

Volume 9, Issue 2

October 2016

ISSN: 1946-1836

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

In this issue:

- 4. A Comparison of Open Source Tools for Data Science**
Hayden Wimmer, Georgia Southern University
Loreen M. Powell, Bloomsburg University

- 13. Exploratory Study of Effects of eLearning System Acceptance on Learning Outcomes**
Biswadip Ghosh, Metropolitan State University of Denver

- 24. Leakage of Geolocation Data by Mobile Ad Networks**
Christopher Snow, Pace University
Darren Hayes, Pace University
Catherine Dwyer, Pace University

The **Journal of Information Systems Applied Research (JISAR)** is a double-blind peer-reviewed academic journal published by **ISCAP**, Information Systems and Computing Academic Professionals. Publishing frequency is currently quarterly. The first date of publication was December 1, 2008.

JISAR is published online (<http://jisar.org>) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (<http://conisar.org>)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

2016 AITP Education Special Interest Group (EDSIG) Board of Directors

Scott Hunsinger
Appalachian State Univ
President

Leslie J. Waguespack Jr
Bentley University
Vice President

Wendy Ceccucci
Quinnipiac University
President – 2013-2014

Nita Brooks
Middle Tennessee State Univ
Director

Meg Fryling
Siena College
Director

Tom Janicki
U North Carolina Wilmington
Director

Muhammed Miah
Southern Univ New Orleans
Director

James Pomykalski
Susquehanna University
Director

Anthony Serapiglia
St. Vincent College
Director

Jason Sharp
Tarleton State University
Director

Peter Wu
Robert Morris University
Director

Lee Freeman
Univ. of Michigan - Dearborn
JISE Editor

Copyright © 2016 by the Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Scott Hunsinger, Editor, editor@jisar.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

JISAR Editorial Board

Ronald Babin
Ryerson University

Teko Jan Bekkering
Northeastern State University

Gerald DeHondt II

Meg Fryling
Siena College

Biswadip Ghosh
Metropolitan State University of Denver

Audrey Griffin
Chowan University

Muhammed Miah
Southern University at New Orleans

Monica Parzinger
St. Mary's University

Alan Peslak
Penn State University

Doncho Petkov
Eastern Connecticut State University

Bryan Reinicke
Rochester Institute of Technology

Karthikeyan Umapathy
University of North Florida

Leslie Waguespack
Bentley University

Peter Wu
Robert Morris University

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
Dept. of Business Education and 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 model's 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, & Pflieger, 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 Demšar (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.

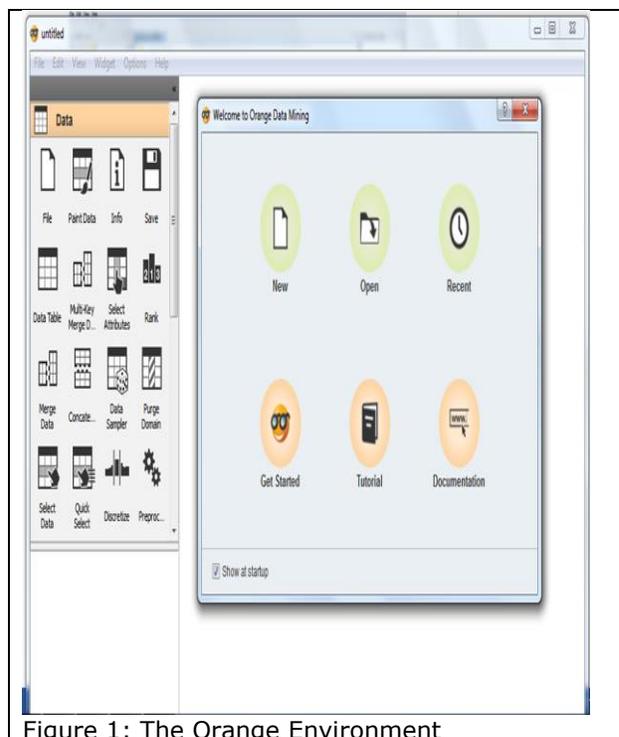


Figure 1: The Orange Environment

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.

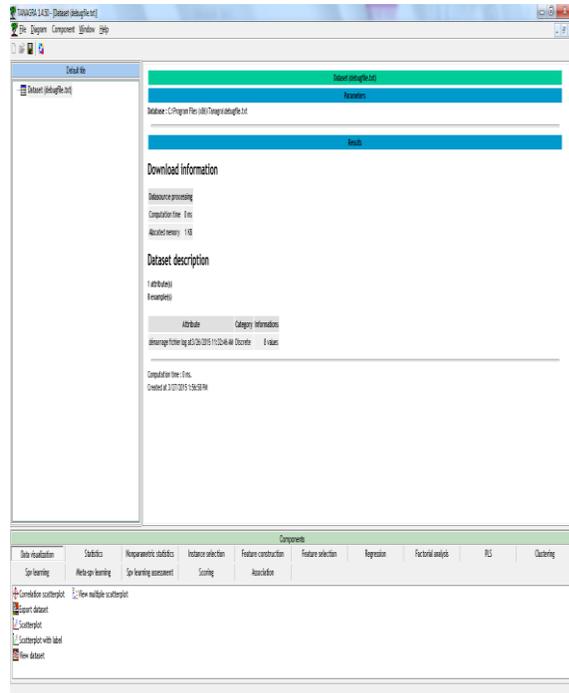


Figure 2: The Tanagra Environment

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.

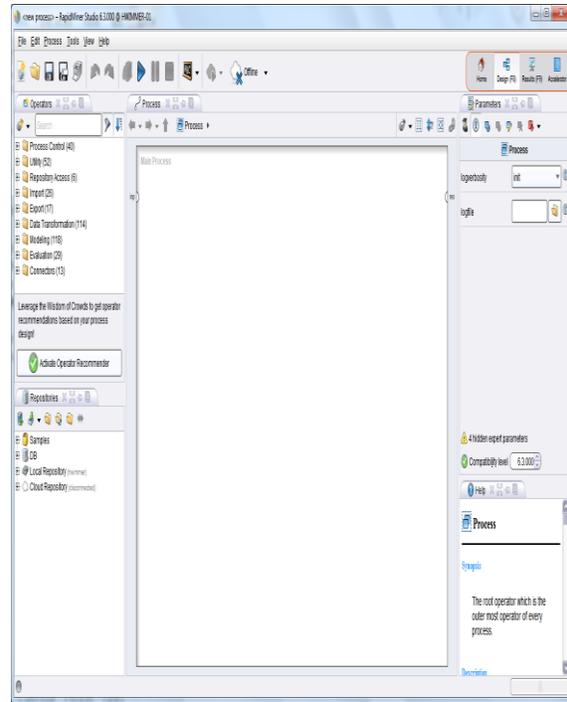


Figure 3: The Rapid Miner Environment

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.

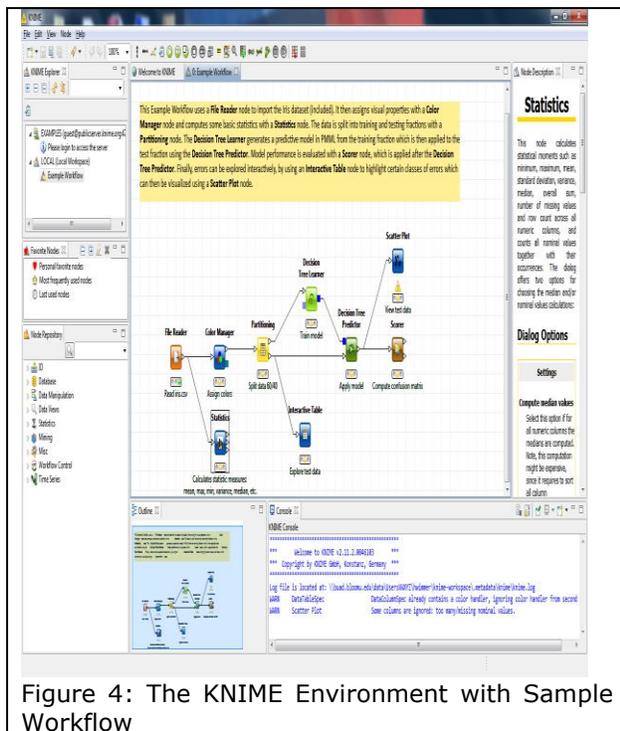


Figure 4: The KNIME Environment with Sample Workflow

R
R is a free and open source package for statistics and graphing. R is traditionally command line; however, there are many freely 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 R Hadoop. R Hadoop 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.

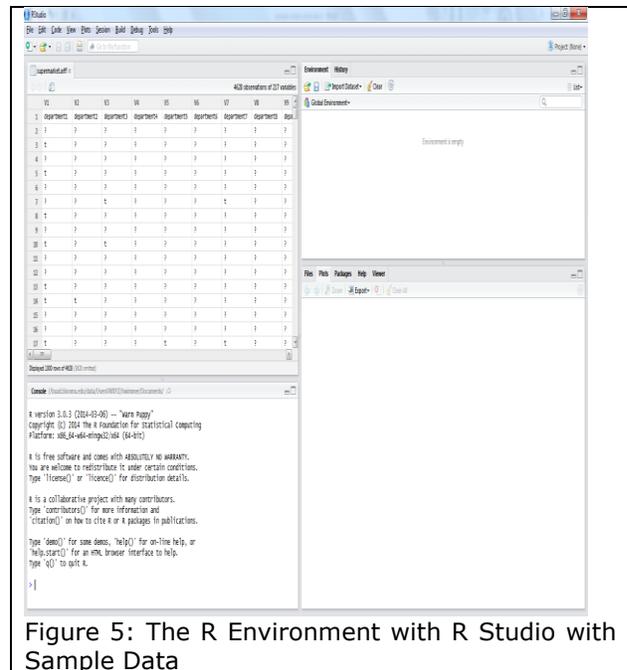


Figure 5: The R Environment with R Studio with Sample Data

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.

