JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 13, Issue 1
March 2020
ISSN: 1946-1836

In this issue:

iscap
Information Systems & Computing Academic Professionals

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

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

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

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

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

An Empirical Study of Post-Production Software Code Quality When Employing the Agile Rapid Delivery Methodology

Laura F. Poe
laurapoe@verizon.net
Longwood University
Farmville, VA

Elaine Seeman
seemane@ecu.edu
East Carolina University
Greenville, NC

Abstract

In response to the business need to adopt a faster delivery model to enable them to stay ahead in the marketplace, organizations implementing Agile practices expect to deliver projects faster and with higher quality. Widespread assumptions of increased code quality for software implementations using Agile require empirical investigation. The purpose of this paper is to evaluate software delivery with an emphasis on the quality of the software code. The outcome of this research will assist business leaders with making informed decisions on selecting a successful project methodology. While numerous factors can impact project delivery, this case study of DigiTek LLC evaluates their software development project teams' software delivery hours to the number of defects encountered during development and after implementation to production. Teams using the traditional Waterfall methodology had slightly higher production code quality when compared to teams using the Agile methodology across similar software development products. Companies planning to adopt Agile should evaluate the impacts to code quality and consider other factors as part of the decision to transition.

Keywords: Software engineering; case study; Agile; code quality; rapid delivery; Waterfall

1. INTRODUCTION

This paper compares Agile and Waterfall project methodologies and the quality of software code based on the number of production defects relative to software development hours. The adoption of the Agile methodology in software development projects has been considered a means to stay ahead of technology trends that are sweeping the industry. Agile is a methodological response to the rigid requirements and design processes of earlier methodologies that often lead to fixed project scope, significant modifications to design late in the software development life cycle, and customer dissatisfaction. Success has been primarily determined by the delivery speed and the ability to change design and scope based on customer input. Additionally, the quality of the software code should be analyzed to determine its software reliability in conjunction with rapid delivery. A quantitative, empirical analysis is necessary to determine the effect of Agile practices on the quality of code being implemented. This study was performed to evaluate Agile's impact on code quality by directly comparing the Waterfall defect rate to the defect rate using Agile. The benefit of the case study is the ability to obtain valid project results in a live industrial setting.

The emergent Agile methodology professed as delivering higher quality products demands a comparative research study to evaluate the accuracy of the claims and to classify characteristics leading to higher quality software code. While numerous factors can determine team performance, such as team diversity of skill set as described in Lee and Xia's study on Software development agility (2010), methodological success and team performance can be measured based on the quality of the product delivered. This paper discusses the process of conducting a quantitative study with eight teams employed at DigiTek LLC in order to obtain actual measurements of production code quality after release by teams using both Waterfall and Agile methodologies. The advantage of an empirical study is the delivery of reliable results that can be used to compare and contrast industry averages. At this juncture, few empirical studies have been performed that provide a side-by-side analysis between Waterfall and Agile projects for code quality impacts using a multidimensional construct. The metrics used provide consistent and equal factors for statistical comparisons.

We intend to provide the following contributions on agile literature. The results of the study will be helpful to practitioners, who are seeking faster software delivery releases, to choose the appropriate project methodology for their organization dependent upon complexity of software development team structures.

The sections of this paper are divided as follows: the fundamental differences between Waterfall and Agile and highlights the main reasons organizations choose the Agile methodology; issues related to Agile and software testing; the scope of the case study and research questions; and the results of the case study with recommendations.

2. THEORETICAL BACKGROUND

Waterfall and Agile Approaches

The emergence of the Agile methodology in the 1990's went largely undetected by the software development industry. Few companies pursued the methodology to deliver new functionality or code enhancements. The shift towards rapid delivery methodologies over the last decade demanded that the software development community question the efficiency of traditional software processes and models (Sampaio, Vasconcelos, & Sampaio, 2004).

Competition in the technology sector was a distinct motivator for companies and led to market driven changes in methodology to accommodate the delivery of technological advancements at near record paces. Agile organizations have the ability to react swiftly and decisively to sudden shifts in overall market conditions, such as the emergence of new competitors and new technologies (Holbeche, 2015). In an effort to deliver products more quickly, organizations evaluate optimizations to internal processes, and as a result, shift from the traditional Waterfall project methodology to Agile. The shift to Agile requires changes to project team member roles and transitions software development from the step by step approach to the combined step of design, development, and testing. Typically correlated with IT and software development, Agile is being implemented as a model across organizations as a means for tracking and delivering work.

The Waterfall method associated with the Software Development Life Cycle had been uncontested for years until the rise of Agile. IT projects using Waterfall operated in a backwards scheduling approach by selecting a go-live, implementation date, before the project began. Once the date was determined, the project manager and team would identify the tasks required to implement the project, allocating time for each of the activities. The Waterfall methodology requires that each phase of the project reach a stage of completion before the next phase begins. Agile practices allow the team to decide how much work can be accomplished per iteration before beginning the work. These distinctly differing models of delivery management maintain the same end goal of releasing a quality product.

IT managers often cite the reduction of time spent in initial planning, leading to an evolving and more efficient process as an advantage of Agile (Dyba et al. 2008). Evidence suggests that the rigid development processes of the Waterfall methodology results in rework, customer dissatisfaction (Dybe et al. 2008), and missed implementation dates. Moreover, project team members often endure increased working hours necessary in order to meet the go-live dates. In addition, date-driven projects tend to go over budget due to increased resources and project hours for a successful implementation.

Methodological Distinctions

Simple, key methodological distinctions between the Agile and Waterfall methodologies (Table 1) can be summarized in the overall management of

the project. Practitioners utilizing the Waterfall methodology are accustomed to intense up-front project planning. With the Agile methodology, the focus moves to a more independent model with self-managing teams. A hybrid approach is usually appropriate for project planning to incorporate the known factors, such as the size of the project and known future requirements (Serrador et al. 2015).

Agile Components

In the traditional waterfall methodology, each phase of the project is distinctly separate. The System Development Life Cycle, i.e. requirements, design, development, testing, and deployment, requires that each phase is completed individually before moving on to the next phase. The software must pass a number of quality checks after completing a phase and before moving on to the next phase. Each phase of the waterfall method has a specific deliverable, and the project has a predetermined go-live date. However, a major disadvantage is the difficulty returning to a previous phase and making significant changes. For instance, if a new requirement is found during testing, the requirement could impact the earlier design phase deliverables, causing changes that result in project delivery delays. Additionally, software products are not delivered incrementally, and the final product is implemented as a single release. Waterfall methodologies are well suited for predictable environments, i.e. heavily regulated, but are cumbersome, bureaucratic, and lack the capability to succeed in environments comprised of high uncertainty and change (Beck 1999). Traditional measurements of project success focused on the time, budget, and product quality (Atikinson 1999).

The Agile approach to software development originated to overcome the disadvantages of the waterfall methodology, primarily to shorten the development time for software and get the product to the customer / market as quickly as possible. Work is broken down into small, iterative cycles, known as sprints, with code releases built into the iterations. Each increment builds upon the previous until the total software product is complete. Agile is comprised of repeated practices that enable teams to work in a faster paced environment, such as behavior driven development, test automation, continuous integration, and continuous deployment. Each code release contains small portions of development necessary to build the larger framework. The measurement of a successful project is the overall end product delivered the customer and the subjective variable of the customer's satisfaction of the product's quality (Jugdev et al. 2005).

According to Serrador and Pinto (2015), projects utilizing the Agile methodology's iterative approach, report consistently higher project success. However, Agile software development emphasizes that teams should be self-managed and without direction on the implementation of leadership (Moe et al. 2009). A further challenge is the lack of synchronization of interdependencies among multiple Agile teams working on the same overall project (Melo et al. 2013). Teams with multiple autonomous members tend to find difficulty with the principle of the Agile manifesto, "the most efficient and effective method of conveying information to and within a development team is face-to-face conversation" (Beck 2001).

Planning Ceremonies

Agile claims vast reductions in delivery based on the ability to adjust to changing business demands by operating in iteration cycles. However, the upfront planning is crucial to the team's success and is a process that cannot be eliminated from the development cycle. Planning ceremonies are used to determine the software release schedule and identify the chunks of work that fill the team's backlog. Each backlog item must be fully refined for the team's consumption and execution. The Agile terminology labels traditional project requirements as stories, which are groomed by working with the business stakeholder representing the team to fully convey the expected behaviors of the system. The depth of the groomed stories can be a determining factor in the success of the software code and can, also, contribute to the number of defects found in pre-production. Rigorous empirical analysis of the impact of story grooming is a key measurement of the effectiveness of agile software development.

Quality Impacts

From a practical perspective, using Test-Driven Development (TDD) and Behavior-Driven Development (BDD) approaches require that testing is performed simultaneously with development. TDD necessitates the creation of automated test scripts while developers are building the product. The test scripts become the requirements documents. BDD operates similarly to TDD; however, the language of the test scripts are basic programmable statements that can be automated using special software. BDD was designed to eliminate the complexity of building automated testing scripts. Specific phases of testing, such as full integration and user

acceptance testing, are not formally used. The product is accepted by the product owner, a business stakeholder, who typically previews a demo of the final version of the product before the code is released to production.

Along with the iterative development and testing, eXtreme Programming (XP) is frequently used in software development projects. The goal of XP is to accomplish significantly faster development through the collaboration of developers programming the same functionality simultaneously. During XP, stakeholders are present to review the product and provide immediate feedback. Favorable results have been reported in studies using XP as part of the Agile methodology, although few case studies have been performed providing documented, empirical evidence (Layman et al. 2006).

3. WATERFALL STRUCTURAL COMPONENTS

In contrast, waterfall projects follow several phases of testing to include a minimum of the following: unit/system, integration, regression, and user acceptance testing. Testing is a formal part of the project life cycle with documentation and traceability to the business requirements. While Agile appears to have a gap in the lack of full integration and traditional user acceptance testing, the prevailing industry argument maintains that smaller increments require less integration and software can be validated through repeatable automation tests. The Agile version of acceptance testing is achieved through the business' acceptance of the product at the end of the sprint demo.

Formalized Business Requirements

Testing is a validation that the system works according to the business requirements. Waterfall business requirements documents created traceability between test cases and specific requirements. Rather than producing business requirements documents, Agile creates stories that are recorded and serve as the business requirements. Many Agile tools, such as VersionOne, JIRA, GitHub, etc. provide linkages from test scripts to stories, thus solving for the traceability of requirements to testing. Due to the rapid timeline, acceptance tests are often foregone as too timely and requiring business resources that may not be readily available. When acceptance tests are performed, they are process driven and tend to be end-to-end. External systems are typically required to be fully functional and require considerable work to set up the environment properly before test execution (Rogers 2004).

Quality of software code is primarily measured by evaluating the number and criticality of production defects. To produce high software quality and reduce the number of defects, testing is seen as the solution for employing higher levels of code quality. Thus, the focus in Agile has shifted from scrum principles to ensuring test-driven development (TDD) and behavior-driven development (BDD) practices. The use of the Agile methodology affects perceived software quality through its impacts on internal performance (Kong 2007). When complemented with pair programming, an eXtreme Programming method wherein developers work side by side, teams are more productive and produce fewer defects (Rico et al. 2009). Test-driven development is posited to be "100 times more efficient than traditional methods when combined with continuous integration" (Rico et al. 2009). Rico provides return on investment (ROI) values to validate the claim, but does so without an empirical side by side comparison of the number of defects to software development hours.

4. CASE STUDY SCOPE AND RESEARCH OBJECTIVES

Research performed in previous studies of Agile has found that performance is linked with the effectiveness of teamwork coordination in software development teams (Moe 2009). The quality of the software code is of critical importance to organizations seeking to maintain a competitive advantage in the marketplace. This research seeks to provide an analysis of the defect rate of the Agile methodology and compare it to the defect rate using the Agile methodology across eight separate projects.

DigiTek LLC is a consulting firm that works with public and private sector companies. Data used for this study was gathered during consulting activities for two different clients, an insurance company and financial institution, and a total of eight project teams. This resulted in eight separate project teams from which to evaluate. The analysis was performed using three Waterfall and five Agile project teams of similar size. Each of the teams worked within the same line of business, each with separate functionality being developed and implemented.

The following research questions were raised:

1. Is the Agile Methodology able to complete a comparable number of software development hours while achieving higher levels of software code quality when compared to similar efforts using the Waterfall Methodology?

2. How is the criticality of defects impacted by Agile versus Waterfall?

Research Method

To establish comparable measurements between the Waterfall and Agile projects, software development hours were recorded along with production and pre-production defects. Agile projects operated in two-week sprint cycles for a total of twelve weeks, and Waterfall projects operated for a total of twelve weeks. Included in the study were two Agile project teams and two Waterfall project teams.

The following assumptions were made for the Agile teams participating in the study: Extreme Programming (XP) was used; stories were created and properly groomed; level of skill sets of developers, testers, and supporting team members were not a determinant in the study; and equal skill set was assumed across all participating teams.

Limitations to the study were recognized, as each of the teams worked on different software platforms and functionality. Due to the nature of rapid delivery methodologies, pre-production defects are not normally recorded. Agile teams work closely together to develop, test, develop, and test in eXtreme Programming sessions, fixing defects immediately. The ability to fix the defects and gain stakeholder approval real-time propels the team for faster turnaround of the final product. As a result, Agile teams have reduced reliance on pre-production metrics for determining quality. The focus becomes on the successful completion of iterative product releases and production code quality. For the purposes of this study, pre-production defects were recorded and used for the quantitative analysis. However, criticality was not recorded for pre-production defects regardless of methodology.

Case Study Procedure

Results from both a large financial institution as well as an international insurance company operating with Agile software development teams as well as Waterfall teams were used to evaluate the choice of project methodology's impact on code quality. Quality measurements were based on the number and criticality of production defects relative to the number of software development hours. Level of Agile maturity was measured for each team based on the following: grooming ceremonies were performed, software was released at the end of each iteration/sprint, each team had an assigned product owner, each team had an assigned scrum master, and daily scrums were held.

Participants included eight project teams, three of which utilized Waterfall and five that utilized Agile. Software development hours and defects were measured across the teams for a twelve-week period. With differing numbers of developers per project team, the ratio of software development hours to code quality allowed for an equivalent measurement. By employing a direct comparison of Agile and Waterfall projects, the analysis avoids the subjectivity and relativity of industry defect averages. Each of the project teams tracked their pre-production and production defects for code release during the twelve-week period. Additional measurements were performed to determine the impact of the level of agile maturity on the number of production defects. Product owners provided a level of acceptance of the product prior to each release to production by giving a go or no-go decision to release.

5. RESULTS AND ANALYSIS

Raw Defect Analysis

Pre-production defects, which were recorded for all teams during the twelve-week cycle, were much higher for Waterfall teams than for Agile teams, 43 to 13 respectively, with double the percentage of defects per software development hours, 1.37% to 0.63%, as shown in Table 2. This indicates a higher level of testing prior to implementation in order to prevent production defects. The amount of testing automation utilization was not considered in the quantitative analysis.

Defects by Project Methodology					
Project Methodology	Total Combined Development Hours	Total Number Prod Defects	Total Number Pre-Prod Defects	Percentage of Prod Defects to Dev Hours	Percentage of Pre-Prod Defects to Dev Hours
Waterfall	3140	19	43	0.61%	1.37%
Agile	2076	21	13	1.01%	0.63%
Full project cycle of 12 weeks					

Table 2
Defect analysis for Waterfall and Agile projects based on software development hours

Likewise, production defects were recorded for all teams during the twelve-week cycle. The total number of production defects for Waterfall compared to Agile were relatively the same, 19 to 21. However, when factoring in the number of software development hours for each project methodology, the percentage of production defects relative to software development hours for the Waterfall teams was nearly half the percentage of defects found in the projects using Agile. Waterfall projects had 0.61% of product

defects to development hours whereas Agile suffered 1.01%. The analysis indicates that the Waterfall methodology yields higher software development quality due to the time spent testing in pre-production regions.

Defects by Criticality

Total production defects recorded by all teams were further broken down by criticality. The criticality of defects was recorded as high, medium, or low, and criticalities were determined by the project managers. A general definition of high criticality represented defects that inhibited major functionality or contained significant user experience impacts. Medium criticality is assigned to defects with impacts to core functionality where alternative solutions exist. Low criticality defects do not prevent users from performing any intended functionality, such as font sizes on graphical user interfaces.

When reviewing the criticality by project methodology, fewer Waterfall defects were considered *high* criticality (Figure 1 and Figure 2). The results for Agile had the most difference when considering the *high* criticality defects, which made up 33% of their production defects compared to 26% for Waterfall. The *high* plus *medium* combined group signify defects that must be fixed prior to going to production. In considering the *high* plus *medium* criticality, the Agile projects yielded the same percentage of defects with 52%.

