Machine Learning for the Developing World

MARIA DE-ARTEAGA, Machine Learning Department, Heinz College, Auton Lab, Carnegie Mellon University
WILLIAM HERLANDS and DANIEL B. NEILL, Machine Learning Department, Heinz College, Event and Pattern Detection Laboratory, Carnegie Mellon University
ARTUR DUBRAWSKI, Auton Lab, Robotics Institute, Carnegie Mellon University

Researchers from across the social and computer sciences are increasingly using machine learning to study and address global development challenges. This article examines the burgeoning field of machine learning for the developing world (ML4D). First, we present a review of prominent literature. Next, we suggest best practices drawn from the literature for ensuring that ML4D projects are relevant to the advancement of development objectives. Finally, we discuss how developing world challenges can motivate the design of novel machine learning methodologies. This article provides insights into systematic differences between ML4D and more traditional machine learning applications. It also discusses how technical complications of ML4D can be treated as novel research questions, how ML4D can motivate new research directions, and where machine learning can be most useful.

CCS Concepts: • **General and reference** → **Surveys and overviews**; *Reference works*; • **Computing methodologies** → **Machine learning**; *Artificial intelligence*; • **Information systems** → *Data mining*; • **Applied computing** → *Health informatics*; *Computers in other domains*; *Computing in government*; *Education*;

Additional Key Words and Phrases: Global development, developing countries

ACM Reference format:
Maria De-Arteaga, William Herlands, Daniel B. Neill, and Artur Dubrawski. 2018. Machine Learning for the Developing World. *ACM Trans. Manage. Inf. Syst.* 9, 2, Article 9 (August 2018), 14 pages.
https://doi.org/10.1145/3210548

1 INTRODUCTION

Six billion people live in developing countries. The unique challenges faced by these regions have long been studied by researchers in fields ranging from sociology to statistics and ecology to economics. As machine learning (ML) methodologies quickly mature, researchers are increasingly

This material is based upon work supported by NSF Graduate Research Fellowship DGE-1252522, NSF awards IIS-0953330 and IIS-1563887, and DARPA awards FA8750-12-2-0324 and 750-14-2-0244.
Authors' addresses: M. De-Arteaga (Corresponding author), Machine Learning Department, Heinz College, Auton Lab, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213; email: mdeartea@andrew.cmu.edu; W. Herlands, Machine Learning Department, Heinz College, Event and Pattern Detection Laboratory, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213; email: wherland@andrew.cmu.edu; D. B. Neill, Machine Learning Department, Heinz College, Event and Pattern Detection Laboratory, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213; email: neill@cs.cmu.edu; A. Dubrawski, Auton Lab, Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213; email: awd@cs.cmu.edu.

https://doi.org/10.1145/3210548

turning to ML to support development research and practice. For example, supervised ML methods can provide expert decision support for health care in regions with few resources, while deep learning techniques can analyze satellite imagery to create novel economic indicators. Yet, there are notable challenges for applying ML in the developing world. Availability of data, computational capacity, and Internet accessibility are often markedly more limited than in developed countries.

The confluence of ML's immense potential with the practical challenges posed by developing world settings has inspired a growing body of research in ML for the developing world (ML4D). In this article, we explore this burgeoning literature. We suggest a formal definition of ML4D in order to scope out the extent of research, and provide a survey of prominent application topics. We suggest best practices drawn from the literature for how to design and implement ML4D projects that advance critical development objectives. Throughout this article, we focus on research that substantially relies on ML. Thus, we will mostly avoid discussion of purely Big Data or information technology topics, which have been well surveyed in previous publications (Ali et al. 2016; Hilbert 2016; Kshetri 2014; World Bank 2014).

The remainder of the article is organized as follows. Section 2 formally defines the area of "machine learning for the developing world". Section 3 provides an overview of the most prominent applications of ML to development research. Section 4 provides a road map for ML4D research, describing how ML can most effectively contribute to addressing the complexity of developing world problems. In Section 5, we describe how developing world challenges can inspire novel ML methodological research. Section 6 follows with the conclusions.

