U.S. Department of Transportation

Creating a Data Science Competition for Intersection Safety Systems

Insights from the U.S. DOT Intersection Safety Challenge Stage 1B System Assessment and Virtual Testing

www.its.dot.gov/index.htm

Final Report – February 2026

FHWA-JPO-26-005

Notice

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-JPO-26-005		

4. Title and Subtitle	5. Report Date
Creating a Data Science Competition for Intersection Safety Systems Insights from the U.S. DOT Intersection Safety Challenge Stage 1B System Assessment and Virtual Testing	February 2026
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Mohammad Goli, Haley Townsend, Stephen Scarano, Anand Seshadri, Caden Young, Claire Silverstein, Peiwei Wang, Karl Wunderlich	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Noblis Inc. 500 L'Enfant Plaza, S.W., Suite 900 Washington, D.C. 20024	
	11. Contract or Grant No. 693JJ321D000021

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Intelligent Transportation Systems (ITS) Joint Program Office (JPO) 1200 New Jersey Avenue, S.E., Washington, DC 20590	FINAL
	14. Sponsoring Agency Code

15. Supplementary Notes

Work Performed for: Norah Ocel (ITS JPO; Task Order Manager)

16. Abstract

To assess the technological maturity of data fusion and artificial intelligence (AI) capabilities of intersection safety systems (ISS), the U.S. DOT designed the Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing as a data science competition. For this data science competition, the U.S. DOT collected and provided real-world sensor data from a controlled test intersection at the Federal Highway Administration's (FHWA's) Turner-Fairbank Highway Research Center (TFHRC) in McLean, VA. The data was collected over several months of operation in 2023 and 2024 from multiple roadway sensors of different types (e.g., camera, Light Detection and Ranging [LiDAR], radar, thermal camera) and covered multiple scenarios, including non-conflicts, potential conflicts, and actual collisions between vehicles and pedestrians. Please note that surrogate pedestrians were used when necessary, and no one was harmed in the data collection process. The purpose of this document is to describe the impetus, assumptions, practical considerations, evolution, and lessons learned for designing and facilitating a data science competition for AI-based ISS, grounded in experiences from the U.S. DOT Intersection Safety Challenge Stage 1B.

17. Keywords	18. Distribution Statement
Safety, Artificial Intelligence, Machine Learning, Sensors, Data Fusion, Calibration, Intersection, Prediction, Evaluation, Pedestrians	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	25	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

List of Figures

Executive Summary

Purpose of Document

The purpose of this document is to describe the impetus, assumptions, practical considerations, evolution, and lessons learned for designing and facilitating a data science competition for artificial intelligence (AI)-based intersection safety systems (ISS), grounded in experiences from the U.S. Department of Transportation (DOT) Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing.

Analysis

Participating teams' algorithms were assessed on their ability to quickly and accurately perform three key technical elements of ISS operations: (1) the detection, classification, and localization of a number of pedestrians and vehicles in the intersection; (2) path prediction for those identified pedestrians and vehicles; and (3) conflict prediction among the identified pedestrians and vehicles. A hybrid evaluation approach—combining ground truth-based testing and qualitative expert review—was used for evaluating the submissions.

To form the foundation for assessment across the three key technical elements as part of the data science competition, multi-sensor, multi-condition data was collected at a single signalized four-way intersection at the Federal Highway Administration's (FHWA's) Turner-Fairbank Highway Research Center (TFHRC) in McLean, VA. Data was collected from several roadside sensors and traffic control devices, including from eight visual cameras, five thermal cameras, two Light Detection and Ranging (LiDAR) sensors, and four radar sensors. Traffic Signal Phase and Timing (SPaT) and vehicle and/or pedestrian calls to the traffic signal controller (if any) were also collected. Participants were provided training data (unlabeled sensor data) and validation data (labeled sensor data) to establish a baseline with their systems. The test dataset provided to participants consisted of sensor data and reflected only potential pre-conflict conditions at the intersection. Participants were expected to submit their results in a pre-specified format. Submissions were evaluated against well-established metrics, with precedents in competitions resembling Stage 1B.

Summary of Findings

The U.S. DOT Intersection Safety Challenge Stage 1B data science competition structure effectively balanced fairness, technical rigor, and scenario complexity, allowing meaningful differentiation of the assessment of each team's capabilities. The results offer a valuable snapshot of the current state of the field, helping to identify both promising approaches and areas where additional research and development are most needed. Stage 1B created a structured competition to rigorously evaluate the algorithmic capabilities of ISS solutions, focused on perception and prediction. It enabled comparative insights into team performance and assessment reproducibility. The constraints of a pre-collected dataset limited evaluation of complete end-to-end systems in their operational settings. Therefore, future ISS efforts should consider real-time, dynamic testing to better gauge real-world readiness.

Chapter 1. Introduction

The U.S. Department of Transportation (DOT) launched the U.S. DOT Intersection Safety Challenge ("the Challenge") in early 2023. The Challenge was a joint effort led by the Office of the Assistant Secretary for Research and Technology (OST-R) / the Advanced Research Projects Agency – Infrastructure (ARPA-I) and the Intelligent Transportation Systems (ITS) Joint Program Office (JPO), with contributions from modal partners. The goal of the Challenge was to transform roadway intersection safety through the innovative application of emerging technologies that identify and mitigate unsafe conditions involving vehicles and other road users (e.g., pedestrians, bicyclists, wheelchair users) at intersections. The Challenge kicked off with Stage 1A: Concept Assessment in which the U.S. DOT received 120 innovative concept papers from external entities describing proposed intersection safety systems (ISS). From these submissions and after an extensive judging process, 15 were selected for Challenge prize awards. The U.S. DOT announced the winners of the Stage 1A prize awards at the 2024 Transportation Research Board (TRB) Annual Meeting in Washington, D.C. These 15 winning teams were then invited to participate in the follow-on Challenge Stage 1B: System Assessment and Virtual Testing Primary Track Competition [1-3].

