A Comparison of Open Source Tools for Data Science

Hayden Wimmer
Department of Information Technology
Georgia Southern University
Statesboro, GA 30260, USA

Loreen M. Powell
Department of Information and Technology Management
Bloomsburg University
Bloomsburg, PA 17815, USA

Abstract

The next decade of competitive advantage revolves around the ability to make predictions and discover patterns in data. Data science is at the center of this revolution. Data science has been termed the sexiest job of the 21st century. Data science combines data mining, machine learning, and statistical methodologies to extract knowledge and leverage predictions from data. Given the need for data science in organizations, many small or medium organizations are not adequately funded to acquire expensive data science tools. Open source tools may provide the solution to this issue. While studies comparing open source tools for data mining or business intelligence exist, an update on the current state of the art is necessary. This work explores and compares common open source data science tools. Implications include an overview of the state of the art and knowledge for practitioners and academics to select an open source data science tool that suits the requirements of specific data science projects.

Keywords: Data Science Tools, Open Source, Business Intelligence, Predictive Analytics, Data Mining.

1. INTRODUCTION

Data science is an emerging field which intersects data mining, machine learning, predictive analytics, statistics, and business intelligence. The data scientist has been coined the "sexiest job of the 21st century" (Davenport & Patil, 2012). The data science field is so new that the U.S. bureau of labor and statistics does not yet list it as a profession; yet, CNN's Money lists the data scientist as #32 on their best jobs in America list with a median salary of $124,000 (Money, 2015). Fortune lists the data scientist as the hot tech gig of 2022 (Hempel, 2012). The volume of data has exploded (Brown, Chui, & Manyika, 2011); however, a shortfall of skilled data scientists remain (Lake & Drake, 2014) which helps justify the high median salary.

Data science is an expensive endeavor. One such example is JMP by SAS. SAS is a primary provider of data science tools. JMP is one of the more modestly priced tools from SAS with the price for JMP Pro listed as $14,900. Based on the high price point of related software, data science efforts are out of reach for small and medium business as well as many local and regional healthcare organizations where efforts bring competitive advantages, improved performance, and cost reductions. The shortfall of data scientists detail the need for higher education to provide training programs; nonetheless, the high cost of data science software is a barrier to classroom adoption. Based on the aforementioned shortfalls, open source based solutions may provide respite. This paper seeks to provide an overview of data science and the tools required to meet the needs of organizations, albeit higher education,

business, or clinical settings as well as provide insight to the capabilities of common open source data science tools.

2. BACKGROUND

This section begins with introducing the term Data Science and prominent aspects of data science, namely data mining, machine learning, predictive analytics, and business intelligence. Next, open source software is introduced and skills of the data scientist are framed based on an industry certification.

Data Science

Data science is a revived term for discovering knowledge from data (Dhar, 2013); yet, the term has come to encompass more than merely traditional data mining. A universally accepted definition does not yet exist. It is generally agreed that a data scientist combines skills from multiple disciplines such as computer science, mathematics, and even art (Loukides, 2010). The data scientist must combine techniques from multiple disciplines which include, but are not limited to, data mining, machine learning, predictive analytics, and business intelligence. Regardless of the tool, extracting knowledge from data, particularly for predictive purposes, is at the heart of the data science field. The data scientist collaborates with domain experts to extract and transform data as well as provide guidance in the analysis of the results of data science activities.

Data Mining

Data mining (DM), commonly referred to as knowledge discovery in databases (KDD) and an integral aspect of data science, is the process of extracting patterns and knowledge from data such as pattern discovery and extraction (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The Cross Industry Standard Process for Data Mining (CRISP-DM), one of the leading data mining methodologies, divides the data mining process into 6 steps (Chapman et al., 2000; Wirth & Hipp, 2000). First, a business understanding of the project is developed followed by an analysis and understanding of the current data resources. Third, data pre-processing is performed to format the data suitable to data mining applications and algorithms. Next, models based on the data are generated. Model generation may be automatic via machine learning, semi-automatic, or manual. The models are then evaluated for performance and accuracy. The final step is deployment of the model(s) to solve the mission identified in the first step when developing a business understanding of the project.

