NIST IAD DSE Pilot Evaluation Plan

Version 1.8, Updated on September 21, 2016

1 Introduction

This document describes the plan for the National Institute of Standards and Technology (NIST) Information Access Division (IAD) Data Science Evaluation (DSE) Series Pilot Evaluation to be held in fall 2016. This pilot evaluation is a continuation and extension of the Data Science Pre-Pilot Evaluation that was run in 2015, and precedes the first full evaluation of the DSE series. The primary goals of the pilot are to:

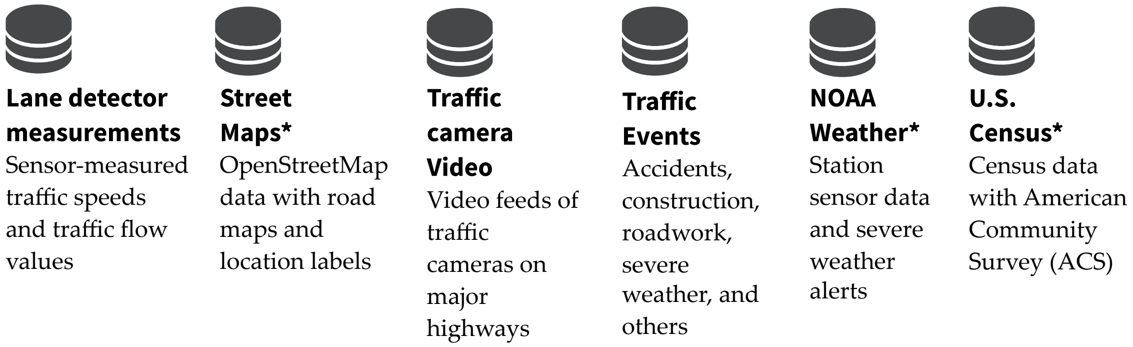
- further develop and exercise the evaluation process at NIST in the context of Data Science,
- provide participants the opportunity to exercise the evaluation process prior to participating in larger-scale evaluation,
- serve as an archetype for the development of future evaluation tasks, datasets, and metrics,
- establish baseline performance measurements,
- bring to the fore new measurement methods and techniques that might be applied to a broad range of use cases, regardless of data type and structure.

The pilot evaluation is set in the automotive traffic domain. This is because the domain is relevant to a broad audience and because large amounts of data are publicly available. However, it is expected that many of the algorithms and techniques to be evaluated (as well as the evaluation approaches and metrics themselves) will generalize to other domains.

The tasks in this evaluation are briefly enumerated below, with specific application to the traffic domain as an example of each case.

1. **Cleaning.** Detecting and correcting errors, omissions, and inconsistencies in data or across datasets, for example, detecting and correcting errors in traffic lane detector data flow values.
2. **Alignment.** Relating different representations of the same object, for example, matching traffic events with traffic video segments containing those events.
3. **Prediction.** Making decisions about future values of a variable of interest, for example, inferring the number and types of traffic incidents in an upcoming time interval based on historical data.
4. **Forecasting.** Determining future (time-stamped) values of a variable within a given time interval, e.g., determining traffic flow values for a set of lane detectors and their corresponding timestamps.

For the purpose of this document, the difference between prediction and forecasting resides in the type of variables under scrutiny: a set a values for the prediction task and a timeseries for the forecasting task. Each task can be completed independently; however, the tasks were designed so that the output of the cleaning task may be used as an input to the remaining tasks. As a result, submissions will be in multiple stages to better evaluate the individual components and the pipelining of the components. The different stages are described more in Section 8. The evaluation sets a number of rules in Section 9, some of which are repeated and emphasized in the task descriptions when appropriate.



Lane detector measurements Sensor-measured traffic speeds and traffic flow values

Street Maps* OpenStreetMap data with road maps and location labels

Traffic camera Video Video feeds of traffic cameras on major highways

Traffic Events Accidents, construction, roadwork, severe weather, and others

NOAA Weather* Station sensor data and severe weather alerts

U.S. Census* Census data with American Community Survey (ACS)

Figure 1: Six types of data are available in the pilot evaluation: Lane detector measurements, Maps, Traffic camera video, Traffic events, Weather data, and U.S. Census data. Each dataset marked with a * above will be made available through its respective organization's url. The other datasets will be supplied through an Amazon S3 Storage bucket available only to registered participants.

2 Data for Traffic Use Case

The section presents data used for the traffic use case in this pilot evaluation. Six datasets form a common core of data that are available for training and development for all tasks in the traffic domain (with the exception of the cleaning task, which limits the data available; see Section 3.1 for details). Figure 1 illustrates the common core of data. Appendix A provides more details about the content of traffic-related datasets, including size estimates and the organization of the data on the Amazon S3 storage bucket[1].

Traffic camera video have been provided courtesy of the Maryland Department of Transportation. Traffic detector data and traffic events have been provided courtesy of the CATT Lab (The Center for Advanced Transportation Technology Laboratory).

In addition to submitting systems, participants are required to submit system descriptions that describe the implementations and the data sources used. Specifications of system descriptions are given in Section 7.

3 Cleaning

In the cleaning tasks, participants are asked to hypothesize and process errors within a set of records. These tasks are common to a wide range of Data Science problems including outlier detection and correction, handling of missing values, and elimination of duplicates. Such problems are relevant to a variety of applications including database integration, and patient matching. Cleaning may also serve as a precursor to downstream tasks such as prediction, forecasting, or classification.

3.1 Cleaning in the traffic domain

In the traffic domain, the task of *cleaning* is applied to a set of traffic flow values from two different perspectives, which form two different tasks:

1. **Error Detection.** Classification of each traffic flow value and speed according to the degree (a confidence value) to which these values are hypothesized to be erroneous; the intent is that the confidence values might serve as input to a subsequent filtering process (although filtering will not be required for the pilot evaluation).

[1] See `https://aws.amazon.com/s3/` for information about Amazon S3 Storage and the Disclaimer on page 18.

2. **Error Correction.** Finding and replacing traffic lane detector data flow values that are hypothesized to be erroneous; the intent is for the correct measurement values to be provided in all cases regardless of the original values provided.

These two tasks are summarized more specifically below:

Summary 1: Cleaning, traffic: error detection task

This task involves determining which traffic lane detector data flow values were hand-manipulated to be erroneous.

Input. Traffic lane detector measurements with erroneous traffic flow values and speeds.

Output. A confidence value for each traffic lane detector measurement, where a greater confidence value indicates a greater belief that the measurement was manipulated to be erroneous.

Trial File(s). `cleaning_test_`*yy*`_`*mm*`.tsv`.

Submission File(s). `detection_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}.tsv`.

Summary 2: Cleaning, traffic: error correction task

This task involves replacing traffic lane detector data flow values that are believed to have been manipulated to be erroneous.

Input. Traffic lane detector measurements with erroneous traffic flow values and speeds.

Output. Traffic lane detector measurements with corrected traffic flow values.

Trial File(s). `cleaning_test_`*yy*`_`*mm*`.tsv`.

Submission File(s). `correction_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}_{`*metric*`}.tsv`.

The input to both of these tasks is a set of time-stamped and location-specific traffic lane detector data flow values and speeds, a subset of which have been to be erroneous. In the error detection task, for each specific flow value and speed, a classification and confidence value are to be calculated on the basis of other measurements and, accordingly; these then may serve as guidance for replacement of errors with corrected values. The error correction task requires the replacement of flow values that are hypothesized to be erroneous with ones that are believed to be correct. (Values that are not hypothesized to be erroneous may be left unchanged.)

For these tasks, it is important to note that errors were synthetically introduced. See Appendix D.1 for information on how errors are introduced into the test data for the cleaning task.

3.1.1 Training data, cleaning: traffic, both tasks

The error detection and error correction tasks are unsupervised, meaning that participants will not have access to non-erroneous traffic lane detector data. However, other non-erroneous data will be available for these tasks, namely:

- **Traffic lane detector inventory.** The traffic lane detector inventory provides detector metadata, including each detector's lane ID, zone ID, road, direction, and coordinates.
- **OpenStreetMap data**. This data includes the maps available from OpenStreetMap.org.

