A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

*** "International Journal of Computers and Their Applications is abstracted and indexed in INSPEC and Scopus."**

Volume 25, No. 1, March 2018 **ISSN 1076-5204**

International Journal of Computers and Their Applications

A publication of the International Society for Computers and Their Applications

ISCA Headquarters···•···P. O. Box 1124, Winona, MN 55987 USA···•···Phone: (507) 458-4517
E-mail: isca@ipass.net • URL: http://www.isca@isca-hq.org.

Editor's Note: March 2018

It is my distinct honor, pleasure and privilege to serve as the Editor-in-Chief of the International Journal of Computers and Their Applications (IJCA). I have a special passion for the International Society for Computers and their Applications.

I would like to begin this volume by giving a review of this past year. In 2017 we had 33 articles submitted to the International Journal of Computers and Their Applications. We currently have 4 that are still under review. As a reminder, the journal will not be accepting articles that are less than 6 pages. The authors of these papers will be encouraged to submit their papers to ISCA conferences.

We are still working towards getting IJCA online. Hopefully we can end up with a nice repository soon.

I look forward to working with everyone in the coming years to maintain and further improve the quality of the journal. I would like to invite you to submit your quality work to the journal for consideration of publication. I also welcome proposals for special issues of the journal. If you have any suggestions to improve the journal, please feel free to contact me.

Frederick C. Harris, Jr.
Computer Science and Engineering
University of Nevada, Reno
Reno, NV 89557, USA
Phone: 775-784-6571
Email: Fred.Harris@cse.unr.edu

This year we have 4 issues planned (March, June, September, and December). We begin with a special issue from the best papers at the ISCA Fall 2017 SEDE Conference, followed by a special issue from CAINE 2017. We have a proposal for the best papers from the ISCA Spring Conference cluster (CATA/BICOB) which will appear in the December issue. The other issue is being filled with submitted papers.

I would also like to announce that I begun a search for a few Associate Editors to add to our team. There are a few areas that we would like to strengthen our board with, such as Image Processing. If you would like to be considered, please contact me via email with a cover letter and a copy of your CV.

Frederick C Harris, Jr.
Editor-in-Chief
Email: Fred.Harris@cse.unr.edu

Guest Editorial:
Special Issue from ISCA Fall--2017 SEDE Conference

This Special Issue of IJCA is a collection of five refereed papers selected from the SEDE 2017: 27th International Conference on Software Engineering on Data Engineering, held during October 2--4, 2017, in San Diego, CA, USA.

Each paper submitted to the conference was reviewed by at least two members of the International Program Committee, as well as by additional reviewers, judging the originality, technical contribution, significance and quality of presentation. After the conferences, five best papers were recommended by the Program Committee members to be considered for publication in this Special Issue of IJCA. The authors were invited to submit a revised version of their papers. After extensive revisions and a second round of review, the five papers were accepted for publication in this issue of the journal.

The papers in this special issue cover a broad range of research interests in the community of computers and their applications. The topics and main contributions of the papers are briefly summarized below.

PETER J. GRAZAITIS of the U.S. Army Research Laboratory, Aberdeen Proving Ground, MD and MARK R. CAMMARERE of Technology Service Corporation Trumbull, CT, USA presented in their paper “Visualization of Dynamic Fluid Characteristics for Hose and Pipeline Systems in Army Operational Environments” the challenges that the Army must deal with when building out the supply chain from theater entry point to the forward operating bases and resupply points. In addition to security and concealment factors, Army warfighters must factor in equipment availability, terrain elevation changes that impact the hydraulic pressures in the line, and limited lift resources to haul the equipment to the field. All of these factors drive the selection of the route that will be used to meet the desired flow rates and demands of deployed forces.

ESTEBAN SEGARRA and BRADFORD TOWLE JR. of Florida Polytechnic University, in Lakeland Florida, USA, proposed and evaluated in their paper “Application of an Augmented Reality Device as a Rangefinder and Odometry Source” that the technology used in augmented devices can be harnessed as a viable sensor for a robotic platform. With this in mind, several test results were presented demonstrating the feasibility of using Augmented Reality (AR) sensors. Experimentation with the HoloLens concluded that the device is capable of sustaining odometry, localization, and acquiring distance measurements while under movement.

CONNOR SCULLY-ALLISON, HANNAH MUÑOZ, VINH LE, SCOTTY STRACHAN, ERIC FRITZINGER, FREDERICK C. HARRIS, JR. and SERGIU DASCALU of University of Nevada Reno, USA, introduced in their paper “Advancing Quality Assurance Through Metadata Management: Design and Development of a Mobile Application for the NRDC” the design, implementation, and impacts of a cross-platform mobile application that facilitates the collection of metadata for in-situ sensor networks and provides tools assisting Quality Assurance processes on remote deployment sites. Specifically, the proposed application allows for the near-real time update and centralized storage of contextual metadata.

MARCO X. BORNSCHLEGL, KEVIN BERWIND, and MATTHIAS L. HEMMJE of the University of Hagen, Hagen, Germany described in their paper “Modeling End User Empowerment in Big Data Analysis and Information Visualization Applications” how handling the complexity of data requires new techniques with regard to data access, visualization, perception, and interaction for supporting innovative and successful information, informed decision making, and business strategies. After deriving and qualitatively evaluating the conceptual IVIS4BigData model, a set of conceptual end user empowering use cases for each IVIS4BigData processing stage were modeled in their paper

TIMOTHY OLADUNNI and SHARAD SHARMA of Bowie State University, Bowie, MD, USA presented in their paper "Deep Learning and Hedonic Pricing" an overview of deep learning and hedonic pricing (which suggests that the price of a differentiated commodity is a function of the implicit prices of its composite attributes). They then developed a predictive algorithm based on this concept and compared the performance of deep learning with a ridge regression and came up with an improvement of prediction by 27%.

As guest editors we would like to express our deepest appreciation to the authors and the program committee members of the conference these papers were selected from.

We hope you will enjoy this special issue of the IJCA and we look forward to seeing you at a future ISCA conference. More information about ISCA society can be found at http://www.isca-hq.org.

Guest Editors:

Frederick C. Harris, Jr, University of Nevada, Reno, USA, SEDE 2017 Conference Chair
Sergiu M. Dascalu, University of Nevada, Reno, USA, SEDE 2017 Program Chair
Sharad Sharma, Bowie State University, USA, SEDE 2017 Program Chair

March 2018

Visualization of Dynamic Fluid Characteristics for Hose and Pipeline Systems in Army Operational Environments

Peter J. Grazaitis*
Aberdeen Proving Ground, MD 21005 USA

Mark R. Cammarere†
Trumbull, CT 06611 USA

Abstract

When the Army deploys in Brigade or larger size units (whether for combat operations or humanitarian relief), the demand for bulk fuel and water is tremendous. Depending on the situation, that demand could be thousands to tens of thousands of gallons per day. Often, deployment will be to austere and environmentally-hostile regions (i.e., Afghanistan). Unlike commercial supply chains, the Army often has to build out the supply chain from theater entry point to the forward operating bases and resupply points. As the primary provider of fuel to US military and coalition partners in a theater of operation, the Army will plan and establish fuel and water distribution pipe and hose line systems to push the large quantities of fuel and water that are needed to meet its own, other service and coalition partner's demands. In planning the routes for these hose and pipeline systems, Army petroleum and water warfighters must deal with several constraining factors to achieve the required push. In addition to security and concealment factors, Army warfighters must factor in equipment availability, terrain elevation changes that impact the hydraulic pressures in the line, and limited lift resources to haul the equipment to the field. All of these factors drive the selection of the route that will be used to meet the desired flow rates and demands of deployed forces.

Key Words: US army, fuel and water distribution, assault hose line system, inland petroleum distribution system, tactical water distribution system, early entry fluid distribution system.

1 Problem

With all the modern computational capabilities currently available (e.g., laptops and tablet computers, etc.), conducting a course of action planning analysis for a hose or pipeline route (referred to as a *trace*), is still done the way it has been since WWII – with paper maps, overlays and pencils. Calculators and spreadsheets are also often used to support some of the complex hydraulic equations – the primary one being the Darcy-Weisbach equation [4] for dynamic head loss (loss of dynamic pressure due to friction between the fluid and the pipe or hose wall) sometimes written as: [1]

$$H_f = (0.031)\cdot f\cdot[(L\cdot Q^2) / d^5] \text{ (feet)}, \quad (1)$$

by assuming that the average fluid velocity (since it is not constant across the pipe interior) can be modeled as a very thin turbulent layer at the wall combined with largely constant laminar flow across the rest of the cross-section. H_f is the pressure loss (feet of head), f is the (dimensionless *Darcy* friction factor), L is the length of pipe or hose (feet), d is the interior diameter (inches) and Q is the flow rate (GPM). When the friction factor (f) is known (see next paragraph), then the head loss can be characterized as a Net Available Head (NAH) right triangle (Figure 1a) [6], where the y-extent is the NAH at the output of a pump station and the x-extent is the length of pipe or hose (across flat terrain as per equation (1)) that reduces this head to the minimum suction of the next station. By sliding this triangle across a profile of the trace elevation (Figure 1b) [6], the pump stations can be located along the trace.

With regard to the use of paper, pencil, calculator and spreadsheet – there are several factors that complicate the use of Darcy-Weisbach as described above. For example, friction factor (f) is a function of the Reynolds number (Re) – which in turn is a function of flow rate (Q), interior diameter (d) and fluid kinematic viscosity (K, centistokes): [1]

$$\text{Re} = (3160\cdot Q) / (d\cdot K), \quad (2)$$

where kinematic viscosity (K) is also a function of temperature (Figure 2a) [7]. Standard (steel or aluminum) pipe curves (Figure 2b) [8] can be used to characterize f as a function of Re. However, friction is also a function of pipe or hose wall roughness (standard curves have an assumed roughness baked in) so characterization of friction head loss using f-values based on (for example) worst- and/or best-case interior wall roughness is also a valuable consideration (friction factor information is generally held quite close by hose line manufacturers). Friction factor (f) as a function of pipe or hose wall roughness can be calculated using the Colebrook equation: [5]

* U.S. Army Research Laboratory. Email: peter.j.grazaitis.civ @mail.mil.

† Technology Service Corporation. Email: mcammarere@tsc.com.

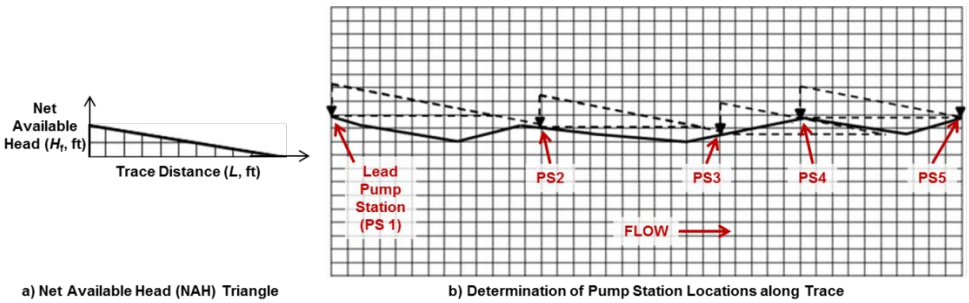

Figure 1: Use of Darcy-Weisbach to determine Trace pump station locations

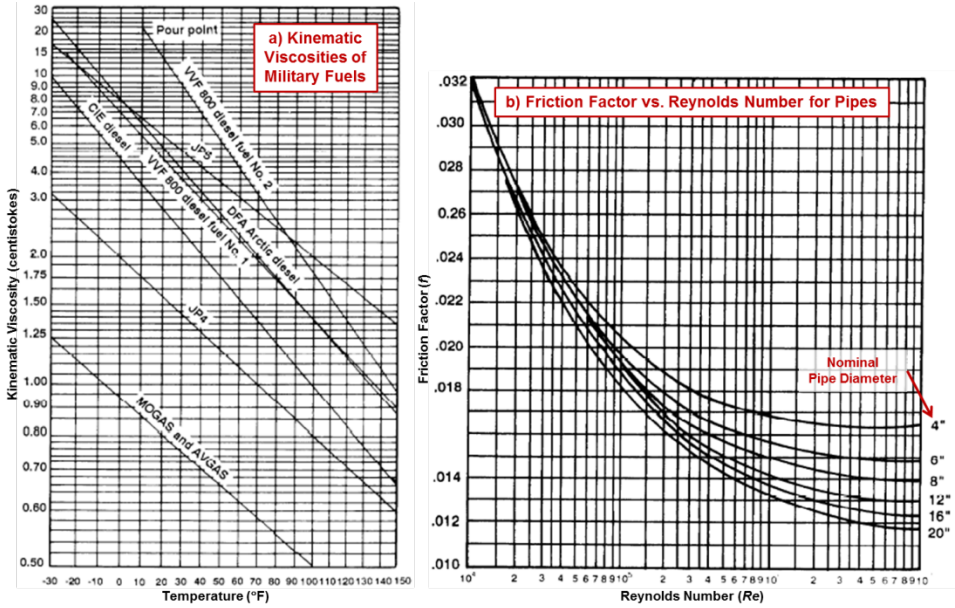

Figure 2: Kinematic viscosity, Reynolds number and friction factor information

$$1 / f^{1/2} = -2 \cdot \log[(\varepsilon / d / 3.7) + (2.51 / (Re \cdot f^{1/2}))], \quad (3)$$

where ε is the RMS interior wall roughness. The form of the equation requires an iterative solution approach. One such approach is the Serghide's [12] solution. Figure 3 plots friction factor (f) over a range of Reynolds Numbers (Re) for a 6" diameter conduit. In the figure, the value of $\varepsilon = 0$ inches for the best- and 0.00707 inches (derived from [13]) for the worst-case. Additional complicating factors that make hand calculation difficult (or are simply not considered) include: pressure corrections for fluid specific gravity (also varies with temperature), and (especially for the IPDS system) the presence and placement of additional components (e.g. valves, elbows, vent assemblies, etc.) whose dynamic friction loss must be obtained using means other than direct application of Darcy-Weisbach.

Also, although traces can certainly be hand drawn on overlays, there is no directly-linked capability to visualize how the dynamic and static pressures change along the proposed pipe or hose line trace due to elevation shifts. Today, the Army can at best develop one or two course of action plans for each proposed route via pencil and paper maps. Furthermore, these plans are only coarse estimates of expected flow rates, equipment needed to implement the hose or pipeline system, etc. due to operational time constraints. Calculation errors are common due to the complex fluid dynamic equations involved (or the inability to consider some of the complex factors). Therefore, deployment and implementation of a pipe or hose line system is often trial and error exercise until an acceptable flow rate is achieved with the equipment available – a manually intensive and potentially error prone process that often results in trial and error system installation at the end of the day. This process is complicated by the fact that the bulk fluid supply lines are typically nonexistent upon theater entry and need to be developed and redeveloped as the battlespace expands over time and the locations of retail points change. the day. This process is complicated by the fact that the bulk fluid supply lines are typically nonexistent upon theater entry and need to be developed and redeveloped as the battlespace expands over time and the locations of retail points change.

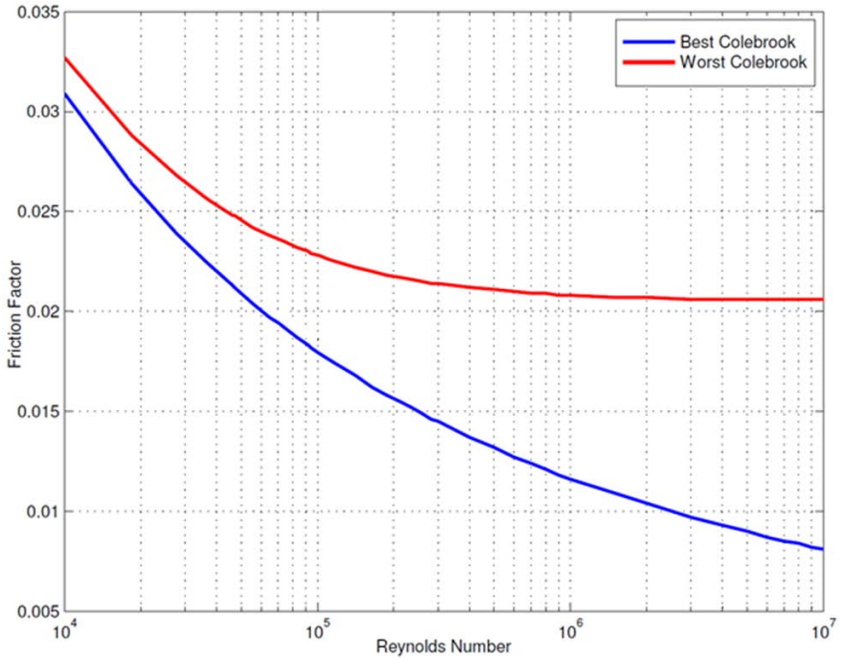

Figure 3: Postulated best- and worst-case friction factor for 6" conduit using Colebrook equation

Finally, due to battlespace dynamics and limited equipment, these hose and pipeline systems may need to be picked up and moved or expanded periodically. So, the planning process is an ongoing, time constrained task for Army petroleum and water warfighters.

Besides the constraints of time, limited equipment and complex calculations; terrain, weather, and threat conditions create additional constraining factors in implementing the best trace. Changes in elevation impact the pressures and flow in a hose or pipeline trace that traverses hilly terrain. These changes may require additional pumps as well as pressure control devices (check valves and pressure reducers) along the trace. Generally, trace routes that are as level as possible are preferred. However, identifying such routes can be difficult, and one rule-of-thumb is to follow other infrastructure that already has some type of terrain improvements (e.g. roads, railways, power lines, commercial pipelines, etc.) [11]. (But this can conflict with other constraints that recommend avoidance of urban and population centers, for example. [9])

Although elevation change (slope) is a major constraint [2], others should also be considered including: road and river crossings [2], areas prone to flooding[10], concealment from the enemy [9], etc. These factors and others should be weighed when considering a trace route and many, depending on operational and construction time constraints, may not be included in trace route planning cycles.

2 Objective

To address these shortcomings, researchers at the U.S. Army Research Laboratory represented the planning exercise as a constrained object type problem. This allows it to be approached as an optimization problem – finding a balance of all constraining factors to achieve a trace route that minimizes the impact(s) of all constraining factors as much as possible. In this case, the objects are the hose or pipe line system components and the optimal route is one that achieves the desired flow rate with the available equipment while addressing other constraining factors such as security/concealment, terrain, transport, and time. Army researchers also wanted to ensure end product ease of use by petroleum and water warfighters, while allowing them to incorporate their subject matter expertise to quickly create traces and (most importantly) to visualize the impact of constraints (particularly elevation changes) on the trace pressure profiles. Army researchers wanted the tool to allow petroleum and water warfighters to quickly edit trace routes, create new ones, and provide hydraulic analyses in interactive, visual ways that go beyond simple presentation of calculated mathematical values. Rich visualization was a primary consideration for the Army researchers' choice of a map-centric tool – from specifying parameters for any particular pipe or hose line system, to creating trace routes on a digital map, to visualizing the dynamic and static pressure changes along the trace as well as what equipment is needed and where it needs to be placed along the trace if the system is implemented.

Army researchers felt this approach would address usability issues as well as allow Army petroleum and water warfighters to quickly determine when the desired flow rates had been achieved with the on-hand equipment while also addressing other requirements such as time, security, and costs. In the end, Army researchers wanted tool use to be as natural as for any Microsoft application. It was also important that the business practices of the software tool be as similar as possible to those that are taught in the Army's petroleum and water

schoolhouse. So in the end, U.S. Army researchers developed a tool that would: 1) automate the pipe or hose line trace planning process, 2) follow the soldiers' business practices to comport with their schoolhouse training, 3) adhere to best commercial software practices for apps found on laptops, tablets and smart phones, 4) incorporate human factor usability principles, 5) visually display constraints on the digital map, 6) convey complex analytics in a visual and intuitive way for rapid situational awareness, and 7) provide performance characteristics that allow Army petroleum and water warfighters to develop many potential courses of action based on their experience and subject matter expertise.

3 Approach

To ensure the software tool achieved these objectives, a spiral development approach was used to get prototypes in the hands of petroleum and water warfighters early and often – to use and evaluate the tool on hypothetical traces, provide user feedback on what they did and didn't like, and to make recommendations for new tool capabilities. This feedback was then used to incorporate changes and add capabilities to help ensure an end product that was easy to use, required minimal training, quickly conveyed to the soldiers how the pipe or hose line system would perform, and allowed the soldier to create and evaluate many different courses of action in a short period of time.

The resulting tool is the Petroleum and Water Trace Locator (PAWTL) [10]. Army researchers wanted PAWTL to support all the Army's bulk petroleum and water hose and pipe line systems in the inventory as well as the Early Entry Fluid Distribution System (E2FDS) currently under development and acquisition. Army researches built physical attribute (e.g., equipment and implementation doctrine) information into the PAWTL tool for the Army's three existing bulk fuel and water distribution systems: the Inland Petroleum Distribution System (IPDS), the Assault Hose Line System (AHS), and the Tactical Water Distribution System (TWDS). A 'custom' system option gives Army petroleum and water warfighters the ability to define customized systems using Commercial-off-the-Shelf (COTS) components to support the proposed E2FDS which is still in development for acquisition. The PAWTL database also includes fluid characteristic information (e.g., specific gravity, kinematic viscosity, etc.) for water and all the bulk fuel products in the Army inventory: Jet Fuel 4 (JP-4), Jet Fuel 5 (JP-5), Jet Fuel A (Jet A), Jet Fuel A-1 (Jet A-1), Jet Fuel B (Jet B), Jet Fuel 8 (JP-8), Jet Fuel 100/130 (100/130), Motor Gasoline (MOGAS), Diesel Fuel 1 (DF-1), Diesel Fuel M (DFM) and Diesel Fuel 2 (DF-2). Figure 4 depicts the PAWTL *Trace Design Parameters* dialog, where users select a particular hose or pipeline system, its operating parameters (ambient temperature, working pressure and flow rate), and the fuel product that will be pushed through the system. PAWTL uses visual selection and drop-down menus to allow Army petroleum and water warfighters to quickly define the system's operating parameters.

PAWTL uses the department of defense's Commercial Joint Mapping Tool Kit (CJMTK) to provide the functionality for terrain visualization and spatial analysis. A terrain import capability gives Army petroleum and water warfighters the ability to plan pipe and hose line traces using digital terrain maps that they are visually familiar with from their basic military training (although a PAWTL technology refresh effort currently underway is employing global data sets that will make the import capability unnecessary). PAWTL also uses an on-board terrain suitability engine [3] that can be used to analyze terrain and visually display user selected constraints such as slope, urban areas, road networks, bodies of water, etc. as generally advantageous or disadvantageous areas with regard to siting the pipe or hose line trace route based on doctrinal or performance considerations. Visually conveying and profiling information in this manner, allows Army petroleum and water warfighters to quickly assess what terrain is or is not suitable. A typical analysis is depicted in Figure 5. Other thematic or analytic layers that could be added to provide a more robust view of constraints of concern (e.g. forested areas, line of sight from known enemy locations, flood zones, etc.) are depicted in Figure 6.

Traces drawn on the PAWTL's digital maps are very similar to manual drawings on acetate overlays (except that they are drawn with a mouse instead of a pencil), but with the added advantage that trace edits can be made very quickly. PAWTL automatically performs the hydraulic analysis for the selected hose or pipe line system based on the defined operational parameters and product to be pushed along the specified route. PAWTL's visual presentation provides Army petroleum and water warfighters with an immediate feel for the trace's viability by showing the trace on the map, pump locations and a summary of the numbers of components required if the trace were deployed (see Figure 7).

Most importantly, PAWTL incorporates the critical hydraulic pressure analytics, digital elevation profile, and doctrinal information along the proposed hose or pipe line route and presents that information in a visual manner that allows petroleum and water warfighters to quickly assess the viability of the proposed trace – including comparison of alternatives. This is shown in the *Trace Pressure Analysis* presentation of Figure 8. This presentation provides a visual and intuitive side-by-side comparison of 2 proposed hose line configurations along the same trace route. It provides visualization of the static and dynamic pressures along the trace that allow Army petroleum and water warfighters to intuitively understand and correlate with elevation changes and the locations of components. This allows those warfighters to employ their subject matter expertise to quickly assess if the proposed route for the selected system is sufficient to support Army requirements for bulk fuel or water.

Finally, PAWTL provides an equipment report which lists all required components and their Military Grid Reference System (MGRS) locations along the trace if the trace were implemented. This information is important for doctrinally required route reconnaissance missions to ensure that the terrain along the route will indeed support the component placement required to deploy the hose or pipeline system. A

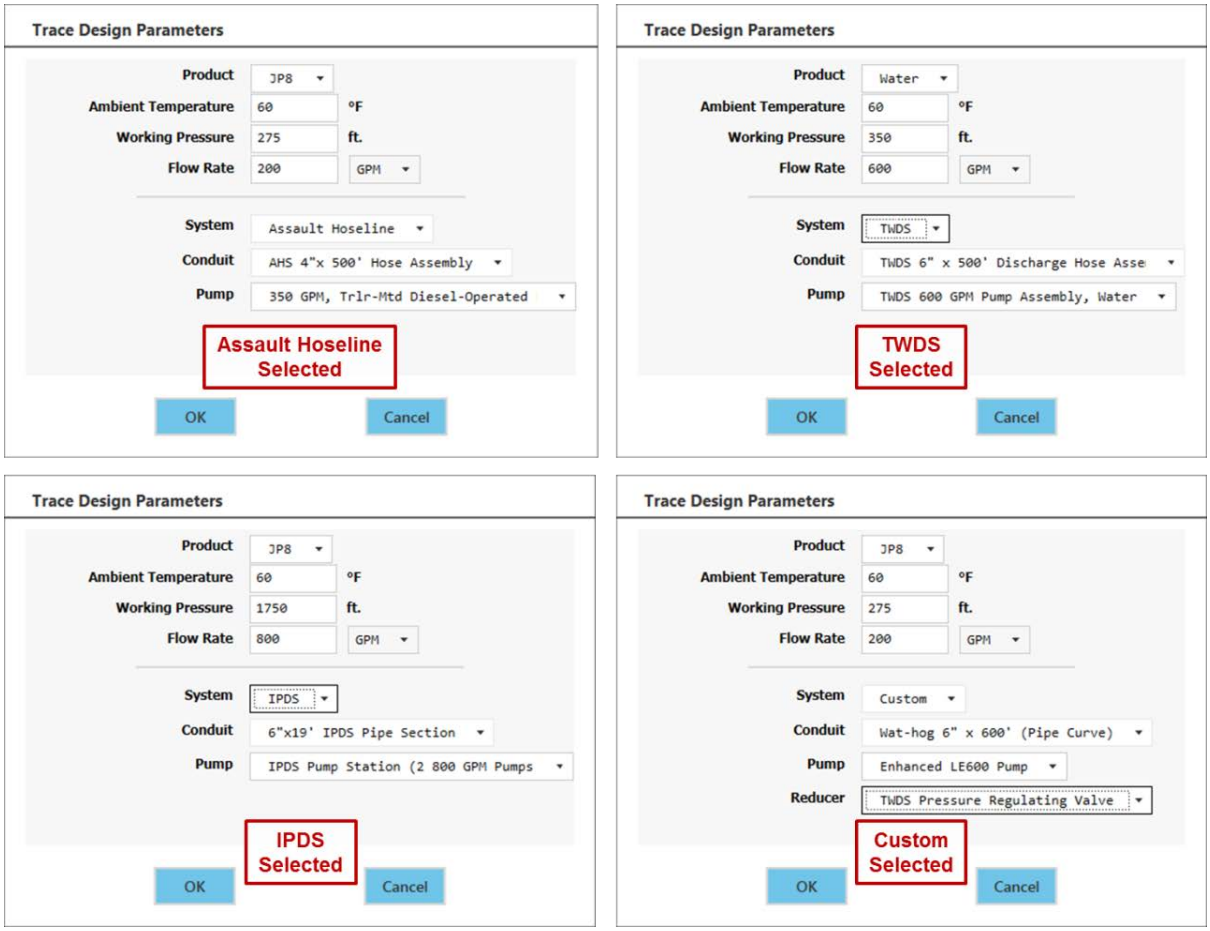

Figure 4: PAWTL trace design parameters for AHS, TWDS, IPDS and 'custom' system selection

sample of this report (which can also be used to support trace implementation) is shown in Figure 9.

