A Virtual Dairy Farm Brain

Di Liang[a], Hector Delgado[b], Victor Cabrera[c*]

[a] Department of Dairy Science, University of Wisconsin Madison, dliang7@wisc.edu

[b] Department of Dairy Science, University of Wisconsin Madison, hector.delgadorodriguez@wisc.edu

[c] Department of Dairy Science, University of Wisconsin Madison, vcabrera@wisc.edu

* Corresponding author

Abstract: *We are developing a "Virtual Dairy Farm Brain" by applying Precision Farming, Big Data analytics, and the Internet of Things. This is a trans-disciplinary research and extension project that engages multi-disciplinary scientists, dairy farmers, and industry professionals. Dairy farms have embraced large and diverse technological innovations such as sensors and robotic systems, and procured vast amounts of constant data streams, but they have not been able to integrate all this information effectively to improve whole farm decision-making. Consequently, the full impact of all this new Smart Dairy Farming is not being fully realized. It is imperative to develop a system that can collect, integrate, manage, and analyze on-farm and off-farm data in real-time for practical and relevant actions. To date, we have successfully implemented real-time, integrated, Big Data streams from 3 farms in Wisconsin. The warehouse connects daily cow, herd, farm, weather, and economic data. This involves cleaning and normalizing the data as well as retrieving data on demand. We are using the state-of-the-art database management system from the University of Wisconsin-Madison Center for High Throughput Computing. We demonstrate our predictive analyses concept by providing some illustrations of practical applications using integrated data streams such as studying the evolution of the feed efficiency, milk income over feed cost, mastitis incidence and severity, and survival analyses. We are securely advancing towards our overarching goal of developing our "Virtual Dairy Farm Brain." This is an ongoing innovative project that is anticipated to transform how dairy farms operate.*

Keywords: data warehouse, data integration, simulation, modeling, optimization

Introduction

Greater productivity, rather than a larger dairy herd, will be the primary aim to meet the increased demand for dairy products in the foreseeable future due to increased regulations and consumer preferences. The challenge is then to optimize the dairy farm system to become even more efficient. Milk production is a function of genetics and management (Shook, 2006) and significant advancements have been made in genetics (Shook, 2006), nutrition (VandeHaar and St-Pierre, 2006), reproduction (Moore and Thatcher, 2006), health (LeBlanc et al., 2006), cropping systems (Tilman et al., 2002), and management that have resulted in productivity risen by 40% during the last 50 years (Oltenacu and Broom, 2010). However, new challenges arise with these changes and additional improvements are required. Farm management evolves constantly and good-quality, real-time, and integrated data can help herd managers to optimize the production system. Additional improvements will push the bio-physiological limits of the cow and the whole system. Furthermore, improvements in some management areas could be counterproductive for other management areas. For instance, more productive cows have reproductive problems (Lucy, 2001), may be less efficient converting feed to milk, and produce more waste per unit of milk as they become bigger in size (Manzanilla-Pech et al., 2016). Adding to this complexity, market patterns and weather conditions are highly unpredictable. The market of dairy products has become much more uncertain (Jesse and Cropp, 2008) and as such, the dairy industry is experiencing unprecedented market and milk income minus feed costs volatility (Gould, 2017). There is a need for decision-support tools and projections that account for biological, price, and weather uncertainties inherent to the production system (Mirando et al., 2012). To achieve this, real-time, integrated Big Data analytics decision-making is vital (Wolfert et al., 2017).

Dairy farms have embraced technological innovations and nowadays count with massive permanent data streams. A tremendous amount of dairy farm related data is constantly generated and includes on-farm data such as milking, feeding, or reproduction, and off-farm data such as weather and prices. Animal scientists are part of the Big Data revolution (Madrigal, 2012). The introduction of new technologies on the farm, together with the Internet of Things to manage livestock production using the principles and techniques of process engineering, becomes Precision Livestock Farming (Wathes *et al*., 2008; Bewley and Russell, 2010). However, since these new technologies evolve rapidly, dairy farm managers and animal scientists are not ready to overcome the new challenge and therefore take full advantage of the opportunities (Madrigal, 2012).

Dairy producers use separate software tools to visualize and interpret all these data streams and to make isolated decisions. It is not surprising that some dairy farms do even have dedicated computers to a specific farm software when all these data streams are inter-related. Consequently, it is difficult, if at all possible, integrate these data streams for practical applications. Dairy farmers have not been able to efficiently integrate them to support improvements of the whole-farm management and decision-making. It is crucial to develop a system that can collect, integrate, manage, analyze, and project on-farm and off-farm data in real-time for practical and relevant actions. Some managers would occasionally merge datasets with specific purposes and punctual analyses, but these are not permanently integrated and the analyses become inconsistent throughout time. The lack of integration and its subsequent analysis and projection generate different problems such as: delays in optimal actions; increased risk of mistakes and failure; lack of awareness of a changing environment; sub-optimal use of on- and off-farm resources; narrow vision of opportunities for improvement; and ultimately sub-optimal profitability and consequently decreased sustainability and resilience.

In light of the above discussion, our objective in this manuscript is two-fold and consists of: 1) develop an integrated real-time Big Data warehouse for the dairy farming operation and 2) demonstrate the feasibility and application of integrated real-time data visualization and Big Data analyses to support optimal decision-making. A follow up third objective that is beyond the scope of this paper consists of fully implementing the real-time data analytics and developing farm-specific customizable decision support tools for practical application on dairy farms.

Materials and Methods

The Virtual Dairy Farm Brain team is currently collecting on- and off-farm data on a permanent basis from selected Wisconsin dairy farms and other sources. All these Big Data flow are being stored in a server located at the University of Wisconsin-Madison premises. These data are being normalized and integrated into a data warehouse on real-time to visualize and perform integrated artificial intelligence analytics for improved decision-making.

The team

Due to the challenges and opportunities and within the complexity involved in this research framework, the team conformation is a critical factor for successful achievement of the stated objective. We deemed necessary to strongly connect two scientific disciplines, dairy and computer science, that are rather dissimilar, but highly synergetic and complementary for this purpose. The team includes 3 faculty from Dairy Science, one with expertise in management, one in genetics, and the other in nutrition. The team also includes 3 faculty from Computer Science, one with expertise in database management, one in artificial intelligence analytics, and the other in the intersection of both, database management and data analytics. Important part of the team are 4 postdocs, 2 in either department, all of whom are well versed on data analytics. The 4 postdocs interact permanently and complement very well among themselves. Two additional master students, one in each department, and a few undergrad students support the work of the postdocs and faculty. The team also includes 2 experts in server maintenance and one expert on data collection, data warehouse design, and database maintenance. The team has vast experience in data analytics related to and/or applicable to dairy farm management, even though previous experiences had not, or only partially, integrated real-time Big Data streams from dairy farms. Examples include dynamic programming optimization for cow replacement decisions (Cabrera, 2010; Kalantari et al., 2015); Markov chains to select the best herd-level reproductive programs (Giordano et al., 2012) and individual cow reproductive management (Cabrera, 2012); machine learning to predict insemination outcomes (Weigel, 2016; Shahinfar et al., 2014); Monte Carlo stochastic simulation to improve feeding efficiency (Kalantari et al., 2016); and constructing complex machine learning pipelines using natural language (Leo John et al., 2017).

The farms and the data

Farm managers from 3 prominent dairy farms in Wisconsin are collaborating and are sharing all their data streams for the Virtual Dairy Farm Brain project. The data being collected and the software or services used for such collection are summarized in Table 1 and Fig. 1 and, in general, include:

Herd management that is used to keep records of routine and operational activities in the herd, e.g., reproduction, calvings, health such as vaccinations, presence of diseases, and treatments.

Milking system collects the data recorded during the milking process, e.g., milk volume, milk conductivity, milking speed.

Genetic/Genomic keeps records of the genetic data of all animals, calves, heifers, and cows, e.g., pedigree, total performance index, net merit.