To assess each participating teams' data fusion and artificial intelligence (AI) capabilities of their individual intersection safety systems (ISS), the U.S. DOT designed the Stage 1B Primary Track as a data science competition. This competition was the first of its kind undertaken by the ITS JPO. For the Stage 1B competition, the U.S. DOT collected and provided real-world sensor data from a controlled test intersection at the Federal Highway Administration's (FHWA's) Turner-Fairbank Highway Research Center (TFHRC) in McLean, VA. The data was collected over several months of operation in 2023 and 2024 from multiple roadway sensors of different types (e.g., camera, Light Detection and Ranging [LiDAR], radar, thermal camera) and covered multiple real-world scenarios, including non-conflicts, potential conflicts, and actual collisions between vehicles and pedestrians. Please note that mechanical dummies (also referred to as surrogate road users) were used when necessary, and no person was harmed or put at risk during the entire data collection process. Stage 1B kicked off in April 2024, and after evaluation and judging of the thirteen submissions received in October 2024, a further set of cash prize awards were announced for the top ten teams in January 2025 during the 2025 TRB Annual Meeting [3].

What are Intersection Safety Systems (ISS)?

The U.S. DOT saw an opportunity to adapt already-available automated driving systems (ADS) and automated vehicle (AV) technologies, including sensing, perception, real-time path planning and decision-making, to the roadway intersection infrastructure setting. For the purposes of this Challenge, an ISS is a system that identifies, predicts, and mitigates unsafe conditions involving vehicles and other road users (e.g., pedestrians and cyclists) at an intersection in real-time. An ISS is anticipated to deploy emerging, low-cost sensors at intersections to improve sensing, use multi-sensor data fusion and AI to improve situational awareness and anticipate safety threats, and issue warnings and/or modify intersection signalized control settings to improve road user safety while minimizing disruptions to optimal traffic flow.

Evaluation Approaches for ISS

Evaluating the performance of machine learning systems in intelligent transportation applications—such as object detection, classification, localization, trajectory prediction, and conflict prediction—requires carefully designed methodologies that balance rigor, scalability, and real-world relevance. Several complementary evaluation strategies have emerged in the field, each offering distinct advantages and limitations.

Ground Truth Evaluation

One common approach is ground truth-based centralized evaluation, in which participants submit predictions on a reserved test dataset annotated in advance, typically by a central authority such as a transportation agency. These predictions are evaluated using standardized metrics, such as mean Average Precision (mAP) for detection, classification, and localization; Average Displacement Error (ADE) for trajectory prediction; and the F2 Score for conflict prediction assessment. This ground truth evaluation method facilitates objective comparison across systems by ensuring uniform evaluation conditions. It is particularly well-suited for leaderboard generation and large-scale competitions. Its limitations include a potential lack of generalizability to real-world conditions, risk of model overfitting to the test set distribution, and the resource burden associated with creating high-quality annotations and evaluation scripts [4].

Qualitative Expert Review/Evaluation

An alternative approach is the qualitative expert review model. In this approach, teams are responsible for generating their own annotations (either on provided data subsets or on newly collected data), executing their own models, and submitting comprehensive reports describing their methodology, evaluation process, and results. Submissions are then assessed through expert peer review using criteria such as annotation quality, model robustness, reproducibility, and innovation. This approach encourages deeper engagement with the data and promotes transparency and creativity. It also provides valuable insights into labeling ambiguities and real-world deployment challenges. This method also introduces subjectivity, allows participants the opportunity to selectively present only favorable data while ignoring contradictory evidence, can be time-intensive for both participants and reviewers, and may complicate direct comparisons due to differences in evaluation protocols or data usage.

Hybrid Evaluation

A hybrid evaluation framework is also possible, combining the objectivity of ground truth-based evaluation with the depth and flexibility of qualitative expert review. In such models, a standardized test dataset may serve as the primary evaluation basis, while supplemental reports may be used to highlight innovative methods, real-world readiness, or robustness to edge cases. These qualitative factors could contribute to the overall score, serve as the basis for bonus recognition, or warrant secondary awards—enhancing the evaluation process without undermining comparability.

Taken together, these approaches offer a spectrum of options for assessing ISS, from highly controlled benchmarking to exploratory and context-aware evaluation. The choice of evaluation method depends on the goals of the assessment—whether the emphasis is on standardization, practical deployment insights, or a balance of both—as well as the resources available to support the challenge.

Benefits of Hybrid Evaluation Approach

After careful consideration of various evaluation methodologies, a hybrid evaluation approach (skewed more heavily toward the ground truth evaluation) was used for evaluating the participating teams' submissions to the U.S. DOT Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing. This approach model offered multiple strategic advantages aligned with both the objectives of the Challenge and the needs of the participants.

First and foremost, the ground truth component of the hybrid evaluation approach created a level playing field for all participants by providing a common dataset with high-quality labels. This ensured that teams were evaluated solely on the performance of their algorithms for detection, classification, localization, path prediction, and conflict prediction—rather than on their ability to deploy hardware or collect and curate data. By focusing the evaluation on core perception and prediction tasks, this approach encouraged innovation in the underlying data fusion and AI problem space, rather than in supporting infrastructure. Additionally, teams were required to submit revised concept papers as part of the evaluation process. While these papers had a smaller impact on the overall Stage 1B score, they served an important role in evaluation by describing and recognizing novel approaches and creative methodologies.

The decision to utilize a hybrid evaluation approach was also inspired by precedents set by other federal agencies, such as the unmanned aerial vehicle, glider, and ground (UG2)+ Prize Challenge organized by the Intelligence Advanced Research Projects Activity (IARPA), the xView Challenge developed by the Defense Innovation Unit (DIU) under the U.S. Department of Defense (DoD), and the Passenger Screening Algorithm Challenge developed by the U.S. Department of Homeland Security (DHS), all of which have used similar approaches in their technical challenges to great effect. These examples demonstrated that centralized data collection can not only streamline the evaluation process but also elevate the scientific rigor and reproducibility of the results obtained.

