WEED RESEARCH An International Journal of Weed Biology, Ecology and Vegetation Management

DOI: 10.1111/wre.12255

Big Data for weed control and crop protection

F K VAN EVERT*, S FOUNTAS†, D JAKOVETIC‡§, V CRNOJEVIC‡§, I TRAVLOS¶ & C KEMPENAAR*

*Wageningen University & Research, Wageningen, The Netherlands, †Natural Resources Management and Agricultural Engineering, Agricultural University of Athens, Athens, Greece, ‡BioSense Institute, Novi Sad, Serbia, §Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia and ¶Faculty of Crop Science, Agricultural University of Athens, Athens, Greece

Received 27 September 2016
Revised version accepted 22 March 2017
Subject Editor: Rob Freckleton, Sheffield, UK

Summary

Farmers have access to many data-intensive technologies to help them monitor and control weeds and pests. Data collection, data modelling and analysis, and data sharing have become core challenges in weed control and crop protection. We review the challenges and opportunities of Big Data in agriculture: the nature of data collected, Big Data analytics and tools to present the analyses that allow improved crop management decisions for weed control and crop protection. Big Data storage and querying incurs significant challenges, due to the need to distribute data across several machines, as well as due to constantly growing and evolving data from different sources. Semantic technologies are helpful when data from several sources are combined, which involves the challenge of detecting interactions of potential agronomic importance and establishing relationships between data items in terms of meanings and units. Data ownership is analysed using the ethical matrix method to identify the concerns of farmers, agribusiness owners, consumers and the environment. Big Data analytics models are outlined, together with numerical algorithms for training them. Advances and tools to present processed Big Data in the form of actionable information to farmers are reviewed, and a success story from the Netherlands is highlighted. Finally, it is argued that the potential utility of Big Data for weed control is large, especially for invasive, parasitic and herbicide-resistant weeds. This potential can only be realised when agricultural scientists collaborate with data scientists and when organisational, ethical and legal arrangements of data sharing are established.

Keywords: neural network, graphical model, support vector machine, multivariate regression, data ownership, data sharing, semantics.

van Evert FK, Fountas S, Jakovetic D, Crnojevic V, Travlos I & Kempenaar C. (2017). Big Data for weed control and crop protection. *Weed Research* **57**, 218–233.

Introduction

Food production must increase by 70% in order to feed a world population that is expected to reach 9.6 billion by 2050 (Foley, 2011; Foley *et al*., 2011). This challenge is even greater, when we take into account the scarcity of new arable land, the effects of climate change on agricultural production and the societal demand for decreasing the environmental impact of agriculture (Foley *et al*., 2011). Weed management will be of crucial importance, given that crop yield losses caused by weeds (about 32%) are higher than those caused by either pests (18%) or pathogens (15%) (Oerke & Dehne, 2004).

Farmers have access to three categories of data-intensive technologies to help address the above

Correspondence: F K van Evert, Wageningen University & Research, PO Box 16, Wageningen 6700 AA, The Netherlands. Tel: (+31) 317 480573; Fax: (+31) 317 418 094; E-mail: frits.vanevert@wur.nl

mentioned challenges: (i) Farm Management Information Systems (FMIS), which refer to a planned system for collecting, processing, storing and disseminating data in the form needed to carry out a farm's operations and functions (Fountas *et al*., 2015); (ii) precision agriculture, which is the scientific domain that deals with management of spatial and temporal variability to improve economic returns and reduce environmental impact (Blackmore *et al*., 2003); and (iii) agricultural automation and robotics, which is the process of applying robotics, automatic control and artificial intelligence techniques at all levels of agricultural production (Zhang & Pierce, 2013).

However, it is not just farmers who will use Big Data solutions for weed control. In several European countries, the number of invasive plants (IAS, invasive alien species) has significantly increased during the last decades (De Almeida & Freitas, 2012; Pyšek *et al*., 2012). Big Data solutions have been developed to prevent further spread of IAS (Pěknicová & Berchová-Bímová, 2016). For example, areas vulnerable to invasive weeds were identified using species distribution data and data on local environmental conditions in conjunction with species distribution models, GIS software and statistical tools (Guisan & Zimmermann, 2000; Thuiller *et al*., 2009). These vulnerable areas can then be subjected to monitoring. Early *et al*. (2016) provided the first global, spatial forecast of weed invasions in the 21 century by analysing spatial data for the factors that determine introduction and establishment of IAS. Big Data analysis has also been used to predict the spread of IAS with particular preferences for soil, water and temperature, such as *Oxalis pes-caprae* L. (bermuda buttercup), *Solanum elaeagnifolium* Cav. (silverleaf nightshade) and *Taraxacum* spp. (Travlos *et al*., 2008; Luo & Cardina, 2012; Travlos, 2013b). Finally, Big Data analysis will result in a better understanding of the biology and ecology of several parasitic weeds like *Orobanche* spp. and *Phelipanche* spp., which in turn will enable better management (Song *et al*., 2005; Prider *et al*., 2012).

Plant invasions on global and regional scales pose severe ecological, agricultural and health concerns resulting in considerable economic losses. *Ambrosia artemisiifolia* L. (common ragweed) is an important agricultural weed, especially in spring-sown crops, such as sunflower, maize, sugarbeet and soyabean. A main problem with this plant is its enormous production of highly allergenic pollen grains, generating huge medical costs and reduced quality of life among the allergic population (Fumanal *et al*., 2007). The highly allergenic pollen causes sensitisation of up to 60% of the allergic population, with annual medical costs of these allergies amounting to, for example, €110 million in Hungary and €88 million in Austria (Gerber *et al*., 2011). The European Aeroallergen Network (EAN) pollen database (https://ean.polleninfo.eu/Ean/) holds information from more than 600 pollen-monitoring stations from all over Europe. EAN data have been used to identify large local permanent or expanding populations of ragweed (Šikoparija *et al*., 2009; Thibaudon *et al*., 2010). Combined with other data sources, this can lead to early detection and eradication in new areas and the development of a sustainable management strategy of *A. artemisiifolia* in several invaded or potentially susceptible habitats.

Data-driven innovations have already revolutionised several sectors of the economy. The promise that a similar revolution in agriculture may provide benefits is contributing to a growing interest in the application of Information and Communication Technology (ICT) in agriculture. Data collection, data modelling and analysis, and data sharing have become core challenges, an opportunity for innovation and a growth area for commercial development. Vast amounts of data are collected with proximal, airborne or satellite-based sensors, *in situ* sensors (i.e. soil moisture sensors), on-farm weather stations and instrumented farm equipment. This qualifies as Big Data according to the definition of De Mauro *et al*. (2016), namely information assets that are characterised by high volume, high velocity and high variety and that require specific technology and analytical methods for its transformation into value.

In addition, there is a need to share data across the supply chain, both to increase the efficiency of the supply chain and to respond appropriately to agricultural standards, such as integrated crop and weed management. Consumer pressure for more information about agronomic practices creates technical and business model opportunities, if the right architectures, analytical tools and data presentations can be developed. The growth of open data and linked data provides opportunities to integrate data from multiple sources and thus to provide new insights and new services. The combination and proper analysis of Big Data from previous records in a wide area, together with specific measurements and data from field history, can result in the quick evaluation and management of herbicide-resistant weeds. This can be further accompanied by decision-support systems, to find the ideal tailor-made solutions for each case.

The tools provided by precision agriculture and other information technologies have not yet moved into mainstream agricultural management. In general, adoption of technological innovations depends on characteristics of the innovation (e.g. cost, complexity), the innovator and his or her socio-economic

background (e.g. preferences and educational level of farmer), the perceived usefulness and ease of use (Rogers, 1995). This has been confirmed for agricultural innovations (Pedersen *et al.*, 2004; Kutter *et al.*, 2011; Lawson *et al.*, 2011; Fountas *et al.*, 2015). In agriculture in general, the adoption of innovations is also highly dependent on the knowledge support system in place (Straub, 2009).

The aim of this study was to provide an overview of technologies relevant to the application of Big Data for weed control and crop protection, to highlight noteworthy examples and to indicate the work that is still needed to increase the exploitation of Big Data. The remainder of this study is structured as follows. In the following three sections, we describe the building blocks for Big Data in weed control and crop protection, namely data (Big Data capture, storage and sharing), data analytics (Big Data analytics) and thirdly delivering information to farmers (delivery of actionable information to farmers). We then discuss existing decision-support systems for weed control and crop protection and describe opportunities for further development (current applications of Big Data for weed control and crop protection). Conclusions and recommendations are given in the final section.

Big Data capture, storage and sharing

Precision agriculture is an information-intensive, cyclic activity, which can be divided into data collection, data analysis, decision-making and evaluation of decisions (Fig. 1) (Fountas *et al.*, 2006). It is useful to characterise decisions based on the planning horizon and to distinguish strategic, tactical and operational decisions. An example of a strategic decision is whether or not to use precision agriculture; an example of a tactical decision is which crops to include in the rotation; finally, operational decisions have to be made on a day-to-day basis regarding the timing of field operations and the amounts inputs used.

