大數據、資料探勘、機器學習、數位鑑識、大數據鑑識。**

1.前言

現今的企業不但面臨與詐欺和貪汙相關的巨大挑戰，在數位時代更曝露於新型威脅，包括身分竊盜、隱私資料竊盜、智慧財產權竊盜和資料外洩等等。為了因應大數據環境數位鑑識的課題，對大數據鑑識科技必須有深入的了解。本文的目的主要在於闡述大數據鑑識的概念和原理，包括大數據、資料探勘、機器學習，以及相關的演算法和大數據分析等基礎概念，進而介紹數位鑑識和大數據鑑識的內涵。最後整理大數據鑑識所面臨的挑戰，並介紹目前常用的數位鑑識工具及其鑑識分析技術。

概念、挑戰工具和技術

2. 大數據：術語、概念和原理

2.1知識階層模型與大數據

知識階層模型（DIKW Model）代表資料（data）、資訊（information）、知識（knowledge）和智慧（wisdom）。知識階層模型說明兩個重要的概念：即解釋資料、資訊、知識和智慧的意義，以及描述從資料整理歸納資訊、再從資訊挖掘知識、以及再從知識轉化為智慧的過程。資料乃生活中的每一事實，與上下文的背景較無關係。資訊除了包括這些事實外，還包含上下文的背景和觀點，讓我們可以看出資料之間的關係。知識包括資訊以及資料間的關係模式（pattern）。了解了模式之後，才能進一步從中獲得新的知識。從理解這些模式與行為之間會發生的因果關係，進而獲得知識即所謂的智慧，這也是從知識轉化成智慧的附加價值。大數據是從複雜的資料，結構化、半結構化或非結構化的資料中，將其轉換成可理解的結構，發現隱藏於資料中的模式，萃取可操作的知識（actionable knowledge）及預測未知的數據，以輔助支援或取代決策之用。DIKW模型如圖1所示。

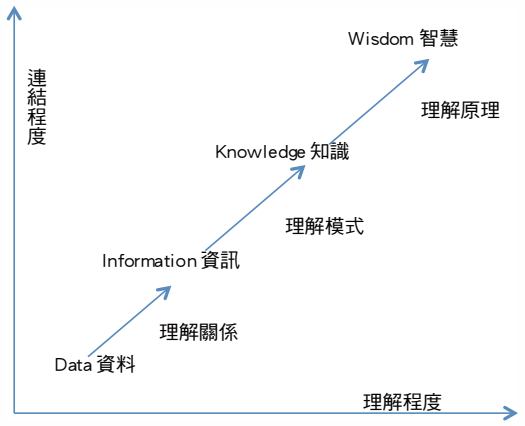

圖1.DIKW模型（資料來源：http://www.systems-thinking.org/dikw/dikw.htm）

D = Data：符號、信號等沒有上下文的事實資料。

I = Information：可以用來理解資料間關係的資訊（提供何人、何事、何地、何時、及多少的答案）。

K = Knowledge：可以作為理解資料中的模式、描述資料與資訊的運用，並可解答如何的知識。

W = Wisdom：能整合知識和資訊，以及理解模式為何發生的因果關係，進而洞察事實真相。

2.2大數據定義和概念

大數據乃從組織內部和外部的數據來源（包括系統、使用者、應用程式和感測器等等）蒐集彙整成龐大且複雜的（軌跡）數據所成的集合。大數據包括四個特色（4個V），即（1）容量（volume）：數據非常龐大，從兆位元組到千兆位元組；（2）速度（velocity）：數據以指數級數的速度快速增加；（3）多樣化（variety）：數據存在的形式非常多元，包括結構化、半結構化和非結構化的數據；（4）價值（value）：大數據的挑戰為確定哪些是有價值的資料，以便能進而獲取、轉換、萃取和分析數據。帝茂羅（De Mauro, et al., 2016）等提出了一個大數據的新定義：大數據是以高容量、高速度和多樣性為特徵的資訊資產，需要特殊的技術和分析方法才能將大數據轉變化為價值。換言之，大數據需要新一代的技術和處理架構，才能藉由高度之技術獲取資料、發現知識和進行模式分析，從大量各式各樣的數據中有效地萃取知識，獲取隱藏於大數據中的價值。大數據的分析技術包括資料探勘、機器學習演算法和技術等等，才能從資料中獲取支援決策的資訊和提升競爭優勢。

