Roundtable on Data Management: Preparing for Machine Learning

Summary Report

Cambridge, Mass.

October 15–16, 2019

Moderated by:

Dr. Sergio Caballero
Research Scientist
MIT Center for Transportation & Logistics

MIT Center for
Transportation &
Logistics

Table of Contents

Executive Summary

On October 15–16, 2019, MIT's Center for Transportation and Logistics (CTL) hosted representatives from 18 organizations for a roundtable on data management for machine learning in the supply chain. Six sessions focused on: 1) the importance of data, 2) managing organizational transformation, 3) organizational data governance, 4) data collection, 5) data wrangling, and 6) data visualization. Each session began with a short presentation followed by wide-ranging discussions. To encourage candor, statements have not been attributed to named organizations.

The first session focused on how the rapidly declining costs of data storage, bandwidth, and computation foster greater supply and demand for data, and promote the development of fundamentally new data processing methods such as machine learning. However, getting value from data requires transforming organizations to have better data management and data governance.

The next two sessions covered organizational issues of transformation and governance. Although data management seems to belong to IT, roundtable participants suggested that business stakeholders should own the data because they know where it comes from, what it means, who needs to see it, and how often it needs updating. Harmonizing both data and KPIs—creating one source of the truth—enables greater and more consistent use of data but requires either stakeholder alignment or a top-down mandate. Firms represented at the event were prioritizing digital projects through use cases and strategic imperatives.

The final three sessions covered the data supply chain within the organization. First, data collection gathers diverse types of data from diverse sources as needed for various business decisions. Second, data wrangling cleanses, harmonizes and processes data for business users. Third, visualization requires user-centered design of dashboards and graphics to support different types of users: executives, operational managers, and front-line people. The goal is timely and actionable information that does not overwhelm the user. Finally, pilot projects in these areas must be converted to robust production systems to avoid technical debt.

Many of the discussions highlighted the key role of people, especially data scientists and data stewards. Hiring and retaining these essential specialists is challenging but can be achieved through tactics such as providing incentives and formal career paths for individuals. Finally, change management at all levels drives support for and use of data-driven decision-making processes.

Most companies were in the early stages of digital transformation—much work remains for the future. Participants visited CTL's Computational and Visual Education (CAVE) Lab to see how advanced visualization can be used on complex supply chain challenges.

After almost two days of presentations, deep dives, and discussions, participants cited key takeaways and planned next steps. Participants planned a range of future applications including end-to-end integration, a shift from descriptive to more predictive and prescriptive analytics, and more sophisticated statistical models. Future roundtables may delve into issues such as the cross-functional challenges of data management and the need for more effective ways to visualize data.

Importance of Data

The roundtable began with a session on the increasing importance of data. Steep declines in the costs of computing, bandwidth, and storage have enabled the exponential growth of volumes of data and its use in managing companies and their supply chains. Growing ponds, lakes, and oceans of data are fuel for machine learning. As that fuel becomes ever cheaper and more plentiful, so computational engines become cheaper and more powerful. However, harnessing data requires an understanding of how data can be used and then developing the organizational processes and technological infrastructure as well as training people to use it.

How Data Fuels Machine Learning

The way organizations use data is changing. In the past, creating software for computers required explicit step-by-step programming that codified existing knowledge. Subject matter experts and programmers worked together to design and code explicit rules and algorithms. However, this strategy of application development suffered from Polanyi's paradox: People know much more than they can explain. For example, people can easily recognize faces, drive vehicles, and understand customers' verbal requests, but they can't write the step-by-step algorithms by which they perform these basic everyday activities.

Machine learning (ML) is a fundamentally different, data-driven approach to creating applications. ML takes large volumes of data that exemplify the specific desired patterns of the application and feeds them into pattern-finding algorithms such as least squares, random forests, or neural networks that can learn the rules and algorithms from the data. Machine learning, especially deep learning on large volumes of data, can produce significantly more accurate results than can traditional statistical algorithms.

Vast amounts of data and very fast computers are essential to successful ML applications, and both have become available in recent years. For example, the data might consist of thousands or millions of warehouse images that depict subjects such as a pallet, a shelf, a box, a forklift, a person, or a loading dock. When that data is used to train a neural network, the result is a computer program that can recognize these objects when given new images, such as those coming from the cameras of an autonomous forklift or robot. The approach is much like training a child by pointing at different things in the world and naming them.

While traditional learning approaches depend on gathering lots of expertise to program a computer, machine learning depends on gathering lots of data to train a computer. That data can come from database tables, sensors, scanners, text, images, audio, video, and so on. However, a major challenge in moving from programming to ML comes from managing all that data. To be effective in learning applications, the data needs to be clean and well organized. Moreover, for data to be effective in an ongoing operation such as a supply chain, the data needs to be collected in an ongoing and organized way. Consequently, companies need data management and governance of that data.

Business Applications of Data

Three broad categories of business applications depend on data. The first is operational efficiency, in which data enables visibility and resource optimization, as well as process improvement, to reduce costs and increase quality. Second, data can be used to improve the customer experience through segmentation, targeting, and providing visibility throughout the order-to-delivery cycle. Third, data enables new business models. For example, a retailer described its new "click-and-collect" service that merges previously separate online and physical retail business models. The service depends on timely data to drive 3,500 forecasting models for the combined demand in both channels, which ensures adequate in-store inventories to support good customer service for all shoppers.

A survey sent to participants prior to the roundtable asked them, "What are the business needs that require data as an input?" Nearly all companies (93%) answered "to inform decision making" and "to gain visibility into our operations (operational dashboards)." More than two-thirds of companies (71%) needed data "to gain understanding about the business (reporting)," while nearly two-thirds (64%) said "to support hypothesis/data," and 57% indicated "to identify new business opportunities." A more modest number of companies (43%) used data "to reduce risk," and 14% had data needs not listed in the poll questions.

Various participants described what they were doing for some of the less-popular options, such as hypothesis testing and risk management. For example, a retailer does A-B testing for proposed changes to store design, product assortment, and new products or services. The retailer implements the idea at a few stores and then measures sales patterns against similar, unchanged stores to see if the change has the hypothesized effect. Other participants described how they use data to manage risks, such as supplier material risks, risks of missed or delayed shipments, cold chain shipment risks and payment risks with customers. In each case, historical data was gathered on the measurable conditions (e.g., descriptors of each customer) and subsequent risk outcomes (e.g., non-payment by each customer) were fed into a risk model and prediction engine.

