



A prospective study on the application of Data Science in agriculture

Kleber Xavier Sampaio de Souza, Sônia Ternes, Stanley Robson de Medeiros Oliveira, Maria Fernanda Moura, Luís Gustavo Barioni, Roberto Hiroshi Higa, Maria do Carmo Ramos Fasiaben

Embrapa Informática Agropecuária, Campinas, SP, Brasil,
kleber.sampaio@embrapa.br, sonia.ternes@embrapa.br,
stanley.oliveira@embrapa.br, maria-fernanda.moura@embrapa.br,
luis.barioni@embrapa.br, roberto.higa@embrapa.br, maria.fasiaben@embrapa.br

RESUMO

A quantidade e diversidade de dados disponíveis têm o potencial de causar profundas transformações na maneira que se realiza pesquisa e se propõe inovações na agricultura. Na chamada era do Petabyte, caracterizada pela ubiquidade de sensores e computadores, armazenamento quase infinito, computação em nuvem, robótica e IoT, a demanda e as oportunidades para aplicação da computação científica são extraordinárias, tanto na extração do conhecimento quanto na compreensão dos mecanismos associados a sistemas complexos. Este artigo apresenta um estudo prospectivo com base no estado da arte e enumera algumas áreas nas quais a aplicação da Ciência de Dados resultaria em grande benefício para pesquisadores, agricultores e agentes públicos.

PALAVRAS-CHAVE: Computação científica, Aprendizado de máquina, Modelagem e simulação, Agricultura, Redes de sensores.

ABSTRACT

The amount and diversity of data available have the potential to cause profound transformations in the way how research is conducted and innovations in agriculture are proposed. In the so-called Petabyte era, characterized by the ubiquity of sensors and computers, almost infinite storage, cloud computing, robotics and IoT, the demand and opportunities for scientific computing applications are extraordinary, both in extracting knowledge and in understanding mechanisms associated with Complex systems. This paper

conducts state-of-the-art research and enumerates some areas in which the application of data science would be of great benefit to researchers, farmers, and public agents.

KEYWORDS: Scientific computing, Machine learning, Simulation and modeling, Agriculture, Sensor networks.

INTRODUCTION

The enormous amount of data generated every day by humanity, which in accordance to IBM¹ amounts 2.5 exabytes (2.5 quintillion bytes) of data has caused a profound impact in research, development and innovation in agriculture. This occurs especially because agriculture deals with data ranging from nanoscale (DNA, RNA, proteins) to macroscale (remote sensing data from satellites, climate data, , agricultural census and global production databases), and have to combine them with data collected by sensor networks, harvesters, actuators etc.

In terms of complexity, the amount of data to be analyzed (terabytes or petabytes), its dimensionality, heterogeneity and dispersion in cloud storages characterizes what is called Big Data. This study performs literature review and proposes a framework to structure the actions of research, development and innovation of Data Science in agriculture.

STATE OF THE ART SURVEY

Data Science

The abundance of data is causing an enormous impact in the way how science is performed, not only in agriculture, but in general. In 2008, Wired Magazine editor, Chris Anderson (ANDERSON, 2008) published a paper on how the abundance of data renders the scientific method obsolete. The rationale behind this affirmation is that theories are inherently incomplete. The Newtonian model of the universe did not explain all questions and was supplanted by Einsteinian model, which in turn does not answer questions of the subatomic world.

For example, Google translator does not understand language semantics, nor Watson – IBM's computing system capable of answering questions in natural language – understands the meaning of the questions of the television show Jeopardy. Everything is mathematically calculated by machine learning algorithms that do not understand the nature of the data they are dealing with. Despite this, translations only get better and Watson defeated human players in Jeopardy.

