Transit Module 22: Harnessing Social Media and Big Data Technologies for Transit Business Intelligence

Ken Leonard: ITS Standards can make your life easier. Your procurements will go more smoothly and you'll encourage competition, but only if you know how to write them into your specifications and test them. This module is one in a series that covers practical applications for acquiring and testing standards-based ITS systems.

I am Ken Leonard, director of the ITS Joint Program Office for USDOT and I want to welcome you to our newly redesigned ITS standards training program of which this module is a part. We are pleased to be working with our partner, the Institute of Transportation Engineers, to deliver this new approach to training that combines web-based modules with instructor interaction to bring the latest in ITS learning to busy professionals like yourself.

This combined approach allows interested professionals to schedule training at your convenience, without the need to travel. After you complete this training, we hope that you will tell colleagues and customers about the latest ITS standards and encourage them to take advantage of the archived version of the webinars.

ITS Standards training is one of the first offerings of our updated Professional Capacity Training Program. Through the PCB program, we prepare professionals to adopt proven and emerging ITS technologies that will make surface transportation safer, smarter, and greener which improves livability for us all. You can find information on additional modules and training programs on our web site www.pcb.its.dot.gov.

Please help us make even more improvements to our training modules through the evaluation process. We look forward to hearing your comments. Thank you again for participating and we hope you find this module helpful.

Susan Bregman: Today we're going to talk about Module 22, which is Harnessing Social Media and Big Data Technologies for Transit Business Intelligence. I'm Susan Bregman. I'm the principal and founder of Oak Square Resources, and I've worked in the public-transportation industry for more than 30 years. I'm the author of *TCRP Synthesis 99*, which was about uses of social media in public transportation, and was also a researcher for the *TCRP Report 179*, use of web-based writer feedback to improve public transit services.

I was also co-editor and a contributor to the book, *Best Practices in Transportation Agency, Use of Social Media*. And for ten years, I was editor of the *Transit Wire*, a blog to share news about emerging transit technology. Now I'm going to ask Manny to introduce himself.

Manny Insignares: I'm Manny Insignares. I have 30 years program management and system engineering experience, and I've helped numerous transportation agencies worldwide to plan, develop, and deploy intelligent transportation systems, including transit. I will be covering Learning Objective Number 3 and learning Objective Number 4.

Susan Bregman: As a first step, I will define how transit providers use business intelligence. Business intelligence combines information from multiple sources to enable organizations to make data-driven decisions. For transit providers, business intelligence can help inform and support service planning, operations and investment decisions. Potential sources of data for business intelligence includes both qualitative and quantitative data sources and data that is generated by the agency, and data that is external to the agency.

So qualitative data may include customer surveys and panels, focus groups, and stakeholder interviews. And these are typically agency generated. Quantitative data may include information from automatic passenger counters, automated vehicle location data, general transit feed specification files, which show vehicle location for the real-time options and also route and schedule information, and electronic fare payment system datasets. And those are typically generated within an agency.

And finally, agencies may want to use external data sources, which could include social media posts, which we'll be talking about today, and also some of the files that you're already pretty familiar with, like census files and other public datasets.

Business intelligence can help agencies in a number of ways. It can help transit operators meet their mandated reporting requirements, provide greater transparency in reporting to internal and external audiences, provide input for planning, operations and capital investments, and support briefings for senior staff and boards of directors.

Right now what I want to do is give you a few examples of how agencies might choose to use Big Data when addressing various agency goals. So for example, an agency might want to improve customer satisfaction. To do that, they might conduct surveys and focus groups, possibly establish an online customer panel, and then analyze social media posts to better understand customer sentiment. To help an agency improve service reliability for bus operations, the agency might review internal data on on-time performance and travel time, and then also look at social media posts to help identify specific locations, just as an example, where bus routes might be prone to delay.

Susan Bregman: As another example, an agency might want to improve maintenance at its rail stations, and the agency might take actions like reviewing internal maintenance records, analyzing social media posts where customers might be flagging a broken fare-ticketing machine, for example. Sometimes agencies also take things a step farther by encouraging customers to report these kinds of issues via social media, via text messages, or sometimes through a mobile feedback application. Finally, agencies might want to find ways to improve transparency in performance reporting.

Here, an agency might develop key performance indicators, or KPIs, from available data sources like some of the ones we've already discussed, and also report those KPIs via an online performance dashboard, and performance dashboards is something that Manny's going to be talking about in a little bit. So now it's time to have a poll.

