University of Texas at El Paso
DigitalCommons@UTEP

Open Access Theses & Dissertations

2019-01-01

A Bottom-Up Modeling Methodology Using Knowledge Graphs For Composite Metric Development Applied To Traffic Crashes In The State Of Texas

Daniel Michael Mejia
University of Texas at El Paso, dmmejia2@gmail.com

Follow this and additional works at: https://digitalcommons.utep.edu/open_etd

 Part of the Computer Sciences Commons

Recommended Citation
Mejia, Daniel Michael, "A Bottom-Up Modeling Methodology Using Knowledge Graphs For Composite Metric Development Applied To Traffic Crashes In The State Of Texas" (2019). *Open Access Theses & Dissertations*. 114.
https://digitalcommons.utep.edu/open_etd/114

This is brought to you for free and open access by DigitalCommons@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of DigitalCommons@UTEP. For more information, please contact lweber@utep.edu.

A BOTTOM-UP MODELING METHODOLOGY USING KNOWLEDGE GRAPHS FOR COMPOSITE METRIC DEVELOPMENT APPLIED TO TRAFFIC CRASHES IN THE STATE OF TEXAS

DANIEL MICHAEL MEJIA

Doctoral Program in Computer Science

APPROVED:

Natalia Villanueva-Rosales, Ph.D., Chair

Monika Akbar, Ph.D.

Ruey Long Cheu, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

Copyright ©

by

Daniel Mejia

2019

Dedication

To my family. For the Glory of God.

A BOTTOM-UP MODELING METHODOLOGY USING KNOWLEDGE GRAPHS FOR COMPOSITE METRIC DEVELOPMENT APPLIED TO TRAFFIC CRASHES IN THE STATE OF TEXAS

by

DANIEL MICHAEL MEJIA, BSCS, MSCS

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2019

Acknowledgments

En aquel tiempo, I thought that pursuing a Ph.D. would be fun; it was fun, but it also took a lot of hard work, dedication, and discipline. I have always wanted to pursue a doctorate degree and would have not been able to achieve this huge milestone in my life had it not been for the grace of God and my family. I would first like to give thanks to God for giving me the strength to continue on with my education and opening doors for me in my life. I would like to give thanks to my parents, whom without their support I would have achieved success. My parents have given me the support, guidance, and love that I have needed to never give up and push myself each day. I would like to thank my siblings for always supporting me in my educational journey. I would like to thank all of my friends for their support during my studies. It is through my family and friends that I am continuously motivated to do the very best that I could possibly do and achieve excellence.

I would like to acknowledge my Ph.D. faculty advisor Dr. Natalia Villanueva-Rosales for her insight and support with this research. I would also like to thank Dr. Kelvin Cheu and Dr. Monika Akbar for their feedback and guidance with this work. A very huge thank you to Diego Aguirre for our many conversations that helped guide me as a researcher and a professional; moreover, the support-system that we established with each other is something I am truly grateful for. Thanks to all of the members of iLink for their support in my research. Thank you to all of the participants of the user evaluation study conducted as part of this research. Thank you to the UTEP writing center for their editing and revision to this dissertation. Thank you to all of the faculty and staff of the Computer Science department that helped me throughout this educational journey. This work used resources from Cyber-ShARE Center of Excellence, which is supported by the National Science Foundation grant number HRD-0734825.

Abstract

Data is a key factor for understanding real-world phenomena. Data can be discovered and integrated from multiple sources and has the potential to be interpreted in a multitude of ways. Traffic crashes, for example, are common events that occur in cities and provide a significant amount of data that has potential to be analyzed and disseminated in a way that can improve mobility of people, and ultimately improve the quality of life. Improving the quality of life of city residents through the use of data and technology is at the core of Smart Cities solutions. Measuring the improvement that Smart Cities provide usually relies on data collection and analytics before and after the implementation of such solutions.

Through a methodological approach, implicit information about mobility, in particular traffic crash data, can be discovered and used in the interpretation and dissemination of information through different data views, such as metrics and narratives, thus fostering the gain of knowledge. In this work, a novel modeling methodology for traffic crashes was developed, namely the Bottom-Up Modeling (BUM) methodology. This methodology integrates publicly available mobility data, proposes a data model implemented in a knowledge graph that includes semantic annotations, and produces a composite metric called the Critical Composite Index (CCI). The CCI uses weighted criteria values to make each crash comparable to others with similar data provided. The BUM methodology was applied to model traffic crashes between different geographic locations in Texas.

The resulting methodology enables the creation of metrics for use by many stakeholders, particularly non-domain experts. The use of the BUM methodology to generate different perspectives of the crash data is addressed through the generation of different data views (i.e.,

metrics and data narratives). Moreover, an ontology was developed based on the knowledge graph to formalize the proposed data model to, verify logic consistency and infer implicit information with the use of a generic description logic reasoner. The CCI metric was evaluated by comparing against currently used frequency metrics and a user-evaluation survey. Evaluation results show that the CCI provides improved knowledge gain over currently used metrics. This work contributes to data science research, using an interdisciplinary approach that involves Computer Science techniques, mathematics and domain expertise to address complex challenges, such as those in converting cities to Smart Cities.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 MOTIVATION

Smart Cities is an area of research that transcends multiple fields across sciences including social sciences and engineering. In the Computer Science perspective, the contribution of this research gap being filled is understanding how data can be transformed into knowledge through a bottom-up approach that uses data points as a source for the representation of facts. Data is at the center of understanding real-world issues. Data can be discovered from many sources and has the potential to be represented interpreted in a multitude of ways. As a result of data being represented interpreted in many ways, it may become difficult to understand such data regardless of any person's expertise in the domain. Data provides factual information without analysis whereas useful knowledge has a clear way of being interpreted and disseminated (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The transformation of data to knowledge occurs when data is analyzed and denoted with information that can be understood by humans for dissemination and further analysis. Mirroring a concept known as exploratory data analysis (Vila et al., 2016) Smart Cities can be improved; by understanding facts, it is possible for that same data to be used to represent knowledge in a singular way (e.g., through a composite metric). This work will focus on enhancing the area of Computer Science through a unique data modeling technique, analysis, and metric development to give deeper insight into the improvement of data views – a focus on the inputs that influence the results (output) (Gil & Garijo, 2017), data representation, and data modeling issues and solutions for the domain of Smart Cities, specifically Smart Mobility. Smart mobility, like Smart Cities is not well defined, however it can generally be regarded as using advanced technology and data to help understand and improve movement of people about a city (Neirotti, De Marco, Cagliano, Mangano, & Scorrano, 2014).

Smart Cities are forward-looking "System-of-Systems" (Naphade, Banavar, Harrison, Paraszczak, & Morris, 2011) and are guided by the idea of incorporating technology into the everyday lives of citizens (Caragliu, del Bo, & Nijkamp, 2011) within a city to improve the productivity, sustainability, and efficiency of services and resources for the ultimate goal of improving the quality of life of its citizens. The need for Smart Cities stems from rapid urbanization of both large metropolitan and smaller urban areas; it is expected that by 2050, 68% of the world population will live in a city (UN, 2018). The integration of technology comes in many forms including, but not limited to information and communications technologies (ICT), data analysis, large data sets, real-time systems, the Internet of Things (IoT), and user-centric data. By connecting many different forms of technologies, cities are tasked with improving overall sustainability, efficiency, and productivity.

Previous work in Smart Cities has primarily focused on addressing challenges that are usually immediately visible to researchers, citizens, and policymakers without data analysis and appropriate data representation (Dameri, 2014). The challenges are often seen, and a solution is developed with limited information to determine the improvement that the solution would provide or if it were actually necessary. Research has focused on determining best practices to address various problems; these best practices are often reported by the developing Smart City itself (Neirotti et al., 2014). Research is being done to understand challenges that Smart Cities have and finding solutions to them; however, solutions should be validated by data so that they can be measured, compared, and transferred to other geographic locations.

Throughout the world, there have been many different types of metrics created to measure safety, efficiency, and effectiveness, but none of them are widely used beyond the city or country that

initiates them. Many of the countries that have adopted any sort of metrics to measure city performance use the idea of a *triple bottom line*, which has a focus on environmental, economic, and social issues (Mori & Yamashita, 2015) as represented in Figure 1.1.

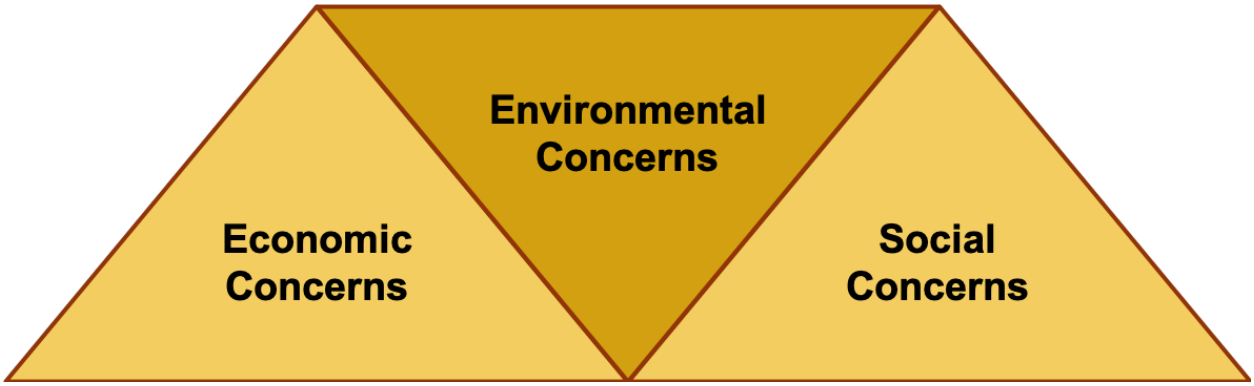


Figure 1.1 *Triple Bottom Line*

The *triple bottom line* plays an important role in the development of metrics and lays the foundation necessary for building standard metrics. Based on work by Larios et al. (Larios, Gomez, Mora, Maciel, & Villanueva-Rosales, 2016), this research proposes the idea of the *four principles of Smart Cities* that are based on the *triple bottom line* (Mori & Yamashita, 2015) as a mechanism to develop new metrics.

The *four principles of Smart Cities* are the following:

1. Efficiency & Productivity – Resources, data, applications, ideas, and research will efficiently, productively, and effectively produce knowledge that enables growth in real-world solutions for the improvement of economic, environmental, and social concerns

2. Reusability – Resources, data, applications, ideas, and research can be reused in multiple Smart Cities focus areas for the improvement of economic, environmental, and social concerns
3. Sustainability – Resources, data, applications, ideas, and research will be sustainable in that they have longevity to be efficiently reused and modified for the continued growth and understanding of Smart Cities for the improvement of economic, environmental, and social concerns
4. Improved Quality of Life – Resources, data, applications, ideas and research will yield solutions for the improvement of economic, environmental, and social concerns that are efficient, productive, reusable, and sustainable – which will improve the quality of life of dwellers in Smart Cities

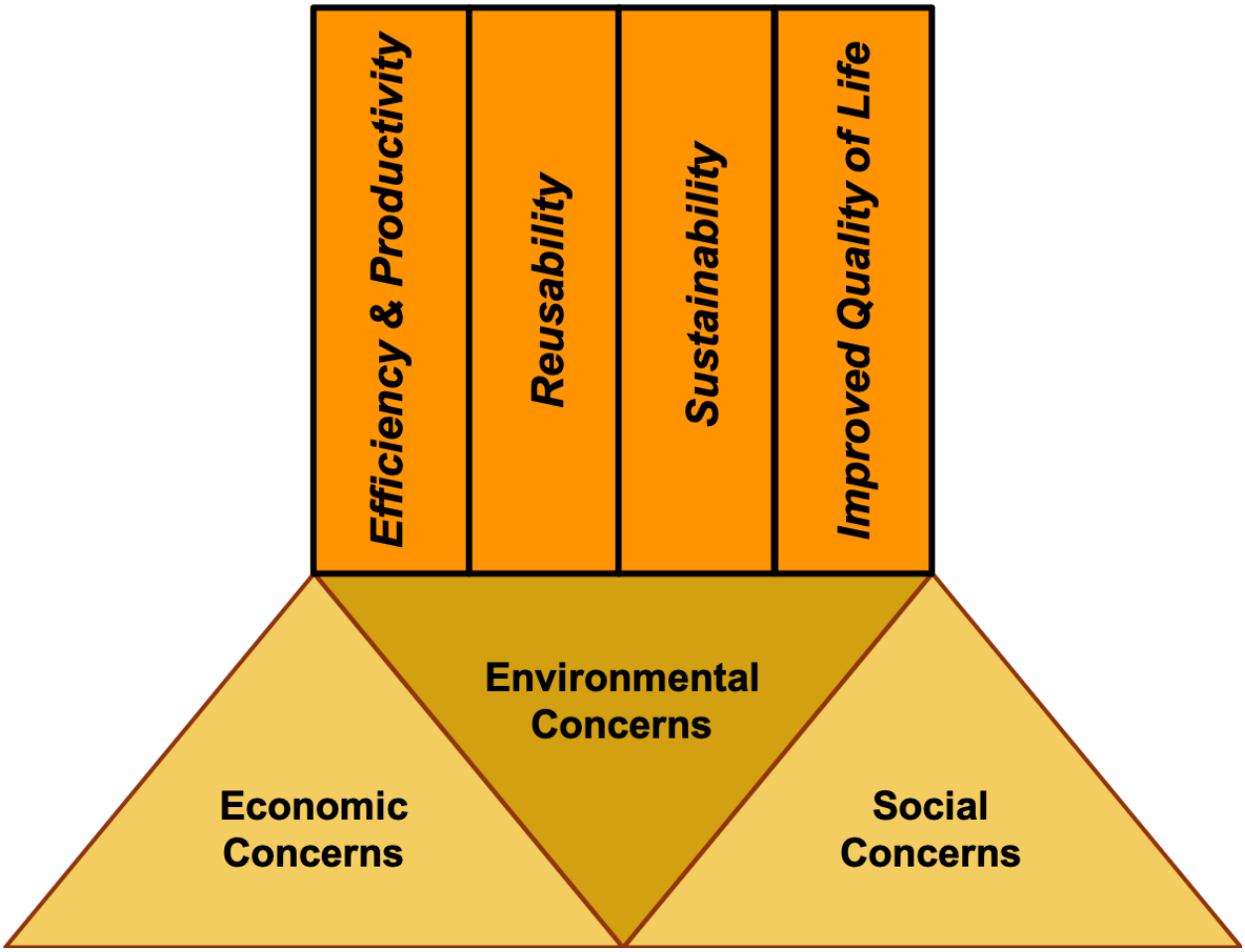


Figure 1.2 *Four principles of Smart Cities* built on the *triple bottom line*

Figure 1.2 shows how the *four principles of Smart Cities* are built upon the *triple bottom line*. The *four principles of Smart* Cities are based on the *triple bottom line* since each of the four principles: efficiency & productivity, reusability, sustainability, and improved quality of life are focused on the improvement of economic, environmental, and social concerns. Many Smart Cities metrics are subjective, insofar as they represent the perception of people about the services that are currently provided and how they improve their everyday lives (Albino, Berardi, & Dangelico, 2015). Current metrics do not necessarily measure how any given solution aligns with the *four principles of Smart Cities*. Though it is important to consider how services are

provided to citizens, it should not be the driving factor behind most metrics; it does not allow for analysis that raw data would.

Since Smart Cities have more than one specific domain focus of research, there are metrics to represent transportation, energy consumption, water use, health, economy, land usage, and several other areas (McMahon, 2002). The International Organization for Standardization (ISO) has the most widely accepted metrics with respect to improving the quality of life, specifically ISO 37120, which has standards for many areas including the economy, education, finance, health, safety, and transportation (Steele, 2014).

Since the late 1990s and early 2000s, there has been a pattern of developing metrics for transportation systems. The metrics that are defined are generally related to the *triple bottom line* principles of focusing on the impacts of the economy, environment, and social concerns (Mihyeon Jeon & Amekudzi, 2005; Mori & Yamashita, 2015). Metrics are a critical form of representing knowledge in a way that is easy to understand. The knowledge gained from the metrics must be derived using a standard, transferrable processing technique, or methodology. Computer Science provides the techniques necessary to develop a methodology to describe a specific domain clearly as shown in Figure 1.3.

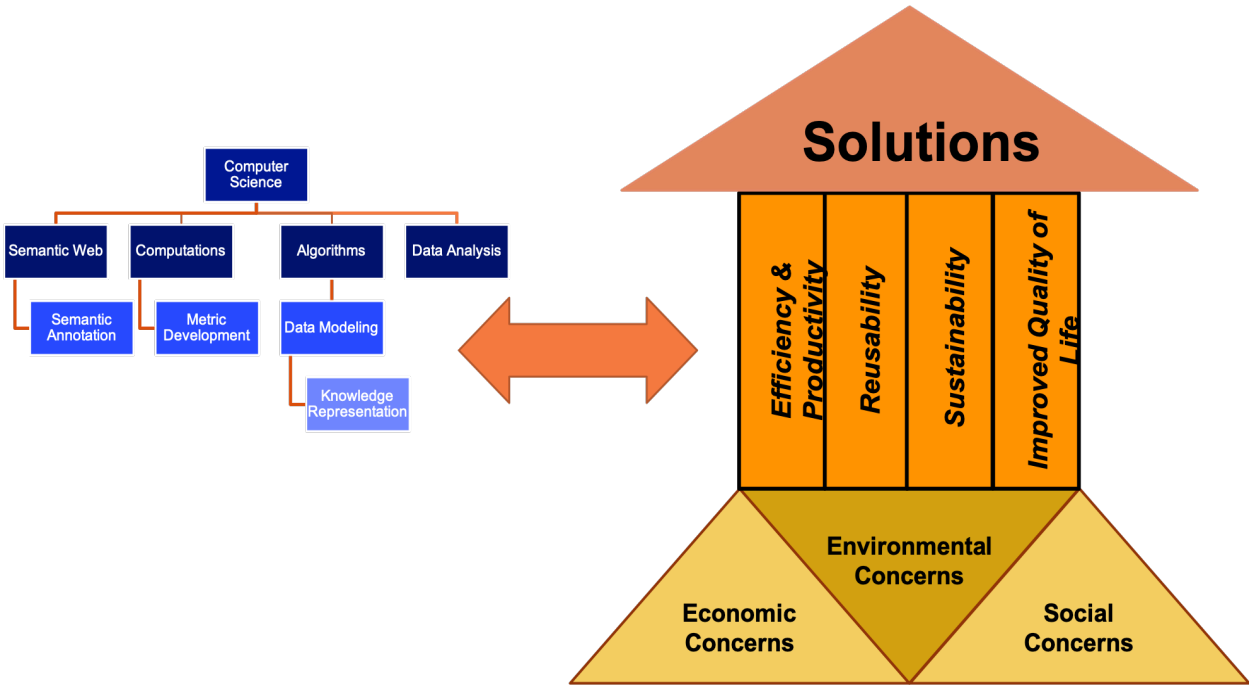


Figure 1.3 Smart Cities domain and the contribution of relevant Computer Science topics yielding improvement

Figure 1.3 shows the intertwining relationship of Computer Science and Smart Cities. As a basis, Computer Science provides a mechanism to represent and model data. In this work, Semantic Web for semantic annotation, computation for metric development, algorithms to provide data modeling and knowledge representation and data analysis is used to describe traffic crashes data. As a result, Computer Science yields the improvement of Smart Cities and in turn Smart Cities provides information for Computer Science problems. Through the development of novel data representation techniques and algorithms, Computer Science yields improvement to Smart Cities through advanced data representation for knowledge gain which are founded on the *triple bottom line* focus of improving economic, environmental, and social concerns. The concerns presented by the *triple bottom line* are built upon to show that through the use of appropriate practices,

Smart Cities should be efficient & productive, reusable, sustainable, and improve the quality of life; through these principles, solutions can be made.

Smart Cities is a broad area of research that has many possible domains that can be focused on, including Smart Mobility. Based on work done by Mihyeon Jeon and Amekudzi (Mihyeon Jeon & Amekudzi, 2005), the United States Department of Transportation (USDOT) has a mission to "serve the United States by ensuring a fast, safe, efficient, accessible and convenient transportation system… and enhance the quality of life of the American people." Table 1.1 shows the work of Mihyeon Jeon and Amekudzi (Mihyeon Jeon & Amekudzi, 2005) describing the mission statements of several selected State Departments of Transportation (DoT); Texas was an addition to the report for additional comparison (TxDOT, 2017). Out of the 16 Departments of Transportation selected, 15 of them explicitly mention safety in their mission statement. 10 out of the 16 explicitly mention that they want to be effective or efficient in delivering proper transportation services to its residents.

Table 1.1 Mission Statements of selected states in the United States with reference to “safety,” “effective,” or “efficient” (Mihyeon Jeon & Amekudzi, 2005)

State Department of Transportation	Mention “Safety” in Mission Statement	Mention “Effective” or “Efficient”
United States DoT	YES	YES
Texas DoT	YES	NO
Florida DoT	YES	NO
Georgia DoT	YES	YES
Indiana DoT	YES	YES
Louisiana DoT	YES	NO

Michigan DoT	NO	NO
Montana DoT	YES	YES
New Jersey DoT	YES	NO
New York DoT	YES	YES
Nevada DoT	YES	YES
Oregon DoT	YES	YES
Rhode Island DoT	YES	YES
South Dakota DoT	YES	NO
Vermont DoT	YES	YES
West Virginia DoT	YES	YES

In the United States, it is clear that safety is at the forefront of most Departments of Transportation; moreover, efficiency and effectiveness are also important to the states. In the focus area of Smart Mobility, safety should be the cornerstone of a majority of its metrics because it is needed to improve the reliable understanding of events on the roadways. Moreover, the Federal Highway Administration has five major goals to achieve between 2016 and 2021 as shown in Figure 1.4: Highway Safety, Improved Mobility, System Performance, Environmental Sustainability, and to prepare for the future (United States Department of Transportation (USDOT), 2016).

Figure 1.4 The Federal Highway Administration (FHWA) goals as defined in (O'Rourke, Beshers, & Stock, 2015)

There are many factors that relate to metrics with respect to traffic crashes, including the severity of the crash. A traffic crash is defined as, "a harmful event occurring on a traffic way that produces injury, death or damage." (Texas Department of Transportation, 2016, 2017) Severity metrics are shown by describing the effects of the crash that occurred, such as injury severity, fatalities, damage, or with respect to the position of the vehicle (Laureshyn, Svensson, & Hydén, 2010; National Safety Council & ANSI, 2017). Furthermore, the frequency of crashes on the roadway and its location also play a role in safety metrics (Hermans, Brijs, Wets, & Vanhoof, 2009).

In the city of El Paso, Texas there were more than 96,500 traffic crashes between 2014 and 2018 that were reported to law enforcement agencies (e.g. local police, sheriff, state trooper) (TxDOT, 2018). The traffic crashes range from minor fender benders to multiple fatalities in a single

crash. Throughout the world, and especially the United States, departments of transportation are working towards providing safe roadways and travel for those who use its roads. The development of semantically annotated metrics is beneficial in providing a larger composite understanding of safety and efficiency in the roadways. The Semantic Web provides meaning to data through the use of vocabularies and semantic descriptions. It provides meaning to data in a way enables humans and computers to work together (Berners-Lee, Hendler, & Lassila, 2001). Department of transportations throughout the United States, as well as road users, will be able to use the proposed metric because data provides a clearer view of what is occurring on the roadways. The knowledge gained through the transformation of data to knowledge will allow for policymakers to understand what is occurring on roadways and a way to use that information to make improvements with respect to safety and efficiency.

This work will enhance the way that data is cleaned, tracked, and used for improved interoperability and support of ad-hoc decisions. These improvements will be implemented through the development of a Bottom-Up Modeling (BUM) methodology and the development of a novel metric.

Through the BUM methodology the transformation of data to knowledge improve the understanding of Smart Cities metrics, namely Smart Mobility. Along with data integration, data analysis and data views, the outcomes of the BUM methodology will provide new insight to traffic crashes by providing a singular standard technique of classifying and comparing them.

1.2 THESIS STATEMENT & RESEARCH QUESTIONS

Thesis statement: The analysis of public available data, such as traffic crashes, can be improved by *1) Creating a systematic, reproducible approach for the representation and integration of relevant data, and 2) introducing a novel metric to describe data amendable to generate information with more details than current metrics, in the case of traffic crashes this includes people involved, location, time, and the external circumstances related to it.*

Research Questions:

Q1. *What do semantically-enhanced data models contribute to data representation and data integration for metric development of publicly available mobility data targeting non-domain expert stakeholders?*

Q2. *How can a modeling methodology contribute to the transformation of data to useful knowledge represented for supporting decision making by non-domain experts?*

Q3. *How can very large public data-sets be used to contribute to quantifiable metrics that are both reusable and comparable to other geographic locations and improve data views?*

a. *How does semantically annotated data (input) improve metric development (output)?*
b. *How do elements of data-sets (input) affect the way the information they provide is understood (output)?*

1.3 RESEARCH GOAL

Goal 1 The goal of this research is to create a systematic approach for modeling publicly available mobility data-sets that enables: i) formal descriptions of information

embedded in data, ii) quantitative and qualitative information extraction, and iii) knowledge discovery for decision making.

The objectives to achieve this goal are:

Objective 1 Create a modeling methodology to formally describe publicly available mobility data-sets using knowledge graphs for knowledge discovery

Task 1.1 Create a bottom-up methodology (BUM) that uses knowledge graphs to semantically describe information of publicly available mobility data-sets

Task 1.2 Apply the BUM methodology to publicly available data about traffic crashes in the state of Texas

Task 1.3 Investigate the use of knowledge graphs as a data-model representation of publicly available data in the context of traffic crashes throughout the State of Texas

Task 1.4 Evaluate the proposed data model with respect to its ability to integrate data, represent formal descriptions through logical rules that enable consistency checking and the use of inference services, and transferability to a mobility dataset from another state.

Objective 2 Create a novel metric aiming to improve understanding of publicly available mobility data-sets by users regardless their level of domain expertise

Task 2.1 Develop a metric that can represent perspectives of domain experts and non-domain experts (i.e., commuters)

Task 2.2 Evaluate the proposed metric with respect to comprehension,

knowledge gain and improvement perceived by different stakeholders.

1.4 RESEARCH CONTRIBUTIONS

The contributions of this work are at the intersection of Computer Science, Mathematics and Smart Cities (mobility transportation) and contributes to the increasing area of Data Science through the use of qualitative and quantitative methods (Waller & Fawcett, 2013).

1.4.1 Computer Science Contributions

From the Computer Science perspective, the major areas that this work contributes to are: Data to Knowledge Processing Algorithms, Data Analytics, semantic domain representation, and Knowledge Representation.

1.4.1.1. Semantic-Based Knowledge Representation

Figure 1.5 represents an application view of data science in terms of Computer Science, Math and a domain of Transportation. Within Computer Science, the area of knowledge representation has been used through the application of knowledge graphs based to model traffic crash data and enable the use of logical rules to obtain inferences about such data. Using the domain of traffic crashes, mathematical metric equations have been computed to develop a metric (CCI). Knowledge representation is humanly understood through data narratives. Furthermore, ontologies and semantic reasoning has provided the integration of data for qualitative and quantitative analysis. Through the use of semantic modeling, knowledge graphs and thus ontologies, qualitative reasoning in the form of a CCI has been intertwined with quantitative data representation in data narratives.

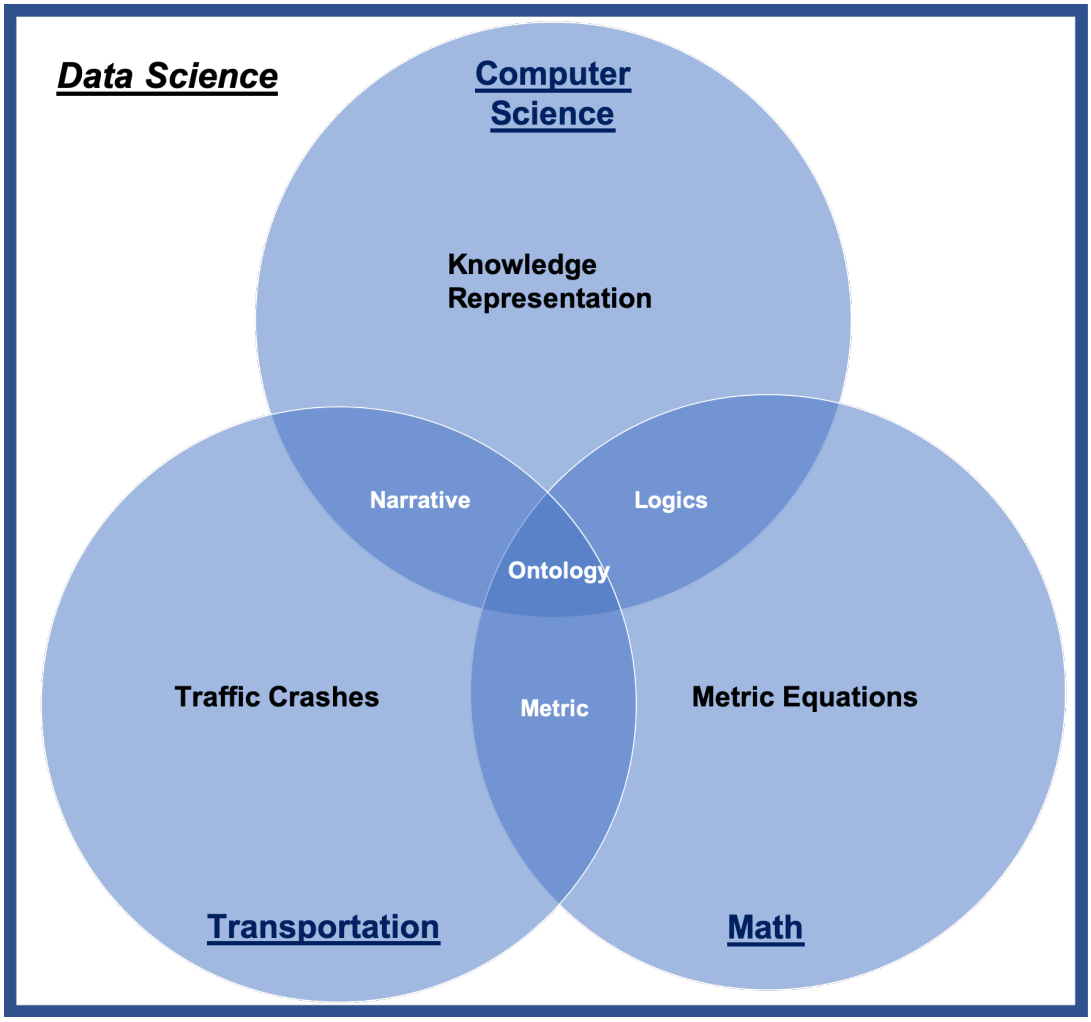


Figure 1.5 VENN diagram illustrating the integration of disciplines in this work in the area of Data Science

1.4.1.2. Data Modeling Methodology

The BUM methodology developed in this work contributes to Data Science by providing means to convert data to knowledge with a bottom-up approach. The BUM methodology is a formalized five-step process that begins with a manual analysis of a given domain. The applied domain used for this work to model a knowledge graph was traffic crashes. The modeling of a knowledge graph has been highlighted by a step-by-step process that takes data-sets and their respective data points and builds them into a cohesive knowledge graph; the knowledge graph is representative

of the data that is available. As part of the BUM methodology, a generic parser was developed for combination of relevant data-sets for the creating of novel metrics in the domain area of Smart Mobility. The parser was created to introduce data transformation through the use of a knowledge graph in the same way as the knowledge graph to ensure high data interoperability. An assessment of information loss was performed to ensure no information was lost when the data was transformed and consumed by the transformation process. Moreover, this work will yield improvements towards traffic crash analysis by first understanding them on a case-by-case basis through data algorithms and knowledge processing.

This work uses techniques to model domain specific data in a way that data does not lose its original meaning, but instead goes through a process to leverage untapped meaning to extract information; in this research it will be the Critical Composite Index (CCI).

1.4.1.3. Knowledge Representation

This knowledge graph was derived into a concrete model that expands the data-sets and enhances the data view (Gil & Garijo, 2017) which takes into greater consideration the inputs and its effect on the end results and maintains the flexibility to change as data-sets are updated. Through this focus, it has been shown that using domain specific data (i.e., historical traffic crash data) improves the knowledge of data views. The understanding of data views is improved by using the raw data as a controlled input and using the output of it by both creating a method to transform it into knowledge and use the data to create a dynamic novel metric, based on adjustable weights.

The BUM methodology is generic insofar that the results gained from it can be transferred, reused, and compared for potential uses in other domain areas. Through the use of the BUM methodology that has been developed, the disciplines involved in this work as part of data science have been improved through data representation, interpretation, and dissemination of information.

1.4.1.4. Contribution to Data Analytics and Metric Creation

From the data analytics perspective, aligned to the application of mathematical principles, the BUM methodology uses publicly available and validated mobility data to ensure the outputs of the methodology are representative of the initial data. In this research, the CCI was an outcome of the developed knowledge graph and the traffic data parser. In the Smart Mobility research area, a majority of the metrics are frequency based whereas the CCI represents more detailed, real-world data that can provide insight for both non-domain and domain experts. As a result of the developed metric, this work shows that the BUM methodology is capable of providing contextualized insights about the data that foster a better understanding of the raw data. Data can be viewed in large scale or refined to a specific crash type. Furthermore, this work describes the importance of using a bottom-up model with respect to a given domain to properly map data inputs and relationships to create usable knowledge graphs for the dissemination of knowledge. This work contributes to data processing as it centers on using real-world data that is transformed with the intent to improve the way traffic crashes are understood and policies that are derived from traffic crashes; thus, improving the quality of life.

Data is the critical factor of describing real-world events and problems. Through the BUM methodology and CCI, the acquisition of untapped knowledge from data is fostered in a way that

can be used in other domains and research areas. The work presented improves the depth of knowledge in the area of Computer Science by contributing to modeling, integration, and retrieval algorithms and metric development for the transformation of data to knowledge.

1.4.2 Smart Mobility (Domain) Contributions

From the domain-specific perspective, this work contributes to Smart Cities, particularly Smart Mobility. A novel mobility metric (CCI) was developed based on historical data that is transferable, reusable, measurable, and sustainable for multiple uses in the research area. The creation of a novel metric for the domain of Smart Mobility provides the ability to understand the severity of traffic crashes on a case-by-case basis. Understanding traffic crashes, evaluating them, and comparing them to one another allows for the understanding of problems on the roadway. The knowledge gained from traffic crash comparison directly links to improving the quality of life because new discoveries can be made yielding new solutions. The improvement of the quality of life stems from understanding what is occurring in the real-world. This research leverages real-world data as a domain focus and provides new insight that has not yet been discovered (i.e. additional possible factors of traffic crashes). Smart Cities use technology and data to drive a the "forward-looking System-of-Systems" that is necessary for the improvement of the quality of life; in this research, data and its respective modeling methodology provides the techniques needed to improve the way the real-world is understood so that improvements can be discussed by policymakers, researchers, non-domain and domain experts.

The proposed selection of the *Four Principles of Smart Cities* and how they align with the *triple bottom* line provide a mechanism to describe individual traffic crashes which begin to describe the affect that it has on the economy, environment, and social concerns. The selected *Four*

Principles of Smart Cities introduce the use of resources, data, solutions and ideas in a standard way that is efficient and productive, sustainable, reusable, and improve the quality of life. This research leverages those ideas in that the data used is publicly available thus providing the efficiency and productivity towards costs and services to compute a CCI; the data can be compared to many data sets within and outside of the State of Texas (i.e. Pennsylvania) for several years and is expected to last onward; the CCI can be reused, refined over time for additional computation and knowledge gain; and the information that the CCI provides improves the understanding of real-world issues and lays the foundation for improving the issues that are faced. Moreover, the CCI is suitable for both non-domain and domain experts since it uses factual data and produces a metric compared to a static scale. The CCI describes the localized issues in a given geographic area by describing the severity of a traffic crash in terms of the event itself, the people involved, and the apparent external circumstances.

1.5 ORGANIZATION

This dissertation is organized into 7 chapters. Chapter 2 will describe the background and related work of Smart Cities, transportations metrics, and modeling of knowledge graphs. Chapter 3 will describe the research methodology. Chapter 4 will describe the results of the research. Chapter 5 will discuss the evaluation of the results. Chapter 6 will discuss the research questions and goal. Chapter 7 will describe the final conclusions and future work.

Chapter 2: Background and Related Work

2.1 SMART CITIES

Research in Smart Cities is an ongoing global endeavor. There are multiple definitions of Smart Cities that both complement and contradict each other. Giffinger (Giffinger, 2007) defines a Smart City to be a forward-looking, self-decisive, and aware city that is powered by the citizens understanding their role in the "System-of-Systems" (Naphade et al., 2011) as shown in Figure 1.2.

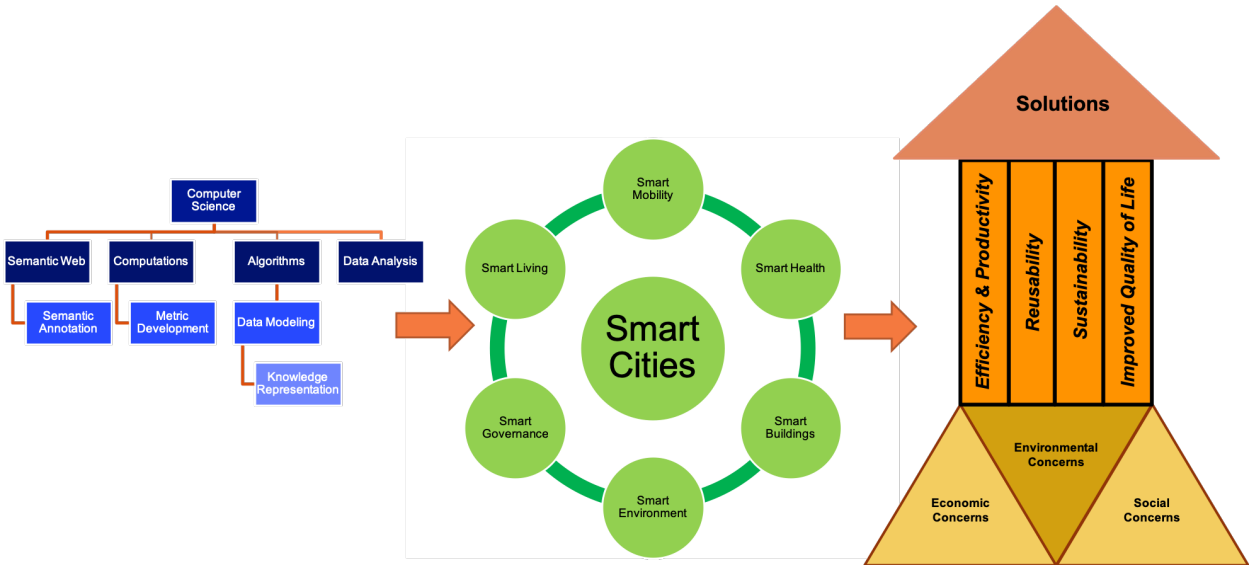


Figure 2.1 Smart Cities "System-of-Systems" (Naphade et al., 2011) and the technology surrounding the city

Caragliu and Nijkamp (Caragliu et al., 2011) describe Smart Cities as a technology-driven infrastructure to improve economic, political, social, and urban development. Each of these definitions can arguably be correct in their description of Smart Cities, where one is driven by people and the other by technology (Dameri, 2013). In either case, both of these definitions allude to improve the quality of life for the citizens that live in the city (Vlacheas et al., 2013).

Although many definitions of Smart Cities are arguably correct, there are not a lot of definitions that include data as a driving factor to Smart Cities. Intuitively, it is necessary to have a grasp on the problems, infrastructure, and potential solutions to improve the smartness of a city; this can only be done by understanding the larger scheme through data.

In a wide search of the literature, there is not a single specific definition of Smart Cities that is accepted worldwide nor the role that data can have on driving improvement to cities. Moreover, an accepted individual metric in a research area is also not accepted. Data is not discussed to be important for tangible applications because it does not immediately provide results or improvements to a city. The implications that data has on Smart Cities research often takes a long time to propagate itself up to a significant level for broad dissemination. Though it is not often popular to discuss raw data, it is critical to have data as a way to explore what is occurring in the real world. Data provides a way to give factual (as best as possible) and trusted information that can be modeled for the improvement of the issue that it represents.

The need to model data is intended to understand relationships and discover new insights of which it represents. The transformation of data into a data-model provides the semantics that is needed for computers and humans alike. Through the implementation of data-models, structured and semi-structured data gain additional information by describing how it can become related to other seemingly unrelated data and become interoperable. Through data interoperability, the way data is interpreted can be enhanced. Through the process of modeling, and enhancing with semantics, data undergoes a critical transformation into usable knowledge that is necessary for understanding any given domain.

Smart Cities is a broad research domain that is commonly separated into several smaller focus areas. The focus areas of Smart Cities include, but are not limited to: Smart Mobility, Smart Buildings, Smart Health, Smart Living, and Smart Governance. Each of these research focuses contributes to the larger understanding of Smart Cities.

To understand Smart Cities as a whole, it is critical to understand each focus area individually. Through the introduction of real-world data, Smart Mobility is explored on the basis understanding traffic crashes at a deeper level. Although it may be appropriate to model a Smart City based on what researchers, policymakers, or developers find useful, it can be difficult to measure its success on an application that is abstract. In this research, a bottom-up approach is used; knowing what information and data is available and using it to make metrics useful to others. Using this approach, this research will model only a specific piece of Smart Cities research: Smart Mobility.

Traffic crashes are events that occur on roadways and affect the way that people move about a city; moreover, any given traffic crash has an immediate effect on the people involved and others surrounding the crash. Each individual crash that occurs has some type of effect on people. When traffic crash data is presented to people it is commonly shown in frequency of crashes and its type. Although frequency of traffic crashes may be appropriate in some domains, it does not provide enough context needed for Smart Mobility enhancements.

Frequency measurement in traffic crashes are surface level statistics that do not provide insight to traffic crashes at an individual level. Since traffic crashes have become part of everyday life, a frequency representation of those crashes does not mean much to people, especially those in

large states with more vehicles on the road, such as Texas. It is necessary to use the data that is available as a mechanism for people to realize how any given traffic crash may affect them; this is done through the development of a metric. Many current metrics are often disguised as traffic crash frequency and is widely accepted based on current trends and the mission that a city has with respect to improving its smartness (Neirotti et al., 2014).

We propose that Smart Cities should be driven by data and use the tangible data as way to understand problems. For example, by looking at a particular incident it is possible to remove the abstraction of what may actually be occurring on a roadway. In Texas, there is data that can be used to look at individual traffic crashes, describe them in a clear way and use that information to disseminate knowledge to both non-domain and domain experts.

The *triple bottom line* is a good candidate to model a Smart City, but it lacks on the specificity that is required for disseminating useful information to its citizens. Most citizens would like an improvement in the economy, environment, and social impacts, but how can improvements be shown? The way to evaluate improvements with any of these ideas is to measure it over time with a standard measurement that will be useful for more than policymakers. Work by Larios et al. (Larios et al., 2016) describes characteristics of building Smart Cities solutions with data as a core asset having scalability, interoperability, modularity, resiliency, and security. Each of these focuses will be intertwined into the exploitation of bottom-up modeling and knowledge gathering. This is critical to measure 'before, during, and after' results of Smart Cities solutions.

Available data will be explored to drive the development of new metrics as a context for the research being done. The data that is used will be transformed and semantically annotated so that

additional information may be derived from it. The focus of the metrics that are developed will be to improve the understanding of traffic crashes with respect to the event itself, those involved, and the external circumstances. This metric will be able to be used as a basis to describe safety, movement of people and the effect that each of these plays in the quality of life of people.

In Smart Cities, technology plays a significant role in its development. Most of the development stems from creating solutions that would be beneficial to the *four principles of Smart Cities*. Based on characteristics by Larios et al., the *four principles of Smart Cities* add value to building resources, ideas, solutions, and metrics to Smart Cities so that they may become more reusable, efficient, sustainable, and improve the quality of life. Although some of the most popular technology solutions and advancements associated with Smart Cities is IoT, mobile applications, and sensors, each of these is only possible because of their reliance on data. Although technology and people can be thought of to be mutually exclusive, both are necessary for Smart Cities that are driven by data. Data provides the information and knowledge that is necessary for human understanding and computer computations. It is through data that Smart Cities are able to become safe, sustainable, productive, effective, predictive, and efficient cities based on people and enhanced by technology.

Smart Cities have many models for development based on the subjectivity of the developer. Models can be based on improving the economy, government, environment, transportation, safety, and other important city issues. Along with all of the models that are developed for a developing a city, comes a large number of metrics to measure any of the models focuses. Most of the focuses of Smart Cities research attempts to address entire areas directly instead of separating each area into smaller groups that can be looked into individually (Anthopoulos,

Janssen, & Weerakkody, 2015). Since work is done on such a large scale, it is difficult to get a clear understanding of any of the focus areas.

2.2 DATA SCIENCE IN SMART CITIES

Data is one of the most crucial aspects in the development of Smart Cities. In 2012, the Obama administration launched an initiative called the "[Big] Data Research and Development Initiative" which aims for increased research and development using data (Li, Cao, & Yao, 2015). Data in many cases can be one of the main focuses in Smart Cities research, not only because of the initiative in the United States but because of the significance it plays for everyday citizens. Data science poses the idea of helping everyday people understand what is occurring in the world beyond what is known.

Smart Cities are intended to be highly connected digital (and technological) cities with massive amounts of data from IoT, sensors, and historical data. However, in some case studies, a large amount of the data remains on hard disks or on reports (Lim, Kim, & Maglio, 2018). Some data is never processed for actual value creation which is needed for consumers of the data. Data needs to be processed systematically and consider the context in which it will be used for proper results (Provost & Fawcett, 2013).

Data is continuously becoming more prevalent in our everyday lives. Since 2013, more than 90% of all the world digitized data has been recorded (Al Nuaimi, Al Neyadi, Mohamed, & Al-Jaroodi, 2015). The increase in data is likely caused by a combination of rapid urbanization as well as more data sources. This data has the potential for cities to gain insight into new information that was not available before (Hashem et al., 2016). Data Science is a leading area of

research not only in Smart Cities but in many areas because it allows us to learn more from the transformation of that data to knowledge. Provost and Fawcett (Provost & Fawcett, 2013) state that "Data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data." Furthermore, Li, Coa, and Yao (Li et al., 2015) write that through analyzing data sets, even with large complexity the conversion of data into valuable information can be done quickly. Smart Cities research makes way for new applications to create new values or metrics for stakeholders involved, including but not limited to citizens and domain experts (Lim et al., 2018).

Smart Mobility provides a research area that has great potential to be expanded by data science. Not only are real-time systems important for improvement to mobility, but historical data is also critical. Analyzing historical factors such as their cause and the effects that it played on people, may provide a way for future work in reducing traffic congestion and providing alternate routes (Hashem et al., 2016). The impact of having useful knowledge is necessary to produce these types of services in the future by expanding the *precision* of identifying the issues apparent in a traffic crash and improving the *recall* of retrieving an ample amount of information to support understanding of traffic crashes (Brewster, Alani, Dasmahapatra, & Wilks, 2004).

Through the transformation of data to knowledge people have the opportunity to easily design new ideas and representations of that data into easily used forms for dissemination (Dobre & Xhafa, 2014). Moreover, data science involves the need to have principles, processes, and techniques to understand that data (Provost & Fawcett, 2013). The implementation of a bottom-up approach being done in this work follows the idea of having a process and technique needed to understand data in a meaningful way, through the transformation of data to knowledge.

Data provides four key components for people to understand (Al Nuaimi et al., 2015; Tapadinhas & Gartner, 2014):

1. *Descriptive*. What happened?
2. *Diagnostic.* Why did it happen?
3. *Predictive.* What will happen?
4. *Prescriptive.* What should I do?

These four components are critical for data science and research relating to Smart Cities. The work being done with this research further takes these ideas and puts them into practice by producing a model for the development of traffic metrics. The metrics are derived from traffic crashes reported by official policing agencies. This research focuses on a subset of the four key components by taking data from these traffic crashes and through the model give a description of traffic crashes, provide a diagnostic of it through competency questions, and gives everyday users the opportunity to make their own decisions about what they should do with the given knowledge.

It is through data that knowledge is acquired and non-domain experts are able to gain insight into what is being presented to them. Data science provides a perspective that is necessary to enable a design representation of data and develop new values or metrics for users (Lim et al., 2018). It is necessary to use data science as a platform to understand Smart Cities because it provides a way to use data as a way to improve the quality of life for its citizens (Hashem et al., 2016) by understanding the *four principles of Smart Cities*.

2.3 DATA MODELS & KNOWLEDGE REPRESENTATION

In Semantic Web research and applications, ontologies are the most popular way to represent data and knowledge amongst data sources. Ontologies provide a foundation to entity representation that is needed in many domains. An ontology is a form of knowledge graph that represents data and its relationship between entities or actions. Ontologies are formalized representations of a domain area that provide semantics to data. Ontologies use the Resource Description Framework (RDF); a formalized schema that defines the rules on knowledge representation. RDF is a framework for representing things on the semantic web (Cyganiak, Wood, & Lanthaler, 2014). RDF describes concepts of a domain in a triple; triples consist of a *subject*, *predicate*, and *an object*. Figure 2.2 shows the representation of concepts in a knowledge graph and how they can be described in terms of an RDF triple; knowledge graphs represent concepts as a relationship between entities. In this work, ontologies are seen as a specific type of knowledge graphs that enable the description of a domain with semantics. Knowledge graphs provide context to data that is necessary for describing data without the formalism that is inherent with ontologies. This research will use the term *Knowledge graphs* as a way to describe relationships between entities and data, as a way to broaden the use of data-modeling.

Knowledge graphs are large collections of interconnected entities (Arenas, Cuenca Grau, Kharlamov, Sar Unas Marciuška, & Zheleznyakov, 2015) that model a domain by defining the relationships between elements of the domain. They play a significant role in understanding how real-world concepts can be linked to one another and any insufficiencies that there may be in representing the information. They can be considered a domain representation since they describe relationships at a high level, thus improve the understanding of relationships between various concepts that describe the domain (Mejia, 2017). Knowledge graphs are critical to

understand relationships between data entities in a domain. For example, Smart Mobility can be modeled in terms of traffic crashes. Describing traffic crashes in terms of the characteristics and data available in the crash can provide insight to understanding the traffic crash because it integrates data together. They have played an important role in data integration research. Semantic Web research has focused on adding semantics (or contextual meaning) to data.

Figure 2.2 Representation of relationships in a triple, the building blocks in knowledge graphs using RDF

Figure 2.2 shows how relationships in a knowledge graph are represented as triples, with a domain term (*subject*) with some relationship (*predicate)* between another term (*object*). The purpose of having this type of relationship is that in general, all concepts of a domain can be represented and easily show how they relate to other concepts. By relating concepts to one another a deeper understanding of data is intended to be exposed. Figure 2.3 is an example of a real representation of a subject, predicate and object for traffic crashes. This example highlights the importance of understanding how entities can be related to one another in real-world applications.

Figure 2.3 Representation of a relationship between *CRASH* and Crash_ID

Through the representation of data in a contextual way, describing data becomes available through knowledge graphs. Data heterogeneity can typically be described by four individual categories: structure, syntax, system, and semantics. Many data-sets are often structured in ways that are unlike other data-sets; data syntax is different from data-set to data-set; the system on which the data is represented on can be different from one system to another; and data can be semantically different from one another. Two data sets may say the same thing, but mean something different based on its context (Buccella, Cechich, & Rodriguez Brisaboa, 2003). Having data heterogeneity is not an optimal way of understanding a domain. Developing a model that provides data interoperability is necessary to fully understand any particular domain. Data interoperability within a domain that has disjoint data sets is often difficult to achieve without a proper data model. Cruz and Xiao (Cruz & Xiao, 2005) describe the elimination of data heterogeneity as *semantic data integration*. The representation of data as concepts within a knowledge graph is the essential process of *semantic data integration* and provides the creation of Linked Open Data. For example, traffic crash data is divided into three large data-sets that describe the crash event, the drivers, and any other people that were involved in the crash, respectively. Alone these data sets can be used for statistical purposes, but they do not provide contextual meaning to non-domain and domain experts. Through the process of *semantic data integration,* the data sets are linked together, and additional knowledge can be gained from it.

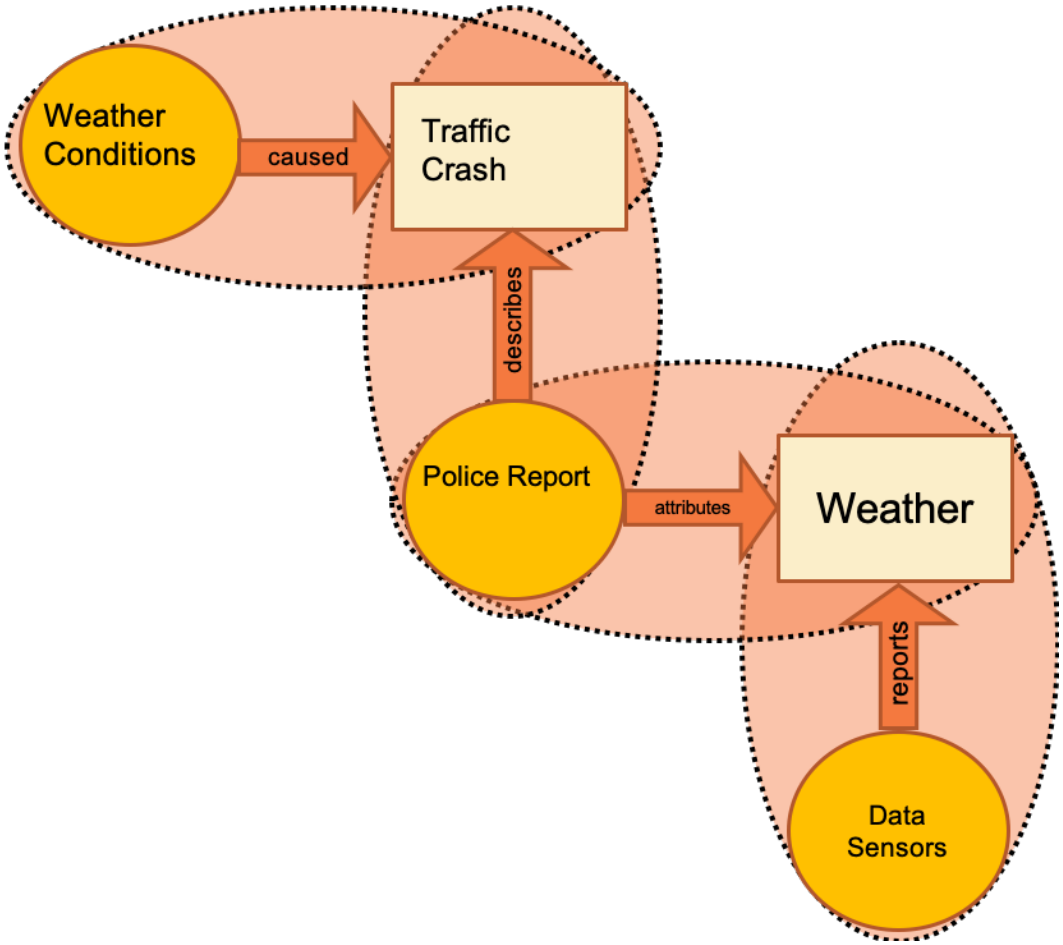


Figure 2.4 Linked data used to gather new information in traffic crash reporting

Linked Open Data is the idea that data can be interconnected with each other in a "Web of data" (Jain, Hitzler, Yeh, Verma, & Sheth, 2010). The Semantic Web and knowledge graphs are built with the purpose of having data that can be shared across domains. Similarly, Linked Open 'Usable' Data is the idea that the data shared across domains are useful; however, most Linked Open Data is not easily linked to each other (Jain et al., 2010). This research is attempting to exploit a relationship between data modeling through knowledge graphs and the idea of Linked Open 'Usable' Data by linking heterogenous data together as shown in Figure 2.4.

Knowledge graph modeling can have different levels of abstraction and clarity based on the way it describes any given domain. Knowledge graphs are typically the medium that is necessary for sharing data across the web since they represent knowledge and logical relationships for a specific domain (Spyns, Meersman, & Jarrar, 2000). Furthermore, since knowledge graphs are not always fully representative of an entire domain because of domain expert differences, the closer a knowledge graph gets to becoming agreed upon, the more shareable and reusable it will be in the domain area (Spyns et al., 2000). To integrate data and have it linked to other knowledge graphs is difficult because there may be possible differences in the development requirements, or functionality of the knowledge graph (Ziegler & Dittrich, 2007). Given large domains, this research will provide a model that will begin to produce improved interoperability for published data.

To illustrate the use of knowledge graphs in this work, Figure 2.5 that was obtained as part of this work provides a high-level representation of Smart Cities and Smart Mobility. Figure 2.5 has nodes in blue (Smart Cities and weather) that represent domain areas; the green node (Smart Mobility) represent the focus areas; the orange nodes (Metrics, Issues) represent the sub areas of focus in each Smart focus areas; the yellow node (Traffic Crashes) represents a specific concern of Smart Mobility; the grey nodes represent specific data attributes about traffic crashes and weather. The solid lines show the direct relationships between concepts whereas the dotted lines show potential relationships that can be obtained by using logic rules.

