The CSAE Framework for Data-Driven Policing

Why It Matters and How It Works In Practice

POLITIE
REPHRAIN
Protecting citizens online
UK Research and Innovation
University of BRISTOL
UCL
UNIVERSITY OF BATH
THE UNIVERSITY of EDINBURGH
KING'S College LONDON

The CSAE Framework for Data-Driven Policing

Why It Matters and How It Works In Practice

Erik van de Sandt Arthur van Bunningen Jarmo van Lenthe

January 28, 2026

January 28, 2026

Erik van de Sandt
Primary and corresponding author.
Email address: erik.van.de.sandt@politie.nl; ev18710@bristol.ac.uk.
National Police; The Netherlands.
University of Bristol; United Kingdom.
National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online; United Kingdom.

Arthur van Bunningen
National Police; The Netherlands.

Jarmo van Lenthe
National Police; The Netherlands.

The first version of the white paper was presented at *the Third INTERPOL-UNICRI Global Meeting on AI for Law Enforcement* on November 25, 2020, and published by the UK's National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) in March 2021[1]. This second edition is based on the original idea by Erik van de Sandt, Arthur van Bunningen, Jarmo van Lenthe and John Fokker. 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.

Navigating the Data-Driven Swamp: CSAE Foundations & Practice

Around the year 2016, the High Tech Crime Unit of the Dutch national police was engaged in a number of large-scale transnational investigations that generated massive amounts of evidence. The biggest concern was not merely a technical problem, but the absence of concepts in words and images that would enable the investigators to closely collaborate with their international peers. After years of working on a common language, CSAE emerged: a comprehensive framework for data-driven policing that serves the public interest. What makes CSAE truly unique is that it did not originate as a top-down strategy, but emerged from operational practice itself: from frontline police work, driven by concrete investigative necessity, and subsequently scaled into a framework that now serves our organization as a whole.

Since then, CSAE (pronounced as 'see-say'), its community and associated tools have been pivotal in virtually all game-changing transnational investigations against serious and cyber organized crime that the Dutch police initiated and/or participated in. Because of these successes, CSAE also gained ground in the hierarchal organizational structure of the national police in the Netherlands and internationally. After publication of the first edition in 2021[1], researchers of the Dutch Ministry of Justice and Security, an independent committee for police reform, a newly appointed director for data-driven policing and top managers of the Dutch national police have all stated that CSAE is the way forward to become a more agile and effective police organization.

With the official adoption of the framework by the Dutch national police, CSAE enters a new phase with new problems. One of the challenges is that developing, transitioning to, and executing data-driven policing is - in the words of philosopher Donald Schön - not hard high ground. Rather, the new paradigm of data-driven policing is a swampy lowland. In other words, becoming a data-driven organization is not a simple, manageable problem that can be solved by merely technical solutions. Aspiring data-driven practitioners and managers will enter a swamp of trial and error while constantly facing problems of great societal concern.

This new edition of CSAE is a co-publication of the police and leading UK universities, emphasizing both the framework's operational relevance and its academic rigor. It presents an updated report from the swamp, focusing on the framework's foundations and use in practice. In due course, follow-up editions will address the transition to a data-driven organization, including topics such as organizational design, governance and structure, and rolling out and scaling CSAE. The framework hopefully provides words and images to everyone in the public and private sector that makes a similar journey. So that they can say what they see.

Last but not least, the authors have asked me to thank all academics and practitioners who provided feedback, shared their experiences, and - most importantly - embraced and helped scale CSAE. This work would not have been possible without your valuable contributions.

Happy reading!

Wilbert Paulissen
Deputy Commissioner of the National Police of the Netherlands.

Executive Summary

Data-driven policing is essential to address the growing complexity of modern crime
This white paper consists of three chapters. In the first chapter, we explain that data-driven policing is an inevitable paradigm for police organizations. Modern policing - such as police intelligence and criminal investigations - is becoming increasingly complex. A major contributing factor is advanced technologies used by criminals to commit and conceal crimes. Even when digital evidence is successfully collected, many law enforcement agencies are left with large, unstructured datasets that are difficult to translate into actionable and factual police reports for strategic or operational purposes. To address this, agencies are increasingly turning to 'smart' technologies such as advanced statistics and artificial intelligence, which are expected to significantly enhance both operational policing and strategic decision-making. Experience has taught us, however, that the transition to data-driven policing is nothing less than a paradigm shift for law enforcement agencies.

CSAE is a comprehensive conceptual framework for data-driven policing
The second chapter presents the core foundational concepts of data-driven policing, developed and put into practice by operational experts over a decade and grounded in existing law-enforcement and industry standards. The associated framework is called CSAE (pronounced as 'see-say'), an abbreviation of our business process for data-driven policing that stands for Collect, Store, Analyze and Engage. Besides a business process, CSAE's core concepts are a public interest philosophy, harmonization objectives, mixed methods methodology and job disciplines.

CSAE guides police-wide coherence and local practice
The third chapter subsequently describes and explains how these five foundational core concepts work together in practice. Because the concepts are interdependent, their alignment is essential in data-driven policing. The chapter shows how CSAE fosters police-wide coherence by providing a shared structure and vocabulary that enables different units to align their approaches. Because the CSAE framework operates at a high level of abstraction, it applies to virtually all police organizations. At the same time, its practical implementation is inherently context-specific. Putting CSAE into practice is therefore not a one-size-fits-all endeavor and requires translating the framework into locally appropriate practices that reflect operational realities and contextual factors. Throughout this white paper, we therefore include numerous examples drawn from real-world operations.

CSAE provides a unifying approach for different police philosophies
It is because of our commitment to a public interest philosophy that we share CSAE's comprehensive framework with a broader academic and practitioner audience. Although CSAE is already a proven practice, we recognize that the development and integration of data-driven policing is still in its early stages. Given how much remains to be done before data-driven methods become an established field within various philosophies of modern policing - e.g., community policing, intelligence-led policing or problem-oriented policing - we hope that CSAE provides a unifying framework for the algorithmic era, fosters harmonization among police organizations and promotes collaboration with the academic world and private sector.

Table 1: The CSAE framework consists of five core concepts that relate to why, where to, what & when, how and who.

CSAE Core Concepts	Relates to:	Short Description
Public Interest Philosophy	Why	Data-driven policing must be value-driven, ensuring that ethics, principles and valuesguide practice and reinforce, rather than undermine, police legitimacy
Harmonization Objectives	Where to	Besides legal and organizational harmonization, police must also strive for technical harmonization internally and with partners as they transition to data-driven policing
Business Process	What & When	Data-driven policing requires a structured, clearly defined business process that consists of collecting data, storing information, analyzing intelligence and engaging crime through fact-based interventions
Mixed-Methods Methodology	How	A data-driven methodology that integrates qualitative and quantitative sources, methods and techniques is essential to achieving valid, reliable and credible results
Job Disciplines	Who	Data-driven policing requires the coordinated involvement of three interrelated job disciplines: (1) domain experts - e.g., criminal investigators, intelligence analysts and data analysts; (2) technical experts - software developers, digital investigators and data engineers; and (3) numerical experts - statisticians, mathematicians and data scientists

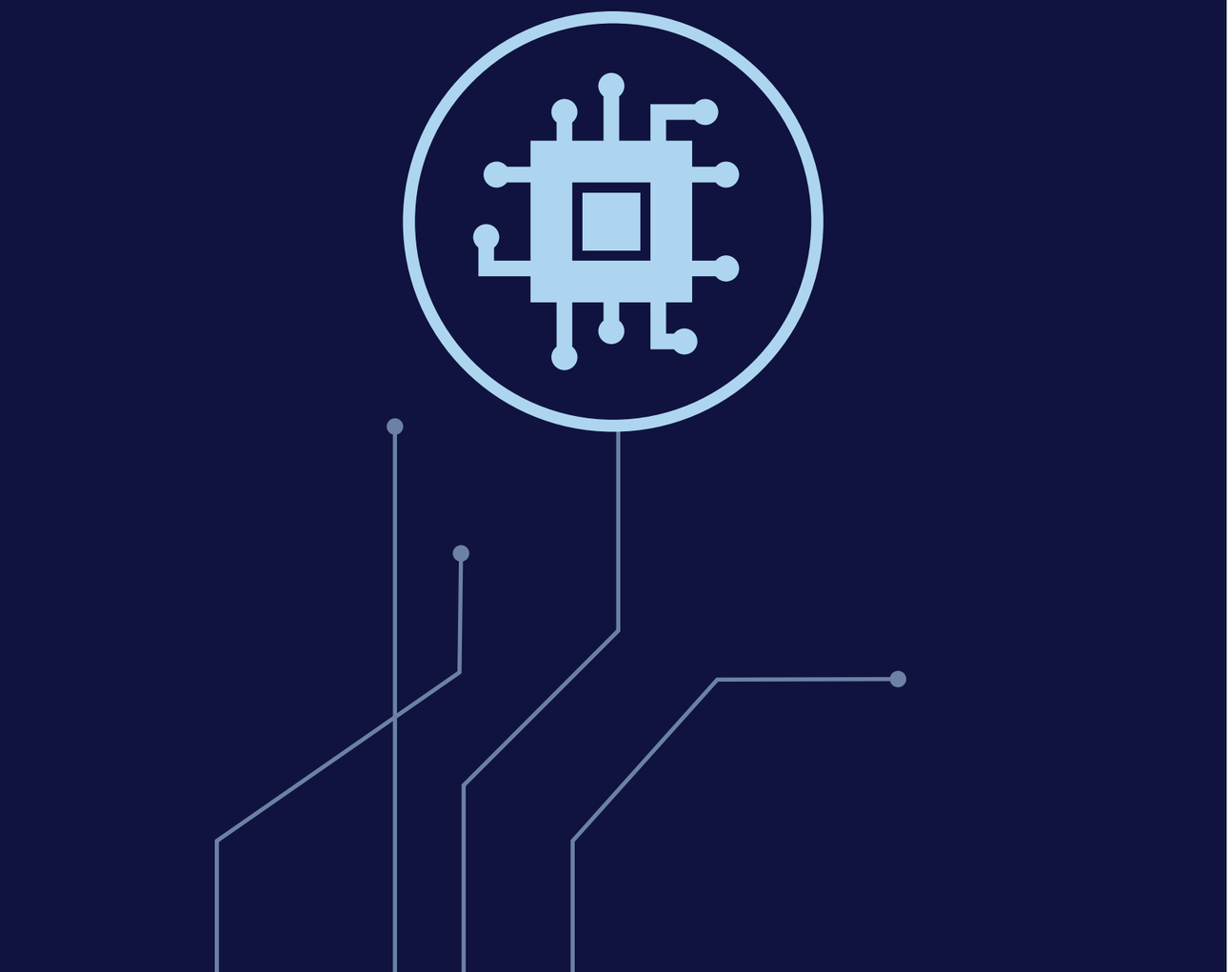
Contents

List of Figures

List of Tables

1

The Data-Driven Paradigm



In this white paper, we introduce the rationale, key concepts and practical application of CSAE (pronounced ‘see-say’): a comprehensive and unifying framework for data-driven policing. As the conceptual understanding and documentation of this emerging type of policing are currently lacking, this second edition aims to bridge that gap. It incorporates new insights drawn from state-of-the-art industry standards and years of experience in high-tech operations against various types of serious and organized crime.

Our focus is not on explaining what advanced statistics and artificial intelligence (AI) are or detailing what kind of algorithms police organizations apply in their work. Instead, we argue that a framework like CSAE is crucial for managing and guiding the inevitable adoption and integration of AI in modern policing philosophies and daily police work. As such, CSAE ultimately addresses how police interact with technology and make decisions within a structured and accountable framework.

We are fully aware that understanding CSAE concepts is only the first step. The real challenges arise in the practical application and transition to data-driven policing, where the primary barriers are organizational, cultural and political. In due course, we will publish follow-up editions addressing the transition to a data-driven organization, including topics such as organizational design, governance and structure, and rolling out and scaling CSAE. Lastly, *italicized* terms are used throughout the paper to draw the reader’s attention to CSAE’s key concepts.

Police intelligence and criminal investigations as examples of data-driven policing
Why emphasize specific subfields of policing namely, law enforcement, and more specifically police intelligence and criminal investigations while placing less focus on frontline policing (like emergency response, order maintenance and service provision)? Research indicates that algorithms are currently most prevalent in law enforcement than these other core functions[2]. Some scholars even suggest that data-driven policing represents a new paradigm, because of the significantly different data-driven methods employed in high-profile investigations into organized crime[3]. Notably, many of these investigations have been conducted especially in the Netherlands using CSAE principles [4]. Because investigations have a retrospective and real-time focus on historical and current crimes rather than future events, we do not discuss predictive policing. Predictive policing involves the use of algorithms to optimize the allocation of police resources to prevent future crimes, but it is also controversial (see e.g.,[5] [6] and the text box in Section 3.5).

1.1 A Short History

We describe and explain CSAE through the lens of a specific subfield of policing: criminal investigations. The recent history of investigations helps to clarify where the police come from, where they stand today, and where they are heading. Criminal investigations consists of three interrelated disciplines. The first is the domain knowledge perspective of - what we nowadays call - traditional criminal investigations, while the second discipline is the technical perspective of digital investigations that was introduced in the eighties of the last century[7]. In the 2000s, with the rise of intelligence-led policing, law enforcement adopted a numerical discipline based on relatively simple statistical methods[8] ; this was recently followed by the introduction of advanced statistical techniques and AI[9]. The integration of domain knowledge, technical and numerical approaches leads to data-driven policing.

Three different perspectives on the same document
Imagine a lengthy text document of a suspect. How do different investigative perspectives review this document for evidence? The criminal investigator and police intelligence analyst with domain knowledge analyze the content of the document. They identify key details such as names, dates, locations and events mentioned, establish chronological order and sequence of events, determine the suspect’s purpose and motivation, understand the intended audience, and make links to the suspect’s activities. The digital investigator, with a technical ‘bits and bytes’ perspective, examines the metadata of the document. This includes looking at the file format, version history, creation and modification dates, file paths and system locations, and author information. The data scientist converts the text of the document into numerical data for statistical analysis. They highlight words that are unique to the document, cluster its topics and calculate the frequencies and distributions of e.g., city names.

From traditional to digital investigations
The goal of criminal investigations is the attribution of crimes to suspects - determining who did what - for prosecution purposes, and law enforcement agencies (LEAs) have a monopoly on this process of bringing suspects to justice. Before the digital era, all criminal investigations were traditional, offline investigations and

forensics. The focus of these investigations was mostly on the human and physical factors of crime, thus very much social and behavioral in nature, combined with sound understanding - i.e., *domain knowledge* and *subject matter expertise* - of criminogenic factors in the physical world. These investigators have a deep level of qualitative knowledge which includes experiential knowledge, expert judgment and gut instinct[10]. They gain this knowledge by using qualitative methods and techniques like observations, interrogations and eavesdropping for their truth-seeking process, supported by traditional forensic sciences that focus on physical, tangible evidence from the offline world such as ballistics, DNA analysis, fingerprint analysis and pathology[3]. As the world became increasingly digitalized, crime became enabled, 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 - an umbrella term for digital forensics, OSINT and cyber crime investigations - 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 digitalization of police work, such as the merge between traditional and digital investigations, is still an ongoing process for many law enforcement departments around the world[11] [12].

The limitations of digital investigations

Today, criminal investigations face an effectiveness crisis[13] [14] [15] [16]. Investigations are too labor 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[17]. Yet academics and practitioners predicted the end of the 'Golden Age of Digital Forensics' years ago[18] [19]. They foresaw a situation in which evidence would be permanently out of reach to investigators, or - when successfully retrieved - could not be properly analyzed because of data management issues. These predictions have come true. The security practices employed by criminal actors to obstruct or evade police investigations have become democratized, and law enforcement agencies have been unable to effectively address these developments[20]. In other words, because criminal actors have access to a broad range of legitimate and illegitimate - mostly technical - security products and services, law enforcement agencies face increasing difficulty in collecting the evidence necessary to build cases against criminal suspects. In the early 2010s, the concept of Big Data gained prominence within police organizations. This highlighted that, while traditional and digital forensic methods can retrieve evidence, they are not equipped to process the volume, variety, velocity and veracity of large data sets (the '4Vs of Big Data') into timely, relevant, accurate, and actionable reports. Thus, law enforcement agencies are unable to create a fifth V: value, which represents the operational and strategic utility necessary to overcome the effectiveness crisis in modern policing. This is understandable: data problems do not stop when evidence is collected. Data sets might be encrypted or come in unknown formats, while pieces of evidence might be hidden like a needle in a digital haystack or be false because of deception tactics,to name just a few criminal countermeasures[20].

Figure 1.

