COOPERATIVE & AUTOMATED TRANSPORTATION PROGRAM

TEXAS DEPARTMENT OF TRANSPORTATION

What are the Challenges of CAT Data?

CAT Brief

Strategic Initiatives and Innovation Division

November 2024

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

AV	Automated Vehicle
CAT	Cooperative and Automated Transportation
CAV	Connected and Automated Vehicle
CV	Connected Vehicle
FOIA	Freedom of Information Act
GPS	Global Positioning System
IoT	Internet of Things
ITS	Intelligent Transportation System
RSU	Roadside Unit
SME	Subject Matter Expert
TTI	Texas A&M Transportation Institute
TxDOT	Texas Department of Transportation
V2X	Vehicle-to-Everything

1. Introduction

The Texas Department of Transportation's (TxDOT's) transportation network is the largest in the nation, with more than 80,900 miles of roadways, 271 general aviation airports, and 78 million tons of cargo moving through its seaports. For TxDOT, continuing to investigate innovative and economical methods for creating a safe and reliable transportation system for all users is critical. The integration of connected vehicles (CVs), automated vehicles (AVs), and other emerging transportation technologies into the state's multimodal transportation system offers numerous potential benefits to the traveling public. To support the efficient integration of these technologies, TxDOT launched the Cooperative and Automated Transportation (CAT) Program based on the CAT Strategic Plan and Program Plan.

The statewide CAT program supports the agency's mission of Connecting You With Texas. This program explores emerging technologies to address some of the state's greatest challenges in the areas of safety, mobility, environment, and funding, while executing strategies to accommodate disruptive changes and emerging technology trends. TxDOT seeks to proactively integrate CAT initiatives into transportation projects, from planning, design, and construction to operations and maintenance, rather than respond reactively to its proliferation in the multimodal transportation system.

A key initiative is to fully understand the challenges associated with CAT data.

1.1 Project Background

The rapid growth of CAT infrastructure[1] and related technologies in CVs/AVs is generating an exponential increase in data volume, velocity, and variety. To effectively manage and leverage CAT data, it is crucial for TxDOT to develop a robust digital infrastructure to handle the vast amounts of data produced, and this presents significant data management challenges to TxDOT.

This CAT brief, "What are the Challenges of CAT Data?" comprehensively explores and documents these challenges. Key areas of investigation include data storage, collection, sharing, user management, analytical needs, and standardization. By proactively addressing these challenges, TxDOT aims to harness the potential of CAT technologies to support its core functions of delivering a safe and reliable transportation system.

> **CAT data is any information integrated into, generated, or disseminated by Intelligent Transportation System (ITS) infrastructure directly or indirectly for the purpose of transportation automation, connectivity, and intelligent decision-making to improve safety, mobility, efficiency, and reliability of people and goods movement.**

This CAT Brief is also part of a series of initiatives related to CAT data integration within TxDOT. **Figure 1** illustrates how this brief relates to other CAT data initiatives, highlighting their specific purposes and interconnections.

[1] CAT infrastructure refers to the physical and digital assets required to support communication, data exchange, and coordinated actions within the CAT ecosystem, directly or indirectly influencing transportation system users and operational capabilities. Source: *CAT Brief: What is CAT Infrastructure?*

