International Journal of Open Source Software and Processes
Volume 12 • Issue 1 • January-March 2021

A Survey of Open Source Statistical Software (OSSS) and Their Data Processing Functionalities

Gao Niu, Bryant University, USA
Richard S. Segall, Arkansas State University, USA
Zichen Zhao, Yale University, USA
Zhijian Wu, New York University, USA

ABSTRACT

This paper discusses the definitions of open source software, free software and freeware, and the concept of big data. The authors then introduce R and Python as the two most popular open source statistical software (OSSS). Additional OSSS, such as JASP, PSPP, GRETL, SOFA Statistics, Octave, KNIME, and Scilab, are also introduced in this paper with function descriptions and modeling examples. They further discuss OSSS's capability in artificial intelligence application and modeling and Popular OSSS-based machine learning libraries and systems. The paper intends to provide a reference for readers to make proper selections of open source software when statistical analysis tasks are needed. In addition, working platform and selective numerical, descriptive and analysis examples are provided for each software. Readers could have a direct and in-depth understanding of each software and its functional highlights.

KEYWORDS

Artificial Intelligence, General Public License, Graphical User Interface, Integrated Development Environment, Machine Learning, Open Source Software, Python, R, Statistical Software

1. INTRODUCTION

In this paper, the authors discuss the most popular open source statistical software with its creation history, target practitioners, and statistical usage examples. Although Programming languages such as Java, C++ can also perform statistical analysis with intensive coding, the authors limit discussion to the software specifically designed for statistical analysis.

The motivation of this paper started from a research insight book of Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities (Segall & Niu, 2020), edited by the first two authors of this book. The book introduces OSSS, presents multiple applications and discusses research opportunities. This paper summarizes the information, extends the discussion to

DOI: 10.4018/IJOSSP.2021010101

a broader statistical processing functionality such as machine learning and artificial intelligence. We first introduced an artificial intelligence (AI) techniques categorization, and surveyed the popular OSSS designed for AI applications.

The objective of this paper is to create a reference for the readers and guide them to make proper selection of open source software when a statistical analysis task is in demand. The discussion includes the background information, research areas that the software designed for, and the overview of how to use the software.

First section of the paper introduces the definition OSSS, several similar type of software such as Free Software and Freeware are compared, both traditional software and open source software development are summarized. Second section of the paper presents multiple popular OSSS, such as R, Python and etc., designed for statistical applications are presented. Third section of the paper presents OSSS designed for AI are presented. Popular AI techniques are categorized and briefly described, and OSSS designed for AI processing are presented. The authors focus on creating an overview of all open source statistical software in this paper.

The article introduces current machine learning data processing platform. Readers can be benefitted from this short reference of Open Source Statistical Software.

2. BACKGROUND

2.1 How Open Source Software, Free Software, And Freeware Differ

2.1.1 Open Source Software (OSS)

Open Source Software (OSS) is a type of computer software in which source code is released under a license in which the copyright holder grants users the rights to study, change, and distribute the software to anyone and for any purpose. (Wikipedia (2019a))

For software to be considered "Open Source", it must meet ten conditions as defined by the Open Source Initiative (OSI). Of these ten conditions, it's the first three that are really at the core of Open Source and differentiates it from other software. These three conditions are according to the Open Source Initiative (2007):

1. **Free Redistribution**: The software can be freely given away or sold.
2. **Source Code**: The source code must either be included or freely obtainable.
3. **Derived Works**: Redistribution of modifications must be allowed.

The other conditions are: (Open Source Initiative (2007))

4. **Integrity of The Author's Source Code**: Licenses may require that modifications are redistributed only as patches.
5. **No Discrimination against Persons or Groups**: no one can be locked out.
6. **No Discrimination against Fields of Endeavor**: commercial users cannot be excluded.
7. **Distribution of License**: The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.
8. **License Must Not Be Specific to a Product**: the program cannot be licensed only as part of a larger distribution.
9. **License Must Not Restrict Other Software**: the license cannot insist that any other software it is distributed with must also be open source.
10. **License Must Be Technology**: Neutral: no click-wrap licenses or other medium-specific ways of accepting the license must be required.

Macaulay (2017) discussed benefits of open source software that are summarized in Figure 1 below.

2.1.2 Open Source License

According to Wikipedia (2019f) an open source license is a type of license for computer software and other products that allows the source code, blueprint or design to be used, modified and/or shared under defined terms and conditions. This allows end users and commercial companies to review and modify the source code, blueprint or design for their own customization, curiosity or troubleshooting needs.

Open-source licensed software is mostly available free of charge, though this does not necessarily have to be the case.

Licenses that only permit non-commercial redistribution or modification of the source code for personal use only are not considered generally as open source licenses.

2.2 Free Software or Freeware

Unlike the Open Source term, Free Software only has 4 "Freedoms" with its definition and are numbered 0-3 as created by the Free Software Foundation (FSF) (2019a) as follows:

The freedom to run the program for any purpose (Freedom 0)

The freedom to study how the program works and adapt it to your needs (Freedom 1)

The freedom of redistribution of software (Freedom 2)

The freedom to improve the program and release your improvements to the public to benefit the while community. (Freedom 3)

Although not explicitly outlined as a freedom, access to source code is implied with Freedoms 1 and 3. You need to have the source code in order to study or modify it. Figure 2 illustrates the relationship and overlap of these properties of Free Software with Open Source Software and was drawn using Drake (2019) discussion of the difference between free and open source software.

Figure 3 compares the features of freeware versus shareware that illustrates the later has fewer features than the former freeware.

2.3 Free Open Source Software (FOSS)

Not all software is free and Free Open Source Software (FOSS) is both free and open. The Free Software Foundation (FSS) (2019b) provides a searchable directory of over 15,000 free software packages.

Free and open-source software (FOSS) is software that can be classified as both free software and open-source software. That is, anyone is freely licensed to use, copy, study, and change the software in any way, and the source code is openly shared so that people are encouraged to voluntarily improve the design of the software. This is in contrast to proprietary software, where the software is under restrictive copyright licensing and the source code is usually hidden from the users. (The Free Software Foundation (FSS) (2019b))

Open source software (OSS) is a type of computer software that had its code released to the public. St. Laurent (2008) indicated that users have the right to study, change and redistribute the software under the copyright granted by the software license holder. Closed source or proprietary

Figure 1. Benefits of Open Source Software (OSS) (derived from Macaulay (2017))

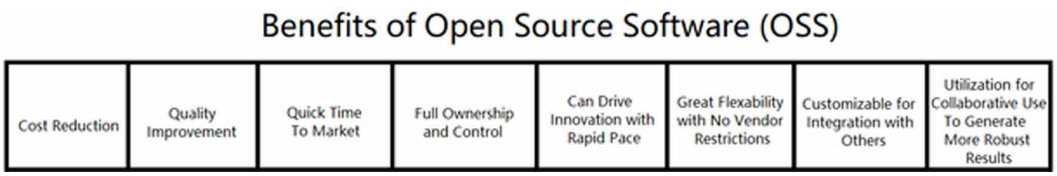


Figure 2. Comparisons of features of Open Source Software (OSS) versus Free Software

Figure 3. Comparison of the features of Freeware and Shareware

Features of Freeware and Shareware	
Freeware	**Shareware**
No cost to acquire	Limited Free Trial Period
Normally shared with no source code	
Unable to modify	Probably Limited Features of Full Version
Not Proprietary Protected	

software can only be modified and maintained by the people, teams and organizations who own the software. Microsoft Office and Adobe Photoshop are well-known proprietary software.

