Agricultural Systems 195 (2022) 103298

Contents lists available at ScienceDirect

Agricultural Systems

journal homepage: www.elsevier.com/locate/agsy

Big data in agriculture: Between opportunity and solution

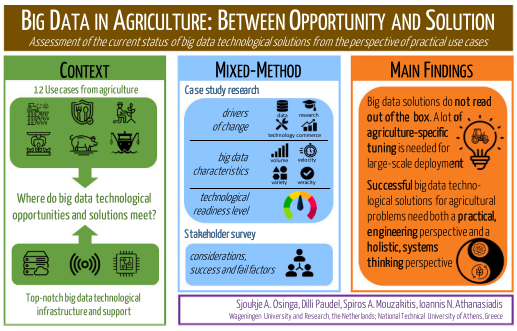
Sjoukje A. Osinga [a,*], Dilli Paudel [a], Spiros A. Mouzakitis [b], Ioannis N. Athanasiadis [a]

[a] Wageningen University and Research, P.O. box 47, 6700 AA Wageningen, The Netherlands
[b] National Technical University of Athens, Iroon Polytechniou str, 15780 Athens, Greece

HIGHLIGHTS

- Technology is not sufficiently conquered yet: the big data 4Vs still pose substantial challenges, even with top-notch technological facilities.
- The majority of state-of-the-art use cases pursues a relatively modest technological maturity level of the aspired solution.
- To stakeholders, solutions should be affordable, make use of already owned datasets, and pay specific attention to clarity of inputs and results.
- Big data solutions are not yet out-of-the-box and depend much on the domain. The transition to agriculture-specific solutions is emerging.
- Successful big data solutions for problems in agriculture need both a practical engineering and a holistic systems-thinking perspective.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Guillaume Martin

Keywords:
Big data solutions
Precision Agriculture
Case study
Stakeholders
Technological maturity level
Mixed-method approach

ABSTRACT

CONTEXT: Big data applications in agriculture evolve fast, as more experience, applications, good practices and computational power become available. Actual solutions to real-life problems are scarce. What characterizes the adoption of big data problems to solutions and to what extent is there a match between them?
OBJECTIVE: We aim to assess the conditions of the adoption of big data technologies in agricultural applications, based on the investigation of twelve real-life practical use cases in the precision agriculture and livestock domain.
METHODS: We use a mixed method approach: a case study research around the twelve use cases of Horizon 2020 project CYBELE, varying from precision arable and livestock farming to fishing and food security, and a stakeholder survey (n = 56). Our analysis focuses on four perspectives: (1) the drivers of change that initiated the use cases; (2) the big data characteristics of the problem; (3) the technological maturity level of the solution both at start and end of the project; (4) the stakeholder perspective.
RESULTS AND CONCLUSIONS: Results show that the use cases' drivers of change are a combination of data-, technology, research- and commercial interests; most have at least a research drive. The big data characteristics (volume, velocity, variety, veracity) are well-represented, with most emphasis on velocity and variety. Technology readiness levels show that the majority of use cases started at experimental or lab environment stage and

* Corresponding author.
E-mail addresses: sjoukje.osinga@wur.nl (S.A. Osinga), dilli.paudel@wur.nl (D. Paudel), smouzakitis@epu.ntua.gr (S.A. Mouzakitis), ioannis.athanasiadis@wur.nl (I.N. Athanasiadis).

https://doi.org/10.1016/j.agsy.2021.103298
Received 12 June 2021; Received in revised form 1 October 2021; Accepted 5 October 2021
Available online 21 October 2021

aims at a technical maturity of real-world small-scale deployment. Stakeholders' main concern is cost, user friendliness and to embed the solution within their current work practice.
The adoption of better-matching big data solutions is modest. Big data solutions do not work out-of-the-box when changing application domains. Additional technology development is needed for addressing the idiosyncrasies of agricultural applications.
SIGNIFICANCE: We add a practical, empirical assessment of the current status of big data problems and solutions to the existing body of mainly theoretical knowledge. We considered the CYBELE research project as our laboratory for this. Our strength is that we interviewed the use case representatives in person, and that we included the stakeholders' perspective in our results.
Large-scale deployments need effective interdisciplinary approaches and long-term project horizons to address issues emerging from big data characteristics, and to avoid compartmentalization of agricultural sciences.
We need both an engineering perspective – to make things work in practice – and a systems thinking perspective – to offer holistic, integrated solutions.

1. Introduction

Today, big data is ubiquitous, machine learning applications are thriving, artificial intelligence appears in everyday conversations, internet of things is present even in household appliances. Businesses and organizations are increasingly managed through cloud computing and high performance computing is progressively accessible as a service (Rodríguez-Mazahua et al., 2016). Opportunities and benefits are becoming omnipresent, as we can organize operations more effectively by means of high computational power, and analyze huge amounts of data by means of tailored machine learning algorithms which result in profitable insights, patterns, and decisions. As high fidelity sensory devices are becoming more accessible, data records are accumulated and open up opportunities for tackling food security problems more accurately and inclusively (Nature Food, 2020). However, reality is not always cooperative. Seemingly unimportant irregularities may become practical obstacles that stand in the way of theoretical success. This article aims to look specifically at how far apart big data technologies and practical applications are from one another within the agricultural domain.

Looking at agriculture, it is easy to imagine that technological developments can bring benefits not only to the sector, but to society as a whole (Chavas et al., 2010). Compared with other domains, the agricultural sector is characterized by relatively low operational efficiency and small managerial power due to farm size limitations, a high level of uncertainty because of weather and environmental conditions, and a volatile balance between food supply and demand due to growing and breeding times of crops and livestock (Eurostat, 2018; Huirne, 2002; Kamble et al., 2020; Newton et al., 2020). This makes farming in general a risky endeavor (Allen and Lueck, 1998). More effective operations, reduced uncertainties, and real time decision-support could revolutionize agriculture to a great extent (Cockburn, 2020). Food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic, social, and ecological benefits.

Big data technologies have been introduced already in agricultural applications (Astill et al., 2020; Cockburn, 2020; Ip et al., 2018; Kamble et al., 2020; Kamilaris et al., 2017; Lokers et al., 2016; Pylianidis et al., 2021; Saiz-Rubio and Rovira-Mas, 2020; Verdouw et al., 2019; Wolfert et al., 2017). Kamilaris et al. (2017) reported thirty-four case studies covering a wide range of big data analytics applications within the sector. They concluded that, at that time, big data analysis had not yet been widely applied in agriculture, that it was still at an early development stage. They reported also several barriers, including lack of human resources and expertise, limited availability of reliable infrastructure (Sawant et al., 2016), and lack of standardization and governance related to agricultural data (Nandyala and Kim, 2016; Nativi et al., 2015). Next to the technological advancements, business models should be sufficiently attractive for solution providers and stakeholders (Kempenaar et al., 2016; Sonka, 2016). Big data applications in smart farming is also related to socio-economic challenges Wolfert et al. (2017). According to the aforementioned reviews, most big data applications had not been fully adopted by their intended users and were either in an early development stage or had a limited scope in the sense that they did not sufficiently address the risks related to agriculture.

Since then, big data technologies have developed further, as more experience, algorithms, good practices and computational power become available (Oussous et al., 2018). Several new big data and artificial intelligence applications for the agricultural domain have been developed all over the world (Lezoche et al., 2020). The focus of this article is to study the conditions of the adoption of big data technologies in agricultural applications, based on the investigation of twelve real-life practical use cases in the precision agriculture and livestock sector. Are data-driven solutions effective in addressing the intrinsic characteristics of the agricultural applications? What level of maturity have big data solutions reached? Do end-users or stakeholders see the added-value to adopt them?

There is no straightforward answer to these questions. "Big data technology" is not a well-defined concept, but rather an ecosystem of several technologies that may address various types of problems (Cui et al., 2020; Jagadish, 2015). Finding the match between enabling technologies and successful applications is by no means trivial, as recent review articles from various sectors demonstrate (Kuo and Kusiak, 2019; Saggi and Jain, 2018).

1.1. Perspectives

Therefore, to investigate the status of the adoption of big data solutions in the agricultural domain and the challenges that go with it, we take on a pragmatic approach. We consider four perspectives to study twelve use cases of a specific research project on big data in agriculture. Together, these perspectives contribute to shape our understanding of this adoption within the scope of the research project.