Machine Learning

Machine learning (ML), employed as a method in data science, is the process of programming computers to learn from past experiences (Mitchell, 1997). ML seeks to develop algorithms that learn from data directly with little or no human intervention. ML algorithms perform a variety of tasks such as prediction, classification, or decision making. ML stems from artificial intelligence research and has become a critical aspect of data science. Machine learning begins with input as a training data set. In this phase, the ML algorithm employs the training dataset to learn from the data and form patterns. The learning phase outputs a model that is used by the testing phase. The testing phase employs another dataset, applies the model from the training phase, and results are presented for analysis. The performance on the test dataset demonstrates the models ability to perform its task against data. Machine learning extends beyond a statically coded set of statements into statements that are dynamically generated based on the input data.

Predictive Analytics

Predictive analytics, a cornerstone of data science efforts, is the process of employing empirical methods to generate data predictions (Shmueli & Koppius, 2010). Predictive analytics frequently involve statistical methods, such as regression analysis, to make predictions based on data. Predictive analytics has a wide range of applications from marketing, finance, and clinical applications. A common marketing application is customer churn analysis which seeks to determine which customers may switch to a competing provider and make special offers in order to retain these high-risk of churn customers. Finance applications include predicting customer profitability or risk management as employed by the insurance industry. Clinical applications include clinical decision support, determining which patients are at risk for hospital readmission, or medication interaction modeling.

Business Intelligence

Business intelligence, or BI, combines analytical tools to present complex information to decision makers (Negash, 2004). BI is part of data science efforts frequently as output of such efforts. Business intelligence tools integrate data from an organization for presentation. One such example is providing executive

management with dashboards which provide a view of the organization's operations. Decision makers employ this information to make strategic or operational decisions that impact the objectives of the organization. One goal of business intelligence is presentation of data and information in a format that can be easily understood by decision makers. Business intelligence includes key performance indicators (KPI) from the organization, competitors, and the marketplace. BI efforts have capabilities such as online application/ analytical processing (OLAP), data warehousing, reporting, and analytics.

Open Source Software

Open source has, in the minds of many, come to be synonymous with free software (Walters, 2007). Open source software is software where the development and the source code are made publically available and designed to deny anybody the right to exploit the software (Laurent, 2004). Open source generally refers to the source code of the application being freely and openly available for modifications. Two such examples of open source licenses are the GPL, or general public license (GNU.org, 2015a), and GNU(GNU.org, 2015b). Anyone can develop extensions or customizations of open source software; though, charging a fee for such activities is typically prohibited by a public license agreement whereby any modifications to the source code automatically become public domain. Communities emerge around software with developers worldwide extending open source software.

Techniques of the Data Scientist

Data science employs a myriad of techniques. Industry has long offered certifications in topics such as business intelligence; however, data science certifications are relatively new. One leading certification is EMC's Data Science Associate (EMC, 2015). This certification follows 6 key learning areas: 1) Data Analytics and the Data Scientist Role, 2) Data Analytics Lifecycle, 3) Initial Analysis of Data, 4) Theory and Methods, 5) Technology and Tools, and 6) Operationalization and Visualization. The theory and methods section focuses on specific methods employed by the data scientist while technology and tools relates to big data technologies such as Hadoop. Methods identified include: K-means clustering, Association rules, linear regression, Logistic Regression, Naïve Bayesian classifiers, Decision trees, Time Series Analysis, and Text Analytics. A brief description of each, as adapted from by EMC is:

