The Data-Laundromat?

Public-Private-Partnerships and Publicly Available Data in the Area of Law Enforcement

*Thilo Gottschalk**

Law enforcement increasingly relies on complex machine learning approaches to support investigations. With limited knowledge and funding LEAs often depend on opaque private-public collaborations. Failure to provide legal bases on the national level paired with shortcomings both in the GDPR and Directive EU-2016/680 (LED) result in severe risks for fundamental rights of EU citizens. To overcome these risks an interdisciplinary discussion is required. This paper hence sheds light on technical challenges and misconceptions as well as legal shortcomings to foster a common understanding of the challenges to find out how they might be addressed. To do so, the author searches for common ground of 'public availability' and reviews currently used technical approaches and common processing constellations. Based on the outcomes, the author proposes a change in the LED and discusses a centralised institution to govern access to novel data driven technology.

Keywords: law enforcement; public-private partnership; data protection; GDPR; LED

I. Introduction

If data is the oil of 21st century, data analysis is the refinery to make it usable. While the General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED) set up general rules for the processing of personal data, the resulting protection remains fragile when both frameworks are brought together in public-private partnerships. One area where this is the case is the realm of 'publicly available data'. With limited IT-knowledge and budgets, law enforcement agencies (LEAs) often put themselves in the hands of external data scientists that lack the necessary knowledge of the legal implications of their approaches in particular when it comes to publicly available data. Unfortunately, neither the GDPR nor the LED or data protection authorities lay down sufficiently precise rules for collaboration, leaving the participating parties alone in a legal vacuum. For example, as of the time of writing – well over one year after the GDPR and LED stepped into force – the EDPB website on police and justice does not provide any content at all.[1]

In combination with novel data driven approaches - such as artificial intelligence - these collaborative approaches have the capability to severely undermine the fundamental rights of EU citizens.

At the same time citizens increasingly start to use and request privacy supporting techniques such as encryption, cryptocurrencies, VPN's or the TOR network[2] The regulation of access is hence often increasingly governed by technical rather than legislative measures. As a consequence, data that is not subject to so-called *privacy enhancing technologies* (PETs) is often seen as fair game for data processing - not only in the area of law enforcement. While chilling ef-

DOI: 10.21552/edpl/2020/1/6

* Thilo Gottschalk, Center for Applied Legal Studies, Karlsruhe Institute of Technology (KIT). For correspondence: <thilo.gottschalk@kit.edu>. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.740558. All Internet links in this article were last accessed on 12 March 2020.

1 <https://edpb.europa.eu/our-work-tools/general-guidance/police-justice-guidelines-recommendations-best-practices_en>.

2 The Verge, 'Apple argues stronger encryption will thwart criminals in letter to Australian government' (2018); Microsoft Corp, 'OneDrive Personal Vault brings added security to your most important files and OneDrive gets additional storage options' (2019) <https://www.microsoft.com/en-us/microsoft-365/blog/2019/06/25/onedrive-personal-vault-added-security-onedrive-additional-storage/>.

fects of surveillance have been long proven[3] the amount publicly available information still increases steadily due to the so-called privacy paradox.[4] This can result in valuable information for law enforcement agencies[5] that need efficient tools to handle increasing illicit use of novel technologies[6].

Current legal research (II) widely focuses on general problems of data processing but does seldom embed these problems in practical processing constellations. Technical research on the other hand is usually driven by the analysis of specific practical problems and mostly neglects legal implications. Interdisciplinary research is almost non-existent indicating that the GDPR's cry for *privacy-by-design* has not yet been heard by all relevant parties. This article aims to be a very first step in closing this gap. Although focused on the area of law enforcement, the identified problems and solutions are partly transferable to processing constellations in other contexts. The get the full picture, this article illuminates a common misconception on public availability (III), possible data sources and types (IV), available analytical methods (V) and the processing constellations (VI). The author proposes a centralised data-clearinghouse solution that ensures access to data driven analytical methods while safeguarding the fundamental rights to privacy and data protection of the data subjects (VII) followed by a conclusion of the findings (VIII).

II. Related Work

The complexity of the legal framework paired with complex technological issues resulted in a lack of interdisciplinary research that would be necessary to identify and address current and future challenges. As a consequence, related research is usually focused on novel data processing approaches such as artificial intelligence (AI) in the context of the GDPR, the context of law enforcement or multi-legislation frameworks are only covered very sparsely (eg Brkan[7]). Most of the research is either focused on general societal and technical challenges[8] related to the use of algorithms or focused on the commercial context[9]. Privacy related issues are seldom covered in detail. The general implications and needs for protection of privacy have been extensively researched with regard to public and non-public data, though.[10] Beyond that, not only extraction of publicly available data but also production of such (eg astroturfing) has been covered[11]. Purtova[12] and Brkan[13] identified the

3 Jonathon W Penney, 'Internet surveillance, regulation, and chilling effects online: a comparative case study' (2017) 6(2) Internet Policy Review 394; Jonathon W Penney, 'Chilling Effects: Online Surveillance and Wikipedia Use' (2016) <http://dx.doi.org/10.15779/Z38SS13>; Alex Marthews and Catherine E. Tucker, 'Government Surveillance and Internet Search Behavior' (2017) <https://dx.doi.org/10.2139/ssrn.2412564>.

4 A Acquisti and J Grossklags, 'Privacy and rationality in individual decision making' (2007) 3(1) IEEE Security & Privacy 1.

5 Internal Revenue Service (IRS), 'Social Media Research Request' (2018) <https://www.fbo.gov/spg/TREAS/IRS/NOPAP/2032H8-RFI-MEDIA/listing.html>; United States District Court, ED California, *United States v Marcos Paulo de Oliveira-Annibale* (2019) <https://www.justice.gov/opa/press-release/file/1159711/download>; United States District Court, CD California, *United States v Tibo Lousee, Klaus-Martin Forst and Jonathan Kalla* (2019) <https://www.justice.gov/opa/press-release/file/1159706/download>.

6 Europol, 'Internet Organised Crime Threat Assessment' (2018) <https://www.europol.europa.eu/iocta-report>.

7 Maja Brkan, 'Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond' (2019) 27(2) International Journal of Law and Information Technology <https://doi.org/10.1093/ijlit/eay017>.

8 Vlad Krotov and Leiser Silva, 'Legality and Ethics of Web Scraping' (2018) <https://bit.ly/33tN6k7>.

9 Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2017) <https://dx.doi.org/10.2139/ssrn.3038939>; Ross Anderson et al, 'Measuring the Changing Costs of Cybercrime' (2019) The 2019 Workshop on the Economics of Information Security (WEIS 2019) <https://pure.tudelft.nl/portal/files/54190531/Measuring_the_Changing_Cost_of_Cybercrime_WEIS_1.pdf>.

10 Penney, 'Chilling Effects: Online Surveillance and Wikipedia Use' (n 4); Marthews and Tucker (n 4); Julia Hoernle, 'Juggling more than three balls at once: multilevel jurisdictional challenges in EU Data Protection Regulation' (2019) 27(2) International Journal of Law and Information Technology 142 <https://doi.org/10.1093/ijlit/eaz002>; Christian Rückert, 'Cryptocurrencies and Fundamental Rights' (2019) 5 Journal of Cybersecurity; Christian G Rückert, 'Zwischen Online-Streife und Online-(Raster-)Fahndung - Ein Beitrag zur Verarbeitung öffentlich zugänglicher Daten im Ermittlungsverfahren' (2017) 129(2) ZStW 302; Shoshana Zuboff, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilization' (2015) Journal of Information Technology; Daniel J Solove, 'A Taxonomy of Privacy' (2005); Daniel J Solove, 'I've Got Nothing to Hide and Other Misunderstandings of Privacy' (2011) <https://bit.ly/2vsVCTX>; Lilian Edwards and Lachlan Urquhart, 'Privacy in Public Spaces: What Expectations of Privacy do we have in Social Media Intelligence?' (2015) <https://academic.oup.com/ijlit/article/24/3/279/2404493>.

11 Amelia Johns and Niki Cheong, 'Feeling the Chill: Bersih 2.0, State Censorship, and 'Networked Affect' on Malaysian Social Media 2012–2018' (2019) <https://journals.sagepub.com/doi/pdf/10.1177/2056305118821801>; Blake Miller, 'Automated Detection of Chinese Government Astroturfers Using Network and Social Metadata' (2019) <https://dx.doi.org/10.2139/ssrn.2738325>.

12 Nadezhda Purtova, 'Between the GDPR and the Police Directive: Navigating Through the Maze of Information Sharing in Public-Private Partnerships' (2018) International Data Privacy Law <https://dx.doi.org/10.2139/ssrn.2930078>.

13 Brkan (n 8).

gap between the GDPR and LED when it comes to joint controllership of LEAs and private parties but did not touch upon the topic from an interdisciplinary perspective. All of the research shows (potentially) negative effects of all sorts of surveillance on society in some form. However, it is seldom differentiated between possible processing scenarios. The lack of differentiation between constellations seems to result in less efficient discussions and misunderstandings on the legal and technical side (eg developers assuming that the GDPR is always applicable). The precise evaluation of processing contexts is necessary to appropriately address challenges in a manner that balances LEA requirements and the protection of fundamental rights. This paper hence tries to open the discussion for a more interdisciplinary approach and to provide a look at the 'bigger picture'.