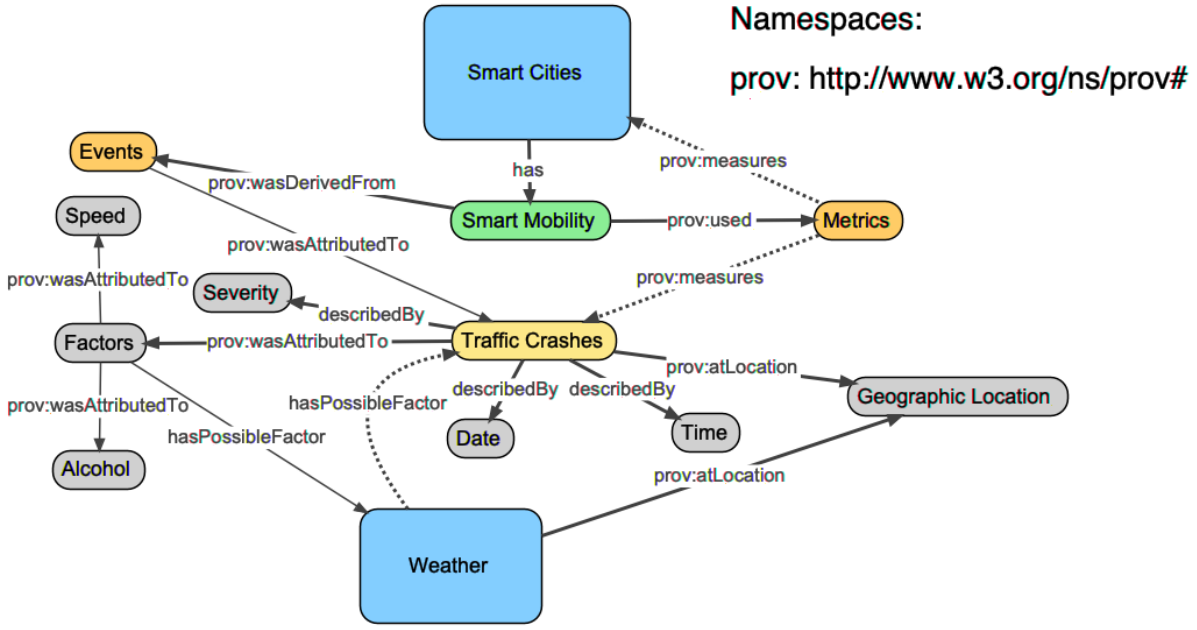


Figure 2.5 Semantically annotated knowledge graph model of Smart Cities with a focus on Smart Mobility

There is no single way of formally representing any given domain. Domains are often guided by domain experts, and since many have different points of view, there are likely to be different representations of the same domain (Noy & McGuinness, 2001). Typically, a knowledge graph is representative of a single domain; from that single domain, many concepts are described and shown to be in a relationship with other concepts. The representation of a domain in a knowledge graph is intended to describe how data and ideas are linked together. Knowledge graphs do not change data that is collected; it allows for a common understanding of information and promotes the reuse of information at a larger scale (Noy & McGuinness, 2001). The knowledge graph presented in this work can be linked to other data through the use of the PROV Ontology (Lebo, Sahoo, & McGuinness, 2013) and the road accidents ontology (Dardailler, 2012). The provenance ontology contains elements that map to person in the knowledge graph developed in this work; moreover, the road accidents ontology uses many of the same vocabulary terms such

as event, location. These open ontologies can introduce provenance of data sources in the knowledge graph developed in this work as well as show common vocabularies. For example, the knowledge graph in this work has a main element of Crash that may use the provenance ontology to link to the road accidents ontology; Crash prov:wasDerivedFrom CRIS (TxDOT, 2018); Crash prov:wasInfluencedBy rao (Road Accidents Ontology) (Dardailler, 2012). In addition, events shown in Figure 2.5 can be mapped to prov:Events. By introducing additional ontologies, it is fosters additional interoperability with other data sources, knowledge graphs, and standardized vocabularies.

2.4 SMART CITY METRICS

Understanding and measuring the data that is found for comparison and growth is critical to the improvement of technology and society. According to Meadows, "Indicators (Metrics) are natural, everywhere, part of everyone's life… Indicators (Metrics) arise from values (people measure what is cared about), and they create values (people care about what they measure)" (Meadows, 1998).

Metrics can be found in many places and are necessary to measure the high quality of life for people. In the case of Smart Mobility, metrics can be developed in a similar way; for the purpose of giving everyone, domain and non-domain experts alike the ability to understand what the metric means and how it can be measurable, comparable, and useful to them. Cheu and Balal (Cheu & Balal, 2018) write that some of the common metrics found that relate to traffic describe specific events, including, but not limited to:

- number of crashes per year (crash rate)
- number of injuries per year (injury rate)

- number of fatalities per year (fatality rate)
- number of crashes per 100,000-population per year
- number of injuries per 100,000-population per year
- number of fatalities per 100,000-population per year
- number of crashes per million-vehicles per year
- number of injuries per million-vehicles per year
- number of fatalities per million-vehicles per year
- number of crashes per million-VMT per year
- number of injuries per million-VMT per year
- number of fatalities per million-VMT per year

Of these metrics, none describe traffic crashes in a clear way and is not intuitively meaningful to city residents nor provide information that is useful in making decisions. These metrics provide frequency statistics, but do not provide contextual meaning to traffic crashes. These metrics do not immediately relate to the non-domain expert because they do not provide enough context for them. Furthermore, these metrics are not comparable amongst different places without qualifying the results. Comparing the total number of crashes per year in Texas does not likely provide a way to understand how it differentiates with the total number of crashes per year in Rhode Island. Moreover, current metrics do not describe additional information within the metric itself such as what caused the crashes or if there is a possibility that the geographic location played a factor with it.

Metrics are crucial to understanding the *four principles of Smart Cities* proposed in this work. Many researchers determining the efficiency, productivity, reusability, and sustainability of a

city come up with different results since there is not one way to measure these metrics nor the smartness of a city (Lazaroiu & Roscia, 2012). A majority of the research in determining metrics for a city takes into account all of the different domains; from this, a composite type of metric is attempted to be developed (Lazaroiu & Roscia, 2012; Neirotti et al., 2014; Steele, 2014).

Work by Lazaroiu and Roscia (Lazaroiu & Roscia, 2012) aims to develop a sustainability metric of Smart Cities by addressing the economy, mobility, environment, people, living conditions, and governance. This work is done through the development of a model based on fuzzy logic. Each of these areas has sub-areas that are associated with it, from which are given a set of different possible metrics. These types of metrics may appear appropriate to determine some type of metric for a city, but since the indicators that are given are based on the opinion of a set of individuals and not based on objective data, it is not necessarily an accurate representation of the city.

One of the most accepted metrics in the world is the ISO 37120:2014 sustainable development of communities – metrics for city services and quality of life (Steele, 2014). The ISO has developed a set of metrics to assess a cities performance and measure progress with the intent to improve the quality of life. The items being looked at by the ISO 37120:2014 standards include, but are not limited to the economy, education, finance, health, safety, and transportation. These standards are intended to be useful for research that is being done on a large scale to compare cities globally. These standards have been represented by using knowledge graphs, with the intent to formally describe the relationships between the standards and the data that it may contain (Khazei & Fox, 2017).

The ISO 37120:2014 is the first of its kind for city standards; particularly on a global scale. The standards presented by the ISO are significant advances in the way that Smart Cities are measured, but the issue lies in its overwhelming generalization. The metrics that are described are not exhaustive enough to be confident in its measurements. The standards are broad enough to be used as a measurement for an entire city, but it is difficult to show how these standards are computed. This standard would greatly be improved by developing a clear methodology for computation of the multiple sub-domains that are part of it.

2.4.1 International Metrics

The National Statistical Institute of Italy (National Statistical Institute of Italy, 2001) reported the initiatives and metric types that they were going to be promoting in the country, with the main focus of determining a way to improve the environment. The type of metrics that they will be addressing are in air quality, energy consumption, green areas, noise, transportation, waste, and water. Specifically, transportation metrics focus on the infrastructure and management processes with respect to the environment. Although the metrics presented a focus on environmental sustainability, it can also help understand the way people move around a city and the role that humans play in transportation. Unlike other metrics, this grouping details the ways that metrics will be measured. Metric measurement is the central way to understand progress and makes it useful for comparisons (Hiremath, Balachandra, Kumar, Bansode, & Murali, 2013; National Statistical Institute of Italy, 2001).

Smart Cities metrics is a global research area, including countries from Europe and North America. A report done by Hernandez-Moreno and De Hoyos-Martinez (Hernandez-Moreno & De Hoyos-Martinez, 2010) shows that the executive branch of government in Mexico has

developed a plan for sustainability of resources and goods, similar to that of the *triple bottom line*. The metrics that they are concerned about are two-fold with respect to transportation; the first is the infrastructure and the second is people. The maintenance of sustainable infrastructure is one of the major metric groups that is being measured. Additionally, measuring the way that people are intertwined with the infrastructure is a key metric for understanding movement (Hernandez-Moreno & De Hoyos-Martinez, 2010). Connecting both the infrastructure and the users of the infrastructure shows the need to measure the way that humans affect transportation services. Not only does sustainability play a vital role in the metrics, but safety is also a key factor for decision making (Hernandez-Moreno & De Hoyos-Martinez, 2010). Much of the European Union has adopted the PAS 180:2014 standards that were developed by the British Standards Institute (BSI). The standards laid out for Smart Cities, attempt to define terms that are used including interoperability and metrics; they are described as systems efficiently working together and having a defined measurement method and scale (Institute, 2014). According to the British Standards Institute, data is able to drive city-wide change. Moreover, data provides a marketplace for new information and services to create new value within the city (Institute, 2014).

In the United States and Canada, there is work being done to improve the sustainability of cities by focusing on key metrics in specific focus areas. The focus areas include, but are not limited to the economy, transportation, safety, and the environment (Mihyeon Jeon & Amekudzi, 2005). The metrics listed in the report by Mihyeon Jeon and Amekudzi (Mihyeon Jeon & Amekudzi, 2005) show that the metrics for each of the focus areas have metrics that can be quantified and compared. For comparison purposes, composite indices are often used to see progress (Yigitcanlar & Dur, 2010). Having quantifiable metrics become the corner-stone for

understanding the progression, development, and "smartness" of a city. Particularly, a majority of individual states throughout the United States have a mission to promote safety, effectiveness, and efficiency for their roadways and transportation systems. As a result of what is being promoted throughout the United States, this work will focus on using data as a foundation for metric development. The development of these metrics will take into account the way that data is represented and transforms it into usable knowledge.

Throughout the world at least 1.2 million people are killed each year due to road crashes (World Health Organization, 2004) and up to 50 million additional people suffer injuries (Hermans, Brijs, et al., 2009); of all the deaths, more than half of them are between 15-44 years old (World Health Organization, 2004). The United States Department of Transportation has issued a five-year Research, Development and Technology Strategic Plan that describes their research and development priorities.

There are many metrics throughout the world that have been developed. The literature shows that a majority of the metrics have come from Europe and Canada. The metrics range from measuring the number of vehicles on a roadway to the number of crashes per a given number of vehicles (Ma, Shao, Ma, & Ye, 2011). The selection of metrics to describe safety and mobility is different from researcher to researcher. Relatively, a safety metric is based on the number of crashes and fatalities on the roadway (Hermans, Van den Bossche, & Wets, 2009).

2.4.2 Metric Significance

For USDOT, the four main focuses are to promote safety, improve mobility, improve infrastructure, and preserve the environment (United States Department of Transportation

(USDOT), 2016). In the United States, the problems with rapid urbanization are developing quickly. It is predicted that over the next 30 years the population will increase by 70 million and the economy will double (United States Department of Transportation (USDOT), 2016). With increases in the population, problems contain to increase such as traffic, pollution and economic issues (Al Nuaimi et al., 2015). With rapid growth and the problems that come with it, it is critical to examine the way that safety and mobility can be improved and measured. Figure 2.6 shows the need to have significant metrics as a way to show how Smart Cities are improving from the solutions that are developed.

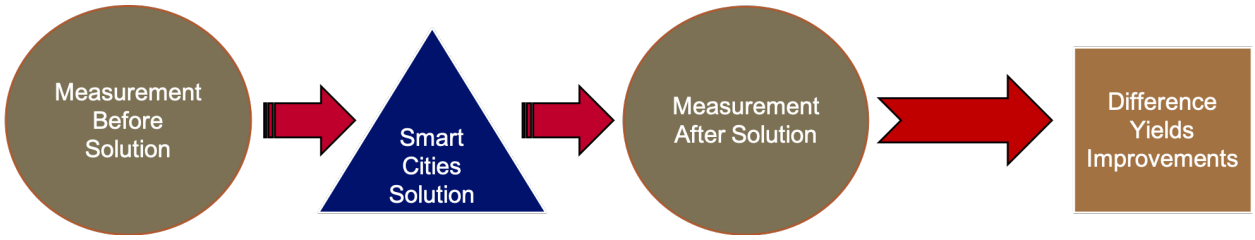


Figure 2.6 The significance of metric measurement to determine improvements of Smart Cities solutions developed

Metrics are necessary and crucial to help the way the world is understood. Whether a metric is for blood pressure, weather temperature, vehicle speed, blood sugar, or traffic crash measurement, they give people a way to possess a concrete understanding of abstract information. Metrics can provide a way for policy-makers to understand what changes, if any, need to be made with concrete evidence to support that claim. According to the work by Hoornweg et al. (Hoornweg, Nunez, Palugyai, Villaveces, & Longfellow, 2007), there are 12 major characteristics that a metric has: must have a clear objective, be relevant to the objectives, be measurable and replicable, statistically representative of the city, comparable and standardized, potential to predict, effective, economical, interrelated to society, consistent and

sustainable. Since there are many characteristics to a well-rounded metric, they are often challenging to develop; thus, this research will focus primarily on metrics that will focus on understanding the safety and mobility of people. Moreover, metrics should also promote the buildup of a Smart City insofar as it being scalable, interoperable, modular, resilient, and secure (Larios et al., 2016).

The Federal Highway Administration (FHWA) is attempting to address national highway challenges over the next five years. The goals that the FHWA are looking toward improving are (United States Department of Transportation (USDOT), 2016):

- Highway safety
- Improving the mobility of people and goods
- Maintaining infrastructure integrity
- Enhancing system performance
- Promoting environmental sustainability
- Preparing for the future

Similarly, the National Highway Traffic Safety Administration (NHTSA) has a mission to save lives, prevent injuries and reduce economic costs due to road traffic crashes, through education, research, safety standards, and enforcement (United States Department of Transportation (USDOT), 2016). Other departments within the USDOT including, but not limited to the Federal Aviation Administration (FAA), Federal Transit Administration (FTA), and the Intelligent Transportation Systems – Joint Program Office (ITS) all have a focus on improving safety and mobility (United States Department of Transportation (USDOT), 2016). The different goals of improvement are interconnected with each other in that improved safety on roadways will likely-

yield improved mobility, performance, and sustainability. Through the development of metrics, these goals can be measured over time to see where additional improvements may be needed.

For metrics to be considered valuable, it must be measurable (Fox, 2015). Measurable metrics derive from removing subjectivity and introducing a standard that can show progression or digression of the domain over time. Metrics are derived from facts and reveal new information (Hermans, Van den Bossche, et al., 2009) by nature. This research attempts to transform qualitative information into a quantitative measurement. In general, information that is useful to researchers such as road safety, efficiency, productivity and the effect that these areas have on the quality of life are qualitative measurements. It is difficult to describe road safety, efficiency, or productivity from an outside perspective. Without determining a way to provide a metric of road safety, efficiency, or productivity they are only abstract thought problems.

The transformation of qualitative data into useful quantitative data is necessary to standardize a way to understand what is truly going occurring on the roadways. *The key to transforming individual data sets into useful knowledge is developing a significant metric.* Metrics without significance are just another arbitrary number of some event or set of situations. Safety metrics provide meaningful monitoring of road-safety and its developments over time (Gitelman, Doveh, & Hakkert, 2010). Most of the time index values are developed so that it is easy to understand for human interpretation, as well as provide a comparable value over time or against other places (Hermans, Van den Bossche, et al., 2009). By producing metrics that can be clearly interpreted by diverse groups of stakeholders, the results may improve decision-making process on future actions for increased safety; thus help understand how the improvements are actually increasing safety over time (Hiremath et al., 2013).

This work is intended to develop metrics as a way to enable broad domain measurements such as safety, effectiveness, and efficiency to understand it at a definite level (Hiremath et al., 2013). By focusing on a specific domain, such as transportation, the metrics developed will move from an abstract understanding of transportation to concrete knowledge about movement in-and-around a city, thus contributing to understanding Smart Cities as a whole.

Chapter 3: Research Methodology

3.1 APPROACH

The research is done using a bottom-up approach that focuses on data that has been collected from crashes provided by an official source (e.g., a government agency). The approach used in this research mirrors an exploratory data analysis approach in which insight is maximized by the data-set, important information is extracted, and data is analyzed then modeled (NIST/SEMATECH, 2013). The data provides a historical reference as well as a provenance trace that enable users to trust the data. By establishing a way to link data from the primary data model, the effectiveness of this method is compared to a top-down approach (where a model is made prior to data collection). The top-down approach is beginning with a large abstract problem or idea and then a solution is developed for that problem.

Many of the methods used in a top-down approach are trial and error until a refined solution is established (Cosgrave, Arbuthnot, & Tryfonas, 2013). A bottom-up approach focuses on "what is known" and how it can be improved or understood deeper. Both methods of addressing Smart Cities research are valid options, however since this research focuses on data, a bottom-up approach is best suited.

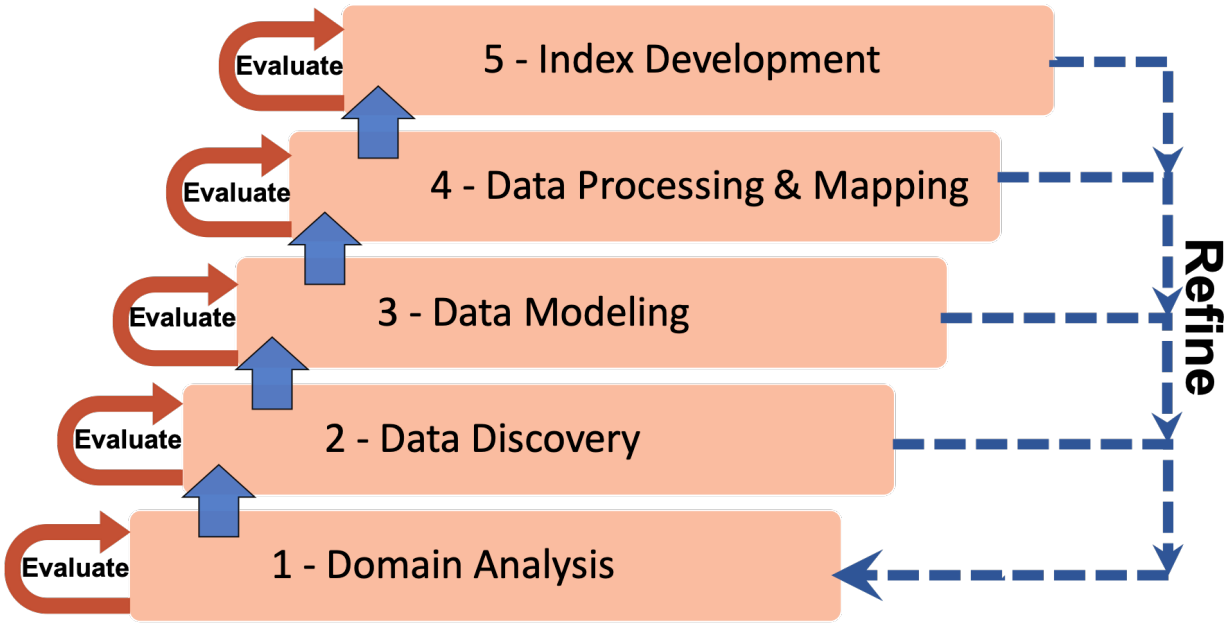


Figure 3.1 Graphical representation of the BUM methodology being used

The approach shown in Figure 3.1 is iterative, where data, relationships, computations, and analysis can be made on what data is available. The approach is iterative insofar that it is evaluated at every stage prior to continuing onto the next. The BUM methodology will allow for the answering of research and stakeholders questions step by step and determine the quality of the approach at each stage of development for enhanced insight.

The approach follows the idea that data drives the creation of knowledge which promotes the transformation of cities into Smart Cities. The BUM methodology focuses on using publicly available mobility data to provide a model that uses linked data.

The remainder of this chapter describes each one of the methodology steps using the domain of traffic crashes in the State of Texas.

3.2 DOMAIN ANALYSIS

The domain analysis was conducted through manual learning of the domain of traffic crashes. In traffic engineering, safety is the top priority for engineers creating roadways (Roess, Prassas, & McShane, 2011). Safety and crash analysis are integral pieces of understanding traffic. There are many types of traffic crashes that occur. Traffic crashes are classified in many ways. The Texas Department of Transportation (TxDOT) uses the American National Standard Institute (National Safety Council & ANSI, 2017) as a guide for traffic crashes. Analysis on crashes are usually done based on one point of view; traffic crashes can occur in a wide variety of ways including, but not limited to, by person injury severity, damage severity, first harmful event, location, and the number of vehicles involved in a crash (National Safety Council & ANSI, 2017). Traffic crashes can occur in many situations and understanding traffic crashes is one of the main concerns of traffic engineers because it is their priority; as part of traffic engineering is the inclusion of semiautomatic and fully automatic studies which use data collection as a key foundation of understanding traffic (Roess et al., 2011). This work will leverage the idea of integrating data, thus integrating various points of view of traffic crash analysis. This analysis is done to incorporate a wide variety of data that explores linked open data and other domains to improve how traffic crashes can be viewed on the basis of data.

3.3 DATA DISCOVERY

There are many different sources of data that can be used to understand traffic crashes, but it is critical to retrieve information from trusted sources. The main source of data came from the Texas Department of Transportation Crash Records Information System (CRIS) (TxDOT, 2018). The data source was determined based on the work of Torres (Torres, 2016). CRIS is main data

source used in this research to maintain a single structure of the implementation process; additional data can be incorporated for additional data interoperability.

The data used in this work is reported by law enforcement agencies to TxDOT for every traffic crash reported in the State of Texas. The data was originally collected by the law enforcement agencies after a traffic crash occurred. The process of collecting traffic crash data is performed by individual officers or through Special Traffic Investigators. The location of the crash is determined based on approximation of the received emergency call coordinates. The collection of traffic crash data in the States of Texas is guided by the definitions used by the American National Standard Institute (National Safety Council & ANSI, 2017) and published guides by TxDOT (Texas Department of Transportation, 2017). The guidelines used provide law enforcement agencies a standard format to report traffic crashes. The police report has a list of all of the information that needs to be acquired as part of the traffic crash investigation. After the traffic crash investigation, each officer reports it into a computer system (e.g., Lonestar (Software, 2013)) which then is uploaded into CRIS (Shields, 2018; TxDOT, 2018). The BUM methodology presented in this research provides the systematic approach needed to transform original law enforcement agency data into more understandable information for dissemination of knowledge.

Through the evaluation process of this data, TxDOT discloses that every crash is reported cannot be deemed one hundred percent accurate (TxDOT, 2018); however, with respect to this research, this information is considered to be trusted. The data was acquired through the TxDOT using the Crash Records Information System (CRIS) (TxDOT, 2018). The process of data discovery can

be transferred to different domains for reuse of the BUM methodology; furthermore, additional data can be added for additional data analysis.

For this research, data has been collected primarily for the State of Texas. The State of Texas is being selected as a primary data location because of its large size, data volume, possibility to compare rural and urban areas, and standard reporting practices. Furthermore, the data being used is reported by authority agencies, such as local police departments and state authorities by using the Crash Records Information System (TxDOT, 2018). In Figures 3.2 – 3.14, the data shown are the data points for the data-sets available; it is not a specific traffic crash data value.

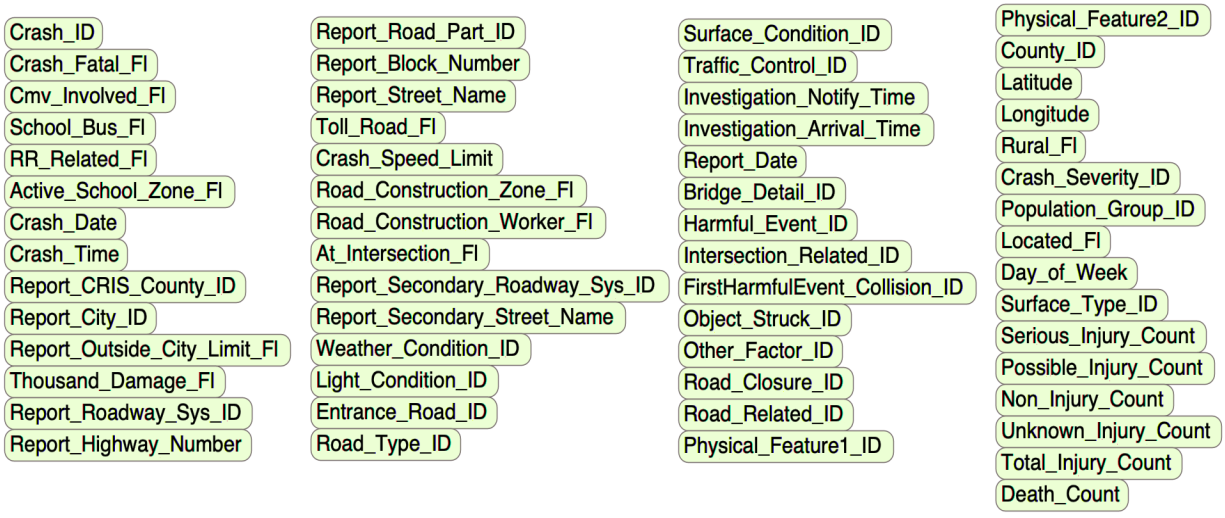


Figure 3.2 Data points from crash-related data (green nodes)

Figure 3.2 shows the initial data points that are contributed to the data set that describes crashes reported by official government agencies (e.g. police, DoT) (TxDOT, 2018). Each of the data points in this data set is related to each other insofar as they come from the same data set and each individual traffic crash will have a value in each of the data points. The data points that are

represented in Figure 3.2 describe the crash data set. The crash data focuses on specific factors related to the traffic crash. Each traffic crash will have values for each data point that will extensively describe what occurred as part of the major investigation by official government agencies.

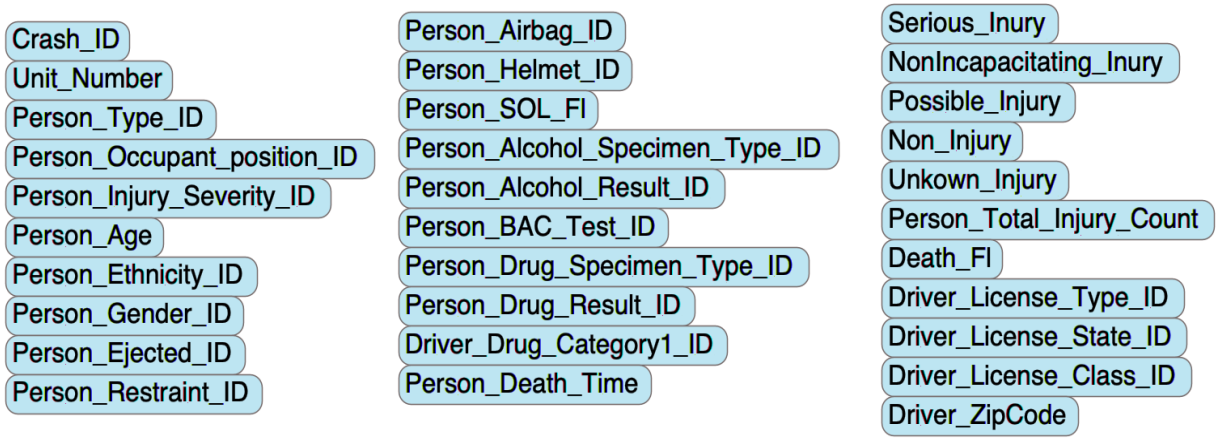


Figure 3.3 Data points from primary persons data set (blue nodes)

Figure 3.3 shows the initial data points that are related to the primary persons involved in the investigated traffic crash. The primary person is the driver of the vehicles as well as any pedestrian that may have been involved in the crash. For crashes containing more than one vehicle (or person not in a single vehicle crash), there will be more than one data value to describe the persons involved. The persons will be connected together by the Crash_ID and differentiated by the Unit_Number describing each person individually.

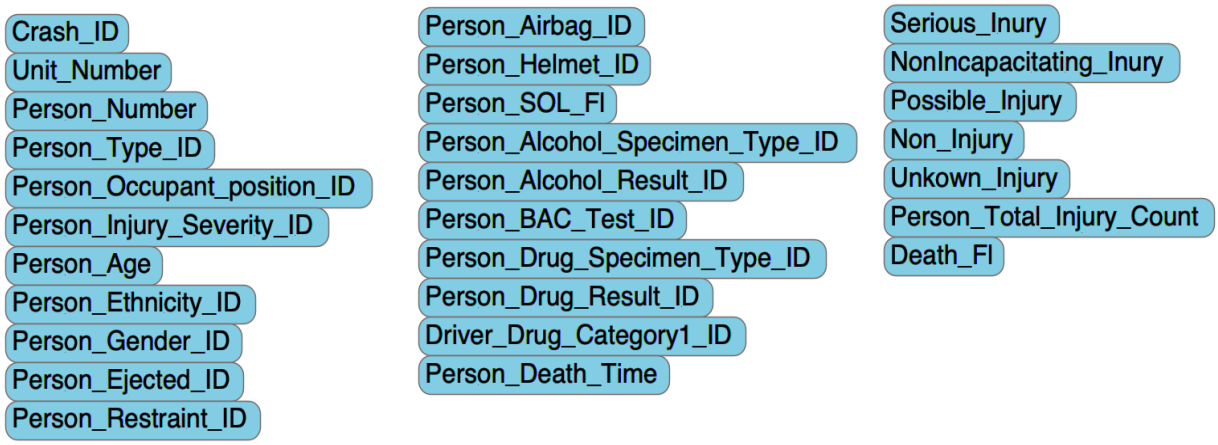


Figure 3.4 Data points from secondary person data set (blue nodes)

Figure 3.4 shows the initial data points for secondary persons involved in a traffic crash. Secondary persons include those who are any type of passenger in the vehicle. Secondary persons involved in the traffic crash are linked to the crash by Crash_ID and separated by vehicle using its Unit_Number. Each unit will have a Person_Number to differentiate each individual in the vehicle that was part of the traffic crash. The Person_Number describes the passengers in terms of their seated position in the vehicle.

3.4 DATA (KNOWLEDGE GRAPH) MODELING

Knowledge graphs are the basis to which data can be processed and mapped in this work. In Semantic Web research, ontologies, of which is a subset of a knowledge graph, is the pivotal tool that describes domains. The use of a knowledge graph in this research can be directly compared to an ontology. The use of a knowledge graph provides the semantics needed between relationships without the formalization that is inherent in ontologies. Noy and McGuinness (Noy & McGuinness, 2001), describe the development of ontologies – thus knowledge graphs, using multiple approaches; one of the suggested approaches is to use a bottom-up approach. The

bottom-up approach follows the idea that was described through the research methodology. In interdisciplinary research, knowledge graphs provide the semantics and content needed to describe a data-set without needing ontology experts.

The relevant literature reveals that the use of knowledge graphs is widely accepted as a mechanism for describing domains and can be developed using a bottom-up approach. Moreover, the BUM methodology shows that the knowledge graph can be developed iteratively which mimics that of ontology development (Noy & McGuinness, 2001). Modeling a knowledge graph based on data is a systematic practice that provides flexibility that ontologies do not have.

3.4.1 Data Source and Primary Data

As a result of having data guiding the bottom-up modeling of traffic crashes as well as the development of a metric, it provides a foundation to enable access to knowledge bases by rendering the knowledge in different modalities (i.e. natural language text and raw data).

The developed model is a knowledge graph, which serves as a high-level data model. This model shows the relationships between modeled entities and attributes in terms of data as well as determine any missing data that may need to be represented. The knowledge graph was developed taking an iterative approach. An iterative approach was based on the idea of a software iterative design pattern in which the knowledge graph was built focusing on individual components, then added to those components as a modification until the entire graph was complete (Bass & John, 2003). Using an iterative approach provides a way for the graph to expand and modify as additional data is included in it. The knowledge graph comes directly from the data that was discovered. The data provided a basis to begin classifying commonalities

amongst itself as well as show the relationships that it had with other data points. By understanding the knowledge graph at each phase of development, knowledge can be gained to better describe the data.

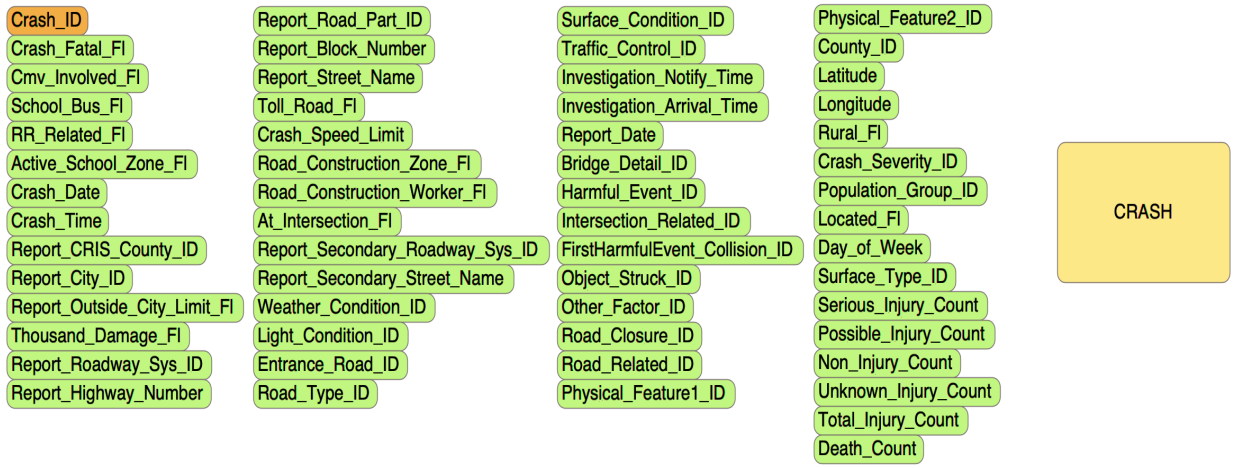


Figure 3.5 Data points and their relation to a larger entity – *CRASH* (yellow node)

Figure 3.5 shows the initial stages of modeling the data beyond their initial data points. The orange node represents a data point that is common amongst multiple data sets; green nodes are specific to crash-related data. The modeling of the data begins by describing which data points are common amongst any or all of the data sets; by modeling the data in this way, the data points can be generalized as belonging to *CRASH*. *CRASH* is a larger class that contains all of the data points from the crash data set. As shown in the figure, Crash_ID is identified as a different color node to describe that it is the linking data point to other data sets.

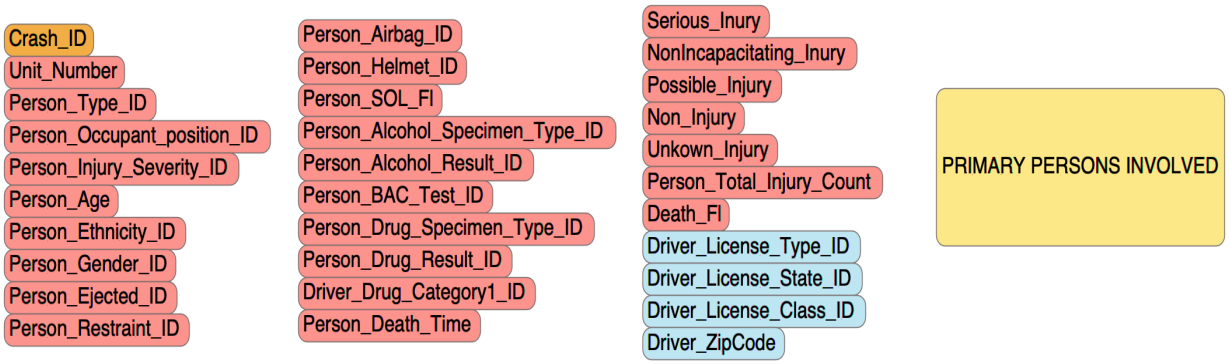


Figure 3.6 Data points and their relationship to a larger entity – *PRIMARY PERSONS INVOLVED* (yellow node)

Figure 3.6 shows the initial stages of modeling the primary persons data set – the primary persons involved in the traffic crash. The orange node represents a data point that is common amongst multiple data sets; pink nodes represent data that is common amongst primary and secondary persons involved; light blue nodes represent data unique to the primary persons data set *PRIMARY PERSONS INVOLVED* is a larger class that contains all of the data points related to the primary persons data set. Each crash has a primary person involved per unit. Each unit is either a driver of a vehicle involved, bicyclist, or pedestrian. The nodes in pink are data points that are the same as the secondary persons involved data set. Light blue represents data that is unique to the primary persons data set as they reflect the driver of the vehicle. The primary person data-set also contains Crash_ID which links the primary person to the traffic crash.

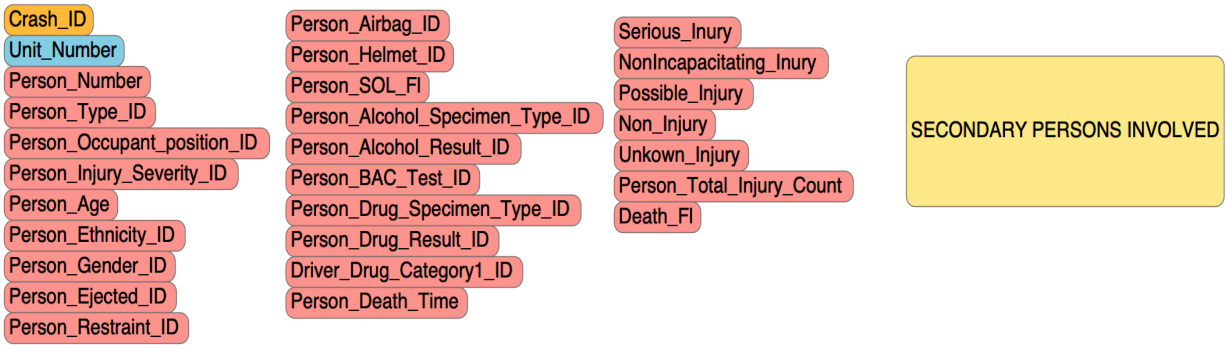


Figure 3.7 Data points and their relationship to a larger entity – *SECONDARY PERSON INVOLVED* (yellow node)

Figure 3.7 shows the initial stages of modeling the secondary person data set. The data set is based on describing each individual person involved in the traffic crash who is not the driver of the vehicle but involved in the crash. The orange node represents a data point that is common amongst multiple data sets; the blue node represent data from the secondary person data set; pink nodes represent data that is common amongst primary and secondary persons involved. *SECONDARY PERSON INVOLVED* is a larger class that is composed of all of the data points inside of the secondary person data set. The secondary persons are described by specifying which unit (vehicle) they are in and the position in which they were seated. The blue nodes represent data from the secondary person data-set that differentiates between a primary person and a secondary person involved in the traffic crash. The secondary person data-set also contains Crash_ID which links the secondary person to the traffic crash.

3.4.2 Generalization of Data Sets – Linked Data

Data was separated into generic entities to promote the integration of the various data sets. The data separation was done into three major concepts: Location, Investigation, and Person. Location contained all of the data points that are representative of a geographic location or

alludes to location; all other data points in the CRASH data-set were automatically part of Investigation. This design was decided to ensure a separation of location-based semantics and investigation semantics. Both the primary person and secondary person data-set were merged into a single concept, Person. Person semantically describes a person whether they were the driver of a vehicle or not. If a person was not a driver, information regarding the driver are left empty, as specific data about drivers were not collected about passengers.

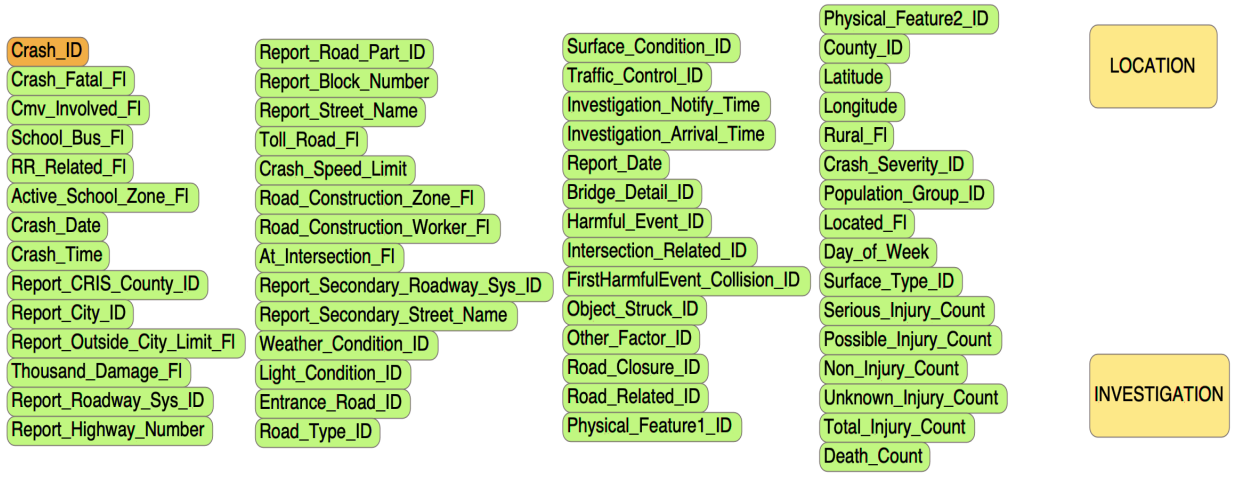


Figure 3.8 Data points from the crash data generalized into two larger entities – *LOCATION*, and *INVESTIGATION* (yellow nodes)

Figure 3.8 shows the logical progression of having all of the data points in the crash data-set and generalizing it to either a location or investigation. The *LOCATION* data-set will contain the data points that describe the physical location of a given traffic crash, including coordinates, streets, cities, intersections and school zones. The *INVESTIGATION* data-set will contain information regarding the investigation results of the local government agency on the traffic crash.

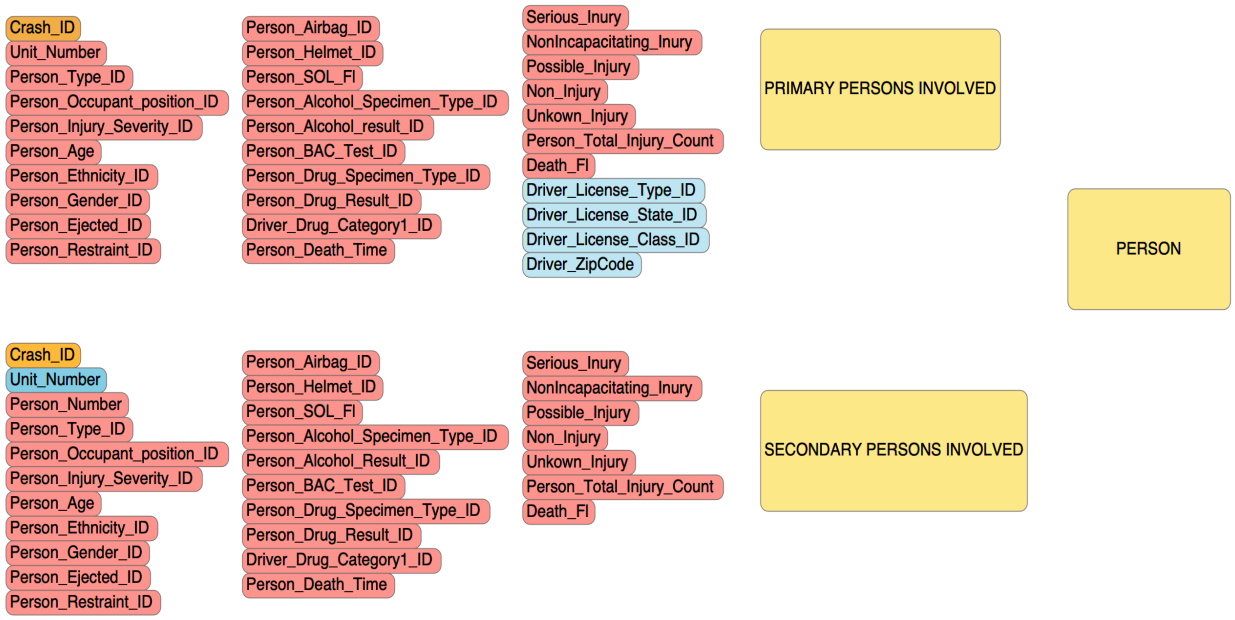


Figure 3.9 Data points from *PRIMARY PERSONS INVOLVED* and *SECONDARY PERSON INVOLVED* – generalized into a *PERSON* involved in the crash

Figure 3.9 shows the logical model combination of the primary persons data set and the secondary person data set into a person data-set that holds all of the information regarding both primary and secondary persons involved in a specific traffic crash distinguished by the Crash_ID. The *PERSON* data set is a generalized class that is composed of the *PRIMARY PERSONS INVOLVED* and *SECONDARY PERSON INVOLVED*. Since each of the people involved in a traffic crash is described individually, the data set is transformed to describe each person individually.

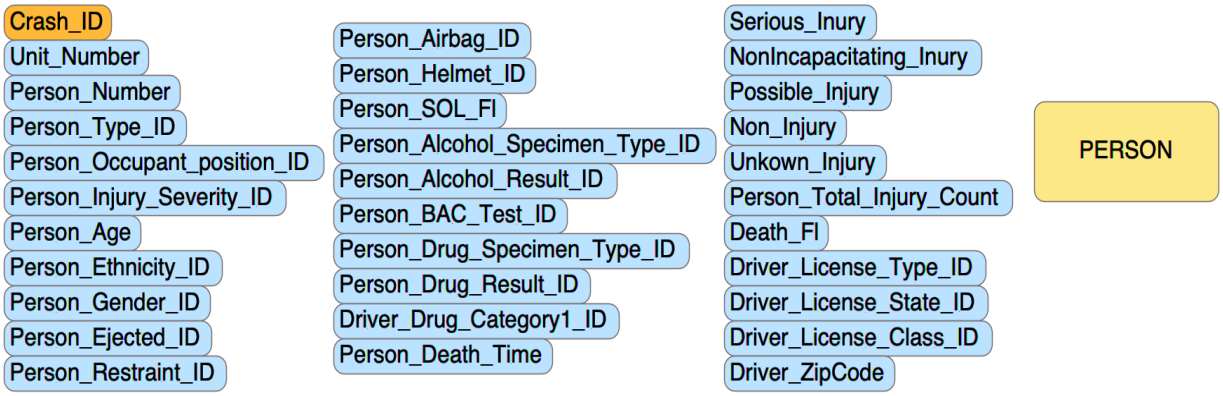


Figure 3.10 Generalization of data points from primary and secondary persons data sets into *PERSON* (yellow node)

Figure 3.10 shows the combination of all of the primary and secondary data points into a unified *PERSON* data set. The data-set will contain values that will not have any information for secondary persons, such as Driver license information and driver zip code since secondary persons are not drivers; in these cases, secondary persons will be described will empty values. To ensure interoperability, it is necessary to combine all of the persons involved in the traffic crash into a single data set.

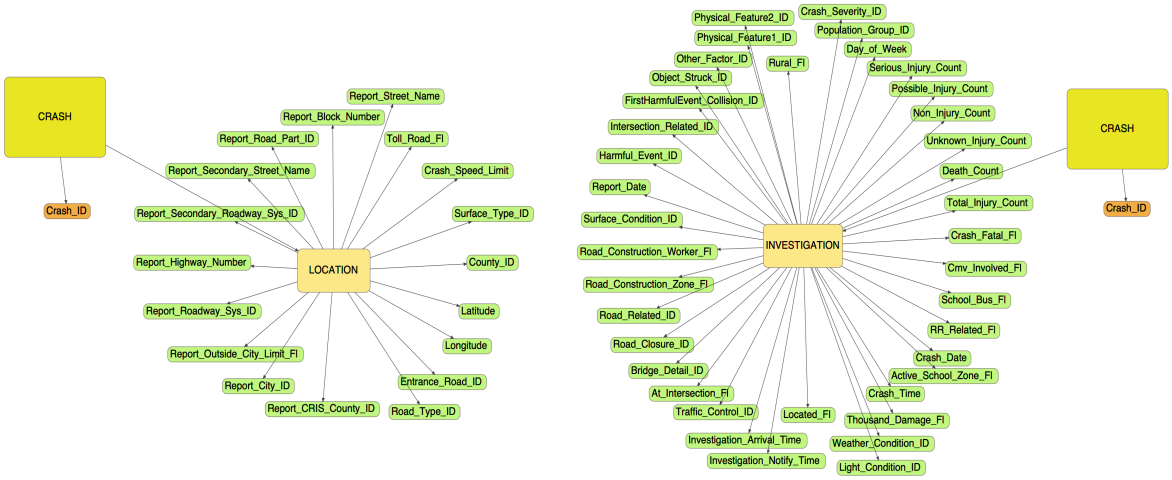


Figure 3.11 Data points from the crash data set with relation to *LOCATION* and *INVESTIGATION*

Figure 3.11 shows the next stage of the model by expanding and linking the generalized idea of a *LOCATION* to the individual data points. Each *CRASH* is linked to a *LOCATION* which contains the data points that describe it. Furthermore, the investigation is linked to the data points that describe the investigation of the traffic crash. Each crash is linked to an investigation which has the data points that describe it. Moreover, each *CRASH* contains a unique Crash_ID that differentiates from a different traffic crash. Figure 3.11 illustrates the distribution of data points into two different datasets (i.e., LOCATION and INVESTIGATION), both providing a Crash_ID.

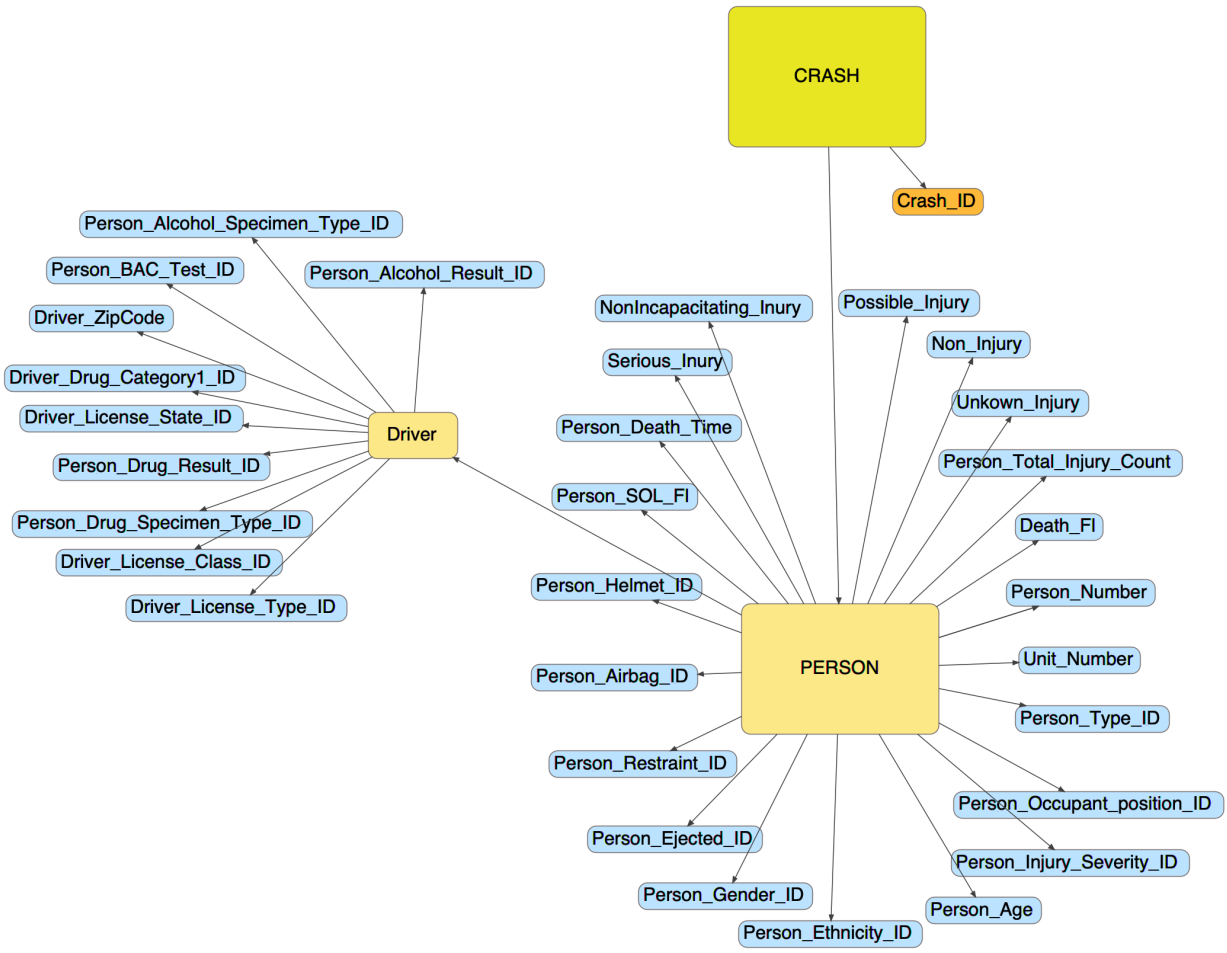


Figure 3.12 Data points from the primary and secondary person with relation to *PERSON*

Figure 3.12 describes the expansion of the persons involved in the traffic crash and the relationship it has to the individual data points that describe it. The data points that are specific to a driver is also explicitly expressed and is linked to the persons involved in the crash. As a result, each crash is linked to persons involved. The *CRASH* has a unique Crash_ID that is necessary to differentiate between which crashes contain which persons involved.

3.4.3 Unified Data Sets

Data sets were unified (or linked together) by discovering the commonalities between the different data sets. The concept of CRASH became a concept that linked the data together, and it was identified by a unique crash id. Each crash is represented the same way, as a knowledge graph. The difference between each crash knowledge graph is dependent on the number of people involved in the crash.

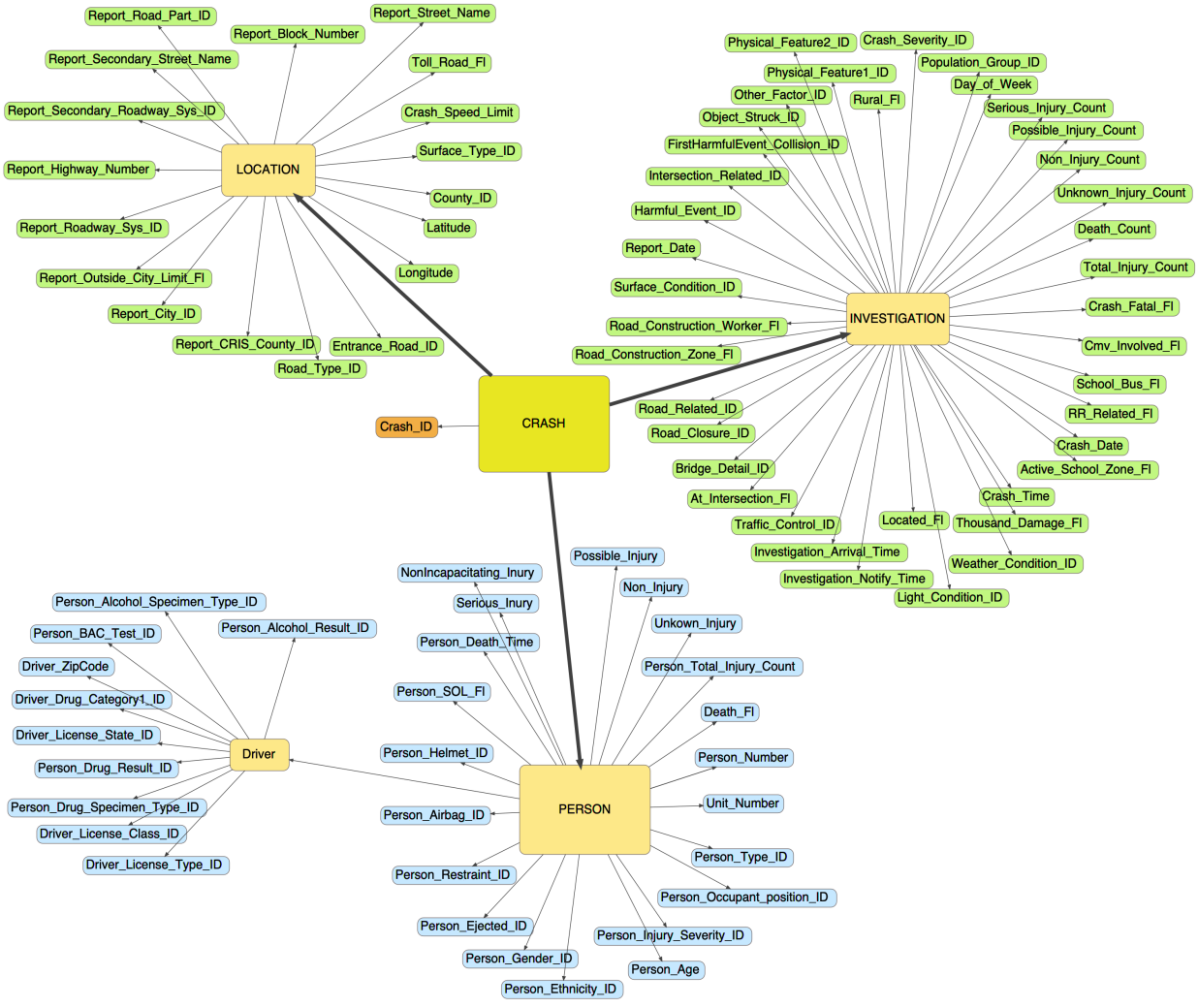


Figure 3.13 Data points of *CRASH* with relation to *LOCATION*, *INVESTIGATION*, and *PERSON*

Figure 3.13 shows the expansion and linking of *CRASH* to the *LOCATION*S, *INVESTIGATION*, and *PERSON* for any given traffic crash. The linking is necessary to show that the three major data sets have a unique Crash_ID in common and are composed in a similar way.