Importantly, the U.S. DOT has access to a well-instrumented, controlled roadway intersection at TFHRC, equipped with a variety of sensors, which allows for the generation of rich, multi-modal datasets. Leveraging this infrastructure enabled the Challenge organizers to provide the participants with high-quality, synchronized data that would otherwise be prohibitively difficult and costly for individual teams to acquire on their own. This approach also lowered the barrier to entry by eliminating the need for participants to perform their own sensor deployment and calibration (including complex intrinsic and extrinsic procedures), reducing both the technical burden and financial cost. By removing these early-stage logistical hurdles, the Challenge was made more available to a broader and more assorted set of teams.

Additionally, this approach will serve as a steppingstone toward more realistic deployment scenarios in future ISS activities. Teams could incrementally build their capabilities, starting from algorithm development on standardized data, and gradually progressing toward live system integration and deployment. This phased strategy reduced early-stage risk and enabled more robust safety solutions to emerge over time.

In summary, the hybrid evaluation approach offered a practical, balanced, and technically sound foundation for the Stage 1B assessment of the Intersection Safety Challenge. It aligned with best practices, reduced unnecessary overhead for participants, and helped focus attention on the fundamental technical challenges of developing a robust effective ISS.

Chapter 2. Methodology

This section focuses on the ground truth evaluation that was used for the data science competition portion of the U.S. DOT Intersection Safety Challenge Stage 1B hybrid evaluation approach.

Setting up the Data Science Competition

The U.S. DOT set up Stage 1B of the Intersection Safety Challenge as a data science competition in which all participating teams' ISS were assessed against the same test dataset. The test dataset was created from the real-world sensor data collected from the controlled test intersection at the FHWA's TFHRC that was reserved for evaluation.

Three Technical Elements

Algorithms submitted by participating teams were assessed on their ability to quickly and accurately perform three key technical elements of ISS operations: (1) the simultaneous detection, classification, and localization of a number of pedestrians and vehicles in the intersection; (2) path prediction for those identified pedestrians and vehicles; and (3) conflict prediction between and among the identified pedestrians and vehicles. "Potential conflict" refers to scenarios that intentionally created a risk of or actual occurrence of two road users (i.e., pedestrians and vehicles) occupying the same location in the intersection facility at the same time (i.e., a collision). Non-conflict refers to scenarios that intentionally avoided that risk and the actual occurrence of two road users occupying the same location on the intersection facility at the same time. Note that no actual human pedestrians or bicyclists were put at risk of being involved in a collision during the full data collection campaign.

Conflict mitigation and intervention, another critical element of the overall ISS concept, was not assessed in the Stage 1B data science competition but may be a focus of potential future activities on ISS prototyping.

Competition Data

To assess the three key technical elements of ISS operations as part of the Stage 1B data science competition, multi-sensor, multi-condition data was collected at a single signalized four-way controlled intersection at the FHWA TFHRC Smart Intersection facility in McLean, VA from October 2023 through March 2024. The data was collected from 20 roadside sensors and traffic control devices, including eight closed-circuit television (CCTV) visual cameras, five thermal cameras, two LiDARs, and four radar sensors. Additionally, the traffic signal phase and timing (SPaT) data and vehicle and/or pedestrian calls to the traffic signal controller (if any) were also collected.

Data was collected for various potential conflict-based experimental scenarios and non-conflict-based experimental scenarios. Each data collection experimental scenario included a range of experimental conditions, where certain factors varied within a set of desired levels/values, such as vehicle speeds, road user types and speeds, road user props, and time-of-day conditions. Moreover, additional variability was

introduced in the data collection execution for both potential conflict- and non-conflict-based scenarios to improve the alignment with real world-equivalent intersection scenarios as much as possible. Details on the conditions executed for each scenario, as well as additional variabilities introduced, were not provided to the teams as they were part of the Stage 1B Challenge. Sensor data was organized by each "run," which is defined as the roughly one to two minutes during which various scripted and non-scripted vehicular and other road user movement occurred on or adjacent to the controlled roadway intersection.

The majority of the collected data was provided to participants for use in developing, training, and validating their ISS algorithms. The total size of the dataset is approximately 1 TB and is available at the ITS DataHub.[1] As noted previously, a portion of the collected data, referred to as the test dataset, was held back and reserved for evaluation of the ISS concepts by the U.S. DOT. This test dataset may be made available to the public in the future.

The U.S. DOT aimed to collect as much realistic intersection operating data as possible, while working within the constraints of a large-scale operating facility, including schedule, cost, and other considerations. A few general notes about the data collected from the intersection:

- **Realistic behaviors:** The dataset contains real vehicles and other road users, although the traffic volume was generally lower than what would be expected at a comparable public real-world intersection given constraints in the controlled testbed. The pedestrians and bicyclists were instructed to behave as they normally would and to follow all applicable traffic rules. Similarly, driver behaviors were aimed at modeling realistic driver behaviors.
- **Variety of road users:** The U.S. DOT placed emphasis on collecting data that included a variety of road users using both mobility improving (e.g., manual and motorized wheelchairs) and impeding (e.g., carrying a large box) props.
- **Surrogate road users to ensure safety:** Surrogate pedestrians (i.e., robotic pedestrian dummies) and surrogate bicyclists were used in a number of runs to provide a variety of use cases, especially when safety was a potential concern. The surrogate pedestrians and bicyclists wore thermal vests to help make them detectable by the thermal cameras.
- **Variety of conditions:** The data was collected over a series of months, covering a range of weather conditions (e.g., sunshine, rain, wind, snow), as well as during both daytime and nighttime conditions.
- **Sensor considerations:** Sensors were mounted at different angles to resemble a realistic intersection configuration. The data was collected using real sensors that sometimes malfunctioned. While a variety of quality control procedures were in place, some of the data still contain errors (e.g., discoloration, blurring) or are missing for one or more sensors entirely for a given run, reflective of potential real-world situations.

