Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems

Article *in* Journal of Agricultural & Food Information · July 2019

DOI: 10.1080/10496505.2019.1638264

CITATIONS
42

READS
5,437

3 authors:

Nicoleta Tantalaki
University of Macedonia
6 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Stavros Souravlas
University of Macedonia
48 PUBLICATIONS **257** CITATIONS
SEE PROFILE

Manos Roumeliotis
University of Macedonia
135 PUBLICATIONS **715** CITATIONS
SEE PROFILE

Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems

Tantalaki Nicoleta[1], Souravlas Stavros, Roumeliotis Manos

156 Egnatias Str., 54636, Thessaloniki, Greece

Abstract

Developments in crop growth modelling and the progress in the use of tools to monitor and collect data from farms paved the way to reveal a new world of insight to support decision-making in precision agriculture. In this paper, we provide a review on the research dedicated to applications of data science techniques and especially machine learning techniques in relevant agricultural systems. Big data technologies and high-performance computing create new opportunities for data intensive decision-making, enabling producers to improve productivity and show potential to support several agricultural domains. A review on works in agriculture that employ the practice of big data analysis to solve various problems is performed, opportunities and promising areas of use are revealed. However, big data analysis has not yet been widely adopted in agriculture. The high volume and complexity of the data produced nowadays pose challenges in successfully implementing precision agriculture. Machine leaning seems promising to cope with big data but it needs to reinvent itself to meet existing challenges. We present research trends and barriers that need to be overcome to realize the potential of big data analysis to revolutionize agriculture.

Keywords: big data, machine learning, real-time analytics, precision agriculture, smart farming

*Corresponding author
Email address: nicoleta@uom.gr (Tantalaki Nicoleta)

1. Introduction

By 2050, the world's population is expected to be 34% larger than today. According to the Food and Agriculture Organization of the United Nations (2010), to keep up with rising population, global food production must increase by 70% in order to feed the world. This poses the challenge of improving agricultural productivity, while lowering its environmental footprint.

Advancements in crop growth modelling, progress in the use of tools to monitor and collect information from farms in a less labor-intensive manner, and global navigation systems give rise to precision agriculture, in which precise measurements at local points and data-intensive approaches support decision-making (Kamilaris et al., 2017). Over the past decade, machine learning techniques have been deployed across precision agriculture to provide more accurate solutions, mainly because of the capability to handle highly complex and non-linear agricultural problems (Liakos et al., 2018; Morota et al., 2018). Machine learning techniques have different potential, are of different complexity and computational requirements, and continually evolve. As model complexity increases, more data must be collected.

The vast amounts of data produced in the context of Internet of Things (IoT) can support strategic decision-making by increasing models' accuracy and generalization abilities. Big data is prominent in a variety of fields; it is increasingly applied to several agricultural domains including precision agriculture. While agronomic models will play a role in the interpretation of data, big data transforms agriculture from model- to data-driven. Learning from these massive data collections is likely to identify significant opportunities (Coble et al., 2018; Rao, 2018). Machine learning and data mining techniques are expected to be instrumental in meeting the challenges facing global agriculture, by taking advantage of big data.

However, the collection and analysis of large, complex, heterogeneous data, coming from the variety of sources encountered in agriculture, cannot be accomplished with traditional machine learning methods such as linear regression.

Although learning from big data seems promising, it faces severe challenges.

To address such knowledge gaps, we initially tried to verify the usefulness and applicability of different machine learning techniques in precision agriculture, to explore where the field has been, where it is now, and where it is likely to go. However, the large potential for big data in several agricultural domains other than precision agriculture, and the fact that big data analysis is a developing area, led to our examination of the opportunities and the challenges from its use. Ultimately, our objectives with this work are:

1. to provide a comprehensive look at the evolution of different prominent analytical techniques used in precision agriculture with regard to common themes, possibilities, computational aspects and current limitations;
2. to examine how the most promising field, machine learning, is applied to support precision agriculture;
3. to enhance the awareness for the potential implications of big data analytics in agriculture, presenting existing opportunities and promising areas of applications, and
4. to shed light on the factors that delay big data adoption in agriculture, providing future directions, open issues, and research trends to speed up adoption.

The rest of this paper is organized as follows: In the next section, we examine the precision agriculture paradigm and the evolution of a number of data analysis methods that accompany precision agriculture to support decision making. Special attention is paid to machine learning techniques, along with their possibilities and limitations. In section 3 the application of big data in agriculture is presented and the most promising areas are revealed, as we approach the big data era. Opportunities and challenges that arise along with open trends are further discussed. Section 4 concludes our work.

2. Precision Agriculture

In traditional agriculture, crops have been treated under the assumptions of uniform soil, nutrient, moisture, weed, and insect conditions. This has several times led to over-applications or under-applications of pesticides, irrigation, fertilizers, and other treatments (Wang & Li, 2013). The advent of Global Positioning Systems (GPS) and Global Navigation Satellite Systems (GNSS) enabled the practice of precision agriculture, "a management strategy that uses information technologies to bring data from multiple sources to bear on decisions associated with crop production" (National Research Council, 1997). The factual base of precision agriculture is the spatial and temporal variability of soil and crop factors between and within fields. The goal of collecting geo-referenced data is to generate more accurate descriptions of system aspects to inform decisions.

New challenges to the successful implementation of precision agriculture stems from technology advances and the huge increase of data both in number of records and variables. The augmented possibilities for data storage, high-throughput, and fully automated technologies have been rapidly generating large-scale data in agricultural settings. Despite the fact that scaling up to big data adds another layer of complexity, this challenge to efficiently extract insight from big data can be tackled by using techniques from machine learning and data mining (Morota et al., 2018).

In the remainder of this section, we initially analyze the methods used to obtain insight in the field of precision agriculture. Then, we will discuss the possibilities and challenges that arise from the cooperation of these methods with IoT technologies and big data to enhance agriculture.

2.1. Producing insight

The evolution of agricultural system models in precision agriculture is ongoing, making attempts to map inter- and intra-field variability, identify underperforming areas, and develop effective decision support systems. Parametric methods proved to be successful in extracting variables designed for local conditions, but

they have limited applicability in a broader operational setting. matter how much data a parametric model is given, the parameters needed remain the same. In the case of agriculture, they are characterized as sensor-specific models (Verrelst et al., 2015), because an explicit choice of sensor bands has to be made. However, in non-parametric approaches, the number of parameters is flexible, no assumption on the shape of the density is made, and the functions form is learned from the training data. The model learns from the data and can combine different data types. No prior choice of specific bands has to be made and all bands can be used to develop a model. Several machine learning approaches belong to this category.

The adoption of remote sensing in agriculture has led to the systematic collection of data volumes over large geographical areas. These data are not homogeneous and come from a variety of sources with different spatial, temporal, and spectral modalities. Machine learning techniques (based on non- and semi-parametric structures) are commonly used in the literature as they are flexible and capable of processing a large number of inputs and handling non-linear problems. The available data play a major role in the development of these models. The more complicated the problem to be solved, the more data are required.

There is extended literature comparing parametric and non-parametric approaches in several agricultural applications (Ali et al., 2015; Verrelst et al., 2015). Comparing their predictive accuracy, non-parametric methods are more suitable for yield predictions in agricultural planning (Gonzalez-Sanchez et al., 2014) and in general, the weaknesses of parametric regression seem to outweigh its strengths (Verrelst et al., 2015). Greater attention is paid to machine learning data-driven approaches that have shown their versatility in different contexts by fusing data from several sources. These approaches are discussed below.

2.1.1. Systems' evolution

Most crop models from pre-precision agriculture literature and during its dawn were typically based on linear regression analysis, calculations of root mean

square error, and mean error. Multiple linear regression techniques using interaction terms have been considered with improvement over strictly linear models (Drummond et al., 1995; Khakural et al., 1999; Kravchenko & Bullock, 2000). Multiple linear regression and linear mixed models have been used in soil mapping, where the variability of a target soil property is explained by its relationships with other soil and climate factors, with shortcomings like autocorrelation and non-linearity between variables (Meersmans et al., 2008). In agricultural practices, a variety of interrelated factors influence crop production. In complex situations where data are not linearly related and several outliers exist, linear regression models were not found useful in understanding yield response and did not provide accurate predictions even within subfield regions thought to be homogeneous (Drummond et al., 2003; Lambert et al., 2004; Sadler et al., 2007).

The high complexity and non-linearity of problems faced in agriculture required methods able to approximate complex mappings by integrating data coming from different sources and exploiting the information contained in the obtained reference samples. These methodologies are represented by machine learning techniques (Ali et al., 2015; Coble et al., 2018). Artificial neural networks, support vector machines, decision trees, and random forests are common machine learning techniques, frequently applied for agricultural management purposes.

Artificial neural networks (ANNs) have been largely applied for pattern classification and attribute mapping in agriculture since the mid-1990s (Wang, 1994; Kimes et al., 1999). Then, ANNs also appeared to be a promising technique in regression problems related to agriculture, such as estimations of crop features, (e.g. bark volume (Diamantopoulou, 2005)), rainfall and temperature values (Bendre et al., 2016), soil properties (Lahoche et al., 2003), water content to support irrigation (Bendre et al., 2016), optimization of fertilization rates (Pokrajac & Obradovic, 2001) and yield (Pantazi et al., 2016). As an example, Lahoche et al. (2003) presented a method for predicting field soil properties and variability of soil components. ANNs proved well-adapted and more accurate than usual linear models but the resulting prediction models cannot be univer-

sal. Surveys refer to ANNs as a powerful tool for crop yield estimation, as the relationship between variables is not known and is complex (Pande & Varma, 2008; Paswan & Begum, 2013). Nevertheless, the performance and accuracy of these black boxes are highly determined by their configuration and the user usually has no role, except providing the input data. Large training datasets are also needed for accurate results, a procedure which is time-consuming.

Advanced neural networks, such as adaptive neuro-fuzzy inference system (ANFIS) and extreme learning machine (ELM), arose to address several drawbacks of the conventional ANNs (Moreno et al., 2014; Jafari et al., 2016; Deka et al., 2018). ANFIS avoids the fuzziness characteristic of agriculture information and is known for its powerful generalization ability. ELMs have good generalization performance and learn much faster than conventional ANNs. More efforts, though, should be made to prove the reliability of such methods and uncover the hidden knowledge from these black boxes in future work. The ANNs impractical application, complexity, and high computational cost, led to alternative solutions, simpler to train, like support vector machines (SVMs), decision trees (DTs) and random forests (RFs) with great potential for precision agriculture applications (Sridharan & Gowda, 2017; Chlingaryan et al., 2018).

Support vector machines (SVMs), unlike other kernel methods, have good intrinsic generalization ability and are relatively robust to noise in the training data. Rumpf et al. (2010) proposed a procedure for the early detection of sugar beet diseases based on SVM, using spectral vegetation indices. Classification accuracy between healthy and diseased leaves reached 97%. Recently, support vector regression was used for the retrieval of continuous vegetation attributes, estimations, and soil mapping; however, this required more computational training time than other regression techniques (Tuia et al., 2011; Ali et al., 2015; Verrelst et al., 2015).

Decision Trees (DT) have been used more frequently in classification applications, but target variables can also take continuous values (e.g., regression trees that predict yield responses from soil variables) (Chlingaryan et al., 2018). The term Classification And Regression Tree (CART) is an umbrella term used

to refer to both of the above cases, classification and regression. Zheng et al. (2009) used a CART model to predict yield variability responses to variations of soil properties and management practices. In their study, regression trees were used to predict yield responses from soil and agronomic variables, and classification trees were used to identify the most important soil and management variables affecting yield. CART proved to be robust with low prediction error to predict yield. Gradient boosting is a technique typically used with decision trees (especially CART trees), allowing greater flexibility and predictive performance in modelling the data (Colin et al., 2017).

Random forests (RFs) correct trees' habit of overfitting to their training dataset and are popular in applications with attributes' mapping, demonstrating high efficiency and consequently higher prediction performance (Were et al., 2015; Rahmati et al., 2016; De Castro et al., 2018). RFs have relatively small training time (lower than SVMs and ANNs) and an easy parameterization.

Deep Learning (DL) is a quite promising technique that extends classical ANN by adding more complexity (" depth") into the model. The most identified use of DL is image classification (Kamilaris & Prenafeta-Boldú, 2018). For example, convolutional neural networks (CNNs) belong to deep neural networks and are widely used in image recognition, as they use a mathematical process known as convolution to analyze images in non-literal strategies (Andrea et al., 2017). This allows such networks to identify even things that are partially obscured. Nevertheless, such complex neural networks with a huge number of features and fully connected dense layers are prone to overfitting. Deep learning requires large datasets to work well and appropriate infrastructure. Moreover, setting up a neural network using deep learning techniques is much more tedious than using RFs and SVMs but deep learning models are able to locate important features themselves through regularization (possibly in combination with cross-validation).

2.1.2. Comparisons and discussion

There is extensive literature on more machine learning techniques (e.g., Bayesian networks) that have been used in different applications in precision agriculture (Ali et al., 2015; Verrelst et al., 2015; Chlingaryan et al., 2018; Liakos et al., 2018). The interest in identifying the optimal technique for ensuring accuracy and stability for a given agriculture application is great and keeps evolving.

There are several works that try to compare these techniques. For instance, in Pasolli et al. (2011), SVRs proved robust in the presence of outliers and noise with greater estimation accuracy (for soil moisture) when compared to ANNs. In Were et al. (2015), SVR and ANN models (used to map the patterns of soil organic carbon stocks (SOC)) yielded similar performance, and the authors argued about the importance of data quality. Ruß (2009) evaluated different regression techniques to find suitable models that achieve high accuracy and high generality in terms of yield prediction capabilities. Neural networks, despite their site-dependency, proved robust, but the SVR model used was more accurate while being computationally less demanding. Dou & Yang (2017) recommended ELMs and ANFIS to estimate carbon flux in terrestrial ecosystems. These models were compared to ANNs and SVMs and proven to have higher robustness and flexibility. ANNs (and deep learning methods), RFs and SVMs have mainly been reported as classifiers, producing high accuracies (Chinchuluun et al., 2009; Nitze et al., 2012; Raczko & Zagajewski, 2017). Cereda (2016) characterized deep neural networks as the most promising architectures for segmentation tasks of agricultural images.

To conclude, we comment on reliability issues, threads of analyses, and computational aspects of the techniques described:

- ANNs have been applied widely to remote sensing applications in many agricultural fields, being more accurate and well adapted than linear models. Nevertheless, they remain intransparent models, contain a high level of complexity in computational processing (power consumption is increased to unacceptable levels), are prone to overfitting, and demand large datasets

and a time-consuming parameter tuning procedure. Advanced machine learning techniques like ANFIS and ELMs address several neural networks approaches drawbacks, have good generalization ability, and very high performance. Nevertheless, they are also black boxes that demand more research to prove their reliability.

- DTs are the easiest to understand but are too sensitive to small changes in the training dataset, occasionally unstable, and also prone to overfitting.
- SVMs and RFs have great potential for agriculture applications as they are fast, accurate, and require fewer training samples compared to ANNs. They are easier to implement, robust to noise in training data, and less prone to overfitting compared to other methods. SVMs and RFs have increased popularity especially in classification, which is widely used in agriculture.
- DL methods seem very promising in agricultural applications that demand image classification; these achieve recognition accuracy at high levels but require large amounts of data and substantial computational cost.

The decision on which technique to choose depends on the dataset available and the problems complexity. Existing studies that search for the best technique for specific agricultural aspects do not always present the same conclusions and cannot provide global solutions. More studies comparing the aforementioned algorithms with different and extensive training samples over time are needed.

2.2. Contributions to Precision Agriculture

The above review demonstrates machine learning techniques used in several agricultural applications targeting crops, soil, weeds and diseases, and weather/climate change. Next, we review how the aforementioned machine learning methods support precision agriculture. The organization is based on the applications target:

Crops. Applications targeting crops are mainly cases of yield estimation and/or the recognition and estimation of crop features. These applications

provide valuable information for forecasting yields to plan harvest schedules and make market plans, to apply fertilizers variably according to the crops needs, or other crop management practices.

Yield estimation is one of the most important topics in precision agriculture (Ruß, 2009; Zheng et al., 2009; Gonzalez-Sanchez et al., 2014; Pantazi et al., 2016). Accurate and timely forecast of yield is required for marketing, storage, and transportation decisions. Machine learning methods, capable of handling non-linear relationships, can process a large number of inputs. Inputs from different sensing systems like soil (e.g., salt, organic matter) or climate characteristics can be combined to predict yield accurately and provide crop recommendations. As an example, Pantazi et al. (2016) proposed the integration of high sampling resolution multi-layer data on soil and crop by using supervised neural networks to predict the spatial distribution of wheat yield with high accuracy. The estimation of crop features values is also needed in order to better understand the environmental dynamics at a region of interest (Diamantopoulou, 2005; Tuia et al., 2011).

Crop recognition is used for the automatic identification and classification of crop species in a fast and cost-effective manner, avoiding the use of human experts (Kimes et al., 1999; Nitze et al., 2012; Moreno et al., 2014; Raczko & Zagajewski, 2017). Precise crop map building by pixel classification is relevant for the development of precision agriculture where crops are monitored by remote sensing. As an example, Moreno et al. (2014) tried to create accurate thematic maps of soybean using ELM for the classification of remote sensing hyperspectral data. Machine learning techniques, especially RFs, SVMs, and DL methods, have increased popularity in image classification.

Soil. Applications for soil include the estimation of soil components, temperature, and soil moisture content. The knowledge of spatial variability of soil components helps understanding variabilities in production. Accurate estimations of soil properties are needed to optimize soil management, make nutrient planning and take land-use decisions (Lahoche et al., 2003). Land management practices in agriculture may also target the prevention of floods and landslides

to reduce negative environmental impacts.

Carbon storage estimation has gained increasing attention in recent years (Meersmans et al., 2008; Were et al., 2015), due to its interaction with the earths climate system. Maintaining and increasing SOC stocks through improved land use and management practices can help to counteract increasing atmospheric carbon dioxide concentrations. As an example, Were et al. (2015) developed and evaluated support vector regression (SVR), artificial neural network (ANN), and random forest (RF) models for predicting SOC stocks in Kenya using auxiliary data like soil, climatic, topographic, and remotely-sensed data. Prediction maps of SOC stocks were created.

Soil moisture plays also an important role in crop yield variability. Its monitoring enhances the understanding of water exchange rate at the atmosphere/ground interface and has motivated the development of airborne and satellite microwave sensors (Pasolli et al., 2011). Soil moisture estimations can provide high-resolution maps of water content and in tandem with temperature and weather estimates can contribute to the enhancement of irrigation systems and the maintenance of the climatological balance (Rahmati et al., 2016; Hassan-Esfahani et al., 2017).

Soil is highly heterogeneous, with complex mechanisms that are difficult to understand and interpret. For instance, the amount of water in a given place is affected by several geo-environmental factors. Machine learning techniques, when representative models are used, can provide a low cost and reliable solution for the accurate estimation of soil conditions, since they are well-adapted for modelling complex behaviors including several factors of importance. In this way, the time-consuming conventional soil measurements can be avoided.

Diseases and weed detection. Uniformly applying pesticides or fertilizers over an area of interest leads to high financial and significant environmental cost. Residues in crops, water contamination, and impacts on ecosystems are just some of the consequences of this practice. Machine learning techniques can combine various parameters and perform complex non-linear modeling of crop yield dependence on nutrients to have optimal agro-chemicals input targeted in

terms of time and place (Pokrajac & Obradovic, 2001).

Plant diseases are often associated with several physiological and visual modifications of their host plants, but their visual monitoring at early stages in the field is time-consuming and expensive. Alternative evaluation methods like hyperspectral imaging and non-imaging sensors have proven to be useful for detection of early-stages of vegetation stress, identifying small differences in vegetation cover abundances, or measuring leaf pigment concentrations. For precision plant protection, disease detection methods must facilitate an automatic classification of the diseases. Machine learning techniques seem promising to solve this complex agricultural problem, providing classification models to make a prediction for new unlabeled data (Rumpf et al., 2010; Jafari et al., 2016). As an example, Jafari et al. (2016) trained an ANFIS model to distinguish healthy roses from the infected ones to limit the use of excessive chemicals. Digital infrared thermography was used to visualize spatial and temporal changes in the surface temperature of infected and non-infected rose plant leaves.

Apart from diseases, weeds are also a serious threat for producers. They are difficult to be detected between crops, but remote sensing technologies allow us to build accurate classifiers, in order to distinguish weeds from both diseased and healthy crops (Cereda, 2016; Andrea et al., 2017; De Castro et al., 2018). Tools detecting and removing weeds can then minimize the need for herbicides and human intervention.

Weather and climate change. Weather and climate conditions have a profound influence on the growth and yield of crops, affect fertilizer and irrigation requirements, and incidence of pests and diseases. Extreme weather conditions can damage the whole production and cause serious soil erosion, while crop quality during movement from field to storage or to market is severely affected by unpredictable changes in weather. Climate Smart Agriculture is a term that refers to simultaneously improving farm productivity and incomes, increasing adaptive capacity to climate change effects, and reducing green house gas emissions from farming, with the use of an integrated set of technologies and practices (Rao, 2018). Most persistent issues in this category of applica-

tions evolve around capturing the huge heterogeneity of interdisciplinary data. The development of models combining historical data with data collected from several meteorological stations seems promising in providing accurate and in time forecasts to support producers and mitigate the weather effects (Bendre et al., 2016; Deka et al., 2018). As an example, Bendre et al. (2016) collected and analyzed daily minimum and maximum temperature, humidity, and rainfall data from a weather station over a ten-year period to forecast rainfall and support farmers decisions on crop pattern selection and water management. The neural network model used in their study showed a considerable potential of data fusion but the authors mentioned the need for rapid advances in platforms and tools to handle large data for future predictions.

While agriculture is strongly affected by climate and weather conditions, it is also one of the economic sectors that strongly affects climate change itself. The relationship between agriculture and climate change is two-way. Precision agriculture can lower emissions by better targeting inputs to spatial and temporal needs of the fields. Improved soil, water, fertilizer, and pest management can significantly reduce greenhouse gas emissions, while maintaining similar yields and reducing production costs. Advanced machine learning techniques proved valuable to imitate the complex, nonlinear issues in the ecological, climatological, and environmental fields (Dou & Yang, 2018). Table 1 groups by target the agricultural applications mentioned in the text and the machine learning techniques used to support them.

2.3. Challenges and Limitations

Data analysis is a mature scientific field that provides the ground for the development of numerous applications related to agriculture. Machine learning techniques have been applied in multiple precision agriculture data-intensive applications. Data-driven models show promise for automating and significantly improving accuracy, computational efficiency, and cost of various tasks in precision agriculture but some issues are raised (Guo et al., 2014; Ali et al., 2015; Chi et al., 2016; Jones et al., 2017; Chlingaryan et al., 2018):

Table 1: Agricultural applications, organized by main target and data modelling techniques

Reference	**Application**	**Main target**	**LR/MLR**	**ANN**	**Advanced NN**	**DT**	**RF**	**SVM**	**DL**
Drummond et al. (1995)	Yield estimation	Crop	✓						
Khakural et al. (1999)	Yield estimation	Crop	✓						
Kimes et al. (1999)	Type classification, mapping	Crop	✓	✓					
Kravchenko & Bullock (2000)	Yield estimation	Crop	✓						
Diamantopoulou (2005)	Features estimation	Crop	✓	✓					
Ruß (2009)	Yield estimation	Crop		✓		✓		✓	
Zheng et al. (2009)	Yield estimation	Crop				✓			
Tuia et al. (2011)	Features estimation, mapping	Crop						✓	
Nitze et al. (2012)	Type classification	Crop		✓			✓	✓	
Gonzalez-Sanchez et al. (2014)	Yield estimation	Crop	✓	✓		✓		✓	
Moreno et al. (2014)	Type classification, mapping	Crop			✓				
Pantazi et al. (2016)	Yield estimation	Crop		✓					
Andrea et al. (2017)	Weed detection	Crop							✓
Raczko & Zagajewski (2017)	Type classification	Crop		✓			✓	✓	
Wang (1994)	Type classification, mapping (land-suitability assessment)	Soil		✓					
Lahoche et al. (2003)	Properties estimation, mapping	Soil	✓	✓					
Meersmans et al. (2008)	Properties estimation (SOC), mapping	Soil	✓						
Pasolli et al. (2011)	Properties estimation (moisture)	Soil		✓				✓	
Were et al. (2015)	Properties estimation (SOC), mapping	Soil		✓			✓	✓	
Rahmati et al. (2016)	Properties estimation (water content), mapping	Soil					✓		
Hassan-Esfahani et al. (2017)	Properties estimation (water content)	Soil		✓					
Pokrajac & Obradovic (2001)	Fertilizer optimization	Weeds/Diseases		✓					
Rumpf et al. (2010)	Disease detection	Weeds/Diseases						✓	
Cereda (2016)	Growth status recognition & Weed detection	Weeds/Diseases				✓	✓	✓	✓
Jafari et al. (2016)	Disease detection	Weeds/Diseases			✓				
De Castro et al. (2018)	Weed detection	Weeds/Diseases					✓		
Bendre et al. (2016)	Rainfall & Temperature estimation	Weather/Climate Change	✓	✓					
Deka et al. (2018)	Temperature estimation	Weather/Climate Changes			✓			✓	
Dou & Yang (2018)	Carbon flux estimation	Weather/Climate Changes		✓	✓			✓	

- Spatial variability: Spatial variability is of crucial importance to understand the interaction of important variables that affect crop variability. One serious limitation of using the aforementioned models is that they assume homogeneity; fields are not usually homogeneous, leading to false assumptions in yield simulations. The appropriate size of management zone should be carefully identified by experts when making decisions regarding input usage.

- Temporal variability: Appropriate technique selection for each given dataset may vary a lot from one year to another, since new data is always incorporated. Methodologies for optimal data fusion and for making models able to exploit information at different temporal scales are needed to improve the temporal consistency and accuracy of the estimation process.

- Variable selection: It is difficult to establish a constant set of attributes that guarantee good results all the time for all techniques, while some agricultural datasets may also be difficult to model for any technique due to high complexity of the crop behavior. It might make sense that adding more features to the total set of possible features would increase models performance, but a large number of irrelevant features simply increases the possibility to overfit. The challenge for next generation models includes not only modelling the known factors affecting crop yield but also incorporating all of the important factors. This requires the collection of large and suitable datasets that describe the production process.

- Datasets availability: Complicated models with many features compared to the training examples, are likely to overfit. The application of machine learning methods combined with sensing technologies, conducted on small areas with small samples of data, leads to a low ability to generalize the learned parameters to areas with different characteristics. The availability of large datasets from diverse sources is necessary to achieve better generalization.

Table 2: Strengths and weaknesses of machine learning (ML) techniques in precision agriculture

Strengths	**Weaknesses**
• Cope well (generally) with data coming from a diversity of data sources (e.g. historical data from repositories with data sensed or monitored) • Do not demand preconceived relationships from theory (no a priori assumptions on variable relations are made) • Produce adaptive (non-linear) models, ideal for complex relationships • Support accurate and customized decision-making with: – yield estimations – crop type and features recognition – soil properties estimation – disease and weed detection – weather and climate change predictions	• Require filtering out low quality and misleading data • Require data transformation and aggregation techniques • Make strong assumptions about data; – Assume homogeneity while there is inter- and intra-field variability – Cannot permanently guarantee good results • Need large amount of field data -spatial and temporal- e.g. via measurement campaigns • Cannot apply a guiding theory on them; difficult to include all factors of importance (variable selection) • Demand training that can be computationally and timely demanding • Demand expert knowledge e.g. for tuning

Farming systems are affected by various factors like environmental conditions, soil characteristics, managing of crop diseases and weeds, and water availability. Lack of data restricts the capabilities of existing models to include factors of importance and be accurate enough to gain users confidence in their abilities to provide reliable results. Even when large datasets are available, machine learning techniques that focus on learning from data still have to face challenges that are going to be analyzed in the following section. Table 2 overviews the strengths of machine learning techniques used in precision agriculture processes as well as their weaknesses and challenges that have to be faced.

The amount of data collected on farms through sensors like yield monitors, drones, or portable devices has increased dramatically over the last decade. The availability of high quality spectral, spatial, and temporal resolution data can lead to refined and robust models. The types of data have also changed, because, apart from simple numerical values, data may include qualitative measures, images, or videos. The desire to collect information on soil and crop variability and respond to such variability on a fine-scale has become the goal of precision agriculture. The use of big data aims at supporting this goal but several new challenges are posed.

3. Big Data in Agriculture

The emergence of trends like the IoT, robotics, and cloud computing allowed for an increase in the volume, velocity and variety of data generated in agriculture. Big data is a term that describes the data characterized by the three V's dimensionsvelocity, volume and variety. Metadata capturing management practices and technologies, such as seeding depth, seed placement, cultivar, machinery diagnostics, time and motion, dates of tillage, planting, scouting, spraying, and input application are considered big data in agriculture (Ma et al., 2015). We consider that big data analysis in agriculture does not need to satisfy all three dimensions. Based on precision agriculture applications, the use of sensors and GNSS to create spatial variability maps can lead to high volumes of data. The highest volume appears in remote sensing applications because of the large sizes of the images used. Taking into consideration that weed and disease detection require urgent action, relevant projects and alert systems demand high velocity. Nevertheless, soil and crop related approaches for production estimations do not demand immediate actions and rarely have to deal with data of high velocity. Decisions on weather forecasting (e.g., decisions for irrigation or fertilization) also need to be made at almost real time. Papers referring to weeds and diseases, dealing with production security, do not have to access a variety of data to address a problematic issue. On the other hand, modeling of weather and

climate change needs various data sources to provide accurate forecasting and support producers tasks (Kamilaris et al., 2017). Innovative technical and analytical strategies have been developed to cope with such data and are gradually gaining popularity.

Both big data and precision agriculture derived from the advent and application of information and communication technologies (ICT), yet they are not synonymous. Precision agriculture involves site-specific application of inputs and the use of yield monitors. On the other hand, to satisfy the volume and variety characteristics of big data, observations from numerous farm fields, spread over time and space, are needed. Apart from precision data from each field, there is a need for access to additional sources of data naturally residing outside the field. Moreover, precision agriculture employs graphical comparisons of field maps as its dominant method of analysis. However, identifying complex interactions across several production factors and multiple years requires more sophisticated methods (Sonka & Cheng, 2015; Lakkakula, 2016). Analytics is a major differentiating feature of big data and methods for their implementation are presented in subsection 3.2.2..

Despite these differences, precision agriculture provides an input for big data for analytics. Big data analytical platforms in the cloud, and machine learning techniques that drive artificial intelligence, when fully realized, can help (Ali et al., 2015; Kamilaris et al., 2017; Shekhar et al., 2017; **?**). We can consider precision agriculture and big data as complementary to each other. In the following subsection, we group the agricultural big data systems into three categories, we describe each category and present application examples and promising areas of big data application based on the research conducted.

3.1. Agricultural big data systems

Agricultural big data systems can be divided into three categories. We describe each category and present application examples and promising areas of big data application, based on the research conducted. Most of the applications found in the literature use advanced machine learning techniques. Different approaches

are used in several agricultural areas. Specifically, these systems take advantage of IoT technologies but have different scopes. As such, we divide them into three domains:

- **Advanced sensor technology systems** - refers to systems that collect data to characterize spatial and temporal variability in the production system and determine actions to be taken in field. Prescriptive analysis is conducted to determine necessary farm interventions.

- **Risk management systems** - refers to systems that use advanced analytic techniques to manage the risk of crop failure. These systems attempt to make risk management specific to field location, soil type, and desired yields and assess the most probable risks on a given farm. Weather and climate change adaptation and mitigation are common matters of interest in such systems.

- **Agricultural management systems** - refers to systems that provide smart farming solutions. They address farm needs like accounting, food market access and traceability, and wireless linking of farm managers, operators, consumers, and stakeholders, to provide support for better management practices.

3.1.1. Advanced sensor technology systems

Remote sensing provides efficient ways to collect information over very large geographical areas. The availability of spatially and temporally referenced input and output data, incorporating the effects of climate and soil on yields, allows rapid and accurate estimation of production relationships and surpasses the traditional experimental approach (Coble et al., 2018; Chi et al., 2016). Production systems, deploying robotics, advanced sensors, and big data analytics, enables farmers to manage their farms on much smaller, and consequently more precise, scales (Lesser, 2014; Wolfert et al., 2017; Weersink et al., 2018). The resultant analysis supports the automation of several agriculture procedures.

Several firms are active with precision agriculture trials using environmental sensors and big data analytics software to maximize yields at a reduced cost (NEC, 2014; Shendar, 2014). For instance, Monsanto purchased Climate Corporation for its weather data and modeling technology, and John Deere bought Precision Planting to increase the machine learning capabilities of its farm equipment (Carolan, 2017; Weersink et al., 2018). There are also open source projects using such systems. For example, Handsfreehectare (Handsfreehectare, 2019) is a project that aims to use solely automated machines in order to grow arable crops remotely.

By aggregating the results of big data analysis from remote sensing across a large number of farms for a long period of time, valuable trends that were not obvious before can be revealed, and hidden structures and common features from variations in management practices can be explored. In this way, improved customized management practices to local conditions can be then provided (Shekhar et al., 2017). As an example, seed varieties may perform variably in a range of soil types or given different weather patterns. Big data platforms and their underlying algorithms can analyze this information and determine which varieties achieve maximum yield across various soil types and growing conditions.

Moreover, advances in high-throughput genotyping offer the opportunity to select breeding results on-demand. Genomics is a very promising domain of big data application in agriculture, although it is not directly related to to precision agriculture. Precise genetic engineering or genome editing makes it possible to change a crops genome. Big data technologies can assist and speed up plant breeding and are expected to enhance crop yield and disease treatment (Kamilaris et al., 2017; Shekhar et al., 2017; Hu et al., 2018; Weersink et al., 2018). Machine learning can help to identify genomic regions of agronomic value and especially SVM and deep learning techniques are commonly used in the literature. These techniques seem promising to breeding analysts, due to their capability of capturing complex, nonlinear relationships in biological big data (Chen et al., 2018; Hu et al., 2018; Ma et al., 2018).

Using images from remote sensing instruments like drones or satellites is a recent practice to approximate agriculture problems by image analysis (Alves & Cruvinel, 2016; Stratoulias et al., 2017). Generating accurate but also timely maps with high spatial resolution using image samplings remains a scientific challenge in agriculture. Cai et al. (2018) used Landsat multi-temporal scenes and took advantage of the short-wave infrared bands that were proven to be extremely useful in efficiently identifying differences between crops. They combined the new data with spectral data from Landsat satellites over a 15-year period and used supercomputers to handle the huge amount of data. Their Deep Neural Network (DNN) managed to distinguish the crops studied with 95% overall accuracy just two to three months after planting. Their approach seems promising enough to be scaled up to large geographic extents.

Object recognition and classification from aerial and satellite imagery using DNN is a promising area of big data application in agriculture. Crops are systems with increasing complexity in shape and appearance. CNNs have shown excellent capabilities in extracting useful information from images. It has also been shown that the learned features obtained from pretrained CNN models can generalize properly even in different domains for those in which they were trained (Penatti et al., 2015). In-time accurate maps derived from high spatial resolution satellites can determine growing conditions or threats to support best farm management practices on a local scale (Cai et al., 2018).

3.1.2. Risk management systems

Management of risk due to field location, soil type, and mainly to heat stress or freeze is a matter of crucial importance in agriculture. A specific circumstance for farming is the influence of the weather and especially its volatility. Merging datasets is a key operation for data analytics in this case. Regional climate models are used to combine information from global models with regional and local meteorological records to provide climate information for smaller spatial units and support real-time adaptation to climate and weather changes (Lesser, 2014; Chi et al., 2016; Rao, 2018).

For instance, Bendre et al. (2015) presented a scenario of big data application in rainfall forecasting taking advantage of large datasets. The results displayed considerable potential of data fusion in precision agriculture. Climate Corporations platforms use in-season imagery and DL to help producers identify issues early and take action to protect and improve yield (Carbonell, 2016).

Weather-driven crop yield variability across the regions of a country can be properly captured via suitably designed weather indices. Biffis & Chavez (2017) demonstrated how a big data mediated machine learning method (CART) can be employed to mine satellite data and identify optimal weather indices for agricultural food-related weather risk management. The use of big data promises to facilitate the agriculture insurance policies. Weather satellites with wide-area coverage, used in tandem with accurate big data machine learning algorithms that combine the collected meteorological data with auxiliary data (e.g., planting/production records over different areas) can be employed to predict possible crop failures across large regions (Rao, 2018). In this way, insurers can improve the prediction of potential crop performance beyond what weather alone might allow. When greater insight and understanding of crop production risk is developed, better risk management solutions and personalized insurance policies can be offered, and the risk can be priced accurately. Machine learning techniques are valuable in such systems due to their ability to handle heterogeneous data and capture nonlinear and high-order interactions in dynamic environments (Biffis & Chavez, 2017). Several attempts have been made recently to establish insurance programs in developing countries (ACREAfrica, n.d; Raju et al., 2016).

3.1.3. Agricultural management systems

ICT enables farmers to exchange information, establish cooperation, and collaborate. As farmers get connected, software management systems emerge. Agricultural management systems arise to provide accounting services, linking farmers with farm managers and operators, and give benchmarking abilities to farmers by connecting them. Their aim is to help farm operators and agribusi-

nesses around the world collect, integrate, and analyze huge amounts of data from different sources to support their business decisions. Such systems provide smart farming solutions. Smart farming is a term that extends precision agriculture by basing management tasks not only on field-specific data, but also on data enhanced by context and situation awareness, triggered by real-time events (Wolfert et al., 2017). For instance, studies conducted in developing world small farms indicate that farmers are not able to sell harvests due to oversupply or lack of necessary information (Kshetri, 2014). Tools for better yield and demand predictions can enable crops to be integrated to the international supply chain (Kamilaris et al., 2017).

Towards this direction, Singh et al. (2018) proposed a big data analytics approach that collects and analyzes social media data using vector machines to identify issues related with supply chain management in food industries. Their approach led to a cluster of words informing supply chain decision makers about ways to improve various segments of food supply and thus support production planning and scheduling. Social media text analytics can inform decision makers about improving various segments of food supply chain management (Kamilaris et al., 2017; Singh et al., 2018). Demand and supply are affected by many unpredictable factors that interact in a complex manner. Analyzing and interpreting details on market behavior and consumers preferences from several sources in real time can assist producers in making better and faster decisions to satisfy customer requirements. Moreover, spatial mining techniques on collected data can be used to identify regions susceptible to possible disasters (e.g., severe weather conditions), to predict locations inappropriate for sensitive crops, and update supply the chain accordingly (Shekhar et al., 2017).

There are also more holistic approaches (Kaloxylos et al., 2012, 2014; Stojaspal, 2014; Jayaraman et al., 2016; Van Rijmenam, 2017) that, apart from the support of typical farming procedures (e.g. using a meteorological service), could also back producers' participation in international markets and improve profitability through transmission of farm data to agronomists, suppliers, and consultants. This can be accomplished by combining data that come from differ-

ent sources and by providing convenient communication between stakeholders along the food chain that interact, having access to the same datasets in the cloud.

Table 3 presents the opportunities provided by the use of big data in agriculture and the potential benefits. Collecting and analyzing big data generated by automated systems, including digital images and other data from ground sensors, unmanned systems, or remote sensing satellites, and their combination with already existing data pose challenges to successful implementation of precision agriculture. Emerging fields of data mining and machine learning methods are promising approaches to gain insight from such data (Kamilaris et al., 2017; Shekhar et al., 2017; Ip et al., 2018; Morota et al., 2018; Weersink et al., 2018). These methods can help analyzing bigger and more complex data, in order to uncover hidden patterns and reveal trends fast and accurately. The potential of these techniques in big data analysis have not been adequately appreciated in agriculture for a number of reasons examined below.

3.2. Challenges of big data adoption in agriculture

Most of the available open platforms (systems that allow users to alter the code such that the software functions differently than the original programmer intended) mentioned previously result from recent projects; their challenge is still broadly broad adoption, to determine final success. Many of them may still be under development and have not reached their full potential yet. There are several publications describing precision agriculture but reports with evaluation of the economics of big data adoption in agriculture are much less numerous (Lokers et al., 2016; Kamilaris et al., 2017). However, systems' marketplace adoption can be monitored as a means to assess whether there are benefits from their technology.

Several big data applications seem to be suited to large farms and industries (i.e., Climate Corp and Monsanto) that already use data in their decision-making and have access to data captured from machinery, greater access to capital, and resources (Carbonell, 2016). Smaller intercropped fields, though,

may require more manual labor and less mechanized processes. However, there is little research examining this assumption (Jakku et al., 2016; Fleming et al., 2018). Big data could potentially be very useful for non-industrial farming practices, but emerging moral and ethical questions about access, cost, and support should be addressed to realize this benefit. During this initial phase, benefits from data are not so large for the farmers. Concerns are held among growers, that the benefits and risks of big data related developments will be unevenly distributed. Concerted efforts are needed to lay the foundations required for everyone who wants to participate to be able to participate (Jakku et al., 2016). This involves at least improving access to Internet connections even in very remote areas, and by not excluding smaller farmers (e.g., due to high start-up costs and complex contract arrangements) (Fleming et al., 2018). A longer discussion on these concerns is beyond the scope of this paper. The interested reader can check (Carbonell, 2016; Jakku et al., 2016; Fleming et al., 2018).

Expectations from the big data adoption in agriculture are high but this adoption is relatively slow. Next, we describe the challenges that have to be faced to leverage the value that big data have to offer in agriculture.

3.2.1. Collection challenges

In agriculture applications, big data comes from various sources. Combining data from a variety of sources raises concerns about matters of data quality and data fusion, and the access to collected big data raises concerns about security and privacy.

Data-driven methods demand clean and relevant data to be utilized. Incomplete datasets, destroyed data, and the presence of outliers or biases in the training set affect models accuracies. The assessment of data quality demands significant human involvement and expert knowledge. Nevertheless, even semi-automated approaches are not practical when it comes to large volumes of data. Event monitoring and real-time processing of streams of data that are highly applied in agriculture further deteriorate data quality. Techniques like outlier detection, data transformations, cross-validation, and bootstrapping are valu-

able tools in data quality management, but until recently, data quality research has primarily focused on structured data stored in relational databases and file systems.

IoT data in agriculture usually comes in streams from sources in geographical proximity and are more likely to be correlated. The spatiotemporal correlation of data permits advanced sensing techniques like compressive sensing to minimize the sampling rate and consequently the network traffic load (Lee & Choi; Zheng et al., 2017), but demand real-time anomaly detection algorithms (Chen et al., 2015). Research on data quality management is ongoing and computational techniques to tackle the aforementioned challenges are needed (Gudivada et al., 2017).

The traditional multi-source data fusion just handles structured data (Halevy et al., 2006). As we have already mentioned, the availability of large datasets is necessary to help data-driven models achieve better generalization. Recent advancements in cloud computing and distributed processing for voluminous data computing could help integrate resources in different scales but this is insufficient. New methods are also needed to tackle the challenges of data fusion, representation, and cleansing but this still remains an obstacle for the exploitation of IoT big data in agriculture (Nowak et al., 2012). Deep learning models have been shown to be very effective in integrating data from different sources and can handle successfully representation problems like the semantic gap (Srivastava & Salakhutdinov, 2012; Chi et al., 2016; Zhang et al., 2018).

The practice of big data collection also raises concerns over access and security. The ability of researchers to conduct large scale and big data oriented research strongly depends on the availability of farm data. Ag-Analytics (Woodard, 2016) is a platform that supports stakeholders for this purpose. Recently, more big datasets are becoming publicly available (OADA, 2016; Global Open Data for Agriculture and Nutrition, 2010). Nevertheless, data sharing demands special attention to matters of data privacy and security (Sykuta, 2016). Recommendations for governing security, data ownership, data protection, and data use should be set by farm alliances and agriculture technology providers.

Moreover, federal legislation protecting farm data is also required.

3.2.2. Analysis techniques challenges

Big data needs extraordinary techniques to efficiently process its large volume within limited run times. Hypothesis testing and machine learning are the most commonly used ways for data analysis (Peters et al., 2014). Agricultural analysis is largely statistical. Its main intention is to understand the underlying system through an analysis of observations. Such approaches start with a theory and lead to one or more hypotheses. Statistical significance tests try to extract conclusions for the population using small samples of as much as possible high quality data. Nevertheless, in the case of big data, the sample used may represent even the entire population. The underlying concept of big data relies more on correlation and less on causation (Karimi, 2014; Coble et al., 2018). However, there are examples where the two are effectively combined (Peters et al., 2014).

Machine learning techniques do not use preconceived relationships from theory but begin with the data, to examine possible relationships among variables (Chen & Zhang, 2014; Karimi, 2014; Coble et al., 2018). Nevertheless, they cannot be a panacea to all challenges posed by big data. The collected datasets are large and complex making it difficult to deal with typical machine learning techniques. Such techniques often perform poorly when applied to agricultural data. Scalable and parallel techniques are needed to cope with voluminous data. Moreover, big data collected in agriculture violate common assumptions underlying several machine learning and analytics methods, such as the independence and identical distribution of data (i.i.d assumptions). Big data in agriculture exhibits spatio-temporal autocorrelation, has heterogeneity and high dimensionality, is nonstationary, and usually has to be processed in a real-time manner (Fan et al., 2014; Shekhar et al., 2017; Coble et al., 2018).

For instance, if we consider adjacent plots, we will find similar soil-type, climate, and precipitation. Models that are inaccurate or inconsistent with the dataset may be extracted, if we ignore auto-correlation during data analysis. A

variety of spatiotemporal methods used for traditional data could be extended to handle agricultural big data (Shekhar et al., 2015; Li et al., 2016; Golmohammadi et al., 2018).

The unstructured streaming data received from several diverse agricultural sources are multi-dimensional. Having many dimensions gives rise to accumulated error terms and there is no guarantee that every dimension is especially useful for performing analysis (Fan et al., 2014; Shekhar et al., 2017; Coble et al., 2018). Statistical and machine learning techniques do not lower the dimensionality of the problem in a deterministically exact way and they exhibit the "curse of dimensionality" (Krishnamurthy, n.d.; Li et al., 2016). The are several techniques in handling high-dimensional data like Principal Component Analysis and Incremental Singular Value Decomposition, but most of them are based on dimension reduction and usually fail to extract the core value from massive big data (Krishnamurthy, n.d.; Chen & Zhang, 2014).

Moreover, machine learning models are not appropriate for non-stationarity (i.e., cannot be used when new climate and weather patterns or new and improved crops arise), but big data in agriculture exhibit non-stationarity. Most methodologies learn through historical datasets. Consequently, models trained on specific observations should be combined with mechanistic models based on theory and domain knowledge to explore explanatory relationships (Shekhar et al., 2017). Since many agricultural applications are time-sensitive and depend on data freshness, developed models must be retrained to reflect the evolution of data. For learning from high speed streams of data, online learning could be integrated with traditional techniques and theory (Li et al., 2016).

Several current advanced machine learning models have gained a considerable amount of interest as promising frameworks for handling big data in agriculture. Deep learning is of crucial importance in providing predictive analytics solutions for large-scale datasets, especially with the increased processing power and the advances in graphics processors. DNNs can work with thousands of parameters, but complex models can overfit easily. Increasing the dataset to model the interactions among production variables at different locations and

seasons could relieve the overfitting problem but can be unrealistic and costly. RFs using multiple predictor functions (to avoid using just one overfitted function) and kernel methods like SVMs could be useful solutions to avoid overfitting, but SVMs suffer from serious scalability problems in both memory use and computation time (Chen & Zhang, 2014; Morota et al., 2018). To speed up learning, the novel learning algorithm ELM is proposed to deal with high velocity of data; it is able to provide extremely fast learning speed and achieve better generalization.

To handle the big datasets that accompany precision agriculture, analytics methods must scale up in parallel and distributed ways, avoiding high computational complexity. Advances in cloud computing and parallel/distributed architectures can help towards this direction. Cloud computing can be used to integrate sources in different locations, and then the data input can be partitioned into a distributed and parallel architecture. The combination of machine learning and parallel training implementation techniques provides potential ways to process big data. Developed models, though, should be compatible with parallel computing; unfortunately, not all algorithms can be distributed or implemented in parallel form. As a successful example, Parallel SVM (PSVM) (Chang, 2011) reduces memory and time consumption and in Baldominos et al. (2014) a scalable machine learning service is introduced for stream processing and real-time analysis.

3.2.3. Computing infrastructure challenges

Big data demand not only novel analytical paradigms to extract information, but also compatible parallel computing frameworks and novel wireless solutions, that may be too elaborate for an individual farmer. In farm management several technical challenges exist, as diverse and high-dimensional data streams from sensors should be ingested in real time, delivered and analyzed usually in short time, to meet the demands posed by several agricultural applications (Chi et al., 2016). Real-time analysis platforms are needed to deal with online remote sensing data and combine it with offline data from one or more dis-

tributed data centers. Precision agriculture relies heavily on event monitoring that demands data stream processing and consequently requires lower latency and higher bandwidth. Also the amount of disk input/output (I/O) has to be minimized.

The parallelism of Hadoop (the most used open source implementation of MapReduce) (Apache Software Foundation, 2019b) is suitable for batch processing and products for performing advanced analytics on stored big data have largely been built over Hadoop. Nevertheless, Hadoop is not appropriate for nearly real-time routines due to its disk I/O intensiveness. In-memory computing eliminates significant amount of disk I/Os and thus reduces data processing time, enabling the immediate analysis of live data. Recently, a number of real-time stream processing engines like Apache Storm (Apache Software Foundation, 2018), Spark Streaming (Foundation, 2019) and Flink (Apache Software Foundation, 2019a) have been developed for this purpose. Modules that can work with these systems like Mlib (Apache Software Foundation, 2019d) and GraphLab (Low et al., 2012) provide common machine learning operations, while others like Tensorflow (Tensorflow, 2018) are designed to build sophisticated machine learning models like DNNs in real-time.

3.2.4. Storage and interpretation challenges

Results are delivered to target destinations like databases, micro-services, and messaging systems via supported application programming interfaces. Cloud computing, apart from realizing the needed scalability for machine learning algorithms, can also enhance storage capacity through necessary infrastructure. Big data storage led to the development of NoSQL databases (Hbase (Apache Software Foundation, 2019c), Cassandra (Apache Software Foundation, 2016), and MongoDB (MongoDB Inc, 2019)). Many big data tools rely on open source software solutions, which dramatically reduces costs, but expenditures related to hardware, its maintenance, and the training of potential users are matters still to be faced.

Massive amounts of data cannot be interpreted by producers. Big data and

its resultant analysis will not have much impact unless it is understood, adopted, and adapted by farmers and other managers. Visualization is a key component of services intended to enhance precision agriculture. Techniques are needed to make analytics work for producers and help them act on events as they happen. Visualization should be considered as early as possible and in tandem with prior interdisciplinary domain expert knowledge, provide accurate and on time support for decision making. Needed action should be clearly provided, as we cannot expect producers to hire predictors, analysts, and decision-makers for their fields. Research on real-time processing of large volumes of data combined with visualization tools providing interactive exploration is still in progress but very promising (Chi et al., 2016; Wachowiak et al., 2017).

Ultimately, large-scale analysis of agricultural data to support business analytics in high scale and speed necessitates investments in cloud infrastructures, while big data processing demands advanced techniques of parallel and distributed computing. Most traditional machine learning techniques are not inherently efficient or scalable to handle the challenges posed by big data in agriculture. Several current advanced learning methods, though, seem promising enough. Research funded by public organizations and work for the common good are needed to make such tools and techniques enter the public domain (and become open-sourced) (Carbonell, 2016). Once big data research evolves and matures, it is expected that machine learning, given its history, will manage to tackle the challenges posed by big data. Successful applications of big data, though, will be determined not only by technology but by organizational and managerial factors, as well. Strong multidisciplinary engagement by producers with agricultural economists, biologists, computer scientists, and government organizations is needed to make the needed scientific advancement. Table 3 presents the aforementioned challenges that arise from the use of big data analytics in agriculture and mentions the requirements needed to address them (costs).

Table 3: Opportunities, challenges and cost-benefit analysis of BD (analytics) adoption in agriculture

Opportunities	**Benefits**
• Characterize spatial and temporal variability in soil, crop, and environmental characteristics on precise scales • Determine growing conditions and identifying needs and threats in (near) real time • Predict yield, weather, and threats (e.g. extreme climate conditions, infections) accurately and on time	• Automation of agricultural procedures • Accurate and timely decision making in Precision Agriculture
• Explore hidden structures and extract common features on farms across large regions and time scales	• Better personalized on-farm management practices • Improved food access and supply chain management
• Identify wanted traits (e.g. stress tolerance) and dissect their genetics	• Precise and effective genome editing for plant breeding
Challenges	**Costs**
• Data quality issues	• Data quality management techniques required
• Data heterogeneity (data from multiple sources, with different formats, different time points)	• Data preparation, fusion and representation techniques required
• Data availability	• Data initiatives and producers cultural change needed
• Data security holes and privacy concerns	• Laws and regulations needed
• Spatiotemporal autocorrelation of data	• Scalable spatiotemporal methods and explanatory space-time analysis
• High-dimensionality of data	• Spurious correlations, noise accumulation (wrong statistical inference, false conclusions, wrong discoveries) • Effective dimension reduction methods, variable selection methods and large datasets required • Combination of empirical and mechanistic models required
• Non-stationarity of data and velocity	• Stream processing and online learning techniques required • Real-time analysis frameworks for in memory computations (batch and stream) required • Combination of empirical and mechanistic models required
• Voluminous datasets	• Computational cost • Memory cost • Scalable analytics methods in parallel and distributed ways required • Parallel/distributed infrastructure in the cloud required
• Data Interpretation	• Advanced visualization techniques required • Strong multidisciplinary engagement required

4. Conclusions

Challenges of agricultural production are increasing, making the need to understand the complex agricultural ecosystems more imperative than ever. Machine learning techniques are widely applied in precision agriculture due to their capabilities to mine information hidden in agricultural data. Several techniques of increasing complexity and computational requirements are identified along with their strengths and limitations, examining their use and contribution in precision agriculture applications and categorized appropriately. Based on the 29 publications (Table 1), the reader can be informed about where the development of agricultural system models has been and where it could go in the future.

The increasing availability of data through advancements in ICT seems promising for enhancing innovation on strategic decision-making by increasing models' accuracy and generalization ability. Without employing the data generated by precision agriculture practices, it is difficult to predict if big data will have significant impact. On the other hand, learning from massive data is expected to bring significant opportunities and transformative potential for precision agriculture. We consider precision agriculture and big data as complementary fields. Most of the systems studied have not been applied in farming practice yet or if they have, they have not proven their value. As the time for big data is coming, the collection and analysis of datasets is difficult to be dealt by traditional learning methods. These methods are not inherently efficient or scalable enough to work well with large-volume agricultural data exhibiting features like heterogeneity, high dimensionality and spatiotemporal autocorrelation. We provide a discussion about the challenges of learning from big data and several corresponding solutions in recent researches. Advanced machine learning methods like convolutional neural networks and big data processing techniques offer higher accuracy, robustness, flexibility, and generalization performance. Big data also requires considerable technical skills to handle analysis methods, frequently demanding real-time processing, and parallel/distributed infrastruc-

tures. A cost-benefit analysis of big data analytics adoption in agriculture is provided.

Big data are expected to play an important role on several agricultural domains apart from precision agriculture. The outlook for big data and machine learning in agriculture is very promising. High-performance scalable learning systems for data-driven discovery can turn farm management systems into artificial intelligence systems, providing richer real-time recommendations and automation of several agricultural procedures. Emerging fields of advanced machine learning and data mining combined with open datasets and policy frameworks are expected to be instrumental in helping meet the challenges of agricultural production in terms of productivity, environmental impact, food security, and sustainability.

References

ACREAfrica (n.d). URL: https://acreafrica.com.

Ali, I., Greifeneder, F., Stamenkovic, J., Maxim, N., & Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, *7*, 16398–16421. doi:10.3390/rs71215841.

Alves, G. M., & Cruvinel, P. E. (2016). Big data environment for agricultural soil analysis from ct digital images. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (pp. 429–431). doi:10.1109/ICSC.2016.80.

Andrea, C., Mauricio Daniel, B. B., & Jos Misael, J. B. (2017). Precise weed and maize classification through convolutional neuronal networks. In *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)* (pp. 1–6). doi:10.1109/ETCM.2017.8247469.

Apache Software Foundation (2016). Cassandra-manage massive amounts of data, fast, without losing sleep. URL: http://cassandra.apache.org/.

Apache Software Foundation (2018). Apache storm. URL: http://storm.apache.org/.

Apache Software Foundation (2019a). Apache flink-stateful computations over data streams. URL: https://flink.apache.org/.

Apache Software Foundation (2019b). Apache hadoop. URL: http://hadoop.apache.org.

Apache Software Foundation (2019c). Hbase. URL: https://hbase.apache.org/.

Apache Software Foundation (2019d). Mllib is apache spark's scalable machine learning library. URL: https://spark.apache.org/mllib/.

Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014). A scalable machine learning online service for big data real-time analysis. In *2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD)* (pp. 1–8). doi:10.1109/CIBD.2014.7011537.

Bendre, M., Thool, R., & Thool, V. (2016). Big data in precision agriculture through ict: Rainfall prediction using neural network approach. In *S. Satapathy, Y. Bhatt, A. Joshi, and D. Mishra (Eds.) Proceedings of the International Congress on Information and Communication Technology.* (pp. 165–175). doi:10.1007/978-981-10-0767-5_19.

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015). Big data in precision agriculture: Weather forecasting for future farming. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 744–750). doi:10.1109/NGCT.2015.7375220.

Biffis, E., & Chavez, E. (2017). Satellite data and machine learning for weather risk management and food security. *Risk Analysis*, *37*, 1508–1521. doi:10.1111/risa.12847.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote Sensing of Environment*, *210*, 35–47. doi:10.1016/j.rse.2018.02.045.

Carbonell, I. M. (2016). The ethics of big data in big agriculture. *Internet Policy Review*, *5*. URL: https://ssrn.com/abstract=2772247.

Carolan, M. (2017). Publicising food: Big data, precision agriculture, and co-experimental techniques of addition. *Sociologia Ruralis*, *57*, 135–154. doi:10.1111/soru.12120.

Cereda, S. (2016). *A comparison of different neural networks for agricultural image segmentation.*. Master's thesis Politecnico di Milano. URL: https://www.politesi.polimi.it/bitstream/10589/133864/3/tesi.pdf.

Chang, E. Y. (2011). Psvm: Parallelizing support vector machines on distributed computers. In *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception* (pp. 213–230). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-20429-6_10.

Chen, C., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347. doi:10.1016/j.ins.2014.01.015.

Chen, P.-Y., Yang, S., & McCann, J. (2015). Distributed real-time anomaly detection in networked industrial sensing systems. *Transactions on Industrial Electronics*, *62*, 3832–3842. doi:10.1109/TIE.2014.2350451.

Chen, S., Wu, C., & Yongmao, Y. (2018). Analysis of plant breeding on hadoop and spark. *Advances in Agriculture*, *8*. doi:10.3390/agriculture8060075.

Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., & Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, *104*, 2207–2219. doi:10.1109/JPROC.2016.2598228.

Chinchuluun, R., Lee, W., Bhorania, J., & Pardalos, P. (2009). Clustering and classification algorithms in food and agricultural applications: A survey. In *.M. Pardalos and P.J. Papajorgji (Eds.)* (pp. 433–455). volume 25. doi:10.1007/978-0-387-75181-8_21.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61 – 69. doi:https://doi.org/10.1016/j.compag.2018.05.012.