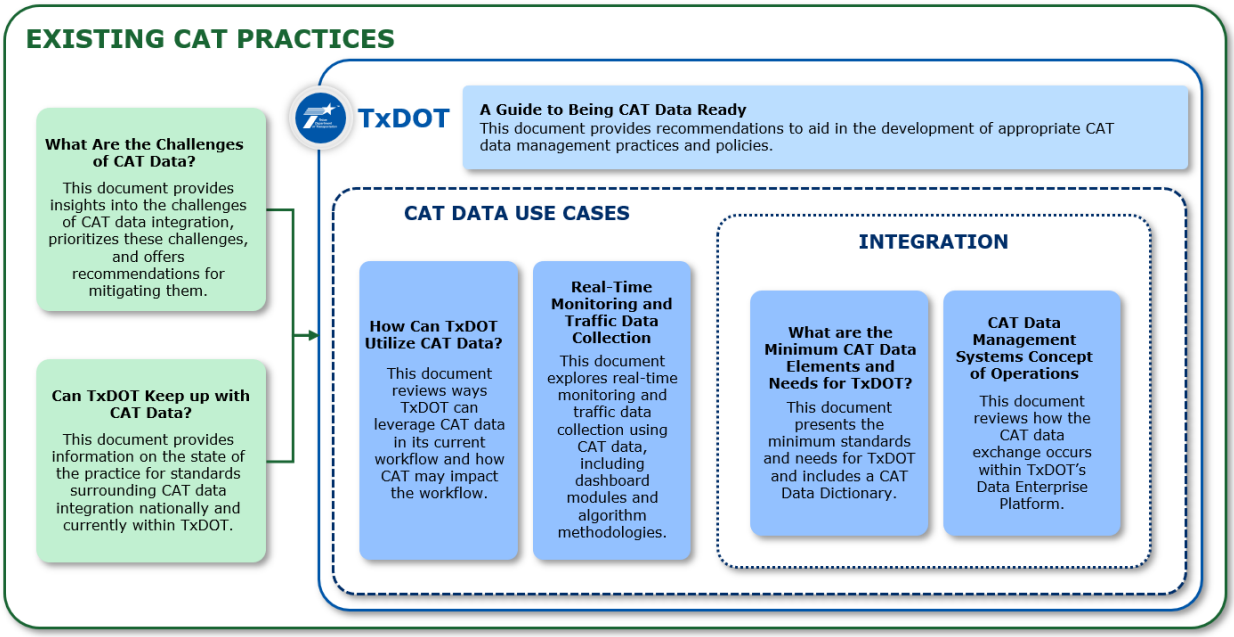

Figure 1: CAT Program Data-Related Initiatives

Figure 1 outlines a framework for TxDOT to effectively utilize and manage CAT data. It starts with establishing a foundation by assessing current practices ("Can TxDOT Keep Up with CAT Data?") and identifying challenges ("What are the Challenges of CAT Data?"). This informs the development of a guide ("A Guide to Being CAT Data Ready") for implementing CAT data use cases ("How can TxDOT Utilize CAT Data?"), such as real-time monitoring ("Real-Time Monitoring and Traffic Data Collection"). Finally, the framework addresses data integration ("What are the Minimum CAT Data Elements and Needs for TxDOT?" and "CAT Data Management Systems Concept of Operations") to ensure TxDOT is prepared to manage CAT data.

1.2 Methodology

To identify the key challenges associated with CAT data, a comprehensive literature review was conducted, focusing on CAT and emerging technology efforts nationwide (see **Appendix A: References**). Interviews were also conducted with subject matter experts (SMEs), including representatives from TxDOT divisions, the Texas A&M Transportation Institute (TTI), and transportation data vendors (see **Appendix C: Subject Matter Experts**). Based on the insights gained from the literature review and expert interviews, the top ten CAT data challenges were identified as follows:

Storage - Storage of CAT data	**Reliability and Coverage/Availability** - Gaps in data coverage
Data Latency - Time to receive CAT datasets	**Standardization Architecture/Format** - Receiving CAT data in a standardized format

Receiving - Integrating CAT data into TxDOT systems	**Analytical/Processing Needs** - Requirements for a trusted method of analyzing and processing CAT data
Sharing - Sharing CAT data with other TxDOT systems and other stakeholders	**User/Access** - Managing user and access levels of CAT data
Privacy/Security - Keeping CAT data private and secure	**Quality** - Ensuring that data collected are accurate

More information about each data challenge can be found in **Appendix B: Data Challenges**.

1.3 Organization of this CAT Brief

This CAT Brief is organized into the following sections:

- **Section 1: Introduction** – Provides an overview of the project background and summarizes the information presented in this document.
- **Section 2: CAT Data Challenges** – Identifies the challenges associated with CAT data.
- **Section 3: Prioritization of Challenges** – Presents a ranking of different CAT data challenges based on their likelihood and potential impact, as assessed in coordination with SMEs.
- **Section 4: Mitigation of CAT Data Challenges** – Reviews potential mitigation options to address challenges.

2. CAT Data Challenges

Table 1 provides a detailed overview of the challenges associated with CAT data. These challenges are often interconnected and can span multiple categories. While the relevance of specific challenges may vary depending on the CAT data application, a thorough review of all challenges was conducted to ensure a comprehensive understanding.

Cost is a significant challenge across all categories and has been carefully incorporated into the evaluation of likelihood of occurrence and potential impact of these challenges.

Table 1: CAT Data Challenges

Category	Challenges
Storage	▪ **Storage Capacity:** CAT data's massive volume can strain storage capabilities. ▪ **Data Redundancy:** Multiple sources may contribute to duplicate data and enlarged storage needs. ▪ **Cost Management**: Data's value needs to be balanced against storage costs. ▪ **Environmental Impact:** Excessive data storage can have negative environmental effects.
Data Latency	▪ **Impaired Decision-Making:** Latency hinders real-time decisions. ▪ **Delayed Emergency Response:** Delays in collecting data can affect emergency response. ▪ **Data Synchronization Issues:** Latency complicates maintaining data consistency. ▪ **Over Source or Under Source:** There is risk of procuring unnecessary data or not having data in proper latency for specific use cases requirements. ▪ **Infrastructure Limitations:** Limited infrastructure hinders data collection, especially in rural areas. ▪ **Rural Disadvantage:** Rural areas are more susceptible to latency issues due to infrastructure gaps.
Receiving	▪ **Data Overload:** High-frequency data can overwhelm systems. ▪ **Lack of Transparency:** Third-party data sources lack transparency, generating distrust in analyses. ▪ **Resource Constraints:** Limited resources affect the ability to manage and analyze received CAT data.
Sharing	▪ **Compliance:** Data sharing must comply with third-party vendor terms and conditions. ▪ **Data Quality:** Shared data must be accurate and reliable. ▪ **Proprietary Data:** Data sharing must be balanced with Freedom of Information Act (FOIA) requests for proprietary data. ▪ **Interoperability:** Shared data should be accessible and usable by all stakeholders. ▪ **Privacy Concerns when Sharing:** Agencies must be aware of CAT data that may expose sensitive location information when sharing.