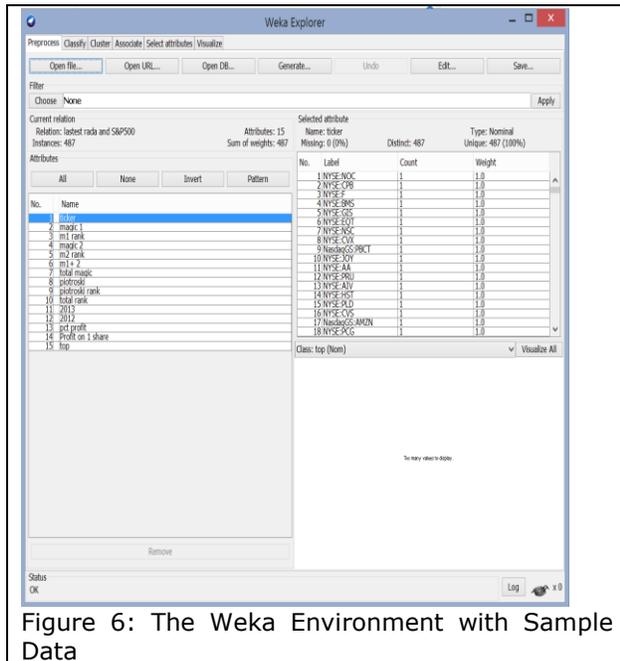


Figure 6: The Weka Environment with Sample Data

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 tool 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.
- Goppta, 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/gdata/mine/>
- 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.

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2015 Meritorious Paper. The acceptance rate is typically 15% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2015.

Exploratory Study of Effects of eLearning System Acceptance on Learning Outcomes

Biswadip Ghosh
bghosh@msudenver.edu
Computer Information Systems
Metropolitan State University of Denver
Denver, Colorado 80217, USA

Abstract

End-user learning is an important element of Information Systems (IS) projects. End-user learning of software applications can constitute roughly 5% to 50% of project budgets. To lower costs and make learning more convenient for the end-users, organizations are largely utilizing online systems for the electronic delivery of such learning programs, referred to as Technology Mediated Learning (TML). In this learning format, before the end-users are able to immerse themselves in the actual learning program, they are first required to adopt and use an online learning system. Currently published IS research has two mature streams of publications: one stream focused on models of technology acceptance and usage that is based on the TAM (Technology Acceptance Model) model and a second stream based on the TML framework consisting of learning content, structures and outcomes. This research study aims to build and validate an empirical model extended from the TML framework with constructs from TAM. This extended model is validated and relationships are tested using survey data collected from an e-learning system used for teaching spreadsheet and database management software applications. The results indicate that the acceptance and usage of the e-learning system and the learning outcomes of mastering office productivity applications is related to individual characteristics and facilitating conditions that boost perceived ease of use and perceived usefulness. The results of this study have implications for both the TAM and TML research streams and also the design and use of e-learning for software applications by IS practitioners.

Keywords: Technology Mediated Learning; Learning Outcomes and Technology Acceptance Model.

1. INTRODUCTION

End-User learning is one of the most pervasive methods for developing human resources within modern organizations. Majority of end-user learning deals with teaching end-users how to use computer applications and gain tool operational knowledge to do their assigned jobs in the organization. There are three targeted goals of most end-user learning programs (Gupta, et.al., 2010): (1) skill-based goals (tool procedural) that target the user's ability to use the system, (2) cognitive goals (tool conceptual or business procedural) that focus on the use of the system to solve business problems and (3) meta-cognitive goals that focus on building the individual's belief regarding their own abilities

with the computer applications. To lower costs and make learning more convenient and schedule friendly for employees, the use of online learning systems for the electronic delivery of end-user learning has become popular (ASTD, 2013). Recent reports suggest that upwards of 40-50% of end-user learning is conducted through technology mediated learning (TML) systems (ASTD, 2013). Technology-mediated learning environments (TML environments) are environments "in which the learner's interactions with the learning content (readings, assignments, exercises), peers, and the program instructions are mediated through advanced information technology" (Alavi and Leidner, 2001, p.2). In addition to commercial organizations, many universities also leverage

technology based learning systems to teach students popular software and commercial systems such as enterprise resource planning systems (e.g. SAP) and office productivity software, such as spreadsheet software and personal database management software.

However, there is continuing frustration with technology mediated learning as the success of these e-learning programs is highly dependent on the student's acceptance and correct use of the system to manage their learning process. As the variety of e-learning systems grow, identifying the critical factors related to users' perceptions and acceptance of e-learning technology continues to be an important issue (Mun and Hwang, 2003). To this end, studies of user perceptions of these learning systems and understanding factors supporting effective use of these systems (Mun and Hwang, 2003) have become increasingly essential to improve awareness of acceptance and utilization (Lau and Woods, 2008). The currently published research has primarily focused on finding answers to the adoption problem by investigating individuals' decisions on whether or not to adopt e-learning systems that appear to promise substantial benefits (McFarland and Hamilton, 2006; Venkatesh, et.al., 2003).

Some studies have applied the Technology Acceptance Model (TAM) to understand effects of the pedagogical design of such e-learning systems. The focus has been on understanding the impact of learning system features such as learning activities, security, information and service quality, interactivity and responsiveness, learner control and the ability to self-organize their learning on the user's acceptance of those systems (Selim, 2003; Pituch and Lee, 2006; Roca and Gagne, 2008; Sun, et.al., 2008). However, the ultimate effectiveness of any e-learning system is not its utilization, but the learning outcomes it produces. Although e-learning research has attracted much attention over the last decade, suitable frameworks to assess e-learning program outcomes have yet to emerge despite a variety of models and variables characterized in these studies (McGill & Klobas, 2009). This research gap calls for more innovative and comprehensive approaches to fully understand the factors affecting e-learning program acceptance and program outcomes and the need for validated measurement models of the learning outcomes of e-learning systems.

Research Goals

The focus of this research study is to answer the question "Does the level of acceptance and use

of features and capabilities of an online learning system impact learning outcomes?" To answer this question, the paper extends the TML framework with constructs from the TAM model – perceived ease of use and perceived usefulness and measures their impacts on learning appropriation and outcomes.

The goals of this study are:

1. To develop and empirically validate an extended TML research model that also includes the users' learning system usage behavior and the facilitating conditions supporting such usage.
2. To measure the impacts of the usage behavior and facilitating conditions on the users' learning outcomes.

2.BACKGROUND THEORY

With the popularity of TML adoption and an increase in cloud based courseware, there is vast diversity in these online learning systems, which employ various platforms and software architectures that pose a variety of challenges (Bensch and Rager, 2012). Information technology deployed in typical learning programs is used as a primary structural element in the learning process (e.g. simulations or exercises that are part of the learning process) or as a secondary tool in the learning process (e.g. computer based tests and quizzes). However, the actual use of the features and capabilities of an online learning system have been found to differ across groups of users (Bekkering and Hutchison, 2009). Individual differences play a role in what features of these systems are used and how the systems can impact each end-users' learning process and outcome (Gupta, Bostrom and Anson, 2010). The current research stream of IS end-user learning has studied the impact of the above learning structures on different learning outcomes along with various confounding factors such as the individual's learning style, their motivation to participate and their interest in the learning content (Bostrom, et.al., 1990; Nogura and Watson, 2004).

A comprehensive TML research framework is elaborated in Gupta and Bostrom (2009). In the TML framework, the learning structures (or scaffolds) support the delivery of the learning content, such as the rules, resources and methods, the level of detail in the instructions given to participants, the guidance provided by the facilitator and the nature of the facilities and equipment used in the learning session. While the TML model incorporates technology as a

structural element of learning delivery, it does not take into account the usage behavior of the specific capabilities of the learning platform by the individual users. Individual differences can impact learning outcomes by generating a different mental response to the learning content and influencing their interactions with the learning delivery structures (Bekkering and Hutchison, 2009). Learning style of the user plays an important role in the user's conformance to the learning tasks embedded in the online learning system (Bohlen and Ferratt, 1997). For example, abstract learners perform better than users with concrete learning styles in online technology based learning. The user's motivation and attitudes also have been found to influence learning performance in the TML context (Szajna, B. and Mackay, J.M., 1995; Yi and Davis, 2003). Such results support the need to merge additional constructs into the TML framework to represent the user's technology acceptance and usage behavior.

The technology acceptance model (TAM) is one of the most widely used models used in Information systems research to study the adoption and usage intentions of users of systems. TAM's roots are from the theory of planned behavior and the theory of reasoned action (Ajzen, 1988; Ajzen, 1991). TAM was developed by Davis (1989) to explain the determinants of the intention to use computer systems. Two key components that were used in the original model are – perceived usefulness and the perceived ease of use of any technology. Perceived usefulness is referred to as the "degree to which a person believes that using a particular system will enhance their performance" (in a job or activity). The perceived ease of use defines the "degree to which a person believes that using a particular system would be free of effort". It is posited in the original TAM that actual intention to use a system will positively depend on both of these constructs. TAM has been validated over a wide category of information systems and user domains and proven to give reliable and valid results (Venkatesh, et. al., 2003). The simplicity and compactness of TAM provides the necessary constructs for this research study to extend the TML framework and develop a model to build a measure of learning outcomes.

Prior studies have applied TAM to examine the acceptance and effectiveness of e-learning system use (eg, Lau and Woods, 2008). In spite of its popularity and considerable empirical support, e-Learning researchers have also extended TAM with other socio-technical

constructs, such as computer self-efficacy, enjoyment and modeled their impact on intention to use through the TAM variables (Agarwal and Karahanna, 2000; Davis, 1993).

Researchers have extended TAM with other socio-technical constructs, such as computer self-efficacy, enjoyment and modeled their impact on intention to use through the TAM variables (Agarwal and Karahanna, 2000; Davis, 1993). Researchers have introduced subjective norm (SN), such as social influence into the Technology Acceptance Model (TAM) for its application to real world organizations (Madon, 2000; Malhotra and Galletta, 1999). The construct of social influence is operationalized in terms of certain processes (internalization, identification and compliance) and field data provided evidence of the reliability and validity of the proposed constructs, factor structures and measures. Musa, Meso and Mbarika (2005) added external variables of Accessibility and Exposure to Technologies (AET) and Perceptions of Socio-economic Environment (PSEE) to extended TAM in a study of technology adoption in Sub Saharan Africa.