Open source software is popular to statistical analysis practitioners, not only because it is free, but also because it is more adaptive to the current rapidly developing academic research advancement environment.

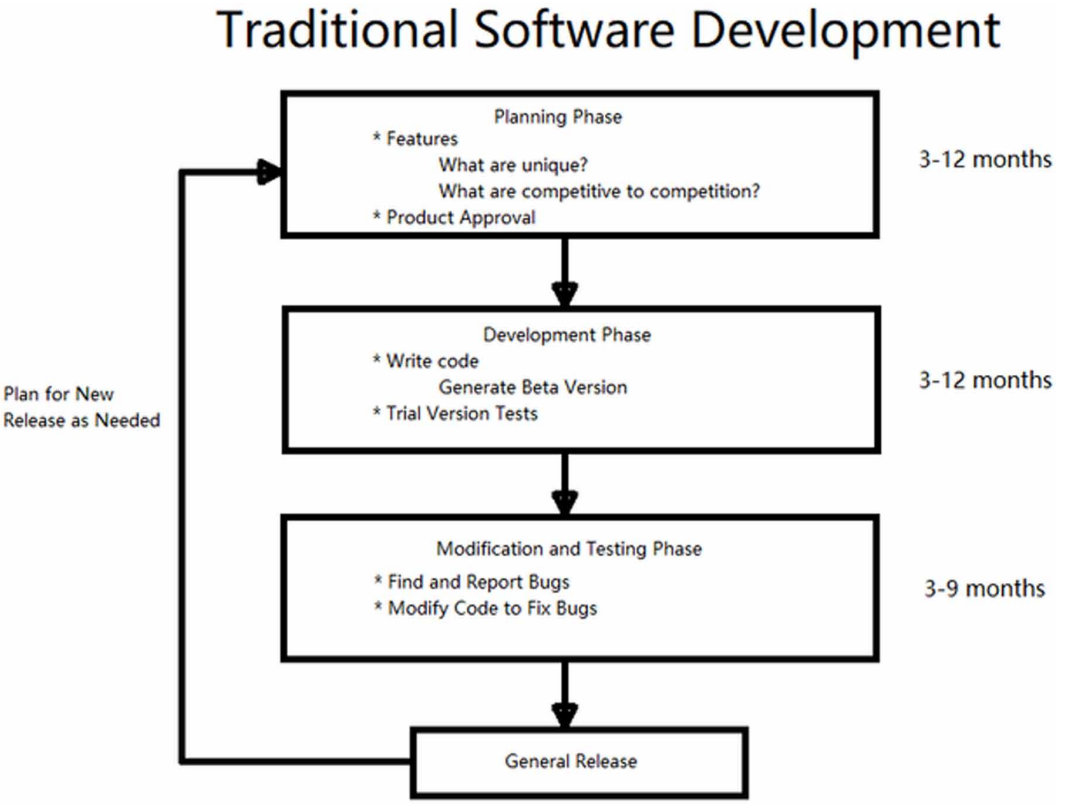
2.4 Open Source Software Development versus Traditional Software Development

The traditional method for software development includes a planning phase, development phase, modification and testing phase before general release followed by potential plans for new releases as needed as shown in Figure 4. The time line for this cycle typically is anywhere form nine months to almost three years.

Saini & Kaur (2014) performed an extensive review of open source software development life cycle models, and Mandal et al. (2011) preformed open incremental model for an Open Source Software Development Life Cycle Model 'OSDLC'.

The Open Source Development Life Cycle (OSDLC) as discussed by Linux Foundation (2011), Haddan (2008), and Goldman & Gabriel (2005) entails multiple internal users/developers who each provide improvements to the open source software (OSS) prior to its release to worldwide users/developers. These external users then subsequently provide an additionally enhanced source code

Figure 4. Traditional software development cycle



for a new release version for potentially additionally improved source code upon feedback from both internal and external users. Figure 5 below illustrates the Open Source Development Cycle.

Freeman et al. (2018) discussed several of the best open source software for software development from which the following Table 1 was derived.

Next, we first introduce the two most popular Open Source Statistical Software (OSSS) R and Python along with its Integrated Development Environment (IDE) and Graphical User Interface (GUI). Then, additional OSSS, like JASP, PSPP, GRETL, SOFA Statistics, Octave, KNIME and Scilab, are introduced with description of their functions and modeling examples. Software selection and description are consistent with the book chapter named *Introduction to the Popular Open Source Statistical Software (OSSS)* by the authors (Wu, Zhao, & Niu, 2020).

3. OPEN SOURCE STATISTICAL SOFTWARE

3.1 R

R is arguably the most popular open source statistical software. It has a strong statistical analysis capability and graphical visualization functionality. This section provides an overview of the software R by introducing most used packages, its popular IDEs and its functionalities. R was initially written by Ross Ihaka and Robert Gentleman from the University of Auckland, New Zealand (The R Foundation, 2020). R is licensed under GNU General Public License. As of September 2020, the most current version is 4.0.2, it was released on Jun, 22nd, 2020. The official website of the software is https://www.r-project.org/.