The first perspective focuses on identifying the drivers of change behind the application, or in other words, the motivations behind the choice for adopting a big data technology solution. When the initiative comes from the technology provider with a promising technology, looking for a suitable domain area to apply it to, the focus maybe more on the technology than on the problem itself. The resulting application may be more technology-oriented than practical for the end user. Alternatively, when the need originates from someone in the domain area itself, the leading goal is to obtain a suitable application to solve an existing problem, regardless of whether the technology of this application is established or novel. Likewise, when researchers are involved, it is likely to assume that they are mainly interested in proof-of-principle solutions. A different outcome may be expected when there is a commercial interest that requires a more profitable, hence practical application. Therefore, the driver of change may have a different impact on the successfulness of the adoption and is worth investigating.

The second perspective is to investigate the big data characteristics of the problem: volume, velocity, variety and veracity (Laney, 2001). Each of these 4 Vs has different associated challenges that may prohibit

the adoption of big data applications. For example, a volume-related data challenge can be addressed by improving the data storage and processing methods, which is mainly a technical issue that relates to availability and efficient use of infrastructure and processing power (Ang and Seng, 2016). The same holds for a velocity-related challenge, where processing speed is important to allow for real-time response, which is also primarily a technical issue. On the other hand, a variety-related data challenge needs a customized way of pre-processing the data which may concern semantic issues, for which domain expertise is needed, which makes it a content-related issue (Jagadish, 2015). The same is true for veracity, where also a domain expert should be involved to resolve issues related to data cleaning or interpretation.

The third perspective assesses the maturity of the intended solution, or in other words, the ambition level of the developers in terms of practical applicability. This is partly related to the drivers of change: when the intention is to develop a prototype solution in a research context, then this will probably not qualify for a practical solution that will be adopted on a large scale. Likewise, when the intention is to find a solution for a real-world problem situation, then successful adoption of this solution is more likely.

The fourth perspective concentrates not on the problem domain or the solution providers, but on end-users and stakeholders. This perspective is included to provide an additional viewpoint. The drivers of change, big data characteristics and maturity of the solution are investigated from the viewpoint of the application developers and domain experts. Stakeholders and end users cover a broader range, and may have a different opinion on what they consider a successful application.

We investigate these four perspectives applying a mixed method approach around the twelve use cases developed by the CYBELE project (Perakis et al., 2020). CYBELE is an EU-funded H2020 project in which an interdisciplinary consortium of partners in the fields of high-performance computing, big data, cloud computing, and internet of things co-develop big data solutions for real-world use cases related to several facets of agriculture: from precision arable and livestock farming, to fishing and food security. CYBELE can be considered a laboratory in which the potential of big data solutions is examined by investigating to what extent big data solutions can be tailored to solve actual problems in agriculture, as both domain experts and technical providers are actively collaborating to deploy practical solutions. Our mixed-method approach consists of a case study research and a survey. For the case study research, we interviewed in-person the use case representatives of the CYBELE project. The survey addressed a broad network of stakeholders and end users. This mixed-method approach was selected to provide a more in-depth and timely understanding of the dynamics in this field, rather than a literature review.

1.2. Research steps

The research focus of this article is to assess the conditions of the adoption of big data technologies in agricultural applications. We make this assessment by means of the four perspectives introduced above, each highlighting a different aspect of the adoption conditions. Therefore, the research steps are (1) to identify the drivers of change behind the use case applications; (2) to identify the big data characteristics of the use case applications; (3) to identify the maturity level of the intended solutions for the use case applications; and (4) to identify stakeholders' considerations for adopting big data solutions.

The remainder of this article is organized as follows. Section 2 explains the mixed-method approach in further detail. Section 3 presents the results. Section 4 discusses our findings, and Section 5 is the conclusion.

2. Methods

Our mixed method approach consists of a case study research and a survey. The relationship between perspectives and how they are operationalized is presented in Fig. 1, and explained in the subsequent subsections. The first three perspectives center around the use case problems and intended solutions themselves and are studied by means of a case study research. The last perspective concentrates on the stakeholders' considerations, and is studied by means of a survey.

In the remainder of this section, we explain the methodological details of the case study research (2.1), the survey (2.2), and the four perspectives (2.3). We also present a short overview of the twelve CYBELE use cases (2.4).

2.1. Case study research

The case study research consisted of the following steps. First, we studied the use case documentation as available within the CYBELE project. We then carried out semi-structured interviews with each use case representative concerning both the existing problem situation and current solution (if available), the expected future big data solution, and how this is assumed to change the problem situation. By 'use case representatives' we mean representatives from the project partners who had brought the use cases to the project. The representatives were either directly from the respective company or organization, or researchers with close connections there.

We conducted a semi-structed interview with each use case representative, by means of Skype, or live where possible (once) in March-May 2019. For every interview, one or two representatives from the same use case were present. During each interview, workflow diagrams were discussed and drawn together through a shared screen. The result was a set of at least two workflow diagrams (the current problem situation and the desired big data solution), or more when appropriate. For each workflow diagram, also the people involved, and their roles and tasks were discussed. All interviews were recorded (audio and screen recordings, for the live interview audio only). The resulting interview documents and workflow diagrams were further commented and revised by the use case representatives in order to correct any mistakes or to clarify issues. All interview data and resulting diagrams are extensively documented in a public CYBELE project report (Athanasiadis et al., 2020).

We then performed a three-stage analysis and assessed (1) the drivers of change, (2) the big data characteristics and (3) the technological readiness level of each use case. These analyses were based on the authors' interpretation of the use case descriptions and of the interviews in which they were discussed. The results were presented at subsequent project meetings where they could be verified and further refined.

2.2. Stakeholder survey

We carried out a survey among the stakeholders that are part of CYBELE's extended network, in order to analyze the stakeholder perspective. These stakeholders form a wider group than the intended end users of CYBELE applications, aimed at intermediary service providers. Note that we asked the stakeholders' opinions *during* the CYBELE project, not afterwards, as the project is still ongoing.

The stakeholder survey was constructed and distributed among relevant stakeholders and possible future users of CYBELE, contacted through CYBELE's extended network. The survey ran between 19th of November and 17th of December 2019 and had 61 respondents. The survey consisted of seven close-ended Likert-scale questions, with additional options for open responses. The survey questions elaborated on what stakeholders consider a reason to adopt a solution, which relates to (their) drivers of change. They were asked to evaluate factors that can potentially improve productivity or profit within a domain. Those factors were related to the big data 4 Vs. They were also asked for critical success factors and to mention factors that hinder adoption. These questions relate to the maturity level of a solution.

The survey questions are presented in Appendix 1. For the analysis of

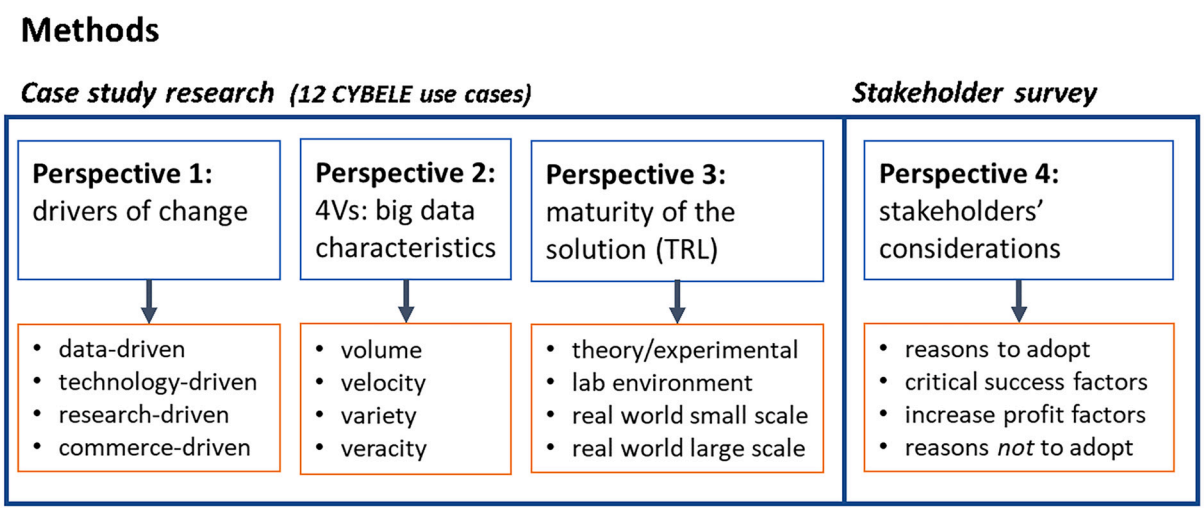


Fig. 1. Operationalization of our mixed method approach

the stakeholder survey, we used standard statistical indicators. The full survey questions and analysis can be found in Athanasiadis et al. (2020) and is also discussed in Mouzakitis et al. (2020).