4 Conclusion

The PAWTL tool has been validated by comparing manually-generated traces with those created using the tool. Although work in this area is ongoing (not yet officially documented), to date these comparisons have produced very encouraging results (less than 10% variation) when considered by Army subject matter experts. Evaluation of usability during the spiral development process and during two field evaluations resulted in very positive feedback. Training for the Army petroleum and water warfighters that participated in these evaluations consisted of about a one-hour training session followed by an additional hour of hands-on use working a problem for evaluation. The PAWTL tool has been transitioned to Program Manager – Petroleum and Water Systems (PM-PAWS) and will become part of the E2FDS program of record when it is fielded.

ARL researchers are also looking at the potential to add a constraint optimization capability to the PAWTL for evacuation and retrieval of a hose or pipeline system. Evacuation and retrieval of fuel hose line systems is also manually intensive and laborious, since all fuel in the line must be recovered prior to retrieval. A ten-mile hose line system can contain as much as 5.38 million gallons of fuel. Optimizing the route (prior to implementation) to also support pigging operations for hose line evacuation would also provide a great benefit to bulk fuel operations in the area of operations. This would represent a new PAWTL capability that would perform route optimization for not only delivering but also evacuating and retrieving the hose line once it is no longer needed.

The PAWTL tool currently supports the planning and layout of all Army hose and pipeline systems of record. Furthermore, its fuel and water framework either already includes or can be made to support all the constraint and analytic modules required to design a theater-wide petroleum (or fluid) distribution supply chain. The concept would be to expand PAWTL's capabilities to allow Army decision makers to quickly conduct course of action analyses to support contingency planning for bulk fuel and water in regions where the U.S. Army may have to deploy for either combat or humanitarian assistance operations. The framework upon which the PAWTL is built has the ability to layout terminal heads for ship-to-shore or rail deliveries, bag farms, water processing plants, convoy routes, and refueling points. This

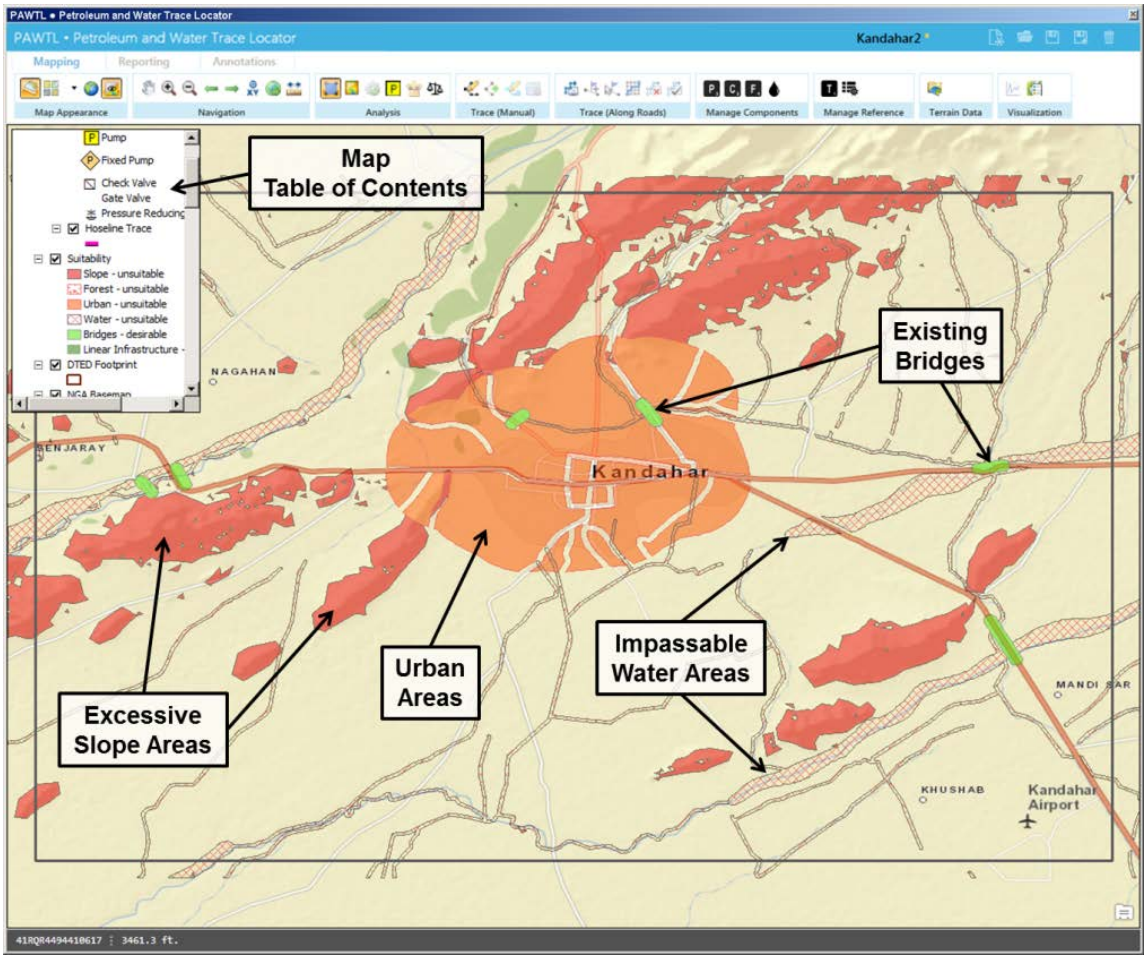

Figure 5: Terrain constraint analysis overlaid on PAWTL map display

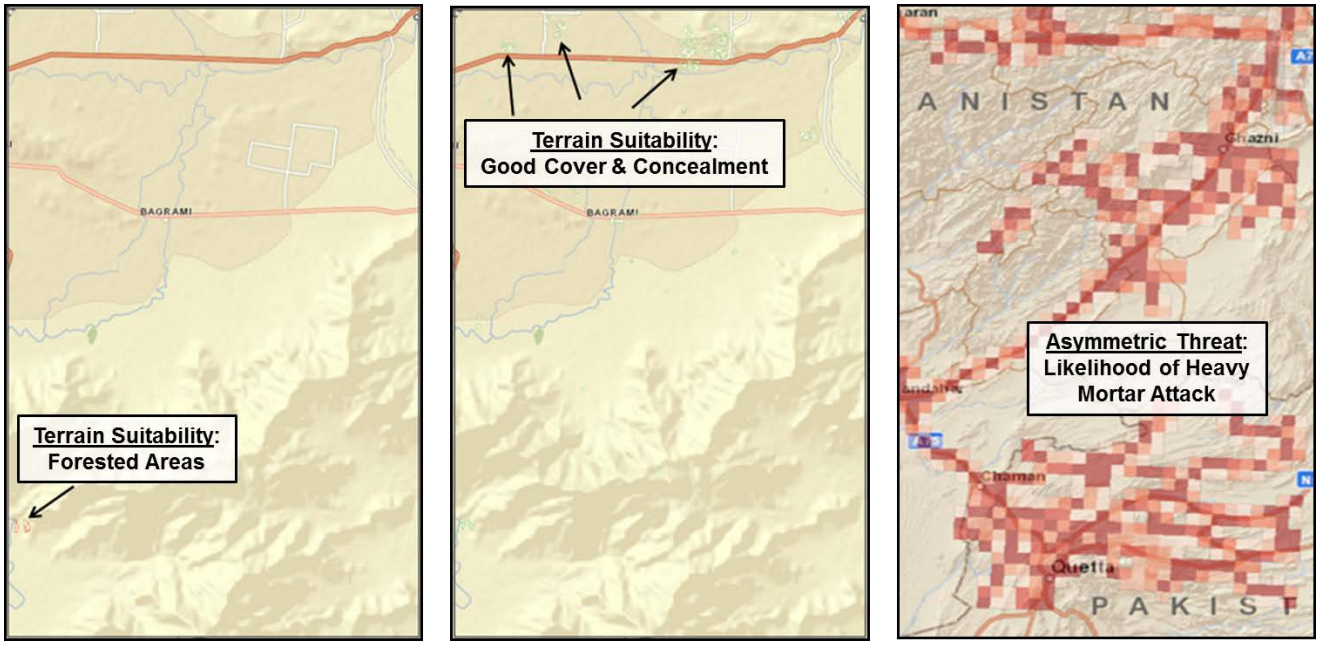

Figure 6: Examples of additional battlespace constraint factors

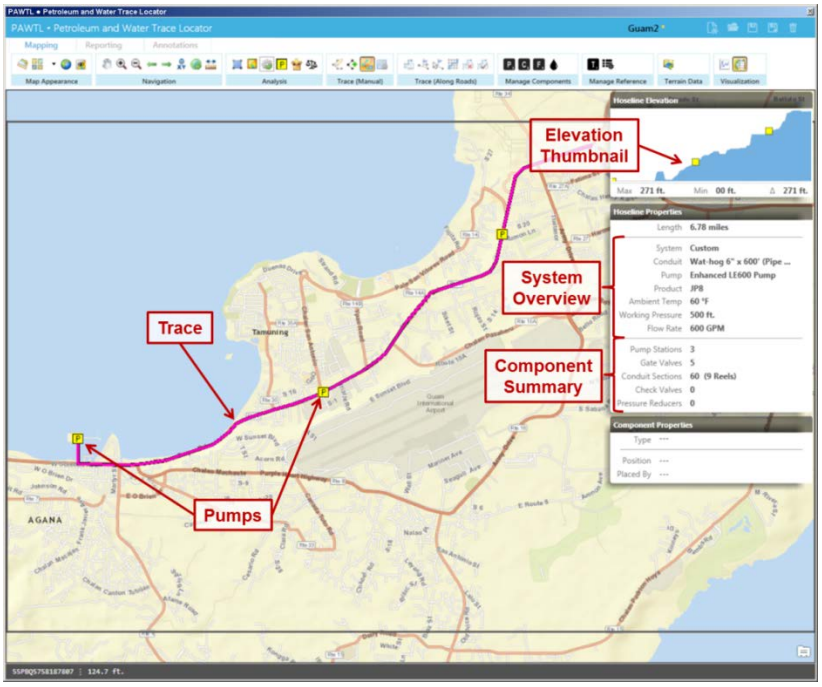

Figure 7: Completed trace analysis showing elevation thumbnail, system overview and component summary

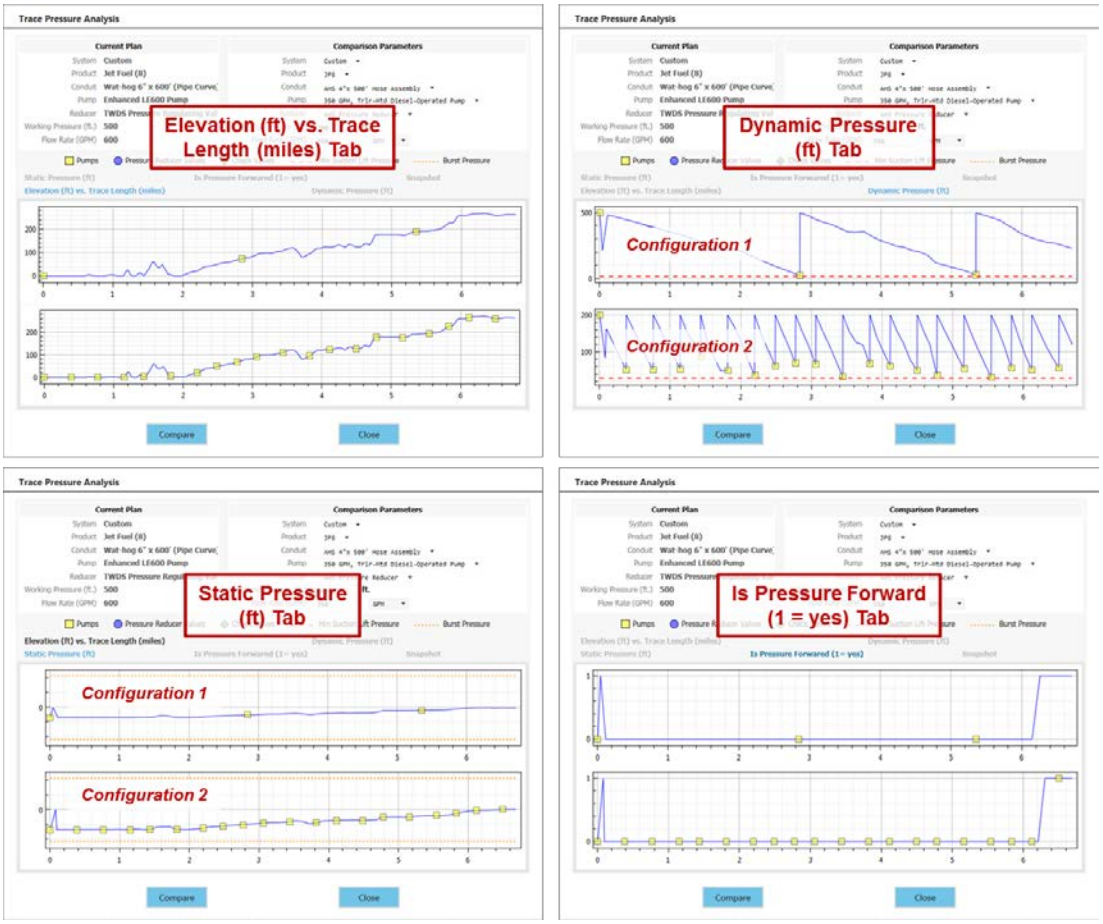

Figure 8: PAWTL trace pressure analysis dialog showing elevation, dynamic and static pressure profiles

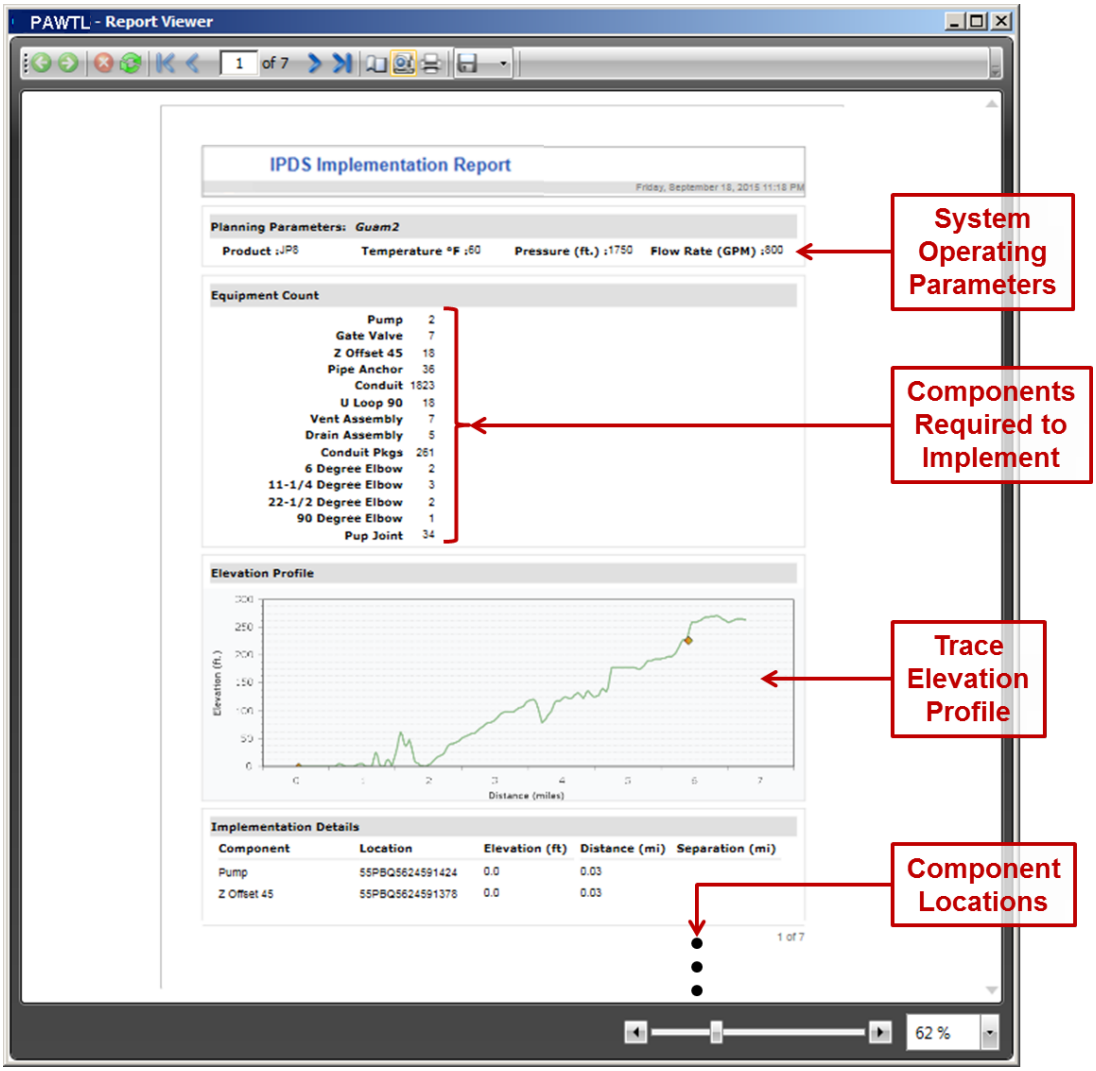

Figure 9: PAWTL equipment report showing operating parameters, component list, elevation profile and individual component locations

would allow a theater-wide bulk fuel and water supply chain to be modeled in an optimal way.

References

[1] Army Field Manual FM 5-482, *Military Petroleum Pipeline Systems*, HQ Dept. of the Army, p. A-2, 26 August 1994.

[2] Army OPROJ TSE-X-03-00-A, IPDS Operations and Planning Guide, TACOM, (Not an official document, intended as a guide or quick reference.), p. 79, 30 July 2010.

[3] Mark R. Cammarere, "A Configurable Suitability Engine for Non-Homogeneous, Multi-Resolution Data," Presentation to the NorthEast Arc User Group, Nashua, NH, 6-7 October 2009.

[4] Yunus A. Cengel and John M. Cimbala, *Fluid Dynamics Fundamentals and Applications*, McGraw-Hill, New York, NY, p. 346, 2010.

[5] Yunus A. Cengel and John M. Cimbala, p. 357, 2010.

[6] Combined Arms Support Command (CASCOM), Proposed Assault Hoseline System (AHS) Work Package document (excerpt), provided by Program Manager-Petroleum and Water Systems (PM-PAWS) via private communication.

[7] FM 5-482, p. C-6, 26, August 1994.

[8] FM 5-482, p. C-7, 26 August 1994.

[9] FM 5-482, pp. 3-4 and 3-5, 26 August 1994.

[10] Peter J. Grazaitis and Mark R. Cammarere, "Petroleum and Water Trace Locator (PAWTL)," *Proc. ESRI International User Conference*, San Diego, CA, 26 July 2012.

[11] IPDS Student Handbook "Guide," FORSCOM Petroleum Training Module, (Not an official document, intended as a guide or quick reference.), p. 2-2, January 2005.

[12] T. K. Serghide's Implementation of Steffensons' Accelerated Convergence Technique is Reported to have Appeared in *Chemical Engineering*, March 5, 1984.

[13] US Army TARDEC Fuels and Lubricants Research Facility Southwest Research Institute® (SwRI®), Interim Report TFLRF No. 401, Testing of Prototype Conduit, January 2010.

Peter J. Grazaitis is an Operations Research Analyst with the U.S. Army Research Laboratory with over forty years of experience working for the Army as a civilian employee. He has bachelor of science degree from the University of Scranton in Physics. He also has a Master of Science degree from The Johns Hopkins University in Computer Science. Mr. Grazaitis has been conducting applied research for the Army in the area of Logistics and Sustainment for over twenty-five years. He received the U.S. Army's commander's award for superior meritorious service as a software technical manager for phase I of the Joint Logistics Advanced Concept Technology Demonstration project in 1997. He received an Army Small Business Innovative Research (SBIR) Program achievement award as the contracting officer representative supporting technology transition to the Army in 2010. He went on to receive the U.S. Small Business Administration's Tibbetts Award in 2015 in recognition of his outstanding contribution to the SBIR program for technology transition. Today, Mr. Grazaitis continues his research into understanding and addressing theater level operational issues with the Army's supply chain.

Mark Cammarere is a System Engineer with over 34 years of experience in radar/sonar signal processing algorithms design, tactical decision aid design, software development, Geographic Information Systems (GIS) and information technologies. He joined Technology Service Corp (TSC) in 1988 and is a principal staff engineer in TSC's Trumbull, CT operation. Mr. Cammarere received B.S.E.E. and M.S.E.E. degrees from Syracuse University in 1983 and 1986, respectively. He is the Principal Investigator (PI) and program manager for TSC's programs related to the PAWTL tool.

Application of an Augmented Reality Device as a Rangefinder and Odometry Source

Esteban Segarra* and Bradford Towle Jr. *
Department of Computer Science, FL Polytechnic University Lakeland, FL, 33805, USA

Abstract

Augmented reality devices can now address several challenging problems dealing with spatial understanding between reality and virtual environments. Robotics must overcome very similar challenges in order to safely navigate in the real world including localization and accurate spatial understanding of obstacles. Currently computationally expensive algorithms are used to continually map and localize the robot. This paper proposes that the technology used in augmented devices can be harnessed as a viable sensor for a robotic platform. With this in mind, several test results are presented here demonstrating the feasibility of using Augmented Reality (AR) sensors. This includes getting the raw distance data, using the built-in odometry, creating a map, and finally deploying an autonomous robotic platform with the AR device. From the results of these tests, a more comprehensive understanding can be found regarding the use of an AR device as a robotic sensor. Experimentation with the HoloLens concludes that the device is capable of sustaining odometry, localization, and acquire distance measurements while under movement. **Key words**: augmented reality; robotics; robot operating system.

1 Introduction

Augmented reality (AR) technology [14] is currently advancing faster than in previous years and can now address several challenging problems dealing with spatial correlation between reality and a virtual environment [2]. The first obstacle is mapping the environment in order to detect objects in reality. This means that an AR device must be able to detect range in the environment with respect to itself. Another issue arises when tracking the movement of the device itself. Tracking the position and rotation change with accelerometers will result in an additive error regardless of the sensor's quality. This necessitates the incorporation of localization, which uses the sensors to correct the position data provided by on board sensors. After an AR device has successfully overcome the above two obstacles, it must deal with integrating virtual elements into reality while respecting depth and occlusion of physical objects.

Like AR devices, robots must be spatially aware and be able to sense obstacles in the environment. To accomplish this robots will often use computationally expensive algorithms to ensure their position is correct and they have an accurate understanding of their environment [7]. The objective of this paper is to use an AR device as a sensor for a robot that will provide reasonably accurate odometry and sensory information [10]. For this experiment the Microsoft™ HoloLens was used as a range finder and odometry source for a Pioneer 3DX. The goals for this objective are: getting reasonable sensor data, using it to create a map with simultaneous localization and mapping (SLAM), and finally navigating safely with data only from the HoloLens itself. This paper has the following structure: a short section on related works and similar experiments, followed by a description of our methodology and results, a brief section on future works, and finally concluding with some closing statements.

2 Related Work

Current techniques being used for navigation or SLAM utilize specialized sensors such as LIDAR, a combination of algorithms and simple cameras, or AMCL. LIDAR systems commonly depend on a laser mounted on a motor that spins at a high speed to give accurate readings of the environment, which could run the risk of mechanical failure and usually have a high start-up cost for robotic implementation. One such example of an industrial-quality LIDAR system include the Hokuyo URG-04LX system, which can provide refresh rates up to 10,000 Hz and a maximum range of 5.6 meters [1].

The use of optical sensors for range-finding and obstacle avoidance is not a recent technology, with some early editions including the easily obtainable Xbox 360 Kinetic for creating an artificial laser scan [9].

These sensors proved to be useful for mapping an area with images and as a component for SLAM localization [6]. However, these sensors lack odometry capabilities and have to be provided through computationally expensive software algorithms or a dedicated odometry sensor [11]. Some, such as the Xbox 360 sensor, cannot operate in all environments effectively.

Through technological advancements, technologies such as the GyroWand provide accurate ray tracking/laser scan capabilities and odometry with the combination of technology such as dead reckoning with inertial measurement units (IMU). The Microsoft™ HoloLens uses a more refined version of the technology used in the GyroWand [5].