Figure 5. Open Source Software (OSS) development cycle

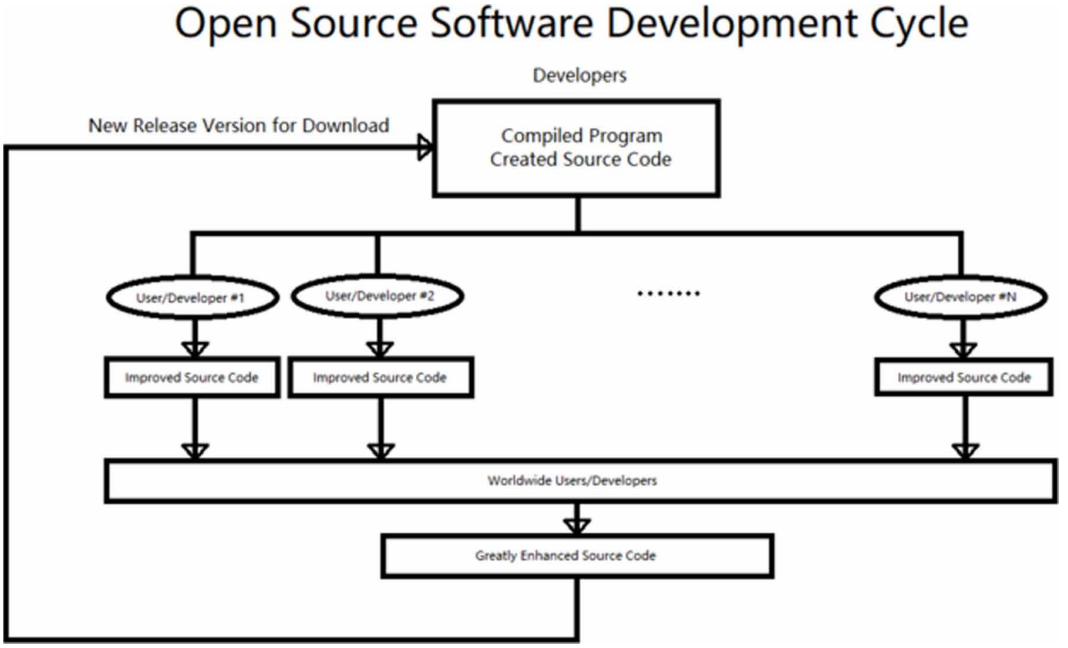


Table 1. Open Source Software for software development

Open Source Software Name	Features
Tuffle Framework	Suite of tools to help develop. test, and deploy smart contracts to the Ethereum blockchain.
Blockstack	Set of application development tools for building blockchain-based decentralized applications (dapps) on the Bitcoin blockchain.
Julia	High performance dynamic programming language for numerical computing.
Taucharts	Data-focused JavaScript charting library

As of September 7th, 2020, there are 16,236 CRAN packages (Contributed Packages, 2020). The authors select a few popular packages, categorize them by functionality, and list in table 2 based on Awesome R (2019).

Table 3 lists out popular R IDEs. Among them, RStudio, Jamovi and architect are further discussed with more details.

3.1.1 RStudio

RStudio is one of the most popular IDEs. The working platform includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial edition. It runs on the desktop (Windows, Mac and Linux) and in browsers that connected to RStudio Server or RStudio Server Pro (RStudio, 2019). RStudio was founded in 2008, and its initial public release was in 2011. As of September 2020, the most current desktop version of the RStudio was released on Aug 11th, 2020,

Table 2. Selected popular R packages by category

R Package Function Area	Package Names
Data Manipulation	dplyr, Data.table, readr, rlist, vroom, DataExplorer
Visualization	ggplot2, ggfortify, lattice, gganimate, plot3D
Database	RODBC, DBI, odbc, RMySQL, ROracle, RSQLite, RHive
Statistical Analysis and Machine Learning	Bigrf, C50, forecast, prophet, gbm, glmnet, kernlab, lasso2, maptree, randomForest, tree, mcmc, coda, igraph, network, ggmap, Remap, spacetime, spatstat, tigris
Finance	PerformanceAnalytics, fAssets, tseries, scorecard

Table 3. Selected popular R IDEs

R IDE	Short Description
RStudio	One of the most popular R user interfaces
Jamovi	GUI similar to SPSS and focuses on Bayesian and Frequentist analysis
Architect	IDE that focuses on the need of data scientist
StatET	Eclipse based IDE for R
Revolution R Enterprise	Free to academics, commercial software focus on big data.
R Commander	GUI
Deducer	Friendly user interface with direct data editing function
Radiant	Browser based user interface based on Shiny
Bio7	Focuses on ecological modeling
RTVs	R tools for visual studio

the version number is 1.3.1073. RStudio Server Pro and RStudio Shiny Server Pro were announced in 2013, and RStudio Connect was delivered in 2016.

RStudio is a popular and strong IDE which improves user experience of R significantly. Original R software has almost no user interface function. All syntax needs to be programmed and executed in order to reach the output. RStudio provides a powerful coding support with recommended codes, error check which enhances users' productivity. With the help of RStudio, source data is accessed more visually and directly. The IDE is also linked with online databases. Users could utilize its automated package lookup, download and install functions instead of searching and downloading packages manually. RStudio supports more interactive graphics with RShinny. RStudio is well-received by researchers and practitioners from many areas, such as statistics, biostatistics, mathematics, actuarial science, finance, engineer, business and etc.

3.1.2 Architect

Architect is an Eclipse-based cross-platform IDE for R. Architect is a fully open source software. It is available on Windows, Mac and Linux. As of September 2020, the most current version of Architect is 0.9.11. The official website is https://www.getarchitect.io/

Architect provides a user-friendly interface that supports all data science tasks from statistical analysis to report generation. The working plate platform includes integrated R console, object browser, data viewer, graphing tools, package development, workspace management, and debugging tools. Architect is fully embedded in the Eclipse ecosystem (Open Analytics NV, 2020), and it can

be used to work in multiple programming languages such as R, Python, Julia, Scala, C++ etc. It can also connect to NoSQL database which makes big data analysis easier.

3.1.3 Jamovi

Jamovi is a free open source GUI for R. It improves the functionality of R in two aspects. On one hand, the GUI component of Jamovi makes programming easier, especially for users who are not familiar with R syntax. Muenchen (2018) indicated that Jamovi added functions and methods of programming that other software has, such as SPSS and SAS. Lakens (2017) wrote that Jamovi was developed by a group of developers who used to work on JASP, thus the user interface and functionality have a lot of similarities. It can be operated on Windows, MacOS, Linux and ChromeOS. The official website is https://www.jamovi.org/. As of September 2020, the most current version of the software was released on August 31st, 2020, the version number is 1.6.0. Jamovi's official website (2020) mentioned that it intends to be a free open source version of the costly statistical products.

Jamovi is adaptive to various data formats. SPSS, SAS and Stata files can be imported to Jamovi directly. The user interface of Jamovi has live data management, users could edit and modify data directly without syntax. Edelsbrunner (2017) indicated that, in Jamovi, data is dynamically linked with its analysis, and test results are automatically updated after the data has been edited or modified. The analytical results are edited such that tables, results can be easily copied and pasted into other editing software, such as LaTex, Word, PowerPoint and etc.

3.2 Python

Python is one of the most popular open source programming software used for automation, artificial intelligence, application, websites as well as statistical analysis for big data. This section provides a brief overview of the software and several popular IDEs, such as Python IDEs, PyCharm, Spyder and Jupyter. Python was created by Guido van Rossum in the early 1990s from Stichting Mathematisch Centrum in Netherlands (Python Software Foundation, 2020). Python is licensed under General Public License (GPL)-compatible. The difference is that GPL compatible makes it possible to combine Python with other software that are released under the GPL. As of September 2020, the most current version of Python is 3.8.5. The official website of the software is https://www.python.org/.

3.2.1 PyCharm

PyCharm is one of the most popular Python IDEs. According to Taft (2010), PyCharm was developed by the Czech company JetBrains. It runs on Windows, Mac and Linux. As of September 2020, the most current version of PyCharm is 2020.2.1 and it was released on August 25th, 2020.

The basic text editor for Python is called IDLE (Integrated DeveLopment Environment) which provides limited support in code editing. PyCharm organizes files by projects and provides a user interface that can be modified with high flexibility.