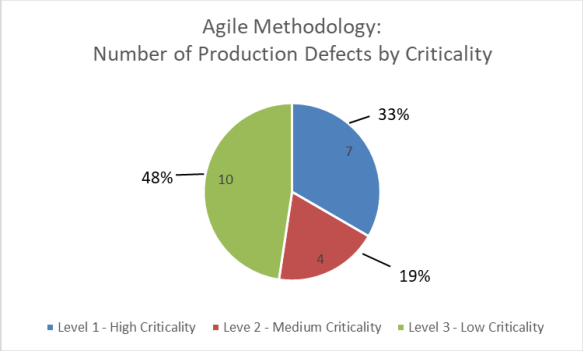


Figure 1
Defect analysis by criticality for Agile Methodology

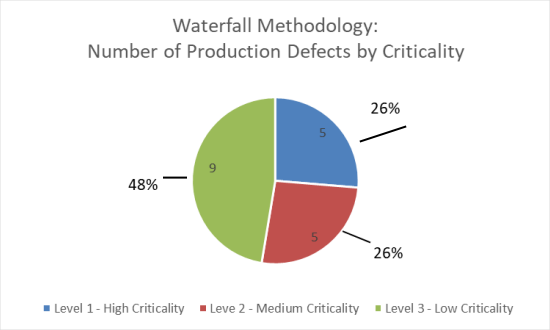


Figure 2
Defect analysis by criticality for Waterfall Methodology

The analysis between the two methodologies indicates that projects using the Waterfall methodology are more likely to spend more time testing, resulting in higher numbers of pre-production defects found and remediated prior to implementation. Agile's rapid delivery methodology finds fewer defects in pre-production but has double the production defects as a percentage of software development hours after implementation. The cause of the disparity can be related to numerous factors, such as the longer period of time spent testing in Waterfall projects contrasted with the targeted test cases performed during the Agile sprint cycle.

Agile project teams rely more heavily on automated test cases and spend less time on user acceptance testing. User acceptance testing, as well as manual testing, can allow the user/tester to focus efforts on attempts to break the application. Automated test scripts tend to focus on the simple, 'happy path', test cases rather than creating test scenarios for multiple permutations of the functionality. Automated test cases can be more accurate than manually running a test, because they are less subject to error. However, automated test cases sacrifice the depth of the testing.

6. CONCLUSIONS AND RECOMMENDATIONS

The overall results of this study suggest that Waterfall projects spend more time during the testing cycle and identify more defects prior to production. Agile projects have smaller pieces of functionality being delivered and focus on the speed of delivery, thereby shortening the volume of testing. Practitioners making selections between the two methodologies should consider the depth of pre-production testing. The Agile methodology can increase the software development quality through more expansive testing measures prior to releasing the software

code to production. The use of automation simplifies and shortens the testing cycle but should not be relied upon solely for testing all permutations of the changed functionality.

Future studies of automation could determine the impact of automated testing and behavior driven development on the code quality measurement. Additionally, this study did not review customer satisfaction of the products delivered but merely the defects.

7. REFERENCES

Atikinson, R. (1999). Project management: cost, time, and quality, two best guesses and a phenomenon, it's time to accept other success criteria. International Journal of Project Management, 17, 337-342.

Beck, K. (1999). Extreme Programming Explained. Boston: Addison-Wesley.

Beck, K. (2001, February 11). Principles Behind the Agile Manifesto. Retrieved October 25, 2017, from Manifesto for Agile Software Development: http://agilemanifesto.org/principles.html

Dyba, T., & Dingsoyr, T. (2008). Empirical studies of agile software development: a systematic review. Information Software Technology, 50(9), 833-859.

Holbeche, L. (2015). The Agile Organization: How to Build an Innovative, Sustainable and Resilient Business. London: Kogan Page Limited.

Jugdev, K., & Muller, R. (2005). A retrospective look at our evolving understanding of project success. Project Management Journal, 36(4), 19-31.

Kong, S. (2007). Agile Software Development Methodology: Effects on Perceived Software Quality and the Cultural Context for Organizational Adoption. Ann Arbor: ProQuest Information and Learning Company.

Layman, L., Williams, L., & Cunningham, L. (2006). Motivations and measurements in an agile case study. Journal of Systems Architecture, 52, 654-667.

Melo, C. d., Cruzes, D. S., Kon, F., & Conradi, R. (2013). Interpretative case studies on agile team productivity and management. Information and Software Technology, 55, 412-427.

Moe, N. B., Dingsoyr, T., & Dyba, T. (2009, November 20). A teamwork model for understanding an agile team: A case study of a Scrum project. Information and Software Technology, 52, pp. 480-491.

Rico, D. F., & Sayani, H. H. (2009). The Business Value of Agile Softare Methods: Maximizing ROI with Just-in-Time Process and Documentation. Fort Lauderdale: J. Ross Publishing.

Rogers, R. O. (2004). Acceptance Testing vs. Unit Testing: A developer's perspective. LNCS (3134), 22-31.

Sampaio, A., Vasconcelos, A., & Sampaio, P. F. (2004, November). Assessing Agile Methods: An empirical study. Journal of the Brazilian Computer Society, 10(3).

Serrador, P., & Pinto, J. K. (2015). Does Agile work? - A quantitative analysis of agile project success. International Journal of Project Management, 33, 1040-1051.

Appendix

	Waterfall methodology	Agile methodology
Core delivery process	Follows the system development life cycle with requirements analysis, design, development, testing, implementation, and support pre-defined cycles with a fixed end date	Iterative requirements, design, and development with continuous integration and deployment activities and iterative release cycles
Stakeholder Involvement	Fixed project scope with stakeholder approver before design and development can begin; stakeholder review of final product	Changing scope per sprint iteration with stakeholder input on requirements and prioritization; stakeholder review of final product
Development	Based on fixed requirements and overall architectural design	Performed while determining requirements and test-driven
Testing	Largely manual execution, tied directly to business requirements, performed as several phases (unit, integration, user acceptance, performance)	Automated, behavior-driven, performed as part of development
Documentation	Business and system requirements are written and approved by stakeholders; testing scenarios and results are documented with traceability to requirements	Requirements are loosely documented as sprint stories; testing scripts are sometimes tied to stories

Table 1
Methodological distinctions between Waterfall and Agile project methodologies

Conceptualization of Blockchain-Based Applications: Technical Background and Social Perspective

Jason Xiong
xiongjj@appstate.edu
Department of Computer Information Systems
Appalachian State University
Boone NC

Yong Tang
tangyong@uestc.edu.cn
Center of Cyberspace and Security,
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Dawn Medlin
medlinbd@appstate.edu
Department of Computer Information Systems
Appalachian State University
Boone NC

Abstract

There has been an increasing amount of research into blockchain when blockchain is receiving increasing enthusiasm from both the practitioners and scholars. It has been revolutionary in bringing trustless computing and immutable decentralized ledger of digital assets and transactions to businesses, organizations, and individuals. The decentralized nature of blockchain-based systems and applications eliminates intermediaries, saves costs, and enhances efficiencies. In its infancy, blockchain demonstrates high disruptive potentials in many areas such as finance, healthcare, education, and real estate, among others. This research presents the conceptualization of blockchain-based applications from the Information Systems perspective. The concepts of blockchain, smart contract, ICO, Dapps, DAOs, DAC, DAS, and AM are discussed and analyzed. The technical background and social perspectives of the blockchain-based applications are discussed and analyzed as well. The paper contributes to a framework of blockchain level from the Information Systems perspective.

Keywords: Blockchain, Decentralization, Cryptocurrency, Bitcoin, Conceptualization, Smart Contract.

1. INTRODUCTION

Blockchain, a distributed ledger technology, is attracting attention from industries to academics as a disruptive technology with tremendous potentials throughout a vast range of applications. With a short history of less than a decade (Gupta, 2017; Schlegel, Zavolokina, & Schwabe, 2018; Sompolinsky & Zohar, 2018), blockchain is believed to have brought an unprecedented decentralization revolution to not only the technology field but also to the structure

of human society (Baruffaldi & Sternberg, 2018; Lu & Zheng, 2018; Sadhya & Sadhya, 2018; Swan, 2015; Tapscott & Tapscott, 2016a).

As the underlying technology of bitcoin and other cryptocurrencies introduced by Satoshi Nakamoto (Nakamoto, 2008), blockchain is an open-source technology with tremendous successes regarding worldwide acceptance, trading volumes, and applications (Joseph Cook, 2014; Mandjee, 2014; Tapscott & Tapscott, 2016b; Underwood, 2016). Some researchers suggest that blockchain is expanding as a disrupting force with applications in finance, health, government, society, business, politics and more (Agyepong, 2016; Beck & Müller-Bloch, 2017; Eldred, 2016; Hurlburt, 2016; Lee, James, Ejeta, & Kim, 2016; Mougayar, 2016; Post, Smit, & Zoet, 2018; Sadhya & Sadhya, 2018; Swan, 2015; Wörner & Bilgeri, 2016; Yue, Wang, Jin, Li, & Jiang, 2016).

Recently, increased attention has been paid to big data analytics (Günther, Rezazade Mehrizi, Huysman, & Feldberg, 2017; Loebbecke & Picot, 2015). The blockchain serves as the backbone of big data analytics, and its unparalleled potentials and challenges should never be underestimated. As blockchain startups emerge and the involvement of significant technology companies increases, the technological ecosystem has significantly evolved with the support of venture capitalists and organizations (Friedlmaier, Tumasjan, & Welpe, 2016). In 2016, a stunning $500M venture capital was fueled into blockchain initiatives, which indicate high confidence in this area of investments (Asatryan, 2017).

Researchers from engineering, business, and the social sciences are investigating the innovations of the fast-evolving blockchain industries. While opportunities exist with the adoption and innovations surrounding blockchain, others call for clarity of the revolutionary issue of blockchain. Challenges exist in the rise of blockchain development, especially after scandals like bitcoin stealing, abuses in the areas of illegal drug trading, and money laundering.

These instances have often created a negative reputation for both bitcoin and cryptocurrencies. The dark side of blockchain, together with the fluctuations in bitcoin values have led some to believe in a conspiracy theory that bitcoin is a Ponzi scheme that only benefits the initial investors.

However, as a cryptocurrency, bitcoin has been built upon the confidence and adoption of all investors. It does not fit the definition of a Ponzi scheme in many ways, especially given the fact that the bitcoin ecosystem does not pay rewards for new recruitments or participation. The value of bitcoin only depends on supply and demand. Just like other technologies, the underlying technology of blockchain does not take a side and is not controlled by any other party, including the initial investors.

Therefore, it is essential to clarify the misunderstandings and to build confidence for blockchain decision-makers (Beck & Müller-Bloch, 2017). This confidence requires more research into blockchain and its innovative ecosystem (Lindman, Chalmers, & Rossi, 2017), including the archetypes and the economic perspective of the blockchain (Catalini & Gans, 2016; Walsh, O'Reilly, Gleasure, Feller, Shanping, & Cristoforo, 2016).

Blockchain is a fast-evolving field with both excitement and doubt. Though this application is immature, many innovations can harvest the uniqueness of blockchain. Several industries are leading the applications of blockchain, but gaps exist as well. Therefore, academicians have begun to address these gaps in research through a systematic and interdisciplinary approach from the fields of technology, economics, social sciences, business, and philosophy.

This research offers a conceptualization of blockchain-based applications from the IS perspective. Overall, the research question is ***How to conceptualize constructs of applications based on blockchain from the IS perspective***?

The paper is organized as follows. After the introduction, the second section provides a conceptual perspective literature review of blockchain. Key concepts, including blockchain, smart contracts, ICO, Dapps, DAOs, DAC, DAS, and AM are introduced. The third section provides the technical background review of the blockchain. Different technical concepts, including distributed Nodes, proof of work, and security are discussed. The fourth section provides an analysis of the social perspective of blockchain. The fifth section presents the different conceptual levels of blockchains. The paper ends with a conclusion and future research.

2. TECHNICAL BACKGROUND OF BLOCKCHAIN

Blockchain is a database recording of all historical transactions with consensus among all parties and without a central authority. Distributed nodes blockchain is a distributed ledger chronologically stored in nodes that provide verification and storage services for the entire network.

Distributed Nodes
The nodes can be any computing entities such as computers, servers in the cloud, Internet of Things (IoT) devices, or specially designed chips running within mining pools. The nodes communicate with each other in a peer-to-peer method and gain rewards by providing confirmation services of transactions that are occurring within the network.

To fake transactions, one needs to control 51% of the nodes, which is almost impossible for an active blockchain with a large number of nodes. The nodes are designed to run as a self-governed organization without relying on a specific central node, thus avoiding single-point failures within centralized systems.

All distributed nodes maintain identical records ensuring that the transactions are stored equally and therefore are resistant to attacks and are free of traditional reconciliation and audition.

Proof of work
Proof of work is a computation of adding new transaction information into a blockchain. Using brute force, this process requires tremendous computing power. In bitcoin, it is compensated by rewarding bitcoins to the computers or miners. This process grants the economic values to proof of work for the mined coins. Since all transactions are required to be verified before merging into a block, the blockchain relies heavily on proof of work.

Thus, a single correct chain with all verified historical transaction records is guaranteed without the possibility of multiple blockchains with more than one version. For public blockchains like bitcoin, there is an abundance of nodes running as nodes anytime, while for some private blockchains, specific nodes are needed to be deployed and keep running in order to verify transactions happening in the network.

Proof of work provides a trustless consensus among multiple parties. However, it has shortcomings. For instance, proof of work requires a large amount of computing power consumption, slow speeds, and has a risk of a 51% attack. New consensus algorithms have been proposed, such as proof of stake, proof of activity, and proof of capacity to address these shortcomings. For example, instead of providing computation services to get incentives in proof of work, nodes could be used to invest coins to verify new transactions in proof of stake.

Security
Cryptography provides the underlying security for blockchain. However, there are still security concerns. Verbücheln (2015) uses cryptographic proof to replace the need for the involvement of a trusted third party.

All previous transactions are hashed as well as all-new transactions are also hashed as a Merkle tree. The Merkle tree is a tree of hashed values of pairwise transactions allowing a fast location of a specific transaction or the identification of a modified transaction.

A nonce is brutally calculated using SHA-256 begins with a number of zero bits (Gilbert & Handschuh, 2003). In this way, the proof of work is achieved by the CPU time and the consumed electricity. All blocks are linked by hash values in order of time, and an attacker needs to modify one block and all blocks afterward, thus making it practically impossible and uneconomical since the same computation efforts are required to create a fake blockchain.

Thus, rational choice would be an honest miner. Meanwhile, the pseudonymous characteristic of blockchain ensures the users' privacy and identity are protected without risk of exposure to potential attackers or trackers.

In the scenario of decentralized energy trading, Zhumabekuly Aitzhan and Svetinovic (2016) discussed an approach combining blockchain, multi-signatures, and anonymous encrypted messaging streams. This method indicates that blockchain-based systems can utilize other security and privacy methods to provide application-level enhanced protection.

3. CONCEPTUAL PERSPECTIVE OF BLOCKCHAIN

Blockchain
Blockchain originated as the open ledger of all transactions for bitcoins stored across nodes in the decentralized peer-to-peer networks that can exist beyond geographic boundaries and authoritative controls. Blockchain can be viewed as a giant public accessible registry to record information, assets, and transactions, which are verifiable and transparent for all (Beck, Avital, Rossi, & Thatcher, 2017).

Blockchain is a continuously growing time-stamped record. However, the conception of blockchain is far richer than technology implementation as a distributed ledger technology. The richness of blockchain and the potentials for business, as well as the political aspects of human society, brings new level concepts like smart contracts, Dapps, DAO, DAC, DAS, and AM into play (Swan, 2015).

Smart Contracts
Smart contracts are digitized agreements between two or more parties programmed on a blockchain (Fairfield, 2014). Unlike the paper-based contracts, which are agreed upon by parties and legalized by authorities, smart contracts are coded as running programs that can be automatically executed once the preset conditions are met, thus allowing exchanges of a digital or physical asset.

Due to the decentralized mechanism of blockchain, the contracts are ensured to be honored and the whole process is executed without relying on certain authorities that require validation. Smart contracts can be coded in commonly used procedural languages as well as logic-based languages (Idelberger, Governatori, Riveret, & Sartor, 2016).

Running on the underlying blockchain, smart contracts allow the parties to be humans, machines, organizations, and even other contracts. This feature dramatically enriches the concept of contracts and dramatically increases the features of some applications. The disruptive potentials of blockchain largely rely on how innovative applications of smart contracts are (Peters & Panayi, 2016).

Initial Coin Offerings (ICO)
ICOs are known traditionally as Initial Public Offerings (IPOs) in order to sell shares in exchange for funds from investors. Similarly, for a cryptocurrency startup, it can sell its cryptograph coins or tokens to initial investors to raise funds for a specific product.

