Science and Technology Law Review

Volume 16 | Number 2 Article 4

2013

# A Taxonomy and Classification of Data Mining

Liane Colonna

# A Taxonomy and Classification of Data Mining

*Liane Colonna**

## I. Introduction

Data is a source of power, which organizations and individuals of every form are seeking ways to collect, control and capitalize upon.[1] Even though data is not inherently valuable like gold or cattle, many organizations and individuals understand, almost instinctively, that there are great possibilities in the vast amounts of data available to modern society. Data mining is an important way to employ data by dynamically processing it through the use of advancing technology.

The common usage of the term "data mining" is problematic because the term is used so variably that it is beginning to lose meaning.[2] The problem is partially due to the breadth and complexity of activities referred to as "data mining." This overuse, especially from the perspective of those lacking a scientific background, creates a befuddlement and alienation of the topic. As such, individuals seem to haphazardly refer to data mining without a genuine understanding of what this technology entails.

This paper seeks to demystify data mining for lawyers through a clarification of some of its intricacies and nuances. The goal is to explain how data mining works from a technological perspective in order to lay a foundation for understanding whether data mining is sufficiently addressed by the law. A central ambition is to look beyond the buzzword and to take a realistic view of the core attributes of data mining. In an effort to understand if there is a need for new legal models and solutions, particular attention will be paid to exploring whether data mining is a genuinely new concept or whether it is a case of "the emperor's new clothes."

Another equally important goal is to establish a common vocabulary between the various stakeholders and to create a more precise use of the language of data mining. This approach is referred to as "semantic management." Ostensibly with a more refined set of terms for understanding data mining comes the potential for a greater understanding of the concerns that arise from the technology.

This paper begins with a general introduction to data mining. Then, it attempts to differentiate data mining from other closely related fields and

---

* Liane Colonna is a third-year doctoral candidate at the Swedish Law and Research Institute located at Stockholm University. She is writing her dissertation on the legal implication of data mining.

1. *See generally* Alexander Furnas, *Everything You Wanted to Know About Data Mining but Were Afraid to Ask*, The Atlantic (Apr. 3, 2012), http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388.

2. Liane Colonna, *Data Mining and the Need for Semantic Management*, *in* Internationalisation of Law in the Digital Information Society: Nordic Yearbook of Law and Informatics 2010–2012, at 335 (Dan Jerker B. Svantesson & Stanley Greenstein eds., 2013).

forms of data processing. Next, a basic taxonomy of data mining is provided. Finally, a classification of different data mining applications is afforded to the reader in an effort to highlight how data mining can be applied in different contexts.

## II. The Terminological Inexactitude of Data Mining

Because "data mining" is a nebulous term that is subject to numerous subjective definitions, its meaning frequently depends on the user and the context in which it is used. When used by a layman, the term data mining likely refers to a "[t]ype of database analysis that attempts to discover useful patterns or relationships in a group of data"[3] or "us[e of] mathematical formulas to sift through large sets of data to discover patterns and predict future behavior."[4] These definitions, while providing a useful starting point for understanding the concept, tend to oversimplify data mining.

In 1992, Frawley and Piatetsky-Shapiro, first used the term data mining in a scientific context when they referred to it as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."[5] Later, Han and Kamber elaborated by referring to data mining as "the extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) information or patterns from data in *large* databases."[6] Han and Kamber also opine that data mining is a complete misnomer because the goal is to mine for knowledge, not merely data.[7] Notwithstanding their suggestion that data mining would have been more appropriately named "knowledge mining from data," they adopted the phrase in the title of their well-known

3. *Data Mining*, Merriam-Webster Dictionary, http://www.merriam-webster.com/dictionary/data%20mining (last visited Sept. 1, 2013).

4. *Learning to Live with Big Brother*, Economist (Sept. 29, 2007), http://www.economist.com/node/9867324. *See also Data Mining*, Wikipedia, http://en.wikipedia.org/wiki/Data_mining (defining data mining as "the analysis step of the 'Knowledge Discovery in Databases' process, or 'KDD,' a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets.") (last visited Aug. 26, 2013).

5. W. Frawley, G. Piatetsky-Shapiro & C. Matheus, *Knowledge Discovery in Databases: An Overview*, AI Magazine, Fall 1992, at 57, 58.

6. Jiawei Han & Micheline Kamber, Data Mining: Concepts and Techniques 5 (Diane D. Cerra et al. eds., 2001). *See also* David J. Hand et al., The Principles of Data Mining (Mass. Inst. of Tech. 2001), *available at* ftp://ftp.sbin.org/pub/doc/books/Principles_of_Data_Mining.pdf (defining data mining as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.").

7. Han & Kamber, *supra* note 6, at 5.

book *Data Mining: Concepts and Techniques* because of the term's popularity.[8]

Some scientists, such as Harper and Jonas, have crafted more narrow definitions that focus solely on the predictive nature of data mining.[9] Other scientists, namely coming from the field of statistics, adopt a definition of data mining that emphasizes the importance of mathematical considerations.[10] Rather humorously, some scientists define data mining as "[t]orturing the data until it confesses . . . and if you torture it enough, you can get it to confess to anything."[11] Still other scientists, those coming from a machine-learning perspective, contend that data mining is the application of machine learning *per se*.[12]

The definition of data mining is also subjected to various legal interpretations. A 2004, report from the U.S. General Account Office ("GAO") defined the data mining as: "The application of database technology and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results."[13]

This definition creates ambiguity with respect to the extent that data warehousing, online analytical processes (OLAP), and data visualization, described more in depth below, are included within the meaning of data mining.

In 2004, the Department of Defense Technology and Privacy Advisory Committee also produced a definition of data mining. This definition includes "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."[14] This placed very basic forms of data processing, such as search and query analysis, within the definition of data mining.

Subsequently, in 2005, the U.S. Congressional Research Service defined data mining as "the use of sophisticated data analysis tools to *discover* previously unknown, valid patterns and relationships in large data sets. . . [such

8. *Id.*

9. Jeff Jonas & Jim Harper, *Effective Counterterrorism and the Limited Role of Predictive Data Mining*, 584 POL'Y ANALYSIS 1, 2 (2006), *available at* http://www.cato.org/sites/cato.org/files/pubs/pdf/pa584.pdf.

10. Jeff Jonas, *What is Data Mining? Depends Who You Ask . . .*, JEFF JONAS (Sept. 08, 2006), http://jeffjonas.typepad.com/jeff_jonas/2006/09/what_is_data_mi.html.

11. *Id.*

12. Christoph Schommer, *A Unified Definition of Data Mining*, CoRR (Sept. 16, 2008), http://arxiv.org/pdf/0809.2696.pdf.

13. U.S. GEN. ACCT. OFF., GAO-04-548, DATA MINING: FEDERAL EFFORTS COVER A WIDE RANGE OF USES (2004).

14. U.S. DEP'T OF DEF. TECH. & PRIVACY ADVISORY COMM., SAFEGUARDING PRIVACY IN THE FIGHT AGAINST TERRORISM 4 (2004).

as] statistical models, mathematical algorithms, and machine learning methods."[15] It further elaborates that data mining "consists of more than collecting and managing data, it also includes analysis and prediction."[16] This definition makes a helpful distinction between "data mining" and simpler forms of data processes. However, this definition fails to address whether quasi-advanced techniques, such as OLAP that are not discovery or "data driven," constitute data mining.

Then, the 2007 Federal Agency Data Mining Reporting Act ("the Act"), which requires federal agencies to report to Congress every year on their data mining activities, also defined data mining. Interestingly, the Act narrowly defined the term by limiting it to predictive, pattern-based analyses, despite the breadth of the above-mentioned definitions floating around the United States government.[17] In other words, the Act's definition is limited to "searches, queries or analyses" that are conducted for the purpose of identifying predictive patterns or anomalies.[18] This implies that research in electronic databases producing a summary or description of the data set is not data mining.

## III. The Need for Semantic Management

The semantic muddle surrounding the concept of data mining has legal implications. For example, if a statute's definition of data mining does not precisely use words such as "prediction" and "description," it may create easily exploited legal loopholes. Conversely, it may be important to distinguish data mining from other forms of data processing, in order to avoid an overly inclusive statute.

Other legal implications of the semantic mismanagement of data mining include: failure on behalf of the law to respond to important legal concerns raised by the technology, and unintentional gaps in protection resulting from an overly simple notion of data mining. For example, because the level of automation/human intervention that occurs in data mining is not necessarily consistent across different systems, laws that limit the role of automated decision-making may not apply data mining in a sufficiently obvious manner. Furthermore, laws that aim to protect individuals from the misuse of their personal data might be inapplicable because of the fact that data mining often creates fragments and abstractions of data that may not obviously constitute "personal data." Even assuming the applicability of privacy and data protection laws, they still may fail to provide sufficient safeguards against data

15. J.W. Seifert, Cong. Research Serv., RL31798, Data Mining: An Overview (2005).

16. *Id.*

17. Federal Agency Data Mining Reporting Act of 2007, 42 U.S.C. § 2000ee-3 (Supp. I 2007).

18. *Id.* at § 804 (b)(1)(A).

mining because of a fundamental misunderstanding of the nature of the technology.[19]

The lack of clarity with respect to the way the term data mining is applied creates a shaky base for the law's application. In order to think intelligently about the regulation of data mining, there is a critical need to better articulate what precisely this technology actually concerns and in what particular contexts it is applied. With a common vocabulary to express the particular features of data mining and more contextual understanding of its applications, it becomes possible to think more intelligently about the law surrounding data mining and the interests at stake.

## IV. A Data Mining Anecdote

At the very outset, it is useful to start with a *canonical anecdote* about data mining in order to better understand the potential of this advancing technology. According to the story, after mining sales records in a point-of-sales system for a large supermarket company, a marketing manager discovered that "customers who purchase diapers are likely to also purchase beer."[20] This example is famous among data mining analysts, since it is an example of unpredictable knowledge found in a huge dataset. After all, no one expected the relationship between diapers and beer to emerge.[21] The story continues by explaining how the marketing manager was able to exploit this unpredictable knowledge in his/her decision making process. After discovering the beer/diaper pattern, the marketing manager decided to place the beer shelves closer to the diapers shelves, which resulted in a significant increase in beer sales.[22] Through data mining, the supermarket company was able to achieve a competitive advantage over others without such knowledge.[23]

## V. Distinguishing Data Mining from Other Disciplines

In 1995, the first ACM Conference on Knowledge Discovery and Data Mining was held in the United States.[24] Six years later, MIT's Technology

19. Liane Colonna, *Data Mining and Its Paradoxical Relationship to the Purpose Limitation Principle, in* Computers, Privacy and Data Protection - Reloading Data Protection (Serge Gutwirth et al. eds.) (forthcoming Spring 2014).

20. Deheon Lee & Myong Ho Kim, *Data Mining, in* Database and Data Communication Network Systems, (Corneilius T. Leondes ed., Elsevier Science 2002); K.A. Saban, *The Data Mining Process: At a Critical Crossroads in Development*, 8:2 J. Database Mktg. 157, 158 (2001).

21. Lee & Ho Kim, *supra* note 20, at 42.

22. *Id.*

23. *Id.*

24. Illhoi Yoo et. al., *Data Mining in Healthcare and Biomedicine: A Survey of the Literature*, 36:4 J. Med. Sys. 2431, 2431 (Aug. 2012) ( "The first ACM Con-

Review listed data mining as one of the ten emerging technologies that would change the world.[25] Today, references to data mining are ubiquitous and the so-called "data scientist," a person who can find golden nuggets of knowledge from mountains of data, has been called one of the "sexiest" jobs of the twenty-first century.[26]

Data mining is an extension of traditional data analysis and statistical approaches.[27] It incorporates analytical techniques drawn from a range of disciplines including, but not limited to statistics, visualization, pattern recognition, and areas of artificial intelligence such as machine learning and neural networks.[28] This interdisciplinary nature of data mining has created confusion with respect to the terminology that surrounds it.[29] The result is an imbroglio of perspectives, vocabularies, and analytical tools, where concepts are easily misinterpreted, conflated, or used imprecisely, especially by scientific outsiders. Important perceptions about the technology are lost in translation and situations are created where an individual may think that he or she is talking about data mining when he or she is technically discussing something else such as knowledge mining from databases, pattern analysis, big data, or analytics.[30]

Accordingly, this section seeks to do the following: distinguish data mining from other closely related scientific fields, bring a more precise meaning to the term data mining, and highlight data mining as a distinct area

---

ference on Knowledge Discovery and Data Mining (a.k.a. SIGKDD) was held in the USA in 1995, and the term 'Data Mining' was first registered for the 2010 Medical Subject Headings (MeSH 1) in late 2009.").

25. *The Technology Review Ten*, MIT TECH. REV., Jan.-Feb. 2001, at 97.

26. *Data, Data Everywhere*, ECONOMIST (Feb. 25, 2010), http://www.economist.com/node/15557443.

27. Joyce Jackson, *Data Mining: A Conceptual Overview*, 8 COMMC'NS ASS'N INFO. SYS. 267, 267–68 (2002).

28. *Id.*; *See also* JIAWEI HAN & MICHELINE KAMBER, DATA MINING: CONCEPTS AND TECHNIQUES 9 (Morgan Kaufman Publishers, 2d ed., 2006) ("Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing) and advanced data analysis (involving data warehousing and data mining). . . . With numerous database systems offering query and transaction processing as common practice, advanced data analysis has naturally become the next target.").

29. HAN & KAMBER, *supra* note 6, at 29 ("[D]ata mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.").

30. *See id.* at 5.

of research. Brief references to legal concerns raised by the technology are made throughout the section to raise awareness about some of the issues.

### A. Knowledge Discovery in Databases

The phrase "knowledge discovery in databases" ("KDD"), coined by Gregory Piatetsky-Shapiro in 1989, has a broader meaning than data mining. KDD denotes the entire process of using unprocessed data to generate information that is easy to use in a decision-making context.[31] While there are numerous ways to classify the steps in KDD, Fayyad et al. organized it into five segments: data selection, data preprocessing, data transformation, data mining, and interpretation/evaluation.[32] It is important to remember that while there may be a nominal canonical order, the entire process of KDD is iterative and dynamic; therefore, analysts can jump from one phase to another in an instant.[33]

Because of the similarity of the processes of data mining and KDD, the two terms have been used interchangeably. However, many researchers believe data mining to be only one of the major steps of the KDD process.[34] Thus, data mining can be considered as part of the KDD process.[35] More specifically, data mining in the KDD process involves choosing the data-mining task and technique/algorithm and then applying it to the cleaned data set in order to discover certain previously unknown characteristics of the data.[36] This will be elaborated upon below.

From a legal perspective, making the distinction between KDD and data mining is important because KDD includes not just the processing of data, but it also includes the selection, storage, and human interpretation of the results. Data mining, however, in the very narrow sense, technically only includes the processing of the data through largely automatic means. This is a distinction that has consequences in the law. For example, different rules attach to the storage of data and the processing of data. Furthermore, by blurring the two together, important understandings about the level of automation involved in the processes might be lost.

---

31. NAT'L RESEARCH COUNCIL, COMM. TECHNICAL & PRIVACY DIMENSIONS OF INFO. FOR TERRORISM PREVENTION & OTHER NAT'L GOALS, PROTECTING INDIVIDUAL PRIVACY IN THE STRUGGLE AGAINST TERRORISTS: A FRAMEWORK FOR PROGRAM ASSESSMENT 186 (National Academic Press, 2008).

32. U. Fayyad et. al., *From Data Mining to Knowledge Discovery in Databases*, AI MAGAZINE, Fall 1996, at 41.

33. HAN & KAMBER, *supra* note 6.

34. *Id.* at 666.

35. Usama Fayyad et al., *supra* note 32, at 39.

36. Usama Fayyad et al., *From Data Mining to Knowledge Discovery: An Overview, in* ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING 9 (Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth & Ramasamy Uthurusamy eds.,1996).

## B. Statistics

Data mining uses statistical tools and methods, but it differs from statistics in several ways. First, data mining is usually conducted on huge volumes of data, whereas statistics generally uses a small sample of data.[37] Hand explains, "[s]tatisticians have typically not concerned themselves with data sets containing many millions or even billions of records."[38] Huge data sets lead to problems with which statisticians have not usually had to deal with in the past. For example, despite the recent dramatic increases in memory capacity, data will not always fit into the main memory of the computer.[39] In this respect, data mining can be separated from statistics because data mining focuses much more on scalable techniques that work for very large datasets than statistics.[40]

Furthermore, while most of the data in statistics are "flatted" (i.e. two-dimensional), various forms of data, often with a high dimensionality, can be mined (such as text, audio, video, etc.).[41] Likewise, the data applied in a data-mining situation is often in a state of constant evolution, which is different from the conventional statistical situation.[42] For example, supermarket transactions or phone calls occur every day and may need to be analyzed in real-time to bring an added value in the decision-making context.[43]

Data mining is generally "secondary" data analysis, which means the data has already been collected for some other purpose: the data is often a byproduct of an operational system.[44] Statistics, on the other hand, is "primary" data analysis because it controls the data it collects.[45] As such, data mining tends to rely on incomplete data to a larger extent than statistics. For instance, it is likely for the real-world data to be mined to have missing

37. Yoo et al., *supra* note 24, at 2432 (explaining that while statistics typically uses a sample of data (a few thousand records at most) drawn from a population, data mining typically uses data encompassing the entire population).

