Big Data Analytics in Healthcare: Investigating the Diffusion of Innovation

by Diane Dolezel, EdD, RHIA, CHDA, and Alexander McLeod, PhD

Abstract

The shortage of data scientists has restricted the implementation of big data analytics in healthcare facilities. This survey study explores big data tool and technology usage, examines the gap between the supply and the demand for data scientists through Diffusion of Innovations theory, proposes engaging academics to accelerate knowledge diffusion, and recommends adoption of curriculum-building models. For this study, data were collected through a national survey of healthcare managers. Results provide practical data on big data tool and technology skills utilized in the workplace. This information is valuable for healthcare organizations, academics, and industry leaders who collaborate to implement the necessary infrastructure for content delivery and for experiential learning. It informs academics working to reengineer their curriculum to focus on big data analytics. The paper presents numerous resources that provide guidance for building knowledge. Future research directions are discussed.

Keywords: big data; analytics; curriculum; Diffusion of Innovations theory; informatics; business intelligence

Introduction

Healthcare decision making is a data-intensive process. Data are generated in Centers for Medicare and Medicaid Services (CMS) value-based purchasing reports and in the ubiquitous use of electronic health records, smart sensors, and mobile devices.[1] Data stores gather data from quality management measurements, the provision of connected healthcare, and population healthcare research. Analyzing healthcare data could provide financial benefits. A McKinsey Global institute study estimated that effective use of big data analytics could decrease national healthcare expenditures by approximately 8 percent annually.[2] However, the analysis of big data requires talented data scientists. By 2018, the projected demand for data scientists could exceed the supply by 50 to 60 percent, and 1.5 million more managers would be needed to apply the analysis results effectively.[3] In fact, the American Health Information Management Association (AHIMA) 2018 career map lists job titles for data analytics that include quality manager, mapping specialist, data integrity specialist, population health analyst, data analyst, and vice president of data management and analytics.[4] An AHIMA research study indicated that academics and industry representatives perceive a lack of talent available for data analytics jobs.[5]

Clearly, our nation has a shortage of data scientists.[6–8] Meeting this demand is problematic because advanced data analytics skills take years to develop, and working professionals with these skills have a high attrition rate.[9] Because of this, organizations should be proactive to prioritize building their big data science talent capacity, or they will risk losing their competitive edge.[10–12] Researchers have suggested that academics can help narrow this gap by expanding data science course offerings.[13, 14] This purpose of this paper is to survey current big data tool and technology usage. This paper examines the gap between the supply and the demand for data scientists through the Diffusion of Innovations (DOI) theory, proposes engaging academics to accelerate the diffusion knowledge of these skills, and recommends adoption of curriculum-building models to deliver that knowledge to future data scientists.

Literature Review

Human activities produce big data.[15] Big data can be characterized by the Four Vs, which are volume, variety, velocity, and veracity.[16] Volume is measured in terabytes (10^{12}) or greater. For example, most US companies have 100 terabytes of data stored, and among a world population of 7 billion, 6 billion people own cell phones.[17] The type of data produced varies; it can be structured, semistructured, or unstructured. Examples of unstructured data sources are the 420 million wearable health monitors or the 30 billion pieces of Facebook content.[18] Data velocities range from slow batch processing to fast data streaming, and an estimated 18.9 billon data network connections were made in 2016. Regarding veracity, one out of three business leaders do not trust the quality of their data.[19]

Healthcare data could include medical images, mobile device data, social media data, emails, cell phone data, audio data, and video data. Thus, special tools and technologies are needed to process these data sets. One popular suite of tools is Apache Hadoop, an open-source framework consisting of Hadoop Distributed File System, Pig, Hive, and HBase. Related to Hadoop is MapReduce, the Hadoop data processing model, which uses Apache Pig and Hive, and Microsoft's DryadLINQ.[20] When selecting one of these tools, analysts must take several considerations into account. For instance, Hadoop tools can process many types of structured and unstructured data, but these tools lack speed and are best suited to crunching retrospective (i.e., not real-time) data with batch processing.[21] To process large amounts of concurrent transactional data for online web and mobile users, the Apache Cassandra NoSQL distributed database, originally developed by Facebook, is a better choice.[22, 23] Table 1 shows tool names and usages and their web links.

Within the healthcare field, the use of big data analytics originated with scientists and government researchers. For instance, the world's largest biomedical research center, the US Department of Health and Human Services, houses the National Institutes of Health (NIH).[24] The NIH's Big Data to Knowledge (BD2K) initiative of 2012 established the Center of Excellence in Data Science to train data scientists to use computer science and statistics to derive value from clinical data.[25] The NIH Informatics for Integrating Biology and the Bedside (i2b2) program supports several National Center for Biomedical Computing organizations that are developing frameworks for translational research,[26–28] such as a cancer imaging archive and a cancer genome atlas.[29] Similarly, CMS is applying the Apache Hadoop/MapReduce software framework to create a Medicare and Medicaid reporting database[30] and is collaborating with Oak Ridge National Laboratory to evaluate data visualizations tools for Medicare claims data.[31]

Upon adoption by clinical data analysts, big data analytics quickly gained popularity. It was utilized to forecast readmissions due to heart failure[32] and to identify and manage high-cost patients.[33] MapReduce tools helped to speed up medical image analysis, and Kaiser Permanente's removal of Vioxx from the market, due to adverse effects, was driven by analysis of clinical and cost data.[34, 35] However, the adoption of big data analytics is hampered by the lack of skilled analysts. To close this knowledge gap, universities should add classes in data science, data analytics, and statistics to their curricula. Universities should form business alliances with IBM, Teradata, Microsoft, SAP, SAS, Tableau, RapidMiner, and other software companies. For instance, the SAP University Alliance provides business and predictive analytics along with a rich variety of classroom materials.[36] Likewise, SAS provides a university edition of its software, and RapidMiner offers free educational licenses.

