Journal of the Midwest Association for Information Systems

In Memory of Daniel J. Power 1950 - 2021

Table of Contents

Volume 2021, Issue 1, January 2021 www.JMWAIS.org

Contact

Daniel J. Power, Ph.D.
Professor of Management and Information Systems
Department of Management
University of Northern Iowa
Cedar Falls, IA 50613
(319)273-2987
daniel.power@uni.edu

Rassule Hadidi, Dean
College of Management
Metropolitan State University
Minneapolis, MN 55403-1897
(612)659-7295
rassule.hadidi@metrostate.edu

Journal of the Midwest Association for Information Systems (JMWAIS) at http://jmwais.org is a double-blind, peer-reviewed, quality focused, and open-access online journal published by the Midwest United States Association for Information Systems at http://www.mwais.org/. The infrastructure for online publication of this journal is currently provided by the Metropolitan State University, St. Paul, Minnesota.

Journal of the Midwest Association for Information Systems

Table of Contents

Editorial Board

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 1

Date: 01-31-2021

Impacts of the Global Health Crisis on the Use of Information Technologies

Daniel J. Power
University of Northern Iowa, Daniel.Power@uni.edu

Rassule Hadidi
Metropolitan State University, Rassule.Hadidi@metrostate.edu

Abstract

The 2020 novel coronavirus pandemic has impacted our lives in many ways. This article examines the rapid adoption and use of traditional and "state-of-the-art" information technologies intended to help cope with the pandemic. People and organizations have adopted and used IT tools for collaboration, communication, surveillance and monitoring, remote working, and cloud-based applications for one major reason – necessity. There is no viable alternative to maintain our civilized society. These and other information technologies have helped people continue to work, to socialize, to communicate, to entertain, to visit doctor's office, to shop, and live. Experience with these technology adaptations has demonstrated that we need more and better IT solutions, more technology literacy, better public health surveillance, and better preventative measures to minimize harms from health crises to find a new normal. In the future, many people will likely choose to work and learn remotely, and organizations and governments must upgrade their digital capabilities and the skills of employees. IT can increase the robustness and adaptability of our economic and social systems as well as our well-being.

Keywords: Pandemic, rapid change, Information Technology adoption

DOI: 10.17705/3jmwa.000062

1. Introduction

One year ago, Hadidi and Power (2020) asserted that the technology adoption rate is becoming much faster, and that technology adoption has "changed from an approximately normal curve to a skewed curve with more people adopting new technology quickly." On March 11, 2020, the World Health Organization announced that COVID-19 is a pandemic. The health crisis has encouraged and promoted faster adoption of innovative technologies. Our personal, organizational, and societal experiences during 2020 reinforce our conclusions about technology adoption. A major external disruptive event can significantly alter technology adoption. A global pandemic is a serious, unexpected event. The magnitude of the current disruption is hard to comprehend. In 2020, there were globally more than 85.5M cases of the Coronavirus disease, 60.4M people recovered, and 1.85M people died (https://www.worldometers.info/coronavirus/ accessed 1/3/2021). These are conservative numbers. Some would call the Coronavirus Pandemic a Black Swan event (Taleb, 2007); others would call it an unanticipated disaster. This article reviews major changes that have occurred from the rapid adoption and increased use of Information Technology (IT).

In the past 25 years, we have experienced smaller-scale health crises like Ebola virus disease (EVD) and the human immunodeficiency virus (HIV/AIDS). More than one hundred years ago, the often-mentioned Spanish flu or the 1918 flu pandemic caused by the H1N1 influenza A virus was much deadlier and more widespread. The 1918 flu pandemic lasted from February 1918 to April 1920 and infected 500 million people – about a third of the world's population in four successive waves. Estimates are that 50 million people died during the 1918 pandemic in 26 months. The Black Death plague of 1347 to 1351 was the deadliest pandemic recorded in human history. The Black Death pandemic resulted in the deaths of 75–200 million people in Eurasia and North Africa, cf., en.wikipedia.org/wiki/Black Death.

Society survived the Black Death and the 1918 flu without IT, but with a tremendous loss of lives. The current pandemic has been far less virulent than either of those global health crises, but our response has not been rapid and robust. Our current systems and processes that incorporate IT are more robust than 100 years ago and our medical and genetics knowledge are much more sophisticated, but we can and are making improvements. Implementation of Information Technology is reducing some of the negative consequences and reducing our vulnerability from a rapidly spreading, deadly virus.

The novel coronavirus pandemic has impacted our lives in many ways. This article narrowly examines only the first and second-order changes from the rapid adoption and use of traditional and "state-of-the-art" information technologies.

2. Impacts on Individuals

The pandemic has affected young and old alike. In particular, when it comes to the social, and physical well-being and education of children, this pandemic has been devastating. Technology has helped somewhat to remedy some of the difficulties. However, the availability and utilization of technology by children is neither uniform nor consistent across various geographic areas and the economic status of families. High speed Internet simply is not available to all families and in all geographic areas around the country. For children with disabilities, this situation is even much worse. During this time of social distancing and isolation, we need to take steps and be more prepared for future pandemics and make sure the needed technologies are more widely available and accessible to all individuals and families and, in particular, families with fewer resources and younger children.

More students will continue learning through hybrid IT-based instruction, perhaps alternating between classes in person and virtually on Zoom or other platforms. Teams are using more collaboration and scheduling tools. Distributed teams are more widely accepted. Teleworking has increased significantly (Messenger, Vadkerti, and Uhereczky, 2020). The increase in telework has resulted in expanded data collection from monitoring remote workers and remote students. Many people have had to rapidly acquire new skills to be able to learn and work at home. Working and learning from home will likely continue as a popular alternative once the pandemic subsides. It will be interesting to study and see the long-term implications of using new technologies in real estate, transportation, retailing, environmental sustainability, and other domains of commerce.

More people in businesses, agencies, and other organizations are having synchronous decision-making meetings. Zoom, Microsoft Teams, Google Meet, Go To Meeting, and other conferencing tools have increased in popularity and use. Videotelephony is widely accepted and understood. Working from home or remote locations using technology is more widely accepted. Evidence shows that use of conferencing tools has significantly increased in the past year. For example, Zoom launched in 2011 and had 30 million users in 2014 according to: https://usefyi.com/remote-work-statistics/#863. Zoom surpassed 300 million daily Zoom meeting participants in 2020, a 50% increase from the prior month (200 million). For comparison, in December 2019, Zoom reported 10 million meeting participants (https://usefyi.com/remote-work-statistics/#851).The Zoom Cloud Meetings app on the App

Store ranked #3 worldwide among the non-gaming app publishers (Statista). Table 1 summarizes some usage statistics for Zoom.

Number of People Using Zoom	300 Million Daily Meeting Participants
Number of Users	200 Million
Number of Corporate Customers	265,400
Number of Schools Using Zoom	100,000
Number of Zoom Customers with more than 10 Employees	81,900
Number of Zoom Installs in June 2020	71.2 Million
Number of times Zoom was downloaded from the App Store in Q2 2020	94 Million

Table 1. Zoom facts (as of July 2020), source https://expandedramblings.com/index.php/zoom-statistics-facts

Data shows that there are more apps for ordering food products and for other types of shopping. The delivery economy is growing, especially in large cities. Online to offline food delivery has facilitated consumer access to prepared meals and enabled food providers to keep operating (Li, Mirosa, and Bremer, 2020).

Location aware apps are assisting with contact tracing. Healthcare professionals and patients have embraced telemedicine and app-based monitoring of wellness information. The findings of a recent study suggest that "telemedicine and virtual software are capable of decreasing emergency room visits, safeguarding healthcare resources, and lessening the spread of COVID-19 by remotely treating patients during and after the COVID-19 pandemic" (Bokolo, 2020, p. 1). Hakim, Kellish, Atabek, et al (2020) suggest that telemedicine and virtual software platforms may be helpful to manage the pandemic. In this crisis, information technology has opened a new frontier in mental health support and data collection. There are thousands of mental health apps available in the iTunes and Android app stores. Torous, Myrick, Rauseo-Ricupero, et al (2020) suggest that increasing investments in digital mental health today will improve access to mental health care in the future.

Rapid adoption of thermal imaging and video surveillance systems monitor temperatures, enforce mask restrictions, and detect social distancing violations, cf., Lopeztello and Wulffson (2020). Sensors for temperature checks are more widely used in airports and office buildings. There is greater interest in Internet of Things (IoT) enabled sensors and devices that are used to increase the efficiency of the appliances in a smart building and make it more efficient, sustainable, and safer. IT is increasingly used for automatically controlling heating, ventilation, air conditioning, lighting, security and other systems of a building.

Barcoding allows vaccine information to be documented in an Electronic Health Record (EHR) instantly and accurately. Immunization Information Systems (IISs), otherwise known as immunization registries, are confidential, population-based, computerized databases that record all immunization doses administered by participating providers to persons residing within a given geopolitical area. They offer an opportunity for confidential, secure, centralized, and immediate access to immunization records for authorized providers.

More seniors over age 65 are adopting information technologies. Etkin (2020) asserts "In 2020, many older adults' own devices with Internet capabilities and are able to use them to video chat with family members and friends, order groceries, consume content online and even exercise." Families are having FaceTime and Zoom gatherings. Virtual events have provided some new social activities.

3. Impacts on Organizations

Health organizations are increasing the use of IT. For example, bioinformatics is an important, interdisciplinary field that develops methods and software tools for understanding biological data, especially large and complex data sets. IT is used for gene sequencing and genetic engineering of vaccines. McCullers and Dunn (2008) noted "The introduction of genetic engineering has fueled rapid advances in vaccine technology and is now leading to the entry of new products in the marketplace." Increased adoption of health information technologies has been a major consequence of the health crisis. Health IT provides many opportunities for improving and transforming healthcare. Adoption of health IT is reducing human errors, improving clinical outcomes, facilitating care coordination, improving practice efficiencies, and collecting and tracking data over time, cf., https://www.ncbi.nlm.nih.gov/pmc/.

Robotic process automation is changing workflows, including document generation and payment processing. Industrial robots and automated manufacturing are changing production and distribution processes.

There is greater use and acceptance of Artificial Intelligence. According to McKendrick (2020), "KPMG is applying AI approaches to rapidly analyze contractual obligations and termination clauses, as industries face supply chain delays, cancelled events and other roadblocks. ... KPMG also reports developing AI-based tools to supplementing employee and customer call centers to analyze and triage issues and questions."

More government services are provided online rather than in person. For example, online portals, increased social media use, AI and robotics, cf., Charlton, 2020. Governments have employed digital platforms, and big data analytics.

Organizations are adopting many information technologies including online health care; blockchain-based epidemic monitoring platforms; robots that deliver food and medications and that screen people's temperatures; online education platforms and home-based working solutions; and robotics and 3D-printing technologies to manage social distancing in manufacturing plants.

The impacts of COVID-19 on organizations are deep, broad, and potentially permanent. The impact is not limited to the increase in the use of technology and how employees continue doing their jobs. The pandemic is impacting consumer behavior, marketing, human resource management, competition, supply chains, and possibly even corporate social responsibility and sustainability.

As the pandemic forced organizations' employees to work from home, their consuming habits changed. No longer could people easily go to restaurants at their work locations or close by. Their shopping habits changed. Online shopping significantly increased and this could become a permanent habit for many consumers. Shops have to modify their marketing strategies to fit the consumers' behavior during and possibly even after the pandemic. Organizations need to be aware of implications for their human resources activities including leave policies, remote work policies, technology training, and mental health issues. The pandemic has significantly increased competition between online and brick and mortar entities. Due to altered consumer habits, the practice of online shopping could potentially remain very strong even after the pandemic. We have all experienced supply chain issues for many essential products. Let us hope that companies learned a lesson to incorporate newer technologies in their supply chain operations and move towards smart supply chain platforms. Let us also hope that at least some good will come out of this devastating pandemic in areas such as corporate social responsibility and sustainability. He and Harris (2020) clearly articulate the potential movement of organizations toward more robust corporate social responsibility and sustainability.

4. Conclusions

Clark (2020) notes "When we look back on the current health crisis, there's no doubt that we'll learn that it resulted in a number of innovations: new drugs and medical devices, improved healthcare processes, manufacturing and supply chain breakthroughs, novel collaboration techniques." He argues "crisis demands movement and change – the pace of ideation, decision making, and implementation all increase dramatically." The increased use of information technology is likely a permanent change. Our new normal will be a mix of both face-to-face and information technology-mediated activities. Better supply-chains, more automation, and Artificial Intelligence will create a more flexible economy.

Information technologies have helped people and organizations continue to function. IT helps us to work, to socialize, shop, and live. Experience with these technology adaptations and supports has demonstrated that to find a new normal we need more and better IT solutions, more technology literacy, better public health surveillance, and better preventative measures to minimize harms from any future health crises. In the future, more people will likely work and learn remotely, and organizations and governments must upgrade their digital capabilities and the skills of employees.

As a species, we have an extreme dependence upon human contact which is both an economic and social strength and weakness. We have observed how an infectious disease can spread by human contact. The deadlier the disease the more we need to isolate, quarantine, and use Information Technology intermediation. At some point the disease will be "controlled", but we cannot return to the processes and behaviors of the past. Businesses, schools, and colleges are reopening, but life will never return to the old normal. Society will fashion a new normal that makes greater use of IT.

Unfortunately, the impact of the COVID-19 pandemic will last a very long time. Let us hope that during these terrible times of death, sickness, and lost livelihoods, we have all learned a significant lesson. Information Technology is important and relevant for the well being of all of us.

5. Overview of the Contents of this Issue

This issue of the journal includes three traditional research articles and a tutorial.

Joey George and Alastair Robb, in a very timely article, look at deception detection in digital communication among people. In particular, they look at individuals in the US and Australia to determine if people in one culture can detect deception of people in the other. They look at full audiovisual, video only, audio, and text communication. Their study finds that both Americans and Australians can detect deception at approximately the same rate across both cultures. Further, their study finds that individuals were able to more accurately detect deception for full audiovisual types of communication.

John Muraski, Jacob Iverson, and Kimberly Iverson, look at the challenges organizations are facing in finding skilled IT talent. Specifically, they look at the Northwest Wisconsin region and the creation of a "New Digital Alliance." This innovative initiative is funded by local companies and universities. The main purpose of this collaborative alliance is to attract, develop, and retain IT talent for the Northern Wisconsin region. This collaborative network could potentially serve as a model for other regions in the country.

Loknath Sai Ambati and Omar El-Gayar investigate the performance of machine learning (ML) techniques used in human activity recognition (HAR). They specifically look at the more commonly used approaches of Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Stochastic Gradient Descent, Decision Tree, Decision Tree with entropy, Random Forest, Gradient Boosting Decision Tree, and NGBoost procedures. This study evaluates the ML techniques using accuracy, precision, recall, F1 score, support, and run time performance measures on several HAR datasets. The authors highlight the importance of adopting an appropriate ML technique based on the specific HAR requirements and the characteristics of the associated HAR datasets.

Vlad Krotov and Matthew Tennyson in their tutorial demonstrate how to use "rvest" and "xm12" packages for Web Scraping. Specifically, they use simple examples to demonstrate how these R-based packages can be used to retrieve data from sites such as Bayt.com which is a prominent employment Web site in the Middle East.

We appreciate and wish to acknowledge the contributions of reviewers for this issue of the journal, including Gaurav Bansal (University of Wisconsin, Green Bay), David Biros (Oklahoma State University), Queen Booker, (Metropolitan State University), Mari Buche (Michigan Technological University), Sean Eom (Southeast Missouri State University), Joey George (Iowa State University), Yi "Maggie" Guo (University of Michigan, Dearborn), Bryan Hosack (Penske Logistics), Jakob Iverson (University of Wisconsin, Oshkosh), Anoop Mishra (University of Nebraska, Omaha), Alanah Mitchell (Drake University), Shana Ponelis (University of Wisconsin, Milwaukee), Kevin Scheibe (Iowa State University), and Neetu Singh (University of Illinois, Springfield).

6. References

Bokolo, A., (2020). Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 Pandemic. *Irish Journal of Medical Science*, https://doi.org/10.1007/s11845-020-02299-z

Charlton, E., (2020). How governments are communicating online during the COVID-19 crisis. World Economic Forum, May 5, 2020 at URL https://www.weforum.org/agenda/2020/05/how-coronavirus-is-accelerating-the-move-to-digital-government/

Clark, L., (2020). Innovation in a Time of Crisis. Harvard Business Publishing, March 26, at URL https://www.harvardbusiness.org/innovation-in-a-time-of-crisis/

Etkin, K., (2020). OPINION: Technology's Impact on the Pandemic. Nextavenue, April 13, at URL https://www.nextavenue.org/technology-impact-pandemic/

Hadidi, R. and D. J. Power, (2020). Technology Adoption and Disruption -- Organizational Implications for the Future of Work. *Journal of the Midwest Association for Information Systems* | Vol. 2020, Issue 1, pp. 1-7.

Hakim, A. A., Kellish, A. S., Atabek, U. et al. (2020). Implications for the use of telehealth in surgical patients during the COVID-19 pandemic. *The American Journal of Surgery*, 220(1), 48-49. https://doi.org/10.1016/j.amjsurg.2020.04.026

He, H. and Harris, L., (2020). The impact of Covid-19 pandemic on corporate social responsibility and marketing philosophy. *Journal of Business Research*, Volume 116, pp. 176-182. https://doi.org/10.1016/j.jbusres.2020.05.030

Li, C., Mirosa, M., & Bremer, P. (2020). Review of online food delivery platforms and their impacts on Sustainability. *Sustainability*, 12(14), 5528. doi: http://dx.doi.org/10.3390/su12145528

Lopeztello, A. and Wulffson, T., (2020). Thermal Cameras to Fight Coronavirus in the Workplace. *Security Magazine*, August 12, at URL https://www.securitymagazine.com/articles/93037-thermal-cameras-to-fight-coronavirus-in-the-workplace

McKendrick, J. (2020). 3 Ways The Covid-19 Crisis Has Opened Minds About Technology. Forbes, April 17 URL https://www.forbes.com/sites/joemckendrick/2020/04/17/3-ways-the-covid-19-crisis-has-opened-minds-toward-technology/#1ca9426a174a

McCullers, J. A., & Dunn, J. D. (2008). Advances in vaccine technology and their impact on managed care. *Pharmacy and Therapeutics:* a peer-reviewed journal for formulary management, 33(1), 35–41 at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730064/

Messenger, J., Z. Vadkerti and A. Uhereczky, (2020). Practical Guide on Teleworking during the COVID-19 pandemic and beyond. International Labour Organization (ILO), July 16, at URL https://www.ilo.org/travail/whatwedo/publications/WCMS_751232/lang--en/index.htm

Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable, Random House, ISBN 978-1400063512

Torous, J., Jan Myrick, K., Rauseo-Ricupero, N. and Firth, J. (2020). Digital mental health and COVID-19: Using technology today to accelerate the curve on access and equity tomorrow. *JMIR Mental Health* 7(3): e18848.

Author Biographies

Daniel J. Power is a Professor of Information Systems and Management in the College of Business Administration at the University of Northern Iowa. Power is the Founding Editor of the *Journal of the Midwest Association for Information Systems* (JMWAIS), He also edits DSSResources.com, PlanningSkills.com, and Decision Support News. His research interests include the design and development of decision support systems and how these systems impact individual and organizational decision behavior.

Rassule Hadidi is Dean of the College of Management, Metropolitan State University, Minneapolis, Minnesota. His current research areas of interest include online and blended teaching and learning pedagogy and its comparison with face-to-face teaching; curriculum development and quality assessment; cloud computing and its applications for small and medium-sized enterprises; and quality of online information. He has served as the president as well as the At-Large Director of the Midwest Association for Information Systems and is the founding Managing Editor of the *Journal of the Midwest Association for Information Systems*. He is a member of the Board of Directors of the Society for Advancement of Management.

This page intentionally left blank.

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 2

Date: 01-31-2021

Comparing Australians' and Americans' Abilities to Detect Deception across Cultures and Communication Media

Joey F. George
Iowa State University, jfgeorge@iastate.edu

Alastair Robb
University of Queensland, a.robb@business.uq.edu.au

Abstract

The reach of global communication is expanding through the growing availability of smartphones. Smartphones are particularly popular for texting and voice/video calls, and their affordability means that more and more people around the world can now communicate with each other. Yet with the spread of global communication also comes increased exposure to deceptive communication. Can people in one culture accurately detect deception across cultures? And does the communication media they use play a role in their detection accuracy? We attempt to answer these two research questions in a study of Australian and US judges who were asked to detect deception in Australians and Americans, across four different media: full audiovisual, video only, audio only, and text. We found that both Australians and Americans could accurately detect deception at about the same rate across both cultures, and they were better at detection when exposed to full audiovisual stimuli compared to text.

Keywords: Computer mediated communication; deceptive communication; national culture

DOI: 10.17705/3jmwa.000063

1. Introduction

The Pew Research Center reports 81% of Americans owned smartphones in 2018 (Taylor & Silver, 2019), and the three most popular uses for them were text messaging, internet use, and voice/video calls (Smith. 2015). More than 5 billion people worldwide had mobile devices in 2018, and over half of these were smartphones. Further, while people in developed countries surveyed were more likely to have smartphones (76%), almost half of those in developing countries had them (45%) (Taylor & Silver, 2019). While internet use is ubiquitous in developed countries, its use is growing rapidly in emerging economies (ITUNews, 2018). In short, the world is becoming increasingly connected through smartphones, and aside from internet access, the primary purpose of smartphones is communication. Accordingly, people are able to communicate with each other all over the world, in real time and at relatively low cost.

Everyday normal communication includes a deceptive component (DePaulo, Kashy, Kirkendol, Wyer & Epstein, 1996), whether that communication takes place in face-to-face encounters or over the phone. While most deceptive communication research has been conducted in North America, there is no reason to believe international cross-cultural communication will be any less deceptive. While it is widely recognized that on average people are able to successfully detect deception only about 54% of the time (Bond & DePaulo, 2006), not as much is known about how well people are able to detect deception in cultural groups other than their own. Only five published studies have explicitly investigated deception detection across cultures (Al-Simadi, 2000; Bond & Atoum, 2000; Bond, Omar, Mahmoud, & Bonser, 1990; Castillo, Tyson, & Mallard, 2014; George, Gupta, Giordano, Mills, Tennant & Lewis, 2018). Four of the five studies found that people could accurately detect deception in other cultural groups. These studies included seven national cultures, as defined by Hofstede (1980): Americans (USA), Australians, Colombians, Indians, Jordanians, Malaysians, and Spaniards.

Bond and colleagues (1990) argued that there were two ways to look at the question of culture and deception detection: a "universal cue hypothesis" and a "specific-discrimination hypothesis." The former posits that, given the universal nature of deception, people should be able to detect it just as easily in other cultures as in their own. The latter argues that deception and its detection are both learned and hence depend heavily on the cultural context and language in which that learning takes place. Hence people should have a difficult time accurately detecting deception across cultures. The evidence to date supports the universal cue hypothesis, but the evidence is limited. Additional cultures and the deception detection abilities of their members need to be studied for a more complete understanding of the relationship between culture and deception detection. While it is impractical, and likely implausible, to compare all national cultures to each other, it is possible to compare a few. To date, all studies of deception detection across cultures have featured members of cultures that differ a great deal from each other. This study focuses on two national cultures that are very similar, people living in the US and people living in Australia. By establishing baselines, based on extreme differences and on extreme similarities, we will be able to make inferences about the how detecting deception across a host of cultures might compare.

As the statistics on smartphone and internet use show, this increasing interconnectedness of people of different cultures is a direct result of increased use of communication and information technologies. We know a great deal about deception and its detection in real-time face-to-face environments, we know less about the relationship between detection and the computer-mediated communication mode (e.g., skype, SMS texting, voice-over-IP, email) over which the deceptive communication takes place. While several studies have found media differences between face-to-face and computer-mediated communication (e.g., Dunbar, Jensen, Burgoon, Kelley, Harrison, Adame, & Bernard, 2015; Van Swol, Braun & Kolb, 2015), the evidence of a direct relationship between media and detection accuracy in computer-mediated modes is limited (Burgoon, Stoner, Bonito, & Dunbar, 2003; Burgoon, Blair, & Strom, 2008; Dunbar, Jensen, Tower, & Burgoon, 2014; George, Marett & Tilley, 2008; George, Tilley & Giordano, 2014; Hancock, Woodworth, & Goorha, 2010; McHaney, Gupta & George, 2018; Rockmann & Northcraft, 2008; Zhou & Zhang, 2007). However, given the differences between face-to-face communication and email or smartphone-based videoconferencing, it follows that the accuracy of deception detection would depend on the medium over which it is conveyed.

To further investigate the issues of culture, media and deception detection, we created a stimulus set that mixed honesty and dishonesty, as well as media, using individuals from Australia. We had already created such a stimulus set for US English. Altogether, we have created and tested five such stimulus sets as part of a multi-year program of study. (The other three are Indian English, Spanish, and Hindi.) We then exposed Australian and American judges to the Australian and US English stimulus sets. This comparison of American and Australian stimulus sets is novel, as this is the first

cross-cultural study dealing with deception detection that we are aware of that compares two similar national cultures. Our research questions are: 1) Can individuals of one culture accurately detect deception in individuals from a similar but different culture? and 2) Is there a relationship between deception detection and media? The rest of the paper is organized as follows: First we review the literature on deception, media and culture. We then present our research methods, our findings, and a discussion of the implications of our results.

2. Literature Review

We define deception as "a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver" (Buller & Burgoon, 1996, p. 205). In general, people are not very good at detecting deception, with an accuracy rate of around 54%, barely better than chance (Bond & DePaulo, 2006). While deceptive communication has been studied for decades, much of what we know has been learned in a North American context, based on dyadic real time communication. Until recently, neither communication media nor national culture were key aspects of the study of deception and its detection. In the next sections, we review the relevant literature on deception and culture and on deception and media. But first we present a brief primer on culture and its dimensions.

2.1 Culture

National culture is defined as "the collective programming of the mind which distinguishes the members of one human group from another" (Hofstede, 1980). Hofstede (1980) was one of the first to study national culture and to determine its specific dimensions. There are currently six dimensions (Hofstede, 2016): 1) power distance, 2) individualism vs collectivism, 3) masculinity vs femininity, 4) uncertainty avoidance, 5) long-term vs. short-term orientation, and 6) indulgence vs restraint. Power distance is a measure of how tolerant people are of the unequal distribution of power in a society. Individualism vs collectivism reflects the extent to which a society is tightly- or loosely-knit. Masculinity vs femininity has less to do with gender than it does with whether a society is based on competition and assertiveness or cooperation and nurturing. Uncertainty avoidance, as the name suggests, is the extent to which members of a society can tolerate uncertainty. Long- vs. short-term orientation is a measure of the extent to which societies prepare for the future. Finally, indulgence, the newest dimension, reflects the allowance of free gratification of human drives, focusing on enjoying life and having fun, while restraint leads to the suppression of such gratification. Hofstede's original studies (Hofstede, 1980) were conducted between 1967 and 1973 and resulted in the definition of the original four dimensions.

2.2 Differences in Deception across Culture

Several studies have reported differences in how deception is viewed across cultures. Although most of these studies have focused on differences between Western and Eastern cultures (especially East Asian cultures), cultures from all over the world have been investigated. Table 1 provides a sampling of this work.

Table 1. A sampling of study results related to differences in deception across cultures.

Study	Countries	Select Findings
Triandis et al 2001	Multiple national cultures	In business negotiations, members of collectivist cultures are more dishonest than members of individualistic cultures
Seiter & Bruschke 2007	China & US	Americans experienced more guilt over lying than Chinese participants.
Fu et al. 2011	China & US	Chinese participants perceived lying more favorably than Americans for modest behavior
Bessarabova 2014	Russia & US	Russians lied more often than Americans to help the underperforming in-group members.
Hamilton & Kirwan 2013	Ireland & US	Online dating profiles of Irish males were found more deceptive than American profiles.
Banai et al. 2014	Israel & Kyrgyzstan	Kyrgyzstanis more likely to endorse ethically questionable negotiation tactics (i.e., pretending, deceiving, & lying) than Israelis.