2.3. Perspectives

The methods and argumentation used for each of the perspectives is presented in the next four subsections.

2.3.1. Drivers of change

The first perspective is concerned with the driver of change of the use case, i.e. what the motivation was to apply big data technologies and data-driven analytics. From literature, we know that changes in farm practice with respect to decision making are likely to occur in response to 'trigger events', after which new options are considered and pursued (Sutherland et al., 2012). Push and pull mechanisms can serve as such triggers as well (Wolfert et al., 2017).

We operationalized these motivations into four categories: (a) *data-driven*, meaning that more data sources or higher-quality data regarding the use case domain are available that form an opportunity to solve a problem. For example, higher resolution images can increase the quality of an image-based decision. (2) *technology-driven*, meaning that solving the problem becomes more efficient when more advanced techniques are applied. For example, involving parallelization or high-performance computing can speed up the execution of a required model which leads to a faster decision. (3) *research-driven* in the sense that involved academic researchers who have a strong link with the problem domain see a solution to a problem from the perspective of a research innovation or a funding opportunity. (4) *commerce-driven,* when parties involved see an opportunity to gain competitive advantage in offering an advantageous solution for a typical problem in a certain sector.

There is not necessarily a single driver of change, instead it is likely that a combination of drivers of change is applicable.

2.3.2. Big data characteristics

As a second perspective, we analyzed each use case problem in terms of the big data characteristics (Laney, 2001). The complexity of handling big data is associated with its typical characteristics, the big data 4 Vs: volume, velocity, variety and veracity. For an extensive definition of each of them with examples from the agro-environmental domain, see Lokers et al. (2016). Characteristic *Volume* implies that the data needed to solve the problem is very large in size, and increasingly growing. Characteristic *Velocity* implies that decision-making has to happen in real time, so the data and models necessary to make that decision have to be processed very fast. Characteristic *Variety* implies that data of multiple formats needs to be processed together for decision-making. Characteristic *Veracity* implies that there is uncertainty involved in the data required for decision-making. With more data sources becoming available, the fidelity of the data is not always guaranteed.

When more than one qualification applied, we assigned only the two most dominant ones, in which case we prioritized them as most important and second important.

2.3.3. Maturity of the solution

As a third perspective, we looked at the maturity of the proposed big data solution for the problem. A solution will be more likely to be adopted when it was the intention of the use case representatives to achieve a sufficient level of applicability. To assess the maturity of the solution, we applied a scheme originally developed by NASA and adopted by the European Commission to estimate the maturity of EU-funded research and innovation projects (EARTO, 2014). We used the technological readiness level (TRL) scheme to identify whether a solution is aimed to be at experimental level only, at prototype stage in a lab environment, to be deployed in a small-scale real-world environment, or to reach real-world large-scale deployment.

In order to apply the TRL scale, running from 1 to 9 to assess how ready a certain technology is, it needs to be adapted for the intended application (EARTO, 2014). Our TRL interpretation is summarized in Table 1, where we group the scale levels into four ascending categories. The lowest category of TRL relates to applications that are merely experimental and have no other intention than to test a conceptual model or theory. The next TRL category indicates that the intended solution is to produce a proof of concept or a prototype, but not to take this out of the lab situation. The third TRL category consists of big data solutions that aim for deployment in the real world, but on a small scale, perhaps still in a controlled environment. The highest TRL category is reserved only for applications that are operating in real world environments and have all potential to be adopted on a large scale.

By means of this scale we assigned each use case a TRL level for the current technological maturity state of the proposed solution. We estimated the technological maturity that is expected to be achieved in 2-3 years, which is the duration of the CYBELE project. This 2-3 year ambition level may be lower than the ultimate ambition level of a use case, but that is considered outside the scope of CYBELE and this article. All use cases were assigned TRL scores that refer to the desired big data solution only.

Table 1
Interpretation of NASA's Technology-Readiness Level (TRL) to indicate the technological maturity of each case study's proposed big data solution.

TRL	Stage	Explanation
1-2	Theory; experimental stage	*No system at all, but problem has been explored theoretically* TRL 1: there is a conceptual model TRL 2: there is a model ready to implement
3-4-5	Lab environment; prototype stage	*There is a system, but so far only for scientific explorations* TRL 3: system is a proof of concept TRL 4: system has been tested in controlled lab environment TRL 5: system has been simulated in a relevant but still controlled environment
6-7	Real world; small scale deployment	*The system is deployed to the real world on a small scale* TRL 6: small-scale experimental application in a controlled, but real world environment TRL 7: small-scale experimental application in a commercial real world environment (no longer controlled)
8-9	Real world; large scale deployment	*The system is deployed in the real world on a large scale* TRL 8: wide range commercial trial in the real world TRL 9: system is in full commercial use on a large scale

2.3.4. Stakeholder considerations

It is important to include the intended stakeholders of an information system in an early stage to be aware of their needs (Jayashankar et al., 2019; Penn et al., 2019). Indeed, the CYBELE research project is set up in such a way that end users and stakeholders are consulted at regular intervals, while the components of the solutions are being developed.

The survey that we used for this article was designed while we already had knowledge of the use case descriptions and interviewed the use case representatives, which tailored our questions. What interested us primarily was to see how the stakeholders' view corresponded to or differed from what the use case representatives had indicated regarding the three other perspectives.

2.4. CYBELE use cases

The case study research is centered around the twelve CYBELE use cases related to precision agriculture and precision livestock farming applications. A short description of what each use case entails is provided in Table 2. More elaborate descriptions can be found in Appendix II.

3. Results

We present our results per perspective in four sub sections.

3.1. Perspective 1: Drivers of change

The twelve use cases are approximately equally driven by data, technology, research, and commercial interests, and most often a combination of these (Table 3). Most use cases are at least motivated by a research interest. Noteworthy is that only three out of twelve use-cases are commerce-driven.

Data-driven use cases seek to incorporate new data streams, that were either (a) higher quality or higher resolution datasets (e.g. use case 5); (b) existing, but unexplored datasets, such as already measured sensory data (e.g. use cases 6 and 10); or (c) incorporating new data sources, as in the case of crowd-sourced data (e.g. use case 1) or by deploying new sensors (e.g. use case 4). *Technology-driven* use cases aim to solve problems more efficiently by employing more advanced technologies. Use cases 3, 9, and 11 employed technologies related to the real-time access and processing of voluminous data sources. A *research drive* is

Table 2
A summary of the CYBELE use cases from agriculture and livestock domains

Nr	Short name	Aim (within the project scope of 2 to 3 years)
1	Organic soya yield prediction	To develop methods to increase accurate soya yield prediction, including soil analysis and weather data. Use crowd-sourced datasets provided by soybean growers to develop advanced machine learning algorithms to predict the protein content.
2	Prevent food safety incidents	To demonstrate the capacity of HPC solutions for supporting complex deep learning and machine learning prediction models to estimate food safety risks for dairy products, nuts product, sugar, fruits and vegetables, in order to minimize risks and mistakes in future.
3	Prevent frost/hail damage	To develop an early warning system that can help farms to prevent damage on their grapes through protective methods for frost and hail. The farmers could benefit highly when they know such events will happen.
4	Develop agro-robots	To develop fleets of small, autonomous vehicles and carry out farmer tasks. The range of operations to be delivered include soil chemical analysis, identify soil/crop condition, (plant/weed) identification, individual plant harvest readiness assessment and plant level automated harvesting.
5	Optimize crop yields forecast	Improve the existing instrument for crop yield monitoring (e.g. early warning & anomaly detection), index-based insurance (index estimates) and farmer advisory services. Include parcel-specific data associated with advanced weather forecasts and computations and additional satellite imagery data.
6	Pig weighing optimization	An accurate estimate of the live weight of slaughter pigs is useful to the farmer, especially for knowing when to send the pigs to the slaughterhouse, and to more accurately diagnose and treat diseases. Goal is to infer the live weight of the pigs via video images with deep convolutional neural networks.
7	Sustainable pork meat quality	Sustainable pig production and global food challenges require producing with optimal productivity, health and welfare of the pigs. These can be obtained by on-farm data sources and data from slaughterhouses that are available but not yet fully utilized. By integrating and analyzing these data sources at a large scale, these goals can be easier obtained, resulting in higher meat quality and better conditions for the pigs.
8	Improve health and welfare of pigs	This case aims to improve the detection of health, welfare and performance problems at fattening pig farms through better use of available sensor and farm data.
9	Integrate fish fleet data	To utilize the potential of commercial fishing vessels, which can collect real time data with sensors and IT equipment on board, to improve ecosystem-based fisheries management to locate fish and to avoid overfishing. To this end, data from the digital logbooks from ships, satellite-based vessel monitoring systems, environmental data from satellite based imaginary systems, and more are used.
10	Optimize fishing vessel on-board data	This case aims to optimize the database system on board of a particular commercial Belgian fishing vessel which has advanced data collection systems but no integrated system to process these.
11	Machine vision for fish catch detection	This case is to support visual-based processing of the catch using an RGB camera instead of manual sorting, which is labour-intensive and gives no information on detection of species, undersized fish, non-commercial fish and non-commercial benthic animals.
12	Aquaculture monitoring	Optimize the process of fish feeding, because feed is a high cost factor, and wasted food is deposited in the seabed and generates an environmental impact on the surrounding area. Aerial images of fish farms taken from drones will be processed, combined with weather information and sensor measurements.