Dairy Herd Improvement (DHI) data records monthly test-day visit variables, e.g., milk volume, somatic cell count, milk fat and protein content, total amount of fat and protein in the current lactation.

Feed monitoring records information related to the nutrition process of the animals individually or grouped by pens, e.g., diet composition, average consumption per day, dry matter intake, cost per kg of ration.

Milk processor data reports the milk composition in each milk shipping load, e.g., milk fat and protein content, somatic cell count.

All collected data are stored in the Center for High Throughput Computing (HTCondor server, http://chtc.cs.wisc.edu) from the University of Wisconsin-Madison. Before storing, all data are properly converted to readable format and checked for duplication. These files populate a multidimensional database system. Multidimensional databases are advantageous because facilitate the implementation of different analyses and visualization that otherwise would be very complicated (Chaudhuri and Dayal, 1997). All different data streams being collected and their analysis and validation are depicted in Fig. 1 and 2.

Table 1. Data collected and software or service used by the 3 participant farms for the Virtual Dairy Farm Brain project.

Data collected	Farm 1	Farm 2	Farm 3
Herd management	DairyComp, http://web.vas.com/updates/dairycomp		
Milking system	SmartDairy[1], AlPro[2], DairyPlan[3]	SmartDairy –pen level	SmartDairy –cow level
Genetic/Genomic	Enlight, https://www.enlightdairy.com		
DHI data	AgSource, http://agsource.crinet.com		

Feed monitoring	Feed Supervisor[4]	TMR Tracker[5]	FeedWatch[6]
Economic data	Understanding Dairy Markets, http://future.aae.wisc.edu		
Milk processor	Foremost[7]	Grande[8]	Grande

[1] https://boumatic.com/us_en/products/smartdairy-us
[2] http://www.delaval.com.au/en/-/Product-Information1/Management/Systems/ALPRO/
[3] https://www.gea.com/en/products/dairy-plan-c21.jsp
[4] http://www.supervisorsystems.com/software/category_30dfe68ab33c/product_368d35405169/
[5] https://digi-star.com/solutions/2-2/TMR_Tracker
[6] http://dairyone.com/feedwatch/
[7] http://www.foremostfarms.com
[8] http://www.grande.com/Pages/Welcome.aspx

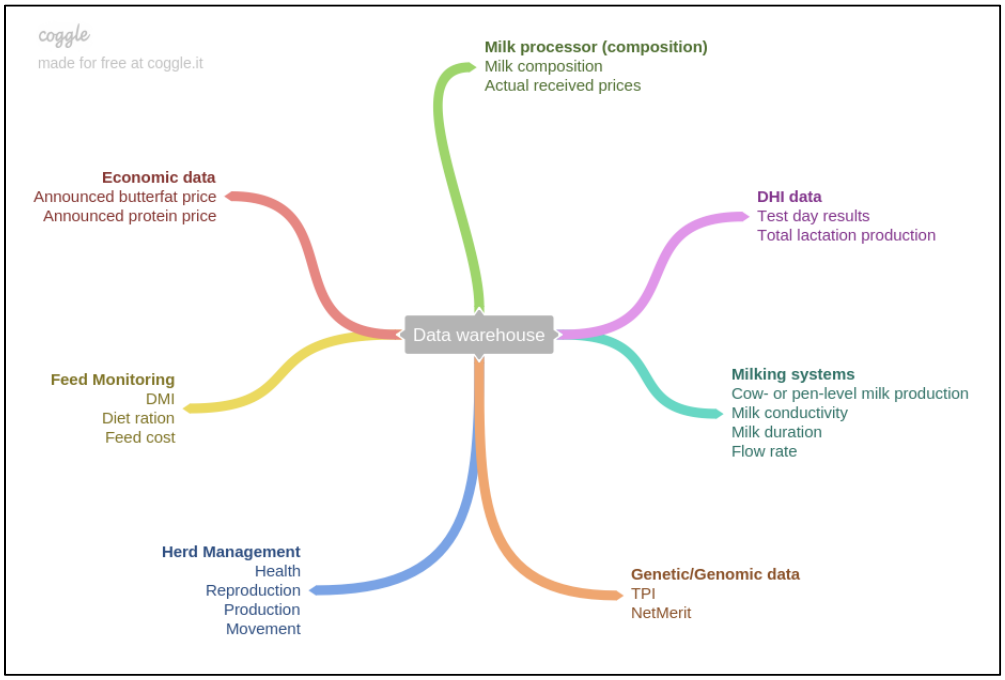

Figure 1. Real-time data stream sources currently collected by the Virtual Dairy Farm Brain project.

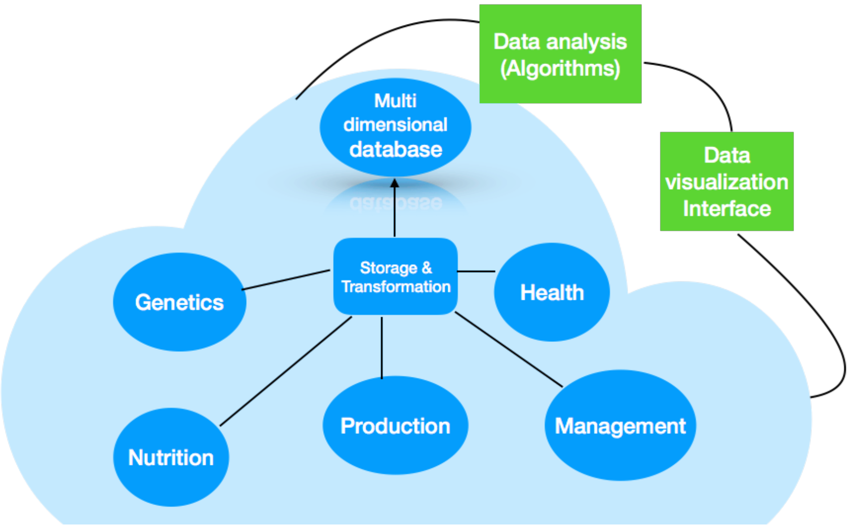
Data analysis (Algorithms)
Multi dimensional database
Data visualization Interface
Storage & Transformation
Genetics
Health
Nutrition
Production
Management

Figure 2. Schema of data flow and transformation for the Virtual Dairy Farm Brain project.

Data normalization and editing

We follow different normalizing and editing steps to match the data, select the variables, and develop readable datasets. The key variables selected to match the data from individual animals are cow identification number (ID), date of birth (BDAT), and electronic identification number (EID). Although cow ID is the most commonly variable used in the daily farm management decision-making, it may not be properly included in some cow-level data sources. Instead, EID is always recorded in the cow-level data because the radio-frequency chip reader identifies EID directly from cows. To identify the cows, it is necessary to use combinations of the different variables depending on the software to match. For example, connecting the herd management software (DairyComp) with the milking system software (SmartDairy) needs a 2-step matching process because SmartDairy does not record individual cow's ID properly. The first step is to match the EID in SmartDairy with the EID in DairyComp. The second step is to find the matching ID in DairyComp based on the cow's EID. In addition, EID is a character variable in DairyComp that first needs to be normalized into a numeric variable. Some of the EID in DairyComp may also start with the characters "USA," which needs to be converted to "840" before matching with the EID from other sources. To avoid the potential cow ID duplication, BDAT is also used when matching cows ID. A summary illustration of the different variables used to match the software is depicted in Table 2.

Table 2. Key variables used to match different datasets from 5 different data sources in Farm 3.

Data Source	DairyComp	SmartDairy	Enlight	FeedWatch	AgSource
Cow ID (ID)	NUM[1]	NUM	NUM	-	NUM
Electronic ID (EID)	CHAR[2]	NUM	-	-	NUM
Date of birth (BDAT)	NUM	-	CHAR	-	-
Lactation (LACT)	NUM	-	-	-	NUM
Pen	NUM	NUM	-	NUM	-
Date	NUM	NUM	-	NUM	-

[1] Numeric, [2] Character.