2.3 資料探勘和機器學習

2.3.1 資料探勘（Data Mining）

米斯拉將資料探勘定義為從現有的資料庫解析獲取有意義的洞見及分析企業用戶消費紀錄的過程（Mishra, 2016）。因此，資料探勘是一個從資料中挖掘知識，即從大量資料中萃取有用的、未知的、及具潛在應用價值的模式或資訊的過程。資料探勘是以複雜的數學演算法挖掘隱藏於資料間的關係，包括識別資料中的模式（pattern）、關聯性和群集關係等。資料探勘的方法可分為兩個類別：描述型（descriptive） 和預測型（predictive）。描述型的方法是找到資料中的特徵模式，主要包括分群（clustering）和關聯規則分析（association rule analysis）。預測型方法則是使用一些變量（variables）來預測未知的數據或未來資料的分類，主要包括分類技術（classification）、回歸分析（regression）、時間序列分析（time-series analysis）和預測技術（prediction）。

2.3.2 機器學習（Machine Learning）

機器學習是運用績效評估導引從數據推導決策模型的訓練學習技術(Gollapudi, 2016)。概念（concept） 是物件、符號或事件，因具有共同的特徵模式而組成的一種集合。電腦善於歸納概念或模式。我們可以訓練電腦從已知類別的標記資料（labeled data）中學習或訓練及歸納推導決策模式。也就是說機器學習是一種從資料中自動學習與分析，獲得規律，並利用所獲得的規律對未知資料進行預測的演算法。機器學習模式可分為兩種：監督式學習和非監督式學習。

2.3.2.1監督式學習（Supervised Learning）

監督式學習是從已知類別標記的歷史資料，運用訓練演算法推導機器分類模型，並以訓練得到的機器分類模型預測未來資料的類別（Tiwary, 2015）。監督式學習可從過往資料的特徵來對未來資料做預測判斷。例如垃圾郵件自動分類， 首先觀察目前信箱的信件，將郵件分為垃圾或非垃圾郵件，再將這些已知標記的電子郵件資料，輸入學習演算法訓練推導垃圾郵件的機器分類模型，再運用機器分類模型判定未來的郵件是否為垃圾郵件。監督式學習主要包含分類和迴歸兩種預測方法。

2.3.2.2 非監督式學習（Unsupervised Learning）

非監督式學習主要的功能是分析和挖掘未標記資料中的潛藏結構（Kaluža, 2016）。非監督式學習是指在沒有任何標記資料訓練的情形下，尋找資料間的關聯和型態的技術。例如，亞馬遜公司使用協同過濾的機器學習技術，根據客戶購物紀錄與其他客戶以前購物紀錄的相似性，推導出客戶可能會喜歡的產品，作為購物推薦的依據。分群和關聯規則分析演算法是非監督式學習的典型範例。

2.4 大數據分析（Big Data Analytics）

大數據分析主要經由分析大數據以提供過去、現在和未來的統計資料和有用的洞察見解，以期能做出更好的企業決策（Ankam, 2016）。大數據分析的意涵在於大數據資料集的處理，以及獲取有意義且可運用的知識。這種知識可以是模型、關聯或有用的洞察見解，諸如可以幫助一個組織了解目前市場趨勢、顧客偏好、甚至潛在的商機。大數據分析可分為描述性分析、診斷性分析、預測性分析和指導分析等四種，如圖2所示。

