REPHRAIN

Protecting citizens online

Towards Data Scientific Investigations:

A Comprehensive Data Science Framework and Case Study for Investigating Organized Crime and Serving the Public Interest

Prepared by Erik van de Sandt [*†‡§], Arthur van Bunningen[†], Jarmo van Lenthe[†], John Fokker[¶]

March 2021

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

REPHRAIN
Protecting citizens online

Executive summary

Big Data problems thwart the effectiveness of organized crime investigations.
The practice of attribution - who did what - currently faces heavy weather as law enforcement operations are becoming increasingly complex. One of the reasons for this effectiveness crisis are the multitude of security practices by organized crime in today's Information Age - deception, deletion and encryption to name a few - that thwart all phases of the investigation process. Even when law enforcement are able to retrieve evidence via digital forensics, Big Data problems arise. The variety, velocity, veracity and volume of the collected evidence work as a countermeasure when law enforcement agencies are unable to process data into factual police reports. A frequently proposed solution is the introduction of 'smart' data science technologies to support criminal investigations.

The need for a common data science framework for law enforcement agencies.
Experience has taught us, however, that the transition to - what we call - data scientific investigations is nothing less than a paradigm shift for law enforcement agencies, and cannot be done alone. Legal, organizational and technical harmonization with other public partners requires a common understanding, including shared language, of data scientific operations. This white paper therefore presents guiding principles and best practices for data scientific investigations of organized crime, developed and put into practice by operational experts over several years, while connecting to existing law enforcement and industry standards such as, but not limited to, the Intelligence Cycle and CRISP-DM. The associated framework is called CSAE (pronounced as 'see-say'), an abbreviation of Collect, Store, Analyze and Engage. In this paper, we share CSAE's comprehensive data science approach with a broader academic and practitioner audience, including the details of our public interest philosophy, methodology, business process and associated policy agendas.

CSAE is methodology, business process, policy agenda and public interest philosophy.
Our data science methodology is a simple mixed-methods approach that combines qualitative and quantitative sources from both the criminal and safety and security communities while using a variety of mixing purposes, research designs and strategies. This approach, that we named Quadrant, runs like a thread through all phases of our business process as it creates foresight, hindsight, insight and oversight for both business intelligence and operational purposes, and to both run and change the business of data scientific investigations. Applying this approach to gain, and subsequently formalize, a strategic business and data understanding about a particular crime theme, including appropriate responses, is also the first step to give direction to the business process. The business process itself consists of four phases, more specifically obtaining data in Collect, warehousing information in Store, creating intelligence in Analyze and executing lawful interventions based on facts in Engage. Besides outlining the details of each phase and giving practical examples, we further argue how CSAE can also be used as a model to improve policy-making on international relations, legal and workforce issues. Throughout this paper, we explain our public interest philosophy and highlight the ethical design choices related to - amongst others - explainability, privacy and victim assistance that law enforcement agencies will face when implementing CSAE.

The future of CSAE and data scientific investigations.
It is because of this public interest philosophy that we share our comprehensive data science framework for investigating organized crime. Although CSAE is a proven practice, we acknowledge that the development and integration of data scientific operations in law enforcement agencies is still in its infancy. Because so much needs to be done before data scientific investigations become an established field of study, we hope that CSAE promotes harmonization between law enforcement agencies as well as research on and with law enforcement by academics. CSAE is a living document and will continue to be updated and improved when the academic world, industry and law enforcement provide feedback on concepts and implementation. As CSAE is put into greater practice, additional lessons learned will be integrated into future versions. This will ensure CSAE is meeting the needs of law enforcement in the dynamic and challenging environment of crime and investigations. After all, the effectiveness of criminal investigations affects society as a whole as security is a cornerstone of liberal democracies that are governed by the rule of law.

REPHRAIN
Protecting citizens online

Contents

List of Figures

List of Tables

*Primary and corresponding author (ev18710@bristol.ac.uk); †National Police, The Netherlands; ‡University of Bristol, United Kingdom; §National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN), United Kingdom; ¶Private security industry.

This white paper was presented at the Third INTERPOL-UNICRI Global Meeting on AI for Law Enforcement on November 25, 2020. Disclaimer: The views and opinions expressed in this white paper are those of the authors and do not necessarily reflect the official policy or position of the Dutch National Police.

REPHRAIN
Protecting citizens online

1 | Introduction: Adding A New Layer to Investigations

As depicted in Figure 1, criminal investigations can be simplified as an onion that consists of three interrelated layers. The core of this model is the social and behavioral perspective of - what we nowadays call - traditional investigations, while the second layer is the technical perspective of digital investigations that was introduced in the eighties of the last century [1]. Recently, a numerical layer with advanced mathematical/ statistical methods and techniques has entered the law enforcement arena [2]. The integration of social and behavioral, technical and numerical approaches leads to a new field of study: data scientific investigations. Because documentation on this novel type of investigations is absent, we present CSAE (pronounced as 'see-say'): a comprehensive data science framework for investigating organized crime.

Figure 1: Investigations can be represented as an onion model that consists of a traditional, digital and numerical layer. Each layer is an addition to, and should be integrated with, the previous layer, while contributing to the core business of investigations: attribution of who did what for prosecution purposes. As depicted in the Venn diagram of Figure 2, the overlap between the layers constitutes data scientific investigations.

From traditional to digital investigations

The goal of criminal investigations is the attribution of suspects of crime - who did what - for prosecution purposes, and law enforcement agencies (LEAs) have a monopoly on this process of bringing suspects to justice. Before our digital era, all criminal investigations were - what we now call - traditional, offline investigations and forensics. The focus of these investigations were mostly on the human factor of crime, thus very much social and behavioral in nature. Investigators used methods and techniques like observations, interrogations and eavesdropping for their truth-seeking process, supported by 'traditional' forensic sciences that focus on physical evidence from the offline world such as ballistics, DNA analysis, fingerprint analysis and pathology. When crime was enabled and assisted by, and focused on, information technologies (IT), evidence moved to hardware and online environments. As a result, digital investigations appeared, supported by digital forensics such as computer-, network- and mobile forensics, to recover evidence from data carriers. While being an additional layer to the core of traditional investigative methods and techniques, the introduction of digital investigations and forensics was nothing less than a paradigm shift when looking at the legal, organizational and technical effects on the whole legal justice system. In fact, the merge between traditional and digital investigations has been an ongoing process for many law enforcement departments around the world [3, 4].

The limitations of digital investigations

Today, criminal investigations face an effectiveness crisis [5, 6, 7, 8]. Investigations are too labour and time intensive, while outcomes - i.e., successful attribution, arrest and prosecution - are uncertain. Too many crimes go unsolved and too few suspects are apprehended, and this is a major problem: doing attribution poorly undermines the state's credibility, its effectiveness, and ultimately its liberty and security [9, p.4]. Yet academics and practitioners predicted the end of the 'Golden Age of Digital Forensics' more than a decade ago [10, 11]. They foresaw a situation in which evidence is permanently out of reach to investigators, or - when successfully retrieved - cannot not be properly analyzed because of data management issues. These predictions have come true. The security of organized crime has democratized, and law enforcement agencies have been unable to deal with these practices [12]. In other words, because professional criminals have access to a broad range of legitimate and illegitimate administrative, physical and technical services and products of a protective nature, law enforcement agencies have difficulties to collect the necessary evidence to build cases against suspects of crime. Even when law enforcement agencies are able to retrieve evidence with traditional and digital forensics, these same forensic methods and techniques are not able to process the variety, velocity, veracity and volume of data sets (commonly known as the 4Vs of Big Data) into timely, relevant, accurate and actionable reports. This is understandable: the security controls of professional criminals do not stop when data is collected. Evidence in data sets might be encrypted, come in unknown formats, be hidden as a needle in a digital hay stack, or be false because of deception tactics, to name just a few criminal countermeasures [12].

REPHRAIN
Protecting citizens online

The introduction of data science and comprehensive framework

So besides social and behavioral and technical perspectives, criminal investigations need a new layer that regards evidence as being *measurable*. In other words, a numerical approach is necessary to deal with the nature and extent of today's evidence, and require the introduction of new mathematical and statistical methods and techniques that produce *statistical sight* and *probabilistic evidence*. Since long, scholars have regarded a shift from digital forensics to *intelligent forensics* as the way forward to deal with the current effectiveness crisis [13][11, pp.26-27][14, pp.224-225][12]. As depicted in Figure 2, the integration of social and behavioral, technical and numerical perspectives is what we call *data scientific investigations*. While a numerical approach is an additional layer to traditional and digital investigative methods and techniques, the introduction and integration is - again - nothing less than a paradigm shift for law enforcement agencies, and requires major organizational, technical and legal reforms.

Although scholars rightly argue that attribution in general is a nuanced process and not a simple problem [9, p.7], a framework for data scientific operations is so far missing. We therefore present the CSAE model: a comprehensive data science framework for investigating organized crime that supports law enforcement agencies to become value-driven information technology organizations that serve and protect the interests of liberal democracies. Pivotal in this approach is the CSAE business process that consists of four phases:

1. *Collect* in which not only evidence from previous operations is obtained, but also new strategic data sets are lawfully acquired;

2. *Store* in which these multiple data inputs are warehoused, and converted into information;

3. *Analyze* in which related information points are combined with knowledge, and becomes intelligence; and

4. *Engage* in which intelligence is refined into facts that are used for lawful actions against crime.

The goal of this white paper is to present the first comprehensive framework on data scientific operations, with an emphasis on criminal investigations of organized crime, based on state-of-art industry standards and our experiences from operations against organized cyber crime. Without applying advanced mathematical methods and techniques, the business process is further applicable to agencies and departments that only conduct traditional and/or digital investigations of organized crime. Be- sides its operational utility, the framework can also be used for business intelligence (BI) purposes (see Section 3.1). Ultimately, we hope that CSAE not only promotes harmonization between law enforcement agencies to confront today's crimes more effectively, but also to conduct research with and on law enforcement. Regarding the former type of research, simple Internet search queries - e.g., 'H2020 tools for law enforcement' - reveal the many publicly funded projects that promise to make software for law-enforcement agencies. While we have reviewed many proposals and finished products from such public and private consortia, we have never encountered such tools in action in our own or partner agencies. In our opinion, software adoption might be increased when these consortia follow a standardized framework. Lastly, research on law enforcement agencies strengthens liberal democracies. We believe that the debate about the usage, scope and limitations of data science in criminal investigations becomes more specific, thus improves, when scholars use CSAE for critical thinking about the risks of data scientific investigations such as biases, privacy and surveillance [15, pp.33-34].

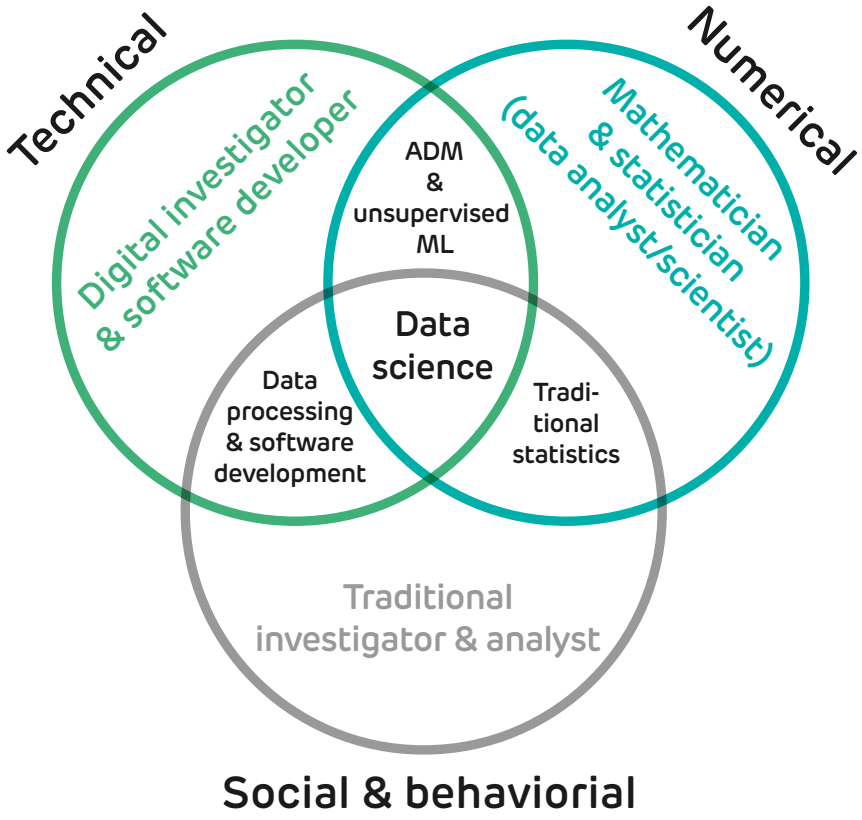


Figure 2: Traditional, digital and mathematical/statistical approaches stand on their own, but also overlap. Digital investigators may extract evidence from data carriers of suspects. Officers with a mathematical background - frequently called data scientists - subsequently create advanced statistical models to gain sight over these data sets. The results of these models are then interpreted by data analysts, and subsequently used by traditional investigators with domain- specific knowledge of crime and appropriate responses. The Venn diagram is further explained by the text box in Section 2.2.

Reading guide

While CSAE is applicable to other organized crime themes such as child sexual abuse material (CSAM), counter-terrorism and drugs trafficking, this white paper is build around examples related to investigations of organized cyber crime. The paper has the following structure:

- Section 2 - ***Background of Study*** **describes how the revolutionized security practices of criminals negatively affect the quality and quantity of today's evidence. The section further explains why data scientific investigations and technical harmonization between law enforcement agencies are interrelated objectives to gain and maintain the upper-hand in the cat-and-mouse game between criminals and law enforcement;**

- Section 3 - ***Description and Explanation of the CSAE Model*** **presents our business process for data scientific investigations by describing all steps from start to finish in detail, including our data science methodology and practical examples.**

- Section 4 - ***CSAE as a Model for Public Policy*** **demonstrates how the model supports, shapes and structures law enforcement agencies and their policy agendas on international relations (IR), public-private partnerships (PPP), law and human relations (HR).**

The paper is of a multidisciplinary (i.e., socio-techno-legal) nature, and intends to serve a broad audience of academics, private security researchers and legal practitioners, including the increasing number of law enforcement professionals with a technical and numerical background. Throughout this study, we explain the philosophy behind CSAE, and show how the framework aims at serving the public interest. The business process, methodology, philosophy and policies are not theoretical, top- down invented concepts. On the contrary, the framework has been developed by operational experts over several years, while connecting to existing law enforcement and industry standards (see Appendix B). As a result, the business process is a proven product in practice. Law enforcement officers who implemented CSAE have successfully developed a range of data science models, forensic tools and research methods and techniques. A last remark is that *italic* characters are an invitation to focus the reader's attention on CSAE's key concepts.

2 | Background of Study

As most large organizations in today's Information Age, law enforcement agencies face Big Data challenges related to the variety, velocity, veracity and volume of data sets. Yet the underlying reasons differ from e-commerce business, and are unique to the world of organized crime fighting. data scientific investigations are a reaction to ever-changing crime characteristics. The empirical justification for data scientific investigations are the *technical computer security practices* of criminals, called *deviant security* [12]: a game-changer in the cat and mouse game between those who trespass the law and those who enforce the law. This section goes beyond the well-documented 4Vs of Big Data, and describes how deviant security practices negatively affect the quality and quantity of evidence. Yet every modus operandi (MO) comes with all kinds of minor, major and even unavoidable weaknesses. While traditional and digital investigations exploit respectively human and technical weaknesses, this section subsequently explains why data science methods and techniques are the 'exploit' that LEAs need to improve the evidential quality of large data sets. This section finishes by arguing that the introduction of data science must go hand in hand with the objective of technical harmonization between law enforcement agencies to gain and maintain the upper-hand in the cat-and-mouse crime game.

2.1 How Organized Crime and Law Enforcement Negatively Affect Evidence

In today's Information Age, organized crime has many countermeasures at their disposal to thwart investigations. Law enforcement agencies are subsequently confronted with large, abstract and fragmented data sets that are riddled with security measures of professional criminals that affect the nature and volume of data. At the same time, most agencies cannot process these data sets into workable evidence. This is not only a legal and technical, but very much an organizational challenge as well. There is no comprehensive approach that inspires law enforcement agencies how to process today's evidence. As a result, law enforcement not only have to overcome the security practices applied by career criminals, but also their own, self created barriers and those of significant others in the safety and security community, that benefit criminals in the protection of crime.