Table 3
The drivers of change for the 12 use cases. Most use cases report more than one driver, which are not prioritized

Nr	Use Case ID	Data-driven	Technology-driven	Research-driven	Commerce-driven
1	Organic soya yield prediction	√			√
2	Prevent food safety incidents	√	√		
3	Prevent frost/hail damage		√	√	
4	Develop agro-robots	√		√	
5	Optimize crop yields forecast	√			
6	Pig weighing optimization			√	
7	Sustainable pork meat quality			√	√
8	Improve health and welfare of pigs			√	
9	Integrate fish fleet data	√	√		
10	Optimize fishing vessel on-board data	√			
11	Machine vision for fish catch detection		√		
12	Aquaculture monitoring			√	√

typical for use cases where innovation is still in an early stage and big data methods are explored for their potential, such as in use cases 4 and 6. A *commercial* interest was manifested when a company was already involved, which is the case for use case 1, 7 and 12, or when there was potential in the asset value in e.g. the data collection of the use case.

3.2. Perspective 2: Big data characteristics

Velocity and Variety were found to be two of the most essential big data characteristics for the use cases (Table 4). *Velocity* is a strong prerequisite for real-time decision-making and it is considered important for more than half of the use cases we investigated. For example, in use case 4 a robot on a tomato field needs to decide whether a tomato should be picked or not. This requires very fast processing of the data coming from the various robot sensors to make the right decisions.

The ability to handle data *variety* is necessary when combining data from various sources. This prerequisite is important for use cases that have a research focus, such as use cases 2, 10 or 12. For example, in use case 12 (aquaculture monitoring), in-situ sensors are to be integrated with drone video footage, and feed management systems. This poses challenges on how to integrate and analyze data of multiple modalities.

Volume was considered as important for only one use case, the same one that reported only data as the driver of change (use case 5). Despite not being the main characteristic, volume remains relevant for several other use cases. In use case 1, satellite images of ever increasing temporal and spatial resolution are processed to increase the granularity of decision-making, which also increases the need for more storage capacity.

Veracity was found relevant for use cases whose predictions depend on uncertain data, such as weather data or data coming from multiple sources, e.g. in use cases 1 and 3. For example, in use case 3 that aims to prevent frost and hail damage, it is very important to have accurate, local weather forecasts, since hail is very difficult to predict.

3.3. Perspective 3: Maturity of the solution

We measured the maturity of the use case solutions both at the beginning of the project and the intended one at the end of the project using the Technology Readiness Level (TRL) (EARTO, 2014). All use cases started from a low maturity level (experimental or prototype stage), and the majority aimed for a real-world small-scale deployment (Table 5).

Three use cases aim for prototype-level deployment at the end of the project in a controlled lab environment. Two out of three were driven by a research interest. For example, use case 6 which aims to identify individual pigs in a pen and estimate their weight from streaming video data.

Most of the use cases (n=7) aspire to small-scale real-world deployments at the end of the project. They aim to make predictions out of real-world data, but do not consider all the potential uncertainties and vulnerabilities such as local adaptations or missing data that might occur in a large-scale deployment.

Only two use cases aim for the highest TRL, that of large-scale real-world deployment. Both use cases 7 and 9 aim at improving existing systems for commercial practice, and the technical challenge lies in velocity. This ambition level fits well with the CYBELE project that offers state-of-the-art supercomputer infrastructure and support, which can facilitate to realize such objectives.

3.4. Perspective 4: Stakeholder considerations

The majority (61%) of the stakeholder survey respondents (n=56) have a background in agricultural farming; circa 20% have a research background; circa 7% are from technology and knowledge providers; the remaining respondents are policy makers, other companies, or public entities. Most respondents self-identified as being familiar with precision livestock and precision agriculture farming (circa 70%). Almost half of the respondents has no experience with big data in agriculture. The rest assess themselves as novice (25%), competent

Table 4
The big data characteristics of the 12 use cases. More than one characteristic may be assigned to each use case, and the large-sized indicator represents the predominant one.

Nr	Use Case ID	Volume	Velocity	Variety	Veracity
1	Organic soya yield prediction	√		√	√
2	Prevent food safety incidents			√	√
3	Prevent frost/hail damage				√
4	Develop agro-robots		√	√	
5	Optimize crop yields forecast	√			
6	Pig weighing optimization			√	√
7	Sustainable pork meat quality	√	√		
8	Improve health and welfare of pigs	√	√		
9	Integrate fish fleet data	√	√		
10	Optimize fishing vessel on-board data			√	
11	Machine vision for fish catch detection		√		
12	Aquaculture monitoring		√	√	√

Table 5
Results of the analysis with respect to TRL; the starting point represents the TRL at the start of the project; the arrow indicates the ambition level for the duration of the project (3 years).

Nr	Case Study	1	2	3	4	5	6	7	8	9
1	Organic soya yield prediction			|-	----	----	->			
2	Prevent food safety incidents			|-	----	----	->			
3	Prevent frost/hail damage			|-	----	----	->			
4	Develop agro-robots		|-	----	----	-->				
5	Optimize crop yields forecast				|--	----	---	->		
6	Pig weighing optimization		|-	----	-->					
7	Sustainable pork meat quality				|--	--–	---	---	->	
8	Improve health and welfare of pigs				|--	----	---	->		
9	Integrate fish fleet data		|-	----	----	----	---	---	->	
10	Optimize fishing vessel on-board data		|-	----	----	----	---	->		
11	Machine vision for fish catch detection		|-	----	----	-->				
12	Aquaculture monitoring		|-	----	----	----	->			

(18%) or expert (9%) in big data technologies.

Stakeholders' most important reason to opt for a big data solution was its commercial value (n=23). Data, technology, and research drivers of change seem to be equally important to them (n=18, 19, 18). They are least driven by research (n=4) when considering to adopt a big data solution for making better decisions for their organizations.

When stakeholders consider whether a big data solution could improve productivity, they indicate that the most important factors to them are access to visualizations and data-analytics (n=22), instant delivery of notifications and alerts (n=20), very fast data processing capabilities (n=20), and out-of-the-box data availability (n=18). The least important factor is big data infrastructures (n=7).

Most stakeholders (n=28) indicated that financial attractiveness is important for adopting a solution. Stakeholders are also concerned that big data solutions will not be easy to embed in their current workflow, that they lack staff, or that their staff is not skilled enough to work with big data solutions.

4. Discussion

We structure the discussion section as follows. In 4.1, we discuss our findings directly related to our perspectives. In 4.2, we state the added value of our work. In 4.3, we combine our insights into a list of lessons learned. Lastly, in 4.4, we discuss the limitations of the study.

4.1. Findings related to our perspectives

4.1.1. Perspective 1: Drivers of change

Our findings suggest that most use cases were data-driven, triggered by the opportunities provided through the availability of new, high-quality or more diverse datasets. Also, they mainly had a research interest, and were primarily concerned with developing the technological solution rather than making it available to end-users. In the literature we find confirmation that big data technology is mostly seen as a driving force for companies or organizations to create more value out of their data (Kuo and Kusiak, 2019; Saiz-Rubio and Rovira-Mas, 2020). At the same time, it is broadly recognized that data alone are not enough to make applications successful (Kamble et al., 2020; Lezoche et al., 2020; Saggi and Jain, 2018; Wysel et al., 2021). Wysel et al. (2021) summarize this requirement concisely as "Creating value from data requires a community of stakeholders, a faciliatory system, and data on, and for, the community". Our findings show that end user or stakeholder involvement, although certainly accounted for in the project, was not yet a main priority for the use cases, at least not within the scope of this project.