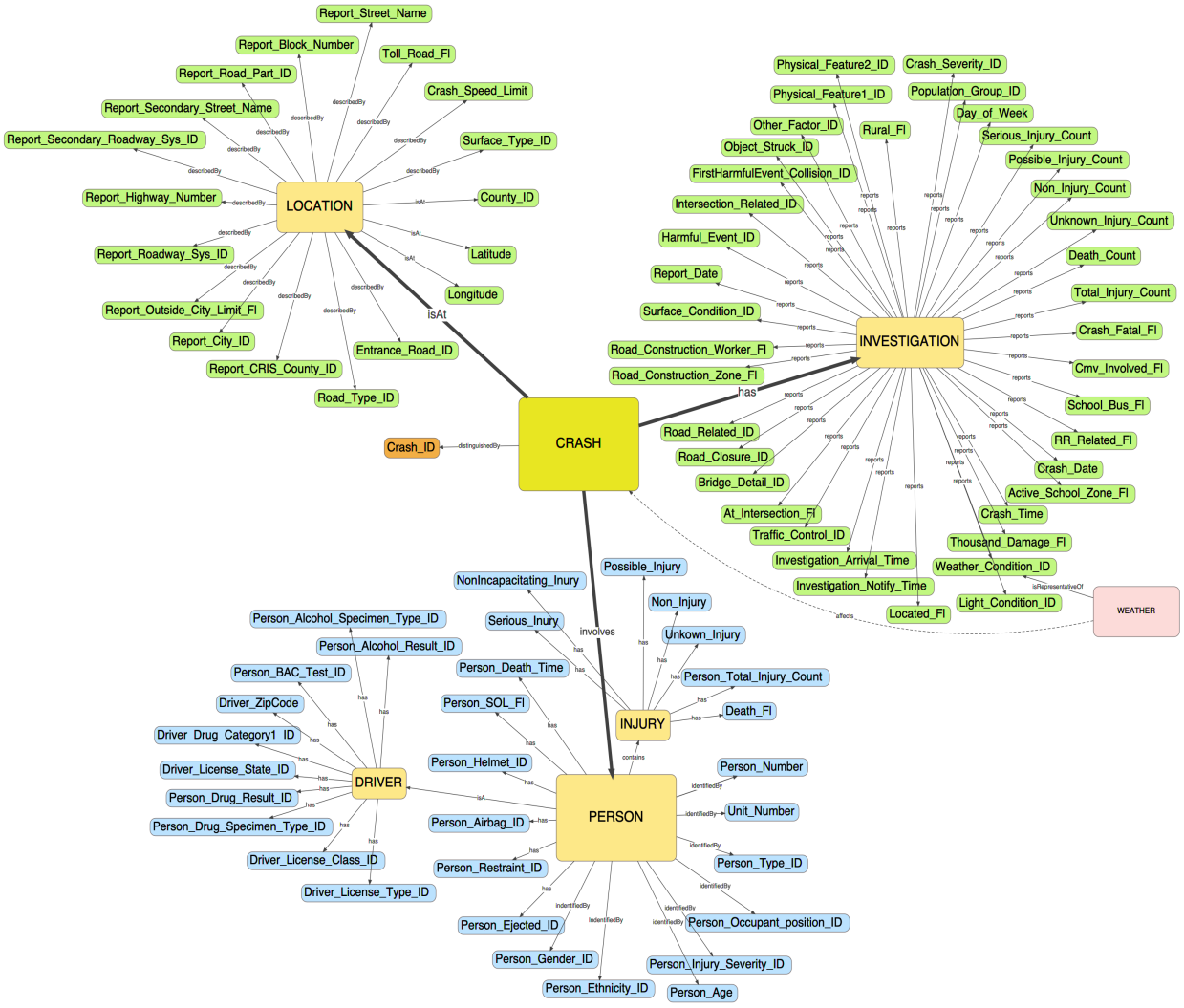


Figure 3.14 Data points with named relation to *CRASH*, *LOCATION*, and *PERSON*

Figure 3.14 continues the expansion of the model by including possible additional data sets that may be relevant. The additional data sets that can possibly be included is WEATHER data that can be linked together with Weather_Condition_ID to describe the weather during the time of the

crash in greater detail. Furthermore, the relationships between data points and larger entities are named to describe the relationship themselves.

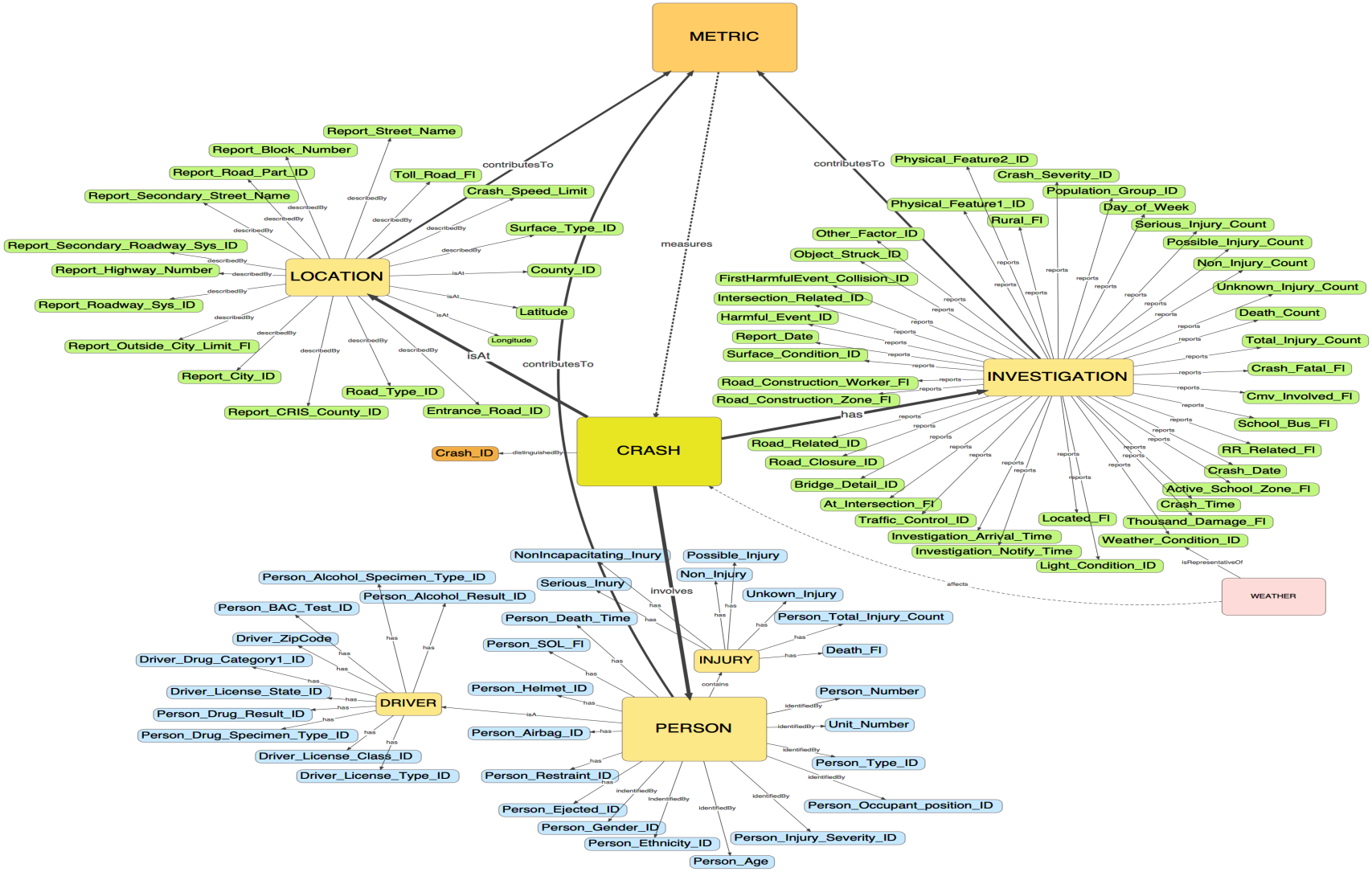


Figure 3.15 All data points with relation to an overarching METRIC

Figure 3.15 describes the final step in the bottom-up model. The model is intended to provide a unique way to develop a new metric that is driven my data itself. This model shows that location, investigation, and persons are derived from individual data points. *LOCATION*, *INVESTIGATION*, and *PERSON* contribute to the development of a METRIC. The METRIC then measures individual crashes to provide a deeper understanding of what the data is contributing and complete the transformation of data to knowledge. The data model will serve as a foundation to determine any additional open problems that can be addressed. By continued use of this model, a process can be developed that is supported by data science to create new ways to measure a domain and make them more effective (Darema, 2004). Moreover, by using this a modeling technique, understanding the process for transforming data to knowledge leverages a maintainable and reusable model (Müller, Reichert, & Herbst, 2007).

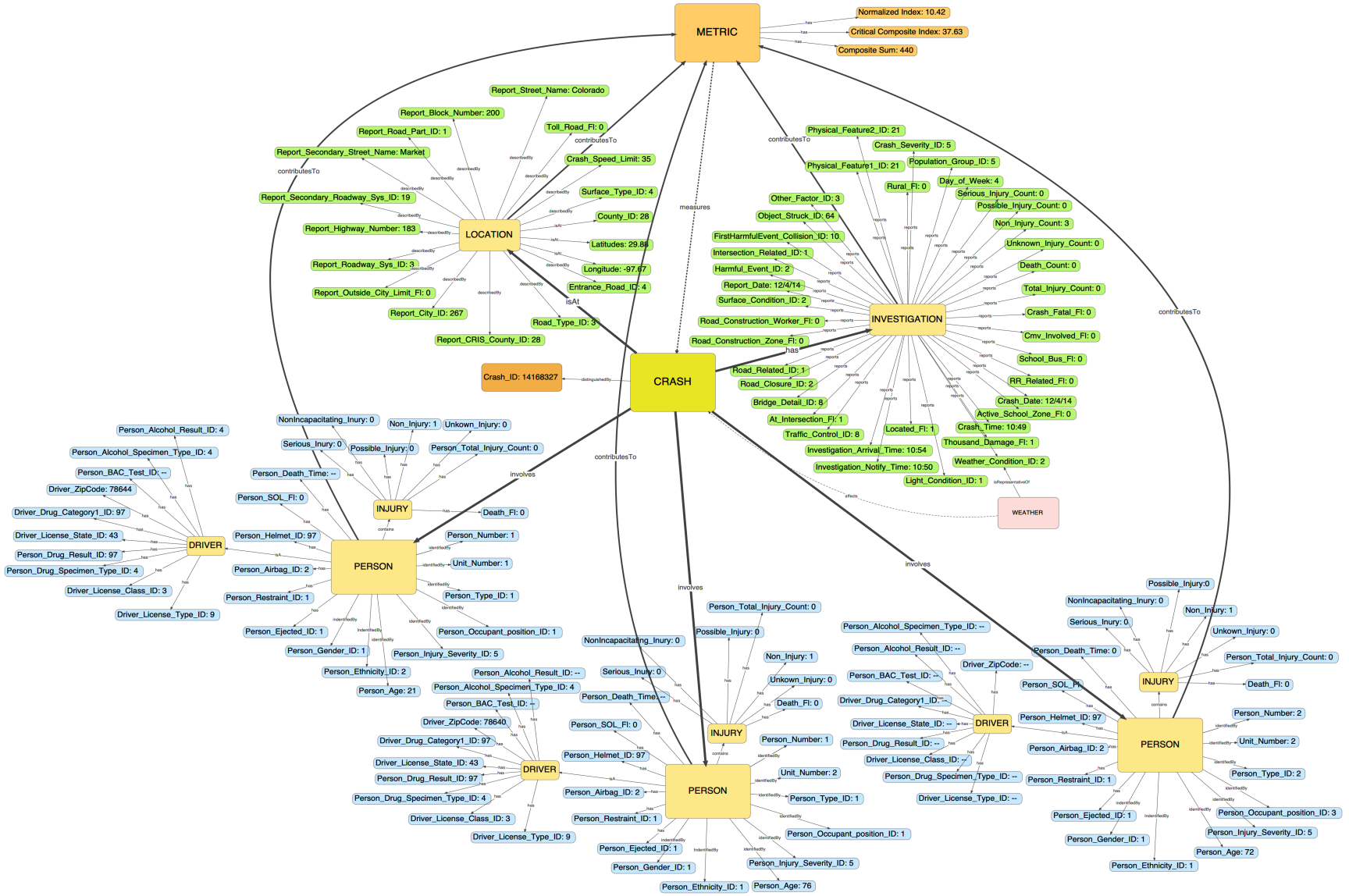


Figure 3.16 Example data representation of a traffic crash, Crash_ID: 14168327

Figure 3.16 shows a full representation of an individual traffic crash. This figure expands on what was shown in Figure 3.15 insofar as it describes the actual values for the data points and shows multiple people in a particular traffic crash. In this traffic crash, there were three individuals involved; two drivers and one additional passenger in the second vehicle. The representation of the people involved is shown in nodes with different shades of blue to differentiate the three individuals.

The traffic crash described in Figure 3.16 is representative on one individual traffic crash. This representation shows that new information can be gathered because the entire crash itself contributes to the metric development by way of the location, investigation, and persons involved. Through the data gathered from the traffic crash, a metric is developed to consider the inputs using weights to produce a generalization of the traffic crash, described by the CCI. The CCI does not consider all of the data points in the traffic crash, but instead is determined based on a selected group of data points that can be modified to fit the need of additional data or domain expert evaluation.

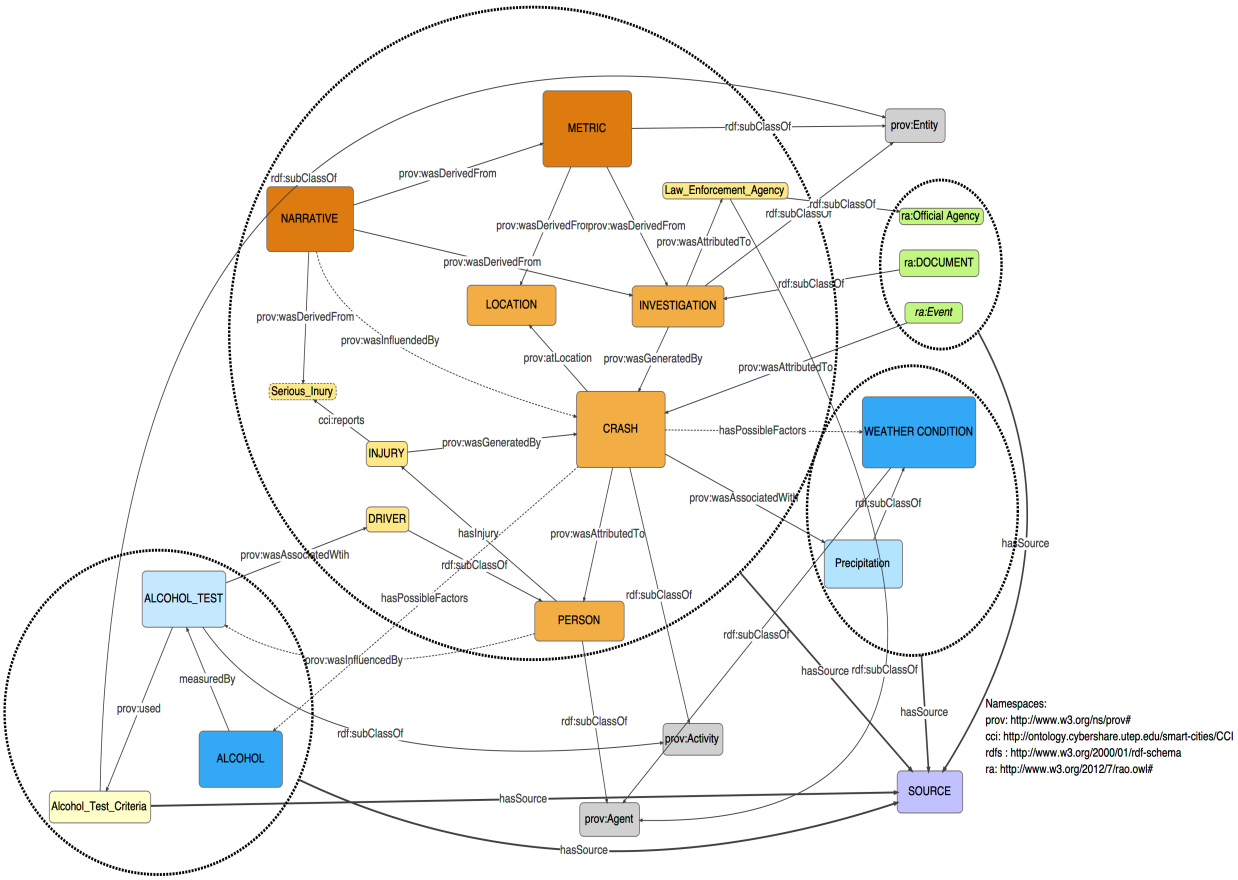


Figure 3.17 CCI Knowledge Graph with semantic mappings to other vocabularies and ontologies to indicate additional relationships

Figure 3.17 shows the representation of a traffic crash and its relationship to location, investigation and person. All of these concepts are well-defined as part of the data sets used in this research. The circles surrounding the concepts are based on using the same data sources. In addition to the concept's location, investigation, and person are three major concepts called weather, alcohol, and narratives. The three additional concepts (Narratives, Weather Condition, and Alcohol) are representative of how additional information can be gathered and the inferences that can be made. In addition to the concepts, a domain ontology called, the road accident ontology (Dardailler, 2012) has been incorporated to link external data into the CCI ontology

being developed. Moreover, a W3C standard upper-ontology called the provenance ontology (PROV-O)(Lebo et al., 2013) has been added to introduce additional formal semantics. An ontology was developed based on the knowledge graph and a the HermiT reasoner (Information Systems Group Oxford University, 2016) was run to create inferences on the data.

3.5 IMPLEMENTATION

The data for this research comes from the Crash Records Information System (CRIS) (TxDOT, 2018). For the years 2014-2018, there are 33 data files each containing approximately 99,000 traffic crashes in each file. There are 33 data files containing approximately 182,000 primary persons involved in those crashes in each file. Additionally, there are 33 data files containing approximately 80,000 secondary persons (non-drivers) involved in those crashes in each file. There are over 23.5 million individual data values of crashes, persons, and secondary persons in the traffic crash data sets being considered.

From the implementation perspective, multipurpose parsers were created to handle different CSV file inputs from multiple sources. The parsers clean and process the raw data into a numerical form, which may be used for future machine learning, then gives a JSON output of the cleaned data. The data provided is the source of factual evidence insofar as it is considered to be true for the development of the BUM methodology.

3.5.1 Data Processing and Mapping

After developing a knowledge graph representative of the available data, it was processed to fit the knowledge graph by using python scripts. The code was originally developed in Java, but after evaluating the scope of the project, it was deemed that python would be a better fit for this

research. Though the data was successfully transformed in Java, the solution was not optimal because of the massive amounts of data that was being parsed; the limitations inherent to the programming language caused memory overflows.

The code written for this work is generic insofar that only a few modifications would be needed to expand it beyond the state of Texas and the data provided through data discovery. Through the use of python, the data could be mapped to a singular metric. The data is representative of the knowledge graph insofar as describing the data points and its respective values. In this stage, the mapping was evaluated for accuracy to the original knowledge graph. The data mapping process takes an approach that can be used for other domain areas based on the data that is given. The data was transformed into integer values that will be useful for machine learning algorithms. The data that is transformed will be representative of the values that come from the raw data provided by the sources.

Parsed data was introduced into a NoSQL database as individual JSON documents. The JSON information can then be retrieved through appropriate queries. The raw JSON files have been uploaded into a MongoDB NoSQL database (MongoDB, 2018) for competency questions. This research shows the process of the dissemination of large heterogeneous data sets by linking them together and transforming it into data that can be queried, thus having knowledge can be captured from it.

The following process is the methodology for the transformation of data into JSON. This methodology is based on the steps developed for a bottom-up methodology. The first portion of the program reads all of the CSV files that were given as an input. Each year of traffic crashes is

presented on up to seven different CSV files. The files are broken up by months from January to December of each of the years being studied.

Using the same design principles developed in the original Java program, the program was translated into Python. Using Python allowed for a lightweight design and allowed leveraging internal libraries that would be useful for data management and cleaning.

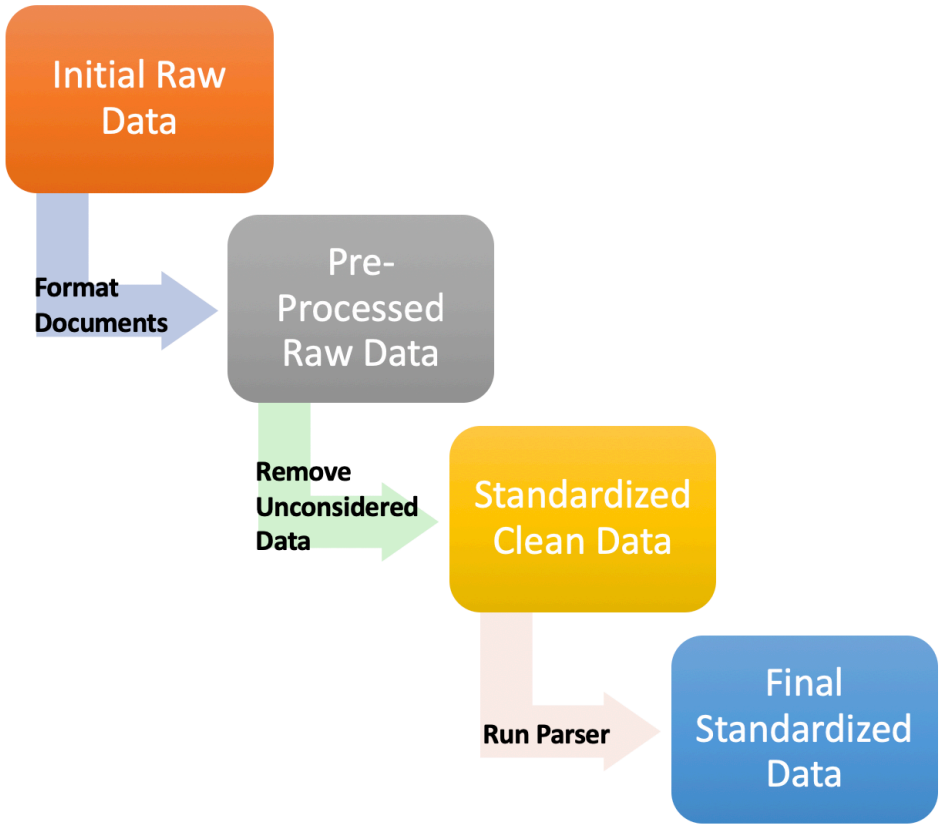


Figure 3.18 Data transformation process

The files retrieved are in a .csv file format; there are instances where portions of the data values within the file contain commas. Without changing any of the values of the data contained in the files, the commas are removed for improved efficiency of the data parser developed; this stage of

the development is done prior to introducing the data set into the parser, thus becoming a pre-processed data-set as shown in Figure 3.24.

After the data has been pre-processed, the files are introduced into the parser. The data is cleaned by removing data that will not be considered in the development of the metric; this data includes repeated information, reporting agency-specific information, information that cannot be compared to a majority of the traffic crashes, and information that does contribute to the development of a new metric (A full list of removed items can be found in the appendix).

Each data set is cleaned and standardized individually. Once all of the data sets have been cleaned, they are combined into similar files (primary persons and secondary persons) and traffic crashes throughout all of the years being viewed. At this stage, an additional file is created that represents each of the data points as an integer for the purpose of using machine learning algorithms to make predictions on it.

Once the data for all of the years has been read and stored, a set of columns are removed by first determining the values that are unnecessary then removing the entire column of that data. Data values were determined to be unnecessary if they were unreported for all traffic crashes or provided repeated information. Once all of the data that is unnecessary has been removed, all of the data from the crashes data set, primary persons data set, and secondary person data set are stored in their own individual files, respectively. The crashes data set undergoes an additional step to transform non-integer values into integers for the purpose of creating a file of values that is representative of the raw data in numerical values. The implementation of the data processing

has a complexity of O(n); this is because each of the rows (each traffic crash) is read individually and accessed at once, further retrieval of the data is in constant time.

The primary and secondary person data sets are combined into a single CSV file using a *pandas* library in Python (Python Software Foundation (PSF), 2019). The output file contains the union of all of the data from both files. Once all of the crash data sets of the multiple years and the primary and secondary person data sets of multiple years has been combined together, the process of transforming it into a JSON file can proceed.

The CSV files containing all of the information from the crashes and persons, respectively, are read and converted into JSON. The final output is two JSON files; one with all of the crash data and the other with all the involved persons data.

3.5.2 Index Development

The final step of the transformation process was the development of the CCI for the application domain of traffic crashes. The CCI was first developed by determining a set of criteria that would be considered as part of the CCI. The entire data-set consisted over a hundred different data points for each traffic crash. The criteria were selected based on the data points that was complete (not missing information for traffic crashes) and what may be important to traffic crash metric development (Cheu & Balal, 2018; National Safety Council & ANSI, 2017; Texas Department of Transportation, 2016, 2017). The criteria chosen for the weighted points focused on three parts: the crash event, the crash location, and circumstances of the traffic crash. The criteria chosen are not the only possible criteria that can be expressed by the CCI, the CCI

provides access to change, add, or remove any selected criteria, then be recomputed with those accepted changes.

Severity, by definition is subjective because it differs from one person to another; there is not a single standard to which a traffic crash is currently evaluated against (Cheu & Balal, 2018; National Safety Council & ANSI, 2017; Texas Department of Transportation, 2016, 2017). The weighted values were determined based on intuitive reasoning for traffic crashes – the weights described are not necessarily valid because there has not been a consensus on how traffic crashes are evaluated. This part of the process is subjective to the individuals that assign the value. The weights are subjective because of the many different perspectives that domain experts may have on a traffic crash. The four sample weighted criteria used in this research is intended to show the generic and adjustable weights that produce results that can be interpreted for both domain and non-domain experts. In practice, immediate severity classification is subjective to those reporting and receiving a traffic crash at the 911 center (Shields, 2018).

The amount of subjectivity that occurs when attempting to classify severity of a traffic crash can be slightly mitigated through additional computations. A machine learning methodology can be developed to do the following: using the data available in this research, each domain expert would individually classify the severity of the traffic crash based on each of the data points for each traffic crash with respect their professional opinion. Each individual traffic crash would be described as minor, moderate, major, or severe by each individual domain expert. This provides a training data set that could be used to converge on appropriate weight values for each of the selected criteria; using the weighted computed values by a machine learning methodology, each traffic accident could be evaluated and its CCI could be determined. This research did not use

this method to compute weights because it requires several thousand domain experts analyzing several thousand traffic crashes for a significant training data set, which is not feasible for this research. However, the CCI can still be adopted and its particular weight values can be adjusted by a governing body for traffic crash reporting. As part of this research, the BUM methodology provides the standard process to attempt to close the gap and introduce objectivity in traffic crash metrics. Although there is a high level of subjectivity, the weights in the samples have been naïvely been developed to conform to reasonable values, as shown by the user evaluation study.

The computation of the CCI is similar to f a decision tree. Decision trees are a model that is used to represent classification and regression applications (De Oña, López, & Abellán, 2013). In a classification method, decision trees are typically useful for finite analysis, meaning there is a finite set of terms to be evaluated. The development of a decision tree is based on binary rules, and based off of those rules, a decision is made, then the tree continues to be traversed until some predefined leaf is visited. From the root node of a decision tree, rules can be explicitly defined for the continued traversal of the tree. Although decision trees are useful, they provide more subjectivity in the work than the CCI computation. Similar to the CCI development, decision trees subjectively provide weight values, however they introduce additional subjectivity by generalizing rules of the decision tree, such as differentiating weights based on gender, age, and season (fall, winter, spring, summer); though these rules may be useful in some contexts, they do not directly affect the severity of a traffic crash. Through the implementation of the CART method (Abellán, López, & de Oña, 2013), which is commonly used by decision trees, it is intended to determine a well-defined classification (or reason for the crash).

Using a similar approach of decision trees, the CCI uses weights that have also been subjectively determined. The subjectivity of the weights cannot be removed as in other research because there is a determined severity outcome level that is used for training purposes; Furthermore, traffic crash severity is subjective in practice, as a result some subjectivity is included for a severity computation. The CCI computes the severity of traffic crashes based on data. Like decision trees, the CCI has a finite set of rules (limited to the amount of data available and their corresponding weights) and from those rules the CCI is computed. Though some decision trees are not limited to binary decisions, they all converge into a single leaf node that provides a solution. The CCI uses the novel approach of mimicking a decision tree on the basis of having criteria-weight (rule-decision) relationship and is given a value based on a defined rule. The CCI computation however, does additional computations to determine a CCI, whereas decision trees do not perform additional computations.

The remainder of the process in the CCI computation removes all human reasoning, introducing a standardization computation mechanism. The weighted values were mapped to the data-set to create a composite sum, composite index, and a normalized index. The index development can be modified to fit different weighted samples for domain experts to use as an input, including adding or removing weighted criteria.

The file containing the list of data points and their associated weights are matched with each individual data point from the crash data set. A list of weights is shown in the appendix; the weights can be manipulated to describe situations with more significance with a higher weight.

Based off of the weight, a composite sum is developed to maintain the total weight sum of severity for a particular traffic crash; all weights are between 0-100. Each crash also has a composite index which is computed – Figure 3.18 shows how the CCI is implemented with respect to the current weights that are given; when a crash has completed the evaluation, the composite sum and composite index is stored along with the Crash_ID.

$$g(W) = \sum_{i=1}^{|W|} (w_i) \qquad for\ criteria\ \neq Death\ Count\ or\ Injury\ Count$$

$$h(V, W) = \sum_{i=1}^{|W|} (v_i * w_i)\ for\ criteria = Death\ Count\ or\ Injury\ Count$$

$$f(V, W) = \frac{g(w) + h(v, w)}{\sum_{i=1}^{|V|} \max(a_i)} * 100$$

Equation 3.19 Formulas to compute a CCI

The formula shown in Equation 3.19 is separated into three parts:

1. g(W): The summation of every *associated criteria weight* (w_i) instance where the criteria value is not equal to death count of injury count
2. h(W, V): The summation of every *instance value* (v_i) multiplied by the *associated criteria weight* (w_i), where the *criteria* is equal to death count or injury count
3. f(V, W): the value of g(W) + h(V,W) is computed the be the composite sum. The composite sum divided by the summation of the maximum possible *associated criteria value* ($\max(a_i)$) multiplied by 100 for normalization

This formula is used to ensure that the values are normalized within the range of [0,100], where absolute worst case possible will be equal to 100 and that the best case (no crash) is equal to 0. However, the values that are greater than 50.1 are considered to be severe because it will include physical injury or death. Collisions, in general, can quickly move from being minor to extreme. Through this metric, a crash is representative of its investigation, the external circumstances that may have possibly contributed, and its effect on people; all of these factors are necessary when describing a composite style index.

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Crash
> 20 – 40	Moderate Crash
> 40 – 50	Major Crash
> 50 +	Severe Crash

Figure 3.20 CCI Severity Chart

Minor Crash
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Crash
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Crash
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Crash
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Figure 3.21 CCI Severity Chart Including Common Crash Features

Once all of the crashes have been given a computed composite sum and composite index, a normalization function is computed for each crash to determine its severity with respect to all other crashes in the data set. These values determined by the computation can then be mapped back to Equation 3.19 and Equation 3.22, which can be adjusted based on the CCI values. Traffic crash severity is commonly determined based on single factors or points of views (e.g. with respect to vehicle, with respect to persons involved, with respect to location) (National Safety Council & ANSI, 2017). The range values presented in Figure 3.20 are based on a combination of values presented by the National Safety Council of No damage/injuries (minor crash), other damage/injuries (moderate crash), functional damage/injuries (major crash), and disabling damage/fatalities (severe crash).

$$y(v_i, w_i) = \mathrm{g}(w_i) + \mathrm{h}(v_i, w_{i,})$$

$$k(x_i) = \frac{y(v_1, w_1) \dots y(v_n, w_n) - \min(y(v_1, w_1) \dots y(v_n, w_n))}{\max(y(v_1, w_1) \dots y(v_n, w_n)) - \min(y(v_1, w_1) \dots y(v_n, w_n))} * 100$$

Equation 3.22 Formula to compute the normalization

Based on the formulas described in Equation 3.22, the normalization is the following:

1. $y(v_i, w_i)$: $g(w_i) + h(v_i, w_i)$ is the composite sum for an individual traffic crash
2. $k(x_i)$: The composite sum of an individual crash minus the smallest composite sum of the set of all considered traffic crashes; divided by the difference of the maximum composite sum of the set of all considered traffic crashes and the smallest composite sum of the set of all considered traffic crashes; multiplied by 100 for normalization

$k(x_i)$ is a normalized value between a range of [0,100], where 0 has a worst composite sum (crash severity) than 0% of all crashes in the data set and where 100 has a worst composite sum (crash severity) than 100% of all crashes in the data set. This formula ensures that all of the crashes in the data-set are relative and can be compared to each other. Both the formulas shown in Equation 3.20 and Equation 3.22 are expressed as an example in Table 3.3. Table 3.3 is an example of a single traffic crashes. The instances values (v) are shown to be as human readable values, however the formula used is based on integer values that are identical representations of the human readable form.

Table 3.3 shows a single traffic crash by its *instance values* (v), *associated criteria weight* (w) & *possible criteria weight* (A) with the computations of the CCI and Normalized Index

Criteria	Instance Value (V)	Associated Criteria Weight (W)	Possible Criteria Weight (A)
Fatal Crash	YES	80	80
Commercial Vehicle Involved	NO	0	50
School Bus Involved	NO	0	50
Railroad Involved	NO	0	50
Active School Zone	NO	0	50
Crash Time	9:03 PM	40	60
At Least $1000 In Damages	YES	50	50
Type of Line Division	2 LANE, 2 WAY	40	50
Construction Zone	NO	0	50
Construction Workers Present	NO	0	50
Crash At Intersection	NO	0	50
Weather Conditions	CLEAR	0	90
Light Conditions	DARK, NOT LIGHTED	70	70
Surface Conditions	DRY	10	90
Harmful Event (Object Struck)	HIT MOVING VEHICLE	85	90
Number of Serious Injuries*	1	85	85
Number of Non-Incapacitating Injuries*	0	0	55
Number of Possible Injuries*	2	4	2
Number of Non-Injuries*	0	0	0
Number of Unknown Injuries*	0	0	2
Number of Fatalities*	2	190	95
	g(W)	375	
	h(V,W)	279	
	COMPOSITE SUM: g(xi) + h(xi)	654	
	Maximum Possible Criteria Weight: ai		1169
	Possible Criteria Weight for Injuries/Death Count are per person		
	Critical Composite Index	55.945	
	min(Composite Sum For All Data)	45	
	max(Composite Sum For All Data)	3835	
	k(x)	16.069	

3.5.3 NoSQL Database Implementation

The NoSQL database that is being used for this research is MongoDB (MongoDB, 2018). A NoSQL database was chosen for this work because of the flexibility that it provides. Moreover, the traffic crash data is document based, being that it is in a JSON format, and can be modeled in many different ways that may be necessary. In this work, traffic crash data has been placed into a single database with three collections.

Table 3.4 Shows the Collisions Database and the collection of data it contains

Collisions Database	
event (collection)	Contains all of the crashes and their respective composite critical index
people (collection)	Contains all of the people involved in the crashes

Table 3.4 describes how data will be separated and stored within the crashes database. The Collisions Database holds two collections: *event* and *people*. The use of these collections is based on the two main clusters of crash data and to improve the efficiency of the queries. As the data size increases, it is necessary to separate the data types to ensure a reasonable query time. An additional collection *complete* is used to maintain a backup of all of the data but is not used as part of the standard query process. The event collection contains each individual crash that occurred on the roadway and its computed CCI. The people collection contains each individual person that was involved in all of the crashes. The complete collection contains both every individual crash that occurred on the roadway and its computed CCI as well as each individual person that was involved in all of the crashes. For each of the collections, each of the entries is an individual document that contains the respective information.

Queries can be done on the crashes collection if users do not want information about people; queries can be done on the people collection if users do not want information about individual crashes. Queries can also be done where an aggregation of both collections can be done to produce a single document output that can be interpreted. As a basis to the work, the NoSQL database provides the storage and query functionality to provide services to a real-world application in both visualization and competency questions.

Additionally, the JSON-LD document for each traffic crash can be transformed into an OWL document by introducing it into the OWL API (Manchester, 2011). Transforming JSON-LD into OWL is not a trivial process. Projects from the Cyber-ShARE center of excellence, including GOWL and additional data narratives project have begun to transform data using the OWL API, however a complete method has not been developed; these works use alternative methods including intermediate Java objects as a way to begin the transformation. The resulting JSON-LD documents are yet to be compatible with ontology editors such as Protégé (Musen, 2015). JSON-LD provides the ability to semantically annotate data, however, the complex structure of a JSON-LD document compatible with an ontology editor or triple store nears human unreadability because of instance, object property, and data property syntax.

In this research, the process of transforming JSON-LD has begun by annotating the JSON into a standard JSON-LD format. The individual JSON-LD documents have been given two major additional fields that provide the context required for becoming linked data. First, each individual traffic crash has been annotated with an ontology "@id" represented by a URI (e.g. http://ontology.cybershare.utep.edu/smart-cities/CCI#[*Crash_ID*], where Crash_ID is an individual traffic crash identification). Secondly, each traffic crash has a "@type", namely crash. For each traffic crash, its type is annotated to be representative of standard W3C syntax (e.g. [http://www.w3.org/2002/07/owl#NamedIndividual, http://ontology.cybershare.utep.edu/smart-cities/CCI#Crash]), where each traffic crash is a named individual (described by owl) and a Crash. The addition of the "@id" and "@type" provide the annotation to describe each individual traffic crash as an individual instance of an ontology; this process is replicated for each traffic crash in the data-set.

3.6 COMPETENCY QUESTIONS

Competency questions have been developed based on informal communication with the El Paso Police Department and workers at the 911 call center (Shields, 2018). Through NoSQL databases, questions can be queried in such a way that will be useful to understand specific information. Where metrics are usually generalizations to understand in a broad way, competency questions play a crucial role in finding information for future researchers. The questions being asked can be an aggregation of information from the different collections in the database or individual queries.

Competency questions are crucial in understanding scientific data because they provide a way to answer questions about the underlying data (Azzaoui et al., 2013). The questions themselves show that the system is functional and complete insofar as being able to answer questions on the data. For this research, the competency questions can answer information about traffic crash data that has been collected to give insight into what has occurred on the roadways. There are many competency questions that can be asked; a sample set of competency questions is included below and classified by those that can be answered by current metrics and those that can be answered with the outcomes of the BUM methodology, including the use of metrics and generation of inferences from the formal description of the model.

3.6.1 Competency Questions answered by Current Metrics

1. How many traffic crashes occurred in *[city]* between [*Year*] and [*Year]* ?

(Such that *city* is a city in the State of Texas and *Year* is in the range [2014, 2018])

2. How many fatal crashes occurred in *[city]* in *[Year]*?

(Such that *city* is a city in the State of Texas and *Year* is in the range [2014, 2018])

3. How many traffic crashes occurred on a *[day of the week]* in *[Year]* in *[County]*?

(Such that *day of the week* is [Sun-Sat], *Year* is in the range [2014, 2018], and *County* is a County in Texas)

4. How many crashes occurred in the State of Texas in [*Year]*?

(Such that *Year* is in the range [2014, 2018])

5. How many crashes occurred at an intersection in [*Year]* in Texas?

(Such that *Year* is in the range [2014, 2018])

6. How many fatal crashes occurred at an intersection in [*Year]* in Texas?

(Such that *Year* is in the range [2014, 2018])

3.6.2 Competency Questions answered by the BUM methodology

7. How many crashes were classified as [*Severity*] by the CCI in Texas between [*Year*] and [*Year]*?

   (Such that Severity is minor, moderate, major, or severe; city is a city in Texas; and Year is in the range [2014, 2018])

8. How many crashes occurred on Rpt_Hwy_Num 10 (I-10) and were classified as [*Severity*] by the CCI in [*City*] in [*Year*]?

   (Such that Severity is minor, moderate, major, or severe; city is a city in Texas; and Year is in the range [2014, 2018])

9. How many crashes occurred on Rpt_Rdwy_Sys_ID 1,2,3,14 (Interstates, US Highways, State Highways, Spurs) and were classified as [*Severity]* by the CCI in Texas in [*Year*]?

   (Such that Severity is minor, moderate, major, or severe; city is a city in Texas; and Year is in the range [2014, 2018])

10. Where are all of the traffic crashes involving a fatality in El Paso, County, Texas in 2014?
11. Show information for both the traffic crash event and people involved in traffic crash Crash_ID 13630135.

Table 3.5 lists additional queries to be answered by the knowledge graph used in this work. These queries illustrate the use of external vocabularies such as the Provenance Ontology and the use of standard reasoners that use the rules encoded in the knowledge graph to generate inferences.

Table 3.5 Describes the queries and characteristics on an ontology

Query		Characteristics
12.	Find the individuals of Crash that are included in a Narrative	Conjunctive query, transitive relations, inference
13.	Find all individuals that are Entity, Agent, and Activity according to the PROV ontology.	Ontology Integration, Upper-Level ontology usage,
14.	Find the individuals that have a blood alcohol level above the legal limit	Conjunctive query, inference using logical reasoning
15.	Find the individuals that have alcohol as a possible causing factor	Conjunctive query, transitive relations,
16.	Find the individuals of Person that are included in a Narrative	Conjunctive query, transitive relations

Chapter 4: Results

The El Paso, TX area poses a unique challenge given its position as a border with Mexico. By using a knowledge graph as a high-level data model, it has been shown that data sets have relationships that may be useful in the development of traffic safety and efficiency metrics (Mejia, 2017). This idea leverages the Linked Open Data principles for research. A knowledge graph can be expanded and used as a way to link data together. Tim Berners-Lee has four main principles of Linked Open Data, "1) Use URIs as names for things, 2) Use HTTP URIs so that people can look up those names, 3) When someone looks up a URI, provide useful information, 4) Include links to other URIs. so that they can discover more things" (Berners-Lee & W3C, 2009). Through the implementation of a knowledge graph, traffic crash data and the CCI is open to become linked to other sources and domains for additional ways metrics can be used to expand the knowledge, specifically in traffic crashes.

4.1 Data Representation

The data sets used were separated into individual traffic crashes; based on the specifications of MongoDB NoSQL each traffic crash is its own individual document (MongoDB, 2018). As a result of having each traffic crash reported in Texas for the years of 2014-2018 approximately 3,027,861 documents are being stored and are accessible; each of the documents have 73 data points (e.g. Crash_ID, Crash_Date, Crash_Time).

In addition to the NoSQL database, a geographical mapping application was developed. The application was developed using the Google Maps API (Google, 2018). As an example, shown in Figure 4.1, a query was run using a Python script that returned a JSON document from which the

Google Maps interpreted and geographically mapped it. A data view was developed through the introduction of a visual aspect of traffic crash reporting.

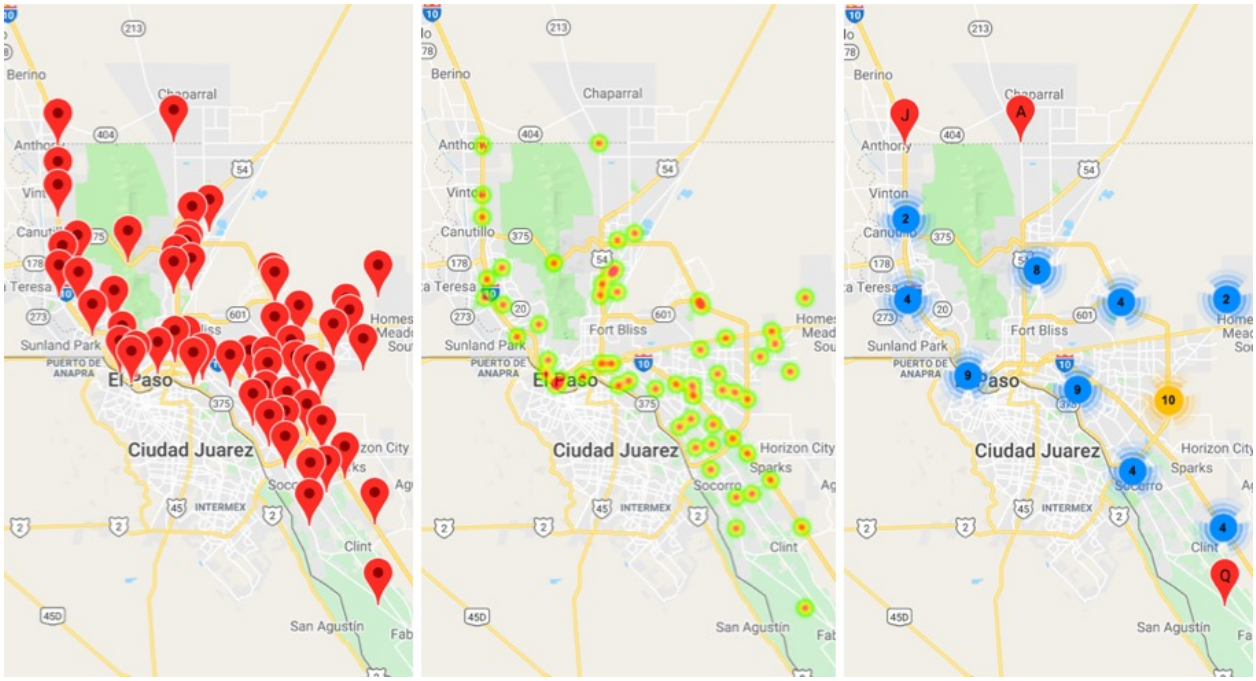


Figure 4.1 Map of all crashes involving a fatality in El Paso County, TX in 2014.

Figure 4.1 shows a map of all of the crashes involving a fatality in El Paso County, TX in 2014. Individual data points of each crash (left); heatmap of the traffic crashes (middle); clustering and count of all of the crashes per generalized location (right). This figure is a visual map representation of all of the crashes queried as part of the competency question 10: *"Where are all of the traffic crashes involving a fatality in El Paso, County, Texas in 2014?"*. The locations were determined by the reported geographic location from the crash investigation. Moreover, the information provided by the visualization answers the competency by showing where the traffic crashes in question occurred as well as shows additional mapping techniques of the data that may be useful for domain experts.

4.2 Answering Competency Questions

Competency questions are necessary to show the completeness of an application or system. Moreover, they can also be used to gather new information that is needed for advanced analysis and understanding of a domain. The competency questions for this research are expanded from simple frequency metrics to questions that require more in-depth responses and visualizations. The answering of competency questions occurred through queries of the data on the MongoDB NoSQL database (MongoDB, 2018). Table 4.1 shows a listing of defined competency questions and its respective answers.

Table 4.1 Sample set of competency questions queried on traffic crash data

#	Question	Answer
1	How many traffic crashes occurred in Austin between 2014-2018?	81,134
3	How many traffic crashes occurred on a Friday in 2015 in Travis County?	3,212
4	How many crashes occurred in the State of Texas in 2014?	555,206
7	How many crashes were classified as 'severe' by the CCI in Texas between 2014-2018?	9,227
8	How many crashes occurred on Rpt_Hwy_Num 10 (I-10) and were classified as 'severe' by the CCI in El Paso in 2014?	3
4	How many traffic crashes occurred in the State of Texas between 2014-2018?	3,027,861
10	Where are all of the traffic crashes involving a fatality in El Paso, County, Texas in 2014?	See Figure 4.1

11	Show information for both the traffic crash event and people involved in traffic crash Crash_ID 13630135.	See Figure 4.2

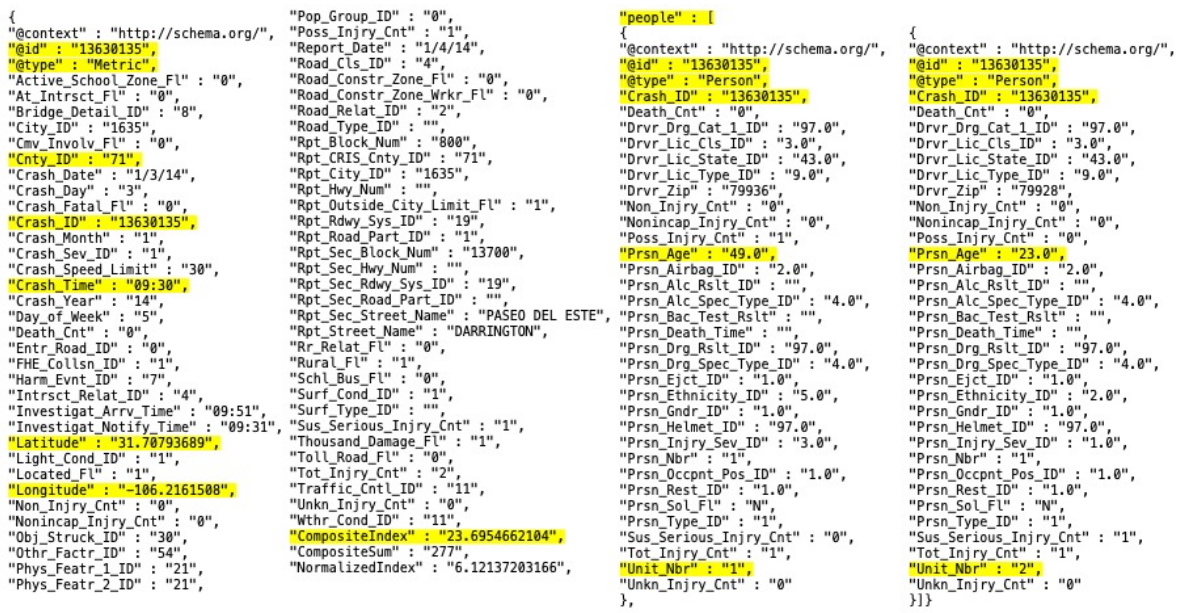


Figure 4.2 Traffic crash data for a specific traffic crash event; Crash_ID: 13630135

Figure 4.2 shows a specific traffic crash (Crash_ID: 13630135), its details and the people involved in the crash. Each crash has the same data points associated with it as described by the BUM methodology. This information highlights a few major points of interest with the traffic crash including, id, the type, Cnty_ID, Crash_ID, Crash_Time, Latitude, Longitude, and CompositeIndex. Furthermore, in each crash are the people involved; each of the people involved in the crash has their own set of relevant data. Some of the highlighted points are the following: semantic id, type, Crash_ID, Prsn_Age, and Unit_Nbr.

The queries show in Figure 4.3-4.6 are semantically based queries that uses a logical reasoner, HermiT reasoner (Information Systems Group Oxford University, 2016), to check for the consistency of an ontology and provides inferences.

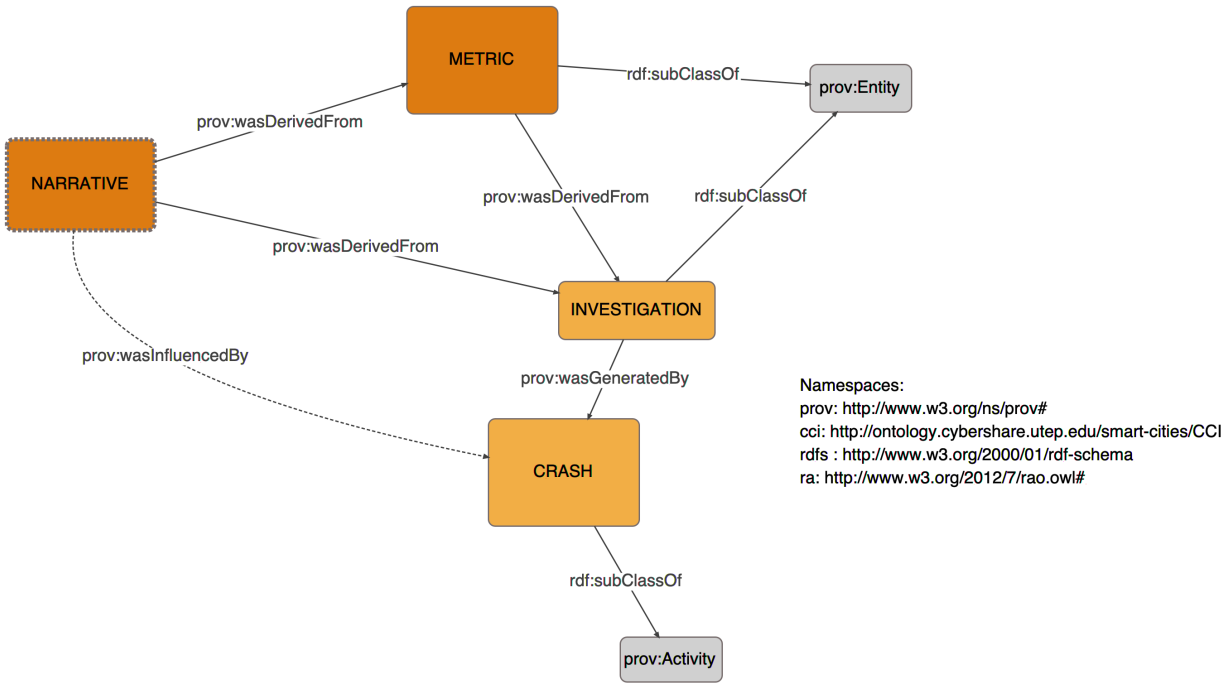


Figure 4.3 Graphical representation of a query that requires the use of an inferenced relationship (dotted line).

Figure 4.3 is based on a query from Table 3.5, Find the individuals of Crash that contribute to a Narrative. The query begins by acquiring all of the individual traffic crashes. As a result of running the HermiT reasoner, the acquired individuals have an inferred link to the concept NARRATIVE. This link is inferred by the reasoner through the transitive relationships between CRASH and NARRATIVE. An inference is a link between concepts that is not explicitly asserted. From the traffic crashes acquired, each of them has an investigation that was generated by the individual traffic crashes, this is linked together by the provenance relationship

prov:wasGeneratedBy. NARRATIVE is linked directly to INVESTIGATION through an asserted relationship, prov:wasDerivedFrom. Through this link it is inferred that the narrative of a traffic crash was influenced by traffic crash itself.

