MACHINE LEARNING CAPSTONE
Complete Documentation

Personnel Directory

- Stakeholders
    - Sandra Batista
        - Email: sbatista@usc.edu
        - Organization: USC Viterbi School of Engineering
    - Kevin Lee
        - Email: kevin@mobilizedconstruction.com
        - Organization: Mobilized Construction
    - Jens Egholm pederson
        - Email: jens@mobilizedconstruction.com
        - Organization: Mobilized Construction
- Developers
    - Andrew Zolintakis
        - Email: azolinta@usc.edu
        - Organization: USC Viterbi School of Engineering
    - Ernest Xu
        - Email: jiachanx@usc.edu
        - Organization: USC Viterbi School of Engineering
    - Lisa Steinhubl
        - Email: steinhub@usc.edu
        - Organization: USC Viterbi School of Engineering

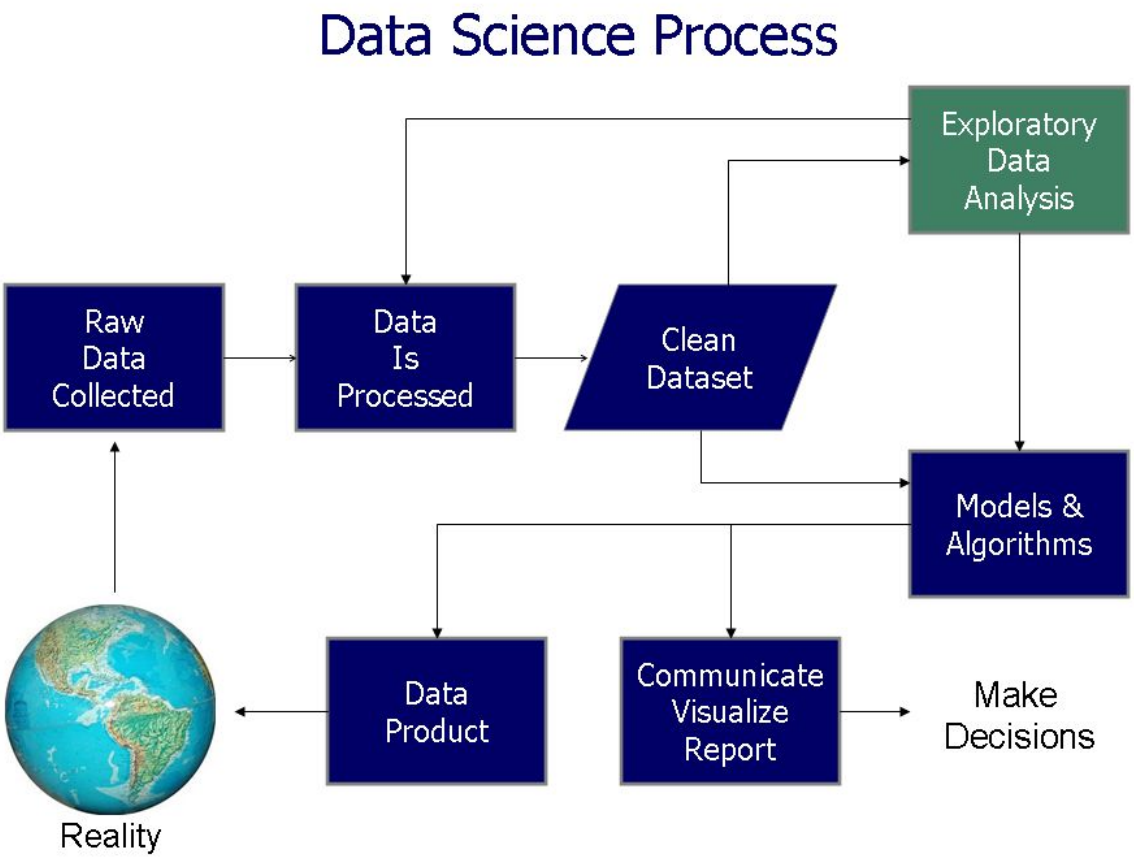
GitHub File Directory [link]

