Chapter 32

Summary of Potential and Exploitation of Big Data and AI in Bioeconomy

Caj Södergård

Abstract In this final chapter, we summarize the DataBio learnings about how to exploit big data and AI in bioeconomy. The development platform for the software used in the 27 pilots was a central tool. The Enterprise Architecture model Archimate laid a solid basis for the complex software in the pilots. Handling data from sensors and earth observation were shown in numerous pilots. Genomic data from crop species allows us to significantly speed up plant breeding by predicting plant properties in-silico. Data integration is crucial and we show how linked data enables searches over multiple datasets. Real-time processing of events provides insights for fast decision-making, for example about ship engine conditions. We show how sensitive bioeconomy data can be analysed in a privacy-preserving way. The agriculture pilots show with clear numbers the impact of big data and AI on precision agriculture, insurance and subsidies control. In forestry, DataBio developed several big data tools for forest monitoring. In fishery, we demonstrate how to reduce maintenance cost and time as well as fuel consumption in the operation of fishing vessels as well as how to accurately predict fish catches. The chapter ends with perspectives on earth observation, machine learning, data sharing and crowdsourcing.

32.1 Technologies for Boosting Sustainable Bioeconomy

Big data and AI have the potential to boost—in a sustainable way—biomass production within agriculture, forestry and fishery. Biomass means raw material for food, biomaterials and energy. For this, data is gathered in several ways: through satellites, airplanes and drones; from sensors in fields, air and ocean as well as from sensors in agriculture machinery, forest harvesters and fishing vessels. In addition, there is other data to be utilized, like weather forecasts and market prices. When all these data sources are integrated, analysed through various models and visualized, huge opportunities are created. These solutions are able to support the end users—farmers, forest owners, fishermen and other stakeholders—in their decisions and thus increase

C. Södergård (✉)

VTT Technical Research Centre of Finland, 02044 VTT, Finland

e-mail: Caj.Sodergard@vtt.fi

© The Author(s) 2021

C. Södergård et al. (eds.), *Big Data in Bioeconomy*,
https://doi.org/10.1007/978-3-030-71069-9_32

biomass production as well as decrease costs and the burden on the environment, as demonstrated in the numerous pilots in this book.

As the DataBio pilots in the three sectors utilize similar big data solutions, we created a *development platform* for the software to be used in the 27 pilots as described in Chap. 1. The platform and its assets are on the cloud and can be used by developers of bioeconomy services after the end of the project to accelerate their developments. The platform assets are gathered together in the DataBio hub (<https://www.databiohub.eu/>) and consist of 101 software *components*, of which 62 components from 28 partners were used in the 2 trial rounds conducted in 2018 and 2019 for the 27 pilots. The assets also include 65 *data sets*, of which 45 were created in DataBio and partly openly published. In addition, we collected components into 45 software *pipelines* grouped into 7 generic ones. The pipelines consist of components from the project partners and open-source components. They show how the components are interconnected. The descriptions of the *pilot systems* and the trial results are published as publicly available reports on the website (<https://www.databio.eu>). The reports are cross-linked to the hub providing a more detailed and multi-view description of the single assets, e.g. which components and datasets have been used in which pilot.

The DataBio project significantly matured already existing components during the project by adding, e.g. new user interfaces and new APIs. As a result, the technology readiness level (TLR) of the components grew with 2.7 units during the project being on average 7 on a scale from 1 to 9. When the project finished, many components were well on their way towards TLR 8 that means “system complete and verified”. One factor behind this achievement is that we applied in the planning stage a solid *enterprise architecture* model. This modelling was needed as a basis for the extensive and complex software to be constructed for 27 pilots. We adopted Archimate, which is based upon the Unified Modelling Language (UML), to create 580 diagrams, which described interfaces, subordinates and deployment environments of the components as well as the integration of components into pipelines. In addition to serving the system design, the visual models helped to communicate the pilot designs across the project team. As shown in Chap. 9, we developed a measurement system to evaluate how efficient and comprehensive the software models are.

Digital bioeconomy benefits from the rapid development of *sensors* and more widely from the emerging Internet of Things, which is expected to grow annually with two digit numbers and exceed \$1 trillion in 2022. Highly accurate sensors measuring environmental conditions at farms have enabled precision agriculture. As pointed out in Chap. 3, our DataBio pilots were able to utilize autonomous, sun-powered and wireless sensing stations from our partners measuring plenty of properties from the air, crops and soil. We also show how smart tractors equipped with telemetry tools can support current farm work as well as enable new business models.

