Data Literacy to Support Human-centred Machine Learning

Conference or Workshop Item

Data Literacy to Support Human-centred Machine Learning

Annika Wolff
Department of Computing and Communications, The Open University
Milton Keynes, UK
annika.wolff@open.ac.uk

Daniel Gooch
Department of Computing and Communications, The Open University
Milton Keynes, UK
daniel.gooch@open.ac.uk

Gerd Kortuem
Department of Computing and Communications, The Open University
Milton Keynes, UK
Gerd.kortuem@open.ac.uk

Abstract
In the past, machine learning applications were mostly developed and deployed in specialist situations where the outputs would be either read by experts, or else interpreted for the public, with the methods hidden from view. In the current data driven society, the general public are increasingly interacting with complex data sets and the outputs of machine learning technologies. Within the domain of the smart city, non-experts are also being brought closer to the design process itself. This paper explores whether improving the overall data literacy of a society can instill within that society a set of core competences that improve the capacity of non-experts in machine learning to engage with machine learning outputs in a more knowledgeable way, or to provide insight and differing perspectives into the design of machine learning applications.

Author Keywords
Data literacy; machine learning; smart cities; competences; human-data interaction

ACM Classification Keywords
H.1.2. Information systems: User/Machine Systems; Human factors

INTRODUCTION

We are now living in a time where interaction with complex data is an everyday occurrence. Where once machine learning was the domain of experts, the general population are coming much closer not just to the outputs of machine learning, but also in some cases to being part of the design process itself. This paper explores the scenarios in which non-experts and machine learning interact and proposes that data literacy is becoming an important skill to support these interactions. Data literacy can facilitate *non-experts* to participate in design of machine learning applications and can facilitate machine learning *experts* to ensure that they target their expertise towards solving real problems.

Non-experts as consumers of machine learning outputs

These days it is becoming increasingly common that the outputs of machine learning algorithms are being targeted across large populations of non-expert users. People contribute data through their actions, leaving digital traces of their lives and habits. Data, by itself, has little value. The value is provided through interpretation of one or more data sets in a given context. Collected data is processed and presented in a variety of different ways. For example, machine learning algorithms are applied to data to make predictions on consumer habits, and to provide recommendations of web-resources, books, music, movies and to help consumer decision-making [8]. In the field of learning analytics, predictions about student performance is provided to a range of stakeholders, such as students, teachers and faculty staff, to provide actionable insights to improve the outcomes of learning. In the smart city, citizens interact with applications that use complex data analysis to help them live more efficient and sustainable lives, for example predicting their heating and other energy needs, or providing intelligent information about the city transport infrastructure to help them move around more efficiently.

When machine learning outputs are designed for the end-user experience, a common approach is to shield the non-expert end-users from the complexities of the operational algorithms [1]. This might be a practical approach for simple web recommendations about books and movies. However, this shielding of end-users can have detrimental effects in some situations. Non-expert users put themselves at risk by not understanding how their data might be used, in particular how data that is consciously provided might be combined with additional data sources to make inferences about them, or situations in which data can be derived through observation and about which they may be unaware. This issue is at the heart of the emerging field of study of Human-Data Interaction [11], which investigates how greater transparency can benefit users giving them better insight into the use - or sometimes misuse - of their personal data. Other reasons why shielding end-users from the machine learning process is not always a good idea is that non-experts might misuse machine learning applications and their outputs [1]. It may also limit opportunities for eliciting feedback of non-expert users for improving algorithms.

Non-experts as makers of machine learning technology

There is an increasing push to involve citizens of a smart city in bottom-up design of innovations. The rationale behind this bottom-up approach is that citizens in this scenario are the domain experts, with clearer insight into their own local problems [9][2].

Moving from top-down to bottom-up smart city innovation shifts citizens from a role as passive users/consumers of technology and contributors of data, through to active participants in identifying problems that could be solved with data and in some cases as innovators who shape and implement solutions to urban problems [12].

These active participants and innovators are analogous to *makers*, or hobbyists (e.g. see [9]) using data as part of a broader DIY tool-kit to individually, or collaboratively, design solutions to urban problems through the process of *bricolage*. Citizen-led smart city innovation is facilitated by the opening up of a large number of data sets [6][7] that can be used as part of this bricolage process.

But these approaches are often based on complex data analysis - skills that the citizens may not have. A *maker* needs a range of ad-hoc competencies depending on the focus of their innovation. These include (but are not limited to) knowing how to find and use the open data that is published about their environment, how to generate and use data, e.g. from sensors, and how to analyse data, including knowledge of machine learning techniques.

Machine learning experts as drivers of innovation

Traditionally, machine learning applications have been both instigated and designed not by 'engaged citizens' but by machine learning experts. It is quite common for such experts to apply techniques in a domain they are unfamiliar with. Therefore, during the design process the machine learning experts will elicit input from a small number of specialists from the application domain. This synthesis between domain expertise and machine learning expertise is critical, since reliance only on the machine learning experts in the design process can lead to solutions which do not fully address end-user needs [1]. Very rarely does an individual encompass both the required machine learning knowledge and the domain knowledge. Inclusion of some domain specialists allows the machine learning experts to capture and utilise domain knowledge to help in framing the problem to be solved, to provide insight into locating and understanding available data and to facilitate feature selection, choice of analysis and interpretation of the output of analyses performed.

