Transcript of Webinar

U.S. Department of Transportation's Safety Data Initiative Request for Information

Held on Thursday, November 15, 2018, 2:00 p.m.

Operator: Ladies and gentlemen, thank you for standing by and welcome to the Safety Data Initiative RFI conference. At this time, all participants are in listen only mode. If you should require assistance during the call, please press star then zero. I would now like to turn the call over to your host, Jason Broehm. Please go ahead.

Jason Broehm: Thank you. Thank you all for joining us today for this webinar about the U.S. Department of Transportation's Safety Data Initiative and our Request for Information, or RFI. My name is Jason Broehm, and I have a few housekeeping notes before walking through the agenda, and introducing our first speaker. First, hopefully, you are all seeing that we have three poll questions on the screen at this time, and we would ask you to take a minute or so to enter your responses just so we can see how many people we have with us today and the nature of the audience. That would be helpful to us. I should mention that webinar participants' phonelines are muted for this call, but if you think of questions during the presentation, we encourage you to type them into the chat pod. We will keep a running list of questions and answer them after the presentations. We are recording today's webinar and will post it on our Safety Data Initiative webpage following the webinar. I believe David Winter here has provided that in the chat pod. If you could take a few seconds to complete the poll questions, then we will move forward.

[pause]

David Winter: It looks like most people are either data science or analytics services firm or data technology/solution solver. It looks like, as far as the skills offered, the largest number of people are predictive analytics followed by data visualization. For the most part, it's a single person viewing. We have a few people calling in where there are two or more people in the room.

Jason Broehm: Okay. Thank you, David. This is Jason, again. The purpose of today's webinar is to share information about the Safety Data Initiative, provide an overview of the RFI, and answer your questions about the RFI. We will start with Derek Kan, our Under Secretary of Transportation for Policy for introductory remarks about the vision behind the Safety Data Initiative. Next, David Winter, the director of the Federal Highway Administration's Office of Highway Policy Information, will provide an overview of what we are doing as part of the Safety Data Initiative. Then Dan Morgan, the Department's Chief Data Officer and Acting Chief Technology Officer, will walk through the RFI, the specific questions we have posed in it, and what we are hoping to learn from the responses we receive. Finally, we will leave plenty of time to answer your questions about the RFI. Next I will turn to our Under Secretary of Transportation for Policy, Derek Kan, for opening remarks.

Derek Kan: Thank you very much, Jason. Safety is a DOT's number one priority, and we have made significant progress in transportation safety over the years. However, we continue to experience challenges. Traffic fatalities increased in both 2015 and 2016, erasing years of traffic safety gains. In 2016, we suffered 37,461 traffic fatalities on U.S. roadways with around 2.5 million estimated nonfatal injuries. In 2017, there was a modest decrease to 37,133 traffic fatalities. But highway safety remains a huge challenge, and we must continue to address this problem. One problem is that we do not fully

understand why traffic fatalities has increased in recent years. What it boils down to is we need better tools to help us understand the factors contributing to this trend.

Derek Kan: While we at the DOT have hundreds of data sets at our disposal, we are not fully leveraging the information they contain to improve safety. The Department's data is siloed, and it comes in at different tempos. To give you a sense, we have over 700 or 800 different data sources, and these various data sources are analyzed separately. Many are only published annually, and many of them have been collected and organized in the same way for years, if not decades. Yet recent innovations in data science provide us with the opportunity to do so much more. We have all seen the growth of "big data" sources, which can provide huge volumes of anonymous, near real-time data, and by improving real-time data, we can create a closer feedback loop, which helps us better understand roadway operating conditions and other factors relevant to risk. Innovations such as advanced data analytics and machine learning are helping pull valuable information out of these data streams.

Derek Kan: We have an opportunity to gain new safety insights with advanced analysis and traditional and new "big data" sources. And we can provide decision-makers and front-line safety professionals with new tools and near real-time information so interventions can be focused and effective. And at the core of it, what we need to do is evolve from a retrospective, cross-sectional, years old, if not decade-old, analysis to predictive analytics to better target and understand emerging risks.

Derek Kan: What we really need to do is integrate existing databases and new private sector data sources to gain insight into safety questions so we can create new data visualizations and tools that allow rapid and rigorous analysis by decision-makers through clear, compelling, and interactive maps, graphs, and charts. And then we need to use all the tools to apply advanced analytics to identify risk patterns and start pattern matching and develop insights that help policymakers, city managers, and safety professionals to anticipate and mitigate safety risk to reduce injuries and fatalities.

Derek Kan: We publicly launched the Safety Data Initiative in January of this year, and we have been taking steps toward advancing the vision I just described. My office is leading the initiative and we are collaborating across the Department. This includes colleagues from the CIO's [Chief Information Officer] office, the Bureau of Transportation Statistics, the [Federal] Highway Administration, the National Highway Traffic Safety Administration, FRA, [Federal] Motor Carrier [Safety Administration], the [National] Transit Administration, and PHMSA. And over the last year, we have undertaken several projects; we have launched a data visualization challenge; and launched numerous outreach activities.

Derek Kan: I went to end by saying we believe what we are doing here is groundbreaking. We are changing the way we think about technology and data as it pertains to safety. For everybody on the call, who's a data scientist, a data analyst, a practitioner in the space, we invite you to partner with us to help save lives. There are few things we can do that are more meaningful than improved safety on our roadways. Next, I will turn it over to David winter who is the director of the Federal Highway Administration's Office of Highway Policy Information. David has played a key role in the Safety Data Initiative, and he's going to briefly describe each of these efforts.