Additional Considerations

While the Stage 1B data science competition was limited in its ability to assess real-time conflict predictions and mitigations, it was designed so that teams had to rely on their algorithm's performance rather than on manual labeling of the test dataset. This was ensured by having a First Data Submission within 72 hours of the test dataset release, thereby constraining the amount of time available to teams to

[1] To view a sample of the Intersection Safety Challenge Stage 1B data and for instructions to access the full dataset, please visit: https://data.transportation.gov/Roadways-and-Bridges/Intersection-Safety-Challenge-Stage-1B-Sample/vq7s-mv3v/about_data.

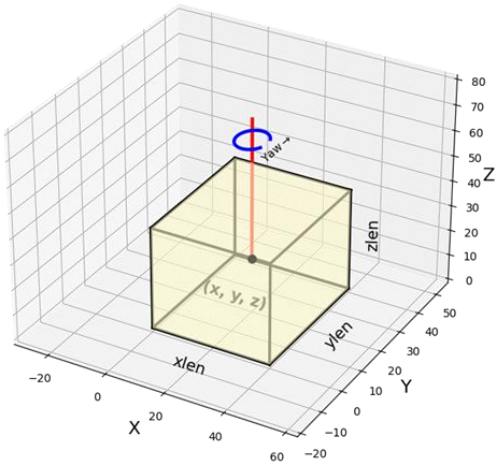
hand label each test run. Without receiving feedback on their First Data Submission performance, teams were also required to submit a Final Data Submission six weeks after the initial test data release, allowing them to have more time to work on and with the data and to finalize their individual technical approaches.

Due to the complexity and range of requirements for the Stage 1B data science competition, the technical contractor support team to the U.S. DOT created a custom portal to manage the participant team accounts, data sharing, and submissions.

Deriving Ground Truth

Participants were given three large distinct sets of data: training, validation, and test data. For the purposes of Stage 1B, training data refers to sensor output data that was not labelled. This data was provided to the participants for training and establishing a baseline with their systems. Validation data refers to data that included annotations (with ground truth). The validation data included runs with the same types of sensor information included in the training data, but with accompanying ground truth labels for which more information is provided later in this section. Furthermore, binary labels of "conflict" or "no conflict" for each validation run were provided. Validation data allowed participants to potentially verify their system performance and prepare for the test data release. The training and validation data were released incrementally over a period of weeks through a secure data management platform. Test data included additional runs with similar information to the validation data, however, in the test data the data labels were withheld for evaluation. Therefore, the participants only received unlabeled sensor output data as part of the test dataset.

As mentioned, the first technical element that participants were evaluated on was the detection, classification, and localization of the moving road users of interest in each scenario tested. For the purposes of ground truth, these properties of the road users are represented by annotated three-dimensional (3D) bounding boxes with fields: (subclass, x_center, x_length, y_center, y_length, z_center, z_length, yaw). **Figure 1** shows a visualization of an example ground truth bounding box on the frame of reference. The parameters x_center, y_center, and z_center represent the centroid of the bounding box within the 3D coordinate space, shown as x, y, and z in the figure. The length variables represent the dimensions of the bounding box. Each dimension was represented in meters, with yaw within a range of [0, 360) degrees.



Source: U.S. DOT

Figure 1. Ground Truth Bounding Box Example (Axes in meters). Source: U.S. DOT

The ground truth path prediction label was determined by these fields evaluated over a set of timestamps provided to participants. The ground truth binary "conflict" versus "no conflict" label was determined analytically using the surrogate safety measure of time-to-collision (TTC), which is the time when two road users' paths are expected to intersect if their velocities do not change. If this value was calculated to be less than or equal to a specified TTC threshold at any point during a particular run, a label of "conflict" was assigned to the run. Otherwise, a label of "no conflict" was assigned to that run.

Subclass refers to the road user categorical label for each 3D bounding box. The list of potential subclasses comprises a variety of vehicle, real actor pedestrians and bicyclists, and surrogate pedestrian and bicyclist labels (e.g., "Adult Using Stroller") and was provided to the participants so they knew which potential subclasses would be evaluated.

Participants were expected to submit their results for each technical element following the provided submission format, within a defined evaluation area of interest. As mentioned earlier, the evaluation area of interest was a 3D coordinate system placed in the frame of the LiDAR sensors deployed at the intersection.

Labeling Ground Truth

Initially, there were discussions around using global positioning system (GPS) data for ground truth. The data was collected by using high-precision (real-time kinematic [RTK]-corrected) GPS receivers on moving objects. The data also consisted of high-precision (RTK-corrected) GPS readings of pre-defined locations around the intersection (stationary GPS reading). While the GPS data seemed to be initially suitable to this use case, upon closer inspection, issues with both the real actor and surrogate road user GPS receivers resulted in inconsistent errors when displaying elevation readings, which negatively impacted calibration and ground truth efforts. To improve the accuracy of pedestrian and vehicle positioning, additional ground truth generation techniques were explored and implemented.

Ultimately, positioning ground truth data was derived by labelling 3D bounding boxes within the intersection area of interest for each validation dataset and test dataset run. The LiDAR data was used to support ground truth generation. The LiDAR data was provided in packet capture (PCAP) files. These files were converted to point clouds, then stitched together using calibration extrinsic parameters to form a cohesive 3D representation of the intersection. LiDAR point clouds were processed and labeled for key objects in the evaluation area, then exported with timestamps for validation and testing.