Deviant security negatively affects the quantity and quality of evidence

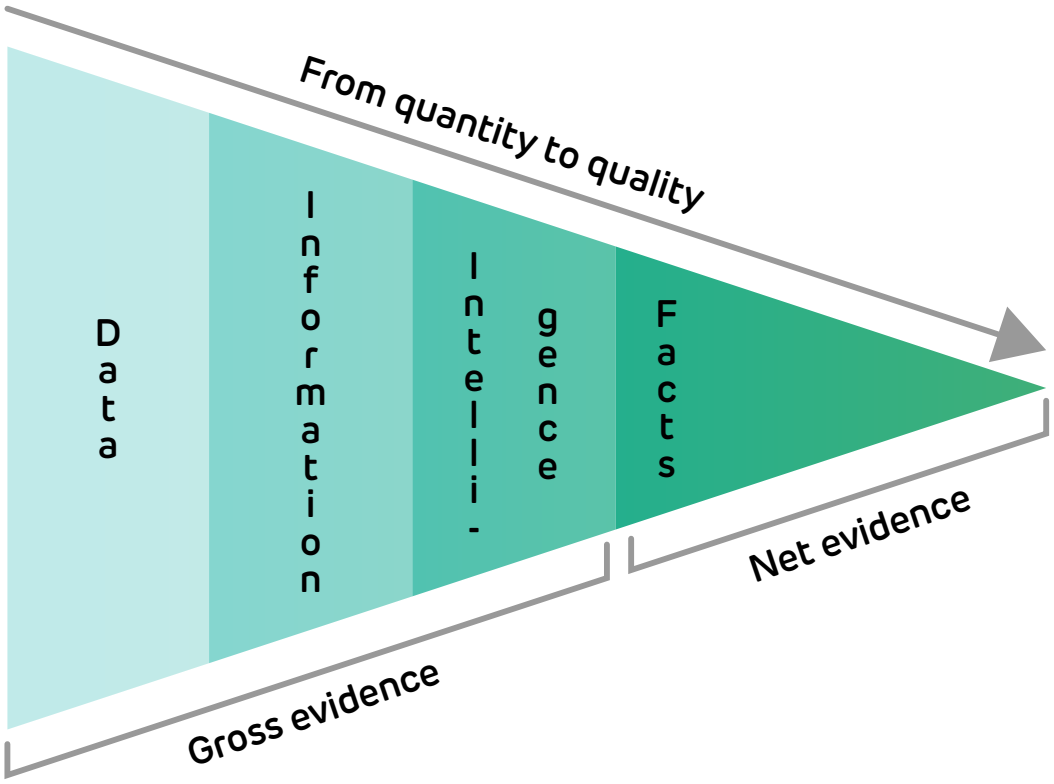
Police investigations face an effectiveness crisis: operations are too labour and time intensive with poor outcomes [5, 6, 7, 8]. One of the underlying reasons is the democratization of technical computer security, and a vast and highly accessible underground economy that provides professional offenders with the means to increase the protection of the criminal and his/her crimes [12]. Organized criminals have their own deviating version of defense and offense *security in depth*, consisting of layers of administrative, physical and technical controls of a preventive, deterrent, detective, corrective, recovery and compensating nature. In other words, the security of professional criminals does not stop after LEA has breached the first defensive layers of criminal organizations and retrieved the evidence they need by exploiting human and technical weaknesses in MOs. On the contrary, (parts of) the retrieved data are also riddled with countermeasures that affect the *availability, confidentiality* and *integrity* of evidence. While these three concepts are legitimate technical computer security objectives (commonly known as the CIA triad [16, p.2]) and well-accepted among the cyber security community, criminal organizations apply the CIA triad in their own deviant manner with all its consequences for the quantity and quality of today's evidence.

Availability of evidence is not only limited by countermeasures that promote data volatility - such as secure deletion - or distribution tactics that result in partial data sets. Availability of data is in today's Big Data era also increased. Not only have information technologies the ability to create and retain evidence for unlimited periods [17], but also do many career criminals generate large amounts of data by centralizing and sticking to the same online/offline locations without interruption, for a considerable amount of time. To conclude, the security goal of availability results in large, but fragmented data sets. Confidentiality leads to unlinkability, more specifically, the anonymity of the criminal and the unobservability of crimes. This security goal is not only achieved by technical countermeasures such as access control or encryption, but also by administrative (i.e., soft, management-oriented) controls like security policies that rely on association deniability, data minimization and external secrecy. As a result, entities - e.g., email accounts, IP addresses, monikers - cannot be interlinked to the criminal and/or his/her crimes. The integrity of data is also harmed. Deception, for example, hides the real and shows the false, and affects the authenticity and non-repudiation of findings. Such disinformation may not only create false leads, but also fake exculpatory evidence, and therefore has a major effect on the efficiency and effectiveness of investigations.

Such deviant countermeasures remain active after data have come into possession of law enforcement. As a result, the volume of data sets increases while the quality decreases. Because deviant security is not only defined and shaped by those who apply security (i.e., the criminals), but also by those who are confronted by it (i.e., the safety and security community), the absence of a solution that deals with these practices benefits the security of criminals. The next paragraph highlights some of the current incapabilities of law enforcement agencies to transform raw data into *facts*: statements about reality that go beyond reasonable doubt.

Evidence processing: the current state of affairs

Today's challenge for law enforcement agencies is not only the collection of evidence, but also processing, analyzing and presenting it [9, p.6]. From our experience, most law enforcement agencies have no comprehensive approach for producing high-quality products, including factual police reports, out of multiple large, abstract data sets. As a result, agencies may go after the low-hanging fruit - the individuals that apply little security and whose identity and activities are revealed with a minimum effort - to the advantage of serious and organized criminals whom generally have invested in better security controls and whose crimes will therefore go unpunished [18, pp.97-104]. In reality, virtually all law enforcement agencies only process the *circumstantial* and *factual evidence* they need to build a case, i.e., *net evidence* (with intelligence gathering as a separate business instead of being an integral part of an evidence flow as depicted in Figure 3). Historically, all evidence of traditional investigations was net evidence: investigators put all relevant findings directly in a police report. In today's Information Age, digital investigations generate high volumes of *gross evidence* - data sets that contain links to historic, current and future crimes - yet only a relatively small preselected part of the data is processed for prosecution purposes, and as such, becomes net evidence. While there might be legal reasons why agencies do not fully index, and have access to, all gross evidence, the truth is that most agencies miss the organizational and technical means to do so. There is no common understanding of, and consensus about, a common business process, methodology, philosophy, policies and associated language to mine all the collected data. Let alone that they have the technical resources to normalize large raw data sets in unknown formats from other agencies, load those sets in their police systems, conduct advanced analyses on these sets and select suitable targets for investigations.



But there is a real need for law enforcement agencies to develop these data processing capabilities. Suspects of child sexual abuse material, for example, often collect thousands, and sometimes millions, of indecent images. For reasons of efficiency, Dutch case law prescribes that only a few representative images should be put in police reports and added to the case file, and these images become net evidence. Yet the remainder of material, i.e., gross evidence, may contain investigative leads about historic, current and future crimes against the most vulnerable among us. The absence of standardized procedures to process evidence fuels the effectiveness crisis with more unsolved crimes as a result. The current situation pushes towards the creation of a single connected ecosystem dominated by the few public nodes that do have the resources to collect, process and analyze investigative data, often helped by closed-source, proprietary technologies of a small number of private security companies [22].

In today's world of safety and security networks that evolve around information capitalism [23], we have experienced how this situation specifically affects less equipped agencies that share their evidence with third parties but are not able to process any received data sets with links to criminals and crimes in their own jurisdiction. Scarce resources are wasted because of duplicate efforts and/or expensive commercial products to mine evidential data sets, and even lead to situations in which agencies are technically unable to process their hard-earned evidence for investigative purposes.

Figure 3: CSAE builds upon the common understanding within the intelligence community about the relationship between data, information and intelligence (commonly known as the DIKI continuum [19, p.I-2][20, p.16][21, pp.70-74]), while connecting to the world of criminal investigations with its own unique legal principles, organizational structures and current big data challenges. This means that the CSAE phases follow the transformation of evidence: from data in Collect, information in Store, intelligence in Analyze to the last and additional step of facts in the Engage phase.

2.2 The Need for Data Science and Harmonization

While there is consensus among academics and legal practitioners that data scientific investigations are needed to process today's nature and volume of evidence, the transition from digital to data scientific investigations might for many law enforcement agencies be a bridge too far. For a start, a common business process for data scientific investigations is so far missing. As a result, efforts of law enforcement might go in vain, and lead to vendor lock-ins on core data science technologies. To prevent this from happening, law enforcement agencies must take mutual collaboration to a next level and strive for technical harmonization to develop core data science technologies. To do so, there needs to be a common language and understanding about, and a clear business process for, data scientific investigations.

The need for data scientific investigations

When evidence is collected, raw data have to be processed, evaluated and aggregated into information, intelligence and ultimately factual police reports about who did what beyond a reasonable doubt. In each step, the quantity should decrease while the quality should increase until net evidence - i.e., court admissible evidence - remains. Moreover, the gross evidence residue - i.e., data, information and intelligence from an investigation that is *not* used in court proceedings - is still highly valuable for law enforcement agencies. This kind of evidence contains many important leads about past, current and future crimes. A lack of standardization in investigations [10], combined with incomplete or conflicting processes and inefficient use of resources, results in overlooking or missing crucial evidence that was present in the already existing data sets, i.e., *unknown knowns*. If the gross evidence set is used more effectively, i.e., if a normalized set of parameters existed, then early indicators could be identified to initiate a new or aid a current, Investigation at a much earlier stage.

Since long, academics and practitioners alike have acknowledged that digital forensics and associated workflows - like [26, 27, 28] - are not sufficient to scale the process and analyses of these exponentially growing and highly abstract data sets and produce timely, accurate, relevant and actionable outcomes, or any results at all [10, 11, 13]. Intelligent forensics that rely on advanced analytics - e.g., artificial intelligence, natural language processing and social network analysis - have been proposed as the way forward [13][11, pp.26-27][14, pp.224-225][29] . What these 'big data' technologies bring to the investigative table is a so far missing numerical valuation of evidence. Adding mathematics, including advanced statistics, to the existing layers of traditional and digital investigations results in *data scientific investigations*. With data science, law enforcement agencies will be better equipped to deal with the ever-changing organized crime characteristics in today's Information Age, including the increased security of professional criminals. Data science allows law enforcement agencies to formulate data-informed strategic priorities as they instantly know what the broad topics are of e.g., victim complaints on any given time. Agencies will further learn what organized crime hotspots they need to pro-actively acquire a data position, discover previously undefined entities related to child sexual abuse, make probability statements whether the identified author of text A is also the so far anonymous author of text B, or predict what the impact will be on the larger criminal network when a money launderer is arrested. As a result, data science will lead to better allocation of police resources, operational efficiency and professional judgement, while creating opportunities to improve accountability of organizations, occupational health of staff and decrease physical intrusion of suspects and victims.

But we have experienced how law enforcement agencies around the world have difficulties in understanding how to legally, organizationally and technically achieve the transition to data scientific investigations. Although existing frameworks range from architecture to intelligence (see Table 1) [30], a common business process for offensive legal interventions in general is missing, let alone one for data scientific investigations. While the broader cyber security community has been working on the 4Vs of Big Data, associated models are inspirational, yet have several limitations for law enforcement agencies.

A numerical approach in criminal investigations?

As depicted in **Figure 2**, traditional statistics within law enforcement are generally a cross-over between a social and behavioral and numerical understanding of crime. Statistics have been used for both decision-making - e.g., policy- making and intelligence-led policing - and criminal investigations, especially the interpretation of forensic evidence such as DNA analysis. So besides traditional methods of *arguments* and *scenarios*, investigators have indeed also been using *probabilities* with a statistical foundation as a normative framework for evidential reasoning **[24]**. **Figure 2** further depicts an overlap between technical and numerical perspectives in law enforcement. This cross-over generally refers to situations where unsupervised machine learning (ML) algorithms and automated/autonomous decision- making (ADM) are applied, thus without the intervention - e.g., manual checks - of human beings (in other words, without influence of the social and behavioral perspective). Yet the usability of this technical-numerical cross-over for investigative purposes is questionable. Results of unsupervised learning never stand on their own, and always need to be explained by social and behavioral experts, while automated decision-making by artificial intelligence (AI) raises serious ethical concerns, and is therefore deemed undesirable in the Netherlands **[25]**.

When taking a closer look at the models that are most adjacent to offense (i.e., cyber threat intelligence models) we learn that these models are generally developed for the private security industry, and therefore differ on desired outputs (i.e., mission objectives) compared to public law enforcement agencies. More specifically, most cyber threat intelligence models are *not offensive* in nature while investigations by law enforcement agencies are per definition *offensive* in nature. Investigative powers exploit weaknesses in *modi operandi* and harm important assets related to the criminal and his/her crimes [12]. Ultimately, these vulnerabilities should be exploited with traditional, digital *and* numerical investigative methods and techniques to confront crime in both an efficient and effective manner. But because a public interest philosophy and conceptual framework for data scientific investigations by agencies with offensive capabilities are absent, law enforcement agencies risk that private security vendors exploit this knowledge gap and offer all-in-one solutions to LEAs that do not necessarily serve the public interest, nor deliver what they promise, while promoting vendor lock-ins on core technologies.

Harmonization between law enforcement agencies.

To prevent vendor lock-ins, and associated problems of over spending, high switching costs and loss of innovation, law enforcement agencies should take mutual collaboration to a next level. As depicted in Figure 4, many legal and organizational problems between agencies are already addressed by harmonizing policy agendas, laws, operations and organizations. This development occurs to the extent that different public and private agencies nowadays come together at a single physical location - and increasingly online environments as well - to investigate crime, share and analyze evidence, and/or execute a joint operation: in other words, organizational integration. However, the ultimate of national and international collaboration is *technical harmonization* between law enforcement agencies of liberal democracies governed by the rule of law [12]. Technical harmonization is about shared ethics and resources, and establishing technically uniform norms, criteria, methods and principles to process evidence by law enforcement agencies.

Continuum of Cyber Security Models

Categories	Architecture	Passive defense	Active defense	Intelligence	Offense
Description	The planning, establishing, and upkeep of systems with security in mind	Systems added to the architecture to provide reliable defense or insight against threats without consistent human interaction	The process of analysts monitoring for, responding to, and learning from adversaries internal to the network	Collecting data, exploiting it into information, and producing intelligence	*Offensive* legal countermeasures by *law enforcement agencies*, and *lawful* self-defense actions against an adversary *by others*
Associated models	National Institute of Science and Technology (NIST) 800 Series; Purdue Enterprise Reference Architecture; Payment Card Industry Data Security Standard (PCI DSS)	Defense in Depth; National Institute of Science and Technology (NIST) 800 Series; NIST Cybersecurity Framework	The Active Cyber Defense Cycle; Network Security Monitoring	The Intelligence Cycle; the Cyber Kill Chain; the Diamond Model of Intrusion Analysis; CSAE	CSAE

Table 1: According to industry expert Robert M. Lee, there are five categories of actions that contribute to cyber security [30]. Each category has its own recommended industry models and standards except Offense on which governments largely have a monopoly. We therefore add the following wordings in ***italic*** *to Lee's continuum:* ***'Offensive legal countermeasures by law enforcement agencies'*** *and* ***'lawful*** *self-defense actions against an adversary* ***by others'****. In doing so, we stress that victims and other plaintiffs can take lawful self-defense actions as well such as, but not limited to, notice-and-takedowns of malicious servers, placing decoy data on corporate networks to confuse intruders who want to steal valuable information [31], or starting civil lawsuits against criminal defendants [32,33,34]. We further regard CSAE as an associated model for the categories Intelligence and Offense.*

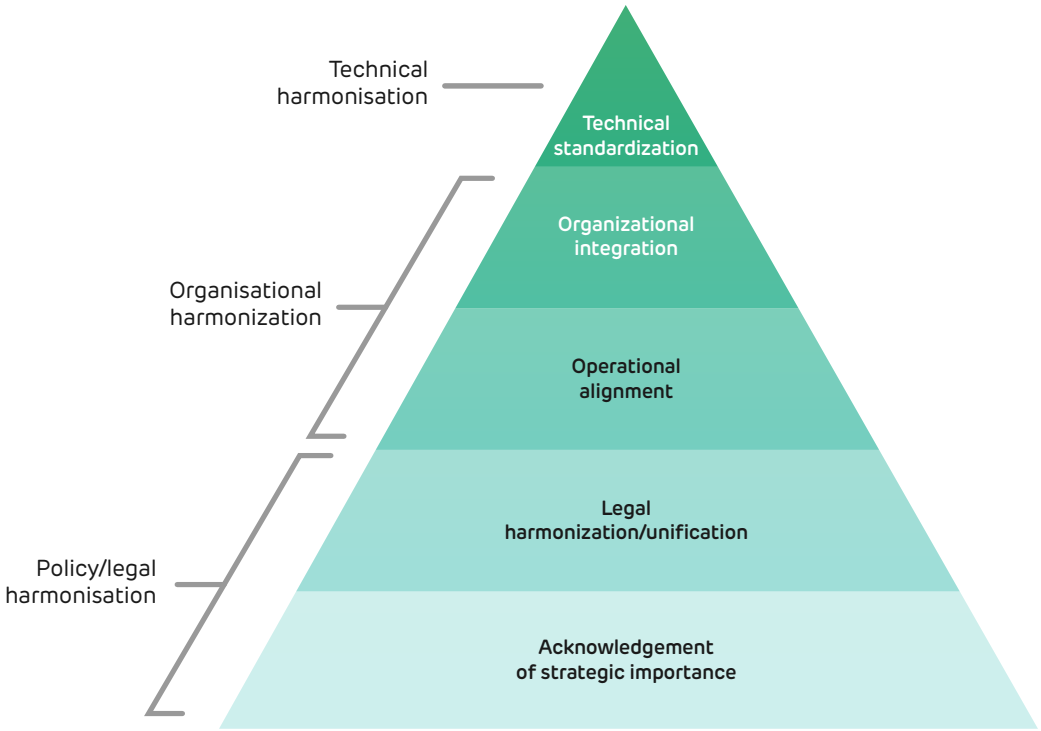


Figure 4: These five stages represent the different steps in collaboration between law enforcement agencies, and (other) public and private partners [12]. Acknowledgement of strategic importance and legal harmonization occur on a legal/ policy level, operational alignment and organizational integration occur on an organizational level of harmonization while technical standardization is an outcome of technical harmonization.

So while there is consensus among academics and industry experts that there is a need for standardization in technical processes, terminology and ontologies [14][35, pp.91-92] [36, p.52][37], the term harmonization is purposely used in this white paper. Standardization as a starting point might well be too rigid for public agencies. Legal oversight bodies, like legislators and the judiciary, need to be able to oversee and regulate technologies, including algorithms, as they see fit. Harmonization between law enforcement agencies implies alignment, fine-tuning and collaboration while respecting diversity. The term further relates to a degree of agility and flexibility which is needed for agencies in a dynamic world with rapidly changing technologies and (geo)politics. Harmonization may also lead to more rigid standardization but only when legal practitioners, including legislators and the judiciary, decide that this is needed. Moreover, law enforcement agencies not only need models that promote technical harmonization, but also legal and organizational harmonization.