The Microsoft HoloLens has the advantage of being able to output odometry through the use of its built-in sensory equipment and use its visual cameras for use in localization of the real world [4]. It was because of the after mentioned features we initially chose the HoloLens for use in robotic navigation, localization, mapping, and crash avoidance [3].

*(Department of Computer Science. Email: esegarra3278,btowle)@floridapoly.edu

3 Methodology

To test our theory and goals for integrating and utilizing the HoloLens as a sensor, we will use a robotic system configured for experiments with the AR system. This platform must accomplish the goals of being able to successfully move the HoloLens around an arena, providing a means of collecting and broadcasting information, and reliably control the system without major interruptions.

Our system consists of three major hardware parts and two software-side components.

4 Hardware Operation

HoloLens: The Microsoft™ HoloLens comes integrated with all sensory equipment. This includes an IMU, four environmental understanding cameras, one depth camera, mixed reality capture, four microphones, and a light ambience sensor [8]. These systems come together to form a spatial understanding in the environment in which it is located. The HoloLens will acquire and transmit the distances based on a spatial mesh created in Unity, to ROS, in the format of a laser scan message.

Computers: The computer used for collecting and broadcasting the information collected by the HoloLens is a 2015 HP EliteBook running on Ubuntu 16.04 LTS. When in operation, the computer was mounted on top of the Pioneer 3 DX. The computer was additionally configured to function as a mobile hotspot location for external computers to communicate with the computer, interact with the robot, and collect information from the HoloLens.

Another computer, the ROS client, was used for interacting with the ROS server computer. The ROS client was an Asus G75VW laptop used to run the move base program, teleoperation, gmapping, and other monitoring services for the ROS service. Data received from the EliteBook was wirelessly received through a hotspot connection.

Robotic Platform: The robotic platform of choice for experimentation is a Pioneer 3 DX robot powered by three 12V lead-acid batteries. The device does not come with an on-board computer and must be controlled through the use of an external computer. The Pioneer 3 DX has two large wheels and a free-moving coaster to achieve holonomic control over the robot. The robot can carry up to 23 kg, well suited for our purposes.

5 Software–side Operation

ROS: Robot Operating System (ROS) is a software-side operating system that creates a framework that can interact with a robot using C++ or Python packages. For the purposes of our experiment, we used ROS version Kinect. ROS is used to control robot motors, record and analyze sensor data, and utilize incoming data to be used for advanced algorithms with the robot [12]. For the purposes of this paper, we created a ROS server which received data from the HoloLens such as odometry transforms and laser scan data and forwarded it to a ROS client. The ROS server was given commands remotely through a ROS client on a separate computer which then ran the processing algorithms (such as gmapping, safe wander, and move base).

Unity and HoloLens Toolkit: Unity 3D is a game engine that is used to project the data provided by the HoloLens Toolkit into virtual space and provides the means for creating a ray cast vector to calculate distances. The HoloLens Toolkit creates a virtual mesh that is a low-polygon approximation of the objects in real-life and contains the distances between the HoloLens and the environment around it. The Unity engine can then create ray cast vectors to find the distances and convert them into ROS-usable laser scans messages.

Through a ray casting method of calculating distances through the Unity engine, a laser scan can be created. The HoloLens can measure up to 10 m effectively and a minimum distance of 20 cm roughly has a 5 cm resolution at these distances. The laser scan has a flat-plane vision range of 180 degrees of seamless vision and the horizontal angle can be adjusted for a lower or higher plane of vision. For the purposes of experimentation with the HoloLens, the angle of vision was adjusted two degrees below the level plane.

Odometry of the HoloLens was outputted through the Unity Engine as well. This was done by creating transform messages through Unity, which was then broadcast to a ROS server via a ROS bridge web socket. Transform messages contain information about the position and location of a robot from the physical world. The position and location of the robot are then used in a virtual environment.

6 Tests Applied

Sensory Data test: The first test determines the feasibility of finding the distance to the spatial mesh in Unity, using a ray casting, in order to create a virtual laser scan. This will test for the accuracy, geometry, and location of the points provided through the HoloLens's geometrical mesh. Readings from this test include an array of distances from the points in front of the robot and positional marker of where the robot is located.

Odometry Test: The second test determines the validity of using the HoloLens as an odometry sensor. The test will check the ability of the HoloLens to maintain its position and direction in the virtual environment and confirm the odometry of the HoloLens does not reset such to confuse the ROS odometry. Readings from this test will include a map of measured distances, the distances in the real world based on the ROS Cartesian plane, HoloLens odometry readings, and the Pioneer's odometry readings.

Mapping with SLAM: The third test will demonstrate the abilities of the HoloLens in creating a boundary map as well as keeping its odometry in the arena. This technique is called simultaneous localization and mapping (SLAM) and entails the combination of the odometry and mapping of the robot. The HoloLens will be mounted on top of the robot and driven around to create the map of the arena. A successful mapping would include a rough outline of the arena in a 2D perspective with accurate positioning of the robot.

Safe Wanderer: The fourth test will involve using the HoloLens as a primary sensor for an autonomous program. The challenge will include utilizing the HoloLens's virtual laser scan and odometry to be able to safely navigate a loop around the arena without colliding with the sides of the walls. The robot must keep a safe distance from the wall given the data being sent from the HoloLens. The algorithm is set to turn towards an opening if there was an obstacle in front of the robot, 0.5 meters away. A successful test of the safe wanderer program would be the successful traversal of the arena without collisions.

Autonomous Movement: Once the previous tests are complete, the final test will be conducted where an autonomous algorithm called move base will utilize the HoloLens as a localization sensor. A map will be given to the algorithm, constructed from previous test runs with the HoloLens, and odometry will be provided by the HoloLens for navigating the arena. The robot must reach a set goal which will be provided by a ROS utility called RVIZ. A successful test will include being able to navigate the arena without collisions, localizing its position in the map, and finding its way to the given goals.

7 Experimental Results

7.1 Test 1: Sensory Test

An arena was set up with the use of several cushion chairs made to form a small box with an opening at the side. Testing proceeded with the use of a modified TurtleBot teleoperation or a joystick controller to position the robot at several locations. Figures 1 and 2 are the results from the testing of the HoloLens for distance gauging and geometrical representation.

These early testing periods included using the HoloLens with a leveled angle and no other utilities running. Odometry was not yet implemented at this point nor were the mapping utilities attempted for use at this point. The shapes of the wheeled cushion chairs were clearly seen in the laser scan on the HoloLens. This proves that the HoloLens can simulate a laser scan through Unity by calculating ray casts to formulate a distance from the environment.

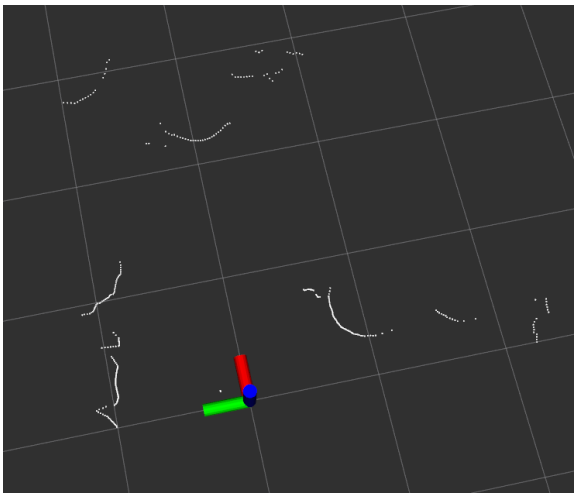

Figure 1: Initial scans of laser scan

Figure 2: Physical location of robot

It was noticed that the HoloLens detected walls differently during varying lighting conditions and the similarity of coloring from the floor to the walls of the arena. This led to changing the arena configuration and adjusting lighting around the environment of the arena. The HoloLens would sometimes momentarily lose spatial recognition and have to adjust for the different lighting condition. An advantage of the HoloLens is that the sensor has a reduced potential for seeing through glass, a common problem on similar sensors [13].

7.2 Test 2: Odometry Testing

A test of the sensor odometry was performed on a measured arena. The test was performed several times through keyboard-teleoperated control at a minimal velocity of 0.2 m/s and an angular velocity of 0.1 rad/s. One difficulty during this phase was adjusting the transform messages from the Unity coordinate plane (yzx format) into the ROS coordinate plane (xyz format). The following table shows the results concluded from the test. Figure 3 represents the distances measured and tested for.

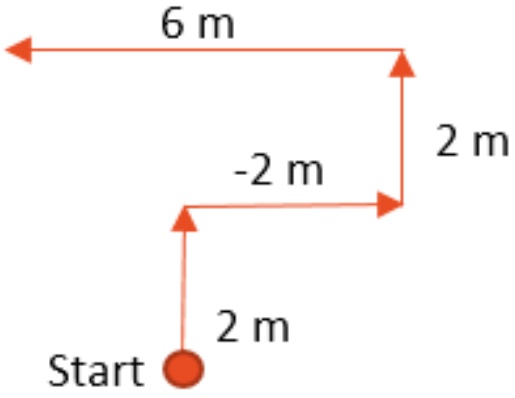

Figure 3: Map of traversal

Figures 4 and 5 describes the difference from the ideal point from the perspective of reality, the HoloLens, and the odometry of the Pioneer 3. It can be observed that the Pioneer has an additive error as the robot is moved from the points described in Figure 3. It can be noticed that the HoloLens follows a trend closely to the error in reality, with larger deviations in the Y-axis.

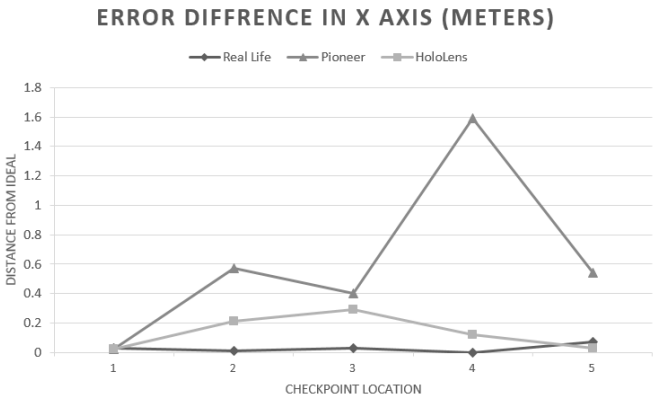

Figure 4: X axis error graph

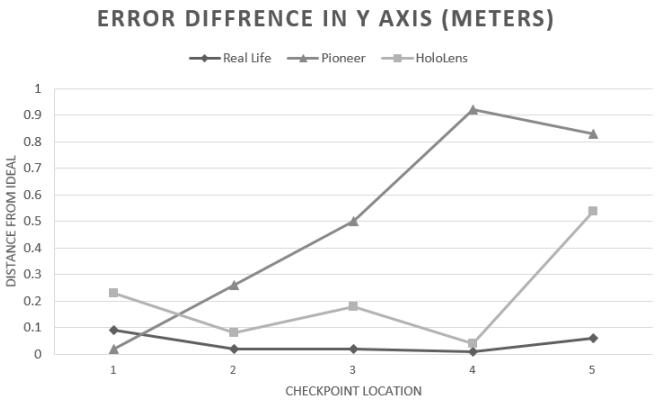

Figure 5: Y axis error graph

Ideal Distance (x, y, z) (meters)	Real World (meters)	HoloLens (meters)	Odometry (meters)
2, 0, 0	2.03, -0.09, 0	2.02, 0.23, 0	2.02, 0.02, 0
2, -3, 0	2.01, -2.98, 0	1.79, -3.08, 0	2.57, -2.79, 0
4, -3, 0	3.97, -3.02, 0	3.71, -3.18, 0	4.40, -2.5, 0
4, 3, 0	4.00, 2.99, 0	3.88, 2.96, 0	2.41, 3.92, 0
0, 0, 0	0.07, -0.6, 0	0.03, -054, 0	0.54, -0.83, 0

Figure 6: Measurement table

7.3 Test 3: SLAM Testing

SLAM testing was conducted after odometry was established with the robot. Teleoperation of the robot was done and was used to navigate several arenas. In this early version of the odometry data packages, the arenas were traversed approximately five times for the SLAM mapping utility, gmapping, to build the map correctly. It was noticed that the HoloLens had difficulty identifying the cardboard arena (arena 1). A two degree change was added to the ray cast angle direction in order to improve the identification process, however it still had difficulty in discovering the arena. To mitigate this problem, another arena, arena 2, was designed so that the HoloLens could identify it through a different material. Figures 7 and 9 show the arenas constructed from mapping.

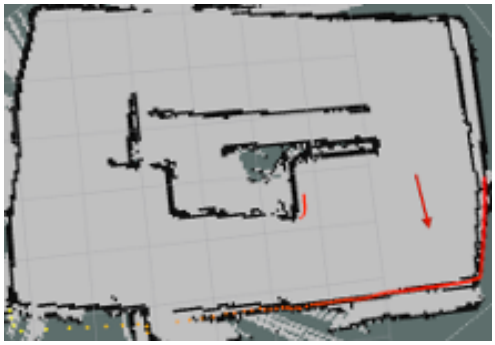

Figure 7: 2D map of Arena 1

Figure 8: Arena 1 Layout

Figure 9: 2D Map of Arena 2

Figure 10: Arena 2 Layout

7.4 Test 4: Safe Wandering Testing

Safe Wandering testing was done through the use of a program designed to seek out a clear route for traversal of 0.5 meters and turn right or left when a distance greater than 1.2 meters was in view. The robot was set to wander under the same testing conditions as done in the SLAM testing arena. During successful trials, the safe wanderer program ran around the arena

for six loops without interruption or human intervention.

During initial testing, the robot would occasionally lose odometry and laser scan data causing the robot to collide head-on into the walls. This was fixed through reducing the speed to 0.3 m/s and the angular velocity of 0.5 rad/s. The reason for this was the HoloLens lost internal localization and deleted the existing spatial mesh to remap its environment causing the virtual laser scan to detect open space. During later testing, the laser scan was reliable and the robot did not collide with the walls. This proved that the robot could reliably use the HoloLens laser scan and odometry.

7.5 Test 5: Move Base Testing

Autonomous navigation was achieved through a commonly used ROS program, move base. A map was rendered previously through the use of safe wander or through teleoperation. The robot was then placed at a starting position, given a starting indicator, and waypoints (marked white) were added to the navigation stack for the move base algorithm to navigate around the arena. The move base algorithm was then used to calculate a path around the arena and get odometry and laser scan data from the HoloLens sensor. Figures 11 and 12 show the robot navigating the arena with the HoloLens.

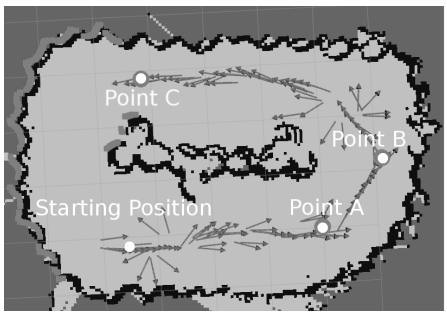

Figure 11: Path followed by the robot while base move is active

Figure 12: Robot Moving with Move Base

8 Further Testing

Three aspects of this experiment caused concern. First, only one test had been done on the odometry. Second, the results in the Y-Axis were ambiguous as to whether or not the HoloLens odometry had additive error. Third, the connection between Unity and ROS was not optimized for object oriented coding. We rewrote the ROS Bridge communication module to better integrate with Unity allowing for separate instantiations of the publisher/subscriber class. This in turn, allowed different game objects in unity to communicate directly with ROS. After this was finished we tested the robot on a more complex odometry course in order to validate our previous findings. The test was performed three times and the results clearly shows the odometry on the HoloLens is far more accurate than conventional motor encoders. The test layout was performed by running a course as shown in Figures 15, 16, and 17. As seen in Figures 13 and 14, display the average error for all trials.

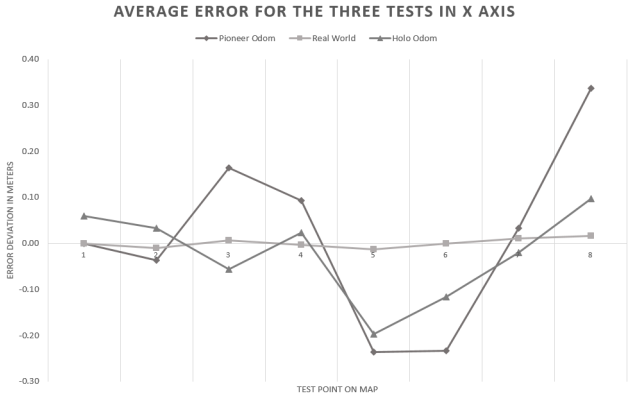

Figure 13: Average Error on X Axis

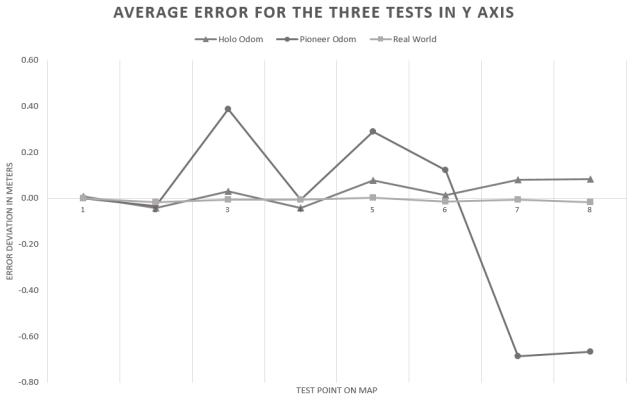

Figure 14: Average Error on Y Axis

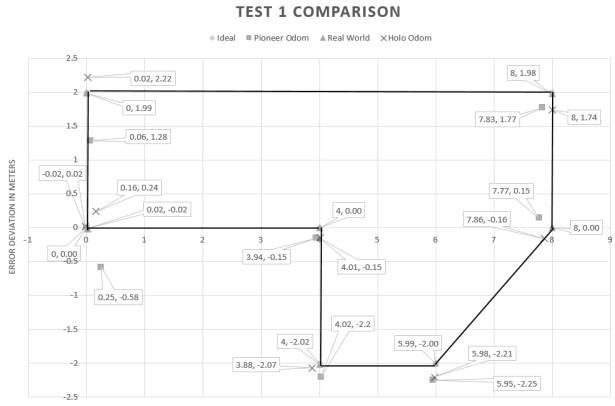

Figure 15: Map of the Odometry Points from Test 1

Ideal	Pioneer Odom	Real World	Holo Odom
0,0	0.0,0.0	0,0	-.02,.02
4,0	3.94,-0.15	4,0	4.01,-.15
4,-2	4.02,-2.2	4.0,-2.02	3.88,-2.07
6,-2	5.95,-2.25	5.99,-2	5.98,-2.21
8,0	-7.77,0.15	8,0	7.86,-.16
8,2	7.83,1.77	8,1.98	8,1.74
0,2	0.06,1.28	0,1.99	0.02,2.22
0,0	0.25,-0.58	0.02,-0.02	0.16,.24

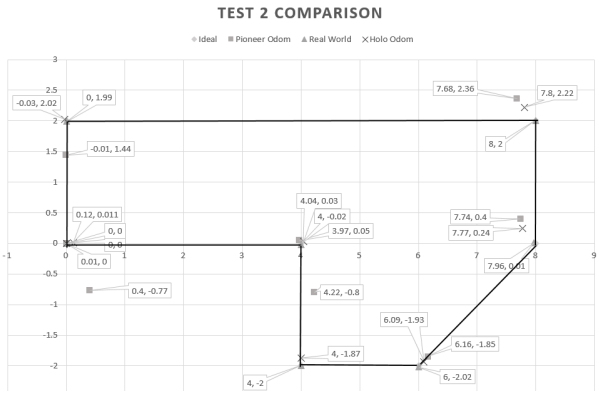

Figure 16: Map of the Odometry Points from Test 2

Ideal	Pioneer Odom	Real World	Holo Odom
0,0	0.0,0.0	0,0	0,0
4,0	3.97,-0.05	4,-0.02	4.04,0.03
4,-2	4.22,-0.8	4.0,-2	4,-1.87
6,-2	6.16,-1.85	6,-2.02	6.09,-1.93
8,0	7.74,0.4	7.96,0.01	7.77,0.24
8,2	7.68,2.36	8,2	7.80,2.22
0,2	-0.01,1.44	0,1.99	-0.03, 2.02
0,0	0.40,-0.77	0.01,0	0.12,0.01

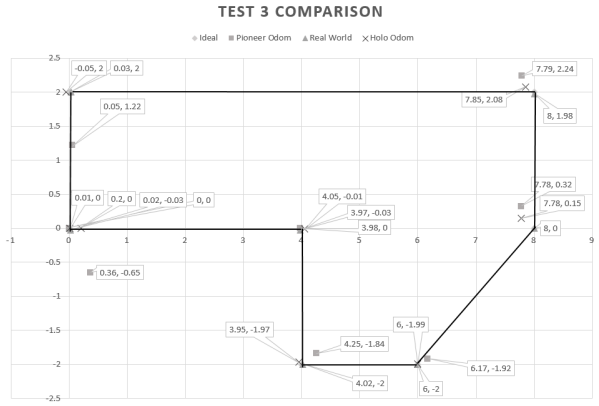

Figure 17: Map of the Odometry Points from Test 3

Ideal	Pioneer Odom	Real World	Holo Odom
0,0	0.0,0.0	0,0	0.2,0
4,0	3.98,0	3.97,-0.03	4.05,-0.01
4,-2	4.25,-1.84	4.02,-2	3.95,-1.97
6,-2	6.17,-1.92	6,-2	6,-1.99
8,0	-7.78,0.32	8,0	7.78,0.15
8,2	7.79,2.24	8,1.98	7.85,2.08
0,2	0.05,1.22	0.03,2	-0.05, 2
0,0	0.36,-0.65	0.02,-0.03	0.1,0

9 Conclusion and Future Work

This paper has presented the feasibility of applying the Microsoft™ HoloLens towards a robotic application such as autonomous driving and SLAM mapping. The tests done in this paper demonstrated that the HoloLens has the sensor capability to allow a robot obstacle detection, environmental mapping, reliable odometry, and autonomous navigation. The accuracy of the odometry was acceptable and outperformed the on-board odometry provided by the Pioneer 3 servos and was reliable enough for autonomous navigation. From these results, we conclude that the HoloLens can be used as a feasible robotic sensor and integrate powerful augmented reality capabilities into the robotic system.

Future research will involve investigating additional applications of the HoloLens for improved user interface and comparing its performance against current sensor navigation systems. This will include a comparison of the HoloLens against a commercial-grade LIDAR system and observing any differences in speed, accuracy, and reliability. An application of the Unity 3D space shall be used in experimentation for controlling a robot and providing a pathfinding solution from Unity to the robot in the real world using the HoloLens gesture tracking and positioning system. Currently, an experiment is under way using the HoloLens as an AR windshield for autonomous vehicles. This will simulate the experience of being inside an autonomous vehicle from a first person perspective with additional infographics of the robot and/or of the environment.

Acknowledgments

This research was supported by the National Aeronautics & Space Administration through the University of Central Florida's NASA FLORIDA SPACE GRANT CONSORTIUM.

References

[1] ACKERMAN, Evan: "Sweep is a 250 LIDAR with a Range of 40 Meters That Works Outside". In: *IEEE Spectrum* (2017). http://spectrum.ieee.org/automaton/robotics/robotics-hardware/sweep-lidar-for-robots-and-drones

[2] CUI, Nan ; KHAREL, Pradosh ; GRUEV, Viktor: "Augmented reality with Microsoft HoloLens holograms for near infrared fluorescence based image guided surgery". In: *Proc. of SPIE Vol* Bd. 10049,8 Feb. 2017

[3] EVANS, Gabriel ; MILLER, Jack ; PENA, Mariangely I. ; MACALLISTER, Anastacia ; WINER, Eliot: "Evaluating the Microsoft HoloLens through an augmented reality assembly application". In: *SPIE Defense+ Security* International Society for Optics and Photonics, 2017, S. 101970V–101970V

[4] GARON, Mathieu ; BOULET, Pierre-Olivier ; DOIRONZ, Jean-Philippe ; BEAULIEU, Luc ; LALONDE, Jean-François: "Real-Time High Resolution 3D Data on the

HoloLens". In: *Mixed and Augmented Reality (ISMAR-Adjunct), 2016 IEEE International Symposium on* IEEE, S. pp. 189–191, 2016

[5] HINCAPIÉ-RAMOS, Juan D. ; ÖZACAR, Kasim ; IRANI, Pourang P. ; KITAMURA, Yoshifumi: "GyroWand: An Approach to IMU-Based Raycasting for Augmented Reality". In: *IEEE computer graphics and applications* 36(2):pp. 90-96, 2016

[6] KAMARUDIN, Kamarulzaman ; MAMDUH, Syed M. ; SHAKAFF, Ali Yeon M. ; ZAKARIA, Ammar: "Performance analysis of the microsoft kinect sensor for 2D simultaneous localization and mapping (SLAM) techniques". In: *Sensors* 14 (2014), Nr. 12, S. 23365–23387

[7] LIM, TY ; YEONG, CF ; SU, ELM ; CHIK, SF ; CHIN, PJH ; TAN, PH: "Performance Evaluation of Various 2-D Laser Scanners for Mobile Robot Map Building and Localization". In: *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 8 (2016), Nr. 11, S. 105–109

[8] N.A.: "HoloLens Hardware Details". In: *Microsoft Developer Center* (2017). `https://developer.microsoft.com/en-us/windows/mixed-reality/hololens_hardware_details`

[9] OMARA, Hesham Ibrahim Mohamed A. ; MOHAMED, Khairul Salleh Mohamed S.: "Indoor mapping using kinect and ROS". In: *Agents, Multi-Agent Systems and Robotics (ISAMSR), 2015 International Symposium on* IEEE, 2015, S. 110–116

[10] SAMSURI, Saiful B. ; ZAMZURI, Hairi ; RAHMAN, Mohd Azizi A. ; MAZLAN, Saiful A. ; RAHMAN, Abdul Hadi A.: "Computational Cost Analysis of Extended Kalman Filter in Simultaneous Localization and Mapping (EKF-SLAM) Problem for Autonomous Vehicle". In: *ARPN Journal of Engineering and Applied Sciences* 10 (2015), Nr. 17, S. 153–158

[11] SANTOS, Joao M. ; COUCEIRO, Micael S. ; PORTUGAL, David ; PROCHA, Rui: "A sensor fusion layer to cope with reduced visibility in SLAM". In: *Journal of Intelligent & Robotic Systems* 80 (2015), Nr. 3-4, S. 401

[12] THOMAS, Dirk: "What is ROS?". In: *ROS* (2014). `http://wiki.ros.org/ROS/Introduction`

[13] ZHANG, T ; CHONG, ZJ ; QIN, B ; FU, JGM ; PENDLETON, S ; ANG, MH: "Sensor fusion for localization, mapping and navigation in an indoor environment". In: *Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2014 International Conference on* IEEE, 2014, S. 1–6

[14] ZLATANOVA, Sikya: "Augmented reality technology". In: *GISt Report No. 17, Delft, 2002, 72 p.* (2017)

Bradford Towle Jr. is currently teaching at Florida Polytechnic University as an assistant professor. He graduated from Montana State University with a bachelor degree in computer engineering. He went on to receive his masters and doctorate degrees at the University of Nevada, Reno in Computer Science and Computer Science and Engineering respectively. Dr. Towle focused primarily on software engineering and game design for his masters degree and proceeded to study robotic control architectures for his doctorate. He has taught computer science at Landmark College, which focuses on students with learning disabilities, and Keene State College before coming to Florida Polytechnic. His current research involves demonstrating the feasibility of an augmented reality device simulating a passenger awareness system for autonomous vehicles as well as integrating augmented reality into robotic testing.