2 DEFINITION

To clarify the scope of this article and to facilitate future discussions, we present a definition of "machine learning for the developing world" (ML4D). Although we acknowledge the limitations of the term "developing world," this article does not attempt to resolve the ongoing discussion on how development should be defined (Kothari 2005; Myrdal 1974; Sen 1988). We consider ML4D to have four key properties:

(1) Applications and data are geographically constrained to developing countries, as defined by the United Nations Development Program (United Nations 2014).
(2) Problems being addressed concern a critical development area for the region of interest.
(3) Problems being addressed or contextual elements necessitate solutions that differ from existing or plausible solutions in developed regions, and the proposed solution effectively addresses these differences.
(4) Proposed solutions substantially use ML as an integral element of the projects.

The first and second criteria are critical since they ensure that we are exclusively considering development issues in the developing world. This is important since many development issues also arise in developed nations. For example, work on violent crime and quality of health care pertain to nearly all countries. Yet, we believe that it is important to concentrate on *developing regions* since these tend to be under-studied in the ML literature and they differ systematically in multiple ways from developed countries. The third criterion emphasizes these differences. In some cases, the problem itself might be unique to developing regions, e.g., eradicating malaria. In other cases, the problem might be found throughout the world, but the context in developing regions makes it such that viable solutions must be fundamentally different in these areas. An example of the latter is preventive care for non-communicable diseases. While such care is needed everywhere, successful solutions in developed nations may not succeed in developing ones. This is a common challenge given that the context and available resources in developing regions tend to be fundamentally different. The unique need for novel solutions in the developing world is precisely our motivation

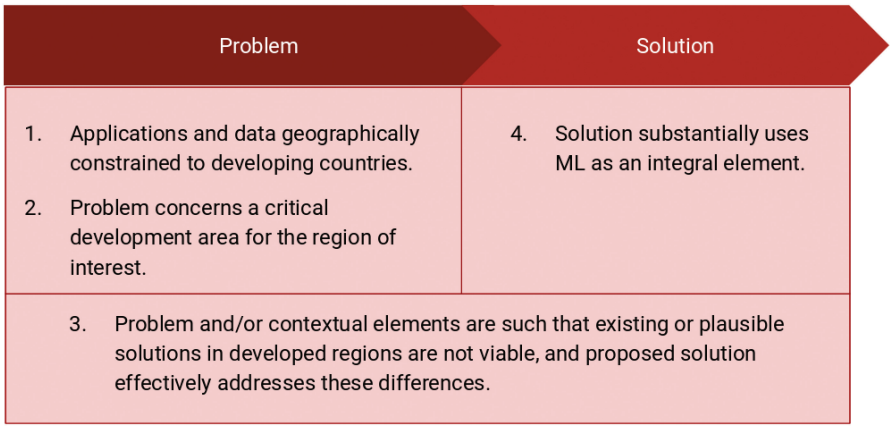

Fig. 1. Four key properties of ML4D. These properties concern both the problems that are being addressed and the solutions that are being proposed.

to study ML4D. The fourth element serves to focus the scope of our methodology. The four key conditions of ML4D can be understood as concerning both the problems that are being addressed and the solutions that are being proposed, as illustrated in Figure 1.

It is important to note that ML4D may or may not involve novel statistical or machine learning methodology. While in some cases development challenges may lead to theoretical developments, the application of existing methods in novel ways may often be the appropriate ML4D approach. Thus, rather than being a subfield of ML, ML4D should be understood as an interdisciplinary domain in the intersection of ML and development research.

Finally, we note that the importance of aligning research projects with development goals highlights the difference between simply applying ML to developing world datasets and ML4D. The latter must have a broader appreciation for the context in which results would be used in order to provide solutions that can effectively contribute toward achieving development objectives.

3 SURVEY OF PROMINENT LITERATURE IN ML4D

In this section, we provide a high level survey of prominent topics in the literature along with a number of illustrative papers. Specifically, we consider five pillars of development: social, economic, health, environmental, and institutional. This classification stems from the four pillars that the UN used for many years (social, economic, environmental, and institutional)[1] and an additional pillar for health given the central role of global health in the development studies. This is not intended to be an exhaustive literature review, nor is this classification the only possible classification of ML4D application domains. Rather, we hope to provide an understanding of the variety, focus, and direction of the most prominent topics in ML4D.