III. Publicly Available Data

To start off, it is necessary to think about the meaning of public availability. Compliance with legal regulations is only possible if the subjects of the law understand what is required from them. The crux in the current discussion is, that public availability is often falsely used synonymous with 'not protected in any way'. This section will discuss possible origins of this misconception and how it may be addressed in the future. *Public availability* can be defined in multiple contexts. The term originates from far before great scale data processing existed and is used in a broad variety of contexts. It now needs to be transferred to the context of data processing where it currently seems to lack a common understanding at least between legal practitioners and computer scientists/developers. It is necessary to have an interdisciplinary consensus what we mean with that term and what its legal (and ethical) implications are. One way to do so is to find the lowest common denominator in different interpretations. The two most relevant interpretations stem from the legal and the technical context.

1. Legal Interpretation

To date there is no globally valid legal definition of public availability. Given the global scope of the World Wide Web (WWW) a common consensus of the understanding would be desirable. To ensure this the creation should hence stem from overarching governing bodies such as the European legislator or international treaties. Scholars and jurisdiction have already started to carve out a common understanding at least on the European level. Different legal systems such as the US American one of have similar discussions. The basis of these discussions (EU/US) is different and should not be mixed thoughtlessly. Unfortunately, it appears that exactly this mixing takes place through the back door of international entanglement in the area of information technology and computer science where the different foundations of the discussion are quickly overlooked. The following section will start with the European legal interpretation, followed by a quick excursion to the US legal discussion. While both systems are generally independent in the legal theory and reflected in the statutes themselves, the underlying perception, interpretation and execution of these statutes is influenced by a global discussion that does not include a clean distinction between legal systems. For example, a developer with no legal background will not be able to identify detailed differences and limitations between the US and EU legal frameworks. The lack of differentiation can be a source for data processing approaches that are noncompliant with the European legal framework. To overcome this lack of differentiation in the general discussion, it is necessary to raise awareness of the different implications of the respective legal frameworks, their backgrounds and the different discussions that are led within them. The following Sub-sections III.1.a and III.1.b will point out the differences between the US and EU legal discussions that may be a source for misconceptions described in Section III.2 with regard to the processing and the protection of publicly available personal data.

a. EU Legal Interpretation

The term 'public availability'[14] appears in various European laws and contexts[15] but none provides a concrete legal definition. In the area of law enforcement, the inclusion of the term in Article 17 of the Europol

14 or variations such as 'publicly available' or '(manifestly) made public'.

15 Regulation (EU) No 596/2014 of the European Parliament and of the Council of 16 April 2014 on market abuse (market abuse regulation) and repealing Directive 2003/6/EC of the European Parliament and of the Council and Commission Directives

Regulation indicates that in contrast to US legislation - the European legislator does not see publicly available data as a 'carte blanche' for data processing from a data protection perspective. Similarly, in a commercial context protection may be derived from the so-called Database Directive.[16] Widening that protection, the Court of Justice of the European Union (CJEU) ruled that Terms-of-Use (TOUs) can be sufficient to disallow web scraping in cases where the Directive EU-96/9 is not applicable.[17] Where it is applicable, it generates a *sui generis* protection for the database. Speciously, the commercial context seems to be irrelevant when it comes to data processing for law enforcement purposes - however, it can quickly become relevant where private companies take care of the data collection and analysis for LEAs. The core of the understanding appears to be that data is *publicly available* if access to it is not limited to a specified group of persons. The necessary limitations itself are yet to be discussed.[18] Some scholars require the data to be free (ie no remuneration required)[19] others argue for a definition without this feature[20]. The inclusion of (monetary) compensation however should not be included for multiple reasons. First, it blurs the line between public and non-public data. Eventually, everyone pays to access data in one way or the other - be it direct access fees, service provider fees, or provision of data for advertisement. Secondly, geo-blocking and similar measures would result in different data categorisations within Europe. If I can freely access an Italian newspaper through a free VPN does that make data publicly available *for me* but not for non VPN-users? Similarly, some companies or institutions (eg universities) may be able to freely access data that others have to pay for - again resulting in different classifications. Last, and most importantly - if we connect the level of protection with public availability based on monetary remuneration, we couple the protection of fundamental rights of individuals to economic interests of third parties. Two things that do not necessarily go well together. In contrast to the US, the discussion about the scope of the term is, however, subsidiary. The key question under EU legislation is if data are related to natural persons and hence fall under the scope of Articles 7 and 8 of the Charter of Fundamental Rights of the European Union (EU-CFR), specified in the GDPR and LED. The rights to privacy and data protection are governed by a *material scope* - ie the relation to a natural person - rather than an *areal scope* like in the US. The mere public availability of data is no feature that can completely disable the protection of a natural person through Articles 7 and 8 EU-CFR. It plays however, an important role in the weighing of interests and can certainly influence the degree of protection (see III.3).

b. US Legal Interpretation

While the US perspective seems irrelevant on first sight, it appears that in the discussion between legal and technical persons the lines of argumentation are often influenced by concepts that originally come from the US. While these argumentations are seldom easily transferrable to the EU level, they still need to be understood to understand where certain misconceptions (ie publicly available data is fair game) stem from. On constitutional level data protection/privacy is embedded in the 4th Amendment that particularly protects 'persons, houses, papers, and effects' against 'unreasonable searches and seizures'. In contrast to Article 7 EU-CFR the protection is hence not automatically granted if data is 'related to a natural person' but rather depends on the reasonable expectation of privacy[21] of the affected person. If that expectation exists, is regularly determined by areal considerations. In earlier decisions the US Supreme Court decided that there cannot be

2003/124/EC, 2003/125/EC and 2004/72/EC Text with EEA relevance [2014]; Regulation (EU) 2016/794 of the European Parliament and of the Council of 11 May 2016 on the European Union Agency for Law Enforcement Cooperation (Europol) and replacing and repealing Council Decisions 2009/371/JHA, 2009/934/JHA, 2009/935/JHA, 2009/936/JHA and 2009/968/JHA [2016]; General Data Protection Regulation (EU 2016/679) [2016]; Law Enforcement Directive (EU 2016/680)' (2016); Regulation (EU) 2019/631 of the European Parliament and of the Council of 17 April 2019 setting CO2 emission performance standards for new passenger cars and for new light commercial vehicles, and repealing Regulations (EC) No 443/2009 and (EU) No 510/2011' [2019].

16 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996].

17 Case C-30/14 *Ryanair v PR Aviatio*n [2015] ECLI:EU:C:2015:10.

18 Michael Dallmann and Philipp Busse, 'Verarbeitung von öffentlich zugänglichen personenbezogenen Daten - Datenschutzrechtliche Voraussetzungen und Grenzen' (2019) Zeitschrift für Datenschutz 394.

19 ibid.

20 Edwards and Urquhart (n 11).

21 Supreme Court of the United States, *United States v Katz* (1967) <http://scdb.wustl.edu/analysisCaseDetail.php?cid=1967-043-01/>.

a reasonable expectation of privacy if you are in public, namely with regard to trashcans on the curbs[22] and the observation from a helicopter[23]. American legal scholars repeatedly argued and criticised that this distinction is not sufficient[24] and more recent jurisdiction tries to broaden the protection through the extension of the *reasonable expectation of privacy*. For example, in *Florida v Jardines*[25] the US Supreme Court decided that the publicly accessible front porch of a building is still protected under the 4th Amendment due to the reasonable expectation of privacy of the resident. However, even with this broadened scope the general assumption that there is no reasonable protection of privacy in public spaces is generally held up to this day. The interpretation resulted in *public availability* being almost synonymous with the omission of protection under US legislation. Transferring this to the internet, US scholars try to find areal analogies to narrow down the scope of public availability and in consequence broaden the space where a reasonable expectation of privacy - and hence protection of the data subject - can be assumed.

Kerr[26] discusses the applicability of current US computer trespass laws such as the Computer Fraud and Abuse Act (CFAA)[27] on data scraping measures and derives the reasonable expectation of privacy from classic trespassing norms and social custom. He argues that the courts need to identify the norms that apply best but also need to set up '*normative policy decisions about what understandings should govern the Internet*'[28] to shape future (computer) trespass norms. In *US v Auernheimer*[29] the defendant was convicted of *unauthorised access* for collecting information from a website of US telecommunication provider AT&T which was accessible on a *hard to guess website* that was not intended to be accessed. Although the data was publicly accessible the court stated that analogous to a home where 'the front door is left open or unlocked' the data was still protected under the CFAA. The defence argued that the information was made available to everyone and the general public was authorised to view the information. As Kerr[30] points out, norms for the WWW are hard to identify due to two different narratives. First, generally everyone can go to any website. Second, many website owners/persons do not want everyone to access their sites. He then argues, that due to the open nature of the web, the access of a website should still be deemed authorised even if (so called) 'speed bumps' (eg hidden addresses, cookies, and IP-blocks) are bypassed. In his line of argumentation, the authorisation line would be crossed when access is gained by bypassing an authentication requirement that creates a sufficient barrier (eg password protection). Under US current legislation this definition of public availability would imply that such publicly available data is not protected by the 4th Amendment without such speed bumps. In contrast to this scholarly view, courts increasingly start to presume that publication of data does not necessarily preclude an expectation of privacy[31] nor commercial protection[32]. This is accompanied by legislative actions on state level to foster data protection such as the Californian Consumer Data Protection Act[33]. Having said that, as a consequence of the jurisdiction reaching back to the 60s and 70s there is a deeply rooted conception that public availability results in the omission of constitutional protection not at least by leading data scientists.

