analytics

MDPI

Editorial

Data Stream Analytics

Jesus S. Aguilar-Ruiz [1,*], Albert Bifet [2] and Joao Gama [3]

[1] School of Engineering, Pablo de Olavide University, ES-41013 Seville, Spain
[2] AI Institute, University of Waikato, Hamilton 3216, New Zealand; abifet@waikato.ac.nz
[3] INESC TEC, FEP-University of Porto, 4200-464 Porto, Portugal; jgama@fep.up.pt
* Correspondence: aguilar@upo.es

The human brain works in such a complex way that we have not yet managed to decipher its functional mysteries. It has five main channels that act as information input: the senses. Sight, hearing, taste, smell, and touch generate information that flows from their corresponding receptors, i.e., the eyes, ears, tongue, nose, and skin, that help us understand the world around us. In short, the brain transforms information flows into knowledge and has the ability to store it for later use, i.e., it learns and memorizes.

To these five basic senses (identified by Aristotle), more recent ones (thermoception, nociception, proprioception, etc.) could be added, but it is not the purpose of this work to delve into the senses. The purpose of this study is to explore how the senses act as an input information channel to a complex organ that processes information. In fact, there are a multitude of specific receptors that send information to the brain, but do not interact with the outside world (e.g., neural sensors to control the head tilt, kinesthetic receptors to detect stretching in muscles and tendons, or receptors to measure oxygen levels in arteries).

Lastly, all the sensitive receptors transform the information via sensory nerves into electrical impulses, types of "data" that the brain can process. However, information is extremely rich, even more than we think. For instance, humans may be able to smell over 1 trillion scents [1,2]. In addition, the mechanisms that allow the electrical impulse to be initiated are not yet well understood (the Nobel Prize in Physiology or Medicine 2021 was awarded jointly to David Julius [3] and Ardem Patapoutian [4] for their discoveries of receptors for temperature and touch).

The brain is the cornerstone of human intelligence, i.e., of natural intelligence. Artificial intelligence, on the other hand, tries to emulate natural intelligence, but unlike the latter, it is not nurtured by different artificial channels of information, but through a single digital channel, it feeds a digital brain that can extract knowledge from information, i.e., it also learns and memorizes. However, the way in which the brain analyzes huge amounts of information translated into electrical impulses in a very short time is an enigma. The brain is, in essence, the foremost expert in data analytics.

Emulating the human brain means being able to process independent and fast streams of information, called *data streams* in the field of artificial intelligence. Independence means that the system can simultaneously handle the input of various data streams and perform some kind of aggregation to enable processing of the complete data set. Since the 1960s, researchers have attempted to address physical aspects (hardware) that contribute to the more efficient processing of multichannel data [5]. However, the main concern does not lie in this aspect (independence), but in how to approach the analytical functionality for a single data stream (speed), since a crucial premise for these types of data is to assume that the data arrive so quickly that the machine is not able to process it all at once. This scenario implies introducing *incrementality* into the analysis, a very important factor which, indeed, is another virtue of the human brain.

Early works in the 1990s already considered that querying a data stream is an incremental process, but a portion of data stream became temporarily resident in the database [6].

Citation: Aguilar-Ruiz, J.S.; Bifet, A.; Gama, J. Data Stream Analytics. *Analytics* **2023**, *2*, 346–349. https://doi.org/10.3390/analytics2020019

Received: 3 April 2023
Accepted: 6 April 2023
Published: 14 April 2023

As data become more significant, the debate between exact versus approximate computation emerges [7] and gain momentum from the work of Domingos & Hulten [8]. However, progress was already reported in a discipline known as *incremental learning*, which was consistent with the typology of data streams [9,10].

Incremental learning extracts knowledge from new incoming data continuously and updates the learned model dynamically. Therefore, each new model M_{i+1} is a function f of the current model M_i and the new coming data D_i, which can contain one or more new examples. When incremental strategies are applied to time series, at each instant t, a new single example e_t is available (constant period). However, in the case of data streams, at instant t, we may have zero, one or more new data (indefinite period). This property makes the processing of data streams much more complex than that of time series.

In addition, a time Δt elapses from the arrival of D_{i-1} to the arrival of D_i. This is the maximum time for the function f to process D_i and generate M_{i+1}. If the function f uses computationally expensive methods, such as those based on deep learning, there will not be enough time to perform the model update and the learning will lose quality. This results in the same situation as when a difficult concept is explained to a person very quickly: *overflow*. Modulating learning times is as important for the human brain as it is for data stream analytics.

