CHAPTER 3

Big Data: The Latest Tool in Fighting Crime

Benjamin C. Dean

Benjamin C. Dean
President, Iconoclast Tech

A confluence of trends around digital technologies, data collection, and data analysis over the past two decades has brought new opportunities and challenges to public and private organizations alike. Digital technologies and data analysis can and are increasingly used to identify "bad actors" so as to detect and deter or prevent fraud, money laundering, bribery, terrorism, regulatory non-compliance, and other criminal activities. A variety of techniques are now used including profiling, metadata collection, network analysis, data fusion, and predictive analytics. While powerful when used properly, data and data analysis are still subject to statistical and economic limitations. Organizations require people with new skills and a realistic understanding of what these technologies can and cannot do to be able to effectively deploy these technologies and analytical techniques.

After briefly defining relevant terms and outlining trends that have driven advances in digital technologies, this chapter provides an overview of ways in which organizations are taking advantage of advances in digital technologies and data analysis to profile, track, and mitigate malicious actors. Case studies are provided throughout to illustrate the strengths and weaknesses of each of these methods. The final section provides some recommendations based on the issues raised throughout the chapter.

Definitions

"Bad actors" are defined as those individuals or entities whose activities are in contravention of the laws or policies of the United States and other authorities. Examples of such actors include transnational criminal organizations and human traffickers; those conducting financial crimes such as counterfeiting, money laundering, and fraud; terrorists and terrorist organizations; and malicious actors in cyberspace, which encompasses

An illegal diamond dealer from Zimbabwe displays diamonds for sale in Manica, near the border with Zimbabwe. *Photo credit:* Reuters/Goran Tomasevic.

threats emanating from a range of entities—from nation-states to individual actors.[111]

"Digital technologies" are defined as technologies that "fulfil the function of information processing and communication by electronic means, including transmission and display, or use electronic processing to detect, measure and/or record physical phenomena, or to control a physical process."[112] Data, the plural of datum, are information in binary form that can be digitally transmitted or processed.

Technology Trends

Three technological trends related to digital technologies have dovetailed over the past two decades: Faster and cheaper computing power, commonly referred to as Moore's Law, has seen the price of a fixed amount of computing power halved approximately every eighteen months.[113] Network bandwidth has become faster, doubling every nine months.[114] Growing data storage has seen the cost of data storage halved approximately every twelve months.[115]

These technological advances have the potential to create new opportunities for governments and corporations. The adoption of big data analytics has grown in parallel with these advances and has allowed for increased use and experimentation to help increase tax transparency, decrease corruption, counter terrorism, and reduce fraud.

At the same time, these same technologies are also enabling state and non-state actors to promote

111 Department of Defense, *Identity Activities Joint Doctrine,* Note 2-16, August 3, 2016, http://www.dtic.mil/doctrine/notes/jdn2_16.pdf.

112 This is an adaptation of the definition from the Organisation for Economic Co-operation and Development *Glossary of Statistical Terms* (2004) for information and communication technology goods.

113 Ibid.

114 Dan Geer, "Data and Open Source Security," nominal delivery draft for Recorded Future, October 21, 2014.

115 Ibid.

Table 3.1: DoD Categories of Identity Attributes

Biographical	Biological	Behavioral	Reputational
Identity Attribute Sub-Elements			
Core personal Addresses Employment Educational Military service Family Cohabitants Aliases	Individual static Physical attributes (hair/eye color) Scars, marks, tattoos Familial Group Fingerprints, iris, face, palm print, voice, and DNA	Financial transactions Law enforcement records Digital personas Social affiliations Commercial transactions Media consumption/production Body language (gait, posture, eye movements, hand gestures, typing patterns) Micro-expressions (brief involuntary facial expressions)	Judicial judgements Sworn statements Public licenses Financial (historical) Community observations Employer evaluations

Source: DoD, *Identity Activities*.

violent ideologies; obtain and transfer illicit funds; recruit and train personnel; arrange transport, arms, and equipment; and sustain operational communications."[116] The impacts of these crimes can be costly for public and private organizations alike.

Opportunities of Data and Digital Technologies

Advances in digital technologies around collection, analysis, and secure storage over the past two decades have thus simultaneously brought immense opportunities and significant challenges. Many organizations are now taking advantage of advances in digital technologies and data analysis to profile, track, and mitigate malicious actors. This section examines some of the ways in which these technologies and data analysis are being used for this purpose.