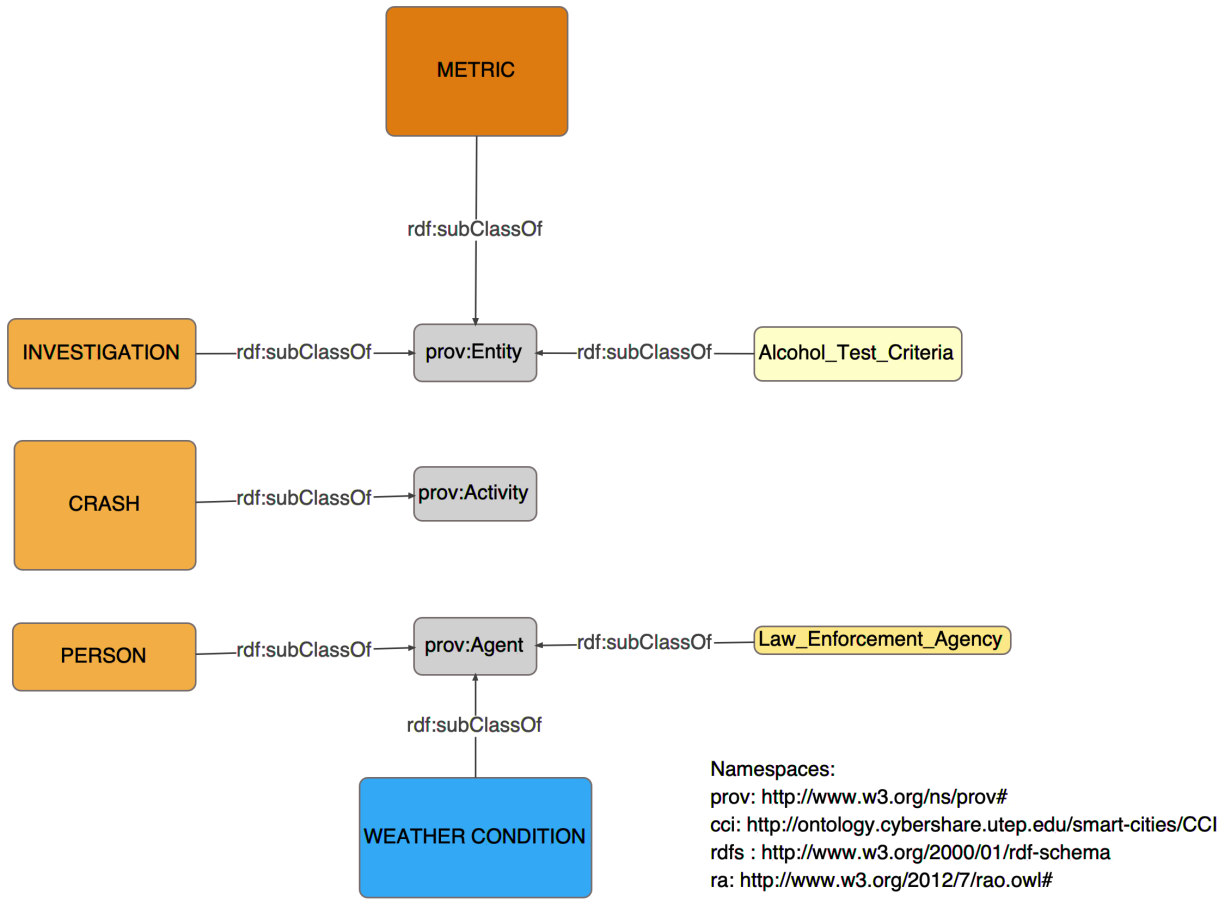


Figure 4.4 Graphical representation of a query that uses concepts of an upper-level ontology (PROV-O)

Figure 4.4 is based on a query from Table 3.5, Find all of the individuals that are Entity, Agent, and Activity. The query acquires all of the individual instances of traffic crash, investigation, person, weather condition, alcohol test criteria, metric, and law enforcement agency that are part of the upper-level ontology concepts: Entity, Agent, and Activity as described in the PROV-O

ontology. As a result of classifying each individual instance from the concepts as being a subclass of either Entity, Activity, or Agent, individuals are returned as: belonging to Entity – all traffic crash investigation instances, Alcohol test criteria instances, and traffic crash metric instances; belonging to Activity – traffic crash instances; and belonging to Agent - traffic crash person instances, weather condition instances, and law enforcement agency instances. This query illustrates interoperability with other data sources and systems that use the PROV-O upper level ontology (Belhajjame et al., 2013).

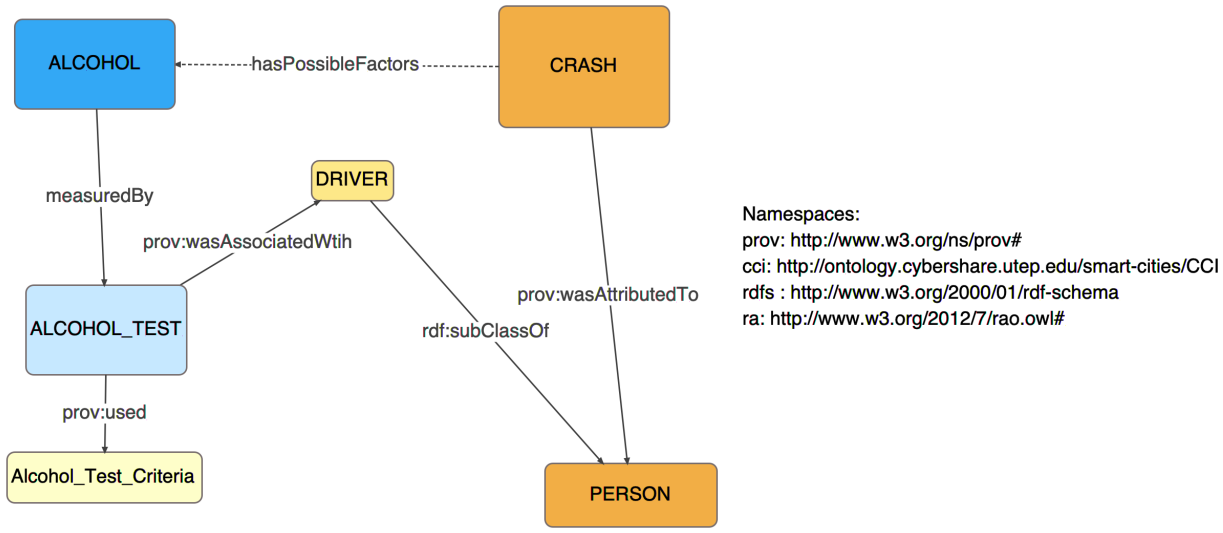


Figure 4.5 Graphic representation of a query that uses inferences to identify possible factors in a traffic crash

Figure 4.5 is based on queries Table 3.5, find the individuals that have blood alcohol level above the legal limit and find the individuals that have alcohol as a possible causing factor. The query begins by acquiring each individual traffic crash. As a result of running the HermiT reasoner, the acquired traffic crashes have an inferred link to ALCOHOL. This means that a relationship between CRASH and ALCOHOL exists, however it is not explicitly asserted. The link between

CRASH and ALCOHOL is attributed to a transitive relationship between PERSON and ALCOHOL; this relationship is determined because an individual person in a traffic crash can be a driver. In some cases, a traffic crash requires a blood alcohol test of a driver in a traffic crash. The driver in the traffic crash is given an alcohol test and based on a set of criteria an alcohol reading is conducted. This query expands to provide logical assertions as part of the alcohol test criteria that determines if a driver had a blood alcohol concentration over the legal limit of 0.08 (Texas)([TABC], 2019); Alcohol is measured by the result of the alcohol test, and if someone is over the legal limit, Alcohol may have been a possible factor in the crash. These results demonstrate the use of external knowledge not initially found in the knowledge graph but that can be mapped to the concepts ore relationships in the knowledge graph given its relevance to the domain to discover additional knowledge.

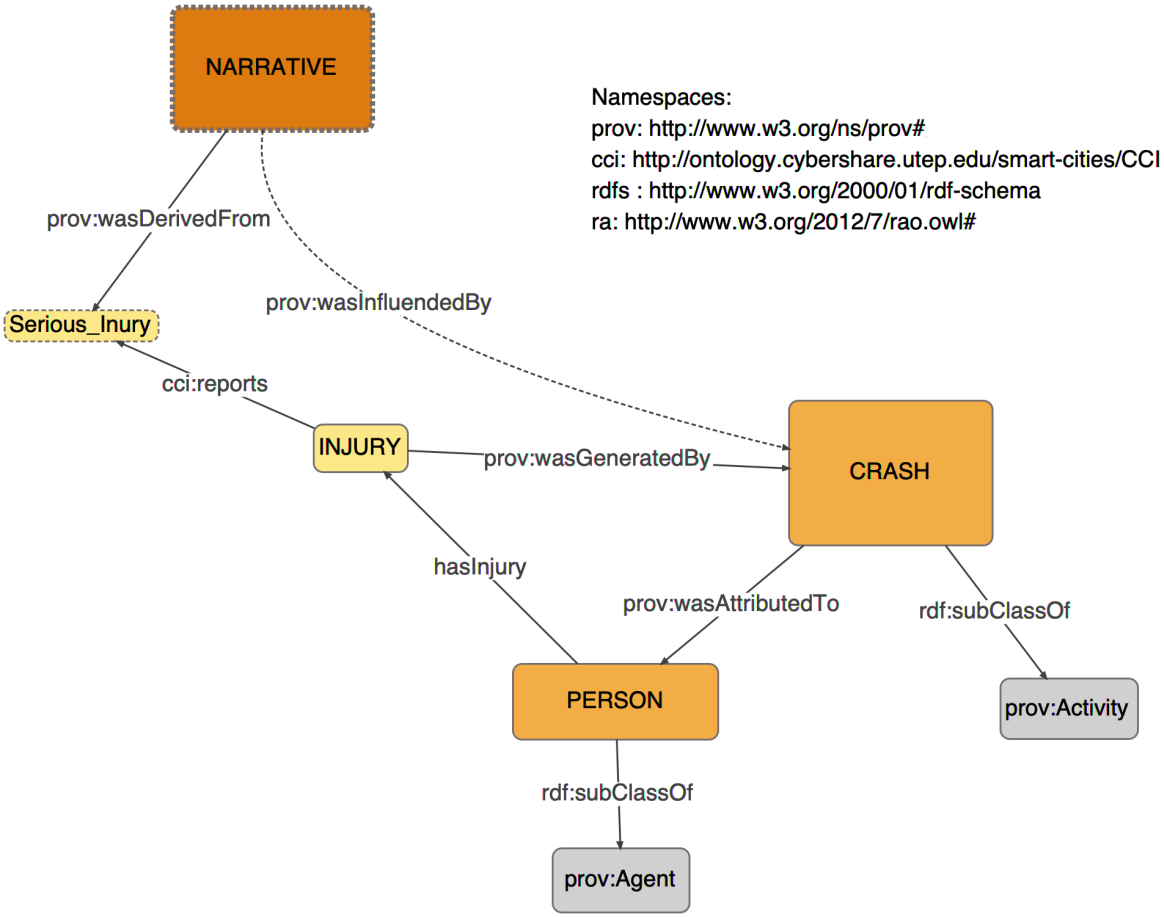


Figure 4.6 Graphic representation of a query that requires inferences to link a narrative and a traffic crash through a person involved

Figure 4.6 is based on a query Table 3.5, Find the individuals of PERSON that contributes to a NARRATIVE. The query begins by acquiring all of the traffic crash person instances. The acquired individuals have a relationship to the concept INJURY (i.e. a person has an injury). This person injury is reported as an injury type (i.e. Serious Injury). The traffic crash narrative instance is derived from the injury type which is information that can be explored through a data narrative.

The information gathered through the inferences and relationships described provide a foundation for data to be mapped to additional ontologies and knowledge graphs. Reusing data loses semantic because there is not direct access to the data source, but through proper semantic mapping semantics are redefined. By mapping to other ontologies, additional ontologies can use the CCI that was developed through the BUM methodology. The mapping of additional ontologies enables data interoperability; the work done through this ontology provides foundational knowledge that can be accessed outside of the ontology. The addition of semantic annotation to the data-sets contribute to semantic annotations in a real-world domain for the purpose of increased knowledge gain in data narratives and metric representation.

These queries show the significance of integrating multiple data sets into an aggregated set of information for the dissemination of information. Through the model developed by the BUM methodology, data heterogeneity has been removed and the data sets are interoperable.

The competency questions and respective answers in this work do not provide statistical information that can be gathered from other database query applications. The answers to the competency provide information that completes the transformation of data to knowledge. Through in-depth queries, all of the data available can be leveraged to make more informed decisions. Furthermore, as additional fields or data are added to the data set, the model can easily evolve and continue to provide information to users in the form of competency question answering and map visualization, whereas traditional relational databases or even ontologies cannot provide that flexibility.

4.3 METRIC RESULTS

The comparison table available in Appendix C shows a detailed comparison chart between the two sample critical weights used to compare two critical composite weight indices. Both of the sample weights are representative of possible points of view for an individual severity of a traffic crash. The weights are assigned to relevant columns determined by factors that may affect the health and welfare of individuals in a crash; the weights are given from 0 to 100, where 0 is the least severe and 100 is most severe, thus giving the least and most weight for that category respectively. The CCI was developed based off of the given weights. The CCI describes the severity of crashes that occur on the roadways. The weights were determined to weight heaviest on fatalities, and injuries, then consider external circumstances that may have contributed to the traffic crash such as weather and light conditions, and finally the location of the traffic crashes. Weight sample two has high value weights for each of the criteria as a comparison between a realistic weight sample.

4.4 TRAFFIC CRASH CASE STUDIES

4.4.1. Traffic Crash Case One – Minor

In this traffic crash case, a ‘minor’ crash is examined, based off of the CCI described by Figure 3.18.

```
{
"@context" : "http://schema.org/",
"@id" : "15575237",
"@type" : "Metric",
"Active_School_Zone_Fl" : "0",
"At_Intrsct_Fl" : "0",
"Bridge_Detail_ID" : "8",
"City_ID" : "244",
"Cmv_Involv_Fl" : "0",
"Cnty_ID" : "101",
"Crash_Date" : "1/2/17",
"Crash_Day" : "2",
"Crash_Fatal_Fl" : "0",
"Crash_ID" : "15575237",
"Crash_Month" : "1",
"Crash_Sev_ID" : "5",
"Crash_Speed_Limit" : "65",
"Crash_Time" : "03:04",
"Crash_Year" : "17",
"Day_of_Week" : "1",
"Death_Cnt" : "0",
"Entr_Road_ID" : "0",
"FHE_Collsn_ID" : "1",
"Harm_Evnt_ID" : "7",
"Intrsct_Relat_ID" : "4",
"Investigat_Arrv_Time" : "03:12",
"Investigat_Notify_Time" : "03:05",
"Latitude" : "29.68822175",
"Light_Cond_ID" : "4",
"Located_Fl" : "1",
"Longitude" : "-95.03169053",
"Non_Injry_Cnt" : "1",
"Nonincap_Injry_Cnt" : "0",
"Obj_Struck_ID" : "39",
"Othr_Factr_ID" : "54",
"Phys_Featr_1_ID" : "21",
"Phys_Featr_2_ID" : "21",
"Pop_Group_ID" : "6",
"Poss_Injry_Cnt" : "0",
"Report_Date" : "1/2/17",
"Road_Cls_ID" : "2",
"Road_Constr_Zone_Fl" : "0",
"Road_Constr_Zone_Wrkr_Fl" : "0",
"Road_Relat_ID" : "4",
"Road_Type_ID" : "2",
"Rpt_Block_Num" : "12700",
"Rpt_CRIS_Cnty_ID" : "101",
"Rpt_City_ID" : "244",
"Rpt_Hwy_Num" : "225",
"Rpt_Outside_City_Limit_Fl" : "0",
"Rpt_Rdwy_Sys_ID" : "3",
"Rpt_Road_Part_ID" : "5",
"Rpt_Sec_Block_Num" : "",
"Rpt_Sec_Hwy_Num" : "146",
"Rpt_Sec_Rdwy_Sys_ID" : "3",
"Rpt_Sec_Road_Part_ID" : "1",
"Rpt_Sec_Street_Name" : "SH 146",
"Rpt_Street_Name" : "SH 225",
"Rr_Relat_Fl" : "0",
"Rural_Fl" : "0",
"Schl_Bus_Fl" : "0",
"Surf_Cond_ID" : "1",
"Surf_Type_ID" : "5",
"Sus_Serious_Injry_Cnt" : "0",
"Thousand_Damage_Fl" : "1",
"Toll_Road_Fl" : "0",
"Tot_Injry_Cnt" : "0",
"Traffic_Cntl_ID" : "20",
"Unkn_Injry_Cnt" : "0",
"Wthr_Cond_ID" : "5",
"CompositeIndex" : "19.6749358426",
"CompositeSum" : "230",
"NormalizedIndex" : "4.88126649077",
"people" : [
    {
        "@context" : "http://schema.org/",
        "@id" : "15575237",
        "@type" : "Person",
        "Crash_ID" : "15575237",
        "Death_Cnt" : "0",
        "Drvr_Drg_Cat_1_ID" : "97.0",
        "Drvr_Lic_Cls_ID" : "3.0",
        "Drvr_Lic_State_ID" : "43.0",
        "Drvr_Lic_Type_ID" : "9.0",
        "Drvr_Zip" : "77058",
        "Non_Injry_Cnt" : "1",
        "Nonincap_Injry_Cnt" : "0",
        "Poss_Injry_Cnt" : "0",
        "Prsn_Age" : "54.0",
        "Prsn_Airbag_ID" : "3.0",
        "Prsn_Alc_Rslt_ID" : "",
        "Prsn_Alc_Spec_Type_ID" : "4.0",
        "Prsn_Bac_Test_Rslt" : "",
        "Prsn_Death_Time" : "",
        "Prsn_Drg_Rslt_ID" : "97.0",
        "Prsn_Drg_Spec_Type_ID" : "4.0",
        "Prsn_Ejct_ID" : "1.0",
        "Prsn_Ethnicity_ID" : "1.0",
        "Prsn_Gndr_ID" : "2.0",
        "Prsn_Helmet_ID" : "97.0",
        "Prsn_Injry_Sev_ID" : "5.0",
        "Prsn_Nbr" : "1",
        "Prsn_Occpnt_Pos_ID" : "1.0",
        "Prsn_Rest_ID" : "1.0",
        "Prsn_Sol_Fl" : "N",
        "Prsn_Type_ID" : "1",
        "Sus_Serious_Injry_Cnt" : "0",
        "Tot_Injry_Cnt" : "0",
        "Unit_Nbr" : "1",
        "Unkn_Injry_Cnt" : "0"
    }
]
}
```


Figure 4.7 Traffic crash data for traffic crash event; Crash_ID: 15575237

The values from a traffic crash identified by its Crash_ID: 15575237 was a minor crash. The crash described by its JSON representation in Figure 4.7 occurred in La Porte, TX which is just outside Houston, TX, shown in Figure 4.8. This crash occurred on January 2, 2017, at 3:04 am.

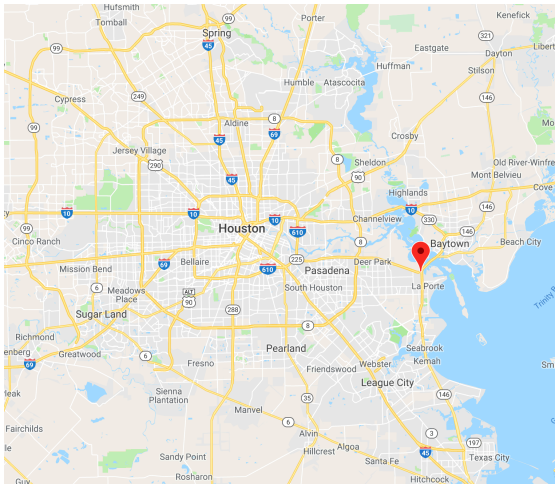


Figure 4.8 Geographic location of Crash_ID: 15575237

This crash was computed to be a minor crash. This crash was computed to be minor based on the criteria values presented and the weights in weighted sample one. It is intuitively clear that this crash is minor because there were no fatalities or injuries reported; furthermore, the time of day has minimal impact on others on the road. Though the crash is minor, weather conditions specifies that it was foggy, thus increasing the CCI of the crash. Through the understanding of the investigation, external circumstances, and the effect that the crash has on people, the CCI was determined to be 19.67 (minor). Furthermore, this crash was worse than 4.88% of all crashes in the entire data set.

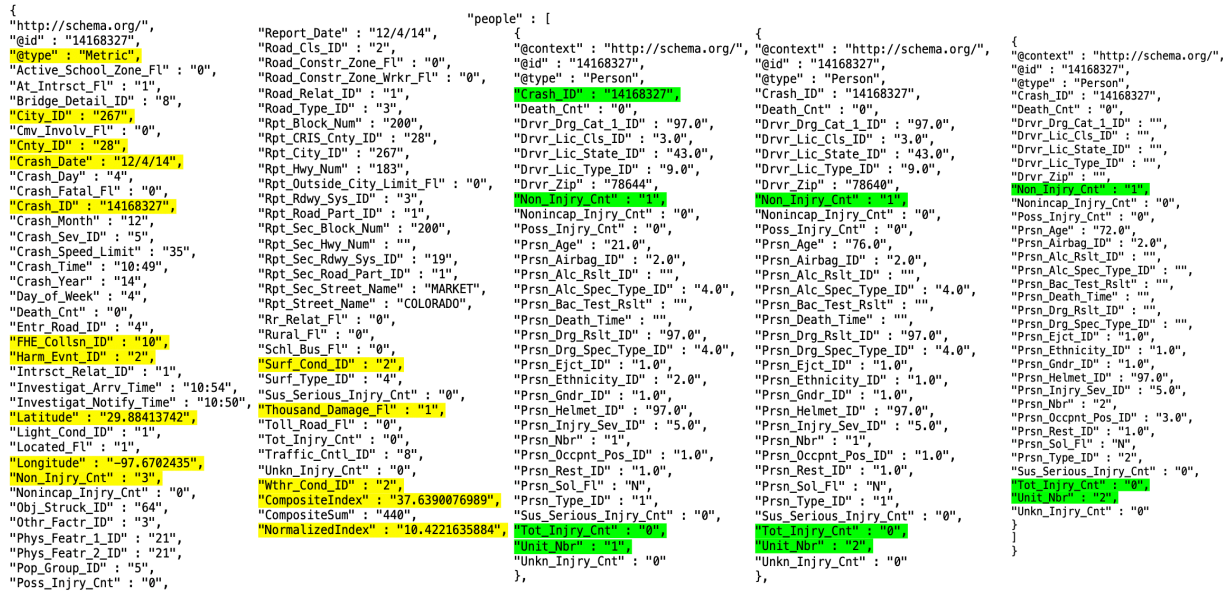
Table 4.2 CCI comparison table of a traffic crash (Crash_ID: 15575237) for four different weighted samples

Critical Composite Index Comparison	
Weight Sample One **{** **"@context" : "http://schema.org/",** **"@id" : "15575237",** **"@type" : "Metric",** **"CompositeIndex" : 19.6749358426,** **"CompositeSum" : 230,** **"Crash_ID" : "15575237",** **"NormalizedIndex" : 4.88126649077** **}**	**Weight Sample Two** **{** **"@context" : "http://schema.org/",** **"@id" : "15575237",** **"@type" : "Metric",** **"CompositeIndex" : 62.7227722772,** **"CompositeSum" : 1267,** **"Crash_ID" : "15575237",** **"NormalizedIndex" : 0.63731509092** **}**

Table 4.2 compares the two different sample weighted value tables for this work. Since the severity and importance of various factors are considered to be subjective, it is necessary to understand how additional sample weighted value tables describe the same crash. The values in weight sample two are test values to show the difference of using values that have significantly higher weight.

4.4.2. Traffic Crash Case Two – Moderate

In this traffic crash case, a ‘moderate’ crash is examined , based off of the CCI described by Figure 3.18.



```
{
"http://schema.org/",
"@id" : "14168327",
"@type" : "Metric",
"Active_School_Zone_Fl" : "0",
"At_Intrsct_Fl" : "1",
"Bridge_Detail_ID" : "8",
"City_ID" : "267",
"Cmv_Involv_Fl" : "0",
"Cnty_ID" : "28",
"Crash_Date" : "12/4/14",
"Crash_Day" : "4",
"Crash_Fatal_Fl" : "0",
"Crash_ID" : "14168327",
"Crash_Month" : "12",
"Crash_Sev_ID" : "5",
"Crash_Speed_Limit" : "35",
"Crash_Time" : "10:49",
"Crash_Year" : "14",
"Day_of_Week" : "4",
"Death_Cnt" : "0",
"Entr_Road_ID" : "4",
"FHE_Collsn_ID" : "10",
"Harm_Evnt_ID" : "2",
"Intrsct_Relat_ID" : "1",
"Investigat_Arrv_Time" : "10:54",
"Investigat_Notify_Time" : "10:50",
"Latitude" : "29.88413742",
"Light_Cond_ID" : "1",
"Located_Fl" : "1",
"Longitude" : "-97.6702435",
"Non_Injry_Cnt" : "3",
"Nonincap_Injry_Cnt" : "0",
"Obj_Struck_ID" : "64",
"Othr_Factr_ID" : "3",
"Phys_Featr_1_ID" : "21",
"Phys_Featr_2_ID" : "21",
"Pop_Group_ID" : "5",
"Poss_Injry_Cnt" : "0",
"Report_Date" : "12/4/14",
"Road_Cls_ID" : "2",
"Road_Constr_Zone_Fl" : "0",
"Road_Constr_Zone_Wrkr_Fl" : "0",
"Road_Relat_ID" : "1",
"Road_Type_ID" : "3",
"Rpt_Block_Num" : "200",
"Rpt_CRIS_Cnty_ID" : "28",
"Rpt_City_ID" : "267",
"Rpt_Hwy_Num" : "183",
"Rpt_Outside_City_Limit_Fl" : "0",
"Rpt_Rdwy_Sys_ID" : "3",
"Rpt_Road_Part_ID" : "1",
"Rpt_Sec_Block_Num" : "200",
"Rpt_Sec_Hwy_Num" : "",
"Rpt_Sec_Rdwy_Sys_ID" : "19",
"Rpt_Sec_Road_Part_ID" : "1",
"Rpt_Sec_Street_Name" : "MARKET",
"Rpt_Street_Name" : "COLORADO",
"Rr_Relat_Fl" : "0",
"Rural_Fl" : "0",
"Schl_Bus_Fl" : "0",
"Surf_Cond_ID" : "2",
"Surf_Type_ID" : "4",
"Sus_Serious_Injry_Cnt" : "0",
"Thousand_Damage_Fl" : "1",
"Toll_Road_Fl" : "0",
"Tot_Injry_Cnt" : "0",
"Traffic_Cntl_ID" : "8",
"Unkn_Injry_Cnt" : "0",
"Wthr_Cond_ID" : "2",
"CompositeIndex" : "37.6390076989",
"CompositeSum" : "440",
"NormalizedIndex" : "10.4221635884",
"people" : [
{
"@context" : "http://schema.org/",
"@id" : "14168327",
"@type" : "Person",
"Crash_ID" : "14168327",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "3.0",
"Drvr_Lic_State_ID" : "43.0",
"Drvr_Lic_Type_ID" : "9.0",
"Drvr_Zip" : "78644",
"Non_Injry_Cnt" : "1",
"Nonincap_Injry_Cnt" : "0",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "21.0",
"Prsn_Airbag_ID" : "2.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "2.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "5.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "1.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "1",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "0",
"Unit_Nbr" : "1",
"Unkn_Injry_Cnt" : "0"
},
{
"@context" : "http://schema.org/",
"@id" : "14168327",
"@type" : "Person",
"Crash_ID" : "14168327",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "3.0",
"Drvr_Lic_State_ID" : "43.0",
"Drvr_Lic_Type_ID" : "9.0",
"Drvr_Zip" : "78640",
"Non_Injry_Cnt" : "1",
"Nonincap_Injry_Cnt" : "0",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "76.0",
"Prsn_Airbag_ID" : "2.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "1.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "5.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "1.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "1",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "0",
"Unit_Nbr" : "2",
"Unkn_Injry_Cnt" : "0"
},
{
"@context" : "http://schema.org/",
"@id" : "14168327",
"@type" : "Person",
"Crash_ID" : "14168327",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "",
"Drvr_Lic_Cls_ID" : "",
"Drvr_Lic_State_ID" : "",
"Drvr_Lic_Type_ID" : "",
"Drvr_Zip" : "",
"Non_Injry_Cnt" : "1",
"Nonincap_Injry_Cnt" : "0",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "72.0",
"Prsn_Airbag_ID" : "2.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "",
"Prsn_Drg_Spec_Type_ID" : "",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "1.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "5.0",
"Prsn_Nbr" : "2",
"Prsn_Occpnt_Pos_ID" : "3.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "2",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "0",
"Unit_Nbr" : "2",
"Unkn_Injry_Cnt" : "0"
}
]
}
```


Figure 4.9 Traffic crash data for traffic crash event; Crash_ID: 14168327

The traffic crash identified by its Crash_ID: 14168327 was a moderate crash. The crash described by its JSON representation in Figure 4.9 occurred in Lockhart, TX which is just outside Austin, TX, shown in Figure 4.10. This crash occurred on December 4, 2015, at 10:49 am.

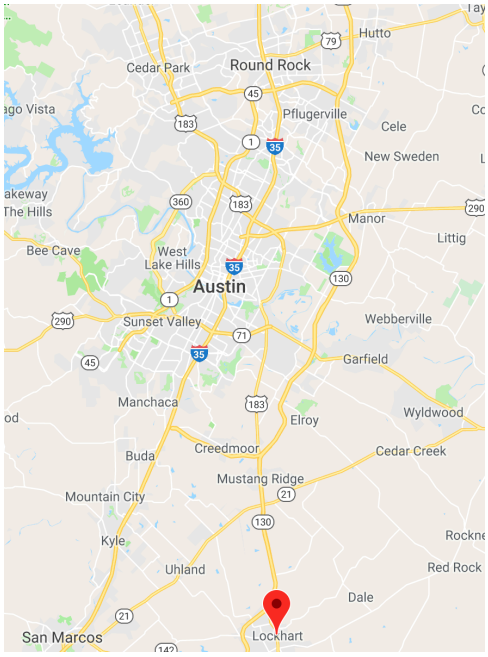


Figure 4.10 Geographic location of Crash_ID: 14168327

This crash was computed to be a moderate crash. This crash was computed to be moderate based on the criteria values presented and the weights it held. It is intuitively clear that this crash is moderate as well as matches the CCI because there were no fatalities or injuries reported, but the surface conditions were wet and a vehicle collided with another vehicle in motion; furthermore, the time of day has an impact on others on the road. Moreover, the crash is computed to be moderate because weather condition specifies that it was wet, thus increasing the CCI of the crash. Through the understanding of the investigation, external circumstances, and the effect that the crash has on people, the CCI was determined to be 37.63 (moderate). Furthermore, this crash was worse than 10.42% of all crashes in the entire data set.

Table 4.3 CCI comparison table of a traffic crash (Crash_ID: 14168327) for four different weighted samples

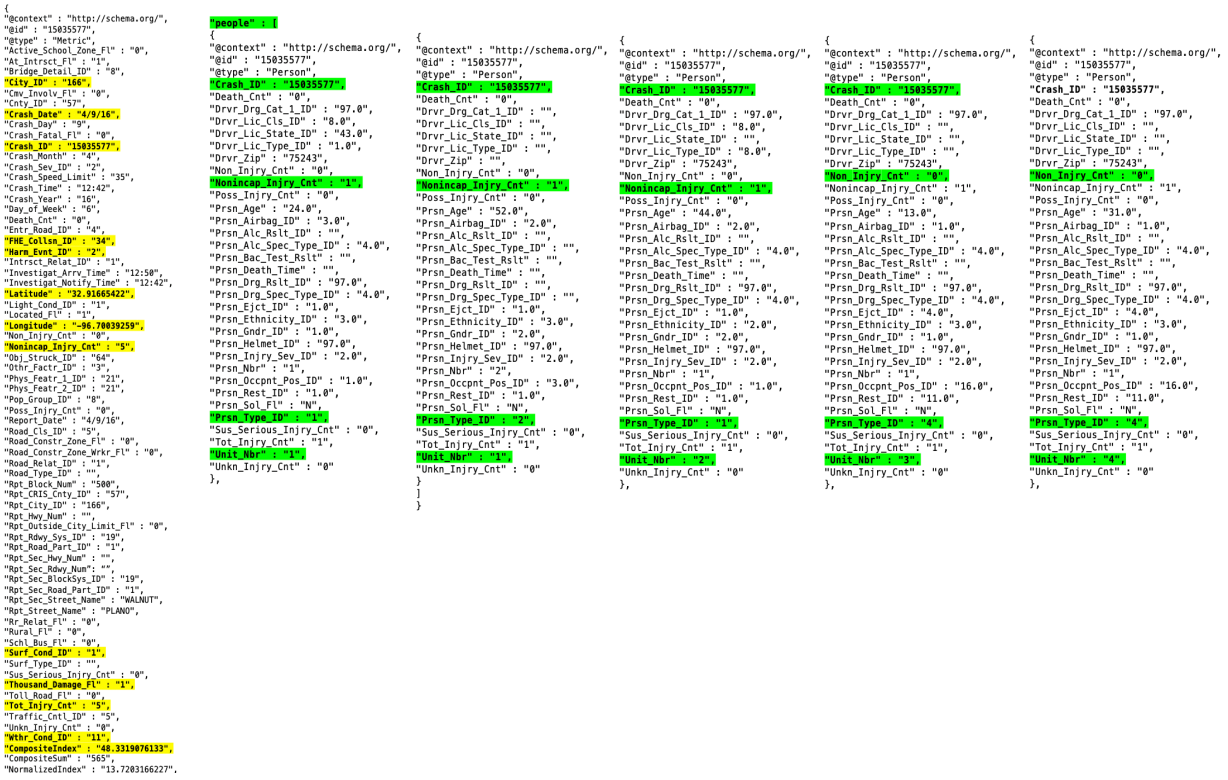
Critical Composite Index Comparison	

Weight Sample One
{
"@context" : "http://schema.org/",
"@id" : "14168327",
"@type" : "Metric",
"CompositeIndex" : 37.6390076989,
"CompositeSum" : 440,
"Crash_ID" : "14168327",
"NormalizedIndex" : 10.4221635884
}
	Weight Sample Two
	{
	"@context" : "http://schema.org/",
	"@id" : "14168327",
	"@type" : "Metric",
	"CompositeIndex" : 82.1287128713,
	"CompositeSum" : 1659,
	"Crash_ID" : "14168327",
	"NormalizedIndex" : 2.92931064725
	}

Table 4.3 compares the two different sample weighted value tables for this work. Since the severity and importance of various factors are considered to be subjective, it is necessary to understand how additional sample weighted value tables describe the same crash. The values in weight sample two are test values to show the difference of using values that have significantly higher weight.

4.4.3. Traffic Crash Case Three – Major

In this traffic crash case, a ‘major’ crash is examined, based off of the CCI described by Figure 3.18.



{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Metric",
"Active_School_Zone_Fl" : "0",
"At_Intrsct_Fl" : "1",
"Bridge_Detail_ID" : "8",
"City_ID" : "166",
"Cmv_Involv_Fl" : "0",
"Cnty_ID" : "57",
"Crash_Date" : "4/9/16",
"Crash_Day" : "9",
"Crash_Fatal_Fl" : "0",
"Crash_ID" : "15035577",
"Crash_Month" : "4",
"Crash_Sev_ID" : "2",
"Crash_Speed_Limit" : "35",
"Crash_Time" : "12:42",
"Crash_Year" : "16",
"Day_of_Week" : "6",
"Death_Cnt" : "0",
"Entr_Road_ID" : "4",
"FHE_Collsn_ID" : "34",
"Harm_Evnt_ID" : "2",
"Intrsct_Relat_ID" : "1",
"Investigat_Arrv_Time" : "12:50",
"Investigat_Notify_Time" : "12:42",
"Latitude" : "32.91665422",
"Light_Cond_ID" : "1",
"Located_Fl" : "1",
"Longitude" : "-96.70039259",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "5",
"Obj_Struck_ID" : "64",
"Othr_Factr_ID" : "3",
"Phys_Featr_1_ID" : "21",
"Phys_Featr_2_ID" : "21",
"Pop_Group_ID" : "8",
"Poss_Injry_Cnt" : "0",
"Report_Date" : "4/9/16",
"Road_Cls_ID" : "5",
"Road_Constr_Zone_Fl" : "0",
"Road_Constr_Zone_Wrkr_Fl" : "0",
"Road_Relat_ID" : "1",
"Road_Type_ID" : "",
"Rpt_Block_Num" : "500",
"Rpt_CRIS_Cnty_ID" : "57",
"Rpt_City_ID" : "166",
"Rpt_Hwy_Num" : "",
"Rpt_Outside_City_Limit_Fl" : "0",
"Rpt_Rdwy_Sys_ID" : "19",
"Rpt_Road_Part_ID" : "1",
"Rpt_Sec_Hwy_Num" : "",
"Rpt_Sec_Rdwy_Num": "",
"Rpt_Sec_BlockSys_ID" : "19",
"Rpt_Sec_Road_Part_ID" : "1",
"Rpt_Sec_Street_Name" : "WALNUT",
"Rpt_Street_Name" : "PLANO",
"Rr_Relat_Fl" : "0",
"Rural_Fl" : "0",
"Schl_Bus_Fl" : "0",
"Surf_Cond_ID" : "1",
"Surf_Type_ID" : "",
"Sus_Serious_Injry_Cnt" : "0",
"Thousand_Damage_Fl" : "1",
"Toll_Road_Fl" : "0",
"Tot_Injry_Cnt" : "5",
"Traffic_Cntl_ID" : "5",
"Unkn_Injry_Cnt" : "0",
"Wthr_Cond_ID" : "11",
"CompositeIndex" : "48.3319076133",
"CompositeSum" : "565",
"NormalizedIndex" : "13.7203166227".
"people" : [
{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Person",
"Crash_ID" : "15035577",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "8.0",
"Drvr_Lic_State_ID" : "43.0",
"Drvr_Lic_Type_ID" : "1.0",
"Drvr_Zip" : "75243",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "1",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "24.0",
"Prsn_Airbag_ID" : "3.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "3.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "2.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "1.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "1",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "1",
"Unit_Nbr" : "1",
"Unkn_Injry_Cnt" : "0"
},
{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Person",
"Crash_ID" : "15035577",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "",
"Drvr_Lic_Cls_ID" : "",
"Drvr_Lic_State_ID" : "",
"Drvr_Lic_Type_ID" : "",
"Drvr_Zip" : "",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "1",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "52.0",
"Prsn_Airbag_ID" : "2.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "",
"Prsn_Drg_Spec_Type_ID" : "",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "3.0",
"Prsn_Gndr_ID" : "2.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "2.0",
"Prsn_Nbr" : "2",
"Prsn_Occpnt_Pos_ID" : "3.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "2",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "1",
"Unit_Nbr" : "1",
"Unkn_Injry_Cnt" : "0"
}
]
}
{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Person",
"Crash_ID" : "15035577",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "8.0",
"Drvr_Lic_State_ID" : "",
"Drvr_Lic_Type_ID" : "8.0",
"Drvr_Zip" : "75243",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "1",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "44.0",
"Prsn_Airbag_ID" : "2.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "1.0",
"Prsn_Ethnicity_ID" : "2.0",
"Prsn_Gndr_ID" : "2.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "2.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "1.0",
"Prsn_Rest_ID" : "1.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "1",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "1",
"Unit_Nbr" : "2",
"Unkn_Injry_Cnt" : "0"
},
{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Person",
"Crash_ID" : "15035577",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "",
"Drvr_Lic_State_ID" : "",
"Drvr_Lic_Type_ID" : "",
"Drvr_Zip" : "75243",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "1",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "13.0",
"Prsn_Airbag_ID" : "1.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "4.0",
"Prsn_Ethnicity_ID" : "3.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "2.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "16.0",
"Prsn_Rest_ID" : "11.0",
"Prsn_Sol_Fl" : "N",
"Prsn_Type_ID" : "4",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "1",
"Unit_Nbr" : "3",
"Unkn_Injry_Cnt" : "0"
},
{
"@context" : "http://schema.org/",
"@id" : "15035577",
"@type" : "Person",
"Crash_ID" : "15035577",
"Death_Cnt" : "0",
"Drvr_Drg_Cat_1_ID" : "97.0",
"Drvr_Lic_Cls_ID" : "",
"Drvr_Lic_State_ID" : "",
"Drvr_Lic_Type_ID" : "",
"Drvr_Zip" : "75243",
"Non_Injry_Cnt" : "0",
"Nonincap_Injry_Cnt" : "1",
"Poss_Injry_Cnt" : "0",
"Prsn_Age" : "31.0",
"Prsn_Airbag_ID" : "1.0",
"Prsn_Alc_Rslt_ID" : "",
"Prsn_Alc_Spec_Type_ID" : "4.0",
"Prsn_Bac_Test_Rslt" : "",
"Prsn_Death_Time" : "",
"Prsn_Drg_Rslt_ID" : "97.0",
"Prsn_Drg_Spec_Type_ID" : "4.0",
"Prsn_Ejct_ID" : "4.0",
"Prsn_Ethnicity_ID" : "3.0",
"Prsn_Gndr_ID" : "1.0",
"Prsn_Helmet_ID" : "97.0",
"Prsn_Injry_Sev_ID" : "2.0",
"Prsn_Nbr" : "1",
"Prsn_Occpnt_Pos_ID" : "16.0",
"Prsn_Rest_ID" : "11.0",
"Prsn_Sol_Fl" : "NN",
"Prsn_Type_ID" : "4",
"Sus_Serious_Injry_Cnt" : "0",
"Tot_Injry_Cnt" : "1",
"Unit_Nbr" : "4",
"Unkn_Injry_Cnt" : "0"
},

Figure 4.11 Traffic crash data for traffic crash event; Crash_ID: 15035577

The traffic crash identified by its Crash_ID: 15035577 was a major crash. The crash described by its JSON representation in Figure 4.11 occurred in Garland, TX which is just outside Dallas, TX, shown in Figure 4.12. This crash occurred on April 9, 2016, at 12:42 pm.

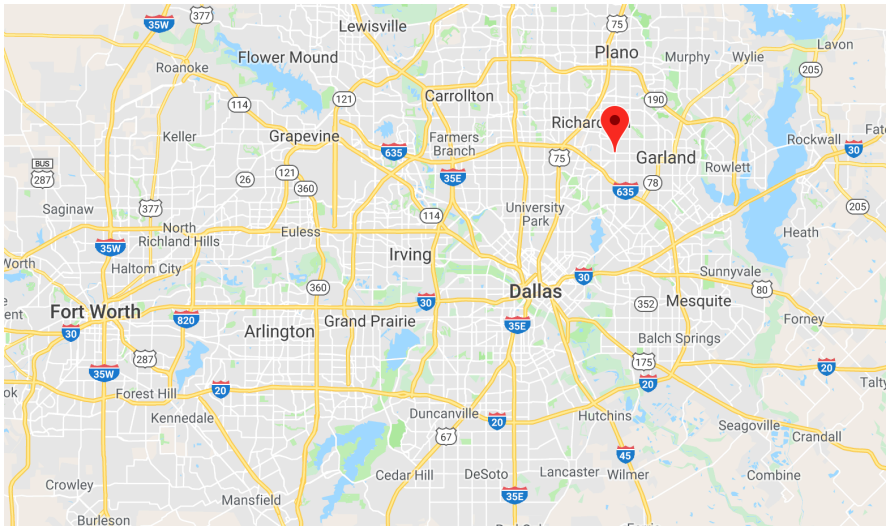


Figure 4.12 Geographic location of Crash_ID: 15035577

This crash was computed to be a major crash. This crash was computed to be major based on the criteria values presented and the weights it held. It is intuitively clear that this crash is major as well as matches the CCI because there were five non-incapacitating injuries without fatalities. The surface conditions were dry, and at least one vehicle collided with another vehicle in motion; there were a total of four vehicles involved in a crash. Furthermore, the time of day has an impact on others on the road. Although the weather conditions were clear, the crash is computed to be major because of the number of non-incapacitating injuries associated with the crash, thus increasing the CCI of the crash. Through the understanding of the investigation, external circumstances, and the effect that the crash has on people, the CCI was determined to be 48.33 (major). Furthermore, this crash was worse than 13.72% of all crashes in the entire data set.

Table 4.4 CCI comparison table of a traffic crash (Crash_ID: 15035577) for four different weighted samples

Critical Composite Index Comparison	
Weight Sample One	**Weight Sample Two**
{	{
"@context" : "http://schema.org/",	"@context" : "http://schema.org/",
"@id" : "15035577",	"@id" : "15035577",
"@type" : "Metric",	"@type" : "Metric",
"CompositeIndex" : 48.3319076133,	"CompositeIndex" : 90.4455445545,
"CompositeSum" : 565,	"CompositeSum" : 1827,
"Crash_ID" : "15035577",	"Crash_ID" : "15035577",
"NormalizedIndex" : 13.7203166227	"NormalizedIndex" : 3.91159445711
}	}

Table 4.4 compares the two different sample weighted value tables for this work. Since the severity and importance of various factors are considered to be subjective, it is necessary to understand how additional sample weighted value tables describe the same crash. The values in weight sample two are test values to show the difference of using values that have significantly higher weight.

4.4.4. Traffic Crash Case Four – Severe

In this traffic crash case, a 'severe' crash is examined , based off of the CCI described by Figure 3.18.

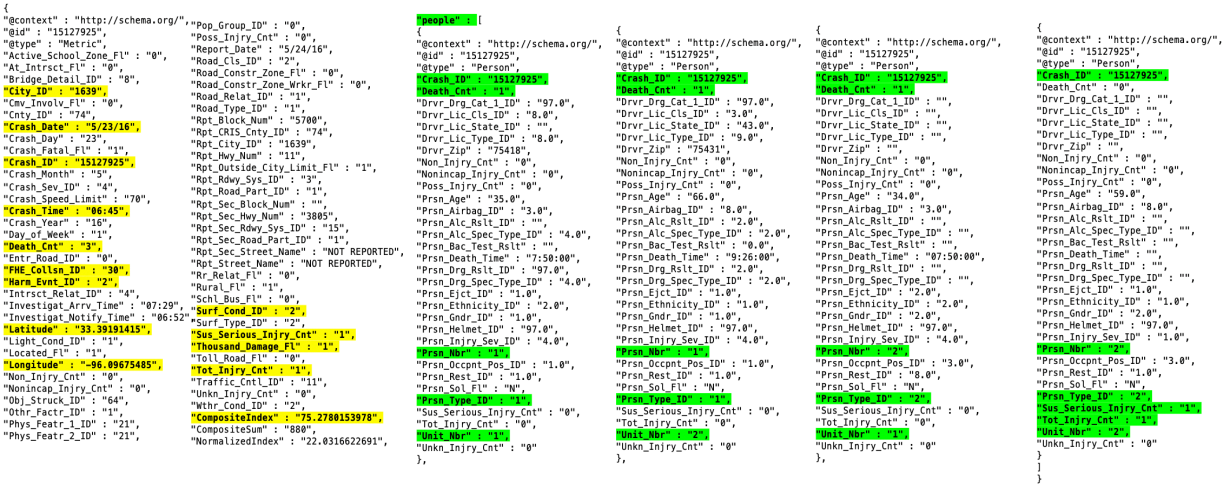


Figure 4.13 Traffic crash data for traffic crash event; Crash_ID: 15127925

The traffic crash identified by its Crash_ID: 15127925 was a severe crash. The crash described by its JSON representation in Figure 4.13 occurred in Rural Fannin County, which is northeast of Dallas, TX, shown in Figure 4.14. This crash occurred on May 23, 2016, at 6:45 am.

Figure 4.14 Geographic location of Crash_ID: 15127925

This crash was computed to be a severe crash. This crash was computed to be severe based on the criteria values presented and the weights it held. It is intuitively clear that this crash is severe as well as matches the CCI because there were three fatalities and one serious injury sustained as a result of the crash. The surface conditions were wet, and at least one vehicle collided with another vehicle in motion; there were a total of two vehicles involved in the crash. Furthermore, the time of day has an impact on others on the road since it was in the morning. The crash occurred during rainy weather conditions. Moreover, the crash is computed to be severe because of the fatalities and serious injury associated with the crash, thus increasing the CCI of the crash. Through the understanding of the investigation, external circumstances, and the effect that the

crash has on people, the CCI was determined to be 75.27 (severe). Furthermore, this crash was worse than 22.03% of all crashes in the entire data set.

Table 4.5 CCI comparison table of a traffic crash (Crash_ID: 15127925) for four different weighted samples

Critical Composite Index Comparison	
Weight Sample One	Weight Sample Two
{	{
"@context" : "http://schema.org/",	"@context" : "http://schema.org/",
"@id" : "15127925",	"@id" : "15127925",
"@type" : "Metric",	"@type" : "Metric",
"CompositeIndex" : 75.2780153978,	"CompositeIndex" : 92.2772277228,
"CompositeSum" : 880,	"CompositeSum" : 1864,
"Crash_ID" : "15127925",	"Crash_ID" : "15127925",
"NormalizedIndex" : 22.0316622691	"NormalizedIndex" : 4.12793077238
}	}

Table 4.5 compares the two different sample weighted value tables for this work. Since the severity and importance of various factors are considered to be subjective, it is necessary to understand how additional sample weighted value tables describe the same crash. The values in weight sample two are test values to show the difference of using values that have significantly higher weight.

4.5 Data Views From Multiple Perspectives – Narratives

Based on the work of Gil and Garijo, (Gil & Garijo, 2017) data narratives are "containers of information about computationally generated research findings… A set of narrative accounts that are automatically generated to be human consumable renderings of the record and entities." This work explores narratives based on two major human perspectives: non-domain experts and domain experts. The narratives of both perspectives are similar, as it expresses a majority of the same information, however domain-expert narratives contain additional information that may not be useful for a non-domain expert. The approach used to develop the narratives follows the work

of Gil and Garijo in that it is outputted by computational research findings for human consumption.

Figure 4.15 shows the three templates that were used to create the data narratives as part of this research. The four different data narratives for each of the four case studies (minor, moderate, major and severe) have three different perspectives. The first, full list form, is a complete list view which may be appropriate for researchers and domain experts. This amount of information presented can provide quick lookup information without having to read large amounts of text as well as presents all of the data as part of the narrative. The second, readable expert, provides the same information as the full list form, however it is in paragraph form. Paragraph form allows for experts to give reports about traffic crashes in complete sentences. The third, readable non-domain expert provides a subset of information from the expert in human readable paragraph text. Paragraph text provides non-domain experts to read about a traffic crash without all of the information that may be unnecessary to them. The narratives in this research were designed based on the six elements of narratives described by Gil and Garijo, 1) Human-readable, 2) Customizable, 3) Persistent, 4) Accurate, 5) Inspectable, and 6) Publishable (Gil & Garijo, 2017). The narratives in this research meet the standards expressed as they provide human-readability, customization based on different perspectives, persistent in that the narrative is backed by evidence (i.e., raw data), accurate by means of representing the actual information in the data, inspectable in that it can be used to understand and make analysis (i.e., through queries in Mongo DB) on the CCI, and publishable insofar as it can be used as a sample for future work (i.e. they can be included in reports).

All of the templates have variables in *italic* that are representative of the values that are dynamically adjusted based on each individual traffic crash. These narratives are based on the query of multiple traffic crashes. The amount of information provided in each data narrative represents the different perspectives of users who may use traffic crash data. The amount of information given in each data narrative is representative of a possible example of the necessary information for the users who will use the traffic information, it is not the only templates that can be used; each of the data narratives may additionally be adjusted to fit the needs of a given user.

Full List Form Template

"The following traffic crash occurred.\n"
"The traffic crash has a Crash_ID of " + *crash_id* +".\n"
"The crash occurred in " + city + " located in "+ county +" county Texas on "+*crash_date*+" at "+ *crash_time* + ". \n"
"The coordinates of the crash are: (" +*lat*+","+*long*+").\n"
"The crash occurred on a "+ *road_type* +" roadway.\n"
"The crash has a Critical Composite Index of: "+ *composite_index* +".\n"
"Was the crash fatal: " + *crash_fatal* +".\n"
"The crash involved a harmful event including, but not limited to hitting a: "+ *harm_event* +".\n"
"The vehicle " + *obj_struck* + ".\n"
"The light conditions were reported as: " + *light* + ".\n"
"The weather conditions were reported to be " + *weather* +".\n"
"The surface conditions were reported to be " + *surface* + ".\n"
"Was the crash in a construction zone: " + *construction_zone* + ".\n"
"Was the crash around construction workers "+ *construction_worker* + ".\n"
"Was the crash in an active school zone: " + *active_school_zone* + ".\n"
"Was the crash involve a school bus: " + *bus* + ".\n"
"Was the crash at an intersection: " + *intersection* + ".\n"
"Was the crash at railroad: " + *rail* + ".\n"
"Was a CMV involved: " + *cmv* + ".\n"
"Did the crash involve at least $1000 in damages: "+ *damage* + ".\n"
"The crash had a total of: " + *death_count* + " deaths.\n"
"The crash had a total of: " + *serious_count* + " serious injuries.\n"
"The crash had a total of: " + *nonincap_count* + " non-incapacitating injuries.\n"
"The crash had a total of: " + *no_injury_count* + " non-injuries.\n"
"The crash had a total of: " + *unknown_count* + " unknown injuries.\n"

Readable Expert Template

"The following traffic crash occurred. \n"

"The traffic crash has a Crash_ID of " + *crash_id* +". The crash occurred in " + *city* + " located in " + *county* +" county Texas on "+ *crash_date*+" at "+ *crash_time* + ". \n"
"The coordinates of the crash are (" +*lat*+","+*long*+"). The crash occurred on a "+ *road_type* +" roadway. The crash has a Critical Composite Index of "+*composite_index*+" which is classified as a " + *cci_string*+" crash. \n"
"The crash was " + *crash_fatal* +"fatal. The crash involved a first harmful event when a vehicle hit a "+*harm_event*+". The lighting was " + *light* + ", "the weather conditions were " + *weather* +", and the road surface conditions were " + *surface* + ".\n"
"The crash was " + *construction_zone* + "in a construction zone and there were "+ *construction_worker* + " construction workers around.\n"
"The crash was " + *active_school_zone* + "in an active school zone. A school bus was " + bus + " involved in the traffic crash. \n"
"The crash " + *intersection* + ". The crash " + *rail* + ".\n"
"The crash " + *cmv* + ". The crash caused "+*damage* + ".\n"
"The crash had a total of " + *death_count* + " deaths, "+ *serious_count* + " serious injuries, "+ *nonincap_count* + " non-incapacitating injuries, " + *no_injury_count* + " non-injuries, and " + *unknown_count* + " unknown injuries.\n"

Readable Non-Domain Expert
"The following traffic crash occurred. \n"
"The traffic crash has a Crash_ID of " + *crash_id* +". The crash occurred in " + *city* + " located in " + *county* +" county Texas on "+ *crash_date*+" at "+ *crash_time* + ". \n"
The crash has a Critical Composite Index of "+*composite_index*+" which is classified as a " + *cci_string*+" crash. \n"
"The crash was " + *crash_fatal* +"fatal. The lighting was " + *light* + ", "the weather conditions were " + *weather* +", and the road surface conditions were " + *surface* + ".\n"
"The crash was " + *active_school_zone* + "in an active school zone. A school bus was " + bus + " involved in the traffic crash. \n"
"The crash " + *intersection* + ". The crash " + *rail* + ".\n"
"The crash " + *cmv* + ". The crash caused "+*damage* + ".\n"
"The crash had a total of " + *death_count* + " deaths, "+ *serious_count* + " serious injuries, "+ *nonincap_count* + " non-incapacitating injuries, " + *no_injury_count* + " non-injuries, and " + *unknown_count* + " unknown injuries.\n"

Figure 4.15 Data narrative template

The narratives presented in Figures 4.16-4.19 show how the consumption of raw data can be used to transform into human understandable information beyond the CCI. The approach used however is open for expansion to introduce more definite rules for data to be viewed in a structured text.