In addition to sensor data, *earth observation* data forms the second underpinning of digital bioeconomy as shown in Chaps. 2 and 4. Almost all DataBio pilots have used freely accessible Sentinel 2 satellite data that is offered by the European Space Agency ESA. A third data category, *genomic data* from crop species of agricultural interest, opens unprecedented opportunities to predict *in-silico* plant performance and traits like yield as well as abiotic and biotic resistance. This has, as discussed

in Chap. 6, impressive applications in plant breeding, where *genomic selection* is a new paradigm allowing to bypass costly and time-consuming field phenotyping by selecting superior lines based on DNA information.

With this variety of data sources in bioeconomy, methods for integrating them are crucial. *Linked data* is a one such technology for integrating heterogeneous data. In Chap. 8, we show how we with linked data can query, for example, how fields with a certain crop intersect with buffer zones of water or the amount of pesticides used in selected plots. The semantic RDF database—triplestore—enabling these functions in DataBio has over 1 billion triples making it one of the largest semantic repositories related to agriculture. Such *knowledge graphs* are important in environmental, economic and administrative applications, but constructing links manually is time and effort intensive. Links between concepts should therefore be *discovered* automatically. In DataBio, we developed a system for discovery of RDF spatial links based on topological relations. The system outperforms state-of-the-art tools in terms of mapping time, accuracy and flexibility.

Bioeconomy applications often require *real-time processing* of sensor data as a key pillar. We demonstrate in Chap. 11 how detected situations and events provide useful real-time insights for operational management, such as preventing pest infestations in crops or machinery failures on fishing boats. In addition to being real time, data is frequently sensitive. Data might then not be made available, because of concerns that the data becomes accessible to competitors or to others that could misuse the data. In Chap. 12, we show that it is possible to handle *confidential data* as part of data analytics, combining open data and confidential data in a way that both provides business value and preserves data confidentiality. As an example, we were able to analyse high-precision data on the location and time for fishing catches without the fishery shipping companies revealing to each other where and when they got the catches.

The pilot chapters in this book show how the technologies described above and in Part I – IV of the book were deployed to meet the performance and user experience needs of each pilot.

32.2 Agriculture

As stated in previous chapters, there are high expectations on smart and precision agriculture—the forecasted market value worldwide in 2023 is over 23 billion US dollar. Smart agriculture utilizes big data technologies, Internet of Things and analytics in the various stages of the agriculture supply chain. The examples in this book illuminate the importance of smart agriculture for productivity, environmental impact, biodiversity, food security and sustainability.

In the *precision farming* pilots in Chap. 15, we achieved a significant reduction in costs of up to 15% for pesticides, 30% for irrigation and up to 60% for fertilization. These economic savings are at the same time environmental benefits. Furthermore, in yet another precision farming pilot (Chap. 18), the experiences showed

the benefits with optimal variable application of nitrogen fertilizers based on satellite monitoring of the farm fields. It is expected that the precision farming results achieved will be further improved as more data is collected to further train the models. In Chap. 17 on sorghum and potato phenology, big data allowed a more accurate prediction of yield and other plant characteristics in comparison with approaches currently in use. This improved yield prediction will help the farmers, but also the processing industry, to enhance their sales planning. In Chap. 16, we report a four times reduction in breeding time and a five times reduction in breeding costs for sorghum by applying *next-generation sequencing technologies*, and genomic prediction and selection modelling, allowing to select superior cultivars based on genetic merit derived from whole-genome DNA information. This technology can easily be scaled up to other crop species and animal husbandry.

In the *insurance* pilot in Chap. 19, we introduce new computational tools for getting more insight about the risk and the impact of heavy rain events for crops. For example, potato crops are very sensitive to heavy rain, which may cause flooding of the field due to lack of run-off and saturation of the soil. This may cause the loss of the potato yield in just a few days. A more accurate insurance assessment will encourage bigger agricultural investments. The pilot results point on possibilities to strongly reduce manual ground surveys, thus decreasing insurance costs for the farmers. To support the authorities in common agricultural policies (CAP) *subsidies control*, we achieved excellent results as reported in Chap. 20. As an example, we detected fully automatically 32 crops with 97% accuracy on areas of 9 million ha encompassing 6 million parcels in Romania. Overall, the results showed that authorities can benefit from the use of continuous satellite monitoring instead of random and limited controls. While conventionally only about 5% of the applications are cross-checked either by field sampling or by remote sensing, the methodology developed in this pilot allows checking the compliance of the farmer declarations for all agricultural parcels above 0.3 ha.