Technically, this factor is named *scalability*, and becomes a bottleneck when data are very massive. Distributed computing can mitigate this problem, providing interesting solutions such as *edge computing* [11,12].

Incremental learning is one of the great challenges of machine learning. Humans learn incrementally and with very limited supervision, i.e., little by little, and by means of trial and error. When children conjugate an irregular verb wrongly (*anomaly*), applying a learned grammatical rule, they will have to hear the correct irregular form several times, to discover the mistake and fix it thereafter. Although, it will not be easy while they hear alternately the correct form from adults and the incorrect form from other children (*inconsistency*). It is more difficult to learn in the presence of anomalies and inconsistencies, since both are certain type of *noise*. When the anomaly becomes regular, a change in trend occurs, which is called a *concept drift* [13].

Humans are continually confronted with concept drift. For example, the perception of a political leader can change based on the news published about his or her actions. To do so, our brain processes information related to the relevant issues to our ideology (education, health, economy, etc.). The change can be smooth (it lasts longer), abrupt (immediate), gradual (it oscillates before happening), and even recurrent (it had already taken the value to which it is changing). Consequently, these concept drifts might modify voting intentions in political elections.

Formally, the ultimate goal of machine learning is to infer a function that maps some input space into an output space. Traditional approaches assume that both spaces are fixed and predefined before learning. Current challenging applications go beyond this setting. For example, in Earth monitoring applications using sensor networks, the input space may evolve over time. Sensors may stop sending information and be replaced by new ones, eventually with different characteristics.

We humans are continually expanding our input space. Liking or disliking a type of wine is an extremely simple perception. We need new elements (*feature evolution*) such as acidity or tannin structure to enrich the perception, and we need training to make the new features become relevant.

In standard machine learning applications, the output space consists of a single variable with a well-defined domain. However, there are applications that require predicting a multidimensional vector or a multi-label set; for example, in an industrial process, when the decision involves manipulating several control valves to mitigate an anomaly.

Humans face ignorance poorly, which occurs when the output space is open, something that, in fact, happens throughout our lives (*concept evolution*). In Europe, several types of potatoes are available in the markets. When visiting Peru, however, travelers discover

hundreds of types, with different colors, textures, flavors, and, therefore, gastronomic possibilities. We could not learn about it until it appeared. Learning systems handle situations of uncertainty (concepts emerge and fade over time) in the output space with difficulty.

One of the unpleasant consequences of learning is forgetting. Unfortunately, it is an award for aging, but it is also due to disuse. Incremental methods tend to forget information they were fed during the learning process. Forgetting is also a natural way to reinforce what was learned. Sleep is extraordinarily important for memory and necessary for proper consolidation of new learning. How the brain forgets unnecessary memories is unclear. However, the retrieval of a target memory may lead to retrieval-induced forgetting of currently irrelevant competing memories [14]. In fact, the process operates at two levels in the brain: a cleaning of irrelevant information when memorizing, and a blocking of irrelevant information when remembering [15]. Perhaps a review of Richard Semon's theory of memory, originally published in 1904 [16], could help to algorithmically approach the complexity of memorizing, remembering, and forgetting.

One of the most challenging problems in supervised data stream learning is the availability of labelled examples. In some cases, labels are available with a delay, e.g., forecasting electricity prices for the next day; the ground truth is available in 24 h. In other cases, labels are not easily available. There are several research opportunities on semi-supervised learning for data streams and on active learning from data streams. By leveraging both labeled and unlabeled data, semi-supervised learning techniques can improve model performance and reduce the need for extensive manual labeling. This approach can be particularly beneficial in situations where obtaining labeled data is expensive or time-consuming. Active learning from data streams is another promising research direction that can help mitigate the scarcity of labeled data. Active learning techniques involve selecting the most informative instances from the data stream for labeling, thus reducing the labeling effort and improving model performance. By intelligently prioritizing which instances to label, active learning can help optimize the trade-off between labeling costs and model performance.

In recent times, artificial intelligence, in general, and machine learning, in particular, have a high societal impact. The European Commission is promoting proactive actions in regulating autonomous decision support systems in two directions: privacy and explainability. These regulations aim to ensure that AI-based systems maintain user trust, adhere to ethical guidelines, and remain transparent in their decision-making processes.