ICO is an innovative financing schema for blockchain-based startups and is different from crowdfunding. In an ICO, a preset target is identified and agreed upon by all investors. If, and only if the specified target is met, the ICO is declared a success and the startup formally becomes operational. Otherwise, the ICO fails, and all investments are returned. The key to a successful ICO is the acknowledgment and acceptance of the campaign.

Considering the open competition among ICOs and the driving forces of financial incentives, the market will automatically evolve a natural selection for competitive ICOs and eliminate inferior products. All is done through a blockchain without the involvement of traditional brokers, underwriters, central exchange markets, or regulating bodies, and without any significant costs.

The Ethereum (Wood, 2014), a project of a decentralized application platform, launched a successful ICO where 18 million dollars was raised that offered large returns for initial investors. The appearance of blockchain-based ICOs provides hints for how traditional financing activities can be changed and how decentralized economies can work, thus requiring serious discussions of present business leaders, innovators, and regulators.

Decentralized Applications (Dapps)
Dapps are services running on blockchain and are decentralized applications. Applications range from finance, banking, e-commerce, social networks, file sharing, property sharing, among others (Agyepong, 2016; Guo & Liang, 2016; Peters & Panayi, 2016), which generally have respective counterparts in traditional centralized cyberspaces.

As innovations based on this blockchain technology are created, these Dapps demonstrate the enormous business potential in a much broader scope beyond that of the financial industry. Many startups are providing innovative Dapps solutions to disrupt established business models and traditional business processes (Raval, 2016).

Dapps can also be designed to revolutionize sharing economic models (Puschmann & Alt, 2016) by merging blockchain with the IoT (Huckle, Bhattacharya, White, & Beloff, 2016).

Decentralized Autonomous Organizations (DAOs)
DAOs allow multiple parties to reach an agreement on internal structures, rules, and collective missions. Internal organization constitutions and external laws can enforce the authority of the agreement.

Traditionally, it is challenging to build temporary, geographically distributed organizations. Powered by blockchain, the whole life cycle of an organization can be implemented as multi-party smart contracts. DAOs are innovative for societal issues by redefining and reconstructing the mechanisms of an organization.

DAOs are open-sourced, transparent, run in an automated environment by codes without controls from dominating centers, and thus the collective intelligence can be utilized and maximized into actions that are free of trust issues. Blockchain-based DAOs also provide trust and identity for sharing economically-based applications (Jarvenpaa & Teigland, 2017; Puschmann & Alt, 2016).

Decentralized Autonomous Corporation (DAC)

For-profit organizations, especially those business-orientated commercial corporations, can be reinvented as Decentralized Autonomous Corporation, or DAC, on the blockchain. The essentials of corporate governance and operations can be fully programmed in contracts deployed on the blockchain with a full set of functionalities and capabilities in order to conduct business with external entities.

The DACs are natural alternative business forms for people to conduct pure global or semi-business activities with self-defined corporation constitutions and autonomous business processing while remaining free from bureaucratic costs. With the emerging of DAC, there is a lack of legal regulations for DACs. This remains a challenge for the blockchain community, government, and lawmakers.

Decentralized Autonomous Society (DAS)
DAS is a collection of entities connecting and interacting with each other in order to exchange resources within certain structures. Since the individual entities are based on blockchain, they are autonomously running as sets of smart contracts in a manner of decentralization, and without human interference.

This is not an updated highly autonomous system built today to speed up processes, but rather a massive and pervasive DAO and DAC that will define a fundamentally completely new DAS.

Automatic Markets (AM)
AM are the future driving forces through which resources can be allocated. Trades among DAOs and DACs can create an AM in which ownerships are exchanged and resources are consumed. For instance, and in relation to smart properties, the underlying resources encoded as smart properties can be rights, options, and utilities, as well as physical or non-tangible goods.

Trade in an automatic market is realized once a smart contract is satisfied with preset conditions. The signals can be outcomes of other smart contracts, the output of legendary systems, as well as real-time data from machine networks or the Internet of Things (IoT). With emerging DAOs and DACs, automatic markets are inevitably bringing new business models and impacts on the traditional centralized economy paradigms.

4. PUBLIC, CONSORTIUM, AND PRIVATE BLOCKCHAIN

When designing a blockchain-powered system, it is important to choose the right blockchain solution. In terms of permission and accessibility, it is possible to deploy the system over either a public chain, private chain, or a consortium chain.

Public Blockchain
The underlying blockchain of bitcoin is a typical public blockchain with equal accessibility for all participants. The identical version of blocks is stored in a distributive manner, crossing all nodes and not relying on specific nodes. The nodes are free to leave or join anytime without significant impacts on the running performance of the blockchain. The information stored on public blockchains is transparent without geographic or organizational restrictions.

Private Blockchain
Opposite to public blockchains, private blockchains are ledgers running in a closed environment and usually within an organization. Private blockchains are only transparent for permitted participants according to access controls.

The whole computation facilities and software are owned by organizations, providing an isolated and secure blockchain infrastructure that is built to support advanced applications. Since the blockchain is restricted to an organization, the data shared in the blockchain is suitable for sensitive data.

Consortium Blockchains
Consortium blockchain offers limited access to selected organizations that are identified as consortiums. A consortium blockchain is maintained and accessible by participants within the consortium with possible controlled accessibility to outsiders.

	Public	Consortium	Private
Permission	Permissionless	Permitted	Permitted
Identity	Pseudonymous	Non-anonymous	Non-anonymous
Data confidentiality	Low	High	High
Nodes ownership	All	Members	Organization
Governance	Decentralized	Decentralized	Centralized
Maintenance	free	Shared by participants	Organization
Mining cryptocurrency	Bitcoin	Not necessary	Not necessary
Efficiency	Low	High	High
Use scenario	Public	Organizational collaboration	Internal process
Example	Bitcoin	Bank clearing services	City e-government

Table 1. Comparisons of public, consortium, and private blockchains

In Table 1, comparisons of the public, consortium, and private blockchains are presented. The differences between these three kinds of blockchains require decision-makers to decide which type of blockchain is suitable for their business models.

For private blockchains and consortium blockchains, the blockchain infrastructures are owned and controlled by pre-selected participants within in a single organization or organizations in the consortium, however, on the contrary, public blockchains are fully open for anyone and not owned by specific participants.

This difference of ownerships brings misunderstandings of believing the private or consortium blockchains are compromised blockchains, for the centralization of ownerships and controls which is against the nature of the decentralization of blockchain. However, this misunderstanding is rooted in the misinterpretation of the meaning of decentralization which is more about decentralized transaction processing rather than the technology implementation.

In other words, the private and consortium blockchains are still decentralized ledgers. This is from the nature of decentralized transactions processing and how data is shared. The decentralization of underlying blockchain infrastructure is not necessary and certain full or semi control of accessibility is necessary and indispensable for some scenarios.

5. SOCIAL PERSPECTIVE OF BLOCKCHAIN

Blockchain technology and innovative applications are pushing their way into many domains of human society with promising benefits. Before its full implementation into governments, businesses, and societies where individuals are relying on systems, applications, infrastructures, and algorithms powered by blockchain, it is important to study the impacts and implications related to both positive and negative consequences.

As a decentralized public ledger, the trust in an untrusted environment is achieved by algorithms running on machines without relying on human judgments. This machine trust can avoid any human or organization errors as well as malicious damages.

Additionally, this revolution is a strong advancement for human society, which has been suffering the high costs and inconveniences of maintaining and ensuring hierarchic management structures that only provided authoritative trust systems. Now, due to the adoption of blockchain, current trust systems can be partially or entirely replaced by algorithmic ensured trust systems.

For example, by analyzing the functions of a cryptocurrency-based monetary system, it can provide the monetary authority and work as the clearinghouse, while needs outside solutions for resort lender (Guo & Liang, 2016; Hayes, 2016; Peters & Panayi, 2016). Thus, a technocracy requires no human interventions and can be free of human weakness, frailties, and limitations.

Intelligent process automation

Intelligent process automation in relation to blockchain allows for transactions and verification of digital assets, which can be automatically processed by smart contracts and other decentralized applications.

This advantage can save time and cost for service providers and consumers as well as provide more efficiencies. A highly automatic environment can free employees from repetitive procedures and allow them to participate in more creative and fulfilling activities.

However, the automation process can also bring changes to job positions and responsibilities. Though the overall effects may appear positive and tempting and the changes also appear unstoppable, there are certainly challenges that exist.

6. CONCEPTUAL LEVELS OF BLOCKCHAIN

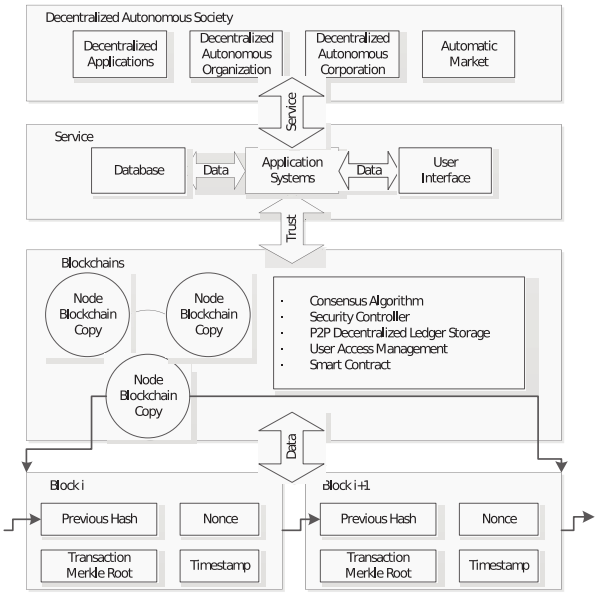


Figure 1. The Framework of Blockchain Level

To summarize the above discussions of blockchain conceptions, we propose a four-level conceptional framework.

In Figure 1, we illustrate the conceptional framework of blockchain in different levels from an underlying block structure to the decentralized autonomous society: (1) In the block level, blocks are chained in chronicle order with data of the previous block hash, the Merkle root of transactions, timestamp, and the mining nonce. (2) In the blockchain level, the whole blockchain is verified by independent computing nodes providing consensus, security, ledger storage, access control, and the running environment of smart contracts. (3) In the service level, blockchain and a traditional database is integrated into application systems with user interfaces to provide services of certain functionalities. (4) All blockchain services then form the background to support the decentralized autonomous society level in which decentralized applications, organizations, corporations, and markets are gathered with unprecedented business, management, organizational, and social values.

7. CONCLUSIONS AND FUTURE RESEARCH

This research presents the conceptualization of blockchain-based applications from the Information Systems perspective. Based on the conceptional framework we proposed, there are many questions that remain unanswered. Based on the proposed conceptual levels of blockchain, we further plan to provide a systematic mapping and provide several potential research questions for IS researchers in the future.

8. REFERENCES

Agyepong, S. (2016). Facilitating Financial Inclusion Using Ict : Lessons From M-Pesa and E-Zwich. Paper presented at the Twenty-Fourth European Conference on Information Systems (ECIS), Istanbul, Turkey.

Asatryan, D. (2017). Blockchain VC Investment Nears $500M in 2016. In: http://bankinnovation.net/2017/03/blockchain-vc-investment-nears-500m-in-2016/.

Baruffaldi, G., & Sternberg, H. (2018). Chains in Chains - Logic and Challenges of Blockchains in Supply Chains. Paper presented at the 51st Hawaii International Conference on System Sciences, HICSS 2018, Hilton Waikoloa Village, Hawaii, USA, January 3-6.

Beck, R., Avital, M., Rossi, M., & Thatcher, J. B. (2017). Blockchain Technology in Business and Information Systems Research. In: Springer.

Beck, R., & Müller-Bloch, C. (2017). Blockchain as Radical Innovation : A Framework for Engaging with Distributed Ledgers. Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences.

Catalini, C., & Gans, J. S. (2016). Some Simple Economics of the Blockchain. SSRN Electronic Journal. Retrieved from http://www.nber.org/papers/w22952.pdf http://www.ssrn.com/abstract=2874598

Eldred, M. (2016). Blockchain Thinking and Euphoric Hubris [Letter to the Editor]. IEEE Technology and Society Magazine, 35(1), 39.

Fairfield, J. (2014). Smart Contracts, Bitcoin Bots, and Consumer Protection. Washington and Lee Law Review Online, 71(2), 35-50. Retrieved from http://scholarlycommons.law.wlu.edu/wlulr-online/vol71/iss2/3

Friedlmaier, M., Tumasjan, A., & Welpe, I. (2016). Disrupting industries with blockchain : The industry , venture capital funding , and regional distribution of blockchain ventures. In.

Gilbert, H., & Handschuh, H. (2003). Security analysis of SHA-256 and sisters. Paper presented at the International workshop on selected areas in cryptography.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems, 26(3), 191-209. doi:https://doi.org/10.1016/j.jsis.2017.07.003

Guo, Y., & Liang, C. (2016). Blockchain application and outlook in the banking industry. Financial Innovation, 2(1), 24. Retrieved from http://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0034-9

Gupta, V. (2017). A Brief History of Blockchain. Harvard Business Review, 2-4. Retrieved

from https://hbr.org/2017/02/a-brief-history-of-blockchain

Hayes, A. (2016). Decentralized Banking: Monetary Technocracy in the Digital Age. In P. Tasca, T. Aste, L. Pelizzon, & N. Perony (Eds.), Banking Beyond Banks and Money: A Guide to Banking Services in the Twenty-First Century (pp. 121-131). Cham: Springer International Publishing.

Huckle, S., Bhattacharya, R., White, M., & Beloff, N. (2016). Internet of Things, Blockchain and Shared Economy Applications. Procedia Computer Science, 98, 461 - 466. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877050916322190

Hurlburt, G. (2016). Might the Blockchain Outlive Bitcoin? IT Professional, 18(2), 12-16.

Idelberger, F., Governatori, G., Riveret, R. e., ,gis, & Sartor, G. (2016). Evaluation of Logic-Based Smart Contracts for Blockchain Systems. In J. J. Alferes, L. Bertossi, G. Governatori, P. Fodor, & D. Roman (Eds.), Rule Technologies. Research, Tools, and Applications: 10th International Symposium, RuleML 2016, Stony Brook, NY, USA, July 6-9, 2016. Proceedings (pp. 167-183). Cham: Springer International Publishing.

Jarvenpaa, S., & Teigland, R. (2017). Trust in Digital Environments : From the Sharing Economy to Decentralized Autonomous Organizations. Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences.

Joseph Cook, R. (2014). Bitcoins: Technological Innovation or Emerging Threat? The John Marshall Journal of Information Technology & Privay Law, 30(3), 535-570. Retrieved from http://repository.jmls.edu/jitpl/vol30/iss3/4

Lee, K., James, J. I., Ejeta, T. G., & Kim, H. (2016). Electronic Voting Service Using Block-Chain. Journal of Digital Forensics, Security and Law, 11(2), 123-136.

Lindman, J., Chalmers, G., & Rossi, M. (2017). Introduction to Open Digital Services and Platforms Minitrack. Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences, Hawaii.

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. The Journal of Strategic Information Systems, 24 %6(3), 149 - 157 %&.

Lu, Y., & Zheng, X. (2018). Block Chain Based Double Auction Design. Paper presented at the 24th Americas Conference on Information Systems.

Mandjee, T. (2014). Bitcoin, Its Legal Classification and Its Regulatory Framework. Journal of Business {\&} Securities Law, 15(2), 1.

Mougayar, W. (2016). The Business Blockchain: Promise, Practice, and Application of the Next Internet Technology: John Wiley \& Sons.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Www.Bitcoin.Org.

Peters, G. W., & Panayi, E. (2016). Understanding Modern Banking Ledgers Through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money. In P. Tasca, T. Aste, L. Pelizzon, & N. Perony (Eds.), Banking Beyond Banks and Money: A Guide to Banking Services in the Twenty-First Century (pp. 239-278). Cham: Springer International Publishing.

Post, R., Smit, K., & Zoet, M. (2018). Identifying Factors Affecting Blockchain Technology Diffusion. Paper presented at the 24th Americas Conference on Information Systems.

Puschmann, T., & Alt, R. (2016). Sharing Economy. Business and Information Systems Engineering, 58(1), 93-99.

Raval, S. (2016). Decentralized applications: harnessing Bitcoin's blockchain technology: " O'Reilly Media, Inc.".

Sadhya, V., & Sadhya, H. (2018). Barriers to Adoption of Blockchain Technology. Paper presented at the 24th Americas Conference on Information Systems.

Schlegel, M., Zavolokina, L., & Schwabe, G. (2018). Blockchain Technologies from the Consumers' Perspective: What Is There and Why Should Who Care? Paper presented at

the 51st Hawaii International Conference on System Sciences, HICSS 2018, Hilton Waikoloa Village, Hawaii, USA, January 3-6, 2018.

