AIP Conference Proceedings, Volume 3109, Issue 1, April 2024, Article number 030010
DOI:10.1063/5.0204918

Big Data Applications in Food Supply Chains

Emel Aktas[1, a)]

Author Affiliations
[1]*Cranfield University, Cranfield School of Management, College Road, Cranfield, MK43 0AL, United Kingdom*

Author Emails
[a)] Corresponding author: emel.aktas@cranfield.ac.uk

Abstract. Food supply chains are characterized by innovation, not only in products but also in processes. This paper aims to identify big data applications in the food and drink sector and present its findings as a state-of-the-art literature review. Academic databases were searched using 'food' or 'drink' and 'big data' keywords. Scholarly publications from 2015 onward are identified and presented in broad categories of demand prediction and retail operations optimization. The review recognized big data applications as a great opportunity for food supply chains. The applications aimed 1) to understand the customer base and inform marketing communications strategy, 2) to predict demand and organize retail operations to meet this demand, and 3) to optimize prices, assortment, and inventories based on demand patterns. Applications in this review focused more on descriptive and predictive analytics than prescriptive analytics, possibly due to the emergent nature of these applications. Descriptive analytics applications focused on capturing data, summarizing the status quo, and developing customer segments which can then be managed using varying marketing strategies. Predictive analytics applications focused on demand prediction with novel approaches proposed by the machine learning community. Prescriptive analytics applications aimed at promotion optimization and pricing for profit maximization. Cognitive analytics applications extracted customer reviews from online stores to inform which products should be marketed in what way. The review offers managerial insights on circumstances where big data analytics could prove beneficial. Managerial implications suggest that data integrators enable big data applications by ensuring the data collected are accurate, timely, and complete to inform descriptive, predictive, and prescriptive analytical models.

INTRODUCTION

The food and drink industry is among the EU's largest sectors, considering jobs as well as the value generated. It has doubled its exports to €90 billion over the last decade. The term 'food and drink industry' comprises food manufacturing companies, food retailers, restaurants, cafes, and other related businesses. Food legislation is harmonized across the EU single market, making it possible to produce in different EU countries under the same standards and sell in different countries without being subject to border checks. The industry is highly fragmented, with activities ranging from packaging to preparing, transporting, and serving food or drink products, excluding primary production.

The industry as a whole (including manufacturing and retail) employs 4.5m people with an average turnover of €3.7m per enterprise. With 290k enterprises and a total turnover in excess of €1.06t, the EU food and drink sector is 1.5 times the size of the sector in the USA [1].

The wealth of data available in the food and drink sector allows the building of advanced analytical models. Practitioners usually have preferences for a specific method when they face a business problem that requires modeling. This favorite method depends on not only the technical background and experience of the modeler but also the software available. With the growing research in analytics, many alternative models are available for a range of purposes. Big data can address many concerns in food supply chain management, among which food authentication is a growing area of application. Food authentication is an analytical process for validating the label information on the origin and production process [2]. Big data analytics is a promising avenue for the future of food authentication

owing to the amount of data collected and analyzed from multiple sources. Authenticity indicators are various. Four main methods of authentication are as follows: 1) Rare earth elements could be used for authentication along with precious metals such as gold, silver, platinum, and palladium. 2) Microbial fingerprinting establishes the microbes existing in food products to help geolocate their origin. 3) Similarly, metabolomic fingerprinting profiles the metabolites in the cells of fresh produce. 4) Finally, sensory analysis tries to correlate chemical data of food products sensorial attributes such as smell, taste, and texture.

'Local data production for local data consumption' is a concept for providing a range of Internet of Things (IoT) services without a mobile network coverage or a cloud-based storage server that can also process big data [3]. While local governments operate on limited budgets, they also have the relationships and communication efficiency in mobilizing local companies to build such platforms.

Artificial Intelligence (AI) and big data applications by the soft drinks giant Coca Cola include scanning social media and monitoring customer sentiment, reading product codes with TensorFlow for verifying procurement transactions, image processing of selfies taken by consumers to create an engaging experience, and using augmented reality in bottling plants to address operational issues [4].

This review aims to capture and present recent big data and analytics applications in food supply chains. For this purpose, a systematic literature review was conducted using relevant keywords. The papers identified were organized under retail operations such as customer segmentation, assortment planning, and availability (demand sensing, stock deployment, substitution, replenishment). The applications reviewed were built on internal data of the company (eg point-of-sales, customer footfall) and external data (eg social media, weather, or other publicly available data). The review develops suggestions on how different actors in the food supply chain (retailers, logistics service providers, manufacturers) can benefit from big data and analytics within and outside the company. It presents a mapping of descriptive, predictive, and prescriptive analytics to forecasting and optimization models, providing operational insights to food companies.

LITERATURE REVIEW

Big Data Applications Terminology

Big data are digitally produced in large quantities. They are collected via sensors and other devices connected to the Internet. It is possible to locate the origin of the data as well as when they were produced, in real-time or near real-time [5]. For a data set to be considered "big", it must be too large to be stored and processed using regular computers running end-user software. We witness an abundance of data generated by organizations as well as by persons who are using, for example, wearable devices. The digital world is overflowing with images, videos, and text generated by real persons, as well as AI chatbots.

5Vs and Sources of Structured and Unstructured Data

5Vs is a short-hand to describe big data. It refers to the "volume", which is the size of the data sets. The second V refers to "velocity", which refers to how fast the data are generated. The third V, "variety", is related to the multiple forms the data can take. Although data flowing from supply chain operations are mainly structured and transactional, it is now possible to incorporate unstructured data such as voice recordings, images, social media comments, along with readings of a range of sensors (eg temperature recorded at 15-minute intervals in a refrigerated container over its journey from China to Rotterdam). Many of the models for food supply chain management incorporate geospatial and temporal data, especially for demand forecasting and distribution operations planning. The fourth V, "variability or veracity" show how the same terminology may mean different things under different circumstances. For example, TSP frequently indicates the Travelling Salesperson Problem in the supply chain management context, but it can easily mean teaspoonful in the food preparation and cooking context. That is why when analyzing data, the analyst should never take variable names for granted and always verify with the owner of the data their meaning as well as their acceptable values. Additionally, many companies repeatedly report inaccuracies in the master data, which makes it difficult to rely on the data for making decisions. Such inaccuracies occur due to separate databases overlooking parts of the business but not having been integrated. A typical example is around the naming of products. Depending on the customers of a company, product names can change slightly, generating duplicate records in the database and making it quite challenging to access true sales information. Finally, the fifth V, "value", shows the return on investment in big data collection, storage, and analysis hardware and software. Big data has great potential to assist in supply chain

decision-making at strategic and operational levels. Investing in digital transformation to allow an entire big data architecture to operate at an organizational level requires this last V, "value", to be quantified as much as possible.