David Winter: Thank you, Derek. I'm going start with a quick overview of the Safety Data Initiative pilot projects and outreach activities that I'll describe in my presentation. I will start with the data visualization that our National Highway Traffic Safety Administration, or NHTSA, developed. Then I will describe the data integration and analysis projects we have undertaken. Finally, I will highlight several outreach activities, including a data visualization challenge, currently underway.

David Winter: NHTSA maintains the Fatality Analysis Reporting System, or FARS, which is a nationwide census of fatal injuries suffered in motor vehicle crashes. Traditionally, NHTSA has reported these data in tabular format in a series of fact sheets, but NHTSA has started to experiment with new ways of presenting these data. NHTSA has used Tableau visualization software to develop an interactive visualization of its 2016 traffic fact sheet focusing on speeding. Over the summer, NHTSA released a beta version of this visualization on its website, and a link is available from the Safety Data Initiative webpage, which as Jason mentioned, is in the chat pod [https://www.transportation.gov/policy/transportation-policy/safety/safetydatainitiative]. The visualization includes a map showing the percent of traffic fatalities that were speeding-related by state along with a ranking of states across the bottom. It allows the user to explore data by sorting data into a variety of ways, including by individual State, month, date, time of day, and roadway type. Visualizations like this can provide new insight into presenting data in more interactive ways, which will help policymakers and the public.

David Winter: As Derek mentioned earlier, DOT maintains hundreds of data sets and reports, and the vast majority of these are not integrated. We have an opportunity to integrate these datasets with each other and with data from other sources, such as other Federal agencies and the private sector, to gain insights. We initiated several projects to integrate data sources for analysis. An earlier pilot project focused on better understanding the relationship pedestrian fatalities may have with the characteristics of the transportation system and the built environment in Los Angeles, California. We integrated various public sector data sets, including data from DOT's Federal Highway Administration and NHTSA, as well as data from the Environmental Protection Agency, and the U.S. Census Bureau. We found that in urban areas, traffic on non-access controlled arterial roadways was found to significantly increase pedestrian fatality risk. We also found a strong association between employment density in the retail sector and increased pedestrian fatality risk. The work from this project was published in the December 2018 issue of the journal *Accident Analysis & Prevention*. We believe that the methods we used could be applied to other local governments to help them identify locations with the highest risk of pedestrian fatalities and target safety improvement efforts accordingly.

David Winter: Another pilot project is exploring the opportunities to estimate police-reported traffic crashes in near real-time by combining crowdsourced crash data from Waze, a private sector source, with crash data submitted to the DOT by the State of Maryland through the National Highway Traffic Safety Administration's, or NHTSA's, Electronic Data Transfer pilot, or EDT. We're using DOT's Volpe National Transportation Systems Center as a fee-for-service contractor to provide machine learning techniques on these data using DOT's prototype Secure Data Commons, which involves remote machine access to computational resources. We're using software programs such as ArcGIS, R, and Python. So far, DOT has learned that the models supported with Waze data provide reasonably good estimates of police-reported crashes, which has laid the foundation for future development of a national crash count tool. Work on this project is continuing, using data from other states.

David Winter: Our rural speed pilot project is an ongoing research effort to understand the contribution of prevailing speed, speed limit, and average travel speed to the prevalence and severity of crashes on rural highways. The pilot further seeks to understand the relationship roadway design and traffic volumes have with speed and crash outcomes. We used an existing indefinite delivery/indefinite quantity contract, or IDIQ contract, through the Federal Highway Administration's Turner-Fairbank Highway Research Center to obtain analytical services from the Texas A&M Transportation Institute. The contractor has performed statistical analysis using R and geospatial tools to geospatially conflate crash

data from the Highway Safety Information System to the Federal Highway Administration's National Performance Management Research Data Set, or NPMRDS, speed data. The Federal Highway Administration purchases the speed data from the private sector for non-safety applications.

David Winter: As part of the Safety Data Initiative, we have been reaching out to stakeholders to draw on their knowledge, experience, and creativity. In June, we launched the Solving for Safety Visualization Challenge. This three-stage challenge asks participants to come up with innovative ways to visualize data that will reveal insights into serious crashes on our roads and rail systems while improving our understanding of transportation safety. We received ideation proposals from 54 Solvers, and we selected five semi-finalists to advance to Stage II. They are currently developing their initial proposals for analytical visualization tools into proofs-of-concept. The semi-finalists are: Arity, an Allstate company, is using connected vehicle and driver behavior data to explore the relationship between driving behavior and road design. Ford Motor Company is combining crash data with connected vehicle and driver behavior data to determine crash risks, test solutions, and evaluate results. Uber is combining its web-based tool that visualizes geolocation data sets, and historical speed data collected from Uber trips with NHTSA's FARS data to better visualize traffic safety. The University of Central Florida is integrating real-time and static traffic data and using predictive analytics to help diagnose real-time safety conditions. And finally, VHB is using pedestrian avatars and applying game theory techniques to help State and local transportation professionals see potential safety improvements for the pedestrian perspective.

David Winter: In June, we convened a Safety Data Forum to engage a diverse group of stakeholders in discussion about opportunities to leverage data analytics tools to predict and prevent fatalities and injuries. Participants included representatives of data and technology firms, universities, national safety organizations, and Federal, State, and local governments. Several DOT leaders and staff spoke about the vision for the Safety Data Initiative, the transportation safety challenges we are trying to solve, and the various projects I described earlier in the presentation. We also invited several State and local government representatives to speak about their data integration, visualization, and analytical efforts to highlight several best practices in using data to improve transportation safety. We are seeking to identify best practices from across the country and share information about these as well as what we are learning through our pilot projects with State and local transportation agencies that could use them.