In other words, the goal of harmonization must be incorporated in legal, organizational and technical policy agendas, and the CSAE business model supports, structures and aligns to this new policy objective. A practical implementation of this approach is that law enforcement agencies should initiate consortia that help them to develop their own data schemes, software and analytical models as much as possible, and subsequently share these technologies with like-minded public partners with fewer resources. This decentralized approach not only opens an opportunity to reduce evidential problems, but also to close the digital divide between law enforcement agencies as investigative resources are more equally and fairly distributed among public key players in the safety and security community. Like other work on (cyber) security models that try to achieve technical harmonization and standardization within the larger community [38, p.49], this paper is also an open invitation for academics and industry experts to strengthen and further develop the model. Lastly, we stress that harmonization between law enforcement agencies is not about binding rules or obligations to mandatory share any collected evidence with third parties.

We foresee that a data science approach will lead to significant changes in how law enforcement agencies collect data, store information, analyze intelligence and engage in lawful acts. Public debates about the reinforcement of oversight and strengthening *transparency, explainability* and *judicial review* for evidence collection, exchange and usage between key players within the security community are not new, but may include the underlying technologies, algorithms and methodological choices in the future as well [39]. To achieve the goal of technical harmonization between LEAs, and to structure future debates on data scientific investigations, a common understanding of related processes and language is needed [35, pp.91-92][36, p.52].

3 | Description and Explanation of the CSAE Model

While Figure 2 and Section 2.2 describe the three main disciplines and roles in data scientific investigations, this section starts with describing the data science methodology that is applied in all phases of the CSAE business process. After methodology, the importance of gaining and formalizing a strategic business and data understanding is explained. Finally, this section then describes the specifics of the CSAE business process. More specifically, the implementation details are given how to transform evidence - i.e., visualized as cycle in Figure 5 - from data, information, intelligence, to factual police reports, in respectively the Collect, Store, Analyze and Engage phases.

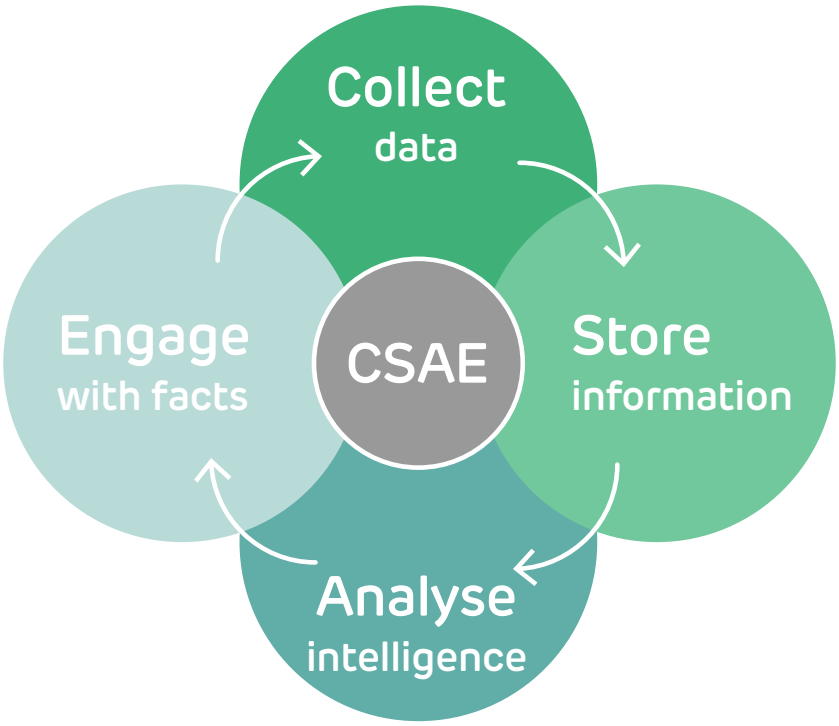


Figure 5: CSAE is a circular business process for data scientific investigations, and dovetails with existing legal principles and organizational structures of law enforcement agencies. Initially, we visualized the CSAE business process as a Rube Goldberg machine (see Appendix A). After all, law enforcement agencies are notorious for performing seemingly simple tasks in an indirect and overly complicated way. Yet we noticed how the image board of Appendix A promotes understanding and discussion among academics and practitioners alike

3.1 Internalize Data Science Methodology

From a CSAE point of view, criminal investigations and academic research bear several similarities. Both are truth-seeking processes to make transparent factual statements. Furthermore, both law enforcement officers and academics create sight - *foresight, hindsight, insight and oversight* - by de- scribing and explaining crime. A deep, sound understanding about reality is only possible with a robust *methodology*, more specifically, understanding participants and data, qualitative and quantitative methods and techniques, and research findings. So what is CSAE's data science methodology? While views on what data science encompasses sometimes differ among academics and practitioners, there is generally consensus that the field is of a multidisciplinary nature, and includes mathematics, computer science and engineering, and behavioral and social sciences. Our methodological model - called *Quadrant*, see Table 2 - brings these various numerical, social and behavioral and technical disciplines together. It stresses the need for a *mixed methods approach*, using a variety of participants and data sources from the criminal and safety and security community, to produce valid, reliable and credible outcomes. Quadrant serves two interrelated purposes. Firstly, the methodology creates strategic, tac- tical and operational sight on crime themes, appropriate reactions to these phenomena (such as, but not limited to, criminal investigations), and how the broader cat-and-mouse game between criminals and law enforcement evolves. Secondly, the methodology creates business intelligence (BI): strategic, tactical and operational sight which is needed to optimize each phase of the CSAE business process accordingly.

Running example: profit driven cyber attacks on financial institutions

This section describes the CSAE business process using a running example. The hypothetical case is based on the profit driven advanced persistent threats (APTs) of organized cyber criminals, and used for illustrative purposes only. In the case, actors have been launching campaigns against financial institution targets in Asia, Europe and the United States for several years. They send phishing emails to bank employees, install malware to control the network, and transfer money via various means such as SWIFT and personal bank accounts. The private security industry has tried to monitor and mitigate the attacks, but so far, the larger cyber security community lags behind events. Using elements from the running example, the next sections illustrate how CSAE supports data scientific investigations against these kind of organized cyber attacks.

Quadrant	Qualitative methods and techniques	Quantitative methods and techniques
Participants and data sources from the organized crime community	Interrogations of suspects; debriefings of convicts; conversations with criminal informants; review of criminal writings; observations of online criminal platforms; listening to intercepted conversations.	Topic modeling on written texts; predictive modeling; social network analyses on communications.
Participants and data sources from the safety and security community	Round table discussions with investigators; interviews with industry experts; reviews by analysts; small questionnaires; literature studies.	Large surveys among industry experts; topic modeling on victim complaints; automated IoC extraction; user statistics of (forensic) software.

Table 2: This matrix presents various examples of methods and techniques. Be aware that associated research designs do not necessarily stand on their own, but that the quantitative and qualitative methods/techniques can be combined into a mixed-methods approach.

Participants and data sets: criminal and safety and security communities

Key players of investigations should be regarded as *participants* who are either from the organized crime community or the larger safety and security community. To gain sight, LEA must be embedded in both networks. The former group consists of active and non-active offenders, more specifically high-risk individuals, criminals, suspects, convicts and former convicts, in other words, individuals who have first-hand experience with, and participated in, an organized crime community. These communities are not homogenous. Within organized cyber crime there are, for example, Portuguese (i.e., Brazilian), Chinese, English, Farsi and Russian-language communities with their own subcultural characteristics. Each of these participant groups create their own large *data sets*. These data sets consist of financial, social and technical communications, such as money laundering databases, chat groups, and/or NetFlow traffic.

The second group of participants that has experiential relevance are those who are confronted by organized crime and are part of the larger safety and security community: academics, analysts, attorneys, diplomats, investigators, judges, legislators, municipal officials, policy-makers, private security researchers, public prosecutors and (victim) witnesses. Such stakeholders all have a role in fighting organized crime. Investigations should support, learn from and be an addition for stakeholders. In cyber security, participants may work within computer emergency response teams (CERTs), corporate security of large private enterprises, cyber hotlines, dedicated cyber crime investigation units of national law enforcement agencies, internet service providers (ISPs), non-governmental organizations (NGOs) and universities [40].

Other participants are those whose assistance is sought by the cyber security community such as victims, witnesses, netizens and general Internet users: they too are confronted by organized cyber crime [41, pp.88-89]. Each of these participant groups create their own large data sets - think of cyber threat intelligence, police reports, network traffic and victim complaints - that can be used by LEA in a quantitative manner.

Mixed methods approach: qualitative and quantitative methods and techniques

How should sight be extracted from participants and associated data sets? Via a *mixed methods* approach: a combination of qualitative and quantitative methods and techniques that produce valid, reliable and credible sights. Firstly, there is qualitative research. Especially investigators and analysts are familiar with associated methods and techniques such as interviews, observation, small surveys with open questions and literature review. Qualitative research is conducted when a problem needs exploration, when theory is absent, and when a complex, detailed understanding of an issue is necessary [42, pp.39- 40]. Because the researcher *is* the instrument in qualitative research, *credibility* is of utmost importance and concerns the knowledge and experience of the researcher related to participants, methods and techniques, interpretation of results, and the field of study in general.

Evidence about the world based on observation and experience can also be expressed in a numerical fashion. Associated methods and techniques are of a quantitative nature, and include traditional statistics such as frequency counts, simple/multiple regression analyses and correlation tests, but also supervised and unsupervised artificial intelligence.

These analyses can be conducted by statisticians and mathematicians. Generally, quantitative researchers use *tools*, and therefore *validity* and *reliability* are key terms. In other words, methods and techniques and associated instruments should measure what they are suppose to measure and produce identical outcomes when the same data sets are used.

These different qualitative and quantitative research practices may stand on their own. We can conduct statistical analysis over previously seized cryptomarkets, distribute questionnaires among netizens, interview private security researchers, interrogate suspects and observe the dark web to strengthen our knowledge about other online markets. Yet the strength of Quadrant is to combine qualitative and quantitative approaches by using a variety of mixed-methods purposes, research designs and mixing strategies [43, 44]. As explained in the next section, the multidisciplinary research findings of Quadrant allows law enforcement agencies to make strategic, tactical and operational decisions about what is needed in Collect, Store, Analyze and Engage, and examples how to apply Quadrant in practice are given throughout this white paper.

3.2 Gain and Formalize Strategic Business and Data Understanding

Because investigating organized crime is complex and requires long-term dedication and resources, law enforcement agencies must first gain a strategic business and data understanding about a particular organized crime theme, including appropriate responses. In other words, agencies must identify and prioritize criminal threats and associated investigative objectives, and formalize these findings in an official program.

Gain strategic business and data understanding

Before any data sources are collected, LEAs have to gain a general business understanding on a strategic level. In other words, a *strategic business and data understanding* refers to a sound and deep understanding of a particular organized crime theme, including how that phenomenon manifests itself in associated data sets. With this in mind, law enforcement agencies should first explore the nature and extent of an organized crime theme on a macro level. Gaining a strategic business and data understanding is a constant process that generates high-abstract statements about the *who* and *what* of a particular crime theme, respectively organized crime communities and their activities. In the selected case, a distinction can be made whether IT is the focus of criminals or merely enables or assists criminals in their activities. Associated underground economies are not homogenous, global and universal market places, but differ in offered products/services, members, size, language and very much in culture. Motivations might differ as well: crime can be driven by financial, sexual, thrill-seeking or political motives. Cyber crime ranges from being very low-tech to very high-tech in nature: from simple to advanced methods and techniques for commission and protection, from local to global threats, from small handwork projects to large-scale processes, from opportunistic to targeted attacks, from loosely to closely organized, and from small to big impacts. The conclusion might be the following: profit-driven APTs are generally computer-focused crimes, executed by predominantly Russian-language organized groups that have the capabilities and resources to launch global, scalable and targeted attacks against high-value victims. Not surprisingly, each crime theme generates its own unique evidence and is therefore represented differently in data. Based on - amongst other things - the available resources, needs and objectives of stakeholders and organizational vision/mission statements, decisions have to be made *how* to engage against criminal communities and their activities, or in other words, what the organization's focus will be on a particular crime theme.

Formalize findings in a strategic roadmap Based on the strategic business and data understanding about organized crime, law enforcement agencies draw a strategic roadmap that formulates a vision, mission and strategy what to achieve, how and with who. In other words, the roadmap provides overall direction to investigation divisions with attainable goals, initiatives and criteria

Not a two-discipline show

One might think that these mixed method approaches are a two-men show between those with social and behavioral backgrounds and their colleagues with numerical backgrounds. After all, they have respectively the necessary qualitative and quantitative skills to execute Quadrant. In practice, professionals with a technical background play an equally important role. Digital investigators generally have a unique business and data understanding about technical aspects of organized crime, while data engineers and software developers contribute to Quadrant by processing the necessary data sets and developing tailor-made forensic tools.

for guidance such as indicators of progress and time-bounds. In the selected case, the focus is on profit-driven Russian-language organized crime groups that are largely autonomous in commission and protection and target high-value victims, and a few crime-as-a-service providers that support these groups. The responses are, for example, long-term reactive public-private investigations against the autonomous groups with an emphasis on damage mitigation and victim assistance, and mid-term proactive public- public investigations of malafide security providers with an emphasis on disruption [12]. Besides Quadrant, the phases of harmonization (see Figure 4) will help to make decisions about strategic partnerships to achieve the goal of technical harmonization. APTs require substantial support of the private security industry, while malafide providers generally require support of international law enforcement partners.
The roadmap must further allocate resources to implement the plans. Because the CSAE business process is the practical implementation of this roadmap, resources need to be allocated to the management of working *with* CSAE ('run the business', i.e, process management) and working *on* CSAE ('change the business', i.e., social and technical innovations that require project management) which is further explained in Section 4.2.

3.3 Obtain Data in Collect Phase

The next step is to implement the roadmap which starts with the first phase of CSAE: the collection of raw and uninterpreted observations and measurements, i.e., *data* [20, p.16]). Legislators acknowledge that starting with evidence gathering from scratch after an incident without any possibility to crossmatch found entities with previously collected data sets is very hard, and in some occasions, even impossible because of the security practices of organized crime. Therefore, LEA are generally allowed by law to store evidence of previous operations for future purposes, i.e., evidence retention. As a result, what you collected in the past, very much determines the success of future operations. The Collect phase is about promoting a process in which agencies structurally and systematically think about what data sets are needed as input to achieve their desired goals in the Engage phase. Lastly, we stress the targeted nature of Collect. In other words, quality of data sets is far more important than quantity. In practice, this means that LEA may focus on data in motion and at rest that are exclusive to the criminal community. Generally, many online hotspots where demand and supply of criminal products and services meet in the underground economy are off the beaten track to law-abiding citizens, and therefore contain no to few non-criminal members [12].

Quadrant for strategic business understanding

Quadrant helps law enforcement agencies to gain a strategic business and data understanding, and thus promotes strategic decision-making on what threats and objectives the organization should focus on. *Exploratory designs* are suitable for LEA agencies that do not yet have a thorough business and data understanding. In a typical research process, these designs start with qualitative methods and techniques that are followed by more traditional quantitative methods and techniques. For example, traditional analysts may first conduct a literature review by reading academic papers and cyber threat intelligence reports about profit-driven APTs. They might try to find answers to the following macro level questions. Which known profit-driven APTs and associated underground economies (e.g., Chinese, English, Russian) pose the largest threat to national security? Which APT MOs - e.g., ransomware, DDoS extortion, CEO-fraud - cause the most damage? Which supportive criminal services/ products - e.g., bullet-proof hosting, malware crypting, online money laundering - are vital to the value chain of these attacks? Based on these qualitative findings, traditional statisticians may quantify the number of historical victims complaints and/or investigations related to these threats. LEA agencies that are able to conduct *data-driven research* (as compared to *theory-driven* research and relying on *intuition* of law enforcement officers) may have a different starting point. They may begin with *unsupervised* quantitative approaches, thus using numerical data and advanced mathematical methods, that are followed by qualitative approaches. These agencies apply, for example, topic modeling on selected criminal conversations associated to these or similar threats. This form of unsupervised statistical machine learning identifies so far undiscovered patterns: in this case, the major themes that these career criminals discuss in conversations. Because quantitative results are descriptive and can be quite abstract, the next step is to let strategic and data analysts interpret the discovered topics via e.g., round table discussions. Lastly, strategic business and data understanding must include the views, needs and objectives of other stakeholders. After all, investigations are merely one of the instruments in the toolbox against organized crime.

Identify, prioritize and locate strategic data sources

After gaining a business and data understanding of - amongst others - different actor communities (who) and associated threats (what), the next step is to identify, prioritize and locate associated *strategic data sources*. We label the data sources as being of a strategic nature when these sets help law enforcement agencies to achieve their long-term and overall objectives. Strategic data sets are not in the possession of law enforcement agencies yet, and are a vital addition to existing *operational data* sets of past investigations that have already been processed in Store. Organized cyber criminals are highly depending on a range of legitimate and il- legitimate services and products, and tend to centralize, for a considerable time, on the same online location [12]. Some of these hotspots are indeed strategic data sources, and fill-in major investigative gaps that occur when law enforcement agencies merely work with operational evidential data sets from incident-driven investigations. LEA should acquire a lawful data position on the strategic hotspots of the larger criminal community's social, technical, financial and legal infrastructure. The investigative efforts against criminal hotspots like Alphabay, Hansa Market and Silk Road fit within this approach. From a business understanding, US and Dutch agencies understood that in order to achieve their long term objectives - i.e., prosecute major drugs vendors on online English-speaking cryptomarkets - they should first acquire a strategic data position on these marketplaces that offer a secure social, financial and technical infrastructure to professional criminals.

Law enforcement agencies may acquire strategic data sources:

- *Nationally* within a LEA's jurisdiction or internationally, i.e., from abroad; collected by either
- *Public or private organizations;* and with a
- *Reactive stance* when a strategic data source is already in possession by another organization, or a proactive stance when no party has so far collected the data set.