Curriculum Models

The overall curricular design must be considered when courses are developed. The Association for Computing Machinery (ACM) and IEEE information technology (IT) curriculum model was the prototype for an information systems curriculum redesign that included the addition of a data analytics major and additional interdisciplinary courses.[37] This study presented lessons learned when a data analytics major was added to an undergraduate business curriculum. After confirming that this new major was consistent with their goal to provide relevant experience, the faculty employed a business roundtable to design the coursework. Roundtable members selected Python and R for assignments, set up internships, encouraged existing faculty to update their skills, and proposed hiring new faculty.[38] Within this model, the program mission, career goals, and program competencies were first considered, and then course designs were planned.[39] In a second study, the ACM MSIS 2006 model was employed to add a big data analytics course with a pedagogical (active learning) approach to a business school curriculum.[40] Specifically for healthcare, the Office of the National Coordinator for Health Information Technology (ONC) provides many curriculum resources.[41] The ONC resources cover population health, care coordination, interoperable health IT systems, value-based care, healthcare data analytics, and patient-centered care. The materials include lecture slides, assignments, recordings, and data sets. The ONC topics covered embrace natural language processing, database design, object-oriented programming, and clinical analytics.

Despite organizations' need to inform the building of capacity for advanced data analytics, most relevant research was focused on retrospective literature reviews, and few workplace surveys collected data from end users.[42–44] One study identified a need for health professionals with data governance skills who can collaborate with quality managers.[45] Another study presented big data analysis, informatics, and data governance as important skills for the future workplace and recommended earning a Certified Health Data Analyst credential, but did not specify technologies that were being used in the workplace.[46] These studies do not provide enough guidance for academics working to infuse big data analytics into their curriculum. This paper addresses those limitations by providing practical data on the usage of big data analytics and suggesting curriculum-building models, course content, and teaching resources. It provides guidance for healthcare organizations, academics, and industry leaders as they collaborate to increase the number of data scientists.[47–49] The recommendations for implementing these planned changes are derived from Rogers's Diffusion of Innovations theory.

Research Questions

The research questions are as follows:

1. What are the technologies utilized for big data analysis?
2. What are the skills needed for big data analysis?
3. What are the tools utilized for big data analysis?

Research Model

For this study, Rogers's Diffusion of Innovations theory was applied to the investigation of the diffusion of big data analytics in healthcare.[50] This theory has been used to guide the implementation of mobile devices by nursing students,[51] to improve the provisioning of diabetes care,[52] and to understand why organizations adopt information systems.[53] It was applied to a bachelor of science in nursing program to facilitate the development of competencies for conducting evidence-based studies in a senior research class.[54] The nursing program's implementation plan included partnership with community agencies.

According to DOI theory, an innovation is a new product that provides advantages to a group or market.[55] The entry of the product into the market initiates the diffusion of the innovation, which can occur through planned communication. For this study, *innovation* refers to the emergence of big data analytics skills in the healthcare workplace, and diffusion is the process for disseminating these skills. Currently, big data analytics implementation is slow because of a lack of human talent, which creates a knowledge gap. The knowledge gap needs to be narrowed with planned transmissions of knowledge related to big data analytics. Collaboration between healthcare organizations, academics, and industry leaders to facilitate curriculum building and to provide the necessary hands-on experience and infrastructure is recommended. A survey of key users was conducted to form a baseline to better understand the specific tools and technologies needed in the workplace.

Methodology

The aim of this descriptive study was to identify the skills, tools, and technologies used in the healthcare field to conduct big data analytics. The study site was a large university in the southern United States. Data were collected with a web survey developed by the researchers. Data were cleaned and analyzed using Microsoft Excel and uploaded to SPSS 24 for statistical analysis. Demographics and descriptive statistics were generated.

Participants

After institutional review board approval was granted, respondents' email addresses were obtained by the study researchers from the Definitive Healthcare database, a subscription healthcare database available at the study site. Participants were selected on the basis of having job titles that (potentially) reflected a broad knowledge of big data analytics usage, such as chief information officer, chief operations officer, chief executive officer, chief medical officer, chief nursing officer, director of IT, and health information management (HIM) director. Participants were recruited by email with a link to the online Qualtrics survey. The survey was available in the spring and summer of 2018. In total, 17,972 emails were sent, and 3,810 email addresses were returned as invalid. A total of 123 participants started the survey, and 112 participants completed the entire survey. All survey responses were analyzed. Regarding the response rate, 112 of 14,162 emails is a 0.79 percent response rate, which is very low.

Instrument

The survey instrument is included in the appendix. One question used a seven-point Likert scale (1 = never, 7 = very frequently) and invited respondents to rate their usage of big data infrastructure, tools, and technologies in their workplace. Other questions invited participants to state how frequently they used big data skills and to select the types of data storage, tools, and statistical and data visualization software used at work. Demographic questions asked about participants' years of healthcare experience, their type of workplace, their job position, and the country and state where they worked. Data were analyzed in SPSS 24 to determine the frequency of current workplace use of big data tools and technologies, the types of organizations, and the type of jobs of participants. Descriptive statistics were generated to describe respondents' demographics.