The research shows that what some see as deceptive, and hence inappropriate, others might see as a perfectly acceptable practice. These findings imply that accurate detection of deception across cultures might be difficult. Accurate detection becomes more complicated if there is a lack of agreement on what constitutes deception in the first place.

2.3 The Universal Cue Hypothesis

Only five published studies have explicitly investigated the relationship between deception detection and culture (Al-Simadi, 2000; Bond & Atoum, 2000; Bond, et al., 1990; Castillo, et al., 2014; George et al, 2018). The first study (Bond, et al., 1990) found that people were not able to accurately detect detection in other cultural groups. The authors argued their findings implied support for the "specific-discrimination hypothesis," whereby deception and its detection were dependent on cultural and language-based behaviors that people learned as they learned to communicate. Such a conclusion would have been expected in light of differing cultural definitions of deception, as illustrated by Table 1. The other studies all found that people could indeed accurately detect deception across other cultures, supporting the universal cue hypothesis, that everyone regardless of culture engages in similar behaviors during deception. These behaviors would be universally regarded as indicative of deception. Further, three studies (Al-Simadi, 2000; Castillo, et al., 2014; George et al, 2018) found that in some cases, people were even better at detecting deception in other cultural groups than in their own. The cultural comparisons conducted in these four studies are shown in Table 2.

Table 2. Cultural comparisons across four studies that found evidence of the ability to successfully detect deception across cultures.

	USA	Australia	Spain	India	Jordan	Malaysia	Colombia
USA	-----			x [1]			
Australia	x [current]	-----					x [3]
Spain	x [4]		-----				
India	x [4]			-----	x [1]		
Jordan	x [1]				-----		
Malaysia					x [2]	-----	
Colombia							-----

Studies: [1] Bond & Atoum, 2000; 2] Al-Simadi, 2000; [3] Castillo et al 2014; [4] George et al 2018

Across the available empirical evidence from these studies, we see three patterns (Table 3). Some groups of judges have been better at detection deception in their own group than in others; some have been better with members of other cultures; and for one group, all three sets of judges were equally accurate. In general, based on the findings from these studies, the universal cue hypothesis seems to hold – people can accurately detect deception in their own group, and in other cultures, apparently using similar indicators of deception. Given these findings, we would expect that the answer to our first research question -- Can individuals of one culture accurately detect deception in individuals from another culture? -- would be affirmative. That leads to our first hypothesis:

Hypothesis 1: Members of a national culture will be able to accurately detect deception among members of their own culture and in members of other cultures.

Table 3. Patterns of comparative deception detection success across cultures in four studies.

Better in own group (compared to other groups)	Better with other group (compared to their own group)	No differences
Americans (judging Americans & Jordanians) (Bond & Atoum, 2000)	Jordanians (judging Jordanians & Malaysians) (Al-Simadi, 2000)	Indians, Americans & Jordanians (judging Indians) (Bond & Atoum, 2000)
Jordanians (judging Americans & Jordanians) (Bond & Atoum, 2000)	Malaysians (judging Jordanians & Malaysians) (Al-Simadi, 2000)	
Spaniards (judging Spaniards & Americans) (George et al, 2018)	Australians (judging Australians & Colombians) (Castillo et al, 2014)	
Indians (judging Indians & Americans) (George et al, 2018)	Americans (judging Americans, Spaniards & Indians) (George et al, 2018)	

But the findings reveal some interesting comparative outcomes. In the case of members of two particular cultures, Americans and Jordanians, individuals sometimes did better with their own culture, and sometimes they did better with other cultures. Why is that the case? It could be due to different stimulus materials across studies, or it could be due to differences in the cultures that were compared. For example, American judges did better with their own group when compared to Jordanians (Bond & Atoum, 2000), and they did better with the other groups when compared to Spaniards and Indians (George et al, 2018). Based on Hofstede's measures, the seven national cultures that have been studied in this context differ dramatically from each other on some cultural dimensions but not on others (Table 4). For example, Colombia and Australia differ widely on power distance, individualism/collectivism, and uncertainty avoidance, but their scores are very similar for femininity/masculinity (both are masculine) and short/long-term orientation (both tend towards short-term). Jordanian culture tolerates a more unequal distribution of power in society than does Spanish culture, and Jordanians are more collectivist, more tolerant of uncertainty, and more short-term oriented than Spaniards.

Table 4. Cultural Dimension Scores from geert-hofstede.com/dimensions.html (sorted by scores on Individualism/Collectivism, from most individualist to most collectivist).

Cultural Dimension	USA	Australia	Spain	India	Jordan	Malaysia	Colombia
Power Distance (PD)	40	36	57	77	70	100	67
Individualism/ Collectivism (IC)	91	90	51	48	30	26	13
Masculinity/ Femininity (MF)	62	61	42	56	45	50	64
Uncertainty Avoidance (UA)	46	51	86	40	65	36	80
Long Term Orientation (LT)	26	21	48	51	16	41	13

To chart the differences across these cultures, we devised a simple (and somewhat crude) measure, which we call the Hofstede score difference index. In comparing two cultures, we subtract the Hofstede scores for one culture from the other for each dimension, and we sum the absolute values of those differences. The results, for past cultural comparisons, plus the difference in Australian and US cultures, are shown in Table 5 and Figure 1.

Table 5. Hofstede score difference index for cultural comparisons (possible range: 5 – 500).

USA-AUS	Malaysia-Jordan	India-Jordan	USA-Jordan	USA-Spain	USA-India	AUS-Colombia
16	93	96	117	137	139	148

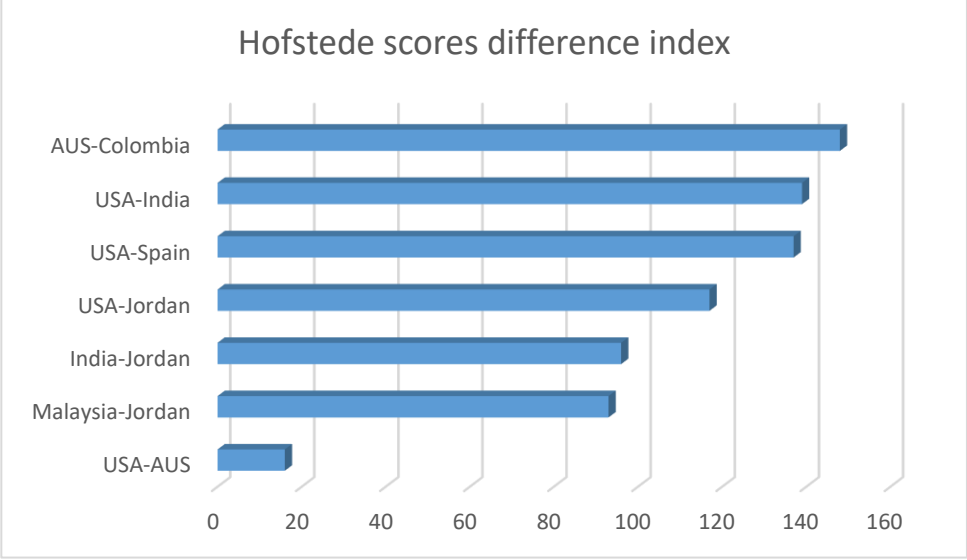


Figure 1. Hofstede score difference index for cultural comparisons

Out of this set of national cultures, the two cultures that are most different are Australians and Columbians. However, no two are more similar than Australia and the U.S. Both are low power distance, highly individualistic, masculine, short-term orientation countries, and both have mid-level scores for uncertainty avoidance. Whether the specific-discrimination or universal cue hypothesis holds, due to the close similarity of their cultures, Australians and Americans should be equally able of accurately detecting deception both within and across their cultures. Thus, we hypothesize:

Hypothesis 2: Australian and US judges will be able to detect deception equally well both within and across their cultures.

2.4 Deception & Media

The past decade has seen a dramatic increase in the number of studies investigating the relationship between media and deception detection. Most of these studies have investigated one medium at a time, with a particular focus on the development of an automated tool for detecting deception within a specific medium (see Zhou, Burgoon, Zhang, & Nunamaker, 2004). Another set of studies has investigated the direct relationship between media and detection by comparing multiple media in a single study. The findings from these studies have been mixed. While some have found direct effects (Burgoon, et al., 2003; Burgoon, et al., 2008; Dunbar, et al., 2014; Zhou & Zhang, 2007), others have found evidence of a mediated relationship (George et al, 2008; George et al, 2014; Hancock, et al., 2010; Rockmann & Northcraft, 2008).

There has also been evidence of media differences in the few studies that have investigated media and culture. In fact, the key difference between Bond's 1990 and 2000 studies was that the experimental stimuli had no sound in the former study, while those in the latter study did have sound. The availability of sound helps explain the differences in findings between the studies -- Bond concluded that people could not detect deception across cultures after the 1990 study, but he concluded the opposite after the 2000 study. George and colleagues (2018) also found evidence of media effects. Looking at the relationship between media and deception detection accuracy, they found that veracity judges were less successful at deception detection when judging video-only communication, compared to full audiovisual, audio-only, and text communication. Given this pattern of findings, even with the few studies that have been conducted, we would expect to find differences in detection accuracy, depending on media. But which media should be best for detecting deception?

Leakage theory asserts that deception is cognitively and emotionally complicated, making the process difficult to control, so deceivers often leak cues in the form of verbal and non-verbal behaviors (Ekman 1985; Ekman and Friesen, 1969). The leakage of cues is what allows deception to be detected at all. If those being deceived are observant to the verbal and non-verbal behaviors of others, deceivers stand a better chance of getting caught. According to such media theories as Media Synchronicity Theory (Dennis, Fuller & Valacich, 2008), different media have different capabilities. These capabilities include transmission velocity, symbol set variety, parallelism, rehearsability, and reprocessibility. For example, face-to-face communication should be high in transmission velocity and symbol set variety and low in parallelism (the capability to send and receive messages across multiple channels simultaneously), rehearsability (the capability to carefully plan and edit a message before sending), and reprocessibility (the capability to examine a message carefully as much as needed). Two-way SMS texting, at the other extreme, would be moderate in transmission velocity, low in symbol set variety and parallelism, and high in both rehearsability and reprocessability. Different media capability combinations should render some media able to transmit more cues to deception, compared to others. Compared to texting, face-to-face communication (or its electronic equivalent, videoconferencing) should provide more cues to deception, given that texting can transmit verbal communication only. Rao and Lim (2000) linked a medium's capability to transmit the maximum number of cues to more success in deception detection (Table 6). A medium's capability to transmit a variety of cues influences the accuracy of deception detection, such that the availability of more cues should be associated with more accurate deception detection.

As shown in Table 6, for the 14 cues to deception that are listed, all 14 can be detected in full audiovisual media, such as videoconferencing. Nine can be detected in audio; seven can be detected in written media; and five can be detected on video-only media. If detection accuracy is improved when more cues to deception are available, which leakage theory implies, then the use of full audiovisual media should result in the most accurate deception detection. Hence:

Hypothesis 3: Media that can transmit more cues to deception will be associated with more accurate deception detection, compared to media that transmit fewer cues.

Table 6. Cues to deception across various media (from Rao and Lim 2000)

Behavior	Audio video	Video Only	Audio Only	Written Media
Visual				
Pupil dilation	Detectable	Detectable		
Blinking	Detectable	Detectable		
Facial segmentation	Detectable	Detectable		
Adaptors	Detectable	Detectable		
Bodily segmentation	Detectable	Detectable		
Paralanguage				
Response length	Detectable		Detectable	Detectable
Speech errors	Detectable		Detectable	Detectable
Speech hesitations	Detectable		Detectable	
Pitch	Detectable		Detectable	
Verbal				
Negative statements	Detectable		Detectable	Detectable
Irrelevant information	Detectable		Detectable	Detectable
Immediacy	Detectable		Detectable	Detectable
Leveling	Detectable		Detectable	Detectable
General Discrepancy	Detectable		Partially detectable	Partially detectable

3. Research Methods

The study had two primary aspects, the creation of the stimulus materials and the experimental sessions in which judges were asked to determine the veracity of the stimulus materials. Both aspects are described below.

3.1 Stimulus Materials and Measures

The two stimulus sets used in this study were created in a similar fashion (Figure 2). In both cases, students were asked to attend an experimental session, where they would be interviewed, and to bring along a personal résumé. The sessions were held in the same rooms recruiters would use for job interviews. They were met by an experimenter, who reviewed their résumés and asked them to complete an application for a fictional scholarship that might be offered by their college. They were told it was all right to make themselves look good on the application, but they were not asked to be dishonest. The researcher then collected the applications, which were provided to an interviewer. For the US English stimulus set, each student was interviewed by another student via VoIP. For the Australian English stimulus set, the researcher who greeted the student conducted the interview. In both instances, students were asked about the information on the applications, regardless of whether it was true or not. In many cases, students then had to defend information they knew was false.

All interviews, of 20 US students (from the panhandle of Florida) and 21 Australian students (from Queensland), were recorded. Researchers reviewed all of the recorded material in order to create the stimulus sets. The résumés acted as ground truth, so by comparing the contents of the résumé to the application, researchers could tell what information on the applications was false. The researchers were looking for parts of each interview that were false and for parts that were true. They selected 16 snippets, half of which were true and half of which were false. These snippets were selected from the set of all 20 (USA) or all 21 (AUS) recorded interviews. The final stimulus set for each national culture consisted of a total of 32 recorded snippets. In addition to balancing them in terms of honesty, the researchers also balanced them in terms of media. Eight snippets were selected for each of four media representations: full audiovisual, video-only, audio-only, and text. For video-only snippets, the audio portion of the snippet was removed. For audio-only, the video portion was removed. The text snippets were transcribed from the interviews. The 32 snippets were then randomly

ordered to complete the stimulus set. The US English stimulus set was created for a doctoral dissertation (Lewis, 2009), based on recordings created in an earlier study (Tilley, 2005). The Australian English stimulus set was created by the authors during the summer of 2016 at a major Australian university. Each stimulus set was used to create a questionnaire in Qualtrics.

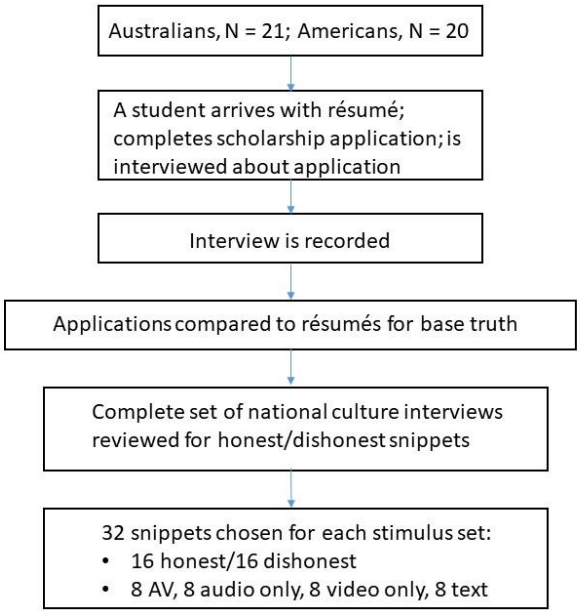


Figure 2. Flowchart of stimulus materials preparation

3.2 Veracity Evaluation

A different set of students was recruited to be veracity judges in a set of separate experimental sessions, later in 2016. A total of 36 undergraduate students enrolled in the business school of a major Australian university, and 40 undergraduate students at a Midwestern American business school, were recruited to judge the veracity of the stimulus materials (Figure 3). Half were randomly assigned to view the US English stimulus set, and half observed the Australian English set. Of the 36 participants at the Australian university, 19 self-identified as Australian. Seven of those were exposed to the US English stimulus set; 12 were exposed to the Australian stimulus set. At the US university, five students did not self-identify as Americans. Of the 35 who did, 16 were exposed to the US English stimulus set; 19 were exposed to the Australian set. It was important to the study, where the veracity judges needed to be representative of a particular national culture, that only the judgments of students who identified as either Australian or American be analyzed.

Each judge was asked to watch, listen to and/or read each of the 32 snippets in one or the other stimulus set and to then indicate the veracity of each on a 7-point scale that ranged from '1' for 'very honest' to '7' for 'very dishonest.' If the participant selected 5, 6 or 7, the next window that opened asked the participant to describe why he or she believed the snippet to be dishonest. Participants averaged about one hour in completing this part of the study. They were asked to answer a series of questions that measured five of Hofstede's cultural dimensions (Hofstede, 1980; 2008; Hofstede, Hofstede, Minkov, & Vinken, 2008; Srite & Karahanna, 2006). The dimensions were measured on a 5-point scale, varying from '1' for 'strongly agree' to '5' for 'strongly disagree.'

4. Analysis & Results

Before we present the analysis of detection accuracy across cultures, using repeated measures logistic regression, we present the results of measuring the cultural dimension scores of both Australian and American study participants.

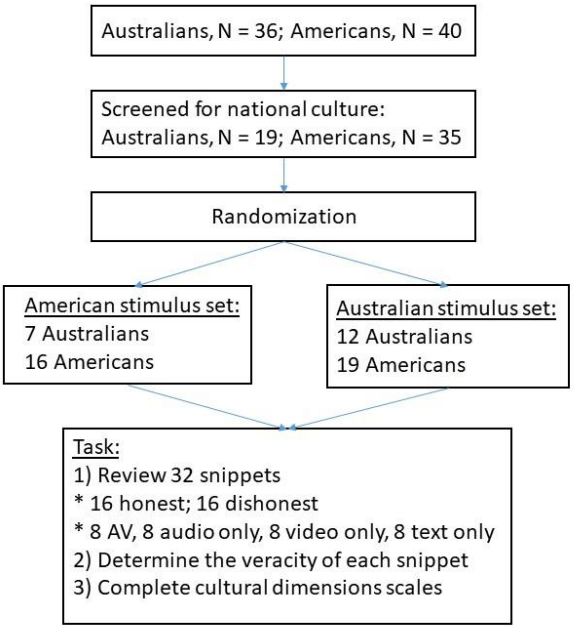


Figure 3. Experimental flowchart

4.1 Tests for Hofstede's Cultural Dimensions

Culture is a macro-level construct, so it often lacks precision in explaining individual-level behavior (Srite & Karahanna, 2006). Accordingly, it is inappropriate to use country scores, as developed by Hofstede (1980), or an individual's national citizenship, to predict individual behavior based on cultural values and beliefs (Furner & George, 2012; Straub, Loch, Evaristo, Karahanna, & Srite, 2002). Therefore, we measured cultural dimensions for each individual study participant.

We used measurement items published by Srite and Karahanna (2006) to measure Hofstede's dimensions. To test for the adequacy of four of the scales, we used factor analysis. We used exploratory factor analysis because of past psychometric issues with the scales (Lewis, 2009; Furner & George, 2012; George et al, 2018). There were serious problems with the six items used to measure uncertainty avoidance, and four other items were problematic, so all 10 of these items were dropped. The factor analysis results, based on varimax rotation and three forced factors, are shown in Table 7. Once the problematic items had been pruned, the reliabilities for the remaining items were reasonable (MF: .782; PD: .557; IC: .690). We averaged the remaining items for each scale to compute scores for the dimensions. We did not measure indulgence vs restraint. For the long- vs short-term orientation dimension, we used the following formula, where the m variables refer to items in the scale (Hofstede, 2008; Hofstede, et al., 2008) and C refers to a constant:

$$LT = 40(m18 - m15) + 25(m28 - m25) + C(ls)$$

Table 8 reports the US and Australian scores for the four cultural dimensions we could measure. Australians and Americans participants scored virtually the same on these four cultural dimensions, underlying the similarity of the cultures. One way ANOVA tests showed no statistically significant differences between Australians and Americans for each dimension. The first three dimensions were measured on a five-point scale. For individualism/collectivism, higher numbers indicate individualism, so the scores for both nationalities are in line with Hofstede's measures. For masculinity/femininity, lower numbers are masculine, so again, the results are consonant with expectations. The scores on long/short term orientation are also in line with expectations, towards the short-term orientation end of the scale. The only questionable scores are for power distance, which should be higher, indicating less tolerance for inequality. Despite the latter finding, the sample of Australians and Americans seems to be largely representative of what Hofstede's measures (Hofstede, 2016) indicate they should be.

Table 7. Factor analysis results for four of Hofstede's cultural dimensions for the Australian sample

	1	2	3
MF3	**.794**	.107	.276
MF2	**.793**	.092	.013
MF1	**.737**	.126	.147
MF5	**.727**	-.013	.108
IC3	-.014	**.757**	.173
IC5	-.186	**.661**	.158
IC1	.247	**.630**	-.006
IC2	.392	**.605**	-.151
IC4	.013	**.592**	-.317
PD3	-.087	-.291	**.641**
PD5	.262	.184	**.597**
PD1	.023	.016	**.563**
PD7	.330	.396	**.554**
PD2	.295	-.026	**.524**

Table 8. Measured scores on Hofstede's cultural dimension for Australians and Americans.

Cultural Dimension	Australia	USA
Individualism/Collectivism (IC)	2.95	3.02
Masculinity/Femininity (MF)	1.81	1.87
Power distance (PD)	2.07	2.12
Long Term Orientation (LT)	39.41	59.56

4.2 Tests for Relationships between Deception and Culture and Media

We had three hypotheses. The first predicted that representatives of a national culture would be able to successfully detect deception in both members of their own culture and in members of other cultures. As we have seen, people tend to be only as good as chance at accurate deception detection, so a score of 50% or higher indicates successful detection. No statistical test is needed to determine if H1 is supported – it will be supported if detection accuracy levels are at 50% or above for both cultures being judged, that of the judge and that of the other culture. H2 asked about the relative accuracy of judges for members of their culture and for members of other cultures. It predicted that there would be no relative differences in detection accuracy, given the comparison of similar American and Australian cultures. Due to the repeated measures of the design and the binomial nature of the dependent variable, H2 will be tested with repeated measures logistic regression. Finally, H3 predicted that there would be a media difference in detection accuracy, such that media that can transmit the fewest cues to deception would be associated with less accurate detection than would media that can transmit more cues. H3 will also be tested with repeated measures logistic regression, for the same reasons as H2. With data similar to ours, both George et al (2018) and McHaney et al (2018) used repeated measures logistic regression for testing their hypotheses.

We analyzed the relationships between culture and deception detection accuracy, and between media and deception detection accuracy, using repeated measures logistic regression, in SPSS Version 23. The GENLIN command was used, with a binomial distribution and logit as the link function. Repeated measures were used, as each participant answered 32 different questions. Given that each of the 54 judges was asked to respond to 32 snippets, 1728 responses were generated.

Logistic regression was used since the dependent variable, whether or not the veracity judgment was correct, was binomial (correct or incorrect). The 7-point scale on which veracity was originally measured was collapsed into a discrete variable, where scores of 1 to 3 were considered a judgment of truth, and scores of 5 to 7 were considered a judgment of dishonesty. The 228 responses of '4' (i.e., undecided), at the center of the 7-point scale, were omitted from the analyses. The total number of responses omitted represented 13.20% of the total number of responses received. (Australian judges answered '4' 74 out of 608 times, or 12.20%; US judges answered '4' 154 out of 1120 times, or 13.75%.) The predictive

factors were communication media and stimulus set.

Overall, participants correctly distinguished between honest and dishonest snippets 53.90% of the time (51% correct for US English and 56% correct for Australian English). Australians successfully detected deception in both Australian (57%) and American (50%) snippets; Americans successfully detected deception in both American (51%) and Australian (56%) snippets. The results of the logistic regression analysis showed that the culture of the treatment was statistically significant ((X^2 (1, N = 1500) = 4.591, $p \leq 0.032$). It was more difficult to accurately detect deception in the US English stimulus set than in the Australian English set. The Australian and the American judges all had trouble with the American snippets. Their success with the American snippets was no different from chance (two-tailed tests: USA judges, USA snippets: t(436) = 0.287, p < .774; AUS judges, USA snippets: t(197) = .071, p < .943). Success with the Australian snippets was better than chance for both sets of judges (USA judges, AUS snippets: t(530) = 2.887, p < .004; AUS judges, AUS snippets: t(337) = 2.470, p < .014).

The logistic regression analysis also showed a statistically significant difference for media (X^2 (3, N = 1500) = 14.402, $p \leq 0.002$). A Bonferroni matched pair test ($\alpha < .05$) showed only one statistically significant difference: detection was more accurate in audiovisual snippets (60% correct) than in text (47% correct). The audio only snippets had an accuracy rate of 53%, while video only snippets had an accuracy rate of 54%. There was no interaction between culture and media.

When analyzing the data separately for the two groups of judges, we found that there were no statistically significant differences for culture. There were differences for media, though. While Australians were 50% accurate with the US English stimulus set and 57% accurate with the Australian English set, the differences were not statistically significant. For media, they were better with full audiovisual (63% accuracy) compared to video only (50%) ((X^2 (3, N = 534) = 8.778, $p \leq 0.032$). Analysis of the data from the perspective of US judges was similar. They were 51% accurate with the US English stimulus set and 56% accurate with the Australian set, but the differences were not statistically significant. For media, US judges were better with full audiovisual (58%) than they were with text (47%) ((X^2 (3, N = 966) = 10.119, $p \leq 0.018$).

We found, then, that both Australian and US judges were able to accurately detect deception in both the US English and Australian English stimulus sets, with success rates at or above 50% for each cultural group. However, there were no statistically significant differences in their detection accuracy across cultures. We also found that each group of judges was better at detection in one medium over another: audiovisual beat video-only for the Australians, while audiovisual beat text for the Americans. Overall, the judges were more successful with audiovisual snippets than they were with text. The answer to both of our research questions -- 1) Can individuals of one culture accurately detect deception in individuals from another culture? and 2) Is there a relationship between deception detection and media? -- is yes.

5. Discussion

The reach of global communication has spread rapidly in the past decade, due in part to the availability of smartphones. Smartphones allow relatively inexpensive access to the internet and access to people all over the world through text messaging and voice/video calls. And with that increase in communication comes an increase in exposure to deception, within our own culture and across the world's many cultures. Can we discern deception on the part of people from cultures other than our own? And the communication media that we are using make a difference in how easy it is for us to discern deception?

Based on our findings in this study, the answer to both of these research questions is affirmative (Table 9). Australian judges were able to discern deception in the communication of both Australians (at 57%) and Americans (although the accuracy rate for the US English stimulus, at 50%, is low and just meets the threshold for determining if deception detection is successful (Bond & Atoum, 2000; Bond, et al., 1990)). US judges were also able to detect deception in the communication of both Americans (51%) and Australians (56%). Hypothesis 1 is supported. Neither group of judges was better with one group or the other, supporting Hypothesis 2. Our findings provide additional support to the 'universal cue hypothesis,' which says that people can successfully detect deception, regardless of the culture of the sender or the receiver. Based on the results of this study, the Australians would fill an additional cell in Table 3, in the 'no difference' column, as would the Americans. Table 3, with these additions, is reproduced here as Table 10. The reason for the finding of 'no difference' no doubt lies in part in the extreme similarity between Australian and US cultures. This finding may seem obvious to some, but there have been no prior tests of deception detection abilities across national cultures, where the cultures were so similar. All prior studies involved national cultures that varied widely from each other.

Table 9: Summary of hypotheses tests

H1: Members of a national culture will be able to accurately detect deception among members of their own culture and in members of other cultures.	*Supported:* Australians successfully detected deception in Australian (57%) & American (50%) snippets; Americans successfully detected deception in American (51%) and Australian (56%) snippets.
H2: Australian and US judges will be able to detect deception equally well both within and across their cultures.	*Supported:* The differences in detection across cultures were not significant for either Australians or Americans.
H3: Media that can transmit more cues to deception will be associated with more accurate deception detection, compared to media that transmit fewer cues.	*Supported:* Detection success was higher with full audiovisual snippets (60%) than with text (47%).