In the selected case, officers may identify, prioritize and locate the relevant data sets from past investigations by other law enforcement agencies. Chances are that associated strategic data sets on malafide security providers and APT groups - i.e., communications between providers and their clients - have already been collected nationally or internationally. Foreign law enforcement agencies may have, for example, investigated a bulletproof hoster and money launderer that both had a large Russian- language customer base. After all, virtually all profit-driven cyber crimes, including APTs, rely on servers and money laundering. Likewise, a thorough business and data understanding might reveal that most existing data sources of law enforcement agencies are related to suspects of APTs, but not their victims because law enforcement agencies do not have intrusion/detection systems on the net- works of potential victims, nor do victims of cyber crime easily file a complaint when a breach is not made public.

'Why is my HUMINT report regarded as (numerical) data?'

Law enforcement agencies may lawfully collect knowledge from human sources via interpersonal contact, a practice known as human intelligence gathering or HUMINT. During discussions, we frequently noticed that HUMINT officers did not see their reports as data, but as intelligence (see Section 3.5). This argument makes indeed sense from the perspective of the individual author of such a report. He/she knows that certain names in a HUMINT reports refer to locations, and certain numbers to telephone lines. Linking these information points, combined with the author's knowledge, makes the report indeed an intelligence product. But this standpoint is not correct from a technical and organizational standpoint when the report is not processed in Store. In such a case, the report is considered data as shown in the following example. Imagine that all these HUMINT report are printed out and randomly stored in a physical room (as compared to warehoused in Store). Although officers can manually search these reports, they miss - amongst other things - the metadata about who wrote the report, when and where, and for what purpose; entity linking of locations, telephone numbers and names of natural persons within the text; and links and relationships to previous reports and investigations. It now becomes clear that HUMINT reports are a data set of raw, uninterpreted observations to the vast majority of the organization. Moreover, these unstructured text documents can be processed as numerical data in the Store phase, and as such becomes measurable when advanced mathematical approaches are applied in the Analyze phase. Natural language processing (NLP), for example, is an automatic document classification technique to analyze large amounts of manually written text documents, like HUMINT reports.

In such cases, private feeds with indicators of compromise (IoCs) from specific APT-style attacks on victims might become a much-needed strategic data source (hence reactive public-private investigations against APTs). To conclude, the collection of strategic data sources must be very targeted, and is therefore indeed more about quality than quantity [9, pp.11-12, 32], and diversity of data sources than more of the same [38, p.45].

Assess resources and risks

After agencies have identified, prioritized and located the strategic data sources, the next step is to assess what is needed to collect these sources. This requires thinking about the available and required resources, and associated costs and benefits of:

- *Data formats* - Scrapping of cryptomarkets may generate HTML files, lawful intercept of chat channels PCAPs and preservation of a money laundering database SQL databases.
- *Tools and techniques* - Besides digital-forensic software to collect data in motion or at rest, hardware is needed to e.g., store strategic data sources like data carriers. Very large data sets of remote third parties require an upload portal, while live interception of evidence on a locus delicti may demand a law enforcement vehicle with the necessary equipment.
- *Expertise* - The execution of investigative powers to collect these sources - e.g., preservation, information requests, undercover - requires hard skills such as digital-forensic and tactical knowledge. Yet gathering existing strategic data sources from other public and private parties inside and outside one's jurisdiction asks for soft skills such as collaboration, corporate attitude and/or intercultural communication.

Potential risks should be discussed, including any related contingency plans. Will execution of any investigative powers notify a criminal collective or harm ongoing operations of other law enforcement agencies? Will a full but detectable collection of a criminal data set create the risk that an online hotspot returns in a better protected form as compared with a less complete yet undetectable collection method?

Collect and review strategic data sources.
The last step in going from a business to a data understanding involves the collection and review of data. This step also constitutes the overlap between the Collect and Store phases as data have to be collected, described, explored, verified, and transformed into information. In practice, this means that from a legal perspective, police reports have to be written about the collection of new evidence or the transfer of existing evidence, including opening criminal cases. Who collected what, when, where, how and why have to be documented and archived (i.e., chain of custody). From a technical perspective, the integrity of evidence has to be ensured. Because the collected data consists of raw and uninterpreted observations and measurements [20, p.16], the next step is to turn raw and uninterpreted data into information, in other words, go from Collect to the Store phase.

Quadrant in the Collect phase

The CSAE's mixed methods approach - i.e., Quadrant - helps LEA to identify, prioritize and locate strategic evidential data sources. In the running example, qualitative methods include manually observing online hotspots of Russian-language criminal communities to gain expertise in deciding what other platforms are important and what not as criminals frequently discuss competing products/services. This method is relatively straight forward, yet other approaches may require a consecutive chain of multiple analyses. A quantitative approach may begin with automatically extracting Internet domains that are mentioned in previously seized storage servers of criminal groups, or laptops and mobile devices of individual suspects. This identification process is followed by statistically weighting (i.e., prioritizing) the identified Internet domains on e.g., frequency, temporal consistency and recency. With the help of proprietary cyber security products, the domains can now be categorized into groups with labels such as Business, Government or News. The result is a categorized list with the most important domains to criminals while illegitimate Internet domains can be distinguished from legitimate domains. Law enforcement agencies may acquire an evidential position on the top illegitimate domains, while the latter top domains of legitimate companies provide guidance for public-private partnerships **(see also Section 4.3)**. After all, these companies are either a victim target, or their products/services are popular among criminals, and heavily misused for criminal purposes.

3.4 Warehouse Information in Store Phase

During the Store phase, the collected data sets are normalized, and subsequently converted into *information*: data that have been put in context and empowered with meaning, which gives it greater relevance and purpose [20, p.16]. In practice, this means that certain combinations of letters in data sets are recognized, labelled and stored as names of natural persons, certain number sequences as tele- phone numbers, and combinations of letters and numbers as bank accounts. This section first explains why a data warehouse strategy supports data scientific investigations as compared to data lakes.
The section then describes the related steps of extract-transform-load (ETL). While ETL is generally a relatively straightforward process, CSAE has a distinct public interest philosophy on Store.

Traditionally, law enforcement has been using proprietary products for Store, resulting in a multitude of tools, processes and standards of doing ETL. *Explainability* is of utmost importance to law enforcement, yet agencies face the risk that ETL processes and other algorithms become blackboxes because of commercial confidentiality [15, p.26]. Third party tools can, for example, prioritize or withhold information, without the larger criminal justice system knowing about these decisions. Criminal investigations need a flexible range of tools to process ever-changing data sets without being dependent on the services of a private monopolist or having substantial switching costs. For this reason, data schemas, ETL tools and associated processes must be regarded as *core technologies*, and be owned and managed by law enforcement agencies themselves. Our Store approach not only increases the legal explainability and chain of custody, but also the organizational maneuverability and technical agility, thus ultimately the independency, of LEAs. After all, power depends upon who owns code [45, p.534].

Data warehouse (or data lake?)

When making the transition to data scientific investigations, law enforcement agencies have a choice to make between opting for a *data warehouse or a data lake*. The latter strategy implies that all data are kept irrespective of the source and its structure. Data are thus kept in their raw, and often unstructured form, and only transformed when data are need to be used. Thus, data lakes follow the steps of extract-load-transform (ELT). While data lakes are generally associated with big data and data science, investments in the highly-structured repositories of ETL-warehouses pay off when looking at the legal, organizational and technical requirements of law enforcement agencies. Because ETL-procedures include assessing and - when needed – cleaning evidence *before* it is stored, warehouse strategies ensure data quality such as accuracy, consistency and relevancy [38, p.47]. In other words, warehouses are better equipped to respect legal principles such as the *origin* and *integrity* of evidence (e.g., chain of custody, see also Section 4.1). Structured data are also more easy to use and understand for those who work in the Analyze phase. As depicted in Figure 6, the largest group within law enforcement agencies are generally traditional investigators and analysts who focus on the human factor of crime, but have less knowledge about numerical and technical aspects of evidence. Especially during an emergency incident - think of a disruptive APT-style attack on vital critical infrastructure - they and their digital and mathematical peers have no time to get acquainted with unstructured evidence in data lakes and subsequently transform these data sets, especially when findings need to be shared as soon as possible among partners for hit/no-hit purposes. Lastly, the unique core business of law enforcement - i.e, attribution of individual suspects - becomes easier in the Analyze and Engage phases when data from different sources are transformed into a common data structure. For example, ETL-procedures transform SMS, email and chat into 'messages' that can then be used as input for advanced analyses such as authorship attribution. These arguments further link data warehouses to the objective of technical harmonization between law enforcement agencies. While developing and managing warehouses are indeed more labour-intensive than data lakes, this downside is mitigated when LEAs invest in technical harmonization, more specifically share the burden of structuring data with other agencies and use a common data scheme/ ontology (see also Sections 2.2 and 4.3).

Extract

Firstly, data files should be extracted from hardware, i.e., data carriers. Hardware and data files come in different shapes and sizes: client databases of crime-as-a-service providers, intercepted telecommunications, mailboxes of suspects, malicious software, operating systems of command and control (C&C) servers, video material, written reports of covert observations - the list is endless. Full extraction of complete data sets is not always necessary, and is in many instances even undesirable. *Partial extraction* strategies are preferable because of the legal principle of data minimization (see also Section 4.1), and data engineers should only extract those parts of data sets that contain entities that have the potential to solve historic, current and future crimes. The next step is to *manually* or *automatically* extract entities from the selected data sets. Traditionally, most police systems only support manual entity extraction, and force officers to structure their essentially unstructured text documents (such as the HUMINT reports example in Section 3.3). In practice, this means that officers not only add metadata about the text document such as the origin of the source, but also link entities and extract relations within the text. Of course, many entities are nowadays found in structured data sets such as metadata - e.g., timestamps, IPs and domain names - that are generally stored in separate database tables. As such, these entities are easy recognizable by any entity extraction tool. But entities might also be incorporated in text field tables of databases. In the selected case, one might think of text messages between the conversation between malafide providers and their clients that contain names of Internet domains, locations and/or organizations. Moreover, officers may overlook entities when manually linking entities. Therefore, automated entity extraction is needed to recognize these entities in non-structured data sets.

Transform

The next step is that the extracted entities are transformed into a uniform form. For example, suspects of crime and investigators alike may write down the same telephone number in many different ways. Therefore, the extracted entities first have to be cleaned according to a certain data scheme or ontology. Telephone numbers, for example, are mapped to country codes with a fixed sequence of digits. Furthermore, duplicate records - i.e., information points - need to be removed as well as conversing time zones to calculate timestamps. Establishing relations and adding enrichments may be considered during this phase, such as providing the telecommunication provider to the extracted telephone number (indeed, this example implies that a list of telecommunication providers may well be a strategic data source). The data schemes and ontologies for law enforcement purposes are a core technology. They are one of the foundations to achieve technical harmonization and should therefore be *closed-open source*, or in other words, open source to a closed community of trusted partners. Why are such schemes not completely open source? Investigations ensure that data schemes and ontologies constantly expand. Related to the selected case, there is an information asymmetry between the cyber criminal and cyber security community. As a result, cyber criminals go at great length to increase the amount and quality of information about the capabilities and operations of the cyber security community, and subsequently adjust their behavior to their findings [12]. Thus, any suggested expansion of a data scheme by law enforcement might well give away what their current strategic focus and associated investigations are about.

Load

After the data set is transformed into information points, the normalized and converted evidence is loaded and stored in a number of databases. These databases may have different functionalities as the technical tools over these databases answer different questions in the Analyze phase.

Some databases promote simple, fast queries while other databases allow more advanced analyses. Databases are also shaped by various legal and organizational requirements. Legal framework generally prescribe rules about removal of evidence after a number of years or linking information points from various investigations. Organizations may also have certain demands. Law enforcement agencies generally have, for example, different authority levels to access evidence. As a result, the overlap between Store and Analyze ensures that ETL processes are executed in close collaboration with, and to the benefit of, the Analyze phase. Contrary to the Store databases and Analyze tools, ETL processes including ETL tools are core technologies, and should therefore be owned and managed by law enforcement agencies. Multiple inputs from a variety of proprietary and open-source tools are normalized, converted and stored to create multiple outputs for, again, a variety of proprietary and open-source tools. This *pluggable architecture* approach creates *interoperability*, avoids a vendor lock, and therefore promotes competition and innovation on the forensic software market. An independent warehouse that has exclusive rights and control over its own ETL tools has the ability to ingest evidence *from* and *to* any public or proprietary tool. Obsolete databases and tools are easily scrapped, while new databases and tools are just as easily adopted in the Store architecture: just a single pipeline has to be build by the ETL tools to extract, transform and load evidence to the new database and tool.

Quadrant In the Store phase

One might think that *all* collected evidence must be warehoused, but nothing could be further from the truth. Quadrant helps to determine what data should be extracted out of a larger data set. Statistics over case law might reveal that images, for example, are not relevant to prove crimes of e.g., malafide service providers in court. Moreover, a round table discussion with in-house warehouse experts may show that storage of visual material absorbs too many scarce resources or the absence of software to automatically recognize and subsequently extract entities such as locations, persons and text. Whether full or partial extraction is applied, the next step is to determine what entities are of interest. In the following example, we demonstrate how a *triangulation design* is used to confirm, cross-validate or corroborate findings. A security researcher may state during an interview that career criminals use a new payment method. A simple database query confirms the existence of this newly discovered payment method in existing databases. In this example, it is important for the Store phase that the interviewer asks the security researcher what the formatting is of the associated transaction codes. Data engineers can then extract and transform transaction codes out of the warehoused data sets so that these entities become information points. Lastly, a literature review or small survey among experts may help to determine in what databases the transactions can be loaded. These transactions are now ready to be used in the next CSAE phase, i.e., for Analyze purposes. The transactions can also be used for Store monitoring and management purposes (i.e., business intelligence). Statistics may show, for example, the total number of all financial transactions or percentage of the new payment transactions compared to other payment methods.

3.5 Create Intelligence in Analyze Phase

After the strategic data sources are normalized, converted and loaded into a range of analytical tools, the transformation of information to intelligence begins. All efforts in the Analyze phase rely on, again, a mixed methods approach in which analysts and investigators may provide qualitative input for AI models but always check numeric results. Moreover, the Analyze phase consists of a continuous cycle of reduction, coordination, enrichment and investigation. Warehoused information points are reduced by - amongst others - AI models to find suitable targets. After a legal, organizational and technical assessment, the target entities are enriched with new information points until a satisfactory result is achieved. After evaluating and analyzing the results, the interrelated information points and gained knowledge about these targets has become *intelligence* [20, p.16]. The intelligence picture will lead to new investigative leads, i.e., the collection of new data and information, that in turn have to be reduced which leads to new target entities that require enrichment.

Reduce information points to targets

With hypothetically millions of information points that relate to thousands of potential targets, analysts are unable to intellectually grasp the full picture of relations between entities. Simple keyword queries will not identify organized crime or central key players in complex networks, but rather those suspects who have few resources, including knowledge assets, to protect their crimes and their identity. Besides such issues related to the nature of identified targets, manual target selection is a labour-intensive job that requires knowledge and experience. It is, in other words, not a scalable process, especially not for law enforcement agencies with high employee turnover. What is needed are intelligent strategic, tactical and operational analyses, using a variety of sources and the previously mentioned mixed methods approach. The strategic, tactical and operational reduction models further follow the narrative account of criminal investigations setting out who did what, when, where, how and why, i.e., the 5W1H of attribution [46, pp.256, 269] [47, p.308][36, pp53-54].

Strategic reduction models build on the sights of the strategic business and data understanding of Section 3.2, and are about statements about large populations on a macro level. In the selected case, these analyses focus on 'the top tier Russian-language cyber criminal underground' or 'industries with high-value victims of APTs'. An example on the 'what' are text analyses that reveal the topics that are discussed by certain populations, like bank fraud, bulletproof hosting or money laundering. These sights not only provide input to determine thematic policy priorities, but also to subsequently develop reduction models on a tactical level. *Tactical reduction models* are about statements on a meso level, i.e., organizational clusters and specific niche communities. Imagine that a strategic reduction model highlights the importance of money laundering via cryptocurrencies to Russian-language groups that extort large private enterprises. A related tactical reduction model on the 'who' is automated role identification based on the unique argot that money launderers use during criminal conversations, and subsequent prioritization of only those money laundering schemes that specialize in cryptocurrencies. A prioritized scheme can be expressed as a cluster that consist of a number of entities like nicknames, email accounts, domain names and IP addresses. Therefore, tactical reduction models provide input to reduction analyses on a micro level. *Operational reduction models* are about statements on a micro level, i.e., individuals. For instance, an operational reduction model on the 'when' are time patterns of the individuals behind the money laundering scheme. We have experienced that reduction models become more granular when components of the 5W1H are combined. For example, 'when', 'where' and 'who' could be linked in a reduction model that combines time stamps with travel movements and criminal roles within schemes.

Data science algorithms, including reduction algorithms, should not become blackboxes. They must be *fair*, and therefore - amongst others - be *explainable*, and - when necessary - made *transparent* to other legal practitioners in the criminal justice system such as the judiciary. In other words, proprietary code that leads the deployment of far-reaching investigative powers and attribution will likely not serve the public interest. Statistical reduction models - like authorship analysis or native language identification - may produce probability ratios. Generally, the more variables a model incorporates, the lower the probability scores. This fact may generate perverse incentives for the private security industry. Models with few variables will produce higher scores at face value but are actually less valid. The development of these models should rather be done by in-house data scientists and/or academics. Moreover, as compared to traditional manual investigations that create *random errors*, algorithms may generate structurally flawed outcomes. Therefore, we stress that all actionable outcomes have to be checked manually by analysts and investigators in close collaboration with those colleagues whom developed the models.