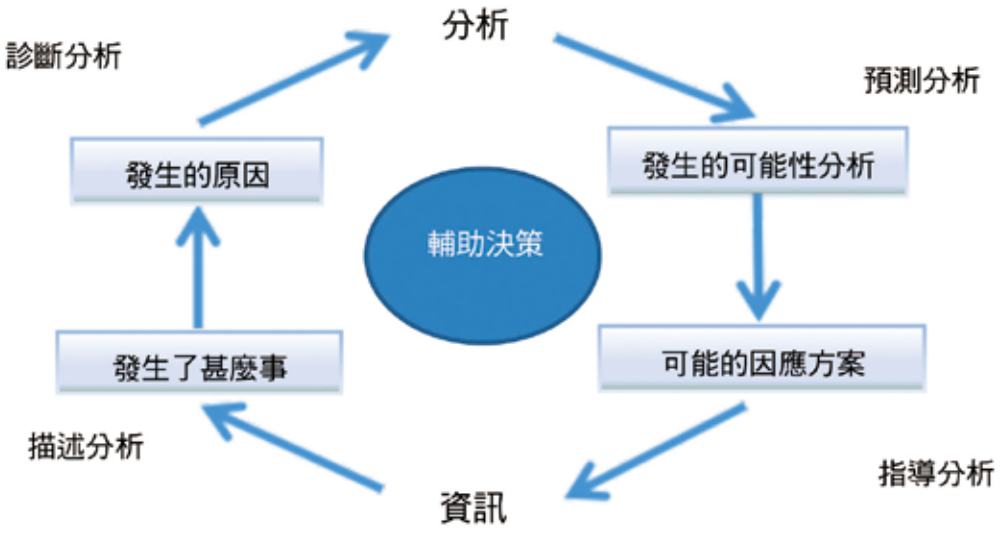

圖2. 資訊、分析與決策（資料來源：Corcoran, 2016）

（1）描述性分析（Descriptive Analytics）：描述性分析使用數據集合和資料探勘來提供對過去的了解，並可回答像「曾經發生過什麼」之類的問題。

（2）診斷性分析（Diagnostic Analytics）：診斷性分析用於發現和確定「為什麼發生」以識別和驗證兩個事件之間的因果關係。

（3）預測性分析（Predictive Analytics）：預測性分析使用統計模型和預測技術來推知未來，回答「什麼可能會發生」之類的問題。

（4）指導分析（Prescriptive Analytics）：指導分析使用優化和模擬等演算法來建議可能解決方案及其結果，以回答「我們該怎麼辦」之類的問題。

目前企業所面臨的許多問題，正驅使企業組織必須採取大數據及資料分析導向的策略，才能克服，如表1所示：

表1. 大數據分析解決方案

企業問題	範例
優化業務營運	銷售，定價， 獲利力，效率
識別企業風險	客戶流失，詐欺，呆帳
預測新的商機	向上銷售， 交叉銷售，最好的顧客與前景
遵守法律法規要求	反洗錢， 公平貸款，巴塞爾協II-III， 薩班斯-克斯利法案

（ 資料來源：EMC, 2015 ）

2.5 大數據分析技術

常用的資料探勘和機器學習演算法如表2所示，主要用於挖掘大數據分析中特殊型態或預測資訊。

表2. 常用的資料探勘和機器學習演算法

演算法	描述
分群（Clustering）	分群是一種將類似的物件歸屬同一組的非監督式學習演算法。相似度的計算方式如毆基里德距離（Euclidean distance） 和k-平均（K-Means）演算法 等。
分類（Classification）	這類演算法通常是監督式的，需要分類演算法和有助於辨識組別的訓練資料集。簡而言之，分類演算法可以將每個資料點或物件歸到最適合的類別。
迴歸分析（Regression）	迴歸分析用於估計自變量與應變量間的關係。例如運用相關的指標預測銷售量等重要決策所需資訊。

關聯規則（Association Rule）	這類演算法通常用於探索或發現存在於大數據集中物件間的關係。常見的例子包括購物籃分析、點擊流分析和詐欺偵測等等。

3. 數位鑑識及原理

3.1數位鑑識（Digital Forensics）

數位鑑識是指從數位媒體中蒐集、分析和保存涉及司法案件的數位證據之相關技術和方法。麥克肯密錫（McKemmish, 1999）將數位鑑識定義為：以法庭可接受的方式來識別、保存、分析和呈現數位證據的過程。因此，數位鑑識是使用能以科學驗證的方法，來蒐集、驗證、識別、分析、解釋、紀錄和呈現數位證據，證明犯罪。

3.2數位鑑識工作（The Digital Forensics Process）

數位鑑識工作可以整理組織成為一系列步驟。山盟提出了八個步驟的數位鑑識模型，為現場實作提供了良好的參考（Sammons, 2014）。八個步驟為：

（1）搜索權限（Search Authority）：搜索權是鑑識的第一個步驟。在刑事案件中，通常需要搜索票或傳票；在民事案件中，則須有法院的授權。

（2）監管鏈（Chain of Custody）：有紀錄良好的監管鏈對於保持證據的完整性是必要的。通常監管鏈透過表格、報告、證據單據、紀錄和證據標示進行紀錄。每次證據經手過程，皆應記錄下來。