Profiling

Profiling is the act or process of extrapolating information about known identity attributes (traits and tendencies) pertaining to an individual, organization, or circumstance.[117] Identity attributes can be categorized in four ways: biographical, biological, behavioral, and reputational.[118] Identity attributes can subsequently be organized into multiple sub-elements to support data collection, analysis, and management. The US Department of Defense (DoD) has developed at least five hundred such data types and sub-types associated with identify attributes (see table 3.1).[119]

If the attributes commonly associated with a particular category of bad actor can be identified, a "signature" (or "fingerprint") can be constructed for that actor. Subjects' profiles can then be compared against this signature to flag potentially undesirable actors and activities.

Box 3.1. The Total Information Awareness Project and Its Ancestors

In 2002, the Information Awareness Office of the Defense Advanced Research Projects Agency (DARPA), led by Dr. John Poindexter, began developing the Total Information Awareness project (later the Terrorism Information Awareness project). The project was premised on the belief that terrorist activity has an information signature.[120] It was hoped that by identifying these signatures, patterns of activity or transactions that

116 Department of Defense, *Identity Activities.*

117 Adapted from the Merriam-Webster Learner's Dictionary full definition of "profiling."

118 Department of Defense, *Identity Activities.*

119 Ibid.

120 John Poindexter remarks, *Overview of the Information Awareness Office*, DARPATech 2002 Conference, Anaheim, California, August 2, 2002, https://fas.org/irp/agency/dod/poindexter.html.

analysts had predetermined were associated with terrorist attacks could be used to scan through databases (containing phone calls, emails, text messages, rental car reservations, credit card transactions, prescription records, etc.) to investigate past terrorist incidents and preempt potential incidents in the future.[121] Profiling by determining which individuals exhibited attributes previously associated with terrorists was considered essential to preempting potential incidents.

Following congressional concerns about the project, linked to privacy issues, the Total Information Awareness project was defunded in 2003.[122] Components of the project were later transferred from DARPA to other government agencies including the Advanced Research and Development Activity.[123] One of these components was the core architecture, later named Basketball, which was described as a "closed-loop, end-to-end prototype system for early warning and decision-making."[124] Another component was Genoa II, later renamed Topsail, which analyzed domestic call metadata to help analysts and policy makers anticipate and preempt terrorist attacks.[125]

Today, the ancestors of these elements of the Total Information Awareness project live on in the counterterrorism-related activities of intelligence agencies, law enforcement authorities, and the private companies that develop these services for public authorities. In spite of long-standing issues with regard to the effectiveness of profiling for counterterrorism purposes, both for methodological[126] and practical[127] reasons, a new generation of machine learning and artificial intelligence techniques is now being applied in the hope of overcoming these prior issues.[128]

Profiling has been used for many decades. Advances in technologies are making it more practical and cheaper to integrate identity attribute data from many sources into a single or multi-layered profile. However, some forms of profiling—by their nature—create privacy and civil liberty concerns. Ensuring that adequate oversight is in place to avoid infringing upon relevant legislation is essential to the success of profiling activities.

Metadata

At the most basic level, metadata are data that provide information about other data, giving people an understanding of what the data constitute. For instance, statisticians use metadata to help data users understand characteristics of data. For survey data, this might include the sample population, the unit of analysis, and the reference period. For a more practical example, when a phone call is made the data can be considered the content of the call itself. The metadata of the call would include the caller, the recipient, the time of the call, and the location of the call.

Metadata are typically divided into the following categories:[129]

- **Descriptive metadata**, which describe a resource for purposes such as discovery and identification, e.g., title, abstract, author, and keywords.
- **Structural metadata**, which indicate how compound objects are put together, e.g., how pages are ordered to form chapters.
- **Administrative metadata**, which provide information to help manage a resource, e.g., the origin of data as well as whether and/or how the data may have been altered. There are several subsets of administrative data; two that are sometimes listed as separate metadata types are *rights management metadata*, which deal with intellectual property rights, and *preservation metadata*, which contain information needed to archive and preserve a resource.

121 Shane Harris, *The Watchers: The Rise of America's Surveillance State* (New York: Penguin Books, 2010).

122 Federation of American Scientists, *Congressional Record: September 24, 2003 (House) H8500-H8550*, 2003, https://fas.org/sgp/congress/2003/tia.html.

