# Weatherlink

Literature Review

Combining demographic, meteorological, and traffic data for insights into emergency response time.

The George Washington University
Alper Cetinkaya
Aditya Sanjay Gujral
Tharun Saravanan
Matthew Winchester

## 1. Introduction

Are you more likely to die in a car crash if it is raining? Is a fire put out slower at night? Does an ambulance arrive too late to neighborhoods with lower median income? These are the types of questions that Weatherlink seeks to answer. By combining traffic reports, accident response times, severe weather instances, and demographic information, we seek to understand and visualize where the intersection of these data sets indicate real problems impacting public safety. This will help government organizations better allocate resources, provide training data and awareness for first responders, and shed light on inefficiencies and inequalities when comparing public safety in areas of disparate socioeconomic status. The combination of these data sources in a way that will prove useful is a technical challenge. The volume of weather data, traffic reports, and demographic census data is large, and datasets vary in format. We will need to fuse data, fill in sparse data gaps, and be able to clean and prepare the data so it can be presented and visualized quickly and adapted to answer various questions. We identified and studied eight papers on related topics to understand current practices. We describe the problems these papers identified, discussed solutions they presented, and mention how we will apply them in our own project.

## 2. Problem space

Following data processing, the data must be analyzed in a meaningful way that appropriately fits the trends present in the data. In the case of traffic accidents, crash frequency can be especially volatile given the nature of weather-related accidents, and thus may not follow a linear pattern. The model analyzing the data trends will need to account for sudden transitions in both weather and census data resulting in larger, more systematic issues pertaining to traffic accidents. Another issue in the same vein is choosing an effective model to make these predictions. For example, while neural networks could provide

high accuracy in classifying the data, approaching the problem with a deep learning model could make more accurate predictions. Thus, choosing a suitable model itself presents a challenge.

Critical issues in the domains of spatio-temporal data analytics and traffic safety were covered in two papers. One of these papers surveys the landscape of spatial and spatio-temporal data analytics, highlighting the need for efficient tools and systems to capture, manage, and analyze vast volumes of spatial and temporal data generated by location-based services and various sources. The other one focuses on the challenge of bridging the gap between statistical modeling of crash risk and prescriptive modeling for safer route optimization in the context of rising global traffic accidents. It aims to leverage data-driven approaches to enhance road safety by examining the disconnect between these two research streams and proposing strategies to integrate their findings effectively.

The paper “Data Integration: The Teenage Years” emphasizes the need to effectively combine data from many sources, especially in large organizations and web search engines. The Information Manifold project mentioned in the paper aims to offer a consistent query interface for many data sources and streamline access and integration. Additionally, it highlights the significance of exact source descriptions and presents the Local-as-View (LAV) method to facilitate this process. The paper examines efforts to enhance adaptive query processing in dynamic situations, lower manual labour, and semi-automate schema mapping. In conclusion, the research provides insights into the advancement, difficulties, and future prospects in data integration, highlighting the field's expanding significance in the contemporary data-driven world.

The paper “A Framework for Scalable Real-Time Anomaly Detection over Voluminous, Geospatial Data Streams” focuses on providing a robust framework to detect anomalies in the geospatial data and tries to support analytical activities by providing resources for analytical activities like visualization. According to the paper, the anomaly is a data point outside the norm e.g. inconsistent sensor readings or an irregular event. The paper uses individual anomaly detector implementations to classify events as anomalous. Observations comprise n-dimensional tuples with each dimension representing a feature of interest. Features include temperature, air pressure, humidity, etc.A feature may also have linear or non-linear relationships with each other; for instance, there may be a relationship between temperature and precipitation at certain geographic locations. These relationships are then used to classify our data into anomalous or normal with a degree of irregularity

Ingesting all of the available weather, traffic, and census data will be no small task. Data from different years varies greatly in size, both in number of files and the size of each file. Processing this data into a useful form so we can visualize it easily will require batch processing these large datasets and storing the data in a way that we can reuse it many times to perform different kinds of analysis and queries. The process will need to be fault tolerant, both for the patch processing, and then storing of the data. If a single process fails while we are ingesting data, we want to avoid having to invalidate the entire dataset and starting over. We will need to homogenize the data, so that it is cleaned into a format that can be queried.

## 3. Solutions

The first paper, "A Survey on Spatio-temporal Data Analytics Systems," provides a comprehensive overview of the landscape of spatial and spatio-temporal data analytics, contributing to the solution of challenges in this domain. By surveying the existing ecosystem, it aids in the identification of efficient tools and systems required for capturing, managing, and analyzing the substantial volumes of spatio-temporal data generated by location-based services and diverse sources.

The second paper, focusing on "Crash Risk Modeling and Safer Routing," takes on the challenge of enhancing traffic safety by bridging the gap between statistical modeling of crash risk. It tackles the pressing issue of rising global traffic accidents by leveraging data-driven approaches. By identifying the disconnect between these two research streams, it offers a path toward more effective integration of research findings. This integration can lead to practical strategies for optimizing routes and reducing crash risks in real-time, thus addressing the alarming increase in traffic-related deaths and economic losses. The authors introduce descriptive analytical methods, including data summarization, visualization, and dimension reduction, as tools to promote safer routing. It also offers code to assist practitioners and researchers in data collection and exploration.

The findings from the paper "Data Integration: The Teenage Years" are very helpful for a project that uses several datasets to examine how the weather affects first responders. The paper's emphasis on data integration, source descriptions, schema mapping, and adaptive query processing will help out with integrating multiple datasets for the project.