The feed monitoring system keeps records of the diets supplied to a group of animals (pen-level). It does not record individual cow-level feeding. Consequently, individual cow's feed intake can only be approximated as the average consumption of the pen. For such calculation, cow's pen allocation is needed, which is extracted from the management software. All the 3 participant farms use DairyComp as the primary farm management software to schedule and record all the different farm events. We extract and merge the ID of the animals registered in this software with the other software as presented in Table 2.

Management software

We separate different variables or events recorded in the herd management software DairyComp by their type to better design the data warehouse, facilitate data processing, and posterior analyses. Consequently, we classified variables as Health, Reproduction, Production, and Management. An illustration of the Health events on Farm 3 extracted from DairyComp is presented in Table 3.

Table 3. Events classified as Health events recorded in the herd management software (DairyComp).

Event	Frequency
CULTURE	15,361
CYSTIC	479
DIARRHEA	3,892

ILL MISCELLANEOUS	1,564
INJURY	614
KETOSIS	2,103
LAME	11,579
LEFT DISPLACED ABOMASUM	622
MASTITIS	13,856
MASTITIS NO TREATMENT	7,221
METRITIS	2,342
MILF FEVER	327
PINK EYE	784
PNEUMONIA	8,482
RIGHT DISPLACED ABOMASUM	40
RETAINED PLACENTA	900

However, within each farm and even inside a farm throughout time, events can be named differently. This is a flexibility provided by DairyComp that becomes a major challenge in the data cleaning process. Consequently, we identify and normalize the events of the 3 participant farms by checking the language patterns in the events or variable remarks. This process will be automatized in the future. An illustration is presented in Table 4.

Table 4. Different names of same variables used by participant farms, their description, and standardized name to *Event1*.

Farm 1	Farm 2	Farm 3	Description	Event1
CULTURE	CULTURE	CULTURE	Culture sample	CULTURE
CYST	CYSTIC	CYSTIC	Cystic ovary	CYSTIC
DA	DA	RDA-LDA	Displaced abomasum	DA
SCOURS	DIARHEA	DIARRH	Diarrhea	DIARHEA
INJURY	INJURY	INJURY	Injury	INJURY
KETOSYS	KETOSIS	KET	Ketosis	KETOSYS
FEET	LAME	LAME-LAME1	Lameness	FEET
MAST	MAST-MASTEVL	MAST-MASTNT	Mastitis	MAST
MET	METR-ACUTEMET	MET	Metritis	METR
MLKFVR	MF	MF	Milk fever	MF
RESP	PNEU	PNEU	Respiratory problems	RESP
PREV	PREV	PREV	Prevention	PREV
RP	RP	RP	Retained placenta	RP
POSILAC	BSTART	BSTON	rBST hormone start	BST START
POSSHT	BSTOP	BSTOFF	rBST hormone stop	BST STOP
MOVER	MOVE	MOVE	Move animal	MOVE
PROST	PGF	PG90	Prostaglandin	PG90

To avoid inconsistencies in the data, we adapted the St-Onge et al. (2002) data editing process. Only animals with at least 1 recorded calving and with information for milk production (DHI records) are included in final database. Table 5 shows a fragment of the data editing process.

Table 5. Number of records deleted at each step of editing procedure.

	Records	
Editing criteria	Removed	Remaining
Initial number of animals with records.		34,031
Animals with no registered calvings (heifers or sold before calving).	11,343	22,688
Animals with productive life before year 2007.	4,804	17,884
Animals that calved for the first time but there is no further information for milk production (removed from the farm immediately after calving).	1,775	16,109
Outlier animals with age at first calving before 18 months or after 44 months of age	5	16,104

Construction of the data warehouse

Ten tables of data from different dairy production aspects are initially extracted from the different data sources. These tables are: 1) animal information, 2) genetics (DairyComp), 3) genetics (Enlight), 4) reproduction, 5) health, 6) management, 7) every milking records, 8) lactation-level production summary, 9) DHI records, and 10) feed monitoring. The records are then matched for the selected cows and herds following the criteria presented in Table 2 using SAS® 9.4 (SAS, 2018).

We edit and clean the data to avoid input errors and incorrect calculations. Most of this process is now automated, although a significant portion of the cleaning and transformation work was first manually done. As such, the first integrated prototype database has been developed in Microsoft Access 2016 (Fig. 3). The main goal was to integrate datasets from different sources allowing queries using the integrated data streams to facilitate, first, descriptive analytics and, second, predictive tool models. Key variables are assigned on Herd, ID and BDAT for establishing one to one or one-to-many relationships within the datasets (e.g., one animal from the animal file has 4 different lactations, and each lactation has several records in the DHI dataset). An illustration is provided in Table 6, which shows the number of records included in each relational database.

Table 6. Summary of the number of records included in the relational database

Tables	Farm 1	Farm 2	Farm 3
Animal information	2,621	2,344	11,063
Genetics (DairyComp)	2,621	2,344	11,063
Genetics (Enlight)	-	1,149	6,561
Reproduction	31,643	34,904	146,677
Health	17,555	1,992	102,304
Management	34,414	20,384	481,819
Every milking records	-	252,853	1,826,301
Lactation-level production summary	6,557	4,913	24,623
DHI records	65,655	40,569	249,239
Feed monitoring	-	1,524	2,019

The nature of the relations and the different data sources integrated are shown in Fig. 3. For this prototype database, 5 data sources were integrated: herd management (DairyComp), DHI records

(AgSource), genetics (Enlight), every milking records (SmartDairy, ALpro, DairyPlan) and feed monitoring (FeedSupervisor, TMR tracker, FeedWatch).

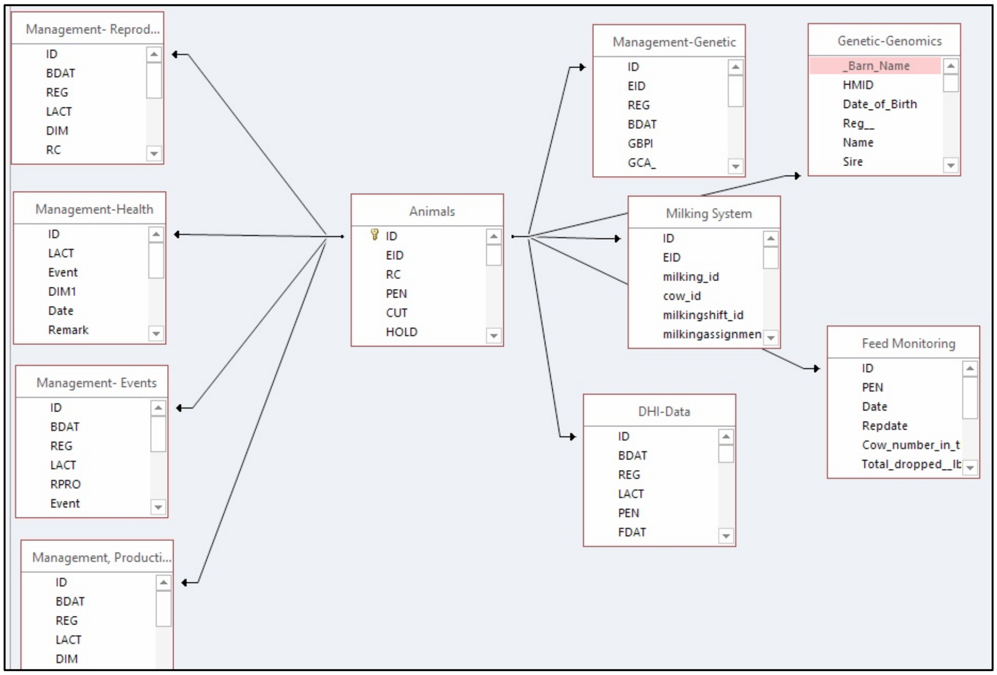

Figure 3. Relational schema of data streams extracted from the different software and services included in the database.

Data preparation and analysis

Descriptive and predictive data analyses need further preparation on the information stored in the prototype database, such as removing the missing data, removing outlier data, or understanding the language pattern of the event's remarks. To estimate the mastitis withdrawal period after mastitis treatment in Farm 3 (discussed later), the data preparation followed the procedure in Fig. 4 and was conducted in R 3.4.1. (R Core Team, 2017). The first step after retrieving all the milking records from cows that had mastitis ("MAST") from the prototype data warehouse was to sort the extracted data by date in ascending order. The next 3 steps were to remove mastitis events that occurred before the data were available from the milking system (May 2016), occurred in dry cows, or affected cows that were either removed or dried soon after the case of mastitis. The next step consisted of cleaning up the remaining data due to errors in milking records such as missing EID, incorrect date, outliers, etc. The following step was to categorize the mastitis case by severity according to the event remarks. Then, finally, we estimated the withdrawal period by using the aggregated data from the 3 following milking recordings after the mastitis event occurred.

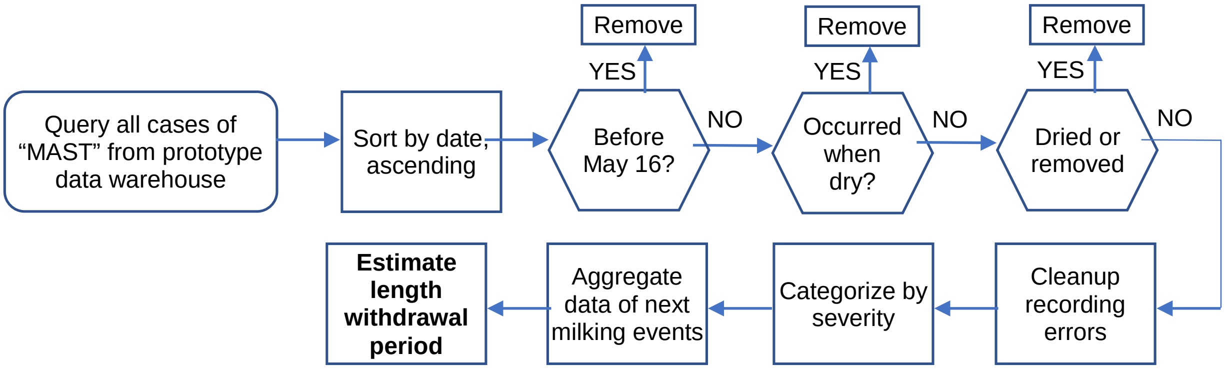

Figure 4. A flowchart of data preparation process to estimate the milk withdrawal period after mastitis antibiotic treatment in Farm 3.

Results and Discussion

We have developed a real-time prototype integrated database that connects the most important data streams in 3 Wisconsin farms. This is a forward-looking task in which data are constantly flowing to the database in real-time at the moment data are being collected on the dairy farms by different software or service systems. Since it is necessary background and historical data for our initial analyses and under the possibility to also collect historical records, we included these historical records in our database, to the extent possible, at the beginning. We were able to retrieve historical data from cows that started production in 2007. This collection allows to trace back the available information for retrospective and longitudinal studies. The distribution of recorded lactations and the average age at the calving by lactation are described in Tables 7 and 8.

Table 7. Distribution of recorded lactations by dairy farm from 2007 to 2017.

Lactation	Farm 1	Farm 2	Farm 3
1	2,648	2,353	11,098
2	1,780	1,370	7,068
3	1,097	641	3,891
4	609	321	1,693
5+	423	228	865

Table 8. Average age and standard deviation at the moment of freshening by lactation and dairy farm for cows that calved between 2007 to 2017.

Lactation	Farm 1	Farm 2	Farm 3
1	24.01±1.35	24.32±2.01	23.52±1.97
2	36.9±2.19	37.45±2.95	36.87±2.98
3	49.97±2.88	50.93±3.96	49.94±3.6
4	62.99±3.48	64.46±4.54	62.83±4.47
5+	76.03±4.21	77.11±4.99	75.53±5.27

The distribution of age at first calving and culling for the animals included in the database are summarized in Fig. 5.

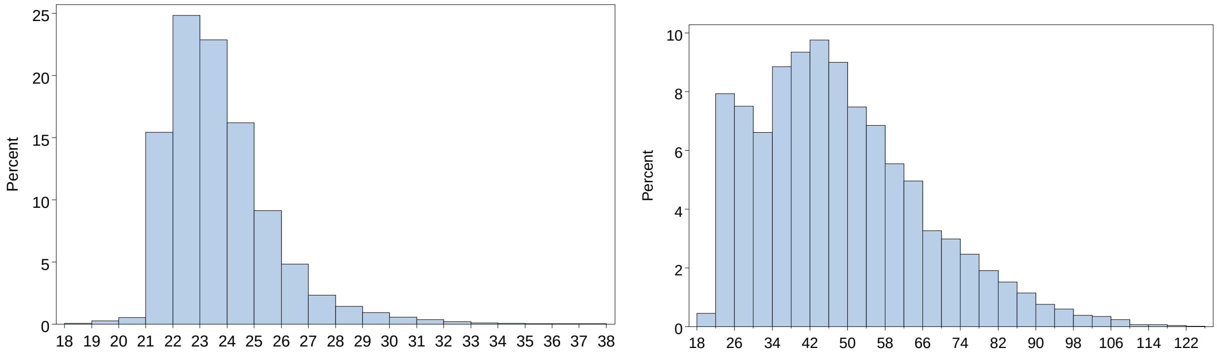

Figure 5. Distribution of average age at the moment of first calving in months (left) and distribution average age in months at culling (right) for the 3 studied dairy farms between years 2007 and 2017.

Some productive, performance, health, and reproductive indicators for the 3 farms for year 2017 are summarized in Table 9. As seen, Farm 2 has the highest milk production (16,249 kg/cow per lactation), the shortest cow lifetime (44 months), and the highest culling rate after 1st calving (43.17%). Farm 1 has the highest average of somatic cell counts (286,000 cells/ml), an indicator of subclinical mastitis, which is consistent with having the greatest percentage of clinical mastitis cases in the 1st month of lactation (22.2%). Farm 3 is in the middle among the 3 farms with respect to dry period length (57 days) and calving to conception interval (117 days). With this integrated and now becoming real-time database, it is possible to identify factors influencing these patterns and their effects on overall performance and profitability.

Table 9. Performance of the 3 studied herds during year 2017

Productive and performance indicator	Farm 1	Farm 2	Farm 3
Average milk production per cow (kg/cow per lactation)	14,240	16,249	14,094
Mature equivalent production at 305 days (ME305, kg)	12,864	15,161	13,724
Average lactation length (days)	335	342	341
% Culling rate (after the 1st calving)	25.9	43.17	35.5
% Death (after the 1st calving)	4.5	3.64	4.9
Culling age (months)	52	44	48
Udder health indicators			
Average bulk tank somatic cells count (1,000 cells/ml)	286	113	125
Clinical mastitis in 1st lactation month (%)	22.2	20.1	16.4
Reproductive indicators			
Average dry period length (days)	54	63	57
Calving to conception interval (days)	113	129	117

Practical Uses of the Integrated Real-Time Database Warehouse

Online query and dashboard from warehouse

A working-in-progress outcome is the development of a web-portal to retrieve and visualize real-time cow- and herd-level data from our Big Data live warehouse. At the moment, users can plot on demand an individual cow's daily milk production and compare it with the daily herd-level milk production across a selected time period. Users can combine those data with cow's recorded health, reproduction events, or other management actions. For example, this online data retrieval is allowing to quantify an individual cow's daily milk production change after such cow was moved from one pen to another. This specific application assists dairy producers in modifying and optimizing their cow grouping strategies. Furthermore, this online system is also proving to be extremely useful for research purposes in supporting real-time Big Data analytics within our vision of the "Virtual Dairy Farm Brain" framework that comprises a highly inter-disciplinary team of researchers who require diverse datasets retrieved efficiently and effectively on demand as needed.

Feed efficiency and income over feed cost

Feed efficiency is defined as milk produced divided by the dry matter feed consumed. Daily feed efficiency can be calculated by integrating the daily cow-level (Farm 3) or pen-level (Farm 2) milking records (from the milking system software, SmartDairy), pen-level dry matter intake (from the feed monitoring software, FeedWatch in Farm 3 and TMR tracker in Farm 2), and cows' daily pen allocation (from the herd management software, DairyComp). An illustration for Farm 3 is depicted in Fig. 6.

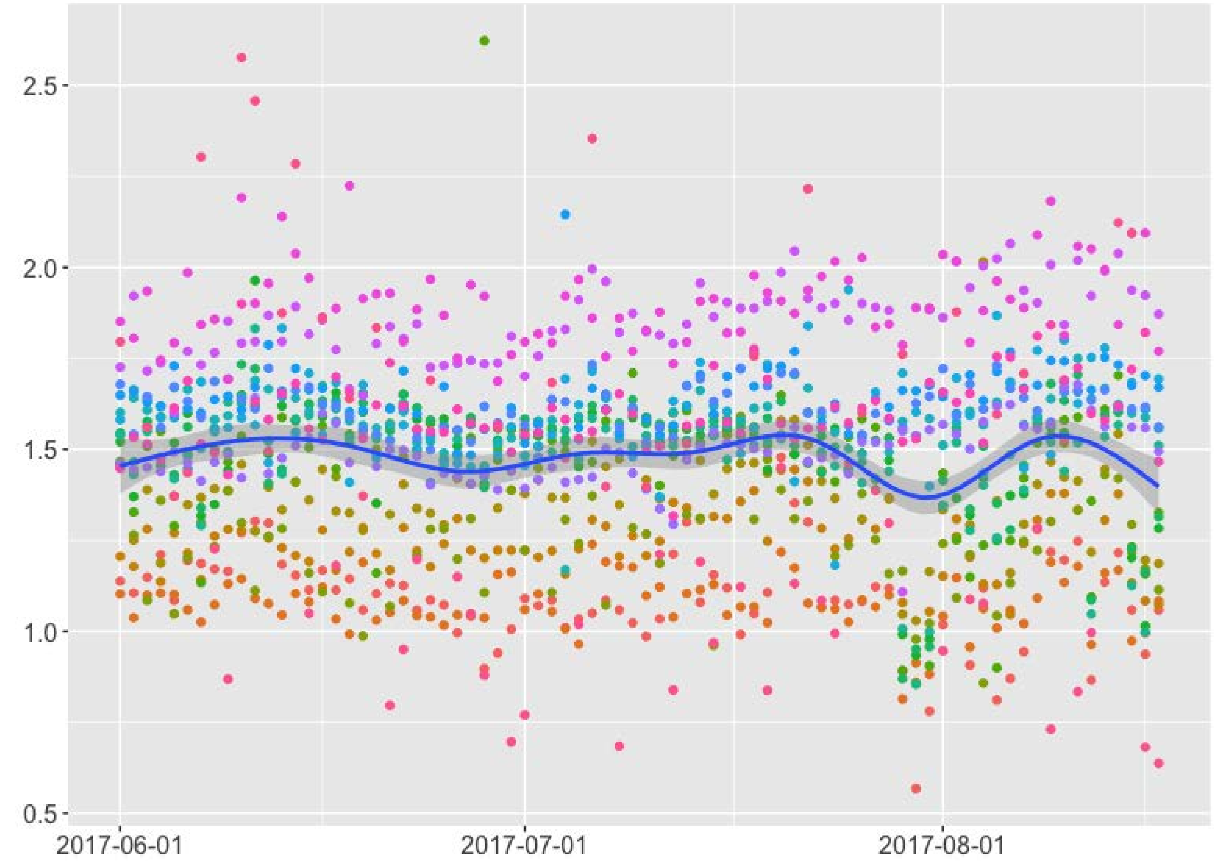

Figure 6. Daily pen (dots) and farm (line) feed efficiency for Farm 3 for selected days in 2017. Feed efficiency = kg milk /kg dry matter feed consumed.

Furthermore, milk income over feed cost can also be calculated, defined as the difference in revenue from milk sales minus the investment in feed (an illustration for Farm 2 is depicted in Fig. 7). To achieve this, additional data integration is required. Milk composition data (butterfat and protein content) is extracted from the milk processor database (Grande). Milk component prices is collected from an external economic database (Understanding Dairy Markets, http://future.aae.wisc.edu). Feed costs are finally collected from the feed monitoring software. Milk income over feed cost, which is the largest determinant of dairy farm profitability (Kalantari et al., 2016), is a key performance indicator that can assist with the effective management policies such as longevity or production goals for tactical or strategical critical farm decisions. It is clear by looking at Fig. 7 that there is a seasonal pattern on milk income over feed cost with much greater variability (uncertainty) during the months of August and September, the end of the summer season in which cows might have had experienced heat stress. Later in the year, variability decreases and milk income over feed cost increases, which would be a function of increased productivity, improved feed efficiency, better market conditions or a combination of all or some of these factors.

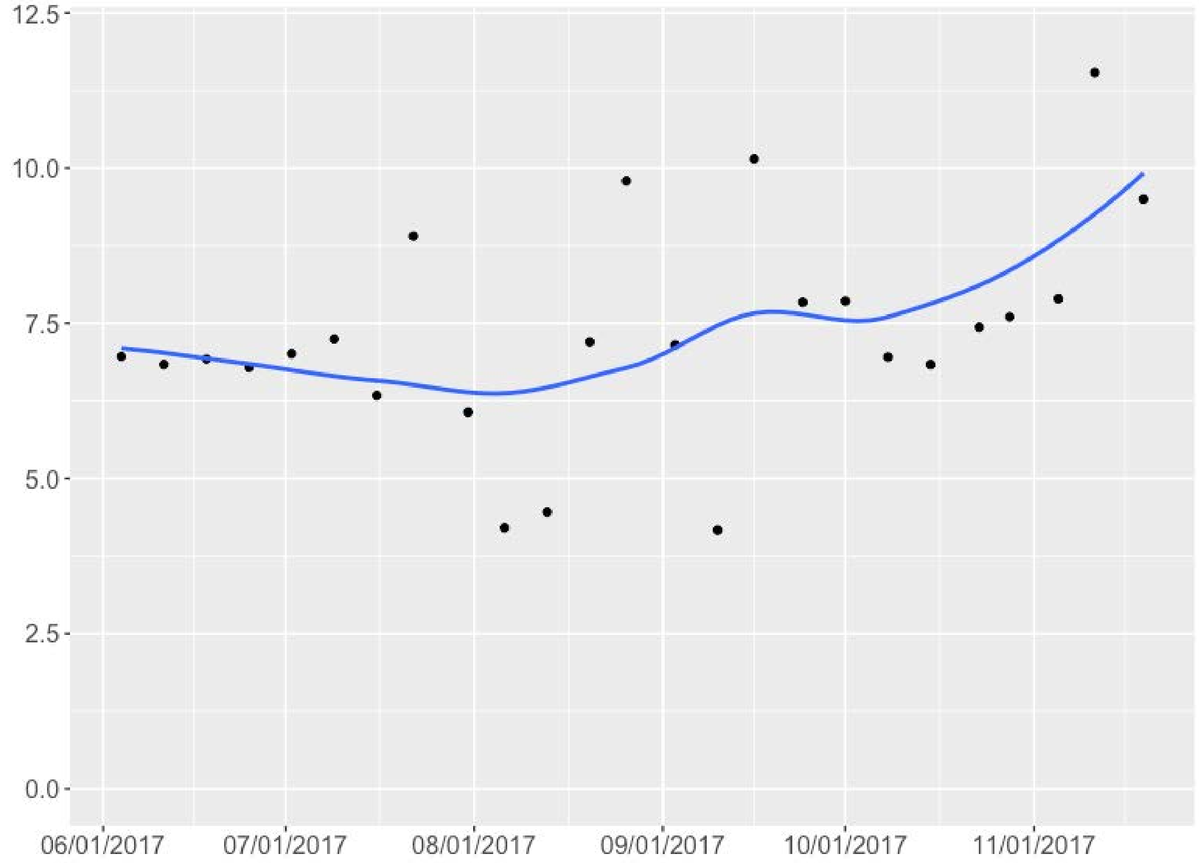

Figure 7. Weekly farm (dot) milk income over feed cost ($/cow per day) in Farm 2 for selected months in 2017. The line is the trend of farm-level milk income over feed cost.

Having access to real-time integrated data streams from the farm (production, performance) and external sources that affect the farm (weather, market) provides farmers a great competitive advantage for improved systematic and data-based decision-making. This rather simple, but powerful illustration demonstrates the real and practical possibility to effectively integrate large data streams and use them in real-time for practical purposes.

Mastitis incidence and antibiotic treatment withdrawal period

Another example of the use of the integrated database at the management level, is the analysis of the withdrawal of milk due to antibiotic treatment. Distribution of milk withdrawal period length for cows diagnosed with mastitis of different severity levels (mild, moderate and severe) and prescribed with antibiotic treatment in Farm 3 is depicted in Fig. 8. This analysis requires the integrated cow-level mastitis records from the herd management software (DairyComp) and cow-level milking records from the milking system (SmartDairy). In the near future, we will combine these results with survival analysis and profitability indicators. This will help farmers understand the time of survival of animals after similar cases of mastitis and their profitability, so managers can anticipate their actions at the time they encounter a new case of mastitis.

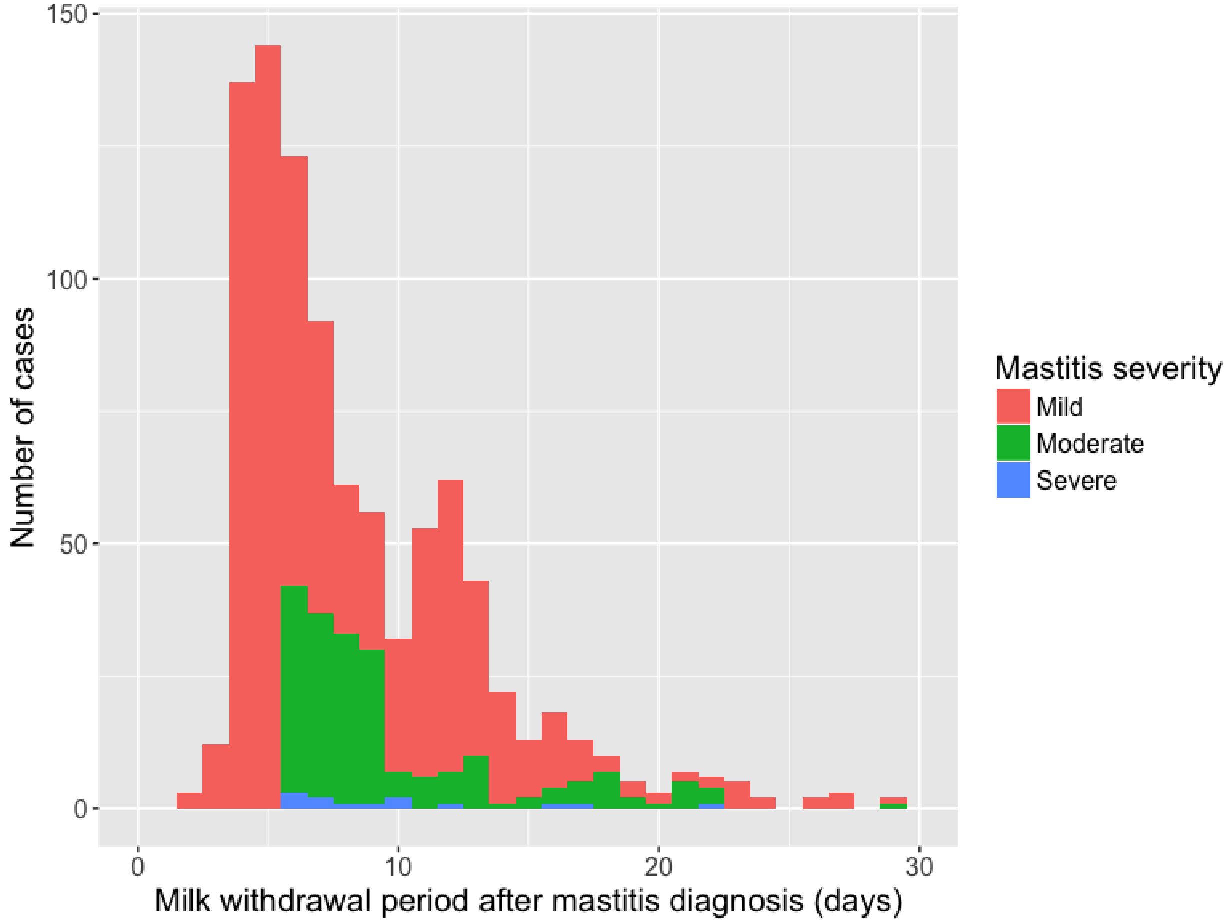

Figure 8. Milk withdrawal period length after mastitis diagnosis between 2016 and 2017 in Farm 3.

Real-time datasets provide the great opportunity of learn from current and past data and use artificial intelligence (e.g., machine learning) to anticipate actions, likely even before they occur. In the specific case of mastitis, for example, it will be impactful to have a probabilistic estimate incidence based on historical cow-level records and current situation, so actions can be optimized earlier. We anticipate this process to be dynamic and interactive, meaning that the prediction accuracy will improve as more integrated real-time data is available. Similar to farmers gaining experience and expertise throughout time, we expect the system to “learn” as it goes, so projections, and consequently decisions, will improve over time. In order for this to occur, we need the functional real-time integrated database and high sophisticated data analytics that include artificial intelligence.

Survival analysis

Knowing the reasons and times when cows are leaving the herd is critical for dairy farm production management and decision-making (Cabrera, 2010). Within our dynamic integrated framework, we will combine survival analysis with other data analysis. As above alluded, survival probability curves will be part of the analysis of mastitis and other diseases. Survival analysis will be used as an additional indicator to optimize decision-making at the time the disease is detected or when there is a high risk for the animal. For example, an illustration of survival analysis follows. We calculated survival probabilities with Proc Lifetest (SAS 9.4) for cows removed from their herds between 2007 and 2016 and that calved at least once in their lifetimes (Fig. 9). The curves indicate that on average 50% of the animals are being removed from their respective herds at around 1,394 days of age (45.85 months). Farm 2 has the most aggressive culling pattern, where 50% of its cows are removed before reaching 1,144 days of life, whereas Farm 1 has the least aggressive culling pattern in which 50% of the animals are removed by 1,509 days of age, a difference of 365 days. Farm 3 is in between, in which 50% of the animals are removed at 1,417 days of age. From this simple but enlightening analysis, it is possible to infer that projections are going to be different by farm. Consequently, these probabilistic farm-specific conditions should be included in any predictive or optimization analysis.

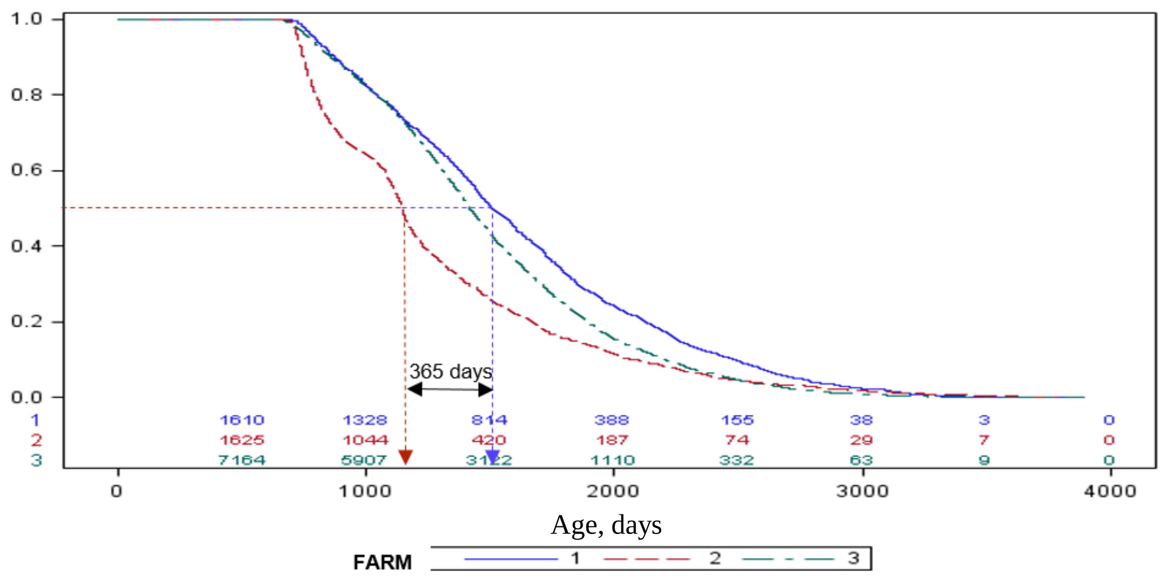

Figure 9. Survival probability curves for adult cows that calved at least once between years 2007 and 2016.

The importance of longevity (longer survival of the cows in the herd) is explained by 2 main factors. First, farmers that allow their animals stay for longer time in the herd will incur in a lower cost of replacement. This is because the overall cost of replacements is amortized over a longer time of production. And second, herds with longer longevity will allow cows to reach more productive lactations (Essl, 1998; Heikkilä *et al*., 2012). On the other hand, farmers who cull more aggressively allow for faster genetic improvement. Although, longevity has an undeniable importance for the success of the farm, milk production is the most important revenue in the profit equation. Around 90% of dairy farm revenues comes from milk production (VandeHaar, 2006). With the information collected in our live database, we can observe that the average mature equivalent expected milk production to 305 days (ME305 in kg; Table 9) for Farm 2 is 15,161, for Farm 3 is 13,724, and for Farm 1 is 12,864. The difference of ME305 between Farm 2 and Farm 1 is 2,297 kg. With this information, a new important question arises, does the higher volume of milk obtained per lactation in Farm 2 compensates for the shorter expected lifetime?

Another illustrative analysis applied to survival curves can be associated to mastitis incidence, under the knowledge that mastitis is one of the most frequent causes for animal removal (Bascom and Young, 1998; Bar et al., 2008). As an illustration, we selected all the animals that were removed in this herd because of "mastitis" (n = 857) between 2007 and 2016 to study the impact of clinical mastitis on longevity of productive cows in Farm 3. Then, we counted the number of episodes recorded of clinical mastitis per animal (3.7±2.73). For this farm, to consider a new mastitis episode, there must be at least a difference of 14 days between cases. Finally, we classified the animals by the number of lactation when they presented the first episode of mastitis. The survival probability curves from the moment of the first episode of clinical mastitis to the culling by lactation is displayed in Fig. 10. It is clear that animals that were diagnosed with a case of mastitis for the first time in lactation 1 or 2 present a higher life expectancy than animals that were diagnosed with mastitis for the first time during lactation 3 or higher. Once again, anticipation of what could potentially happen with an animal in the future according to current status (mastitis case) and historical events is of upmost importance for optimal dairy farm management. Although currently we do not understand all the factors intervening in the final decision of removing an animal in relationship with a case of mastitis, this is a working in progress.

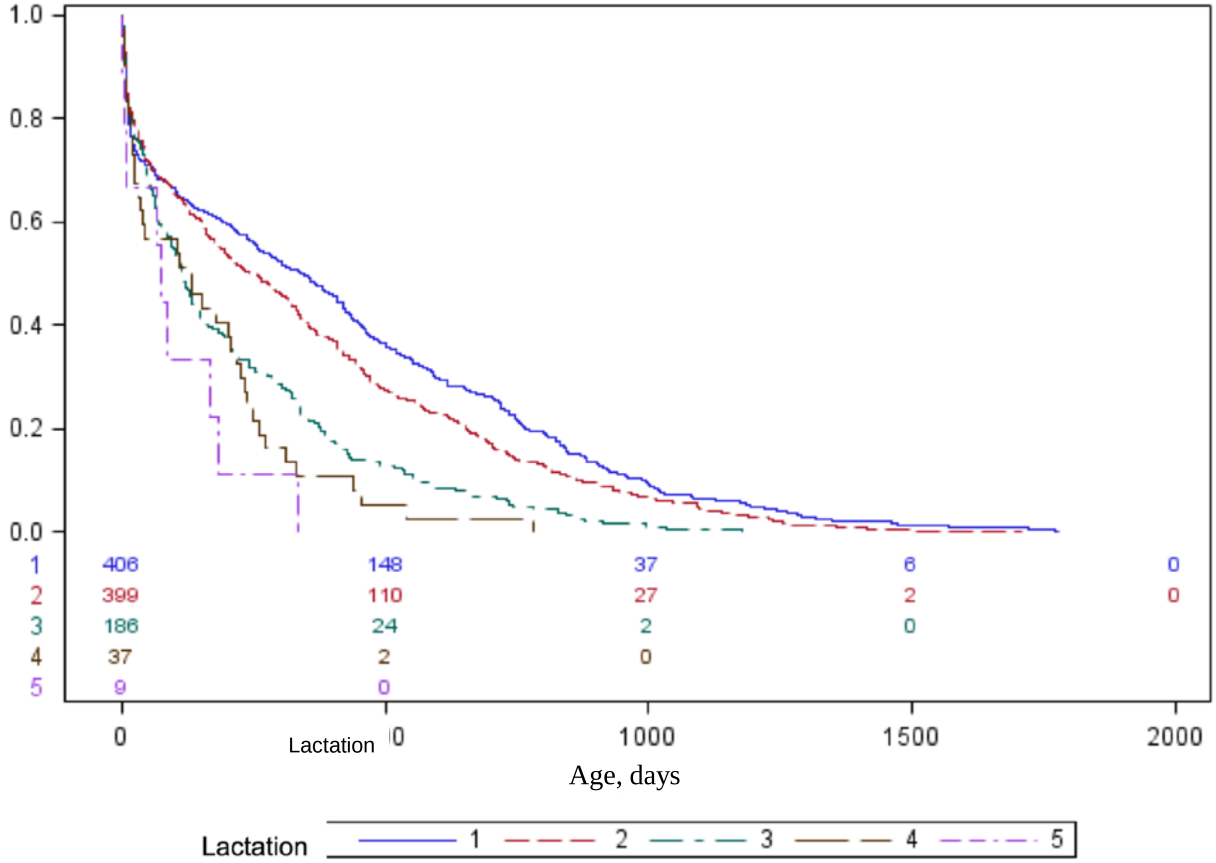

Figure 10. Survival probability curves for animals diagnosed with clinical mastitis for the first time by lactation for Farm 3.

We also envision to combine survival analysis with critical voluntary and economic culling decisions. To date, we have developed an improved daily Markov chain model inspired in Cabrera (2012) and Giordano et al. (2012). Different than previous research, the uniqueness of this model is that it is connected to cow and farm-level data streams within our integrated live database warehouse. Consequently, its projections are based on current status of the specific herd and are being updated in real-time. The model simulates each cow status using herd and cow transition probabilities matrices, which are critical. These matrices pertain to involuntary culling, mortality, pregnancy, abortion, diseases, etc. (Cabrera, 2012, De Vries, 2006) and in the past, had been used as averages of the industry (e.g., Giordano et al., 2012 using data from Pinedo et al., 2010) recognizing that these probabilities are herd and cow specific. Survival curves will be real-time created and connected with our Markov-chain model to optimize decisions of voluntary replacement, disease treatment, or differentiated breeding (as suggested in Cabrera, 2018).

In the near future, best herd-managers and top decision-makers will use real-time predictive tools to estimate the impact of their decisions. As the decision-making process becomes more complex, it is expected that decision-making will be a mix of human and computer factors with the help of Big Data analytics (Wolfert et al., 2017).

Conclusions

We have successfully implemented a real-time, integrated, Big Data live warehouse in 3 farms in Wisconsin. The warehouse connects daily cow, herd, farm, weather, and economic data streams. The work involves cleaning and normalizing the data as well as storing and providing an efficient way to query, visualize, and retrieve data on demand. We are using the state-of-the art database management contained at the University of Wisconsin-Madison Center for High Throughput Computing. We provide some illustrations of practical applications using integrated data streams on stored data such as studying the evolution of the feed efficiency, milk income over feed cost, mastitis incidence and severity, and survival analyses. These are proof of concept that are being connected with more sophisticated predictive models using artificial intelligence and Big Data analytics. We are securely advancing to our overarching goal of developing our conceptualized "Virtual Dairy Farm Brain." This is an ongoing innovative project that is anticipated to transform how dairy farms operate.

References

Bar, D., Gröhn, Y.T., Bennett, G., González, R.N., Hertl, J.A., Schulte, H.F., Tauer, L.W., Welcome, F.L. and Schukken, Y.H. (2008) Effects of repeated episodes of generic clinical mastitis on mortality and culling in dairy cows. *Journal of dairy science*, *91*(6): 2196-2204.

Bascom, S. S., & Young, A. J. (1998) A summary of the reasons why farmers cull cows. *Journal of Dairy Science*, *81*(8): 2299-2305.

Bewley, J.M. and Russell, R.A. (2010) *Reasons for slow adoption rates of precision dairy farming technologies: evidence from a producer survey*. In Proceedings of the First North American Conference on Precision Dairy Management.

Cabrera, V. E. (2010) A large Markovian linear program to optimize replacement policies and dairy herd net income for diets and nitrogen excretion. *Journal of dairy science*, *93*(1): 394-406.

Cabrera, V. E. (2012) A simple formulation and solution to the replacement problem: A practical tool to assess the economic cow value, the value of a new pregnancy, and the cost of a pregnancy loss. *Journal of dairy science*, *95*(8): 4683-4698.

Cabrera, V. E. (2018) Invited review: Helping dairy farmers to improve economic performance utilizing data-driving decision support tools. *animal*, *12*(1): 134-144.

Chaudhuri S., Dayal U. (1997) An Overview of Data Warehousing and OLAP Technology. *SIGMOD 26*

De Vries, A. (2006) Economic value of pregnancy in dairy cattle. *Journal of dairy science*, *89*(10): 3876-3885.

Essl, A. (1998) Longevity in dairy cattle breeding: a review. *Livestock Production Science*, *57*(1): 79-89.

Giordano, J. O., Kalantari, A. S., Fricke, P. M., Wiltbank, M. C., & Cabrera, V. E. (2012) A daily herd Markov-chain model to study the reproductive and economic impact of reproductive programs combining timed artificial insemination and estrus detection. *Journal of dairy science*, *95*(9): 5442-5460.

Gould, B. (2017) LGM_Dairy Analyzer, unpublished software system. *University of Wisconsin-Madison Dept. of Ag. and Applied Economics,* http://future.aae.wisc.edu/lgm_analyzer/.

Heikkilä, A. M., Nousiainen, J. I., & Pyörälä, S. (2012) Costs of clinical mastitis with special reference to premature culling. *Journal of dairy science*, *95*(1): 139-150.

Jesse, E., & Cropp, B. (2008) Basic milk pricing concepts for dairy farmers. *Economic Research*, *180*:200.

Kalantari, A. S., Cabrera, V. E., & Solis, D. (2015) A comparison analysis of two alternative dairy cattle replacement strategies: optimization versus simulation models. *Economía Agraria*, *18*: 12-24.

Kalantari, A. S., Armentano, L. E., Shaver, R. D., & Cabrera, V. E. (2016) Economic impact of nutritional grouping in dairy herds. *Journal of dairy science*, *99*(2): 1672-1692.

LeBlanc, S. J., Lissemore, K. D., Kelton, D. F., Duffield, T. F., & Leslie, K. E. (2006) Major advances in disease prevention in dairy cattle. *Journal of dairy science*, *89*(4): 1267-1279.

Leo John, R. J. L., Potti, N., & Patel, J. M. (2017) Ava: From Data to Insights Through Conversations. Eight Biennial Conference on Innovative Data Systems Research (CIDR). *Chaminade (CA), USA, 8-11Jan, 2017.*

Lucy, M. C. (2001) Reproductive loss in high-producing dairy cattle: where will it end?. *Journal of dairy science*, *84*(6): 1277-1293.

Madrigal C. A. (2012) The Perfect Milk Machine- How big data transformed the Dairy Industry. *The Atlantic (online article).* *https://www.theatlantic.com/technology/archive/2012/05/the-perfect-milk-machine-how-big-data-transformed-the-dairy-industry/256423/*

Mirando, M. A., J. M. Bewley, J. Blue, D. M. Amaral-Phillips, V. A. Corriher, K. M. Whittet, N. Arthur, and D. J. Patterson. (2012) Reinventing extension as a resource—What does the future hold?. Journal of Animal Science 2012(90):3677–3692.

Manzanilla-Pech, C. I. V., Veerkamp, R. F., Tempelman, R. J., van Pelt, M. L., Weigel, K. A., VandeHaar, M., ... & Hanigan, M. (2016) Genetic parameters between feed-intake-related traits and conformation in 2 separate dairy populations—the Netherlands and United States. *Journal of dairy science*, *99*(1): 443-457.

Moore, K., & Thatcher, W. W. (2006) Major advances associated with reproduction in dairy cattle. *Journal of Dairy Science*, *89*(4): 1254-1266.

Oltenacu, P. A., & Broom, D. M. (2010) The impact of genetic selection for increased milk yield on the welfare of dairy cows. *Animal welfare*, *19*(1): 39-49.

Pinedo, P. J., De Vries, A., & Webb, D. W. (2010) Dynamics of culling risk with disposal codes reported by Dairy Herd Improvement dairy herds. *Journal of dairy science*, *93*(5): 2250-2261.

R Core Team (2017) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., & Weigel, K. (2014) Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *Journal of dairy science*, *97*(2): 731-742.

Shook, G. E. (2006) Major advances in determining appropriate selection goals. *Journal of Dairy Science*, *89*(4): 1349-1361.

St-Onge, A., Hayes, J. F., & Cue, R. I. (2002) Economic values of traits for dairy cattle improvement estimated using field-recorded data. *Canadian journal of animal science*, *82*(1): 29-39.

Tilman, D., K. G. Cassman, P. A. Matson, R. Naylor, and S. Polasky. (2002) Agricultural sustainability and intensive production practices. *Nature* 418(6898):671-677.

Vandehaar, M. J. (2006) Alimentation, gestion et croissance des génisses laitières de remplacement. In *Proceedings 30 Symposium sur le bovin laitiere, Quebec*.

VandeHaar, M. J., & St-Pierre, N. (2006) Major advances in nutrition: relevance to the sustainability of the dairy industry. *Journal of Dairy Science*, *89*(4):1280-1291.

Wathes, C. M., Kristensen, H. H., Aerts, J. M., & Berckmans, D. (2008) Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall?. *Computers and electronics in agriculture*, *64*(1): 2-10.

Weigel, K. A. (2016) Machine learning methods for finding useful information in big, messy data sets. *Presentation held at the 31st American Dairy Science Association (ADSA) Discover Conference on Food Agriculture: Big Data Dairy Management. Chicago, USA, 1-4 Nov. 2016.*

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017) Big Data in Smart Farming–A review. *Agricultural Systems*, *153*: 69-80.