Trends: More Volumes, Velocities, and Varieties of Data

Current data volumes are huge. As of 2018, the average day saw 3.5 billion Google searches, 700 million Facebook comments, 656 million Tweets, 24 million Walmart customer transactions, and 20 million UPS parcel deliveries. The volume, variety, and velocity of data continues to grow. The Internet-of-Things (IoT) will bring many possible kinds of sensors attached to many kinds of objects, machines, locations, or people. GPS and location services will bring streams of data on the movements of vehicles, people, and goods wherever they happen to be. Regulatory requirements such as ELDs (Electronic Logging Devices) in trucking and tighter product tracking in pharmaceuticals also increase the volume of supply and demand for data.

By 2020, 21 billion devices will be connected online, including billions of smartphones, smart speakers, and wearables that each have suites of motion, audio, video, and GPS sensing devices. Data velocity will increase as more devices stream their sensor, event, and activity data to the cloud. The increasing velocity will help satisfy customers' and employees' increasing expectations for real-time notifications and live data on their mobile devices.

The declining costs of bandwidth enable the transfer of more data more frequently—both from sensors and to users. The declining costs of data storage permit the accumulation of more data for analytics and machine learning. The declining costs of computation enable more sophisticated analytics, more frequent updates of complex models, and more real-time analyses and visualizations. The combined impact of falling costs is a rapidly growing potential for using data in business that is only bounded by an organization's ability to collect, wrangle, and visualize the meaning contained in all that data. However, realizing these capabilities is not easy.

Top Challenges

The poll of the roundtable participants also surveyed the data-related challenges that participant companies were facing. The leading challenges were data quality and data reliability, each cited by 64% of attendees. Next came data availability with 57%. For 43% of participants, leveraging existing data is challenging, while 36% of the companies were experiencing data usability issues, and 29% were experiencing other unlisted challenges. However, subsequent discussions during the roundtable revealed that the hardest problems weren't the technological ones; they were the organizational issues. Thus, many of the discussions during the two days delved into how companies were managing processes and people to change their organizations. In turn, these efforts in governance and change management could address many of the top data challenges related to data quality, reliability, and availability.

Journey of the Organization

Sessions two and three of the roundtable discussed organizational issues: managing transformation through use-case-based prioritization and data governance. Both sessions began with a short presentation by a technology manufacturing firm that has done extensive work in this area and hosted one of the previous roundtables. The presenter talked about a range of key decisions and management processes regarding data architecture, data ownership, data harmonization, data initiative prioritization, and governance policies. In the discussions that followed, participants shared their experiences, successes and challenges on these issues.

Most companies were in the early stages of these kinds of efforts. When surveyed on "Where is your company in your digital transformation journey?" no company considered itself "advanced" and only 14% of companies indicated they were "somewhat advanced." Nearly two-thirds (64%) were in the "developing" stage, while 7% were in the "early pilot" phase, and 14% were "thinking about" their transformation journey. No company said it was in a "not started, not thinking about it" phase.

At the roundtable, many companies admitted they were in the early innings or first miles of a marathon. This included embryonic efforts on data governance, establishing nomenclature, defining one source of the truth, and debating the choice of metrics. Some companies alluded to being in the early stages of a succession of levels of data use, which start with describing the current condition, then predicting future events, and finally prescribing actions to optimize the future. One company shared that it had tried to create an enterprise-wide data lake back in 2000 but failed. Another company said that early efforts in this area had been like random Brownian motion.

Overall, the presenter for the two organization-focused sessions said there were four critical success factors for data governance: 1) selecting the right organization model, 2) embedding data governance into the existing drive for organizational performance improvement, 3) establishing data governance metrics, and 4) automating and scaling the data governance process. The company aligned its efforts with five program-level processes: 1) setting transformation priorities, 2) planning the scope and items of data for release, 3) defining a backlog of efforts through user stories, 4) managing the impact on the architecture, and 5) planning for sprints in their agile development process.

Building the Right Data Architecture

In the kickoff presentation for the organization transformation session, the presenter discussed alternatives for how the organization's data might be organized. Data architecture choices spanned a decentralization-centralization spectrum. At the one extreme, a centralized architecture enables IT to completely harmonize all data, bring all the data into one harmonized data lake, and then hope that everyone uses the resource. Centralized architecture is a great approach if the organization can drive top-down development and enforce tool migration. However, it may fail if one size does not fit all. At the other extreme, a localized architecture lets each subgroup optimize specialized data definitions tailored to its context. However, localized development creates many data ponds, an approach which can fail to deliver ROI because it can't get to scale.

A technology manufacturing firm chose a federated or mixed architecture to combine both extremes and to fit the company's large scale and different business units. Data governance and IT are centralized in the enterprise, create the platform that everyone uses, and have some investment in data science. The individual data science teams work with the business units to tailor the solutions to the local context. It's important for everyone to work together to provide value to the individual business and to leverage the shared platform and data.

A retailer described how its data architecture was organized along the flows from suppliers to customers. The company currently operates three data warehouses: one for the supply-side, one for retail items that merchants use to create UPC codes and track merchandise, and one for customers who sign up for a loyalty program. However, as the company has moved toward integrating e-commerce into physical retailing, it is experiencing the need to harmonize and connect all these data warehouses together. This effort will first align the item data and customer data for e-commerce retail and then expand from there.

Choosing Data Owners and Data Managers: Business or IT?

In response to the poll question, “Who in your organization is responsible for data strategy and results?” nearly two-thirds (64%) of attendees said that there was no single point of accountability. In 21% of companies, the responsibility was held by a Chief Data Officer, in 7% by the Chief Information Officer and in 7% by another C-level executive. No company in the survey had a Chief Analytics Officer or a Chief Data Scientist. Similarly, there was no universal way to manage the data. The survey also asked, “Which department manages the data in your company?” Half of the respondents (50%) said data was managed by IT, a Center of Excellence managed the data in almost a third of companies (29%), and only 7% gave the task to operations. An entity not listed in the questionnaire managed the data for 14% of participants.

The group debated the options for ownership of data. On one hand, IT clearly provides the technological infrastructure to move, store, process, and deliver data. On the other hand, IT does not know how to define business metrics or the business meaning of the incoming data, which data is needed for which metrics, and which users need different data products. Thus, business units need to own and govern the data, even if IT is acting like a contract manufacturer or public warehouse in managing it. One firm noted that it did not realize value from its data until the supply chain drove transformation and told IT what to do. “If your business units don’t take ownership of data curation and governance, there’s an order of magnitude difference in accuracy and ability to use your data,” the manufacturer said.

The second related issue is where to place the data scientists and their teams within the organization. At two firms, the data science department was not housed on either the business or IT side but was part of a central service set up as a Center of Excellence (COE). Several participants from a range of product and service companies reported split or matrix arrangements between the functional and technical sides of the organization. Including members of the data science department in other functional teams helped these technical specialists hear users’ day-to-day issues and see how products were used in daily workflows. “You have to have skin in the game to define a dashboard’s integrated market view—it has to live in the functional team,” said one apparel company.