- K-means clustering – an unsupervised method learning method which groups data instances. K-means is the most popular algorithm for clustering where the data is grouped into K groups.
- Association rule mining - an unsupervised method to find rules in the data. ARM is commonly used as market basket analysis to determine which products are commonly purchased together.
- Linear regression – used to determine linear functions between variables.
- Logistic Regression – used to determine the probability an event will occur as a function of other variables.
- Naïve Bayesian classifiers – used for classification and returns a score between 0 and 1 of the probability of class membership assuming independence of variables.
- Decision tree – classification and prediction method to return probability of class membership and output as a flowchart or set of rules for determining class membership.
- Time Series Analysis – accounts for the internal structure of time series measurements to determine trends, seasonality, cycles, or irregular events.
- Text Analytics – the processing and representation of data in text form for analyzing and model construction.
- Big Data Processing – the processing of large volume datasets using techniques such as distributed computing, distributed file systems, clustering, and map reduce (i.e. Hadoop)

The aforementioned data science techniques will be the basis for comparison of open source tools.

3. OPEN SOURCE TOOLS FOR THE DATA SCIENTIST

This section covers current reviews on open source data science tools. Following the review, open source tools are compared based on the industry data science certification, EMC's Data Science Associate.

Current Reviews

In 2005 a special workshop on open source data mining was conducted by SIGKDD (Goethals, Nijssen, & Zaki, 2005) where different topics within data mining were presented with frequent item set mining the most represented. While algorithms and methods were discussed no tools for practitioners were reviewed. Open source tools were reviewed by Chen, Ye, Williams, and

Xu (2007) where 12 prominent data mining tools and their respective functionalities were detailed. ADAM (Rushing et al., 2005), Alpha Miner (Institute, 2005), ESOM (Ultsch & Mörchen, 2005), Gnome Data Miner (Togaware, 2006), KNIME (Berthold et al., 2008), Mining Mart (Zücker, Kietz, & Vaduva, 2001), MLC++ (Kohavi, John, Long, Manley, & Pfleger, 1994), Orange (Demšar et al., 2013), Rattle (Williams, 2009), Tanagra (R. Rakotomalala, 2008) ,Weka (Hall et al., 2009; Holmes, Donkin, & Witten, 1994), and Yale (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006) were compared. The open source tools were compared on general characteristics (i.e. language), data source capabilities, functionality, and usability.

Advancing to 2008, Zupan and Demsar (2008) reviewed open source data mining tools including R (Ihaka & Gentleman, 1996), Tanagra (R. Rakotomalala, 2008), Weka (Hall et al., 2009; Holmes et al., 1994), YALE (Mierswa et al., 2006), Orange (Demšar et al., 2013), KNIME (Berthold et al., 2008), and GGobi (Swayne, Lang, Buja, & Cook, 2003). Saravanan, Pushpalatha, and Ranjithkumar (2014) reviewed Clementine (Khabaza & Shearer, 1995), Rapid Miner (Mierswa et al., 2006), R (Ihaka & Gentleman, 1996), and SAS Enterprise Miner (Cerrito, 2006). While the aforementioned reviews provide insight into the tools available and a look at functionality, the dimensions required for a prominent industry data science certification were not fully represented. This work extends the current literature by providing a feature base comparison of open source data science toolkits from a practitioner perspective based from a prominent industry certification, the EMC Data Science Associate.

Tool Selection

Tools that intersect multiple reviews are Tanagra, Orange, KNIME, Weka, and Yale (now Rapid Miner). In addition to academic literature, from a practitioner standpoint, 2 websites that mention top open source data mining are included. The first from The New Stack discusses 6 open source data mining toolkits which include Orange, Weka, Rapid Miner, JHepWork, and KNIME (Goopta, 2014). The second internet based source, from Tech Source, discusses 5 open source tools which include Rapid Miner, Weka, Orange, R, KNIME, and NTLK (Auza, 2010). Tools that were included in 2 or more of the academic or practitioner sources are included; Orange, Tanagra, Rapid Miner/ YALE, Weka, and KNIME. In addition to the aforementioned tools, R is added since the EMC certification in data science is heavily based in R.

