



Artificial Intelligence and Big Data

Daniel E. O'Leary, University of Southern California

AI Innovation in Industry is a new department for *IEEE Intelligent Systems*, and this first article will examine some of the basic concerns and uses of AI for big data (AI has been used in several different ways to facilitate capturing and structuring big data, and it has been used to analyze big data for key insights). In future articles, we'll present some case studies that analyze emerging issues and approaches that integrate AI and big data.

What Is Big Data?

Michael Cox and David Ellsworth¹ were among the first to use the term *big data* literally, referring to using larger volumes of scientific data for visualization (the term *large data* also has been used). Currently, there are a number of definitions of big data. Perhaps the most well-known version comes from IBM,² which suggested that big data could be characterized by any or all of three "V" words to investigate situations, events, and so on: volume, variety, and velocity.

Volume refers to larger amounts of data being generated from a range of sources. For example, big data can include data gathered from the Internet of Things (IoT). As originally conceived,³ IoT referred to the data gathered from a range of devices and sensors networked together, over the Internet. RFID tags appear on inventory items capturing transaction data as goods are shipped through the supply chain. Big data can also refer to the exploding information available on social media such as Facebook and Twitter.

Variety refers to using multiple kinds of data to analyze a situation or event. On the IoT, millions of devices generating a constant flow of data results in not only a large volume of data but different types of data characteristic of different situations. For example, in addition to RFID, heart monitors in patients and location information from phones all generate different types of

structured data. However, devices and sensors aren't the only sources of data. Additionally, people on the Internet generate a highly diverse set of structured and unstructured data. Web browsing data, captured as a sequence of clicks, is structured data. However, there's also substantial unstructured data. For example, according to Pingdom,⁴ in 2011 there were 555 million websites and more than 100 million blogs, with many including unstructured text, pictures, audio, and video. As a result, there's an assemblage of data emerging through the "Internet of People and Things"⁵ and the "Internet of Everything."

Velocity of data also is increasing rapidly over time for both structured and unstructured data, and there's a need for more frequent decision making about that data. As the world becomes more global and developed, and as the IoT builds, there's an increasing frequency of data capture and decision making about those "things" as they move through the world. Further, the velocity of social media use is increasing. For example, there are more than 250 million tweets per day.⁴ Tweets lead to decisions about other Tweets, escalating the velocity. Further, unlike classic data warehouses that generally "store" data, big data is more dynamic. As decisions are made using big data, those decisions ultimately can influence the next data that's gathered and analyzed, adding another dimension to velocity.

Big data isn't just volume, variety, and velocity, though; it's volume, variety, and velocity at *scale*. As a result, big data has received substantial attention as distributed and parallel computing has allowed processing of larger volumes of data, most notably through applications of Google's MapReduce.

MapReduce and Hadoop

MapReduce⁶ has been used by Google to generate scalable applications. Inspired by the "map" and

“reduce” functions in Lisp, MapReduce breaks an application into several small portions of the problem, each of which can be executed across any node in a computer cluster. The “map” stage gives subproblems to nodes of computers, and the “reduce” combines the results from all of those different subproblems. MapReduce provides an interface that allows distributed computing and parallelization on clusters of computers. MapReduce is used at Google for a large number of activities, including data mining and machine learning.

Hadoop (<http://hadoop.apache.org>), named after a boy’s toy elephant, is an open source version of MapReduce. Apparently, Yahoo (<http://developer.yahoo.com/hadoop>) is the largest user (developer and tester) of Hadoop, with more than 500 million users per month and billions of transactions per day using multiple petabytes of data.⁷ As an example of the use of the MapReduce approach, consider a Yahoo front page that might be broken into multiple categories—such as advertisements (optimized for the user), must-see videos (subject to content optimization), news (subject to content management), and so on—where each category could be handled by different clusters of computers. Further, within each of those areas, problems might be further decomposed, facilitating even faster response.

MapReduce allows the development of approaches that can handle larger volumes of data using larger numbers of processors. As a result, some of the issues caused by increasing volumes and velocities of data can be addressed using parallel-based approaches.

Contributions of AI

Like Big Data, AI is about increasing volumes, velocities and variety of data.

Under situations of large volumes of data, AI allows delegation of difficult pattern recognition, learning, and other tasks to computer-based approaches. For example, over one-half of the world’s stock trades are done using AI-based systems. In addition, AI contributes to the velocity of data, by facilitating rapid computer-based decisions that lead to other decisions. For example, since so many stock trades are made by AI-based systems rather than people, the velocity of the trades can increase, and one trade could lead to others. Finally, variety issues aren’t solved simply by parallelizing and distributing the problem. Instead, variety is mitigated by capturing, structuring, and understanding unstructured data using AI and other analytics.

Generating Structured Data

AI researchers have long been interested in building applications that analyze unstructured data, and in somehow categorizing or structuring that data so that the resulting information can be used directly to understand a process or to interface with other applications. As an example, Johan Bollen and Huina Mao⁸ found that stock market predictions of the Dow Jones Industrial average were improved by considering the overall “sentiment” of the stock market—this is an unstructured concept, but based on structured data generated from Google.