The original TAM model (Davis, 1989) provides a suitable and parsimonious framework for this research study to extended and develop a model to incorporate measures of the perceived usefulness and ease of use and usage of the eLearning System features and their impact on the constructs from the TML model.

3. RESEARCH MODEL

The research constructs are defined in the following subsections. The dependent variable in the model is Learning Outcomes (LO). The independent variables are the TML system (modeled as a formative second order construct consisting of learning system features, content and structures. The Individual characteristics (IC) and Facilitating Conditions (FC) are derived from the TML framework and are also independent variables in the model. The perceived ease of use, the perceived usefulness and the usage are constructs adopted from the TAM model and used in this research.

The research model is displayed in Figure 1.

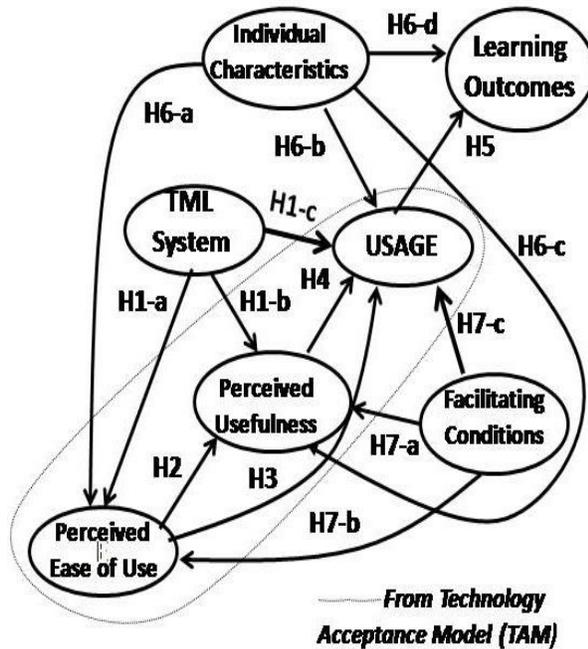


Figure 1. Research Model extended from TML framework

Learning Outcomes

Learning outcomes (LO) focus on the mental awareness and judgment of the end-user and the levels of application of acquired knowledge towards operating business functions (Gupta, et.al, 2010). The learning outcomes is a formative construct that consists of three types of outcomes – skill based, cognitive and meta-cognitive. There are three targeted goals of most learning programs: (1) skill-based goals (tool procedural) that target the user’s ability to use the software, (2) cognitive goals (tool conceptual or business procedural) that focus on the use of the system to solve business problems that are outside of the learning program and (3) meta-cognitive goals that focus on building the individual’s belief regarding their own abilities with the system (Gupta, et.al, 2010). Skill based goals of learning focus on collecting procedural know how or the nuts and bolts of using the system, such as spreadsheet or database management software (Gupta, et.al., 2010). These include creating a new sheet, building formulae and utilizing various features of the application. Cognitive training goals focus on the metal awareness and judgment of the user to transfer the learning to new situations, such as applying the software application to solve a new problem different from what was used in the learning. Finally, meta-cognitive goals focus on enhancing the learner’s ability to understand his/her own

learning and information processing procedure and confidence (Gupta and Bostrom, 2010).

TML System

As the use of TML in learning programs intensifies, the need to list the features of such applications as a component of the overall learning system is more important. System features mentioned in the research stream refer to responsiveness and quality (Lee, Yoon & Lee, 2009), feedback and facilitation of communications about assigned instructional work (Putuch & Lee, 2006), flexibility, autonomy and user control of the learning process and steps (Piccoli, Ahmad and Ives, 2001).

The TML system is characterized by the user features that establish learning structures to support the delivery of learning content. Learning content (LC) refers to instructional methods that encourage students to accomplish learning goals. These allow end-users to fill gaps in their understanding and builds skills (skill focus) and knowledge about how they can use the system to improve their productivity (cognitive focus). “Soft skills” are also developed that allow members to learn collective beliefs and norms that help them develop confidence and knowledge in solving future business problems. Learning structures (LS) refer to the scaffolds that support the delivery of the learning content. Also referred to as appropriation support (Gupta, et.al, 2010), they include the rules, resources and methods that support the elements of the collaborative learning session. For this research study, the learning structures include level of detail in the instructions given to participants, the guidance provided by the facilitator and the nature of the facilities and equipment used in the learning session.

Individual Characteristics

People prefer learning methods based on their specific learning styles (Nogura and Watson, 2004). Individual differences influence the formation of mental models, which effects the learning process. “States” are general influences on performance that vary over time and include temporal factors such as motivation level and interest level (Bostrom, et.al., 1990). “Traits” are static aspects of information processing affecting a broad range of outcomes. Cognitive traits refer to learning styles such as a preference for procedural or abstract knowledge and an exploratory or reflective approach to instructional content delivery format (Bostrom, et.al., 1990; Nogura and Watson, 2004). For this research study, the Individual

characteristics (IC) variable is measured using motivation and interest as states and individual learning style as traits. Both intrinsic motivation and extrinsic motivation influences the learner's state and is measured in the survey.

Facilitating Conditions

Facilitating conditions are environmental factors that refer to the users' perceptions of resources and support to use the technology (Venkatesh, et. al., 2008). Such factors support the individual's belief that an organizational and technical infrastructure exists to support use of the system. In the context of a learning system, facilitating conditions include resources, accessibility, compatibility with other systems, infrastructure quality and support (McGill and Klobas, 2009; Venkatesh, et.al., 2008).

Perceived Usefulness & Ease of Use

Two key components were used in the original TAM model – perceived usefulness and the perceived ease of use of any technology innovation. The UTAUT model includes two components – Performance Expectancy and Effort Expectancy (Venkatesh, Thong and Xu, 2012). Performance Expectancy (PE) is referred to as the "degree to which a person believes that using a particular system will enhance their performance" (in a job or activity). Effort Expectancy (EE) defines the "degree to which a person believes that using a particular system would be free of effort". It is posited that actual usage of a system will positively depend on both of these constructs (Venkatesh, et. al., 2003).

Usage

Actual usage behavior is captured in the research model as Usage. Both behavioural intentions and actual usage behavior to use the technology are part of the original TAM and the UTAUT models (Venkatesh, et. al., 2003). While behavioral intentions imply the plans and intentions to use the system, actual usage behavior refers to the duration, frequency and intensity of the use of the system (Venkatesh, et.al., 2008).

4. RESEARCH HYPOTHESES

The research hypotheses are listed below. Given the exploratory nature of this study, rather than be parsimonious, the emphasis is to model and test various possible relationships across constructs in the TML and TAM models.

TML System Features Support Usage

Based on the review of previous research studies, we find that e-learning system features such as quality, information quality, interface

presentation style influences the perceived usefulness of the system to the student (Seddon, 1997). The perceived usefulness of an e-Learning system is related to the users' perceptions regarding the potential benefits of the system in delivering the learning content and teaching the application and whether the learning structures imposed by the system fit the learners' preferences. Likewise, the perceived ease of use of a system refers to the users' belief that using the system will be free of effort (Venkatesh, et.al., 2003). In the context of e-learning, ease of use includes the notion that the system will not require a great deal of extra effort to operate or impose any additional cognitive burden during the learning process (Lin, 2009).

The features of the e-learning system can help reduce the cognitive burden on a student by making the learning content more accessible and providing reminders and quicker feedback to pace the student learning activities. The e-learning systems support the student's learning in several ways such as by providing reminders about assignments that are due, providing feedback on submitted assignments, displaying performance summaries and providing hints and demonstrations. Certain features of the e-learning system such as those that enable the student to exercise control over the learning pace, sequence and content delivery can help lower a student's resistance towards using the e-learning system (Picolli, et.al., 2001).

H1-a: TML system features have a positive effect on perceived usefulness. That is greater the perceived TML system feature value, the higher the perceived usefulness.

System features that have high ease of use encourage greater usage, which sustains a higher sense of system usefulness.

H1-b: TML system features have an positive effect on perceived ease of use. That is greater the perceived TML system feature value, the higher the perceived ease of use.

Using an e-learning system proves to be effective if it increases the students' efficiency by reducing their time and cost of learning and/or improving their performance/score. The greater the perceived value of the e-Learning system features, the greater the usage of the system. Therefore, we have:

H1-c: TML system features have a positive effect on System Usage. That is greater the perceived TML system feature value, the higher the system usage.

H5: The higher the Usage of the e-learning System the higher the Learning Outcomes.

TAM Framework

These three hypotheses come directly from the TAM model (Davis, 1989) and can be stated as below. These three hypotheses are also included in this study and will be tested in the context of e-Learning in this study.

H2: Perceived ease of use of the TML system have a positive effect on the perceived usefulness of the TML system.

H3: Perceived ease of use of the TML system have a positive effect on the usage of the TML system.

H4: Perceived usefulness of the TML system have a positive effect on the usage of the TML system.

IC Supports Usage & Outcomes

Individual characteristics (IC) represent the cognitive aspects of human activities that are often referred to as "learning ability" and influence learning outcomes directly through the formation of mental models or indirectly through interactions with the e-learning system (Olfman et al. 2000). Motivation theory suggests that individual behavior is determined by two fundamental types of motivation: extrinsic (utilitarian) motivation and intrinsic (hedonistic) motivation (Ryan and Deci, 2000). Motivation theory has been used often to understand individuals' e-learning use and learning behavior (Igbaria, et.al., 1996; Tharenou, 2001). The results of their empirical study suggested that computer-based training is more effective than lecture-based training except for assimilators, who appear to learn equally well under either method (Sein et al., 1989).

H6-a: Individual Characteristics have a positive effect on perceived ease of use.

H6-b: Individual Characteristics have a positive effect on e-learning system usage.

H6-c: Individual Characteristics have a positive effect on perceived usefulness.

Individual differences influence the formation of mental models, which represent the outcomes of

the training process (Gupta, et.al., 2010). "States" are general influences on performance that vary over time and include temporal factors such as motivation level and interest level while "traits" are static aspects of information processing affecting a broad range of outcomes over time (Bostrom, et.al., 1990). Therefore, we have

H6-d: Individual Characteristics have a positive effect on Learning Outcomes.