Where does the data come from?

The data in precision agriculture originate from many sources. Crop and soil management data describe the operations that are carried out in the field: tilling, planting, fertilisation, crop protection, weed management and harvest, along with the details such as date, kind of seed or fertiliser or chemicals used, as well as the amounts and the manner in which they are applied. The volume of this information is very small, just a few hundred bytes ha^{-1} $year^{-1}$ (Table 1, Fig. 2), and it is often recorded manually by the farmer in a Farm Management Information System (FMIS). Another kind of information concerns samples of soil and plants that are sent to a laboratory for analysis of texture, chemical composition and potential presence of pathogens and weeds. Yields are recorded at the end of the season and will certainly show up on the receipts sent by the cooperative or private buyer to whom the product is shipped.

By far the largest amount of data results from automatic recording with electronic sensors. These include automated weather stations on farms, soil moisture sensors and an increasing number of sensors attached to quads, tractors, harvesters and (semi-)autonomous ground and aerial vehicles (Table 1, Fig. 2).

Data storage

Once collected, data must be physically stored and organised in such a way that it can be queried. In the 1960s, relational databases evolved as the standard to model and store data, in part because relational databases can model alternative types of databases, such as hierarchical and network databases. The behaviour of

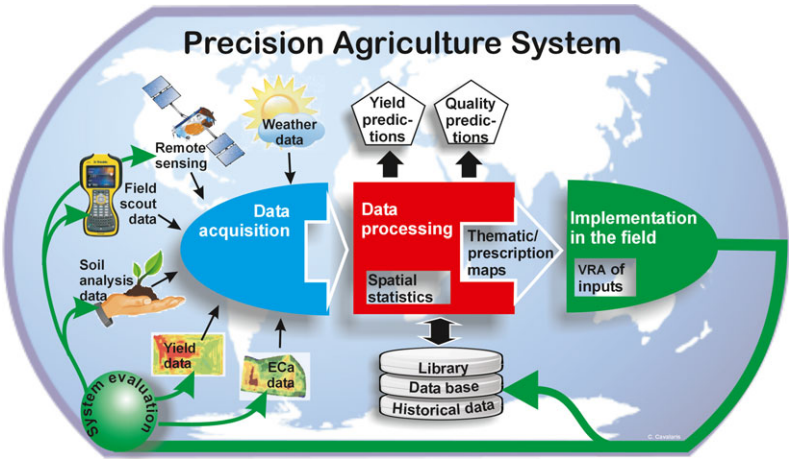

Fig. 1 Presentation of a precision agriculture system (courtesy: University of Thessaly, Volos, Greece).

Table 1 Volume of data produced by selected data sources. The area of the circles in Figure 2 is related to the volume of data

Symbol	Type	Subtype	Single measurement			Volume MB ha^{-1} year^{-1}
			Size bytes	Resolution m^2	Frequency year^{-1}	
CM	Crop management		400	50 000	1	0.00008
AW	Weather station		21 0240	500 000	1	0.00420
NA	Reflectance	N-Sensor ALS	16	152	10	0.01056
GS	Reflectance	Greenseeker	16	38	10	0.04224
OX	Reflectance	OptRx	24	38	10	0.06336
NP	Reflectance	N-Sensor	216	152	10	0.14256
WL	Weed locations	–	10	1	5	0.50000
UM	Reflectance	Multispectral camera on UAV	16	0.01	10	160
UH	Reflectance	Hyperspectral camera on UAV	160	0.01	10	1600
G1	Generic	1 measurement per m^2	4	1	1	0.04000
G2	Generic	1 measurement per cm^2	4	0.0001	1	400

CM: Data on crop, cultivar, planting date, seed rate, fertilizer applications, pesticide applications – here it is assumed that 100 single-precision numbers suffice to describe this information, resolution is a field of 5 ha. AW: Hourly measurements of temperature, precipitation, solar radiation, wind speed, air humidity, one station suffices for a farm of 50 ha; NA: Vegetation index SN, time, location; a working width of 50 m is assumed, one measurement per second, driving speed of 10 km/h; GS: Vegetation index V1, time, location; OX: Reflectance in three bands, V1, V2, time, location; NP: As NA, but intensity of incoming and reflected light in 56 bands; WL: This information could be produced by a robot that records the location and species of weeds, on average 1 weed per m^2; UM: Reflectance with multispectral (four bands) camera on UAV; UH: Reflectance with hyperspectral (40 bands) camera on UAV; G1: Generic measurement, once per m2; G2: Generic measurement, once per cm^2.

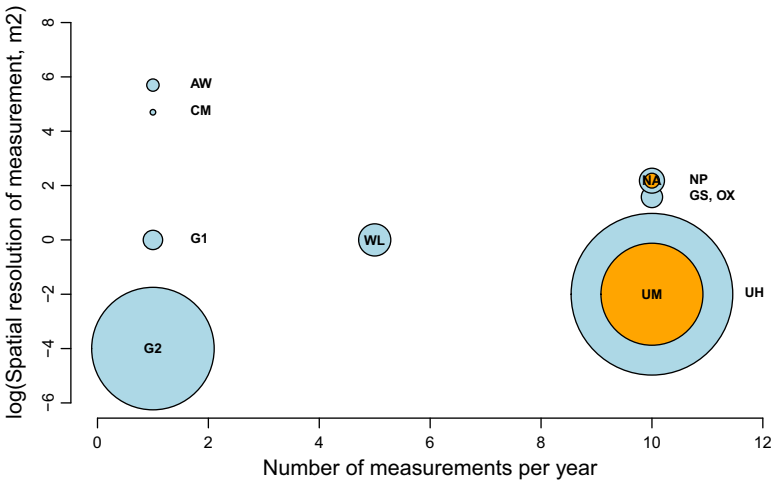

Fig. 2 Overview of spatial and temporal characteristics of common measurements. Horizontal axis: frequency of the measurement (year^{-1}); vertical axis: spatial resolution of the measurement (^{10}log(m^2)). See Table 1 for explanation of symbols.

relational data can be fully described using set theory (Codd, 1970). There is ample literature using the relational model to store agricultural data, including work started by decision support system for agrotechnology transfer (DSSAT) and continued by the International Consortium for Agricultural Systems Applications (ICASA) (Hunt *et al*., 1994; White *et al*., 2013) and the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig *et al*., 2013), but also by others (Van Evert *et al*., 1999a,b; Steiner *et al*., 2009).

In Big Data applications, specific requirements with respect to storing and searching tend to make the use of relational databases difficult. For example, the data usually have to be distributed across several machines due to its volume, and it may moreover be constantly growing and evolving. In such situations, it might be challenging to 'partition' a relational database management system (RDBMS) across multiple machines and maintain it as new data continue to pour in (Marz & Warren, 2015). Moreover, searching can be slow in very large relational databases. To cope with these challenges, several alternatives for RDBMS are employed with Big Data systems, including NoSQL ('not only' Structured Query Language) databases like key value stores (e.g. Riak, http://basho.com/products/riak-kv/), document stores (e.g. MongoDB, http://www.mongodb.com), or distributed storage (e.g. Google's Bigtable) (Chang *et al*., 2008).

Big Data storage and querying may be made more efficient through a lambda-architecture (Marz & Warren, 2015). With a lambda-architecture, arbitrary 'views' (queries) are *pre-computed* over the stored data; that is, they are made ready before an actual request for them is posed. This ensures that the needed information can be retrieved quickly when a request is placed for a specific view. Clearly, pre-computing these views takes a certain amount of time (say, some hours). Consequently, new data are available only a few hours after arrival in the system. This can be compensated through an additional system component (termed the speed layer), which is responsible for processing new ('incremental') data. While this system part still has to provide high-speed querying on the data, this is required only on the *data increment*, and not on the entire data set, hence allowing for an efficient system overall.

Linked data

Applications of Big Data typically involve the challenging task of establishing relationships between data items of different provenance. For example, the term 'wheat yield' may refer to 'yield-as-harvested' (e.g. 11.6 Mg ha^{-1}, moisture content not known), to 'dry matter yield' (e.g. 10 Mg ha^{-1}), or to 'yield adjusted to market-standard moisture content' (e.g. 11.35 Mg ha^{-1}). The meaning of the term 'yield' is slightly different in each case, and it would be an error to use them interchangeably. Similarly, yield may be expressed using units of t ha^{-1}, but also g m^{-2} or dt ha^{-1} (in common use in Germany). Again, errors will occur if units are not taken into account.

When data are stored in table format (e.g. in a database, spreadsheet, or text file), the names of columns typically give an indication of the meaning and the units of the data, but this is rarely conclusive. The manual intervention that is almost always needed to bring data from two or more sources together constitutes a significant barrier to the application of Big Data in agriculture.