Unlike most other tasks in the pilot evaluation, additional data sets are not allowed to be used. This point is highlighted in Rule 1.

Rule 1: Cleaning tasks rule: allowable data

Participants may only use data from the **traffic lane detector inventory**, **traffic lane detector measurements**, and the **OpenStreetMap** data sets to develop their systems for both the error detection and error correction tasks.

3.1.2 Test data, cleaning: traffic, both tasks

The traffic lane detector measurements in the core data serve as the test set for the cleaning task. In these tasks, participants are allowed to interact with the traffic lane detector data. This is an exception to a general rule in Section 9 that restricts participants' interaction with the test data (in general, participants are not allowed to interact with the test data). Furthermore, since these traffic lane detector data serve as training data—as opposed to test data—for the other tasks, participants will be allowed to interact with these traffic lane detector data.

It is important to note that the traffic speed measurements (given in miles per hour) of the sensors also contain errors. See Appendix B.2 (Figure 3 on page 23) for the organizational structure of the core data (where the cleaning test data is located) and where to find the filtering test data on Amazon S3. More information on the construction of the test data is provided in Appendix D.1.

For both error detection and error correction, a **trial** is a single measurement indexed by a *trial_id* and identified by a *lane_id* and a *timestamp* indicating when the measurement was taken. An erroneous measurement is referred to as a *target trial*. If a trial is not a target trial (the provided flow measurement is the correct value), it is considered a *non-target trial.*

3.1.3 Trial file(s) (system input), cleaning: traffic, both tasks

The **system input** consists of trials in multiple tab-delimited files, each containing all the lane detector outputs for a single month, and each trial file will be named `cleaning_test_`*yy_mm*`.tsv`, where *yy* is the two-digit year and *mm* is the two-digit month during which the measurements were taken. The measurements will be sorted first by the measurement timestamp and then by the lane id. In each file, each line is a trial, and each trial contains the following tab-separated fields:

1. *trial_id*,
2. *lane_id*: the identifier of the lane detector,
3. *measurement_start*: the timestamp indicating when the measurement was taken.
4. *speed*: the provided mean speed of vehicles in miles per hour,
5. *flow*: the provided traffic flow value,
6. *occupancy*: the percentage of time a vehicle was above the sensor,
7. *quality*: a flag indicating the quality of the measurement.

In this data, each timestamp follows the ISO 8601 format of combined date and time as `YYYY-MM-DDThh:mm:ss.ssssss±hh:mm`,

3.1.4 Submission file(s) (system output), cleaning: traffic, error detection

For each trial file, `cleaning_test_`*yy*`_`*mm*`.tsv` the corresponding submission file for the error detection task is named `detection_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}.tsv`, where:

- *yy*: the two-digit year,
- *mm*: the two-digit month,
- *team*: a unique, consistent identifier for the team, which should also include the name of the participating organization or institution as well as a string that identifies the team within that organization, no underscore characters,
- *submission*: a number indicating the submission version.

Regarding **system output format**, each file will have the same number of lines in the same order as the corresponding test file, and consist of a two values per line, each separated by a tab character. These two values are:

1. *trial_id*,
2. *confidence_value*: a real-number indicating the belief that the flow value is erroneous, with a higher value indicating a greater belief that the value is erroneous[2].

System output results must be provided in multiple files, one submission file per trial file, using UTF-8 encoding.

3.1.5 Submission file(s) (system output), cleaning: traffic, error correction

For each trial file `cleaning_test_`*yy*`_`*mm*`.tsv`, the corresponding submission file for the error correction task is named `correction_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}_{`*metric*`}.tsv`, where:

- *yy*: the two-digit year,
- *mm*: the two-digit month,
- *team*: a unique, consistent identifier for the team, which should also include the name of the participating organization or institution as well as a string that identifies the team within that organization, no underscore characters,
- *submission*: a number indicating the submission version,
- *metric*: “cp” if the system is optimizing over metric $cost$, or “ca” if the system is optimizing over $cost_{alt}$.

Regarding **system output format**, each submission file will have the same number of lines in the same order as the corresponding test file, and consist of two values per line, each separated by a tab character. These two values are:

1. *trial_id*,
2. *cleaned_flow*: the corrected traffic flow for the corresponding measurement.

[2] In some cases one may wish to interpret the confidence value as a *log likelihood ratio*.

3.1.6 Performance metrics, cleaning: traffic, error detection task

For the error detection cleaning task, the primary performance metric will be the minimum of a decision cost function that weighs misses and false alarms. Any flow measurement that is erroneous that is classified as erroneous is a *true positive*. Any flow measurement that is erroneous that is classified as correct is a *miss*. Any flow measurement that is not erroneous but is classified as erroneous is a *false alarm*. Any flow measurement that is not erroneous and is classified as correct is a *true negative*.

Each trial's confidence value will be converted to a decision by comparing the confidence value to a certain threshold; a trial with a confidence value greater than or equal to the threshold will be interpreted as a decision of `true`, indicating that the system's belief is that the flow value is erroneous; each other confidence value will be converted to a decision of `false`, indicating the belief that flow values is not erroneous.

For any given threshold τ, a Decision Cost Function (DCF) representing a linear combination of the miss and false alarm rates at a given threshold τ can be computed. The decision cost function for this task will be:

$$DCF_{ed}(\tau) = c_{miss} * P_{target} * P_{miss\,|\,target}(\tau) + c_{fa} * (1 - P_{target}) * P_{fa\,|\,nontarget}(\tau) \tag{1}$$

where

$$P_{miss\,|\,target}(\tau) = \frac{|\text{misses}(\tau)|}{|\text{target trials}|}$$
$$P_{fa\,|\,nontarget}(\tau) = \frac{|\text{false alarms}(\tau)|}{|\text{non-target trials}|}$$

and $P_{target} = 0.0312$, $c_{miss} = 1$, and $c_{fa} = 1$. P_{target}, the target prior used to compute system performance, is not necessarily the same as the prior probability of target trials in the data.

To account for the different thresholds, the DCF value for all possible thresholds τ will be taken and the minimum score will be used. Hence the filtering cost for a system on this task is:

$$C_{det} = \min_{\tau} \left(DCF_{ed}(\tau)\right) \tag{2}$$

To improve the intuitive meaning of C_{det}, it will be normalized by dividing it by $C_{default}$, defined as the best cost that could be obtained without processing the input data (i.e., by either classifying every value as erroneous or classifying every value as non-erroneous). Since the better of these two systems will classify every value as non-erroneous, the default cost is:

$$C_{default} = c_{miss} * P_{target} \tag{3}$$

Hence, the performance metric, or the cost, is the normalized cost function, c_{norm}, which is:

$$\begin{aligned} cost = C_{norm} &= \frac{C_{det}}{C_{default}} \\ &= \frac{\min_{\tau}\left(c_{miss} * P_{target} * P_{miss\,|\,target}(\tau) + c_{fa} * (1 - P_{target}) * P_{fa\,|\,nontarget}(\tau)\right)}{c_{miss} * P_{target}} \end{aligned} \tag{4}$$

For the error detection task, it is possible that only a subset of the submitted values will be scored.

3.1.7 Performance metrics, cleaning: traffic, error correction task

For the error correction cleaning task, the primary metric is the Mean Absolute Error (MAE), where n is the number of trials, and for each trial i:

- $\widehat{fl_i}$ is the estimated traffic flow.
- fl_i is the correct traffic flow.

The performance metric, the mean absolute error in flow, is the cost function:

$$cost = MAE = \frac{\sum_{i=1}^{n} |\widehat{fl_i} - fl_i|}{n} \tag{5}$$

Furthermore, for additional analysis, an implicit detection decision will be extracted from the submissions.

A secondary metric will calculate an alternative cost ($cost_{alt}$) with a discounted penalty for a changes to a measurement, where x_i is the provided flow value for trial i.

$$cost_{alt} = \frac{\sum_{i=1}^{n} \left(1 - c_d * \min\left(1, \frac{|\widehat{fl_i} - x_i|}{c_{flmax}}\right)\right) |\widehat{fl_i} - fl_i|}{n} \tag{6}$$

where for this evaluation, the values of the two constants are $c_{flmax} = 20$ and $c_d = 0.4$. Additional details on this metric are provided in Appendix D.2.