In Table 10, Australians appear in two columns, 'better with other groups' and 'no difference.' (Americans appear in all three columns.) Clearly, the ability of members of a cultural group to detect deception across other cultural groups is not dependent on the culture of the judge. Judges from a particular culture perform at different levels, sometimes being better detectors of deception in their own group, sometimes better with other groups, and sometimes being equally good across groups. To some extent, the pattern of detection success seems to depend on the differences between the cultures of the judge and those being judged. For cultures that are very similar, like Australia and the US, group members are equally good at detecting deception both within their group and in the other group. Where the differences are more extreme (Table 5 and Figure 1), differences between cultures seem to be associated with one group doing better at detection with the other group than with their own. What might account for this outcome? One possibility: A bias against foreigners, in particular a bias against foreigners speaking in a second language (Bond & Atoum, 2000; Evans & Michael, 2013; Castillo et al., 2014). As Castillo and colleagues report in their 2014 study of Australians judging Australians and Colombians, "the difference in response bias across cultures was in the direction that suggests a tendency to greater suspicion of people from another culture – i.e., Colombian clips, in particular, those speaking in a second language" (p. 79). Such a bias, conscious or not, might motivate veracity judges to be suspicious of foreigners, leading to better detection of deception among members of those groups, as compared to their own group. Differences across cultures would influence a judge's detection success, where extreme differences would result in better detection with the other group than with his or her own.

Table 10. Patterns of comparative deception detection success across cultures in four studies

Better in own group (compared to other groups)	Better with other group (compared to their own group)	No differences
Americans (judging Americans & Jordanians) (Bond & Atoum, 2000)	Jordanians (judging Jordanians & Malaysians) (Al-Simadi, 2000)	Indians, Americans & Jordanians (judging Indians) (Bond & Atoum, 2000)
Jordanians (judging Americans & Jordanians) (Bond & Atoum, 2000)	Malaysians (judging Jordanians & Malaysians) (Al-Simadi, 2000)	*Australians (judging Americans) (current study)*
Spaniards (judging Spaniards & Americans) (George et al, 2018)	Australians (judging Australians & Colombians) (Castillo et al, 2014)	*Americans (judging Australians) (current study)*
Indians (judging Indians & Americans) (George et al 2018)	Americans (judging Americans, Spaniards & Indians) (George et al, 2018)	

We also found a main effect for media. As mentioned previously, there is some evidence of a direct relationship between media and deception detection, but many studies have found that the relationship is mediated. We found evidence of a direct effect, however. Across both groups of judges, those watching full audiovisual snippets were better at detecting deception than those who viewed text. This finding supports Hypothesis 3, that deception detection would be more successful with media that transmit more cues to deception than with media that transmit fewer cues. According to Rao and Lim (2000), full audiovisual media transmit 14 cues to deception, and text transmits half as many (7). What is perhaps more interesting is that media effects differed with each group of judges. While US judges were better at detection with full audiovisual compared to text, Australian judges were better at detection with full audiovisual than

with video-only communication. Based on Rao and Lim's work (2000), video-only modes of communication convey only five cues to deception. These findings are consonant with earlier findings regarding culture, deception and media. In the first study conducted by Bond and colleagues (1990), the researchers concluded that people could not accurately detect deception across other cultures, and their videos had no sound. In Bond's second study (with Atoum, 2000), the researchers concluded that people could detect deception across other cultures, but this time, the videos had sound. While it is interesting that media effects differed across cultures, the reasons why are not clear.

6. Limitations, Implications and Future Research

Although the unit of analysis was the veracity judgment, a limitation of this study was the relatively small number of participants who self-identified as being from either Australia or the US. Of the 76 participants in the study, we were not able to use data from 22 of them (which translates into 704 veracity judgments). We would also have preferred a more balanced sample of US and Australian judges.

Our findings have two implications for research. First, we have provided additional evidence that the universal cue hypothesis holds, even when the cultures being compared are very similar. Previous research found support for the hypothesis when the cultures being compared varied quite a bit from each other. Apparently, the cues that senders give off when deceiving are recognizable to members of both similar and dissimilar national cultures (as well as to members of their own cultures). Second, we now have an additional data point for cultural comparison with the inclusion of the Australian sample and the creation of our Australian English stimulus set. Unlike the other cultures and stimulus sets, the Australians were selected because of their similarity to other cultures, not because of their differences. We now have a broader spectrum from which to study cultural dimensions and deception detection.

Our findings also have implications for practice. The universal cue hypothesis holds across several cultural comparisons. For interrogators or judges, or anyone else whose job involves determining if someone is telling the truth, in most cases they can rely on the indicators of deception they have learned, in life or through training. What they have learned to look for in detecting deception works as well for people from any other culture as for people from their own. Specifically, for American interrogators, they should approach deception detection in Australian interviewees the way they approach the task with American interviewees, and vice versa. Our findings about media support the premise that media with the most cues are associated with better deception detection. When it is possible to select the best medium for interviews, where detecting the truth is important, interrogators should choose the medium that conveys the most cues to deception. Full audiovisual, such as videoconferencing, is more effective than text-based media, such as email. For Australian judges, as opposed to the entire sample of participants, full audiovisual was better than video without sound, again supporting the more-cues-is-better for detection accuracy thesis.

7. Conclusions

We now have enough cross-cultural deception detection comparisons to move beyond the question of whether people can successfully detect deception across cultures. They can. Now we can try to better understand why there are differences in detection outcomes: Why one group can detect deception better in another group compared to their own, while yet another group is better at detecting deception in their own group compared to others. And as this study has shown, in some cases groups are equally good at detection in both their group and another. A bias against foreigners is one possible explanation for some of these findings, especially when the foreigner, who typically speaks a different language, is now speaking "your" language. A selective comparison of more cultures and languages can be conducted to help produce a more complete understanding. From the few studies that have investigated media, culture and deception, we do have consistent evidence of media effects. Video-only communication seems to be the worst for deception detection. However, it is interesting to note that media effects seem to differ across judges from different cultural groups. Additional research is called for to investigate this intriguing finding.

References

Al-Simadi, F.A. (2001). Detection of deceptive behavior: a cross-cultural test. *Social Behavior & Personality*, 28(2), 455-462.

Banai, M., Stefanidis, A., Shetach, A., & Özbek, M.F. (2014). Attitudes toward ethically questionable

negotiation tactics: a two-country study. *Journal of Business Ethics*, 123(4), 1-17.

Bessarabova, E. (2014). The effects of culture and situational features on in-group favoritism manifested as deception. *International Journal of Intercultural Relations,* 39, 9-21.

Bond, C.F. & Atoum, A.O. (2000). International deception. *Personality and Social Psychology Bulletin*, 26(3), 385-395.

Bond, C.F., Omar, A., Mahmoud, A., & Bonser, R.N. (1990). Lie detection across cultures. *Journal of Nonverbal Behavior*, 14(3), 189-204.

Bond, C.F., & DePaulo, B.M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3), 214-234.

Buller, D.B., & Burgoon, J.K. (1996). Interpersonal deception theory. *Communication Theory*, 6(3), 203-242.

Burgoon, J.K., Stoner, G.M., Bonito, J.A., & Dunbar, N.E. (2003). Trust and deception in mediated communication. *Proceedings of the 36th Hawaii International Conference on System Sciences*, 1-11.

Burgoon, J.K., Blair, J.P., & Strom, R.E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. *Human Communication Research,* 34(4), 572-599.

Castillo, P.A., Tyson, G., & Mallard, D. (2014). An investigation of accuracy and bias in cross-cultural lie detection. *Applied Psychology in Criminal Justice*, 10(1), 66-82.

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, tasks, and communication processes: a theory of media synchronicity. *MIS Quarterly*, 32(3), 575-600.

DePaulo, B.M., Kashy, D.A., Kirkendol, S.E., Wyer, M.M., & Epstein, J.A. (1996). Lying in everyday life. *Journal of Personality & Social Psychology*, 70(5), 979-995.

Dunbar, N.E., Jensen, M.L., Tower, D.C. & Burgoon, J.K. (2014). Synchronization of nonverbal behaviors in detecting mediated and non-mediated deception. *Journal of Nonverbal Behavior*, 38(3), 355-376.

Dunbar, N.E., Jensen, M.L., Burgoon, J.K., Kelley, K.M., Harrison, K.J., Adame, B.J. & Bernard, D.R. (2015). Effects of veracity, modality, and sanctioning on credibility assessment during mediated and unmediated interviews. *Communication Research*, 42(5), 649-674. https://doi.org/10.1177/0093650213480175

Ekman, P., & Friesen, W.V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, 32(1), 88-105.

Ekman, P. (1985). *Telling Lies: Clues to Deceit in the Marketplace, Marriage, and Politics*. New York: W.W. Norton.

Evans, J.R. & Michael, S.W. (2013). Detecting deception in non-native English speakers. *Applied Cognitive Psychology*, 28(2), 226-237.

Fu, G., Heyman, G.D., & Lee, K. (2011). Reasoning about modesty among adolescents and adults in China and the U.S. *Journal of Adolescence*, 34(4), 599-608.

Furner, C.P and George, J.F. (2012). Cultural determinants of media choice for deception. *Computers in Human Behavior* 28, 1427-1438.

George, J. F., Gupta, M., Giordano, G., Mills, A.M., Tennant, V.M., and Lewis, C.C . (2018). The effects of communication media and culture on deception detection accuracy. *MIS Quarterly*, 42(2), 551-575.

George, J.F., Marett, K., and Tilley, P.A. (2008). The effects of warnings, computer-based media, and probing activity on successful lie detection. IEEE *Transactions on Professional Communication,* 51(1), 1-17.

George, J.F., Tilley, P., and Giordano, G. (2014). Sender credibility and deception detection. *Computers in Human Behavior*, 35, 1-11.

Hamilton, N.F. & Kirwan, G. (2013). A cross-cultural comparison of deception in online dating profiles using language analysis. In *Cyberpsychology and New Media: A Thematic Reader*, A. Power and G. Kirwan (eds.), Psychology Press, 49-59.

Hancock, J.T., Woodworth, M.T. & Goorha, S. (2010). See no evil: The effect of communication medium and motivation on deception detection. *Group Decision and Negotiation*, 19(4), 327-343.

Hofstede. G. (2016). *National Culture*. Accessed 7/20/16 at https://www.geert-hofstede.com/national-culture.html

Hofstede, G. (1980). *Culture's Consequences: International Differences in Work-Related Values*, Beverly Hills, CA: Sage Publications.

Hofstede, G. (2008). *Values Survey Module VSM08*. Accessed 9/29/2009 at www.geerthofstede.nl.

Hofstede, G., Hofstede, G.J., Minkov, M. & Vinken, H. (2008). *Values Survey Module 2008*. Accessed 9/29/2009, at www.geerthofstede.nl.

ITUNews, (2018). New ITU statistics show more than half the world is using the Internet. Accessed 10/2/20 at https://news.itu.int/itu-statistics-leaving-no-one-offline/

Lewis, C. *To Catch a Liar: A Cross-Cultural Comparison of Computer-Mediated Deceptive Communication*. (2009). Unpublished doctoral dissertation, College of Business, Florida State University.

McHaney, R., George, J.F. & Gupta, M. (2018). Deception detection: An exploration of annotated text-based cues. *Journal of the Midwest Association for Information Systems*, 2018 issue 2, article 2, 5-17.

Rao, S.R. & Lim, J. (2000). The impact of involuntary cues on media effects. *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 1-10.

Rockmann, K.W. & Northcraft, G.B. (2008). To be or not to be trusted: The influence of media richness on defection and deception. *Organizational Behavior and Human Decision Processes*, 107(2), 106-122.

Seiter, J.S. & Bruschke, J. (2007). Deception and emotion: The effects of motivation, relationship type, and sex on expected feelings of guilt and shame following acts of deception in United States and Chinese samples. *Communication Studies*, 58(1), 1-16.

Smith, A. (2015). U.S. smartphone use in 2015. Pew Research Center. Accessed 7/18/16 at http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

Srite, M. & Karahanna, E. (2006). The role of espoused national cultural values in technology acceptance. *MIS Quarterly*, 30(3), 679-704.

Straub, D., Loch, K., Evaristo, J.R., Karahanna, E. & Srite, M. (2002). Toward a theory-based measurement of culture. *Journal of Global Information Management* (l), 13-23.

Taylor, K. & Silver, L. (2019). Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. Accessed 10/2/20 at https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/.

Tilley, P.A. *Training, Warning and Media Richness Effects on Computer-Mediated Deception and Its Detection*. (2005). Unpublished doctoral dissertation, College of Business, Florida State University.

Triandis, H.C., Carnevale, P., Gelfand, M., Robert, C., Wasti, S.A., Probst, T., Kashima, E.S., Dragonas, T., Chan, D., Chen, X.P., Kim, U., De Dreu, C., Van De Vliert, E., Iwao, S. & Schmitz, P. (Global Deception Research Team). (2001). Culture and deception in business negotiations: a multilevel analysis. *International Journal of Cross Cultural Management*, 1(1), 73-90.

Van Swaol, L.M., Vraun, M.T., & Kolb, M.R. (2015). Deception, detection, demeanor, and truth bias in face-to-face and computer-mediated communication. Communication Research, 42(8), 1116-1142. https://doi.org/10.1177/0093650213485785

Zhou, L., Burgoon, J.K., Zhang, D. and Nunamaker, J.F. Jr. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior*, 20(3), 381-402.

Zhou, L. and Zhang, D. (2007). Typing or messaging? Modality effect on deception detection in computer-mediated communication. *Decision Support Systems*, 44(1), 188-201.

Author Biographies

Joey F. George is the John D. DeVries Endowed Chair in Business and a Distinguished Professor in Business in the Ivy College of Business at Iowa State University. His bachelor's degree in English is from Stanford University (1979), and he earned his doctorate in management from the University of California Irvine in 1986. Dr. George's research interests focus on the use of information systems in the workplace, including deceptive computer-mediated communication, computer-based monitoring, and group support systems. He is also the Associate Dean for Research. He was recognized with the AIS LEO Award for Lifetime Achievement in 2014.

Alastair Robb is a Senior Lecturer and Discipline Leader in the Business Information Systems Discipline at University of Queensland Business School. His qualifications are in Business and a PhD in Commerce (Information Systems) from the University of Queensland. Previously Dr. Robb has held positions in the automotive industry and was a Director of a Credit Union. His research interests lie in the fields of Distributed Ledger Technology, XBRL, deceptive behaviour in communications, and data quality. In his downtime Alastair studies Art History with a particular interest in the Italian Renaissance Masters.

This page intentionally left blank.

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 3

Date: 01-31-2021

Building Collaboration Networks and Alliances to Solve the IT Talent Shortage: A Revelatory Case Study

John Michael Muraski
University of Wisconsin – Oshkosh, muraskij@uwosh.edu

Jakob Holden Iversen
University of Wisconsin – Oshkosh, iversen@uwosh.edu

Kimberly Jean Iversen
NEW Digital Alliance, kim@newdigitalalliance.org

Abstract

As companies increasingly face challenges finding sufficient numbers of skilled IT workers, regions around the country have attempted different strategies to address the gap. In Northeast Wisconsin, the primary strategy has been the formation of a formal organization, the NEW Digital Alliance, charged with attracting, developing, and retaining IT workers in Northeast Wisconsin, funded by local companies and universities. In this paper, we will explore collaborative networks and the innovative effect they have on solving the IT talent pipeline challenge in a specific geographic region. Specifically, we explore the role of collaboration maturity and present a new comprehensive framework that may help understand and direct new regional collaborative efforts. The findings suggest that an alliance of business, education, and economic development partners can move a region forward in ways that are difficult for single players to achieve. We find that the Northeast Wisconsin region has been able to achieve coordination between K-12, higher education, and employers to improve on awareness of the problems that each part of the talent pipeline is facing. With an increase in collaboration maturity, the organization was able to relatively easily transition to virtual activities as well as assemble new constellations of collaborative efforts in short order when faced with the COVID-19 crisis.

Keywords: Collaborative Networks, IS Recruitment, Case Study

DOI: 10.17705/3jmwa.000064

1. Introduction

In the early part of the 2010's, businesses in the Northeast Wisconsin region faced significant challenges recruiting and retaining IT talent. As a result, they joined forces with area universities, K-12 schools, non-profits, and economic development agencies to form what was originally known as the NEW IT (Information Technology) Alliance is now referred to as the NEW Digital Alliance[1]. This paper explores the events that led to the formation of this Alliance, the increasing maturity of the Alliance, and the impact of the Alliance on IT talent in the region from 2015 through the COVID-19 outbreak.

Then as now, organizations struggled to fill the growing demand for information technology (IT) jobs. In fact, demand for IT talent is expected to grow 12.1% from 2019 to 2029 with an estimated 48,941 new IT-related jobs opening each year (U.S. Bureau for Labor Statistics (BLS), 2020). The IT field is projected to see the third-fastest growth after health care and community and social support occupations. At the same time, interest in IT-related academic majors has not kept pace with the rapidly increasing demand (National Center for Educational Statistics, 2018). This imbalance is causing significant talent shock in the marketplace and causing companies and colleges to explore heretofore unexplored options and possibilities.

Individual organizations have adopted comprehensive talent management, recruitment, and acquisition programs. These programs have included recruitment software and applicant tracking systems, social networking sites, video interviews, as well as formalized onboarding and training programs (Jose, 2019). Traditionally, these efforts have been carried out by individual companies. Even with these efforts, IT talent issues have consistently ranked as the second or third most worrisome concern of CIO's over the past 7 years (Kappelman et al., 2020). The inability to solve these issues independently led individual leaders from these companies to seek a more collaborate approach to the talent shortage.

Similarly, individual colleges and universities have attempted to address the challenges of increasing enrollment through the development of new majors, minors, and certificate programs. Some of these programs include artificial intelligence, big data analytics, business intelligence and analytics, cybersecurity, ERP, Internet of Things (IoT) (Case et al., 2019). As with individual organizations, these colleges and universities often worked in isolation.

In response to this talent shortage, the Northeast Wisconsin Digital Alliance (NEW Digital Alliance) was founded in 2015 to address the talent shortage in the Northeast Wisconsin area. This article explores the establishment and maturation of the NEW Digital Alliance. Guided by the work of Schilling (2015), who focused on technological collaboration and innovative outcomes, as well as Morgan (2012), who developed a collaboration maturity model, we explored three key areas. First, inflection points that highlight the transition from one phase of maturity to the next were identified. Second, the relationship between the NEW Digital Alliance maturity and the overall collaboration network maturity was explored. Finally, innovative outcomes from the collaboration were identified. As we face the ongoing challenges of the COVID-19 crises, the findings of our research can be applied to new networks being developed to face new challenges.

This paper is organized as follows. The next section covers the theoretical background around collaboration networks and alliance formation as well as collaboration maturity. Section 3 describes the single-case study research method that was used to guide the research as well as an overview of the case environment. Section 4 provides a detailed case description and analysis. Section 5 provides a discussion on the findings. Finally, the conclusion includes a summary, limitations, and direction for further research.

2. Theoretical Background

In an effort to explore and understand our case study as well as develop research questions, we explored established literature on collaboration networks and alliance formation as well as collaboration maturity. While studying these individual theories, we identified that prior research had not integrated them into a comprehensive framework. Data from the case study helped showcase concepts from these theories to explain how a regional collaboration can form and mature.

[1] For clarity, in the rest of this paper we will refer to the organization by its current name, NEW Digital Alliance, or simply 'the Alliance.'

2.1 Collaboration Networks and Alliance Formation

Uncertainty motivates firms to enter into alliances (Frankort et al, 2016). Schilling (2015) identified the relationship between a technology shock, alliance formation, collaboration network, and the resulting innovation outcomes (Figure 1). In this model, a triggering event, such as the introduction of a new technology, facilitates the formation of an alliance of firms as well as the larger collaboration network. Consequently, the alliance formation directly positively contributes to the development of the collaboration network. The alliance formation contributes to, supports, and generally guides the collaboration network. Both the alliance formation and collaboration network result in innovation outcomes. For Schilling (2015), the network included firms, government labs, universities and other organizations that together represented significant components of the global technology network. Schilling (2015) measured the innovation outcomes by the number of patents issued.

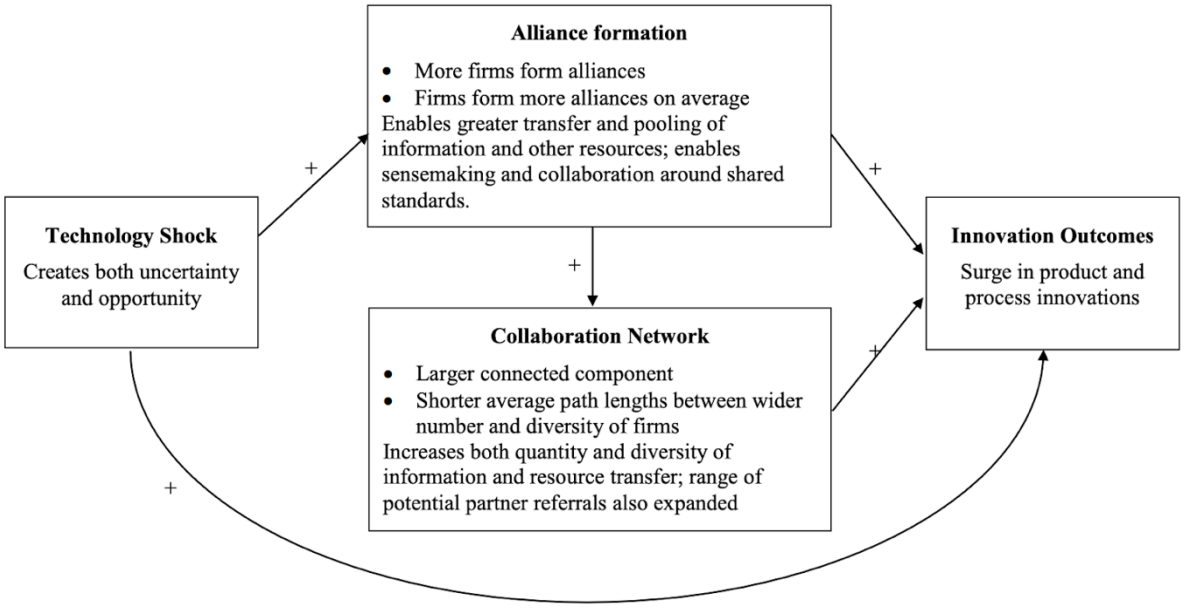


Figure 1: Technology shocks, technological collaboration, and innovation outcomes. Adopted from Schilling (2015).

This model has also been used to explore interorganizational knowledge transfer (Milagres & Burcharth, 2019). Specifically, technology shock was identified as a significant environmental uncertainty that facilitates knowledge transfer. Similarly, the technology shocks and resulting innovation was used to explore technology alliances and market cycle (Martynov, 2019). Finally, this model was used as a foundation to show that the growth of an inventor network is associated with innovative impacts (Argyres, et al., 2020).

2.2 Collaboration Maturity Model

All organizations mature over time and have unique characteristics at each stage of maturity. The notion of a model to describe this growing maturity was first developed and popularized at the Software Engineering Institute at Carnegie Mellon University in the 1980s and 1990s. This work resulted in the Capability Maturity Model aimed at describing the maturity of software development processes (Paulk et al., 1993). Morgan (2012) used the maturity concept to identify an emergent collaboration maturity model for describing stages of a collaboration ecosystem progression. In the Morgan (2012) model, collaborative networks progress along five maturity stages that include: Unaware, Exploratory, Defined, Adoptive, and Adaptive. The primary goal is to reduce the strategic value gap and deliver greater business value. The strategic value gap represents the gap between current maturity and business value achieved from a fully adaptive organization. As the collaborative network matures, the strategic value gap declines and the overall business value increases. Figure 2 shows the five levels of maturity and provides a definition of each level.

The five phases of collaboration maturity can be assessed along five characteristics: goals and objectives, organizational culture, process, technology, and governance.

2.2.1 Goals & Objectives

Goals and objectives represent the outcomes that an organization hopes to achieve. Morgan (2012) identified the following organizational levels: company, department, employee, and customer. In the Unaware phase, goals and objectives have not been stated at any level. During the Exploratory phase, they have been explored. In the Defined phase, goals and objectives have been defined. During the Adoptive phase, they have been formally communicated. Finally, in the Adaptive phase, goals and objectives have been adapted to the particular situation of the collaborative network.

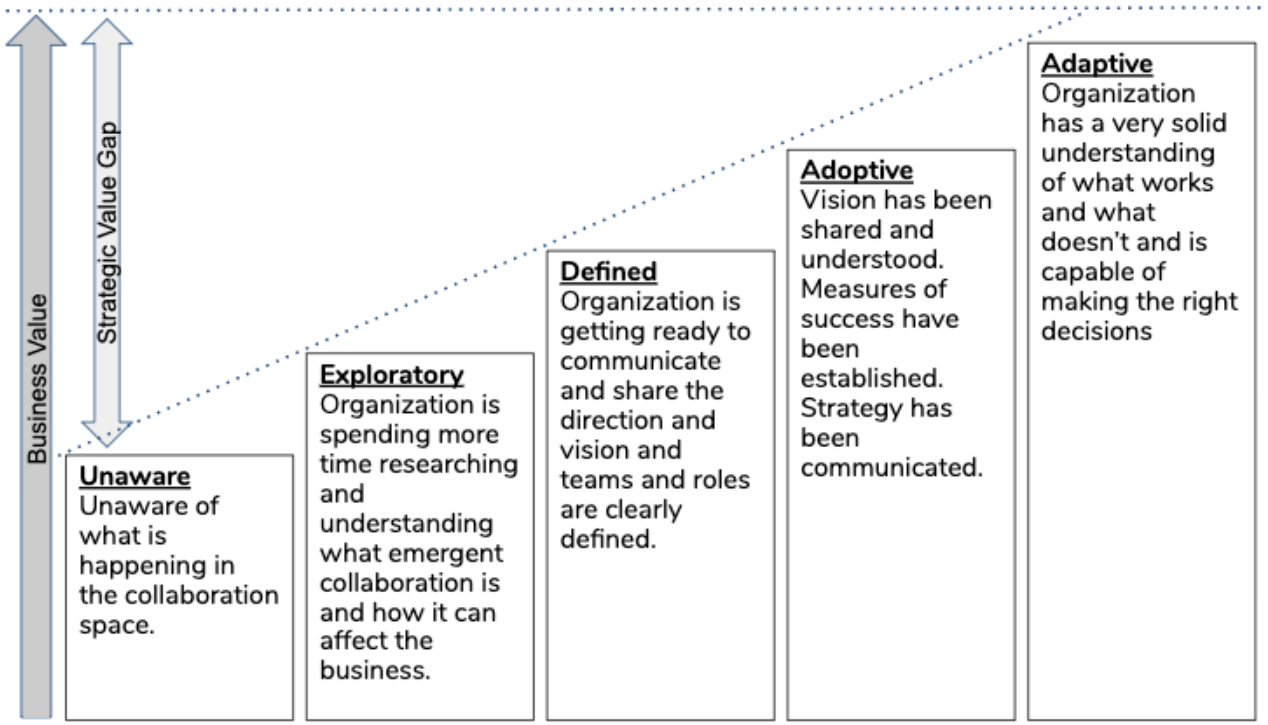


Figure 2: Emergent Collaboration Maturity Model. Adopted from Morgan, 2012.