Sompolinsky, Y., & Zohar, A. (2018). Bitcoin's Underlying Incentives. Commun. ACM, 61(3), 46-53. Retrieved from http://doi.acm.org/10.1145/3152481

Swan, M. (2015). Blockchain: Blueprint for a new economy: O'Reilly Media, Inc.

Tapscott, D., & Tapscott, A. (2016a). Blockchain Revolution: How the Technology Behind Bitcoin Is Changing Money, Business, and the World: Penguin.

Tapscott, D., & Tapscott, A. (2016b). The impact of the blockchain goes beyond financial services. Harvard Business Review.

Underwood, S. (2016). Blockchain beyond bitcoin. Communications of the ACM, 59(11), 15-17.

Verbücheln, S. (2015). How Perfect Offline Wallets Can Still Leak Bitcoin Private Keys. Paper presented at the Proceedings of Mediterranean Conference on Information Systems (MCIS).

Walsh, C., O'Reilly, P., Gleasure, R., Feller, J., Shanping, L., & Cristoforo, J. (2016). New kid on the block: a strategic archetypes approach to understanding the Blockchain. Paper presented at the 37th International Conference on Information Systems (ICIS).

Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper, 151(2014), 1-32.

Wörner, D., & Bilgeri, D. (2016). The Bitcoin Ecosystem: Disruption Beyond Financial Services? Paper presented at the Proceedings of the Twenty-Fourth European Conference on Information Systems (ECIS), Istanbul, Turkey.

Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. Journal of Medical Systems, 40(10), 218.

Zhumabekuly Aitzhan, N., & Svetinovic, D. (2016). Security and Privacy in Decentralized Energy Trading through Multi-signatures, Blockchain and Anonymous Messaging Streams. IEEE Transactions on Dependable and Secure Computing, PP(99), 1-14.

Addressing issues with EMR resulting in workarounds: An exploratory study

Sushma Mishra
mishra@rmu.edu
Computer and Information Systems Department
Robert Morris University
Moon Township, PA 15108 USA

Kevin Slonka
slonka@pitt.edu
Information Sciences Department
University of Pittsburgh
Greensburg, PA 15601 USA

Peter Draus
draus@rmu.edu

Natalya Bromall
bromall@rmu.edu

Computer and Information Systems Department
Robert Morris University
Moon Township, PA 15108 USA

Kelli Slonka
kstanisl@conemaugh.org
Pulmonary Medicine Department
Conemaugh Memorial Medical Center
Johnstown, PA 15905 USA

Abstract

The goal of this study is to understand how the use of Electronic Medical Records (EMR) systems impacts the experience of the healthcare service provider in delivering patient care. Surveying and interviewing the hospital workers helped to answer this question and to determine the underlying factors that lead to healthcare providers' deviating from the protocol measures. Finally, specific recommendations were made in improving both healthcare providers' and patients' experience with EMR.

Keywords: Electronic Medical Records (EMR), Electronic Health Records (EHR), healthcare information systems, survey, health care providers, workarounds

1. INTRODUCTION

It is impossible to underestimate the importance of Electronic Medical Records (EMR) systems. In the past two years, 86% of all physicians' offices used EMR in some form. For the hospitals, this number is 97% (National Coordination of Health IT, 2018).

While the adoption of EMR increases, there are multiple issues with its use. Nurses, as the top category of EMR users (Sackett et al., 2006) report encountering with these issues daily: entering a patient's record, sending prescriptions, integrating the data from multiple EMRs, and such. In many cases, the inability of the nurses to perform critical operations promptly can lead to policy compliance and workarounds (Dudding, Gephart, and Carrington, 2018).

This study is the continuing step of the research by Draus et al., 2019, where the authors found the general areas of concerns related to the use of EMR. In this study, interviews were conducted with various healthcare providers as the primary users of EMR, to determine the specific areas in EMR that need improvement. According to Dudding, Gephart, and Carrington (2018), some issues in EMR, such as constant interruptions, changes in communication patterns and workflow, may lead to healthcare providers' using workarounds. Another goal of the study was to determine such issues and give recommendations on how they can be improved.

The following research questions were answered in the study:

RQ1: How does use of EMR impact the experience of the healthcare service provider in delivering patient care?

RQ2: What are the underlying factors that lead to health care providers taking off protocol (not prescribed by EMR) measures to get their job done?

RQ3: How can the issues with EMRs be addressed to make a health care service provider's experience better to improve patient outcome?

2. REVIEW OF LITERATURE

Machinery has existed and been used in hospitals in the form of life-supporting devices for quite some time. Information technology use in the medical field, however, has only recently blossomed. In the late 1990s, with the release of the American Academy of Colleges of Nursing's "The Essentials of College and University Education for Professional Nursing," the use of information technology in hospitals for informatics, the category under which Electronic Medical Records (EMR) fall, began. The core concept of this document was placing strategic value upon healthcare informatics, specifically in the nursing field. The impact was counter to its standard view as a mere resource, and in opposition of the federal funding, trend instituted four years earlier lowering the amount of money available for implementing informatics systems (Sackett, Jones, & Erdley, 2005).

There is somewhat of a debate on the definition of the term EMR. The confusion with the term is due to a similar term existing, Electronic Health Record (EHR). The EMR system is defined as a "digital version of the paper charts in the clinician's office" (Garrett & Seidman, 2011, para 5). Whereas the EMR is merely a digital version of a paper chart, the EHR builds upon this by allowing use by many different healthcare providers, such that all providers involved in a patient's care can record and share information. The list includes providers that are not Primary Care Physicians (PCP), such as sports medicine clinicians, chiropractors, etc. (Garrett & Seidman, 2011).

The Health Insurance Portability and Accountability Act was the impetus for the adoption of EMR systems (Government Publishing Office, 2018; Adler-Milstein & Jha, 2017), but despite the government mandate electronic systems offer the medical field many benefits. Friedman, Parrish, & Ross (2013) noted many benefits due to electronic systems making records available quickly and from anywhere: the ability for providers to measure disease level/distribution, report and investigate notifiable diseases, have access to complete, longitudinal patient records, and gain timely access to patient records. While these specific benefits are notable, one must be able to see the forest despite the trees. Improved efficiency, better patient care, and higher patient safety are the higher-level contributions achieved through the benefits above (Institute of Medicine, 2003).

EMR Adoption and Champions

In four years, from 2008 to 2012, it was found that EMR systems have been adopted (at a

minimal level) at a rate that rose from 9% to 44%. Although this increase in adoption seems to suggest a positive trend, it is important to note that the percentage of hospitals with only a basic EMR system is much lower than the percentage of hospitals that do not have any electronic system at all (27.3% versus 56%). In only one case do we see that the percentage of hospitals that have adopted a basic EMR system is higher (by 4.9%) than those that have no electronic system: major teaching hospitals (members of the Council of Teaching Hospitals). Although the outlook may look dire, many hospitals do use electronic systems for essential public health functions: submitting syndromic data to public health agencies, submitting data to immunization registries, and submitting lab reports to public health agencies is done by more than 50% of hospitals (DesRoches et al., 2013).

To increase adoption of electronic systems, it is essential to understand which user groups are the champions of such projects. Healthcare professionals (not administrators), such as nurses (Sackett, Erdley, & Jones, 2006) and nursing leaders (Edwards, 2012), are significant drivers of adoption by impacting their "reception, design, development, and implementation" (p. 111). There are other champions of electronic system adoption, such as PCPs, due to the electronic system's ability to help with PCP efficiency and workload productivity (Bae & Encinosa, 2016; Xierali et al., 2013). Other healthcare professionals can also drive adoption of electronic systems, such as dentists, pharmacists, physical therapists, and allied health professionals. It is crucial, however, to note that adoption is more vehemently pursued by professionals when they have a specific need for such systems; but, even so, sometimes the requirement does not outweigh the perceived negative factors of EMR implementation for some professionals (Acharya, Schroeder, Schwei, & Chyou, 2017; Fuji, Galt, Siracuse, & Christoffersen, 2011; Wang & Biedermann, 2012; Yung, 2017).

Acceptance Issues

Despite the numerous user groups with the ability to pursue the implementation of EMR systems, such systems are not always immediately accepted within the institution. The most common issues are "limited or no access to computers, fear of change, nurses too busy to use computers, and nurses don't like computers" (Sackett, Erdley, & Jones, 2006, p.251; Gesulga, Berjame, Moquiala, & Galido, 2017). The common theme with these issues is the healthcare professional's ability to use technology. A key acceptance factor that emerges is user-friendliness (Aldosari, Al-Mansour, Aldosari, & Alanazi, 2018; Sidek & Martins, 2017; Gagnon et al., 2014). Healthcare professionals' information knowledge is at a low level; education is "urgently required" (Syoubuzawa, Yamanouchi, & Takeda, 2006, p. 819), especially for those professionals 30 years of age and above. Not only should professionals be educated on the EMR systems themselves, but due to the nature of the information with which they work, it is critical to understand the ease accessing and distributing this information, which can lead to the blurring of ethical lines (Aluas, 2016). As long as healthcare professionals are adequately trained, most would recommend electronic systems over the traditional paper systems (Choi, Chung, & Lee, 2006).

EMR Deficiencies and Workarounds

Although the satisfaction of healthcare professionals is essential, patient safety is an equally critical higher-level contribution of EMR systems. Bar, Rask, & Becker (2018) found that patient safety events can be significantly decreased by implementing EMR systems produced by a single vendor. One, however, should not ignore the previously exposed implementation issues. When healthcare professionals view the electronic system as adding "new additional steps in [their] workflow" (Patterson, 2018, p. 281), they find ways to bypass such steps in order to maintain their situational awareness of their patient load, such as resorting to using non-electronic means (e.g., paper, whiteboards, etc.) (Stevenson, Israelsson, Nilsson, Petersson, & Bath, 2016). Harkening back to patient safety, such "workarounds" are typically viewed by non-healthcare professionals as a cause for concern (Halbesleben, Wakefield, & Wakefield, 2008; Meeks et al., 2014; Stutzer & Rushton, 2015). Healthcare professionals, on the other hand, merely view these workarounds as "work patterns [... created] to accomplish a crucial work goal within a system of dysfunctional work processes that prohibit the accomplishment of that goal or makes it difficult" (Morath & Turnbull, 2005, p. 52). This pattern of behavior confirms the Adaptive Structuration Theory, which states that groups will evolve the technology to fit their needs (Barrett, 2017) better. Though previously described as unfavorable, workarounds are not always such. Barrett & Stephens (2017) argue that the use of workarounds can lead to lower resistance to

using EMR systems and increased perception of their success.

Many studies had investigated the reasons for the use of the EMR system workarounds by healthcare professionals. Heavier workloads, work interruptions, and altered communication patterns were exposed as crucial motives behind the use of workarounds (Dudding, Gephart, & Carrington, 2018; Assis-Hassid, Grosz, Zimlichman, Rozenblum, & Bates, 2019; Rathert, Porter, Mittler, & Fleid-Palmer, 2019). One study suggested that more errors are made due to the drop in critical thinking skills promulgated by built-in automated processes/instructions in EMR systems (Pagulayan, Eltair, & Faber, 2018). On the other end of the spectrum, there are healthcare professionals who refuse to use EMR systems due to their suboptimal design. Instead, they hire medical scribes to shadow them throughout the day and handle the secretarial duties of entering data into the EMR system. This delegation of work to secretarial staff, too, can be viewed as a workaround (Schiff & Zucker, 2016). Others, however, have negative feelings toward EMR systems due to the belief that said systems were designed more for hospital administrators than to assist healthcare professionals in providing better patient care (Eason & Waterson, 2013). The contradiction in sentiments is related to the professional's expected versus perceived user-friendliness of the electronic system, due to the system's enforced workflow that causes users to enter data in non-intuitive ways that are vastly different than how one interacts with patients (Rathert, Porter, Mittler, & Fleid-Palmer, 2019). Although these studies do not agree on every aspect, one theme is at the heart of them all: poor EMR design.

3. DATA COLLECTION AND ANALYSIS

Semi-structured interviews were conducted on eight health care workers at two geographic locations (metro and rural) working at various sizes of hospitals (Level 1 Trauma, small regional, etc.). Audio recordings of the interviews were reviewed for themes for each interview questions by at least two separate individuals on the research team. These themes were then collated and assigned to each research question. The demographic information of the participants is provided in Appendix 1. The themes are discussed below in the results section.

4. RESULTS

RQ1: How does use of EMR impact the experience of the healthcare service provider in delivering patient care?

It is clear from the data that the impact of EMR is perceived as more negative than positive by most of our health service provider participants. Participants identified several issues with EMR that lead to frustration within the workforce providing direct care. Our data suggests five emergent themes about issues with EMR that negatively impact patient care (Table 1): Data schema does not match reality, navigation of the systems cumbersome, lack of appropriate fidelity, systems do not mirror actual workflow, and poor communications with other systems in and outside of the hospital or setting.

Respondents noted that the system allowed a lot of data redundancy, which means entering the same information multiple times in many steps. It is easy to get lost or be unable to retrieve pertinent information from the system as there are too many options available from which to choose and it does not make intuitive sense.

The organization of data flow in the systems and the schema did not quite match how information was generated and consumed by the system and providers. The second theme was the most dominant theme for this research question as practically every participant mentioned this problem. The system has multiple entry points to complete charting specific procedures and sometimes it is impossible to navigate to all the right spots. One participant observed:

> "When we do pulmonary function studies, you could have the printout from the PFT machine… That physical paper that gets printed out gets read by the doctor; he interprets it via the telephone… they dictate it [...]. That dictation goes one place in that chart, but the scanned paper that's printed from the machine with the graphs and the actual test goes in a separate place in that chart. So many people who aren't familiar with the charting system don't know where to look, and for some reason, they can't merge the two in the chart somewhere."

Final Themes		count
1-1	Data schema doesn't match reality	5
1-2	Navigation of the systems cumbersome	8
1-3	Lack of appropriate fidelity	5
1-4	Systems do not mirror actual workflow	3
1-5	Poor communication with other systems	3

Table 1. Themes about issues with EMR

The third theme, lack of appropriate fidelity, is about the system demanding too much time to get any work done. The system sometimes requires too many details, which seems unnecessary, or not enough details, which is worrisome, to complete a patient's file. As a participant shared,

> "Sometimes I'll listen to a patient [...], and they'll be so wheezy, and you know no one's been listening to them because everyone's just been charting the same damn thing because you can just click [Use Previous Values]. It happens all the time."

The fourth theme for this research question, systems do not mirror actual workflow, suggests counterintuitive systems in terms of how the process flows on the actual floor of a hospital versus how it moves on the system. The actual workflow requires certain things to be done in order to move on to other things, whereas the systems do not mirror the real sequence of events, which leads to entering similar information multiple times.

The final theme for this research question, poor communication with other systems, entails challenges due to systems within the same facility unable to communicate appropriately. The systems that need to be linked in a way that data can flow seamlessly for cost-effective patient care are more like silos that have limited asynchronous connectivity causing more harm than good for patient outcomes. As shared by a respondent:

> "The ABG machine was an hour off because of daylight savings time. So when the results crossed over, it looked like I drew the blood gases an hour later or earlier than what I actually did, which then wouldn't correlate if someone went to change that person's oxygen."

All of these themes identified from our data suggest how and why systems can cause frustration for providers, the time lag in right care, and inefficiency in workflow and compromised data quality in patient charts. These are some of the reasons that encourage "workarounds" in such settings.

RQ2: What are the underlying factors that lead to health care providers taking off protocol (not prescribed by EMR) measures to get their job done?

As discussed in the previous section, working with EMR leads to multiple issues which can demoralize service providers and encourage them to take "shortcuts" to get the job done. For research question 2, four categories of workarounds were identified (Table 2). The first category suggests that EMR issues acknowledged by providers negatively impact the service provider's attitude to patient care. The service provider's priority becomes meeting the system's requirements (which are complex) over interacting with patients directly. This shift in opinion on the part of the provider impacts patient-caregiver connection.

The second category of responses was around poor interface design of the systems leading to "workarounds," such as not scanning medications and following the right protocols. The systems are counterintuitive and repetitive and many times slow in providing timely help to the patient. Scanning medications is an area where workarounds due to bad design is standard. As shared by a participant:

> "A medication I administer a lot is called Duoneb. So say I'm out of Duoneb but I have an Albuterol and I have an Atrovent, you know, put those two together, and it makes Duoneb. You're able to press this little ID code button and say 'unable to scan' and administer whatever medication that you want to. In the past, we've had to let go nurses because they said that they were giving certain medications that they weren't."

Final Themes		Count
2-1	System negatively impacts service provider's attitude to care	3
2-2	Poor interface design	6
2-3	System demands/charting slows down patient care	6
2-4	System for liability purposes more than patient care	4

Table 2. Categories of Workarounds

The third category of factors leading to workarounds is about system demands on the service providers, which slows down actual patient care. The functional requirements of systems in terms of detailed charting and multiple entries of similar things leads to situations where actual values of vitals are noted down and charting is performed after the shift. The delay in recording the data and charting leads to less accurate data entry, less time to spend with patients, and being less confident about being able to meet all demands to get the job done.
As shared by a respondent:

> I write down the vitals so that I can get it in when I have time. Hours after writing it down, when I get to charting in the systems, I should be able to read my handwriting and guess the right time of the day when I noted these numbers! How else am I going to get through the shift?