Determining Sensor Data Cutoff Times for Prediction

The test dataset reflected only potential pre-conflict conditions at the intersection, i.e., runs were cut off at certain timestamps to exclude the potential conflict period so participants could not "cheat" by manually labeling the correct conflict label from visual inspection. Participants were required to predict vehicle, pedestrian, and bicyclist paths and potential conflicts after the cutoff point for a specified prediction time horizon for each run within the provided test dataset.

Sensor data for test dataset runs was cut off at a cutoff point based on varied proximity to a conflict defining moment (i.e., TTC) or based on key intersection traversing movements of road users within the experimental scenario, meaning each run did not have its sensor data cut at the exact same time. This was done to ensure that there was enough sensor data provided for each run to allow participants to be able to predict the subsequent paths of pedestrians, bicyclists, and vehicles, and any potential resulting conflicts. Participants were asked to submit predictions of the position of each road user within the run for five seconds, or 50 timestamps, after the sensor data was cut off.

To determine an estimate of the cutoff point, the TTC was calculated a priori by the Challenge organizers. During this calculation process, a conflict point within the intersection was identified by calculating the point of potential collision based on trajectories and velocities derived from latitude and longitude data for each road user of interest. Distance to the point was calculated, velocity was derived, and the time needed for each road user to reach that conflict point was determined. The TTC was analyzed to determine if two road users would have overlapping positions if they did not change their velocities within a certain time threshold, indicating an impending potential conflict.

Providing Submission Formats

The expected submission formats as well as example submission files were provided to the participant teams for their detection, classification, and localization, path prediction, and conflict prediction. These formats were provided so that the teams knew how to format their own individual results against the provided test dataset and could therefore prepare a pipeline to output their results appropriately before the official test dataset was released.

Submission Format for Detection, Classification, and Localization

Participants received transformation parameters for transforming their algorithm outputs into the ground truth frame of reference. They were also provided ground truth timestamps, corresponding to the sensor frame. For each run within the test dataset, a comma-separated values (CSV) file containing the following information was expected to be submitted by the participant teams:

- Indices: Timestamps
- Columns: [subclass, x_center, x_length, y_center, y_length, z_center, z_length, z_rotation]

The columns labeled “center” and “length” refer to the specific geometric center coordinate of the submission bounding box and length of the submission bounding box, respectively for each road user detected. Z_rotation (or yaw) refers to the rotation of the box around the z axis. Each dimension was represented in meters, with yaw within a range of [0, 360) degrees. Each road user detected within the evaluation region would have an associated row within the submission data and evaluated one-to-one with the ground truth.

Submission Format for Path Prediction

For path prediction, participants were expected to provide up to three predicted paths for road users after the cutoff time. The column “path_ID” was used to group the locations of individual road users together to form their paths. Each predicted path for a road user was assigned a confidence score, with all scores summing to 1.0 per road user. Participants were expected to submit a CSV file for each run within the test data that included the following indices and columns:

- Indices: Timestamps
- Columns: [path_ID, subclass, x_center, y_center, z_center, x_length, y_length, z_length, confidence_score]

Submission Format for Conflict Prediction

For conflict prediction, participants were expected to provide a single CSV file with one row for each test data run. For each data run, participants had to provide the binary classification of “conflict” or “no conflict” between two road users (based on TTC) in that particular scenario. In addition to denoting this

binary classification when submitting their results for the specified test data, the following attributes were expected to be included in the above-mentioned CSV file for runs in the test data for which a conflict was identified:

- The timestamp of the conflict
- Subclass of road user 1 involved in the conflict
- Subclass of road user 2 involved in the conflict

The CSV submitted was expected to include the following indices and columns:

- Indices: Run_ID
- Columns: [conflict_no_conflict_label, timestamp_conflict, road_user1_subclass, road_user2_subclass]

Selecting Evaluation Metrics

Evaluation metrics for object detection and path prediction tasks are well-established, comprehensible, and able to capture information specific to the respective Stage 1B evaluation tasks. For instance, precision, recall, F1 metric, and average precision (AP) are commonly referenced in the object classification literature [5]. Similarly, path prediction studies typically make use of final displacement error (FDE) and average displacement error (ADE) [4]. While the object tracking and path prediction literatures are saturated with numerous approaches, mean Average-Precision (mAP) and ADE have been featured in open competitions resembling Stage 1B [6].

Evaluation Metrics for Detection, Classification, and Localization

The detection, classification, and localization capabilities of each team's submission were collectively evaluated by the mAP metric. mAP scores capture the Average Precision (AP)—interpreted as the estimated area under the precision-recall curve—across several precision thresholds. While AP captures both precision and recall information, which are typically in tension, it can only be calculated after tabulating the true positives, false positives, and false negatives of an algorithm across all classes at a particular threshold value (interpreted as the strictness of localization).

MAP enables evaluation of AP without choosing an arbitrary precision threshold and provides a single metric encapsulating the performance of multiple classes. Additionally, mAP was calculated at both the class and subclass levels, weighted equally to capture coarse (class) and fine (subclass) classification information. An algorithm's capacity to distinguish between vehicles, pedestrians, and bicyclists was considered critical for the Challenge even if it failed to distinguish between subclasses of those road users.

MAP is not the only possible metric for evaluating detection, classification, and localization performance, but it is a staple of the computer vision literature [5]. Additionally, mAP has been tested in similar contexts to the Stage 1B competition: the well-known Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset successfully benchmarked online submissions for localization tasks using mAP and provided a roadmap to a tested implementation [4].

Evaluation Metrics for Path Prediction

Path prediction proficiency in each scenario's runs was evaluated by the ADE: the Euclidean distance between the predicted class location and the actual class location, averaged across all timestamps.

Teams could submit up to three predicted paths, weighted by confidence. While ADE is a common and intuitive metric in the trajectory prediction literature [6], the formulation used for Stage 1B enabled comparison between predictions made under different levels of uncertainty. ADE was weighted to compromise between class and subclass identification, again distinguishing between coarse and fine classification.