And the question is: which of the following is not a source of data for business intelligence? And the choices are: A) Automatic passenger counters, or APC; B) Social media posts; C) Electronic fare collection systems or EFCS; or D) None of the above. So now take a couple of minutes and choose the correct answer.

And the answer—the correct answer is: D) None of the above. All of the data sources listed can be used to support transit decision making, this includes automatic passenger counters, social media posts, and electronic fare collection systems data.

Moving on to Learning Objective 2, I am going to define social media platforms and their applications to public transportation. So social media. You probably all know what social media is, but just to make sure everyone's on the same page, social media platforms are web-based or mobile applications that encourage users to interact with one another in real-time. Many users also find social media provides an opportunity to influence, or at least try to influence other users. Social media, which is also called social networking, includes different types of applications. And I'm going to be talking about those in a couple of minutes.

The platforms are mostly owned by private companies with proprietary formats, and they're not consistently regulated. Social media posts can share information, and sometimes misinformation, so it's important to keep that in mind when doing an analysis of social media posts. And finally, social media is still evolving, and the platforms continue to change.

When I did the research for the TCRP Synthesis on Social Media Platforms, it was probably about 2010, and at that time, maybe three dozen agencies were using social media. Now just ten years later, I would say that almost every transit operator uses social media, at least in some ways. Some use it better than others, some use it more than others, but almost everybody is using social media these days. So there are different ways to categorize social media platforms based on their characteristics. And I'm going to talk about some of the ways to do that with the caveat that many platforms don't fit neatly into a single category, but have characteristics of several.

The social networks. These are the most widely known types of social media, and this is what people think about when you say "social media." Social networks, they allow people to connect with each other online, to share information, comments and media, and there are both personal and professional networks. And certainly you're all familiar with Facebook, with Twitter and LinkedIn, which is generally considered a professional network. The internet is well-suited for visual media, and media-sharing networks allow users to post photos and videos, including live videos, online.

And while most platforms, including Facebook and Twitter, allow customers to post photos and videos, the platforms that are considered media-sharing networks, like Instagram, like YouTube, like Vimeo; they are designed specifically for posting those types of media. Discussion forums are like they sound; they enable users to find, share and discuss their opinions about information. And Reddit is a well-known example, as is Quora.

With content curation platforms, users can collect and share content from multiple online sources. Pinterest and SlideShare are examples of this. Pinterest, in particular, is often used as creative inspiration for everything from wedding planning to recipes, as customers take inspiration from various applications across the internet. Customer Review Networks; you're probably familiar with these things, apps like Yelp and TripAdvisor. This is where users generate reviews and share opinions about goods and services. Finally, blogging and publishing networks. These are platforms for individuals or organizations to post commentary or news. They typically focus on a particular topic, and they're usually longer than other social media posts, and less time sensitive. Examples would include Tumblr, Medium, and Blogger.

So most transit operators use social media platforms like these for outbound communications, meaning the agency is, in effect, broadcasting the information to its users. And—or to an audience. And audiences may include riders or customers, stakeholders, the media, first responders, public officials, and community members. And agencies use social media for multiple purposes. And I'm going to talk about those in a minute. But first I want to also say that outbound communications, like the ones I'm going to talk about, typically are not used to support business intelligence activities. But it's important to understand these in the context of social media and seeing how business intelligence does use social media.

So the most common way that agencies use social media, I would say, is for service updates and alerts. For example, here the MBTA Commuter Rail in the Boston Metro Area is providing an alert on Twitter about a train that's running behind schedule. In Sound Transit, in the state of Washington, the agency is also using Twitter here to post an alert about an out-of-service elevator.

Social media is also very commonly used for emergency communications, and certainly that has become especially important during the COVID-19 pandemic. Agencies also can use social media to communicate during weather events, natural disasters, like earthquakes or hurricanes, and to share public safety information. These posts tend to be time sensitive, so Twitter and Facebook are good options for agencies.

Here's an example of a COVID-19 communication. This is from the New York City Subway, which is providing a service alert, and also a safety message to remind riders to wear a mask when they're traveling. The MBTA Transit Police uses Facebook to share public-safety messaging, and also to seek public feedback on persons of interest, as they're doing here. Agencies also use social media for marketing. And these are typically not time-sensitive posts, but they are designed more to generate good will for an agency.

