Thesis Proposal: Digital Twin Optimization Through Machine Learning Models for Steps Towards a Smart City

Tim Andersen, Hailee Kiesecker

September 2020

1 Introduction

Since the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), also known as COVID-19, pandemic has started, there has been a slight decrease in air pollution across the United States. Ada County in Boise, Idaho's air pollution levels have gone down 9%-12% in the recent months [7], current date being Wednesday June 25th, 2020. With Boise going into phase 3 of our reintegration traffic is picking back up and with a lot of people doing remote work now, traffic levels have the potential to increase past that of what they were before the start of quarantine, due to workers taking breaks at different times of day than what they would normally be allowed within their places of work. While air pollution caused by traffic is relativity low compared to other pollutant sources as can be seen in figure 1 [17], we should not discredit it as a pollutant and efforts should be made to minimize its affects on the O-zone.

Traffic that remains congested produces higher levels of pollution into the air, "For example, Sjodin et al. (1998) showed up to 4-, 3- and 2-fold increases in CO, HC and NOx emissions, respectively, with congestion (average speed of 13 miles per hour, mph; 1 mph=1.61 km per hour) compared to uncongested conditions (average speed, 38–44 mph)" [22]. Those who live closer to busy roadways are more likely to get ill due to air pollution or have other related health respiratory issues [22]. This means in Boise, Idaho, citizens who live by these intersections Eagle and Fairview, Eagle I-84 WB Ramp, Chinden Glenwood, Chinden and Curtis/V.M.P. Franklin Milwaukee, Eagle Ustick, etc. [1] are at a much higher risk of having respiratory issues and thus might not be able to fight off respiratory diseases and viruses such as COVID-19, as effectively as someone who lives further away from major roadways [7].

If Boise is able to optimize key traffic lights to allow more traffic flow, less congestion, potentially supplying cleaner air quality for its citizens, then we should look into implementation of such technology to allow for it [13]. A potential solution in optimizing key traffic lights within Boise, Idaho is through the use of digital twins.