Target high-value organized crime networks and suspects

Operational intelligence packages (or: target package) can be created *proactively* and *reactively* during the preparatory investigative phase [48]. The former consists of targets identified by LEA themselves, generally based on several of the above mentioned operational reduction models. As explained, targets can also be delivered to LEA by others, thus reactively. Academics and private security researchers may provide target lists based on their own reduction models. Foreign police agencies may formally and informally point to potential targets. Victims may file a complaint and as such provide a target as well. Whether complaints come from high value victims of targeted attacks - like key players in national critical infrastructure - or in bulk files from citizens after a ransomware campaign: the notification of witnesses and victims about crime is not only an opportunity to restore justice, but also vital for avoiding *police biases* and increasing the business understanding of the underground economy.

An operational intelligence package describes and explains the modus operandi of a specific target. More specifically, the target package should provide a comprehensive socio-technical overview of the nature and extent of a subject's 5W1H, and the interrelation of three components that give LEA jurisdiction to start an investigation:

- *Offenders, including their capabilities*, who are located in points of attack origin (or: source country);
- *Technical, financial and legal infrastructure*, located in points of attack linkage (or: transit country);
- *Victims* who are located in points of attack occurrence (or: destination country).

Based on this socio-technical overview of the target's MO, the identified acts should then be legally categorized as criminal offenses. Substantive law criminalizes certain behavior and prescribes when law enforcement agencies have jurisdiction. A good target package presents reasonable grounds for suspicion that crimes are committed. The target package should also outline *short-term investigative opportunities*, like the location of a malicious server, and *long-term investigative* objectives during the Engage phase, e.g., seizure of the financial infrastructure and prosecution of the main suspects. To conclude, the preparatory investigative phase in Analyze produces timely, accurate, relevant and actionable socio-technical-legal outcomes, i.e., operational intelligence packages.

Coordinate legal, organizational and technical issues

The target package that was compiled in the preparatory phase needs an internal review. Internal checks and balances for that provide accountability, auditability and scrutiny have traditionally already been incorporated in all workflows of law enforcement agencies. Yet our business process recommends a built-in moment for review and decision-making - called Coordination - where the preparatory investigative phase ends and investigations begin. Coordination means a technical, legal and organizational assessment of the operational intelligence package. Law and technology set hard requirements with organizational effects. Simultaneously, law enforcement agencies also have their own organizational policies and mechanisms.

The overall review should include an inventory of technical, legal and organizational resources, and associated opportunities and challenges. The legal assessment includes a review of what information is needed to open a case, what the status is of each individual piece of information, where the information comes from and who the owner is (i.e., chain of custody), and what authority levels are to view and share information, and which jurisdictions are affected. Based on the long-term objectives, the required investigative powers should be reviewed. This relates closely to the technical assessment. Does the agency in charge have the resources to execute technical investigative powers like lawful intercept? What kind of data sources will presumably be collected, and how will these sources be processed in the Store phase and reviewed in the Analyze phase? The answers to these technical and legal questions contribute to the organizational assessment that decides who should execute the target package. This could be another law enforcement agency, but also an intelligence or security service, watchdog or CERT. These agencies may have regional, state, federal or international jurisdiction, be located in another country, or solely provide analyses and coordination like Europol or INTERPOL. Other options include public-private consortia that consist of academics, civil servants or private security researchers who provide unique capabilities, data or services. Thus, Coordination in the Analyze phase very much steers the tactical and operational efforts of LEA yet require a degree of agility as well. Because criminal networks are dynamic and many details are still unknown to law enforcement during the Analyze phase, investigations rarely work out as planned.

Enrich target package

After the coordinate phase, target entities are enriched with *internal* and *external sources* to create a criminal cluster on a micro level. A target package may consist of entities like domain names, cryptocurrency wallets, email addresses, monikers and IP-addresses. In this phase, we would like to discover related entities to the ones that came out of the reduction models or were provided by third parties. In the running example, an Internet domain is related to a botnet of an APT. In this case, we have just a single entity: a domain name. Based on external, open Internet sources, we can conclude that the domain name was registered with a particular email address. This email address was also used to register other, so far unknown historical domain names. These historical domain names with time stamps are then queried in the warehoused internal sources, and are linked to two Russian-speaking individuals that shared the domain name in their communications. Native language identification shows that one individual is likely a Ukrainian-speaking individual. Open source cyber threat intelligence reports of the private security industry show that the two associated user accounts are involved in email spamming and specialize in targeting large business enterprises in the Anglo-American world. What we learn from this simplified example is how enrichment relates to the 5W1H - who: two individuals, what: spamming, when: based on timestamps, why: financial motives, how: botnet, and where: based on IP addresses of the domain names, location of victims and language-usage by suspects.

We further learn that important external data sources - like open and closed Internet sources - are *not* necessarily strategic data sources that need to be obtained in Collect and subsequently ingested in Store. External sources can also be queried during the Analyze and Engage phases, either automatically via APIs of public and/or private databases, or manually by interviewing human sources or searching open and closed Internet sources. A single email address may result in five domain names, each domain name may have multiple associated IP addresses, these IP addresses may be associated to multiple hosting providers and/or clients, and so on. Moreover, technical harmonization ensures that analytical charts can be shared with (inter)national law enforcement partners who enrich entities with their internal data sets and/or external data sets they have access to, and subsequently return their findings. Ultimately, enriching entities by using both internal and external sources means that findings may grow exponentially, and hopefully increase the number of investigative opportunities.

Quadrant In the Analyze phase: the Hyperion method

Traditional analysts are generally great qualitative researchers. Based on their knowledge and experience, they manually reduce large amounts of data with targeted database queries, and create network charts during the enrichment phase of Analyze. These network charts are visualizations of interrelated entities called *clusters*, and provide an oversight of suspects, (mis)used technical, legal and financial infrastructure, and/or victims of organized crime. Analysts further write intelligence reports about their findings, i.e., their knowledge about the targeted cluster such as descriptions, explanations and/or predictions. How can analysts ensure that their intelligence products - i.e., the interrelated information points of, and knowledge about, the cluster - are preserved, become input for new analyses, and are proactively presented to a broader law enforcement audience?
The answer to these questions is Hyperion: an analytical model developed by Dutch academics and police analysts **[49]**. The method prescribes that analysts, after finishing their product, add a 'cluster entity' to the chart. Besides a description of the cluster, the new entity holds several properties based on a limited number of options that provide answers about 'scenes' that consist of:

- **Social-cultural qualities (who): which individuals belong to the cluster and what is their interrelation;**
- **Business qualities (what): what is the core business of the cluster, what crimes are committed, and/or criminal market is served; and**
- **Temporal/spatial qualities (when/where): what timelines and online/offline locations are associated with the cluster.**

From a technical perspective, Hyperion is about *annotation*. The cluster itself becomes a new information point with properties and attributes - i.e., descriptive metadata - that needs to be warehoused in a (separate intelligence) database within Store. From the CSAE's standpoint on data science methodology, Hyperion is a mixing strategy - i.e., *data transformation* - as it provides a taxonomic scoring system in which qualitative data (i.e., knowledge of traditional analysts about clusters) are numerically coded, while quantitative data are transformed into narrative about criminal clusters. As such, Hyperion has multiple advantages to create intelligence in the Analyze phase. Firstly, Hyperion ensures that qualitative sights are preserved to agencies irrespective of employee turnover, and are proactively made available to other analysts while duplicate efforts are avoided. When an analyst who works on a criminal cluster stumbles upon an entity that has already been linked to a previously identified cluster, the Hyperion entity will appear with the full chart and description of the latter cluster. Secondly, Hyperion allows to make statements about organized crime on multiple levels. Individual entities relate to statements on an operational level, clusters on a tactical level, and scenes on a strategic level. It may become apparent how previously distinct social-cultural and business scenes come together. In the selected case, bulletproof hosters located in the Netherlands may collaborate with Russian-speaking autonomous groups via English-language cryptomarkets. Thirdly, Hyperion promotes the mixing strategy of *typology development*. The qualitative research of analysts may yield a typology on money laundering clusters. While specific variables of a typology are not necessarily part of the taxonomic scoring system, these variables are useful for automated cluster detection. Manual review of these automatically detected clusters will result in the discovery of new variables that strengthen the typology. Lastly, Hyperion allows social network analyses (SNA) to calculate the most central and/or dense cluster within the larger network. In other words, Hyperion supports target selection. After all, SNA answers the question what cluster must be targeted to have the biggest impact on the larger criminal underground.

Conduct investigative reasoning and cycle

Enrichment brings new relevant entities to light that will raise multiple investigation questions. Therefore, the enriched cluster in general and the new entities specifically are input for *argumentative, probabilistic* and/or *scenario* (also known as *narrative*) *reasoning* based on *theory-driven* and/or *data-driven hypotheses*. The former method of reasoning is to provide arguments and counterarguments that are potentially presented in court. Narrative methods consider the construction and comparison of scenarios of what may have happened, while probabilistic methods show the connections between the probability of hypothetical events and the evidence. Entities may point to new forensic leads, and as a result, investigative powers might be deployed to collect new evidence and confirm, revise or reject the hypotheses. In the running example, an IP address of a server may be discovered during the enrichment phase. Is the server malicious, and what are the counterarguments that it is not? If so, will preservation contribute to the investigation or are other, more proportional and/or less obtrusive interventions an option? Related metadata may be requested from an ISP such as financial and subscriber details of the server. Based on the results of such an information request, a natural person who paid the server may become of interest. In what scenario fit these new findings? The preserved data on the server contains details that partially match previously preserved C&C servers. What is the probability that these servers are part of the same APT? Ideally, data scientific investigations apply all three methods in an integrated manner. Scenarios explain who did what, while arguments are used to support or attack these scenarios with evidence. Arguments to and from scenarios can subsequently be placed in the context of probability. An argument can have a strength, measured by probability, which expresses a degree of uncertainty [28].

It further becomes apparent that the Analyze phase is a continuous cycle of reduction, target selection, coordination, enrichment and investigative reasoning that also involves Collect and Store when investigative powers generate new evidence. Enrichment generally leads to the collection of new data sets. These data sets have to be stored, normalized and converted into information in Store. The new and existing information points have to be reduced again to narrow down the investigation. New hypotheses are formulated and tested, and based on these results, investigative reasoning and available resources, decisions have to be made about which entities are added or removed from the investigation. In the running example, investigations of organized crime facilitators will generally also include data about their partners and clients.

Because these individuals and their crimes could reach thousands, choices have to be made who or what to investigate. Reduction models may help to identify the most suitable targets based on e.g., the nature and extent of the crimes these clients commit, or where the clients are located. Slowly, the investigation takes shape as conclusions are established beyond a reasonable doubt. We acknowledge that in practice, investigations must take advantage of opportunities as they arise. That notion not only makes the investigative cycle highly dynamic and constantly subjected to change, but the investigation objectives as well. After all, the cycle works towards, and overlaps with, the next and last phase of Engage.

Evaluate process

Process evaluations determine if theories on CSAE and related implementation are successfully followed or need to be updated. These evaluations are needed to increase the efficiency of data scientific investigations. The distinction between *theory* and *implementation failures* is pivotal. What both failures have in common is that the results are not expected. The difference is that the latter failure relates to poor implementation practices such as a low number of collected strategic evidence sources due to a lack of trained staff. Theory failures occur when processes are correctly implemented, yet expected results are not found because the theory behind the processes and intelligent models are incorrect. Applied methods and techniques may not be valid or reliable, and therefore produce *measurement errors*. As previously described, traditional investigative methods and techniques are predominantly of a qualitative nature. In such a case, investigators and analysts themselves *are* the instrument, and therefore errors are generally random and therefore largely unavoidable such as individual mistakes, failures and accidents that have an effect on expected outcomes. But with the shift to data scientific investigations that use algorithms, *systematic errors* may occur. Because these errors are consistent and repeatable, process evaluations are vital in data scientific investigations and must apply to the full business process: the input of Collect, the activities of Store and Analyze, and the output of Engage.

3.6 Execute Lawful Interventions in Engage Phase

When the investigation cycle of the Analyze phase is conducted to a sufficient level, operations work towards the execution of their objectives. Traditionally, investigations are deployed for attribution only - namely: who did what for prosecution purposes - and has few possible outcomes that go beyond punitive sentencing. Because points of attack origin, linkage and occurrence may not be in the same jurisdiction, or even be located in safe havens, bringing suspects to justice can take several years, and is in some instances even impossible. Therefore, LEA must formulate additional outcomes and related outputs that go beyond attribution. Indeed, attribution for prosecution purposes is inextricably linked to *repression of crime*. Opposite to repression is *prevention of crime*. However, prevention and repression of crime do not form a dichotomy, but are rather two ends of the same continuum. CSAE formulates four outcomes and related outputs that go from prevention to repression of organized crime. The central idea behind this model is that law enforcement agencies should affect the intangible and tangible assets of criminals by deploying their monopoly on offensive investigative powers, exploiting their unique data position and serving a wide audience of stakeholders, including citizens.

Damage mitigation

The first and most preventive outcome is *damage mitigation*. Police investigations may generate data about organized crimes that are still in the preparation and pre-activity phase. Moreover, data may also be available about ongoing crimes in the activity or post-activity phase that are about to target other industries or jurisdictions in the imminent future. In both situations, there is a need for key players in the safety and security community to be aware about these threats. Mitigation is focused on helping these stakeholders to increase their security against threats (including misuse of their services/ products) before they turn into successful attacks. As such, harm will be limited or even prevented. In other words, damage mitigation aims at generating actionable threat intelligence for a range of public and private actors, and as a consequence, reduce the effectiveness and conversion rate of future attacks on potential victims. Related outputs and audiences include warnings in media outlets for the general public, threat analyses for private industries, and actionable sight for potentially vulnerable individuals, groups and organizations about upcoming attacks.

Victim assistance

The next objective is a key task of many law enforcement agencies: give help to those who are in need of help, i.e., *victim assistance*. This particular outcome does not focus on potential victims and threats, but actual victims and past and ongoing attacks. In such a situation, victims are detected, identified and subsequently notified about their victimization and MOs. Help is being offered to stop the attacks and/or further damage to citizens, but also property such as systems, software and personal- and company data. In the selected case, LEA may discover victim information in the system of a preserved C&C server. These victims might not be aware that they are victimized, and notifying them via CERTs may help to limit further damage. In turn, victims may file a complaint and/or provide additional data about the attacks. (Stolen) company and personal data on such servers are not only assets to their lawful owners, but also to criminals. The two victim-based approaches - damage mitigation and victim assistance - mostly affect criminal assets related to crime (as compared to the criminal). The role of the police in this approach is to support the safety and security of the general public, public institutions and the corporate sector.

Quadrant In the Engage phase

Quadrant helps to determine what effective interventions are against organized crime. Members of the private (security) community may indicate in online surveys what they need from law enforcement to successfully mitigate threats, and/or increase the costs of committing and protecting crime. In the selected case, calculations of revenues and operating costs, using financial transactions from a seized customer relationship management (CRM) database of a bulletproof hoster, may reveal that margins are very small, and that an increase in transaction or operating costs maybe a viable pressure point to disrupt revenue and demand [50]. Interrogations of cooperating suspects may contain questions about how to identify and exploit so far unidentified financial weaknesses in MOs. The joint review of above mentioned results to create new or consolidated variables and data sets is called data consolidation/merging. The consolidated variables and data sets of this mixing strategy are then usable for purposes of further research on, for example, the financial aspects of restrictive deterrence (i.e., limiting the frequency, magnitude or seriousness of offenses).

Disruption of criminal business processes

Research shows that investigative powers are always disruptive to MOs, even when not put in action because even the mere threat of investigations affect the criminal business process [12]. So, after the previous two victim-based approaches comes the first offender-based approach of *disruption*. In other words, hindering the processes of committing crime and protecting crime and the criminal. Outputs and interventions should go beyond existing interventions that aim at skimming profits of the commission of crime such as seizing financial assets like hard cash or cryptocurrencies. Disruption may not only consist of increasing the commission costs [51], but also the total costs related to protection such as taking down vital criminal servers or adding false positives into the criminal process. In such cases, professional criminals will either over spend on security or under protect assets. Other disruptive interventions may focus on bringing either collective underground economies or individual MOs to a suboptimal level by respectively promoting market failures via increase of information asymmetries [52], or disturbing the balance between commission and protection via e.g., targeted messages to individual suspects to act in a particular way or omit certain behavior.

Attribution

The last and most repressive goal of investigations is *attribution* for prosecution purposes and alternative sanctions. An important remark is that prevention of potential offenders as a primary goal of investigations has no place on the continuum. Investigations differ from supervision of compliance, and exclusively deal with suspects and violations of substantive law as compared to behavior of high-risk but still law-abiding groups. Indeed, repressive punishments may hold general and specific deterrent effects. They send respectively an important message to the general public, and help to avoid reoffending by the suspect and further damage inflicted to victims. Nevertheless, alternative sentences - e.g., community services, financial transactions, fines, official warnings and probation - for low-threat offenders such as young offenders hold punitive effects, and are therefore primarily forms of repressive punishments. Yet these sanctions are less far-reaching than incapacitation, with more emphasis on other goals of punishments such as deterrence, rehabilitation, restitution and restorative justice.

These last two offender-based approaches harm assets related to the criminal and his crimes. The latter is directed at the physical subject of the criminal in case of an arrest, while disruption damages criminal tangibles (i.e., seized cash money) and intangible assets (i.e., lowering of the status and/or reputation of a member in the criminal community).

Evaluate impact

Ultimately, CSAE is about increasing the effectiveness of investigations. Periodic and objective *impact evaluations* are necessary to measure the immediate effects of specific Engage interventions against the threats as described in the roadmap of Section 3.2. Impact evaluations - as compared to the process evaluations of Section 3.5 - are about the causal relationship between the intervention and the outcome of interest (i.e., *causality*). Indeed, impact evaluations are in close relation with the process evaluations as described in Section 3.5: when implementation of CSAE is (in part) a failure, it is difficult to find out about the effects of data scientific investigations. The outcomes of impact evaluations support evidence-based policy, and provide input to improve the quality and effectiveness of the CSAE roadmap. Impact evaluations also help to understand which interventions are most cost-effective in a given situation, and how to allocate scarce resources more efficiently. Research questions related to the selected case include: does the arrest of few members of a specific criminal organization stop their cyber attacks on financial institutions? Are these members replaced by others, and/or does the group resume their activities in time? Similarly, does automated IoC sharing increase the security of stakeholders, and what feedback do these public and private partners give to LEA? These specific cause-and-effect questions draw on - amongst others - mixed-method approaches which, again, underlines the importance for law enforcement agencies to become familiar with Quadrant. A last remark is that impact evaluations fit within the public interest philosophy of CSAE as these evaluations are a means to demonstrate results of data scientific investigations and to be more accountable as law enforcement agencies to core constituencies of liberal democracies.