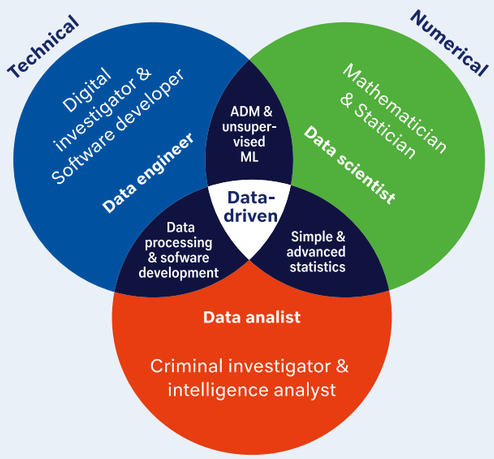


Figure 1. Venn diagram of data-driven investigations
Domain knowledge, technical and numerical approaches stand on their own, but also overlap. Digital investigators may extract and process evidence from data carriers of suspects. Practitioners with a mathematical background - e.g., data scientists - 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 criminal investigators with domain knowledge of crime and appropriate responses. When these disciplines work together fully, that is what we call data-driven. Yet a possible future of police work is one in which AI agents make autonomous decisions (ADM) derived from machine-learning (ML) patterns, reducing the need for human domain expertise. The Venn diagram is further explained in Section 2.5.

Data processing: the gross data problem

We argue, like other academics[17] [3], that today's challenge for law enforcement agencies is not so much the collection of data, but the steps afterwards which is essentially about the digitalization of police work such as processing, analyzing and presenting findings. In investigations, most law enforcement agencies collect and directly process the *circumstantial* and *factual evidence* they need to build a case, i.e., *net evidence*. This is understandable: historically, all evidence of traditional investigations was net evidence. Investigators put all relevant findings directly in a police report. Creating police intelligence is generally a separate business (instead of being an integral part of an evidence flow as depicted in Figure 5). Since the Big Data era, digital investigations have been generating 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 data, the truth is that most agencies miss the organizational and technical means to do so. The under-utilization of gross data described above is primarily the result of legal, organizational and technical fragmentation.

Data processing: the institutional capacity problem

In practice, law enforcement agencies are essentially 'siloed' systems, characterized by legal, organizational and technical fragmentation[21] [22]. As described in Section 1.2, there is no common understanding of, and consensus about, a business process, methodology, philosophy, policies and associated concepts and language to process all the gross evidence from multiple investigations. Let alone that agencies have the technical resources to normalize large raw data sets in unknown formats from partner agencies, load those sets in their police systems, conduct advanced analyses on these sets and select suitable targets for investigations. 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[20] [23].

The need for, and purpose of, data-driven policing and investigations

Academics and practitioners alike have acknowledged that digital forensics and associated workflows - like[24] [25] [26] - are not sufficient to scale the processing of exponentially growing and highly abstract data sets and produce timely, accurate, relevant and actionable outcomes, or any results at all[27] [18] [19]. New ways of policing - referred to in terms like *algorithmic policing, big data policing* and, indeed, *data-driven policing* - that use advanced statistics and artificial intelligence have been proposed as the way forward. What these technologies bring to the investigative table is a so far missing numerical valuation of evidence. As depicted in Figure 1, adding mathematics to the existing layers of traditional and digital investigations results in *data-driven investigations*.

Based on a review of multiple large-scale, data-driven investigations in the Netherlands, scholars argued that these technologies will find their way into (1) the workflows of primary and secondary police work - such as the use of drones in investigations and data processing of collected evidence, respectively - and (2) the steering and management of police work, including *intelligence-led policing* and *business intelligence* (see text boxes in Sections 1.2 and 2)[3]. Besides its impact on operational work and strategic decision-making, we also firmly believe that data-driven policing should strengthen the legitimacy of police agencies within liberal democracies. Yet for all of these purposes - especially the core objectives of effectiveness and legitimacy [28] - police will need a comprehensive framework as explained in the next section.

Data-driven policing leads to precision policing

Many police managers believe that data-driven policing will increase efficiency, i.e., 'do more with less'. However, in our experience, data-driven policing today is more about organizational adaptability and thus effectiveness in the algorithmic era: the ability to ask new questions that could not be answered before, and to ask familiar questions in smarter, more effective ways. Data-driven police work allows formulating data-informed strategic priorities as police organizations instantly know the broad topics of victim complaints at any given time. Agencies can learn what organized crime hotspots they need to proactively target, discover previously undefined information entities related to child sexual abuse, make probability statements about whether the identified author of text A is also the so far anonymous author of text B, or assess the impact on the larger criminal network when a money launderer is arrested. Thus, based on our experience, using algorithms and the CSAE framework closely aligns with *precision policing*[29] [30]. Efficiencies are currently, in our opinion, mainly a positive side effect of CSAE's strategic harmonization objectives, such as shared data schemas, ETL tooling and forensic software that promote technical *interoperability* within law enforcement agencies and between partners. We acknowledge that many law enforcement agencies are nowadays experiencing 'the between times of AI'. Advanced statistics e.g., regression analysis and AI e.g., natural language processing are used to support current activities but have not yet transformed the police system itself, such as bots taking over certain police tasks for reasons of efficiency[3].

1.2 Why a Common Framework for Data-Driven Policing?

We frequently encounter police investigators who claim to work in a data-driven manner. While we do not doubt their intentions or work, these colleagues often cannot clearly articulate what this emerging discipline entails or how their day-to-day work fits into a broader conceptual framework. In other words, they lack a foundational understanding and a shared language to explain what they are doing, what they aim to achieve, and what they need to reach those goals. This absence of a comprehensive framework and common terminology for data-driven policing is precisely why we developed CSAE.

Absence of evidence-based frameworks for data-driven policing

Scholars have pointed to the lack of standardization in criminal investigations for many years (see e.g.,[18]). The practical consequences of this gap should not be underestimated. Without a common framework,

shared understanding or unified language, international strategic and operational meetings are often plagued by communication problems. Those who have attended such meetings will recognize the image: individuals presenting an unstructured mixture of investigative methods and techniques, strategic and operational goals, steps from implicit business processes, and various legal, organizational and technical details - to an audience from diverse cultural and professional backgrounds. Unfortunately, this situation is not much better at the national level. Different units, departments, divisions, and even individual teams often develop their own working cultures and internal languages, leading to divergent ways of 'doing things'. Unsurprisingly, this makes internal collaboration challenging - and external collaboration even more so.

Over-reliance on non-policing concepts
Yet today's complex investigations demand exactly that: internal and external collaboration. As a result, the need for common concepts, words and a shared framework is greater than ever. In the past, police forces have borrowed concepts from the military and intelligence domains, such as the Intelligence Cycle (see Appendix B of the first edition of our white paper[1]) and the DIKI model (see text box in Section 2.3). With the rise of data-driven investigations, law enforcement is once again searching for language to describe and structure their work. Increasingly, we see agencies borrowing terminology from the private sector - this time from the world of Big Data, such as CRISP-DM. In Appendix B of the first edition of our white paper we explain why such borrowing can be problematic[1]. Concepts developed for private-sector analytics often do not align with the unique objectives and the legal, organizational and technical environments of law enforcement agencies. Blindly adopting these 'alien' frameworks risks undermining effectiveness, clarity and police legitimacy. The CSAE framework was designed specifically to address these shortcomings and to provide a structure tailored to the needs of policing in the algorithmic era.

A single, comprehensive framework for and by police
What is needed is a comprehensive, unifying framework for different forms of data-driven policing, developed by and for the police. Comprehensiveness means that a framework consists of multiple components that are also well-aligned with one another. We believe that our journey toward developing a comprehensive framework for data-driven policing will be similar to that of other law enforcement agencies. Our journey began with a big data challenge and ethical concerns about potential vendor lockin tied to the key technologies required to process and analyze that data. Therefore, we aimed to develop ETL and advanced analytical tools ourselves, in collaboration with our partners. This technical harmonization between (inter)national partners is essentially a strategic objective. However, to process and analyze data together using mutual data schemas, tools and models, a shared business process is necessary. Very soon, we - and others who have embarked on a similar journey - realized that a data-driven methodology is essential. After all, it is necessary to determine and measure what is needed at each phase of the business process. When applying both the business process and the methodology, you encounter the job disciplines already in place as well as those needed for the future. As initial successes emerge, government officials and police leadership may take interest. At that point, it becomes important to explain how these innovations fit within the broader history and practice of policing. One of these elements is the police's commitment to the public values of liberal democracies. From the outset and throughout the development and implementation of CSAE, we have encountered normative and value-driven questions. Ultimately, a public interest philosophy becomes the unifying element that connects all core concepts of data-driven policing, forming a truly comprehensive framework.

1.3 Why and For Whom is This Publication?

A new algorithmic era is taking shape, and only by understanding data-driven policing can we maintain control over it. By fostering public debate on data-driven policing, this CSAE white paper seeks to strengthen liberal democracies. Created by operational experts and based on established standards, the framework welcomes feedback and collaboration from scholars and industry specialists to refine and improve it.

Discussion strengthens policing in liberal democracy
We hope that CSAE will not only promote harmonization between law enforcement agencies, but will also align legislation, policy and research to the real world practices of policing, and as a result, strengthens liberal democracies. Indeed, public debates about reinforcing oversight and strengthening transparency, explainability and judicial review are not new. However, the debate should include the underlying technologies, algorithms and methodological choices as well [31]. We believe that the debate about the usage, scope and limitations of data science in criminal investigations will become more specific, thus improves, when scholars use CSAE for critical thinking about the risks of data-driven investigations such as biases, privacy and surveillance[32].

Understanding privacy and its practical application in data-driven policing
CSAE can help academics and practitioners interpret and apply abstract legal concepts such as 'privacy' in practical terms. During the Collect phase, law enforcement gathers data. Given today's vast data sources, *data minimization* is crucial: only necessary data should be collected, while acknowledging that criminals generate massive amounts of data themselves since the Big Data era (see also Figure 5). In the Store phase, collected data must be securely stored according to *data protection* laws, with strict access controls and retention limits. The principle of data minimization continues here, requiring the deletion of unnecessary data. During the Analyze phase, access to evidence is controlled through legal authorization and role-based access. The European principle of *purpose limitation* ensures data is analyzed only for the original investigative purpose.

Finally, in the Engage phase, law enforcement must legally *disclose* evidence to involved parties (suspects, victims or witnesses), other agencies, or authorized third parties, following laws on chain of custody and data subject rights.

By the police, for the police & a broad audience
This white paper adopts a multidisciplinary approach - integrating socio-technical and legal perspectives - and is aimed at a diverse audience, including academics, private-sector experts and police professionals. Throughout this paper, we explain the philosophy behind CSAE, and show how the framework aims at serving the public interest. The business process, harmonization objectives, methodology, job disciplines and public interest philosophy are not theoretical, top-down invented concepts. On the contrary, the framework has been developed over a decade by operational experts and is grounded in existing policing philosophies and industry standards (see Appendix B of the first edition of our white paper[1]). As a result, the framework has been proven in practice. Law enforcement officers who implemented CSAE have successfully developed a range of data-driven models, forensic tools and methods and techniques. This paper is an open invitation for academics and industry experts to critique, strengthen and further develop CSAE.

Is data-driven policing truly a paradigm shift, or old wine in new bottles?
While this paper refers to data-driven investigations as a 'paradigm shift' and top management often calls it a 'game changer' some academics and practitioners question this[3]. They argue that even before the digital era, the phases of Collect, Store, Analyze and Engage have long defined policing practice; what has changed are merely the methods, such as digital forensics, statistical analysis, and now AI. Others see data-driven policing as a continuation of intelligence-led policing, which also uses data to shift from reactive to proactive approaches targeting 'hot spots', 'hot groups' and 'hot times'[8]. We partly agree: CSAE builds on these longstanding, often unwritten best practices. However, intelligence-led policing lacks a defined business process, methodology, job disciplines, harmonization objectives, and public-interest philosophy. Many forces have therefore adopted the military Intelligence Cycle and DIKI model, which do not align with the police's unique objectives, legal frameworks, and organizational structures ((see for example Appendix B of the first edition of our white paper[1], and Section 2.3). Similarly, problem-oriented and community policing emphasize the Analyze and Engage phases respectively, but will increasingly need to become data-driven. In conclusion, CSAE does not replace these established paradigms but connects and strengthens them, providing a unifying framework for the algorithmic era explicitly designed for the realities and responsibilities of modern policing.

2

Core Concepts of CSAE

We have identified five core concepts that serve as the foundational pillars of data-driven policing. CSAE initially began with our ethical objections toward several software vendors who not only wanted to provide advanced analytical tools, but also to normalize very sensitive police data. Therefore, the idea arose to strive for *data sovereignty*, and build ETL and advanced analytics tools in-house (see for example[33]), and with our most trusted partners. This motivation marked the beginning of our *public interest philosophy*.

The ambition to develop core technologies in-house subsequently led to the idea of *technical harmonization* within and between law enforcement agencies. However, as we closely collaborated with national and international partners on legal, organizational and technical harmonization, it became clear that a *business process* was needed. After all, do we need to harmonize on data collection, information storage, intelligence analysis or engagement on crime with facts? Soon, we discovered that a fourth core concept was required to put data-driven policing into operational practice: a *methodology* to measure what to collect, store, analyze and engage. A final core concept - already introduced in Section 1 and illustrated in Figure 1 - concerns professional roles involved in data-driven policing, specifically the *job disciplines* responsible for its implementation.

Applicability of CSAE in traditional police intelligence and criminal investigations
Over the years, we learned that many agencies and departments are not yet ready to apply advanced statistical methods and techniques. However, we noticed that the framework is still very applicable to agencies and departments that only conduct traditional and/or digital investigations. This is not surprising: virtually all law enforcement agencies have (unwritten) public interest philosophies, harmonization objectives, business processes and working methods. CSAE provides a coherent structure for organizing these core concepts, whether an agency relies solely on qualitative investigative approaches or is gradually transitioning to more advanced, quantitative methods. Besides its operational utility, CSAE can also be used for business intelligence (BI) purposes (see Section 2.4).

Table 1: The CSAE framework consists of five core concepts that relate to why, where to, what & when, how and who.

CSAE Core Concept	Relates to:	Short Description
Public Interest Philosophy	Why	Data-driven policing must be value-driven policing, and the ethics, principles and values of a public interest philosophy strengthens police legitimacy
Harmonization Objectives	Where to	Besides legal and organizational harmonization, police must also strive for technical harmonization internally and with partners as they transition to data-driven policing
Business Process	What & When	Our common business process cycle for data-driven policing consist of collecting data, storing information, analyzing intelligence and engaging crime through fact-based interventions
Quadrant Methodology	How	The Quadrant methodology ensures valid, reliable and credible results through a mixed-method approach that combines qualitative and quantitative sources, methods and techniques
Job Disciplines	Who	Three interrelated job disciplines play a key role in data-driven policing: (1) domain experts - e.g., criminal investigators, intelligence analysts and data analysts; (2) technical experts - software developers, digital investigators and data engineers; and (3) numerical experts - statisticians, mathematicians and data scientists

2.1 Why: Public Interest Philosophy for Police Legitimacy

Throughout this paper, we explain our public interest philosophy and emphasize its ethics, principles and values which are crucial when implementing CSAE. We argue that any framework should incorporate a public interest philosophy, as this provides the *legitimacy* of data-driven policing, in contrast to the *effectiveness* provided by the other concepts. At the same time, the public interest philosophy must serve as a common thread throughout all concepts, influencing the choice of harmonization partners, applied methods and techniques, every step of the business process, and the staff involved.

Data-driven policing is value-driven policing

Our motivation to develop CSAE stemmed from ethical concerns about private security vendors processing and analyzing highly sensitive evidence. Do we really want platforms that are built, operated and accessed by a few private companies[34] - companies that, as academics argue[21], "hide behind trade secrecy and nondisclosure agreements, ultimately circumventing public-sector transparency requirements and lowering police accountability by making it harder for scholars to study, regulators to regulate, and activists to mobilize for or against specific practices"? No wonder that alongside overly optimistic potential benefits of algorithms, scholars, politicians and the public have so many concerns about data-driven policing[2].

For the better or worse, data-driven policing is very much social[21]. In other words, law enforcement and stakeholders all have the power to craft it, both for effectiveness as for legitimacy. While the CSAE concepts of harmonization objectives, business processes and methodology ensure the effectiveness of data-driven policing, it is ethics, values and principles that provide its legitimacy. These should not be treated as standalone concepts but rather as integral parts of a broader public interest philosophy that runs as a common thread through the other concepts of CSAE. Concerns about how police use their power are always present in debates about policing. In liberal democracies, police rely on the legitimacy they gain from public trust, cooperation and compliance. This raises important questions about how much the public actually views the police as legitimate[35]. So without such a philosophy, any framework becomes valueless, and this is what sets law enforcement in liberal democracies apart from authoritarian regimes with little legitimacy. In other words, data-driven policing must be *value-driven policing*.