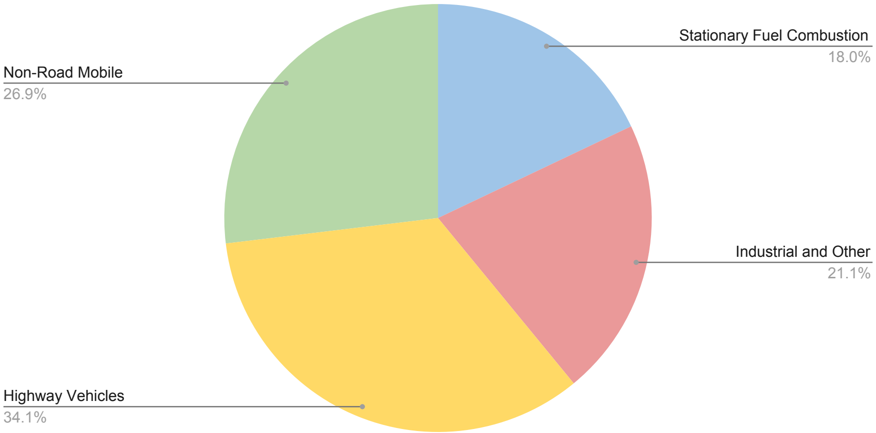


Figure 1: Combined Pollution Sources of CO and NOx

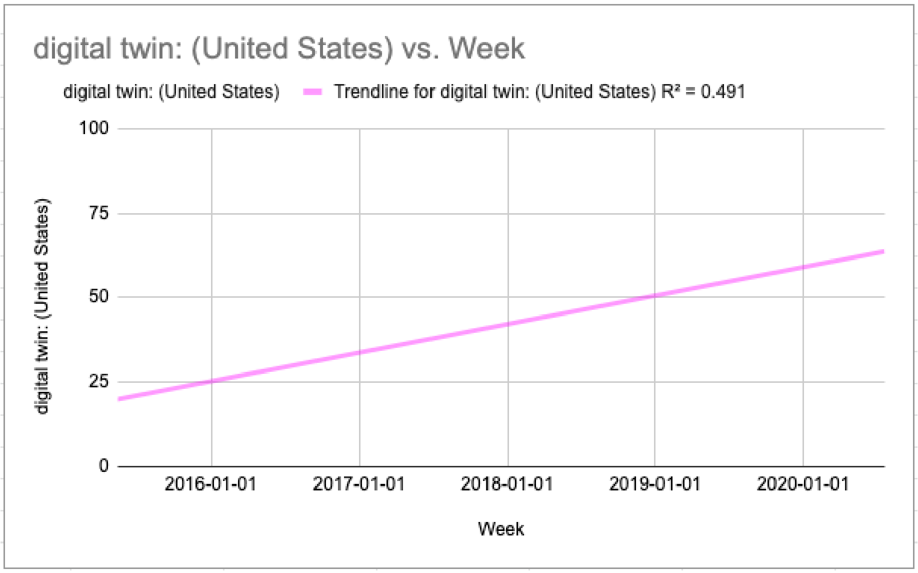


Figure 2: Digital Twin Popularity

1.1 Digital Twins

Our solution to optimizing the current Boise, Idaho traffic system to limit pollution is through the use of digital twins on traffic light signals. We have labeled a successful traffic light signal as a light that limits stop and go traffic that is able to flush out cycles of cars while limiting the number of accidents. It is also able to adjust accordingly to the current traffic demands. The reasoning for limiting stop and go traffic for an optimized successful traffic light is due to the increase of start-stop engines being produced. A start-stop system can be defined as an engine that is able to reduce CO2 emissions by turning off the combustion engine while a vehicle is not moving [15]. Idling cars for long periods of time will slowly become a thing of the past with start-stop technology and instead the combustion from the starting of an engine will be the core source of air pollution at a stopped intersection, this is why we feel in the future it will be better to have longer stops and longer goes that flush out traffic.

> *An optimized model of a digital twin traffic light intersection, is one that is able to limit stop and go traffic. As well as get as many cars through the intersection within one cycle and minimize number of accidents caused by the intersection.*

Digital twins are a new form of technology introduced in 2002 by Greives [5] as a Product Life-cycle Management System. This recently became fruitful in 2010 when National Aeronautics and Space Administration (NASA) used the digital twin ideal to develop a road-map of current technology to be used in the future for high fidelity aircraft digital twins [11] and continues to be explored and improved by companies like Cityzenith [10]. Greives definition of a digital twin is, “a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level. At its optimum, any information that could be obtained from inspecting a physical manufactured product can be obtained from its Digital Twin”[5]. In our words, the purpose of digital twins is to understand the limitations and potential of the physical entity, that allows the largest potential of predicted benefits and minimizes unpredictable, undesirable outcomes. This can be in the form of preventing unexpected malfunctions or downtime as well as finding new opportunities for improvement.

Throughout the years a lot of research has been done in the field of digital twins, rising in key word interest steadily since January 2016 (view 2). Digital twins are a part of the new frontier of industry 4.0 which can be defined as the fourth revolution of industry. The digital twin can currently be used to identify fault diagnostics of physical entities through different means of implementation including deep transfer learning [21], creating greater predicted benefits [16], among other things [3], [2], [12], etc.. The current digital twin research spans through key frameworks of how to create a digital twin, to keeping the previous states of the twin secure through different means such as Blockchain and encryption. Throughout our research however, we discovered that due to the relatively new nature of the research there is a huge variety of definitions and classifications of what a digital twin is. A majority of the current scholarly articles spend much of their time explaining what their version of a digital twin is and how it can benefit companies in the future, along with their potential, few go into case studies of digital twin implementation. Something missing from the current research field of digital twins is how to make them autonomously optimized; with around only one article addressing the issue [14] in which a novel framework was suggested, tested, and produced promising results from the creation of a digital twin for a petrochemical factory. This framework presents how a digital twin can modify current settings and or tools that will allow it to improve itself without the use of an expert constantly having to reevaluate the data. This is done through use of machine learning models being used in unison with the digital twin.