Three Roles

Through the above analysis we have revealed three different roles that humans may take when interacting with machine learning. These are:

1. Humans as consumers – who need skills to interpret machine learning outputs that are increasingly presented as part of their every day life and to protect the use of their personal data.
2. Humans as makers – who need the skills to integrate machine learning approaches into broader overall strategies for identifying and solving real-world problems.
3. Humans as machine learning experts –who must combine their strong technical data skills with knowledge of the domain of the data.

Skills gap

From the above scenarios, it is possible to identify a potential skills gap. On one side, there are the machine learning experts, who without careful guidance, may pose and solve problems that do not meet user needs. However, once they have a clearly-defined problem to solve they should be able to identify appropriate

domain specialists and other stakeholders to help them in the design process. Alternatively, there are domain specialists and stakeholders, in the form of smart city citizens, who have good insight into their problems, but not the requisite understanding of what machine learning is or how it can be applied to data. This has consequences for their ability to instigate the design. However, if they had an improved higher level understanding of machine learning they could potentially quite easily either identify and solicit input from appropriate machine learning specialists, or use their knowledge to acquire the specific skills needed. In addition, an informed citizenry would be better protected from misuse of their data.

The case for data literacy

The term *data literacy* is used to broadly describe the set of abilities around the use of data as part of everyday thinking and reasoning for solving real-world problems. Data literacy is increasingly considered to be a life skill, as daily interactions with data become evermore commonplace [4] and individuals more frequently make judgments from data and make decisions regarding the use of their own personal data [3].

There is no single clear definition of data literacy. However, it is possible to identify commonalities amongst them. Mandinach and Gummer [5] propose a definition of data literacy in the context of supporting teachers to use student data to improve their practice, as a type of learning analytics. In their view, data literacy is:

> *"the ability to understand and use data effectively to inform decisions. It is composed of a specific skill set and knowledge base that enables educators to transform data into information and ultimately into actionable knowledge. These skills include knowing how to identify, collect, organise, analyse, summarise and prioritise data. They also include how to develop hypotheses, identify problems, interpret the data, and determine, plan, implement, and monitor courses of action."*

Vahey et al. [10] propose that:

> *"data literacy includes the ability to formulate and answer questions using data as part of evidence-based thinking; use appropriate data, tools, and representations to support this thinking; interpret information from data; develop and evaluate data-based inferences and explanations; and use data to solve real problems and communicate their solutions."*

It appears from these definitions that data literacy is not intended to describe the specific skills for collecting and analyzing data, but instead a set of core competences which allow people to understand how to use data to answer questions and from which they can make more informed judgements when assessing the outputs of data analysis.

Therefore, our first proposal is to make a *data-inquiry process* more explicit within definitions of data literacy as a way of framing a set of data literacy competencies. We suggest that the PPDAC inquiry cycle is an appropriate starting point for exploring data literacy in this context. PPDAC stands for: **P**roblem, **P**lan, **D**ata, **A**nalysis and **C**onclusion. It is an approach that has

been developed for teaching statistical thinking in New Zealand schools [11], and is typical of a data analysis cycle, but with a focus on application to real world problems. Like other types of inquiry, the stages represent part of an iterative cycle in which the conclusions might prompt further questions and analysis, often of increasing complexity as the problem is being solved. Sometimes, in answering one question, a completely new question or problem is identified which triggers a completely new inquiry process. This same cycle can be applied to many different types of data analysis, including statistical analysis, visual analysis and machine learning methods. Both the specification of the problem and the attributes of the selected data will inform the choice of an appropriate method of analysis.

We propose that within the inquiry framework there is a hierarchy of knowledge and skills related to working with data within a real-world context, ranging from a foundational level of understanding of inquiry, through to very specialist technical skills for hands-on data handling. Figure 1 shows a first step towards mapping the space of data literacy skills. More specialist knowledge of individual methods and tools, such as specific machine learning techniques, should be both learned and applied within this framework. This ensures that analysis is undertaken in the broader context of the problem-setting, data collection and interpretation and in drawing conclusions from the output of the analysis. In addition, the ability to create explanations from data analysis are very important in order to accurately convey the output of the analysis in a way that others can also draw conclusions from it.