Category	Challenges
Privacy/Security	▪ **Data Sensitivity:** CAT data's high granularity, numerous variables, and frequent collection pose heightened privacy risks. ▪ **Encryption:** Robust encryption is crucial for data protection. ▪ **Compliance:** Legal and regulatory requirements for data privacy and security must be met, including retention and deletion policies. ▪ **Limited Experience:** The novelty of CAT data can lead to a lack of established best practices for security and privacy management. ▪ **Privacy Concerns when Sharing/Receiving:** Agencies must be aware of CAT data that may expose sensitive location data when sharing and receiving data.
Reliability and Coverage/ Availability	▪ **Data Availability:** Lack of widespread sensor and communication infrastructure can limit data access, especially in rural areas. ▪ **Data Procurement:** Market instability and vendor challenges (e.g., Wejo's bankruptcy in 2023) affect data acquisition.
Standardization Architecture/ Format	▪ **Lack of Standards:** CAT data, especially vehicle telemetry and weather data, lacks standardization across vendors. ▪ **Procurement Challenges:** Standardized procurement language is needed to ensure data meets TxDOT's needs and provides transparency. ▪ **Legacy System Integration:** Integrating CAT data into older digital systems can be inefficient. ▪ **Other Industries:** Other industries such as the insurance industry are setting standards to meet their needs and may not be aligned with transportation agency standard practices or needs. ▪ **Measurement Details:** It is crucial to determine and understand whether the data is measured through observation, inference, or derivation.
Analytical/ Processing Needs	▪ **Data Analysis:** Non-standardized data requires more effort for cleaning and validation. ▪ **Transparency:** Lack of transparency in data sources and processing raises quality concerns. ▪ **New Roles:** Data analytics expertise is needed to manage CAT data effectively.
User/Access	▪ **Interoperability:** Shared data should be accessible and usable by all stakeholders. ▪ **Policy Awareness:** Users must understand and adhere to TxDOT's Data Classification policy.

Category
Quality

A clear understanding of implications and potential solutions is essential to navigate these challenges effectively. The following sections will provide a detailed analysis of each challenge, rank them based on their likelihood of occurrence and potential impact, and offer actionable recommendations for mitigation.

3. Prioritization of Challenges

CAT data challenge categories are further divided into subcategories, detailed in **Appendix B: Data Challenges**. These subcategories were evaluated on their likelihood of occurring and potential impact, yielding an overall *CAT Data Challenge Category Score*, which is used to prioritize challenges.

Challenges are ranked based on their "Likelihood of Occurring" and "Impact Ranking" (e.g., High-Medium). For example, a High-Medium ranking indicates a high probability of encountering the challenge and a medium level of effort required to mitigate it, with no associated safety risk or liability.

Each challenge subcategory's priority is determined by combining the Likelihood and Impact scores (as defined in **Table 2**). Averaging subcategory scores yields an overall score for each data challenge category, with higher scores signifying higher priority. These rankings inform mitigation strategies but recognize that many actions will address multiple, interconnected categories.

While determined by SMEs and thus subjective, these scores aid in prioritizing different CAT data challenges. **Table 3** presents the Categorical scores.

Table 2: Ranking Categories

	Ranking	Definition	Score
Likelihood of Occurring	High	High likelihood of occurring.	3
	Low	Low likelihood of occurring.	1
Impact	High	If this challenge occurs, there will be a high cost, including safety risk to the public and/or liability to TxDOT.	3
	Medium	If this challenge occurs, it will not have safety risks/liabilities but could have high or medium cost/effort to mitigate.	2
	Low	If this challenge occurs, no major efforts are required to mitigate it or there is no impact to the public or TxDOT. There are no safety concerns if it occurs.	1

Table 3: CAT Data Challenge Category Scores

Category	Score
Data Latency	Use Case Specific
Standardization Architecture/Format	5.5
Receiving	5.3
Quality	4.6
Analytical/Processing Needs	4.3
Privacy/Security	4.2
Storage	4.0
Sharing	4.0
User/Access	3.9
Reliability and Coverage/Availability	3.8

*Note: Category Scores are based on combining subcategory scores of the likelihood of occurring (high (3), medium (2), and low (1)) and impact if they do occur (high (3), medium (2), and low (1)) and averaging them for each challenge category. See **Table 2** for more details.*

The analysis suggests that "Standardization Architecture/Format" is the highest-priority challenge category for CAT data integration. All these categories are interconnected with one another. For example, a standardized architecture and format would improve data analytics, storage maintenance, privacy/security processes, and understanding of the received CAT data. Similarly, challenges in data reception and privacy/security are linked to multiple other categories. These interdependencies suggest that addressing one challenge can positively impact others.