123 Mark Williams Pontin, *The Total Information Awareness Project Lives On*, MIT Technology Review, 2006, https://www.technologyreview.com/s/405707/the-total-information-awareness-project-lives-on/.

124 Shane Harris, "TIA Lives On," National Journal, February 23, 2006, https://web.archive.org/web/20110528231531/http://shaneharris.com/magazinestories/tia-lives-on/.

125 Ibid.

126 Jonathan Rae, "Will It Ever Be Possible to Profile the Terrorist?" *Journal of Terrorism Research* 3, no. 2 (2012): DOI: http://doi.org/10.15664/jtr.380.

127 William Press, "Strong Profiling Is Not Mathematically Optimal for Discovering Rare Malfeasors," *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 6 (2008): 1716-1719.

128 Aline Robert, "Big Data Revolutionises Europe's Fight against Terrorism," Euroactiv.fr, June 23, 2016, https://www.euractiv.com/section/digital/news/big-data-revolutionises-europes-fight-against-terrorism/.

129 Jenn Riley, *Understanding Metadata: What Is Metadata, and What Is It For?*, National Information Standards Organization, 2004, http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Box 3.2. The Panama Papers: Tracking Tax Evasion through Analysis of Large Datasets[130]

In early 2016, a network of journalistic outlets began releasing stories collectively known as the Panama Papers. The stories centered on a law firm, Mossack Fonseca, which had facilitated tax avoidance or evasion for many decades. An unknown person with access to the firm's internal communications began leaking this information to a journalist at the *Süddeutsche Zeitung*. At least 2.6 terabytes of data were leaked.

So overwhelmed was the newspaper that received this enormous amount of data that it enlisted the help of the International Consortium of Journalists, who in turn fed the data to over four hundred other journalists. Entirely new kinds of journalistic teams had to be assembled to secure (e.g., encrypt), scan, index, search, store, order, distribute, edit, and share the data across continents. Making sense out of the data required skills in data visualization and graphics.

Some governments are now using these large databases—and the metadata they contain—to connect the dots and crack down on tax evasion. For instance, Denmark recently paid approximately US$1.3 million for a leaked dataset from the Panama Papers containing information on potential Danish tax evaders.[131]

Metadata are data about data. The metadata associated with data contained in large datasets can be analyzed, and potentially used as inputs to visualizations, to provide an analyst or audience with a better understanding of the contents of a large dataset.

Network Analysis

The origins of network analysis lie in the mid-1700s with Swiss mathematician Leonhard Euler, whose work led to graph theory.[132] In essence, graph theory is concerned with nodes (which could be people, devices, organizations, or other entities) and links between those nodes, which, in sum, represent a network.

Once mapped, a network can be analyzed to determine characteristics of specific nodes, e.g., those that have the most direct connections to other nodes (degree centrality, degree distribution), those that are best connected in the network (betweenness centrality), or those that have best access to the network (closeness centrality). The entire network (or "network topology") can be characterized by how efficiently information can be exchanged (efficiency), the density of links between nodes in a network (modularity), and many other attributes.[133]

After many years of theoretical development, network analysis capabilities were greatly enhanced by technological advances surrounding telephony since the 1980s, computing advancements during and since the 1990s, and the emergence of online social networks in the 2000s. These advances provided both the computational capability and data sources necessary to undertake large-scale network analysis.

Box 3.3. Network Analysis and Mapping Out Criminal or Terrorist Organizations

Much has changed since the 1990s, when Harvard University Professor Malcolm Sparrow lamented that "the concepts of network analysis are highly pertinent to many forms of intelligence analysis and are currently being used seldom, if at all."[134] Spurred-on, in particular, by the overhauling of intelligence activities following the attacks on September 11, 2001, network analysis and metadata collection have been increasingly used as tools for mapping out criminal or terrorist networks and organizations, identifying central individuals, and monitoring communications of individuals in these networks.

One publicly available example of network analysis put into practice for such purposes is a 2002 paper by Valdis Krebs entitled "Mapping of Terrorist Cells."[135] Krebs constructed a network graph—based on publicly available information—of those who hijacked flights on September 11, 2001.

Unfortunately, there is limited publicly available information on the workings of terrorism-related work undertaken by government

130 Information primarily taken from Alan Rusbridger, "Panama: The Hidden Trillions," *New York Review of Books*, Issue 27, October 2016.