Similar to mAP, ADE is present in the broader trajectory prediction literature as well as specifically in the KITTI trajectory prediction benchmark [4], a close analog to Stage 1B's path prediction assessment. At the time of drafting the evaluation plan, the choice was made—considering the variable uncertainties of launching a novel competition—to parallel existing similar benchmarks.

Evaluation Metrics for Conflict Prediction

Scoring for conflict prediction in each scenario's runs was based primarily on F2 score. F2 score was selected for the purposes of capturing both precision and recall information, biased towards road safety, while collecting challenging, speculative collision identifications without penalty. The F2 portion of a participant team's score for this evaluation component was selected to reward recall, i.e., the identification of a true conflict, over precision, the accuracy of positive predictions. It was determined that conflict prediction evaluation needed to reflect that, for the purposes of collision detection in an intersection setting, false negatives were a greater safety risk than false positives.

A small percent (10 percent) of the conflict prediction score was predicated on the incorporation of the timestamp and road user subclasses involved in a predicted conflict. As long as participant teams indicated these pieces of information for each run ID for which they predicted a conflict and the ground truth reflected a conflict, they received the entirety of the 10 percent score. The additional 10 percent of credit given to teams with timestamp and road user subclass information provided the U.S. DOT insight into the feasibility of estimating potential conflict details without unduly docking team scores.

Chapter 3. Key Insights

This section summarizes key insights from designing, planning, implementing, and evaluating the U.S. DOT Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing.

Overall Effectiveness

The structure of the U.S. DOT Intersection Safety Challenge Stage 1B proved largely effective in evaluating key technical elements of an ISS, particularly within a controlled virtual testing context. By using standardized sensor data from the real-world operation of a four-way signalized intersection at the FHWA TFHRC, the Challenge allowed participating teams to test and demonstrate their capabilities in detection, classification, localization, path prediction, and conflict prediction.

Most teams successfully met both the 72-hour and 6-week submission deadlines, suggesting that the competition design was feasible and conducive to iterative development under realistic constraints. Although differences between initial and final submissions were generally minor, they still reflected meaningful team engagement and improvement without the benefit of real-time feedback from the organizers. While detection, classification, and localization were relatively well-handled across teams, the tasks of predicting paths and potential conflicts proved considerably more complex. This complexity was further compounded by Challenge constraints, such as cutoff times for sensor data and the absence of real-time system updates, making predictive tasks more difficult than what might be expected in live deployment scenarios. Nonetheless, the competition setting ensured a level playing field and allowed for fair performance comparisons. Key observations and insights regarding the overall effectiveness of the Stage 1B data science competition are summarized below:

- **Task-based performance variability clarified technical maturity.** Detection, classification, and localization tasks were the most successfully executed across the board. These foundational capabilities showed a level of technical maturity in most team approaches. In contrast, path and conflict prediction tasks revealed greater variability and challenge. This suggests that while foundational perception capabilities are advancing, more work is needed in predictive modeling and decision support within ISS systems.
- **Challenge constraints successfully focused on algorithmic performance.** The competition relied on pre-collected, standardized sensor data collected from a controlled intersection testbed, rather than on teams deploying their own sensor setups or interacting with live systems. This removed variability in data quality and ensured that all teams worked from an identical dataset. While this limited the realism of real-time perception and system adaptation, it was a deliberate choice that enabled consistent benchmarking. By providing the same multi-modal sensor inputs to all participants—without real-time feedback—the Challenge emphasized algorithmic performance over hardware tuning, making it possible to isolate core technical capabilities across different team solutions.
- **Performance gaps were uncovered.** Certain test elements consistently revealed technical limitations. For example, the surrogate child subclass was notably difficult for teams to detect and predict accurately, indicating a potential area for further research. Performance also varied depending on situational factors. For example, teams performed better during daytime runs compared to nighttime runs. Right-turn scenarios yielded better results than left-turn scenarios.

Slower-moving road users were classified and localized more accurately, while faster-moving users were easier to predict in terms of trajectory. These findings highlight the importance of scenario variation in benchmarking ISS performance.

- **Results highlighted potentially optimal sensor fusion approaches for ISS.** Teams that employed multi-sensor fusion approaches demonstrated higher performance compared to those that used single-sensor type approaches.

In summary, the Stage 1B data science competition structure effectively balanced fairness, technical rigor, and scenario complexity, allowing meaningful differentiation of the assessment of each team's capabilities. While the virtual and non-real-time (asynchronous) nature of the competition reduced operational realism, it nonetheless served as a robust platform for evaluating the core functions of ISS technologies. The results offer a valuable snapshot of the current state of the field, helping to identify both promising approaches and areas where additional research and development are most needed.

Challenge Structure Retrospective

The U.S. DOT Intersection Safety Challenge Stage 1B featured a well-designed structure to evaluate the core components of a number of individual intersection safety systems. A central strength of this design was its ability to create a level playing field between the multiple teams' submissions by mandating the use of a standardized, real-world sensor dataset labeled with road user trajectories and conflict events provided by the U.S. DOT (note that test dataset labels were withheld for scoring).

The use of a single shared dataset also removed variability caused by team-specific sensor configurations or intersection settings, promoting fairness and reproducibility. The use of the same dataset also emphasized the effectiveness of algorithms over reliance on specialized hardware or proprietary inputs. Moreover, this uniform dataset offered the U.S. DOT insight into how well the teams' algorithms could integrate with existing intersection sensor infrastructure, highlighting their real-world adaptability. By eliminating the need for teams to perform their own sensor calibration, the Challenge reduced technical barriers, enabling more focus on algorithmic development.

The U.S. DOT used a hybrid system testing and evaluation model that leaned heavily on ground truth-based scoring against this test dataset, an approach that emphasized objective and quantifiable performance. This model ensured a consistent and rigorous evaluation across all participating teams on core system capabilities, including detection, classification, localization, path prediction, and conflict prediction. Teams were scored against the same reference data, reinforcing scientific rigor while lowering the barrier for participation, especially for those without access to expensive sensors or large-scale data collection or testbed infrastructure.

