Photo: Have a nice day (AdobeStock)

2

Digital agriculture
Definitions and technologies

Kleber Xavier Sampaio de Souza | Stanley Robson de Medeiros Oliveira | Carla Geovana do Nascimento Macário | Julio Cesar Dalla Mora Esquerdo | Maria Fernanda Moura | Maria Angelica de Andrade Leite | Helano Póvoas de Lima | Alexandre de Castro | Sônia Ternes | Inácio Henrique Yano | Edgard Henrique dos Santos

Introduction

Advances in information processing and in the areas of nanotechnology, biotechnology, and cognitive science are promoting a convergence between sciences, which is currently called Nano-bio-info-cogno. The report commissioned by the National Science Foundation of the United States named *Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology and Cognitive Science* (Roco; Bainbridge, 2003), was prepared by more than 100 scientists and pointed out the synergy between nanotechnology, biotechnology, information technology, and cognitive science as the segment with the greatest potential for advancement in innovation. This report highlights that systemic approaches using mathematics and computing will, for the first time, enable us to understand the functioning of complex systems in the natural world such as the human mind, stellar explosions, social interactions, organs of the human body, and the natural phenomena involved with agriculture.

Agriculture works directly with three of these areas, nanotechnology, biotechnology, and information technology. In fact, it has been influenced and fueled by the staggering growth of data acquisition capacity coming from different sources. This data ranges from the cell scale, such as information obtained by analysis in the field of "omics" sciences (genomics, proteomics, transcriptomics and metabolomics), to the macroscopic scale, which includes socioeconomic data and data obtained through remote sensing devices, such as satellites. At the local scale, this can also come from farms equipped with agricultural equipment and sensors.

Digital agriculture is an increasingly connected and remote operation that surveys and processes large amounts of data collected in all links of the production chain: pre-production, production and post-production. This involves different types of digital technologies: sensors embedded in orbital, suborbital, airborne, or autonomous systems (drones, agricultural machines) which can be installed directly in the field or in different 'things' (Internet of Things – IoT) along the production chain. The technologies involve telecommunication systems, global positioning, control, management, analysis software (data analytics), and actuators.

Data from these technologies are now collected not only by conventional means, but also from collaborative platforms or social media (citizen science), among others. Their accumulation poses a challenge to storage, search, and retrieval systems, while also impacting processing and retrieval methods.

The abundance of data creates a great gap in terms of management and analysis capacity, and consequently in terms of the production of knowledge from them. This precipitates into a complex scenario, where transforming data into information and knowledge assumes a strategic role in all sectors of the economy, including agriculture, a strategic sector for Brazil. All these data need to be integrated, preprocessed, and analyzed so that the required knowledge to establish digital agriculture can be extracted.

This chapter presents the concepts related to digital technologies that are used throughout this book. This is done in a consolidated way in order to facilitate the readers' understanding and access.

Digital technologies

The digital technologies presented here are divided into five groups. In the first group there are technologies linked to the organization and representation of information. In the second there are the mathematical and statistical modeling techniques involving biological, social, and environmental phenomena. In the third, the application of artificial intelligence in agriculture. In the fourth group, sensor and robotics technologies. In the fifth and last group there are technologies in which applications interact with agriculture, such as cloud computing and blockchain.

Organization, representation, and information access

The volume of information and diversity of formats (DNA data, satellite images, sensor data) in which this information is presented represent an enormous challenge to its organization and reuse. It becomes necessary to annotate, classify, structure and provide access mechanisms so that information can be found and reused in the future, which is the purpose of the technologies in this section.

Thesaurus – According to ANSI/NISO Z39.19-2005 (National Information Standards Organization, 2010), thesauri are controlled vocabularies, arranged in such a way that the relationship between their terms is clearly identified and standardized. Terms are composed of one or more words and selected from natural language to be included in a thesaurus. In the National Agricultural Thesaurus (Thesagro), for example, the word *mite* is related to the word *Arachnid*, so that *Arachnid* is the broader term (BT) and mite is the more specific term (NT - Narrower Term). There are also other terms subordinate to *Arachnid* and which, therefore, are NT of *Arachnid*, such as *spider* and *scorpion*. BT and NT are forms of vertical relationship between terms used in thesauri; there are also horizontal associations between terms, expressed as a related term (RT - Related Term). For example, in Thesagro there are the terms *Acaricide* and *Tick* associated with the RT of *Mite*. The first is a counter agent for mites, belonging to the hierarchy that starts with *Pesticide*, and the latter belongs to the hierarchy of *Animal Parasite*. Thesauri, therefore, form a web of relationships between terms, and this web helps to find the information one is looking for. The terms and their hierarchy can be used to organize the content of websites on the internet and to expand the searches that are performed on a determined content. For example, when searching for *Acarus*, documents that speak of *tick* or *Acaricide* can also be retrieved.

Ontology – An ontology is formally defined as a common vocabulary for sharing information about a particular knowledge domain. The ontology includes machine-interpretable definitions of the basic concepts of this domain and the relationships between these concepts (Noy; Mcguiness, 2001).

