The Columbia

SCIENCE AND TECHNOLOGY LAW REVIEW

www.stlr.org

DATA MINING AND DOMESTIC SECURITY: CONNECTING THE DOTS TO MAKE SENSE OF DATA

By K. A. Taipale*

Abstract

Official U.S. Government policy calls for the research, development, and implementation of advanced information technologies for aggregating and analyzing data, including data mining, in the effort to protect domestic security. Civil libertarians and libertarians alike have decried and opposed these efforts as an unprecedented invasion of privacy and a threat to our freedoms.

This Article examines these technologies in the context of domestic security. The purpose of this Article is not to critique or endorse any particular proposed use of these technologies but, rather, to inform the debate by elucidating the intersection of technology potential and development with legitimate privacy concerns. This Article argues that security with privacy can be achieved by employing value-sensitive technology development strategies that take privacy concerns into account during development, in particular, by building in rule-based processing, selective revelation, and strong credential and audit features. This Article does not argue that these technical features alone can eliminate privacy concerns but, rather, that these features can enable familiar, existing privacy protecting oversight and control mechanisms, procedures and doctrines (or their analogues) to be applied in order to control the use of these new technologies.

[*] B.A., J.D., New York University, M.A., Ed.M., LL.M., Columbia University; Executive Director, Center for Advanced Studies in Science and Technology Policy.

The author would like to thank the editorial board of the Columbia Science and Technology Law Review and Eben Moglen, Daniel Solove, Paul Rosenzweig, Daniel Gallington, Usama Fayyad, and David Jensen, whose insights, comments, or work helped inform this Article. The views and any errors are solely those of the author.

> Further, this Article argues that *not* proceeding with government funded research and development of these technologies will ultimately lead to a diminution in privacy protection as alternative technologies developed without oversight are employed in the future since those technologies may lack the technical features to protect privacy through legal and procedural mechanisms.
>
> Even if it were possible, controlling technology through law alone, for example, by outlawing the use of certain technologies or shutting down any particular research project, is likely to provide little or no security and only brittle privacy protection.

Table of Contents

Knock, knock.
"Who's there?"
"FBI. You're under arrest."
"But I haven't done anything."
"You will if we don't arrest you," replied Agent Smith of the Precrime Squad.[1]

Prelude

On September 11, 2001, nineteen foreign terrorists launched a brazen attack on American soil by hijacking four civilian airliners and crashing them into the twin towers of the World Trade Center, the Pentagon, and a Pennsylvania field, killing more than 3,000 innocent victims.[2] In the aftermath, the U.S. government was chastised for the apparent inability of its security services and law enforcement to "connect the dots" and prevent the attack.[3]

[1] *See Minority Report* (20th Century Fox 2002) ("precogs" predict who will commit murder in the future thus allowing for their preemptive arrest); Charles Piller & Eric Lichtblau, *FBI Plans to Fight Terror With High-Tech Arsenal*, L.A. Times, July 29, 2002, at A1 ("By Sept. 11, 2011, the FBI hopes to use artificial-intelligence software to predict acts of terrorism the way the telepathic precogs in the movie Minority Report foresee murders before they take place.").

[2] In addition to the deaths, the terrorist attack on the World Trade Center towers has been variously estimated to have caused between $50 billion and $100 billion in direct economic loss. Estimates of indirect losses exceed $500 billion nationwide. General Accounting Office U.S. Congress, GAO-02-700R, Review of Studies of the Economic Impact of the September 11, 2001 Terrorist Attacks on the World Trade Center (2002), *available at* http://www.gao.gov/new.items/d02700r.pdf.

[3] *See*, *e.g.*, CNN.com, *Senator: U.S. didn't connect 'dots' before 9/11* (May 15, 2002) ("A key question, [Senator] Graham said, would be 'why these dots weren't seen and connected'"), *at* http://www.cnn.com/2002/US/05/15/inv.fbi.terror/?related; *Administration, agencies failed to connect the dots*, USA Today, May 17, 2002, at 1A, *available at* http://www.usatoday.com/news/washington/2002/05/17/failure-usatcov.htm. In addition to media chastisement, there has been internal criticism of this failure to connect the dots. *See F.B.I. Chief Admits 9/11 Might Have Been Detectable,* N.Y. Times, May 30, 2002, at A1 (At a news conference in May 2002, the FBI Director himself said "But that doesn't mean there weren't red flags out there or dots that should have been connected."); *see also* Joint Inquiry Into the Intelligence Community Activities Before and After the Terrorist Attacks of September 11, 2001 House Permanent Select Comm. on Intelligence & Senate Select Comm. on Intelligence, H. Rep. No. 107-792, S. Rep. No. 107- 351 (2002) [hereinafter Joint Inquiry Report] (refers at least ten times to the intelligence communities failure to "connect the dots"), *available at* http://www.gpoaccess.gov/serialset/creports/911.html; *see*, *e.g.*, *id.* at 62, 335, 337, 340 and Additional Views of the Members of the Joint Inquiry at 6, 33, 45, 67, 106.

"Connect the dots" refers to the child's game in which discrete data points (i.e., numbered dots) are transformed into a single image by drawing links between the dots in the correct order, thus creating information (the picture) from the data (the dots). Connecting the dots has become a metaphor for discovering the "big picture" from seemingly unrelated facts.

In response, the U.S. Department of Justice and the FBI have undertaken to reorganize their mission from the traditional role of investigating and prosecuting crime that has already occurred to that of preventing future acts of terrorism.[4] To help support this reorientation, the government has made the updating of information technology and systems a priority, and information sharing and automated analysis technologies have become part of official government information technology development policy.[5]

In fact, development of these technologies is already mandated by law as the Homeland Security bill signed by President Bush on November 25, 2002 contains provisions that specifically make it the responsibility of the Undersecretary for Information Analysis and Infrastructure Protection at the Department of Homeland Security to "establish and utilize . . . a secure communications and information technology infrastructure, including data mining and other advanced analytical tools, in order to access, receive, and analyze data and information in furtherance of the responsibilities under this section . . . ."[6]

Further, the Congressional Joint Committee Inquiry into the Terrorist Attacks of September 11, 2001 specifically highlights the need for these tools and recommends their development. For example, the Joint Inquiry found:

> The facts surrounding the September 11 attacks demonstrate the importance of strengthening the Intelligence Community’s ability to detect and prevent terrorist attacks in what appears to be the more common, but also far more difficult, scenario [in which there is no single, specific piece of information that would have provided advanced warning of the attacks]. Within the huge volume of intelligence reporting that was available prior to September 11, there were various threads and pieces of information that, at least in retrospect, are both relevant and significant. The degree to which the Community was or was not able to build on that information to

This Article concerns itself with information technologies that can identify and make known useful patterns in data both by connecting known "dots" (that is, by analyzing links from or relationships to a known subject, object or event) and by identifying unknown "dots" and relationships (that is, by discovering or revealing new patterns in data) in order to reveal the existence of higher level things (that is, organizations and activities) based on lower level data (that is, people, places, objects, and transactions).

[4] *See* U.S. Department of Justice, Fact Sheet: Shifting from Prosecution to Prevention, Redesigning the Justice Department to Prevent Future Acts of Terrorism (2002), *available at* http://www.usdoj.gov:80/ag/speeches/2002/fbireorganizationfactsheet.htm.

[5] *See, e.g.,* White House, National Strategy for Homeland Security (2002), *available at* http://www.whitehouse.gov/homeland/book/index.html; Joab Jackson, White House IT Strategy Emphasizes Info Sharing, Wash. Tech., July 16, 2002 (quoting Steve Cooper, Chief Information Officer of the Homeland Security Office, stating "Information sharing and data mining are integral IT components of the White House's . . . national strategy for homeland security"), *available at* http://www.washingtontechnology.com/news/1_1/regulation/18604-1.html.

[6] Pub. L. No. 107-296, § 201(d)(14) (2002). *But cf. infra* note 28 (describing various legislative proposals to restrict or prohibit the development or use of these technologies).

> discern the bigger picture successfully is a critical part of the context for the September 11 attacks and is addressed in the findings that follow.[7]

The report goes on to note:

> At the FBI, information access continues to be frustrated by serious technology shortfalls. The Bureau's Deputy Assistant Director for Counterterrorism Analysis told the Joint Inquiry:
>
> > There were a variety of problems in sharing information, not only with other agencies, but within the Bureau itself. This was and is largely attributable to inadequate information technology. In a nutshell, because the Bureau lacks effective data mining capabilities and analytical tools, it has often been unable to retrieve key information and analyze it in a timely manner - and a lot probably has slipped through the cracks as a result.[8]

Based on these findings, the Joint Inquiry Report specifically recommends the development and implementation of, among other things, information sharing and data mining technologies:

> [T]he Director of National Intelligence shall develop the Intelligence Community component of the strategy, [which] should encompass specific efforts to: . . . improve and expand the use of data mining and other cutting edge analytic tools; and to develop [the] capability to facilitate the timely and complete sharing of relevant intelligence information both within the Intelligence Community and with other appropriate federal, state, and local authorities.[9]
>
> . . . .
>
> Congress and the Administration should insure the full development within the Department of Homeland Security of an effective all-source terrorism information . . . center [that will] implement and fully utilize data mining and other advanced analytical tools, consistent with applicable law.[10]

[7] Joint Inquiry Report, *supra* note 3, at 6.

[8] *Id.* at 341.

[9] *Id.* errata print, at 4.

[10] *Id*. at 5—6.

This Article examines data aggregation and automated analysis technologies, in particular, those technologies popularly referred to as "data mining,"[11] and the related privacy concerns arising in the context of employing these techniques in domestic security.[12]

Introduction

New technologies do not determine human fates; rather, they alter the spectrum of potentialities within which people act.[13] Advanced information technologies for aggregating and analyzing large datasets, including data mining,[14] have already enabled new business opportunities by turning large volumes of corporate data into competitive business opportunities[15] and have expanded scientific investigation possibilities by

[11] In general usage, the term "data mining" is used in two distinct ways – both to refer to the overall process of finding useful patterns in data and to describe a specific step within such processes in which heuristic discovery algorithms are run against the data.

This Article generally follows the popular convention and uses the term "data mining" generically to refer to the overall process of applying automated analysis to data to gain knowledge, that is, to describe the overall process of finding useful information in datasets ("to make sense of data"). In Part II, in which data mining is described in some detail, data mining as overall process is distinguished from data mining as a particular step in the process.

It should be noted that among those with technical expertise in the field, the overall process is more properly referred to as "knowledge discovery" and the term "data mining" is reserved for use in describing the particular step of applying algorithms to data to extract rules for a descriptive or predictive model. *See infra* note 81. Strictly speaking, data mining refers to the development of descriptive or predictive models. The application of developed models to new data – that is, pattern-matching to identify new subjects – is decision-making (or inference). *See infra* notes 98, 99 and accompanying text.

[12] *See infra* note 40 (distinguishing general law enforcement usage of these technologies from domestic security applications).

[13] Robert McClintock & K. A. Taipale, Educating America for the 21st Century, Institute for Learning Technologies 2 (1994), *available at* http://www.taipale.org/ilt/ILTplan.html.

[14] *See supra* note 11; discussion *infra* Part II.

[15] *See*, *e.g.*, Ronald J. Brachman et al., *Mining Business Databases*, 39 Comm. of the Ass'n for Computing Machines 42—48 (1996) ("Ad hoc techniques – no longer adequate for sifting through vast collections of data – are giving way to data mining and knowledge discovery for turning corporate data into competitive business advantage."); Michael J. A. Berry & Gordon Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support (1997); Jill Dyche, e-Data: Turning Data into Information with Data Warehousing (2000) (discussing how data mining and knowledge discovery technologies have been widely adopted for marketing, sales and customer relationship management ("CRM") applications, among others).

Other current business uses include screening for financial investments, detecting fraud in credit card transactions, troubleshooting manufacturing and production processes, and monitoring telecommunications and other complex networks (for example, triggering alarm episodes). *See* Brachman, *supra*, at 45—47. A simple example of data mining techniques that should be familiar to most readers can be experienced at amazon.com, which uses "association rules" to suggest books, CDs and other products that a user might be interested in purchasing on return visits based on correlations between past purchases and purchases by other users. *See infra* text accompanying notes 102—103.

automating data reduction procedures to enable scientists to effectively exploit immense datasets.[16]

Many now argue that these new tools should be employed to detect and prevent terrorism,[17] sometimes with naïve optimism that technology can "solve the terrorism problem."[18] Others, of course, raise the specter of an Orwellian Big Brother[19] amassing "dossiers" on all Americans with attendant loss of civil liberties and freedom if these technologies are permitted to be used for national security or law enforcement purposes.[20] Some argue that technology is neutral and can be contained by legal regulation:

> As a general principle, technology is neutral with regard to privacy. It is the rules governing the use of technology that matter. Privacy advocates and civil libertarians are right to focus attention on the rules under which technology operates, but to dismiss these kinds of technological advances as inherently destructive of privacy is mistaken. Within the proper set of rules, we can protect privacy while using technology to modernize government systems for domestic defense. [21]

[16] *See*, *e.g.*, Usama Fayyad et al., *Mining Scientific Data*, 39 Comm. of the Ass'n for Computing Machines 51 (1996) [herinafter Fayyad et al., Mining]. The data problem in scientific investigation arises from the widening gap between data collection capabilities and the ability to analyze data manually. Data mining techniques allow the exploitation of large datasets through the partial automation of such analysis, particularly through data reduction. Examples of scientific applications include cataloging sky objects, finding volcanoes on Venus, biosequencing DNA databases, detecting tectonic activity from satellite data, and predicting weather phenomena. *See id.* at 51—57.

[17] *See*, *e.g.*, Shane Ham & Robert D. Atkinson, Progressive Policy Institute, Using Technology to Detect and Prevent Terrorism (2002), *available at* http://www.ppionline.org/douments/IT_terrorism.pdf; Markle Foundation, Protecting America's Freedom in the Information Age: A Report of the Markle Foundation Task Force (2002), *available at* http://www.markletaskforce.org/ [hereinafter Markle Report]; *see also supra* text accompanying notes 7—10 (citing the recommendations of the Congressional Joint Inquiry Report).

[18] *Contra* Ham, supra note 17, at 1 ("The purpose of this issues brief is not to provide a comprehensive blueprint of how technology alone can solve the terrorism problem – it can't."). *See also* Markle Report, *supra* note 17, at 2 ("Technology is not a panacea.").

[19] *See*, *e.g.*, Susan Baer, *Broader U.S. Spy Initiative Debated*, Balt. Sun, Jan. 5, 2003, at 1A ("ominous and Orwellian, conjuring up visions of Big Brother"); Press Release, ACLU, Big Brother is No Longer a Fiction, ACLU warns in New Report (Jan. 15, 2003), *available at* http://www.aclu.org/Privacy/Privacy.cfm?ID=11612.

[20] *See*, *e.g.*, William Safire, *You Are a Suspect*, N.Y. Times, Nov. 14, 2002 ("[TIA] has been given a $200 million budget to create computer dossiers on 300 million Americans."), *available at* http://www.nytimes.com/2002/11/14/opinion/14SAFI.html. Safire has been credited with triggering the "anti-TIA" stampede. *See* Stuart Taylor, Jr., *Big Brother and Another Overblown Privacy Scare*, Atlantic Online (Dec. 10, 2002) ("hyperventilated William Safire . . . in a . . . column that helped touch off a frenzy of similar stuff"), *at* http://www.theatlantic.com/politics/nj/taylor2002-12-10.htm.

[21] Ham, *supra* note 17, at 11.

Others see the technology itself as a powerful "monster" that once unleashed might not be controlled.[22] Still others dismiss privacy concerns as premature given that the technology is immature and that current attempts are experimental or "years away from final implementation."[23]

As always, the true potentials lie somewhere in between the extremes. However, like with many debates about developments in technology, this one also suffers from a lack of technical understanding on both sides that leads to the issue being presented as a false dichotomy – a choice between security *or* privacy.

It is the thesis of this Article that although information aggregation and analysis technologies, including specifically data mining, do raise legitimate and compelling privacy concerns,[24] these concerns can be significantly mitigated by incorporating privacy values in the technology development and design process itself.[25] Thus, by building in technical features that support privacy protecting implementation policies (as well as existing laws, doctrines, and due process procedures) these technologies can be developed and employed in a way that leads to increased security while protecting privacy interests.[26]

Indeed, it is an underlying assumption of this Article that *not* proceeding with government funded research and development of these technologies (in which political oversight can incorporate privacy protecting features into the design of the technologies)

[22] *See, e.g.,* Jay Stanley & Barry Steinhardt, ACLU, *Bigger Monster, Weaker Chains: The Growth of an American Surveillance Society* (2003), http://www.aclu.org/Privacy/Privacy.cfm?ID=11573&c=39.

[23] *See*, *e.g.*, Hiawatha Bray, *Mining Data to Fight Terror Stirs Privacy Fears*, Boston Globe, Apr. 4, 2003, at C2 (quoting Heather MacDonald of the Manhattan Institute), *available at* http://www.boston.com/dailyglobe2/093/business/Mining_data_to_fight_terror_stirs_privacy_fears+.shtml.

[24] *See* discussion *infra* Part III.

[25] *See* Ben Shneiderman & Anne Rose, *Social Impact Statements: Engaging Public Participation in Information Technology Design*, in Batya Friedman, Human Values and the Design of Computer Technology ("Constructive criticism and guidelines for design could help protect us against the adverse ramifications of technology such as . . . dissatisfaction with privacy protection." *Id*. at 118); *see also*, Julie E. Cohen, *Symposium: The Law and Technology of Digital Rights Management: DRM and Privacy*, 18 Berkeley Tech. L.J. 575, 609—617 (2003) (arguing for building privacy protection into DRM code (in addition to law) by employing value sensitive design and development strategies. "[B]oth judicial and regulatory sanctions are second-best strategies for ensuring effective [privacy] protection for all users. A far more effective method of ensuring that information users actually enjoy the privacy to which they are entitled would entail building privacy into the design of DRM technologies in the first instance."). *See generally* Batya Friedman et al., *Value Sensitive Design: Theory and Methods* (Draft of June 2003), *at* *http://www.ischool.washington.edu/vsd/vsd-theory-methods-draft-june2003.pdf* ("Value Sensitive Design is a theoretical grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.").

[26] For a detailed discussion of the legal processes and mechanisms that could be applied in implementing these technologies while protecting privacy and civil liberties and how these mechanisms interact with technological design, see Paul Rosenzweig, Heritage Foundation, Proposal for Implementing the Terrorism Information Awareness System (2003), *at* http://www.heritage.org/Research/homelanddefense/lm8.cfm (setting out a proposed legal and procedural framework for implementation that is designed to exploit built-in technical characteristics like those described in Part IV of this Article).

will ultimately lead to a diminution in privacy protection as alternative technologies developed without oversight (either through classified government programs or proprietary commercial development) are employed in the future, since those technologies may lack the technical features required to support legal and procedural mechanisms to protect privacy and civil liberties.[27] Thus, this Article draws a distinction between laudable legislative efforts to provide for oversight of these programs and ill-conceived efforts to kill funding for particular research programs or outlaw specific technologies.[28] Consequently, it is my view that the recent defunding of DARPA's

[27] In addition, technologies developed without broad oversight may not be effective or may contain technical flaws. "If research programs are either classified or proprietary, then the resulting algorithms will not get wide review within the technical community. The history of data mining is that early algorithms often contain serious technical flaws that are only revealed after years of testing and analysis within the community. . . . Without [such] review, flawed algorithms are far more likely to make their way deep into the technical computing infrastructure of the U.S. intelligence community." E-mail from David Jensen, Research Professor of Computer Science and Director of the Knowledge Discovery Laboratory, University of Massachusetts (Sept. 8, 2003) (on file with the author).

An additional benefit of developing technologies that incorporate privacy protecting features through government research and development projects is that they will provide opportunities to improve privacy protection more generally throughout society as privacy protecting procedures based on such features can then be voluntarily or legislatively adopted in the private sector.

[28] A full and open public debate and Congressional oversight of government research and development programs is wholly appropriate and necessary to insure ultimate public acceptance of the use of these technologies for domestic security purposes. Further, specific Congressional authorization prior to implementation in any particular agency and for any specific use should also be considered. *See infra* note 58 and accompanying text. However, the public debate and such Congressional oversight and authorization ought to be based on a sound understanding of the technologies and their potential impact on privacy. *See* Rosenzweig, *supra* note 26, at 9; *see also* Daniel J. Gallington, Rational Steps in the Information Technology, National Security and Privacy Debate (Potomac Inst. for Policy Studies, Waypoint Issue Paper, 2003) ("We agree that there should be a broad based rigorous public debate on the intended, unintended, and perhaps unnecessary tradeoffs between national security and privacy, and that the debate should be responsibly informed – technologically and constitutionally."), *available at* http://www.potomacinstitute.org/pubs/030213-waypoint.pdf.

Unfortunately, to date, Congressional action does not seem to reflect an informed understanding of the technologies, nor a very sophisticated understanding of the privacy concerns involved. For an analysis of the implicated privacy interests, *see infra* Part III. Thus, simplistic legislative initiatives have sought to outlaw particular technologies, such as "data mining," or techniques, such as the use of "hypotheticals," rather than attempt the more difficult task of reconciling how these technologies might be employed to provide both privacy and security.

For example, The Data-Mining Moratorium Act of 2003, S. 188, 108th Cong. (2003), *available at* http://thomas.loc.gov/cgi-bin/query/z?c108:S.188:, does not even define "data-mining." The Executive Committee of the Special Interest Group on Knowledge Discovery and Data Mining of the ACM believes that these legislative efforts "do not reflect a sound understanding of data mining science, technology or applications." Executive Committee SIGKDD of the ACM, Data Mining is NOT Against Civil Liberties (June 30, rev'd July 28, 2003) [hereinafter SIGKDD of the ACM], *available at* http://www.acm.org/sigkdd/civil-liberties.pdf.

And, the Citizens' Protection in Federal Database Act of 2003, announced by Senator Wyden on July 29, 2003, seeks to prohibit the "search or other analysis for national security, intelligence, or law enforcement purposes of a database based solely on a hypothetical scenario or hypothetical supposition of who may commit a crime or pose a threat to national security." S. 1484, 108th Cong. §4(a) (2003) [hereinafter CPFDA], *available at* http://thomas.loc.gov/cgi-bin/query/D?c108:1:./temp/~c108rlm0lN::. Obviously, this kind of broad prohibition of a commonly used investigative technique would prevent the search of any database using even traditional modus operandi or psychological or behavioral profiling techniques. S*ee infra* note 66 and accompanying text. It would also prohibit any methods based on human developed hypotheses, as well as existing uses of automated pattern analysis to detect money laundering and insider stock trading or to select income tax returns for audit. *See supra* note 41.

Additionally, legislative efforts to restrict or eliminate funding for specific implementations or particular research programs seem either ineffective or short-sighted. For example, on October 1, 2003, President Bush signed the Department of Homeland Security Appropriations Act, 2004, which includes language purporting to prohibit the use of funds for the CAPPS II program (*see* discussion *infra* Part II for a detailed discussion of CAPPS II) until the General Accounting Office ("GAO") has reported to the Congress that the program meets certain criteria specified in the bill. H.R. 2555, 108th Cong. § 519 (2003). However, in a separate "Signing Statement," President Bush has already declared that this provision is ineffective under INS v. Chadha, 462 U.S. 919 (1983):

> To the extent that section 519 of the Act purports to allow an agent of the legislative branch to prevent implementation of the law unless the legislative agent [GAO] reports to the Congress that the executive branch has met certain conditions, the executive branch shall construe such section as advisory, in accordance with the Chadha principles.

Statement on Signing the Department of Homeland Security Appropriations Act, 2004, Pub. Papers (Oct. 1, 2003), *available at* http://www.whitehouse.gov/news/releases/2003/10/20031001-9.html.

Also on October 1, 2003, President Bush signed the Department of Defense Appropriations Act, 2004, which included language prohibiting the use of funds for the Terrorism Information Awareness ("TIA") (*see* discussion *infra* Part II for a detailed discussion of TIA) or any successor program:

> SEC. 8131. (a) Notwithstanding any other provision of law, none of the funds appropriated or otherwise made available in this or any other Act may be obligated for the Terrorism Information Awareness Program: *Provided,* That this limitation shall not apply to the program hereby authorized for Processing, analysis, and collaboration tools for counterterrorism foreign intelligence, as described in the Classified Annex accompanying the Department of Defense Appropriations Act, 2004, for which funds are expressly provided in the National Foreign Intelligence Program for counterterrorism foreign intelligence purposes.
>
> (b) None of the funds provided for Processing, analysis, and collaboration tools for counterterrorism foreign intelligence shall be available for deployment or implementation except for:
>
> (1) lawful military operations of the United States conducted outside the United States; or
>
> (2) lawful foreign intelligence activities conducted wholly overseas, or wholly against non-United States citizens.
>
> (c) In this section, the term `Terrorism Information Awareness Program' means the program known either as Terrorism Information Awareness or Total Information Awareness, or any successor program, funded by the Defense Advanced Research Projects Agency, or any other Department or element of the Federal Government, including the individual components of such Program developed by the Defense Advanced Research Projects Agency.

Pub. L. No. 108-87, § 8131, 117 Stat. 1054, 1102 (2003). Further, the Joint Explanatory Statement included in the conference committee report specifically directed that the Information Awareness Office (the DARPA program manager for TIA, see discussion in Part II infra) be terminated immediately. "The conferees are concerned about the activities of the Information Awareness Office and direct that the Office be terminated immediately." 149 Cong. Rec. H8755—H8771 (Sept. 24, 2003).

However, President Bush declared in another "Signing Statement" that the classified annex referred to in § 8131(a) would not be considered part of the signed act, therefore anything mentioned in the annex would not be subject to the data mining restriction. Statement on Signing the Department of Defense Appropriations Act, 2004, Pub. Papers (Oct. 6, 2003), *available at* http://www.whitehouse.gov/news/releases/2003/10/20031001-2.html. In addition, the Intelligence

Authorization Act for Fiscal Year 2004, as cleared by Congress on November 21, 2003, explicitly approves data mining for foreign intelligence and requires that the attorney general and director of central intelligence report publicly on the privacy implications of data mining within one year. H.R. 2417, 108th Cong. (2003).

In any case, the defunding of TIA and the shut down of IAO has not resolved these issues. *See infra* note 43 and text accompanying notes 187—194. Defunding TIA has merely eliminated the most visible and focused opportunity around which open public debate could have developed appropriate policy for the use of these technologies, including the development of technical features to mitigate privacy concerns. *Id*. Note also that the Department of Defense Technology and Privacy Advisory Committee (more information can be found at http://www.sainc.com/tapac/) originally commissioned to examine the privacy implications of TIA has been asked by the Secretary of Defense to continue its activities and deliver its report despite the defunding of TIA and IAO. *See infra* note 238.

Another issue raised by these early legislative forays is the effort to limit the application of these technologies to activities conducted against "non-United States citizens," following an historic and traditional distinction between how information relating to U.S. and non-U.S. persons is treated. *See* Department of Defense Appropriations Act, 2004 § 8131(b)(2); *see also* Consolidated Appropriations Resolution, 2003, Pub. L. No. 108-7, § 111(c)(2)(B), 117 Stat. 11, 536 (limiting deployment to "activities conducted wholly against non-United States persons").

These arbitrary distinctions (arbitrary in a technical sense, that is, based on a legal categorization not on characteristics of the data itself) about how information is to be treated based on whom it relates to or where it is collected are increasingly difficult to sustain in the context of international terrorism in which potentially relevant information exists within data sets of otherwise innocuous data, and transactional data relating to U.S. and non-U.S. persons is commingled. *See* Markle Foundation, Creating a Trusted Information Network for Homeland Security: Second Report of the Markle Foundation Task Force 18 (Dec. 2003) [hereinafter, Second Markle Report] (suggesting that new rules are required to replace the "previous 'line at the border' that largely defined the distinctive rules for foreign and domestic intelligence").

Nevertheless, some commentators have suggested that, based on "thirty years of experience in dealing with 'U.S. Person' information" in the context of foreign signal intelligence ("SIGINT") gathering, existing SIGINT oversight regimes and policies may be applicable to the use of data aggregation and automated analysis technologies in the domestic security context. Securing Freedom and the Nation: Collecting Intelligence Under the Law Before the House Permanent Select Committee on Intelligence (Apr. 9, 2003) (testimony of Daniel Gallington, Senior Fellow, Potomac Institute for Policy Studies). U.S. Person data in foreign SIGINT is currently handled by exceptional procedures ("minimization") under NSA/CSS United States Signal Intelligence Directive 18 ("USSID 18") (July 27, 1993), *available at* http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB23/07-01.htm. However, others have suggested that the changed nature of the data subject to analysis – from non-U.S.-Person-centric data in traditional foreign SIGINT to commingled or undifferentiated in domestic security or commercial databases – requires the adoption of a new rule rather than trying to apply procedures developed to manage exceptions. *See, e.g.,* K. A. Taipale, Technology, Security and Privacy, Presentation at The Potomac Institute for Policy Studies slide 26 (Dec. 2, 2003), *available at* http://www.taipale.org/presentations/PIPS-TSP-120203.htm [hereinafter Taipale, Privacy].