Esteban Segarra is an undergraduate at Florida Polytechnic University studying computer engineering with a concentration in machine intelligence. He has been working on various projects related with robotics and Machine Learning including American Sign Language to visual representation. Currently he is researching the use of machine intelligence for robotics in space exploration. Through the use of AR technology, he would be able to simulate a scenario and its agents for the use of machine intelligence for control.

Advancing Quality Assurance Through Metadata Management: Design and Development of a Mobile Application for the NRDC

Connor Scully-Allison*, Hannah Muñoz*, Vinh Le*, Scotty Strachan*,
Eric Fritzinger*, Frederick C. Harris, Jr.*, and Sergiu Dascalu*
University of Nevada, Reno, Reno, NV, 89509, USA

Abstract

In this paper we present the design, implementation, and impacts of a cross-platform mobile application that facilitates the collection of metadata for in-situ sensor networks and provides tools assisting Quality Assurance processes on remote deployment sites. Created in close conjunction with scientists and data managers working on environmental sensor networks, this paper details the software requirements, specifications, and implementation details enabling the recreation of such an application. In a discussion on how this software improves on existing techniques of logging contextual metadata and quality assurance information, we show that this application represents a significant improvement over-existing-methods. Specifically, the proposed application allows for the near-real time update and centralized storage of contextual metadata. Compared to prior methods of logging, often physical notebooks with pen and paper or program comments on embedded field sensors, the method proposed in this paper allows for contextual information to be more tightly bound to existing data sets, ensuring use of collected data past the lifetime of a specific research project.

Key Words: Data management, data science, mobile application, sensor networks, software engineering, cross platform mobile development.

1 Introduction

Individual researchers and one-off projects dominate the model of data collection in traditional climate/environmental research [9]. In traditional research, single use data proves extremely effective at answering a singular project's research questions and fulfilling research requirements attached to funding streams. However, despite the short-term success of such a model, a clear problem arises when another research team wishes to use this previously collected data, or when data need to be integrated into larger syntheses [3]. This problem drives the need for complete, accurate, and usable metadata.

Metadata is considered a major part of the data lifecycle. Creation of metadata include information surrounding the data set, such as format, file names, and measurement units, and information about the experiment producing the dataset, such as documenting data processing steps and contextual information to the data [13]. Its purpose is to make the datasets quick and easy to understand [20]. Unfortunately, that is not always the case. Traditionally, scientific metadata is kept in notebooks and papers, a holdover from days before computers. This creates fractured data, where it's hard to relate the electronic data sheets to hand written documents [18].

Due to the narrow focus of typical projects, only the original researchers intimately know how the data was generated. Metadata is often non-standardized, incomplete, and stored in temporary formats. Some communities and organizations have developed their own metadata standards; however, several different metadata standards can exist within any given discipline [7]. Eventually, over time -- or given enough distance -- the value of these data sets is diminished to other researchers and the public. It becomes harder to recover and ascertain contextual information that is essential to decoding it and metadata standards rarely address long-term preservation [18]. Methods for uniform quality assurance and metadata collection are being recognized as the next major challenge for data intensive science as collection becomes increasingly automated and results globally disseminated [21].

This paper proposes a mobile application that manages and maintains quality assurance metadata about data collected from remote sensor networks. The Quality Assurance (QA) Application described in this paper represents a positive step forward into modern data collection models by centralizing, modernizing, and standardizing contextual metadata for environmental sensor systems. The QA App gives technicians and researchers a tool for dynamic modification and creation of contextual information relating to hundreds of live data streams in a statewide sensor network.

Continuing from here, this paper is structured as follows: Section 2 presents a survey of scholarly works related to the app developed; Section 3 details the software specifications and use cases of the application; Section 4 discusses the architectural and user interface design of this application; Section 5 discusses how the application was implemented; Section 6 evaluates the success of implementation and Section 7 explores ideas for future developments.

2 Related Work

At the broadest level, Quality Assurance refers to the preventive maintenance and management process employed to reduce inaccuracies in data automatically logged by sensors [5]. Although many works touch on the idea of Quality Assurance,

* Department of Computer Science. Email: cscully-allison, hannahmunoz, vle@nevada.unr.edu, scotty@dayhike.net, ericf@unr.edu, fred.harris@cse.unr.edu, dascalus@cse.unr.edu.

the most seminal motivating work published on the subject comes from a 2013 paper "Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data" [4]. In this paper, the authors put forth a comprehensive, generalized set of practices to optimize QA on environmental field sensors. They suggest that QA procedures be automated, well documented, and complete metadata maintained alongside data. This work presents Quality Assurance as a "process oriented" approach to data management, this strongly implies that no single software solution can provide effective QA but can only rather aid the QA process performed by humans. Accordingly, a significant number of background works on this subject study the analysis and creation of software that best facilitates good QA practices.

A further example of a motivating work indicating the need for Quality Assurance practices in the paper "Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis" [1]. This paper outlines the development of a software environment that "addresses technical challenges related to accessing and using heterogeneous sensor data from within the Kepler scientific workflow system." Within the bounds of their end-to-end examination of this existing sensor network workflow, the authors indicate on several occasions that a clear need exists for quality assurance practices with in-situ sensor networks. They also acknowledge that existing software solutions used to facilitate these practices are not well developed. Our QA application intends to fill this proposed gap.

Outside of motivating works driving research in QA, there also exists several works that explore the practical or theoretical implementation of Quality Assurance processes and software. One, "Meta-information concepts for ecological data management" represents an early survey of many data management needs for ecological sensor data collection [12]. This paper suggests exactly at which stage in the data collection pipeline to place QA processes and software. Another paper, "Anomaly detection in streaming environmental sensor data: A data-driven modeling approach" shows a more practical approach to Quality Assurance by suggesting that expected shifts in the quality of data can be anticipated by using data-driven modeling techniques. Although the approach taken by the authors of this paper is not representative of the approach we undertook, it shows that there exists a strong interest in applying modern software engineering techniques to the problem of Quality Assurance.

Along similar lines, in the paper, "Automatic processing, quality assurance and serving of real-time weather data" the authors demonstrate that there exists a strong interest in developing software to automate and streamline the process of Quality Assurance on environmental sensor data [22]. The authors of this paper propose a software to manage and utilize statistical metadata that can indicate the quality of data through uncertainty values. The concept of collecting metadata in a standardized format for quality assurance strongly reflects the goal of the Quality Assurance software developed and detailed in this paper for the NRDC. However, our approach is unique in that it is oriented towards metadata collection as an end goal rather than as a supplement to Quality Control computations. Taken together, the above papers generally indicate a strong research interest external to the Nevada Research Data Center (NRDC) in the development of Quality Assurance software, however there also exists a well-documented interest in developing software within this organization as well.

The NRDC is a data management organization dedicated to the, "storage, retrieval, and analysis of research data that is relevant to the needs and interests of the state of Nevada" [15]. Conceived apart of a NSF Track 1 project, the NRDC represents the collaborative efforts of top research institutions, including the University of Nevada Reno, the University of Nevada Las Vegas and the Desert Research Institute [14, 16]. It presently supports the data sets of five projects and works actively with external research networks to disseminate and preserve data for continued research.

References to the considerations of a Quality Assurance system for the NRDC appear in early literature proposing practices and architecture for its predecessor project NCCP [11]. These works present Quality Assurance and Quality Control (QC) as crucial elements of any large scale environmental research project. They also impress upon the reader a need for a standardized and centralized set of tools which enable universal comprehension of data being collected. From this specific need to improve on existing QA practices, a quality assurance application was conceived.

3 Software Specifications

The QA App was developed with many functional and nonfunctional requirements in mind. These requirements were decided on after extensive talks with several data management experts and stakeholders. Detailed in Table 1, these requirements guided design and development of core functionalities for the QA application. Using an agile development method, these requirements went through several iterations before settling into the current list.

The nonfunctional requirements set many constraints on development, but most importantly dictated that the system should be multiplatform, upload and download data at only one point and perform logins with an SSH certificate. This set of requirements informed development by cementing the software and architecture used to implement the app and indicate how it should interface with the backend server.

Using these functional requirements, a series of use cases were constructed and mapped in a use case diagram, found in Figure 1. This process informed the principal design phase of this applications construction and was frequently referenced or tweaked alongside the software specifications through the implementation phase. The description of each use case follows:

- LogIn

Field technicians must log into the app. Once logged in, technicians are given a list of projects they are associated with. This helps reduce the amount of unnecessary data downloaded. Technicians can also add new entries and upload them to the

Table 1: Functional requirements

Functional Requirements	Description
Input Data	The user shall be able to enter new data into the QA app.
Upload Data	The user shall be able to upload data to a secure database.
Read Data	The user shall be able to download entries from the database to their mobile device and view previous data entries.
Edit Data	The user shall be able to edit previous entries and upload the change data to the database.
Navigate Data	The user shall be able to move between different screens of the app, and input data.
User Authentication	The user will be able to authenticate themselves to access secure functionalities.
Delete Server Data	The user, with proper authorization levels, shall be able to delete data stored on the database.
Upload Photos	The user will be able to launch the device's camera and upload a photo from their photo gallery on their device.
Save Unsynced Data	The user shall be able to save data locally on their device to upload it to the server at a later point

server. Administrative technicians, once verified through the log in, are given the ability to edit or delete entries already synced to the server.

- SyncToServer
  Connects to the server and uploads new data entries found on the phone. Then, downloads new data found on the server. The app can manually be synced by pressing the synchronization icon on the header bar on the front page.

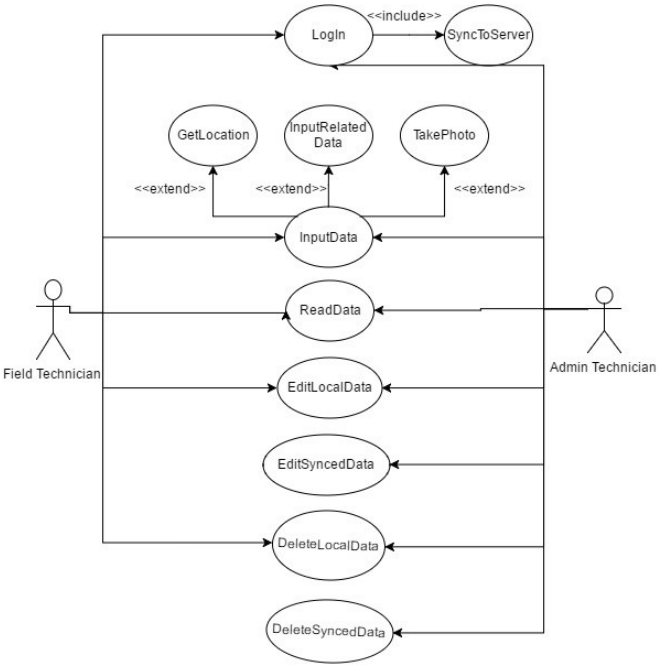

Figure 1: Diagram showing use cases and their actor interactions for the quality assurance application

- InputData
  Allows the user to input new data. Opens a blank template for whichever dataset they are choosing to input. Once finished, it is saved to local memory until synced to the server.
  - TakePhoto
    Opens the phone's camera app to take a picture that can be uploaded alongside the data set, like the System in Figure 2. Not every data set can have a photo.
  - GetLocation
    Uses the phone's GPS to fill out latitude and longitude coordinates. Only two types of data sets need GPS location.
- ReadData
  All users must be able to view the data, regardless of whether or not they are logged in. This is so users who are not a part of the project, but are interested in the data, can view it. To read the data, users need only to navigate to their desired object and click on the name.
- EditLocalData
  Users are allowed to edit entries that have not yet been synced to the server. Users can navigate to unsynced data entries and select the edit button to change them.
- EditSyncedData
  Administrative technicians are allowed to edit data already synced to the server. If an admin is logged in, they can edit entries by navigating to it and clicking the "Edit" button. If the user is not an admin, this button will be greyed out. The changes will be uploaded the next time the app is synced to the server.

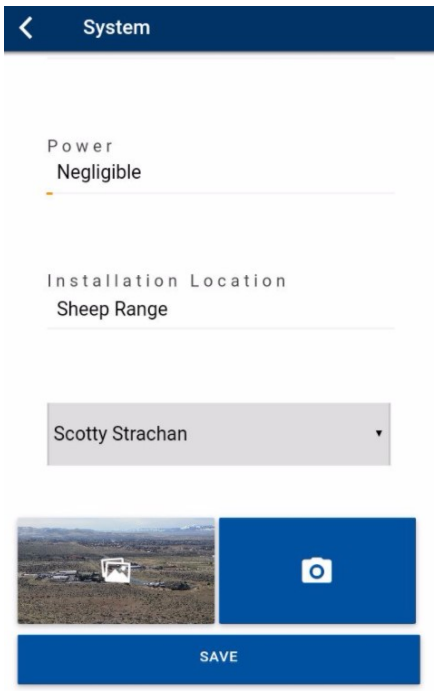

Figure 2: An example of inputting data with a picture from the phone's camera

- DeleteLocalData
  Users can delete entries that have not yet been synced to the server. Users can navigate to unsynced data entries and select the delete button to remove them.

- DeleteSyncedData
  Administrative technicians are allowed to delete data entries already on the server. If an admin is logged in, they can delete entries by navigating to it and clicking the "Delete" button. If the user is not an admin, this button will be greyed out. The changes will be uploaded the next time the app is synced to the server.

4 Software Design

4.1 Architectural Design

When designing the Quality Assurance Application, there were several key requirements that shaped the development process. First, the QA application must be able to structure the enormous amounts of service entries and access them in a timely manner. To achieve this, the application utilizes a data access hierarchy that narrows down the amount of data queried. Second, the application must handle the situation that there is no available internet or cell signal in the immediate area. The application manages this problem by storing the changes locally and allows users to commit these changes when they reach an appropriate area. Finally, and most importantly, the application cannot function as a singular client-side application without any support. The QA application consists of a client and server with the client utilizing a Model-View-Controller(MVC) architectural pattern and the server utilizing a microservice architecture.

The client-side application utilizes a MVC architectural pattern as shown in Figure 3. This architectural pattern was executed while utilizing the Google Angular 3 framework alongside the C# and Javascript programming languages. The model section of the client-side application is the central nexus of interaction with the main NRDC System. This contains routines that would handle the HTTP communication with the server end, as well as the manipulation of locally stored data. The view section is tasked with the primary task of satisfying the first requirement and showcasing the data in a hierarchal format. In this section, the functionality to navigate through the hierarchy, view imagery, and handle conflicts syncing with the database is handled. Finally, the controller section serves the primary master and control portion of that application that dictates the behavioral actions that result from the interactions made by the user. This section is where the application would issue the command to shift the page views, create the transitional effects in-between views, and initiate the interaction with the model section.

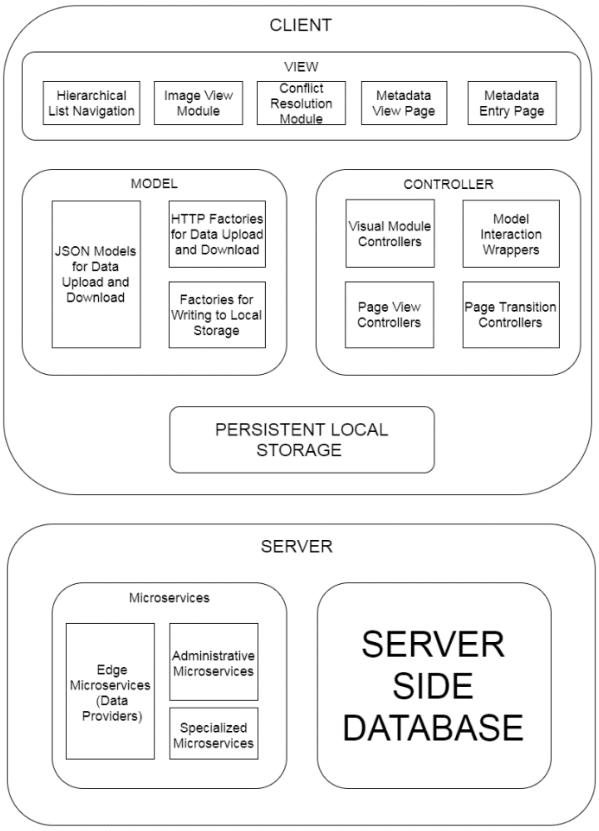

Figure 3: High-level block diagram detailing the architecture of the QA app. Client and server are connected via http calls from the mobile application to the Edge Microservices

The server side of the application utilizes a microservice architecture where each web service is independent of each

other and are combined to create greater functionality. These services can each be classified into three main groups: Edge Microservices, Administrative Microservices, and Specialized Microservices. The Edge Microservices are infrastructure services that perform the role of data provider from the data base to the client application. The Administrative Microservices perform the role of security and provides varying level of access to the hierarchy, based on the individual's status within the project. The Specialized Microservices provide complex functionality to the application that are outside that of the Edge Microservices. These functionalities can include anywhere from photo compression and searching to file conversions.

4.2 UI Design

Throughout development, the User Interface design of the Quality Assurance application transformed drastically. Initial designs for a prototype implementation of this application, visible in Figure 4, were relatively simplistic, utilitarian and focused on the highest level of sensor network metadata that would be managed: a "project". A project described a specific cluster of in-situ sensors networked together that collected data associated with answering the question of a single research project. Given the broad scope of "a project," the metadata collected and stored was general in scope, requiring only one page and a few form fields. As development on this application expanded so too did the UI design to better accommodate the technical and personal needs of user stakeholders.

The first major design shift was a visual one. The application transitioned out of the simple Flat Design of the initial prototype into a Material Design variant with the addition of more primary color contrasts, subtler form fields and floating action buttons. This design choice gave the application a more modern and commercial feel that people have come to expect from a high quality mobile application.

A second design shift, more critical to the proper functioning of the application itself, was the inclusion of a hierarchical navigation structure. While the details of this structure are explored in Section 5, from a design standpoint it was crucial to enable the fast traversal of this hierarchy by making each item on every navigation level a large, clickable button that only leads to a sub list of items contained by the item clicked. Figure 5 contains an example of one of these sub lists. By placing a link to the information about the previously clicked item in the top right corner, the design of this list navigation negates the possibility of users unintentionally opening information pages that will slow their navigation. This streamlined design enables users to find the lowest level component they want in seconds, and then open information pages for the editing and viewing of related metadata.

Finally, at a late stage in development stakeholders expressed a strong need for the display of saved pictures showing the makeup of a sensor system or the structure of a specific component. This need drove the development of a dedicated image viewing component which retrieved images from a remote or local source and displayed them on appropriate pages. The addition of this component required significant planning as each image had many management functions associated with it: saving the image to the database, saving the image locally, delete locally, delete on the database, open an enlarged view of

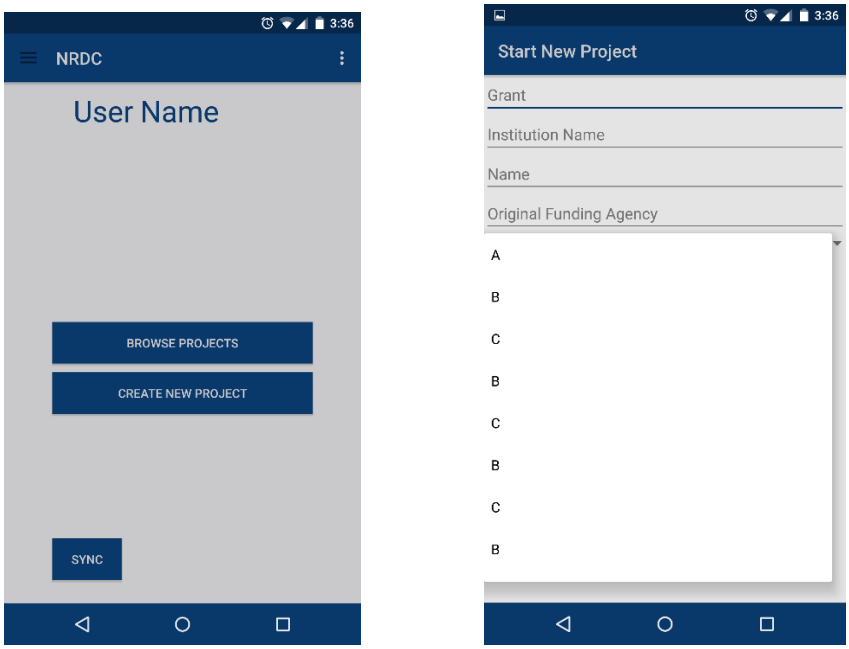

Figure 4: Screenshots of the initial UI design for the NRDC QA application. Left, the main screen shows the basic functionality implemented in the prototype. From here users can browse projects or create a new project. Right, the "Start New Project" interface was used to input relevant metadata about the sensor research project being documented

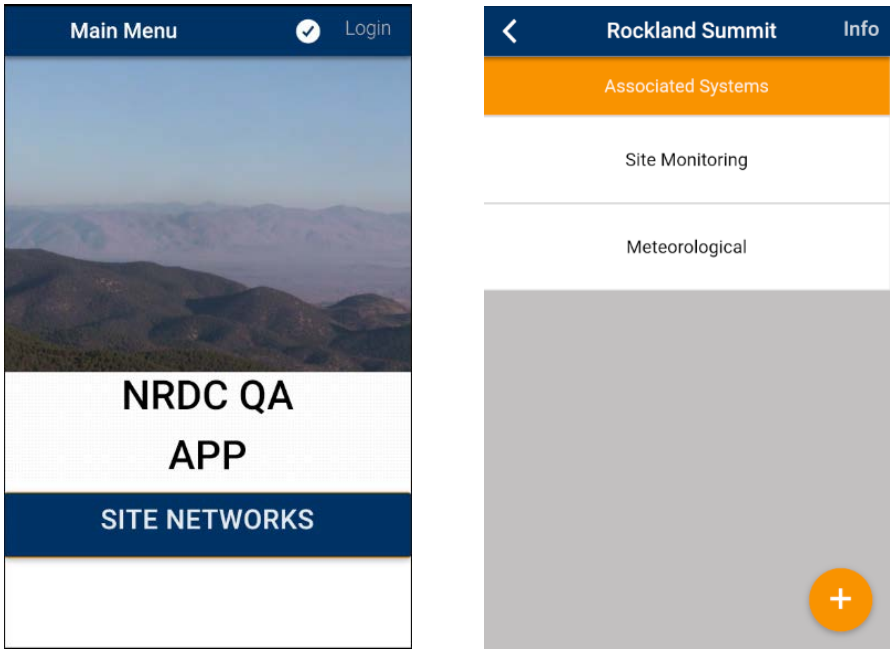

Figure 5: Screenshots indicating the final UI design of the Quality Assurance Application. On the left is an updated and ascetically pleasing main menu screen. On the right, we see an example of a sub-list in the hierarchal navigation structure. Here we see that "Site Monitoring" and "Meteorological" are systems associated with the "Rockland Summit Site". Clicking on either item will populate a new list of "Deployments" that are in the chosen system

the image, and more. This suite of functionality had to be added without cluttering the limited real-estate of a mobile screen. We overcame this problem with the inclusion of a floating action button which expanded into an array of buttons that each perform one of the above stated functions.

Finally, the interface provides users with a simple straightforward medium to quickly access to the forms required by technicians. Visually reminiscent of Google's material design, this application takes cues from public facing software to provide a refined interface to encourage smoother adoption of this app among unskilled mobile technology users. Large buttons and clickable lists simplify use for technicians wearing gloves when performing maintenance on sites in high elevations or in colder months.

5 Implementation

Currently, the NRDC QA application is comprised of a front-end system developed in the Ionic Framework, and a back end comprised of essential and independent web services communicating through a centralized hub [10]. The data transferred between the two are stored inside a Microsoft Virtual Environment with Microsoft SQL Server 2012 as the primary database management. These two main components together allow for a seamless interface between the main databases and the client application.

For the front-end system, the QA application was built with the Ionic Framework [8]. This framework utilizes HTML and CSS as a wrapper to manipulate the interface, as well as Javascript to apply functionality. Once completed, the HTML, CSS, and Javascript are then compiled into the appropriate codebase: either Apple or Android.

A wide collection of libraries and modules were used to simplify development at various stages of implementation. Primarily, Google's AngularJS was used as a structural framework for the Javascript codebase and allowed for a more object-oriented approach to manipulation of the HTML and interaction with microservice APIs [6]. Additionally, Node Packet Manager (NPM) and Bower were used as the main package managers for this application. They ensured libraries, assets, and utilities were regularly updated and organized.

Organization of the QA app follows a pattern representative of the hierarchical organization of existing sensor networks managed by the NRDC and associated institutions. At the topmost level, the application presents the user with a selection of Site Networks: a representation of several data collection sites connected by their similarity of purpose or project associations. From there the user selects a site associated with that network, a system associated with that site, and a deployment associated with that system, ending with a hardware component associated with that deployment. This workflow of "tunneling" down into atomic components gives data scientists a logical means of quickly locating the exact sensor network element they seek by leveraging their knowledge of existing infrastructure.

At any point in the navigation of the sensor network hierarchy, a user can add a new entry to the list of displayed entries or view the details of existing ones. Whether choosing to display or create, the user will be greeted with the same page. If displaying

data about an existing item, the page will be populated with data about the selected element. If the user chooses to create a new item, the form fields on this page are blank and ready for input.