A system to address these shortcomings and to make automated matching of data possible has been proposed (Berners-Lee *et al.*, 2001). They proposed the name 'semantic web', but the name currently used is Linked Data. An introduction to recent developments is available (Allemang & Hendler, 2011). Linked Data is built on a number of principles. First, every 'thing' is given a name: a uniform resource identifier (URI). Second, this name preferably is a uniform resource locator (URL) which you can type into a web browser and then will give you information about the thing. In the case of the above example, 'dry matter yield' would have a different name (perhaps http://ld.example.org/dry-matter-yield) than 'yield adjusted to market-standard moisture content' (perhaps http://ld.example.org/yield-standard-moisture-content). Third, information about 'things' is given in the form of triples, basically simple sentences of the form <thing1 > <thing2 > <thing3 > , where the meaning of each part of a sentence can be looked up. If we take 'ex:' as shorthand for http://ld.example.org/, we can for example create the following triples:

ex:my-measurement	ex:has-quantity	ex:yield-dry-matter-basis
ex:my-measurement	ex:has-units	ex:mg-per-ha
ex:my-measurement	ex:has-quantitative-value	10.0::double precision

Measurements expressed using Linked Data technology can be combined without manual intervention, regardless of where they were collected or where they were stored, as long as they are described using the same concepts (or when a mapping exists between concepts). This highlights the importance of shared vocabularies or ontologies. A number of ontology development efforts are under way. Of particular interest to the domain of weed control and crop protection are the Global Agricultural Concept Scheme (GACS), which combines AGROVOC, the CAB Thesaurus and the NAL Thesaurus into one ontology (http://tester-os-kktest.lib.helsinki.fi/gacsdemo/gacs/en/), the Plant Ontology (Jaiswal *et al.*, 2005) and Crop Ontology (Shrestha *et al.*, 2010). Unfortunately, anyone trying to use these ontologies will quickly find that many concepts are not yet included, which limits their immediate usefulness.

The advent of Linked Data has led to the development of databases that are optimised to store triples. Examples are RDF4J (http://rdf4j.org) and Virtuoso (http://virtuoso.openlinksw.com). Tools such as D2RQ (http://www.d2rq.org) offer the capability to access relational databases as if they contained triples.

Ownership and sharing of data

Big Data applications typically involve several data owners. For research data, the issue of ownership, archiving and sharing has received ample attention (King, 2007; White & Van Evert, 2008). The consensus is that the scientific method calls for sharing data liberally, although care should be taken to respect concerns such as privacy of people, the need to protect rare species and habitats by withholding details about location, and the need to publish before sharing (Duke & Porter, 2013).

Sharing research data is, of course, not the same as sharing data from commercial farms. Tellingly, a survey of Danish and US farmers showed that many are even reluctant to use cloud-based storage (Fountas *et al*., 2015). However, a decrease in public funding for agricultural research in recent years has resulted in fewer scientific experiments in agricultural sciences. This is at a time when increasingly there is a need for long-term experiments (LTEs) to investigate issues such as climate change, where the effects can be expected to become visible over a long time horizon (White & Van Evert, 2008). The scarcity of new experiments also reveals the need of extensive exploitation of already available data to develop efficient integrated weed management that benefits farmers and the environment. Intensively monitored farms may be the LTEs of the future. When on-farm collected data becomes an important vehicle for scientific progress, some of the arguments that apply to sharing scientific data become applicable to sharing farm data. The discussion about sharing farm data that does not proceed beyond the obligation of scientists to make research data available does not do justice to the topic. A framework to discuss ownership and sharing is needed.

Ethics is the branch of philosophy that examines the rights and duties of people in a systematic way. It seeks to answer questions such as 'what is the right way to act'. Ethics has no ready-made answer for our specific question whether farmers should share production data and with whom. Here, we hypothesise that ethical reasoning can help to structure the argument and can thus contribute to finding a resolution that is acceptable to parties involved. We will focus on the ethical matrix which was proposed by Mepham (2005), following work by (Beauchamp & Childress, 2001).

The ethical matrix has two dimensions. The first (column) dimension consists of the three broad categories in which Mepham (2005) summarises the major ethical theories. These categories are *well-being* (related to utilitarianism: the greatest benefit to the largest number of people), *autonomy* (related to deontology: do as you would be done by) and *fairness*. The second (rows) dimension describes the parties that are affected by the issue at hand. In our case, the parties with ethical standing are farmers, owners of agribusinesses, consumers and the living environment (biota).

The ethical matrix is used to record concerns that exist about a new situation that is envisaged. In our case, that situation is 'data collected on commercial farms is shared'. Each concern about this situation is entered in the cell of the ethical matrix that is at the cross between the party affected and the category of the concern. A possible listing of concerns that is about sharing farm data is shown in Table 2.

Big Data analytics

Once the relevant data have been properly prepared and stored, knowledge valuable to users is extracted through data analytics. Conventionally, agricultural applications use standard statistical methods, such as regression, analysis of variance (ANOVA) and principal component analysis (PCA). Big Data applications require new methods. First, standard statistics may be inadequate to deal with the large number of variables typically found in Big Data applications, and these variables may be related in a complex, non-linear manner. Second, even the implementation of simple methods is not straightforward when extremely large data sets are involved. In other words, devising and implementing a numerically efficient 'Big Data PCA' is a non-trivial task (Balcan *et al*., 2014). At least two steps must be considered: adopting an appropriate machine learning model (e.g. a neural network), and secondly training the model using an appropriate algorithm (e.g. a gradient descent method). A third step consists of measures to ensure privacy, which is of high relevance in agriculture.

Machine learning models

The goal of a machine learning task is to learn the relation between input and output, given a set of training data. For example, given training data (X_i, Y_i) where $i = 1,\ldots,n$, and where a pair (X_i, Y_i) represents measured environmental parameters and the yield for a certain past season i (Brdar *et al*., 2011), the goal is to learn the function f, $Y = f(X)$, which fits best (in a certain sense) the available training data. A possible approach is to find f which minimises the average squared error loss:

$$\sum_i |Y_i - f(X_i)|^2,$$

but many other forms of losses are also possible. For computational tractability, one needs to restrict f to a certain class of functions (e.g. polynomials of order at most m), such that the above minimisation is feasible. Generally, the choice of the loss function and the admissible function class determine different machine learning approaches or models.

Three machine learning models are widely used and relevant in agricultural applications (Kastens & Featherstone, 1996; Baral *et al*., 2011; Brdar *et al*., 2011; Rahaman *et al*., 2015; Agrimetrics, 2016) and Big Data (Davies & Frigola, 2014; Hsieh *et al*., 2014; Najafabadi *et al*., 2015). These are first, neural networks (NNs, see also the related concept of deep learning (Najafabadi *et al*., 2015)), second, (nonlinear)

Table 2 Ethical matrix applied to the envisaged situation: 'data collected on commercial farms are shared'

	Well-being	Autonomy	Fairness
Farmers	Income Liability	Choose what to share hoose how to farm	Receive payment or other benefit for data as compensation for shift in economic power No backlash from government access to data
Agribusiness owners		Innovate with data-based methods	Equitable trading of data
Consumers	Safe food, high quality	Traceability for informed purchasing decisions	Sufficient, affordable food
Biota	Conservation	Biodiversity	Sustainability

Each row lists concerns that pertain to a stakeholder group; concerns are grouped by the three broad categories of Mepham (2005). For each concern identified, an attempt is made to determine how it will be affected by the envisaged new situation. For *farmers*, income is a direct measure of farmers' well-being. Sharing data with scientists will lead to new scientific insights that will in turn allow farmers to improve profitability and sustainability of their business. On the other hand, sharing data with businesses may increase the economic power of those businesses and compromise the ability of farmers to sell at attractive prices. A farmer may risk liability suits, for example when records show that equipment malfunctioned and (unintended) contamination of the environment occurred. The autonomy of a farmer could be compromised when he or she loses control over the flow of data. Also, the farmer's sense of identity may be compromised when critical farming decisions are made by consultants or decision-support software. On the other hand, new insights resulting from sharing data with scientists may provide the farmer with more options to manage the farm and to make better decisions. Fairness requires that a farmer be compensated for the value that others derive from the data he or she shares. Farmers are concerned that governments may use farm data to argue for stricter controls on, for example, emissions of nutrients and chemicals. *Agribusinesses* need access to farm data to generate income. For developing innovative data-based services (autonomy), agribusiness is dependent on the ability to access farm data. Businesses require an equitable regulatory framework (fairness) with respect to acquiring, storing, transferring and using farm data. *Consumers* will benefit from an increase in safety and quality of food made possible by new insights when farm data are shared with scientists. Tracking and tracing the origin of products and the agricultural practices used to produce them with special focus on the pesticides used (and especially residual herbicides) will allow consumers to make informed purchasing decisions (autonomy). New insights derived from sharing farm data with scientists will lead to an increase in the availability of food and increase fairness for consumers. Finally, *wildlife and the living environment* will benefit (fairness) when sharing data lead to new scientific knowledge and less pollution (well-being), preservation of biodiversity and native species, and when threatened species and rare breeds are preserved (autonomy).

support vector machines (SVMs) with kernels and third, graphical models (*GMs*). Other relevant models include (group)-sparsity and other structured models (Slavakis *et al.*, 2014), models involving spatial data (Vatsavai *et al.*, 2012) and linear and non-linear dimensionality reduction and clustering methods (Kashyap *et al.*, 2015).

