Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI

Andreas Holzinger[1,2(✉)], Peter Kieseberg[3,4], Edgar Weippl[3,5], and A Min Tjoa[6]

[1] Holzinger Group, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
a.holzinger@hci-kdd.org
[2] Institute of Interactive Systems and Data Science, Graz University of Technology, Graz, Austria
[3] SBA Research, Vienna, Austria
[4] University of Applied Sciences St. Pölten, St. Pölten, Austria
[5] Christian Doppler Laboratory for Security and Quality Improvement in the Production System Lifecycle, TU Wien, Vienna, Austria
[6] Information & Software Engineering Group, Institute of Information Systems Engineering, TU Wien, Vienna, Austria

Abstract. In this short editorial we present some thoughts on present and future trends in Artificial Intelligence (AI) generally, and Machine Learning (ML) specifically. Due to the huge ongoing success in machine learning, particularly in statistical learning from big data, there is rising interest of academia, industry and the public in this field. Industry is investing heavily in AI, and spin-offs and start-ups are emerging on an unprecedented rate. The European Union is allocating a lot of additional funding into AI research grants, and various institutions are calling for a joint European AI research institute. Even universities are taking AI/ML into their curricula and strategic plans. Finally, even the people on the street talk about it, and if grandma knows what her grandson is doing in his new start-up, then the time is ripe: We are reaching a new AI spring. However, as fantastic current approaches seem to be, there are still huge problems to be solved: the best performing models lack transparency, hence are considered to be black boxes. The general and worldwide trends in privacy, data protection, safety and security make such black box solutions difficult to use in practice. Specifically in Europe, where the new General Data Protection Regulation (GDPR) came into effect on May, 28, 2018 which affects everybody (right of explanation). Consequently, a previous niche field for many years, explainable AI, explodes in importance. For the future, we envision a fruitful marriage between classic logical approaches (ontologies) with statistical approaches which may lead to context-adaptive systems (stochastic ontologies) that might work similar as the human brain.

A. Holzinger et al. (Eds.): CD-MAKE 2018, LNCS 11015, pp. 1–8, 2018.
https://doi.org/10.1007/978-3-319-99740-7_1

Keywords: Machine learning · Knowledge extraction
Artificial intelligence · Explainable AI · Privacy

1 Introduction

Artificial intelligence (AI) has a long tradition in computer science, reaching back to 1950 and earlier [24]. In the first three decades, industry, governments and the public had extremely high expectations to reach the "mythical" human-level machine intelligence [9,17]. As soon as it turned out that the expectations were too high, and AI could not deliver these high promises, a dramatic "AI winter" affected the field; even the name AI was avoided at that time [8].

The field recently gained enormous interest due to the huge practical success in Machine Learning & Knowledge Extraction. Even in famous journals including Science [12] or Nature [15] the success of machine learning was recently presented. This success is visible in many application domains of our daily life from health care to manufacturing. Yet, many scientists of today are still not happy about the term, as "intelligence" is not clearly defined and we are still far away from reaching human-level AI [18].

Maybe the most often asked question is: "What is the difference between Artificial Intelligence (AI) and Machine Learning (ML) – and is deep learning (DL) belonging to either AI or ML?". A formal short answer: Deep Learning is part of Machine Learning is part of Artificial Intelligence: DL $\subset$ ML $\subset$ AI

This follows the popular Deep Learning textbook by Ian Goodfellow, Yoshua Bengio & Aaron Courville (2016, see Fig. 1):

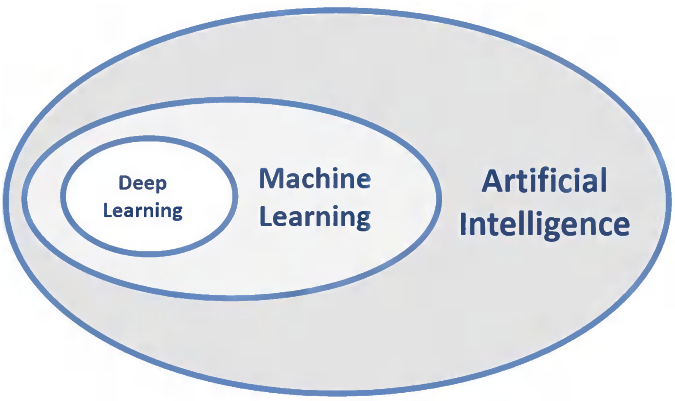

Fig. 1. A question most often asked: What is the difference between AI, ML and DL, see also [6].

2 Trend Indicators

Industry as Trend Indicator

Many global industrial players from Amazon to Zalando have now concerted international efforts in AI. The topic is so hot, that e.g. Google Brain has recently itself renamed to Google AI. Start-ups are emerging at an unprecedented rate - AI spring is here.