On the element creation screen, visible in Figure 2, the user inputs information into blank form fields that expand or contract to fit the size of the input data. The user can also choose to upload a related picture or get their location via their phone's GPS. This functionality enables field technicians to upload accurate location data about sites they are working on with the touch of a button. Once all necessary information is entered, the new metadata entry can be saved locally. And, once the user is done adding new entries, they can upload them en-masse to the server for storage in the database.

On the view screen the user is presented with a few different options compared to the creation screen. Principally she can no longer save an entry, only edit or delete with proper permissions, and there appears a floating orange icon in the bottom right corner visible in Figure 6. From the submenu which this button populates, users can add two different types of metadata about an entry, a document and a service entry. Documents allow users to add related files to a metadata entry. Service Entries are entered when scientific equipment is repaired or replaced.

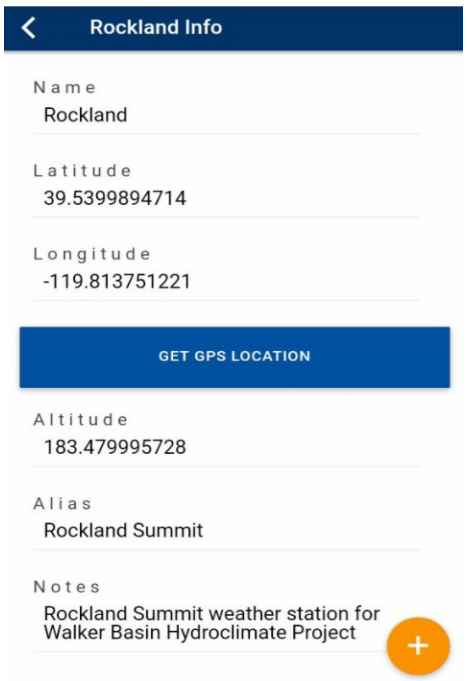

Figure 6: An example of retrieving latitude and longitude coordinates from the phone's GPS

6 Discussion

6.1 NRDC Impact

The successful implementation of the QA application changes the face of quality assurance in the field significantly for existing projects in the NRDC. The inclusion of a dedicated application impacts the workflow of sensor technicians and researchers by substantially augmenting current data management capabilities. Data stewards performing QA on sensor networks benefit from this application in several ways over traditional methods: uniform data entry, centralized QA data storage with synchronization and a usable interface facilitating the utility of the above benefits.

The problem of uniform an accurate data entry naturally occurs in any system reliant upon human interaction as the primary interface between a means of measurement and the means of logging. This problem is further exacerbated when technicians are deployed to remote areas, often equipped with only a notebook. It can be very hard to meaningfully restrict metadata and service logging, as different people are going to include different data that they find relevant to a QA expedition. The use of form fields significantly normalizes data input by restricting users to only give information deemed necessary and sufficient to detail the quality assurance practices performed. An example of these forms can be seen in Figure 7. In the case of intrinsically non-structured data, the option to attach documents is provided. This enables a diversity of data input methods.

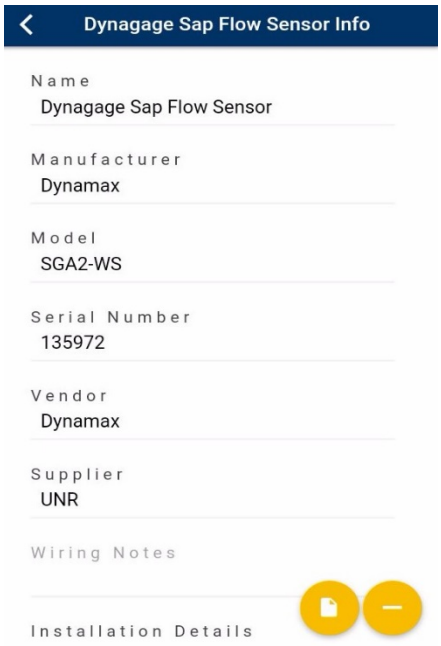

Figure 7: An example data entry page containing info about a single data sensor. The floating action button in the lower right-hand corner provides the option to add a service entry when clicked.

Previously, QA-relevant metadata collected and maintained by technicians was held in a decentralized heterogeneous collection of notebooks, spreadsheets and program comments. This proved problematic internally, as audits and reviews of quality assurance processes could not be effectively performed in a timely manner. Externally, this lack of centralized provenance-tracking metadata damages the integrity of collected data, as logs are not comprehensively tied to data

streams. This limits the re-usability of collected data. With a central repository and dedicated backend infrastructure, the QA app significantly improves the maintenance of QA and metadata logs by providing a centralized, organized database to store this information and bind it to existing data streams. With the frontend mobile component automatically syncing with remote servers, users need not worry about any logistics of storing and formatting QA data for future use. They now only need to perform their normal maintenance and installation practices and fill out the form fields detailing their work.

6.2 Broader Impact

It has been well documented that an issue of paramount importance to the greater scientific community is the need for collected data to be accessible, findable, inter-operable and reusable. [9, 20-21] As scientific research has become increasingly collaborative with the growth of internet technologies, the free and easy distribution of data across the internet has not kept pace. [2, 17, 19] While there are many factors that contribute to this data need/accessibility gap, one prominent problem is a lack of identifying metadata accompanying datasets. [17, 21] That is the goal of this application and we believe that a need still exists in the wider scientific community for the easy creation and upload of identifying metadata, so a survey was created to demonstrate this.

To provide more significant validation and show the broader applicability of this software to the national scientific community a survey was conceived and solicited to Environmental Scientists. The survey was comprised of several questions that sought to evaluate how findable, accessible, interoperable, and reusable respondent's datasets were. Most questions were bound by a Likert scale of 1 to 5.

The pool of potential respondents came from two working groups – called "Clusters" -- for the collaborative organization Environmental Science Information Partners (ESIP). Specifically, members of the Envirosensing Cluster and the Documentation Cluster were emailed with the survey. After hosting the survey for one week 12 responses were acquired.

Using the responses collected from these domain experts we made some preliminary conclusions about how data scientists are managing their metadata. Primarily, we observed that there still exists a strong need for software like the Quality Assurance app to automate and speedup metadata documentation processes. Secondarily, we observed that although metadata was being digitized, it was not being digitized effectively.

Concerning the first conclusion, a few questions demonstrated the current gap in technology that limits effective metadata collection. When asked how managers document their metadata, the majority of respondents replied with either "Field Notebook/ Pen and Paper" and "Comments on Data Logger Scripts". When we juxtapose these responses against a later question that asks people to estimate the length of time it takes to get metadata into a machine-readable format, a picture emerges showing that scientists are collecting and digitizing this data but not in the most effective way. Upwards of six respondents indicated that it takes them "hours" and up to "weeks" to digitize their metadata.

When scientists are extensively using program comments and field journals to collect their data, they are not leveraging the advantages of mobile technology has to offer in the internet age. By comparison, the Quality Assurance app enables the fast and easy input of formatted data which is digital from the start and immediately uploaded to a central database. This application could save many data scientists and technicians hundreds of hours over the course of a year which are now wasted on basic data input.

With this survey and the extensive background works that clearly demonstrate a need to this sort of automated workflow, we have demonstrated that a strong need still exists for software like the QA app presented in this paper.

7 Future Work

Work on the Quality Assurance application will continue in the interest of expanding the present functionality detailed. First and foremost, further work will be done to help audits and administration of QA practices. Presently actions performed on the application are primarily user agnostic. They are performed with no considerations or limitations based upon the present user of the app. This can be problematic when it is necessary to track down the specific user who performed a given preventative maintenance on a given piece of equipment. A paper trail can prove immensely useful to any project on the scale of those which the NRDC helps maintain.

Outside of internal growth of the application itself, the data collected and maintained by the Quality Assurance application will be used to provide increased functionality to a companion quality control web application. QA and QC are often referred as nearly the same entity in discussions of data management. Where QA is concerned with ensuring that data streams have little opportunity to fail in their logging through constant maintenance and monitoring, QC is concerned with handling data that has been logged in error and attempting to correct those mistakes. The data stored via this QA app can be used to give context to any errors that might be discovered by an automated quality control service.

Acknowledgement

This material is based in part upon work supported by The National Science Foundation under grant number IIA-131726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Derik Barseghian, Ilkay Altintas, Matthew B. Jones, Daniel Crawl, Nathan Potter, James Gallagher, Peter Cornillon, Mark Schildhauer, Elizabeth T. Borer, Eric W. Seabloom, and Parviez R. Hosseini, "Workflows and

Extensions to the Kepler Scientific Workflow System to Support Environmental Sensor Data Access and Analysis," *Ecological Informatics* 5, DOI: http://dx.doi.org/10.1016/ j.ecoinf.2009.08.008, 1:42-50, 2010.

[2] Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan, "Research Objects: Towards Exchange and Reuse of Digital Knowledge," *Nature Poceedings*, DOI: http://dx.doi.org/10.1038/ npre.2010.4626, February 2010.

[3] W. Bishop and T. H. Grubesic, "Metadata," *Geographic Information*, Springer Geography, Springer, Cham, 2016.

[4] John L. Campbell, Lindsey E. Rustand, John H. Porter, Jeffery R. Taylor, Ethan W. Dereszynski, James B. Shanley, Corinna Gries, Donald L. Henshaw, Mary E. Martin, and Wade M. Sheldon, "Quantity is Nothing Without Quality," *BioScience*, 63(7):574-585, 2013.

[5] ESIP, "Federation of Earth Science Information Partners," http://wiki.esipfed.org/index.php/Sensor_Data_Quality# Quality_Control_ .28QC.29_on_data_streams (Accessed on January 12, 2018).

[6] Google, "Angularjs," https://angularjs.org/, Last accessed June 13, 2017.

[7] Sean Gordon, and Ted Habermann, "The Influence of Community Recommendations on Metadata Completeness," *Ecological Informatics* 43:38-51, 2018.

[8] Ionic, "Ionic," https://ionicframework.com/, Last accessed June 13, 2017.

[9] John Kratz and Carly Strasser, "Data Publication Consensus and Controversies," *F1000 Research*, DOI: http://dx.doi.org/10.12688/f1000research,3979.2, pp.1-21, October 2014.

[10] V. D. Le, M. M. Neff, R. V. Stewart, R. Kelley, E. Fritzinger, S. M. Dascalu, and F. C. Harris, "Microservice-Based Architecture for the NRDC," *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, pp. 1659-1664, July 2015.

[11] Michael J. McMahon, Frederick C. Harris, Sergiu M. Dascalu, and Scotty Strachan, "S.E.N.S.O.R. Applying Modern Software and Data Management Practices to Climate Research," 2011.

[12] William K. Michener, "Meta-Information Concepts for Ecological Data Management," *Ecological Informatics* 1, DOI: http://dx.doi.org/10.1016/ j.ecoinf.2005.08.004, 1:3-7, January 2006.

[13] William K. Michener, "Quality Assurance and Quality Control (QA/QC)," *Ecological Informatics*, Springer, Cham, pp. 55-70, 2018.

[14] NEXUS, "Solar Energy Water Nexus," https://solar nexus.epscorspo.nevada.edu/, Last accessed June 13, 2017.

[15] NRDC, "Nevada Research Data Center," http:// sensor.nevada.edu/NRDC/, Last accessed June 13, 2017.

[16] NSHE, "Epscor Nevada," https://epscorspo. nevada.edu, Last accessed June 13, 2017.

[17] Dominique G. Roche, Loeske E.B. Kruuk, Robert Lanfear, and Sandra A. Binning, "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13, DOI: http://dx.doi.org/ 10.1371/jour nal.pbio.1002295, 11:1-12, November 2015.

[18] Carly A. Strasser and Stephanie E. Hampton, "The Fractured Lab Notebook: Undergraduates and Ecological Data Management Training in the United States," *Ecosphere* 3, 12:1-18, 2012.

[19] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE*6, DOI: http://dx.doi.org/10.1371/journal.pone.0021101, 6:1-21, June 2011.

[20] Michael C. Whitlock, "Data Archiving in Ecology and Evolution: Best Practices," *Trends in Ecology & Evolution* 26, DOI: http://dx.doi.org/10.1016/ j.tree.2010.11.006, 2:61–65, November 2011.

[21] Mark D. Wilkinson, Micheal Dumontier, IJsbrad Jan Aalbersberg, Gabrielle Appleton, Myles Axton, and Arie Baak, "The Fair Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, 2016.

[22] Matthew Williams, Dan Cornford, Lucy Bastin, Richard Jones, and Stephen Parker, "Automatic Processing, Quality Assurance and Serving of Real-Time Weather Data," *Computers & Geosciences 37*, DOI: http://dx.doi.org/10.1016/j.cageo.2010.05.010, 3:353-362, March 2011.

Connor Scully-Allison received his B.A. in Philosophy in 2012 from the University of Nevada, Reno (UNR). Accepted into the master's program at UNR for Computer Science and Engineering in 2015, he is currently working as a research assistant on the Track 1 Nexus Project for the Cyber-Infrastructure lab located in the College of Engineering. His research interests include Human Computer Interaction, High Performance Computing, and Software Engineering. He has published two conference papers since 2016. As of June 2018, Connor holds a position as a student fellow for the Earth Science Information Partners (ESIP) Organization.

Hannah Muñoz graduated in 2016 with a B.S. in Computer Science from the University of Nevada, Reno. She started on her journey to her M.S in Computer Science with her advisers Dr. Sergiu Dascalu and Dr. Frederick Harris, Jr. in 2017. She hopes to finish her degree in 2018. Currently, Hannah works in the Cyber-infrastructure Lab where she helps develop applications that enable earth scientists to further their research.

Hannah has written and published two conference papers during her time as a Master's student. Her interest lies in mobile development and developing software to help in scientific analyses.

Vinh Le graduated from the University, Reno with a B.S in Computer Science and Engineering in 2015. Vinh is a Graduate Research Assistant affiliated with the Cyber Infrastructure Lab at the University of Nevada, Reno. He currently aims to earn a Master of Science in Computer Science and Engineering by 2018 and his research interests consist primarily of Software Engineering, Internet Architecture, and Human-Computer Interaction.

Scotty Strachan is the Director of Cyberinfrastructure in the Office of Information Technology at the University of Nevada, Reno. Strachan graduated from the University of Nevada, Reno in 2001 with a bachelor's degree in geography and minor in economics. After spending some additional years as a geotechnical consultant and project manager, he returned to the University and completed a M.S. and Ph.D., both in geography, along with a graduate minor in business administration. Strachan's primary research interests lie in mountain ecosystems and observational networks, and he relies heavily on the integration of information technologies with research to accomplish his goals of producing useful, long-term science.

Eric R. Fritzinger received his B.S. (2003) and M.S. (2006) from the University of Nevada, Reno. After spending several years in the field of medical robotics, he returned to UNR to participate in a state-wide project studying the effects of climate change in the Great Basin. He has worked on model and data interoperability as well as management and organization of environmental sensor data. He is currently the lead developer for the Nevada Research Data Center (NRDC), based out of UNR's Computer Science and Engineering Department.

Frederick C. Harris Jr. received his BS and MS degrees in Mathematics and Educational Administration from Bob Jones University, Greenville, SC, USA in 1986 and 1988 respectively. He then went on and received his MS and Ph.D. degrees in Computer Science from Clemson University, Clemson, SC, USA in 1991 and 1994, respectively.

He is currently a Professor in the Department of Computer Science and Engineering and the Director of the High-Performance Computation and Visualization Lab at the University of Nevada, Reno, USA. He has published more than 200 peer-reviewed journal and conference papers along with several book chapters. His research interests are in parallel computation, computational neuroscience, computer graphics, and virtual reality.

He is also a Senior Member of the ACM, and a Senior Member of the International Society for Computers and their Applications (ISCA).

Sergiu Dascalu is a Professor in the Department of Computer Science and Engineering at the University of Nevada, Reno, USA, which he joined in 2002. In 1982 he received a Master's degree in Automatic Control and Computers from the Polytechnic University of Bucharest, Romania and in 2001 a Ph.D. in Computer Science from Dalhousie University, Halifax, NS, Canada. His main research interests are in the areas of software engineering and human–computer interaction. He has published over 180 peer-reviewed papers and has been involved in numerous projects funded by industrial companies as well as federal agencies such as NSF, NASA, and ONR.

Modeling End User Empowerment in Big Data Analysis and Information Visualization Applications

Marco X. Bornschlegl*, Kevin Berwind*, and Matthias L. Hemmje*
University of Hagen, 58097 Hagen, Germany

Abstract

Handling the complexity of relevant data (generated through information deluge and being targeted with Big Data technologies) requires new techniques with regard to data access, visualization, perception, and interaction for supporting innovative and successful information, informed decision making, and business strategies. As a response to increased graphics performance in computing technologies and Information Visualization, the Information Visualization Reference Model was developed many years ago. Since then and due to further developments in Information Systems as well as Data Analysis and Knowledge Management Systems in recent years, this model has been adapted for addressing the recent advancements. Thus, the hybridly refined and extended IVIS4BigData Reference Model was recently derived from the original model to cover the new conditions of the represent situations with advanced visual interfaces providing opportunities for perceiving, managing, and interpreting Big Data analysis results to support insight and emerging knowledge generation. After deriving and qualitatively evaluating the conceptual IVIS4BigData model, a set of conceptual end user empowering use cases for each IVIS4BigData processing stage will be modeled in this paper. These conceptual use cases, that contain all activities of their respective IVIS4BigData process stage, serve as a base for a functional, i.e. conceptual as well as technical IVIS4BigData system specification supporting end users, domain experts, as well as for software architects in utilizing IVIS4BigData.

Key Words: IVIS4BigData, advanced visual user interfaces, distributed big data analysis, information visualization, end user empowerment.

1 Introduction and Motivation

As a response to increased graphics performance in computing technologies and information visualization, Card et al. [10] developed the information visualization (IVIS) Reference Model. Due to further developments in Information Systems as well Data Analysis and Knowledge Management Systems in recent years, this reference model has been adapted for covering the recent advancements. Therefore, Bornschlegl et al. perfomed the Road Mapping of Infrastructures for Advanced Visual Interfaces Supporting Big Data workshop [7, 8, 5]. Industrial researchers and practitioners working in the area of Big Data, Visual Analytics, and Information Visualization were invited to discuss and validate future visions of Advanced Visual Interface infrastructures supporting Big Data applications in Virtual Research Environments. Within that context, the IVIS4BigData Reference Model, that is illustrated in Figure 1, was presented [5] and qualitatively evaluated [4] within the road mapping activity [6].

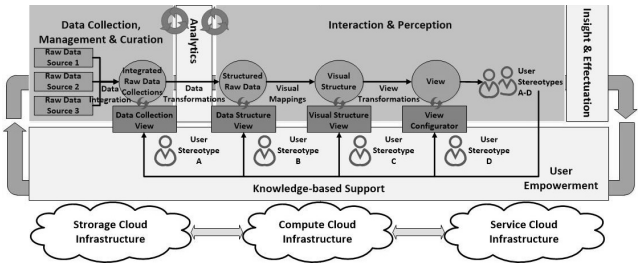

Figure 1: IVIS4BigData Reference Model [5]

However, as *"users are increasingly willing and, indeed, determined to shape a software they use to tailor it to their own needs"* [2], the model has to be enriched with so-called user empowerment features which enable system configuration without programming or reprogramming [2, 12]. Thus, this research will address these issues with a special focus on modeling end user empowerment to support distributed Big Data analysis based on the IVIS4BigData Reference Model.

In summary, user empowerment within human computer interaction based on the IVIS4BigData Reference Model and the existing state-of-the-art in the relevant areas of computer science as well as established open ICT standards still has to be achieved. In the remainder of this paper, we are going to address this challenge by means of modeling corresponding conceptual use cases and user empowerment support features based on the IVIS4BigData Reference Model.

*Faculty of Mathematics and Computer Science, EMail: (marco-xaver.bornschlegl, kevin.berwind, matthias.hemmje) @fernuni-hagen.de

2 Relevant State of the Art

Into IVIS4BigData, the IVIS pipeline is segmented in a series of data transformations. Furthermore, due to the direct manipulative interaction between different user stereotypes within the single process stages and their adjustments and configurations of the respective transformations by means of user-operated controls, each segment in the IVIS4BigData pipeline needs to support an interactive user empowerment, i.e., system configuration workflow allowing to configure the transformations and visualizations in the different phases.

For this reason, the IVIS pipeline will be divided into four consecutive process stages that empower end users to define, configure, simulate, optimize, and run each user empowered phase of the pipeline in an interactive way. Arranged in the sequence of the IVIS pipeline, each process stage contains all actions of its interactive and user empowered transformation configuration workflow between the two IVIS phases. Starting from raw data on the left side (Figure 1), the four consecutive IVIS4BigData process stages, where each stage represents a certain IVIS4BigData transformation (*Data Integration*, *Data Transformation*, *Visual Mapping*, and *View Transformation*), are defined the following way:

- Data Collection, Management, and Curation: Harmonization and Semantic Integration of individual, distributed, and heterogeneous raw data sources into a uniform schema (Data Integration from local source schemata into a global integrated mediator schema).
- Analytics: Distributed and cloud-based Big Data analysis of the integrated raw data (Data Transformation).
- Visualization: Definition and creation of a visualization based on the structured data (Visual Mapping).
- Perception and Effectuation: Facilitation of the interaction with appropriate views of the generated visual structures to enable suitable interpretations of the Big Data analysis results (View Transformation).

2.1 Architectural Concepts of End User Empowerment

For enabling end-users *"to articulate incrementally the task at hand"* [14], *"the information provided in response to their problem-solving activities based on partial specifications and constructions must assist users to refine the definition of their problem"* [14]. To realize this interaction Figure 2 represents the different types of interaction with functional arrows between the cross-functional Knowledge-Based Support layer and the corresponding layers above as introduced in Fischers and Nakakojis [14] multifaceted architecture. This means, user empowerment interactions and features as illustrated in Figure 2, are utilized to derive a functional conceptual use case framework for the resulting user empowered configuration workflow use cases of the different IVIS4BigData process stages.

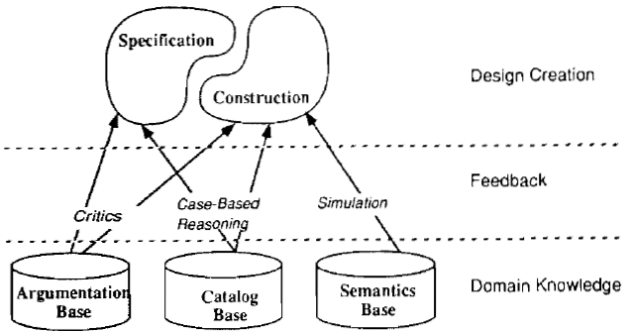

Figure 2: Elements of a multifaceted architecture [14]

In this user empowerment architecture model, five central elements (specification, construction, argumentation base, catalog base, and semantics base) can be identified, *"that assist end users to refine the definition of their problem in their problem-solving activities based on partial specifications and constructions"* [14].

Derived from an end users configuration perspective, a domain independent problem solving, i.e., user interface configuration process can be divided in three layers. The ***Design Creation*** layer contains the construction and specification components that represent the interactive part of this process utilizing the three static components argumentation base, catalog base, and semantics base within the lowest **Domain Knowledge** layer. Moreover, the **Feedback** layer in the middle of this architecture represents the interactive user actions (critics, case-based reasoning, and simulation), that are initiated during the specification or construction process. In addition to this architectural illustration and to emphasize the importance of the construction and specification elements, Fischer and Nakakoji defined a process based illustration of the whole design process, that is outlined in Figure 3.

In this process, *"starting with a vague design goal, designers go back and forth between the components in the environment"* [14]. Thus, *"a designer and the system cooperatively evolve a specification and a construction incrementally by utilizing the available information in an argumentation component and a catalog and feedback from a simulation component"*

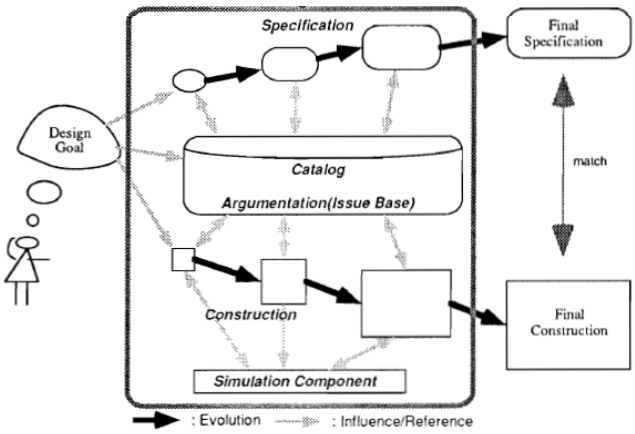

Figure 3: Construction and specification of design in multifaceted architecture [14]

[14]. As a result, a matching pair of specification and construction is the outcome.

2.2 Architectural Modeling Approach

Based on the described user empowerment principles and architectural features, a generic conceptual use case framework is now defined. It covers all activities within an IVIS4BigData process stage. As outlined in Figure 4, this framework will be the context for the design of corresponding IVIS4BigData conceptual use cases including user empowerment.

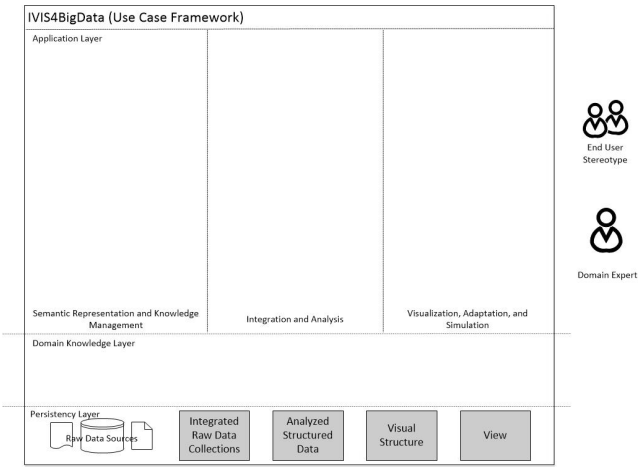

Figure 4: IVIS4BigData use case framework based on IVIS4BigData and end user empowerment architectural tiers

From a vertical perspective, the conceptual user empowerment use case framework consists of three layers according to Fischer and Nakakoji [14]. Although it contains all elements of their multifaceted architecture, the layers differ from the constitutive IVIS4BigData model that is to be extended with user empowerment. Whereas the middle **Domain Knowledge** layer accord to the original model, Fischer and Nakakoji defined a separate layer especially for feedback. Not in alignment to the original model, this framework utilizes the **Application** layer, that represents Fischer's and Nakakoji's Design Creation layer, also for feedback and adaptation. Moreover, to cover the new situation of distributed processing, an additional **Persistency** layer is defined underneath both layers.