2.2.2 Organizational Culture

Organizational culture can be defined as ensuring that collaboration is valued and incentivized. Morgan (2012) identified the following components of organizational culture: leadership, organizational change management, ensuring evangelists, fostering openness, and establishing mutually beneficial value. In the Unaware phase, the organizational culture components are not addressed and have no sponsorship. During the Exploratory phase, these components begin to be identified and discussed. In the Defined phase, many of the organizational culture components are defined and secured. During the Adoptive phase, the components are initially communicated and executed. Finally, in the Adaptive phase the components are formally communicated and adapted.

2.2.3 Process

Process refers to the required organizational changes that will occur to support the maturing collaboration. Morgan (2012) identified the following components of process: escalation, automation, and information management. In the Unaware phase, the process components are not in place nor considered. During the Exploratory phase, processes and plans are identified. In the Defined phase, process components are developed and communicated. During the Adoptive phase, the process components are implemented. Finally, during the Adaptive phase, these components are generally adapted as circumstances change.

2.2.4 Technology

Technology can be defined as the technological advances that enable communication, information sharing, and collaboration. Morgan (2012) identified the following components of technology: tool selection, integration, training, adoption, and maintenance and upgrades. In the Unaware phase, technology is not considered or addressed. During the

Exploratory phase, problems and opportunities are identified and technology is considered. In the Defined phase, strategies for addressing technology are developed. During the Adoptive phase, the road map for technology is communicated and further developed. Finally, during the Adaptive phase, training occurs, and technology is continuously adapted.

2.2.5 Governance

Governance can be defined as ensuring employees understand the policies of the organization. Morgan (2012) identified the following components of governance: best practices, guidelines, policies, oversight team, and social service level agreements. In the Unaware phase, governance does not exist. During the Exploratory phase, governance is recognized as needed. In the Defined phase, approaches are identified, and teams selected. During the Adoptive phase, governance components are created and developed. Finally, in the Adaptive phase, governance components are evaluated and evolved on an ongoing basis.

This model was used to develop the book The Future of Work, Build Better Leaders and Create a Competitive Organization (Morgan, 2014). In addition, this collaboration framework was used to provide insight into the adoption of social collaboration software (Komarov et al., 2014) as well as to explain the social impact of knowledge work (Getto et al., 2014).

2.3 Research Questions

Morgan (2012) highlighted that business value is directly related to organizational maturity. Schilling (2015) focused on the value of collaboration. In considering the collaboration maturity model as well as collaboration networks and alliance formation, we identified the following research questions:

1. What are the inflection points that highlight transition from one phase to another phase?
2. How does maturity of the alliance formation contribute to the maturity of the collaboration network.
3. What innovation outcomes occurred? How were they measured?

3. Research Methodology & Case Environment

In an effort to explore and understand our case study as well as develop research questions, we explored established literature on collaboration networks and alliance formation as well as collaboration maturity. While studying these individual theories, we identified that prior research had not integrated them into a comprehensive framework. Data from the case study helped showcase concepts from these theories to explain how a regional collaboration can form and mature.

3.1 Research Methodology

This revelatory case study seeks to investigate the development and maturation of a collaborative network. As research of regional collaborative networks is limited, it is suitable to conduct the study from an exploratory perspective utilizing a single case (Eisenhardt, 1989; Yin 2017). Case study methodology is appropriate under specific conditions. Yin (2017) notes that a case study is an "empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between the phenomenon and context are not clearly evident" (p. 13).

3.2 Case Environment

This case is set in the Northeast Wisconsin economic development region (New North). This region is made up of 18 counties covering roughly a quarter of the state of Wisconsin. Figure 3 shows a map of the region. The primary economic development organization for the region is New North Inc, which was established with a mission "to be the catalyst for regional prosperity for all through collaborative action". The primary industries in the region include transportation equipment manufacturing, dairy product manufacturing, foundries (pulp, paper, paperboard manufacturing and converting), electrical equipment manufacturing, machinery manufacturing, and fabricated metal product manufacturing. Manufacturing in the region has become increasingly reliant on technology and 25% of the workforce is estimated to work in advanced manufacturing industries (New North 2020).

Figure 3: Map of the 18 counties that make up Northeast Wisconsin

Having been involved in the IT community in the region for several decades, the authors have observed the changing demands for IT professionals in the region and the desire to collaborate at the regional level. In the late 1990s, demand for talent was strong and students responded by enrolling in IT-related programs in very large numbers. For example, in 1999-2000 the Information Systems major at University of Wisconsin Oshkosh was the largest in the college. However, student demand dropped sharply and by 2004, the major had shrunk to 10% of its size 5 years prior and it stayed at that level for about 10 years.

One of the authors was involved in the very early days of the NEW Digital Alliance in 2014 and 2015 when it was just a loose affiliation of companies, universities, K12 schools, and interest groups, and was able to follow the early stages of the formation of the collaboration. Another of the authors was a founding member of Women in Technology and chair of its WIT4Girls program aimed at increasing opportunities for girls ages 5-18 to experience IT. She was later hired as the first director of the NEW Digital Alliance and has been instrumental in the increased level of collaboration in the region as well as the maturation of the collaboration efforts. The final author has extensive consulting, strategic planning, and IT project management experience in a number of companies and organizations in the region before joining UW Oshkosh as a faculty member. He has since been involved in several NEW Digital Alliance committees and initiatives. Because of their close connections, the authors have had many informal conversations about the IT community in the region that have impacted the direction of the Alliance, and several other programs in ways that cannot be teased out formally.

As we discussed the recent efforts in the region, we realized that our experiences provided us with a unique vantage point from which to provide an inside account of the formation and shaping of strong regional collaborative efforts aimed at strengthening the local IT community for both employers, students, and IT professionals.

4. Case Description & Analysis

In Northeast Wisconsin, over the past several years, the need for an increased number of technology-educated and technology-trained employees has been recognized (Matzek, 2018). Several organizations started to explore this issue but without a central hub for this developing network, activities were not well coordinated.

Through the formation of the formal NEW Digital Alliance with a dedicated full-time director, the region has been able to move up the levels of maturity. The Alliance now plays a central role in coordinating and organizing a number of activities as well as being a central hub for information about all the events in the region through a website, monthly newsletter, and robust social media presence. This section will describe the development of the organization and how it fits in the collaboration maturity model. As part of this narrative, we have identified several inflection points of particular

significance that signaled maturation of the Alliance. Figure 4 shows a timeline of all the inflection points. This section is organized by stage on the collaboration maturity model.

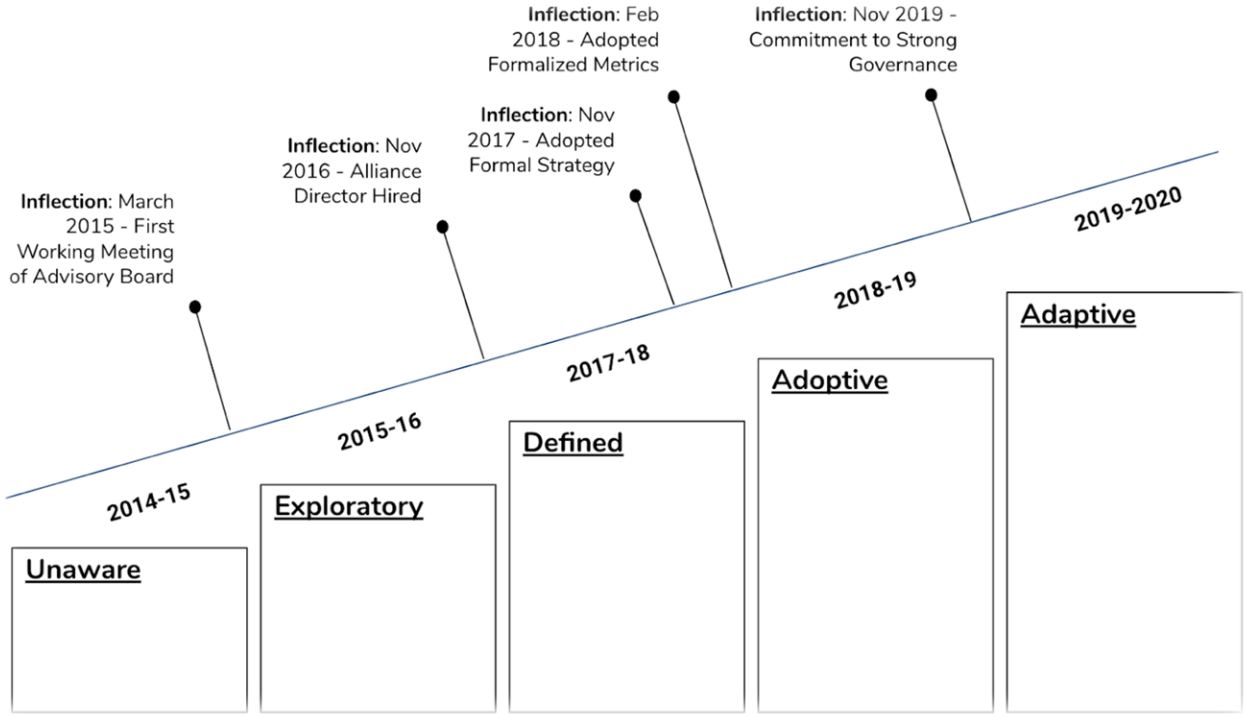


Figure 4: Reflection Points Between Maturity Levels

4.1 Unaware Stage (prior to March 2015)

In the early 2010s, IT leaders in Northeast Wisconsin businesses gradually realized that it had become difficult to find the necessary IT talent. In spite of this, educational institutions did not see an uptick in students to meet the demand. However, with no shared understanding of the problem among businesses, companies largely felt they were alone with the problem of finding IT talent, and educational institutions felt no pressure to systematically increase offerings or attempt to draw more students. With no collaboration, and moreover, no real sense that there was a problem, the region was firmly in the Unaware stage as shown in Table 1.

In early 2015, a small group of IT leaders in the region started discussions about how to solve the problem across the region so as to avoid cannibalizing IT employees from each other but instead increase the supply of IT talent in the entire region. We identify the first meeting of this group in March 2015 as the first inflection point towards moving from the Unaware stage towards exploration of a collaborative approach to solving the IT talent crisis.

Characteristics	**Assessment**	**Description**
Goals & Objectives	Vague	Vague realization that problems with finding IT talent are shared across many companies. No specific goals or objectives.
Culture	Unaddressed	Not addressed.
Process	None	Nothing in place.
Technology	Unaware	Nothing in place.
Governance	None	Ad hoc leadership team.

Table 1. Unaware Stage (Prior to March 2015)

4.2 Exploratory Stage (2015-16)

For about two years, efforts to solve the IT talent problem were driven by various individuals and organized groups. The NEW Digital Alliance grew out of these efforts and was envisioned as a forum to coordinate the disparate efforts across the region that were hampered by the lack of dedicated focus on collaboration and coordination. Each stakeholder group had specific motivations to be included. Businesses sought to improve the talent pipeline, thus reducing recruitment and hiring costs. Universities sought to raise enrollments in these high-value STEM majors. Several of the founders, nearing the end of their active careers, sought to positively impact the region.

Table 2 shows the primary characteristics of the efforts during this period that culminated with the hiring of a director for the Alliance. We characterize this as the Exploratory phase as the focus was on considering goals, identifying leaders, seeking sponsorships, and recognizing the need for additional governance.

4.2.1 First Meeting - March 2015

In 2015, Northeast Wisconsin business and educational institutions partnered to form what first became known as the NEW IT Alliance and later the NEW Digital Alliance. From the first meeting in March 2015, the group invited business partners, higher education, K-12 schools, non-profits, workforce development agencies, economic development agencies, chambers of commerce and others to participate in roundtable discussions on how to solve the problem. At these quarterly meetings it became clear that many organizations in the region were working to solve the same issue but with limited coordination. This alliance began to explore the issue of the IT Talent Pipeline and potential paths for approaching this issue. Meetings often had 40-50 people in attendance, and all were involved in working on various initiatives to help solve the problem. The meetings consisted mostly of updates on various initiatives going on around the region with an attempt at establishing collaboration where appropriate. Appendix A includes a list of the partners in the region that helped form the Alliance and have collaborated since then.

4.2.2 Survey of IT Talent Issues

One of the first formal activities of the organization included working with Northeast Wisconsin Educational Resource Alliance (NEW ERA), a collaborative of all the higher education institutions in the region, to conduct a survey of businesses in the region to determine the extent of the problem of hiring IT talent. The NEW ERA survey showed that demand existed for an additional 3,000 IT employees by 2021. One of the issues that became clear from the survey and the follow-up meetings around the region to present the results to groups of businesses was that many businesses were unaware that the problem of finding IT talent extended beyond their own organization. The survey results raised the awareness that a broader problem existed.

4.2.3 Volunteer Efforts Continue

The community leaders who were informally leading the regional discussions realized in 2016 that they were hamstrung by not having someone who could dedicate full-time attention to the issues raised. It became a problem that little progress was being made between the infrequent meetings. So, a consortium of businesses and higher education institutions decided to pledge enough money to hire a full-time director for the Alliance.

During 2016, the Alliance began to coordinate activities offering a central hub to the growing distributed network of organizations, businesses, and educational institutions focused on this challenge. The group discussed how to achieve open communication of goals and activities, sponsorship of the entire network, governance of the network, and goals and objectives.

4.2.4 Hiring of Director

This process concluded in November 2016 with the hiring of a full-time director for the organization, which we identified as the second inflection point, as it allowed the organization to have a person dedicated to coordinating the various efforts and increasing collaborative maturity in the region. The director has a background in IT having worked 13 years at a large local employer. As a founding member of Women in Technology Wisconsin and leader of their WIT4Girls initiative, she had deep connections to the broader IT community in the region.

Characteristics	Assessment	Description
Goals & Objectives	Initial Goals Defined	These goals included the hiring of the full-time director (Nov. 2016), legal registration of the organization, and funding to ensure initial and ongoing operations.
Culture	Developing	Collaborative networks forming and growing across the region.
Process	Developing	Ad hoc leadership group calling meetings approximately quarterly.
Technology	Unaddressed	Email and ad hoc file sharing among the members.
Governance	Developing	Leadership team becomes more defined.

Table 2. Exploratory Stage (2015 - 2016)

4.3 Defined Stage (2016-17)

We identify the period from the end of 2016 through 2017 as crucial for the organization. Having hired a director, the focus in this period was primarily on formalizing most aspects of the organization. This included major changes to processes, technology, strategy, and governance. This period moved the organization from the Exploratory stage, which is primarily about planning and into the Defined stage, which is characterized by having put plans into action. By the end of 2017, the organization was vastly different from what it looked like a year earlier. See Table 3 for summary of characteristics of Defined stage.

4.3.1 Initial Director Actions - November 2016 –2017

During the director's first year, most of the effort was focused on establishing the organization and its governance. However, the first NEW Connect IT job and Career Fair was also conducted in November 2017. Hiring a director who was able to dedicate full-time attention to the organization led to a dramatic increase in maturity with processes being defined and technology resources established. During 2017, the following activities were accomplished:

- Established website with Job Board
- Established initial sponsorship amounts for supporting organizations.
- Launched monthly newsletter
- Established social media channels and hired outside resource to manage social media presence ensuring regular activity
- Implemented Wild Apricot for event management and communication.
- Hired a college intern to support marketing and social media efforts.

Towards the end of 2017, the organization was well established with a strong presence in the community a website with strong content, social media presence across multiple channels, and a regular newsletter.

4.3.2 Adopted Strategy - Fall 2017

The director led the Executive Committee through a series of discussions to determine a formal strategy for the organization. This culminated in November 2017 with the formal adoption of a strategic plan for the organization that defined three pillars and three key audiences. Figure 5 lays out the key elements of the strategy. We identify the adoption of the strategy as an inflection point as it provided direction and focus to the efforts of the organization over the next several years.

The strategy was later refined to define three pillars related to increasing the talent pipeline: Attract, Develop, and Retain. The strategy was initially conceived to be very broad in terms of audiences. It included convincing high school students to study IT, keeping IT professionals in the region, attracting IT professionals from other regions, and inspire working or under-employed adults to pursue an IT career.

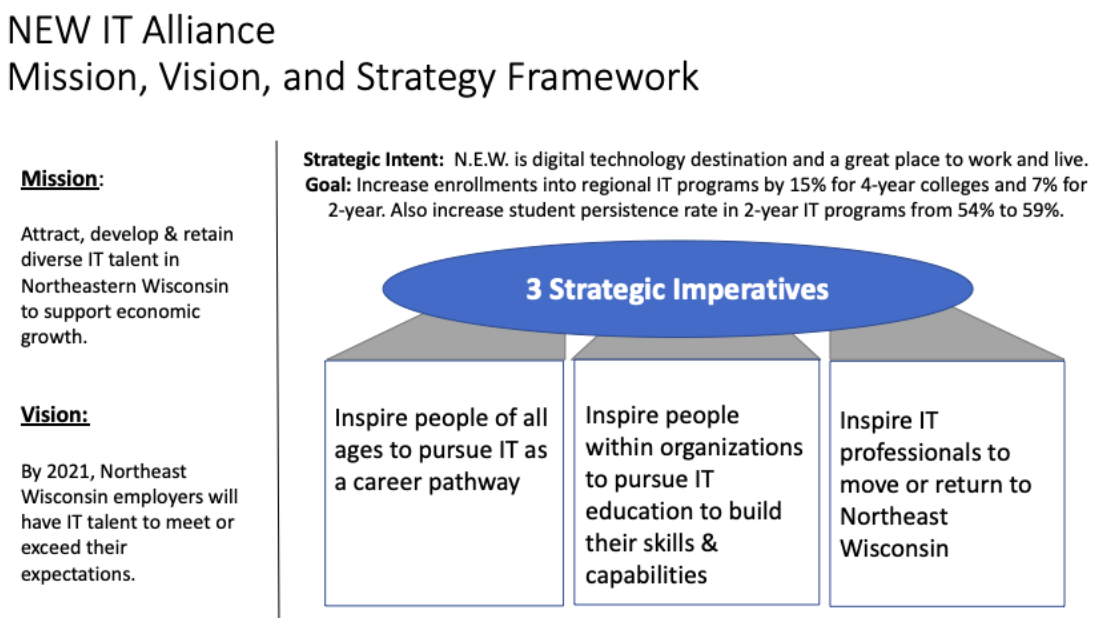


Figure 5: NEW IT Alliance strategic framework adopted in late 2017 and metrics adopted in early 2018.

4.3.3 Defined KPIs - Q1 2018

The next inflection point followed relatively quickly after the establishment of the strategy as the Executive Committee defined formal KPIs for one of the three pillars in the strategic framework: Develop. The goal was set to increase enrollments in regional IT programs by 15% for 4-year institutions and 7% for 2-year institutions. It also recognized the problem of students dropping out of 2-year programs at high rates by looking to increase the persistence rate in these programs from 54% to 59%.

Characteristics	**Assessment**	**Description**
Goals & Objectives	Defining throughout the period	Adopt strategy and set metrics.
Culture	Developing	Strong collaboration with partners throughout the region. Founders and Director setting the tone for the organization.
Process	Defined to Adoptive	• Social media presence through a contract with an external partner. Launched monthly newsletter. • Hired college intern. • Held NEW Connect IT (job and career fair). • Invoicing handled by parent organization (New North). • No CRM system for membership tracking and invoicing. Spreadsheets and manual processes used instead.
Technology	Selected - in some areas	• Adopted Google as collaboration platform set up as personal accounts and not with a dedicated business domain. • Website launched. Adopted software to manage events and newsletter mailing list as well as collaboration.
Governance	Defined	• Formalized executive committee with clear roles responsibilities. • Formed Talent and Marketing Committees • Established sponsorship amounts for supporting organizations

Table 3. Defined Stage (2016 - 2017)

4.4 Adoptive Stage (2018-19)

With the adoption of a strategic framework and defined metrics, the organization signaled a readiness to move into the Adoptive stage where the focus shifts to collecting metrics, executing on plans, and strengthening governance. Table 4 shows the characteristics of the organization during this period as collaboration moved from the Exploratory through Adoptive towards the Adaptive stage.

4.4.1 Data Collection

Having defined the KPIs, the organization set out to collect the data necessary to determine progress. This was done through an annual survey to each higher education institution for data to show whether enrollment was increasing or decreasing. Figure 6 shows the data for 2015-2020 for both 4-year universities as well as 2-year technical colleges in the region where data is available for each of the five years.

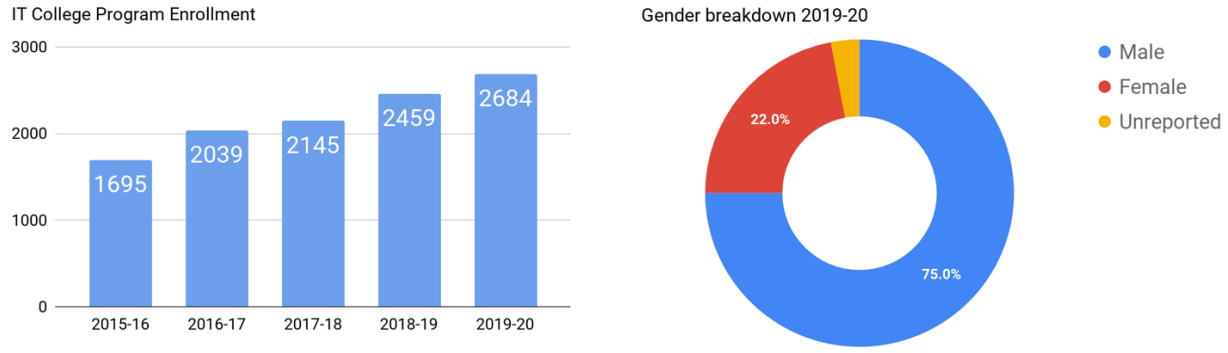


Figure 6: Enrollment in college IT programs

Simultaneously, the Alliance launched a collaboration with Microsoft to collect data from the K-12 system. That survey launched in 2018 as a pilot in two counties before expanding to the entire 18 county region in 2019. This survey collected data on both course offerings and enrollments as well as what barriers exist within the K-12 schools to offering more CS coursework. Figure 7 shows an example of data available in the dataset. This is the percent of high school students enrolled in a few standard CS classes in 2018-19 and 2019-20. As can be seen, most of the courses exhibited significant growth year-over-year. In the period, the average number of CS courses run by each district also increased from 3.4 to 4.2. Gender breakdown across the four traditional computer science courses mirror the trend in higher ed with about 76% of students identifying as male.

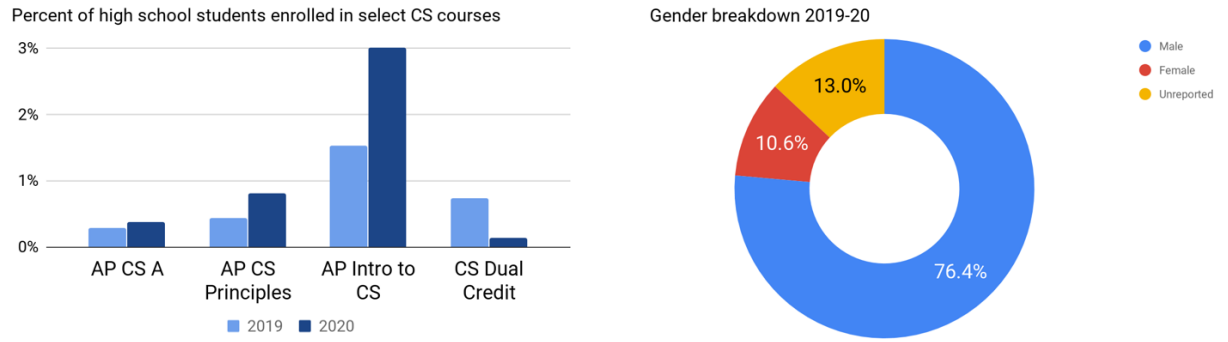


Figure 7: Enrollment in high school computer science classes

4.4.2 Promoting IT Careers

The Alliance was recognized to receive several grants to help boost focus on the K-12 space. The Wisconsin Department of Public Instruction awarded $14,000 to support developing IT Career Pathways from the high school level through college and into careers. Microsoft granted a $25,000 grant to help grow digital learning opportunities for underserved populations. The result was a map connecting high school and college educational opportunities in the region to five specific IT careers (DPI, 2019).

In fall 2018, the Alliance launched an effort to promote the IT profession in the region by creating a series of videos featuring local IT professionals talking about what they find exciting about their jobs and working in the Northeast Wisconsin region. This resulted in a series of 10 short videos uploaded to YouTube and made available to schools and others in the region to use for the promotion of IT.

4.4.3 Commitment to Stronger Governance - Fall 2019

Towards the end of 2019, it became clear that governance needed to be improved. At that time, the organization had the following committees: Executive, Talent, Higher Ed, and Marketing. The Marketing committee had become dormant due to lack of engagement from member volunteers whereas the other committees met regularly. Meetings were planned and led by the director. It became clear that there was a need for more engagement from committee members. This led to an inflection point when the Executive Committee members pledged to become more active in various activities of the organization, including participating in the other committees. The Higher Ed committee also elected a chair to help the director set the agenda and provide direction for the committee's work. In the Talent committee, no volunteer for the position stepped forward.

At the same time, several companies announced plans to drop their membership citing lack of time to engage and internal financial constraints.

In January 2020, the data collection efforts came to fruition with the publication of key data from K-12 and higher education on a dedicated page on the NEW Digital Alliance website (www.newdigitalalliance.org). This Factsheet page showed data from the first two years of K-12 data as well as four years of data from the higher education institutions. Both sets of data indicated growth among students engaged in IT education. It also highlighted challenges still ahead - including demand outpacing supply and a gender imbalance where the number of men far exceed the number of women. Figures 6 and 7 show some of the published data.

Characteristics	**Assessment**	**Description**
Goals & Objectives	Adoptive	• Expanded with a new strategy. • Launched survey to companies. • Continuing to collect K-12 and Higher Ed data. • Planning IT Summit in June 2020 to report on the state of IT across the region from K-12, Higher ed, and companies.
Culture	Exploratory/Defined	Lack of engagement may be holding back progress both within the Alliance as well as across the larger community in the region.
Process	Adoptive	• NEW Connect IT event planning smoother due to experience. • Social media presence expanded with additional channels and in-housing some management and content creation to interns. • Launched videos of IT Professionals. • Launched Insights on Technology with local publishing partner. • Added a high school intern to help with website support and set up regular working sessions with interns.
Technology	Adoptive/Adaptive	Switch from Google to Microsoft. Standardized on Microsoft as a collaboration platform to ensure consistent organizational support and access.

Governance	Adoptive	Governance strengthens significantly in this period: • Formed Higher Education committee. • Discussions with the Executive committee on strengthening engagement. Each member signed up for additional work. However, limited actual follow-through. • One executive joined the Higher Education committee, which also elected a chair so as to not have the director be solely in charge of the meetings.

Table 4. Adoptive Stage (2019)

4.5 Next Steps (2020 and beyond)

In early 2020, the organization was at a crossroads. The need for IT talent was still strong and a major concern for many companies in the region. Enrollment in IT courses in high schools across the region had increased significantly and barriers to further growth had been identified. At the higher education level, enrollments had started to slowly increase. However, company engagement with the organization had dropped with several member companies announcing they were not renewing their membership. Similarly, there was a general lack of engagement from member volunteers.

Even with these setbacks, the Alliance moved forward on several fronts. Table 5 highlights the movement toward the Adaptive stage.

4.5.1 New Strategy and Renamed Organization - February 2020

To help increase the number of member organizations, improve on engagement from member organizations and stabilize the financial outlook for the organization, the Alliance embarked on a strategic planning effort during the first quarter of 2020. This culminated with a decision to broaden the scope of the organization and a rebranding. By renaming the organization from the NEW IT Alliance to the NEW Digital Alliance, the idea was that the scope of the organization would be broadened to not just focus on traditional IT roles but to expand to also focus on the digital skills needed to be successful in any career.