A further refinement to address this issue has been proposed that would create a new category of information, "Terrorist Threat Information," in which relevant U.S. and non-U.S. person information would be commingled but where U.S. person information would be protected using selective revelation strategies. *See* discussion *infra* note 62 and accompanying text; *see also*, Daniel Gallington, Better Information Sharing and More Privacy in the War on Terrorism – A New Category of Information is Needed (Potomac Inst. for Policy Studies, Waypoint Issue Papers, July 29, 2003), *available at* http://www.potomacinstitute.org/research/072903-project_guardian.cfm; Potomac Institute for Policy Studies, Oversight of Terrorist Threat Information: A Proposal (June 25, 2003), *available at* http://www.potomacinstitute.org/pubs/Guardian_Proposal%20_0703.pdf. *Cf.* Homeland Security Presidential Directive/HSPD-6, 39 Weekly Comp. Pres. Doc. 1234—1235 (Sept. 16, 2003) (outlining procedures for integrating information about individuals who are known or suspected terrorists within the Terrorist Threat Integration Center ("TTIC"), the all source intelligence fusion and analysis center announced by the President in January 2003. *See* White House Fact Sheet, "Strengthening Intelligence to

Information Awareness Office ("IAO") and its Terrorism Information Awareness program and related projects will turn out to be a pyrrhic 'victory' for civil liberties as this program provided a focused opportunity around which to publicly debate the rules and procedures for the future use of these technologies and, importantly, to oversee the development of the appropriate technical features required to support any concurred upon implementation or oversight policies to protect privacy.[29]

This Article is premised on a belief that we face one of two inevitable futures – one in which technologies are developed with privacy protecting values and functions built into the design or one in which we rely solely on legal mechanisms and sanctions to control the use of technologies that have been developed without regard to such protections. To the extent that "code is law," that is, to the extent that the features and technical constraints built into the technology itself enable or constrain certain behaviors,[30] (including, in this case, the ability to protect privacy or constrain government

Protect America," Washington, DC: The White House, *available at* http://www.whitehouse.gov/news/releases/2003/01/20030128-12.html. TTIC's role and responsibilities are set out in the classified Director of Central Intelligence Directive (DCID) 2/4 (effective May 1, 2003)). *See also* Daniel Gallington, The New Presidential Directive on "Screening" Terrorist Information (Potomac Inst. for Policy Studies, Waypoint Issue Paper, Oct. 6, 2003). *See also* Second Markle Report, *supra*, at 19 (suggesting that HSPD-6 and TTIC may have "radically changed the balance of liberties" without "significant public debate on this fundamental question [i.e., maintaining the U.S. person distinction]").

[29] For a more detailed description of the IAO and its TIA and related projects, see *infra* Part II. For a discussion of the defunding of TIA and the shutting down of the IAO, see *supra* note 28 and *infra* the text accompanying notes 191—198. On the potential negative consequence of such defunding, see *infra* note 197 and accompanying text. For a discussion of the technical features required to support policy, see *infra* Part IV.

[30] That "code is law" has already become cliché. *See* Lawrence Lessig, Code and Other Laws of Cyberspace (1999) ("[Code] constitute[s] a set of constraints on how you behave. . . . The code or . . . architecture . . . constrain[s] some behavior by making other behavior possible, or impossible."). Lessig actually postulates that behavior is controlled (regulated or constrained) through a dynamic interaction of legal rules, social norms, market forces and architecture (or code). *Id*. at 83—99.

See William J. Mitchell, City of Bits: Space, Place and the Infobahn 111 (1995) ("Out there on the electronic frontier, code is the law."); *see also*, Joel R. Reidenberg, Lex Informatica: The Formulation of Information Policy Rules Through Technology, 76 Tex. L. Rev. 553, 554—555 (1998); James Boyle, Foucault in Cyberspace: Surveillance, Sovereignty, and Hardwired Censors, 66 U. Cin. L. Rev. 177, 191 (1997); Lawrence Lessig, Reading the Constitution in Cyberspace, 45 Emory L. J. 869, 896—897 (1996).

To those involved or who have followed the development of information technology over the past two or three decades, the notion that code is law seems banal. *See* Marc Rotenberg, Fair Information Practices and the Architecture of Privacy: (What Larry Doesn't Get), 2001 Stan. Tech. L. Rev. 1, P6 (2001).

> Even before the recent protests over architectures of surveillance, many others have observed the relationship between design and methods of social control. *See generally* Jeremy Bentham, Panopticon (1971); Jacques Ellul, The Technological Society (1964); David Burnham, The Rise of the Computer State (1983); Michel Foucault, Discipline and Punish: The Birth of the Prison (1995); Oscar H. Gandy, The Panopticon Sort: A Political Economy of Personal Information (1993); Gary T. Marx, Undercover: Police Surveillance in America (1988).

Rotenberg, *supra,* at *n.6.*

See also, Neal Kumar Katyal, Architecture as Crime Control, 111 Yale L.J. 1039, 1047 (2002) (discussing the use of physical architecture – structural and space design – as an effective alternative form from extra-legal use[31]) then simply relying on regulatory legal sanctions to control use of whatever technologies may become available in the future seems a second-best strategy that provides little or no security and brittle privacy protection.[32] A more effective strategy for insuring the protection of privacy and civil liberties interests is to build features that support those values into the technologies in the first place. And, it is only through involvement in and oversight of government sponsored research projects that public interest concerns can be incorporated into the development process.

Although technologies do not themselves determine human fates, their design does constrain opportunities, and their development is rarely value-neutral as technological systems are themselves social constructions[33] and therefore reflect social values and interests during the development process. Thus, as argued in Part IV infra, attention to privacy concerns at the design and development stage can produce data aggregation and analysis technologies that build in privacy protecting features that provide intervention and control points for existing legal process and mechanisms to function. This Article does not argue, however, that technical features alone can eliminate privacy concerns, but rather that "code" developed without features to support privacy protecting implementations will not be constrained by "law" alone.[34]

Even if it were desired, it is unrealistic to believe that the development or application of these technologies can be prevented easily through legislation since there is

of crime control; design mechanisms discussed include: (1) creating opportunities for surveillance, (2) instilling a sense of territoriality, (3) building community and avoiding isolation, and (4) protecting targets); Neal Kumar Katyal, Digital Architecture as Crime Control, 112 Yale L.J. 2261 (2003) (applying these four principles of realspace architecture design to the problem of security in cyberspace). That "code is law"—that is, that technology has regulatory impact—should not be confused with the notion "that software code and legal code are somehow regulatory substitutes." *See* R. Polk Wagner, The Case Against Software 3, draft of Aug. 1, 2003 (developing an analytic framework for the evaluation of regulatory policy in cyberspace based on the premise that code is complementary to, not a substitute for, law), *available at* http://www.law.upenn.edu/polk/wagner.against_software.pdf.

[31] This Article argues that automated analysis technologies should be designed to enable existing privacy protecting legal mechanisms, procedures, and doctrines (i.e., "law") to function (or adapt), not that built-in technical features (i.e., "code") alone can eliminate privacy or civil liberties concerns.

[32] Privacy protection is "brittle" in an engineering sense, meaning that any breach will result in catastrophic failure. If technologies are developed without privacy protecting features built in but outlawed for law enforcement or domestic security purposes and then the laws are changed in the future, for example, in response to a new terrorist attack, the then-existing technologies will not be capable of supporting implementation policies that provide any privacy protection.

Contra Ham & Atkinson, *supra* note 17 (arguing that "technology is neutral with regard to privacy" and can be controlled "within the proper set of rules").

[33] *See generally* Wiebe E. Bijker, Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change (1997); Trevor J. Pinch & Wiebe E. Bijker, *The Social Construction of Facts and Artifacts*, in The Social Construction of Technological Systems (Wiebe E. Bijker et al. eds., 1994) (technological development as social construction); Shaping Technology/Building Society (Wiebe E. Bijker & John Law eds., 1992).

[34] *Cf.* Lessig, "Code and the Law of Cyberspace," *supra* note 30, at 6 ("We can build, or architect, or code cyberspace to protect values that we believe are fundamental, or we can build, or architect, or code cyberspace to allow those values to disappear. There is no middle ground.").

a strong commercial imperative for their continued development and use in the private sector [35] and a strong political (as well as practical) imperative for their eventual application for domestic security.[36]

The practical reasons driving development are the same in both the private and public sector – "vast data volume, fewer analytic resources."[37] The practical need for developing data mining techniques is a direct result of the growth in data volumes. Traditional database analysis relies on specific queries formulated by individual database analysts familiar with the particular data and database structure. This manual analysis is slow, expensive and highly subjective,[38] and no longer able to manage the size and dimensionality of current data collection methods. Databases are increasing in two ways: (1) size, that is, the number of records or objects in the database and (2) dimensionality, that is, the number of fields or attributes to an object. As databases grow manual data analysis becomes impractical. Thus, the need to scale up human analytic capabilities through computational automation is driven by a practical (and unrelenting) imperative.[39]

The notion that powerful analytical tools developed for commercial and scientific application will not eventually be used for terrorism prevention (or, for that matter, general law enforcement purposes[40]) seems unrealistic, particularly since these

[35] *See*, *e.g.*, Berry & Linoff, *supra* note 15; Dyche, *supra* note 15.

[36] *See supra* Prelude (discussing the development and use of these technologies as official government policy); *infra* notes 41—42 (pointing out that these technologies are already being widely adopted for such uses).

[37] *See* David Jensen, Data Mining in Networks, Presentation to the Roundtable on Social and Behavior Sciences and Terrorism of the National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice, slide 16 (Dec. 1, 2002), *available at* http://kdl.cs.umass.edu/people/jensen/papers/nrcdbsse02.html.

[38] From a civil liberties perspective it must be noted that traditional human directed analysis is prone to human error and bias in decision making. *See infra* note 118 (pointing out that in certain contexts automated actuarial methods are superior to human judgment).

[39] *See* Usama Fayyad et al., *From Data Mining to Knowledge Discovery in Databases*, 17 AI Magazine 37, 38 (Fall 1996) [hereinafter Fayyad et al., Databases]; Jensen, *supra* note 37 ("A frequent theme in assessments of the technical capabilities of the U.S. intelligence community is how the volume of available data is increasing much faster than the analytical resources to analyze data.").

> Currently one of [intelligence agencies'] significant problems is managing a flood of data that may be relevant to their efforts to track suspected terrorists and their activities. . . . There are well-known examples in which planned terrorist activities went undetected despite the fact that evidence was available to spot it – the relevant evidence was just one needle in a huge haystack.

Committee on Science and Technology for Countering Terrorism, National Research Council, Making the Nation Safer: The Role of Science and Technology in Countering Terrorism (2002), *available at* http://www.nap.edu/html/stct/index.html.

See also John M. Poindexter, Finding the Face of Terror in Data, N.Y. Times, Sept. 10, 2003, at A25 ("The amount of data available to the federal government far exceeds the human capacity to analyze it.").

[40] Obviously, there are additional considerations involved in thinking about how such tools might be employed in law enforcement more generally. These concerns generally have to do with the circumstances

technologies are already being used in a wide variety of law enforcement contexts. First, generic data mining tools are available for (or are built into) all major commercial database applications. As government agencies upgrade their database applications, these tools are becoming widely available regardless of whether government research into domain specific (i.e., law enforcement or domestic security) applications is hindered or not.[41] Second, the private sector is developing domain specific technologies (that is, applications developed specifically for law enforcement purposes) to aggregate and mine data using both link analysis and pattern-matching[42] in criminal investigations and these technologies are already being adopted and employed in a variety of law enforcement environments.[43]

under which raw data is collected, accessed and analyzed, i.e., what information databases should be made available to whom, under what conditions or legal constraints, and for what law enforcement purposes. To the extent that there is a relationship between law enforcement strategies and privacy concerns, the lesser the crime the greater the hurdle for any new technology or wider use that implicates privacy concern.

However, this Article is primarily concerned with the meta-issues involved in applying these data aggregation and analysis techniques to find actors who are hidden among the general population and who have the potential for creating harms of such magnitudes that a consensus of society requires that government adopt a preventative rather than investigative approach. *See generally* Editorial, *The Limits of Hindsight*, Wall St. J., July 28, 2003, at A10 (responding to the release of the Joint Inquiry Report, *supra* note 3, "But even more important is recognizing that terrorism cannot be treated as a law enforcement issue, in which we wait until the bad guys actually pull the trigger before we stop them.").

It is beyond the scope of this Article to address precisely where that line is to be drawn by delimiting particular types of crimes that meet that criteria, or by specifying which government organs or agencies should be permitted particular uses. This Article makes a general assumption that there is a category of criminal – terrorist – for which aggressive preventative law enforcement strategies are appropriate and tries to identify how and where data aggregation and analysis technology intersect with other policy considerations, particularly privacy.

Compare Rosenzweig, *supra* note 26, at 2 (calling for an absolute statutory prohibition on the use of certain applications of these technologies for non-terrorism investigations) with the developments detailed in notes 41 and 43, *infra* (adoption and application of link analysis and pattern-matching for general law enforcement purposes).

[41] For example, Oracle Corp.'s latest version of the Oracle Database 10g includes "embedded data mining capabilities to help decision makers quickly sift through massive amounts of corporate data to uncover hidden patterns, predict potential outcomes, and minimize business risks." Oracle, *Oracle 10g Data Mining*, *at* http://www.oracle.com/ip/index.html?dm_home.html. These tools are currently being used by government agencies in the normal course of business, for example, the Internal Revenue Service is using Oracle's built in data mining tools to improve their audit selection process. *Id.*

Another example is SPSS Inc.'s popular "Clementine" data mining framework, which is also being used to build predictive models for a variety of purposes. *See SPSS, Clementine Data Mining* (to "search for non-compliant tax payers" and focus security resources on "likely threats"), *at* http://www.spss.com/spssbi/clementine/ (last visited Jan. 21, 2003); Dennis Callaghan, *Analytic Detective Work*, eWeek, Sept. 1, 2003 (for general law enforcement purposes: "The Richmond (Va.) Police Department has deployed Chicago-based SPSS' Clementine data mining software to help snuff out crime before it happens, prevent property crimes from escalating into more violent crimes, and even gain insight into how the drug trade operates.").

[42] *See infra* Part II.

[43] *See*, *e.g.*, Jim Goldman, Google for Cops: Revolutionary software helps cops bust criminals (TechTV broadcast Apr. 12, 2003, modified Apr. 17, 2003), *available at* http://www.techtv.com/news/scitech/story/0,24195,3424108,00.html (describing the use of CopLink, a commercial product that allows police departments to link their databases together and search them

simultaneously using artificial intelligence analytics); *see also* Gareth Cook, Software Helps Police Draw Crime Links, Boston Globe, July 17, 2003, at A1 ("The Boston Police Department is rolling out a powerful new computer program built to find hidden connections among people and events almost instantly, allowing detectives to investigate murders, rapes, and other crimes far faster than they can today").

Likewise, law enforcement authorities in the District of Columbia, Virginia, Maryland, Pennsylvania and New York have announced a major initiative to share data across jurisdictions using "powerful tools to quickly detect links among people and events." Spencer S. Hsu, Crossing Lines to Fight Terrorism, Wash. Post, Aug. 6, 2003, at B2, *available at* http://www.washingtonpost.com/wp-dyn/articles/A21710-2003Aug5.html.

New York and Pennsylvania, together with Alabama, Connecticut, Florida, Georgia, Kentucky, Louisiana, Michigan, Oregon, South Carolina, Ohio and Utah have also announced that they would participate in the Multistate Antiterrorism Regional Information Exchange System, known as MATRIX, which aggregates data from both commercial and government sources in a private database and uses analytical algorithms to find links among data items. The U.S. Justice Department recently provided $4 million and the Department of Homeland Security another $8 million to expand the MATRIX program nationally. "The system, developed for Florida by [a private contractor], combines information about persons and property from commercial databases with information from criminal records databases to identify potential terrorists by using a sophisticated algorithm, said Jim Burch, acting deputy director for policy at the U.S. Bureau of Justice Affairs." William Welch, Matrix Taps Databases, 18 Wash. Tech. No. 11 (Sept. 1, 2003). Several states have since quit the program, some due to financial constraints, and others because of privacy concerns. *See*, *e.g.*, John Murawski, *States bow out of anti-crime database in Boca*, Palm Beach Post, Oct. 4, 2003, at 1C. However, seven of the original participating states remain in the pilot program.

The ACLU has recently questioned whether the MATRIX program is an attempt by the federal government to replace "an unpopular Big Brother initiative [i.e., TIA] with a lot of Little Brothers." *See* Press Release, ACLU, What is The Matrix? ACLU Seeks Answers on New State Run Surveillance Program (Oct. 30, 2003), *available at* http://www.aclu.org/Privacy/Privacy.cfm?ID=14257&c=130; ACLU, The MATRIX: Total Information Awareness Reloaded, Data Mining Moves into the States, an ACLU Issue Brief (last accessed Oct. 29, 2003), *at* http://www.aclu.org/Files/OpenFile.cfm?id=14253.

In another project, IBM is developing a data mining system, known as Matchbox, for the federal government that provides for "the sharing of private and confidential information using secure hardware." Murawski, *supra*.

"Pattern recognition" applications (that is, searches using models or patterns developed through data mining) are also already in practical use in law enforcement. *See* discussion *infra* Part III. For example, an agency of the Department of the Treasury – the Financial Crimes Enforcement Network – uses data mining and pattern matching decision-making to search for evidence of money laundering in large databases of financial transactions. *See* Jensen, *supra* note 37, at 5—8; U.S. Congress, Office of Technology Assessment, Information Technologies for the Control of Money Laundering, OTA-ITC-630 (Sept. 1995), *available at* http://www.wws.princeton.edu/~ota/ns20/year_f.html.

The Securities and Exchange Commission, through the self-regulating stock exchange agencies, uses data mining applications to search for patterns of insider trading. *See*, *e.g.*, Daniel Franklin, Data Miners, Time Magazine Online (Dec. 23, 2002), *at* http://www.time.com/time/globalbusiness/article/0,9171,1101021223-400017,00.html?cnn=yes; Press Release, Innovative Use of Artificial Intelligence: Monitoring NASDAQ for Potential Insider Trading and Fraud, American Association for Artificial Intelligence (July 30, 2003), *available at* http://www.aaai.org/Pressroom/Releases/release-03-0730.html:

> [T]here is a growing need for better tools to monitor the market for suspicious activity that warrants closer inspection. The millions of trades, wire stories, and SEC filings each day makes it impossible for humans alone to sift through all the data to perform surveillance. To mine these vast stores of data, NASD has harnessed computers to sweep through all the data, identify and link items of potential interest, then present them to human analysts for further review.

Thus, another practical problem with efforts to simply block particular government research and development projects or outlaw specific technologies over privacy concerns is that "the genie is already out of the bottle." Resisting developments that have already occurred or will occur elsewhere regardless of whether any particular government project (for example, Terrorist Information Awareness ("TIA")[44]) is shut down seems futile and counter-productive.[45]

Further, even if it were possible to prevent eventual adoption by blocking particular government research projects, to do so would intentionally preserve inefficiencies in the methods for legitimate intelligence and law enforcement analysis. To hobble the government's ability to take advantage of our national technical prowess in information technology at a time when American security is at stake seems a dereliction of the responsibility of a civil society to protect its citizenry against violence from others in addition to protecting individual liberties.[46] This is particularly the case when the opposition is in the main premised on a general misunderstanding of both the technology

Another example of data mining technologies in use by private sector entities in order to comply with law enforcement regulatory requirements is the use by banking and financial services companies of automated analysis technologies to comply with the USA PATRIOT Act, Pub. L. No. 107-52, 115 Stat. 272 (2001), requirements relating to money laundering and terrorist financial activities. For example, Wachovia Corp. plans to deploy "SAS' Anti Money Laundering solution to find patterns in transaction that could indicate suspicious activity." Callaghan, *supra* note 41.

In addition, pattern-recognition (and deviation analysis) is widely used in both the private sector and in government agencies to detect fraud or illegal activity in insurance claims, illegal billing, telecommunications crime, and to detect and prevent computer intrusions.

[44] The TIA project is discussed in greater detail in Part II *infra*. Funding for TIA and any "successor program" was eliminated in the Department of Defense Appropriations Bill, 2004, H.R. 2658, 108th Cong., see *supra* note 28. However:

> If anybody thought data mining was going to go away with TIA, they were sorely mistaken," said Steven Aftergood, director of the project on government secrecy at the Federation of American Scientists. "Data mining as a concept is commonplace in the private sector and in various parts of the intelligence community.

William New, *Data Mining in the "Nooks and Crannies,"* Nat'l J., Oct. 6, 2003, *available at* http://nationaljournal.com/pubs/techdaily/features/issues/issu031006.htm.

[45] *See* Rosenzweig, *supra* note 26, at 5 n.12 ("The commercial development of TIA-like technology demonstrates another flaw in the critics' argument: They are attempting to sweep back the tide."); New, *supra* note 44.

[46] "In a liberal republic, liberty presupposes security; the point of security is liberty." *See* Thomas Powers, *Can We Be Secure and Free?,* 151 Public Interest 3, 5 (Spring 2003). Powers goes on to argue that the politicization of the civil liberties debate has resulted in a false dichotomy – a choice between liberty or security – that is inconsistent with the liberal political foundation on which this country was founded. *Id.* at 16—20 "From [Madison's] point of view, it is clear that there is not so much a 'tension' between liberty and security as there is a duality of our concern with security, on the one hand, and with liberty, on the other." *Id.* at 21. *See also* Rosenzweig, *supra* note 26, at 23 (concluding (after citing John Locke) that "the obligation of government is a dual one: to protect civil safety and security and to preserve civil liberty.").

and current development programs and the likely impact these technologies may have on privacy and civil liberties. [47]

In any case, this Article assumes that these technologies will continue to be developed and that eventually practical efficiency[48] and availability[49] will compel adoption by government for certain domestic security or law enforcement purposes.[50] Therefore, it seems short-sighted for those concerned about privacy and civil liberties to oppose government research and development efforts in these areas,[51] since it is only through involvement in and oversight of these government sponsored projects that privacy interests can be incorporated into the development process. To meet legitimate privacy concerns, technical features supporting privacy protection need to be built into the architecture of the technologies from the start.[52]

[47] *See* Heather MacDonald, *Total Misrepresentation*, Weekly Standard, Jan. 27, 2003 ("To call the [media descriptions of TIA] caricatures of the Pentagon project is too charitable. Their disconnect from reality was total."), *available at* http://www.weeklystandard.com/Content/Public/Articles/000/000/002/137dvufs.asp; Taylor, *supra* note 20.

Even within the technical community there is significant divergence in understanding what these technologies can do, what particular government research programs entail, and the potential impact on privacy and civil liberties of these technologies and programs. *Compare* Letter from Public Policy Committee of the Association for Computing Machinery to Senators John Warner and Carl Levin (Jan. 23, 2003) (expressing reservations about the TIA program) [hereinafter Public Policy Committee of the ACM], *available at* http://www.acm.org/usacm/Letters/tia_final.html, *with* SIGKDD of the ACM, *supra* note 28 (defending data mining technology and expressing concern that the public debate has been ill-informed and misleading); *see also* Rosenzweig, *supra* note 26, at 3 and n.5 ("Few people, including many . . . critics, seem to understand what the TIA program entails or how it would work. . . . [but] those with the seeming greater familiarity with the technologies are less apocalyptic in their reactions.").

[48] The requirement for automated analysis of large datasets is a function of the inability of traditional manual analysis to manage the size and dimensionality of existing data collection methods. It would be an unusual polity that demanded accountability from its representatives for being unable to "connect the dots" from existing datasets to prevent terrorist acts yet denied them the available tools to do so. "While technology remains one of this nation's greatest advantages, it has not been fully and effectively applied in support of U. S. counter terrorism efforts. Persistent problems in this area include . . . a reluctance to develop and implement new technical capabilities aggressively." Joint Inquiry Report, *supra* note 3, at xvi.

[49] *See supra* notes 41, 43 (detailing widespread use and availability).

[50] This Article makes the general assumption that there is a category of criminal – terrorist – for which aggressive preventative law enforcement strategies are appropriate but it is beyond the scope of this Article to address precisely where that line is to be drawn by delimiting particular types of crimes that meet that criteria, or by specifying which government organs or agencies should be permitted particular uses. *See supra* note 40.

[51] Especially since the "public debate has not [adequately] distinguished between the research and development of data mining technology and the application and use of these technologies by specific agencies on specific data for specific purposes." *See* SIGKDD of the ACM, *supra* note 28. Thus, much of the opposition is based on suppositions the truth or falsity of which is the very subject of the proposed research. *See* Rosenzweig, *supra* note 26, at 2.

[52] *See supra* note 25 (discussing value sensitive design); *infra* Part IV (discussing technology development strategies for addressing privacy concerns in data aggregation and analysis tools).

This Article examines data aggregation and analysis technologies, particularly data mining, and the privacy policy implications of their employ for certain domestic security purposes. This Article is not an attempt to critique or endorse any particular government program (for example, TIA[53] or Computer Assisted Passenger Pre-Screening II ("CAPPS II"),[54] which are discussed in Part II infra) or to make specific policy or legal recommendations for a particular implementation. Rather, it attempts to highlight the intersection of technological potential and development with certain policy (particularly, privacy) concerns in order to inform the debate.

However, in that regard this Article does proffer certain overriding principles that should govern development and implementation of these technologies in order to help achieve security with privacy.[55] These principles include that these technologies:

- be used only as investigative, not evidentiary, tools;[56]
- be used only for investigations of activities about which there is a political consensus that aggressive preventative strategies are appropriate;[57] and

That use of these technologies for particular application in particular agencies:

- be subject to strict congressional oversight and review before implementation;[58]

[53] The TIA program (earlier called the Total Information Awareness program, *infra* note 152) was a counter-terrorism project of the Information Awareness Office ("IAO") of the Defense Advanced Research Projects Agency ("DARPA") specifically intended to research technologies suitable for detecting and identifying terrorists and preempting terrorist acts. *See* discussion *infra* Part II.

DARPA (previously "ARPA") is the primary research and development unit of the U.S. Department of Defense. *See generally* DARPA, DARPA HOME, *at* http://www.darpa.mil/. DARPA funded the initial development of ARPANET, the precursor to the Internet, in the 1960s. See Barry M. Leiner et al., *A Brief History of the Internet*, *at* http://www.isoc.org/internet/history/brief.shtml.

[54] Computer Assisted Passenger Pre-Screening ("CAPPS II") is a Transportation Safety Administration project designed to pre-screen passengers to assess threat levels to aviation security. *See* discussion *infra* Part II.

[55] *See generally* Powers, *supra* note 46 (discussing the duality of security and liberty within the liberal political tradition that informed the Founding Fathers); ISAT 2002 Study, *Security with Privacy*, Dec. 13, 2002 (discussing the purely technical aspects of security with privacy), *available at* http://www.taipale.org/references/isat_study.pdf. This document was formerly available at http://www.darpa.mil/iao/secpriv.pdf.

[56] Data mining (particularly, pattern-matching) should not automatically trigger significant adverse law enforcement consequences for individuals such as "black-listing" or arrest without further review and analysis using traditional methods and procedures of corroboration. Data mining should be considered an investigative tool that can help focus law enforcement resources on potentially useful areas or subjects, but not as a determinant of guilt or innocence. *See infra* text accompanying notes 108—118 (discussing post-processing and decision-making), 119—128 (describing the uses of data mining in domestic security).

[57] *See supra* notes 4, 40 and accompanying text. Compare Rosenzweig, *supra* note 26, at 2, 22 (calling for an absolute statutory ban on the use of certain of these technologies except for anti-terrorist, foreign intelligence or national security activities) with the general law enforcement applications already in use described *supra* in notes 41 and 43.

- be subject to appropriate administrative procedures within executive agencies where they are to be employed;[59]

- be subject to appropriate judicial review in accordance with existing due process doctrines;[60] and

That development and design of these technologies include features to support privacy and civil liberty protections, including specifically:

- rule-based processing functions; [61]

- selective revelation functions;[62] and

- strong credential and audit functions.[63]

Part I of this Article provides an introduction to data aggregation and analysis technologies, in particular, data mining. Part II examines certain government initiatives as paradigmatic examples of development efforts in these areas. Part III briefly outlines the primary privacy concerns and the related legal framework. Part IV suggests certain technology development strategies that could help ameliorate some of the privacy concerns. And, Part V concludes by restating the overlying principles that should guide development in these technologies.

[58] *See supra* note 28; *see also* Rosenzweig, *supra* note 26, at 9 ("in light of the underlying concerns over the extent of government power, . . . formal congressional consideration and authorization of the use of [data mining] technology, following a full public debate, should be required before the system is deployed"); Paul Rosenzweig & Michael Scardaville, Heritage Foundation, *The Need to Protect Civil Liberties While Combating Terrorism: Legal Principles and the Total Information Awareness Program* 9 (2003).

[59] *See* Rosenzweig, *supra* note 26, at 9—12 (describing administrative procedures for use); Gallington testimony, *supra* note 28 (suggesting the applicability to these technologies of existing procedures and oversight regimes used to manage foreign signal intelligence).

[60] *See* Rosenzweig, *supra* note 26, at 15, 21—22 (describing judicial review for breaking the anonymity barrier and calling for severe administrative and criminal sanctions for misuse or abuse and for a private right of civil action by individuals aggrieved by misuse).