1.2 Further Background

It is to be understood that the term digital twin is not simply a *digital model* which is defined as a manual flow of data from physical object to digital object. Nor is it a *digital shadow* which is defined as a manual flow of data from a digital to a physical object with a automatic flow of data from physical object to digital object, creating a virtual mirror effect. A fully functional digital twin has high levels of fidelity that being a high degree of similarity between the physical entity and the digital one achieved through an automatic flow of data between digital and physical.

1.2.1 Digital Entity

A digital entity is the digital representation of a physical entity. It receives all of its information from sensors, IoT, 5G, and other external tools see section **??** that monitor the Physical entity and communicate it back to the digital one. It is possible for the digital entity of a digital twin to have a digital model attached to it that is able to show a visual of what the physical entity looks like. This can also be shown in augmented reality for a 3D model if the designer so chooses [18].

1.2.2 Physical Entity

A physical entity can be described as the physical system. This for example can be a wind turbine, an engine, a factory floor, etc. Physical entities are monitored by sensors and other external tools (or if you rather linkage) to send back to the Digital entity. It is important to note that these external tools can be used for communication across both physical and digital. Unless within a purely digital simulation.

1.2.3 Use Cases

Anything physical has the potential to be developed into a digital twin. However, the main focus of digital twins currently rest within manufacturing and healthcare fields along with the potential of smart cities. "The International Data Corporation (IDC) forecasts that companies investing in digital twins will see improvements of 30 percent in cycle times of their critical processes in the next five years."[19].

Using digital twins' information about its environment and how important each attribute is to the overall system, we could produce a step towards a smart city digital twin that is able to optimize itself, at least in regard to traffic flow. A smart city is essentially a digital twin, “a city can be defined ‘smart’ if it enhances the quality of living of its citizens by applying synergy of inhabitants’ knowledge, traditional-modern communication infrastructures, information technology, efficient use of natural resources and participatory good governance” [8]. One of the more promising forecasted capabilities of digital twins is a digital city that would be able to be a fully synchronized hub of information that will allow better flow of traffic, cleaner energy sources, with more connected citizens. The current extent of implementation towards a smart city digital twin is equivalent to a skyscraper digital twin located in Singapore. “The digital twin draws on IoT sensors, big data and cloud computing, combined with 3D models, geospatial datasets and BIM. Virtual Singapore was co-developed with the French firm Dassault Systèmes, by leveraging its existing software platform” [3]. It is useful in monitoring real time foot traffic, climate and temperature, and informing how a change would affect the system.

2 Digital Twin Foundation Implementation

Use of digital twins on physical assets like wind turbines and manufacturing equipment already exist and are being used in industry [4]. A base blueprint or overall structure of digital twins that has been presented to the research community and are described bellow.

The three-dimension model architecture for a digital twin is described as including a physical entity, digital entity, and a connection. This connection is characterized as the communication between physical and virtual [9]. The three-dimension model can be extended into a five-dimension model architecture.

With five-dimension, we have the physical asset, digital asset, Sensors, Digital Data Model, and the Connection Model. It is shown in A.Y.C Nee’s paper on digital twin driven prognostics and health management for complex equipment that using the five-factor implementation improves accuracy and optimization due to how it is composed [20].

In regard to the five-dimension framework of a digital twin the sensors digital data model and connection model all have factors feeding into them. Both the three and five dimensions share the same inputs in regard to the digital entity which are: The geometry, physics, behavior, and rules.

The five-dimension framework can be described more complexly than it’s three-dimension counterpart. In short there is a physical asset model(composed of one part), virtual equipment model(composed of four parts), services model(composed of five parts), digital twin data model(composed of five parts) and a connection model(composed of six parts) that integrates all four models together, view the table bellow.

Definitions Table for 5 Factor Framework

Aspect	Parts	Description
Physical Entity	One part composed of different aspects that make up the physical entity	Composed of different physical parts that make up the physical asset
Virtual Entity	Geometric model, physics model, behavior model, and rule model	The virtual entity is a high fidelity model of the physical counterpart
Services	Function, Input, Output, Quality and State	Goal is to optimize the physical asset and ensure high fidelity of virtual asset
Digital Twin Data Model	Data from physical, virtual and services, domain knowledge, fused data of all section aspects	Includes collected data from both physical and virtual entities for fusion to help enrich data observations
Connections	Connection between services and DT data, physical entity and DT data, virtual entity and DT data, physical entity and services, virtual entity and services, finally physical entity and virtual entity, each connection has 4 parts; Data source, Unit, Value, Scope, Sampling interval	Creates a fully connected model from digital to physical and physical to digital

3 Novel Framework

A summary of [14] a Machine Learning Based Digital Twin Framework for Production Optimization in Petrochemical Industry can be described bellow. This journal article is the basis for our execution of a machine learning optimized digital twin for a traffic light system that will provide steps towards a smart city in the future.