38. D.J. Hand, *Data Mining: Statistics and More?* 52 AM. STATISTICIAN 112-118 (1998).

39. *Id.*

40. Toon Calders & Bart Custers, *What is Data Mining and How Does it Work, in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY: DATA MINING AND PROFILING IN LARGE DATABASES 29 (2013).

41. Yoo et al., *supra* note 24.

42. Hand, *supra* note 38.

43. *Id.*

44. Calders & Custers, *supra* note 40, at 28. "Unlike in statistics, where the data is collected specially with the purpose of testing a particular hypothesis, or estimating the parameters of a model, in data mining one usually starts with historical data that was not necessarily collected with the purpose of analysis, but rather as a by-product of an operational system."

45. *Id.* at 44.

values and noise, while traditional statistics seldom considers it.[46] It has been remarked that "[d]ata mining results are inherently soft or fuzzy as the data is generally both incomplete and inexact."[47]

Data mining can generate hypotheses on its own from data-driven discovery instead of having the user state which hypothesis needs to be checked against the data.[48] As such, hypotheses generated by data mining do not have the same status as those in statistics.[49] That is, they may be considered less reliable because they lack a basis in sound theory and could occur purely by chance.

Yoo et.al. explain that statisticians follow the conventional scientific method.[50] That is, they construct a hypothesis, collect data, and then test the hypothesis on the data collected. This involves a process of reasoning from the general (i.e. a hypothesis) to the specific (i.e. data).[51] Unlike the conventional scientific method, the data mining method involves an exploration of a dataset without a hypothesis in order to discover hidden patterns from data. This involves a process of producing the general (i.e. knowledge or an evidence-based hypothesis) from the specific (i.e. data).[52] The automation of the scientific inquiry means that data mining is not limited by the creativity of humans to come up with a relevant hypothesis.[53]

Despite these differences, it is important to remember that statistical procedures play a major role in data mining. This is particularly true when it comes to developing and assessing models. Most of the machine-learning algorithms use statistical tests when they construct rules and when they correct models.[54]

---

46. HAN & KAMBER, *supra* note 6, at 289.

47. Lawrence J. Mazlack, *Imprecise Causality in Mined Rules*, *in* 2639 LECTURE NOTES IN COMPUTER SCIENCE 581 (Guoyin Wand ed., Qing Liu, Yiyu Yao & Andrzej Skowron eds., 2003).

48. Calders & Custers, *supra* note 40, at 28.

49. *Id.*

50. Yoo et al., *supra* note 24.

51. *Id.*

52. *Id.* at 2433.

53. Paul de Hert & Rocco Bellanova, *Data Protection from a Transatlantic Perspective: the EU and US Move Towards an International Data Protection Agreement?*, CIVIL LIBERTIES, JUSTICE & HOME AFFAIRS 22 (2008), http://works.bepress.com/rocco_bellanova/7; *see also* Steven Wiley, *Hypothesis-Free? No Such Thing: Even So-called 'Discovery-Driven Research' Needs a Hypothesis to Make Any Sense*, SCIENTIST (May 1, 2008), http://www.the-scientist.com/?articles.view/articleNo/26330/title/Hypothesis-Free—No-Such-Thing.

54. Jackson, *supra* note 27.

There are major legal implications from the fact that data mining relies on secondary data analysis and can generate hypotheses automatically. This technological reality calls into question some of the basic fundaments of data protection law. For example, the purpose limitation principle cannot be met because a data miner cannot meaningfully inform the individual of a specific and legitimate purpose for the data processing in advance of data mining.[55] Likewise, since data mining relies on secondary data analysis, which challenges rules concerning the accuracy of data, the mined data is incomplete, missing values, and taken out of context.

### C. Visualization

As the name suggests, visualization targets the visual representation of large-scale data collections to help people understand and analyze information.[56] It seeks to provide a picture of discovered patterns and relationships in various forms in order to engage the information-processing abilities of human analysts.[57] Ware explains:

> Information visualization enables mental operations with rapid access to large amounts of data outside the mind, enables substituting of perceptual relation detection for some cognitive inferencing [sic], reduces demands on user working memory, and enables the machine to become a co-participant in a joint task, changing the visualizations dynamically as the work proceeds.[58]

Data visualization is related to data mining to the extent that it concerns an important way for identifying and understanding the deeper relationships discovered with data mining. Information visualization ("infovis"), for example, permits the exploration of data before modeling.[59] Bertini and Lalanne explain:

> While information visualization . . . targets the visual representation of large-scale data collections to help people understand

55. Colonna, *supra* note 19.

56. Enrico Bertini & Denis Lalanne, *Surveying the Complementary Role of Automatic Data Analysis and Visualization in Knowledge Discovery*, VAKD PROC. ACM SIGKDD WORKSHOP ON VISUAL ANALYTICS AND KNOWLEDGE DISCOVERY: INTEGRATING AUTOMATED ANALYSIS WITH INTERACTIVE EXPLORATION, 2009, at 12, 16.

57. COLIN WARE, INFORMATION VISUALIZATION: PERCEPTION FOR DESIGN xvii (2d ed. 2004) ("What information visualization is really about is external cognition, that is, how resources outside the mind can be used to boost the cognitive capabilities of the mind. Hence the study of information visualization involves analysis of both the machine side and the human side. Almost any interesting task is too difficult to be done purely mentally.").

58. *Id.*

59. Bertini & Lalanne, *supra* note 56.

> and analyze information, data mining, on the other hand, aims at extracting hidden patterns and models from data, automatically or semi-automatically. In its most extreme representation, infovis can be seen as a human-centered approach to knowledge discovery, whereas data mining is generally purely machine-driven, using computational tools to extract automatically models or patterns out of data, to devise information and ultimately knowledge.[60]

The core of infovis combines human flexibility, creativity, and general knowledge with computer storage capacity and computational power during the knowledge discovery process. Keim elaborates, "[t]he basic idea . . . is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data."[61] He further explains that unlike data mining, "visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters."[62]

Understanding the relationship between visualization and data mining is important from a legal standpoint because it concerns the level of automation that takes place in data mining. That is, even if data mining is highly automated, a human still has a role in the interpretation of the end result. An understanding of the role of humans in the process of data mining is critical because certain laws prohibit decisions being made by fully automated processes.

### D. Pattern Recognition

Pattern recognition is a task that is performed both by humans and by artificial systems.[63] It is one of the most important functionalities for intelligent behavior.[64] Broadly, the term pattern recognition "could cover any context in which some decision or forecast is made on the basis of currently available information."[65]

Majii et.al. define pattern recognition more narrowly as "the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions

---

60. *Id.*

61. Daniel A. Keim, *Information Visualization and Visual Data Mining*, 7:1 IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 100 (2002),

62. *Id.*

63. Azriel Rosenfeld & Harry Wechsler, *Pattern* Recognition: Historical Perspective and Future Directions, 11 INT'L. J. IMAGING SYS. TECH. 101 (2000).

64. *Id.*

65. Pradipta Maji & S.K. Pal, *Introduction to Pattern Recognition and Data Mining*, *in* ROUGH-FUZZY PATTERN RECOGNITION: APPLICATIONS IN BIOINFORMATICS AND MEDICAL IMAGING 1 (2012).

about the categories of the patterns."[66] Majii goes on to explain "the discipline of pattern recognition essentially deals with the problem of developing algorithms and methodologies that can enable the computer implementation of many recognition tasks that humans normally perform."[67] The motivation for developing machines to perform pattern recognition is to perform recognition tasks more accurately, faster, and economically than humans and to release humans from the "drudgery resulting from performing routine recognition tasks repetitively and mechanically."[68]

One of the key goals of data mining is to uncover previously unknown patterns in the data set. Data mining is a broader field than pattern recognition, which applies other types of techniques in addition to pattern recognition techniques.[69] It is important to keep this distinction in mind when applying law to data mining because limiting the definition of data mining to pattern recognition may result in an under-inclusive statute.

### E. Machine Learning

Machine learning is a broad subfield of artificial intelligence.[70] It is closely related to pattern recognition, and many of the same techniques apply. There is also a large overlap between machine learning and statistics.[71] For example, machine-learning algorithms have a sound mathematical basis, and many directly incorporate statistics into their algorithms. However, statistics is more oriented towards testing hypotheses, whereas machine learning is more centered on making predictions.[72]

More specifically, machine learning is concerned with the development of algorithms and techniques for building computer systems that can automatically improve with experience ("learn").[73] Computers learn new knowledge in a variety of ways, most notably from either supervised or unsupervised data sets.[74] Supervised learning induces models from sets of

66. *Id.*

67. *Id.*

68. *Id.*

69. NAT'L RESEARCH COUNCIL, *supra* note 31, at 188.

70. Tu Bao Ho, Saori Kawasaki & Janusz Granat, *Knowledge Acquisition by Machine Learning and Data Mining*, 59 STUDIES IN COMPUTATIONAL INTELLIGENCE (SCI) 69, 71 (2007).

71. V.G. Ivancevic & T.T. Ivancevic, *Introduction: Human and Computational Mind*, 60 STUDIES IN COMPUTATIONAL INTELLIGENCE (SCI) 1, 121 (2007).

72. Sally Jo Cunningham, *Machine Learning and Statistics: A Matter of Perspective* (1995), *available at* http://www.cs.waikato.ac.nz/ml/publications/1995/Cunningham95-ML-Stats.pdf.

73. Ho et al., *supra* note 70, at 71.

74. *Id.*

training data and these models can be used to classify other unlabeled data.[75] With unsupervised learning (or "learning without a teacher"), a form of inductive reasoning, there is no model or hypothesis before running the analysis: the aim is to identify and explore regularities and dependencies in data.[76] Other learning techniques include semi-supervised learning, reinforcement learning, "transduction," and "learning to learn," which will not be elaborated upon here.[77]

Machine learning and data mining are closely connected because they share the goal of finding novel and useful knowledge in data.[78] Also, a large amount of data mining applies machine-learning methods.[79] Machine learning has been very important in the context of the data-mining arena for many years, and it is increasingly proving more useful as the volume of data grows exponentially over time.[80]

A fundamental difference between machine learning and data mining exists in the volume of data being processed: the fact that data mining involves huge data sets is an important criterion with respect to distinguishing it from machine learning.[81] The type of input data is also different. In traditional machine learning the data is more static and error-free than in a data-mining scenario.[82] Moreover, data mining does not necessarily require rule or pattern extraction.[83] Finally, data mining seeks to extract data for human comprehension, whereas machine learning seeks to extract data to improve the program's own understanding.[84]

Understanding the role of machine learning in data mining is critical since many of the machine-learning techniques give rise to the largest societal concerns. This has to do with concerns about the way machines learn to make decisions. In other words, if the foundation for the machine-learning process is faulty then there is a greater likelihood for false positives, the

75. Padraig Cunningham, Matthieu Cord & Sarah Jane Delany, *Supervised Learning*, MACHINE LEARNING TECHNIQUES FOR MULTIMEDIA: CASE STUDIES ON ORGANIZATION AND RETRIEVAL, 21, 21 (2008).

76. Lars Kai Hansen & Jan Larsen, *Unsupervised Learning and Generalization*, PROC. OF THE IEEE INT'L CONF. ON NEURAL NETWORKS, 1 (1996).

77. Ivancevic & Ivancevic, *supra* note 71, at 122.

78. ROB SULLIVAN, INTRODUCTION TO DATA MINING FOR THE LIFE SCIENCES, 71 (Humana Press 2012).

79. *Id.*

80. *Id.*

81. Lee & Ho Kim, *supra* note 20, at 45; Ho et al., *supra* note 70, at 71.

82. Lee & Ho Kim, *supra* note 20, at 44.

83. Ivancevic & Ivancevic, *supra* note 71, at 121.

84. Margeret Rouse, *Machine Learning*, WHATIS.COM (Aug. 24, 2013), http://whatis.techtarget.com/definition/machine-learning.

generation of an incorrect inference, or false negatives, the failure to generate a crucial inference.[85]

## F. Information Retrieval

Information retrieval pertains to getting back or retrieving information stored in various data repositories in response to a defined or specified database query.[86] Holsheimer explains that in traditional information retrieval one must "talk" to a database by specifically querying for information.[87] An example of information retrieval using traditional query and report tools is the use of simple search queries in Lexis or Westlaw legal databases to find relevant cases and statutes.[88] Taipale explains, "[t]he 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."[89]

In the case of data mining, however, it is possible to just "listen" to a database because the data mining algorithms and techniques are capable of formulating thousands of hypotheses on their own.[90] The results are not explicit but rather are implicit and nonobvious: data mining reveals knowledge that does not exist *a priori*.[91] In other words, data mining and information retrieval can be distinguished based on the type of information need that is handled by each type of technology. Choenni et. al. explains that "[i]n general, data mining is capable of handling an information need that has a higher degree of vagueness and incompleteness than information retrieval."[92]

The fact that data mining automatically and serendipitously reveals novel and implicit information from datasets has critical implications for data protection law. This is because it can be hard (or at least impracticable) for data miners to provide notice to individuals, which is a basic requirement of many data protection laws, about the details of the new information that

85. J.W. SEIFERT, DATA MINING: AN OVERVIEW, CONG. RESEARCH SERV. RL31798, at 1 (2004).

86. K.A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 2, 22 (2003).

87. Lita van Wel & Lambér Royakkers, *Ethical Issues in Web Data Mining*, 6 ETHICS & INFO. TECH. 129, 131 (2004).

88. Lee Tien, *Privacy, Technology and Data Mining*, 30 OHIO N.U. L. REV. 389, 394 n.22 (2004).

89. Taipale, *supra* note 86, at 22.

90. Wel & Royakkers, *supra* note 87, at 131.

91. Taipale, *supra* note 86, at 22.

92. Sunil Choenni, Robin Bakker, Henk Ernst Blok & Robert de Laat, *Supporting Technologies for Knowledge Management*, 9 KNOWLEDGE MGMT. & MGMT. LEARNING 89, 91 (2005).

arises from mining the dataset.[93] Because an individual is likely to be unaware about the new information that is created as a result of data mining, he or she will unlikely be able to control the flow of this "swelling river" of personal data in direct contravention of many laws.[94]

## G. Information Fusion

Boström defines information fusion as "automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making."[95] Essentially, it involves exploiting the synergy in the information acquired from multiple sources such as sensors, databases, and information gathered by humans.[96] The goal is to produce a result that would be better than if such sources were used individually.[97]

There are, at least, two major differences between data mining and information fusion. First, data mining usually uses data from only one source: even if the data used for mining comes from several sources, it is typically preprocessed into one source so that data mining techniques and algorithms do not have to handle the problem with many sources.[98] Information fusion, however, strives to fuse information from several sources into one source.[99] This means that techniques from information fusion may very well be used prior to applying data mining, but they are not a prerequisite for data min-

93. Richard Huebner, *Barriers to Adopting Privacy-Preseving Data Mining*, NORWICH UNIV. (2006), *available at* http://www.aabri.com/OC2012Manuscripts/OC12006.pdf.

94. Gehan Gunasekara, *The 'Final' Privacy Frontier? Regulating Trans-Border Data Flows,* 17 INT'L J. L. INFO. TECH. 147, 158 (2009).

95. Henrik Boström, et al., *On the Definition of Information Fusion as a Field of Research*, ch. 3.1.1 (2007) *available at* http://www.his.se/PageFiles/18815/Information%20Fusion%20Definition.pdf; *see also* Belur V. Dasarathy, *Information Fusion - What, Where, Why, When, and How?*, 2 INFO.FUSION 75, 75 (2001) ("Information fusion encompasses theory, techniques and tools conceived and employed for exploiting the synergy in the information acquired from multiple sources (sensor, databases, information gathered by human, etc.) such that the resulting decision or action is in some sense better than would be possible, if these sources were used individually without such a synergy exploitation.").

96. *Id.*

97. *Id.*

98. T. Löfström, R. Konig, U Johansson, L. Niklasson, M. Strand & T. Ziemke, *Benefits of Relating the Retail Domain to Information Fusion*, PROC. OF THE 9TH IEEE INT'L CONF. ON INFO. FUSION 1 (2006).