The snowball effect of a multi-stakeholder approach with multiple outcomes

In practice, these multiple outcomes do not only stand on their own, but are also communicating vessels. During an investigation of a profit-driven advanced persistent threat that was similar to the selected case, law enforcement officers and private security researchers wrote a joined threat analysis for financial institutes about the observed attack based on preserved C&C servers. The report contained many IoCs and was distributed by the Dutch GovCERT to other national GovCERTs that subsequently sent the report to the associated members of their national financial Information Sharing and Analysis Centres (FI-ISACs). This allowed banks to take necessary countermeasures against the threat. Based on the enclosed IoCs and TTP, some financial institutes discovered that they were not potential but actual victims, and requested the assistance of the private security company that wrote the report. Other actual victims were discovered through analysis of the incoming bots of preserved C&C servers, and were notified by their national CERT. Identifying the victims and helping to remove the infections before financial harm was done does not only assist victims but is also disruptive: the organized crime group made upfront investment costs to infect and monitor the machines, but did not yield any profits yet. A suspect was in sight during the execution of these interventions, yet this did not harm the attribution process. This observation might be a pattern: many organized criminals do not necessarily stop their criminal activities after LEA interventions. During another operation against ransomware, the Dutch police and a private security company offered decryption keys from preserved C&C servers to victims for free. Yet the suspects continued their activities even after they keys were published. They were forced to alter their MO to a suboptimal level in which new mistakes, failures and vulnerabilities occurred that provided further evidence against them.

4 | CSAE as a Model for Public Policy

For most law enforcement agencies, the transition to data scientific investigations is nothing less than a paradigm shift, and requires major organizational, technical and legal reforms that can only be achieved with the support of long term policy agendas. Because research shows that non-technical factors are the biggest obstacle to technical innovation projects within law enforcement agencies [53], this section describes how legal/administrative-political, organizational/human relations and international relations/public-private partnerships policy agendas may foster data scientific operations. CSAE may also reform policy-making itself: besides serving traditional top-down policy-making formulated by the executive branch, our framework allows *bottom-up, data-driven policy-making*, and helps agencies to give meaning to abstract legal terms like 'privacy' and policy rhetoric such as 'public-private partnerships'. Besides creating the right conditions, policy agendas must also ensure that data scientific operations align with, and when possible, strengthen the foundations of liberal democracies such as inclusiveness, privacy and sovereignty.

4.1 Legal and Administrative-Political Agendas

We experienced how the lack of a common comprehensive approach for data scientific investigations does not improve the quality of internal and public legal debates. Too often, professionals struggle to formulate the corresponding legal questions on a sufficiently detailed level, and may therefore end up with incorrect answers to their concerns. We therefore explain why data science methodologies is grounded in existing and new legal principles. We further show how CSAE fine-tunes the ongoing debate about the hard to capture legal meaning of the term privacy, and how data science may become a means to promote privacy in criminal investigations.

Existing and new legal principles for data scientific investigations

Whether data scientific investigations should be subjected to new legal principles is a matter of academic, legislative and public debate. Yet we noticed the following line of reasoning during internal legal discussions about the design and usage of data science methodologies in criminal investigations. Firstly, data science methodologies must align with the *fundamental human rights* that are cornerstones of liberal democracies.

In other words, research designs - selecting data, developing and executing methods and techniques, interpreting outcomes - must follow the lines of e.g., checks, balances and fairness, and avoid abuse of power, arbitrary decisions and discrimination. Secondly, data science methodologies should be placed in existing empirical and normative frameworks for evidential reasoning. The empirical framework refers to principles of 'good' research such as validity, reliability and credibility. The normative framework refers to transferring empirical findings to a legal environment. These frameworks go hand in hand. After all, the normative methods of arguments and scenarios are associated with empirical qualitative research and *non-probabilistic evidence*, while probabilities are associated with empirical quantitative research and *probabilistic evidence* [54]. Similar to traditional statistics, intelligent forensics that use, for example, machine learning are means of probabilistic reasoning about uncertainty of investigative conclusions. This means that the principles of probabilistic evidence apply as well, such as *likelihood ratios* and avoidance of *false causality*, *prosecutor's fallacy* or *survivorship bias*.

At the same time, intelligent forensics are a game-changer that may require reconsidering current accountability and legal scrutiny mechanisms. For example, algorithms may produce systematic errors, i.e., inaccurate results. To avoid such errors, legal scholars recommend to avoid automatic unsupervised decision-making, and introduction of the new legal principle of *human-in-command* [55]. The latter principle prescribes that all data science models require human interaction such as manual checks of results. Accuracy of evidence relates to the principle of *integrity* and its associated attribute of *accountability*. Current 'traditional' accountability mechanisms within law enforcement agencies focus predominantly on the integrity of staff, digital forensic software and data, but less on advanced statistical and mathematical methods and techniques. So besides explainability and transparency, legal scholars have also pled for *algorithmic accountability* as part of legal legal scrutiny [39, 56, 57]. At the same time, data science can also create opportunities to increase accountability by revealing previously unseen patterns of how suspects, victims and witnesses are treated and decisions are made [15, p.28].

Collect and privacy: data minimization

Collect is *de jure* and *de facto* targeted, and about quality of evidence rather than quantity. Law enforcement agencies are not only regulated by laws like penal and procedural codes during the collection of evidence, but also by the labour-intensive process of criminal investigations that requires considerable administrative, physical and technical resources. Investigators must prove the criminal nature of strategic data sources and obey the legal procedures around targeted scraping, lawful intercept and preservation. Still, strategic data sources are in today's Information Age rather large. The legal principle of data minimization is relevant to promote privacy in the Collect phase as this principle prescribes that investigative powers should collect as few data as possible. Indeed, the current academic debate focuses whether data minimization in the Collect phase is still possible [39], especially because organized criminals themselves apply data maximization. After all, they outsource parts of commission and protection to relatively few centralized, malafide providers who have automated business processes, large historical client databases and keep law-abiding outsiders at a distance [12].

Store and privacy: data retention and protection

Whatever legal means for collection are used, once possessed by law enforcement, data will have a status according to the relevant data protection acts. Legal issues in Store evolve around requirements like *data retention and protection* that are generally set forward in data protection acts, civil law and industry standards. These laws and regulations may mention technical computer security requirements such as access control to prevent unauthorized access to data and associated legal terminology such as due diligence and due care. Again, the principle of data minimization is relevant as this principle prescribes that data related to certain data subjects must be deleted after a certain period of time (i.e., data retention). A topic for discussion is, for example, how data science can help to select entities that should remain to be stored for an extended period, or deselect entities relatedto e.g., victims and witnesses that should be deleted because of data retention laws.

Analyze and privacy: purpose limitation

Where Store meets Analyze, the issue of legal access - i.e., authorization - to evidence by law enforcement officers and third parties via rule- and role-based access control is of importance to achieve privacy. Associated safeguards are generally found in data protection acts, but procedural laws may also play a role, for instance, when evidence is transferred from one investigation to another. Moreover, the principle of *purpose limitation* is relevant to privacy in the Analyze as the goal of processing evidence for analyses purposes generally has to be similar as the initial purpose of collection. Data science models on unobtrusive camera observations may help to record only suspicious activities or blur individuals who are not related to the crime under investigation so that these events or persons cannot be analyzed for different purposes in time.

Engage and privacy: disclosure

Penal codes and procedural laws are relevant when writing police reports. After all, law enforcement actions must have a legal base, and indeed, requires paperwork. This includes proving intent of crimes of suspects based on penal codes, and the deployment of interventions as set in procedural laws. Privacy in Engage relates to the legal principle of *disclosure* of evidence to others, either being data subjects (e.g., suspects, witnesses, actual and potential victims), public law enforcement agencies (including CERTs), and third parties (e.g., public and private security researchers). Although associated legal requirements on disclosure – e.g., chain of custody – are generally set forward in data protection acts, certain rights of data subjects are codified in procedural laws when it concerns ongoing investigations. Secure multi-party computation (also known as privacy-preserving computation) is important to disclose evidence as, for example, an IoC feed to national critical infrastructure (NCI). Data science may refine the jointly computed functions over inputs from investigations (e.g., all evidential data sets that contain IoCs) and NCI (e.g., only few specific IoCs related to a certain threat) while keeping those inputs private.

4.2 Organizational and Human Relations Agendas

The shift to data science has been nothing less than a management revolution in the private sector [58], and law enforcement will be no exception when it comes to decision-making for business intelligence purposes or operations against organized crime. Leadership can measure, and therefore manage, what and how much is collected, stored, analyzed and engaged, and why. They can further make *data-informed decisions* based on both empirical findings *and* intuition, and execute more effective interventions. Yet the transition from traditional and digital to data scientific investigations will require hands-on - and in some cases hands-off - leadership to steer organizations and staff.

Organizations

We frequently experienced how even managers and officers of the same agency end up with many misunderstandings because of the absence of a common business process for investigations. CSAE provides a single common language for law enforcement officers within the same agency, but also between agencies (of other countries), and will therefore foster collaboration between them. We distinguish two governance models within CSAE. Firstly, there is the management of those who work *with* CSAE ('run the business'), and the second is the management of those who work on CSAE ('change the business').

Working with CSAE means managing a *process*, and allows operational lower level (i.e., first-line) management to draw *workflows* within each phase on a more detailed level. The deliverables of each phase should be the start of the workflow of the next phase. As such, the full workflow allows to assess how work - i.e., the transformation of evidence as depicted in Figure 3 - flows through the CSAE business process, moving from person to person and from task to task, as part of a broader look at how to improve operations. CSAE allows middle management to formulate *key performance indicators*, and steer on quantitative and qualitative accountability in each phase. What and how many data sets were collected in Collect, and why? What and how many information points were warehoused in Store, and why? What and how many analytical models were developed and subsequent targets generated during Analyze, and why? What and how many interventions were executed during Engage, and why? Lastly, strategic top (i.e., upper) management may use CSAE to identity opportunities for harmonization between LEAs and/or cross-functional collaboration, and receive input for data-driven decision-making. Such a strategic task force may not only consist of the heads of e.g., investigation divisions (for Engage and organizational issues), but also less obvious, but nowadays equally important heads of IT (for Store and technical matters) and legal counsel (for regulatory issues).

Because organized crime constantly evolves, any governance model for working with CSAE should be able to identify and address the innovations that the process needs. Therefore, working on CSAE is also necessary as new data science models and tools have to be developed for Collect, Store, Analyze and Engage. Working on CSAE requires the management of *projects*, and related governance models should be able to - amongst others - prioritize projects, assess their legal, technical and organizational requirements, allocate resources, manage and execute the projects, and ultimately deliver products that are incorporated in the workflow of those who work with CSAE.

Lastly, we noticed that working with and on CSAE requires different forms of governance that compete with the current traditional, top-down hierarchal structures of law enforcement agencies. In other words, CSAE is not merely about technical innovation, but goes hand in hand with *social innovation*. The characteristics of today's crime and related responses, including data scientific investigations, means that day-to-day strategic and operational decision-making becomes increasingly complex. *Collective leadership* promotes the input of different views, shared decision-making and group accountability, creativity and involvement.

While all team members should identify innovative needs, committees that consist of individuals who 'want' to contribute and have hands-on experience and various backgrounds may, for example, prioritize projects based on a set of pre-defined criteria. When projects deliver products/services that are integrated into a workflow, an individual needs to be responsible and accountable for management and support. In other words, *ownership* is important as well.

People

Attribution is indeed a team sport that requires more skills and resources than any single mind can offer [9, p.7]. What kind of team is required to work with the CSAE business process? Firstly, CSAE helps HR departments to understand the number and nature of job families and required skill sets of current and future employees. As depicted in Figure 6, CSAE requires a continuum of staff that ranges from a few *technology-enabling specialists* in Store and ends with a high number of *business-impacting generalists* in Collect and Engage (the latter two phases rely on the same procedural laws that are executed by generally the same kind of operators, see also Section 4.3). Because CSAE provides oversight into the products, services, methodologies and outcomes of each phase, HR departments can easily put standards on required hard skills and competencies for current and future employees and explain to them what they will do and where they fit in the process. Yet soft skills and related competencies such as collaboration, discipline, punctual and resiliency should not be overlooked [59, pp.65-66]. Of specific importance is the soft skill of communicating data science methodology - i.e., data, methods and techniques and findings - to a non-technical audience of legal practitioners within the criminal justice system and general public [9]. For example, officers need to be able to communicate with those who either provide their work input or who receive their work output. In other words, investigators need to understand the disciplines of those who work with them on the 'left and right side' of the business process. There is also a need for data science-savvy functional managers who have the ability to understand and communicate with a range of backgrounds, disciplines and skills, and ask questions about methodology and related limitations. Such managers may allow themselves to be overruled by the data [58, p.8], but keep relying on intuition for decision-making. After all, attribution is not only science but also an art, and will always require the intuition of experienced managers and operators [9, p.7].

These insights and oversights in staff, skills and competencies will help HR departments to deter- mine which trainings are needed. Because knowledge, methodologies and tools are largely developed in-house, trainings as well as sprints will be organized by experienced staff. This approach not only saves scarce financial resources but also collects input - e.g., user feedback - from peers, while minimizing the 'not invented here' syndrome and maximizing cross-functional cooperation at the same time. We experienced how active outreach to the public academic sector and staff who work on cross-functional cooperation create an atmosphere in which continuous knowledge transfer occurs between peers. In other words, the structured approach of CSAE increases the maturity level of staff in a variety of disciplines, and achieve *career agility*: individuals with a self-reflective, iterative career path who are able to respond to change, optimize creativity and have a growth mindset [60].

Lastly, we stress the need for *cognitive diversity* and *social inclusion* to avoid biases in e.g., choosing data sets, applying methods and techniques and interpreting research outcomes, and to promote better decision-making via collective leadership. This is indeed an example that *social innovations* are inextricably linked to technical innovations, and as a result, leads to more efficient and effective investigations. data scientific investigations do not only require diversity in critical thinking by those with social and behavioral expertise whom collect evidence via interactions with suspects such as interrogations, eavesdropping or observations, and/or subsequently qualitatively analyze - i.e., explain and interpret - evidence.

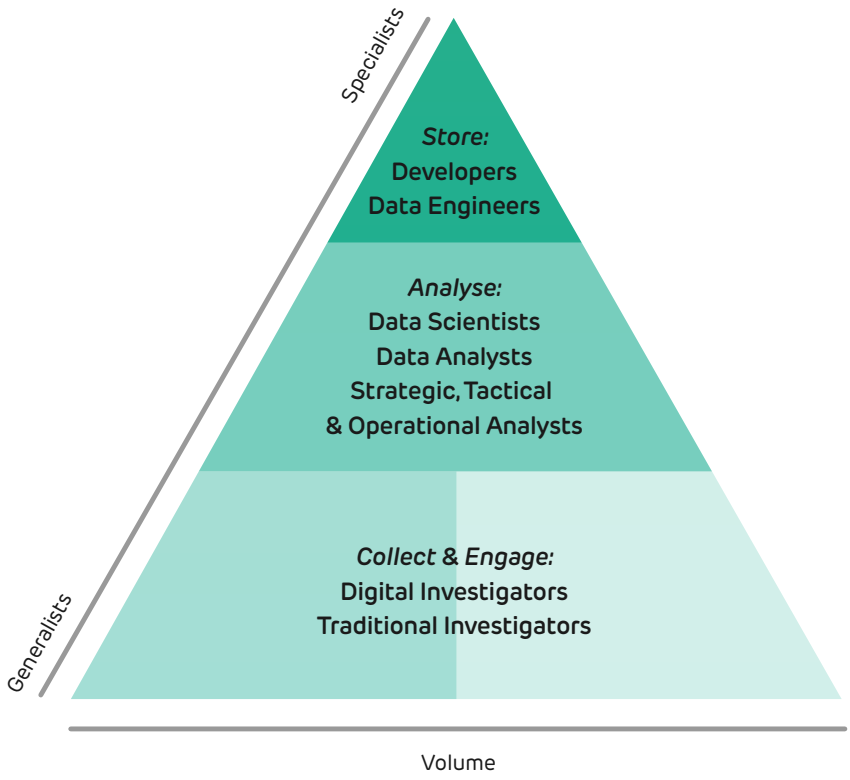


Figure 6: Store needs relatively few developers and data engineers, Analyze is conducted by an ascending number of data scientists, data analysts and strategic, tactical and operational analysts, while Collect and Engage are executed by a large pool of digital and traditional investigators, including financial investigators, HUMINT operators and OSINT specialists. The overlaps in Figures 2 and Appendix A show that data scientific investigations occur in respectively the professionals' cross-discipline and cross-phase collaboration

On the contrary, diversity is also needed within the group of professionals with technical and numerical backgrounds. After all, digital investigators and data engineers make decisions what evidence sets are extracted related to an ever-changing range of suspects and crimes, while developers, data scientists and data analysts consciously and unconsciously put their views and values into software code and analytical models that target human beings and their behavior. Besides investing in ethics for data scientific investigations [61], we believe that there is no better way to achieve diversity in critical thinking when law enforcement agencies ensure that they are socially inclusive organizations with people of all abilities, ages, ethnicities, gender, sexual orientations and walks of life.

4.3 International Relations and Public-Private Partnerships

Harmonization is achieved via collaboration with external partners, i.e., international relations and public-private partnerships. CSAE helps LEA to identify *what* they need in a particular phase, and subsequently determine *who* can help out. Thus any collaboration - whether on an international/national and private/public level - should serve and improve the business process of data scientific investigations.

