Invited review: Big Data in precision dairy farming

C. Lokhorst¹,²†, R. M. de Mol¹ and C. Kamphuis¹

¹Wageningen Livestock Research, PO Box 338, 6700AH Wageningen, The Netherlands; ²Van Hall Larenstein University for Applied Science, PO Box 1528, 8901BV Leeuwarden, The Netherlands

(Received 18 January 2018; Accepted 19 November 2018)

Insight into current scientific applications of Big Data in the precision dairy farming area may help us to understand the inflated expectations around Big Data. The objective of this invited review paper is to give that scientific background and determine whether Big Data has overcome the peak of inflated expectations. A conceptual model was created, and a literature search in Scopus resulted in 1442 scientific peer reviewed papers. After thorough screening on relevance and classification by the authors, 142 papers remained for further analysis. The area of precision dairy farming (with classes in the primary chain (dairy farm, feed, breed, health, food, retail, consumer) and levels for object of interest (animal, farm, network)), the Big Data-V area (with categories on Volume, Velocity, Variety and other V’s) and the data analytics area (with categories in analysis methods (supervised learning, unsupervised learning, semi-supervised classification, reinforcement learning) and data characteristics (time-series, streaming, sequence, graph, spatial, multimedia)) were analysed. The animal sublevel, with 83% of the papers, exceeds the farm sublevel and network sublevel. Within the animal sublevel, topics within the dairy farm level prevailed with 58% over the health level (33%). Within the Big Data category, the Volume category was most favoured with 59% of the papers, followed by 37% of papers that included the Variety category. None of the papers included the Velocity category. Supervised learning, representing 87% of the papers, exceeds unsupervised learning (12%). Within supervised learning, 64% of the papers dealt with classification issues and exceeds the regression methods (36%). Time-series were used in 61% of the papers and were mostly dealing with animal-based farm data. Multimedia data appeared in a greater number of recent papers. Based on these results, it can be concluded that Big Data is a relevant topic of research within the precision dairy farming area, but that the full potential of Big Data in this precision dairy farming area is not utilised yet. However, the present authors expect the full potential of Big Data, within the precision dairy farming area, will be reached when multiple Big Data characteristics (Volume, Variety and other V’s) and sources (animal, groups, farms and chain parts) are used simultaneously, adding value to operational and strategic decision.

Keywords: dairy chain, data analytics, data characteristics, data mining, expectations

Implications

Insight into current scientific applications of Big Data within the precision dairy farming area may help us to understand the inflated expectations around Big Data. In total, 142 papers were selected, and the analyses demonstrated that the full potential of Big Data has not yet been reached. The papers focussed on farm-related animal data that were based on the Volume and Variety categories, and that were predominantly analysed with supervised classification methods. The dairy sector can possibly benefit more from Big Data if other V categories, such as Velocity, will be incorporated and if more animals, groups, farms and chain parts will be involved.

Introduction

John Mashley introduced the term ‘Big Data’ to the high-tech community in the early 1990s (Lohr, 2013). However, the term was not further defined, and it took some years before the three characteristics (Volume, Velocity and Variety) identified by Laney (2001) became associated with Big Data (De Mauro et al., 2016) to form today’s mainstream description of Big Data. Still, the use of the term Big Data was barely noticeable before 2011, to explode thereafter (Sonka and Cheng, 2015a). This increased use of the term Big Data is associated with advances in computer storage and processing power, a sharp reduction in costs of sensors and communication and recent developments in the Internet of Things. These advances result in the use of many sources
Big Data analytics is an increasingly used term within the agro-food domain, where Big Data is described as highly valuable once it is established. For example, Kempenaar et al. (2016) explored the field of Big Data in a Dutch context, using the dairy-milk production chain as a use case. According to them, the value of Big Data lies in the information and (new) insights that organisations can draw from it, rather than in the data as such. They raised the perspective that Big Data represents a disruptive innovation, potentially changing organisations dramatically. They concluded that the use of Big Data, once established, will support smart decisions and management, but that the creation of an integration platform for Big Data analysis was still far too ambitious, as well as being much more complex than expected. Despite this complexity, Big Data also is of interest and raises high hopes within the livestock sector: for example, the Animal Task Force, a European public–private platform promoting a sustainable and competitive livestock sector in Europe, published a number of themes with research priorities (Animal Task Force, 2014). One of these themes included resource efficiency, where the topic of Big Data, in combination with phenotyping and precision livestock farming, was identified with a high priority to meet policy goals within the European community. Finally, Big Data is an emerging research field within the precision dairy farming area. There were only a small amount of papers addressing Big Data explicitly in earlier international conferences on precision livestock farming, with one paper in the European conference on precision livestock farming in 2013 (Berkmans and Vandermeulen, 2013) and two papers in 2015 (Guarino and Berkman, 2015). There was a larger amount of papers having Big Data as a specific topic during the first international Precision Dairy Farming 2016 conference (Kamphuis and Steeneveld, 2016). At that conference Dias et al. (2016) addressed the creation of value with data from pasture-based farming systems, Van der Waaij et al. (2016) used machine learning to predict individual cow feed intake, Verhoosel and Spek (2016) examined the semantics for Big Data applications, Harty and Healy (2016) used Big Data advanced analytics to optimize health and fertility and Bahr et al. (2016) went into the field of data-driven smart feeding. During the most recent European Conference on Precision Livestock Farming (Berkmans and Keita, 2017), there was an entire session specifically devoted to Big Data. All in all, Big Data is increasingly a theme of interest and high expectations for the precision dairy farming area. Full potential of Big Data is expected to occur when data with different V characteristics (Volume, Variety, Veracity) are used, that originate from more animals, groups, farms and chain parts in adding value to operational and strategic decisions. However, to validate if these high expectations are justified, it is essential to identify the current scientific applications of Big Data in the precision dairy farming area. Insight into these scientific activities will help to identify at what stage Big Data is within the Gartner hype cycle (Fenn and Raskino, 2008) with regard to the precision dairy farming area, and whether this can be supported by science. To do so, a literature review has been conducted to find relevant scientific papers addressing the topic of Big Data within the precision dairy farming area.

Based on these developments, Animal requested an update to be given of the scientific status of Big Data in precision dairy farming. Therefore, the goal of this review paper is to give that scientific background and to see whether Big Data has overcome the peak of inflated expectation. Precision dairy farming is interpreted as a subset of precision livestock farming, specifically focussing on the dairy sector.

Material and methods

To structure the selection and review of relevant scientific papers on Big Data within the precision dairy farming area, a conceptual model was constructed. This conceptual model was based on the task and skills mapping from Gartner (Fenn and Raskino, 2008). According to this mapping, three core areas were identified to be present in a Big Data project or environment: domain understanding: information technology (IT) to generate data; and skills in data analytics. These three core areas were defined more specifically for this current review into: the precision dairy farming area, the Big Data-V area, and the Big Data analytics area. For each of these three, relevant levels and categories were identified.

Big Data in the precision dairy farming area

The production network in the dairy sector is quite complex, and there are several potential places in that network where decisions could be improved or supported by the use of the Big Data concept. To identify where Big Data was applied within the precision dairy farming area, the dairy chain was categorized into dairy farm, food, retail and consumer as main levels. Added to this were the main partners for dairy farmers, which come from the feed, breed and health organisations. With these categories, the most important stakeholders who might invest and benefit from Big Data, as far as concerning the production side of precision dairy farming, were identified. The financial, legal and governmental aspects of the dairy production network were not taken into account.
Within each level, three sublevels were created to identify the object of interest: the first is the animal sublevel, where data are used originating from individual animal level. The second is the farm sublevel, where a group of cows on a farm was the observed unit. Since data may originate from more than one farm and more than one part of the production chain, the third sublevel was seen as the network. Within each main level (dairy farm, food, retail, consumer, feed, breed, health), the sublevels (animal, farm, network) were classified in the literature review. These main and sublevels will provide basic insights into the type of questions and problems that are currently being addressed with a Big Data approach. In addition to these main levels and sublevels, specific remarks were made for some papers to identify unexpected and specific questions that were addressed.

The Big Data-V area

Big Data is an interplay of various V’s, and analytics (Sonka and Cheng, 2015a). The Big Data-V area focuses on various V’s, where the three most commonly used V’s defined by Laney (2001) are Volume, Variety and Velocity:

- **Volume**: this aspect clearly links directly to the Big component (Sonka and Cheng, 2015a), and it is this aspect that is often emphasised in the media since it is easy to impress with really large numbers. However, the threshold of what volume should be called Big is not specifically defined (Sonka and Cheng, 2015a). This is a highly subjective aspect that varies between industries and applications. To be called Big, this may be any data set whose size exceeds the ability of typical software used to capture, store, manage and analyse, or any data set challenging the constraints of a system capability or business needs (Zaslavsky et al., 2012; Sonka and Cheng, 2015b; Wolfert et al., 2017). Adding to this, data that we consider Big today may not be considered Big tomorrow due to the continuous advances in processing, storage and other system capabilities (Zaslavsky et al., 2012).

- **Variety**: This aspect links to the Data component (Sonka and Cheng, 2015a). Laney (2001) identified the variety of incompatible formats, non-aligned structures and inconsistent semantics as the greatest barrier to effective data management. This aspect of Big Data refers to the expansion of what data really are and includes the widening range of data types and data sources that are available and that need to be handled (Devlin, 2012; Zaslavsky et al., 2012).

- **Velocity**: This aspect refers to the Data component and links to how frequently data are generated (Zaslavsky et al., 2012), the increasing speed of data arrival and processing (Devlin, 2012) and the capability to understand and respond to events as they occur, even in real-time situations (Sonka and Cheng, 2015a).

Over the years, many more V’s have been added: veracity – the need to trust the data; variability – the variance in meaning of data in sentiment analyses; visualization – the creation of complex graphs that include many variables of data while remaining understandable; and value – the value of relying on data-driven analyses for decision-making have been mentioned too (Zaslavsky et al., 2012; Van Rijmeman, 2017). Devlin (2012) adds virality, viscosity and validity, but does not specify any of these terms.

None of the aforementioned V’s used to describe Big Data-V are easy to quantify, imposing a challenge to objectively score which Big Data-V category is fulfilled. For this review, we classified papers into the following four Big Data categories: Volume, Variety, Velocity and Other V’s. For the Volume category, not only the amount of data used was taken into account but also the scalability of the approach. For the Variety category, it was relevant whether or not different types of data were used, for example, milk yield and milk temperature. For the Velocity category, it was checked whether or not an immediate real-time response was required in the application. No restriction or specification was made for the last V category (Other V’s), only the use of such a word. Per paper, individual remarks were stored if they could contribute to a better quantification of the V category that was addressed.

The Big Data analytics and aspects area

Data analytics is needed to transform Big Data into information, knowledge and action. Both traditional statistical techniques and data-driven machine-learning techniques can be used to make sense of Big Data and to translate it into timely and valuable information. Statistical techniques can be seen as a subset of machine-learning techniques, where subsequently machine-learning techniques are a subset of data-mining techniques. Machine-learning algorithms are used to automatically learn patterns and make inferences from data. Those algorithms can be classified according to algorithm type (Murthy et al., 2014). Within the area of data analytics for this review paper, the following four categories of machine learning were used (Murthy et al., 2014):

- **Supervised learning**: With supervised learning, the outcome of interest is known for each record used for model development. In other words, the data used for model development are labelled. Within this category, papers were classified into regression and classification. For regression, the outcome variable has a numerical value. Possible techniques involve linear regression, polynomial regression, use of radial basis functions, multivariate adaptive regression splines or multilinear interpolation. For classification, the outcome variable is categorical (e.g. binary yes/no). Possible techniques include neural networks, decision trees, naive Bayes model or support vector machines.

- **Unsupervised learning**: With unsupervised learning, the outcome of interest is unknown (unlabelled) for each record used for model development. Within this category, papers were classified on clustering and dimensionality reduction. Clustering techniques include K-means,
Gaussian mixture modelling, spectral clustering or hierarchical clustering. Dimensionality-reduction techniques include, for example, principal component analysis or independent component analysis.

- **Semi-supervised classification:** Here, a mixture of small amounts of labelled data is fused with large unlabelled data sets by using, for example, active learning, transfer learning or co-training (Khormashahi et al., 2013).

- **Reinforcement learning:** Here, a mapping function is learnt to maximise a reward function. This technique is used, for example, in Markov decision processes or Q-learning methods.

To get insight into the type of data used for Big Data, the selected papers were also classified on the following data type aspects (according to Murthy et al., 2014):

- **Time-series** data are sequences of values or events obtained over repeated measurements of time, for instance sensor data of activity or behaviour of dairy cows. Scaling of the existing time-series model algorithms might be needed, and this is successfully done for dynamic time warping.

- **Streaming** data are constantly arriving, for instance from remote sensors or surveillance systems, and should be processed in an online fashion. Offline algorithms can be approximated by distributed machine-learning algorithms for streaming data.

- **Sequence** data consist of sequences of ordered elements or events that are recorded with or without a concrete notion of time.

- **Graph** data where problems are modelled as graphs, like in social networks or biological networks. Algorithms may include a matrix representation and therefore necessitate large matrix solvers that have to be adapted in case of Big Data.

- **Spatial** data can be place-related data or remote-sensing data. These data are not independent because of the spatial relationship, and this fact can be exploited by the algorithms.

- **Multimedia** data, such as images, videos, audio and text mark-ups, need special digital-signal-processing techniques for image segmentation and motion-vector analysis.

Once the three core areas (precision dairy farming, Big Data category V and Big Data analytics and aspects) were defined and made more specific as described above, a literature search was performed in Scopus. The search was limited to papers published between 1994 and June 2017. The search string was ‘machine learning OR statistical learning OR big data OR data mining OR neural network OR decision tree OR support vector machine OR k-means cluster* OR component analysis OR semi-supervised learning OR reinforcement learning & subject area = agricultural OR veterinary OR decision sciences’. This search resulted in 1442 scientific peer-reviewed papers. After a first screening by the present authors, the majority of the papers were considered irrelevant as they were not directed to livestock or dairy cattle. A total of 348 papers remained for further evaluation. After a classification by the present authors on fitness to the review, 142 remaining papers were assigned to precision dairy farming area, Big Data category V and Big Data data analytics and aspects areas. The rejected papers were dedicated to other fields of cattle research, like young stock, feeding cattle, dairy processing, genetics or other livestock species, and not on dairy cattle. Per paper, remarks on, for example, accuracy, speed, robustness, scalability or interpretability of the methods used were made. Results are presented as number of papers per core area, and summarized in tables. For the interaction between the areas, cross-tabs were created with the number of papers and the percentage of these papers for all combinations of precision dairy farming area, Big Data-V, Big Data analytics and data type. This resulted in a number of interactions based on a too limited number of papers, disabling any further exploration or interpretation. However, there were some interactions of specific interest, worthwhile to describe. These results were most interesting when the percentage of papers per area deviated from the overall row/column average percentage in the crosstab. For example, in total 54% of the papers dealt with Classification as analytics, but for the Health area this was much higher: 79%. Only large deviations from the overall row/column averages are described in the result section.

**Results**

*Publication dates of papers*

In Figure 1, the number of papers published per year from 1994 onwards is presented. Note that the number of papers for 2017 includes only those published in the first 6 months. A small number of papers published per year from 1994 to 2005, where the peak in 1995 are papers that originated mainly from one author. From 2007 onwards, there is a trend of a gradual increase in the amount of papers published. In more recent years, ~20 papers published per year have been dedicated to Big Data in the precision dairy farming area.

*Precision dairy farming area*

The results of the classification of the papers per area within precision dairy farming are summarized in Table 1. Since a few papers addressed more than one main or sublevel, 153 classifications were made based on the 142 papers. With 58% of the classifications, the dairy farm stakeholder is by far the biggest, followed with 33% of the classifications for the health stakeholders. Papers involving stakeholders from feed, breed and food, being directly involved in the production chain, are represented limited with 3%, 4% and 2%, respectively. Only one paper involved the retail level and no paper involved the consumers level.

Table 1 also shows that the majority of papers analyse Big Data at the animal sublevel (83% of the classifications), whereas Big Data were analysed at the farm or the network sublevel in only 11% and 6% of the classifications.
respectively. From the papers that addressed more than one main level at the same time, seven papers were oriented on the dairy farm–health interaction at the animal sublevel. Other interactions were feed–breed, dairy farm–breed and breed–health, with all but one paper at the animal sublevel; that paper was oriented on the dairy farm–food interaction at the network sublevel. Within the main health level, there was one paper addressing both the animal and farm sublevels. The remainder of the papers (n = 135) focussed only on one main or sublevel at a time.

The Big Data-V categories
Since some papers covered more than one Big Data-V category, a total of 203 classifications into these Big Data-V categories were made. There were 57 papers addressing the Volume and Variety categories at the same time. The vast majority with 119 (59%) of the 203 classifications addressed the Volume category, and 76 (37%) of the classifications addressed the Variety category. Only eight (4%) of the classifications addressed the Other V’s category and none of the papers addressed the Velocity category. The Other V’s were present in the papers of McQueen et al. (1995), Ortiz-Pelaez and Pfeiffer (2008), Cole et al. (2012), Garcia (2013), Yoder et al. (2014), Van Der Weerd and de Boer (2015), Vonder et al. (2015) and Hermans et al. (2017).

Big Data analytics area
There were 154 different classifications of the 142 papers into categories and levels within categories of the Big Data analytics area, due to papers relating to machine-learning techniques from more than one main category. Results presented in Table 2 show that 87% of the classifications used supervised learning techniques, and only 12% used unsupervised learning. Semi-supervised learning was used only by Yao et al. (2016), and no papers were found on reinforcement learning. Within supervised learning, classification was used in 64% of the papers. The other 36% of the papers used regression as supervised learning technique.

The paper of Moya et al. (2015) is the only paper that addresses three different machine-learning techniques, reporting on supervised classification, unsupervised clustering and dimensionality reduction. The papers from Martinez-Ortiz et al. (2013), Garcia (2013), Caraviello et al. (2006) and Lacroix et al. (1997) combine supervised regression and classification techniques. Moreover, three papers (Sloth...
et al., 2003; Caccamo et al., 2015; Kayano and Kida, 2015) combine the supervised regression techniques with the unsupervised dimensionality reduction techniques. Only two papers (Yoder et al., 2014; Brotzman et al., 2015) combine the unsupervised clustering and dimensionality reduction techniques. Three different combinations between supervised and unsupervised analytics were reported. Shahriar et al. (2016) combined classification and clustering, Schultz et al. (2016) combined regression and clustering and Nielen et al. (1995) combined classification and dimensionality reduction techniques.

From the 142 papers, only 114 could be classified based on the data type (see Table 3). The vast majority of classified papers (61%) used time-series data, followed by 18% of the classified papers using multimedia data. Sequence, graph and spatial data accounted for 9%, 5% and 5%, respectively. The use of streaming data was very limited, involving only one out of the 114 classified papers (1%).

There were six papers that combined time-series data with another data type. Williams et al. (2016), Homburger et al. (2014) and Nadimi et al. (2012) combined time-series with spatial data, whereas Martinez-Ortiz et al. (2013), Graunke et al. (2013) and Gargiulo et al. (2012) combined time-series with multimedia data. In other papers, graph data were combined with sequence data (Nohuddin et al., 2010; Cole et al., 2012), multimedia data (Guzhva et al., 2015; Guzhva et al., 2016) and spatial data (Dawson et al., 2015).

Interactions between precision dairy farming area, Big Data-V and Big Data analytics

Based on the cross-tabs results the following interactions of specific interest are worthwhile to describe. Papers using spatial data (86%) and multimedia data (81%) were more than the overall average (66%) adressed to the Big Data-V category Volume. Moreover, papers on the dairy farm main level (64%) adressed more than average (58%) the Big Data-V category Volume, and this was less for the papers on the health main level (51%). Papers using spatial data (14%) and multimedia data (19%) were less than the overall average (31%) adressed to the Big Data-V category Variety. However, 38% of the papers using time-series data adressed the Big Data-V category Variety. Papers on the health main level (47%) were more than average (38%) addressing the Big Data-V category Variety, and this was less for papers on the dairy farm main level (32%). This is contrary to Big Data-V category Volume.

When multimedia data were involved, supervised regression occurred relatively more often in the papers: 38%, compared with the expected 27%. There was also a trend that supervised regression was used more at the dairy farm level (47%) and less at the health level (13%), compared with the expected 33%. On the other hand, supervised classification occurred more when time-series were involved (66%) and less when multimedia data were involved (46%), compared with the expected 57%. Supervised classification was used relatively less, with 47% at the dairy farm level and relatively more with 79% at the health level, compared with the expected 54%.

Discussion

As expected, the number of papers on Big Data within the precision dairy farming area has increased in the past decade. This coincides with the predictions of Gartner’s yearly reports on the hype cycle. According to Fenn and Raskino (2008), the ‘hype cycle’ is a branded graphical presentation developed by the Gartner firm for representing the maturity, adoption and social application of specific technologies. Recent publications of the hype cycle show that Big Data is currently passing the peak of inflated expectations, and evidence on useful applications of Big Data will come. Therefore, it is also expected that more scientific papers on Big Data applications in the precision dairy farming area will be published over the coming years.

During the review process, it became clear to the present authors that there is a thin line between Big Data and precision dairy farming and that these two aspects are easy to confuse (Sonka and Cheng, 2015a). Certainly, the use of
technologies (sensing and others) is increasing on dairy farms (Rutten et al., 2013; Caja et al., 2016). The number of variables being recorded has also increased, ranging from milk quality parameters to cow behaviour, and new methods required to analyse this data are currently under investigation (Mottram, 2016). Thus, precision dairy farming does include key elements of Big Data, but there are differences between the two terms. One of these differences is that precision dairy farming usually focuses on information on individual cows, or at farm level, but the Volume category of Big Data-V requires observations from many animals or farms (Sonka and Cheng, 2015b). Moreover, data from individual farms, or other agricultural stakeholders, are not likely to possess the entire range of data sources needed to generate new insights. Additional sources of data naturally reside and originate outside the farm (Sonka and Cheng, 2015b), and our current review confirms that these are not yet fully explored. Traditionally, research in precision dairy farming focusses on one factor (e.g. reproduction or udder health). However, to identify complex interactions across several factors and multiple years requires much more sophisticated tools (Sonka and Cheng, 2015b). Getting access to and aggregate data from several sources may be a huge challenge within the current precision dairy area. These challenges include data ownership, intelligent processing and analytics and clear business models (Kempenaar et al., 2016; Wollfert et al., 2017).

Results of the first core discipline (Big Data within the precision dairy farming area) showed surprisingly more papers at the animal sublevel than at the farm sublevel. The present authors expected that Big Data would be most valuable when data of different dairy farms and parts of the supply chain were combined. Most of the published papers are at the animal sublevel, which suggests that people expect to gain a lot from animal data. Within these animal data, it can be seen that most of the published papers deal with health issues. This seems to be in contrast with the development of funded research projects in the Netherlands, where there is a strong focus on breed- and feed-oriented research projects. From the breed perspective, the focus is on precision phenotyping, where animal performance data are collected in different production environments and used to develop new phenotypes, focussing on animal health and behaviour or proxies for complex traits such as resilience or efficiency, to be used in breeding programmes. For feed, the focus is on precision feeding of individual animals. The Big Data papers seemed to focus on the relationship between health and feed, since changes in health directly impact the feeding of animals. Recently more papers have been published with a focus on the farm sublevel and the production network. This will be more in line with the present authors expectations. It is also interesting that the published papers did not cover the (precision) dairy aspects of food companies, retailers and consumers. This may be explained by precision dairy farming’s focus on individual animals that live in a group and an increased involvement of feed, breed and food advisers. Retailers and consumers are not directly involved yet in precision dairy farming. Thus, we can expect Big Data papers focussing on the food-company perspective to increase in the coming years.

The expectation that Big Data will address more V’s at the same time is not supported by our review. Until now, most of the published papers dealt with Volume or Variety. The large number of papers using Volume is as expected, since data are gathered more at the animal sublevel, and spatial and multimedia data has become easier to process. Developments in the Internet of Things and data platforms will stimulate this even more. These developments also improve the integration of data coming from multiple sources and multiple types. Much to the present authors’ surprise, there were no papers dealing with Velocity. Thus, it seems that Velocity was not considered as important, or there was no demand for studies on real-time decision-making on dairy farms. All other V’s are hardly addressed by the published papers. When they did appear in the reviewed papers, this was more indicative rather than actually involving that aspect in the data analytics. The present authors expect that successful Big Data applications will also support operational decisions. For example, the decision to inseminate a cow will be based not only on the increased activity, lower milk production and increase in milk temperature (De Mol et al., 2007), but also will be based on the actual semen price and milk price. These additional data need to come from sources other than the dairy farm itself. The current papers on Big Data contribute to the scientific knowledge, allowing for the development of new products, services or management strategies.

Data mining is also used in the context of dairy data analysis. Lokhorst et al. (1999) investigated the potential of data mining to benchmark dairy farms at the farm level. The concept of data mining was used to find new insights and knowledge when data from several farms were brought together. Although the development of new insights is also a promise of Big Data, the reviewed papers show a more structural analysis based on assumptions (and biological relevance) and supervised learning techniques. This search might need the combined expertise of domain knowledge and data analytics, as also suggested by Gartner (Fenn and Raskino, 2008), when mapping Big Data tasks and skills into the three core disciplines of domain understanding, data science and IT skills. Classification and regression techniques are preferred in the reviewed papers, which seemed to originate from the more classical statistical approaches in data analytics used within the animal sublevel. In the precision dairy farming area, it is generally expected that data sets will contain missing data and/or that only a subset of data will be available. To tackle these issues, semi-supervised learning techniques might be appropriate, however, these techniques were not used in the current papers.

With regard to Big Data analytical aspects, it is logical that time-series are used. This is especially so when the individual animal (cow) level is used and production data are analysed. As such, time-series data were well represented in the current papers. We also see the appearance of spatial and multimedia data in more recent published papers. Use of
camera techniques, including three-dimensional (3D) and multispectral cameras, have become possible not only through advances in camera technology, but also the ability to easily acquire, exchange and handle data in an internet-based environment. Examples include the use of 3D cameras to determine body composition (Fischer et al., 2015) and locomotion scoring (Viazzì et al., 2013). The data obtained from these developments also become available for Big Data analytics. The main driving force is the availability of the technology itself, rather than the specific need for data. Thus, in Big Data projects, there is still a lot of pragmatism. If data are available, we use them.

The papers published so far have focussed on traditional questions that occur in precision dairy farming. The question then is whether the full potential of Big Data is being realized. Full potential is being able to use data from different V categories and from more animals, groups, farms and chain parts, adding value to operational and strategic decisions. Is it possible that, in this challenging environment, other questions can be asked that currently are not foreseen. How to formulate these more data-driven questions and in which research environments and teams this should be done cannot be answered in this review paper. This will not be the exclusive area of domain experts anymore.

**Conclusion**

Based on the literature research and analysis of the 142 reviewed papers, the following conclusions can be drawn:

- Big Data is a topic of research in the precision dairy farming area, with a steady increase of papers in the past decade.
- The number of published papers focussed on the animal sublevel (83%) exceeded those focussing on the farm or network sublevel (11% and 6%, respectively). At the animal sublevel, topics on the dairy farm level prevail against those on the health level (58% vs. 33%, respectively).
- From the Big Data-V categories, the Volume category is most favoured (59%) of the published papers, followed by 37% of published papers that deal with the Variety category. The Velocity category is not dealt with in the reviewed papers.
- Papers dealing with supervised learning exceed those dealing with unsupervised learning. Within supervised learning, 64% of the papers dealt with classification issues, and exceed the papers dealing with regression methods.
- Time-series are used in 61% of the papers, and are mostly dealing with animal-based farm data. Recently, multimedia data are appearing increasingly in published papers.

Based on these results and the expectations of the present authors, it can be stated that the full potential of Big Data in the precision dairy farming area has not been realised. Full potential is expected to occur when data with different V characteristics are used, that originate from more animals, groups, farms and chain parts in adding value to operational and strategic decisions.

**Acknowledgements**

The study is part of the project KB-27-001-001 Big Data for healthy resources utilization, which is funded by the Dutch Ministry of Agriculture, Nature, and Food Quality. The full list of the 142 classified papers can be seen in the supplementary Material S1. The remarks made during the review process, and the crosstabs can be obtained on request by the authors.

**Declaration of interest**

There is no conflicts of interest.

**Ethics statement**

The review did not needed the involvement of an ethical committee and does not conflict with legislation issues.

**Software and data repository resources**

None of the data were deposited in an official repository.

**Supplementary material**

To view supplementary material for this article, please visit https://doi.org/10.1017/S1751731118003439

**References**


Big Data in precision dairy farming


Fenn J and Raskine M 2008. Mastering the hype cycle: how to choose the right innovation at the right time. Harvard Business Press, Cambridge, MA, USA.


Sonka S and Cheng YT 2015a. Big data: more than a lot of numbers!. Farmdoc Daily 5, 201.