Here, for example, the Los Angeles Metro, L.A. Metro, is using a photo from its archives of busses doing training in the Los Angeles River Basin. They're using Instagram, and it's not a time-sensitive Tweet—a post, it's really just there to generate goodwill for the agency and to engage people. As you can see, it's already—it accrued 3,700 likes.

Transit operators also use social media for customer service in real-time to address comments and complaints. Here's an example from SEPTA in Philadelphia where they're communicating with a rider who was complaining about a 45-minute commute that ended up taking three hours. It's another way that agencies can use to connect with their riders. Sometimes agencies use social media to solicit feedback from its customers. Here, Long Beach Transit is using Instagram to let people know that they are seeking feedback on an agency project.

And finally, agencies may use social media just for general announcements. They may be job listings, or press releases, and this is where social media posts can complement, but shouldn't necessarily replace, traditional agency communication channels. So, for example, the MTA in New York City is using Flickr to share information about its new fare payment media. And DART, Dallas Area Rapid Transit, uses Facebook to share press releases.

So those were examples of agency-generated social media. Now I want to talk about customer-generated social media. And I use "customer" broadly to refer to transit riders, stakeholders and other people who just have opinions about public transportation that they are sharing in the social space. These user-generated posts typically fall into three broad categories: questions, complaints, or compliments.

And this is often the source of information for data-mining activities, because this unfiltered feedback from the public, from customers, and agencies can analyze this information to answer internal questions. So, for example, customers may be asking questions of an agency generating a little back and forth here. The Toronto Transit Commission is communicating with a rider and providing information about a medical emergency in the field.

Customers also social media to complain about transit, and this is probably the most common type of customer communication. Here's an example where the MBTA is speaking with a rider about the status of an escalator at a Downtown station. Finally, some customers use social media to compliment transit agencies. Here's an example in Vancouver from a TransLink rider who takes the opportunity to thank the agency and its workers for working to get—to resolve an issue.

Crowdsourcing is another way the transit operators, among others, can solicit ideas and feedback on a specific topic. There are also mobile applications that create platforms for subscribers to share information with one another, and those are known as peer-to-peer communications. So examples of these include the transit app, which is a mobile application that complements its real-time data feeds with crowdsourced information about vehicle location. Pigeon is an app that Google developed for crowdsourced information about transportation. And Clever Commute is a mobile application that allows customers to share information about service on New Jersey Transit, Long Island Railroad, and Metro North Services.

So now it's time for another poll, and the question is: which of these is not a source of social media data for business intelligence? And the choices are: A) Agency marketing posts; B) Customer complaints; C) Customer questions; or D) Peer-to-peer communications. And now select your proper answer.

And the answer is: A) Agency marketing posts. Marketing social media posts can generate goodwill for an agency, but they are not used to inform data-driven decisions. Customer complaints can provide valuable data to an agency. Customer questions can also provide valuable data, as can peer-to-peer communications. So now I'm going to turn it over to Manny, and he will lead off with Learning Objective 3.

Manny Insignares: Thank you, Susan. We're going to go into Learning Objective 3: Define Big Data in relation to social media for transit. So we're going to talk a little bit about what is Big Data. We'll talk about the characteristics of Big Data, including variety, volume and velocity. And we will certainly go over those terms in detail. We'll review the sources of transit-related Big Data, including those that are internal to the agency, and talk about external data sources. We'll talk about the characteristics of social media datasets. And we'll get an overview of social media data standards that are emerging. Before I go into the topic of Big Data, I did want to say that much of what will be discussed and described in this Learning Objective, Learning Objective 3, and the next Learning Objective, Learning Objective 4, apply to what I'll characterize as small data, or data that isn't as connected.

This, what we're going to present here, can be used in a number of ways to solve basically data, data translation, and trying to do data analysis. So if you're a smaller agency, you don't have all this stuff in it, this still applies for what you're going to do. And it's valuable to you to get information and insights from your data.

So what is Big Data? What makes it big? Well, large volume of data, that is, some of it is structured and unstructured. Also, with using current technologies, it is often difficult to process the data that is timely and as needed. So these are the things that characterize it as Big Data. We use three Vs here—the three Vs are: variety, volume, and velocity—to describe these large datasets. So if we look at the picture here on the left-hand side, at the very top you have things like videos, unstructured text, photographs, you have spreadsheets, and perhaps some GIS file, some map information.