[61] *See infra* Part IV. Rule-based processing controls how data from multiple sources with potentially different access rules, permissions and privacy restrictions can be processed.

[62] *Id.* Selective revelation protects privacy by separating identity from transactional or behavioral data to protect anonymity or otherwise by incrementally revealing additional data to limit intrusion on privacy.

[63] *Id.* Strong credentialing and audit features should seek to make "abuse difficult to achieve and easy to uncover" by providing secure access control and tamper-resistant evidence of where data goes and who has had access to it. Rosenzweig, *supra* note 26, at 21.

Part I. Data Mining: The Automation of Traditional Investigative Techniques

Understanding technology development generally should always begin with an understanding of the underlying real world process to which the technological solution is to be applied and the particular characteristics of the technology being considered. Data mining is one of a number of tools that can more broadly be classified as *sense-making* applications – that is, software tools that bring meaning to vast amounts of raw data.[64] In the context of law enforcement, data mining is no more than the computational automation of traditional investigative skills – that is, the intelligent analysis of myriad "clues" in order to develop a theory of the case.[65]

The popular view of investigation in law enforcement is that there must first be a specific crime and that law enforcement then follows particularized clues or suspicions after the fact. In reality, investigators often look for criminal patterns or hypothetical suspects in order to anticipate future crime. For example, investigators may use pattern recognition strategies to develop modus operandi ("MO") or behavioral profiles, which in turn may lead either to specific suspects (profiling as identifying pattern) or to crime prevention strategies (profiling as predictor of future crime, resulting, for example, in stakeouts of particular places, likely victims, or potential perpetrators).[66]

The application of data mining technologies to domestic security is the attempt to automate certain analytic tasks to allow for better and more timely analysis of existing datasets with the intent of being able to prevent terrorist acts by identifying and cataloging various threads and pieces of information that may already exist but remain unnoticed using traditional means of investigation. Further, it attempts to develop predictive models based on known or unknown patterns to identify additional people, objects, or actions that are deserving of further resource commitment or attention.[67]

[64] *See generally* M. Mitchell Waldrop, *Can Sense-making Keep Us Safe?*, MIT Tech. Rev., 43 (Mar. 2003), *available at* http://www.technologyreview.com/articles/waldrop0303.asp. Other sense-making technologies include data-visualization, statistics, modeling, etc. In technical usage, sense-making technologies are considered "knowledge discovery" technologies.

[65] *See, e.g.*, Cook, *supra* note 43, at A1 (describing the use of CopLink by the Boston Police Department, "[t]hough the program is bound to alarm some privacy advocates with its relentless drive to find even the most subtle connections between people and events, officers point out that the software does nothing police don't already do, and it is still the police - not the machine - deciding what leads are worth following.").

[66] *See generally* Brent Turvey et al., Criminal Profiling: An Introduction to Behavioral Evidence Analysis (2d ed. 2002); Ronald M. Holmes & Stephen T. Holmes, Profiling Violent Crimes: An Investigative Tool (3d ed. 2002).

Note that the CPFDA, *supra* note 28, as proposed would seem to prohibit matching any such profile against a database (§4(a) prohibits the use of any queries based on "hypotheticals").

[67] *See supra* notes 40, 50 (discussing the need for prospective identification and preventative strategies in the context of terrorism). From a civil liberties perspective, further scrutiny based on behavioral profiling would seem less problematic than current analogs, for example, racial or national origin profiling. From a security perspective it would allow concentrating resources on more likely targets.

A. *Data Mining: An Overview*

Data mining is the process of looking for new knowledge in existing data. The basic problem addressed by data mining is turning low-level data, usually too voluminous to understand, into higher forms (information or knowledge) that might be more compact (for example, a summary), more abstract (for example, a descriptive model), or more useful (for example, a predictive model).[68] At the core of the data mining process is the application of data analysis and discovery algorithms to enumerate and extract patterns from data in a database.[69]

A formal definition of data mining is "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data."[70] Each aspect of this definition has important implications for our purposes in trying to understand what data mining is and in distinguishing it from previously familiar data-processing and database query technologies.

Extracting *implicit* information means that the results of data mining are not existing data items in the database.[71] Traditional information retrieval from a database returns arrays consisting of data from individual fields of records (or entire records) from the database in response to a defined or specified database query.[72] The results of the traditional database query are explicit in the database, that is, the answer returned to a query is itself a data item (or an array of many items) in the database. Data mining techniques, however, extract knowledge from the database that is implicit – knowledge that "typically [does] not exist a priori" is revealed.[73] Data mining generally identifies

[68] *See* Fayyad et al., Databases, *supra* note 39, at 37; *see also* Jensen, *supra* note 37, at slide 22 ("A key problem [for using data mining for counter-terrorism] is to identify high-level things – organizations and activities – based on low-level data – people, places, things and events.").

[69] *See* Bhavani Thuraisingham, Data Mining: Technologies, Techniques, Tools and Trends 110—112 (1999) ("Now we come to the important part of data mining, and that is the algorithms and techniques employed to do the mining.").

[70] William J. Frawley et al., *Knowledge Discovery in Databases: An Overview*, 13 AI Mag. 57, 70 (1992); *see also* Advances in Knowledge Discovery and Data Mining 6 (Usama Fayyad et al. eds., 1996) [hereinafter Fayyad et al., Overview]; *infra* note 80 (distinguishing the evolution of the terms data mining and knowledge discovery).

[71] *See generally* Fayyad et al., Overview, *supra* note 70, at 3—9; Richard J. Roiger & Michael W. Geatz, Data Mining: A Tutorial-based Primer 318 (2003) (discussing the difference between heuristics and statistics for inductive problem solving).

[72] A "query" is a search of a database for all records satisfying some specified condition. Returns from a traditional database query are sometimes referred to as "tuples." The term originates as an abstraction of the sequence: single, double, triple, quadruple, quintuple, . . . n-tuple. *See* Joseph S. Fulda, *Data Mining and Privacy*, 11 Alb. L.J. Sci. & Tech. 105, 106 (2000) (citing a definition of tuples as "finite ordered sequences of arbitrary elements").

[73] Thomasz Imillienski & Heikka Mannila, *A Database Perspective of Knowledge Discovery*, 39 Comm. of the Ass'n for Computing Machines 60 (1996).

patterns or relationships among data items or records that were not previously identified (and are not themselves data items) but that are revealed in the data itself.

Thus, data mining extracts information that was *previously unknown*. That is, data mining employs complex techniques[74] that can provide answers to questions that have not been asked (or elicit questions for problems that have not been identified). As discussed below in Part III, it is this aspect – the creation of new knowledge without previously particularized suspicion – that creates the most unease among privacy advocates in the context of law enforcement use of these techniques.[75]

Finally, and most importantly, data mining extracts information that is *potentially useful*. In order to be actually useful, the results must be appropriately analyzed, evaluated, interpreted, and applied within the specific domain to which they relate before being acted upon.[76] This process includes checking for and resolving conflicts with previously known or derived knowledge, as well as determining decision-making thresholds for specific applications – that is, establishing a *confidence interval* for determining how well the discovered pattern describes or predicts within the formulated goals.[77]

If unconstrained by limits, the number of potential patterns elicited through data mining is potentially infinite.[78] Thus, for data mining to be useful, the knowledge

[74] Data mining can include the use of simple heuristics, complex algorithms, artificial intelligence, fuzzy logic, neural networks and genetic-based modeling; and can involve natural language text mining, machine learning or intelligent agents. However, the purpose of this Article is to provide a general overview so the term "algorithm" is used generally throughout to include any of these analytic approaches or tools. *See also infra* note 100 (distinguishing *functions* and *representations*).

[75] Critics of employing data mining techniques for law enforcement often refer to this aspect as allowing for "fishing expeditions." See, for example, the lead-in to the comment by Edward Tenner in Bray, *supra* note 23, and Alex Gourevitch, *Fishing Expedition*, The American Prospect Online, June 26, 2003, *available at* http://www.prospect.org/webfeatures/2003/06/gourevitch-a-06-26.html.

Of course, "fishing expedition" as a metaphor for a non-particularized search only applies to amateur fishermen. Any experienced angler knows that a successful fishing expedition is generally targeted at a specific species and is based on explicit as well as implicit domain knowledge about the species, its habitat and habits, exactly what supporters of data mining for law enforcement would say can be applied in the search for terrorists and exactly what the research and development efforts aim to test.

Further, describing data mining as a "fishing expedition" conflates the development of the "fishing net" (descriptive or predictive models) with its particular application. *See infra* text accompanying notes 97—98.

[76] Extracted (i.e., discovered) knowledge can be used directly, incorporated into another system for further action, or simply documented and reported to interested parties. Fayyad et al., Databases, *supra* note 39, at 42. "Knowledge" in the context of knowledge discovery or data mining is a purely user-oriented utility function and by no means absolute. *See* Fayyad et al., Overview, *supra* note 70, at 8—9. Thus, the important distinction to be drawn between data mining to develop a model or describe relationships from post-processing decision-making, that is, the application of the model to new data. *See infra* text accompanying notes 108—118 (discussing post-processing and decision-making).

[77] *See infra* note 104 and accompanying text (discussing "confidence intervals").

[78] *See infra* note 105. Knowledge discovery from data is essentially a statistical undertaking. Early efforts at data mining were often ridiculed in statistics for producing endless, useless regressions. This is because any data set (including randomly generated data) will eventually show patterns that appear to be

discovery process must constrain the data mining step so that results meet a measure of *interestingness* relative to the original goal. *Interestingness* is the term-of-art for the overall measure of pattern value – combining validity, novelty, usefulness, and simplicity.[79]

Thus, a central question to be answered before actually implementing any system for data mining in the context of domestic security is whether the technology is useful in identifying potential terrorist actors. If it is (or potentially is), then the question becomes whether it can do so in a way that protects (or enhances, relative to the alternatives) privacy. Despite a long heritage rooted in statistical analysis, artificial intelligence, and related fields, data mining is not a perfected technology and its usefulness in identifying terrorists from among the general population has not been proven. However, the question of its ultimate efficacy for a particular purpose is not grounds for opposing its research, development, or testing for that very purpose.[80]

B. *Data Mining and The Knowledge Discovery Process*

Technically speaking, the term "data mining" refers to a particular step in the knowledge discovery process.[81] The steps that compose the knowledge discovery

statistically significant, but, in fact, are not. The "art" of statistics is hypothesis selection – the aim of KDD is to automate (to the degree possible) the application of this art.

[79] Fayyad et al., Overview, *supra* note 70, at 8. *See generally* Abraham Silberschatz & Alexander Tuzhilin, *What Makes Patterns Interesting in Knowledge Discovery Systems*, IEEE Transactions on Knowledge and Data Engineering 970 (1996), *available at* http://www.computer.org/tkde/tk1996/k0970abs.htm.

[80] Opposition to research on the basis that it "might not work" is an example of what has been called the "zero defect" culture of punishing failure, a policy that stifles bold and creative ideas. At least one commentator has characterized such opposition to risk-taking as "downright un-American." David Ignatius, *Back in the Safe Zone*, Wash. Post, Aug. 1, 2003, at A:19 (discussing the knee-jerk opposition to a "terrorist futures market").

Obviously, opposition to research on technologies where the development or testing process is itself inherently ethically suspect or destructive of other values, for example using human subjects to test deadly drugs, or where even the successful outcome might be socially undesirable, for example human cloning, is legitimate. Thus, in the case of data mining, care should be taken to protect privacy during development, for example, by insuring that access to "real" data is restricted until privacy issues are resolved or efficacy tested.

However, opposition to data mining because it might be successful (that is, it might actually be useful in identifying terrorists hidden among the general population) is problematic given both the high stakes involved in domestic security applications, the potential for limiting privacy impacts, and the many legitimate and beneficial other uses of the technology. Other beneficial uses include, for example, medical drug design experimentation, biological micro-arrays, as well as countless business and scientific applications. *See, e.g., supra* notes 15—16 (discussing business and scientific applications); SIGKDD of the ACM, *supra* note 28 (noting medical and life-saving applications); Bloom, *infra* note 285 (TIA likely to lead to improvements in general search technology). In addition, as noted *infra* in Part IV, certain related technologies are being developed to manage intellectual property (rules-based processing) and spam (analytic filtering).

[81] Historically, "data mining" was the term mostly used by statisticians, data analysts and the MIS database community. "Knowledge discovery in databases" (KDD for short) was coined by Gregory

process are (1) pre-processing (including goal identification; data collection, selection, and warehousing; and data cleansing or transformation), (2) "data mining" itself, and (3) post-processing (including interpretation, evaluation and decision-making or action).[82]

1. Pre-Processing

The first (and perhaps most important) step in pre-processing – *goal identification* – involves understanding the domain in which knowledge discovery methodologies are to be applied and identifying desired outcomes.[83] In the context of domestic security, the security goal might be to identify potential terrorists from among the general population by searching for electronic footprints in databases of personal and transactional data.[84] However, in the broader context of employing these technologies within a liberal democracy, the overall goal should be to do so while protecting privacy and civil liberty values.[85]

Next in pre-processing comes *data collection*, *selection,* and *warehousing*, which involve assembling data that is to be mined into a single dataset for subsequent processing.[86] Historically, data mining applications generally required that data to be mined be aggregated in a single database generally referred to as a data warehouse.[87]

Piatetsky-Shapiro in 1989 and became popular in the artificial intelligence and machine learning community. The business press popularized "data mining" to the point that a current web search for data mining yields an order of magnitude more hits than does one for KDD. Although KDD and data mining are somewhat interchangeable in usage, the "knowledge discovery process" is generally used for describing the overall process, including data preparation and post-processing, while data mining is used to refer to the step of applying algorithms to the cleansed data. *See* Gregory Piatetsky-Shapiro, *Knowledge Discovery in Databases: 10 Years After*, SIGKDD Explorations, Jan. 2000, at 59.

[82] Roiger, *supra* note 71, at 147—174; Fayyad et al., Databases, *supra* note 39, at 42.

[83] Fayyad et al., Databases, *supra* note 39, at 42.

[84] That is, to "connect the dots" among the electronic data trails left behind when terrorists engage in mundane activities of everyday life in preparation for terrorist actions. Obviously, many, if not all, preparatory steps leading up to an attack are legal. However, they may become suspicious when combined in a particular way in a particular context. *See infra* text accompanying notes 125—127 (discussing searching for terrorist pre-cursor acts and the use of multiple relational identifiers in an "ensemble classifier" to reduce error).

[85] It is a premise of this Article that protecting civil liberties and security are dual obligations of the liberal democratic state. *See supra* note 46*; see also* Chloe Albanesius, *Officials Defend Idea of Data Mining; Experts Weigh Options*, Nat'l J.'s Tech. Daily (Dec. 2, 2003) (quoting Kim Taipale, "The goal is security with privacy . . . . Security and privacy are not dichotomous rivals to be traded one for another in a zero-sum game; they are dual objectives, each to be maximized within certain constraints.").

[86] Fayyad et al., Databases, *supra* note 39, at 40—42. These processes are also sometimes referred to as "data integration" or "data fusion." The privacy implications of data aggregation or integration are discussed in Part III *infra*.

[87] The concept of data warehousing – assembling data about customers from several internal databases and combining that with external data – is a relatively common practice in the commercial sector. For example, a financial institution might combine basic account holder information from one database

However, current research and development efforts are aimed at developing techniques for "virtual" data aggregation in which a single query or intelligent agent negotiates access to multiple distributed databases on local terms. Under this approach, instead of importing data and standardizing it for processing centrally, an intelligent "prospecting agent" accesses distributed databases over a network and adapts to the local database conditions or requirements, both for database access and for data processing.[88] Further (and importantly for maintaining privacy protections in domestic security applications) data mining itself does not require a single, massive database. "Provided that certain (very low) size thresholds are exceeded to provide statistical validity, data mining techniques can be applied to databases of a wide variety of sizes."[89] Virtual aggregation

with transactional records from another. These are then combined with additional personal or demographic data from external sources, perhaps credit reporting agencies, into a single database to be mined. Historically, information sharing among government agencies and between federal and state sources has been weak or non-existent. *See, e.g.*, Peter Paulson, Unisys.com, *The Enemy Within - Necessitating Vertical and Horizontal Data Integration*, *available at* http://www.unisys.com/public_sector/insights/insights__compendium/enemy__within.htm (last visited Dec. 15, 2003). Prominent among the goals in re-engineering law enforcement efforts to prevent terrorism is improving information sharing and agency interoperability. *Id.; see also* USDOJ Fact Sheet, *supra* note 4; Len Silverston, "Terrorism: A Call for Data Integration" *available at* http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=1819; Homeland Security Information Sharing Act, HR 4598 passed by the House of Representatives 422-2, June 26, 2002, referred to Senate Judiciary Committee June 27, 2002, *available at* http://thomas.loc.gov/cgi-bin/bdquery/z?d107:HR04598:@@@X; William Mathews, FBI to Build Data Warehouse, Computing Week (June 3, 2002), *available at* http://www.fcw.com/fcw/articles/2002/0603/news-fbi-06-03-02.asp. *But see* U.S. DOJ, Office of Inspector General, Audit Division, The Federal Bureau of Investigation's Efforts to Improve the Sharing of Intelligence and Other Information, Audit Report 04-10 (Dec. 2003) (concluding that the FBI still doesn't sufficiently share intelligence information about terrorism within its own ranks or with other agencies); Shane Harris, Report Finds Information Sharing Still a Problem for FBI, GovExec.com (Dec. 22, 2003), *available at* http://www.govexec.com/dailyfed/1203/122203h1.htm; *see also* Hsu, *supra* note 43 (discussing "data sharing" among the District of Columbia and four major Eastern states.). *See generally* the discussion of the MATRIX program, *supra* note 42.

[88] The technology and policy implications of techniques employing virtual data aggregation are discussed *infra* Part III and IV.

Developing technologies to access distributed databases and information sources was among the primary goals of the TIA-related Genisys program within IAO. *See infra* notes 173—174 and accompanying text. "[The Genisys program] aims to develop a . . . database architecture and algorithms that would allow analysts and investigators to more easily obtain answers to complex questions by eliminating their need to know where information resides or how it is structured in multiple databases." *IAO Report to Congress regarding the Terrorism Information Awareness Program* (May 20, 2003) in response to Consolidated Appropriations Resolution, 2003, No. 108-7, Division M, §111(b) [signed Feb. 20, 2003] [hereinafter IAO Report]. The program "aims to create technology that enables many physically disparate heterogeneous databases to be queried as if it [sic] were one logical 'virtually' centralized database." *Id.* at A-11.

[89] Jensen, *supra* note 37, at slide 18. Thus, for privacy purposes, rather than assuming a single, massive database (or even a single "virtual" database of equally accessible distributed databases) a more practical approach would be to distinguish between primary and secondary datasets with different access rules or privacy protections for each. For example, a particular intelligence agency might data mine its own local database (which might be populated with real world data or hypothetical data based on known attributes) to which it has unrestricted access. Once a predictive model is developed from the primary dataset it may or may not be applied to subsequent secondary datasets (whether additional government

(as contrasted with a single, massive database) provides important technical support for certain privacy protections, for example, by allowing for different privacy standards to be applied before access is granted to additional databases or to particular information.[90]

Once data has been collected or aggregated, traditional data mining practices required that the data be *cleansed* or *transformed* – purging or correcting data that is redundant, unreliable, or otherwise unusable, and standardizing data for processing.[91] Generally, cleansing and transformation are considered distinct steps. Cleansing involves eliminating noise and dealing with missing data, while transformation involves data normalization, type conversion, and attribute selection. Thus, cleansing involves tidying up the database, while transformation involves manipulating the data itself to more easily match the intended processing or algorithm to be employed.[92]

Data cleansing and transformation are key requirements to achieving useful results from current data mining applications.[93] However, data mining does not necessarily require "clean" data. There are existing statistical techniques that can be used to compensate for known data problems, including missing or noisy data.[94] And, data mining itself can be used to develop data cleansing algorithms.[95]

databases or to commercial databases) based on that model's particular characteristics (for example, its "intrusiveness" on privacy and its "reasonableness" as a predictor).

The point is not that there is no privacy concern, but that there are subsequent intervention points – enabled by the technology and its process – to apply legal process controls. Legal and administrative procedures to review or authorize the "reasonableness" of further investigation or action can parallel the existing structure of investigative, reasonable suspicion or probable cause standards currently required for increasingly intrusive law enforcement actions such as opening an investigation, engaging in a physical search, or arrest, etc. *See* Rosenzweig, *supra* note 26, *passim* (proposing a procedural implementation structure and distinguishing among both the types of databases to be accessed (e.g., government versus commercial) and the type of methodology to be employed (e.g., subject- or pattern-based query)).

[90] *See* Jensen, *supra* note 37, at slide 18 (discussing a multi-tiered approach to databases consisting of primary and secondary datasets). A distributed architecture has important technical, security and privacy implications. *See infra* text accompanying Part II. Further, current research using enhanced data mining models premised on using iterative, multi-pass inference accessing different types or amounts of data on each pass (i.e., supporting a distributed architecture) show that "substantially smaller amounts of data can be accessed in later stages of inference." *See* David Jensen et al., *Information Awareness: A Prospective Technical Assessment*, ACM SIGKDD '03, Aug. 2003 [hereinafter Jensen, Technical Assessment]. This early research provides additional support against privacy critiques premised on methodologies requiring a single, massive database.

[91] Fayyad et al., Databases, *supra* note 39, at 41 Fig. 1; Roiger, *supra* note 71, at 153—161.

[92] Fayyad et al., Databases, *supra* note 39, at 41 Fig. 1; Roiger, *supra* note 71, at 153—161. An example of the former would be eliminating redundant records from the database, while an example of the latter might be combining two attributes of low predictive power (for example, price and earnings) into a single attribute of higher predictive value (such as price/earnings or P/E ratio).

[93] *See generally* Mauricio Hernandez & Salvatore J. Stolfo, *Real World Data is Dirty: Data Cleansing and the Merge/Purge Problem*, J. of Data Mining and Knowledge Discovery (1998); Dorian Pyle, Data Preparation for Data Mining (1999).

[94] For example, Bayesian probability can readily handle missing data. *See, e.g.*, Marco Ramoni & Paolo Sebastiani, *Robust Learning with Missing Data*, 45 Machine Learning 147 (2001). *But see* Jensen, *supra* note 37, at slide 24 (noting that "in relational data, *fragmentary data* (e.g., some missing . . . associations or . . . transactions) can cause errors in naive methods for data mining.").

For many purposes, standard pre-processing and warehousing (that is, *data integration* or *data fusion*) are in themselves sufficient to enable traditional query-based search strategies or group comparisons to provide useful results from the aggregated and cleansed data without engaging in actual "mining" (without looking for unknown patterns).[96] However, where the data does not provide explicit solutions, or where the data is too voluminous for traditional deductive query methods, data mining strategies are applied to extract implicit knowledge. A key challenge for counter-terrorism is to extract implicit relational knowledge – information about relationships between data – that is not explicit in the data.

2. Data Mining: Descriptive and Predictive Modeling

The data mining step itself consists of the application of particular algorithms[97] to the cleansed data in order to elicit, identify, or discover certain previously unknown characteristics of the data, including descriptive and predictive patterns or relationships, which emerge from the data itself.[98] Strictly speaking, data mining involves developing the descriptive or predictive model (identifying the rules); while applying the model to new data (predictive profiling or pattern-matching) is decision-making (or *inference*, that is, using the discovered knowledge to infer or predict something), a part of post-processing.[99]

For expository purposes, this Article describes two basic algorithm types here: *clustering* and *association rules*.[100] Clustering includes both classifying data into pre-

[95] *See, e.g.,* Isabelle Guyan, *et al.*, *Discovering Informative Patterns and Data Cleaning in* Fayyad et al., Overiew, *supra* note 70, at 181, 187—193 ("We propose data cleaning algorithms and analyze experimental results.").

[96] *See* Tal Z. Zarsky, *Mine Your Own Business*, 1 Yale J. L. & Tech. 8 (2003) (using traditional analytic tools on warehoused data is "adequate for many businesses . . . that are not interested in the additional expense data mining would entail"); *see also* Jesus Mena, Investigative Data Mining for Security and Criminal Detection 39—74 (2003); Fayyad et al., Databases, *supra* note 39, at 40 (discussing online analytical processing (OLAP) tools that allow for multidimensional data analysis as a popular approach for analysis of warehoused data); Roiger, *supra* note 71, at 318 (discussing query tools, OLAP, and visualization tools); the discussion of CAPPS II in Part II, *infra* (distinguishing CAPPS II, a data integration project, from TIA, a data mining research and development project); Eric Lichtblau, *Administration Creates Center for Master Terror 'Watch List'*, N.Y. Times, Sept. 16, 2003, at A20 (discussing the Terrorist Screening Center to be administered by the FBI to consolidate terrorist 'watch lists' from various federal agencies); Press Release, The White House, New Terrorist Screening Center Established, (Sept. 16, 2003), *available at* http://www.whitehouse.gov/news/releases/2003/09/20030916-8.html.

[97] *See supra* note 74 (describing the use of the term "algorithm" to describe a variety of analytics).

[98] Fayyad et al., Databases, *supra* note 39, at 39.

[99] *See* Jensen, *supra* note 37, at slide 12.

[100] *See* Fayyad et al., Databases, *supra* note 39, at 44—45. Additional functional types include *regression, summarization* and *change and deviation detection*. *Id*. *See also The KDD Process for*

existing categories (although this strategy is more properly called "classification"), and the mapping of data to new categories that are created during the data analysis process and that are determined by the data itself (also called "unsupervised clustering").[101]

Association rules are used to discover interesting associations between data attributes and include techniques to describe dependencies between data, find links among data, and model sequential patterns in data.[102] Dependency modeling (showing dependencies among variables), link analysis (developing association rules that describe or predict when certain variables occur together), and sequence analysis (showing sequential patterns) are the core techniques at the heart of data mining research for domestic security.[103]

In both clustering and association rules, the lower level data reveals higher level order or patterns (knowledge) within the constraints and according to the rules of the applicable algorithm. Thus, algorithms need to be developed and selected based on

Extracting Useful Knowledge from Volumes of Data, 39 Comm. of the Ass'n for Computing Machines 27, 31—32 (Usama Fayyad et al. eds, 1996) [hereinafter Fayyad et al., KDD Process]. *Cf.* Roiger, *supra* note 70, at 34—51 and Fig. 2.1. In general, data mining uses mainly known techniques from machine learning, pattern recognition, and statistics. *See also supra* note 74.

In addition*, functions* should be distinguished from *representations*, which can include "decision trees and rules, linear models, nonlinear models (e.g. neural networks), example-based methods (e.g., nearest-neighbor and case-based reasoning methods), probabilistic graphical dependency models (i.e., Bayesian networks), and relational attribute models." Fayyad et al., KDD Process, *supra*, at 32. Model *functions* determine what the algorithms do with the data (i.e., cluster, look for links or associations, etc.) and *representations* determine the method used (for example, following a decision tree or using a case-based example).

Functions and representation are also referred to in the literature as *outcomes* (i.e., classification, clustering, association, deviation detection, etc.) and *techniques* (neural networks, inductive logic programming, decision trees, nearest neighbor, etc.). *See* Thuraisingham, *supra* note 69, at 105—113.

[101] Fayyad et al., Databases, *supra* note 39, at 31. Closely related to clustering is *probability density estimating*, which is estimating from the data the joint multivariate probability function of all the variables or fields in the database. *Id*.

An example of classification in the context of domestic security might be to query a database to identify all Saudi citizens in the United States on student visa attending flight school who also spent time in Afghanistan during the period of Taliban rule (predefining each of the data attributes). An example of unsupervised clustering might be to query the data with the names of the nineteen hijackers and find that various overlapping subgroups can be clustered as Saudi citizens, student visa holders, flight school attendees, etc. (without predefining the categories).

[102] *Id.* at 31—32.

[103] *See infra* Part III (discussion of DARPA's Evidence Extraction and Link Detection program and the Scalable Social Network Analysis program).

Examples in the context of domestic security might be identifying financial relationships between the particular terrorists based on dependency models (for example, if all terrorists in a particular cell received their funding from a common source or common pattern), determining organizational structure through link analysis models (for example, if certain terrorists used the same addresses or called the same phone numbers), and predicting events through time sequence analysis (for example, if communication chatter increased among certain known terrorists or channels prior to a terrorist act then future observations of increased chatter may predict an impending terrorist act). For an example of the latter, see Megan Lane, *How Terror Talk is Tracked*, BBC News Online (May 21, 2003), *available at* http://news.bbc.co.uk/2/hi/uk_news/3041151.stm.

detailed domain-based knowledge and data familiarity in order to avoid irrelevant, misleading, or trivial attribute correlations.[104] Even relevant correlations need to be analyzed as to their significance and usefulness within the context of the original goals. "Blind application of data mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns."[105]

In general, there are two distinct data mining approaches or methodologies – *top-down* or *bottom-up*.[106] The top-down approach begins with a particular hypothesis and seeks to validate it. The hypothesis can be developed from having initially mined the data using a bottom-up approach or can be developed from some real world knowledge. The bottom-up approach analyzes the data and extracts patterns on which a hypothesis or model can be based. The bottom-up approach can be *directed* or *supervised* (where you have some idea what you are looking for) or *undirected* or *unsupervised* (where you have no idea what you are looking for).[107]

3. Post-Processing

Post-processing consists first of interpreting and evaluating the discovered patterns and determining their usefulness within the applicable domain context. The knowledge discovery process can involve several iterations of analyses of the same data

[104] See the discussion of *confidence* and *support* in Roiger, *supra* note 71, at 78—79 and Fayyad et al., Overview, *supra* note 70, at 410—411. Also, even valid correlations can be either too broad or too narrow to be useful in context. An example of the former might be "terrorists tend to have two eyes," while an example of the latter might be "terrorists tend to use boxcutters;" *see also*, Fayyad et al., Overview, *supra* note 70, at 25 discussing *confidence intervals* ("In some applications, it is crucial to attach confidence intervals to predictions produced by the KDD system. This allows the user to calibrate actions appropriately."). Determining appropriate confidence intervals for predictive models used in counter-terrorism is a crucial research area. "DARPA's goal for [TIA] is to find out what is possible. If [TIA] programs cannot extract terrorist signatures from a world of noise . . . there is no reason to proceed. However, if the technology works in a realistic simulation, its advantages for protecting the Nation against terrorism can be weighed against its potential for reducing privacy." IAO Report, *supra* note 88, at A-11. In other words, the confidence interval for these technologies must be developed through research and development before an informed public debate can determine their appropriateness for particular uses. *See als*o SIGKDD of the ACM, *supra* note 28.