Results

A total of 112 respondents participated in the study. The respondents worked in 38 different states within the United States. The majority were from Texas (n = 25), California (n = 11), or New York (n = 7). The ranges of healthcare work experience were 1 to 5 years (1.8 percent), 6 to 10 years (13.4 percent), 16 to 20 years (18.8 percent), and more than 20 years (66.1 percent). The primary workplace organizations were hospitals (64.29 percent) and academic medical centers (13.39 percent). Most of the respondents were chief nursing officers (*n* = 36) or HIM directors (*n* = 28), or they held other job positions (*n* = 19; e.g., chief performance officer, director of clinical informatics). The work demographics are presented in Table 2.

Technologies Needed for Big Data Analytics

For this study, the overall level of use of big data analytics was defined as very frequently (daily), frequently (one or two times a week), occasionally (a few times a month), rarely (a few times every three months (i.e., every quarter), and never (not used at all). Approximately 63 percent reported either very frequent or frequent overall use of big data analytics at work. The most commonly used technologies (i.e., very frequent use) were statistical analysis (47.6 percent), data mining (39 percent), data visualization (34.1 percent), Structured Query Language (28.0 percent), and Java (26.8 percent).

Other technologies in use (by less than 10 percent of respondents) were artificial intelligence, C/C++ programming, cryptography, and parallel processing. A few respondents used stream processing to handle large packets of data. The most frequently used streaming technologies were SQLstream s-Server (*n* = 20) and IBM InfoSphere Streams (*n* = 6), followed by Apache Kafka (*n* = 2), Apache Spark Streaming (*n* = 2), and SAP HANA (*n* = 2). Table 3 presents technologies needed for big data analytics having more frequent use.

Skills Needed for Big Data Analytics

In addition to the skills tied to the technologies previously listed, Table 4 presents the database skills needed for big data analytics having more frequent responses. Respondents could select all or none of these databases. For relational databases, most respondents used Microsoft SQL Server (*n* = 48), Oracle (*n* = 23), or MySQL (*n* = 11). The most popular nonrelational databases were Apache Cassandra (*n* = 11) and Redis (*n* = 8). The nonrelational (NoSQL) responses in the “other” category (*n* = 4) encompassed Meditech, QlikView, InterSystems Cache, and SAP SQL Anywhere.

Tools for Big Data Analytics

Table 5 presents tools for big data analytics. The top three data science tools were IBM Apache Zookeeper (*n* = 25), Tableau (*n* = 13), and IBM Infosphere (*n* = 11). Additional tools used included these Apache tools: Pig programming, Hadoop HDFS file system, Hive query language, HBase database, and Mahout machine learning algorithm. Other data science tools in use were Dryad, JAQL query language, and Jaspersoft BI Suite. Statistical analysis was primarily conducted on SAS (*n* = 18), IBM SPSS (*n* = 13), Minitab (*n* = 7), MATLAB (*n* = 6), or R statistical software (*n* = 6). A few respondents used JMP (*n* = 3), Statistica (*n* = 5), or Stata (*n* = 2).

The most utilized data mining and analysis tools were SAS Enterprise Miner (*n* = 12), IBM SPSS Modeler (*n* = 9), Dryad Parallel Processing (*n* = 9), IBM Watson Analytics (*n* = 9), and R software (*n* = 5). Other data mining and analysis tools applied were Konstanz Information Miner (*n* = 2) and QlikView (*n* = 1). The top data visualization tools were Tableau (*n* = 16), Google Analytics (*n* = 15), Microsoft Power Business Intelligence (*n* = 14), and Oracle Visual Analyzer (*n* = 8). Additional data visualization tools employed were Highcharts (*n* = 1) and SAP Analytics Cloud (*n* = 5).

Limitations

As with all research, this study has several limitations. First, the response rate was 0.79 percent, which is very low. Because the recruitment emails contained survey links, the emails may have been blocked by security software at some facilities. Second, most respondents were from Texas, California, or

New York. For this study, the sample size was small, and the respondents were geographically focused. A larger sample with more states represented could have different results, as tool usage may vary significantly by geography. Third, 64 percent of the respondents worked at hospitals; thus, the results may not generalize to other healthcare settings.

Most of the incomplete responses occurred when participants stopped in the section on technologies for streaming data. Consequently, future researchers should consider reducing the number of questions as one way to increase the completion rate.

Discussion

It is challenging to hire people with data science talent. According to Glassdoor, data scientist jobs are the one of the six best jobs in America, with six-figure salaries and more than 4,500 job postings.[56] A LinkedIn report ranked data scientists' and big data developers' jobs among the top five emerging jobs in the United States for 2017.[57] The same report indicated that data scientist jobs postings increased 650 percent from 2012 to 2017, and big data developer job postings increased 550 percent.[58] Similarly, Indeed.com reported a 75 percent increase in data scientist job postings from 2015 to 2018.[59]

As universities struggle to reengineer their curricula to churn out more data scientists to meet the growing workplace needs, results from this study provide guidance on workplace needs. Results showed that data mining, data visualization, and SQL were the most frequently used technologies. Most respondents used Microsoft SQL Server, Oracle, or MySQL databases, while the most popular nonrelational database was Apache Cassandra. Respondents indicated that statistical analysis was primarily conducted on SAS or IBM SPSS software. The preferred data mining tools were SAS Enterprise Miner, IBM SPSS Modeler, and Dryad Parallel Processing. Moreover, Tableau, Google Analytics, and Microsoft Power Business Intelligence were the top data visualization tools used.