4.1.2. Perspective 2: Big data characteristics

From our findings we learn that all four big data characteristics are relevant challenges for our use cases, and that the predominant ones are velocity and variety. This implies that working with big data still poses substantial technical challenges that cannot be overcome solely by scaling up technical infrastructure facilities and support, as is explained by Jagadish (2015). Especially the requirement to use the knowledge from a domain expert – in case of a variety or veracity challenge – is generally considered a bottleneck (Coleman et al., 2016). This is one of the key reasons reported in our interviews of why it remains a very big step to move from a real-world small-scale to a large-scale deployment – which also relates to our third perspective. Several of our case studies suffer from low quality or unbalanced datasets and it was a challenge for them to reproduce prototype results in controlled environments.

Consider as an example an experimental farm. Data are generated in a very much controlled manner, yielding complete datasets. On the contrary, in a real-world deployment involving several farms, data sources can be expected to be incomplete, which essentially changes the type of problem to be solved. There are successful examples of scaling up data-driven solutions, like the Copernicus project that brings together huge amounts of detailed satellite imagery, together with services to disclose these data (EU Earth Observation programme, 2021). However, there are less successes of such magnitude in agriculture. Large-scale agricultural applications involve high data uncertainties, due to missing data, noise, mismatches in scales, scopes, and formats (Lezoche et al., 2020). Data are segregated in silos, that still need to be integrated, and linking these data requires further investments before getting any additional insights.

4.1.3. Perspective 3: Maturity of the solution

The research context of CYBELE gives rise to expectations. Given that CYBELE facilities and support are state-of-the-art, it may be expected that the ambition level of the use cases is of the highest level as well. From our results we observe that this is not the case. Apparently, the transition from a lab environment to a real-world environment is a very large step, even when close support from big data experts and data infrastructures are available (Lezoche et al., 2020; Wysel et al., 2021). It requires complementary expertise and additional skills and experience that most use cases do not have in-house. It also requires new partnerships that take long time to be effective (Wysel et al., 2021).

On the other hand, the CYBELE research context may have a limiting effect as well. One limitation could be on the use case owners' ambition level, because of the 2-3 year scope of the project. There are more generic pitfalls when it comes to limiting case studies to a funded project, for example the fact that there is inequality in research funding which is an intrinsic bias to the results (Li et al., 2017). On the other hand, if we had instead approached several companies to do our research, we would probably also have had biased results, because of the differences between them (application area, startup or established company, and so on). The CYBELE context was perhaps limited but well-defined and in a way equal for all use cases. Despite the limitations, and given that alternative ways could have been used to reach results, our approach gained valuable insights.

4.1.4. Perspective 4: Stakeholder considerations

The stakeholder survey emphasized the end user viewpoint: their main interest is to gain advantage by means of affordable solutions that can be embedded in practice. There is a gap between the level of most current technical solutions and the actual needs of potential end users. The more use cases diverge from similar developments in other fields of application, the bigger such a gap is.

Often, stakeholders are not included at all in the development of agricultural decision-support systems, despite good intentions (Mir and Padma, 2017). But also if they are, apparently there is no ideal match yet between the project partners' push solutions in the form of relevant, up-to-date scientific knowledge on the one hand, and the stakeholders' knowledge on the other, as is recommended by Martin (2015). In part, we should attribute this discrepancy in our results to the way we presented the stakeholder survey. The survey was constructed after the case study and contained closed questions on adoption rather than more open questions to see how they would have expressed their needs themselves, which would have been a more participatory approach (Van Meensel et al., 2012). Also, the use case representatives' goal was to achieve a big data solution for their own problem; they did not frame this goal to aim for large-scale adoption or to address stakeholders' needs, a lack of conguency also described by (Bundy et al., 2018).

However, this article puts emphasis on investigating the driving factors and conditions for the adoption of big data solutions from different perspectives. Our results mainly show where the use case representatives stand, and where the stakeholders stand. Despite the inherent limitations of the survey, we still observed that the stakeholders had different perspectives and drivers from the use case representatives.

4.1.5. Added value of this research

Our research provides some useful insights mostly already known from literature, but its key strength is that our results rely on an empirical analysis of case studies. This practical, pragmatic approach is an added value to the theoretical body of knowledge.

We highlight the kick-starting conditions for successful adoption of big data solutions rather than the solutions themselves, therefore we focus on the status quo. To do this, we use the CYBELE project as a laboratory, in which all variables are similar to the participants, like in a controlled environment. The partners have access to facilities, network, HPC capacity and funding, within which they develop their solutions. Even in these circumstances, we see that the solution developers are cautious: apparently, the adoption is not easy.

Since big data solutions are not widely used, the use cases of CYBELE highlight the most important problems or challenges at hand. These challenges are often domain-specific. A generalized solution may be more difficult to achieve or less useful as a solution. These observations are practical and empirical rather than theoretical, and we can consider this an added value as well.

4.2. Lessons learned

4.2.1. Technology transfer is hard in the agricultural domain

The adoption of big data solutions depends largely on the agricultural domain. We can relate this to the concept of technology transfer: the process of applying known technologies to new and novel applications (Lane, 1999). It is possible to have a direct transfer of mature technological advancements from other fields to agriculture (Hayter et al., 2020). For example, consider computer vision that has been successfully applied in a variety of domains (Peregud and Zharovskikh, 2020), with production-level software libraries available: it has been relatively straightforward to employ this in agricultural use cases as well.

However, the challenges reported by our twelve case studies are highly related to the agricultural subdomain they originate from. We observe that in most use cases further research is required to extend and adapt established big data technologies and make them useful in agriculture. Translating general big data and artificial intelligence technology into meaningful applications in agriculture requires still further development.

4.2.2. Domain-aware and customizable solutions are called for

Another insight is that most use cases are highly specialized. The compartmentalization of the agricultural domain itself, as manifests from datasets, models and practice, hinders the widespread exploration of various big data and artificial intelligence advancements, and prohibits technology transfer across applications with similar problems. Agricultural applications are too diverse and local to have universal solutions that can be applied across several systems (let them be fields, farms, or food chains). Solutions that work in one case are not easily transferrable to another one (Saggi and Jain, 2018). Further research is needed on how to customize solutions, and also how to transfer knowledge from one domain to another.

4.2.3. Big data solutions require more than technology

Literature suggests that advancing the maturity level of big data technology is a matter that goes beyond technological challenges (Ang and Seng, 2016; Lezoche et al., 2020; Saggi and Jain, 2018; Wysel et al., 2021). The link between digital agriculture and economic, business and institutional arrangements has been highlighted by Klerkx et al. (2019). Similarly, socio-economic issues, organization and governance, suitable business models for data sharing, and attention for the entire supply chain have been suggested to have priority over developing big data technology itself (Wolfert et al., 2017).

4.2.4. Stakeholders' needs are difficult to identify

When developing big data applications, it is important to take the expectations of the intended end users into account at an early stage (Jakku and Thorburn, 2010). In CYBELE, the end users (other than use case representatives) are scheduled to be heard regularly in various stages of the project. Even under these conditions, the stakeholder views appear to be different from the use case representatives' views. This means that the adoption of solutions that match well with stakeholder needs is still a challenge.

4.2.5. A systems-thinking approach is required to co-develop integrated solutions

Neither the engineering, nor the end-user perspective alone are sufficient for real-world big data applications in agriculture. Rather a systems-thinking approach is required, able to co-develop integrated solutions. This is a long process, where progress happens in small steps. It requires agricultural and big data experts to engage in a holistic and interdisciplinary process, develop a common language to communicate effectively, and work together for better understanding the uncertainties of the multi-faceted agricultural systems.

4.3. Limitations

Our research concentrated on the analysis of twelve use cases. The selection of CYBELE partners and case studies involves some bias that limits this research. For example, the use case partners may have tailored their ambition level to what they could realistically deliver within the scope of the project. Still, the use cases may serve as demonstrative examples of the status of big data in the agricultural domain. CYBELE is a collaborative project, funded by a competitive grant by the EU Horizon 2020 programme. Therefore, the selected use cases are at the forefront of state of the art in big data technology and the agriculture domain in Europe. The thorough selection procedure serves as a quality attribute of involved partners. Also, the breadth of the case studies topics, and the diversity of big data technologies involved, increase our confidence that both the agriculture and big data domains have been representatively covered.

Similarly, this work is also limited by the fact that survey respondents were contacted through the CYBELE consortium network. However, from the survey results we can infer that the survey respondents represent the agricultural domains well enough, as they include stakeholders from different backgrounds.

5. Conclusion

Having studied the use case problems, the solutions, and the stakeholders viewpoint from all four perspectives leads us to the conclusion that the adoption of big data solutions is still modest.

Despite the unencumbered access to top-notch infrastructure and support, most use cases pursue a relatively modest technological maturity level of the aspired big data solution. The distance of big data solutions to solve actual problems at large scale seems too far to bridge within a three year project. Use cases revealed that it is not easy to advance from middle to high technology readiness levels, but that this is rather a process that takes time, as new problems arise when scaling up.

Big data technology has not sufficiently been conquered yet: the big data characteristics still pose substantial challenges, even when excellent technological facilities and support are available. Big data solutions do not work out-of-the-box when changing application domains, and additional technology development is needed for addressing the idiosyncrasies of agricultural applications. The adoption conditions of large-scale, agriculture-specific big data systems are emerging, and a systems-thinking approach is required to co-develop big data solutions for addressing agricultural systems uncertainties and food security challenges.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been co-funded by the CYBELE project, a European Commission research program under grant agreement 825355. We are particularly grateful to the interviewees from the CYBELE case study partners, and the stakeholders of our survey. We would like to thank Simon Ntanopoulos, Giannis Tsapelas and Dimitris Askounis (NTUA) for their help with the organization and analysis of the survey.

Appendix I – Stakeholder Survey Questions

Nr	Question
	What is your domain?
1	Agri-food industry / Agriculture farming / Livestock farming / Policy maker / Researcher / Insurance or Marketing / General public or consumer / Other
	Are you familiar with Precision Agriculture/ Precision Livestock Farming?
2	Yes / No / Other
	What is your organization's current position regarding the application of data analytics and machine learning techniques for agriculture and livestock farming?
	Aspirant: We have no experience with these applications but may wish to have in future
	Novice: We are experimenting with pilot applications that are not used in practice yet
	Competent: We have some applications running that are currently used in practice
3	Expert: We have many applications running that are currently used in practice
	Consider the following reasons to adopt a solution that involves analyzing information coming from installed sensors, cameras, satellite images etc. in order to make better predictions and analytics for your organization. How important do you think each of them is?
	Prioritize the following options 1- 4; 1 (most important) – 4 (least important) :
	1) Data orientation: there are additional data available that we don't use yet
	2) Technology orientation: there are techniques available that we don't use yet
	3) Research orientation: there is a research interest in finding a solution for this problem
4	4) Commercial orientation: there is an opportunity to gain competitive advantage
	How important do you think each of the below factors is for improved productivity/ profit?
	Rate each on a 5 point scale; 1 (not important) – 5 (very important) :
	1) Physical technical infrastructure (e.g. sensors for data collection)
	2) Data availability (to be able to find the data needed out-of-the-box)
	3) Data Storage infrastructure (to be able to store a very large amount of data)
	4) Very fast data processing capabilities
	5) Access to visualizations & data-analytics (turn available data in actionable information)
	6) Access to customizable data visualization & analytics (to have access to an earlier stage in the analysis)
5	7) Instant delivery of notifications and alerts
	How important do you think each of the below attribute is to adopt a technological solution relevant to Precision Agriculture/ Precision Livestock Farming?
	Rate each on a 5 point scale; 1 (not important) – 5 (very important) :
6	Cost / Precision of results / Time / Number of functionalities / Customization / Interoperability
	Other than the above, can you think any more factors that are critical for success of a technical out-of-the box solution?
7	Open question
	What would be a reason for your organization NOT to adopt practices related to Precision Agriculture/ Precision Livestock Farming?
	Multiple-answer checkbox :
	1) Lack of finance – We expect not to be able to afford this solution
	2) Limited scope – Big data analytics/ML probably solves our problem only partially
	3) Hard to embed – Our business processes are not suitable to work with this solution
	4) Long term maintenance – Not confident that we can use this solution for the long term
	5) Lack of realism – Our situation will always be more complicated
	6) Personnel/Skills – Our staff cannot work with this solution
	7) Security concerns / Potential privacy issues
8	8) Other:

Appendix II - CYBELE use cases detailed description

1 Organic Soya yield prediction

In the EU, soybean is the mostly used plant protein for animal husbandry. Currently, the EU imports most soybean from Brazil, Argentina and the US, but soybean cultivation within the EU is growing rapidly. Unlike the US and the rest of the world, GM (Genetically Modified) products are strongly regulated in the EU, meaning that the EU must rely on its own production and increase its efficiency. There is large room for technical improvement in cultivation and processing phases, aiming at increasing production efficiency and decreasing the environmental impact. One innovative way of doing this is to optimise the protein production instead of optimising the production of soybean in general.

Soybean yield prediction is currently based on remote sensing technology by analyzing time series of spectral vegetation indices (NDVI). For CYBELE, the aim is to develop methods that include other parameters to increase accurate yield prediction, such as soil analysis and weather data. Using crowd-sourced datasets provided by soybean growers, advanced machine learning algorithms are developed to associate the available data with prediction of protein content. Consequently, insight will be gained in what the field conditions are to produce a high protein yield. The CYBELE solution intends to show the farmers per field heat maps for advice on which seeds to plant, when and which fertilizer to use and when to harvest.

2 Prevent food safety incidents

It is a common belief that Risk assessment is a very critical part of a food safety system in order to prevent food safety incidents in the supply chain. Today, Quality Assurance and Safety Experts that are working in food companies are using risk estimation approaches that are based on static data such as literature and guidelines published by National Authorities. Such risk estimation approaches are not taking into account the emerging and increasing risks of the global supply chain and cannot predict the risk. This results in several serious food safety incidents that may impact public health, can cause large financial loss for farmers and industries and can damage their “brand” and lose customers.

For CYBELE, the intention is to demonstrate the capacity of the HPC solutions proposed in the project for supporting complex highly deep and machine learning prediction models for dairy products, nuts product, sugar, fruits and vegetables. Through data analytics and prediction models, farmers and food industries could minimize mistakes and future risk via good agriculture practices.

3 Prevent frost/hail damage

Horticultural crops, such as apple, citrus, peaches or persimmons trees, are sensitive to frost and hail events and protecting them from the effects of low temperature and hail damage is crucial. Climate conditions influence the probability of occurrence for these events, together with other issues such as vegetation present, topography and soil type with relevance at local scale. Therefore, early warning systems at local scale with a suitable spatial resolution on frost and hail occurrence and their associated risks are relevant for agriculture. Frost and hail forecasts may help farmers to reduce any possible injuries to their crops since protection methods can be used.

CYBELE aims to establish an early warning system that can help farms to prevent damage effects through protective methods for frost and hail. In this case, the focus is on climate predictors that are either correlated with frost or hail occurrence and can then be used for planning risk prevention operations. Temporal series of instability indexes such as pressure, height, temperature, dew temperature, among others, will be analyzed. Models based on machine learning techniques, using as input data the climate instability indices are being built. Validations will be performed against data collected in the field where the fall of hail was verified.

4 Develop agro-robots

Dictated by the weather, farming tasks have often to be carried out within a short time window. Consequently, equipment has increased in size to complete the work rapidly.

One alternative solution is for farmers to manage fleets of smaller, autonomous vehicles and carry out the tasks as required. The range of operations to be delivered include soil chemical analysis, hyperspectral imaging (HSI) of soil/crop condition, real time object level (plant/weed) identification, individual plant harvest readiness assessment (particularly for soft fruits) and plant level automated harvesting, currently not possible because it would be massively labor intensive.

The ultimate goal is for minimally sized equipment like a small tractor or a scouting vehicle to carry the sensory devices. Such sensor ‘transporters’ will be combined with a network of ‘actuator’ devices such as plant level harvesters, precision soil enrichment vehicles or cultivation/planting equipment. It is envisaged that a pair (at least) of systems will operate in tandem on a given task with the sensory elements passing over the crop relaying measurement data to a central location. The data is then processed to identify plant, weed, readiness for harvest etc., generating the inputs for the actuator to harvest the appropriate plant.