32.3 Forestry

Big data technologies have potential to replace traditional practices in forestry, even if this may require legislative changes in many countries. The reporting and monitoring of forest carbon fluxes and sustainability are increasingly in demand, and big data online platforms provide optimal tools for this. Big data and AI allow development of entirely new types of forest monitoring. DataBio developed several tools for forest owners and other stakeholders. In the work of Chap. 23, an open version of Finland's national Metsään.fi resource database was developed and got around 11 million visits in a year. The mobile crowdsourcing service *Laatumetsä*, which is connected to Metsään.fi, makes it possible for the forest owner and citizen to easily report forest damages and control quality of implemented forest operations. In 2019, the Big Data Value Association (BDVA) selected this solution as the second best success story of big data projects funded by the European Commission.

As discussed in Chap. 24, DataBio developed a *forest inventory system* that estimates forest variables and their changes based on remote sensing data and field surveys. Overall, the pilot demonstrated the benefits of big data use in forest monitoring through a range of forest inventory applications. In addition, the pilot highlighted (1) the technical transferability of online platform-based forest inventory services and (2) importance of local involvement in fine-tuning services to meet local needs. The pilot presented in Chap. 25 shows that it is possible to use field data combined with drone images to assess the *health of forest stands*. Once we obtain these local models, it is possible to extend them to larger areas at the regional or national level. The chosen tree species, despite their economic interest, required the systems to operate at the limits of the capacity of current earth observation technologies.

In Chap. 25, we report our results on forest observation from satellites for government decision-making. Because of our work, the Czech Republic changed its national legislation with updated calamity zones. The maps produced by the DataBio method help the forest owners to *optimize* timber harvesting, process resources and *fight bark beetle calamity*.

32.4 Fishery

As for the other two sectors described above, the fishery pilots demonstrated that the fishing industry can benefit from big data and AI for a more cost effective and sustainable activity. As discussed in Chap. 29, we were able to demonstrate the potential to reduce maintenance cost and time as well as fuel consumption in the *operation of fishing vessels* with better utilization of sensor information and intelligent data analysis. Both the *energy consumption model* and the *species distribution models* help optimize the route and fuel saving decisions as well as the time at sea. The DataBio engine fault predicting tool was installed on one oceanic Tuna fishing vessel and tested in real operations.

The pilot in Chap. 30 demonstrated the potential of using physical and biological parameters like catch area, season, moon phase and fish species to forecast *catch volumes*. This helps to reduce fuel consumption, stock management and to a certain extent to estimate patterns in fish prices. The decision support system has been installed on several pelagic vessels.

End users have been actively participating and giving feedback during the whole project period. Seven fishing companies have been involved in the project to test the framework and give feedback to ensure the most useful implementation including installation on the vessels.

On the other hand, the fishing industry is still in the beginning of the digital transformation and needs to overcome several obstacles before a wider scale adoption of digital technologies can take place.

32.5 Perspectives

Earth observation data is central in the applications described in this book. The freely available Sentinel satellite images offered by the European Space Agency ESA through the Copernicus Programme are used by most pilots in DataBio with good success. However, it was noted that cloudy conditions in satellite images can disturb the image analysis used for decision support, like determining the harvesting time for a crop. Therefore, it is important to have secondary sources of information as well as strong models and filtering algorithms to compensate for the disturbances.

Machine learning and data-driven artificial intelligence models are largely used for prediction and image recognition, as described earlier in this book. Advances in algorithms, like artificial neural networks and deep learning, have radically raised the accuracy of these methods. However, these data-driven methods require that extensive volumes of *labelled training data* are available. For example, data from several years might be needed in reliable crop detection. Some labelled data, like farmer's declarations and manual field observations, are costly and time consuming to obtain. As more labelled data gathers—for example, from data sharing practices, modelling and simulations—the methods used in precision agriculture and prediction of yield and fishing catches become increasingly accurate enabling better economy and sustainability. Furthermore, current artificial neural networks need in some applications to be complemented with more transparent understandable methods to create trust in the machine created recommendations. Long-range forecasts like prediction of grain and fish market prices remain challenging. However, the forecasts are continuously improving and might be useful to stakeholders even if they contain uncertainties.

One of the main hurdles in data-driven bioeconomy is the lack of *standardized data exchange and sharing*. For instance, sensors on-board fishing vessels typically demand proprietary interfaces to be built to get access to its readings. Therefore, currently, a lot of resources are needed to collect data from a large fleet of vessels. The European initiatives to create common data spaces and data infrastructures for vertical sectors, like agrifood, are highly needed. It is important to develop them also for other bioeconomy sectors like forestry and fishery.

Crowdsourcing, involving land and forest owners, as well as citizens in general, provides valuable complementing information about natural resources. However, we found that it requires a great deal of motivating actions to get, e.g. forest owners and others visiting and moving around in the forests, to participate.

Big data and artificial intelligence have to be applied to a much larger extent than currently for a more sustainable bioeconomy. The DataBio results can here offer a stepping stone for future developments, where the DataBio pipelines and solutions are scaled up to serve diverse business models and societal needs.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.