By default, the scoring metric will be the cleaning metric *cost*. Submitting two versions of a cleaning system, one for each metric is encouraged. Although the main metric is the *cost*, both metrics will be computed on all submissions for analysis purposes.

For the error correction task, it is possible that only a subset of the submitted values will be scored.

4 Alignment

In this task, participants are asked to relate different instances of the same object. The task is applicable to several different problems in data science, e.g., a word with the corresponding visual object, an object appearing in different images or videos, or time stamps associated with two different time series.

4.1 Alignment in the Traffic Domain

In the traffic domain, the goal of the alignment task is to analyze video from camera feeds to detect an event and match it to a separate inventory of traffic events. This task may be divided into two steps:

- From video, detect the occurrence of one or more traffic events.
- Match the detected events to events in the event instances list.

Summary 3: Alignment task

This task involves matching traffic events with the traffic video segments containing those events.

Input. Video segments from traffic cameras and traffic event data around that camera.

Output. A confidence value for each (v = video segment, e = traffic event) pair, where a greater confidence value indicates a greater belief that v and e refer to the same event.

Trial File(s). `alignment_{camera_name}_test.tsv`.

Submission File(s). `alignment_subm_{camera_name}_{team}_{submission}.tsv`.

For this task, a *recorded* event is a traffic event that is present in the video and a *reported* event is a traffic event present in the traffic event inventory. Systems must detect events recorded in the video and match them to reported traffic events.

Table 2 in Appendix B.2 lists and summarizes the relevant event types that are to be detected in one or more of the pilot evaluation tasks. The types of traffic events that are to be detected in the alignment tasks are the event types:

- Accidents and Incidents,
- Obstructions,
- Device Status.

Note that some recorded events may not be reported, e.g., if the situation did not last long, did not disturb traffic, or resolved itself on its own. For example, a disabled vehicle may not be reported because the driver restarted the car after a few minutes. Systems should not link recorded events with events that are not reported in the traffic event inventory.

Each video segment has accurate time information, and the location of the source camera of each video segment is given in latitude and longitude. Some video feeds may have the direction the source camera is facing water-marked (“E” for East, for example). When matching video segments with desired events, *all of the timestamps from those recent events have been removed.* The remainder of the event information will be provided; the order of the events will be changed so that that implicit timestamp information is not provided.

The output of this task is a confidence value for each video segment and traffic event pair. All pairs must be evaluated, and must be evaluated independently of traffic events (see Rule 2), e.g. for any videos v_a and v_b and for any traffic events e_c and e_d, to compute a confidence for (v_a, e_c), a system may consider (v_b, e_c) but not (v_a, e_d).

A (video, event) pair is considered a match when the recorded event is visible during a part or the whole video segment. If the recorded event overlaps multiple video segments, every video segment that contains some part of the recorded event will be considered a match.

Rule 2: Alignment Task Rule: Independent Trials

All (v = video segment, e = traffic event) pairs must be **evaluated independently of the traffic events**, e.g. for any videos v_a and v_b and for any traffic events e_c and e_d, to compute a confidence for (v_a, e_c), a system may consider (v_b, e_c) but not (v_a, e_d).

4.1.1 Training data, alignment: traffic

All the common core data described in Section 2 will be available for training and development purposes.

An additional "Traffic Event" dataset will also be provided for training. This training set involves a subset of traffic cameras, denoted as training cameras, and for each of those cameras, the list of traffic events with timestamps will be provided. In more detail, this additional training data consists of:

- **Video.** 15-minute video segments from the training cameras, available from the core data, at `core_data/traffic_videos/<camera_name>`. See Appendix B.2 (Figure 3) for the organizational structure of the core data.
- **Reported events.** All traffic events reported as having taken place at a distance less than $d = 500$ meters from each training camera's location. These traffic events will include timestamps. These events will be supplied in the alignment subfolder of the training data. See Appendix B.2 (Figure 4, page 23) for the organizational structure of these additional training sets.

4.1.2 Test data, alignment: traffic

Participants will be tested on different traffic cameras, which are denoted as the test cameras. Test cameras may not be the same cameras present in the training data.

A (video segment, reported event) pair is referred to as a **trial**. When the video segment contains a recorded event that corresponds to the given reported event, this is referred to as a *target trial*. If a trial is not a target trial, it is considered a *non-target trial*.

The test data consists of video segments and traffic events, similar to the training data described in Section 4.1.1, but the events provided will not have timestamps. In order to consider all traffic events reported in the vicinity of a camera, the test data will consist of:

- **Video.** 15-minute video segments from the source test cameras. These test video segments will be supplied in the `test/alignment/videos/<camera_name>` folder.
- **Reported events.** All events reported as having taken place at a distance less than $d = 500$ meters from each test camera's location. *All original timestamp data will have been removed.*

See Appendix B.2 (Figure 5) for the organizational structure of the test data.

4.1.3 Trial file(s) (system input), alignment: traffic

The **system input** consists of one trial file per video camera, with trials listed one per line. This file named `alignment_`{*camera_name*}`_test.tsv` will have the following tab-separated fields:

1. *trial_id*,
2. *video_id*,
3. *event_id*.

4.1.4 Submission file(s) (system output), alignment: traffic

For every test camera, the corresponding submission file is named as follows:
`alignment_subm_`{*camera_name*}`_`{*team*}`_`{*submission*}`.tsv`, where:

- *camera_name*: the test camera name,

- *team*: a unique, consistent identifier for the team, which should also include the name of the participating organization or institution as well as a string that identifies the team within that organization, no underscore characters,
- *submission*: a number indicating the submission version.

Regarding **system output format**, each test camera's submission file will list results for all (video, event) pairs, and will index each pair by trial id. The results must be listed one per line as two tab-separated fields:

1. *trial_id*: from the test file,
2. *confidence value*: the computed confidence between the event and the video segment. Confidence scores are real numbers between 0 and 1, inclusive ($0 \leq \text{confidence value} \leq 1$).

All trials, must appear in the submission file. Participants are encouraged to submit the trials in the order of the trials in each test file, but at a minimum, participants must include the *trial_id* in each row. Files should be submitted without header rows.

System output results must be provided in a single file per test camera, using UTF-8 encoding.

4.1.5 Performance metrics, alignment: traffic

When a system outputs a non-match for a target trial, the resulting error is called a *miss*, and when a system outputs a match for a non-target trial, the resulting error is called a *false alarm*.

Each trial will be treated as a match or non-match by comparing the system output to a certain threshold; trials greater than or equal to the threshold will be considered matches, and all others will be considered non-matches. By using the sorted system outputs as thresholds, the system's misses and false-alarms can be calculated at all possible *a posteriori* thresholds. The performance metric will be based on a decision cost function ($DCF_{alignment}$) representing a linear combination of the miss and false alarm rates at a threshold τ. The decision cost function for this task will be:

$$DCF_{alignment}(\tau) = c_{miss} * P_{target} * \frac{|\text{misses}(\tau)|}{|\text{target trials}|} + c_{fa} * (1 - P_{target}) * \frac{|\text{false alarms}(\tau)|}{|\text{non-target trials}|} \tag{7}$$

For this task, the $c_{miss} = 1$, $c_{fa} = 100$, and $P_{target} = 0.5$. These values should be used for calibration only and should not be interpreted as insight in the composition of the dataset.

The overall performance metric will be the minimum DCF value obtained considering all τ. Hence the cost for a system on the alignment task is:

$$cost = \min_{\tau} \left(DCF_{alignment}(\tau) \right) \tag{8}$$

For this task, it is possible that only a subset of the submitted values will be scored.

5 Prediction task

In this task, participants are asked to estimate the value of a variable or multiple variables of interest at future times. The task is applicable to several different problems in data science, e.g., hypothesizing potential stock market events from sentiments expressed in social media or determining the likelihood of a team winning a game.