¹ <https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>

Therefore, Anderson claims that instead of constructing (incomplete) theories supported by mental exercise over previously established theories, using data simply to validate them, one can analyze mathematically and computationally the data and contextualize what was learned a posteriori. This data intensive way of doing science is what has been called Data Science (DHAR, 2013; HEY et al., 2009; BELL et al., 2009). There are, of course, conflicting opinions: Marzzocchi (2015) analyzes the proposition from epistemological point of view and argues that “even the computational approach involves the testing of certain assumptions”, but it is undeniable that computational analysis of massive data plays a complementary role in scientific discovery.

A report to the president of the United States (PITAC, 2005) advocates that Computational Science “has become the third pillar” of scientific research, along with theory and experimentation. This report highlights examples linked to agriculture, such as the use of biomolecular simulations to explore research areas considered impossible via experimentation. Architectural modeling of plants and bioinformatics will allow, when sufficiently developed, the reduction of field experiments.

Synthetic biology, the design and construction of new parts, devices and systems not existing in real world as well as the redesign of existing biological systems to perform desired functions, is another area benefited from the combination of information technology (IT) with nanotechnology. The report Nanofrontiers: visions for the future of nanotechnology (SCHMIDT, 2007), organized by National Science Foundation and National Institutes of Health of the USA, pointed out the importance of IT for the organization of a library of nanometric world (Nano Library), and for treatment of information generated by nanosensors.

Another report by a Canadian initiative called ETC Group (2007) predicts that biological engineers of the future will start their work on laptops instead of their laboratories. In accordance to this group, there is no barrier to the synthesis of plants and animals, and this will occur as soon as someone decides to finance it. Synthetic biology also will change intellectual property protection, because it will involve not only the organisms and biological processes designed, but also the software used in the process of creation.

In 2008, Jay Keasling, from UC-Berkeley, succeeded in inserting genes of *Artemisia annua* into a yeast cell and then reprogrammed some genes to transform sugar in artemisin, a powerful anti-malaria drug (INTERLANDI, 2008), reducing substantially its production costs. Besides the influence on biological research, Data Science has the potential to cause impacts on farming, as discussed in the next section.

Smart Farming and Internet of Things

In the macro scale, the Food and Agriculture Organization of the United Nations (FAO) predicts that world population will be around 9,6 billion people by 2050, increasing demand on agricultural production by 70% until then (ALEXANDRATOS; BRUINSMA, 2012). The need for production increase has to be combined with restricting factors such as deceleration of productivity growth, pressure to reduce deforestation, limited availability of arable land and water and climate change. From a global perspective, FAO estimates that water withdrawal for use in agriculture is around 70%. Agricultural processes will have to increase productivity while reducing the use of natural resources. These constraints require optimized use of resources.

Smart farming is applied in many sectors, such as tracking of farm vehicles, arable farming, livestock monitoring, greenhouses and stables, fish farming, forestry and storage monitoring of tanks (TOWARDS..., 2015). It also depends on sensing technologies, telematics and geographic positioning and communication systems. Internet of Things (IoT) has gained especial attention recently. IoT comprises a global network of machines and devices capable of interacting with each other (LEE; LEE, 2015). There are many applications of IoT in agriculture. For instance, the use of RFIDs² and temperature tags can capture temperature, humidity, animal feed rate and drinking water flow to evaluate animal health and comfort (ENABLING..., 2017).

Sensor networks data analysis in smart farming applications often require Data Science expertise to make sense of the huge volume of data, filter out noise and calculate appropriate interventions. A sensor network linked to an irrigation system, for instance, collect data on temperature and humidity from local weather stations and on moisture from soil sensors; filter out noise from faulty sensors; mix information from weather forecast service for that location; and calculates the amount of irrigation need. The decisions on field actuators could either be a fixed set of rules for every possible situation, or dynamic, through the use of machine learning algorithms and cognitive computation (KELLY III, 2015). A cognitive irrigation system proposes the amount of irrigation needed based on calculations and probabilities, but learns from the human behavior if the amount of irrigation actually applied is different.