[105] Fayyad et al., Databases, *supra* note 39, at 39*; see also* Fayyad et al., Overview, *supra* note 70, at 4. The misuse of statistical tools for data dredging is also referred to as "overfitting the model." These problems are well known, and technical methods (many derived from statistical theory) are available to manage these problems. These methods include randomization tests, cross-validation, ensemble methods, pruning, and penalized evaluation functions. *See* Jensen, *supra* note 37, at slide 12.

For a technical description of the overfitting problem and various techniques for controlling induced errors, see David Jensen and Paul R. Cohen, *Multiple Comparisons in Induction Algorithms*, 38 Machine Learning 309—338 (2000).

[106] *See* Thuraisingham, *supra* note 69, at 109—110 and Figure 7-3.

[107] *Id*. See the examples of classification and unsupervised clustering in note 101 *supra*. Compare the distinction drawn in the text accompanying notes 124—128 *infra* between subject-based and pattern-based applications.

either to increase granularity, for example, by generating sub-categories of clusters or exposing weaker links, or to eliminate certain patterns that are judged to be non-useful within the domain context.[108] Subsequently applying the discovered knowledge to new data sets in order to discover additional relationships, identify particular subjects, or predict future actions is the key step in the knowledge discovery process and is dependent on determining a *confidence interval* (that is, an acceptable error rate) for decision-making (or drawing an inference) in the particular context.[109]

In domestic security applications, different uses of derived knowledge raise different levels of privacy concerns. For example:

a. incorporating discovered knowledge into the domestic security intelligence system (for example, by looking for like occurrences in other agencies' data sets or using the model as an alert threshold for new data),

b. taking action based on discovered knowledge (for example, by referral to law enforcement agencies for follow up investigation of suspects), or

c. simply documenting the discovered knowledge and reporting it to interested parties (for example, by alerting the law enforcement or the public to be on the look-out for certain persons, things or behaviors or briefing political leaders about potential scenarios),

each raise significant but different policy issues in particular contexts that should be subject to public debate and legal procedural protections prior to implementation.[110] The knowledge discovery process provides multiple intervention points for political or legal control of decision-making requiring a more nuanced debate than the all-or-nothing dichotomy between privacy and security that has emerged.[111]

Importantly, data mining results cannot be evaluated independently of the entire knowledge discovery process.[112] The additional steps involved in pre-processing, incorporation of appropriate prior and subsequent domain knowledge into process development, and post-processing itself, are essential to ensuring that useful knowledge

[108] The latter could be considered not of sufficient *interestingness* (*see supra* note 78) within the context of identified goals – it may be unable to detect terrorists (not useful for the primary purpose) or may be too invasive of privacy (for example, a pattern match that required mining health records or that generates an unacceptable level of false positives).

[109] *See supra* text accompanying note 104.

[110] *See* Rosenzweig, *supra* note 26, at 9 (calling for congressional consideration and authorization, following full public debate, before implementation). The notion that public policy related privacy concerns vary in accordance with the potential consequences is explored in Paul Rosenzweig, *Civil Liberty and the Response to Terrorism – Myth and Reality,* 42 Duq. L. Rev. (forthcoming 2004) (draft at 10—14, on file with the author). See also the discussion in the penultimate paragraph of Part III(B), *infra.*

[111] *See supra* discussion in Introduction; *see also* SIGKDD of the ACM, *supra* note 28.

[112] *See* Ronald J. Brachman & Tej Anand, *The Process of Knowledge Discovery in Databases: A Human-centered Approach*, in Fayyad et al., Overview, *supra* note 70, at 37—58.

is derived from the data. It might be more appropriate to view the automated part of the knowledge discovery process (i.e., data mining in the narrow sense) as transforming data into information and that it is the proper post-processing – the expert interpretation and application of the information to a domain for decision-making – that transforms information into knowledge on which action can be based (inference), thus completing the process.

Obviously, the dangers of relying on the automated analysis and output without adequate human oversight or post-processing are especially troublesome in applications involving counter-terrorism or other law enforcement situations where the consequences triggered by the knowledge discovery process can have significant effects on individual liberties. Therefore, a guiding principle in the application of these technologies should be that data mining not be used to automatically trigger law enforcement consequences, such as "black-listing," without further review and analysis.[113] Data mining is a descriptive and predictive tool that should be used to identify patterns or relationships, or identify subjects, for further investigation using traditional means subject to traditional rules of due process.[114] Data mining for terrorism prevention or law enforcement purposes must be considered an investigative, not evidentiary, tool.[115]

Further, data mining is not a substitute for human analytical decision-making.[116] Rather, data mining is a powerful computational tool that can help support human analysts in synthesizing new knowledge, forming and testing hypotheses, and creatively developing models, including valid behavior profiles. Human analysts can thereby draw

[113] Needless to say, even investigation or further scrutiny can be seen as invasive of privacy interests. Thus the need to apply strict procedural and technical constraints to keep the number of false positives low. *See* Jensen, *supra* note 37, at slide 40 ("the problem of false positives emphasizes the need for overall control, oversight, and auditing by expert human analysts.").

[114] Rosenzweig, *supra* note 26, at 8—16 (describing how existing administrative procedures can be mapped or adapted to use of these technologies).

[115] The result of an automated analysis or pattern match should not trigger any significant adverse consequences (for example, "black-listing" or arrest) automatically without further review. The results of a pattern-match should be used only as a predicate for further investigation, that is, as a tool to help allocate investigative resources, not as a determinant of guilt or innocence. *See supra* note 56; Rosenzweig, *supra* note 26, at 16. *Cf.,* European Data Directive, 95/46/EC, 1995 O.J. (L281) 31:

> Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.

This provision has generally been interpreted to require human review of automated decision.

[116] *See* Jensen, *supra* note 37, at slide 39:

> A common myth is that the models produced by data mining algorithms will replace human analysts and decision makers. However, the last two decades of work with artificial intelligence systems — including data mining systems — have shown that these systems are usually best deployed to handle mundane tasks, thus freeing the analyst to focus on more difficult tasks that actually require his or her expertise.

insight from and conduct investigative analysis of large or distributed datasets that may contain relevant information hidden within vast amounts of irrelevant data. The higher levels of human analysis to be applied in pre- and post-processing include theory formation, hypothesis of new relationship or patterns of activity, filtering what is useful from background, and searching for clues that require a large amount of highly specialized domain knowledge.[117] Nevertheless, there are specific contexts in which automated analysis is superior to human judgment and eliminates human error or bias.[118]

C. *Data Mining and Domestic Security*

The information problem facing U.S. intelligence and law enforcement in preventing future terrorist acts is to some extent the same as that documented earlier in this Article with respect to business and scientific applications – that is, large data volumes and limited analytic resources.[119] However, compounding the problem is the fact that relevant data (that is, information about terrorist organizations and activities) is hidden within vast amounts of irrelevant data and appears innocuous (or at least ambivalent) when viewed in isolation. Individual data items – relating to people, places, and events, even if identified as relevant – are essentially meaningless unless viewed in context of their relation to other data points. It is the network or pattern itself that must be identified, analyzed, and acted upon.[120] Thus, data mining for domestic security requires development of additional capabilities because existing techniques were primarily developed to analyze propositional data – to analyze transactional data from

[117] *Cf.* Fayyad et al., Mining, *supra* note 16, at 52.

[118] The use of probabilistic models developed through data mining can substantially improve human decision-making in some contexts. *See* Jensen, *supra* note 37, at slide 39. Using probabilistic models can focus human attention and resources, can outperform humans in certain limited contexts (for example, in certain clinical medical diagnostic applications), and can encourage an institutional culture of hypothesis testing and probability assessment. *Id*. *See generally* Amos Tversky & Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, 185 Science 1124; Judgment under Uncertainty: Heuristics and Biases (Daniel Kahneman et al. eds., 1982) (both describing biased heuristics used in human judgment); Robyn M. Dawes et al., *Clinical versus Actuarial Judgment*, 243 Science 1668 (describing how statistical/actuarial methods often outperform human judgment in certain diagnostic contexts); Zarsky, *supra* note 96, at 47—48 ("There is no convincing reason to suppose that decisions made by software are inferior to the ones made by humans (and . . . there are several occasions where the opposite is true.")).

[119] *See supra* text accompanying notes 15—16, 37—39.

[120] *See* Jensen, *supra* note 37, at slides 21, 22 (identifying the key challenge for counter-terrorism as "analyzing relational data"). An example of how relational data analysis can be useful for counter terrorism can be seen in the analysis of "betweeness" in email traffic. "By looking for patterns in email traffic, a new technique can quickly identify online communities and the key people in them. The approach could mean terrorists or criminal gangs give themselves away, even if they are communicating in code or only discussing the weather." Hazel Muir, *Email Traffic Patterns can Reveal Ringleaders*, New Scientist, *available at* http://www.newscientist.com/news/news.jsp?id=ns99993550 (Mar. 27, 2003).

unrelated subjects to make inferences about other unrelated subjects – and may be poorly suited for relational analysis in the context of domestic security.[121]

Post-hoc analysis of the September 11 terror network shows that these relational networks exist and can be identified, at least after the fact.[122] Research and development efforts in knowledge discovery technologies seek to provide the tools to identify these networks *ex ante*.[123]

Here it is again useful to distinguish between the process of knowledge discovery and the component step of data mining. Knowledge discovery can be used in two distinct ways – by following a subject-based inquiry, or by following a pattern-based inquiry.[124] Subject-based inquiries begin with the specification of a particular data subject, for example, an identified individual, and attempt to develop a more complete picture of that individual, his activities and his relationship to other individuals, places or events. Pattern-based inquiries seek to identify individual people, places, or things based on matching a hypothesized pattern or model. In either case, data mining (in its narrow sense) can be used as the technology for identifying links (in the subject-based inquiry) or for developing descriptive or predictive model (to be used for pattern-based inquiries).

What emerges then are actually three discrete applications for automated analysis in the context of domestic security:

- first, "subject-oriented link analysis," that is, automated analysis to learn more about a particular data subject, its relationships, associations, and actions;

- second, "pattern-analysis" (or "data mining" in the narrow sense), that is, automated analysis to develop a descriptive or predictive model based on discovered patterns; and,

- third, "pattern-matching," that is, automated analysis using a descriptive or predictive model (whether itself developed through automated analysis or not) against new data to identify other related (or "like") data subjects (people, places, things, relationships, etc.).

The policy question then becomes not one of what technology is employed but one of specific application – that is, what data is it permissible to access (for example, which additional government databases or commercial databases), using what

[121] *See* discussion *infra* notes 177—180 and accompanying text (distinguishing between traditional commercial data mining of propositional data and the relational analysis requirements for domestic security applications); *see also* Jensen, *supra* note 37, at slide 24 ("Data mining for counter-terrorism requires a set of new capabilities that are not found in current commercial tools.").

[122] *See* Vladis E. Krebs, *Uncloaking Terrorist Networks*, First Monday (mapping and analyzing the relational network among the September 11 hijackers), *at* http://www.firstmonday.dk/issues/issue7_4/krebs/.

[123] *See infra* notes 149—187 and accompanying text.

[124] Rosenzweig, *supra* note 26, at 6. *Cf. supra* notes 106—107 and accompanying text (compare the discussion of directed and undirected, and top-down and bottom-up, approaches).

methodology (for example, using a subject-based or pattern-based query), and for what purpose (for example, distinguishing between investigation of a known or suspected terrorist and prescreening passengers for air travel). In a general privacy context, subject-based inquiries are related to the problem of *data aggregation* and pattern-based inquiries to *non-particularized suspicion*.[125]

Another important point to understand about application of knowledge discovery techniques to domestic security is related to refining goal identification in the context of preventing terrorist activities. Because spectacular terrorist events[126] are rare, they may be too infrequent for data mining techniques to extract useful patterns. Thus, the focus of automated analysis should be on lower level, frequently repeated events that together may warrant further attention or resource commitment. These activities include "illegal immigration, money transfers, operating front businesses, and engaging in recruiting activity."[127] By combining multiple independent models aimed at identifying each of these lower level activities in what is commonly called an *ensemble classifier,* the ability to make inferences about (and potentially disrupt) the higher level, but rare, activity – the terror attack – is greatly improved.[128]

Part II. Data Aggregation and Data Mining: An Overview of Two Recent Initiatives

This section considers two government programs that have attracted significant public attention and opposition from privacy and civil liberty advocates:

- first, the Computer Assisted Passenger Pre-Screening ("CAPPS II") program now being developed by the Transportation Security Administration,[129] and

[125] *See* discussion *infra* Parts II, III; *infra* notes 238—240 and accompanying text; Rosenzweig, *supra* note 26, at 5—6 (distinguishing procedures to be used to safeguard privacy depending on type of database to be accessed and type of inquiry used); *see also* Taipale, Privacy, *supra* note 28, at slides 22—32 (discussing a "calculus of reasonableness" relating scope and method of inquiry with sensitivity of data and level of threat).

[126] Spectacular in the sense that they are rare, diverse and evolving. "Tomorrow's attacks are unlikely to look like today's." Jensen, *supra* note 37, at slide 35.

[127] *Id.*

[128] *See id.* at slide 25. Organizations can be inferred from the observation of organizational activity. *See generally* Krebs, *supra* note 122. An *ensemble classifier* uses multiple independent models to enhance descriptive or predictive accuracy. That is, by combining multiple independent models that each use different sets of relationships and have particular error rates, the composite result can achieve greater overall reliability. In the context of domestic security applications, because of the relational nature of the analysis, this may actually reduce false positives. *See* text accompanying *infra* notes 292—398. False positives flagged through a relationship with a single "terrorist identifier" will be quickly eliminated from further investigation. On the other hand, a true positive is likely to exhibit multiple relationships to a variety of independent identifiers. *See* Jensen, *supra* note 37, at slide 40.

[129] *See* Press Release, U.S. Department of Homeland Security, Transportation Security Administration, *TSA's CAPPS II Gives Equal Weight to Privacy, Security* (Mar. 11, 2003), *available at* http://www.tsa.gov/public/display?theme=44&content=535. But compare the information available from

- second, the recently terminated Terrorism Information Awareness ("TIA") project of the Defense Advanced Research Projects Agency at the Department of Defense.[130]

This Article briefly reviews these two programs – not to criticize or endorse any particular program but as paradigmatic examples which illustrate recent developments and help to understand actual practices, potentials, and concerns involving data aggregation and automated analysis. [131]

For the reasons set forth earlier, this Article accepts as a given that these technologies will be employed in some form for domestic security or other law enforcement purposes.[132] Therefore, this Article examines these two programs – one an implementation in the test phase and the other a terminated research and development project that was investigating new technologies to be employed in the future – in order to better understand the specific technical solutions being applied or proposed to meet particular legitimate security and law enforcement needs and the related privacy implications.

Electronic Privacy Information Center ("EPIC") on their "Passenger Profiling" page *available at* http://www.epic.org/privacy/airtravel/profiling.html.

[130] *See* Terrorism Information Awareness Program (TIA): System Description Document Version 1.1, *available at* http://www.taipale.org/references/tiasystemdescription.pdf (July 19, 2002) and the summary available at http://www.taipale.org/references/iaotia.pdf. These documents were previously available from the Information Awareness Office (IAO) web site at DARPA, (http://www.darpa.mil/iao/) but were removed following the defunding of the IAO and TIA. *See supra* note 28; *see also* IAO Report, *supra* note 88, §111(b).

Cf. Jay Stanley, ACLU, Is the Threat from 'Total Information Awareness' Overblown? (2002), *available at* http://www.aclu.org/Privacy/Privacy.cfm?ID=11501&c=130; Total Information Awareness, EPIC, *at* http://www.epic.org/privacy/profiling/tia/; Total Information Awareness, Electronic Frontier Foundation ("EFF") *at* http://www.eff.org/Privacy/TIA/.

Note that TIA was but one of several programs within the IAO relating to data aggregation and automated analysis and described in more detail below. The TIA program itself was the research and development effort to integrate technologies from several other DARPA programs and elsewhere (as appropriate) "to better detect, classify, and identify potential foreign terrorists." IAO Report, *supra* note 88, at 3. Many critics and news reports have conflated TIA with other, related research projects of the IAO with the result that TIA had come in popular usage to stand for an entire subset of IAO programs.

[131] For an overview of competing views of TIA, compare Stanley, *supra* note 130, with Paul Rosenzweig & Michael Scardaville, Heritage Foundation, The Need to Protect Civil Liberties While Combating Terrorism: Legal Principles and the Total Information Awareness Program (2003), *available at* http://www.heritage.org/Research/HomelandDefense/lm6.cfm *and* Michael Scardaville, Heritage Foundation, No Orwellian Scheme Behind DARPA's Total Information Awareness System (2002), *available at* http://www.heritage.org/Research/HomelandDefense/wm175.cfm. *See also* Dan Farmer & Charles C. Mann, *Surveillance Nation*, MIT Tech. Rev. 34, 46 (2003); Simon Garfinkel, *Database Nation* (2001).

[132] *See supra* text accompanying notes 37—39, 48 (discussing efficiency and the need to manage increased data volumes); *supra* note 41, 43 (detailing a variety of current applications).

In any case, information sharing and automated analysis are already official government policy. *See supra* Prelude.

A. *CAPPS II: An Overview*

CAPPS II is the second-generation[133] automated airline passenger screening system currently being developed by the Transportation Safety Agency ("TSA") of the Department of Homeland Security.[134] The CAPPS II program is being implemented under authority granted by Congress following the September 11, 2001 terrorist attacks.[135]

The stated purpose of CAPPS II is to (1) authenticate identity, and (2) perform a terrorist risk assessment of airline passengers prior to airport screening.[136] Under CAPPS II, the TSA will receive from airlines at the time a reservation is made the passengers full name, address, date of birth, and phone number.[137] TSA will then query commercial

[133] CAPPS II is a follow on to CAPPS the original passenger screening program developed in the late 1990's. See § 307 of the Federal Aviation Reauthorization Act of 1996, Pub. L. No. 104-264, directing the FAA to continue to assist airlines in developing a computer-assisted passenger profiling system. *See infra* note 134.

CAPPS II is intended to address some of the short-comings of CAPPS I. Among other changes, CAPPS II is to be administered directly by TSA whereas CAPPS I was administered by the airlines themselves. *See generally* Joe Sharkey, *A Safer Sky or Welcome to Flight 1984?*, N.Y. Times, Mar. 11, 2003, *available at* http://www.nytimes.com/2003/03/11/business/11ROAD.html; Mathew L. Wald, *U.S. Agency Scales Back Data Required on Air Travel*, N.Y. Times, July 31, 2003.

[134] The TSA Office of National Risk Assessment began testing the CAPPS II system at three undisclosed airports in March 2003. *See* Leslie Miller, *Feds Testing Air-Passenger Check System*, Associated Press, Feb. 28, 2003, *available at* http://www.govtech.net/news/news.phtml?docid=2003.02.28-42025. However, in response to the public reaction and comments to the original program announcement the TSA has scaled back public testing. *See* 68 Fed. Reg. 45265-45269 (Aug. 1, 2003); Michael Delio, *CAPPS II Testing on Back Burner*, Wired News, June 13, 2003 ("[A] TSA spokesman confirmed Friday that the agency has decided to delay further public testing of the CAPPS II until a privacy policy . . . can be crafted."), *available at* http://www.wired.com/news/privacy/0,1848,59252,00.html.

[135] Under § 136 of the Aviation and Transportation Security Act, signed into law on November 19, 2001, Congress directed the Secretary of Transportation to ensure that CAPPS "or any successor system, (i) is used to evaluate all passengers before they board an aircraft; and (ii) includes procedures to ensure that individuals selected by the system . . . are adequately screened." 49 U.S.C. § 44903(j)(2)(A) (2003).

However, the Department of Homeland Security Appropriations Bill, 2004, *supra* note 28, enacted October 1, 2003, contains language purporting to restrict funding for CAPPS II until the Government Accounting Office reports to Congress that the CAPPS II program meets certain criteria set forth in the bill. President Bush has declared in a separate Signing Statement, *supra* note 28, that such restrictive language is ineffective under the doctrine set forth in *Chadha*, *supra* note 28.

[136] TSA, *supra* note 129.

[137] Originally the TSA characterized this data as information that airlines routinely collect as part of the Passenger Name Record (PNR) in the normal course of making a reservation, however, privacy advocates have pointed out that airlines do not currently collect that information (for example, the PRN routinely contains the travel agent's address and phone number by default, not the passenger's, and date of birth information has never been collected). *See, e.g.*, Edward Hasbrouck, *Total Travel Information Awareness*, *at* http://www.hasbrouck.org/articles/travelprivacy.html (last visited Dec. 15, 2003).

databases to verify identity and government databases to generate a "threat score."[138] Threat scores are to be color classified as green (allowing for standard airport security procedures), yellow (subjecting the passenger to heightened airport scrutiny), and red (resulting in referral to law enforcement).[139]

Despite widespread criticism of the CAPPS II program as a vast "data mining" exercise,[140] it should be clear to the reader of this Article that CAPPS II is a data aggregation program involving a traditional subject-based query to one or more commercial and government databases. That is, the TSA intends to use a standard query method to interrogate external databases to verify identity and internal databases to make a threat assessment for a particular specified individual. In a strict technical sense, CAPPS II does not involve data mining but instead involves data matching against a watch list or data aggregation to confirm identity. The CAPPS II system does not itself profile, conduct surveillance, or data mine.[141]

Further, according to the Director of TSA, CAPPS II will only access information to which government is already legally entitled and it will do so in conformity with the Privacy Act of 1974.[142] Additionally, TSA does not intend to retain any information generated or returned by the queries beyond the termination of the passenger flight.[143]

Nevertheless, CAPPS II raises privacy concerns. Data aggregation itself (regardless of whether the data is amassed in a central database or through virtual integration) creates its own privacy concerns, but those concerns are distinct from those

[138] TSA, *supra* note 129.

[139] *Id.*

[140] *See* Press Release, ACLU, CAPPS II Data-Mining System Will Invade Privacy, ACLU Warns (Feb. 27, 2003), *at* http://www.aclu.org/Privacy/Privacy.cfm?ID=11956&c=130; Roy Mark, *TSA Books Data Mining Program*, Internetnews, Mar. 4, 2003, *available at* http://dc.internet.com/news/article.php/2013781; EFFector, *CAPPS II on the Defensive?* ("CAPPS II is yet another government anti-terrorist data-mining program that would try to analyze public and private databases in search of terrorist activity patterns."), *available at* http://www.eff.org/effector/HTML/effect16.07.html (Mar. 14, 2003); *see also* Deborah Pierce, Opinion, *Law and Technology: CAPPS II*, Seattle Press, Mar. 11, 2003 (calling CAPPS II a "vast surveillance system" that should be thought of as "TIA's slightly smaller brother"), *available at* http://www.seattlepress.com/article-10116.html.

[141] *See* Ben H. Bell, Director of the Office of National Risk Assessment, TSA, Presentation: Congressional Briefing on the Office of National Risk Assessment (ONRA) and the Computer Assisted Passenger Pre-Screening (CAPPS II) Program (Mar. 7, 2003).

Note that identification verification and threat assessment each could themselves be based on the results of a "data-mined" pattern, a human developed "terrorist profile," or a previous determination based on other information. However, the use of data mining (or any heuristically derived pattern) is not a requirement of the CAPPS II program.

[142] *See* Admiral Jim Loy, Administrator, Transportation Security Administration, Remarks as Prepared for Delivery at the Privacy and American Business Conference (Mar. 13, 2003). For purposes of this Article, these government statements are accepted at face value. *But see* Hasbrouck, *supra* note 137. Also, despite this assurance, TSA has sought a Privacy Act exemption for the CAPPS II program. *See infra* note 223.

[143] *Id.; see also* TSA, *supra* note 129; Wald, *supra* note 133.

created by data mining (or pattern-matching).[144] From a technology point of view, CAPPS II is a data integration (or data matching) project, not a data mining project.[145] The policy issues involved are determining what information can or should be gathered from passengers prior to flight and which databases TSA can or should have access to in order to authenticate identity and assess potential threats (and, what criteria or methodology is used to make those assessments). Nevertheless, the program does not require the use of "new" technologies or techniques. The question is simply whether the TSA can systematically and indiscriminately automate a process that it probably has the legal authority to engage in on an individualized basis.[146] To the extent that maintaining certain government inefficiencies helps protect individual rights from centralized state power, the question involved in CAPPS II is one of increased government efficiency.[147]

B. *Terrorism Information Awareness: An Overview*

Unlike CAPPS II, which intends to query databases with respect to an identified individual to either confirm data (identity) or assess threat level, TIA was a research and development project specifically intended to develop and integrate data mining (and other

[144] The privacy issues involved in data aggregation and data mining, including the distinction between database privacy and pattern-matching (or non-particularized suspicion) are discussed in Part III *infra*. With respect to privacy issues arising from data aggregation, see *infra* text accompanying notes 244—251.

In response to public criticism, TSA is reviewing CAPPS II for privacy concerns and has announced that details of the program, including what the system will do and what databases it will access, will be disclosed through public notice in the Federal Register before any future public testing or implementation. *See* Wald, *supra* note 133; Delio, *supra* note 134.

Beyond questioning the details of the program, some privacy advocates have raised the threshold question of whether it is appropriate – even if the government has legal access to the databases – for any such individualized scrutiny to be triggered by the making of a reservation for airline travel. Aspects of that question are currently being litigated. *See* Gilmore v. Ashcroft, No. 02-3444 (N.D. Ca. 2002) (challenging the legality of requiring ID for air travel generally), information *available at* http://cryptome.org/freetotravel.htm.

[145] Thus, while criticism of the program may be justified on other grounds, opposing it as a vast data mining program is misguided and inappropriate. *See* SIGKDD of the ACM, *supra* note 28 ("[R]ecent public portrayals . . . appear to create the misguided impression that these programs are synonymous with data mining technology, and that data mining technology is potentially a major threat to civil liberties.").

[146] Whether TSA has the legal authority to require identification prior to travel is being litigated in *Gilmore v. Ashcroft*. *See supra* note 144. Some form of passenger prescreening is clearly authorized under 49 U.S.C. §44903(j)(2)(A), as discussed *supra* in note 135. But see the restrictions set out in the Department of Homeland Security Appropriations Bill, 2004, enacted October 1, 2003, discussed *supra* in note 28.

[147] *See infra* text accompanying notes 244—247 (discussing government inefficiency as protective of individual liberty).

related) technologies[148] in order to enable appropriate government agencies to prospect among large data volumes, from both government and commercial databases, and to find patterns and relationships that may help identify terrorists before they strike.[149]

Although TIA was attacked as an attempt to build "dossiers" on 300 million U.S. citizens[150] to track their everyday activities,[151] or to build a "supercomputer" to surveil the entire population,[152] it was actually something quite different and far less ominous, despite its unfortunate initial choice of name, logo, and motto.[153] According to the IAO documents, TIA was an experimental prototype system that consists of three parts – language translation technologies, data search and pattern recognition technologies, and advanced collaborative and decision support tools:

> Together, these three parts comprise the Total [sic] Information Awareness (TIA) project.
>
> The language translation technologies will enable the rapid translation of foreign language publications and give intelligence analysts the capability to quickly search for clues about emerging terrorist acts. The intelligence community believes it can find evidence of terrorist activities in open source foreign language publications. Rapid translation technologies will help intelligence analysts search a significant amount of material in a much shorter period than is possible today.
>
> The research into data search and pattern recognition technologies is based on the idea that terrorist planning activities or a likely terrorist attack could be uncovered by searching for indications of terrorist activities in vast

[148] For a description of the various technology projects that were within IAO related to the TIA program, see IAO Report, *supra* note 88, app. A. *See also* text accompanying notes 168—187 *infra* (describing TIA and certain related programs involving data aggregation and data mining).

[149] TIA was actually a broad research program encompassing several dozen different subprograms to research and develop "tools to better detect, classify, and identify potential foreign terrorists. TIA's research and development goal was to increase the probability that authorized agencies of the United States can preempt adverse action." IAO Report, *supra* note 88, at 3.

[150] *See* Safire, *supra* note 20.

[151] *See, e.g.*, Audrey Hudson, *Homeland bill a supersnoop's dream*, Wash. Times, Nov. 15, 2002 (TIA "will allow the federal government to track the email, Internet use, travel, credit-card purchases, phone and bank records of foreigners and U.S. citizens").

[152] *See, e.g.*, Cheri Preston, *Big Brother or Terror Catcher?*, ABC News ("It's called Total Information Awareness, a supercomputer can that [sic] keep track of . . ."), *available at* http://abcnews.go.com/sections/scitech/DailyNews/cybershake030124.html (Jan. 24, 2003).

[153] The program was initially called the "Total" Information Awareness program and the original logo for the IAO featured the Illuminati all-seeing eye symbol (also present on the one dollar bill) gazing on the entire world with the motto "Scientia Est Potentia," Latin for "Knowledge is Power" featured prominently. Needless to say, this was a public relations error. *See, e.g,* Hendrik Hertzberg, *Too Much Information*, New Yorker, Dec. 9, 2002, at 45; *see also* note 34 *infra*.

> quantities of transaction data. Terrorists must engage in certain transactions to coordinate and conduct attacks against Americans, and these transactions form patterns that may be detectable. Initial thoughts are to connect these transactions (e.g., applications for passports, visas, work permits, and drivers' licenses; automotive rentals; and purchases of airline ticket and chemicals) with events, such as arrests or suspicious activities. For this research, the TIA project will use only data that is legally available and obtainable by the U.S. Government.
>
> A major challenge to terrorist detection today is the inability to quickly search and correlate data from the many databases maintained legally by our intelligence, counterintelligence, and law enforcement agencies. The collaborative reasoning and decision-support technologies will help solve existing coordination problems by enabling analysts from one agency to effectively collaborate with analysts in other agencies.[154]

Despite these measured assurances from government managers, other descriptions of the project, also from DARPA's own IAO web site, suggest a more centralized master database likely to cause increased concern among privacy advocates. For example:

> Technically, the TIA program is focusing on the development of: 1) architectures for a large-scale counter-terrorism database, for system elements associated with database population, and for integrating algorithms and mixed-initiative analytical tools; 2) novel methods for populating the database from existing sources, create innovative new sources, and invent new algorithms for mining, combining, and refining information for subsequent inclusion into the database; and, 3) revolutionary new models, algorithms, methods, tools, and techniques for analyzing and correlating information in the database to derive actionable intelligence.[155]

Nevertheless, for purposes of this Article, either description is adequate since the primary intent is clear. As stated in a report by the ACLU, "[t]he overall concept is clear. 'The purpose of TIA would be to determine the feasibility of searching vast quantities of data to determine links and patterns indicative of terrorist activities,' as Under Secretary of Defense Edward C. "Pete" Aldridge put it."[156]

[154] *See* DARPA, *Information Awareness Office and Terrorism Information Awareness Project*, *available at* http://www.taipale.org/references/iaotia.pdf. This document was previously available from DARPA *at* http://www.darpa.mil/iao/ioatia.pdf.

[155] This paragraph was previously available at http://www.darpa.mil/iao/TIASystems.htm (last visited Sept. 2003). However, it has subsequently been removed.

[156] Stanley, *supra* note 130.

1. Massive Database or Distributed Architecture?

The technical differences between aggregating data from disparate sources into a centralized database (the traditional "data warehouse" approach) or integrating data sources by accessing individual distributed databases (conceptually, a "virtual" warehouse) has technical, security, and privacy implications.

First, there is a practical inefficiency in centralizing data. Since data is collected locally for local uses, the procedures for collecting, verifying, and updating information are also local. Thus, any effort to centralize data would likely fail because of an inability to keep the centralized database up-to-date and useful for its primary local purpose, unless, of course, all local storage and management were eliminated and data were only entered and stored in the single central database. Obviously local database owners, whether government agencies or commercial entities, are not going to cede control and management over their database architecture and data methodologies – which are being maintained to meet their own particular local needs – to a central authority. Nor could a central authority develop, maintain, or manage a single database architecture that satisfied such diverse local needs.[157]

Second, security concerns for a central database are different than for a system of distributed databases. For example, the central database represents a single target for potential attack or source for other abuse. On the other hand, a distributed system, in which individual databases remain in the control of their local hosts, runs the risk of exposing the database queries themselves to many sources, either alerting suspects or invading privacy when it becomes known that a particular subject was queried.

Third, a distributed database architecture supports privacy protection by diversifying control. A centralized database puts control of privacy and access rules in a single government agency, whereas a distributed system maintains local control and, to some extent, local accountability. Local access control and individual privacy rules can be enforced and tracked at many points in the system and there is no single point of control to be exploited either by attack or by misuse.[158] Additionally, a distributed system provides multiple audit trails under independent control.

[157] This raises another interesting issue: the data aggregation need for domestic security applications may be much narrower than the need for local data collection. "Another persistent myth is that data mining for counter-terrorism requires [more] data collection. . . . [However,] some techniques developed by data mining researchers could actually reduce data collection." Jensen, *supra* note 37, at slide 19; *see also* Jensen, *Technical Assessment*, *supra* note 89, at § 2 (proposing an enhanced implementation model using iterative data searches that shows "that substantially smaller amounts of data can be accessed in later stages of inference."). The point here is that data mining itself (including active learning approaches) focuses attention on what data is most useful and what data is irrelevant for the particular purpose, in this case, finding terrorists. Thus, effective use of data mining is likely to focus investigators' attention on narrower, relevant datasets than to lead to uncontrolled data gathering exercises. The point of these technologies is to make the haystacks smaller and more relevant, not increase their size.

[158] *See* Jensen, *supra* note 37, at slide 18 ("This approach keeps institutional control of databases distributed, providing a bulwark against both outside intruders and widespread institutional misuse."); *see also* note 89 *supra* (suggesting that this approach provides more intervention points for legal or political control).

For these reasons, as well as for current and historical political reasons, it is unlikely that any system consisting of a single, centralized massive database will (or should) be implemented.[159]

Further, ongoing research and development in database applications is primarily focused on distributed architectures. Technologies to standardize structured data (such as XML)[160] and to allow direct analysis of unstructured data (such as semantic analysis and natural language applications),[161] as well as improvements in statistical methods, machine

[159] As discussed throughout this Article, the strong negative reaction to TIA engendered by early descriptions of a massive centralized government database containing "all information about everyone" makes it unlikely that the political will to support such an effort could be mustered. Further, there is an historical precedent for resisting this approach. *See* Garfinkel, *supra* note 131, at 13—15. In 1965, the Bureau of the Budget proposed that instead of each federal agency investing in its own computers, storage technology, and operations personnel, that the US government should instead build a single National Data Center. While the original motivation was to cut costs, other benefits became clear and were publicly discussed. *Id.* The Princeton Institute for Advanced Study enthusiastically supported the project as improving security of data and privacy. A central database would also make the data more accurate, it was argued, allowing for redundant and incorrect records to be purged quickly and efficiently ("cleansed"). Legal protection for privacy would also be streamlined with one federal law that would protect the records, provide for privacy, and promote accountability of the database managers. *Id.*

After an initial enthusiastic reception by the media and the public, an attack on the idea of a centralized database appeared in the New York Times Magazine that began a series of Congressional hearings and public opposition. By 1968, the Bureau of the Budget backed off of the idea, and the National Data Center was never built. *Id.;.see also* Note*, Privacy and Efficient Government: Proposals for a National Data Center*, 82 Harv. L. Rev. 400 (1968).

Thus, prior to the development of wide area network technologies, and particularly the Internet, much privacy protection in the U.S. was actually achieved through the inefficiency of storing information in many, unconnected databases. An unfortunate side effect that was not anticipated is that the political need for a strong federal privacy law never materialized. This has led to the current situation in which network technologies have suddenly overtaken the available legal protections by creating a single "virtual" data set from multiple, distributed databases all subject, if at all, to different and sometimes conflicting regulation.

New calls for centralization are unlikely to occur since such a development would run counter to current information technology development trends and information theories that are geared towards exploiting distributed data architectures. In any case, network technologies obviate the need.

Indeed, the Second Markle Report, *supra* note 28, calls for the creation of a trusted, decentralized network. The Report identifies weaknesses of a centralized system as: (1) susceptibility to a single point of failure or attack, (2) designed to flow information up to leadership not down to operational entities, and (3) not supportive of real-time operations (because, among other things, the difficulty of maintaining centralized data up-to-date). *See id.* at 14, Ex. B. Instead, the Report proposes a distributed peer-to-peer network architecture in which local data repositories are accessed through a common data layer kept independent from applications, and uses directory- and web-services to manage and access data. *Id.* at 13—17 and Ex. D.

[160] *See, e.g.*, Mark Songini, *IBM Pushes Out Virtual Database Technology*, Computerworld, Feb. 6, 2003, *available at* http://www.computerworld.com/databasetopics/data/software/story/0,10801,78258,00.html; Lisa Vaas, *Virtual Databases Make Sense of Varied Data*, eWeek, Mar. 31, 2003, *available at* http://www.eweek.com/article2/0,3959,981992,00.asp (discussing XML-based virtual database technologies to enable access over heterogeneous databases and information sources).

[161] *See, e.g.*, Cathleen Moore*, Search Tools Look for Context*, InfoWorld, Feb. 28, 2003, *available at* http://www.infoworld.com/article/03/02/28/HNsearch_1.html (discussing context based search); Darryl K. Kraft, *IBM Takes Search to New Heights*, eWeek, Aug. 11, 2003, *available at*

learning and artificial intelligence methods for processing non-standard data will continue this trend, further obviating the need for centralization of data.[162] In addition, recent data mining research using sophisticated models premised on iterative, multi-pass pattern-matching against distributed data sources further undermines any need for centralization.[163]

Although the distinction between a massive centralized database and virtual aggregation through a distributed database architectures is often dismissed as irrelevant for privacy concerns,[164] it is in fact fundamental for developing legal procedures and mechanisms that take advantage of the distinct characteristics of distributed architectures in order to apply existing due process and other protections to the use of these technologies. By conceiving an architecture (and implementation strategy) based on accessing primary datasets for analysis and secondary datasets for investigation,[165] procedural rules that have proven effective in the "real world" can be applied to the use of these new technologies. By developing rule-based processing technologies – technologies that label individual data items with attributes that specify how it can be accessed and according to what rules – a distributed architecture can significantly protect privacy interests.[166] Further, current data mining research has shown that enhanced application models premised on iterative, successive access to distributed data may actually lower the amount of data ultimately required to be analyzed to achieve comparable results thus resulting in less privacy intrusion in the first place.[167]

http://www.eweek.com/print_article/0,3668,a=46085,00.asp (discussing IBM's use of artificial intelligence techniques to search unstructured data).

[162] *See* Jensen, *supra* note 37, at slide 17 (commenting that "exclusive access to data . . . may be declining in importance."). The implication being that analysis of available disparate data is increasing in importance relative to exclusive or central control over a database.

[163] *See* Jensen, Technical Assessment, *supra* note 90, at §§ 2.2, 6.

[164] *See, e.g.*, Electronic Frontier Foundation, EFF Review of May 20 Report On Total Information Awareness, Executive Summary ("does it really matter that there is no 'real' centralized database?"), *available at* http://www.eff.org/Privacy/TIA/20030523_tia_report_review.php (last visited Dec. 15, 2003); ACLU, Total Information Compliance: The TIA's burden under the Wyden Amendment, at 6 (2003), *available at* http://www.aclu.org/SafeandFree/SafeandFree.cfm?ID=12650&c=206 (that the system uses a distributed database architecture "is a distinction without a difference" that "makes no difference for privacy").

[165] *See supra* note 89.

[166] There are actually two aspects of rule-based processing. First, intelligent agents that negotiate access and processing directly with a distributed database before receiving a result, and second, data labeling (meta-data, that is, data about the data) that accompanies the data item when it is returned in response to the query. Meta-data, including rules for processing or privacy protection, can be based on the particular data item itself, or by classification or source, etc. So, for example, a data item might be tagged as belonging to a U.S. or a non-U.S. person, resulting in different standards for processing depending on its designation. *See, e.g.*, Potomac Institute for Policy Studies, *Oversight of Terrorist Threat Information*, *supra* note 28. Or, a data item might be tagged as coming from a medical database, thus requiring different handling and legal compliance than a data item from a state driver's license database, etc.

[167] *See, e.g.*, Jensen, Technical Assessment, *supra* note 90, at § 3.3:

2. TIA and Data Mining

The TIA program was one of several related programs in the IAO.[168] Many of these programs had little privacy concern, for example, efforts to develop capabilities for machine translation that would allow for the access of information regardless of language.[169] Others may have raised privacy concerns but are outside the scope of this Article, for example, programs to develop face or gait recognition technologies.[170] The TIA program itself was the "systems-level" program of the IAO that "aim[ed] to integrate information technologies into a prototype to provide tools to better detect, classify, and identify potential foreign terrorists [with the goal] to increase the probability that authorized agencies of the United States [could] preempt adverse actions."[171] As a systems-level program, "TIA [was] a program of programs whose goal [was] the creation of a counterterrorism information architecture" by integrating technologies from other IAO programs (and elsewhere, as appropriate).[172]

Among the other IAO programs that were intended to provide TIA with component data aggregation and automated analysis technologies were the Genisys program, Genisys Privacy Protection, Evidence Extraction and Link Discovery, and Scalable Social Network Analysis.

a. Genisys – The Genisys program was aimed at developing technologies for virtual data aggregation in order to support effective analysis across heterogeneous databases as well as unstructured public data sources, such as the World Wide Web.[173] Genisys:

> aim[ed] to create technology that enable many physically disparate heterogeneous databases to be queried as if it [sic] were one logical

> [A] system could access different amounts or types of data on each pass. For example, assume that an initial prediction could be made based on relatively innocuous data (i.e., data which are not considered highly sensitive). Then those initial results could be used to limit the types of more sensitive data examined in subsequent passes, either resulting in fewer data items being accessed per object or in the same data items being accessed for fewer individuals. . . . [W]e examine how multi-pass inference can be used to achieve higher accuracy with lower data utilization rates than single-pass inference.

[168] *See* IAO Report, *supra* note 88 (providing a detailed description of the component programs of IAO).

[169] *See id.* at B-9.

[170] *See id.* at A-18, A-22.

[171] *See id.* at 3.

[172] *Id.*

[173] *See* IAO Report, *supra* note 88, at A-10.

> "virtually" centralized database. The technology, mediation, refers to the use of intelligent software agents that would relieve analysts from having to know all the details about how to access different databases, the precise definitions of terms, the internal structure of the database and how to join information from different sources, and how to optimize queries for performance. . . . They would be able to use all the databases to which they have access as a federation – a new 'megadatabase' would not be created. Information from other sources such as the web or semi-automated collection systems would be somewhat easier to convert to structured data and that would help TIA increase its coverage.[174]

In addition, Genisys aimed to develop "simulation and pattern-matching technology" that could extract "terrorist signatures from a world of noise."[175]

b. Genisys Privacy Protection – This program sought to research and develop technology:

> to ensure personal privacy and protect sensitive intelligence sources and methods in the context of increasing use of data analysis for detecting, identifying and tracking terrorist threats. . . .[This program would develop] technologies that enable greater access to data for security reasons while protecting privacy by providing critical data to analysts while not allowing access to unauthorized information, focusing on anonymized transaction data and exposing identity only if evidence warrants and appropriate authorization is obtained for further investigation, and ensuring that any misuse of data can be detected and addressed.[176]

In other words, the Genisys Privacy Protection program was seeking to develop the technologies that this Article advocates: rule-based processing, selective revelation and strong audit.[177]

c. Evidence Extraction and Link Discovery – The objective of this program was "to develop technology for 'connecting the dots' – starting from suspicious activities noted in intelligence reports."[178] The technologies developed under this program would support "subject-based" queries, in which a particular subject – person, place, or thing – is the

[174] *Id.* at A-11.

[175] *Id*.

[176] *Id*. at A-12.

[177] *See infra* Part IV.

[178] IAO Report, *supra* note 88, at A-14.

starting place for investigation, and the technology automates the process of finding key relationships or associations based on that subject.

It should be noted that DARPA drew a particular distinction in describing this technology from the more general data mining technologies as "currently understood in the commercial sector." According to DARPA, existing "commercial data-mining techniques are focused at finding broadly occurring patterns in large databases, in contrast to intelligence analysis that consists largely of following a narrow trail and building connections from initial reports of suspicious activity."[179]

Commercial data mining techniques are generally applied against large transaction databases in order to classify people according to transaction characteristics and extract patterns of widespread applicability. The problem in counter-terrorism is to focus on a smaller number of subjects within a large background population and identify links and relationships from a far wider variety of activities.[180]

Commercial data mining is focused on classifying propositional data from homogeneous databases,[181] while domestic security applications seek to detect rare but significant relational links between heterogeneous data.[182] In general, commercial users have been concerned with identifying patterns among unrelated subjects based on their transactions in order to make predictions about other unrelated subjects doing the same. Intelligence analysts are interested in identifying patterns that evidence organization or activity among related subjects in order to expose additional related subjects or activities.

Thus, in this program, DARPA was seeking to develop technologies to extract evidence of terrorist activity (Evidence Extraction), discover links from that evidence (Link Discovery), and develop models from those links to focus resources on particular patterns that may identify terrorist activity (Pattern Learning).[183] As described above in Part I, the distinction (and related challenges) between relational and propositional data mining is well known in knowledge discovery research. [184]

[179] *Id*. But note that domain specific (i.e., developed for law enforcement) technologies, such as CopLink described *supra* note 42, are specifically aimed at extracting links from subject-oriented queries. Additionally, research in both commercial and scientific database architectures are increasingly focused on supporting relational analysis. For example, a database architecture used for particle physics applications emphasizes "objects" that "inherit their properties from parent objects in a vast family tree [and] can bring important connections to light more efficiently than traditional [database methods]," and is attracting attention for national security and intelligence uses. *See* Wade Roush, *Managing Antiterror Databases*, MIT Tech. Rev., June 2003.

[180] IAO Report, *supra* note 88, at A-14.

[181] That is, patterns in a database of like-transactions, for example, book sales.

[182] That is, links among a wide variety of different subjects, activities, transactions, and associations.

[183] IAO Report, *supra* note 88, at A-15.

[184] *See* Jensen, *supra* note 37, at slides 21—33, distinguishing between propositional and relational data mining and calling for additional research to develop:

> [a] new synthesis of first-order (or high-order) logics and probabilistic reasoning, so that we can construct and use statistical models of relational data. In the past two decades, some work in inductive logic programming, social network analysis, and Bayesian networks has produced some good foundations, but years of additional work is [sic]

d. Scalable Social Network Analysis – This program was seeking to develop "techniques based on social network analysis for modeling the key characteristics of terrorist groups and discriminating these groups from other types of societal groups."[185] Preliminary *post hoc* analysis of the 9/11 hijackers "showed how several social network analysis metrics changed significantly in the time immediately prior to September 11; this change [sic] could have indicated that an attack was imminent."[186] This program intended to "develop algorithms and data structures for analyzing and visualizing the social network linkages."[187]

3. TIA related programs: Summary

The TIA program intended to research, test and develop technologies "that [automate] many lower level functions that can be done by machines guided by human users [to give] the users more time for the higher analysis functions which [sic] require the human's ability to think."[188] "In today's world, the amount of information that needs to be considered far exceeds the capacity of the unaided humans in the system."[189]

To do so, the TIA program researched, developed, and integrated technologies to virtually aggregate data, to follow subject-oriented link analysis, to develop descriptive and predictive models through data mining or human hypothesis, and to apply such models to additional datasets to identify terrorists and terrorist groups.[190]

4. TIA related programs: Epilogue – A Pyrrhic "Victory"

As previously noted, the Department of Defense Appropriations Bill signed into law by President Bush on October 1, 2003 contained provisions prohibiting the use of funds for TIA or any "successor program," thus effectively eliminating the IAO at

> needed before we will have tools that match the level of development of current commercial tools for propositional data.

Id. at slide 27.

[185] IAO Report, *supra* note 88, at A-16.

[186] *Id.*; *see also supra* note 120 (discussing the analysis of email traffic to identify organization and leadership roles); *cf supra* note 122.

[187] IAO Report, *supra* note 88, at A-16.

[188] *Id.* at A-2.

[189] *Id.* at A-1.

[190] See *id.* at A-1—A-6.

DARPA.[191] At first hailed as a "victory" for civil liberties,[192] it has become increasingly apparent that the defunding is likely to be a pyrrhic victory. As argued earlier, not proceeding with a focused government research and development project (in which Congressional oversight and a public debate could determine appropriate rules and procedures for use of these technologies and, importantly, ensure the development of privacy protecting technical features to support such policies) is likely to result in little security and, ultimately, brittle privacy protection.

Indeed, following the demise of IAO and TIA, it has become clear that similar data aggregation and automated analysis projects exist throughout various agencies and departments not subject to easy review.[193] Further, many of the projects formerly under the IAO have been moved directly to intelligence agencies and other classified programs, including the Army Intelligence and Security Command ("INSCOM").[194] The very legislation that eliminated IAO funding itself contains a classified annex purportedly detailing how such technologies can be developed and used in foreign intelligence. Yet, President Bush has stated in a Signing Statement[195] that such classified annex would not be considered part of the signed act, "which means that anything mentioned in the annex is not subject to the data mining restriction."[196]

It is my firm belief that defunding IAO and shutting down TIA was an ill-conceived and short-sighted action that effectively eliminated the most visible and accountable government program in which appropriate privacy-protecting procedures and policies could be developed.[197] In particular, eliminating the Genisys Privacy Protection

[191] *See supra* note 28; *see also* Carl Hulse, *Congress Shuts Pentagon Unit Over Privacy*, N.Y. Times, Sept. 26, 2003, at A20 ("A Pentagon office that became steeped in controversy over privacy issues . . . was shut down by Congress today as the Senate passed and sent to President Bush a $368 billion military measure that eliminates money for it."). The Joint Explanatory Statement included in the conference committee report on H.R. 2658, *supra* note 28, specifically directed that the IAO be closed quickly. "The conferees are concerned about the activities of the Information Awareness Office and direct that the Office be terminated immediately." *Id.*

[192] "From a standpoint of civil liberties, this is a huge victory." Senator Wyden, *quoted in* Hulse, *supra* note 191.

[193] *See, e.g.*, New, *supra* note 44.

[194] *Id.*

[195] *See supra* note 28.

[196] *See* New, *supra* note 44. See also H.R. 2417, *supra* note 28, in which data mining for foreign intelligence is specifically authorized. The former TIA projects Genisys and Genoa II are believed to be included in the classified annex to the Defense Appropriations Bill, *supra* note 28. *See, e.g.,* Major General Paul Nielson, *Protection of Personal Privacy in the Development of Advanced Technologies to Fight Terrorism*, Air Force Oral Statement to the TAPAC 6 (Nov. 20, 2003) ("[Genisys and Genoa II] may be affected by the classified annex of the Defense Appropriation Bill."), *available at* http://www.sainc.com/tapac/library/Nov2003/PaulNielsen-Statement.pdf.

[197] This view is apparently shared by several privacy advocates whose efforts helped eliminate the program. For example, David Sobel of the Electronic Privacy Information Center was quoted in New, *supra* note 44, as saying that although closing TIA was a "very important action," it now might be harder to detect such activity because "[t]o some extent [data aggregation and automated analysis programs have] gone underground." James Dempsey, Executive Director of the Center for Democracy and Technology, is

project may severely impact the effectiveness of any future privacy protecting implementation or oversight policies because the required technical features to support such policy may never be developed.[198]

Part III. Data Aggregation and Analysis: Privacy Concerns

This section discusses certain privacy issues as they relate to the use of data aggregation and automated analysis technologies for domestic security or law enforcement purposes.

A. *Privacy Concerns: An Overview*

News headlines remind us daily that we are in imminent danger of "losing our privacy."[199] And commentators on the left and the right, with little else in common, agree.[200] But privacy means different things to different people.[201]

It is beyond the scope of this Article to definitely define privacy or reconcile competing views. Rather, this Article seeks to distill from the debate certain core principles and examine how they intersect with technology developments relating to data aggregation and data mining in the context of domestic security in order to understand how technological development and implementation strategies can be designed to help ameliorate privacy concerns. The aim is not to critique or suggest particular legal frameworks or new laws, nor to criticize specific structures or programs, but to try to identify core privacy concerns that might be addressed through the application of certain guiding principles both to the development of these technologies and to the policies governing their use.[202]

To that end, "it is important to distinguish between the concept of privacy and the right of privacy. . . . Privacy as a concept involves what privacy entails and how it is to be valued. Privacy as a right involves the extent to which privacy is (and should be) legally protected."[203] Thus, this Article is not so much concerned with how these particular

reported as being "disappointed that Congress cut all funding for TIA because that meant such research could migrate to less accountable agencies." Drew Clark, *Privacy: Business Records Called An Open Door To Data Mining*, Nat. J.'s Tech. Daily, Oct. 23, 2003. *See also infra* note 341.

[198] *See infra* Part IV discussing technical features that can support privacy policy.

[199] *See, e.g.*, Editorial, *America's Number One Fear In The 21st Century Is Loss Of Personal Privacy*, St. Petersburg Times, Nov. 3, 1999 at 18A.

[200] *See, e.g.*, Stanley & Steinhardt, *supra* note 22; Safire, *supra* note 20.

[201] "Privacy is a value so complex, so entangled in competing and contradictory dimensions, so engorged with various and distinct meanings, that I sometimes despair whether it can be usefully addressed at all." Robert C. Post, *Three Concepts of Privacy*, 89 Geo. L.J. 2087, 2087 (2001).

[202] *See infra* Part IV and Conclusion; *supra* Introduction.

[203] Daniel J. Solove & Marc Rotenberg, Information Privacy Law 25 (2003).

technologies fit within the current legal structure, but rather how they may impact core privacy interests and whether procedural and technical features can be developed to lessen their impact.[204]

Privacy and security are often seen as being in tension and requiring balance. The metaphor of balance suggests that when new developments create disequilibrium, and a new fulcrum needs to be sought to restore balance. Privacy advocates believe that new information technologies have upset the balance and that stronger protection for individual privacy is the required response – they would move the fulcrum towards "more privacy." Those concerned with security argue that global terrorism has upset the balance and that new methods and technologies, even if they impact on historic privacy concerns, need to be employed – they would move the balance towards "more security." The metaphor of balance has led to the dichotomous ideological choice being presented as privacy *or* security.[205]

However, privacy and security might be more properly considered dual obligations – to protect civil security *and* to preserve civil liberty.[206] Then, achieving security while also protecting privacy means recognizing "that security need not be traded off for liberty in equal measure and that the 'balance' between them is not a zero-sum game."[207]

This Article argues that achieving security while protecting privacy requires building technology structures (and related implementation strategies) that reflect both values.

1. Information Privacy Law

Although privacy interests are said to be as "old as civilization" itself,[208] the modern notion of a legally protected privacy right is a relatively recent, and illusive,

[204] Thus, the question is not so much whether the government is only accessing data to which it is already entitled, but whether there is an unreasonable incremental invasion of a privacy interest as a result of applying automated aggregation and analysis technologies to the data.

[205] *Cf.* Etzioni, The Limits of Privacy 184 (1999) ("Once we accept the concept of balance, the question arises as to how we are to determine whether our polity is off balance and in what direction it needs to move, and to what extent, to restore balance.").

[206] *See* Powers, *supra* note 46, at 5, 21 ("In a liberal republic, liberty presupposes security; the point of security is liberty.); Paul Rosenzweig, *Defending the Pentagon's Information Awareness Program* Fox News, Sept. 8, 2003, *at* http://www.foxnews.com/story/0,2933,96694,00.html; Albanesius, *supra* note 85 (quoting Kim Taipale, "The goal is security with privacy . . . [that] does not mean balancing security and privacy but maximizing the set of results you want within [given] constraints.").

[207] Rosenzweig, *supra* note 26, at 23 (rejecting the notion that privacy should be considered a trump value).

[208] *See*, *e.g.*, Will Thomas DeVries, *Annual Review Of Law And Technology: III. Cyber Law: A. Privacy: Protecting Privacy in the Digital Age*, 18 Berkeley Tech. L.J. 283, 284 (2003) (citing the Qur'an and the Old Testament).

conception; in the main a creature of twentieth century legal developments.[209] The word privacy never appears in the U.S. Constitution or the Bill of Rights.[210] Most commentators credit a hallmark law review Article by Samuel Warren and Louis Brandeis[211] for establishing a right to privacy as a tradition of common law based on property rights.[212] The U.S. Supreme Court has since recognized a constitutionally protected privacy interest in various contexts, relying on the First, Third, Fourth, Fifth, Ninth and Fourteenth Amendments.[213]

Privacy law has generally developed in response to new technologies.[214] As emerging technologies have challenged existing doctrines through new forms of intrusion, new principles have emerged and new laws created.[215] As a result, U.S. privacy law is "disjointed and piecemeal,"[216] best described as *bricolage* – the notion of using whatever material or tools happen to be available.[217] Privacy law consists of a

[209] *See* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 Stan. L. Rev. 1393, 1430 (2001) ("Throughout this century, a distinctive domain of law relating to privacy has begun to develop.").