5.1 Prediction Task in the Traffic Domain

In the traffic domain, participants will develop a system that can predict the number and types of traffic events by type for a given (geographical bounding, interval of time) pair. This task will consider only a subset of the available event types, given in Table 2 (page 20). The set of traffic event types considered for this task is denoted $\mathcal{E}$ and consists of:

- Accidents and Incidents.
- Roadwork.
- Precipitation.
- Device Status.
- Obstruction.
- Traffic Conditions.

Summary 4: Prediction task

This task involves predicting the number and types of traffic incidents in a region over a time period.

Input. geographical bounding boxes and time intervals.

Output. predicted counts for each specified type of traffic event.

Trial File(s). `prediction_test.tsv`

Submission File(s). `prediction_subm_`{*team*}`_`{*submission*}`.tsv`.

The task output will be a list of counts of traffic events for each event type listed above.

5.1.1 Training data, prediction: traffic

All the common core data described in Section 2 will be available for training and development purposes.

5.1.2 Test data, prediction: traffic

The test data consists of a series of trials, where each **trial** is a (location, time interval) pair. For the prediction task, time intervals will be time intervals in the past that occur after the times in the training data.

5.1.3 Trial file(s) (system input), prediction: traffic

The **system input** consists of a series of trials in the file file `prediction_test.tsv`, one trial per line. Each trial is specified with the following tab-separated fields:

1. *trial_id*,
2. *nw_lat*: the decimal latitude coordinate of the North-West corner of the bounding box,
3. *nw_lon*: the decimal longitude coordinate of the North-West corner of the bounding box,
4. *se_lat*: the decimal latitude coordinate of the South-East corner of the bounding box,

5. *se_lon*: the decimal longitude coordinate of the South-East corner of the bounding box,
6. *start*: timestamp of the start of the time window,
7. *end*: timestamp of the end of the time window.

Each latitude and longitude coordinate is in decimal degrees, and each timestamp follows the ISO 8601 format of combined date and time as `YYYY-MM-DDThh:mm:ss.ssssss±hh:mm`. In the trial file:

- The area of each bounding box is at least 0.25 km^2 and at most 8 km^2.
- Time intervals are at least 3 hours long and at most 31 days long.

5.1.4 Submission file(s) (system output), prediction: traffic

System submissions must be provided in a single file using standard UTF-8 encoding and named `prediction_subm_{`*team*`}_{`*submission*`}.tsv`, where:

- *team*: a unique, consistent identifier for the team, which should also include the name of the participating organization or institution as well as a string that identifies the team within that organization, no underscore characters,
- *submission*: a number indicating the submission version.

Regarding **system output format**, each line of the submission file must contain the output of a trial by producing the following tab-separated fields:

1. *trial_id*: from the test data,
2. *n_accidents_incidents*: the predicted number of traffic events of the *Accidents and Incidents* type,
3. *n_roadwork*: the predicted number of traffic events of the *Roadwork* type,
4. *n_precipitation*: the predicted number of traffic events of the *Precipitation* type,
5. *n_device_status*: the predicted number of traffic events of the *Device Status* type,
6. *n_obstruction*: the predicted number of traffic events of the *Obstruction* type,
7. *n_traffic_conditions*: the predicted number of traffic events of the *Traffic Conditions* type.

Results for all trials must be submitted. Participants are encouraged to submit the trials in the order of the trials in each test file, but at a minimum, participants must include the *trial_id* in each row. Files should be submitted without header rows.

5.1.5 Performance metrics, prediction: traffic

Task performance will be scored with the average of each trial's Root Mean Squared Error (RMSE).

In this metric, $\mathcal{E}$ is the set of traffic event types and there are $|\mathcal{E}|$ types of events. The RMSE is computed for each trial i, where

- $\widehat{e_i}$ is the predicted count of events of type e,
- e_i is the true count of events of that type.

$$RMSE(i) = \sqrt{\frac{\sum_{e \in \mathcal{E}} (\widehat{e_i} - e_i)^2}{|\mathcal{E}|}} \tag{9}$$

An event is considered to be in the test window if there is any overlap between the test window and any of the event's *created*, *start* or *closed* times[3]; i.e. if any of *created*, *start*, or *closed* times are in the test window, or if the test window is enclosed in an interval made from any combination of the *created*, *start*, or *closed* times.

The RMSE for all the trials is then averaged to get a cost, where n is the number of trials:

$$cost = \frac{1}{n} \sum_{i=1}^{n} RMSE(i) \tag{10}$$

6 Forecasting task

In this task, participants are given time series measurements as training data and are asked to forecast future values of these time series. The time series measurements serve as training data and the future values to forecast are the test measurements. Forecasting in time series is a common task that has many applications, including in weather and economics. Forecasting may be viewed as a regression or classification problem where the desired values occur in the future.

6.1 Forecasting Task in the Traffic Domain

In the traffic domain, the goal of the forecasting task is to leverage past traffic information and current conditions (using the other datasets) to forecast traffic flow values from future values in the traffic lane detector data.

Summary 5: Forecasting task

This task involves determining future traffic flow values.

Input. a list of lane detectors and times to forecast traffic flow.

Output. for each lane detector and timestamp, a forecasted traffic flow value in vehicles per minute.

Trial File(s). `forecasting_test_`*yy*`_`*mm*`.tsv`.

Submission File(s). `forecasting_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}.tsv`.

Systems will be required to forecast the flow values for series of trials, where each trial is a (*lane_id*, *timestamp*) pair. For each trial, the system must output a single flow value that represents the number of vehicles that have passed within a predetermined number of seconds. For each lane detector, that predetermined number of seconds is specified in the interval field. For the forecasting task, the interval will always be 60 seconds, meaning that the traffic flow value is often the number of vehicles to have passed through the detector in the previous minute.

6.1.1 Training data, forecasting: traffic

All the common core data described in Section 2 will be available for training and development purposes. Of particular use will be the lane detector measurements that are also used for the cleaning task. However, some of the lane detectors in the test data will not have any training data available.

[3] See Appendix A.2 for a description of the events fields.

6.1.2 Test data, forecasting: traffic

In this task, the test data consists of the trials. A **trial** is a measurement indexed by a trial_id and specified by a lane_id and a measurement timestamp. For the forecasting task, the trial timestamps are in the past but are after the timestamps for measurements of those detectors in the training set.

6.1.3 Trial file(s) (system input), forecasting: traffic

The **system input** consists of trials in multiple tab-delimited files, each file containing all of the trial measurements for a single month, and each trial file is named `forecasting_test_`*yy*`_`*mm*`.tsv`, where *yy* is the two-digit year and *mm* is the two-digit month during which flow values are desired. In each file, each line is a trial, and each trial contains the following tab-separated fields:

1. *trial_id*,
2. *lane_id*: the identifier of the lane detector,
3. *measurement_start*: the timestamp indicating the time for which to forecast the flow measurement.

Following the format in the cleaning task, each timestamp follows the ISO 8601 format of combined date and time as `YYYY-MM-DDThh:mm:ss.ssssss±hh:mm`. In the forecasting task, some of the lane detectors in the test trials do not have any training data.

See Appendix B.2 (Figure 5) for the organizational structure of the test data and where to find it on Amazon S3.

6.1.4 Submission file(s) (system output), forecasting: traffic

For every trial file `forecasting_test_`*yy*`_`*mm*`.tsv`, the corresponding submission file is named `forecasting_subm_`*yy*`_`*mm*`_{`*team*`}_{`*submission*`}.tsv` where:

- *yy*: the two-digit year,
- *mm*: the two-digit month,
- *team*: a unique, consistent identifier for the team, which should also include the name of the participating organization or institution as well as a string that identifies the team within that organization, no underscore characters,
- *submission*: a number indicating the submission version.

Regarding **system output format**, each submission file will have the same number of lines in the same order as the corresponding test file, and consist of a two values per line, each separated by a tab character. These two values are:

1. *trial_id*,
2. *forecasted_flow*: the forecasted traffic flow for the corresponding measurement in vehicles per minute.

Flow values for all trials must be provided. Participants are encouraged to submit the trials in the order of the trials in each test file, but at a minimum, participants must include the *trial_id* in each row. Files should be submitted without header rows.