Universities should use the results from this study to inform their choice of curricular content. As a first step, academic administrators should create a project plan to better manage the changes. The plan should address how to locate faculty, design courses, update the curriculum, and procure course materials. Implementing this plan presents administrative and financial challenges in many areas. First, where will administrators find the faculty with the computer science, data analytics, and business intelligence skills needed to these courses? Obviously, they could train existing faculty or hire new faculty. Unfortunately, hiring from outside the university places faculty recruiters in competition with industry recruiters, who can offer the applicants higher corporate salaries and bigger benefits packages.[60] Correspondingly, training existing faculty is expensive, and faculty learning must be reinforced with hands-on practice on complex, expensive software systems. This requirement creates stress for the faculty who must maintain their teaching load while they learn new systems and develop new analytics courses.

Second, making curricular changes requires standardization of course syllabi if content is to be managed across several courses. However, a study of more than 30 big data syllabi identified diverse subjects and content among the courses.[61] Fortunately, guidance for standardization is available in the AHIMA Draft 2018 Graduate Curriculum Guidance,[62] which presents domains of practice, Bloom's taxonomy levels, and suggested learning resources. For instance, for statistical analysis AHIMA recommends using SAS, Python, SPSS, and R, and for data visualization AHIMA suggests Tableau, QlikView, and GIS mapping tools.

Third, problem-based assignments focused on current data science scenarios must be created.[63, 64] Data streams, like Twitter or stock market data, could be analyzed to create business intelligence.[65, 66] According to researchers, SAS Enterprise Miner is useful for teaching data analytics, and Tableau provides big data visualization.[67, 68] Additionally, the Apache software project provides free tools, examples, and documentation for many of the Apache products. Furthermore, the Health Information

Management Systems Society (HIMSS) Data and Analytics Task Force provides complimentary introductory materials on its Big Data 101 web pages.[69] In summary, many resources are available to guide project planning for curriculum reengineering. The most challenging issues, however, are finding qualified faculty and creating a change management plan while working with a limited budget.

Future Studies

Future research should examine more healthcare industries to determine which big data technologies are in use, and this information should guide course content and class development. This study focused on the current usage of big data analytic technical skills, such as database skills and the skills needed to use big data analytics in healthcare. In the future, researchers could expand the study to consider critical thinking, analytic thinking, data visualization, and communication skills, which would be measured with different instrument scales. Studies should be conducted to identify lessons learned at universities where new data science courses or curricula have been implemented. Faculty should be surveyed regarding their perceptions of their readiness to take on the challenges of teaching these new classes. Moreover, businesses and professional organizations are a vital part of solving the shortage of big data skills. Because the cost of many data science tools is prohibitive, the industry should produce free academic software licenses and provide free online training materials to reduce the financial burden for universities. Academics should collaborate with firms that could provide internships, scholarships, and hands-on data camps for university students. One suggestion is to create internship classes that are taken for college credit. Future researchers should investigate industry, business, and professional organizations’ attitudes and willingness to partner with academics. For example, from the industries’ perspective, what benefits would industry leaders like to see emerge from these collaborations?

Conclusion

This study provides data collected from executives on the current usage of big data tools and technologies in the workplace. This information is valuable because mountains of data must be analyzed to provide the knowledge for improving healthcare decision making, and the data scientists needed to perform the analysis are in high demand. This constant need for improvement drives the diffusion of big data science skills to the healthcare market. Currently, there is a knowledge gap in the pool of healthcare job applicants, and this gap is not expected to narrow without aggressive, planned dissemination of big data analytic skills to future job seekers.

Diane Dolezel, EdD, RHIA, CHDA, is an assistant professor in the Department of Health Information Management at Texas State University in San Marcos, TX.

Alexander McLeod, PhD, is an associate professor in the Department of Health Information Management at Texas State University in San Marcos, TX.

Notes

1. Fang, Ruogu, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. “Computational Health Informatics in the Big Data Age: A Survey.” *ACM Computing Surveys* 49, no. 1 (2016): 1–36. doi:10.1145/2932707.
2. Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Richard Dobbs, Charles Charles Roxburgh, and Angela Byers. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute, 2011.

3, Ibid.

4. AHIMA. “AHIMA Career Map.” 2018. Available at https://my.ahima.org/careermap.
5. Jackson, K., C. L. Lower, and W. J. Rudman. “The Crossroads between Workforce and Education.” *Perspectives in Health Information Management* 13 (Spring 2016): 1–11.
6. Alharthi, Abdulkhaliq, Vlad Krotov, and Michael Bowman. “Addressing Barriers to Big Data.” *Business Horizons* 60 (2017): 285–92.
7. Oachs, Pamela K., and Amy L. Watters. *Health Information Management: Concepts, Principles and Practice*. 5th ed. Chicago: AHIMA, 2016.
8. Schiller, Shu, Michael Goul, Lakshmi Iyer, Ramesh Sharda, and David Schrader. “Build Your Dream (Not Just Big) Analytics Program.” Presented at the Twentieth Americas Conference on Information Systems, Savannah, Georgia, 2014.
9. Fernandes, L., M. O’Connor, and V. Weaver. “Big Data, Bigger Outcomes: Healthcare Is Embracing the Big Data Movement, Hoping to Revolutionize HIM by Distilling Vast Collection of Data for Specific Analysis.” *Journal of AHIMA* 83, no. 10 (2012): 38–43.
10. Taylor, Erin, Janice Genevro, Peikes Peikes, Kristin Geonnotti, Winnie Wang, and David Meyers. *Building Quality Improvement Capacity in Primary Care: Supports and Resources* (AHRQ Publication No. 13-0044-2-EF). Rockville, MD: Agency for Healthcare Research and Quality, April 2013.
11. World Health Organization. “Tobacco Free Initiative (TFI): Capacity Building and Initiatives.” 2006. Available at http://www.who.int/tobacco/control/capacity_building/background/en/.
12. UN Global Working Group on Big Data for Official Statistics. Task Team on Skills, Training and Capacity Building. *Analysis of Big Data Survey 2015 on Skills, Training and Capacity Building*. https://unstats.un.org/unsd/trade/events/2015/abudhabi/presentations/day2/02/Analysis_of_Big_Data_Survey_2015_on_Skills_Training_and_Capacity_Building%20v1%200.pdf
13. Asamoah, D., R. Sharda, A. Zadeh, and P. Kalgotra. “Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course.” *Decision Sciences Journal of Innovative Education* 15, no. 2 (2017): 161–90.
14. Cegielski, C., and L. Jones-Farmer. “Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-method Study.” *Decision Sciences Journal of Innovative Education* 14, no. 1 (2016): 91–118.
15. Alharthi, Abdulkhaliq, Vlad Krotov, and Michael Bowman. “Addressing Barriers to Big Data.”
16. Feldman, B., E. Martin, and T. Skotnes. “Big Data in Healthcare: Hype and Hope.” Dr. Bonnie 360°, October 2012.
17. IBM. “The Four V’s of Big Data.” 2018. Available at https://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg.
18. Ibid.
19. Ibid.