2. Technical Interpretation

Aside from the legal interpretations of the term it is unclear how far these discussions are heard in tech-

22 Supreme Court of the United States, *California v Greenwood* (1988) <https://caselaw.findlaw.com/us-supreme-court/486/35.html>.

23 Supreme Court of the United States, *Florida v Riley* (1989) <https://caselaw.findlaw.com/us-supreme-court/488/445.html>.

24 Lance E Rothenberg, 'Re-Thinking Privacy: Peeping Toms, Video Voyeurs, and Failure of the Criminal Law to Recognize a Reasonable Expectation of Privacy in the Public Space' (2011) 49(5) American University Law Review <https://bit.ly/2vy0mrC>.

25 Supreme Court of the United States, *Florida, Petitioner v Joelis Jardines* (2013) <https://www.leagle.com/decision/insco20130326e71>.

26 Orin S Kerr, 'Norms of Computer Trespass' (2016) 116(4) Columbia Law Review <https://columbialawreview.org/content/norms-of-computer-trespass/>.

27 US Computer Fraud and Abuse Act (CFAA) (2012).

28 Kerr, 'Norms of Computer Trespass' (n 27) 1155.

29 Supreme Court of the United States, *United States v Andrew Auernheimer* (2014) <https://cite.case.law/f3d/748/525/>.

30 Kerr, 'Norms of Computer Trespass' (n 27); Orin S Kerr, *Computer Crime Law - Summer 2018 Case Supplement* (2018).

31 Supreme Court of the United States, *Carpenter v United States* (2018) <https://supreme.justia.com/cases/federal/us/585/16-402/>.

32 United States District Court, N.D. California, *HiQ Labs Inc. v LinkedIn Corp* (2017 (ongoing)) <https://epic.org/amicus/cfaa/linkedin/LinkedIn-Opening-Brief.pdf>.

33 CA State Senate, CA State Assembly, 'California Consumer Privacy Act' (2018) <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375>.

nical realms. With major data-science hubs in the America and Asia and the global entanglement of research in this area the understanding of public availability seems to be influenced on a global level and ignores peculiarities of the respective European legal systems. Due to a lack of discussion within the realm of data and computer scientists of these issues, it appears that public accessibility lacks a common understanding and regularly depends on individual perceptions. In the broadest manner, data may be publicly available if it is not subject to a state-of-the-art protection. For professional data scientists the scope of publicly available is hence likely to be broader than in the legal understanding - either in the US or the EU.

The complexity and uncertainty of the legal discussions of the term, as well as unresolved questions around the applicability of data protection in the area of novel data driven approaches seem to trigger the urge to go with a simple solution - namely, cherry picking the most favourable arguments or completely ignoring the discussion - reflected by the absence of discussion in technical papers on data science.

3. Conclusion

The influence of the US legal discussion that drips into technical realm, paired with general uncertainty in the European legal discussion seems to result in a blurred understanding of 'public availability' and a widespread misconception about the consequences of public availability. One reason for this could be the fact that developers and data scientists usually work in an international context while legal scholars are often caught in national contexts and perceptions. This can be tackled on multiple levels. First, data scientists must become more aware of legal implications of their processing and understand differences between legal systems. It must be acknowledged that public availability does not omit the protection of personal data but is merely one building block in the assessment of legitimate interests for a lawful processing. With regard to the differences between the EU and US it must be acknowledged that both protective approaches have their right to exist, are coherent in themselves and reflect the respective societal perceptions. Mixing them up however will most certainly result in lowered factual protection of fundamental rights in the European Union. On the other side, legal scholars need to gain better in-depth understanding of technical approaches and possibilities to foster the discussion about public availability in a way that reflects technical reality. In addition, to that globally valid guidelines and best practices would be desirable to foster a common international understanding in the long term (cf the Robots.txt[34]).

IV. Data Sources

Electronic information is usually stored in databases, ie a somewhat structured set of data held in a computer. The following section will give a quick overview of some relevant sources of information and shortly discuss specific risks related to them. For many of them the legal implications are not yet conclusively covered and further research is required. Independently of the source, the processing of publicly available data generally constitutes an interference with fundamental rights to data protection and privacy if it can be related to natural persons[35].

1. Open Data

Open data is data that is made freely accessible without limitations regarding the purpose of the processing. Such data usually pre-processed and ideally does not contain information relatable to natural persons. Such data is hence not rated personal data under the European data protection framework. However, where such data is connected with personal data it can quickly become related to a natural person. For example, if you live in the German town of Soest-Paradiese and are under 17 - you are most likely (76%) female.[36] The publicly available demographic data it-

34 Google Inc. 'Formalizing the Robots Exclusion Protocol Specification' (2019) <https://webmasters.googleblog.com/2019/07/rep-id.html>; Martijn Koster, 'ANNOUNCE: A Standard for Robot Exclusion' (1994) <http://1997.webhistory.org/www.lists/www-talk.1994q3/0017.html>.

35 Penney, 'Internet surveillance, regulation, and chilling effects online: a comparative case study' (n 4); Johns, and Cheong (n 12).

36 Based on open-data from <https://opendata.gelsenkirchen.de>.

self is anonymous however it allows the creation of very exact profiles and predictions. While data scientists are aware of these issues,[37] anonymisation measures are however usually related to single data sets. There is no oversight of combination possibilities for public data sets, often rendering the best anonymisation techniques obsolete[38]. In addition, open data is usually prepared for further processing and accessible through application programming interfaces (APIs), making this data particularly easy to access and process. It can hence easily be used to gather intelligence or support ongoing investigations (eg crime statistics[39], research hubs[40] or open data spaces[41]). While European authorities are rather careful with statistics and data on criminal activity, US pendants are much less reluctant in this regard. For example, data.gov provides lists of every arrest in New York City going back to 2006. The data provided in this context can be of high value when it comes to law enforcement operations but also bears great risks with regard to biased interpretation or misuse for illicit profiling. Having said that, it must be acknowledged that open data can be extremely valuable in various contexts and their publication is hence generally desirable. Data providers, ie controllers, should not be accountable to control access to these data sets. However, the legislator must acknowledge that processing of such information in a law enforcement context bears particular risks (of profiling) and needs to be regulated.

2. Surface Web

The information that is available on publicly accessible websites is almost infinite. The layman understanding of publicly available websites usually refers to the so-called surface web. The available data types are almost infinite and the specific risks for the data subject strongly depend on the specific information that is extracted. In this context the surface web only covers information that is accessible through screen-scraping. In general, scraping software accesses websites similarly to a human interaction. There are various approaches available and the scraping usually targets specific data points instead of the whole website, to exclude advertisements or irrelevant information. Targeting specific information however is a difficult task when unstructured data is accessed. Website-designs and structures change and require scrapers to adapt accordingly. This adaptation often relies on machine learning/AI. It must hence be acknowledged that the scraped data may lack specific information which raises questions with regard to data quality and accuracy that are particularly important in the area of law enforcement. The website-owner usually has very limited power to limit such scraping activities, it is hence that the scraping party has to be deemed the individual controller of the collection.

3. Social Media

A sub-section of the surface web is social media (eg Instagram, Snapchat, Facebook, Tinder). Social connections have always been an important investigative approach, with the shift from real-life to electronic communication these connections are often easily accessible and generate valuable insights for law enforcement. [42] Some of the currently existing networks allow users to limit the reach of their content to certain user groups (everyone, network participants, friends, friends of friends). The public availability for such restricted data hence often depends on factual barriers that these settings eventually raise. Data on social networks are easily relatable to natural persons and often give insights in particularly sensitive areas of a persons' life such as religious or political beliefs or sexual preferences. Accessing social media data is hence bears severe risks to the fundamental rights of the data subject. While data

37 H Silva et al, 'A Re-Identification Risk-Based Anonymization Framework for Data Analytics Platforms' (2018).

38 Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'Estimating the success of re-identifications in incomplete datasets using generative models' (2019) <https://doi.org/10.1038/s41467-019-10933-3>; Timothy Morey, Theodore Forbath and Allison Schoop, 'Customer Data: Designing for Transparency and Trust' (2015) <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>; Latanya Sweeney, 'Simple Demographics Often Identify People Uniquely' (2000) <https://dataprivacylab.org/projects/identifiability/paper1.pdf>.

39 BKA, 'Polizeiliche Kriminalstatistik (Germany)' (2018) <https://www.bka.de/DE/AktuelleInformationen/StatistikenLagebilder/PolizeilicheKriminalstatistik/pks_node.html>.

40 'SSRN' <https://www.ssrn.com>; 'Google Scholar, Google LLC'.

41 'Data.gov' <https://www.data.gov/open-gov/>; 'Open Data Austria' <https://www.data.gv.at/>; 'Open Data Gelsenkirchen' <https://opendata.gelsenkirchen.de/>.

42 Markus Huber et al, 'Social Snapshots: Digital Forensics for Online Social Networks'; Hamaira Arshad et al, 'A multilayered semantic framework for integrated forensic acquisition on social media' (2019) 29 Digital Investigation 147 <https://doi.org/10.1016/j.diin.2019.04.002>.

on social media may be *manifestly* made public, this cannot be re-interpreted as consent or abandoning fundamental rights protection.