David Winter: Now that we have described the Safety Data Initiative's vision and activities, I would like to introduce Dan Morgan, DOT's Chief Data Officer and Acting Chief Technology Officer. He will provide an overview of the RFI, the specific questions we have posed in it, what we are hoping to learn from responses to the RFI, and how this information is vital to our desire to improve our ability to use data to improve transportation safety.

Dan Morgan: Thanks so much, David. So, I think based on what you've heard from the Under Secretary and from David's recounting of our experience with Safety Data Initiative to date, you can see that we are trying to take a systematic approach to safety. It is far more than just hotspots of existing outcomes or just twisting our data around in different, new, and creative ways. Data integration is at the core of what we're trying to do, not only to identify outcomes that are happening, but risk patterns to be able to prioritize risks and develop solutions. In order to do that, we have to improve the way we collect, manage, integrate, analyze, and use that data. We know that the data environment and safety environment involves a multitude of stakeholders and data sources. You heard from David about Census data, EPA data, private-sector data from providers like Waze or the National Performance Management Research Data Set, which is a purchased dataset from INRIX, I believe. Those kinds of merging of public and private sector data sets each moving with different tempos and different approaches to collection

and different purposes still can be integrated, but it creates different kinds of analytical challenges. But it also creates many opportunities. So, when we look at her changing data environment, we have many new capabilities, technologies, and tools available to us. And in the RFI, we sort of describe different technology capabilities. It may not be the world's most complete inventory of the different things we're using in the agency. We have a variety of traditional data management analysis tools and we have been making significant investments in more modern and open source tools as well.

Dan Morgan: Obviously, transportation is in a very exciting moment where we starting to see a proliferation of data available from new sources that is increasingly timely and contemporaneous. It is increasingly granular and fast-moving and we have such an opportunity here to better understand how to apply those new data sources to get the job done. As I said, the Department's technology environment is changing rapidly, and we have been making smart investments in the kinds of technologies that we need to get insights from our data. But our capacity at the heart of many of these new analytic tools is a bit limited. And one of the reasons we have gone out with this RFI is to understand the abilities that are out there in industry so that we can get a better understanding of how to ask for the kind of help that we need. We know the data integration can be time-consuming. This is not a simple join. This is merging survey data with administrative data. This is bringing fast-moving data from a provider like Waze where we are collecting incidents around the country every two minutes through their connected citizens program with slower moving data that may be available inside the agency. Even the police crash reports that we get which don't always come in in the most highly method. Right? Even if a crash occurs that hour, you still have to wait for the law enforcement officer to get to a place where they can actually complete and enter that complete crash report, and then it has to go through a bunch of local and state administrative systems before it can even get to the U.S. Department of Transportation. So, our data supply chain can be really long. And that creates other kinds of challenges. But that is why we are looking for those kinds of different approaches as we move through.

Dan Morgan: I see there are a couple of questions in the chat pod so before I leave the slide, I can jump into it. I will leave actual questions about the RFI response towards the end here. The list of data sources that may be currently available, you can go to data.gov and filter for the U.S. Department of Transportation data catalog. That's the best way to get a few of our data sources. Many of our data sources are public and you can actually see how we release them. That may not be the way we store them inside the agency, but you can see what is in them.

Dan Morgan: Another question here was speaking more about the tools that are in use today and what is working and what is not. I think we have, I joke, we're the U.S. Department of Transportation so we have one of everything. So, if you can name a data warehousing and business intelligence tool, it is probably in use somewhere in the Department. One of the challenges that we have actually is trying to figure out how to work across those different boundaries at the agency and either virtualize or centralize, depending on how we go, our data sets, to be able to pull them together. We also have many, many different copies and uses of non-DOT data as we have talked about, using things like the American Community Survey or Longitudinal Employer Household Dynamics or unemployment statistics can be challenging in that not only are they not here inside agency, but they may be curated differently by our other Federal and non-Federal partners. And that creates challenges and understanding the concepts and utility. So, we describe number of the tools in the actual text of the RFI. So I would encourage you to review that section to have a better understanding of the tools that are in use today.

Dan Morgan: Another question here is around whether we are planning to create a data warehouse collecting data from different state DOTs. We have several data warehouses already inside the agency. And most of our data, as I mentioned earlier, is born locally and is moved to the state and is ultimately reported to us. Different datasets have different reporting requirements and reporting regimes. Generally, that's for a good reason. And so there are a number of different approaches that we have been taking to sort of move across our silos. But I think, I'm not sure that centralization is always the right answer. I think we want to pick something that is deeply relevant to our stakeholders and our data users. So I think we want to be user centric in our approach to maturing our current data warehousing strategy.

Dan Morgan: The approximate size of the data? We are not even anywhere near petabytes, we are in gigabyte stage. There are some pretty large datasets from research that may be bigger so I would say gigabyte, terabyte kinds of sizes, but not petabytes.

Dan Morgan: For external data sources, are there limitations based on security and privacy concerns? Yes. Obviously, we have a unique responsibility in the public sector to respect privacy. Confidential business information is another kind of challenge that we may have. And of course, security matters to us quite deeply as part of the various laws with which we must comply. I think as we look through, we will get some more detail moving forward. I will move to more slides in a little bit. But I think overall, you will hear a little bit more about what we think in terms of privacy and security and some of those kinds of things. And of course, we want to make sure that we hear from you about how best to manage some of those concerns as we try to move forward and really change the way we look at the way we use data for safety.