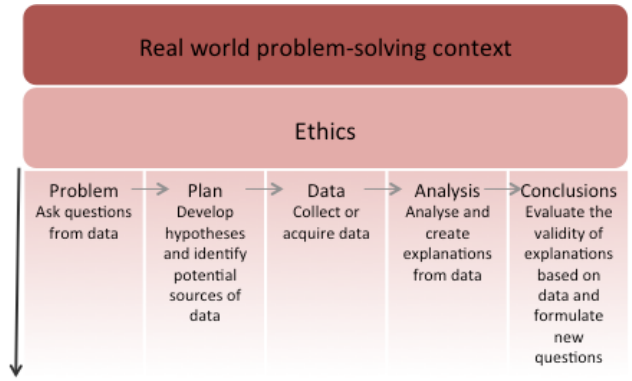

Figure 1. The space of data literacy skills. The arrow within PPDAC activities reflect more specialised data handling skills

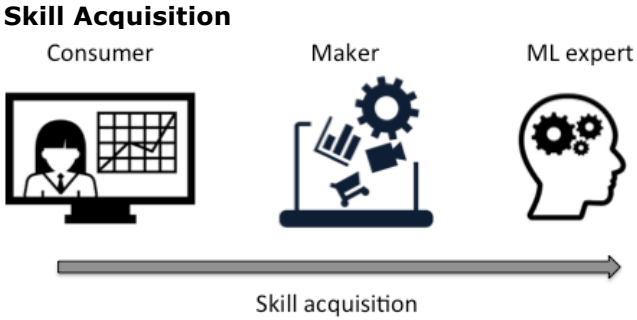

Figure 2. Skill acquisition to transition between roles.

Figure 2 shows how it is possible to transition between the three different user roles that were previously identified, through the acquisition of specialist skills. The foundational skills identified in figure 1 are a prerequisite for each of these roles, whereas the

acquisition of specialist skills is a necessary condition for changing roles.

We propose that raising the level of data literacy amongst the population will have a benefit for machine learning applications, particularly within domains where the outputs are targeted towards the general public. These benefits are summarized below:

- To facilitate non-experts to frame solutions to problems that require the application of machine learning methods to complex data sets, even where they do not themselves have the specific skills required to implement the solution
- To provide enough working knowledge of machine learning and its applications to more critically assess who it is applied to personal data
- To facilitate non experts to begin to acquire specific data skills for solving their own problems
- improve communication between domain experts and machine learning experts during design
- over time, create a more data literate generation of machine learning specialists who are less focused on methods and more considerate towards end-user needs and to solving real-world problems.

Conclusions

In this paper we discuss situations in which non-experts in machine learning more regularly interact with machine learning applications. From this, we identify three distinct user roles. Firstly, there is the passive *consumer* of machine learning outputs, which is commonly targeted towards non-experts. Secondly, there are active *makers* of applications that use applications of machine learning to complex data to solve problems, we have identified situations in which non-experts are more likely to be involved in the design of such applications, namely in citizen-led smart city innovation. Finally, there are the machine learning experts, who have in-depth knowledge of machine learning methods but may have limited knowledge of the application domain. We propose that human-centred machine learning is made possible by ensuring that humans acquire a set of higher level competences and conceptual understanding related to a cycle of data collection and analysis, which we link to the concept of *data literacy.* We make the case that creating a more data literate society will help each type of identified user to better fulfill their roles.

REFERENCES

1. S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*,*35*(4), 105-120.
2. D. Gooch, A. Wolff, G. Kortuem, and R. Brown. 2015. Reimagining the role of citizens in smart city projects. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 1587-1594). ACM.
3. H. Haddadi, H. Howard, A. Chaudhry, J. Crowcroft, A. Madhavapeddy, and R. Mortier. 2015. Personal Data: Thinking Inside the Box. *Critical Alternatives 2015*, Aarhus, Denmark
4. House of Lords Select Committee on Digital Skills. 2015. Make or Break: The UK's Digital Future. Retrieved May 8, 2015 from: http://www.publications.parliament.uk/pa/ld201415/ldselect/lddigital/111/111.pdf

5. Ellen Mandinach and Edith Gummer. 2013. A systemic view of implementing Data Literacy in Educator Preparation. *Educational Researcher*, Vol 42 No. 1 pp 30-37.

6. D. McAuley, H. Rahemtulla, J. Goulding and C. Souch. 2014. How Open Data, data literacy and Linked Data will revolutionise higher education. Retrieved May 8, 2015 from: http://pearsonblueskies.com/2011/how-open-data-data-literacy-and-linked-data-will-revolutionise-higher-education/

7. A. Ojo, E. Curry and F.A. Zeleti. 2015. A Tale of Open Data Innovations in Five Smart Cities. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 2326-2335). IEEE.

8. L. Steen. 1999. Numeracy: The new literacy for a data-drenched society. *Educational Leadership*, 57(2), 8–13.

9. J. G. Tanenbaum, A.M. Williams, A. Desjardins and K. Tanenbaum. 2013. Democratizing technology: pleasure, utility and expressiveness in DIY and maker practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2603-2612). ACM.

10. P. Vahey, L. Yarnall, C. Patton, D. Zalles, and K. Swan. 2006. Mathematizing middle school: Results from a cross-disciplinary study of data literacy. *American Educators Research Association Annual Conference,* 5.

11. C. Wild and M. Pfannkuch. 1999. Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3), 223-265

12. A. Wolff, D. Gooch, U. Mir, J. Cavero and G. Kortuem. 2015. Removing barriers for citizen participation to urban innovation. In: *Digital Cities 9*, Limerick.