Figure 1: The Orange Environment

Orange

Orange is an open source data mining, visualization environment, analytics, and scripting environment. Figure 1 shows the Orange environment. Widgets are used as the building blocks to create workflows within the Orange environment. Widgets are categorized as Data, Visualize, Classify, Regression, Evaluate, Associate, and Unsupervised. Data widgets enable data manipulation such as discretization, concatenation, and merging. Visualization widgets perform graphing such as plotting, bar graphs, and linear projection. Classification widgets are at the heart of the Orange functionality and can be employed for multiple decision trees such as C4.5 and CART, k-nearest neighbor, support vector machines, Naïve Bayes, and logistic regression. Regression widgets have logistic and linear regression as well as regression trees. Evaluation widgets contain standard evaluations such as ROC curves and confusion matrices. Associate widgets have association rule mining (ARM) capabilities while unsupervised capabilities include k-means clustering, principle component analysis (PCM), as well as a host of other capabilities. The Orange environment, paired with its array of widgets, supports most common data science tasks. Support for big data

processing is missing; on the other hand, Orange supports scripting in Python as well as the ability to write extension in C++. Finally, creating workflows is a supported feature via linking widgets together to form a data science process. Figure 1 illustrates the Orange environment.

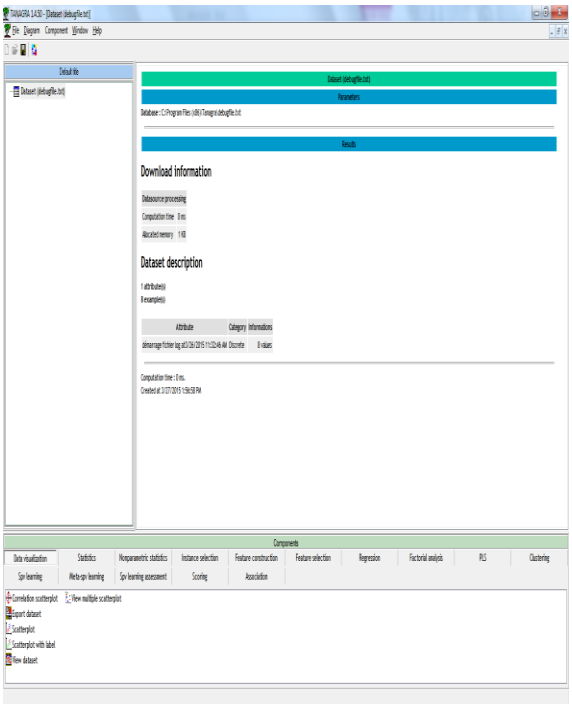

Figure 2: The Tanagra Environment

Tanagra

Tanagra claims to be an open source environment for teaching and research and is the successor to the SPINA software (R Rakotomalala, 2009). Capabilities include Data source (reading of data), Visualization, Descriptive statistics, Instance selection, Feature selection, Feature construction, Regression, Factorial analysis, Clustering, Supervised learning, Meta-Spv learning (i.e bagging and boosting), Learning assessment, and Association Rules. Tanagra is designed for research and teaching; conversely, use in for profit activities is permitted based on the license agreement. One statement in the license agreement specifically addresses commercial use. The translated statement "The software is primarily for teaching and research. Anyone still can load and use, including for profit, without payment and royalties." Tanagra is full featured with multiple implementations of various algorithms (3 for A-Priori alone). Developed in Delphi, extending will prove difficult. Additionally, capabilities for big data processing are not mentioned. Finally, workflows are possible via the diagram menu where tasks may be added and processed in order. Figure 2 illustrates the Tanagra environment.