（3）映像和雜湊值（Imaging and Hashing）：由於原始物件很容易被修改甚至破壞，實務上必須盡可能避免在原始物件進行檢驗分析。因此，一般需準備原始物件的映像檔（或稱鑑識映像檔），並在映像檔上進行所有的檢驗工作。雜湊值本質上是特定文件、媒體片段…等等的「數位指紋」。數位指紋可以用來檢驗比對原始證據和鑑識映像檔內容的一致性，兩者之雜湊值完全相同，表示映像檔未被破壞或汙染。

（4）驗證工具（Validated Tools）：數位鑑識取證的工具，無論是硬體、軟體或儲存媒體，在使用之前都必須驗證其準確性。尤其是新的鑑識取證工具及更新版軟體更應加以驗證。且這些驗證過程都必須記錄下來。

（5）重複性（Repeatability）：鑑識工作必須確保鑑識結果準確無誤。鑑識結果必須具備可重複性，即任何獨立的測試者運用相同的工具、相同的步驟都應該得到相同的鑑識結果。

（6）分析（Analysis）：檢驗者用他們的技巧、經驗和工具來尋找和解釋在媒體上的數位軌跡及其成因，並在檢驗工作結束時提出判讀意見。通常這種意見以可能性的程度（例如極不可能、不可能、可能、非常可能…等等）表達，而不是以絕對的是或否來表達。

（7）報告（Reporting）：檢驗者也必須提出一個清楚易懂的檢驗報告。內容應包含： 檢驗工作過程概要、檢驗的原始證據清單、用於檢驗分析的方法和使用的工具、檢驗的重要發現、結論和任何相關的證據。簡言之，報告必須以清楚且精準的方式撰寫。

（8）專家解讀（Possible Expert Presentation）：對非技術人員，如法官等，介紹複雜的科技及其操作細節，是一件不容易的事。必要時可由相關的專家幫忙針對鑑識結果提供淺顯易懂的解說。

4. 大數據鑑識與鑑識大數據

4.1大數據鑑識（Big Data Forensics）

傳統電腦鑑識大多聚焦在一般的資料來源，例如行動裝置和筆記型電腦。大數據鑑識是從大數據系統（big data system）中處理數據的識別、蒐集、分析和展示呈現（Sremack, 2015）。大數據鑑識並非取代傳統的電腦鑑識，而是加強電腦鑑識技術和知識，以便能在大量且分散的大數據系統中進行鑑識處理工作。大數據鑑識（big data forensics）是一種新型態的電腦鑑識，就像大數據是一種為處理大量且複雜多元資料而發展的新型資料處理方法，大數據鑑識的主要目標是從分散的檔案系統、大規模檔案庫以及相關應用程式系統中蒐集證據資料。

4.2大數據鑑識vs.傳統鑑識

由於大數據系統往往儲存大規模的巨量資料，其鑑識過程所需具備的條件已超越了傳統電腦鑑識的需求。例如傳統電腦鑑識使用MD5和SHA-1工具來驗證資料，不但耗時且鑑識過程必須將系統停機進行取證，往往導致干擾系統的正常運作。由於大數據系統牽涉範圍極為廣泛，實務上調查人員不能應用這些技術進行大數據系統的鑑識工作。

4.2.1大數據資料蒐集方法

鑑識調查的證據乃儲存在系統的數位資料。這些資料可以是檔案的內容、詮釋資料（metadata）、被刪除的檔案、在主記憶體中的資料、硬碟的殘存空間內容等等。電腦鑑識技術期望能擷取所有相關的資訊，包括被刪除的資訊。在傳統電腦鑑識中，系統可以被停機進行離線取證，藉由移除硬碟及創造一份供鑑識使用的映像檔來操作。但由於大數據系統是龐大、牽涉範圍廣泛、且複雜的分散式系統，可能無法像傳統電腦鑑識將整個大數據系統停機進行離線取證。因此大數據系統證據的蒐集通常藉由邏輯檔案複製或使用查詢語言指令等方式，針對鑑識標的進行取證工作（Sremack,2015）。

4.2.2大數據資料驗證

傳統電腦鑑識主要以MD5和SHA-1來證明所蒐集資料的原始性，通常其需耗費大量的時間來計算所蒐集資料的MD5和SHA-1數位簽章（數位指紋）。然而在大數據系統取證，要計算數