4. Deep Web

The term 'deep web' generally describes content that cannot be accessed through conventional search engines such as Google, Bing or Yahoo - ie content that has not been indexed by these search engines. The exclusion from indexing is mostly based on best practices such as the robots.txt and factual technical barriers such as encryption and password protection. The fact that something is not indexed has no influence on public availability as long as there are no additional measures beyond the robots.txt. Similar to the surface web, data cannot be categorised but needs to be evaluated on a case-by-case basis. This becomes particularly important if data becomes publicly available through faulty configuration of web servers. Identification of *accidental* publication is similarly difficult as identification of *manifest* publication and cannot be interpreted as consent of the data subject.

5. Dark Web

The so-called 'dark web' is a part of the internet that can only be accessed through specific software that uses anonymising functionalities. The most popular browsers/networks in this regard are the TOR[43] and I2P[44], with the latter being much less used. Both intend to create a network that is resistant to government interference (eg censorship) to ensure free flow of information between users (clients). Client connections are steered through various hops (relays) and directed to the target (surface-)web server. The user hence does disclose his IP address to the provider of the server and the browsers additionally suppresses cookies and other re-identification measures. In addition, both networks provide the possibility for so called hidden-services. In this scenario the targeted webserver also does not have to disclose its IP-addresses and can remain anonymous. As it is difficult to locate these hidden-services, the analysis of the publicly available information - ie darknet websites becomes ever more important. The mere requirement of a specific software does not contradict the public-availability of the dark web, the same is true for pseudo-access restrictions such as unrestricted forum registration. Hidden services are often misused for illicit purposes such as providing platforms for trading of drugs, weapons or counterfeit money and seem to be relatively short living[45] with darknet marketplaces usually closing down in under a year.[46] Closing down, however, regularly just means vendors and customers moving to the next marketplace.[47] With limited longevity the processing and in particular storage of publicly available data becomes particularly important to avoid the decay of relevant evidence. At the same time, the extensive processing of this information source can state a severe interference with fundamental rights to privacy and data protection of the network users.

6. Publicly Available APIs

Besides these sources many companies provide publicly available APIs[48]. While some providers of APIs require registration (API-authorisation-keys) these APIs still have to be deemed publicly available if the registration is not further restricted to a specific group of users. However, API access can of course also be limited to specific persons. Provision of APIs is usually driven by economic considerations. APIs are hence usually created to allow additional software/functions to be usable with the respective platforms. These APIs provide specific insights and access to functions and content available through the data providers. Looking at the use of APIs it remains difficult to determine the controllership-attributes of the requesting and the responding party. While the request is governed by A, the content of the response and the accessed database itself is controlled by B. The

43 'The Onion Router' <https://www.torproject.org/>.

44 'Invisible Internet Project' <https://geti2p.net>.

45 Amirali Sanatinia et al, 'A Privacy-Preserving Longevity Study of Tor's Hidden Services' (2019).

46 European Monitoring Centre for Drugs and Drug Addiction, 'Drugs and the darknet: perspectives for enforcement, research and policy' (2017) <http://www.emcdda.europa.eu/publications/joint-publications/drugs-and-the-darknet>.

47 Anirudh Ekambaranathan, Sarah Meiklejohn and Andreas Peter, 'Using Stylometry to Track Cybercriminals in Darknet Forums' (2018) <https://pdfs.semanticscholar.org/aebe/d4be444386dc4beb88f8e63c75e81c14a7b8.pdf>; Xiao H Tai, Kyle Soska and Nicolas Christin, 'Adversarial Matching of Dark Net Market Vendor Accounts'.

48 eg docs.shapeshift.io, de.openlegaldata.io, reddit.com/dev/api, github.com/public-apis.

CJEU cases C-210/16 (Facebook fanpage), C-40/17 (like button on external website) as well the Article 29 Working Party Opinion on controllership[49] state that the access to APIs most likely has to be deemed a joint controllership by the requesting and providing party. Regarding the risks for the data subject it is debatable if API access should be preferred over screen-scraping from a privacy perspective. Data provided through APIs is usually well structured, while screen-scraping is in general more prone to errors and is likely to result in a lower data quality - increasing the risk for the data subject. The joint controllership constellation provides the data subject with two accountable parties. Since the third party will often remain unknown to the data subject, an additional point of contact to exert data protection rights is positive. At the same time, the accountability of the joint controllership is limited to the actual access/collection of the data through the API. Any further processing will usually remain under the sole control of the collecting party. However, a known initial point of contact at least eases identification of the respective third party.

7. P2P

There are myriads of peer-to-peer (p2p)-networks that can provide investigators with publicly available information.[50]

Classic p2p-sharing-networks such as Kazaa, Limewire or Torrents are usually publicly available to allow all users to participate in the file-sharing. Since these networks mainly raise questions with regard to IP law, this section will focus on a specific p2p-driven system - cryptocurrencies. Cryptocurrencies usually rely on blockchain technology to overcome the issue of lacking trust between the participating actors in the network. Cryptocurrencies such as Bitcoin or ZCash have been subject to increasing interest in research and law enforcement. Given the lack of trust in a central actor, all transactions must be verified by the network itself. As a consequence, all transactions need to be verifiable and hence visible to the network participants. Since participation is not limited to a specific group, the data transaction data is deemed publicly available. Based on cryptographic approaches some virtual currencies allow obfuscation of the transactions thereby trying to maintain anonymity of the users. Having said that, a long line of research has shown that even so-called privacy coins often produce only limited levels of anonymity.[51] With cryptocurrencies speciously being anonymous, they are often the currency of choice for illicit transactions. At the same time, they are becoming increasingly popular in daily life - be it as speculative financial investment or actual substitution of a fiat currency. *Follow the money* remains a valid principle in cyber-based investigations and LEAs are keen to use this principle on virtual currencies. Performing transactions requires the user to broadcast transaction-message to a node in the network. This message is then spread across the network through other participating nodes (so-called 'Gossip Protocol'). The straight-forward approach is based on classic man in the middle (MITM) attack. The investigating party (LEA or private party) sets up one or more nodes in the network to analyse the network traffic and store IP-addresses of the participants[52] often in combination with additional techniques, such as the analysis of so called simplified-payment-verification (SPV) messages of Bitcoin clients.[53] This approach however is limited by the fact that the initial message can still be sent through additional network layers, such as TOR with some cryptocurrencies integrating such protective measures in their protocol itself (eg Bitcoin[54]). Besides metadata (eg IP-addresses, connection times, etc) the content of the network communication (ie transaction messages) is analysed as well. Based on this data, the pseudonymous or

49 WP29, 'Opinion 1/2010 on the concepts of 'controller' and 'processor (WP 169)' (2010) <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf>.

50 Claudia Peersman et al, 'iCOP: Live forensics to reveal previously unknown criminal media on P2P networks' (2016) 18 Digital Investigation <https://doi.org/10.1016/j.diin.2016.07.002>; Marc Liberatore et al, 'iCOP: Live forensics to reveal previously unknown criminal media on P2P networks' (2016) 7 Digital Investigation <https://doi.org/10.1016/j.diin.2010.05.012>; Yee-Yang Teing et al, 'Forensic investigation of P2P cloud storage services and backbone for IoT networks: BitTorrent Sync as a case study' (2017) 58 Computers & Electrical Engineering 350 <https://doi.org/10.1016/j.compeleceng.2016.08.020>.

51 Fergal Reid and Martin Harrigan, 'An Analysis of Anonymity in the Bitcoin System' in Yaniv Altshuler et al (eds), *Security and Privacy in Social Network*s (Springer New York 2013); George Kappos et al, 'An Empirical Analysis of Anonymity in Zcash' (2018).

52 Alex Biryukov, Dmitry Khovratovich and Ivan Pustogarov, 'Deanonymisation of clients in Bitcoin P2P network' (2014) abs/1405.7418 CoRR.

53 Arthur Gervais et al, 'On the Privacy Provisions of Bloom Filters in Lightweight Bitcoin Clients' (Proceedings of the 30th Annual Computer Security Applications Conference, 2014).

54 Hackernoon.com, 'Bitcoin Upgrades with Dandelion: The Transaction Privacy Protocol' (2019).

pseudo-anonymous transaction details can be subject to further analysis such as clustering (eg multi-input heuristic[55]) that aim to attach multiple addresses to a single (known or unknown) entity. Transaction data on permissionless blockchains has to be deemed publicly available. The data can reveal sensitive financial information about the data subject and can hence state a severe interference with fundamental rights. Given the complexity of the system itself, limited technical knowledge of the average user and the analytical approaches, users can not be expected to foresee third party processing of their transactions. As pointed out before, the public availability can hence not be reinterpreted as consent for processing beyond what is necessary to perform the transaction nor as *manifest publication*. Beyond direct access to the blockchain, some services make additional information on transactions publicly available via APIs[56] (see IV.6).

8. Conclusion

With technology developing in all areas of life, novel public data sources arise on a daily basis. Be it wrongfully configured Internet of Things (IoT) devices or blockchain-based cryptocurrencies and smart contracts. The combination of different data sources can easily be (mis-)used to create profiles on natural persons, which can be of interest for law enforcement. It is hence that data analysis in this area is often particularly focused on being able to create such profiles.