Neural networks (NNs) have proved successful in speech recognition and image and natural language processing (Xie *et al.*, 2014). Their name indicates a resemblance in structure to actual, biological neural networks. Namely, with NNs, function f is modelled as a functional composition of basic computational elements, for example. neurons, where each neuron consists of a linear activation function, parameterised by a weight vector w and a non-linear transfer function s (e.g. a sigmoid function (Bishop, 2006; Hinton *et al.*, 2006)). The neurons are organised in layers (the number of layers is the depth of a NN), each of which has a certain width. Common loss functions are squared error and cross-entropy loss, and a popular numerical algorithm for training NNs is back propagation and its variants (Bishop, 2006; Hinton *et al.*, 2006).

Neural networks have been used in many agricultural use cases, including prediction of yield (Baral *et al.*, 2011) (tactical decision), farmers' risk preferences (Kastens & Featherstone, 1996) (strategic or tactical), and site-specific herbicide management (SSHM) (Eddy *et al.*, 2008) (operational). With advances in imaging technology and computer processing speed, methods like the one proposed by Eddy *et al.* (2008) seem promising for real-time detection and mapping of weed species for SSHM in agriculture.

Support vector machines (SVMs) with kernels – Given a training data set, SVMs seek a function f which makes at each data point an error of at most ε, where ε is a predefined small positive number, as explained by Smola and Schölkopf (2004). SVMs were initially proposed for linear models. A non-linear version can be made by first transforming input X into a (higher dimensional) feature space through a non-linear mapping $\Phi(X)$ and then applying standard linear SVMs over features $\Phi(X)$. This can be done without ever explicitly calculating features $\Phi(X)$; thus, there is no need to work directly in the (usually very high dimensional) feature space. Namely, function f can be expressed as a linear combination of inner products with data points X_i's. That is, it is only needed to define (and subsequently compute) a Kernel function $K(x_1,x_2)$, which defines the inner products $<\Phi(x_1)$,

$\Phi(x_2)>$ in some feature space. Many easily computable functions (e.g. polynomial kernels, exponential kernels) turn out to be valid kernel functions (valid inner products in some feature space), which makes SVMs very efficient in practice.

Support vector machines have been widely used in agricultural applications. Examples include prediction of agricultural yields based on relevant environmental parameters (Brdar *et al.*, 2011) (tactical decisions) and detection and classification of plant diseases (Rumpf *et al.*, 2010) (operational decisions).

Graphical models (GMs) In a standard setting, *GMs* are not concerned with input–output modelling, but rather they model interdependencies among a set of (input) variables (Jordan, 2004; Wainwright & Jordan, 2008). Variants exist that model input–output relations also, such as conditional random fields (CRFs) (Lafferty *et al.*, 2001; Wytock & Kolter, 2013). Differently from NNs and SVMs, CRFs adopt a probabilistic framework; that is, they model input–output relations through a (conditional) probability distribution $P(Y \mid X)$, rather than an explicit function of the form $Y = f(X)$. However, the underlying principle is similar: given a training data set (X_i, Y_i), $i = 1,\ldots,n$, one again learns the probability distribution $P(Y \mid X)$ from a certain class of distributions (e.g. Gaussian), by minimising an appropriately defined loss function. Once the 'best' distribution $P(Y \mid X)$ is learned (the training has been completed), one can perform inference (e.g. predict a new output Y_i based on a given new input X_i) by finding a maximum *a posteriori* estimate of Y_i, that is by finding Y_i which minimises $P(Y \mid X_i)$ viewed as a function of Y.

With CRFs and with *GMs* in general, the key object of the (joint) probability distribution of interest is associated with a graph, whose nodes are the individual variables X_i's and Y_j's. Then, the probability distribution is defined as a product of factors associated with graph cliques (all-to-all connected subsets of nodes). The graph structure allows for natural modelling of complex phenomena in applied fields, as noted in Jordan (2004), for example for modelling spatial or spatio-temporal processes (Vatsavai *et al.*, 2012). Introducing the graph formalism and structure with *GMs* turns out to be quite useful in constructing efficient numerical algorithms to perform inference. The so-termed belief propagation-type algorithms. (e.g. Wainwright & Jordan, 2008) and their various approximations (e.g. Donoho *et al.*, 2009) are very frequently used methods for inference over *GMs*.

Graphical models have significant potential for modelling in Big Data agricultural applications. A variant of *GMs*, spatial random fields, can be used to model various spatial phenomena, such as the location prediction problem, for example prediction of spatial distribution of a disease within a field (Vatsavai *et al.*, 2012). This may correspond with tactical or operational decisions. *GMs* are also used for modelling traits' interdependencies with large scale phenotyping (Rahaman *et al.*, 2015; Agrimetrics, 2016).

Missing data and variable spatio-temporal resolutions are two specific challenges that arise with machine learning in agriculture. There are many situations where certain planned data entries are missing (e.g. a sensor malfunctioned, or cloud cover prevented acquisition of satellite imagery). It is thus necessary to have machine learning models that can cope with missing data; see for example Slavakis *et al.* (2014) for a class of such robust models. When confronted with a lack of data on lifecycle stages during the development of a model to predict the spread of invasive weeds, Kueffer *et al.* (2013) addressed this problem using life cycle data of naturally established individuals to improve the accuracy of predictions about the distribution range of the invasive weeds (Ramírez-Albores *et al.*, 2016). Agricultural data often have variable resolution in time and space (Table 1). Models are needed which can effectively treat such data (Klein *et al.*, 2015).

Numerical algorithms: parallel and distributed optimisation

Once the prepared data are ready for processing, and an appropriate machine learning model with parameter set w has been adopted, a numerical algorithm is used to produce the set of parameters w^* which best explain the data. Usually, this task is performed by solving an optimisation problem, namely that of minimising an appropriately defined loss function (e.g. a squared loss) with respect to model parameters w:

$$\text{minimise } f(w; D),$$

parameterised by the available data $D = \{(X_i, Y_i), i = 1,\ldots,n\}$. For example, with classification tasks, the loss function f can be logistics or hinge loss (Bishop, 2006). When function f is convex, this optimisation problem can in principle be efficiently solved by standard numerical optimisation methods (e.g. gradient descent or Newton method.) However, in Big Data applications, a major challenge is that the size of the data set D and possibly the dimension of the unknown parameter set w are so large that the problem cannot be solved in a reasonable time with standard numerical optimisation methods on a single standard computer. Therefore, there is a need to develop *parallel and distributed* optimisation methods that partition the problem of interest into multiple smaller problems, each of

which is solved by a separate processor (Jakovetić *et al.*, 2014; Slavakis *et al.*, 2014). There are now parallel and distributed methods that can solve huge problems. For example, a (convex) logistics loss problem with an order of 70 000 data points of size 20 000 (real numbers) was solved in less than 10 s using 40 parallel processes (Facchinei *et al.*, 2015).

Several challenges arise when designing Big Data algorithms. The first challenge, scalability, refers to how computational time reduces when the number of processors is increased. A naive consideration would imply that the time decreases linearly with the number of processors. However, the delays due to interprocessor communications cause more complicated (and less efficient) scaling (Hong *et al.*, 2015). A second challenge is that in many agricultural applications, analytics should be able to respond in real time to changes in the sensed data (e.g. weed emergence, weather changes, plant stress). This is also true in the case of evaluation of herbicide-resistant weeds, where a rapid decision and response are required (Travlos, 2013a). That is, algorithms should be able to quickly adapt their solutions based on the changes in the incoming streaming data. This can be, in many cases, accomplished through *online learning* and *stochastic optimisation algorithms* (e.g. stochastic gradient descent) (Duchi *et al.*, 2011). Essentially, such methods allow for computationally inexpensive solution updates (e.g. a gradient descent step) accounting for only newly acquired data samples, as opposed to revisiting all the data samples at each algorithm iteration. A third challenge, privacy preservation, is discussed in more detail in the next section on *Data analytics under privacy constraints*.

Several commercial and open source Big Data software libraries and platforms exist. Apache Hadoop (Apache, 2016) includes several modules: (i) Hadoop Distributed File System (HDFS) – a distributed file system for high-throughput access to application data, (ii) YARN – a framework for job scheduling and computer cluster resource management, (iii) MapReduce – a system for parallel processing of large data sets, (iv) Mahout – a machine learning tools library and (v) Spark – a compute engine and a programming model which supports a wide class of tasks, including machine learning, stream processing and graph computation. GraphLab (Low *et al.*, 2010) is a framework for developing efficient and provably correct parallel machine learning algorithms, very expressive for asynchronous iterative algorithms with sparse computational dependencies. Package pbdR (Ostrouchov *et al.*, 2012) is a software package for Big Data based on the R programming language. PAIRS is a platform specifically designed for handling geo-spatial data which has been used for agriculture-related Big Data (Klein *et al.*, 2015).