20. Watson, H. “Tutorial: Big Data Analytics: Concepts, Technologies, and Applications.” *Communications of the Association for Information Systems* 34, no. 65 (2014): 1248–68.
21. Sagiroglu, Seref, and Duygu Sinanc. “Big Data: A Review.” Presented at the IEEE 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, California, May 20–24, 2013.
22. Amster, Ari. “Cassandra vs. Hadoop Use Cases: A Comparative Look.” Qubole, January 28, 2016. Available at https://www.qubole.com/blog/cassandra-vs-hadoop/.
23. Chen, Min, Shiwen Mao, and Yunhao Liu. “Big Data: A Survey.” *Mobile Networks and Applications* 19, no. 2 (2014): 171–209. doi:10.1007/s11036-013-0489-0.
24. National Institutes of Health. “About NIH.” 2018. Available at https://www.nih.gov/about-nih.
25. Margolis, Ronald, Leslie Derr, Michelle Dunn, Michael Huerta, Jennie Larkin, Jerry Sheehan, Mark Guyer, and Eric D. Green. “The National Institutes of Health’s Big Data to Knowledge (BD2K) Initiative: Capitalizing on Biomedical Big Data.” *Journal of the American Medical Informatics Association* 21, no. 6 (2014): 957–58. doi:10.1136/amiajnl-2014-002974.
26. Partners Healthcare. “Informatics for Integratging Biology and the Bedside (i2b2).” 2018. Available at https://www.i2b2.org/about/index.html.
27. Murphy, S., and A. Wilcox. “Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2).” *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 2, no. 2 (2014). doi:10.13063/2327-9214.1074.
28. Murphy, Shawn, Alyssa Goodson, Michael Mendis, Marykate Murphy, Lori Phillips, Yanbing Wang, and Christopher Herrick. “Bringing Healthcare Analytics to Where Big Data Resides Using a Distributed Query System.” *International Journal on Computer Science & Information Systems* 11, no. 2 (2016): 237–40.
29. The White House. “Fact Sheet: Big Data across the Federal Government.” March 29, 2012. Available at https://obamawhitehouse.archives.gov/the-press-office/2015/12/04/fact-sheet-big-data-across-federal-government.
30. Ibid.
31. Ibid.
32. Bradley, Elizabeth H., Leslie Curry, Leora I. Horwitz, Heather Sipsma, Yongfei Wang, Mary Norine Walsh, Don Goldmann, Neal White, Ileana L. Piña, and Harlan M. Krumholz. “Hospital Strategies Associated with 30-Day Readmission Rates for Patients with Heart Failure.” *Circulation: Cardiovascular Quality and Outcomes* 6, no. 4 (2013): 444–50. doi:10.1161/CIRCOUTCOMES.111.000101.
33. Bates, David W., Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. “Big Data in Heath Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients.” *Health Affairs* 33, no. 7 (2014): 1123–31. doi:10.1377/hlthaff.2014.0041.
34. Raghupathi, Wullianallur, and Viju Raghupathi. “Big Data Analytics in Healthcare: Promise and Potential.” *Health Information Science and Systems* 2, no. 3 (2014): 1–10. doi:10.1186/2047-2501-2-3.
35. Belle, Ashwin, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian. “Big Data Analytics in Healthcare.” *BioMed Research International* (2015): 1–16. doi:10.1155/2015/370194.
36. Week, John, Alexander McLeod, Mark G. Simkin, and Bret Simmons. “Toward a User Commitment Continuum.” Presented at the Americas Conference on Information Systems, Lima, Peru, 2010.
37. Wymbs, C. “Managing the Innovation Process: Infusing Data Analytics into the Undergraduate Business Curriculum (Lessons Learned and Next Steps).” *Journal of Information Systems Education* 27, no. 1 (2016): 61–74.

38. Ibid.
39. Ibid.
40. Asamoah, D., R. Sharda, A. Zadeh, and P. Kalgotra. “Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course.”
41. Office of the National Coordinator for Health Information Technology. “Health IT Curriculum Resources for Educators.” 2018. Available at https://www.healthit.gov/topic/health-it-resources/health-it-curriculum-resources-educators.
42. Alonso, Susel Gongora, Isabel de la Torre Diez, Joel J. P. C. Rodrigues, Sofiane Hamrioui, Miguel Lopez-Coronado. “A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector.” *Journal of Medical Systems*.” 41 (11):183. doi: 10.1007/s10916-017-0832-2.
43. Archenaa, J., E. A. Mary Anita. “A Survey of Big Data Analytics in Healthcare and Government.” *Procedia Computer Science*. 50:408-413. doi: 10.1016/j.procs.2015.04.021.