In another application, firms have begun to investigate the impact of unstructured data issues such as a firm’s reputation. For example, Scott Spangler and his colleagues⁹ reviewed how some firms are analyzing a range of different types of data to provide continuous monitoring of a range of activities, including generating structured measures and assessments of firms’ and products’ reputations, while

in other work I investigated issues such as monitoring and auditing financial and other data streams (in fraud detection, for example).¹⁰

Structuring data has taken multiple approaches. Philip Hayes and Steven Weinstein¹¹ developed a system for use at Reuter’s News Service to help categorize individual news articles. The resulting system categorizes unstructured news articles into around 700 categories, recognizing more than 17,000 company names with an accuracy of 85 percent. As another approach, researchers have begun to generate analysis of unstructured sentiment contained in blogs, Twitter messages, and other text.¹² The nature of those different opinions can be used to investigate a range of issues. For example, after an advertisement has been run, there’s structured transaction information, such as when the ad ran, where it ran, and so on. That transaction information could be aligned with previously unstructured data, such as the number of tweets that mention the ad, along with corresponding positive or negative sentiment in those messages. In addition, AI research often examines what other available data can provide structure. For example, Efthymios Kouloumpis and his colleagues¹³ investigated Twitter messages and found that hashtags and emoticons were useful for ascertaining sentiment. Once the data has been structured, enterprises want to use data mining to develop insights into these kinds of big data—however, some limitations exist that hinder such analysis.

Some Limitations of Current AI Algorithms

Xindong Wu and his colleagues have identified the top 10 data mining algorithms.¹⁴ Unfortunately, available algorithm sets often are non-standard

and primarily research-based. Algorithms might lack documentation, support, and clear examples. Further, historically the focus of AI has largely been on single-machine implementations. With big data, we now need AI that's scalable to clusters of machines or that can be logically set on a MapReduce structure such as Hadoop. As a result, effectively using current AI algorithms in big data enterprise settings might be limited.

However, recently, MapReduce has been used to develop parallel processing approaches to AI algorithms. Cheng-Tao Chu and his colleagues introduced the MapReduce approach into machine learning to facilitate a parallel programming approach to a variety of AI learning algorithms.¹⁵ Their approach was to show that they could write the algorithms in what they referred to as a *summation approach*, where sufficient statistics from the subproblems can be captured, aggregated, and solved. Using parallel processing, they obtained a linear speedup by increasing the number of processors.

Consistent with that development, along with Hadoop, there's now a machine learning library with capabilities such as recommendation mining, clustering, and classification, referred to as *Mahout* (Hindi for a person who rides an elephant; see <http://mahout.apache.org>). Accordingly, this library can be combined with Hadoop to facilitate the ability of enterprises to use AI and machine learning in a parallel-processing environment analysis of large volumes of data.

Parallelizing Other Machine Learning Algorithms

AI researchers are increasingly drawn to the idea of integrating AI capabilities into parallel computing.

For example, Tim Kraska and his colleagues,¹⁶ as well as others, have initiated research on issues in machine learning in distributed environments. However, AI researchers might not be familiar with issues such as parallelization. As a result, teams of AI and parallel computing researchers are combining efforts.

As part of the MapReduce approach, the “map” portion provides subproblems to nodes for further analysis, to provide the ability to parallelize. Different AI approaches and different units of analysis potentially can influence the extent to which algorithms can be attacked using MapReduce approaches and how the problems can be decomposed. However, in some cases, algorithms developed for single-machine environments can be extended readily to parallel processing environments.

Although the system in Hayes and Weinstein¹¹ was developed prior to MapReduce developments, we can anticipate implementing it in such an environment. Because the algorithm categorizes individual news stories independently, one approach to decomposing the data into subproblems would be to process each news story separately in a cluster. As another example, Soo-Min Kim and Eduard Hovy¹² analyzed sentiment data at the sentence level, generating structured analysis of unstructured data. If sentences are processed independently, then subproblems can be developed for the sentence level. Similarly, if the unit of analysis is hashtags or emoticons, then subproblems can be generated for those artifacts. If the task is monitoring transactions or other chunks of data,⁹ then individual transactions and chunks can be analyzed separately in parallel. As a result, we can see that AI algorithms designed for single-machine environments might have emergent

subproblem structures useful for parallelization.

Emerging Issues

There are a number of emerging issues associated with AI and big data. First, unfortunately, the nature of some machine-learning algorithms—for example, iterative approaches such as genetic algorithms—can make their use in a MapReduce environment more difficult. As a result, researchers such as Abhishek Verma and his colleagues¹⁷ are investigating the design, implementation, and use of genetic algorithms and other iterative approaches on Hadoop.

Second, it follows that with big data there will also be *dirty data*, with potential errors, incompleteness, or differential precision. AI can be used to identify and clean dirty data or use dirty data as a means of establishing context knowledge for the data. For example, “consistent” dirty data might indicate a different context than the one assumed—for example, data in a different language.