4.5.2 Technology Transition - Google Docs to Microsoft SharePoint

Early on, the organization had adopted Google's collaboration tools - Google Docs, Drive, Hangout etc. - to support the organization. However, it had become increasingly clear that this choice wasn't sustainable as the accounts were set up as personal Google accounts making it difficult to provide professional branding and strongly separate organizational business from personal accounts. In addition, some companies blocked key Google technologies from their networks making it difficult to collaborate across companies. The Alliance decided to switch to a comprehensive platform and chose Microsoft Teams and SharePoint for collaboration instead. While most of the work to transition to the new platform fell to the director and the two interns, the value of the collaborative network established by the organization became evident here as well. When the process hit a snag, a local technology training firm and member organization made one of their SharePoint trainers available for several hours of consultation to help fix the problem.

4.5.3 And Then the World Changed

Like all organizations, the NEW Digital Alliance was affected by the transition to remote and virtual operations as a result of the COVID-19 pandemic that hit in early 2020. This meant that almost all in-person activities and events from March through at least the end of the year would be conducted virtually. Because of the increase in maturity along all dimensions, the organization was able to transition and continue operating with little disruption. The recently adopted collaboration technologies in the form of Microsoft Office 365/SharePoint as well as Zoom allowed for all meetings to be transitioned to virtual instead of in person.

The NEW Digital Alliance was also well positioned to take advantage of the collaborations and networks previously established to set up new related collaborative efforts, such as a survey launched in collaboration with internship coordinators at regional colleges (most of whom had not previously collaborated) to examine the impact of the pandemic on summer internships for college students. The Alliance also quickly pulled together a broad-based committee to look at opportunities for displaced workers to re-skill into IT roles. Both of these efforts show the organization having matured and able to play to the strengths it had built up over the previous years in establishing collaborative networks across the region.

Characteristics	Assessment	Description
Goals & Objectives	Adoptive/Adaptive	Expanded with the new strategy to also include digital readiness of the entire workforce as well as innovation across the region.
Culture	Defined	Significant effort required to increase engagement and evangelism on behalf of the organization and IT community in the region.
Process	Adoptive/Adaptive	Ability to pivot in face of change is a testament to mature processes.
Technology	Adaptive	Transition to Office 365, SharePoint, Teams and Zoom. Email addresses transitioned to new domain names instead of personal Gmail accounts.
Governance	Adoptive/Adaptive	• Working towards 501c.3 status. • Exploring tiered investment options for various types of organizations (higher education, not-for-profit, and for-profit). • The Higher Education Committee works on setting goals and objectives for the committee that is more independent of the Alliance with the purpose of making the committee more valuable to the members from the higher education institutions.

Table 5. Moving to Adaptive Stage (2020)

4.6 Innovative Outcomes

Discrete activities represent the primary measurement of innovation outcomes. These activities originate from both the Alliance directly as well as from the overall collaborate network. Activities have been tracked through identification and self-reporting since early in 2017. Table 6 provides an overview of activities from 2017 through 2020. Many events in 2020 were not even scheduled given the COVID-19 restrictions. See Appendix B 2019 Alliance and Network Events for the full list of 2019 events.

	2017[2]	2018	2019	2020	Total
Alliance	4	10	12	16 (3 cancelled)[3]	42
Network	16	52	97	40 (2 cancelled)	205
Total	20	62	109	56 (5 cancelled)	247

Table 6. Alliance and Network Activities

There are several key Alliance events. First, NEW Connect IT represents a yearly conference targeted toward high-school and college students. This full-day event allows students to learn about technology careers and the educational pathways to pursue those careers, meet with higher education institutions, and talk to many area employers. Second, quarterly TechTalks bring presentations of the newest technology to colleges and universities in the region. Third, the annual Tech Talent Summit provides the state of the Digital ecosystem in Northeast Wisconsin. This event includes data, successes, and various panel discussions for all Alliance members and the community at large. Finally, a quarterly CS Advisory Board meeting facilitates addressing the IT gaps in high schools through facilitation of advisory boards. All current events can be viewed at https://newdigitalalliance.org/events/.

5. Discussion

In this case study, we explore our experience working and interacting with the NEW Digital Alliance and the related collaborative effort that is evolving across Northeast Wisconsin. Significant effort and resources have been focused on building the IT Talent pipeline in the region. As predicted by Schilling (2015), collaborative behavior is induced by a

[2] Calendar starts halfway through the year.

[3] Cancelled events are still listed in the calendar.

major technology shock. In Northeast Wisconsin, the trigger event was the robust growth of job demand. That is, the common realization that many companies were facing talent shortages. Before this point, each company thought they were unique in their difficulties to identify and attract technical talent. As a shared understanding of the problem emerged, this talent demand shock led to an informal collaboration network. With the hope of developing innovative outcomes relating to talent, a formal alliance was created. This relationship, in light of the original Schilling model, can be seen in Figure 8.

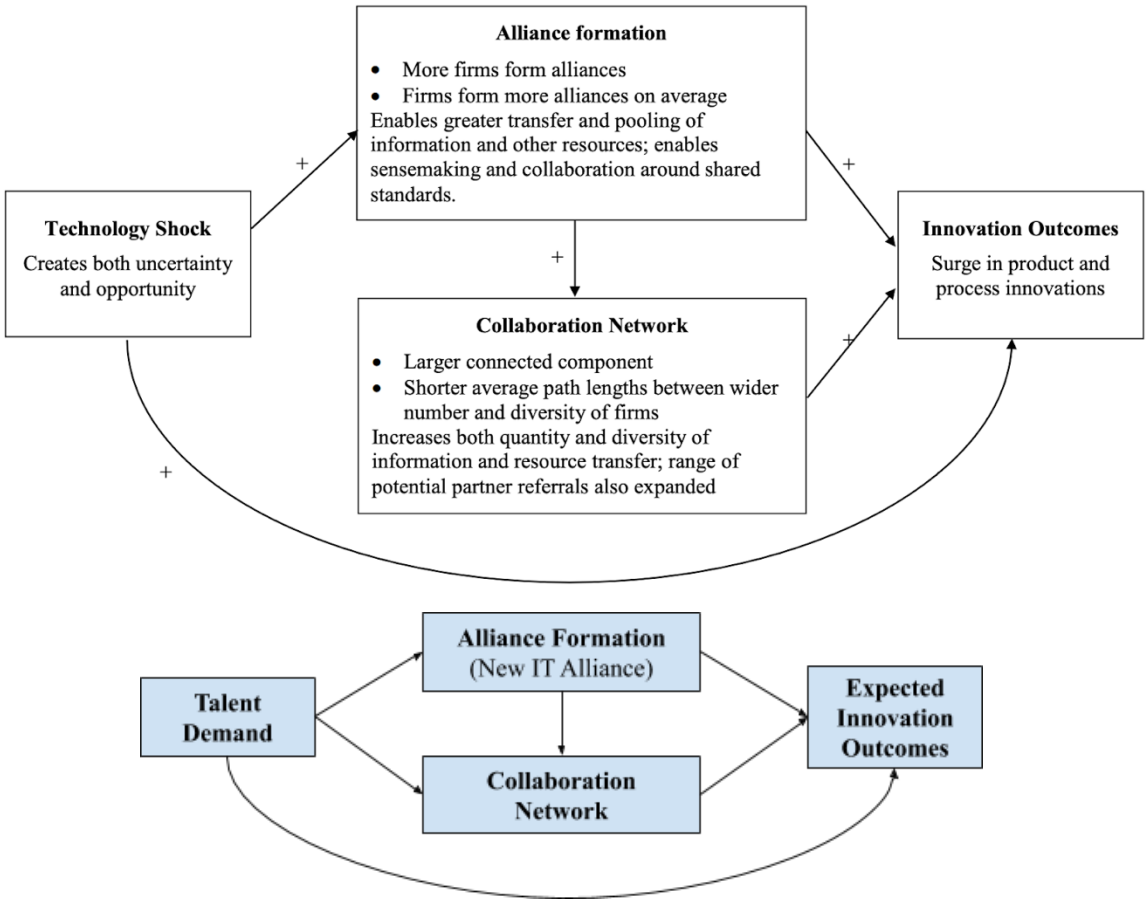


Figure 8: (Top) Original Schilling (2015) model and (Bottom) NEW Digital Alliance Collaboration

As noted earlier, the Northeast Wisconsin collaborative network is transitioning from the Adoptive phase into the Adaptive phase. Many projects are underway targeting distinct groups. In addition, new opportunities are continually identified. These are expected actions of a collaboration network. By moving to a more mature state, the network should increase both capabilities and value. Following guidance from the collaboration maturity model, the NEW Digital Alliance should recognize the need to mature and articulate strategies in support of culture, process, technology, and governance.

5.1 What are the inflection points that highlight transition from one phase to another phase?

In reviewing the previous five years of activities and maturity, several inflections points have been identified that signaled a transition in maturity. First, in 2015, an initial meeting occurred with senior IT leaders from across the Northeast Wisconsin region. This resulted in the recognition of collaboration potential and opportunities for collaboration and began the Exploratory phase of maturity. Second, in 2016, a director was hired. This resulted in the organization becoming more defined. The implementation of a strategic plan in 2017 shifted the NEW Digital Alliance and related collaboration space into an Adoptive phase. This phase was further enhanced in 2018 with the development of specific strategic-related KPIs. In 2019, a strengthened governance framework was established signifying the shift from Adoptive to Adaptive maturity. As can be seen in Figure 4, these inflection points tended to occur at the beginning of the move towards a new level of maturity and we contend that the inflection points were major drivers towards achieving the next level of maturity.

COVID-19 and the resulting economic and health uncertainty represent a new shock of the collaboration ecosystem of Northeast Wisconsin. The impact on the current alliance is unknown. Initially, IT-related hiring remains strong and sponsoring companies remain involved and committed to collaboration. However, it is unclear if companies will continue to be able to commit financially to the effort during the downturn in the economy. It is also unclear what grant opportunities may be available and whether the NEW Digital Alliance will be able to access them to help establish re-skilling efforts as we transition out of the pandemic.

5.2 How does maturity of the Alliance contribute to the maturity of the collaboration network?

While serving as the central point for collaboration and event organization, the ongoing focus on goals and objectives, culture, process, technology, and governance by the NEW Digital Alliance positively impacted the related collaboration network. Ongoing strategy and KPI development were driven by a board representing regional companies, economic development organizations, and higher education institutions. These representatives have a vested interest in the collaboration network remaining viable and impactful. Their involvement also helped ensure alignment between the activities and actions of both the Alliance and their respective organizations. Similarly, technology maturation facilitated greater collaboration. Finally, committees and subcommittees set goals and objectives in alignment with the regional strategies directed by the Alliance.

5.3 In what ways are innovative outcomes impacted by maturing collaborative alliances and networks?

Innovative outcomes are difficult to measure. While traditional metrics such as the number of companies involved in the NEW Digital Alliance, number of events, number of attendees, or job postings can offer insight into results of the Alliance and the collaborative network, they do not specifically represent innovative outcomes. As an example, a single job opening that is filled may or may not have resulted from the work and effort of the Alliance and related collaboration network. Anecdotally, companies will continue supporting the Alliance and partaking in the collaboration network if they feel they receive value for the time and money they are spending. This culture of loose affiliation has allowed the collaboration to take hold and grow but may also represent a barrier to future maturity. This culture also impedes formal tracking of innovative outcomes. This lack of measurable outcomes represents a challenge to the Alliance.

6. Conclusion

Given the tremendous impact of COVID-19 on our workplace, educational institutions, and society at large, understanding collaboration, collaboration networks, and innovative outcomes is critical. This timely study explores the formation and maturity of an alliance and related collaboration networks and the impact on innovative outcomes to solve real challenges and problems. Combining the concepts of collaboration networks and collaboration maturity, we present a first attempt at a new comprehensive framework that may help understand and direct new regional collaborative efforts. Further research will be needed to fully realize this framework and explore its applicability to different regional settings.

There are two key limitations to this study. First, this paper only explores a single collaboration network in a single geographic region. It would be beneficial to explore additional networks in the same geographic area or similar types of network in different geographical areas to increase generalizability. Second, this study is based on personal experiences and knowledge of the authors. Wider interviews and surveys could yield richer results and greater insight into decisions throughout the maturation of the NEW Digital Alliance and the associated collaboration network.

There are several paths for future research. First, issues of diversity (gender, racial etc.), have not been addressed in this research. As an innovative outcome to solve the talent shortage, research into how the Alliance and related network can target a more diverse base of participants in the technology community would be warranted. Efforts are under way to collect more data on diversity of IT students and the IT workforce in the region. Second, this study occurred mostly during periods of strong economic growth (2015-2020). Different economic conditions are likely to cause alliance and network developments to proceed differently. Finally, understanding the impact of working "safely from home" on collaboration alliances and networks could be explored. While the pandemic will pass, the long-term impacts may involve continued remote working for employees and members of organizations.

By identifying key inflection points, maturation relationship between a formal alliance and collaboration network, and impact on innovative outcomes, we hope to provide guidance to those struggling to solve new challenges. The value of collaboration can result in positive and innovative outcomes. This is required now more than ever.

7. References

Argyres, N., Rios, L.A., & Silverman, B.S. (2020). Organizational change and the dynamics of innovation: Formal R&D structure and intrafirm inventor networks. *Strategic Management Journal*, *41*(11), 2015-2049.

Case, T., Dick, G., Granger, M., & Akbulut, A.Y. (2019). Invited Paper: Teaching information systems in the age of digital disruption. *Journal of Information Systems Education*, *30*(4), 287-297.

Wisconsin Department of Public Instruction (DPI) (2019). Regional Career Pathways, New North. Retrieved December 9, 2020, from https://dpi.wi.gov/pathways-wisconsin

Eisenhardt, K.M. (1989). Building theories from case study research. *The Academy of Management Review*, *14*(4), 532–550. http://doi.org/10.2307/258557

Frankort, H.T.W., Hagedoorn, J., & Letterie, W. (2016). Learning horizon and optimal alliance formation. *Computational and Mathematical Organization Theory*, *22*(3), 212-236. http://doi.org/10.1007/s10588-015-9203-z

Getto, G., Franklin, N., & Ruszkiewicz, S. (2014). Networked rhetoric: iFixit and the social impact of knowledge work. *Technical Communication*, *61*(3), 185-201.

Jose, S. (2019). Innovation in recruitment and talent acquisition: A study on technologies and strategies adopted for talent management in IT sector. *International Journal of Marketing and Human Resource Management*, *10*(2), 1-8.

Kappelman, L., Johnson, V., Maurer, C., Guerra, K., McLean, E., Torres, R., Snyder, M., & Kim, K. (2019). The 2019 SIM IT issues and trends study. *MIS Quarterly Executive*, *19(*1), 69-104. https://doi.org/10.17705/2msqe.00026

Komarov, M., Kazantsev, N., & Grevtsov, M. (2014). Increasing the adoption of social collaboration software. *2014 IEEE 16th Conference on Business Informatics*, 54-59. https://doi.org/10.1109/cbi.2014.36

Martynov, A. (2019). Sequencing of emphases on technology alliances and internal R&D: The effects of the market cycle. *Long Range Planning*, *52*(1), 117-133. https://doi.org/10.1016/j.lrp.2018.10.004

Matzek, M. (2018). Hire Tech: Recruiting to Fill the Many Faces of IT Roles. *Insights on Technology*, November 2018. Insight Publications. http://www.insightdigital.biz/i/1045475-november-2018/23

Milagres, R., & Burcharth, A. (2019). Knowledge transfer in interorganizational partnerships: What do we know? *Business Process Management Journal*, *25*(1), 27-68. https://doi.org/10.1108/bpmj-06-2017-0175

Morgan, J. (2012). *The collaborative organization: A strategic guide to solving your internal business challenges using emerging social and collaborative tools*. McGraw-Hill.

Morgan, J. (2014). *The future of work: Attract new talent, build better leaders, and create a competitive organization*. John Wiley & Sons.

National Center for Educational Statistics. (2018). Digest of Education Statistics, 2018. Retrieved September 26, 2020, from https://nces.ed.gov/programs/digest/d18/ (See Table 322.10).

New North (2020), *New North website.* Retrieved December 17, 2020. https://www.thenewnorth.com

Paulk, M.C., Curtis, B., Chrissis, M.B., & Weber, C.V. (1993). Capability maturity model, version 1.1. *IEEE Software*, *10*(4), 18-27. https://doi.org/10.1002/0471028959.sof589

Schilling, M. A. (2015). Technology shocks, technological collaboration, and innovation outcomes. *Organization Science*, *26*(3), 668-686. https://doi.org/10.1287/orsc.2015.0970

U.S. Bureau of Labor Statistics. (2020). Employment by Major Occupational Group. Updated September 1, 2020. Retrieved December 16, 2020, from https://www.bls.gov/emp/tables/emp-by-major-occupational-group.htm

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage Publications.

Appendix A: Collaboration Network

Organization	**Collaborative Role**
Amplify Oshkosh	Local organization started by the Oshkosh Chamber of Commerce to promote the confluence and capabilities of technology in Oshkosh.
Central Wisconsin IT Alliance (CWITA)	Group of action focused employers in central Wisconsin working together to enhance the image of IT careers and position the region as a hub for IT opportunities.
Chambers of Commerce	Advocates of the local business and industry community, many of whom have a focus on workforce development.
Code.org	National nonprofit dedicated to expanding access to computer science in schools and increasing participation by women and underrepresented minorities.
Cooperative Education Service Agency (CESA)	Serve educational needs in all areas of Wisconsin by serving as a link between school districts, and between school districts and the state.
Department of Public Instruction (DPI)	State organization that sets standards for public schools in Wisconsin.
Department of Workforce Development (DWD)	State organization, developing youth and adult apprenticeships in IT
Employers	Organization with IT talent needs, many of whom are engaging with and providing financial support for one or more non-profit organizations within the collaboration network.
Microsoft	Microsoft's TechSpark is focused on regional internet connectivity, digital skills development, career skills development, nonprofit support and digital business transformation.
NEW Digital Alliance	Regional non-profit, funded by local employers, to help attract, develop & retain diverse IT talent in Northeastern Wisconsin to support economic growth.
NEW Manufacturing Alliance	Group of manufacturers, educators, workforce development, chambers of commerce and state organizations working to promote manufacturing in the Northeast WI region.
New North	Regional marketing and economic development organization representing the 18 counties of Northeast Wisconsin.
Northeast Wisconsin Education Alliance (NEW ERA)	Alliance that fosters regional collaboration among public colleges and universities in Northeast Wisconsin.
School Districts	Many of the school districts across the region were represented.
TEALS	National non-profit whose mission is to get computer science in every high school, with a focus on AP level CS classes.
Women in Technology (WIT) Wisconsin	Regional non-profit focused on initiatives designed to attract, grow and retain women and girls in technology related careers.

Appendix B: 2019 Alliance and Network Events

January 2019
Jan 17 Business Intelligence Best Practices
Jan 24 NEW CS Advisory Board
Jan 24 Ideas Amplified: Exploring the World of Virtual Reality & Augmented Reality

February 2019
Feb 05 Salesforce Technology: Wisconsin Non-Profit Group (Appleton) Kickoff Meeting
Feb 07 Counselors for Computing: Professional Development in Emerging Careers
Feb 21 CS Fundamentals Workshop
Feb 21 LinkedIn Best Practices
Feb 21 Amplify Member Mixer
Feb 21 Fox Valley Business Intelligence and Analytics Meetup
Feb 22 WIT@Work Breakfast Series - How to Leverage Social Media for Talent Acquisition and Talent Retention
Feb 28 Foxconn R&D center in Green Bay topic of Tech Council network luncheon on Feb. 28

March 2019
Mar 05 EDCi Taste of Technology
Mar 12 NEW CS Advisory Board
Mar 12 Salesforce Nonprofit User Group Meeting
Mar 12 WIT Networking and Social Event
Mar 14 Fox Valley Business Intelligence and Analytics Meetup
Mar 18 Wisconsin Tech Summit 2019
Mar 19 Virtual/Augmented Reality: Business Applications
Mar 20 Ideas Amplified: The Ins & Outs of E-Recycling
Mar 28 LinkedIn Best Practices

April 2019
Apr 02 Fox Valley Business Intelligence and Analytics Meetup
Apr 07 Hack Appleton
Apr 09 Salesforce for Nonprofits User Group - Problem-Solving Workshop
Apr 23 Building Blocks to a Highly Optimized Website
Apr 23 Building Blocks of Stellar Digital Marketing
Apr 25 Amplify Oshkosh- What is Blockchain and How Does it Work?
Apr 26 WIT@Work Breakfast Series - How to Stay Current With The Latest Technology
Apr 26 Innovation Challenge on the Aging Community
Apr 27 Innovation Challenge on the Aging Community
Apr 30 TechTalk 2019 @ Moraine Park Technical College

May 2019
May 09 Fox Valley Business Intelligence and Analytics
May 14 Building Blocks to a Highly Optimized Website
May 17 WIT@Work Breakfast Series - Learn how to persuade, inform and plan through enhanced communications
May 21 Midwest Association for Information Systems 14th Annual Conference (MWAIS 2019)
May 23 WIT4Girls Community Info Event
May 30 A Beginner's Guide to Salesforce Trailhead, In Action!

June 2019
Jun 08 Headway Open House Happy Hour
Jun 09 Robotics & STEM Camp: Session 1
Jun 11 Building Blocks to Stellar Digital Marketing
Jun 13 2019 Digital Learning Opportunities in NEW
Jun 13 6 Tips for Increasing and Managing Leads
Jun 13 Salesforce for Non-profits Group – Reports & Dashboards Workshop and Problem Solving
Jun 13 Fox Valley Business Intelligence and Analytics

Jun 16 Robotics & STEM Camp: Session 2
Jun 17 FVTC Technology Camp
Jun 17 Tech Titans Mobile App Academy
Jun 18 First Meeting of the Green Bay "Titletown" PUG
Jun 19 Microsoft Dynamics 365/CRM User Group - Security
Jun 19 Amplify Celebration Event
Jun 23 Video Game Programming Camp: Level 1 (Session 1)
Jun 24 FVTC Technology Camp
Jun 25 CRM A-Z: From Selection and Implementation to User Adoption and ROI

July 2019
Jul 07 Video Game Programming Camp: Level 1 (Session 2)
Jul 11 WIT Annual Meeting
Jul 15 FVTC GirlTech Summer Camp
Jul 16 Ryan Tischer - Strategic Cloud Engineer at Google
Jul 18 Fox Valley Business Intelligence and Analytics
Jul 21 Video Game Programming Camp: Level 2
Jul 23 Northeast Wisconsin Collaborative Intern Event - Members Only!
Jul 23 How to Do Customer Interviews & Gain Valuable Insights
Jul 30 IT Career Pathways Discussion
Jul 30 Code.org CS Fundamentals Intro Workshop
Jul 31 Code.org CS Fundamentals Deep Dive Workshop

August 2019
Aug 05 2019 NSA/NSF GenCyber Teacher Camp
Aug 06 THAT Conference
Aug 08 Amplify Member Mixer
Aug 12 Startup WI Week Happy Hour - Green Bay Community Partnership
Aug 21 Tech and Tailwinds
Aug 22 Advancing Cybersecurity in the Industry, Energy, Water Nexus
Aug 22 Fox Valley BI & Analytics - Google Cloud Platform's Big Query and BI & Analytics
Aug 24 2019 UWGB Cyber Teacher/Educator Workshop
Aug 27 IT Career Pathways Discussion

September 2019
Sep 09 The Story of Starting a Cryptocurrency Exchange
Sep 12 Fox Valley BI & Analytics - Data Visualization Best Practices Part II
Sep 17 Challenges and Opportunities: The Future of the Internet of Things in Wisconsin and the Midwest
Sep 17 NEW AITP - Navigating Shift Creek: A Guide to Career Transition in Today's Business World
Sep 17 ReactJS: Managing State with Hooks & Context
Sep 18 Microsoft Dynamics 365/CRM User Group: Sept. 2019 – Training Topic SharePoint and OneNote Integrations
Sep 18 RAMS - Documentary Screening to Benefit the Boys & Girls Club Arts Initiative
Sep 19 Titletown HDI September Meeting - Cloud/Cyber Security
Sep 19 Ideas Amplified: Data Ethics
Sep 26 Code.org CS Fundamentals Intro Workshop
Sep 26 TechTalent Summit - Members Only!
Sep 27 WIT @ Work Breakfast Series - Feel to Heal: The Truth about Health and Happiness – September

October 2019
Oct 04 Great Lakes Analytics Conference
Oct 09 IT Leadership Academy: Critical IT Communication Skills
Oct 10 Pulp, paper and packaging: Future of industry topic of Oct. 10 Tech Council luncheon in Appleton
Oct 10 Pulp, paper and packaging: Future of Industry
Oct 10 Fox Valley BI & Analytics - Panel Discussion "How are you Managing Self Service BI and Analytics?"
Oct 15 NEW CS Advisory Board: IT Career Pathways
Oct 15 NEW AITP – Design Thinking: From Theory to Application
Oct 18 Code.org CS Fundamentals Deep Dive Workshop

Oct 22 React.js 101 - Build Your First React App! (Beginners)
Oct 25 WIT @ Work Breakfast Series October - Intelligent Automation -- Successes through Automation
Oct 29 Envision 2019

November 2019
Nov 04 Meet the Meetups
Nov 04 Meet the Meetups
Nov 06 NEW IT Alliance TechTalk 2019 @ UW Oshkosh
Nov 08 Startup Wisconsin Week
Nov 14 Fox Valley BI & Analytics - Building a Data Insights Practice or Center of Excellence
Nov 15 Ethical IT Conference
Nov 19 NEW AITP presents Brent Walkow, Comedian/Entertainer
Nov 21 NEW Connect IT Job and Career Fair
Nov 22 WIT @ Work Breakfast Series - Productivity Beyond the Information Age – November

December 2019
Dec 03 UX Design and Development Education - Community Discussion
Dec 07 Northeast Wisconsin Code Camp 2019
Dec 11 Microsoft Girls Workshop: FVTC Hour of Code
Dec 12 Holiday Member Mixer - Amplify Oshkosh
Dec 12 Fox Valley BI & Analytics - End of Year "All Things Data" Networking Event
Dec 16 Digital Fertilize Presents: 2019 Year-End Celebration

Author Biographies

John Michael Muraski is an assistant professor of information systems in the College of Business at the University of Wisconsin – Oshkosh. After 20 years of working and consulting in industry, he transitioned into higher education and earned his DBA at the University of Wisconsin – Whitewater. Over the last 10 years, he has taught at both the undergraduate and graduate level and led the development of several new programs, include the ERP and Business Analysis programs at UW Oshkosh. Dr. Muraski conducts research into two main areas: (a) new technology characteristics that enhance infusion between a software and an employee in an organizational context and (b) challenges and opportunities relating to high school and college student reluctance to explore technology-related educational pathways.

Jakob Holden Iversen is a Professor of Information Systems with 20+ years of experience in higher education teaching, research, and administration. He currently serves as Associate Dean for the College of Business at University of Wisconsin Oshkosh. He earned his Ph.D. at Aalborg University in Denmark with a focus on Software Process Improvement. He has been at UW Oshkosh since 2000 where he has taught a range of courses at both the undergraduate and graduate level in Information Systems. Throughout his time at the University of Wisconsin – Oshkosh he has worked on creating new degree programs in collaboration with other departments as well as other universities in the University of Wisconsin System. This included leading the development of the Interactive Web Development Management major, the Information Systems minor, and participation in the creation and redesign of a number of other degree programs. As a scholar he has published in leading academic journals including MIS Quarterly, and is the co-author of two textbooks on Mobile app development and C# programming. A recent focus for his research and teaching has been in the area of mobile app development and usage.