In ontologies, the relationships between domain concepts are made explicit so that they can be interpreted by computers. Each concept contains its attributes, for which there are possible values. For example, the concept "vehicle", which is a class, contains the subclasses "car" and "motorcycle". A car usually has four wheels and a motorcycle has two. So, the attribute number of wheels would be four for a vehicle and two for a motorcycle. Both the motorcycle and the car have a manufacturer attribute and any other attributes one would want to add and enrich the information contained in the ontology for its reuse. As ontologies provide a common machine-processable language, an agent can automatically browse several websites that work on the same subject – for example car parts – and add the information provided by them for price comparison. This is greatly facilitated when multiple sites use the same ontology to describe their parts.

Big Data – The term Big Data includes data sets where sizes go beyond the capacity of data management systems to process them. It is usually data from various sources, such as mobile devices, body sensors, social media, emails, electronic medical records, genomics and geospatial sensor data, among many others. This variety of sources, the amount of data, and the speed at which the data arrives for processing generate what is called the "three Vs" of Big Data: volume, velocity and variety, to which sometimes "veracity" and "value" are also added. The definition encompasses structured, semi-structured, and unstructured data, although the treatment of unstructured data by systems that process Big Data is much more common (Dedić; Stanier, 2017). Big Data applications appear all the time: when analyzing posts on social networks about a certain subject to see their repercussion; when analyzing Google searches to identify outbreaks of flu pandemics. Given the inadequacy of traditional database management systems in processing Big Data, solutions were developed by companies that traditionally operate with large volumes of data, such as Google and Cloudera, which developed MapReduce, Flume, and Sqoop. MapReduce (Dean; Ghemawat, 2008) is an algorithm developed by Google with a free implementation developed by the Apache Foundation, called Hadoop (White, 2012). It operates by distributing large datasets to be processed on multiple computers in parallel (possibly thousands of computers) and then consolidating answers. Apache Flume[1] was originally developed by Cloudera to manage large volumes of log file data, but it has been extended to process events from web sources such as Twitter and Facebook. Apache Sqoop[2] is a tool to efficiently transfer data between structured, semi-structured, and unstructured data sources. It is an interesting tool for bringing data from external sources, such as relational databases, into Hadoop's distributed file system. MapReduce, Flume, and Sqoop are just examples of systems that were developed to handle Big Data, and they are not the only systems capable of handling data with large volume, variety, and velocity.

API – An Application Programming Interface (API) is a way for two applications to talk to each other. A requesting application triggers the execution of another so that its own task is completed, in other words, the requesting application needs the second one as a provider for its functioning. The communication intermediary between the two applications is the API, which defines protocols, routines and tools so that the message is delivered to the provider application and the response returns to the requesting

[1] Available at: http://flume.apache.org

[2] Available at: https://sqoop.apache.org

application. A Web API operates on the internet using the usual protocols for exchanging information, such as HTML, XML, or JSON. As an example of API applied to the agricultural area, we can mention the AgroAPI platform from Embrapa Agricultural Informatics (2020). It provides a series of information and models that can be coupled with software, web systems, or mobile applications from companies, including startups, public, and private institutions. Each API is allowed free use of up to 1,000 requests per month. An example is API Agritec, part of AgroAPI, which gathers information on planting time, fertilization, productivity, agricultural zoning and cultivars for five agricultural crops. Another example is the SATVeg API, which uses satellite data to generate the visualization of the evolution over time of NDVI and EVI vegetation indices for all of South America. These indices make it possible to observe variations in green biomass on the land surface, and may help implementing the Brazilian Forest Code (Código Florestal), or monitoring the cycle of certain agricultural crops, among other land cover and land use dynamics.

Mathematical modelling and statistics

The representation of natural phenomena through models is an integral part of the scientific method. This section is dedicated to showing the representation through models used in the scientific method. It also conceptualizes Data Science, which emerged from a confluence of various branches of expertise to extract knowledge from increasingly abundant masses of data.

Mathematical model – A model arises from the need to understand a phenomenon in the physical world and predict its behavior in a given situation. A model is always an abstraction of what happens in the real world, a simplification of what happens in reality so that a system can be understood and quantified (Torres; Santos, 2015). In Bassanezi (2002), a mathematical model consists of transforming reality into mathematical problems, which are solved and interpreted in light of what happens in the real world. Building a mathematical model involves several steps: a) conceptualization, which occurs after initial observations about a problem, formulation of hypotheses to explain its functioning, and a first selection of which variables, processes, and interactions are considered relevant. An important task occurs during the conceptualization, which is the simplification of the model in terms of variables and interactions that are essential for the representation of the problem, since the phenomena of the natural world, especially the biological ones, are extremely complex; b) mathematical formalization, which is the translation of the problem into mathematical language. There are many different approaches to this translation, such as differential equations, Bayesian equations, stochastic systems, finite difference equations, and agent-based systems, each with its advantages and limitations. Its choice depends on the nature of the problem that is being modeled; c) parameter estimation, which involves discovering which numerical values are guiding the elaborated mathematical formulation. These parameters can be obtained through experimental measures, and the adoption of experimental statistical techniques adds greater reliability; d) simulation and prediction, which is the moment when the system of equations is solved analytically, or the model is run on a computer. As biological problems usually involve control and regulation mechanisms, the analytical solution of the models is almost always impossible, which makes the computational approach the most frequent to solve a model; e) model validation, in which the system response is verified for each scenario of the input values of the model variables. This answer has to coincide both in terms of the trajectory of the system and with the values obtained in the experimental measurements. Therefore, it is at this point that one evaluates how close the model is representing reality while its accuracy is being measured. Another desired feature of the model is its capability to predict new facts and unknown relationships that can be verified in the real world; f) model refinement, where the validity of the model results is criticized in terms of the trajectories of the modeled system when confronted with the real world while its accuracy is evaluated. When models deviate from what was expected, it may be due to some hypotheses that either have not been considered or that are false. There may also have been an error in obtaining the data that fed the construction of the model or an error in its mathematical formulation. In this case, new hypotheses and/or new variables and a re-verification of the mathematical model may be proposed.