Third, since data visualization was one of the first uses of big data, we would expect AI to further facilitate additional developments. One approach could include capturing expert visualization capabilities in a knowledge base designed to facilitate analysis by other users as big data permeates the enterprise. Another approach is to make intelligent data visualization apps available, possibly for particular types of data.

Fourth, as flash-storage technology evolves, approaches such as in-memory database technology becomes increasingly feasible¹⁸ to potentially provide users with near-real-time analyses of larger databases, speeding decision-making capabilities. With in-memory approaches, business logic and algorithms can be stored with the data, and AI research can include developing

approaches to exploit that technology. However, it's likely that as technology increases the ability to handle more data faster, there will also be interest in even larger datasets and even more kinds of data, such as that available from audio and video sources.

Fifth, up to this point, when we've talked about big data, we've taken a more traditional approach that treats big data as being information typically available in a database, as a signal or text format. However, looking forward we can anticipate that big data will begin to include more audio- and video-based information. Natural language, natural visual interpretation, and visual machine learning will become increasingly important forms of AI for big data, and AI-structured versions of audio and video will be integrated along with other forms of data.

Although recent use of the term "big data" has grown substantially, perhaps the term will someday be found inappropriately descriptive, once what is seen as big data changes with computing technology and capabilities: the scale of big data of today is likely to be little or small data in 10 years. Further, the term is likely to splinter, not unlike the term *artificial intelligence*, as different approaches or subdomains gain attention.


In any case, right now big data is enabling organizations to move away from intuitive- to data-based decision making. Ultimately, enterprises will use big data because it creates value by solving new problems, as well as solving existing problems faster or cheaper, or providing a better and richer understanding of those problems. As a result, a key role of machine learning and AI is to help create value by providing enterprises with

intelligent analysis of that data, and capturing structured interpretations of the wide variety of unstructured data increasingly available. ■

References

1. M. Cox and D. Ellsworth, "Managing Big Data for Scientific Visualization," *Proc. ACM Siggraph*, ACM, 1997, pp. 5-1-5-17.
2. P. Zikopoulos et al., *Harness the Power of Big Data*, McGraw-Hill, 2013.
3. K. Ashton, "That 'Internet of Things' Thing," *RFID J.*, 22 June 2009; www.rfidjournal.com/article/view/4986.
4. Pingdom, "Internet 2011 in Numbers," tech. blog, 17 Jan. 2012; <http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers>.
5. C. Chu et al., "Map-Reduce for Machine Learning for Multicore," *Proc. Neural Information Processing Systems Conf.*, Neural Information Processing Systems Foundation, 2006; <http://books.nips.cc/nips19.html>.
6. J. Dean, and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Comm. ACM*, vol. 51, no. 1, 2008, pp. 107-113.
7. E. Baldeschwieler, "Hadoop @ Yahoo," 2009 Cloud Computing Expo, presentation, 2009; www.slideshare.net/ydn/hadoop-yahoo-internet-scale-data-processing.
8. J. Bollen and H. Mao, "Twitter Mood as a Stock Market Predictor," *Computer*, vol. 44, no. 10, 2011, pp. 91-94.
9. S. Spangler et al., "COBRA—Mining Web for Corporate Brand and Reputation Analysis," *Web Intelligence and Agent Systems*, vol. 7, no. 3, 2009, pp. 243-254.
10. D.E. O'Leary, "Knowledge Discovery for Continuous Financial Assurance Using Multiple Types of Digital Information," *Contemporary Perspectives in Data Mining*, Information Age Publishing, 2012, pp. 103-122.
11. P. Hayes and S. Weinstein, "Constru-TIS: A System for Content-based Indexing of a Database of News Stories," *Proc. 2nd Conf. Innovative Applications of Artificial Intelligence*, Assoc. for the Advancement of Artificial Intelligence (AAAI), pp. 49-64.
12. S. Kim and E. Hovy, "Determining the Sentiment of Opinions," *Proc. COLING Conf.*, Assoc. for Computing Linguistics, 2004, article no. 1367; doi:10.3115/1220355.1220555.
13. E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good, the Bad, and the OMG!" *Proc. 5th Int'l AAAI Conf. Weblogs and Social Media*, AAAI, 2011, pp. 538-541.
14. X. Wu et al., "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, vol. 14, no. 1, 2008, pp. 1-37.
15. UK Future Internet Strategy Group, "Future Internet Report," tech. report, May 2011. https://connect.innovateuk.org/c/document_library/get_file?folderId=861750&name=DLFE-34705.pdf.
16. T. Kraska et al., "MLbase: A Distributed Machine Learning System," *Proc. 6th Biennial Conf. Innovative Data Systems Research*, 2013; www.cs.berkeley.edu/~ameet/mlbase.pdf.
17. A. Verma et al., *Scaling Simple and Compact Genetic Algorithms Using MapReduce*, Illinois Genetic Algorithms Laboratory (IlligAL) report no. 2009001, IlligAL, Univ. of Illinois at Urbana-Champaign, 2009.
18. H. Plattner and A. Zeier, *In-Memory Data Management*, Springer, 2011.

Daniel E. O'Leary is a professor at the Marshall School of Business at the University of Southern California. Contact him at oleary@usc.edu.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.