3.1 Health

Health care is one of the most critical topics for ML4D. Automated medical diagnosis is one way of transferring medical expertise across the globe. A number of papers propose automated diagnostic systems in remote areas and urban hospitals for a plethora of diseases from dengue fever to cataracts (Caicedo-Torres et al. 2016; Green and Flaxman 2010; Pathak and Kumar 2016; Robertson and DeHart 2010; Silberman et al. 2010; Waidyanatha et al. 2013).

[1]We acknowledge that the UN now uses a more complex categorization, but we use the old categorization for simplicity.

For sustainable medical treatment, high-quality medical data is necessary as well. Anomaly detection is used in McCarthy et al. (2013) and Chen et al. (2011) to identify data-entry errors, while Brunskill and Lesh (2010) considers optimizing health worker routes. Additionally, Jain et al. (2016) studies how to track infants using novel fingerprint technology to ensure they are getting the proper vaccinations.

In addition to diagnosis at the point of care, Chretien et al. (2008) discusses the importance of early detection of disease outbreaks. Researchers have considered subset scanning (Neill 2012) and other spatial analysis techniques to search for emerging clusters of symptoms or anomalous health patterns. Real-time implementations of these systems have been tested in Sri Lanka and India (Dubrawski et al. 2009; Waidyanatha et al. 2013). The use of alternative data sources to detect emerging epidemics, like social media (Chen and Neill 2014; Chunara et al. 2012) and phone calls (Rehman et al. 2016; Tatem et al. 2009), has also been studied. At a more macro-level, Green (2009) investigates which factors are associated with increased diarrheal disease in multiple countries.

3.2 Institutional

A critical objective in the institutional focus is understanding and modeling violence. Yet, violence in the developing world often occurs in difficult to access rural areas. In order to prevent such atrocities, "crisis mapping" provides spatio-temporal analyses of violence, protests, or environmental disasters (Ziemke 2012). By applying ML analytics to crisis maps over time, a number of researchers have been able to predict violence in urban environments (Benigni and Furrer 2012; Chen and Neill 2014; Parvez et al. 2016). Similarly, De-Arteaga and Dubrawski (2017) detected emerging patterns of sexual violence in El Salvador using spatiotemporal techniques, and Chen and Neill (2015) discovered patterns of human rights violations using social media data.

Institutional corruption is also a pernicious problem in the developing world (Gray and Kaufmann 1998). Huysmans et al. (2006) provides insight on the relationships between country-level macro-economical factors and corruption. In an attempt to halt such activity, Grace et al. (2016) developed an automated risk classification system for fraud and collusion in international development contracts, and Cantú and Saiegh (2011) uses ML to detect electoral fraud in Argentina's democratic process.

3.3 Economic

Economic outlook is central to development research. Indeed, the primary means of classifying a country's development status is per capita gross national income (GNI) (United Nations 2014). Considering macroeconomic features, Hidalgo et al. (2007) uses network analysis to understand how a country's products and exports impact their capability for future growth. Other studies focus on microeconomics, exploring factors that lead to defaulting on cell phone debts (Yigzaw et al. 2010) or experiencing famine (Mwebaze et al. 2010).

High-quality economic data is rare in the developing world, where government censuses are often incomplete. Instead, researchers and companies have turned to crowd-sourcing for data collection (Blumenstock et al. 2016) or prediction methods that use a variety of supplementary indicators (McBride and Nichols 2015). An alternative approach, proposed by Jean et al. (2016), uses novel transfer learning methods to use satellite images to predict poverty. Similarly, Saavedra and Romero (2016) uses satellite data to estimate the prevalence of illegal mining activity in Colombia.