Machine learning can also be used for identification of crop diseases (BARBEDO, 2016) and in robotics. Astrand and Baerveldt (2002) presented an autonomous agricultural robot that recognized crop rows and identified a single crop among weed plants. Once recognized, a weed tool removed weeds from the row of crops. The University of Sydney has

² RFID - radio frequency identifiers

also developed a prototype called RIPPA (Robot for Intelligent Perception and Precision Application). Instead of physically removing weeds, this robot shoots them with a micro-dose of pesticide (KING, 2017), thereby using only 0.1% of the volume usually consumed in whole field spraying.

Harper Adams University is working on machine vision to recognize weeds and then using laser to kill the weed by heating the growing point (BLACKMORE, 2014). Machine vision is also at the heart of drones in agriculture (MARINELLO et al., 2016). Drones have been used, for instance, in the context of integrated pest management, to evaluate spectral variations in plants to identify spots where problems are likely occurring, such as insect or pathogen infestations (BAYLIS, 2014). Moreover, the surgical elimination of weeds satisfies the increasing consumer's demand for food with less agrochemicals. Consumer awareness and water supply has a direct impact on public policies, as discussed in next section.

Public policies

In the context of social awareness, democratic governments are compelled to discuss and substantiate their decisions based on scientific evidences (DAVIES; NUTLEY, 2001). However, improving data-driven decision support requires the ability to integrate data from different sources and formats and the production of evidence based on robust data analysis techniques (WILSDON; DOUBLEDAY, 2015). Thus, the production of evidence to support the formulation of public policies may include several techniques of scientific computing to assess quantitatively the available options in terms of costs and goals.

In Brazil, one of such policies is the Climatic Risk Agricultural Zoning (ZARC), an instrument of agricultural policy and risk management in agriculture (ARIAS et al., 2015). Developed by the Ministry of Agriculture, Livestock and Supply (MAPA), Embrapa and partner research institutions, ZARC uses a series of simulation and analysis techniques. Assad et al. (2013) assess how climate change is expected to affect ZARC for cotton in Brazil until 2040.

Applications of Data Science in modeling and analysis, digital agriculture and public policies discussed until now form the main axis of the prospective areas discussed in the sequel.

PROSPECTIVE AREAS FOR DATA SCIENCE IN AGRICULTURE

As discussed in previous sections, Data Science has been applied in many different areas and scales, ranging from the nanoscale of synthetic biology to the macroscale of agricultural zon-

ing. The remainder of this paper will enumerate some prospective areas which would benefit from the application of Data Science in research, development and innovation in agriculture.

Integrated data modeling and analysis

- Integration of agricultural databases, including zootechnical, climatic, meteorological, crop production, animal production etc.;
- Systemic analysis of data from advanced biology, considering interactions between genes, proteins, metabolic pathways and regulatory networks. This analysis includes the mathematical simulation of biological processes (PEDERSEN et al., 2014), the application of machine learning processes to search for patterns and the cognitive assistants anchored in domain ontologies;
- Application of statistical and machine learning algorithms in genome-wide association studies (GWAS) to search for genomic markers corresponding to certain characteristics, such as resistance to diseases or tenderness of meat in animals. This search occurs in molecular markers of the type single nucleotide polymorphisms (SNPs);
- Analytical models for genetic evaluation, traditional and genomic, for large animal populations (millions of animals): genetic merit evaluations of breeders, screening characteristics of market interest, depend on the adjustment of complex statistical models to data collected from the population under evaluation.
- Land use change dynamics modeling, involving the use of different temporal and spatial scales in the same model, which can even be treated by different approaches and usually demand high computational capacity (SOARES-FILHO et al., 2010).

