



HAL
open science

Machine learning for agri-food processes: 4 learning from data, human knowledge and interactions| Book chapter Part of ISBN: 9780323984836 Current developments in biotechnology and bioengineering 2022

Nathalie Mejean-Perrot, Alberto Tonda, Nadia Boukhelifa, Ilaria Brunetti, Anastasia Bezerianos, Evelyne Lutton

► **To cite this version:**

Nathalie Mejean-Perrot, Alberto Tonda, Nadia Boukhelifa, Ilaria Brunetti, Anastasia Bezerianos, et al.. Machine learning for agri-food processes: 4 learning from data, human knowledge and interactions| Book chapter Part of ISBN: 9780323984836 Current developments in biotechnology and bioengineering 2022. 2023. hal-04298862

HAL Id: hal-04298862

<https://hal.science/hal-04298862>

Preprint submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Current developments in biotechnology and bioengineering
2022 | Book chapter | Writing - original draft, Conceptualization, Methodology, Supervision,
Writing - review & editing, Validation
Part of ISBN: [9780323984836](https://doi.org/10.1007/978-3-319-84836-6)

Machine learning for agri-food processes: learning from data, human knowledge and interactions

Nathalie Mejean Perrot^{1*}, Alberto Tonda¹, Nadia Boukhelifa¹, Ilaria
Brunetti², Anastasia Bezerianos³, Evelyne Lutton¹

¹Unité MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, France; ²UR 767

ECODEVELOPPEMENT, INRAE; UMR 8568 CIRED, CNRS, CIRAD, AgroParisTech, Université Paris-Saclay, France;

³Université Paris-Saclay, CNRS, INRIA, France

*corresponding author: nathalie.mejean@inrae.fr

Abstract:

This chapter presents three examples of data-based machine learning on time series. The common denominator of these case studies is the sparseness of data, making machine learning results fragile and inaccurate. We show how human expertise can be effectively mobilized for building useful systems, for instance useful decision support systems, able to better meet the needs of the agri-food chain. The design and analysis of different features of machine learning coupled with human knowledge enables us to sketch future human-centered machine learning systems. This approach is very relevant for the modeling of agri-food systems, because human expertise, skills and know-how are rich and numerous, but often implicit, data are heterogeneous -- big and sparse -- and processes are complex and deeply conditioned by human needs and interactions.

DOCUMENT DE
TRAVAIL

24 **Keywords:** Machine learning, interactive machine learning, knowledge integration,
25 visualisation, interactive systems, stochastic methods, agri-food processes, expertise
26 formalisation.

27 **1. Introduction**

28

29 Artificial Intelligence (AI) techniques are now commonly used in the agri-food domain
30 (Kavalenko (2020)) and are expected to be more and more used. From a recent study of
31 Accenture Research, AI has the potential to noticeably increase profitability of industries in the
32 next decades (Hunkefer, 2017). While improving the productivity, AI is also and above all
33 described as a means to meet the urgent requirements of the agroecological transition: reduce
34 environmental impact, reduce wastage, increase traceability across the supply chain, provide
35 markets with safe, high quality products to meet consumer demands (Vilani, 2017).

36 Current machine learning (ML) techniques, in particular Deep Learning, have taken off
37 thanks to the availability of a huge amount of data (big data). This has been made possible in
38 the food chain thanks to IoT (Internet of things) (Misra, 2020) for instance. Challenges
39 associated with IoT have been highlighted in a recent review of the GODAN (Global Open
40 Data for Agriculture and Nutrition) to build actors' confidence in a sustainable food system
41 (Serazetdinova et al., 2018). But when data is sparse, incomplete or inaccurate, solutions can
42 still be found (Perrot et al., 2016)(Perrot & Baudrit, 2012). Algorithms have to be adapted,
43 regarding uncertainty management in particular. Solutions based on the use of complex
44 stochastic optimisation heuristics (Lutton et al., 2016) have been proposed. Another way to
45 deal with this issue is to rely on human knowledge and expertise (implicit as well as explicit)
46 and build more and more rich and adapted human-ML interactions (Boukhelifa et al., 2017).

47 This chapter deals with the integration of human knowledge and user-interactions into AI
48 methods in the case of time dependent data sets. Three examples show how human-centered
49 approaches can be built to deal with sparse data:

- 50 ● A complex model, based on a Long-Short-Term Memory Network, built from time
51 series data thanks to ML: it is shown that human-driven choices have a drastic
52 influence on the quality of the results.
- 53 ● A tool to predict grape berries quality, thanks to a Dynamic Bayesian Network
54 coupled with expert rules. In this system, expert knowledge has been integrated into
55 the model thanks to an elicitation process, making it explicit into the structure of the
56 model through the determination of the variables dependencies and discretization.
- 57 ● Finally, the last example deals with implicit knowledge, such as the priorities experts
58 give to model objectives as they explore trade-offs. It has been shown that such
59 implicit knowledge is able to provide extremely useful information if mobilized,
60 thanks to an appropriate data visualization. It describes an exploratory analysis that
61 explicates how experts used an interactive visualization system to explore time series
62 datasets in the food domain.

63 **2. A Long-Short-Term Memory Network model for biscuit baking**

64 This case study describes how machine learning can be used to model a dynamic process,
65 such as biscuit baking. Despite the efficiency of the technique, we show how an appropriate,
66 human-curated choice of the training set can dramatically improve the final results. Long-
67 Short-Term Memory (LSTM) networks are specific to the field of Artificial Neural Networks
68 (ANNs). LSTM networks are specifically tailored for machine learning of time series, where
69 the outputs of a system are not just a function of their inputs, but also of an internal state. The
70 state itself can be seen as dependent on the historical series of all inputs seen by the system up
71 to that point in time. We present an application of LSTM networks to the modeling of biscuit
72

73 baking. Starting from 16 real-world time series of biscuit baking, gathered by the United
74 Biscuits company under different conditions, we show how the proposed LSTM network can
75 correctly predict unseen values. Remarkably, the network is also able to reproduce a dynamic
76 behavior up to variations that might be overlooked as noise.

77 The process of baking biscuits in industrial ovens involves a considerable number of different
78 biochemical and physical phenomena, such as denaturation of proteins, Maillard reactions,
79 and gelatinization of starch. Due to the complex interactions between these phenomena,
80 creating a physically accurate mathematical model of the biscuit baking process is a
81 challenging task. An alternative to a mechanistic model is to use a data-driven approach, a
82 machine learning technique, to derive a black-box model of the whole process from
83 experimental data. In order to assess generality, the model should then be tested on unseen
84 data. Such an approach could also potentially be effective at modeling outputs that are
85 traditionally harder to describe mathematically, such as the color of the biscuits. While most
86 machine learning techniques are ill-equipped to deal with time series, there is a sub-category
87 of algorithms specifically designed to tackle dynamic problems. LSTM networks are currently
88 among the state of the art in the field.

89 This case study proposes the use of a LSTM network to model the biscuit baking process.
90 Starting from a training dataset of real-world time series of biscuit baking, collected by the
91 company United Biscuits, the proposed approach learns the dynamics of two output variables
92 of interest, color and weight loss, and it is then tested on an unseen test dataset. This section
93 provides minimal information on biscuit baking and LSTM networks.

94 **2.1 Biscuit baking**

95 During the process of biscuit baking, raw biological materials are transformed into a final
96 product that must satisfy multiple criteria. For example, the color of the product must be

97 pleasant enough to entice customers, or its thickness must be within given thresholds to not
98 create issues for packaging. The industrial transformation process from dough to biscuit is
99 usually performed resorting to tunnel ovens: interestingly, even if the process has been
100 thoroughly studied and can be precisely controlled, there are complex coupled biochemical
101 and physical phenomena not completely understood and controlled (Savoie et al., 1992).

102 Phenomena involved in biscuit baking include gelatinization of starch, denaturation of
103 proteins, and Maillard reactions, that give browned food its distinctive flavor. Such
104 biochemical reactions are linked to water activity inside the biscuits, temperature, and
105 humidity (Wade, 1988). Depending on the structure of the industrial baking oven, convection,
106 radiation and conduction also contribute to baking, to different degrees. Describing
107 mathematically the global heat-mass transfer is not simple, because very little information on
108 the thermal properties of commercial dough is accessible, and the characteristics of the
109 product dynamically change during the process. Furthermore, even if the control variables are
110 known and it would be useful to represent the process, it is extremely difficult to
111 mathematically describe the evolution of sensory properties of biscuits, such as formation of
112 color, loss of moisture, and change in mass.

113 Given this complexity, it is not surprising that several approaches have been proposed to
114 model and control the industrial baking process, ranging from fuzzy logic (Perrot et al., 2000;
115 Perrot et al., 2006), to heat-transport models (Sablani et al., 1998; Trystram et al., 1993), to
116 models tackling air properties in tunnel ovens (Mirade et al., 2004). The United Biscuits, Inc.
117 collected 16 time series of biscuit cooking under different conditions, in the scope of the
118 DREAM FP7 European Project (2009-2013). The oven used during the experiments features
119 four different zones, with different temperatures. During the cooking process, biscuits are
120 slowly moved from one zone to the next on metal trays, while the heat flux in the oven is
121 manually regulated by an employee. The considered input variables are: the heat flux

122 measured in the top part of the oven (t_f in W/m^2), the heat flux measured in the bottom part of
123 the oven (b_f in W/m^2), the nominal heat flux in the current zone of the oven (z_c in, W/m^2),
124 and the nominal heat fluxes in the previous zones of the oven that the biscuit tray has already
125 passed (z_{p1} - z_{p4} in W/m^2). The considered output variables are: the color of the biscuits based
126 on the reflected light measured, L of the CIELAB system (c), and the weight loss of the
127 biscuits, measured (w_l in g). Each variable is measured every 5 s, with each baking process
128 lasting 350 s, with a total of 70 points per time-series. Color is always measured on the same
129 individual reference biscuit during the whole time series, weight loss is taken as an average on
130 the same 3 reference biscuits during the experiment. Additionally, the initial conditions of
131 variables c , and w_l are used as inputs during the experiments.

132 Out of the 16 time-series, several are repetitions of an experiment under the same conditions
133 (in groups of 3, 3, 2, 3, 2, 3 time series, respectively). Table 1 summarizes the features of the
134 dataset. Figure 1 shows an example of time series, highlighting the non-negligible differences
135 even between repetitions under the same conditions. Another notable feature is that output
136 variable w_l presents a behavior that, at a first glance, seems extremely noisy.

137 *Table 1 about here*

138 *Figure 2 about here.*

139

140

141

142 **2.2 Long-Short-Memory networks**

143 LSTM networks (Hochreiter and Schmidhuber, 1997; Gers et al., 1999) are a category of
144 ANNs, belonging to the class of Recurrent Neural Networks (RNNs) (Hopfield,1987).

145 Classical ANNs (Rosenblatt, 1958) are machine learning approaches loosely inspired by
146 neural networks in the brain that can work as general function approximators. ANNs are
147 composed by a series of units called artificial neurons connected to each other, able to receive
148 and send signals. Usually, the signal at a connection between artificial neurons is a real
149 number, and the output of each artificial neuron is calculated by a non-linear function of the
150 weighted sum of its inputs. Like most machine-learning approaches, ANNs can approximate
151 an unknown function by learning the appropriate weights in the artificial neurons from a
152 training dataset featuring several combinations of inputs and outputs for a target phenomenon.
153 In order to evaluate the generalization ability of trained models, ANNs are then usually tested
154 on a dataset of unseen values, called test dataset or test set.