V. Analytical Methods

Following the data sources, it is necessary to evaluate the data analysis concepts that are applied to the (publicly available) data. The following section will focus on three general steps in the analytical approach. The mere extraction of publicly available data[57] is often relatively easy and research is hence more focused a specific analysis and learning models to make the processing more efficient or create novel insights. Depending on the source data and protocols, different approaches are available.[58] For example websites can be scraped for text content[59] or for media content[60]. Similarly, there are myriads of approaches for other networks, such as blockchain-based cryptocurrencies.[61] All these approaches have in common that they are increasingly driven by machine learning and similar AI approaches.[62]

1. Framing

Prior to any processing it is necessary to detect and define the problem that needs to be solved. The problem must be sufficiently narrowed down to allow efficient processing. At this point it is already crucial to define the problem without any unwanted bias. This becomes ever more important where (external) data scientists or developers define the problem based on limited information provided by LEAs. Cross-disciplinary description of a problem will most certainly go along with a loss of information and different perceptions of a problem. For example LEAs request a tool to find drug dealers, the data scientist does not know what substances are illicit drugs under the relevant legislation. The developed algorithm hence accidentally also targets innocent persons. After defining the problem the data scientist needs to identify the most meaningful approach/model to solve it. It is necessary to be able explain why a certain model was used to generate evidence/intelligence from the data. In this regard, it needs to be discussed

55 Sarah Meiklejohn et al, 'A Fistful of Bitcoins: Characterizing Payments among Men with no Names' (Proceedings of the 2013 Conference on Internet Measurement Conference, 2013).

56 eg shapeshift.io, changelly.com.

57 R Mitchell, *Web Scraping with Python: Collecting More Data from the Modern We*b (2018); Jussi Myllymaki, 'Effective Web data extraction with standard XML technologies' (2002) 39(5) Computer Networks <https://doi.org/10.1016/S1389-1286(02)00214-1>.

58 Johnson, Faustina, Gupta, Santosh Kumar, 'Web Content Mining Techniques: A Survey' (2012) <https://pdfs.semanticscholar.org/74a7/c3bb6c249726c006412a25297a1427b54920.pdf>.

59 Ekambaranathan, Meiklejohn and Peter (n 48).

60 Gerold Laimer and Andreas Uhl, 'Key-Dependent JPEG2000-Based Robust Hashing for Secure Image Authentication' (2008) 1 EURASIP Journal on Information Security 1:1-1:19; S S Kozat, R Venkatesan and M K Mihcak, 'Robust perceptual image hashing via matrix invariants'; R Venkatesan et al, 'Robust Image Hashing' (2000); Di Wu, Xuebing Zhou and Xiamu Niu, 'A novel image hash algorithm resistant to print-scan' (2009) 89(12) Signal Processing 2415.

61 Dmitry Ermilov, Maxim Panov and Yury Yanovich, 'Automatic bitcoin address clustering'; Masarah Paquet-Clouston, Bernhard Haslhofer and Benoit Dupont, 'Ransomware Payments in the Bitcoin Ecosystem' (2018); Andrew Miller et al, 'An Empirical Analysis of Linkability in the Monero Blockchain' (2017) abs/1704.04299 CoRR.

62 Mikkel Harlev et al, 'Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning' (Proceedings of the 51st Hawaii International Conference on System Sciences, 2018) <http://hdl.handle.net/10125/50331>.

which approach generates the lowest risk for the fundamental rights of the affected data subjects (data protection) and the suspect (criminal procedure). For example, to create a chain of evidence, generate high evidential value and a verifiable outcome in criminal trials it may be required to process and store additional publicly available data - thereby broadening the scope and interference with fundamental rights of thirds. In the end, these questions can be narrowed down to the weighing of the stately interest of prosecution against the fundamental rights to privacy and data protection of the data subjects. Due to the pace and complexity of novel analytical approaches it appears that the legal systems increasingly lag behind, resulting in an insufficient discussion how to balance these interests. By default, the European and, more importantly, national legislators have not yet been able to create sufficient legal bases that create some legal certainty but rather appear to paralysed in sight of the quickly changing technological possibilities. The lack of legislation in this evermore important area however does not result in a lowered use of novel approaches but rather in a grey zone where anything that is available appears to be usable either through third parties or by LEAs themselves. In this uncertainty, the decisional power lies in the hands of the data scientist who has to decide which model is the most feasible for the current use. It is then his to decide if an individual system needs to be set up or if pre-customised (ie AutoML) approaches can be used. This decision is seldom driven by legal considerations but rather factual ones, eg the knowledge of the data scientist or budgetary limitations. The end-user of the outcomes will seldom be able to understand the underlying model and methodology. He is hence unable to evaluate the evidential value, risks for fundamental rights or even the compatibility with a legal basis but rely on the data scientist. In the following, I will try to depict some of the available analysis models to depict the complexity of the underlying data science. Note that each approach requires extensive additional legal research as well as case-by-case evaluation for the specific use case and can only be a starting point for additional research.

2. Collection

Following the framing, data can be collected according the used model. Where publicly available data is used as training data, it must be acknowledged that the training material in itself might be flawed already. Public data is likely to incorporate pre-existing biases of the user - especially when social media or similar data sources are used. It is hence necessary to collect training data in a manner that ideally excludes unwanted bias.[63] The collection process is steered and limited by legitimate interests of the controller and the data subjects. Legitimate law enforcement interests can only exist within their legal barriers - even if they are inherited by third parties. Other interests are likely to be broader, eg due to the high legal requirements in the financial sector and the lowered risks to the data subjects in comparison to law enforcement activity. Where multiple purposes are pursued, it is favourable to create *Chinese walls* between datasets as far as necessary and possible. Subsets can internally be combined where necessary (eg as a training set for machine learning (ML) approaches) - data provision however must adhere to legal limitations in the specific context. These limitations must be defined prior to data provision and reflected on the technical level (ie only anonymous statistics in pre-investigation phases and only based on a relevant subset).

3. Analysis

As described in Section IV, many valuable sources of publicly available information provide information in unstructured form. The following section will discuss three promising methods to analyse publicly available text, image and cryptocurrency data.

Analytical approaches nowadays often rely on machine learning - a subset of AI. Pursuant to the High-Level Expert Group on AI (HLEG-AI) and the European Commission, artificial intelligence systems *display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goal.*[64] Where AI is applied

63 Karen Levy and Solon Barocas, 'Designing Against Discrimination in Online Markets' (2017) 32(3) Berkeley Technology Law Journal <https://doi.org/10.15779/Z38BV79V7K>; Matthias Leese, 'The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union' (2014) 45(5) Security Dialogue <https://doi.org/10.1177%2F0967010614544204>.

64 European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe' (2018) <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF>; High-Level Expert Group on Artificial Intelligence, 'A definition of AI: Main capabilities and disciplines' (2019) <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341>.

in any form (eg clustering, reinforced learning) to generate intelligence for law enforcement purposes associated risks need to be taken into consideration.[65] When it comes to machine learning, two major concepts should be known - *Convolutional Neural Networks (CNNs)* and *Recurrent Neural Networks (RNNs)*. When data is analysed based on these concepts, it is necessary for users to understand their risks and advantages to be able to put outcomes into perspective. CNNs, for example, are especially powerful for pattern detection and are hence particularly helpful in image processing with fixed inputs. To be able to review CNNs outcomes, the reviewer must be aware which patterns can be detected (eg edges or symbols), which filters were used and where shortcomings might be. RNNs on the other hand do not rely on fixed inputs and are hence easily applicable to data sequences such as text or speech. One example for RNNs is the so-called *Natural Language Processing* (NLP). It describes a method to analyse unstructured text data and is hence particularly important to analyse textual content from the surface and/or dark web. In this approach the text is usually split into *tokens* (eg words, characters), that are then often subject to *stemming* (ie reducing inflectional forms in a text) to ease analysis. These reduced tokens are then usually be vectorised based on different models. (eg ngram, bigram). For programming languages like Python there are already libraries like NLTK[66], SpaCy[67], Gensim[68] or TensorFlow[69]. All of these models have different, though overlapping, use cases and most of them come with pre-trained statistical models and word vectors that allow data scientists to quickly analyze text data. However, if these libraries are used it is necessary to understand their shortcomings. If the analytical outcomes don't have sufficient evidential value, the interests of the affected data subject will usually outweigh interests of the controller/LEA. In addition, ML approaches usually provide probabilistic outcomes that require evaluation by the end-user. Failing to understand the underlying libraries easily results in wrongful suspicions and must hence be open to scrutiny. For example, the analysis wrongfully identifies participants of a political discussion as suspects in an investigation because the term 'stealing' appears above average in comparison to the underlying statistical model or the authorship attribution wrongfully connects multiple posts in a forum to a single entity because the style of writing of two entities matches with 95% certainty.[70]

Example - Cryptocurrency Analysis: Although exchange services nowadays fall under AML-Regulation and hence have to comply with KYC, many owners of keys remain pseudo anonymous. Driven by various interests there is a long line of research targeted at de-anonymizing participants in cryptocurrency networks - this is, of course, also in the interest of LEAs. The public availability of the blockchain is the starting point for multiple analytical approaches such as clustering or outlier detection. Cryptocurrencies generally allow participants to create an unlimited number of private/public-key combinations which are roughly comparable to bank account numbers with no names attached to them. A starting point for investigations is to find subsets of public keys that are likely to belong to the same entity. Based on the underlying protocols of cryptocurrencies this so-called *clustering* aims to do so by checking (all) transactions based on certain assumptions (heuristics). For Bitcoin one of these assumptions is that if a transaction stems from multiple input addresses, they are likely to belong to the same entity (multi-input heuristics[71]). This assumption is based on the fact that Bitcoin transactions require a private key that (ideally) is only known by the owner of the address (similar to a TAN/PIN when issuing a classic transaction). Many of the currently existing cryptocurrencies are based on the Bitcoin protocol (eg Bitcoin Cash, Litecoin) making this approach applicable in many contexts. But also, more privacy-centric cryptocurrencies like ZCash[72] or Monero[73] have been shown to allow at least partly de-anonymisation of participants even for transactions across multiple blockchains.[74] Due to a lack of *ground truth*, the ac-

65 Levy and Barocas (n 64); Karen Hao, 'This is how AI bias really happens—and why it's so hard to fix' (2019) Technology Review (MIT) <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/>.