Data analytics under privacy constraints

A very important issue with data analytics in agricultural applications where multiple parties (e.g. farmers) are involved is that of *data privacy*, as farmers may not be willing to disclose or share their private data or practices. On the other hand, exploring hidden knowledge from all parties' data can clearly yield improved solutions with respect to the solutions based on parties' individual data sets.

There have been significant advances in *privacy-aware data analytics* (Dreier & Kerschbaum, 2011; Fung & Mangasarian, 2013; Sarwate & Chaudhuri, 2013; Yan *et al.*, 2013; Duchi *et al.*, 2014; Weeraddana *et al.*, 2014; Xie *et al.*, 2014; Nozari *et al.*, 2016). However, a significant amount of research and development is still needed to devise tools which simultaneously ensure the following: 1) very high degrees of privacy, 2) handling huge data volumes and 3) handling very generic machine learning models or tasks.

We consider the following conceptual system model (Fig. 3). There is a group of N parties, for example farmers, each holding its own private data D_i. Parties outsource *messages* m_i (D_i) related to their private data to an *analytics provider*. (One can think of m_i (D_i) as a 'disguised' version of D_i.) Subsequently, the provider processes messages from *all parties* and sends the obtained *result* to all parties. One can think of this result as a 'disguised' version of the optimal solution that a (hypothetical) provider would compute, if it had available all data D_i's and from which each party can reconstruct this optimum. We assume that there is also an *adversary* who, based on the observed messages (and possibly the observed result), attempts to recover the parties' private data. The goal is to design messages and the provider's analytics, such that the method is computationally feasible and the parties can reconstruct the optimum from the result, while the adversary cannot (or is at least unlikely) to discover private data D_i's.

Existing works to solve the described problem can be broadly categorised into two classes (Weeraddana *et al.*, 2014): (i) *cryptography-based approaches* and (ii) *non-cryptography-based* approaches.

Cryptography-based approaches – With this class, each party creates message m_i (D_i) as an actual, classical encryption (e.g. homomorphic encryption (Xie *et al.*, 2014)) of D_i with a privately known key. Subsequently, analytics is performed *over the encrypted data* (for example, via secure multiparty computation (Xie *et al.*, 2014)) and the encrypted result is sent back to

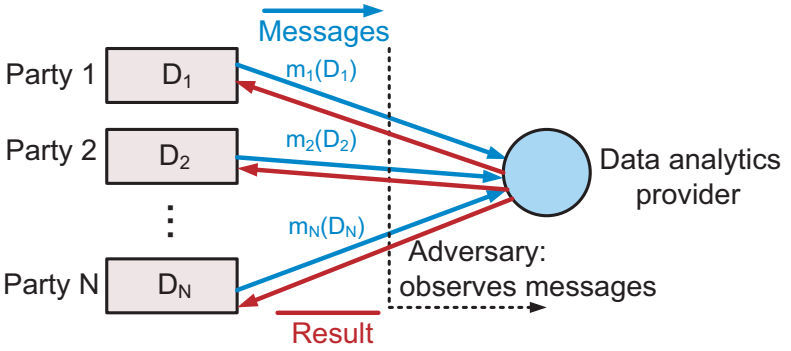

Fig. 3 Conceptual system model underlying privacy-aware analytics.

all parties, which then decrypt the result. This approach obviously allows for a 'perfect privacy', in the sense that the adversary cannot reconstruct private data (without having the parties' private keys) in a feasible amount of time. The price of ensuring 'perfect privacy' is the large amount of computer time needed to generate solutions. The reason is that each non-encrypted bit of information corresponds to a large sequence (perhaps a thousand) encrypted bits, and hence, any arithmetic operation over the encrypted symbols is much costlier than the equivalent operation in the standard, non-encrypted domain. Currently, for many machine learning models, cryptography-based approaches are computationally unfeasible. As an illustration, solving a linear program (LP) with 282 unknowns and 180 constraints (a moderate size problem) with current cryptography-based solutions as of 2011 is estimated to take seven years (Dreier & Kerschbaum, 2011). However, for relatively simple models and tasks, cryptography-based approaches might be good solutions. For example, Xie *et al.* (2014) propose a cryptography-based system for neural networks and conjectures their practical feasibility for inference tasks (e.g. performing classification of a data point for an already trained neural network) and perhaps also learning tasks for simple models (e.g. training a moderate size neural network with a small number of layers).

Non-cryptographic approaches achieve some degree of privacy through *algebraic data transformations*. That is, each message m_i (D_i) represents some (deterministic or random) algebraic transformation of D_i. Naturally, such transformations are sought to be 'non-invertible', in the sense that the adversary cannot (or is very unlikely) to recover private data by observing the messages. For example, Dreier and Kerschbaum (2011) and Fung and Mangasarian (2013) address solving LPs, where each party multiplies its own real-valued private data vector D_i by a *privately generated random matrix*. The main advantage of the non-cryptography-based solutions with respect to the cryptography-based ones is that data analytics is performed *directly over the transformed, but non-encrypted data* (real numbers, vectors and matrices). Hence, they do not introduce huge computational overheads of performing algebraic operations over encrypted sequences. However, this in general comes at the cost of a certain *information leakage, that is,* no 'perfect privacy' is ensured. Currently, for models like LPs, there exist efficient methods to solve moderate sized problems with a very low information leakage (Dreier & Kerschbaum, 2011). As an illustration, an LP with 282 unknowns and 180 constraints can be solved within 25 min (compared with the 7 years' execution time of cryptography-based approaches), while ensuring that the adversary can guess the problem solution with the chance lower than 10^{-1408} (Dreier & Kerschbaum, 2011). Further research is needed to devise methods which handle very general models and huge data scales.

Besides the described 'algebraic transformation' approaches, there are non-cryptographic approaches like the methods based on the notion of *ε-differential privacy* (Sarwate & Chaudhuri, 2013; Duchi *et al.*, 2014; Nozari *et al.*, 2016). Further, other more elaborated models than the one considered in Fig. 3 are certainly relevant and have been studied. For instance, parties can be arranged in a network (e.g. induced by their geographical proximity or by their business relations), where each party itself possesses its own data analytics (computing) resources, that is not only the data but also the actual analytics algorithm is distributed over the N parties. The parties then collaboratively solve the common learning task through exchanging messages along links in the network. The adversary can observe messages from all (or a subset of) links (Yan *et al.*, 2013; Nozari *et al.*, 2016). In such setups, interestingly, some standard, 'general-purpose' iterative distributed methods, like the alternating direction method of multipliers, exhibit certain privacy-enabling properties (Weeraddana *et al.*, 2014).

Delivery of actionable information to farmers

The goal of Big Data analytics in weed control and crop protection is to provide actionable information

for better management decisions by farmers and their agri-food partners. The information can be used in strategic, tactical and operational decisions, and at different spatial scales on the farm (fields, management zones or grids within field, individual plants, for example Christensen *et al.* (2009)), or at regional or agri-food chain level. But no matter what type of decision is made or what the scale of application is, digital information must be presented to farmers in a straightforward and comprehensible way.

The simple computerised record-keeping solutions of early years have evolved into comprehensive Farm Management Information Systems (FMIS). Sørensen *et al.* (2010) defined an FMIS as a planned system for collecting, processing, storing and disseminating data in the form needed to carry out a farm's operations and functions. Essential FMIS components include specific farmer-oriented designs, dedicated user interfaces, automated data processing functions, expert knowledge and user preferences, standardised data communication and scalability. It has been stressed that the evolution of FMIS must take into account the social aspects of business processes (Fountas *et al.*, 2015).

There is not always a smooth path to commercial availability, even for systems that have already shown their potential in a research setting. In the Netherlands alone, several commercial initiatives to develop geo-information system (GIS) platforms for use in agriculture have failed during the last 10–20 years. However, a system called 'Akkerweb' (in English: Farm Maps; see www.akkerweb.nl) is currently gaining traction. Akkerweb is the product of a public–private partnership between Agrifirm, the largest farmers' cooperative in the Netherlands, and Wageningen University & Research, the leading agricultural research organisation in the Netherlands. Akkerweb allows geo-data acquisition, management, visualisation and use at the farm level, in combination with a standard FMIS (Kempenaar *et al.*, 2014b, 2016). The roots of Akkerweb can be traced to the development in 2012 of a decision-support system for control of plant parasitic nematodes NemaDecide (Been *et al.*, 2004, 2007). Akkerweb offers GIS functionality and a number of general free for use applications ('apps'), such as a cropping scheme app, a satellite data app and a sensor data app, to visualise and analyse soil and crop data and to generate task (prescription) maps. Akkerweb also contains several subscription-based apps for variable rate application of pesticides and fertilisers. The success of Akkerweb is due to the combination of its ICT infrastructure and its science-based content, the bottom-up development with users in the driver's seat, and the effective cooperation between a farmers' cooperative, a research institute and an IT company with the know-how to build and maintain the required software. Akkerweb is an open platform, in the sense that third parties can also use the Akkerweb platform to develop and offer fee-based services. Today, data of ca. 30 000 parcel crop years are stored using Akkerweb.