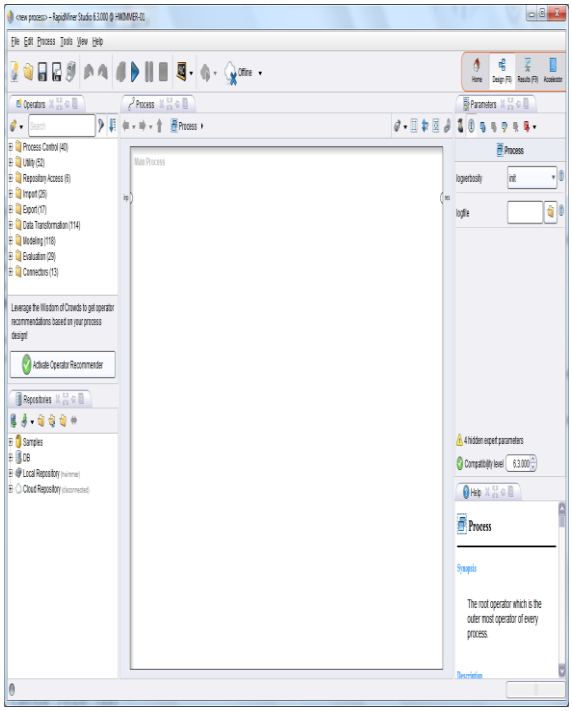

Figure 3: The Rapid Miner Environment

Rapid Miner

Rapid Miner, formerly Yale, has morphed into a licensed software product as opposed to open source; nevertheless, Rapid Miner community edition is still free and open source. Rapid Miner has the ability to perform process control (i.e. loops), connect to a repository, import and export data, data transformation, modeling (i.e. classification and regression), and Evaluation. While many features are available in the open source version certain features are not enabled. One such example is data sources. The open source version only supports CSV and MS Excel and no access to database systems. Aside from data connectivity, memory access is limited to 1GB in the free starter version. Rapid Miner is full-featured with the ability to visually program control structures in the process flows. Additionally, modeling covers the important methods such as decision trees, neural networks, logistic and linear regression, support vector machines, Naïve Bayes, and clustering. In some instances, such as k-means clustering, multiple algorithms are implemented leaving the data scientist with options. Big data processing, Rapid Miner's Radoop, is not available in the free edition. Finally, the ability to create workflows is

well implemented in the Rapid Miner environment which is shown as figure 3.

KNIME

KNIME is the Konstant Information Miner which had its beginnings at the University of Konstanz and has since developed into a full-scale data science tool. There are multiple versions of KNIME each with added capabilities. Much like Rapid Miner, advanced capabilities and tools come at a price. Functionalities include univariate and multivariate statistics, data mining, time series analysis, image processing, web analytics, text mining, network analysis, and social media analysis. Commercial extensions as well as an open source community provide extensions that may be purchased or downloaded. KNIME provides an open API and is based in the Eclipse platform which facilitates developers extending functionalities. Additionally, support for Weka analysis modules and R scripts can be downloaded. KNIME boasts over 1000 analytics routines, either natively or through Weka and R (KNIME.org, 2015). Big data processing is not included in the free version but may be purchased as the KNIME Big Data Extension. Support for workflows is built in to all versions and illustrated in figure 4.