5 Optimize crop yields forecast

Crop yield monitoring can be used as a tool for agricultural monitoring (e.g. early warning & anomaly detection), index-based insurance (index estimates) and farmer advisory services to facilitate precision agriculture and helping to provide greater yields and contributing to better food security. These forecasts are traditionally conducted over a consistent grid of 25 25km, or 10*10km grid, on which basis all data are at rest and already available.

With the advance of the big data in agriculture, more data become available on a level of lower spatial resolution, i.e. farmer parcels. This poses big computational challenges, huge processing times and complex architectures for the productivity training model. In this use case the parcel specific data associated with advanced weather forecasts and computations (weather data interpolation, crop growth model) will be prepared for computations on an HPC enabled infrastructure. Also, with additional satellite imagery data, the added value of calculating farmer’s parcel-specific crop productivity estimates is explored.

6 Pig Weighing Optimization

An accurate estimate of the live weight of slaughter pigs is useful to the farmer, especially for knowing when to send the pigs to the slaughterhouse, which can save the farmer a lot of money. Another reason is that it helps to more accurately diagnose and treat diseases. The latter can potentially lead to a lower use of antibiotics, which is important for combating the spread of multi-resistant bacterial strains in farm animals as well as humans. Some large pig producers have staff employed for the sole task of performing manual weightings. This practice is very laborious and time consuming, making it unfeasible for most.

On these more common herds, being able to infer the live weight of the pigs indirectly via e.g. video images would be of great value. In traditional image processing, image features such as lines and edges are extracted based on e.g. sharp colour contrasts. This means that the image processing is very sensitive to variations in lighting conditions, making it less than ideal for applications in variable real-world environments. Deep convolutional neural networks (CNNs) can be used to achieve this goal with greater probability of success, a task however which is very computationally and memory intensive. This use case has as goals: (1) To estimate the mean and standard deviation of the live weight of grower/finisher pigs in a pen based

on video images; (2) To track the weight of individual pigs in a pen based on video images; (3) To incorporate the growth curve estimated by the CNNs in previously developed models for early warning of diarrhea.

7, 8 Sustainable pork meat quality; Improve health and welfare of pigs

Sustainable pig production and global food challenges require producing with optimal productivity, health and welfare of the pigs. The pig farmer is becoming a manager of growing farms with several thousands of fattening pigs. The usage and fusion of all data generated throughout the lifetime and after slaughter is the future way to be able to improve the health and welfare of the pigs. To work on fulfilling the potential of each pig through its life also increases the quality of the end-product for the market and the consumers.

For CYBELE, there are two main goals: (1) improve carcass and meat quality by using and linking on-farm related factors and slaughterhouse data at a large scale. In general, the use case representative aims to bring data and techniques together to enlarge the impact; (2) improve the detection of health, welfare and performance problems at fattening pig farms through better use of available sensor and farm data.

9, 10, 11 Integrate fish fleet data; Optimize fishing vessel on-board data; Machine vision for fish catch detection

During the last decade, fisheries management in the EU increasingly succeeded in rebuilding overfished stocks and preventing overfishing. These successes stem mainly from the increased availability of data and better analysis methods that enabled to assess, and thus provide more precise management for an increasing number of commercially exploited fish stocks. Despite this positive trend, the state of the largest part of the marine ecosystem, including most fish stocks, remains largely unknown causing that little ecosystem-based management has been put in place. Most marine data is collected by means of scientific surveys on research vessels, which is expensive and happens only on small scale. Commercial fishing vessels have a much wider spatiotemporal coverage of the seas, and the increased usage of sensors and IT equipment on board of commercial fishing vessels allow these vessels to collect many data. However, due to the lack of sufficient processing capacity and adequate database systems, fishers nor fisheries managers make optimal use of these data.

Within the CYBELE project, three goals are aimed for, that make up the three use cases respectively: (1) Integrate the data from the digital logbooks of the entire fleet (research and commercial), that comprise daily landing data of commercial fish stocks and location data from the satellite-based vessel monitoring (VMS) system, with environmental data from satellite based imaginary systems and data collected specifically for fisheries management purposes. By means of data mining models that require training on appropriate computer hardware, managers can optimize the quota uptake of the fleet, and fishers can use more precise information about the location of hotspots of fish; (2) Optimize the database system on board of a particular commercial Belgian fishing vessel who has advanced data collection systems but no integrated system to process these data; (3) Visual-based processing of the catch using an RGB camera. Currently, catches are sorted out manually after being discharged on a conveyor belt. This is very labour-intensive sorting process causes that only the commercial part of the catch is reported while no information is collected on the part of the catch that is thrown back in sea. Detection of species, including undersized fish, non-commercial fish and non-commercial benthic animals, through implementing camera technology on the conveyor belt may fill this gap.

12 Aquaculture monitoring

Aquaculture is probably the fastest growing food-producing sector and now accounts for more than 50 percent of the world's fish that is used for food. One of the main issues in commercial aquaculture is the lost food when the fish are fed. This not only increases the cost of the produced fish (feed cost is a major cost component that accounts for approximately 70% of the operational expenses of the farm) but furthermore, this wasted food is deposited in the seabed and generates an environmental impact on the surrounding area. It also causes failures to comply with EU legislation. Another challenge is maintaining the farm in a good condition. If the cages are not in the correct positions, have deformations, anti-bird nets not placed correctly, etc. this usually leads to damages, financial losses and uncontrolled escapes to the environment.

The CYBELE project will optimize feeding, evaluate impact on the environment and evaluate the status of the infrastructure in open sea aquaculture. It will make use of image processing technology in order to process aerial images of fish farms taken from drones. This will be combined with other data such as weather information and sensor measurements (mainly related to Oxygen and current speed) and machine learning methods.

References

Allen, D.W., Lueck, D., 1998. The nature of the Farm*. The J. Law Econ. 41 (2), 343–386.

Ang, L.M., Seng, K.P., 2016. Big sensor data applications in urban environments. Big Data Res. 4, 1–12. https://doi.org/10.1016/j.bdr.2015.12.003.

Astill, J., Dara, R.A., Fraser, E.D.G., Roberts, B., Sharif, S., 2020. Smart poultry management: Smart sensors, big data, and the internet of things. Comput. Electron. Agric. 170, 105291. https://doi.org/10.1016/j.compag.2020.105291.

Athanasiadis, I., Osinga, S., Paudel, D., Mouzakitis, S., Ntanopoulos, S., Pelekis, S., Tsitouras, S., 2020. CYBELE deliverable D1.3 requirements, methodology and MVP (Version b), technical report, CYBELE project public deliverable. Ref.Ares 458577, 2020 available online. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5cb6a0a03&appId=PPGMS.

Bundy, J., Vogel, R.M., Zachary, M.A., 2018. Organization-stakeholder fit: A dynamic theory of cooperation, compromise, and conflict between an organization and its stakeholders. Strateg. Manag. J. 39 (2), 476–501. https://doi.org/10.1002/smj.2736.

Chavas, J.-P., Chambers, R.G., Pope, R.D., 2010. Production economics and farm management: a century of contributions. Am. J. Agric. Econ. 92 (2), 356–375.

Cockburn, M., 2020. Review: application and prospective discussion of machine learning for the management of dairy farms. Animals *10* (9). https://doi.org/10.3390/ani10091690. Article 1690.

Coleman, S., Gob, R., Manco, G., Pievatolo, A., Tort-Martorell, X., Reis, M.S., 2016. How can SMEs benefit from big data? challenges and a path forward. Qual. Reliab. Eng. Int. 32 (6), 2151–2164. https://doi.org/10.1002/qre.2008.

Cui, Y.S., Kara, S., Chan, K.C., 2020. Manufacturing big data ecosystem: A systematic literature review. Robot. Comput. Integr. Manuf. 62 https://doi.org/10.1016/j.rcim.2019.101861. Article 101861.

EARTO, 2014. The TRL Scale as a Research & Innovation Policy Tool, EARTO Recommendations. *European Association of Research and Technology Organisations*. https://www.earto.eu/event-calendar/earto-innovation-school-on-innovation-and-trls/.

EU Earth Observation programme, 2021. Copernicus: Europe's eyes on Earth. https://www.copernicus.eu/en/access-data.

Eurostat, 2018. Statistics Explained - Small and large farms in the EU - statistics from the farm structure survey.

Hayter, C.S., Rasmussen, E., Rooksby, J.H., 2020. Beyond formal university technology transfer: innovative pathways for knowledge exchange. J. Technol. Transfer. 45 (1), 1–8. https://doi.org/10.1007/s10961-018-9677-1.