System output results must be provided in multiple files, one submission file per trial file, using UTF-8 encoding.

6.1.5 Performance metrics, forecasting: traffic

Task performance will be scored with the mean of each trial's absolute error in flow. This is the same metric as used for the cleaning task. In this metric, n is the total number of measurements in the test set, which is also the number of trials, and for each trial i:

- $\widehat{fl_i}$ is the estimated traffic flow,
- fl_i is the correct traffic flow.

The performance metric, the Mean Absolute Error (MAE) in flow, is the cost:

$$cost = MAE = \frac{\sum_{i=1}^{n} |\widehat{fl_i} - fl_i|}{n} \tag{11}$$

For this task, it is possible that only a subset of the submitted values will be scored.

7 System Descriptions

For each system output submission to a DSE Pilot evaluation task, a description of each system is requested. Each system description should provide:

- **System Name.** The name of the system.
- **Task.** The task(s) the system(s) was used for: cleaning, alignment, prediction, or forecasting.
- **Team Affiliation.** The names of the submitting team.
- **System Summary.** A brief summary the system in a few sentences (at most a paragraph).
- **Algorithmic Information.** A description of additional details about the methods and algorithms used (1-2 paragraphs).
- **Data Sources.** A brief description of the data sources used. This data includes the training data sources as well as additional data fed to the algorithms. Mention which of the provided data sets were used as well as whether there were any external data sets used. Cite any external sources.
- **Development Data.** (optional) a brief description of any "development sets" that were generated and any brief results.

8 Schedule

The key dates for the NIST Pilot evaluation are given below in Table 1 (page 17). In particular, there will be two phases to the pilot, each corresponding to a different *workflow* (one with errorful lane detector measurements and one with true lane detector measurements), as indicated in Figure 2. In the figure, the *erroneous lane detector measurements* are the measurements that contain errors and are the measurements that participants are being asked to process in the cleaning task. The *cleaned lane detector measurements* are the lane detector measurements that have been corrected by a system solving the cleaning task and are the system output of the cleaning task. The *true lane detector measurements* are the ground-truth lane detector measurements, which are the answer to the cleaning task.

In order to compare the performance of systems using the erroneous detector data to those systems using the ground-truth traffic data, the submission process will be scheduled into two phases:

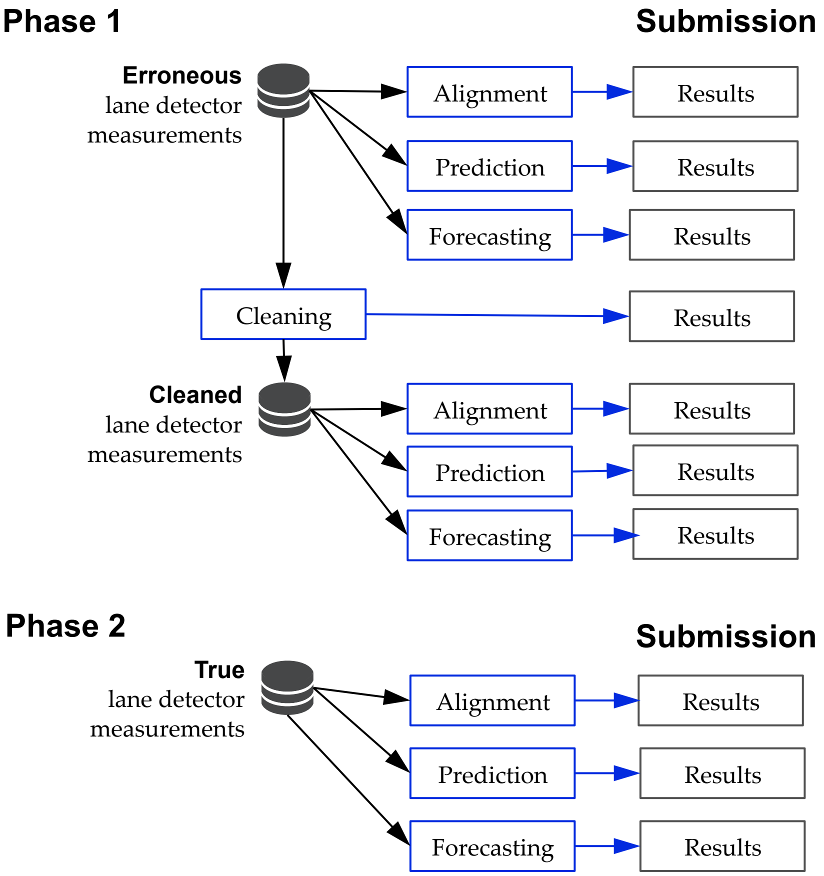


Figure 2: The pilot organizes the tasks into two phases, one in which lane detector measurements are "erroneous," and one in which "true" lane detector measurements are provided. This pipeline approach will allow for contrastive runs where the true output and the system output are used in place of the original traffic detector data.

- **Phase 1.** Participants are given the common core data along with the erroneous lane detector data and will be asked to submit system outputs from the four tasks using this data as input. Additionally, participants will be encouraged to submit the results of the alignment, prediction, and forecasting tasks using the traffic detector data output from the cleaning task (the cleaned data).
- **Phase 2.** Participants are given the error correction cleaning task truth data and will be asked to run the same systems for the alignment, incident, and flow tasks, using the error correction cleaning task truth data as input.

This schedule has two submission dates:

- **The first submission.** (where the bulk of the time is given), Participants complete tasks using the erroneous lane detector data. Note that participants will be submitting results using both the erroneous data as well as (optionally) the output data from the cleaning task. The traffic lane detector data output by the system for the cleaning task is referred to as the cleaned data.
- **The second submission.** (after the first submission deadline has passed), Participants use the lane detector ground truth data to complete the alignment, prediction, and forecasting tasks.

Table 1: IAD DSE Pilot Evaluation Schedule.

Registration opens	July 5, 2016
Register to the evaluation by	July 31, 2016
Release of training data and cleaning test data	August 1, 2016
Release of the rest of the test data	October 28, 2016
First submission deadline (all tasks)	November 28, 2016
Release of ground-truth traffic detector data*	November 29, 2016
Second submission deadline	December 6, 2016
Release of initial results	January 12, 2017
Workshop (location TBD)	March 2017 (Tentative, to be confirmed)

(*) The ground-truth traffic detector data may not be released. Furthermore, that ground truth may only be released to participants that provide sufficient submissions for the tasks in this evaluation. In particular, it is necessary that participants submit a task other than the cleaning that utilizes the traffic lane detector data and provides submissions that use the erroneous traffic detector data and the cleaned traffic detector data with it. If the ground truth for the cleaning task is not released, then the second submission will not be required.

9 Rules

The evaluation is subject to the following rules and restrictions:

- While participants **will** be allowed to use outside data, these data **must be publicly available**, and participants must include references to the data sources used (and how to obtain them) when submitting evaluation results. No internal or proprietary data are allowed to be used. The only **exception** to this rule are both cleaning tasks (error detection and error correction), for which the available data is restricted, see task-specific rules below.

- Participants may not interact with the test data in any way, e.g., reading the test files or watching the test videos is prohibited. The only **exception** to this rule are both cleaning tasks (error detection and error correction), where participants can interact with the traffic lane detector measurements.

- Each participating site must send one or more representatives who have working knowledge of the evaluation system to the evaluation workshop. Representatives must give a presentation on their system(s) and participate in discussions of the current state of the technology and future plans. Workshop registration information will be distributed to registered evaluation participants when available.

- For the **alignment task**, all (v = video segment, e = traffic event) pairs must be **evaluated independently of the traffic events**, e.g., for any videos v_a and v_b and for any traffic events e_c and e_d, to compute a confidence for (v_a, e_c), a system may consider (v_b, e_c) but not (v_a, e_d).

- Dissemination:

  - NIST may generate and disseminate table, statistics, and charts of all system results for conditions of interest, but these will not contain the site names of the systems involved.

Dissemination may include but is not limited to posting on NIST's website. Participants may publish or otherwise disseminate these results, unaltered and with appropriate reference to their source.

