BIG DATA AND ARTIFICIAL INTELLIGENCE: EXAMPLES OF APPLICATIONS FOR THE VINE AND WINE SECTOR

Pascal Neveu[1]; Coraline Damasio[1]
[1] MISTEA, INRAE, l'Institut Agro, Montpellier University, Montpellier 34060, France
Contact : Pascal.Neveu@inrae.fr

Big Data and Artificial Intelligence (AI) are two fast-growing technological trends. The potential of Big Data and AI is mainly based on their convergence, which can offer numerous possibilities for the vine and wine industry. Indeed, throughout the wine production chain, data is multiplying with the use of smartphones, sensors, connected equipment, satellites, social networks or even with the use of computerized services (traceability, technical or economic decision support systems, computerization of declarations, etc.). What to do with all the data produced? How can the data be processed to make the best use of this information for the benefit of stakeholders in the sector and consumers? One answer lies in the association of Big Data with AI methods that will enable the production of new knowledge, provide assistance in making the right decision or simply carry out a diagnosis. But in order to be able to fully develop this potential, it is crucial to structure the data and make links between datasets.

The Big Data of the wine industry

The sources of data for the Big data viticulture are multiple and proliferating. In the field of data science, Big Data is characterized by words beginning with "V" and the wine Big Data is no exception to this rule but with its specificities.

- The ever-increasing "Volume" of data is linked to the development of technologies, the use of the Web and the reduction in the cost of storing information that potentially allows its reuse.
- "Variety" is the strongest characteristic of the Big Data in the field of vine and wine. Beyond the different types of information (text, measurement, image, spectrum, simulation...), these domains need to take into account numerous and extremely diverse sources of data, coming from:
  - at different scales ranging from territory to genomics through different levels such as farm, plot or plant.
  - spatial and temporal information that can describe areas, trajectories, transformations, cultural or winemaking operations on heterogeneous sites and installations.
  - interactions and exchanges that occur at all these scales with the environment (soil, water, climate, pests, fermentations ,...) during the different stages.
  - evolutions on tools, instrumentation, equipment, driving modes, products, rules, etc.
  - transformations that go from the plot to the bottle and the characterization of products at different stages up to tasting
  - social, economic, environmental, health impacts…

Beyond the data, it is also the variety of actors who produce and manipulate them, which is a challenge in itself. These players (winegrowers, oenologists, agronomists, geneticists, biologists, industrialists, distributors, experts, consumers, etc.) have very heterogeneous vocabularies and

modes of data production. This great variety makes it difficult to integrate these data. Therefore, it requires new approaches to structure, link and exploit all these potentially available data.

- "Velocity" corresponds to the need to be able to manage and analyze large masses of data in real time. In agriculture, this aspect will become more important with the multiplication of data sources (sensors, crowd sourcing) as it is already the case in high-throughput phenotyping platforms.
- Other "V's" can be mentioned: "Veracity" and "Validity" which characterize the quality of data sources; "Visualization" which is a problem for large complex datasets; "Visibility" of relevant facts or elements that need to be extracted from large volumes of data; and finally the analysis of all these data can pose problems of invasion of the privacy of the actors in the chain.

Finally, the real challenge of Big Data is the Value that can be obtained from it. This challenge depends on our ability to structure the Big Data. This will allow us to reuse and aggregate data in different ways but also to be able to link them with other data and feed the AI algorithms. In other words, innovation around Big Data is essential to create value, and this requires data structuring as well as the development of new approaches to data analysis and artificial intelligence.

Applications of Artificial Intelligence

Artificial Intelligence (AI) is the theory and development of computer systems capable of performing tasks that normally require intelligence. More broadly, AI is the name given to algorithms (and the programs and systems that implement and use these algorithms) that produce or mimic "intelligent" behaviour.

The best known methods to the general public are those based on machine learning, in particular "Deep Learning", which has been successfully developed over the last ten years. The majority of these methods require a large number of examples to successfully parameterize a decision support system in order to reproduce expert decisions. Impressive results have been obtained mainly in image classification, for example to recognize weeds from photos, but also to contribute to more complex applications. DeepMind company has developed an application capable of beating the world champion of the game of Go based on a very large number of games played with humans and then on millions of games between computers. These spectacular results suggest a multitude of applications for the wine industry, which must however be approached with caution. Indeed, automatic learning methods still raise many questions that are the subject of fundamental research, particularly on their sensitivity to the dataset used in the learning phase. Some of these methods require a very large amount of data (often thousands or even millions of examples) which are not always available for applications in viticulture or winemaking.

Moreover, artificial intelligence is not limited to learning techniques. It encompasses other methods that can also contribute to enhancing the value of agricultural Big Data. Without giving an exhaustive list, we can mention the stakes related to the representation of knowledge that can be of different natures (expert or imprecise). This is an essential step in the development of decision-support tools based on efficient algorithms to implement complex reasoning integrating numerous constraints.