- Speed Bump Tracker iOS App [link]
- Raw Data Files [link]
    - Kenya Labeled Data File: *kenya_oct_15_data_labeled.csv*
    - Kyrgyzstan Raw Data File: *kyrgyzstan.csv*
    - Los Angeles Raw Data File No.1: *los_angeles_1.csv*
    - Los Angeles Raw Data File No.2: *los_angeles_2.csv*
    - Los Angeles Raw Data File No.3: *los_angeles_3.csv*
    - Los Angeles Raw Data File No.4: *los_angeles_4.csv*
    - Los Angeles Raw Data File No.5: *los_angeles_5.csv*
    - Los Angeles Raw Data File No.6: *los_angeles_6.csv*
    - Los Angeles Raw Data File No.7: *los_angeles_7.csv*
    - Los Angeles Raw Data File No.8: *los_angeles_8.csv*

    - Los Angeles Raw Data File No.9: *los_angeles_9.csv*
    - Los Angeles Raw Data File No.10: *los_angeles_10.csv*
    - Los Angeles Raw Data File No.11: *los_angeles_11.csv*
    - Los Angeles Raw Data File No.12: *los_angeles_12.csv*
    - Los Angeles Raw Data File No.13: *los_angeles_13.csv*
    - Los Angeles Raw Data File No.14: *los_angeles_14.csv*
    - Los Angeles Raw Data File No.15: *los_angeles_15.csv*
    - Los Angeles Raw Data File with Video: *los_angeles_video.csv*
    - R script to calculate incremental change in vertical acceleration: *calculate_jolt.R*
    - R script to find possible locations of speed bumps: *find_speedbump.R*
- Documentations [link]
    - Machine Learning Capstone Deliverable 1: *ML Capstone Deliverable 1.pdf*
    - Machine Learning Capstone Deliverable 2: *ML Capstone Deliverable 2.pdf*
    - Machine Learning Capstone Deliverable 3: *ML Capstone Deliverable 3.pdf*
    - Machine Learning Capstone Deliverable 4: *ML Capstone Deliverable 4.pdf*
    - Machine Learning Capstone Deliverable 5: *ML Capstone Deliverable 5.pdf*
    - Machine Learning Capstone Deliverable 6: *ML Capstone Deliverable 6.pdf*
- Exploratory Data Analysis [link]
    - Exploratory Data Analysis on Kenya Labeled Data File [link]
        - Input Labeled Data File: *kenya_oct_15_data_labeled.csv*
        - R Markdown File: *kenya_oct_15_data_processed.Rmd*
        - Processed Data File: *kenya_oct_15_data_processed.csv*
        - Output EDA File: *kenya_oct_15_data_processed.pdf*
    - Exploratory Data Analysis on Kyrgyzstan Raw Data File [link]
        - Input Raw Data File: *kyrgyzstan.csv*
        - R Markdown File: *kyrgyzstan.Rmd*
        - Output EDA File: *kyrgyzstan.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.5 [link]
        - Input Raw Data File: *los_angeles_5.csv*
        - R Markdown File: *los_angeles_5.Rmd*
        - Output EDA File: *los_angeles_5.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.10 [link]
        - Input Raw Data File: *los_angeles_10.csv*
        - R Markdown File: *los_angeles_10.Rmd*
        - Labeled Data File: *los_angeles_10_labeled.csv*
        - Output EDA File: *los_angeles_10.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.11 [link]
        - Input Raw Data File: *los_angeles_11.csv*
        - R Markdown File: *los_angeles_11.Rmd*