3.2.2 Spyder

Spyder is a Python IDE that designed for researchers, scholars, data analyst, engineers and etc. It provides a strong support for data analysis tasks with its built-in integration of many popular scientific packages, such as NumPy, SciPy, Pandas, IPython and more. Spyder runs on Windows, Mac and Linux. Spyder is included in Anaconda, and it is pre-installed in Anaconda Navigator. Anaconda is not an IDE, but a python distribution.

Spyder provides clean user interface and visualization. The key features within the IDE include Editor, IPython Console, Variable Explorer, Help, Static Code Analysis, Profiler, Projects, File Explorer, Find in Files, Online Help, History Log, Internal Console and etc.

3.2.3 Jupyter

Project Jupyter is a non-profit organization that "exists to develop open-source software, open standard, and services for interactive computing across dozens of programming languages." (Jupyter, 2019). The two main platforms JupyterLab and Jupyter Notebook are both web-based was evolved from IPython, and was created by Perez (2014).

3.3 JASP

JASP is free and open source statistical analysis, image process software. The most recent version of JASP is 0.13.1 released July 16th, 2020. It can be operated on Windows, MacOS X, and Linux. The programming languages used to develop this software are C++, R, and JavaScript. It can be considered as an open source version of SPSS. It is user friendly, and can be used interchangeably with SPSS. In the welcome page, JASP claims that it "aims to be a complete and full featured alternative to SPSS". The license is GNU Affero General Public License. JASP is developed and funded by several universities and research funds. The office website is https://jasp-stats.org/

JASP specializes and supports Bayesian statistical analysis and frequencies inference. For the Bayesian analysis, the software uses prior observations to estimate the posterior estimations and make inferences. The software has user-friendly Bayesian t-tests, Bayesian correlation analysis, Bayesian Linear Regression and more. The software produces APA format graphs, which can be used widely in academic publications.

JASP has a balanced user interface. Users could perform descriptive, frequency, and Bayesian analysis directly by selecting statistical assumptions, as mentioned by Wagenmakers, et al., (2018) and Love, et al., (2015). The results will be calculated and displayed on the right-hand side window automatically.

3.4 PSPP

PSPP was published in 1998, and was written in C. As of September 2020, the most current version of the software was released on September 2020, the version number is 1.4.1 by Ben Pfaff. It is also a replacement software for SPSS. One important advantage of selecting PSPP is that the free version of the software already includes variety of advanced statistical packages. The official website of the software is https://www.gnu.org/software/pspp/

The software has three different modes. The Terminal Mode has clean user interface, no additional windows overlapping each other. The Graphic User Interface Mode is user-friendly and has limited syntax usage. Users can input, define, modify, and analyze data by clicking options within the interface. Also, it is adaptive to other spreadsheet applications which makes data transformation across software easily. The third mode is Non-Interactive mode which allows users access the source code directly. It has flexible output format, such as Unicode Text with UTF-8 encoding, PDF, HTML and more (GNU PSPP Screenshots, 2018).

PSPP user interface includes traditional options, such as file, edit, view, window, and help, which provide practitioners commonly used functions to input, output data and modify platform appearances. In addition, PSPP also has data, transform, analyze, graphs, and utilities as its core data analysis functions. Data function allows users to modify data and data file directly, such as sort cases and split files. Transform function allows user perform calculation and modification on the data in various levels. For example, the compute function under transform allows users to calculate the sum of two variables. Analyze function includes the core preprogramed statistical analysis, such as descriptive analysis, cluster analysis, regression etc. Current version of preprogramed SPSS also produces limited but high-quality graphs, such as Scatterplot, Histogram, and Barchart.

3.5 GRETL (GNU Regression, Econometrics and Time-series Library)

GRETL is short for GNU Regression, Econometrics and Time-series Library. It is an open source statistical resource package and mainly serves the field of econometrics. GRETL has been reviewed multiple times by *Journal of Applied Econometrics*, such as Baiocchi and Distaso (2003), Yalta and Yalta (2007) and Mixon and Smith (2006). The earliest version of the software was released on January 31st, 2000. As of September, the most current version is 2020d, which was released on August 6th, 2020. It was written in C language. GRETL has GUI (graphical user interface) and command-line interface. It is adaptive to multiple operating systems, such as Windows, MacOS, and Linux. There are multiple language environments available, such as English, Chinese, French, German and etc. GRETL's GNU license is GPLv3. The office website: gretl.sourceforge.net

GRETL's native scripting language is Hansi. Its add-ons (packages) also need to be written by Hansi. However, the software is adaptive and able to work together with other statistical software including R, Python, Stata, and Julia.

In addition to GRETL's own data format XML, own binary databases (allowing mixed data frequencies and series lengths). It also supports the most popular data format, such as CSV, Excel, Stata.dta files, SPSS.sav files and etc. GRETL is adaptive and able to exchange data and results with other popular statistical software, such as R, Octave, Python and Stata. GRETL's output is in the format of LaTex (GNU Regression, Econometrics and Time-series Library, 2019).

GRETL user interface includes traditional options including file, data, view, and help. In addition, GRETL has Tools, Sample, Variable and Models as its core data analysis functions selected from its top tool bar.

3.6 SOFA (Statistics Open for All) Statistics

SOFA statistics is an open source statistical software. It is short for Statistics Open for All. SOFA can produce many well-designed graphs for presentation. It can also perform basic statistical analysis and used on multiple operating systems including Windows, MacOS, and Linux. As of September 2020, the most current version of the software was released on November 24th, 2019, the version number is 1.5.3. The major feature of SOFA statistics is its user-friendly interface, which could connect to the commonly used database directly, such as MySQL, MS Access (mdb), Microsoft SQL Server etc. (SOFA Statistics, 2019) It also has several other easy-to-use features, such as direct data entry, spreadsheet management, and graph share functions. The office website of the software is https://www.sofastatistics.com/home.php

The statistical analysis is limited in SOFA. They are under the tab of Statistics from the bottom left main user interface.

SOFA statistics preprograms eight types of charts. Its Graphical User Interface makes the graphical design process user-friendly. The graphs are featured by its professional design and color choice. It is adaptive to most presentation platforms and can be directly and easily implemented.

3.7 Octave

Octave is a free and open source software, which is designed for intensive numerical computation. It is used to model linear and nonlinear data. Octave has efficient matrix calculation algorithm (GNU Octave, 2020). Miao (2017) discussed that Octave is widely used in engineer and academic researches. For example, NASA used Octave to develop the flying object connecting system. Octave is very flexible in numerical programming and provides user-friendly visualization solutions. Octave was first published in 1988. As of September 2020, the most current version of the software was released on January 31st 2020, the version number is 5.2.0. Octave was written in C programming language, C++ and Fortran. It is adaptive to Windows, MacOS, Linux, and BSD (GNU Octave, 2020). The Octave's GNU license is GPLv3. There are multiple language environments available,

such as English, Spanish, Chinese, French and etc. The official website of Octave is https://www.gnu.org/software/octave/.