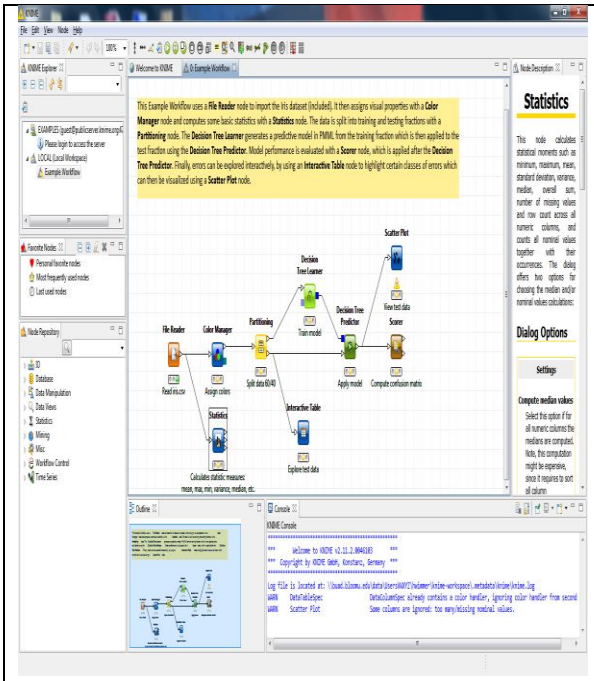

Figure 4: The KNIME Environment with Sample Workflow

R

R is a free and open source package for statistics and graphing. R is traditionally command line; however, there are many feely available open source tools that integrate into R. One such example is R Studio which provides a graphical user interface for R. R can be employed for a variety of statistical and analytics tasks including but not limited to clustering, regression, time series analysis, text mining, and statistical modeling. R is considered an interpreted language more so than an environment. R supports big data processing with RHadoop. RHadoop connects R to Hadoop environments and runs R programs across Hadoop nodes and clusters. Natively, visual features are not available making creating workflows challenging, especially for a novice; still, its broad community provides many graphical utilities such as R Studio shown as figure 5.

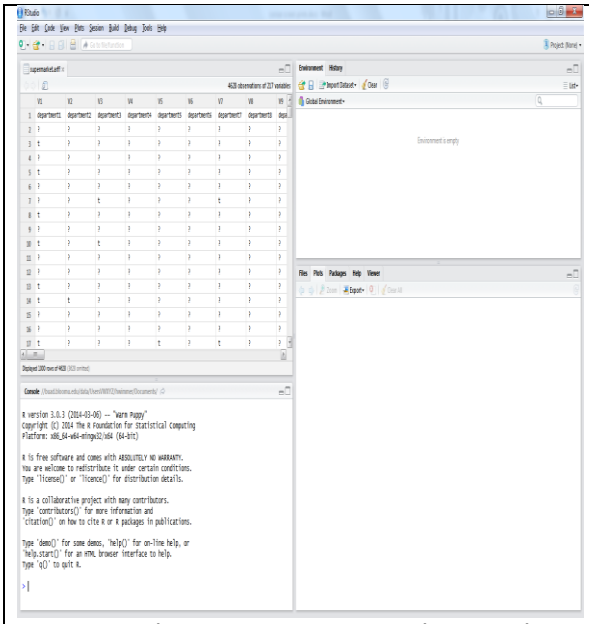

Figure 5: The R Environment with R Studio with Sample Data

Weka

Weka, or the Waikato Environment for Knowledge Analysis, is licensed under the GNU general public license. Weka stems from the University of Waikato and is a collection of packages for machine learning and is Java based. Weka provides an API so developers may use Weka from their projects. Weka is widely adopted in academic and business and has an active community (Hall et al., 2009). Weka's community has contributed many add-in packages such as k-anonymity and l-diversity for privacy preserving data mining and bagging and boosting of decision trees. Tools may be downloaded from a repository and via the

package manager. Weka is java based and extensible. Weka provides .jar files which may be built into any Java application permitting custom programming outside of the Weka environment. The basic Weka environment with sample data is illustrated as figure 6. For big data processing, Weka has its own packages for map reduce programming to maintain independence over platform but also provides wrappers for Hadoop. Weka has workflow support via its Knowledge Flow utility.