Figure 4.16 describes the data narrative for the traffic crash: Crash_ID: 15575237. This narrative was programmatically created based on template and specific data from the crash to provide human readable text about the traffic crash.

```
Full List Form
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15575237.
The crash occurred in LA PORTE located in Harris county Texas on
1/2/17 at 03:04.
The coordinates of the crash are: (29.68822175,-95.03169053).
The crash occurred on a 4 OR MORE LANES, DIVIDED roadway.
The crash has a Critical Composite Index of: 19.6749358426.
Was the crash fatal: NO.
The crash involved a harmful event including, but not limited to
hitting a: FIXED OBJECT.
The vehicle HIT MEDIAN BARRIER.
The light conditions were reported as: DARK, LIGHTED.
The weather conditions were reported to be FOG.
The surface conditions were reported to be DRY.
Was the crash in a construction zone: NO.
Was the crash around construction workers NO.
Was the crash in an active school zone: NO.
Was the crash involve a school bus: NO.
Was the crash at an intersection: NO.
Was the crash at railroad: NO.
Was a CMV involved: NO.
Did the crash involve at least $1000 in damages: YES.
The crash had a total of: 0 deaths.
The crash had a total of: 0 serious injuries.
The crash had a total of: 0 non-incapacitating injuries.
The crash had a total of: 1 non-injuries.
The crash had a total of: 0 unknown injuries.
The crash had alcohol as a possible factor: NO.


Readable Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15575237. The crash occurred in
LA PORTE located in Harris county Texas on 1/2/17 at 03:04.
The coordinates of the crash are (29.68822175,-95.03169053). The
crash occurred on a 4 OR MORE LANES, DIVIDED roadway. The crash has
a Critical Composite Index of 19.6749358426 which is classified as a
severe crash. Alcohol is not expected to be a possible factor in
this crash.
The crash was not fatal. The crash involved a first harmful event
when a vehicle hit a FIXED OBJECT. The lighting was DARK, LIGHTED,
```

```
the weather conditions were FOG, and the road surface conditions
were DRY.
The crash was not in a construction zone and there were no
construction workers around.
The crash was not in an active school zone. A school bus was
involved in the traffic crash.
The crash did not occur in an intersection. The crash did not occur
at a railroad crossing.
The crash  did not involve a commercial vehicle. The crash caused
at least $1000 in damages.
The crash had a total of 0 deaths, 0 serious injuries, 0 non-
incapacitating injuries, 1 non-injuries, and 0 unknown injuries.


Readable Non-Domain Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15575237. The crash occurred in
LA PORTE located in Harris county Texas on 1/2/17 at 03:04.
The crash has a Critical Composite Index of 19.6749358426 which is
classified as a severe crash. Alcohol is not expected to be a
possible factor in this crash.
The crash was not fatal. The lighting was DARK, LIGHTED, the weather
conditions were FOG, and the road surface conditions were DRY.
The crash was not in a construction zone and there were no
construction workers around.
The crash was not in an active school zone. A school bus was
involved in the traffic crash.
The crash did not occur in an intersection. The crash did not occur
at a railroad crossing.
The crash had a total of 0 deaths, 0 serious injuries, 0 non-
incapacitating injuries, 1 non-injuries, and 0 unknown injuries.
```

Figure 4.16 Data narrative of Crash_ID: 15575237

Figure 4.17 describes the data narrative for the traffic crash: Crash_ID: 14168327. This narrative was programmatically retrieved from the raw data to provide human readable text about the traffic crash.

```
Full List Form
The following traffic crash occurred.
The traffic crash has a Crash_ID of 14168327.
The crash occurred in LOCKHART located in Caldwell county Texas on
12/4/14 at 10:49.
```

The coordinates of the crash are: (29.88413742,-97.6702435).
The crash occurred on a 4 OR MORE LANES, UNDIVIDED roadway.
The crash has a Critical Composite Index of: 37.6390076989.
Was the crash fatal: NO.
The crash involved a harmful event including, but not limited to hitting a: MOTOR VEHICLE IN TRANSPORT.
The vehicle NOT APPLICABLE.
The light conditions were reported as: DAYLIGHT.
The weather conditions were reported to be RAIN.
The surface conditions were reported to be WET.
Was the crash in a construction zone: NO.
Was the crash around construction workers NO.
Was the crash in an active school zone: NO.
Was the crash involve a school bus: NO.
Was the crash at an intersection: YES.
Was the crash at railroad: NO.
Was a CMV involved: NO.
Did the crash involve at least $1000 in damages: YES.
The crash had a total of: 0 deaths.
The crash had a total of: 0 serious injuries.
The crash had a total of: 0 non-incapacitating injuries.
The crash had a total of: 3 non-injuries.
The crash had a total of: 0 unknown injuries.
The crash had alcohol as a possible factor: NO.

Readable Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 14168327. The crash occurred in LOCKHART located in Caldwell county Texas on 12/4/14 at 10:49.
The coordinates of the crash are (29.88413742,-97.6702435). The crash occurred on a 4 OR MORE LANES, UNDIVIDED roadway. The crash has a Critical Composite Index of 37.6390076989 which is classified as a moderate crash. Alcohol is not expected to be a possible factor in this crash.
The crash was not fatal. The crash involved a first harmful event when a vehicle hit a MOTOR VEHICLE IN TRANSPORT. The lighting was DAYLIGHT, the weather conditions were RAIN, and the road surface conditions were WET.
The crash was not in a construction zone and there were no construction workers around.
The crash was not in an active school zone. A school bus was involved in the traffic crash.
The crash occurred in an intersection. The crash did not occur at a railroad crossing.
The crash did not involve a commercial vehicle. The crash caused at least $1000 in damages.
The crash had a total of 0 deaths, 0 serious injuries, 0 non-incapacitating injuries, 3 non-injuries, and 0 unknown injuries.

```
Readable Non-Domain Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 14168327. The crash occurred in
LOCKHART located in Caldwell county Texas on 12/4/14 at 10:49.
The crash has a Critical Composite Index of 37.6390076989 which is
classified as a moderate crash. Alcohol is not expected to be a
possible factor in this crash.
The crash was not fatal. The lighting was DAYLIGHT, the weather
conditions were RAIN, and the road surface conditions were WET.
The crash was not in a construction zone and there were no
construction workers around.
The crash was not in an active school zone. A school bus was
involved in the traffic crash.
The crash occurred in an intersection. The crash did not occur at a
railroad crossing.
The crash had a total of 0 deaths, 0 serious injuries, 0 non-
incapacitating injuries, 3 non-injuries, and 0 unknown injuries.
```

Figure 4.17 Data narrative of Crash_ID: 14168327

Figure 4.18 describes the data narrative for the traffic crash: Crash_ID: 15035577. This narrative was programmatically retrieved from the raw data to provide human readable text about the traffic crash.

```
Full List Form
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15035577.
The crash occurred in GARLAND located in Dallas county Texas on
4/9/16 at 12:42.
The coordinates of the crash are: (32.91665422,-96.70039259).
The crash occurred on an Unknown roadway.
The crash has a Critical Composite Index of: 48.3319076133.
Was the crash fatal: NO.
The crash involved a harmful event including, but not limited to
hitting a: MOTOR VEHICLE IN TRANSPORT.
The vehicle NOT APPLICABLE.
The light conditions were reported as: DAYLIGHT.
The weather conditions were reported to be CLEAR.
The surface conditions were reported to be DRY.
Was the crash in a construction zone: NO.
Was the crash around construction workers NO.
Was the crash in an active school zone: NO.
```

Was the crash involve a school bus: NO.
Was the crash at an intersection: YES.
Was the crash at railroad: NO.
Was a CMV involved: NO.
Did the crash involve at least $1000 in damages: YES.
The crash had a total of: 0 deaths.
The crash had a total of: 0 serious injuries.
The crash had a total of: 5 non-incapacitating injuries.
The crash had a total of: 0 non-injuries.
The crash had a total of: 0 unknown injuries.
The crash had alcohol as a possible factor: NO.

Readable Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15035577. The crash occurred in GARLAND located in Dallas county Texas on 4/9/16 at 12:42.
The coordinates of the crash are (32.91665422,-96.70039259). The crash occurred on an Unknown roadway. The crash has a Critical Composite Index of 48.3319076133 which is classified as a major crash. Alcohol is not expected to be a possible factor in this crash.
The crash was not fatal. The crash involved a first harmful event when a vehicle hit a MOTOR VEHICLE IN TRANSPORT. The lighting was DAYLIGHT, the weather conditions were CLEAR, and the road surface conditions were DRY.
The crash was not in a construction zone and there were no construction workers around.
The crash was not in an active school zone. A school bus was involved in the traffic crash.
The crash occurred in an intersection. The crash did not occur at a railroad crossing.
The crash did not involve a commercial vehicle. The crash caused at least $1000 in damages.
The crash had a total of 0 deaths, 0 serious injuries, 5 non-incapacitating injuries, 0 non-injuries, and 0 unknown injuries.

Readable Non-Domain Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15035577. The crash occurred in GARLAND located in Dallas county Texas on 4/9/16 at 12:42.
The crash has a Critical Composite Index of 48.3319076133 which is classified as a major crash. Alcohol is not expected to be a possible factor in this crash.
The crash was not fatal. The lighting was DAYLIGHT, the weather conditions were CLEAR, and the road surface conditions were DRY.
The crash was not in a construction zone and there were no construction workers around.

```
The crash was not in an active school zone. A school bus was
involved in the traffic crash.
The crash occurred in an intersection. The crash did not occur at a
railroad crossing.
The crash had a total of 0 deaths, 0 serious injuries, 5 non-
incapacitating injuries, 0 non-injuries, and 0 unknown injuries.
```

Figure 4.18 Data narrative of Crash_ID: 15035577

Figure 4.19 describes the data narrative for the traffic crash: Crash_ID: 15127925. This narrative was programmatically retrieved from the raw data to provide human readable text about the traffic crash.

```
Full List Form
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15127925.
The crash occurred in RURAL FANNIN COUNTY located in Fannin county
Texas on 5/23/16 at 06:45.
The coordinates of the crash are: (33.39191415,-96.09675485).
The crash occurred on a 2 LANE, 2 WAY roadway.
The crash has a Critical Composite Index of: 75.2780153978.
Was the crash fatal: YES.
The crash involved a harmful event including, but not limited to
hitting a: MOTOR VEHICLE IN TRANSPORT.
The vehicle NOT APPLICABLE.
The light conditions were reported as: DAYLIGHT.
The weather conditions were reported to be RAIN.
The surface conditions were reported to be WET.
Was the crash in a construction zone: NO.
Was the crash around construction workers NO.
Was the crash in an active school zone: NO.
Was the crash involve a school bus: NO.
Was the crash at an intersection: NO.
Was the crash at railroad: NO.
Was a CMV involved: NO.
Did the crash involve at least $1000 in damages: YES.
The crash had a total of: 3 deaths.
The crash had a total of: 1 serious injuries.
The crash had a total of: 0 non-incapacitating injuries.
The crash had a total of: 0 non-injuries.
The crash had a total of: 0 unknown injuries.
The crash had alcohol as a possible factor: NO.
```

Readable Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15127925. The crash occurred in RURAL FANNIN COUNTY located in Fannin county Texas on 5/23/16 at 06:45.
The coordinates of the crash are (33.39191415,-96.09675485). The crash occurred on a 2 LANE, 2 WAY roadway. The crash has a Critical Composite Index of 75.2780153978 which is classified as a severe crash. Alcohol is not expected to be a possible factor in this crash.
The crash was fatal. The crash involved a first harmful event when a vehicle hit a MOTOR VEHICLE IN TRANSPORT. The lighting was DAYLIGHT, the weather conditions were RAIN, and the road surface conditions were WET.
The crash was not in a construction zone and there were no construction workers around.
The crash was not in an active school zone. A school bus was involved in the traffic crash.
The crash did not occur in an intersection. The crash did not occur at a railroad crossing.
The crash did not involve a commercial vehicle. The crash caused at least $1000 in damages.
The crash had a total of 3 deaths, 1 serious injury, 0 non-incapacitating injuries, 0 non-injuries, and 0 unknown injuries.

Readable Non-Domain Expert
The following traffic crash occurred.
The traffic crash has a Crash_ID of 15127925. The crash occurred in RURAL FANNIN COUNTY located in Fannin county Texas on 5/23/16 at 06:45.
The crash has a Critical Composite Index of 75.2780153978 which is classified as a severe crash. Alcohol is not expected to be a possible factor in this crash.
The crash was fatal. The lighting was DAYLIGHT, the weather conditions were RAIN, and the road surface conditions were WET.
The crash was not in a construction zone and there were no construction workers around.
The crash was not in an active school zone. A school bus was involved in the traffic crash.
The crash did not occur in an intersection. The crash did not occur at a railroad crossing.
The crash had a total of 3 deaths, 1 serious injury, 0 non-incapacitating injuries, 0 non-injuries, and 0 unknown injuries.

Figure 4.19 Data narrative of Crash_ID: 15127925

Chapter 5: Evaluation

This chapter will discuss the evaluation of the results. The research conducted through this work is highly interdisciplinary, with a focus on Computer Science; having both core and applied research practices. The core portion of this research is the BUM methodology that was developed whereas the applied research portion is the outcome of the data to knowledge transformation, the Critical Composite Metric. Applied research evaluation practices are supported by relevant literature in Computer Science and interdisciplinary research methods by Julie Klein (Klein, 2008), Aaron Sloman (Sloman, 2016), Bertrand Meyer et al. (Meyer, Choppy, Staunstrup, & van Leeuwen, 2009), and the Research Evaluation Committee of Informatics Europe (Europe, 2008). This work adheres to the premise that Computer Science is a discipline that combines areas from sciences and mathematics to improve theoretical and applied research to make an impact on people (Europe, 2008; Klein, 2008).

The key element of this research describes the significance that a bottom-up model can be formed through bottom-up techniques. Furthermore, this research explores the way that the development of interdisciplinary advancements and standards can be improved to link knowledge across domains through data science. Through readily available data, information can be gathered, modeled, and mapped for improved modularity and access. This research serves as a model to show the importance of linked data for an interdisciplinary view, within the context of Smart Mobility; moreover, create effective impact on the people who use this interdisciplinary work (Klein, 2008). Furthermore, this research shows the importance of defined methodologies to improve personalized, contextualized access to linked data and how it can be useful for the improvement of everyday life. The access stems from the information that non-domain and

domain-experts can gain from the information without having to study the specifics of the traffic crash, unless necessary.

The BUM methodology follows a similar approach to a domain ontology design as described by Noy and McGuinness in Ontology 101 (Noy & McGuinness, 2001). In their work, they describe a seven step ontology design guideline: 1) Determine the domain and scope of the ontology, 2) Consider reusing existing ontologies, 3) Enumerate important terms in the ontology, 4) Define the classes and hierarchy, 5) Define the properties of the classes, 6) Define the facets, and 7) Create instances (Noy & McGuinness, 2001). The BUM methodology uses a five-step process which also includes determining and analyzing the domain. Step 1 in the BUM methodology is comparable to step 1 of ontology 101, both processes aim to identify the domain and scope of the data model. Step 2 of the BUM methodology introduces data discovery as a necessary technique to use data to form the knowledge graph, this step is not explicitly listed as part of Ontology 101. Step 3 of the BUM methodology covers steps 2-6 in Ontology 101 and is the most similar process given that this step is data modeling and Ontology 101 provides data modeling guidelines. However, the BUM methodology is designed to be generic in that it does not always have to have a formal ontology description, but instead uses other domain areas as a way to improve the data model. In this work, an ontology is used as part of the design and is mapped, but it is not always necessary if semantics are not required for a particular domain. Step 4 of the BUM methodology is similar to step 7 of Ontology 101 in the creation of data instances in ontologies, insofar as the BUM methodology introduces data points into the knowledge graph. Step 5 of the BUM methodology extends Ontology 101 by providing guidelines on creating the CCI metric. The metric that is developed provides extensibility for users to understand the data without the need or expertise of understanding the full semantic representation of such data.

5.1 INDEX DEVELOPMENT

The CCI was developed based on the definition of an index by Earl Babbie (Babbie, 2012), "[An index] is a type of composite measure that summarizes and rank-orders several specific observations and represents some more-general dimension." An index itself does not define a single traffic crash, but instead conceptualizes the real-world event into a value that is representative of a real-occurrence. By introducing an index into a real-world scenario, such as traffic crashes, you introduce the "process of specifying observations and measurements that give concepts definite meaning for the research purpose" (Babbie, 2012). An index is inherently a general way to understand occurrences of some event in a relevant way.

The CCI was developed on the basis of being reliable and valid; both of which are standard index evaluation descriptions. First, reliability is based on simply observing the same results or values multiple times (Babbie, 2012); this was accomplished by ensuring that the data is the same for each index computation and the methodology for the index value computation was the same. Secondly, validity refers to having a numerical value that measures a the real meaning of a concept (Babbie, 2012), which is clear in the CCI. Since index values are composite measurements of a concept, it is necessary to choose the proper variables that will contribute to the value; the four major criteria for selecting variables are face validity, unidimensional, general or specific, and variance (Babbie, 2012). Face validity ensures that the variable makes logical sense; unidimensional ensures that the metric measures a single concept; general or specific ensures that the metric measures a concept in either a general or specific way – the CCI is a general measurement; and variance ensures that the metric describes a large variety of events such as traffic crashes. The CCI was developed based on the central idea of measuring traffic crashes with respect to the crash, people, and external circumstances.

5.2 Case Study Evaluation

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Crash
> 20 – 40	Moderate Crash
> 40 – 50	Major Crash
> 50 +	Severe Crash

The above chart, shown previously as Figure 3.20, is used to evaluate traffic crashes using the CCI. The CCI is a composite measurement of traffic crashes with respect to the crash event, the people involved, and the external circumstances of the crash.

Traffic Crash Case One – Minor

Crash_ID: 15575237 occurred in La Porte, TX which is just outside Houston, TX on January 2, 2017, at 3:04 am. The CCI for this traffic Collision is: 19.67; this value describes that the studied crash is a minor crash because the CCI is between 0 and 20. By further exploring this crash it is shown that the CCI is intuitively representative of the crash because there was one vehicle with one person involved without any reported injuries.

Traffic Crash Case Two – Moderate

Crash_ID: 14168327 occurred in Lockhart, TX which is just outside Austin, TX, on December 4, 2015, at 10:49 am. The CCI for this traffic Collision is: 37.63; this value describes that the studied crash is a moderate crash because the CCI is greater than 20 and less than 40. By further exploring this crash it is shown that the CCI is intuitively representative of the crash because

there were two vehicles with damage and three people involved in the crash with no injuries reported.

Traffic Crash Case Three – Major

Crash_ID: 15035577 occurred in Garland, TX which is just outside Dallas, TX on April 9, 2016, at 12:42 pm. The CCI for this traffic Collision is: 48.33; this value describes that the studied crash is a major crash because the CCI is greater than 40 and less than 50. By further exploring this crash it is shown that the CCI is intuitively representative of the crash because three vehicles were involved and damaged in the traffic crash. Moreover, the five people involved in the crash suffered non-incapacitating injuries; however, there were no reported fatalities.

Traffic Crash Case Four – Severe

Crash_ID: 15127925 occurred in Rural Fannin County, which is northeast of Dallas, TX on May 23, 2016, at 6:45 am. The CCI for this traffic Collision is: 75.27; this value describes that the studied crash is an extremely major crash because the CCI is greater than 50. By further exploring this crash it is shown that the CCI is intuitively representative of the crash because two vehicles were involved and damaged in the traffic crash. Moreover, four people were involved in the traffic crash; three of the people involved are deceased and one other sustained serious injury.

For each of the case studies done, a CCI was computed to describe. By manually inspecting the CCI severity value and comparing it with the information for each one of the cases, the CCI describes each of the four traffic crashes appropriately based on the injuries, fatalities, environment, and conditions. The CCI gives insight to the traffic crash without the need to

explore each traffic crash on a data-level and provides a means to classify crashes in an objective way.

As a compliment to the CCI, competency questions can be used to describe the specific circumstances of the traffic crash; additionally, data narratives can be developed to provide additional insight. Moreover, by mapping the CCI onto a geographic map, predictions can be made about future traffic crashes while considering geographic locations and conditions. By introducing a standard way to describe traffic crashes non-domain experts, domain experts, and policy makers can begin to introduce changes into roadways for improvement.

5.3 User Evaluation Study

The *User Evaluation Study* was used to evaluate the CCI and the CCI severity chart; a user evaluation study consists of a sample group of individuals evaluating and commenting on a particular domain for individual perspectives that can be interpreted (Jones, Baxter, & Khanduja, 2013). The evaluation study was done using an online based survey of 107 subjects (100 non-domain experts and 7 domain experts) at least 18 years old; each of the subjects self-reported themselves to be non-domain or domain experts. Since there were not enough domain expert participants in this study to show clear comparison between them and non-domain experts, each subject is treated equally; the domain expert specific questions were removed from the survey report in this section, however, they are available in the appendix. The user evaluation study is intended to capture the increase in knowledge for people (Sloman, 2016).

Domain experts were defined as:

- Individuals with extensive knowledge of traffic crash reporting

- Traffic Investigator
- Traffic Police Officer
- Traffic Engineer
- Other traffic professional

Non-domain experts were defined as:

- Commuter without extensive knowledge of traffic crash reporting

Gathering information from both non-domain and domain experts is necessary for understanding the different points of view that stems from the knowledge that a person may have. The research survey was done according to the techniques described by Jones, Baxter, & Khanduja (Jones et al., 2013) and Krosnick (Krosnick, 1999). Jones, Baxter, & Khanduja first describe two major important features to a successful survey, aesthetics and question order(Jones et al., 2013); moreover, Krosnick (Krosnick, 1999) also makes reference to question order being critical in a successful survey.

Firstly, aesthetic is important to continuously attract subjects to take and complete the survey (Jones et al., 2013). In this research, the user evaluation study was conducted using an online survey software, QuestionPro ("QuestionPro," 2019); this software was made available through The University of Texas at El Paso (UTEP). As a result of using QuestionPro, the aesthetics was predetermined to fit the image set by UTEP.

Secondly, question order was determined by introducing questions that were easier and quicker to answer (Jones et al., 2013; Krosnick, 1999). Moreover, the questions were separated into

sections that correspond to the focuses that the CCI would be evaluated against. The CCI and its severity chart were evaluated based on four focuses:

1. Comprehension of CCI
2. Knowledge Gain & Perception
3. Improvement from current metric reporting
4. CCI Improvements

Based on the four evaluation focuses, high-level questions were developed to determine what information should be gathered and how this can improve the state-of-the-art in Computer Science. The questions are separated in the following way:

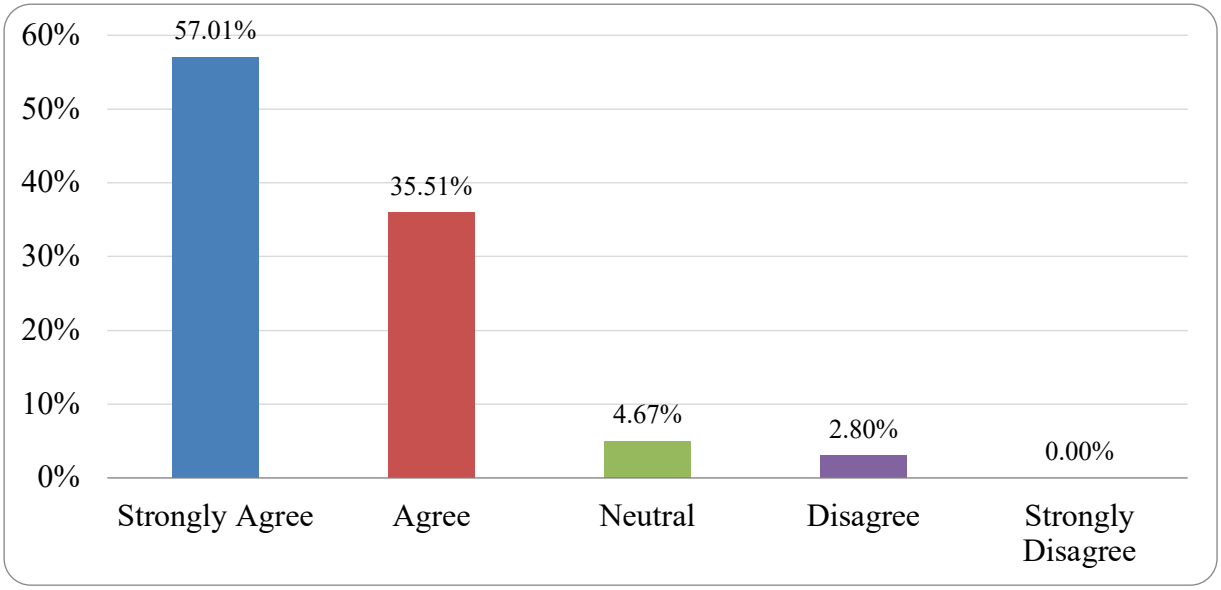
- Informed Consent & Descriptions (Appendix) – Questions: 1-4
- Expert Specific (Appendix) – Questions: 5-7
- Comprehension of CCI – Questions: 8, 9, 12, 13
- Knowledge Gain & Perceptions – Questions: 10, 11, 14 – 17
- Improvement from current metric reporting – Question: 18
- CCI Improvements – Questions: 19 – 21

5.3.1. Comprehension of CCI

The Critical Composite Index for this traffic Crash is: **15.82 - Minor Crash**.

- The data shown describes the following:
- Crash occurred at 1:17 pm
- At least $1000 of damage
- Weather conditions were reported clear
- Light conditions were reported as daylight
- The road conditions were reported as being dry.
- The vehicle hit a fixed object
- One (1) person was involved without any reported injuries or fatalities

Do you agree that the classification of **Minor Crash** accurately represents this traffic crash?



	Answer	Count	Percent
1.	Strongly Agree	61	57.01%
2.	Agree	38	35.51%
3.	Neutral	5	4.67%
4.	Disagree	3	2.80%
5.	Strongly Disagree	0	0.00%
	Total	107	100%

Mean: **1.533**	Confidence Interval @ 95%: **[1.397 - 1.669]**	Standard Deviation: **0.718**	Standard Error: **0.069**

Figure 5.1 Survey Question & Results 8

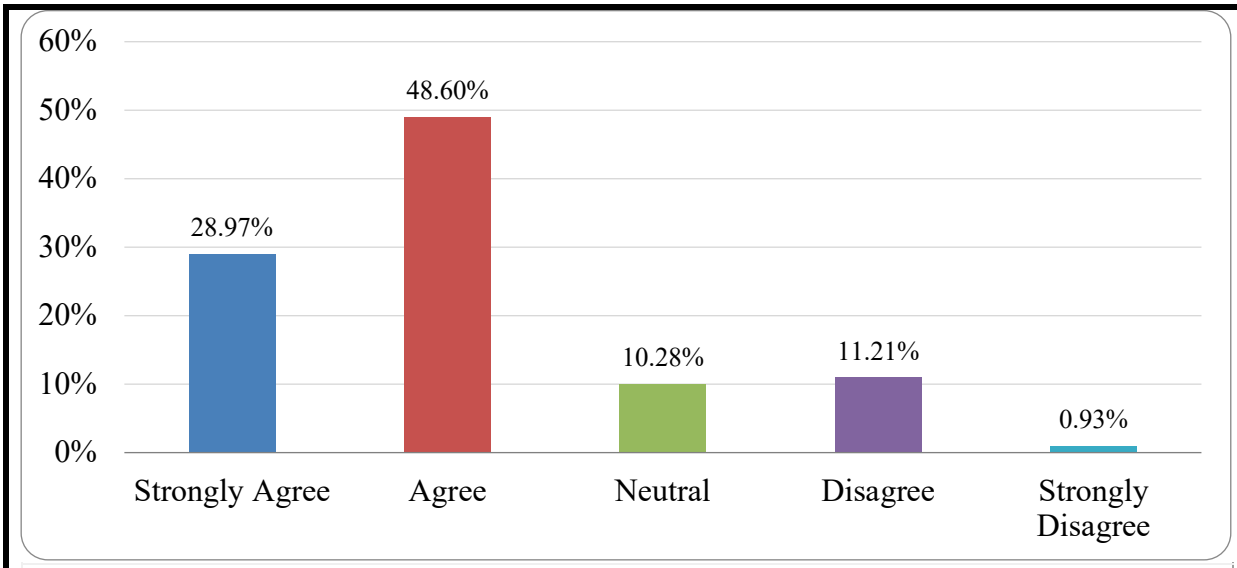
The results shown in Figure 5.1 are based on the question, “Do you agree that the classification of Minor Crash accurately represents this traffic crash?” This question is intended to provide insight to user comprehension of the CCI, its classification and the supporting severity charts. Based on the results, 57.01% of all respondents strongly agree that the described traffic crash is minor and an additional 35.51% agree that the described traffic crash is minor. 92.52% of all respondents agree or strongly agree that the traffic crash described is minor. Furthermore, based on the sampled results, the mean is 1.533 which is between strongly agree and agree; with a 95% confidence that the true population mean would strongly agree or agree that the described traffic crash is minor. The results from this question show a majority of users understand this traffic crash to accurately represented as a minor crash.

The above crash computed a Critical Composite Index of: **27.80 - Moderate**

The data shown describes the following:

- Crash occurred at 1:28 pm
- At least $1000 of damage
- Light conditions were reported as daylight
- Weather conditions were reported to be clear
- The road conditions were reported as being dry
- A vehicle hit a moving vehicle
- Six (6) people were involved in the traffic crash
- One (1) person involved was seriously injured
- Five (5) people did not sustain any injuries

Do you agree that the classification of **Moderate Crash** accurately represents this traffic crash?



	Answer	Count	Percent
1.	Strongly Agree	31	28.97%
2.	Agree	52	48.60%
3.	Neutral	11	10.28%
4.	Disagree	12	11.21%
5.	Strongly Disagree	1	0.93%
	Total	107	100%

Mean: **2.065**	Confidence Interval @ 95%: **[1.883 - 2.248]**	Standard Deviation: **0.964**	Standard Error: **0.093**

Figure 5.2 Survey Question & Results 9

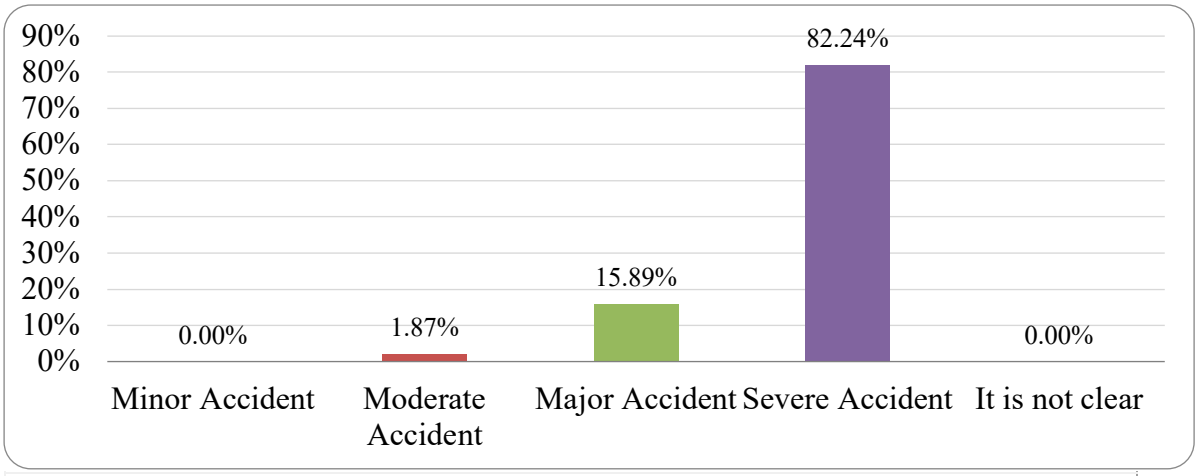
The results shown in Figure 5.2 are based on the question, "Do you agree that the classification of Moderate Crash accurately represents this traffic crash?" This question is intended to provide insight to user comprehension of the CCI, its classification and the supporting severity charts. Based on the results, 28.97% of all respondents strongly agree that the described traffic crash is moderate and an additional 48.60% agree that the described traffic crash is moderate. 77.57% of all respondents agree or strongly agree that the traffic crash described is moderate. Furthermore, based on the sampled results, the mean is 2.065 which is classified to be agree; with a 95% confidence that the true population mean would agree that the described traffic crash is

moderate. The results from this question a majority of roadway users understand this traffic crash to accurately represented as a moderate crash.

The data shown describes the following:

- Crash occurred at 6:45 am
- At least $1000 of damage
- The traffic crash was reported to occur in daylight
- The weather was reported to be raining
- The road conditions were reported as being wet
- A vehicle hit another moving vehicle
- Four (4) people were involved in the traffic crash
- Three (3) of the people involved are deceased as a result of the crash (Fatal Crash)
- One (1) person sustained serious injuries

Based on your understanding of the Critical Composite Index and traffic crash knowledge; What severity would you use to best classify this traffic crash?



	Answer	Count	Percent
1.	Minor Crash	0	0.00%
2.	Moderate Crash	2	1.87%
3.	Major Crash	17	15.89%
4.	Severe Crash	88	82.24%
5.	It is not clear	0	0.00%
	Total	107	100%

Mean: **3.804** Confidence Interval @ 95%: **[3.720 - 3.888]** Standard Deviation: **0.444** Standard Error: **0.043**

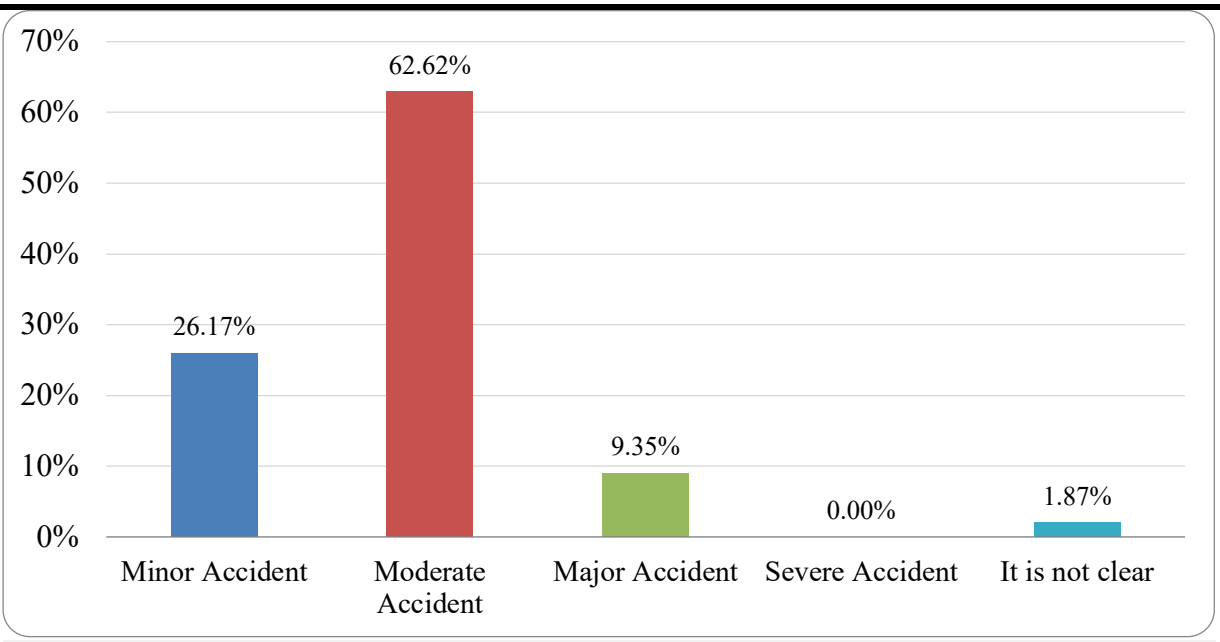
Figure 5.3 Survey Question & Results 12

The results shown in Figure 5.3 are based on the question, “Based on your understanding of the Critical Composite Index and traffic crash knowledge; What severity would you use to best classify this traffic crash?” This question is intended to provide insight to user comprehension of the CCI, its classification and the supporting severity charts. Based on the results, 82.24% of all respondents classified this traffic crash as a severe crash. Based on the CCI of 75.27, the severity of the crash is considered to be a severe crash. Furthermore, based on the sampled results, the mean is 3.804 which is between a major crash and a severe crash; with a 95% confidence that the true population mean would classify the described traffic crash as nearly a severe crash; this result matches the result of the CCI. The results from this question a majority of roadway users accurately classify traffic crashes in a similar way that the CCI would.

The data shown describes the following:

- Crash occurred at 12:42 pm
- At least $1000 of damage
- The traffic crash was reported to occur in daylight
- The road conditions were reported as dry
- A vehicle hit another moving vehicle
- Five (5) people were involved, all of whom suffered non-incapacitating injuries

Based on your understanding of the Critical Composite Index and traffic crash knowledge; What severity would you use to best classify this traffic crash?



	Answer	Count	Percent
1.	Minor Crash	28	26.17%
2.	Moderate Crash	67	62.62%
3.	Major Crash	10	9.35%
4.	Severe Crash	0	0.00%
5.	It is not clear	2	1.87%
	Total	107	100%

Mean: **1.888**	Confidence Interval @ 95%: **[1.752 - 2.024]**	Standard Deviation: **0.718**	Standard Error: **0.069**

Figure 5.4 Survey Question & Results 13

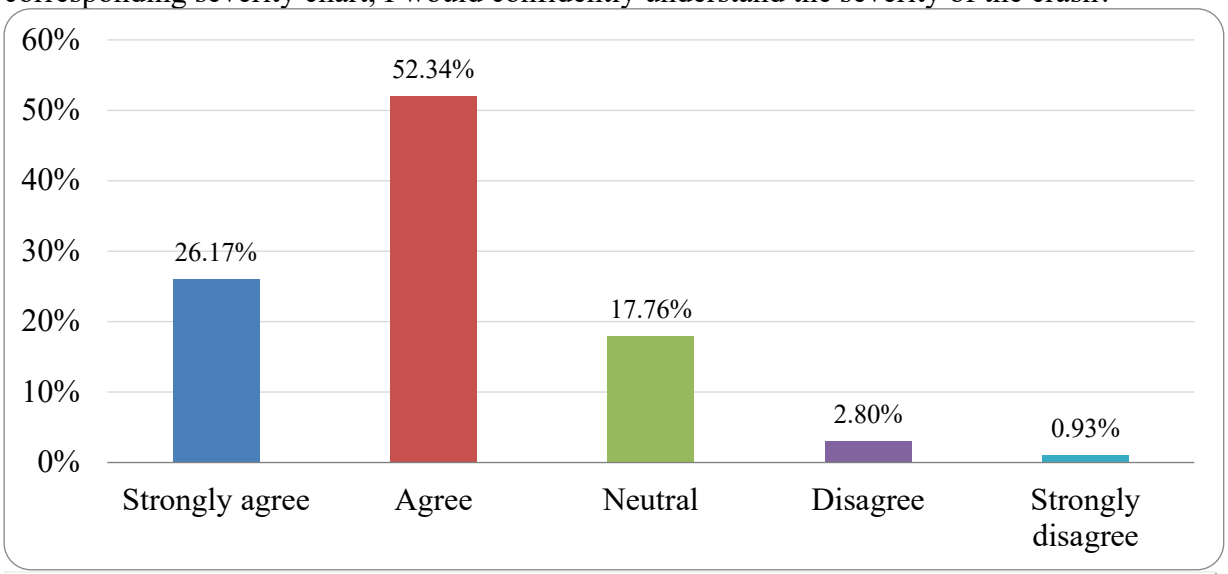
The results shown in Figure 5.4 are based on the question, “Based on your understanding of the Critical Composite Index and traffic crash knowledge; What severity would you use to best classify this traffic crash?” This question is intended to provide insight to user comprehension of the CCI, its classification and the supporting severity charts. Based on the results, 62.62% of all respondents classified this traffic crash a moderate crash. Furthermore, based on the sampled results, the mean is 1.888 approximately a moderate crash; with a 95% confidence that the true population mean would classify the described traffic crash as nearly a moderate crash. Based on

the CCI of 48.33, the severity of the crash is considered to be a major crash. The results from the question and the CCI are not matches. Since the results of the survey do not match the CCI, it can be deduced that the number of persons involved in the traffic crash, even with injuries do not reflect an increased severity for participants of the survey compared to the significance provided by the weighted scale. However, based on the case studies described in Chapter 4, this traffic crash would be considered moderate using weighted sample three. Though these results are not a match, additional information can be gained on how to improve the weighting scale of non-incapacitating injuries.

Based on the results shown in Figures 5.1-5.4, survey participants are clearly able to comprehend the CCI without direct knowledge of how it is computed. This comprehension stems from a combination of experience of traffic crashes and the representation of the CCI based on sample traffic crash descriptions.

5.3.2. Knowledge Gain & Perception

If a traffic crash were to be reported and given a Critical Composite Index value & corresponding severity chart, I would confidently understand the severity of the crash?



	Answer	Count	Percent
1.	Strongly agree	28	26.17%
2.	Agree	56	52.34%
3.	Neutral	19	17.76%
4.	Disagree	3	2.80%
5.	Strongly disagree	1	0.93%
	Total	107	100%

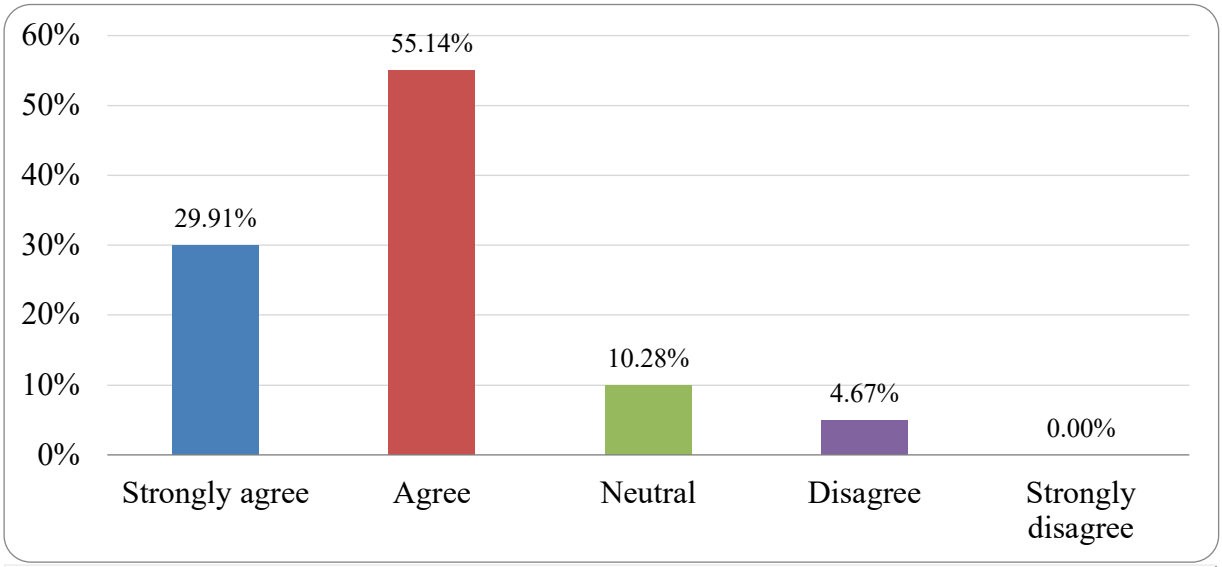
Mean: **2.000** Confidence Interval @ 95%: **[1.848 - 2.152]** Standard Deviation: **0.801** Standard Error: **0.077**

Figure 5.5 Survey Question & Results 10

The results shown in Figure 5.5 are based on the question, “If a traffic crash were to be reported and given a Critical Composite Index value & corresponding severity chart, I would confidently understand the severity of the crash?” This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the supporting severity charts. Based on the results, 52.34% of all respondents agree that they would confidently understand the severity of a traffic crash if given a CCI and severity chart. An additional 26.17% strongly agree

that they would understand the severity of a traffic crash if given a CCI and severity chart. 78.51% of all respondents agree or strongly agree that they could confidently understand the severity of a crash if given a CCI and severity chart. Furthermore, based on the sampled results, the mean is 2.000 which is exactly equal to agree; with a 95% confidence that the true population mean would strongly agree – agree that they would understand a traffic crash if given a CCI and severity chart.

Do you agree that the Critical Composite Index & Severity Classification (Minor, Moderate, Major, Severe) is an appropriate representation of traffic crashes - considering the applicable details of the crash?



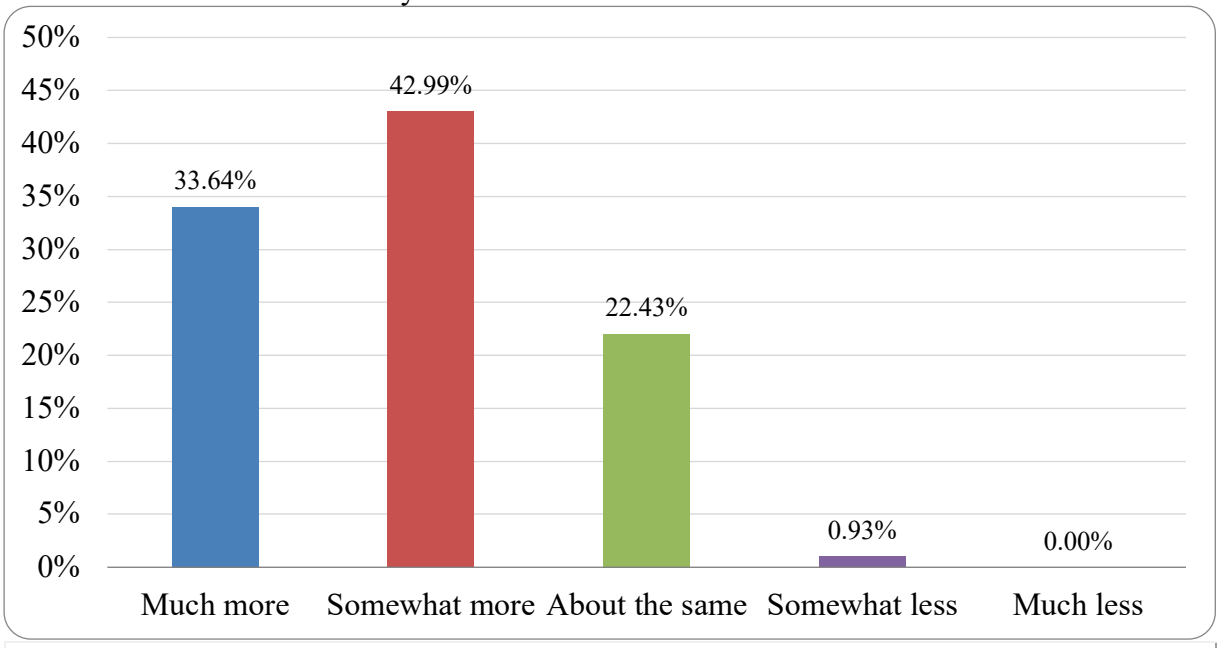
	Answer	Count	Percent
1.	Strongly agree	32	29.91%
2.	Agree	59	55.14%
3.	Neutral	11	10.28%
4.	Disagree	5	4.67%
5.	Strongly disagree	0	0.00%
	Total	107	100%

Mean: **1.897**	Confidence Interval @ 95%: **[1.752 - 2.042]**	Standard Deviation: **0.764**	Standard Error: **0.074**

Figure 5.6 Survey Question & Results 11

The results shown in Figure 5.6 are based on the question, “Do you agree that the Critical Composite Index & Severity Classification (Minor, Moderate, Major, Severe) is an appropriate representation of traffic crashes - considering the applicable details of the crash?” This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the supporting severity charts. Based on the results, 55.14% of all respondents agree that the CCI and severity classification is an appropriate representation of traffic crashes. An additional 29.91% strongly agree that the severity classification is an appropriate representation of traffic crashes. 85.05% of all respondents agree or strongly agree that the CCI and severity classification is an appropriate representation of traffic crashes. Furthermore, based on the sampled results, the mean is 1.897 which is between strongly agree and agree; with a 95% confidence that the true population mean would strongly agree to agree that they find the CCI and severity classification is an appropriate representation of traffic crashes. The results from this question describe that a majority of roadway users find the CCI and severity chart to be an appropriate way to classify traffic crashes.

All traffic crash data collected comes from the Texas Department of Transportation: Knowing that the data used comes from a reliable source, how much does it improve your trust of the Critical Composite Index being a useful way to classify, understand, and compare traffic crashes in a standard way?



	Answer	Count	Percent
1.	Much more	36	33.64%
2.	Somewhat more	46	42.99%
3.	About the same	24	22.43%
4.	Somewhat less	1	0.93%
5.	Much less	0	0.00%
	Total	107	100%

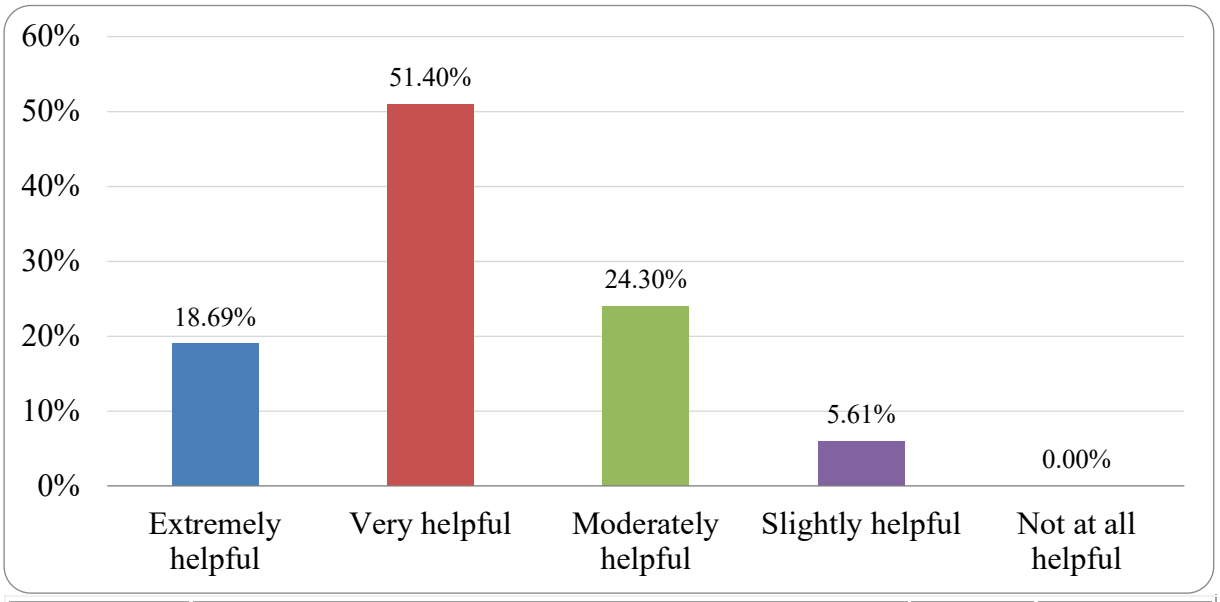
Mean: **1.907**	Confidence Interval @ 95%: **[1.760 - 2.053]**	Standard Deviation: **0.771**	Standard Error: **0.075**

Figure 5.7 Survey Question & Results 14

The results shown in Figure 5.7 are based on the question, "All traffic crash data collected comes from the Texas Department of Transportation: Knowing that the data used comes from a reliable source, how much does it improve your trust of the Critical Composite Index being a useful way to classify, understand, and compare traffic crashes in a standard way?" This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the

supporting severity charts. Based on the results, 42.99% of all respondents agree that knowing the CCI data is retrieved from a reliable source improves their trust somewhat more of it being a useful way to classify, understand, and compare traffic crashes in a standard way. An additional 33.64% of all respondents agree that knowing the CCI data is retrieved from a reliable source their trust much more of it being a useful way to classify, understand, and compare traffic crashes in a standard way. A total of 76.63% of all respondents agree that knowing the CCI data is retrieved from a reliable source improves their trust much more of it being a useful way to classify, understand, and compare traffic crashes in a standard way. Furthermore, based on the sampled results, the mean is 1.907 which is between much more and somewhat more on knowing CCI data is retrieved from a reliable source improves their trust of it being a useful way to classify, understand, and compare traffic crashes in a standard way; with a 95% confidence that the true population mean has improved trust in the CCI usage. The results from this question describe that a majority of roadway users find the CCI data source is important to determining their trust of using it.

How helpful would the Critical Composite Index improve the way you would compare traffic crashes to one another?



	Answer	Count	Percent
1.	Extremely helpful	20	18.69%
2.	Very helpful	55	51.40%
3.	Moderately helpful	26	24.30%
4.	Slightly helpful	6	5.61%
5.	Not at all helpful	0	0.00%
	Total	107	100%

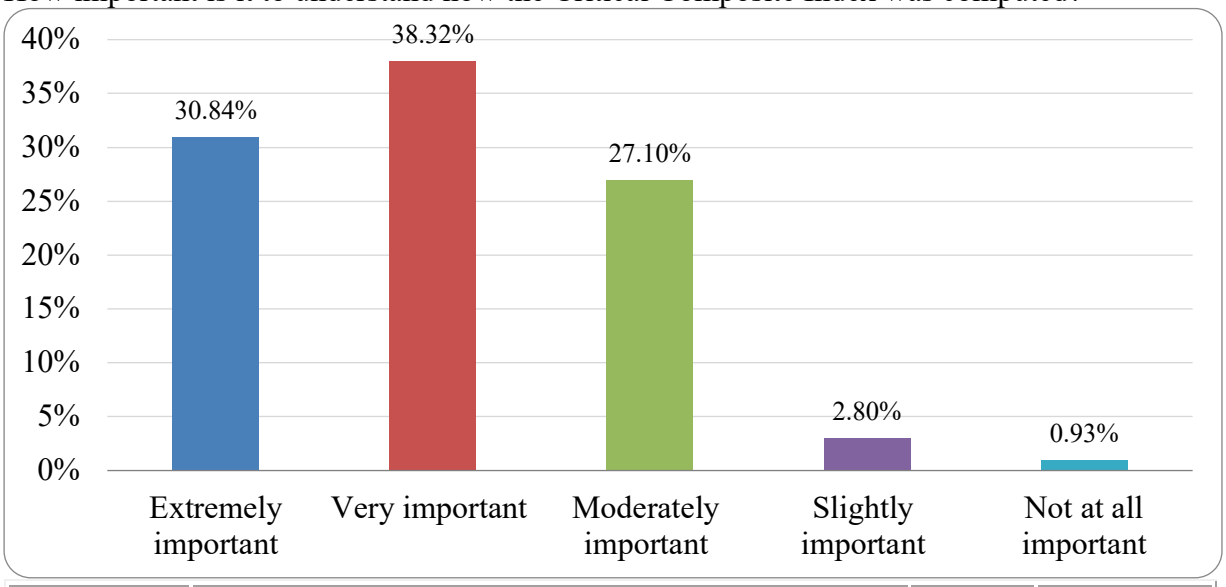
Mean: **2.168**	Confidence Interval @ 95%: **[2.018 - 2.319]**	Standard Deviation: **0.795**	Standard Error: **0.077**

Figure 5.8 Survey Question & Results 15

The results shown in Figure 5.8 are based on the question, “How helpful would the Critical Composite Index improve the way you would compare traffic crashes to one another?” This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the supporting severity charts. Based on the results, 51.40% of all respondents find the CCI Very helpful with another 18.69% finding it extremely helpful as a way to compare traffic crashes to one another. 94.39% of all respondents find the CCI to be moderately to extremely helpful as a way to compare traffic crashes to one another; 95% confidence that the

true population mean finds the CCI approximately very helpful as a way to compare traffic crashes to one another. The results from this question describe that a majority of roadway users find the CCI as a helpful way to compare traffic crashes to one another.

How important is it to understand how the Critical Composite Index was computed?



	Answer	Count	Percent
1.	Extremely important	33	30.84%
2.	Very important	41	38.32%
3.	Moderately important	29	27.10%
4.	Slightly important	3	2.80%
5.	Not at all important	1	0.93%
	Total	107	100%

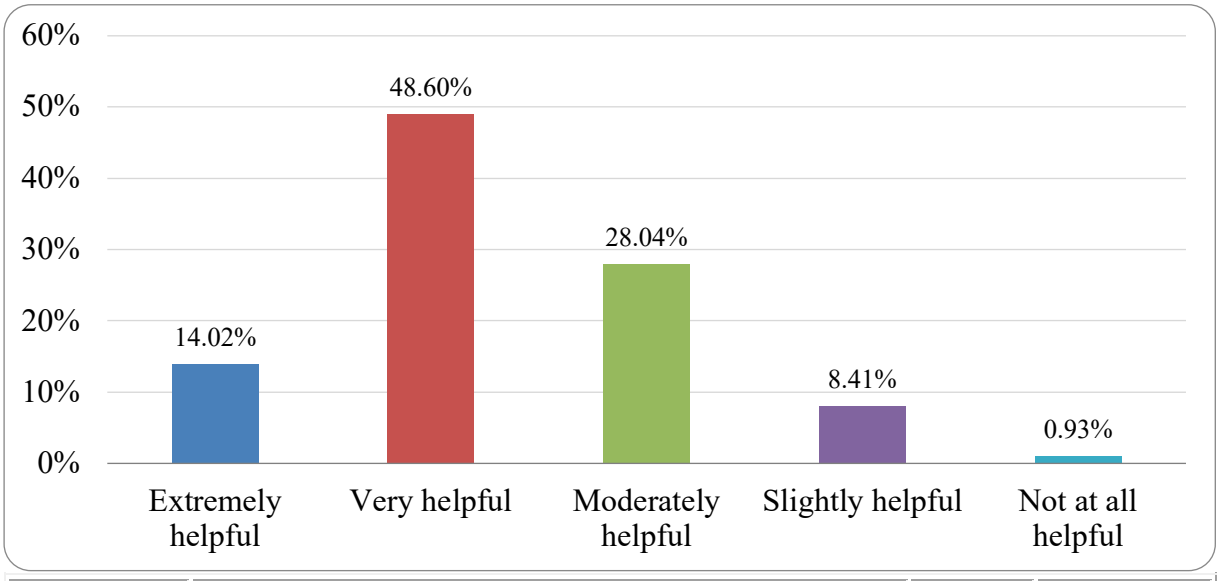
Mean: **2.047**	Confidence Interval @ 95%: **[1.879 - 2.214]**	Standard Deviation: **0.884**	Standard Error: **0.085**

Figure 5.9 Survey Question & Results 16

The results shown in Figure 5.9 are based on the question, "How important is it to understand how the Critical Composite Index was computed?" This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the supporting severity charts. Based on the results, 38.32% of all respondents find the computation of the CCI to be

very important with an additional 30.84% to be extremely important. 96.26% of all respondents find the computation of the CCI to be moderately to extremely important to know. 95% confidence that the true population mean finds the CCI computation methodology approximately very important to know. The results from this question a majority of roadway users are interested in understanding the way the CCI was computed.