Public interest philosophy for police legitimacy

We view a public interest philosophy as a means to enhance *police legitimacy*. Police legitimacy is a multifaceted concept that extends beyond mere effectiveness; it also depends on the lawfulness of police actions, procedural justice, transparency, accountability, ethical conduct and alignment with social values. Police legitimacy refers to the public and stakeholders' perception that the police have the rightful authority to enforce the law and that their actions are justified, fair and appropriate. This legitimacy is crucial because it influences public trust and cooperation, both essential for effective policing. When elements of police legitimacy are present, the public is more likely to perceive the police as a legitimate authority, fostering greater cooperation and support for law enforcement efforts[36].

We argue that the legitimacy of data-driven policing will depend on both existing and new ethics, principles and values that may not yet have been integrated within law enforcement (see also the text box below). Generally, perceived neutrality and trust in decision-making motives by the general public increase police legitimacy[36]. However, aspects of data-driven policing such as centralization and standardization which enhance effectiveness and reduce organizational silos, may also make the police appear more distanced, impersonal, formal and decontextualized - what has been termed the '*abstract police*'[37] [38]. Experts therefore rightly argue that law enforcement should be transparent about how algorithms are used to increase perceived neutrality, and explain, justify and account for algorithm-based interventions to build trust in their motives[21] [6]. This illustrates one of the many complex ethical, legal and societal dilemmas associated with data-driven policing.

Existing ethics, principles and values for data-driven methodologies

We observed that many existing frameworks remain highly relevant to the design and use of data-driven methodologies in criminal investigations. We noticed the following line of reasoning during internal discussions. Firstly, methodologies must align with the *fundamental human rights* and *civil rights and liberties* 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 checks, balances and fairness, and avoid abuse of power, arbitrary decisions and discrimination. Secondly, methodologies should be placed in existing empirical and normative research methodologies. An 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*[39]. Similar to traditional statistics, data-driven methodologies 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, we acknowledge the need to explore new ethics, principles and values for data-driven policing, as outlined in the text box in Section 2.4.

2.2 Where To: Technical Harmonization

Law enforcement agencies must set strategic harmonization objectives to overcome their fragmented silos that hinder data-driven policing[20]. Harmonization involves continuously aligning legal, organizational and technical aspects, which facilitates legitimate, efficient and effective collaboration both within law enforcement agencies and with their partners. In fact, CSAE, as a common and comprehensive framework for data-driven policing, is itself an example of harmonization.

Figure 2.

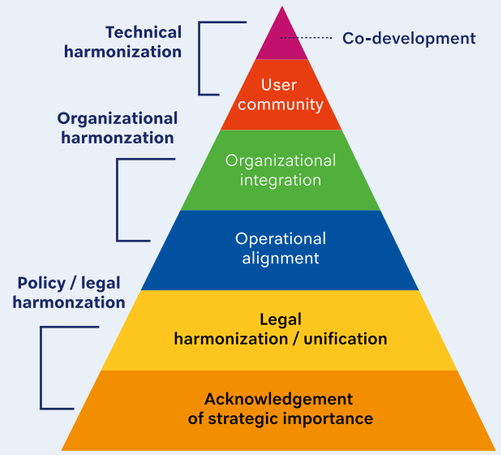


Figure 2. Pyramid chart of collaboration

These six stages represent the different steps in collaboration between law enforcement agencies, and (other) public and private partners[20]. 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 creating user communities around, for example, common data schemas and co-developing forensic software are outcomes of technical harmonization.

What is harmonization?

While there is consensus among academics and industry experts that there is a need for standardization in technical processes, terminology and ontologies [40] [41] [42] [43], 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, policy makers and the judiciary, decide that this is needed. Legal and organizational harmonization are prerequisites for technical harmonization; however, technical harmonization efforts also impact the other two layers. In other words, legal, organizational, and technical harmonization are interconnected and influence each other. 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.

Legal harmonization

The first step in legal harmonization is *acknowledgement of strategic importance* within an organization or with external (inter)national partners, at the administrative, policy and/or political level. This requires a clear understanding of an organization's objectives, capabilities and its relevance to the effective fight against crime. In practice, this foundational step is often overlooked, resulting in partnerships with ill-suited actors or in function creep as collaboration progresses to a new level. Signs of this phase come in different forms. Agreements are made on a policy level, such as organizing periodic meetings, sharing best practices on investigations or simply being present on international conferences hosted by the other party. Secondly, acknowledgement of strategic importance includes agreements made on an operational level. Thus, even countries that cannot legally or politically conduct joint investigations or share information entities with other nations are still able to make agreements about fighting domestic points of a crime threat that are a problem to the other party. Lastly, nations can make agreements to work towards the next step in international collaboration: *legal harmonization* and *legal unification*. Before parties can share investigative data and/or conduct joint investigations, relevant substantive and procedural laws and data protection acts have to be harmonized. On an international level, legal harmonization of laws occurs on both a bilateral and multilateral level such as Mutual Legal Assistance Treaties (MLATs).

Organizational harmonization

Legal harmonization and unification provide the foundation for organizational harmonization. This enables closer collaboration on operations - both nationally and internationally. Within countries, this may take the form of interagency cooperation and technical assistance between federal/national and local partners. Across borders, it includes mechanisms such as parallel investigations, joint investigations and the use of dedicated police liaison officers. The first phase of organizational harmonization is *operational alignment* between partners, as investigations increasingly require coordination, deconfliction and, at times, convergence.

For most agencies, the current endpoint of both national and international collaboration is *organizational integration*: agencies working together at a shared physical location and/or online platform to investigate crimes, share and analyze evidence and conduct joint operations. Examples at the federal/national and local

levels include fusion centers, joint task forces and public-private partnerships, while Europol and Eurojust, the Five Eyes Law Enforcement Group and INTERPOL are international examples of organizational integration.

Technical harmonization

Data-driven policing will lead to significant changes in how law enforcement agencies collect data, store information, analyze intelligence, and engage in lawful activities against crime. The concept of *technical harmonization* within law enforcement has been a longstanding topic of discussion[44]. However, its understanding and implications have become more crucial than ever to address current data processing challenges and facilitate the adoption of data-driven policing. Technical harmonization is indeed a part of digitalization of police work, and essential for ensuring interoperability, efficiency and effective collaboration across law enforcement agencies. From our experience, it encompasses several interrelated key concepts. Firstly, this type of harmonization is about *development* of data schemas, forensic tools and analytical models (i.e., innovation - 'change the business') and *usage* of these products including technical deployment, management and support ('run the business' - management aspects that we will publish in a future edition about the transition to a data-driven police organization). These aspects apply broadly to *data* (including data schemas), *applications* and *software*, and *infrastructure* and *hardware architectures* within law enforcement agencies, and must be harmonized with the *business architecture* of policing, more specifically the CSAE business process described in Section 2.3[45]. Technical harmonization aims to establish shared, technically uniform *criteria*, *methods* and *principles*. This process operates on a continuum that spans from initial *harmonization* efforts to eventual *standardization* across policing practices.

Harmonization: a strategic objective across interdependent technical, legal and organizational domains

A practical illustration of the strategic nature of harmonization, and of how technical, organizational and legal harmonization are interdependent, is the need for police organizations to establish consortia within their own organizations and with external partners. These consortia should develop key technologies such as data schemas, ETL tools and analytical models. From a public interest philosophy, these technologies can then be shared with like-minded public partners that have fewer resources as a form of capacity building and distributive justice but that are nevertheless strategically important in the fight against crime. Technical harmonization not only improves operational efficiency, but also ensures consistency and interoperability across law-enforcement systems. At the same time, sharing data schemas, tools and models with other organizations raises legal questions, including issues of accountability, governance and transparency. Consequently, changes in one form of harmonization will inevitably affect the others.

2.3 What & When: Business Process

A business process in law enforcement is essential to ensure the systematic transformation of raw data into facts. This structured approach promotes efficiency, accountability and alignment with legal, organizational and technical harmonization objectives, enabling effective and informed decision-making throughout all operational phases.

Figure 3.

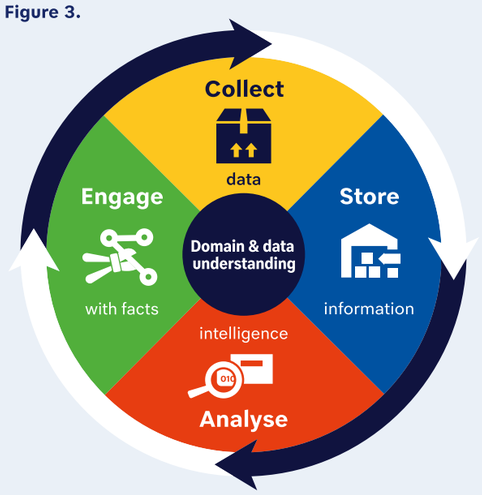


Figure 3. Business process

The CSAE business process is a continuous process cycle as depicted above. Appendix A contains a detailed image board of the business process. We noticed how this image board promotes understanding and discussion among academics and practitioners alike.

DIIF Continuum

CSAE builds upon the common understanding within the military and intelligence community about the relationship between data, information, knowledge and intelligence - commonly known as the DIKI continuum [46] [47]. At the same time, our framework connects to the world of policing and criminal investigations with their own unique legal principles, organizational structures and current big data challenges. This means that CSAE follows the transformation of police records (policing) or evidence (investigations), and therefore introduces the DIIF continuum as shown in Figures 5 and 4: from raw data in Collect, information in Store, intelligence in Analyze, to facts in the Engage phase. This last and additional step - i.e., facts: statements about reality that go beyond reasonable doubt and are bound by principles as accountability, due diligence and neutrality - is what sets DIIF apart from DIKI. Police must engage with facts when interfering in the lives of its citizens, whether the latter are suspects, victims or witnesses. Sanctioning without facts, but merely based on intelligence, will lead, and unfortunately has led, to harming the lives of innocent citizens and thus undermining the foundations of liberal democracies[48]. Indeed, the DIIF continuum implicates that fact-based interventions are intelligence-led (see the text box in Section 1).

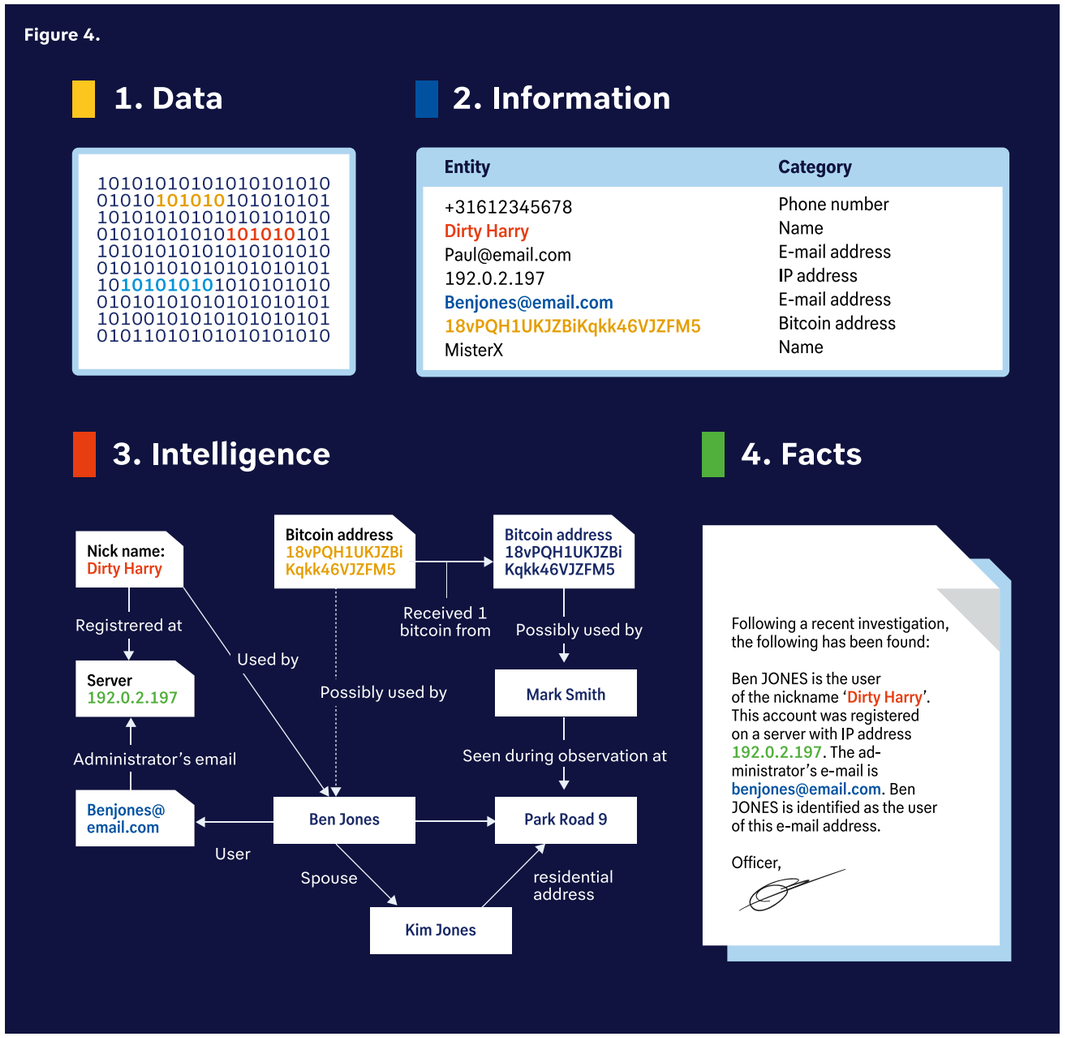
Figure 4. The DIIF model

Data is a collection of raw and uninterpreted observations and measurements. On its own, data has no inherent meaning; it is simply the recording of values in various formats. When data is processed, organized and interpreted, it becomes *information*: meaningful and contextualized entities. When information entities are evaluated, analyzed and linked together, and we understand their relationship, it becomes *intelligence*[47]. When intelligence insights and/or interpretations are rigorously validated, reproducible as objective, verifiable truths, and sufficiently substantiated to stand as *evidence* in court, intelligence becomes *fact*.

Knowledge: the human factor in DIIF

So, where is 'knowledge' positioned in the DIIF continuum represented as 'understanding' in Figure 3? From a CSAE perspective, knowledge is not a separate step between information and intelligence, but represents the 'human factor' that runs as a thread through all steps of the DIIF continuum[49]. In other words, human beings domain experts, technical experts, numerical experts, managers must have a solid understanding of the data, information, intelligence and facts they are working with, and they acquire that knowledge and understanding through the use of another core component of CSAE: the Quadrant mixed-methods methodology (see next Section 2.4). This human factor means much more, it confirms again the idea that data-driven policing is fundamentally social, and very much shaped for better and worse by humans[21].

Figure 4.



Business Process
We have linked the DIIF continuum to a straightforward, four-step, circular business process: obtaining data in the Collect phase, warehousing information in Store, creating intelligence in Analyze, and executing lawful interventions based on facts in Engage. These phases align with existing legal, organizational and technical structures of law enforcement agencies. Typically, law enforcement agencies have operational 'Engage' teams that also collect operational data, supported by a few dedicated digital forensic 'Collect' teams. IT departments store the collected data, transform it into information, and manage associated police systems. Police organizations also have analytics departments that analyze the stored information and turn it into intelligence for decision-making, monitoring, or preparation of new investigations. The outcomes of investigations by Engage teams generate data that must be stored, making CSAE a circular process.

Figure 5.

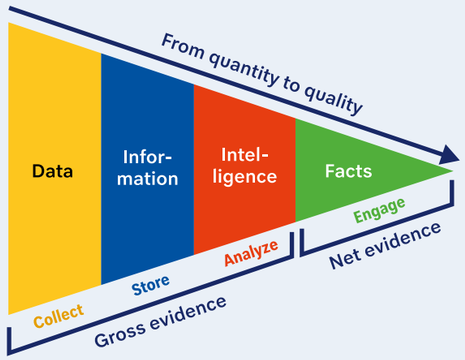


Figure 5. Transformation of evidence
When evidence is collected during data-driven investigations, raw data have to be processed, evaluated and aggregated into information, intelligence and ultimately factual police reports about who did what beyond a reasonable doubt. *Data minimization* is therefore not binary but a continuum: from *data maximization* at one end - mirroring how crime phenomena since the Big Data era have generated vast amounts of data - to progressive reduction toward minimization.

The foundation of successfully working with the CSAE business process is gaining domain and data understanding. This refers to a fundamental understanding of crime and how it manifests in data. Although this understanding is essential in all phases, we have observed that many analysts and investigators in the Analyze and Engage phases have a solid domain understanding but lack data understanding. Conversely, those in the Store phase, such as data engineers and software developers, often have data understanding but lack domain insights into crime and appropriate responses.