44. Belle, Ashwin, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian. “Big Data Analytics in Healthcare.”
45. Desai, Anna. "Scanning the HIM Environment: AHIMA’s 2015 Report Offers Insight on Emerging Industry Trends and Challenges" *Journal of AHIMA* 86, no.5 (May 2015): 38-43. http://bok.ahima.org/doc?oid=107636#.XQLBqFxKg2w
46. AHIMA. “HIM Reimagined.” http://www.ahima.org/about/him-reimagined.
47. Pence, Harry E. “What Is Big Data and Why Is It Important?” *Journal of Educational Technology Systems* 43, no. 2 (2014): 159–71. doi:10.2190/ET.43.2.d.
48. Taylor, Erin, Janice Genevro, Peikes Peikes, Kristin Geonnotti, Winnie Wang, and David Meyers. *Building Quality Improvement Capacity in Primary Care: Supports and Resources*.
49. World Health Organization. “Tobacco Free Initiative (TFI): Capacity Building and Initiatives.”
50. Orr, C. “Diffusion of Innovations, by Everett Rogers (1995).” 2003. Available at https://teamlead.duke-nus.edu.sg/vapfiles_ocs/2011/edu/Diffusion_of_Innovations_by_Everett_Rogers_1995.pdf.
51. Doyle, Glynda J., Bernie Garrett, and Leanne M. Currie. “Integrating Mobile Devices into Nursing Curricula: Opportunities for Implementation Using Rogers’ Diffusion of Innovation Model.” *Nurse Education Today* 34, no. 5 (2014): 775–82. doi:http://dx.doi.org/10.1016/j.nedt.2013.10.021.
52. Lein, Angela, and Yi-Der Jiang. “Integration of Diffusion of Innovation Theory into Diabetes Care.” *Journal of Diabetes Investigation* 8, no. 3 (2017): 259–60. doi:10.1111/jdi.12568.
53. Mustonen-Ollila, Erja, and Kalle Lyytinen. “Why Organizations Adopt Information System Process Innovations: A Longitudinal Study Using Diffusion of Innovation Theory.” *Information Systems Journal* 13, no. 3 (2003): 275–97.
54. Schmidt, Nola, and Janet Brown. “Use of the Innovation–Decision Process Teaching Strategy to Promote Evidence-based Practice.” *Journal of Professional Nursing* 23, no. 3 (2007): 150–56. doi:10.1016/j.profnurs.2007.01.009.
55. Orr, C. “Diffusion of Innovations, by Everett Rogers (1995).”
56. Columbus, Louis. 2018. “Data Scientist Is the Best Job In America According Glassdoor’s 2018 Rankings.” *Forbes*, January 29, 2018. Available at

https://www.forbes.com/sites/louiscolumbus/2018/01/29/data-scientist-is-the-best-job-in-america-according-glassdoors-2018-rankings/#178c29155357.

57. LinkedIn. “LinkedIn’s 2017 U.S. Emerging Jobs Report.” 2017. Available at https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report.
58. Ibid.
59. Sasso, Michael. “This Is America’s Hottest Job.” Bloomberg, May 18, 2018. Available at https://www.bloomberg.com/news/articles/2018-05-18/-sexiest-job-ignites-talent-wars-as-demand-for-data-geeks-soars.
60. Chiang, R., P. Goes, and E. Stohr. “Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline.” *ACM Transactions on Management Information Systems* 3, no. 3 (2012). doi:10.1145/2361256.2361257.
61. Friedman, A. “Measuring the Promise of Big Data Syllabi.” *Technology, Pedagogy and Education* 27, no. 2 (2018): 135–48. doi:10.1080/1475939X.2017.1408490.
62. AHIMA Council for Excellence in Education. “Draft 2018 Graduate Curriculum Guidance.” Available at https://www.ahima.org/~/media/AHIMA/Files/2018/Academic_Curricula_Resources/John%20updated%20version/Draft_2018_Graduate_Curriculum_Guidance_06-25.ashx?la=en.
63. Iskandaryan, Mike. “Developing the Right Workforce at the Right Time for Big Data Analytics.” *Journal of AHIMA*, November 17, 2015. Available at https://journal.ahima.org/2015/11/17/developing-the-right-workforce-at-the-right-time-for-big-data-analytics/.
64. Chiang, R., P. Goes, and E. Stohr. “Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline.”
65. Sigman, Betsy Page, William Garr, Robert Pongsajapan, Marie Selvanadin, Mindy McWilliams, and Kristin Bolling. “Visualization of Twitter Data in the Classroom.” *Decision Sciences Journal of Innovative Education* 14, no. 4 (2016): 362–81.
66. Splunk. 2018. Available at https://www.splunk.com/.
67. Zadeh, A., S. Schiller, K. Duffy, and J. Williams. “Big Data and the Commoditization of Analytics: Engaging First-Year Business Students with Analytics.” *e-Journal of Business Education & Scholarship of Teaching* 12, no. 1 (2018): 120–37.
68. Belle, Ashwin, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian. “Big Data Analytics in Healthcare.”
69. Health Information Management Systems Society (HIMSS) Data and Analytics Task Force. “Big Data 101: Moving from Health IT Data to Action (Series).” 2014. Available at https://www.himss.org/big-data-101-moving-health-it-data-action-series.