Current applications of Big Data for weed control and crop protection

It is useful to consider strategic, tactical and operational decisions separately and outline the specificities for each of these.

Strategic decisions

NemaDecide is a system to support strategic decisions on the control of plant pathogenic nematodes (Been *et al.*, 2004, 2007). The system is based on a model of the population dynamics of nematodes and takes into account the presence of host plants, specific crop rotations, soil analysis data and efficacy of control methods. GeoNema (Haverkort & Kempenaar, 2016) is the NemaDecide decision-support system in a GIS platform accessible via Akkerweb. Farmers can apply soil analysis data from laboratories in combination with the decision support, to decide on optimal crop rotations and control strategy, for example to make a task map for site-specific control.

In weed and disease control, the use of population model information to make strategic decisions in crop rotation management is less advanced. A DSS on weed control would be especially useful if it can contribute to effective weed control methods that minimise the development of herbicide-resistant weeds. Decision support might also be given in the form of predicting which mix of cultivars of certain crops is most likely to maximise yield and minimise risk (Marko *et al.*, 2016). In this study, the effect of weeds, pests and diseases on yield was not explicitly considered, but this could be included.

The data that need to be collected in order to support strategic decisions on weed control include the occurrence of weeds (kind or species, density), combined with soil data, management, crop yield and, of course, weather. These data can be used to derive risk factors for weeds and to determine how effective weed control measures are. Several model-based decision-support systems for weed management in arable crops are already available, taking advantage of the tank of Big Data (Berti *et al.*, 2003; Rydahl, 2004; Parsons *et al.*, 2009).

The algorithms that will be useful to support strategic decisions in crop protection include Bayesian

parameter estimation methods which can be used to (better) parameterise population dynamics models, possibly using inverse modelling. It may also be possible to use NNs to develop non-mechanistic models of cause and effect, especially in the case of many factors with weak influence, such as soil pH, soil organic matter content, CEC and soil texture.

Tactical decisions

In several countries, decision support for weed control is available in the form of recommendations for herbicide selection, herbicide rate and time of application, for example Crop Protection Online (Murali *et al.*, 1999; see also https://plantevaernonline.dlbr.dk/cp/documents/InfoFactSheet2.pdf) in Denmark and Gewis (http://www.agrovision.nl/sectoren/teelt/producten_voor_de_teler/crop/gewis/) in the Netherlands. These recommendations are based on data on weed species, crop sensitivity and climatic conditions. These kinds of systems can also be used to optimise the application of fungicides and insecticides, whether or not in combination with early warning systems for infection of crops by diseases. Attempts are now being made to put these decision-support systems in GIS-platforms. This is the case for variable rate application of soil herbicides. The Akkerweb recommendation for variable rate application of soil herbicides uses data on spatial variation in soil organic matter, CEC and pH, in combination with data on soil moisture, sensitivity of the crop to the herbicide, climate conditions, FMIS data and weed maps, in order to make task maps for variable rate application taking into account the relevant spatial variation. This has resulted in a reduction in herbicide use of 10–20% compared with uniform treatment of the field when applied at a resolution of 10–30 m^2 grids (Kempenaar *et al.*, 2014a).The power of this method can also be illustrated with an effort in which data on the application of fertiliser N and resulting maize yields are pooled across experiment sites and years. This has led to more specific and better fertiliser recommendations (Sawyer, 2010).

The data that are needed for Big Data supported decisions in weed control include spatial information on weed occurrence, landscape position, soil characteristics and weather. Data on weed occurrence are traditionally collected on an intermittent basis in experiments (Gerhards *et al.*, 1996). Nowadays, this information can be obtained on a large scale by logging the occurrence of weeds as they are detected by weed-detection software fed by cameras on robots or on spray booms. Alternative methods include using a camera mounted on an unmanned aerial vehicle (Perez-Ortiz *et al.*, 2016). Landscape position can be obtained from a Digital Elevation Model (DEM) on the basis of the logged position.

Useful algorithms include probabilistic reasoning to estimate parameters of population dynamics models. Also useful will be neural networks (NN) to link cause and effect where insufficient knowledge about underlying mechanisms is available.

Operational decision

Autonomous robotic weed control depends on accurate information on the position and determination of weeds in crops. Although much scientific progress has been made in this area (e.g. Bakker *et al.*, 2006, 2010), certainly in the field of algorithms for weed-detection (Eddy *et al.*, 2008), commercial use is still limited (Merfield, 2016). The aim of these kinds of efforts is illustrated by a prototype weed control robot that is truly autonomous (Van Evert *et al.*, 2011). In this robot, weed detection is combined with an autonomous platform and a mechanical weed control device that uses the information to destroy the tap root of *Rumex obtusifolius* L. (broad-leaved dock) in grasslands with high accuracy.

The most useful data are also the hardest to obtain: labelled images. Labelling can be performed by outlining the weed or by simply noting whether a weed is present or not. Typically, labelling is performed by humans and is extremely time-consuming. For operational decisions, the algorithms that are most useful for classification are SVM and NN.

The cases above illustrate how data can be used to obtain actionable information for weed control and crop protection. We expect this will grow in the future when more data layers, models and data analytics become available. The model parts, either statistical models or agronomic models, will become better when farmers share data to better estimate the parameters of the models.

Conclusions and recommendations

In this paper it was argued that a new conceptual model for weed control and crop protection should be developed which consists of three elements: (i) capture and store data, (ii) analyse data and (iii) generate recommendations. We put forward the view that integrated solutions for weed control and crop protection are needed. Such integrated solutions require simultaneous advances in agricultural science, in ICT, in collaboration between supply chain partners (co-innovation), in respecting the interests of all parties involved, and in legal frameworks. In the area of

science, new knowledge is needed which will allow us to use historical data to predict the occurrence (time, location, severity) of weeds, pests and diseases. Research is needed on the interaction between real-time data collection on weed occurrence, soil and climatic conditions during the growing seasons. These data should be the basis for building models for the physiology and behaviour of weeds at given climatic conditions, which should be organised in a systematic way and use the appropriate Big Data analytics to deliver the best decisions. Technical advances are needed to allow us to integrate data from various sources. Here, the most likely avenue to success is through semantic technologies and the most pressing need is for appropriate ontologies to be developed. Any integrated solution will require the collaboration of supply chain partners, even if they are many, and even if they are commercial competitors. In the Netherlands and some other countries, farmers' cooperatives play an important role in establishing effective working relationships between supply chain partners. This example may need to be emulated by farmers elsewhere, and indeed by the many enterprises, large and small, that are offering services. We have made the case that safeguarding the interests of all partners will be helpful in establishing successful collaboration. In case of conflicting interests, ethical reasoning may help to reach understanding between parties. Data sharing protocols may need to be developed that can be used as templates in commonly occurring situations. Agreements between parties must be formalised in legally binding contracts and national and international law must be in place to support this. Creating protocols and reaching agreements ultimately is based on trust; this trust has to be earned by the parties that want to be involved.

In recent years, significant advances have been made in developing general-purpose tools and methods for Big Data capture, storage and analysis, as well as some emerging customised systems and applications in the agri-food domain. An interdisciplinary effort is required to overcome remaining challenges and fully realise Big Data opportunities in agriculture. In the case of weeds, many opportunities may arise, especially for invasive, parasitic or herbicide-resistant weeds. This effort requires the involvement of agricultural experts, of computer and data science experts, as well as advances in terms of organisational, ethical and legal arrangements.

Acknowledgements

FKVE has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 664388. DJ is supported by the Serbian Ministry of Education, Science, and Technological Development, Grant no. 174030. CK gratefully acknowledges financial contributions of the Dutch Ministry of Economic Affairs and associated local governments and companies through Topsector Agri & Food.

References

Agrimetrics (2016) Agrimetrics. Available at: http://www.agrimetrics.co.uk/ (last accessed 18 April 2017).

Allemang D & Hendler J (2011) *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier, Amsterdam, The Netherlands.

Apache (2016) Apache Hadoop. Available at: https://hadoop.apache.org/ (last accessed 18 April 2017).

Bakker T, Van Asselt CJ, Bontsema J, Müller J & Van Straten G (2006) An autonomous weeding robot for organic farming. In: *5th International Conference on Field and Service Robotics*, Vol. 25. Springer, Port Douglas, Australia.

