**Big Data and Athletics**  
Wayne Winston, Indiana University


**Digital Agriculture, Big Data and Cognitive Computing**  
Dr. Robin Lougee, IBM Research

Big Data is a timely theme for the 31st American Dairy Science Association’s Discover Conference on Food Animal Agriculture. The phenomenon known as Big Data is a revolution in our time. Thanks to a convergence of social and technological trends, data is being increasingly generated all around us, all the time. For instance, in the span of just two years from 2014 to 2015 more data was created than in the entire previous history of the human race. Innovators are tapping this unprecedented wealth of data to provide insights and business models that are changing how we live, work and play.

The proliferation of Big Data is ushering a new era of computing, which we call “cognitive computing.” Cognitive computing is a comprehensive set of capabilities designed to augment human intelligence and foster new discoveries across industries. For example, in human health care cognitive computing can advance precision medicine by ingesting patients’ electronic medical history and relevant medical literature, performing cohort analysis, identifying micro-segments of similar patients, evaluating standard-of-care practices and available treatment options, ranking therapies by relevance, risk and preference, and ultimately recommending the most effective treatments. Today IBM’s Watson for Oncology system is using cognitive computing to help provide oncologists the assistance they need to make more informed treatment decisions for patients.

Data is disrupting entire industries. Applications of big data technologies and cognitive computing hold the promise of enabling new approaches to difficult challenges. Initial results to perennial challenges of food safety and supply chain traceability in agriculture are encouraging.

Robin Lougee is the IBM Research Lead for Consumer Products & Agriculture. Robin serves on the Advisory Committee for the World Agritech Investment Summit and chairs the 2017 Syngenta Crop Challenge Award in Analytics prize committee. She is an industrial research scientist with a strong track record of delivering innovation to IBM and its customers. Robin pioneered the creation of COIN-OR, an open-source foundry for computational operations research, and led its growth to an independent non-profit that has served the scientific and business community for over 10 years. She was elected to the Board of INFORMS, the largest society in the world for professionals in the field of operations research, management science and analytics, Chair of the INFORMS Computing Society, and President of the Fora of Women in ORMS. She is an Associate Editor of *Surveys in Operations Research*. Robin earned a Ph.D. in Mathematical Sciences from Clemson University in 1993.

**Big Data Dairy Management: Big Picture**  
Miel Hostens & Bonny Van Ranst, University of Ghent

Miel Hostens is a post-doc assistant position on herd health management focusing on the optimization of productive and reproductive performances in small and large herds with an emphasis on nutrition at the Ambulatory Clinic of the Department of Reproduction, Obstetrics and Herd Health.
Bonny Van Ranst obtained his degree in Veterinary Medicine at the State University of Ghent, where he started working in the Ambulatoric Clinic afterwards. He left the faculty six years later to continue his work in dairy practice. An important part of his time went to the development of dairy management and herd health software. First there was Argos and then Uniform farm management software. Next to that he founded DDW (Dairy Data Warehouse). Nowadays his focus lies on the development of MmmooOgle, a cloud based big data dairy application. He is, as an independent academic consultant, a member of the Department of Reproduction, Obstetrics and Herd Health (prof. dr. Geert Opsomer) of the Faculty of Veterinary Medicine of Ghent. During his career he always kept a very close contact with daily dairy practice as he owns, together with his brother, several dairy farms in Belgium and France. He has been actively involved all through his career in a lot of lecturing for veterinarians and farmers.

Legal Issues for Ag Data
Todd J. Janzen, President, Janzen Agricultural Law LLC

The farm press has been abuzz in recent years about the issues surrounding agricultural data—that stream of data coming off the farm from fields, machinery, livestock equipment, and other places. Although it is common for these ag technology providers to claim “the farmer owns his data,” the reality is not so clear. This presentation will take a deeper look at what “ag data” is from a legal perspective, and whether the current law protects farmers’ privacy and security.

The presentation will also explain the efforts by farm organizations and ag technology providers to address farmers’ privacy concerns with ag data. The recent creation of the Ag Data Transparency Evaluator tool is one means for farmers to determine what really happens to the farmer’s data after it leaves the farm and is stored on an off-farm cloud-based server. The Evaluator is a web-based research tool that can help farmers understand the key provisions in their contracts with ag tech providers regarding ownership, usage, and sharing of data generated by technology products.

Todd Janzen serves as the administrator for the Evaluator. Todd will share what he has learned from the Ag Data Transparency Evaluator’s first few months of operation.

Todd J. Janzen serves as the Administrator for the Ag Data Transparency Evaluator, a national effort to bring transparency to contracts between farmers and technology providers. Todd is also an attorney and founder of Janzen Agricultural Law LLC, a law firm dedicated to serving the needs of farmers, ag technology providers, and agribusinesses. Todd grew up on a grain and livestock farm in Kansas. Todd is former chair of the American Bar Association’s Agricultural Management Committee and the Indiana State Bar Association’s Ag Law Section. Todd regularly publishes articles on the Janzen Ag Law Blog, which are reprinted on Farm Journal’s Agweb, Precision Farming Dealer magazine, and Hoosier Ag Today.

How Big Data Drive the Genomics Revolution
George R. Wiggans, Agricultural Research Service, USDA (retired)

The advent of low-cost genotyping of DNA markers in 2008 has led to a revolution in dairy cattle breeding. The volume and variety of single-nucleotide polymorphisms (SNPs) that have become available have created a need for high-speed processing. The SNP markers have facilitated rapid calculation of genomic evaluations with accuracy equivalent to that for a traditional bull evaluation based on early daughter records or greater than that for a cow evaluation based on several lactations. The availability of genomic evaluations early in an animal’s life has enabled a halving of the generation interval because bulls and cows can be selected as parents of the next generation as soon as they reach sexual maturity. The result is a doubling of the rate of genetic improvement. Another benefit of genomic selection is that all genotyped animals can receive an evaluation for a trait once enough animals have records and have been genotyped. Consequently, more of the factors that determine an animal’s economic worth can be considered.

Genomics programs rely on processing massive volumes of diverse data. Traditional evaluations for yield traits are based on over 75 million animals with nearly 140 million lactation records. Over 1.5 million animals have been genotyped. Theoretically, the largest source of information is the bovine genome with nearly 40 million locations where genetic differences between animals can be detected. Currently 60,671 SNP markers are used to track the inheritance of chromosomal segments. Those SNPs were determined from 23 genotyping chips with from 2,900 to 777,962 SNPs. If SNPs that are used in genomic evaluation are not included in an animal’s genotype, they must be imputed (inferred from
relatives and animals with similar haplotypes, which are groups of genes inherited together from one parent). Imputation is a major consumer of computer resources.

Genotypes are checked extensively to ensure that the DNA came from the intended animal and that the pedigree information is correct. Each genotype is compared with all the others to determine if an animal’s parents are correct and to ascertain if other parent-offspring relationships were not included in the pedigree (often the correct parent). Efficient computational methods are required to make such comparisons.

Because genomic evaluations determine an animal’s economic value, owners want to receive this information as soon as possible, which has created a need for high processing velocity. As a consequence, most animals are genotyped before they are a month old, and evaluations for new animals are released weekly. Evaluations for all animals except bulls that are being marketed are updated monthly; evaluations for marketed bulls are released 3 times each year.

These genomic tools based on big data enable the selection of dairy cattle (and other species) to meet the current and future needs of the human population for healthy animals with efficient production and minimal adverse impact on the environment.

George Wiggans was a Research Geneticist for 38 years with the Animal Genomics and Improvement Laboratory (formerly the Animal Improvement Programs Laboratory), Agricultural Research Service, USDA, in Beltsville, MD, and retired in June 2016. He received a B.S. degree in dairy science and M.S. and Ph.D. degrees in animal breeding from Cornell University. He has made numerous contributions to improving the accuracy of genetic evaluation procedures for economically important traits of dairy animals (cattle and goats). He is recognized internationally as an authority on projection and standardization of yield records, genomic evaluation of dairy cattle, traditional genetic evaluation of dairy goats, computer analysis and application of an animal model to large data sets, and development of genomic evaluations based on DNA markers. Dr. Wiggans currently serves as a technical advisor for the Council on Dairy Cattle Breeding.

A Farmer Cooperative Perspective of “Big Data”
James R. Zimmerman, Dairy Farmers of America, Inc. and Dairy One Cooperative, Inc.

Dairy Farmers of America is a farmer-owned cooperative with 8,500 member farms and 14,000 members across the United States. DFA operates 41 processing facilities that manufacture all types of dairy products for the benefit of its members. DFA is a milk marketing cooperative and dairy foods processor dedicated to delivering value to members through secure markets, competitive pricing and strong participation throughout the entire dairy chain.

Dairy One Cooperative, Inc. is a farmer owned cooperative with a mission to create and provide management information to the agricultural community. Dairy One is affiliated with DFA and part of the DFA members’ services group. Dairy One measures many physical activities on farms so that farms can in turn manage their day-to-day activities and develop long term strategy. Dairy One is a practitioner of integrating data for farmers.

Data creation and use by dairy farmers continues to increase in importance for improved profitability and sustainability. DFA and Dairy One come from a viewpoint that producers own the data they create and that we are the stewards of that data.

The DHIA/milk recording system in the U.S. is one of the original “big data” systems in the world. It has led to an ever increasing gains in animal efficiency including milk production and reductions in greenhouse gas emissions. The DHIA system is built on a foundation of solid data collection that begins with unique cattle ID and certified data collection procedures.

An example of a system that produces large amounts of data for dairy farmers is the milk payment system that is used to pay producers, bill milk customers, and provide farm management data. The system is governed by the FDA and USDA to insure accuracy throughout the process.

DFA and Dairy One have data controls in place to ensure that producer data is only shared with whom the producer designates. This is done through the use of RAC codes for DHIA information and password protection for producer payment information.

There are a variety of issues facing farm data and technology today. One of those is characterized by a significant “divide” in how data is used by farmers. The dairy industry is made up a very diverse population of producers, some of whom use very little data for managing their farms and those that are very data intense. Meeting the needs and desires of the diverse population is a challenge for all information providers.

As new sources of data are developed for the dairy industry they become more complex to manage at the farm. Farmers desire to have data in a single system (or few systems) compiled for them and mined for them so they can manage their business activities and not spend a lot of time on data interpretation. Video is a newer source of data on
farms that is bringing new benefits for management and security, but is difficult to manage the large amount of data. Farmers are seeking assistance in the monitoring and analysis of video footage. Also as new sources and forms of data are created, early adopters risk market obsolescence of those systems. As we look to the future producers are looking for data systems that will act as a “data hub” for incorporation of many sources of data. To date that role has been filled by the major cattle management software platforms. Outside of the barn new data sources related to crop production and nutrient management are exploding. This area is likely more dynamic at the current time than in the barn. This is being pushed through the development of crop yield monitors, resource use models, and the ability to discern variation in soils.

To date the dairy industry has focused much on cattle management and less on cropping and nutrient management. It is my belief that there needs to be more of a continuum of data connecting the soil, water, and animal management to continue to gain efficiencies and sustainability of the long term.

Jamie Zimmerman has spent his lifetime in the dairy industry. He grew up on dairy farm in the Finger Lakes region of New York State and following graduation from Cornell University with a degree in Agricultural Economics he spent seven years managing the family farm. He joined Dairylea Cooperative in 1989 first working as a farm-level consultant and then serving in a variety of positions including field supervision and farm services development. During this time, he earned an MBA from Syracuse University. He is responsible for developing one of the first price risk management programs in the U.S. for the benefit of Dairylea producers. In 1997 he assumed the position General Manager of Dairy One Cooperative, Inc. and continues today as CEO & General Manager of the organization. Additionally, Jamie serves as Vice President, Farm Practices with Dairy Farmers of America. Jamie has served in leadership roles with a number of industry organizations including National DHIA. He currently serves on the Environmental Issues Committee of the National Milk Producers Federation.

Using Data to Drive Monitoring, Intervention, Oversight and Decision Making in Feedlots and Calf Grower Operations
Calvin W. Booker, DVM, MVetSc, Feedlot Health Management Services Ltd.

Large-scale data collection and routine analysis using epidemiologic tools is important when providing animal health and production consulting services to commercial feedlots and calf grower operations. In herd health and production medicine, epidemiologic analysis of health and production data is parallel to thorough clinical examination in individual animal medicine. However, rather than using epidemiological tools in an ad hoc manner, such as in response to a disease outbreak, systematic use of data is integral to making informed data-based decisions in a sustainable business or scientific model. Accurate information to characterise animal health and production is crucial to practicing credible food production medicine.

To successfully apply epidemiological principles, information requirements must be defined and accurate. Efficient data collection methods must be developed with outcome measures, desired outputs, and analysis methods defined prior to collecting data. This requires a comprehensive understanding of the production system of interest. Subsequently, there is a wide range of possibilities in how collected data may be used, depending on the application of interest or opportunity available. The most common uses in production medicine include monitoring and surveillance, forecasting, research trials, economic and sensitivity modelling, retrospective analysis and disease investigation.

Monitoring and surveillance is key to assessing animal health and production outcomes over time and requires a defined monitoring system, defined critical levels at which management action will be taken, and predefined management/intervention actions. Examples include monitoring respiratory disease morbidity in feedlot pens by days on feed or the level of total serum protein at arrival in dairy calf grower operations. Collecting simple descriptive information is straightforward; however, in order to properly interpret this information, it is necessary to appropriately establish the population of interest. The data collected for monitoring animal health and production measures also provides the basis for predicting animal health and production outcomes and costs. For example, predicted morbidity rates in some calves can be used to forecast resources required to meet anticipated animal health and production needs.

Epidemiologic analysis of large-scale data sets can be used to generate animal health and production research questions. Field trials in commercial production settings are extremely valuable for assessing animal health and production strategies provided that they are based on sound scientific study design and conducted using standardized and unbiased procedures. Economic modelling uses data collected from research trials or monitoring systems to determine whether changes in protocol or interventions are economically sustainable based on current market prices. Sensitivity modelling can be used to investigate the dynamics of market variability on the economic sustainability of animal health
and production outcomes. Data collected for routine monitoring and surveillance can also be used retrospectively to track short and long-term trends.

In summary, practical application of large-scale data collection and epidemiologic analysis is crucial to sustainable food animal health and production as it forms the basis for establishing long-term productive relationships with commercial feedlot and dairy calf grower operators.

**Dr. Calvin W. Booker** received his Doctor of Veterinary Medicine (D.V.M.) in 1989 and his Master of Veterinary Science (M.Vet.Sc.) in 1992 from the Western College of Veterinary Medicine, Saskatoon, Saskatchewan. Dr. Booker completed an internship at the Ambulatory Clinic, New York State College of Veterinary Medicine, Cornell University, Ithaca, New York in 1990 and a residency at the Field Service Clinic, Western College of Veterinary Medicine, Saskatoon, Saskatchewan in 1992. From 1990 to 1992, Dr. Booker collaborated on research efforts with Feedlot Health Management Services (Feedlot Health), which is a professional services company that provides comprehensive veterinary and production consulting services, as well as computerized individual animal data collection systems and management/execution tools, to feedlot and calf grower operations throughout Canada, the United States, and Mexico. In addition, Feedlot Health has extensive research capabilities that focus on identification, development, and evaluation of applied technologies in commercial feedlot and calf grower production systems. Dr. Booker became part of the Feedlot Health team in 1992 as an epidemiologist, data analyst and production consultant. Dr. Booker is currently a Managing Partner in the firm, with responsibilities that include managing the Feedlot Health production consulting business, directing research and development activities, and building the company growth and development plan. Dr. Booker was the 2006 recipient of the American Association of Bovine Practitioners’ Merial Preventive Medicine Award – Beef and the 2010 recipient of the Academy of Veterinary Consultant’s Consultant of the Year Award. Dr. Booker is also an Affiliate Faculty Member in the Department of Clinical Sciences at Colorado State University.

The Role of DHI in Big Data
Scott Tripp, DHIA

1. There are multiple companies within the “DHI” organizations.
2. Each organization has its own path and opportunity to change.
3. The Council of Cattle breeding has benefited from information that has funnel to them from the DHIA’s.
4. IF fewer herds test, how does good data flow to the council?
5. How does information flow back to the FDA on health and drug use on cattle?

**Dr. Scott Tripp** is a Dairy management consultant who works for DHI Provo where he collaborates on new software to manage dairies. He is also the Clinician for Daisy Farms a research Dairy facility in Paris Tx. Dr. Tripp has been awarded the Excellency in Preventive Medicine Award by the American Association of Bovine Practitioners and is the recipient of the Master Sales Award by the Monsanto Dairy Business. Dr. Tripp is also a regional and national judge for the “Dairy Challenge” (a senior college student evaluation of dairies) as well as a national judge on dairy reproduction for the Dairy Cattle Reproduction Council.

Biological Modelling: The key to unlocking the riches in precision livestock data
Friggens, N. C., Martin, O., Thorup, V. M., and Puillet, L.

This talk develops arguments for the benefits of placing precision livestock data in a systemic biology context, and the benefits of this for predictive modelling. There is a large untapped potential to combine precision livestock data to provide biologically meaningful phenotypes. First generation precision livestock solutions are generally developed for the purposes of monitoring one particular condition, and very often using only one measure. Whilst there may be sound commercial or applicability reasons for doing this, there is almost always not full mapping of the condition of interest onto the measure being used to monitor it. One major challenge for the development of a second-generation PLF is to move from single-event monitoring to combined measures that allow improved prediction of animal states, i.e. precision phenotyping. Recent work has shown that combinations of precision livestock measures can provide a fuller, and more robust, description of complex traits. A simple example of this is that energy balance measures can be derived directly from body weight and condition score data without the need to measure intake, provided that the high-frequency measures
are available. Other examples are methods to move from relatively crude binary classification of disease states to degree of infection measures.

Further, long-term trajectories of performance measures can be used to characterize individual variability in the underlying biological functions. Using a simple systemic model we have shown how body weight trajectories can be decomposed into components such as growth and maternal investment. This approach may have practical utility in identifying animals that have the appropriate partition of resources to different life functions (e.g. production vs body reserves) according to the local production environment. It also provides necessary information for biological models that can be used to predict the consequences on herd-level efficiency and resilience of future genetic selection and management strategies. A recently published model (Puillet et al., 2016. Genet. Sel. Evol 48:72) has clearly demonstrated that the key traits needed to increase short-term milk production efficiency and lifetime efficiency are not the same. This model describes animals in terms of their ability to acquire resources (i.e. variation in intake capacity) and their allocation of acquired resources between different life functions. It demonstrates the utility of biological models to explore genotype environment interactions, and to identify the selection and management strategies that will ultimately allow the optimum balance to be found between production efficiency and robustness.

Nicolas Friggens is head of a research unit at INRA working on Systemic Modelling Applied to Ruminants (MoSAR), and Visiting Professor at SRUC. Following a PhD from the University of Edinburgh he worked at the Scottish Agricultural College and then at Aarhus University, Denmark. He has a background in nutrition and modelling, with a focus on the interplay between body reserves and other life functions. He has led projects evaluating the effects of dairy cow genotype and nutrition on health and reproductive performance, and industry collaborations developing in-line monitoring technology and decision support software to improve dairy cow health and reproduction management. The aims of MoSAR, his current unit, are to develop: improved modelling tools to predict animal performance according to nutritional environment and physiological state, and methodology to characterise differences between animals in their adaptive capacity and robustness.

Applied Analyses and the Importance of Data Quality

Michael W. Overton, Elanco Knowledge Solutions

There is a booming interest from consultants, researchers and industry to examine or investigate the relationships between disease, growth, management, and other factors on the lactation, reproduction and culling performance in dairy cattle and heifers via observational analyses. Using cow-level data retrieved from on-farm record systems rather than controlled, prospective approaches allows the discovery of new knowledge with significantly less expense. The resulting information can be extremely valuable provided the data are collected, analyzed and interpreted correctly in the context of the confounding and biases that are commonly encountered. These large dairy data sets create both a great opportunity and a great responsibility.

Evaluating these observational data includes a number of important steps. First, there is the validation of the accuracy and completeness of the data that are being examined. In any statistical evaluation of data, an awareness of the inherent biases and confounding likely present is important but especially critical when the data are observational data recorded by commercial dairies for management purposes rather than for research. Thoughtful consideration of the most appropriate statistical approaches considering the data types and the question posed, followed by appropriate selection of the independent variables for model building is required. An especially important issue for the evaluation of observational data is that of correlation vs. causation. Often, the temporal relationships between independent variables are poorly understood or unknown. Finally, data analyses should follow a logical, planned approach using subject matter expertise rather than blindly allowing complicated statistical programs to “examine” the data and develop the best fitting model using stepwise processes.

Statistical approaches used for observational data vary considerably based upon the nature of the dependent variable, the nature of the independent variable(s), and the question being posed of the data. Summarization of data may utilize frequency histograms, bar charts, pie charts, etc. while statistical approaches to examining the relationship between dependent and independent variables include a wide variety of approaches such as scatterplots, correlation coefficient, t-test, basic regression, ANOVA, linear regression, logistic regression, and Cox Proportional Hazards models. Regardless of the approach taken, the pursuit of significant p-values can lead to the development and use of inappropriate or inaccurate statistical models. Instead of simply pursuing and reporting the final p-values for only those variables with values less than or equal to 0.05, effect sizes and confidence intervals should be used to convey both the magnitude and relative importance of associations observed within carefully constructed statistical models.
Dr. Michael Overton received his B.S. and D.V.M. from North Carolina State University and practiced veterinary medicine for 8 years in North Carolina. After a move to California to complete a Dairy Production Medicine Residency and his Masters of Preventive Veterinary Medicine degree, he worked as a Dairy Production Medicine Specialist at UC Davis-VMTRC in Tulare, CA for 6 years. Then, he joined the University of Georgia – College of Veterinary Medicine where he served as Professor of Dairy Production Medicine and chief of service for the food animal program for about 7 years. In May 2012, Dr. Overton left his tenured position at the University of Georgia to assume a Dairy Analytics position with Elanco Knowledge Solutions. In this role, Dr. Overton is responsible for developing economic models and tools for both internal and external customers, for providing consultative services to dairies and their consultants, and for building analytical capabilities for the global Elanco team. Throughout his professional career, Dr. Overton worked extensively in the areas of reproductive management, transition management, analysis of on-farm records, and economic decision making. He has been active in service to the dairy industry and travels frequently to speak and consult in the U.S. and internationally. He has authored or co-authored over 100 peer-reviewed, proceedings or industry publications on various topics regarding dairy production medicine. Dr. Overton lives in Athens, Georgia with his wife Carol. Their son, Justin, works for ANC, a turnkey technology service provider in Atlanta. Their daughter, Audrey, is a recent graduate from at Mercer’s Georgia Baptist College of Nursing and is married and living in Blairsville, Ga.

Data Analysis Techniques: Time Series and Statistical Process Control
Albert De Vries, University of Florida

If you believe the internet, then time series is the new big data. A univariate time series is a sequence of measurements of the same variable generated over time. Often the measurements are made at regular time intervals. Time series data are not necessarily independent, and order matters. Big data is used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors, and/or computer-based transactions (Megahed and Jones-Farmer, 2013). Volume, variety, and velocity (3V’s) are the three main characteristics that distinguish big data from the data that was generated in the recent past. While we may not have truly big data in dairy production yet, examples of newer data that appear big are generated by sensors, genomics, and milk analysis. Sometimes these data appear in time series and one of the main interests of the dairy farmer is the identification of unplanned changes in production processes with time series output. Examples are alerting cows or groups that perform other than expected. Unplanned changes need to be detected as soon as possible so true problems can be corrected, but not every deviation from expectation warrants investigation because of costs. Statistical process control (SPC) is the use of statistically based methods to evaluate a process or its output to achieve or maintain a state of control. SPC charts are the traditional means to monitor time series data to detect unplanned changes, so a review of their key principles and performance measures is useful to understand their potential role in monitoring big data. SPC charts essentially monitor deviations from expectation. An important improvement for application of SPC to dairy production (big) data is that the expected performance is likely not a simple average. Expected variation in performance can be, for example, diurnal or seasonal, or known effects of breed, stage of lactation, parity, and treatment. More traditional techniques to model expectation include regression, dynamic linear models, and (double) exponential smoothing. Deviations from expectation are monitored with Shewhart-type, cusum, or EWMA control charts. Multivariate versions of these charts allow for monitoring multiple performance variables together, such as changes in milk production and dry matter intake. Naïve Bayesian networks are one technique to combine deviations from multiple sources of time series data, combined with other sources of data. Collectively, the techniques to monitor time series data for unplanned changes have become more sophisticated over time. Application of traditional SPC charts to monitor big data has been a challenge, but the core principles still apply. New techniques developed in engineering and statistics now combine data mining methods such as machine learning and neural nets with control charts. Several examples of these newer techniques are becoming applied in animal production, and examples from engineering with potential applications in dairy production will be discussed.

Albert De Vries is currently an associate professor in the Department of Animal Sciences at the University of Florida. He grew up on a dairy and swine farm in the Netherlands. He went to Wageningen University where he received a BS and MS in Animal Science with a minor in agricultural economics. In 2001 he received a Ph.D. in Animal Sciences at the University of Minnesota with a focus on dairy science, applied economics, operations research, and statistics. His Ph.D. dissertation was on the application of statistical process control techniques to monitor dairy herd production. At UF
since 2001, his interests are in optimization of culling and replacement strategies, statistical process control, economics of reproduction and genetics, data mining, and financial analysis of dairy farming. Albert De Vries is engaged in research, teaching and Extension.

Machine Learning Methods for Finding Useful Information in Big, Messy Data Sets
Kent Weigel, Department of Dairy Science, University of Wisconsin – Madison

Researchers and practitioners in dairy management and related fields are increasingly confronted with big, messy data sets and asked to make meaningful inferences or practical recommendations. However, the task of turning data into information is not trivial, and numerous pitfalls exist. The definition of “big data” is subjective, but essentially it means that the data set is too large to permit visual inspection of patterns or anomalies. Likewise, the definition of “messy data” is subjective, but essentially it means that the data set contains some combination of missing observations, empty cells, confounding, biologically implausible values, or calibration failures. Simple statistical approaches, such as linear regression or mixed linear models, may be unable to capture the complexities associated with the aforementioned characteristics, particularly if the number of potential explanatory variables is large.

Machine learning is a branch of artificial intelligence that focuses on making data-driven predictions in big, messy data sets using highly flexible models or algorithms. Countless options exist, as regards specific algorithms and parameter choices, and the optimal approach may differ widely from one application to another. Broadly speaking, methods can be divided into the categories of supervised or unsupervised learners, noting that a hybrid (semi-supervised learning) category is emerging. Supervised learning relies on creation of a training set, in which explanatory variables (inputs) and dependent variables (outputs) are known – these are referred to as labeled data. This training set is used to select an appropriate algorithm and tune its parameters, with the goal of minimizing some error or entropy function. Once the algorithm and its parameters are chosen, its performance is evaluated in an independent testing set, which must be totally blind to the model training and tuning processes. In contrast, unsupervised learning relies on an unlabeled training set in which the outputs or dependent variables are unknown or nonexistent. The objective of unsupervised learning is to discover hidden patterns or structures in the data, for the purpose of clustering observations into groups, detecting anomalies, or reducing dimensionality of the data set using some type of distance function.

While machine learning methods offer great potential, due to their flexibility and ability to accommodate missing values, unbalanced data, and highly nonlinear relationships among variables, these algorithms can lead to erroneous inferences or conclusions if used improperly. Design of the cross-validation scheme is critical, and this usually relies on randomly dividing the data set into a given number of folds or subsets, such that a portion of the data are used for training and tuning the model and its parameters, while the remainder of the data are used for validating model performance. Complete independence between the training and testing sets is critical; otherwise predictive ability of the algorithm will be exaggerated. Overfitting the training data is a common pitfall, both in regression (continuous data) and classification (binary or categorical data) applications. A highly complex model may fit the training data perfectly, but this comes at the expense of performance in the testing data and generalization to future data sets. Nonetheless, machine learning methods have gained solid footing in fields such as dairy management and animal genetics in recent years, and with growing expertise and experience these approaches will be a valuable addition to our analytical toolboxes.

Dr. Kent Weigel is Professor and Chair of the Department of Dairy Science at the University of Wisconsin-Madison. He also serves as Extension Dairy Genetics Specialist and is a key technical consultant for the National Association of Animal Breeders and many other partners in the dairy genetics industry. His research focuses on methods and strategies for genomic selection of dairy cattle, as well as genetic improvement of production efficiency, health, and fertility using genomic testing, advanced reproductive technologies, and novel breeding strategies. Dr. Weigel has published more than 200 peer-reviewed journal articles on various aspects of genetic and genomic improvement of dairy cattle and has given lectures to academic, industry, and producer audiences in more than thirty countries.

Value of Information
Dr. Tyler Mark, University of Kentucky

As an economist working in precision agriculture, I see new technologies that come to market weekly. One of the key questions that I often get is “Does the information that this technology provides add value to my operation?” My answer is often that it depends! However, this is a key question that producers, processors, and equipment manufacturers
must answer. We have had a number of different revolutions take place in the agricultural sector over the last 100 years. They include the industrial revolution with the introduction of tractors, the introduction of GMO seed in the 1990’s, and now the next revolution may just be “Big Data”.

The amount of data that the agricultural sector is generating through precision technologies is staggering. However, raw data in itself has little to no value. For there to be value the data must be processed and utilized in management decisions. Only at this point will data have value. Within the precision dairy sector there are a wide range of questions that producers, researchers, consultants, and companies are trying to address. If you ask a producer, they are probably trying to increase milk yield or improve reproduction. Researchers are also looking at these questions, but are also trying to address questions that are more abstract, such as “Will the adoption of this technology increase the quality of life for the producer and by how much?”

Within agricultural economics and the decision sciences disciplines we have a wide range of methodologies that can be implement to assess the value for these different questions. Some of these methods involve investment analyses, mathematical programming models, simulation models and technical efficiency analysis. However, the problem is much more complicated than just assessing the value of each technology individually. Individual technologies may have some marginal value from its usage, but in precision agricultural we find that more value can be extracted with the optimal bundle of technologies for your farm. This is a very complex question to answer, because each dairy will be different.

However, before we can even assess the value of each technology there are some important barriers that need to be addressed. Some of these include the legal framework to protect the data, ability to store and transmit the massive amounts of data being produced, quality data, and equipment capable of collecting accurate data. If this can be accomplished than there is a wide range of individuals that could find value in the data. At the farm level I do expect to see efficiency gains and marginal increases in profitability. I expect more substantial gains will most likely be realized at the processor or consumer level.

In conclusion, the value that can be extracted from the data collected will be dependent on the industry’s ability to remove barriers and understand how the data can be utilized in a whole system approach. Right now the precision agricultural sector and the precision dairy sector operate independently. Going forward we are going to have to find ways to merge these two sectors to answer 21st century problems.

Dr. Tyler Mark is a Production Economist at the University of Kentucky. His applied research interests include broadband availability in rural areas, precision agriculture, precision dairy, dairy policy, renewable energy feedstocks, and hemp economics. Current projects include factors that impact the profitability of Kentucky farmers, broadband internet impact on precision agriculture data transmission, economics of hemp production in Kentucky, dairy policy in the Southeastern United States, and economics of precision dairy technologies. In 2014 he was awarded the C-FARE/AAEA Early Career Professional Leadership Award and 2016 the Golden Quill Award for Most Outstanding Journal Article in the Journal of ASMRA.

Dairy Data - Guide to Avoid Potential Pitfalls
Steve Eicker, Valley Ag Software

A number of recent technological advancements have coalesced to provide an incredible opportunity for the dairy industry: Computational power (hardware and software), Internet access, access to more data, and advances in sensors and battery life. Regardless of the new power, there are still some fundamental issues that exist:

- Correlation is NOT causality
- Data quality issues
- Confounding
- Timing and speed
- Data ownership, access
- Economics

Unfortunately, very few of these issues have been mitigated by the new technology. Instead, the opportunities might have attracted people less familiar with the potential issues listed above. We will use some examples to try to demonstrate some potential risks as people try to adopt these tools.
Steve Eicker is the co-founder (with Connor Jameson) of Valley Ag Software. He earned a BS in mathematics, an MtS in Computer Sciences and a DVM from Michigan State University. He formerly taught at University of Wisconsin at Madison, and then at Cornell University, where he currently serves as one of the directors of the Summer Dairy Institute. He works with DHI organizations, veterinarians and consultants, pharmaceutical and genetics companies, and progressive dairy farms world-wide. He frequently lectures on economics, and data-driven decision processes.

Data Visualization
John C. Hart, University of Illinois at Urbana-Champaign

Data visualization provides a high bandwidth connection between the data managed by a computer system and the human brain, facilitated by visual communication. The 1987 Visualization in Scientific Computing report by McCormick, DeFanti and Brown showed long ago that visualization is a necessary component of computer applications, quoting Richard Hamming that “the purpose of computing is insight, not numbers.” Such insight is even more necessary now in the era of big data, when more than 90% of our data has been generated in the past two years. Modern data science methods such as machine learning and data mining provide the computational tools for extracting useful information from warehouses of data. However, as Fred Brooks emphasized in his 1999 Turing Award lecture, the combination of computers with humans is greater than computers alone, and data mining is most effective when coupled with human analysis, reasoning and intuition, and data visualization facilitates that coupling.

As we apply data science to specific disciplines, including dairy science, one of the main challenges is workforce development. Data science and data visualization can be learned classically through books and classrooms, but modern online courses have greatly expanded the educational options for learners to understand this new field and develop these new skills. The Data Visualization MOOC available on Coursera provides the opportunity to learn the fundamentals of data visualization, including computer graphics, user interaction, human perception and cognition, as well as the skills to determine which visualization charts and graphs to use based on the data, and how to form data visualization systems to best leverage the abilities of the human in the equation. This data visualization course is part of a Data Mining Specialization, as well as a Master’s degree program in computer science focusing on data science offered by the University of Illinois at Urbana-Champaign on the Coursera platform. This new online degree program is an effort to serve educational mission of the University of Illinois to develop the skills needed for a new workforce for the state, nation and world to adapt to the critical need for more data scientists, a job title that didn’t even exist just a decade ago.

John C. Hart is a Professor in the Department of Computer Science at the University of Illinois, Urbana-Champaign where he studies computer graphics and data visualization. Prof. Hart teaches a Coursera Data Visualization course with over 100,000 learners from over 200 different countries as part of its Data Mining Specialization. He is also Executive Associate Dean of the Graduate College, and the designer and acting director of the new stackable Master of Computer Science in Data Sciences degree delivered online through the Coursera platform. Prof. Hart is a past Editor-in-Chief of the ACM Transactions on Graphics and is currently the graphics area editor for ACM Books, which recently published The VR Book by Jason Jerald.

Educating Tomorrow’s Leaders
Alan Fahey, University College Dublin

Big Data has been defined as “datasets whose size is beyond the ability of typical database tools to capture, store, manage, and analyse” (McKinsey Global Institute). The development of precision farm technologies and the use of genomic technologies in dairy breeding programmes now means that the global dairy industries have entered the “Big Data Era”. The utilization of Big Data offers the dairy industry the opportunity to utilize data driven decisions to improve efficiency and sustainability of dairy enterprise. One of the challenges that faces the dairy industry in the “Big Data Era” is the training and development of Data Scientists with a knowledge of the dairy industry. The objective of this paper is to better understand the skills that are required to be a Big Data Scientist and how Animal and Dairy Science university departments can contribute to educating the next generation of Big Data Scientist.

Studies have shown that the term Data Scientist is not a well-defined term and can mean different things to different companies. Analysis of over 2700 Big Data job advertisements showed that companies had diverse requirements for Big Data Scientists and that different job postings required different skillsets. However, the one common skillset that
was required for all Big Data jobs was the requirement to know how the data and the resulting analyses could impact on the business. Therefore people employed as Data Scientist should have a knowledge of the dairy industry.

Most graduate Data Analytics programs around the world are driven by Computer Science, Engineering and Statistics departments with little input from Agriculture programs. Dairy and Animal Science departments have the opportunity to meet education requirements for tomorrow’s Data Science leaders through collaboration with Computer Science, Mathematics and Statistics, and Engineering departments. Undergraduate and postgraduate student exposure to these skills will help them see the potential of Big Data application to their field of study.

Alan Fahey is an Associate Professor of Animal Breeding and Statistics in the School of Agriculture and Food Science at University College Dublin. He graduated with a degree in Agriculture Science (Animal Science major) from University College Dublin. He then moved to the Animal Sciences department at Purdue University and graduated with an MS in dairy cattle breeding and a PhD in poultry genetics. Prof Fahey currently teaches undergraduate and postgraduate courses in Animal Breeding and Statistics at University College Dublin. His current research interests are in the area of validating and developing selection indexes for alternative dairy production systems. He enjoys interacting with farmers and dairy industry professionals in the development of dairy breeding and management solutions.

Precision Dairy Monitoring Technologies
Jeffrey Bewley, Associate Extension Professor, Dairy Systems Management, University of Kentucky Department of Animal and Food Sciences

What is Precision Dairy Farming? Precision Dairy Farming is the use of technologies to measure physiological, behavioral, and production indicators on individual animals to improve management strategies and farm performance. Technologies included within Precision Dairy Farming range in complexity from daily milk yield recording to measurement of specific attributes (e.g. fat content or progesterone) within milk at each milking. Many Precision Dairy Farming technologies, including daily milk yield recording, milk component monitoring (for example, fat, protein, and SCC), pedometers, automatic temperature recording devices, milk conductivity indicators, automatic estrus detection monitors, and daily body weight measurements, are already being utilized by dairy producers. Other Precision Dairy Farming technologies are being developed to measure rumen pH and contractions, feeding behavior, body condition score, animal positioning and activity, locomotion, lying behavior, odor, progesterone, heart rate, greenhouse gas emissions, and respiration rate. With Precision Dairy Farming, the trend toward group management may be reversed with focus returning to individual cows through the use of technologies. The main objectives of Precision Dairy Farming are maximizing individual animal potential, early detection of disease, and minimizing the use of medication through preventive health measures. Precision Dairy Farming is inherently an interdisciplinary field incorporating concepts of informatics, biostatistics, ethology, economics, animal breeding, animal husbandry, animal nutrition, and process engineering.

Benefits of Precision Dairy Farming. Benefits of Precision Dairy Farming technologies include increased efficiency, reduced costs, improved product quality, minimized adverse environmental impacts, and improved animal health and well-being. These technologies are likely to have the greatest impact in the areas of health, reproduction, and quality control. Realized benefits from data summarization and exception reporting are anticipated to be higher for larger herds, where individual animal observation is more challenging and less likely to occur. As dairy operations continue to increase in size, Precision Dairy Farming technologies become more feasible because of increased reliance on less skilled labor and the ability to take advantage of economies of size related to technology adoption.

A Precision Dairy Farming technology allows dairy producers to make more timely and informed decisions, resulting in better productivity and profitability. Real time data can be used for monitoring animals and creating exception reports to identify meaningful deviations. In many cases, dairy management and control activities can be automated. Alternatively, output from the system may provide a recommendation for the manager to interpret. It’s important to remember that information obtained from Precision Dairy Farming technologies is only useful if it is interpreted and utilized effectively in decision making. Integrated, computerized information systems are essential for interpreting the mass quantities of data obtained from Precision Dairy Farming technologies. This information may be incorporated into decision support systems designed to facilitate decision making for issues that require compilation of multiple sources of data.

The Advantage of Precision Dairy Farming. Historically, dairy producers have used experience and judgment to identify outlying animals. While this skill is invaluable and can never be fully replaced with automated technologies, it is limited to human perception of a cow’s condition. Often, by the time an animal exhibits clinical signs of stress or illness, it is too late to intervene. These easily observable clinical symptoms are typically preceded by physiological responses
evasive to the human eye (e.g. changes in temperature or heart rate). Thus, by identifying changes in physiological parameters, a dairy manager may be able to intervene sooner. Technologies for physiological monitoring of dairy cows have great potential to supplement the observational activities of skilled herdspersons, which is especially critical as more cows are managed by fewer skilled workers. Consequently, Precision Dairy Farming may change the way cows are managed in the future.

Dr. Jeffrey Bewley is currently an Associate Extension Professor in Dairy Systems Management at the University of Kentucky. Jeffrey’s primary interests are the application of precision dairy farming technologies, economics of decisions on dairy farms, milk quality management, dairy cow comfort and well-being, records management and benchmarking, systems troubleshooting, and strategic dairy business management.

Opportunities for managing health and well-being using precision technologies
C. S. Petersson-Wolfe, T. H. Swartz, N. M. Steele and B. T. Dela Rue

Disease prevention and treatment is a constant focus in the management of a dairy herd. All medium and large sized dairy operations have reported at least one case of clinical mastitis, lameness, retained placenta, reproductive problems, or milk fever with an even higher percentage having at least one cow with a health problem (USDA, 2007). Of these diseases, those that are the most prevalent throughout the United State dairy industry are clinical mastitis occurring in 95%, lameness at 88% and reproductive problems at 84% of all size operations (USDA, 2007). The costs associated with these diseases can be anywhere from $200 for a case of clinical ketosis or mastitis, to more than $300 for an identified case of lameness (Kelton et al., 1998). The clinical state of these diseases is easily identifiable and therefore, it is easy to associate the cost to the overall dairy operation when they occur. However, the true cost of disease is unknown because of detection limitations in diagnosing disease in the subclinical stages. For this reason and the associated effect on animal well-being, recent research has been focused on assessing animal health with more sensitive indicators that are suitable for use before animals become clinically ill (von Keyserlingk et al., 2009).

Historically, dairy producers have focused much of their health management efforts on the treatment of disease. To date, physical observation, and urine or blood tests have been used as indicators of animal health but this can be costly and time consuming, which indicates a need for a more rapid and continuous measure of health (von Keyserlingk et al., 2009). However, more recently, dairy producers have adopted a more proactive health management strategy and in response, advanced technology tools for monitoring herd health have been developed (LeBlanc et al., 2006). Developments in on-farm monitoring technologies have allowed producers to enter the realm of “Precision Dairy Farming”. The data generalized from Precision Dairy Farming can allow for the early detection of disease and this capability can improve cow comfort and animal well-being (Dawkins, 2003, von Keyserlingk et al., 2009).

Precision Dairy Farming is defined as a collection of technological advances that can measure physiological, behavioral, and production indicators on individual animals (Bewley, 2010). These advancements provide management tools to identify problem areas related to reproduction, nutrition, dairy calf management and feeding, dairy cattle health, mastitis and milk quality. Technologies currently existing monitor a variety of outcomes including daily milk yield, milk components (e.g. fat, protein, and SCC), animal activity, rumen and milk temperatures, milk conductivity, milk replacer intake, estrus detection, and daily body weight measurements (Bewley, 2010). The opportunities for managing health and well-being using precision technologies is advancing at a rapid rate and is limited only by the speed at which researchers can evaluate and discern the masses of data that are collected from these advancements. The objective of this talk is to discuss the current state of knowledge regarding disease identification as it relates to the dairy industry both from the standpoint of lactation and non-lactating animals, as well as young stock.

Dr. Christina Petersson-Wolfe is an Associate Professor of Dairy Science at Virginia Tech. She completed her B.S. (Dairy & Animal Science) at Penn State University, M.Sc. (Epidemiology) at the University of Guelph and Ph.D. (Animal Science) at Ohio State University in 2006. Her research interests are focused around mastitis prevention, disease detection and animal well-being. Currently, she has a heavy Extension appointment where she works directly with stakeholders in the field, while also maintaining an active research program.

Technology and Data Collection
Brian Houin, Homestead Dairy, Plymouth, IN
Here at Homestead Dairy, we collect a lot of data to make management decisions. With the help of various technologies, we can collect and analyze that data. We can now manage a large number of cows better than we could manage a small number of cows 20 years ago.

Starting at birth, we use a brix refractometer to be sure that the colostrum each calf receives is of the best quality. We then pull blood samples at day 3-4 to make sure each calf gets great passive transfer. Each morning we weigh each calf born, and take a DNA sample and send it to the lab. By the time the calf is a month old we know her complete future production. By the time the heifer is 6 months of age we have 4 weights to track growth at each stage. We then weigh the heifers prior to breeding to keep track of each grower. At the heifer breeding facility we have the Cow Manager heat detection system. It ensures that heat detection is not an issue and the proper time of insemination is followed. It also has rumination to allow us to catch animals that may be sick.

When each animal calves, Time, man on duty, and assisted or normal calving, are recorded to ensure we have correct information to start treatment protocols as soon as possible if necessary. We then have the AFI milk program with the AFI lab, on all fresh cows, so we can find any health issues as quickly as possible. With activity monitoring, milk monitoring, lying time, fat, and protein of the milk, we have all the information we need to keep cows healthy and reaching their full potential. We also have Dairy Comp with pocket cowcard, to keep track of the complete history of each animal, and can manage large groups of cows very quickly. Having the pocket cowcard enables our compliance on protocols to be extremely high. All information is wirelessly uploaded into the dairy comp program. We also have Feedwatch to monitor recipes and intakes for all groups on our farm.

We are currently building a large robotic dairy. Initially installing 24 Lely A4 Robots. We will be moving to a stress free environment for the cows having the ultimate cow comfort. This adds another software program to run and gather information. With this dairy hopefully cows will live a longer and happier life!

Brian Houin was born and raised on a dairy farm in north central Indiana. He attended Purdue University. Brian returned to the family farm - Homestead Dairy in Plymouth, Indiana – joining the family business in 2003, where he heads the heifer operation and manages the milkers. Homestead milks 3300 cows in 3 locations and has a grower unit about 30 miles away from the main farm. On the 3,000 plus acres they produce all corn silage and shelled corn. Today ten computers including three hand held systems allow them to monitor performance of every animal in real time.

Sensor Data and Use: Reproduction
Stephen LeBlanc, University of Guelph, Canada

Effective and efficient insemination, and timely diagnosis of pregnancy are critical to management of dairy herds. However, detection of estrus by observation of cows’ behavior is challenging and less than optimal in many herds. Detection of the substantial transient increase in activity that is associated with estrus can be used to largely replace observation of cow behavior. Pedometers or accelerometers for estrus detection were among the first and now most widely adopted sensor technologies in the dairy industry. These tools have been validated in several modest-scale studies to detect 60 to 90% of estrus events. Several large-scale field studies in a few herds are consistent in finding that overall herd reproductive performance is comparable between management based on activity monitoring or synchronization programs for timed insemination. However, their relative performance varied among herds, but the reasons for these differences have not been well quantified. Activity-based estrus detection is a viable tool for management of reproduction in commercial dairy herds, although it needs to be complemented with interventions for timed insemination for approximately 20% of inseminations.

Additional commercially-available technologies for reproductive management include systems for automated measurement of progesterone in milk, for diagnosis of pregnancy based on pregnancy-associated glycoproteins (PAG) in milk or blood, for detection of calving based on body temperature or activity, and for measurement of rumination as an aid in estrus detection or calving. There is some published performance data on each of these. However, most sensor-based tools for management of reproduction would benefit from additional independent, published validation data under field conditions. Moreover, the full promise of these technologies will only be realized if they can be integrated to improve the accuracy and timeliness of alerts, and if cow- and herd-level economic decision support are added to the cow-level task-based information that they presently provide.

Stephen LeBlanc is a Professor in the department of Population Medicine at the Ontario Veterinary College, and Research Program Director – Animal Production Systems at the University of Guelph. He received a BSc(Agr) in Animal Science from McGill University in 1992, and a DVM (in 1997) and DVSc (in 2001) from the University of Guelph. After
five years of private dairy veterinary practice, he joined the faculty at the University of Guelph where he teaches veterinary and agriculture students and provides clinical farm service. His research focuses on transition dairy cow metabolic and reproductive health and management, including field validation of precision technologies. With graduate students and collaborators, this work has resulted in over 90 peer-reviewed papers.