The paper presents a theoretical framework for the petrochemical industry where machine learning can be used on a digital twin model smart factory to improve production control, as at the time there was a low level of efficiency, intelligence, and sustainability in the product design, manufacturing and service phases [14]. The paper then goes on to list some research gap questions for the industry, those being along the lines of current data processing methods being isolated and fragmented, as well as few methods to achieve fast and effective interaction between the virtual modes and real environments [14]. The paper then expresses what a successful petrochemical enterprise is and concludes that a digital twin connection between the physical and cyber world is needed for better production. They define success as "can provide a broad variety of high-quality products while keeping manufacturing and distribution costs low to meet customer expectations and needs" [14].

The paper defines its digital twin as being used mainly for production control, that is, able to run eliminating the dependency for expert experience and knowledge. The digital twin iteratively and dynamically generates models according to the changes in the physical environment so that one single model does not crucially influence the factory as a whole. It is noted in the paper that the proposed framework can be used within other process manufacturing industries as a way to improve their economic benefits through production control.

At face value the framework is composed as follows; the basic framework of the factories production line is constructed, this includes the physical factory (production units, chips, production services systems, and environment), the digital factory (virtual model, simulation, validation, digital simulation systems, and operation processes) and the mapping between them (real time data, IoT data collection,production report, etc.). The digital twin model is then trained by machine learning using historical big data of existing industrial systems, as well as production and operation systems. The digital twin is then evaluated and optimized using different series of evaluation indexes. The final optimized model is combined with input market demand and the optimal solution is created by real time big data on the digital twin model, this then feeds back into the distributed control system to guide the production control. That is where the expert knowledge is eliminated. The digital twin will be iteratively trained and optimized from continuously changing data creating a constant loop of optimization between digital and physical systems, or rather "digital twin practice loops" [14]. (Similar to figure 4)

The rest of the paper goes into detail on how this digital twin framework will be implemented as well as describing in more detail what each part of their system is. The five outlined steps to create the digital twin framework are shortened down to; preparation and data collection, data feature engineering, model training and validation, tryout and optimization and finally model online deployment. The last of which is the final created digital twin connecting with the petrochemical industrial IoT to obtain necessary real-time data as input and then output the control command directly to the production line.

The paper concludes with a case study on the MAYA factory from northern china. Four machine learning models were used on its historical data; random forest, adaboost, lightgbm, and xgboost. It was found that the lightgbm had better predicting accuracy and comprehensive performance, therefore it was chosen to be used in the above outlined digital twin framework. Five important controllable indicators were selected for real time control that the digital twin could manipulate at any given time, riser outlet temperature, liquid temperature of fractionator, stabilized tower bottom temperature, settler pressure and regenerator pressure. It was found that when the yields of a specific product are set as the goal of the machine learning model that the digital twin can effectively optimize production control of the MAYA factory. The paper continues on about different benefits of machine learning based digital twins within different manufacturing industries as well as smart cities.

If we take the novel framework presented in [14], we believe that it is possible to implement it onto a city's current traffic light system for optimization to take steps into a smart city digital twin.

4 Thesis Statement

The purpose of this research will be to optimize a digital twin through machine learning to take a step towards a high-fidelity smart city digital twin. We believe that using a novel digital twin framework on a traffic light system could allow traffic congestion levels in a city to drop. Creation of a singular digital twin for one traffic intersection should produce enough data for our machine learning framework to optimize that intersection. See section Digital Twins for our definition of a successful traffic light system.