Bakker T, van Asselt K, Bontsema J, Muller J & van Straten G (2010) Systematic design of an autonomous platform for robotic weeding. *Journal of Terramechanics* **47**, 63–73.

Balcan M-F, Kanchanapally V, Liang Y & Woodruff D (2014) Improved Distributed Principal Component Analysis. Available at: arxiv.org/abs/1408.5823. (last accessed 18 April 2017).

Baral S, Kumar Tripathy A & Bijayasingh P (2011) Yield prediction using artificial neural networks. In: *Computer Networks and Information Technologies: Second International Conference on Advances in Communication, Network, and Computing, CNC 2011, Bangalore, India, March 10–11, 2011. Proceedings*. (eds VV Das, J Stephen & Y Chaba), 315–317. Springer, Berlin, Heidelberg.

Beauchamp TL & Childress JF (2001) *Principles of biomedical ethics*. Oxford University Press, USA.

Been TH, Schomaker CH & Molendijk LPG (2004) NemaDecide: a decision support system for the management of potato cyst nematodes. In: Decision support systems in potato production. Symposium proceedings Potato modelling conference, Edinburgh, March 2003. Wageningen (eds DKL Mackerron & AJ Haverkort). Academic Press, Wageningen.

Been TH, Schomaker CH & Molendijk LPG (2007) NemaDecide, a decision support system for the management of potato cyst nematodes. *Phytopathology* **97**, S152.

Berners-Lee T, Hendler J & Lassila O (2001) The semantic web. *Scientific American* **248**, 28–37.

Berti A, Bravin F & Zanin G (2003) Application of decision-support software for postemergence weed control. *Weed Science* **51**, 618–627.

Bishop CM (2006) *Pattern recognition and machine learning*. Springer, New York.

Blackmore S, Godwin RJ & Fountas S (2003) The analysis of spatial and temporal trends in yield map data over six years. *Biosystems Engineering* **84**, 455–466.

Brdar S, Ćulibrk D, Marinković B, Crnobarac J & Crnojević V (2011) Support vector machines with features contribution analysis for agricultural yield prediction. In: *Proceedings of the Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment* (EcoSense 2011), 43–47. Belgrade.

Chang F, Dean J, Ghemawat S *et al*. (2008) Bigtable: a distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)* **26**, 4.

Christensen S, Sogaard HT, Kudsk P *et al*. (2009) Site-specific weed control technologies. *Weed Research* **49**, 233–241.

Codd EF (1970) A relational model of data for large shared data banks. *Communications of the ACM* **13**, 377–387.

Davies A & Frigola R (2014) Probabilistic Models for Big Data. Available at: http://www.rogerfrigola.com/doc/bigdata.pdf. University of Cambridge. (last accessed 18 April 2017).

De Almeida J & Freitas H (2012) Exotic flora of continental Portugal–a new assessment. *Bocconea* **24**, 231–237.

De Mauro A, Greco M & Grimaldi M (2016) A formal definition of Big Data based on its essential features. *Library Review* **65**, 122–135.

Donoho DL, Maleki A & Montanari A (2009) Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **106**, 18914–18919.

Dreier J & Kerschbaum F (2011) Practical privacy-preserving multiparty linear programming based on problem transformation. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 916–924. IEEE.

Duchi J, Hazan E & Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12**, 2121–2159.

Duchi JC, Jordan MI & Wainwright MJ (2014) Privacy aware learning. *Journal of the Association for Computing Machinery* **61**, 38.

Duke CS & Porter JH (2013) The ethics of data sharing and reuse in biology. *BioScience* **63**, 483–489.

Early R, Bradley BA, Dukes JS *et al*. (2016) Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature Communications* **7**, 12485.

Eddy PR, Smith AM, Hill BD, Peddle DR, Coburn CA & Blackshaw RE (2008) Hybrid segmentation - Artificial Neural Network classification of high resolution hyperspectral imagery for Site-Specific Herbicide Management in agriculture. *Photogrammetric Engineering and Remote Sensing* **74**, 1249–1257.

Facchinei F, Scutari G & Sagratella S (2015) Parallel selective algorithms for big data optimization. *IEEE Transactions on Signal Processing* **63**, 1874–1889.

Foley JA (2011) Can we feed the world and sustain the planet? *Scientific American* **305**, 60–65.

Foley JA, Ramankutty N, Brauman KA *et al*. (2011) Solutions for a cultivated planet. *Nature* **478**, 337–342.

Fountas S, Wulfsohn D, Blackmore BS, Jacobsen HL & Pedersen SM (2006) A model of decision-making and information flows for information-intensive agriculture. *Agricultural Systems* **87**, 192–210.

Fountas S, Carli G, Sorensen CG *et al*. (2015) Farm management information systems: current situation and future perspectives. *Computers and Electronics in Agriculture* **115**, 40–50.

Fumanal B, Chauvel B & Bretagnolle F (2007) Estimation of pollen and seed production of common ragweed in France. *Annals of Agricultural and Environmental Medicine* **14**, 233–236.

Fung GM & Mangasarian OL (2013) Privacy-preserving linear and nonlinear approximation via linear programming. *Optimization Methods and Software* **28**, 207–216.

Gerber E, Schaffner U, Gassmann A, Hinz H, Seier M & MÜLLER-SCHÄRER H (2011) Prospects for biological control of *Ambrosia artemisiifolia* in Europe: learning from the past. *Weed Research* **51**, 559–573.

Gerhards R, Sokefeld M, Knuf D & Kuhbauch W (1996) Mapping and geostatistical analysis of weed distribution in sugarbeet fields for site-specific weed management. *Journal of Agronomy and Crop Science-Zeitschrift Fur Acker Und Pflanzenbau* **176**, 259–266.

Guisan A & Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* **135**, 147–186.

Haverkort AJ & Kempenaar C (2016) Recent advances in biotechnology and information technology in the potato industry. In: Proceedings Crop Protection in Northern Britain, 183–190.

Hinton GE, Osindero S & Teh Y-W (2006) A fast learning algorithm for deep belief nets. *Neural computation* **18**, 1527–1554.

Hong M, Razaviyayn M, Luo Z-Q & Pang J-S (2015) A Unified Algorithmic Framework for Block-Structured Optimization Involving Big Data." Available at: http://arxiv.org/abs/1511.02746 (last accessed 18 April 2017).

Hsieh C-J, Si S & Dhillon IS (2014) A divide-and-conquer solver for kernel support vector machines. *Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research (PMLR)* **32**(1), 566–574, 2014.

Hunt LA, Jones JW, Hoogenboom G *et al*. (1994) Input and output file structures for crop simulation models. In: *Crop Modeling and Related Environmental Data: A Focus on Applications for Arid and Semiarid Regions in Developing Countries*. (eds PF Uhlir & GC Carter), 35–72. CODATA Commission on Global Change.

Jaiswal P, Avraham S, Ilic K *et al*. (2005) Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comparative and Functional Genomics* **6** (7–8). doi: 10.1002/cfg.496.

Jakovetić D, Xavier J & Moura JM (2014) Fast distributed gradient methods. *IEEE Transactions on Automatic Control* **59**, 1131–1146.

Jordan MI (2004) Graphical models. *Statistical Science* **19**, 140–155.

Kashyap H, Ahmed HA, Hoque N, Roy S & Bhattacharyya DK (2015) Big data Analytics in Bioinformatics: A Machine Learning Perspective. Available at: https://arxiv.org/abs/1506.05101 (last accessed 18 April 2017).

Kastens TL & Featherstone AM (1996) Feedforward backpropagation neural networks in prediction of farmer

risk preferences. *American Journal of Agricultural Economics* **78**, 400–415.

Kempenaar C, Hejting S & Michielsen JM (2014a) Perspectives for site specific application of soil herbicides in arable farming. In: *12th International Conference on Precision Agriculture (ICPA)*, Sacramento, CA, USA.

Kempenaar C, Van Evert FK & Been T (2014b) Use of vegetation indices in variable rate application of potato haulm killing herbicides. In: *12th International Conference on Precision Agriculture (ICPA)*, Sacramento, CA, USA.

Kempenaar C, Van Evert FK, Been T, Kocks CG & Westerdijk CE (2016) Towards data-intensive, more sustainable farming: advances in predicting crop growth and use of variable rate technology in arable crops in the Netherlands. In: *13th International Conference on Precision Agriculture (ICPA)*, St. Louis, MO, USA.

King G (2007) An introduction to the Dataverse network as an infrastructure for data sharing.

Klein LJ, Marianno FJA, Albrecht CM, Freitag M & Hamann HF (2015) PAIRS: A scalable geo-spatial data analytics platform. In: *2015 IEEE Conference on Big Data 1290–1298*.

Kueffer C, Pyšek P & Richardson DM (2013) Integrative invasion science: model systems, multi-site studies, focused meta-analysis and invasion syndromes. *New Phytologist* **200**, 615–633.

Kutter T, Tiemann S, Siebert R & Fountas S (2011) The role of communication and co-operation in the adoption of precision farming. *Precision Agriculture* **12**, 2–17.

