Available online at www.sciencedirect.com

SCIENCE DIRECT®

Government Information Quarterly 21 (2004) 461–480

ELSEVIER

Government Information Quarterly

Data mining and the search for security: Challenges for connecting the dots and databases[☆]

Jeffrey W. Seifert[*]

Congressional Research Service, Library of Congress, Washington, DC 20540-7450, USA

Available online 28 October 2004

Abstract

Data mining is emerging as one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. However, compared to earlier uses of data mining by government, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Three of the higher profile initiatives include the now defunct Terrorism Information Awareness (TIA) project, the recently canceled Computer-Assisted Passenger Prescreening System II (CAPPS II), and the Multistate Anti-Terrorism Information Exchange (MATRIX) pilot project. This article examines the evolving nature of data mining for homeland security purposes, the limitations of data mining, and some of the issues raised by its expanding use, including data quality, interoperability, mission creep, and privacy.
Published by Elsevier Inc.

[☆] The views expressed in this article are those of the author and do not necessarily reflect the position of the Library of Congress or the Congressional Research Services.

[*] Fax: +1 202 707 0781.
E-mail address: jseifert@crs.loc.gov.

0740-624X/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.giq.2004.08.006

1. Introduction

Since the September 11, 2001, terrorist attacks, government officials have continued to grapple with the questions of whether the attacks could have been prevented and what can be done to increase the government's awareness and knowledge of terrorist activity. As evidenced by congressional inquiries into so-called intelligence failures and the hearings held by the National Commission on Terrorist Attacks Upon the United States,[1] a significant amount of attention appears to be focusing on how to better collect, analyze, and disseminate information. In doing so, technology is commonly and increasingly looked upon as both a tool and in some cases a substitute for human resources.

One such technology that is playing a prominent role in homeland security initiatives is data mining. Similar to the concept of homeland security,[2] while data mining is widely mentioned in a growing number of bills, laws, reports, and other policy documents, an agreed upon definition or conceptualization of data mining appears to be generally lacking within the policy community. While data mining initiatives are usually purported to provide insightful, carefully constructed analysis, at various times data mining itself is alternatively described as a technology, a process, and/or a productivity tool. In other words, data mining, or factual data analysis, or predictive analytics, as it also is sometimes referred to, means different things to different people.

For example, in a proposed bill to require executive branch agencies to report on their data mining activities to Congress, a very contextually specific definition is offered. S.1544, the Data-Mining Reporting Act of 2003, identifies data mining as meaning:

> a query or search or other analysis of 1 or more electronic databases, where-
>
> (A) at least 1 of the databases was obtained from or remains under the control of a non-Federal entity, or the information was acquired initially by another department or agency of the Federal Government for purposes other than intelligence or law enforcement;
> (B) the search does not use a specific individual's personal identifiers to acquire information concerning that individual; and
> (C) a department or agency of the Federal Government is conducting the query or search or other analysis to find a pattern indicating terrorist or other criminal activity.[3]

In contrast, in its March 2004 report examining the Terrorism Information Awareness (TIA) project, the Department of Defense's Technology and Privacy Advisory Committee (TAPAC) purposely used a broad conceptualization of data mining to inform its research. The TAPAC report defines data mining to include "searches of one or more electronic databases of information concerning U.S. persons by or on behalf of an agency or employee of the government."[4]

In testimony before a House subcommittee in March 2003 regarding the use of data mining in government program audits, the General Accounting Office (GAO) defined data mining as "analyzing diverse data to identify relationships that indicate possible instances of previously undetected fraud, waste, and abuse."[5] However, just over a year