Statistical model – Statistics is the basis of the scientific method, which can be summarized as follows: a) definition of the problem to be studied; b) formulation of one or more hypotheses to be tested; c) conducting experiments to test the formulated hypotheses; d) statistical analysis of the data obtained; e) interpretation of results and obtaining conclusions, that is, obtaining a descriptive or inferential statistical model that proves or not the original hypotheses. As example (Snedecor; Cochran, 1967): the problem to be studied was the variability in the calcium concentration of turnips; the hypotheses concerned the behavior of the variability of calcium in plants and, specifically, in the leaves of each plant; in the experiment, four plants were randomly chosen, and then three leaves of each plant were randomly selected. Two 100 mg samples were taken from each leaf, determining the amount of calcium in each sample through microchemical processes ; the data were submitted to an analysis of variance according to the model of the formulated hypotheses; the analysis concluded that, statistically, at a 5% significance level, the variability in the leaves of each plant is more important than the variability in the whole plant, and that the ideal model (the hypotheses raised) statistically represents reality. Each of these variability effects is estimated according to the initial hypotheses, the calculated estimates show whether the model is adequate or not to the formulated hypotheses based on an accepted margin of error, which in this case was 5%. In general, biological processes are inherently complex and the variability of each observed factor needs to be estimated. That means that the number of observed variables is enormous, and sometimes not all variables are known. Those not known will introduce a greater error to the estimated model; remembering that the model is accepted after the estimates are statistically proven, and that the model as a whole has an error, which is also estimated. In this scenario, and in many others, such as general gas theory or natural selection (Fisher, 1934), the arguments are built on statistical grounds.

Data Science – Data Science is an interdisciplinary field focused on the processes and systems for extracting knowledge or insights from data in various forms, structured or not. It incorporates techniques and theories from the most diverse areas of knowledge such as computing, engineering, mathematics, statistics, economics, data mining, and artificial intelligence in order to collect, process, integrate, and analyze data so as to create data products and services (Amaral, 2016). Data Science is not restricted to analyzing large volumes of data (Big Data analytics). Small (Small Data) and large (Big Data) data repositories are important aspects of this research area. Small Data includes simple information, which is in the database of any company or small rural property. Small Data includes research results, consumer or rural producer data, data on agricultural properties, e-mails with information on management practices, data containing agricultural production volume per period, among others. It usually consists of structured data ready for analysis. Big Data, on the other hand, refers to (mainly) unstructured data, originating from multiple sources and which should be collected, aggregated, and analyzed in order to generate managerial information. Among the Data Science applications are digital marketing, which elaborates personalized advertisements based on information obtained through user profiles and their browsing history in companies. Further examples are recommendation systems, which are based on the pattern of pages visited or products purchased to suggest new products, and bank customer credit rating systems, which consults track record and existing scores in credit protection companies to calculate the likelihood of a customer defaulting.

Artificial intelligence

Pattern recognition and machine learning technologies, including deep learning, are an integral part of the many systems used today, such as autonomous cars and voice recognition systems. These new technologies analyze large data sets and learn patterns from them that allow, for example, to identify objects or anticipate the next word to be spoken in a sentence. On the other hand, fuzzy logic is used when the rules of a system are explained directly, without the use of machine learning, but still allows a certain degree of imprecision.

Pattern recognition – A pattern, as understood within the concept of pattern recognition, can be the representation of a handwritten number, a number written on a house, an orange, a car, a pronounced word, sequences of temperature measurements, pressure and rain, stock market value sequences, as well as many other things we want a computer system to learn to recognize. It is for this reason that many of the important pattern recognition problems can be characterized either as waveform classifications (sounds, temperature measurements, action values, etc.) or as classification of geometric figures, as with images (Fukunaga, 1990). Our brain is specially designed to recognize patterns. In the very early years of our existence, we learned to differentiate sounds, words, what is a cat and how it is different from a dog and so many other things. What pattern recognition in a computational system should achieve is to learn to differentiate the data presented to it, an activity that is computationally known as classification. To process this data, several techniques are used for pattern recognition, such as decision trees, random forests, k-nearest neighbors, support vector machines and neural networks (Bishop, 2006). The application of pattern recognition techniques can indicate, for example, that a given sequence of temperature values is within normality, that a stock listed on the stock exchange is on a downward trajectory, that the handwritten number on a paper is 3, or that the object in a certain position in an image is an orange.