Coble, K. H., Mishra, A. K., Ferrell, S., & Griffin, T. (2018). Big Data in agriculture: A challenge for the future. *Applied Economic Perspectives and Policy*, *40*, 79–96.

Colin, B., Clifford, S., Wu, P., Rathmanner, S., & Mengersen, K. (2017). Using boosted regression trees and remotely sensed data to drive decision-making. *Open Journal of Statistics*, *7*, 859–875. doi:10.4236/ojs.2017.75061.

De Castro, A. I., Torres-Snchez, J., Pea, J. M., Jimnez-Brenes, F. M., Csillik, O., & Lpez-Granados, F. (2018). An automatic random forest-obia algorithm for early weed mapping between and within crop rows using uav imagery. *Remote Sensing*, *10*. doi:10.3390/rs10020285.

Deka, P. C., Patil, A. P., Kumar, P. Y., & Naganna, S. R. (2018). Estimation of dew point temperature using svm and elm for humid and semi-arid regions of india. *ISH Journal of Hydraulic Engineering*, *24*, 190–197. doi:10.1080/09715010.2017.1408037.

Diamantopoulou, M. J. (2005). Artificial neural networks as an alternative tool in pine bark volume estimation. *Computers and Electronics in Agriculture*, *48*, 235 – 244. doi:10.1016/j.compag.2005.04.002.

Dou, X., & Yang, Y. (2017). Modeling and predicting carbon and water fluxes using data-driven techniques in a forest ecosystem. *Forests*, *8*, 498. doi:10.3390/f8120498.

Dou, X., & Yang, Y. (2018). Comprehensive evaluation of machine learning techniques for estimating the responses of carbon fluxes to climatic forces in different terrestrial ecosystems. *Atmosphere*, *9*, 83. doi:10.3390/atmos9030083.

Drummond, S., Sudduth, K., Joshi, A., Birrell, S., & Kitchen, N. (2003). Statistical and neural methods for site-specific yield prediction. *Transactions of the ASAE*, *6*, 5–14. doi:10.13031/2013.12541.

Drummond, S. T., Birrell, S. J., & Sudduth, K. A. (1995). Analysis and correlation methods for spatial data. *American Society of Agricultural Engineers. Annual Meeting*, *95*, 1335.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*, 293–314. doi:10.1093/nsr/nwt032.

Fleming, A., Jakku, E., Lim-Camacho, L., Taylor, B., & Thorburn, P. (2018). Is big data for big farming or for everyone? perceptions in the australian grains industry. *Agronomy for Sustainable Development*, *38*, 24. doi:10.1007/s13593-018-0501-y.

Food and Agriculture Organization of the United Nations (2010). How to feed the world in 2050. URL: http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/\How_to_Feed_the_World_in_2050.pdf.

Foundation, A. S. (2019). Spark streaming-apache spark. URL: http://spark.apache.org/streaming/.

Global Open Data for Agriculture and Nutrition (2010). A global partnership advocating for food security. URL: http://godan.info.

Golmohammadi, J., Xie, Y., Gupta, J., Li, Y., Cai, J., Detor, S., & Shekhar, . S. (2018). *An Introduction to Spatial Data Mining.*. Technical Re-

port Department of Computer Science and Engineering University of Minnesota. URL: https://www.cs.umn.edu/sites/cs.umn.edu/files/tech_reports/18-013_0.pdf TR 18-013.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*, 313–328. doi:10.5424/sjar/2014122-4439.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*, 1–20.

Guo, H., Wang, L., Chen, F., & Liang, D. (2014). Scientific big data and digital earth. *Chinese Science Bulletin*, *59*, 5066–5073. doi:10.1007/s11434-014-0645-3.

Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases* VLDB '06 (pp. 9–16). VLDB Endowment.

Handsfreehectare (2019). Homepage. URL: http://www.handsfreehectare.com/.

Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., & Mckee, M. (2017). Spatial root zone soil water content estimation in agricultural lands using bayesian-based artificial neural networks and high-resolution visual, nir, and thermal imagery. *Irrigation and Drainage*, *6*, 273–288. doi:10.1002/ird.2098.

Hu, H., Scheben, A., & Edwards, D. (2018). Advances in integrating genomics and bioinformatics in the plant breeding pipeline. *Agriculture*, *8*. doi:10.3390/agriculture8060075.

Ip, R., Ang, L., Seng, K., Broster, J., & Pratley, J. (2018). Big data and machine learning for crop protection. *Computers and Electronics in Agriculture*, *151*, 376–383. doi:10.1016/j.compag.2018.06.008.

Jafari, M., Minaei, S., Safaie, N., & Torkamani-Azar, F. (2016). Early detection and classification of powdery mildew-infected rose leaves using anfis based on extracted features of thermal images. *Infrared Physics Technology*, *76*, 338–345. doi:10.1016/j.infrared.2016.03.003.

Jakku, E., Taylor, B., Fleming, A., Mason, C., & Thorburn, P. (2016). *Big data, big trust and collaboration: Exploring the socio-technical enabling conditions for big data in the grains industry.*. Technical Report CSIRO. doi:10.4225/08/5852da5ef0628 eP164134.

Jayaraman, P., Yavari, A., Georgakopoulos, D., Morshed, A., & Zaslavsky, A. (2016). Internet of things platform for smart farming: Experiences and lessons learnt. *Sensors*, *16*, 1884. doi:10.3390/s16111884.

Jones, J., Antle, B. B., J.M, Boote, K., Conant, R., Foster, I., Godfray, H., & Wheeler, . T. (2017). Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural Systems*, *155*, 269–288. doi:10.1016/j.agsy.2016.09.021.

Kaloxylos, A., Eigenmann, R., Teye, F., Politopoulou, Z., Wolfert, S., Shrank, C., & Dillinger M. and... Kormentzas, Z. (2012). Farm management systems and the future internet era. *Computers and Electronics in Agriculture*, *89*, 130–144. doi:10.1016/j.compag.2012.09.002.

Kaloxylos, A., Groumas, A., Sarris, V., Katsikas, L., Magdalinos, P., Antoniou, E., Politopoulou, Z., & ... Maestre Terol, C. (2014). A cloud-based farm management system: Architecture and implementation. *Computers and Electronics in Agriculture*, *100*, 168–179. doi:10.1016/j.compag.2013.11.014.

Kamilaris, A., Kartakoullis, A., & Prenafeta-Bold, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*, 23 – 37. doi:10.1016/j.compag.2017.09.037.

Kamilaris, A., & Prenafeta-Boldú, F. (2018). Deep learning in agriculture: A

survey. *Computers and Electronics in Agriculture*, *147*, 70–90. doi:10.1016/j.compag.2018.02.016.

Karimi, H. (2014). *Big data: techniques and technologies in geoinformatics*.

Khakural, B., Robert, P., & Huggins, D. (1999). Variability of corn/soybean yield and soil/landscape properties across a southwestern minnesota landscape. In *P.C. Robert, R.H. Rust, W.E. Larson (Eds.)* (pp. 573–579). doi:10.2134/1999.precisionagproc4.c51.

Kimes, D. S., Nelson, R. F., Salas, W. A., & Skole, D. L. (1999). Mapping secondary tropical forest and forest age from spot hrv data. *International Journal of Remote Sensing*, *20*, 3625–3640. doi:10.1080/014311699211246.

Kravchenko, A., & Bullock, D. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, *92*, 75–83.

Krishnamurthy, S. (n.d.). Dealing with high-dimensionality in large data sets. part 1: Foundations and basics. URL: http://www.quantuniversity.com/High-Dimensions.pdf.

Kshetri, N. (2014). The emerging role of big data in key development issues: Opportunities, challenges, and concerns. *Big Data Society*, . doi:10.1177/2053951714564227.

Lahoche, F., Godard, C., Fourty, T., Lelandais, V., & Lepoutre, D. (2003). An innovative approach based on neural networks for predicting soil component variability. In *Proceedings of the 6th International Conference on Precision Agriculture and Other Precision Resources Management* (pp. 803–816). URL: http://www.grignon.inra.fr/economie-publique/publi/innovative_approach.pdf.

Lakkakula, P. (2016). Spotlight on economics: What is big data in the context of agriculture? URL: https://www.ag.ndsu.edu/news/columns/spotlight-on-economics/

spotlight-on-economics-what-is-big-data-in-the-context-of\
-agriculture/.

Lambert, D. M., Lowenberg-Deboer, J., & Bongiovanni, R. (2004). A comparison of four spatial regression models for yield monitor data: A case study from argentina. *Precision Agriculture*, *5*, 579–600. doi:10.1007/s11119-004-6344-3.

Lee, D., & Choi, J. (). Learning compressive sensing models for big spatio-temporal data. In *Proceedings of the 2015 SIAM International Conference on Data Mining* (pp. 667–675). doi:10.1137/1.9781611974010.75.

Lesser, A. (2014). *Big data and big agriculture.*. Technical Report Analyst. URL: https://gigaom.com/report/big-data-and-big-agriculture/ last time accessed on November 23rd, 2018.

Li, S., Dragicevic, S., Antn Castro, F., Sester, M., Winter, S., Coltekin, A., & Cheng, . T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. doi:10.1016/j.isprsjprs.2015.10.012.

Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*, 2674. doi:10.3390/s18082674.

Lokers, R., Knapen, R., Janssen, S., van Randen Y., & Jansen, J. (2016). Analysis of big data technologies for use in agro-environmental science. *Environmental Modelling Software*, *84*, 494–504. doi:10.1016/j.envsoft.2016.07.017.

Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., & Hellerstein, J. M. (2012). Distributed graphlab: A framework for machine learning and data mining in the cloud. *Proc. VLDB Endow.*, *5*, 716–727. doi:10.14778/2212351.2212354.

Ma, W., Qiu, Z., Song, C.-Q., J., & Ma, C. (2018). Deepgs: Predicting phenotypes from genotypes using deep learning. *Planta*, *248*, 1307–1318. doi:10.1007/s00425-018-2976-9.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, *51*, 47–60. doi:10.1016/j.future.2014.10.029.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., & Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of soil organic carbon (soc) at the regional scale (flanders, belgium). *Geoderma*, *143*, 1–13. doi:10.1016/j.geoderma.2007.08.025.

MongoDB Inc (2019). Mongodb-the database for modern applications. URL: https://www.mongodb.com/.

Moreno, R., Corona, F., Lendasse, A., Graña, M., & Galvão, L. S. (2014). Extreme learning machines for soybean classification in remote sensing hyperspectral images. *Neurocomputing*, *128*, 207 – 216. doi:https://doi.org/10.1016/j.neucom.2013.03.057.

Morota, G., Ventura, R., Silva, K. M., F. F, & Fernando, S. (2018). Big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of Animal Science*, *96*, 1540–1550. doi:10.1093/jas/sky014.

National Research Council (1997). *Precision Agriculture in the 21st Century: Geospatial and Information Technologies in Crop Management*. Washington, DC: The National Academies Press. doi:10.17226/5491.

NEC (2014). Nec and dacom collaborate on precision farming solution to maximize yields and reduce costs. URL: https://www.nec.com/en/press/201410/global_20141023_03.html.

Nitze, I., Schulthess, U., & Asche, H. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification.

In *Proceedings of the 4th Geobia* (pp. 35–40). URL: https://pdfs.semanticscholar.org/a10e/79e21be020daab308d0fb5aafe3b3efa5adf.pdf?_ga=2.13100983.1606770651.1555587080-865103835.1554664684.

Nowak, R., Biedrzycki, R., & Misiurewicz, J. (2012). Machine learning methods in data fusion systems. In *13th International Radar Symposium* (pp. 400–405). doi:10.1109/IRS.2012.6233354.

OADA (2016). Open ag data alliance. URL: http://openag.io.

Pande, J., & Varma, B. (2008). Survey of crop yield estimation models with emphasis on artificial neural network model. In *Proceedings of the 2nd National Conference: INDIACom-2008 Computing for Nation Development*. URL: https://www.bvicam.ac.in/news/INDIACom%202008%20Proceedings/pdfs/papers/60.pdf.

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., & Mouazen, A. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57 – 65. doi:https://doi.org/10.1016/j.compag.2015.11.018.

Pasolli, L., Notarnicola, C., & Bruzzone, L. (2011). Estimating soil moisture with the support vector regression technique. *IEEE Geoscience and Remote Sensing Letters*, *8*, 1080–1084. doi:10.1109/LGRS.2011.2156759.

Paswan, R., & Begum, S. A. (2013). Regression and neural networks models for prediction of crop production. *International Journal of Scientific Engineering Research*, *4*, 98–108. URL: https://www.ijser.org/researchpaper/Regression-and-Neural-Networks-Models-for-Prediction-of-\Crop-Production.pdf.

Penatti, O. A. B., Nogueira, K., & dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 44–51). doi:10.1109/CVPRW.2015.7301382.

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., & Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, *5*, 1–15. doi:10.1890/ES13-00359.1.

Pokrajac, D., & Obradovic, Z. (2001). Neural network-based software for fertilizer optimization in precision farming. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)* (pp. 2110–2115 vol.3). volume 3. doi:10.1109/IJCNN.2001.938492.

Raczko, E., & Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral apex images. *European Journal of Remote Sensing*, *50*, 144–154. doi:10.1080/22797254.2017.1299557.

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of gis-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at mehran region, iran. *CATENA*, *137*, 360 – 372. doi:10.1016/j.catena.2015.10.010.

Raju, K., Naik, G., Ramseshan, R., Pandey, T., Joshi, P., Anantha, K., & Kesava Rao, A. D. K. C. (2016). Transforming weather index-based crop insurance in india: Protecting small farmers from distress. status and a way forward. URL: https://core.ac.uk/download/pdf/78386894.pdf.

Rao, N. (2018). Big data and climate smart agriculture - review of current status and implications for agricultural research and innovation in india. In *Proceedings Indian National Science Academy.*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2979349.

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., & Plumer, L. (2010). Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, *74*, 91 – 99. doi:https://doi.org/10.1016/j.compag.2010.06.009.

Ruß, G. (2009). Data mining of agricultural yield data: A comparison of regression models. In *P. Perner (Ed.)* (pp. 24–37). doi:10.1007/978-3-642-03067-3_3.

Sadler, E., Jones, J., & Sudduth, K. (2007). Modeling for precision agriculture:how good is good enough, and how can we tell? In *Proceedings of the 6th European Conference on Precision Agriculture* (pp. 241–248). URL: https://www.ars.usda.gov/ARSUserFiles/50701000/cswq-0314-sadler.pdf.

Shekhar, S., Jiang, Z., Ali, R., Eftelioglu, E., Tang, X., Gunturi, V., & Zhou, X. (2015). Spatiotemporal data mining: A computational perspective. *ISPRS International Journal of Geo-Information*, *4*, 2306–2338. doi:10.3390/ijgi4042306.

Shekhar, S., Schnable, P., LeBauer, D., Baylis, K., & VanderWaal, K. (2017). Agriculture big data (agbd) challenges and opportunities from farm to table. URL: https://pdfs.semanticscholar.org/c815/75e059a826f39b47367fceaac67a8f55fb07.pdf.

Shendar, N. (2014). Zadara storage helps farm intelligence build petabyte-scale 'big data for crops' analytics service in the aws cloud. URL: https://zadara.com/blog/2014/03/20/zadara-storage-helps-farm-intelligence-build-petabyte-scale\-big-data-for-crops-analytics-service-in-the-aws-cloud/.

Singh, A., Shukla, N., & Mishra, N. (2018). Social media data analytics to improve supply chain management in food industries. *Transportation Research. Part E: Logistics and Transportation Review*, *114*, 398–415. doi:10.1016/j.tre.2017.05.008.

Sonka, S., & Cheng, Y.-T. (2015). Precision agriculture: Not the same as big data but... URL: https://farmdocdaily.illinois.edu/2015/11/precision-agriculture-not-the-same-as-big-data.html.

Sridharan, M., & Gowda, P. (2017). Application of statistical machine learning algorithms in precision agriculture. In *Proceedings of the 7th Asian-Australasian Conference on Precision Agriculture*. doi:10.5281/zenodo.895303.

Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep boltzmann machines. In *Proceedings of the Neural Information Processing Systems Conference*. URL: https://papers.nips.cc/paper/4683-multimodal-learning-\with-deep-boltzmann-machines.pdf.

Stojaspal, J. (2014). Precision agriculture: From real-time farming data to meta analysis. URL: https://www.tu-auto.com/precision-agriculture-from-real-time-farming-data-to-meta-analysis/.

Stratoulias, D., Tolpekin, V., De By, R., Zurita-Milla, R., Retsios, V., Bijker, W., Alfi Hasan, M., & Vermote, E. (2017). A workflow for automated satellite image processing: From raw vhsr data to object-based spectral information for smallholder agriculture. *Remote Sensing*, *9*, 1048. doi:10.3390/rs9101048.

Sykuta, M. (2016). Big data in agriculture: Property rights, privacy and competition in ag data services. *International Food and Agribusiness Management Review*, *19*, 57–74. URL: https://www.ifama.org/resources/Documents/v19ia/320150137.pdf.

Tensorflow (2018). An end-to-end open source machine learning platform. URL: https://www.tensorflow.org/.

Tuia, D., Verrelst, J., Alonso, L., Perez-Cruz, F., & Camps-Valls, G. (2011). Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geoscience and Remote Sensing Letters*, *8*, 804–808. doi:10.1109/LGRS.2011.2109934.

Van Rijmenam, M. (2017). John deere is revolutionizing farming with big data. URL: https://datafloq.com/read/john-deere-revolutionizing-farming-big-data/.

Verrelst, J., Camps-Valls, G., noz Marí, J. M., Rivera, J. P., Veroustraete, F., Clevers, J. G., & Moreno, J. (2015). Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties -a review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *108*, 273 – 290. doi:10.1016/j.isprsjprs.2015.05.005.

Wachowiak, M., Walters, D., Kovacs, J., Wachowiak-Smolkov, R., & James, A. (2017). Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. *Computers and Electronics in Agriculture*, *143*, 149–164. doi:10.1016/j.compag.2017.09.035.

Wang, F. (1994). The use of artificial neural networks in a geographical information system for agricultural land-suitability assessment. *Environment and Planning A: Economy and Space*, *26*, 265–284. doi:10.1068/a260265.

Wang, N., & Li, Z. (2013). 8 - wireless sensor networks (wsns) in the agricultural and food industries. In D. G. Caldwell (Ed.), *Robotics and Automation in the Food Industry* Woodhead Publishing Series in Food Science, Technology and Nutrition (pp. 171 – 199). Woodhead Publishing. doi:https://doi.org/10.1533/9780857095763.1.171.

Weersink, A., Fraser, E., Pannell, D., Duncan, E., & Rotz, S. (2018). Opportunities and challenges for big data in agricultural and environmental analysis. *Annual Review of Resource Economics*, *10*, 19–37. doi:10.1146/annurev-resource-100516-053654.

Were, K., Bui, D., & Dick, S.-B., B (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. *Ecological Indicators*, *52*, 394–403. doi:10.1016/j.ecolind.2014.12.028.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming-a review. *Agricultural Systems*, *153*, 69–80. doi:10.1016/j.agsy.2017.01.023.

Woodard, J. (2016). Big data and ag-analytics: An open source, open data platform for agricultural environmental finance, insurance, and risk. *Agricultural Finance Review*, *76*, 15–26. doi:10.1108/AFR-03-2016-0018.

Zhang, L., Xie, Y., Xidao, L., & Zhang, X. (2018). Multi-source heterogeneous data fusion. In *Proceedings of the International Conference on Artificial Intelligence and Big Data (ICAIBD)*. doi:10.1109/ICAIBD.2018.8396165.

Zheng, H., Chen, L., Han, X., Zhao, X., & Ma, Y. (2009). Classification and regression tree (cart) for analysis of soybean yield variability among fields in northeast china: The importance of phosphorus application rates under drought conditions. *Agriculture, Ecosystems Environment*, *132*, 98 – 105. doi:10.1016/j.agee.2009.03.004.

Zheng, H., Li, J., Feng, X., Guo, W., Chen, Z., & Xiong, N. (2017). Spatial-temporal data collection with compressive sensing in mobile sensor networks. *Sensors*, *17*, 2575. doi:10.3390/s17112575.