CHAPTER 11

# Data Analytics

The term **data analytics** refers broadly to the processing of data to infer patterns, correlations, or models for prediction. The results of analytics are then used to drive business decisions.

The financial benefits of making correct decisions can be substantial, as can the costs of making wrong decisions. Organizations therefore invest a lot of money to gather or purchase required data and build systems for data analytics.

## Bibliographical Notes

[Kimball et al. (2008)] and [Kimball and Ross (2013)] provide textbook coverage of data warehouses and multidimensional modeling.

[Mitchell (1997)] is a classic textbook on machine learning and covers classification techniques in detail. [Goodfellow et al. (2016)] is a definitive text on deep learning. [Witten et al. (2011)] and [Han et al. (2011)] provide textbook coverage of data mining. [Agrawal et al. (1993a)] introduced the notion of association rules.

Information about the R language and environment may be found at www.r-project.org; information about the SparkR package, which provides an R frontend to Apache Spark, may be found at spark.apache.org/docs/latest/sparkr.html.

[Chakrabarti (2002)], [Manning et al. (2008)] and [Baeza-Yates and Ribeiro-Neto (2011)] provide textbook description of information retrieval, including extensive coverage of data-mining tasks related to textual and hypertext data, such as classification and clustering.

Definitions of statistical functions can be found in standard statistics textbooks such as [Bulmer (1979)] and [Ross (1999)].

[Zhuge et al. (1995)] describes view maintenance in a data-warehousing environment. [Chaudhuri et al. (2003)] describes techniques for fuzzy matching for data cleaning, while [Sarawagi et al. (2002)] describes a system for deduplication using active learning techniques.

[Fayyad et al. (1995)] presents an extensive collection of articles on knowledge discovery and data mining. [Kohavi and Provost (2001)] presents a collection of articles on applications of data mining to electronic commerce.

[Agrawal et al. (1993b)] provides an early overview of data mining in databases. Algorithms for computing classifiers with large training sets are described by [Agrawal et al. (1992)] and [Shafer et al. (1996)]; the decision-tree construction algorithm described in this chapter is based on the SPRINT algorithm of [Shafer et al. (1996)]. [Cortes and Vapnik (1995)] introduced several key results on Support Vector Machines, while [Cristianini and Shawe-Taylor (2000)] provides textbook coverage of Support Vector Machines.

An efficient algorithm for association rule mining was presented by [Agrawal and Srikant (1994)]. Algorithms for mining of different forms of association rules are described by [Srikant and Agrawal (1996a)] and [Srikant and Agrawal (1996b)]. [Chakrabarti et al. (1998)] describes techniques for mining surprising temporal patterns. Techniques for integrating data cubes with data mining are described by [Sarawagi (2000)].

[Ng and Han (1994)] describes spatial clustering techniques. Clustering techniques for large datasets are described by [Zhang et al. (1996)]. [Breese et al. (1998)] provides an empirical analysis of different algorithms for collaborative filtering. Techniques for collaborative filtering of news articles are described by [Konstan et al. (1997)].

[Chakrabarti (2000)] provides a survey of hypertext mining techniques such as hypertext classification and clustering.

## Bibliography

**[Agrawal and Srikant (1994)]** R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", In *Proc. of the International Conf. on Very Large Databases* (1994), pages 487–499.

**[Agrawal et al. (1992)]** R. Agrawal, S. P. Ghosh, T. Imielinski, B. R. Iyer, and A. N. Swami, "An Interval Classifier for Database Mining Applications", In *Proc. of the International Conf. on Very Large Databases* (1992), pages 560–573.

**[Agrawal et al. (1993a)]** R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1993), pages 207–216.

**[Agrawal et al. (1993b)]** R. Agrawal, T. Imielinski, and A. N. Swami, "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Volume 5, Number 6 (1993), pages 914–925.

**[Baeza-Yates and Ribeiro-Neto (2011)]** R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, 2nd edition, ACM Press (2011).

**[Breese et al. (1998)]** J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", In *Procs. Conf. on Uncertainty in Artificial Intelligence*, Morgan Kaufmann (1998), pages 43–52.

**[Bulmer (1979)]** M. G. Bulmer, *Principles of Statistics*, Dover Publications (1979).

**[Chakrabarti (2000)]** S. Chakrabarti, "Data Mining for Hypertext: A Tutorial Survey", *SIGKDD Explorations*, Volume 1, Number 2 (2000), pages 1–11.

**[Chakrabarti (2002)]** S. Chakrabarti, *Mining the Web: Discovering Knowledge from HyperText Data*, Morgan Kaufmann (2002).

**[Chakrabarti et al. (1998)]** S. Chakrabarti, S. Sarawagi, and B. Dom, "Mining Surprising Patterns Using Temporal Description Length", In *Proc. of the International Conf. on Very Large Databases* (1998), pages 606–617.

**[Chaudhuri et al. (2003)]** S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Robust and Efficient Fuzzy Match for Online Data Cleaning", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (2003), pages 20–26.

**[Cortes and Vapnik (1995)]** C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, Volume 20, Number 3 (1995), pages 273–297.

**[Cristianini and Shawe-Taylor (2000)]** N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, Cambridge University Press (2000).

**[Fayyad et al. (1995)]** U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, MIT Press (1995).

**[Goodfellow et al. (2016)]** I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press (2016).

**[Han et al. (2011)]** J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann (2011).

**[Kimball and Ross (2013)]** R. Kimball and M. Ross, "The Data Warehouse Tookit: The Definitive Guide to Dimensional Modeling", John Wiley and Sons (2013).

**[Kimball et al. (2008)]** R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, "The Data Warehouse Lifecycle Toolkit", John Wiley and Sons (2008).

**[Kohavi and Provost (2001)]** R. Kohavi and F. Provost, editors, *Applications of Data Mining to Electronic Commerce*, Kluwer Academic Publishers (2001).

**[Konstan et al. (1997)]** J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News", *Communications of the ACM*, Volume 40, Number 3 (1997), pages 77–87.

**[Manning et al. (2008)]** C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press (2008).

**[Mitchell (1997)]** T. M. Mitchell, *Machine Learning*, McGraw Hill (1997).

**[Ng and Han (1994)]** R. T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", In *Proc. of the International Conf. on Very Large Databases* (1994), pages 144–155.

**[Ross (1999)]** S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, Harcourt/Academic Press (1999).

**[Sarawagi (2000)]** S. Sarawagi, "User-Adaptive Exploration of Multidimensional Data", In *Proc. of the International Conf. on Very Large Databases* (2000), pages 307–316.

**[Sarawagi et al. (2002)]** S. Sarawagi, A. Bhamidipaty, A. Kirpal, and C. Mouli, "ALIAS: An Active Learning Led Interactive Deduplication System", In *Proc. of the International Conf. on Very Large Databases* (2002), pages 1103–1106.

**[Shafer et al. (1996)]** J. C. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining", In *Proc. of the International Conf. on Very Large Databases* (1996), pages 544–555.

**[Srikant and Agrawal (1996a)]** R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1996), pages 1–12.

**[Srikant and Agrawal (1996b)]** R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", In *Proc. of the International Conf. on Extending Database Technology* (1996), pages 3–17.

**[Witten et al. (2011)]** I. H. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 3rd edition, Morgan Kaufmann (2011).

**[Zhang et al. (1996)]** T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1996), pages 103–114.

**[Zhuge et al. (1995)]** Y. Zhuge, H. Garcia-Molina, J. Hammer, and J. Widom, "View Maintenance in a Warehousing Environment", In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1995), pages 316–327.