Applications for digital agriculture

- Integrated analysis of data from sensor networks of different nature with the objective of identifying patterns and computing the best intervention in human, environmental and economic terms. Interventions could be assisted by cognitive agents;
- Applications developed in the logic of games (serious games) with the purpose of training the user in a particular product, service or procedure (MICHAEL; CHEN, 2006);

- Development of simulators of agricultural systems or their components using "digital twin" technology, which integrates sensors, simulation and actuators in an interactive way to reproduce the behavior of the real system;
- Augmented reality systems incorporated into smartphones so that producers can view on-screen additional information about the components of their farms;
- Pest and disease alert systems based on climate conditions, computed from dispersion and susceptibility models using private and public sensor networks data;
- Recommendation systems that analyze users' interaction with systems and can suggest new content based on previous similar behaviors;
- Use blockchain technology in the development of applications for agri-food traceability (TIAN, 2016). The technology has many applications and has also been proposed to monitor contract length and transparency on Carbon Dioxide Removal (CDR) contracts (COFFMANN; LOCKLEY. 2017).⁶

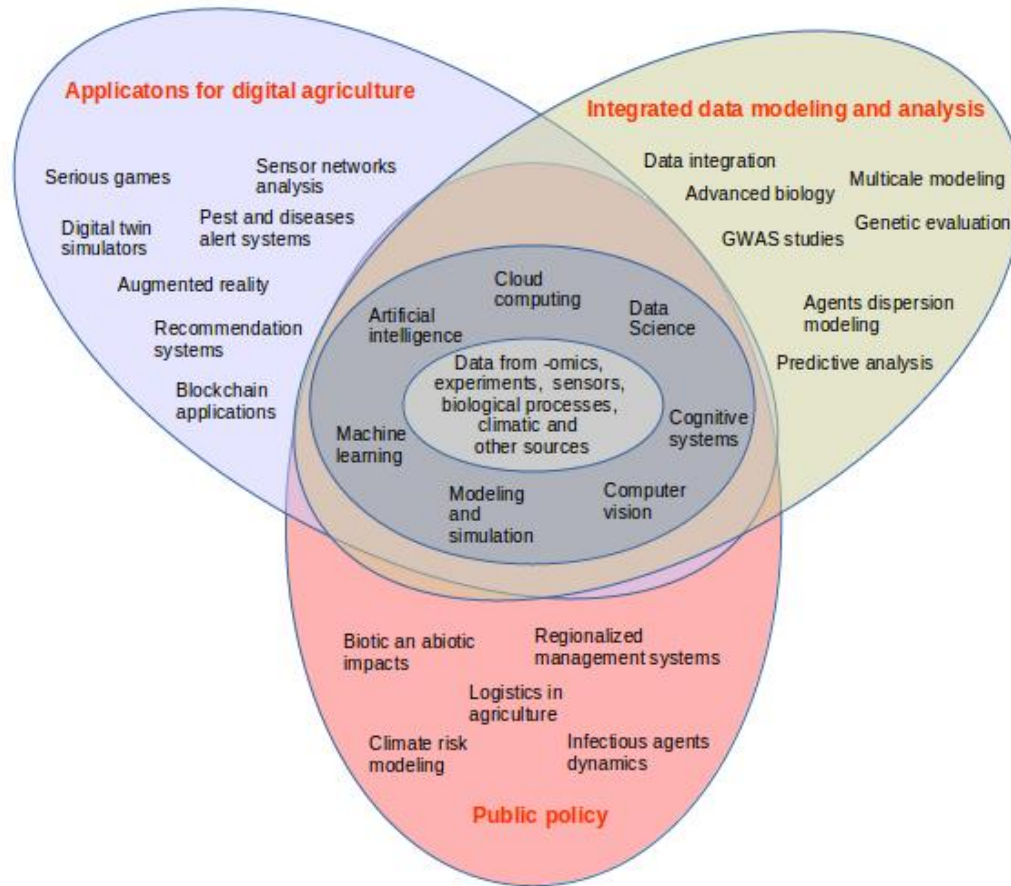
Public policy

- Modeling, simulation and optimization of agricultural production systems in response to risks and climate trends and evaluation of their environmental, social and economic sustainability;
- Predictive analysis of impacts on productivity caused by biotic or abiotic agents. The analysis could include both structured and unstructured data such as agricultural news;
- Data-driven regionalized management systems, considering aspects such as zoning, insurance and sustainable land management;
- Issuance of alerts and establishment of blocking areas to contain the spread of infectious agents in the national territory, based on predictive models (CUNNIFFE et al., 2015);
- Analysis of opportunities in logistics solutions and their impacts on the sustainability and competitiveness of agricultural production. From this analysis, the economic and environmental impacts of investments in infrastructure and logistics can be estimated.

