Life at the edge: accelerators and obstacles to emerging ML-enabled research fields

Soukayna Mouatadid
Department of Computer Science
University of Toronto
soukayna@cs.toronto.edu

Steve Easterbrook
Department of Computer Science
University of Toronto
sme@cs.toronto.edu

Abstract

Machine learning is transforming several scientific disciplines, and has resulted in the emergence of new interdisciplinary data-driven research fields. Surveys of such emerging fields often highlight how machine learning can accelerate scientific discovery. However, a less discussed question is how connections between the parent fields develop and how the emerging interdisciplinary research evolves. This study attempts to answer this question by exploring the interactions between machine learning research and traditional scientific disciplines. We examine different examples of such emerging fields, to identify the obstacles and accelerators that influence interdisciplinary collaborations between the parent fields.

1 Introduction

In recent decades, several scientific research fields have experienced a deluge of data, which is impacting the way science is conducted. Fields such as neuroscience, psychology or social sciences are being reshaped by the advances in machine learning (ML) and the data processing technologies it provides. In some cases, new research fields have emerged at the intersection of machine learning and more traditional research disciplines. This is the case for example, for bioinformatics, born from collaborations between biologists and computer scientists in the mid-80s [1], which is now considered a well-established field with an active community, specialized conferences, university departments and research groups. How do these transformations take place? Do they follow similar paths? If yes, how can the advances in such interdisciplinary fields be accelerated?

Researchers in the philosophy of science have long been interested in these questions. In [2], Kuhn suggested that scientific research operates within a set of traditions, called paradigm. When new problems arise that cannot be solved under the existing paradigm, a paradigm shift occurs, and a new research field emerges. [3] studied the dynamics and evolution of scientific fields using a network-based analysis and found that the cross-fertilization of established research fields often precedes the emergence of new fields. Several other studies have explored the question of how scientific collaboration networks appear, grow and fade away [4, 5, 6]. In all these cases, the examined emerging field results from interactions between various well-established parent fields. However, never before has a single field had such a rapid impact, and as much potential for impact, on a variety of scientific disciplines as AI. Consequently, several studies have examined the opportunities resulting from the application of ML to scientific discovery [7, 8, 9, 10]. Yet, the rate of progress, when applying ML, can vary from one scientific field to another. The study of collaborations between ML and existing fields can highlight common patterns and shed light on key factors that accelerate the emerging field or instead hold back its progress.

Our goal is to move beyond anecdotal evidence of the impact of ML on traditional fields and the benefits of interdisciplinary research for science. Instead, in the next section, we draw on the literature from a wide range of ML-enabled research fields and identify the main factors capable of slowing

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

down or accelerating progress in these emerging fields. We summarize these insights as a table of the key factors impacting the rate of progress in interdisciplinary fields.

2 Places to intervene in the system

2.1 Obstacles

Publicly available datasets are a pre-requisite for interdisciplinary research involving AI. However, after examining several emerging data-driven fields, data-related obstacles can be identified. In fields such as health-care [11] and cyber-security [12], privacy concerns present an obstacle that can stall the adoption of ML technologies. In other fields such as marketing and legal research areas [13], the 'creepy stalker factor' associated with data collection practices can also slow down interdisciplinary collaborations. Finally, in a number of traditional scientific disciplines, such as climate modelling and cancer research, datasets present unique aspects like non-stationarity and paucity of representative samples, that differ from data science problems usually encountered in commercial applications [14].