- Participants may not publish or otherwise disseminate comparisons of their performance results with those of other participants without the explicit written permission of each such participant. Furthermore, publicly claiming to "win" or suggest a ranking the evaluation is strictly prohibited. Any misrepresentation of the evaluation or its results is also strictly prohibited.
- The results reported by NIST are not to be construed, or represented, as endorsement of any participant's system, or as official findings on the part of NIST or the U.S. Government.

Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this evaluation plan in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

References

[1] 2010 u.s. census, 2015. URL `http://www.census.gov`.

[2] Maryland chart traffic cameras, 2015. URL `http://www.chart.state.md.us/travinfo/trafficcams.php`.

[3] NOAA's integrated surface hourly, 2015. URL `http://www.ncdc.noaa.gov/isd`.

[4] NOAA, 2015. URL `http://www.ncdc.noaa.gov/swdi`.

[5] J. Neal Lott. The quality control of the integrated surface hourly database. In *84th American Meteorological Society Annual Meeting*, volume 7.8, Seattle, WA, 2004. American Meterological Society. URL `http://www1.ncdc.noaa.gov/pub/data/inventories/ish-qc.pdf`.

[6] Photograph, National Archives Identifier 546711 (Artist Yoichi R. (Robert) Okamoto). Looking south from beltway bridge over the potomac. the capital beltway circles the virginia-maryland suburbs and provides high speed access to points in the district, 5/1973, May 1973. DOCUMERICA: The Environmental Protection Agency's Program to Photographically Document Subjects of Environmental Concern, 1972–1977 Record Group 412: Records of the Environmental Protection Agency, 1944–2006.

A Pilot Data Description

We describe the different data sets available for the DSE Pilot Evaluation in more detail, supplementing the overview provided in Section 2. The documentation associated with each data source provides a detailed description of all of its corresponding fields.

A.1 Lane detector measurements

Traffic detector data provided courtesy of the CATT Lab (The Center for Advanced Transportation Technology Laboratory). For the lane detector measurements, there are two data components.

- **dse_p2016_lane_detector_inventory: Lane Detector Inventory.** List of all traffic lane detectors. Each detector is uniquely identified by its `lane_id` value, and each detector inventory gives the location of the detector (in decimal latitude and longitude coordinates), the source organization for the measurements of those detectors, the time interval between scheduled measurements, and other relevant information.
- **dse_p2016_lane_detector_measurements: Lane Detector Measurements.** Measurements from traffic sensors in locations in the DC Metro area and the Baltimore area. Traffic sensors are placed on both directions of the highways, in each lane. Lane and zone (multiple lanes of the same road going in the same direction) data are provided. The measurements include the following attributes (among others):
  1. *Flow:* the number of vehicles to have passed through the lane detector since the last scheduled measurement.
  2. *Speed:* the average vehicle speed since the last measurement.
  3. *Occupancy:* the average percent of time a vehicle was in front of the detector since the last measurement.
  4. *Quality:* a data quality field.

In the lane detector measurements, a traffic flow value refers to the number of vehicles that have passed within a predetermined number of seconds. For each lane detector, that predetermined number of seconds is specified in the interval field and is often 60 seconds, i.e., the traffic flow value is often the number of vehicles to have passed through the detector in the previous minute.

A.2 Traffic events

Traffic events provided courtesy of the CATT Lab (The Center for Advanced Transportation Technology Laboratory). For the Traffic events, there is one data component.

- **dse_p2016_event_instance_measurements: Traffic Events Instances.** A traffic event is defined as a situation that involves traffic in the ways described in Table 2. Prominent features of an event are injuries, damage to vehicles, hazards to persons, failure of equipment, closure of one or more lanes, debris or roadkill on the road or shoulder, or any obstruction on roadway. Each traffic event listing includes the following fields (among others):
  1. Description.
  2. Location, both in formatted text (the intersection) and in decimal latitude and longitude.
  3. Times the event was started, created, confirmed, and closed.
  4. The type and subtype of the traffic event.

  A traffic event is reported in the inventory when it is outlined by or to authorities (911 calls, road kill pickups, etc.). It is confirmed when an authority arrives at the scene (the police arrives, etc.) and deemed over when it was closed by authorities. For an accident, this typically indicates when all lanes have been reopened, damaged vehicles have been removed, and all responders have left the scene. This dataset has data from 2003 to 2015, collected from the DC-Maryland-Virginia area, and is approximately 200 MB.

Table 2: Traffic event types and subtypes for event types used in the pilot evaluation tasks. For each event type, rather than listing all possible event subtypes, the different kinds of event subtypes are summarized.

Traffic Event Type	Traffic Event Subtype
Accidents And Incidents	Abandoned vehicle, accident, accident involving a semi trailer, accident involving a truck, disabled vehicle, hazardous material spill, incident, injury accident, minor accident, multi vehicle accident, numerous accidents, serious accident, vehicle on fire
Device Status	Sign down, traffic lights not working, power failure
Obstruction	Animal struck, debris on roadway, downed cables, drawbridge open, fallen trees, obstruction on roadway, subsidence, animals on roadway
Precipitation	Snow, rain, hail, frost, precipitation cleared
Roadwork	Bridge construction, bridge maintenance operations, construction work, emergency maintenance, overgrown grass, overgrown trees, paving operations, road construction, road maintenance operations, road marking operations, road widening, storm drain, water main work, work in the median, work on underground services
Traffic Conditions	Traffic congestion, stopped traffic, traffic congestion cleared

Table 2 contains a summary of the event subtypes for the event types used in the pilot evaluation tasks.

A.3 Traffic camera video

For the traffic camera video data, there are two data components. Video provided courtesy of the Maryland Department of Transportation.

- **dse_p2016_traffic_camera_inventory: Camera Inventory [2].** A list of all traffic cameras with their locations, described both in text (the intersection) and in decimal latitude and longitude.
- **dse_p2016_traffic_camera_videos: Traffic Camera Video Feeds [2].** Consecutive video segments from traffic cameras in Maryland with start times. Most segments are 15 minutes long. The traffic cameras may be remotely operated by humans, who can rotate the camera and zoom, which happens when the human operator chooses to look at a traffic situation. Some cameras may have watermarks indicating the direction the camera is facing (E for East, SW for South-West, etc), or the current time.

A.4 U.S. Census

For the U.S. Census data, there are two data components.

- **dse_p2016_2010_us_census: 2010 U.S. Census [1].** Publicly available information including population counts; age, income, and occupation demographics; and household demographics in summary files and PUMS (Public Use Microdata Sample).
- **dse_p2016_american_community_survey: American Community Survey (ACS) [1].** A more frequent survey providing statistics on transportation and commutes, such as the average com-

mute length, the percentage of people who carpool, and the percentage of people who use public transportation. There are 1-year, 3-year, and 5-year surveys as summary Files and PUMS, like the U.S. Census Data.

A.5 OpenStreetMap

For the OpenStreetMaps, there is one component.

- **dse_p2016_openstreetmap_maps: OpenStreetMap Maps.** Publicly-available map data from from OpenStreetMap, describing the road network in the DC-MD-VA area as well as locations including airports, public transportation stations, and buildings that host large events. These maps also support lookup by latitude and longitude coordinates.

A.6 NOAA weather

For the NOAA Weather data, there are two components.

- **dse_p2016_integrated_surface_data: NOAA Integrated Surface Data (ISD) [3].** A dataset of measurements from weather stations in the DC-MD-VA area with a variable number of measurements. Measurements include station information, temperature, air pressure, weather condition, precipitation, and other elements. The ISD set is quality-controlled. The quality control does not state that it is free of errors or missing data; only that others have looked at it to try to improve the quality of the data. Lott [5] discusses the quality control process that are used in the ISD to check for formatting errors and outliers.
- **dse_p2016_severe_weather_data_inventory: NOAA Severe Weather Data Inventory (SWDI) [4].** A compilation of many types of severe weather, including storms, hail, tornados and lighting strikes.

B Pilot Data organization structure

This section presents the organizational structure of the data that will be provided to participants and includes an indexing of each of the data sources and the organization structure of the Amazon S3 bucket.

B.1 Data indexing catalog

To aid in the organization, each dataset is labelled with an index id (letters and numbers), and a brief name. The indices of the data is provided in Table 3.