66 'Natural Language Toolkit' <https://www.nltk.org/>.

67 'spaCy' <https://spacy.io/api/doc>.

68 'Gensim' <https://radimrehurek.com/gensim/apiref.html>.

69 'Tensorflow' <https://github.com/tensorflow/docs>.

70 Ekambaranathan, Meiklejohn and Peter (n 48).

71 Meiklejohn et al (n 56).

72 Kappos et al (n 52).

73 Abraham Hinteregger and Bernhard Haslhofer, 'An Empirical Analysis of Monero Cross-Chain Traceability' (2018) abs/1812.02808 CoRR.

74 ibid; Haaroon Yousaf, George Kappos and Sarah Meiklejohn, 'Tracing Transactions Across Cryptocurrency Ledgers' (2018) abs/1810.12786 CoRR.

curacy of the underlying assumption currently cannot be proven, resulting in unclear evidential value.[75] In trials the existing uncertainty can at least partly be overcome by pairing 'probabilistic evidence' with factual evidence - ie checking probabilistic outcomes against deterministic analysis of seized devices. To overcome the (potential) lack of accuracy clustering, outcomes can be enriched with further data - often from other publicly available sources. This does not change the accuracy of the clustering but can increase the information density of the dataset and hence be an indicator for the plausibility of the analysis. Higher information density is hence an important selling point for commercial tools in the area of cryptocurrency analytics. Besides connecting it with publicly available data the information density can also be coupled with additional analytical methods such as outlier/anomaly detection[76] or ML cluster-classification.[77] This bundle of analytical and enrichment processes often requires extensive computational effort and is hence seldom conducted by LEAs directly. Consequently, these approaches usually take place prior to and/or independently of law enforcement activity/investigations. When such techniques are used, supervision and review is only possible if the underlying technological concepts are understood. In addition, it is necessary to discuss the evidential value of such outcomes. It is, for example, yet not clear if outcomes of clustering are sufficient to raise an initial suspicion to start an investigation, or if this clashes with the general prohibitions for fully automated decisions laid down in Article 22 GDPR or Article 11 LED.

4. Conclusion

Framing, collection and analysis are all highly critical steps in law enforcement work. Nowadays some of these steps are outsourced to private parties with different appreciation and understanding of fundamental rights. It is hence necessary to discuss if society accepts that private parties define what material may be relevant in investigations. The key question in this regard is if processing for law enforcement purposes can constitute an acceptable purpose under the GDPR. The flexible legal bases in Article 6 GDPR may be interpreted to allow all kinds of data processing as long as the fundamental rights of the data subjects are sufficiently protected. In this light, the meaning of the purpose would be reduced to a mere measure for the scope of processing (purpose limitation). However, this argumentation oversees that the purpose must still be lawful - ie requires a legal basis. Since public security is a domain of public sovereignty, the GDPR is unable to provide such a legal basis. In addition, the GDPR does constitute a preventive ban on processing of personal data that is subject to special legal permission - ie specific legal bases. It is hence currently up to member states to provide a sufficiently clear national legislation for law enforcement data analysis. To my best knowledge, this is currently not the case in any Member State. Instead, data processing ever so often is wrongfully based either on the GDPR itself or on general clauses in national law. The latter will seldom conform with the requirement for sufficient precision of the law and hence only allows minor interferences with fundamental rights. Due to uncertainties in the analysis, the required amounts of data and the ability to connect data points, the use of AI-driven analysis is unlikely to constitute such a minor interference. Where national legal bases are provided, it is at least necessary to equip supervisory authorities and end-users with sufficient manpower and knowledge to be able to review the provided data. This is currently not the case. The reliance of expert witnesses in criminal trials to overcome these issues is not sufficient as they will only be able to review outcomes in the individual case but not for cases that never go to trial.

VI. Processing Constellations

Beyond the data sources, the context, and the methodology of processing it is important to consider the specific constellations in which these actions take place. There are uncountable possible constellations for the analysis of publicly available data - often consisting of many players, that knowingly or unknow-

75 Michael Fröwis et al, 'Safeguarding the Evidential Value of Forensic Cryptocurrency Investigations' (2020) Forensic Science International: Digital Investigation <https://doi.org/10.1016/j.fsidi.2019.200902>.

76 Masarah Paquet-Clouston, Bernhard Haslhofer and Benoit Dupont, 'Ransomware Payments in the Bitcoin Ecosystem' (2018) arXiv preprint arXiv:180404080.

77 Francesco Zola et al, 'Cascading Machine Learning to Attack Bitcoin Anonymity' (2019).

ingly work together to generate intelligence or evidence. Many of these constellation result in opaque legal situations making it hard for supervising authorities, data subjects and defendants to review or challenge the underlying processes to the degree envisioned in criminal procedure as well as in data protection law.

1. Relevant Actors

To understand the constellations and evaluate accountability of the actors, it is necessary to identify the controller and processor in the respective constellations. Pursuant to Article 4(7) GDPR controller means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data. In contrast, under the LED only *competent authorities* can be controllers. Article 4(8) GDPR and Article 3(9) LED identically define the 'processor' as a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller. In addition, some players in this realm (eg Europol, Interpol) currently completely fall outside the scope of unified data protection regulation. The Europol Regulation[78] does not differentiate between controllers and processors as it does not fall under European-bodies regulation[79] but is based on an individual legal framework. Similarly, Interpol does not fall under EU legislation at all, only being subject to their own 'Rules on the Processing of Data'[80] that also do not know the concept of controllers and processors.

2. Single Party Processing

LEAs historically gathered information on their own and have it evaluated by their internal experts, data forensics and so on. In these cases, the LED and its respective national transposition remains fully applicable for each step. It is hence necessary that data collection and processing have a sufficient legal basis either in police law or criminal procedure codes. Similarly, private parties can process data for forensic purposes outside law enforcement activity (eg for forensic research) under the GDPR. The single party is hence controller and processor. This constellation does not raise additional risks and the processing rules in this constellation are comparatively well defined, although they have not yet been adapted to novel data driven technologies. As a consequence, law enforcement practitioners often perceive data protection as a heavyweight impractical instrument that is limiting any data driven investigation efforts. This is true to an extent - due to the lack of a legal basis there is no balancing weight for data protection legislation but rather a prohibition to use novel data driven approaches at all.

Example 1: The LEA lawfully acquires a picture from a publicly available source and requires internal experts on image analytics to identify if the picture was photoshopped or not.

Example 2: Internal experts evaluate content of a seized device.

3. Controller-Processor

In this constellation the LEA outsources certain tasks to a private party but remains in control of the processing. In this case the legal limitations of LEAs are usually conveyed to the private partner through contractual agreements. In this scenario the LEA is the controller and the forensic institute the commissioned data processor. This processing approach is also relatively unproblematic and is governed by the criminal procedure code and the LED through the contractual agreement.

Example 1: LEA assigns a forensic institute to analyse a lawfully seized mobile phone.

Example 2: The LEA identifies a relevant transaction on the blockchain. Due to a lack of internal expertise, an external data forensics expert is charged to analyse the specific transaction flow for the specific investigation. The expert subsequently analyses the data based on a specific contractual agreement and for an explicitly defined purpose.

78 Regulation 2016/794 (n 16).

79 Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (Text with EEA relevance) [2018].

80 Interpol General Assembly, 'Interpol Rules on the Processing of Data - III/IRPD/GA/2011 (2016)' (2016).

4. Controller-Controller

When it comes to novel data driven investigation methods many commercial parties aim to provide their services to LEAs, often as subscription models. The cooperation still includes contractual agreements between the parties that require data providers to ensure that only lawfully collected information is provided. However, LEAs do not have direct control of the data processing of the private party. In comparison to the controller-processor relation, the private party acts independently and also does not fall under the LED nor under criminal procedure codes. This concept appears to be compelling because at first sight the legal limitations of private parties to gather data are much wider and flexible than the ones for LEAs. Legal limitations of LEAs can only be imposed on the party contractually. Where contracts are not sufficiently precise, the collection and analysis of the data is remains governed by the GDPR which provides a bouquet of possible legal bases for data processing. Usually the initial processing of the private controller is based on Article 6(1) lit. d, e or f GDPR. These legal bases are only speciously applicable, though. Article 6(1) lit. d GDPR applies if the processing is necessary to protect vital interests of the data subject or of another natural person. Since the processing takes place for LEAs and hence to protect public - it can be argued that it is necessary to protect vital interests of the data subjects and natural persons. Article 6(1) lit. e GDPR requires necessity of the processing for a task carried out in the public interest or in exercise of official authority vested in the controller. Since law enforcement activity is generally deemed within the public interest, lit. e may also be applicable. In addition, the gathered data and analytical approaches are often used for other purposes such as research. Article 6(1) lit. f GDPR requires the controller to have legitimate interest in the processing that needs to weighed against interests of the data subjects. This legal basis is arguably the most common one, and usually the following line of argumentation enfolds. First, the economic interests of the private party are generally deemed legitimate. Second, the processing for law enforcement purposes which could arguably also a legitimate interest. Third, the data subjects' interests are not worthy of protection due to their public availability (see misconception of public availability in Section III). Fourth, even if the data subjects' interests are put into consideration, the eventual access to the data is still at the hands of LEAs. For example, the law enforcement officer will only search the database for a specific individual 'Suspect X' and hence only receive related information. As a consequence, private data provides usually classify their processing as lawful speciously deemed lawful from a GDPR perspective. Unfortunately, none of the arguments above changes the fact that the EU is unable to conclusively govern - ie allow - data processing for law enforcement purposes in the GDPR since the area of public security is still subject to national sovereignty. This means that the private processing for law enforcement purposes usually lacks any legal basis. The common perception, however, seems to be that processing limitations only indirectly arise from additional issues such as admissibility of so generated evidence in a trial due to the uncertain legal situation and a lack of understanding the underlying data analysis. As observed in other contexts, this increases the risk for *parallel construction*.[81] Last but not least, the collection and analysis of publicly available data often follows a 'multi-purpose' business model. The ability to extract information from publicly available data is of course not only relevant for LEAs but also for many others (eg research, financial institutions, etc). The purpose of the initial collection hence often defined by interests of multiple parties and can be broken down to 'processing for the purpose of providing tailored insights into data for the respective customers'. As a consequence, the initial data processing scope can easily be expanded and remain compliant with the GDPR. While this may work in private contexts, the provision of outcomes to LEAs remains incompliant with the initial purpose (Article 5 (1) lit. b GDPR) and lacks a legal basis.