Table 1

Tools Names and Usages

Tools	**Usage**	**URL**
Hadoop Distributed File System (HDFS)	Distributed file processing system for large data sets	https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
Apache Pig	MapReduce data scripting language for extracting, transforming, and loading data into data stores	https://pig.apache.org/
Apache Hive	SQL query language; used with HDFS and HBase data warehouses	https://hive.apache.org/
Apache HBase	Nonrelational (NoSQL) database storage used on HDFS; modeled after Google's Big Table	https://hbase.apache.org/
MapReduce	Hadoop programming model distributes task across servers	https://hadoop.apache.org/docs/stable/index.html
IBM Apache Hadoop Zookeeper	Synchronizes data delivery to servers	https://www.ibm.com/analytics/hadoop/zookeeper
Apache Cassandra	Nonrelational (NoSQL) database storage	http://cassandra.apache.org/

Table 2

Work Demographics (*n* = 112)

Characteristic	**Number**	**Percentage**
Work organization		
Academic educational institute	4	3.57
Academic medical center	15	13.39
Community health organization or clinic	1	0.89
Healthcare provider	5	4.46
Healthcare system corporate office	6	5.36
Hospital	72	64.29
Long-term/post-acute care	3	2.70
Other	5	4.50
Job position		
Chief executive officer	7	6.25
Chief information officer	3	2.68
Chief operations officer	6	5.36
Chief medical officer	6	5.36
Chief nursing officer	36	32.14
Director of IT	7	6.25
HIM director	28	25.00
Other	19	16.96

Table 3

Technologies Needed for Big Data Analytics

Technology	*Number	Percentage (%)
Artificial intelligence	7	8.5
C/C++ programming	5	6.1
Cryptography	8	9.8
Data mining	32	39.0
Data visualization	28	34.1
Java	22	26.8
Parallel processing	6	7.3
Structured Query Language	23	28.0
Statistical analysis	39	47.6

* Number = frequency of those responding to this question, participants may respond to several choices or no choice for each question, thus percentages may not add to 100%

Table 4

Database Skills Needed for Big Data Analytics

Skill	**Number**	***Percentage (%)**
Relational		
IBM DB2	8	7.1
Microsoft SQL Server	48	42.9
MySQL	11	9.8
Oracle	23	20.5
PostgreSQL	3	2.7
SAP Hanna	5	4.5
Teradata	3	2.7
Other	8	10.7
Nonrelational (NoSQL)		
Apache Cassandra	11	9.8
Couchbase	5	4.5
Apache Hadoop/MapReduce	2	1.8
Apache CouchDB (document database)	4	3.6
Apache HBase	4	3.6
MongoDB (document database)	6	5.4
Redis	8	7.1
Other	4	25.9

* Percentage = percentage of those responding to this question, participants may respond to several choices or no choice for each question, thus percentages may not add to 100%