We are extraordinarily far from being able to deal algorithmically with volumes of information similar to those processed by the brain. One of the most interesting situations occurs when our senses are focused on a single object: for example, tasting a new dish. The brain processes what it sees, what it smells, and when we put it in our mouth, what it hears, tastes, and feels by touch with the tongue or lips (much of this information was poorly encoded quantitatively). These flows, with different typologies and speeds, help us identify (or create) concepts related to the sensory experience, although somewhat later, with some delay, we will form (or modify) other concepts associated with the bill.

Artificial intelligence is still a long way from natural intelligence. A better understanding of the brain will provide more insight into the learning process, and it will then be possible to emulate the findings with machine intelligence. Data streams analytics should provide models with sufficient stability and plasticity to be able to produce good results at any time, independently of how fast, irregular, or changing the incoming information is. Important issues were highlighted, but not all challenges, some inherited from traditional machine learning, that will become relevant in the near future, such as imbalance [17] or the curse of dimensionality [18] in data streams, were listed. The research opportunities in the field of *data stream analytics* are innumerable and challenging.

Author Contributions: All the authors have contributed equally to writing the draft and revising the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bushdid, C.; Magnasco, M.O.; Vosshall, L.B.; Keller, A. Humans can discriminate more than 1 trillion olfactory stimuli. *Science* **2014**, *343*, 1370–1372. [CrossRef] [PubMed]
2. McGann, J.P. Poor human olfaction is a 19th-century myth. *Science* **2017**, *356*, eaam7263. [CrossRef] [PubMed]
3. McKemy, D.D.; Neuhausser, W.M.; Julius, D. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* **2002**, *416*, 52–58. [CrossRef] [PubMed]
4. Woo, S.-H.; Lukacs, V.; de Nooij, J.C.; Zaytseva, D.; Criddle, C.R.; Francisco, A.; Jessell, T.M.; Wilkinson, K.A.; Patapoutian, A. Piezo2 is the principal mechonotransduction channel for proprioception. *Nat. Neurosci.* **2015**, *18*, 1756–1762. [CrossRef] [PubMed]
5. Murray, G.L.; Macefield, B.E.F. Processing efficiency of interacting data streams. *Nucl. Instrum. Methods* **1968**, *62*, 122–124. [CrossRef]
6. Hartzman, C.S.; Watters, C.R. A relational approach to querying data streams. *IEEE Trans. Knowl. Data Eng.* **1990**, *2*, 401–409. [CrossRef] [PubMed]
7. Henzinger, M.; Raghavan, P.; Rajagopalan, S. Computing on data streams. In *External Memory Algorithms*; Abello, J.M., Vitter, J.S., Eds.; DIMACS Series in Discrete Mathematics and Theoretical Computer Science; U.S. DIMACS: New Brunswick, NJ, USA, 1999; Volume 50.
8. Domingos, P.; Hulten, G. Mining high-speed data streams. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 20–23 August 2000.
9. Schlimmer, J.C.; Granger, R.H., Jr. Incremental learning from noisy data. *Mach. Learn.* **1986**, *1*, 317–354. [CrossRef]
10. Lange, S.; Zeugmann, T. Incremental learning from positive data. *J. Comput. Syst. Sci.* **1996**, *53*, 88–103. [CrossRef]
11. Khan, W.Z.; Ahmed, E.; Hakak, S.; Yaqoob, I.; Ahmed, A. Edge computing: A survey. *Future Gener. Comput. Syst.* **2019**, *97*, 219–235. [CrossRef]
12. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge Computing: Vision and Challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [CrossRef]
13. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 1–37. [CrossRef]
14. Anderson, M.C.; Bjork, R.A.; Bjork, E.L. Remembering can cause forgetting: Retrieval dynamics in long-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **1994**, *20*, 1063–1087. [CrossRef] [PubMed]
15. Josselyn, S.A.; Tonegawa, S. Memory engrams: Recalling the past and imagining the future. *Science* **2020**, *367*, eaaw4325. [CrossRef] [PubMed]
16. Semon, R.W. *Die Mneme als Erhaltendes Prinzip im Wechsel des Organischen Geschehens*; Wilhelm Engelmann: Leipzig, Germany, 1911; Volume 7101.
17. Wang, S.; Minku, L.L.; Yao, X. Resampling-based ensemble methods for online class imbalance learning. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1356–1368. [CrossRef]
18. Wellinger, R.E.; Aguilar–Ruiz, J.S. A new challenge for data analytics: Transposons. *BioData Min.* **2022**, *15*, 9. [CrossRef] [PubMed]