3.4 Social

Understanding population dynamics is critical for allocating resources. As previously mentioned, researchers have turned to cell phone data for analyzing population mobility and the features of the underlying social networks. Hill et al. (2010) uses network analysis to compare cell phone usage

behaviors between students in the United States and Kenya. Similarly, Rubio et al. (2010) presents a comparative analysis of human mobility and sparsity of social networks between developed and developing countries. Also focused on human mobility, Wesolowski and Eagle (2010) uses mobile phone data to test prominent anthropological and economic theories regarding migration patterns in slums. Indeed, cell phone records have become so fundamental to social and population analysis that a number of studies have focused on understanding the nature of cell phone use (Blumenstock et al. 2010; Frias-Martinez et al. 2010).

Education is another facet of social life that has been the subject of extensive ML work. In order to ensure high-quality education, intelligent tutoring systems promise to provide low-cost, individualized education curricula for students (Nye 2015). Brunskill et al. (2010) provide an example of such a system that uses active learning to improve student learning when several students need to share computers. While such tools hope to directly improve the quality of education, other studies use retrospective analyses to better understand indicators of a student's success or failure (Mgala and Mbogho 2015; Moussavi and McGinn 2010).

In addition to formal education, access to information is increasingly important and is often facilitated by expanding Internet connectivity in the developing world (ITU UNESCO 2016). Yet, Barnard et al. (2010) discuss the challenges of information access in small communities of South Africa who speak niche languages. They suggest a variety of bootstrapped natural language technologies (such as text-to-speech and automated speech recognition) to enable such populations to access resources on the Internet and elsewhere. Similarly, Farrell et al. (2010) explore collaborative filtering to enable navigation in the *Spoken Web*, a tool that enables individuals with low levels of literacy to use the Internet (Parikh 2010).

3.5 Environment

Much of the current environmental studies literature using ML focuses on understanding and mitigating natural disasters. While all regions of the globe face pernicious weather and dangerous tectonic activity, the developing world is particularly impacted by these events due to inadequate building construction (Alcántara-Ayala 2002) and dearth of government resources available for disaster response (Ferris 2010). A number of researchers have studied the susceptibility for flooding or landslides in Southeast Asia (Tehrany et al. 2014; Tien Bui et al. 2012). These studies were motivated by disasters over the last decade and relied on data from previous floods and landslides to identify areas of high risk. In order to create faster response mechanisms, Kapoor et al. (2010) use cell tower activity data to detect and predict seismic events in the Democratic Republic of Congo. Other papers have also explored the use of satellite data to automatically discover land use and land cover (Jia et al. 2017) and agricultural land availability (Chakraborty et al. 2016).

Instead of focusing on large environmental disasters, Ermon et al. (2015) use Markov decision processes to model migratory herder trajectories in space and time, looking specifically at how they respond to environmental shocks. Such work may benefit from other environmentally related projects, such as those proposed by creators of the "causality workbench" (Guyon et al. 2010). Using causal modeling methods can further help policy makers plan and mitigate environmental shocks to indigenous peoples and urban populations in the developing world.

4 ADVANCING GLOBAL DEVELOPMENT WITH MACHINE LEARNING

Having reviewed prominent ML4D literature in Section 3, we now consider how best practices from that literature can provide a road map for using ML to advance global development goals. First, in Section 4.1, we argue that it is essential for researchers to consider the local context of their project in order to ensure proper alignment between development needs and the project's

technological and institutional objectives. Then, in Section 4.2, we provide a three-step road map to better conceptualize how ML4D can be used to advance global development goals.

4.1 Ensure that ML4D Research Considers Local Context

Within any ML4D project, it is important that proposed solutions should be aligned with the available means and local expertise. For example, if the goal is to inform policy, the technical methodology should provide results that policy makers can feasibly act upon. Consider the work in Mgala and Mbogho (2015), which is motivated by teacher shortages and attempts to inform data-driven interventions by predicting which students are likely to fail later in their academic career. Their outcome indicates that 70% of the students require intervention. With a shortage of teachers, it seems unlikely that the schools will be able to react effectively to this information. If the researchers had taken into account this constraint when deciding what the output of their algorithm would be, the potential impact of their research might have been greater.

Potential infrastructure constraints, such as computational capacity and the need to travel long distances for repairs, should also be taken into account. For example, Pathak and Kumar (2016) use a low-cost device to diagnose cataracts that can be easily acquired and maintained by a hospital or clinic in a low-resource region. Countries and regions may differ substantially in quantities such as the proportion of individuals with access to a mobile phone, and these differences may make a proposed solution feasible in some regions and infeasible in others.