So one of the things that characterizes Big Data is that it comes from multiple sources. Could be within your agency, again, or it could be some is in your agency, some of it is with planning organizations, some of it may be from some other government source, perhaps the Census Bureau or wherever you might collect your demographic data, as well as private sources, if you're buying it. And the data comes in multiple formats. Text data we're familiar with. We look at photos and videos and these PDF files that you're familiar with coming up in your browser, database formats, CSV, which stands for comma-separated variables, and spreadsheet information.

There's a wide variety of formats for this information, so the data isn't in one format you can process, it's in a variety of different formats. Some of it is structured, and that means that you're able to look at it with different fields and structured in terms of what kind of data type it is; is it numeric data, is it textual only? And there's unstructured, which means things like photographs or videos or text, for example, in social media is unstructured. It's just typed in from whatever the person's situation is.

The next V stands for volume. And as you can see on the picture there, we have Big Data, like brontobytes, which means big in terms of volume, terabytes, and petabytes. These are new terms. They're new terms to me. But in the old days we had things like megabytes, and well, terabytes we have nowadays, but terabyte is 10-to-the-12 bytes, and petabytes is a little bit bigger, and brontobytes is 10 to the 27th power. But this is just the beginning and it certainly goes upwards from here.

The third V is velocity and describes the speed that's required to convert the inputs into the outputs that you need. And for streaming data, such as video, there's continuous conversion from the inputs to the outputs. Here are some examples of the three Vs as they apply to transit-related data. So on the left-hand column we have a Data Description, then we talk about the three Vs: the variety in the second column; the volume, referring to the storage, or how big the data is; and velocity, how frequently do you need to update it? How quickly do you have to convert some input into an output?

Let's take the example of vehicle location. If you make 100,000 trips per year, we're going to characterize this data as structured, and if you collect 50 bytes of data every five seconds, in other words, you know the position of the bus every five seconds, or of the train, then you're going to have to have 3.6 gigabytes of data per year. Your scheduled data—this example is from SEPTA, you have structured in the GTFS format, and this data is provided seasonally, which is fairly typical of transit agencies. That data's 21 megabytes.

Next example is video from 300 cameras. Video is the variety; it's its own format, and it's streamed, so whenever you take a picture, it gets transmitted right away. So it's always on in terms of how frequently do you need it, do you need to see the picture immediately. And 300 cameras' worth of video takes up about 1.2 terabytes of data. Then the last example is Geographic Information System files. These are those map files that you can read into your software to view a picture of the routes and the stops. This data is structured, and for NJT Bus, New Jersey Transit Bus, one of these files takes up 40 megabytes. And, again, it is provided seasonally. So that gives you an idea of how the three Vs can be applied when you're looking at how it corresponds to different kinds of data.

So we have now a look at transit-related data that comes from internal sources. Susan went through some great examples, but this includes ridership, surveys, focus groups, data from your passenger counts, vehicle location, your GTFS data, which is your schedule data, as well as the real-time data associated with it, which tends more to relate to vehicle movements, estimated time of arrivals, and your fare-payment data.

In your external data sources, since this is a course about social media, your social media posts, as well as Census Bureau files, traffic data. Perhaps you're gathering data from webpages and storing it, because you need it to do some kind of analysis. So this is a way to characterize what's internal to you versus what is external. And external will include social media data. Characteristics of social media datasets: Social media, moreover, is written in natural language. So it is unstructured text. People simply will write whatever they want into their little app, and communicate it however they feel like communicating. There is no need for it to be structured in any particular way. It is difficult to categorize, it is uncategorized, it is entirely up to whatever everyone wants to write. It is voluminous because of the sheer number of folks that are using social media and communicating it.

And social media datasets come in a wide variety of formats. As Susan was describing, things like Pinterest and some of the ones that have to do with picture and video sharing. You have formats of data that go with those, like JPG, or J-PEG, which is for photographs; and MP3, which is for audio; and MP4, which is for video. So there's a variety of formats from text, and a lot of them that have to do with photo and with video.

Social media data standardization is a challenge, but standards are emerging. As we said, social media is unstructured, may include natural texts, images, video, databases, as we have described before, and social media platforms are mostly owned by private for-profit entities. So they may use a proprietary format. Some social media have application-programming interfaces, called APIs, that allow third parties or users to download the data, but this is not true across all social media. So these are some of the challenges in trying to come up with a way to standardize essentially how you connect social media with other data, and social media with other social media datasets.