Descriptive, Predictive, Prescriptive, Cognitive Analytics

Over the last few years, "analytics" has become a popular term to represent methods that support decision-makers. Big data analytics, as a subcategory of analytics, refers to using existing methods and developing new ones that can handle big data sets to provide decision support to managers for a range of business problems. Figure 1 shows a pasta supply chain whilst introducing descriptive, predictive, prescriptive, and cognitive analytics applications. The supply chain is conceptualized like a waterfall, with raw materials flowing from primary producers (agricultural production) to manufacturers (processing, industrial refining) to retailers (distribution) to consumers. Activities at the beginning of this flow are referred to as upstream activities, and those closer to the end customer (consumer) are referred to as downstream activities. A manufacturer would be considered as downstream for a primary producer (eg farmer) and upstream for a retailer. Similarly, a retailer would be a downstream company for the manufacturer.

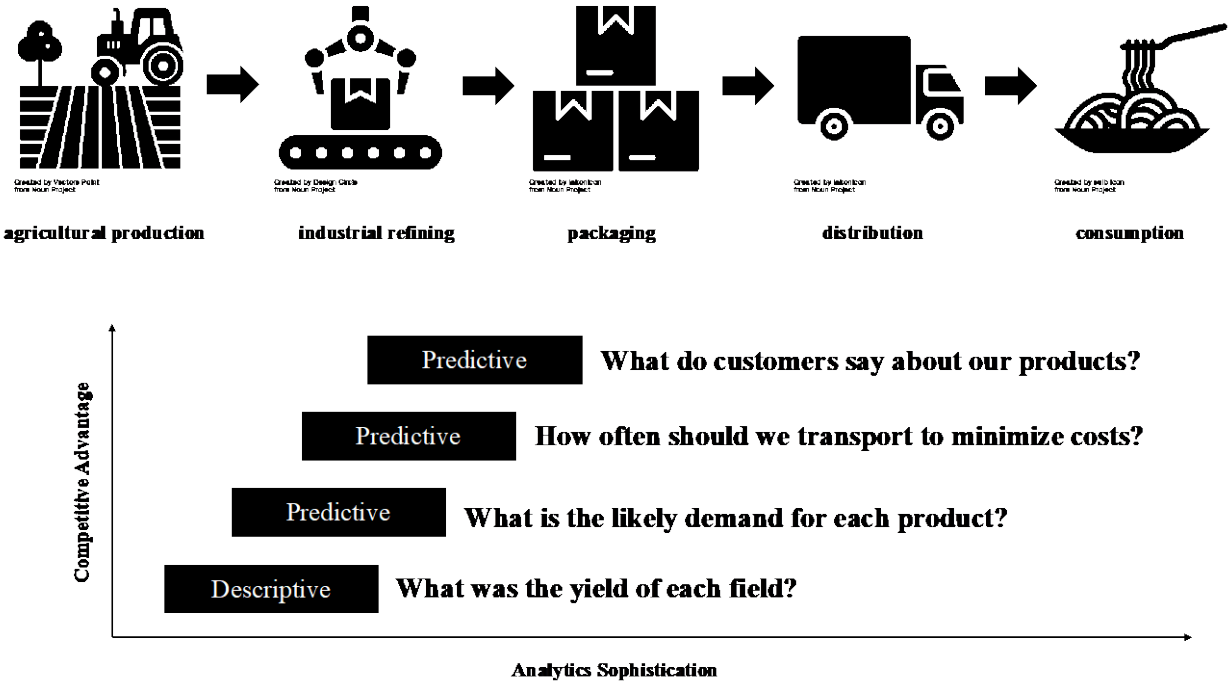

FIGURE 1. Examples of supply chain analytics

Descriptive analytics is concerned with answering what happened and why it happened. A typical question could be the yield from each wheat field and historical differences that can explain the variation in yields. Predictive analytics, on the other hand, is concerned with what might happen in the future. Typical time series models or causal models, such as linear regression or more sophisticated models, such as neural networks, can be used for this purpose. The question of forecasting demand for each product, which is typically referred to as a stock-keeping unit (SKU), would be answered by predictive analytics techniques. The next level of sophistication in analytical modeling is prescriptive analytics, which is concerned with how the desired level of performance can be achieved. The approaches in this domain usually comprise mathematical, linear, and nonlinear optimization and simulation models. For the conceptual supply chain given in Figure 1, the minimum transportation cost that meets the demand would be found by mathematical modeling, typically using a variant of the vehicle routing problems. The last level of sophistication, maybe the most difficult one, is cognitive analytics, which aims to learn dynamically from unstructured data such as social media posts or customer reviews left on a website. Natural language processing could be used to automate the process of understanding what customers are saying about a certain product category. It is most likely followed by data mining where customers could be grouped under meaningful categories depending on the reviews they share about the product on online platforms. Natural language processing means analyzing text generated by humans to

arrive at meaningful conclusions. It may provide insights on emotions shared by humans as well as relationships between different entities which are not immediately obvious. Typical applications include building chatbots for addressing consumer inquiries, analyzing social media posts to understand consumers' responses to new products, or using machine translation of the text to produce labels in multiple languages. In the context of food and drink supply chains, researchers are interested in consumer reviews to capture customer sentiment and predict demand.

Advanced predictive models can be used, for example, to estimate the age of Madeira Wine, one of the finest fortified wines from Portugal [6]. Wine age prediction helps the manufacturer to understand the complex wine chemistry and how the wine aging process develops. Hence, the manufacturer can monitor the process and evaluate whether the wines follow the desired path of aging and take action in cases of deviations from the expected aging process. Bottling starts once the wine aging process is complete. Descriptive analytics can be used to classify wines and detect frauds or adulterations.

Many alternative models are available for prediction purposes, and the model selection itself becomes a matter for the modeler. A predictive analytics comparison framework can be used to assist this 'model selection' decision [7]. The comparison framework has an analytics domain and a data domain, along with a comparison engine. The analytics domain is rich with four classes of predictive methods, each with several competing methodologies:

1. Methods to select variables such as a range of regression techniques augmented with genetic algorithms
2. Regression methods that handle multicollinearity (e.g., ridge regression)
3. Methods that measure latent variables (path models)
4. Ensembles that combine outcomes of decision trees.

This predictive model comparison framework facilitates model selection based on performance. This comparison framework is tested on wine age prediction using real-life data [7]. It is found that the predictive methods recommended by the framework depend on the predictors used.

Based on the data available for a predictive modeling exercise, the framework recommends a different model for the same prediction task [7]. What is more, applying the framework on various real-world data sets showed that the best modeling approach may be different from what is expected before the analysis is performed. It may use different variables. For example, quantified phenolic and furanic compounds in wine samples can be quantified [7]. It is not easy to predict how these compounds will interact with each other. Even new compounds may emerge over time. Their framework was then used to select the best-performing model to predict wine age. The framework can accommodate many competing models for the same prediction task, and practitioners can use it to select a high-performing predictive model.

METHODOLOGY

This state-of-the-art literature review focused on big data applications in food supply chains. The purpose of the review was to identify current big data and analytics applications and present a synthesis of how descriptive, predictive, and prescriptive analytics are used to improve operations and efficiency. To identify papers to be reviewed, the following terms were searched in academic databases such as Scopus and INFORMS PubsOnline.

- drink AND 'big data'
- beverage AND 'big data'
- food AND 'big data:
- 'point of sales' OR 'point-of-sales' AND 'big data'
- 'point of sales' OR 'point-of-sales' AND 'food'
- 'point of sales' OR 'point-of-sales' AND 'drink'

The review focused on papers published from 2015 onwards to capture the trends in big data applications. A total of 38 peer-reviewed journal articles, ten conference proceedings, and one book chapter were identified to be included in the review. The papers are organized based on the application.

FINDINGS

Demand Sensing

Real-time data from various supply chain processes can help capture short-term trends as they emerge so that prediction accuracy is improved. Demand sensing considers not only sales data but also external data that could be collected from social networks, websites, and online platforms. To be able to create an online platform, initially, the

ontology must be developed. In computer science and information science, an ontology describes the concepts and variables in a specific subject area. Categories and properties of data should be named and defined. How these categories interact with each other should be defined as relationships. An ontology could represent several stakeholders in the food supply chain: producer, buyer, educator, seller, supplier, transporter, planner, administrator, researcher, processor, investor, and consumer. For example, a consumer buys, orders, eats, enjoys, prefers certain foods and drinks, each of which has nutrition facts [8]. A consumer's personal preferences govern what foods and drinks he or she will consume. A consumer's health constraints would limit the types of foods he or she can consume. The power of such platforms is the ability to host hundreds, even thousands of food products for consumers and to leverage the big data collected from the platform to inform demand forecasts and devise logistics routes to meet the demand.

New Product Development

New product introductions are common for the food sector, and big data have a role to play in improving the new product development process. After the UK introduced the Soft Drinks Industry Levy, a tax on drinks that contain sugar to address concerns about increasing obesity and Type 2 diabetes in society, many soft drink manufacturers had to develop new products with less sugar. In response to the sugar tax, a fruit juice manufacturer incorporated big data analytics into the development of new products [9]. They collected nutritional data about the products, specifically the sugar content. They combined this nutritional data with prices from retailers and customer reviews from online portals. In terms of internal data, they used cost and the estimated environmental impact of a product in comparison to a range of alternatives already in the market. This big data analytics approach that combined a range of internal and external sources helped the manufacturer to decrease the new product development cost by 33%. Comparably, the new product development time was shortened by 11%.

Forecasting, Market Analysis, and Promotions

Demand forecasting in fresh food supply chains is particularly important since such products should be offered to consumers as fresh as possible before they start to deteriorate. Both underestimation and overestimation of demand affect the revenue with little chance of being able to sell products whose demand has been overestimated. The implications for food waste also raise ethical questions since the products have been grown, packaged, shipped, and displayed with huge water, energy, and carbon footprints, all detrimental to the environment. Large-scale data analysis improves the sales planning for perishable goods [10].

Customer segmentation is a typical marketing analytics exercise to management strategies that differentiate service provision. The key idea behind segmentation is to divide consumers into smaller, homogenous groups and then target each group with customized offers. While it is commonly used for segmenting consumers, it is also possible to segment customers in a B2B setting. Customer loyalty, which is associated with customer lifetime value, is assessed based on recency, frequency, and monetary attributes. It can be used to identify groups of customers based on their scores across these attributes [11].

The marketing literature is rich with discrete-choice models. These models have massive choice sets, such as those available in supermarkets that stock hundreds of products in many categories, including soft drinks. Random projection, a tool for dimension reduction in machine learning, can be used to estimate the parameters of these models [12]. Choice sets include products in different package sizes and categories. For example, the 330 ml Coke Zero can is a product sold in singles, in packages of four, eight,12, and 24. It is sold along with many other soft drink products, which also have various packaging alternatives. The size of the choice set in a model grows exponentially when we assume households purchase combinations or bundles of brands. Dimension reduction is needed to handle exponentially growing choice sets. It is the process of reducing the dimensions in a data set whilst retaining the most meaningful features. Dimension reduction in this context means aggregating all Coke Zero products.

Random utility maximization models are widely used in choice-based demand models [13]. These models assume that a customer's probability of choosing a product is proportional to the utility of that product. In other words, the customer rationally chooses the product with the highest utility. However, humans have bounded rationality, and their choices depend on the time available to make the decision as well as the cognitive limitations of the mind.

Multinomial logit models consider individual choices in a product category, analyzing the impact of market actions (price, promotions, product features) on choice [14]. These models assume that irrelevant alternatives are independent of each other. It is another way of saying that the probability of a customer choosing a product depends on the

perceived utility of this product in comparison to the utilities of other alternatives. Hence, eliminating some of the unchosen alternatives from the choice set shouldn't affect the selection of the product as the best option.

Multinomial probit model relaxes this independence of irrelevant alternatives assumption at an increased cost of estimation (it cannot scale as the choice set grows). Hence, the classical demand models such as multivariate logit and probit fail to accommodate large data sets due to their mathematical complexity. As a remedy, a conditional restricted Boltzmann machine can be used to capture consumer decision patterns [15]. The conditional restricted Boltzmann machine is an extension of the restricted Boltzmann machine; it estimates the dependence across categories purchased together [16].

Especially in the food retail context, the market basket of a shopper consists of many products purchased during a trip to the grocery store. The market basket at the end of the trip shows multicategory choices (bread and fruits, for example). The choices for products presented in multiple categories tend to be interdependent, which means that choosing a product in a category affects choosing another product in another category (eg hot dog buns purchased along with hot dogs). Two categories are considered complements if purchases in one category increase the purchases in another category. They would be considered competing products if the purchases in one category reduce the purchases in another.

The conditional restricted Boltzmann machine is an extension of [16]'s technique. It captures the time effects of purchases across multiple categories. It is possible to incorporate customer heterogeneity into the estimation procedure. This technique can model nonlinear relationships between products purchased by a large number of customer groups. For example, [15] were able to find out that the promotion of pasta was associated with an increase in purchases of pasta sauce and a decrease in purchases of soup, while the promotion of dog food was associated with a decrease in purchases of soup, pasta sauce, and pasta. For purchases in categories that do not depend on purchases in other categories, past purchases are expected to predict future purchases. Retailers can use this model to describe and predict the shopping behavior of individual consumers. Such predictions could be an input into personalized marketing activities.

Customer Sentiment Analysis

Advances in information technology have significantly changed how services are provided. The task boundary is shifted toward the customer; many tasks that were previously performed by the service provider (eg taking the order at a restaurant) are now performed by the customer [17]. This led to the coining of the term 'customer-facing technologies'. These technologies are at the forefront of the food service business as they facilitate customers' ordering and restaurant's fulfilling the orders.

Customer-facing technologies significantly affect customer experience, especially in the hotels, restaurants, and catering sectors. The number of restaurants with online and offline self-order technologies is increasing. One of these technologies, tabletop devices offered in restaurants, allows customers to change their dining experience by increasing the customer's participation in the food service provision [18].

Another application of these self-order technologies is online ordering via websites and mobile apps. These self-order technologies require the customer to place an order using their own device (mobile phone, tablet, or personal computer) and pay before receiving the service. These technologies reduce the waiting time for customers, but they do not necessarily lead to job cuts [19]. If customers are sensitive to waiting, these technologies will help manage customers' expectations. But for a fine dining experience, waiting time is expected; hence, the technology may even have an adverse effect on that type of food service business. Wait sensitivity refers to the impatience of customers and dissatisfaction with delays. Customers with high wait sensitivity would be dissatisfied when they need to wait to be served or service is delayed. It is proportional to income level, ie customers with high income are sensitive to waiting (they do not wish to wait to be served). If a restaurant serves customers with high wait sensitivity, it should use online self-ordering technologies (mobile apps) to manage service times. This will give the customer visibility of the service process and allow ordering in advance of arriving at the restaurant. For restaurants with customers who have low wait sensitivity, offline technologies such as self-service kiosks should be used [19]. These are in-store machines where customers can order food on their own. The customer is likely to wait for the machine to become idle before they can place their order in the restaurant.

The type of channel (online vs offline and self-service vs full service) affects the quantity and the quality of items purchased by consumers [20]. Systembolaget is Sweden's government-run monopoly seller of alcohol. In the stores of Systembolaget, the service provision was changed from a human serving the customer behind a counter to customers buying their alcohol using self-service machines. Using 160,168 orders of alcoholic and non-alcoholic beverages made by 56,283 customers, [20] demonstrated a disproportionately large increase in the sales of products

that had names that were very difficult to pronounce. The authors concluded that the increase in sales was due to the reduction in social friction thanks to self-service checkouts. A second experiment showed that when customers ordered pizzas online, they chose more complex products than the orders placed over the phone. This was another indicator of reduced social friction when using self-service ordering technology online. Such findings are critical to the design of an omnichannel experience for customers.

The Impact of Nutrition Information and Public Health Interventions on Sales

Food and drink market is characterized by a large number of products and rapid changes in what is on offer (eg new product introductions and discontinuations of underperforming products). Using nutrition data from food packaging, it is possible to link food consumption to nutrition consumption. Big data comes into play to harvest and analyze large volumes of data that are publicly available.

In a recent application, [21] collected 3,193,171 records of 128,283 products from six major UK supermarket websites using a weekly extraction routine written in Python. This database is available to answer questions around not only nutrition facts but also the dynamism in the introduction of new products and cancellation of existing products. For example, the salt, fat, and saturated fat content of ready meals sold at a lower price tier were much lower than full-price alternatives. The longitudinal analysis showed that nutritional formulation of the pizzas offered in these supermarkets changed for 11% of the products offered. The product turnover was also high: approximately 30% of the pizzas were either no longer available to purchase after six months or new product introductions.

The USA's Supplemental Nutrition Assistance Program (SNAP) was introduced to enhance the dietary quality of low-income Americans to curb obesity, diabetes, and other diet-related diseases [22]. Nielsen Homescan Panel produces a longitudinal data set that features food and beverage product purchases of approximately 60k households. The panel data includes volume, price, and retailer, but it cannot capture bulk-bought products due to a lack of packaging / unique product identifier codes in such purchases. Bulk products are those that are sold by weight, and without packaging; hence, they do not have a barcode and therefore cannot be included in the analysis. Typical examples are onions, apples, and grains sold by weight.

This data shows that households purchase healthy foods less often and in lower quantities [22]. The mean household purchases exceeded the recommendations issued by the Dietary Guidelines for Americans for sodium and saturated fat. Analyses such as this one can inform public health interventions promoting or incentivizing healthier products since high consumption of saturated fat and sodium is associated with poor health outcomes such as obesity or high blood pressure.

Fresh food and local food supply chains are becoming popular as the demand for fresh, organic, and nutrition-sensitive products increases [5]. Big data and analytics can connect industrial marketing (product, price, promotion, place) with food security (availability, accessibility, affordability, appeal) to develop strategies that affect consumer demand. Product, what is available on the market for sale, is related to the availability aspect of food security. For example, the availability of gluten-free alternatives is necessary for people suffering from coeliac. The place is associated with accessibility: do the people have access to fresh and nutritious products? Food deserts emerge in urban areas where it is difficult for residents to buy affordable and good-quality fresh food. Price is related to the affordability of food. Marketing decisions on pricing and discounts directly affect how affordable food products will be to consumers. Finally, promotions are related to the appeal of products. These exclude discounts, which are considered a part of price decisions. The frequency of advertising can affect the demand for specific food products.

Indeed, the impact of nutrition information on sales is a relevant question for food and beverage supply chains. Sugary drinks are defined as energy-dense, nutrient-poor, nonalcoholic water-based beverages with added sugar. Typical examples include sugar-sweetened soft drinks, sports drinks, and flavored water, which provide little nutritional value but increase the risk of obesity and other related health issues [23].

In an online experiment with 3,034 Australian adults, [23] investigated the impact of point-of-sale nutrition information about sugary drinks on customers' purchase decisions. The experiment also included health warnings about sugary drinks. Customers who saw the sugar level, the product's health star rating, or a health warning associated with the recommended daily sugar intake chose sugary drinks at a significantly lower rate. In this experiment, the participants were asked to imagine that they were about to buy a packaged drink for immediate consumption. The point-of-sale signs [23] used in the experiment were shown to positively impact consumer choices in a simulated environment since the participants were asked to imagine a scenario in which they were buying a drink to consume immediately in the presence and absence of the signage. The idea could be tested in actual retail and food service settings.

There are other policies to divert consumers' purchases from sugar-sweetened beverages to zero-sugar alternatives. The UK has introduced the 'sugar tax' in April 2018. The main rationale behind this policy was to help reduce the level of sugar in soft drinks and lower childhood obesity. In anticipation of this regulation, Jamie's Italian, a national restaurant chain, increased the price of nonalcoholic sugar-sweetened beverages by £0.10, corresponding to a price increase of around 3.5% [24]. Jamie's Italian informed customers that the additional £0.10 would be donated to the Children's Health Fund. At the same time, a documentary titled 'Jamie's Sugar Rush' was also aired on television. Analyzing point-of-sales data from 37 Jamie's Italian restaurants (2 million sales of soft drinks) [24] compared the sales of sugar-sweetened beverages before and after the introduction of the levy. They found that a customer, on average, bought 11% fewer sugary drinks after the price increase. Considering the scale of the business, this reduction had a significant impact on population health.

Retail Operations

Pricing

Supermarkets in the US carry, on average, 40k SKUs, and 35% of category sales (more than $129b) occur while products are on promotion [25]. Category sales represent the sales of products in different categories, such as soft drinks, cereals, dairy, etc. Promotion efficiency measures how much extra profit is generated by the promotion. While categories such as fresh seafood have low efficiency, categories such as frozen meat have high efficiency. The efficiency varies by store location, competitors, number of products in a category, and how often people buy these products [25].

Since a large proportion of fast-moving consumer goods are sold on promotion, planning sales promotions lends itself to analytical approaches such as optimization. Typically, it is the category manager's task to decide when to promote each product and which price level to use. Grocery category managers can use an optimization model with multiplicative and additive demand functions to estimate the promotion lift and post-promotion dip [26]. Grocery sales data shows each transaction with a timestamp and a list of purchased items. The demand models in [26] reached high accuracies (the Mean Absolute Percentage Error ranging from 11.5% to 18.7% for coffee, tea, chocolate, and yogurt), and the optimization model is expected to increase the profits by up to 5%.

Price promotions can be offered to all customers without any differentiation. The promotion's value and format can be customized for a customer when the customer uses a loyalty card. Using a data set of sixteen soft drinks brands, [27] ran a conjoint analysis with 790 respondents to model soft drink demand when the products had non-linear price promotions, some of which were in the form of quantity discounts. Producers of consumer packaged goods, such as soft drinks, cereals, and cooking oils, often offer several package sizes of the same product at a lower unit price for a larger package. This is the typical example of nonlinear pricing, and nonlinear price promotion is similar; the discount applied during the promotion does not need to be the same across all package sizes; it could be 50% off for a small package and 33% off for a large package.

Product substitution is a common phenomenon in food retail. Many manufacturers produce multiple substitutable products (consider Coke Zero vs Diet Coke), and coordination of price promotions should consider consumers' substitution patterns. Indeed, what to promote, when to promote, and how long to promote are current and relevant research questions. In the promotion planning area, [28] set out to answer the following research question: how should a company producing substitutes coordinate price promotions? To answer this question, a grocery chain in Chicago, Dominick's, provided data spanning eight years. Critical to answering this question was the presence of at least two different types of related products listed in the assortment. The analysis focused on the most popular product of a given category and the related products by the same firm. The first product differed only in package size (eg Cheerios 10 oz versus Cheerios 15 oz). The second related product differed both in size and some other characteristics (eg canned versus bottled soda). The data set had 24,222 observations of sales quantities and prices for 66 products in 22 categories. It was possible to show the link between price promotions of the most popular product affected the sales of the related products [28].

Assortment and Shelf-space Planning

Many factors affect consumers' ability to find and purchase products: eg, the number of products available in a category, in which aisle they are located, and where they are displayed on the shelf. Prescriptive analytics can be used to support shelf-space planning and maximize the return on shelf space allocated to products [29].

Retailers often choose to convert existing store formats. Conversion in this context means taking an existing store and refurbishing it in a way that its purpose changes. A typical example is Walmart, which converts existing discount stores into supercenters 60% of the time [30]. Discount stores are usually small and carry a few hundred SKUs, whereas supercenters are large and carry tens of thousands of SKUs. This conversion inevitably changes the assortment breadth and depth in the converted large formats. Walmart's supercenter conversion led to a 41% increase in weekly revenues, but the increase was attributable to the larger expenditure of the existing consumers rather than more customers [30]. The demand increase was more pronounced in food categories, mainly as a result of an increased number of purchases.

Inventory Planning and Fulfilment

Perishability is a typical aspect of inventory management in food supply chains, but it is also applicable to highly innovative sectors such as pharmaceuticals and consumer electronics. Maybe not perishability, but obsolescence is relevant for industries that witness changes in consumer preferences, examples of which include books, records, garments, or perfumes. When demand variability is high, the adaptive base stock policy is recommended [31].

Using aggregate, firm-level data, [31] found that inventory purchases depended on current sales, changes in sales forecasts, and sales growth. Firms with high sales growth tended to react less to forecasted changes in sales growth compared with firms with moderate sales growth. Moreover, purchasing constraints such as the minimum order quantities imposed by supply chain contracts or transportation efficiencies may affect inventory holding costs.

One of the first self-service applications was vending machines. With the advent of the IoT, it is now possible to monitor the sales rate of products sold from vending machines in real-time. Machine-to-machine communication between vending machines and cloud-based storage systems eliminates humans from the data flow process [32]. The real-time stock position of the vending machine can trigger a restocking activity and schedule it in the next round of replenishment visits.

As urbanization increases, many more people live in megacities, those with a population of more than 10m. Especially in megacities of emerging countries, consumers behave differently compared with those in developed countries. Small retailers fulfill the daily needs of these consumers [33] by removing some of the complexities associated with going to a superstore and picking a few products from aisles of hundreds of products. These small retailers, which are usually family-owned, independent stores, dominate the food retail sector in emerging countries, and the distribution of consumer goods becomes similar to the last-mile problem, which is concerned with distributing goods to end-consumers.

The relationship between retail distribution and sales is complex. Retailers stock products because they expect them to sell in high quantities and the products sell well because they are widely available to consumers [34]. Using monthly sales of alcoholic beverages sold in Systembolaget, [34] estimated the effect of store size on the market share of products sold. Product-level sales depended on the number of products stocked. Small stores stock fewer products, and the demand is distributed to fewer competing products; therefore, the volume per product is high enough to make the distribution efficient.

With the advent of e-commerce and mobile apps, on-demand meal ordering platforms such as Just Eat in the UK, Grubhub in the US or Uber Eats across the world have become part of consumers' dining experience. While individual restaurants cannot scale up as online orders and delivery requests increase, these platforms can achieve high efficiency and release restaurants from the burden of last-mile delivery. One of the most challenging problems in last-mile logistics is meal delivery because a typical order must be delivered within minutes after the food is ready. A meal delivery routing problem was formulated and tested with real-life instances from Grubhub in [35]. They conducted an extensive numerical study using instances from the Grubhub meal delivery routing problem instance set, which contained tens of thousands of instances derived from real-life historical data. Their formulation achieved up to a 17% reduction in click-to-door time, which is significant when considering that the majority of this click-to-door time is spent for the preparation of the ordered meal.

Supply chain coordination is a well-researched area in supply chain management with a range of mathematical models from contract design (eg wholesale price, revenue sharing, quantity discounts) to information sharing, which has been facilitated by advancing information technologies and increasing availability of data. Using data from a leading consumer packaged goods manufacturer in the US beverage and snack food sector, [36] showed that downstream sales data improved the forecast accuracy by 7% to 80% for all products included in the analysis where forecasts were produced using autoregressive integrated moving average time series method.

DISCUSSION

Companies that are considering adopting tabletop technologies need to factor in a transition period when customers will learn to use the device and possibly provide staff assistance in the first few instances of ordering. Such technologies not only improve the customer service experience but also increase the productivity of restaurants by increasing the average sales per check and reducing the meal duration [37]. This can be translated to reducing staffing levels of waiters without compromising service levels in restaurants.

Deep discounts reduce sales [38]. The negative effect of discounts is more pronounced in credence goods such as organic foods; retailers of such products should avoid offering deep discounts, which raise questions about the quality of the product.

When observed repeatedly, purchasing behavior is a good proxy for determining consumption patterns and developing marketing strategies [39]. Factors outside the control of the company, such as seasonal events (holidays and festivals) or product positioning (price and labels) lead to variations in demand. Factors representing customer characteristics such as education, age, gender, marital status, willingness to try new things, and experience determine a customer's loyalty. When loyalty is low, exogenous factors change purchase patterns. That's why considering these factors together improves forecast accuracy.

Considering the traffic effects in densely populated megacities, switching to a presale strategy could benefit many fast-moving consumer goods manufacturers and distributors in emerging markets. Without the data collected in [33], it is not possible to decide which strategy is better since pre-sales is costlier than van-sales at the outset, but when its impact on sales is considered, the conclusion is that the costlier strategy is also the most profitable because of the uplift in revenues.

A high level of substitution between products results in promoting one product at a time [28]. When customers prefer the brand over the product type, substitute products should not be promoted at the same time. In other words, if the consumers are loyal and buying the product for the brand, then the firm should not promote substitutes of the product at the same time to prevent the substitute product from stealing sales (cannibalizing) from the product of interest. Promoting substitute products at the same time should be considered if the product itself is more important than the brand of the product for the consumer.

Selected methods and purposes in descriptive analytics

- Nuclear density estimation was used to find out the concentration of food and beverage outlets in a city [40]
- Statistical analysis was used to investigate the impact of different distribution policies on profit [33], to establish the impact of store size on revenues [30], to understand which firms could benefit from coupons [41]. *Selected methods and purposes in predictive analytics*
- To predict demand, long short-term memory networks [42], conditional restricted Boltzmann machine [15], random utility maximization [13], time-series autoregressive integrated moving average models [10] were used

Selected methods and purposes in prescriptive analytics

- Optimization was used to find the best location for mobile collection points [43], to minimize the cost of meal deliveries whilst meeting service constraints [35], to minimize the expiration of perishable products [44].

Selected methods and purposes in cognitive analytics

- IoT module in vending machine sends data to AppServer & Web API to capture real-time stock information to minimize stock-outs [32].
- Natural language processing was used to understand customer preferences and sentiment [45]–[47].

CONCLUSIONS

This review identified big data applications as a great opportunity for businesses in the food supply chain. The applications aimed 1) to understand the customer base and inform marketing communications strategy, 2) to predict demand and to organize retail operations to meet this demand, and 3) to optimize prices, assortment, and inventories. Applications documented in this review employed descriptive and predictive analytics more than prescriptive

analytics. One of the reasons for this unbalanced representation of different analytical approaches could be the emerging nature of these applications.

Descriptive analytics applications focus on capturing data and summarizing the current status or developing customer segments which can then be managed using varying marketing strategies. Predictive analytics works focused on demand prediction. The machine learning community is devising novel approaches that outperform existing methods, for example, in prediction tasks. Prescriptive analytics approaches are mainly aimed at promotion optimization and pricing decisions for profit maximization. Cognitive analytics applications in the review extracted customer reviews from online stores to inform which products should be marketed in what way.

Food supply chain-related studies that use big data applications focus on marketing analytics and forecasting, most probably due to the availability of large data sets and well-established, robust models. These big data applications necessitate a strong information technology architecture as well as a willingness to integrate with upstream and downstream supply chain players. Investment in information technology alone is not sufficient to coordinate the supply chain and increase its performance. Firms in food supply chains can achieve excellent performance through data-driven and collaborative supply chain operations [48]. We are witnessing more and more cloud-based applications allowing real-time data sharing to achieve this data-driven and collaborative operations management.