B.2 Data organization structure on AWS

Figure 3 gives the overall structure of the data in the common core.

Figure 4 shows the organizational structure of the additional training data provided for the alignment task described in Section 4.1.1.

Figure 5 shows the organizational structure of the test data for the forecasting task described in Section 6.1.2.

Table 3: Data indices of all pilot evaluation datasets, providing a catalog of the pilot evaluation. The indices are labeled with abbreviations to aid in searching for the data. Each dataset marked with a * will be made available through its respective organization's url rather than being provided by NIST; nevertheless, these sets are included in the catalog.

Index ID (Data set name)	Brief Description
dse_p2016_lane_detector_inventory	Metadata on traffic lane detectors by lane id. See Appendix A.1 for a more detailed description.
dse_p2016_lane_detector_measurements	Timestamped flow and speed measurements output by the lane detectors. See Appendix A.1 for a more detailed description.
dse_p2016_event_instance_measurements	Timestamped located traffic events of various types. See Appendix A.2 for a more detailed description.
dse_p2016_traffic_camera_inventory	Text file liking camera ids, camera names, and locations. See Appendix A.3 for a more detailed description.
dse_p2016_traffic_camera_videos	Segments of recorded traffic camera video whose files are named with the video date and start times. See Appendix A.3 for a more detailed description.
dse_p2016_2010_us_census*	The 2010 U.S. Census data. See Appendix A.4 for a more detailed description.
dse_p2016_american_community_survey*	The American Community Survey (ACS) is data collected in 2011, 2013, and 2015 that supplements the 2010 U.S. Census. See Appendix A.4 for a more detailed description.
dse_p2016_openstreetmap_maps*	Publicly-available maps and relevant map data. See Appendix A.5 for a more detailed description.
dse_p2016_integrated_surface_data*	NOAA Integrated Surface (ISD) is weather data on various weather stations including temperature. See Appendix A.6 for a more detailed description.
dse_p2016_severe_weather_data_inventory*	NOAA Severe Weather Data Inventory (SWDI) is a compilation of storms and other severe weather events. See Appendix A.6 for a more detailed description.

C Traffic zones

Lane detectors at the same location are aggregated into **traffic zones**. Figure 6 illustrates this notion of a traffic zone.

D Cleaning Tasks: Additional Information

This section provides additional details on the test data and the metrics specific to the cleaning task and anomaly detection tasks.

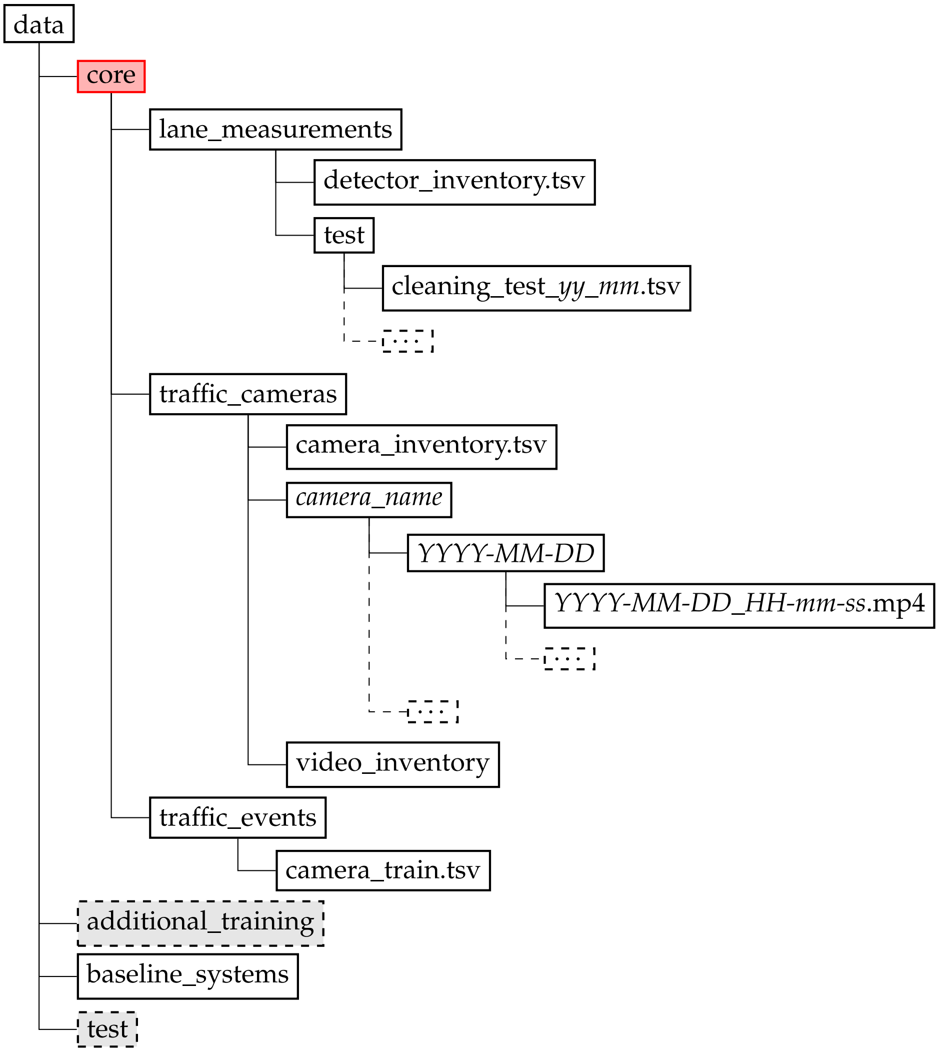


Figure 3: Organizational hierarchy for the pilot core data on Amazon S3. See other file hierarchy figures in the Appendix for the dotted border greyed out sub-hierarchies.

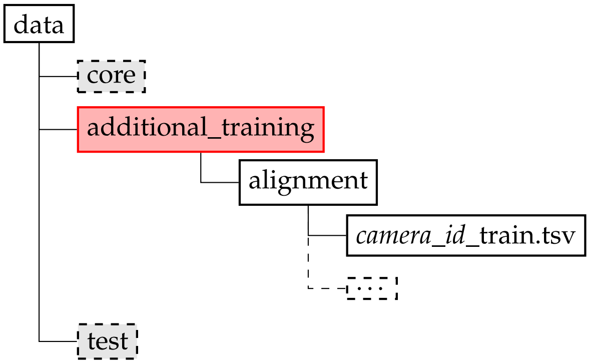


Figure 4: File hierarchy for the additional training data provided for the alignment task. See other file hierarchy figures for the dotted border greyed out sub-hierarchies.

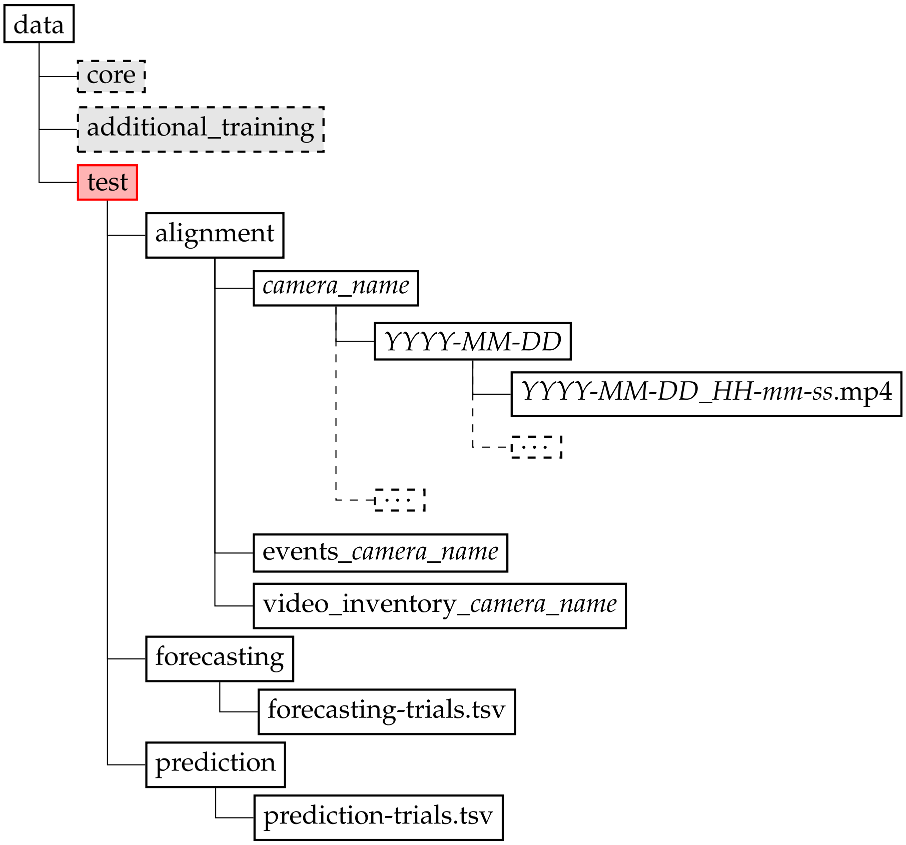


Figure 5: File hierarchy for the evaluation test data. Note that the cleaning task test data is in the "core" section rather than in the "test" section.