以兆位元組計資料的數位簽章，極可能因為緩不濟急而行不通。替代方案大多以總量管制，蒐集電腦日誌記錄，以及其他描述性資訊來證明資料的原始性。

5. 大數據鑑識工作的挑戰

綜合而言，執行大數據鑑識工作時，可能會遇到下列挑戰：

（1） 數據處理和分析：由於大數據往往是半結構化和非結構化的數據，鑑識工作必須將大量的檔案、電子郵件、貼文以及其他訊息資料進行分析與鑑識，以辨識詐騙及發現可疑的行為跡證。

（2）大數據的資料量多到難以處理：大數據鑑識需要以高效率的分析方法來處理數十億筆的原始資料及上億種的資料組合，運用傳統的電腦鑑識技術可能無法預先辨識大數據系統中的異常跡證，以作為預防性的犯罪管理。

（3）大數據鑑識分析方法的採用：及早運用大數據鑑識方法即時分析各種大數據檔案，預先偵測及處理犯罪行為，可以有效減少或避免以獲取金錢為目的之非法犯罪活動，例如洗錢、賄賂、貪汙、購物詐騙、採購或拍賣詐欺等等。

（4）大數據鑑識人員之短缺：目前符合資格（擁有資料科學技巧、實際操作工具經驗、能夠執行在大數據資料鑑識調查案件、及資料鑑識分析）的鑑識分析人員，為數很少且培養不易。

（5）大數據鑑識作業標準之缺乏：在鑑識實務中，大數據鑑識專家必須具備的資格與大數據鑑識的作業標準，目前仍缺乏共同的標準與鑑識作業之指導方針。

（6）大數據鑑識人員之證照與法律議題：對於大數據鑑識人員取得證照的要求，即便在先進的國家如美國，聯邦及各州的規範也不盡相同。且法庭上對於數位證物的必須具備的條件，各州也有不同規範。

（7）數位證據偵測與取證工具的認可：相較於實體證據，數位鑑識不但範圍更為廣泛，且更具個人機密性及移動性，因此鑑識人員也需要具備各種不同的訓練與輔助工具。此外鑑識工具的認可，對於數位證據之偵測與取證而言，也是非常重要的（Goodison, E. et al., 2015）。

（8）大數據證據之辨識：從大量、多樣的大數據資料中，辨識出具有證據價值的資料，比一般電腦鑑識更具挑戰性。

（9）雲端鑑識：對於雲端平台，例如IaaS雲端而言，蒐集與萃取證據可能極具挑戰性。主要乃因在雲端平台上，其網路流量非常大而且速度極快，蒐集及處理相關的證據，需要更先進的硬體及軟體工具。

（10）大數據蒐集方法：由於大數據系統是大型且複雜的系統，它們可能無法像傳統鑑識方法一般，被停機進行離線取證進行鑑識工作（Sremack, 2015）。

（11）大數據驗證：傳統鑑識主要倚賴MD5和SHA-1來證明所蒐集資料的原始性，通常需花費許多時間計算所蒐集數據資料之MD5和SHA-1數位簽章。然而在大數據系統取證，要計算數以兆位元組計資料的數位簽章，極可能因為緩不濟急而行不通（Sremack, 2015）。

6. 大數據鑑識分析之工具和技巧

6.1數位鑑識工具和認證

在數位鑑識實驗室中，經常使用之硬體和軟體鑑識工具，詳列如表格3（Sammons, 2014）。

表3. 數位鑑識工具

工具名稱	用途	公司及網址（URL）
Forensic Toolkit	多種用途的鑑識工具（存取、辨識、尋找、報告、抹除等等）	Access Data Group, LLC http://accessdata.com
EnCase	多種用途的鑑識工具（取證、驗證、搜尋、產出報告、抹除等等）	Guidance Software, Inc. http://www.guidancesoftware.com
SMART for Linux	多種用途的鑑識工具（取證、驗證、搜尋、產出報告、抹除等等）	ASR Data, LLC http://www.asrdata.com/forensic-software/
X-Ways Forensics	多種用途的鑑識工具（取證、驗證、搜尋、產出報告、抹除等等）	X-Ways Software Technology AG http://www.x-ways.net/forensics/
Helix3 Pro	多種用途的鑑識工具（取證、驗證、搜尋、產出報告、抹除等等）	e-fense, Inc. http://www.e-fense.com/products.php
Softblock, Macquisition, Blacklight	麥金塔系統多種用途的鑑識工具	BlackBag Technologies, Inc. https://www.blackbagtech.com/forensics.html
Mac Marshal	麥金塔系統多種用途的鑑識工具	Forward Discovery, Inc. http://www.forwarddiscovery.com/Raptor