Kimberly Jean Iversen Kim Iversen is the founding Director for the Northeast Wisconsin Digital Alliance (formerly NEW IT Alliance), a position she has held since November, 2016. In that time, she has established the NEW Digital Alliance as the leading regional IT & Digital collaborative network focused on growing the digital talent pipeline and the digital ecosystem in the region. In her role as director for the NEW Digital Alliance, Iversen works to coordinate the efforts across K-12 and Higher Ed to attract, retain, and develop students interested in joining the IT and digital workforce in the 18-county region of Northeast Wisconsin. Prior to joining the NEW Digital Alliance, Iversen spent 13 years at Kimberly-Clark in various roles within the IT program management office, including an effort to resolve IT resource constraints globally within KC projects. Iversen was a founding member of Women in Technology Wisconsin, where until recently she also chaired the WIT4Girls committee, which is committed to introducing young women in 6-12th grade to the exciting opportunities in IT. Iversen holds a Bachelor of Science in Biochemistry from Oklahoma State University, along with a Master of Science in Molecular Biology and Human Genetics from Aarhus University, Denmark, and a Master of Science in Information Systems from the University of Wisconsin Oshkosh.

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 4

Date: 01-31-2021

Human Activity Recognition: A Comparison of Machine Learning Approaches

Loknath Sai Ambati
Dakota State University, LoknathSai.Ambati@trojans.dsu.edu

Omar El-Gayar
Dakota State University, Omar.El-Gayar@dsu.edu

Abstract

This study aims to investigate the performance of Machine Learning (ML) techniques used in Human Activity Recognition (HAR). Techniques considered are Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Stochastic Gradient Descent, Decision Tree, Decision Tree with entropy, Random Forest, Gradient Boosting Decision Tree, and NGBoost algorithm. Following the activity recognition chain model for preprocessing, segmentation, feature extraction, and classification of human activities, we evaluate these ML techniques against classification performance metrics such as accuracy, precision, recall, F1 score, support, and run time on multiple HAR datasets. The findings highlight the importance to tailor the selection of ML technique based on the specific HAR requirements and the characteristics of the associated HAR dataset. Overall, this research helps in understanding the merits and shortcomings of ML techniques and guides the applicability of different ML techniques to various HAR datasets.

Keywords: Human Activity Recognition, Machine Learning, Performance, Healthcare, Benchmark.

Please note: A prior version of this article received a best paper award submitted for the 2020 Midwest Association for Information Systems (MWAIS) in Des Moines, Iowa which was cancelled due to COVID-19 pandemic. The article has been expanded and received a second round of reviews. We congratulate the authors.

DOI: 10.17705/3jmwa.000065

1. Introduction

The popularity of wearable technology has increased over the recent years (Iqbal et al. 2018). Applications such as self-management aimed at managing disease condition, and self-care for facilitating health and wellbeing have adopted wearable technology to improve health and wellbeing for users. Most of these wearable devices contains sensors such as accelerometers, gyroscopes, magnetometers, heart rate sensors and similar sensors embedded for successful human activity recognition (HAR). The availability of data coupled with the wide ranging applications of HAR resulted in HAR garnering significant attention in academia and in practice (Qin et al. 2020).

In that regard, machine learning and data mining techniques have proved beneficial in extracting features and classifying HAR data (Ramasamy Ramamurthy and Roy 2018). Most of the HAR applications in the market today are striving to improve their performance by utilizing ML techniques and have demonstrated success in terms of performance metrics such as classification accuracy and processing speed (Meyer et al. 2016). Further, HAR data are often characterized by a number of attributes, such as activity type, sensor type, preprocessing steps, and position of sensor on a specific body area. Such diverse characteristics makes HAR particularly challenging and is a persistent driver for ongoing research. Specifically, prior research has mainly focused on developing and improving novel ML models for a number of activities in unique environments and populations (Wang et al. 2019), e.g., elderly individuals in a home care environment (Chen et al. 2017). Further, most of HAR literature is concentrated around improving the HAR performance by considering a single dataset and a specific ML classification technique which limits the generalizability of the findings (Baldominos et al. 2019; Nabian 2017). Although some attempts have been made to compare various ML techniques on multiple HAR datasets (Dohnálek et al. 2014; Li et al. 2018), their focus is often limited to either improving feature learning or finding optimal techniques with the best tradeoff between speed and accuracy rather than a comprehensive approach that could be employed to understand the performance of ML techniques and map them to the characteristics of various HAR data sets.

Accordingly, this research study aims to analyze the performance of different ML classification techniques using various HAR datasets. As HAR sensor data sets can vary significantly with respect characteristics such as sampling frequency, type of activities performed, number of sensors, sensor types and sensor positions, these variations in characteristics have been demonstrated to impact ML techniques hyper parameter tuning, classification performance, and run time. (Baldominos et al. 2019; Dohnálek et al. 2014; Nabian 2017; Wang et al. 2019). This research extends the understanding of the performance of ML classification techniques on HAR data. The significance of this research is both theoretical and practical. From a theoretical point of view, this research helps to understand the merits and shortcomings of ML techniques that could help future researchers figure out how to improve the ML classification techniques for HAR datasets. From a practical point of view, this research helps in guiding the applicability of different ML classification techniques to HAR datasets. Altogether, the research on HAR performance improvement can remarkably facilitate self-management and self-care interventions. In addition, these improvements extend beyond the medical and healthcare domains to other context, wherever the detection of human activity is vital.

The remainder of the paper is organized as follows: a brief literature review is presented in section 2, while section 3 describes the methodology including the characteristics of the dataset and the details of the analysis process. Section 4 illustrates the results obtained from the analysis and section 5 summarizes and discusses the results by comparing with extant literature. Finally, section 6 concludes by summarizing the key contributions, limitations, and suggests directions for future research.

2. Literature Review

2.1 Human Activity Recognition

Raw data obtained from the wearable sensors undergo a number of steps as demonstrated by the Activity Recognition Chain (ARC) model (Bulling et al. 2014) for classifying human activities as shown in Figure 1. In this model, the first step involves sampling the raw data obtained from different sensors with multiple dimensions, before it undergoes preprocessing, segmentation, feature extraction, and finally, classification. Among these steps, feature extraction requires deep domain expertise in the field. Therefore, researchers tend to depend on domain experts for feature engineering and extraction. Utilizing the resultant engineered features with ML and deep learning techniques, the activities are classified into specific human activities (Saha et al. 2018).

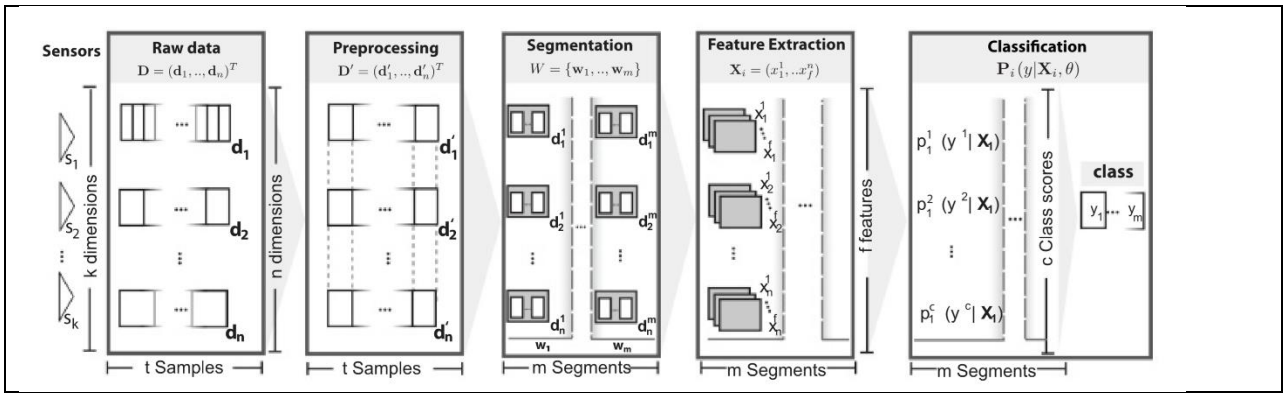


Figure 1: Activity Recognition Chain (ARC) Model (Bulling et al. 2014)

HAR research focuses predominantly on classifying human activities using ML techniques or/and preprocessing the data (Baldominos et al. 2019; Jain and Kanhangad 2018; Nabian 2017; Ronao and Cho 2017; Sousa et al. 2017). HAR using smartphones is a popular sub-field where there is abundant of literature that deals with improving HAR classification using various innovative pre-processing and ML techniques (Anguita et al. 2013; Jain and Kanhangad 2018; Micucci et al. 2017; Nakano and Chakraborty 2017; Ronao and Cho 2014, 2017). Few studies tried to improve the HAR classification obtained from inertial sensors using hyper parameter tuning of ML techniques (Gaikwad et al. 2019; Garcia-Ceja and Brena 2015; Seto et al. 2015). Others have also focused on the problems and difficulties associated to segmentation and proposed solutions to tackle these problems (Kozina et al. 2011; Oresti Banos 12:19:30 UTC). These studies have shown a partial effect of segmentation on the performance of ML techniques. Similarly, few studies demonstrated that the window size affects the HAR classification, e.g., 1-2 second interval results in optimal tradeoff between accuracy and recognition speed (Banos et al. 2014; Ni et al. 2016).

2.2 Comparative analysis

Most of the HAR literature is concentrated around improving the performance HAR using a single dataset and a specific ML technique which limits the generalizability of the findings. There are some studies that tried to consider various ML techniques (Akhavian and Behzadan 2016). These types of analyses compare various ML techniques to identify the most suitable technique for a HAR dataset (Baldominos et al. 2019; Nabian 2017). These studies used a single dataset to understand the relation between few HAR characteristics such as sensor position, type of activity and hyper parameter tuning on ML performance. Recent studies on HAR has shown evidence that no prior preprocessing of raw sensor data has shown reasonable ML performance especially in a comparative study (Dohnálek et al. 2014). Although some attempts have been made to compare various ML techniques on multiple HAR datasets (Dohnálek et al. 2014; Li et al. 2018), their focus is often limited to either improving feature learning methods or finding optimal techniques with the best tradeoff between speed and accuracy. Accordingly, there is a need for a comprehensive study to evaluate and benchmark the performance of various ML techniques with different HAR datasets and map the characteristics of various HAR datasets to appropriate ML techniques. Prior work on HAR data partly tried to address this gap by comparing multiple HAR datasets with the accuracy score of different ML techniques (Ambati and El-Gayar 2020). We aim to extend this work by collectively considering multiple HAR datasets, the type of activities being classified, the performance of an expanded portfolio of ML techniques, and the use of an expanded set of performance metrics to get more insights in understanding the ML techniques and their relation to HAR data in conjunction with the extant literature. These insights can help future researchers in designing a robust and comprehensive framework/model depending on the HAR application.

3. Methodology

3.1 Datasets

We used three HAR datasets in a manner that captures the diversity of characteristics commonly present in various datasets. The first two datasets (Pampa2 and mHealth) are from the University of California, Irvine (UCI) data repository.

The datasets were chosen in such a way that they are distinct in terms of sensors utilized, sampling frequency, activity environment and similar attributes, and are utilized in prior research (Anguita et al. 2013; Gaikwad et al. 2019; Garcia-Ceja and Brena 2015; Nakano and Chakraborty 2017). This makes these datasets unique and appropriate for utilizing them to benchmark various ML techniques. The third dataset is selected from the SWELL project supported by the Dutch national program COMMIT (Shoaib et al. 2014). Table 1 presents the data sets and their characteristics. 3D accelerometers, 3D gyroscope, and 3D magnetometer are the common sensors employed in all the three datasets. These sensors have become a basic functionality for wearable devices that attempt to recognize human activity. 3D accelerometer helps in recognizing the speed with which the user is moving in all three dimensions, 3D magnetometer helps in recognizing the orientation of the user with respect to earth's magnetic north, and 3D gyroscope helps in recognizing the angular velocity of the user. Other than these three sensors, each dataset has some unique sensors when compared to each other. For example, Pamap2 dataset has a heart rate monitor and a temperature sensor, while mHealth has an ECG sensor and SWELL has a linear acceleration sensor. With respect to data collection, Pamap2 relies on wireless IMU's, while mHealth uses wearables, and SWELL uses smartphones to collect the data. All activities in a particular dataset are conducted for approximately the same amount of time and are represented evenly in the data sets. Therefore, data imbalance does not constitute an issue. When data size is considered, Pampap2 dataset is the largest dataset with a 519,185-record size, while SWELL stands second with a 189,000-record size, and MHealth being the smallest with 102,959 record size.

Dataset	Sensors	Sensor Position	Activities performed	Dataset description	Sampling Frequency
Pamap2	3 Colibri wireless IMUs (inertial measurement units) and BM-CS5SR (HR monitor) – Accelerometer, Gyroscope, magnetic sensor and temperature sensor.	wrist, chest and side ankle.	lying, sitting, standing, walking, running, cycling, Nordic walking, watching TV, computer work, car driving, ascending stairs, descending stairs, vacuum cleaning, ironing, folding laundry, house cleaning, playing soccer and rope jumping.	9 subjects (1 female and 8 male) aged 27.22 (+-) 3.31 years performed the 12 mandatory activities and 6 optional activities for 2-3 minutes.	100 samples/sec
Mhealth	accelerometer, a gyroscope, a magnetometer and ECG (Shimmer2 [BUR10] wearable sensors).	chest, right wrist and left ankle	L1: Standing still (1 min), L2: Sitting and relaxing (1 min), L3: Lying down (1 min), L4: Walking (1 min), L5: Climbing stairs (1 min), L6: Waist bends forward (20x), L7: Frontal elevation of arms (20x), L8: Knees bending (crouching) (20x), L9: Cycling (1 min), L10: Jogging (1 min), L11: Running (1 min), L12: Jump front & back (20x)	10 volunteers of diverse profile performed 12 physical activities for about 1 min	50 samples/sec
SWELL	accelerometer, a gyroscope, a magnetometer, and a linear acceleration sensor (Samsung Galaxy SII (i9100) smartphone).	upper arm, wrist, two pockets, and belt position	walking, sitting, standing, jogging, biking, walking upstairs and walking downstairs	10 participants performed 7 activities for 3-4 minutes. All are male with ages 25-30.	50 samples/sec

Table 1. Dataset Characteristics

3.2 Analysis

We compared different ML techniques using a number of HAR datasets (Pamap2, mHealth and SWELL) across various ML performance metrics. The ML techniques are Naïve Bayes, Support Vector Machine (SVM) with linear kernel, K-Nearest Neighbor (KNN), Logistic Regression, Stochastic Gradient Descent (SGD), Decision Tree, Decision Tree with entropy, Random Forest, Gradient Boosting Decision Tree (XGBoost), and NGBoost algorithm. Although, deep learning techniques such as neural network based algorithms are attracting popularity over the recent years, they tend to over fit in the case of HAR data (Jobanputra et al. 2019). Moreover, the runtime of each dataset over various ML techniques is already high when run on Python Jupyter Notebook using eight-generation intel i7 processor considering the data is not extensively preprocessed, therefore, applying neural network-based techniques on these large HAR datasets would significantly increase runtime. Considering these circumstances, neural network techniques are not implemented in this research. Although accuracy is the most popular ML performance metric in HAR (Li et al. 2018), we utilized additional metrics such as precision, recall, F1 score, support and runtime to facilitate an in-depth analysis. Table 2 shows the description of ML metrics employed for evaluating the ML performance.

All the three datasets are minimally preprocessed by addressing the missing values and excluding the data during the transient stage (transition from one activity to another) based on timestamp and standardizing the format of the data such

that it would be easier to implement the ML techniques and interpret the results obtained from the analysis. Specifically, the data is standardized in a manner such that each row represents the sample values for each sensor for a specific sampling time, and each column represents a sensor except for timestamp, subject ID, and classification activity. After standardizing the format of each dataset with minimal preprocessing, we split the data into training (70%) and test data (30%) for each dataset. Once all the datasets are split into training and test data, we train (including hyperparameters tuning) of the various ML techniques using each of the datasets. Then, we evaluate each ML technique with the considered performance metrics on each dataset.

ML Metrics	Accuracy	Precision	Recall	F1 score	Predicting Run Time
Definition	The ratio of number of correct predictions to the total number of predictions.	The ratio of number of correctly predicted positive values to the total predicted positive values.	The ratio of correctly predicted positive values to the total number of positive values.	Harmonic mean of precision and recall.	Time taken for target classification using test data.
Formula	(TP+TN)/(TP+TN+FP+FN)	TP/(TP+FP)	TP/(TP+FN)	2*(Precision*Recall)/ (Precision+Recall)	-

Table 2. Description of ML metrics

Where:

- True Positive (TP) - Number of correctly predicted positive values.
- True Negative (TN) - Number of correctly predicted negative values.
- False Positive (FP) – Number of predictions that interpret negative values as positive values.
- False Negative (FN) – Number of predictions that interpret positive values as negatives.

To investigate the potential tradeoff between classification performance and prediction runtime, we identify the Pareto efficient ML techniques for each of the datasets. Pareto efficiency is a concept where no individual criterion can be declared better without a sacrifice in one of the other criterion (Bokrantz and Fredriksson 2017). Accuracy is used as the metric representing classification performance, while prediction run time is measured in seconds.

3.2.1 Classification performance by activity

To get a better understanding of the performance of the various ML techniques using the various datasets, we considered three cases as follows.

Individual activities: In this case, the target variable is represented as a categorical variable where each category representing one activity such as sitting, standing, running, and lying for each dataset. This type of grouping is very popular in comparative analysis for understanding each activity and the respective effect of ML technique. It provides for the most fidelity as all activities are accounted for. However, it allows for the number of activities (classes) to vary among the three datasets which may compound the comparative analysis of the performance of various ML techniques.

Grouped activities: To address the aforementioned issue and considering that differentiating between sitting and standing, and between walking fast and running, and similar differentiation can be very difficult to obtain (Gjoreski et al. 2014), we conducted another set of experiments where all the activities in each dataset are divided into two categories namely locomotive activities and stationary activities. Activities where the user is staying idle with no physical movement such as sitting, standing, and lying are considered stationary activities. All other activities which require the user to perform a physical movement such as, but not limited to walking, running, jumping, climbing stairs and similar activities are categorized as locomotive activities. It is assumed that since all the locomotive activities share a similarity that the sensor movement is dynamic and similarly, stationary activities share a similarity that all the sensor movement would be idle, it should alleviate the problem of differentiating similar activities. This categorization should guide us towards understanding more towards the type of activity and how it is going to affect the ML techniques and their performance, respectively.

Common activities: Another possibility for standardizing the activities across datasets while maintaining as much fidelity as possible (in terms of the number of activities/classes) considered, we conducted an additional experiment where we included the activities that are common in all three datasets. The common activities in all the datasets are walking, sitting, standing, running, cycling, and climbing stairs.

4. Results

4.1 Classification of individual activities

Table 3 depicts the performance of the various ML techniques on the three data sets. With respect to accuracy, the performance of ML techniques irrespective of the datasets in the order of best performance are XGboost, Random Forest, KNN, SVM, Decision Tree with entropy, Decision Tree, NGBoost, Logistic Regression, Naïve Bayes, and SGD. There are three exceptions to this observation, Naïve Bayes technique performed better in the case of SWELL when compared to Logistic Regression, SGD performed better in the case of Mhealth when compared to Naïve Bayes technique, and Random Forest, KNN, and SVM performed better than XGboost in the case of mHealth.

With respect to precision, recall, and F1 Score, the performance of ML techniques irrespective of the datasets in the order of best performance are KNN, SVM, Random Forest, XGboost, Decision Tree with entropy, Decision Tree, Logistic Regression, Naïve Bayes, and SGD. There are two exceptions to this observation, Naïve Bayes technique performed better than Logistic Regression in the case of SWELL and SGD performed better than Naïve Bayes in the case of the mHealth dataset.

The performance of ML techniques irrespective of the datasets in the order of least runtime, Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, and KNN. There is one exception to this observation, XGboost has lower run-time when compared to Naïve Bayes in the case of Pamap2 dataset.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.901	0.999	0.999	0.9	0.92	0.999	**0.999**	0.999	0.999	0.936
	SWELL	0.879	0.996	0.998	0.847	0.855	0.975	0.977	0.995	0.999	0.881
	MHealth	0.521	0.965	0.991	0.629	0.738	0.911	0.918	0.939	0.934	0.875
Precision	Pamap2	0.91	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.85	0.98	0.98	1	1	0.88
	MHealth	0.52	0.97	0.99	0.63	0.72	0.91	0.92	0.94	0.94	0.87
Recall	Pamap2	0.90	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.86	0.98	0.98	1	1	0.87
	MHealth	0.52	0.97	0.99	0.63	0.74	0.91	0.92	0.94	0.93	0.87
F1 Score	Pamap2	0.90	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.85	0.98	0.98	1	1	0.88
	MHealth	0.55	0.97	0.99	0.62	0.65	0.91	0.92	0.94	0.93	0.87
Predicting Run-Time (s)	Pamap2	7.281	639.2	12,860	0.145	0.15	0.182	**0.149**	1.518	6.923	401.765
	SWELL	1.026	303.271	6,527	0.062	0.046	0.087	0.072	0.599	2.638	281.942
	MHealth	0.398	232.583	488.51	0.021	0.02	0.03	0.024	0.262	2.625	122.888

Table 3. Performance metrics of ML techniques for individual activities

4.2 Classification of grouped activities

As shown in Table 4, all ML techniques performs better using the Pamap2 dataset on all performance metrics except the runtime when compared to the other two datasets. When we consider accuracy, precision, recall and F1 score, the general trend that is followed by the ML techniques irrespective of the datasets is, Logistic Regression, SGD, Naïve Bayes, NGBoost Decision Tree, Decision Tree with entropy, SVM, KNN, Random Forest, and XGboost. There is one exception where Logistic Regression performed better than Naïve Bayes in the case of Pampap2 dataset.

When prediction run-time is considered, the general trend followed by the ML techniques irrespective of the dataset in the order of the shortest to longest runtime is Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, NGBoost, and KNN. As expected, (with only two classes), the performance metrics of ML techniques for grouped activities is better when compared to the individual activities.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.966	1	0.999	0.999	0.999	1	1	1	1	1
	SWELL	0.99	0.999	0.999	0.97	0.97	0.998	0.998	0.999	0.999	0.998
	MHealth	0.989	0.991	0.999	0.834	0.814	0.998	0.998	0.999	0.999	0.996
Precision	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.83	0.80	1	1	1	1	0.99
Recall	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.83	0.81	1	1	1	1	0.99
F1 Score	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.82	0.80	1	1	1	1	0.99
Predicting Run-Time (s)	Pamap2	0.712	21.831	12965	0.09	0.053	0.102	0.108	0.596	1.058	118.688
	SWELL	0.343	11.216	6505.16	0.025	0.029	0.048	0.044	0.265	0.57	46.751
	MHealth	0.069	52.799	486.397	0.006	0.008	0.014	0.012	0.109	0.217	10.065

Table 4. Performance metrics of ML techniques for grouped activities

4.3 Classification of common activities

As shown in Table 5, all ML techniques perform better using the Pamap2 dataset on all performance metrics except the runtime when compared to the other two datasets. When we consider accuracy, precision, recall and F1 score, the general trend that is followed by the ML techniques irrespective of the datasets is, SGD, Logistic Regression, Naïve Bayes, NGBoost Decision Tree, Decision Tree with entropy, SVM, KNN, Random Forest, and XGboost. There is one exception where Logistic Regression performed better than Naïve Bayes in the case of Pampap2 dataset.

When prediction run-time is considered, the general trend followed by the ML techniques irrespective of the dataset in the order of the shortest-longest runtime is Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, NGBoost, and KNN. The performance metrics of ML techniques for common activities are better when compared to the individual activities and slightly lower when compared with the grouped activities.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.945	0.999	0.999	0.984	0.987	0.999	0.999	1	1	0.999
	SWELL	0.939	0.998	0.999	0.924	0.931	0.991	0.993	0.998	0.999	0.964
	MHealth	0.931	0.99	0.998	0.789	0.829	0.988	0.99	0.999	0.999	0.959
Precision	Pamap2	0.95	1	1	0.98	0.98	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.78	0.82	0.99	0.99	1	1	0.95
Recall	Pamap2	0.95	1	1	0.98	0.99	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.79	0.83	0.99	0.99	1	1	0.95
F1 Score	Pamap2	0.95	1	1	0.98	0.98	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.78	0.82	0.99	0.99	1	1	0.95
Predicting Run-Time (s)	Pamap2	1.374	50.455	3603.943	0.076	0.074	0.082	0.078	0.661	1.987	305.698
	SWELL	0.959	120.28	4645.316	0.05	0.046	0.077	0.066	0.514	2.221	203.605
	MHealth	0.121	31.922	152.443	0.01	0.008	0.013	0.013	0.109	0.557	28.638

Table 5. Performance metrics of ML techniques for common activities

5. Discussion

When Pamap2 dataset is employed, all the ML techniques performed to their best for all possible groupings of activities. The relatively large size of the data resulting from the higher sampling frequency, and the additional sensors (temperature sensor and heart rate monitor) utilized in the dataset, positively affected ML performance. This leads us to conclude that the overall improvement of ML performance metrics tends to be associated with the number of sensors and higher sampling rate employed to collect the HAR data. Generally, tree-based algorithms such as Random Forest, NGBoost, and XGboost outperformed Naïve Bayes, SGD and Logistic Regression (with an exception of KNN and SVM, as their runtime is very high for real time usage) in terms of ML performance metrics thereby attesting to the claims made by other studies that Tree based techniques perform better than other techniques in the field of HAR (Sánchez and Skeie 2018).

When performance metrics of ML techniques for individual activities and common activities are compared with grouped activities, all the ML techniques performed better when activities are grouped as locomotive or stationary activities. This supports the assertion that it is particularly challenging to differentiate similar activities among a particular group (stationary and locomotive). Although, this observation is expected, this comparison provides an additional dimension for comparing ML techniques behavior.

Although, previous studies achieved accuracies up to 0.97 and F1 score of 0.84 with just the wrist position using deep learning techniques (Baldominos et al. 2019), this study obtained much higher accuracies and F1 scores using less complex ML techniques compared to neural network based techniques. However, in every dataset utilized, a combination of three or more sensor positions were employed. Therefore, evaluating the performance for each sensor position separately would give more insights but drastically increases the complexity of the analysis given the additional consideration for the number and location of sensors. This can be further explored in the future research.

When run-time is analyzed, it is expected that the dataset having more data (both in terms of features as well as data collected) would take more time to run a particular ML technique. Accordingly, in all considered scenarios, the predicting run time for any ML technique is highest when Pamap2 dataset is employed, followed by SWELL, and mHealth. Usually, all the ML techniques take more time for training and takes less time for predicting. Naïve Bayes technique on the other hand, took the least time for training the model but took a relatively long time for prediction using the test data. Naïve Bayes model size (with respect to the number of model parameters to be estimated) is relatively small compared to the other ML techniques considered. Moreover, depending on the conditional independence assumption being true, the model converges very fast resulting in a low training run time and less data being required for training compared to the other datasets considered (Mark 2015). If we consider any real time application that requires recognition of human activity, the main concern for designing the application would be minimizing prediction run time while maintaining acceptable classification performance. This puts Naïve Bayes ML technique at a disadvantage. Further, Logistic Regression tend to have the shortest run time while KNN has the longest run time in most of the cases considered.

Considering the tradeoff between classification performance represented using accuracy we find that DT with Entropy appears on the pareto efficient frontier regardless of the data set. DT with entropy is also the sole ML technique that is Pareto efficient for the Pamap2 dataset. Random Forest and Logistic Regression are Pareto efficient for SWELL and MHealth, while XGBoost is Pareto efficient for SWELL only. The prevalence of tree-based ML techniques such as DT with entropy, Random Forest, and to some extent XG Boost further supports prior research tree based techniques perform better than other techniques in the field of HAR (Sánchez and Skeie 2018). Although, KNN has a relatively long run time, it exhibits the best classification performance irrespective of the HAR dataset. Interestingly, KNN and SVM are in effect Pareto efficient for MHealth. However, run time for these two techniques is in the order of four magnitude larger than the run time for the other techniques rendering a steep tradeoff between run time and classification performance. This, considering very high run time of SVM and KNN, neither of these techniques may be suitable for real time applications.