The significance of latency is use case specific. A lack of understanding the use case needs for data latency may result in a challenge when deploying the use case application. Latency should be considered for each stage of data: collection, sending to TxDOT application, and time to process and use the data. For applications requiring sub-second latency, this challenge may be ranked higher. Use cases that can tolerate a few seconds of latency have a medium ranking, while those not requiring near real-time data may have a lower priority.

4. Mitigation of CAT Data Challenges

Since these challenges are interdependent, working to mitigate one challenge may impact other categories. **Table 4** presents recommendations for addressing these challenges and indicates which categories would be affected by each recommendation.

Table 4: Recommendations for Mitigating CAT Data Challenges

#1: Coalition of Agencies for Standards and Requirements	Potential Challenges Impacted
Forming an interagency coalition focused on CAT data standardization could significantly benefit TxDOT. Such a coalition would have greater influence on vendors, encouraging them to provide comprehensive information for utilizing their datasets, while also facilitating the exchange of best practices and lessons learned.	• Receiving • Privacy/Security • Standardization Architecture/Format • Analytical/Processing Needs • Quality
#2: Review of Legacy Systems	**Potential Challenges Impacted**
CAT data's high volume and various formats pose integration challenges for legacy systems such as Lonestar. A feasibility review, including compatibility testing, is needed to assess integration possibilities. Alternatively, TxDOT could explore the option of implementing a new system or segregating data in an enterprise data platform for on-demand access.	• Receiving • Standardization Architecture/Format • Analytical/Processing Needs • Quality • Sharing
#3: Contract Requirements for Procuring CAT Data	**Potential Challenges Impacted**
CAT data procurement faces challenges such as unknown sources and processing procedures. TxDOT should continue to require vendors to disclose such information and collaborate with others to set industry standards, driving advancements despite potential vendor reluctance. The connection between the use case requirements and the source-procurement mechanism is important to understand along with who/what entity in TxDOT has the authority to determine or change these requirements.	• Receiving • Privacy/Security • Standardization Architecture/Format • Analytical/Processing Needs • Quality
#4: Consider Edge Processing	**Potential Challenges Impacted**
Increased CAT data can be managed using edge processing and only storing needed data. Edge processing can clean up data before it is received by TxDOT's Data Enterprise Platform.	• Storage • Receiving • Privacy/Security • Standardization Architecture/Format • Analytical/Processing Needs

#5: Aggregated Data	Potential Challenges Impacted
When reviewing a use case for CAT data, TxDOT should determine whether aggregated data will be suitable. There are four forms of data: raw, concealed raw, aggregated, and synthesized. TxDOT is focused on protecting and respecting privacy by not procuring or managing non-anonymized data and using aggregated data. There are different levels of aggregation layered on top of one another that must be understood for each use case. Aggregation may include temporal, spatial, source, and reporting aggregation. Understanding if a sample of data or all available data points are required will provide the desired actions and insights.	• Privacy/Security • Analytical/Processing Needs • Sharing
#6: Prepare for Increased Storage	**Potential Challenges Impacted**
The increase in CAT data requires scalable and cost-effective storage solutions. Data lifecycle, volume, performance, scalability, and security are crucial considerations for planning storage needs.	• Storage • Receiving • Privacy/Security
#7: Data Best Practices for Data Storage	**Potential Challenges Impacted**
CAT data can be managed by defining best practices and developing use cases. Determining the required data, its frequency, and data life cycle ensures the relevance of collected data variables. Developing data retention practices and implementing a Data Quality Assurance process can reduce unnecessary data storage.	• Storage • Receiving • Privacy/Security • Standardization Architecture/Format • Analytical/Processing Needs
#8: Policy Review and Modifications	**Potential Challenges Impacted**
The integration of emerging CAT data requires a comprehensive review of existing policies, including data retention, storage, privacy, and access. Data integration policies need to be adaptable to support TxDOT's sustainable and responsible management of CAT data.	• Privacy/Security • Receiving • Sharing • User/Access • Storage
#9: Review How CAT Will Impact all Business Areas	**Potential Challenges Impacted**
The emergence of CAT data requires a review of affected divisions, potentially altering infrastructure design and requiring new skill sets from new hires or training existing employees. TxDOT may need to modify its business model, particularly in areas of data governance, data management, training programs available, organizational structure, and resource allocation, to fully benefit from these emerging technologies.	• User/Access • Analytical/Processing Needs

#10: Virtual Sensors and Roadside Units (RSUs)	Potential Challenges Impacted
Virtual sensors and RSUs have the potential to reduce capital and maintenance costs associated with CAT data collection. They also reduce fieldwork risks, address workforce concerns, and offer a more sustainable data collection approach. These tools can fill gaps in ITS physical infrastructure, providing better coverage and insights.	• Analytical/Processing Needs • Quality • User/Access • Reliability and Coverage/Availability

TxDOT should conduct cost-benefit analyses for all CAT data deployments and develop effective communication strategies to increase public awareness and support. For successful mitigation, it is crucial to address each challenge during the planning phase of CAT use cases. Specific actions will vary depending on whether data are sourced internally or externally. Reducing risks where possible and strategically addressing necessary challenges will help balance costs and benefits.