Funding as Trend Indicator

Worldwide, enormous grants are now fostering AI research generally and machine learning specifically: DARPA in the US or Volkswagen Stiftung in Germany are only two examples. The European Union targets for a total of 20 BEUR bringing into AI research in the future across both, public and private sectors. Health is one of the central targets, which is easy to understand as it is a topic that affects everybody. The primary direction was set in the last Horizon2020 initiative: The goal is to develop an European AI ecosystem, bringing together knowledge, algorithms, tools and resources available and making it a compelling solution for users, especially from non-tech sectors (such as health). The aim is to mobilize the European AI community including scientists, businesses and start-ups to provide access to knowledge, algorithms and tools. On the EU agenda are particularly ELSE aspects, where ELSE stands for Ethical, Legal and Socio-Economic issues.

At the same time there is the ELLIS initiative (https://ellis-open-letter.eu) which urges for seeing machine learning at the heart of a technological and societal artificial intelligence revolution involving multiple sister disciplines, with large implications for the future competitiveness of Europe. The main critique is that currently Europe is not keeping up: most of the top laboratories, as well as the top places to do a PhD, are located in North America or Canada; moreover, ML/AI investments in China and North America are significantly larger than in Europe. As an important measure to address these points, the ELLIS initiative proposes to found a *European Lab for Learning & Intelligent Systems* (working title; abbreviated as "ELLIS"), involving the very best European academics while working together closely with researchers from industry, ensuring to have economic impact and the creation of AI/ML jobs in Europe. This mission is meanwhile supported by IFIP TC 12.

In the UK the House of Lords (see the report by Wendy Hall and Jerome Presenti from October, 15, 2017: bit.ly/2HCEXhx) is convinced that the UK can lead in AI by building on a historically strong research program, which proposes five principles [19]: 1. AI should be developed for the common good and benefit of humanity. 2. AI should operate on principles of intelligibility and fairness. 3. AI should not be used to diminish the data rights or privacy of individuals, families or communities. 4. All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside AI. 5. The autonomous power to hurt, destroy or deceive human beings should never be vested in AI.

Conferences as Trend Indicator

A good indicator for the importance of machine learning is the conference on Neural Information Processing Systems (NIPS) - which is now trying to re-name itself. This conference was first held in Denver in December 1987 as a small meeting. The conference beautifully reflects the success of statistical learning methods attracting more and more researchers from machine learning (see Fig. 2).

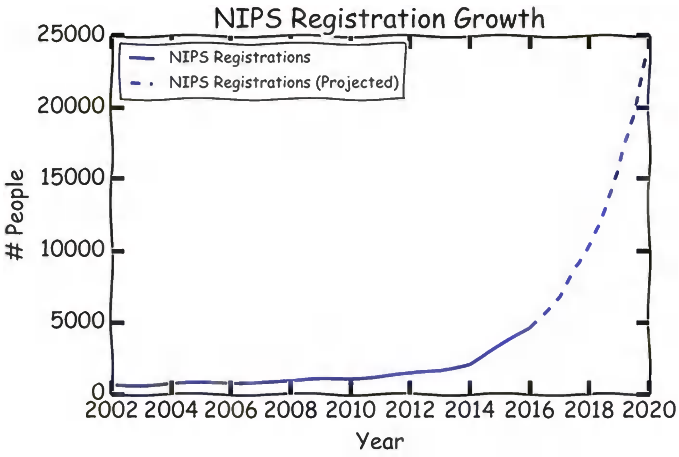

Fig. 2. NIPS 2017 in Long Beach was the most popular ML conference yet, attracting over 8,000 registered attendees, following the 2016 event with 6,000 registered attendees in Barcelona (image taken from Ian Goodfellow's tweet on June, 15, 2018).

3 Main Problems Today

A major issue with respect to explaining machine learning algorithms lies in the area of privacy protection: Trust is one of the core problems when dealing with personal, and potentially sensitive, information, especially when the algorithms in place are hard or even impossible to understand. This can be a major risk for acceptance, not only by the end users, like e.g. hospital patients, or generally in safety-critical decision making [10], but also among the expert engineers that are required to train the models, or, in case of an expert-in-the-loop approach, partake in the daily interaction with the expert system [14]. One option is to include risk management practice early in the project to manage such risks [11]. Trust and Privacy are actually a twofold problem in this regard; an example from the medical domain shall illustrate this: The patients need to be able to trust the machine learning environment that their personal data is secured and protected against theft and misuse, but also that the analytical processes working on their data are limited to the selection they have given consent to.