The importance of gross evidence in the overall evidence flow
As shown in Figure 5, each step in the evidence-processing chain should see a decrease in quantity but an increase in quality until only net evidence i.e., court-admissible evidence remains. Moreover, the gross evidence residue data, information, and intelligence not used in court proceedings remains highly valuable for law enforcement agencies, as it contains important leads about past, current and future crimes. Access to gross data is therefore a necessary condition for data-driven policing, which may require new regulations that, in turn, will impact the organizational and technical infrastructure of police agencies. Ultimately, how police process and use gross evidence fundamentally implicates public interest considerations about the tradeoff between operational effectiveness and police legitimacy.

2.4 How: Mixed-Methods Methodology

How do law enforcement officers gain a domain and data understanding, and generate knowledge about what to collect, store, analyze, and engage with? In other words, how do law enforcement officers and academics create *sight* to describe and explain crime, and to understand what is legally, organizationally and technically needed to optimize the CSAE business process? Such a deep, sound understanding of reality is only possible with a robust methodology. Our methodological model, called Quadrant (see Table 2), is a mixed-methods approach that uses diverse participants and data sources from the criminal and safety & security community to produce outcomes that are valid, reliable, and credible, and therefore accurate.

Figure 6.

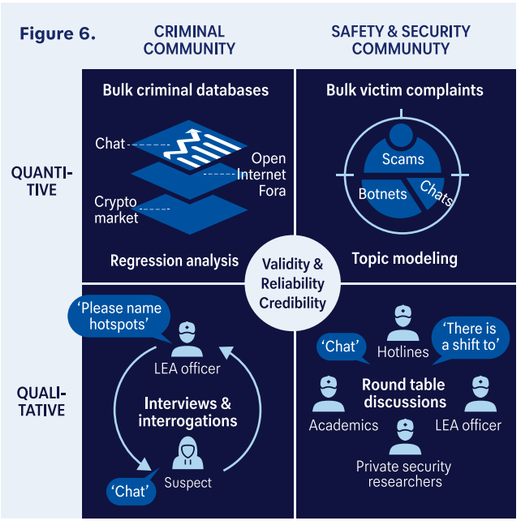


Figure 6. Quadrant methodology
This visual example includes both data sources and the corresponding methods and techniques.

Subjects & data from the criminal and safety & security communities
Key players in policing are individuals from either criminal networks or the broader safety and security community. Law enforcement agencies (LEAs) must be embedded in both types of networks to gain comprehensive *insight, oversight, hindsight* and *foresight*.

The first group includes active and inactive offenders, such as high-risk individuals, criminals, suspects, convicts and former convicts. These participants have direct experience within criminal communities, which are not homogeneous. Each subgroup generates large datasets, including financial, social and technical communications, such as money laundering databases, chat group logs or NetFlow traffic.

The second group of subjects includes those confronting organized crime as part of the larger safety and security community, such as academics, investigators, legislators, municipal officials, policy-makers, private security researchers, victims and witnesses. Each of these stakeholders also generates substantial datasets, like raw research data from academics, threat intelligence feeds from private security, or bulk complaints from victims, which LEAs can use in a quantitative manner.

Additionally, it is helpful to distinguish between *internal* and *external sources of data*. Internal sources come from within law enforcement, while external sources are obtained from outside entities. Using external sources is crucial for avoiding *blind spots* and *confirmation biases*.

New ethics, principles and values for data-driving methodologies?
While the text box in Section 2.1 explains that many existing ethics, principles and values remain highly relevant to the design and use of data-driven methodologies, we argue that data-driven policing may also require the development of new ones. Data-driven algorithms should not operate as black boxes. They must be *fair, explainable* and *transparent* to other stakeholders such as judges, legislators or the public. Additionally, unlike *random errors* by human beings, algorithms can lead to *systematic errors*. Therefore, it is crucial to conduct *model evaluations* before deployment of algorithms, and to manually review actionable results. This latter principle of human oversight can take different forms, commonly referred to as *human-in-command*, *human-in-the-loop* and *human-on-the-loop*[50]. This principle prescribes that all algorithmic models require human interaction such as manually checking the accuracy of results. Human-in-command therefore relates to *integrity* and its associated attribute of *accountability*. Because traditional accountability mechanisms within law enforcement agencies focus predominantly on the integrity of staff, digital forensic software and data, legal scholars have also pled for *algorithmic accountability* for data-driven methods and techniques as part of legal scrutiny [51] [31] [52]. At the same time, data science can also create opportunities to increase accountability by revealing previously unseen patterns in police treatment of suspects, victims and witnesses, as well as in police decision-making[32].

Mixed methods approach: qualitative & quantitative methods and techniques How should sights be extracted from participants and associated data sets? Through a mixed methods approach that combines qualitative and quantitative techniques to produce valid, reliable and credible results.

Qualitative research, familiar to most investigators and analysts, includes methods such as interviews, observations, open-ended surveys and literature reviews. This approach is used when exploring a problem, when theory is absent, or when a complex, detailed understanding of an issue is needed [53]. Since the researcher is the key instrument in qualitative research, their credibility - stemming from domain knowledge, experience and the ability to interpret results - is crucial, and relates to qualitative research trustworthiness.

Table 2: This matrix presents various examples of methods and techniques. It is important to note that the associated research designs do not necessarily stand on their own; rather, quantitative and qualitative methods and techniques can be combined in a mixed-methods approach.

Quadrant	Sources from the criminal community	Sources from the safety & security community
Quantitative methods & techniques	Topic modeling of written texts and images; time series anomaly detection on interception metadata; social network analysis of communications	Large surveys among industry experts; text clustering on victim complaints; automated indicators of compromise (IoC) extraction; threat intelligence feeds; user statistics of (forensic) software
Qualitative methods & techniques	Interrogations of suspects; debriefings of convicts; conversations with criminal informants; review of criminal writings; observations of online criminal platforms; listening to intercepted conversations	Roundtable discussions with investigators; interviews with academics; reviews of police files by analysts; short questionnaires among industry experts; literature reviews

Quantitative research expresses observations and experiences numerically and involves methods such as frequency counts, regression analyses, correlation tests and both supervised and unsupervised AI techniques. These analyses are typically conducted by statisticians, mathematicians and increasingly data scientists. In quantitative research, *validity* (measuring what is intended) and *reliability* (producing consistent results) are critical. While traditional quantitative research typically involves testing predefined hypotheses on selected datasets, AI also enables exploratory analysis - identifying patterns or correlations in data without requiring pre-formulated hypotheses.

While qualitative and quantitative methods can be used independently, the strength of the Quadrant approach lies in their combination. By integrating these methods through mixed-methods designs, purposes and strategies[54] [55], Quadrant enables law enforcement agencies to make informed strategic and operational decisions across the Collect, Store, Analyze and Engage phases. Practical examples of how to apply Quadrant in policing are provided in the next section.

A truly multidisciplinary approach
One might think that these mixed-method approaches are a two-person show between those with domain knowledge and their colleagues with numerical backgrounds. After all, they possess the necessary qualitative and quantitative skills to execute the Quadrant methodology. However, in practice, professionals with technical backgrounds play an equally important role. Digital investigators generally have a unique understanding of the technical aspects of crimes, while data engineers and software developers contribute to CSAE by processing necessary data sets and developing tailor-made forensic tools. Moreover, what we learned at the Dutch High Tech Crime Unit is that those with numerical and technical backgrounds must be treated as equals to police officers[56]. They should sit in the same operational workspace with their colleagues who have domain knowledge backgrounds, and participate in operations. Only then, everybody will grow towards the data-driven center as depicted in Figure 1 and discussed in the next section.

2.5 Who: Job Disciplines

A final core concept concerns the job disciplines involved in data-driven policing. By harmonizing these specific roles - domain, technical and numerical - the organization develops the *job capabilities* necessary to move beyond traditional forensics and into the algorithmic era. Figure 1 in Section 1.1 shows a simplified history of three interrelated disciplines within policing. The figure further explains that data engineers, data scientists and data analysts are new job titles that fit within existing job disciplines for those with a background in domain, numerical and technical expertise. While most domain experts hold police-specific roles, many technical and numerical specialists have job titles that closely resemble those found in the private sector.

Co-designing, building & using algorithmic models
We distinguish between (1) designing (i.e., formulating, conceptualizing), (2) building (i.e., developing) and (3) using algorithmic models. In general, during the first step, domain, technical and numerical experts discuss potential algorithmic models, based on the Quadrant methodology. Next these models have to be build. Data engineers and software developers then normalize necessary data sets, while data scientists develop the models. Lastly, domain experts use the data-driven models in practice.

Domain experts
Domain experts in policing include emergency dispatchers, community officers, police constables, complaint intake officers, threat assessment analysts, surveillance specialists, and traffic, waterway and airport police officers. In investigative contexts, domain experts include detectives and special agents; operational and intelligence analysts; and specialists such as financial investigators, OSINT experts, liaison officers, public–private partnership advisors, informant handlers, and various forensic professionals - including experts in ballistics, DNA, and toxicology. As described in Section 3, those with strong *data literacy* and the ability to translate quantitative sights into actionable strategic and/or operational outcomes take on new roles as data analysts or data investigators. Business operations specialists (e.g., communications, finance and HR officers) form another category of domain experts, specifically in the area of business intelligence.

Technical experts
Over the years, many technical roles have become unique to the police such as digital investigators, lawful interception specialists, incident response experts, forensic software developers, and various digital forensic professionals, including mobile, network and cloud forensics experts. There are also technical roles that closely mirror those in the private sector because the underlying skills - along with the tools, methods, and data challenges - are largely transferable. These include systems administrators, Ops, DevOps and BizDevOps engineers, full-stack and frontend developers, and cybersecurity specialists. Yet we have noticed that some of these seemingly similar roles are, in practice, quite police-specific - such as ETL developers and data engineers - since processing police data also requires a domain and data understanding of crime.

The unique data environment of police work
Most people can imagine that domain experts within the police have a unique job. After all, the police hold a monopoly over their core functions. However, many technical and quantitative roles within the police have the same job titles as in the private sector, while the actual work differs greatly because of the nature of police data. In principle, there are few legal constraints on the scale and types of data that may be collected when this serves legitimate policing purposes, such as truth-seeking investigations. In practice especially during high-profile operations police data engineers and data scientists are confronted with large volumes of highly obscure data formats that are often undocumented and poorly structured by criminal operators, while the importance and urgency of the investigations leave them no option but to process these datasets.

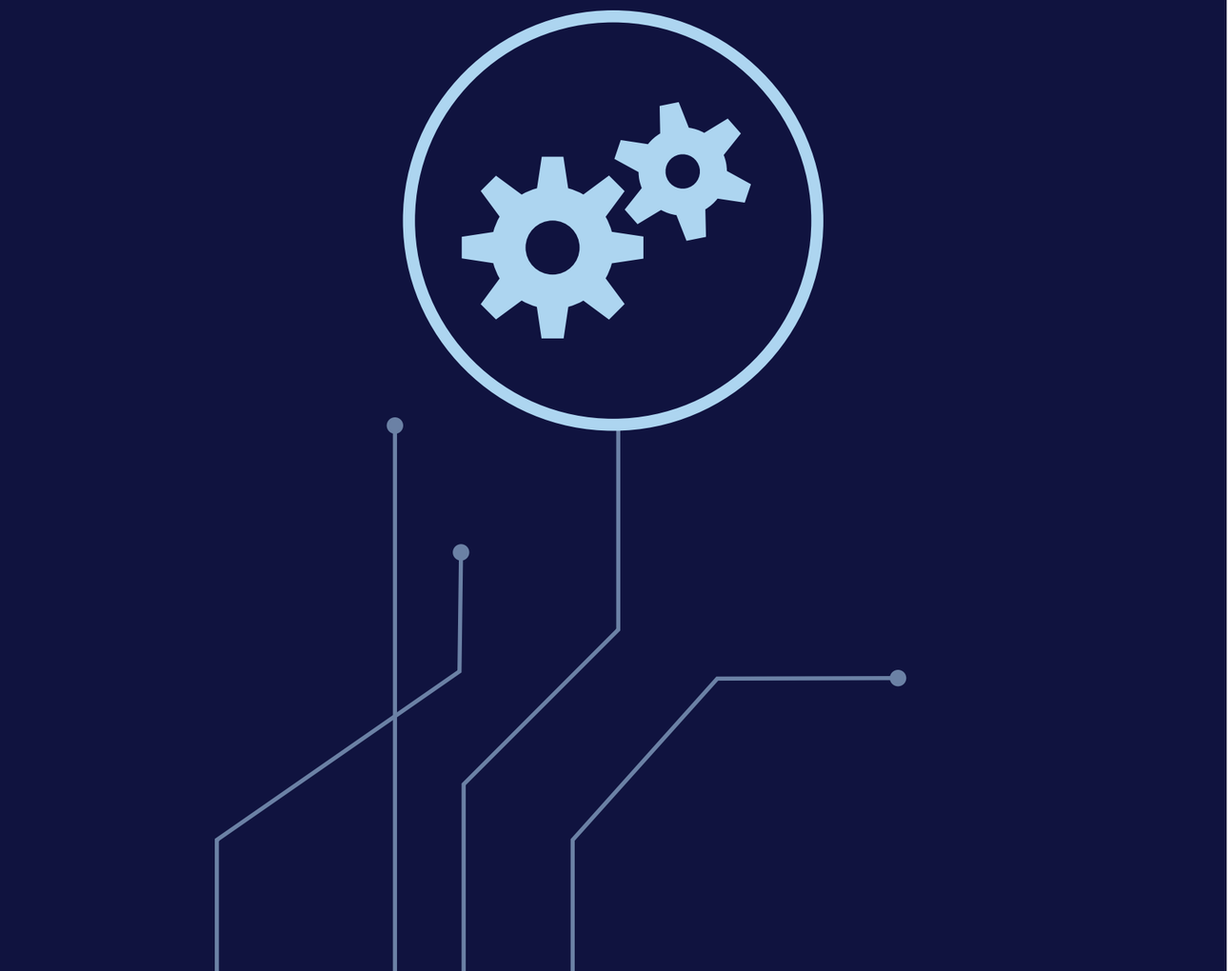
Numerical experts
Job titles include mathematicians, statisticians, and data scientists, as well as more specialized roles such as research scientists, machine learning engineers and natural language processing engineers. Not all numerical experts are responsible for developing new data-driven models. For example, applied data scientists typically run existing models, while machine learning operations (MLOps) engineers focus integrating these models into police systems. Similar to several roles of technical experts, all these job titles closely resemble those in the private sector. In the coming years, we expect that some numerical roles will also become unique to the police - for example, investigative machine learning engineers or lawful interception data scientists - as these positions demand a highly specific domain and data understanding of crime.

Management & leadership
For most law enforcement agencies, the shift to data-driven policing marks a fundamental change in organizational paradigm. Non-technical factors frequently present the greatest barriers to technical innovation - including the implementation of CSAE[57] - highlighting that data-driven policing is, above all, a social transformation[21]. This shift involves changes in organizational design, governance and structure, carried out by various managerial and leadership roles. As this is a topic in its own right, these roles and responsibilities will be addressed in a future edition.

CSAE coaches
A pivotal role is that of *CSAE coaches* (also known in the private sector as *analytics translators*[58]). These guides have extensive strategic and operational domain knowledge and sit at the center of the Venn diagram shown in Figure 1. They should launch data-driven policing initiatives and guide the various disciplines through the CSAE process, while also taking a strategic role: addressing operational challenges to management and ensuring that initiatives align with existing or new organizational strategies, programs, and projects. CSAE coaches play a critical role in leading cross-functional teams, ensuring that data-driven models are not only developed but also translated into actionable operational outcomes. Given the diverse skill set they require extensive domain knowledge, moderate numerical and technical expertise, an entrepreneurial mindset, and project management skills establishing in-house training programs is essential. Such programs are an example of the kind of organization-wide initiatives needed to support the transition toward a truly data-driven police force a topic we will explore in a future edition.

3

CSAE in Operational Practice



In this section, we place the business process at the center and demonstrate how CSAE, along with other core concepts, functions in practice. These concepts - public interest philosophy, methodology, job disciplines and harmonization objectives - are deeply interdependent. Their alignment is essential for data-driven policing.

Operationalizing CSAE is not a one-size-fits-all endeavor; it requires translating core concepts into a *local CSAE variant* that accounts for context-specific factors. We are also fully aware that transitioning to data-driven policing is not primarily a matter of understanding these concepts. In practice, the primary barriers are organizational, cultural and political - not conceptual. Success depends less on the elegance of data schemas and more on the willingness of the organization to break down silos, shift longstanding investigative habits, and secure the political mandate necessary for large-scale technical harmonization.

3.1 Gain Domain and Data Understanding

Before initiating any operational steps within the CSAE process, all job disciplines - including management - must establish a foundational strategic domain and data understanding of the criminal landscape. While leadership requires a broad strategic view, domain experts must possess a deep operational understanding of their specific crime domain and how that criminal activity manifests within datasets.