Machine learning – This is a process closely related to pattern recognition (previous topic), as what is desired during machine learning is for the computer to learn from the patterns presented to it. According to Bishop (2006), machine learning and pattern recognition are two facets of the same field of knowledge, with pattern recognition originating from engineering and machine learning from computing. For this reason, the algorithms between pattern recognition and machine learning are also shared. Generally, it is possible to divide machine learning into supervised, when starting from a previously defined set of labeled data one wants to find a function that is capable of predicting unknown labels; and unsupervised, which seeks to identify groups or patterns from the data, without a specific objective to be achieved (Russel; Norvig, 2020). These two concepts are defined as follows.

Unsupervised learning – In this type of learning, the data set used does not have any type of label. The objective of this type of learning is to detect similarities and anomalies between the analyzed objects. The process of grouping objects into similar classes is called clustering. This procedure is also known as data segmentation, as it partitions large datasets according to the similarity between subsets. The objects that are more similar to the characteristics imposed by the domain must be allocated in the same group, while those less similar must be allocated in different groups. The similarity between objects must be obtained by algebraic measures, such as the Euclidean distance, for real values; or by simple correspondence, for nominal values. These algorithms can be divided into two more general classes, according to the heuristic used to construct the groups.

The first class refers to partitional algorithms, which, usually with linear execution computational cost, operate in an iterative way based on the previous definition of the desired number of groups and the definition of representative objects of each group, known as centroids. In each iteration, each object is associated with the centroid and, consequently, with its most similar group. The centroids of the groups are then recalculated for the next iteration. The algorithm reaches its point of convergence when the centroids are no longer changed between one iteration and another, that is, when the groups are well defined, considering the similarity measure used. The k-means is in this subclass of algorithms (Macqueen et al., 1967), and is considered one of the ten most influential algorithms in data mining (Wu, 2008).

The second class is hierarchical algorithms, which have a computational cost of execution that is normally quadratic and therefore do not require the identification of initial representatives or the desired number of groups. Thus, in a single execution, *n* nested partitions can be generated for the same set of *n* instances, containing from 1 to *n* groups each, constituting a cluster hierarchy (Han; Kamber, 2006). Two distinct strategies can be used to build this hierarchy: the agglomerative one, which initially considers each instance of the dataset as a group, merging pairs of groups in each iteration; and the divisive, which initially considers all samples belonging to a single group, dividing them into smaller groups in each iteration (Hastie et al., 2009).

Supervised learning – The process of supervised machine learning consists of presenting a large amount of previously classified data to a computer and making it learn from that data. Learning happens by modifying system parameters as more and more examples are presented to it. These parameters are numbers for which their values are unknown. So, the learning task is to find out which values make the system get it right most often. For each example, it is verified whether the system learned to correctly classify that example. If it is right, the system reinforces the parameters that allowed this correct classification through their weights. Otherwise, it calculates what correction the system must undergo so as not to make this error and negatively adjust the weights that led to that wrong answer. One can then imagine a system with many knobs or buttons that have to be turned just the right amount for that system to finally get the answer right. However, instead of turning the knobs ourselves, we have algorithms that do it in a controlled way in order to make learning happen. The main learning paradigms can be listed as follows:

a) **Symbolic (decision trees):** a decision tree is a flowchart-like structure in which each internal node represents a "test" in an attribute, each branch represents the test result, and each leaf node represents a class label (decision made after computing all attributes). The paths from root to leaf represent classification rules (Quinlan, 1986).

b) **Based on instances (*k*-NN or k nearest neighbors):** the main idea of *k*-NN is to determine the classification label of a sample based on the *k* neighbor samples from a training set. Among the *k* examples, there is the most frequent class. This class is attributed to the new example (Fukunaga; Narendra, 1975).

c) **Based on statistical learning (Support Vector Machines – SVM):** the simplest way to partition an *n*-dimensional Euclidean space is through hyperplanes. The SVM classifier is also based on this strategy, however, it uses a special type, the optimal separating hyperplane. It is a hyperplane that divides classes, maximizing the margin of separation between them (Vapnik, 1995, 1998).

d) **Committee-based:** it is the field of machine learning that builds a group of classifiers, called base classifiers, in order to be more accurate than the best elements of the group. The simplest approach based on this algorithm is simple majority voting, in which several classifiers are combined into one voting strategy. As a result, the response that receives the highest number of votes is considered the committee's response (Han; Kamber, 2006). Random Forest is an example of this type of approach. It is a classification and regression technique that consists of a set of decision trees combined to solve classification problems (Breiman, 2001).

e) **Connectionist (Artificial Neural Networks – ANN):** they are computational models inspired by the central nervous system (particularly the brain), capable of performing machine learning, as well as pattern recognition. An example of a connectionist model is the deep learning technique, detailed as follows.