Raptor	Linux系統取證與預覽工具	Forward Discovery, Inc. http://www.forwarddiscovery.com/Raptor
Forensic Dossier	硬體取證	Logicube, Inc. http://www.logicube.com/
Forensic Hardware Tools	防止寫入、橋接、儲存、取證等等	Tableau, Inc. http://www.tableau.com/ Wiebetech, Inc. http://www.wiebetech.com/home.php

（資料來源：Sammons, 2014）

6.2 認證

美國刑事鑑識實驗室主管及實驗室認證委員學會（ASCLD/LAB）被公認為是鑑識實驗室認證的世界領導者。除了ASCLD/LAB，美國測試和材料協會（ASTM）國際組織也提供在鑑識科學的各種標準與規範，包括數位鑑識。

6.3常用的大數據鑑識技巧

各種不同形式的資料集，不論是結構化的、半結構化或非結構化的，都可以藉由大數據的解決方法來進行處理。例如藉由查察大數據中的惡意活動、不尋常的事件，或尋找不經常發生但重要的型態，可以偵測並進而預防詐騙和網路犯罪。以下為常用的大數據鑑識分析技術：

（1）連結分析（Link Analysis）：連結分析是一種用來發現和評估數據間的關係和連結數據的分析技術，包括組織、群眾和交易活動。例如連結分析技術可以用來分析電信事業一般個人用戶的電話紀錄、行動電話帳單及用戶存取地點的紀錄，幫助公司繪製出其用戶的活動軌跡；另外藉由連結分析，企業也能夠發現”潛藏的”關係和可疑員工與供應商間的資訊外洩管道；此外大規模數據連結分析也可以用來追蹤複雜的金融交易，視覺化實體間的關係，進而辨認出不尋常的型態。

（2）社群網路分析（Social Network Analysis ）：社群網路分析主要以網路理論，分析個體間的網路關係。節點代表社群中的個體。連結代表個體間的關係，如朋友關係、親屬關係、及組織階層等等。例如社群網路分析以及連結分析可以幫助我們辨識出詐騙案中相關聯的團體、利益衝突關係、圍標操作關係、及其他詐騙行為。

（3）概念分群（Concept Clustering）：概念分群主要將相似的個體或行為歸類在具有個別意涵的群組，並可藉由概念分群辨識出異常行為或個體。概念分群可以有效率地運用在大數據分析，包括結構化的、半結構化或非結構化資料的處理。

（4）情感分析（Sentiment Analysis）：情感分析又稱為行為分析，主要運用文字分析技術來辨識和萃取貼文作者的看法、感情狀態、及內在的情感商數等主觀上的資訊。情感分析可以判定文章所表達的意見是正向的、負面的或是中性的。例如文字探勘、情感分析和概念分群的運用可以從交易、電子郵件、及文字檔案資料辨識詐騙、賄賂、浪費和濫用等行為。

（5）數據視覺化（Data Visualization）：可以從大數據中辨識出”潛藏的型態”。 數據視覺化技術已經被證實十分有效，因為相較數字或文字而言，人類更能夠從圖像中吸收大量資訊。例如將相關的調查結果以視覺化技術呈現，能更有效的辨識詐騙犯罪。

（6）關聯視覺化和交易流分析（Relationship Visualization & Transaction Flow Analysis）：可以應用在分析社群網路，凸顯人群、個體、事件之間關係型態和互動形式。關聯視覺化和交易流分析可以應用在不同的調查案件，幫助釐清個人、數據和物件間的連結與關係。

（7）安全上的大數據分析和科技（Big Data Analytics and Technologies for Security）：大數據科技例如Hadoop生態系、 NoSQL 資料庫、串流分析和複雜事件處理技術等，能夠高速分析大規模及異質性的大數據。大數據分析技術可以應用在網路流量分析、日誌檔案及金融交易資料分析，將相關聯的各種資訊清楚地呈現，以協助辨識可疑的活動和異常行為（Cárdenas et al., 2014; Kandanoor, B., 2015）。

（8）雲端鑑識平台FaaS（Forensic-as-a-Service）：雲端鑑識平台的優勢在於能讓數位鑑識人員將大量的日誌檔案儲存在雲端檔案或資料庫，以便於能隨時隨地的進行資料存取和證據搜尋工作。雲端鑑識平台的運用將使鑑識人員可以利用MapReduce的計算模型來處理日常的鑑識工作。