131 Glyn Moody, "Panama Papers: Denmark to Pay $1.3M Plus for Leaked Data to Probe Tax Evasion," Ars Technica, September 9, 2016, http://arstechnica.com/tech-policy/2016/09/panama-papers-denmark-payout-data-tax-evasion-probe/.

132 Greg Satell, *How the NSA Uses Social Network Analysis to Map Terrorist Networks*, DigitalTonto, June 12, 2013, http://www.digitaltonto.com/2013/how-the-nsa-uses-social-network-analysis-to-map-terrorist-networks/.

133 Linton C. Freeman, *Centrality in Social Networks: Conceptual Clarification*, *Social Networks* 1 (1978/79): 215-239.

134 Malcolm K. Sparrow, "Application of Network Analysis to Criminal Intelligence," *Social Networks* 13, no. 3 (September 1991): 251-274.

135 Valdis Krebs, "Mapping Networks of Terrorist Cells," *Connections* 24, no. 3 (2001): 43-52.

agencies.[136] One instance that is known is the US National Security Agency's bulk-telephony metadata collection program. This program uses network analysis to identify and link suspect individuals based on metadata collected from their call records.[137] Network analysis methods are also used for social media monitoring, which allows analysts to link profiles associated with terrorist-related content to other profiles that have interacted with the original profile.[138]

The use of metadata and network analysis provides a powerful combination for understanding how entities interact and the emergent behavior networks of entities. Social networks have created a new source of data, and associated metadata, which are used by intelligence and law enforcement agencies in their counterterrorism activities.

Data Fusion

Data fusion describes the process by which several datasets are brought together from multiple sources to create a new, singular dataset. The Joint Directors of Laboratories, which pioneered a multi-level data fusion model in the early 1990s, defines data fusion as a "multi-level, multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from several sources."[139]

The advantages of data fusion mainly involve enhancements in data authenticity or availability.[140]

The field of data fusion has developed to address four broad challenges associated with data inputs: data imperfection, data correlation, data inconsistency, and disparateness of data form.[141] Different algorithms are used to address these varying challenges. No single data fusion algorithm is capable of addressing all of them.

Different combinations of these challenges will arise depending on the use case in question due to the various data inputs being used. It is crucial to identify which of these challenges may be present up-front because, if they are not rectified, any error introduced will be magnified in later output.[142]

Predictive Analytics and Machine Learning

Predicative analytics uses statistical techniques to derive a probabilistic score for the likelihood an entity will perform a given action in the future. The analysis is typically based on its current and past profile attributes relative to a comparable population.

In the past, regression techniques have been a mainstay of predictive analytics. Regression involves determining a relationship (correlation) between a dependent variable and an independent variable in a given population. There are many regression models (e.g., linear, logistic, probit) that might be used depending on the phenomenon under examination.

In recent years, machine learning techniques have become increasingly popular for predictive analytics. Machine learning involves the application of induction algorithms, which intake specific instances and produce a model that generalizes beyond those instances.[143] Rather than program a computer to perform a certain task, machine learning involves inputting data into an algorithm that then leads the computer to change its analysis technique.

There are two broad categories of machine learning algorithms: supervised and unsupervised. The former uses labelled records to sort data inputs into known outputs. The latter does not use labelled records so the outputs are not known ex ante. The algorithm explores data, finds some structure, then uses this to determine the outputs. This is particularly useful for use cases like fraud detection or malicious network activity, where the phenomenon to be detected is too rare or its outward characteristics are unknown. Unsupervised learning algorithms are better at searching for anomalies, which signal significant deviation from some sort of "normal."

Machine learning and other more advanced analytical techniques have been deployed for many years to assess consumer credit[144] and detect credit

136 Steve Ressler, "Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research," *Homeland Security Affairs* 2, Article 8 (July 2006), https://www.hsaj.org/articles/171.

137 "Documents on NSA Efforts to Diagram Social Networks of US Citizens," *New York Times*, September 28, 2013, http://www.nytimes.com/interactive/2013/09/29/us/documents-on-nsa-efforts-to-diagram-social-networks-of-us-citizens.html.

138 Ibid.

139 F.E. White, Data Fusion Lexicon, Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego, California, 1991.

140 Bahador Khaleghia, Alaa Khamisa, Fakhreddine O. Karraya, and Saiedeh N. Razavi, "Multisensor Data Fusion: A Review of the State-of-the-Art," *Information Fusion* 14, no. 1 (2013): 28-44.