        - Labeled Data File: *los_angeles_11_labeled.csv*
        - Output EDA File: *los_angeles_11.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.12 [link]
        - Input Raw Data File: *los_angeles_12.csv*
        - R Markdown File: *los_angeles_12.Rmd*
        - Labeled Data File: *los_angeles_12_labeled.csv*
        - Output EDA File: *los_angeles_12.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.13 [link]
        - Input Raw Data File: *los_angeles_13.csv*
        - R Markdown File: *los_angeles_13.Rmd*
        - Labeled Data File: *los_angeles_13_labeled.csv*
        - Output EDA File: *los_angeles_13.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File No.14 [link]
        - Input Raw Data File: *los_angeles_14.csv*
        - R Markdown File: *los_angeles_14.Rmd*
        - Labeled Data File: *los_angeles_14_labeled.csv*
        - Output EDA File: *los_angeles_14.pdf*
    - Exploratory Data Analysis on Los Angeles Raw Data File with Video [link]
        - Input Raw Data File: *los_angeles_video.csv*
        - R Markdown File: *los_angeles_video.Rmd*
        - Labeled Data File: *los_angeles_video_labeled.csv*
        - Output EDA File: *los_angeles_video.pdf*
- Experimentations of sklearn Models on Los Angeles Data [link]
    - Decision Tree Classifier: *sklearn_DTClassifier.py*
    - Extra Trees Classifier: *sklearn_ETClassifier.py*
    - Gradient Boosting Classifier: *sklearn_GBClassifier.py*
    - Logistic Regression: *sklearn_Logistic.py*
    - Multi-Layer Perceptron Classifier: *sklearn_MLPClassifier.py*
    - Random Forest Classifier: *sklearn_RFClassifier.py*
    - Exhaustive Grid Search for Hyperparameter Tuning: *sklearn_CVGrid.py*
    - Randomized Search for Hyperparameter Tuning: *sklearn_CVRandomized.py*
- Optimal sklearn.DTClassifier Model on Los Angeles Data [link]
    - Optimal Decision Tree Classifier: *sklearn_DTClassifier.py*

Development Methodology: Data Science Process

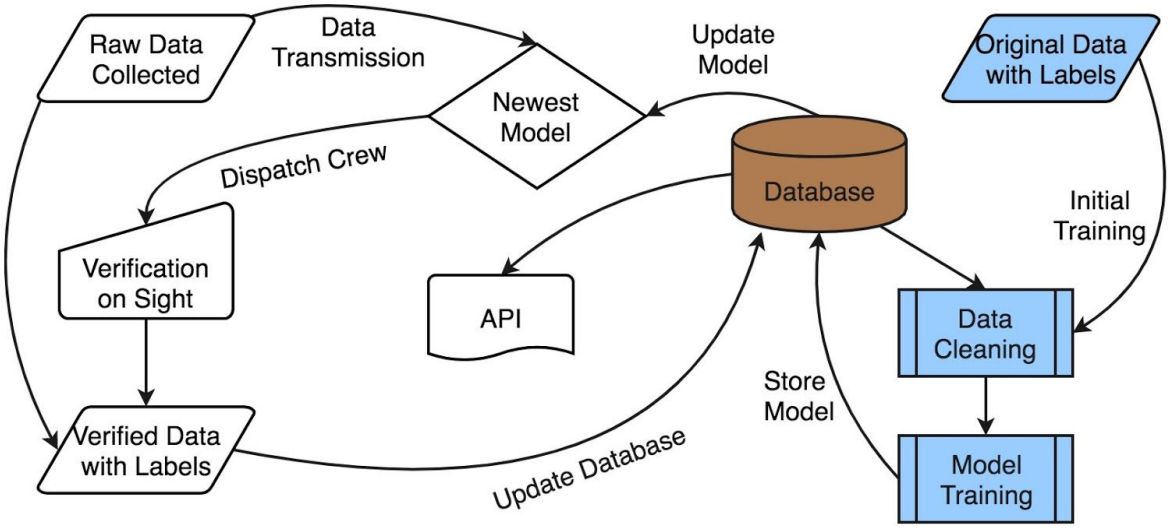


- Data Collection
  - Android App Developed by Mobilized Construction
  - iOS App Developed by USC Development Team: *Speed Bump Tracker*
  - iOS App from the App Store: *Gauges* [link]
- Exploratory Data Analysis
  - Data Verification
    - Step 1: run *find_speedbump.R* on a raw data file to find possible locations of speed bumps using normal distribution
    - Step 2: verify the labels of speed bumps by cross-referencing with timestamps, GPS coordinates, and record videos
  - Data Cleaning
    - *Date*: the timestamp of a data point
    - *Latitude*: the latitude component of the GPS coordinate
    - *Longitude*: the longitude component of the GPS coordinate
    - *Speed*: traveling speed of the vehicle
    - *speedbump*: whether this data point is a speedbump
    - *forw_accel*: forward acceleration (front and back) of the vehicle