Example: Company A scrapes, stores and analyses darkweb data over two years to be able to create and sell insights on darknet activity based on Article 6(1) lit. f GDPR. Customer 1 is a pharma company who needs insights on illicit trading with their drugs. Customer 2 is a LEA investigating drug dealers. Customer 3 is a cryptocurrency exchange service that requires information on pseudo-anonymous transactions to be able to comply with AML regulations. Company

81 Natasha Babazadeh, 'Concealing Evidence: 'Parallel Construction,' Federal Investigations, and the Constitution' (2018) <https://ssrn.com/abstract=3319387>.

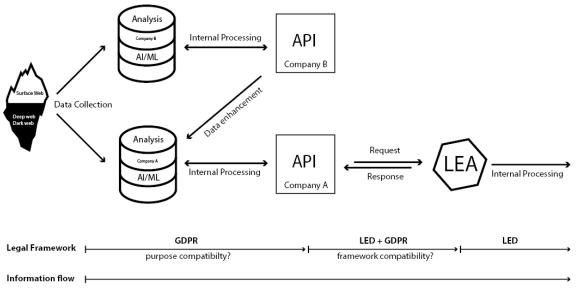

***Figure 1**. Processing constellation example*

A gathers and analyses data based on its customer requirements. All customers are provided with relevant information based on the whole dataset. None of the customers knows or understands the underlying machine learning algorithm that was used to identify the provided information due to commercial secrets.

5. Joint Controllership

In the section above, the data collection and analysis happened independently. If LEAs and private companies team up to process data, this often results in a factual joint controllership situation - ie both parties jointly control the purposes and means of the processing (cf Article 26(1) GDPR). Unfortunately, it remains unclear if the concept of a joint controllership is applicable between the legal frameworks of the GDPR and the LED. The LED only allows *competent authorities to be controllers*, making it impossible for private parties to fall under this Directive. As described above data processing relies on the use of APIs to outsource computational approaches or extend functionality in general (cf Maltego, IBM Analyst Notebook, Palantir). In these cases, the API provider usually defines the possible requests and underlying data processing - the requesting party further specifies the request for an individual case - in short, both parties control the processing (see Figure 1). The use of APIs hence usually results in a joint controllership.[82] Where LEAs are involved these constellations fulfil all conditions of joint controllership except one. Consequently, even if LEAs and private parties collaborate, they *currently* have to be classified as individual controllers under their respective legal frameworks. An overarching joint controllership is not yet foreseen in the European data protection framework. A 'practical' joint controllership between private and public parties hence currently results in the same risks discussed in Section VI-D. In this context it is necessary to discuss if the provisions of the GDPR or the LED need to be adapted to cover (practical) joint controllership constellations. Since the GDPR does not cover for law enforcement purposes in general, the lacking reflection of these types of cooperation currently results either in illicit and uncontrolled data processing or severely constraints the LEAs capability to access novel data driven methods.

Example 1: Company A and LEA decide to start a cooperation. Both identify the specific requirements and limitations and conduct the processing accordingly.

Example 2: Company A is individually scraping and analysing data that it assumes to be valuable for LEA customers (single controller). Outcomes of the processing are provided to customers through an API (joint control).

Example 3: Company A provides a software that integrated an API from Company B that allows on demand (live) analysis of websites (joint control with LEA+B).

6. Conclusion

Private-public processing constellations, are often used to circumvent constitutional limitations based on the legal uncertainty revolving around the applicability of legal frameworks. It is hence necessary to identify if the GDPR is at all applicable in these circumstances. In Recital 88, the GDPR at least acknowledges the 'legitimate interests of law-enforcement authorities'. This could be seen as an indicator that such interests are *limited to* competent authorities. Otherwise it could just have referred to 'law enforcement purposes'. Article 23 GDPR particularly allows the restriction of data subjects' rights for the purposes of prevention, investigation, detection or prosecution of criminal offences. The Regulation hence at least knows the possibility that data processed under the

82 Case C-40/17 *Fashion ID GmbH & Co.KG gegen Verbraucherzentrale NRW e. V.* [2018] ECLI:EU:C:2019:629; Case C-210/16 *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein gegen Wirtschaftsakademie Schleswig-Holstein GmbH* [2018] ECLI:EU:C:2018:388.

GDPR can be relevant to law enforcement. At the same time, the EU (and hence the GDPR) is not empowered to govern the core area of public security of the Member States. These limitations hence only concern secondary processing types such as general administration, traffic regulation, handling of environmental disasters, etc.

The state-of-the-art (SOTA) analysis of publicly available data for investigative purposes is, however, part of the core area of law enforcement and can hence not be governed by the GDPR. In reality, however, the initial processing of data by private parties if usually based on the GDPR and the companies legitimate economic interests. The collection in itself can hence often be lawful if it is also conducted for other (GDPR-compliant) purposes. Since initial data processing for law enforcement purposes is not compliant with the GDPR the provision of data to LEAs can also not constitute a lawful *further processing* under Article 5 lit. b GDPR. Consequently, the constellations as described above are widely incompatible with the GDPR and constitute illicit data processing. Such limitations thwart LEAs from much needed capabilities to access SOTA analytical methods, in particular when it comes to publicly available data. From a legal perspective it would hence be desirable to open up the LED insofar as to allow joint controllership constellations under the law enforcement framework and provide novel legal bases for data driven investigations. This would create certainty with regard to the scope of processing help balancing the use of SOTA technology[83] against the fundamental rights of the data subjects. The flexibility provided by the GDPR, however, is not sufficiently precise to be applicable in a law enforcement context. Even if the GDPR would be deemed applicable, further inconsistencies between the two frameworks would arise. For example, Article 14 GDPR requires the controller to provide information about the processing (eg contact details, purposes of processing, data categories). The GDPR also includes exemptions from this requirement, such as that the requirement is not applicable where it proves impossible or would involve a disproportionate effort (Article 15(5) lit b GDPR). The LED on the other hand does generally limits information obligations but does not know *effort* as a relevant factor. In contrast to the GDPR, Article 25 LED requires detailed logging of *collection, alteration, consultation, disclosure including transfers, combination and erasure*.[84] Due to the vast amount of affected data subjects and the often pseudonymous nature of the data, the private controller will usually be able to substantiate that the direct provision of information to data subjects is at least a disproportionate effort. In this scenario, the GDPR requires the controller to take appropriate measures to protect the rights and freedoms and legitimate interests of the data subjects and especially names the possibility to make information about the processing publicly available. It is yet unclear which degree of detail and which channels of publication are necessary. In one of the first cases concerning publicly available data the respective authorities suggested that the controller could have used SMS messages or commercial television spots to provide information to the data subjects.[85] These inconsistencies expressly underline that the EU legislator did not foresee cooperation involving both frameworks. In addition, the LED only allows *competent authorities* to be controllers of law enforcement data processing - currently excluding private parties from this possibility. This does not result in a regulatory gap since the requirements in themselves are conclusive. It does however result in a gap between practical requirements and legal possibilities. LEAs regularly lack sufficient legal bases, technical knowledge and budget to use such novel data driven technologies. Consequently, it appears that some kind of 'shadow-market' has evolved, where private parties provide data that was gathered under their own control to LEAs. It comes as no surprise that both parties happily want to believe that such approaches are compatible with the existing data protection framework as LEAs urgently require access to the data analysis and private companies access new markets. Having said that, even if the (API-)access of LEAs to private data is restricted, such types of cooperation are currently incompatible with the European data protection framework. From a legal perspective the current system basically results in a prohibition of LEAs to access SOTA data driven technology. From a practical perspective, however, the legal uncertainty in this area appears to result in

83 Europol (n 7).

84 Jarj Sajfert and Teresa Quintel, 'Data Protection Directive (EU) 2016/680 for Police and Criminal Justice Authorities' (2019).