The last category in this section is about perceiving systems and charting more as a liability reduction tool than improving patient care vector. Service providers are mandated to follow system requirements and providers strictly are discouraged from using their judgment to document patient data. As shared by a service provider:

> Has the patient changed position? Okay, my patient is sedated and knocked out straight for three days. They are not in moving position. But you can't keep pulling your information from the previous assessment! But nothing has changed! It is a very big circle of keep doing exactly the same thing. It's redundant. More for liability than care.

The motivations for "workarounds" in healthcare settings is mostly providers trying to do the right thing at the right time, even if it requires cutting corners to get the job done. The EMR system is perceived as a useful tool in service delivery, but it exists more to protect an organization's liability exposure and less to improve care for the patient.

RQ3: How can the issues with EMRs be addressed to make a health care service provider's experience better to improve patient outcome?

Our data suggested three things to improve healthcare service providers' experience with EMRs (Table 3). The first suggestion is to develop a universal platform for EMRs such that no matter which organization they work with, they are familiar with the interfaces and know precisely where to find any piece of information. They spend years learning one system, and with a quick change of system or a job they have to relearn everything completely. A respondent shared her experience about shifting from one EMR to another:

> "It was kind of a nightmare for a couple of weeks. They cleared out the hospital to make sure that it would go easier. They sent patients home or to other hospitals."

The second suggestion is around improving charting experience. Data suggests that it is frustrating to efficiently chart and provide excellent patient care at the same time. Service providers typically do charting later after they have engaged with the patient or after the shift is over. Most EMRs have dense menus, redundant information requirements, and lack of clarity in effectively navigating the system to make the right decision at the right time for service providers. As shared by a respondent:

> "I've noticed the older nurses, like; they'll stay after work for two hours to chart. Sometimes they get paid, sometimes they don't. And you never, ever, ever see a young nurse do that. [...] Two hours… I was so pissed when I found out they get paid. That's overtime!"

Improvements in system	Count
Develop a universal platform for EMR	6
Make charting more convenient	10
Develop system interoperability	5

Table 3. Suggested Improvements in EMR

The final theme for this research question is to develop system interpretability to be able to acquire real benefits of EMR systems. It is helpful when service providers can pull patient records at different functional units for the same organization. As shared by a respondent:

> "When my dad had his episode with his seizures, we went to three different facilities, and you know what, they were actually all [within the same health system]. I know at least two of them had Epic, I don't know if [the third] does so I can't answer that. Every single facility we went to asked my dad what happened. And I get it; they want to hear the story from him. But some of them didn't even know he had a history of seizures where he had been seizing for 10-15 years, so that was frustrating."

Our participants agreed that EMR is a good thing. EMR makes service providers accountable for their actions and provides continuity of care to patients. However, it is crucial that the experience of using EMR is more positive than negative for service providers. Negative experiences with poorly designed system interfaces lead to workarounds in healthcare settings.

5. DISCUSSIONS

In the sociotechnical framework of information systems, our results indicate that all problems in using EMR are primarily in the technical sphere of the context and are systemic. EMR is a tool created for health service providers to make better patient care available. The EMR system, however, is disappointing in the apparent intent (more liability oriented), in its design, in its inability to follow the actual workflow, and in lack of support for seamless flow of data creating frustration for users of this systems. These kinds of frustrations lead to "workarounds," which are, for the most part, service providers' trying to do the right thing at the right time for the patient. As one of the participants observed:

> "A doctor will put in for a triple dose; pharmacy sees it and says, 'I don't recommend that.' So I call down to the pharmacy and say, 'This patient cannot breathe. They're chugging away at a respiratory rate of 40/min. I need this dose approved so I can document it.' In reality, I've already started that treatment because I know it's an effective dose, and I have an order from the doctor."

It is essential that these problems are addressed and we have systems that have a universal platform and interoperability to support continuity of care. Charting is critical to provide an excellent patient outcome and the current cumbersome and confusing charting is detrimental to patient care. As shared by a participant:

> "For every single breathing treatment that I give, I have to put a code [...], and it charges the patient for one breathing treatment. So say I'm seeing 30 patients… I have to charge for oxygen; I have to charge for every time I put a pulse-ox on their finger… nothing is automatically charged. It all relies on humans. Think about it; humans make mistakes. [...] The number of times that they order nebs in the ER, maybe once every couple months… they give them daily, and they should be charging for every single one. Think about all of the money that's missed."

Additional insight into the usage of EMR in this study is about extra challenges in using such systems by a relatively older population of healthcare service providers. This segment of users have less confidence in the use of technology and take more time than younger employees, which leads to increased anxiety and more chances of errors in the systems. All participants agreed that older workers find the system more challenging, feel frustrated in asking the same questions many times, and stay back longer to complete their regular shift work on the system. As a participant shared:

> "I have known people who have quit over the implementation of the computer system that we have. They would rather leave their job than have to have learned that new system."

Our results provide several recommendations for practitioners in this field:

- Enable health care service providers on the floor of hospitals by providing a powerful tool such as EMR that aids them in doing their work and not hinder the day-to-day work. (Themes 1-2, 1-4, 2-1, 2-2, 2-3, 2-4)
- Allow a realistic amount of time for system usage purposes per shift to health care workers so that the focus on the quality of care is not diluted in the

process of using the system. (Theme 1-3, 2-1)
- Create a universal platform that helps retrieval or data input effortlessly and service providers could use it from anywhere. (Theme 1-5, 2-1, 2-3)
- Periodically revisit protocols and procedures around the usage of EMR. Assess if the system mirrors the workflow and assists the provider, not take them away from the patient's bedside. (Themes 1-1, 1-4, 2-1, 2-2, 2-3, 2-4)

If the system makes service providers stay for extra hours rather than leave after their shifts are over to complete charting, it will eventually not result in good patient outcomes. There should be real-time provision for charting in any service provider's shift. Overall, the technical issues with EMR could be fixed, the vision of having "one record per person" in universal EMR is possible, and it could eventually bring the total cost of healthcare down in the country.

6. CONCLUSION

This study has implications for theory and practice in the healthcare field. The results of this study identify issues with EMR leading to workarounds in healthcare settings daily. The results in this study support and significantly contribute to the body of knowledge in EMR and workaround research, adding much more workaround detail than previous studies, and it can inform many more studies to refine the understanding of these issues further. Practitioners, such as healthcare administrators, could use these results in their settings to further refine the protocol of EMR usage. EMR vendors could use the results to improve their products that are being used in the industry.

7. REFERENCES

Acharya, A., Schroeder, D., Schwei, K., & Chyou, P.H. (2017). Update on electronic dental record and clinical computing adoption among dental practices in the United States. *Clinical Medicine & Research, 15*(3-4), 59-74.

Adler-Milstein, J. & Jha, A.K. (2017). HITECH act drive large gains in hospital electronic health record adoption. *Health Affairs, 36*(8), 1416-1422.

Aldosari, B., Al-Mansour, S., Aldosari, H., & Alanazi, A. (2018). Assessment of factors influencing nurses acceptance of electronic medical record in a Saudi Arabia hospital. *Informatics in Medicine Unlocked, 10,* 82-88.

Aluas, M. (2016). Ethical and legal considerations of healthcare informatics. *Applied Medical Informatics, 38*(3-4), 91-98.

Assis-Hassid, S., Grosz, B.J., Zimlichman, E., Rozenblum, R., & Bates, D.W. (2019). Assessing EHR use during hospital morning rounds: A multi-faceted study. *PloS ONE, 14*(2), 1-15.

Bae, J. & Encinosa, W.E. (2016). National estimates of the impact of electronic health records on the workload of primary care physicians. *BMC Health Services Research, 16,* 172-182.

Bae, J., Rask, K.J., & Becker, E.R. (2018). The impact of electronic medical records on hospital-acquired adverse safety events: Differential effects between single-source and multiple-source systems. *American Journal of Medical Quality, 33*(1), 72-80.

Barrett, A.K. (2018). Technological appropriations as workarounds: Integrating electronic health records and adaptive structuration theory research. *Information Technology & People, 31*(2), 368-387.

Barrett, A.K. & Stephens, K.K. (2017). Making electronic health records (EHRs) work: Informal talk and workarounds in healthcare organizations. *Health Communication, 32*(8), 1004-1013.

Choi, E.Y., Chung, E.J., & Lee, H.S. (2006). Users' satisfaction with the electronic nursing record system. *Studies in Health Technology & Informatics, 122*, 855.

DesRoches, C.M., Charles, D., Furukawa, M.F., Joshi, M.S., Kralovec, P., Mostashari, F., Jha, A.K. (2013). Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. *Health Affairs, 32*(8), 1478-1485.

Draus, P., Mishra, S., Slonka, K., Bromall, N., Slonka, K. (2019). Healthcare professionals' perception of the usability of electronic medical records. Submitted for publication in Issues in Information Systems, 2019.

Dudding, KM., Gephart, S.M., & Carrington, J.M. (201). Neonatal nurses experience unintended consequences and risks to patient safety with electronic health records. *Computers, Informatics, Nursing, 36*(4), 167-176.

Eason, K. & Waterson, P. (2014). Fitness for purpose when there are many different purposes: Who are electronic patient records for? *Health Informatics Journal, 20*(3), 189-198.

Edwards, C. (2012). Nursing leaders service as a foundation for the electronic medical record. *Journal of Trauma Nursing, 19*(2), 111-114.

Friedman, D.J., Parrish, R.G., & Ross, D.A. (2013). Electronic health records and US public health: Current realities and future promise. *American Journal of Public Health, 103*(9), 1560-1567.

Fuji, K.T., Galt, K.A., Siracuse, M.V., & Christoffersen, J.S. (2011). Electronic health record adoption and use by Nebraska pharmacists. *Perspectives in Health Information Management, 8*(3), 1-9.

Gagnon, M.P., Ghandour, E.K., Talla, P.K., Simonyan, D., Godin, G., Labrecque, M., Rousseau, M. (2014). Electronic health record acceptance by physicians: Testing an integrated theoretical model. *Journal of Biomedical Informatics, 48,* 17-27.

Garrett, P. & Seidman, J. (2011). EMR vs. EHR – What is the difference? Retrieved from https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference

Gesulga, J.M., Berjame, A., Moquiala, K.S., & Galido, A. (2017). Barriers to electronic health record system implementation and information systems resources: A structured review. *Procedia Computer Science, 124,* 544-551.

Government Publishing Office. (2018). Code of federal regulations: Title 45 – Public Welfare: Part 162 – Administrative Requirements. Washington, DC.

Halbesleben, J.R.B., Wakefield, D.S., & Wakefield, B.J. (2008). Work-arounds in health care settings: Literature review and research agenda. *Health Care Management Review, 33*(1), 2-12.

Institute of Medicine of the National Academies. (2003). *Key capabilities of an electronic health record system: Letter report* (ISBN 0-309-55877-8). Washington, DC: The National Academies Press.

Meeks, D.W., Smith, M.W., Taylor, L., Sittig, D.F., Scott, J.M., & Singh, H. (2014). An analysis of electronic health record-related patient safety concerns. *Journal of the American Medical Informatics Association, 21*(6), 1053-1059.

Morath, J.M. & Turnbull, J.E. (2005). *To do no harm*. San Francisco, CA: Jossey-Bass.

Pagulayan, J., Eltair, S., Faber, K. (2018). Nurse documentation and the electronic health record: Use the nursing process to take advantage of EHRs' capabilities and optimize patient care. *American Nurse Today, 13*(9), 48-54.

Patterson, E.S. (2018). Workarounds to intended use of health information technology: A narrative review of the human factors engineering literature. *Human Factors, 60*(3), 281-292.

Rathert, C., Porter, T.H., Mittler, J.N., & Fleid-Palmer, M. (2019). Seven years after meaningful use: Physicians' and nurses' experiences with electronic health records. *Health Care Management Review, 44*(1), 30-40.

Sackett, K., Jones, J., & Erdley, W.S. (2005). Incorporating healthcare informatics into the strategic planning process in nursing education. *Nursing Leadership Forum, 9*(3), 98-104.

Sackett, K.M., Erdlet, W.S., & Jones, J. (2006). The Western New York regional electronic health record initiative: Healthcare informatics use from the registered nurse perspective. *Studies in Health Technology and Informatics, 122,* 248-252.

Schiff, G.D. & Zucker, L. (2016). Medical scribes: Salvation for primary care or workaround for poor EMR usability? *Journal of General Internal Medicine, 31*(9), 979-981.

Sidek, Y.H. & Martins, J.T. (2017). Perceived critical success factors of electronic health record system implementation in a dental clinic context: An organisational management perspective. *International Journal of Medical Informatics, 107,* 88-100.

Stevenson, J.E., Israelsson, J., Nilsson, G., Petersson, G., & Bath, P.A. (2016). Vital sign documentation in electronic records: The development of workarounds. *Health Informatics Journal, 24*(2), 206-215.

Stutzer, K. & Rushton, C.H. (2015). Ethics in critical care: Ethical implications of workarounds in critical care. *AACN Advanced Critical Care, 26*(4), 372-375.

Syoubuzawa, S., Yamanouchi, K., & Takeda, T. (2006). Nursing information processing abilities: A comparison of nursing managers and staff nurses. *Studies in Health Technology and Informatics, 122,* 819.

Wang, T. & Biedermann, S. (2012). Adoption and utilization of electronic health record systems by long-term care facilities in Texas. *Perspectives in Health Information Management,* 1-14.

Xierali, I. M., Hsiao, C.J., Puffer, J.C., Green, L.A., Rinaldo, J.C.B., Bazemore, A.W., Phillips Jr., R.L. (2013). The rise of electronic health record adoption among family physicians. *Annals of Family Medicine, 11*(1), 14-19.

Yung, A. (2017). Adoption of electronic health record system in community-based physiotherapy clinics: A pilot case study. *Studies in Health Technology & Informatics, 234,* 395-400.

APPENDIX 1

Participants' Demographic Information

Participants	**Job description, age, years of experience**	**Primary shift, employment status, gender**
Participant 1	Assistant at detox unit, 31 years 4 years	Evenings 3-11, Full time, Female
Participant 2	Nursing assistant, 21 years, 3 years	day shifts, part time, female
Participant 3	patient care technician, 21 years, 3 years	both day and night shifts, casual, Male
Participant 4	patient care technician, 20-30 range, 2 years	Day shift, casual, Female
Participant 5	Respiratory Therapist, 23 years old, 3 years	Night shift, Full time, Female
Participant 6	Respiratory Therapist, 23 years old, 3 years	Day shift, Full time, Female
Participant 7	Respiratory Therapist, 32 years old 12 years	Day shift, Full time, Female
Participant 8	Respiratory Therapist, 26 years old 4 years	Day shift, Full time, Male

Literary Analysis Tool: Text Analytics for Creative Writers

Austin Grimsman

Douglas M Kline
klined@uncw.edu
Information Systems

Ron Vetter
vetterr@uncw.edu
Computer Science

Curry Guinn
guinnc@uncw.edu
Computer Science

University of North Carolina Wilmington
Wilmington, NC 28403

Abstract

Creative writers struggle with obtaining reliable and consistent readers for their draft works. Human reviewers are notoriously inconsistent across different reviewers, and a single reviewer's feedback can vary significantly over time. Additionally, there are logistical issues with human feedback. We apply text analytics techniques to literary works with the goal of aiding writers in revisions. For a set of text, the Literary Analysis Tool (LAT) provides objective statistics, windowed statistics over the length of the text, and mood analysis. The LAT provides quick feedback, at any time, in an absolutely objective manner. We present the feedback of a small set of creative writers. The results indicate that text analytics has a place in the creative writing process.

Keywords: Creative writing, literary analysis, text analytics.

1. INTRODUCTION

Creative writers typically follow this general cyclical process: draft, get review, revise, get review, revise… Reviews are necessary to make sure that what the author intends to convey is what they actually conveyed.

Historically, reviews are performed by a human, preferably not the author, who lacks the distance necessary to be objective. However, human reviewers have disadvantages. They are subject to personal biases & moods resulting in inconsistent and conflicting feedback. Furthermore, performing a good review is time consuming, mentally demanding, and requires patience, enthusiasm, and experience. As a result, good reviewers are hard to find, and it may take a long time to get a review.

We present the Literary Analysis Tool (LAT) which attempts to give useful feedback and addresses some of the disadvantages of human reviewers. The LAT uses standard text analytics techniques applied to the unique domain of creative writing. The LAT has the advantages of being on-demand and perfectly objective, but cannot realistically

compare to the experience and nuanced analysis of a human reviewer. Furthermore, a computational analysis might ease the burden on human reviewers by analyzing tedious items such as punctuation, length of sentences, and indirect language.