Figure 1 shows schematically the interaction among the prospective areas. The inner oval represents data from sensors, experiments, biological processes, climatic or any other sources. In the next level, one can find the mathematical and computational disciplines which will be jointly applied to build applications. The other ovals represent each of the prospective

areas with their corresponding specific applications.

Figure 1 – Interaction among prospective areas for data science in agriculture.



Source: authors

CONCLUSION

In the era of Big Data, Data Science becomes the protagonist of digital transformation in agriculture. Several opportunities and directions for research have emerged as a consequence of a myriad of devices, standards, technologies, systems, models, algorithms, concepts and equipments. In particular, Data Science methods can be applied in a number of prospective areas, including research, digital farming public policies, marketing and commercialization. More specifically, such methods can: i) benefit researchers and practitioners from a better and more comprehensive data modeling and analysis tools applied to biological processes; ii) assist farmers to have more precision in their interventions with sensor networks – they would spend less and would have access to restricted markets that require stringent use of agrochemicals; and, iii) uphold governments to increase precision in global forecasts.

REFERENCES

- ALEXANDRATOS, N.; BRUINSMA, J. World agriculture towards 2030/2050: the 2012 revision. ESA Working paper No. 12-03. Rome, FAO. 154p. 2012.
- ANDERSON, C. The end of theory: The data deluge makes the scientific method obsolete. **Wired**, v. 16, n. 7, 2008.
- ARIAS, D.; MENDES, P.; ABEL, P. (Coord.). **Revisão rápida e integrada da gestão de riscos agropecuários no Brasil**: caminhos para uma visão integrada. Brasília, DF: Banco Mundial, 2015. 76 p. il. Disponível em: <<http://documents.worldbank.org/curated/en/717561467986362017/pdf/AUS12876-PORTUGUESE-REVISED-PUB-OUO-9-Riscos-Agropecu%C3%A1rios-no-Brasil-World-Bank-Group-paginas-compressed.pdf>>. Acesso em: 9 ago. 2017.
- ASSAD, E. D.; MARTINS, S. C.; BELTRAO, N. E. M.; PINTO, H. S.. Impacts of climate change on the agricultural zoning of climate risk for cotton cultivation in Brazil. **Pesq. agropec. bras.** [online]. 2013, vol.48, n.1, pp.1-8. ISSN 0100-204X. <http://dx.doi.org/10.1590/S0100-204X2013000100001>.
- ÅSTRAND, B.; BAERVELDT, A. J. Autonomous Robots (2002) 13: 21. doi:10.1023/A:1015674004201
- BARBEDO, J. G. A. Expert Systems Applied to Plant Disease Diagnosis: Survey and Critical View. **IEEE Latin America Transactions**, v. 14, n. 4, p. 1910-1922, 2016.
- BAYLIS, A. Is precision agriculture ready to tackle pest management?. **Outlooks on Pest Management**, v. 25, n. 6, p. 350, 2014.
- BELL, G.; HEY, T.; SZALAY, A. Beyond the Data Deluge, **Science**, vol. 323, no. 5919, pp. 1297–1298, 2009, doi: 10.1126/science.1170411.
- BLACKMORE, S. Farming with robots 2050. In: Presentation delivered at Oxford Food Security Conference, 2014.
- COFFMANN, D.; LOCKLEY, A. Carbon dioxide removal and the futures market. **Environmental Research Letters**, v. 12, n. 1, p. 015003, 2017.
- DAVIES, H.; NUTLEY, S.; Evidence-based policy and practice: moving from rhetoric to reality. Third International, Inter-disciplinary Evidence-based Policies and Indicator Systems Conference, University of Durham, July 2001.
- DHAR, V. Data Science and Prediction. **Communications of the ACM**, v. 56, n. 12, p. 64-73, 2013.
- ENABLING the smart agriculture revolution: the future of farming through the IoT perspective. [S.l.]: Beecham Research, 2016. 23p. Available at: <<http://www.beechamresearch.com/download.aspx?id=1051>>. Accessed: 15 May 2017.
- ETC Group. Extreme Genetic Engineering: an Introduction to Synthetic Biology. 2007.
- HEY, T.; TANSLEY, S. & TOLLE, K., ed., *The Fourth Paradigm: Data-Intensive Scientific*