Restaurants, cafes, and catering firms need to understand their customer base well and optimize their range and service accordingly. Certain kinds of economic activity could be inhibited by personal interactions because customers wish to avoid embarrassing situations such as not being able to pronounce the name of a product [20]. In the presence of such effects, deploying online ordering or self-service checkouts has the potential to increase sales.

'Data integrators', those companies that generate and integrate data from point-of-sale or IoT devices, play a key role in connecting demand with the focal company and have a great potential to facilitate big data analytics applications in forecasting, promotion planning, and revenue maximization.

Stakeholders in the food supply chains respond positively to technological advances that address key concerns of the sector: food safety and traceability, healthy and functional foods, and environmental and social sustainability. New technologies such as big data and smart stocking are used to reduce waste and optimize logistics and distribution.

The economic value of artificial intelligence in supply chain management is estimated to be in the order of hundreds of billions of US dollars, with the highest share in predictive maintenance, followed by yield optimization [49]. However, significant skill gaps exist in working with models and algorithms on large data sets. We are fast approaching a phase where many repeated decisions will be delegated to artificial intelligence-based automated systems, necessitating food companies to upskill their workforce.

Review studies are limited by the papers selected for the review pool. Future research could update the search and capture the papers published in the last year. The impact of the Coronavirus Pandemic on consumers' food purchase decisions lends itself as an appealing area for research.

REFERENCES

1. ECSIP, "The competitive position of the European food and drink in-industry," http://ec.europa.eu/DocsRoom/documents/15496/attachments/1/ translations, 2016, accessed: 2021-10-18.
2. G. P. Danezis, A. S. Tsagkaris, V. Brusic, and C. A. Georgiou, "Food authentication: State of the art and prospects," Current Opinion in Food Science, vol. 10, pp. 22–31, 2016.
3. Y. Shoji, K. Nakauchi, W. Liu, Y. Watanabe, K. Maruyama, and K. Okamoto, "A community-based IoT service platform to locally disseminate socially-valuable data: Best effort local data sharing network with no conscious effort?" in 2019 IEEE 5th World Forum on Internet of Things (WF-IoT). IEEE, 2019, pp. 724–728.
4. S. Vyas, S. S. Jain, I. Choudhary, and A. Chaudhary, "Study on the use of AI and big data for a commercial system," in Amity International Conference on Artificial Intelligence (AICAI). IEEE, 2019, pp. 737–739.
5. L. Dube´, A. Labban, J.-C. Moubarac, G. Heslop, Y. Ma, and C. Paquet, "A nutrition/health mindset on commercial big data and drivers of food demand in modern and traditional systems," Annals of the New York Academy of Sciences, vol. 1331, no. 1, pp. 278–295, 2014.
6. R. Rendall, A. C. Pereira, and M. S. Reis, "Advanced predictive methods for wine age prediction: Part I–A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods," Talanta, vol. 171, pp. 341–350, 2017.
7. R. Rendall and M. S. Reis, "Which regression method to use? making in-formed decisions in "data-rich/knowledge poor" scenarios–the predictive analytics comparison framework (PAC)," Chemometrics and Intelligent Laboratory Systems, vol. 181, pp. 52–63, 2018.

8. A. Kanak, ˙I. Arif, A. M. Karadeniz, and S. Ergu¨n, “Kebap: A semantic key enabling beverage and appetite platform enriched with virtual reality,” in IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 205–210.
9. S. Jagtap and L. N. K. Duong, “Improving the new product development using big data: A case study of a food company,” British Food Journal, vol. 121, no. 11, pp. 2835–2848, 2019.
10. J. Huber, A. Gossmann, and H. Stuckenschmidt, “Cluster-based hierarchical demand forecasting for perishable goods,” Expert Systems with Applications, vol. 76, pp. 140–151, 2017.
11. N. R. Maulina, I. Surjandari, and A. M. M. Rus, “Data mining approach for customer segmentation in B2B settings using centroid-based clustering,” in 16th International Conference on Service Systems and Service Management, ICSSSM, 2019.
12. K. X. Chiong and M. Shum, “Random projection estimation of discrete-choice models with large choice sets,” Management Science, vol. 65, no. 1, pp. 256–271, 2019.
13. S. Jagabathula and P. Rusmevichientong, “The limit of rationality in choice modelling: Formulation, computation, and implications,” Management Science, vol. 65, no. 5, pp. 2196–2215, 2019.
14. P. Aurier and V. Mejia, “Multivariate logit and probit models for simultaneous purchases: Presentation, use, appeal and limitations,” Recherche et Applications en Marketing (English Edition), vol. 29, no. 2, pp. 75–94, 2014.
15. F. Xia, R. Chatterjee, and J. H. May, “Using conditional restricted Boltzmann machines to model complex consumer shopping patterns,” Marketing Science, vol. 38, no. 4, pp. 711–727, 2019.
16. H. Hruschka, “Analyzing market baskets by restricted Boltzmann machines,” OR Spectrum, vol. 36, no. 1, pp. 209–228, 2014.
17. M. M. Davis, J. Field, and E. Stavrulaki, “Using digital service inventories to create customer value,” Service Science, vol. 7, no. 2, pp. 83–99, 2015.
18. A. M. Susskind and B. Curry, “An examination of customers’ attitudes about tabletop technology in full-service restaurants,” Service Science, vol. 8, no. 2, pp. 203–217, 2016.
19. F. Gao and X. Su, “Omnichannel service operations with online and offline self-order technologies,” Management Science, vol. 64, no. 8, pp. 3595–3608, 2018.
20. A. Goldfarb, R. C. McDevitt, S. Samila, and B. S. Silverman, “The effect of social interaction on economic transactions: Evidence from changes in two retail formats,” Management Science, vol. 61, no. 12, pp. 2963– 2981, 2015.
21. R. A. Harrington, V. Adhikari, M. Rayner, and P. Scarborough, “Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure,” BMJ Open, vol. 9, no. 6, p. e026652, 2019.
22. A. H. Grummon and L. S. Taillie, “Nutritional profile of supplemental nutrition assistance program household food and beverage purchases,” The American Journal of Clinical Nutrition, vol. 105, no. 6, pp. 1433– 1442, 2017.
23. M. Scully, B. Morley, M. Wakefield, and H. Dixon, “Can point-of-sale nutrition information and health warnings encourage a reduced preference for sugary drinks?: An experimental study,” Appetite, vol. 149, p. 104612, 2020.
24. L. Cornelsen, O. T. Mytton, J. Adams, A. Gasparrini, D. Iskander, C. Knai, M. Petticrew, C. Scott, R. Smith, C. Thompson et al., “Change in non-alcoholic beverage sales following a 10-pence levy on sugar-sweetened beverages within a national chain of restaurants in the UK: interrupted time series analysis of a natural experiment,” Journal of Epidemiology & Community Health, vol. 71, no. 11, pp. 1107–1112, 2017.
25. M. Trivedi, D. K. Gauri, and Y. Ma, “Measuring the efficiency of category-level sales response to promotions,” Management Science, vol. 63, no. 10, pp. 3473–3488, 2017.
26. M. C. Cohen, N.-H. Z. Leung, K. Panchamgam, G. Perakis, and A. Smith, “The impact of linear optimization on promotion planning,” Operations Research, vol. 65, no. 2, pp. 446–468, 2017.
27. J. R. Howell, S. Lee, and G. M. Allenby, “Price promotions in choice models,” Marketing Science, vol. 35, no. 2, pp. 319–334, 2016.
28. M. Sinitsyn, “Managing price promotions within a product line,” Marketing Science, vol. 35, no. 2, pp. 304–318, 2016.
29. T. Bianchi-Aguiar, E. Silva, L. Guimara˜es, M. A. Carravilla, J. F. Oliveira, J. G. Amaral, J. Liz, and S. Lapela, “Using analytics to enhance a food retailer’s shelf-space management,” Interfaces, vol. 46, no. 5, pp. 424–444, 2016.
30. M. Hwang and S. Park, “The impact of Walmart supercenter conversion on consumer shopping behaviour,” Management Science, vol. 62, no. 3, pp. 817–828, 2016.
31. C. R. Larson, D. Turcic, and F. Zhang, “An empirical investigation of dynamic ordering policies,” Management Science, vol. 61, no. 9, pp. 2118–2138, 2015.