Text Analytics

The LAT uses traditional statistical based text analytics described by (Salton, Automatic Text Processing, 1989) and (Salton & McGill, Introduction to modern information retrieval, 1983). This type of analysis is essentially based on word frequencies, with the supposition that higher frequencies indicate importance.

This type of analytics can be improved by counting words with a known purpose or meaning. For example, we might count the occurrences of words in this set {death, dead, dying, mortality}. A high count might indicate a focus on death and general dark tone to a passage.

The Linguistic Inquiry and Word Count (LIWC) application takes this approach (Pennebaker, Booth, & Francis, 2007). The LWIC application has a large number of word lists, categorized by linguistic purpose (pronouns, verbs, prepositions etc.) as well as concepts such as "Family", "Anger", "Time", "Sexuality", "Death", etc.

LIWC also counts:

- First person singular and plural pronouns
- Third person singular and plural pronouns
- Parts of speech: articles, nouns, verbs, prepositions, conjunctions, negations, etc.
- Past, present, and future tenses.

Less astonishing analytics can also be performed that are very helpful for a creative writer. Metrics such as word counts and spelling check are now standard in word processors. Additional metrics that are helpful to a creative writer include: length of paragraph, frequency of words, counts of punctuation usage, and words repeated within a window, e.g., the word "really" was used twice within 5 words.

In this exploratory study, the LAT does not implement more sophisticated methods such as Natural Language Processing or deep learning models. If the easily-implementable statistical methods prove useful to writers, more sophisticated measures can be pursued.

2. PROBLEM STATEMENT

Writing is a nebulous creative process. Good finished products can be accomplished individually, but external feedback is acknowledged as beneficial. A writer's perception of their work is influenced by their journey in writing. External review is the only real way for a writer to know what they have written, rather than what they think they wrote.

Writers have traditionally relied on human reviewers. However, humans bring their own personal biases and moods to the reading. Different reviewers can provide widely variable and often opposite feedback. Relying on a single reviewer might result in positive / negative, so multiple reviews are advisable.

Unfortunately, reviewing is difficult and time consuming. It takes substantial mental effort for a focused reading with the added cognitive load of assessing on many dimensions what is being read. Good reviewers should be experienced, honest, and willing to spend the time to perform the review.

For the writer, obtaining external reviews is an external dependency over which they have little control. Overly harsh/soft feedback from an overly depressed/happy reviewer can significantly change the direction of revisions. A long delay in a review can also impact the writing process.

In short, humans are not ideal reviewers. Attributes of a truly ideal reviewer include:

- Instant feedback, on demand
- Consistency – same feedback for same work
- Feedback on technicalities: punctuation, word counts, etc.
- Feedback on style: passive voice, flowery language
- Feedback on mood
- Feedback over the entire passage, as well as the throughout the reading

The LAT system is an attempt to address these issues using straightforward text analysis techniques.

3. LAT SYSTEM

The goal of the LAT system is to assess the usefulness of standard text analytics for creative writers. Considering this goal, the choice was made to forego application integration with word processors. Application integration (with, for example MS Word) would have been dependent on particular word processors and versions, and would have required deployment across users machines.

HTML, CSS, and Javascript proved to be a good prototyping platform, and the entire system was

deployed as a single page web site. Deployment was a matter of sending a URL to a user. Appendix A gives a screenshot of the LAT system. Users paste a passage on the left-hand textbox, then choose an analysis button. Results are shown in the right-hand textbox. Graphical representations are shown below the buttons and can be seen by scrolling down to them.

To explore the utility of text analytics for creative writers, the LAT implemented very standard, easily implemented metrics:

- LIWC Analysis
- Language Tone Analysis
- Paragraph check
- Word Frequency
- Punctuation
- Indirect Language
- Word Proximity
- Moving Window Language Tone

Table 1: LIWC parts-of-speech analysis for Poe's "The Cask of Amontilado"

Word Count: 2348
Function Words: 1332
Total pronouns: 359
Personal pronouns: 261
1st pers singular pronouns: 158
1st pers plural pronouns: 24
2nd pers pronouns: 33
3rd pers singular pronouns: 86
3rd pers plural pronouns: 13
Impersonal pronouns: 98
Articles: 240
Common verbs: 271
Auxiliary verbs: 171
Past tense: 142
Present tense: 81
Future tense: 25
Adverbs: 71
Prepositions: 317
Conjunctions: 121
Negations: 46
Quantifiers: 41
Numbers: 35
Swear words: 0

LIWC Analysis

LIWC analysis provides counts of words in many categories. The categories include parts-of-speech, as well as general categories of meaning. Parts-of-speech output is a count of the number of instances of, for example: number of words, total number of pronouns, 1st person singular pronouns, 3rd person plural pronouns, impersonal pronouns, verbs, past tense verbs, adverbs, prepositions, conjunctions, quantifiers, negations, numbers, and swear words.

Where LIWC really shines is in relating words to abstract concepts and emotions. For example LIWC provides a count of words in the passage that relate to family, such as {kin, mom, dad, husband, son, family, uncle, relative, etc.}, and variations on those. As another example, LIWC provides a count of words related to anger, such as {agitate, angry, bitter, cruel, destroy, fury, jerk, mad, jealous, revenge, etc.} Details on how these word lists were developed can be found here (Tausczik & Pennebaker, 2010).

Table 2: LIWC Concept Analysis for Poe's "The Cask of Amontilado"

Social processes: 235
Family: 2
Friends: 8
Humans: 8
Affective processes: 142
Positive emotion: 77
Negative emotion: 65
Anxiety: 13
Anger: 13
Sadness: 14
Cognitive processes: 265
Insight: 32
Causation: 20
Discrepancy: 17
Tentative: 21
Certainty: 21
Inhibition: 9
Inclusive: 108
Exclusive: 45
Perceptual processes: 63
Seeing: 12
Hearing: 34
Feeling: 14
Biological processes: 62
Body: 37
Health: 12
Sexual: 3
Ingestion: 12
Relativity: 302
Motion: 50
Space: 151
Time: 103
Work: 14
Achievement: 24
Leisure: 13
Home: 5
Money: 7
Religion: 2
Death: 9
Assenting: 10
Nonfluent words (er, umm): 3
Filler words: 2

Language Tone Analysis
One drawback of the LIWC analysis is the amount of detail – LIWC breaks down to 64 different constructs. To address this, we provide a more aggregated, and faster executing, analysis.

Table 3: Language Tone output for Edgar Allen Poe's "The Cask of Amontilado"

Word Count: 2348
Positive Emotions: 77
Negative Emotions: 65
Total Emotional Intensity: 142
Cognitive Mechanisms: 265
Motive Concerns: 63
Perceptual/Personal Processes: 116

Paragraph, Frequency, Punctuation, and Indirect Language, Word Proximity
Tables 4 through 7 give the output for the LAT's analysis of Paragraphs, Word Frequency, Punctuation and Indirect Language, respectively.

Table 4: Paragraph Analysis

Paragraphs: 89
Words: 2348
Avg Paragraph length: 26.4

Table 5: Word Frequency

said: 24
amontillado: 16
upon: 15
ugh: 15
fortunato: 14
will: 13
us: 10
one: 8
replied: 8
let: 8
friend: 7
yes: 7
luchresi: 6
go: 6
back: 6
long: 6
catacombs: 6
bones: 6
must: 5
<shortened for space>

Depending on the audience and the writer's intention, longer or shorter paragraphs may be desirable. Feedback on word frequency may be a sign to the writer of the actual emphasis on certain topics. Table 6 shows punctuation, and it is interesting to note that Poe used 28 semicolons in his short story The Cask of Amontillado. Table 7 shows counts of indirect language. The number of adverbs is one imperfect measure. Examples of indirect language are {generally, commonly, presumably, could, might, etc.}.

Table 6: Punctuation

Periods: 177
Commas: 158
Colons: 0
Semicolons: 28
Apostrophes: 3
Quotation Marks: 166
Exclamation Marks: 49
Brackets: 0
Parentheses: 0
Braces: 0
Hyphens: 36
Ellipses: 0
Em Dashes: 29

Table 7: Indirect Language

9 indirect language phrases.
71 adverbs.

Table 8 shows the output for a five-word window. As an example, the word "ugh" 73 times in a moving window of five words throughout the text. Changing the window to 2, shows that "ugh" appears 28 times directly next to itself, specifically, in the passage that includes 15 instances in succession.

Table 8: Word Proximity, 5 word window

the: 162
he: 79
ugh: 73
i: 50
of: 22
yes: 18
ha: 13
a: 10
true: 8
in: 8
and: 8
punish: 6
you: 6
as: 6
it: 6
mason: 6
are: 4
to: 4
not: 4
tell: 2
will: 2
sign: 2
with: 2

Windowed Tone Analysis
The LAT provides a way to view the dynamics of a passage from beginning to end. This is

implemented as a windowed analysis where the window is adjustable. Appendix B gives an example by displaying a count of positive and negative emotion words with a window size of 500. The x-axis represents the word location in the passage, the blue line represents frequency (count) of negative-emotion words, and the tan line represents frequency of positive-emotion words. The window is adjusted at the beginning and end, for example, the window at word 1 is 250 words, representing the first 250 words, the window at word 2 is the first 251 words, with the first full 500-word window occurring at word 250.

For those familiar with the short story The Cask of Amontillado, the negative emotions surpass the positive at the point when Fortunado endures a fit of coughing, and the narrative becomes immediately darker with words including {health, ill, kill, die, caution, buried, repose, serpent, fangs, skeletons, etc.}

Appendix C displays a similar graph, this time displaying the frequency of perceptual words exemplified by colors, smells, sounds, touch, and body parts (especially sensory organs). The frequency of perceptual words again makes sense for this short story, building up through exposition and as the characters move into the crypt, then descending through dialog and action.

4. RESULTS

To assess the utility of the LAT, we performed a small study by enlisting a convenience sample of creative writers. The enlisted writers self-assessed their experience level, and ranged from inexperienced to published experience writers. Nine candidates were invited, with a clear expectation that the assessment would require at least an hour of their time, and also require learning the relatively simple LAT software. Six ultimately completed the assessment. It took approximately six weeks for the six writers to complete the assessment.

Efforts were made to not bias the writers, but interactions were inevitable, mainly in helping use the LAT. Writers were instructed to try the LAT's various analyses on existing works that they knew well, and also their own work. Free-form experimentation with the LAT was to last at least 30 minutes, after which they could complete a short (12 question survey).

Respondents rate the various analysis features on a 5-point likert scale where 1 was "Not Useful" and 5 was "Extremely Useful". Raters varied in their overall opinions, with the average across-the-board ratings mean ranging from 2.11 for one rater to 4.11 for another. The results are shown in Table 9. Due to the limited sample size, all ratings are shown, in ascending order, along with the mean rating.

Table 9: Writer Assessments of the LAT

feature	Ratings	Mean
Language Tone	2, 3, 3, 3, 4, 5	3.33
LIWC Analysis	1, 2, 3, 4, 4, 5	3.17
Paragraph Check	2, 2, 2, 3, 3, 5	2.83
Word Frequency	2, 2, 3, 3, 5, 5	3.33
Punctuation	1, 2, 2, 3, 4, 4	2.67
Indirect Language	2, 2, 3, 3, 4, 4	3.00
Word Proximity	2, 2, 3, 4, 4, 5	3.33
Overall Usefulness	2, 2, 3, 4, 4, 5	3.33
Would you seriously consider using the Literary Analysis Tool to assist in your creative writing process?	4 – Yes 2 – No	
Would you seriously consider using the Literary Analysis Tool to assist your writing process if it were further refined in the future?	6 – Yes 0 - No	

5. CONCLUSION

The last question from the survey was encouraging, in that writers found at least some potential for a useful system. The most highly rated features were the Language Tone Analysis, the Word Frequency, and the Word Proximity report. The lower rated features included the Punctuation report and the Paragraph report.

The results must be moderated by consideration of the low sample, and the intangible, subjective nature of rating "usefulness to the writing process". However, the prototyping process has appeared to clearly work, indicating that a more polished application merits consideration.

We also received informal feedback regarding absolute versus relative metrics. For example, knowing that the average paragraph length is 65 words, an absolute metric, is difficult to interpret and take action on. In contrast, seeing a graph of the number of positive-emotion words is more

helpful, in that it compares a point in the passage with previous and following narrative. Perhaps, a metric such as paragraph length would be better presented as a relative metric, pointing out paragraphs that are unusually small or large.

In the future, a more polished system might be integrated into the users preferred word processor, with reports available through a context-sensitive menu for a particular selection of text. This would certainly avoid the cut-and-paste necessary steps necessary for the current LAT. With this improvement, writers might more fully integrate feedback solicitation into their writing process. It could be that the extra steps necessary did not allow writers to take advantage of the on-demand, immediate benefits of an automated analysis.

6. REFERENCES

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. Palo Alto, CA: Xerox Palo Alto Research Center.

Ananiadou, S. (2009). Adding Value to Scholarly Communications & Repositories through Text Mining. Manchester, UK: The University of Manchester. Retrieved 1 1, 2013, from https://indico.cern.ch/event/48321/contributions/1992204/attachments/957243/1358661/OAI6_SA.pdf

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1995). Representing documents using an explicit model of their similarities. Journal of the American Society for Information Science, 46(4), 254-271.

Charniak, E. (1996). Statistial Language Learning. MIT Press.

Clifton, C., Cooley, R., & Rennie, J. (2004). TopCat: data mining for topic identification in a text corpus. IEEE Transactions on Knowledge and Data Engineering, 16(8), 949-964.

Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. nature: international journal of science, 430(6998), 453-456. doi:10.1038/nature02634

Mischne, G. (2007). Applied text analytics for blogs. Universiteit van Amsterdam. Retrieved from https://pure.uva.nl/ws/files/4378014/47196_mishne_thesis.pdf

Pang, B., & Lee, L. (2008). Opinon Mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1-135.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Operator's Manual Linguistic Inquiry and Word Count: LIWC2007. Austin, TX: The University of Texas at Austin and The University of Auckland, New Zealand.

Ramsay, S. (2003). Reconceiving Text Analysis: Toward an Algorithmic Criticism. Literary and Linguistic Computing, 18(2), 167-174. Retrieved from https://doi.org/10.1093/llc/18.2.167

Salton, G. (1989). Automatic Text Processing. Reading: Addison-Wesley Publishing Company.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. McGraw-Hill.