[210] *Cf.* Alaska Const., art. I, § 22 ("The right of the people to privacy is recognized and shall not be infringed.") and Calif. Const. art. I, § 1 ("All people are by nature free and independent and have inalienable rights. Among these are enjoying and defending life and liberty, acquiring, possessing, and protecting property, and pursuing and obtaining safety, happiness, and privacy.").

[211] Samuel Warren & Louis Brandeis, *The Right to Privacy*, 4 Harv. L. Rev. 193 (1890).

[212] *See* Solove & Rotenberg, *supra* note 203, at 3, 63—64.

[213] *See* James P. Nehf, *Recognizing the Societal Value in Information Privacy*, 78 Wash. L. Rev. 1, 33 (2003). For example, the Court has recognized the privacy interest in association (NAACP v. Alabama, 357 U.S. 449, 462 (1958)), politics (Watkins v. United States, 354 U.S. 178, 198—199 (1957)), and anonymity in public expression (Talley v. California, 362 U.S. 60, 64 (1960)); *see also* Solove & Rotenberg, *supra* note 203, at 20—21.

[214] "The modern evolution of the privacy right is closely tied to the story of industrial-age technological development - from the telephone to flying machines. As each new technology allowed new intrusions into things intimate, the law reacted - slowly - in an attempt to protect the sphere of the private." DeVries, *supra* note 208, at 285.

[215] *Id*. Even the Warren and Brandeis article, *supra* note 2011, was alleged to have been written in response to the development of new technology – the "instantaneous" camera and the tabloid newspapers that were beginning to intrude in the "private" lives of public figures. See William L. Prosser, *Privacy*, 48 Cal. L. Rev. 383, 383 (1960) (asserting that the article was written because Warren was upset when the newspapers had a "field day on the occasion of the wedding of a daughter."); *cf.* Robert Ellis Smith, *Ben Franklyn's Web Site: Privacy and Curiosity from Plymouth Rock to the Internet*, Privacy J. (2000), at 118—119 (asserting that Warren was reacting to a number of articles reporting on the dinner parties thrown by his wife at their home, including a "wedding breakfast for his cousin."). Regardless of the particular subject of the offending newspaper article, the Warren and Brandeis article was written in response to the emergence of new journalism technologies. *See* Solove, *supra* note 209, at 1431 (asserting that the Warren and Brandeis article "raised alarm at the intersection of yellow journalism . . . and . . . instantaneous photography'.").

[216] DeVries, *supra* note 208, at 285.

[217] Solove, *supra* note 209, at 1430.

"mosaic of . . . tort law, constitutional law, federal and state statutory law, evidentiary privileges, property law, and contract law."[218]

It is beyond the scope of this Article to review the history and full breadth of privacy law.[219] Further, although there are many aspects to information privacy law, this Article is most concerned with data protection and related interests.[220] For historical reasons, database privacy protection in the United States has relied to a great extent on the inefficiencies inherent in the distributed nature of our information infrastructure.[221] Having never built a centralized data repository,[222] the United States never developed a single comprehensive approach to data protection in either the public or private sphere.[223]

[218] *Id.*

[219] For a detailed discussion of privacy law and its conception, see generally Prosser, *supra* note 215; Nehf, *supra* note 213; Solove, *supra* note 209; Solove & Rotenberg, supra note 203 at 1—61; A. Michael Froomkin, *Symposium: Cyberspace and Privacy: A New Legal Paradigm? The Death of Privacy?* 52 Stan. L. Rev. 1461 (2000); Smith, *supra* note 215. *But see* Etzioni, *supra* note 205 (arguing that individual privacy needs to be balanced with communal goods); Rosenzweig, *supra* note 26, at 23 (arguing that the protection of privacy is not an absolute value); Richard A. Posner, The Economics of Justice 232—242 (1983) (arguing from an economic perspective that the individual right to privacy stems from a desire "to manipulate the world . . . by selective disclosure of facts . . . [in order] to mislead those with whom [the individual] transacts [business]" and is therefore economically and socially inefficient.).

[220] In other words, this Article is most concerned with concepts of privacy specifically as they relate to information contained in electronic form in databases. *See generally* Paul M. Schwartz & Joel R. Reidenberg, Data Privacy Law: A Study of United States Data Protection 5 (1996) (distinguishing connotations of "data protection").

[221] S*ee supra* note 159.

[222] *Id.*

[223] Thus, the general approach in the U.S. has been to deal with privacy on a piecemeal basis. "In contrast to the rest of the developed world, the U.S. has no strong, comprehensive law protecting privacy – only a patchwork of largely inadequate protection." Stanley & Steinhardt, *supra* note 22, at 15; Schwartz, *supra* note 216, at 1—3, 6—12.

With respect to restrictions on the federal government's ability to collect data through search or surveillance, the applicable laws are: U.S. Const. amend. IV (prohibiting unreasonable searches); Title III (governing electronic surveillance), 18 U.S.C. §§ 2510 et seq. (2003), as amended by the USA PATRIOT Act, Pub. L. No. 107-52 (2001); the Electronic Communications Privacy Act (governed access to stored communications), 18 U.S.C. §§ 2701 et seq. (2003), as amended by the USA PATRIOT Act; the Privacy Protection Act (protecting publishers), 18 U.S.C. § 2000aa; the Foreign Intelligence Surveillance Act of 1978, Pub. L. No. 95-511, 92 Stat. 1783 (providing a separate regime for "foreign intelligence"), 50 U.S.C. §§ 1801 et seq., as amended by the USA PATRIOT Act. *See generally* Electronic Frontier Foundation, Privacy – Surveillance & Wiretapping Archive, *at* http://www.eff.org/Privacy/Surveillance/ (last visited Oct. 2003); Congressional Research Service Reports, Privacy: An Overview of Federal Statutes Governing Wiretapping and Electronic Eavesdropping (updated Aug. 1, 2001); Congressional Research Service Reports, The Foreign Intelligence Surveillance Act: An Overview of the Statutory Framework (updated Apr. 29, 2002); Congressional Research Service, Electronic Briefing Book Terrorism – Wiretapping Authority, *available at* http://www.congress.gov/brbk/html/ebter130.html; U.S. Attorneys Manual, Electronic Surveillance, 9-7.000, *available at* http://www.usdoj.gov/usao/eousa/foia_reading_room/usam/title9/7mcrm.htm; USDOJ Computer Crime and Intellectual Property Section, Searching and Seizing Computers and Gathering Electronic Evidence Manual (July 2002), *available at* http://www.cybercrime.gov/s&smanual2002.htm; Department of Justice, Attorney General's Guidelines on General Crimes, Racketeering Enterprise and Domestic Security/Terrorism Investigations (May 2002), *available at* http://www.usdoj.gov/olp/generalcrime2.pdf [hereinafter DOJ Investigations].

Data gathered by the U.S. government in the ordinary course of providing government services is governed by the Privacy Act of 1974, 5 U.S.C. § 552a, as amended, which restricts the collection, use, and dissemination of personal information by federal agencies and allows individuals the right to access and correct personal data. *See generally* Nehf, *supra* note 213, at 35—45; Solove & Rotenberg, *supra* note 203, at 473—484. The Privacy Act also contains certain procedural restriction on "matching" information from several government databases and for sharing data among agencies, requiring certain inter-agency agreements. *See* The Computer Matching and Privacy Protection Act of 1988, Pub L. No. 100-503, § 1, 102 Stat. 2507 (1988) (appears as a note amending the Privacy Act in 5 U.S.C. § 552a (2003)). The Privacy Act contains exemptions for both computer matching and for inter-agency data sharing for national security and law enforcement purposes. *See* 5 U.S.C. §§ 552a(b)(7), 552a(a)(8)(B)(vi), 552a(j) (2003); s*ee also* Sean Fogarty & Daniel R. Ortiz*,* Limitations Upon Interagency Information Sharing: The Privacy Act of 1974, *in* the Markle Report, *supra* note 17, at 127—132. Note that in connection with the CAPPS II program, discussed *supra* Part II, the TSA is seeking an exemption from the Privacy Act for use of the Aviation Security Screening Records (ASSR) passenger information database that would be central to the CAPPS II system. *See* 68 Fed. Reg. 10, 2101—2103 (Jan. 15, 2003).

There are also numerous narrowly applicable laws on privacy and data protection that generally protect specific types of personal information held by the federal government and provide procedural protection for their disclosure or sharing. For example, U.S. Census data is protected under 13 U.S.C. § 9 (2003); certain medical records collected for research purposes under 42 U.S.C. § 242(m) (2003); educational records under 20 U.S.C. § 1232(g) (2003) (*but see* USA PATRIOT Act amendments, 20 U.S.C. 1232(g)(j)); and tax records under The Tax Reform Act of 1976, Pub. L. No. 94-455, 90 Stat. 1590. With respect to state governments, the Driver's Privacy Protection Act of 1994, 18 U.S.C. § 2721 (2003), regulates the use and disclosure of personal information from state motor vehicle records. There is a broad exemption for use by any government agency, including law enforcement, for use in carrying out its functions.

There are also a number of sector specific laws restricting the collection, use or disclosure of personal information by private sources. Among these, the Fair Credit Reporting Act, 15 U.S.C. §1681 et seq. (2003), regulates the use of information by credit reporting agencies, the Video Privacy Protection Act of 1988, 18 U.S.C. § 2710 (2003), prohibits the disclosure of video rental records, the Cable Communications Policy Act of 1984, 47 U.S.C. §551 (2003), limits disclosure of cable television subscriber data, and the Telecommunications Act of 1996, 47 U.S.C. §222 (2003), limits the use and disclosure of customer proprietary network information. Additionally, individually identifiable health information is protected by the Department of Health and Human Services "Standards for the Privacy of Individually Identifiable Health Information," 45 C.F.R. §§ 160, 164 (2003) pursuant to the Health Insurance Portability and Accountability Act of 1996, 42 U.S.C. § 1320(d) (2003) (in note).

Compare the comprehensive approach to data protection in Europe where there are two important supra-national policies covering both the public and private sectors. The Council of Europe's "Convention on Data Protection," European Treaty Series No. 18 (1981), and the European Union's "Data Directive," Council Directive 95/46/EC, O.J. L281:31 (1995), both of which recognize a fundamental "right to privacy" and restrict collection and uses of personal data. For a national enactment of the EU Directive, see, for example, the Data Protection Act, 1988, c.29 (Eng.). For an overview of international privacy law, see Solove & Rotenberg, *supra* note 203, at 687-783. For more on the EU Directive, see Schwartz & Reidenberg, *supra* note 221, at 12-17. *See also* Peter P. Swire & Robert E. Litan, None of Your Business: World Data Flows, Electronic Commerce, and the European Data Directive (1998); *Universal Declaration of Human Rights*, G.A Res. 217A U.N. GAOR, 3d Sess. (1948); *The International Covenant on Civil and Political Rights*, G.A. Res. 2200A (XXI), U.N. GAOR, 21st Sess., Supp. (No. 16) at 52 (1966). Both documents recognize a right of "privacy" in international law. For the full text of the resolution and the covenant, see Official Records of the General Assembly, 21st Session, and Supplement No. 16 (A/6316) 49.

It is beyond the scope of this Article to fully delineate the existing legal or regulatory structure relating to privacy, either in the U.S. or elsewhere. For a general overview of related U.S. law, see Gina Stevens, Congressional Research Service, Privacy: Total Information Awareness Programs and Related Information

The development of digital network technologies – allowing for easy and efficient data aggregation or virtual integration from these distributed sources – challenges this regulatory structure and creates new privacy concerns.[224]

2. Privacy Interests in Information

Before addressing the particular privacy concerns expressed with regard to data aggregation and data mining technologies in the context of domestic security specifically, it is helpful to review the general conceptual privacy interests that may be at stake.

Chief Justice Warren and Justice Brandeis framed their seminal privacy argument in terms of the "right to be left alone," and early conceptions of privacy generally followed that line.[225] With the emergence of electronic record keeping, two distinct lines of privacy theory developed: the first, based on traditional notions of "private space" is concerned with surveillance and physical intrusion, the second, based on control of information about the self, is concerned with self-determination or autonomy.[226] In *Whelan v. Roe*, the Supreme Court first recognized the right to information privacy finding that there are at least two kinds of individual interests to be protected: "One is the individual interest in avoiding disclosure of personal matters, and another is the interest in independence in making certain kinds of important decisions."[227] In a footnote to the opinion, the Court quotes Professor Kurland:

> The concept of a constitutional right of privacy still remains largely undefined. There are at least three facets that have been partially revealed, but their form and shape remain to be fully ascertained. The first is the

Access, Collection, and Protection Laws, *available at* http://www.fas.org/irp/crs/RL31730.pdf (Feb. 14, 2003). For an overview of general information privacy law, see Solove & Rotenberg, *supra* note 203.

For our purposes, it is sufficient to have a general understanding of the different approaches as well as to recognize that because of the complex and inconsistent structure of data regulation, individual data items in distributed databases may be subject to significantly different protections or other legal requirements based on what type of data, about whom, collected how, stored where, etc. This complexity is relevant to the discussion in Part IV *infra*, describing *rule processing technologies* for controlling access to or use of particular data items when aggregating or integrating data from a variety of sources.

[224] *See* Nehf, *supra* note 213, at 8—29; Garfinkel, *supra* note 131 (discussing concerns with the "database nation"); Roger Clark, *Information Technology and Dataveillance*, 31 Comm. of the ACM 498—512 (1988) (coining the term "dataveillance" to describe how database stores of personal information have facilitated new surveillance practices). *Cf.* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 Stan. L. Rev. 1393, 1398 (2001) (proposing that the Orwellian "Big Brother" metaphor be replaced with one based on Kafka's "The Trial" suggesting that the database problem is a "more thoughtless process of bureaucratic indifference, arbitrary errors, and dehumanization, a world where people feel powerless and vulnerable, without meaningful form of participation in the collection and use of their information" rather than the more traditional concern of secrecy or surveillance).

[225] *See* DeVries, *supra* note 208, at 286—287 ("In essence, they argued for a 'right to be let alone'.").

[226] *See* Cohen, *supra* note 25, at 577—579.

[227] 429 U.S. 589, 599 (1977) (citations omitted).

> right of the individual to be free in his private affairs from governmental surveillance and intrusion. The second is the right of an individual not to have his private affairs made public by the government. The third is the right of an individual to be free in action, thought, experience, and belief from governmental compulsion.[228]

These three "facets" can be seen to correspond with the notions of anonymity, secrecy, and autonomy. Anonymity is the interest in not being associated by government surveillance with one's "private affairs" or activities. Secrecy is the interest in not having those private affairs revealed or "made public.″ And autonomy is the interest in being "free in action."[229]

To many, the issue of information privacy is wrapped in philosophical musings of power, control, and human dignity. Thus, information privacy, in the words of one commentator, "is the claim of individuals, groups, or institutions to determine for themselves when, how, to what extent information about them is communicated to others."[230] Or, in the words of another, "'informational privacy' [is used] as shorthand for the ability to control the acquisition or release of information about oneself."[231] Underlying these articulations is the notion that individuals have a right to control information that defines themselves – and a corresponding right to not be judged on incomplete information (that is, "profiled").[232] Reputational information is held to have power because it may constrain opportunity.[233]

Others argue that such an individual right to hide reputation or "to conceal personal facts . . . presents opportunities for exploitation through misrepresentations" and

[228] *Id.* at 599, n.24 (quoting Philip B. Kurland, *The Private I: Some Reflections on Privacy and the Constitution*, Univ. of Chicago Mag. 7—8 (Autumn 1976)).

[229] This taxonomy was first suggested to me by Professor Eben Moglen, Columbia Law School, October 2002. *Cf.* Daniel J. Solove, *Conceptualizing Privacy*, 90 Calif. L. Rev. 1087, 1094 (2002):

> Although the extensive scholarly and judicial writing on privacy has produced a horde of different conceptions of privacy, I believe that they can be discussed under six headings: (1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control of personal information; (5) personhood; and (6) intimacy. These headings often overlap, yet each has a distinctive perspective on privacy.

[230] Alan Westin, *Privacy and Freedom* (1967) (excerpted in Solove & Rotenberg, *supra* note 203, at 28).

[231] Froomkin, *supra* note 219, at 1463.

[232] *See, e.g.*, Jeffrey Rosen, The Unwanted Gaze: The Destruction of Privacy in America 8 (2000) (where Rosen argues that a "central value of privacy" is to protect individuals "from being misidentified and judged out of context in a world of short attention spans, a world in which information can easily be confused with knowledge").

[233] *See* Paul M. Schwartz*, Internet Privacy and the State*, 32 Conn. L. Rev. 815, 825 (2000) (discussing the "autonomy trap . . . that it leads to a reduced sense of the possible."); *see also* Zarsky, *supra* note 96, at 34—41 (discussing manipulation and threats to autonomy from the commercial uses of data mining technologies).

is socially inefficient.[234] In part to counter such arguments, which tend to pit an individual right to conceal against the common good[235] (that is, the collective need to reveal), some argue that privacy itself should be considered a collective, not individual, good.[236] This argument is premised on a belief that privacy is a protective cocoon within which the individual can develop and make autonomous "decisions about speech, belief, and political and intellectual association" to the greater benefit of the community as a whole.[237]

I do not intend in this Article to resolve this philosophical debate. For our purposes, it is accepted that informational privacy, however conceived, is an important and valued individual right, the protection of which creates a greater social good. The question is: can its core interests – the autonomy of the individual – be protected through technical means?

3. Privacy concerns relating to Information Technology and Domestic Security

Distilled to a simple taxonomy, the significant privacy concerns voiced in opposition to the technologies with which this Article is concerned are primarily two: those that arise from the aggregation (or integration) of data and those that arise from the automated analysis of data that may not be based on any individualized suspicion.[238]

[234] Posner, *supra* note 219, at 232—233. *See also* Richard Murphy*, Property Rights in Personal Information: an Economic Defense of Privacy*, 84 Geo. L. J. 2381, 2382 (1996) ("In grossly oversimplified terms, the consensus of the law and economics literature is this: more information is better, and restrictions on the flow of information in the name of privacy are generally not social wealth maximizing, because they inhibit decisionmaking, [sic] increase transaction costs, and encourage fraud."). *See generally* Richard A. Posner, *The Economics of Privacy*, 71 Am. Econ. Rev. (1981); Richard A. Posner, *Privacy, Secrecy and Reputation*, 28 Buff. L. Rev. 1 (1979); Richard A. Posner, *The Right to Privacy*, 12 Ga. L. Rev. 393 (1978).

[235] *See generally* Etzioni, *supra* note 205 (articulating a "communitarian public philosophy of privacy").

[236] *See* Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. Cal. L. Rev. 1083, 1115—1116 (2002) ("Privacy is often viewed as an individual right. . . . The problem with viewing rights in purely individualistic terms is that it pits individual rights against the greater good of the community, with the interests of society often winning out because of their paramount importance when measured against one individual's freedom.").

[237] *See* Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 Stan. L. Rev. 1373 (2000) (excerpted in Solove, *supra* note 203, at 33—34); *see also* Nehf, *supra* note 213, at 69—74 (section titled "Moving Towards a Societal View of Privacy"), 74—81 (section titled "Defining Characteristics of Societal vs. Individual Concerns").

[238] *Cf. Safeguarding Privacy in the Fight Against Terrorism*, The Report of the [Department of Defense] Technology and Privacy Advisory Committee (Confidential Draft Dec. 16, 2003 at 20, on file with the author) (forthcoming Jan. 2004) [hereinafter, TAPAC Report] ("We believe the privacy risks that government data mining projects pose can be divided into six broad categories: (1) chilling effect and other surveillance risks; (2) data aggregation; (3) data inaccuracy; (4) data misuse; (5) false positives; and (6) risks associated with the act of data processing.").

The former might be called the "database" problem,[239] and the latter the "mining" problem.[240] The former is implicated in subject-based inquiries that access distributed databases to find more information about a particular subject, and the latter is implicated in the use of pattern-matching inquiries, in which profiles or models are run against data to identify unknown individuals.[241]

Additional concerns are that the technology will not work for the intended purpose (providing either a false sense of security by generating false negatives or imposing civil liberties costs on too many innocent people by generating false positives),[242] that the technology is subject to potential abuse, or that it will be vulnerable to attack.[243]

B. *Data Aggregation: The Demise of "Practical Obscurity"*

The efficiencies inherent in data aggregation itself cannot be denied; indeed, it is these efficiencies that provide the impetus for developing and employing data aggregation technologies in the first place.[244] Nor can the impact of this efficiency on privacy be denied.[245] New technologies that provide easy access to distributed data and efficiency in processing are obviously challenging to a system that is at least partially based on protecting certain rights by insisting on inefficiencies. On the one hand there is a need to "connect the dots" and on the other hand the notion of a free society is at least partially built on keeping the power to "connect the dots" out of the control any one actor, particularly the central government.[246] Making access to data easier and more efficient

[239] *See* Nehf, *supra* note 213, at 8—29, 32 (discussing "the database privacy problem"and how "[o]ur fears are heightened by a vague awareness of the absolute enormity of information residing in databases. . . ."). *See generally* Garfinkel, supra note 131; Solove, *supra* note 236 (characterizing the database problem as akin to the bureaucratic nightmare in Kafka's The Trial). In the TAPAC Report, *supra* note 238, this is referred to as the "data integration" risk.

[240] Or, the "fishing expedition" problem. *See supra* note 75. The TAPAC Report, *supra* note 238, discusses this concern in terms of its chilling effect on individual behavior that may result from data analysis not based on individual suspicion. "People act differently if they know their conduct *could be* observed." *Id.* at 20.

[241] *See* discussion *supra* notes 124—128 and accompanying text.

[242] *See, e.g.*, Public Policy Committee of the ACM, *supra* note 47 (in particular, the section on "Personal Risks");TAPAC Report, *supra* note 238, at 20 ("false positives" and "data inaccuracy").

[243] *See, e.g.*, Public Policy Committee of the ACM, *supra* note 46 (in particular, the section on "Security Risks"); TAPAC Report, *supra* note 238, at 20 ("data misuse" and "risks associated with data processing").

[244] *See supra* text accompanying notes 37—38, 47 (discussing the need to manage data volumes efficiently).

[245] *See supra* note 239.

[246] The very theory and structure of the U.S. political system built on checks and balances – among the federal branches and between the federal government and the states – recognizes that inefficiencies

(in a sense, lowering the transaction cost of data use) magnifies and enhances government power.[247]

Interestingly, the Supreme Court addressed the issue of data aggregation almost 15 years ago, albeit in contrapose to the problem at hand. In *Department of Justice v. Reporters Committee for Freedom of Press,*[248] the court held that raw FBI criminal data (in this case, a "rap sheet") that was officially part of the public record did not have to be disclosed to a reporter's Freedom of Information Act request because the aggregation of public records in one place negated the "practical obscurity" that protected those records in the world of distributed paper records.

Justice John Paul Stevens opined: "[T]here is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives and local police stations throughout the country and a computerized summary located in a single clearinghouse of information."[249] Thus, in weighing the relative rights of two private parties – the free press rights of the reporter and the privacy rights of the individual – the Court held that the interest in privacy expectations created by inefficiencies in data acquisition was a recognized and protectable interest.

The question that has not been definitively determined as yet is whether that same analysis, when applied to government aggregation or integration of previously discrete, distributed sources of information – each which it may have the perfect legal right to access individually – is itself problematic under the Fourth Amendment's right to be free from "unreasonable" search.[250] This issue becomes particularly heightened when

resulting from distributed power sharing are required to keep political power in check. *See generally* Edwin Corwin, Understanding the Constitution (2d ed. 1961).

[247] *See generally* Solove, *supra* note 236. *But cf.* Rosenzweig, *supra* note 26, at 10, discussing subject-oriented knowledge discovery:

> The only conceivable objection to the use of [knowledge discovery] in [the form of subject-oriented query] is an objection against enhanced government efficiency. This is not an objection to be slighted as negligible – the entire conception of separation of powers and checks and balances that inheres in our constitutional system is premised on the notion that barriers to effective government action enhance individual liberty. . . . [However,] in the unique context of terrorist threats that hold the potential for mass destruction, it appears advisable to relax our disapproval of an efficient government if suitable controls can be implemented (citations omitted).

[248] 489 U.S. 749, 780 (1989).

[249] *Id.* Note, this proposition has been cited by some to support a new legal standard for disclosure of public records – "that access to electronic records should be roughly equivalent to their availability on paper," thus building back into electronic distribution the inherent inefficiencies of distributed paper records. David Brin, The Transparent Society 76 (1998).

[250] However, as noted by Amitai Etzioni, the Fourth Amendment provides a balanced conception of privacy in that it does not privilege privacy because the "prohibition on unreasonable searches is not accorded more weight than the permission to conduct reasonable searches." Etzioni, *supra* note 205, at 206 n.77.

The legislative branch has also taken note of the privacy concerns with data integration. See the restrictions on data "matching" contained in the Computer Matching and Privacy Act, 5 U.S.C. § 522a (2003), discussed in *supra* note 223.

combined with the concern expressed below regarding queries that are not based on individualized suspicion. However, for purposes of developing guiding principles for technology development this question does not need to be settled, it is sufficient to recognize that the concern raised by data aggregation is legitimate and to suggest technology development and implementation strategies to mitigate the impact. Thus, the relevant question for subject-driven inquiries becomes to what extent technical efficiencies are to be allowed. Here, the critical juncture is the determination of subject. In other words, based on the identity of the subject and the purpose of the inquiry, what is the permissible scope of query – to what data can the query be directed and what standard of association must be met by any automated processing. A separate, but related, question is what consequences are triggered from the result – the more draconian the potential consequences, the higher the burden for use.

For example, a subject-based inquiry that seeks additional information about a known terrorist – already lawfully subject to investigation – and their associations or activities is fundamentally different than an inquiry to confirm identity of an air traveler and check against a "threat-list" triggered only by the making of a travel reservation. However, the policy issues involved do not turn so much on what kind of technology or even what methodology is to be used, but rather on what the standard for initial query ought to be and how that relates to what data should or should not be accessible and for what ultimate purpose.

The point here is not to minimize the privacy concerns but to isolate the issues so that technical and procedural protections can be built in. In the context of data aggregation and subject-based queries, *rule-based processing*, which can allow for incremental (rather than all-or-nothing) access to distributed data (and differential processing of such data) are part of the technical solution.[251] Rule-based processing allows the incorporation into the technology itself of policy judgments as to how, when, where, and for what purpose, particular information can be accessed.

C. *Data Analysis: The "Non-particularized" Search*

As noted earlier, a significant concern for privacy advocates in connection with data mining is that the search for previously unknown information may not be based on individualized or particular suspicion.[252] Rather, the data itself may be mined in order to discover certain patterns or relationships,[253] and then the pattern may be matched against

[251] *See infra* Part IV.

[252] *See supra* note 75 and accompanying text (discussing "fishing expeditions"). The privacy concern is that data analysis not based on individual suspicion amounts to "surveillance" and that surveillance may "chill" individual behavior. According to the TAPAC Report, *supra* note 234, at 20, "potential knowledge is present power" and awareness that government may analyze activity is likely to alter behavior; "people act differently if they know their conduct *could be* observed." The risk is that protected rights of expression, protest, association, and political participation may be affected by encouraging "conformity with a perceived norm, discouraging political dissent, or otherwise altering participation in political life." *Id.* at iii.

[253] *See supra* Part II and text accompanying notes 96—106.

new data to identify additional subjects for further processing.[254] For those opposed to the use of these technologies this amounts to a search "led by investigators with no clear idea how to identify real terrorist threats"[255] "put[ting] the government in the business of maintaining constant surveillance on millions of people"[256] – "a sharp departure from the principle that you have the right to be left alone unless your government has just cause."[257] Pattern-matching, it is contended, "investigate[s] everyone, and most people who are investigated are innocent."[258]

Although much of the concern behind these criticisms is legitimate, there are technical and procedural subtleties missing from the critics' analysis. First, as described above, a distinction must be drawn between the development of descriptive and predictive models (data mining in the narrow sense), which may employ undirected data mining techniques to model normative behavior, and their subsequent application to new data to find additional like occurrences or deviations (pattern-matching).[259]

Unlike in commercial applications, pattern development for domestic security or intelligence purposes usually involves analyzing actual (or hypothesized) terrorists or terrorist activity in order to discern whether there are identifying characteristics that can reveal a descriptive or predictive pattern that can then be used to identify other terrorists or related events.[260] To the extent that computational "data mining" is used to automate the task of extracting patterns, the data to be analyzed generally still relates to particular terrorists, terrorist activities, or related analogs – the intent of data mining is to uncover connections that may not be obvious from manual observation.[261] The popular

[254] Pattern-matching is discrete from data mining. Data mining is used to develop descriptive or predictive models (based on patterns). The application of patterns or models to new data is part of post-processing or decision making about how to use the discovered knowledge. *See supra* text accompanying note 98.

[255] Bray, *supra* note 23 (Edward Tenner commenting that, "[w]hen people do data mining, they don't really know what they are looking for").

[256] *Id.* (lead-in to comment by Kate Corrigan, ACLU Legislative Council).

[257] *Id.* (quoting comment by Kate Corrigan, ACLU Legislative Council).

[258] Solove, *supra* note 236, at 1109 (quoting Priscilla M. Regan, Legislating Privacy 90 (1995)).