That said, there is a natural alignment between ML4D techniques and development goals. ML allows computers to perform tasks that could previously only be achieved by highly skilled experts. Thus, areas of the world with a deficit of these specialists are ideal settings to deploy such new technologies. In the developed world, many people fear the possibility of artificial intelligence taking away their jobs (YouGov Omnibus 2016). But in many areas of the developing world, machines provide the promise of performing jobs that are fundamental for society and difficult to fill currently.

Examples of this are not hard to find. In regions with few or no specialized doctors, the need for ML is palpable. When people live hours away from the nearest hospital, they often will only visit for emergencies, when their health has already deteriorated and diseases are in later stages. It is not realistic to expect people to travel long distances when experiencing mild symptoms. However, inexpensive ML-enabled devices can provide high specificity early diagnoses of potentially debilitating diseases. Such devices can inform patients when to visit the hospital before their medical condition deteriorates. Similarly, teacher shortages are common in rural areas of the developing world (UNESCO 2015). If machines could augment and support human teaching responsibilities, this could help increase literacy and sharpen STEM skills, paving a road to improve development.

4.2 Road Map for ML4D

Given the strong alignment between development needs and ML approaches, we provide a road map to strengthen the application of ML in this domain. This road map details three technical stages where ML4D can play an essential role and meaningfully contribute to global development. We believe this lays out a clear vision for how the field of ML4D can flourish in the coming years. Although we present these stages as successive, important work can be, and has already been, done in any of the stages; we provide examples of such work below. Finally, while ML4D does not necessitate novel ML methodological research, we extend this road map in Section 5 by linking each stage with specific areas of novel ML research that could be pursued within the context of ML4D (see Figure 2).

(1) **Improve Data Reliability.** While many data resources exist in developing regions, much of the data is incomplete, noisy, biased, or otherwise "messy." For example, reliable

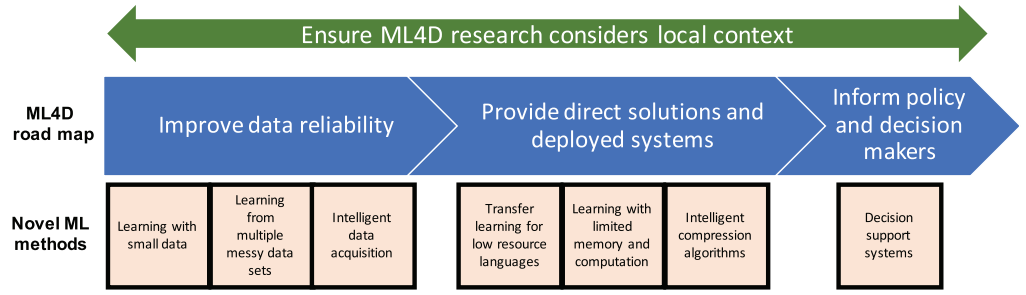

Fig. 2. Top: road map for how ML4D can be used to foster development solutions (see Section 4.2). Bottom: corresponding ML areas where ML4D can foster novel research (see Section 5).

economic indicators are fundamental to evaluating a country's existing policies and understanding potential macro-level effects of proposed initiatives. Motivated by this, Jean et al. (2016) uses satellite imagery to estimate consumption expenditure and asset wealth. With a similar goal in mind, Blumenstock et al. (2016) proposes aggregating data collected via mobile phones by a network of local contributors into reliable economic indicators. In Saavedra and Romero (2016), ML is used to create a map of illegal mining in Colombia, allowing the authors to estimate the effect of a policy change. Lum et al. (2010) uses a Bayesian model to combine 15 different sources of data, all known to be biased and incomplete, to estimate the number of homicides and disappearances in Casanare, Colombia, between 1998 and 2007. Similarly, Silva et al. (2010) use a referral-based sampling technique to estimate systematic enforced disappearances in Punjab, India, between 1984 and 1996.