International Relations

CSAE helps to structure international relations. In practice, we experienced how our agencies collaborated with an ascending order of intensity and descending number of potential international partners on either i) solely Collect/Engage, ii) Analyze and Collect/Engage, or iii) all CSAE phases (again, Collect and Engage are merged because of legal and organizational reasons: both phases rely on the same procedural laws that are executed by generally the same kind of operators). Most agencies currently work with international partners solely on Collect/Engage. With all its consequences, they basically receive input and share output with their partners via organizational harmonization, more specifically operational alignment and, in few instances, via organizational integration. Yet we noticed that the step towards technical harmonization - i.e., working on Collect, Store, Analyze *and* Engage - is a major one. Besides additional resources, the number of potential international partners is limited. Some agencies do not have the resources to develop software and data science models, or to frequently travel and participate in international sprints. The pool of potential partners is also limited because of geopolitics and fundamental human rights.

Operational sprints and Organized Crime Field Labs to learn CSAE and achieve technical harmonization

We noticed that a major challenge is not so much to explain the theory of CSAE, but to implement and scale the framework to a large user base in practice. One of the organizational tools to gain hands-on experience are *sprints* on each of the four phases on a national and international level. During these sprints, colleagues from various law enforcement departments and backgrounds solve a specific problem in a time slot. Sprints may focus on working with CSAE - think of generating targets packages on a specific crime - as well as working on CSAE such as developing a new data science model or analytical tool. In our experience, positive side-effects are that agencies can draw upon each other resources, while individual participants are able to network and learn about other specializations. A structural and more systematic approach to learn CSAE and achieve technical harmonization in practice are *Organized Crime Field Labs*. These specially designed environments include a structured but flexible problem-solving space, an inclusive facilitative process and a custom-made accountability structure that support collaborative design processes **[62]**. We expect that such environments for experimenting with, learning about and innovating in collaborative governance are very suitable to learn CSAE while working towards shared objectives and taking actions against real organized crime problems.

After all, software code and ontologies show what data sources law enforcement agencies have in their possession, what investigative breakthroughs are realized and what innovations they are working on. Moreover, data science must not only respect standards of existing laws, but also support the underlying values of liberal secular democracies that are governed by the rule of law. We therefore stress that international partnerships are not necessarily about the development of technical harmonization and/or data scientific investigations. In many instances, collaboration on Collect/Engage and Analyze may be beneficial as well, especially as this may work as a stepping-stone towards technical harmonization and data scientific investigations. When agencies miss a numerical investigative approach, they frequently have the resources to interpret statistical outcomes and manually enrich entities because of their unique cultural knowledge about a specific underground economy.

Public-Private and Public-Public Partnerships

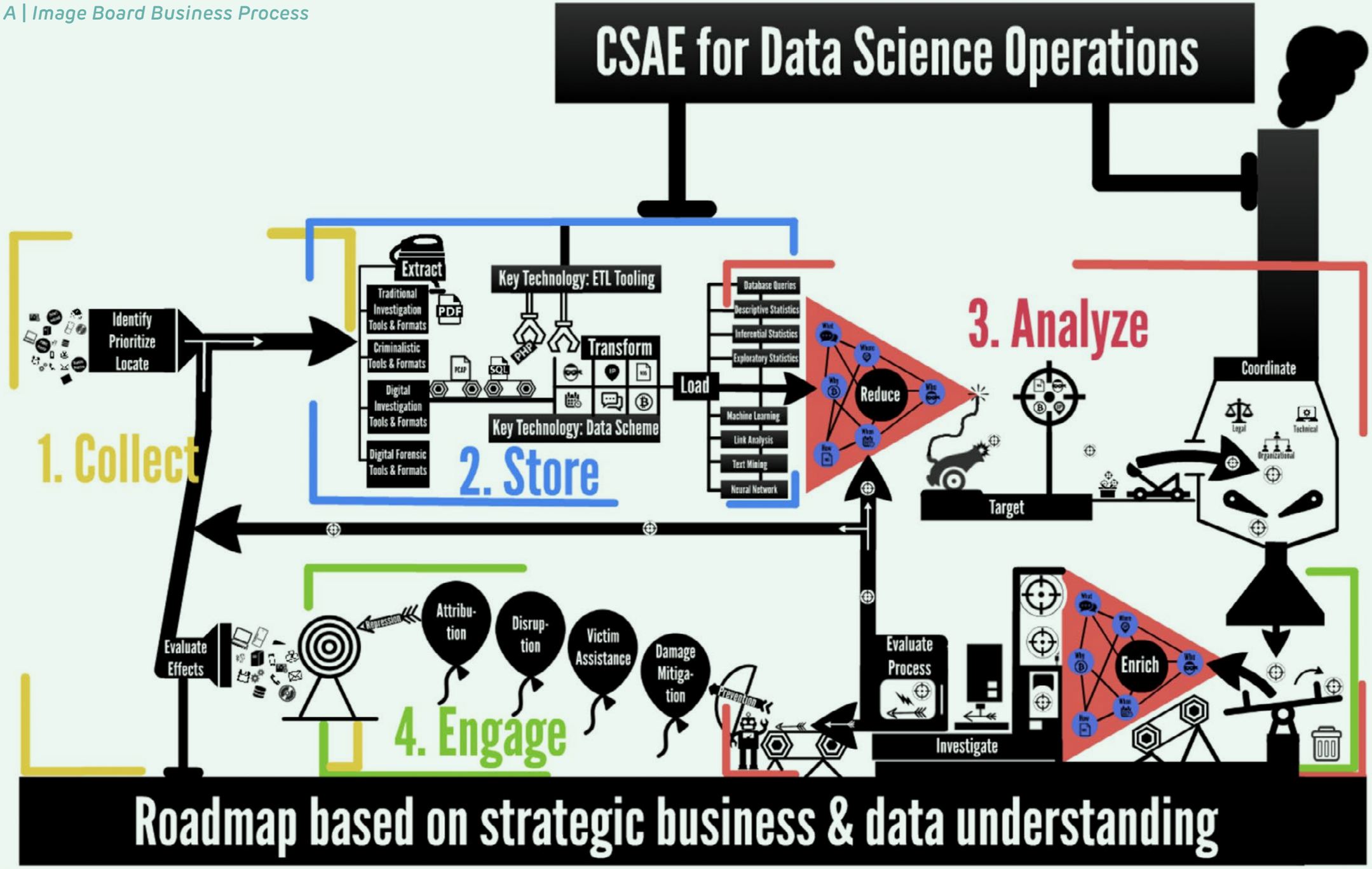
As explained in Section 1, the intelligent software that is developed by public and private consortia is generally not adopted by law enforcement agencies. We believe that CSAE might not only professionalize the development and adoption of such free forensic tools, but also the collaboration with private and public partners in general as we noticed how current partnerships suffer from the absence of a common business process, methodology and public interest philosophy.

In practice, this means that PPPs should solely serve the public interest in general, and the needs of the CSAE business process specifically. In the running example, public and private partners may find each other in the goal of protecting the integrity of the global financial system against profit-driven APTs. The next question is subsequently how partnerships can contribute to either Collect, Store, Analyze and/or Engage. After all, each phase has its own specific legal, organizational and technical challenges that agencies cannot solve alone.

In Collect, we need to know what APTs the private and public sector determine as a threat, and what they have identified as associated strategic data sources. Public and private organizations might have unique data sets such as malware feeds that will help agencies to achieve their Engage objectives. In Store, partnerships may help agencies to build conversions, ontologies and pipelines to normalize ingested data sets. As- sociated warehouse tools may also include data science models to filter noise related to, for example non-suspects, victims and witnesses. In Analyze, potential victims - e.g., financial institutions - may help to build reduction models that serve damage mitigation and victim assistance. Partners could also directly provide target packages on APT groups, while application programming interfaces (APIs) of open and closed data bases from the private security industry will help LEAs to enrich target entities. Lastly, public and private partners can help LEAs in the Engage phase. These actions may range from writing and/or distributing cyber threat intelligence reports about APTs and identification of actual victims to takedowns of infrastructure and attribution of suspects.

REPHRAIN
Protecting citizens online

Appendix A | Image Board Business Process



Appendix B | Related Data Mining and Intelligence Standards

Collect	Store	Analyze	Engage
Intelligence Cycle			
CRISP-DM			
		Kill Chain; Diamond: ATT&CK	
	STIX / TAXII		MISP
		Q Model	

Table 3: The reviewed data mining and (cyber threat) intelligence frameworks provide helpful lessons for the CSAE phases of Collect, Store, Analyze and Engage.

This appendix describes how CSAE is built on the Intelligence Cycle and Cross Industry Standard Pro- cess for Data Mining (CRISP-DM): two business processes that are a well-accepted within respectively the law enforcement community and corporate sector. Because the previously mentioned 4V-problem has been ever-present in cyber security community [14, pp.224-225], six taxonomies, shared standards and ontologies within cyber threat intelligence are reviewed that operate on a lower layer of abstraction than the CRISP-DM and Intelligence Cycle business processes. Notwithstanding that none of the reviewed models forms a complete description of a business process for data scientific investigations (see Table 3), that is not to say that the CSAE model cannot learn from these frameworks. On the contrary, our model learns from, and incorporates parts of, these state-of-art industry standards, while simplifying the business process for data scientific investigations into four clear phases.

Intelligence Cycle: Collect, Store and Analyze

The closed feedback loop of the Intelligence Cycle (also known as the Intelligence Process) consists of six steps, i.e., i) planning and direction, ii) collection, iii) processing and exploitation, iv) analysis and production, and v) dissemination and integration, while a sixth component - feedback and evaluation - is applicable in all phases [19, p.I-6]. The end- product of the cycle are assessments and reports with inherent uncertainties and levels of confidence, summarizing the analyzed information for decision makers in military and intelligence agencies.

While the steps of the Intelligence Cycle are incorporated into the CSAE business process, our model also differs from the Cycle on several points. CSAE is specifically tailored to the needs of *law enforcement* with its own distinctive objectives, legal principles and organizational structures (as compared to military and intelligence agencies). This means, amongst others, that CSAE does not stop at the Analyze phase that equals steps iv and v of the Intelligence Cycle. While intelligence products are indeed important for decision-making in law enforcement, most notably in intelligence-led policing [21], such products must also link in a smooth manner to an additional step in law enforcement - i.e., the operational objectives of the Engage phase - where officers write factual police reports that are the basis of lawful actions against crime and hold statements about reality beyond a reasonable doubt. To stimulate that smooth transition between phases and the collaboration between departments of law enforcement agencies, all CSAE phases overlap with the next phase. Those in charge of Collect have a shared responsibility with the Store department that raw data become information, just like those in charge of Store have to agree with the intelligence (i.e., analysis) department that the tools in which information is loaded are suitable for reduction and enrichment purposes. Lastly, the Intelligence Cycle does not take today's 4Vs as a starting point, and therefore its process is not tailored to, nor explains, the usage of data science methods and techniques for strategic and operational purposes. We therefore review the well-established data science standard of CRISP-DM.

CRISP-DM: Collect and Store

CRISP-DM is a methodology and process model to guide data science efforts and consists of six phases: i) business understanding, ii) data understanding, iii) data preparation, iv) modeling, v) evaluation and vi) deployment [37, 63]. CSAE learns from all of these steps in the following manner. While both the Intelligence Cycle and CRISP-DM stress the need for continuous feedback and evaluation, CSAE distinguishes *process evaluation* from *impact evaluation* as both types of evaluation are necessary to improve the business process as well as the impact on crime. The fourth and sixth step - i.e., modeling and deployment - are relevant to respectively Analyze and Engage of CSAE, descriptions of these CRISP-DM phases are rather limited (which actually points towards the need for a clear vision on data science methodology and developing analytical models). Therefore, CSAE has primarily adopted CRISP-DM components that relate to Collect and Store, more specifically business and data understanding, and data preparation. The strength of CRISP-DM - i.e., its independence of both the industry sector and the technology used - is also its limitation when it comes to its usability for law enforcement purposes. One of the alterations to make CRISP- DM components more suitable for law enforcement purposes is that CSAE distinguishes strategic from operational business and data understanding. The former refers to a sound and deep comprehension of a particular organized crime theme on a macro level, including how crime phenomena are manifested in associated data sets, and associated strategic law enforcement objectives. Operational business and data understanding is similar to the description of CRISP-DM, and linked to the Collect phase of CSAE. This kind of understanding refers to grasping crime dynamics on a meso and micro level, i.e., crime characteristics on respectively a community and individual level. The necessary addition of strategic business and data understanding is needed because the crime business differs from e- commerce that CRISP-DM initially focused on. In other words, criminals are for a number of reasons not your regular e-customers that visit a webshop (i.e., traffic). Responses to organized crime differ as well, and have little to do with search engine optimization and bounce/conversion rates. We therefore review several Analyze models for (cyber) crime in the next paragraph. These models are developed by the private sector, yet public policy priorities are highly dynamic and change on a regular basis, and different guiding principles and appraisal criteria are at play as compared to the private sector (see e.g., [64]). Accuracy is for the public sector, for example, more important than efficiency. Because LEA have offensive capabilities - i.e., investigative powers to breach the security of suspects to collect evidence - the Collect and Engage phases will differ from defenders in the safety and security community. Moreover, the responsibilities that come with these powers - i.e., effectively fighting organized crime, while protecting fundamental human rights - point towards higher, broader and more complex public interests that law enforcement have to serve in each phase as compared to private interests. Therefore, examples that underline the public interest philosophy of CSAE are given throughout this paper.

Kill Chain, Diamond and ATT&CK: Analyze

There are several taxonomies that describe and structure cyber criminal MOs. Kill Chain describes the various phases of advanced cyber attacks and provides a taxonomy to analyze and confront these threats [65]. The Diamond Model of Intrusion Analysis integrates and complements the phased approach of Kill Chain by broadening the technical and socio-political perspective and complex relationships between adversaries and their capabilities, technical attack infrastructure and victims [66]. ATT&CK is a knowledge base of adversary tactics, techniques, and procedures (TTPs) based on real-world observations by private and public security researchers [67], and can be seen as a more in-depth and more actionable iteration of the Kill Chain. CSAE's data science methodology learns from the quantitative research principles of these models - i.e., validity and reliability - while providing accuracy, effectiveness and efficiency. Yet the associated analytical models are developed for defensive countermeasures against cyber attacks, while CSAE promotes to look at the broader underground economy, including various providers of crime-as-a- service and their clients, during their investigations. This means that investigators must not only interested in different actors, but also different phases of a variety of MOs. For example, Kill Chain and Diamond do not mention the preparations, pre-activities and post-activities and related services to commit crime, nor is there a focus on the protection of crime and the criminal. Police investigators are, for example, interested to know which malafide hoster or money launderer are used by professional criminals. Therefore, not only the used technical infrastructure and social relations of adversaries have to be mapped, but also, when relevant, the financial and legal infrastructure. In other words, analysis models for law enforcement purposes should produce entity relations of complete MOs. This includes entities related to all commission and protection practices of suspects, their technical, financial and legal infrastructure, and their victims, from preparation and pre-activity phase to the activity and post-activity phase.

Lastly, the value of these analytical frameworks is their alignment to standardized Store data ontologies as the next paragraph shows. Because of the related challenge to model knowledge of law enforcement officers, related degree of trust and uncertainty about their findings and subsequently express it in an ontology [35, p.98], we share such a method - called Hyperion - in Section 3.5.

STIX/TAXII and MISP: Store and Engage

The Standardizing Cyber Threat Intelligence Information with the (STIX) is a standardized language to represent structured information about cyber threats. It has been developed so it can be shared, stored, and otherwise used in a consistent manner that facilitates automation and human assisted analysis [68]. Trusted Automated eXchange of Indicator Information (TAXII) is a collection of services and message exchanges to enable the sharing of information about cyber threats across product, service and organizational boundaries. It is a trans- port vehicle for STIX structured threat information and key enabler to widespread exchange [69]. The strength of STIX/TAXII is that it promotes standardized, structured data representations while using the previously described Kill Chain model (in CSAE terms: aligning Store to the Analyze phase). Having the capability to integrate with TAXII, the Malware Information Sharing Platform (MISP) subsequently provides the design and implementation of a collaborative threat intelligence sharing platform [70]. Sharing intelligence to others is regarded as Engage because it creates an output that allows potential victims to increase their security against future attacks. The connection between STIX/TAXII, Kill Chain and MISP show, on a higher level of abstraction, that confronting crime is a process, and associated phases should not be separate, isolated silos in law enforcement agencies, but need to connect to one another, and ideally have overlap. Therefore, CSAE model promotes a business process that flows horizontally through vertically organized agencies. The overlap, as depicted in Figures 5 and Appendix A, ensures that departments within law enforcement agencies have a shared responsibility where the output of one phase becomes the input of the next phase, thus from Collect to Store, from Store to Analyze, from Analyze to Engage, and from Engage back to Collect as CSAE is a circular process model. Lastly, CSAE appreciates STIX' standpoint not to lose human judgement and control, and the assurance that information is not only machine-parsable but also human-readable [68, pp.6, 12]. We therefore review another framework - Q Model - that addresses the human factor of crime and investigations.

Q Model: Analyze and Engage

A limitation of the previously mentioned Store and Analyze models is the predominantly technical perspective and audience, and limited focus on the core business of investigations: attribution. The Q Model explains, guides and improves the art and science of attribution of cyber attacks on a strategic, tactical and operational tactical level in the Analyze and Engage phase [9]. More specifically, Q Model is build around applying qualitative methods and techniques to answer six fundamental questions to investigations - who, what, when, where, why and how, also known as 5W1H - and stresses the importance of policy themes like public relations and communications. CSAE adopts this approach in its Analyze and Engage phases. For example, CSAE makes 5W1H a central element of the Analyze phase, and aligns analytical models that are based on 5W1H with the core objective of law enforcement agencies in the Engage phase: attribution of who did what for prosecution purposes. The qualitative methodology of Q Model is further integrated in CSAE's data science methodology. While many new innovations may lie in (advanced) statistical tools, the Q model stresses that investigations are ultimately conducted by human beings. CSAE combines the qualitative research designs, with traditional and advanced quantitative methods an techniques. This combination is CSAE's mixed-methods data science approach that aims at producing credible, valid and reliable statements about reality that go beyond reasonable doubt [71, p.350][43, pp.233-234]. Mixed-methods have the potential to respect legal principles: algorithms that may affect suspects, witnesses and victims should not be blackboxes to legal practitioners, nor should (semi-)automated decision-making be possible without human checks and balances. Statistical outputs should be examined by human beings, and/or backed by evidence that is retrieved from qualitative methods and techniques such as interrogations, interviews, observations and/or small surveys.