A tiered submission model added further value. The first 72-hour submission period encouraged rapid, automated solutions and discouraged manual data labeling, preserving the integrity of the Challenge. Meanwhile, the six-week final submission window gave teams ample time to refine and mature their solutions, allowing both quick prototyping and robust system development to be evaluated.

In addition to quantitative scoring, a qualitative component in the form of revised concept papers gave teams the opportunity to share system-level context, motivations, and design rationale. While these papers had a limited influence on the final scores, they helped provide a more holistic view of each team's individual approach.

Strengths of the Challenge Structure

- **Clear expectations:** While the Challenge required significant upfront planning, this effort paid off. The same information was shared with all participating teams at the same time, allowing for equal access to information. All teams knew what to expect—including data and submission formats—and could prepare for their competition submission even before the final comprehensive test dataset was released. Similarly, the U.S. DOT evaluators knew exactly what to expect from the submissions before the final submission deadline.
- **Focused on algorithms:** Providing and assessing all teams on the same datasets ensured fairness and allowed evaluation based on algorithm performance, rather than hardware or setup. This also helped to lower the barrier-to-entry for teams without extensive resources for custom data collection using sensors of their own.
- **Reduced calibration requirement:** Teams were provided with calibrated sensor data, removing the need for time-consuming calibration—but still had the flexibility to recalibrate if they chose.
- **Enabled consistent, quantitative scoring:** Using ground truth-based evaluation enabled consistent, objective measurement of perception and prediction capabilities, promoting fairness.
- **Prevented "cheating" while allowing iterative refinement:** The first 72-hour window incentivized automation and discouraged "cheating" in the form of manual data handling while the final 6-week submission window allowed for iterative improvement and deeper system development.
- **Allowed granular performance comparisons:** Since all teams submitted results on the same curated dataset that included various conditions, this enabled performance comparisons across specific conditions (e.g., night, left turns, child surrogates) and functional areas (e.g., detection vs. conflict prediction).
- **Provided space for explanation:** The qualitative input (i.e., the updated concept papers) allowed teams to explain their reasoning and design logic, enriching the evaluation with system-level insight.
- **Demonstrated scalability and comparability:** The Challenge structure successfully facilitated participation from thirteen teams and allowed them the creative freedom to meet the Challenge objectives. It also helped to highlight gaps in capability and areas for future research.
- **Protected participants' Intellectual Property (IP):** Since the participating teams only had to submit their results and updated concept papers, they did not have to worry about losing IP related to their algorithms. This may have encouraged broader participation, including from private industry.

Limitations of the Challenge Structure

- **Did not allow for real-time, (synchronous) system-in-the-loop assessment:** Use of static datasets limited the ability to evaluate system latency, responsiveness, and live decision-making, which are critical features for operational deployment.
- **Limited the evaluation to a single intersection:** The dataset was collected for a single controlled test intersection. This may have led to model overfitting to this one specific relatively simple intersection layout and reduced model generalizability.
- **Restricted hardware innovation:** The standard dataset precluded teams from demonstrating skills in sensor selection, mounting, calibration, or data labeling—critical components in end-to-end ISS deployment.
- **Restricted modeling flexibility:** Teams were required to follow a binary conflict/non-conflict labeling scheme based on fixed time-to-collision thresholds, which constrained the opportunity to implement more nuanced or context-aware definitions of conflict.

- **Limited in its ability to create and share new products:** Because participating teams were only required to submit their results and updated concept papers, the U.S. DOT did not gain access to their code, algorithms, or other related products and could not share them. In this case, the data-driven results were deemed more valuable than acquiring the code for this Challenge, so this was not considered a major limitation.

In summary, the U.S. DOT Intersection Safety Challenge Stage 1B successfully created an objective, structured competition to rigorously evaluate the algorithmic capabilities of ISS solutions, focused on perception and prediction. It enabled comparative insights into team performance and assessment reproducibility. However, the constraints of a pre-collected dataset limited the evaluation of real-time responsiveness, end-to-end system integration, and operational variety. Future ISS activities should consider testing real-time capabilities and allowing innovation in sensor-related workflows to better reflect operational realities. Future efforts may also benefit from blending objective scoring with broader, more dynamic evaluation elements to better gauge real-world readiness and innovation.

Technical Lessons Learned

This section summarizes technical lessons learned from planning, designing, executing, and evaluating the U.S. DOT Intersection Safety Challenge Stage 1B data science competition.

- **Anticipate and mitigate sensor calibration complexity early in the process.** The calibration process for visual camera, thermal camera, LiDAR, and radar sensors was labor-intensive and required tailored methods for each modality. Challenges included poor thermal image resolution, inaccurate GPS location data for pedestrians, and being limited to default, pre-determined sensor placements in the field. These factors directly impacted ground truth quality and downstream usability. Future work should reserve adequate time for calibration and consider using high-precision GPS systems and improved sensor placements, especially for critical pedestrian tracking.
- **Engage vendors early and secure data access agreements for proprietary sensors like radar.** Accessing and preparing data from certain sensors—particularly radar—can require substantial time, effort, and coordination with vendors. In Stage 1B, working with radar data involved navigating proprietary restrictions and developing custom processing workflows. Future challenges should prepare in advance by engaging vendors early and securing necessary data access agreements, especially if radar or similarly constrained modalities are to be included.
- **Use advanced surrogates or other techniques to better represent pedestrians and bicyclists.** Due to necessary restrictions on using human subjects in safety-critical scenarios, surrogate pedestrians and bicyclists (i.e., pedestrian and bicyclist dummies) were deployed in the data collection process. While necessary, this introduced new challenges. The surrogate pedestrians and bicyclists could not mimic the full range of natural human movement patterns, impacting the realism of detection and prediction tasks. Additionally, to improve thermal camera performance, thermal vests were added to surrogate pedestrians and bicyclists—an effective but imperfect workaround. Future efforts should consider more advanced surrogates or simulation techniques to better approximate human behavior and improve sensor compatibility.
- **Validate submission formats.** Adherence to formatting requirements was essential for automated scoring, which relied on structured outputs aligned with the ground truth labels. While offering teams pre-built example submission templates helped reduce formatting errors, some teams still deviated from the requirements, introducing processing delays. Automated emails were sent to the teams confirming their valid submissions and notifying them if a submission was invalid. In the case of an invalid submission, the email shared basic information regarding why

their submission was invalid (e.g., “missing localization data”). Future competitions should continue specifying and enforcing structured templates, enhance input validation in the submission portal, and provide more detailed error feedback to the teams when possible.