Dan Morgan: So how many users and who, internal versus external, to anticipate? I don't think we actually know. I would say outside of the FAA, we are a 12,000-employee agency and a good chunk of those folks are smart at statistics and programming and want to be radically more productive with their data. So, I think we can have a conversation. I'm not sure that is necessarily relevant right now at this stage of where we are in terms of understanding what industry they have available to us. So, I think you will hear more about what questions we are asking you in a little bit. So, let's move ahead.

Dan Morgan: Our ideal end state here is to try to overcome the challenges that we have with data integration. So, we want to find a way to you some place where data integration progresses permit the repeatable integration of data for analysis. And we want that integrated data to be useful both internally and externally. As part of this RFI, we want access to partners and vendors that can perform quick analyses to using innovative techniques to answer research questions that broaden our understanding of transportation safety through predictive insights. Now when we talk about the vendors and part opportunity, we will get into little more detail about how that would like for us. This is not purely a technology solicitation. I repeat, this is not purely a technology solicitation. It's very important to us that we are looking for folks that can build teams that include transportation subject matter experts, data scientists, engineers and technologists so that we can figure out how we can best take advantage of the datasets that we are looking for here. So, when we laid it out in RFI, we laid out with some of our objectives are. We know that we need to get to a place where we can enable a more efficient, effective, frequent, and rapid exchange of data and information between us and the rest of the community. That means a few different things. Sometimes, that means that the private sector has superior data that is moving quickly, and we need to find a way to ingest or harness that data, whether that be actually bringing it on board into a cloud environment or datalink, or if that means that we access a range of application programming interfaces from some of these private-sector data providers.

Dan Morgan: I think that when we look at faster moving data, we think we might be able to understand risk or negative safety outcomes in a much more timely manner. When we talk about the exchange of data and information, as Derek pointed out earlier, some of our datasets are published our early. Some of them have significant lag time as we work through data quality issues in the extended data supply chain. It's very important for us to be able to find ways to move more quickly so we can react to changing safety data risks. We also know that DOT will not be the sole source of data. One question in the chat pod is about whether we are interested in using social media data and all those other kinds of things. As somebody pointed out earlier that it's a bit rife with security and privacy risks. But I think we are not limiting the question of what constitutes safety data. I think it's very important, as we think through how best to address and prioritize risk, we look at as many datasets as possible to help us save lives. What we are looking for as part of this RFI, our teams or partnerships or individuals, individual organizations that might be most successful or ways to structure a request for information so we can get those kinds of types of diverse teams to bring other datasets on board. We know data ownership will vary depending on the data and their use. Knowing that we have confidential business information or some other kinds of datasets that might be, intellectual property rights might limit our ability to harness that data. So understanding that we have a diverse ecosystem of data providers. We know that we have data integration challenges. We know that we need to build our governance, use accessibility and technology standards to a place where we can actually succeed in this diverse environment. And we of course want to balance stewardship and access with encouraging open, available, and transparent data. We don't want the analytics that we provide to be a surprise. We want them to be replicable, and we want them to be useful to the transportation community writ large, not just the U.S. Department of Transportation.

Dan Morgan: As it relates to how we might approach the market, in the event that we would issue a solicitation based on what we hear from the RFI responses, we wonder about how we might achieve a balance for sufficient flexibility for a multitude of on-demand projects that can be short-term, spot analyses, to long-term, researching, multi-stage kinds of projects. It will depend on the state activity and the safety topic, the participants -- and by that we mean state and local DOTs or economic development agencies or metropolitan planning organizations or transportation safety stakeholders -- as well as the data sets and the analytical tools that are used. We wonder how we might best achieve flexibility for collaboration and exchange between U.S. DOT and members of the transportation community. Derek has laid out a vision that says we are doing this kind of work in service to the transportation community at large. So, a key feature of what might we look for as we move forward with an RFI is our ability to engage with state and local partners around the country. And ultimately, we are looking for a task management approach that will effectively manage and leverage complex relationships that go beyond a traditional government to vendor relationship. We don’t view this as a transaction. We view this as partnership if we were able to go to market. And we are looking for you, the respondents, to help us understand how we might best structure a vehicle to achieve these goals.

Dan Morgan: If we do it right, we will build our capacity to get to insights faster. We will expand our capacity to be prospective and be predictive and build upon our tradition of reporting and understanding what has happened so that we can know was is happening, what might happen, where things are changing, and direct our resources, whether they be Federal, State, or local to the things that are of highest priority. At the end of this arrangement, we should be much better at handling data quality and integration challenges. We know to some extent, I think we feel that data sources from some of these nontraditional or nongovernment providers may be superior. They may have higher quality or less missingness. But we know that's not true. Our experience to date has shown us that is not

true. Sensors get fixed. Sensors sometimes fall off and do not transmit data. We still have to deal with missingness and lack of standards across a wild range of datasets. It is important for us to be able to cope with those kinds of things and not assume that we are going to get the highest quality data from any of these other kinds of data sources. Lastly, I think we really want to mature into a place where we can deploy shared data service models with adequate security and privacy protections. We want a community of collaborators. And that means that it is not just inside the four walls of the U.S. Department of Transportation. If we do it just for us, we have not hit the mark. It has to be for the larger transportation community. And that involves all stakeholders -- state and local government, academic institutions, the private sector -- working together to save people's lives.