For the expert, on the other hand, there is the need to trust the environment that their input to the system is not manipulated later on. Furthermore, usability is a fundamental factor for successfully integrating experts into AI systems, which, again, requires the designers of the interfaces to understand the fundamentals of the system in place. Here it must be noted that usability and security are often considered fundamental opposites, hence research in the so-called area of *usable security* [3] is urgently needed.

A topic closely related to the issue of security and privacy, but still different in nature, is the issue of fingerprinting/watermarking information [22]. Many approaches in utilizing data for data driven research face the problem that data must be shared between partners, i.e. data sets are either sent to a central analysis repository for further processing, or directly shared between the partners themselves. While the earlier approach allows for some kind of control over the data by the trusted third party operating the analysis platform, in the later one, the original owner potentially gives up control over the data set. This might not even be a problem with respect to privacy, as the data shared with the other partners will in most cases obey data protection rules as put forth by various regulations, still, this data might be an asset of high (monetary) value. Thus, when sharing the data with other partners, it must be made sure that the data is not further illegally distributed. A typical reactive approach to this problem is the implementation of so-called *fingerprints* or *watermarks*; these can also be used to embedded information that helps to detect collusion in deanonymization attacks [13,21]. Both terms, fingerprinting and watermarking, are often used synonymously by authors, while others differentiate them as watermarks being mechanisms that prove the authenticity and ownership of a data set and fingerprints actually being able to identify the data leak by providing each partner with the same basic set marked with different fingerprints.

Throughout the past decades, watermarking and fingerprinting of information has gained a lot of attention in the research community, most notably regarding the protection of digital rights in the music and movie industries [23]. Approaches for marking data have also been put forth (e.g. [1]) and while a lot of them exist nowadays, most of them only focus on marking whole data sets and fail with partially leaked sets. Thus, in order to provide transparency with respect to privacy, as well as explainability, we propose that a fingerprinting mechanism within data driven research requires the following criteria:

1. **Single Record Detection:** The detection of the data leak should be possible with only one single leaked (full) record. This is a major obstacle for most algorithms that rely on adding or removing so-called *marker*-records from the original data set.
2. **Collusion Protection:** Several partners being issued the same fingerprinted data set might collude in order to extract and remove the fingerprints, or even frame another partner. The fingerprinting algorithm is required to be stable against such kinds of attacks.
3. **High Performance:** In order to make this protection mechanism usable, it must not require a lot of resources, neither with respect to calculation time

(for both, the generation of the fingerprint, as well as the detection), nor with respect to additional storage requirements.
4. **Low distortion:** The algorithm must not introduce a large amount of additional distortion, thus further reducing the value of the data used in the analysis.

The development of novel techniques in this area is thus another open problem that has a high potential for future research. When developing new solutions contradicting requirements including future improvements in "counter-privacy", aka. forensics [2], have to be considered.

Last, but not least, the need to understand machine learning algorithms is required to deal with distortion: Due to novel regulations in the European Union, especially the General Data Protection Regulation (GDPR), the protection of privacy has become extremely important and consent for processing personal information has to be asked for rather narrow use cases, i.e. there is no more "general consent". Thus, research labs tend to consider anonymizing their data, which makes it non-personal information and thus consent-free to use. Still, as it has already been shown [16], many standard anonymization techniques introduce quite a large amount of distortion into the end results of classical machine learning algorithms. In order to overcome this issue, additional research in the area of Privacy Aware Machine Learning (PAML) is needed: The distortion needs to be quantified in order to be able to select the anonymization algorithm/machine learning algorithm pairing that is ideal with respect to the given data set. Explainable AI can be a major enabler for this issue, as understanding decisions would definitely help in understanding and estimating distortions. In addition, algorithms (both, for anonymization and machine learning) need to be adapted in order to reduce the distortion introduced, again, a task where the black-box characteristics of machine learning nowadays is an issue. Thus, explainable AI could be the key to designing solutions that harness the power of machine learning, while guaranteeing privacy at the same time.

4 Conclusion

To provide an answer to the question *"What are the most interesting trends in machine learning and knowledge extraction?":* the most interesting ones are not known yet. What we know is that the driver for the AI hype is success in machine learning & knowledge extraction. A promising future approach is the combination of ontologies with probabilistic approaches. Traditional logic-based technologies as well as statistical ML constitute two indispensable technologies for domain specific knowledge extraction, actively used in knowledge-based systems. Here we urgently need solutions on how the two can be successfully integrated, because to date both technologies are mainly used separately, without direct connection.

The greatest problem, however, is the problem of black box algorithms. These make machine decisions intransparent and non-understandable, even to the eyes of experts, which reduces trust in ML specifically and AI generally.