Nonetheless, there are a number of international efforts on big data standardization. We reference a report that is from ISO on Big Data. This report is published in 2014, and we're unaware of any updates. But at the time there was a scan—a worldwide scan across all standards organizations—to try to put in one place what activities were occurring related to bringing together different datasets, including social media. So we have two pages here that we describe—the first one talks—we have the ISO/IEC JTC Steering Committee Number 32 (ISO/IEC JTC 1/SC 32) on data management and interchange, including database languages, multi-media object management, metadata, and electronic business, or e-Business.

Steering Committee Number 38 talked about the standardization for Interoperable Distributed Application Platform and Services, including web services and service-oriented architectures and cloud computing. The ITU has a Cloud Computing for Big Data activity. And the W3C, the Worldwide Web Consortium, has an activity on web and semantic related standards for markup, structure, query, semantics and interchange. On Page 2 of this list, we have the Open Geospatial Consortium. Geospatial-related standards for specifying the structure or the query in process—processing of location-based data.

The Organization for the Advancement of Structured Information Standards is carrying on an activity related to information access and exchange. Transaction Processing Performance Council has developed benchmarks for Big Data systems. And the TM Forum is developing and enabling enterprises, service providers and suppliers to continuously transform data to succeed in the digital economy. As you can see, this covers a wide breadth of kinds of information, but it is not specific, and really provides you only an overview of what you might be able to do.

Now, the activity. And here's the question: Which of the below is not one of the three V characteristics of Big Data? Your answer choices are: A) Velocity; B) Viscosity; C) Variety; and D) Volume. We'll take a minute to enter your selection.

And the correct answer is: B) Viscosity. Viscosity is not one of the three Vs of Big Data, but a useful measure for assessing the quality of maple syrup and ketchup. A) Velocity, is not correct. Velocity is one of the three Vs, and refers to the speed required to convert input data into output data. C) is incorrect; it is one of the three Vs. Variety refers to the diversity and inconsistency in the structure and unstructured data that is present in Big Data. And D) Volume, is incorrect. It is one of the three Vs. Volume refers to the quantity of data and also to the growth rate of that data.

Our next Learning Objective, Learning Objective 4, we'll understand the process for applying Big Data analytics and social media to inform transit business intelligence. So we tried to put everything together and describe a process that you can go through to analyze your data. And then before we get back into this, just to reiterate, your data doesn't have to be Big Data, it can be smaller data, it can all be structured, it can all be within your organization, so you can still take advantage of what we're presenting here as a template that you can use to be able to process your data, so you can get better insights from that for business intelligence.

So we're basically going to go through a process here, and then we're going to talk about a couple of issues. And in the process, we'll talk about: data acquisition; data preparation; data analysis; with an emphasis on data science techniques; data presentation, basically looking at visualizations, dashboards; and then we'll talk about policy issues and technical issues. So the first step to doing analysis on the data is you have to go get your data. And this is an essential step, because this is the step that is needed so that you can do any of the other steps.

So what does it mean to acquire data? There are a number of ways of acquiring data. You may have recorders, or gathering information about natural events, for example, sensors. You may record human-made events. There could be some kind of data entry that's going on, and people are entering the data. It could be collection of data from social media. But however, it is that you're gathering the data, you have an opportunity to collect it so that you can do something with it to analyze it, and to gain some insights from it.

One of the key questions is: "What data do I have?" You can look at your internal sources and your external sources to look at the problem that you're trying to understand better. "What data do I need, but that I do not have? Do I have to do some data scraping? Do I have to use an application-programming interface? How much will this new data cost me? Are there sources of the data that's from the government, for example, that are free of cost? What are my storage requirements in terms of volume, keeping it secure, does it go in the cloud, do I have to work with my IT group so I can keep the data in-house? And where do I store my data?"

At this point, if we look at the diagram on the left, we see these data coming in and it's Big Data here and brontobytes, terabytes, and at the bottom it shows the Data Lake. We want to introduce this term because it's widely used to describe a repository for Big Data. Data Preparation: This is the step where you want to remove data that is incomplete, data that is incorrect, data that is out of range of your analysis. So you want to ask, "Do I have the right data? Is it the right granularity in terms of precision? Does it give me the coverage I need? What is the content of the data? What are the specifics that I need in terms of data content? Geographic region and data; do I need GPS files, GIS, Geographic Information System, data? What are my timeframes that I'm looking at?"