It provides useful advice on how to successfully merge different datasets, so the weather information and first responders' reactions may be successfully combined for analysis. The paper also discusses issues with data quality and ambiguity, which are essential when working with real-world information. By utilizing the ideas in the paper, the project will quickly comprehend how the weather affects first responders, resulting in better decision-making and improved emergency response methods.

The techniques described in the paper “A Framework for Scalable Real-Time Anomaly Detection over Voluminous, Geospatial Data Streams” will be useful for cleaning the geolocation data associated with the datasets. These datasets can then be combined. The model defined in the paper is highly scalable making it an ideal choice to work with a giant dataset. The models will autonomously identify anomalies within the defined geographical extents. These anomalies might represent unusual patterns or events in either the weather data or first responders' response data that are potentially linked to weather conditions. The specificity of the classifications can be controlled by adjusting the granularity of the geographical extents or the anomaly detection thresholds.

One potential approach to the issue of non-linearity in the data is utilizing the cusp catastrophe approach. The cusp catastrophe model applies a high-order probability density function to analyze the data, giving it the capability to analyze both linear and non-linear relationships (Theofilatos, 473). In the analysis by Theofilatos, traffic flow was seen to have a “strong-non linear effect” on the number of crashes, while rainfall intensity had a “linear relationship” on the number of crashes (Theofilatos, 476-477) through the use of the model. Despite the fact that cusp models do not provide as robust of a conclusion as more traditional models, we can utilize cusp models as a strong tool to indicate the possibility of a catastrophe.

This could be supplemented by a more robust algorithm in analyzing the data in the form of a clustering algorithm. In an analysis by Gutierrez-Osorio regarding road accident forecasting algorithms, classification algorithms and decision trees were both found to have a high amount of interpretability, at the cost of precision and accuracy in trend analysis. Deep learning algorithms were susceptible to overfitting the presented data. Analyses utilizing Clustering Algorithms however, presented high performance and accuracy, making them a strong choice as a model for our data.

MapReduce is a programming model used for distributed computing. For large datasets like the various sources we have identified for our project, implementing a MapReduce framework will help break down computation into phases that will be distributable and scalable. We can take advantage of the

processing power of many systems to work in parallel on ingesting weather data, cleaning traffic data, and aggregating census data into a form that will be easy to query for visualizations. MapReduce also helps with fault tolerance, and if our system were scaled to a vast network of systems, our overall implementation would not need to change due to the adaptability of MapReduce to any number of nodes. Computation is just one part of the problem however, we will have to store the results of batch processing for visualization.

A distributed file system like Hadoop Distributed File System (HDFS) will allow us to store a large amount of data. We can store the data in HDFS before and after it is processed. HDFS allows for scalability, and data is replicated in chunks across multiple nodes, for reliability. When we do processing and query for visualization, HDFS will allow us to capitalize on the bandwidth and processing power of multiple machines at the same time.

## 4. Proposal

The papers on analytic applications in road traffic safety and spatial-temporal data will serve as valuable resources for gaining a comprehensive understanding of large-scale spatio-temporal data analysis. Additionally, they will aid in the decision-making process regarding the selection of appropriate technologies and architectural frameworks for our project.

The techniques outlined in the papers will help clean the geolocation data based on the scalable model mentioned in the paper. The integration methodologies mentioned in the paper would help merge the multiple datasets to find insights

The cusp catastrophe approach allows for a more comprehensive analysis of trends in the data, and clustering algorithms provide high performance and accuracy in modeling and forecasting road accidents. By using aspects of both models, we can create a more robust framework that improves our ability to identify complex patterns and dependencies within the data, and produce more successful predictions in our accident forecasts.

Existing open source technologies like Hadoop have ecosystems that mesh MapReduce and distributed file systems together. We plan to leverage these and related frameworks so our system can handle large amounts of data flexibly and quickly, while being fault tolerant and scalable. This will allow us to focus on answering and visualizing the public safety questions we are trying to solve while creating a solution that can scale both in terms of processing power and data storage.

## References


1. Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. Journal of Traffic and Transportation Engineering (English Edition), 7(4), 432–446. doi:10.1016/j.jtte.2020.05.002
2. Theofilatos, A. (2019). Utilizing Real-time Traffic and Weather Data to Explore Crash Frequency on Urban Motorways: A Cusp Catastrophe Approach. Transportation Research Procedia, 41, 471–479. doi:10.1016/j.trpro.2019.09.078
3. Alam, M. M., Torgo, L., & Bifet, A. (2022). A survey on spatio-temporal data analytics systems. ACM Computing Surveys, 54(10s), 1-38.
4. Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M. A., Mohabbati-Kalejahi, N., Vinel, A., ... & Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modelling. Sensors, 20(4), 110
5. Data Integration: The Teenage Years Alon Halevy Google Inc. halevy@google.com Anand RajaramanKosmix Corp .anand@kosmix.com Joann Ordille Avaya Labs joann@avaya.com
6. A Framework for Scalable Real-Time Anomaly Detection
over Voluminous, Geospatial Data Streams
Walid Budgaga, Matthew Malensek, Sangmi Lee Pallickara, and Shrideep Pallickara
Computer Science Department, Colorado State University, Fort Collins, CO, USA
7. Jeffrey Dean, Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc.
https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
8. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. 2010. The Hadoop Distributed Filesystem. Apache Foundation. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf