Analytics of the Future
Predictive Analytics

Summary Report

Cambridge, Mass.

November 18, 2020

Moderated by:

Dr. Matthias Winkenbach
Mr. Jim Rice
Ms. Katie Date

ctl.mit.edu

MIT Center for Transportation & Logistics

Table of Contents

Executive Summary

MIT's Center for Transportation and Logistics (CTL) hosted a virtual roundtable for its Supply Chain Exchange partners in which leading companies discussed predictive analytics. The event combined presentations from academia and industry with sharing by all attendees of their experiences, challenges, and ideas. To encourage candor, no statements in this report have been attributed to any specific company.

A short presentation summarized key concepts and the main algorithmic methods (see Appendix) for doing predictive analytics, including decision trees, random forests, k-nearest neighbors, support vector machines, artificial neural networks, regression, and time series.

During the roundtable, participants introduced themselves and described their firms' uses of predictive analytics; this initial discussion showed the diversity of use cases for predictive analytics in supply chains. Companies listed various applications in demand forecasting, predicting the timing of events (e.g., driver availability, container unloading, and shipment events), and anomaly or risk prediction (e.g., manufacturing scrap rates, anomalous orders, and service failures). Applications for forecasting predominated in 70% of the companies, a pre-roundtable CTL survey found.

One company, a maker of technology products, presented its approach to predictive analytics, which included processes for identifying target projects, prioritizing them, developing minimum viable products in order to get feedback, and then iterating to create tools that address business users' needs. The company centralized its data, analytics, and optimization efforts to provide enterprise-level management of data strategy and application development. The company's team for analytics included both technical staff and "data translators" who bridge the gap between technology and business.

Discussions often focused on the challenges of predictive analytics in companies. Prevalent obstacles included data availability (e.g., quantity of samples, the right variables, and quality), organization maturity, and alignment of data science projects to organizational needs.

Ultimately, predictive analytics is a journey with a beginning but no ending. Companies can always find new sources of data and new applications for using that data to reduce costs, improve reliability, and add value.

Predictive Analytics in Supply Chains

Throughout the roundtable, participants described how they use predictive analytics in their supply chains. Some of the applications and roundtable discussions blurred the boundaries between descriptive, predictive, and prescriptive analytics. (Descriptive analytics provide an understanding about the present; predictive analytics provide insights into the future; and prescriptive analytics provide recommendations about actions.) This blurring occurs because managers ultimately want to use data to guide action, which is inherently prescriptive. However, guiding actions often requires descriptive analytics to understand the situation and predictive analytics to forecast what might happen next, in order to optimize the action. Overall, most of the applications discussed at the roundtable fell into three categories: forecasting demand, predicting the timing of events, and foreseeing risks or disruptions.

Forecasting Demand

A pre-roundtable CTL survey (described in section 3.1) found that 70% of companies selected demand forecasting as being the area in which their organization predominantly employed predictive models. During the roundtable, several companies from a diverse range of industries mentioned demand forecasting as one of their main applications for predictive analytics. The reason is because the anticipated volume of business affects many activities in procurement, manufacturing, warehousing, distribution, and retail, such that demand forecasts play a central role in planning in all these areas.

Predicting the Timing of Future Events

Four companies described how they use predictive analytics for estimating the timing of future events. For example, one carrier predicts when drivers will be available for the next load. Another carrier predicts the timing of unloading of containers and has reached 83% accuracy with just two months' worth of data. Both carriers can use these predictions for improving staff scheduling. The third company, an enterprise software company, predicts the timing of shipping events, especially going into the holidays. Finally, a manufacturer, uses predictive analytics to forecast the submission of large orders in the deal pipeline.