In more detail, the Application layer, where all design specification and construction functions of the IVIS4BigData process stage workflow are located as well as the functions to run predefined workflows, is separated into three areas. Whereas the **Semantic Representation and Knowledge Management** area of this layer contains all activities to configure the data, analysis, or visualization workflow models depending on the IVIS4BigData process stage, the central **Integration and Analysis** area includes all functions to execute the resulting interactive workflow of each stage. This central area also includes the central transformation of each IVIS4BigData process stage (*Data Integration*, *Data Transformation*, *Visual Mapping*, or *View Transformation*) within the use cases. Finally, the **Visualization, Adaptation, and Simulation** area includes all activities to support the construction and specification process. In combination with the **End User Stereotype** and the **Domain Expert**, where both of them interacting with the activities within these three areas, the Workflow layer represents the central element of this framework. In this way, this central layer combines all Design Creation and Feedback activities of Fischer's and Nakakoji's Multifaceted Arch. in a more detailed and effective way.

The Domain Knowledge layer in the middle of this framework contains all data, analysis, or visualization catalogs and other types of domain knowledge that can be utilized within the IVIS4BigData process stages. In this way, this layer, based on its elements, will support the activities of the central application layer with regard to the respective IVIS4BigData transformation of each process stage.

Finally, the Persistency layer contains the data sources as well as the phases of the IVIS4BigData pipeline and is responsible for the data persistence during the interactive transformation between two consecutive IVIS4BigData phases. As a result of the ability to utilize distributed architectures and cloud services for data storage as well as for data processing, this layer emphasizes the importance to manage and control the data during and after an IVIS4BigData process stage.

3 Modeling Conceptual End User Empowering Use Cases

After deriving the IVIS4BigData process structure including the description of their interactive configuration workflow objectives between each two IVIS phases and the definition of the interactive configuration use case framework, the gap between the architectural and the functional mapping of the process stages still exists. To close this gap and to derive a set of conceptual user empowerment use case for each IVIS4BigData process stage, each process stage is considered from a functional perspective. These conceptual configuration use cases, that contain all configuration activities of their respective process stage, serve as a base and a functional system description for end users, domain experts, as well as for software architects utilizing IVIS4BigData.

3.1 Configuration of the Data Collection, Management, and Curation Phase

The first conceptual configuration use case in the sequence of the IVIS-4BigData process stages, describes functions that can configure how to integrate distributed and heterogeneous raw data sources into an uniform raw data collection, by means of Semantic Integration. Thus, as illustrated in Figure 5, several functions are provided to facilitate the main configuration functionality of this use case and the first IVIS4BigData transformation **Data Integration**.

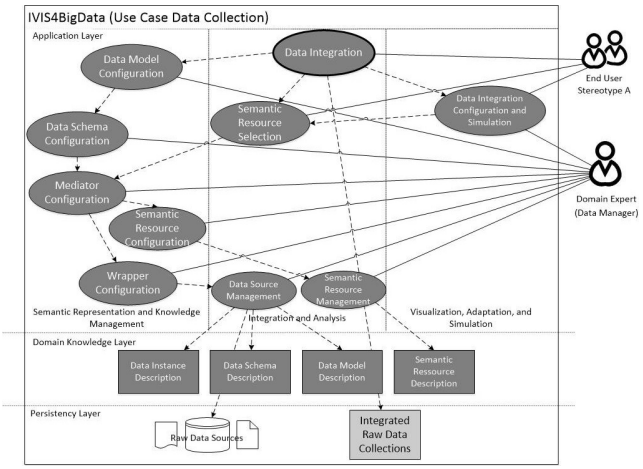

Figure 5: Configuration support in IVIS4BigData use case data collection, management, and curation

Starting from raw, already processed, or stored data of previous IVIS4BigData process iterations, the configuration functions of the application layer empower end users to select individual data sources as well as the semantic representation of the entire data of the connected data sources to design, configure, and finally create and manage integrated data sets as an intermediate result and preliminary input for the next consecutive use case in the next IVIS4BigData process phase. Within the application layer, beginning with the distributed and heterogeneous data sources, the **Data Schema Description**, and the **Data Model Description** within the Domain Knowledge Layer.

Whereas the Data Instance Description provides information about technical attributes (like e.g. data type, data host address, data port, data source log-in information, or supported communication protocol) for the physical connection, the Data Schema Description contains information about the data structure (like e.g. table names, columns, or property lists) and the Data Model Description contains information about data content (like e.g. data model, representation, syntactical relationships, or constraints) for logical connections of data sources. The Data Source Management function is located at the lowest level within this use case to emphasize that there is no relation to raw data within the data sources at this level of abstraction although this function accesses information of the Domain Knowledge Layer and provides it to the subsequent functions. Based on this base functionality, hereafter the process of the Data Integration is segmented in a technical and a logical path.

From a technical perspective, the **Wrapper Configuration** function, located at the Semantic Representation and Knowledge Management area, provides access to the data of data sources by exporting some relevant information about their schema, data and query processing capabilities. Moreover, the **Mediator Configuration** function represents the second step of the technical path of the data integration. In addition to the functionality of the Mediator Configuration, the defined mediator combines both paths to a resulting logical path.

To *"exploit encoded knowledge about certain sets or subsets of data to create information for a higher layer of applications"* [22], and to store the data provided by the wrappers in an unified view of all available data with central data dictionary by the utilization of Semantic Integration, the mediator relies on the information of the logical function **Semantic Resource Configuration**. It configures the data sources from a semantic perspective with focus on their logical content based on the available semantic resources, which can be configured within the **Semantic Resource Management** function. For the management of the semantic resources, this function relies on the **Semantic Resource Description** within the Domain Knowledge Layer, providing information (like e.g. full text information or other type of metadata) about the content of the connected raw data

sources. Based on the unified views of all available source data within the mediator, the **Data Schema Configuration** and the **Data Model Configuration** function consider the data from a target perspective. Whereas the Data Schema Configuration function aim of specifying the data structure and type, the Data Model Configuration function focuses on the definition of the data model from a content perspective of the resulting integrated data.

Before focusing on the functions for the end users, that are responsible for the execution of this IVIS4BigData workflow, the **Data Integration Configuration and Simulation** function at the Visualization, Adaptation and Simulation area enables domain experts as well as end users to configure and simulate individual raw data integration workflows as well as to store the workflows for the essential data integration depending on the raw data sources data and the data integration purpose. Based on the functions with enhanced capabilities, where domain experts are able to configure technical details of the connected data sources as well as of their semantic representation and the resulting data model within IVIS4BigData, end users are empowered to select and integrate the data of connected heterogeneous and distributed raw data sources. With the function **Semantic Resource Selection** end users are able to select data sources with their semantic representation of their respective data. Finally, the function **Data Integration**, that represents the first transformation of the IVIS4BigData pipeline, integrates the data by utilizing the configured and stored data integration workflows and provides the resulting integrated raw data set to the the **Integrated Raw Data Collection** process phase at the Persistency Layer.

3.2 Configuration of the Analytics Phase

The conceptual configuration use case for configuring the Analytics phase of IVIS4BigData, that is illustrated in Figure 6 and is located at the second position in the sequence of the IVIS4BigData process stages describes functions to end users as well as to domain experts to facilitate the essential technical **Big Data Analysis**. This main functionality represents the second IVIS4BigData transformation and transforms the integrated and unstructured raw data in analyzed structured data.

Starting from the integrated raw data of the heterogeneous and distributed raw data sources, the functions of the application layer empower end users to select unstructured raw data sets, configure and simulate Big Data analysis workflows, execute the configured workflows, and export the resulting analyzed and structured data for the consecutive use case. Before focusing on

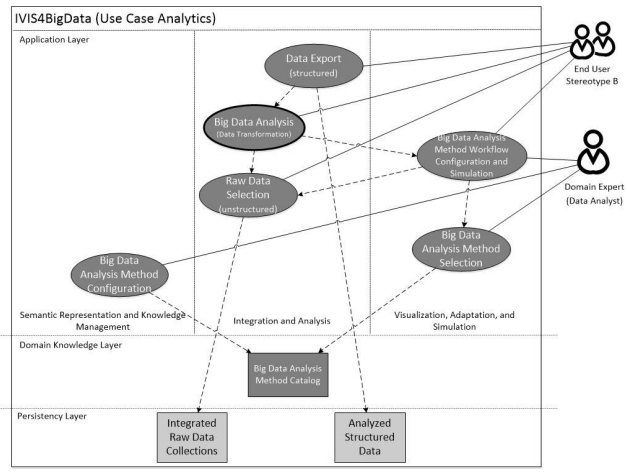

Figure 6: Configuring the conceptual IVIS4BigData use case analytics

the functions for the end users, that are responsible for the execution of this IVIS4BigData workflow, two central configuration functions for domain experts are considered within the application layer at first. Starting with the **Big Data Analysis Method Configuration** function at the Semantic Representation and Knowledge Management area, that resorts to the **Big Data Analysis Method Catalog** within the Domain Knowledge layer, this function enables domain experts to configure Big Data analysis algorithms for the utilization in IVIS4BigData. Afterwards, these algorithms can be selected by the **Big Data Analysis Method Selection** function at the Visualization, Adaptation, and Simulation area, relying on configured algorithms within the Big Data Analysis Method Catalog. The last function at the Visualization, Adaptation, and Simulation area **Big Data Analysis Method Workflow Configuration and Simulation** enables domain experts as well as end users to configure and simulate individual Big Data analysis workflows as well as to store the workflows for the essential analysis depending on the source data and the analysis purpose.

Thus, after configuration and simulation of analysis algorithms and methods, the end users are empowered to perform their Big Data analysis with the aid of three functions at the Integration and Analysis area. First of all, the function **Raw Data Selection** enables the selection of the integrated but unstructured data of the heterogeneous and distributed raw data sources. Afterwards, the main function **Big Data Analysis**, that represents the second transformation of the IVIS4BigData pipeline, utilizes the configured and stored analysis workflows to transform the unstructured data to structured data. Finally, the **Data Export** function provides the resulting structured data to the analyzed **Structured Data** process phase at the Persistency Layer for the consecutive use case.

3.3 Configuration of the Visualization Phase

The third conceptual configuration use case in the sequence of the IVIS4BigData process stages describes functions to transform analyzed and structured data into visual structures. As illustrated in Figure 7, several functions are provided to end users as well as to domain experts to facilitate the main functionality of this use case and the third IVIS4BigData transformation **Visualization**.

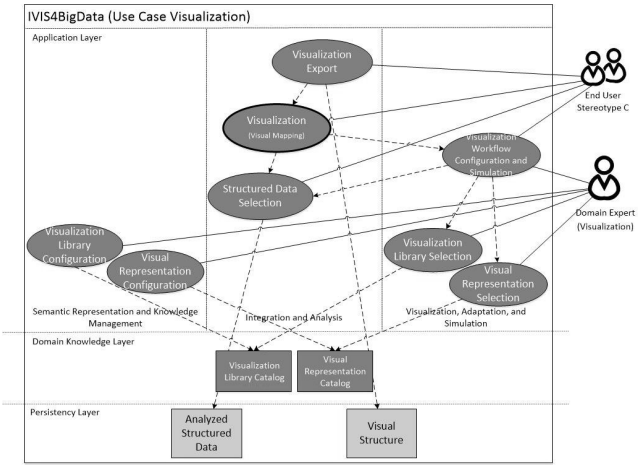

Figure 7: Configuring the conceptual use case visualization in IVIS4BigData

Starting from the structured and analyzed data of the heterogeneous and distributed data sources, the functions of the application Layer empower end users to select structured data sets, configure and simulate Big Data visualization workflows, execute the configured workflows, and export the resulting visual structure for the consecutive use case. Similar to the previous configuration use case supporting the Analytics phase in IVIS4BigData, several central configuration functions for domain experts are considered within the Application layer.

Starting with the **Visual Representation Configuration** function in the Semantic Representation and Knowledge Management area, that resort to the **Visual Representation Catalog** within the Domain Knowledge layer, this function enables domain experts to configure suitable visual representations (like e.g. Word Cloud, Tree Map, Sunburst Chart, Choropleth Map or Small Multiples) depending on the respective data structure within the analyzed and structured data process stage. Moreover, the **Visualization Library Configuration** function based on using the **Visualization Library Catalog**, enables domain experts to configure visualization libraries (like e.g. D3.js, Charts.js, Dygraphs, or Google Charts) for utilization in IVIS4BigData.

Afterwards, visual representations as well as the visualization libraries can be selected by the **Visual Representation Selection** and **Visualization Library Selection** functions at the Visualization, Adaptation and Simulation area, by making use of configured visual representations and visualization libraries within the catalogs. The last function in the Visualization, Adaptation and Simulation area is **Visualization Workflow Configuration and Simulation**. This function enables domain experts as well as end users to configure and simulate individual Big Data visualization workflows as well as to store the workflows for the essential Big Data visualization depending on analyzed and structured source data as well as on the visualization and analysis purpose.

After the configuration and simulation of the visualization methods, the end users are empowered to perform their Big Data visualization with the aid of three functions in the Integration and Analysis area. First, the function **Structured Data Selection** enables the selection of integrated and structured raw data in the heterogeneous and distributed data sources. Afterwards, the main function **Visualization**, that represents the third transformation of the IVIS4BigData pipeline, utilizes configured and stored visualization workflows to transform analyzed and structured data to visual structures. Finally, the **Data Export** function provides resulting visual structures to the **Visual Structure** process phase at the Persistency Layer for the consecutive use case.

3.4 Configuration of the Perception and Effectuation Phase

The final use case Perception and Effectuation, which is illustrated in Figure 8, is located at the fourth position in the sequence of the IVIS4BigData process stages. This use case describes functions for end users as well as for domain experts to facilitate the generation of suitable views. The main functionality **View Transformation**, that represents the fourth IVIS4BigData transformation, transforms the visual structure into interactive views, whereby end users are empowered to interact with analyzed and visualized data of the heterogeneous and distributed raw data sources for perceiving, managing, and interpreting Big Data Analysis results to support insight.

Starting from integrated, analyzed and visualized raw data of the heterogeneous and distributed data sources, the configuration functions of the Application Layer empower end users to select visualized data sets, generate suitable views, and interact with visualized data of the heterogeneous and distributed raw data sources. As well as the previous configuration use cases and before focusing on the functions for the

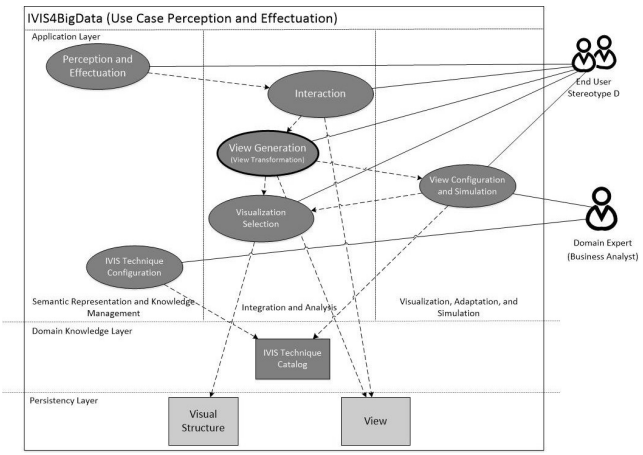

Figure 8: Configuration of the conceptual use case perception and effectuation in IVIS4BigData

end users, that are responsible for the execution of this IVIS4BigData workflow, one central configuration function for domain experts is considered within the Application layer at first. This **IVIS Technique Configuration** function in the Semantic Representation and Knowledge Management area, that uses the **IVIS Technique Catalog** within the Domain Knowledge layer, enables domain experts to configure visualization technologies (like e.g. SVG, VRML, or X3D) for the utilization in IVIS4BigData.

The **View Configuration and Simulation** function within the Visualization, Adaptation, and Simulation area enables domain experts as well as end users to configure and simulate individual and suitable views as well as to store the views for the essential interaction and perception of the visualized data depending on the analysis purpose. Thus, after the configuration and simulation of the visualization technology, end users are empowered to perform their interaction and perception with the aid of three functions within the Integration and Analysis area. First, the function **Visualization Selection** enables the selection of the visual representation of the integrated and analyzed heterogeneous and distributed data sources. Second, the main function **View Generation**, that represents the fourth transformation of the IVIS4BigData pipeline, utilizes the configured and stored views to transform the visual structure to an interactive view. Third, the **Interaction** function enables end users to perceive, manage, and interpret Big Data Analysis results to support insight.

Finally, with the aid of the **Perception and Effectuation** function within the Semantic Representation and Knowledge Management area, the emergent knowledge process, which is symbolized by the outer loop of IVIS4BigData, can be achieved by actively managing the insights created by effectuating data and integrating these effects into the knowledge base of the analysis process [5].

4 Qualitative Alignment Evaluation, Discussion, Conclusion and Outlook

As presented and described within this paper, IVIS4BigData represents an adaptation of the original ***IVIS Reference Model*** [9], in combination with Kaufmann's ***BDM Reference Model*** [16] to cover the new conditions of the present situation with distributed and heterogeneous data sources, cloud computing technologies as well as modern advanced visual user interface opportunities for perceiving, managing, and interpreting Big Data analysis results to support insight. Nevertheless, IVIS4BigData aligns with other scientific reference models. This alignment will be qualitatively evaluated and discussed within this section from a horizontal perspective on IVIS4BigData, with focus on the evolution from raw data sources on the left side to the human perceiver on the right side, that represents the maturity level of the technical Big Data analysis. Moreover, the qualitative evaluation regarding alignment with existing models is also discussed from a vertical perspective on IVIS4BigData, where multiple visually-interactive user interface views enables a direct manipulative interaction between end user stereotypes with single process stages and the adjustments of the respective IVIS4BigData transformations by user operated user interface controls, that represents the incremental evolution from data to wisdom between each process stage.

4.1 Qualitative Alignment with Big Data Analysis Process Models

From an abstract and non technical perspective with focus on Big Data analysis from distributed and heterogeneous raw data sources on the left side to human perception on the right side, IVIS4BigData aligns with Ackoff's ***DIKW Hierarchy*** [1] and Bellinger et al.'s processed-based illustration of this hierarchy [3]. Both models, that describes the transformation process from data to wisdom *"as a content of the human mind"* [1, 20] instead from an information systems perspective, outline the evolution from data to wisdom as a sequential process. Descending from wisdom, which is located at the top of the DIKW Hierarchy, there are understanding, knowledge, information, and, at the bottom, data. *"Each of these single categories include the categories that fall below it, because there can be no wisdom without understanding and no understanding without knowledge"* [1].

The suggestions of both abstract models, *"that moving from data to information requires **understanding relations**, moving from information to knowledge involves **understanding patterns**, and moving from knowledge to wisdom involves **understanding principles**"* [20], align with the transformations of IVIS4BigData, where each transformation (***Data Integration***, ***Data Transformation***, ***Visual Mapping***, and ***View Transformation***) is configured, simulated, and applied on a certain evolution level of the raw source data within the consecutive IVIS4BigData process phases. Moreover, due to the characteristic of the IVIS4BigData Use Case Framework, where certain domain experts, that represent the Data Analyst role, support the end users in the process of specifying and creating the interactive workflow of each IVIS4BigData process stage, IVIS4BigData considered Rowley's [20] principles in a broader way.

The arrangement of this domain expert role within the Use Case Framework illustrates their unique role in the specification and creation process, by supporting the end user with the specific domain knowledge (understanding relations, understanding patterns, and understanding principles) of the respective data transformation and their enhanced capability, enables end users *"to articulate incrementally the task at hand"* [14]. By concluding the first two IVIS4BigData transformations (Data Integration and Data Transformation), because the first transformation has only been added to cover the new conditions of the present situation to integrate distributed and heterogeneous data sources, Table 1 outlines how these three principles of the evolution from data to wisdom within the DIKW process can be assigned to the IVIS4BigData transformations.

Table 1: Assignment of DIKW principles to IVIS4BigData transformations

DKIW Hierarchy	**Rowley's Theses**	**IVIS4BigData Transformations**
Data → Information	Understanding Relations	Data Integration, Data Transformation
Information → Knowledge	Understanding Patterns	Visual Mapping
Knowledge → Wisdom	Understanding Principles	View Transformation

Moreover, IVIS4BigData also aligns with North's DIKW-based ***Knowledge Staircase*** [19], that is building up on both Ackoff's DIKW Hierarchy, as well as with Fayyad et al. ***Knowledge Discovery in Databases (KDD) Process*** [11]. Whereas the Knowledge Staircase is separated in an operational and a strategic Knowledge Management area, the KDD Process focuses only on the operational part of extracting useful information (knowledge) from data [11].

Table 2: Assignment of KDD process phases to IVIS4BigData transformations

KDD Process Phases	**IVIS4BigData Transformations**
Selection, Preprocessing	Data Integration
Transformation	Data Transformation
Data Mining	Visual Mapping
Interpretation / Evaluation	View Transformation

The upper strategic Knowledge Management area aligns to the ***Insight & Effectuation Layer*** of IVIS4BigData where no process steps are currently located because *added value* is rather generated from knowledge, which is a *"function of a particular perspective"* [17] and will be generated within this layer by combining the analysis results with existing knowledge where end users *"learn how to apply this knowledge and convert it into skills and turning their skills in actions, because only actions of persons deliver measurable results for an organization"* [21].

In contrast to the upper part of the Knowledge Staircase, their lower operational part as well as the whole KDD Process align with the technical process stages of IVIS4BigData, even if KDD focuses mainly on Data Mining use cases that have mostly been used by Statisticians or Business Analysts. Due to the characteristics of IVIS4Big-Data, that aims to represent a generic framework for supporting distributed Big Data analysis applications to support emerging knowledge generation in interdisciplinary research collaboration as well as in cross-domain and cross-organizational Business Intelligence and corresponding innovation scenarios, Table 2 describes how the certain process steps of the KDD Process, can be assigned to the data transformations of the introduced IVIS4BigData Reference Model.

In this way, the aim of IVIS4BigData aligns with the unifying goal of the KDD process, to extract high-level knowledge from low-level data in the context of large data sets, *"which are typically to voluminous to understand and digest easily"* [11]. Even if the KDD process focuses on Data Mining and relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data

[11], IVIS4BigData can represent a generic framework where common Big Data frameworks can easily be integrated and utilized in addition to those techniques.

4.2 Qualitative Alignment with Human-Computer Interaction and Knowledge Management Models

In addition to the influence of Fischer's and Nakakoji's [14] multifaceted architecture and with focus on the major adaptations of IVIS4BigData in relation to the original IVIS Reference Model, where multiple visually-interactive user interface Views enable a direct manipulative interaction between end user stereotypes with single process stages and the adjustments of the respective transformations by user-operated user interface controls to provide *"the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way to the 'right' person"* [13], the design of the use case framework was strongly influenced by Nonaka's and Takeuchi's ***SECI Model of Knowledge Dimensions*** [17] as well as by the ***HACE Theorem*** of Wu et al. [23].

From a global perspective, the SECI Model of Knowledge Dimensions shows different modes of knowledge conversion and the evolving spiral movement of knowledge through the SECI process [18]. Thus, this model stresses the importance of repeated conversion of explicit knowledge to tacit knowledge and vice versa to generate new knowledge, and highlights the mutual complementary nature of tacit and explicit knowledge [15]. Moreover, Nonaka and Takeuchi remark that *"it is important to note that the movement through the four process stages of knowledge conversion forms a spiral, not a circle"* [17, 18].

Therefore, IVIS4BigData enables end users to interact with the certain process stages instead of interacting only with the results at the end of the Data Analysis process. Furthermore, end users are able to move to any process stage to view, configure, simulate, or execute each data transformation repeatedly. This characteristic of IVIS4BigData clarifies that not only interacting with the final results of the whole IVIS process stages represents a SECI circle, which is illustrated by the circular iteration around the whole set of layers. It clarifies that IVIS4BigData is not aiming at supporting solely a one time process because the results can be used as the input for a new process iteration. Moreover, as outlined in Figure 9 and in addition to the global SECI circle, each of those functional arrows of IVIS4BigData between the cross-functional Knowledge-Based Support layer and the corresponding layers above, represent individual SECI circles as well.

In this illustration, that outlines the alignment of the SECI Model of Knowledge Dimensions in relation to the

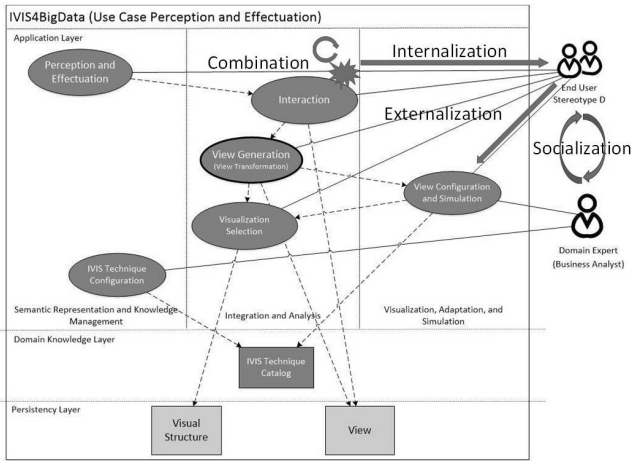

Figure 9: Integration of the SECI model of knowledge dimensions to the IVIS4BigData use case framework

Human-Computer Interaction of IVIS4BigData, SECI circle will be described by means of the Perception and Effectuation use case exemplarily. Starting with the discussion between end user and domain expert ***(Socialization)***, where both individuals try to convert new tacit knowledge by sharing their experiences [17], both end user or domain expert configure and simulate ***(Externalization)*** individual and suitable views for the essential interaction and perception of the visualized data depending on the analysis purpose. With this process, they externalized their tacit knowledge that allows it to be shared by others and it becomes the basis of new knowledge [17]. Afterwards, an end user is either able to perform his analysis purpose with the aid of the functions within the Integration and Analysis area, or to perform the configuration and simulation process once again.