Harmonization: The Journey to One Source of the Truth

A global technology manufacturer explained how organizational layers of data, platforms, and applications had become fragmented. Inconsistent definitions and loosely connected data operations created scaling issues because “my data and your data can’t talk to each other.” Thus, the company wanted to take a step back, harmonize the data to a canonical model, and manage the data as a corporate asset. The core rationale for harmonizing was simple visibility: “We wanted to digitize parts of our processes —that’s the easy win.”

A key part of getting to one source of the truth is agreeing on what the truth is. For instance, how exactly are last week’s sales measured? What constitutes a sale and exactly when does “last week” start and end? Some industries have a standardized data model, but many do not. The exact definition of the standard, however, is less important than using some unified standard. As one participant put it, “As long as everyone agrees to row in the same direction, it’s good.” Ultimately, the different functions in the company that use sales data must find some alignment and standardization—a process that can take time or mandates from leadership. In the meantime, one company was focusing on local harmonization of the supply chain—“cleaning our own house first”—with the intent to expand the effort across the organization. The hope was that if the supply chain group takes the lead and demonstrates the benefits of harmonization, it can drive broader adoption across the firm.

Agreeing on the definitions of key performance indicators (KPIs) can be a challenge for governance. Because KPIs are often used to measure people’s job performance, people have very strong opinions about them. Gaining alignment can be a tough conversation, even in cases of erroneous math. One company found a logical error in how discounts were calculated from POS data. It took them months of arguing with the leadership to get agreement on a correction. To solve such situations, some companies use top-down approaches to force better data management while others had to bring all the stakeholders in the room to get alignment.

KPIs also apply to the governance process itself. A participant from a technology company summarized the three levels of metrics his company uses to assess the success of data governance. The first measures are operational utilization defined by the numbers of data assets, business functions, and business metrics that are using harmonized data. The second covers data quality metrics such as conformity, consistency, availability, and quality of content. Finally, the third level measures organizational impact as estimated from surveys and assessments of improvements on business metrics and goals. The company is also working on a data quality monitoring platform.

Another key element of harmonization is master data management. Unfortunately, in many cases, an organization's different functions, geographies, business units, projects, suppliers, customers, and third-party data providers often use different identifiers or have inconsistent terminology for identical types of items. Gaining an accurate picture of the quantities and activities associated with each type of item means harmonizing the data across the different synonymous identifiers and terms.

Prioritization of Projects with Use Cases, Strategic Imperatives (and Politics)

Growing volumes and varieties of available data—along with the growing libraries of algorithms and possible business metrics—create virtually unlimited possibilities for data science projects vying for the limited resources of the organization and the limited attention of the firm's data scientists. Companies can use data to fix almost anything, but they can't fix everything all at once.

A technology manufacturer suggested project prioritization based on use cases created around user experiences within a story for the future data-driven organization. The company looked at what was most important in the supply chain and built "day-in-the-life" user experience scenarios. This story had five experiential elements featuring different supply chain people: 1) making the right commitment to customers, 2) getting the right parts in place at the right time, 3) keeping the customer informed if something went wrong, 4) managing supplier ecosystem, and 5) digitizing the work to drive performance. These five use cases then drove the project efforts.

An apparel company initially prioritized projects based on who screamed the loudest. They wanted customers for the data science teams and some quick, visible wins but then stopped to consider whether those efforts were actually the most important to pursue. That led them to prioritize efforts around two strategic imperatives of the entire organization: product line diversification and getting closer to the customer.

Of course, participants admitted that politics still plays a role. Sometimes the biggest hammer in the room determines what gets nailed down first. A company might expect solid use cases during prioritization but accept the squishier use case of a top executive. Two companies argued that part of a leader's role was to define how to prioritize and what to prioritize.

Governance Policies: Security, Privacy, and Metrics

Data governance also includes policies related to data collection and handling. During the two days, participants discussed three categories of policies. The first helped ensure the data which should be collected gets collected in an accurate manner. The second category helped ensure that data that should not be collected doesn't get collected. Finally, the third category helped ensure that the data that was collected was not mishandled.

An estimated 70–80% of data issues stem from human error and, in turn, inaccurate data can cause a 25% increase in labor costs. A technology manufacturer said that strong data governance helps ensure data is correct at the source. However, that accuracy requires compliance with data entry policies, which some companies said could be a struggle. That was especially true of distant suppliers in Asia. A technology manufacturer said performance contracts can help. For example, a third-party logistics provider (3PL) measures compliance (data entry and errors) of shippers and carriers and has incentives or penalties to motivate the correct behavior. A pharmaceutical maker noted that some compliance problems arise from obstacles to entering data, such as the physical challenge of keying in data on the factory floor while wearing work gloves. In short, participants emphasized that compliance should be made as easy as possible.

Second, a related data governance policy concerned what data should not be collected. This issue involves PII (personally identifiable information) on consumers. This issue arises in the context of both consumers' sensitivity about privacy issues and regulations such as GDPR (the EU's General Data Protection Regulation) and CCPA (the California Consumer Privacy Act). Because collecting data on consumers has legal and reputational risks, at least one company had very restrictive governance policies for PII that it calls "privacy by design." The company only collects PPI data if there is a very clear use case that makes sense and that has value for both the organization and the customer. Other participants, however, wondered about the lost opportunities if companies did not collect data on their customers or created too many obstacles to data use. In response, the company with the restrictive policies felt that transparency is important, and if it is not possible to make the customer feel good about how their data was going to be used, the company did not want to collect it.

Third, data governance policy also includes information security—ensuring that sensitive data won't be lost or stolen. This task includes the organization's proprietary data was well as sensitive data from suppliers, customers, and information vendors. Two companies mentioned the information security strategy of prohibiting data downloads. "No data on laptops!" said one of the enterprises. Instead, users' applications and dashboards access live data and metrics as needed. As a side benefit, only using live data from the managed platform helps prevent the creation and circulation of uncontrolled derivative copies of the organization's data. Such a policy ensures that everyone is using the same single source of truth and that the source is secure.

Project Challenge: Daily Imperatives vs. Long-Term Improvements

One general frustration between business and data science teams was the mismatch between everyday operational time scales and data science project time scales. Executives and operations managers need data and answers today, but a thorough data science project can take two to 12 months to implement. Broader enterprise efforts can take even longer. Worse, the long process of gathering data, cleaning it, and developing the models means that data scientists may have nothing to show for their work until quite late in the project.