Deep learning – The deep learning technique, or deep neural networks, is a machine learning technique in which the model chosen for the learning algorithm is an artificial neural network with many layers. Neural networks were inspired by the way neurons function in biological systems, operating in a parallel and decentralized way (Marblestone et al., 2016). Typically, a neural network can contain more than 100 layers, arranged one after the other or in parallel. Each of these layers is composed of one or more neurons, interconnected so that the result of neurons that are in one layer feeds the input of neurons that are in the posterior layer. The neural network training method often employs an algorithm called backpropagation. Since it is associated with each neuron, there is a weight that that neuron represents in the response, this algorithm compares the system response with the value that should have been and distributes the error by recalculating the values of the weights of the neurons backwards. There are many neural network architectures available, such as feedforward networks, convolutional networks, recurrent networks and restricted Boltzmann machines, among many others (Goodfellow et al., 2016). The architecture chosen for the network depends on the problem to be solved: forward connected networks are used in both classification and regression problems; convolutional networks, for image classification problems; recurrent networks, for problems involving sequences, such as natural language processing; and the restricted Boltzmann machines are applied for dimensionality reduction, a task with a large amount of input variables, and the most significant ones are identified. This list of problems for each network is not exclusive, it is only to serve as an example, as a constrained Boltzmann machine can be used to solve other problems, such as regression and classification, like with other neural network architectures. The deep learning area includes a great deal of art involved in selecting a given architecture for a problem, as well as in the parameterization of models.

Fuzzy sets and fuzzy logic – The classical set theory defines a class of objects with binary membership as a set, that is, each element may or may not belong to the set (∈ or ∉). Zadeh (1965) founded the concept of fuzzy sets (FS) as a class of objects in which each element has a continuous degree of membership, admitting any value between zero and one. This concept allows addressing real-world problems, where membership criteria and boundaries between classes are not precisely defined (that is, they are fuzzy). An element can have different "degrees of membership" for various sets. Analogous to the theory of classical set, a whole class of logical operations is derived from fuzzy sets, called fuzzy logic. Those that operate with fuzzy logic are called Fuzzy Rules Based Systems (FRBSs). They are inference systems whose logical components are expressed through FS. Typically, it is composed of a fuzzy database (input and output variables), an inference mechanism and a fuzzy rule base of the "IF A then B" type, whose linguistic terms are FS (Klir; Yuan, 1995).

Earth and study sensors

Sensors and actuators are at the heart of digital agriculture, as they enable to perceive what is happening in the environment, and take appropriate actions. Sensors can be orbital, such as satellites, which allow the collection of geospatial data, or proximal data, such as sensors installed on rural properties and linked to the Internet of Things devices. When computing is fully integrated with the sensors of an environment and distributed in that environment, we have Ubiquitous Computing.

Ubiquitous computing – The term Ubiquitous Computing was proposed by the scientist Weiser (1991), of Xerox's Palo Alto Research Center (PARC), to refer to a computing paradigm proposed for the 21st century. In this paradigm, computing should be everywhere, hence the term ubiquitous, and invisible to its user. To explain the concept, Weiser considers that written language was the first ubiquitous technology, because before it, information was restricted to people's memories. With his invention, anyone who knows how to read is able to understand what is written, therefore, independent of the memory of the person who wrote it.

The concept of computing everywhere is different from taking a notebook anywhere, because even at that point what you take with you is computational power and the focus therefore is on the computer. With ubiquitous computing, computers operate at a distance, with no physical contact with users. The interaction with these computers in the environment could be done by recognition of presence, voice and gestures by sensors installed in the environment, displays and projectors. Ubiquitous computing also implies more intelligence on the part of the computer, as its sensors would have to perceive what is happening in the environment and take actions to facilitate the task of users who are in it, activating services, without the user having required them. For example, when entering your office and looking for a certain document saved on paper, the system would point out where you have stored that document in the past. The system could also bring the project you were working on into a meeting room so it could be presented. Obviously, ubiquitous computing would present new challenges in terms of privacy and security, as the first example means that the system was watching your every step when you saved that document in the past, while the second means that the system would have access to all your files and would transfer only the files needed for the presentation. In addition to privacy and security, there are also other challenges, such as bringing together pieces of hardware and software from multiple manufacturers whose software would have to be integrated and driven by a larger system. Although there is no system that fully implements the idea of ubiquitous computing, some technologies try to approach this ideal, such as speaker systems that hear what is being said and perform tasks such as adjusting the lighting, playing a favorite song or perform an internet search. In agriculture, the concept of ubiquitous computing has been used in the application of agrochemicals. In this application, existing sensors close to the leaves would guide the electronics embedded in the sprayers in order to control the greatest possible coverage, using the least amount of liquid.