（9）人工智慧分析（Artificial Intelligence Analysis）：模型基礎的分析方法，能夠藉由結合各種取證方法所得到的資料和各種來源的資料，達到自動化鑑識的效果。使用基於人工智能分析規則的方法，可以將商業規則邏輯、法規、及訴訟需求等等，建立自動化系統來協助辨識和解讀資料（Ernst & Young LLP,2013）。

7. 結論

本論文說明大數據和數位鑑識的基本概念，以及資料探勘和機器學習的概念及相關的演算法。由於大數據分析大多使用上述的演算法，資料探勘和機器學習已成為大數據鑑識工作重要的方法，為大數據鑑識人員必備的技能和知識。本論文也介紹了數位鑑識和大數據鑑識的概念和原理，並歸納大數據鑑識的各種挑戰。此外本文也整理目前常用的電腦鑑識工具和技術，以提供有興趣的研究人員繼續鑽研的參考。

二十一世紀智能社會正邁入一個以數據主導的數位經濟時代，新的數位科技如大數據科

技、物聯網、雲端運算和行動網路等等不斷地推陳出新，近年來大數據活動和社群媒體事件呈現以指數級數方式增加的爆炸性增長。此外我們也正面臨網路犯罪的快速增長，各種威脅及詐騙也不斷地擴散及潛藏於雲端和數位裝置的社群媒體中。數位鑑識分析急需採用新的工具、硬體和軟體、以及新的方法和程序來因應大數據鑑識的需求與挑戰。雖然數位鑑識工具和科技在發現異常的型態、行為和趨勢進而偵測潛在的詐騙犯罪，扮演著關鍵的角色，然而除了具備先進工具和技術外，也需要訓練成熟的技術團隊，才能有效地進行大數據的鑑識工作。因而，充實精進鑑識科學教育和訓練也是極其重要且迫切的一環。FACT

參考文獻

1.Ankam, V.（2016）. Big Data Analytics, Packt Publishing, Print ISBN-13: 978-1-78588-469-6.

2.Cárdenas, A.A., Pratyusa, K. M., and Rajan, S. P. （2014）. "Big Data Analytics for Security, " Retrieved from https://www.infoq.com/articles/bigdata-analytics-for-security

3.Corcoran, M.（2016）. "The Five Types of Analytics, " Retrieved from http://www.informationbuilders.es/sites/www.informationbuilders.com/files/intl/co.uk/presentations/four_types_of_analytics.pdf?redir=true

4.Mauro D. A., Greco, M., and Grimaldi, M.（2016）. "A formal Definition of Big Data Based on Its Essential Features," Library Review, Vol. 65 Issue: 3, pp.122 – 135.

5.EMC Educational Services.（2015）. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley & Sons, January 27.

6.Ernst & Young LLP.（2013）. " Forensic Data Analytics, " http://www.ey.com/Publication/vwLU

7.Gollapudi, S.（2016）. Practical Machine Learning, Packt Publishing, Print ISBN-13: 978-1-78439-968-9.

8.Goodison, S. E., Davis, R. C., and Jackson, B. A.（2015）. "Digital Evidence and the U.S. Criminal Justice System: Identifying Technology and Other Needs to More Effectively Acquire and Utilize Digital Evidence," Santa Monica, CA: RAND Corporation, Retrieved from http://www.rand.org/pubs/research_reports/RR890.html.

9.Kaluža, B.（2016）. Machine Learning in Java, Packt Publishing, Print ISBN-13: 978-1-78439-658-9.

10.Kandanoor, B.（2015）. "Big Data - Helping Security Analytics, " Retrieved from https://www.linkedin.com/pulse/big-data-helping-security-analytics-bharat-kandanoor?articleId=6008687518009094144

11.McKemmish, R.（1999）. "What is forensic computing? " Trends & issues in crime and criminal justice, no. 118, Canberra: Australian Institute of Criminology.

12.Mishra, P.（2016）. R Data Mining Blueprints, Packt Publishing, Print ISBN-13: 978-1-78398-968-3.

13.Sammons, J.（2014）. The Basics of Digital Forensics, 2nd Edition, Syngress.

14.Sremack, J.（2015）. Big Data Forensics – Learning Hadoop Investigations, Packt Publishing, Print ISBN-13: 978-1-78528-810-4.

15.Tiwary, C.（2015）. Learning Apache Mahout, Packt Publishing, Print ISBN-13: 978-1-78355-521-5, March 30.