155 ANNs have success stories in applications ranging from games (Silver et al., 2016) to image
156 classification (Sermanet et al., 2013), they are designed to model processes for which the
157 outputs depend exclusively on the current inputs. In dynamic processes, however, the outputs
158 are also a function of an internal state that is itself dependent on the history of inputs until that
159 point. RNNs try to address this issue, by adding connections between units to form directed
160 cycles. Thanks to this feedback mechanism, RNNs exhibit dynamic temporal behavior and are
161 used in issues where the sequence of inputs is relevant for the outputs, such as speech
162 recognition or natural language processing.

163 LSTM networks are one of the most successful paradigms of RNNs: in a LSTM network,
164 each unit is considerably more complex than a simple artificial neuron in an ANN (Figure 2).
165 A LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is
166 responsible for storing values over an arbitrary time interval, while each gate regulates the
167 flow of values that goes through the connections of the LSTM: the input gate controls the
168 extent to which a new value flows into the cell, the forget gate controls the extent to which a
169 value remains in the cell and the output gate controls the extent to which the value in the cell

170 is used to compute the output activation of the LSTM unit. Thanks to the ability of storing
171 information over variable intervals of time, LSTM networks currently represent the state of
172 the art in several domains, such as speech recognition (Xiong et al., 2017).

173 *Figure 2 about here.*

174 In a study by the authors, the 16 time-series were split into a training set (12 time series) and a
175 test set (4 time series), the latter of which was unseen by the LSTM network during the
176 training phase. The test set had been selected among the repetitions of experiments in
177 conditions already present in the training set. All the variables had been normalized by
178 subtracting the mean and scaling to unit variance on the basis of the values contained in the
179 training set. After a few tentative runs, the parameters of the network were configured as
180 follows: eight inputs (all previously described input variables plus the initial conditions for the
181 two output variables), 50 units in a single hidden layer, two outputs (all output variables); tanh
182 activation function (hyperbolic tangent), 3000 training epochs (iterations of the optimization
183 process for the weights), RMSprop gradient descent optimizer (Hinton et al., 2014). All the
184 codes of the machine learning algorithm were implemented in the Keras (Chollet et al., 2015)
185 and scikit-learn (Pedregosa et al., 2011) Python libraries.

186 The final model had excellent fitting on the test set, with mean squared error $MSE = 0.015$
187 and $R^2 = 0.9863$. An interesting result was that, visually, the model was able to reproduce
188 trends in unseen data that at a first glance might be mistaken for noise. For example, in Figure
189 3, the model is able to closely predict the behavior of the weight loss showing that the signal-
190 to-noise ratio is better than what a human expert could have considered from a superficial
191 analysis of the data.

192 While the results were quality-wise satisfying, it was important to remark that the correct
193 choice of the training set could make a considerable difference in the final generalization

194 capabilities of the model. In a second set of experiments, the same model previously
195 described was tested on a leave-one-out cross-validation, being iteratively trained on all
196 available time series except one, and tested on the one being left out. From the results
197 reported in Table 2, it is noticeable how the performance on some of the time series is subpar,
198 probably because the type of information they contain cannot be extrapolated from the rest of
199 the data, and thus represent a unique contribution that the machine learning model needs in
200 order to properly characterize the phenomenon.

201

202 *Table 2 about here.*

203 *Figure 3 about here.*

204

205 **3. Prediction of grape berries quality for decision making**

206

207

208 The grape berries maturation is a complex process relying on physicochemical and
209 biochemical reactions. These reactions depend on multiple factors of which the climate is the
210 most influential, especially in the last weeks preceding the harvest. Since berries maturity
211 plays a major role in determining wine potentialities, to anticipate the maturation process and
212 determine the right harvesting date is a significant challenge for the wine industry. Different
213 indicators to evaluate the maturation state can be considered, which might be chemical, and
214 thus exactly measurable as the content of sugar, or sensory characteristics, as the seeds color,
215 which requires an expert evaluation on a symbolic ordinated scale. Sensors have been
216 developed in last decade measuring some grape characteristics as color, sugar content or
217 aromatic potentialities, (Ben Ghazlen et al., 2010; Geraudie et al., 2010). Nevertheless, those
218 analysis are most of time realized in laboratory, time consuming and generally expensive for a
219 close monitoring of the grape berries maturity and never predictive. For the grape maturity

220 prediction, a model has been developed by (Baudrit et al., 2015; Perrot et al., 2015) linking
221 chemical indicators to weather conditions on Cabernet Franc grape berries.

222 3.1 Grape berries maturation

223 Experimental data were generated from vineyard plots located in the Loire Valley region
224 followed by the IFV institute over several years with weekly sample collection. It covered
225 years from 1989 to 2017, with plots distributed between the two geographical places: “Anjou”
226 and “Touraine” for a total of 30 parcels including between 2 to 5 points by kinetics for each
227 parcel according to the year of the experiment. As inputs, meteorological conditions over
228 seven days were supplied by Meteo France meteorological stations located near and/or on the
229 parcels. :

- 230 • Temperature (°C) labeled “t”, $\{\sum_{i=1 \text{ to } 7 \text{ days}} (t_{\min,i}+t_{\max,i})\div 2\}$,
- 231 • Rainfall (mm) labeled “pl”, $\{\sum_{i=1 \text{ to } 7 \text{ days}} (pl_i)\}$,
- 232 • Relative humidity (%) labeled “hr” $\{\sum_{i=1 \text{ to } 7 \text{ days}} (hr_{\min,i}+hr_{\max,i})\div 2\}$.

233 The solar radiation (in hours) over seven days, labeled “Ins”, $\{\sum_{i=1 \text{ to } 7 \text{ days}} (Ins_i)\}$ was only
234 given by one meteorological station located at Montreuil-Bellay, in the center of the area of
235 study. As outputs, physicochemical and sensory measurements were achieved:

- 236 • Physico-chemical measurements selected for this study were those defined by the
237 experts as essential: sugar(s) en g/l, total acidity (ac) in equivalent H₂SO₄ g/L and
238 malic acid (ac_m) in g/L, (Barbeau, 2003; Riou, 1994). Their variations during a
239 week (between two points) were also considered: Var_s; Var_ac; Var_ac_m. Each
240 week, a lot of two-hundred berries of Chenin, with pedicels, were randomly picked
241 up from each parcel at each ripening stage according to the method of Vine and
242 Wine French Institute (ITV-France) (Cayla et al., 2002) in order to limit the effects

243 of the grape heterogeneity. With the lot of two-hundred berries of each sampling, a
244 crushing was realized with a blender, then the must was filtered through a
245 Whatman paper filter. Reducing sugars concentration (g/l) was measured with a
246 refractometer; total acidity (g/l eq. H₂SO₄) by the titration method and malic acid
247 (g/l).

248 • Sensory measurements were also achieved on the berries managed by the IFV
249 institute and the ESA institute. Several measurements were described on the grape
250 berries, the juice and the global flavor using an ordinated scale range from 0 to 5.
251 For the DBN model were chosen to select an integrative flavor indicator: the
252 global aromatic intensity (IntGloAro) to complete the physicochemical
253 predictions.

254 **3.2 Expert knowledge handling**

255 Three scientists and two winegrowers experts working on the areas were interviewed during
256 one or two sessions (2–3 h). Each of the elicitation sessions was attended by one expert and
257 one or two interviewers. To build the interview, adapted methods proposed by Sicard et al.,
258 (2011) were applied. The elicitation process was based on a set of predetermined structured
259 open-ended questions used to direct the interviews. Questions were designed according to
260 techniques based on survey methods with the aim of optimizing the expression of expert
261 knowledge. Particular attention was paid to context reinstatement. This involved having the
262 expert think about and describe the feelings during the episodes being recalled.

263 **3.3 The Dynamic Bayesian Model (DBN)**

264 The model used was a Dynamic Bayesian Network (DBN), a probabilistic graphical model
265 able to describe phenomena developing over time (Jensen & Nilsen, 2010; Pearl, 1988). The

266 structure of a DBN is an oriented graph, representing correlations between the variables was
 267 created by interacting with human experts. Once the structure of a DBN was fixed, it was then
 268 possible to compute its parameters starting from a dataset. The parameters were conditional
 269 probability tables, assessing the probability for variables taking a specific value, knowing the
 270 values of the variables they depend on. For specific application, the values of the variables
 271 needed were discretized. Differently from a classical Bayesian Network, a DBN makes it
 272 possible to estimate the variable values over several subsequent time steps. In the present
 273 case, each time step was equivalent to two weeks in the grape ripening processes. DBNs have
 274 been successfully adopted for several agri-food applications (Baudrit et al., 2015; Perrot et al.,
 275 2015).

276 More formally, a DBN is a graph-based model of a joint multivariate probability distribution,
 277 capturing properties of conditional independence between variables. Like a BN, a DBN is a
 278 directed acyclic graph (DAG) where the nodes represent variables, and the missing arcs
 279 represent conditional independence between variables. In DBNs in particular, nodes $X(t) =$
 280 $(X_1(t), \dots, X_n(t))$, represent n discrete random variables, indexed by time t , providing a
 281 compact representation of joint probability distribution P for a finite time interval $[1, \tau]$. In
 282 other words, the joint probability P can be written as the product of the local probability
 283 distribution of each node and its parents, as follows:

$$284 \quad P(X(1), \dots, X(\tau)) = \prod_{i=1}^n \prod_{t=1}^{\tau} P(X_i(t)|U_i(t))$$

285 Where $U_i(\cdot)$ denotes the set of all parents of node $X_i(\cdot)$, and $P(X_i(\cdot)|U_i(\cdot))$ describes the
 286 conditional probability function associated with random variable $X_i(\cdot)$ given the values of
 287 $U_i(\cdot)$. $X_i(t)$ is termed “slices”, and it represents the set of all variables at time t . This
 288 factorization of the joint probability distribution, based on information from the graph, makes

289 it possible to straightforwardly represent large models, and use them for practical
290 applications. In other words, DBNs represent the beliefs of possible trajectories of the
291 variables involved in a dynamic process.

292 In order to make the problem treatable, DBNs assume the first-order Markov property: the
293 parents of a variable in time slice t must appear in either slice $t - 1$ or t . As a consequence,
294 for the first-order homogeneous Markov property, the conditional probabilities are time-
295 invariant, meaning that $P(U(t)) = P(U(2)) \forall t \in (1, \tau)$. In order to fully specify a DBN, we
296 will then need to define the intra-slice topology (within a time slice), the inter-slice topology
297 (between two time slices), as well as the parameters (i.e., conditional probability functions)
298 just for the first two time slices. The structure of a model can be explicitly built on the basis of
299 knowledge available in the literature and parameters can be automatically learned without a
300 priori knowledge on the basis of a dataset, a process termed parameter learning. The
301 techniques for learning DBNs are generally extensions of the techniques for learning BNs.
302 Specialized literature reports several methods to learn the structure or the parameters of a
303 DBN from substantial and/or incomplete data (Geiger & Heckerman, 1997; Heckerman,
304 1999). In this work, the topology of the graph is obtained from expert knowledge; for
305 parameter learning, we consider the simplest and most commonly adopted methodology,
306 simply evaluating the co-occurrence rate of values of variables in the training data.