REPHRAIN
Protecting citizens online

Nomenclature

4V	*Variety, velocity, veracity and volume*
5W1H	*Who, what, when, where, why and how*
ADM	*Automated or autonomous decision-making*
AI	*Artificial intelligence*
API	*Application programming interface*
APT	*Advanced persistent threat*
BI	*Business intelligence*
CC	*Command and control*
CERT	*Computer emergency response team*
CIA	*Confidentiality, integrity and availability*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CRM	*Customer relationship management*
CSAM	*Child sexual abuse material*
CWAE	*Collection Warehouse Analysis Engagement*
ELT	*Extract-load-transform*
ETL	*Extract-transform-load*
FI-ISAC	*Financial Information Sharing and Analysis Centre*
HR	*Human relations*
HUMINT	*Human intelligence*
IoC	*Indicator of compromise*
IR	*International relations*
ISP	*Internet service provider*
IT	*Information technologies*
LEA	*Law enforcement agency*
MISP	*Malware Information Sharing Platform*
ML	*Machine learning*
MO	*Modus operandi*
NCI	*National critical infrastructure*
NGO	*Non-governmental organization*
NIST	*National Institute of Science and Technology*
NLP	*Natural language processing*
PCI DSS	*Payment Card Industry Data Security Standard*
PPP	*Public-private partnerships*
SNA	*Social network analyses*
STIX	*Structured Threat Information eXpression*
TAXII	*Trusted Automated eXchange of Indicator Information*
TTP	*Tactics, techniques and procedures*
US	*United States of America*

REPHRAIN
Protecting citizens online

References

[1] M. Pollitt, "A history of digital forensics," in IFIP Advances in Information and Communication Technology, vol. 337 AICT. Springer, Berlin, Heidelberg, 2010, pp. 3–15.

[2] M. I. Pramanik, R. Y. Lau, W. T. Yue, Y. Ye, and C. Li, "Big data analytics for security and criminal investigations," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 4, p. e1208, jul 2017. [Online]. Available: http://doi.wiley.com/10.1002/widm.1208

[3] M. Nouh, J. R. Nurse, H. Webb, and M. Goldsmith, "Cybercrime Investigators are Users Too! Understanding the Socio-Technical Challenges Faced by Law Enforcement," in Proceedings of the 2019 Workshop on Usable Security (USEC) at Network and Distributed System Security Symposium (NDSS). Internet Society, feb 2019. [Online]. Available: http://arxiv.org/abs/1902.06961

[4] V. S. Harichandran, F. Breitinger, I. Baggili, and A. Marrington, "A cyber forensics needs analysis survey: Revisiting the domain's needs a decade later," Computers and Security, vol. 57, pp. 1–13, mar 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0167404815001595

[5] N. Kop, "Van opsporing naar criminaliteitsbeheersing. Vijf strategische implicaties." Boom Lemma uitgevers, Den Haag, Tech. Rep., 2012.

[6] Tweede Kamer der Staten-Generaal, "Naar een effectieve en toekomstbestendige opsporing. Een eerste voortgangsnota Juni 2016 (bijlage bij 29628,nr.643)," Vergaderjaar 2015-2016, 2016. [Online]. Available: https://zoek.officielebekendmakingen.nl/blg-774894.pdf

[7] Ministerie van Veiligheid en Justitie, "Herijkingsnota. Herijking realisatie van de nationale politie," 2015.

[8] Tweede Kamer der Staten-Generaal, "Contouren voor een effectieve, toekomstbestendige opsporing (bijlage bij 29628,nr.593)," Vergaderjaar 2015-2016, 2015.

[9] T. Rid and B. Buchanan, "Attributing Cyber Attacks," Journal of Strategic Studies, vol. 38, no. 1-2, pp. 4–37, jan 2015. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01402390.2014.977382

[10] S. L. Garfinkel, "Digital forensics research: The next 10 years," Digital Investigation, vol. 8, no. 7, pp. 64–73, 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/ S1742287610000368

[11] N. Beebe, "Digital forensic research: The good, the bad and the unaddressed," in Advances in digital forensics V. Fifth IFIP WG 11.9 International Conference on Digital Forensics, G. Peterson and S. Shenoi, Eds. Orlando, FL: Springer, 2009, pp. 17–36. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-04155-6{_}2

[12] E. H. A. Van de Sandt, The Deviant Security Practices of Cyber Crime. Leiden: Brill Academic Publishers, 2021. [Online]. Available: https://brill.com/view/title/60184

[13] A. Irons and H. Lallie, "Digital Forensics to Intelligent Forensics," Future Internet, vol. 6, no. 3, pp. 584–596, sep 2014. [Online]. Available: http://www.mdpi.com/1999-5903/6/3/584

[14] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," Computers & Security, vol. 72, pp. 212–233, jan 2018. [Online]. Available: https://www-sciencedirect-com.bris.idm.oclc.org/science/article/pii/S0167404817301839

[15] Centre for Data Ethics and Innovation, "AI Barometer Report," Centre for Data Ethics and Innovation, Tech. Rep., 2020. [On- line]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf

[16] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," Tech. Rep., 2004. [Online]. Available: https://dl.acm.org/citation.cfm?id=1026492

[17] K. D. Haggerty, Richard V. Ericson and R. V. Ericson, "The surveillant assemblage," British Journal of Sociology, vol. 51, no. 4, pp. 605–622, dec 2000. [Online]. Available: http://doi.wiley.com/10.1080/00071310020015280

[18] W. Landman, R. Kouwenhoven, and M. Brussen, "Kijk naar het systeem. Begrijpen en beïnvloeden van opsporingspraktijken," Politie en Wetenschap, Tech. Rep., 2020. [Online]. Available: https://www.politieenwetenschap.nl/publicatie/politiewetenschap/2020/kijk-naar-het-systeem-348/

[19] U.S. Joint Chiefs of Staff, "Joint Intelligence. Joint Publication 2-0," U.S. Department of Defense, Tech. Rep., 2013. [Online]. Available: https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf

[20] Organization for Security and Co-operation in Europe, OSCE Guidebook Intelligence-Led Policing. Vienna, Austria: Organization for Security and Co-operation in Europe, 2017. [Online]. Available: https://www.osce.org/chairmanship/327476

[21] J. H. Ratcliffe, Intelligence-led policing, 2nd ed. London & New York: Routledge Press, 2016.

[22] R. Anderson, "Privacy versus government surveillance: where network effects meet public choice," in 13th Annual Workshop on the Economics of Information Security, State College, PA, 2014. [Online]. Available: http://weis2014.econinfosec.org/papers/Anderson-WEIS2014.pdf

[23] D. S. Wall and M. Williams, "Policing diversity in the digital age: Maintaining order in virtual communities," Criminology and Criminal Justice, vol. 7, no. 4, pp. 391–415, nov 2007. [Online]. Available: http://crj.sagepub.com/cgi/doi/10.1177/1748895807082064

[24] B. Verheij, F. Bex, S. T. Timmer, C. S. Vlek, J.-J. C. Meyer, S. Renooij, and H. Prakken, "Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning," Law, Probability and Risk, vol. 15, no. 1, pp. 35–70, mar 2016.

[25] Tweede Kamer der Staten-Generaal, "Beantwoording schriftelijke vragen AI bij de politie," 2020. [Online]. Available: https://www.rijksoverheid.nl/documenten/kamerstukken/2020/02/18/tk-beantwoording-schriftelijke-vragen-ai-bij-de-politie

[26] P. Hunton, "The growing phenomenon of crime and the internet: A cybercrime execution and analysis model," Computer Law & Security Review, vol. 25, no. 6, pp. 528–535, nov 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026736490900154X

[27] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response (NIST Special Publication 800-86)," National Institute of Standards and Technology, Gaithersburg, Tech. Rep., 2006. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-86/final

[28] R. Rowlingson, "A Ten Step Process for Forensic Readiness," International Journal of Digital Evidence, vol. 2, no. 3, 2004. [Online]. Available: https://www.utica.edu/academic/institutes/ecii/publications/articles/A0B13342-B4E0-1F6A-156F501C49CF5F51.pdf

[29] D. Quick and K. K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," Digital Investigation, vol. 11, no. 4, pp. 273–294, dec 2014.

[30] R. M. Lee, "The Sliding Scale of Cyber Security," SANS, Tech. Rep., 2015.

[31] D. Bradbury, "FBI Document-Seeding Tactics Echo Decades-Old Hacker-Hunting Trick - Infosecurity Magazine," jan 2020. [Online]. Available: https://www.infosecurity-magazine.com/infosec/fbi-documentseeding-tactics/

[32] B. Fung, "How Microsoft killed off a massive botnet, with trademark law," jul 2013. [Online]. Available: https://www.washingtonpost.com/news/wonk/wp/2013/07/24/how-microsoft-killed-off-a-massive-botnet-with-trademark-law/?utm{_}term=.cf8e40260d07

[33] J. Meisner, "Microsoft Names Defendants in Zeus Botnets Case; Provides New Evidence to FBI," jul 2012. [Online]. Available: https://blogs.microsoft.com/blog/2012/07/02/microsoft-names-defendants-in-zeus-botnets-case-provides-new-evidence-to-fbi/

[34] R. D. Boscovich, "Microsoft Reaches Settlement with Piatti, dotFREE Group in Kelihos Case," oct 2011. [Online]. Available: https://blogs.microsoft.com/blog/2011/10/26/microsoft-reaches-settlement-with-piatti-dotfree-group-in-kelihos-case/

[35] V. Mavroeidis and S. Bromander, "Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence," in 2017 European Intelligence and Security Informatics Conference (EISIC). Athens, Greece: IEEE, sep 2017, pp. 91–98. [Online]. Available: http://ieeexplore.ieee.org/document/8240774/

[36] E. W. Burger, M. D. Goodman, P. Kampanakis, and K. A. Zhu, "Taxonomy Model for Cyber Threat Intelligence Information Exchange Technologies," in Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security - WISCS '14. New York, New York, USA: ACM Press, 2014, pp. 51–60. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2663876.2663883

[37] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000, pp. 29–39. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133

[38] S. Brown, J. Gommers, and O. Serrano, "From Cyber Security Information Sharing to Threat Management," in Proceedings of the 2nd ACM Workshop on Information Sharing and Collabora- tive Security - WISCS '15, 2015.

[39] D. Broeders, E. Schrijvers, and E. Hirsch Ballin, "Big Data and Security Policies: Serving Security, Protecting Freedom," 2017.

[40] R. Broadhurst, "Developments in the global law enforcement of cyber-crime," Policing: An Inter- national Journal of Police Strategies & Management, vol. 29, no. 3, pp. 408–433, 2006.

[41] D. S. Wall, "The Internet as a Conduit for Criminal Activity," in Information Technology and the Criminal Justice System, A. Pattavina, Ed. Thousand Oaks, CA: Sage Publications, Inc., 2015, pp. 77–98.

[42] J. W. Creswell, Qualitative Inquiry & Research Design. Choosing Among Five Approaches., 2nd ed. Sage, 2007.

[43] ——, Research design. Qualitative, Quantitative, and mixed methods approaches, 2nd ed. SAGE Publications, Inc., 2009.

[44] V. J. Caracelli and J. C. Greene, "Data Analysis Strategies for Mixed-Method Evaluation Designs," Educational Evaluation and Policy Analysis, vol. 15, no. 2, pp. 195–207, jun 1993. [Online]. Available: http://journals.sagepub.com/doi/10.3102/01623737015002195

[45] L. Lessig, "The Law of the Horse: What Cyberlaw Might Teach," Harvard Law Review, vol. 113, pp. 501–549, 1999.

[46] M. Innes, "Investigation order and major crime inquiries," in Handbook of Criminal Investigation, T. Newburn, T. Williamson, and A. Wright, Eds. Willan Publishing, 2007, ch. 10, pp. 255–276.

[47] A. Broeders, "Principles of forensic identification science," in Handbook of Criminal Investigation, T. Newburn, T. Williamson, and A. Wright, Eds. Willan Publishing, 2007, ch. 12, pp. 303–337.

[48] College of Policing, "Investigation process," 2019. [Online]. Available: https://www.app.college. police.uk/app-content/investigations/investigation-process/

[49] P. A. C. Duijn and T. Vis, "Hyperion Whitepaper. Analysemethode inwinning." Nationale politie, Tech. Rep., 2017.

[50] A. Noroozian, J. Koenders, E. Van Veldhuizen, C. H. Ganan, S. Alrwais, D. McCoy, and M. Van Eeten, "Platforms in everything: Analyzing ground-truth data on the anatomy and economics of bullet-proof hosting," in Proceedings of the 28th USENIX Security Symposium, 2019, pp. 1341–1356. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/ presentation/noroozian

[51] L. Allodi, F. Massacci, and J. M. Williams, "The Work-Averse Cyber Attacker Model: Theory and Evidence From Two Million Attack Signatures," in Workshop on the Economics of Information Security, San Diego, CA, jun 2017. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=2862299

[52] A. Mell, "Promoting Market Failure: Fighting Crime with Asymmetric Informa- tion," 2015. [Online]. Available: https://docs.google.com/viewer?a=v{&}pid=sites{&}srcid= ZGVmYXVsdGRvbWFpbnxhbWVsbGVjb258Z3g6NWYyMTE1YjI1NzdlNjFmMQ

[53] S. Ernst, H. ter Veen, J. Lam, and N. Kop, "Leren van technologisch innoveren "De techniek is niet zo spannend"," Politieacademie, Kennis & Onderzoek, Tech. Rep., 2019. [Online]. Available: https://www.politieacademie.nl/kennisenonderzoek/Onderzoek/ Documents/ 19115190507DIGIPublicatieLerenvantechnischinnoveren.pdf

[54] B. Verheij, "To catch a thief with and without numbers: arguments, scenarios and probabilities in evidential reasoning," Law, Probability and Risk, vol. 13, no. 3-4, pp. 307–325, sep 2014.

[55] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy Artificial Intelligence," European Commission, Tech. Rep., 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc{_}id=60419

[56] N. Diakopoulos, "Accountability in algorithmic decision making," Communications of the ACM, vol. 59, no. 2, pp. 56–62, jan 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid= 2886013.2844110

[57] A. Koene, C. Clifton, Y. Hatada, H. Webb, and R. Richardson, "A governance framework for algorithmic accountability and transparency," European Parliamentary Research Service, Tech. Rep. April, 2019. [Online]. Available: http://www.europarl.europa.eu/ RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

[58] A. McAfee and E. Brynjolfsson, "Big Data:The Management Revolution," Harvard Business Review, vol. 90, no. October, 2012. [Online]. Available: https://hbr.org/2012/10/ big-data-the-management-revolution

[59] C. S. D. Brown, "Investigating and Prosecuting Cyber Crime: Forensic Dependencies and Barriers to Justice," International Journal of Cyber Criminology, vol. 9, no. 1, pp. 55–119, 2015.

[60] M. Konstant, "The Rise of the Agile Careerist," sep 2015. [Online]. Available: https: //www.linkedin.com/pulse/rise-agile-careerist-marti-konstant

[61] F. Dechesne, V. Dignum, L. Zardiashvili, and J. Bieger, "AI & Ethics at the Police: Towards Responsible use of Artificial Intelligence in the Dutch Police," Leiden University Center for Law and Digital Technologies, TU Delft Institute of Design for Values, nationale politie, Leiden/Delft, The Netherlands, Tech. Rep., 2019.

[62] M. Waardenburg, M. Groenleer, and J. De Jong, "Designing environments for experimentation, learning and innovation in public policy and governance," Policy and politics, vol. 48, no. 1, pp. 67–87, jan 2020.

[63] IBM, "CRISP-DM Help Overview." [Online]. Available: https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

[64] Prime Minister's Strategy Unit, "The Strategy Survival Guide," Cabinet Office, London, United Kingdom, Tech. Rep., 2004.

[65] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," Lockheed Martin Corporation, Tech. Rep., 2011. [Online]. Available: https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf

[66] S. Caltagirone, A. Pendergast, and C. Betz, "The Diamond Model of Intrusion Analysis," The Center for Cyber Intelligence Analysis and Threat Research, Hanover, MD, Tech. Rep., 2013. [Online]. Available: https://www.activeresponse.org/wp-content/uploads/2013/07/diamond.pdf

[67] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "MITRE ATT&CKTM: Design and Philosophy," Tech. Rep., 2018.

[68] S. Barnum, "Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX), Version 1.1, Revision 1," MITRE, Tech. Rep., 2014.

[69] J. Connolly, M. Davidson, M. Richard, and C. Skorupka, "The Trusted Automated eXchange of Indicator Information (TAXIITM)," Tech. Rep., 2012.

[70] C. Wagner, A. Dulaunoy, G. Wagener, and A. Iklody, "MISP - The design and implementation of a collaborative threat intelligence sharing platform," in WISCS 2016 - Proceedings of the 2016 ACM Workshop on Information Sharing and Collaborative Security, co-located with CCS 2016. Association for Computing Machinery, Inc, oct 2016, pp. 49–56.

[71] W. Newton Suter, Introduction to Educational Research. A Critical Thinking Approach, 2nd ed. Sage, 2012.

REPHRAIN

Protecting citizens online

UK Research and Innovation

University of BRISTOL

UCL

THE UNIVERSITY of EDINBURGH

KING'S College LONDON

UNIVERSITY OF BATH