Facilitating Conditions Support Usage

Facilitating conditions include objective factors in the environment that help to make the act of using the e-learning system easier to do (Venkatesh, et.al., 2003). An important influence on the user's usage of the e-learning system is the support provided (Gupta, et.al., 2010). These include technical support, instructor guidance, specialized computer resources and ready to use labs and assistance with system usage. The focus of support is to influence the interaction of the learners with the learning content and methods structures. In fact, the effect of facilitating conditions increases with experience as experienced users of technology find multiple avenues for help and support and certain groups of users attach more importance to receiving help and assistance (Venkatesh, et.al., 2003). The need for support may gradually fade as learners become more independent, confident and competent with the e-learning system.

H7-a: Facilitating conditions have a positive effect on perceived usefulness. That is higher the perception of the facilitating conditions, higher the perceived usefulness.

H7-b: Facilitating conditions have a positive effect on perceived ease of use. That is higher the perception of facilitating conditions, higher the perceived ease of use.

H7-c: Facilitating Conditions have a positive effect on TML system usage.

5. METHODOLOGY

A survey was developed to measure the research constructs. The survey consists of multiple items for each construct and uses a 5 point Likert scale (1 being strongly disagree and 5 being strongly agree) to measure user responses to each item. The survey is included in the Appendix. Two of the seven constructs –

Learning System Features (TML System) and Individual Characteristics (IC) are formative constructs. The data collection approach consisted of surveying business school students, who used an online e-learning system, "MyITLab" (www.myitlab.com) to learn to use spreadsheet and database software applications.

MyITLab is a feature rich learning application that allows users to complete a variety of simulated tutorial exercises and case studies with Microsoft excel and access software packages. The system is accessed through a web browser and has no client installation requirements. While some parts of the system can be cumbersome and requires extensive scaffolding, such as initial registration, login and a properly configured browser for accessibility, yet the major benefits of using the system are quick feedback on assignments, interactive help on various procedural aspects of Excel and Access software and organization of the learning process.

There were 10 chapters of assignments (5 chapters of Excel and 5 chapters of Access) that covered features of Excel and Access software. Each week's assignment consisted of tutorial exercises that were executed inside a simulated environment representing the particular application features of interest for that week. The tutorials typically consisted of 20-30 activities each week and each activity was individually executed and submitted for grading. Hints for help was available for each activity in three forms – as a voice only clip describing the step by step instructions, as text-based instructions that appeared on a status text box and a computer animation showing exactly how the activity was to be performed. Thus the tutorials supported the tool-procedural skill based learning. Each week a case study was assigned that required the students to prepare an Excel or Access document to solve a business problem and upload the document into MyITLab for auto grading and feedback. This was the applied portion of the learning, which addressed business procedural outcomes.

A pilot survey was conducted to ascertain the content validity and clarity of the survey items. The final survey was completed with 200 users of MyITLab and reliability and validity of the survey instruments has been calculated (Table 2). A total of 139 completed surveys were collected for a response rate of 70%. The demographics of the respondents are presented in Table 1. The students were mostly in their 2nd or 3rd year of college and had had some

prior experience with using Excel (3.17 years on average), but minimal experience with Access (1.14 years on average). The students used the MyITLab system on average for 3.57 hours a week for the 10 weeks of the semester. Most favored learning styles identified by the students were learning by doing and least favored style was learning by feeling. Note that some users selected multiple preferred learning styles.

Table 1: Demographic Variables (n = 139)

Variable	Min	Max	Mean	S.D.
Years of College Edu (years)	2	6	2.56	1.12
Prior Excel Use (years)	0	8	3.17	1.95
Prior Access Use (years)	0	6	1.14	2.25
MyITLab Usage (Hours /wk)	1	16	3.57	2.12
Gender	Male: 88		Female: 51	
Preferred Learning Styles	Learn by Doing (86); Learn by Thinking (57), Learn by watching (34); Learn by Feeling (8)			

6. RESULTS

The 139 completed surveys collected from the study were analyzed with Smart PLS and results are presented in Tables 2 and 3.

Table 2: Construct AVE, Composite Reliability, R-square, Cronbach Alpha

Construct	AVE	Composite Rel	R-sqr	Cronbach Alpha
Perceived Ease of Use	0.7077	0.9061	0.5068	0.8623
Facilitating Conditions	0.6662	0.8870	0.6213	0.8283
Individual Characters	n/a	n/a	0.5702	n/a
Learning Outcomes	0.6810	0.8949	0.7401	0.8431
Perceived Usefulness	0.8012	0.9416	0.5008	0.9176
System Features	n/a	n/a	0.6203	n/a
Usage	0.6329	0.8726	0.6516	0.8030

The seven constructs have measurement validity as seen from Table 2 with high AVE and R-square values. The reliability measures for the constructs are represented by Composite reliability and Cronbach's Alpha and the high scores on these measures indicate adequate reliability. Compared with coefficient alpha, which provides a lower bound estimate of internal consistency, the composite reliability is a more rigorous estimate of the reliability. The

recommended levels for establishing a tolerable reliability are above the 0.70 threshold and above 0.80 for strong reliability. Consequently, evidence for internal consistency and construct reliability are supported by these results.

After the measurement model was validated, Smart PLS was used to test the paths between constructs and determine the support for the study hypotheses. Table 3 lists the results of the hypotheses testing. The greater the acceptance of the TML system features results in a higher perceived ease of use, higher perceived usefulness and system usage (H1a-c supported). The greater the Perceived ease of use, the higher the perceived usefulness of the system (H2 supported), but not the usage of the system (H3 not supported). Higher perceived usefulness did increase system usage (H4 supported). Individual characteristics was found to affect the level of perceived ease of use (H6a), the level of system usage (H6b), the level of perceived usefulness (H6c) and the level of learning outcomes (H6d). The level of facilitating conditions was found to support the level of perceived ease of use (H7b supported). But facilitating conditions did not support the level of perceived usefulness (H7a not supported), nor the level of usage of the system (H7c not supported). The greater the level of system use, the higher learning outcomes (H5 supported).

Table 3: Hypothesis Testing and T-values

Hypothesis	Path Coeff	T-Val	Support
H1-a: TML -> PEU	0.3224	2.2584	YES
H1-b: TML -> PU	0.5487	3.1829	YES
H1-c: TML -> USE	0.4234	2.5073	YES
H2: PEU -> PU	0.1433	1.9729	YES
H3: PEU-> USE	0.1367	1.2229	NO
H4: PU -> USE	0.2845	2.7436	YES
H5: USE -> OUT	0.2824	2.3550	YES
H6-a: INDV -> PEU	0.2216	2.8847	YES
H6-b: INDV -> USE	0.5032	4.2795	YES
H6-c: INDV -> PU	0.3593	2.7311	YES
H6-d: INDV -> OUT	0.6240	5.7402	YES
H7-a: FC -> PU	0.0416	0.2355	NO
H7-b: FC -> PEU	0.3732	3.7731	YES
H7-c: FC -> USE	0.1623	1.2577	NO

7. DISCUSSION

The goals of this study were twofold: to develop and empirically validate an extended TML research model that also includes the users' learning system usage behavior and the facilitating conditions supporting such usage. Secondly to use that model to measure the impacts of those constructs on the usage

behavior and facilitating conditions on the users' learning outcomes.

The study found that the features of the e-learning system are significantly related to the perceived usefulness and ease of use of the system and also its usage. Perceived usefulness of the e-learning system drives greater usage of the system. Moreover, the lack of perceived ease of use by the users does not inhibit system usage, as the ease of use and usage do not show a significant relationship. Facilitating conditions like technical support, computing resources and instructions about e-learning system increase the perceived ease of use for the users. But such conditions do not impact the perceived usefulness of the e-learning system nor the ultimate usage of the system. Individual characteristics is the most important factor that has the strongest supported relationship in determining usage of the e-learning system and impacting learning outcomes of the user.

The results of the study suggest that more support needs to be provided to users during the initial adoption phase of the e-learning system. Users can be engaged by things like group workshops, proactive technical support and one on one sit down help to get started. All these reduce the cognitive load on the users and increases the perception of the e-learning system as being easy to use. After the initial adoption, the usage and learning outcomes are strongly impacted by individual characteristics. The usage intensity and learning outcomes are governed by the set of features that appeals to each individuals learning style and habits. This calls for a future follow on study to evaluate which features of the e-learning system are favored by what types of learners. Moreover, can adequate personalization of the e-learning system, that can support individual learning habits and preferences, be achieved that can impact learning outcomes?

8. REFERENCES

- Agarwal, R. & Karahanna, E. (2000) Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage, *Management Information Systems Quarterly*, 24(4), 665-694.
- Alavi, M. & Leidner, D.E. (2001). Research Commentary: Technology-Mediated Learning - A Call for Greater Depth and Breadth of