In this example, new knowledge is generated during interacting with the analysis results, that can be converted into more complex and systematic sets of explicit knowledge ***(Combination)*** [17]. Nevertheless, the generation of new knowledge can occur at each activity within an IVIS4BigData process stage, because new knowledge can not be limited to a certain activity. Finally, to complete the SECI circle, the end user embodies the generated explicit knowledge into tacit knowledge ***(Internalization)*** [17] before he is able to start a new SECI process iteration.

Finally and in addition to the SECI Model of knowledge Dimensions, the architecture design of the use case framework is also influenced by ***(HACE Theorem)*** of Wu et al. [23], which describes a multi-tier processing framework to provide end users actions with considerations on data accessing and computing (Tier I), data isolation and domain knowledge (Tier II), and

Big Data mining algorithms (Tier III) to overcome the limitations of singe end users (blind men) to his local region or the part of information he collects during the Big Data analysis process (elephant in this scenario) [23]. In this way and as outlined in Figure 10, each interactive workflow that is described by the generic use case framework which covers all of the activities within an IVIS4BigData process stage, aligns with the multi-tier HACE processing framework.

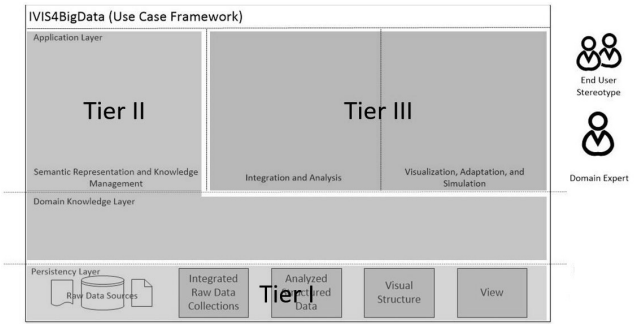

Figure 10: Integration of the HACE multi-tier processing framework to the IVIS4BigData use case framework

As outlined in this illustration, ***(Tier I)*** of the HACE framework, that focuses on data access and arithmetic computing procedures because Big Data is often stored in different distributed locations and data volumes may continuously grow [23], can be assigned to the ***Persistency Layer***, that contains the data sources as well as the phases of the IVIS4BigData pipeline and is responsible for the data persistence during the interactive transformation between two consecutive IVIS4BigData process stages. ***(Tier II)***, that centers around semantics and domain knowledge for different Big Data applications for the simple reason that such information can provide additional benefits to the mining process, can be assigned to the middle ***Domain Knowledge Layer***, that contains all data, analysis, or visualization catalogs and other types of domain knowledge that can be utilized within the IVIS4BigData process stages as well as to the ***Semantic Representation and Knowledge Management*** area within the ***Application Layer*** where all activities to configure the data, analysis, or visualization workflow models, depending on the IVIS4BigData process stage, are located. Finally, ***(Tier III)*** is represented by the ***Visualization, Adaptation, and Simulation*** and ***Integration and Analysis*** areas where all activities to support the construction and specification process as well as all functions to execute the resulting interactive workflow of each stage are included.

4.3 Discussion and Conclusions

Summarizing the results of the IVIS4BigData design process as well as the state of the art and technology review, the characteristics of a variety of established reference models for Big Data analysis and Knowledge Management were considered within the whole design process of the IVIS4BigData Reference Model. In this way, IVIS4BigData considers and integrates a broad range of beneficial standards and recommendations within its horizontal and vertical perspectives, that represent the technical Big Data analysis as well as the direct manipulative interaction between end user stereotypes with single process stages and the adjustments of the respective IVIS4BigData transformations by user-operated user interface controls to support the horizontal evolution from data to wisdom between each process stages. Therefore, IVIS4BigData can represent a sustainable reference model for for distributed visual Big Data analysis to cover the new conditions of the present situation with advanced visual interface opportunities as well as context awareness for perceiving, managing, and interpreting distributed Big Data Analysis results to support insight, e.g., in Virtual Research Environments, i.e., virtual labs to be used in e-Science, industrial research, and corresponding education and training in which the users of the system become empowered co-designers to give domain workers more independence from computer specialists.

4.4 Outlook and Future Activities

After presenting [5] and qualitatively evaluating [4] the IVIS4BigData Reference Model in a road mapping activity [6], where all of the experts agreed that *"this model can represent a framework for their research as well as a generic framework for distributed Big Data analysis applications to support Business Intelligence"* [4], this research presents a generic process structure framework and use case scenarios to close the existing gap between the architectural and the functional mapping of IVIS4BigData. Thus, this research empowers end users to perceive, manage, and interpret Big Data analysis results to support insight and emerging knowledge generation, based on a functional system description for end users, domain experts, as well as for software architects.

However, what is still missing and can be considered as a remaining challenge for this research for achieving an usable and sustainable implementation of IVIS4BigData and its functional system description, is the design of a conceptual **Service-Oriented Architecture (SOA)** that must ensure easy operability as well as a certain flexibility for special accommodations by their customers. Moreover, after designing the

architecture and to evaluate the generic process structure framework and use case scenarios, the integration of certain presented Advanced Big Data Applications of the road mapping activity in 2016 [6] as bookable service modules within this framework is intended.

5 Acknowledgments and Disclaimer


This publication has been produced in the context of the EDISON project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675419. However, this paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Ackoff, R. From data to wisdom. *Journal of Applied Systems Analysis 16* (1989), 3–9.

[2] Ardito, C., Buono, P., Costabile, M. F., Lanzilotti, R., and Piccinno, A. End users as co-designers of their own tools and products. *Journal of Visual Languages & Computing 23*, 2 (2012), 78 – 90. Special issue dedicated to Prof. Piero Mussio.

[3] Bellinger, G., Castro, D., and Mills, A. Data, information, knowledge, and wisdom, 2004.

[4] Bornschlegl, M. X. IVIS4BigData: Qualitative evaluation of an information visualization reference model supporting big data analysis in virtual research environments. In *Advanced Visual Interfaces. Supporting Big Data Applications*, vol. 10084 of *Lecture Notes in Computer Science*. Springer International Publishing, 2016, pp. 127–142.

[5] Bornschlegl, M. X., Berwind, K., Kaufmann, M., Engel, F. C., Walsh, P., Hemmje, M. L., Riestra, R., and Werkmann, B. IVIS4BigData: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments. In *Advanced Visual Interfaces. Supporting Big Data Applications*, vol. 10084 of *Lecture Notes in Computer Science*. Springer International Publishing, 2016, pp. 1–18.

[6] Bornschlegl, M. X., Engel, F. C., Bond, R., and Hemmje, M. L. *Advanced Visual Interfaces. Supporting Big Data Applications*, 1 ed., vol. 10084 of *Lecture Notes in Computer Science*. Springer International Publishing, 2016.

[7] Bornschlegl, M. X., Manieri, A., Walsh, P., Catarci, T., and Hemmje, M. L. Road mapping infrastructures for advanced visual interfaces supporting big data applications in virtual research environments. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI 2016, Bari, Italy, June 7-10, 2016* (2016), pp. 363–367.

[8] Buono, P., Lanzilotti, R., and Matera, M., Eds. *AVI '16: Proceedings of the International Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2016), ACM.

[9] Card, S. K., Mackinlay, J. D., and Shneiderman, B. Information visualization. In *Readings in Information Visualization*, S. K. Card, J. D. Mackinlay, and B. Shneiderman, Eds. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 1–34.

[10] Card, S. K., Mackinlay, J. D., and Shneiderman, B., Eds. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[11] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI magazine 17*, 3 (1996), 37.

[12] Fischer, G. In defense of demassification: Empowering individuals. *Human-Computer Interaction 9*, 1 (1994), 66–70.

[13] Fischer, G. Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2012), AVI '12, ACM, pp. 287–294.

[14] Fischer, G., and Nakakoji, K. Beyond the macho approach of artificial intelligence: empower human designers - do not replace them. *Knowledge-Based Systems 5*, 1 (1992), 15 – 30.

[15] Hoe, S. L. Tacit knowledge, nonaka and takeuchi seci model and informal knowledge processes. *International Journal of Organization Theory and Behavior* (2006), 490–502.

[16] Kaufmann, M. Towards a reference model for big data management, 2016.

[17] Nonaka, I., and Takeuchi, H. *The knowledge-creating company: how Japanese companies create the dynamics of innovation.* Oxford University Press, 1995.

[18] Nonaka, I., Toyama, R., and Konno, N. SECI, ba and leadership: A unified model of dynamic knowledge creation. *Long Range Planning 33* (2000), 5–34.

[19] North, K. *Wissensorientierte Unternehmensführung*, 2., aktualisierte und erw. aufl. ed. Gabler-Lehrbuch. Gabler, Wiesbaden, 1999.

[20] Rowley, J. The wisdom hierarchy: Representations of the dikw hierarchy. *J. Inf. Sci. 33*, 2 (Apr. 2007), 163–180.

[21] Sain, S., and Wilde, S. Review of soft skills within knowledge management. In *Customer Knowledge Management*. Springer Berlin Heidelberg, 2014, pp. 7–55.

[22] Wiederhold, G. Mediators in the architecture of future information systems. *Computer 25*, 3 (March 1992), 38–49.

[23] Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on 26*, 1 (Jan 2014), 97–107.

Marco X. Bornschlegl is a PhD Student at the University of Hagen (FernUniversität in Hagen FUH). He has several years experience in different IT management positions and in various national and international IT infrastructure projects (like, e.g., connection of decentralized project sites, implementation of an identity management solution, e-mail and data center migration, automation, or cloud authentication projects) within large international constructing, engineering, and services groups. Since 2016, he is CISO and leads the business unit IT-Services as an authorized representative with general power of attorney of STRABAG BRVZ GmbH & Co. KG, an internal service provider of STRABAG SE. In this context, he is responsible for the worldwide IT infrastructure within the STRABAG Group including data center operation, hosting and cloud services, network services, connectivity services, client services, and security topics.

Marco Xaver Bornschlegl received his B.Sc. degree in applied computer science and his M.Sc. degree in information technology at the Mannheim university of applied sciences and was a fellow in the Klaus Murmann Fellowship programme of the foundation of the German economy (Stiftung der Deutschen Wirtschaft, sdw), which is directed toward outstanding, socially committed students and doctoral candidates who exhibit potential for taking on leadership responsibilities in business and society, until successful completion of his studies. Furthermore, he received two additional scholarships (Mannheim Model Mid-Tier Industry Scholarship), from leading companies in the metropolitan area Rhein-Neckar. He completed his bachelor and master studies with honors and for his master thesis he received an award from a leading German engineering association (IfKom e.V.).

Kevin Berwind graduated in Business Informatics (Bachelor) followed a by Master of Science degree in Business Informatics majored in "Information Management and Consulting". He is a PhD student at the University of Hagen at the chair of Prof. Dr. Matthias Hemmje. His research interests includes Big Data infrastructures (especially Hadoop), Big Data Management, Big Data Processes, Data Mining, Cloud Computing.

Kevin Berwind is working as a consultant for the BridgingIT GmbH in the business portfolio line SAP (particularly Business Intelligence). He is also a member of the Center of Excellence Analytics within the BridgingIT GmbH and is among others responsible for the connection between research and teaching. Before he joined the BridgingIT GmbH he worked for Accenture as Business & Technology Delivery Consultant. He is also a lecturer at the University of Ludwigshafen am Rhein for business informatics.

Matthias L. Hemmje is a full professor at the University of Hagen (FernUniversität in Hagen FUH), one of the leading distance universities in Germany. He is involved in FUH's research clusters on Virtual Information and Knowledge Environments, Technology Enhanced Learning and E-Education, Knowledge-based Virtual Collaboration Environments, as well as Long term Archival and Digital Preservation.

He has 20 years of experience in IT R&D on national and international level, has been author and co-author in more than 100 publications, is consulting as Senior Expert Consultant for the German Ministry for Research and Education (BMBF), the European Commission (EC), the German Research Foundation (DFG), and several R&D spin-offs. Earlier affiliations include 15 years at the German National Research Center of Computer Science (GMD) and Fraunhofer Integrated Publication and Information Systems Institute (FhI IPSI) in Darmstadt, as well as University Professorships at Ludwig-Maximilians-University in Munich and University of Duisburg.

With respect to applications, he has led or participated in leading functions in various national and international projects on the above R&D topics. In particular, he has successfully led and is still leading relevant R&D activities within several EU projects including research activities on user interfaces, usability, information visualization, information access, archival, preservation, innovation road-mapping, big data and visual analytics, e.g., in the H2020 projects SENSECARE, METAPLAT, EDISON, RAGE as well as FP7 projects SCIDIP-ES, APARSEN, SMART VORTEX, as well as FP6 projects VIKEF, FAIRWIS, FAIRSNET, INNOVANET, FADIVA.

H2O Deep Learning for Hedonic Pricing*

Timothy Oladunni† and Sharad Sharma†
Bowie State University, Bowie, MD 20715 USA

Abstract

This work focuses on a deep learning approach to hedonic pricing theory. Deep learning is a branch of machine learning that uses high level abstraction to improve the predictive capability of a neural network. The hedonic pricing is an econometric concept that suggests that the price of a differentiated commodity is a function of the implicit prices of its composite attributes. We developed a hedonic pricing predictive algorithm. Model generalizability improvement was attained using L1 and L2 regularizations. Performance evaluation was based on r-squared and mean squared errors. Deep learning algorithm improved the r-squared value of prediction by 27% as compared to the ridge regression. Our experiment shows that while deep learning may be preferred in predicting the price of a differentiated commodity like a real estate property, ridge regression is more suitable in explaining the concept of *hedonics* based on the set of implicit prices of its characteristic features. An improved housing evaluation mechanism provides a more consistent and reliable evaluation.

Key Words: Deep learning, neural network, regularization, LASSO, ridge regression, hedonic pricing model, housing prices prediction, hedonic pricing theory.

1 Introduction

For the most part of the last decade, the housing market was at its bubbling phase in the United States. During the bubble, would be mortgagors (qualified and unqualified) were willing to take a leap of faith to finance houses. Mortgage lenders found themselves biting off more than they could chew. The volatility of the housing market erupted into an unprecedented rate of foreclosure in the United States. Reuben and Lei in an article in *Lincoln Institute of Land Policy* pointed out that in nearly fifty years, some local government witnessed a drop of 15 percent in revenue. For example, in Arizona State, by 2009 as its housing prices tumbled by 20 percent, its per capital income dropped by 23.5 percent. During this period, Texas saw a decline of 17 percent in its revenue [27]. The pattern was the same in all the fifty states.

Addressing the international conference on Housing Markets, Financial Stability and the Economy in 2014, Mr. Min Zhu, Deputy Managing Director, International Monetary Fund (IMF) remarked that detecting an over-valuation in real estate market should be the concern of the global economic community [11]. The present housing valuation methodology has been blamed as one of the main reasons why the housing market collapsed. The methodology is based on a sales comparison approach. A sales comparable method is a manual method of house appraisal used to estimate the value of residential houses in the United States and some other countries.

During mortgage loan origination, banks hire certified real estate appraisers to provide professional advice and estimate the market value of a subject property. This is a necessary step for the following reasons; (i) protect banks from giving a mortgage that is more than the market value of a property, (ii) prevent sellers from taking advantage of ignorant buyers, (iii) reduce fraud etc. The appraiser contacts the homeowner, goes to the site, takes pictures and obtains other relevant information of the property. He returns to his office or home and manually calculates the estimated market value of the subject property. The opinion of the appraiser is based on the average price of at least three (occasionally banks requests for more) recently sold properties. Comparable properties are required to have similar characteristic features as the subject property. In addition, comparable properties should be in the same submarket with the subject property. However, adjustments are made if there are differences in features or submarkets.

Just like every manual system, the major setback with the manual value estimation approach is that, it is open to price manipulation and circumvention. Sales comparison approach is based on the economic principle of commodity substitution. Supposing we have two properties X and Y that have the same features (geographical, structural etc.) in all aspects. In an arm's length transaction, the principle of commodity substitution assumes that; a customer will not purchase X more that the market value of Y since the utility of both properties are the same. However, if there is a slight difference between the two, a price difference commensurate with the variation of the features will be added or subtracted. This is known as price adjustment in the real estate world. For instance, supposing

* Extended paper of the proceedings of the 26th International Conference on Software Engineering and Data Engineering. Oct 2 - 4, 2017, San Diego, CA, USA.

† Department of Computer Science. Email: oladunnit0423@students.bowiestate.edu, ssharma@bowiestate.edu.

property X does not have a basement while property Y has a basement. To balance the sales, an adjustment will be made in the price of X to compensate Y for the basement.

Obviously, a large error in one or more of the three comparable properties of a sales comparison methodology will have a significant effect on the predicted value of the subject property. For example, assuming properties A, B, C and D are in the same real estate submarket. Supposing A, B, and C are recently sold properties with the same feature as the subject property D. Using the sales comparison approach, the market value of D is the average price of A, B, and C, assuming other factors remain the same. Obviously, manipulating either A, B or C, will affect the estimated market value of D. Now, if the system generates a wrong value of D, during another sale, D will be used as a comparative property. Thus, the error will have an uncontrollable snowballing, far-reaching and unpredictable consequential effect on the entire submarket. This is one of the challenges of a sales comparison approach.

We hereby propose a *learning algorithmic mechanism of the hedonic pricing theory*. Considering the devastating impact of the housing crisis of the last decade and its consequential damage to the economy, an automated house pricing system is a necessity in the United States. The reverberated effect of the housing crisis demonstrated that the economies of the United States and most developed nations are intrinsically linked to the housing market. The proposed scheme is superior to a sales comparison approach.

Unlike the present sales comparison approach, in the proposed scheme, an error on one property will have a negligible effect on the appraised value of a subject property. The learning algorithm uses thousands of properties to determine the price of a property, (Present appraisal method uses three properties). Also, our approach eliminates the manual feature adjustment methodology of the sales comparison approach, (Manual adjustment is easily prone to human error, manipulation and circumvention).

Furthermore, an automated system will make it more practicable for mortgage underwriters to verify the price of a property, thereby reducing fraud. Real estate appraisers on their part will have a more reliable and efficient instrument to estimate the values of properties. Enhanced property evaluation mechanism will also improve smooth economic transactions in the housing market, thus providing a stable market. Real estate agents, property appraisers and other professionals in the housing market will find it easier to give professional advice to the public. An automated value estimation system also provides tools for local governments that depend on property taxes to finance their budgets. In general, housing is an economic commodity.

We will discuss related works in the next section. Section 3 will be about the experiment, while result and analysis will be in Sections 4 and 5, respectively. We will conclude the paper in Section 6.

2 Literature Review

2.1 Related Work.

The econometric concept of hedonic pricing theory has been applied to different areas of economic transaction in estimating the prices of varieties of differentiated commodities. Mitropoulou, et al.; proposed the use of hedonic pricing index to estimate the price of an Infrastructure as a Service (IaaS) model [22]. Using self-assessed value land data Joshi, et al; proposed the use of hedonic pricing method in quantifying the accessibility to agricultural lands in Nepal based on its economic value [13]. Hedonic pricing has been proposed for estimating the price of cod in Northeast US. The researchers argued that hedonic pricing is superior to the existing pricing methodology [17].

Using the hedonic theory, Famuyiwa and Kayode, investigated the relationship between pricing and rents. The study used linear, linear - log, and log – log functional forms. Performance evaluation was based on R-Squared, Adjusted R-Squared, Akaike information and Schwarz criteria. Experimental results showed that log– log functional form had the best performance [6].

Hedonic pricing has been used to estimate the price per acre of Wyoming agricultural parcels. The study showed that scenic view, elk habitat, sport fishery productivity and distance to town are the most important statistical features of amenities. The outcome of the experiment suggests that hedonic pricing is suitable in measuring the values of amenities. Datasets were obtained from the Geography Information System (GIS) [3]. John et al; proposed a multimarket hedonic model. The proposed methodology was based on interregional studies of wages, price of houses and the location of amenities. The experiment showed that amenities have a significant effect on housing and wages. Datasets were obtained from 1980 census [10]. Using datasets of houses in Christchurch, New Zealand, a researcher compared the performance of neural network predicting capability with the traditional hedonic regression [19]. In our previous work, we developed a multiple listing system with predictive capability using MVC architecture and linear regression [23].

2.2 Theoretical Background

2.2.1 Hedonic Pricing Theory. Rosen (1974) postulated the hedonic pricing theory in a seminar paper. According to Rosen, the market value of a differentiated commodity is a function of its composite characteristic features. By implication, the theory suggests that if we can decompose a differentiated commodity to its composite characteristic features, we can estimate its market value. The theory further argues that the utility a consumer derives from the consumption of a differentiated commodity is based on its characteristic features. Utility is the

satisfaction derived from the purchase of a commodity. Each differentiated commodity has a set of implicit prices with respect to its characteristic features. Thus, the set of implicit prices of each differentiated commodity has a direct relationship with its utility or hedonics [26].

Hedonikos, is a Greek word for pleasure. Therefore, hedonics-pricing theory shows the relationship between the price of a differentiated commodity and its utility or pleasure. This implies that a consumer pays for the pleasure or utility they derive from the purchase of a commodity through its characteristic features. Houses, laptops, cars etc. are examples of differentiated commodities. In an arm's length transaction, if an able and willing consumer decides to purchase a property that has a swimming pool instead of the one without a swimming pool, then he is paying an additional cost for the pleasure he will derive from a swimming pool. In a mathematical world, a partial differentiation of the price of the property in reference to the swimming pool is the additional cost a buyer incurs for purchasing a house with a swimming pool.

Mathematically;

$$\frac{\partial y}{\partial x_k}(x) = \frac{\partial f}{\partial x_k}(x) \quad (1)$$

The *implicit price* as demonstrated in equation (1) is the unit increase in y because of the addition of attribute *xk* (swimming pool). While Rosen did not specify the actual functional form of the hedonic pricing theory, however, assuming y is the value of a differentiated commodity, f an unknown function, x characteristic feature and $\in$ the stochastic error, mathematically;

$$y=f(x) + \in \quad (2)$$

2.2.2 Deep Learning. Deep learning is an emerging field of machine learning research built on the success of an artificial neural network. Artificial neural network was designed to overcome the hyperplane requirement debacle of the perceptron learning algorithm. An artificial neural network is a collection of artificial neurons and nodes built after the functioning system of the brain. Just like the brain, artificial neural networks consist of an interrelated network of neurons [9]. A neural network comprises of input, hidden and output layers. The input layer is connected to the data source, which processes the information and sends the output to the hidden layers. The hidden layer sends its output to the input of the output layer. Depending on the application, a neural network can produce a regression or classification output. The weighted links between the neurons are approximate values, which can be adjusted based on the input and expected output. A neural network can be a feedforward of feedback. A feedforward is an acyclic graph while the feedback is cyclic. A cyclic feedback graph provides a part for a signal to travel in both direction. Thus, network learning is dynamic. However, as computation becomes more complex and costly with large datasets, an artificial neural network becomes less efficient. Deep neural network is a stack of layers of neural network. Just like the neural network, the input layer receives the raw data, processes it and an output is fed as input into the next layer. The receiving layer learns an abstract from the output of the previous layer. The chain continues until the final output.

2.2.3 Over-Fitting. One of the challenges faced in predictive modeling is over-fitting. Over-fitting is a phenomenon whereby, trained predictive models perform very well in training datasets, however, their performances decline when applied to the testing dataset, holdout dataset or future real dataset. Thus, the testing mean-square error (MSE) is greater than the training MSE; a typical case of diminishing return. This implies poor generalizability performance of the hypothesis [1]. The selected hypothesis has a lower in-sample-error but a higher out-of-sample-error. For the most part, the performance of a learning algorithm improves as the size of the dimension increases at the training stage, however, high dimensions do not always guarantee high performance. This situation is known as the curse of dimensionality in most literature.

Several reasons are responsible for over-fitting and its subsequent decline in the performance of a learning algorithm. One major concern is the presence of noise in the training dataset. Noise can be deterministic or stochastic. Deterministic noise is a function of the complexity of the target function. Stochastic noise is a random noise in the dataset. While the deterministic noise of the same model is constant if the same data {x, y} is generated several times, however, the stochastic noise varies. On the other hand, two different models learning from the same data {x, y} will have different deterministic noise. The learning algorithm in addition to learning from the predictive variable of the training dataset also learns the characteristic of its unique noise [16]. Learning algorithms always train on what it is given to train, thus coefficient estimate is awarded to the noisy variables during training.

Another reason for overfitting may be a relatively large size of the characteristic features p. This problem escalates when there are enormous numbers of categorical variables. Categorical variables in the dataset include; County, Advertising, City names, Style, Cooling type, Parking, Pool and Subdivision. The presence of these large numbers of categorical variables posed a major challenge to our computation. For example, the dataset comprises of 4,555 subdivisions in the 31 counties comprising of 430 cities. In the real estate world, subdivisions are geographical submarkets. Submarkets reflect the spatial homogeneity of the housing market. Real estate professionals divide each county of every state to several submarkets to ease the burden of computation. Using a sales comparison approach, during house valuation, appraisers assign a subject property to its closest submarket. All categorical variables were converted to numeric variables using dummy variables. An n level categorical variable requires n -1 dummy variables; the base line is considered the first level. This has a significant impact on the dimensionality of the dataset. Thus, for a reliable test MSE, the dataset needs some regularization.

2.2.4 Regularization. Researchers have come up with

different methodologies of dealing with over-fitting. One of the most popular approaches is the use of subset selection in a linear model. Subset selection may be in the form of best subset selection or stepwise subset selection (forward or backward). While this approach is good, however, it turns out that as the size of the explanatory variables increases, it becomes computationally costly and inefficient. Dropout and batch normalization are other forms of improving the generalizability of a deep neural network.

In a dropout regularization, during training, probabilistic approach is used to 'drop' some nodes and its connections. A small drop out value is more effective than a large value; a large value of dropout may be a result in under-fitting. Fanyu [4] proposed a dropout approach for clustering heterogeneous objects in a deep learning cyber security research. The experiment showed that a dropout generalizability approach achieved a better result in clustering heterogeneous data. Ishan et. al., [14] applied dropout approach to reduce the effect of noise in the CIFAR-10 and MNIST datasets. The researchers claimed that the proposed approach improved the generalizability of the dataset.

Batch normalization as the name suggest, achieves generalizability using normalization approach. It normalizes the activation of previous layers at every batch. Mean and standard deviation of the activation are kept at zero and 1 respectively. Lionel et. al. [25] argued that a combination of batch normalization with exponential linear units improved the performance of a gesture and sign language predictive models. The experiment was based on datasets from the Corpus NGT, Corpus VGT and LAP RGB-D . Rui et. al., [28] applied batch normalization approach to improve the performance of a CNN blurred image classifier.