Even when an interdisciplinary approach proves to be productive, collaborations between parent fields can slow down when the two fields grow separately in complexity and their disciplinary boundaries solidify [15]. It becomes then harder for respective researchers to specialize in one field, let alone both fields. In other cases, technical challenges can hold back productive collaborations as is the case for quantum machine learning. In this case, challenges such as quantum states preparation and measurements still hinder the adoption of QML technologies by both ML and quantum computing researchers. In a different context, disciplines such as climate modelling for instance, involve the study of various physical processes interacting on different space and time scales. Using machine learning techniques to model such processes in a modular way through specialized collaborations (e.g., oceanography and ML experts, atmosphere and ML experts, etc) contradicts the entanglement that characterizes non-linear climate dynamics. As David Randall puts it: "The trouble with modular models is that nature is not modular" [16]. Physical processes involved in the climate system are subject to intertwined feedback loops and cannot be formulated independently of one another.

In addition, ML researchers proposing ML-based solutions for problems in different fields can sometimes report findings and insights that were long known in the parent field, which reduces the acceptability of ML research in that field. Finally, another phenomenon slowing down progress in interdisciplinary fields is the fast-paced competition in machine learning research, which encourages winning challenges by conducting empirical studies on public datasets, rather than developing insight and understanding [17]. These factors, combined with the lack of interpretability of ML models [18, 19], hinders the acceptance of other fields for machine learning solutions. In addition to the fields' complexity, the traditions of the parent fields also play a crucial role in slowing down collaborations. When the parent fields are steeped in different traditions around the communication of results, the language used, the frequency of publications, the length of publications, etc, the objectives of young researchers become misaligned and these differences send a signal discouraging interdisciplinary research projects. In particular, ML fields are characterized by fast and frequent publication cycles in conference venues which require relatively short papers compared to more traditional scientific disciplines where lengthy journal submissions are considered standard.

Another obstacle is the hype surrounding the emerging data-driven fields. When the application of ML technologies to a scientific discipline is hyped through over-promising and is followed by under-delivery of results and insights, real progress is slowed down by a breach of trust between the two fields. The case of quantum supremacy illustrates this point [20]: although real progress is being made in the field, researchers expect it will take several more years before quantum computers show their worth [21]. And when real progress is achieved, there could be a perceived risk by researchers in the traditional field that they will be replaced by AI. This perceived risk can implicitly push scientists to dismiss ML solutions that are otherwise valid. This rejection is further worsened by some issues plaguing the machine learning research fields recently: namely, reproducibility and hyper-parameter tuning of ML models [22, 17]. Researchers in scientific fields are aware of the ML reproducibility crisis, and may dismiss valid ML solutions as a consequence.

Finally, another set of obstacles is faced by researchers publishing findings at the intersection of ML and another traditional field. These obstacles have to do with the supply of skilled reviewers to evaluate the interdisciplinary contributions. Sculley and colleagues [17] argue that it takes years to train skilled reviewers in ML areas. This means it would take even longer to train reviewers who

have an understanding of two different fields, who would recognize the potential of interdisciplinary research, and not force researchers working at the boundaries to re-present their work as more *centric* to just one of the parent fields [23].

2.2 Accelerators

Just as there are obstacles to the progress of data-driven interdisciplinary research fields, there are also factors than can accelerate progress (see Table 1). One of the key accelerators for an emerging field is the availability of comprehensive and well-written papers that set the vision for the new field, summarize advances in the parent fields for a mixed audience, use unifying language and metaphors [15], emphasize the complementarity of the parent fields [21], highlight the novel and important research questions as well as their priority [24] and identify research 'low hanging fruits'. These 'low hanging fruits' are easily achievable research breakthroughs available to the first come to the boundary research and include problems in one field solvable using tools from the second field in a straightforward manner. Identifying such problems can attract young researchers to the emerging field. Another accelerator is the adoption of a theory-guided data science paradigm [25, 26] and thoughtful ML principles [27]. The advantage of adopting such approaches is to leverage the knowledge accumulated in scientific principles by introducing scientific consistency and enabling scientific interpretability in data-driven models [25].