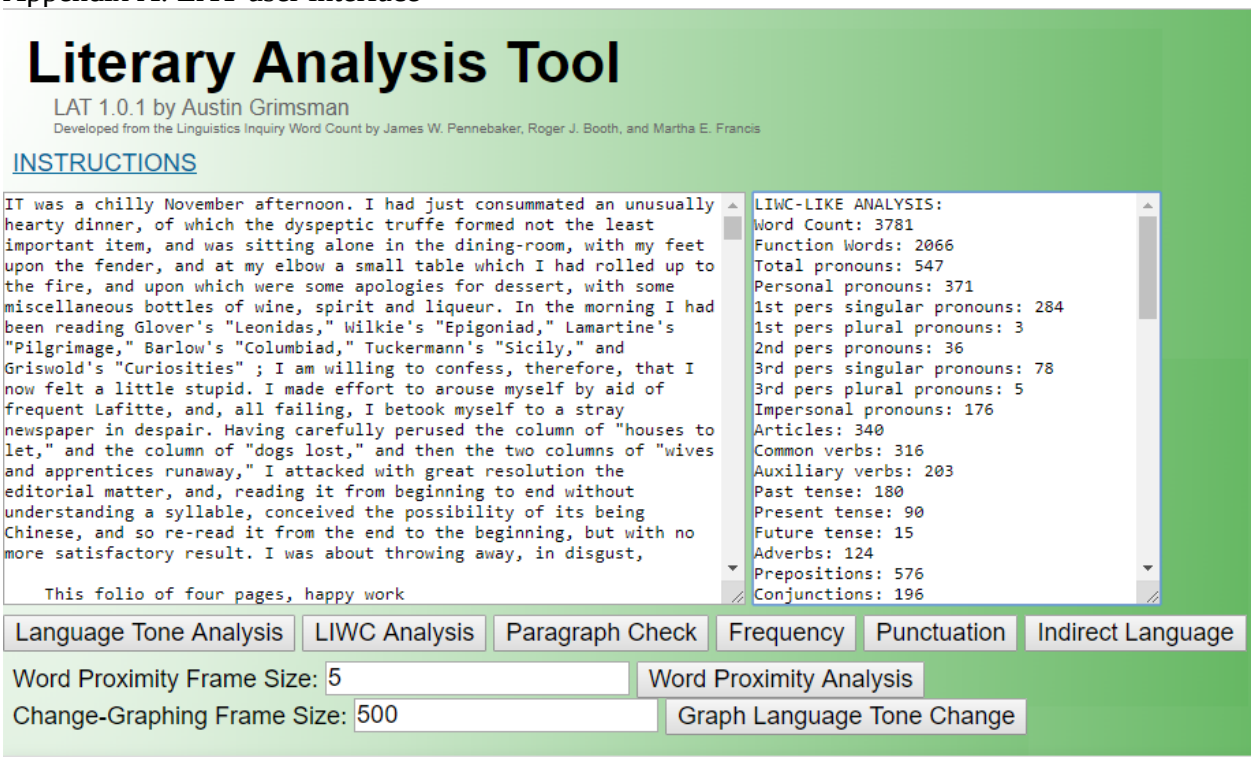
Shutova, E. (2010). Models of metaphor in NLP. ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (pp. 688-697). Uppsala, Sweden. Retrieved from https://dl.acm.org/citation.cfm?id=1858752

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

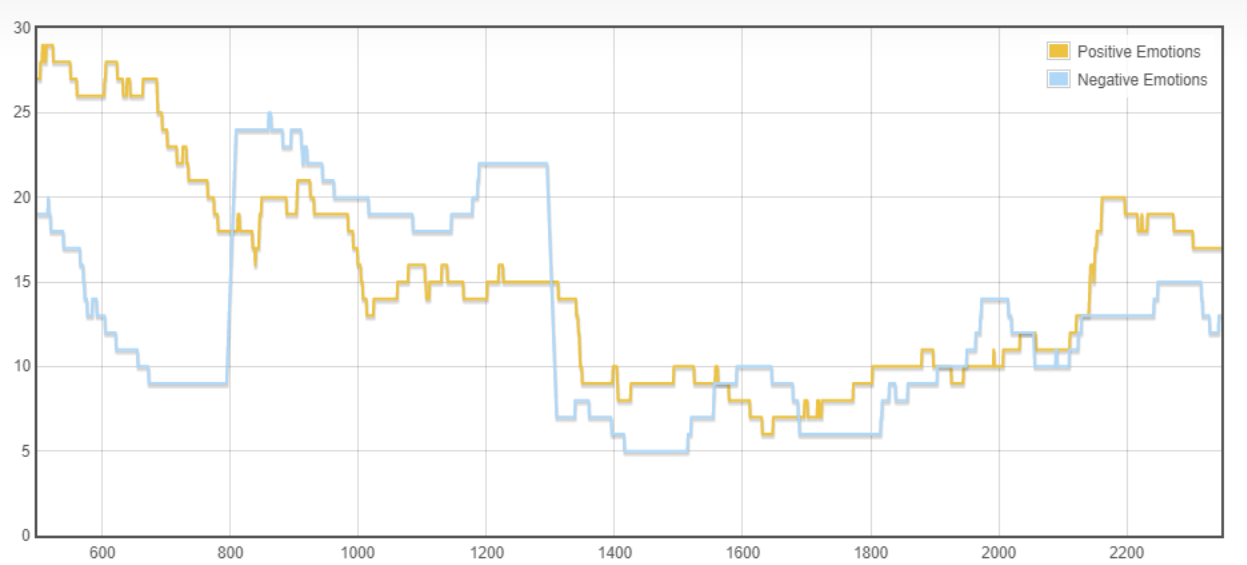
Williams-Whitney, D., Mio, J. S., & Whitney, P. (1992). Metaphor production in creative writing. Journal of Psycholinguistic Research, 21(6), 497-509.

Appendices and Annexures

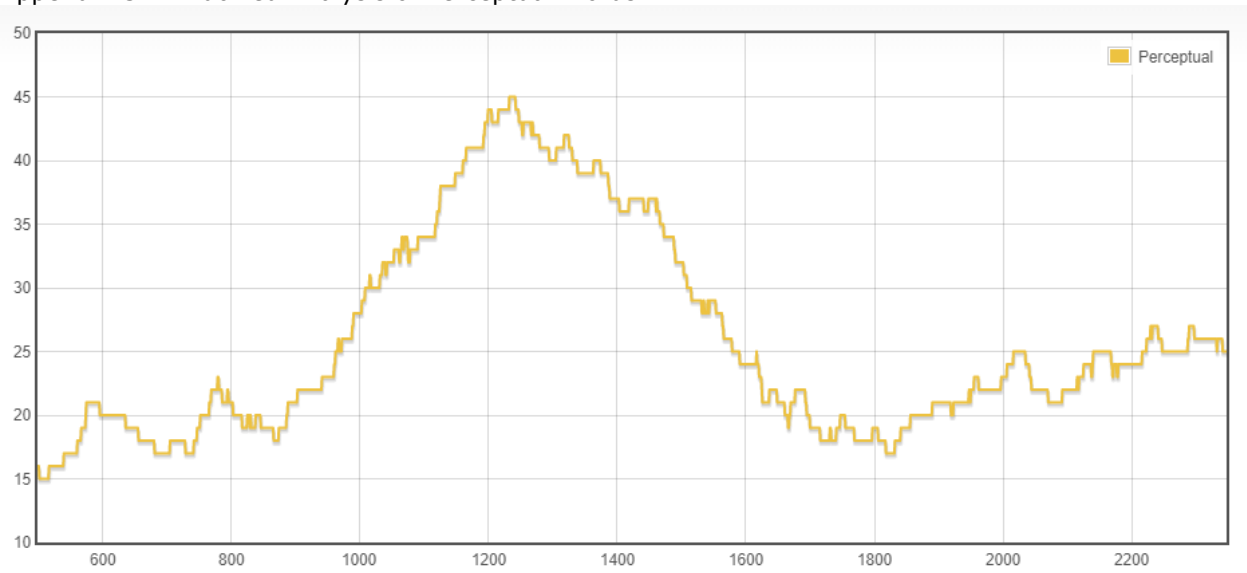
Appendix A: LAT user interface



Appendix B: Windowed Analysis of Positive and Negative Emotions



Appendix C: Windowed Analysis of Perceptual Words



Privacy Considerations Throughout the Data Life Cycle

James J. Pomykalski
pomykalski@susqu.edu
Sigmund Weis School of Business
Susquehanna University
Selinsgrove, PA 17870, USA

Abstract

Due to recent technological advances in the Internet-of-Things (IoT) and data analytics algorithms, the issues of data privacy are again at the forefront of many enterprise strategies. Data privacy has been seen, throughout the years as a legal, public policy and, most recently, a risk management issue. In this paper, we begin the discussion of data privacy at the operational level by addressing operational initiatives and activities to be implemented. More specifically, these initiatives and activities are discussed across a data life cycle model. The focus of this paper is to examine the characteristics of each data life cycle stage and propose a series of organizational data management processes to address data privacy.

Keywords: Privacy, Strategy, Data Management, Risk Management, Organizational Processes, Consumer Data

1. INTRODUCTION

Despite the long history on the discussion of the issue of the "right to privacy", the privacy issue still remains largely ill defined. Nissenbaum (2009) discusses privacy through the lens of contextual integrity because of the changing nature and culturally varying use of the term (Nissenbaum, 2009). Brandeis and Warren (1890) offered the most succinct argument for the right to privacy in 1890 when, against a growing backdrop of technological advances, they argued that privacy is simply "the right to be left alone".

In today's business environment privacy, specifically data privacy is an important issue for many consumers and enterprises. Some commentators have suggested that enterprises should "abandon privacy as an organizational framework" (Bamberger & Mulligan, 2015, p. 22); largely due the issue of contextual integrity (Nissenbaum, 2009). However, many enterprises are addressing privacy through the development of a data strategy, which is becoming a necessary part of an overall enterprise strategy. "Data was once critical to only a few back-office processes, such as payroll and accounting. Today it is central to any business, and the importance of managing it strategically is only growing." (DalleMule & Davenport, 2017, p. 121).

A data strategy "helps by ensuring that data is managed and used like an asset. It provides a common set of goals and objectives across projects to ensure data is used both effectively and efficiently" (SAS, Inc., 2018, p. 4). Specifically, DalleMule and Davenport's (2017) data strategy framework includes both a data offense and data defense component.

"Data defense and offense are differentiated by distinct business objectives and the activities designed to address them" (DalleMule & Davenport, 2017, p. 114). The data offense component deals with the use of data for increasing profitability, revenue, and customer satisfaction, while the data defense approach focuses on minimizing downside risk for the enterprise. Data defense includes activities that ensure compliance with regulations and the integrity of the data; including privacy (DalleMule & Davenport, 2017); the focus of this paper.

However, in order to execute an effective data strategy an enterprise must understand the path data takes within the enterprise. Chisholm (2015) has proposed a seven-stage data life cycle (DLC). The DLC attempts to characterize the activities performed on data as it helps to inform the business processes within an enterprise. This paper examines the activities that take place in each of the phases of the DLC and proposes measures and activities necessary for an enterprise to manage data privacy effectively.

Before proceeding further, it is important to discuss the use of terms data and information. The author acknowledges that data and information have distinct meanings, namely that information is data placed in context. However, in this paper, especially in reference to the work of other scholars, the words data and information are used interchangeably. In addition, at some point during the data life cycle data does become information therefore the privacy of the data as well as the privacy of derived information are essentially synonymous. Future work is already underway to examine the contextual nature of data privacy.

This paper serves as an initial examination of the of the characteristics of each phase of the Chisholm DLC model and proposes a series of organizational data management processes, using known practices as well as existing regulations, to address the privacy of data within an enterprise.

This paper begins by briefly reviewing the literature around data privacy in terms of public policy, legal, and risk management frameworks. In section three, the Chisholm model (Chisholm, 2015) is described in more detail. In sections four through nine, privacy concerns/issues across each of the seven stages of the DLC are discussed in more detail. Section 10 briefly addresses future work in this area, while the final section presents a summary and some conclusions from this work.

2. LITERATURE REVIEW

In the early 1970s, the privacy literature concentrated on privacy as a public policy issue; that citizens had the right to privacy. Regan has stated that privacy should be framed as "a common value, a public value, and a collective value" (Regan, 1995, p. 241). She further argued that society is better when privacy is considered. Nissenbaum (2009) argues that privacy is a social good. She states that the use of analytics can breach the privacy of individuals and others, such as in the 23andMe (Belluz, 2014) or the Target case (Duhigg, 2012). The 23andMe case involves the matching of genetic testing material by police to privately developed DNA databases. The Target case related the use of data to identify pregnant customers. Target used this information to send coupons to these customers; one customer was a 17-year-old girl whose father was not aware of her pregnancy. Both of these cases serve as examples when examining the individual stages of the Chisholm mdoel.

The legal discussion of information privacy per-dates the work of both Regan (1995) and Nissebaum (2009). Following on the work of Brandeis and Warren (1890), privacy was defined further by Westin (1967) where he initiated the concept of "informational self-determination" which means that individuals should have the right to determine the extent of the use of their information. While this does not show privacy as a public policy issue, it does assert that individuals should control privacy. This issue of consent, one of the key principles within the European Union's General Data Protection Regulation (GDPR) (Information Commisioner's Office (ico.)), is examined later.

The issue of privacy has become more concerning with the acceleration of information technologies (Shaw, 2009; Johnson & Miller, 2009). Shaw (2009) believes that technological advances have made people reconsider the public/private distinction and has led the legal profession to become more concerned as well. Solove (2005) created a taxonomy of activities that invade the privacy of individuals. This taxonomy includes 16 privacy harms ranging from information aggregation, insecurity of information, and disclosure and exposure. Another issue, data capture through means of surveillance, deals with the on-going capture of data on individual; Zuboff (2019) updates and extends the discussion on nature of surveillance activities undertaken with information technology.

The underlying principles of informational self-determination are the foundation of the European Union's General Data Protection Regulation (GDPR) that went into effect in May 2018. One aspect of this regulation focuses on data protection as it relates to ensuring people can trust an organization to use their data fairly and responsibly; this is a practical level of the fundamental right to privacy (Information Commisioner's Office (ico.), n.d.). The GDPR is built on seven underlying principles of data protection:

1. Lawfulness, fairness and transparency—enterprises have identified an appropriate bases for data collection and processing, the enterprise considers how processing data will impact individuals—which includes the issue of consent—and that enterprises are open and honest informing the public on data usage,
2. Purpose limitation—enterprises have clearly identified and documented the purpose for data collection and processing,
3. Data minimization—enterprises collect only the data they specify; data must have a specified business purpose,
4. Accuracy—enterprises build in checks for data accuracy throughout the life cycle,
5. Storage limitation—enterprises define, maintain and adhere to data collection and retention policies,
6. Integrity and confidentiality—enterprises maintain strict security measures to protect an individual's data, and
7. Accountability—requires enterprises to take responsibility for personal data, in their possession, and demonstrate/document compliance.

These underlying principles are key to the GDPR and the issue of data protection and privacy. Each of these principles have an impact on data across the data life cycle.
Finally, a third area that warrants discussion is the practical operations and policies used by enterprises; namely C-level executives like Chief Privacy Officers (CPO); although a recent report shows that many enterprises do not have a formal CPO and are therefore struggling to deal with data privacy operational policies (Howard, 2020)

Hilliman states, "data governance oversight [especially in terms of privacy and security] should exist in an interdisciplinary and accountable setting" (Hilliman, 2013, p. 136). Bamberger and Mulligan (2015) examine the practical aspects used by, both US and European, CPOs in safeguarding privacy. These CPOs view their role as strategic; CPOs spend time attempting to integrate "privacy concerns throughout decision making about firms goals, products, and services ensuring a voice on privacy matters is heard at the [executive] table" (Bamberger & Mulligan, 2015, p. 78). Another vital part of the job of a CPO is to interpret and understand the external environment in which their enterprise operates. CPOs spend time interacting and building relationships with privacy regulators, including the governmental and NGO privacy groups.

"To the extent privacy governance requires the dynamic, 'learning' approach that many [CPOs] described, privacy is increasingly framed as a part of the evolving practice of risk management" (Bamberger & Mulligan, 2015, p. 81). In this way, privacy governance is about creating governance structures for operational managers, then the role of the privacy team is to monitor and audit decision-making activities. As one CPO said it, "my team is not responsible for compliance, they're responsible for enabling the compliance of the business" (Bamberger & Mulligan, 2015, p. 84).

3. CHISHOLM DATA LIFE CYCLE MODEL

A data life cycle model fits well into this strategic view of privacy. The data life cycle does not necessarily define all the specific processes involved in data management, however, it does provide "high-level", i.e., strategic, understanding of the activities within that stage regarding enterprise data.

While the Chisholm (2015) model is not the only data life cycle model available (National Network of Libraries of Medicine (NNLM), nd), the seven stage model best fits this research effort for three primary reasons.

First, the Chisholm model has generic applicability and specifically addresses the issue of data governance; the other DLC models address the needs of handling library and/or (National Network of Libraries of Medicine (NNLM), nd) research data. Data governance consists of the development of "a system of decision rights and accountabilities for information-related processes" (Thomas, 2014, p. 3).

Second, the Chisholm model has clearly described the separation of the stages and the associated activities. The model phases depict "logical dependencies and not actual data flows" (Chisholm, 2015, "Critique"); data are harder to capture and are informed by enterprise business processes.

Finally, the model phases focus on specific data governance activities that are unique to each phase; these are shown in Figure 1 below.

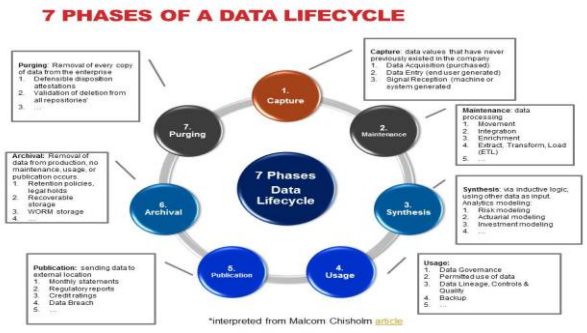


Figure 1: Data Governance in the Data Life Cycle

The focus and the importance, for its applicability, of the Chisholm model is to ensure sound data governance principles in every stage of the DLC (Chisholm, 2015, "Critique"). In the following sections, each phase of the Chisholm model—Figure 1—is introduced and the specific privacy concerns/issues in terms of data governance are addressed.

4. CAPTURE

The first stage, capture, deals with the initial, and original, entry of the data into the enterprise. Chisholm specifically mentions three primary means for data capture: data acquisition, data entry, and signal acquisition. Data acquisition and signal acquisition involve the capture/acquisition of data by an enterprise from outside of the firewalls (Chisholm, 2015), data entry is data generated internally.