Lafferty J, McCallum A & Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the eighteenth international conference on machine learning (ICML)*. Vol. 1, 282–289.

Lawson LG, Pedersen SM, Sorensen CG *et al.* (2011) A four nation survey of farm information management and advanced farming systems: a descriptive analysis of survey responses. *Computers and Electronics in Agriculture* **77**, 7–20.

Low Y, Gonzalez J, Kyrola A, Bickson D, Guestrin C & Hellerstein JM (2010) GraphLab: A New Parallel Framework for Machine Learning. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. Available at https://arxiv.org/abs/1408.2041 (last accessed 18 April 2017).

Luo J & Cardina J (2012) Germination patterns and implications for invasiveness in three *Taraxacum* (Asteraceae) species. *Weed Research* **52**, 112–121.

Marko O, Brdar S, Panic M, Lugonja P & Crnojevic V (2016) Soybean varieties portfolio optimisation based on yield prediction. *Computers and Electronics in Agriculture* **127**, 467–474.

Marz N & Warren J (2015) *Big Data: principles and best practices of scalable realtime data systems*. Manning Publications, Greenwich, CT, USA.

Mepham B (2005) *Bioethics: an introduction for the biosciences*. Oxford University Press, Oxford, UK.

Merfield CN (2016) Robotic weeding's false dawn? Ten requirements for fully autonomous mechanical weed management *Weed Research* **56**, 340–344.

Murali NS, Secher BJ, Rydahl P & Andreasen FM (1999) Application of information technology in plant protection in Denmark: from vision to reality. *Computers and Electronics in Agriculture* **22**, 109–115.

Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R & Muharemagic E (2015) Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2**, 1.

Nozari E, Tallapragada P & Cortés J (2016) Differentially private distributed convex optimization via objective perturbation. In: *American Control Conference (ACC)*, Boston, MA, USA.

Oerke E-C & Dehne H-W (2004) Safeguarding production—losses in major crops and the role of crop protection. *Crop Protection* **23**, 275–285.

Ostrouchov G, Chen W-C, Schmidt D & Patel P (2012) pbdR –Programming with Big Data in R. Available at: http://r-pbd.org/ (last accessed 18 April 2017).

Parsons DJ, Benjamin L, Clarke J *et al.* (2009) Weed Manager—a model-based decision support system for weed management in arable crops. *Computers and Electronics in Agriculture* **65**, 155–167.

Pedersen SM, Fountas S, Blackmore BS, Gylling M & Pedersen JL (2004) Adoption and perspectives of precision farming in Denmark. *Acta Agriculturae Scandinavica Section B-Soil and Plant Science* **54**, 2–8.

Pěknicová J & Berchová-Bímová K (2016) Application of species distribution models for protected areas threatened by invasive plants. *Journal for Nature Conservation* **34**, 1–7.

Perez-Ortiz M, Pena JM, Gutierrez PA, Torres-Sanchez J, Hervas-Martinez C & Lopez-Granados F (2016) Selecting patterns and features for between- and within-crop-row weed mapping using UAV-imagery. *Expert Systems with Applications* **47**, 85–94.

Prider J, Correll R & Warren P (2012) A model for risk-based assessment of *Phelipanche mutelii* (branched broomrape) eradication in fields. *Weed Research* **52**, 526–534.

Pyšek P, Chytrý M, Pergl J, Sadlo J & Wild J (2012) Plant invasions in the Czech Republic: current state, introduction dynamics, invasive species and invaded habitats. *Preslia* **84**, 575–629.

Rahaman MM, Chen D, Gillani Z, Klukas C & Chen M (2015) Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in Plant Science* **6**, 619.

Ramírez-Albores JE, Bustamante RO & Badano EI (2016) Improved predictions of the geographic distribution of invasive plants using climatic niche models. *PLoS ONE* **11**, e0156029.

Rogers EM (1995) *Diffusion of innovations*, 4th edn. The Free Press, New York, NY, USA.

Rosenzweig C, Jones JW, Hatfield JL *et al.* (2013) The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology* **170**, 166–182.

Rumpf T, Mahlein AK, Steiner U, Oerke EC, Dehne HW & Pluemer 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.

Rydahl P (2004) A Danish decision support system for integrated management of weeds. *Aspects of Applied*

Biology. *Advances in Applied Biology: Providing New Opportunities for Consumers and Producers in the 21st Century 72*, **4**, 3–53.

Sarwate AD & Chaudhuri K (2013) Signal processing and machine learning with differential privacy: algorithms and challenges for continuous data. *IEEE signal processing magazine* **30**, 86–94.

Sawyer JE (2010) Experiences Developing and Using a Regional Corn Nitrogen Response Trial Database. In: *ASA, CSSA, SSSA International Meetings*, Long Beach CA. Available at: https://scisoc.confex.com/scisoc/2010am/webprogram/Paper59153.html (last accessed 18 April 2017).

Shrestha R, Arnaud E, Mauleon R *et al*. (2010) Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants* **2010**, plq008. doi: 10.1093/aobpla/plq008

Šikoparija B, Smith M, Skjøth CA *et al*. (2009) The Pannonian plain as a source of Ambrosia pollen in the Balkans. *International Journal of Biometeorology* **53**, 263–272.

Slavakis K, Giannakis GB & Mateos G (2014) Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *IEEE signal processing magazine* **31**, 18–31.

Smola AJ & Schölkopf B (2004) A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222.

Song W, Zhou W, Jin Z *et al*. (2005) Germination response of Orobanche seeds subjected to conditioning temperature, water potential and growth regulator treatments. *Weed Research* **45**, 467–476.

Sørensen CG, Fountas S, Nash E *et al*. (2010) Conceptual model of a future farm management information system. *Computers and Electronics in Agriculture* **72**, 37–47.

Steiner JL, Sadler EJ, Wilson G *et al*. (2009) Stewards watershed data system: system design and implementation. *Transactions of the ASABE* **52**, 1523–1533.

Straub ET (2009) Understanding technology adoption: theory and future directions for informal learning. *Review of Educational Research* **79**, 625–649.

Thibaudon M, Hamberger C, Guilloux L & Massot R (2010) Ragweed pollen in France: origin, diffusion, exposure. *European Annals of Allergy and Clinical Immunology* **42**, 209.

Thuiller W, Lafourcade B, Engler R & Araújo MB (2009) BIOMOD–a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369–373.

Travlos IS (2013a) Competition between ACCase-inhibitor resistant and susceptible sterile wild oat (*Avena sterilis*) biotypes. *Weed Science* **61**, 26–31.

Travlos IS (2013b) Responses of invasive silverleaf nightshade (*Solanum elaeagnifolium*) populations to varying soil water availability. *Phytoparasitica* **41**, 41–48.

Travlos I, Paspatis E & Psomadeli E (2008) Allelopathic potential of Oxalis pes-caprae tissues and root exudates as a tool for integrated weed management. *Journal of Agronomy* **7**, 202–205.

Van Evert FK, Spaans EJA, Krieger SD, Carlis JV & Baker JM (1999a) A database for agroecological research data: I. Data model. *Agronomy Journal* **91**, 54–62.

Van Evert FK, Spaans EJA, Krieger SD, Carlis JV & Baker JM (1999b) A database for agroecological research data: II. A relational implementation. *Agronomy Journal* **91**, 62–71.

Van Evert FK, Samsom J, Polder G *et al*. (2011) A robot to detect and control broad-leaved dock (*Rumex obtusifolius* L.) in grassland. *Journal of Field Robotics* **28**, 264–277.

Vatsavai RR, Ganguly A, Chandola V, Stefanidis A, Klasky S & Shekhar S (2012) Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In: *Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data*, 1–10. ACM.

Wainwright MJ & Jordan MI (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**, 1–305.

Weeraddana PC, Athanasiou G, Jakobsson M, Fischione C & Baras JS (2014) On the Privacy of Optimization Approaches. Available at: https://arxiv.org/abs/1210.3283 (last accessed 18 April 2017).

White JW & Van Evert FK (2008) Publishing agronomic data. *Agronomy Journal* **100**, 1396–1400.

White JW, Hunt LA, Boote KJ *et al*. (2013) Integrated description of agricultural field experiments and production: a The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture* **96**, 1–12.

Wytock M & Kolter JZ (2013) Sparse gaussian conditional random fields: algorithms, theory, and application to energy forecasting. In: ICML (3), 1265–1273.

Xie P, Bilenko M, Finley T, Gilad-Bachrach R, Lauter K & Naehrig M (2014) Crypto-Nets: Neural Networks Over Encrypted Data. Available at: https://arxiv.org/abs/1412.6181. (last accessed 18 April 2017).

Yan F, Sundaram S, Vishwanathan S & Qi Y (2013) Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering* **25**, 2483–2493.

Zhang Q & Pierce FJ (2013) *Agricultural Automation: fundamentals and Practices*. CRC Press, Boca Raton, FL, USA.