Octave can be considered as an open source version of MATLAB, many functions available in MATLAB can also be executed directly in Octave. However, there are some subtle syntax between Octave and MATLAB. Octave also has ample free packages can be downloaded from Octave Forge. Octave has community packages and external packages. Community packages are designed and maintained by Octave Forge and Octave Developers. External packages are developed by a third party which fulfill the requirements for hosting at Octave (Octave Forge, 2019).

Octave has multiple features. Basic mathematical operators, including sin, cos, tan, exp, log, and floor, can be performed. Octave is an interpreted programming language. Similar to C++ and Java, Octave can be used to define and redefine variables. Traditional mathematical algorithms such as inverse, transpose can be operated on arrays, vectors and matrixes. Although Octave focuses on data operation, it can also be connected to graphing software, for example, GUNPLOT, to generate graphs in multiple dimensions and make comparisons for variable sensitivity tests.

3.8 KNIME (Konstanz Information Miner)

KNIME, also called Konstanz Information Miner (Berthold, et al., 2008), is a free open source software for data processing, statistical analysis, and report generation. It was developed by a team of engineers from the University of Konstanz (German) in 2004 (End to End Data Science, 2019). As of September 2020, the most current version is 4.2.0, which was released on July 13th 2020. The software was written in Java, and it can be operated on Windows, MacOS, and Linux. KNIME is licensed under GNU General Public License. Its official website is https://www.knime.com/.

Tiwari & Sekhar (2007) indicated in 2006, KNIME was introduced to the pharmaceutical research, and it also started rapidly gaining its popularity in business and financial area. Sieb, Meinl, & Berthold (2007) noted that KNIME's data mining and machine learning process are through data pipelining. Users can explicitly make variable selection, executive specific analysis, and validate the results. The software has a well-designed Graphical User Interface, which allows for direct data input, variable transformation, graphical design selection and statistical analysis. The software includes more than 2000 modules, which is called "nodes" in KNIME, for workflow construction. KNIME is considered as the open source version of its proprietary counterpart SAS.

KNIME's GUI allows users to select data source, process data through nodes (modules) and then get the result directly. The adaptive input and output data format include xls, csv, doc, ppt, and pdf. Users need to make selections within the node repository to manage the project. The selection can be simply executed by clicking the corresponding node and dragging the node into the middle top window to create a workflow structure.

3.9 Scilab

Scilab is a free open source software. It is designed for engineer and academic researchers. It can be considered as an open source version of MATLAB. It is adaptive to Windows, MacOS, and Linux. Its official website is https://www.Scilab.org/. Scilab is licensed under GNU General Public License. As of September 2020, the most current version is 6.1.0, which was released on February 25th 2020.

The software has more than 1500 mathematical functions provided by math and simulation features (Scilab, 2019). Continuous and discrete mathematical optimization calculations can be executed through its advanced algorithms efficiently. Scilab also has strong statistical analysis and modeling capability. Scilab is featured by its 2-dimension and 3-dimension visualization functions. Scilab is driven by syntax, and most of its statistical and modeling work are processed through sophisticated codes.

4. ARTIFICIAL INTELLIGENCE (AI) TECHNIQUES OVERVIEW

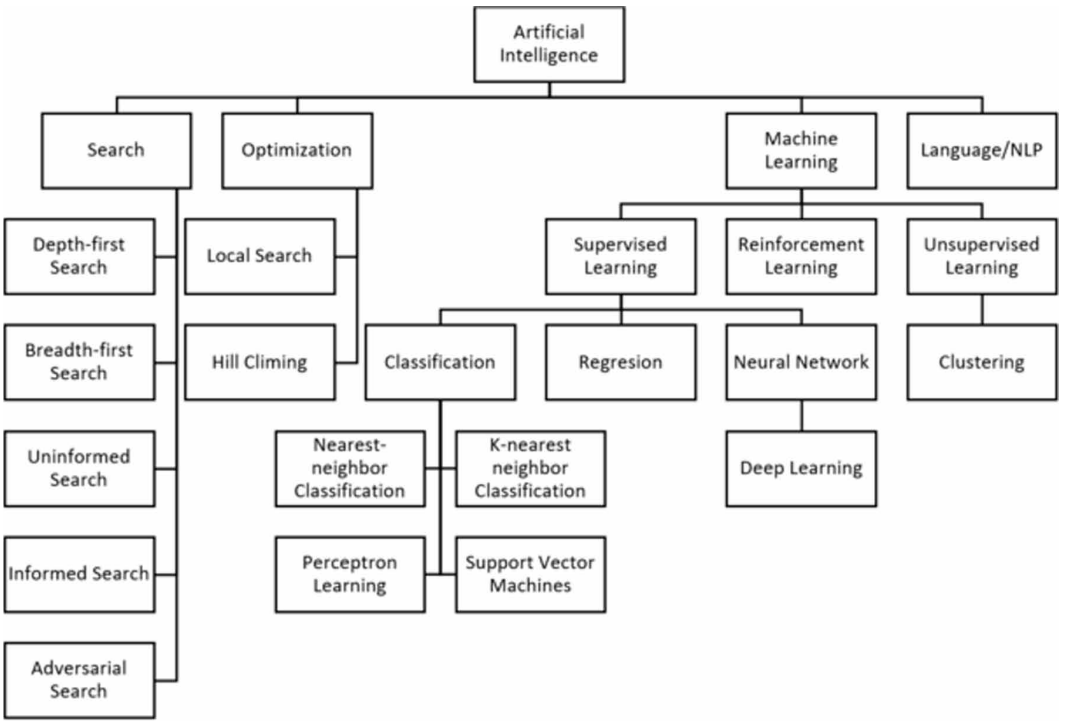
Artificial Intelligence (AI) has been increasingly influencing wide range of areas in the modern world. It is generated by machines and algorithms, in comparison to natural intelligence that generated by humans and animals. Search, optimization, machine learning (ML) and language processing are the four popular artificial intelligence techniques. Machine learning techniques are gaining popularities in the big data analysis. Applications such as search, optimization, machine learning and languages all can be programmed in open source software such as Python, R, and etc. Traditional software such as SAS, Matlab are enhancing its capabilities to program AI and machine learning modeling, open source software is taking the major role as the processing platform. This section briefly introduces the major AI techniques and its categorizations. Figure 6 shows some key AI techniques and its relationships.

Machine learning is one of the most popular used techniques for statistical analysis. It is a modeling process that captures intelligence from data and algorithms. Three major categories are supervised learning, unsupervised learning and reinforcement learning. Supervised learning is when input and output are known and used for training and testing. Unsupervised learning only involves with input data. Reinforcement learning is a process that Classification, regression and neural network are typical and popular supervised learning techniques. Nearest-neighbor classification, k-nearest neighbor classification, perceptron learning and support vector machines (SVM) are popular are popular methods of classifications.