Dan Morgan: If you are wondering what would we want from this, from this solicitation, we think these are the kinds of deliverables we would go after. All of this is in the RFI. So sometimes it might be recruiting new data into the Department, identifying, collecting, analyzing new safety data and datasets and it does not have to be just from government. Once we have got the data, we have to figure out how to integrate it, connect it, fuse it, knowing that it might be administrative, it might be crowd sourced, it might be survey data. All of those things work together. Things we have learned about crowdsource signals, they tend to be better concentrated on interstate than on local roads. They tend to be louder doing rush-hour than the wee hours of the morning. Correcting for that bias, being able to handle the kind of bias I think they are all important things we need to think through as we bring on these new kinds of data sources. Of course, if we look at expanding our universe of data to things that include the private sector, we may need help facilitating the establishment of data management and sharing agreements across the transportation community or the nontraditional transportation community. We think we might need help actually conducting the analysis and using analytical tools, some of which may the in the Department, and some of which we may need to go get. We think we need help conducting exploratory analysis, doing the hard work of data science, and ultimately creating data products that make people feel something and to help them take action about where risk might be emerging. We need data visualization to help us see the data, understand the complexities of that data, but also manage to protect privacy and meet our security goals so that we can still communicate with the rest of the transportation community. And ultimately, my boss would make sure that I remind you that we are trying to be a shared services organization, and we believe that the best way to get to collaboration across our silos is to sponsor shared services that provide our users, whether inside or outside the Department, the ability to be radically more productive with their data. I will take a quick pause here before we go into the questions that we are asking you, the potential respondents to this RFI, to address a couple of things in the chat pod.

Dan Morgan: State DOTs are at the forefront on data collection. They do not have a data sharing infrastructure. Most of their data are being unused. Could be one of the focus areas of this RFI?

Dan Morgan: The answer is yes. As I mentioned, it is not just about the Department itself. It is about the transportation community writ large. Including state and local governments is part of this. We actually solicited feedback from the states as we developed this RFI to see if we were asking the right kinds of questions. So, I think that's definitely an area where we can go. And certainly, the Department has a great deal of technical assistance programs out there to assist the states today as they work to mature their data sharing and management practices in a variety of domains, whether it’s roadway data, crash data, motor carrier safety data. I could go on.

Dan Morgan: Another question is if the goal is to reduce risk, what attributes identify reduced risk?

Dan Morgan: Certainly, you are right. I think there are a number of different ways we can quantify risk. I think there's operational risk, there's built environment risk, a number of different domains in which risk can be reduced. Not all treatment are necessarily engineering treatments or operational treatments to the actual vehicles themselves. Generally speaking, it's going to be a systematic approach. There's a really great website that Federal Highways [the Federal Highway Administration] has put together about data driven safety analysis and thinking through systematic approaches. It involves the interaction of people in their environment and the way that we have engineered it, and the way that we try to counteract some of the human error that goes on in developing crashes. David, I'm not sure if you want to add anything to sort of how we might quantify risk from a safety perspective. It's okay if you don't.

David Winter: No, no, I think that you did a good job.

Dan Morgan: Okay. Good. Everything I learned, I learned from David. I'm glad I did a good job.

Dan Morgan: How long do you expect to reach this ideal state?

Dan Morgan: The expectation is that we are going to move, I think we are in that walk stage of the crawl, walk, run. We want to move fast. We want to be the best fast. So, I don't think we will get there uniformly very quickly, but I think we have been pursuing demonstration projects to try to move us forward quickly and accelerate our learning so we can get to the ideal state faster. If we did this just as a big research project, and waited until the end to learn everything and write a report, we'd never get there.

David Winter: And I would say that as far as our data infrastructure is concerned, we're definitely making progress with the move to the cloud of most of our data systems, or a lot of our data systems, and then purchasing additional software system to help with the data integration and helping with data governance and management. So, I think we are making good progress, but it is slow going.

Dan Morgan: There have been a couple of questions about the current state of the Department's data environment, and I would encourage folks to take a look at page 5 of the RFI where we describe the state of the Department's data environment. Pages 5 and 6 do a really great job of laying it out. If you have specific questions about what's laid out on pages 5 and 6, we'd be glad to answer those.

Dan Morgan: On the last piece, yes, cloud is available and we're still learning how to use cloud effectively. But I will point out that is going to take a team approach to build what we are looking for. It will take subject matter expertise from transportation. Just having the cloud and just having the best data science platform is not enough to get where we need to go. So, things that we are looking for and questions we are going to ask industry I think will help us, help you understand how we are going to pursue it.

Dan Morgan: Our overarching questions are actually for you. We would love to hear from industry in your response to the RFI how you think we should consider our focus areas, what frameworks and strategies are best to achieve our ideal end state. You have heard about our challenges with integration, visualization, and insight. We are interested in hearing from you based on your experience and expertise how you think we should go about that. You have seen from us what kind of deliverables we think we would be looking for underneath this RFI if we were to go underneath a procurement if we were to go out, what are we missing? What additional deliverables should we be considering based on our focus areas, and our ideal end state? We're asking you to describe your capabilities in data integration,

connecting and fusing multiple data sources. Understand that our situation here is beyond just simple join. We have challenges with probabilistic and deterministic linkage. We have challenges of dealing with data attributes like space and time. We have geospatial and geostatistical kinds of analyses. But it is not just about putting points on a map. We want to be able to understand your experience in dealing with this with high variety kind of data, and how you think we should ask for assistance in overcoming our challenges with data integration. And we would ask you to go ahead and share examples of where you have been successful in providing services that employ data science platforms and cloud data and data analysis products and services with ongoing governance and management to help us make sure that these data integration efforts are not one offs, that the data janitor work is something we can reuse and build community around. We are looking for you to describe where you have either seen or actually made success happen with the exchange and use of information amongst a number of contributing parties.