Shrinking the coefficient to zero or constrained is another alternative choice of model generalizability. This is achieved by adding a shrinkage penalty to the residual sum of squares error. Regularization improves the performance of a model by optimizing its bias-variance tradeoff. It is also very computationally efficient when compared to the best subset selection. Furthermore, it eliminates multi-collinearity of explanatory variables [18]. L1 and L2 norms are the most popular regularization techniques. Babajide, et. al., [2] proposed a L1 and L2 regularizations approach for improving the predictive accuracy of the MNIST, NORD and Reuters texts classifiers. In another study on improving the sales predicted model for a point-of-sale algorithm, Yuta and Katsutoshi [15] proposed a L1 deep learning regularization approach. The authors claimed that the model had an 86% forecasting accuracy. The experiment demonstrated a superior performance of a L1 deep learning approach as compared to a logistic regression.

i. L1 Regularization

Shrinking some small weighted predictors to zero is the main idea behind the L1 regularization. As explained, some features constitute noise, which increases both the complexity and over-fitting of the model. Using L1 regularization, the sum of the absolute weight is multiplied by a hyper-parameter known as lambda and added to the residual sum of squares error (RSS) [12]. One of the basic applications of this concept is the Least Absolute Shrinkage and Selection Operator (LASSO).

For a response variable *y* with characteristic features *x*, a total number of characteristic features *p* and coefficient estimate β, the residual sum of square error (RSS) of the predictive model can be mathematically represented as;

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad (3)$$

In a regression modeling, optimization is attained by keeping the value of RSS as low as possible during testing. However, LASSO goes a step further by imposing a penalty of L1 norm.

Mathematically, LASSO is represented as;

$$LASSO = RSS + \lambda \sum_{K=1}^{p} |\beta_K| \quad (4)$$

Where RSS, β_k, p is the residual sum of square, coefficient values and number of predictors respectively. The l_1 norm of the vectorized coefficient β_k of lasso is given as;

$$||\beta||_1 = \sum |\beta_k| . \quad (5)$$

In a neural network, with an objective function $F(\beta, X, Y)$, an L1 regularizes the model by adding a $\lambda||w||_1$ to the function, where *w* is the weight. The L1 regularized function imposes a constant penalty on all weights regardless of magnitude. This explains why smaller weights are eliminated in an L1 regularization.

ii. L2 Regularization

While L1 regularization shrinks some variables to zero, L2 regularization shrinks them towards zero but retains all the variables. L2 regularizes by multiplying a hyper-parameter known as lambda penalty to the squared weight of the coefficients. The result is added to the residual sum of squares error. Large weights (positive or negative) are more penalized than smaller weights. This is referred to as ridge regression in linear regression and weight decay in neural network [29].

Mathematically, ridge is represented as;

$$RSS + \lambda \sum_{k=1}^{n} \beta_k^2 \quad (6)$$

Where RSS, β_k, k is the residual sum of square, coefficient values and number of predictors respectively. The l_2 norm of the vectorized coefficient β_k of an L2 is given as

$$||\beta||_2 = \sqrt{\sum_k^n \beta_k^2} \quad (7)$$

3 Experiment

Applying the hedonic econometric tool of pricing, the price *Y* of a real estate property at the market equilibrium, is a function of its composite attributes *X*. We experimented this concept by *regressing the price Y of a house H over its characteristic features X*. Thus, using equation 2, of the

hedonic pricing theory as shown above, we develop a house-predictive algorithm.

3.1 Data Exploratory

3.1.1 Datasets. Ten thousand two hundred and twenty-six datasets of houses listed and sold in Maryland, Washington DC, Philadelphia and some parts of Delaware were considered for the experiment. Datasets were single family properties extracted from the Multiple Listing Services (MLS) [21]. Only arm's length transactions and sold houses were considered. The MLS is the repository of properties listed and sold by real estate agents in a metropolitan area. It is a medium where buyers and sellers meet through their agents. Status of properties changes from active to sold after the consummation of transactions. Sold properties must have passed house appraisal and inspection tests. Majority of the transactions on the MLS required financing. We assume that datasets obtained from the MLS are accurate, reliable and complete. Properties considered had a minimum price of $7,000.00 and maximum price of $9,100,000.00 with a mean value of $397,311.00 and a median value of $308,500.00.

3.2 Learning Algorithm

The general function of the hedonic theory was mathematically expressed in equation 2. In our previous experiment, we investigated the functional form using both linear and nonlinear learning algorithms. Our experiment showed that both models are competitive [24]. In this study, using hedonic price definition, we trained two other learning algorithms, namely; deep neural network and ridge regression. Dataset was pre-processed to make it suitable for the predictive model. Figure 1 shows the architectural design predictive algorithm.

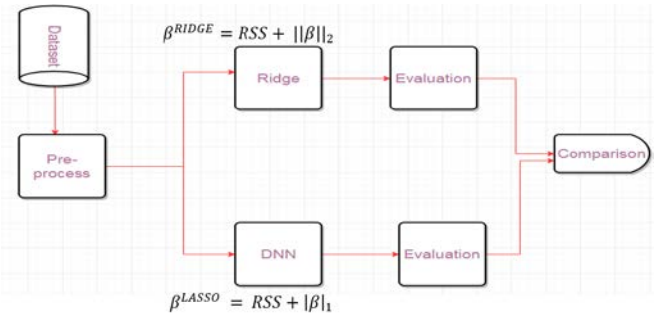

Figure 1: The predictive model

A 10-fold cross validation approach was used to train and test the ridge regression, while a 70:30 split validation approach was used to train and test the deep neural network.

3.2.1 Deep Neural Network Predictive Model. Deep learning algorithm was split-validated into 70:30 ratio for training and testing respectively. Performance was measured using mean squared error (MSE) and r-squared. The deep learning algorithm was a multi-layer feed forward artificial neural network. It was back propagated using a stochastic gradient. Rectifier form was used as the activation function. Unlike tanh and sigmoid activation functions, rectifier overcomes the gradient vanishing debacle. To improve the predictability of the model, we use L1 as the regularization hyper-parameters. The algorithm was run on a H2O cluster.

Other types of deep neural network (DNN) include convolutional neural network (CNN) and recurrent neural networks (RNN) which are more suitable for image and sequential data respectively. H2O deep learning is our preferred choice because our goal is to regress transactional (tabular) dataset.

Luiz, et al. [20] developed a H2O based deep learning predictive model to predict the college attrition rate in Brazil. According to the study, the model had a 71.1% accuracy. Using H2O deep learning, Silvio, et al., [5] studied the federal expenditures of the Department of Research and Strategic Information (DIE) of Brazil. The researchers developed a H2O deep learning predictive model to prioritize actions based on expenditures. Performance outcome was based on a mean square error (MSE) with a value of 0.0012775. Some progress has been made in using H2O as an efficient tool for analyzing big data, however findings have shown that the assumption does not hold in all cases. For instance, Katarina [8] experiment on predictive modeling of energy datasets shows that local learning with support vector regression (SVR) output performs regular support vector regression and H2O deep learning. Performance evaluation was based on prediction accuracy and computation time.

H2O deep learning is a feed forward multiple layer neural network based on stochastic gradient descent and back propagation. The architecture comprises of input layer, multiple hidden layers and output layer neural networks. Depending on the application, the hidden layer can be in large or small numbers. Using R as the analytical tool, datasets are imported into H2O distributed frame using H2O. ImportFile() function. HTTP REST API carries the path arguments to the H2O instances. Tasks are shared among the peer nodes. Datasets are passed into the clusters as chunks. The design is a typical distributed key-value store architecture. Figure 2 shows the architectural design of the deep learning.

As shown in the figure, pre-processed datasets are passed into input node of the deep neural network of the H2O. Datasets were split for training and evaluation. Some parameters which need to be set before training begins are known are hyper-perimeters. This includes; the activation type, the number of epochs, number of hidden layers, if variable importance is necessary and how to deal with missing values. In this category is the regularization method and value, loss function and how to evaluate the performance of the model. We used different values of epoch to observe the accuracy of the model. Epoch values include 1, 5, 10, 15 and 20. Performance evaluation was done for each epoch value.

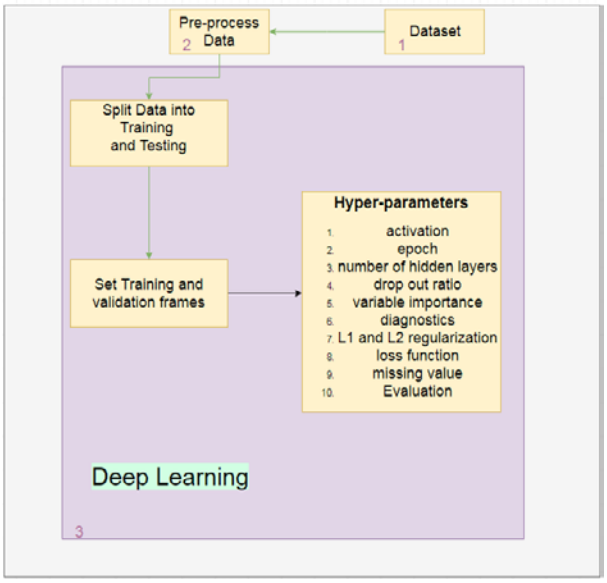

Figure 2.: Deep learning architecture

3.3.1 Ridge Regression. We also trained a ridge regression also using the same dataset and the cross-validation approach. Suppose Y, X, β and ϵ are the response, explanatory, coefficient estimate and stochastic errors respectively. In an OLS, all things being equal, the relationship between the three parameters can be expressed as;

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \epsilon \quad (8)$$

Equation 8 shows a multiple linear regression equation with p explanatory variables. In a traditional statistical term, we say that the null hypothesis of the equation suggests that there is no relationship between explanatory variables X and response variable Y. While the alternative hypothesis assumes the opposite; there is a relationship between X and Y. Using a statistical approach, we can determine whether there is sufficient evidence to show that the estimated value of the coefficients β are far from zero. This can be determined using the *t-statistic* test. T- Statistics uses a probability value (p-value) to determine whether to reject or fail to reject the null hypothesis. Most statistical experiments have been based on 0.05. A threshold of 0.05 implies that any variable with a p value less than 0.05 has a significant effect on the response variable Y; meaning that the relationship is not due to random chance. However, if the p-value is more than the threshold, then it is assumed that we do not have sufficient evidence to show that associations exist or do not exist between the two variables. The error term ϵ in equation 8 suggests that there may not be a perfect estimate of the value of parameters β. Experimental outcome of the estimated value of the accuracy of the of coefficient value β can be measured using the residual sum of square error (RSE). RSE shows the strength of the accuracy of the modeling parameters. In other words, it reflects the fitness of the model showing how the response variable is far from the true regression line.

Ridge regression is an ordinary least square (OLS) based on L2 regularization. As explained, L2 norm applies penalty to each explanatory variable included in the training set. While ridge regression has the capability of shrinking the values of each parameters towards zero, it does not shrink it to zero. Thus, it keeps the interpretability advantage of a linear regression while improving the fitness of the model. Te, et al. [30] argued that despite Fuzzy Neural Networks (FNNs) predictive capability its major setback has been its inability to cope with noisy variables. To overcome this challenge the authors proposed a ridge regression Extreme Learning Fuzzy System (RR-EL-FS). The outcome of the experiment shows that there is an improved performance of RR-EL-FS when compared with FNN. In order to solve the problem of zigzags when tracking facila landmarks in a video, Zhenye, et al. [7] proposed an improved convolutional neural network (CNN) model using ridge regression. The authors claimed that ridge regression transforms adjacent error into bias errors, proposed model showed an improvement with a 300-VW dataset. The algorithm was run using a mobile device (iPhone 5s) at 150 Fps.

As suggested by hedonic pricing theory, price was set as the response variable while other parameters were set as the predictive variables. Figure 3 shows the schema of a ridge regression. As shown in the diagram, a ridge regression is an ordinary least square (OLS) regression with the residual sum of squares (RSS) regularized using the L2 norm.

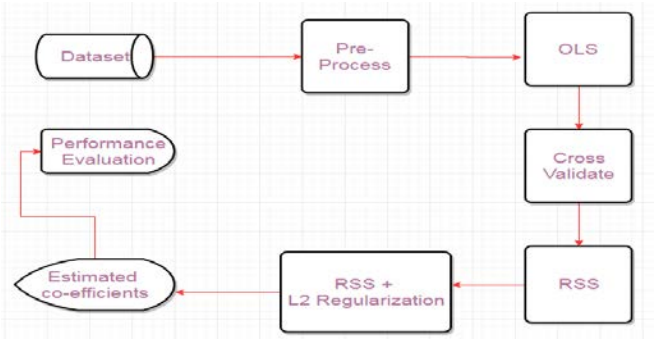

Figure 3: Ridge regression architecture

Normalization, removal of extraneous and outlier data was done at the preprocessing stage. Regularization was necessary in the computation because of the large number of categorical variables.

4 Result

4.1 Deep Learning Result

We computed the r-squared for each epoch of the deep learning algorithm. R-squared is a measure of the fitness of the model. Figures 4 and 5 shows the results of the experiment using r-squared and MSE vs epoch respectively.

A high r- squared suggests that the model fits very well, while a low r-squared suggest that the fitness of the model is very poor. An r-squared takes value between zero and positive one.

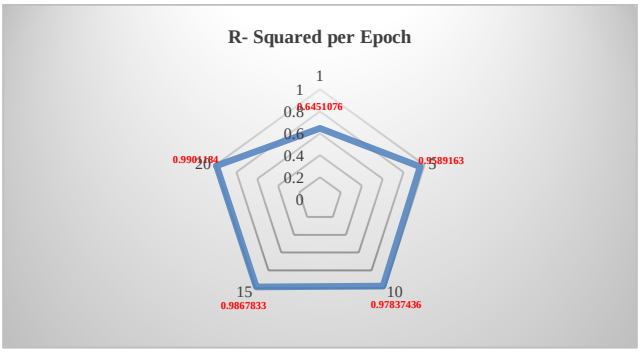

Figure 4: R-squared per epoch. Model has highest r-squared value of 0.99 at the 20th epoch

Exact zero or one r-squared values are rare in experiment. Using the r -squared result, Figure 1 shows that the model has the best and worst performances when the epoch of the learning algorithm was at 20 and 1 respectively. At this point, the results suggest a 0.99011 and 0.6451 r –squared value respectively.

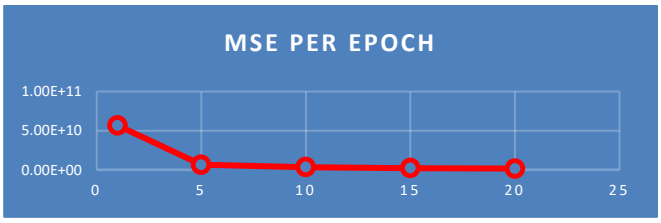

Figure 5: MSE vs Epoch. The mean squared error (MSE) per epoch circle

We also computed the mean squared error (MSE) per epoch circle. A MSE measures the average deviation of the predicted value from the expected outcome. A large value of MSE shows that the predicted value is very far from the expected value, while a low MSE suggests a good performance. Our result suggests that at the 20th epoch the MSE was at 1.56863168E9. This is the best performance epoch which is in line with the result of the r-squared.

4.2 Ridge Regression Result

We also measured the result of the ridge regression using r-squared and mean squared error. The result is shown in Figure 6.

The left-hand side of the y-axis shows the coefficient estimates of the variables. 23 variables are represented with their corresponding serial numbers at the left tip of the chart. Numbering and variables names are explained using the key below. The lower part of the x-axis shows the log lambda value, while the upper side shows the total number of variables (23). For each lambda value, 23 variables are considered.

Key

1. County, 2. Advertising, 3. AgeRestricted Housing, 4. BathsHalf, 5. BathsFull, 6. Basement, 7. Beds, 8. CityName, 9. DaysOnMarket 10. DomProperty, 11. Stories, 13. Style, 14. PostalCode, 15. Longitude, 16. Latitude, 17. YearBuilt, 18. ListingArea, 19. Cooling, 20. LotAreaAcre, 21. Parking, 22. Pool, 23 Subdivision

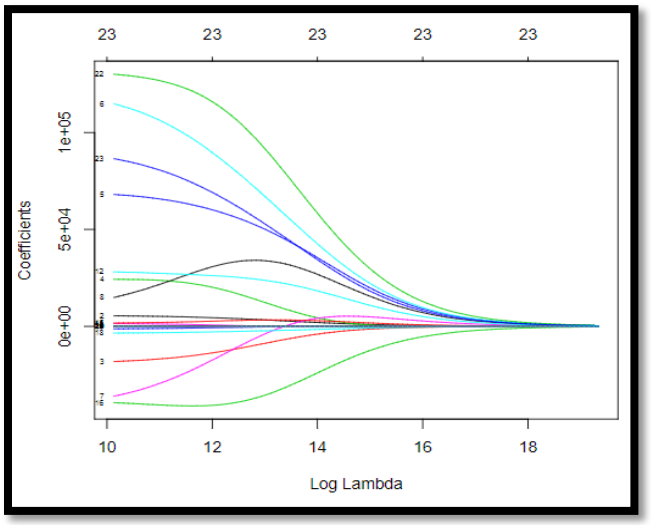

Figure 6: Ridge coefficient vs Log Lambda

The ridge regression showed an r-squared of 0.779, while the mean squared error was at 3.4522165816E+10. We plotted the ridge coefficient against the log lambda of the tuning parameter. Figure 3 shows that all the twenty-three explanatory parameters were considered in the model. Each curve is an estimate of the coefficient values of each of the twenty-three variables. The green line represents variable 22 which is the swimming pool. Variables 1 and 4 represents county and half baths respectively. When the value of lambda becomes zero, model coefficients become typical values of ordinary least square regression. On the other hand, when the value of lambda becomes extremely large, at this point, all the coefficients of the ridge regression become essentially zero, this is the null model stage

The graph suggests that variables 22, 6, 23 and 5 have the highest weight or influence on the value of a house respectively. Houses with swimming pools may be the most expensive in the counties considered for the study. Thus, in an arm's length transaction, if an able and willing consumer decides to purchase a property that has a swimming pool instead of the one without a swimming pool, then he is paying an additional cost for the pleasure or *hedonics* he will derive from the swimming pool.

5 Analysis

The result shows that deep learning performed better in terms of predictability with an r-squared value of 0.99 at the 20th epoch. Also, the r-squared performance of the ridge regression was 0.779.

6 Conclusion

We have used the deep learning - hedonic pricing approach to predict the price of a real estate property based on the H2O

deep learning, this was our preferred choice of deep learning architecture. H2O deep learning is a feed forward multiple layer neural network based on stochastic gradient descent and back propagation. The architecture comprises of input layer, multiple hidden layers and output layer neural networks. R was used for computation. Datasets were imported into H2O distributed frame using H2O. ImportFile() function. Model generalizability improvement was attained using L1 regularizations. Performance evaluation was based on r-squared and mean squared errors.

Using the hedonic pricing theory, our experiment demonstrated that a deep learning has a superior predictive capability in estimating the value of a real estate property. It also shows that a ridge regression model has an interpretability and inferentiality advantage. While deep learning may be preferred in predicting the price of a differentiated commodity like a real estate property, ridge regression is more suitable in explaining the concept of utility based on the set of implicit prices of its characteristic features.

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*, 1st ed., California: AMLbook.com, 2012.

[2] B. O. Ayinde and J. M. Zurada, "Deep Learning of Constrained Autoencoders for Enhanced Understanding of Data", *IEEE Transactions on Neural Networks and Learning Systems*, 99:1-11, 2018.

[3] C. T. Bastian, D. M. McLeod, M. J. Germino, W. A. Reiners, and B. J. Blasko, "Environmental Amenities and Agricultural Land Values: a Hedonic Model Using Geographic Information Systems Data", *Ecological Economics*, 40(3):337-349, March 2002.

[4] F. Bu, "A High-Order Clustering Algorithm Based on Dropout Deep Learning for Heterogeneous Data in Cyber-Physical-Social Systems", *IEEE Access*, 99:1, 2017.

[5] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, "Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning", 15th IEEE International Conference on Machine Learning and Applications, Anaheim, 2016.

[6] F. Famuyiwa and G. K. Babawale, "Hedonic Values of Physical Infrastructure in House Rentals", *Journal of Facilities Managemen*, pp. 211-230, 2014.

[7] Z. Gan, L. Ma, C. Wang, and Y. Liang, "Improved CNN-Based Facial Landmarks tracking via Ridge Regression at 150 Fps on Mobile Devices," in *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Shanghai, China, 2017

[8] K. Grolinger, M. A. M. Capretz, and L. Seewald, "Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources", IEEE International Congress on Big Data (BigData Congress), San Francisco, 2016.

[9] G. Hacheling, *Mastering Machine Learning with Scikit-Learn*, Birmingham: Packt Publishing, 2014.

[10] J. P. Hoehn, M. C. Berger, and G. C. Blomquist, "A Hedonic Model of Interregional Wages, Rents, and Amenity Values", *Journal of Regional Science*, 27(4):605-620, November 1987.

[11] IMF, "Housing Markets, Financial Stability and the Economy", [Online]. Available: http://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514, 5 June 2014.

[12] G. James, D. Witten and T. Hastie, *An Introduction to Statistical Learning*, New York: Springer, 2013.

[13] J. Janak, M. Ali, and R. P. Berrens., "Valuing Farm Access to Irrigation in Nepal: A Hedonic Pricing Model", *Agricultural Water Management*, 181:35-46, 2017.

[14] I. Jindal, M. Nokleby, and X. Chen, "Learning Deep Networks from Noisy Labels with Dropout Regularization", IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016.

[15] Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales", IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016.

[16] M. Kuhn and K. Johnson, *Applied Predicitve Modeling*, New York: Springer, 2013.

[17] M.-Y. Lee, "Hedonic Pricing of Atlantic Cod: Effects of Size, Freshness, and Gear", *Marine Resource Economics*, 3:259, 2014.

[18] C. Lesmeister, *Mastering Machine Learning with R*, Birmingham: Packt Publishing, 2015.

[19] V. Limsombunchai, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network", in New Zealand Agricultural and Resource Economic Society, Blenheim, 2004.

[20] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. C. Victorino, and M. Holanda, "Early Prediction of College Attrition Using Data Mining", 16th IEEE International Conference on Machine Learning and Applications, Cancun, Mexico, 2017.

[21] MRIS, Metropolitan Regional Information Systems, Inc, 2016. [Online]. Available: http://www.mris.com/. [Accessed 15 June 2016].

[22] P. Mitropoulou, E. Filiopoulou, C. Michalakelis, and M. Nikolaidou, "Pricing cloud IaaS services based on a hedonic price index", *Computing*, 98(11):1075, 2016.

[23] T. Oladunni and S. Sharma, "Predictive Real Estate Multiple Listing System Using MVC Architecture and Linear Regression," ISCA 24th International Conference on Software Engineering and Data Engineering (SEDE 2015), San Diego, 2015.

[24] T. Oladunni and S. Sharma, "Hedonic Housing Theory – A Machine Learning Investigation", The 15th IEEE International Conference on Machine Learning and Applications (IEEE ICMLA'16), Los Angeles, 2016.

[25] L. Pigou, M. V. Herreweghe, and J. Dambre, "Gesture and Sign Language Recognition with Temporal Residual

Networks", IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, 2017.
[26] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82(1):35-55, 1974
[27] K. Rueben and S. Lei, "What the Housing Crisis Means for State and Local Governments", *Lincoln Institute of Land Policy*, 2010.
[28] R. Wang, W/ :o. R/ Qom amd K/ Wi, "Blur Image Classification based on Deep Learning", IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, 2017.
[29] J. F. Wiley, *R Deep Learning Essential*, Birmingham: Packt Publishing, 2016.
[30] T. Zhang, Z. Deng, K. S. Choi, J. Liu, and S. Wang, "Robust Extreme Learning Fuzzy Systems using Ridge Regression for Small and Noisy Datasets", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, 2017.

Timothy Oladunni is a computer scientist with a background in Electrical Engineering. He earned his masters and doctoral degrees in Computer Science from the Bowie State University. He is presently an adjunct professor at the same university. His research interest includes: machine learning, statistics, artificial intelligence, pattern recognition and software engineering.

Sharad Sharma is an Associate Professor in the Department of Computer Science and Director of the Virtual Reality Laboratory at the Bowie State University. He has received a Ph.D. in Computer Engineering from Wayne State University, Detroit, MI in 2006 and M.S. from University of Michigan, Ann Arbor, MI in 2003. His research focus is on modeling and simulation of multi-agent systems (MAS) and multi-user virtual reality (MUVR) environments for emergency response and decision-making strategies. His work is motivated by the need of research in real-time agent navigation for reaching a goal in emergency situations like evacuation. He is involved in developing new data and visualization methods for course of action planning, visualization, training, and assessment. He is also exploring socio-cultural issues in Collaborative Virtual Environments (CVE) for emergency response and decision making in dense urban environments. His research interests include Software Engineering, Virtual Reality, Augmented Reality, Human-Computer Interaction, Data Science, and Data Visualization.

Instructions for Authors

The International Journal of Computers and Their Applications is published multiple times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration

1. Email your manuscript to the Editor-in-Chief, Dr. Fred Harris, Jr., Fred.Harris@cse.unr.edu.

2. Illustrations should be high quality (originals unnecessary).

3. Enclose a separate page (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.

B. Manuscript Style:

1. The text should be **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month and year.
5. Figures should be captioned and referenced.

C. Submission of Accepted Manuscripts

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-Chief.
2. The submission may be on a CD/DVD or as an email attachment(s) . **The following electronic files should be included:**
    - Paper text (required).
    - Bios (required for each author). Integrate at the end of the paper.
    - Author Photos (jpeg files are required by the printer, these also can be integrated into your paper).
    - Figures, Tables, Illustrations. These may be integrated into the paper text file or provided separately (jpeg, MS Word, PowerPoint, eps).

3. Specify on the CD/DVD label or in the email the word processor and version used, along with the title of the paper.

4. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges

After a manuscript has been accepted for publication, the contact author will be invoiced for publication charges of **$50.00 USD** per page (in the final IJCA two-column format) to cover part of the cost of publication. For ISCA members, $100 of publication charges will be waived if requested.

January 2014