Discovery, Microsoft Research, Redmond, Washington,. 2009.

INTERLANDI, J. Power 2009: Jay Keasling, Molecular Biologist. **Newsweek** 19/12/2008. <http://www.newsweek.com/power-2009-jay-keasling-molecular-biologist-83051>. Acessado em 10/05/2017.

KELLY III, J. E. Computing, cognition and the future of knowing: how humans and machines are forging a new age or understanding. IBM Global Services, Sommers, NY, 2015.

KING, A.. The Future of Agriculture. **Nature**, v. 544, n. 7651, p. S21-S23, 2017.

LEE, I.; LEE, K. The Internet of Things (IoT): applications investments and challenges for enterprises. *Business Horizons*, v. 58, n. 4, p. 431-440, 2015.

MARINELLO, F.; PEZZUOLO, A.; CHIUMENTI, A.; SARTORI, L. Technical analysis of unmanned aerial vehicles (drones) for agricultural applications. **Engineering for Rural Development**, 15, 2016.

MAZZOCCHI, F. Could Big Data be the end of theory in science?. **EMBO reports**, p. e201541001, 2015.

MICHAEL, D.; CHEN, S. Serious Games: Games That Educate, Train and Inform. Course Technology, Mason, USA, 2006.

CUNNIFFE, N. J.; STUTT, R. O.; DESIMONE, R. E.; GOTTWALD, T. R.; GILLIGAN, C. A. Optimising and communicating options for the control of invasive plant disease when there is epidemiological uncertainty. **PLoS computational biology**, v. 11, n. 4, p. e1004211, 2015.

PEDERSEN, M.; OURY, N.; GRAVILL, C.; PHILLIPS, A. Bio Simulators: a web UI for biological simulation. **Bioinformatics**, v. 30, n. 10, p. 1491-1492, 2014.

PITAC - President's Information Technology Advisory Committee. Computational Science: Ensuring America's Competitiveness. Report to the President. 2005.

SCHMIDT, K. F. Nanofrontiers: visions for the future of nanotechnology. Project on Emerging Technologies. Woodrow Wilson International Center for Scholars. 2007.

SOARES-FILHO, B. S.; RODRIGUES, H. O.; COSTA, W. L. Modeling Environmental Dynamics with DINAMICA EGO. Belo Horizonte, 2010. 120 p.

TIAN, F. An agri-food supply chain traceability system for China based on RFID & blockchain technology. In: **Service Systems and Service Management (ICSSSM), 2016 13th International Conference on. IEEE**, 2016.

TOWARDS smart farming: agriculture embracing the IoT vision. [S.l.]:Beecham Research, 2015. 35 p.

WILSDON, J.; DOUBLEDAY, R. (Ed.). Future directions for scientific advice in europe. Cambridge: Centre for Science and Policy, 2015. 178p. Disponível em: <<http://sro.sussex.ac.uk/54004/4/future-directions-for-scientific-advice-in-europe-v10.pdf>>. Acesso em: 20 abr. 2017.