Several companies recommended a strategy of generating small wins from easier projects before tackling bigger ones. The goal was to give the organization some confidence in the data science team and create buy-in. A second approach was to split the team's time, allocating some to short-term projects that fulfill the immediate needs of the business while dedicating some time every week to long-term efforts, such as cleaning data that will ultimately deliver greater strategic value. Regardless of the approach, the data science team needed to do its own public relations to keep the organization interested in (and willing to fund) the long-term vision.

A related challenge arises from the natural tendency to seek expedient solutions rather than invest the time and money to carry out projects the right way. That is, a quick win may show the merits of data science, machine learning, and data-based decision making. However, a quick-win implementation might only be a prototype that requires further development, thereby creating the problem of "technical debt." That is, fragile, incomplete, and non-scalable short-term solutions that eventually need to be revisited and create interim costs and quality issues along the way. A product manager said that his company has to carefully plan its project sprints to ensure adequate testing.

Managing Organizational Change

Throughout the day, participants talked about how change management reared its stubborn head as an obstacle to the adoption of data-driven business strategies and systems. In some cases, users reject an initiative. A maker of business products described how the company created a great tool to calculate the profitability of its business customers. But when the company's salespeople were shown the tool, they saw it as an attack on unprofitable customers. Although that was never the intent of the tool, salespeople perceived it that way. The developers realized that part of change management includes working with or convincing users of the good intentions behind such an effort.

In other cases, leadership can be the obstacle. For example, some companies noted that the business side is reluctant to own data. A typical perspective is that "data is an IT thing," or business users don't want the added responsibilities or costs. Showing leaders why they need to be responsible for business data in order to derive business value is a change management issue. One company invited an industry consultant to convince its people why the business needed to take ownership of data. Another company created an aligned set of KPIs in one dashboard, showed great business results, and proved the value of better data management to skeptical leaders.

In many cases, change management is a matter of marketing the new system to users and leadership. Use cases, stories, and small wins can help everyone appreciate how the proposed system can make their jobs easier and more productive on the metrics that matter. In other situations, though, participants said the only way to drive change is through a top-down push.

Finally, part of the change management process also implies assessing what changes will be needed and if the change is wise. For example, one company is considering streaming its SAP data into Snowflake instead of SAP HANA. However, many in the company have built native tools in HANA and are used to that system, so the proposed switch would require added resources for replacing those tools and retraining. Companies should study whether the benefits of a proposed change exceed the costs of that change.

Differences Among Organizations

Different companies manage data in different ways, and this was reflected in the various data management approaches used by participant enterprises. Many roundtable participants represented large, established organizations with decades of accumulated legacy systems, including disparate systems from mergers and acquisitions. For them, change was hard, especially if they faced the "curse of success" that made others in the organization see no reason for change. Larger organizations needed strong top-down leadership to drive change, especially changes such as harmonization that require greater cross-organization coordination and standards.

A few participants from smaller companies or subsidiaries did not have large teams of data scientists and technical specialists. In these organizations, employees tend to wear multiple hats. One of the smaller companies asked the group about the advisability of outsourcing data science projects. Outsourcing was possible, came the reply, but such projects needed to be carefully scoped to stay in budget and carefully structured so that the knowledge created by the contractor would be transferred and retained by the firm. This structuring included defining the deliverables and perhaps hiring the contractor full time if the project demonstrated enough value. Thus, outsourcing should be used to build capacity rather than deliver a black-box system that no one in the organization understands. One participant noted than interns and graduate students could be a very beneficial, low-cost resource.

Companies that were fast-growing digital natives had the advantage of innate familiarity with technology but experience the challenge of how to scale during growth. Small startups can run on tribal knowledge and loose bottom-up processes that "fail fast" to solve problems and add functionality. When a company is small, it can onboard a new employee quickly. As a company grows, however, and thousands of people are adding columns and data types, the result can be a deep hole of technical debt. That cleanup and coordination requires people and support from the top. The "fail-fast" ethos of a startup must morph into a "bend-but-don't-break" motto of a larger, more interdependent firm.

Journey of the Data: Collection, Wrangling, and Visualization

The next three roundtable sessions delved into data collection, data wrangling, and data visualization. This was essentially a source-make-deliver sequence of activities for a supply chain for data. In this supply chain view, some partners and functions might be suppliers of raw data (e.g., retail POS, transportation GPS), some functions might be consumers of data (e.g., finance, inventory managers, customer service), and some must "manufacture" the right visualizations from the collected data. Presentations and discussions covered how different companies were handling each of these three stages.

Types of Data and Applications

When polled about the types of data they were collecting, every participant confirmed that they were collecting structured data (tables, records), while 57% were also collecting machine-generated data (e.g., from sensor, RFID, devices). About 50% of respondents were using web logs and clickstreams, and the same proportion of respondents were collecting event data. Forty-three percent were using unstructured data such as audio, images, and video, and the same number of participants collected social media data. A final 21% of companies were using spatial or GPS data. The participants shared stories about how they were using some of the lesser-used data sources.

A number of participants had applications for using data derived from cameras and computer vision. One company used video cameras to monitor conveyor belts for flow and to automatically detect jams—a task typically done by people. Two companies cited efforts to use computer vision in grocery stores to monitor planogram accuracy and stockouts. One company's pilot-stage introduction of robots to store aisles yields new sources of timely data from retail environments and enables new applications. A fourth computer vision application uses OCR (optical character recognition) technology to scan paper documents such as contracts and converts the paperwork into searchable text and for subsequent machine learning initiatives.

Participants identified free-form text as an important type of unstructured data about suppliers and customers. A 3PL is converting unstructured emails from shippers and carriers into more structured forms that can then trigger automated replies or processing. A manufacturer searches and analyzes website text about the industry and suppliers to monitor the health of the supply base and find new prospective suppliers. Another category of free-form text was social media data that's typically processed for sentiment analysis. A 3PL was analyzing reviews of shippers' facilities by truckers, and an apparel maker was capturing social media data to study applications related to new product introductions. These kinds of applications typically use NLP (Natural Language Processing), which is a type of machine learning designed to glean meaning from human speech and writing.

Companies reported using other kinds of data such as Excel files, weblogs, and spatial data. An apparel company was analyzing Excel files to extract BOM (bill of materials) data and estimate the company's environmental footprint. Web logs were being used to improve inventory, customer service, and transportation supply and demand. Vehicle GPS and telematics data was being used for visibility, descriptive analytics, and predictive analytics although companies said that this type of data is messy.