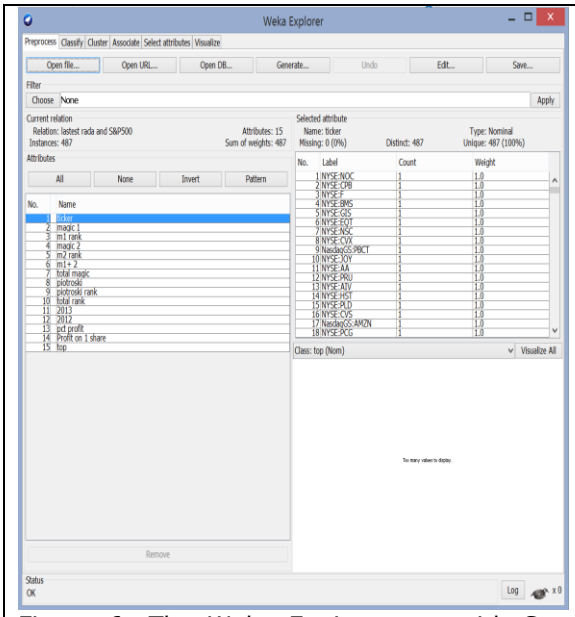

Figure 6: The Weka Environment with Sample Data

Comparison Matrix

The comparison matrix shows the open source tools and their support for common data science techniques. Based on the matrix, WEKA offers the most support on an open source basis; however, each software tool has unique features and strengths. While R is a close second, R requires more in-depth technical skills to execute basic tasks. Tools like Rapid Miner, KNIME, Orange, and Tanagra provide more visual approaches; however, there is an associated cost. KNIME requires a complicated installation process. Along those lines, Tanagra was developed for teaching and research; therefore, its capabilities may be outside the reach of the lay-person. Rapid Miner has a simple installation; however, much functionality is removed from the open source version. Similar to Rapid Miner, Orange's visual approach and widget functionality introduces a simplified approach to creating data science tasks. One advantage to Rapid Miner is the availability of commercial support. Prior to adopting a tool in a data science project it is important to consider the skills of the data scientists and domain experts, the scope of the project, future growth, and available budgetary constraints to name a few

	Orange	Tanagra	Rapid Miner	KNIME	R	Weka
K-means Clustering	Yes	Yes	Yes	Yes	Yes	Yes
Association Rule Mining	Yes	Yes	Yes	Yes	Yes	Yes
Linear Regression	Yes	Yes	Yes	Yes	Yes	Yes
Logistic Regression	Yes	Yes	Yes	Yes	Yes	Yes
Naïve Bayesian Classifiers	Yes	Yes	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes	Yes	Yes
Time Series Analysis	No	No	Some	Yes	Yes	Yes
Text Analytics	Yes	No	Yes	Yes	Yes	Yes
Big Data Processing	No	No	No	No	Yes	Yes
Visual WorkFlows	Yes	Yes	Yes	Yes	No	Yes

4. CONCLUSION

Data science is one of the most in demand professions available with projected growth and shortfalls in supply driving up salary for the position. Efforts in data science are challenging with high software costs that are prohibitive to small and medium size organizations whether in a business or a clinical environment. Data science provides a competitive advantage to business and can be employed to lower the costs of healthcare and has the potential to improve quality of life for patients. Training the next generation of data scientist in an academic setting is challenging due to shrinking academic budgets for software. In order to address these issues, this work provides an overview of the open source tools available to the data scientist.

The definition of data science varies; therefore, this paper defines data science as the intersection of data mining, machine learning, predictive analytics, and business intelligence. Techniques of the data scientist are extracted from one of the available industry certifications. We highlight reviews already available in the

academic literature in order to extend the current literature. Open source tools are selected via the intersection of reviews in academic literature and practitioner websites. Each open source tools is described detailing its history and capabilities. A matrix is presented detailing the capabilities of each of the open source tools available. Future research will include exploring implementations of open source data science tools, comparison of algorithmic efficiency and accuracy, as well as furthering a clear definition of the data science field.

5. REFERENCES

Auza, J. (2010). 5 of the best and free open source data mining software. Retrieved from http://www.junauza.com/2010/11/free-data-mining-software.html

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2008). *KNIME: The Konstanz information miner*: Springer.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'. *McKinsey Quarterly, 4*, 24-35.

Cerrito, P. B. (2006). *Introduction to data mining using SAS Enterprise Miner*: SAS Institute.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chen, X., Ye, Y., Williams, G., & Xu, X. (2007). A survey of open source data mining systems *Emerging Technologies in Knowledge Discovery and Data Mining* (pp. 3-14): Springer.