Foreseeing Risks or Disruptions

Several companies use predictive analytics in the context of risks, such as outliers and disruptions that potentially occur at many points in their supply chains and organizations. In manufacturing, two companies where looking at yield and scrap rates. On the transportation side, a carrier was predicting service failures that could cause a load not to be delivered. This problem also had a time prediction aspect, namely forecasting when the shipper would have a load available versus when the carrier's network would have capacity for that load. On the sales side, an enterprise technology product maker was predicting anomalous or "disruptive orders" that could affect the supply chain. Although its sales staff are expected to understand the life cycle of the deals they are working on, predictive analytics could help forecast the timing of the order, especially on behalf of newer, inexperienced sales staff. Finally, on the customer side, two companies were using predictive analytics to identify potential customer churn or defections.

A Leading Organization's Approach

A large technology hardware, software, and service company shared its extensive efforts in using data science and predictive analytics, which were part of its company-wide four-year digital transformation journey. At the roundtable, the company showed its data science portfolio, which listed 16 initiatives spanning classification, forecasting, clustering, anomaly detection, optimization, and simulation. These initiatives served corporate functions including planning, procurement, manufacturing, and logistics. Although these initiatives also included descriptive and prescriptive analytics projects, they illustrate the breadth of applications of data science and analytics to supply chain organizations. The company's approach involved creating a process for developing analytics and organizing people to achieve the aims of the digital transformation.

Process for Creating Analytics

One of the biggest roadblocks to using any kind of analytics is deciding what the target applications should be. To do this, the company answered a much wider set of questions. They didn't start their thinking with "What [analytics] organization do we want?" Instead they asked, "What do we want the business to be?" Creating a vision of the future of the whole company led to envisioning a set of "experiences," which are how the company's employees experience their day-to-day work lives. The result is a set of technology use cases to support where the company is currently and where it wants to be in the future. The process for creating analytics also required asking what was possible with the data and technology before proposing initiatives to executives.

The goal of the company's ongoing process is to map as many opportunities as possible. Doing this involves getting input broadly across the executive level and the engineering level to map the many challenges that the company is facing. The result is a long list of ideas, from little ones to grand schemes. The next step is to prioritize them.

To prioritize the targets, the company uses a multidimensional quadrant approach. The first dimension is business value — projects need to solve the highest value problems, otherwise they're just exercises in data science. The second is ease of implementation — something is hard to implement if it requires multiple data sets, social media, and lots of work to accomplish properly. Next, the company assesses two more characteristics: innovativeness of the project and ease of adoption (i.e., that the team needing it can easily do change management). "Innovative" is not an essential requirement, but innovative projects help keep data scientists interested, which is important for their job satisfaction and retention.

The company's development and deployment process emphasizes creating a minimum viable product (MVP) rather than perfecting the product. The MVP may not have all the bells and whistles, but it does provide valuable feedback from the business users that then guides or redirects the development effort. Early adopters might be a select few in the organization, but they help ensure the project ultimately creates something useful. In some cases, the project morphs over time as business needs, processes, or KPIs change.

Some efforts cut across applications and functions. For example, forecasting plays such a pivotal role in so many areas of the company's activities that the company built a center of excellence around it. Rather than buy a vendor's solution, the company decided to build its own tool to incorporate the many good practices that the company had learned from different areas. The resulting tool combines numerous classic forecasting models as well as cutting-edge advanced machine learning techniques such as neural networks. The tool offers automation for ease of use by less-technical users, but it also provides power users (such as data scientists) with access to the internal technical elements. The purpose is to enable all of the company's team members to use time series forecasting in their day-to-day jobs.

The People Side of the Equation

The company centralized its data analytics into one organization during its digital transformation. Specifically, the team was conceived to tackle and solve the types of business problems that use data analytics, data science, and optimization. The team handles data, analytics, and automation. The team also manages both data quality and data availability across multiple sources of data and information. By centralizing in this way, the team can manage more problems and get deeper into the solutions, such as an organization-wide shared toolkit for forecasting. The centralized approach created an internal consulting practice with knowledge of both the technology and the business.

The company's digital transformation organization — as it has grown over the last few years — now consists of two different categories of people. First, it has the data science workers that develop the technical solutions. Second, it has "data translators" who are the bridge between the business and the technology. Data translators understand both the business and the technology, which means that they can explain the technology solutions to the business and the business' need to technologists. This second category of team members are crucial to envisioning new initiatives, getting alignment, selling initiatives, and driving adoption.

Challenges

Many of the presentations and discussions highlighted key challenges in creating and deploying predictive analytics. A CTL survey taken shortly before the roundtable asked respondents to fill in the blank: "In my opinion, the biggest barrier for my organization to use predictive analytics effectively is..." Answers included:

* "Technical knowledge and knowledge application"
* "Understanding of predictive models, data availability and alignment on use"
* "Data and Integration," "Getting hold of reliable data," "Data quality"
* "The extent to which history is not a predictor of the future"
* "Alignment to business benefit"
* "We respond to disasters and there are a lot of variables including no-notice disasters"
* "IT department"

Discussions during the roundtable touched on many of these issues.

Organizational Maturity Survey Results

One of the first challenges in creating and using predictive analytics is related to the level of understanding and maturity in the organization regarding the technology. A pre-roundtable survey of CTL's supply chain partners asked three questions on this topic with responses on a 7-option scale from "strongly agree" to "strongly disagree." The statement that "the supply chain organization of my company is frequently using state-of-the-art predictive analytics tools in its decision making" had responses with more than a third (38%) on the "disagree" end of the spectrum and less than half (46%) on the "agree" end of the spectrum. The question, "People in my organization have a clear understanding of the difference between descriptive, explanatory, predictive, and prescriptive analytics" also had more than a third (38%) on the "disagree" or "strongly-disagree" end of the spectrum but more than half (62%) that somewhat agreed or agreed. Finally, "People in my organization understand the purpose, potential applications and specific limitations of predictive models" had only 31% on the disagree side and 62% on the somewhat-agree side. The mixed results suggest that organizations are spread out in their journeys toward understanding and using predictive analytics, but most are making progress toward using the technology (which was also echoed in the examples shared at the roundtable).

Big Data Ideals vs. Little Data Realities

Data was called the #1 roadblock to predictive analytics, with five companies making substantive comments about the problem. Simply having enough data was a challenge. Projects don't necessarily fail because they lack the right methods — they fail because they lack sufficient data, the participants said. Lack of data was especially true for deep learning neural network methods that require a lot of data to make accurate predictions. Only the very largest e-commerce organizations have high enough volumes of data for some methods and prediction problems.

Data scarcity can affect parts of a predictive analytics project or limit its scope. A manufacturer looking at supplier ingredient quality and product yield noted that although they have data from millions of units of production, the much smaller numbers of bulk batches of supplier ingredients create a shortage of data for analytics at the ingredient level. Similarly, a carrier noted that they may have sufficient data for their analytics on their biggest customers and highest-volume activities but not for the smaller customers or specific lanes.

The small number of data samples is only part of the data scarcity roadblock. Without data on the right variables or features, the model will fail to differentiate classes of conditions or know the sources of variation that most affect a forecast or prediction. Explained one participant: "So if you're trying to separate, for example, customers based on their spend versus kinds of products ordered, and if those two dimensions don't work, you need to figure out what alternate dimensions or other ways for any of these methods to work." Getting the right variables means being cognizant of the possible drivers for a prediction, and that's often more challenging than it sounds.

One essential type of data required for most predictive analytics methods are the labels needed for supervised learning methods. Unfortunately, much of the data in the world and in supply chains is unlabeled. Getting those labels can depend on human expertise (and labor) to recognize, classify, score, tag, or provide feedback on what happened as a precursor to training a predictive analytics system. For example, predicting whether a given delivery route is "good" or "bad" for driver performance might require asking human drivers their opinions and rationale about different routes to both label the data and understand salient route features (intersections, congestion, etc.). Similarly, sales staff might know what makes a forthcoming customer order anomalous or disruptive. Data on convenient routes or disruptive orders can be tagged by people and then clustered and classified by machine to help understand what makes a good or bad example of these things.

Another part of the data roadblock is data quality in terms of cleanliness and orderliness. "I think many of the companies are not at a point where data is completely structured, organized, of high quality, and in a single source," said one participant. "This is something that we have to constantly refine" said another.

In some cases, key parts of the supply chain are simply a data black hole. A carrier noted that although modern fulfillment warehouses can provide copious data from scanners, many older warehouses in the world, especially for transloading, often rely on manual processes. "That was the biggest black hole from our perspective, to understand why some containers wait a day in the warehouse or maybe six days in that warehouse."

The carrier did develop a potential solution to this problem: they installed cameras around the warehouse, collected video data of warehouse operations, and computed visual analytics. It's still a hard problem because work in such warehouses occurs in fits and starts, here and there, as workers and goods circulate in the facility. However, the technique shows promise. Another logistics company was pursuing the same type of visual analytics solution: using video from the last mile and deliveries. Video cameras have the potential to provide literal visibility onto the previously dark corners and edges of supply chains to provide the missing data needed to do analytics, including predictive analytics.

Organizational Issues and Alignment

Numerous stories and discussions at the roundtable centered on the challenges of using predictive analytics in a business environment. Dr. Winkenbach was surprised that such a vast majority of survey respondents were predominantly focusing their predictive analytics efforts on demand forecasting. In theory, the technology has many other potential applications, such as in procurement, risk management and predictive maintenance. Discussions of the causes of this pattern and related issues highlighted the gap between technology and business.

One participant suggested that the prevalence for forecasting applications could be due to common misunderstandings about "predictive" versus "prescriptive" analytics. Business people say they support predictive analytics but everything they do is actually prescriptive, she said. When analysts explained about all the data that the business would need to gather and clean to make a true prescriptive system, the business people felt that it was too much. "We just want a forecast," they said. Sometimes, based on the way business described what they wanted to achieve, the analytics people proposed delivering a real time recommendation engine, but the business people felt overwhelmed and said, "Well, no, I just want a decision rule. Just give me five decision rules or break down my segment of customers in such a way that I can do X."

Part of the organizational challenge is in selling the idea of predictive analytics to non-technical executives and users. At one carrier, selling even the seemingly obvious potential of predictive maintenance in vehicles faces obstacles because the business people don't understand the algorithms or because there's a difference between what the business wants to see now versus in the future. In some cases, innovative predictive analytics solutions may be harder to explain than simple ones. In other cases, business leaders might hear "leading-edge" buzzwords and think they want the buzzword feature. However, the buzzword may not mean what they think it means, or the feature does not serve the actual interests and priorities of the business.

Although business people might understand the general concept of predictive analytics, they may need to see more concrete potential applications in their own field as well as understand the worst-case consequences of not having predictive analytics. They need to see how the technology can really benefit them and change the way they do things. These challenges highlighted the need for data translators (described in the previous section) who understand the technology, the business, and the relationships between them. Data translators can bridge the two sides to help align an analytics initiative to the business and explain solutions to help bring business people on board.

Finally, managing adoption is another challenge at the boundary of organizational and technological issues. Some users want certain capabilities earlier than others do; partial roll-outs can incorporate feedback loops; and some capabilities take more development than others do. Two companies stressed agility — using early roll-outs of preliminary solutions to get feedback rather than waiting for the perfect solution. The deeper point is to not ignore the adoption issue and thereby run the risk of creating solutions that can't be used.

In addition to explaining predictive analytics projects to decision makers is the need to develop overall strategies for data and development. Thus, another important aspect of data science in the organization is having an enterprise data strategy that replaces a federated, disparate, or siloed view of the data. A logistics company explained how data is like fuel and is becoming a strategic asset. An enterprise view of data is an effective way of enabling solutions in different verticals and across functions.

A related issue is the enterprise's technology development strategy and managing the balance or allocation of responsibilities among the corporate IT organization, any centralized data science group, business units, and external vendors. Some companies mentioned collaborations with vendors while others mentioned in-house efforts. Two companies mentioned using centers of excellence around analytics that can bring together both technical expertise and skilled "data translators."

The Never-Ending Journey to the Future

The technology company that shared its approach to analytics at the roundtable noted that although there is a beginning to these efforts, there is no end. Predictive analytics is an ongoing effort rather than a one-and-done project. Similarly, a carrier noted the never-ending opportunity for improvements from using analytics: "Every time we go into it, we discover something very specific that we can identify as an improvement with our shippers. So it's been very valuable to us."

The presenter from the technology company saw a pitfall in thinking about the journey for machine learning and its relationship to human knowledge. He said the objective of any machine learning or data-driven solution is not to encode human knowledge because that's likely to keep doing things the way they've been done before, rather than making them better. Instead, by taking a true data driven approach, a company might come up with different and better solutions. Sometimes changing the definition of what is being sought or changing the KPI can change the way of thinking about the problem and can lead to solving it in a new way.

The journey is also continuing through roundtables like this one. This roundtable was part of a longer series of events covering the broad and evolving topic of analytics and data in the supply chain. Four other events over the last two years have covered areas including AI/machine learning, digital transformation, data management, and robotic process automation. The next event will cover prescriptive analytics. Through this and future workshops, CTL and its supply chain partners will continue drilling down and focusing on specific applications and facilitating sharing of experiences.

Appendix: Predictive Analytics Methods

Predictive analytics fits into a spectrum of analytic methods that help convert data into: an understanding about the present (descriptive analytics), insights into the future (predictive analytics), and recommendations about actions (prescriptive analytics). Dr. Matthias Winkenbach, Director of the Megacity Logistics Lab at the MIT Center for Transportation and Logistics, provided a short overview of several types of numerical methods used for predictive analytics. The introduction provided a common terminology, definitions of different analytic approaches, and a few of the main issues when applying predictive analytics. The short presentation was not intended to dive into technical details and did not discuss the pros and cons of various methods. Slides from this portion of the presentation are available to CTL partners.

The Language of Data Analytics

The language of data analytics has been heavily influenced by the computational challenges of recognizing objects in images and other patterned data. Within a dataset, the known or measured variables in the data sample are often called the "features," "attributes," or "dimensions" of that data. The values that are sought as answer to the analytic process (e.g., the forecast demand, predicted arrival time) are often called "labels." Datasets where the labels are known are called labeled data and data without labels is unlabeled data.

Supervised learning and unsupervised learning are two key categories of machine learning. Supervised learning (often used for predictive analytics) requires labeled data for training. The labeled values provide the "supervision" or feedback needed to see if the predictive model is getting the correct answer during training with labeled data. Unsupervised learning (often used for pattern recognition, data clustering, and reducing the dimensionality of data) can use unlabeled data. Supervised versus unsupervised learning is more of a spectrum than a black-or-white dichotomy — semi-supervised learning works on partially-labeled data. A third category of machine learning called reinforcement learning — used for prescriptive analytics and game-playing systems —has the machine model trying various actions and getting positive or negative reinforcement based on the outcomes of the actions.

Machine learning processes typically have two phases: training and inference. The labor-intensive and computer-intensive training phase uses pre-existing data to build an analytic model that best reflects the patterns hidden in the data and the objectives of the effort (e.g., minimum forecast error). Typically, the training phase attempts to estimate or find and refine a set of model parameters that provide the lowest error or best behavior of that model. Then, the inference phase uses the trained model with new data to generate the needed predictions or other analytic data products.

One noteworthy challenge in building analytic models is called overfitting. Overfitting causes the model to learn false patterns created by statistical noise, spurious details, and non-representative outliers in the data. An overfit model will have less error on the training data (seems very good) but then worse error when applied to new data (which really is very bad). A major part of the "art" of data science lies in managing this issue through using some of the training data to test the model, tuning the parameters, and understanding the limits of the quantity and quality of the data.

From Decision Trees to Random Forests

The first type of method Dr. Winkenbach presented was decision trees. Decision trees build a prediction or decision model that looks like a forking flow-chart of if-then statements. Each statement might be very simple. A decision tree for predicting late deliveries might have simple rules such as, "if the route > 20 miles," "if destination = rural," or "if weather = snow." The sophistication of the model comes from the nesting of the rules that carefully subdivide the space of conditions into predictions, such as "late" and "not-late" expected delivery times.

Random forests are a sophisticated and popular method that builds on the decision tree concept. The "forest" is a collection of decision trees and the random aspect refers to how the method tries various combinations of the feature variables and data samples in an attempt to find the most predictive variables and the best possible trees while reducing the risks of overfitting. Random forests aggregate the outputs of their constituent trees to (hopefully!) provide a robust prediction or analytic output.

K-Nearest Neighbors

K-nearest neighbors is a commonly used predictive method because it's relatively simple to implement. New data values are compared to existing labeled data by looking at those historical data samples in which the conditions were most similar (nearest neighbors) to the current new value. By looking at some number (K) of these nearest neighbors, the model assumes the future will be similar to what happened when conditions were similar in the past. The size of the neighborhood controls the smoothing — too large a neighborhood will miss important details but too small a neighborhood causes overfitting.

Support Vector Machines

Support vector machines (SVM) are a more sophisticated means of finding some vector, line, or plane in the feature space that best divides the dataset among the labeled classes (e.g., loyal customers versus defecting customers). Special tricks and higher-dimensional spaces can help SVM handle cases with a curved boundary between the labeled classes or where the classes seem to overlap.

Artificial Neural Networks

Artificial neural networks are a very sophisticated and successful machine learning technology modeled on the networks of neurons in the brain. The model uses layers of neurons between the input data and output result. Each neuron computes a mathematical function that is a weighted combination of either the data or the outputs of the previous layer of neurons. Each neuron then passes its output to more neurons in the next layer or to the final output. With many neurons in each layer and many layers in the whole network, the sophistication of the total mathematical function between input data and output result is almost unlimited. However, as powerful as neural networks can be, they suffer from being true "black boxes" in that it's almost impossible to interpret how or why a neural network produced the answer it did.