Data structuring and linking

Building the datasets needed to train artificial intelligence algorithms requires the use of many data sources and requires these sources to be clearly described and interoperable. For this purpose, the organization and description of the data from these sources must meet various requirements:

- objects (plots, tanks, wines, farms, surveys, etc), actions (operations, additions, treatments) and problems (freezes, breakdowns, etc) must be formally identified in a unique, unambiguous and standardised manner (URI, DOI, RFID, etc). Variables must be subject to naming conventions.
- Standardised metadata must clearly describe the data in order to understand it, i.e. indicate the What? Where? How? When? as well as managing them (computerically and administratively) and finally to be able to link them. In order to ensure that different data complement or enrich each other, it is necessary to ensure that we are "talking" about the same thing. The use of computerized semantic resources (dictionaries, thesauri, taxonomies, ontologies) makes it possible to formalize the concepts and terms associated with the data. Finally, the relationships between concepts must be formalized in ontologies and will enable links to be made, for example between a tank and a wine, or an investigator and a farm.

Linked data aims to promote the publication and use of structured data, not in the form of silos of data isolated from each other, but by linking them together to form a global set of information (the data are organized in a graph) based on Web standards. This organization allows the incorporation of knowledge into the datasets that is essential for the quality of the data and to promote their (re)use. These data graphs containing a wealth of important information are ready to be exploited and analyzed to discover new knowledge.

Examples of AI applications for vines and wine

BigDataGrapes is a European project that aims to help European companies in the wine and natural cosmetics sectors to become more competitive on international markets. In particular, it aims to help companies in the wine value chain to benefit from Big Data tools, by providing decision support resulting from the real-time and transversal analysis of important and diverse data sources. For example, as part of the BigDataGrape project, works based on deep learning techniques have been carried out to develop a tool capable, from images, of automatically counting the number of leaves per vinestock. This tool can be used in several application frameworks, such as to feed agronomic models, particularly 3D representation models, or for precision viticulture applications, particularly to limit the use of inputs. This work was carried out by the CNR in Italy and the LEPSE and MISTEA units.

As part of the BigDataGrapes project, the CNR teams in Italy have also developed machine learning methods for predictive data analytics on wine data, collected from online social networks of wine passionate users (discussed in deliverable 4.3 of BigDataGrapes project). The data available consisted of a general description of the wine, user-generated data (notes and comments) as well as user profile information. A set of algorithms was deployed to assess the potential market penetration of a given wine in a new country. They estimated this penetration capability by learning a model, from user-generated contents, to be able to predict wine ratings in a target country from wine characteristics (grape variety, origin, aromas, etc).

Within the same project, the University of KULeuven is collaborating with MISTEA to develop visual analytics tools to explore large datasets and to present meaning emerging from data. A tool is being developed to visualize variables related to viticulture and winemaking for different wines over several years. The user will be able to select targeted aromas or specific characteristics (grape variety, years of interest, etc.) and compare wines produced using selected variables from vine to wine. To achieve this goal, the preliminary work was semantic data linking in order to connect them in an intelligent way, to guarantee a vine-grape-wine continuum, based on experimental data from Pech Rouge (INRAE unit). This tool will enable to answer several questions such as the influence of climatic effects or the winemaking process on the aromatic composition of a wine.

As part of a collaboration between MISTEA and NYSEOS, a study was conducted to better understand the environmental effects on the aromatic profiles of wines. Measurements of aromatic compounds on a learning set of several dozen different wines were used. The wines studied came from different sites and different vintages and 3 grape varieties were used. For a given grape variety, it is always the same type of practices and the same type of winemaking that was carried out regardless of the site or the vintage. The large number of aromatic compounds in the wine as well as the high heterogeneity of the data makes it difficult to characterise the aromatic profiles and interpret them using classical methods. To conduct this study, the data were structured in the form of a graph, and the idea followed was to look for a minimum subset (a key from the computer point of view) of compounds capable of discriminating different wines. This can be interpreted as the aromatic imprint of a wine. This work led to the development of a tool capable of learning how to distinguish wines and trying to mimic the behaviour of an expert. The tool can learn the aromatic fingerprint of hundreds, even thousands of different wines. It can then establish links between subsets of aromatic compounds and different datasets describing environmental conditions.

Conclusion and outlook

Many decision support tools, based on artificial intelligence methods and relying on Big Data tools, are currently under development. For example, we can cite a tool for the early detection of disease (flavescence dorée) using AI techniques (Chouette company) or a virtual wine shop service (Matcha company). The major issue for the development of this type of solution remains data. The wine-growing environment is in the midst of a technological transition and many players in the sector need to benefit from new advances to meet challenges such as adaptation to climate change or the reduction of inputs. To achieve this goal, it is crucial to be able to gather data and make them accessible to wider communities in a not complex or redundant way, but this is a real issue and a fundamental step for the development of all current and future intelligent tools.