The second set of accelerators revolves around datasets. Publicly available benchmark datasets can accelerate collaborations between ML and scientific disciplines. Ebert and colleagues [28] highlight 10 key characteristics of ideal benchmark datasets to build bridges between geoscience and data sciences. These characteristics include high impact, active research area, means for evaluating success and citability. Sim and colleagues [29] discuss the importance of benchmark datasets developed by research community consensus to show agreement on which problems are important, concentrating the community's attention on them, and hence increasing the scientific maturity of emerging research fields. However, with respect to the ML community's focus on winning the benchmark competition versus advancing the understanding, these benchmark datasets should be accompanied with an expiration date, after which the benchmarks should retire. Such a precaution helps prevent overfitting on these datasets, which results in wasted effort from non-generalizable results.

Reproducibility of published empirical work accelerates and sustains progress in emerging fields as it enables accurate judgement of the improvement offered by new methods [22]. To promote reproducibility, Sculley and colleagues [17] suggest that sharing experimental notes and records can help with issues of multiple hypothesis testing and post-hoc explanations. They also argue that complete empirical evaluations are likely to involve large groups of collaborators. In this context, improved credit assignment can incentivize such collaborations. In the case of interdisciplinary fields, published articles should acknowledge the contributions and inspirations, however subtle, from the other field, whenever it's relevant. Also, flexible paper structures can accelerate collaborations progress by destroying the barriers of disciplines' traditions. Alternative paper formats can be adopted, with different versions (short, long, with or without code, data and analysis, reviews and answers) overlayed in a centralized system [30]. Reviewers for these interdisciplinary collaborations can be trained through specialized MOOCs [31]. Finally, taking inspiration from the field of quantum machine learning, interdisciplinary emerging fields should provide open access of relevant resources to early adopters. Companies such as Google, Intel and Microsoft, as well as startups like Rigetti computing, IonQ and Quantum Circuits made quantum computers accessible via the cloud to early adopters. This open access of resources can accelerate progress in two ways. First, early business adopters can provide valuable feedback and revenue stream for startups (for QML, such opportunities include financial modelling, chemistry and route optimization modelling) [32]. Second, the full ML and quantum computing revolution will be carried with a new generation of students and researchers that will get to play with practical machines and contribute to a paradigm change as quantum computers require different programming languages and fundamentally different ways of thinking about what programming is [21].

Table 1: Obstacles and accelerators to collaborations between ML and scientific disciplines

	Obstacles	Accelerators
Research	• Complexity and technical challenges in parent fields; • Non-modularity and non-stationarity in scientific principles and disciplines [16]; • Focus on winning challenges on public datasets, as opposed to developing insights and understanding [17]; • Hyper-parameter tuning and multiple hyper-parameter testing [17]; • Cheap development but expensive maintenance of ML systems [33]; • CACE principle (changing anything changes everything) [33]; • Research solutions providing tiny accuracy benefits accompanied with massive increase in system complexity [33];	• Publish survey articles setting the direction of the emerging field; • Use unifying language and metaphors [15, 34, 35]; • Emphasize complementarity of parent fields; • Identify research low hanging fruits; • Highlight of important and novel questions; • Establish ML best practices (similarly to software engineering best practices) [36]; • Adopt a theory-guided ML paradigm [25]; • Open access to code and data resources [37]; • Demonstrate impact on other fields beyond the parent fields; • Establish reliable baselines [22];
Data	• Privacy concerns [37]; • Creepy stalker factor; • Specificities of datasets in scientific disciplines such as non-stationarity, paucity of labelled representative samples, spatial and temporal correlations, high dimensionality, multi-resolution and interest in rare events [14]; • Limitations and price of traditional computational resources;	• Publicly available datasets; • High volume and quality of data; • Awareness of strengths and weaknesses in data collection processes [14]; • Identification of pre-processing steps [14]; • Benchmarks with 10 key properties: problems challenging for data scientists, data science generality and versatility, rich information content, hierarchical problem statements, means for evaluating success, quick start guide, context and citability [28]; • Retirement of benchmark datasets [29]; • Alternative sources of data (e.g., GAN simulations);
Community	• Lack of reproducibility [22]; • Traditions of parent fields; • Gap in size between ML theory and applications communities; • Gap in size between models' development and models' applications communities • Hype [38]; • Perceived risk/threat of being replaced; • Open code and prisoners' dilemma [31]; • Lack of trained reviewers due to rapid advances in ML [17]; • Publication of interdisciplinary research in traditional conferences; • Reviewers forcing boundary research to become centric research [23]; • Reviewers crunch during conference cycles [30];	• Reproducibility [39, 40]; • Standard requirements for empirical evaluations [16]; • Sharing of experiment code notes and records; • Academic licenses for open code [31]; • Acknowledgement of contributions and inspirations; • Credit assessment and attribution [17]; • Open access to rejected papers and reviews; • Flexible paper structures; • Availability of different versions of the same article in a centralized repository [30]; • Open review systems [41]; • Online training for reviewers [31]; • Options for publication venues and flexible conference formats; • Increase in size and diversity of the models' development community; • University training of modelers with more degrees of freedom compared to government and private laboratories focusing on applications [42]; • Open access to resources for early private and business adopters; • Alternative sources of funding through research start-ups and industrial research;