And there are challenges to mixing data, for example, if I'm collecting two datasets, and one of those datasets is given to me in five-minute increments, and I'm merging it with another dataset that gives me the data in one-hour increments, decide how to put those together by mixing and matching geographic information, and I have some that is based on a county, for example, demographic data, and I have some other data where I have stop locations, and some other data that is given by zip code, or some kind of zip-oriented tabulation data that many people in transportation use.

"How do I bring those together and harmonize those geographically? Are there standards available?" This is a step where you might want to consider mapping your data to a standard, and specifically writing down how I convert some dataset into the standardized format, and I can do this across my datasets, and if I use the standard, I have all my data harmonized, the standard helps me do that. That's one of the key reasons we have standards is to be able to allow you to bring together different datasets.

"What data scrubbing do I have to do? What filtering? Do I want to remove outliers? How do I handle missing data? How do I remove data that's out of range? And how do I handle null data?" So let's say I'm processing weather information, or I'm processing bus data, but let's say that that day in the weather station example maybe the power went out. Or in the bus example, maybe my AVL system stopped recording some information, for whatever reasons, and there are many reasons why this happens.

"What do I do with that gap where there are just a stream of nulls? There's no data." So these are things that have to be considered when you're trying to mix and match data, is that you have a complete dataset. You have to decide whether you're going to get rid of it or not, if you have missing data in your record sets. At this point, you may want to define rules for doing your sentiment analysis, or doing—developing a topic maps. What kinds of words are you looking for, or what is the importance of those words that you're going to use? And this may help you create the linkages between your disparate datasets, and especially if you're going to be doing some data mining that includes social media.

In the next step we do data analysis. With data analysis, we will interpret the relationships between data to gain some insights about the problem or the solution. Data analysis techniques include: data mining; data visualization; creation of topic maps; sentiment analysis; there's data similarity analysis; stochastic analysis, where you use statistics and probability to try to understand the data; and data correlation. We also have some techniques that we've used for a very long time in transportation, that revolve around our artificial intelligence and machine learning. Things like image processing, facial recognition, automated license plate recognition, and predictive analytics have been in use for a very long time.

Once you have gained your insights, you're ready to present the information. So data presentation is a process of using your results of analysis to provide some explanation or to make a claim about your data. Agency dashboards draw data from multiple sources to share KPIs, the Key Performance Indicators. So these are the things you care about the most in your operations: ridership, service performance, financials, customer satisfaction, looking at maintenance records and electronic fare payment, and there are a number of ways to show the data.

On the left-hand side, we show a couple of different—the very top one is a dashboard from MBTA; that we'll go into more detail in the next couple of slides, but you also have things like heat maps. And you have trend analysis, you have bar charts, and you can have just other kinds of charts and graphics to show, essentially, "How are you doing? What's the consistency of what you're looking at?"

So just to summarize what the Big Data steps are: We begin with data acquisition; you prepare your data, get rid of missing data, outliers, out of range. Conduct your data analysis to understand how the data go together. And lastly, put it all together for presenting the information to folks that may not have sort of the intimate knowledge or the detailed knowledge that you might have, but you still want them to come away with the insights, and there are graphs and graphical maps and charts that are easy to understand, and easy to describe your insights.

Here's an example of an MBTA Dashboard, and it's looking at service reliability, and you can see there that, at the very top on the left, you have average reliability. And you can see a trend, eighty percent in December, and then eighty percent the last seven days, seventy-nine percent the last thirty days. To me this looks very good. And then below it, there's a chart that you can see hovers right around the eighty percent mark, which is what you would expect, so you have very consistent reliability for MBTA on the C Line. So this supports the ability, and shows you an example of presenting the information in an easy to understand way, to basically get at the point that your data is highly consistent, at least for Line C.

The URL is in the Student Supplement, if you want to go look at it in more detail. This particular dashboard does not include any social media posts, or social media information. Another example is busstat.nyc. This is a project that was sponsored by the NYU Center for Urban

Science, and it was a capstone project of a Master's Program, sponsored by TransitCenter. And in this example, they were looking for some new ways to communicate along the lines of a dashboard, easy ways to understand travel time and new metrics, new KPIs for folks looking at transit data.