By understanding and mitigating these challenges, TxDOT can effectively leverage CAT data to improve transportation systems while ensuring responsible and efficient data management.

Appendix A: References

AECOM. (2020). *TxDOT Statewide TSMO Data Platform.*

Amazon Web Services. (2021*). Designing Next Generation Vehicles Communication with AWS IoT Core and MQTT: Implementation Guide OR AWS Whitepaper.* Amazon Web Services.

Bennet, S. (2023). *AI Provides Vital Assistance to TMS Companies.* Transport Topics.

Burbano, L. (2021). *What Is A Smart Parking System? Functionalities and Benefits.* Tomorrow City.

Correos, E. (2023). *Connected Cars and Data Privacy: Challenges and Solutions in a New Automotive Era.* Telecom Review.

Crow, M., Avery, P. A., & Fuller, J. (2023). *Achieving Efficiencies within ODOT with the Event Streaming Platform.*

Masello, L., Castignani, G., Sheehan, B., Murphy, F., & McDonnell, K. (2022). *On the road safety benefits of advanced driver assistance systems in different driving contexts.* Transportation Research Interdisciplinary Perspectives.

NCHRP. (2024). *Using Cooperative Automated Transportation Data for Freeway Operational Strategies.* NCHRP Research Report 1080.

Okas, J. (2023, Feb 6). *Data Management: The Challenge of Connected Cars*. https://real-wireless.com/data-management-the-challenge-of-connected-cars/

Sahar, D. (2018). *Challenges of Securing Connected Car.* Upstream. https://upstream.auto/resources/upstream-chalk-talks-challenges-of-securing-connected-cars/

Smalley, S. (2023). *As car hoover up more and more driver data, is it time to regulate the industry?* The Record. https://therecord.media/connected-cars-hoover-up-data

The Intelligent Transportation Society of America. (2024). *ITS Technology Use Case Library.* Washington, D.C.

Appendix B: Data Challenges

Challenges vary based on data acquisition methods. Balancing costs and benefits requires mitigating risks when possible and strategically addressing necessary challenges.

Table 5: Storage Data Challenge Subcategories

Storage	**Storage of Cooperative and Automated Transportation (CAT) data**	
Subcategories	**Description**	**Score (ranking and value)**
Data Growth	The most evident challenge is the exponential growth of data. As more data are generated through various channels, organizations and individuals need to find scalable storage solutions.	High-Medium (5)
Data Redundancy and Duplication	Data can be stored in multiple locations, leading to redundancy and increased storage costs. Deduplication (removing duplicate copies of repeating data) and data synchronization will minimize this issue.	High-Medium (5)
Data Backup and Recovery	It is critical to ensure data are backed up and can be restored in case of data loss due to hardware failures, data corruption, or disasters.	Low-Low (2)
Scalability	Data storage solutions need to scale easily to accommodate growth. Scaling can be complex, especially for on-premises storage systems.	Medium-Medium (4)
Data Retention and Lifecycle Management	Defining data retention policies and automating the lifecycle of data from creation to archiving and deletion can be complex and require careful planning.	Medium-Medium (4)
Cost Management	It is crucial to manage storage costs, which can escalate rapidly. Balancing performance and capacity needs with budget constraints is an ongoing challenge.	High-High (6)
Hardware and Software Compatibility	Storage solutions often involve a mix of hardware and software components. It is essential to ensure these components are compatible and optimized for performance.	Medium-Medium (4)
Data Archiving and Retrieval	Managing long-term data archiving and ensuring that archived data can be retrieved when needed can be a complex task.	Low-Low (2)
Environmental Concerns	Energy efficiency and the environmental impact of data centers and storage solutions are growing concerns as data storage infrastructure expands.	Medium-Medium (4)