Contribution and acknowledgments

Following the recommendations in [17], a brief contribution statement is presented: primary author led the research, wrote the first draft of the paper and edited the final version of the paper. Secondary author took part in all discussions of the ideas presented, and edited the final version of the paper.

References

[1] Teresa K Attwood and David J Parry-Smith. *Introduction to bioinformatics*. Prentice Hall, 2003.

[2] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

[3] Mark Herrera, David C Roberts, and Natali Gulbahce. "Mapping the evolution of scientific fields". In: *PloS one* 5.5 (2010), e10355.

[4] Bruce W Herr II, Russell J Duhon, Katy Börner, Elisha F Hardy, and Shashikant Penumarthy. "113 years of physical review: Using flow maps to show temporal and topical citation patterns". In: *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE. 2008, pp. 421–426.

[5] Loet Leydesdorff. "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals". In: *Journal of the Association for Information Science and Technology* 58.9 (2007), pp. 1303–1319.

[6] Mark EJ Newman. "The structure of scientific collaboration networks". In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[7] Eric Mjolsness and Dennis DeCoste. "Machine learning for science: state of the art and future prospects". In: *science* 293.5537 (2001), pp. 2051–2055.

[8] Yolanda Gil, Mark Greaves, James Hendler, and Haym Hirsh. "Amplify scientific discovery with artificial intelligence". In: *Science* 346.6206 (2014), pp. 171–172.

[9] Hiroaki Kitano. "Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery". In: *AI magazine* 37.1 (2016), pp. 39–49.

[10] Andrea Mannocci, Angelo A Salatino, Francesco Osborne, and Enrico Motta. "2100 AI: Reflections on the mechanisation of scientific discovery". In: (2017).

[11] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[12] Sumeet Dua and Xian Du. *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[13] Omer Tene and Jules Polonetsky. "A theory of creepy: technology, privacy and shifting social norms". In: *Yale JL & Tech.* 16 (2013), p. 59.

[14] Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. "Machine Learning for the Geosciences: Challenges and Opportunities". In: *arXiv preprint arXiv:1711.04708* (2017).

[15] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. "Neuroscience-inspired artificial intelligence". In: *Neuron* 95.2 (2017), pp. 245–258.

[16] David Randall. "The evolution of complexity in general circulation models". In: *The De-705 velopment of Atmospheric General Circulation Models. Complexity, Synthesis and Com-706 putation* (2010), p. 272.

[17] D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. "Winner's Curse? On Pace, Progress, and Empirical Rigor". In: (2018).

[18] Alun Preece. "Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges". In: *Intelligent Systems in Accounting, Finance and Management* (2018).