141 Ibid.

142 With thanks to Daniel Meisner, senior director, Platform, head of Open Data and Ecosystems, Thomson Reuters, for pointing this out.

143 Ron Kohavi and Foster Provost, "Glossary of Terms," *Machine Learning* 30 (1998): 271–274, http://ai.stanford.edu/~ronnyk/glossary.html.

144 Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo, "Consumer Credit Risk Models via Machine-Learning Algorithms," MIT

card fraud.[145] Such practices, previously also used in matchmaking on online dating sites, are now beginning to find applications in such varied areas as graduate recruitment.[146]

Box 3.4. Use of Predictive Analytical Techniques to Improve Policing Outcomes

The field of predictive policing seeks to use advances in data collection and analysis to identify instances of increased crime risk and develop/deploy an associated prevention strategy to mitigate and/or reduce those risks.[147] Varying levels of success for these initiatives have been observed; the extent of success has been linked in part to the specific use case, the phenomena under examination, and the relative operational capabilities and resources of the law enforcement agency in question.

One study[148] used a randomized controlled field trial to evaluate the effectiveness of an Epidemic Type Aftershock Sequence (ETAS) crime forecasting model as compared with the existing best practice used by crime analysts in a district.[149] Trials were held with the Los Angeles Police Department (United States), where analysts traditionally used a COMPSTAT (computer statistics) policing model, and with the Kent Police Department (United Kingdom), where analysts traditionally used an intelligence-led policing approach.

Overall, the study found that ETAS models outperformed analysts' and their traditional techniques. For instance, in the United Kingdom (UK), the analyst predicted 6.8 percent (Maidstone, England) and 4.0 percent (Sevenoaks, England) of crimes successfully compared with 9.8 percent and 6.8 percent, respectively, by the ETAS model. In the United States, the analyst successfully predicted 2.1 percent of crimes compared with 4.7 percent for the ETAS model. Relative to the amount of patrol time allocated to certain hotspots, ETAS-predicted locations were expected to experience 7.4 percent fewer crimes (on a mean of 58.17 crimes per division) per week in the absence of patrol. Analysts' use of traditional methods was expected to yield half the reduction (~3.7 percent) at equivalent patrol levels.

Another study[150] evaluated the effectiveness of the first version of the Chicago Police Department's Strategic Subject List (SSL) predictive policing program. The program's goal was to use social network analysis methods to identify people at risk of gun violence. These people were then to be referred to local police commanders for preventive intervention in the hopes of reducing future crimes linked to gun violence.

The predictive model ended up identifying only 1 percent of the eventual homicide victims (3 out of 405 victims). The program did, however, result in SSL subjects being more likely to be arrested for a shooting.[151] This last finding was thought to indicate that the police used the list as a resource to pursue criminals after the fact, rather than in accordance with the intended purpose: to intervene before crimes took place. The lesson to acknowledge from this case is that the outcomes from using technology, like predictive analysis, will be only as good as the organizational arrangements that allow insights to be acted upon appropriately.

Machine learning techniques have become increasingly popular for predictive analytics. Unsupervised learning algorithms in particular allow for the identification of rare phenomena that may previously have been difficult to identify in large datasets. As with any technology, one key to

Sloan School of Management and Laboratory for Financial Engineering, 2010, https://dspace.mit.edu/openaccess-disseminate/1721.1/66301.

145 Richard J. Bolton and David J. Hand, "Unsupervised Profiling Methods for Fraud Detection," Imperial College, London, via CiteSeerX, 2001, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.5743.

146 Laura Noonan, "Deutsche Uses Koru's 'Dating Site' Tech to Enhance Match with New Recruits," *Financial Times,* September 7, 2016, https://www.ft.com/content/b83108fe-72b4-11e6-bf48-b372cdb1043a.

147 Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, and John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*, Rand Corporation, 2013, http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf.

148 G. O. Mohler, M. B. Short, Sean Malinkowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, and P. J. Brantingham, "Randomized Controlled Field Trials of Predictive Policing," *Journal of the American Statistical Association* (2015): DOI: 10.10.1080/01621459.2015.1077710.

149 ETAS models are analogous to those used for seismic activity. Using an Expectation-Maximization algorithm, as crimes occur in real time, the model adjusts the probabilities of future crime hotspots similar to the way that one might model aftershocks following an earthquake (if one incident occurs in a hotspot, it is more likely that others will follow).