Table 5

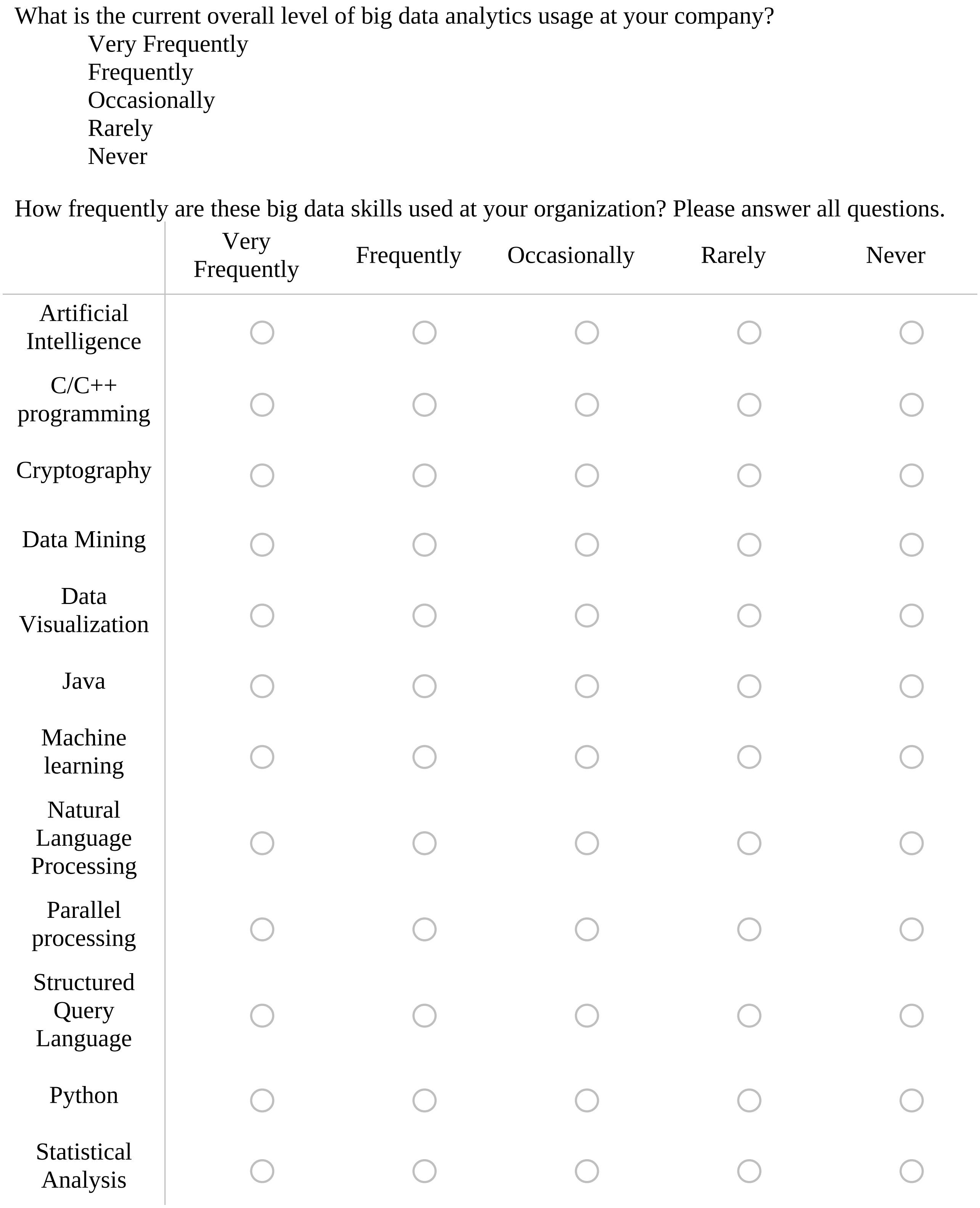
Tools for Big Data Analytics

Tool	Number	*Percentage (%)
Data science		
Apache Pig	1	0.9
Apache Hadoop HDFS	3	2.7
Apache Hive	3	2.7
Apache HBase	2	1.8
Dryad	1	0.9
JAQL	5	4.5
Jaspersoft BI Suite	1	0.9
IBM Infosphere	11	9.8
Apache Mahout	2	1.8
Tableau Desktop and Server	13	11.6
IBM Apache Zookeeper	25	2.7
Statistical analysis		
R	6	5.4
JMP	3	2.7
Minitab	7	6.3
MATLAB	6	5.4
SAS	18	16.1
IBM SPSS	13	11.6
Stata	2	1.8
Statistica	5	4.5
Data mining and analysis		
SAS Enterprise Miner	12	10.7
IBM SPSS Modeler	9	8.0
Dryad Parallel Processing	9	8.0
IBM Watson Analytics	9	7.1
R software	5	5.4
Konstanz Information Miner	2	0.9
QlikView	1	0.9
Data visualization		
Tableau	16	14.3
Google Analytics	15	13.4
Microsoft Power Business Intelligence	14	12.5
Oracle Visual Analyzer	8	7.1
SAP Analytics Cloud	5	4.5
Highcharts	1	0.9

* Percentage = percentage of those responding to this question, participants may respond to several choices or no choice for each question, thus percentages may not add to 100%

Figure 1

Survey Instrument



What is the current overall level of big data analytics usage at your company?

Very Frequently
Frequently
Occasionally
Rarely
Never

How frequently are these big data skills used at your organization? Please answer all questions.

	Very Frequently	Frequently	Occasionally	Rarely	Never
Artificial Intelligence	○	○	○	○	○
C/C++ programming	○	○	○	○	○
Cryptography	○	○	○	○	○
Data Mining	○	○	○	○	○
Data Visualization	○	○	○	○	○
Java	○	○	○	○	○
Machine learning	○	○	○	○	○
Natural Language Processing	○	○	○	○	○
Parallel processing	○	○	○	○	○
Structured Query Language	○	○	○	○	○
Python	○	○	○	○	○
Statistical Analysis	○	○	○	○	○

Indicate which relational databases are in use at your organization. Select all that apply.

IBM DB2
Microsoft SQL Server
MySQL
Oracle database
PostgreSQL
SAP Hanna
Teradata
Other please specify

Indicate which NOSQL non-relational databases are in use at your organization. Select all that apply.

Apache Cassandra
Couchbase
ArangoDB
Apache Hadoop MapReduce
Apache CouchDB – document db
Apache HBase
MongoDB – document db
Redis
Other please specify

Indicate which big data tools are in use at your organization. Select all that apply.

Apache Pig programming
Apache Hadoop HDFS distributed file system
Apache Hive Query Language
Apache HBase column-oriented database
Apache Avro data serialization
Dryad
Hortonworks Data Platform
JAQL query language
Jaspersoft BI Suite
IBM InfoSphere
Karmasphere studio and analyst (Hadoop)
Apache Mahout machine learning algorithms
Pentaho business analytics
Skytree Server
Tableau Desktop and Server
Talend Open Studio
IBM Apache Zookeeper
Other please specify

Indicate which stream processing tools are in use at your organization. Select all that apply.

Apache Kafka
Apache Spark Streaming
IBM InfoSphere Streams
SAP HANA
SQLstream s-Server
StreamCloud

Apache Storm
Other please specify

Indicate which data analysis tools are in use at your organization. Select all that apply.

Microsoft Dryad – parallel computing
IBM SPSS Modeler
IBM Watson Analytics
Konstanz Information Miner (KNMINE)
Apache MapReduce – parallel computing
R
RapidMiner
SAS Enterprise Miner
Waikato Environment for Knowledge Analysis (Weka)/Pentaho
Other please specify

Indicate which statistical analysis tools are in use at your organization. Select all that apply.

R
SAS JMP
Minitab
Matlab
SAS
SPSS
Stata
Statsoft Statistica
Other please specify

Indicate which data visualization tools are in use at your organization. Select all that apply.

FusionCharts
Google Analytics
Highcharts
IBM Watson Analytics
Microsoft Power BI
Oracle Visual Analyzer
Qlikview
SAP Analytics Cloud
Tableau
Other please specify