IoT – The internet of things (IoT) is defined by the International Telecommunication Union (ITU 2012) as a global infrastructure for information society, enabling advanced services through the interconnection of things (physical and virtual), based on interoperable information and communication technologies, whether these structures are in place or evolving. From the point of view of the internet of things, the ITU defines things as objects in the physical or virtual world that can be identified and integrated into communication networks. Virtual objects are included in the IoT through physical things linked to devices, which in turn have mandatory communication capabilities. Communication between devices can be performed through a communication network (with or without an intermediary gateway) or directly between devices, without a communication network, in the latter case, direct communication between devices is required. When communication between devices takes place through a gateway, it must provide at least two network technologies, either to integrate devices, such as ZigBee, Bluetooth, Wi-Fi or LoRa, or to integrate devices to the communication network, such as 2G, 3G, LTE, satellite networks or others. Devices also have the sleep mode and automatically return to save energy. This capability is especially important for sensors that are installed in remote locations that do not have direct connection to electricity, as is the case with some agricultural monitoring sensors. Objects connected to the IoT network can range from people or animals with RFID tags to pacemakers and other hospital devices for individual use, agricultural implements, cell phones, surveillance cameras, humidity and atmospheric pressure sensors, rain gauges, cars with embedded sensors and many others. All things connected to the IoT are required to have an internet address, that is, an IP address. With this address, things can be accessed by any machines connected to the internet. This access at any time makes things connected to the IoT vulnerable in two points: security and privacy. Concern with security creates the need to implement requirements to ensure the confidentiality and integrity of information, both in the data and in the services that process this data. The issue of privacy also needs to be supported by the IoT, as the data that travels through the IoT system can transit sensitive information linked to the owners or users of the connected things.   Protecting the privacy of such data must take place during data transmission, aggregation, storage, processing and mining. In agriculture, RFID sensors have been used to identify and track animals in the field.

Robotics – The term robot comes from the word robota, which means servant in the Czech language. Josef Čapek proposed it to his brother Karel to be used in the fiction play Rossum's Universal Robots, published in 1920 (Szabolcsi, 2014). In this play, machines with human behavior and appearance perform work. Nowadays, robots take on various forms and functions. In manufacturing, they take the form of arms to perform repetitive tasks such as welding, or dangerous tasks such as decontamination in nuclear facilities. Military, agricultural and space exploration robots are often vehicles with wheels or wings. Robotics is a research area that combines efforts from multiple areas, such as computer engineering, information engineering, mechanical engineering, electronic engineering, biology, as well as in the social sciences as robots must assume behaviors suitable for human interaction. A robot's degree of autonomy can range from remote control to fully autonomous operation. Depending on the task or the degree of autonomy, the robot needs: to have computer vision to build a global representation of the environment it is in and the objects within the field of view; have a control system to perform the desired task, which may or may not include artificial intelligence; have actuators that will move the parts according to the control and implement a user interface. It may also need devices that implement the sense of touch, hearing and smell. There are several advanced robots today: the Asimo, developed by Honda, is one of the most evolved humanoid-looking robots. It can walk on uneven surfaces, talk to several people at the same time, open bottles and pour liquid into a glass, in addition to mastering several simultaneous conversations with different people. NASA created Robonaut2 and sent it to the International Space Station to help carry out dangerous or even mundane tasks. In agriculture, robots often take the form of an off-road vehicle, such as the See and Spray robot, developed by Blue River to detect weeds and selectively apply crop protection products, only on those plants, avoiding the planted crop.

Geospatial data – Also called geographic data, they belong to a particular class of data that describe facts, objects and phenomena of the terrestrial globe, associated with its location on the land surface, at a particular moment or period (Câmara et al., 1996). Geospatial data are fundamentally identified from others by their spatial component, which associates each entity or phenomenon with a location translated by a geodetic territorial reference system. Geotechnology is the name given to a special category of technologies used for the process of acquiring, visualizing, processing, analyzing and/or making available geospatial data. In this context, technologies such as remote sensing, the Global Positioning System (GPS), topography, Geographic Information Systems (GIS), geographic databases, among others, are classified as geotechnologies. When geospatial information is derived from one or more geotechnologies, it is called geoinformation. Finally, geoprocessing is the process of applying one or more geotechnologies to acquire, process, visualize, analyze and/or make available spatially referenced data, in order to generate geoinformation. Geospatial data are used in agriculture, for example, to monitor the crop of a particular commodity, in which a sequence of satellite images is analyzed over time in a given region to determine how much will be produced.

GIS – Geographic Information Systems (GIS) is one of the main technologies for visualization, analysis and treatment of geographic data. There are several definitions for GIS, from the most complex to the simplest. Pires et al. (1994) define GIS as a system that performs the computational treatment of geospatial data, storing, managing and retrieving information. These systems are widely used in decision environments, providing users with facilities to combine information from a given region. The main difference between GIS and a conventional information system is that GIS can store both the descriptive attributes of the data and the geometries of different types of geographic data. The following are the main characteristics of GIS: to input and integrate, in a single database, textual spatial information and other data sources, such as satellite images and GPS data; and offer mechanisms to combine the various information, through manipulation and analysis algorithms, as well as to consult, retrieve and visualize the contents of the geographic database. The approach traditionally used to organize geospatial data in a GIS is its distribution in layers, or information planes, where each layer addresses a different theme for a given geographic region. For example, a satellite image of a region is a layer, as well as the municipalities in that region, their geomorphology and their hydrology. Each layer is internally represented using logical structures specific to each GIS and is stored in different files, according to the format of the system used. In agriculture, GIS can be used to create a digital model of a rural property from the measurements made using GPS at various points on that property.

Converging technologies

Digital agriculture incorporates concepts that were originally developed for other areas, such as blockchain and cloud computing, which converge for the solution of agricultural problems. The reuse of these technologies came from the need to store data remotely, to better process the data, and also to meet a recurrent demand in agriculture, which is the traceability of its products and processes.