- Research, *Information Systems Research* 12(1), 1-10.
- ASTD (2013), ASTD State of the Industry Report. <http://www.astd.org>, ASTD.
- Bekkering, E. & Hutchison, D. (2009), A Follow-up Study of Using Remote Desktop Applications in Education, *Information Systems Education Journal*, 7(55). 1-13.
- Behsch, S. & Rager, M. (2012). Cloud-Based Online Learning Platforms, Proceedings of BIS 2012 Workshops, LNBIP 127, 165-176.
- Bohlen, G.A. & Ferratt, T.W. (1997). End User learning: An experimental comparison of lecture versus computer-based learning, *Journal of End User Computing*, 9(3), 14-27.
- Bostrom, R.P., Olfman, L. & Sein, M.K. (1990). The Importance of Learning Style in End-User Learning, *MIS Quarterly*, 14(1), 101-119.
- Davis, F.D. (1989). Perceived usefulness, Perceived ease of use and user acceptance of information technology, *Management Information Systems Quarterly*, 13(3), 319-340.
- Davis, F.D. (1993), User acceptance of information technology: System Characteristics, User perceptions and behavioral impacts, *International Journal of Man-Machine Studies*, 38, 475-487.
- Gupta, S. & Bostrom, R.P. (2009), Technology-Mediated Learning: A Comprehensive Theoretical Model, *Journal of the Association for Information Systems*, 10(9), 686-714.
- Gupta, S., Bostrom, R.P. & Anson, R. (2010), Do I matter? The Impact of Individual Differences on Learning Process, *Proceedings of the 2010 Special Interest Group on MIS's 48th Annual Conference on Computer Personnel Research*, 112-120.
- Gupta, S., Bostrom, R.P. & Huber, M. (2010). End-User Training Methods: What we Know, Need to Know, *The Database for Advances in Information Systems*, 41(4), 9-39.
- Igbaria, M., Parasuraman, S., and Baroudi, J.J. (1996), A Motivational Model of Microcomputer Usage, *Journal of*
- Management Information Systems*, 13(1), 127-143.
- Lau, S. H. & Woods, P. C. (2008). An investigation of user perceptions and attitudes towards learning objects. *British Journal of Educational Technology*, 39(4), pp. 685-699.
- Lee, B.C., Yoon, J.O., Lee, I. (2009). Learner's Acceptance of e-Learning in South Korea: Theories and results, *Computers and Education*, (53), pp. 1320-1329.
- Lin, H.F. (2009). Examination of Cognitive Absorption influencing the intention to use a virtual community, *Behavior and Information Technology* 28(5), 421-431.
- Madon, S. (2000). The Internet and Socio-Economic Development: Exploring the Interactions, *Information Technology and People*, 13(2), 85-101.
- Malhotra, Y., & Galletta, D.F. (1999). Extending the Technology Acceptance Model to Account for Social Influence: Theoretical Bases and Empirical Validation, Proceedings of 32nd Hawaii International Conference on Systems Sciences Big Island, HI.
- McGill. T.J. & Klobas, J.E. (2009). A task-technology fit view of learning management system impact, *Computers and Education* 52(2), 496-508.
- McFarland, D. J., & Hamilton, D. (2006). Adding contextual specificity to the technology acceptance model. *Computers in Human Behavior*, 22, 427-447.
- Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, 2(3), 192-222.
- Mun, Y. Y., & Hwang, Y. (2003). Predicting the use of Web-based information systems: Self-efficacy, enjoyment, learning goal orientation, and TAM. *Human-Computer Studies*, 59(4), 431-449.
- Musa, P.F., Meso, P. & Mbarika, V. (2005), Towards sustainable Adoption of Technologies for Human Development in Sub-Saharan Africa:

- Precursors, Diagnostics and Prescriptions, *Communications of the Association for Information Systems*, 15(33), 592-608.
- Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic Skills IT training. *MIS Quarterly*, 25(4), 401-426.
- Noguera, J.H. & Watson, E.F. (2004). Effectiveness of Using an Enterprise System to teach process centered concepts in business education. *Journal of Enterprise Information Management*, 17(1), 56-74.
- Olfman, L. & Pitsatron, P. (2000). End-user training research: Status and models for the future. In R.W. Zmud (ed.), *Framing the domains of IT management : projecting the future-- through the past* (pp. 129-146), Pinnaflex Education Resources Inc., Cincinnati, Ohio.
- Piccoli, G., Ahmed, R. & Ives, B. (2001). Web-Based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic IT Skills Training, *MIS Quarterly*, 25(4), 401-426.
- Pituch, K. & Lee, Y. (2006). The influence of system characteristics on eLearning use, *Computers and Education*, 47(2), 222-244.
- Roca, J.C. & Gagne, M. (2008). Understanding eLearning continuance intention in the workplace, A self-determination theory perspective, *Computers in Human Behaviour*, 24(4), 1585-1604.
- Ryan, R.M. & Deci, E.L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions, *Contemporary Educational Psychology*, 25, 54-67.
- Sein, M. K. & Bostrom, R. P. (1989). Individual differences and conceptual models in training novice users. *Human Computer Interaction*, 4(3), 197-229.
- Selim, H.M. (2003). An Empirical Investigation of Student Acceptance of Course Websites, *Computers and Education*, 53(3), 588-598.
- Sun, P.C., Tsai, R.J., Finger, G., Chen, Y.Y., & Yeh, D. (2008). What drives a successful eLearning? An Empirical investigation of the critical factors influencing learner satisfaction, *Computers and Education*, 50(4), 1183-1202.
- Szajna, B. & Mackay, J.M. (1995). Predictors of Learning Performance in a Computer-User Learning Environment: A Path-Analytic Study, *International Journal of Human-Computer Interaction*, 7(2), 167-185.
- Tharenou, P. (2001), "The Relationship of Training Motivation to Participation in Training and Development", *Journal of Occupational and Organizational Psychology*, 74(5), 599-621.
- Venkatesh, V., Thong, J.Y.L. & Xu, X. (2012), Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology, *MIS Quarterly*, 36(1), 157-178.
- Yi, M.Y. & David, F.D. (2003). Developing an Validating an Observational Learning Model of Computer Software Learning and Skill Acquisition, *Information Systems Research*, 14(2), 146-170.
- Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D. (2003), User Acceptance of Information Technology: Toward a Unified View, *Management Information Systems Quarterly*, 27(3), 425-478

Editor's Note:

This paper was selected for inclusion in the journal as the CONISAR 2015 Best Paper. The acceptance rate is typically 2% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2015.

Appendix –Survey Questionnaire

Construct & Sources	Survey Items
<p>TML System</p> <p><i>Gupta, et.al. (2010); Gupta (2009);</i></p>	<ol style="list-style-type: none"> 1. The output from MYITLab was presented in a useful format. 2. The information about Excel and Access from MyITLab is accurate. 3. MyITlab graded my assignments in a fair manner. 4. I am satisfied with the management of assignments in MyITLab. 5. I am satisfied with the way MyITLab gave feedback on assignments. 6. I am satisfied with the way MyITLab accepted my assignments online.
<p>Perceived Usefulness</p> <p><i>Venkatesh, et.al. (2003); Venkatesh, et.al. (2008); Davis (1989)</i></p>	<ol style="list-style-type: none"> 1. Using MyITLab enhanced my effectiveness in learning. 2. Using MyITLab increased my productivity in the course 3. I found MyITLab to be very useful in the learning process 4. MYITLab fit my study habits and practices.
<p>Perceived Ease of Use</p> <p><i>Venkatesh, et.al.2003); Venkatesh, et.al.2008); Davis (1989)</i></p>	<ol style="list-style-type: none"> 1. It was very easy for me to learn to use MyITLab. 2. It was easy to find information about MyITLab 3. I found MyITLab to be very easy to use. 4. It was easy for me to become skillful at using MyITLab.
<p>Individual Characteristics</p> <p><i>Nogura and Watson (2006); Ryan & Deci (2000); Bostrom (1990)</i></p>	<ol style="list-style-type: none"> 1. I was motivated to learn as much as I can from this class. 2. I was very interested to take this class. 3. I was excited about learning the skills that were covered 4. I worked hard on this project only to get a better grade
<p>Facilitating Conditions</p> <p><i>Venkatesh, et.al. (2003); Venkatesh, et.al. (2008)</i></p>	<ol style="list-style-type: none"> 1. I had the resources necessary to use MyITLab 2. I had all the support necessary to use MyITLab 3. I am satisfied with the documentation of MyITLab 4. I am satisfied with the facilities and equipment that were available for my use in the learning process.
<p>Usage</p> <p><i>Venkatesh, et.al. (2003); Venkatesh, et.al. (2008); Davis (1989)</i></p>	<ol style="list-style-type: none"> 1. I believe that I used MyITlab quite extensively. 2. I used MyITLab more frequently compare to other learning systems. 3. I relied on MyITlab to successfully complete this course 4. Once I started working with MYITlab, I found it hard to stop.
<p>Learning Outcomes</p> <p><i>Gupta, et.al. (2010); Gupta (2009); Pituch and Lee (2006)</i></p>	<ol style="list-style-type: none"> 1. MyITlab helped me to improve my proficiency of Excel and Access. 2. My ITLab allowed me to grow my knowledge of the applications. 3. MyITLab challenged me to develop new knowledge beyond my existing knowledge of features of Excel and Access to solve problems 4. I am now confident that I can finish an assigned task with Excel and Access. 5. I now understand how I can navigate Excel and Access

Leakage of Geolocation Data by Mobile Ad Networks

Christopher Snow
csnow@pace.edu

Darren Hayes
dhayes@pace.edu

Catherine Dwyer
cdwyer@pace.edu

Seidenberg School of Computer Science & Information Systems
Pace University
New York, NY

Abstract

Mobile ad networks connect advertisers with mobile app publishers, to improve the suitability of ads shown to app users. These ad networks send metadata about mobile users and their devices to advertisers, who then use this metadata to select appropriate ads. This research demonstrates how mobile networks leak location data and other sensitive information from mobile phones by sending plaintext, unencrypted transmissions. It is therefore possible that geolocation information, associated with a user, could be captured by government and private sector entities, as well as by nefarious actors. An experiment was designed to discover how iPhone applications ("apps") transmit unencrypted geodata to identify a user's location. This research revealed that several popular mobile apps disclose the location of an iPhone by means of its UDID (serial number); this primarily occurred through mobile ad networks.

Keywords: Mobile Privacy, Mobile Advertisements, Geodata, iPhone Forensics, Edward Snowden, NSA

1. INTRODUCTION

Consider this: a person can no longer leave the country without a number of people, other than friends or family, knowing about it. It is only possible though if you pack-up, get in a car that is not wired to GPS, and leave your cellphone behind. Even then, some type of surveillance camera is monitoring and recording you. Sounds impossible, doesn't it? Mobile devices have completely taken away a user's sense of privacy, more specifically a user's sense of locational privacy. Modern smartphone owners have a subliminal sixth sense that "big brother" is always watching. From social media, to photos

and mobile applications, there is always at least one person that a smart device user has never met who knows exactly where that smart device user is. Smart devices, such as the iPhone, the Microsoft Surface Tablet, and even a car's navigation system are always collecting, storing, analyzing, and sending data about the user's location in the form of longitude and latitude metadata, otherwise known as geodata.

Documents released by Edward Snowden in 2015 describe Operation BADASS - a U.S. government program that included the retrieval of metadata transmitted by mobile ad networks, including user location information from mobile

devices and computers. The National Security Agency and Department of Defense allegedly captured geodata from these devices via Wi-Fi, iPhone UDID, and electronic paper trails (SPIEGEL ONLINE, 2013).