99. *Id.*

ing.[100] The second difference between data mining and information fusion is "that information fusion techniques normally are applied to solve problems online, while data mining is most often used offline; i.e. time constraints are more important for information fusion."[101]

Zhou et. al further explains:

> Both information fusion and data mining are processes of generating knowledge. According to the size of the original data collection, data mining mainly uses artificial intelligence or statistical methods to speculate and search for the potential complexity relationship or the model contained in the data, rather than paying much attention to the data sources. Information fusion stresses the analysis of the information from multiple sensors or data sources, identifying and estimating the objectives or making a comprehensive judgment.

In the logical inference point, data mining and information fusion are two opposite processes. Data mining, which studies and summarizes knowledge from the original data, is a process of induction. Information fusion, which uses already existing knowledge and experience to deal with the data from different areas of the unknown world, is a process of deduction.[102]

From a legal perspective, information fusion and data mining are often clumped together which may not be intentional. The subtle differences between the two should be understood and possibly reflected in the law. The interplay between the two fields is relevant because the use of information fusion in data mining seems to imply a higher level of efficiency and invasiveness (if personal data is concerned) which may have legal consequences.

## H. High Performance Computing and Super Computing

Supercomputing or high-performance computing ("HPC") is used for high calculation intensive tasks, such as problems involving quantum mechanical physics, weather forecasting, global warming, and molecular modeling.[103] Xie et al. describes supercomputing as "the interface between hardware and software, which directly affects the scalability, program-

100. Professor Boström graciously helped explained the difference between data mining and information fusion in an email exchange. The record is on file with the author.

101. Löfström, *supra* note 98, at 1-2.

102. ZUDE ZHOU, SHANE (SHENGQUAN) XIE, DEJUN CHEN, FUNDAMENTALS OF DIGITAL MANUFACTURING SCIENCE, 172 (Springer 2012).

103. Qingyu Zhang & Richard S. Segall, *Commercial Data Mining Software, in* DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK 1245, 1246 (Oded Mairmon & Lior Rokach eds., 2d ed. 2010).

mability, and availability of the system."[104] Supercomputers have contributed to significant improvements in the context of national security, economic and social developments, and have also played a critical role in scientific discovery and innovation.[105] The value of supercomputing is exemplified by a procedure conducted by the Children's Hospital of Pennsylvania. In this procedure, the hospital took MRI scans of a child's brain in just seventeen seconds, while a similar procedure using supercomputers would have taken seventeen minutes — assuming the patient does not move.[106]

Because of the enormous size of the data and the amount of computation involved in data mining, high-performance computing is a critical component for data mining applications. In other words, data mining is not possible without sufficient computational resources because it is so data-intensive. This is especially true as data mining often involves the processing of data in real-time, which requires a high-performance system where computation can be completed fast. However, if the computation takes too long the analysis may become worthless. It has been stated that "[d]ata mining with these big, superfast computers is a hot topic in business, medicine and research because data mining means creating new knowledge from vast quantities of information, just like searching for tiny bits of gold in a stream bed."[107]

Understanding the relationship between supercomputers and data mining is important since it connects to the issue of information equality. Because supercomputers are expensive and not widely available to the public, those that have access to them – and thus the capacity to run highly efficient data mining operations – are in a position to abuse their power and may discriminate and profile people. As such, the law could reflect upon the issue of supercomputers on the one hand, and data mining on the other. The law could also take a more holistic approach; an easy answer to some of the societal concerns raised by data mining is to outlaw all supercomputers.

## I. Data Warehousing

Data warehousing is described as "computer systems designed to manage data for analysis and to assist management in decision making."[108] A data

104. Xianghui Xie et al., *Evolution of Supercomputers*, 4 FRONTIERS COMPUTER SCI. CHINA 428, 429 (2010).

105. *Id.* at 428.

106. Zhang & Segall, *supra* note 103, at 1246 (citing E. Sanchez, *Speedier: Penn Researchers to Link Supercomputers to Community Problems*, COMPASS (Sept. 17, 1996), *available at* http://www.upenn.edu/pennnews/features/1996/091796/research).

107. *Id.*

108. S. SUMATHI & S.N. SIVANANDAM, *Introduction to Data Mining and its Applications*, STUD. IN COMPUTATIONAL INTELLIGENCE 21, 22 (Janusz Kacprzyk ed., 2006).

warehouse is an infrastructure specifically designed for "query, analysis and reporting."[109] Gray explains:

> A question often asked is 'why make the expense and effort of keeping copies of data that exists in other systems?' Many people can see the value in making all your data available from a single source, but the real answer is primarily to do with structuring the data such that it is most suitable for analysis.[110]

Essentially, the construction of a data warehouse is an important pre-processing step for data mining that involves data cleaning and data integration.[111] The creation of a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task that takes years and costs millions of dollars.[112] Basically, data warehousing is the process of compiling and organizing data into one common database. Data mining, on the other hand, is the process of extracting meaningful data from that database.[113] Li explains, " . . . data mining is finding the proverbial needle in the haystack, where the needle is the desired piece of intelligence and the haystack is the large data warehouse which is built up over a long period of time."[114]

It is still important to remember that a data warehouse is not a requirement for data mining.[115] For example, instead of using a data warehouse, the data to be mined can be directly extracted from one or more operational or transactional databases.[116] While this might save time and money, it might not resolve all of the data integrity problems.[117]

Data warehouses are concerned with the collection and storage of data. Data mining – while reliant on data warehouses – is concerned with the actual processing of data. The difference between "collection," "storage," and "processing" of data often has important implications in the law; whether this should continue must be further reflected upon. It is also worth reflecting on

109. *Id.*

110. *Id.*

111. Jackson, *supra* note 27, at 267, 268.

112. Paul Gray & Hugh J. Watson, *Professional Briefings. . .Present and Future Directions in Data Warehousing*, 29 DATABASE FOR ADVANCES IN INFO. SYS. 83, 84–88 (1998).

113. HAN & KAMBER, *supra* note 6, at 5, 12.

114. Yongchang Li, *An Intelligent, Knowledge-based Multiple Criteria Decision Making Advisor for Systems Design* (Jan. 9, 2007) (unpublished Ph.D dissertation, Georgia Institute of Technology) (on file with Georgia Institute of Technology Library).

115. Jackson, *supra* note 27, at 268.

116. *Id.*

117. Li, *supra* note 114, at 81.

whether processing data, in highly structured data warehouses, should be subject to additional regulations because the efficiency could lead to greater invasion into individual's private life.

### J. Exploratory Data Analysis

The term "exploratory data analysis" characterizes the notion that statistical insights and modeling are driven by data.[118] These notions were reinforced in the early 1970s using a litany of ultra-simple methods by dispelling the traditional dogma that "one was not allowed to look at the data prior to modeling."[119] Data mining embraces the idea that, pictures and numerical summaries of data, are necessary to understand how rich a model the data can support.[120]

Some definitions of data mining consider it exclusively as a form of exploratory data analysis. Other definitions of data mining take a broader view and include hypothesis-driven techniques as well as data-driven techniques. Whether true data mining is always data driven, hypothesis driven, or both, it is a question that needs further inquiry.

### K. Online Analytical Processing

Online Analytical Processing ("OLAP"), is often used to describe the various types of query driven analysis taken when analyzing the data in a database or a data warehouse.[121] OLAP provides the selective extraction and viewing of data from different points of view, generally referred to as dimensions.[122] Jackson explains, "the traditional query and reporting tools describe 'what' is in a database, while OLAP is used to answer 'why' certain things are true in that the user forms a hypothesis about a relationship and verifies it with a series of queries against the data."[123]

The essential distinction between OLAP and data mining is that OLAP is a data summarization/aggregation tool, while data mining allows the automated discovery of implicit patterns and interesting knowledge that is hiding in large amounts of data.[124] In other words, the main difference between data

118. John F. Elder, IV & Daryl Pregibon, *A Statistical Perspective on Knowledge Discovery in Databases*, *in* ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING 83, 84 (Usama M. Fayyad et al. eds., 1996).

119. *Id.* at 84–85.

120. *Id.* at 85.

121. *See generally* MICHAEL J. A. BERRY & GORDON S. LINOFF, MASTERING DATA MINING: THE ART AND SCIENCE OF CUSTOMER RELATIONSHIP MANAGEMENT, 147–48 (Robert M. Elliot ed., 2000).

122. Usama Fayyad, *The Digital Physics of Data Mining*, COMM. OF THE ACM 62, 64 (2010).

123. Jackson, *supra* note 27, at 270.

124. HAN & KAMBER, *supra* note 6, at 94.

mining techniques and OLAP is that to use OLAP an analyst must know precisely what he or she is looking for, while in data mining an analyst needs only a vague idea about what he or she is looking to discover. OLAP is user-driven, meaning "it merely provides the user with the tools to quickly generate the aggregates in the data he or she selects to be displayed and presents them in a convenient display."[125]

OLAP answers questions such as: "Who are our top 100 best customers for the last three years?" and "Which customers defaulted on their mortgages last two years?"[126] Whereas data mining answers questions like: "Which 100 customers offer the best profit potential?" and "Which customers are likely to be bad credit risks?"[127]

Understanding the difference between OLAP and data mining is important because the two are often conflated. Sometimes when people are talking about data mining they are really talking about OLAP, and vice versa. It is also connected to the issue described above concerning whether real data mining is hypothesis driven or data driven.

## L. Analytics

Analytics is "the transformation of data to extract useful information and effectively draw conclusions" by the use of statistical modeling, selection of representative subsets of data, and curve fitting against an expected outcome, etc.[128] "Both analytics and data mining are aimed at information that is actionable."[129] However, analytics and data mining are different.

125. Calders & Custers, *supra* note 40, at 29; *see also* PAULRAJ PONNIAH, DATA WAREHOUSING FUNDAMENTALS: A COMPREHENSIVE GUIDE FOR IT PROFESSIONALS 405–06 (John Wiley & Sons, Inc. 2001) ("When an analyst works with OLAP in an analysis session, he or she has some prior knowledge of what he or she is looking for. The analyst starts with assumptions deliberately considered and thought out. Whereas in the case of data mining, the analyst has no prior knowledge of what the results are likely to be. Users drive OLAP queries. Each query may lead to a more complex one and so on. The user needs prior knowledge of the expected results. The process is completely different in data mining. Whereas OLAP helps the user to analyze the past and gain insights, data mining helps the user predict the future.").

126. PONNIAH, *supra* note 125, at 406.

127. *See id.*; *see also* Jackson, *supra* note 27, at 270 (". . . an analyst might want to determine the factors that lead to loan defaults. She might initially hypothesize that people with low incomes are bad credits risks and analyze the database with OLAP to verify or disprove this assumption.").

128. Louie Velocci & David Stewart, *Clarity Through Data—The Practicality of Forensic Data Mining for Valuators*, 2009 J. BUS. VALUATION 21, 23 (2009).

129. Sid Adelman et al., *What is the Difference between Analytics and Data Mining?*, INFO. MGMT. (Feb. 4, 2005), http://www.information-management.com/news/1019393-1.html.

"Analytics usually comes with hypotheses testing. The analyst has something in mind and is looking to answer a question and has a hypothesis about that question."[130] According to some, data mining is an "act of discovery that lacks a hypothesis."[131] In other words, "[d]ata mining is exploring data for trends that cannot be 'defined' where analytics is looking at data for trends that can be defined."[132] Also, data mining differs from analytics because it "uses more complex computer modeling, database analysis, and theoretical modeling which often requires a significant investment in software, computer hardware, and specialized data analysis resources."[133]

It is important to understand the difference between data mining and analytics because the two are often confused. Again, the issue is what is "real" data mining and whether the law should distinguish between hypothesis and data driven processing techniques.

### M. Big Data

Broadly, the term "big data" refers "to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."[134] Although big data is not a new concept, it has become a much-debated subject in recent years because of the development of the Internet, cloud computing and ubiquitous computing, among other things.[135] Furnas explains, "[n]early every transaction or interaction leaves a data signature that someone, somewhere is capturing and storing."[136]

Due to these developments, massive amounts of data are produced continually, which overwhelm traditional data processing technologies.[137] Because data is generated as a constant stream, this data has a huge volume, a heterogeneous, complex, variable nature, and a quick velocity.[138] The relationships in these data are complex and multi-dimensional.

Data mining facilitates the understanding of these very large and complicated data sets. It automates the process of understanding subtle patterns

130. *Id.*

131. *Id.*

132. *Id.*

133. Velocci & Stewart, *supra* note 128.

134. Ann Cavoukian & Jeff Jonas, *Privacy by Design in the Age of Big Data*, PRIVACY BY DESIGN 3 (June 2012), http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf.

135. ZhenXin Qu, *Semantic Processing on Big Data, in* 129 ADVANCES IN INTELLIGENT AND SOFT COMPUTING 43–48, at 43 (David Jin & Song Lin eds., 2012).

136. Furnas, *supra* note 1.

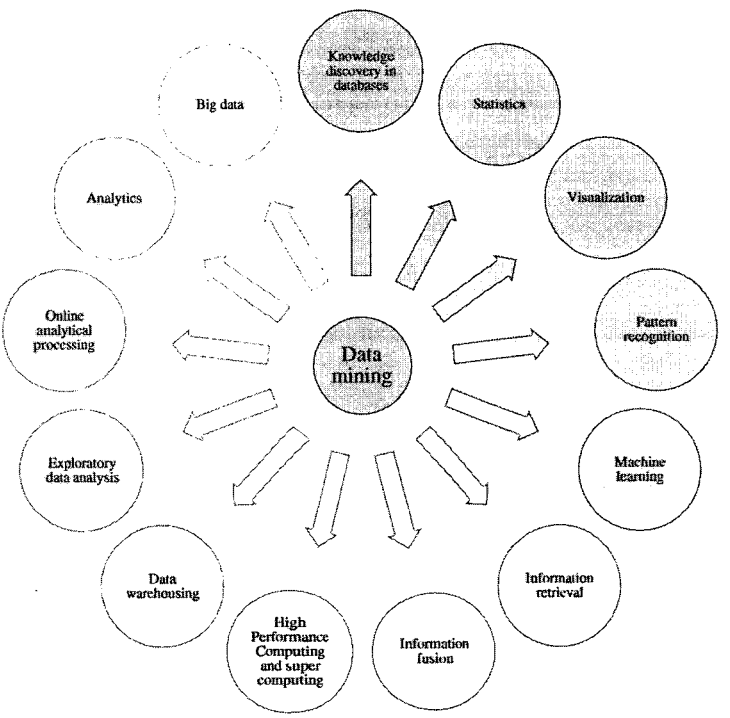
137. Qu, *supra* note 135.

138. Intel, *Vision Paper: Distributed Data Mining and Big Data: Intel's Perspective on Data at the Edge*, INTEL 3 (Aug. 2012), http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/distributed-data-mining-paper.pdf.

and relationships in these huge data sets.[139] It is "a way to see the forest without getting lost in the trees."[140] For example, it can reveal that "one of these things is not like the other," or it can reveal categories and then sort things into pre-determined categories.[141] What is simple with a megabyte of data, however, is not so simple with a terabyte of data, and this is where the nexus between data mining and big data becomes obvious.[142] Big data is not inherently useful. Rather, data mining provides value by uncovering patterns, correlations, and insights.[143]

Big data definitions vary. Just as with some of the other fields above, it is often confounded with data mining, which may result in confusion in the law. These two ideas, while related, are not the same thing. To emphasize, big data is the idea that we have an explosive growth of data; it is the challenge. Data mining, on the other hand, is the techniques and software that can be employed to make sense of the data; it is the solution.

**FIGURE 1: DATA MINING AND ITS CONNECTION TO OTHER DISCIPLINES**




139. Furnas, *supra* note 1.

140. *Id.*

141. *Id.*

142. *Id.*

143. Cavoukian & Jonas, *supra* note 134, at 5.

This figure is a visualization of data mining and its connection to other disciplines. There are other fields that have not been included here. It should not be viewed as an exhaustive picture, but rather as a starting point for understanding the nexus between data mining and other scientific fields. Also, there are many overlaps between the fields listed below, such as the overlap between the fields of pattern recognition and machine learning. These overlaps have not been discussed because the goal here is to understand the relationship between data mining and related fields, and not necessarily to understand the relationship among all of the fields.

## VI. The Process of Data Mining: High-level Goals, Pre-mining Tasks, Algorithms, Heuristics, Data Mining Tasks, Techniques, Models, Patterns and Hypotheses and Decision Making

Before data mining can take place, several important pre-mining tasks must be completed in order to prepare the data to be successfully mined and to remove imperfections in the dataset. As mentioned above, these tasks are part of the broader process of KDD. For example, these tasks include identifying a specific goal that would benefit from the application of data mining to a dataset and then obtaining all the required data. Other pre-processing tasks include cleaning the data by removing noisy (i.e. meaningless) and inconsistent data.