- **Plan robust data storage, distribution, and backend infrastructure to support teams efficiently.** To manage bandwidth, costs, and security, data was distributed to teams using pre-signed Amazon Web Services (AWS) uniform resource locators (URLs) in segmented zip files. A process also had to be developed around managing data submissions and ensuring changes could be made quickly. While the setup worked well, it required robust backend infrastructure including: a continuous deployment pipeline to manage and rapidly release stable software updates as necessary; authentication and user management; time-restricted download link generation; tracking feedback issuance; and manual support dashboards to troubleshoot issues and monitor engagement. Future competitions should plan for these systems in advance and ensure backend stability, particularly for high-volume data exchange scenarios.

Potential Future Enhancements

Building on the lessons from Stage 1B, several enhancements can be implemented to strengthen future intersection safety system activities and similar data-driven competitions. These span improvements in challenge design, technical infrastructure, data management, and participant engagement.

- **Have all data collected, cleaned, and validated in advance.** A key improvement area is ensuring all datasets—including complete ground truth labeling, training data, and sensor calibration—are finalized and validated before launching the competition. This would allow teams full data access from the outset and reduce the risk of mid-challenge disruptions or inconsistencies. Ground truth data should continue to be reserved for evaluation purposes only, with labeled validation subsets released to support teams’ algorithm development and ensure consistency in submission formats.
- **Formalize the platform strategy early on and ensure regulatory alignment.** To manage both regulatory and operational needs, future challenges should establish a dual-platform strategy early on—leveraging Challenge.gov for official postings and a custom cloud platform for data exchange, submission intake, and team interaction. This includes pre-planning access for federal staff, accounting for upload limitations, and addressing potential risks. All phases and submission types should be defined upfront to ensure compliance with Federal regulations and avoid administrative delays.
- **Strengthen cross-contractor coordination.** Future phases should expand cross-contractor team coordination mechanisms—especially between sensor deployment, data collection, data processing, and evaluation teams—to reduce misalignments and minimize development delays.
- **Upgrade equipment and consider other strategies to improve data quality.** Investing in higher-fidelity positioning systems (e.g., Real-Time Kinematic GPS) and improving sensor calibration protocols would enhance data quality. Early vendor engagement and pre-arranged access agreements are essential if radar or similarly constrained sensor modalities are required. Future efforts should also explore incorporating additional safe human subjects and more realistic surrogate road users or simulation-based road users that better replicate human behavior. This is particularly important for detection and path prediction tasks. These kinds of investments in data quality, before data collection begins, are extremely valuable as the datasets live on well past the challenge itself.
- **Enhance evaluation mechanisms and iterative feedback.** The fixed-dataset, no-performance-feedback structure used in Stage 1B, while fair, may have limited teams’ abilities to improve their models. Future efforts could pilot interim scoring opportunities, sandbox instances, leaderboards,

and/or post-submission feedback cycles to promote iterative learning and innovation, particularly for complex tasks like path or conflict prediction.

- **Scale backend infrastructure and automation.** Administration dashboards, user authentication, and real-time monitoring tools were critical in Stage 1B. Future challenges should continue developing and refining robust backend systems like these for managing submissions, tracking downloads, and issuing feedback, especially for large-scale data science competitions involving more teams in the future.

Lessons learned from the U.S. DOT Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing, which was designed as a data science competition, can help inform future related efforts for intersection safety systems research and development, as well as similar data science competitions involving multi-sensor data.

References

[1] U.S. DOT. (2024, January 8). “U.S. DOT Announces Winners of the Intersection Safety Challenge.” https://www.transportation.gov/briefing-room/us-dot-announces-winners-intersection-safety-challenge

[2] U.S. DOT. (2024, December 9). “Intersection Safety Challenge Stage 1B Sample” [Open-Access Data Portal]. ITS DataHub. https://data.transportation.gov/Roadways-and-Bridges/Intersection-Safety-Challenge-Stage-1B-Training-Da/vq7s-mv3v/about_data

[3] U.S. DOT. (2025, January 7). “U.S. DOT Announces Winners of the Intersection Safety Challenge Stage 1B: System Assessment and Virtual Testing.” https://www.transportation.gov/briefing-room/us-dot-announces-winners-intersection-safety-challenge-stage-1b-system-assessment-and

[4] A. Geiger, P. Lenz, R. Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite.” In: *Conference on Computer Vision and Pattern Recognition (CVPR).* 2012.

[5] Z. Zou, K. Chen, Z. Shi, Y. Guo and J. Ye, "Object Detection in 20 Years: A Survey," *IEEE,* pp. 257 - 276, 2023.

[6] X. Guo, M. Adl, B. Abdi and A. Emadi, "Intersection-Specific Trajectory Prediction for Road Users: A Review," *IEEE,* pp. 40054 - 40075, 2025.

U.S. Department of Transportation
ITS Joint Program Office – HOIT
1200 New Jersey Avenue, SE
Washington, DC 20590

Toll-Free “Help Line” 866-367-7487

www.its.dot.gov

[FHWA-JPO-26-005]

U.S. Department of Transportation