307 Once a DBN is fully specified, it can be used to estimate marginal probabilities for target
308 variables, through a process also known as Bayesian inference:

$$309 \quad P(O(t')) = o(t'), \forall t' \in [1, \tau]$$

310 Where X is a set of variables whose values we are interested in predicting, and O is a set of
311 variables whose values are known (for example, in food processing X might be the variables
312 representing the physicochemical properties of a product and O might be the variables

313 representing the observed environmental conditions). In general, given a way of calculating
314 $P(X(t)|O(t'))$ from the knowledge of $P(X(t')|O(t))$, inference in a DBN is performed using
315 recursive operators and Bayes' theorem, updating the belief state of the DBN as new
316 observations become (Murphy, 2002). The DBN previously introduced was evaluated with a
317 leave-one-out cross-validation (LOOCV), where the model was repeatedly trained on the
318 whole dataset, minus one sample, and the remaining sample was used for testing. The
319 procedure was repeated until each sample was used for the testing. Considering the mean and
320 standard deviation, the results of a LOOCV provides a better estimate of the model's
321 capabilities than just considering a random split of the available data between a training set
322 and a test set (Geisser, 1993).

323 For the choices made in this study, before training the model, it was necessary to discretize
324 the real-valued variables in the dataset. However, in order to evaluate the performance of the
325 model's predictions against the ground truth, the results of the model will have to be
326 converted back into real values. Recalling that the predictions of a DBN model for variable
327 x will consist in a series of probabilities P_i for each possible discrete class $i = 0, 1, \dots, n_x$
328 associated with variable x , the predicted outcome can be converted to a real value using the
329 following equation:

$$330 \quad x^{predicted} = \sum_{i=1}^{n_x} \bar{x}_i P_i$$

331 Where \bar{x}_i is the average value of all samples of variable x that fall under class i .

332 The first metric used to evaluate the quality of the predictions against the ground truth is the
333 root mean squared error (RMSE):

334

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i^{predicted} - x_i^{observed})^2}$$

335 Where N is the number of predictions considered for target variable, and $x^{observed}$ indicates
 336 its observed value. In this study, we will also use the relative RMSE (RRMSE) that expresses
 337 the RMSE as a percentage of the range of observed values for the target variable, and it is thus
 338 more informative as an error metric:

339

$$RRMSE = \frac{RMSE}{\max(x_i^{observed}) - \min(x_i^{observed})} \times 100$$

340

341 Inspired by previous work on Cabernet-Franc and Gamay wines (Baudrit et al., 2015), the
 342 network structure predicts physicochemical indicators starting from weather conditions
 343 (Figure 4). Only the climatic variables having an influence on each physicochemical maturity
 344 indicator are selected from expert knowledge and literature. In particular, relative humidity
 345 only affects the two acidities, sunshine influences sugar content, while temperature and
 346 rainfall have an impact on the four variables considered: sugar (s), total acidity (ac), acid
 347 malic (ac_m) and the global aromatic intensity (IntGloAro).

348 *Figure 4 about here*

349 As the DBN needs to be able to capture dynamical variations of the values over time, to better
 350 predict the four variables, it is necessary to define new intermediate state variables. A month
 351 before the harvest, only alterations in the weather caused a significant deviation from an
 352 established trajectory in time (see Figure 5). More formally, considering each
 353 physicochemical variable $x \in \{ac, ac_m, s\}$ at time t and $t + 1$:

354

$$x(t + 1) = var_x(t + 1) + x(t)$$

355 And consequently

356
$$\text{var}_x(t) = x(t) - x(t - 1)$$

357 As already anticipated, the (absolute) value of a variable can be used as an indicator of the
358 current stage of ripening, while the variation, as a function of the climatic variables, will
359 dictate the ripening trajectory.

360 *Figure 5 about here*

361 At time = 0, the value of each variable is observed and known; for the next two-times steps,
362 only the climatic variables are observed and known, while the physicochemical quantities and
363 their variations are predicted by the model.

364 As previously described, to create the CPTs of our DBN model, it is necessary to define the
365 discretization of the continuous variables in the problem. In this context, discretizing variable
366 x amounts to finding several intervals $\{[x_1, x_2), [x_2, x_3), \dots, [x_{n-1}, x_n)\}$ of continuous values
367 such that $x_1 < x_2 < \dots < x_n$, with each interval corresponding to a discrete class.

368 For the climatic variables, the following intervals were defined:

- 369 • $Ins = [[15,30], [30,40], [40,55], [55,60], [60,75]]$
370 • $pl = [[0,10], [10,20], [20,30], [30,45], [45,70], [70,100]]$
371 • $t = [[0,11], [11,15], [15,17], [17,19.5], [19.5,22]]$
372 • $hr = [[60,70], [70,75], [75,80], [80,90], [90,100]]$

373 For the grape sensory variable IntGloAro, the discretization is fixed to 1 inside the sensory
374 scale [1,5]. It is linked to the limit of sensitivity evaluated to be 0.5 by the experts.

375 For the physicochemical variables, an interactive semi-automated discretization approach is
376 developed, based on the notion of co-occurrence between variable values and their variations.

377 The methodology is based on a visualization software, EvoGraphDice, coupled with an

378 evolutionary optimization approach (Boukhelifa et al., 2017). For example, the variation of
379 the variable sugar var_S, is fixed on the basis of the expert description, more focused on the
380 variations of the value of the variable than on the variable itself. The optimal discretization of
381 the sugar is then calculated to ensure the most as possible a homogeneous repartition of the
382 var_S classes of interval for each sugar interval in the data basis. The results of the
383 optimization, and thus, the discretization proposed for the physicochemical variables are
384 presented Table 3.

385 *Table 3 about here*

386 After performing a LOOCV on the dataset, where at each iteration the network is trained on
387 the whole dataset minus one sample, and then tested on that sample, a mean RRMSE for each
388 predicted variable was obtained. For the sensory variable, the RMSE is also calculated but
389 also the percentage of points well classified in the five classes considered with a threshold at
390 0.25 or 0.5, 0.5 being considered as the classical sensory threshold of sensibility for those
391 measurements.

392 *Table 4 about here*

393 The results (Table 4) showed that it was possible to predict with good results the total acidity
394 and the sugar in a range that is satisfying for the experts (10% for the sugar, 6% for the total
395 acidity) and so anticipate the maturation two weeks before. For the malic acid, it seemed to be
396 more complex to have a good prediction with the only variables considered as inputs of the
397 DBN. Probably for this variable, for a better prediction, we would have to define the state by
398 including other parameters or variables.

399 For the sensory variable, results were also relevant with a variable that seemed to be relatively
400 well predicted two weeks before with less good results for a shorter time step. As regards to
401 the integrative and more uncertain measurement represented by this sensory variable, it was

402 possible that the slope for two weeks indicates more the tendency of evolution than the one
403 for one time step, which could explain this result. Nevertheless, results of the sensory
404 prediction were quite good with 72.5% of good prediction at two time steps.

405 **4. Machine learning user interactions to understand how** 406 **agronomy experts explore model simulations**

407
408 Machine Learning ML algorithms build models that are trained to recognize certain types of
409 patterns (Bishop, 2006). Domain experts and decision-makers often rely on these models to
410 reason over new data and to make informed decisions. It is generally possible to determine
411 what predictions the machine learning model is likely to make based on new input data, how
412 domain experts will use those model predictions for reasoning and to make inferences is
413 uncertain. Because domain experts may not fully understand ML models and their domain
414 knowledge may not be fully encoded in the model, conflicts may arise and they themselves
415 may not be consistent in interpreting and responding to ML results (Valdez et al., 2017;
416 Fernandes et al., 2018; Dimara et al., 2018).

417 In previous work, Boukhelifa et al. (2019) conducted an observational study to understand
418 how domain experts use ML models to explore agri-food processes. Multiple interactive
419 sessions were organized where experts from agronomy explored model simulation datasets
420 using an existing exploratory visualization tool (Elmqvist et al., 2008; Cancino et al., 2012)
421 (Figure 6). These exploration sessions were video-recorded and experts' interactions with each
422 other and with the tool were logged. The main exploration task was open-ended, but experts'
423 primary goal was to explore alternative *trade-offs*, such as between the amount of fertilizer
424 supplied and the quantity of crop yield.

425 *Figure 6 about here*

426 These exploration sessions were helpful to the domain experts who, guided by the views
427 proposed by the ML model, found useful insights in the form of interesting correlations,
428 temporal trajectories and trade-offs that they have not considered before. As a group, they
429 were able to guide the ML component to interesting views and to reason about their data.
430 However, the qualitative analysis of the video recordings show that experts often lose track of
431 their analysis steps and therefore the many trade-off scenarios they were trying to compare.
432 Domain experts also appear to structure their investigation into mini-analysis scenarios,
433 during which they explore different hypotheses and research questions. But, when asked, they
434 were not able to give a clear overview of past exploration, or an accurate evaluation of
435 whether their exploration strategy was a robust or exhaustive one (Barczewski et al., 2020).

436 Exploratory Data Analysis EDA tools (Grinstein, 1996), such as the visualization system used
437 in this study, provide different types of visual and statistical methods to analyze the data and
438 to examine them from different viewpoints. However, they offer limited support for viewing
439 the exploration history, for example, by visualizing past analysis steps or data queries (Heer et
440 al., 2008). Little support is typically provided to show high-level information to entice users
441 to reflect upon and make sense of their past exploration. This type of information, called
442 *provenance* (North et al., 2011; Bors et al., 2019; Madanagopal et al., 2019) could provide
443 opportunities to review and share insights, but importantly, it can potentially improve user
444 exploration practices and strategies (Carrasco et al., 2017).

445 In what follows, ML is not only considered as a means to guide visual exploration, but also to
446 structure and help users revisit and reflect on past exploration sessions. This work describes
447 (a) the modelling of the user exploration history of the aforementioned exploration sessions,
448 and (b) the visualization of provenance information to the analysts as high-level views of their
449 past exploration (Barczewski et al., 2000). The main goal is to establish a methodology to
450 automatically detect key analysis stages of the exploration, which correspond to the change of

451 focus in the trade-off analysis space. To detect such changes, unsupervised learning and time
452 series modelling (Hidden Markov Models HMM) are applied to two use cases from
453 agronomy.

454 **4.1 Characterizing exploratory data analysis**

455 There is an established body of work from the cognitive psychology that looks at how people
456 make sense of data during exploratory analysis. Prominent sensemaking theories such as by
457 Klein et al. (2007) and Pirolli et al. (2005) focused on the cognitive processes involved. Their
458 results show that analysts continuously re-frame their research questions (Klein et al., 2007),
459 and interleave new and refined hypotheses in a non-linear fashion (Pirolli et al., 2004). This
460 work builds on these theories and findings, and focuses on different aspects of sensemaking
461 such as uncertainty (Boukhelifa et al., 2017b), alternatives (Liu et al., 2019) and structures
462 within the so-called exploration scenarios (Boukhelifa et al., 2019). These analysis scenarios
463 correspond to the shifts of user focus in the search space at different stages of the exploration.
464 Six types of scenarios are identified, including instances where analysts examine new and
465 refined research questions and hypotheses, and others where they attempt to recap and
466 establish common ground (Boukhelifa et al., 2019; Goyal et al., 2016). The approach, thus far
467 in studying sensemaking activities has been based on the qualitative research methods such as
468 observational studies, walkthroughs and interviews (Creswell, 2002). Although this approach
469 can yield deep insights, it is often time and resource intensive, and findings may be hard to
470 generalize. In a follow-up work (Barczewski et al., 2000), an automatic procedure was
471 implemented to detect the scenarios from the logs of user interactions, and new visual designs
472 to incite the analysts to reflect on their progress and exploration strategies.