If an application is more interested in the accuracy, low on budget for additional sensors such as heart rate monitor, then XGBoost would be an ideal solution with some run time tradeoff compared to Random Forest. Similarly, if an application is more interested in short run time, then Decision Tree with entropy would be an ideal solution with a small tradeoff with the accuracy. But in most of the other cases, DT with Entropy is the optimal performer for any combination of weighted performance metrics selected. There can always be some exceptions such as in an application with very limited data, then KNN or XGBoost might be a better fit depending on the size of data.

In essence, depending on the requirements of the HAR application and data amount, we can choose the sensor types (Shoaib et al. 2014), sensor positions (Baldominos et al. 2019), ML techniques (Dohnálek et al. 2014), sampling frequency (Wang et al. 2019), and similar characteristics based on the insights provided in this research. The key insights pointed out in this study:

- It is relatively difficult to differentiate similar activities among a particular group (stationary and locomotive).
- High sampling frequency improves the ML performance metrics (Accuracy, precision, recall, F1 score, and support), however, it will take a toll on the run-time.
- DT with Entropy stands out to be the optimal performer in most cases of the HAR applications.
- Ensemble techniques outperforms traditional ML techniques in terms of ML performance metrics except run time for HAR data.
- Naïve Bayes technique is efficient when there are more activities involved.
- Naive Bayes technique takes the least time for training the data and build the model but takes a heavy toll in time taken for predicting the test data.

6. Conclusion

In this study, the performance of various ML techniques used for HAR are evaluated using ML performance metrics such as accuracy, precision, recall, F1 score, and run time on multiple HAR datasets. We investigated the relationship of different HAR dataset characteristics to the performance of various ML techniques. Examples include the amount of the data collected, sampling frequency, sensor types, type of activity performed, number of activities performed, and sensor positions. Although, DT with Entropy performed best on most types of HAR data considering its performance metrics across all the datasets, there is no single silver bullet for HAR data. The findings highlight the importance to tailor the selection of ML technique based on the specific HAR requirements and the characteristics of the associated HAR dataset. Future research can analyze the impact of sensor types and positions individually on ML performance. Another potential future research avenue of this study is extending the portfolio of ML techniques to include an investigation of deep learning and (more importantly in the context of wearables, light-weight architectures). Future research could also explore the effect of various pre-processing to further explore the Pareto efficient frontier between run-time performance and classification performance.

7. References

Akhavian, R., and Behzadan, A. H. 2016. "Smartphone-Based Construction Workers' Activity Recognition and Classification," *Automation in Construction* (71), pp. 198–209. (https://doi.org/10.1016/j.autcon.2016.08.015).

Ambati, L. S., and El-Gayar, O. 2020. "A Comparative Study of Machine Learning Approaches for Human Activity Recognition," in *Proceedings of the Fifteenth Midwest Association for Information Systems Conference*, Des Moines, Iowa, May 28, p. 6.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. 2013. "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *ESANN*.

Baldominos, A., Cervantes, A., Saez, Y., and Isasi, P. 2019. "A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition Using Mobile Devices," *Sensors* (19:3), p. 521. (https://doi.org/10.3390/s19030521).

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., and Rojas, I. 2014. "Window Size Impact in Human Activity Recognition," *Sensors (Basel, Switzerland)* (14:4), pp. 6474–6499. (https://doi.org/10.3390/s140406474).

Bokrantz, R., and Fredriksson, A. 2017. "Necessary and Sufficient Conditions for Pareto Efficiency in Robust

Multiobjective Optimization," *European Journal of Operational Research* (262:2), pp. 682–692. (https://doi.org/10.1016/j.ejor.2017.04.012).

Bulling, A., Blanke, U., and Schiele, B. 2014. "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors," *ACM Comput. Surv.* (46:3), 33:1-33:33. (https://doi.org/10.1145/2499621).

Chen, O. T.-, Tsai, C., Manh, H. H., and Lai, W. 2017. "Activity Recognition Using a Panoramic Camera for Homecare," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, , August, pp. 1–6. (https://doi.org/10.1109/AVSS.2017.8078546).

Dohnálek, P., Gajdoš, P., and Peterek, T. 2014. "Human Activity Recognition: Classifier Performance Evaluation on Multiple Datasets," *Journal of Vibroengineering* (16:3), pp. 1523–1534.

Gaikwad, N. B., Tiwari, V., Keskar, A., and Shivaprakash, N. C. 2019. "Efficient FPGA Implementation of Multilayer Perceptron for Real-Time Human Activity Classification," *IEEE Access* (7), pp. 26696–26706. (https://doi.org/10.1109/ACCESS.2019.2900084).

Garcia-Ceja, E., and Brena, R. 2015. "Building Personalized Activity Recognition Models with Scarce Labeled Data Based on Class Similarities," in *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information*, Lecture Notes in Computer Science, J. M. García-Chamizo, G. Fortino, and S. F. Ochoa (eds.), Springer International Publishing, pp. 265–276.

Gjoreski, H., Kozina, S., Luštrek, M., and Gams, M. 2014. "Using Multiple Contexts to Distinguish Standing from Sitting with a Single Accelerometer," in *European Conference on Artificial Intelligence (ECAI)*.

Iqbal, Z., Ilyas, R., Shahzad, W., and Inayat, I. 2018. "A Comparative Study of Machine Learning Techniques Used in Non-Clinical Systems for Continuous Healthcare of Independent Livings," in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, , April, pp. 406–411. (https://doi.org/10.1109/ISCAIE.2018.8405507).

Jain, A., and Kanhangad, V. 2018. "Human Activity Classification in Smartphones Using Accelerometer and Gyroscope Sensors," *IEEE Sensors Journal* (18:3), pp. 1169–1177. (https://doi.org/10.1109/JSEN.2017.2782492).

Jobanputra, C., Bavishi, J., and Doshi, N. 2019. "Human Activity Recognition: A Survey," *Procedia Computer Science* (155), The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology, pp. 698–703. (https://doi.org/10.1016/j.procs.2019.08.100).

Kozina, S., Lustrek, M., and Gams, M. 2011. *Dynamic Signal Segmentation for Activity Recognition*.

Li, F., Shirahama, K., Nisar, M. A., Köping, L., and Grzegorzek, M. 2018. "Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors," *Sensors* (18:2), p. 679. (https://doi.org/10.3390/s18020679).

Mark. 2015. "How to Decide When to Use Naive Bayes for Classification," *Data Science, Analytics and Big Data Discussions*, , October 31. (https://discuss.analyticsvidhya.com/t/how-to-decide-when-to-use-naive-bayes-for-classification/5720, accessed July 13, 2020).

Meyer, J., Schnauber, J., Heuten, W., Wienbergen, H., Hambrecht, R., Appelrath, H., and Boll, S. 2016. "Exploring Longitudinal Use of Activity Trackers," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, , October, pp. 198–206. (https://doi.org/10.1109/ICHI.2016.29).

Micucci, D., Mobilio, M., and Napoletano, P. 2017. "UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones," *Applied Sciences* (7:10), p. 1101.

(https://doi.org/10.3390/app7101101).

Nabian, M. 2017. "A Comparative Study on Machine Learning Classification Models for Activity Recognition," *Journal of Information Technology & Software Engineering* (7:4), pp. 4–8. (https://doi.org/10.4172/2165-7866.1000209).

Nakano, K., and Chakraborty, B. 2017. "Effect of Dynamic Feature for Human Activity Recognition Using Smartphone Sensors," in *2017 IEEE 8th International Conference on Awareness Science and Technology (ICAST)*, , November, pp. 539–543. (https://doi.org/10.1109/ICAwST.2017.8256516).

Ni, Q., Patterson, T., Cleland, I., and Nugent, C. 2016. "Dynamic Detection of Window Starting Positions and Its Implementation within an Activity Recognition Framework," *Journal of Biomedical Informatics* (62), pp. 171–180. (https://doi.org/10.1016/j.jbi.2016.07.005).

Oresti Banos. 12:19:30 UTC. *Evaluating the Effects of Signal Segmentation on Activity Recognition*, Science. (https://www.slideshare.net/orestibl/banos-iwbbio-2014pdf).

Qin, Zhen, Zhang, Y., Meng, S., Qin, Zhiguang, and Choo, K.-K. R. 2020. "Imaging and Fusing Time Series for Wearable Sensor-Based Human Activity Recognition," *Information Fusion* (53), pp. 80–87. (https://doi.org/10.1016/j.inffus.2019.06.014).

Ramasamy Ramamurthy, S., and Roy, N. 2018. "Recent Trends in Machine Learning for Human Activity Recognition—A Survey," *WIREs Data Mining and Knowledge Discovery* (8:4), John Wiley & Sons, Ltd, p. e1254. (https://doi.org/10.1002/widm.1254).

Ronao, C. A., and Cho, S. 2014. "Human Activity Recognition Using Smartphone Sensors with Two-Stage Continuous Hidden Markov Models," in *2014 10th International Conference on Natural Computation (ICNC)*, , August, pp. 681–686. (https://doi.org/10.1109/ICNC.2014.6975918).

Ronao, C. A., and Cho, S.-B. 2017. "Recognizing Human Activities from Smartphone Sensors Using Hierarchical Continuous Hidden Markov Models," *International Journal of Distributed Sensor Networks* (13:1), p. 1550147716683687. (https://doi.org/10.1177/1550147716683687).

Saha, S. S., Rahman, S., Rasna, M. J., Zahid, T. B., Islam, A. K. M. M., and Ahad, M. A. R. 2018. "Feature Extraction, Performance Analysis and System Design Using the DU Mobility Dataset," *IEEE Access* (6), pp. 44776–44786. (https://doi.org/10.1109/ACCESS.2018.2865093).

Sánchez, V. G., and Skeie, N.-O. 2018. "Decision Trees for Human Activity Recognition Modelling in Smart House Environments," *SNE Simulation Notes Europe* (28:4), pp. 177–184. (https://doi.org/10.11128/sne.28.tn.10447).

Seto, S., Zhang, W., and Zhou, Y. 2015. "Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition," in *2015 IEEE Symposium Series on Computational Intelligence*, , December, pp. 1399–1406. (https://doi.org/10.1109/SSCI.2015.199).

Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., and Havinga, P. J. M. 2014. "Fusion of Smartphone Motion Sensors for Physical Activity Recognition," *Sensors* (14:6), Multidisciplinary Digital Publishing Institute, pp. 10146–10176. (https://doi.org/10.3390/s140610146).

Sousa, W., Souto, E., Rodrigres, J., Sadarc, P., Jalali, R., and El-Khatib, K. 2017. "A Comparative Analysis of the Impact of Features on Human Activity Recognition with Smartphone Sensors," in *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, WebMedia '17, New York, NY, USA: ACM, pp. 397–404. (https://doi.org/10.1145/3126858.3126859).

Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. 2019. "Deep Learning for Sensor-Based Activity Recognition: A Survey," *Pattern Recognition Letters* (119), pp. 3–11. (https://doi.org/10.1016/j.patrec.2018.02.010).

Author Biographies

Loknath Sai Ambati is a doctoral student pursuing Ph.D degree in Information Systems with specialization in analytics and decision support at Dakota State University. Loknath is currently working as graduate research assistant at Dakota State University, his research interests include Health Information Technology, Population Health, Health Informatics, and Data Analytics. His work was published in IGI Global, IIS, AMCIS and MWAIS. Besides, he is an active referee of several international journals and conferences like Journal of Intelligent and Fuzzy Systems, IACIS, AMCIS, and HICSS.

Omar El-Gayar, Ph.D. is a Professor of Information Systems at Dakota State University. Dr. El-Gayar has an extensive administrative experience at the college and university levels as the Dean for the College of Information Technology, United Arab Emirates University (UAEU) and the Founding Dean of Graduate Studies and Research, Dakota State University. His research interests include: analytics, business intelligence, and decision support with applications in problem domain areas such as healthcare, environmental management, and security planning and management. His inter-disciplinary educational background and training is in information technology, computer science, economics, and operations research. Dr. El-Gayar's industry experience includes working as an analyst, modeler, and programmer. His numerous publications appear in various information technology related fields. Dr. El-Gayar serves as a peer and program evaluator for accrediting agencies such as the Higher Learning Commission and ABET, as a panelist for the National Science Foundation, and as a peer-reviewer for numerous journals and conferences. He is a member of a number of professional organizations such as the Association for Information Systems (AIS) and the Association for Computing Machinery (ACM).

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 5

Date: 01-31-2021

Web Scraping in the R Language: A Tutorial

Vlad Krotov
Murray State University, vkrotov@murraystate.edu

Matthew F. Tennyson
Murray State University, mtennyson@murraystate.edu

Abstract

Information Systems researchers can now more easily access vast amounts of data on the World Wide Web to answer both familiar and new questions with more rigor, precision, and timeliness. The main goal of this tutorial is to explain how Information Systems researchers can automatically "scrape" data from the web using the R programming language. This article provides a conceptual overview of the Web Scraping process. The tutorial discussion is about two R packages useful for Web Scraping: "rvest" and "xml2". Simple examples of web scraping involving these two packages are provided. This tutorial concludes with an example of a complex web scraping task involving retrieving data from Bayt.com - a leading employment website in the Middle East.

Keywords: Web Scraping, R, RStudio, HTML, CSS, XML, rvest, xml2

DOI: 10.17705/3jmwa.000066

1. Introduction

Data available on the World Wide Web is measured in zettabytes (Cisco Systems, 2017). This vast volume of data presents researchers and practitioners with a wealth of opportunities for gaining additional insights about individuals, organizations, or even macro-level socio-technical phenomena in real time. Not surprisingly, Information Systems researchers are increasingly turning to the web for data that can be used to address their research questions (e.g. see Geva et al., 2017; Gunarathne et al., 2018; Triche and Walden 2018; Vaast et al., 2017)

Harnessing the vast data from the web often requires a programmatic approach and a good foundation in various web technologies. Besides vast volume, there are three other common issues associated with accessing and parsing the data available on the web: variety, velocity, and veracity (Goes, 2014; Krotov & Silva, 2018). First, web data comes in a variety of formats that rely on different technological and regulatory standards (Basoglu & White, 2015). Second, this data is characterized by extreme velocity. The data on the web is in a constant state of flux, i.e., it is generated in real time and is continuously updated and modified. Another characteristic of web data is veracity (Goes, 2014). Due to the voluntary and often anonymous nature of the interactions on the web, quality and availability of web data are surrounded with uncertainty. A researcher can never be completely sure if the data he or she needs will be available on the web and whether the data is reliable enough to be used in research (Krotov & Silva, 2018).

Given these issues associated with "big web data", harnessing this data requires a highly customizable, programmatic approach. One functional and easily-customizable platform for retrieving and analyzing web data is R - one of the most widely-used programming languages in Data Science (Lander, 2014). R can be used not only for automating web data collection, but also for analyzing this data using multiple techniques. Currently, there are more than 16,000 R packages for various data analysis techniques - from basic statistics to advanced machine learning (CRAN, 2020). Some packages are useful for web crawling and scraping, pre-processing, and organizing data stored on the web in various formats.

The following sections introduce the reader to the web scraping infrastructure in R. First, a general overview of the web scraping process is provided. This overview provides a high-level understanding of the steps and technologies involved in automatically sourcing data from the web. Second, the tutorial provides an overview of the "rvest" and "xml2" R packages. These packages are useful for writing web crawling applications and retrieving data from the web in various formats. Third, the article provides simple examples of web scraping code written in R together with a longer and more complex example of automating a web scraping task. The final section discusses implications for researchers and practitioners and the usefulness of the web scraping approach.

2. Web Scraping in R: An Overview

In this tutorial, web scraping is broadly defined as using technology tools for automatic retrieval and organization of data from the web for the purpose of further analysis of this data (Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020). Web scraping consists of the following main, intertwined phases: website analysis, website crawling, and data organization (see Figure 1) (Krotov & Silva, 2018). Although listed in order, these phases are often intertwined. A researcher has to go back and forth between those phases until a clean, tidy dataset suitable for further analysis is obtained.

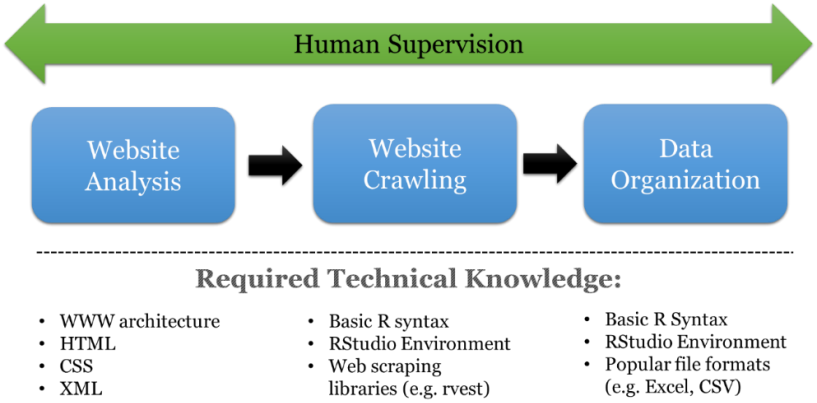


Figure 1. Web Scraping (Adapted from Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020)

The "Website Analysis" phase involves examining the underlying structure of a website or a web repository in order to understand how the needed data is stored at the technical level (Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020). This is often done one web page at a time. This analysis requires a basic understanding of the World Wide Web architecture and some of the most commonly used Web technologies used for storing and transmitting data on the web: HTML, CSS, and XML.

The "Web Crawling" phase involves developing and running a script that automatically browses (or "crawls") the web and retrieves the data needed for a research project (Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020). These crawling applications (or scripts) are often developed using programming languages such as R or Python. We argue that R is especially suitable for this purpose. This has to do with the overall popularity of R in the Data Science community and availability of various packages for automatic crawling (e.g. the "rvest" package in R) and parsing (e.g. the "xml2" package in R) of web data. Furthermore, once data is retrieved using R, it can be subjected to various forms of analysis available in the form of R packages. Thus, R can be used to automate the entire research process – from the time data is acquired to the time visualizations and written reports are produced for a research paper or presentation (the latter can be accomplished with the help of the package called "knitr").

The "Data Organization" phase involves pre-processing and organizing data in a way that enables further analysis (Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020). In order to make further analysis of this data easy, the data needs be to be clean and tidy. Data is in a tidy format when each variable comprises a column, each observation of that variable comprises a row, and the table is supplied with intuitive linguistic labels and the necessary metadata (Wickham, 2014b). A dataset is "clean" when each observation is free from redundancies or impurities (e.g. extra white spaces or mark-up tags) that can potentially stand in the way of analyzing the data and arriving to valid conclusions based on this data. This often requires the knowledge of various popular file formats, such as Excel or CSV. Again, R contains various packages and built-in functions for working with a variety of formats. This is another set of features that make R especially suitable for web scraping.

Most of the time, at least some of the processes within these three phases cannot be fully automated and require at least some degree of human involvement or supervision (Krotov & Tennyson, 2018; Krotov & Silva, 2018; Krotov et al., 2020). For example, "ready-made" web scraping tools often select wrong data elements from a web page. This often has to do with poor instructions supplied by the user of such tools or an ambiguous mark-up used to format data. These tools also often fail to save the necessary data elements in the "tidy data" format (Wickham, 2014b), as data cleaning often requires human interpretation of what this data represents. Moreover, numerous networking errors are possible during web crawling (e.g. an unresponsive web server) that require troubleshooting by a human. Finally, things change so fast on the web! A ready-made tool that is working now may not work in the future due to changes made to a website. This is due to the fact that a ready-made tool may not be fully customizable or modifiable (at least, not from the perspective of a regular user). Thus, changes in the underlying technology behind a website may not be accommodated using the existing features of the tool. All these problems become more acute for large, complicated web scraping tasks, making these "ready-made" tools practically useless. Thus, at least with the current state of technology, web scraping often cannot be fully automated and requires a "human touch" together with a highly customizable approach.

3. R Packages Needed for Web Scraping

Developing custom tools for web scraping requires a general understanding of the web architecture; a good foundation in R programming and RStudio Environment, and at least basic knowledge of some of the most commonly used mark-up languages on the web, such as HTML, CSS, and XML. This tutorial assumed that the readers have the necessary foundation in the aforementioned tools and technologies. If not, then the readers can use the resources listed in Appendix A to get the necessary background knowledge.

Skipping all these foundational knowledge areas, this part of the tutorial focuses on the functionality of two R packages available from the Comprehensive R Archive Network (CRAN): "rvest" and "xml2". Both packages can be downloaded and installed for free from CRAN (see https://cran.r-project.org/). The primary use of the "rvest" package is simulating browser sessions necessary for "crawling" a website. Although the "rvest" package also has some tools for data parsing (e.g. saving HTML code as text), this functionality is often derived from the "xml2" package, which contains a wide variety of tools necessary for parsing data. The functionalities of both packages are explained in detail in the sections below. Short web scraping example that rely on these packages are provided at the end of this section.

3.1 Rvest Package

This section contains an overview of the R package called "rvest". Some tables, examples, and related explanations in this section come from Krotov & Tennyson (2018) and Wickham (2016). There are many features that make "rvest" useful for accessing and parsing data from the web. First, "rvest" contains many functions that can be used for simulating sessions of a web browser. These features come in handy when one needs to browse through many webpages to download

the needed data (this is referred to as the "web crawling" process). Second, "rvest" contains numerous functions for accessing and parsing data from web documents in HTML, XML, CSS, and JSON formats.

Some of the most essential functions of the rvest package are listed in Table 1. Many other functions are available as a part of "rvest" package (Wickham, 2016). One should refer to the official "rvest" documentation to learn more about the package and its usage (see Wickham, 2016).

Table 1. Some Functions of the rvest Package (Reprinted from Wickham, 2016; Krotov & Tennyson, 2018)

Function Usage and Purpose	**Arguments**
Usage: read_html(x, ..., encoding = "") Purpose: This function reads HTML code of a web page from which data is to be retrieved. Can be used for reading XML as well.	• x: A url, a local path, or a string containing HTML code that needs to be read • …: Additional arguments can be passed to a URL using the GET() metod • encoding: specify encoding of the web page being read
Usage: html_nodes(x, css, xpath) Purpose: This function is used to select specific elements of a web document. To select these specific elements one can use CSS elements which contain the needed data or use XPath language to specify the "address" of an element of a web page	• x: A document, a node, or a set of nodes from which data is selected • css, xpath: a name of a CSS element or an XPath 1.0 link can be used to select a node
Usage: html_session(url, ...) Purpose: This function allows to start a web browsing session to browse HTML pages for the purpose of collecting data from them.	• url: address of a web page where browsing starts • …: Any additional httr config commands to use throughout session
Usage: html_table(x, header = NA, trim = TRUE, fill = FALSE, dec = ".") Purpose: This function can be used to read HTML tables into data frames (a commonly used data structure in R). Can be especially useful for reading HTML tables containing financial data.	• x: A node, node set or document • header: if NA, then the first row contains data and not column labels • trim: if TRUE, the function will remove leading and trailing whitespace within each cell • fill: If TRUE, automatically fill rows with fewer than the maximum number of columns with NAs • dec: The character used as decimal mark for numbers

3.2 Xml2 Package

In the past, the rvest package was also used to with XML documents using such functions as xml_node(), xml_attr(), xml_attrs(), xml_text() and xml_tag(). Eventually, these XML functions branched out into "xml2" package designed specifically to work with XML files (including with XML files retrieved from the web). The package has numerous functions for working with XML data. Some of the most commonly used functions of this package together with their commonly used arguments are listed in Table 2 below. Additional details about the functions listed in Table 2 (together with many other functions omitted here) are provided in Wickham et al. 2018.

Table 2. Some Functions of the xml2 Package (Reprinted from Wickham et al., 2018)

Function Usage and Purpose	**Arguments**
Usage: read_xml(x, encoding = "", ..., as_html = FALSE, options = "NOBLANKS") Purpose: This function reads XML and HTML files.	• x: a string, a connection, or a raw vector • encoding: specify a default encoding for the document • ...: additional arguments passed on to method • as_html: optionally parse an XML file as if it's HTML • options: set parsing options for the libxml2 parser
Usage: xml_children(x); xml_contents(x); xml_parents(x); xml_siblings(x) Purpose: These functions are used for navigating through an XML document structure. xml_children returns only elements, xml_contents returns all nodes, xml_parents returns all parents up to the root, xml_siblings returns all nodes at the same level,	• x: a document, node, or node set.
Usage: xml_find_all(x, xpath) Purpose: Finds all XML nodes that match a particular XPath expression	• x: a document, node, or node set. • xpath: a string containing an xpath (1.0) expression
Usage: xml_text(x, trim = FALSE); xml_set_text(x, value) Purpose: Used to extract or replace text in an XML document, node, or node set.	• x: a document, node, or node set • trim: If TRUE will trim leading and trailing spaces • value: character vector with replacement text

4. Simple Web Scraping Examples

The four examples provide below show how to use the "xml2" and "rvest" packages for accessing data in XML, HTML, and CSS format.

The first example comes from Krotov and Tennyson (2018). The example illustrates how "xml2" package is used to extract financial data from an online XML document. The document can be found here:

https://www.sec.gov/Archives/edgar/data/1067983/000119312518238892/brka-20180630.xml

The XML document is a 10-Q statement for Berkshire Hathaway, Inc. posted on EDGAR (an open web database of financial statements of publicly listed companies). Technically, the document is an XBRL (eXtensible Business Reporting Language) instance file. XBRL is an extension of XML language used specifically for interexchange of financial reporting data over the web. Since XBRL is based on XML, one can use the common functions of "xml2" package to work with this data. More specifically, the example below extracts the Net Income data reported by Berkshire Hathaway and saves this data into a data frame named Net_Income_Data.

```
#This example requires xml2 R package
require(xml2)

#Parse XML from an online document found at the URL specified below
URL <- "https://www.sec.gov/Archives/edgar/data/1067983/000119312518238892/brka-20180630.xml"
XML_data <- read_xml(URL)

#Save all nodes related to Net Income or Loss as defined by US GAAP
Nodes <- xml_find_all(XML_data, ".//us-gaap:NetIncomeLoss")

#Convert the node structure into a character vector
Nodes_Vector <- as.character(Nodes)

#Retreive values for all Net Income or Loss elements and save them into a vetcor
#These are dollar values for each Net Income or Loss item reported in the 10Q document
Nodes_Values <- xml_text(xml_find_all(XML_data, ".//us-gaap:NetIncomeLoss"))

#Bind the vectors together as columns and convert the structure into a data frame
#The data frame contains two columns and four rows
Net_Income_Data <- data.frame(cbind(Nodes_Vector,Nodes_Values))
```

The resulting data frame (Net_Income_Data) contains various XBRL attributes of each node containing Net Income or Loss data together with the dollar value of Net Income or Loss (in dollars) reported for each of these items. The definitions of each of the four Net Income or Loss items can be explored further by analyzing the values of attributes for each of the nodes saved in the data frame.