As such, this understanding is the foundation of data-driven policing in practice as depicted in Appendix A and further described in the text box below.

Gain domain understanding
Domain understanding is not reserved solely for investigators; it is a shared requirement. All personnel - including managers, data scientists, and technical experts - must possess a *strategic domain understanding*. At a macro level, this involves recognizing how contemporary crime reflects a complex, globalized landscape shaped by technological advancements and socio-economic shifts. It also requires an understanding of diverse policing philosophies, such as intelligence-led policing or problem-oriented policing, to ensure all actions remain strategically aligned. When focusing on a specific crime theme - such as sexual exploitation within the prostitution sector - domain experts must transition to a *deep operational domain understanding*. This involves the ability to describe and explain the phenomenon at a granular level. Specifically, experts must be able to generate conceptual statements regarding the 'who' and the 'what' of the crime: identifying the specific criminal communities involved and the precise nature of their activities.

Figure 7.

Figure 7. Domain and data understanding
Domain and data understanding is not a standalone phase but the continuous heartbeat of the CSAE cycle. It requires all personnel to stay synchronized with evolving criminal threats and actively share insights across the organization. By institutionalizing knowledge sharing - through lectures, briefings and workshops - agencies transform basic proficiency into deep domain and data understanding. This ongoing education ensures practitioners understand both their own tasks and the broader operational context. Ultimately, this shared understanding aligns the organization and provides the necessary 'navigational corrections' to stay on track as the criminal landscape shifts.

Crime scripts to promote strategic and operational domain understanding
A great product to express domain understanding are crime scripts. A *crime script* is a representation of a specific *modus operandi*, providing a structured sequence of criminal events and a detailed view of the offender's decision-making process[59]. While a *strategic crime script* outlines the general stages of cocaine trafficking such as production, transport, storage and distribution an *operational crime script* details the specific mechanics of a stage, such as the precise steps involved in exploiting overseas container logistics for cocaine transport. Although a crime script is essentially a qualitative product, both quantitative and qualitative sources and methods can also be used as input for developing these crime scripts (see the text box below and [60]). The idea behind starting with a crime script is that domain experts are forced to put their current knowledge to paper, find caveats in their knowledge about crime and set a verifiable knowledge benchmark. Moreover, crime scripts are necessary to create as explained in the next paragraph a data understanding about crime, and to formulate targeted interventions in the Engage phase (see Figure 13). Lastly, *scenarios* are also helpful in data-driven policing; however, we observed that this product is more effective during incidents and ongoing investigations, where the Analyze and Engage phases overlap (see text box in Section 3.4).

Gain data (and tool) understanding
Each crime theme generates its own unique trail of evidence, which is represented differently in data. *Data understanding* refers to the ability to recognize how a phenomenon manifests in data and identify related information entities. This skill is a form of *data literacy*, as it involves the ability to technically explore, understand and communicate data and information. In practice, this step is most successful when investigators use a crime script as a roadmap. For each step in the script, investigators must identify the representative information entities and the specific datasets where these entities reside.

These information points may include highly *parsable entities* like domain names, email addresses or SWIFT/BIC codes, as well as less parsable entities like names of companies, streets and persons (i.e., *named entities*). Gaining a data understanding is intrinsically linked to the Collect and Store phases. It requires police officers to expand their traditional focus - traditionally centered on intelligence and facts - and begin thinking in terms of data and information as well. Lastly, data understanding also involves familiarity with police systems and the various tools available to experts across the different CSAE phases. Tools are the essential gateways that provide access to the raw data, structured information, and synthesized intelligence required to establish facts.

Quadrant for strategic and operational domain & data understanding
Quadrant helps law enforcement agencies to gain a strategic domain and data understanding, and thus promotes strategic decision-making on what threats and objectives the organization should focus on. There distinguish two design approaches. (1) *Exploratory designs* (i.e., qualitative first) are suitable for LEA agencies that do not yet have a thorough domain 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, strategic analysts may first conduct a literature review, read intelligence reports and conduct interviews. Based on these qualitative findings, statisticians may quantify the number of historical victims complaints and/or investigations related to these threats. (2) *Discovery designs* (i.e., quantitative first). LEAs with access to large, normalized data sets such as bulk communications of criminals may have a discovery-driven starting point. They may begin with *unsupervised* quantitative approaches, using large data sets and applying advanced mathematical methods to identify patterns, before transitioning to qualitative interpretation. 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.

By utilizing these mixed-methods approaches, agencies bridge the gap between high-scale computation and nuanced human interpretation. This ensures that strategic decisions are grounded in both the 'breadth' of big data and the 'depth' of domain expertise.

Define a realistic data-driven problem
The synthesis of domain and data understanding is the selection of a specific data-driven problem. This problem may be a 'wicked' societal issue or a 'tame' operational inefficiency; it can be assigned top-down by management or identified bottom-up by experts. Crucially, a gap identified within a crime script - representing a realization that current knowledge is insufficient - is itself a valid data-driven problem, provided that data-driven methods are expected to bridge that gap and make a meaningful difference (see e.g., text box in Section 1.1).

To ensure a problem is viable, we utilize the *strategic triangle*[61]: (1) there must be support from both management and experts; (2) legal, organizational and technical capabilities must be in place; and (3) the initiative must create public value, inherently embracing a public-interest philosophy. This will be discussed in greater depth in future editions on the transition to a data-driven police organization, including topics such as organizational design and governance structures.

Translate findings into a dynamic strategic roadmap
The next step is to create a dynamic strategic roadmap that outlines a pathway for solving the problem in a data-driven manner. While data-driven policing is very much about organizational adaptability to changing crime phenomena (as described in Section 1.1), we also recognized the importance of communicating a clear vision and strategy - defining what to achieve in the Engage phase, how to achieve it, and with whom - both to management and to experts. Rather than being a fixed or lengthy document, such a strategic roadmap provides high-level direction by outlining attainable goals, concrete initiatives, and guiding criteria such as indicators of progress, timelines and the need for management commitment. At the same time, the roadmap is dynamic, constantly being updated based on lessons learned during data-driven processes, projects and programs. For example, crime themes that involve the collection of new types of data may require substantial investments in data engineering and forensic software development, with only a few partners capable of achieving the necessary technical harmonization. When these datasets are collected and stored, unforeseen legal, organizational or technical challenges may arise. Addressing such issues may demand additional resources in Analyze, while emerging data-driven insights into a crime theme may subsequently influence the roadmap and achievable objectives in Engage.

3.2 Obtain Data in Collect

As shown in Figure 8, the first step in this phase is to connect domain and data understanding to the collection of strategic data sources. Current legal frameworks generally permit LEAs to retain evidence from previous operations for the purpose of strategic decision-making and the initiation of new investigations. This cumulative approach is vital; by layering current evidence with historical datasets, LEAs build a 'memory' where the data collected today provides the essential context to amplify the effectiveness of policing tomorrow. The Collect phase encourages agencies to systematically identify large-scale datasets that are structurally required as inputs. By prioritizing these sources based on their strategic importance, agencies can build the data capital necessary to achieve long-term operational objectives and maintain a proactive stance against evolving criminal threats.

Figure 8.

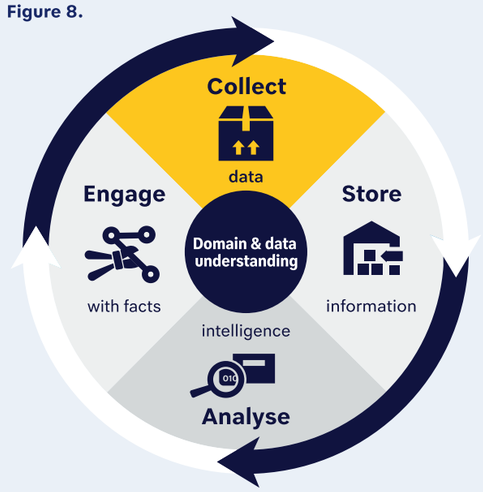


Figure 8. Collect data
Some agencies have dedicated collection teams that identify, prioritize and collect strategic data sources. In contrast, most agencies have departments that primarily support other teams in the Store, Analyze, or Engage phases by assisting with the collection of strategic and operational data sources. These supportive teams contribute specialized expertise - such as digital forensic skills - that the other teams do not possess. Regardless of the organizational setup, those who collect raw data must establish close connections with those responsible for storing and managing it.

What are strategic data sources?
Once an domain and data understanding of crime, actor communities ('who') and associated threats ('what') is established, LEAs must identify relevant strategic data sources. We define a data source as strategic when it directly assists an agency in achieving its long-term, high-level objectives.

While *strategic data sources* are distinct from *operational data sources*, the two often overlap. Operational data typically refers to evidence collected by an investigative team to build a specific case (see Section 3.4). These cases generally focus on isolated incidents and the activities of a limited number of suspects - for example, the contents of a single seized mobile phone. While vital for a conviction, such individual sources often provide significant operational value but limited strategic insight.

However, operational sources can graduate to strategic status when they provide extensive leads concerning so many important leads about other past, current and future crimes and associated communities. For instance, a large-scale seizure of mobile devices across multiple criminal organizations becomes a strategic data source when it enables the mapping of inter-group links and the identification of previously unknown high-value targets.

Strategic investigations against Dark Markets
Organized criminals rely on a number of illegitimate services and products to commit and protect crime, and they tend to centralize their activities at the same online locations for extended periods[20]. When LEAs focus exclusively on incident-driven investigations, they often face significant intelligence gaps. In contrast, identifying these digital hotspots as strategic data sources allows agencies to address the underlying criminal infrastructure. We have learned this lesson during the coordinated actions against criminal hubs like AlphaBay and Hansa Market. Through a deep strategic domain and data understanding of online drugs trafficking, U.S. and Dutch agencies recognized that targeting isolated vendors would yield minimal impact. Such vendors are often too difficult for individual investigative teams to penetrate due to the robust security measures inherent to the platforms.

To achieve their long-term objective prosecuting major drug vendors on English-speaking crypto markets the agencies first had to establish a strategic data position. These marketplaces provide a secure social, financial, and technical infrastructure for professional criminals. By seizing the market infrastructure itself rather than just chasing its users, LEAs gained access to a strategic data source that transformed thousands of isolated criminal transactions into a comprehensive, high-value dataset for future investigations.

Collect approaches for strategic data sources
LEAs often face the initial challenge of determining where to begin their collection efforts. Strategic data sources can be acquired through various channels: *nationally* within an agency's jurisdiction, or *internationally* via mechanisms such as Mutual Legal Assistance Treaties (MLATs). These sources may originate from either *public* or *private* organizations.

It is crucial to note that strategic data sources do not inherently require personally identifiable information (PII). Many essential datasets - both public and private - serve primarily to enrich abstract entities. For instance, commercial datasets often provide the geolocation of IP addresses, while public registries supply the country of origin and carrier information associated with international phone numbers.

Agencies generally adopt one of two stances when collecting these sources:

- *Reactive stance*: engaging with partner organizations or private entities that are already in possession of a strategic dataset; and a
- *Proactive stance*: actively seeking out 'data hotspots' that have not yet been collected by any party.

In proactive operations, the objective is to secure a lawful position within the criminal community's social, technical, financial, or legal infrastructure. This approach allows LEAs to intercept data at the source. Crucially, this strategy must be paired with the targeting and prosecution of the criminal facilitators who maintain the infrastructure that supports these criminal hotspots.

'Why is my HUMINT report regarded as data?'
Law enforcement agencies lawfully gather knowledge through human sources via personal contact, a practice known as Human Intelligence (HUMINT.). In our experience, HUMINT officers often view their reports exclusively as intelligence (see Section 3.4) rather than data. From the perspective of the individual author, this makes sense: they understand the context, they know which names refer to locations, and they recognize which numbers represent phone lines. To the author, the synthesized knowledge in the report is indeed a finished intelligence product.

However, from an organizational and technical standpoint, this view is incomplete. If a report is not technically processed in the Store phase, it remains raw, 'dark' data. To illustrate this, imagine thousands of HUMINT reports printed out and stored randomly in a physical room. While an individual officer can manually search through them, the absence of standardized data structures means the organization loses the ability to perform automated metadata auditing, cross-document entity linking, and the relational mapping necessary to connect separate investigative threads into a single operational picture. For the organization at large, an unprocessed report is merely a collection of uninterpreted observations. Only when these unstructured documents are ingested and harmonized can they be treated as measurable information. By applying advanced mathematical approaches in the Analyze phase such as Natural Language Processing (NLP) for automatic classification the organization can detect patterns across massive volumes of text that would be invisible to any single human reader.

Quality over quantity
In the realm of data-driven policing, the quality of datasets far outweighs their quantity[17]. We advocate for a *targeted collection strategy* that prioritizes data 'at rest' and 'in motion' within environments exclusive to the criminal community. Many digital hotspots where the underground economy thrives are naturally avoided by law-abiding citizens; consequently, these 'off the beaten track' locations contain minimal non-criminal noise[20]. By targeting these specific infrastructures, LEAs can uphold data-minimization principles while maximizing investigative relevance.

Furthermore, strategic value is found in data diversity, rather than simply collecting more of the same[62]. For instance, if an agency already possesses a large money laundering client database containing nicknames and payment records, the most valuable next step is not more financial data, but a complementary strategic source such as intercepted communications. This allows investigators to triangulate the data and shed light on the potentially illegitimate origins (i.e., criminal business activities) behind those payments.

Because the collected data is raw - i.e., uninterpreted observations and measurements[47] - the next step is to turn it 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. Of course, qualitative methods include reading text messages of criminals from existing strategic data sources e.g., decrypted communications what other platforms, products and services are mentioned. 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 domain names (i.e., websites) that are mentioned in existing strategic data sources. This identification process is followed by statistically weighting (i.e., prioritizing) the identified domain names on e.g., frequency, temporal consistency and recency. With the help of proprietary products, the domains can now be categorized into groups with labels such as 'Business', 'Illicit' or 'News' websites. The result is a categorized list with the most important websites to criminals while illegitimate Internet domains can be distinguished from legitimate domains. Law enforcement agencies may acquire an intelligence or evidence position on the most important illegitimate websites. The domains of legitimate companies provide input for public-private partnerships such as streamlining the process of sending lawful information requests or optimizing their fraud detection systems. After all, these companies are either a victim target, or their products/services are popular among criminals and thus heavily misused for criminal purposes.

3.3 Convert Information in Store

During the Store phase (see Figure 9), the collected raw data sets are normalized and subsequently converted into *information* (structured as *entities* and associated *attributes*, such as 'messages'): data that have been put in context and empowered with meaning, which gives it greater relevance and purpose[47]. 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 telephone numbers, and combinations of letters and numbers as bank accounts. This section first explains why a data warehouse strategy supports data-driven investigations as compared to data lakes. The section then describes the related and pivotal steps of extract-transform-load (ETL). Contrary to commercial databases in Store and simple analytical tools in Analyze, we regard ETL processes - i.e., tools and data schemas - as core technologies that should therefore be developed and managed by law enforcement agencies themselves as much as possible.

Figure 9.

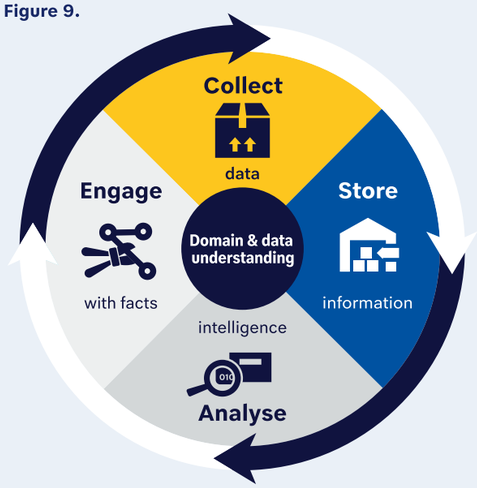


Figure 9. Store information
Staff in the Store phase do not merely process incoming data; they actively curate "entity enrichment datasets" based on their domain expertise. While data engineers initially extract abstract entities - such as domain names, IP addresses, or company names - from raw collections, enrichment adds critical attributes like geographic origin, registration timestamps, or service providers.This added context transforms isolated data points into high-quality information. By proactively layering these attributes onto abstract entities, the Store phase provides the granular, contextualized foundation necessary for advanced modeling in the Analyze phase.

Data warehouse approach
When managing collected data, law enforcement agencies must navigate the choice between a data warehouse and a data lake (see the text boxbelow for more on data lakes). For data-driven policing, we recommend prioritizing the structured architecture of a data ware-house - especially given the stringent legal and organizational mandates inherent to police work - even when immediate warehousing is not technically feasible for all datasets.

In an ETL (Extract, Transform, Load) warehouse approach, data is assessed and cleaned prior to storage. This front-end processing ensures high data quality regarding accuracy, consistency and relevance. Crucially, warehouses are better equipped to uphold legal principles, such as maintaining the chain of custody and ensuring the integrity of evidence.