We believe that after implementation on a singular traffic light intersection the digital twin can then be expanded onto the next light and then the next, so on and so forth until the whole of Boise and its surrounding cities are a part of the digital twin system. Effectively creating a fully optimized traffic light system, a promising big step into its smart city potential. This presents a relativity cost effective solution to limit air pollution levels within the city as all of the major sensors for the traffic lights within Ada county are already in place and their data is already being stored. Using machine learning on this big data produced by the current in place sensors, with some additional streams of information should allow optimization throughout the traffic system without the need for a traffic engineering professional to physically go out and evaluate any given problem intersection. Effectively cutting costs and wasted time.

5 Methods

For the purposes of this thesis it is proposed that one traffic light digital twin intersection be created. Following the given framework from section Novel Framework, and constructed in [14]. The available sensors that are already established by the Ada County Highway District(ACHD) are outlined in Current ACHD Sensors. Additionally to these sensors information for the digital twin would include; current construction projects which can be taken directly ACHD's Roadwork in The Area (RITA) database, current in route city busses that have the potential to block a lane of traffic, time of day, day of week, week of year, road conditions, weather conditions, public and major events or holidays, as well as any additional useful information that we run into during the creation process.

We will try to autonomously report data to a cloud storage provider or local storage depending on the collected data size. For cloud storage, we recommend using google cloud as they have a data flow feature that will allow integration of our machine learning models.

The machine learning model will include streaming real time data from our additional streams of data outlined above as well as the currently established sensors that the Ada County Highway District (ACHD) have in place. Most importantly their Video Image Processor (VIP) sensor information system, view table 5.1.

5.1 Current ACHD Sensors

The Ada County Highway District (ACHD) located in Garden City, Idaho established in the 1970's maintains and operates approximately 2,100 miles of roads and streets in Ada County. It is also the only county wide highway district in Idaho. Five commissioners run the district and are completely decided by the people every two years. Current projects in place by ACHD include but are not limited to, intersection projects, Commuteride; a economically friendly way for people of Ada county to get around, bridge projects, and bike lane projects [6].

ACHD is in control of the roadways, traffic signals, intersections, construction, etc. inside the county limits of Ada. Their current in place sensors for their traffic light system are connected to different types of controllers depending on the intersection. What sensors that are in place also depend on the intersection.

Controllers that ACHD currently have by Traficware include; NTCIP Based Advanced Transportation Controller (ATC), and NTCIP Based TS2 Controller. Their newest controller is Commander 980 ATC. The Commander 980 ATC controller is cross compatible with newer and older versions of NTCIP compliant software. Depending on the type of cabinet that the controller is being stored in different operating modes are able to be in place and is able to run on Linux platforms.

Sensors currently in place at different traffic signal intersections around Ada county supplied by ACHD include; Traficon, Gridsmart, Wavetronix, and Econolite. See 5.1

Sensors

Name	Description	Type
Trafcon	Detects : Counts, Speeds, Classification, Occupancy, Density, Headway, Gap Time Wrong way drivers, sudden speed variant Up to 24 zones VIP3D.1 or Up to 20 zones per camera VIP3D.2 VIP-BIKE Thermal detection for bikers (configure range and frame rate) VIP-PTZ Auto incident detection, if not in camera view operator must zoom to incident area	One camera per approach
Gridsmart	$intersectionview \rightarrow fisheyestyle$ Does Not need to be moved or zoomed by an operator as it has full view The Bell Camera is powered from a single-wire, Power over Ethernet (POE) connection by the GS2 Processor. GS2, is the intelligence of the GRIDSMART System, which runs the GRIDSMART Engine Able to build 3d model of approaching objects, therefore can count cars, trucks, potentially speed, emergency vehicles and accidents. every night data is automatically uploaded to GRIDSMART cloud (Gridsmart Client)	One Camera for Whole Intersection
Wavetronix	Detect vehicles with the reliability of radar electromagnetic wave is much larger than the wavelength of light so radar can propagate through rain, snow, fog and even dust storms without becoming distorted. Dynamic Zones activated based on a vehicle's speed, range and ETA. Only seen when activated by a loop so there can be cars that the intersection Does not know about because they are no longer activating the loop or, they can pose as three cars at once depending on the loop. When SmartSensor Advance detects a vehicle that meets the user-defined criteria, it sends a call to the controller, which then extends the green light.	One camera per approach
Econlite	One of the newest for ACHD Assuming AccuScan 600c average 600ft radar Stop Line Detection lane-marking painted perpendicularly to the driving direction of the road Best used for adaptive control systems Radar detection sensors	One Camera for approach

Whenever a sensor goes off, it is reported to the controller, the controller then decides what the best course of action is to take for the traffic light signal based on its programming.