Regression

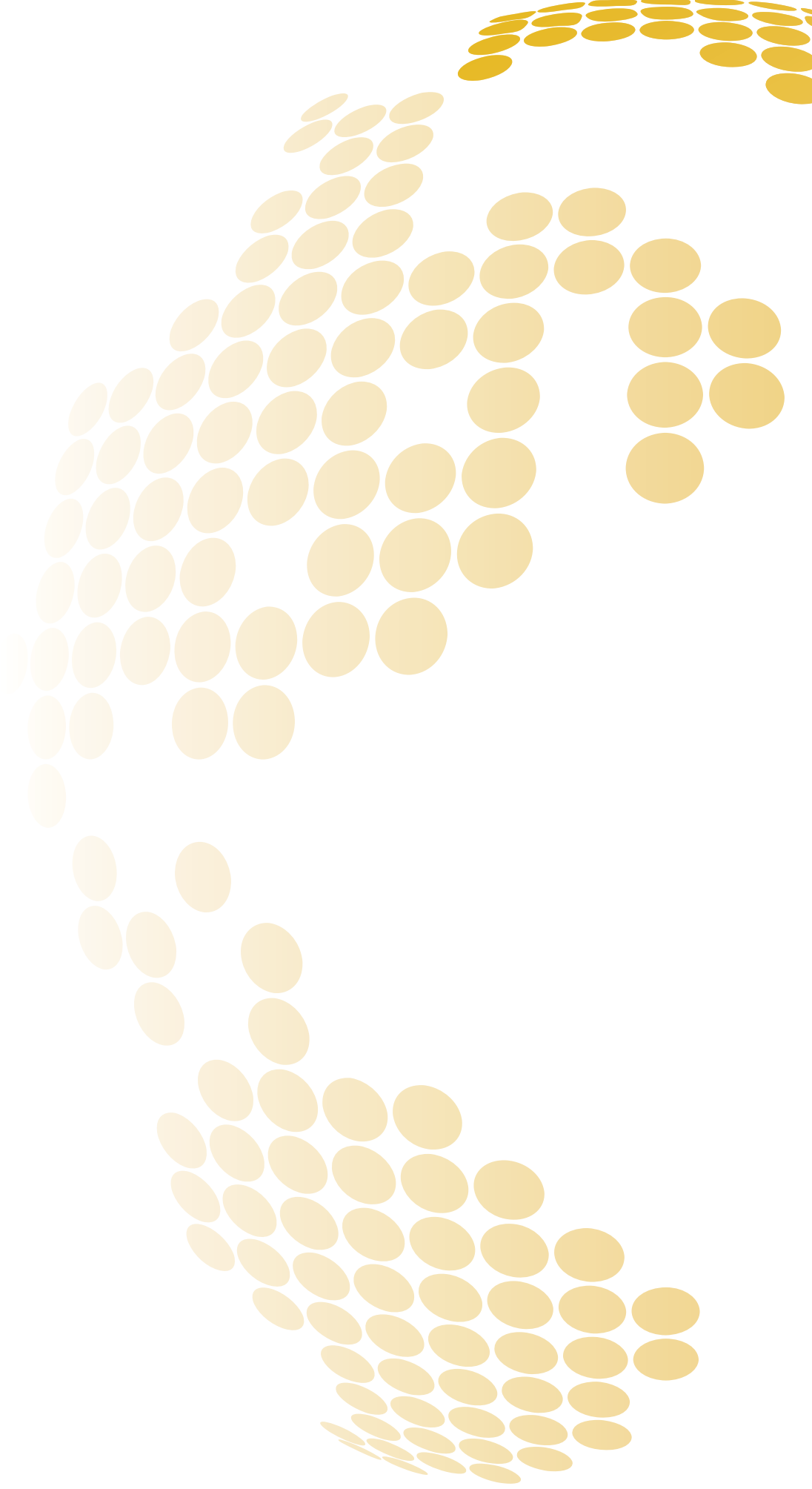
Linear regression (sometimes also referred to as least-squares curve fitting) will be familiar to most people who have taken statistics. It's one of the simplest and oldest methods for predictive analytics. The method finds the best fit mathematical line or curve through the middle of all the data. Regression can be extended to multiple dimensions and various mathematical curves (e.g., the polynomial regression found in Excel). As with every other method, creating too complex a model can run the risk of overfitting.

Time Series

In time-series methods, the features or dimensions in the dataset are typically the lagged values of a single variable (e.g., sales from last week, two weeks ago, three weeks ago, four, five, etc.) The CTL survey results suggested that these methods were popular. The methods often assume some weighting or statistical pattern for how historical values of different ages help forecast the next period's future value. The patterns in the lagged-values can also be used to predict future events (e.g., a progressive pattern of factory machine readings or vehicle sensor readings that precede a breakdown and downtime).

Conclusion

All of these methods have complex sets of issues, pros, and cons. The most basic issue being how they express the geometry of the multidimensional space of the data through rules, lines, formulas, curves, and so forth. More sophisticated methods are not always better. Dr. Winkenbach concluded by advising to never boil the ocean. If the problem is simple enough for a simpler method, then use a simple method.



MIT Center for Transportation & Logistics