32. R. Dijaya, E. Suprayitno, and A. Wicaksono, “Integrated point of sales and snack vending machine based on internet of things for self-service scale micro-enterprises,” in Journal of Physics: Conference Series, vol. 1179, no. 1. IOP Publishing, 2019, p. 012098.
33. Y. Boulaksil and M. J. Belkora, “Distribution strategies toward nano stores in emerging markets: The valencia case,” Interfaces, vol. 47, no. 6, pp. 505–517, 2017.
34. R. Friberg and M. Sanctuary, “The effect of retail distribution on sales of alcoholic beverages,” Marketing Science, vol. 36, no. 4, pp. 626–641, 2017.
35. B. Yildiz and M. Savelsbergh, “Provably high-quality solutions for the meal delivery routing problem,” Transportation Science, vol. 53, no. 5, pp. 1372–1388, 2019.
36. R. Cui, G. Allon, A. Bassamboo, and J. A. Van Mieghem, “Information sharing in supply chains: An empirical and theoretical valuation,” Management Science, vol. 61, no. 11, pp. 2803–2824, 2015.
37. T. F. Tan and S. Netessine, “At your service on the table: Impact of tabletop technology on restaurant performance,” Management Science, vol. 66, no. 10, pp. 4496–4515, 2020.
38. Z. Cao, K.-L. Hui, and H. Xu, “When discounts hurt sales: The case of daily-deal markets,” Information Systems Research, vol. 29, no. 3, pp. 567–591, 2018.
39. M. Kunc, “Market analytics of the rice wine market in Japan: An exploratory study,” International Journal of Wine Business Research, vol. 31, no. 3, pp. 473–491, 2019.
40. X. Lu, H. Xing, M. Yu, and Y. Xu, “Study on spatial distribution of commercial network in Jinan based on POI information: Taking food and shopping network for example,” in Proceedings of the 2nd International Conference on Big Data Research, 2018, pp. 178–181.
41. I. Reimers and C. Xie, “Do coupons expand or cannibalize revenue? evidence from an e-market,” Management Science, vol. 65, no. 1, pp. 286–300, 2019.
42. X. Liu and R. Ichise, “Food sales prediction with meteorological data—a case study of a Japanese chain supermarket,” in International Conference on Data Mining and Big Data. Springer, 2017, pp. 93–104.
43. C. K. Glaeser, M. Fisher, and X. Su, “Optimal retail location: Empirical methodology and application to practice: Finalist–2017 M&SOM practice-based research competition,” Manufacturing & Service Operations Management, vol. 21, no. 1, pp. 86–102, 2019.
44. A. Akkas, V. Gaur, and D. Simchi-Levi, “Drivers of product expiration in consumer packaged goods retailing,” Management Science, vol. 65, no. 5, pp. 2179–2195, 2019.
45. D. Puranam, V. Narayan, and V. Kadiyali, “The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors,” Marketing Science, vol. 36, no. 5, pp. 726–746, 2017.
46. I. Ip, S. Fong, Y. Zhuang, and R. Wong, “Competitive intelligence study on Macau food and beverage industry,” in 7th International Conference on Cloud Computing and Big Data (CCBD). IEEE, 2016, pp. 170–174.
47. A. K. Varudharajulu and Y. Ma, “Feature-based restaurant customer reviews process model using data mining,” in Proceedings of the 2018 International Conference on Computing and Big Data, 2018, pp. 32–37.
48. M. Irfan and M. Wang, “Data-driven capabilities, supply chain integration and competitive performance: Evidence from the food and beverages industry in Pakistan,” British Food Journal, vol. 121, no. 11, pp. 2708– 2729, 2019.
49. M. Chui, J. Manyika, M. Miremadi, N. Henke, R. Chung, P. Nel, and S. Malhotra, “Notes from the AI frontier,” https://www.mckinsey.com/featured-insights/artificial intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning, McKinsey, 2018, accessed: 2021-10-18.

Big data applications in food supply chains

Aktas, Emel

2024-04-09

Aktaqs E. (2024) Big data applications in food supply chains. AIP Conference Proceedings, Volume 3109, Issue 1, April 2024, Article number 030010
https://doi.org/10.1063/5.0204918