//I want to reference a sensor power point here or user guides (I want to include and reference artifacts)

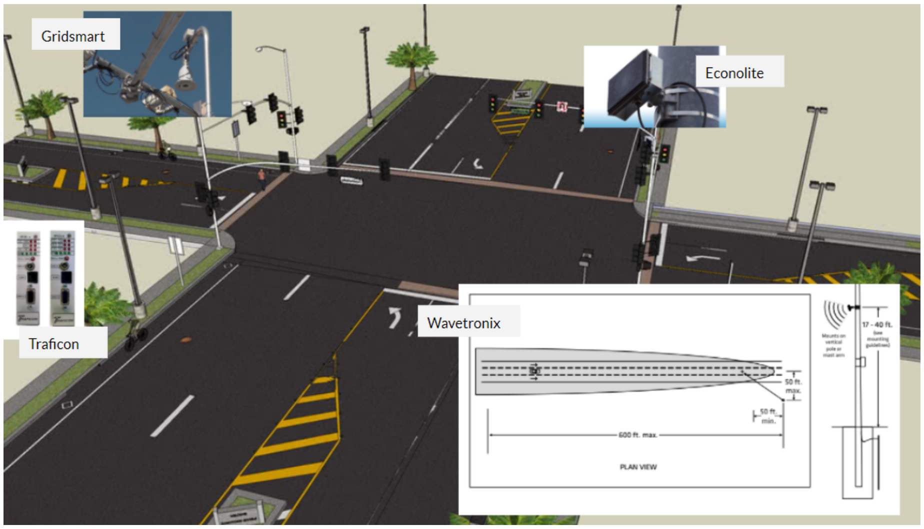


Figure 3: Intro to ACHD Sensor Imagery

5.2 Procedure

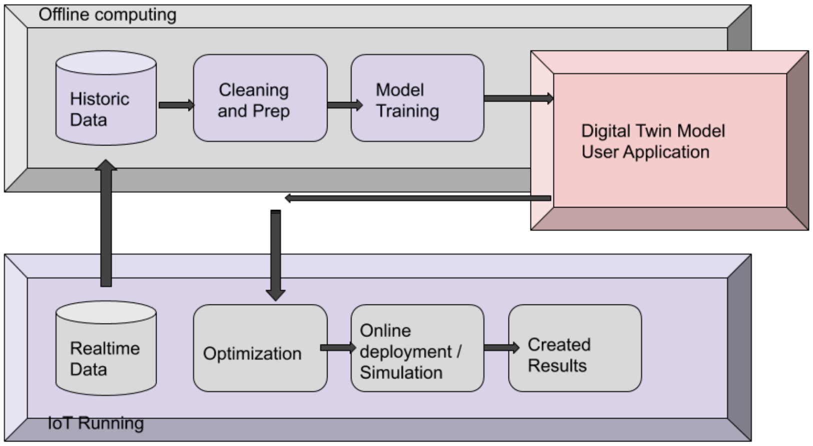


Figure 4: Digital Twin Novel Framework Base

After cleaning and reviewing all data from our multiple data sources, our data input streams will be inputted into a machine learning regression model for training. The training function is defined in equation 1 . Where Y is the yield, F is the training function, X are the controllable independent variables, and Z are the uncontrollable indicators with $t \pm \Delta$ being the given time to get those attributes lag. The attributes being used within this function need to be played around with to see what has the greatest impact on the intersection towards our defined successful traffic light.