In this phase, the enterprise has made a conscious effort to incorporate "captured" data into their infrastructure; the data fits their business needs. The first three GDPR principles—lawfulness, fairness, and transparency; purpose limitation; and data minimization—are most applicable in this phase. The enterprise decides to collect specific data items based on identified appropriate bases for data collection; by identifying and documenting the purpose for the data collection and usage. These criteria can be clearly shown in the documentation of data requirements and business need. In addition, adherence to privacy standards, set by management, can only be achieved through the establishment of clear priorities and formal policies. These policies should proactively define and align the rules for data collection, provide on-going services to data stakeholders, and react to and resolve issues arising from non-compliance (Thomas, 2014). In addition, in terms of transparency of the data capture, enterprises should develop "guidelines that align the interests of companies and their customers, and ensure that both parties benefit from personal data collection" (Morey, Forbath, & Schoop, 2015, p. 100). A survey of 900 individuals—from five different countries—showed that consumers are largely unaware of the type of data that is captured routinely by firms (Morey, Forbath, & Schoop, 2015).

Privacy concerns, due to increased data capture, through surveillance activities, have escalated especially since the late 1980s. In fact, in 1986 only about one percent of all data was digitized. However by 2013 nearly 98% of all data was in digital form (Zuboff, 2019); this increased the opportunity for many enterprises to collect data first and then postulate a purpose second. In her recent book, Zuboff (2019) suggests that this intensive digitization, along with increased use of information technologies, more powerful than those suggested by Shaw (2009), have led to an increase in "surveillance capitalism". Surveillance capitalism is "not technology; it is logic that imbues technology and commands it into action. Surveillance capitalism is a market form that is unimaginable outside the digital milieu" (Zuboff, 2019, p. 15).

5. MAINTENANCE & SYNTHESIS

The next two phases maintenance and synthesis, deal with preparing the data for further analysis. The maintenance phase involves the storage, "integration, cleansing, enrichment, changed data capture, as well as the familiar extract-transform-load processes" (Chisholm, 2015, "Data Maintainence"). New data values are derived in this stage by using deductive logic. While in the synthesis phase, new data values are created by using, not deductive logic but inductive logic which requires the use of expert judgment, experience, and/or intuition as part of the logic. (Chisholm, 2015). The privacy issues and concerns of both of these phases are similar so the discussion covers both of these stages simultaneously.

In these stages, captured data is combined with existing enterprise data. The primary privacy concern at this stage is that of data anonymization or data pseudonymization. A lengthy discussion of the primary difference between an anonymization process and a pseudonymization, the activities in each process, and the risks—to security—associated with each process appears in a document published by the Information Commissioner's Office (Information Commissioner's Office (ico.), 2012). The document provides guidance to enterprises that need to anonymize data, it also helps identify

issues to consider ensuring effective anonymization of personal data, and describes the legal tests required in the GDPR.

The data creation accomplished in these phases is the result of efforts to classify vast amounts of data into homogenous clusters or data discovery through insight creation techniques such as data mining; much of this latter work occurs in the synthesis phase. The establishment and adherence to enterprise and external policies regarding the integrity and confidentiality of this data are of vital importance. There is a higher risk that these new data elements, containing personally identifiable data on specific individuals or groups, increase the need to understand the issues of data anonymization. In addition, the establishment of individual accountability to compliance with the policies needs clear documentation and enforcement throughout the enterprise.

This stage utilizes the enterprise's experience in performing integration, cleansing and enrichment of data through past efforts in creating a data warehouse. Data governance efforts that focus on data privacy, as well as data security are critical at this stage. The data governance efforts may be limited to only certain data—i.e., master data—and responsible data governance personnel will be accountable for:

- access management and security requirements,
- alignment of frameworks and data governance initiatives,
- risk assessment and development of risk management plans,
- Enforcement of regulatory, contractual, architectural compliance requirements,
- identification of stakeholders, establishment of decision rights, and clarification of accountabilities (Thomas, 2014).

Another issue that requires attention in this stage is the development of the use of systems that automate data integration. Recent research has led to the development of machine learning/data mining systems that build in privacy preservation. Clifton et al. (2004) call for the need for further research, development, and use of privacy preserving systems especially in data integration and sharing efforts.

There are two recent examples of data synthesis efforts that draw concern. The first example involves the use of facial recognition software in US airports for the boarding process (Funk, 2019). As passengers board their facial image is compared to images in a database of photos taken from visas, passports, and related immigration applications. While many believe that this increases security, many privacy concerns exist with facial recognition software systems. First, these systems do have a 99 percent accuracy rating for identifying white men, while the error rate for females and people with darker skin tones is as high as 35 percent (Funk, 2019). In this case, females and minorities have an increased likelihood of being targeted for additional screening measures. The second concern raised with the use of facial recognition is the use of the database. Funk (2019) states "Americans should be concerned about whether images of their faces collected by this program will be used by companies and shared across different government agencies. Other data collected for immigration purposes—like social media details—can be shared with federal, state, and local agencies. If one government agency has a database with facial scans, it would be simple to share the data with others" (Funk, 2019, Para 10). In this case, it is not the data that is misused but the technology utilizing the data that could result in privacy violations (Funk, 2019).

In the second, more widely publicized case, an open source DNA database was used by police in California to track down the Golden State Killer (Molteni, 2018). The case involves questions of privacy because of the frequent use of genetic testing through firms like 23andMe and Ancestry. In this case, the police matched the killer's genetic profile, obtained through old crime scene samples, to samples in the open source DNA database. Once a pool of individuals was obtained they used other clues—familiar connections, sex, age, place of residence—to rule out suspects. Eventually their search found a single suspect, which was confirmed through matching his DNA to the crime scene samples (Molteni, 2018). While some people may be comfortable with this police tactic, others, especially privacy experts, are concerned.

6. USAGE

The usage stage of the DLC is described as "the application of data as information to tasks that the enterprise needs to run and manage itself" (Chisholm, 2015, "Data Usage").

Chisholm (2015) goes on to point out that with any data there may be additional permissions that are needed through the data governance structure. Specifically, Chisholm (2015) refers to legal or regulatory restrictions imposed by outside agencies that restrict the use of the data

to certain business processes; as consent issues were addressed in an earlier stage.

The other issue left to consider at this stage is the use of privacy-preserving data models, namely machine learning/data mining, which will further protect the privacy of individuals in this stage. The current state of these privacy-preserving systems is outlined in a recent post by engineers at Dropout Labs (Mancuso, DeCoste, & Uhma, 2018).

In this report, the engineers discuss the impact of machine learning systems on privacy and the ongoing research and development in privacy-preserving systems. The authors view the development of these systems, especially open source systems, as positive and encourage enterprises to examine and utilize these systems.

7. PUBLICATION

Chisholm (2015) describes the data publication stage as the point in time where data is sent to a location outside of the enterprise. Data publication is the stage that represents the point in time where data is beyond recall or correction. This stage also covers the breach of data from internal systems.

Data publication is often the first time that consumers may react to lack of transparency in data capture, synthesis, and usage of their personal data. For example, Duhigg (2012) examines Target's use of personal data to identify pregnant customers, in particular a minor female. Target issued a public apology and quietly withdrew the program.

Data breaches are a form of data publication that cause enterprises major issues. Data breaches are often when consumers find, after it has been compromised, that an enterprise maintained their personal data. Data breaches carry both legal and public relations implications, which cost enterprises in terms of not only dollars but also reputation, brand, consumer trust and, potentially, future sales (Bowers, 2011). Data breaches may also have implications more directly on consumers in the form of identity theft possibilities.

8. ARCHIVAL

Data archival is "the copying of data to an environment where it is stored in case it is needed again in an active production environment, and the removal of this data from all active production environments" (Chisholm, 2015, "Data Archival"). A data archive is a place to store data once its useful life has be exhausted. The data remains as part of the data infrastructure and can be restored if necessary, however, no further maintenance, usage or publication occurs.

While many of the same concerns about security and data anonymization/pseudonymization exist at this stage to those discussed in the maintenance and synthesis stages, one of the specific issues questions facing enterprises with data archival is the development of a data retention policy. The GDPR address the issue of data retention through the storage limitation principle, which states that for personal data enterprises should (1) not keep the data longer than needed, (2) be able to justify the length of retention for the data, (3) have a policy setting group set compliance standards that are well documented (Information Commissioner's Office (ico.), 2012).

9. PURGING

The final stage of the data life cycle is the purging of all instances of the data item from the enterprise and its systems (Chisholm, 2015). Data deletion is difficult for enterprises but, for much of the data in current systems, "the costs of keeping data are higher than you think, and the benefits are lower…there is a chance it will be harmful—like being lost in a breach or subpoenaed in a lawsuit" (Branscombe, 2019, Para 6). Branscombe (2019) goes to state that about a third of data, stored in current data centers, is likely redundant, obsolete or trivial and since it holds no business value; it should be purged.

Purging data can be cost-effective as well as risk reducing. Additional costs may be incurred due to additional anonymization/ pseudonymization processes. In addition, the risk of having the data lost and de-identified often outweigh the loss of this data. In fact, Joshua de Larios-Heiman, chair of the California Lawyers Association Internet & Privacy Law Committee, warns that enterprises should think of this data as uranium rather than oil; as assets that could become toxic (Branscombe, 2019).

10. FUTURE WORK

As discussed in the introduction, the question of the transition of data to information is important in terms of privacy. The application of the enterprise's specific purpose, knowledge, and activities imposed on data can lead to different privacy concerns. The investigation of how these situations can be modeled will be the subject of future research.

This work has two additional natural extensions. First, the focus of future work can shift from looking at data governance activities that help ensure privacy to those that help ensure data quality. Data quality is still a part of the data defense strategy (DalleMule & Davenport, 2017) and is a natural extension of this work on data privacy. Data quality has been cited, as one of the key issues that enterprises face, especially with the use of analytics, in ensuring that the data that is being used is of the highest quality; as defined by the end users (Kwon, Lee, & Shin, B, 2014; Hazen, Boone, Ezell, & Jones-Farmer, 2014). Understanding how users view data quality in each of the different phases of the life cycle is important in understanding how data quality improvement is accomplished.

Second, further delineation of the data governance activities described in each of the seven phases of the DLC needs to be undertaken. This would include the use of current practices found within existing enterprises in managing privacy throughout the DLC. Subsequent work can focus on specific key phases in the DLC individually.

11. SUMMARY/CONCLUSIONS

This paper addressed an initial set of data governance activities and initiatives that need to be addressed within each stage of the data life cycle. While this is not the first work to begin to address these data governance activities, it is the first look at data governance with respect to privacy across particular stages of a data life cycle.

This work illustrates the number of issues and concerns that must be addressed when it comes to data privacy. While, in the US the legal oversight concerning data privacy is scattered because individual state laws are enforced. For example, California will have a new law in effect, the California Consumer Privacy Act (CCPA) (Californians for Consumer Privacy, n.d.), in January of 2020 to address data privacy.

However, the recent enactment of the European Union's (EU) General Data Protection Regulation (GDPR) has come a long way in forcing US enterprises to address privacy more rigorously due to the fact that any enterprise doing business within the EU must adhered to these new standards.

The problem is that while some enterprises are dealing with data privacy as a risk management issue (Bamberger & Mulligan, 2015) there seems to be little progress being made on privacy as a public policy issue.

Overall, this paper focused on the examination of the characteristics of each data life cycle stage and the unique activities that need to be addressed in each stage with regard to data privacy.

12. ACKNOWLEDGEMENTS

The author would like to thank the anonymous editors of this paper for their thoughtful, constructive, and detailed comments and feedback. In addition, conference participants assisted the author in clarifying future work efforts. This version of the paper is not only more readable, but also provides a clearer and better-articulated understanding of privacy activities across the data life cycle.

13. REFERENCES

Bamberger, K. A., & Mulligan, D. K. (2015). Privacy on the Ground: Driving Corporate Behavior in the Inited States and Eurpoe. Cambridge, MA: The MIT Press.

Belluz, J. (2014, December 18). Genetic testing brings families together, and sometimes tears them apart. Vox, p. 2014. Retrieved June 21, 2019, from https://www.vox.com/2014/9/9/6107039/23andme-ancestry-dna-testing

Bowers, T. (2011). Security as Business Risk: How Data Breaches Impact Bottom Lines. Experian.

Brandeis, L., & Warren, S. (1890). The Right to Privacy. Harvard Law Review, IV(5).

Branscombe, M. (2019, July 3). Data Deletion: Your data Strategy' Greatest Defense. CIO Magazine. Retrieved July 24, 2019, from https://www.cio.com/article/3405129/data-deletion-your-data-strategys-greatest-defense.html

Californians for Consumer Privacy. (n.d.). About the Law. Retrieved July 24, 2019, from CAPrivacy.org: https://www.caprivacy.org/about

Chisholm, M. (2015, July 9). Seven Phases of a Data Life Cycle. Information Management, p. n.p. Retrieved May 28, 2018, from https://www.information-

management.com/news/7-phases-of-a-data-life-cycle

Clifton, C., Kantarcioglu, M., Doan, A., Schadow, G., Vaidya, J., Elmagarmid, A., & Suciu, D. (2004). Privacy-preserving data integration and sharing. 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (pp. 19-26). Paris: ACM. Retrieved September 27, 2019

DalleMule, L., & Davenport, T. (2017, May-June). What's Your Data Strategy? Harvard Business Review, pp. 112-121.

Duhigg, C. (2012, February 16). How Companies Learn your Secrets. New York Times, p. 2012. Retrieved from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

Funk, A. (2019, July 2). I Opted Out of Facial Recognition at the Airport—It Wasn't Easy. Wired. Retrieved July 22, 2019, from https://www.wired.com/story/opt-out-of-facial-recognition-at-the-airport/

Hazen, B., Boone, C., Ezell, J., & Jones-Farmer, L. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, 154, 72-80.

Hilliman, C. A. (2013). Data Privacy, Security, and Compliane through Data Governance. In N. Bhansali, Data Governance: Creating Value from Information Assets (First ed., pp. 125-148). Boca Raton, FL: CRC Press.

Howard, J. (2020, January 13). How effective are privacy programs? Retrieved January 14, 2020, from https://jhoward.us/2020/01/13/how-effective-are-privacy-programs/

Information Commisioner's Office (ico.). (n.d.). Introduction to Data Protection. Wilmslow, Cheshire, UK. Retrieved June 26, 2019, from https://ico.org.uk/for-organisations/guide-to-data-protection/introduction-to-data-protection/some-basic-concepts/

Information Commissioner's Office (ico.). (2012). Anonymisation: Managing Data Protection Risk Code of Practice. Wilmslow, Cheshire (UK): Information Commissioner's Office (ico.). Retrieved July 20, 2019, from https://ico.org.uk/media/for-organisations/documents/anonymisation-code.pdf

Johnson, D. G., & Miller, K. W. (2009). Information Flow, Privacy, and Surveillance. In D. G. Johnson, & K. W. Miller, Computer Ethics (pp. 81-108). Upper Saddle River, NJ:: Prentice Hall.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International journal of information management, 34(3), 387-394.

Mancuso, J., DeCoste, B., & Uhma, G. (2018, January 10). Privacy-Perserving Machine Learning 2018: A Year in Review. Retrieved September 27, 2019, from https://medium.com/dropoutlabs/privacy-preserving-machine-learning-2018-a-year-in-review-b6345a95ae0f

Molteni, M. (2018, April 27). The Creepy Genetics Behind the GOlden State Killer Case. Wired. Retrieved July 22, 2019, from https://www.wired.com/story/detectives-cracked-the-golden-state-killer-case-using-genetics/

Morey, T., Forbath, T., & Schoop, A. (2015). Customer data: Designing for transparency and trust. Harvard Business Review, 93(5), pp. 96-105. Retrieved July 24, 2019

National Network of Libraries of Medicine (NNLM). (nd). Data Life Cycles: Research Data Cycles and Guides. Retrieved June 5, 2019, from National Institute of Health: https://nnlm.gov/data/data-life-cycles

Nissenbaum, H. (2009). Privacy in Context: Technology, Policy, and the Integrity of the Social Life. Palo ALto, CA: Stanford University Press.

Regan, P. (1995). Legislating Privacy: Technology, Social Vaues, and Public Policy. Chapel Hill, NC: University of North Carolina Press.

SAS, Inc. (2018). The Five Essential Components of a Data Strategy. Cary, NC: SAS Inc.

Shaw, J. (2009, September). Exposed: The eroision of privacy in the Internet era. Cambridge, MA. Retrieved June 26, 2019, from https://harvardmagazine.com/2009/09/privacy-erosion-in-internet-era

Solove, D. J. (2005). A taxonomy of privacy. University of Pennsylvania Law Review, 154, 477-560.

Thomas, G. (2014, June 11). The DGI Data Governance Framework. Retrieved 2019, from The Data Governance Institute: http://www.datagovernance.com/dgi-data-governance-framework/

U.S. Geological Survey. (nd). Data Management. Retrieved June 5, 2019, from USGS.org: https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle

Westin, A. F. (1967). Privacy and Freedom. New York, NY: Atheneum.

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Fronier of Power. New York, NY: PublicAffairs.