The next example also comes from Krotov and Tennyson (2018). This example illustrates how the "rvest" package can be used for accessing data from an HTML page available on the web:

https://www.sec.gov/Archives/edgar/data/1067983/000119312516760194/d268144d10q.htm

This HTML page also contains a 10-Q statement posted by Berkshire Hathway on EDGAR, albeit from a different period. This time, the goal is to retrieve Balance Sheet data from the statement. This data is available in the form of an HTML table. The HTML table is a part of the HTML web page containing the entire 10-Q statement by Berkshire Hathaway. The table data is retrieved and saved into a data frame called "balance_sheet_data" using the R code below.

```
#This script requires the rvest package to be installed and activated
require(rvest)

#The variable url contains a URL to the 10-Q document published via EDGAR
url <- "https://www.sec.gov/Archives/edgar/data/1067983/000119312516760194/d268144d10q.htm"

#read_html() function from rvest package used to read HTML code from page
tenqreport_html <- read_html(url)

#xpath used to retreive HTML code specifically for balance sheet table
balance_sheet_html <- html_nodes(tenqreport_html,
xpath='/html/body/document/type/sequence/filename/description/text/table[7]')

#HTML code from the balance sheet table is obtained
balance_sheet_list <- html_table(balance_sheet_html)

#Balance sheet data is saved from a list into a data frame
balance_sheet_table <- balance_sheet_list[[1]]
```

Once balance sheet data is saved into a data frame (named "balance_sheet_table" in this example), individual values from the balance sheet can be accessed and used in calculations. For example, certain accounting ratios can be calculated. Alternatively, one can match the balance sheet with other data related to the company and available on the web. The "rvest" package can be used in a similar fashion to obtain quantitative and qualitative data from the same HTML document or other sources on the web.

The code below illustrates how to access top reviews for the iPadPro product listed on Amazon.com. This time, a CSS selector is used to retrieve top reviews for this particular product listed on Amazon. The SelectorGadget tool was used to find an XPath for the CSS element containing the top reviews. The resulting variable review_text is a character vector containing 6 reviews. Note that only a portion of one review was displayed to save space.

```
#Require the rvest package
require(rvest)

#Read HTML code for Apple iPad Pro
ipad_page <-
read_html("https://www.amazon.com/gp/product/B01CGXU0GM/ref=s9_dcacsd_dcoop_bw_c_x_17_w")

#Access top reviews for the product and save them in review_text
review <- html_nodes(ipad_page, xpath='//*[contains(concat( " ", @class, " " ), concat( " ", "a-expander-
partial-collapse-content", " " ))]')
review_text <- html_text(review)

#Display top customer reviews
review_text

## [6] "To fully understand my review, must explain a few things. I come to owning the iPad Pro 9.7\" via
a long lineage of Macs, iPods, iPhones and iPads.
```

The final example is supplied with the "rvest" package (Wickham, 2016). It involves retrieving rating and cast data for "The Lego Movie" found on IMDb website from corresponding HTML or CSS elements. Again, the SelectorGadget can be used to find XPath link associated with each of these elements – this should save time and eliminate the need to be knowledgeable in the particularities of XPath syntax. The "%>%" string in the example below is the so called "pipe" operator used to pass results from one function to another function. Thus, the operator simplified the code by creating a "pipeline" that performs work on data using various functions. The contents of the "rating" and "cast" variables are displayed.

```
library(rvest)
lego_movie <- read_html("http://www.imdb.com/title/tt1490017/")

rating <- lego_movie %>%
  html_nodes("strong span") %>%
  html_text() %>%
  as.numeric()
rating
#> [1] 7.8

cast <- lego_movie %>%
  html_nodes("#titleCast .itemprop span") %>%
  html_text()
cast
#>  [1] "Will Arnett"     "Elizabeth Banks" "Craig Berry"
#>  [4] "Alison Brie"     "David Burrows"   "Anthony Daniels"
#>  [7] "Charlie Day"     "Amanda Farinos"  "Keith Ferguson"
#> [10] "Will Ferrell"    "Will Forte"      "Dave Franco"
#> [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"
```

As one can see from the examples provided this section, "rvest" is a robust package that can be fine-tuned to automatically scrape data for virtually any Information Systems research project requiring web data. The next section contains a more elaborate example of a web scraping project relying on the "rvest" package.

5. A Complex Web Scraping Example

This section contains an example of a web scraping project involving Bayt.com, a leading employment website in the Middle East. At any point in time, the website contains thousands of employment ads from various industries and for

a diverse set of roles. The data collected from the website can be used to answer a number of interesting questions about IT competencies together with competencies in other industries and roles. The example is structured in accordance with the three phases of web scraping discussed previously: Website Analysis, Website Crawling, and Data Organization.

Apart from being larger and more complex, there are three features of this more elaborate example that make it different from the examples provided earlier. First, the dataset retrieved is fairly large. Second, retrieving the dataset involves "crawling" the website one page at a time. This is in contrast to the previous examples where data is accessed from a single web page. Third, the data is saved into a file on the computer after the web scraping task is completed.

5.1 Website Analysis

The web scraping project aimed at retrieving data from Bayt.com starts with the examination of the underlying structure of the website. The structure of the Bayt.com site is fairly typical for a job posting site. One way to browse job postings is by sector (http://www.bayt.com/en/international/jobs/sectors/). When viewing the "Jobs by Sector" page, a list of sectors (such as "Technology and Telecom") is displayed, each as a hyperlink (see Figure 2).

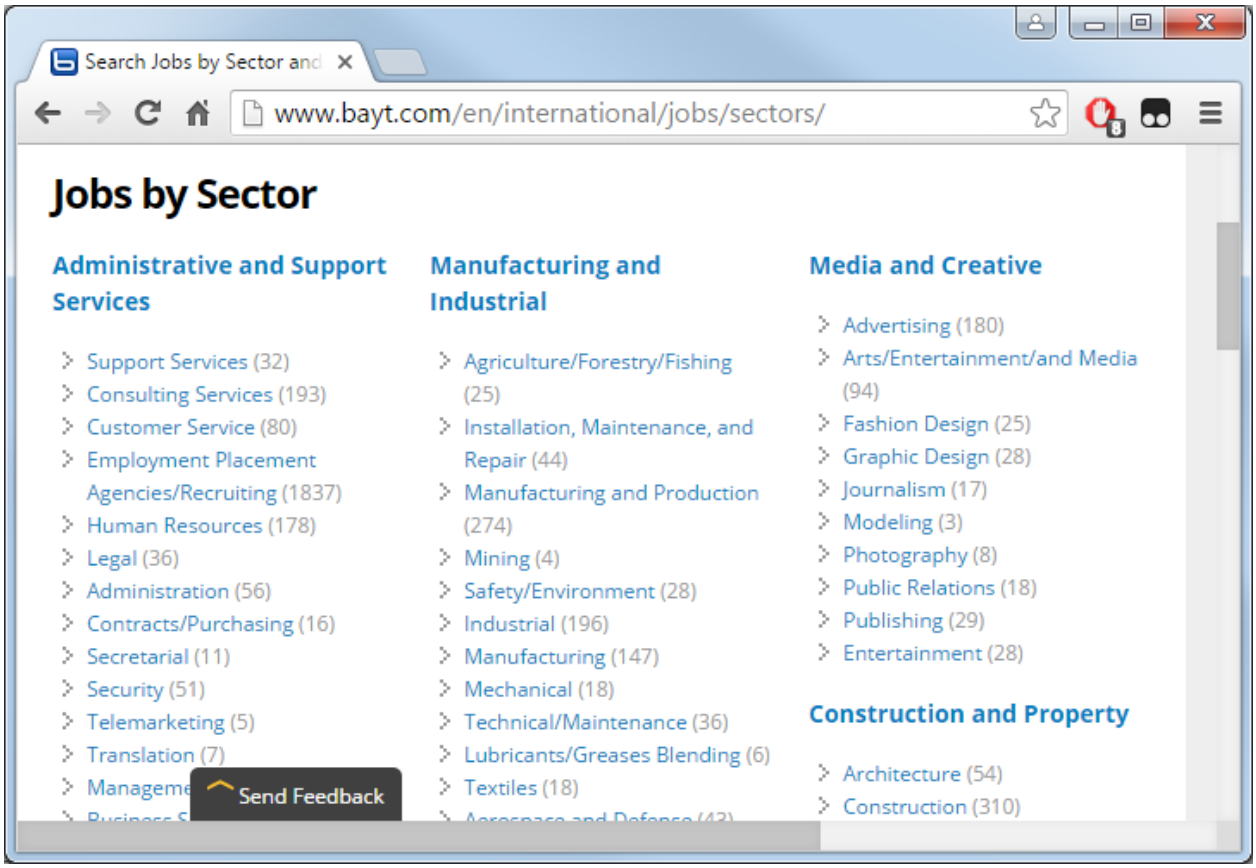


Figure 2. The Bayt.com website when viewing "Jobs by Sector"

Each job sector, therefore, can be accessed using a unique URL. The "Technology and Telecom" and "Banking and Finance" sectors, for example, can be accessed using the following URLs:

https://www.bayt.com/en/international/jobs/sectors/technology-telecom/

https://www.bayt.com/en/international/jobs/sectors/banking-finance/

Note that each URL shares the same base URL, and each is distinguished only by its respective subfolder. This observation will be useful when scraping the website sector by sector later during the Web Crawling phase.

When a particular sector's webpage is viewed, links to the individual job postings are listed, but only in "pages" of twenty at a time. Each page of jobs can be accessed through its own URL, as follows:

https://www.bayt.com/en/international/jobs/sectors/banking-finance/?page=1

https://www.bayt.com/en/international/jobs/sectors/banking-finance/?page=2

https://www.bayt.com/en/international/jobs/sectors/banking-finance/?page=3

On each of these pages, its twenty unique job postings are represented in HTML as a series of "div" elements. Each "div" element contains a hyperlink with the following structure:

```
<a data-js-aid="jobID" href="...">Job Title</a>
```

So, for each job posting, a hyperlink exists that contains all of the information we are scraping. Notice that the element is identified with a distinctive attribute (data-js-aid) that has "jobID" as the value. Therefore, we will need to collect all of the <a> elements that have that distinctive attribute/value pair. Then we will extract the content of the <a> element to get the title, and we will extract the value of the "href" attribute to get the URL.

Also, on each of these pages, there is a <link> element containing the URL to the next page of jobs. The <link> element has the following structure:

```
<link rel="next" href="..." />
```

This will be the element used to determine the URL of the next page of job listings. We will look for a <link> element with a "rel" attribute whose value is "next". Once that element is found, we will use the value of the "href" attribute to determine the URL of the next page of jobs. We will know that we've reached the end of the job listings for the current sector if such an element doesn't exist, at which point we will move on to the next sector of jobs and continue from there.

In this section, we have evaluated the structure of the website, identified how the website can be methodically and programmatically accessed, identified which HTML elements contain our desired information, and inspected the HTML to determine how those elements can be accessed. Therefore, the "Website Analysis" phase is complete.

5.2 Website Crawling

Once the elements that contain the target data have been identified and the respective HTML has been inspected (during the "Website Analysis" phase as discussed in the previous subsection), the data must actually be extracted. In this subsection, we will walk through the script shown in Appendix B step-by-step to demonstrate how the R language can be used to scrape data from the Bayt.com website.

In overview, the script will operate as follows:

1. The necessary library packages will be imported.
2. The working directory will be set.
3. Important input and output variables will be initialized
4. A loop will be set up to iterate through each job sector. At each iteration, one sector will be processed by visiting that sector's webpage.
5. Another loop will be set up to iterate through each page of jobs for that sector. At each iteration, one page of jobs will be processed. A list of the jobs accessible via that page will be collected.
6. Yet another loop will be set up to iterate through that list of jobs. At each iteration, one job posting will be processed. Each job's title and URL title will be extracted. The information will be saved.
7. Finally, once all of the jobs, pages, and sectors have been processed, a file containing all of the collected data will be written.

The remainder of this section will examine and discuss the corresponding lines of the script in detail.

Lines 1-4:

```
#Required packages
require(tm)
require(rvest)
require(XML)
```

These lines are used to load packages into the current session, so that their functions can be called. As it was explained earlier, the purpose of the rvest package is "to make it easy to download, then manipulate, both html and xml" documents (Wickham, 2016). The XML package contains functions for parsing XML documents, and is required by the rvest package. Various functions that are contained in these packages are used throughout the scripts, and will be discussed within the context of those portions of the script in which they are used.

Lines 6-19:

```
#Set working directory
setwd("C:/Users/.../Research/WebScraping")

#Identify job sectors from which to scrape data
sectors <- c("banking-finance",
             "technology-telecom",
             "media-creative")

#Create frame in which data results will be stored
job_data <- data.frame(Sector=character(0),
                       Title=character(0),
                       URL=character(0))

colnames(job_data) <- c("Sector", "Title", "URL")
```

These lines are used to set up the working environment for the script. The working directory specifies the local directory from which any input files will be read and to which any output files will be written. The directory should be set as desired. The "sectors" vector identifies which job sectors will be queried for job postings. Recall from the "Website Analysis" section that each sector has its own URL such that each URL shares the same base address but is distinguished only by its respective subfolder. The values in this vector represent those subfolders, and so they must match the URL subfolder exactly. The "job_data" frame will be used to store all of the extracted information. Each row of the frame will represent exactly one job posting, while the columns of the frame correspond to the bits of information being collected (i.e., "Sector, "Title", and "URL").

Lines 21-25:

```
#For each job sector (each job sector will be processed one at a time)
for (sector in 1:length(sectors))
{
  base_url <- "https://www.bayt.com/en/uae/jobs/sectors/"
  page_url <- paste(base_url, sectors[sector], sep="")
```

These lines are used to loop through each job sector, one at a time. The first line sets up the loop to iterate exactly as many times as there are sectors. The "base_url" variable represents the base URL for all of the job sector websites. That base URL is then concatenated with the name of the current job sector to create the "page_url" variable.

Lines 27-32:

```
  #For each page of jobs within the sector
  repeat
  {
    #Read website for the current sector page
    page_html <- read_html(html_session(page_url))
    cat(paste(page_url, "\n")) #display script progress
```

These lines are used to loop through each page of jobs within the current sector. The "html_session" function (from the rvest package) is used to send an HTTP request for the specified webpage and simulate the functionality of a browser. The "read_html" function is used to get the raw HTML code for the webpage associated with that session. The "cat" function is used to simply display output to the console as the script is executing to indicate to the user which page is currently being processed.

Lines 34-36:

```
    #Retreive the list of job links
    a_elements <- html_nodes(page_html, "a[data-js-aid=\"jobID\"]")
    cat(paste(length(a_elements), " jobs found\n", sep="")) #display progress
```

These lines are used to collect all of the <a> elements that contain the desired information (job title and URL). Recall

from the "Website Analysis" section that each page contains a list of job postings. For each job posting in the list, an <a> element exists, which contains a hyperlink to the respective job posting. Each of these <a> elements is identified with a distinctive "data-js-aid" attribute having "jobID" as the value. The "html_nodes" function is used to retrieve a list of these relevant <a> elements.

Lines 38-44:

```
#For each job within the current page (process one job at a time)
for (i in 1:length(a_elements))
{
  job_href <- xml_attr(a_elements[i], "href")
  job_url <- paste("https://www.bayt.com", job_href, sep="")
  job_title <- html_text(a_elements[i], trim=TRUE)
  cat(paste(" ", i, ". ", job_title, "\n", sep="")) #display progress
```

For each of the <a> elements that were collected in the previous step, we must extract the job title and the URL for the respective job posting. The value of the "href" attribute tells us the URL of the job posting, so we use the "xml_attr" function to extract it. The "href" attribute contains only a relative path, so the Bayt.com domain is prepended to create a complete URL and stored in the "job_url" variable. The content of the <a> element contains the job title, so the "html_text" function is used to extract it, which is stored inside the "job_title" variable. The "cat" function is used to display the progress of the script as it executes by printing the title of the job that was just extracted.

Lines 46-53:

```
  #Consolidate all the information about this job into a single dataframe
  job_info <- data.frame("Sector"=sectors[sector],
                         "Title"=job_title,
                         "URL"=job_url)

  #Write the current job info to the frame where all data is stored
  job_data <- rbind(job_data, job_info)
}
```

At this point in the script, all of the data for the current job posting will have been retrieved. These lines from the script are used to simply write that data to a single-row data frame, and then append that row to the "job_data" frame, where the aggregation of all the data from all the job postings will be stored. The closing brace "}" is used to end the loop that is being used to iterate through each job.

Lines 55-59:

```
#Get the URL for the next page
next_html <- html_node(page_html, "link[rel=\"next\"]")

if(length(next_html)==0)
  break
```

Recall from the "Website Analysis" section that there is a <link> element that contains the URL for the page of jobs. It is distinguished by having a "rel" attribute with the value "next", so the "html_node" function is used to extract that element from the HTML, which is stored in the "next_html" variable. If the element is not found, then the variable will have a length of zero, in which case we will break out of the loop. This will end the execution of the loop that is iterating through each page within the sector, so we will then move on to the next sector.

Lines 61-62:

```
next_xml <- xmlParse(next_html, asText=TRUE)
page_url <- xmlAttrs(xmlRoot(next_xml))["href"]
```

These lines of code will only be reached if we did not break out of the loop during the execution of the previous lines of code, which means that there are still more pages of jobs to process. So, we use the "xmlParse" function to read the HTML as XML, and use the "xmlAttrs" function to extract the value of the "href" attribute, which will contain the URL of the next page of job listings. We update the "page_url" variable with that new URL, and we are then ready for the next iteration of the loop, so that we can process the next page of job listings.

5.3 Data Organization

In the "Data Organization" phase of the web scraping process, the data retrieved during the previous phase is organized into a useful format, such as a spreadsheet. In this case, our data is conveniently collected into a data frame (called "job_data" in the script) that has the exact same structure as a columnar table. That data is written to a comma-separated CSV file in the very last line of the script, which is shown below:

```
#Finally, write the collected data to an output file
write.table(job_data, file="output_data.csv", sep=",", col.names=TRUE, row.names=FALSE)
```

The CSV file can be opened in a spreadsheet program such as Excel. In this example, no further manipulation of the data is necessary. Our data is clean and tidy. However, in general, if additional manipulation of the data is necessary, it would be performed at this stage. Conditional formatting could be added to the spreadsheet to highlight minimum or maximum values; pivot tables could be added to see different views of the data; and so on.

6. Implications for Researchers and Practitioners

We believe that web scraping using R offers a number of advantages to researchers and practitioners. First, when dealing with Big Data, even simple manipulations with quantitative data or text can be quite tedious and prone to errors if done manually. The R environment can be used to automate a number of simple and also complex data collection and transformation processes and techniques based on complex data transformation heuristics (Krotov & Tennyson, 2018). Second, using R for web scraping and subsequent analysis ensures reproducibility of research (Peng, 2011) – something that is central to the scientific method. Any manual manipulation of data may involve subjective choices and interpretations in relation to what data is retrieved and how it is formatted, pre-processed, and saved. Oftentimes, even simple research involving basic data analysis cannot be replicated by other researchers due to the errors made at various stages of data gathering and analysis (e.g. see The Economist, 2016). With R, all aspects of data retrieval and manipulation can be unambiguously described and then reproduced by other researchers by running the script used for data collection. Finally, once web data is retrieved using R, it can be subjected to virtually all known forms of analysis implemented via thousands user-generated R packages available from CRAN. The data can also be used for creating various data products (e.g. via the Shiny tool available with RStudio) – something that can be of relevance for industry researchers.

7. Conclusion

The World Wide Web is a vast repository of data. Many research questions can be addressed by retrieving and analyzing web data. Unfortunately, web data is often unstructured or semi-structured, based on somewhat loose standards, and generated or updated in real time. Retrieving such data for further analysis requires a programmatic approach. Developing web scraping scripts that automate data collection from the web, regardless of which programming language is used for that, requires a good understanding of web architecture and some of the key web technologies, such as HTML, CSS, and XML. As demonstrated in this tutorial, the R environment is an effective platform for creating and modifying automated tools for retrieving a wide variety of data from the web. We believe that the approach to web scraping outlined in this tutorial is general enough to serve as a good starting point for any academic or industry-based research project involving web data. Also, given the brief yet detailed coverage of all fundamental technologies underpinning web scraping in R, this tutorial can also be used by instructors in various Information Systems and Analytics courses to introduce students to web scraping in R.

8. References

Basoglu, K. A., & White, Jr., C. E. (2015). Inline XBRL versus XBRL for SEC reporting. *Journal of Emerging Technologies in Accounting*, 12(1), 189-199.

Cisco Systems. (2017). Cisco Visual Networking Index: Forecast and Methodology, 2016–2021. Retrieved from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

Comprehensive R Archive Network (CRAN). (2020). Retrieved from: https://cran.r-project.org/

Dynamic Web Solutions. (2017). A Conceptual Explanation of the World Wide Web. Retrieved from: http://www.dynamicwebs.com.au/tutorials/explain-web.htm

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using Forum and Search Data for Sales Prediction of High-Involvement Products. *MIS Quarterly,*41(3), 65-82.

Goes, P. (2014). Editor's comments: Big data and IS research, *MIS Quarterly,* 38(3), 3-8.

Gunarathne, P., Rui, H., & Seidmann, A. (2018). When social media delivers customer service: Differential customer treatment in the airline industry. *MIS Quarterly*, 42(2), 489-520.

Krotov, V., and Tennyson, M. F. (2018). Scraping financial data from the web using R language. *Journal of Emerging Technologies in Accounting*, 15(1), 169-181.

Krotov, V. and Silva, L. (2018). Legality and ethics of web scraping. The Twenty Fourth Americas Conference on Information Systems.

Krotov, V. , Johnson, L. and Silva, L. (2020). Legality and ethics of web scraping. *Communications of the AIS*, 47(1), 22.

Lander, J. P. (2014). R for Everyone: Advanced Analytics and Graphics. Boston, MA: Addison-Wesley.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226-1227.

The Economist (2016). "Excel errors and science papers". Retrieved from https://www.economist.com/graphic-detail/2016/09/07/excel-errors-and-science-papers

Triche, J., & Walden, E. (2018). The Use of Impression Management Strategies to Manage Stock Market Reactions to IT Failures. *Journal of the Association for Information Systems*, 19(4), 333-357.

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social Media Affordance for Connective Action: An Examination of the Microblogging Use During the Gulf of Mexico Oil Spill. *MIS Quarterly*, 41(4), 1179-1206.

Wickham, H. (2014a). Advanced R. Boca Raton, FL: CRC Press.

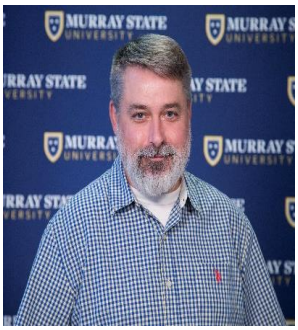
Wickham, H. (2014b). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.

Wickham, H. (2016). Package 'rvest'. Retrieved from: https://cran.r-project.org/web/packages/rvest/rvest.pdf

Wickham, H., Hester, J. and Ooms, J. (2018). Package 'xml2'. Retrieved from https://cran.r-project.org/web/packages/xml2/xml2.pdf

Author Biographies

Dr. Vlad Krotov is an Associate Professor of Management Information Systems at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. Dr. Vlad Krotov received his PhD in Management Information Systems from the Department of Decision and Information Sciences, University of Houston (USA). His teaching, research and consulting work is devoted to helping managers and organizations to use Information and Communication Technologies for analyzing organizational data in a way that enhances organizational performance. His quantitative and qualitative research has appeared in a number of academic and practitioner-oriented journals and conferences, such as: CIO Magazine, Journal of Theoretical and Applied E-Commerce, Communications of the Association of Information Systems, Business Horizons, Blackwell Encyclopedia of Management, America's Conference on Information Systems (AMCIS), Hawaii International Conference on System Sciences (HICSS), International Conference on Mobile Business (ICMB). His research was recognized by the 2016 research was recognized by the 2016 Outstanding Researcher award and 2017 Emerging Scholar award at Murray State University.



Dr. Matthew Tennyson is an Associate Professor of Computer Science at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. Matthew earned his B.S. in Computer Engineering from Rose-Hulman in 1999. After graduating, he worked at Caterpillar, developing embedded systems for various types of earth-moving machinery. In 2004 he earned his M.S. in Computer Science from Bradley University. A few years later, he started pursuing a Ph.D. at Nova Southeastern University, graduating in 2013. Matthew's teaching and research interests include software engineering, programming practice and theory, and computer science education.

Appendix A: Additional Resources

Web Architecture

The following online tutorials are recommended for those who want to learn more about web architecture:

- Dynamic Web Solutions. 2017. A Conceptual Explanation of the World Wide Web. Available at: http://www.dynamicwebs.com.au/tutorials/explain-web.htm
- Tutorials Point. 2017. HTTP Tutorial. Available at: http://www.tutorialspoint.com/http/

The World Wide Web Consortium (W3C) is the ultimate source on the technologies and specifications related to the web:

- World Wide Web Consortium (W3C). 2017. Available at: https://www.w3.org/

HTML

A free online tutorial from w3schools.com on HTML:

- http://www.w3schools.com/html/default.asp

CSS

A free online tutorial from w3schools.com on CSS:

- http://www.w3schools.com/css/default.asp

XML

A free online tutorial from w3schools.com on XML and a number of related technologies:

- http://www.w3schools.com/xml/xml_exam.asp

In addition to providing a rather thorough treatment of XML, the tutorial also has sections devoted to related technologies, such as XML Namespaces, XML Schema, XPath and XLink

R and RStudio

We recommend the following book to people with basic understanding of computer programming but no previous knowledge of R:

- Lander, J. P. 2014. R for Everyone: Advanced Analytics and Graphics. Boston, MA: Addison-Wesley.

The book contains an excellent introduction into various aspects of R language and contains a manual on installing and using RStudio. Much of the examples found in this note are based on this book.

For those already familiar with R, the following advanced text on R can be recommended:

- Wickham, H. 2014. Advanced R. Boca Raton, FL: CRC Press.

Alternatively, one can access various articles and tutorials on R online:

- https://www.r-bloggers.com/how-to-learn-r-2/

For those wishing to get a practical introduction to data science in R and learn various related technologies and statistical techniques, the following online specialization from John Hopkins University is available in Coursera:

- https://www.coursera.org/specializations/jhu-data-science

Appendix B: Example Web Scraping Script

```
#Required packages
require(tm)
require(rvest)
require(XML)

#Set working directory
setwd("C:/Users/mtennyson/Documents/Research/WebScraping")

#Identify job sectors from which to scrape data
sectors <- c("banking-finance",
             "technology-telecom",
             "media-creative")

#Create frame in which data results will be stored
job_data <- data.frame(Sector=character(0),
                       Title=character(0),
                       URL=character(0))

colnames(job_data) <- c("Sector", "Title", "URL")

#For each job sector (each job sector will be processed one at a time)
for (sector in 1:length(sectors))
{
  base_url <- "https://www.bayt.com/en/uae/jobs/sectors/"
  page_url <- paste(base_url, sectors[sector], sep="")

  #For each page of jobs within the sector
  repeat
  {
    #Read website for the current sector page
    page_html <- read_html(html_session(page_url))
    cat(paste(page_url, "\n")) #display script progress

    #Retreive the list of job links
    a_elements <- html_nodes(page_html, "a[data-js-aid=\"jobID\"]")
    cat(paste(length(a_elements), " jobs found\n", sep="")) #display progress

    #For each job within the current page (process one job at a time)
    for (i in 1:length(a_elements))
    {
      job_href <- xml_attr(a_elements[i], "href")
      job_url <- paste("https://www.bayt.com", job_href, sep="")
      job_title <- html_text(a_elements[i], trim=TRUE)
      cat(paste(" ", i, ". ", job_title, "\n", sep="")) #display progress

      #Consolidate all the information about this job into single frame
      job_info <- data.frame("Sector"=sectors[sector],
                             "Title"=job_title,
                             "URL"=job_url)

      #Write the current job info to the frame where all data is stored
      job_data <- rbind(job_data, job_info)
    }
```

```
    #Get the URL for the next page
    next_html <- html_node(page_html, "link[rel=\"next\"]")

    if(length(next_html)==0)
      break

    next_xml <- xmlParse(next_html, asText=TRUE)
    page_url <- xmlAttrs(xmlRoot(next_xml))["href"]
  }
}

#Finally, write the collected data to an output file
write.table(job_data, file="output_data.csv", sep=",", col.names=TRUE, row.names=FALSE)
```