Blockchain – It is a type of distributed database with a storage model that allows permanent and inviolable record keeping. It is known worldwide as being the technology on which bitcoin cryptocurrency was developed, and its origin dates back to 2008, when its author, under the pseudonym Satoshi Nakamoto, published an article on the internet (Nakamoto, 2008) on the creation of a decentralized electronic payment system, secure and based on a peer-to-peer (p2p) network. Blockchain allows encoding the content of a variable-length message to fixed-length data via integrity and authentication protocols based on single-use ciphers, or one-way hashing, (Castro, 2017; Ethereum, 2019).

Each transaction can be understood as an action that can be traced, and which is certified by the network's nodes, and part or all of its content may be confidential. These transactions are grouped similar to a ledger, also used in accounting operations, and because of this characteristic, the set is called a ledger. Ledgers are the basis, within a framework of computational tools, for implementing transaction systems with blockchain technology in corporate environments.

Traceability systems via blockchain provide a secure and distributed way to provide information within an agricultural production chain, or any other agribusiness processes, allowing to track information such as the origin of the product and its inputs, the use of pesticides in farming, among others.

Cloud computing – It refers to a technology that allows to access programs, files and services through the internet, without the need to install programs or store data – hence the allusion to the "cloud". The term is generally used to describe data centers available to many users over the internet (Hayes, 2008). Once it is properly connected to the online service, it is possible to appreciate its tools and save all the work done and access it later, from anywhere, from any computer with internet access, regardless of the platform. The minimum requirement is a computer compatible with the resources available on the internet. For example, a personal computer becomes just a chip connected to the internet, which in this case would represent the "great cloud" of computers, requiring only input devices, keyboard, mouse and monitor.

Cloud computing can be understood as an infrastructure paradigm that allows establishing software as a service, a large set of web-based services, with the objective of providing features that, until then, required large investments in hardware and software, which works through a pay-as-you-go model (Buyya et al., 2009). A typical example of cloud computing is file sync services like Dropbox. When copying or moving a file to this space, it will be duplicated on the application server and on other computers that have the program installed and on which a user accesses his account.

Cloud computing offers several benefits, such as: a) cost reduction: either by reducing expenses with energy, no-break or generator, air conditioning and physical security of the equipment, or by purchasing software and hardware; b) saving space: from the moment the user connects/adheres to cloud services, the storage will be completely virtual; c) flexibility: the services are perfectly adaptable to the company's different needs. If this prediction is underestimated, it is easy to increase the service, readjusting it to the real demand; d) constant updating: technology advances and hardware quickly becomes outdated. When migrating to cloud computing, following technological development becomes a much less exhausting and costly task, since contracted services are constantly updated; e) storage capacity: the ability to backup a vast amount of data, instantly, is as important as the effortlessness of recovering this data at any time, at a considerably low cost; f) increased collaboration: by allowing multiple users to access the same file remotely, cloud computing encourages collaborative work. As updates are done in real time, the data exchange between members of the same team is much faster.

However, cloud storage can generate distrust, especially when it comes to security. After all, the proposal is to keep important information in a virtual environment, and not all companies and individuals are comfortable with this approach.

Final considerations

This chapter introduced the main concepts used in data management, processing and visualization of digital agriculture. It presented digital technologies linked to the organization and representation of information, mathematical and statistical modeling, artificial intelligence, sensors and robotics and convergent technologies such as cloud computing and blockchain. In the next chapters, these technologies are explored in the many applications, built by Embrapa Agricultural Informatics and its partners, in order to provide solutions for an increasingly dynamic and integrated agriculture, such as digital agriculture. As can be seen, based on the list of technologies conceptualized here, the tools used to solve agricultural problems are located at the frontier of technological knowledge.

References

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e big data. Rio de Janeiro: Alta Brooks, 2016.

BASSANEZI, R. **Ensino-aprendizagem com modelagem matemática**. 4 ed. São Paulo: Contexto, 2002.

BISHOP, C. M. **Pattern recognition and machine learning**. Singapore: Springer Science+Business Media, 2006.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001. DOI: 10.1023/A:1010933404324.

BUYYA, R.; YEO, C. S.; VENUGOPAL, S.; BROBERG, J.; BRANDIC, I. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. **Future Generation Computer Systems**, v. 25, n. 6, p. 599-616, June 2009. DOI: 10.1016/j.future.2008.12.001.

CÂMARA, G.; CASANOVA, M.; MEDEIROS, C. B.; MAGALHÃES, G.; HEMERLY, A. **Anatomia de Sistemas de Informação Geográfica**. Campinas: Ed. Unicamp, 1996. 193 p.

CASTRO, A. de. Quantum one-way permutation over the finite field of two elements. **Quantum Information Processing**, v. 16, article number 149, 2017. DOI: 10.1007/s11128-017-1599-6.

DEAN, J.; GHEMAWAT, S. MapReduce: simplified data processing on large clusters. **Communications of the ACM**, v. 51, n. 1, p. 107-113, 2008. DOI: 10.1145/1327452.1327492.