D.1 Test trial construction for both cleaning tasks

For this task, the test data consists of all measurements of a selected subset of lane detectors. The detectors included in the test set will be arbitrarily selected by their *zone_id* (all detectors in a chosen zone will be included), and each detector will be specified by its *lane_id*.

Errors will have been introduced into the test set measurements through modification of two of the attributes: flow and `speed`. How these data are changed is a function of the lane detector and the measurement time.

First, how errors are introduced into the traffic flow values is discussed.

- **Flow errors, August 2015 and earlier.** For each measurement from August 2015 or earlier (all lane detectors), with probability $p_n = 0.03$ uniformly at random, each value will have its flow altered.

- **Flow errors, September 2015 and after.** For each measurement from September 2015 or later, depending on the time and the lane detector id, for time t and lane detector l, with probability $0 < p_{tl} < 1$ uniformly at random, each value will have its flow altered.

Next, how errors are introduced into traffic speed values are discussed.

- **Speed errors, August 2015 and earlier.** For each measurement from August 2015 or earlier (all lane detectors), with probability $p_n = 0.03$ uniformly at random, each value will have its speed altered. The altering of the speed values is independent of the altering of the flow values.

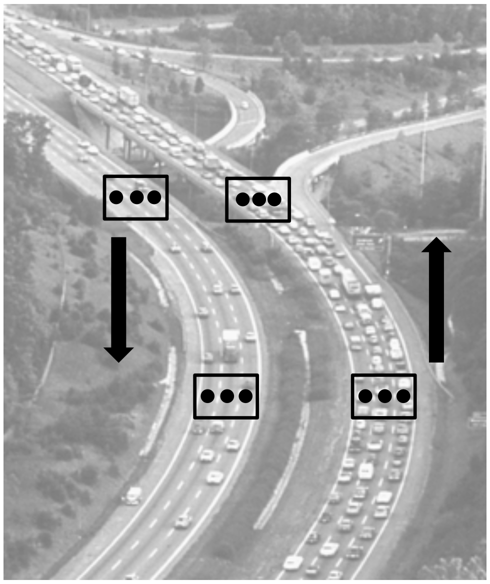


Figure 6: Traffic zones are an aggregation of traffic lanes. The dots are traffic lane detectors and the boxes are the traffic zones. In this figure, there are 12 traffic lane detectors: 4 traffic zones with 3 detectors in each zone. There is one traffic zone detector per road segment in each direction. In the data, there are different numbers of lane detectors in each traffic zone. This photo was obtained from National Archives [6].

- **Speed errors, September 2015 and after.** For each measurement from September 2015 or later, depending on the time and the lane detector id, for time t and lane detector l,with probability $0 < p_{tl} < 1$ uniformly at random, each value will have its speed altered. The altering of the speed values is independent of the altering of the flow values.

D.2 Alternate cleaning error correction task metric explanation

The *cost* is a generalization of the *percentage error* metric for detection tasks[4]. In order to account and adjust for this property of $cost$, a secondary metric will calculate an alternative cost ($cost_{alt}$) with a discounted penalty when measurements are changed from the original provided values. This metric is the alternative metric $cost_{alt}$.

Recall the alternative cost metric for the cleaning task. For this metric, x_i is the value of the provided flow measurement for trial i. This alternative metric, the alternative cleaning cost ($cost_{alt}$), can be defined by the following cost function:

$$cost_{alt} = \frac{\sum_{i=1}^{n} \left(1 - c_d * \min\left(1, \frac{|\widehat{fl_i} - x_i|}{c_{flmax}}\right)\right) |\widehat{fl_i} - fl_i|}{n} \tag{12}$$

When $c_d = 0$ (and $c_{flmax} \neq 0$), $cost_{alt} = cost$. This means that $cost_{alt}$ is one way to generalize the original metric $cost$.

The new metric $cost_{alt}$ is the original cost with a discounted penalty for changes to the measurements. Meaning, that as the system corrects more aggressively, the total error in flow is discounted more, until the system changes the value by more than c_{flmax} in either direction. At that point the

[4] Like its detection counterpart, this metric is sensitive to the proportion of positives (dirty data measurements for this task), and given the (small) amount of positives, this metric may favor more conservative systems.

maximum discount will be applied. The weight of the maximum discount, specified as a percentage of the flow error to discount is specified by the constant c_d. Although the total flow error is discounted, changing the flow value to err more may increase the cost because the error in flow increases the cost more than is decreased by the discount for changing the value.

There are two constants that parameterize this alternative cost function. The first, c_{flmax} specifies the maximum amount of change that results in a decreased weight to the absolute error. For this evaluation, $c_{flmax} = 20$, meaning that changing the flow value by more than c_{flmax} does not decrease the weight to the error. The second constant c_d, specifies the weight (percentage) at which the error is reduced. For this evaluation, $c_d = 0.4$.

In this metric, rather than just taking the arithmetic mean of the absolute errors in flow, each absolute error in flow is multiplied by a weight (or a percentage) that takes into account how much the estimated flow $\widehat{fl_i}$ differs from the provided flow x_i. The greater the absolute difference $|\widehat{fl_i} - x_i|$, the lower the absolute error in flow is weighted.

To be more specific on how the metric adjusts for the changing of flow values,

$$\min\left(1, \frac{|\widehat{fl_i} - x_i|}{c_{flmax}}\right) \tag{13}$$

is the cumulative uniform distribution function from 0 to c_{flmax} (requiring $c_{flmax} > 0$) where the input to the cumulative distribution function is the absolute value of the amount the flow value was changed.

To understand the influence of this metric, some examples are provided.

For example, first suppose that $\widehat{fl_i} = 10$, $fl_i = 20$, $c_{flmax} = 20$ and $c_d = 0.4$.

1. If $x_i = 10$, $cost_{alt} = (1 - 0.4 * 0) * 10 = 10$.
2. If $x_i = 15$, $cost_{alt} = (1 - 0.4 * 0.25) * 10 = 9$.
3. If $x_i = 30$, $cost_{alt} = (1 - 0.4 * 1) * 10 = 6$.

As another example, suppose that $x_i = 10$, $fl_i = 20$, $c_{flmax} = 20$ and $c_d = 0.4$.

1. If $\widehat{fl_i} = 10$, $cost_{alt} = (1 - 0.4 * 0) * 10 = 10$.
2. If $\widehat{fl_i} = 15$, $cost_{alt} = (1 - 0.4 * 0.25) * 5 = 4.5$.
3. If $\widehat{fl_i} = 25$, $cost_{alt} = (1 - 0.4 * 0.75) * 5 = 3.5$.

In summary, the new metric $cost_{alt}$ is the cost with a discounted penalty for corrections to the measurements. Meaning, that as the system corrects more aggressively, the total error in flow is discounted more, up to a change c_{flmax} vehicles per interval. Although the total flow error is discounted, changing the flow value to err more may increase the cost because the error in flow increases the cost more than is decreased by the discount for correcting.