Unbeknownst to many, U.S. government agencies, like the National Security Agency (NSA) or the Department of Justice (DOJ), are silently intercepting foreign and domestic communications from mobile devices and subsequently analyzing these communications. The United States is not the only country performing these covert operations. India, France, and Saudi Arabia all perform similar, if not more invasive, surveillance on their citizens. According to documents leaked by whistleblower Edward Snowden and other anonymous sources, government agencies are collecting electronic data at unfathomable speeds and quantities (Bamford, 2012; MacAskill, 2013).

A leaked presentation from the British intelligence agency Government Communications Headquarters (GCHQ) titled "Mobile Apps Doubleheader: BADASS Angry Birds" shows that government agencies are capturing data collected from private sector communications, typically mobile advertisements, using a program under the codename 'BADASS' (Lee, 2015). The research presented in this paper shows data collection from mobile advertisements is quite possible, demonstrating how insecure mobile advertisement transmissions can be.

2. BACKGROUND

One must analyze U.S. law to understand how the U.S. government has the legal authority to collect metadata from smartphone users. Section 216 of the PATRIOT Act states that the U.S. government has the ability to collect "data without content" under pen register. Thus, data such as IP addresses, geolocation, phone numbers, and URLs can be collected with a court order rather than meeting the higher standards needed to obtain a warrant. Of course, this data can be used to determine more invasive information by performing a search for a URL or using the coordinates on a map (Doyle, 2001). However, analyzing this data superficially does not provide sensitive information about an individual.

Section 216's definition of pen register allows the NSA to legally perform operations, like Operation BADASS, to collect content-less data in large quantities. Operation BADASS indicates

that governments can collect personal information from private sector activities and communications. Figure A1 shows what information is collected through mobile advertising provider Mobclix, now named Axonix. The slide, which is derived from the previously mentioned leaked GCHQ presentation, identifies the fields within the mobile advertiser's HTTP requests. The most important fields noted in the presentation include "&ll" and "&u." The "&ll" field is described in the slide as the field containing longitude and latitude coordinates. While the "&u" field in the request is noted as containing an "IMEI." An IMEI number uniquely identifies a user's cellphone on a GSM network (Lee, 2015). Both of these combined elements allow a user profile to be developed and track movements. A query could be made on Mobclix/Axonix's database for a specific IMEI sequence to display the HTTP requests made by that specific user. This would allow a map to be generated based on the locations where a specific user has made HTTP requests while being served an advertisement. This map could ultimately lead to someone's true identity being revealed based on the patterns of their movement.

A bill introduced in the U.S. House of Representatives on July 8, 2015, titled "The Consumer Privacy Protection Act", would require companies to notify their customers within 30 days if hackers obtained "sensitive information". This bill expands the definition of "sensitive information" to include geolocation data (Davis, 2015). Treating geolocation as "sensitive information" could set a precedent for future legislation to include geolocation as personal information. The bill also seeks to encourage higher security procedures, including the minimization of "sensitive personally identifiable information" stored by companies.

3. RELATED WORK

A report titled "Taming the Android AppStore: Lightweight Characterization of Android Applications" from the networking and security department at EURECOM, a graduate school in France, described research on the Android platform (Vigneri, Chandrashekar, Pefkianakis, & Heen, 2015). Their research sought to analyze network connections when Android applications are launched. The goal of EURECOM's research was to categorize the types of communication connections and develop an application for users to calculate the privacy of an application. They sought to provide users with information about how apps collect personal information. They

tracked a smart device's network connections, through apps, to verify whether these connections were safe. The researchers routed all traffic from these Android applications through a virtual private network (VPN). EURECOM's research uncovered that AdMob and Flurry, which are two popular advertisement companies, made the top 20 list of servers receiving communications from Android applications. From the analyzed requests, EURECOM researchers were able to develop a suspicion algorithm that is used to identify how suspicious an application is. They used factors such as the webutation.com ranking to decide how safe or malicious the HTTP request made by the application was and how many times HTTP requests are made. Their research concluded that many Android applications on the *Google Play Store* make undesirable communications unbeknownst to the user. Additionally, the research concluded that a number of high-ranking applications made excessive requests to advertising companies.

A report released by Zscaler Inc. in early 2014 documented how the Angry Birds app was divulging personal information to third parties. After the accusation, Rovio, the developers of Angry Birds, attributed blame to mobile advertisers (Robertson, 2014). Their privacy policy exclusively states Rovio "may collect and process your location data to provide location related services and advertisements." Additionally, Rovio states they "reserve the right to use and disclose the collected, non-personal data for purposes of advertising by Rovio" (Rovio, 2013). This accusation coincided with the GCHQ presentation of Operation BADASS and it is not hard to surmise that user location data has been leaked since the Angry Birds application does use advertising extensively. The study conducted by Zscaler Inc. looked at 30,000 Android applications and found that 38% leaked a smartphone's unique IMEI/MEID number and 15% of these apps divulged the user's phone number (Robertson, 2014).

4. EXPERIMENTAL RESEARCH

The experimental research documented in this paper sought to discover how a user could be profiled by means of mobile advertisements directed through iPhone apps. Related work has shown that Android applications regularly divulge personal data. With 1.2 million applications available in Apple's App Store, it is hard to believe that all of them are secure when dealing with a user's personal information (Perez, 2014). Additionally, this research seeks

to investigate mobile advertisement security through ad networks.

Experimental Design

Our experimental research simulated Operation BADASS, where application data is captured and recorded.

Figure A2 graphically reflects the communications identified in this experiment. This experiment displays the "Middlemen Servers" used to capture the HTTP and HTTPS requests between the smartphone application and the application server (or advertising agency). The figure also shows metadata being shared between the application and the advertising agency. Essentially, an advertising agency, like Mobclix, could use information loaded into an application by the user, such as age or location, to learn more about the user in order to serve targeted, or more suitable, ads. Typically, communications between the application and an application server are encrypted. However, communications between the application and an advertising agency are not encrypted (Lee, 2015). This means that data gathered about a user, by an advertisement may be leaked and collected by "middlemen," such as the NSA.

Our experiments sought to identify what data on an iPhone is transmitted from applications via insecure HTTP requests. With our framework, multiple iPhone applications are used casually and authentically while a program, called Debookee, analyzes and captures the Internet traffic flowing through a wireless router. This data captured from Internet requests is then analyzed to see if any personal information, specifically data related to location, is recorded.

To test this framework, the following tools were used:

- Debookee by iwaxx
- iPhone loaded with various applications (travel, gaming, dating, shopping, etc.)
- Wireless Internet router
- Internet connected computer running Mac OS X

Experiment

The data for this experiment was obtained in April 2015. First, a controlled environment had to be established before any HTTP requests from the target smartphone were captured. To do this, a wireless router was connected to an Ethernet port to facilitate the connection of mobile devices to the Internet. Subsequently, all

wireless devices connected to the wireless router were documented. Debookey offers a LAN scan to display all connected devices and also a basic description of these devices. Figure A3 and Figure A4 show how the target used is correct by matching the IP addresses (i.e., 192.168.1.103).

With the target identified, the HTTP traffic transmitted by the phone can be captured. The traffic from about 40 applications was analyzed. Three yielded the most useful data and were chosen for further analysis. They included Transit, BestBuy, and Kick the Buddy.

Transit

Transit operates as a NYC subway navigator. The application allows a user to input a starting destination, often their current location, and an end point. The application then shows the user the best subway to take to get from point A to point B. This application was used at 100 Henry Street, Brooklyn, NY, and a request was made for transit directions to downtown Manhattan. The Transit app sends requests to a server with fields matching a user's search terms. Debookey was able to capture the following unencrypted HTTP request when this search was performed:

```
http://us-east-planner1.thetransitapp.com/open  
tripplanner-api-webapp/ws/plan?routerId=10&s  
howIntermediateStops=true&walkReluctance=3.  
75&toPlace=40.711016,- 74.004845&from  
Place=40.697740,-73.993262&time=16:40  
&date=04/18/2015
```

Figure 1: Unencrypted HTTP request from Transit App.

The important fields are bolded, namely "toPlace," "fromPlace," and "time." The "toPlace" and "fromPlace" are easy to distinguish; they are the starting and end coordinates of the query. Figure A5 and Figure A6 show that the coordinates, provided by gps-coordinates.net, in the "toPlace" and "fromPlace" fields match this search. Additionally, time shows the exact military time when the search was performed, as well as the date.

If intercepted, this locational data could easily determine that a student performed the search. A quick search with the term "100 Henry Street" on Google will reveal that the building is used as student housing. Someone intercepting this communication could easily conclude that the search was performed a student in the building. However, identifying the exact student would be difficult because the capture request did not contain any identifying ID or token.

Nevertheless, this data shows that these types of navigational searches on Transit are insecure and could be read by someone unbeknownst to the user.

BestBuy

BestBuy, unlike Transit, does not provide users with navigational information. It does, however, utilize a user's information to locate BestBuy retailers in the area. What was surprising about the data captured from the BestBuy app was that it remembered previous locations. The request in Figure A7 was made after the application was loaded. An interesting "area" field appears in the second line of the request (portions are redacted for privacy). Using gps-coordinates.net, the coordinates pointed to a residence on Long Island. The residence in question happened to be the user's actual permanent address and it also happens to be the last place where the application was run. It is unclear why this request was made, but it appears the application may have been pulling saved data from the application's last use.

A more interesting part of this request is the "cookie" field that also appears in Figure A7 on the fifth line. This cookie field ties back to previous research, specifically operation BADASS. BestBuy appears to be using a cookie to store personal data, which can communicate data to and from the application for advertising and marketing purposes. Therefore, BestBuy presumably has a database where specific cookie tokens could be used in a query in order to see where and when someone has used their app. In this case, it would be easy to tie this specific request back to someone who lives at the residence based on the accurate coordinates the user's phone provided them.

Kick the Buddy

CrazyLion Studios Limited has produced multiple free games that appear on the Apple App Store top charts, including Kick the Buddy. Free games frequently feature more mobile advertisements. Mobile ads are an effective way to make money while keeping the application free for users to download (Hof, 2014). Kick the Buddy features multiple advertisements, as seen in Figure A8 and Figure A9. These advertisements have made Kick the Buddy the perfect candidate to explore further in this controlled experiment.