473 Logging user interaction is common in interactive systems. User interaction logs are often
474 analyzed not only to evaluate how tools are operated by end-users, but also to help the end-
475 users themselves reflect and track their progress. For example, in the context of web

476 browsing, Carrasco et al. (2017) showed that when high level semantic information is shown,
477 users tend to reflect on their browsing habits and are able to infer areas of improvement. Guo
478 et al. (2015) found that visualization of interaction logs improves analysts' performance in
479 finding insights. Like in this study, they found that exploration is composed of multiple
480 chunks which have generic analysis patterns that lead to insight.

481 Analyzing user interaction logs is also used in analytics provenance (North et al., 2011; Bors
482 et al., 2019; Madanagopal et al., 2019) to understand user's reasoning processes, and to
483 support collaborative communication and replication (Ragan et al., 2015). Mining user
484 interactions also serves other purposes than making sense of user exploration, such as to
485 predict users personality traits (Brown et al., 2014), or to detect cognitive biases (Wall et al.,
486 2017). In the context of exploratory data analysis, this work is similar to Aboufoul et al.
487 (2018) and Dung et al. (2016) who used HMMs to model user's search behavior. HMMs are
488 powerful techniques to generate sequences of observations and to learn about the hidden
489 states that produce those observations. In the present study, it is shown that analysis scenarios
490 can be retrieved when considered as hidden states of a Markov chain. Results from the HMM
491 are provided in pseudo real time, which can continuously give high-level semantic
492 information to the analyst.

493 **4.2 Preliminary analysis of two use cases from agronomy**

494 Interaction log data were collected from an observational study based on two real-world use
495 cases in agronomy, one for wheat fertilization and the other for wine fermentation. In each use
496 case, experts from different domains (such as oenology and microbiological engineering for
497 the wine use case) explored model simulation data using a scatterplot matrix (SPLOM)-based
498 tool (Cancino et al., 2012) projected on a large shared tactile display (84" screen, Figure 6).
499 Videos of two exploration sessions were recorded per use case and followed the thematic

500 analysis method (Braun, 2012) to analyze them. Findings from this analysis were reported by
501 Boukhelifa et al., (2019). Figure 7 showed different types of scenarios that were identified.

502 *Figure 7 about here*

503 The visualization tool domain experts used to explore their model simulations data had four
504 key functionalities: **(a)** visual query selection to help experts narrow down their search to
505 important dimensions and value ranges; **(b)** a history bookmark to keep track of previous
506 views (scatterplots) they visited; **(c)** Favorite views album to store interesting scatter plots and
507 findings; **(d)** and a dimension editor to manually specify new dimensions using a
508 mathematical formulae. Alternatively, combined dimensions can be generated automatically,
509 using an interactive evolutionary algorithm, which learns from user interactions and feedback.
510 Log data events were collected pertaining to user visits of scatterplots in the SPLOM, whether
511 this originates directly in the SPLOM through cell selection, or indirectly by retrieving views
512 from the favorites store or the bookmark history.

513 A preliminary analysis of the *scatterplot visits data* showed that the manually identified
514 analysis scenarios often corresponded to localized areas of the search space (Figure 8). For
515 instance, for the wine use case, scenario 1 focused on the changes in the amount of initial
516 nitrogen (N0) at five different stages of the wine fermentation process (shown as a vertical
517 line of colored blue dots for each stage: T0, T25, T50, T75 and T100). In scenario 2, experts
518 examined the relationship between N0 and a target aromatic combination that they entered
519 manually, also with regards to the different stages of fermentation (horizontal line of colored
520 green dots). These initial findings inspired the next analysis step where the *scatterplot visits*
521 *data* was used to cluster user interactions into different analysis scenarios.

522 *Figure 8 about here*

523 In the following section, clustering and machine learning methods are used to detect those
524 scenarios automatically from a sequence of user events (*scatterplot visits data*). Two
525 unsupervised methods were implemented: a clustering method based on a spatiotemporal
526 similarity measure, and a Hidden Markov Model HMM to detect transitions between
527 scenarios. The ground truth in both cases is the manually labelled video dataset. To evaluate
528 the proposed methods, the existing notions of Type I and Type II classification errors were
529 used. However, since the scenarios are chronologically structured, changes of scenarios are
530 more important than knowing the exact identity of a class, say whether it is scenario '1'
531 instead of '3'.

532 **4.3 Spatiotemporal distances to cluster user interaction events**

533 To cluster the scatterplot visits events into different scenarios, there are three steps to follow:

534 **(i) Data preparation:** Since the datasets the experts explored are trade-off datasets and
535 describe biological processes that are dynamic in nature (fermentation and fertilization
536 processes), the data dimensions are grouped into three generic types : objective dimensions
537 (quantities experts would like to optimize, such as through minimization or maximization),
538 parameters (model parameters experts can control or modify) and trajectories (a subset of
539 parameters whose values change over time).

540 **(ii) Distance calculation:** The distance is then calculated between area clicks on the SPLOM
541 (cells) using the Jaccard distance. The result of this step is a distance matrix.

542 **(iii) Clustering:** The DBSCAN algorithm is applied to the distance matrix from step ii. The
543 rationale behind this method is to group user interactions with the SPLOM that are close both
544 spatially (based on the location of cells in the SPLOM) and temporally (based on the time
545 elapsed between two selection events).

546 Figure 9 (bottom timeline) shows the results of the clustering method for the wine use case.
547 The top timeline shows the ground truth data. Each dot represents a user event (a scatterplot
548 selection or visit) and color corresponds to scenarios S1-14 of this use case. The clustering
549 method detects more scenarios than there are in the labelled dataset. When considering
550 scenario transitions only, the method correctly detects 61% of scenario transitions for the
551 wine use case, and only 55% of transitions for the wheat use case. An example of a transition
552 that is correctly detected, is between scenarios 5 and 6 in Figure 9. Cases where the clustering
553 method does not perform well are typically the shorter scenarios where domain experts
554 quickly explore different areas of the search space, more likely to confirm previous
555 knowledge. Another limitation of the clustering method concerns step (i) for the data
556 preparation. In this step, data dimensions are grouped into three types that are identified as
557 pertinent for the different use cases, and more generally when exploring trade-off datasets for
558 dynamic systems. The clustering method is thus highly dependent on the order of dimensions
559 in the SPLOM, yet this order is arbitrary. Moreover, the adopted timescale is also arbitrary
560 and may have a big impact on the clustering performance.

561 *Figure 9 about here*

562

563 An alternative non-supervised clustering method that addresses these limitations is proposed in
564 the next section.

565 **4.4 A Hidden Markov model to detect scenario transitions**

566 Hidden Markov Models HMMs are used in many real-world applications to model sequences
567 of events where the probability of each event depends solely on the state of the previous event
568 (Baum et al., 1966). The basic assumption which underpins HMMs is that observations are
569 created by hidden states whose successions depend on transition probabilities. In

570 unsupervised use cases, studying the observations helps to find the hidden states or patterns in
571 the data. For modelling purposes, it is assumed that the hidden states correspond to the
572 analysis scenarios, and the observations (i.e. scatterplot visits events) fall into a Markov
573 system. The number of hidden states in this study is the hyper parameter, which is set to two
574 as it corresponds, conceptually, to whether or not there is a “change” or “shift” in the
575 exploration strategy or search direction.

576 To build HMM, two types of information are needed, which are extracted from the user
577 interaction logs: the time-delta between observations, and the row and column combination
578 for each scatterplot visited during the exploration. To summarize the performance of the
579 HMM, Figure 10 presents the confusion matrix. Since the goal is to detect transitions between
580 scenarios, rather than the scenario labels themselves, the shape of the path in Figure 10 is
581 more important than the inferred labels (the closer to the diagonal the better). Figure 10 shows
582 that change of scenarios are well detected for ground truth labels between 5 and 9.
583 Overall, the obtained HMM model is able to detect scenario transitions in 91% of cases for
584 the wine use case, and 75% for the wheat use case.

585 *Figure 10 about here*

586

587 **4.5 Visualizing the machine-learned storyline**

588 Besides modelling user exploration into sequences of scenarios, the results of the HMM
589 method can be used by the analysts during or after the exploration. To design this type of
590 visualization, user-centered design methodologies are used (Norman and Draper, 1986) to
591 explore the design space and to gather user requirements. For example, three brainstorming
592 sessions were organized with nine participants in total. Each session lasted roughly two hours.
593 The first two sessions had five participants with design, HCI, or visualization background,

594 and the third session had four participants from an agronomy research centre in France.
595 Participants were either researchers or PhD students. These brainstorming sessions were
596 organized into two parts. First, participants were trained to use the SPLOM-tool, similar to
597 what was proposed in previous work (Boukhelifa et al, 2017a). An ideation part followed in
598 which participants brainstormed about new features they would like to see implemented to
599 support sensemaking of their exploration history. The design ideas were collected and
600 organized using affinity diagramming and thematic analysis. The results are the following
601 high-level user requirements which are ordered by how frequently they were mentioned by
602 the participants: (1) story-tell and author; (2) highlight interesting views; (3) show trends; (4)
603 preview and replay; (5) filter views; (6) compare views; (7) group views; (8) show overview
604 and summary; (9) annotate; (10) save and reuse; (11) steer; (12) initialize; and (13) learn and
605 update.

606 The most frequent user requirements mentioned during the brainstorming sessions
607 corresponded to the storytelling and authoring category. Here participants were interested in
608 tools to automatically create a storyboard of their past exploration, and to annotate it such as
609 by adding tags to places where the exploration branched out, or where they found an
610 important insight. Other participants suggested a git-like visualization that gives both an
611 overview of visited cells and possible branching paths. Inspired by those requirements and
612 findings from previous work (Boukhelifa et al., 2019), a timeline of past exploration, called a
613 “storyline”, has been implemented, where nodes are events linked through time. Figure 11
614 shows a preliminary result of this provenance visualization integrated into an existing
615 SPLOM-tool as a widget, which can be enabled on demand. Analysis scenarios were
616 automatically identified using the HMM method and were then visualized using color.

617 *Figure 11 about here*

618 For the implementation of the storyline visualization, a client-server architecture provided by
619 Flask (a Python web framework) was used. The *server side* handled the logging and
620 modelling components described earlier. User interactions with the SPLOM were collected, as
621 well as the history bookmark and the favorites views album, which were stored in a text file.
622 Using those events (or observations), a HMM detects the hidden states (i.e., the scenarios).
623 Since an unsupervised approach was adopted, the model was applied each time a new set of
624 events are recorded to our log file. The update rate for the timeline widget was arbitrarily set
625 to five minutes, but the user could request an update at any time by pressing the update
626 button. The visualization and rendering were handled on the *client side*, which implemented a
627 storyline widget. To facilitate the sharing of the tool, this widget was integrated into a web-
628 version of the SPLOM-Tool. The storyline widget was composed of linked nodes; each node
629 corresponded to a scatterplot selection event, and its color corresponded to a detected
630 scenario. Clicking on the node rendered the corresponding scatterplot in the zoomed in area of
631 the user interface, and highlighted that cell in the SPLOM through brushing and linking. Since
632 the storyline visualization was not the primary task for domain experts, it was placed at the
633 bottom of the user interface to avoid interfering with the main exploration tasks.