$$Y_t = F(X_{t \pm \Delta} + Z_{t \pm \Delta}) \tag{1}$$

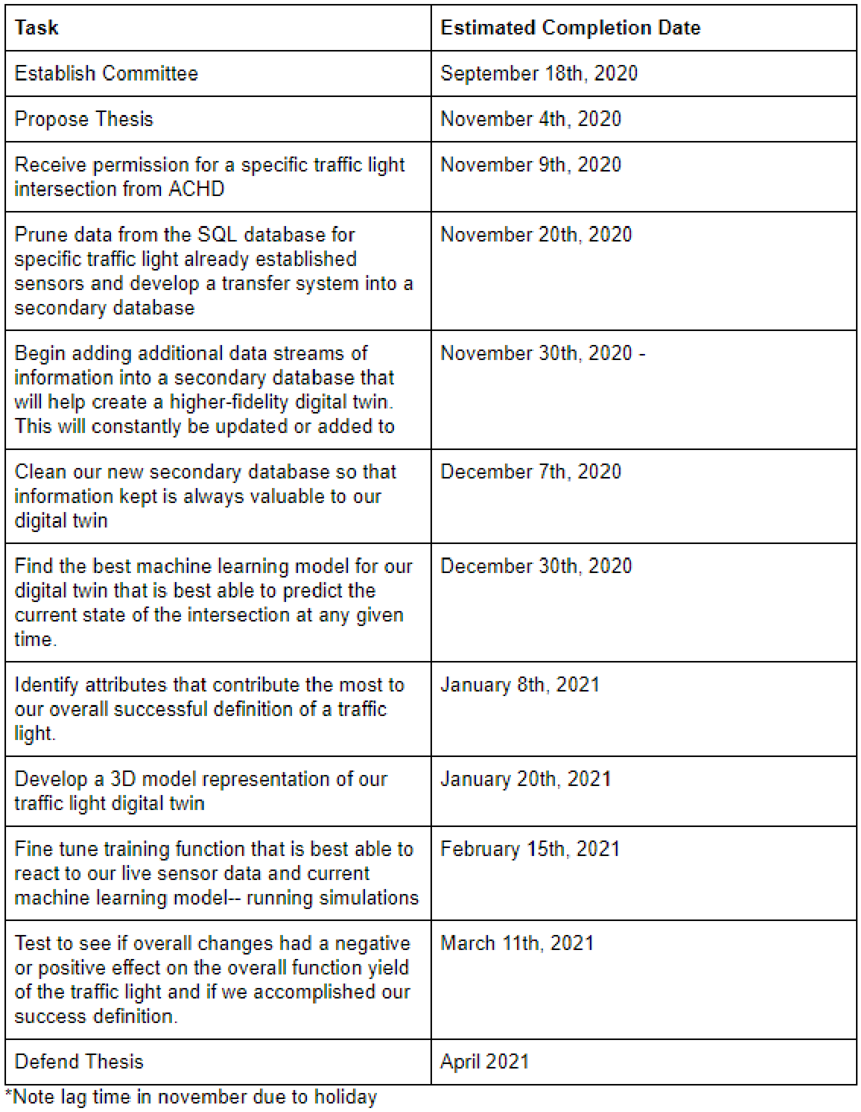
Different machine learning algorithms can be used during the learning process to see which produce the best results. For example, random tree, gradient boosting decision tree, and neural networks, all can be potential functions that have the potential to produce promising results.

A max value algorithm needs to be constructed that will find the root optimization point. After using the base machine learning model on historic data, we can begin to find optimization capabilities given certain conditions on a real time data stream from the physical entity, i.e. simulation. If the goal is to find our *successful traffic light*, then the machine learning target will try to find optimization in settings or attributes which directly relate to that through a search algorithm. Different search algorithms can be tested to see which produce the most optimal results; for example, Breadth First Search, Depth First Search and others. When the attributes are found which directly affect the successful factors, we can modify those settings and run it through our simulator to see the predicted outcome. This should be able to tell us if any optimization and limiting of congestion has taken form.

//I feel like this section is important but not executed correctly, Any suggestions?

This machine learning process is not stagnant, the machine learning model will constantly be looping to see if an improvement can be made at any given time. The current ACHD sensors run at a reporting interval time at 30 seconds to 15 minutes with their most recent technology reporting every 0.1 second intervals however access to this data is not possible yet. It is a goal of this project to be able to run our machine learning loop in as close to real time as possible so that the traffic light is reacting to its environment in the closest time intervals possible to reach our success definition.