A central advantage of the warehouse approach is the use of standardized data schemas across all sources. This standardization is vital for the Analyze phase. For example, ETL procedures can normalize diverse data from SMS, email and instant messaging under a unified 'messages' attribute, while simultaneously tagging whether the source is from a suspect, victim or witness. This allows data scientists to seamlessly aggregate and analyze suspect

communications across multiple platforms for tasks like authorship attribution.

Furthermore, a standardized warehouse fosters organizational trust. Because police agencies rely on numerous domain experts and analysts distributed across different units, the warehouse ensures these users only need to understand a single, common data schema. They can act with the certainty that the data has already been vetted and structured.

While developing and managing a data warehouse is more labor-intensive than a data lake, these costs are mitigated through technical harmonization. By adopting common data schemas and ontologies - as discussed in Section 3.1 - and sharing the burden of data structuring with partner agencies, the long-term benefits of interoperability far outweigh the initial investment. The next paragraphs describe the ETL process in practice and in detail.

Data lakes may fit within a larger Store strategy
While legal requirements such as identifying privileged communications mandate structured data, technical and organizational constraints often prevent immediate warehousing of all datasets. Therefore, we recommend that LEAs develop a strategy to determine when data sets should be warehoused versus when they should be stored in a data lake. In a data lake, all data are retained in their raw, often unstructured form regardless of their source or structure, and are only transformed when needed for use. Data lakes follow an extract-load-transform (ELT) process.

To mitigate this, LEAs must shift their focus to the point of origin. By enforcing organizational schemas at the source either through vendor alignment or internal development data can bypass the complexities of post-hoc normalization and flow directly into the Store layer. This *schema-on-write* approach not only reduces storage overhead but also ensures that data-minimization principles are upheld before the data ever enters the system.

Extract data from source
The initial step involves extracting data from physical or digital hardware — a vast spectrum ranging from client databases of crime-as-a-service providers, intercepted telecommunications, suspects' mailboxes, malicious software, GPS antenna data of police cars, video material and written reports from covert observations.

However, *full extraction* of complete data sets is not always necessary and is often undesirable due to the legal principle of *data minimization*. Therefore, *partial extraction* strategies are preferable, focusing only on data segments that contain entities relevant to solving historical, current and future crimes, while proactively filtering out legally protected content, such as attorney-client privileged communications.

Once the relevant segments are isolated, the focus shifts to entity extraction. Traditionally, this has been a manual task where officers structure inherently unstructured text (such as the HUMINT reports discussed in Section 3.2). This requires the manual addition of metadata - such as source origin - and the laborious linking of entities and their relationships within the text.

In the modern landscape, many entities exist within structured metadata (e.g., timestamps, IP addresses, and domain names), which automated tools can easily ingest into separate database tables. However, critical entities are often 'hidden' within unstructured text attributes, such as a criminal mentioning a specific weapon type or a hidden transaction code in a chat message. Because manual linking is prone to human oversight, *automated entity extraction* is essential. Automated entity extraction identifies these hidden entities within unstructured strings, ensuring that vital information is not lost in the sheer volume of raw text.

ETL as core technologies
While ETL (Extract, Transform, Load) is generally a straightforward process, CSAE has a distinct public interest philosophy regarding data storage in the Store phase. Traditionally, law enforcement has relied on proprietary products for Store, resulting in a variety of tools, processes, and standards for performing ETL. Explainability is crucial for law enforcement, yet there is a risk that ETL processes and associated algorithms, data schemas and tools become black boxes due to *commercial confidentiality*[32].

For instance, third-party tools might prioritize or withhold information without the larger criminal justice system being aware of these built-in decisions. Criminal investigations require a flexible range of tools to process ever-changing data sets, without depending on private monopolies or facing substantial switching costs. Therefore, data schemas, ETL tools, and associated processes should be considered *core technologies*, owned and managed by law enforcement agencies themselves.

Our approach to Store not only enhances legal explainability and chain of custody but also improves organizational maneuverability and technical agility. Ultimately, this approach increases the independence of LEAs from the private sector. After all, decision-making power depends on who controls the code[63].

Transform data into information
Following extraction, entities must be transformed into a uniform format to ensure system-wide interoperability. In raw datasets, the same entity - such as a telephone number - is often recorded in varying formats by different suspects or systems. To resolve this, extracted entities are cleaned and standardized according to a predefined data schema or ontology. This involves, for example, formatting all telephone numbers with international country codes and a fixed digit sequence, deduplicating records, and synchronizing time zones to ensure accurate chronological timestamps.

During this transformation phase (see Figure 9), the data is further refined through relationship mapping and enrichment. A primary example is the identification of the telecommunications provider associated with a specific number. Such enrichments highlight the necessity of maintaining strategic datasets - like provider registries - to add immediate context to raw entities. By transforming these isolated attributes into enriched information points, the system prepares the data for the complex relational queries required in the Analyze phase.

Core technologies: a continuum from open to closed source
Technical harmonization both internally and with external, trusted partners depends on standardized ETL algorithms, data schemas and processing tools. We argue that these *core technologies* should be managed in-house as a strategic continuum, reflecting the public interest philosophy. This continuum ranges from fully open source to closed-source internal models.

For instance, standardized cleaning algorithms, parsers for common file formats (e.g., CSV, JSON, or standard XML) or a user interface can be open-sourced to allow for public review and community-driven improvement. In contrast, proprietary forensic tools are best suited for a 'closed-open source' model transparent only to a trusted community of law enforcement partners, legislators and the judiciary.

Finally, certain technologies, such as specific data schemas and ontologies, must remain strictly closed source. Experience shows that large-scale operations against organized crime lead to a continuous expansion of these schemas. Because criminals actively seek to map law enforcement capabilities[20], any disclosed update to a schema could inadvertently reveal strategic focuses or active investigative directions.

Load information into specialized databases
Once raw data is transformed into information, the normalized evidence is distributed across a suite of specialized databases. Because no single system can answer every investigative question, these databases are selected based on the specific analytical needs of the Analyze phase. For instance, relational databases manage structured case data, while graph databases map complex relationships between entities. Text search engines prioritize document retrieval, content stores handle high-volume multimedia (video, audio, images), and spatial databases process location-based intelligence. This multi-database approach ensures that each query is executed in an environment optimized for speed and accuracy.

Beyond technical performance, databases are also shaped by various legal and organizational requirements. For instance, legal frameworks often dictate rules about the retention or removal of evidence after a certain period or the linking of information points from different investigations. Organizational demands also play a role, such as the need for law enforcement agencies to have different authority levels for accessing evidence. This necessity creates a functional synergy between the Store and Analyze phases: ETL processes are not isolated technical tasks but are designed in close collaboration with analysts and investigators to ensure the output directly serves the analysis.

In this ecosystem, data from diverse proprietary and open-source tools is normalized and stored to serve a wide array of downstream applications. This *pluggable architecture* fosters interoperability and avoids vendor lock-in, driving innovation across the commercial forensic software market. By maintaining an independent warehouse and in-house ETL tools, LEAs gain the flexibility to ingest or export evidence across any platform. Obsolete databases can be phased out without systemic disruption; integrating a new tool only requires a single 'adapter' to bridge the common data schema with the new environment, rather than a complete overhaul of the data pipeline.

Quadrant in the Store phase
The decision of which data to extract from raw datasets into the warehouse is guided by our Quadrant approach, which balances legal necessity against technical and organizational feasibility. For instance, a legal analysis of case law might reveal that images from seized mobile devices are only relevant for proving specific categories of crime. Simultaneously, consultation with warehouse experts might highlight that storing such high-volume visual material consumes disproportionate resources, or that existing tools lack the capability to automatically extract entities- such as locations, faces, or text from those images.

Once the scope of extraction (full or partial) is defined, the focus shifts to identifying specific entities of interest. A *triangulation design* can then be employed to confirm or corroborate findings across different sources. For example, if an academic researcher mentions a new payment method used by career criminals during an interview, a targeted database query can verify whether this method appears in intercepted communications.

This scenario underscores the importance of the *feedback loop*: domain knowledge experts must identify the specific format of the associated transaction codes, enabling data engineers to precisely extract and transform those codes from the warehoused datasets. By turning these raw strings into structured information points, the data is transitioned from the Store phase to being fully operational for the Analyze phase.

3.4 Create Intelligence in Analyze

After establishing domain and data understanding - and successfully completing the Collect and Store phases - the organization begins the transformation of information into *intelligence*. The Analyze phase is not a linear path but a continuous cycle characterized by reduction (filtering out the noise) and enrichment (adding context and depth). In this section, we will demonstrate how this is done in practice across three interrelated processes within Analyze: (1) research, (2) pre-investigations and (3) investigations. In the example provided in this section, the output of a research reduction model serves as input for a preparatory investigative enrichment model. However, in practice, the cycle of reduction and enrichment is typically repeated multiple times within each distinct process of research, pre-investigations and particularly in investigations.

Figure 10.

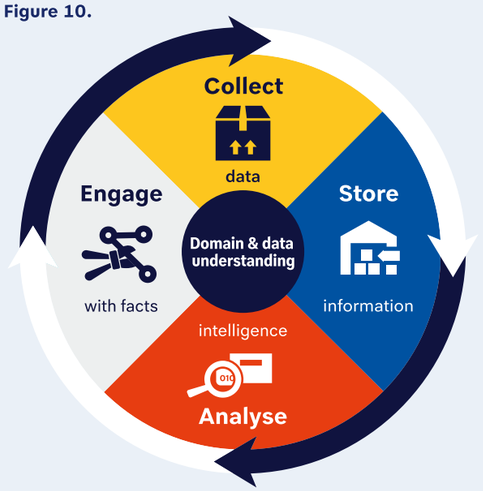


Figure 10. Analyze intelligence
To develop and use analytical intelligence models that ultimately inform the Engage phase, those in the Analyze phase must, based on their domain and data understanding, collect relevant data sources, including data already collected by others. They must pinpoint the essential information entities within these datasets and collaborate closely with the Store phase teams to ensure these entities are properly extracted, structured and indexed for analytical use. This ensures that the data architecture is directly aligned with the investigative needs, moving the organization from passive data storage to purposeful, model-driven intelligence.

Analyze: creating and using intelligence in research, pre-investigations and investigations processes
Research in the Analyze phase focuses on studying the nature and extent of crime and criminal behavior at a global and societal (macro), community (meso) and group/individual (micro) levels. At this stage, intelligence is rooted in theoretical frameworks and empirical findings, rather than a specific legal context. Analysts move from high-level hypotheses to specific associated targets for example, shifting from studying international narcotics transport trends to identifying the most prominent facilitators of maritime cocaine trafficking. The high-level targets are then reprocessed during *pre-investigations*. While the workflow remains similar, the data sources and entities change as the analysis is reframed to create intelligence for a legal context. The process determines which of the top facilitators are most suitable for criminal investigations, ultimately producing a target package for the Engage phase, based on four criteria: (1) does the target meet the elements of crime?, (2) is there organizational and territorial jurisdiction?, (3) what are the short-term investigative opportunities?, and (4) what are the long-term investigative goals as outlined in the Engage phase? The final stage the creation of intelligence for *investigations* occurs at the intersection of Analyze and Engage. This intelligence is based on evidence collected against suspected top facilitators. It serves to guide the overall direction of the investigation, generate investigative leads, and transform analytical hypotheses into substantiated legal facts.

Reduce information to targets for research purposes
In the Store phase, billions of information points exist, representing thousands of potential targets. To navigate this, the first step in Analyze is reduction. It is essential to recognize that even the most basic investigative actions, such as simple keyword queries (‘Ctrl-F’), are fundamentally forms of data reduction. Yet these queries will generally not identify main crime themes, organized crime groups or key players in complex networks, but rather the low hanging fruit: those individuals who are unable to properly protect their crimes and identity (see the text box below and Figure 11)[20].

To find these deeper patterns, analysts use research reduction models. These models consist of calculations that shed light on the who, what, when, where, why and how (5W1H). The output of these reduction models generally consist of a target list with one or more entities related to the 5W1H, such as a place name for the where, a specific word of an MO that relates to the how and/or a nickname for the who. The next step is to enriching the discovered entities with additional information, and - in this example - put them in the specific legal context of the preparatory investigative process within Analyze.

It is important to note that reduction and target selection are not always performed by the analyst. In many cases, the target is already defined by external factors, such as victim complaints with suspect identification, or when incoming information requests from other agencies concerning a specific target. In these scenarios, investigators and analysts bypass the reduction stage and begin directly with the enrichment process to build out the case.

Figure 11.

Figure 11. Intelligence analysts versus criminal investigators
In our experience, the work of intelligence analysts is often not fully utilized by criminal investigators. This disconnect arises from a fundamental discrepancy: intelligence analysts may identify key existing threats that are considered non-actionable, while investigators tend to focus on different targets driven by immediate operational needs. Without a unifying structure, critical insights are frequently overlooked, and the two groups operate in silos. The CSAE framework helps mitigate this issue by providing a structured process, objective quantitative measures and clear harmonization objectives, thereby creating a stronger flow between the Analyze and Engage phases.

Targeting high hanging fruit that is easy to pick
Data-driven analytics, when integrated into the research and pre-investigation phases, achieves superior results compared to manual target selection. Crucially, it resolves the historical tension between intelligence analysts and criminal investigators two groups that often differ in their methods, objectives and organizational cultures as shown in Figure 11. Manual selection is highly dependent on individual experience and is vulnerable to the sunk-cost fallacy, where teams persist with a specific target despite a lack of supporting data simply because of the resources already invested. This misalignment often leads to analysts aiming for ‘high-hanging fruit’ that is strategically significant but operationally unreachable, while investigators gravitate toward ‘low-hanging fruit’ that is easier to catch but has a lower impact on the criminal network. Our data-driven approach mitigates these inefficiencies and reduces confirmation bias by introducing objective scoring and structured workflows.

In the research phase, models first objectively identify the high-impact high-hanging fruit that would otherwise remain invisible. This output then feeds into the pre-investigation phase, where targets are refined to determine which of those high-impact actors are actually accessible and ‘easy to pick’ for a formal investigation. Because these models are integrated with objective scoring criteria, target selection becomes highly scalable, providing structured guidance for even less experienced personnel to ensure the organization remains objective and effective.

Enrich targets with information for pre-investigative purposes
Enrichment in Analyze is learning as much as possible about the entities from the reduction models by linking them to internal *information* stored in the warehouse (Store) and *external information* from closed and open Internet sources (OSINT), private vendors or public organizations. The objective of the pre-investigations process is to assess whether a target is suitable for further investigation. Of course, other functions of policing would enrich targets for their own respective purposes.

Although the data may originate from the same sources used in the research reduction models, data for pre-investigation enrichment models is thus analyzed from a different, specifically legal, perspective. In our experience, these quantitative models serve as an initial step in a more extensive workflow. The models alone may not fully meet all pre-investigative criteria (see text box above), so qualitative methods and techniques - such as manual searches - are also necessary to complete the assessment. The actions

in this workflow often rely on the 5W1H framework as well. For example, determining who the suspects are, what crimes they commit and where they are located can establish elements of the crime, jurisdiction, investigative opportunities and investigative goals.

The primary output is a target package for investigations in the Engage phase, consisting of visual network charts - an *oversight* of suspects, victims, and legal, financial or technical infrastructures and accompanying reports that provide deep *insight* about these charts. By evaluating the four pre-investigative criteria within these documents, analysts create a structured foundation for formal investigations.

Crucially, police organizations should invest in technically cycling these intelligence products back to the Store phase. By adding 5W1H-based metadata and labels - a practice known as *annotation* - analysts ensure that when others encounter these entities, they are immediately linked to established clusters. This process simultaneously provides the high-quality labeled data necessary for future machine learning models to detect similar criminal patterns automatically.

Quadrant in the Analyze phase
After data scientists, data engineers and domain experts identify accessible data sources and entities, they begin exploring research designs. The power of the Analyze phase lies in the ability to combine components of the 5W1H framework to create granular reduction models. For example, an automated role identification model (the 'what') might use natural language processing to detect unique money laundering slang. When this is combined with timestamps ('when') and travel movements ('where'), the model filters the data to isolate only money launderers with a specific modus operandi during a particular time period.

These models act as building blocks that can be used independently or sequentially. The output of one model such as a list of identified money launderers serves as the input for the next. A subsequent model might then ask: "Who are the most popular money launderers among synthetic drug producers?" by analyzing the frequency of mentions of these money launderers in intercepted communications of producers.

The next step is to enrich the profiles of these top money launderers. Again, a group discussion with domain experts can determine which internal and external quantitative data sources should be added. For example, a 'where' model might extract all locations mentioned in intercepted written communications of each money launderer. Associated IP addresses could be enriched with external commercial data sources providing insights such as connection type, GeoIP location and the name of the access provider. Not all quantitative sources will yield useful results or any results at all. A qualitative approach fills the gaps where quantitative data fails by using OSINT findings, interviews with informants, and manual validation of automated results. By treating these models as modular units, the Analyze phase becomes a dynamic environment where intelligence is refined through a continuous cycle of filtering and context-building.

Reduction & enrichment for Engage purposes
So far, we discussed creating intelligence for research and pre-investigative purposes. *Creating intelligence for investigations* relies on the continuous cycle of reduction and enrichment as well, and occurs at the intersection of the Analyze and Engage phases. *Enrichment* of entities related to suspects and crimes with internal and external information generally leads to investigative opportunities. A suspect may be linked to a bank account, phone number or website. The deployment of associated investigative powers to exploit these opportunities leads to *operational collects*, such as sending financial information requests, wiretapping a mobile phone or preserving a server. 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. In this stage, investigators formulate and test new hypotheses, making strategic decisions about which entities to add or remove based on investigative reasoning and available resources. The remaining entities undergo further enrichment, triggering a new wave of investigative powers in a continuous cycle. This process repeats until the intelligence is successfully distilled into the factual statements required for legal objectives. Ultimately, this cycle serves as the bridge where the analytical process overlaps with and transitions into the final phase of Engage.

Investigative reasoning: transforming data-driven intelligence into facts
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. These forms of investigative reasoning are pivotal to transform intelligence into facts. 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. Ideally, data-driven investigations apply all three methods in an integrated manner. A scenario provides the narrative structure, while arguments are used to support or reject these scenarios with evidence. Arguments for 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 [64]. When new leads emerge, investigative powers are deployed to collect fresh evidence, allowing the team to confirm, revise or reject their original theories. This rigorous cycle ensures that the final output is not just a collection of data points, but a verified and cohesive attribution of who did what, providing the structural foundation necessary for the legal demands of the Engage phase.

Coordinate & evaluate processes
Internal checks and balances that ensure accountability, auditability and scrutiny have traditionally been embedded in most law enforcement workflows, including research, pre-investigations and investigations within the Analyze phase. These *coordination mechanisms* involve taking inventory of legal, organizational and technical resources and assessing associated opportunities, challenges and risks when transitioning between stages. Effective coordination ensures that the move from broad research to a targeted investigation is supported by the necessary resources and legal authority.

Part of this coordination involves *process evaluations*, which examine whether workflows are being executed as planned, identify any deviations, and assess the quality of implementation. These evaluations are essential for enhancing the accuracy of data-driven policing. Distinguishing between theory and *implementation failures* is crucial. Both types of failures result in unexpected outcomes, but they differ in nature as explained in the text box below. Identifying the specific type of failure allows for targeted improvements to data-driven policing.

Preventing theory and implementation errors
Implementation failures are linked to poor execution, such as failing to collect strategic data sources due to a lack of trained staff. Conversely, theory failures occur when processes are executed correctly, yet the expected results are not achieved because the underlying logic or models are flawed. This could happen if the methods and techniques used are not valid or reliable, leading to *measurement errors*.

In traditional, predominantly qualitative investigative methods, investigators and analysts themselves are the primary analytical instruments. Errors in such cases are often random and unavoidable, stemming from individual mistakes, failures and accidents. However, with the shift to data-driven policing using algorithms, *systematic errors* can occur. These errors are consistent and repeatable, making process evaluations vital in data-driven policing. These evaluations must be applied to the entire business process: the data input in the Collect phase, the processing in Store and Analyze, and the operational output in the Engage phase

3.5 Execute Fact-Based Interventions in Engage

Once the investigation cycle in the Analyze phase reaches a sufficient level of detail, and intelligence is converted into concrete facts through investigative reasoning, operations can move towards executing their objectives. Lawful interventions against crime are grounded in *facts* - statements about reality that are beyond reasonable doubt and adhere to fundamental principles such as *accountability, due diligence* and *neutrality*. It is important to note that policing and interventions can be carried out by various organizations beyond traditional law enforcement. In these scenarios, the police may act in a supporting role for these stakeholders to execute successful, law-based interventions.

Figure 12.

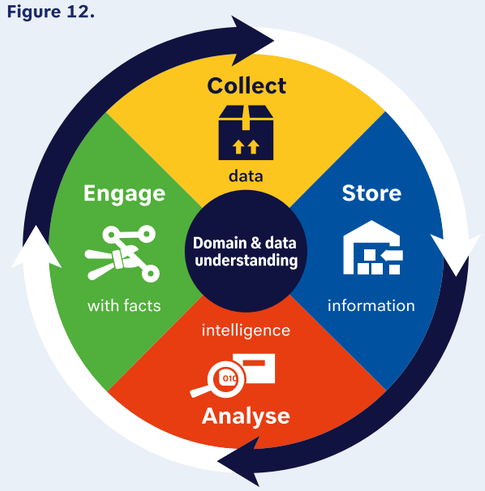


Figure 12. Engage with facts
Those who work in the Engage phase may collect data - sometimes supported by specialized collection teams with expertise such as digital forensics - during criminal investigations and other operational activities. The collected data is then normalized by data engineers in the Store phase and analyzed either by dedicated operational analysts within the Engage team itself or by intelligence analysts in the Analyze phase. These analyses provide intelligence and/or factual evidence that ultimately inform both strategic decisions and operational outcomes such as prosecuting suspects of crime.

A continuum of preventive and repressive engagements
Traditionally, police work has centered on victim protection and repressive actions. In investigations, the primary focus is often attribution: identifying the perpetrator for the purpose of prosecution, with the goal of achieving a punitive sentence. However, as explained in Section 1.1, the modern criminal landscape complicates this, and investigations even face an effectiveness crisis.

Recognizing these limitations, national law enforcement agencies in countries like Australia, Canada, Great Britain, and the Netherlands have expanded their scope[65] [66] [67] [20]. They have developed additional outcomes that move beyond mere attribution. In this context, prevention and repression are not mutually exclusive; they exist as two ends of a continuum. Our framework outlines four related outcomes that span this entire spectrum.

These outcomes are relevant across every phase of the CSAE cycle. Every data point collected, every entity extracted, and every analytical model applied should contribute - directly or indirectly - to achieving these goals. By identifying these outcomes early, specifically during the transition into pre-investigations, investigators can decide whether the goal is to build a case for court (repression) or to proactively inform the general public about crime (prevention).

The snowball effect of a multistakeholder approach & multiple outcomes

In practice, multiple outcomes of CSAE are interconnected and mutually reinforcing. A single investigation can trigger a chain reaction where preventive actions create new opportunities for repression. For example, during an investigation into an advanced cyber attack, law enforcement officers and private security researchers collaborated to produce a joint threat analysis for financial institutions. The analysis included the modus operandi (MO) of the cyber threat including indicators of compromise and what to do against this threat. The Dutch National Cyber Security Center sent the report to similar national governmental computer emergency response organizations (GovCERTs) in other countries. These organizations then distributed the report to their national financial institutions. As a result, banks were able to implement necessary preventative countermeasures against the threat. Some financial institutions, upon reviewing the enclosed evidence, discovered that they were not merely potential victims but had already been compromised. They subsequently sought assistance from the private security company that had contributed to the report. Identifying victims and helping them remove infections before any financial harm was done served a dual purpose. Not only did it assist the victims, but it also disrupted the organized crime group's operations. The criminals had made significant upfront investments to infect and monitor the machines, but were prevented from profiting. Despite these interventions, the suspects continued their criminal activities, albeit with an altered and less effective MO. This forced adaptation led to new mistakes, failures and vulnerabilities, which in turn provided further evidence against them.

Prevent harm for vulnerable individuals, groups, organizations & sectors

The primary preventive outcome is *harm prevention* for *vulnerable individuals*, *groups, organizations* and *entire sectors*. Sue to specific inherent characteristics, certain actors may be at risk of becoming involved in crime - as a victim, witness and/or even as a perpetrator (for the latter, see text at the end of this section). Law enforcement often possesses unique insights into these vulnerabilities and can identify exactly what is needed to bolster resilience. This includes direct interventions, as well as broader efforts to empower 'capable guardians' within a community. For instance, police data may reveal specific recruitment techniques used by organized crime to target youth or identify industries facing imminent ransom threats. In these cases, raising awareness among the at-risk parties is a crucial intervention. The framework also adopts the principles of *situational crime prevention*, focusing on the environment in which crimes occur rather than only the individuals involved. A key outcome is harm prevention through *target hardening*: strengthening the resilience of stakeholders before threats escalate. For example, financial institutions can be informed about the ATM blow-up methods used by cash machine robbers to help prevent such attacks. Thus, the goal of harm prevention is to generate actionable outcomes for a range of public and private actors. This includes issuing warnings through media outlets for the general public, providing threat analyses for private industries, and offering actionable insights to potentially vulnerable individuals, groups, and organizations about emerging threats.

Help & protect victims

The next objective is *helping and protecting victims*, a core responsibility of law enforcement. This outcome focuses not on potential victims and threats, but on *actual victims* and ongoing criminal activities. The process involves detecting and identifying victims, informing them about their victimization, protecting them and understanding the underlying methods of operation (MOs) used in criminal activities to prevent further harm.

In these situations, the role of the police is to enhance the safety and security of citizens, public institutions, and the corporate sector by mitigating further damage. For example, during a cyberattack investigation, law enforcement may discover victim data that the targets themselves are not yet aware of. Proactively notifying them prevents additional exploitation and often turns a passive victim into an active partner; once informed, they may file formal complaints or provide additional data about the attacks.

This outcome often creates a critical bridge back to the earlier phases of the framework: once notified, victims may file formal complaints or provide additional data about the attack. This *victim-sourced intelligence* provides fresh insights into the criminal's MO, which can then be fed back into the Analyze phase to strengthen the overall case.

Disrupt criminal business processes

Building on the previous two victim-based approaches, the first offender-based approach focuses on disruption: hindering the processes of committing and protecting crime. Police interventions, such as investigative powers, are always disruptive to criminal modus operandi (MOs), even when not actively applied, because the mere threat of investigation can affect the criminal business process[20]. Disruption can take several forms and may go beyond measures that merely skim profits from criminal activity, such as seizing financial assets like cash or cryptocurrencies. Disruptive actions may increase the costs of committing and protecting crime [68], forcing professional criminals to either overspend on security or underprotect valuable assets. Other interventions may target the broader underground economy by creating situations in which criminals have incomplete or misleading information[69]. In this way, disruption can also influence criminals to act in certain ways or refrain from specific behaviors.

Quadrant in the Engage phase

In the Engage phase, the Quadrant helps in determining an effective and holistic intervention plan against specific crimes. For example, academics from the safety and security community conducted a research project for the Dutch national police aimed at increasing the operational costs of organized criminals involved in committing and protecting their criminal activities. They analyzed financial transactions from a seized client database of a web hosting service that ignores abuse complaints and allows illegal content to be hosted (also known as bulletproof hosting). The researchers revealed that profit margins were very small. They concluded that increasing transaction and operating costs as a pressure point could disrupt revenue and demand [70]. When suspects are apprehended, the Quadrant approach changes the nature of the interrogation. Instead of focusing solely on guilt, interrogations of cooperating suspects can be used to uncover and exploit previously unnoticed financial weaknesses in their methods of operation (MOs). his creates a powerful feedback loop where research data and interrogation insights are combined to create new variables, measurements, and datasets. This process of *data fusion* allows for further research into *restrictive deterrence*, which analyzes how increased risks or costs influence offenders to reduce the frequency, scale or severity of their operations. Consequently, the Engage phase becomes a source of high quality data that feeds strategic and operational domain and data understanding, ensuring that future interventions are even more precisely targeted.

Impose penalties against offenders & suspects

The last and most repressive goal of policing is imposing penalties against offenders and suspects of crime such as prosecution and alternative sanctions. 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. While traditional prosecution remains the standard, the operational domain also utilizes alternative sentences for low-threat offenders - such as young offenders who have committed an actual offense but do not yet pose a high risk to society. Measures such as community service, financial transactions, fines, official warnings and probation still hold punitive effects and are considered repressive in nature. However, these sanctions are less far-reaching than full incapacitation. They allow for a strategic emphasis on rehabilitation, restitution and restorative justice, aiming to redirect the offender's behavior before they escalate into more serious criminal activities.

In the context of the CSAE framework, this outcome requires the most rigorous data understanding. Every piece of evidence - from the initial extraction to the consolidated research datasets - must be prepared to meet the high evidentiary standards required for legal sanctions.

Predictive policing and potential offenders

It is important to note that the prevention of 'potential offending' does not fall within our continuum, as this would conflict with our public interest philosophy. As we already mentioned in Section 1, predictive policing remains controversial, mainly due to concerns about predicting criminal behavior at the individual level. Instead of viewing "potential offenders" through a lens of future criminality, they should be regarded as vulnerable individuals and groups. Research consistently shows that those at risk of recruitment into crime often face a range of systemic life challenges. By shifting the perspective from 'pre-crime' to 'vulnerability', law enforcement can align its actions with the harm prevention goals discussed earlier. Beyond ethical objections, the European legal principle of purpose limitation creates significant restrictions on the data available for such predictions. Many relevant datasets are collected during criminal investigations focused on suspects of crimes, and the use of this evidence is legally prohibited for targeting high-risk but still law-abiding individuals and group§s.

Evaluate impact

Ultimately, the CSAE framework aims to enhance the effectiveness of policing and police legitimacy. To achieve this, periodic and objective *impact evaluations* are essential for measuring the immediate effects of specific Engage interventions against the threats identified during the domain and data understanding process (see Section 3.1). Unlike the process evaluations discussed in Section 3.4, which focus on the implementation process, impact evaluations assess the actual causal relationship between a police intervention and its outcome.

Impact evaluations are closely tied to process evaluations; if the implementation of CSAE is flawed, determining the effects of data-driven policing becomes challenging.

Engage Matrix	**Prevention** ◄——► **Repression** **Many legal opportunities** ◄——► **Few legal opportunities** **Scalable interventions** ◄——► **Labour-intensive interventions**			
Contribute to Engage ➡ Steps in Crime Script ⬇	Prevent harm for vulnerable individuals, groups, organisations & sectors	Protect victims	Disrupt criminal processes	Impose penalties against offenders & suspects

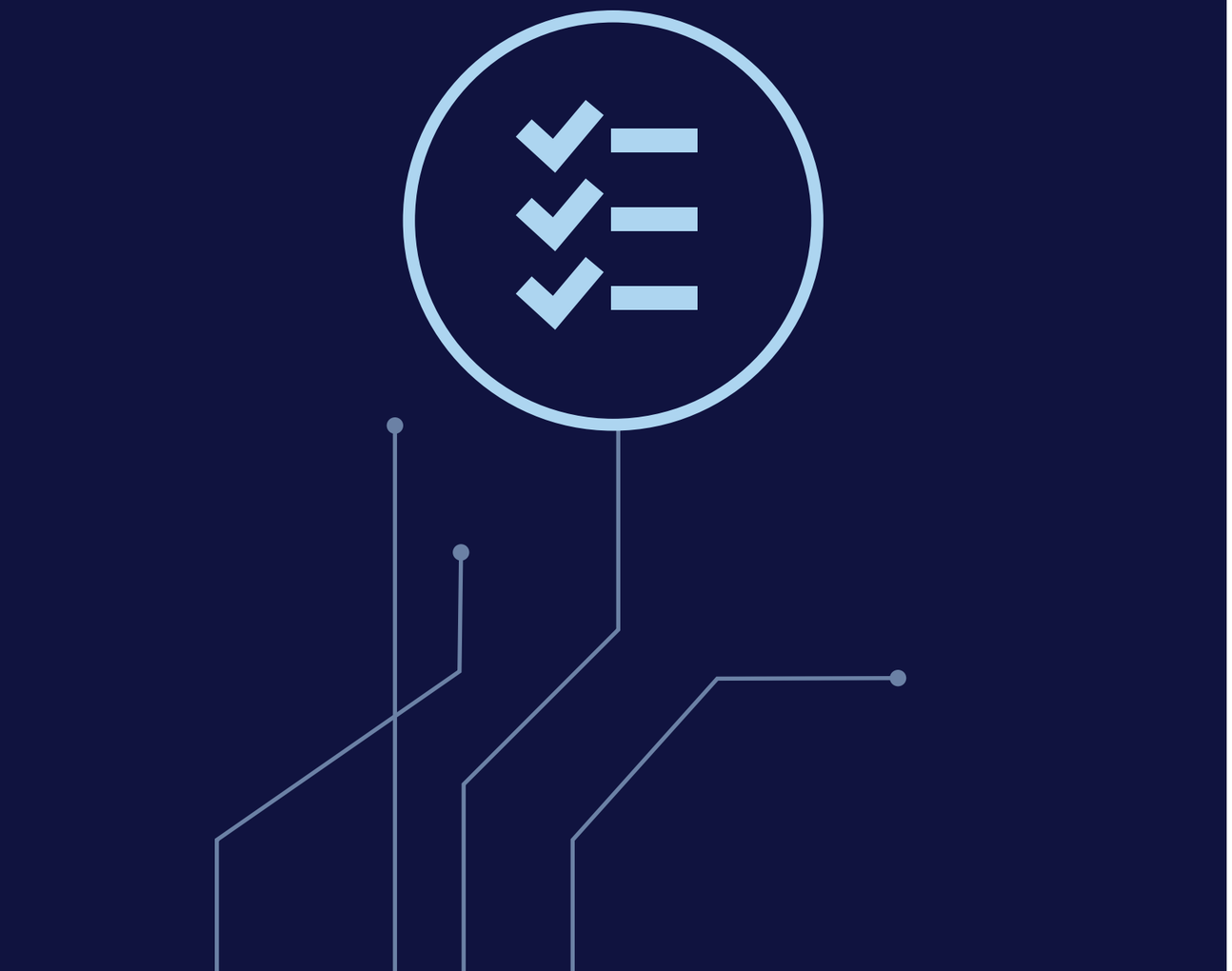
Figure 13. Engage matrix
In Section 3.1, we emphasize the importance of crime scripts to structure domain knowledge. Our Engage matrix links crime scripts (i.e., domain knowledge about a particular crime type) to multiple, interrelated interventions, enabling a fully integrated approach to combating crime. In general, preventive actions offer more legal possibilities and tend to be more scalable than repressive interventions.