So on the left hand, we see this interesting chart that shows actual versus scheduled travel times, and then we have excess wait times, two minutes, the route lateness factors, you have a 39-percent chance that your route will be late, and then the average speed. Want to talk about a couple of issues, policy issues. Let's begin with protecting user privacy. The social media posts in many of the examples that were shown by Susan include specific people's names, or potentially their moniker, and certainly showed their face, and this may be something you have to take into consideration if you're going to use this data, that it is somebody's personal information; even if that person has decided that they want to give their information to the world, your agency may still have some policy about using that data.

Data security—how do you keep the information private? You don't know if that's an issue, but it could be. Again, what is the regulatory environment and what are the policies in your organization? There are organizations and agencies that do not allow anybody to look at social media while at work, so you may not even be able to get to the social media, or you're limited in some way, or you may have to get permission if you're going to do data mining that includes social media. Perhaps the most important thing, though, is you need to understand whether the social media data represents your customer base.

So if you have—your social media is for one demographic, but your service is for a different demographic, then social media may not apply, and you may not—you may be careful how you use it. You still may want to use it, but you may want to understand that restriction and try to prepare the data so you get the right data you need. Lastly, you want to analyze whether you need to take into account multiple languages. Most people will provide social media, and do social media in whatever their native language is. So if you're in a larger, or a potentially larger, metropolitan area, then you may have a diversity of languages being spoken while you—this may not be a consideration somewhere else. But it's something to consider as you look at how the policies are for your agency, and how it applies to social media.

And the technical issues—these are more technical considerations, or just things to think about. We introduced the term "data lake." There's another term that is used to describe partitions of a data lake, they're called "data ponds," that can have limited access. So if you want to restrict your data lake to a subset of users or have some security applied, then you would put that and make that a data pond. You may want to share your resources with another agency, you may want to be able to set up a data pond, or sharing data with them, and them sharing data with you.

Want to look at open source and open data tools. If you're looking down the road of open source and open data, you need to consider whether there's adequate technical support and security available when those tools are implemented. And lastly, you want to look at your resource requirements. Do you have the right skills, storage capacity, hardware, licensing, and whether you want to use in-house versus contracted staff for doing the work.

The activity—Question: which of the below is not a step described in Big Data processing? Your answer choices are: A) Data preparation; B) Data field Quantization; C) Data analysis; or D) Data acquisition. We'll give you a minute to enter your response.

B is the correct answer; data field quantization evaluates elements of the General Relativity Theory to prove gravity exists, and is the basis for the general rule that busses will roll instead of fly. Answer A) is incorrect; it is part of the processing of Big Data. Data preparation is the step of removing data that is incomplete, incorrect, and/or out of range from further analysis. Answer C) Data analysis, is incorrect. Data analysis is the interpretation of relationships between data to gain insights about a problem or solution. And D) is incorrect. Data presentation is the process of using the results of analysis to make a case or to provide an explanation about your data. And with that, I'm going to turn it over back to Susan to go over Learning Objective 5.

Susan Bregman: Before I get into the examples, I want to recap what Manny said about some of the special situations when working with social media data. Social media posts from transit customers, from stakeholders, and from others can be a valuable source of information. This is inbound communications, a valuable source of information of unfiltered feedback. However, there are some caveats with the analysis.

Social media posts use natural language—and I know Manny talked about this—and that requires special analytical techniques to create meaningful datasets. Posts also usually include usernames, and these are typically removed or scrubbed during the analysis to protect user privacy. As Manny just talked about a little bit, some transit agencies also restrict the use of social media by staff. Some just let the marketing folks use social media, or have specific employees who can use social media who are authorized to speak on behalf of the agency. So this may be a consideration when analyzing social media.

And then finally, social media users may not be representative of all transit customers. A small proportion of transit customers in some agencies use social media, especially in small agencies, rural agencies where customers may not have access to smartphones, and they may not be using that kind of technology to communicate. It may be a different situation in larger urban areas or agencies serving college campuses, for example.

That said, it's important to recognize that social media posts, and the analytics of those posts may only represent a small percentage of riders, and it's important for agencies to keep that in mind and maybe supplement that analysis of social media with other types of more traditional analysis, like rider surveys or focus groups. That said, I want to look at a few examples of agencies or researchers that are using social media or Big Data to help answer questions for the agency or solve problems. And I have examples in Chicago, San Diego, London, and the Minneapolis-Saint Paul Twin Cities.