As expected, multiple requests were made to advertising services that transmit advertisements to the user. The following unencrypted HTTP request stood out:

```
http://ads/m/imp?appid=&cid=c666ac880b044f8cac90a19fdc099893&city=Brooklyn[...]&dev=iPhone7%2C2&[...]http%3A%2F%2Fcpp-test.imp.mpx.mopub.com%2Fclick%[...]26app_name%3Dkick%2520the%2520Buddy%253A%2520Second%2520Kick%2520Free%[...]&udid=ifa%3A90E*****CC9-4**C-B7B8-6D24*****B54
```

Figure 2: Unencrypted HTTP request from Kick the Buddy Mobile Game.

The ad was identified as Kick the Buddy: A Second Kick Free, the full name of the application where the ad was found. It is important to note that no advertisements were opened in this experiment; these requests came only from the advertisement found in the application being used. The advertisement was able to determine the correct general location of the user in Brooklyn. Though, perhaps the most valuable piece of information to come from this request is the UDID number, **90E*****CC9-4**C-B7B8-6D24*****B54** (portions of the UDID number have been omitted for privacy).

The UDID is the token assigned to the user specifically. It is used so that Mopub can pull the information they have already collected about the user and subsequently transmit the user with more targeted marketing materials based on like one's gender, age, and/or location. With this UDID number that potentially identified the user, a query was made to Debookee using the discovered UDID to find any other requests that were made by Mopub. Debookee returned with 45 HTTP requests made by Mopub from other applications to serve advertisements to the user. A majority of the requests contained the same information as the Kick the Buddy request, many with less information. Though, there was one alarming result:

```
http://ads.mopub.com/m/ad?v=8&udid=ifa:90E*****CC9-4**C-B7B8-6D24*****B54&id=agltb3B1Yi1pbmNyDAsSBFNpdGUYorkhDA&nv=1.17.2.0&q=m_gender:m,m_age:21&o=p&sc=2.0&z=0400&ll=40.69770885046903,-73.99321115040379&lla=65&mr=1&ct=2&av=2.2.2&cn=Verizon&iso=us&mnc=480&mcc=311&dn=iPhone7%2C2
```

Figure 3: Unencrypted HTTP request from MoPub ad network.

As one can see, the UDID highlighted in bold matches the UDID from the Kick the Buddy HTTP request. There is also gender and age information being transmitted. This information matches the first author, whose phone was used

to collect this data. The advertisement is also identified to come from Mopub. Unfortunately, the HTTP request does not exclusively provide what application generated the ad. But, the information that is stored within this request is extremely personal. Recalling Figure A1, there are similar fields in this request that GCHQ pointed out in their presentation; specifically the "ll=" field for longitude and latitude. The "ll=" field provides the same coordinates from Figure A6, the Transit results. However, it can be inferred that this advertisement request did not come from Transit because the Transit application does not transmit ads to their users. It could, however, have come from another advertisement within Kick the Buddy because the game transmits multiple advertisements. It is unclear how the advertisement was able to identify the user's gender and age as both fields were never provided to any of the applications we researched. It is also important to note that Facebook was not a part of this experiment and the application was not loaded onto the subject smartphone at the time of experimentation.

5. EXPERIMENTAL CONCLUSIONS

This experiment showed that not only mobile advertisers, but also location-aware applications, disclose geodata from mobile phones. The experiment showed applications, like Transit, knowingly using location are sending unencrypted requests. Therefore, it is plausible that the NSA, or any other interested parties, have the ability to gain intelligence regarding location through the interception of mobile advertising transmissions. It is important to note that this was a very small-scale experiment. Imagine if this collection was performed on the entire city of New York by tapping into all of the unencrypted, or open, Wi-Fi hotspots. Unencrypted HTTP requests could theoretically be collected by thousands of people at once. Then requests with the same UDID can be identified as the same people to start building different profiles. The experiment showed identification of targets is possible; 45 other queries were found using the Mopub UDID that was assigned to the first author.

Although the results from this experiment are quite specific, one must consider broader implications of linking data to other data. The information from, for example, one Mopub advertisement may only have information about the user's gender. But, since all Mopub advertisements are linked to one person with UDID, a user's profile can be expanded. Now, not only does Mopub know a user's gender from

application "x" they also now know the same user's location from application "y." Now that a user's coordinate location is known, an analyst can query that location against the HTTP capture database to see if navigation applications like Transit were used to see where the same user traveled.

Another issue that this experiment has unexpectedly uncovered is the question of who is responsible for the data being leaked. It is completely up to the discretion of the app developer and/or advertising companies to ensure the HTTP requests to and from served ads are secure, not the phone manufacturer or wireless router. As seen in Figure A10, HopStop, another subway navigation application like Transit, secures all HTTP requests under HTTPS. This means that any search and serving of data travels through a safer port so man-in-the-middle applications, such as Debookee and WireShark, cannot intercept wireless data.

Google has brought HTTPS encryption to all of their services, including AdMob, its mobile advertising service. In June 2015, a "vast majority" of AdMob's connections were moved over to HTTPS encryption (Google, 2015). In addition, Apple added App Transport Security (ATS) in September 2015 to the iOS 9 operating system. ATS is a framework that imposes network security best practices, for example requiring HTTPS for all network requests, and the use of Transport Layer Security (TLS) 1.2 or higher (Jacobs, 2015).

6. IMPORTANCE OF RESEARCH

This research was able to reproduce how easy it is to collect geodata leaked by mobile apps. These findings validate the leaked Snowden documents from GCHQ.

This research shows government agencies could use the private sector for data collection. The experiment shows that the data displayed in *Figure A1* is accurate. It is not hard to imagine that the American government, or any government, could be collecting the unencrypted data detailed in this paper's findings. The claimed operations of the NSA, outlined by Edward Snowden, are perhaps valid. The private sector is an interesting sector to investigate since companies like Apple and Microsoft are adamant about not facilitating government investigations. But, downloaded applications on smart devices, including the iPhone, are leaking personal data that people would rather not share with law enforcement.

The conclusions drawn from this paper's research also show a defining shift in intelligence gathering. Before the widespread use of smartphones, an important source of information about people and/or their location was through public records or hacking. However, now, there is a plethora of unencrypted personal data flowing through the Internet for anyone to intercept. With that in mind, the key stakeholders in all of this data collection are definitely the mobile advertisers. Mobile advertisement providers like AdMob and Mopub essentially can build entire profiles on a person by correlating all of the information they have on a person. This is one of the most important takeaways from the research conducted in this paper.

7. FUTURE EXPERIMENTS

A future experiment could be to label each incoming transmission, and identify which app it came from. When the HTTP requests were compiled and exported for analysis, it proved difficult to distinguish which apps generated them. Most requests contained the name of the application inside the HTTP requests but some did not.

It may be useful to create multiple applications using different advertising companies, like AdMob or Mopub, to see how much a developer can control the data that can be captured and how easy it is to secure these transmissions. Aside from advertisers, it may prove helpful to also create applications that make HTTP requests back to a self-operated webserver. Transit and BestBuy, for example, both use unsecure HTTP requests back to servers they own. A future experiment might examine how easy it is to secure data transmissions where the app and the hosted server are both self-operated. This could help decide who is responsible for leaking data when encryption is not enabled on transmissions.

A future experiment could also have two participants, one "agent" and one "criminal," both connected to a controlled public wireless router with casual innocent users. In this experiment the designated agent would be capturing the wireless router's unencrypted HTTP requests in an attempt to identify the designated criminal the designated agent has been tracking. This experiment would both look at the possibility of detecting a target on an unencrypted wireless router as well as how much accidental data could be collected from innocent users in a real operation.

8. CONCLUSION

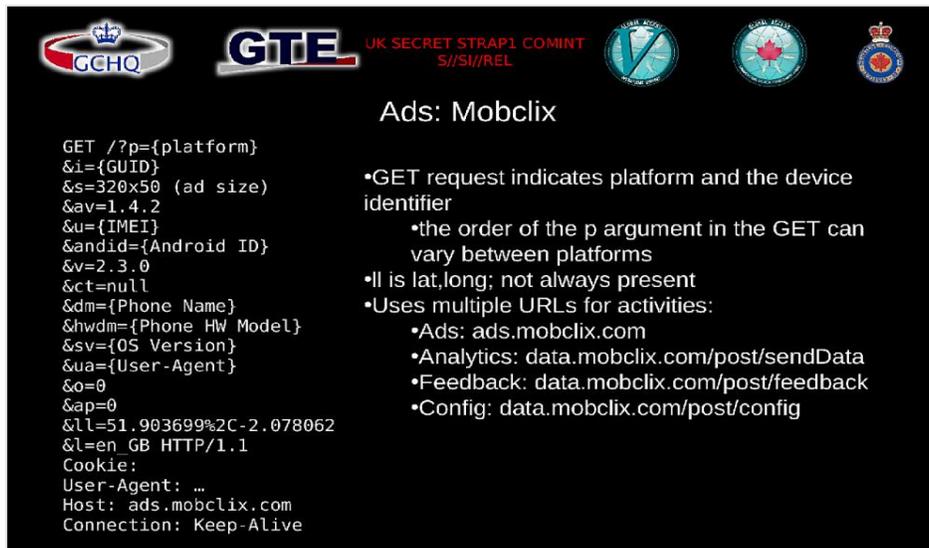
The conclusions drawn from the data presented in literature research and experimental research shows that the U.S. government, along with any other interested parties, does have the ability to target a person and track them through their mobile geodata.

An important finding is that mobile devices are emitting more personal data about the user than the user may realize. Wireless unencrypted personal data is free for anyone to use, whether it is government, a business, or a criminal. In this wireless day and age a hacker no longer has to know a username and password to get personal data, they just essentially have to eavesdrop on the personal data that a user is unintentionally sending. Along the same lines, a business can simply use a tool to analyze social media chatter to find where people are discussing their product instead of sending out an old-fashioned survey. Both cases use the free available location data on Internet that users are willingly sharing, whether or not they know it.