150 Jessica Saunders, Priscilla Hunt, and John S. Hollywood, "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot," *Journal of Experimental Criminology* 12, no. 3 (September 2016): 347–371, DOI 10.1007/s11292-016-9272-0.

151 Ibid.

effectively using predictive analytics is having the appropriate organizational measures in place to act upon the insights gleaned from these techniques.

Blockchain

One of the most interesting and groundbreaking technological innovations of the past decade is the blockchain that underpins bitcoin, a digital currency supported by cryptographic methods (a "cryptocurrency"). One of the key technologies that secures bitcoin is a distributed, publicly available, and immutable ledger commonly referred to as a blockchain.

In very simple terms, a blockchain is a shared database with time-stamped entries. The name is derived from the way in which transactions are grouped together (into a block) and added to the ledger sequentially. Each block is linked to the previous block, thereby making a chain (hence, "blockchain"). That the entries form a chain allows anyone to trace back through the history of transactions to see and confirm what transactions took place between whom and at what time. Three broad types of blockchains have emerged—public, private, and a hybrid of the two.[152] They are differentiated based on their level of centralization/decentralization, their consensus mechanism, and who has read or write ability.

A blockchain is used in bitcoin to prevent the double-spend problem. Before bitcoin, the issue with a digital currency was that someone could spend the same unit of digital currency in multiple places at the same time. A blockchain solves this problem by providing a shared ledger, which ensures that everyone knows and agrees on how much of the digital currency has transacted among users at any point in time.

It is thought that blockchains might provide an effective tool in detecting and preventing corrupt or fraudulent activities. This thinking is premised on the immutability of a blockchain. The immutability prevents any one party from altering past entries, as one might be able to do with paper or digital records.

Box 3.5. Using Blockchain to Address Fraud and Theft

Everledger is a UK-based company that uses public and private blockchains, along with other technologies, to address a novel problem: diamond theft and associated insurance fraud. This problem stems from two factors. First, there previously was not a dependable way to detect if a diamond had been stolen. Moreover, like other luxury goods, proof of ownership remains on paper documents, which are vulnerable to tampering and loss.[153]

Everledger creates a unique, digital "thumbprint" of a diamond, which records its individual set of attributes including color, clarity, cut, and carat weight, as well as forty other metadata points, and links these to the laser inscriptions on the girdle of the stone.[154] It then places this information on the blockchain to create an immutable entry. If stolen, the diamond's original owner can be traced using this entry on the blockchain.

As many organizations that are experimenting with blockchain have found out, there are inherent difficulties using a technology designed to track digital currency transactions for other use cases. Attempting to register physical assets using a digital entry on a blockchain requires a trusted third party. However, bitcoin was designed specifically to remove the need for such a trusted third party through a computationally intensive consensus mechanism.[155] Trust in Everledger therefore becomes paramount, as opposed to bitcoin, where trust is intentionally factored out by design.

Moreover, placing information on any public blockchain—such as the bitcoin blockchain—necessitates making that information publicly available. This might not be appropriate for some sensitive or private information. To overcome this, Everledger uses a private blockchain, with sensitive data such as police reports and policy information kept on the company's Eris-run platform.[156]

152 Vitalik Buterin, "On Public and Private Blockchains," Ethereum Blog, August 7, 2015, https://blog.ethereum.org/2015/08/07/on-public-and-private-blockchains/.

153 Grace Caffyn, "Everledger Brings Blockchain Tech to Fight against Diamond Theft," CoinDesk, August 1, 2015, http://www.coindesk.com/everledger-blockchain-tech-fight-diamond-theft/.

154 "On Blockchain, Diamonds Are Forever," Rakuten Today, October 4, 2016, https://rakuten.today/blog/everledger-blockchain-diamonds-forever.html.

155 Steve Wilson, "Blockchain Plain and Simple," Constellation Research, January 30, 2017, https://www.constellationr.com/blog-news/blockchain-plain-and-simple.

156 Grace Caffyn, "Everledger Brings Blockchain Tech to Fight against Diamond Theft."

The bitcoin blockchain has inspired numerous new projects that all seek to build on the cryptocurrency's original success. However, it must be remembered that the bitcoin blockchain was developed to solve one very specific problem: double-spend. As new projects continue to develop, such as Hyperledger and Ethereum, many new possibilities for applications of distributed/shared ledger technology will emerge.[157]

Shortcomings and Limitations of Data and Digital Technologies

Although the cost of profiling and data fusion are falling due to Moore's Law and other technological advances, there are important economic, statistical, and practical/operational issues that commonly stand in the way of effective deployment of these technologies. As with any tool, use of big data methods will be effective only if those who wield these tools have the requisite knowledge of their applications and shortcomings.