Regression is a traditional statistical method and has been developed for decades. It has been evolved into a model that capable of processing large dataset.

Neural network has a wide range application, from image identification to AI gaming. The basic structure of neural networks has nodes and weights. Deep neural network involves with multiple hidden layers of nodes. The model could be efficiently constructed with hundreds and thousands of

Figure 6. Key artificial intelligence techniques and its categorization (Designed by Niu, 2020)



nodes. Therefore, it stores and models significant amount of data information, and provides effective explanation and predictions. One example is Google's AlphaGo AI defeats Chinese Go Master which is developed based on deep neural networks and tree search. (Silver, et al., 2016) Popular neural network models include convolutional neural networks, recurrent neural networks and etc.

Google map direction is a typical application of search. Techniques includes depth-first search, breadth-first search, uninformed search, informed search and adversarial search.

Optimization involves with make a selection from a set of options. One example of optimization is location selection such as business, hospital, fire department etc. Local search and hill climbing are two systematic optimization techniques. Hill Climbing belongs to the family of local search. It has several variants such as steepest ascent, stochastic, first-choice, random-restart and local beam search.

Language processing and modeling is one important and actively researched artificial intelligence application. Natural language processing (NLP) involves with multiple research aspects, such as automatics summarization, information extraction, language identification, machine language, named entity recognition, speech recognition, and text classification etc. Library NLTK provides rich programming application in Python.

5. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING LIBRARIES AND SYSTEMS

Application programming interface (API) has been evolving into systems that compatible to multiple software platforms. For example, Apache Sparks Ecosystem is compatible to R, SQL, Python, Java and etc. This section discusses the several popular API that are designed for AI/ML processing.

5.1 TensorFlow

TensorFlow is one of the best-known ML and AI platforms. It is a free and open source library developed by Google Brain. TensorFlow can run on central processing unit (CPU) and graphics processing unit (GPU), which has a superior image processing and learning capability. TensorFlow is used for machine learning applications such as deep learning neural networks. (TensorFlow, 2020)

5.2 Theano

Theano is a Python library that evaluates and manipulates matrix-valued mathematical expression. It is developed by Montreal Institute for Learning Algorithms, University of Montreal. Theano provides transparent use of GPU (Bergstra, et al., 2010)which enables deep neural network modeling much faster when processing images compared with CPU.

5.3 Keras

Keras is a neural network library, it is a high-level Application programming interface (API) that runs on top of TensorFlow or Theano. Keras is user-friendly and built in Python. (Thomas, 2017)

5.4 Shogun

Shogun is a free and open-source machine learning library that supports multiple languages, such as Python, R, Octave, Java, C#. Machine learning techniques the library offers includes binary classifier, multiclass classifier, regression, statistical testing, clustering, neural networks and etc. (Shogun, 2020)

5.5 Microsoft Cognitive Toolkit

Microsoft Cognitive Toolkit, also known as CNTK, is an open-source library for neural network and deep learning modeling. It supports Python, C# and C++. It provides commercial-grade distributed deep learning and allows users to program popular neural network models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). (The Microsoft Cognitive Toolkit, 2017)

5.6 Torch

Torch is an open source library specializes in machine learning and provides a wide range of algorithms for deep learning. (Collobert, Kavukcuoglu, & Farabet, 2011) It is a script language based on Lua programming language. Its counterpart PyTorch provides an open source machine learning framework in Python. (From Research to Production, 2020)

5.7 CloudEra

Cloudera is an open source Hadoop distribution and data science workbench uses R, Python or Scala and capable of accessing to Apache Spark and Apache Impala. It is designed for big data processing and is the most popular distribution of Hadoop. (Strata, 2013)

6. CONCLUSION

This paper describes Open Source Software (OSS) and its development history. Similar software, such as Free Software and Freeware. comparisons are discussed. After that, authors introduce a list of popular Open Source Statistical Software (OSSS) with their functionalities and brief development histories. The software selected are up-to-date popular choices for programmers, statisticians, researchers and practitioners.

The paper further discusses Open Source Statistical Software's (OSSS) capability in artificial intelligence application and modeling. Major AI techniques, such as search, optimization, machine learning and natural language processing, and their subcategories are described.

Machine learning has a wide range of application in statistics, finance, social network etc. It is the most popular researched area. Major technology companies invested and developed platforms for machine learning research and application. For example, TensorFlow was developed by Google; CNTK was developed by Microsoft; and CoreML developed by Apple. Popular OSSS-based machine learning libraries and systems are introduced and discussed in this paper.

This paper is served as a short reference for readers to make Open Source Statistical Software (OSSS) selection, and familiarize with the current machine learning data processing platform.

ACKNOWLEDGMENT

The authors refer the reader to more complete descriptions of the above topic in the edited Research Insight book by Segall & Gao (2020) published by IGI Global *Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities*, and its chapters by Segall (2020a) on what is Open Source Software (OSS) and what is big data, Segall (2020b) on Open Source Software (OSS) for big data, and Wu, Zhao & Niu (2020) on Popular Open Source Statistical Software (OSSS).

REFERENCES

Awesome, R. (2019). Retrieved July 18, 2019, from Awesome R: https://awesome-r.com/

Baiocchi, G., & Distaso, W. (2003). GRETL: Econometric Software for the GNU Generation. *Journal of Applied Econometrics*, *18*(1), 105–110. doi:10.1002/jae.704

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., . . . Bengio, Y. (2010, July 6). *Transparent GPU Computing with Theano*. Retrieved from Theano: http://deeplearning.net/software/theano/

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., & Wiswedel, B. (2008). *KNIME: The Konstanz Information Miner. Data Analysis, Machine Learning and Applications*. Springer.

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). *Torch7: A Matlab-like Environment for Machine Learning. In NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*. ACM. Retrieved from https://infoscience.epfl.ch/record/192376?ln=

Contributed Packages. (2020). Retrieved July 18, 2019, from *The Comprehensive R Archive Network*: https://cran.r-project.org/web/packages/

Drake, M. (2017). *The Difference between Free and Open-Source Software*. Retrieved Mary 22, 2019 from https://www.digitalocean.com/community/tutorials/Free-vs-Open-Source-Software

Edelsbrunner, P. (2017, March 23). *Introducing Jamovi: Free and Open Statistical Software Combing Ease of Use with the Power of R*. Retrieved July 18, 2019, from *JEPS Bulletin*: https://blog.efpsa.org/2017/03/23/introducing-jamovi-free-and-open-statistical-software-combining-ease-of-use-with-the-power-of-r/

End to End Data Science. (2019). Retrieved July 18, 2019, from KNIME Official Website: https://www.knime.com/

Free Software Foundation. (2019a). Retrieved May 22, 2019 from https://www.fsf.org/

Free Software Foundation. (2019b). *Free Software Resources*. Retrieved May 22, 2019 from https://www.fsf.org/resources/

Freeman, J., Heller, M., Wayner, P., & Yegulalp. (2018). The best open source software for software development. *InfoWorld*. Retrieved May 22, 2019 from https://www.infoworld.com/article/3306453/the-best-open-source-software-for-software-development.html