6 Schedule



Task	Estimated Completion Date
Establish Committee	September 18th, 2020
Propose Thesis	November 4th, 2020
Receive permission for a specific traffic light intersection from ACHD	November 9th, 2020
Prune data from the SQL database for specific traffic light already established sensors and develop a transfer system into a secondary database	November 20th, 2020
Begin adding additional data streams of information into a secondary database that will help create a higher-fidelity digital twin. This will constantly be updated or added to	November 30th, 2020 -
Clean our new secondary database so that information kept is always valuable to our digital twin	December 7th, 2020
Find the best machine learning model for our digital twin that is best able to predict the current state of the intersection at any given time.	December 30th, 2020
Identify attributes that contribute the most to our overall successful definition of a traffic light.	January 8th, 2021
Develop a 3D model representation of our traffic light digital twin	January 20th, 2021
Fine tune training function that is best able to react to our live sensor data and current machine learning model-- running simulations	February 15th, 2021
Test to see if overall changes had a negative or positive effect on the overall function yield of the traffic light and if we accomplished our success definition.	March 11th, 2021
Defend Thesis	April 2021

*Note lag time in november due to holiday

The above schedule is the plan of completion for my thesis.

References

[1] Ada county's busiest intersections ranked.

[2] Digital twin | GE digital.

[3] *Digital twin TOWARDS A MEANINGFUL FRAMEWORK.* ARUP.

[4] Improving wind power with digital twin technology: Ge renewable energy.

[5] (PDF) origins of the digital twin concept.

[6] Projects and studies.

[7] Traffic is way down because of lockdown, but air pollution? not so much.

[8] Kazi Masudul Alam and Abdulmotaleb El Saddik. C2ps: A digital twin architecture reference model for the cloud-based cyber-physical systems. 5:2050–2062.

[9] Aidan Fuller, Zhong Fan, Charles Day, and Chris Barlow. Digital twin: Enabling technologies, challenges and open research, 2019.

[10] GISuser. Press, Jul 2020.

[11] Edward Glaessgen and David Stargel. The digital twin paradigm for future NASA and u.s. air force vehicles. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. American Institute of Aeronautics and Astronautics.

[12] Edward Glaessgen and David Stargel. The digital twin paradigm for future nasa and us air force vehicles. In *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*, page 1818, 2012.

[13] Jiong Jin, Jayavardhana Gubbi, Slaven Marusic, and Marimuthu Palaniswami. An information framework for creating a smart city through internet of things. 1(2):112–121.

[14] Qingfei Min, Yangguang Lu, Zhiyong Liu, Chao Su, and Bo Wang. Machine learning based digital twin framework for production optimization in petrochemical industry. 49:502–519.

[15] Norbert Mueller, Steffen Strauss, Stefan Tumback, Guo-Chang Goh, and Ansgar Christ. Next generation engine start/stop systems: "free-wheeling". *SAE International Journal of Engines*, 4:874–887, 06 2011.

[16] Q. Qi and F. Tao. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, 6:3585–3593, 2018.

[17] U. S. EPA Office of Air and Radiation. Air quality trends show clean air progress.

[18] Greyce Schroeder, Charles Steinmetz, Carlos Eduardo Pereira, Ivan Muller, Natanael Garcia, Danubia Espindola, and Ricardo Rodrigues. Visualising the digital twin using web services and augmented reality. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pages 522–527. ISSN: 2378-363X.

[19] G. Shao, S. Jain, C. Laroque, L. H. Lee, P. Lendermann, and O. Rose. Digital twin for smart manufacturing: The simulation aspect. In *2019 Winter Simulation Conference (WSC)*, pages 2085–2098, 2019.

[20] Fei Tao, Meng Zhang, Yushan Liu, and A. Y. C. Nee. Digital twin driven prognostics and health management for complex equipment. 67(1):169–172.

[21] Y. Xu, Y. Sun, X. Liu, and Y. Zheng. A digital-twin-assisted fault diagnosis using deep transfer learning. *IEEE Access*, 7:19990–19999, 2019.

[22] Kai Zhang and Stuart Batterman. Air pollution and health risks due to vehicle traffic. 0:307–316.