634 The storyline visualization provided an overview of how domain experts structure their
635 exploratory analysis, and could be helpful for self-reflection and tracking progress. However,
636 more work would be needed to confirm whether indeed such history visualizations
637 encouraged reflection and resulted in a change in exploration strategies. There were three
638 main limitations to this work. First, the log data came from two case studies where domain
639 experts explored trade-offs between the multiple dimensions (or objectives). The ML
640 approach needed to be tested with more use cases and different types of datasets. Second, the
641 visualization tool used in the previous case study relied on a SPLOM representation of the
642 data. The way domain experts structured their exploration might depend on how the tool's

643 user interface was organized. However, the machine learning method can be generalized and
644 applied to other visualization types, since it only requires information about the data
645 dimensions consulted during the exploration and the time of viewing. Third, although cases of
646 branching were observed during the exploration, where experts explored an alternative trade-
647 off subset, the automatic method did not detect multiple, parallel or branching storylines. The
648 machine-learned storyline visualization can be improved in future by detecting the branching
649 and different types of scenarios, and by allowing the experts to augment these storylines with
650 their own annotations, thus integrating their expertise and the insights gained during the
651 exploration.

652 **5. Conclusions and perspectives**

653
654 This chapter offered several viewpoints and approaches for machine learning techniques,
655 learning from the data, expert knowledge and interactions. It showed that the generalization
656 capabilities of some ML models highly depended on the quality of the dataset, the larger the
657 better. However, even a simple expert's choice for building a good quality training set
658 (representativeness, extrapolation ability) makes a considerable difference. It also showed that
659 expertise could be made explicit (in the form of a graph model and variable discretization)
660 and embedded into a complex model to build an efficient decision-making system. Going
661 further, the implicit expertise, i.e., non-conscious skills or knowledge difficult to explain
662 verbally, could be fed into a ML process, thanks to interactive visualization. The analysis of
663 the system proposed, a SPLOM-based visualization tool, led to the design of a new tool, to
664 assist exploratory data analysis. The proposed storyline visualization helps domain experts
665 self-reflect and track their progress when analyzing the complex model simulations.

666 The key message of the work presented in this chapter is that human expertise can be
667 efficiently -- and finely -- nested into data-driven machine learning schemes, which is
668 particularly beneficial in the case of sparse or uncertain data. The three examples presented
669 above outline future components of such interactive systems. This approach is relevant
670 specifically for food-related systems, where at the same time (i) some data still remain
671 unusable, expensive and time consuming to acquire, (ii) human expertise, skills and know-
672 how are rich and numerous, but often implicit, and (iii) there is a strong need for efficient
673 predictive models, decision support systems and knowledge preservation.

674

675 **Acknowledgements**

676

677 Authors acknowledge the support from the partners and funds of the FP7 European DREAM
678 Project: "Design and development of realistic food models with well characterized micro- and
679 macro-structure and composition" and the CASDAR 2015 project coordinate by the "Institut
680 de la Vigne et Vin" (IFV): "Développement et transfert d'un outil d'aide à la décision
681 appliqué à la maturation des baies de raisins rouge et blanc". We acknowledge the IFV and
682 particularly Laurence Guerin and Etienne Goulet, InterLoire, the ESA Angers USC1422
683 GRAPPE and all the partners for the data production and opening. We also acknowledge
684 Hervé Guillemin, Bruno Perret and Daniel Picque from INRAE for their help in this project.
685 The last study is based on a late-breaking extended abstract (Barczewski et al., 2020) co-
686 authored by Antoine Barczewski.

687

688

690 **References**

- 691 Aboufoul, M., Wesslen, R., Cho, I., Dou, W. and Shaikh, S., 2018. Using hidden
692 markov models to determine cognitive states of visual analytic users. In Proceedings
693 of the Machine Learning from User Interaction for Visualization and Analytics
694 Workshop at IEEE VIS.
- 695 Barbeau, G. B.-H. (2003). Comportement de quatre cépages rouges du Val de Loire en
696 fonction des variables climatiques. *J Int Sci Vigne Vin*, 38, 35-40. Boukhelifa, N.,
697 Tonda, A., Trelea, I.C., Perrot, N., Lutton, E. (2017). Interactive knowledge
698 integration in modelling for food sustainability: challenges and prospects. ACM CHI
699 Workshop on Designing Sustainable Food Systems, 2017, NA, France. hal-01604947
- 700 Barczewski, A., Bezerianos, A. and Boukhelifa, N., 2020, April. How Domain Experts
701 Structure Their Exploratory Data Analysis: Towards a Machine-Learned Storyline. In
702 Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing
703 Systems (pp. 1-8).
- 704 Baudrit, C., Perrot, N., Brousset, J.-M., Guillemain, H., Perret, B., Picque, D., . . .
705 Barbeau, G. (2015). A probabilistic graphical model for describing the grape berry
706 maturity. *Computers and Electronics in Agriculture*. *Computers and Electronics in*
707 *Agriculture*, 118, 124-135.
- 708 Baudrit, C., Willemin, P.-H., & Perrot, N. (2013). Parameter elicitation in
709 probabilistic graphical models for modelling multi-scale food complex systems.
710 *Journal of Food Engineering*, 115(1), 1-10. Chollet F. et al., (2015). Keras.
711 <https://github.com/fchollet/keras>.
- 712 Baum, L.E. and Petrie, T., 1966. Statistical inference for probabilistic functions of
713 finite state Markov chains. *The annals of mathematical statistics*, 37(6), pp.1554-1563.
- 714 Ben Ghazlen, N., Moise, N., Latouche, G., Martinon, V., Mercier, L., Besancon, E., &
715 Cerovic, Z. (2010). Assessment of grapevine maturity using a new portable sensor:
716 Non-destructive quantification of anthocyanins. *Journal International des Sciences de*
717 *la Vigne et du Vin*, Vol. 44, pp. 1-8.
- 718 Bishop, C.M., 2006. *Pattern recognition and machine learning*. springer.
- 719 Braun, V. and Clarke, V., 2012. *Thematic analysis*.
- 720 Bors, C., Wenskovitch, J., Dowling, M., Attfield, S., Battle, L., Endert, A., Kulyk, O.
721 and Laramee, R.S., 2019. A provenance task abstraction framework. *IEEE computer*
722 *graphics and applications*, 39(6), pp.46-60.
- 723 Boukhelifa, N., Bezerianos, A., Trelea, I.C., Perrot, N.M. and Lutton, E., 2019, May.
724 An exploratory study on visual exploration of model simulations by multiple types of
725 experts. In Proceedings of the 2019 CHI Conference on Human Factors in Computing
726 Systems (pp. 1-14).

727 Boukhelifa, N., Bezerianos, A., Cancino, W. and Lutton, E., 2017. Evolutionary visual
728 exploration: evaluation of an IEC framework for guided visual search. *Evolutionary*
729 *computation*, 25(1), pp.55-86.

730 Boukhelifa, N., Perrin, M.E., Huron, S. and Eagan, J., 2017, May. How data workers
731 cope with uncertainty: A task characterisation study. In *Proceedings of the 2017 CHI*
732 *Conference on Human Factors in Computing Systems* (pp. 3645-3656).

733 Boukhelifa, N., Tonda, A., Trelea I.-C., Perrot, N., & Lutton, E. (2017). Interactive
734 Knowledge Integration in Modelling for Food Sustainability: Challenges and
735 Prospects. *ACM CHI Workshop on Designing Sustainable Food Systems*.

736 Brousset, J. (2009). *Caractérisation multifactorielle et modélisation de la maturité de*
737 *baies de Cabernet Franc en moyenne vallée de Loire. Rapport InterLoire*.

738 Brown, E.T., Ottley, A., Zhao, H., Lin, Q., Souvenir, R., Endert, A. and Chang, R.,
739 2014. Finding waldo: Learning about users from their interactions. *IEEE Transactions*
740 *on visualization and computer graphics*, 20(12), pp.1663-1672. Claverie, M.,
741 Prud'Homme, P., Mongendre, J., Zabollone, E., Raynal, M., Coulon, T., . . . Forget, D.
742 (2008). *Modélisation statistique de la qualité en viticulture par la méthode PLS Spline*.
743 *VIIe Congrès International des terroirs viticoles*.

744 Cancino, W., Boukhelifa, N. and Lutton, E., 2012, June. Evographdice: Interactive
745 evolution for visual analytics. In *2012 IEEE Congress on Evolutionary Computation*
746 (pp. 1-8). IEEE.

747 Carrasco, M., Koh, E. and Malik, S., 2017, May. popHistory: animated visualization
748 of personal web browsing history. In *Proceedings of the 2017 CHI Conference*
749 *Extended Abstracts on Human Factors in Computing Systems* (pp. 2429-2436).

750 Creswell, J.W., 2002. *Educational research: Planning, conducting, and evaluating*
751 *quantitative* (pp. 146-166). Upper Saddle River, NJ: Prentice Hall.

752 Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A. and Dragicevic, P., 2018. A
753 task-based taxonomy of cognitive biases for information visualization. *IEEE*
754 *transactions on visualization and computer graphics*, 26(2), pp.1413-1432.

755 Dungs, S., 2016. Describing user's search behaviour with Hidden Markov Models.
756 *Bulletin of IEEE Technical Committee on Digital Libraries*, 12, p.2.

757 Elmqvist, N., Dragicevic, P. and Fekete, J.D., 2008. Rolling the dice:
758 Multidimensional visual exploration using scatterplot matrix navigation. *IEEE*
759 *transactions on Visualization and Computer Graphics*, 14(6), pp.1539-1148.

760 Fernandes, M., Walls, L., Munson, S., Hullman, J. and Kay, M., 2018, April.
761 Uncertainty displays using quantile dotplots or cdfs improve transit decision-making.
762 In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*
763 (pp. 1-12).

764 Geiger D., Heckerman D. (1997) A characterization of the Dirichlet distribution
765 through global and local parameter independence. *The Annals of Statistics* 25: 1344–
766 1369.