Privacy Considerations

Strict privacy-related laws have been in place for many decades, in the United States and abroad, to constrain the ability of public and private sector organizations to collect and use personal data. In particular, the European Union's General Data Protection Regulation, which will come into effect in 2018, has specific clauses related to practices such as profiling.

As some of the case studies throughout this chapter have illustrated, large-scale data collection and analysis can often fall foul of privacy laws.

Part of the issue is that anonymized data can be de-anonymized when several data sources are combined.[158] Likewise, non-personally identifiable information can become personally identifiable information—which is treated differently legally—when combined with other data.[159]

A privacy assessment is therefore essential to any initiative using large-scale data collection and analysis to avoid infringing upon privacy laws and civil liberties.

False Positives and Negatives

An important limitation of any profiling effort across relatively large populations is the occurrence of false positives and false negatives. A false positive can be thought of as a false alarm. According to New York University's distinguished professor of risk engineering, Nassim Nicholas Taleb, the "tragedy of big data" is that even though one has more data, it also means one has more false information.[160] More false information makes it harder, and costlier, to correctly identify the desired targets. Reducing the incidence of false positives or negatives becomes more costly as one attempts to eliminate such errors from the predictive analysis.

"... [U]se of big data methods will be effective only if those who wield these tools have the requisite knowledge of their applications and shortcomings."

This may not be an issue in cases where incorrectly identifying and acting upon an entity that is a false positive does not result in enormously meaningful repercussions.[161] However, in instances where there are meaningful repercussions from such an error, the benefits of such predictive profiling may be (substantially) outweighed by the costs.

The Unit of Analysis with Dynamic Profiles in Heterogeneous Populations

The first step in profiling is determining what the unit of analysis should be. In other words, "What do we watch—the farmer, the dog, the chickens, or the coop"?[162] The answer to this question may not immediately be obvious. If the correct unit of analysis is not chosen, however, the rest of the profiling exercise—and the output of any subsequent analysis—is moot.

Moreover, profile attributes are dynamic—they are shaped by many inputs over time and as such can shift depending on the changing circumstances. The true rates of bad actors, which are sentient and

157 See "About the Hyperledger Project," Hyperledger, https://www.hyperledger.org/about and Ethereum, https://www.ethereum.org/, for more information.

158 Latanya Sweeney, "K-anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 5, (2002): 557-570.

159 Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* 57 (2010): 1701.

160 Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder* (New York: Random House, 2012).

161 Daniel Geer, *Measuring Security*, Tutorial, 2007, v2.1:16x07, http://geer.tinho.net/measuringsecurity.tutorial.pdf.

162 Ibid.

thus able to adapt, might also change over time. All these elements require data inputs to be continually tracked and updated, which might not be cheap or practical.

Effective Interpretation of Results and Intervention Strategy

While robust and extensive data analysis might be undertaken with cutting-edge predictive analytic methods, this does not imply that the results of such analysis will subsequently be correctly interpreted and acted upon. There are inherent limitations in using these techniques, and not fully understanding them can have consequences. This is particularly the case when attempting to measure or identify a person's emotions or state of mind.[163]

Even in cases where the analysis is correctly interpreted and understood, an effective prevention or intervention strategy must be developed and deployed to mitigate the identified risk(s).[164] However, the history of predictive policing suggests that developing and deploying these strategies is one of the biggest challenges that initiatives using such data analysis techniques face.

Box 3.6. Gouré, Kellan, and RAND's Vietnam Motivation and Morale Project[165]

During the Vietnam War, to understand whether the US-led carpet bombing campaign was reducing the morale of the Vietcong fighters and North Vietnamese citizens, the RAND Corporation extensively interviewed North Vietnamese prisoners and defectors. Starting in 1964, the original leader of the RAND project, Leon Gouré, interpreted from the sixty-one thousand pages of extensive data collection and analysis (the big data of its day) that the bombing campaign was successful (i.e., the Vietcong's morale was falling). One of his colleagues, Konrad Kellan, later reviewed the interviews in 1965. Kellan postulated a different interpretation, concluding that the opposite (and ultimately correct) outcome was occurring, namely, that the bombing campaign only reinforced the morale of the Vietcong and citizens of North Vietnam.[166]