From Research to Production. (2020, September 30). Retrieved from PyTorch: https://pytorch.org/

GNU Operating System. (2014, April 12). Retrieved July 19, 2019, from GNU: https://web.archive.org/web/20160718094739/http://www.gnu.org/licenses/gpl-2.0.html

GNU PSPP Screenshots. (2018, March 28). Retrieved July 18, 2019, from GNU Operating System: https://www.gnu.org/software/pspp/tour.html

Goldman, R., & Gabriel, R. P. (2005). How to do Open-Source Development. In *Innovations Happens Elsewhere*. Retrieved May 22, 2019 from https://www.dreamsongs.com/IHE/IHE-54.html#pgfId-956812

Haddan, I. (2008). *The Open Source Development Model: Overview, Benefits and Recommendations*. Retrieved May 22, 2019 from http://aaaea.org/Al-muhandes/2008/February/open_src_dev_model.htm

History. (2019). Retrieved July 18, 2019, from Scilab Official Website: https://www.scilab.org/about/company/history

Jamovi. (2020). Retrieved July 18, 2019, from Jamovi Official Website: https://www.jamovi.org/

Jupyter. (2019). Retrieved July 18, 2019, from Jupyter Official Website: https://jupyter.org/

Lakens, D. (2017, March 14). *Equivalence testing in Jamovi*. Retrieved July 18, 2019, from *The Statistician*: http://daniellakens.blogspot.com/2017/03/equivalence-testing-in-jamovi.html

Linux. (2011). *Understanding the Linux Open Source Development Model*. Retrieved May 22, 2019 from http://www.ibrahimatlinux.com/uploads/6/3/9/7/6397792/00.pdf

Love, J., Selker, R., Verhagen, J., Marsman, M., Gronau, Q. F., Jamil, T., . . . Rouder, J. N. (2015, March). Software to Sharpen Your Stats. *Association for Psychological Science Observer*, *28*(3). Retrieved 7 18, 2019, from https://www.psychologicalscience.org/observer/bayes-or-bust-with-new-softwares

Mandal, S., Kandar, S., & Ray, P. (2011). Open Incremental Model: An Open Source Software Development Life Cycle Model 'OSDLC'. *International Journal of Computers and Applications*, *212*(1), 2473–3327.

Miao, B. (2017, April 26). *Octave Introduction and Study*. Retrieved July 18, 2019, from CSDN: https://blog.csdn.net/imbenben/article/details/70768980

Mixon, J. W. Jr, & Smith, R. J. (2006). Teaching undergraduate econometrics with GRETL. *Journal of Applied Econometrics*, *21*(7), 1103–1107. doi:10.1002/jae.927

Muenchen, B. (2018, February 13). *Jamovi for R: Easy but Controversial*. Retrieved July 18, 2019, from r4stats: https://r4stats.com/2018/02/13/jamovi-for-r-easy-but-controversial/

Octave, G. N. U. (2020). Retrieved July 18, 2019, from GNU: https://www.gnu.org/software/octave/about.html

Octave Forge. (2019). Retrieved July 18, 2019, from Source Forge: https://octave.sourceforge.io/

Open Analytics, N. V. (2020). *IDE for Data Science*. Retrieved October 6, 2020, from Get Architect: https://www.getarchitect.io/

Open Source Initiative. (2007). *The Open Source Definition*. Retrieved on May 22, 2019 from https://opensource.org/osd

Perez, F. (2014, July 8). *Project Jupyter*. Retrieved July 18, 2019, from Speakerdeck: https://speakerdeck.com/fperez/project-jupyter

Python Software Foundation. (2020). *History and License*. Retrieved October 6, 2020, from Python Official Website: https://docs.python.org/3/license.html

Regression, G. N. U., & the Econometrics and Time-series Library. (2019, July 2). Retrieved July 18, 2019, from GRETL Official Website: http://gretl.sourceforge.net/

RStudio. (2019). Retrieved July 18, 2019, from RStudio Official Website: https://www.rstudio.com/products/rstudio/

Sani, M., & Kaur, K. (2014). A review of open source software development life cycle models. *International Journal of Software Engineering and Its Applications*, *8*(3), 417–434. Retrieved May 22, 2019, from https://www.researchgate.net/publication/289328296_A_review_of_open_source_software_development_life_cycle_models

Scilab. (2019). Retrieved July 18, 2019, from Predictive Analytics Today: https://www.predictiveanalyticstoday.com/scilab/

Segall, R. S. (2020a). What is Open Source Software (OSSS) and What is Big Data? In R. S. Segall & G. Niu (Eds.), *Open Source Software for Statistical Analysis of Big Data* (pp. 1–49). IGI Global.

Segall, R. S. (2020b). Open Source Software (OSS) for Big Data. In *Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities* (pp. 50–72). IGI Global. doi:10.4018/978-1-7998-2768-9.ch002

Segall, R. S., & Niu, G. (2020). *Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities*. IGI Global. doi:10.4018/978-1-7998-2768-9

Shogun. (2020, September 30). Retrieved from https://www.shogun-toolbox.org/: https://www.shogun-toolbox.org/examples/latest/index.html

Sieb, C., Meinl, T., & Berthold, M. R. (2007). Parallel and Distributed Data Pipelining with KNIME. *The Mediterranean Journal of Computers and Networks*, *3*(2), 43–51.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V., & Hassabis, D. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi:10.1038/nature16961 PMID:26819042

St. Laurant, A. (2004). *Understanding Open Source and Free Software Licensing*., O'Reilly Media, Inc. Retrieved September 10, 2020 from https://people.debian.org/~dktrkranz/legal/Understanding%20Open%20Source%20and%20Free%20Software%20Licensing.pdf

St. Laurent, A. (2008). *Understanding Open Source and Free Software Licensing*. O'Reilly Media.

Statistics, S. O. F. A. (2019). Retrieved July 18, 2019, from *Predictive Analytics Today*: https://www.predictiveanalyticstoday.com/sofa-statistics/

Strata, O. (2013, February 26). *Coudera Accelerates Platform for Big Data with New Enterprises-Required Advancements*. Retrieved from Cloudera: https://www.cloudera.com/about/news-and-blogs/press-releases/2013-02-26-cloudera-accelerates-platform-for-big-data-with-new-enterprise-required-advancements.html

Taft, D. (2010, October 14). *JetBrains Strikes Python Developers with PyCharm 1.0 IDE*. Retrieved July 18, 2019, from *eWeek*: https://www.eweek.com/development/jetbrains-strikes-python-developers-with-pycharm-1.0-ide

TensorFlow. (2020, September 30). Retrieved from https://www.tensorflow.org/

The JASP Team. (2018). *What does JASP stand for?* Retrieved October 6, 2020, from JASP Official Website: https://jasp-stats.org/faq/what-does-jasp-stand-for/