767 Geisser, S. (1993). *Predictive Inference*. New York: Chapman and Hall.

- 768 Geraudie, V., Roger, & J. M., O. H. (2010). Développement d'un appareil permettant
769 de prédire la maturité du raisin par spectroscopie proche infra-rouge(PIR). Revue
770 Française d'Oenologie, 240, 2-8.
- 771 Gers F.A.; Schmidhuber J.; and Cummins F., (1999). Learning to forget:
772 Continual prediction with LSTM.
- 773 Goyal, N. and Fussell, S.R., 2016, February. Effects of sensemaking translucence on
774 distributed collaborative analysis. In Proceedings of the 19th ACM Conference on
775 Computer-Supported Cooperative Work & Social Computing (pp. 288-302).
- 776 Grinstein, G.G., 1996, August. Harnessing the Human in Knowledge Discovery. In
777 KDD (pp. 384-385).
- 778 Guo, H., Gomez, S.R., Ziemkiewicz, C. and Laidlaw, D.H., 2015. A case study using
779 visualization interaction logs and insight metrics to understand how analysts arrive at
780 insights. IEEE transactions on visualization and computer graphics, 22(1), pp.51-60.
- 781 Goelzer, A., Charnomordic, B., Colombié, S., Fromion, V., & Sablayrolles, J. (2009).
782 Simulation and optimization software for alcoholic fermentation in winemaking
783 conditions. Food Control, 20(7), 635-642.
- 784 Heckerman D. (1999) A Tutorial on Learning with Bayesian Networks. MIT Press,
785 Cambridge, MA, USA, 301–354.
- 786 Heer, J., Mackinlay, J., Stolte, C. and Agrawala, M., 2008. Graphical histories for
787 visualization: Supporting analysis, communication, and evaluation. IEEE transactions
788 on visualization and computer graphics, 14(6), pp.1189-1196.
- 789 Hinton G. et al., (2014). Neural Networks for Machine Learning.
790 http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
- 791 Hochreiter S. and Schmidhuber J., (1997). Long short-term memory. Neural
792 computation, 9, no. 8, 1735–1780.
- 793 Hopfield J.J., (1987). Neural networks and physical systems with emergent collective
794 computational abilities. In Spin Glass Theory and Beyond: An Introduction to the
795 Replica Method and Its Applications, World Sci- entific. 411–415.
- 796 ● Hunkefer H. (2017) Accenture Report: Artificial Intelligence Has Potential to Increase
797 Corporate Profitability in 16 Industries by an Average of 38 Percent by 2035.
798 [https://newsroom.accenture.com/news/accenture-report-artificial-intelligence-has-](https://newsroom.accenture.com/news/accenture-report-artificial-intelligence-has-potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-percent-by-2035.htm)
799 [potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-](https://newsroom.accenture.com/news/accenture-report-artificial-intelligence-has-potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-percent-by-2035.htm)
800 [percent-by-2035.htm](https://newsroom.accenture.com/news/accenture-report-artificial-intelligence-has-potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-percent-by-2035.htm)
- 801 Jensen Finn V. and Nielsen Thomas D.. (2010) Bayesian Networks and Decision
802 Graphs, Springer-Verlag. 464p.
- 803 ● Kavalenko O. (2020). Machine Learning and AI in Food Industry: Solutions and
804 Potential. internalreport:[https://spd.group/machine-learning/machine-learning-and-ai-](https://spd.group/machine-learning/machine-learning-and-ai-in-food-industry/)
805 [in-food-industry/](https://spd.group/machine-learning/machine-learning-and-ai-in-food-industry/)

806 Klein, G., Phillips, J.K., Rall, E.L. and Peluso, D.A., 2007, January. A data-frame
807 theory of sensemaking. In *Expertise out of context: Proceedings of the sixth*
808 *international conference on naturalistic decision making* (pp. 113-155). New York,
809 NY: Lawrence Erlbaum Assoc Inc.

810 Liu, J., Boukhelifa, N. and Eagan, J.R., 2019. Understanding the role of alternatives in
811 data analysis practices. *IEEE transactions on visualization and computer graphics*,
812 26(1), pp.66-76.

813 Lutton, E, Perrot, N., Tonda, A. (2016) *Evolutionary Algorithms for Food Science and*
814 *Technology*, John Wiley & Sons.

815 Madanagopal, K., Ragan, E.D. and Benjamin, P., 2019. Analytic provenance in
816 practice: The role of provenance in real-world visualization and data analysis
817 environments. *IEEE computer graphics and applications*, 39(6), pp.30-45.

818 Mirade P.; Daudin J.; Ducept F.; Trystram G.; and Clement J., 2004. Characterization
819 and CFD modelling of air temperature and velocity profiles in an industrial biscuit
820 baking tunnel oven. *Food research international*, 37, no. 10, 1031–1039.

821 Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S. , Upadhyay R. and Martynenko
822 A. , (2020) "IoT, big data and artificial intelligence in agriculture and food industry,"
823 in *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2020.2998584.

824 Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and*
825 *learning*. PhD Dissertation. Berkeley: University of California.

826 Norman, D. A, and Draper, S.W.. *User-Centered System Design: New Perspectives on*
827 *HumanComputer Interaction*. Erlbaum, Hillsdale, NJ, 1986.

828 North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B. and Fink, G., 2011.
829 Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on*
830 *Human Factors in Computing Systems* (pp. 33-36).

831 Pearl J. (1988). *Probabilistic Reasoning in Intelligent systems: Networks of Plausible*
832 *Inference*. Morgan Kaufmann, San Diego. 552p.

833 Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel
834 M.; Prettenhofer P.; Weiss R.; Dubourg V.; Vanderplas J.; Passos A.; Cournapeau D.;
835 Brucher M.; Perrot M.; and Duch- esnay E., 2011. Scikit-learn: Machine Learning in
836 Python. *Journal of Machine Learning Research*, 12, 2825–2830.

837 Perrot, N., De Vries, H., Lutton, E., Van Mil, H.G.J., Donner, M., Tonda, A., Martin,
838 S., alvarez, A., Bourguine, P., van der Linden, E. Axelos, M. (2016). Some remarks on
839 computational approaches towards sustainable complex agri-food systems. *Trends in*
840 *Food Science and Technology*, 48, 88-101.

841 Perrot, N., Baudrit, C. (2012) *Intelligent Quality control systems in food processing*
842 *based on fuzzy logic in Robotics and automation in the food industry: Current and*
843 *future technologies*. Edited by D Caldwell, Italian Institute of Technology, Italy,
844 December 2012 ISBN 1 84569 801 0, Woodhead Publishing Series in Food Science,
845 Technology and Nutrition No. 236

846 Perrot, N., Baudrit, C., Brousset, J. M., Abbal, P., Guillemin, H., Perret, B., & Picque,
847 D. (2015). *A Decision support system coupling fuzzy logic and probabilistic graphical*

848 approaches for the agri-food industry: prediction of grape berry maturity. *PloS one*,
849 10(7), e0134373.

850 Perrot N.; Ioannou I.; Allais I.; Curt C.; Hossenlopp J.; and Trystram G., 2006. Fuzzy
851 concepts applied to food product quality control: A review. *Fuzzy sets and systems*,
852 157, no. 9, 1145–1154.

853 Perrot, N., Trelea, I., Baudrit, C., Trystram, G., & Bourguine, P. (2011). Modelling and
854 analysis of complex food systems: state of the art and new trends. *Trends in Food*
855 *Science & Technology*, 22(6), 304-314.

856 Perrot N.; Trystram G.; Guely F.; Chevie F.; Schoeset- ters N.; and Dugre E., 2000.
857 Feed-back Quality Control in the Baking Industry Using Fuzzy Sets. *Journal of food*
858 *process engineering*, 23, no. 4, 249–279.

859 Pirolli, P. and Card, S., 2005, May. The sensemaking process and leverage points for
860 analyst technology as identified through cognitive task analysis. In *Proceedings of*
861 *international conference on intelligence analysis* (Vol. 5, pp. 2-4).

862 Ragan, E.D., Endert, A., Sanyal, J. and Chen, J., 2015. Characterizing provenance in
863 visualization and data analysis: an organizational framework of provenance types and
864 purposes. *IEEE transactions on visualization and computer graphics*, 22(1), pp.31-40.

865 Raynal, M., Debord, C., Guittard, S., & Vergnes, M. (2010). Epicure, a geographic
866 information decision support system risk assessment of downy and powdery mildew
867 epidemics in Bordeaux vineyards. *Sixth international workshop on the grapevine*
868 *downy and powdery mildew* (p. 144-146). Bordeaux: INRA-ISV.

869 Riou, C. (1994). *Le déterminisme climatique de la maturation du raisin: application au*
870 *zonage de la teneur en sucre dans la Communauté Européenne*. Office des
871 *Publications Officielles des Communautés Européennes*.

872 Rosenblatt F., 1958. The perceptron: A probabilistic model for information storage
873 and organization in the brain. *Psychological review*, 65, no. 6, 386.

874 Sablani S.; Marcotte M.; Baik O.; and Castaigne F., 1998. Modeling of simultaneous
875 heat and water transport in the baking process. *LWT-Food Science and Technology*,
876 31, no. 3, 201–209.

877 Savoye I.; Trystram G.; Duquenoy A.; Brunet P.; and Marchin F., 1992. Heat and
878 mass transfer dynamic modelling of an indirect biscuit baking tunnel-oven. Part I:
879 *Modelling principles*. *Journal of food engi- neering*, 16, no. 3, 173–196.

880 Serazetdinova L, Garratt J, Baylis A, Stergiadis S, Collison M, Davis S. (2019)How
881 should we turn data into decisions in AgriFood? *J Sci Food Agric*; 99(7):3213-9.

882 Sermanet P.; Eigen D.; Zhang X.; Mathieu M.; Fergus R.; and LeCun Y., 2013.
883 Overfeat: Integrated recognition, localization and detection using convolutional
884 networks. *arXiv preprint arXiv:13126229*.

885 Silver D.; Huang A.; Maddison C.J.; Guez A.; Sifre L.; Van Den Driessche G.;
886 Schrittwieser J.; Antonoglou I.; Panneershelvam V.; Lanctot M.; et al., 2016. Mas-
887 tering the game of Go with deep neural networks and tree search. *nature*, 529, no.
888 7587, 484–489.

889 Trystram G.; Fahloul D.; Duquenoy A.; and Al-lache M., 1993. Dynamic
890 modelling and simulation of the biscuit baking oven process. *Comput- ers &*
891 *Chemical Engineering*, 17, S203–S208. doi: 10.1016/0098-1354(93)80230-k. URL
892 [https://doi.org/10.1016/0098-1354\(93\)80230-k](https://doi.org/10.1016/0098-1354(93)80230-k).

893 Valdez, A.C., Ziefle, M. and Sedlmair, M., 2017. Priming and anchoring effects in
894 visualization. *IEEE transactions on visualization and computer graphics*, 24(1),
895 pp.584-594.

896 Vermersh, P. (2006). *L'entretien d'explicitation*. Paris: ESF.

897 Vilani c. (2017). FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE. report:
898 https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

899 Wade P., 1988. *Biscuit, cookies and crackers: The principles of the craft*. Vol. I.

900 Wall, E., Blaha, L.M., Franklin, L. and Endert, A., 2017, October. Warning, bias may
901 occur: A proposed approach to detecting cognitive bias in interactive visual analytics.
902 In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp.
903 104-115). IEEE.

904 Xiong W.; Wu L.; Alleva F.; Droppo J.; Huang X.; and Stolcke A., 2017. *The*
905 *Microsoft 2017 Conversational Speech Recognition System*. Tech. rep.