Davenport, T. H., & Patil, D. (2012). Data scientist. *Harvard Business Review, 90*, 70-76.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., . . . Starič, A. (2013). Orange: data mining toolbox in python. *the Journal of machine Learning research, 14*(1), 2349-2353.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64-73.

EMC. (2015). Data Science Associate. Retrieved from https://education.emc.com/guest/certification/framework/stf/data_science.aspx

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37.

GNU.org. (2015a). GNU General Public License. Retrieved from https://www.gnu.org/copyleft/gpl.html

GNU.org. (2015b). GNU General Public License. Retrieved from http://www.gnu.org/licenses/

Goethals, B., Nijssen, S., & Zaki, M. J. (2005). Open source data mining: workshop report. *ACM SIGKDD Explorations Newsletter, 7*(2), 143-144.

Goopta, C. (2014). Six of the best open source data mining tools. Retrieved from http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.3671

Hempel, J. (2012). The hottest tech gig of 2022: Data scientist. Retrieved from http://fortune.com/2012/01/06/the-hot-tech-gig-of-2022-data-scientist/

Holmes, G., Donkin, A., & Witten, I. H. (1994, 29 Nov-2 Dec 1994). *WEKA: A machine learning workbench.* Paper presented at the Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of computational and graphical statistics, 5*(3), 299-314.

Institute, E.-B. T. (2005). Alphaminer: An Open Source Data Mining Platform. Retrieved from http://www.eti.hku.hk/alphaminer/

Khabaza, T., & Shearer, C. (1995). Data mining with Clementine. *IET*.

KNIME.org. (2015). KNIME Analytics Platform. Retrieved from http://www.knime.org/files/Marketing/Datasheets/KNIME_Analytics_Platform_PDS.pdf

Kohavi, R., John, G., Long, R., Manley, D., & Pfleger, K. (1994). *MLC++: A machine learning library in C++.* Paper presented at the Tools with Artificial Intelligence, 1994. Proceedings., Sixth International Conference on.

Lake, P., & Drake, R. (2014). The Future of IS in the Era of Big Data Big Data *Information Systems Management in the Big Data Era* (pp. 267-288): Springer.

Laurent, A. M. S. (2004). *Understanding open source and free software licensing*: " O'Reilly Media, Inc.".

Loukides, M. (2010). What is data science.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). *Yale: Rapid prototyping for complex data mining tasks.* Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Money, C. (2015). Best Jobs in America. Retrieved from http://money.cnn.com/pf/best-jobs/2013/snapshots/32.html

Negash, S. (2004). Business intelligence. *The Communications of the Association for Information Systems, 13*(1), 54.

Rakotomalala, R. (2008). Tangara. Retrieved from http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

Rakotomalala, R. (2009). Sipina data mining software.

Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R., & Lin, H. (2005). ADaM: a data mining toolkit for scientists and engineers. *Computers & Geosciences, 31*(5), 607-618.

Saravanan, V., Pushpalatha, C., & Ranjithkumar, C. (2014). Data Mining Open Source Tools-Review. *International Journal of Advanced Research in Computer Science, 5*(6).

Shmueli, G., & Koppius, O. (2010). Predictive analytics in information systems research. *Robert H. Smith School Research Paper No. RHS*, 06-138.

Swayne, D. F., Lang, D. T., Buja, A., & Cook, D. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis, 43*(4), 423-444.

Togaware. (2006). The Gnome Data Mine. Retrieved from http://www.togaware.com/datamining/gdatamine/

Ultsch, A., & Mörchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM.

Walters, B. W. (2007). *Understanding Open Source Software*. Paper presented at the ASEE Southeast Section Conference, Louisville, KY.

Williams, G. J. (2009). Rattle: a data mining GUI for R. *The R Journal, 1*(2), 45-55.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Zücker, R., Kietz, J.-U., & Vaduva, A. (2001). Mining mart: metadata-driven preprocessing.

Zupan, B., & Demsar, J. (2008). Open-source tools for data mining. *Clinics in laboratory medicine, 28*(1), 37-54.