Table 6: Receiving Data Challenge Subcategories

Receiving Integrating CAT data into TxDOT systems		
Subcategories	**Description**	**Score (ranking and value)**
Data Security	It is critical to ensure the security of data exchanged between vehicles and infrastructure. Protecting against data breaches, hacking, and unauthorized access is a constant concern.	High-High (6)
Data Privacy	The need for data sharing must be balanced with privacy concerns. Regulations and best practices for handling personal data in connected and autonomous transportation systems need to be established.	High-High (6)
Data Standards	To ensure interoperability, data formats and communication protocols need to be standardized across different manufacturers and transportation systems.	High-High (6)
Data Quality and Reliability	The data received must be accurate and reliable. Poor data quality can lead to incorrect decisions and potentially unsafe situations. TxDOT must be able to receive and process data reliably, even in adverse conditions such as heavy rain, snow, or fog, where sensors and communication systems may be affected.	High-High (6)
Latency	Data transmission and processing must happen with minimal delay, especially for autonomous vehicles that require real-time data for decision-making. High latency can result in accidents or operational inefficiencies.	Use Case Specific
Availability of Data	Lack of vendors for specific data sets can make it difficult to procure data for use cases.	Medium-Medium (4)
Contract for Procurement	Contracts need to be tailored for emerging technologies, which can be challenging given the lack of experience with these technologies and data.	High-Medium (5)
Interference and Connectivity Issues	Interference with wireless communication signals, for example due to physical obstacles or electromagnetic interference, can disrupt data transmission. Ensuring the reliability of infrastructure components, such as roadside sensors and communication networks, is crucial. Downtime or technical issues can disrupt data exchange.	Use Case Specific
Lack of Resources	Documentation (or a point of contact) to assist with data questions can be lacking or non-existent.	High-Medium (5)
Scalability	The infrastructure must be scaled to handle and process a large volume of data traffic efficiently.	High-High (6)
Cross-Vendor Compatibility	Different manufacturers may use different technologies and communication protocols, which can make effective communication between vehicles and infrastructure from different vendors problematic.	Medium-Medium (4)

Table 7: Sharing Data Challenge Subcategories

Sharing	Sharing CAT data with other TxDOT systems and other stakeholders	
Subcategories	**Description**	**Score (ranking and value)**
Data Security	Ensuring the security of data exchanged between vehicles and infrastructure is critical. Protecting against data breaches, hacking, and unauthorized access is a constant concern and ever evolving. Navigating the complex landscape of data regulations and standards across different regions can be difficult. Data-sharing practices must align with local and international laws. Implementing robust cybersecurity measures to protect data during transmission and storage is critical.	High-High (6)
Data Privacy	Balancing the need for data sharing with privacy concerns can be challenging. Regulations and best practices for handling personal data in connected and autonomous transportation systems need to be established. There may be ethical dilemmas related to data sharing, particularly when it involves situations that affect safety, privacy, and individual rights.	High-High (6)
Data Ownership	Determining who owns the data generated can be complex, particularly when data are collected from multiple sources, including vehicle manufacturers, infrastructure providers, and third-party applications.	Low-Low (2)
Data Synchronization and Interoperability	Ensuring data consistency and synchronization across different sources and systems is a challenge. Divergent data practices can result in data conflicts.	High-High (6)
Documentation	Educating the public, policymakers, and industry stakeholders about the benefits and risks of data sharing is essential to garner support and address concerns.	Low-Low (2)
Scalability	As more vehicles become connected and autonomous (or more infrastructure sensors), the infrastructure must scale to handle the increased volume of data traffic.	Low-Low (2)
Compliance	Vehicle-to-everything (V2X) communication and data exchange must conform to national and local regulations.	Low-Low (2)
Data Quality and Reliability	Poor-quality data can lead to incorrect decisions and operational inefficiencies.	High-High (6)

Table 8: Privacy/Security Data Challenge Subcategories

Privacy/Security	Keeping CAT data private and secure	
Subcategories	**Description**	**Score (ranking and value)**
Data Breaches	Cyberattacks and data breaches are a constant threat, and even organizations with strong security measures in place can be vulnerable.	Medium-High (5)
Internal Threats or Human Error	Employees, contractors, or partners with access to sensitive data may intentionally or unintentionally compromise data security. Internal threats can be challenging to prevent and detect. Human error remains a leading cause of data breaches. Inadequate training and awareness can lead to incidents that compromise data security.	Medium-Medium (4)
External Sharing	Many organizations share data with third-party vendors or service providers. Ensuring that these partners maintain data privacy and security standards can be challenging.	High-High (6)
Data Encryption	Encryption is a critical security measure, ensuring that data are encrypted properly, both in transit and at rest, can be complex.	High-High (6)
Compliance	Complying with data residency requirements, which dictate where data can be stored, can be challenging for multinational organizations. Ensuring that data are stored in a manner that complies with relevant laws is crucial. Different industries and regions have specific data privacy and security regulations. Ensuring compliance with these requirements can be complex.	Low-Low (2)
Data Lifecycle	Data must be properly managed throughout its lifecycle, from creation to disposal. Secure data erasure and data retention policies are necessary.	Medium-Low (3)
Access Control	Managing access to sensitive data and ensuring that only authorized users can view or modify it is challenging. The principle of least privilege should be applied but managing user access can be complex in large organizations.	Medium-Medium (4)
Technology Maturity	The proliferation of emerging technologies and Internet of Things (IoT) devices can lack robust security features and can be targeted by attackers.	High-High (6)
Monitoring and Auditing	Regularly monitoring data access and conducting audits to detect anomalies and unauthorized activities can be resource intensive.	Low-Low (2)