906

907

908

909

910

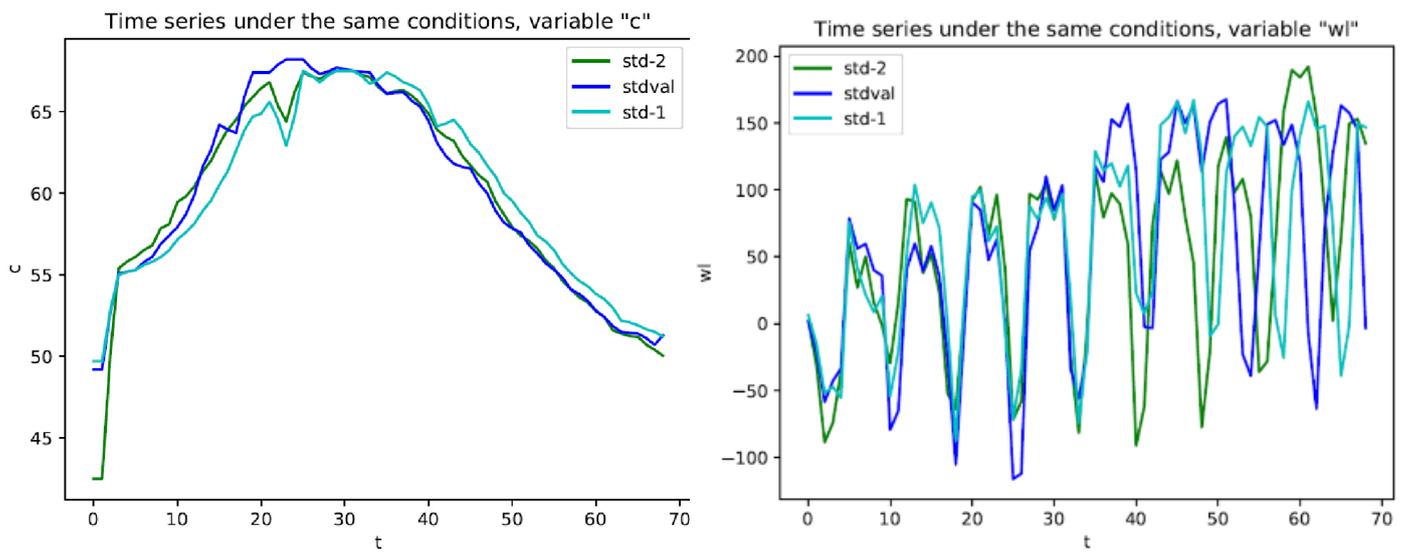
911

912 Figures

913

914

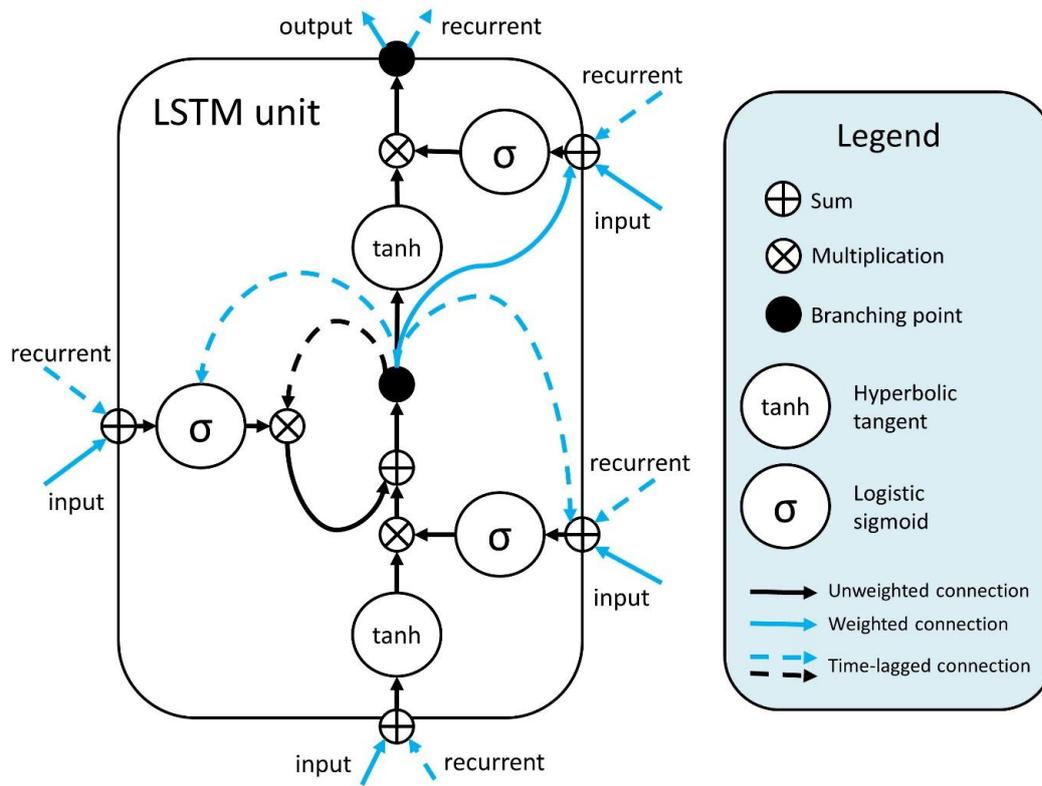
915



922

923 Figure 1 : Three time series describing the change of color (c) and weight loss (wl) during
924 biscuit baking. It is noticeable how, even though the three datasets have been collected under
925 the same conditions, there are relevant differences in the values.

926



927

928

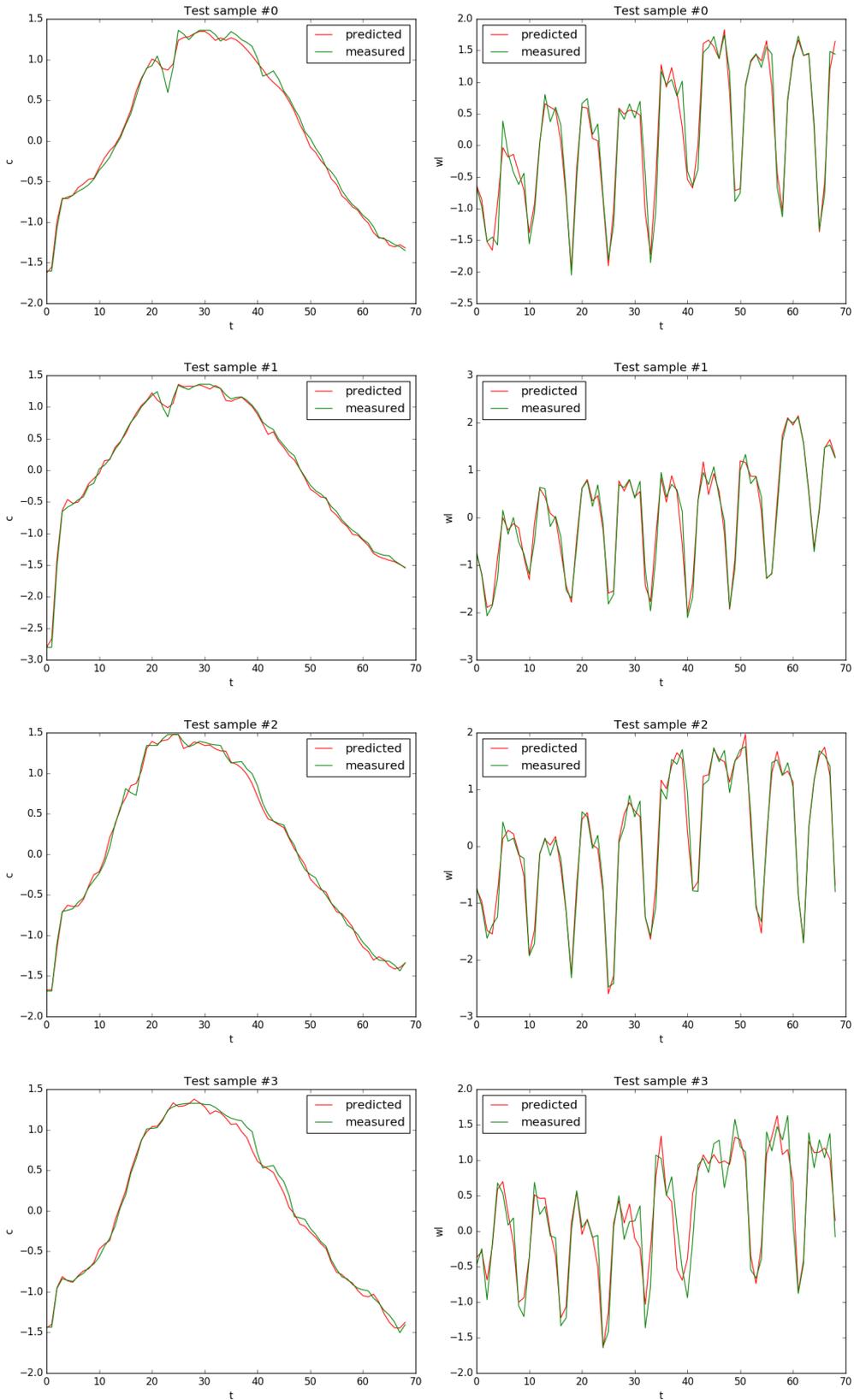
929

930 Figure 2 : Log visualization widget prototype (integrated into the visualization tool) showing
931 a single machine-learned storyline. Nodes indicate scatterplot selections and colour indicates
932 scenarios.

933

934

935

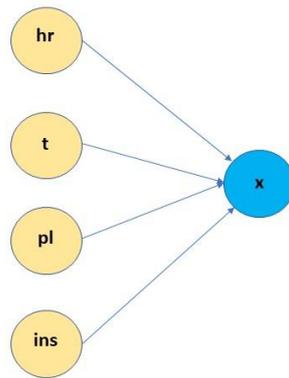


936

937
938
939
940

Figure 3: Models' predictions versus measured values for test samples, unseen during the training process. The model is able to remarkably fit the data, even for parts that might be naively believed to be noisy. The scale is different from the previous plots, as all variables have been normalized.

941



942

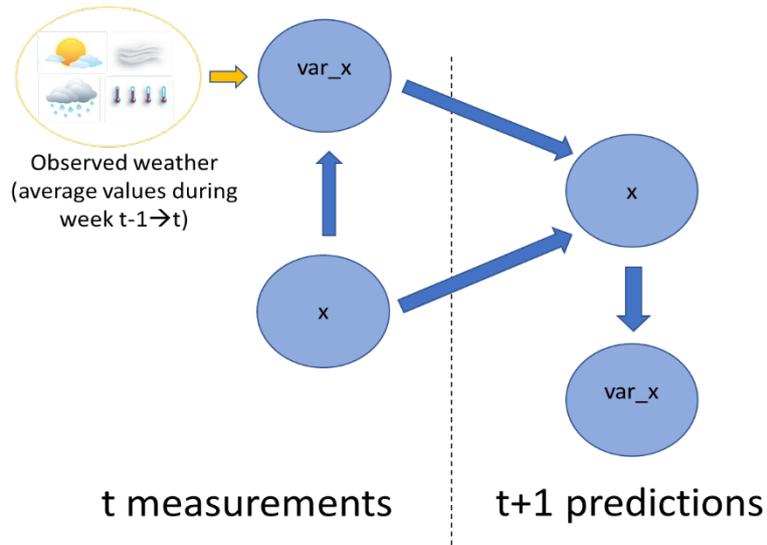
943

944 Figure 4 : Network structure describing the interaction between the climatic variables and the
945 output variables: physicochemical and sensory variables quoted x in this figure.

946

947

948



950

951

952 Figure 5: the dynamical representation of the DBN. It is represented in the form of two
953 generic slices that can be developed on several slices representing the different step times.
954 DBNs assume the first-order Markov property which means that the parents of a variable in
955 time slice t must occur in other slices and the conditional probabilities are time-invariant. The
956 slice representing the time t (t measurements) is concerned at the beginning of the iterations
957 by variables that are measured at time t_0 . The consecutive slice: time $t+1$ is dedicated to
958 predictions. If several slices are added, for example t , $t+1$ and $t+2$, it starts at t_0 with an
959 initialization where variables are measured, followed by two slices predicted $t+1$ and $t+2$,
960 with $t+2$ predicted on the basis of the prediction of $t+1$.