Data mining technically consists of applying a data-mining algorithm to the cleansed dataset to discover hidden patterns and relationships. A data-mining algorithm is a set of heuristics and calculations. Different data mining algorithms perform different tasks. For example, some algorithms are used to classify the data into different groups. Other algorithms are used to visualize the data into a clear picture. This will be discussed below.

A particular enumeration of patterns or models over the data will emerge by applying the algorithms to the dataset. Here, it is important to understand the difference between a model and a pattern. A model is "a global representation of a structure that summarizes the systematic component underlying the data or that describes how the data may have arisen."[144] In contrast, "a pattern is a local structure, perhaps relating to just a handful of variables and a few cases."[145]

The search for patterns forms the foundation of many data mining techniques, which will be discussed at length below. Examples of patterns include: associations (e.g., women who drink more than two glasses of wine a day also suffer from higher rates of breast cancer), sequences (e.g., if pregnant then get a new hair style), classifications (e.g., terrorists buy one-way

144. Penelope Markellou et al., *Knowledge Mining: A Quantitative Synthesis of Research Results and Findings in Knowledge Mining Studies*, *in* 185 Fuzziness and Soft Computing 1, 2 (Spiros Sirmakessis ed., Springer 2005).

145. *Id.*

plane tickets), and forecasting (e.g., a customer is likely to purchase this book based on the past behaviors of other customers). A group of rules derived from patterns in a dataset can be used to create a model. In other words, patterns can be collected together and defined as a mining model.[146]

The models and patterns that emerge after applying the data-mining algorithm can be thought of as automatically-generated hypotheses: they represent theories about the data set, the validity of which must be established through testing, a painstaking process which involves fine tuning the models to make sure they can be trusted. After model or pattern is validated, it can be applied to a set of new data. However, this is not technically "data mining." Instead, this is decision-making and is generally considered part of post processing.[147]

## VII. The Types of Data Mined

### A. Traditional

#### 1. Structured Data

Structured data "refers to the type of data that reside in fixed dimensions/fields," (a dimension or a field represents a specific characteristic of an item such as a social-security number, the name of a patient or the address of an individual).[148] A basic example of structured data includes a relational database.[149] A relational database is a collection of related tables where data are stored in tables made up of rows and columns, which are related to one another by way of common fields.[150] It is a two-dimensional structure and can be contrasted with a multi-dimensional database where data takes on a cube-like structure.

A data warehouse is another typical example of structured data. As explained above, a data warehouse is a repository of data, which is set up by an organization to support strategic decision-making.[151] Typically, this data is modeled as being multidimensional, which offers good support for query and analysis, especially OLAP. The data warehouse coherently stores the historic

146. *See generally* Han & Kamber, *supra* note 6, at 724.

147. Taipale, *supra* note 86, at 28.

148. Bin Zhou, *Keyword Search on Large-Scale Structured, Semi-Structured, and Unstructured Data*, *in* Handbook of Data Intensive Computing 733, 735 (Borko Furht & Armando Escalante eds., Springer 2011).

149. *Id.*

150. Linfeng Wang & Kay Chen Tran, *Database Management*, *in* Modern Industrial Automation Software Design 59, 61 (2006).

151. Mehmed Kantardzic, *Data-Mining Concepts*, *in* Data Mining: Concepts, Models, Methods, and Algorithms 1, 14 (2d ed. 2011).

data of an organization.[152] It is not usually updated and is used to respond to queries from decision makers.[153] Typically, the warehouses are huge and can store billions of records.[154]

A final example of structured data is a transactional database. Unlike the data held in a data warehouse, the data held in a transactional database changes all of the time. It is designed to support the day-to-day transactions that keep an organization functioning.[155]

## B. Untraditional

### 1. Unstructured

Unlike structured data, unstructured data do not reside in any fixed dimensions/fields.[156] It is "wild data" and it is fueling the big-data surge.[157] It does not fit neatly into rows and columns within a database or a spreadsheet because it does not have a specific structure. It can be textual (e.g., generated in a Word document) or non-textual (e.g., generated in a video by a surveillance camera in a department store).[158] This form of data generally requires extensive processing to extract and structure the information contained in it.[159]

152. *Id.* ("The function of the data warehouse is to store the historical data of an organization in an integrated manner that reflects the various facets of the organization and business.").

153. *Id.* ("The data in a warehouse are never updated but used only to respond to queries from end users who are generally decision makers.").

154. *Id.* ("Typically, data warehouses are huge, storing billions of records. In many instances, an organization may have several local or departmental data warehouses often called data marts. A data mart is a data warehouse that has been designed to meet the needs of a specific group of users. It may be large or small, depending on the subject area.").

155. Krzysztof J. Cios et al., *Data*, *in* DATA MINING: A KNOWLEDGE DISCOVERY APPROACH 27, 35 (2007) ("Transactional databases are stored as flat files and consist of records that represent transactions. A transaction includes a unique identifier and a set of items that make up the transaction. One example of a transaction is a record of a purchase in a store, which consists of a list of purchased items, the purchase identifier, and some other information about the sale. The additional information is usually stored in a separate file, and may include customer name, cashier name, date, store branch, etc.").

156. Zhou, *supra* note 148, at 736.

157. Steve Lohr, Op-Ed., *The Age of Big Data*, N.Y. TIMES, Feb. 12, 2012, SR1, at SR 2, *available at* http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

158. Zhou, *supra* note 148, at 736.

159. Kantardzic, *supra* note 151, at 12; *see also* Tengjiao Wang, *Preface to the 2nd International Workshop on Unstructured Data Management (USDM 2011)*, *in* 6612 WEB TECHNOLOGIES AND APPLICATIONS: LECTURE NOTES IN COMPUTER

### 2. Semi-Structured

Semi-structured data is a nebulous term. One commentator describes it as the "duck-billed platypus of the data kingdom,"[160] and another as "just a euphemism" for unstructured data.[161] The term seeks to describe data that includes a mix of structured data, such as metadata ("data about data"), and unstructured data. For example, e-mail, mp3, and video-data can be classified as semi-structured data, even though they contain mainly unstructured data. They meet this classification because they are always attached to some more structured data such as: Author, Subject or Title, Summary, etc.[162]

---

SCIENCE 398, 399 (Xiaoyong Du et al. eds., 2011) ("The management of unstructured data has been recognized as one of the most attracting problems in the information technology industry. With the consistent increase of computing and storage capacities (due to hardware progress) and the emergence of many data-centric applications (e.g. web applications), a huge volume of unstructured data has been generated. Over 80% of world data today is unstructured with self-contained content items. Since most techniques and researches that have proved so successful performing on structured data don't work well when it comes to unstructured data, how to effectively handle and utilize unstructured data becomes a critical issue to these data-centric applications.").

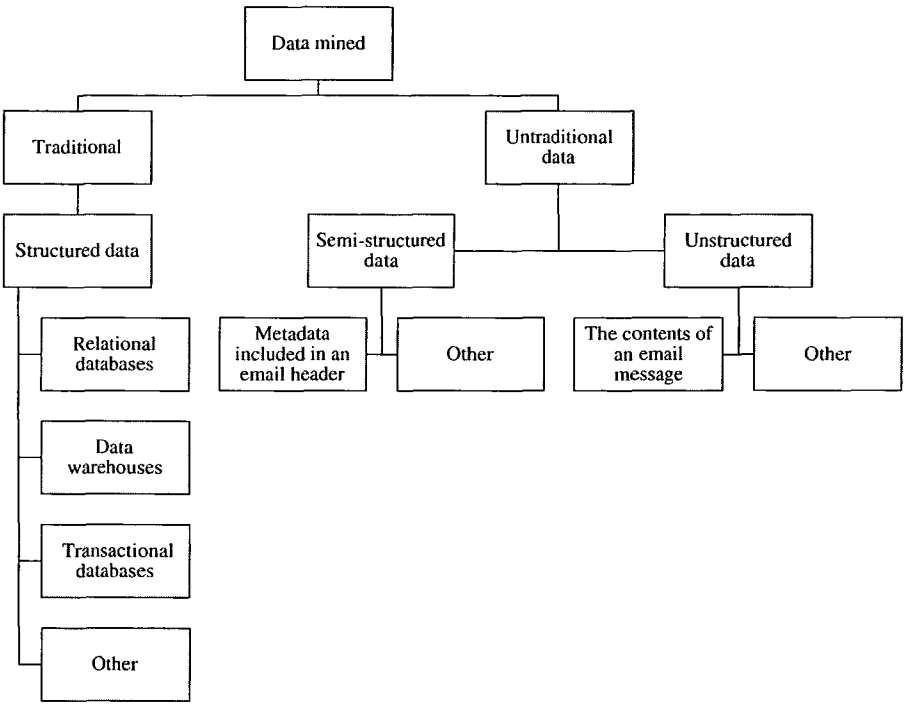
160. Phil Storey, *The Difference Between Structured Data, Semi-Structured Data and Unstructured Data*, DATAWATCH BLOG (Jan. 27, 2012), http://blog.datawatch.com/the-difference-between-structured-data-semi-structured-data-and-unstructured-data/#more-1158.

161. Roberto V. Zicari, *On Analyzing Unstructured Data—Interview with Michael Brands*, ODBMS INDUSTRY WATCH (July 11, 2012), http://www.odbms.org/blog/2012/07/on-analyzing-unstructured-data-interview-with-michael-brands.

162. *Id.*; *see also* Zhou, *supra* note 148, at 736 ("This type of data is a form of structured data which do not conform with a specific data schema or a data model. However, semi-structured data use attributes or tags to separate semantic elements and enforce attribute hierarchies. XML data is one example of the semi-structured data. In recent years, semi-structured data are increasingly popular due to the advent of the Internet. In many situations, full-text documents and relational databases are not able to represent the rich sets of the information."); *see further*, MINELLI ET AL., *What Is Big Data and Why Is It Important?*, *in* BIG DATA, BIG ANALYTICS: EMERGING BUSINESS INTELLIGENCE AND ANALYTIC TRENDS FOR TODAY'S BUSINESSES 578 (2013) ("The term semi-structured data is used to describe structured data that doesn't fit into a formal structure of data models. However, semi-structured data does contain tags that separate semantic elements, which includes the capability to enforce hierarchies within the data.").

**FIGURE 2: DATA MINED**

This figure represents an illustration of the types of data that can be mined.




## VIII. HUMAN ACTORS IN DATA MINING

Data mining involves very complex interdependencies between humans and technology that have hitherto not existed. Data mining is not a simple, stand-alone, "magic-box" technology; rather, it is a complicated socio-technical system.[163] The advanced algorithms applied in data mining are connected to a world of interlinked databases that partly rely on humans to build, train, and apply them.

More specifically, humans play an important role with respect to both the operation and development of a data mining system.[164] From an operational perspective, there are a number of human roles involved in the process of data mining, some of which may be assumed by the same individual (depending on the scale of the project).[165] For example, a project leader may have "the overall responsibility for planning, coordinating, executing, and

163. HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY AND THE INTEGRITY OF SOCIAL LIFE 4–5 (2010).

164. NAT'L RESEARCH COUNCIL, *supra* note 31, at 21–23.

165. Jackson, *supra* note 27, at 271.

deploying the data mining project."[166] Additionally, the data-mining client is the organizational "domain expert that requests the project and utilizes the results, but generally does not possess the technical skills needed to participate in the execution of the more technical phases of the data mining project such as data preparation and modeling."[167] There is also, of course, the data mining analyst whose role is to translate the organizational objectives "into technical requirements to be used in the subsequent development of the data mining model(s)."[168] Another important player is the data mining engineer who not only develops, interprets, and evaluates the data mining model(s) in light of the organizational objectives, but also consults with the data mining client and analyst to assist in achieving positive data mining results.[169] Finally, the IT analyst "provides access to the hardware, software, and data needed to complete the data mining project successfully."[170]

With respect to the role of human interaction with a data mining system, Maulik et al. explains that "[u]ser interaction helps the mining process to focus the search patterns, appropriately sampling and refining the data."[171] He emphasizes the importance of a prior domain-specific knowledge in all phases of a discovery process.[172] As a result of human contribution, the data-mining algorithm performs more effectively.[173]

Data mining is about finding useful information to apply in organizational decision-making. Therefore, it logically follows that the more a data mining analyst understands about the dataset at hand, the problem he or she is investigating, and the sort of pattern he or she is looking for, the more likely he or she will find something useful in the dataset.[174] Legendary data miner Hand explains: "The bottom line is that computing power does not replace brain power. They work hand in hand. The data miner who uses both will be the one who finds the interesting and valuable structures in the data."[175]

---

166. *Id.*

167. *Id.*

168. *Id.*

169. *Id.*

170. *Id.*

171. UJJWAL MAULIK, SANGHAMITRA BANDYOPADHYAY & ANIRBAN MUKHOPADHYAY, MULTIOBJECTIVE GENETIC ALGORITHMS FOR CLUSTERING 68 (2011).

172. *Id*

173. *Id.*

174. David J. Hand, *Protection or Privacy? Data Mining and Personal Data,* 3918 ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING LECTURE NOTES IN COMPUTER SCIENCE 3 (2006).

175. *Id.*

This idea can be described succinctly in the aphorism: "Chance favors the prepared mind."[176] From a developmental perspective, human judgment and expertise play critical roles in shaping how a given system is designed at the outset.[177] For example, humans design the algorithms applied in data mining. Even if some of these algorithms are non-interpretable (i.e. cannot be explained to humans), a human still decides upon the initial design and applies the algorithm in the first place.[178]

## IX. Basic Taxonomy of Data Mining

### A. Verification-Driven Tasks (Deductive, Hypothesis-Driven, Top-Down, Analytic)

Verification-driven data mining (a.k.a. deductive, hypothesis-driven, top-down, analytic), seeks to extract information in the process of validating a hypothesis postulated by a user.[179] This approach begins with a particular hypothesis to be validated by a set of data. Taipale explains that the hypothesis can be developed from using a bottom-up approach to initially mine the data or be developed from real-world knowledge.[180]

Notwithstanding the application of this approach to large data sets, verification-driven data mining is similar to conventional data analysis methods.[181] Specifically, verification-driven data mining is closely associated with traditional databases that rely on query and reporting, multidimensional analysis, and statistical analysis.[182] Anane explains that the purpose of verification-driven data mining "is to validate a hypothesis expressed in terms of the entities and relations that exist in the database. This mode of enquiry is marked by the potential spawning of further queries in response to new insights."[183]

Though verification-driven data mining can be utilized to discover hidden information many commentators do not consider it "true" data mining. This is because with "true" data mining, a computer generates the hypothesis, not the user: the interrogation of the data is completed by the data-mining algorithm rather than by the analyst. Verification-driven data mining has been

176. *Id.*

177. Nat'l Research Council, *supra* note 31, at 21–23.

178. Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 Penn St. L. Rev. 285, 293 (*2011*).

179. Sumathi & Sivanandam, *supra* note 108, at 269, 505.

180. Taipale, *supra* note 86, at 30.

181. Rachid Anane, *Data Mining and Serial Documents*, 35:3 Computers & Human. 300 (2001).

182. Sumathi & Sivanandam, *supra* note 108.

183. Anane, *supra* note 181.

included in this taxonomy because it is part of the iterative, dynamic data mining process (see below for further explanation).

## B. Examples of Verification Driven Tasks

### 1. Query and Reporting

In query and reporting, the goal is to validate a hypothesis expressed by the analyst. For example, "sales of four-wheel drive vehicles increase during the winter season."[184] Query and reporting first requires creation of a query that best expresses a hypothesis.[185] Next, it requires positing the query to the data repository and analyzing returned data to establish whether it supports or refutes the hypothesis.[186] Query and reporting is an iterative process where the initial query is often refined and additional queries are frequently posed.[187]

### 2. Multidimensional Analysis Such as Online Analytical Processing (OLAP)

Multidimensional analysis is used to selectively extract and view data from a myriad of perspectives.[188] For example, an analyst employed by a retail company might be interested in understanding overall sales for the company, as well as sales at a specific time, branch, and location.[189] Multidimensional analysis tools allow the analyst to partition the data for view from a number of perspectives.[190] It also provides the analyst with excellent graphical tools to utilize for further development of his/her understanding of the data.[191]

OLAP, as mentioned earlier in this paper, is a form of multidimensional analysis wherein an analyst is given a number of special tools to manipulate data, allowing him or her to extract, report, or visualize a dataset based on

184. SUMATHI & SIVANANDAM, *supra* note 108, at 199.

185. *Id.*

186. *Id.*

187. *Id.*

188. Anoop Singhal, *An Overview of Data Warehouse, OLAP and Data Mining Technology, in* DATA WAREHOUSING AND DATA MINING TECHNIQUES FOR CYBER SECURITY 1, 7 (2007).