Data Collection

An apparel maker opened the data collection session by describing how its Analytics Center of Excellence (COE) collects data for the organization's data lake. The organization focuses on collecting only the data it needs. These needs are driven by strategic priorities, associated use cases, and the data necessary to support the relevant KPIs. The company uses an intake process for deciding to collect each new data element. COE members meet with stakeholders to get the answers to key questions, such as: 1) What is the data? 2) What is the business value of the data? 3) Who is the sponsor of the data? and 4) How is it going to be used? Communication has to be a full-circle partnership with the business.

Overall, the company ingests both internal and external data. Internal enterprise data comes from operations, sourcing, manufacturing, logistics, retail, product development, merchandising, and finance—every part of the organization is a stakeholder. Internal data collection taps into many different enterprise software systems including SAP ERP, an e-commerce platform, Salesforce, and many scattered pockets of data found in Excel spreadsheets. The company also collects diverse external data from consumers (email, birthday, age, etc.), licensed distributors around the world, retailers (point-of-sale), and supply chain data from partner factories. External data can come with restrictions such as privacy regulations (consumer data) or contractual non-disclosure agreements (supply chain partner data). External data collection also requires the cooperation of the external provider.

The company said it has automated much of its data collection. All the SAP transactional data automatically goes into SAP HANA and into the data lake. Similarly, the collection of POS and clickstream data is automated. However, in some organizations, data generated by activities such as planning and product life cycle management reside in Excel spreadsheets. These spreadsheets are collected through so-called "user-managed data ingestion" in which users must upload Excel files into a folder and AWS Lambda ingests the file into S3.

Data Wrangling

In general, data is messy, owing to shortcomings such as missing values, erroneous outliers, idiosyncratic identifiers, and nonstandard formats. The data from each source needs to be parsed and corrected, which means developing code to do those tasks. A poll of roundtable members found that data preparation took a very high percentage of total project time. Nearly one-third of participants (29%) spent 61–80% of project time on data cleaning. For 36% percent of participants, data cleaning consumed 41–60% of project time, while nearly a third spent 21–40% of project time on this task. Only 7% of respondents reported spending 0–20% of their project time on cleaning up data.

With data typically being siloed and needing to be combined from multiple disparate sources, a participant poll asked: "How often do you need to integrate data from multiple sources?" Over two-thirds (71%) admitted it was "very frequently." For example, an online retailer noted that it has five different sources of data for yard arrivals. Ironically, technology has made the situation worse. One retailer said that in 2000, the company employed an enterprise data warehouse to access all its data. Now, there is no such platform because data comes in many different forms and to access it requires new code to be written. Overall, participants said cleaning data is 80% of the work on any data project and that "there is no silver bullet for data cleaning."

Two companies described how they use a multi-layer architecture to wrangle messy raw data. In the system, raw data accumulates in a landing zone and is then subsequently processed. One company labeled the layers: raw data, standardized data, and conforming data. Another company created a cleansed layer and then a curated layer like a canonical layer. Generally, the final level of data is the official one source of the truth that is sent to users in business teams. Raw and partially processed data is only accessible to those involved in managing, cleaning, and wrangling the data, such as data scientists and IT.

Roundtable participants also discussed the data wrangling challenge of creating and managing reproducible data products. Could companies recreate something like an old forecast value at a later date in order to measure data quality or debug a problematic outcome? Doing so implied being able to trace exactly which bits of raw or processed data went into a given forecast or analytic value. One company uses version numbers on data products. It flags the current version but also keeps the old versions. Similarly, another company uses snapshot version control. Sometimes, the input data changes, too—a company might suddenly decide to run a promotion when all the forecasts are based on there being no promotion. One company handles this change by creating special versions of forecasts for these different scenarios and flagging the currently applicable one. The company can trace any given output result back to the relevant data set.

An online retailer said it has a philosophy of documenting everything, including all database tables, for its data dictionary. The company also uses a fairly standardized approach to naming tables and columns, which helps people understand the semantics of the data. Documentation helps address the problem of derivative data created when people get a copy of some data and apply their own non-standard formula to derive some variant of the data for their own purposes. Documentation does, however, involve extra resources.

A retailer described some of the data cleaning tasks it does. Most of the work is in matching identifiers across different data sources, such as relating social media posts or Google Analytics to product IDs in POS data. A second cleaning task is dealing with nonsensical values. Unless the team can trace the value back or find an expert who knows what it means, they cut the aberrant entry, recreate it, and hope that the bad value never made it into any reports. In one case, an outsourced warehouse had taken liberties with the metrics, and the values of those metrics had to be corrected. A third task concerns missing data, which requires finding a proxy for the value, such as scan data to fill in for missing GPS data. The retailer said that its main enterprise data sources were cleaner than its smaller data sources,

A manufacturer is working to automate and scale its data management and wrangling efforts. "You can't throw people at all the missing and wrong data issues," said the company. It is creating tools for interpolating or correcting wrong or missing fields. Specifically, the company is building a three-element approach—combining a master data management platform, a data catalog and glossary platform, and a data quality monitoring platform—to support machine learning and automation of the data stewardship process. These machine learning algorithms will be trained using historical raw data paired with the manually corrected versions.

Data Visualization and Value Delivery

An online retailer introduced the roundtable's visualization session by suggesting that there are three levels of dashboard user: executives, functional managers, and operational users. The type of user and their needs determine the types of data to show them and the complexity of the visualization. For example, the presenter joked that executives only want to see an emoticon on their dashboards: a happy face or a sad face—a highly aggregated view of whether the organization was running well or not. For executives, pass/fail or red/yellow/green status tiles provide a high-level overview of how the organization is functioning and helps direct their attention to areas of underperformance. Such colored tiles might show the color-coded status of KPIs, the current value of the KPI, and the age of the data.

In contrast, other users of a more technical and functional nature might want graphs of KPIs to see the ebb and flow of activities and outcomes. These plots typically included warning-level and danger-level threshold lines. These users wanted to see not only the current levels of the KPIs but also the rates of change, spikes, previous-period or year-on-year values, exceptions, event frequencies, and whether the daily or weekly patterns looked normal or not. For these users, changes or anomalies matter, even if they have not yet hit any thresholds. Moreover, functional and operational users often need disaggregated KPIs for performance on geospatial, product category, or customer segment divisions; these needs require different kinds of maps, pie charts, and other types of graphics.

Some companies design dashboards in collaboration with external users. For example, a 3PL has to understand what users such as carriers need to do their jobs. The dashboards are designed to provide end-to-end visibility onto logistics, transportation budgets, and carrier performance. Day-to-day operations people want to know when shipments are arriving, so the 3PL slices and dices the data as suits those users. The 3PL said that offering dashboard visualization to customers was a key selling point—an example of how data management capabilities can be used to win new business.

Dashboard design involves input from the user—not only the selection of data or KPIs for the visualization, but also the user's needs for thresholds or alerts that might drive them to action. Part of the design process for these dashboards is determining what is normal, what is abnormal, and how users want to see levels and changes in conditions. It's important to understand the urgency of the data and who is using the data. In other words, what the users want to understand from the data drives the visualization. If the dashboard does not present the data users need to see, they won't use the display.

The group then discussed the differences between visualization for laptop computers versus mobile devices. One participant said the laptop version of its dashboard had 18 full-blown charts—far too large for easy use on a small-screen device. Mobile dashboards need to be much simpler, emphasize a few KPIs, and offer less detailed functionality. Some participants used Tableau for laptop dashboards but said that Tableau was not as good for mobile use. Databox and Domo were better "mobile-native" choices, although Tableau was getting better.

Not all users were equally comfortable with data. Several participants noted that it can be too easy to overwhelm some users. Other users simply don't know what data they need. Both building and deploying dashboards can involve communications and training of users to help them understand data, participate in collaborative design efforts, and make effective use of the results.

A participant cautioned that red/green was a poor color combination for status tiles, threshold lines, and data point plots. These two colors can look indistinguishable to colorblind users (about 4–5% of the population). Orange/blue was a better pairing. Dashboard designers (and data scientists) should test their color schemes with utilities that can simulate the effects of color blindness on the usability of visualizations.

Real-time data and more advanced visualizations play a key role in managing complex systems such as supply chains. These visualizations go beyond the KPI view of performance or the SQL-query approach to analysis. Advanced visualization helps provide intuition or a debugging perspective on what a system is doing. For example, optimizing and operating a complex system like last-mile routing involves all sorts of edge cases and tricks such as avoiding left-hand turns. Visualization of scenarios lets experts and users jointly see and react to a proposed solution and offer insights, such as a last-mile routing that makes a driver go through an unsafe area or causes a delivery time the driver knows is less desirable for that customer.

A digital-native company listed several factors that drive dashboard design:

- the technical sophistication of the user
- the viewing platform (mobile vs. desktop)
- the immediacy of the action the user takes with the data
- the depth of diagnosis or remediation the user might perform

Finally, an e-commerce company described how it uses three tiers of monitoring to create a holistic view of the organization. The lowest tier monitors the status of the company's technology infrastructure: Is the website and network up and running smoothly? The middle tier monitors the models and software: How many product recommendations are they making? The third and top tier monitors aggregate execution at the business level: Is money flowing in and are deliveries going out? Each level catches different types of problems to detect hints of smoke before the house is on fire. High-level alerts, such as an anomalous drop in revenue, can quickly catch cases in which, for example, the website is running well but has a disconnect to the payment processor that prevents orders from going through.

Update Cadence: How Real Is "Real-Time?"

The volume of data available to companies is growing, and much of it derives from streaming sensors and high-speed computing processes that can provide real-time data and updates. However, is the cost of wrangling large volumes of real-time data justified in terms of the value it provides? As one participant said, the value of real-time updates depends on the use case and the pace to which the user is accustomed. Real-time is in the eye of the KPI holder: someone accustomed to getting a weekly report on the following Wednesday might feel that getting daily reports first thing in the morning is "real-time." A manufacturing company commented that real-time updates aren't needed unless they will influence a decision or action.

In contrast, other participants mentioned apps that do need real-time, such as monitoring e-commerce platforms, tracking commodity prices on real-time markets, managing timely deliveries to a schedule, and re-routing drivers for traffic congestion. Moreover, some operational managers want timely tracking of fluctuations in their operations to handle downtime and demand surges. Other functions, however, such as finance, can operate with slower cadences. Some data sources, such as third-party market analytics, might have very slow cadences (e.g., a monthly market share report). System designers need to understand the cadence of the data and the cadence of the user's workflows to optimize the frequencies of data gathering, wrangling, and visualization.

Cadence also affects technology platform choice. Some companies do not store certain types of real-time data, such as traffic congestion or weather updates, in their data lake. Data lakes are geared for planning and strategy, not real-time operations. If immediate action needs to be taken, the data lake is not the right platform. (Although one company said it may need more operational data in the data lake than it originally thought.) Instead, for real-time analytics, companies use streaming data processing platforms such as Apache Kafka and SAP HANA.

A digitally native company argued for more real-time data collection, saying that building a reporting system is similar whether the data is real-time or not. The company cited two benefits of real-time systems. First, the finer granularity of real-time data may reveal subtleties not visible in aggregated data. For example, an online retailer runs weekly forecasts over 12 weeks. Real-time data would let the company see intra-day trends, such as if people are shopping heavily between 10 a.m. and noon, which helps the company to manage its operations more efficiently. Second, today's non-real-time users may want real-time in the future and may need to look back at historical real-time values or use that fast-cadence data for machine learning. The digital-native philosophy sees no harm in gathering more data, only opportunity.

Managing Code

An e-commerce firm uses repository technology commonly utilized by open-source software projects for distributed collaboration. The company stores all the database schemas and code of the entire company in that repository. The repository provides wide access to the metadata and code to support efficient reuse, bug fixing, and enhancement. The company treats database schemas like code and lets anyone modify a table and submit a "pull request" to have the company's DBAs (database administrators) review the change before committing it to broader use. This review process provides a good checkpoint and hopefully helps enforce data conventions and documentation requirements.

The company also created interesting home-grown tools to search of all the metadata, schemas, and code of the organization. With these tools, data engineers and data scientists can reverse engineer how various users' analytics and dashboards are constructed, which database columns they use, and how they use the data. The only limitation is that the tools don't search inside the files of code-like applications such as Excel that many employees use for manipulating data for decision making and other business purposes.

Cloud Services: Silver Lining or Chance of Rain?

Many of the participants were using cloud computing vendors for large-scale data, hosting, and virtual machines for on-demand computing. During the discussions, companies reported using a mix of Microsoft Azure, Google Cloud Platform, and Amazon's AWS. The services provide managed and scalable resources that are needed for big data lakes, computationally intensive data science algorithms, and organization-wide access to data and applications. However, participants reported some problems with these cloud computing services, such as a lack of support from Azure for Python and occasional Google Cloud system failures when data loads are high.

From Pilots to Production

At various points during the two days, the participants touched on the problem of managing the life cycle of data science projects. Data science initiatives often produce a proof of concept or prototype used to demonstrate the potential value of using a particular type of data or model in business. However, these early versions aren't ready for production use. Data scientists often use computational notebook environments, have little formal training in software engineering, and may create inefficient code. "Industrializing" a prototype to create a robust, scalable application takes time and resources. A technology company said it has built about 18 beta releases in the last eight months, but that some of them won't go into production for as long as two years.

The group debated how to go from pilot to production. Three companies expected their data scientists to help put initiatives into operational use—"In our organization, we own from concept to outcome," said one participant. In contrast, a pharmaceutical company splits the duties between different teams. The data scientists who develop the prototypes are not the ones who put them into production because that involves much different volumes of data and different engineering issues.

A technology company spoke of encouraging citizen data scientists among the organization. To support this, the company offers features such as self-services to build data models on the data lake and pre-building data cubes for easy consumption. The company uses Azure because it believes Azure offers better accessibility to these tech-savvy business employees. Reducing barriers to entry enables businesspeople to make more use of the data. In companies with a bottom-up culture, many teams might try to address the same problem simultaneously, leading to the need to manage the amalgamation of the teams or their efforts.

However, some participants expressed concern about citizen data scientists, such as those who use analytics platforms like Alteryx to create their own data-collecting, wrangling, and visualization tools. These specialists sometimes build complex systems and use them in production. One participant joked that they had applications named after the individuals who had created them. However, these ad hoc applications are often not reliable. If the creator goes on vacation or leaves the company, the application crashes and no one knows how to fix it. One manufacturer had worked to regulate Alteryx and pull more of it into the managed data warehouse. Another participant commented that Alteryx was useful for creating a minimum viable product but could not be used forever.

Many of the participants lamented that large portions of their company's data, data processing, and data-related decision making are held in Excel files scattered across the organization. These spreadsheets often contain data of unknown or dubious provenance, age, and processing. These spreadsheets might use undocumented and untested formulas and macros that may contain hidden logic errors. When the creators and users of these spreadsheets share them with others, they spread potentially contaminated data and results across the organization that affects decisions in untraceable and irreproducible ways.

Some participants wanted to eliminate use of Excel at their companies altogether. At least one firm refuses to give Excel to employees unless they can articulate a business purpose. However, other participants defended Excel as the easiest way for the average worker, manager, or executive to do ad hoc data processing and charts for visualization for their job.

Even some data scientists and technology developers find Excel to be a convenient tool for pilots and prototypes. However, other participants cautioned that the computational limits of device-hosted Excel may actually constrain how people think about data; they tend to think much more narrowly, missing the many opportunities to use big data, sophisticated data science algorithms, and elastic cloud computing resources to solve the organization's problems. Thus, some say Excel creates a bad mindset for data science tasks.

The People for the Journey

One myth about data management is that data—by itself—has value. In reality, data is messy, which impairs its value. Even after the data is cleaned, it's not necessarily valuable. Someone has to relate that data to the business and to decisions or actions that can make use of that data. Someone must use mathematically valid and computationally feasible techniques to mine the data and extract the valuable description, prediction, or prescription for a business application. Thus, both the journey of the organization and the journey of the data depend on highly skilled specialists to convert raw data into useful information as well as in helping executives, managers, and front-line workers leverage that information in their daily tasks.

Wanted: Data Scientists

Data scientists play an essential role in making the most of a firm's data by finding patterns in the data, providing statistically valid conclusions, and making sound forecasts from noisy data. Data science tasks can include collecting the data, cleaning the data, modeling the data, selecting among the many data science algorithms, testing and tuning those algorithms, and presenting the results. All of these tasks required collaboration with the business because the data scientist won't know what the data means in business terms, which values are erronous, and what patterns, statistical tests, or predictions are most salient to the business.

The exact job description for a data scientist wasn't clear. The group could not agree on whether laborious data cleaning tasks should be expected of data scientists or not. Some participants thought the data scientist should handle them all, while others offloaded some of the data housekeeping tasks to data engineers, AI engineers, and software engineers.

Many of the participants in the room were hiring data scientists. Participants mentioned hiring data scientists "young," meaning directly from top universities, typically at the master's or PhD level. However, a shortage of qualified people meant that some enterprises were poaching data scientists from other companies, too. The talent shortage also motivated some companies to propose hiring data engineers and building a team to support data scientists, thereby helping retention and enabling the data scientists to focus on the highest-value tasks.

The group discussed which skills data scientists should have. A solid background in math and statistics was seen as an essential prerequisite. Several participants said that better software engineering skills were important for creating robust and feasible solutions that can be put into production at scale. Good communications skills were also very useful for the data scientist's role in collaborating on the needed data, technical approach, interpretation of results, and extraction of value from data. In contrast, knowledge of the business and the functional area were advantageous but not essential because willing individuals can acquire this knowledge from subject matter experts.

The extremely high demand for data scientists made retention a very serious problem. Companies lamented how often they hired good people, trained them, and then saw them leave. The keys to retaining data scientists included giving them hard, interesting, and vital challenges along with the tools and data needed to solve them. Although some data scientists might be attracted to join Wall Street or a startup, supply chain managers can entice data scientists with many interesting high-value, real-world problems, such as network design decisions involving millions in capital expenditures, advanced math for operations research, large-scale operations affecting global supply chains, and cutting-edge analyses and visualizations of the complex dynamics of supply chains. Ensuring that the company had the tools and processes in place needed for data science before hiring the scientist was also a key to retention. Managers of technical teams can help by insulating their data scientists from the annoyances of organization politics and too many low-value staff meetings.

One key retention technique was to have a clear career path for data scientists. The path might include a succession of job titles such as "distinguished engineer" and "principal engineer," with pay and perks rising to be equivalent to the vice president level. Another suggested option for a career path was more like becoming a consultant and being offered a choice of interesting projects.

Wanted: Data Stewards

Two companies highlighted the essential role of data stewardship in data governance to address the data quality, data reliability, and data availability challenges that companies face. Data stewards play key oversight roles in gathering, cleaning, processing, storing, and accessing the data as part of managing the lifecycle of the data. They support both the data scientists and the business users of the data on issues such as privacy, security, and risk management. However, discussions revealed that this key role is not in any job description.

Senior leaders need to recognize the data stewardship is not a technical issue but a business management one. With data ownership comes the responsibility for data stewardship. As with data scientists, data stewardship is also a career path issue that needs to offer employees some incentives to take on the added responsibilities of data governance. At some level, data stewardship is something to be socialized across the organization to the extent that more and more people have roles that create, process, and use data. One manufacturer noted that data governance is a mature domain overall, but it has not infiltrated the supply chain domain yet, where many maintain that "their job is to move boxes."