DEDIĆ, N.; STANIER, C. Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In: PIAZOLO, F.; GEIST, V.; BREHM, L.; SCHMIDT, R. (ed.). **Innovations in enterprise information systems management and engineering**. Berlin; Heidelberg: Springer, 2017. p. 114-122. (Lecture Notes in Business Information Processing, n. 285). DOI: 10.1007/978-3-319-58801-8_10.

EMBRAPA AGRICULTURAL INFORMATICS. **AgroAPI**. Available at: https:// www.agroapi.cnptia.embrapa.br. Accessed on: 18 Jun. 2020.

ETHEREUM. 2019. Available at: https://www.ethereum.org. Accessed on: 22 May 2020.

FISHER, R. A. **Statistical methods for research workers**. 5th ed. Tweeddale Court: Oliver and Boyd, 1934.

FUKUNAGA, K. **Introduction to statistical pattern recognition**. 2nd ed. Boston: Academic Press, 1990. DOI: 10.1016/B978-0-08-047865-4.50007-7.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. **IEEE Transactions on Computers**, v. C-24, n. 7, p. 750-753, July 1975. DOI: 10.1109/T-C.1975.224297.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: The MIT Press, 2016.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2006. 770 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. New York: Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

HAYES, B. Cloud computing. **Communication of the ACM**, v. 51, n. 7, p. 9-11, July 2008. DOI: 10.1145/1364782.1364786.

INTERNATIONAL TELECOMMUNICATION UNION. **ITU-T Y.2060**. Series y: global information infrastructure, internet protocol aspects and next-generation networks: next generation networks – frameworks and functional architecture models: overview of the Internet of things. 2012. Former ITU-T Y.2060 renumbered as ITU-T Y.4000 on 2016-02-05 without further modification and without being republished. Available at: https://www.itu.int/rec/T-REC-Y.2060-201206-I. Accessed on: 17 Apr. 2020.

KLIR, G. J.; YUAN, B. **Fuzzy sets and fuzzy logic**: theory and applications. Upper Saddle River: Prentice Hall, 1995. 574 p.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Oakland. **Proceedings** [...]. Berkeley: University of California Press, 1967. v. 1, p. 281-297. Available at: https://projecteuclid.org/euclid.bsmsp/1200512992. Accessed on: 17 Apr. 2020.

MARBLESTONE, A. H.; WAYNE, G.; KORDING, K. P. Toward an Integration of deep learning and neuroscience. **Frontiers in Computational Neuroscience**, v. 10, n. 94, Sept. 2016. DOI: 10.3389/fncom.2016.00094.

NAKAMOTO, S. **Bitcoin**: a peer-to-peer electronic cash system. 2008. Available at: https://bitcoin.org/bitcoin.pdf. Accessed on: 22 Jan. 2020.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **ANSI/NISO Z39.19-2005 (R2010)**: guidelines for the construction, format, and management of monolingual controlled vocabularies. 2010. Available at: https://groups.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf. Accessed on: 18 Jun. 2020.

NOY, N. F.; MCGUINNESS, D. L. **Ontology development 101**: a guide to creating your first ontology. 2001. Available at: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf. Accessed on: 17 Apr. 2020.

PIRES, M. F.; MEDEIROS, C. M. B.; SILVA, A. B. Modelling geographic information systems using an object oriented framework. In: BAEZA-YATES, R. (ed.). **Computer science 2**. Boston: Springer, 1994. DOI: 10.1007/978-1-4757-9805-0_18.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, p. 81-106, 1986. DOI: 10.1007/BF00116251.

ROCO, M. C.; BAINBRIDGE, W. S. Overview converging technologies for improving human performance: nanotecnologia, biotecnologia, information technology and cognitive science. In: ROCO, M. C.; BAINBRIDGE, W. S. (ed.). **Converging technologies for improving human performance**. Dordrecht: Springer, 2003. p. 1-27. DOI: 10.1007/978-94-017-0359-8_1.

RUSSEL, S.; NORVIG, P. **Artificial intelligence**: a modern approach. 4th ed. New Jersey: Prentice Hall, 2020.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**. 6th ed. Ames: The Iowa State University Press, 1967. 286 p.

SZABOLCSI, R. The birth of the term Robot. **Advances in Military Technology**, v. 9, n. 1, jun. 2014.

TORRES, N. V.; SANTOS, G. The (mathematical) modeling process in biosciences. **Frontiers in Genetics**, v. 6, n. 354, Dec. 2015. DOI: 10.3389/fgene.2015.00354.

VAPNIK, V. N. **Statistical learning theory**. New York: John Wiley and Sons, 1998.

VAPNIK, V. N. **The nature of statistical learning theory**. New York: Springer-Verlag, 1995. DOI: 10.1007/978-1-4757-2440-0.

WEISER, M. The computer of the 21st Century. **Scientific American**, v. 265, n. 3, Sept. 1991. DOI: 10.1038/scientificamerican0991-94.

WHITE, T. **Hadoop**: the definitive guide. Sebastopol: O'Reilly Media, 2012.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.-H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1-37, Jan. 2008. DOI: 10.1007/s10115-007-0114-2.

ZADEH, L. A. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338-353, June 1965. DOI: 10.1016/S0019-9958(65)90241-X.