189. *Id.*

190. Robert M. Coleman, Matthew D. Ralston, Alexander Szafran & David M. Beaulieu, *Multidimensional Analysis: A Management Tool for Monitoring HIPAA Compliance and Departmental Performance*, 17 J. DIGITAL IMAGING 3, 196–204 (Sept. 2004).

191. *Id.*

intuition.[192] For example, an analyst can "pivot" or "rotate" a dataset, meaning he or she may change the dimensional orientation of displayed data. For example, rows and columns may be swapped.[193] An analyst can also "drill up" or "drill down." To "drill up" means to summarize data by climbing up the hierarchy, such as moving from the view of a weekly report to a view of a monthly report. Alternately, the analyst may "drill down." Drilling down allows the analyst summarize the data by climbing down the hierarchy, such as going from a monthly report to a weekly report.[194] Furthermore, the analyst may "slice-and-dice" the data meaning he/she can view it from many perspectives, showing some data fields and hiding others.[195]

### 3. Statistical Analysis

Statistical analysis has long been used to make sense out of data: King David numbered his people, and the Egyptians measured their fields.[196] Essentially, statistical analysis involves an analyst formulating a hypothesis and then testing its validity by running mathematical tests on the collected data.[197] Smith provides an example:

> . . . if an analyst was studying the relationship between income level and the ability to get a loan, the analyst may hypothesi[ze] that there will be a correlation between income level and the amount of credit someone may qualify for. The analyst could then test this hypothesis with the use of a data set that contains a number of people along with their income levels and the credit available to them. A test could be run that indicates[,] for example[,] that there may be a high degree of confidence that there is indeed a correlation between income and available credit. The main point here is that the analyst has formulated a hypothesis and then used a statistical test along with a data set to provide evidence in support or against that hypothesis.[198]

192. CARLO VERCELLIS, BUSINESS INTELLIGENCE: DATA MINING AND OPTIMIZATION FOR DECISION MAKING 80–81 (2009).

193. Sam Y. Sung, Yao Liu, Hui Xiong & Peter A. Ng, *Privacy Preservation for Data Cubes*, 9 KNOWLEDGE & INFO SYS 1, 38–61, 44 (2006).

194. *Id.*

195. *Id.*

196. Arnold Goodman, Chandrika Kamath & Vipin Kumar, *Data Analysis in the 21st Century*, WILEY INTERSCIENCE, Oct. 30, 2007, at 1.

197. J. L. Smith, *One of the Main Differences Between Statistical Analysis and Data Mining*, EZINE ARTICLE, http://ezinearticles.com/?One-of-the-Main-Differences-Between-Statistical-Analysis-and-Data-Mining&id=4578250.

198. *Id.*

Similar to search and query and multidimensional analysis, statistical analysis is used to confirm a hypothesis theorized by an analyst through the use of a "top-down" approach.[199]

### C. Discovery-Driven Tasks (Inductive, Data-Driven, Bottom Up, Synthetic)

Discovery-driven data mining, which is also called inductive, data-driven, bottom-up or synthetic data mining, seeks to analyze the data and extract patterns on which a hypothesis or model can be based.[200] A system based on the discovery-driven approach can generate new concepts from existing information in the database.[201] In other words, an analyst does not need to start with a hypothesis, but rather ask the system to create one.[202] The correlations and patterns discovered in the data are not predictable in advance; the analyst does not know and cannot anticipate with any certainty what correlations and patterns will emerge prior to the application of the data-mining algorithm to the data set.[203]

An advantage of discovery-driven data mining is that an analyst does not need to have any preconceived assumptions about the dataset before she starts to mine the dataset because the analyst is searching for any unanticipated pattern that may exist that she might discover useful.[204] The concern is that scientific proposals that are derived without a preconceived hypothesis are not valuable, reliable or significant because correlations that appear in the data could be totally random.[205] For example, a report on data mining from the U.S. Department of Homeland Security explained that the mere existence of patterns in the data "cannot reveal whether any discovered pattern is meaningful or significant."[206] Similarly, McCarty explains:

---

199. VERCELLIS, *supra* note 192, at 81; *see also* S. SUMATHI & S.N. SIVANANDAM, INTRODUCTION TO DATA MINING AND ITS APPLICATIONS 200 (2006) ("Simple statistical analysis operations usually execute during both query and reporting, as well as during multidimensional analysis. Verifying more complex hypothesis, however, requires statistical operations coupled with data visualization tools.").

200. Taipale, *supra* note 86, at 28.

201. Anane, *supra* note 181, at 300.

202. Gerald Benoît, *Data Mining*, 36 ANN. REV. INFO. SCI. & TECH 265, 271 (2002).

203. Mark MacCarthy, *New Directions in Privacy: Disclosure, Unfairness and Externalities*, 6 J.L. & POL'Y FOR INFO. SOC'Y 425 (2011).

204. J. A. McCarty, *Database Marketing*, WILEY INT'L ENCYCLOPEDIA OF MARKETING (2010).

205. *See* Douglas B. Kell & Stephen G. Oliver, *Here is the Evidence, Now What is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era*, 26 BIOESSAYS 99, 99 (2004).

206. U.S. DEP'T OF HOMELAND SEC., DATA MINING REPORT: DHS PRIVACY OFFICE RESPONSE TO HOUSE REPORT 7 (2006).

> The very real danger with such data-driven approaches is that one may be capitalizing on chance relationships in the data. Without a hypothesis based on sound theory, it is far more likely that one will find relationships because of the search for any relationship; some of the found relationships may exist because of chance occurrences rather than meaningful and robust relationships that will maintain across time.[207]

So, is discovery-driven data mining voodoo science? In a way it is. While discovery-driven data mining can, in some way, be understood as "voodoo science" because the algorithms generate hypotheses automatically, it is also true that these hypotheses must be subsequently and independently evaluated and validated to test their accuracy.[208] In other words, data miners understand the risk in the approach and take steps to evaluate the reliability of their findings.[209]

Discovery-driven data takes two major forms: description and prediction. Insight is the goal of descriptive data mining. For example, this insight could come in the form of a simplification and summarization of the dataset: a marketing manager might want to understand what big spenders look like, or a network administrator might want to see what unusual behavior looks like on the network being monitored. The goal of predictive data mining is to create inferences about future cases based on the patterns revealed in the data set: an insurance underwriter might want to predict the likelihood that a customer will let her policy lapse or a marketing executive might want to predict whether a particular customer would switch brands for a specific product.[210] In short, a descriptive model creates a way to explore the properties of the data examined and a predictive model creates a way to predict new properties.

---

207. McCarty, *supra* note 204.

208. Bart W. Schermer, *The Limits of Privacy in Automated Profiling and Data Mining*, 27 COMPUTERS LAW & SEC. REV. 45, 48 (2011).

209. McCarty, *supra* note 204.

210. Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 PENN ST. L. REV. 285, 292 (2011) ("In a predictive process, the analysts use data mining applications to generate rules based on preexisting data. Thereafter, these rules are applied to newer (while partial) data, which is constantly gathered and examined as the software constantly searches for previously encountered patterns and rules. Based on new information and previously established patterns, the analysts strive to predict outcomes prior to their occurrence (while assuming that the patterns revealed in the past pertain to the current data as well."); *see also* Schermer, *supra* note 208 ("As the name implies, the goal of predictive data mining is to make a prediction about events based on patterns that were determines using known information."); Mark Whitehorn, *The Parable of the Beer and Diapers: Never Let the Facts Get in the Way of a Good Story*, REGISTER, Aug. 15, 2006, http://www.theregister.co.uk/2006/08/15/beer_diapers.

### 1. Supervised, Directed and Predictive Tasks

Predictive data mining involves supervised learning, which is also called directed learning.[211] Supervised learning is essentially a two-stage process.[212] First, an analyst must train the algorithm to recognize different classes of data by exposing it to a series of examples.[213] Second, the analyst must test how well the algorithm has learned from these examples by supplying it with a previously unseen set of data, usually referred to as a test set.[214] After the algorithm learns to effectively identify patterns in a set of training data then it can be used as a model to make predictions in a new set of data where no information about the dataset is known.[215] It is useful to think of predictive data mining as the equivalent to learning with a teacher because the predictive model is inducted from a training set where the answers are already provided.[216]

### 2. Examples of Predictive Tasks Classification

Classification is used to take data and place it into certain predefined groups.[217] In other words, classification is applied to place new data into an existing structure.[218] It is a multi-step process, which consists of creating a model through the use of a set of training data, testing the model through the use of a set of evaluation data, and then applying the model to a new set of data. For example, a financial analyst might seek to construct classification models to categorize bank loan and mortgage applications into risky or safe categories, or a medical analyst might seek to construct a classification model to help define medical diagnosis based on symptoms and health conditions.[219]

The first step, training, builds the actual classification model by analyzing a set of training data where the group assignments are known.[220] Here, the classification algorithm "learns" how to recognize predefined groups

211. ORACLE DATA MINING 26 (Concepts ed., 2008), http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm.

212. *See id.*

213. *Id.*

214. *Id.*

215. Anane, *supra* note 181, at 312.

216. Gary M. Weiss & Brian Davison, *Data Mining*, *in* THE HANDBOOK OF TECHNOLOGY MANAGEMENT 15 (Hossein Bidgoli ed., 2010).

217. Anane, *supra* note 181, at 308.

218. Furnas, *supra* note 1.

219. Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang & Lei Hua, *Data Mining in Healthcare and Biomedicine: A Survey of the Literature*, 36:4 J. MED. SYS. 2431, 2433–34 (2012).

220. *Id.* at 2434.

through the set of examples.[221] If the quality of the training data is high then the quality of the model is also likely to be high.[222] Standard classifiers are trained using two examples (e.g., a blue class looks like this and a red class looks like this).[223] One-class classifiers, however, are trained using examples of only one class (e.g., a red class looks like this).[224] The goal with one-class classification is to find outliers (e.g., data that are not part of the "in" group).[225] The second step, testing, examines the classification model for accuracy.[226] The testing often takes place on a set of testing data that has been set aside from the training data.[227] If the accuracy is acceptable, then the model can be applied to predict a class of objects in a new, unlabeled set of data.[228]

*Examples of Classification Techniques*

The first example of a classification technique is a decision tree.[229] In a decision tree, there are diagrams where the root is a simple question or condition that has multiple answers.[230] Each answer leads to further questions or conditions. Hanmant explains:

> A [d]ecision [t]ree is predictive model that can be viewed as tree, each branch is a classification question and leaves of the tree are partitions of data set with their classification. It divides data on each branch point without losing any of the data. The number of churners and non[-]churners is conserved as we move up or down the tree.[231]

221. *Id.* at 2433.

222. *See* Adam Mazmanian, *What the NSA Can't Do with Your Data (Probably)*, FCW: THE BUS. OF FED. TECH., June 12, 2013, http://fcw.com/Articles/2013/06/12/NSA-risk-assessment.aspx?Page=1 ("Algorithms don't know about common sense" so "[i]f your data is bad, they'll infer the wrong thing.").

223. Richard G. Brereton, *One Class Classifiers*, 25 J. CHEMOMETRICS 225 (2009).

224. *Id.*

225. *Id.*

226. Yoo, *supra* note 219, at 2434.

227. *See generally id.*

228. *See generally id.* at 2433–34 (explaining how banks have used classification models).

229. *Learning Decision Trees*, ARTIFICIAL INTELLIGENCE, http://artint.info/html/ArtInt_177.html (last visited Sept. 2, 2013).

230. *See id.*

231. Hanmant N. Renushe, Prasanna R. Rasal & Abhijit S. Desai, *Data Mining Practices for Effective Investigation of Crime*, 3:3 INT. J. COMPUTER TECH. & APPLICATIONS 865, 868 (2012).

Another example of a classification technique is a neural network.[232] Hanmant explains: "Neural [n]etworks are biological systems that detect patterns, make predictions and learn. The artificial neural networks are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models for historical databases."[233] Neural networks attempt to mimic the human brain and to learn from mistakes just like humans.[234]

*Regression*

Like classification, regression involves taking information from a sample of data and using this information to formulate predictions about behaviors or events in the entire population from which the sample was taken.[235] Regression models are built using a set of training or historic data.[236] Then, the model is tested against a set of evaluation data, and finally, the model is applied to a new set of data to make a prediction.[237]

Regression differs from classification in that it is not a label, but a numeric value, which is attached to the observations and will be predicted to unlabeled, new data.[238] In other words, if classification asks "what class?" then regressions asks "how much?" A simple example of how regression can be used is "to predict the value of a house based on location, number of rooms, lot size, and other factors."[239] Facebook can also use regression to predict the future engagement of a user based on past behavior using factors like "the amount of personal information shared, number of photos tagged, friend requests initiated or accepted, comments, likes etc."[240]

*Example of a Regression Algorithm/Technique*

Linear regression is a classic example of a regression technique. Linear regression models the relationship between variables that are dependent upon

232. M. Seetha, et. al., *Artificial Neutral Networks and Other Methods of Image Classification*, J. THEORETICAL & APPLIED INFO. TECH. 1039.

233. Renushe, *supra* note 231, at 868.

234. *See* Seetha, *supra* note 232, at 1040.

235. S. Fournier, *An Informal Set of Statistics Basics*, HS404B: STATISTICS/REGRESSION, http://people.brandeis.edu/~fournier/HANDOUT.pdf.

236. ORACLE DATA MINING, *supra* note 211, at 53.

237. *Id.*

238. Katharina Morik, *Applications of Knowledge Discovery*, *in* INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 2 (2005).

239. ORACLE DATA MINING, *supra* note 211, at 49.

240. Furnas, *supra* note 1.

one another. It looks for "the 'best' line to fit."[241] For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.[242] Multiple linear regression can be defined as "an extension of linear regressions, where more than two variables are involved and the data are fit to a multidimensional surface."[243]

### 3. Unsupervised, Undirected and Descriptive/Exploratory Tasks

Descriptive data mining is an undirected/unsupervised meaning that the algorithm does not receive any guidance while learning.[244] Descriptive data mining is exploratory in nature.[245] Zarsky explains that "[d]escriptive data mining provides analysts with a better understanding of the information at their disposal, while uncovering hidden traits and trends within the dataset."[246] For example, descriptive data mining can be used to discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in a dataset.

Similarly, Maimon and Rokach state that "[d]escriptive data mining is oriented to data interpretation, which focuses on understanding[,] though visualization[,] for example[,] the way the underlying data relates to its parts."[247] It focuses on, among other things, the intrinsic structure, relations, and interconnectedness of the data and aspires to improve the overall comprehension of the dataset.[248] It is contended that descriptive data mining requires greater user involvement for the interpretation of patterns revealed after the mining is completed.[249]

---

241. S. Sumathi, *Data Mining and Data Warehousing*, STUDIES IN COMPUTATIONAL INTELLIGENCE 421 (2007).

242. *Linear Regression*, YALE UNIV. (Sept. 2, 2013), http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

243. Sumathi, *supra* note 241, at 421.

244. ORACLE DATA MINING, *supra* note 211, at 26.

245. Yoo et. al., *supra* note 219, at 2433.

246. Zarsky, *supra* note 210, at 292; *see* Schermer, *supra* note 208, at 46 ("The goal of descriptive data mining is to discover unknown relations between different data objects in a database. Descriptive data mining algorithms try to discover knowledge about a certain domain by determining commonalties between different objects and attributes. By discovering correlations between data objects in a dataset that is representative of a certain domain, we can gain insight to it.").

247. ODED MAIMON & LIOR ROKACH, DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK 6 (2010).

248. ANDREAS L. SYMEONIDIS & PERICLES A. MITKAS, AGENT INTELLIGENCE THROUGH DATA MINING 15 (2005).

249. Anane, *supra* note 181, at 300.

### 4. Examples of Descriptive Tasks

*Clustering*

Clustering is about clumping together similar things, events or people in order to create meaningful subgroups using automated methods.[250] More specifically, "a cluster is a collection of data objects that are similar to each other within the same cluster, while being dissimilar to the objects in any other cluster.[251] Ponniah provides a very simple example of real-world clustering:

> Take the very ordinary example of how you do your laundry. You group the clothes into whites, dark-colored clothes, light-colored clothes, permanent press, and the ones to be dry-cleaned. You have five distinct clusters. Each cluster has a meaning and you can use the meaning to get that cluster cleaned properly. The clustering helps you take specific and proper action for the individual pieces that make up the cluster.[252]

A classic example of how clustering is applied within the context of data mining is to discover a set of customers that have similar buying behaviors.[253]

Clustering is very different from classification where objects are assigned into predefined classes.[254] In clustering, unlike in classification, the algorithm both defines the classes itself and puts the object into each class

250. Amit Kumar Patnaik, Biswaranjan Nayak & Srinivas Prasad, *Data Mining and Its Current Research Directions*, Sept. 2, 2013, http://ficta.in/attachments/article/55/07%20Data%20Mining%20and%20Its%20Current%20Research%20Directions.pdf.