        - *hori_accel*: horizontal acceleration (left and right) of the vehicle
        - *vert_accel_G*: vertical acceleration (up and down) of the vehicle with natural gravity (-1G)
        - *vert_accel*: vertical acceleration (up and down) of the vehicle without natural gravity
        - *vert_accel_ratio_speed*: the ratio between vertical acceleration (without natural gravity) and traveling speed
        - *sq_vert_accel_ratio_speed*: the ratio between squared vertical acceleration (without natural gravity) and traveling speed
        - *vert_jolt*: vertical jolt of the vehicle (incremental change of vertical acceleration)
        - *vert_jolt_ratio_speed*: the ratio between vertical jolt and traveling speed
        - *sq_vert_jolt_ratio_speed*: the ratio between squared vertical jolt and traveling speed
        - *vert_jolt_mean*: 5-sliding-window mean of vertical jolt
        - *vert_jolt_sd*: 5-sliding-window standard deviation of vertical jolt
        - *vert_jolt_min*: 5-sliding-window minimum of vertical jolt
        - *vert_jolt_max*: 5-sliding-window maximum of vertical jolt
        - *vert_jolt_range*: 5-sliding-window range of vertical jolt
    - Data Visualization
        - Epoch 1: display of vertical acceleration with natural gravity
        - Epoch 2: display of vertical acceleration without natural gravity
        - Epoch 3: display of ratio between vertical acceleration (without natural gravity) and traveling speed
        - Epoch 4: display of ratio between squared vertical acceleration (without natural gravity) and traveling speed
        - Epoch 5: display of vertical jolt (incremental change of vertical acceleration)
        - Epoch 6: display of ratio between vertical jolt and traveling speed
        - Epoch 7: display of ratio between squared vertical jolt and traveling speed
        - Epoch 8: display of sliding-window statistics of vertical jolt
        - Epoch 9: comparative display of two most promising factors
            - Ratio of squared vertical acceleration and speed
            - Ratio of squared vertical jolt and speed (negatively flipped)
- Machine Learning Models
    - Decision Tree Classifier: *sklearn.tree.DecisionTreeClassifier*
    - Extra Trees Classifier: *sklearn.ensemble.ExtraTreesClassifier*
    - Gradient Boosting Classifier: *sklearn.ensemble.GradientBoostingClassifier*
    - Logistic Regression: *sklearn.linear_model.LogisticRegression*

    - Multi-Layer Perceptron Classifier: *sklearn.ensemble.MLPClassifier*
    - Random Forest Classifier: *sklearn.ensemble.RandomForestClassifier*
- Hyperparameter Tuning
    - Exhaustive Grid Search: *sklearn.model_selection.GridSearchCV*
    - Randomized Parameter Optimization: *sklearn.model_selection.RandomizedSearchCV*

Database-Centric Architecture



The graph above demonstrates our understanding of the overall architecture of Mobilized Construction, and how this machine learning capstone project fits into the overall architecture. Mobilized Construction (the Company) possesses a database-centric architecture, because the centerpiece of the Company is the Database that stores correctly labeled historical road data, and the newest machine learning model.

At the beginning of each day, the Company's Database updates the Newest Model Node. Whenever the motion sensors installed on the delivery trucks, etc. transmit new data to the Company, the Newest Model Node finds possible locations of road hazards, such as speed bumps, by identifying anomaly reading. Maintenance crews are dispatched to those possible locations of road hazards, and provide verification on sight. Manual verification is sent back to the Company, along with its corresponding raw data, producing verified data with labels, updating the Database. At the ending of each day, the Database initiates a new cycle of data cleaning and model training, and update the Newest Model Node.

This machine learning capstone project is colored in skyblue in the graph above. The objective of this machine learning capstone project is to produce an initial machine learning model that jump starts the database-centric architecture of the Company.

Academic References

- USC Classes:
    - CSCI 103: Introduction to Programming
    - CSCI 104: Data Structure and Object Oriented Design
    - CSCI 201: Principles of Software Development
    - CSCI 310: Software Engineering
    - CSCI 360: Introduction to Artificial Intelligence
    - MATH 225: Linear Algebra and Linear Differential Equations
    - MATH 125: Calculus I; MATH 126: Calculus II; MATH 127: Calculus III
- Non-USC Tutorials:
    - Data Science Specialization by John Hopkins University on Coursera
    - YouTube Channel: *Siraj Raval*
    - *Deep Learning* published by MIT Press