Kellan attributed his key insight, which allowed him to correctly interpret the body of data, to one interview with a senior Vietcong captain:

> He was asked very early in the interview if he thought the Vietcong could win the war, and he said no. But pages later, he was asked if he thought that the US could win the war, and he said no. The second answer profoundly changes the meaning of the first. He didn't think in terms of winning or losing at all, which is a very different proposition. An enemy who is indifferent to the outcome of a battle is the most dangerous enemy of all.[167]

This reality was something that Gouré had overlooked given his own personal history and biases. The lesson here is that while a large body of data might be available, correctly interpreting the data is an entirely different matter. This has not changed in spite of decades of advances in analytical techniques.

Recommendations

A number of lessons on how to successfully deploy digital technologies and data analytics emerge from the various cases covered in this chapter. These lessons form the basis for the recommendations below.

- Invest in people with the skills and knowledge: A broad skill set is required to correctly secure, scan, index, search, store, order, distribute, and edit data as well as visualize/communicate findings from data analysis. Very rarely does any one person possess all of these skills, so multidisciplinary teams must be formed to successfully use digital technologies and data analysis. Organizations should take this into account when considering the adoption and subsequent use of these technologies.
- Ask whether data analysis is appropriate for answering the desired question: Digital technologies and data analysis are relatively better suited to solving some problems, such as optimization, than others, particularly

163 Malcom Gladwell, "Revisionist History: Saigon, 1965, Podcast Episode 2," 2016, based on Gladwell, "Viewpoint: Could One Man Have Shortened the Vietnam War?" BBC.com, 2013, http://www.webcitation.org/6I1RnuJsR.

164 Perry, McInnis, Price, Smith, and Hollywood, *Predictive Policing*; Greg Ridgeway, "Linking Prediction and Prevention," *Criminology and Public Policy* 12, no. 3 (2013): 545-550; Saunders, Hunt, and Hollywood, "Predictions Put into Practice."

165 Gladwell, "Revisionist History: Saigon, 1965, Podcast Episode 2."

166 Gladwell, "Revisionist History: Saigon, 1965, Podcast Episode 2." It is also worth noting that during the Vietnam War, US Secretary of Defense Robert McNamara became blinded to the reality in the field due to his overreliance on data collection and interpretation. In particular, his focus on the body count blinded him to the other—more important—indicators that the war was not winnable. See Kenneth Cukier and Viktor Mayer-Schönberger, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Eamon Dolan/Mariner Books: 2013). The same methods that worked well in reducing the costs of Ford motorcar production ended up disastrous for the conduct of full-scale war in Vietnam—another lesson that applying effective techniques from one use case does not mean success will occur for other use cases.

167 Gladwell, "Revisionist History: Saigon, 1965, Podcast Episode 2."

those involving behavior or emotions. Before embarking on a data analysis exercise to answer a question, organizations first need to consider whether the techniques they intend on using will be able to generate useful answers. This recommendation also applies to blockchains. Organizations need to consider whether an immutable, publicly available database that requires immense computing power to maintain consensus is superior—given the use case—to relatively more simple, long-standing options in the field of distributed databases.

- Place technology use within a larger strategy: Even if data analysis is correctly done and the results are correctly interpreted and then communicated, the exercise becomes moot if there is not robust implementation/operationalization of the results. Organizations need to understand technology use and data analysis not in isolation but as part of a wider organizational strategy.
- When investing in data analysis technologies, consider all available options: Many data analysis technologies and databases or data sources are open source and freely available. However, in some cases, a custom-built "data analysis solution" might be needed to accomplish organizational goals.
- More data do not necessarily equal better data: A common misconception is that collecting and adding more data results in "better" data. The issue is that beyond a certain point, more data tend to create more noise, which results in "worse" data. Organizations need to consider how much data are required to answer the question they have and determine at what point sufficient data have been collected for useful analysis to be undertaken.

Conclusion

Digital technologies and data analysis have advanced greatly over the past two decades. A variety of techniques are now available including profiling, metadata collection, network analysis, data fusion, and predictive analytics. These techniques can be, and increasingly are, used to profile and track bad actors to detect and deter or prevent fraud, money laundering, bribery, terrorism, and regulatory non-compliance. While powerful when used properly, these technologies are most effective when deployed by organizations in which the staff have appropriate skills and a realistic understanding of just what benefits the technologies can provide.