Dan Morgan: As we look through the changing data environment here at the Department, I think we've tried to lay this out in a number of ways. We have lot of different folks involved in the transportation space. I offer Waze as an example. Through their connected citizens program, they offer the ability for, if states and locals provide information about traffic incidents or construction efforts. In return they can get information about traffic jams, and crashes, and objects on roadways, and vehicles stopped on the side of the road, and those kinds of things. That data is protected. Waze views their schema as confidential business information and holds us to account for protecting their information in exchange, as part of that exchange. They are not the only one in the transportation sector. There's a lot of really great data out there that some of our private sector partners are reluctant to share unless it can be ensured that we can protect it. That's why we have been investing in things like the secure data commons to help us meet the high expectations that some of these confidential information, confidential business information providers are levying on the Department, but then we are also trying to pursue that collaboration space around those kinds of datasets while meeting those allegations. So we are interested in hearing from you. It's not just a governance question. It's really about an assurance question. We would love for you to help us understand places where you have the implemented an approach that allows us to build a data exchange ecosystem. The system is not just a technology. It is a set of processes and policies that go around it. Help us understand how we can find those kinds of triple bottom line partnerships, who you might bring to the table, how we might structure a potential procurement to encourage that kind of innovative partnering so that we can build those kinds of new partnerships.

Dan Morgan: And of course, voluntary agreements that are facilitated that enable effective data integration and analysis across government entities. My favorite example of this in the agency, is the National Transit Map. Once upon a time, Secretary Foxx asked me "how come Google knows where all the transit is, and we don't?" It turns out that Google knows where all of it is because transit agencies have been sharing open data on their portals for years. And we just never went out and figured out how to collect it. So we had to deal with an intellectual property issue where transit agencies actually placed a variety of restrictions around having no derivative products associated with their transit data, or making sure that we could not remix their datasets into a larger dataset. These kinds of challenges prevented us from developing an open picture of where transit was in the country. We were able to figure that out. We are looking for your opinion about other kinds of voluntary arrangements that you can help us with or that you can see as partnership opportunities underneath a potential procurement. So we are looking for feedback from you about what is still on the table. We don't have all the answers.

Dan Morgan: When it comes analysis, we think that all these things are part of the data science mentality that we are trying to pull together. There are two questions for you really. One, describe your experience with inferential statistics, modeling and simulation, forecasting, artificial intelligence and machine learning, and non-statistical approaches like geospatial analysis, analysis around linear referencing systems or root cause investigation, content and survey questionnaire design, audio and video digitization and analysis. There's a lot of opportunity we have not yet fully exploited and we would love to hear places where you have been successful in applying both traditional and advanced and more modern techniques to improve safety or to address a burning issue. And of course, we would love for you to tell us about how you approach exploratory data analysis, especially "big data." To be clear, the vast majority of the data we are talking about is big in terms of variety. Measuring risk in the transportation space is a variety problem first. It involves pulling in so many different sources of data to understand the environment, people's interactions with it, how space is being used, how people are moving between space. Being able to pull those kinds of things together helps us understand where we might see changing risk as well. Sometimes we might want to bring in faster moving data to see if we can see dynamics in a more contemporaneous way. So it's very important for us to be able to look across all of those different kinds of things. But we would love to hear your approach to how you would deal with these kinds of multi-signal, different size, different shape kinds of datasets and still making sense of this to help us measure transportation safety risk.

Dan Morgan: We would love to hear about your visualization capability. We think that visualization is the key to understanding where risk is changing, how is changing, how we might ultimately address some of these safety challenges. We think that risk visualization helps people relate to the story of safety and risk. We would love to hear how you approach visualization of spatial data, how you deal with "big data," and even if you have a portfolio of links or you want to provide some information, provide some samples in your response, you are welcome to do so. We would love for you to provide some examples on the management side for how you actually think about the data pipeline. Describe the beginning to the end of that data pipeline. Describe how your demonstrated ability to work with "big data" like we have talked about. We think that our big issue right now is high variety, high-volume is next, high velocity is last. That is where we are today. We think that may change quickly over the next few years. But that is the state I think in the transportation safety space. And of course, I we have laid out privacy is a special responsibility in the public sector because of the changing nature of the transportation system, and the private sector data sources that are available now, business confidentiality becomes as increasingly important consideration as we move through how we ingest some of these datasets into some of our analyses. And of course, we think that we want to build a community around it so how do we uphold those values and principles while still encouraging collaboration. Places where you have experienced that and successfully managed that. We've seen some of this happen in the health space already as we sequencing the genome or discovering new drugs so I think we are looking for you to tell us what we might be missing here.

Dan Morgan: As we think through what might happen and where we might go after we get some of these things, sometimes a consortium is the right way to go. That is the way, data science is a team sport, and transportation safety is a team sport too. Perhaps a vendor consortium might be more appropriate. What would you recommend? How do we provide for as much flexibility to get the right team in place? Could you describe a place where you have been part of or helped organize a vendor consortium that involved a mix, such as a large nonprofit, small and disadvantaged businesses, nonprofits, universities, quasi-public entities, research consortia, and other applicable vendor types? We are open to many different ways actually. And we would love to hear from you what you think is most appropriate based on what we've laid out. The last piece of this is, if we were to go to a procurement,

we would to hear from you about which kinds of public and private sector arrangements make the most sense for facilitating data analysis, integration, and visualization. We are not just limiting this to a contract. There's also things like broad agency announcements and other kinds of innovative tools that are available. What do you think is the right way to approach this based on what we have laid out and the kind of team experience and partnership that we think we might be looking for? And could you describe any of your potential examples where we can find our own sort of double or triple bottom line, where we can find mutual benefit? Are there arrangements available where there's actually a value exchange and not just to transaction between us and a potential respondent to a procurement. What would that look like, and how would you recommend we go about it? That is it. There are a couple of questions.