251. Mark Last, *Data Mining, in* CYBER WARFARE AND CYBER TERRORISM 358 (Lech Janczeski ed., 2008); *see* Weiss & Davison, *supra* note 216, at 11("There are many reasons to cluster data. The main reason is that it allows us to build simpler, more understandable models of the world, which can be acted upon more easily. People naturally cluster objects for this reason all the time.").

252. PONNIAH, *supra* note 125, at 409-410.

253. Weiss & Davison, *supra* note 216; *see* PONNIAH, *supra* note 125, at 410 ("[T]hink of a specialty store owner in a resort community who wants to cater to the neighborhood by stocking the right type of products. If he has data about the age group and income level of each of the people who frequent the store, using these two variables, the store owner (sic) can probably put the customers into four clusters. These clusters may be formed as follows: wealthy retirees staying in resorts, middle-aged weekend golfers, wealthy young people with club memberships, and low-income clients who happen to stay in the community. The information about the clusters helps the store owner in his marketing.").

254. *See* HAN & KAMBER, *supra* note 6, at 398.

itself.[255] In other words, clustering differs from classification with respect to the fact that it does not assume any prior knowledge: instead of searching for pre-defined classes, the goal is to explore the data set and see what classes emerge.[256] The goal of clustering is to arrange data into previously unknown groups.[257]

It is worth mentioning that even though data mining allows for the discovery of clusters that capture the relevant distinctions between apparent groups in the data set that might not be obvious to the naked eye; it is also true that a human analyst must be able to recognize the value of the discovered clusters in order to do something useful with it. In other words, a cluster might appear in the dataset, but a human has to recognize that this cluster, for example, represents wealthy customers.[258] This is a challenge since it not always easy to identify the meaning of a cluster.[259]

*Example of a Clustering Algorithm/Technique*

K-means is a typical example of a clustering technique.[260] K-means clusters observations into groups without any prior knowledge of those relationships.[261] It is a descriptive technique ". . . used to identify distinct groupings, with the goal of minimizing variability within the cluster and maximizing variability between clusters."[262]

*Association Rule Mining*

The aim of association rule mining is to describe a dataset by finding interesting patterns in the dataset that were previously unknown.[263] Hammant explains: "Association identifies affinities/associations among the collection

255. *See id.*

256. *See id.*

257. *See id.*

258. PONNIAH, *supra* note 125, at 410.

259. *Id.* ("The store owner has to understand that one of the clusters represents wealthy retirees residing in resorts. Only then can the store owner [sic] do something useful with that cluster. It is not always easy to discern the meaning of every cluster the data mining algorithm forms. A bank may get as many as twenty clusters but be able to interpret the meanings of only two. But the return for the bank from the use of just these two clusters may be enormous enough so that they may simply ignore the other eighteen clusters.").

260. Barbara E. McDermott, John F. Edens, Cameron D. Quanbeck, David Busse & Charles L. Scott, *Examining the Role of Static and Dynamic Risk Factors in the Prediction of Inpatient Violence: Variable and Person Focused Analyses*, 32 LAW & HUMAN BEHAV. 325, 330 (2008).

261. *See generally id.*

262. McDermott, *supra* note 260, at 330.

263. HAN & KAMBER, *supra* note 6.

of data as reflected in the examined records. A result is patterns describing rules of association in data."[264]

Association models capture the co-occurrence of items or events in large volumes of data: they establish a correlation between elements in the dataset that were previously unknown to have any connection.[265] For example, an association rule could take the form, "if event X occurs, then event Y is likely to occur 30% of the time."[266] Events X and Y could represent items bought in a purchase transaction or medical symptoms of a given patient, among many other phenomena recorded in a database over time.[267]

Association models are built on a dataset of interest to obtain a better understanding about that dataset. That is, unlike with predictive models, the goal is not to apply the models to a separate dataset, but rather to return a set of rules, which explain how items or events are associated with each other.[268] It is important to note, however, that although association falls under the rubric of descriptive data mining, highly accurate association rules can be used to make predictions.[269] In other words, descriptive data mining can serve as a basis for predictive data mining.[270]

*Examples of an Association Algorithm/Technique*

Bayesian networks are essentially flowcharts that provide the probability that a specific path in the flowchart can be taken.[271] They can be

264. Renushe, *supra* note 231, at 865.

265. ORACLE DATA MINING, *supra* note 211, at 6.

266. *See id.*

267. *See id.*

268. *See id.*

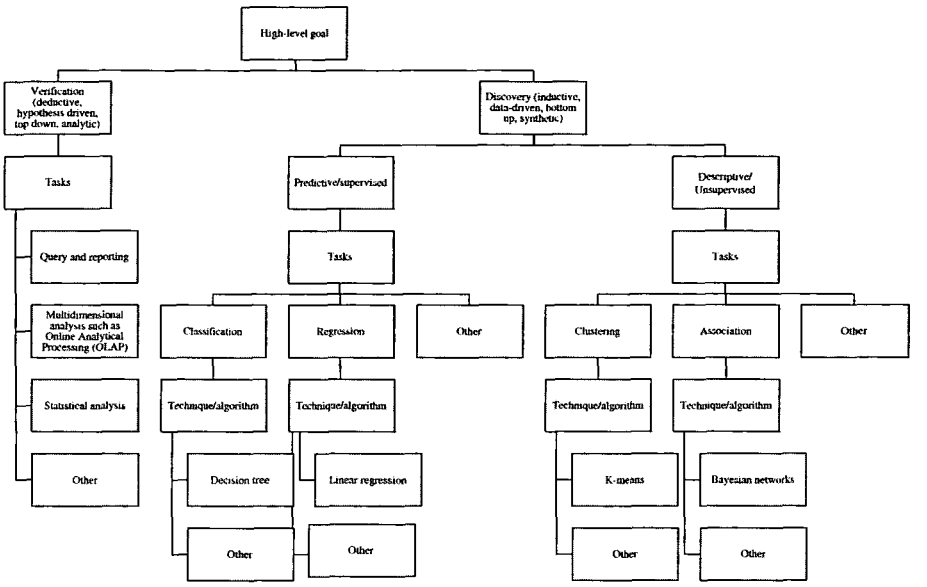
269. *See id.*

270. Furnas, *supra* note 1 ("The patterns detected and structures revealed by the descriptive data mining are then often applied to predict other aspects of the data.") (citing Amazon as a useful example of how descriptive findings are used for prediction).

271. *See* Eugene Charniak, *Bayesian Networks Without Tears*, AI MAGAZINE, Nov. 4, 1991, at 50 ( "The best way to understand Bayesian networks is to imagine trying to model a situation in which causality plays a role but where our understanding of what is actually going on is incomplete, so we need to describe things probabilistically. Suppose when I go home at night, I want to know if my family is home before I try the doors. (Perhaps the most convenient door to enter is double locked when nobody is home.) Now, often when my wife leaves the house, she turns on an outdoor light. However, she sometimes turns on this light if she is expecting a guest. Also, we have a dog. When nobody is home, the dog is put in the back yard. The same is true if the dog has bowel troubles. Finally, if the dog is in the backyard, I will probably hear her barking (or what I think is her barking), but sometimes I can be confused by other dogs barking").

used to draw strong inferences from a dataset.[272]

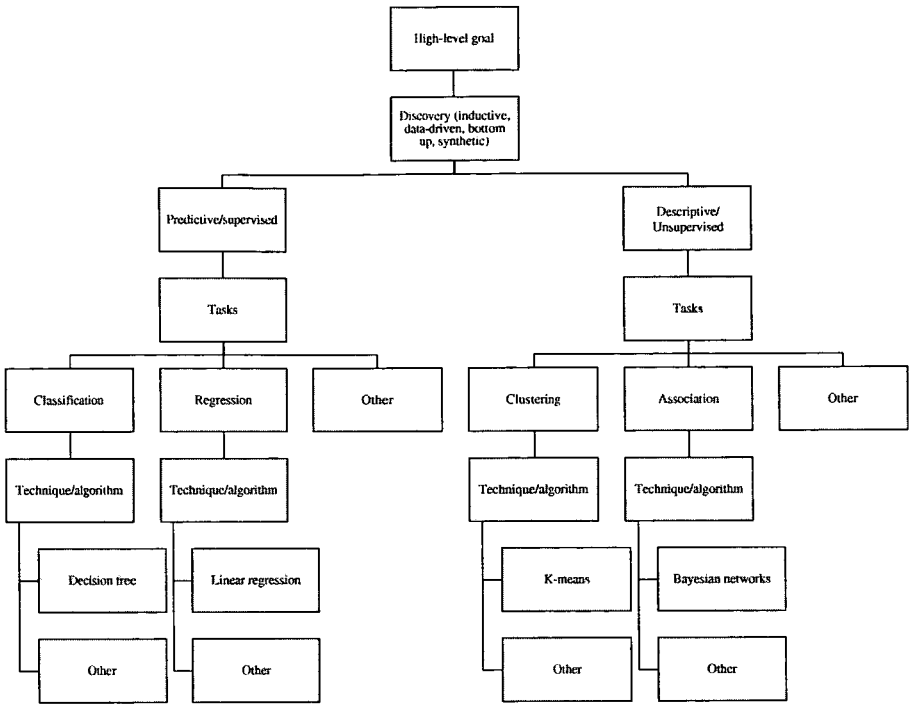
**FIGURE 3: A BASIC TAXONOMY OF DATA MINING**




This diagram is offered as a basic taxonomy of data mining. It is not meant to be a complete taxonomy, but rather it is just a starting point for understanding the concepts that are applied in the context of data mining. There are many more data mining tasks and techniques that exist, which are not included here. It is also important to mention that verification driven data mining is included, which might be controversial because many experts do not consider this to be true data mining.

272. Derya Ersel & Süleyman Günay, *Bayesian Networks and Association Analysis in Knowledge Discovery Process*, J. STAT. & ACTUARIAL SCI. 51 (2012), *available at* http://www.istatistikciler.org/dergi/IstDer120203.pdf.

FIGURE 4: A TAXONOMY OF "TRUE" DATA MINING?




In this diagram, verification-driven data mining was removed from the taxonomy, which may more accurately reflect what many experts consider "true" data mining. It should be noted that even still some experts seem to imply that "true" data mining has a predictive nature and therefore, might exclude descriptive data mining from the picture, too. It is not easy, however, to separate descriptive and predictive data mining tasks because, as mentioned above, descriptive data mining can serve as a basis for predictive data mining.

## X. DATA MINING IS DYNAMIC, ITERATIVE AND SYNERGISTIC

Data mining is a dynamic and iterative process and therefore, it is difficult to classify data mining based on the goals and tasks performed. Verification-driven and discovery-driven data mining work in synergy to produce knowledge and should be viewed as a continuum.[273] Kell and Oliver explain:

> . . .data- and technology-driven programmes [sic] are not alternatives to hypothesis-led studies in scientific knowledge discovery

273. Dmitry Brusilovsky & Eugene Brusilovskiy, *Data Mining: The Means to a Competitive Advantage*, BUSINESS INTELLIGENCE SOLUTIONS, Apr. 2008, http://www.bisolutions.us/The-Means-to-a-Competitive-Advantage.php.

> but are complementary and iterative partners with them. Many fields are data-rich but hypothesis-poor. Here, computational methods of data analysis, which may be automated, provide the means of generating novel hypotheses, especially in the postgenomic era.[274]

Even with respect to the discovery-driven data mining, creating a dichotomy is not easy. Fayyad, who referred to as a "data mining pioneer," explains the synergy that exists between descriptive and predictive data mining:

> There are two sides to data mining, descriptive and predictive . . . Descriptive data mining reorganizes the data, digging deeper into it and pulling out patterns, such as customer similarity, which allows you to create a short description about that group of customers. Predictive data mining looks for the best prediction, such as the best product to pitch to a customer. You won't get much insight, but it increases the performance . . . Using both techniques will give you the best results.[275]

Accordingly, it is important to understand that data mining is a process of retrieving, excluding, comparing reorganizing, digging and pulling, etc. It should be viewed as a whole of its parts, not just its separate parts.

## XI. Types of Data Mining Applications

After reviewing the basic fundamentals of data mining and developing a taxonomy for this advancing technology, it is now important to explore how data mining can be applied in different contexts. With a contextual understanding of data mining, it becomes easier to understand whether it raises legal concerns in one area of application that are not present in another area. In other words, an analysis of data mining applications could further the understanding of whether a one-size model of regulation of data mining is desirable or even possible.

The applications of data mining are widespread and can be broadly classified into five main groups: commercial applications, public sector applications, law enforcement applications, state security applications and personal-private applications. The purpose of this scheme is to find a classification that is related to the structure of current privacy and data protection laws in the United States and the European Union. In other words, this classification scheme is offered as a paradigm, albeit an imperfect one, to better understand how different types of data mining applications raise unique data protection and privacy concerns and, consequently, different balancing of interests.

274. Kell & Oliver, *supra* note 205, at 1.

275. Nathan Segal, *Drilling Down with a Data Mining Pioneer*, Datamation, Nov. 6, 2002, http://www.datamation.com/datbus/article.php/1495951/Drilling-Down-With-A-Data-Mining-Pioneer.htm.

Importantly, the examples provided under the main heading of each application may additionally fall under the rubric of another heading. For example, data mining that takes place in the contexts of "health care" and "higher-education" might be, on the one hand, considered "commercial private sector" activities in the United States and, on the other, "public sector" activities in the European Union. Furthermore, even though a data mining application may initially be classified as a "commercial private sector" application, it is very possible that it can later morph into a "law enforcement sector" application.

## A. Commercial/Private-Sector Applications

Commercial entities have played a major role in the development of data mining. These entities apply data mining in many different ways like the identification of likely consumers of their products, the tailoring of their marketing, and the discovery of future patterns of consumer behavior.[276] Many industries apply data mining including, but not limited to, the banking, insurance, pharmaceutical and retail sectors. These sectors seek to utilize data mining in order to increase profit margins and grow their businesses.

### 1. Contextual Examples

*Retail*

Through the use of scanners and cash registers, the retail industry has been able to capture huge amounts of point-of-sale data.[277] Since information is gathered each time a customer makes a purchase, point-of-sale data or "bar-code data" are often considered the lifeblood of the retail industry.[278] This information includes the transaction time, transaction date, transaction location, transaction data and the point-of-sale location.[279]

Data mining is used to analyze point-of-sale data to provide information on what product combinations were purchased together, when they were bought, and in what sequence.[280] This information is useful to retailers who seek to promote their most profitable products.[281] In addition, it encourages

276. Marie Bienkowski, Mingyu Feng & Barbara Means, ENHANCING TEACHING AND LEARNING THROUGH EDUCATIONAL DATA MINING AND LEARNING ANALYTICS: AN ISSUE BRIEF, U.S. DEPT. OF EDUC. 1 (2012), *available at* http://www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf.

277. PONNIAH, *supra* note 125.

278. Margaret Rouse, *Barcode Data*, WHATIS.COM, Mar. 22, 2011, http://whatis.techtarget.com/definition/barcode-data-point-of-sale-data-POS-data.

279. *Id.*

280. *See Data Mining: What is Data Mining?*, UCLA, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

281. *See id.*

customers to purchase related products that they may have overlooked.[282] Other areas of data mining application lie with understanding the best types of promotions, store plan and arrangement of promotional displays, direct marketing, inventory management, sales trends, seasonal and manpower planning based on busy times.[283]

*Biological*

Data mining is also used to make sense of the flood of biological and clinical data from genomic sequences, DNA microarrays, protein interactions, biomedical images, and disease pathways. For example, "gene sequences isolated from diseased and healthy tissues and then compared in order to identify key differences between the two classes of genes."[284] Data mining can also be used to predict which chemicals might change gene expression and influence certain diseases.[285] Additionally, data mining can be used to identify hidden relationships in biological datasets such as "[i]nfants whose genome has a base pair corruption at position 100,290 on chromosome 11 are four times more likely to suffer the onset of Alzheimer's disease before the age of 55 than is the average person."[286] Similarly, the information derived from the data mining can be used by pharmaceutical companies in the design and production of prescription drugs.[287]

*Financial*

Data mining applications in the financial context are quite varied. Ponniah explains that fraud detection, risk assessment of potential customers, trend analysis, and direct marketing are the primary data mining applications at banks.[288] More specifically, data mining can be used to assess the financial status of a customer and to identify his creditworthiness in order to determine whether the bank should grant the customer a loan.

Ponniah further explains that in the financial area forecasting requirements dominate.[289] That is, there is a desire to forecast stock prices and com-

282. *See id.*

283. PONNIAH, *supra* note 125.

284. JIAWEI HAN, HOW CAN DATA MINING HELP BIO-DATA ANALYSIS? (2002), *available at* http://www.cs.uiuc.edu/~hanj/pdf/biokdd02.pdf.