First, I want to talk about the Chicago Transit Authority (CTA). This was one of the first examples of research on the topic of measuring customer sentiment by analyzing social media and a set of researchers at Purdue University collected and analyzed Tweets that mentioned CTA or the individual subway lines. Their goal was to use social media posts to measure customer satisfaction, and they developed a tool to quantify rider’s sentiment. In other words, they took Tweets and categorized them as positive or negative on a scale based on the content of the Tweet.

And this is very difficult, to be honest, because social media posts use natural language as we've said, but they also use sarcasm, which is very difficult to interpret. They use profanity, and it is a challenge to convert all that information into a dataset for analysis. But the researchers did this, and they were able to determine that something that I think many of us probably long suspected, that riders are much more likely to complain about a situation than to say something nice.

So here are some examples. This was on the morning of July 23, 2011. These are different charts showing the sentiment of Tweets, and the researchers found that negative Tweets spiked at about 9:00 a.m. So they had to determine why these Tweets were spiking. What happened at 9:00 a.m.? So they created a word cloud of the content of the Tweets around that time of day, and as you can see from this image the words "red," "flooded," and "blue" and "train" are some of the most commonly used words that morning. And as it turns out, there were delays on both the Red Line and the Blue Line because of flooding. So that was part of the analysis, where you could see the Tweets actually helped highlight a particular incident on the system.

In San Diego, the Metropolitan Transit System used Big Data to help combat fare evasion on their trolley system. The San Diego trolleys use a barrier-free honor system to collect fares and customers just tap a smartcard before entering on a fare validator on the platform. So MTS contracted with a consultant to analyze fare payment patterns and to determine whether additional enforcement was necessary and, if so, where?

So what they did was they analyzed several Big Datasets. They used GTFS to identify the location of the vehicles. They analyzed the information collected by the fare validators to analyze the smartcard taps. And they used automatic passenger counters to calculate boardings per station. And then the analysis correlated this information to determine locations where additional fare enforcement was required. Locations where ridership outstripped payment, in effect. So this is an example of how an agency used Big Data to solve a problem but this was not an agency using social media.

In London, some researchers tested an approach for analyzing geotagged social media posts. So Twitter, Tweets that were tagged with the location of the smartphone when that Tweet was made. So these were posts made in TFL, Transport for London underground stations. And the goal of the study was to see whether Tweets could be analyzed within a specific geographic area based on the content of the Tweet to see if our advertising could be optimized. So the Tweets were analyzed and categorized based on their geography and based on broad topics of information like sports and entertainment. And this information was intended to provide guidance for advertising campaigns that were targeted to different stations. Right now Twitter no longer supports this level of detailed geotagging, so it may not be possible to use social media for this kind of analysis in the future.

Another example is Metro Transit in the Twin Cities of Minneapolis-Saint Paul. They have a Strategic Initiatives Department, and that department, people in that department draw upon data from multiple sources to support data driven decision making for the agency. For example, how to allocate resources for bus shelters and amenities; how to improve on-time performance, or how to design the transit network to best meet customer needs. So to better determine where to locate and improve amenities at bus shelters, agency staffers looked at multiple data sources. They looked at a customer survey, they looked at their list of facilities to identify where the shelters were. They looked at ridership by stop, and they looked at the demographics of the areas around each stop, because they wanted to develop equity-focused measures to make sure that bus shelters were upgraded in areas where customers really could benefit from that.

Again, this is an example of using Big Data, but it was not an example of using social media to inform this decision. So now it's time for our final activity of the presentation. Based on these examples, analyzing social media data helps inform agency decisions about which of the following? And the choices are: where to upgrade bus shelters; how to understand customer sentiment; where to add fare enforcement; or how to report non-fare revenues. So take a minute to answer, and we'll look at the correct answer soon.

And the correct answer is: B) how to understand customer sentiment. Researchers analyze social media posts to assess CTA customer sentiment. The agency, answer A, where to upgrade bus shelters is incorrect because the agency did not consider social media posts in its analysis. Similarly, the San Diego agency did not use social media to solve its problem. And none of the examples focused on non-fare revenues. Social media was not a source of this data.

So to recap what we talked about today, first we learned how transit operators can use business intelligence tools to make data-driven decisions. We saw examples of agency-generated and customer-generated social media posts. We learned about potential sources of Big Data for use in transportation analysis. We reviewed the process for applying Big Data analytics to social media to inform transit business intelligence, and we reviewed examples of using Big Data to support business intelligence.

Thank you everyone for completing this module, and please use the feedback link below to provide us with your thoughts and comments about the value of the training. Thank you from both me and Manny.