[19] Jaegul Choo and Shixia Liu. "Explainable, Interactive Deep Learning". In: (2018).

[20] Barbara M Terhal. "Quantum supremacy, here we come". In: *Nature Physics* (2018), p. 1.

[21] Will Knight. "Serious quantum computers are finally here. What are we going to do with them?" In: *MIT Technology Review* (2018).

[22] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. "Deep reinforcement learning that matters". In: *arXiv preprint arXiv:1709.06560* (2017).

[23] Brian Wansink and Koert Ittersum. "Boundary research: Tools and rules to impact emerging fields". In: *Journal of Consumer Behaviour* 15.5 (2016), pp. 396–410.

[24] Christian Jakob. "Accelerating progress in global atmospheric model development through improved parameterizations: Challenges, opportunities, and strategies". In: *Bulletin of the American Meteorological Society* 91.7 (2010), pp. 869–876.

[25] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. "Theory-guided data science: A new paradigm for scientific discovery from data". In: *IEEE Transactions on Knowledge and Data Engineering* 29.10 (2017), pp. 2318–2331.

[26] Nicholas Wagner and James M Rondinelli. "Theory-guided machine learning in materials science". In: *Frontiers in Materials* 3 (2016), p. 28.

[27] Yolanda Gil. "Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery". In: *Data Science* Preprint (2018), pp. 1–11.

[28] I Ebert-Uphoff, DR Thompson, I Demir, A Karpatne, M Guereque, V Kumar, E Cabral-Cano, and P Smyth. "A vision for the development of benchmarks to bridge geoscience and data science". In: *7th International Workshop on Climate Informatics*. 2017.

[29] Susan Elliott Sim, Steve Easterbrook, and Richard C Holt. "Using benchmarking to advance research: A challenge to software engineering". In: *Software Engineering, 2003. Proceedings. 25th International Conference on*. IEEE. 2003, pp. 74–83.

[30] David Soergel, Adam Saunders, and Andrew McCallum. "reviews of Open Scholarship and Peer Review: a Time for Experimentation". In: (2013).

[31] D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. "reviews of Winner's Curse? On Pace, Progress, and Empirical Rigor". In: (2018).

[32] Will Knight. "A startup uses quantum computing to boost machine learning". In: *MIT Technology Review* (2017).

[33] D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. "Hidden technical debt in machine learning systems". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2503–2511.

[34] Nicholas M Luscombe, Dov Greenbaum, Mark Gerstein, et al. "What is bioinformatics? A proposed definition and overview of the field". In: *Methods of information in medicine* 40.4 (2001), pp. 346–358.

[35] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. "An introduction to quantum machine learning". In: *Contemporary Physics* 56.2 (2015), pp. 172–185.

[36] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. "What's your ML Test Score? A rubric for ML production systems". In: *NIPS Workshop on Reliable Machine Learning in the Wild*. 2016.

[37] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. "Opportunities and obstacles for deep learning in biology and medicine". In: *bioRxiv* (2018), p. 142760.

[38] Gary Marcus. "Deep Learning: A Critical Appraisal". In: *arXiv preprint arXiv:1801.00631* (2018).

[39] Ricky J Sethi and Yolanda Gil. "Reproducibility in computer vision: Towards open publication of image analysis experiments as semantic workflows". In: *e-Science (e-Science), 2016 IEEE 12th International Conference on*. IEEE. 2016, pp. 343–348.

[40] Victoria Stodden, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John PA Ioannidis, and Michela Taufer. "Enhancing reproducibility for computational methods". In: *Science* 354.6317 (2016), pp. 1240–1241.

[41] David Soergel, Adam Saunders, and Andrew McCallum. "Open Scholarship and Peer Review: a Time for Experimentation". In: *ICML*. 2013.

[42] David A Randall. "A university perspective on global climate modeling". In: *Bulletin of the American Meteorological Society* 77.11 (1996), pp. 2685–2690.