285. Chirag J. Patel & Atul J. Butte, *Predicting Environmental Chemical Factors Associated with Disease-Related Gene Expression Data*, 3:17 BMC MEDICAL GENOMICS 1 (2010), *available at* http:dx.doi.org/10.1186/1755-8794-3-17.

286. ROB SULLIVAN, INTRODUCTION TO DATA MINING FOR THE LIFE SCIENCES 11 (2012).

287. PONNIAH, *supra* note 125.

288. *Id.*

289. *Id.*

modity prices in order to increase profits.[290] Of course, it is also useful to forecast potential financial disasters.[291] Ponniah notes that neural network algorithms are particularly useful in forecasting, options and bond trading, portfolio management, and in mergers and acquisitions.[292]

*Healthcare*

In healthcare, data mining is becoming essential because the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods.[293] For example, data mining can assist healthcare organizations in making customer relationship management decisions.[294] It can also assist physicians with identifying effective treatments because data mining can help identify the patterns of successful medical therapies for different illnesses.[295] Data mining can also be used to give patients access to more affordable healthcare services.[296]

*Telecommunications*

Because telecommunication companies routinely generate and store enormous amounts of high-quality data, have large customer bases, and operate in a rapidly changing and highly competitive environment, it is not surprising that it was one of the first industries to adopt data mining.[297] Data mining is applied to retain customers, better understand customer behavior that indicates increased line usage in the future, discover profitable service packages, and understand which customers are most likely to churn.[298] Data mining is also used to detect fraud, and monitor and maintain telecommunication networks.[299]

290. *Id.*

291. *Id.*

292. *Id.*

293. Hian Chye Koh & Gerald Tan, *Data Mining Applications in Healthcare*, 19 J. HEALTHCARE INFO. MGMT. 64 (2005).

294. *Id.*

295. *Id.*

296. *Id.*

297. Gary M. Weiss, *Data Mining in the Telecommunications Industry*, *in* ENCYCLOPEDIA OF DATA WAREHOUSING AND MINING 486 (John Wang ed., 2d ed., 2008).

298. PONNIAH, *supra* note 125.

299. Weiss, *supra* note 297.

*Higher Education*

Data mining can be applied in higher education for uses, such as ascertaining which students are taking the most credit hours or which students are most likely to return for more classes.[300] It can also be applied to determine which alumni are likely to make larger donations and which types of courses will attract more students.[301] It could further be used to better understand how students learn. Bienkowski et al. explains, "[n]ew computer-supported interactive learning methods and tools —intelligent tutoring systems, simulations, games — have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn."[302]

*Search Engines*

There is a huge amount of information available on the Internet and search engine suppliers, such as Google, rely extensively on data mining.[303] Google not only relies on data mining for the delivery of its customized search service to its user but it also uses data mining to sell targeted ads, among other things.[304] Wang and Liu explain:

> Google uses data mining techniques extensively on the vast universe of their web data. The techniques include probabilistic models for page rank, text mining, spell check, and statistical language translation that may involve hundreds of languages around the globe. . . . Google utilizes millions of variables about its users and advertisers in its predictive modeling to deliver the message to which each user is most likely to respond."[305]

Google even states on its website that "[a]t Google, much of our work on our primary products like search, social, and ads relies on large-scale data mining."[306]

300. Luan Jing, *Data Mining Applications in Higher Education*, SPSS EXEC. REPORT 4 (2004), *available at* http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf.

301. *Id.*

302. Bienkowski et al., *supra* note 276.

303. Chamont Wang & Pin-Shuo Liu, *Data Mining and Hotspot Detection in an Urban Development Project*, 6:3 J. DATA SCI. 389, 390 (2008).

304. *Id.*

305. *Id.*

306. DATA MINING, RESEARCH AT GOOGLE, http://research.google.com/pubs/DataMining.html (last visited Sept. 2, 2013).

*Sports*

Huge amounts of data are collected across all domains of sports, which may be data based on individual player performance, coaching or managerial decisions, game-based events and/or how well the team functions as a whole.[307] Data mining can be used to make sense of this information in a number of ways. For example, coaches can use data mining to identity player patterns that box scores do not reveal, which makes it easier for coaches to determine when and how to position their players for maximum effect.[308]

## B. Public-Sector Applications

Because of the complexity of the public sector, it is difficult to precisely pinpoint its borders. Broadly, it can be understood to include government agencies, ministries, and other types of government and not-for-profit organizations. In Europe, education and healthcare would fall under the rubric of "public sector," although the structure is becoming multifaceted to allow for some commercial providers of education and healthcare, whereas in the United States they generally would not. These organizations often collect huge amounts of information, some of which is highly sensitive personal data, such as data related to child abuse, mental illness, juvenile incarceration, and other personal data.

### 1. Contextual Examples

*Social Security and Social Welfare*

Data mining can be applied with respect to the auditing government benefits and social service programs in order to save huge sums of money through avoiding fraudulent claims and overpayments.[309] Data mining can also be used to identify noncompliance with government mandates, such as compliance with tax rules.[310] Cao explains that "[p]eople working in different communities are increasingly interested in 'what do social security data show' and recognize the value of data-driven analysis and decisions to enhance public service objectives, payment accuracy, and compliance."[311]

307. Robert P. Schumaker, Osama K. Solieman & Hsinchun Chen, *Sports Data Mining: The Field, in* SPORTS DATA MINING 1 (2010).

308. H. Baltazar, *NBA Coaches' Latest Weapon: Data Mining*, 17:10 PC WEEK 69 (2000).

309. Longbing Cao, *Social Security and Social Welfare Data Mining: An Overview*, 42:6 IEEE TRANS.: SMC PART C 837, 838, 840 (2012).

310. *Data Mining: Federal Efforts Cover a Wide Range of Uses*, GOV'T ACCOUNTABILITY OFF., May 2004, http://www.gao.gov/assets/250/242240.html.

311. Cao, *supra* note 309, at 837.

*Human Resources and Internal Operations Management*

Government agencies use data mining not only to uncover fraud and waste, but also to improve and evaluate program performance.[312] For example, data mining can be used to advance internal operations management as evidenced by the U.S. Department of Justice's successful use of data mining to more efficiently allocate agency resources.[313] Data mining can further be used to improve human resources as demonstrated through its application by a human resources department to determine which employees generate value to the organization.[314] Data mining can even be used to predict whether an employee is likely to suffer an accident.[315]

*E-government*

E-government concerns the use of information communication technologies and e-commerce to provide access to government information and delivery of public services to citizens and business partners.[316] E-governance data is huge and e-governance organizations can use data mining to analyze it. For example, data mining can be applied within the context of e-government in order to better understand citizen needs.[317] It can further be used to provide faster access to critical data about service status while increasing the value of information for those who make decisions on different levels of the government.[318] It can also be applied to gain more operational effectiveness and to sustain organizational knowledge.[319]

312. *Data Mining: Federal Efforts Cover a Wide Range of Uses*, GOV'T ACCOUNTABILITY OFF. (2004) *available at* http://www.gao.gov/assets/250/242240.html.

313. *Principles for Government Data Mining: Preserving Civil Liberties in the Information Age*, CONSTITUTION PROJECT 12 (2010), http://www.constitutionproject.org/pdf/DataMiningPublication.pdf.

314. Stephen Baker, *Data Mining Moves to Human Resources*, BUSINESSWEEK, Mar. 11, 2009, http://www.businessweek.com/stories/2009-03-11/data-mining-moves-to-human-resources.

315. *Id.*

316. Yilei Wang, Hui Pan &Tao Li, *The Data Mining of the e-Government on the Basis on Fuzzy Logic, Integration Technology, 2007*, ICIT '07: IEEE INT'L CONFERENCE 774 (2007).

317. M. Hanumanthappa, B. R. Prakash & Manish Kumar, *Applications of Data Mining in e-Governance: A Case Study of Bhoomi Project, in* DATA ENGINEERING AND MANAGEMENT LECTURE NOTES, COMPUTER SCIENCE Vol. 6411, at 212 (2012).

318. *Id.*

319. *Id.*

*Politics*

Data mining is causing an "analytical revolution" in the context of campaigning.[320] For example, in the 2012 United States Election, data mining helped reveal that Obama supporters often "[eat at] Red Lobster, shop at Burlington Coat Factory and listen to smooth jazz" while Romney supporters are "more likely to drink Samuel Adams beer, eat at Olive Garden and watch college football."[321] This kind of information was used by strategists working for each campaign in order to target voters and get backers to the polls on Election Day.[322] Other types of data mining techniques can be used to understand which geographic areas are likely to "make nuanced decisions about how to most effectively deploy resources at every stage of a campaign, from donor prospecting to message targeting. . ."[323] Data mining can also be used in the political arena to understand which politicians have more power than others and to gather connections between different political groups.[324]

## C. Law-Enforcement Sector Applications

Law enforcement has access to huge amounts of data including, but not limited to, offender demographic information, criminal background information, previous investigation files, police arrest records, photographs, video files, bank accounts, credit card statements, call detail and email records, travel and flight itineraries, intelligence reports, open source intelligence findings, service records, hotel and hospital records, and police reports.[325] Law enforcement can use data mining to analyze these data to help "investigate crimes or to enhance their understanding of criminal patterns and behavior."[326] For example, data mining can be used to detect common attributes of crimes, detect relationships between criminals, detect a series of crimes and

320. *How Data Mining is Causing Analytical Revolution for Campaigns' Victory Strategy*, PBS, Sept. 14, 2012, http://www.pbs.org/newshour/bb/politics/july-dec12/victorylab_09-14.html.

321. Charles Duhigg, *Campaigns Mine Personal Lives to Get Out Vote*, N.Y. TIMES, Oct. 13, 2012, http://www.nytimes.com/2012/10/14/us/politics/campaigns-mine-personal-lives-to-get-out-vote.html?pagewanted=all&_r=0.

322. *Id.*

323. *Voters/Donors Analysis and Data Mining*, AZAVEA, http://www.azavea.com/products/cicero/services/political-and-elections-projects/voters-donors-analysis-and-data-mining/ (last visited Sept. 2, 2013).

324. *See* A. Jakulin, W. Buntine, T. M. La Pira & H. Brasher, *Analyzing the U.S. Senate in 2003: Similarities, Clusters, and Blocs*, 17:3 POLITICAL ANALYSIS 291, 292 (2009).

325. F. Ozgul, C. Atzenbeck, A. Celik & Z. Erdem, *Incorporating Data Sources and Methodologies for Crime Data Mining*, INTELLIGENCE AND SEC. INFORMATICS (ISI): 2011 IEEE INT'L CONFERENCE 176–77 (2011).

326. CONSTITUTION PROJECT, *supra* note 313, at 13.

detect cliques and subgroups in criminal networks.[327] Controversially, some commentators also contend that data mining can be used to make predictions about crimes such as the next attack of a serial criminal or a prediction of a missing member in a criminal network.[328]

### 1. Contextual Examples

*Child Abuse*

It is often difficult for states to prevent child abuse from happening. This is because if there is a strong indication about a child's maltreatment, it tends to come from one particular source, and it tends to arrive too late to prevent the abuse.[329] Data mining can assist the state in preventing child abuse through a timely recognition of patterns.[330] For example, researchers from Children's Hospital Boston used data mining to make sense of an anonymous database of 560,000 computerized medical records to develop a model that shows an ability to detect cases of domestic abuse.[331]

*Cybercrime*

Detecting cybercrime is difficult because there is not only a huge amount of information available online, which travels across busy networks, but only a small percentage relates to illegal activities so this information is usually anonymous in nature.[332] It is becoming increasingly apparent that conventional methods of dealing with the large amount of cyberspace activities and their anonymous nature are insufficient, and that manual effort must be coupled with computational efforts such as data mining. For example, Chen proposes using data mining to automatically trace identities of cyber criminals through an analysis of the messages they leave in cyberspace.[333] Chen explains that "[u]nder this framework, three types of message features, including style markers, structural features, and content-specific features, are

327. Ozgul, Atzenbeck & Erdem, *supra* note 325.

328. *Id.*

329. Max Voskob, Rob Howey & Nick Panin, *Data Mining and Privacy in Public Sector Using Intelligent Agents*, CORNELL UNIV. LIBR., Nov. 2003, http://arxiv.org/pdf/cs/0311050.pdf.

330. *See id.*

331. Gene Ostrovsky, *Data Mining to Help Detect Domestic Abuse*, MEDGADGET, Oct. 7, 2009, http://www.medgadget.com/2009/10/data_mining_to_help_detect_domestic_abuse.html.

332. Hsinchun Chen et al., *Crime Data Mining: A General Framework and Some Examples*, 37:4 COMPUTER: IEEE Computer Soc'y, D.C. 50 (2004).

333. Hsinchun Chen et al., *Crime Data Mining: An Overview and Case Studies*, CITESEERX (2003), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.4879&rep=rep1&type=pdf.

extracted and inductive learning algorithms are used to build feature-based models to identify authorship of illegal messages."[334]

*Hate Crime*

Hate crimes not only include a wide variety of criminal behavior, but they differ greatly in their severity and impact on the broader community.[335] Data mining offers the possibility to better understand the nature of hate crimes. For example, one study used data mining to investigate the motives and underlying factors behind hate crimes in a particular geographical region in Turkey.[336] Through the use of data mining, the study was able to deduce hidden information in the criminal data, such as "[w]hen committed in night time[,] [hate crimes] are more planned, individual and professional whereas day time crimes are more spontaneous and co-defendants are mostly from the same family and young."[337]

*Traffic Violations*

Data mining can be used to analyze different types of automobile traffic data. For example, data mining can be used to better understand the kinds of people that commit particular traffic violations and to compare traffic accidents with other variables, such as weather, time, and road type.[338] This information could be used to apprehend offenders, prevent future accidents, and save lives.

*Organized Crime*

Criminals often interact with one another to carry out various illegal activities.[339] Criminal networks take on many different attributes depending on the nature of the underlying criminal objective. Shaikh explains:

> In criminal networks, there may exist groups or teams, within which members have close relationships. One group also may interact with other groups to obtain or transfer illicit goods. Moreover, individuals play different roles in their groups. For example,

334. *Id.*

335. F. Ozgul et al., *Mining Hate Crimes to Figure Out Reasons Behind*, INT'L CONFERENCE ON ADVANCES IN SOC. NETWORKS ANALYSIS & MINING (Aug. 26-29, 2012), *in* COMPUTER, 2012, at 887 (citing J. LEVIN & J. MCDEVITT, HATE CRIMES REVISITED: AMERICA'S WAR ON THOSE WHO ARE DIFFERENT).

336. *Id.*

337. *Id.* at 888.

338. *See* Wei Cheng et al., *The Mining Method of the Road Traffic Illegal Data Based on Rough Sets and Association Rules*, INT'L CONFERENCE ON INTELLIGENT COMPUTATION TECH. & AUTOMATION 856 (2010).

339. M.A. Shaikh & Jiaxin Wang, *Investigative Data Mining: Identifying Key Nodes in Terrorist Networks*, IEEE MULTITOPIC CONFERENCE 201 (2006).

> some key members may act as leaders to control activities of a group. Some others may serve as gatekeepers to ensure smooth flow of information or illicit goods and some act as outliers in a group.[340]

Data mining can be used to facilitate understanding about the nature of criminal networks. This is especially true since analyzing criminal networks often involves processing large amounts of criminal data gathered from a number of different places.[341]

There is also a pressing need to generate knowledge about criminal networks in real-time or close to real-time, which is often difficult to do without automated techniques.[342] For example, data mining can be used to uncover previously unknown information about criminal networks, such as subgroups and patterns of group interaction.[343] Shaikh explains that it might also be possible to connect every-day transactions of criminals, such as applications for passports, visa, car rentals, purchase of airline tickets and chemicals, with events, such as arrests or suspicious activities.[344]

A well-known example of data mining in the context of organized crime is the United States-based COPLINK project. The COPLINK project applies data mining techniques that learn patterns and association to police databases in order to identify connections among suspects, vehicles, crimes, locations and other data to provide investigative leads.[345] For example, classification techniques are used to find the common properties among different crime entities and to classify them into specific groups.[346]

---

340. *Id.*

341. *Id.*

342. *See id.* ("When there is a pressing need to untangle criminal networks, manual approaches may fail to generate valuable knowledge in a timely manner.").

343. Hsinchun Chen et al., *Crime Data Mining: An Overview and Case Studies*, CITESEERX (2003), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.4879&rep=rep1&type=pdf.

344. Shaikh & Wang, *supra* note 339, at 202.

345. JESUS MENA, HOMELAND SECURITY TECHNIQUES AND TECHNOLOGIES 308 (2004).

346. *Id.* at 310 ("The COPLINK crime analysis techniques include association rules, the process of discovering frequently occurring criminal elements in a database. This technique also includes intrusion detection to identify patterns of program execution and user activities as association rules. Another data mining technique is that of classification, the process of finding the common properties among different crime entities and classifying them into groups. Lastly, COPLINK is able to perform clustering, the process of grouping criminal items into classes of similar characteristics.").