Another field that requires more research is the intersection between security (and especially privacy related) research and ML - be it in the form of privacy aware machine learning, where the distortion from data protection mechanisms is mitigated, or rather in the areas of protecting ownership on information or providing trust into the results of ML algorithms. All of these areas could greatly benefit from explainable AI, as the design of novel mechanisms to achieve these security and privacy tasks cannot be soundly done without further insight into the internal workings of the systems they are protecting.

A final remark of applications: According to the ML initiative of the Royal Society the greatest benefit of AI/ML will be in improved medical diagnosis, disease analysis and pharmaceutical development. This on the other hands needs making results transparent, re-traceable and to understand the *causality of learned representations* [4,20].

Consequently, the most promising field in the future is what is called explainable AI [5] where DARPA has already launched a funding initiative in 2016 [7]. This calls for a combination of logic-based approaches (ontologies) with probabilistic machine learning to build *context adaptive systems*.

Acknowledgements. The authors thank their colleagues for valuable feedback, remarks and critics on this editorial introduction. The competence center SBA Research (SBA-K1) is funded within the framework of COMET – Competence Centers for Excellent Technologies by BMVIT, BMDW, and the federal state of Vienna, managed by the FFG. This research was also funded by the CDG Christian Doppler Laboratory SQI and by the KIRAS program of the FFG.

References

1. Agrawal, R., Kiernan, J.: Watermarking relational databases. In: VLDB 2002: Proceedings of the 28th International Conference on Very Large Databases, pp. 155–166. Elsevier (2002)
2. Frühwirt, P., Kieseberg, P., Schrittwieser, S., Huber, M., Weippl, E.: Innodb database forensics: reconstructing data manipulation queries from redo logs. In: 2012 Seventh International Conference on Availability, Reliability and Security (ARES), pp. 625–633. IEEE (2012)
3. Garfinkel, S., Lipford, H.R.: Usable security: history, themes, and challenges. Synthesis Lectures on Information Security, Privacy, and Trust **5**(2), 1–124 (2014)
4. Gershman, S.J., Horvitz, E.J., Tenenbaum, J.B.: Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. Science **349**(6245), 273–278 (2015)
5. Goebel, R.: Explainable ai: the new 42? In: Holzinger, A., et al. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018)
6. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (MA) (2016)
7. Gunning, D.: Explainable artificial intelligence (XAI): Technical report Defense Advanced Research Projects Agency DARPA-BAA-16-53. DARPA, Arlington, USA (2016)
8. Hendler, J.: Avoiding another ai winter. IEEE Intell. Syst. **23**(2), 2–4 (2008)

9. Hernández-Orallo, J.: The Measure of all Minds: Evaluating Natural and Artificial Intelligence. Cambridge University Press, Cambridge (2016)
10. Holzinger, K., Mak, K., Kieseberg, P., Holzinger, A.: Can we trust machine learning results? artificial intelligence in safety-critical decision support. ERCIM News **112**(1), 42–43 (2018)
11. Islam, S., Mouratidis, H., Weippl, E.R.: An empirical study on the implementation and evaluation of a goal-driven software development risk management model. Inf. Softw. Technol. **56**(2), 117–133 (2014)
12. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
13. Kieseberg, P., Schrittwieser, S., Mulazzani, M., Echizen, I., Weippl, E.: An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata. Electron. Markets **24**(2), 113–124 (2014)
14. Kieseberg, P., Weippl, E., Holzinger, A.: Trust for the doctor-in-the-loop. ERCIM News **104**(1), 32–33 (2016)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
16. Malle, B., Kieseberg, P., Schrittwieser, S., Holzinger, A.: Privacy aware machine learning and the right to be forgotten. ERCIM News **107**(10), 22–3 (2016)
17. McCarthy, J.: Programs with common sense. pp. 75–91. RLE and MIT Computation Center (1960)
18. McCarthy, J.: From here to human-level ai. Artif. Intell. **171**(18), 1174–1182 (2007)
19. Olhede, S.: The AI spring of 2018. Significance **15**(3), 6–7 (2018)
20. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. Cambridge, MA (2017)
21. Schrittwieser, S., Kieseberg, P., Echizen, I., Wohlgemuth, S., Sonehara, N., Weippl, E.: An algorithm for k-anonymity-based fingerprinting. In: Shi, Y.Q., Kim, H.-J., Perez-Gonzalez, F. (eds.) IWDW 2011. LNCS, vol. 7128, pp. 439–452. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32205-1_35
22. Shih, F.Y.: Digital Watermarking and Steganography: Fundamentals and Techniques. CRC Press, Boca Raton (2017)
23. Swanson, M.D., Kobayashi, M., Tewfik, A.H.: Multimedia data-embedding and watermarking technologies. Proc. IEEE **86**(6), 1064–1087 (1998)
24. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)