The Microsoft Cognitive Toolkit. (2017, January 22). Retrieved from Microsoft Documentations: https://docs.microsoft.com/en-us/cognitive-toolkit/

The R Foundation. (2020). *Contributors*. Retrieved October 6, 2020, from The R Project for Statistical Computing: https://www.r-project.org/contributors.html

Thomas, R. (2017, January 3). *Big deep learning news: Google Tensorflow chooses Keras*. Retrieved from Fast AI: https://www.fast.ai/2017/01/03/keras/

Tiwari, A., & Sekhar, A. K. (2007, October). Workflow based framework for life science informatics. *Computational Biology and Chemistry*, *31*(5-6), 305–319. doi:10.1016/j.compbiolchem.2007.08.009 PMID:17931570

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., & Morey, R. et al. (2018, February). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. doi:10.3758/s13423-017-1323-7 PMID:28685272

Wikipedia. (2019a). *Free and open source software*. https://en.wikipedia.org/wiki/Free_and_open-source_software

Wikipedia. (2019b). *Freeware*. https://en.wikipedia.org/wiki/Freeware

Wikipedia. (2019c). *Open Source Software*. https://en.wikipedia.org/wiki/Open-source_software

Wu, Z., Zhao, Z., & Niu, G. (2020). Introduction to the Popular Open Source Statistical Software. In R. S. Segall & G. Niu (Eds.), *Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities* (pp. 73–110). IGI Global. doi:10.4018/978-1-7998-2768-9.ch003

Yalta, A., & Yalta, A. (2007). GRETL 1.6.0 and its numerical accuracy. *Journal of Applied Econometrics*, *22*(4), 849–854. doi:10.1002/jae.946

APPENDIX 1

Table 4. List of Open Source Statistical Software (OSSS) and URLs

Open Source Statistical Software	URL
R	https://www.r-project.org/
RStudio	https://rstudio.com/
Architect	https://www.getarchitect.io/
Jamovi	https://www.jamovi.org/
Python	https://www.python.org/
PyCharm	https://www.jetbrains.com/pycharm/
Spyder	https://www.spyder-ide.org/
Jupyter	https://jupyter.org/
JASP	https://jasp-stats.org/
PSPP	https://www.gnu.org/software/pspp/
GRETL	http://gretl.sourceforge.net/
SOFA	https://www.sofastatistics.com/home.php
Octave	https://www.gnu.org/software/octave/index
KNIME	https://www.knime.com/
Scilab	https://www.scilab.org/

APPENDIX 2

Table 5. List of AI and machine learning libraries and systems and URLs

AI and Machine Learning Libraries and Systems	URL
TensorFlow	https://www.tensorflow.org/
Theano	https://pypi.org/project/Theano/
Keras	https://keras.io/
Shogun	https://www.shogun-toolbox.org/
Microsoft Cognitive Toolkit	https://docs.microsoft.com/en-us/cognitive-toolkit/
Torch	http://torch.ch/
CloudEra	https://www.cloudera.com/

Gao Niu is an Assistant Professor in Actuarial Science and Program Coordinator of Actuarial Math Program at Bryant University. He also serves as a Faculty Consultant of the Janet & Mark L Goldenson Center for Actuarial Research at the University of Connecticut. He has a doctorate in actuarial science from the University of Connecticut, is an Associate of the Casualty Actuarial Society and a Member of the American Academy of Actuaries. Dr. Niu has years of experience in academic actuarial research and consulting practice. His research area includes but not limited to the following: big data analytics application in insurance industry, property and casualty insurance practice, predictive modeling, agent based modeling, financial planning, life insurance and health insurance pricing, reserving and data mining.

Richard Segall holds a Bachelor of Science and Master of Science in Mathematics as well as a Master of Science in Operations Research and Statistics from Rensselaer Polytechnic Institute in Troy, New York. He also holds a PhD in Operations Research form University of Massachusetts at Amherst. Dr. Richard S. Segall is a Professor of Computer & Information Technology in the Neil Griffin College of Business at Arkansas State University in Jonesboro, AR and also teaches in the Master of Engineering Management (MEM) Program in the College of Engineering and Computer Science. He is also Affiliated Faculty at the University of Arkansas at Little Rock (UALR) where he serves on thesis committees. He has served on the faculty of Texas Tech University, University of Louisville, University of New Hampshire, University of Massachusetts-Lowell, and West Virginia University. His research interests include data mining, text mining, web mining, database management, Big Data, and mathematical modeling. Dr. Segall's publications have appeared in numerous journals including International Journal of Information Technology and Decision Making (IJITDM), International Journal of Information and Decision Sciences (IJIDS), International Journal of Fog Computing (IJFC), Applied Mathematical Modelling (AMM), Kybernetes: The International Journal of Cybernetics, Systems and Management Sciences, Journal of the Operational Research Society (JORS) and Journal of Systemics, Cybernetics and Informatics (JSCI). He has published book chapters in Encyclopedia of Data Warehousing and Mining, Handbook of Computational Intelligence in Manufacturing and Production Management, Handbook of Research on Text and Web Mining Technologies, Encyclopedia of Information Science & Technology, and Encyclopedia of Business Analytics & Optimization. Dr. Segall was a member of the Arkansas Center for Plant-Powered-Production (P3), is currently a member of the Center for New Boundary Thinking (CNBT) at Arkansas State University, and on the Editorial Board of the International Journal of Data Mining, Modelling and Management (IJDMMM) and International Journal of Data Science (IJDS), and served as Local Arrangements Chair of the MidSouth Computational Biology & Bioinformatics Society (MCBIOS) Conference that was hosted at Arkansas State University. His research has been funded by National Research Council (NRC), U.S. Air Force (USAF), National Aeronautical and Space Administration (NASA), Arkansas Biosciences Institute (ABI), and Arkansas Science & Technology Authority (ASTA). He is recipient of several Session Best Paper awards at World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI) conferences. He is co-editor of four books published by IGI Global: Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities, 2-Volume Handbook of Big Data Storage and Visualization Techniques in 2018 that was selected as IGI Global "Featured Title in Computer Science & Information Technology for 2018-2019", Research and Applications in Global Supercomputing in 2015, and Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications in 2011. Dr. Segall is recipient of Arkansas State University, Neil Griffin College of Business Faculty Award for Excellence in Research in 2019 and 2015, and University Faculty Award for Excellence in Scholarship in 2020.

Zichen Zhao is a graduate student of biostatistics major at Yale University. He has a Bachelor's degree in Actuarial Science from the University of Connecticut. He has worked at multiple investment and insurance companies such as Goldman Sachs, Sunshine Insurance Group. His current research includes but not limited to actuarial science, big data analytics, predictive modeling, machine learning, and statistical modeling.

Zhijian Wu is a graduate student of mathematics major at New York University. He has a Bachelor's degree in Actuarial Science from the University of Connecticut. He worked at multiple insurance and investment companies during his early career. His current research includes but not limited to actuarial science, big data analytics, predictive modeling, machine learning, statistical modeling, and spatial-temporal analytics.