## D. State-Security Applications

Governments around the world have increasingly been applying data mining techniques in an effort to seek out terrorists and secure national borders.[347] They are building databases and deploying data mining to mine law enforcement, communications, and intelligence data.[348] This trend is likely to continue as "[t]he volume of electronically available data is expanding every year along with increased computing power that can be used to exploit it."[349]

Although data mining is increasingly being recognized as one of the most salient technologies in the national security context, it is still in a phase of development. Experts have lined up both for and against data mining in this context.[350] A key argument against data mining in this context is that while governments has access to large volumes of data that can provide clues about national security threats, relevant information about national security threats is typically hidden within vast amounts of irrelevant data that appears innocuous when viewed in isolation.[351]

### 1. Examples

*Cyber Security*

Cyber security involves protecting a nation's computers and networks from threats and attacks, both internal and external.[352] Data mining can be used within this context to safeguard computer and network systems from corruption due to malicious actions that could affect the integrity, confidenti-

347. *See, e.g.*, CONSTITUTION PROJECT, *supra* note 315.

348. *Id.*

349. Stew Magnuson, *Data Mining Not a Panacea for Catching Terrorists, Experts Warn*, NAT'L DEF. INDUS. ASS'N (Feb. 1, 2011), http://www.nationaldefensemagazine.org/archive/2011/February/Pages/DataMiningNotaPanaceaforCatchingTerrorists,ExpertsWarn.aspx.

350. *See* U.S. DEPT. OF DEF. TECH. & PRIVACY ADVISORY COMM., SAFEGUARDING PRIVACY IN THE FIGHT AGAINST TERRORISM (2004) ("[T]he ubiquity of information networks and digital data has created new opportunities for tracking terrorists and preventing attacks."); *cf.*, Bruce Schneier, *Why Data Mining Won't Stop Terror*, WIRED, Mar. 9, 2006, http://www.wired.com/politics/security/commentary/securitymatters/2006/03/70357?currentPage=all ("[T]he promise of data mining is compelling, and convinces many. But it's wrong. We're not going to find terrorist plots through systems like this, and we're going to waste valuable resources chasing down false alarms.").

351. Shaikh & Wang, *supra* note 339, at 201.

352. Bhavani M. Thuraisingham et al., *Data Mining for Security Applications*, IEEE/IFIP INT'L CONFERENCE ON EMBEDDED & UBIQUITOUS COMPUTING 585, 586 (2008), http://www.utdallas.edu/~hamlen/thuraisingham-euc08.pdf.

ality, and availability of information resources.[353] More specifically, data mining can be used to assist in locating anomalous events on a network by building models of normal behavior and automatically detecting significant deviations from it.[354] It could also be used to assist in locating patterns and signatures of previously known attacks.[355] Finally, data mining may even be able to predict cyber-attacks in advance.[356]

*Border & Transportation Security*

Data mining can be used to secure national borders. By analyzing travel information such as traveler identities, images, fingerprints and vehicles used, governments can ostensibly improve state security.[357] A well-known example of how data mining is applied in this context is the United States government's Automated Targeting System ("ATS"). Initially, ATS was a system designed to screen incoming cargo through performing abnormal weight analysis.[358] Its aim was to avert an attack by way of the nation's global trade infrastructure.[359]

ATS has been extended to perform data mining on travelers.[360] The system analyzes data like address information, financial records, "no show" history, ticket purchase information, motor vehicle records, past instances of one-way travel and seating and meal preferences in order to assign individuals who cross the US border with a "risk assessment."[361] The risk assessment is used for a multiplicity of different purposes, such as determining which individuals should be stopped for additional questioning, and the score may be kept on file for up to forty years.[362]

353. A. Al-Shawi, *Data Mining Techniques for Information Security Applications*, 3 WIRES: COMPUTATIONAL STATISTICS, no. 3, 2011, at 221–29.

354. HAN & KAMBER, *supra* note 6, at 659.

355. Al-Shawi, *supra* note 353, at 221–29.

356. B. Thuraisingham, *Data Mining for Malicious Code Detection and Security Applications*, IEEE/WIC/ACM INT'L JOINT CONFERENCES (Sept. 15–18, 2009), 6–7.

357. S. Ahsan & A. Shah, *Data Mining, Semantic Web and Advanced Information Technologies for Fighting Terrorism*, INT'L SYMP. BIOMETRICS & SECURITY TECH. *2 (Apr. 23–24, 2008).

358. *See, e.g.*, D. R. Chambers et al., *How Dangerous Are Measurement Errors to Homeland Security?*, 52 THUNDERBIRD INT'L BUS. REV. 553 (2010).

359. *Id.*

360. Louise Amoore, *Risk Before Justice: When the Law Contests its Own Suspension*, 21:4 LEIDEN J. INT'L L. 847, 854 (2008).

361. *Id.*

362. *Id.*

*Protecting Critical Infrastructure*

Critical infrastructure encompasses a large number of sectors including agriculture, food and water sectors, public health and emergency services sectors, energy, transportation, banking and finance, chemical industry, and postal and shipping sectors.[363] The vulnerability of these sectors makes them a potential security threat because of the essential role the sectors play in society. Data mining is increasingly being applied to monitor these assets. For example, data mining can be used to model normal use behaviors and to distinguish abnormal behaviors from them, which can guide the selection of protective or reactive measures to secure these assets from attacks.[364] Also, data mining can be used by governments to extract open-source data about critical infrastructures, which are often held in the hands of private parties that may not freely communicate with governments.[365]

*Anti-Money Laundering*

Money laundering and terrorist financing pose a serious threat to national security.[366] It is contended that data mining can be used to detect terrorist financing and money laundering. For example, Moorman explains that rule-based systems can be designed to uncover patterns associated with criminal financial activity and that anomaly-based systems can detect specific transactions that deviate from normal behavior.[367] More specifically, data mining can explore operational data related to transactions in a financial organization in order to assess the origin and purpose of these transactions and to detect if they are relevant to money laundering.[368] Data mining can also be used to detect trade-based money laundering which "allows illegal organizations the opportunity to earn, move and store proceeds disguised as legitimate trade."[369]

363. DHS, *The National Strategy for Homeland Security*, July 16, 2002, at 30.

364. Ahsan & Shah, *supra* note 357, at 2.

365. *See* William J. Tolone, Wei-Ning Xiang, Anita Raja, David Wilson, Qianhong Tang & Ken McWilliams, *Mining Critical Infrastructure Information from Municipality Data Sets: A Knowledge-Driven Approach and Its Implications*, *in* EMERGING SPATIAL INFORMATION SYSTEMS AND APPLICATIONS 312 (Brian Hilton ed., 2007).

366. Nhien An Le Khac, Sammer Markos & M-Tahar Kechadi, *A Data Mining-Based Solution for Detecting Suspicious Money Laundering Cases in an Investment Bank*, DBKDA, Apr. 11–16, 2010, at 235.

367. M. Moorman, *Detecting Terrorist Financing*, AM. BANKER, Sept. 24, 2004.

368. George Ibrahim & Manolya Kavakli, *Data Mining in the Investigation of Money Laundering and Terrorist Financing*, *in* SURVEILLANCE TECHNOLOGIES AND EARLY WARNING SYSTEMS: DATA MINING APPLICATIONS FOR RISK DETECTION, 228–41 (Ali Serhan Koyuncugil & Nermin Ozgulbas eds., 2011).

369. *Trade-Based Money Laundering*, ICE (Sept. 2, 2013), *available at* http://www.ice.gov/cornerstone/money-laundering.htm.

*Signals Intelligence*

Signals intelligence ("SIGNIT") is a form of intelligence collection that is derived from the interception of signals, including communications signals, electronic emissions, and telemetry.[370] It is a window into the capabilities, actions, and intentions of a foreign adversary and therefore plays an important role in safeguarding a nation's security and territorial integrity.[371] For example, signals intelligence could provide warning of a military attack, a cyber-attack, or the proliferation of weapons of mass destruction.[372]

Data mining is extremely useful in the context of signals intelligence. This is because data mining can transform massive amounts of communication messages into a higher form of knowledge so that threats can be discerned and actionable intelligence can be provided to national leadership.[373] More specifically, data mining can be used to shed light on connections between individuals, to pick up clues about territorial attacks, and to simply make more efficient use of human investigators.[374]

A notorious case of data mining in the context of signals intelligence is the U.S.'s "Total Information Awareness" ("TIA") Program. The program was established in January 2002 by the Department of Defense through the Defense Advanced Research Projects Agency ("DARPA").[375] The focus of TIA was "to develop and integrate a variety of information technologies enabling early warning and understanding of terrorist activities and ultimately the preemption of terrorist attacks through collaborative decision making at the national security level."[376] The data mining techniques applied in the program included, among others, model matching, hypotheses generation and

370. Judson Knight, *SIGINT (Signals Intelligence)*, *in* ENCYCLOPEDIA OF ESPIONAGE, INTELLIGENCE, & SECURITY, *available at* http://www.espionageinfo.com/Se-Sp/SIGINT-Signals-Intelligence.html.

371. SIGNALS INTELLIGENCE, NSA, *available at* http://www.nsa.gov/sigint.

372. *Centrum för rättvisa v. Sweden*, OBSERVATIONS OF THE GOVERNMENT OF SWEDEN ON ADMISSIBILITY, Apr. 27, 2012, ECtHR, at 22, http://centrumforrattvisa.se/wp-content/uploads/2012/09/Regeringens-svar-FRA.pdf.

373. NAT'L RES. COUNCIL OF THE NAT'L ACADEMIES, GETTING UP TO SPEED: THE FUTURE OF SUPERCOMPUTING 77 (Susan L. Graham, Marc Snir & Cynthia A. Patterson, eds., 2004); *see also* SIGNALS INTELLIGENCE, GOV'T SWEDEN, *available at* http://www.government.se/sb/d/10941.

374. Mazmanian, *supra* note 222.

375. Terrence A. Maxwell, *Information Policy, Data Mining, and National Security: False Positives and Unidentified Negatives*, HICSS 2005, Jan. 3–6, 2005 at 2.

376. John Poindexter, Robert Popp & Brian Sharkey, *Total Information Awareness (TIA)*, IEEE 1.

pattern extraction.[377] TIA was discontinued after public outcry over privacy concerns.

## E. Personal Data Mining

It is not just governments and commercial companies that are interested in data mining: individuals also want to make sense of the huge amounts of data in their lives in order to uncover hidden knowledge about themselves and their world.[378] One can only imagine what 30 years of email might reveal about an individual's habits, behaviors and friends.[379] This information could be used in the fulfillment of personal goals such as becoming more efficient, healthy or knowledgeable about individual strengths and weaknesses: it even may reveal when an individual is most likely to come up with new ideas.[380]

### 1. Lifebrowsers

Lifebrowsers are being offered as a tool to assist individuals to explore their own sets of personal data including e-mails, Web browsing and search history, calendar events, and other documents stored on a person's computer.[381] Data mining techniques are used to sift through personal data and determine what is important to its owner.[382] Horvitz, a scientist at Microsoft who created Lifebrowser, explains that, "[w]e were interested in making local machines private data-mining centers [that are] very smart about you and your memory so that you can better navigate through that great amount of content."[383]

377. Maxwell, *supra* note 375, at 3 (citing Gregory Mack, B. Bebee & G. Wenzel, *Total Information Awareness System Description Document*, DARPA, at 16 (2002)).

378. *See* Liane Colonna, *Mo' Data, Mo' Problems? Personal Data Mining and the Challenge to the Data Minimization Principle*, Workshop Proceedings of "Big Data and Privacy" hosted by the Future of Privacy Forum and the Stanford Law School's Center for Internet and Society, Sept. 2013, *available at* http://www.futureofprivacy.org/wp-content/uploads/Colonna-Mo-Data-Mo-Problems.pdf#!.

379. Anne Eisenberg, *What 23 Years of E-Mail May Say About You*, N.Y. TIMES, Apr. 7, 2012, http://www.nytimes.com/2012/04/08/business/mining-our-personal-data-for-our-own-good.html.

380. *Id.*

381. Tom Simonite, *Microsoft Builds a Browser for Your Past*, MIT TECH. REV., Mar. 15, 2012, http://www.technologyreview.com/news/427233/microsoft-builds-a-browser-for-your-past.

382. *Id.*

383. *Id.*

### 2. Mining Social Media

Social media, such as Twitter and Facebook, offer rich sources of personal information. The information revealed includes everything from individual speech patterns to the topics individuals obsess over to the identity of an individual's "real" friends.[384] For example, personal data mining can be used to "unlock" this information, for example, by better understanding what an individual says when they talk to their friends on Facebook or Twitter.[385]

**FIGURE 5: DATA-MINING APPLICATION DOMAINS, FUNCTIONS & CONTEXTS**

This figure classifies data mining into five main groups: commercial applications, public-sector applications, law-enforcement applications, state-security applications, and personal/private applications. The purpose of this scheme is to find a classification that is related to the structure of current privacy and data protection law in the United States and the European Union. In other words, this classification scheme is offered as a paradigm, albeit an imperfect one, to better understand how different types of data mining applications raise unique data protection/privacy concerns and, consequently, different balancing of interests.

| Application Domain | Examples of mining functions | Examples of contexts |
|---|---|---|
| Commercial sector | Market based analysis<br>Targeted marketing<br>Decreasing customer turnover<br>Detecting fraud<br>Assessing risk<br>Page ranking | • Retail<br>• Biological<br>• Financial<br>• Healthcare<br>• Telecommunications<br>• Higher education<br>• Search engines<br>• Sports |
| Public sector | Fraud detection<br>Risk assessment<br>Reducing recruiting costs<br>Increasing employee retention<br>Improving public health and safety | • Social security and social welfare<br>• Human resources<br>• Internal operations management<br>• E-government<br>• Politics |

384. Christopher Mims, *How to Use Twitter for Personal Data Mining*, MIT TECH. REV., Oct. 13, 2010, http://www.technologyreview.com/view/421201/how-to-use-twitter-for-personal-data-mining.

385. *Id.*

| | | |
|---|---|---|
| Law-enforcement sector | Preventing crime<br>Investigating crime<br>Predicting crime | • Child abuse<br>• Cyber crime<br>• Hate crime<br>• Traffic violations<br>• Organized crime |
| State-security sector | Securing national borders<br>Preventing terrorism<br>Protecting critical infrastructure | • Cyber security<br>• Border and transportation security<br>• Protecting critical infrastructure<br>• Anti-money laundering<br>• Signals intelligence |
| Personal/private sector | To better understand an individual's health, productivity, etc.<br>To uncover hidden information about personal enemies and investigate friends<br>To navigate the digital content of an individual's life | • Lifebrowsers<br>• Mining social media |

## XII. Conclusion

This paper is an attempt to map the granularities and special technological properties of data mining in order to facilitate a better understanding of what data mining means from a legal perspective. The identification of some of the relevant differences and similarities between data mining and other closely related methodologies is intended to provide an understanding of where the boundaries begin to blur and whether there are legal implications created by the obscuring of terms. A general taxonomy is offered in order to provide a clearer understanding of the concept of data mining and to provide a more precise vocabulary to express particular features surrounding the technology. Finally, a classification of different applications of data mining is afforded in an effort to provide the reader with a fuller understanding of how data mining is applied in five different contexts.

Data mining is about identifying patterns and/or relationships in large data sets that are previously unknown. Broadly, it is about satisfying or generating hypotheses and building models for prediction or description. It reveals relationships or patterns that are not obvious by a review of the data with the naked eye or the use of common sense. It requires that some level of automation will be involved. It is a search or an inference rather than a straight-forward computation of predefined quantities like computing the average value of a set of numbers.[386] The results of data mining are not existing data items in the database.[387] That is, information stored explicitly in the

386. Rajni Jain, *Overview of Data Mining*, NCAP, at 2, *available at* http://bioinformatics.iasri.res.in/BAMAST/Lectures/Overview%20of%20Data%20Mining.pdf.

387. *See* Usama M Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth, *From Data Mining to Knowledge Discovery: An Overview*, *in* Advances in Knowl-

database or system catalog is not the subject in data mining.[388] The data mining process is driven by application requirement: the cost of data mining should be rewarded by the benefit.[389] In other words, the results of data mining should lead to some benefit to the end-user. Data mining results are useful when they help in achieving the user's goals, which generally involve some type of decision-making or risk management.[390]

EDGE DISCOVERY AND DATA MINING 11, 12 (Usama M. Fayyad, Gegory Piatetsky-Shapiro, Padhraic Smyth & Ramasamy Uthurusamy eds., 1996).

388. Lee & Ho Kim, *supra* note 20, at 42.

389. *Id.*

390. *Id.*