Huirne, R.B.M., 2002. Strategy and risk in farming. NJAS: Wageningen J Sci. Royal Soc. Agricult. Sci. 50 (2), 249–259. https://edepot.wur.nl/30955.

Ip, R.H.L., Ang, L.-M., Seng, K.P., Broster, J.C., Pratley, J.E., 2018. Big data and machine learning for crop protection. Comput. Electron. Agric. 151, 376–383. https://doi.org/10.1016/j.compag.2018.06.008.

Jagadish, H.V., 2015. Big data and science: myths and reality. Big Data Res. 2 (2), 49–52. https://doi.org/10.1016/j.bdr.2015.01.005.

Jakku, E., Thorburn, P.J., 2010. A conceptual framework for guiding the participatory development of agricultural decision support systems. Agric. Syst. 103 (9), 675–682. https://doi.org/10.1016/j.agsy.2010.08.007.

Jayashankar, P., Johnston Wesley, J., Nilakanta, S., Burres, R., 2019. Co-creation of value-in-use through big data technology- a B2B agricultural perspective. J. Bus. Ind. Mark. 35 (3), 508–523. https://doi.org/10.1108/JBIM-12-2018-0411.

Kamble, S.S., Gunasekaran, A., Gawankar, S.A., 2020. Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications. Int. J. Prod. Econ. 219, 179–194. https://doi.org/10.1016/j.ijpe.2019.05.022.

Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Comput. Electron. Agric. 143, 23–37. https://doi.org/10.1016/j.compag.2017.09.037.

Kempenaar, C., Lokhorst, C., Bleumer, E., Veerkamp, R., Been, T., Evert, F., Noorbergen, H., 2016. Big data analysis for smart farming: Results of TO2 project in theme food security.

Klerkx, L., Jakku, E., Labarthe, P., 2019. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS - Wageningen J. Life Sci. *90-91*, 100315. https://doi.org/10.1016/j.njas.2019.100315.

Kuo, Y.H., Kusiak, A., 2019. From data to big data in production research: the past and future trends. Int. J. Prod. Res. 57 (15–16), 4828–4853. https://doi.org/10.1080/00207543.2018.1443230.

Lane, J.P., 1999. Understanding technology transfer. Assist. Technol. 11 (1), 5–19. https://doi.org/10.1080/10400435.1999.10131981.

Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. *https://www.gartner.com/en/blog*.

Lezoche, M., Hernandez, J.E., Diaz, M., Panetto, H., Kacprzyk, J., 2020. Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. Comput. Ind. 117 https://doi.org/10.1016/j.compind.2020.103187. Article 103187.

Li, J.P., Xie, Y.J., Wu, D.S., Chen, Y.P., 2017. Underestimating or overestimating the distribution inequality of research funding? The influence of funding sources and subdivision. Scientometrics 112 (1), 55–74. https://doi.org/10.1007/s11192-017-2402-2.

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J., 2016. Analysis of Big Data technologies for use in agro-environmental science [Article]. Environ. Model Softw. 84, 494–504. https://doi.org/10.1016/j.envsoft.2016.07.017.

Martin, G., 2015. A conceptual framework to support adaptation of farming systems - development and application with Forage Rummy. Agric. Syst. 132, 52–61. https://doi.org/10.1016/j.agsy.2014.08.013.

Mir, S.A., Padma, T., 2017. Generic Multiple-Criteria Framework for the development of agricultural DSS. J. Decis. Syst. 26 (4), 341–367. https://doi.org/10.1080/12460125.2018.1437501.

Mouzakitis, S., Tsapelas, G., Pelekis, S., Ntanopoulos, S., Askounis, D., Osinga, S.A., Athanasiadis, I.N., 2020. Investigation of common big data analytics and decision-making requirements across diverse precision agriculture and livestock farming use cases. In: International Symposium on Environmental Software Systems, pp. 139–150. Wageningen.

Nandyala, C.S., Kim, H.K., 2016. Big and meta data management for U-Agriculture mobile services [Article]. Int. J. Software Eng. Applicat. 10 (2), 257–270. https://doi.org/10.14257/ijseia.2016.10.2.21.

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O., 2015. Big Data challenges in building the global earth observation system of systems [Article]. Environ. Model. Softw. 68, 1–26. https://doi.org/10.1016/j.envsoft.2015.01.017.

Nature Food, 2020. Systems thinking, systems doing. Nat. Food 1 (11), 659. https://doi.org/10.1038/s43016-020-00190-9.

Newton, J.E., Nettle, R., Pryce, J.E., 2020. Farming smarter with big data: Insights from the case of Australia's national dairy herd milk recording scheme [Article]. Agric. Syst. 181 (13) https://doi.org/10.1016/j.agsy.2020.102811. Article 102811.

Oussous, A., Benjelloun, F.Z., Lahcen, A.A., Belfkih, S., 2018. Big data technologies: A survey. J. King Saud Univers. Computer Informat. Sci. 30 (4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001.

Penn, L., Goffe, L., Haste, A., Moffatt, S., 2019. Management information systems for community based interventions to improve health: qualitative study of stakeholder perspectives. BMC Public Health 19 (1), 105. https://doi.org/10.1186/s12889-018-6363-z.

Perakis, K., Lampathaki, F., Nikas, K., Georgiou, Y., Marko, O., Maselyne, J., 2020. CYBELE – Fostering precision agriculture & livestock farming through secure access to large-scale HPC enabled virtual industrial experimentation environments fostering scalable big data analytics. Comput. Netw. 168, 107035. https://doi.org/10.1016/j.comnet.2019.107035.

Peregud, I., Zharovskikh, A., 2020. Computer vision applications: examples across different industries. Retrieved April 1 from. https://indatalabs.com/blog/applications-computer-vision-across-industries.

Pylianidis, C., Osinga, S., Athanasiadis, I.N., 2021. Introducing digital twins to agriculture. Comput. Electron. Agric. 184, 105942. https://doi.org/10.1016/j.compag.2020.105942.

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C.-A., Sánchez-Cervantes, J.L., Cervantes, J., García-Alcaraz, J.L., Alor-Hernández, G., 2016. A general perspective of Big Data: applications, tools, challenges and trends. J. Supercomput. 72 (8), 3073–3113. https://doi.org/10.1007/s11227-015-1501-1.

Saggi, M.K., Jain, S., 2018. A survey towards an integration of big data analytics to big insights for value-creation. Inf. Process. Manag. 54 (5), 758–790. https://doi.org/10.1016/j.ipm.2018.01.010.

Saiz-Rubio, V., Rovira-Mas, F., 2020. From Smart farming towards agriculture 5.0: a review on crop data management. Agronomy-Basel *10* (2). https://doi.org/10.3390/agronomy10020207. Article 207.

Sawant, M., Urkude, R., Jawale, S., 2016. Organized data and information for efficacious agriculture using PRIDE (TM) model [Article]. Int. Food Agribusiness Manag. Rev. *19* (A), 115–130. <Go to ISI>://WOS:000422721200009.

Sonka, S.T., 2016. Big data: Fueling the next evolution of agricultural innovation. J Innovat Manag. 4 (1), 114–136.

Sutherland, L.-A., Burton, R.J.F., Ingram, J., Blackstock, K., Slee, B., Gotts, N., 2012. Triggering change: Towards a conceptualisation of major change processes in farm decision-making. J. Environ. Manag. 104, 142–151. https://doi.org/10.1016/j.jenvman.2012.03.013.

Van Meensel, J., Lauwers, L., Kempen, I., Dessein, J., Van Huylenbroeck, G., 2012. Effect of a participatory approach on the successful development of agricultural decision support systems: The case of Pigs2win. Decision support systems 54 (1), 164–172. https://doi.org/10.1016/j.dss.2012.05.002.

Verdouw, C., Sundmaeker, H., Tekinerdogan, B., Conzon, D., Montanaro, T., 2019. Architecture framework of IoT-based food and farm systems: A multiple case study [Article]. Comput. Electron. Agric. *165*, 26. Article 104939. https://doi.org/10.1016/j.compag.2019.104939.

Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J., 2017. Big data in smart farming – a review. Agric. Syst. 153, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023.

Wysel, M., Baker, D., Billingsley, W., 2021. Data sharing platforms: How value is created from agricultural data. Agric. Syst. *193* https://doi.org/10.1016/j.agsy.2021.103241. Article 103241.