961

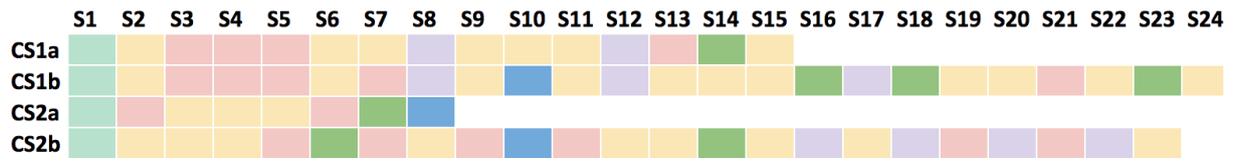
962

963

964

965

973



974

Initial New Refine Compare Alternative Storytelling

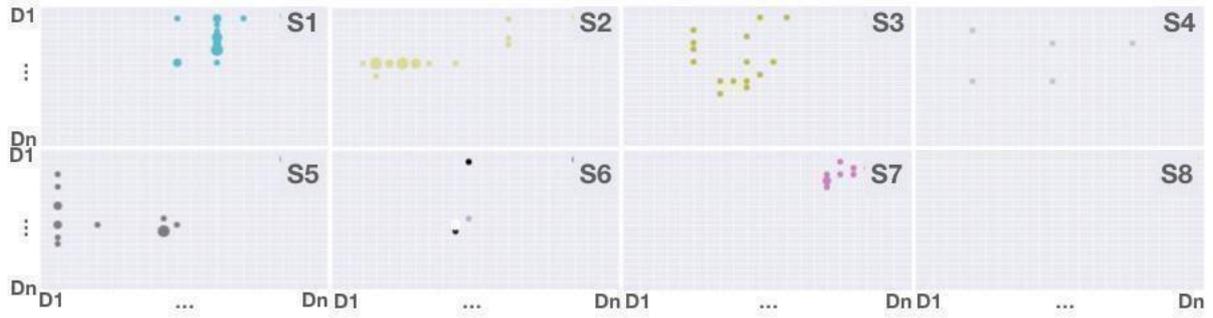
975

Figure 7 : User exploration scenario sequences and types for four sessions with domain experts, as identified from manual video coding and qualitative analysis.

976

977

978



979

980

981

982

983

984

985

Figure 8 : The first eight analysis scenarios of the wine use case S1-8 as identified from manual video coding. Each grid corresponds to one scenario, rows and columns are data dimensions D1-n including any combined dimensions (created manually or automatically). Circles indicate scatterplot visits, and their size the frequency of visits. Analysis scenarios are usually focused on one area of the search space.

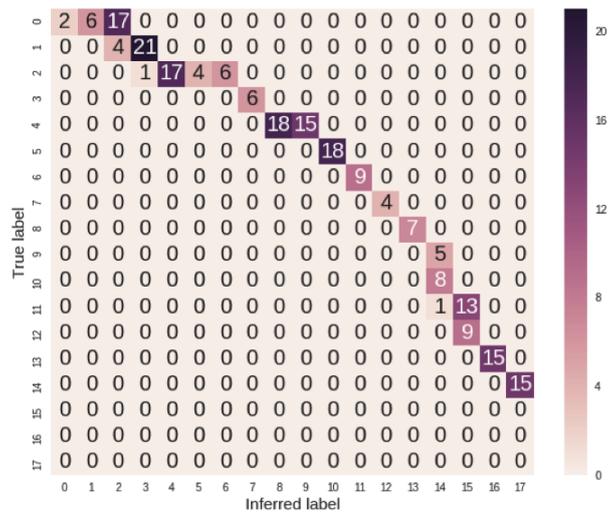
986
987



988
989
990
991

Figure 9 : Results of clustering (bottom timeline), top timeline is the ground truth. Dots are scatterplot selections, and color corresponds to scenarios S1-14.

992



993

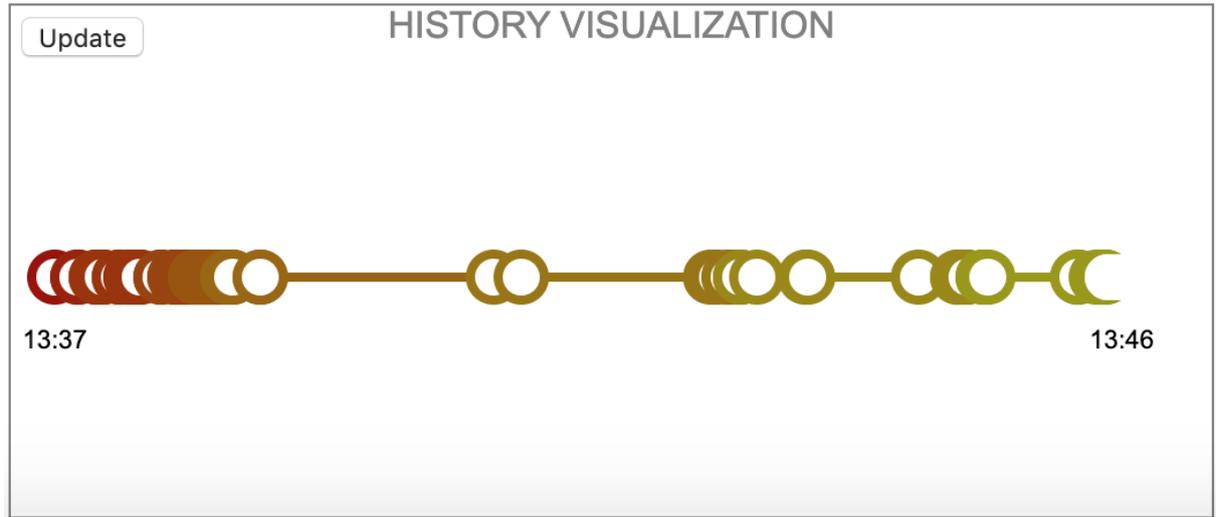
994

Figure 10 : Confusion matrix for the Hidden Markov Model, for the wine use case.

995

996

997



998

999 Figure 11 : Log visualization widget prototype (integrated into the visualization tool) showing
1000 a single machine-learned storyline. Nodes indicate scatterplot selections and colour indicates
1001 scenarios.

1002

1003

1004 Tables

1005

1006

1007 Table 1: Summary of the 16 time series on biscuit cooking gathered by United Biscuits.
 1008 During the experiments, the temperature in different zones of the oven is changed, in order to
 1009 explore several possible behaviors.

ID	Training?	Heat flux (W/m^2)				
		z1	z2	z3	z4	z5
<i>std-1</i>	yes	2500	3500	4000	4000	2000
<i>std-2</i>	yes	2500	3500	4000	4000	2000
<i>stdval</i>	no	2500	3500	4000	4000	2000
<i>T1-1</i>	yes	4000	3500	4000	4000	2000
<i>T1-2</i>	yes	4000	3500	4000	4000	2000
<i>T1val</i>	no	4000	3500	4000	4000	2000
<i>T2-1</i>	yes	2500	3500	4000	4000	3000
<i>T2-2</i>	yes	2500	3500	4000	4000	3000

1010

ID	Training?	Heat flux (W/m^2)				
		z1	z2	z3	z4	z5
<i>T3-1</i>	yes	2500	3500	6000	4000	2000
<i>T3-2</i>	yes	2500	3500	6000	4000	2000
<i>T3val</i>	no	2500	3500	6000	4000	2000
<i>T4-1</i>	yes	2500	3500	4000	6000	2000
<i>T4-2</i>	yes	2500	3500	4000	6000	2000
<i>T5-1</i>	yes	2500	5000	1000	5000	2000
<i>T5-2</i>	yes	2500	5000	1000	5000	2000
<i>T5val</i>	no	2500	5000	1000	5000	2000

1011

1012

1013

1014

Table 2: Results of the leave-one-out cross-validation.

ID	R2	MSE	ID	R2	MSE
<i>std-1</i>	0.9557	0.0491	<i>T3-1</i>	0.6644	0.2535
<i>std-2</i>	0.9789	0.0251	<i>T3-2</i>	0.4579	0.3678
<i>stdval</i>	0.9785	0.0280	<i>T3val</i>	0.0810	1.1728
<i>T1-1</i>	0.9592	0.0279	<i>T4-1</i>	0.4740	0.5685
<i>T1-2</i>	0.5561	0.4572	<i>T4-2</i>	0.9718	0.0311
<i>T1val</i>	0.4674	0.5844	<i>T5-1</i>	0.9584	0.0497
<i>T2-1</i>	0.6942	0.2956	<i>T5-2</i>	0.9859	0.0186
<i>T2-2</i>	0.1237	0.8653	<i>T5val</i>	0.9645	0.0241

1015

1016

1017

1018
 1019
 1020
 1021
 1022
 1023
 1024
 1025

Table 3: Discretization of the physico-chemical variables, fixed by experts for the variation var_X and calculated by optimization for the variable X.

	Discretization variable X	Discretization Var_X
s (sugar)	<ul style="list-style-type: none"> ● Class 0 = $[-\infty, 156.9]$, ● Class 1 = $[156.9, 182.86]$ ● Class 2 = $[182.86, 201.8]$ ● Class 3 = $[201.8, 210]$ ● Class 5 = $[210, 220]$ ● Class 6 = $[220, 230]$ ● Class 7 = $[230, 240]$ ● Class 8 = $[240, +\infty]$, 	<ul style="list-style-type: none"> ● Classe 0 = $[0, 12]$, ● Classe 1 = $[12, 20]$ ● Classe 2 = $[20, 35]$ ● Classe 3 = $[35, +\infty]$
ac (total acidity)	<ul style="list-style-type: none"> ● Class 0 = $[-\infty, 5.47]$ ● Class 1 = $[5.47, 6.33]$ ● Class 2 = $[6.33, 7.94]$ ● Class 3 = $[7.94, +\infty]$ 	<ul style="list-style-type: none"> ● Class 0 = $[-\infty, -1.5]$ ● Class 1 = $[-1.5, -1]$ ● Class 2 = $[-1, -0.6]$ ● Class 3 = $[-0.6, 0]$
ac_m (malic acid)	<ul style="list-style-type: none"> ● Classe 0 = $[-\infty, 3.66]$ ● Classe 1 = $[3.66, 4.6]$ ● Classe 2 = $[4.6, 5.68]$ ● Classe 3 = $[5.68, 6.88]$ ● Classe 4 = $[6.88, +\infty]$ 	<ul style="list-style-type: none"> ● Class 0 = $[-\infty, -2.5]$ ● Class 1 = $[-2.5, -1.5]$ ● Class 2 = $[-1.5, -0.75]$ ● Class 3 = $[-0.75, -0.5]$ ● Class 4 = $[-0.5, 0]$

1026
 1027

1028 Table 4: Results of prediction for the four variables for two time steps of prediction
 1029 (anticipation of two weeks): 1 and 2, labeled X_1 and X_2. RMSE tolerance: for s: 12g/l; for
 1030 ac: 0.5 g/l; for ac_m: 0.5 g/l

1031

1032

Variable	RMSE	RRMSE % ac = [3.4,12.5]; ac_m = [1.7,10] ; s = [144,271.8]
ac_1 (g/l)	0.536	6
ac_2 (g/l)	0.648	7
ac_m_1 (g/l)	0.825	9
ac_m_2 (g/l)	0.867	10
s_1 (g /l)	11.37	8
s_2 (g/l)	12.87	10

1033

Variable	RMSE	Abs(Pred-Obs)<0.5 %	Abs(Pred-Obs)<0.25 %
IntGloAro_1	0.73	47.5	30
IntGloAro_2	0.52	72.5	50

1034

1035

1036