9. REFERENCES

- Bamford, J. (2012, March 15). The NSA is building the country's biggest spy center (watch what you say). Retrieved from http://www.wired.com/2012/03/ff_nsadatacenter/
- Doyle, C. (2001, December). Terrorism: section by section analysis of the USA PATRIOT act. Congressional Research Service, the Library of Congress.
- Davis, W. (2015, July 8). New data breach bill covers photos, geolocation data, other 'sensitive' information. Retrieved July 14, 2015, from http://www.mediapost.com/publications/article/253573/new-data-breach-bill-covers-photos-geolocation-data.html?utm_source=newsletter&utm_medium=email&utm_content=headline&utm_campaign=84298
- Google. (2015, April 17). Inside adwords: ads take a step towards "HTTPS everywhere". Retrieved July 10, 2015, from <http://adwords.blogspot.com/2015/04/ads-take-step-towards-https-everywhere.html>
- Hof, R. (2014, August 27). Study: mobile ads actually do work - especially in apps. Retrieved from <http://www.forbes.com/sites/roberthof/2014/08/27/study-mobile-ads-actually-do-work-especially-in-apps/>
- Jacobs, B. (2015). Apple Tightens Security With App Transport Security. *Code Tutorials*. <http://code.tutsplus.com/articles/apple-tightens-security-with-app-transport-security--cms-24420> Retrieved from <http://code.tutsplus.com/articles/apple-tightens-security-with-app-transport-security--cms-24420>
- Lee, M. (2015, January 26). Secret 'BADASS' intelligence program spied on smartphones. *First Look Media*. Retrieved May 4, 2015, from <https://firstlook.org/theintercept/2015/01/26/secret-badass-spy-program/>
- MacAskill, E. (2013, August 23). NSA paid millions to cover PRISM compliance costs for tech companies. Retrieved from <http://www.theguardian.com/world/2013/aug/23/nsprism-costs-tech-companies-paid>
- Perez, S. (2014, June 2). iTunes app store now has 1.2 million apps, has seen 75 billion downloads to date. Retrieved April 15, 2015, from <http://techcrunch.com/2014/06/02/itunes-app-store-now-has-1-2-million-apps-has-seen-75-billion-downloads-to-date/>
- Robertson, J. (2014, Jan 29). Leaked docs: NSA uses 'candy crush,' 'angry birds' to spy. *SFGate*. Retrieved April 11, 2015, from <http://www.sfgate.com/technology/article/Leaked-docs-NSA-uses-Candy-Crush-Angry-5186801.php>
- Rovio. (2013, October). Privacy policy - rovio entertainment ltd. Retrieved from <http://www.rovio.com/Privacy>
- SPIEGEL ONLINE. (2013, December 30). Interactive graphic: the NSA's spy catalog. Retrieved from <http://www.spiegel.de/international/world/941262.html>
- Vigneri, L., Chandrashekar, J., Pefkianakis, I., & Heen, O. (2015). Taming the android appStore: lightweight characterization of Android applications. *arXiv preprint arXiv:1504.06093*.

Appendix



The slide features logos for GCHQ, GTE, and UK SECRET STRAP1 COMINT S//SI//REL. It displays a GET request and lists several key points about the request's parameters and activities.

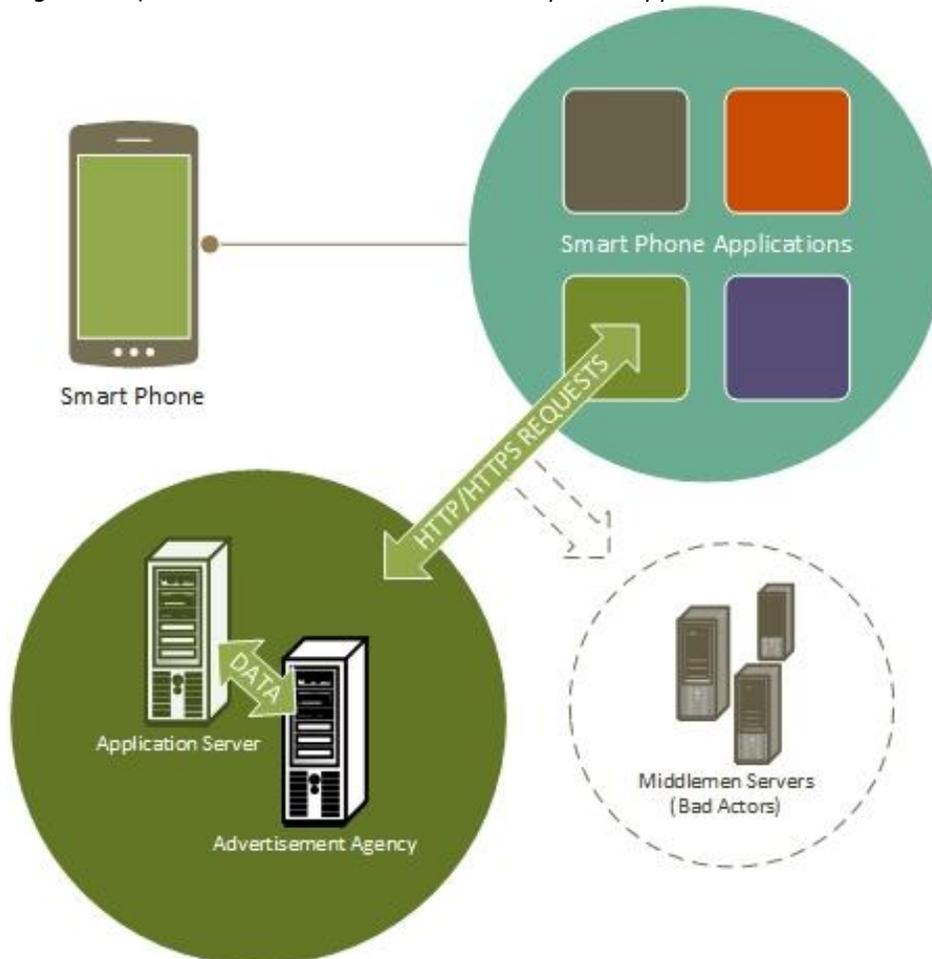
```
GET /?p={platform}
&i={GUID}
&s=320x50 (ad size)
&av=1.4.2
&u={IMEI}
&andid={Android ID}
&v=2.3.0
&ct=null
&dm={Phone Name}
&hwdm={Phone HW Model}
&sv={OS Version}
&ua={User-Agent}
&o=0
&ap=0
&ll=51.903699%2C-2.078062
&l=en_GB HTTP/1.1
Cookie:
User-Agent: ...
Host: ads.mobclix.com
Connection: Keep-Alive
```

Ads: Mobclix

- GET request indicates platform and the device identifier
 - the order of the p argument in the GET can vary between platforms
- It is lat, long; not always present
- Uses multiple URLs for activities:
 - Ads: ads.mobclix.com
 - Analytics: data.mobclix.com/post/sendData
 - Feedback: data.mobclix.com/post/feedback
 - Config: data.mobclix.com/post/config

Figure A1 | Slide from "Mobile Apps Doubleheader: BADASS Angry Birds" GCHQ presentation (Lee, 2015)

Figure A2 | Communications between smart phone applications and servers



IP address	MAC address	Hostname	Role	Vendor
192.168.1.1	00:18:f8:df:ff:e0		Gw	Cisco-Linksys LLC
192.168.1.102	30:d6:c9:05:af:b1			Samsung Electronics C
192.168.1.103	48:e9:f1:ba:83:23		Tgt	Apple
192.168.1.104	00:26:08:ea:9b:41	Christopher's...	Me	Apple

Figure A3 | Debookee screenshot (note IP address 192.168.1.103)

DHCP	BootP	Static
IP Address 192.168.1.103		
Subnet Mask 255.255.255.0		

Figure A4 | iPhone screenshot (note IP address 192.168.1.103)

```

16:24:38 http://api.remix.bestbuy.com/v1/stores(area(40
GET /v1/stores(area(40 ,%20-73 ,%2050))?
Host: api.remix.bestbuy.com
Accept: */*
Cookie: AMCV_F6301253512D2BDB0A490D45%40AdobeOrg=432
User-Agent: buyphone/9.0.5 CFNetwork/711.3.18 Darwin
    
```

Figure A7 | Excerpt of captured HTTP request from BestBuy

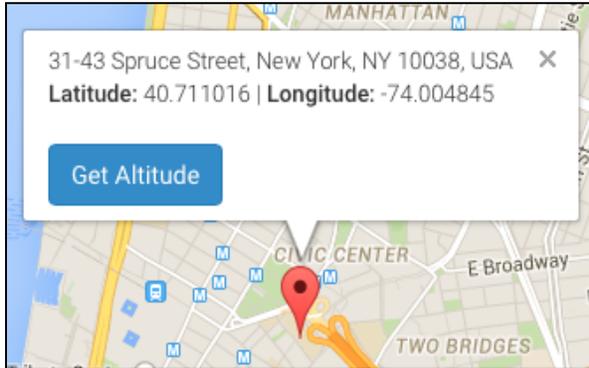


Figure A5 | Destination coordinates, toPlace



Figure A8 | Full screen ad shown on Kick the Buddy game

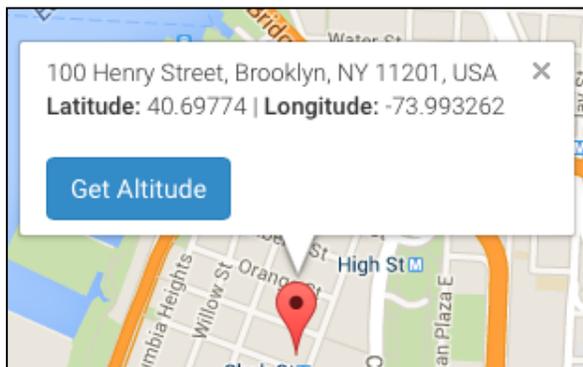


Figure A6 | Beginning coordinates, fromPlace



Figure A9 | Bottom right corner ad shown on Kick the Buddy game

```
22:59:32 H https://www.hopstop.com/... [encrypted]  
22:59:40 H https://www.hopstop.com/... [encrypted]  
22:59:42 H https://www.hopstop.com/... [encrypted]  
22:59:43 H https://www.hopstop.com/... [encrypted]
```

Figure A10 | *Excerpt of captured HTTPS requests*