machines

MDPI

Article

Practical Classification and Evaluation of Optically Recorded Food Data by Using Various Big-Data Analysis Technologies

Tim Jarschel [1,*], Christoph Laroque [1,*], Ronny Maschke [2] and Peter Hartmann [2]

[1] Department of Economics, University of Applied Sciences, 08056 Zwickau, Germany
[2] Faculty of Physical Engineering, University of Applied Sciences, 08056 Zwickau, Germany; Ronny.Maschke.bl6@fh-zwickau.de (R.M.); peter.hartmann@fh-zwickau.de (P.H.)
* Correspondence: tim.jarschel@fh-zwickau.de (T.J.); christoph.laroque@fh-zwickau.de (C.L.); Tel.: +49-375-536-3448 (T.J.); +49-375-536-3221 (C.L.)

Received: 6 May 2020; Accepted: 11 June 2020; Published: 16 June 2020

Abstract: An increasing shortening of product life cycles, as well as the trend towards highly individualized food products, force manufacturers to digitize their own production chains. Especially the collection, monitoring, and evaluation of food data will have a major impact in the future on how the manufacturers will satisfy constantly growing customer demands. For this purpose, an automated system for collecting and analyzing food data was set up to promote advanced production technologies in the food industry. Based on the technique of laser triangulation, various types of food were measured three-dimensionally and examined for their chromatic composition. The raw data can be divided into individual data groups using clustering technologies. Subsequent indexing of the data in a big data architecture set the ground for setting up real-time data visualizations. The cluster-based back-end system for data processing can also be used as an organization-wide communication network for more efficient monitoring of companies' production data flows. The results not only describe the procedure for digitization of food data, they also provide deep insights into the practical application of big data analytics while helping especially small- and medium-sized enterprises to find a good introduction to this field of research.

Keywords: big data; data analysis; food industry; industry analytics; nutrition; laser triangulation; food analysis

1. Introduction

The digital permeation of all areas of our lives and the digital transformation associated with it influence all sectors of the economy. In this context, the impact of new digital technologies on one of the most important industrial sectors in Germany is also increasing: the impact on the food industry. There are 600,000 employees in about 6000 companies that ensure every day that sufficient food is produced in Germany. However, food is not only produced for the people who live within the country, Germany is also the third largest food exporter on the world market. An amount of 33.5% of all food produced in the country is exported to other countries all over the world. In 2017, the industry was able to increase its sales by 4.8% to around 179.6 billion euros compared to 2016 [1]. In particular, the quality of the food is of the utmost importance. Inside and outside of Germany, people have very high expectations and demands on the properties of their food. Therefore, it is really important that the quality of the food that is produced is maintained or even further improved in the future by monitoring the production chains even better. Automated production chains as part of the general digitalization of production techniques can help to successfully meet the above-mentioned challenges and the result of increasing customer demands in the food industry in the future [2].

In order to remain competitive, companies cannot ignore digitization. A large-scale study from summer 2017, carried out by the holding company "Deutsches Interim Management" (DDIM), showed that small- and medium-sized companies in the food industry will only digitalize their processes and effectively promote the appropriate processing of their own data if external circumstances force them to do so [3]. Seventy-eight executives of food producers were interviewed in the course of this study. A large number of them mentioned that it is usually the customers or the suppliers, or even the legislators, who force companies to take action to create the design of digital restructuring processes. In addition, the study illustrated that the companies in the food industry also believe that the market conditions within the entire food industry will change significantly in the future. The trend towards digital networking of all areas of life and the economy will not spare the food industry either. Discussions are already underway as to whether product life cycles in this industry will also shorten in the future and whether the trend towards individualized food products will continue to increase. The accompanying logistical change in combination with the further growing market dynamics will also lead to a complete change of the existing proven structures in the logistics sector of this industry [3].

Based on these circumstances, the Saxon state government defined an innovation strategy in 2017 regarding to the cross-sector future field of "digital communication". One of the sub-projects in this future-oriented field focuses on advanced production technologies in the food industry. Within the scope of this subproject, a measuring system based on laser triangulation has been established in the field of optical detection of food, which can measure static food objects in three dimensions and examine them for their chromatic composition. Based on this, a cluster-based back-end system for the efficient collection, storage, and processing of the data to be analyzed from the various information interfaces was set up within the framework of the analysis and processing of the recorded sensor data. This can be used in a practical environment as an organization-wide server network for monitoring and visualizing of your own production data flows.

2. Related Work

In recent years, digitization has helped to generate, collect, and store many different types of data at an unprecedented speed. The associated benefits are illustrated in an industrial context by using the six basic principles of Industry 4.0 (interoperability, real-time operability, virtualization, decentralization, service orientation, modularity) [4,5]. Traditionally, production data from almost all industrial sectors has been stored in simple databases or data warehouse systems [6]. However, with the emerging technologies of Industry 4.0 (sensors, actuators, industrial computers, robots, wireless devices, cyber physical systems, programmable logic controllers), the amount of data generated is becoming larger and more complex [5]. Many industries, such as logistics or finance, are already taking full advantage of the available technologies and facing the challenge of processing huge amounts of data through cluster-based big data architectures [7,8]. Both, operational as well as strategically interesting application examples in the areas of preventive maintenance, production planning and monitoring, quality assurance and product planning have largely emerged from the technologies of the Industry 4.0 [5]. The interests of many companies from various industries in using business intelligence and big data analytics applications has been growing continuously in recent years [4].

However, the situation is quite different in the food industry. Experts agree that the manufacturers in the food industry should use their own company data much more effectively to support their own decision-making processes. The interrelationships in large amounts of data allow companies to uncover unrecognized potential at the enterprise level, instead of having to look constantly for ways to improve production performance further at the local level [9]. At this point, it is important to stress, that the data of the food industry (scientific data, genomic data, technical data, business data) are as diverse as the industry itself. Big data analysis technologies help to make the processing and evaluation of large amounts of data possible in order to implement intelligent model-based process optimization in the company's operation tools [10]. One of the biggest challenges in dealing with large amounts of data is the variability of the data. Bringing together a wide variety of data sources and making them usable is

probably one of the greatest challenges for the food industry. It is also important to emphasize that this paper analyzes food production data and no customer or logistics information from the food industry. The production of many foods has been highly automated for many years. Especially in some areas of food production, up to 98% of the processes are automated. In some other areas of this industry, only 20% of the production steps are automated. This is a huge contrast and especially in highly automated areas such as the beverage industry, the scope for improvement in production is extremely small. This could be another reason why researchers initially focused on other areas of the food industry rather than on the production part [11]. These two main reasons could be responsible for the fact that so far only a few functioning application examples have been put into practice in some production areas.

But even if the practical use of big data analysis technologies in the food industry is not yet as pronounced as in other industries, there are at least some scientific papers that discuss the basic application areas of the new digital technologies in connection with the possibilities of big data in the food industry. A contribution by researchers from Spain describes the more efficient use of own resources and an improved traceability of the entire value chain, by the development of a digital platform for the monitoring of dairy cattle and feed grain. The information is collected by using edge computing methods (sensors and microcomputers) and evaluated by using big data analysis technologies [12]. Mohamed-Amin Benatia and his team deals with the development of a data-based traceability system architecture to improve food traceability in case of critical incidents along the supply chain and to prevent possible damage to the company's image or high recall costs [13].

In the literature, the potential of big data and business intelligence systems in the food industry is repeatedly referred based on possible areas of applications, such as the implementation of a production management control system. It also pointed out that there are very few good practice examples in this context that have been successfully implemented in reality [14]. This shows that the possibilities of big data in the food industry are slowly being recognized by the companies. However, a closer look at the current literature on the subject reveals the following: due to the large number of different data sources in this industry, many researchers only deal with very specific, niche-like applications. Big data is often discussed, but not every company benefits from the research results. The reasons for this are, that the research work is usually very specific and only the large corporations are in possession of such large data sets, which form the basis for the analyses. This is illustrated by the fact that in recent years more and more literature has been published with regard to increasing agricultural yields by evaluating data from the largest food companies [15,16]. In this context, researchers are also looking at how climate change will affect the profitability of large producers in the future and how these developments can be counteracted in certain regions with regard to the cultivation of modified crops [17]. The use of data for decision making in large agricultural groups is highly developed. Possible applications include detecting plant resistance, predicting flooding and the understanding of pest outbreaks. Attempts are being made to improve risk management and damage control with regard to food cultivation. In addition, food industry groups also evaluate consumer data to understand which foods are particularly popular and why [18]. A completely new field of research in data analysis in the food industry is the analysis of text data. The analysis of text data is intended to give companies more information about food safety and trends in consumer opinions. In addition to the analysis of consumer sentiment, it can also be used to characterize new nutritional patterns in society, on the basis of which new food products can be developed in a targeted manner [19]. However, it is also clear that research in this area has always developed in very specific directions. For instance, a company that has automated only 20% of its own processes will not benefit much from these methods of data analysis. For this reason, this paper deals with a procedure which describes the entry into the development of a multi-purpose big data architecture, from which even small companies can profit in any case across all sectors.

However, it must also be stressed that more and more young start-ups are using modern digital technologies to influence the major food producers, for example, by developing online portals for food traceability to improve data exchange along the value chains [20]. The more efficient use of

their own resources and the increased food safety and quality are exemplary benefits that businesses can already take advantage of. Also, in the area of food product development, the use of big data analysis technologies has been proven to reduce the cost and time required to bring the product to market [21,22]. Big data analysis technologies combined with cloud computing and machine learning have been shown to improve cold chain management along the supply chain, or the systematic planning of reliable raw material demand forecasts [23,24]. The fourth wave of industrialization has reached the food industry and it is important that not only the large food companies benefit from this journey, also small businesses have the right to benefit from this progress, even when they are just starting to analyze their own data.

For inline data acquisition and to feed the big data architecture, a fast and robust detection system is necessary. The direct detection of food products is often done using imaging sensors, spectrum analyzers, spectroscopy, or hyperspectral imaging [10,19]. In addition to the methods just mentioned, the principle of laser triangulation is another very popular method for contactless detection of the geometries of different food objects. Within the scope of this paper, the principle of laser triangulation is chosen, as during the whole project, the acquisition system is used in an inline process. To ensure a high data transfer rate, rapid data acquisition is also necessary. These requirements can be implemented robustly and cost-effectively with the use of the laser triangulation acquisition system [25]. During the process of the laser triangulation, a point is irradiated at an angle onto an object and the offset of the reflection is tracked on a CCD line. The height of the object determines the offset on the detector. The functional principle is shown in Figure 1 [26].

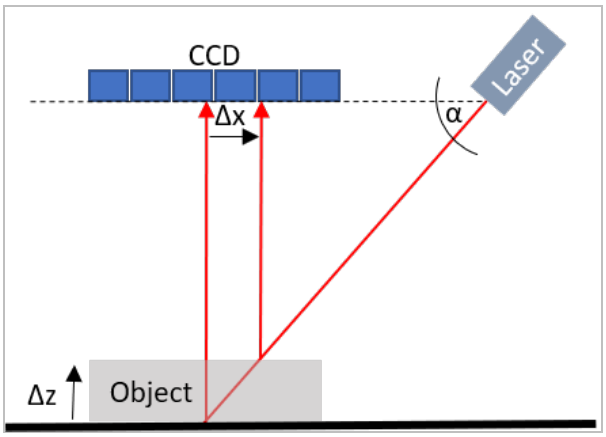

Figure 1. Principle of laser triangulation.

To avoid the influence of perspective, the camera is positioned vertically above the object and the laser is directed at the ground at an angle α. During the adjustment of the system, precise knowledge of the angle α between the lens plane and the laser is crucial. With the relation:

$$\tan \alpha = \frac{opposite\ leg}{adjacent\ leg} \tag{1}$$

follows:

$$\Delta z = \tan \alpha \tag{2}$$

The light-section method as a part of structured illumination basically works according to the principle of laser triangulation and offers the possibility to capture all three dimensions of an object with only one system. A line laser is used instead of the dot laser and a two-dimensional CCD array (camera) is used instead of the CCD line. The camera records the object to be measured in its full width. By means of edge detection, the area of the distorted laser line is detected and measured. It is possible to acquire the information of the height for each point on the line (depending on the resolution of the CCD array) [26].

In addition to methods for the optical recording of geometric and color properties of food objects, researchers are also concerned with the evaluation of digital text data generated in the food industry today. Related use cases are mainly aimed at improving food safety, preventing food fraud, identifying customer nutritional patterns, and evaluating consumer opinions so that the development of new products can be adapted to customer needs [19]. However, most applications of big data analysis technologies in the food industry can be found in product monitoring and in the general improvement of the efficient use of existing resources. Chinese researchers have demonstrated how big data analysis technologies can be combined with the use of a cloud-based data management system [27]. The merger has made it possible to bring together information technology infrastructures of many poultry farms in China. On that basis, important information about the egg production, the flock performance, and the housing conditions of individual farms can be monitored centrally via one system [27]. This is the evidence that in addition to the improved quality of the food products for humans, the housing conditions of the animals and their welfare can also be improved by using big data analysis technologies. The use of the technologies mentioned above and the associated advantages in the food industry, thus also make a significant ethical contribution, which can hardly be demonstrated in any other industry in such a short time after the integration of the corresponding technologies. However, the prerequisite for all of this is that all companies benefit from the successful use of big data technologies, including small- and medium-sized businesses.

3. Experimental Design and the Data Set

In addition to the acquisition of food data regarding dimensions and RGB color values, a near-real-time visualization and evaluation of the data sets should be made possible. The following products are used as food test objects: a green apple, a red apple, a clementine, and a tomato. The test objects we use are optimally suited for the research purposes we pursue due to their long-term and non-cooling-dependent durability and their general diversity in terms of color and surface structure.

The food is recorded by means of laser triangulation. A high-resolution camera positioned vertically above the foodstuff records the laser line distorted by the food. A filter in front of the lens increases the sensitivity of the camera to the laser wavelength (λ = 635 nm). In addition, only the red channel of the camera is used in the software to evaluate the dimensions. This ensures a maximum signal to noise ratio. The edge detection algorithm provides the width of the object to be measured. The deviation of the line from the zero-position calculated with the angle at which the laser is directed at the object delivers the height information of the object. As soon as an object is detected under the camera, a timer starts to count, which is automatically stopped when the object has passed completely. Multiplied by the belt speed at which the food is transported, the information on the length of the object can be calculated. At this point, a region of interest (ROI) is queried in the center of the object and the average value of the three-color channels, red, green, and blue, is formed. The data is stored as a CSV-file and simultaneously sent to an ADC in order to be able to control peripheral devices with the data, or to make the acquired data universally available at the hardware level.

Each newly captured object is saved line by line in the CSV-file regarding to its length, width, height, and the corresponding three RGB color values. The following Table 1 shows the structure of the data set, which forms the basis for further processing of the recorded data in the subsequent analysis procedure.

It has to be mentioned that besides the CSV file format, used during the research work, there are other data formats that are equally well suited for processing large amounts of data. It would also be possible to forward the data in JSON or XML format. The advantage of these two file formats is the simplicity of implementation and the variety of programming languages that can be used to parse data in these two formats. However, a big disadvantage of JSON is the fuzzy definition of the numbers in this format. Therefore, JSON is the basis for many big data projects to form texts out of structured data. However, this is not relevant for our use case. Compared to the CSV file format, the XML file format has a very complex syntax. In many use cases, the code must be rewritten via XML parser so

that the user can continue working with the data afterwards [28,29]. Working with an open source database management system such as MySQL or MariaDB is also possible. A fast reading and writing process of the data is also guaranteed by these technologies. However, the project team decided to use Elasticsearch, which is also a database management system. At the same time, the browser-based analysis tool Kibana provides the ability to analyze all of the data in the Elasticsearch cluster. Such features are not available with a MySQL or MariaDB database technology without having to establish external connections to other programs. In addition, when used commercially, licensing MySQL or MariaDB is hardly affordable for many smaller companies [30].

Table 1. Data record structure of the measured value data.

Time	Length	Width	Height	Red	Green	Blue
2019-11-17T07:00:37.953000	94.34	79.54	84.60	134.25	138.76	58.444
2019-11-17T07:01:38.008638	93.77	79.42	84.52	153.25	157.25	56.46
2019-11-17T07:02:38.094749	93.77	79.54	84.79	134.25	137.01	51.01
2019-11-17T07:03:38.146412	94.93	79.30	84.58	155.01	157.75	61.70
2019-11-17T07:04:38.232388	94.64	79.44	84.59	132.75	136.75	53.72
2019-11-17T07:05:38.317754	94.34	79.39	84.52	152.77	156.51	63.70

In order to gain the best insights from the recorded data, a fully functional large data architecture was set up, which allows data analysis to be carried out almost in real time in order to relate it to recorded values far from the past. The basis of any modern large data architecture is a large storage system for backing up and managing data from different sources in their respective raw formats. In this case, the researchers used a cluster-based solution because, especially in the industrial sector, the amount of data produced is usually so large that even the most powerful computers have difficulty in ensuring efficient data processing. Another important criterion is to ensure automated data integration of the produced data sets into the large data cluster during production operation. In addition to this feature, the overall structure must also offer various possibilities for visualizing the data in the cluster in order to ensure near-real-time data monitoring.

For the requirements set, the Elastic-Stack is used as an optimal basis and is built according to the specifications on three Linux servers [31]. The three core components Logstash, Elasticsearch, and Kibana jointly represent the basic functions of the Elastic-Stack. The core of this big data architecture is Elasticsearch, which functions as a distributed data management system. This tool is based on the Apache Lucene program library, which is similar to an indexing search algorithm. This means that the data is not loaded into the file system by Logstash in its raw format. It has to be parsed beforehand in such a way that it can easily be included in Elasticsearch's data management system. The Logstash tool is used as a persistent data bridge between the CSV-file with the sensor data and the data management system of Elasticsearch. It ensures permanent access to the corresponding CSV-file and at the same time Logstash automatically parses permanently recurring new entries in the CSV-file and forwards them to the Elasticsearch cluster. This is a great benefit of Logstash in cooperation with Elasticsearch, because it creates an automated data flow between both. With the help of different analysis tools, integrated in the browser-based visualization tool Kibana, the data indexed in the Elasticsearch cluster can be examined. In interaction with the data management system, every user of the Elastic-Stack in the same network can create meaningful graphics and illustrate complex issues within the data sets. In summary, Elasticsearch as a database system on the three Linux servers in cooperation with Logstash and Kibana represents the big data analysis technology, which has already been mentioned several times in this paper. The system can be considered as a big data architecture because the size of the indexed data is not a limiting factor. The Elasticsearch database can be extended optionally by adding more Linux servers.

The biggest advantage of the whole Elastic-Stack is its high flexibility. It can be installed on a local machine or in a large cluster of industrial servers, which means that processing several terabytes of data is no challenge. The software runs very stable and is permanently provided with new security

updates. Elastic is also an open-source tool. The software can be downloaded and installed on the user's own infrastructure without paying license fees, which are charged by many other software vendors. The Elastic stack also offers machine learning capabilities and many API clients for popular programming languages like Python, JavaScript, Java, PHP, or GO [31,32]. It must also be emphasized that it is very important that the presented food use case can be transferred into practice as easily as possible, so that the greatest possible number of other researchers can benefit from the project findings themselves.

In addition to using the Elastic-Stack as a cluster-based big data architecture, a Python program is used to visualize the RGB color values. This program automatically reads the RGB color value data from the CSV-file and converts it into a representation that enables the human eye to recognize a concrete color derived from the numerical values. Based on the procedures described, the test setup shown in Figure 2 is required for the optical recording and evaluation of the food data.

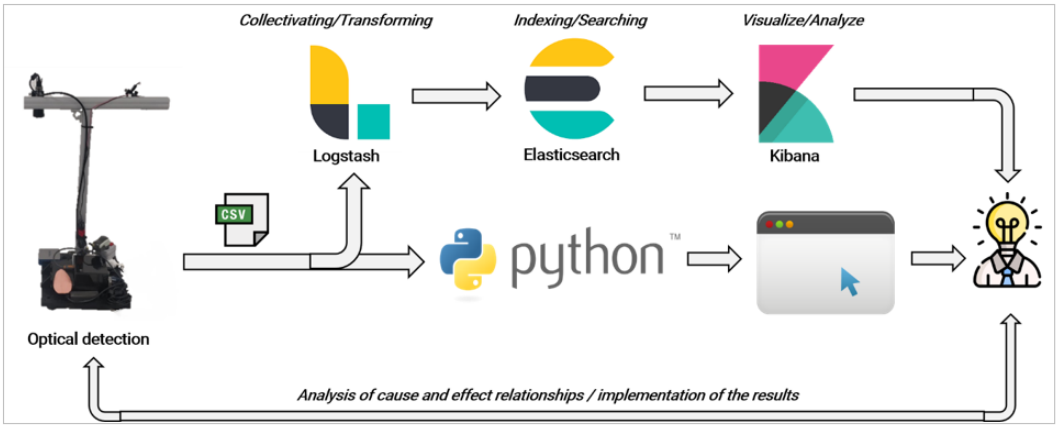

Figure 2. Analysis procedure—recording and evaluation of the measured value data.

4. Methods Applied

For laser triangulation, a camera and a line laser are vertically directed above the object at a known angle to measure the object. Due to the angle, there is an offset in the direction of the viewing direction of the laser, which correlates under an angular relationship with the height of the measurement object. At the beginning of the measurements the system must be calibrated to the conditions. For this purpose, a standard is used whose dimensions are known to the software. The control of the components and the calculation of the dimensions is done with LabView from National Instruments. The control of the ADC and the writing of the CSV-file to the local drive have also been realized with this software environment. If the camera detects a non-zero height, the timer for recording the length information is started and the maximum values of all three dimensions are stored. After the measuring object has passed the laser, the color channels are scanned, and the data storage is initialized. The color channels are each coded with 8 bits. In order to increase repeatability, the color value acquisition is carried out under homogeneous artificial lighting. A pause of currently at least 160 ms is required between the measurement objects to guarantee a smooth process. To transmit the six measured values (L, W, H, R, G, B), each voltage value is present for 10 ms to give the peripheral hardware enough time to read. Depending on the hardware used, time can be saved in this process.

After the object has passed the laser, i.e., the system detects zero height and the system collects the color of the object, it writes to the CSV-file and outputs the data on the ADC in the order, length, width, height, red, green, blue. The output at the ADC is done via 2 analog channels. The first channel counts as an integer from 1 to 6 V, while at the same time the corresponding measured values in a range from 0 to 10 V are present on the second channel. This universal coding into a DC voltage can be interpreted by any peripheral ADC component. Due to the common ground, communication takes place via three lines.

The data in CSV format tapped by Logstash are restructured beforehand within the experimental setup so that the data sets to be indexed can be transferred to the Elasticsearch cluster for later visualization in Kibana. The measurement data are used as raw data to subdivide the objects and their associated data into individual clusters based on specific criteria. The SAS JMP Pro 13 software is used to identify specific factors or parameters within the data, which can then be used to differentiate objects by their dimensions and color values. As a first method to classify the data, 3D scatter diagrams were created. For illustration purposes, in the following Figures 3 and 4 the data concerning the RGB color values and the values of the dimensions are visualized in a 3D scatter diagram in order to clearly divide them into the corresponding groupings of the four different food products.

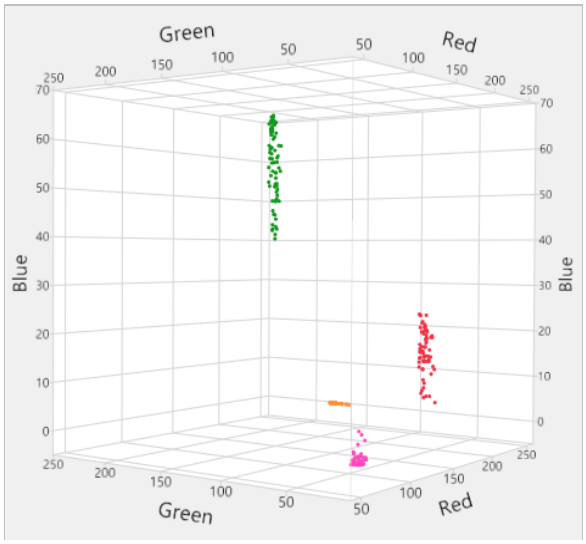

Figure 3. 3D scatter plot of RGB color values.

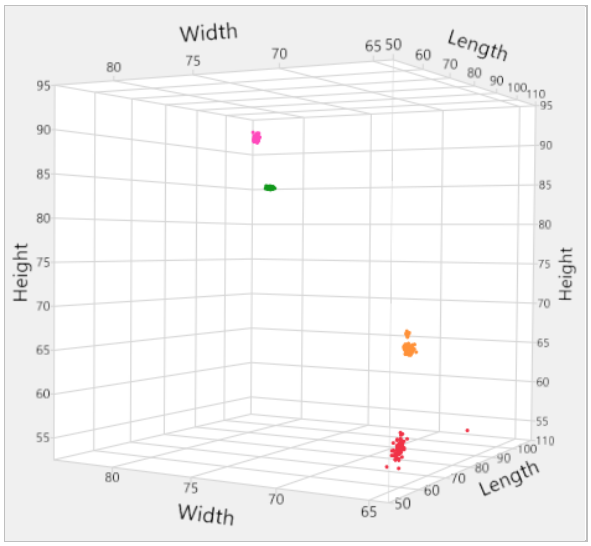

Figure 4. 3D scatter diagram of the dimensional values.

Another possibility for the internal delimitation of measurement data besides the use of 3D scatter plots is the use of decision trees. Using this particular type of predictive model, SAS JMP Pro 13 provides the ability to separate individual groups of measurements by dividing them based on significant differentiators. The aim is to identify and find out groups in the data set of the measured values. For further analysis work, these can be used to ensure automatic grouping of the individual object types based on the recorded measurement values. The significant distinguishing features of the measurement

data are processed by configuring Excel functions so that the dimension values and RGB color values are automatically assigned to their corresponding object type in the CSV-file after acquisition before being indexed together in the Elasticsearch cluster using Logstash. This classification procedure can also be transferred to other food objects without any problems. Generating production rules from decision trees is a common method when it comes to data classification tasks [33]. This method was chosen because the software SAS JMP Pro 13 allows the user to use decision trees to not only divide the data numerically, also to analyze it graphically. Additionally, this increases the value of the statistics model. The r^2-value fits very well to the model and is therefore also a reason for using this classification method. The explanation for this is given in the following section. The analysis of dimension values based on the use of decision trees is summarized in Figure 5.

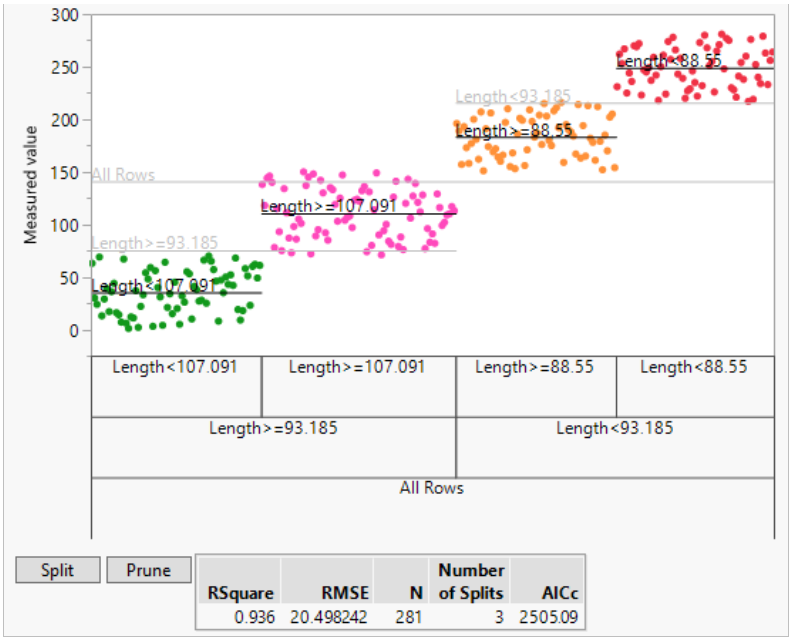

Figure 5. Decision tree of the dimensional values.

It is expressly visible that the four types of objects can be distinguished by their respective lengths. Not only the visualization in the upper part of the figure gives a clear indication that the application of the prediction model of the decision trees to these measured value data leads to an almost unambiguous grouping of the data regarding to the division criterion of length (millimeter). With an r^2-value close to 1, a high quality of fit of the regression line determined for the data set is proven. With the high accuracy of the model, based on the suitable r^2-value, a very good prediction can be made about the differentiation of the measured values according to the selected distribution criteria. The colors of the individual measured value groupings in Figure 5 correspond again to the same scheme as the colors in the 3D scatter plots and are also subdivided according to the respective type of object.

The universality of the scientific or technical approach is based on the CSV-file format in which the food data is stored. In particular, Elastic-Stack and Logstash, make use of this file format to build a data bridge between the data source and the core storage system of the cluster. Logstash reads the CSV file and converts it into the data language of the Elasticsearch cluster. A corresponding configuration file for the execution of this data transfer has to be completely rewritten or adapted to the respective use case for each CSV-file which is to be transferred to the Elasticsearch cluster. This approach is particularly advantageous within a closed network if several users want to index and subsequently evaluate data by using parallel a big data architecture. For example, this would not be a problem in the context of a cross-departmental use case of the Elastic-Stack. Considering the necessary of grouping object types with the mentioned Excel functions in the CSV-file, Figure 5, for example, can be used flexibly as a template for the recording and indexing of further food objects. In this context and

on the basis of the user-independent indexing possibility of data sets into the Elasticsearch cluster, as well as on the basis of the subsequent visualization possibilities of the indexed information by Kibana, the construction of a maximally universally and flexibly applicable big data architecture for the indexing and analysis of food data regarding their dimensions and RGB color values is realized.

5. Results

The three dimensions of the detected food are recorded with an accuracy of up to one millimeter. The system was tested at a belt speed of 0.5 m/s, but it can handle at least twice the speed. With the tested objects, the accuracy the objects are registered as close to 100%.

The food data were collected and divided into several data clusters according to the methods described in the previous Sections 3 and 4. In addition, the Elastic-Stack is fully installed and functional on the three Linux servers that provide the basis for our big data network. The browser-based visualization tool Kibana, based on the data storage and management system provided by Elasticsearch, is also fully operational and can be accessed from any computer within the university network. Logstash was also used to build the data pipeline between the CSV-file being indexed and the Elasticsearch cluster. Once new data is successfully indexed into the cluster, it can be accessed and visualized using Kibana. In a first step, Kibana was used to visualize the raw data and the corresponding recorded measurement values regarding to the dimensions and color of the respective object. The results are summarized in the following Figures 6 and 7.

The changes in the individual values indicate changing objects to be examined over time. Particularly when both visualizations are viewed together, it is possible to see at what time a change in the optical detection of the objects occurred. In addition to the creation of graphics and dashboards for real-time visualization, a data set was examined for corresponding correlations or anomalies using machine learning methods provided by Kibana. The results of this analysis are shown in Figure 8.

Figure 6. RGB color values over time (Kibana).

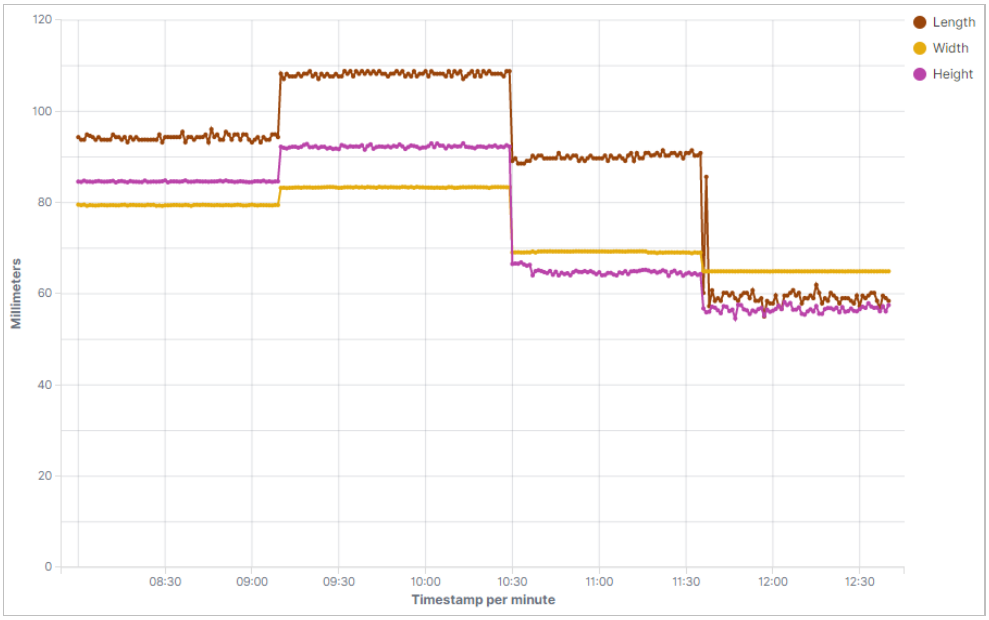

Figure 7. Dimensional values over time (Kibana).

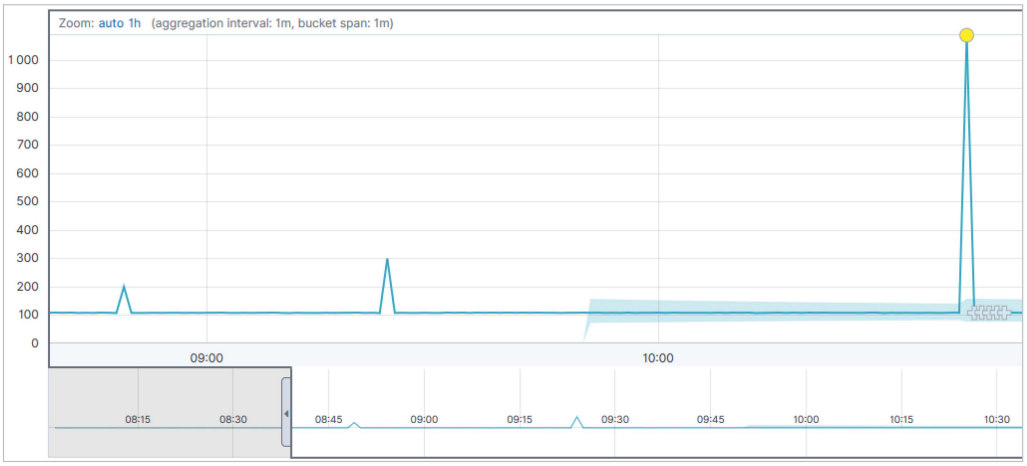

Figure 8. Examination for anomalies in the measured length (Kibana).

By taking measurements, we aim to deliberately distort the underlying data set in order to allow the analyzing algorithm to detect any anomalies in the sequence of our data. The data set used is the measured values of the lengths of the individual objects in relation to the time at which the corresponding values were recorded. The *Y*-axis indicates the length of the objects in millimeters. By looking at Figure 8, it can be seen that approximately halfway through the time, a length value has appeared that is clearly distinguishable from the other recorded values. This global maximum and two of the local maxima were detected by the algorithm for detecting irregularities during the evaluation of our measurement data. This option provided by Kibana for detecting anomalies in its own data sets allows the user to do retrospective analysis of individual measurement series, also for independent parallel recorded measurement series. Subsequently, the indexed data is processed with the help of Kibana in order to be able to realize a concrete detection of the optically detected objects. The basis for this is again the already mentioned CSV-file, which was stored with functions by the previously performed formation of clusters according to the objects, so that the assignment of the measurement data to the corresponding object types is carried out automatically. The following two Figures 9 and 10 show which objects were detected at which time. The objects were detected at minute intervals. The

object types were initially recorded strictly separately. The following Figure 9 shows how many objects of one type were detected in total over time.

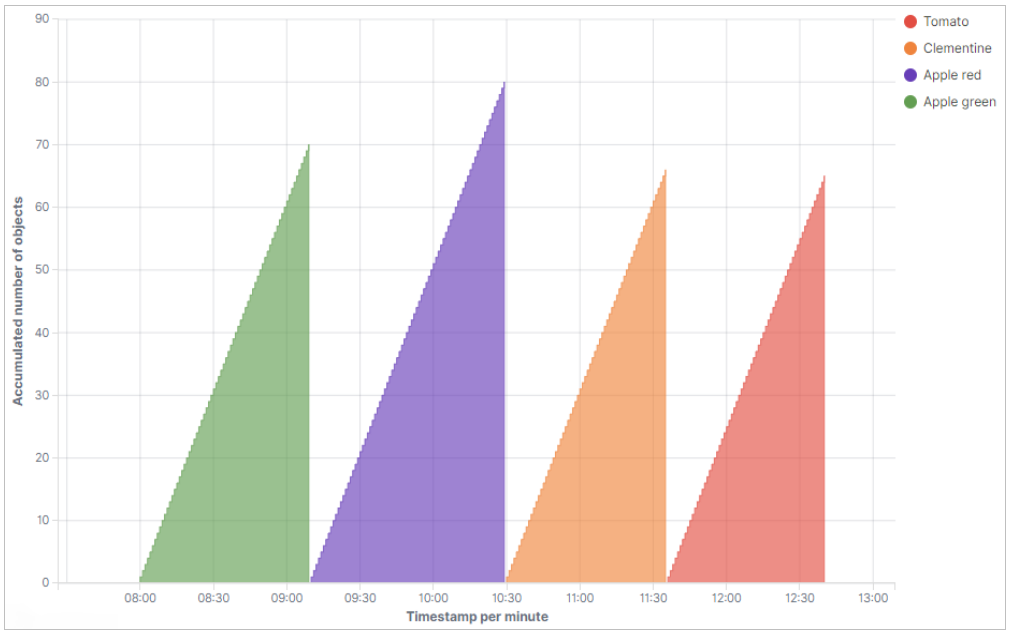

Figure 9. Cumulative number of recorded objects over time (Kibana).

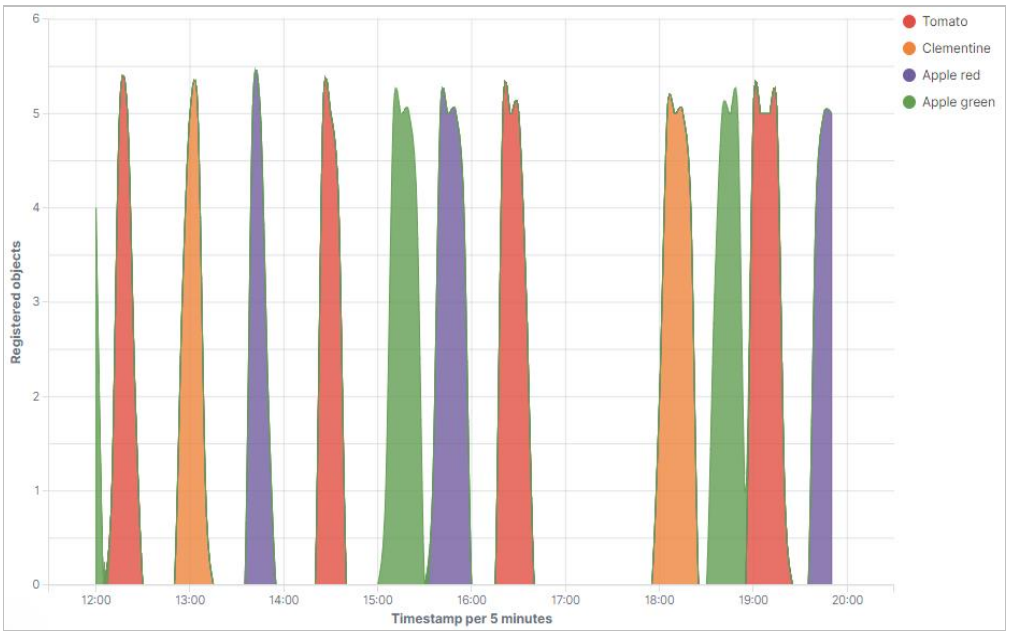

Figure 10. Number of detected mixed objects over time (Kibana).

In the measurement series in Figure 10, the same objects were used as in the measurement series before. However, the object types were now mixed and recorded at irregular intervals. If a group of objects is actively detected, the optical detection of these objects takes place again at intervals of one minute per object. The following diagram in Figure 10 shows the individual recorded groups of object types according to the time course in which the measurements were made. Thus, one of the basic goals of our research project, to establish a near-real-time automated acquisition of food data with subsequent appropriate categorization and visualization of the results, was implemented in a taxonomic way.

In the context of the evaluation of food data, it is now possible to automatically assign the incoming measured value data according to the object types using the delimiting division criteria. Transferred to a possible practice-relevant operation of the test setup, the observer of the test circuit is thus aware at any time which object in which quantity passes through the corresponding process step at what time. The recorded RGB color values should not only be used to ensure the assignment of the object types to the corresponding incoming measurement data. The color values should also be used as the basis for a visualization that shows the outside observer the color value of the object just recorded in the form of a representation that can be quickly understood by human eyes. This ensures that the basic machine information and sensor data can be meaningfully followed up and processed even if the Elastic-Stack fails. For illustration, a tool written in Python displays the color values, the corresponding object type and the time. The use of this software is completely independent of the Elastic-Stack, but also requires permanent access to the CSV file with the measured value data. The combination of Kibana and this tool for real-time detection and visualization of food color values provides the user a complete, past- and present-based overview of the food objects to be processed in their own production. The prerequisite for a supplementary simultaneous use of the software tools is, that both tools in combination with the optical detection of the food are configured in such a way that all programs are involved in chronological harmony regarding to the detection rhythm of the measuring technology or the entire production plant.

6. Discussion

The measuring principle of laser triangulation is very stable and provides high repeatability. Due to the lens used in the camera, only a limited depth of field can be achieved. If the object becomes too high, the laser becomes blurred for the camera and it can lose tracking. Therefore, the focus is set to an average height of the expected measurement objects. Due to its simple components and robust design, it is argued that the optical measuring system is extremely well suited for use in industry and food processing. The costs are low. The entire system is scalable and can be adapted to new conditions. These include larger objects, changed ambient light, higher belt speeds, etc. Further analysis or control options are conceivable at both a hardware and software level. One example is an algorithm which counts the distances between the meat fibers of fish fillets and can indicate on this basis the fat content of the fish. The system allows further output and analysis methods.

Regarding the data analysis based on this measurement system, it must be noted that the exploitation potential of a locally developed big data architecture such as the Elastic-Stack is enormous, in a scientific sense as well as in a business-related context. The successful application of such a technology with the subsequent use of profitable methods of visualization offers great added value in terms of increasing the own company-wide efficiency in dealing with independently collected data. Based on the obtained results, it is apparent how a value chain can be structured regarding to the extraction of information from digital data in the food industry. On the basis of the test setup, it is shown in detail that the near-real-time recording of dimensional values and RGB color values of various food objects can help to monitor its own production process at any time in a comprehensible and transparent manner. In addition to visualizing the RGB color values and object dimensions over time or the cumulative number of detected objects over time, it would also have been possible, for example, to display the number of detected mixed objects from Figure 10 in a corresponding form, so that the cumulative number of detected objects on the basis of time is also apparent. In a next step, the color registration of the objects could also be optimized, for example, by simultaneously determining several color values distributed parallel over the entire surface of the object. This would make it possible to visualize the color characteristics of the food objects with the help of an even more detailed representation. Based on the applied display alternatives it became clear, that a differentiation of the foods among themselves could be made due to the various visualization options offered by Kibana.

SAS JMP Pro 13 pre-filtering and grouping of measurement data has been instrumental in ensuring that individual object types can be distinguished based on specific patterns in the data using clearly

identifiable division criteria, even over a long-term period. Another major goal of the research activities, besides the collection and analysis of food data, is the establishment of a local and fully functional big data architecture in a scientific environment. In recent years, the associated possibilities for monitoring its own production processes in real time have taken on an increasingly comprehensive form, especially in the food industry. For example, at Anuga FoodTec, the leading international trade fair for the food and beverage industry in 2018, the most important digital trends and developments in the industry were discussed, including the goal of almost a complete local networking of all production halls of the respective food manufacturers. What is new here is the establishment of a digital connection between the individual production machines, which are located in a production hall. On the one hand, the objective of complete traceability of the recorded food data can be ensured. On the other hand, the aim of reproducing the individual recorded machine data throughout the entire process chain can also be pursued. Real-time machine-to-machine communication is also popular among manufacturers because this type of communication is the basis for predictive maintenance. The machines independently collect system-relevant information about themselves, especially about the operating and maintenance status, and forward this information to the central maintenance department. In this way, repairs can be carried out on the machines in advance in order to minimize machine downtime and thus increase the overall productivity of the food producer. At last year's Anuga FoodTec, some manufacturers assumed that companies in the beverage industry would have to deal with the challenge of filling beverages in very small batch sizes in the future. In this context, digital interfaces between the machines, their operators and even the customers themselves will be absolutely necessary, because the complexity of the associated individualism can only be tamed if it is possible to automate as many of the information flows along the entire value chain as possible [34].

The greatest exploitation potential of a holistic cluster-based data management system, as provided by Elasticsearch, can be seen in the fact that it is very flexible as an application in the respective organization and enables real-time data processing and visualization. In addition, this arbitrarily scalable way of handling data allows to influence the present and to derive strategies for the future by means of variable settings regarding the management of the indexed data and the related monitoring and to the visualization of streaming data, using all information from the past stored in the cluster. It is also important to mention that the scalability of the described big data architecture is very cost effective. The Elastic-Stack can initially be used as an open-source tool for experimental purposes. In the case of a commercial use, however, license fees are incurred; in return, the scope of functions of the software is expanded, the customer receives a direct support contact and updates are automatically provided by the manufacturer. The issue of system maintenance is thus almost completely covered at software level. On the hardware side, several industrial servers are required to set up the architecture for large data sets. However, it is also sufficient for data volumes of less than 1000 gigabytes to initially use one server with several virtual machines. The advantage of the arbitrary extensibility of the architecture and the adaptation of the system to the individual needs of the customer enable the cost efficiency that comes with the use of open-source technologies.

A further advantage is the following aspect: the functional scope and benefits of a big data cluster can not only be transferred to many different industrial use cases from practice, it can also be used in the context of scientific activities, for example in the network of the University of Applied Sciences Zwickau. This makes it possible to bundle all generated data and information of a project within a big data architecture like Elasticsearch. Each project member is able to independently feed all data and information related to a project into the cluster and identify it by creating an individual index. The real potential of such an architecture lies not only in data storage and subsequent analysis, rather in sharing data and making it universally and fully available to all participants, for example in the context of a project.

7. Conclusions

A setup for laser triangulation was implemented in order to detect food and other measurement objects in three dimensions with an accuracy of up to one millimeter. The measuring process starts automatically after a calibration with the detection of a food object under the measuring system. After successful measurement, the color information is obtained from the object to be measured and the data is stored automatically in a CSV file. Simultaneously a communication on a hardware level is available in the form of two ADC channels. With this it became possible to control peripheral machines especially in inline food production or to choose the correct machines in the next steps of the production line, depending on the size or the color of the food object. Additionally, the system offers space for further analysis options and additional hardware components such as special illuminations, IR cameras, etc.

For further data analysis a powerful big-data architecture is set up and explained in this paper. This paper also explains how different data clusters can be created using recorded dimensional values and RGB color values as a basis for data analysis. Using the SAS JMP Pro 13 software and the illustrations created with it, it is shown how the data volume can be subdivided into individual data groups on the basis of concrete division criteria using the latest technologies for data preparation, based on software-supported clustering. This approach showed above all how powerful modern software-supported cluster methods are in reality and how large the differentiating features within different data groups can be. Based on the cluster methods, it is possible to determine distribution criteria, which make the objects measurably distinguishable. This is the basis for feeding the data to be examined into the cluster-based large data architecture Elasticsearch in the form of a CSV file. The elastic stack was installed in advance on three Linux servers, so that the connection between Logstash, Elasticsearch, and Kibana is fully utilized for the transfer, storage, and processing of the data records. In the subsequent data analyses, Kibana is used to illustrate with the aid of Figures 6–10 how a production step, such as the automated counting of food objects every minute, can be represented in real time on the basis of an automatic assignment of the recorded measurement data to the individual food items. By discussing the results of the data analysis, it becomes clear that there is room for improvement in the future on how any user can use Kibana to configure visualizations and dashboards. The most important thing, however, is the data itself, which must be recorded in the most efficient way possible and first be buffered in order to be available for the subsequent data processing.

In summary, it should also be emphasized that the experimental design used makes it possible to establish a value chain for the collection, processing, and analysis of food data based on two separate sets of measurements. In this context, it is particularly clear that the general synergy effects resulting from the results of building a local large-scale data architecture can also be transferred to other practical use cases. The flexibility of such a system is characterized above all by the fact that several data streams can control a large storage system, to which different users can simultaneously access independently of each other and filter, prepare, or visualize data and information according to their own needs. The investment in a sustainable, very flexible, and comprehensive big data architecture for real-time information provision contributes to a better transfer of knowledge or a more compact presentation of information within its own organization or company. Beyond that in the longer term, this also promotes the necessary know-how in the direct handling of future-oriented big data technologies in companies and consolidates the associated future-oriented knowledge among the employees.

Author Contributions: Conceptualization: P.H., C.L.; methodology: R.M., T.J.; software: T.J.; validation, C.L., P.H.; formal analysis: T.J., R.M.; investigation: T.J., R.M.; resources: R.M.; data curation, T.J.; writing—original draft preparation: R.M., T.J.; writing—review and editing: R.M., T.J.; visualization: T.J.; supervision, P.H., C.L.; project administration: C.L., P.H.; funding acquisition: C.L., P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Saxon State Ministry of Science and Art, grant number: 2047000700.

Acknowledgments: The authors would like to thank the Saxon State Ministry of Science and Art, represented by the State of Saxony, for their partial funding of the research project "Applied research in the future field of digital communication (diKo 19), under which this research was conducted.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Federal Ministry of Economics and Energy: Food Industry. Available online: https://www.bmwi.de/Redaktion/DE/Artikel/Branchenfokus/Industrie/branchenfokus-lebensmittelindustrie.html (accessed on 6 December 2019).
2. A Bite into the Digital Bread—This is How the Digitalisation of the Food Industry Could Look Like. Available online: https://digital-magazin.de/digitalisierung-der-lebensmittelbranche/ (accessed on 6 December 2019).
3. Wittenhagen, J. Lebensmittelindustrie digitalisiert nur wenn nötig. *Food Mag.* **2017**, *31*, 43.
4. Osman, A.A.E. Trajectory Learning Using Principal Component Analysis Costanzo. In *Recent Advances in Information Systems and Technologies*; Rocha, Á., Correia, A.M., Adeli, H., Reis, L.P., Eds.; Springer: Berlin, Germany, 2017; Volume 1, pp. 174–184.
5. Khan, M.; Wu, X.; Xu, X.; Dou, W. Big Data Challenges and Opportunities in the Hype of Industry 4.0. In Proceedings of the IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
6. Latinovic, T.; Preradović, D.; Barz, C.R.; Vadean, A.; Todić, M. Big Data as the basis for the innovative development strategy of the Industry 4.0. In Proceedings of the IOP Conference Series Materials Science and Engineering, Banja Luke, Bosnia and Herzegovina, 9–11 May 2018.
7. Dobos, P.; Tamás, P.; Illés, B.; Balogh, R. Application possibilities of the Big Data concept in Industry 4.0. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kecskemét, Hungary, 7–8 June 2018.
8. Yan, J.; Meng, Y.; Lu, L.; Guo, C. Big-data-driven Based Intelligent Prognostics Scheme in Industry 4.0 Environment. In Proceedings of the Prognostics and System Health Management Conference, Banja Luke, Bosnia and Herzegovina, 9–11 May 2018.
9. Food Industry 'Big Data' Should Be Mined Better. Available online: https://www.foodmanufacture.co.uk/Article/2016/12/05/Food-industry-big-data-should-be-mined-better?utm_source=copyright&utm_medium=OnSite&utm_campaign=copyright (accessed on 6 December 2019).
10. Chiang, L.; Lu, B.; Castillo, I. Big Data Analytics in Chemical Engineering. *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8*, 63–85.
11. Automatica: Fleischindustrie Hat Noch Nachholbedarf. Available online: https://automationspraxis.industrie.de/branchenloesungen/lebensmittel-getraenke/fleischindustrie-hat-noch-nachholbedarf/ (accessed on 27 May 2020).
12. Alonso, R.S.; Sittón-Candanedo, I.; García, Ó.; Prieto, J.; Rodríguez-González, S. An intelligent Edge-IoT platform for monitoring livestock and crops in a dairy farming scenario. *Ad Hoc Netw.* **2019**, *98*, 1–54.
13. Benatia, M.A.; De Sa, V.E.; Baudry, D.; Delalin, H.; Halftermeyer, P. A framework for Big Data driven product traceability system. In Proceedings of the 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 21–24 March 2018.
14. Smys, S.; Tavares, J.M.R.S.; Balas, V.E.; Iliyasu, A.M. *Computational Vision and Bio-Inspired Computing*; Springer Nature: Cham, Switzerland, 2019; pp. 268–274.
15. Belaud, J.-P.; Prioux, N.; Vialle, C.; Sablayrolles, C. Big data for agri-food 4.0: Application to sustainability management for by-products supply chain. *Comput. Ind.* **2019**, *111*, 41–50.
16. Corallo, A.; Latino, M.E.; Menegoli, M. Agriculture 4.0: How Use Traceability Data to Tell Food Product to the Consumers. In Proceedings of the ICITM—International Conference on Industrial Technology and Management, Oxford, UK, 11–13 February 2020.
17. Srinivasan, R.; Giannikas, V.; Kumar, M.; Guyot, R.; McFarlane, D. Modelling food sourcing decisions under climate change: A data-driven approach. *Comput. Ind. Eng.* **2019**, *128*, 911–919.
18. Serazetdinova, L.; Garratt, J.; Baylis, A.; Stergiadis, S.; Collison, M.; Davis, S. How should we turn data into decisions in AgriFood? *J. Sci Food Agric* **2019**, *99*, 3213–3219.
19. Tao, D.; Yang, P.; Feng, H. Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr. Rev. Food Sci. Food Saf.* **2020**, *19*, 875–894.
20. Bernadi, P.D.; Azucar, D. *Innovation in Food Ecosystems: Entrepreneurship for -a Sustainable Future*; Springer Nature: Cham, Switzerland, 2020; pp. 189–221.

21. Hasnan, N.Z.N.; Yusoff, Y.M. Short review: Application Areas of Industry 4.0 Technologies in Food Processing Sector. In Proceedings of the IEEE Student Conference on Research and Development, Selangor, Malaysia, 26–28 Novomber 2018.
22. Jagtap, S.; Duong, L. Improving the new product development using big data: A case study of a food company. *Br. Food J.* **2019**, *121*, 2835–2848.
23. Khanuja, G.S.; Sharath, D.H.; Nandyala, S.; Palaniyandi, B. Cold Chain Management Using Model Based Design, Machine Learning Algorithms and Data Analytics. *SAE Tech. Pap.* **2018**, 1–6. [CrossRef]
24. Nita, S. Application of Big Data Technology in Support of Food Manufacturers' Commodity Demand Forecasting. *NEC Tech. J.* **2015**, *10*, 90–93.
25. Berndt, D.; Bichmann, S.; Breuckmann, B.; Clauß, U.; Glaser, U.; Klattenhoff, J.; Kühmstedt, P.; Noll, R.; Notni, G.; Trostmann, E.; et al. *Leitfaden zu Grundlagen und Anwendungen der Optischen 3-D-Messtechnik*; Fraunhofer-Allianz Vision: Erlangen, Germany, 2003; pp. 4–7.
26. Schuth, M.; Buerakov, W. *Handbuch Optische Messtechnik—Praktische Anwendungen für Entwicklung, Versuch, Fertigung und Qualitätssicherung*; Carl Hanser Verlag: München, Germany, 2017; pp. 363–393.
27. Chen, H.Q.; Xin, H.W.; Teng, G.H.; Meng, C.Y.; Du, X.D.; Mao, T.T.; Wang, C. Cloud-based data management system for automatic real-time data acquisition from large-scale laying-hen farms. *Int. J. Agric. Biol. Eng.* **2016**, *9*, 106–115.
28. Chaitanya: Advantages and Disadvantages of XML. Available online: https://beginnersbook.com/2018/10/advantages-and-disadvantages-of-xml/ (accessed on 26 May 2020).
29. JSON—Its Advantages and Disadvantages. Available online: https://ezeelive.com/json-advantages-disadvantages/ (accessed on 26 May 2020).
30. MySQL: MySQL Products. Available online: https://www.mysql.com/de/products/ (accessed on 26 May 2020).
31. Elasticsearch: Elastic Stack-Features. Available online: https://www.elastic.co/de/elastic-stack/features (accessed on 26 May 2020).
32. Samson, M.; Damon, D.; Dilraj, S.; Onur, E.; Ken, L. Elasticsearch Ratings. Available online: https://www.capterra.com.de/reviews/149304/elasticsearch (accessed on 26 May 2020).
33. Quinlan, J.R. Generating Production Rules from Decision Trees. Available online: https://www.ijcai.org/Proceedings/87-1/Papers/063.pdf (accessed on 20 May 2020).
34. Augsburger, S. Trendbericht: Industrie 4.0. Available online: http://www.lebensmitteltechnik-online.de/Anuga_FoodTec2018/AnugaFoodTec_Trend_No4_Industrie40.html (accessed on 12 June 2020).