Wanted: Cognitive Designers

One tough problem is the tricky balance between overwhelming and under-informing decision makers with too much or too little data, respectively. If there are hundreds of data points and an excessive number of flashing and pulsing exhibits, the human viewer can become overwhelmed. The military discovered that operators of drones can only pay attention for 20 minutes before their cognitive tracking declines. One e-commerce company suggested that companies needed to add a "cognitive designer" or UX (User Experience) person to their teams. A manufacturer said that 20% of cognitive design skills could be taught fairly quickly to the team, but having a specialist who is expert at design could reap big benefits for the company.

The Journey into the Future

The group was glad to see that they were at similar points in their journeys. No company was as behind as they thought they were relative to others. With so many companies in the early stages of their data management journeys, much work remains to be done. That includes work on future applications, visualization of business data, and future roundtables.

Future Applications

Several companies shared their future application plans, which reflect three trends toward more mature and advanced use of data in business. The first was a trend toward end-to-end integration for holistic management—using integrated data to better manage the entire enterprise, not just particular silos. For example, one retailer wants to create one source of truth for all data and drive supply chain activities to support the customer experience. A manufacturer noted that the future will require that functional people sub-optimize their own metrics for enterprise ones. For example, a plant manager who moves the most units possible (their KPI) but creates downstream scrap isn't contributing as effectively to the company as a whole. Or, some people might become slightly less productive if they are also tasked with creating clean data, but the end result will bring higher total performance.

The second trend was a shift from descriptive analytics to predictive or prescriptive analytics for more proactive rather than reactive management. For example, two companies were working on predictive models for service failures to help avoid missed pickups and the subsequent service problems and customer fines that those failures create. A 3PL was working on predicting where drivers will want to drive next to help ensure they get home when they want. A business products maker was working on using images of product shipments to predict the space they will need for shipping.

A third trend was greater sophistication of models. Two companies were working on advanced pricing models that combine internal data with external data, such as syndicated data and social media. A technology manufacturer was working on advanced statistical inventory models for situations where demand that does not follow the normal distribution.

All of these trends call for a growing use of data and more complex math, software, and applications that boost enterprise performance.

CAVE: Future Visualization

At the end of the roundtable, the participants visited CTL's Computational and Visual Education (CAVE) Lab, which consists of an expansive corner-to-corner touchscreen wall display, a large "holotable" like a giant iPad, and a control console. With the system, 10 to 15 people can collaborate by interacting with the various screens, looking at geospatial presentations, dashboards, and other presentations of complex visual data. The CAVE Lab enables natural interaction with analytics and enables visual storytelling. Executives and managers who come to the lab to improve their supply chain networks don't have to change lines of code to change the simulated network; they can just click on the map, for example, to deactivate a facility, and see how that change affects the network. CTL uses the CAVE Lab for sponsored research projects in which MIT CTL researchers work with an organization's data and problems to create both optimization and visualization deliverables. Two demos illustrated what was possible.

The first demo showed research into last-mile route optimization on behalf of a logistics company. During the demo, the large tabletop display replayed a day's worth of GPS telemetry data gathered from a fleet of delivery drivers in an urban area. Watching the dynamic, color-coded traces revealed how drivers progressed on their routes. This revealed inefficient driving patterns caused by possible flaws in the routing algorithm, flaws in the map data, constraints on delivery schedules, or the driver not following the route. The visualization also helped spot dirty data in which a driver appeared to jump miles in an instant. With more data, the visualization could show the temporal patterns in deliveries due to traffic or weather. The intent of the project is to help improve the efficiency and service quality of last-mile delivery.

A second demo showed a distribution network optimization project on behalf of a chemical manufacturer. The analysis addresses the question, How many warehouses does the company need? One optimization challenge was that customers preferred ordering the product from nearby suppliers, which meant that market share depended on distance to the customer. A key feature of the project was that the visual geographic presentation of system on the large table-top touchscreen enabled sales and supply chain executives to collaborate. Sales could contribute their knowledge of customer behavior and the effects of distance on market share. Supply chain could contribute their knowledge of costs and operational issues. And the data and visualization could show how demand is distributed. The group could discuss different sales parameters and network configurations within the simulation and visualize the effects on market share, revenue, and profits. The result was not only a better solution but also greater trust in the solution than what would be achieved through some inscrutable optimization algorithm.

The demos and the discussions during the roundtable showed how data and advanced visualization will play a greater role in the future. Virtual reality (VR) with 3D interfaces could enable managers to literally see the depths of their data and KPIs. Augmented reality (AR) could offer new avenues for overlaying KPIs, exceptions, and other data onto real-life 3D scenes such as pallets in a warehouse, store aisles, or freight yards. The challenge, though, is how to present more data to the user without creating more clutter and confusion. That will require an integrated approach involving data science, graphic designers, and business stakeholders.

Future Roundtables

This roundtable was the third in a series on the topics of machine learning and data management in the supply chain. MIT CTL plans to have additional roundtables on the topic in the future. During the final discussions in the wrap-up of the event, participants shared their preferences for these future roundtables. Overall, these requests fell into three major groups: more sharing of details, more emphasis on the cross-functional aspects of this topic, and more on the subject of visualization.

Many participants wanted more sharing of details within the collegial, non-competing venue of the roundtable format. Some wanted deeper dives into the workflows of initiatives to cover connecting data management to data science, data to value creation, and the product life cycle of creating pilots and production releases. Others asked for details on specific elements, such as use cases or data cleansing. Finally, some participants wanted discussions of higher-level management issues such as engagement models, governance and data structure policy, and the balance between low-risk, use-case driven projects and high-risk, exploratory projects.

Other participants wanted to take a deeper look at the cross-functional issues that are being raised by broader enterprise data management and machine learning initiatives. Three participants wanted to look more at the functional side and the integration of the supply chain back end and sales/marketing front end such as S&OP (Sales & Operations Planning), product substitution engines, and stockout analysis. Future roundtables might also include technologists with business partners so that both sides can understand their respective challenges. Such roundtables would depend on companies bringing the right representatives who can share their insights into these cross-functional issues.

Several participants asked for more depth on visualization and the right way to present information so that people get full value from dashboards. They suggested looking at human-centric analytics, cognitive processing of information, and design principles. Two companies wanted more information on the CAVE Lab and on how other companies visualize business information, especially at scale.

In conclusion, organizations will only get value from their data if employees, managers, and executives can quickly and accurately see the business implications of that data. This roundtable revealed that collecting, managing, wrangling, and governing data are prerequisites to gaining that value, but they aren't sufficient. The visual presentation of data closes the loop and drives people toward the right actions. As one participant said, "That's where the rubber hits the road. How do I use this information—every bit of information?"

MIT Center for Transportation & Logistics

Report recording prepared by:

Andrea Meyer and Dana Meyer
WorkingKnowledge®
3058 3rd St.
Boulder, CO 80304
workingknowledge.com

Edited by the moderators.