Table 9: Standardization/Architecture Format Data Challenge Subcategories

Standardization Architecture/Format	Getting CAT data in a standardized format	
Subcategories	**Description**	**Score (ranking and value)**
Various Data Sources	Data may come from various sources, including different industries, devices, and software platforms, each with its own data format and structure.	High-High (6)
Legacy Systems	Older systems often use proprietary data formats or outdated standards, making data integration and compatibility difficult.	High-High (6)
Versioning	Changes and updates to data standards can create compatibility issues between systems that use different versions of the same standard.	High-High (6)
Data Volume	Handling large volumes of data and transforming those data into a standardized format can be computationally intensive and may require significant processing power and storage.	High-High (6)
Data Synchronization	Ensuring data synchronization across different sources and systems is a challenge. Divergent data practices can result in data conflicts.	High-Medium (5)
Contract for Procurement	Contracts need to be tailored for emerging technologies, which can be challenging given the lack of experience with these technologies and data.	High-Low (4)
Heterogeneous Data	Combining structured, semi-structured, and unstructured data, such as text, images, and video, creates format and standardization challenges.	High-Medium (5)
Cost	Converting existing data into a new standardized format can be costly, in terms of both time and resources.	High-High (6)
Other Industries	Other industries such as the insurance industry are setting standards to meet their needs and may not be aligned with transportation agency standard practices or needs.	High-High (6)
Measurement Details	Determining and understanding if the data is measured by being observed, inferred, or derived.	High-Medium (5)

Table 10: Analytical/Processing Needs Data Challenge Subcategories

Analytical / Processing Needs	Requirements for analyzing and processing CAT data	
Subcategories	**Description**	**Score (ranking and value)**
Data Volume	Connected and Automated Vehicles (CAVs) generate vast amounts of data from sensors, cameras, LiDAR, radar, global positioning system (GPS), and onboard computers. Managing and processing this massive volume of data can be resource intensive.	High-Medium (5)
Various Data Sources	The lack of standards for sensor data creates challenges for analyzing the data properly. Additionally, there is a lack of transparency for the processing that occurred before reaching TxDOT.	High-Medium (5)
Data Variety	Data include not only sensor data but also vehicle telemetry, communications data, environmental data, and more. Combining and making sense of these varied data is challenging.	Medium-Low (3)
Real-Time Processing	Many CAV applications, such as navigation and collision avoidance, require real-time data processing and decision-making. Ensuring low-latency data analytics is essential.	Use Case Specific
Lack of Resources	Hiring personnel who can dedicate time to data analytics for these emerging technologies is costly and difficult. Also, procuring data at the latency required can be costly, if available at all.	High-Medium (5)
Data Privacy	Data can include personal and location-specific information, raising concerns about data privacy and regulatory compliance.	High-Medium (5)
Data Security	Data must be protected from unauthorized access, tampering, and cyberattacks, as compromising the security of these systems can lead to accidents and privacy breaches.	High-Medium (5)
Compliance	Adhering to data privacy and security regulations while collecting, transmitting, and analyzing CAV data is essential but can be a complex task due to evolving regulations.	Low-Low (2)

Table 11: User/Access Data Challenge Subcategories

User / Access	Managing user and access levels of CAT data	
Subcategories	**Description**	**Score (ranking and value)**
Data Privacy	Users may be concerned about their data being collected and shared as part of transportation data. Ensuring that data privacy regulations are adhered to and that users have control over their data can be challenging.	High-Medium (5)
Data Security	Protecting transportation data from unauthorized access, breaches, and cyber threats is a key challenge. Users want assurance that their data are secure.	High-Medium (5)
Data Ownership	Determining who owns the data and who has control over it can be complex, particularly when multiple stakeholders are involved in data collection and management.	Low-Low (2)
Consent Management	Obtaining and managing user consent for data collection and usage is necessary but can be challenging, especially when dealing with real-time data from connected vehicles (CVs).	High-Medium (5)
Access	Ensuring that all user groups, including those with limited digital literacy or resources, have access to transportation data is a challenge.	High-Low (4)
Identity Verification	Verifying the identities of users and ensuring that only authorized individuals or entities access certain data can be a security challenge.	High-Low (4)
Cross-Platform Access	Providing users with seamless access to transportation data across various platforms, devices, and services can be complex, as it requires integration and interoperability.	Low-Low (2)
Data Portability	Allowing users to access and export their data from transportation systems and services, and ensuring data portability, can be a challenge, especially when data is siloed.	Medium-Medium (4)

Table 12: Quality Data Challenge Subcategories

Quality Ensuring that data collected from various sensors, such as GPS, LiDAR, radar, and cameras, are accurate is crucial for the reliable functioning of CV systems.		
Subcategories	**Description**	**Score (ranking and value)**
Data Noise	Sensor data can be affected by noise, interference, and environmental factors, leading to inaccuracies. Filtering out noise while preserving useful data can be challenging.	High-Medium (5)
Data Volume	CVs generate vast amounts of data, which can overwhelm data storage and processing systems. Managing and processing large data volumes efficiently is a challenge.	High-Medium (5)
Data Synchronization	Data collected from multiple vehicles must be synchronized to provide a comprehensive view of traffic conditions and other relevant information. Achieving synchronization can be complex, especially in areas with poor network connectivity.	High-Medium (5)
Data Compression	Reducing the size of data for efficient transmission and storage is necessary, but it must be done without sacrificing data quality or resolution.	High-Medium (5)
Machine Learning for Data Correction	Using machine learning and artificial intelligence to correct and enhance data quality can be challenging, especially when dealing with real-world variability.	Low-Low (2)
Data Validation	Validating data to ensure adherence to expected ranges and formats is necessary for identifying and addressing errors and anomalies.	High-Medium (5)
Data Aggregation	Aggregating data from multiple sources and vehicles into meaningful and actionable information requires advanced data processing and analysis techniques.	High-Medium (5)