If the Critical Composite Index was used in practice (e.g. news media, mobile app) I would find it:



	Answer	Count	Percent
1.	Extremely helpful	15	14.02%
2.	Very helpful	52	48.60%
3.	Moderately helpful	30	28.04%
4.	Slightly helpful	9	8.41%
5.	Not at all helpful	1	0.93%
	Total	107	100%

Mean: **2.336**	Confidence Interval @ 95%: **[2.174 - 2.499]**	Standard Deviation: **0.857**	Standard Error: **0.083**

Figure 5.10 Survey Question & Results 17

The results shown in Figure 5.10 are based on the question, “If the Critical Composite Index was used in practice (e.g. news media, mobile app) I would find it:” This question is intended to provide insight to user knowledge gain and perception of the CCI, its classification and the supporting severity charts. Based on the results, 48.60% of all respondents would find the CCI Very helpful with another 14.02% finding it extremely helpful if adopted in practice. A total of 90.06% of all respondents would find the CCI to be moderately to extremely helpful if adopted in practice; 95% confidence that the true population mean finds the CCI approximately very helpful as a way to compare traffic crashes to one another.

The results from the questions shown in Figures 5.6-5.10 show that roadway users find the CCI as a helpful guide to classify, understand, and compare traffic crashes. Moreover, roadway users gain knowledge from the CCI and may participate in classifying, understanding, and comparing traffic crashes if adopted in practice.

5.3.3. Improvement from current metric reporting

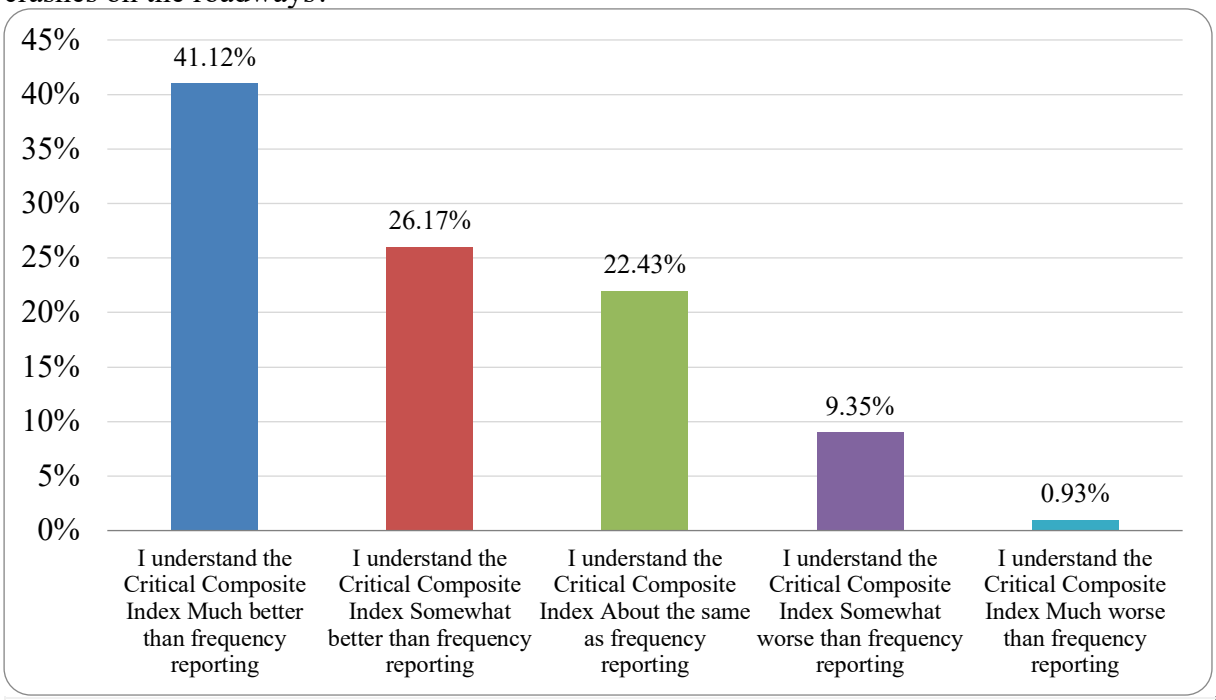
Commonly, traffic crashes are reported to the public as a frequency and use a single scope (e.g. fatality, injury) of measure such as the following:

- number of crashes per year (crash rate)
- number of injuries per year (injury rate)
- number of fatalities per year (fatality rate)

Individual Crashes Critical Composite Index	Frequency Report of Crashes
Crash One - **19.41 Minor** Crash Two - **47.90 Major** Crash Three - **22.24 Moderate** Crash Four - **41.23 Major** Crash Five - **58.16 Severe** Crash Six - **54.31 Severe** Crash Seven - **30.36 Moderate** Crash Eight - **18.13 Minor** Crash Nine - **100.94 Severe** Crash Ten - **38.06 Moderate**	Number of Crashes - 10 Number of Injuries - 37 Number of Fatalities - 1

The data shown above is actual traffic crash data; both the Critical Composite Index and the Frequency Metrics are based on the same set of data.

Which method of reporting, the Critical Composite Index (Case-by-case) or Frequency Report of Crashes (Group of crashes), do you understand better when trying to understand traffic crashes on the roadways?



	Answer	Count	Percent
1.	I understand the Critical Composite Index Much better than frequency reporting	44	41.12%

2.	I understand the Critical Composite Index Somewhat better than frequency reporting		28	26.17%
3.	I understand the Critical Composite Index About the same as frequency reporting		24	22.43%
4.	I understand the Critical Composite Index Somewhat worse than frequency reporting		10	9.35%
5.	I understand the Critical Composite Index Much worse than frequency reporting		1	0.93%
	Total		107	100%
Mean**: 2.028**	Confidence Interval @ 95%**: [1.829 - 2.227]**	Standard Deviation**: 1.050**	Standard Error**: 0.102**	

Figure 5.11 Survey Question & Results 18

The results shown in Figure 5.11 are based on the question, "Which method of reporting, the Critical Composite Index (Case-by-case) or Frequency Report of Crashes (Group of crashes), do you understand better when trying to understand traffic crashes on the roadways?" This question is intended to determine how the CCI is an improvement from current frequency-based metrics. Based on the results, 41.12% of all respondents understand the CCI much better than frequency reporting with an additional 26.17% understanding it somewhat more. A total of 67.28% of all respondents report of an improvement of understanding the CCI over common frequency reporting. 95% confidence that the true population mean understand CCI somewhat more than current frequency reporting. The results show that the CCI gives roadway users more understanding of issues on the road than frequency reporting metrics.

5.3.4. CCI Improvements

Figures 5.12 – 5.14 are based on open response of all of the respondents in the user evaluation survey. Based on the results, keywords were tagged to determine the areas of interest.

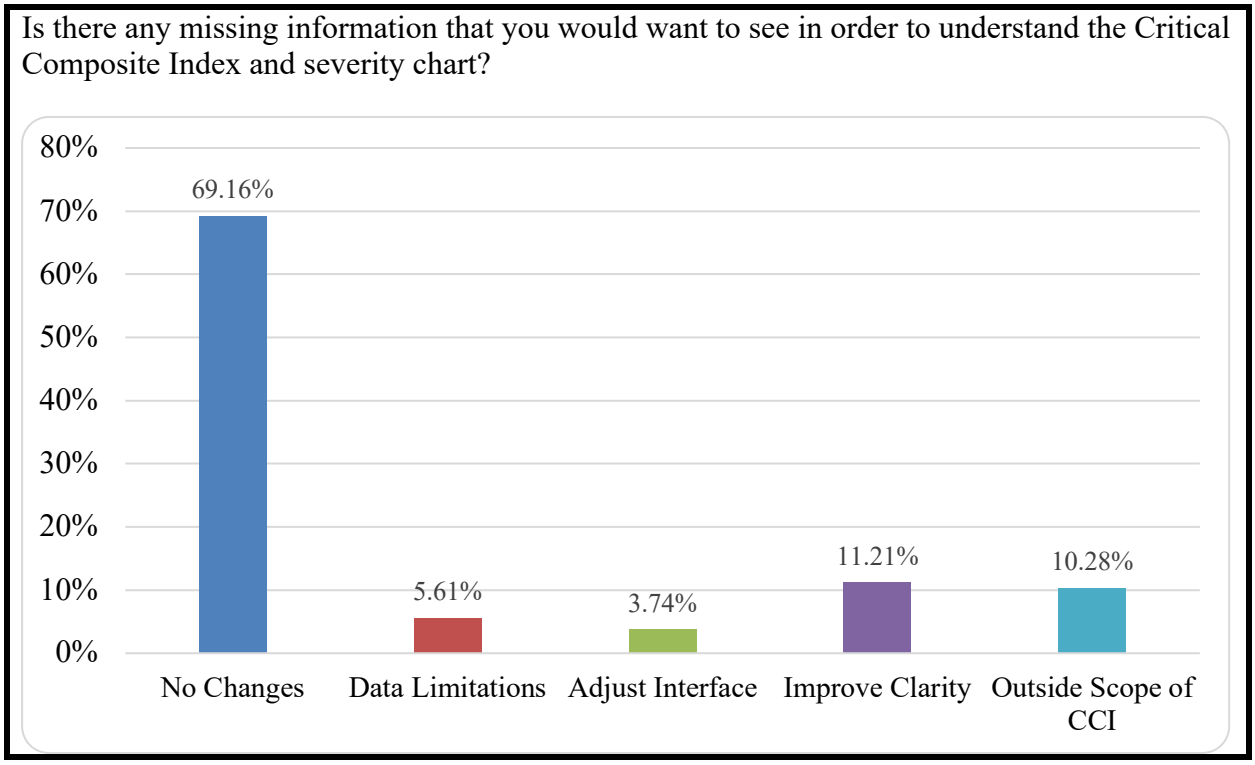


Figure 5.12 Survey Question & Results 19

The results shown in Figure 5.12 are based on the open response question, "Is there any missing information that you would want to see in order to understand the Critical Composite Index and severity chart?" Based on the text analysis, 69.19% of respondents do not contribute any missing information to the CCI. Other areas of improvement include improving the clarity of the description in the survey itself. Additional improvements are limited by data availability and are outside of the scope of this research.

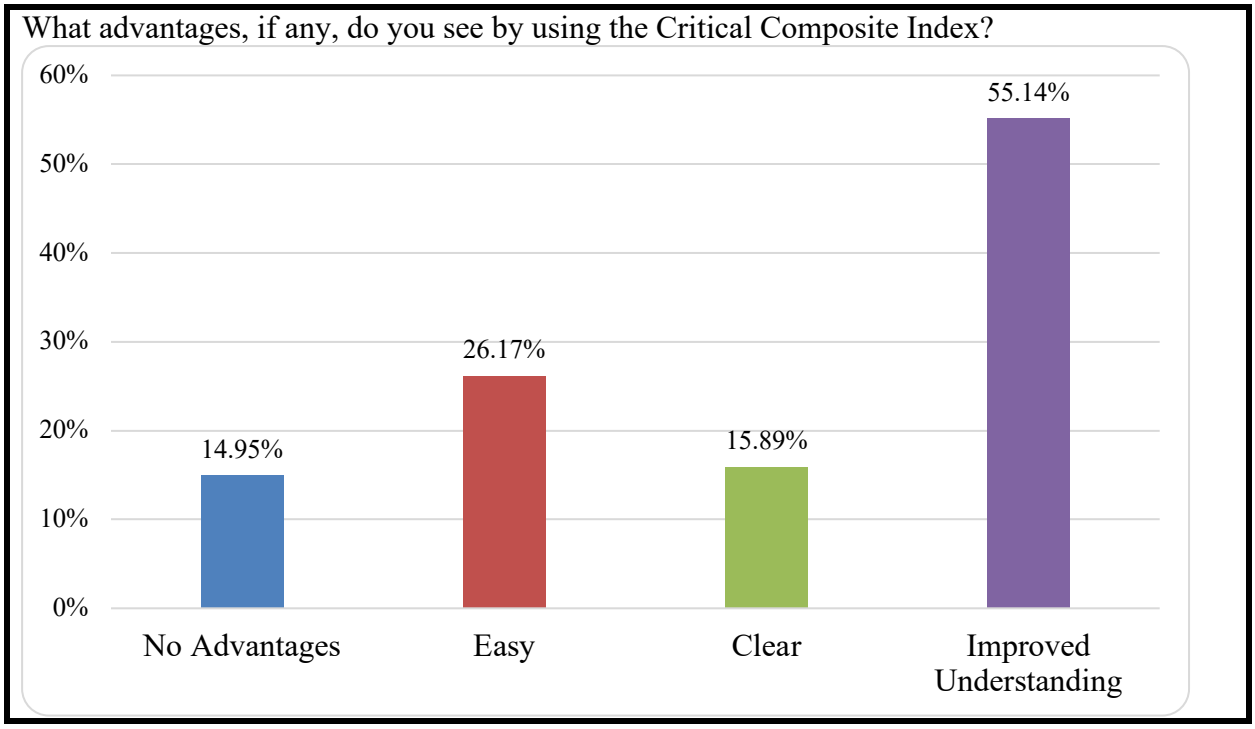


Figure 5.13 Survey Question & Results 20

The results shown in Figure 5.13 are based on the open response question, "What advantages, if any, do you see by using the Critical Composite Index?" Based on the text analysis, 55.14% of respondents find that using the CCI provides improved understanding of traffic crashes. Other advantage areas include improving the clarity and easiness of using such metric. 14.95% of people do not find any immediate advantages of the CCI.

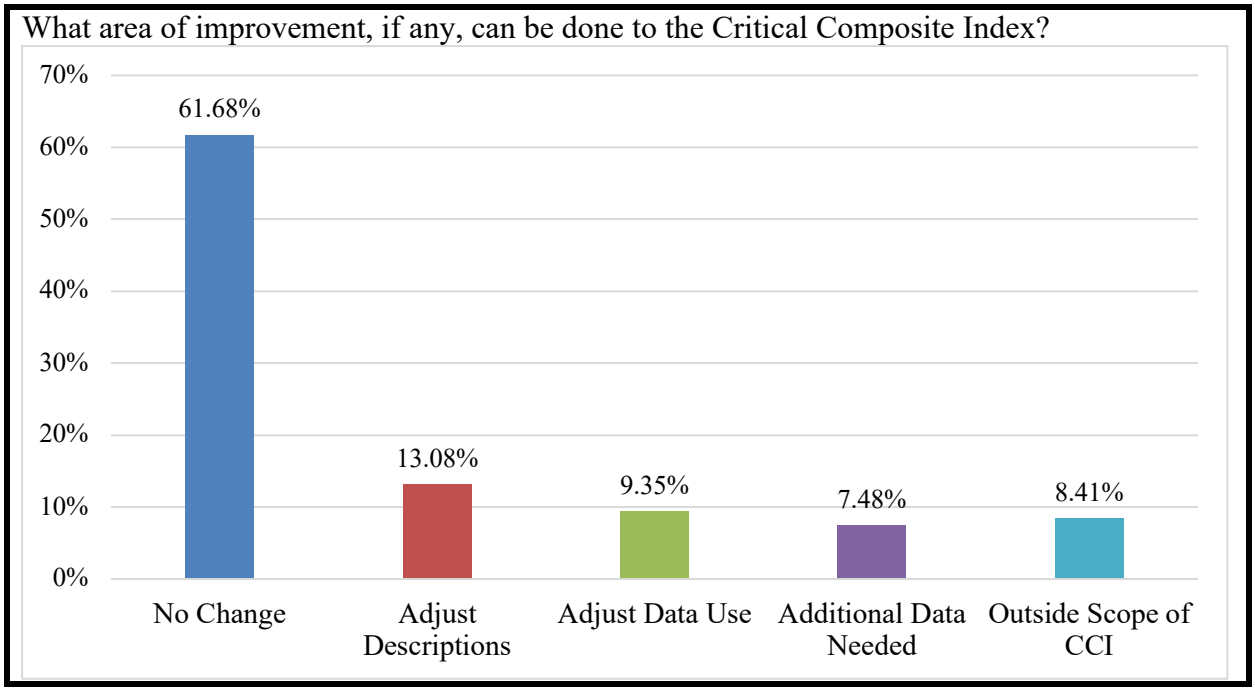


Figure 5.14 Survey Question & Results 21

The results shown in Figure 5.14 are based on the open response question, "What area of improvement, if any, can be done to the Critical Composite Index?" Based on the text analysis, 61.61% of respondents find no changes are needed in the CCI for improvement. Other improvement areas include adjusting the descriptions in practice and how the data is used. Some of the responses require additional data sets and some responses are outside of the scope of this research, such as the impact of insurance and computation of damage costs.

Based on the results shown in Figures 5.12-5.14, the responses to the CCI present the advantages and clarity of the metric. In general, Figures 5.1-5.14 show that the CCI provides additional knowledge to traffic crash reporting to people who use the metric. Additionally, the CCI is comprehendible, provides a gain in knowledge and is an improvement to standard frequency

metric reporting techniques currently used. The full reported results of the CCI is listed in the appendix.

5.4 CCI Vs. Frequency Metrics

The CCI is a combination of a traffic crash event, the people involved, and the external circumstances that may have contributed to the crash. The metric itself, is a study of a particular traffic crash and an index value that is representative of that crash. In contrast, frequency-based metrics including, but not limited to, number of crashes per year (crash rate), number of injuries per year (injury rate), number of fatalities per year (fatality rate), number of crashes per 100,000-population per year, number of injuries per 100,000-population per year, number of fatalities per 100,000-population per year do not describe each crash specifically. Currently, the NHTSA and the FHWA do not study each traffic crash individually nor compare them to each other.

Table 5.1 shows a side-by-side comparison between the BUM methodology components of the CCI and Competency questions against current practice frequency reporting. Using the BUM methodology, the CCI is able to meet standards (blue), competency questions is able to meet standards (blue), and both the CCI and Competency questions work together (green) to provide a better understanding of how it is an improvement to standard frequency metric reporting practices; items in red are not accomplished.

Table 5.1 CCI & Competency Questions vs. Frequency Comparison Table

	BUDD ALGORITHM			*Current Practice*
	Critical Composite Index	Competency Questions	CCI & Competency Questions	Frequency Metrics
Classify Traffic Crashes	YES	NO	YES	NO
Describes Crash Severity	YES	NO	YES	NO
Standard Reporting Technique	YES	NO	YES	YES
Allows for Crash Comparisons	YES	YES	YES	NO
Transferable (Geographic Locations)	YES	YES	YES	NO
Sustainable Description of Crashes	YES	YES	YES	NO
Reports Information About Crash	YES	YES	YES	YES
Considers Multiple Factors in Severity	YES	NO	YES	NO
Representation of Individual Crashes	YES	YES	YES	NO
Describe Multiple Crashes At Once	NO	YES	YES	YES
Describe Individual Crash Fatalities	NO	YES	YES	NO
Describe Individual Crash Injuries	NO	YES	YES	NO

According to the survey results and the evaluation, it can be concluded that the value of producing a CCI for each traffic crash is the following:

1. Use the data that is associated to gain new information
2. Produce standard technique to classify a traffic crash
3. Have a measurable metric to apply to each incident
4. Use the metric to compare to other incidents in the same or different geographic locations
5. Use the same metric to compare to other incidents from the past, present, and future to act as an indicator for improvements or digressions in infrastructure, roads, safety, and severity

Frequency metrics do not provide users to gain new information nor do they classify crashes in a standard way. Moreover, frequency-based metrics cannot act as an indicator for improvement or digression in infrastructure, roads, safety, and severity because it does not consider additional data such as the crash location that is part of the data and used in combination with the CCI.

Though frequency metrics do not measure traffic crashes on a case-by-case basis they do have their purpose. However, using the BUM methodology introduced through this research, additional more complex competency questions can be used to query traffic crash data. Moreover, in combination with the CCI, traffic crash information can be counted based on the values that are presented.

5.5 CCI AS A VALUABLE METRIC

Metrics are evaluated based on their ability to represent total comprehension of the incident severity and the effect that it may have on other commuters. Prior to this work, traffic crash metrics were mere statistics, to which this research can also provide through appropriate competency questions. Moreover, the metrics prior to this work were subjective on person-by-person and case-by-case basis; the metrics developed through this work eliminates multiple individuals determining their own severity or index values.

The CCI is representative of the 12 major characteristics of a valuable metric described by Hoornweg et al (Hoornweg et al., 2007) as shown in Table 5.2.

Table 5.2 Metric evaluation criteria

Metric Evaluation Criteria	CCI
Clear Objective	The CCI has an objective to describe traffic crashes in terms of the traffic event, people involved, and external circumstances to help understand safety and mobility of people
Relevant	The CCI has weighted criteria are based on the crash event, people involved, and external circumstances related to the traffic crash
Measurable & Replicable	The CCI can be measured based on a severity chart that has been developed; It can be replicated for different weighted values
Statistically Representative	The CCI provides description of traffic crashes in many cities in Texas
Comparable	The CCI has been shown to compare crashes between cities through case studies
Standardized	The CCI uses the same weighted values for each traffic crash value that is computed
Prediction Potential	The CCI may provide insight to reoccurring patterns on the roadways with a similar severity index; with a numerical value, it is capable to undergo machine learning algorithms for additional prediction

Effective	The CCI is capable of measuring a large number of traffic crashes throughout the State of Texas relating them all to the same scale for comparison and standardization
Economical	The CCI uses open historical data that has no cost
Interrelated to Society	The CCI provides citizens the opportunity to generalize traffic crashes into a numerical value
Consistent	The CCI is based on weighted values that do not change, thus the values computed are relative to each crash
Sustainable	The CCI has compared values from multiple years using similar data sources; as a result of being used over multiple time periods it is sustainable

The development of a CCI is intertwined with bottom-up modeling, which measure "before, during, and after" results of Smart Cities solutions. Unlike current metrics that describe frequency, the CCI is transferable, comparable, and sustainable. By describing a crash event in El Paso, it can be compared to similar events in Austin, San Antonio, Dallas, Houston or other cities and towns throughout the state. Moreover, the model is open to change based on additional or alternative data throughout the rest of the United States. As traffic crashes continue to occur on roadways, the data collected will continue to be useful to the CCI, which allows it to be sustainable.

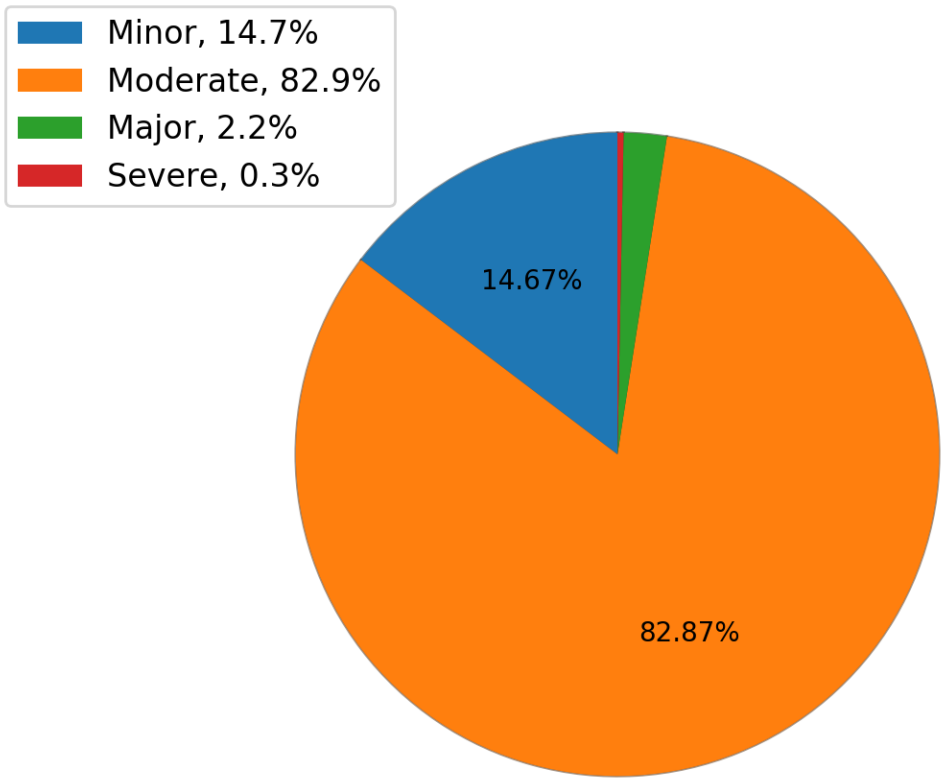


Figure 5.15 Weight Sample One Distribution

Figure 5.15, the distribution of minor, moderate, major, and severe crashes based on the weight samples. The raw data reports five separate severity levels for each traffic crash: Unknown, not injured, possible injury, non-incapacitating injury, suspected serious injury, and fatal.

Table 5.3 CCI Precision & Recall with respect to reported severity categories in traffic crash data

		Reported	
		(Unknown || Poss Injury) - Positive	(Not Inj || NonIncap Inj || Susp Serious Inj || Fatal) - Negative
Actual	Minor - Positive	62,111	-
	!Minor - Negative	443,591	16,628
Precision	12.28%		
Recall	100.00%		
F	21.88%		
		Reported	
		(Not Injured || NonIncap Inj) - Positive	(Unknown || Poss Inj || Susp Serious Inj || Fatal) - Negative
Actual	Moderate - Positive	1,642,239	6,387
	!Moderate - Negative	20,412	10,241
Precision	98.77%		
Recall	99.61%		
F	99.19%		
		Reported	
		(Suspected Injury) - Positive	(Unknown || Not Injured || Poss Inj || NonIncap Inj || Fatal) - Negative
Actual	Major - Positive	66,934	1,967,731
	!Major - Negative	56,509	1,947,498
Precision	54.22%		
Recall	3.29%		
F	6.20%		
		Reported	
		(Fatal) - Positive	(Unknown || Not Injured || Poss Inj || NonIncap Inj || Susp Serious Inj) - Negative
Actual	Severe - Positive	3,512	179
	!Severe - Negative	13,116	1,967,731
Precision	21.12%		
Recall	95.15%		
F	34.57%		

Table 5.3 shows the precision, recall and F score for the CCI compared to the Crash_Severity_ID gathered from the raw data report. For a minor traffic crash, unknown injuries and possible injuries are best matched because it is not clear nor can be determined if these injuries had any effect on the accident severity, thus a minor concern. For a moderate traffic crash, not injured or non-incapacitating injuries are best matched because both types of injuries are either non-apparent or not life threatening, thus a moderate concern. For a major traffic crash, suspected serious injuries is best matched because they can have an effect on life survival and can be generalized to be a major concern. For severe traffic crashes, fatal injuries are best matched because fatalities in a crash denote a severe concern. The CCI does not compute severity based on one individual person involved in the traffic crash, but instead considers every person and details surrounding the traffic crash. The results in Table 5.5 show that minor, major, and severe

accidents have a low accuracy (F) value. The low accuracy value stems from there being a disconnect in how the raw data reports severity and how the CCI computes it. Using the CCI, a traffic crash is defined by more than a single criterion, but instead expanded to understand the totality of the crash. The CCI is based on weights that consider the traffic crash event, people involved, and additional circumstances; moreover, the CCI can be adjusted to consider additional data fields beyond what is considered in this research. The determination of reported severity from the raw data is based on the worst injured person in the traffic crash. The reported severity does not consider other characteristics of the traffic crash beyond the most injured person. The difference in results between minor and moderate traffic crashes and its respective reported severity is based on what is considered in a traffic crash. Having no injuries in a traffic crash does not necessarily describe it as a "minor" event because it does not consider the damage, location, and other circumstances involved. The CCI describes a traffic crash that is minor to essentially having no effect (or unknown effect) on the people or vehicle involved, whereas a moderate crash has more of an effect on the people involved and gives consideration to the location and circumstances surrounding the traffic crash. Though reported minor and moderate traffic crashes do not match exactly, major and severe crashes are more closely represented. Though the CCI considers many criteria, serious injuries and fatalities are highly weighted values for the CCI, thus is closely related to the raw data severity observation. The values provided by the CCI provide more meaning to the traffic crash because it considers multiple factors involved.

5.6 BUM METHODOLOGY TRANSFERABILITY

Information from approximately 120,000 crashes occurred in 2014 in the State of Pennsylvania was successfully ingested by the application created to test the BUM methodology and the generation of the CCI metric. The distribution of the results from the CCI severity (appendix) for Pennsylvania is: 35.8% minor, 61.7% Moderate, 2.3% Major, and 0.2% Severe. Compared directly to the 2014 Texas traffic crash distribution, the results were similar to the State of Texas: 11.8% minor, 85.2% moderate, 2.4% major, and 0.5% severe.

Applying the BUM methodology and CCI to data from the state of Pennsylvania has validated that the methodology is transferable beyond the State of Texas. The BUM methodology and the CCI are generic enough to integrate additional data-sets by applying the same steps proposed. The transferability of this methodology is demonstrated in this work by the introduction of additional traffic crash data sets. Crash data from the State of Pennsylvania was added to describe the transferability of the BUM methodology. The Pennsylvania Department of Transportation (PennDOT) provides publicly available traffic crash data similar to the State of Texas through PennDOT open data ((PennDOT), 2019). Table 5.4 shows a portion of data that is available in Texas and Pennsylvania.

Table 5.4 Texas and Pennsylvania Data Points Comparison

Data Point	Description	Texas	Pennsylvania
Active_School_Zone_Fl	In School Zone	X	X
At_Intrsct_Fl	At Intersection	X	X
Auto_Count	Number of Vehicles		X
Bus_Count	Number of Buses		X
Cmv_Involv_Fl	Comercial Vehicle Involved	X	X
Crash_Fatal_Fl	Fatal Crash	X	X
Crash_ID	Crash ID	X	X
Crash_Time	Time of Crash	X	X
Death_Cnt	Number of Fatalities	X	X
Est_Hours_Closed	Est. Time of Road Closure		X
Harm_Evnt_ID	Object Struck	X	X
Lane_Closed	Lane Closures		X
Light_Cond_ID	Light Conditions	X	X
Medical_Advisory	Medical Emergency	X	
Motorcycle_Count	Number of Motorcycles		X
Non_Injry_Cnt	Number of Non-Injuries	X	X
Nonincap_Injry_Cnt	Number of Nonincapacitating Injuries	X	X
Poss_Injry_Cnt	Number of Possible Injuries	X	X
Private_Dr_Fl	Occurred at Private Road	X	
Road_Constr_Zone_Fl	At Construction Zone	X	X
Rpt_Outside_City_Limit	Outside of City Limits	X	
Schl_Bus_Fl	School Bus Involved	X	X
Small_Truck_Count	Number of Small Trucks		X
Street_Name	Name of Street	X	
Surf_Cond_ID	Surface (Road) Conditions	X	X
Sus_Serious_Injry_Cnt	Number of Serious Injuries	X	X
Unkn_Injry_Cnt	Number of Unknown Injuries	X	X
Wthr_Cond_ID	Weather Conditions	X	X

The data provided by Pennsylvania was used to evaluate the transferability of the BUM methodology and the CCI to a different crash data-set in the United States. The data-set provided by Penn open data has similar data points as Texas; a majority of the data points that were used as part of the weighted samples were available as part of the Pennsylvania crash data-set. Both the BUM methodology and the CCI were applied to the Pennsylvania data-sets.

5.6.1. BUM Methodology Applied to Pennsylvania Crash Data

Step 1. Domain Analysis

The domain considered in Pennsylvania is the same as what was considered in Texas. Traffic crash analysis and mobility metric research was conducted to provide understanding the domain.

Step 2. Data Discovery

The data-set of the 2014 year was used and was acquired from publicly available data-sets from the Pennsylvania Department of Transportation ((PennDOT), 2019). This data-set contains information about 120,000 crashes. As shown in Table 5.4, a majority of the data points were the same in both Texas and Pennsylvania; this is because they both use similar guidelines to collect data. The main differences of the data are that some data points are named differently in each of the states and the specificity of data from state to state.

Step 3. Data Modeling.

The process of modeling additional data is extended from the model created for the State of Texas. Data from the new data source is added to the data model and uncommon data points from either the Texas or Pennsylvania data-sets would remain as empty if there is no applicable data for it.

Step 4. Data Processing & Mapping

Data processing and Mapping of the data from Pennsylvania was nearly identical to that of Texas. The parser created to map the data from the Texas data-set was designed to handle many different data-sets with similar fields. Additional fields included the type of intersection of a traffic crash, additional weather and surface conditions. These criteria were added as part of the weighted sample list that is accessed and integrated into the CCI computation. Additionally, the reporting codes for the State of Pennsylvania differ from that of Texas but were manually

adjusted to meet the criteria values. Moreover, additional fields were also easily added to the parser code. The addition of data from Pennsylvania allowed to show the flexibility of the parser code to handle data-sets outside of the original Texas data-sets used.

Step 5. Index Development

The transferability CCI is demonstrated by modifying the weighted sample criteria (appendix) to represent criteria based on the Pennsylvania traffic crash data. Additional information that was collected by PennDOT was used in addition to the common fields from Texas. The criteria field names for Texas and Pennsylvania are not identical, however, the Pennsylvania criteria names were adapted to meet the criteria names from the Texas data-set. The CCI was computed programmatically for the State of Pennsylvania by modifying approximately 5 lines of code from the parser that was created for the State of Texas (not including file paths). Over 120,000 traffic crashes in 2014 for the State of Pennsylvania were integrated into the system and a CCI was computed for each.

Chapter 6: Discussion

6.1 BUM Methodology

The transformation of data to knowledge through the use of the BUM methodology can be compared to other similar methods that take a top-down approach. A method that uses a bottom-up approach differentiates itself from a top-down method because they are looked at from two different perspectives. In the case of a top-down approach, researchers look at the problems that exist and then attempt to solve those problems. Paul Sabatier (Sabatier, 1986) writes that using a top-down approach takes the perspective of decision makers and in some cases neglect other actors. With respect to Smart Cities research, many scientists, engineers, or policy-makers first look at some 'problem' and begin to take action, without looking first at the actors – in this case, data. This idea that Smart Cities research is typically done top-down is described by Dimitri Schuurman et al. (Schuurman, Baccarne, De Marez, & Mechant, 2012) writing, "Traditional innovation processes start from the belief that an innovation is best developed using a top-down approach." This idea is reinforced by the research done by Marion K. Poetz and Martin Schreier (Poetz & Schreier, 2012) describing that larger organizations have much more control on intellectual property of products and services, thus allowing solutions to conform to their strategic plans.

As top-down approaches focuses on the trickle-down effect of an idea, compared to the BUM methodology that focuses on using factual relevant data to introduce novel ways to understand traffic crashes without introducing an organizational agenda (Sabatier, 1986). There is an idea of the *open innovation with customers* paradigm that suggests that end-users make a relevant contribution to the development process…to solve the needs of the problems that they are facing (Schuurman et al., 2012). The *open innovation with customers* paradigm opens up the need to

have alternate ways to develop solutions, metrics, processes for the needs of Smart Cities which is inherent with the BUM methodology.

The metric developed as a result of the BUM methodology have organically been created through linking of heterogeneous data sets with the intent for them to become interoperable. Moreover, data science has a major focus in studying data and disseminating facts from it (Hey, 2009), which the BUM methodology contributes to. A bottom-up model driven by relevant data for the creation of a knowledge graph is a highly acceptable technique, based off of the literature presented in this section, for the creation of novel metrics and knowledge graph for the dissemination of information. The work done in this research follows the same paradigm of a bottom-up model since it focuses on a larger domain prior to the development of an individual solution.

In this research data was gathered from the CRIS for the State of Texas (TxDOT, 2018). Furthermore, for this research relies on the accuracy of the data source. The data provided to CRIS comes from local law enforcement agencies. The BUM methodology provides the systematic approach to transform original investigation data into useable knowledge. Moreover, the BUM methodology has a possibility to improve the way law enforcement records traffic crash data by introducing a standard way to classify traffic crashes. Through the introduction of a standard classification method, the BUM methodology uses the standard data collection and composite information to classify traffic crashes whereas current use worst injured person to classify crashes. To disseminate knowledge, it is critical that the data is coming from a trusted source. The technique that the data was collected in this research is relatable to the way other large data sets are commonly gathered; the United States government hosts an online open data

listing (U.S. General Services Administration, 2019) that researchers can gather information about many topics. In the realm of traffic, there are a large number of different data sets that can be collected. Much of the data that is collected has been sampled from drivers themselves. The National Personal Transportation Survey (NPTS) has collected data about vehicle miles traveled (VMT) to get a better understanding of the roadways (Massie, Campbell, & Williams, 1995). In contrast to the information that has been gathered, official reported traffic crashes provide a better understanding of the roadways because it provides a larger outlook on a specific area of traffic. Though the data collected does not provide information regarding VMT, it still provides valuable information that can be used for the enhancement of knowledge.

Traffic crash data is directly reported to CRIS by official police agencies and then made available for use. Though official police agencies follow specific procedures when documenting traffic crashes, there is a chance of errors being introduced into the data; this can occur because of simple human error, or because of an investigation not being mapped out properly (Shields, 2018).

In contrast, errors can also be introduced into data that does not directly involve humans such as sensor malfunction. Data processing and mapping involves cleaning data and transforming it into a singular form. Cleaning data is a necessary part of the BUM process because some of the data acquired is not in the same form. Rahm and Do (Rahm & Do, 2000), describe some of the inherent problems with data sets as having illegal values, duplicates, or varying value representations. Through the removal of errors in the data sets, they can easily be integrated into a single form. For this work, JSON-LD was used as a form because it is a W3C standard that is compatible with many programming languages and databases (Sporny, Longley, Kellogg,

Lanthaler, & Lindström, 2014). This work has leveraged standard practices to clean and map data into a standard form.

The BUM methodology has a major outcome of the CCI as a source of data dissemination; however additional outcomes are explored as part of additional data views. The CCI provides information about each traffic crashes and gives individuals the possibility to use the output values of the CCI to compare to one another. The output of maps as shown in the case studies provide a data view that gives information that is useful as a visual representation of traffic crashes. Additionally, data narratives provide human readable text that is useful for individuals who wish to have information about traffic crashes.

Although it is impossible to have complete certainty about the accuracy of any data that was not collected by the researcher, both federal and state government have high standards of accuracy and should be safe to consider accurate. Moreover, traffic crash data has the potential to help people as commuters and improve their knowledge of what is occurring on the roadways. Data plays a significant role in the BUM methodology because it is at the core of potential knowledge. Through the process of discovering data, it is clear that this research follows an accepted approach based on the source of information and the way that information can be used in society.

6.2 Research Questions Discussion

There were three research questions that were being answered through this work. The BUM methodology and metric development conducted through this work provided the answers to the research questions.

6.2.1. Research Question One Discussion

What do semantically-enhanced data models contribute to data representation and data integration for metric development of publicly available mobility data targeting non-domain expert stakeholders?

Advancement in state-of-the-art Computer Science methods and Smart Cities research stems from there being real-world problems to solve. In this research, traffic crashes are at the core of what needs to be understood. This work illustrates how traffic crash analysis can be independent of subjective human reasoning.

In this research, semantic enhanced bottom-up models provide a mechanism to create a metric for traffic crashes that improve the way traffic accidents are understood, as shown by the user evaluation survey. In particular, the contribution of a bottom-up model provides additional semantic meaning for any given traffic crash. The additional meaning improves the understanding of a traffic crash event by introducing additional external data for consideration. The CCI is an outcome of the BUM methodology which provides a standard measurement to improve the understanding of this event for a variety of stakeholders, including non-domain experts. Semantic annotation has provided individual data sets to be linked to outside sources and upper-level ontologies for clear representation of data for future research consumption. Although with many data-sets there may be outliers or incorrect information in them, the data that was considered in this research came from credible sources that are believed to still provide crucial and truthful information that is necessary to understand traffic crashes.

Traffic crashes are known to affect many people throughout the city of El Paso, the State of Texas, the United States, and throughout the world. Having a knowledge graph provides the foundation needed to make informed decisions about the domain area and provides a way for data to be clearly described, mapped, and disseminated. Through the representation of data through a knowledge graph that can be transformed into a metric or a narrative, it paves the way for the clear representation of domain data for many stakeholders including, but not limited to, non-domain experts, domain experts, police, policymakers, and researchers.

The CCI developed through this research is derived directly from data itself. Incorporating a limited amount of subjectivity through weighted criteria, the index itself is a measure of traffic crashes that can be used beyond the State of Texas. Semantically annotated models not only contribute to the actual development of the metric through the modeled traffic crash data, but it also contributes to the way data is presented to people in a way that everyone can understand, which is ultimately the focus of Smart Cities research.

6.2.2. Research Question Two Discussion

How can a modeling methodology contribute to the transformation of data to useful knowledge represented for supporting decision making by non-domain experts?

The transformation of data to knowledge through modeling provides a standard process for non-domain and domain experts to explore data. The standard method introduces a weight-based CCI that is necessary to provide a better insight to traffic crashes; formally described metrics can be transferred between regions and allow side-to-side comparison by comparing the percentage of crashes at each level of severity.

The data model itself gives a layout for the transformation of data to knowledge; it is expected that any data-set can be transformed into knowledge using this approach. This approach leverages the idea of using formalized data transformation algorithms for knowledge processing. Knowledge processing gathers information that is untapped in data to foster knowledge gain. The BUM methodology is a technique that explicitly shows the process of transforming data to knowledge through the use of a defined data model described by a knowledge graph. The BUM methodology enhances the way data is transformed by first discovering the importance of a singular domain, discovering data that is useful to describe such domain, and a standard process to exploit that data into a metric. This research presents a domain of traffic crashes and through the BUM methodology establishes a novel metric, namely the CCI. The CCI presents raw data in a form that is understood by non-domain experts to be useful as shown in the user evaluation survey. Moreover, the CCI is provides clarity and a basis for information dissemination; the CCI can be interpreted in a clear way and is an outcome of useful knowledge as defined in Chapter 2.

Modeling data is not a trivial task, therefore using a standard methodology is helpful to understand how that data can solve problems. Through the practice of bottom-up modeling, additional transportation metrics can be derived by observing the available data; gaps in the data-sets can easily be seen through a bottom-up approach and additional data can be retrieved. Additional metrics that may be useful in transformation is understanding the environmental impact of traffic or understanding which road segments may need to be addressed.

6.2.3. Research Question Three Discussion

How can very large public data-sets be used to contribute to quantifiable metrics that are both reusable and comparable to other geographic locations and improve data views?

a. *How does semantically annotated data (input) improve metric development (output)?*
b. *How do elements of data-sets (input) affect the way the information they provide is understood (output)?*

This research uses data-sets that provide a significant amount of information to create quantifiable, measurable, transferable, comparable and standard metrics. This was shown through the implementation of both Texas and Pennsylvania data sets of largely different sizes that produce an CCI output that is expressed for every traffic crash. Each traffic crash is measured individually and compared against one another for analysis as described by the results of this research. Raw data itself does not have to be interoperable to create metrics; it is through the transformation of data to knowledge through the use of a bottom-up methodology that ensures interoperability in data.

Through the case studies done in this research, it is shown that metrics can be developed based on data that are transferable and comparable. It can be observed that the CCI compares individual crashes to one another and that the normalized index can compare all of the traffic crashes to each other as well. The people involved in traffic crashes provide a lot of information about the crash itself – it gives insight into how the crash can be modeled and studied.
By adding semantics to traffic data, the metric can begin to have a deeper meaning itself. The CCI takes into account the traffic crash severity, the effect of the people involved, and external circumstances into one comprehensive metric. The metrics are improved through this work

through the introduction of semantics that provide a method of introducing meaning to metrics in general. Generalized statistics may not have much impact on people because it does not immediately relate to them. However, providing a scaled metric gives people the ability to cognitively reason about the traffic crash and understand what it means to them.

The CCI gives non-domain experts the ability to understand and make sense of any given crash on the roadway as well as domain experts and policymakers to get a deeper understanding of what is occurring on the roadways. The improvement of metric development based on semantically annotated data is shown by the provenance mapping of the knowledge graph. The CCI is clearly mapped to a provenance ontology which not only improves data view input understanding but ensures that the CCI is derived from reliable sources such as CRIS and Penn Share ((PennDOT), 2019; TxDOT, 2018). The understanding of inputs of a given by a data-set that undergoes a transformation into knowledge and provides a clear output in either competency questions or through the CCI improves the knowledge of data views. Data views promote the need to have clear input and outputs; through this research it has been shown that the BUM methodology produces a way to take large decoupled data-sets (input) as a way to produce a traffic crash metric (output). By understanding the input of data-sets additional information is gathered beyond the traffic crash metric. The data narratives described in this work is a form of data view that provides human readable text about a given traffic crash. Data narratives is an output of the data-sets that introduce additional ways for the same information to be viewed, interpreted for the dissemination of information. By providing critical information through a metric, the transformation of data to knowledge yields the transformation of people to become more knowledgeable about the occurrences on the road.

6.3 GOAL & OBJECTIVES

The BUM methodology and metric development conducted through this fulfilled the goal for this research. A discussion of how the objectives of this research is met is described below.

The goal of this research was to create a systematic approach for modeling publicly available mobility data-sets that enables: i) formal descriptions of information embedded in data, ii) quantitative and qualitative information extraction, and iii) knowledge discovery for decision making.

Objective 1 Create a modeling methodology to formally describe publicly available mobility data-sets using knowledge graphs for knowledge discovery

In this research, the BUM methodology was developed to formalize the transformation of data to knowledge by using a bottom-up modeling approach. This approach yielded the creation of a knowledge graph to represent the domain of traffic crashes in the State of Texas. The knowledge graphs were enhanced by semantic annotation to describe the relationships of data usage in this work to additional data that could be integrated. The use of publicly available mobility data sets provides the openness that is needed for linked open data; through semantics, the data is formally described for additional uses and dissemination.

Through the process of transforming data into usable knowledge, similar approaches were explored in the literature and designed techniques for cleaning and retrieving data efficiently. This was further shown in a process that took initial raw data and passed it through a parser to be mapped in the same form of the knowledge graph. The techniques designed for data retrieving,

parsing and cleaning meet standards for computation since they are transformed into a common usable form (JSON-LD). The BUM algorithm introduces a novel methodology that is accessible to accepted data formats, for the distribution and linkage of data.

The data was semantically annotated by introducing context tags to the JSON, thus transforming it into JSON-LD. The technique used to annotate public data was done through introducing a W3C accepted upper-level ontology that is linked to many other data-sets. Through the integration of traffic crash data to a provenance ontology, it can be easily mapped to other domains for additional usage. Moreover, knowledge graphs can be expanded to include additional ontologies and vocabularies for linked open data. The transformation of the raw data to knowledge is leveraged by the use of having data provide additional meaning.

In this research, a semantic-based data model was developed such that it described concepts and relationships that were developed by data. This was achieved through the implementation of a knowledge graph using data from traffic crashes throughout the State of Texas. Using publicly available data allows for knowledge graphs to be expanded beyond the original scope studied. It provides access to additional knowledge graphs so it can be part of linked data.

This model is a generic framework that is usable, comparable, and transferable throughout the entire State of Texas and provides a description of all of the used data for the entire state. This data model was achieved through the modeling of traffic data throughout the State of Texas.

Objective 2 Create a novel metric aiming to improve understanding of publicly available mobility data-sets by users regardless their level of domain expertise

In this research, novel metrics were developed such that they are measurable, sustainable, transferable, and comparable to describe traffic crashes; this was done with respect to crash details, the effect of the people involved, and external circumstances of the crash. The resulting metric was the Critical Composite Index. The metrics developed were driven by mobility data as a way of ensuring accuracy and creating a factual foundation. The representation of the CCI provides information that has gained knowledge as described by the user evaluation study. It has been shown that the metric provides more information than the current practice of frequency metric reporting. Commuters, researchers and domain experts are able to synthesize and disseminate information from the CCI, as shown in the user evaluation study.

The CCI is representative of all traffic crashes given in the data-set and provide insight for many stakeholders including, but not limited to, commuters, researchers, policymakers, non-domain and domain experts. The CCI provides users to data that is consistently available and provides ability to compare traffic crashes over a sustained period of time. Furthermore, these metrics can be reused and redistributed for multiple geographic locations that use similar data-sets because domain experts can provide their own weighted criteria and values for the Critical Composite Index. Moreover, the metric is reused and redistributed for use throughout the State of Texas, all of which use similar data. The knowledge gained from the CCI provides ability to disseminate information between geographic locations as a standard method of classifying and comparing traffic crashes.

The user evaluation survey shows that the CCI and its respective severity chart provides three major elements: Comprehension of the CCI, knowledge gain, and improvement from current frequency based metric practices. The results from the survey show that the respondents were

able to understand the CCI and gather new knowledge from it. The knowledge gained from the CCI was directly related to understanding actual traffic crash events and classifying them in a way that matched the CCI. Furthermore, discussions with experts at the Texas Department of Transportation (Hernandez, 2018) and the 911 call center in El Paso (Shields, 2018), domain experts have found the BUM methodology to be important and useful. Currently, TxDOT has a focus on improving traffic flow by not only locating traffic crashes but describing the crash based on the way it affects people around them (Hernandez, 2018). The local police in El Paso focuses on the safety of road users as well as understanding the severity of an crash and the way it affects people (Shields, 2018). All of the experts in this domain currently use subjective techniques to describe traffic crashes to the public.

As a result of having a lack of set standards for describing traffic crashes, this work provides a foundation to describe traffic crashes in the transportation domain. Although human subjectivity cannot be taken out of describing traffic crashes completely because it is necessary to provide weights to crash investigation criteria, this work provides a means to standardize the subjectivity amongst users across the city, state, country by adopting this model.

Chapter 7: Conclusions & Future Work

7.1 Conclusions

To understand data, it is critical that it is analyzed in a clear way. The development of techniques to understand data is necessary as a mechanism for people to get useful knowledge out of that data. Without proper techniques to understand data, both non-domain and domain experts of the data domain may not be able to fully capture all that the data can provide to them. Through enhanced modeling techniques provided by the BUM methodology, data can easily be transformed into knowledge. Semantic annotation provides the insight needed to expand data-sets into linked data for possible usage of other research. Furthermore, that same data that has been transformed can be taken further and be used to describe events such as traffic crashes.

In this work, the need to develop a quantitative, comprehensive metric for publicly available mobility data was identified and applied to traffic crashes in the State of Texas. The proposed work created a metric to capture critical components of a traffic crash including the crash itself, the people involved, and the external circumstances of that crash that describe it deeper. The development of this metric began by developing the BUM methodology that took multiple large data-sets of traffic crashes.

The BUM methodology is a bottom-up iterative approach where each stage of the method could be refined based on analyzing the results gathered. The main steps of the method were data discovery, knowledge graph modeling, data processing, data mapping, and index development; it would be through the BUM methodology that a novel metric could then be computed.

The metric describes traffic crashes by taking information that is available (data discovery); modeling it by adding relationships to related concepts and adding semantics to it (knowledge graph modeling); processing the data to be cleaned and in a singular form (data processing); transforming it into a JSON-LD form that was stored in a NoSQL database for the answering of competency questions (data mapping); and finally developing a metric based on weighted criteria values to make each crash comparable to others regardless of location (index development). The process of using data for the development of a CCI is a metric that provides unique insight into traffic crashes. The developed metric solves problems in removing subjectivity in determining a crash metric. Moreover, it provides novel ways to communicate information to non-domain and domain experts.

Data plays a critical role in this research. Data should describe, diagnose, predict, and prescribe. This work tackles the problem of using traffic crash data as a method of describing. A single metric is capable of describing what happened in a traffic crash, but it is difficult to diagnose why it happened and predict what will happen. Allowing for implementation of data narratives, the data may begin to shed information on why the traffic crash happened, what will happen in the future and what actions should be taken to avoid future traffic crashes, based on this data.

As a result of the research done, the BUM methodology provides a way for researchers to expand modeling techniques to include modeling for alternative domains. Additionally, this work has shown the importance of using the CCI as a way to compare traffic crashes beyond a single geographic location. Through the transformation of data to knowledge that was discovered by the development of the CCI, it is possible to begin expanding how highway safety, personal

safety, and movement is affected by traffic crashes. Results show that the CCC metric is a useful way in communicating information to people with respect to traffic crashes.

This work highlights the importance of interdisciplinary research with a foundation of Computer Science in combination with another domain as an area of Data Science. Furthermore, this work has given increased knowledge to the area of Smart Mobility, a Smart Cities research focus area. In practice, the CCI along with relevant competency questions can be transferred to cities to use as a way to standardize the classification, comparison, and analysis of traffic crashes. The knowledge gained from this research will also allow for city users to take part in the understanding and dissemination of knowledge based on guidelines set by the CCI. Through this work it has been shown that bottom-up modeling is critical for the transformation of data to knowledge in Smart Cities research; it provides a way to gain insight to real-world problems that exist and provide solutions that will improve the *four principles of Smart Cities* proposed in this work.

7.2 LIMITATIONS

The limitations with this research are having a small set of contextual standards to compare developed metrics. The number of metrics throughout the world are plentiful, however, they are not fully comprehensive in understanding traffic crashes. In the development of the metric, there are a number of traffic crash events that cause the normalized index to not always be representative such as outlier events from multiple fatality crashes. The CCI is suitable to provide a high-level understanding of each traffic crash that occurs on the road, however competency questions are still required to understand crash specific details.

The ontology expressed in Figure 3.17 was applied to real crash data to demonstrate as a proof-of-concept the use of standard reasoning services such as consistency checking and inferences. However, due current capabilities of available APIs (i.e., OWLAPI (Manchester, 2011)) to ingest generic JSON-LD files, there was not a full programmatic implementation of all JSON-LD traffic crash individuals into the ontology for extended use. This can be further explored to take full advantage of semantic annotations provided in this work.

7.3 FUTURE WORK

This work can be extended by providing a more comprehensive set of data narratives for diverse groups of stakeholders as well as integrating more data related to traffic crashes. Data narratives provide human readable context to data through detailed descriptions of data. The CCI provides a depth of knowledge gain by describing traffic crashes with respect to a value that is mapped to a severity chart. Moreover, the CCI can also be mapped to a generalized description of each severity range. The CCI is unable to describe each traffic crash with detailed descriptions of how the value was determined. Additional research would be done in data narratives to discover how a standard classification technique used in this research can be expanded to incorporate human readable information to gain additional information from the initial data. Each traffic crash would have human readable text to describe it, similar to the usage in the case studies of this work. For example, a traffic crash may be described as: Crash_ID: 15575237 was a minor crash occurring in La Porte, TX which is just outside Houston, TX. This crash occurred on Sunday, January 2, 2017, at 3:04 am. The crash occurred when a vehicle struck a fixed object in the roadway. The reported crash had more than $1000 in damages, no fatalities or injuries with only one person involved. The weather conditions were reported to be foggy in a dark, but lighted area. In addition, work will be done to include visual guidance in the form of photographs of the

traffic accident to provide additional context to commuters. By providing human readable text of the crash, additional information can be gathered for many stakeholders.

Additional work can be done by using the data-sets in this research as a basis for machine learning classification. One of the common machine learning algorithms that could be used as part of the classification of traffic crash data sets is using K-means. K-means is a standard unsupervised learning clustering algorithm (Zhexue Huang, 1998) that can be used to cluster the data in this research into the four levels of severity: minor, moderate, major, and severe. The algorithm would take a sample of the data-set and begin to cluster the data into four different classifications of traffic crashes. From the resulting clusters, additional analysis and prediction can be done on future data-sets. Using the data-sets, additional research can be done to predict the severity of traffic crashes based on scenarios or additional data given as an input.

Additional work can be done to begin expansion of the BUM methodology to additional domain areas outside of traffic crashes. The BUM methodology is intended to be a standard technique to describe multiple domains and gather similar results. The BUM methodology is not domain specific and can be used as a framework to understand other data-sets that have not yet been explored. The data represented through the BUM methodology can be expressed in many ways, however, this methodology standardizes the approach in which such data is used then disseminated. Additionally, customized routes based on traffic crash severity, location safety, and a combination of data-sets can be used to promote more efficient & productive travel for users in a city.

References

(PennDOT), P. D. of T. (2019). Pennsylvania Department of Transportation PennDOT Data Share. Retrieved April 11, 2019, from https://pennshare.maps.arcgis.com/apps/webappviewer/index.html?id=8fdbf046e36e41649bbfd9d7dd7c7e7e

[TABC], T. A. B. C. (2019). Blood Alcohol Percentage Chart. Retrieved April 4, 2019, from https://www.tabc.state.tx.us/enforcement/blood_alcohol_percentage_chart.asp

Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Systems with Applications*, *40*(15), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027

Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, *6*(1), 1–15. https://doi.org/10.1016/j.compchemeng.2017.02.028

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, *22*(1), 3–21. https://doi.org/10.1080/10630732.2014.942092

Anthopoulos, L. G., Janssen, M., & Weerakkody, V. (2015). Comparing Smart Cities with different modeling approaches. *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, *1997*, 525–528. https://doi.org/10.1145/2740908.2743920

Arenas, M., Cuenca Grau, B., Kharlamov, E., Sar Unas Marciuška, ˇ, & Zheleznyakov, D. (2015). Faceted Search over RDF-Based Knowledge Graphs. Retrieved from https://www.cs.ox.ac.uk/files/8303/main.pdf

Azzaoui, K., Jacoby, E., Senger, S., Rodríguez, E. C., Loza, M., Zdrazil, B., … Ecker, G. F. (2013). Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today*, *18*(17–18), 843–852. https://doi.org/10.1016/j.drudis.2013.05.008

Babbie, E. R. (2012). *The Practice of Social Research* (13th ed.). Wadsworth, Cengage Learning. Retrieved from https://books.google.com/books?id=k-aza3qSULoC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Bass, L., & John, B. E. (2003). Linking usability to software architecture patterns through general scenarios. *Journal of Systems and Software*, *66*(3), 187–197. https://doi.org/10.1016/S0164-1212(02)00076-6

Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zhao, J., … McGuinness, D. (2013). PROV-O: The PROV Ontology W3C. *W3C Recommendation*, (April 2013), 153.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). THE SEMANTIC WEB. *Scientific American*, *284*(5), 34–43.