Jason Broehm: Dan, I've been keeping a running list.

Dan Morgan: What did I miss?

Jason Broehm: Let me go back to the beginning. I know you answered some of these earlier. One of the early questions was "Is a sample of Waze data available?"

Dan Morgan: No.

Jason Broehm: Another one, I think you may have partially answered this, but: "Can you share the current technology, tools, and interface standards used in different projects at DOT?" I know you described the current state or environment and said that pages 5 and 6 of RFI describe that well. Is there anything you would like to add to that?

Dan Morgan: In terms of interface standards, it depends on the domain, to be perfectly honest. There are a variety of standards that are out there. I will tell you that we probably do not lag for standards; we lack for conformance. While we have recommendations and guidelines for how to describe things like crashes or crash incidents, or even roadways, sometimes we don't conform. And so generally speaking, not only are the data sets that we are looking at high variety, but the actual elements are high variety too, which creates a number of integration challenges as we've sort of laid out.

Jason Broehm: Another question I think related to the earlier question about the approximate size of data. The question is: "Is data size provided per state of the grand total of data?

Dan Morgan: I think the question was like on a per data set basis versus a grand total. I think grand total-wise, we are talking terabytes.

Jason Broehm: There was a question: "When you talk about data being useful internally and externally, can you clarify what types of external users you intend to use the data and how would this be?"

Dan Morgan: I think it's much more than a website. Websites are part of dissemination for sure. But sometimes it is about really exercising decision support tools. Sometimes it is about developing an actual research paper that is ready for peer review in a refereed journal, and then developing a technical assistance program to road test some of these tools, to hear from real users. Real users can be the transportation safety manager in your city. It can be your Governor's Highway Safety Office. It could be your DOT. It could be your transit agency. It could be any of the traditional transportation stakeholders. It could also be advocacy organizations. It could be nonprofits, who are a big part of the transportation

safety community, who need communication tools to help advocate in communities. It could be, sometimes it is average citizens who just need to be able to be empowered to go to a city meeting, to be able to say "look at what I see?" and communicate effectively with their elected officials. It really depends on the kind of tools and questions we are looking for. I don't know that we have fully thought out exactly what that world would look like. If you think we can do more to describe what that world could look like, as part of your RFI response, you should lay out.

Jason Broehm: Ok, another question was: "Is DOT interested in the development of new tools to meet objectives or is the interest in what existing COTS tools are available in industry?"

Dan Morgan: I think it's a mix. Certainly, there are a lot of commercially available tools. I think we view the analytics space to be very dynamic right now. It is super exciting. Is almost impossible to keep up. I don't care if you're the private sector or the public sector. I think we are interested in hearing from you, if there's something off-the-shelf, you should tell us about it. But I remind you that is what I said at the beginning, this is not purely a technology RFI. So I think you should think through those questions we have laid out. Certainly, technology as a part of it, but it's not the only one. And certainly, I think, we think it is a mix of commercial, off-the-shelf, which you might consider commercially off-the-shelf, but I would tell you that open source tools are commercial off-the-shelf tools as well. We think it's a mix of tools to be successful.

Jason Broehm: There was a question: "How will DOT ensure the data is authoritative, high-quality, and well-managed?

Dan Morgan: You should answer that question in your response.

Jason Broehm: Another question: "What about data governance?"

Dan Morgan: You should answer that question in your response.

Jason Broehm: How are you going to determine user needs?

Dan Morgan: That's an excellent question. I think we are interested in looking for partners to understand user-centric design and user-centric approaches. So I think we are open to understanding that, remembering user needs are outside of the four walls of this building and think through how we do design thinking or user needs discovery across the span of the country can be a bit of a challenge. Right? So, we are looking for ideas, actually, on how best to approach that.

Jason Broehm: Several people had variations of this question: "What do you intend to do with the information gathered from the RFI and ultimately what is the next step?"

Dan Morgan: The first thing is, we did and RFI because weren't sure how to ask for what we wanted. We did the best we could do to describe what our vision was and what we hope for in terms of potential partnerships and the kinds of work that we think we might do under an arrangement. The goal of the RFI really is to understand from you all who is out there that can do it, what are the creative ways we might be able to approach the market, based on what we are looking for, and ultimately, I think we have to come back and look at it inside ourselves to say is there a way or multiple ways for us to go approach the market for the kind of help that we are looking for to advance the Safety Data Initiative. So I think

our initial response release to understand who is out there and then ultimately to determine how best to approach it.

Jason Broehm: So, there's a question about the deadline for the RFI and why it is so tight.

Dan Morgan: We are interested in moving fast.

Jason Broehm: I think we are caught up to the end of the chat pod at this point.

David Winter: There was a couple of about next step in the process.

Dan Morgan: The next step in the process is to review these responses, understand the market's capabilities, and see some of the recommendations that you have in terms of how we might approach this. I think it is a lot of reading for us, which I think we are excited to do. I think we have laid out a really gnarly problem, and there are a bunch of creative people on this webinar, and I'm really just excited to see what we get back.

David Winter: You laid out in your slide that we don't have a set next step right now. We're looking for you. It could be an RFP. It could be an IDIQ. It could be something else. But we are open to suggestions, and we want to hear from you, but we do not have a final next step in mind.