85 UODO, 'The first fine imposed by the President of the Personal Data Protection Office' (2019) <https://uodo.gov.pl/en/553/1009>; WP29, 'Guidelines on transparency under Regulation 2016/679' (2017) <https://bit.ly/2Qn0gu1>.

unregulated cooperation between public and private parties.

VII. Proposal for a Centralised Data Clearinghouse

The previous sections show that LEAs' access to SOTA data driven technologies is limited by multiple factors. First, the legal system does not allow sufficient cooperation with private parties. Second, LEAs lack legal bases to use SOTA analysis themselves. Third, LEAs (often) lack budget and knowledge to be able to make use of SOTA technologies. Fourth, a general legal uncertainty with regard to analytical methods hinders efficient discussion of the aforementioned issues. As shown above, the sometimes perceived 'legal grey zones' do not actually exist. Instead private parties are currently unable to control data processing for law enforcement purposes under the EU data protection legislation. The identified gap is of practical nature and legislators have yet to provide a sufficiently detailed framework to either allow direct use of novel technology (= legal basis) or make it accessible through private-public-cooperation to enable LEAs to conduct necessary cyber-investigations using data driven methods. A solution should ideally address all of the aforementioned factors while the competitiveness of enterprises providing data analysis should not be undermined. In this section I will propose a 'centralised data analysis institution' (CI) as a possible solution to tackle existing challenges in a balanced manner. While undoubtedly complex, it is likely to be still easier than approaching these issues individually on the national level.

1. Setup

A centralised institution to control access and use of novel data driven approaches for LEAs could be designed similarly to other intermediary bodies like Europol or Interpol that already provide communication frameworks (SIENA/I24-7) to share evidence and intelligence. These institutions do not yet officially act as intermediaries when it comes to private public cooperation - although both in fact provide (third party) analytical support to LEAs in some cases. The legal framework must specifically allow cooperation with both sides and include a sufficiently clear legal framework for data collection and analysis. From a technical perspective the CI would be able to provide de facto standards for private parties to provide data, while private parties would still be unable to control data processing for law enforcement purposes. Data provision to the CI would be seen as an intermediary step merely aimed at data analysis rather than law enforcement activity. On the other side the CI could provide LEAs with 'clean' data and ensure only relevant and lawfully gathered information is provided to them. Private companies could provide their services through an API in a formalised way (eg Unified Cyber Ontology (UCO)[86], Cyber-investigation Analysis Standard Expression (CASE)[87]). Since they only act with an intermediary that is not a LEA, their cooperation could be covered by the joint controller model under the GDPR. The CI would hence be an independent organisation with no law enforcement powers, but rather a data cleaning house to ensure the protection of fundamental rights of data subjects. In the long run, such an institution could also act as a clearing house for other institutions, such as banks.

2. Advantages

A centralised platform could foster development and enforcement of the legal framework, technical and operational requirements. A central contact point for all relevant actors (eg developers, data scientists, legal experts, data subjects) would foster development of novel unified standard with regard to processing of publicly available data sources and would be in line with the international cooperation mechanism envisioned in Article 50 lit. a GDPR

a. Ease of Access

Collaborative approaches are often based on API-driven access to analytical methods. To date there is not standardised form for requests and responses for any analytical method. A centralised player could provide technical interfaces (APIs) for LEAs and technical

86 Casework, 'Unified Cyber Ontology (UCO)' (2018) <https://github.com/Ebiquity/Unified-Cybersecurity-Ontology>.

87 Fröwis et al (n 76); Casework, 'Cyber-investigation Analysis Standard Expression (CASE)' (2018) <https://github.com/casework/case>.

partners, fostering access in a formalised and replicable way - eg upload a picture or request analysis of a currency address.

b. Ease of Supervision

To ensure fundamental rights are sufficiently protected, legal limitations need to be enforced on all relevant levels. On the technical level, supervision and enforcement can take place based on evaluation of API requests and responses (eg easy exclusion of particular information from a standardised data format). Analysis of API-use can further help detect unusual/illicit request/response patterns and strengthen the data protection authorities' (eg EDPS) position. Since all requests and responses go through the centralised institution, 'parallel construction' can be prevented. Centralised logging further fosters the defendant's ability to review the underlying data processing. There should also be support by a supervisory authority, eg the EDPS as a central institution could further conduct regular audits and plausibility checks for the provided data. In contrast to uncountable, opaque and often illicit collaborations between LEAs and private parties that are often subject to nondisclosure agreements, so that data subjects seldom even get to know about the processing, a central platform provides a single point of contact for data subjects to be informed about the processing of their personal data, easing the private companies' compliance with transparency requirements.

c. Strengthening Evidential Value

A centralised institution could additionally help to strengthen evidential value. To foster the evidential value of the analytical processes it might be necessary to equip such an institution with additional powers to access external databases for (anonymous) review purposes. This way, probabilistic processing approaches could be measured against ground truth (eg KYC databases of banks, Europol databases etc). Centralised expertise would further ease appointment of experts in trials and strengthen LEAs' training capabilities with regard to novel data driven methods.

d. Adaptive Enforcement of LEA Limitations

The restriction of law enforcement capabilities can be guaranteed through access control - effectively avoiding over-regulation on the one side and strengthening protection of fundamental rights on the other. The necessary restriction of stately power is enforced through the accessibility of data, secured by an independent supranational party.

e. Economic Advantages

From an economic perspective a CI would allow private parties to provide their analytical services to a broader LEA community. Remuneration could relatively easily be determined through API-measurements (pay-per-use model). With remuneration based on actual usage, private companies would may even have a higher incentive to provide high quality data analysis. LEAs would hence be able to gain access to novel data driven analytics without the requirement for expensive long-term contracts or subscription models with individual private parties. A central agency could easily pool money from all LEA participants/member states and would have much better contractual position than individual LEAs. Private companies on the other hand would have central point to provide the law enforcement market with their novel tools in a formalised and certified manner. SMEs would hence bypass lengthy and complex negotiations with multiple LEAs.

f. Standardisation

The standardisation of formats is a long-time issue. Making it necessary to cooperate with a centralised institution will foster development and enforcement of standards in a meaningful way. Data processing providers can continue valuable economic cooperation that fosters the development of novel technologies and strengthens their competitive capabilities. To achieve this, standardised exchange formats for analytical outputs are necessary. A single point of contact would foster the effort to unify exchange formats and ease accessibility to a market for private analysis providers. Data processing methods could additionally be certified/evaluated prior to their use in investigations and outcomes can be checked against other data sources without risks for the data subjects.

g. Interdisciplinary Collaboration

A central analysis intermediary would further constitute a helpful platform to foster interdisciplinary

discussion and development. The insights in analysis usage could for example help to allocate research funding to meaningful channels. Similarly, LEAs would have a specific point of contact to formulate their requirements and challenges instead of hoping for novel tools and powers in a nebulous national market.

3. Challenges

A key question that remains is that private parties would still, although indirectly, gather and analyse data for law enforcement purposes. Given that in most cases, the data processing takes place anyway for other purposes, it could at least be argued that the collaboration with the CI lowers the risk for the data subjects' fundamental rights that arises due to the current opaque processing constellation. In addition, the framework could be constructed in a way that data provision to the CI is only possible where the initial data processing was based on GDPR-covered purposes. With the limited capability of the EU to regulate public security on national level, Member States would still be able to create novel legal bases that allow data processing for law enforcement purposes. However, with a powerful centralised data provider, the incentive to rely on individual contracts with private companies is likely to be limited. In addition, this approach can still be coupled with in-house analytical capabilities of LEAs. Ideally these would be consistent with the standards set up by the centralised institution and could even provide their analytical capabilities through that institution in reverse direction. Requests and data handling must be formalised as much as possible to allow efficient processing and supervision by a CI. It is hence possible that such an approach is unable to reflect complex analytical requests. In this context the CI however could provide specific points of contact to handle such requests. Similarly, requirements for electronic evidence need to be formalised and enforced prior to the (intermediary) provision of information. A key challenge would be the transcription from formal (national) limitations to technical specifications. The centralised introduction of specifications still seems to be more promising than national solo efforts. In addition, if specifications are set up, further development and adaptation for novel approaches is relatively easy.

VIII. Conclusion

As laid out in this article, novel artificial intelligence approaches in law enforcement come along with a number of additional issues. While the individual consideration of these issues is unquestionably important, it is also worth taking a look at the bigger picture. This starts with the common misconception of the consequences of public availability, includes discussions on relevant data sources and analytical methods and is framed by a necessary debate about processing constellations. Only if all levels are put into consideration an overarching consensus can be reached and enforced. Legal practitioners and scholars are hence required to dive into the depths of technical, legal and societal issues at the same time to find suitable solutions. This will only be possible if interdisciplinary collaboration is further intensified. To date, it appears that the lack of interdisciplinary work has resulted in a scattered system prone to misuse and erosion of the fundamental rights to privacy and data protection. A quick solution is not in sight as the data protection legislation seems to struggle with novel technological approaches such as artificial intelligence. Especially in the area of law enforcement it needs to be discussed if changes in legislation are sufficient to address upcoming challenges or if additional - more practical - measures are required. The use of private companies for data analysis can be an efficient way to provide LEAs with access to novel technologies. The current legal framework widely fails to create adequate rules for this access, though. The resulting uncertainty appears to be either misused by private data providers or results in a complete lack of such technologies. The proposal for central data analysis institution may be seen as food for thought for novel approaches to protect fundamental rights to data protection and privacy in a time when data driven investigations are becoming the norm rather than the exception.