Berners-Lee, T., & W3C. (2009). LinkedData, 1–5.

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data Driven Ontology Evaluation. Retrieved from https://eprints.soton.ac.uk/259062/

Buccella, A., Cechich, A., & Rodriguez Brisaboa, N. (2003). An ontology approach to data integration. *Journal of Computer Science & Technology*, *3*(2), 62–68.

Caragliu, A., del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology*, *18*(2), 65–82. https://doi.org/10.1080/10630732.2011.601117

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, *14*(1), 20–26. https://doi.org/10.1109/5254.747902

Cheu, R. L., & Balal, E. (2018). Development of a Comprehensive Metric for Transportation , Environment , and Community Health.

Cosgrave, E., Arbuthnot, K., & Tryfonas, T. (2013). Living labs, innovation districts and information marketplaces: A systems approach for smart cities. *Procedia Computer Science*, *16*(Cser 13), 668–677. https://doi.org/10.1016/j.procs.2013.01.070

Cruz, I. F., & Xiao, H. (2005). The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, *13*(4), 1–18. https://doi.org/10.1.1.60.4933

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation 25 February 2014*, (February), 263–270. https://doi.org/10.1007/s13398-014-0173-7.2

Dameri, R. P. (2013). Searching for Smart City definition: a comprehensive proposal. *International Journal of Computers & Technology*, *11*(5), 2544–2551. https://doi.org/10.1007/s13132-012-0084-9

Dameri, R. P. (2014). Searching for Smart City definition: a comprehensive proposal. *International Journal of Computers & Technology*, *11*(5), 2146–2161.

Dardailler, D. (2012). Road Accident Ontology. *W3C Note*, 1–11.

Darema, F. (2004). Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements. In M. Bubak, G. D. van Albada, P. M. A. Sloot, & J. Dongarra (Eds.), *Computational Science - ICCS 2004* (pp. 662–669). Berlin, Heidelberg: Springer Berlin Heidelberg.

De Oña, J., López, G., & Abellán, J. (2013). Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention*, *50*, 1151–1160. https://doi.org/10.1016/j.aap.2012.09.006

Dobre, C., & Xhafa, F. (2014). Intelligent services for Big data science. *Future Generation Computer Systems*, *37*, 267–281. https://doi.org/10.1016/j.future.2013.07.014

Europe, R. E. C. of I. (2008). Research Evaluation for Computer Science. *Research Evaluation Committee of Informatics Europe*, (May), 1–17.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, *17*(3). https://doi.org/doi.org/10.1609/aimag.v17i3.1230

Fox, M. S. (2015). A Foundation Ontology for Global City Indicators, (May 2015), 1–37.

Giffinger, R. (2007). Smart cities Ranking of European medium-sized cities. *October*, *16*(October), 13–18. https://doi.org/10.1016/S0264-2751(98)00050-X

Gil, Y., & Garijo, D. (2017). Towards Automating Data Narratives. *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17*, (February), 565–576. https://doi.org/10.1145/3025171.3025193

Gitelman, V., Doveh, E., & Hakkert, S. (2010). Designing a composite indicator for road safety. *Safety Science*, *48*(9), 1212–1224. https://doi.org/10.1016/j.ssci.2010.01.011

Google. (2018). Google Maps Platform. Retrieved from https://developers.google.com/maps/documentation/

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., … Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management*, *36*(5), 748–758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

Hermans, E., Brijs, T., Wets, G., & Vanhoof, K. (2009). Benchmarking road safety: Lessons to learn from a data envelopment analysis. *Accident Analysis and Prevention*, *41*(1), 174–182. https://doi.org/10.1016/j.aap.2008.10.010

Hermans, E., Van den Bossche, F., & Wets, G. (2009). Uncertainty assessment of the road safety index. *Reliability Engineering and System Safety*, *94*(7), 1220–1228. https://doi.org/10.1016/j.ress.2008.09.004

Hernandez-Moreno, S., & De Hoyos-Martinez, J. (2010). Indicators of urban sustainability in Mexico.(Report). *Theoretical and Empirical Researches in Urban Management*, (16), 46.

Hernandez, A. (2018). Transvista-ELP-TMC Interview. El Paso.

Hey, T. (2009). *The Fourth Paradigm*. (T. Hey, S. Tansley, & K. Tolle, Eds.) (1st ed.). Redmond, Washington: Microsoft Research. Retrieved from https://www.amazon.com/Fourth-Paradigm-Data-Intensive-Scientific-ebook/dp/B00318D9Y2/ref=sr_1_4?ie=UTF8&s=books&qid=1261412355&sr=8-4#reader_B00318D9Y2

Hiremath, R. B., Balachandra, P., Kumar, B., Bansode, S. S., & Murali, J. (2013). Indicator-based urban sustainability-A review. *Energy for Sustainable Development*, *17*(6), 555–563.

https://doi.org/10.1016/j.esd.2013.08.004

Hoornweg, D., Nunez, F., Palugyai, N., Villaveces, M., & Longfellow, H. W. (2007). City Indicators: Now to Nanjing. *World Bank Working Paper*, (JANUARY 2007), 1–71.

Information Systems Group Oxford University. (2016). HermiT OWL Reasoner The New Kid on the OWL Block. Retrieved April 1, 2016, from http://www.hermit-reasoner.com

Institute, B. S. (2014). PAS 180:2014 Smart cities – Vocabulary. *Pas*, 38.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., & Sheth, A. P. (2010). Linked Data is Merely More Data. *Linked Data Meets Artificial Intelligence. Technical Report SS-10-07, AAAI Press*, 82–86. Retrieved from http://knoesis.wright.edu/library/publications/linkedai2010_submission_13.pdf

Jones, T., Baxter, M., & Khanduja, V. (2013). A quick guide to survey research. *The Annals of The Royal College of Surgeons of England*, *95*(1), 5–7. https://doi.org/10.1308/003588413X13511609956372

Khazei, K., & Fox, M. S. F. (2017). A Public Safety Ontology for Global City Indicators (ISO37120). Retrieved from http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-PublicSafety-Ontology-28apr2017.pdf

Klein, J. T. (2008). Evaluation of Interdisciplinary and Transdisciplinary Research. *American Journal of Preventive Medicine*, *35*(2), S116–S123. https://doi.org/10.1016/j.amepre.2008.05.010

Krosnick, J. a. (1999). SURVEY RESEARCH. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Larios, V. M., Gomez, L., Mora, O. B., Maciel, R., & Villanueva-Rosales, N. (2016). Living labs for smart cities: A use case in Guadalajara City to foster innovation and develop citizen-centered solutions. *IEEE 2nd International Smart Cities Conference: Improving the Citizens Quality of Life, ISC2 2016 - Proceedings*. https://doi.org/10.1109/ISC2.2016.07580773

Laureshyn, A., Svensson, Å., & Hydén, C. (2010). Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis and Prevention*, *42*(6), 1637–1646. https://doi.org/10.1016/j.aap.2010.03.021

Lazaroiu, G. C., & Roscia, M. (2012). Definition methodology for the smart cities model. *Energy*, *47*(1), 326–332. https://doi.org/10.1016/j.energy.2012.09.028

Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV Ontology. *W3C Recommendation*, (April), 1–80. Retrieved from http://www.w3.org/TR/2013/REC-prov-o-20130430/

Li, D., Cao, J., & Yao, Y. (2015). Big data in smart cities. *Science China Information Sciences*, *58*(10), 1–12. https://doi.org/10.1007/s11432-015-5396-5

Lim, C., Kim, K.-J., & Maglio, P. P. (2018). Smart cities with big data: Reference models,

challenges, and considerations. *Cities*, *82*, 86–99. https://doi.org/https://doi.org/10.1016/j.cities.2018.04.011

Ma, Z., Shao, C., Ma, S., & Ye, Z. (2011). Constructing road safety performance indicators using fuzzy delphi method and grey delphi method. *Expert Systems with Applications*, *38*(3), 1509–1514. https://doi.org/10.1016/j.eswa.2010.07.062

Manchester, U. of. (2011). The OWL API. Retrieved from http://owlapi.sourceforge.net

Massie, D. L., Campbell, K. L., & Williams, A. F. (1995). Traffic Accident involvement rates by driver age and gender. *Accident Analysis & Prevention*, *27*(1), 73–87. https://doi.org/10.1016/0001-4575(94)00050-V

McMahon, S. K. (2002). The development of quality of life indicators—a case study from the City of Bristol, UK. *Ecological Indicators*, *2*(1–2), 177–185. https://doi.org/10.1016/S1470-160X(02)00039-0

Meadows, D. (1998). Indicators and information systems for sustainable development. *A Report to the Balaton Group*, 1–25. Retrieved from https://www.iisd.org/pdf/s_ind_2.pdf%0Ahttps://www.iisd.org/pdf/s_ind_2.pdf%5Cnhttp://www.biomimicryguild.com/alumni/documents/download/Indicators_and_information_systems_for_sustainable_develoment.pdf

Mejia, D. (2017). *Integration of Heterogeneous Traffic Data to Address Mobility Challenges In the City of El Paso*. The University of Texas at El Paso.

Meyer, B., Choppy, C., Staunstrup, J., & van Leeuwen, J. (2009). Viewpoint Research evaluation for computer science. *Communications of the ACM*, *52*(4), 31. https://doi.org/10.1145/1498765.1498780

Mihyeon Jeon, C., & Amekudzi, A. (2005). Addressing Sustainability in Transportation Systems: Definitions, Indicators, and Metrics. *Journal of Infrastructure Systems*, *11*(1), 31–50. https://doi.org/10.1061/(ASCE)1076-0342(2005)11:1(31)

MongoDB. (2018). MongoDB. Retrieved December 4, 2018, from https://www.mongodb.com/

Mori, K., & Yamashita, T. (2015). Methodological framework of sustainability assessment in City Sustainability Index (CSI): A concept of constraint and maximisation indicators. *Habitat International*, *45*(P1), 10–14. https://doi.org/10.1016/j.habitatint.2014.06.013

Müller, D., Reichert, M., & Herbst, J. (2007). Data-Driven Modeling and Coordination of Large Process Structures. In R. Meersman & Z. Tari (Eds.), *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS* (pp. 131–149). Berlin, Heidelberg: Springer Berlin Heidelberg.

Musen, M. A. (2015). The protégé project. *AI Matters*, *1*(4), 4–12. https://doi.org/10.1145/2757001.2757003

Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., & Morris, R. (2011). Smarter Cities and Their Innovation Challenges. *Computer*, *44*(6), 32–39.

https://doi.org/10.1109/MC.2011.187

National Safety Council, & ANSI. (2017). American National Standard: Manual on Classification of Motor Vehicle Traffic Crashes. *Association of Transpotation Safety Information Professionals*.

National Statistical Institute of Italy. (2001). Enviornmental Sustainability Indicators in Urban Areas: An Italian Experience. *Joint ECE/Eurostat Work Session on Methodological Issues of Environment Statistics*, (6 November 2006), 1–15. Retrieved from http://www.unece.org/stats/documents/2001/10/env/wp.16.e.pdf

Neirotti, P., De Marco, A., Cagliano, A. C., Mangano, G., & Scorrano, F. (2014). Current trends in Smart City initiatives: Some stylised facts. *Cities*, *38*, 25–36. https://doi.org/10.1016/j.cities.2013.12.010

NIST/SEMATECH. (2013). e-Handbook of Statistical Methods. Retrieved February 19, 2019, from http://www.itl.nist.gov/div898/handbook/

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory*, 25. https://doi.org/10.1016/j.artmed.2004.01.014

O'Rourke, L., Beshers, E., & Stock, D. (2015). *Measuring the Impacts of Freight Transportation Improvements on the Economy and Competitiveness September 2015*.

Poetz, M. K., & Schreier, M. (2012). The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas? *Journal of Product Innovation Management*, *29*(2), 245–256. https://doi.org/10.1111/j.1540-5885.2011.00893.x

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Python Software Foundation (PSF). (2019). Python. Retrieved March 31, 2019, from https://www.python.org/

QuestionPro. (2019). Retrieved February 1, 2019, from https://www.questionpro.com/

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. In *Data Engineering* (Vol. 23, pp. 3–13). New York, New York, USA: IEEE Comput. Soc. https://doi.org/10.1145/1317331.1317341

Roess, R. P., Prassas, E. S., & McShane, W. R. (2011). *Traffic Engineering* (4th ed.). Pearson.

Sabatier, P. A. . (1986). Top-down and Bottom-up Approaches to Implementation Research : A Critical Analysis and Suggested Synthesis. *Cambridge University Press Stable*, *6*(1), 21–48.

Schuurman, D., Baccarne, B., De Marez, L., & Mechant, P. (2012). Smart ideas for smart cities: Investigating crowdsourcing for generating and selecting ideas for ICT innovation in a city context. *Journal of Theoretical and Applied Electronic Commerce Research*, *7*(3), 49–62. https://doi.org/10.4067/S0718-18762012000300006

Shields, T. (2018). 911 Call Center. El Paso.

Sloman, A. (2016). Types of Research in Computing Science Software Engineering and Artificial Intelligence.

Software, L. (2013). Comparison and Analysis.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). JSON-LD 1.0.

Spyns, P., Meersman, R., & Jarrar, M. (2000). STAR Lab Technical Report Data Modelling versus Ontology Engineering Data modelling versus Ontology engineering, *31*(4).

Steele, R. (2014). ISO 37120 standard on city indicators – how they help city leaders set tangible targets, including service quality and quality of life, (October).

Tapadinhas, J., & Gartner. (2014). *Business analytics from basics to value*.

Texas Department of Transportation. (2016). *Classification of Motor Vehicle Traffic Crashes in Texas*.

Texas Department of Transportation. (2017). Classification of Motor Vehicle Traffic Crashes In Texas.

Torres, E. J. (2016). *Ontology-Driven Integration of Data for Freight Performance Measures*. The University of Texas at El Paso.

TxDOT. (2017). Values, Vision, Mission and Goals. Retrieved December 19, 2017, from http://www.txdot.gov/inside-txdot/contact-us/mission.html

TxDOT. (2018). Crash Records Information System. Retrieved January 15, 2018, from https://cris.dot.state.tx.us

U.S. General Services Administration, T. T. S. (2019). DATA.GOV. Retrieved January 31, 2019, from https://www.data.gov/

UN. (2018). *World Urbanization Prospects: The 2018 Revision*.

United States Department of Transportation (USDOT). (2016). Research, Development, and Technology Strategic Plan FY 2017-2021, (December). Retrieved from https://ntlsearch.bts.gov/researchhub/strategicplan.do

Vila, J. J. R., Kozievitch, N. P., Gadda, T. M. C., Fonseca, K. V. O., Rosa, M. O., Gomes-Jr, L. C., & Akbar, M. (2016). Urban Mobility Challenges – An Exploratory Analysis of Public Transportation Data in Curitiba. *Revista de Informát Ica Aplicada*, *12*(1), 1–14. https://doi.org/10.13037/RAS.VOL12N1.145

Vlacheas, P., Giaffreda, R., Stavroulaki, V., Kelaidonis, D., Foteinos, V., Poulios, G., … Moessner, K. (2013). Enabling smart cities through a cognitive management framework for the internet of things. *IEEE Communications Magazine*, *51*(6), 102–111. https://doi.org/10.1109/MCOM.2013.6525602

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

World Health Organization. (2004). World Report on Road Traffic Injury Prevention. *Injury Prevention*, *10*(4), 255–256. https://doi.org/10.1016/j.puhe.2005.09.003

Yigitcanlar, T., & Dur, F. (2010). Developing a sustainability assessment model: The sustainable infrastructure, Land-use, environment and transport model. *Sustainability*, *2*(1), 321–340. https://doi.org/10.3390/su2010321

Zhexue Huang. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/https://doi.org/10.1023/A:1009769707641

Ziegler, P., & Dittrich, K. R. (2007). Data Integration — Problems, Approaches, and Perspectives. In *Conceptual Modelling in Information Systems Engineering* (pp. 39–58). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72677-7_3

Appendix

APPENDIX A – REMOVED DATA COLUMNS

Removed Data Columns	Removed Data Columns2	Removed Data Columns3	Removed Data Columns4
Medical_Advisory_Fl	Street_Nbr	Shldr_Use_Right_ID	Poscrossing_ID
Amend_Supp_Fl	Control	Median_Type_ID	WDCode_ID
Case_ID	Section	Median_Width	Standstop
Local_Use	Milepoint	Rural_Urban_Type_ID	Yield
Rpt_Latitude	Ref_Mark_Nbr	Func_Sys_ID	MPO_ID
Rpt_Longitude	Ref_Mark_Displ	Adt_Curnt_Amt	Investigat_Service_ID
Rpt_Hwy_Sfx	Hwy_Sys_2	Adt_Curnt_Year	Investigat_DA_ID
Rpt_Street_Pfx	Hwy_Nbr_2	Adt_Adj_Curnt_Amt	Investigator_Narrative
Rpt_Street_Sfx	Hwy_Sfx_2	Pct_Single_Trk_Adt	Investigat_Notify_Meth
Private_Dr_Fl	Street_Name_2	Pct_Combo_Trk_Adt	Rpt_Sec_Hwy_Sfx
Rpt_Street_Desc	Street_Nbr_2	Trk_Aadt_Pct	
Rpt_Sec_Street_Pfx	Control_2	Curve_Type_ID	
Rpt_Sec_Street_Sfx	Section_2	Curve_Lngth	
Rpt_Ref_Mark_Offset_Amt	Milepoint_2	Cd_Degr	
Rpt_Ref_Mark_Dist_Uom	Txdot_Rptable_Fl	Delta_Left_Right_ID	
Rpt_Ref_Mark_Dir	Onsys_Fl	Dd_Degr	
Rpt_Ref_Mark_Nbr	Hwy_Dsgn_Lane_ID	Feature_Crossed	
Rpt_Sec_Street_Desc	Hwy_Dsgn_Hrt_ID	Structure_Number	
Rpt_CrossingNumber	Hp_Shldr_Left	I_R_Min_Vert_Clear	
Road_Algn_ID	Hp_Shldr_Right	Approach_Width	
Investigat_Comp_Fl	Hp_Median_Width	Bridge_Median_ID	
ORI_Number	Base_Type_ID	Bridge_Loading_Type_ID	
Investigat_Agency_ID	Nbr_Of_Lane	Bridge_Loading_In_1000_Lbs	
Investigat_Area_ID	Row_Width_Usual	Bridge_Srvc_Type_On_ID	
Investigat_District_ID	Roadbed_Width	Bridge_Srvc_Type_Under_ID	

Investigat_Region_ID	Surf_Width	Culvert_Type_ID	
Road_Part_Adj_ID	Curb_Type_Left_ID	Roadway_Width	
Hwy_Sys	Curb_Type_Right_ID	Deck_Width	
Hwy_Nbr	Shldr_Type_Left_ID	Bridge_Dir_Of_Traffic_ID	
Hwy_Sfx	Shldr_Width_Left	Bridge_Rte_Struct_Func_ID	
Dfo	Shldr_Use_Left_ID	Bridge_IR_Struct_Func_ID	
Street_Name	Shldr_Type_Right_ID	CrossingNumber	
	Shldr_Width_Right	RRCo	

APPENDIX B – RELATIONSHIP TABLE

Entity (Subject)	Relationship (Predicate)	Datapoint/Entity (Object)
METRIC	*measures*	CRASH
CRASH	isAt	LOCATION
CRASH	has	INVESTIGATION
CRASH	involves	PERSON
CRASH	identifiedBy	Crash_ID
LOCATION	isAt	Longitude
LOCATION	isAt	Latitude
LOCATION	isAt	County_ID
LOCATION	describedBy	Surface_Type_ID
LOCATION	describedBy	Crash_Speed_Limit
LOCATION	describedBy	Toll_Road_Fl
LOCATION	describedBy	Report_Street_Name
LOCATION	describedBy	Report_Block_Number
LOCATION	describedBy	Report_Road_Part_ID
LOCATION	describedBy	Report_Secondary_Street_Name
LOCATION	describedBy	Report_Secondary_Roadway_Sys_ID
LOCATION	describedBy	Report_Highway_Number
LOCATION	describedBy	Report_Roadway_Sys_ID
LOCATION	describedBy	Report_Outside_City_Limit_Fl
LOCATION	describedBy	Report_City_ID
LOCATION	describedBy	Report_CRIS_County_ID
LOCATION	describedBy	Road_Type_ID
LOCATION	describedBy	Entrance_Road_ID
INVESTIGATION	reports	Road_Related_ID
INVESTIGATION	reports	Road_Closure_ID
INVESTIGATION	reports	Bridge_Detail_ID
INVESTIGATION	reports	At_Intersection_Fl
INVESTIGATION	reports	Traffic_Control_ID
INVESTIGATION	reports	Investigation_Arrival_Time
INVESTIGATION	reports	Investigation_Notify_Time
INVESTIGATION	reports	Located_Fl
INVESTIGATION	reports	Light_Condition_ID

INVESTIGATION	reports	Weather_Condition_ID
INVESTIGATION	reports	Thousand_Damage_Fl
INVESTIGATION	reports	Crash_Time
INVESTIGATION	reports	Active_School_Zone_Fl
INVESTIGATION	reports	Crash_Date
INVESTIGATION	reports	RR_Related_Fl
INVESTIGATION	reports	School_Bus_Fl
INVESTIGATION	reports	Cmv_Involved_Fl
INVESTIGATION	reports	Crash_Fatal_Fl
INVESTIGATION	reports	Total_Injury_Count
INVESTIGATION	reports	Death_Count
INVESTIGATION	reports	Unknown_Injury_Count
INVESTIGATION	reports	Non_Injury_Count
INVESTIGATION	reports	Possible_Injury_Count
INVESTIGATION	reports	Serious_Injury_Count
INVESTIGATION	reports	Day_of_Week
INVESTIGATION	reports	Population_Group_ID
INVESTIGATION	reports	Crash_Severity_ID
INVESTIGATION	reports	Rural_Fl
INVESTIGATION	reports	Physical_Feature2_ID
INVESTIGATION	reports	Physical_Feature1_ID
INVESTIGATION	reports	Other_Factor_ID
INVESTIGATION	reports	Oject_Struck_ID
INVESTIGATION	reports	FirstHarmfulEvent_Crash_ID
INVESTIGATION	reports	Interection_Related_ID
INVESTIGATION	reports	Harmful_Event_ID
INVESTIGATION	reports	Report_Date
INVESTIGATION	reports	Surface_Condition_ID
INVESTIGATION	reports	Road_Construction_Worker_Fl
INVESTIGATION	reports	Road_Construction_Zone_Fl
WEATHER	isRepresentativeOF	Weather_Condition_ID
WEATHER	*affects*	*CRASH*
PERSON	contains	INJURY
PERSON	isA	DRIVER
PERSON	has	Person_Death_Time
PERSON	has	Person_SOL_ID

PERSON	has	Person_Helmet_ID
PERSON	has	Person_Airbag_ID
PERSON	has	Person_restraint_ID
PERSON	has	Person_Ejected_ID
PERSON	identifiedBy	Person_Gender_ID
PERSON	identifiedBy	Person_Ethnicity_ID
PERSON	identifiedBy	Person_Age
PERSON	identifiedBy	Person_Injury_Severity_ID
PERSON	identifiedBy	Person_Occupant_Position_ID
PERSON	identifiedBy	Person_Type_ID
PERSON	identifiedBy	Unit_Number
PERSON	identifiedBy	Person_Number
DRIVER	has	Person_Alcohol_Result_ID
DRIVER	has	Person_Alcohol_Specimen_Type_ID
DRIVER	has	Person_BAC_Test_ID
DRIVER	has	Driver_ZipCode
DRIVER	has	Driver_Drug_Category1_ID
DRIVER	has	Driver_License_State_ID
DRIVER	has	Person_Drug_Result
DRIVER	has	Person_Drug_Specimen_Type_ID
DRIVER	has	Driver_License_Class_ID
DRIVER	has	Driver_License_Type_ID
INJURY	has	Serious_Injury
INJURY	has	Nonincapacitating_Injury
INJURY	has	Possible_Injury
INJURY	has	Non_Injury
INJURY	has	Unknown_Injury
INJURY	has	Person_Total_Injury_Count
INJURY	has	Death_Fl

APPENDIX C – SAMPLE WEIGHTS

Criteria	Value	Weight	Weight 2	Description
Crash_Fatal_Fl	0	**0**	**90**	NO
Crash_Fatal_Fl	1	**80**	**94**	YES
Cmv_Involv_Fl	0	**0**	**90**	NO
Cmv_Involv_Fl	1	**50**	**93**	YES
Schl_Bus_Fl	0	**0**	**97**	NO
Schl_Bus_Fl	1	**50**	**97**	YES
Rr_Relat_Fl	0	**0**	**87**	NO
Rr_Relat_Fl	1	**50**	**91**	YES
Active_School_Zone_Fl	0	**0**	**94**	NO
Active_School_Zone_Fl	1	**50**	**88**	YES
Crash_Time	0:00	**30**	**93**	Midnight-03:59
Crash_Time	4:00	**30**	**98**	04:00 - 05:59
Crash_Time	6:00	**60**	**89**	06:00 - 08:59
Crash_Time	9:00	**50**	**85**	09:00 - 11:59
Crash_Time	12:00	**45**	**85**	12:00 - 14:59
Crash_Time	15:00	**60**	**88**	15:00 - 17:59
Crash_Time	18:00	**50**	**90**	18:00 - 20:59
Crash_Time	21:00	**40**	**87**	21:00 - 23:59
Thousand_Damage_Fl	0	**0**	**86**	NO
Thousand_Damage_Fl	1	**50**	**94**	YES
Rpt_Road_Part_ID	0	**40**	**91**	OTHER ROAD TYPE
Rpt_Road_Part_ID	1	**40**	**100**	2 LANE, 2 WAY
Rpt_Road_Part_ID	2	**40**	**94**	4 OR MORE LANES, DIVIDED
Rpt_Road_Part_ID	3	**50**	**91**	4 OR MORE LANES, UNDIVIDED
Rpt_Road_Part_ID	94	**0**	**95**	REPORTED INVALID
Rpt_Road_Part_ID	95	**0**	**88**	NOT REPORTED
Road_Constr_Zone_Fl	0	**0**	**92**	NO
Road_Constr_Zone_Fl	1	**50**	**95**	YES
Road_Constr_Zone_Wrkr_Fl	0	**0**	**87**	NO
Road_Constr_Zone_Wrkr_Fl	1	**50**	**97**	YES
At_Intrsct_Fl	0	**0**	**91**	NO
At_Intrsct_Fl	1	**50**	**91**	YES
Wthr_Cond_ID	0	**0**	**97**	UNKNOWN
Wthr_Cond_ID	1	**10**	**87**	DRY

Wthr_Cond_ID	2	**75**	**97**	RAIN
Wthr_Cond_ID	3	**80**	**86**	SLEET/HAIL
Wthr_Cond_ID	4	**90**	**99**	SNOW
Wthr_Cond_ID	5	**90**	**85**	FOG
Wthr_Cond_ID	6	**80**	**86**	BLOWING SAND/SNOW
Wthr_Cond_ID	7	**40**	**85**	SEVERE CROSSWINDS
Wthr_Cond_ID	8	**10**	**96**	OTHER (EXPLAIN IN NARRATIVE)
Wthr_Cond_ID	11	**0**	**90**	CLEAR
Wthr_Cond_ID	12	**5**	**91**	CLOUDY
Wthr_Cond_ID	94	**0**	**91**	REPORTED INVALID
Wthr_Cond_ID	95	**0**	**96**	NOT REPORTED
Light_Cond_ID	0	**0**	**90**	UNKNOWN
Light_Cond_ID	1	**10**	**87**	DAYLIGHT
Light_Cond_ID	2	**20**	**88**	DAWN
Light_Cond_ID	3	**70**	**86**	DARK, NOT LIGHTED
Light_Cond_ID	4	**50**	**87**	DARK, LIGHTED
Light_Cond_ID	5	**20**	**99**	DUSK
Light_Cond_ID	6	**60**	**86**	DARK, UNKNOWN LIGHTING
Light_Cond_ID	8	**0**	**92**	OTHER (EXPLAIN IN NARRATIVE)
Light_Cond_ID	94	**0**	**100**	REPORTED INVALID
Light_Cond_ID	95	**0**	**88**	NOT REPORTED
Surf_Cond_ID	0	**0**	**93**	UNKNOWN
Surf_Cond_ID	1	**10**	**99**	DRY
Surf_Cond_ID	2	**80**	**86**	WET
Surf_Cond_ID	3	**70**	**90**	STANDING WATER
Surf_Cond_ID	5	**80**	**89**	SLUSH
Surf_Cond_ID	6	**90**	**100**	ICE
Surf_Cond_ID	8	**0**	**99**	OTHER (EXPLAIN IN NARRATIVE)
Surf_Cond_ID	9	**85**	**96**	SNOW
Surf_Cond_ID	10	**80**	**94**	SAND, MUD, DIRT
Surf_Cond_ID	94	**0**	**98**	REPORTED INVALID
Surf_Cond_ID	95	**0**	**96**	NOT REPORTED
Harm_Evnt_ID	1	**90**	**92**	PEDESTRIAN
Harm_Evnt_ID	2	**85**	**93**	MOTOR VEHICLE IN TRANSPORT

Harm_Evnt_ID	3	**90**	**88**	RR TRAIN
Harm_Evnt_ID	4	**85**	**96**	PARKED CAR
Harm_Evnt_ID	5	**90**	**94**	PEDALCYCLIST
Harm_Evnt_ID	6	**30**	**94**	ANIMAL
Harm_Evnt_ID	7	**30**	**88**	FIXED OBJECT
Harm_Evnt_ID	8	**30**	**96**	OTHER OBJECT
Harm_Evnt_ID	9	**30**	**86**	OTHER NON COLLISION
Harm_Evnt_ID	10	**90**	**90**	OVERTURNED
Harm_Evnt_ID	11	**0**	**95**	NOT REPORTED
Harm_Evnt_ID	93	**0**	**92**	UNDETERMINED
Harm_Evnt_ID	94	**0**	**88**	REPORTED INVALID
Sus_Serious_Injry_Cnt		**85**	**85**	COUNT
Nonincap_Injry_Cnt		**55**	**86**	COUNT
Poss_Injry_Cnt		**2**	**89**	COUNT
Non_Injry_Cnt		**0**	**94**	COUNT
Unkn_Injry_Cnt		**2**	**96**	COUNT
Death_Cnt		**95**	**87**	COUNT
MAX		**1169**	**1976**	

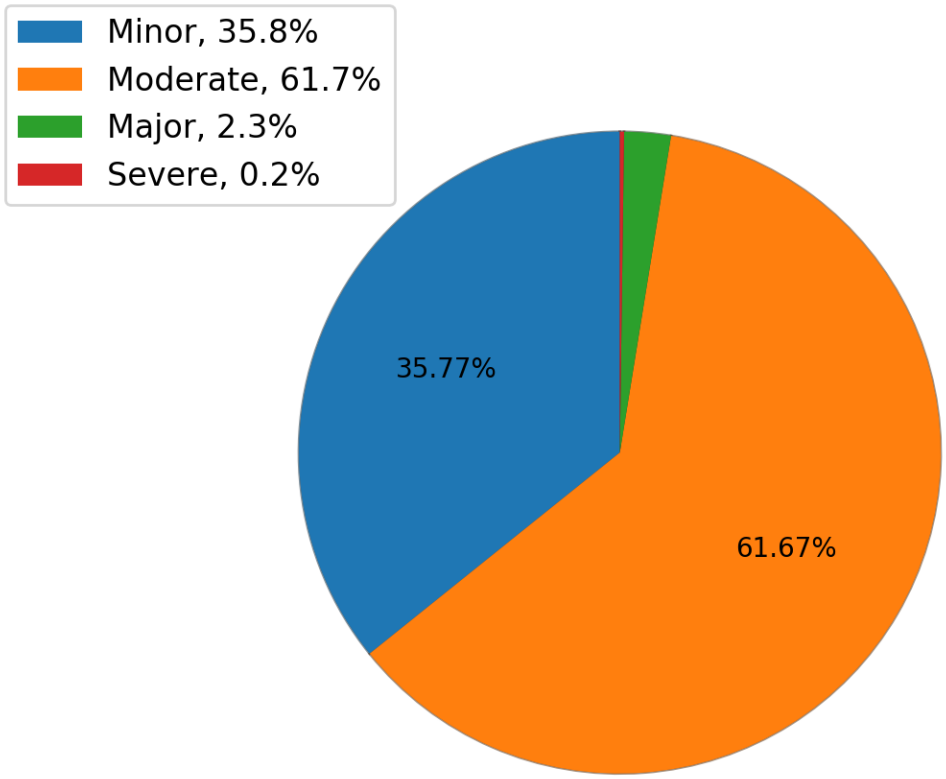
APPENDIX D – PENNSYLVANIA CRITERIA & WEIGHTS

Criteria	Value	Weight	Description
Crash_Fatal_Fl	0	0	NO
Crash_Fatal_Fl	1	80	YES
Cmv_Involv_Fl	0	0	NO
Cmv_Involv_Fl	1	50	YES
Schl_Bus_Fl	0	0	NO
Schl_Bus_Fl	1	50	YES
Active_School_Zone_Fl	0	0	NO
Active_School_Zone_Fl	1	50	YES
At_Intrsct_Fl	2	50	YES
At_Intrsct_Fl	3	50	YES
At_Intrsct_Fl	4	50	YES
At_Intrsct_Fl	5	50	YES
At_Intrsct_Fl	6	50	YES
At_Intrsct_Fl	7	50	YES
At_Intrsct_Fl	8	50	YES
At_Intrsct_Fl	0	0	NO
At_Intrsct_Fl	1	50	YES
Crash_Time	0:00	30	Midnight-03:59
Crash_Time	4:00	30	04:00 - 05:59
Crash_Time	6:00	60	06:00 - 08:59
Crash_Time	9:00	50	09:00 - 11:59
Crash_Time	12:00	45	12:00 - 14:59
Crash_Time	15:00	60	15:00 - 17:59
Crash_Time	18:00	50	18:00 - 20:59
Crash_Time	21:00	40	21:00 - 23:59
Road_Constr_Zone_Fl	0	0	NO
Road_Constr_Zone_Fl	1	50	YES
Wthr_Cond_ID	0	0	UNKNOWN
Wthr_Cond_ID	1	10	DRY
Wthr_Cond_ID	2	75	RAIN
Wthr_Cond_ID	3	80	SLEET/HAIL
Wthr_Cond_ID	4	90	SNOW
Wthr_Cond_ID	5	90	FOG
Wthr_Cond_ID	6	80	Rain & Fog
Wthr_Cond_ID	7	40	Sleet & Fog
Wthr_Cond_ID	8	10	OTHER (EXPLAIN IN NARRATIVE)

Wthr_Cond_ID	11	0	CLEAR
Wthr_Cond_ID	12	5	CLOUDY
Wthr_Cond_ID	94	0	REPORTED INVALID
Wthr_Cond_ID	95	0	NOT REPORTED
Light_Cond_ID	9	0	UNKNOWN
Light_Cond_ID	1	10	DAYLIGHT
Light_Cond_ID	5	20	DAWN
Light_Cond_ID	2	70	DARK, NOT LIGHTED
Light_Cond_ID	3	50	DARK, LIGHTED
Light_Cond_ID	4	20	DUSK
Light_Cond_ID	6	60	DARK, UNKNOWN LIGHTING
Light_Cond_ID	8	0	OTHER (EXPLAIN IN NARRATIVE)
Light_Cond_ID	94	0	REPORTED INVALID
Light_Cond_ID	95	0	NOT REPORTED
Surf_Cond_ID	9	0	UNKNOWN
Surf_Cond_ID	0	10	DRY
Surf_Cond_ID	7	80	WET
Surf_Cond_ID	1	70	STANDING WATER
Surf_Cond_ID	4	80	SLUSH
Surf_Cond_ID	6	90	ICE PATCH
Surf_Cond_ID	5	90	ICE
Surf_Cond_ID	8	0	OTHER (EXPLAIN IN NARRATIVE)
Surf_Cond_ID	3	85	SNOW
Surf_Cond_ID	2	80	SAND, MUD, DIRT
Surf_Cond_ID	94	0	REPORTED INVALID
Surf_Cond_ID	95	0	NOT REPORTED
Harm_Evnt_ID	8	90	PEDESTRIAN
Harm_Evnt_ID	1	85	MOTOR VEHICLE IN TRANSPORT
Harm_Evnt_ID	2	85	MOTOR VEHICLE IN TRANSPORT
Harm_Evnt_ID	4	85	MOTOR VEHICLE IN TRANSPORT
Harm_Evnt_ID	5	85	MOTOR VEHICLE IN TRANSPORT
Harm_Evnt_ID	6	85	MOTOR VEHICLE IN TRANSPORT
Harm_Evnt_ID	3	30	Backing
Harm_Evnt_ID	7	30	FIXED OBJECT
Harm_Evnt_ID	9	30	OTHER OBJECT
Harm_Evnt_ID	0	30	OTHER NON COLLISION
Harm_Evnt_ID	10	90	OVERTURNED
Harm_Evnt_ID	11	0	NOT REPORTED

Harm_Evnt_ID	93	0	UNDETERMINED - FAILED BUSINESS RULE(S)
Harm_Evnt_ID	94	0	REPORTED INVALID
Sus_Serious_Injry_Cnt		85	COUNT
Nonincap_Injry_Cnt		55	COUNT
Poss_Injry_Cnt		2	COUNT
Non_Injry_Cnt		0	COUNT
Unkn_Injry_Cnt		2	COUNT
Death_Cnt		100	COUNT
MAX		974	

APPENDIX E – PENNSYLVANIA CCI DISTRIBUTION FOR THE YEAR OF 2014



APPENDIX F – IRB APPROVAL

Institutional Review Board

Office of the Vice President for Research and Sponsored Projects
The University of Texas at El Paso IRB
FWA No: 00001224
El Paso, Texas 79968-0587
P: 915-747-7693 E: irb.orsp@utep.edu

Date:	February 13, 2019
To:	Daniel Mejia, BS, MSCS
From:	University of Texas at El Paso IRB
Study Title:	[1387186-1] Towards Smart Mobility Through Bottom-Up Data-Driven Modeling Methods For Composite Metric Development In Traffic Accidents
IRB Reference #:	College of Engineering
Submission Type:	New Project
Action:	EXEMPT
Review Type:	Exempt Review
Approval Date:	February 13, 2019

The application for the above referenced study has been reviewed. This study qualifies as exempt from review under the following federal guidelines: **[45 CFR 46.101(b)(2)]**

If Institutional data (secondary or other) will be used for this research project please verify with the applicable department that such data may be used. Additional institutional clearances and approvals may be required. Accordingly, the project should not begin until all required approvals have been obtained.

Exempt protocols do not need be renewed. Please note that it is the Principal Investigator's responsibility to resubmit the proposal for review if there are any modifications made to the originally submitted proposal. This review is required in order to determine if "Exemption" status remains.

This exemption does not relieve the investigators of any responsibilities relating to the research subjects. Research should be conducted in accordance with the ethical principles as outlined in the Belmont Report.

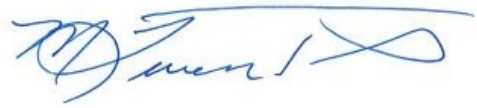
You should retain a copy of this letter and any associated approved study documents for your records.

We will put a copy of this correspondence on file in our office.

Generated on IRBNet

If you have any questions, please contact the IRB Office at irb.orsp@utep.edu or Christina Ramirez at (915) 747-7693 or by email at cramirez22@utep.edu. Please include your study title and reference number in all correspondence with this office.

Sincerely,



Dr. Lorraine Torres, Ed.D, MT(ASCP)
IRB Chair

Generated on IRBNet

APPENDIX G – CRITICAL COMPOSITE INDEX SURVEY

1.

University of Texas at El Paso (UTEP) Institutional Review Board
Informed Consent Form for Research Involving Human Subjects

Protocol Title: Critical Composite Index Survey
Principal Investigator: Daniel Mejia, MSCS
UTEP Department of Computer Science

In this consent form, "you" always means the study subject. If you are a legally authorized representative, please remember that "you" refers to the study subject.

Introduction
You are being asked to take part voluntarily in the research project described below. You are encouraged to take your time in making your decision. It is important that you read the information that describes the study. Please ask the study researcher or the study staff to explain any words or information that you do not clearly understand.

Why is this study being done?
Research is being conducted for the development of a new process that transforms data to knowledge based on reported historical data, namely, traffic accidents in the State of Texas. As an outcome of this research, a new metric (type of measurement) was developed called the Critical Composite Index. The Critical Composite Index attempts to describe individual traffic accidents in a standard way. The Critical Composite Index was developed using a weighted scale based on the following three categories: circumstances of the accident, the people involved in the accident, and any external factors that may have contributed to the accident. For the purpose of this research a traffic accident is defined as: an event where a motor vehicle hits (collides) with an object, person, or another vehicle of any type.

Approximately, 100 people will be enrolling in this study around the city of El Paso.

You are being asked to be in the study because you are:
1) A Domain Expert 18 years or older
a. Individual with extensive knowledge of traffic accident reporting/traffic engineering

2) A Non-Domain Expert 18 years or older
a. Commuter without extensive knowledge of traffic accident reporting/traffic engineering

If you decide to enroll in this study, your involvement will last about fifteen (15) minutes total.

What is involved in the study?
If you agree to take part in this study, the research team will:
1) Provide you with the online survey.

You will:
1) Provide your own device (e.g. computer, smartphone) or use a computer lab device to complete the survey
2) Answer the questions provided in the online survey

What are the risks and discomforts of the study?
There are no known or anticipated risks or discomforts associated with participation.

Are there benefits to taking part in this study?
You are not likely to benefit by taking part in this study. This research may help us determine the benefits of having a standardized metric to describe traffic accidents.

Who is paying for this study?
Not Funded

What are my costs?
There are no direct costs.

Will I be paid to participate in this study?
You will not be compensated for taking part in this research study.

What other options are there?
You have the option not to take part in this study. There will be no penalties involved if you choose not to take part in this study.

What if I want to withdraw, or am asked to withdraw from this study?
Taking part in this study is voluntary. You have the right to choose not to take part in this study. If you do not take part in the study, there will be no penalty or loss of benefit.

If you choose to take part, you have the right stop at any time. However, we encourage you to talk to the researchers so that they know why you are leaving the study.

Who do I call if I have questions or problems?
You may ask any questions you have now. If you have questions later, you may contact Mr. Daniel Mejia at dmmejia2@miners.utep.edu or Dr. Natalia Villanueva-Rosales at nvillanuevarosales@utep.edu.

If you have questions or concerns about your participation as a research subject, please contact the UTEP Institutional Review Board (IRB) at (915-747-7693) or irb.orsp@utep.edu.

What about confidentiality?
Every effort will be made to keep your information confidential; all responses will be kept secure within the UTEP QuestionPro application.

Organizations that may inspect and/or copy your research records for quality assurance and data analysis include, but are not necessarily limited to:
- Department of Health and Human Services
- QuestionPro, which is being used to collect data and your information may be used according to their privacy policy found: https://www.questionpro.com/help/privacy-policy.html

Because of the need to release information to these parties, absolute confidentiality cannot be guaranteed.

The results of this research study may be presented at meetings or in publications; however, your name or other personal identifiers will not be disclosed in those presentations.

Authorization Statement
I have read each page of this paper about the study (or it was read to me). I know that being in this study is voluntary and I choose to be in this study. I will get a copy of this consent by printing a copy for my records.

By checking "I Agree" in the box below, I confirm that I am at least 18 years old and I agree to participate in this research project:

☐ I Agree

2.
About this research:
Research is being conducted for the development of a new algorithm (process) for the transformation of data to knowledge based on reported historical data, namely, traffic accidents in the State of Texas. As an outcome of this research, a new metric was developed called the Critical Composite Index. The Critical Composite Index attempts to describe individual traffic accidents in a standard way.

The Critical Composite Index was developed using a weighted scale based on the following three categories: circumstances of the accident, the people involved in the accident, and any external factors that may have contributed to the accident. For the purpose of this research a traffic accident is defined as: an event where a motor vehicle hits (collides) with an object, person, or another vehicle of any type.

This survey is being conducted to:

1. Discover how the Critical Composite Index is understood based on the information that it provides
2. Determine the clarity of the Critical Composite Index for non-domain and domain-experts
3. Determine if the Critical Composite Index is an improvement from current metric reporting methods
4. Discover any improvements that can be made to the Critical Composite Index

3.The Critical Composite Index is a metric that is intended to describe each individual reported traffic accident. The Critical Composite Index is designed to describe traffic accidents as a way to provide insight into each individual accident in a standard way by considering factors of the accident.

The Critical Composite Index was developed using a weighted scale based on the following three categories: **circumstances** of the accident, the **people** involved in the accident, and any **external factors** that may have contributed to the accident. The weights of the weighted scale has the potential to be changed based on domain experts.

For this research and this survey, a weighted scaled was developed as a general guide for the creation of the Critical Composite Index. The Critical Composite Index described in this survey is a **general example** of possible classification values for any given traffic accident, and may be adjusted for different weights.

The development of the Critical Composite Index used in this research is based on the following criteria and the following importance (weights) highest to lowest:

Table 1. Critical Composite Index Weighted Criteria

Criteria	Description
Circumstances	
Fatal Crash	Deceased Person
Crash Time	Time of Accident
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)
School Bus Involved	School Bus Involved in Accident
Railroad Involved	Accident Occurred at a Railroad
Weather Conditions	Weather Conditions at time of Accident
Light Conditions	Light Conditions at time of Accident
Surface Conditions	Conditions of road surface at time of Accident
People Involved	
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.
Number of Non-Incapacitating Injuri	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.
Number of Unknown Injuries	No Injury reported
Location	
Harmful Event (Object Struck)	Vehicle collided into object
Crash At Intersection	Accident Occurred at an Intersection
Active School Zone	Accident Occurred in an Active School Zone
Type of Lane Division	Number of Lanes and Direction
Construction Zone	Accident Occurred in a Construction Zone
Construction Workers Present	Accident Occurred While Construction Workers were Present
At Least $1000 In Damages	Accident caused at least $1000 in damage

Source: Texas Department of Transportation, 2016

Based on the criteria shown in Table 1, the Critical Composite Index and a severity chart (shown below) has been developed. The Critical Composite Index is a value that is computed based on the above mentioned criteria; each accident is evaluated independently of other accidents, thus each accident has its own Critical Composite Index value and severity classification.

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Accident
> 20 – 40	Moderate Accident
> 40 – 50	Major Accident

> 50 +	Severe Accident

The four developed classifications for measuring traffic accidents are the following: Minor, Moderate, Major, and Severe. Based on the Critical Composite Index value that was computed a severity classification is determined.

The weight scale for the Critical Composite Index can **generally** be described as the following (**Note:** This table should be used as a guide; each traffic accident varies and not all have the same characteristics):

Table 2. Possible commonalities in weight scale description

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Accidents
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

* 4.With respect to my knowledge of Traffic Accidents, I consider myself to be:

◯ A Non-Domain Expert 18 years or older
- Commuter without extensive knowledge of traffic accident reporting/traffic engineering

◯ Domain Expert
- Individual with extensive knowledge of traffic accident reporting/traffic engineering
- Traffic Investigator
- Traffic Police Officer
- Traffic Engineer
- Other Traffic Professional

* 5.

Criteria	Description
Circumstances	
Fatal Crash	Deceased Person
Crash Time	Time of Accident
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)
School Bus Involved	School Bus Involved in Accident
Railroad Involved	Accident Occurred at a Railroad
Weather Conditions	Weather Conditions at time of Accident
Light Conditions	Light Conditions at time of Accident
Surface Conditions	Conditions of road surface at time of Accident
People Involved	
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.
Number of Non-Incapacitating Injuri	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.
Number of Unknown Injuries	No Injury reported
Location	
Harmful Event (Object Struck)	Vehicle collided into object
Crash At Intersection	Accident Occurred at an Intersection
Active School Zone	Accident Occurred in an Active School Zone
Type of Lane Division	Number of Lanes and Direction
Construction Zone	Accident Occurred in a Construction Zone
Construction Workers Present	Accident Occurred While Construction Workers were Present
At Least $1000 In Damages	Accident caused at least $1000 in damage

Do you agree that the above mentioned criteria can accurately classify and describe the severity of an accident?

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

* 6. Based on the following set of criteria, please rank each on the level of importance (1-Not Important to 100-Very Important) when attempting to standardize the way a traffic accident is Classified, Understood, and Compared. (Note: You may have multiple criteria with the same values as others)

Crash Fatalities

Crash Time

Commercial Vehicles Involvement

School Bus Involvement

Railroad Involvement

Weather Conditions

Light Conditions Surface Conditions

Number of Fatalities (Deaths)

Number of Serious Injuries

Number of Non-Incapacitating Injuries

Number of Possible Injuries

Number of Non-Injuries

Number of Unknown Injuries

Harmful Event (Object Struck)

Occurred at an intersection

Occurred in Active School Zone

Type of Lane Division

Occurred in Construction Zone

Occurred with Construction Workers Present

* 7.Additional categories that may be considered when standardizing the classification, understanding, and comparability of traffic accidents are:

* 8.

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Accident

> 20 – 40	Moderate Accident
> 40 – 50	Major Accident
> 50 +	Severe Accident

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Accidents
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Criteria	Description	Value
Fatal Crash	Deceased Person	NO
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)	NO
School Bus Involved	School Bus Involved in Accident	NO
Railroad Involved	Accident Occurred at a Railroad	NO
Active School Zone	Accident Occurred in an Active School Zone	NO
Crash Time	Time of Accident	1:17 PM
At Least $1000 In Damages	Accident caused at least $1000 in damage	YES
Type of Line Division	Number of Lanes and Direction	2 LANE, 2 WAY
Construction Zone	Accident Occurred in a Construction Zone	NO
Construction Workers Present	Accident Occurred While Construction Workers were Present	NO
Crash At Intersection	Accident Occurred at an Intersection	NO
Weather Conditions	Weather Conditions at time of Accident	CLEAR
Light Conditions	Light Conditions at time of Accident	DAYLIGHT
Surface Conditions	Conditions of road surface at time of Accident	DRY
Harmful Event (Object Struck)	Vehicle collided into object	FIXED OBJECT
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.	0
mber of Non-Incapacitating Injur	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.	0
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.	0
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.	1
Number of Unknown Injuries	No Injury reported	0
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.	0

The Critical Composite Index for this traffic Accident is: **15.82 - Minor Accident**.

The data shown describes the following:
Accident occurred at 1:17 pm
At least $1000 of damage
Weather conditions were reported clear
Light conditions were reported as daylight
The road conditions were reported as being dry.
The vehicle hit a fixed object
One (1) person was involved without any reported injuries or fatalities

Do you agree that the classification of **Minor Accident** accurately represents this traffic accident?

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

* 9.

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Accident
> 20 – 40	Moderate Accident
> 40 – 50	Major Accident
> 50 +	Severe Accident

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions

Good weather and road conditions

Major Accidents
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities

Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Criteria	Description	Value
Fatal Crash	Deceased Person	NO
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)	NO
School Bus Involved	School Bus Involved in Accident	NO
Railroad Involved	Accident Occurred at a Railroad	NO
Active School Zone	Accident Occurred in an Active School Zone	NO
Crash Time	Time of Accident	1:28 PM
At Least $1000 In Damages	Accident caused at least $1000 in damage	YES
Type of Line Division	Number of Lanes and Direction	2 LANE, 2 WAY
Construction Zone	Accident Occurred in a Construction Zone	NO
Construction Workers Present	Accident Occurred While Construction Workers were Present	NO
Crash At Intersection	Accident Occurred at an Intersection	YES
Weather Conditions	Weather Conditions at time of Accident	CLEAR
Light Conditions	Light Conditions at time of Accident	DAYLIGHT
Surface Conditions	Conditions of road surface at time of Accident	DRY
Harmful Event (Object Struck)	Vehicle collided into object	HIT MOVING VEHICLE
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.	1
mber of Non-Incapacitating Injur	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.	0
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.	0
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.	5
Number of Unknown Injuries	No Injury reported	0
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.	0

The above accident computed a Critical Composite Index of: **27.80 - Moderate**

The data shown describes the following:

Accident occurred at 1:28 pm
At least $1000 of damage
Light conditions were reported as daylight
Weather conditions were reported to be clear
The road conditions were reported as being dry
A vehicle hit a moving vehicle
Six (6) people were involved in the traffic accident
One (1) person involved was seriously injured
Five (5) people did not sustain any injuries

Do you agree that the classification of **Moderate Accident** accurately represents this traffic accident?

◯ Strongly Agree

- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

* 10.

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Accident
> 20 – 40	Moderate Accident
> 40 – 50	Major Accident
> 50 +	Severe Accident

If a traffic accident were to be reported and given a Critical Composite Index value & corresponding severity chart, I would confidently understand the severity of the accident?

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

11.

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Accidents
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Do you agree that the Critical Composite Index & Severity Classification (Minor, Moderate, Major, Severe) is an appropriate representation of traffic accidents - considering the applicable details of the accident?

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

12. *

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Accidents
Many injuries (including serious injuries)

Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Criteria	Description	Value
Fatal Crash	Deceased Person	YES
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)	NO
School Bus Involved	School Bus Involved in Accident	NO
Railroad Involved	Accident Occurred at a Railroad	NO
Active School Zone	Accident Occurred in an Active School Zone	NO
Crash Time	Time of Accident	6:45 AM
At Least $1000 In Damages	Accident caused at least $1000 in damage	YES
Type of Line Division	Number of Lanes and Direction	2 LANE, 2 WAY
Construction Zone	Accident Occurred in a Construction Zone	NO
Construction Workers Present	Accident Occurred While Construction Workers were Present	NO
Crash At Intersection	Accident Occurred at an Intersection	NO
Weather Conditions	Weather Conditions at time of Accident	RAIN
Light Conditions	Light Conditions at time of Accident	DAYLIGHT
Surface Conditions	Conditions of road surface at time of Accident	WET
Harmful Event (Object Struck)	Vehicle collided into object	HIT MOVING VEHICLE
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.	1
Number of Non-Incapacitating Injuries	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.	0
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.	0
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.	0
Number of Unknown Injuries	No Injury reported	0
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.	3

The data shown describes the following:

- Accident occurred at 6:45 am
- At least $1000 of damage
- The traffic accident was reported to occur in daylight
- The weather was reported to be raining
- The road conditions were reported as being wet
- A vehicle hit another moving vehicle
- Four (4) people were involved in the traffic accident
- Three (3) of the people involved are deceased as a result of the accident (Fatal Accident)
- One (1) person sustained serious injuries

Based on your understanding of the Critical Composite Index and traffic accident knowledge;
What severity would you use to best classify this traffic accident?

- ◯ Minor Accident
- ◯ Moderate Accident
- ◯ Major Accident
- ◯ Severe Accident
- ◯ It is not clear

13. *

Minor Accidents
Few number of people involved
Zero to very few injuries
Occurrence during the day
Not occurring at an intersection
Good weather and road conditions
Moderate Accidents
Multiple people involved
Few non-serious injuries
Any time of the day
Occurring at intersections
Good weather and road conditions
Major Accidents
Many injuries (including serious injuries)
Any time of the day
Occurring at intersections
Poor weather and road conditions
Usually no fatalities
Severe Accidents
Many serious injuries
Traffic accidents during the night (or dark light)
Occurring at intersections
Poor weather and road conditions
Fatalities or a several serious injuries

Criteria	Description	Value
Fatal Crash	Deceased Person	NO
Commercial Vehicle Involved	Large Commercial Vehicle Involved (Semi-truck)	NO
School Bus Involved	School Bus Involved in Accident	NO
Railroad Involved	Accident Occurred at a Railroad	NO
Active School Zone	Accident Occurred in an Active School Zone	NO
Crash Time	Time of Accident	12:42 PM
At Least $1000 In Damages	Accident caused at least $1000 in damage	YES
Type of Line Division	Number of Lanes and Direction	2 LANE, 2 WAY
Construction Zone	Accident Occurred in a Construction Zone	NO
Construction Workers Present	Accident Occurred While Construction Workers were Present	NO
Crash At Intersection	Accident Occurred at an Intersection	YES
Weather Conditions	Weather Conditions at time of Accident	CLEAR
Light Conditions	Light Conditions at time of Accident	DAYLIGHT
Surface Conditions	Conditions of road surface at time of Accident	DRY
Harmful Event (Object Struck)	Vehicle collided into object	HIT MOVING VEHICLE
Number of Serious Injuries	Serious Injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred.	0
Number of Non-Incapacitating Injuries	Non-incapacitating Evident Injury is any injury, other than a fatal injury or serious injury, which is evident to observers at the scene of the crash in which the injury occurred.	5

Number of Non-Incapacitating Injuries		
Number of Possible Injuries	Possible Injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating evident injury. Possible injuries are those which are claimed or reported, or indicated by behavior, but not by wounds.	0
Number of Non-Injuries	No Injury is a situation in which there is no reason to believe that the person received any bodily harm from the motor vehicle traffic crash in which they were involved.	0
Number of Unknown Injuries	No Injury reported	0
Number of Fatalities	Fatal Injury is any injury that results in death within thirty (30) days of the motor vehicle traffic crash.	0

The data shown describes the following:

- Accident occurred at 12:42 pm
- At least $1000 of damage
- The traffic accident was reported to occur in daylight
- The road conditions were reported as dry
- A vehicle hit another moving vehicle
- Five (5) people were involved, all of whom suffered non-incapacitating injuries

Based on your understanding of the Critical Composite Index and traffic accident knowledge;
What severity would you use to best classify this traffic accident?