Dan Morgan: No, there's no milestone after this one. We have not said we would do one thing or another. Knowing that this is due and we want to get some reading done before the holidays, but also know that the holidays are going to be in the middle. You are not getting a Christmas gift of an RFP, I'm pretty sure. So, it will not happen until early next year if something were to result.

David Winter: So, Joe said, "Do you have to response to the RFI to respond to a subsequent RFP?"

Dan Morgan: No.

Jason Broehm: There's another one just before that about taking your definition of "What does a data system ecosystem look like after the RFI?" Or do you need to?

Dan Morgan: I think that's a statement. Okay. I think if you have suggestions for how you might encourage us to do that, that is a great thing to put on your RFI response.

Dan Morgan: I don't know that we had a page limitation. I am trying to look. We do not have a page limitation.

Jason Broehm: I think related to that there was a question about formatting specifications.

Dan Morgan: We do not. Please make it legible.

Jason Broehm: Ok, there's a question here: "If an RFP results from the RFI, what would that timeline look like?"

Dan Morgan: We answered that one.

Dan Morgan: Are there any questions that we did not answer?

David Winter: Ok, so one thing to recap about what we have talked about so far. We are going to post this webinar on the Safety Data Initiative website, which will include audio. We will also post a PDF version of all the slides with the first letters included in the paragraphs so it will be legible. What am I missing?

Dan Morgan: I think those are the big things. There's a question that just popped up about a data virtualization strategy. No, we have not implemented one, but we are thinking about how we might do that.

Jason Broehm: It looks like we have several attendees are typing questions so we'll pause for a moment. There's a question about an attendee list for the webinar for potential partnering purposes.

Dan Morgan: I don't think we can do that. Let us figure out how to do that.

David Winter: Once we post the webinar, they will be able to look at the participant list. The only thing you will not have is the people's email addresses and phone numbers. But I think that we may run into some limitations as far as posting people's emails and addresses or phone numbers.

Dan Morgan: Is the FAA included in this RFI? This is particularly focused on surface transportation safety. So, at the moment, FAA is not considered. But we are of course close friends with them and actually admire many of things they have been able to build, especially in terms of what a data exchange ecosystem really is, and how they've built across the federal government, the FFRDC community, and with the private sector, to build trust and drive the availability of more and more information to help them improve aviation safety. Yeah, super jealous. We wish we could do more of that on our side.

Dan Morgan: Is there a big data platform under consideration? Yup.

David Winter: Saw that coming.

Dan Morgan: It will be Hadoop-based. We do not need to get into which one it is.

Jason Broehm: It looks like we have a couple more questions potentially coming in.

Dan Morgan: All right. Can you leverage anything from the FAA? I think there's a ton we can leverage from the FAA. We think that they are looking at, we think we can leverage models for collaboration. We think we can leverage some of their safety management system approaches. We think that there's plenty of research we can work with as well. How they think about risk is really important too. I think some challenges that make the surface transportation system very different than the FAA system, at least as it currently is, includes the fact that, for the most part, the commercial aviation system would be considered a closed system whereas the surface transportation system is an open system. And the basic qualifications for operating a motor vehicle are pretty different from the basic qualifications for operating an airplane. And the amount of training that is required to operate a motor vehicle is significantly different than the amount of training that is required to operate an airplane, or aircraft if you prefer, whether manned or unmanned. So, I think there are similarities in terms of approaches, but certainly the nature of the system and the systemic approach to improving safety is very different.

Dan Morgan: Yes, we are planning to leverage the current visualization software we have. But we don't think there is one tool to rule them all. I think it is a package of tools. Right? Excluding just on a personal level, I have had conversations with chief data officers of other organizations and large pharmaceutical companies with 38 different technologies as part of their data platform, but that is what they need to discover drugs to save lives. So, they are cool with it. So, I think we want the right tools, whatever they are, the right tools and the best tools.

Dan Morgan: What kind of advanced analytics are currently being done and what are the most popular tools used in this space? I can't speak for, I'll point to some things that David pointed out too earlier. R and Python have been extremely popular for the stuff we have been doing under the Safety Data Initiative. Dealing with challenges of the Waze signal led us to pursuing random forest modeling to try to correct for some of the bias that exists in that crowdsource signal. And to allow us to have a machine to sort of figure out the Waze for us. More traditional approaches, at least in the beginning, on some of the rural safety stuff have been cart kinds of analyses, but then I think we can get to another place where we can get more predictive from those. So, you could name a Python library. I want to be clear that we are looking for industry's perspective on which things might be appropriate. If it is more appropriate for us to describe problem to get better feedback, I think you can say that to us. But I don't think we want to pour machine learning or unsupervised learning on top of a safety problem that isn't suited to that. Right? So, it's not about what kind of advanced analytics are in use so much as what kinds of advanced analytics are appropriate to the safety problem that we are trying to address.

Jason Broehm: I don't see additional questions in the chat pod. I will maybe give people another few seconds and then if we don't see anymore, we will end the webinar. Ok, well thank you all for participating today. We appreciate your time and consideration you will give to the RFI. We look forward to receiving thoughtful responses that will help us in our quest to move forward.

Dan Morgan: Thank you, everybody.

Jason Broehm: All right, and there was one final question: "Are the answers from these questions going to be collected and posted anywhere?" I think that we envision this webinar being the compilation of the questions, and feel free to refer back to that.

Jason Broehm: Thank you all for your time today.

David Winter: Thanks, everyone.

###