The results of these evaluations support *evidence-based policing* and offer insights to improve data-driven policing. They also help identify the most cost-effective interventions, aiding in the efficient allocation of scarce resources.

These cause-and-effect assessments often use mixed methods, underlining the need for law enforcement to develop practical skills in applying the Quadrant. Ultimately, impact evaluations support CSAE's public interest philosophy by showing the tangible outcomes of data-driven policing and strengthening institutional accountability within liberal democracies.

4

Conclusion



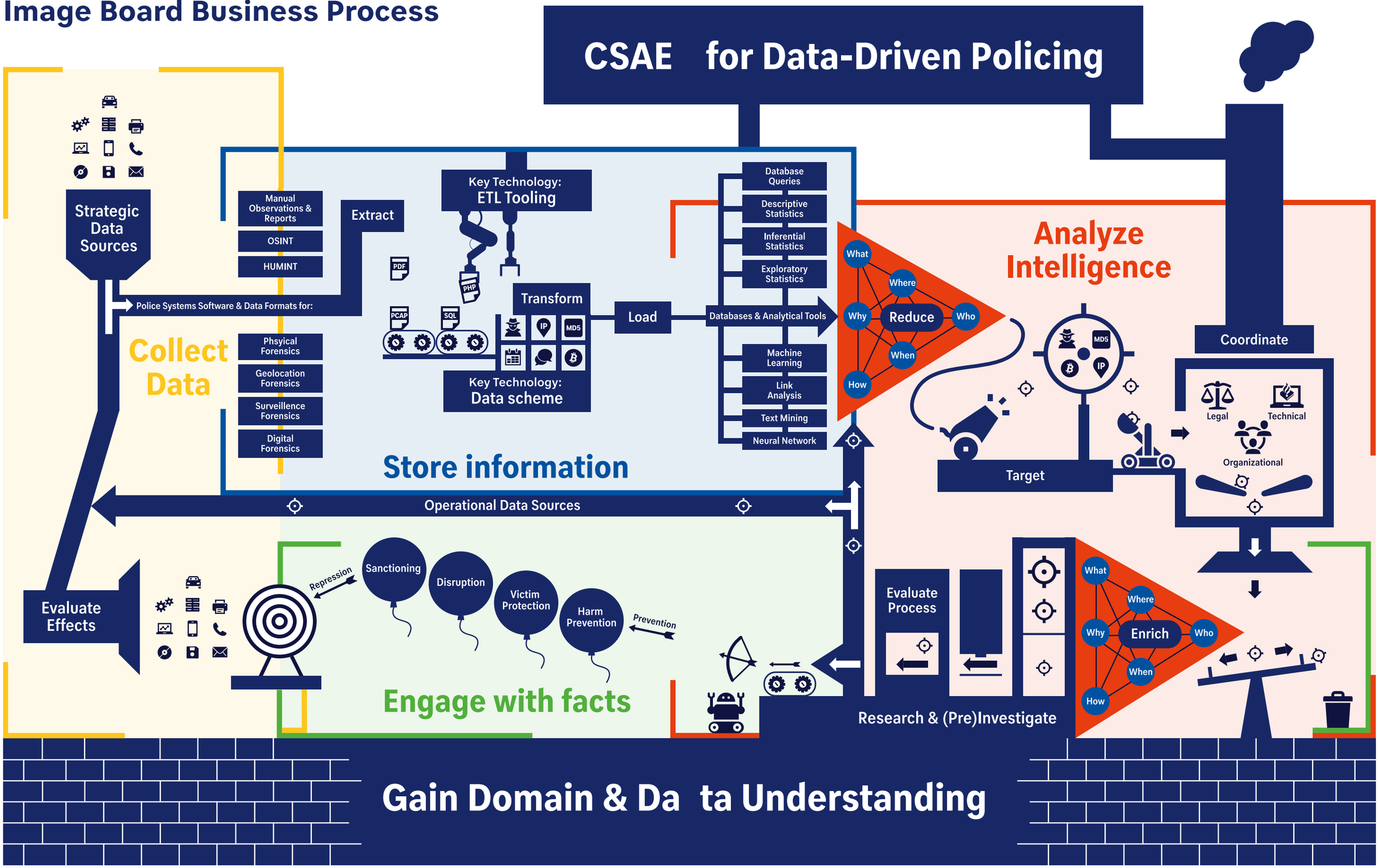
Eras are often clear only in hindsight, yet already we sense a new, algorithmic era of artificial intelligence emerging. Understanding data-driven policing in this context is crucial, for it is through this understanding that we gain the ability to shape and guide its evolution. In this second edition of our white paper, we have incorporated new lessons, best practices and state-of-the-art literature on data-driven policing using the comprehensive CSAE framework. We have explained why this approach matters and how the core concepts of CSAE - public interest philosophy, harmonization objectives, business processes, methodology and job disciplines - interact in practice to address today's criminal threats.

We acknowledge that this new paradigm of policing is still in its early stages. However, CSAE has the potential to serve as a unifying framework for existing policing philosophies - such as community policing, intelligence-led policing and problem-oriented policing - which are all increasingly becoming data-driven. By providing a common structure, the framework helps bridge the gap between traditional investigative intuition and modern algorithmic power.

At the same time, we are fully aware that transitioning to data-driven policing is not primarily a matter of understanding CSAE concepts. In practice, the main barriers are organizational, cultural and political - not conceptual. CSAE is therefore a living framework that evolves as academics, practitioners and law enforcement adapt it to their local contexts and contribute lessons learned. These contributions have directly informed this second edition, ensuring the framework remains grounded in operational reality.

Future editions will address the complex transition toward a fully data-driven police organization, covering essential topics such as organizational design, governance, and the strategies required for rolling out and scaling CSAE across diverse agencies. In this way, we hope that CSAE contributes to the evolving needs of law enforcement within a dynamic and challenging landscape. Ultimately, the effectiveness of policing and police legitimacy affect society as a whole, as public safety and security remain the essential cornerstones of liberal democracies.

Appendix A:
Image Board Business Process



Nomenclature

ADM	**Automated decision making**
AI	**Artificial intelligence**
BI	**Business intelligence**
CSAE	**Collect Store Analyze Engage**
ELT	**Extract-load-transform**
ETL	**Extract-transform-load**
HUMINT	**Human intelligence**
IT	**Information technologies**
LEA	**Law enforcement agency**
ML	**Machine learning**
MLAT	**Mutual Legal Assistance Treaty**
MLOps	**Machine learning operations**
NLP	**Natural language processing**
US	**United States of America**

Endnotes

1 E. Van de Sandt, A. Van Bunningen, J. Van Lenthe, and J. Fokker, "Towards Data Scientific Investigations: A Comprehensive Data Science Framework and Case Study for Investigating Organized Crime Serving the Public Interest," REPHRAIN, Tech. Rep., 2021. [Online]. Available: https://cpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2021/ 03/White-Paper-Towards-Data-Scientific-Investigations.pdf

2 M. Afzal and P. Panagiotopoulos, "Data in Policing: An Integrative Review," *International Journal of Public Administration*, no. ahead-of-print, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01900692.2024.2360586#abstract

3 W. Landman, Politiewerk aan de horizon. Technologie, criminaliteit en de toekomst van politiewerk. Den Haag: Politie Wetenschap, pp.182-186, 2023. [Online]. Available: https://www. politieenwetenschap.nl/publicatie/bijzondere_publicaties/2023/politiewerk-aan-de-horizon-394

4 J. J. Oerlemans and S. Royer, "The future of data-driven investigations in light of the Sky ECC operation," New Journal of European Criminal Law, vol. 14, no. 4, pp. 434–458, 2023. [Online]. Available: https://journals.sagepub.com/doi/10.1177/20322844231212661

5 Y. Lee, B. Bradford, and K. Posch, "The Effectiveness of Big Data-Driven Predictive Policing: Systematic Review," Justice Evaluation Journal, 2024. [Online]. Available: https://doi.org/10.1080/24751979.2024.2371781

6 Wetenschappelijke Adviesraad Politie, "Navigeren in niemandsland. Zeven urgente uitdagingen rondom digitalisering en ai in politiewerk," Tech. Rep., pp.20-21, 2025. [Online]. Available: https://www.wetenschappelijkeadviesraadpolitie.nl/uploads/publications/Navigeren-in-niemandsland.pdf

7 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.

8 J. H. Ratcliffe, Intelligence-led policing. Taylor and Francis, jan 2012.

9 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

10 N. Verma and L. Dombrowski, "Confronting Social Criticisms: Challenges when Adopting Data-Driven Policing Strategies," in CHI 2018 : proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, R. Mandryk, Ed. Montreal: ACM, 2018, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3173574.3174043

11 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

12 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

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

14 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://www.rijksoverheid.nl/documenten/rapporten/2016/06/20/ tk-bijlage-3a-voortgangsnota-versterking-opsporing-juni-2016

15 Ministerie van Veiligheid en Justitie, "Herijkingsnota. Herijking realisatie van de nationale politie," 2015.

16 Tweede Kamer der Staten-Generaal, "Contouren voor een effectieve, toekomstbestendige op-sporing (bijlage bij 29628,nr.593)," Vergaderjaar 2015-2016, 2015.

17 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

18 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

19 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

20 E. Van De Sandt, The Deviant Security Practices of Cyber Crime. Leiden: Brill | Nijhoff, 2021. [Online]. Available: https://brill.com/view/title/60184

21 S. Brayne, Predict and surveil: Data, discretion, and the future of policing. Oxford University Press, 2020. [Online]. Available: https://academic.oup.com/book/33466

22 D. Marciniak, "Data-driven policing: how digital technologies transform the practice and gover-nance of policing," Ph.D. dissertation, University of Essex, 2021.

23 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/

24 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

25 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

26 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

27 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

28 Landelijke Programma Datagedreven Samen Werken, "Landelijke Programmaplan Magazine Datagedreven Samen Werken," Nationale politie, Tech. Rep., 2024.

29 W. J. Bratton and J. Murad, "Precision Policing: A Strategy for the Challenges of 21st Century Law Enforcement," in Urban Policy 2018. Housing ladders, precision policing, the do's and don'ts of city branding - plus homeless shelters, business permitting, and more., W. J. Bratton, S. Eide, S. Goldsmith, M. Hendrix, H. Husock, J. Miller, J. Murad, A. M. Renn, and P. D. Salins, Eds. Manhattan Institute, 2018, ch. 2, pp. 21–38. [Online]. Available: https://manhattan.institute/ article/precision-policing-a-strategy-for-the-challenges-of-21st-century-law-enforcement

30 S. Egbert and E. Esposito, "Algorithmic crime prevention. From abstract police to precision policing," Policing and Society, vol. 34, no. 6, pp. 521–534, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10439463.2024.2326516

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

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

33 M. Schuilenburg and M. Soudijn, "Big data policing: The use of big data and algorithms by the Netherlands Police," Policing: A Journal of Policy and Practice, vol. 17, pp. 1–9, 2023. [Online]. Available: https://doi.org/10.1093/police/paad061

34 T. Linder, "Surveillance capitalism and platform policing: The surveillant assemblage-as-a-service," Surveillance and Society, vol. 17, no. 1-2, pp. 76–82, 2019. [Online]. Available: https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/12903

35 B. Bradford, J. Jackson, and J. Milani, "Police Legitimacy," in The Encyclopedia of Research Methods in Criminology and Criminal Justice: Volume II: Parts 5-8. Wiley, jan 2021, pp. 642–650. [Online]. Available: http://eprints.lse.ac.uk/83543/

36 T. R. Tyler, "Enhancing Police Legitimacy," Annals of the American Academy of Political and Social Science, vol. 593, no. 1, pp. 84–99, 2004. [Online]. Available: https://www.jstor.org/stable/4127668

37 J. Terpstra, N. R. Fyfe, and R. Salet, "The Abstract Police: A conceptual exploration of unin-tended changes of police organisations," Police Journal, vol. 92, no. 4, pp. 339–359, 2019.

38 J. Terpstra, R. Salet, and N. R. Fyfe, “Abstract Police Organisations: Distantiation, Decontextualisation and Digitalisation,” in Policing in Smart Societies, A. Verhage, M. Easton, and S. D. Kimpe, Eds. Palgrave Macmillan, 2022, pp. 9–26. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-83685-6

39 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.

40 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

41 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/

42 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

43 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

44 C. J. C. F. Fijnaut, “De identificatie van de internationale misdadiger in Europa: ontwikkelingen en discussies rond 1900,” in La CVDW. Liber Amicorum Chris Van den Wyngaert, S. Dewulf, Ed. Antwerpen: Maklu, 2017, pp. 205–226.

45 The Open Group, “TOGAF® Standard, 10th Edition – Architecture Development Method,” 2022. [Online]. Available: https://www.opengroup.org/togaf-standard-10th-edition-downloads

46 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

47 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

48 J. Henley, “Dutch government resigns over child benefits scan-dal,” 2021. [Online]. Available: https://www.theguardian.com/world/2021/jan/15/dutch-government-resigns-over-child-benefits-scandal

49 E. van Meijgaarden, “Datagedreven opsporen tegen de Italiaanse maffia met het CSAE-model,” Politieacademie, Tech. Rep., 2023.

50 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

51 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

52 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

53 J. W. Creswell, Qualitative Inquiry Research Design. Choosing Among Five Approaches., 2nd ed. Sage, 2007.

54 Research design. Qualitative, Quantitative, and mixed methods approaches, 2nd ed. SAGE Publications, Inc., 2009.

55 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

56 R. Dudley and D. Golden, “How the FBI Stumbled in the War on Cybercrime,” oct 2022. [Online]. Available: https://www.propublica.org/article/fbi-ransomware-hunting-team-cybercrime

57 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

58 N. Henke, J. Levine, and P. McInerney, “Analytics translator: The new must-have role,” feb 2018. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/ analytics-translator

59 S. P. Chainey and A. A. Berbotto, “A structured methodical process for populating a crime script of organized crime activity using OSINT,” Trends in Organized Crime, vol. 25, pp. 272–300, 2022.

60 Snaphaan, “Connecting the dots: Utilising crime scripting to leverage multimodal data and innovative techniques in a meaningful manner,” Methodological Innovations, 2025. [Online]. Available: https://journals.sagepub.com/doi/10.1177/20597991251336070

61 M. H. Moore, “Creating public value: The core idea of strategic management in government,” International Journal of Professional Business Review, vol. 6, no. 1, pp. 1–2, 2021. [Online]. Available: https://openaccessojs.com/JBReview/article/view/219

62 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.

63 L. Lessig, “The Law of the Horse: What Cyberlaw Might Teach,” Harvard Law Review, vol. 113, pp. 501–549, 1999.

64 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.

65 National Crime Agency, “National Crime Agency Annual Plan 2024-2025: Protecting the public from serious and organised crime,” National Crime Agency, Tech. Rep., 2024. [Online]. Available: https://www.nationalcrimeagency.gov.uk/who-we-are/publications/748-nca-annual-plan-2024-25/file

66 Royal Canadian Mounted Police, “Operational priorities.” [Online]. Available: https://www.rcmp-grc.gc.ca/prior/index-eng.htm

67 Australian Federal Police, “Policing for a Safer Australia: Strategy for Future Capability,” Australian Federal Police, Tech. Rep., 2017. [Online]. Available: https://www.afp.gov.au/sites/default/files/PDF/strategy-for-future-capability.pdf

68 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

69 A. Mell, “Promoting Market Failure: Fighting Crime with Asymmetric Information,” 2015. [Online]. Available: https://docs.google.com/viewer?a=v&pid=sites&

70 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

P-25-0611