[259] *See supra* text accompanying note 99, 124—128. Whether such pattern-matching is "particularized" or not depends on the efficacy of the model. *See infra* notes 266—68 and accompanying text.

[260] See *supra* text accompanying notes 120, 176—79, and 181, discussing data mining techniques in the domestic security context. A subject-based link analysis based on a priori determinations of particular suspicion alone may not always be sufficient to identify loose confederations or independent local cells working towards a common goal, thus the need to develop descriptive and predictive models.

[261] To the extent that the model is based on hypothetical data there should be no privacy concerns at all in its development. A specific research effort of TIA is to develop descriptive and predictive models based on using a "red team of terrorism experts to create synthetic terrorist attack scenarios" and "produce transaction data reflective of [those] attacks" for analysis. IAO Report, *supra* note 88, at A-11. Although this process has been derided (see, e.g., *EFF, supra* note 164, stating that the "TIA R&D is using synthetic data (sort of like 'The Sims' gone wild)"), one can imagine how this approach would be useful in counter-terrorism where future terrorist acts may not resemble those in the past. For example, had the use of

conception that vast amounts of information relating to innocent subjects is mined with no idea as to what the investigator is looking for, on the hope of uncovering "suspicious patterns," is generally false.[262]

But, such generalized undirected data mining of information relating to innocent individuals may indeed be employed in counter-terrorism in certain narrow applications to develop normative models. These models can then be used to contrast against terrorist patterns in order to identify potential terrorists from a general population by subsequently using deviation analysis (that is, to look for "suspicious patterns"). However, although individual data of innocent persons may be processed during such model development, it is not unequivocal that there is a significant privacy impact if such analysis is restricted to developing aggregate categories of "normal" patterns in order to find deviations in future pattern-matches any more than there is from any other statistical analysis of "personal data" that is analyzed or reported in the aggregate, for example social science analysis of census data or medical studies statistically reporting aggregations based on individual cases. Since no individual identifying information (nor any corresponding personal data) is returned to the analyst during the automated development of the model (and there is no scrutiny by any human of any individually identifiable data), it is difficult to discern the privacy implications of the data analysis itself unless privacy is conflated with absolute secrecy. [263]

airliners as missiles to attack U.S. targets been hypothesized prior to 9/11 (based on intelligence previously available, *see* Joint Inquiry Report, *supra* note 3, at 209), transactional data relating to flight schools, student visas, etc. may have been identified as relevant. Thus, the arrest of Zacarias Moussaoui in August of 2001 might have triggered more response.

Another criticism of developing models on hypothetical scenarios is that "the number of 'threat scenarios' that could be imagined is nearly infinite." *ACLU*, *supra* note 164, at 11. However, relational data mining research for domestic security applications recognizes that the nature of specific future acts may be difficult to predict, thus, the intent is to develop models based on uncovering lower level, more frequently repeated activities that are themselves patterns of numerous different specific terrorist actions. *See supra* text accompanying notes 104—106.

[262] *See* Hearings Before the Subcomm. on Technology, Information Policy, Intergovernmental Relations and the Census of the House Comm. on Government Reform, 108th Cong. 1—2 (2003) (written testimony of Dr. Tony Tether, Director for Defense Advanced Research Projects Agency), *available at* http://www.fas.org/irp/congress/2003_hr/050603tether.html (stating, in reference to the undirected mining of data, that "DARPA is not pursuing these techniques."); *see also supra* text accompanying notes 120, 176—79, 181 (noting that such undirected queries are not suited for counter-terrorism goals that require relational analysis).

[263] For example, in the recent Jet Blue incident, although personal information relating to individual flight transactions was analyzed and correlated with household income and other personal data from credit reports, the output of the analysis was the development of aggregate categorical models – for example, "short-trippers," "short notice/short stay," "high spenders, "stranded," "frequent one-way travelers," and "non-descript" – in order to subsequently identify outliers (deviations) based on comparing known terrorist patterns with these normative models. As a technical matter the analysis of the Jet Blue data did not involve the individual scrutiny of any personal data except as the input to develop a statistical model. *See* Philip Shenon, *Airline Gave Defense Firm Passenger Files*, N.Y. Times, Sept. 20, 2003, at A1; Ryan Singel, *JetBlue Shared Passenger Data*, Wired News, Sept. 18, 2003, *available at* http://www.wired.com/news/privacy/0,1848,60489,00.html. The fact that JetBlue may have turned over raw, identifiable personal data to a third party in violation of their own privacy policy in the first place is an entirely different matter with other privacy concerns unrelated to the automated analysis itself. The point is that there are technical means, for example, data anonymization, discussed elsewhere herein, that could be

On the other hand, pattern-matching queries, in which descriptive or predictive models (whether mined from real data relating to terrorists or derived from hypothetical scenarios) are run against new data in order to identify unknown subjects or activities for further investigation, may directly implicate the issue of the non-particularized search.[264] However, by developing technologies that use *selective revelation* (that is, techniques that separate transactional data from identity or otherwise reveal information incrementally) these concerns can be significantly reduced by maintaining anonymity, which in turn protects autonomy – that is, the ability to freely act within the rules of the polity without being surveilled.[265]

Pattern-matching is not inherently a surveillance technology. No individual dossier is created and no individual is scrutinized for suspicious behavior. No person or behavior is individually observed or surveilled by the automated analysis itself.[266] To the extent that valid behavioral or transactional profiles are developed,[267] a search for

used to mitigate these legitimate concerns but that the automated analysis itself is not necessarily an inherent privacy intrusion.

Compare, however, the EU Data Protection Directive, *supra* note 223, which requires that any "processing of personal data" come within its protections. But, even under such a restrictive regime anonymized data-matching has been opined as permissible. *See* Stewart Baker et al., Steptoe & Johnson White Paper, Anonymization, Data-matching and Privacy: A Case Study (Dec. 1, 2003) (on file with the author).

Further, although it is useful to understand the distinction being drawn in this section between directed and undirected uses, the actual use of undirected data mining to develop normative models for domestic security applications in order to identify deviations is limited. Deviation analysis only works well in situations of very precise and constrained normative models with clear deviations that are unlikely to be found in counter-terrorism where the search is not for outliers or deviants from normative models but, rather, for "in-liers," that is, terrorists engaged in generally normative behaviors but whose links or relationships may reveal illegal organization or activity. *See* Taipale, Privacy, *supra* note 28, at slides 16—17.

Interestingly, such generalized deviation analysis has its greatest applicability in domestic security applications not for finding terrorists but for "watching the watchers," that is, by monitoring usage and log files of authorized government users for deviations to prevent abuse. See discussion of logging in Part IV *infra.* Other areas where deviation analysis has applicability include intrusion detection, money laundering and other uses where deviations from well understood normative patterns in defined transaction spaces have efficacy.

[264] *But see infra* text accompanying notes 265—268.

[265] Autonomy is not synonymous with being able to commit or plan terrorist acts in secret without being discovered. Privacy based arguments against technology that can potentially protect against catastrophic terrorist acts must show a privacy impact on legitimate innocent activity.

Thus, a pattern-based query that was 100% effective at identifying terrorists and only terrorists could not be considered "chilling" of constitutionally protected activity since such accuracy would far exceed even a stringent requirement for probable cause – indeed, absolute accuracy (if it were possible) would prove guilt beyond a reasonable doubt. Thus, the policy issue is to decide what accuracy rate for pattern-based queries is appropriate or required under what circumstance.

[266] Again, it is important to maintain the distinction between statistical analysis of data, including personal but non-identifying and transactional data, and observing individualized identifying information.

[267] Valid in the sense that they have passed some policy threshold – for example, reasonable suspicion or probable cause – for use in a general context. *See supra* note 265.

matching behaviors is undertaken. Once matching behaviors are identified, there may be a Fourth Amendment (and due process) issue regarding whether the suspicion is sufficiently reasonable to "particularize" the search – that is, to connect the behavior with identity.[268]

This exposes a particular interesting (and generally accepted) misconception in popular notions of "privacy" and data mining, particularly as it relates to autonomy, pattern-matching, and non-particularized search. Although it is contended that pattern-matching "alters the way government investigations typically occur,"[269] it is unclear that this is so. For example:

> Usually the government has some form of particularized suspicion, a factual basis to believe that a particular person may have engaged in illegal conduct. Particularized suspicion keeps the government's profound investigative powers in check preventing widespread surveillance and snooping into the lives and affairs of all citizens. Computer matches, Priscilla Regan contends, investigate everyone, and most people who are investigated are innocent.[270]

But how is pattern-matching in a database any different than observing behavior in public? A simple example may illustrate the point. Suppose that a police officer observes an individual running on a public street wearing a mask. Due process requires that the officer comply with certain standards of reasonable suspicion and other procedures before taking additional action, not that he close his eyes. If stopping and questioning an individual who is running in the street wearing a mask is reasonable (it may or may not be in the particular circumstance), then why is questioning or investigating someone whose electronic trail indicates a reasonable suspicion of terrorist activity presumptively not? More importantly, does observing the running suspect somehow invade the privacy of the others on the street who are also observed but not questioned?

Obviously, the answer turns on whether one considers the particular database the equivalent of the public street. But that highlights the paradox: to the extent that the question is whether the particular form of data (street observation or database) is subject to expectations of privacy, we are squarely within the traditional Fourth Amendment jurisprudence.[271] Thus, there is no general non-particularized suspicion problem – only

[268] From a civil liberties perspective focusing investigative resources on suspects through behavior profiling would seem preferable to existing methods of using racial or ethnic attributes for screening. From a law enforcement perspective behavior profiling would also be more effective. Further, automatically processing data would seem to enhance privacy over manual screening in which a human, subject to human bias, reviews the data. *Cf. supra* notes 38, 117.

[269] Solove, *supra* note 236, at 1109.

[270] *Id.*

[271] Fourth Amendment protection generally applies where an individual has a "reasonable expectation of privacy." Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., dissenting) ("[T]he rule . . . is that there is a twofold requirement, first that a person have exhibited an actual (subjective) expectation of privacy and, second, that the expectation be one that society is prepared to recognize as 'reasonable.'"). *See*

the same issue encountered before, that is, is the pattern-matching "reasonable" in the particular context of its use. And that question is related to its efficacy, the point of the research and development at issue.[272]

Others, however, argue that there is no protection from "anonymized" data because even if the "authorities had to get a warrant . . . to access our information and discover our identity . . . all the bad effects from [surveillance] would be felt" because "it will always be possible that our innocent activities will coincide with" a pattern being matched and "we would no longer be free." However, that argument raises the issue of false positives (that is, innocent people identified by the model) and not that of non-particularized search.[273] To argue that transactional data should be absolutely secret – as

generally Solove & Rotenberg, *supra* note 203, at 275—322 (discussing the Fourth Amendment and emerging technologies).

Note that in *Katz* the Court also stressed that "[w]hat a person knowingly exposes to the public . . . is not a subject of Fourth Amendment protection." 389 U.S. at 351. Hence*,* in U.S. v. Miller, 425 U.S. 435 (1976) (concerning bank records) and Smith v. Maryland, 442 U.S. 735 (1979) (relating to telephone numbers dialed), the Supreme Court has found that there is generally no Fourth Amendment privacy interest in the business records of third parties to whom information was given. That is, once a party voluntarily discloses information to a third party, the first party no longer has a reasonable expectation of privacy for that information under the Fourth Amendment that would require a warrant. The general rule is that the issuance of a subpoena to a third party to obtain the records of that party does not violate the rights of the original party unless there is a statutory requirement imposing a higher standard for a particular type of information. Determining whether a higher (or lower) standard should be required by statute for certain database transaction records or any other specific type of information is a policy question unrelated to the use of any particular technology or technique. *See infra* note 271.

[272] This point is addressed in greater depth *infra* Part III.D.

[273] To the extent that inhibited behavior is actually terrorist, there is no privacy issue, rather, that is successful deterrence. To the extent that innocent activity is potentially subject to false identification as terrorist activity, that is a problem of false positives. *See* discussion *infra* Part III.D.

The concern that innocent activity may be "chilled" (*see*, *e.g.*, ACLU, *supra* note 164, at 7) by the potential for disclosure even pursuant to a warrant dismisses the relevance of existing Fourth Amendment protections, specifically, the warrant requirement and the judicial determination of 'reasonableness' for revealing identity on the basis of pattern-matched suspicion. *Cf.* Solove, *supra* note 236, at 1124—28 (highlighting the significance of judicial imposition and the warrant requirement in maintaining an "architecture of power" to protect privacy). Warrants "raise the standard of care of law enforcement officials by forcing them to document their requests for authorization" and "the requirement of prior approval prevents government officials from dreaming up post hoc rationalizations." *Id*. at 1126—1127 (citations omitted).

The purpose of my Article is not to defend existing Fourth Amendment procedures as necessarily adequate in general (or sufficient in any particular context) but only to recognize that (1) there are constitutional and statutory mechanisms for protecting due process and privacy in the real world, and (2) that these can be applied in the context of using data aggregation and analysis technologies assuming that certain features are built in. To the extent that additional procedures or standards may be desired for particular types of information, that is a more general policy question unrelated to the thesis of this Article. For a general discussion of statutory regimes imposing higher (and lower) standards for certain searches or activities, see Solove, *supra* note 236, at 1138—51. For example, the Cable Communications Policy Act requires the government to obtain a court order (not merely a warrant) to obtain cable records. 47 U.S.C. § 551(c)(2)(B) (1984) (note that the USA PATRIOT Act, *supra* note 43, amended the Cable Act to apply only to records about cable television service and not other services, such as Internet access or telephone, that a cable operator might provide. Pub. L. No. 107-56, Title II, §211.) E-mail transmissions stored on a third party system are protected under the Electronic Communications Privacy Act. 18 U.S.C. §§ 2510 et seq. (2003); *see also* Rosenzweig, *supra* note 26 (proposing an implementation scheme for TIA that would opposed to anonymous and subject to traditional notions of due process – is not only to privilege privacy as an absolute, it is to extend its reach far beyond existing interests in maintaining individual autonomy for legitimate purposes. Prohibiting pattern-matching queries is more akin to hiding footprints than protecting the disclosure of shoe purchases.[274]

Although the separate issue of false positives is not an insignificant problem (and is addressed in the following subsection), it is qualitatively and quantitatively different than claiming that "everybody is being investigated" through pattern-matching. In reality only the electronic footprints of transactions and activities are being scrutinized – to the extent that there are suspicious footprints there may or may not be consequences to the individual who left them, and there are technical means to make those consequences conform to existing Fourth Amendment or statutory due process procedures.[275] The primary policy issues involved in applying these technologies then are determining what *confidence interval* for the technology and methodology are required to meet the "reasonableness" test, and what procedural protections are imposed between their application and any consequence to the individual.[276]

In any case, this Article argues that the use of *selective revelation* technologies can mitigate the non-particularized suspicion concerns by permitting the imposition of judicial due process between the observed behavior and the act of revealing identity.[277] The automated analysis of potentially relevant transactional data while shielding the exposure of individual identity to a generalized search protects privacy by maintaining anonymity, which in turn preserves autonomy.[278] Rather than minimizing concerns relating to non-particularized suspicion, this Article suggests that traditional due process protection can be built into both the technology and its implementation policy by using

provide additional statutory administrative and judicial procedures and review); Potomac Institute for Policy Studies, Oversight of Terrorist Threat Information: A Proposal, *supra* note 28 (suggesting a higher administrative standard for disclosure of identity or information relating to a U.S. person).

[274] Whether shoe purchases should be disclosed and under what circumstance is a legitimate, but not novel, policy issue – one which existing legal doctrines easily encompass.

[275] For a discussion of statutory regimes imposing higher (and lower) standards or procedures than the Fourth Amendment for certain searches or activities in particular contexts, see Solove, *supra* note 236, at 1138—51; *supra* note 273.

[276] Determining confidence intervals, that is, testing the efficacy of these technologies, was a primary focus of the TIA program's research and development. *See* IAO Report, *supra* note 88, at A-11 ("DARPA's goal for this activity is to find out what is possible.").

[277] The requirement for judicial intervention "particularizes" the suspicion to an individual in exactly the same way that a subpoena or court order for disclosure compels identity in a "John Doe" proceeding. The idea that pattern-matching turns traditional notions of particularized suspicion on its head fails to account for such proceedings or any other in which the unknown identity of an actor is sought by subpoena, warrant or court order for an observed behavior or known attribute. The constitutional issue with pattern-matching, to the extent that there is one, is simply whether there was a reasonable expectation of privacy attached to the locus of the observed behavior. Thus, the type of data and how it was gathered may or may not be constitutionally relevant, but the method of query – directed or undirected – should not be.

[278] *See supra* note 265 and accompanying text.

selective revelation. Under selective revelation, pattern-matching would not lead directly to individual identity without being subjected to the appropriate legal standards. Where matching provides information that is in itself sufficient to meet investigative, reasonable suspicion, or probable cause standards – where the observed match "particularizes" the suspicion[279] sufficiently (that is, reasonably under the circumstances in conformity with Fourth Amendment requirements) – the relevant procedural protection – subpoena, warrant, or court order – can be applied depending on the specific context before identity is revealed or acted upon.[280] Enforcement of these protections would follow the traditional means – pattern-matches determined to be unreasonable would be subject to the exclusionary rule, administrative proceedings, or civil redress.[281] Additionally, in cases of pattern-matching that leads to "adverse, non-punitive collateral civil consequences" (for example, watch-listing) additional administrative procedures requiring notice, time limits, and other due process protections can be devised, including an individual right to appeal adverse administrative review to federal court for *de novo* review.[282]

D. *Data Mining: "Will Not Work."*

Another common criticism is that data mining will not work to help prevent terrorism.[283] This criticism has two prevalent articulations – first, that "dirty data" will lead to mistakes,[284] and, second, that the statistical nature of the analysis will, in any case, return many false positives.[285]

[279] *See supra* note 264—268 and accompanying text.

[280] *See* Stevens, *supra* note 223, at 16—19. Note that Rosenzweig, *supra* note 26, at 15 would require the "interposition of a judicial officer before the barrier of anonymity is broken."; *see also* Solove, *supra* note 236, at 1124—1128 (highlighting the significance of judicial imposition and warrant requirements in maintaining an "architecture of power" to protect privacy).

[281] *See generally* Solove & Rotenberg, *supra* note 203, at 280.

[282] *See* Rosenzweig, *supra* note 26, at 19 (outlining such a structure). "One could, of course, imagine equivalent mechanisms for review that would be equally protective – the proposed is merely one model." *Id.* Additional models, for example, models analogous to those used to manage foreign signal intelligence related to U.S. persons could also be considered. *See*, *e.g.*, Gallington testimony, *supra* note 28.

[283] *See*, *e.g.*, Bray, *supra* note 23. ("I think there are serious scientific questions to be raised about whether it's even feasible."); *see also* Stanley, *supra* note 22 ("The dubious premise of programs like TIA . . . that 'terrorist patterns' can be ferreted out . . . probably dooms them to failure.").

[284] *See*, *e.g.*, Robert Gellman, *Total Info Project is Totally Doomed*, 22 Government Computer News No. 3 (Feb. 10, 2003) ("Records are typically filled with errors, cross-links, and obsolete information. . . . The ancient computer principle of Garbage In, Garbage Out is relevant here."), *available at* http://www.gcn.com/22_3/tech-report/21064-1.html.

[285] A false positive occurs when the analysis identifies innocent subjects as terrorists. *See* Public Policy Committee of the ACM, *supra* note 46, at ¶¶ 11—15; Gellman, *supra* note 280. Another argument put forward against passenger screening in particular is that terrorists will adapt to the screening methods in order to slip through the system. *See* Samidh Chakrabarti & Aaron Strauss, *Carnival Booth: An Algorithm*

As noted earlier, "dirty data" is a recognized problem for data mining applications (as well as other statistical analyses).[286] However, there are two responses to this criticism based on a deeper understanding of the technology, its current development and its likely application in context. First, as previously stated, data mining is a single step in the knowledge discovery process. Pre- and post-processing strategies to lessen the impact of dirty data are standard procedures in the knowledge discovery process already and improved methods are being developed.[287] Second, advances in machine learning and adaptive algorithms are themselves geared towards employing self-correcting methodologies within the data mining process itself.[288]

Nevertheless, these issues are legitimate concerns and present challenges to the application of these technologies for domestic security. However, they seem misguided as the basis for arguing against further research that is specifically aimed at developing methods to overcome these problems and to test efficacy,[289] especially when many of the

for Defeating the Computer-assisted Passenger Screening System, *at* http://www.swiss.ai.mit.edu/6805/student-papers/spring02-papers/caps.htm. Again, this concern should be taken into account when developing applications but does not argue against research and development, or deployment, with adequate accounting for the problem. First, there are obviously ways to defeat any system. Nevertheless, they are worthwhile because they raise the cost of engaging in the terrorist act by requiring avoidance strategies. Not only do such avoidance strategies increase 'costs' to the terrorist but they also provide additional points of potential error on the part of the terrorist that may lead to discovery. Obviously, if we were to take this critique too seriously on its face it would support the conclusion that locks should not be used because locksmiths (or burglars with locksmithing knowledge) can defeat them. Second, to the extent that we are talking about researching adaptive machine learning based algorithms, an important research objective would be to try to anticipate these avoidance methods in application, algorithm and system design, including by building in both variability and random outcomes (for example, by combining random searches with CAPPS II).

[286] *See supra* notes 91—95 and accompanying text.

[287] *See supra* note 94—95 and accompanying text.

[288] *See supra* note 95 and accompanying text. *But see* Jensen, *supra* note 37, at slide 24 (highlighting the problems with fragmentary data in relational analysis).

[289] Making dirty or non-conforming data from multiple sources useful for analysis is among the research goals of TIA, see DARPA, *Report to Congress regarding the Terrorism Information Awareness Program* (2003), *available at* http://www.darpa.mil/body/tia/TIA%20ES.pdf.

Obviously, the development of improvements in any of these areas will also have significant benefit for commercial and consumer applications. *See*, *e.g.*, Howard Bloom, *I Want My TIA*, Wired Magazine, Apr. 2003, *available at* http://www.wired.com/wired/archive/11.04/view.html (arguing that the development of TIA technology will improve general search applications).

Nevertheless, their efficacy for domestic security has yet to proven:

> To be sure, the ultimate efficacy of the technology developed is a vital antecedent question. If the technology proves not to work . . . than all questions of implementation are moot. For no one favors deploying a new technology – especially one that threatens liberty – if it is ineffective. . . . The vital research question, as yet unanswered, is the actual utility of the system and the precise probabilities of its error rate.

Rosenzweig, *supra* note 26, at 4. That is, a prime research goal of TIA was to determine the confidence interval for domestic security application. *See supra* note 272.

examples used by critics – for example, difficulties in resolving identity[290] – are already "solved problem."[291]

The second prong of this concern, that data mining is a statistical analysis and therefore prone to generating false positives,[292] seems a trite observation since all investigative methods begin with more suspects than perpetrators – indeed, the point of the investigative process is to narrow the suspects down until the perpetrator is identified.[293] The question to be tested is whether these technologies can reduce the number of potential suspects sufficiently so that traditional investigative methods applied to the results can identify terrorists before they act.[294] To date, much of the public debate about potential accuracy rates for data mining in domestic security applications,

Given the widespread effective use of these technologies in other areas, for example, life-saving drug design and medical diagnostics, scientific discovery and financial fraud detection, it seems premature to kill research on the mere articulation of questioned success, particularly when those in the field believe the criticism is especially ill-informed as to the technology and its potentials. *See* SIGKDD of the ACM, *supra* note 28. This is particularly so when there are known contexts, for instance, clinical diagnosis, where automated analysis exceeds human capabilities. *See supra* note 118.

[290] *See, e.g.*, Securing the Freedom of the Nation: Collecting Intelligence Under the Law: Hearings Before the House Permanent Select Comm. on Intelligence, 108th Cong. (2003) (testimony of ACLU Legislative Counsel Timothy Edgar), *available at* http://www.aclu.org/SafeandFree/SafeandFree.cfm?ID=12313&c=206:

> One reason why data mining could ultimately prove to be a false security solution is the unreliability of much information in the computer data to be "mined." As the technologists say, "garbage in, garbage out." For example, the Consumer Federation of America and the National Credit Reporting Association found in a new study that 10 percent of credit reports contain errors in names or other identifying information . . . .

[291] *See* Rosenzweig, *supra* note 26, at n.11 (*citing* Jeff Jonas, Center for Strategic and International Studies, *Data-mining in the Private Sector*, July 23, 2003) (the "question of resolving identity – that is, ensuring that data all refer to a single unique individual – is a 'solved problem'."); *see also* Steve Mollman, *Betting on Private Data Search*, Wired News, Mar. 5, 2003 (describing Anonymous Entity Resolution software that identifies individuals listed under different names in separate databases), *available at* http://www.wired.com/news/technology/0,1282,57903,00.html. Additionally, there are statistical methods to correct for data problems. *See supra* note 94.

[292] *See*, *e.g.*, Public Policy Committee of the ACM, *supra note* 47.

[293] "The object of the game is to discover the answer to these three questions: 1st. WHO? Which one of the several suspects did it?" Parker Brothers, Instructions to "CLUE" board game (1963 ed.).

[294] *See supra* note 273 and accompanying text. *See also supra* note 80 (discussing opposition to research on the basis that "it might not work").

particularly in news accounts,[295] is based on wildly simplistic assumptions about the statistical nature of data mining.[296]

Again, the fundamental research question is to determine what *confidence interval* can be applied to data mining results in the domestic security context.[297] To be sure, their efficacy has yet to be proved, "but at this stage the significant step is to recognize the research nature of the program, and thus to avoid strangling nascent technology in its crib by imposing unreasonable and unrealistic 'proving requirement' long before the technology has had a chance to be explored."[298]

That the criticism of data mining's probabilistic nature is often advanced by self-proclaimed statisticians[299] is also interesting. The argument seems to be the following: statistical analysis is good enough for statisticians to use for many noble purposes, including in support of billion dollar policy decisions, life and death diagnostics, et cetera, but it is not good enough for law enforcement. The smuggled assumption in this argument is that law enforcement cannot be trusted to de-couple analysis from action or to take into account and compensate for potential error rates by adopting appropriate procedures. Without addressing the substance of the underlying assumption, there are structural implementation options that can ameliorate both of these concerns as well.

For example, the Markle Foundation Report recommends separating the data analysis (intelligence) function from the law enforcement function for implementation.[300]

[295] *See, e.g.*, Farhad Manjoo, *Is Big Brother Our Only Hope Against Bin Laden?*, Salon.com, Dec. 3, 2002 (suggesting that 80% is a "reasonable but probably still too high" accuracy rate for TIA, resulting in 48 million false positives, a number that represents 20% of the 240 million adult Americans), *at* http://www.salon.com/tech/feature/2002/12/03/tia/index_np.html.

[296] For a technical assessment and response to these "false positive" arguments, see *Jensen, Technical Assessment*, *supra* note 89, at § 1 ("We show how both the accuracy and privacy impact of a hypothetical system could be substantially improved" using an enhanced "simulation model that more closely matches realistic systems.").

These false positive critiques "are based on a naïve model of systems for knowledge discovery and data mining" that uses simplistic assumptions from propositional data analysis in commercial settings to critique relational data analysis for domestic security. *Id*. Among the faulty assumptions identified in this assessment are: (1) assuming the statistical independence of data (appropriate for propositional analysis but not for relational analysis, see text accompanying notes 178—179), (2) using binary (rather than ranking) classifiers, and (3) applying those classifiers in a single pass (instead of using an iterative, multi-pass process). An enhanced model correcting for these assumptions has been shown to greatly increase accuracy (as well as reduce aggregate data utilization).

A variation of the false positive argument is to say that in order to eliminate false positives, the criteria for matching will be tightened to the point where there will be many false negatives, thus undermining actual security by classifying terrorists as "innocent." Again, these arguments are generally based on a simplistic binary methodology that is unlikely to be used in practice where ranking classifiers will inform resource allocations rather than binary classifications exonerating suspects.

[297] *See supra* note 104 (discussing confidence intervals).

[298] Rosenzweig, *supra* note 26, at 4.

[299] *See, e.g.,* Bray*, supra* note 23; Manjoo, *supra* note 295.

[300] Markle Report, *supra* note 17 at 2, 22—24:

By separating intelligence from enforcement, an additional intervention point for legal process control is created. In addition, requiring one agency (or even one bureau within an agency) to pass on intelligence to another for action creates an institutional hurdle and additional administrative control over misuse.[301] Within agencies, administrative procedures can be adopted to manage data use and error correction.[302] In addition, technical methodologies can be adopted to minimize errors and data requirements.[303]

> [The Markle Foundation] Task Force's basic conception is that the Department of Justice and its FBI should be the lead agencies for law enforcement, exercising the power to investigate crimes, charge people with crimes, perhaps take away their liberty, and prepare cases for trial and appeal. The [Department of Homeland Security] should be the lead agency for shaping domestic intelligence products to inform policymakers, especially on the analytical side, so that there is some separation between the attitudes and priorities of intelligence analysis and the different, more concentrated, focus of law enforcement personnel authorized to use force on the street to make arrests and pursue or detain citizens.
>
> We understand that criminal investigation (and counterintelligence) often overlaps with intelligence work. Some overlap is natural and good. But the case for a fundamental separation is strong. Intelligence has much broader purposes than criminal investigation. The operational objectives are different. The training is different. The rules about how to collect, retain, and share information are different. The relationships with sources of information are different.
>
> Therefore the DHS should take the lead in collecting information that is publicly available or voluntarily obtained and in analyzing domestic information and intelligence from all sources and setting overall priorities for new collection efforts, working within an interagency process that will include the FBI and other relevant agencies in the intelligence community. It should coordinate the national organization of homeland security task forces in states, regions, and metropolitan areas across the country. But the FBI should continue to have the responsibility for managing clandestine collection operations, like FISA wiretaps or the recruitment of undercover agents, under the supervision of the Attorney General.