- ◯ Minor Accident
- ◯ Moderate Accident
- ◯ Major Accident
- ◯ Severe Accident
- ◯ It is not clear

* 14.

All traffic accident data collected comes from the Texas Department of Transportation:
Knowing that the data used comes from a reliable source, how much does it improve your trust of the Critical Composite Index being an useful way to classify, understand, and compare traffic accidents in a standard way?

- ◯ Much more
- ◯ Somewhat more
- ◯ About the same
- ◯ Somewhat less
- ◯ Much less

* 15.How helpful would the Critical Composite Index improve the way you would compare traffic accidents to one another?

- ◯ Extremely helpful
- ◯ Very helpful
- ◯ Moderately helpful
- ◯ Slightly helpful
- ◯ Not at all helpful

* 16.How important is it to understand how the Critical Composite Index was computed?

- ◯ Extremely important
- ◯ Very important
- ◯ Moderately important
- ◯ Slightly important
- ◯ Not at all important

* 17.If the Critical Composite Index was used in practice (e.g. news media, mobile app) I would find it:

- ◯ Extremely helpful
- ◯ Very helpful
- ◯ Moderately helpful
- ◯ Slightly helpful
- ◯ Not at all helpful

* 18.

Commonly, traffic accidents are reported to the public as a frequency and use a single scope (e.g. fatality, injury) of measure such as the following:

- number of crashes per year (crash rate)
- number of injuries per year (injury rate)
- number of fatalities per year (fatality rate)

Critical Composite Index > (Greater Than)	Severity
0 – 20	Minor Accident
> 20 – 40	Moderate Accident
> 40 – 50	Major Accident
> 50 +	Severe Accident

Sample Comparison Chart

Individual Accidents Critical Composite Index	Frequency Report of Accidents
Accident One **- 19.41 Minor** Accident Two **- 47.90 Major** Accident Three **- 22.24 Moderate** Accident Four **- 41.23 Major** Accident Five **- 58.16 Severe** Accident Six **- 54.31 Severe** Accident Seven **- 30.36 Moderate** Accident Eight **- 18.13 Minor** Accident Nine **- 100.94 Severe** Accident Ten **- 38.06 Moderate**	Number of Crashes - 10 Number of Injuries - 37 Number of Fatalities - 1

The data shown above is actual traffic accident data; both the Critical Composite Index and the Frequency Metrics are based on the same set of data.

Which method of reporting, the Critical Composite Index (Case-by-case) or Frequency Report of Accidents (Group of accidents), do you understand better when trying to understand traffic accidents on the roadways?

- ◯ I understand the Critical Composite Index Much better than frequency reporting
- ◯ I understand the Critical Composite Index Somewhat better than frequency reporting
- ◯ I understand the Critical Composite Index About the same as frequency reporting
- ◯ I understand the Critical Composite Index Somewhat worse than frequency reporting
- ◯ I understand the Critical Composite Index Much worse than frequency reporting

* 19.Is there any missing information that you would want to see in order to understand the Critical Composite Index and severity chart?

* 20.What advantages, if any, do you see by using the Critical Composite Index?

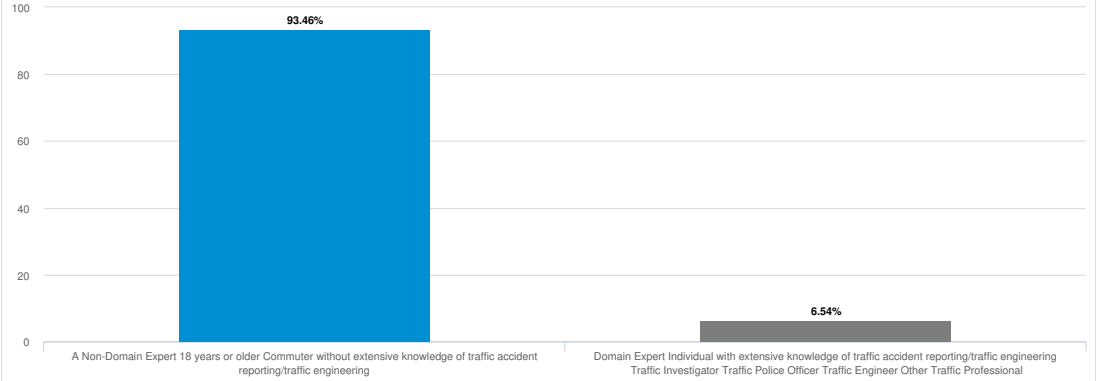
* 21.What area of improvement, if any, can be done to the Critical Composite Index?

APPENDIX H – CRITICAL COMPOSITE INDEX SURVEY RESULTS

Critical Composite Index Survey - Complete - Dashboard

VIEWED	STARTED	COMPLETED	COMPLETION RATE	DROP OUTS	TIME TO COMPLETE
257	107	107	100%	0	13 mins

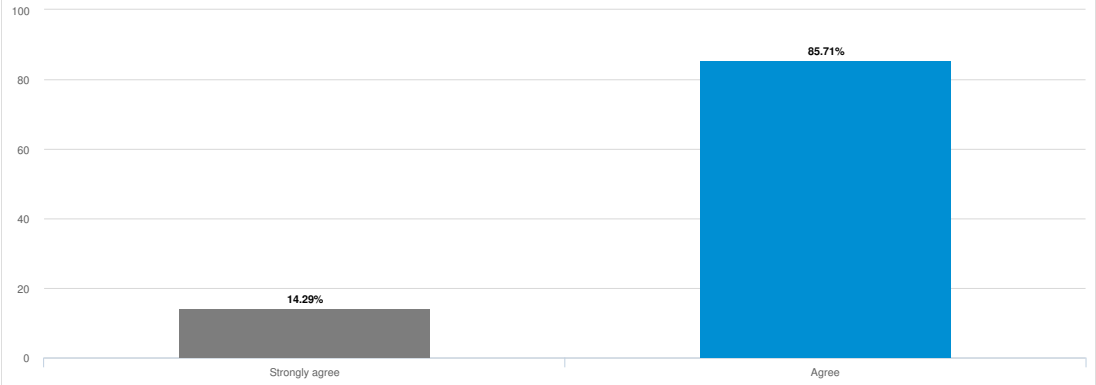
4. With respect to my knowledge of Traffic Accidents, I consider myself to be:



Answer	Count	Percent	20% 40% 60% 80% 100%
A Non-Domain Expert 18 years or older Commuter without extensive knowledge of traffic accident reporting/traffic engineering	100	93.46%	
Domain Expert Individual with extensive knowledge of traffic accident reporting/traffic engineering Traffic Investigator Traffic Police Officer Traffic Engineer Other Traffic Professional	7	6.54%	
Total	**107**	**100 %**	

Mean: 1.065 **Variance: 0.062** **Standard Deviation: 0.248** **Standard Error: 0.024** **Confidence Interval: [1.018 - 1.112]**

5. Do you agree that the above mentioned criteria can accurately classify and describe the severity of an accident?

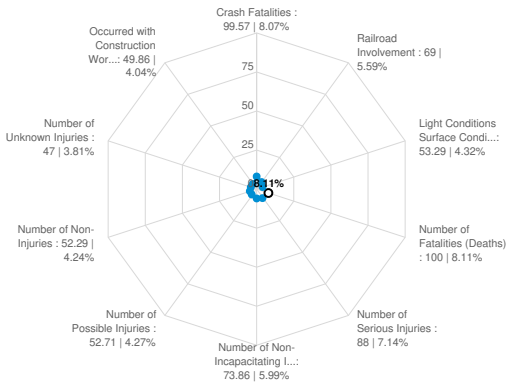


Answer	Count	Percent	20% 40% 60% 80% 100%
Strongly agree	1	14.29%	
Agree	6	85.71%	

Neutral	0	0%
Disagree	0	0%
Strongly disagree	0	0%
Total	**7**	**100 %**

Mean: 1.857 **Variance: 0.143** **Standard Deviation: 0.378** **Standard Error: 0.143** **Confidence Interval: [1.577 - 2.137]**

6. Based on the following set of criteria, please rank each on the level of importance (1-Not Important to 100-Very Important) when attempting to standardize the way a traffic accident is Classified, Understood, and Compared. (Note: You may have multiple criteria with the same values as others)



Powered by AI

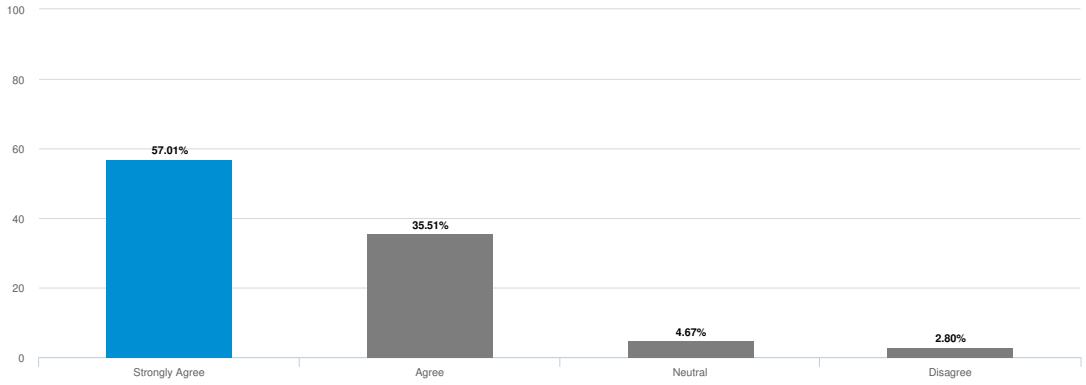
Question	Score	0 — 100
Crash Fatalities	99.57	
Crash Time	61.86	
Commercial Vehicles Involvement	62.14	
School Bus Involvement	67.86	
Railroad Involvement	69	
Weather Conditions	67.57	
Light Conditions Surface Conditions	53.29	
Number of Fatalities (Deaths)	100	
Number of Serious Injuries	88	
Number of Non-Incapacitating Injuries	73.86	
Number of Possible Injuries	52.71	
Number of Non-Injuries	52.29	
Number of Unknown Injuries	47	
Harmful Event (Object Struck)	55	
Occurred at an intersection	60.29	
Occurred in Active School Zone	62.14	
Type of Lane Division	55	
Occurred in Construction Zone	55.86	
Occurred with Construction Workers Present	49.86	
Average	**64.91**	

7. Additional categories that may be considered when standardizing the classification, understanding, and comparability of traffic accidents are:

Additional categories that may be considered when standardizing the classification, understanding, and comparability of traffic accidents are:

03/07/2019	32161620	Location of Accident. Example Interstate vs. local street
02/28/2019	31859408	Visualization, and full attention to whats around you
02/22/2019	31540558	Type of vehicle and its condition
02/22/2019	31542303	type of crash, (rear ended etc)
02/22/2019	31540998	LANE CONFIGURATION
02/15/2019	31097055	Type of vehicles involved and if pedestrians were involved.
02/15/2019	31097116	important

8. The Critical Composite Index for this traffic Accident is: 15.82 - Minor Accident. The data shown describes the following: Accident occurred at 1:17 pm At least $1000 of damage Weather conditions were reported clear Light conditions were reported as daylight The road conditions were reported as being dry. The vehicle hit a fixed object One (1) person was involved without any reported injuries or fatalitiesDo you agree that the classification of Minor Accident accurately represents this traffic accident?



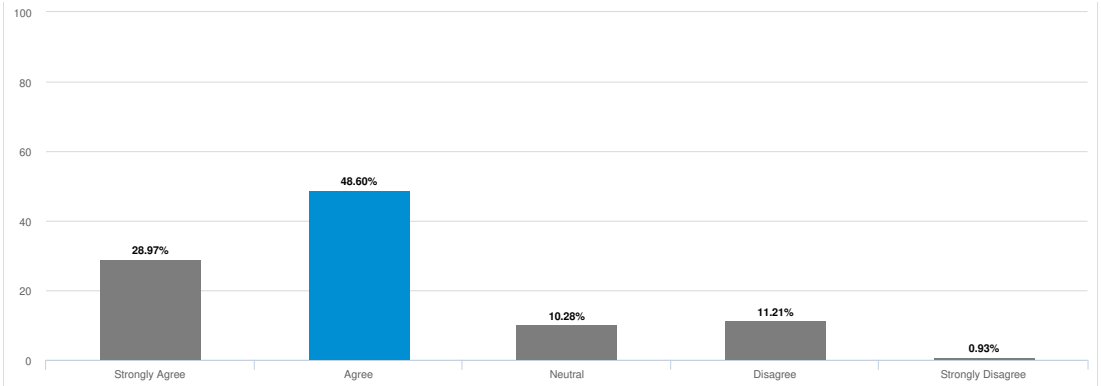
Answer	Count	Percent	20% 40% 60% 80% 100%
Strongly Agree	61	57.01%	
Agree	38	35.51%	
Neutral	5	4.67%	
Disagree	3	2.8%	
Strongly Disagree	0	0%	
Total	**107**	**100 %**	

Mean: 1.533 **Variance: 0.515** **Standard Deviation: 0.718** **Standard Error: 0.069** **Confidence Interval: [1.397 - 1.669]**

9. The above accident computed a Critical Composite Index of: 27.80 - ModerateThe data shown describes the following: Accident occurred at 1:28 pm At least $1000 of damage Light conditions were reported as daylight Weather conditions were reported to be clear The road conditions were reported as being dry A vehicle hit a moving vehicle Six (6) people were involved in the traffic accident One (1) person involved was seriously injured Five (5) people did not sustain any injuriesDo you agree that the classification of Moderate Accident accurately represents this traffic accident?

Top / Bottom Box Score

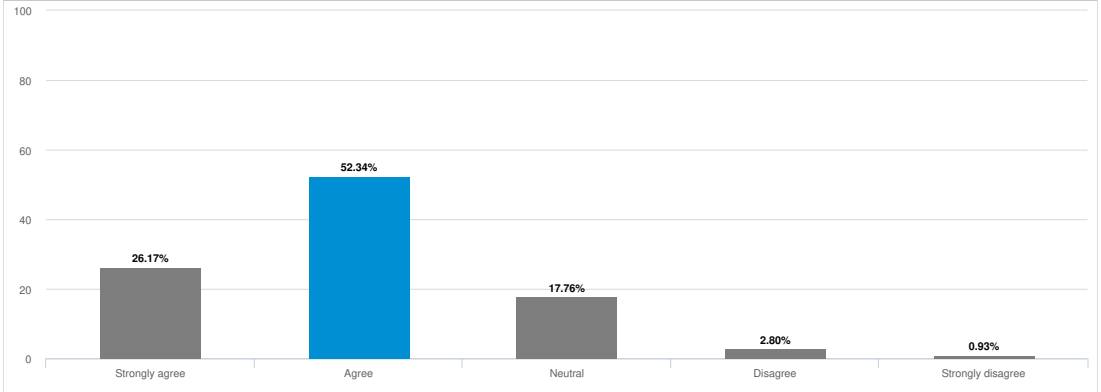
Answer	Count	Percent	20% 40% 60% 80% 100%
Satisfied	0	0%	
Neutral	0	0%	
Unsatisfied	0	0%	
Total	**0**	**0 %**	



Answer	Count	Percent	20%	40%	60%	80%	100%
Strongly Agree	31	28.97%					
Agree	52	48.6%					
Neutral	11	10.28%					
Disagree	12	11.21%					
Strongly Disagree	1	0.93%					
Total	**107**	**100 %**					

Mean: 2.065 **Variance: 0.930** **Standard Deviation: 0.964** **Standard Error: 0.093** **Confidence Interval: [1.883 - 2.248]**

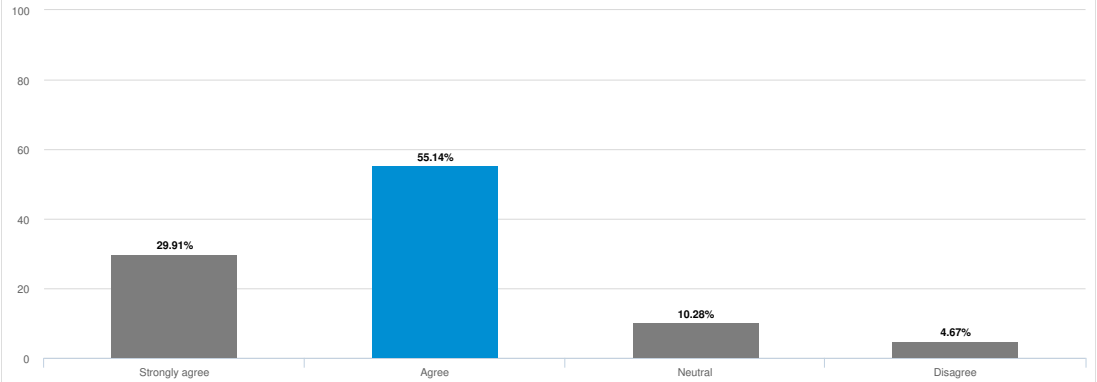
10. If a traffic accident were to be reported and given a Critical Composite Index value & corresponding severity chart, I would confidently understand the severity of the accident?



Answer	Count	Percent	20%	40%	60%	80%	100%
Strongly agree	28	26.17%					
Agree	56	52.34%					
Neutral	19	17.76%					
Disagree	3	2.8%					
Strongly disagree	1	0.93%					
Total	**107**	**100 %**					

Mean: 2.000 **Variance: 0.642** **Standard Deviation: 0.801** **Standard Error: 0.077** **Confidence Interval: [1.848 - 2.152]**

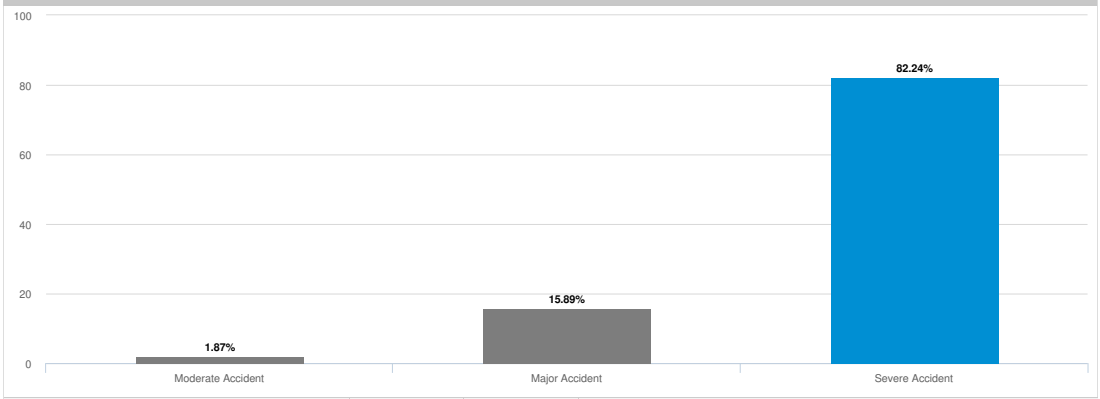
11. Do you agree that the Critical Composite Index & Severity Classification (Minor, Moderate, Major, Severe) is an appropriate representation of traffic accidents - considering the applicable details of the accident?



Answer	Count	Percent	20% 40% 60% 80% 100%
Strongly agree	32	29.91%	
Agree	59	55.14%	
Neutral	11	10.28%	
Disagree	5	4.67%	
Strongly disagree	0	0%	
Total	**107**	**100 %**	

Mean: 1.897 **Variance: 0.584** **Standard Deviation: 0.764** **Standard Error: 0.074** **Confidence Interval: [1.752 - 2.042]**

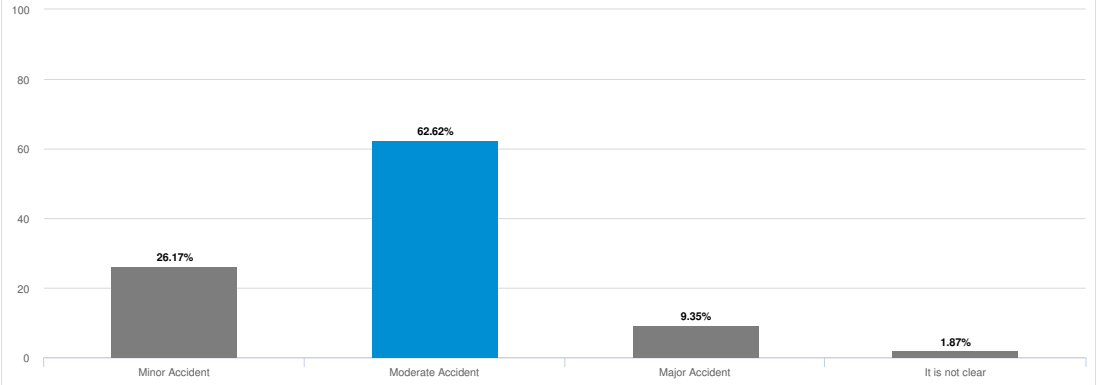
12. The data shown describes the following: Accident occurred at 6:45 am At least $1000 of damage The traffic accident was reported to occur in daylight The weather was reported to be raining The road conditions were reported as being wet A vehicle hit another moving vehicle Four (4) people were involved in the traffic accident Three (3) of the people involved are deceased as a result of the accident (Fatal Accident) One (1) person sustained serious injuries Based on your understanding of the Critical Composite Index and traffic accident knowledge;What severity would you use to best classify this traffic accident?



Answer	Count	Percent	20% 40% 60% 80% 100%
Minor Accident	0	0%	
Moderate Accident	2	1.87%	
Major Accident	17	15.89%	
Severe Accident	88	82.24%	
It is not clear	0	0%	
Total	**107**	**100 %**	

Mean: 3.804 **Variance: 0.197** **Standard Deviation: 0.444** **Standard Error: 0.043** **Confidence Interval: [3.720 - 3.888]**

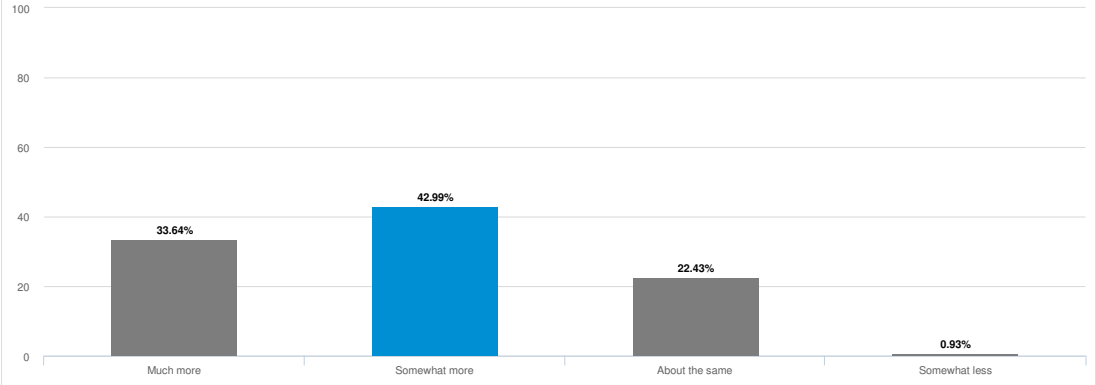
13. The data shown describes the following: Accident occurred at 12:42 pm At least $1000 of damage The traffic accident was reported to occur in daylight The road conditions were reported as dry A vehicle hit another moving vehicle Five (5) people were involved, all of whom suffered non-incapacitating injuries Based on your understanding of the Critical Composite Index and traffic accident knowledge;What severity would you use to best classify this traffic accident?



Answer	Count	Percent	20% 40% 60% 80% 100%
Minor Accident	28	26.17%	
Moderate Accident	67	62.62%	
Major Accident	10	9.35%	
Severe Accident	0	0%	
It is not clear	2	1.87%	
Total	**107**	**100 %**	

Mean: 1.888 **Variance: 0.516** **Standard Deviation: 0.718** **Standard Error: 0.069** **Confidence Interval: [1.752 - 2.024]**

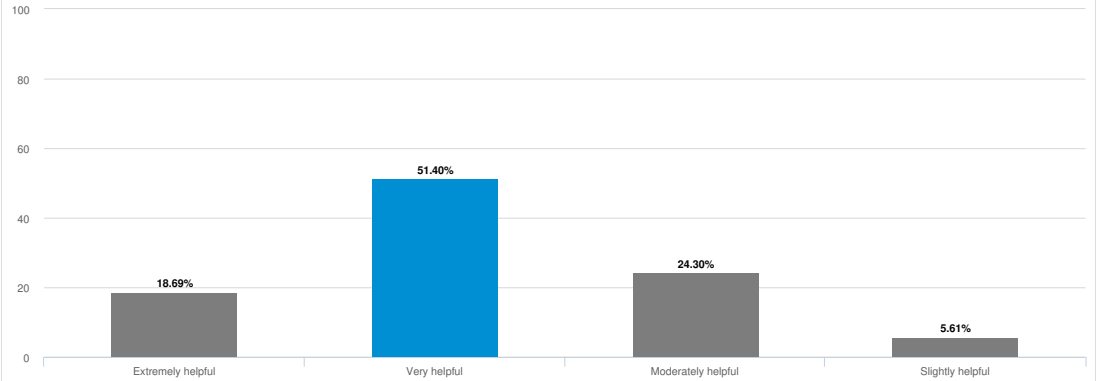
14. All traffic accident data collected comes from the Texas Department of Transportation:Knowing that the data used comes from a reliable source, how much does it improve your trust of the Critical Composite Index being an useful way to classify, understand, and compare traffic accidents in a standard way?



Answer	Count	Percent	20% 40% 60% 80% 100%
Much more	36	33.64%	
Somewhat more	46	42.99%	
About the same	24	22.43%	
Somewhat less	1	0.93%	
Much less	0	0%	

Total			107	100 %	
Mean: 1.907	Variance: 0.595	Standard Deviation: 0.771		Standard Error: 0.075	Confidence Interval: [1.760 - 2.053]

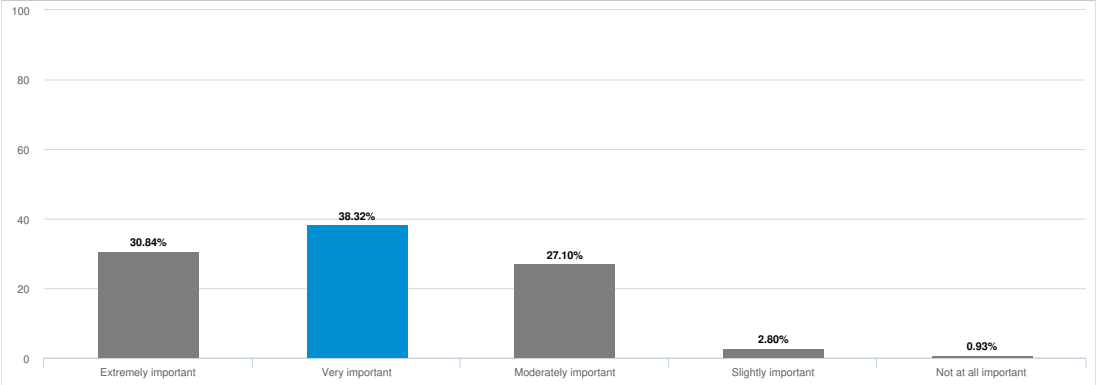
15. How helpful would the Critical Composite Index improve the way you would compare traffic accidents to one another?



Answer	Count	Percent	20% 40% 60% 80% 100%
Extremely helpful	20	18.69%	
Very helpful	55	51.4%	
Moderately helpful	26	24.3%	
Slightly helpful	6	5.61%	
Not at all helpful	0	0%	
Total	**107**	**100 %**	

Mean: 2.168	Variance: 0.632	Standard Deviation: 0.795	Standard Error: 0.077	Confidence Interval: [2.018 - 2.319]

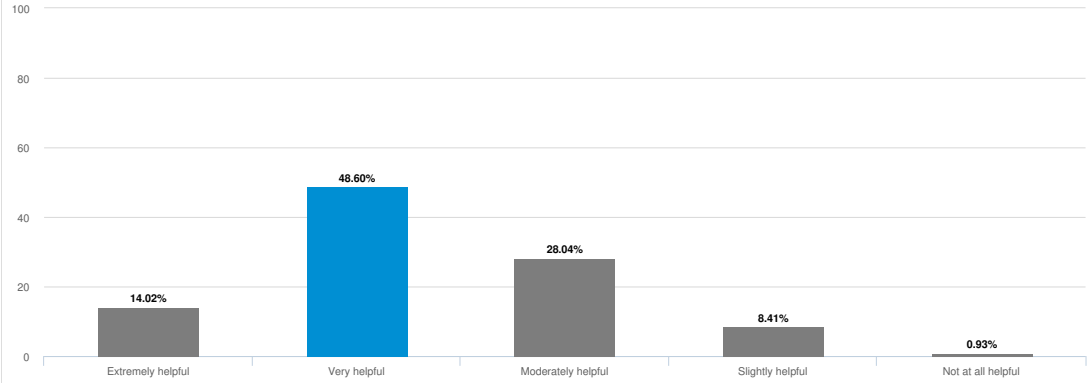
16. How important is it to understand how the Critical Composite Index was computed?



Answer	Count	Percent	20% 40% 60% 80% 100%
Extremely important	33	30.84%	
Very important	41	38.32%	
Moderately important	29	27.1%	
Slightly important	3	2.8%	
Not at all important	1	0.93%	
Total	**107**	**100 %**	

Mean: 2.047	Variance: 0.781	Standard Deviation: 0.884	Standard Error: 0.085	Confidence Interval: [1.879 - 2.214]

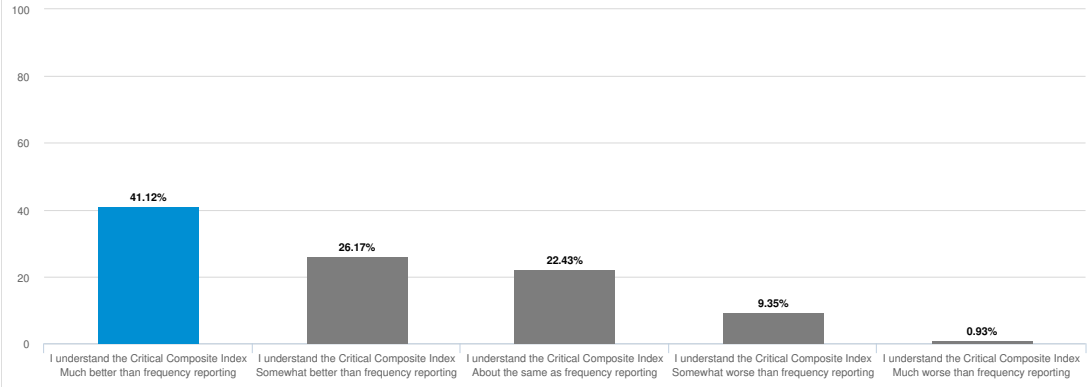
17. If the Critical Composite Index was used in practice (e.g. news media, mobile app) I would find it:



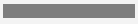
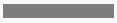
Answer	Count	Percent	20% 40% 60% 80% 100%
Extremely helpful	15	14.02%	
Very helpful	52	48.6%	
Moderately helpful	30	28.04%	
Slightly helpful	9	8.41%	
Not at all helpful	1	0.93%	
Total	**107**	**100 %**	

Mean: 2.336	Variance: 0.735	Standard Deviation: 0.857	Standard Error: 0.083	Confidence Interval: [2.174 - 2.499]

18. Commonly, traffic accidents are reported to the public as a frequency and use a single scope (e.g. fatality, injury) of measure such as the following: number of crashes per year (crash rate) number of injuries per year (injury rate) number of fatalities per year (fatality rate)Sample Comparison Chart Individual Accidents Critical Composite Index Frequency Report of Accidents Accident One - 19.41 Minor Accident Two - 47.90 Major Accident Three - 22.24 Moderate Accident Four - 41.23 Major Accident Five - 58.16 Severe Accident Six - 54.31 Severe Accident Seven - 30.36 Moderate Accident Eight - 18.13 Minor Accident Nine - 100.94 Severe Accident Ten - 38.06 Moderate Number of Crashes - 10 Number of Injuries - 37 Number of Fatalities - 1 The data shown above is actual traffic accident data; both the Critical Composite Index and the Frequency Metrics are based on the same set of data.Which method of reporting, the Critical Composite Index (Case-by-case) or Frequency Report of Accidents (Group of accidents), do you understand better when trying to understand traffic accidents on the roadways?



Answer	Count	Percent	20% 40% 60% 80% 100%
I understand the Critical Composite Index Much better than frequency reporting	44	41.12%	

I understand the Critical Composite Index Somewhat better than frequency reporting	28	26.17%	
I understand the Critical Composite Index About the same as frequency reporting	24	22.43%	
I understand the Critical Composite Index Somewhat worse than frequency reporting	10	9.35%	
I understand the Critical Composite Index Much worse than frequency reporting	1	0.93%	
Total	**107**	**100 %**	

Mean: 2.028 **Variance: 1.103** **Standard Deviation: 1.050** **Standard Error: 0.102** **Confidence Interval: [1.829 - 2.227]**

19. Is there any missing information that you would want to see in order to understand the Critical Composite Index and severity chart?

Is there any missing information that you would want to see in order to understand the Critical Composite Index and severity chart?		
03/08/2019	32213141	Everything seemed to be in order. It seems every scenario can be covered using that index.
03/08/2019	32186669	None
03/07/2019	32182543	I would like to know what the point system is for the index and chart so that I know how deadly an accident is without the decimals.
03/07/2019	32175993	No, information is clear.
03/07/2019	32161620	no comment
03/05/2019	32025495	I believe the Critical Composite Index and severity chart have provided a very clear understanding.
03/05/2019	32024811	No
03/04/2019	32022076	NO
03/04/2019	32020867	No
03/04/2019	32020355	N.A.
03/04/2019	32020092	No
03/04/2019	32020001	NA
03/04/2019	32019904	No
03/04/2019	32018394	N/a
03/04/2019	32019740	No
03/04/2019	32019055	Explain more about the definition of a non-incapacitating injury. I was confused on the original definition of it, I had to look it up online.
03/04/2019	32019107	I believe the Critical Composite Index and severity chart serves its purpose by providing details that are feasible to understand and easily applicable to label.
03/04/2019	32018382	It seems a bit complicated
03/04/2019	32018620	None
03/04/2019	32018146	No
03/04/2019	32018127	None
03/04/2019	32018177	No
03/04/2019	32017692	None
03/04/2019	32017710	None
03/04/2019	32017820	I don't think so.
03/04/2019	32017810	N/A
03/04/2019	32017428	no
03/04/2019	32017294	no
03/04/2019	32017400	no
03/04/2019	32014905	I didn't understand the index >value or
03/04/2019	32000329	The chat does not show a threshold for monetary loss.

02/28/2019	31859458	Everything was good
02/28/2019	31859597	No
02/28/2019	31859478	The severity is understood if you have the chart memorized. Severity chart just shows what is important to individuals
02/28/2019	31859649	no
02/28/2019	31859372	How many accidents happened in a year, the severity of it and where it happened. Then you can compile the data, map it on a map (GIS) and that could help to improve the traffic on that area or just have more security.
02/28/2019	31859365	I noticed numbers on the left side of the chart that classifies the severity of the accidents. I would find more helpful seeing a number at the bottom of each case's rubric that corresponds to that number on the chart. For example: minor accident = 18.
02/28/2019	31859348	Category of injuries, as they a presented in a broad manner in which it was un clear where would it lie on the spectrum
02/28/2019	31859518	it should add maybe the age range of the owner of the crash, and their conditions of driving
02/28/2019	31859521	No
02/28/2019	31859383	Perhaps the traffic in the area
02/28/2019	31859339	There isn't any missing information that I can see.
02/28/2019	31859562	It has good info
02/28/2019	31859712	no i think everything was covered
02/28/2019	31859439	all the information was clear and understandable
02/28/2019	31859519	no
02/28/2019	31859361	A broader scale for the type of accidents, so that no accident lands in the middle of being either major or severe and can be clearly identified.
02/28/2019	31859358	N/A
02/28/2019	31859497	Numbers of vehicles involved.
02/28/2019	31859416	The total number of critically injured people in major accidents and the number of fatalities in severe accidents.
02/28/2019	31859447	More details on how death affects the severity chart
02/28/2019	31859373	No
02/28/2019	31859539	no
02/28/2019	31859377	No
02/28/2019	31859387	Condition the drivers were while operating the motorized vehicle
02/28/2019	31859420	just a better representation of the possible outcomes. such as the answer in question 21
02/28/2019	31859389	To be more precise
02/28/2019	31859415	N/A
02/28/2019	31859408	No
02/28/2019	31859440	N/A
02/28/2019	31859489	What would an accident be called if say, a car crash occurred in the day and there were fatalities?
02/28/2019	31859330	Make bye always be updated
02/28/2019	31859332	no
02/28/2019	31859477	.
02/28/2019	31859423	location of the crash
02/28/2019	31859394	No everything was okay.
02/28/2019	31859344	What was damaged.
02/28/2019	31859393	no
02/28/2019	31859384	no
02/28/2019	31859335	no
02/22/2019	31540558	Not missing but having a more simple or basic definition that could be understood by everyone
02/22/2019	31542303	no
02/22/2019	31540998	NO

02/22/2019	31533072	clearing time; vehicle to pedestrian crashes and fatalities
02/19/2019	31293576	no
02/16/2019	31136113	A chart showing the Critical Cimposite scale.
02/15/2019	31097055	n/a
02/15/2019	31097020	No
02/15/2019	31096819	no
02/15/2019	31097602	Not really
02/15/2019	31097024	I don't think there is.
02/15/2019	31096977	How is the critical composite index calculated?
02/15/2019	31096996	Time of to get thought the traffic.
02/15/2019	31097066	maybe numbers on much each situation would count towards the index, the way it is now its not clear how much more the index is affected by rainy day compared with people involved
02/15/2019	31096830	The location of the accident. More information on exactly where and the road conditions, rather than having "wet road" besides it being wet was the road maintained?
02/15/2019	31096880	No
02/15/2019	31096823	No
02/15/2019	31097033	N/A
02/15/2019	31097115	How each thing is weighted for the index
02/15/2019	31096829	no
02/15/2019	31097116	nothing is missing
02/15/2019	31096870	Police response time, the need to close off streets or avenues because of the accident.
02/15/2019	31097057	No missing information
02/15/2019	31097017	No
02/15/2019	31096848	not that I can think of
02/15/2019	31096868	Structural damage to roadways or buildings that may lead to traffic jams as they're being repaired.
02/15/2019	31096907	No
02/15/2019	31097088	No
02/15/2019	31096836	no
02/15/2019	31089515	If possible, mode of transportation should be included i.e bicycle, motorcycle, car, trucks, semis
02/15/2019	31066577	Final conditions of vehicle
02/14/2019	31062252	No
02/14/2019	31058176	No
02/14/2019	31056549	maybe alcohol use
02/14/2019	31054318	N/A
02/14/2019	31039072	None that I see
02/14/2019	31039696	No

20. What advantages, if any, do you see by using the Critical Composite Index?

		What advantages, if any, do you see by using the Critical Composite Index?
03/08/2019	32213141	Accuracy in any accident. That can be important for legal purposes.
03/08/2019	32186669	None
03/07/2019	32182543	An advantage would be giving the viewer a visual as to how deadly or non-deadly a car accident is.
03/07/2019	32175993	Understanding traffic accidents better, had not heard of it before this survey.
03/07/2019	32161620	no comment

03/05/2019	32025495	The Critical Composite Index helps the reader understand exactly what comprises the ranking.
03/05/2019	32024811	I can easily tell the severity of an accident that occurs.
03/04/2019	32022076	Increased understanding of classifying traffic accidents
03/04/2019	32020867	Immediately visually understandable (colors)
03/04/2019	32020355	N.A.
03/04/2019	32020092	it Is more specific
03/04/2019	32020001	As long as the charts are available a person is able to gauge how bad an accident is.
03/04/2019	32019904	It's a concise way to report types of accidents.
03/04/2019	32018394	N/a
03/04/2019	32019740	You're able to see what kind of accident it was based off the descriptions of each level
03/04/2019	32019055	It was more easier to understand the CCI than the frequency index. Looking at the frequency index chart, I didn't really know what I was looking at despite the description. To me, the frequency index doesn't tell you anything since accidents happen every so often. The CCI simply assigned a number based on the descriptions of the accident categories.
03/04/2019	32019107	I believe the advantages of the Critical Composite Index aid in not only providing the severity of the crash in any possible court case but also to the provided insurance, if any. This aids in placing the accident on a level that is able to fully understand how the accident occurred and what actions are needed to cary out.
03/04/2019	32018382	The relation to the cause of the accident with the road conditions (e.g., wet road)
03/04/2019	32018620	Objective measurements
03/04/2019	32018146	Shows data clearly
03/04/2019	32018127	The understanding of accident severity
03/04/2019	32018177	More understandable
03/04/2019	32017692	Level of severity on a reported accident, in order know if you need to avoid the area when traveling
03/04/2019	32017710	The Type of accident, gives more details on the actual type of accident.
03/04/2019	32017820	I can know how accidents are classified and use it if I'm ever in an accident
03/04/2019	32017810	N/A
03/04/2019	32017428	none
03/04/2019	32017294	more detail
03/04/2019	32017400	better understanding of where major/severe accidents occur and where changes need to be made
03/04/2019	32014905	There are only 4 of them versus 10 ways to classify the accident
03/04/2019	32000329	It could help reporting traffic, as well as be used by maps applications like google maps to help determine a better time estimate on the delay.
02/28/2019	31859458	Advantages that I know a little more about types of traffic accidents
02/28/2019	31859597	No
02/28/2019	31859478	There is more information available about each individual accident
02/28/2019	31859649	It helps you get a better understanding of how traffic accidents are "Measured"
02/28/2019	31859372	You could understand the severity of the accidents and in a way improve traffic in the city
02/28/2019	31859365	Determining the current conditions at the time of the accidents helps to evaluate the causes of the accident.
02/28/2019	31859348	It could be adopted like the richter scale for easier understanding
02/28/2019	31859518	It is way easier specific and faster to categorize the crash. Instead of loosing time in other things
02/28/2019	31859521	It gives better insight to accidents
02/28/2019	31859383	Clear, to the point, easy to understand
02/28/2019	31859339	The advantages of the Critical Composite Index is that it gives an exact description of the accidents and represents in a way most people can understand.
02/28/2019	31859562	It will help people be more informed
02/28/2019	31859712	i think it would benefit a lot of people to know more about this situations.
02/28/2019	31859439	we can understand the garvity of the accident and look forward to solutions for the situation

02/28/2019	31859519	to undertand better how to catalgo the accidents
02/28/2019	31859361	It is way easier to understand and people from all ages would be able to follow on what the CCI is talking about.
02/28/2019	31859358	Learn things I didn't know before about this topic
02/28/2019	31859497	More details
02/28/2019	31859416	It gives a better understanding and breakdown of the type of car crashed that occurred.
02/28/2019	31859447	Like measuring the magnitude of an earthquake, it can help us keep track of possible dangers and how to prevent them according to the CCI.
02/28/2019	31859373	If multiple accidents are reported you are able to see the severity of each individual one instead of seeing the frequencies of what happened. You gain more information on each accident.
02/28/2019	31859539	I believe that it can be useful to know the severity of an accident, the less severe the accident the quicker it can be dismissed. Therefore people may use that when on a commute.
02/28/2019	31859377	having more knowledge about car crashes.
02/28/2019	31859387	Its a simple way to generalize accidents.
02/28/2019	31859420	it helps to determine the severity of road accidents
02/28/2019	31859389	It is specific
02/28/2019	31859415	N/A
02/28/2019	31859408	None
02/28/2019	31859440	It'll be helpful to everyone in the community to follow the Index.
02/28/2019	31859489	It may in the future serve as a guide to classify car accidents
02/28/2019	31859330	You will have more knowledge in car accidents.
02/28/2019	31859332	I see all the different descriptions of the severeness of an accident.
02/28/2019	31859477	.
02/28/2019	31859423	Its very extensive and straight to the point. It will be easier for a lot of people to understand.
02/28/2019	31859394	Being able to classify the severity of an accident an avoid confusion when comparing them
02/28/2019	31859344	Helps understand the severity of the crash based on guidelines.
02/28/2019	31859393	show and explain to everyday commuters the dangers of driving every day
02/28/2019	31859384	Good to see
02/28/2019	31859335	easier to analyze
02/22/2019	31540558	At the same time, it is understandable to have two values (numeric and word description)
02/22/2019	31542303	ranking the accidents and then try to focus on the most severe locations to improve
02/22/2019	31540998	GIVES MORE INSIGHT INTO SPECIFIC COLLISSIONS
02/22/2019	31533072	makes better use of existing data
02/19/2019	31293576	visually easier to classify
02/16/2019	31136113	Outside of giving it a score. Not enough info is provided.
02/15/2019	31097055	in applying this to auto insurance algorithms for determining patterns in accidents.
02/15/2019	31097020	It shows a more detailed representation of the crashes.
02/15/2019	31096819	more factors considered
02/15/2019	31097602	it is more clear
02/15/2019	31097024	It is a faster way to gauge the severity of an accident.
02/15/2019	31096977	It is very simple and clear to understand the data being evaluated.
02/15/2019	31096996	Understanding how severe the accident is.
02/15/2019	31097066	clearer reports ether on the news or official reports for the emergency responders
02/15/2019	31096830	It is easy to use to compare accidents although I would question a lot about how the information was attained.
02/15/2019	31096880	It has more details
02/15/2019	31096823	It let me see the importance of some stuff given

02/15/2019	31097033	It definitely makes it easier to compare one accident to another. It also provides more detail on an accident from a glance.
02/15/2019	31097115	Makes it easier to compare accidents
02/15/2019	31096829	well without the severity chart its useless
02/15/2019	31097116	more easier to separate crashes into categories
02/15/2019	31096870	A more clear understanding of the accident so that I can get a better understanding of what happened.
02/15/2019	31097057	Better understanding
02/15/2019	31097017	The index has a more meaningful message
02/15/2019	31096848	We get to classify the types of accidents better rather than having them reported as crashes.
02/15/2019	31096868	It's much simpler to get a grasp of the important details of a crash. I could see the Index being eventually used in court cases and insurance investigations.
02/15/2019	31096907	Easily readable
02/15/2019	31097088	It is a useful source
02/15/2019	31096836	easier to understand, faster decision making
02/15/2019	31089515	it helps in describing the severity of traffic accident
02/15/2019	31066577	Easy scale to visualize
02/14/2019	31062252	Better understanding of accidents and their reporting.
02/14/2019	31058176	Very clear
02/14/2019	31056549	easier to understand the accident typ
02/14/2019	31054318	N/A
02/14/2019	31039072	Easier to understand due to color scheme
02/14/2019	31039696	I don't know.

21. What area of improvement, if any, can be done to the Critical Composite Index?

		What area of improvement, if any, can be done to the Critical Composite Index?
03/08/2019	32213141	I don’t see any.
03/08/2019	32186669	None
03/07/2019	32182543	It seems to be complex with the variables of how a car accident is rated, so changing the index from a scale that has decimals to a system where each statement has a 0 or 1 answer (i.e "Was the accident during the day? No-0, Yes-1", "How many people were involved? One:1, Two:2, Three:3, Four:4, 5 or more:5" would be an improvement.
03/07/2019	32175993	Perhaps the health of driver or the case in which any substance was used or other factors that may influence driving conditions (just as weather influences).
03/07/2019	32161620	no comment
03/05/2019	32025495	At this time, I do not have a suggestion for the Critical Composite Index.
03/05/2019	32024811	A little more words or guidelines to help people through the process of understanding an accident
03/04/2019	32022076	N/A
03/04/2019	32020867	None
03/04/2019	32020355	N.A.
03/04/2019	32020092	None
03/04/2019	32020001	For the common person i dont feel like this is a good way to describe an accident. But from the data provided and how it is used then i dont believe anything needs to be improved.
03/04/2019	32019904	I don’t know.
03/04/2019	32018394	N/a
03/04/2019	32019740	I think it’s good. Straight to the point
03/04/2019	32019055	I am a bit confused on the numbering/grading system. I think it should just be a simple system of 1-5 (or A, B, C, D, F), without any of the extra fluff incorporated to it. Like what is the difference between 1.45 and 1.39? I also feel, for a minor accident, the total cost of damages should be under $1000. I honestly wouldn't know how much for repairs are, however (how much it would cost to repair a headlight or dent as an example). Also for the specific classifications, it was a little hard to read/located the specific sub-classifications (time of day, weather conditions for example). Maybe if each sub-classification had its corresponding color it would be a bit easier to read.

03/04/2019	32019107	I believe that the situation or the details of the accident could vaguely be included in the Critical Composite Index. Personally experiencing a minor car accident, the car behind me had caught fire on I-54 which was somewhat dangerous since there was low visibility with the oncoming traffic. While the accident was labeled as minor, the conditions were moderately severe.
03/04/2019	32018382	I believe the hour of the day is not important to know the resulting severity of an incident
03/04/2019	32018620	None
03/04/2019	32018146	N/a
03/04/2019	32018127	None
03/04/2019	32018177	None
03/04/2019	32017692	If reported by media, app, etc. how or what would be presented to the public besides the four categories
03/04/2019	32017710	You can add the time frame of the measurement of the accident. was it in a year, month or week etc.
03/04/2019	32017820	Be more specific.
03/04/2019	32017810	N/A
03/04/2019	32017428	none
03/04/2019	32017294	none
03/04/2019	32017400	none
03/04/2019	32014905	add a bit more information to the >value part. or add acronyms using the chart as a key (ex: mi for injuries, si severe injuries) or even a skull and crossbones representing severe and a frown face representing minor injuries
03/04/2019	32000329	Adding a threshold on monetary loss
02/28/2019	31859458	Everything was good
02/28/2019	31859597	Fatalities
02/28/2019	31859478	Unsure how it could be improved
02/28/2019	31859649	I think is good the way it is
02/28/2019	31859372	Traffic
02/28/2019	31859365	Other than adding a number that corresponds to the degree of the accident chart, I do not have any additional suggestions for improvement. I like it a lot.
02/28/2019	31859348	Maybe basing damages as percentage rather than price, because I am a car person and know that I can get charged $200 at a mechanic shop for a fix that I could Achieve for only $30, i.e. replacing a sensor or maybe changing sparkplugs
02/28/2019	31859518	Everything is fine!
02/28/2019	31859521	DNA
02/28/2019	31859383	Adding more criteria to get more accurate results.
02/28/2019	31859339	There isn't any area of improvement that I can see.
02/28/2019	31859562	Adding icons or images
02/28/2019	31859712	i don't think there should be any improvements i think its good just as it is
02/28/2019	31859439	give more examples and real life examples also
02/28/2019	31859519	no
02/28/2019	31859361	I believe just to add a broader range of situations.
02/28/2019	31859358	Include video footage as examples
02/28/2019	31859497	nothing
02/28/2019	31859447	Give us examples of the severity of each scale.
02/28/2019	31859416	I feel like most people rely and trust numbers more than words. So adding something similar to what I mentioned on question 19 would help.
02/28/2019	31859373	I can't think of any improvements.
02/28/2019	31859539	I believe that there shouldn't be weather types involved besides when they are most drastic. If I have minor injuries in harsh weather conditions the accident may not necessarily be caused by the weather itself.

02/28/2019	31859377	None
02/28/2019	31859387	Add more categories to the Critical Composite Index.
02/28/2019	31859420	what if there are certain incidents where it is a combination of multiple factors, such as time of day being under one classification but the weather was not bad and then fell under another classification. this is where i can see issues coming up
02/28/2019	31859389	Care more about the consequences and deceases
02/28/2019	31859415	N/A
02/28/2019	31859408	None needed
02/28/2019	31859440	N/A
02/28/2019	31859489	Be more flexible
02/28/2019	31859330	A bit more of an understanding
02/28/2019	31859332	none
02/28/2019	31859477	.
02/28/2019	31859423	Just maybe add information about the location of the accident.
02/28/2019	31859394	I don't think there are any improvements necessary
02/28/2019	31859344	Taking injuries more into consideration.
02/28/2019	31859393	-
02/28/2019	31859384	none
02/28/2019	31859335	n/a
02/22/2019	31540558	Condensing information
02/22/2019	31542303	N/A
02/22/2019	31540998	I CAN'T THINK OF ANY
02/22/2019	31533072	Stop using the word "accident" to describe crashes or collisions - this absolves drivers of their responsibility.
02/19/2019	31293576	give less importance to the time of day an accident occurred
02/16/2019	31136113	Scale chart and some information on injured, fatalities or damages caused.
02/15/2019	31097055	n/a
02/15/2019	31097020	More combinations of scenarios, the criteria feels a bit too narrow.
02/15/2019	31096819	none
02/15/2019	31097602	cant think of an improvement
02/15/2019	31097024	I think it's good.
02/15/2019	31096977	No improvement, everything is clear and easy to understand.
02/15/2019	31096996	None.
02/15/2019	31097066	more explanation on how to assign points
02/15/2019	31096830	Credibility. How accurate is the data given? Does knowing the amount of accidents really tell you real results?
02/15/2019	31096880	Not only ask if it was at an intersection point, maybe add something like: (if it was at a highway road, school zone,...)
02/15/2019	31096823	none
02/15/2019	31097033	N/A
02/15/2019	31097115	cant think of any
02/15/2019	31096829	none
02/15/2019	31097116	don't see much area for improvement
02/15/2019	31096870	a more colored and visually appealing graph would be useful in knowing where the user is looking at, instead of all cells being one color and having to reference back constantly to remember where the user is at
02/15/2019	31097057	No improvement needed
02/15/2019	31097017	I cannot think of one
02/15/2019	31096848	Have a trial implementation with others and see how others would take in the information as well

02/15/2019	31096868	To move forward with not only the implications of use in the public sector, but also the private, by adding metrics that may be useful to those who might want to use this information for investigative purposes. For example, an accurately reported fatal hit and run might catch the attention of lawyers otherwise unable to reach out to the families of the deceased to ensure full use of the justice system here in America.
02/15/2019	31096907	N/A
02/15/2019	31097088	None
02/15/2019	31096836	a little bit more detail,
02/15/2019	31089515	if its to be adopted for public understanding, the description of each severity should be made simpler.
02/15/2019	31066577	N/A
02/14/2019	31062252	I dont think there needs to be anything more done to this index. It was very knowledgeable and for having very little understanding in this area, I can now say I have a better understanding.
02/14/2019	31058176	Everything is great
02/14/2019	31056549	more knowledge
02/14/2019	31054318	N/A
02/14/2019	31039072	I'm not an expert so I cannot say
02/14/2019	31039696	CCI is too subjective, it should be more detailed by using actual number of people instead of using terms like few or many. Also severity of injury is a subjective way to report injuries.

APPENDIX I – PUBLIC ACCESS

TEX Incidents 2014 (Crash, Person, Secondary Person A-F)
TEX Incidents 2015 (Crash, Person, Secondary Person A-G)
TEX Incidents 2016 (Crash, Person, Secondary Person A-G)
TEX Incidents 2017 (Crash, Person, Secondary Person A-G)
TEX Incidents 2018 (Crash, Person, Secondary Person A-G)
PENN Incidents 2014
Files may be found at: https://minersutep-my.sharepoint.com/:f:/g/personal/dmmejia2_miners_utep_edu/ElXRU4M60yJDug5lXpsz8rkBOGiwZFvC6TiLKbW9G8Dz5w?e=kpJdH6
GitHub – Code: https://github.com/dmmejia2/DissertationCode

Vita

Daniel Mejia is a native of El Paso, TX. He earned his Bachelor of Science degree in Computer Science with a minor in mathematics from The University of Texas at El Paso (UTEP) in May 2015. He further earned his Master of Science degree in Computer Science from UTEP in May 2017. In August 2017 he entered into the Ph.D. Computer Science program.

Daniel participated in the U.S. – Mexico Study abroad program focusing on Smart Cities in the summer of 2016. He also participated in the Interdisciplinary Research Experience for Students (IRES) in the summer of 2018. During these programs he developed a large interest in understanding how Computer Science can play a larger role in understanding and transforming Smart Cities. Daniel was a recipient of the Intel Foundation Scholarship. He has published several peer reviewed papers in conferences including IEEE Smart City Innovation (IEEE SCI) and the International Semantic Web Conference (ISWC); he presented at both the IEEE SCI conference and the ISWC doctoral consortium.

Daniel worked as an associate programmer at GHG Corporation, a lead teaching assistant, and as a research associate with iLink Labs @ Cyber-ShARE. He also served as a guest lecturer in the department of Computer Science.

Contact Information: dmmejia2@gmail.com

This dissertation was typed by Daniel Mejia.