Table 13: Reliability and Coverage/Availability Data Challenge Subcategories

Reliability and Coverage/Availability Subcategories	Gaps in data coverage Description	Score (ranking and value)
Network Infrastructure	The availability and reliability of communication networks is critical for collecting and transmitting data. Ensuring comprehensive coverage, particularly in rural or remote areas, can be challenging.	High-High (6)
Network Latency	Low-latency networks are essential for real-time communication between CVs and infrastructure. High latency can result in delayed data transmission and decision-making, impacting safety-critical applications.	Use Case Specific
Network Congestion	As more vehicles and devices connect to the network, network congestion can affect data transmission and availability. Managing network congestion in high-traffic areas is a challenge.	High-Low (4)
Interference	Radio interference, signal blocking, and electronic jamming can disrupt data transmission in CV systems, posing challenges to data availability and reliability.	Low-Low (2)
Data Transmission Protocols	Implementing efficient and secure data transmission protocols is crucial for data availability. Ensuring that data are transmitted reliably and securely is a complex task.	Medium-Low (3)
Data Coverage Gaps	Identifying and addressing areas with limited or no network coverage, such as tunnels, deep urban canyons, and remote regions, is a significant challenge.	High-Medium (5)
Availability of Data	Lack of vendors for specific data sets can cause challenges in procuring data for use cases.	Medium-Medium (4)
Data Monetization	Balancing the need to monetize CV data with ethical data use and respect for user privacy can be a delicate issue, especially when dealing with user-generated data.	Low-Low (2)
Data Integration	Integrating data from various sources, including different vehicle models and manufacturers, presents challenges related to data compatibility and consistency.	Medium-Low (3)
Data Synchronization	Synchronizing data from multiple vehicles and infrastructure components is essential for creating a coherent and accurate view of the transportation ecosystem.	High-Medium (5)

Table 14: Data Latency Data Challenge Subcategories

Data Latency	Time to receive CAT datasets
Subcategories	**Description**
Decision Making	In situations where real-time decisions are critical, such as managing traffic flow, responding to accidents, or optimizing routes for delivery services, data lag can be problematic. Delayed data can lead to suboptimal decisions and inefficiencies.
Emergency Response	Autonomous vehicles and other emerging technologies need to promptly recognize and respond to emergency vehicles with lights and sirens. Data lag can hinder the timely detection of these vehicles.
Safety Concerns	Autonomous vehicles require immediate access to up-to-date data to make split-second decisions and navigate safely. Data lag can lead to accidents or unsafe driving situations.
Edge Case Scenarios	Autonomous vehicles must be prepared for edge cases, such as unexpected obstacles or challenging road conditions. Real-time data are crucial for identifying and reacting to such scenarios.
Sensor Data Latency	Sensors such as LiDAR, radar, and cameras provide essential input for autonomous vehicles. Any lag in sensor data transmission or processing can affect the vehicle's ability to perceive its environment accurately.
Real-Time Traffic and Environmental Conditions	Real-time traffic and environmental data are needed to make route decisions and avoid congestion, accidents, and road closures. Data lag can result in suboptimal route planning and delays.
Data Synchronization	Ensuring data synchronization and consistency is challenging due to potential latency.
Traffic Signal Recognition	Data lag can lead to missed or misinterpreted signals.
Over Source or Under Source	Procurement of data not useful or the data does not have the proper latency for specific use cases' needs.

Note: No score since use case specific

Appendix C: Subject Matter Experts

Name and Title	Organization
TxDOT Internal SMEs	
Benjamin McCulloch, Strategic Data Scientist Bethany Wyatt, Strategic Management Analyst Charles Tapp, Strategic Management Analyst Matt Sneed, Director of IT Traffic Technology Emilie Schulz, Data Privacy Officer Suresh Sundararajan, Data Management Officer	Texas Department of Transportation
Academy	
Shawn Turner, TTI Senior Research Engineer Martin Michael, Associate Research Scientist	Texas A&M Transportation Institute
Private Organizations / Vendors	
Laura Schewel, CEO	StreetLight/Jacobs
Terri Johnson, Director, Public Sector Services	INRIX
Other Public Agencies	
John Corbin, Automated Vehicle Program Manager John Harding, Team Lead, Connected and Automated Vehicles and Emerging Technologies	FHWA - Operations
Nick Hegemier, Managing Director of Infrastructure	Ohio DOT/Drive Ohio