Calls to create a separate domestic intelligence agency modeled on Britain's MI5 have arisen again in light of perceived difficulties in transforming the FBI from a reactive law enforcement agency to a preemptive intelligence agency. *See, e.g.*, Gary Thomas, *CIA Identity Flap Sparks Debate Over Need for New Domestic Intelligence Agency*, Voice of America, Sept. 30, 2003 ("Former congressman Lee Hamilton is vice chairman of the independent commission investigating the terrorist attacks of September 11, 2001. 'In December we will examine reforms by the FBI and whether we need a new agency to gather intelligence in the United States, what some have called an American version of Britain's MI5,' he said."), *available at* http://www.iwar.org.uk/news-archive/2003/09-30-5.htm .

And, the Fifth Annual Report of the Advisory Panel to Assess Domestic response Capabilities for Terrorism Involving Weapons of Mass Destruction, (the "Gilmore Commission"), released on December 15, 2003, *available at* http://www.rand.org/nsrd/terrpanel/ calls for the establishment of the Terrorist Threat Integration Center, *supra* note 28, as an independent agency to coordinate domestic intelligence. *See also* Albanesius, *supra* note 85 (quoting Kim Taipale, "perhaps there is a 'need for a specific intelligence agency to go after terrorists' with a limited charter").

[301] Administrative procedures requiring documentation and authorization raise the standard of care for use of these technologies (or any procedure). *See* Rosenzweig, *supra* note 26 (setting forth various administrative procedures for controlling use and providing accountability); The Computer Matching and Privacy Act of 1988, Pub. L. No. 100-503, § 1, 102 Stat. 2507 (1988) (requiring formal inter-agency agreements to share data).

[302] *See* Rosenzweig, *supra* note 26, at 17—19 (detailing administrative and judicial procedures to manage data use, error correction, accountability and oversight).

Nevertheless, the purpose here is not to dismiss these concerns but to illustrate where they have applicability in practice and where appropriate procedures or technology development strategies can potentially help avoid privacy concerns. The overriding principle for implementation of any system relying on data mining or knowledge discovery technologies should be that these technologies are considered investigative, not evidentiary, tools and that they meet some reasonable standard for efficacy in the particular context of their use that is subject to appropriate due process procedures and review. Use of automated analytic tools and statistical methodologies should not result in any direct law enforcement outcome without procedural and substantive safeguards for protecting civil liberties, correcting errors, and accounting for misuse. Procedures, administrative organization, and technologies supporting strong credentialing and audit trails to reinforce this result are also recommended.[304]

It must be recognized that no analytic process intended to help prevent future terrorist acts or otherwise predict human behavior is going to be completely effective, regardless of whether the tools used rely to some extent on computational automation, or actuarial methods (for example, inductive reasoning, data mining and behavior profiling), or on traditional "manual" methods (for example, deductive reasoning, hunches and traditional profiling).[305] Recognizing these inherent limitations, however, argues for rigorous testing before implementation and applying strict procedural protections and oversight to new technologies, not for resisting research and development or technology adoption for specious reasons based on presupposed outcomes.

If these technologies are "effective enough to sell books" (as well as designing drugs, investigating scientific phenomena, detecting fraud, diagnosing illness, and a myriad of other uses) why would we not explore their use to protect domestic security in the face of the significant and real threats from terrorism? This can be done even while recognizing that the consequences of error (both false positives and false negatives) in this context can be severe and while insisting that appropriate safe-guards for error rates and other privacy concerns be built in.[306] The argument that the consequences of error in selling books is significantly less than in domestic security applications is an argument for stricter criteria for development, testing, and implementation of these applications – not for abandoning research or outlawing technologies.

[303] *See, e.g.*, Jensen*,* Technical Assessment*, supra* note 89.

[304] *See generally* Markle Report, *supra* note 17 at 22—24; Rosenzweig, *supra* note 26, at 20—21.

[305] Nevertheless, there are contexts in which probabilistic automated analysis has shown itself superior to human judgment, for example, in certain medical diagnostic applications. *See* Fayaad et al., Mining,, *supra* note 16. Therefore, to dismiss out-of-hand the potential for these technologies for domestic security consideration seems at best premature.

[306] *See* Bray, *supra* note 23 (Kim Taipale commenting that "In the real world we use these technologies all the time. . . . If the tool is effective enough to sell books, shouldn't we at least look at it as away to fight crime?"). The point is that we increasingly use these technologies every day to inform significant and consequential commercial, economic, political, medical and other decisions in both the private and public sector. To not explore their use for domestic security or law enforcement is unrealistic. *See also* MacDonald, *supra* note 46, "to say that we cannot even go forward with the preliminary research to see if it works is a mistake."

E. *Security Risks: Rogue Agents and Attackers*

Finally, it is argued that a vast data integration system provides the potential for misuse by insiders or attack from outsiders.[307] However, "to view the potential for misuse as a basis for rejecting new technologies is . . . to despair of technological change and improvement."[308] There is no question that more powerful and more efficient tools can potentially be used for unintended or unauthorized purposes. Nevertheless, there are structural and technical features that can help address these concerns as well.

First, as described above,[309] a distributed database architecture supports certain privacy protections by diversifying control and eliminating a single point of attack. A distributed system permits local institutional control over individual databases and, to some extent, local accountability. Local access control and individual privacy rules can be negotiated, enforced, and tracked at many points in the system. There is no single point of control to be exploited either by attack or by misuse. Additionally, a distributed system provides multiple audit trails under independent control.

Second, institutional separation between intelligence analysis and enforcement action can be designed to lessen power aggregation. Separation could be either between agencies, for example, as suggested in the Markle Report,[310] or within agencies.[311] Separation of function permits imposition of administrative procedures and guidelines for use and imposes institutional barriers against abuse.

Third, strong credential and tamper-proof audit functions should be required for any implementation. Control of audit logs can be assigned to internal or external oversight bodies as well as to individual database owners, thus providing redundant and robust audit trails.

Finally, a key research goal should be (and was within TIA) developing additional security technologies to protect data and systems, including authentication and encryption protocols. Development of these technologies will have great benefit for other applications throughout society, including protection of vital information infrastructure from direct attack.

Addressing legitimate concerns for attack or misuse requires building in strict accountability for use – combining a distributed systems architecture and strong credential checking, distributed accountability and tamper-proof audit trial logging, together with strong laws against abuse and significant sanctions for misuse.

[307] *See, e.g.*, Public Policy Committee of the ACM, *supra* note 47.

[308] Rosenzweig & Scardaville, *supra* note 131.

[309] *See supra* notes 157—167 and accompanying text.

[310] *See supra* note 300 and accompanying text.

[311] *See, e.g.*, Rosenzweig, *supra* note 26, at 20—21 (suggesting internal administrative procedures for compartmentalizing use of these technologies in executive agencies).

Enforcement requires effective monitoring of usage. To help with the latter, automated traffic and usage analysis techniques should be developed and employed as an integral component of any analysis system. Existing data mining techniques can be used to profile (and spot) potential abuse or misuse.[312] Credential and audit features should seek to make "abuse difficult to achieve and easy to uncover."[313]

F. *Summary of Privacy Concerns*

This Article attempts to highlight the core privacy concerns related to the use of these technologies and to show how they intersect with actual likely technology development and implementation paths in order to identify those places where technology development strategies or structural design can help overcome or ameliorate privacy concerns. These technologies are being developed[314] and will be applied in domestic security and law enforcement circumstances.[315] Their use for these purposes raises significant and important privacy and civil liberties concerns. Understanding their weaknesses and strengths and how their characteristics potentially impact privacy and civil liberties in practice is essential in order to guide development towards security with privacy.

Part IV. Building in Technological Constraints - Code is Law

This article argues that privacy concerns relating to data aggregation and data mining in the context of domestic security can be significantly addressed through developing technologies that enable the application of existing legal doctrines and related procedures:

- ♦ First, that *rule-based processing* and a *distributed database architecture* can significantly ameliorate the general data aggregation problem by limiting the scope of inquiry and the subsequent processing and use of data within policy guidelines;

- ♦ Second, that *selective revelation* can reduce the non-particularized suspicion problem, by requiring an articulated particularized suspicion and intervention of a judicial procedure before identity is revealed; and

[312] Thuraisingham, *supra* note 69, at 46 ("[Data] mining can be used to detect intrusions as well as to analyze audit data. . . . One may apply data mining tools to detect abnormal activity."). Interestingly enough, it is in these areas, that is, complex network monitoring, that existing data mining technologies have been most successfully employed. *See, e.g.,* Mena, *supra* note 96, at 301—326.

[313] Rosenzweig, *supra* note 26, at 21.

[314] *See supra* notes 37—39, 48 and accompanying text.

[315] *See supra* Introduction.

- ♦ Finally, that misuse and abuse can be mitigated by employing *strong credential and audit features* and *diversifying authorization and oversight*.

Further, this Article contends that developing these features for use in domestic security applications will lead to significant opportunities to enhance overall privacy protection more broadly in the U.S. (and elsewhere) by making these technical procedures and supporting features available for voluntary or legislated adoption in the private sector.[316] In addition, the aggregation and analysis technologies themselves will have significant beneficial "spill-over" uses for commercial and scientific applications.

This section briefly describes three technical features that this Article has suggested can help ameliorate some of the privacy concerns raised in earlier sections. In particular, these technologies would allow existing legal or other procedural processes to be applied or adopted within a particular implementation in order to control their use within the bounds of existing due process procedures and relevant policy guidelines.[317] While it is beyond the scope of this Article to detail specific technology developments or current research avenues in depth, this section provides a broad overview suggesting thematic solutions to particular privacy problems together with some specific examples of technology research

It should be noted that the IAO's TIA-related programs expressly contemplated development of these technologies, in particular through the Genisys Privacy Protection program[318] and that DARPA has funded basic research in these areas for many years.[319]

> [T]he Genisys Privacy Protection Program [intended to] conduct R&D on technologies that enable greater access to data for security reasons while protecting privacy by providing critical data to analysts while not allowing access to unauthorized information, focusing on anonymized transaction data and exposing identity only if evidence warrants and appropriate authorization is obtained for further investigation, and ensuring that any misuse of data can be detected and addressed.[320]

A. *Rule-based Processing*

To the extent that privacy handling rules can be attached to data or to queries in machine readable form, it becomes possible to enforce privacy restrictions within the framework of automated processing systems. Rule-based processing has two aspects:

[316] *See* Zarsky, *supra* note 96 (discussing the impact of using data mining technologies more generally in society).

[317] *See generally* Rosenzweig, *supra* note 26, passim (setting out an implementation framework premised on exploiting these technical features).

[318] *See* IAO Report, *supra* note 88, at A-12—A-13; *supra* notes 176—177 and accompanying text.

[319] *See* ISAT 2002 Study, *supra* note 55.

[320] IAO Report, *supra* note 88, at A-12.

- First, to the extent that an "intelligent agent" is used for a central query to distributed databases, that software agent must negotiate access and permitted uses with each database. For example, a query marked as pursuant to a warrant might have different local permissions assigned to it than one pursuant to subpoena or administrative authorization. So too, the query might have different access depending on whether the authorizing operator or agency had a particular security clearance or not.

- Second, data items themselves can be labeled with meta-data (data about the data) describing how that item must be processed. Thus, even if a data item is removed or copied to a central database, it retains relevant rules by which it must be processed. For example, a data item may be returned in encrypted form in which only subsequent processing under a warrant or pursuant to a security clearance is permitted. Alternatively, a particular data item may be labeled as belonging to a U.S. citizen or to a foreign national, or its original source labeled as from a particular government or commercial database. In each case different procedures developed to enforce particular policy decision and privacy concerns would apply in its subsequent processing.

Rule-based processing is dependant on research in such areas as *proof carrying code*, *data labeling* and *analytic filtering tools*. In addition, a formal language for expressing privacy rules across different systems must be developed.

1. Proof Carrying Code

In order for intelligent prospecting agents to gain access to a distributed site, the software agent itself must be able to carry certain specifications and proof that those specifications are met. In the example above, the software agent must exhibit to the distributed database that it is seeking access pursuant to a warrant and prove that it meets the technical requirements of that warrant in its search or processing. To do so requires developing technologies called *proof carrying code*.[321] Proof carrying code would allow the distributed (local) server to independently determine whether the query application complies with local privacy (or other) requirements or if it will perform within its stated parameters.

2. Data Labeling

Data labeling can occur either at the data record level or through use of a *wrapper*, that is, software code that contains the data item or record. Labeling (whether direct or through use of a wrapper) presents structure, meta-data, and references to the

[321] *See* George Necula & Peter Lee, *Safe, Untrusted Agents Using Proof-Carrying Code*, LNCS, June 1998 at 1419, *available at* http://www-2.cs.cmu.edu/~necula/papers.html#lncs-abstract.

processing application. Labeling can be static, that is, permanently assigned to the object, or synthetic, that is, assigned by the server to the item when it is requested. The purpose of the label is to specify to some application the rules under which processing can occur. Encrypted wrappers can be used to maintain secrecy except under specified conditions. Challenges to labeling within the context of domestic security involve how to handle derived data (data that is itself the result of a query) and legacy data (pre-existing data that has not been labeled). Possible solutions depend on research in *program semantics*,[322] technologies that interpret what the application requirements are and then label the data accordingly.

In the commercial sector much of the research in data labeling is currently related to digital rights management ("DRM") to protect intellectual property rights in digital assets.[323] In thinking about research avenues and potential applications, one can think about privacy issues involved in using distributed information for domestic security applications as "privacy rights management" issues. Thus, much of the conceptualization involved in rule-based processing for DRM is applicable to protecting privacy in the context of data aggregation and analysis.[324]

3. Analytic Filtering

An example of analytic filtering technology is the Department of Justice developed DCS-1000 application ("Carnivore").[325] DCS-1000 is an analytical filtering tool designed to scan email traffic and only pick out that material that is authorized under the particular search warrant pursuant to which it is being employed. According to the FBI web site:

> The Carnivore device provides the FBI with a "surgical" ability to intercept and collect the communications which are the subject of the lawful order while ignoring those communications which they are not authorized to intercept. This type of tool is necessary to meet the stringent requirements of the federal wiretapping statutes.

[322] ISAT 2002 Study, *supra* note 55, at slides 16—17.

[323] For a general overview of a rule based (mediation) approach to digital rights management issues, see John S. Erickson, *Information Objects and Rights Management: A Mediation-based Approach to DRM Interoperability*, D-Lib Magazine (Apr. 2001), *available at* http://www.dlib.org/dlib/april01/erickson/04erickson.html.

[324] *See also* Cohen, *supra* note 25, at 609—617 (arguing for value sensitive design of DRM technologies). Obviously, the security concerns in domestic security applications may be more stringent.

[325] "Carnivore" was the original name chosen by its developers because it was designed to "get at the meat of the investigation," that is, to retrieve only the relevant data subject to court order. After much public criticism of the program, the FBI renamed the system DCS-1000. However, it appears that the FBI has gone back to using Carnivore. *See* FBI, Carnivore Diagnostic Tool, *at* http://web.archive.org/web/20030623092741/http://www2.fbi.gov/hq/lab/carnivore/carnivore.htm; *The IT Security.com Dictionary of Information Security, at* http://www.itsecurity.com/dictionary/carnivore.htm.

> The Carnivore device works much like commercial "sniffers" and other network diagnostic tools used by ISPs every day, except that it provides the FBI with a unique ability to distinguish between communications which may be lawfully intercepted and those which may not.[326]

The DCS-1000 application, like CAPPS II and TIA, has been the subject of much public debate and criticism from civil libertarians and libertarians.[327] Nevertheless, it is an example of analytic filtering technology.

Within IAO's Genisys program, a "privacy appliance" that would, among other functions, use analytic filtering to automatically screen seemingly harmless queries that might lead to identity inference or other privacy intrusions without conforming to established privacy protecting procedures was being researched.[328] Such a device might, for example, prohibit the search for a single unique identifier that would yield an identifiable return. In addition, the device would include controls for logging and audit.

A related area of research using Bayesian statistical methods for analytical content filtering has been proposed as the solution to the "spam"[329] problem.[330] Existing spam (and non-spam email) is "mined" (computationally analyzed) to develop an adaptive filter based on the Bayesian combination of spam probabilities of individual words. Reported accuracy rates are in the 99.94% range.[331]

[326] *See* FBI, Carnivore Diagnostic Tool, *at* http://web.archive.org/web/20030727211109/http://www2.fbi.gov/hq/lab/carnivore/carnivore2.htm.

[327] *See, e.g.*, Electronic Privacy Information Center, The Carnivore FOIA Litigation, *available at* http://www.epic.org/privacy/carnivore/. *See generally* Solove & Rotenberg, *supra* note 203, at 364—366; H. Judson Jennings, *Carnivore: U.S. Government Surveillance of Internet Transmissions*, 6. Va. J. L. & Tech. 10 (2001); Thomas R. McCarthy, *Don't Fear Carnivore: It Won't Devour Individual Privacy*, 66 Mo. L. Rev. 827 (2001).

[328] *See* Matthew Fordhal, *Researchers Seek to Safeguard Privacy in Anti-terrorism Plan*, Seattle Times, July 14, 2003, *available at* http://seattletimes.nwsource.com/cgi-bin/PrintStory.pl?document_id=135262838&zsection_id=268448455&slug=btprivacy14&date=20030714; *see also* IAO Report, *supra* note 88, at A-13 ("DARPA is examining the feasibility of a privacy appliance . . . to enforce access rules and accounting policy.").

[329] "Spam" refers to unsolicited commercial email (UCE), which is also known as bulk email. Spam is a spiced canned ham manufactured by Hormel (see an official statement on the use of the term for unsolicited email at http://www.spam.com/ci/ci_in.htm) as well as a famous Monty Python skit (see http://www.detritus.org/spam/skit.html). Internet lore traces the use of the word "spam" for unsolicited email to both origins (*see* http://www.turnstep.com/Spambot/glossary.html#spam and http://www.geocities.com/buddychai/Tips/Spam.html).

[330] *See* Paul Graham, *Better Bayesian Filtering*, *at* http://www.paulgraham.com/better.html (Jan. 2003); Paul Graham, *A Plan for Spam*, *at* http://www.paulgraham.com/spam.html (Aug. 2003).

[331] *See* Graham, *Better Bayesian Filtering*, *supra* note 330.

B. *Selective Revelation*

The goal of *selective revelation* is to protect against the revelation of personal information, that is, personally identifying information, while supporting data analysis.[332] This approach uses an iterative, layered structure that reveals personal data partially and incrementally in order to maintain subject anonymity. Initial revelation would be based on statistical or categorical analysis as described in earlier sections. This analysis would be applied to data that was sanitized or filtered in a way so that it did not reveal personally identifying information.[333] Based on initial results, subsequent revelations may or may not be justified. At each step, legal and technical procedures can be built in to support particular privacy policies (or other policies, such as security clearances, etc.).

For example, a directed link analysis based on characteristics of several known terrorists might produce a pattern or reveal additional relationships that appear relevant. A specific query could then be run using the pattern, for example, "search for other occurrences of large quantity chemical purchases and truck rentals." The algorithm would respond by confirming or denying that such other patterns exist without revealing personal identifying information of the transactions. Based on additional queries or subsequent analysis, the analyst could then request permission for additional revealing information (or to access additional data or databases). Depending on the nature and circumstance of the analysis, the type of information sought and from what source, the confidence interval of the pattern-match and other relevant factors, the appropriate legal or administrative procedures would be followed to permit additional revelation or access.

Where initial data analysis provides information that is in itself sufficient to meet investigative, reasonable suspicion, or probable cause standards, the relevant procedural protection – subpoena, warrant, or court order – could be applied depending on the particular circumstances before additional information or identity is revealed or before information is acted upon. In order to satisfy Fourth Amendment concerns, "interposition of a judicial officer before the barrier of anonymity is broken" should be required.[334]

Technology that allows for secure anonymous matching – that is, that can match data between separate databases without revealing the data itself – has already been developed.[335] Using one-way hash functions to convert data into unique but unreadable character strings allows these technologies to compare and update data without revealing the data itself. Thus, two databases – for example, a government watch list and a corporate database – can be compared without exchanging actual data. Only the one-way hash functions are compared and the result is a match without revealing the data. One-

[332] *See generally* ISAT 2002 Study, *supra* note 55, at slide 10; Project Genisys*,* DARPA (indicating that Genisys will: "[C]reate privacy filters, 'aliasing' methods, and automated data expunging agents to protect the privacy of U.S. citizens, and those not involved with foreign terrorists"), *previously available at* http://www.darpa.mil/iao/Genisys.htm.

[333] *Id.*

[334] Rosenzweig, *supra* note 26, at 15. *Cf.* Solove, *supra* note 236, at 1124—1128 (highlighting the significance of judicial imposition and warrant requirements in maintaining an "architecture of power" to protect privacy).

[335] *See, e.g.*, Mollman, *supra* note 291.

way hashes cannot be unhashed "any more than 'a sausage can be turned back into a pig.'"[336] Should the investigation or match warrant, the matching data records can be isolated from the original database without examining any other records.

C. *Strong Credentialing and Audit*

To protect against abuse by insiders (and to identify and track attacks by outsiders) strong tamper-proof audit mechanisms must be built into the architecture.[337] Audit trials must span many distributed databases so technology must be developed to make these logs not only tamper resistant, but also encrypted so that they themselves do not become subject to attack or inadvertent disclosure.[338] In addition, further work must be done to develop methods to search distributed databases without revealing the nature of the query or the results to interception or to the remote database administers.[339]

In order to control accountability and security, DARPA was examining the feasibility of using a *privacy appliance*:

> hardware and software that sits on top of a database, which is controlled by some appropriate oversight authority, and has mechanisms to enforce access rules and accounting policy. The idea is that this device, cryptographically protected to prevent tampering, would ensure that no one could abuse private information without an immutable digital record of their misdeeds. The details of operation of the appliance would be available to the public. Methods such as encryption and distribution would protect both ongoing investigations and the possibility of covering up abuse.[340]

To track where, when and by whom data is accessed, *self-reporting data* technologies (that is, data that, when accessed, reports its location and who is accessing it to an automated log or central tracking system) are also being developed.[341]

[336] *Id.*

[337] *See, e.g.*, Dmitriy Ayrapetov et al., *Improving the Protection of Logging Systems*, UC Berkeley Computer Sci., *available at* http://www.cs.berkeley.edu/~archanag/publications/privacypaper.pdf (last visited Dec. 15, 2003). Where tamper-proof is not technically possible or required, tamper-evident technology may be sufficient.

[338] *See* ISAT 2002 Study, *supra* note 55, at slide 13.

[339] Since that information itself may be valuable to enemies or invasive of privacy requirements for individuals whose data is being queried. Obviously, this is particularly important in connection with government access to commercial databases. *See, e.g.*, Dawn Song et al., *Search on encrypted data*, Proc. of IEEE SRSP, May 2000.

[340] *See* IAO Report, *supra* note 88, at A-13.

[341] *See* IAO Report, *supra* note 88, at A-12. For certain information, particularly private sector databases (for example, credit reports or medical records) it can be envisioned that such technology would eventually enable automatic notification to an individual if and when a third party accessed their private

D. *Additional Research Areas*

Additional research areas that transcend each of the above areas are general computer security, user authentication and encryption. Also, a common language for expressing privacy and other related rules across systems needs to be developed. Additional tools to check and report compliance need to be developed as well.

E. *Development Imperative*

Among the responses to the IAO Report[342] has been the notion that DARPA was paying "lip service" to privacy concerns and that technologies supporting privacy protection may not have been developed or effective.[343] However, it should be noted that the imperative to develop some of these features is inherent to government data sharing and is in the intelligence community's own best interests.

Government agencies are reluctant to share data because of security and liability concerns – both to protect the specific data item and to protect sources and methods that may be revealed from disclosure of the item. Thus, technical means to enforce privacy and security at the database access and data return level is vital for the agencies themselves. Essentially, there can be no effective distributed database architecture for sharing data securely without developing some version of these technology functions. This Article argues that it is vital to insist that privacy values, in addition to security concerns, be reflected in those developments.

Part V. Conclusion

New technologies present new opportunities and new challenges to existing methods of law enforcement and domestic security investigation and raise related civil liberties concerns. Although technology is not deterministic, its development follows certain indubitable imperatives. The commercial need to develop these powerful analytic technologies as well as the drive to adopt these technologies to help ensure domestic security is inevitable.

For those concerned with the civil liberties and privacy issues that the use of these technologies will present, the appropriate and useful course of action is to be involved in

information, essentially fully automating the existing "credit check monitoring" services that notify clients when someone accesses their credit report.

[342] *Supra* note 88.

[343] *See, e.g.*, EFF, *supra* note 164 (arguing that the IAO Report was "[l]ip service paid to privacy and civil liberties. . . . [T]he Report emphasizes privacy protection technologies, like automated audit trails, selective revelation, and anonymization. But the probable effectiveness of these technologies is not discussed.")

guiding the research and development process towards outcomes that provide opportunity for traditional legal procedural protection to be applied to their usage. To do so requires a more informed engagement by both sides in the debate based on a better understanding of the actual technological potential and constraints.

Overall implementation strategies and related policies must be developed under the principles that these technologies will only be used as investigative, not evidentiary, tools, and only for investigations of activities about which there is a political consensus that aggressive preventative strategies are appropriate.

Specific implementation and deployment must be subject to strict congressional oversight and review, be subject to appropriate administrative procedures within executive agencies where they are to be employed, and be subject to appropriate judicial review in accordance with existing due process doctrines.

Technology development strategies to support these principles must be incorporated into the development process itself. Specific technical features that protect privacy by providing opportunities for existing doctrines of due process and reinforcing procedures to function effectively, including rule-based processing, selective revelation, and secure credentialing and tamper-proof audit functions, must be developed.

Resistance to technological developments, particular research projects, on ideological grounds, or because the current generation of technologies cannot do what the equally ideological proponents claim they can do, is not a viable long-term strategy, particularly when faced with real and significant threats to security. While these security threats do not require (or justify) abandoning liberty to pursue security, neither does the pursuit of liberty justify abandoning security for absolute privacy. Development strategies that provide both are possible. A failure to develop and deploy technologies that take advantage of our national technical advantages is a dereliction of our responsibility to seek both liberty and security.

In any case, a failure to engage constructively with government research projects aimed at legitimate security needs will lead to having our civil liberties determined in the future in large part by technologies that were either developed to sell books or predict fashion, or developed through government research conducted in secret to avoid overblown criticism.[344] Only by insisting that important policy considerations relating to

[344] Although critics sometimes referred to the Terrorism Information Awareness project as a "secret Pentagon project" it was for the most part discussed openly and in public and included open discussion of the privacy concerns from the outset. See, for example, the presentation by Dr. John Poindexter, Director, Information Awareness Office, DARPA, at DARPATech 2002 Conference, Anaheim, California (Aug. 2, 2002), available at http://www.taipale.org/references/poindexter.html. Prior to the recent defunding of the IAO, detailed descriptions of TIA and its related projects (including the Genisys Privacy Protection project) were available from the DARPA web site at http://www.darpa.mil/oia/. Those documents have since been removed and, as discussed above in Part II, many of the TIA related projects unfortunately have now been moved to classified programs in intelligence agencies with little opportunity for public debate about their development.

A more appropriate criticism of TIA might be that it was "lacking in good public relations sense." *See* note 152 *supra*; Grant Gross, *U.S. agencies defend government data mining plans*, ComputerWorld, May 7, 2003 (quoting Congressman William Clay (D-Mo.) as saying, "I would like each of you, [TIA and CAPPS II program directors], to go back to your respective agencies and figure out what you can do to help build confidence in your activities . . . and make this process a little easier on the American public." Anthony Tether, director of DARPA, was also quoted as saying, "DARPA is not developing a system to profile the American public. Nothing could be further from the truth. The mistake we at DARPA made is we were so stunned by some of the outrageous comments that we didn't do anything about it for some time."), a*vailable* privacy and civil liberties be included in the development process itself at an early stage can one hope that a system subject to traditional human factor control – through laws, norms, and market forces – intersecting with technologically enabled intervention points will emerge. To the extent that code is law, the code must be developed with the same guiding principles as the related law and policy.

at http://www.computerworld.com/databasetopics/data/datamining/story/0,10801,81014,00.html; *see also*, Ted Levanthal, *Cyber Security: Experts Debate Federal Role in Protecting Cyber Networks*, Nat'l J.'s Tech. Daily, Oct. 21, 2003 :

> Former Defense spokeswoman Victoria Clarke . . . defended DARPA's controversial Terrorism Information Awareness (TIA) program . . . "TIA was a good idea, a good objective, but had lousy execution," she said. "The TIA people were tone deaf, did not explain their motives well, and the program was shut down. And that's unfortunate, in my opinion."