(2) **Provide Direct Solutions and Deployed Systems.** ML can be built into systems that directly impact development issues. For example, Brunskill et al. (2010) applies active learning to maximize engagement in multi-user educational tools, motivated by low-resource schools where it is common for several students to share a single monitor. In the health care domain, Baba et al. (2015) proposes a predictive model to enable low-cost preventive medicine in countries where standard preventive medicine is cost prohibitive. Jain et al. (2016) give infants an identity using their fingerprints, even though it was previously believed that fingerprints could only be used for children at least 2.5 years old. Improved infant identification helps prevent switching babies in hospitals, makes it easier to identify missing children, and improves accuracy of vaccination records.

(3) **Inform Policy and Decision Makers.** Whether it is through knowledge-discovery models that improve our understanding of a phenomenon, or through predictive models that inform proactive policies, ML can be integrated as an essential component of decision-support systems. For example, when Zanzibar was analyzing the viability of a malaria elimination campaign, Tatem et al. (2009) used anonymized mobile phone records to characterize travel patterns between Zanzibar and mainland Tanzania, where malaria was still endemic. Also, in the public health domain, Dubrawski et al. (2009) proposes a spatiotemporal model for early detection of disease outbreaks in real time. A similar approach is used in De-Arteaga and Dubrawski (2017) to discover emerging patterns of sexual violence in El Salvador, with the goal of helping policy makers gain a better understanding of this phenomenon and create policies to counter it.

5 OPPORTUNITIES FOR ADVANCING MACHINE LEARNING THROUGH ML4D

When considering ML4D, the limitations and difficulties of successfully deploying technologies in the developing world are often highlighted (Ali et al. 2016; Hilbert 2016), as is the immense potential of computer science and ML to positively impact these areas of the world (Dias and Brewer 2009; Quinn et al. 2014). In this section, we discuss a topic that receives less attention: the opportunities that the developing world presents to ML. We propose that infrastructure limitations which are often lamented as deterrents for ML tools in these regions should instead be considered cutting-edge ML research questions. Indeed, we suggest that many of the challenges in ML4D are the consequence of research being done for developed countries and only later applied to developing settings.

A common skeptical argument against ML4D is that conditions in the developing world make it almost impossible for ML to thrive there, while more technologically advanced developed countries constitute a better fit for deployment (Weber and Toyama 2010). That is true to an extent. Yet, it is only true because the technology has been built assuming the conditions of developed countries, and thus current ML tools encode the infrastructure and cultural conditions of developed regions. It should come as no surprise that the same tools may not work as effectively when such criteria are no longer met.

It is critical to not only consider environmental constraints of the developing world in the application phase of a project. Instead, we must treat them as integral aspects of the research. This is quite standard for how ML and computer science research is often done. For example, stochastic optimization methods, convex relations in non-convex objectives, and various matrix decompositions are all the result of power, memory, and computational constraints. Similarly, numerous statistical techniques have been developed for common issues with missing, noisy, and biased data.

The problem is not the conditions of the developing world, but a belief in what we term "trickle-down machine learning." Developing world conditions only become roadblocks to research when we have assumed that research for the developed world should necessarily apply to the developing one. The fact that people in rural areas of developing countries do not usually own smartphones is only a limitation when we have performed research assuming the widespread use of smartphones. As we argue in Section 4.1, ML has immense potential to bridge the development gap, specifically in low-resource settings lacking highly skilled human specialists. Since such areas have quite different conditions than developed countries, it is important to modify our research approaches accordingly.

Several examples show us what is possible when environmental conditions in developing countries are considered research questions rather than roadblocks. Poor data quality is often seen as a problem with ML4D, yet Lum et al. (2010) use such a setting to propose a novel method. They use a custom Bayesian model to combine information available in 15 datasets, all known to be incomplete and biased, and estimate the number of murders and disappearances in Casanare, Colombia, between 1998 and 2007. Similarly, Brunskill et al. (2010) are motivated by recognizing that students often share a computer in classrooms in developing regions. Within this setting, they develop an active learning system to maximize engagement in multi-user educational tools.

Yet there is much room for improvement. Overcoming common limitations in developing countries has the potential to inspire cutting-edge ML research. Below, we present seven developing world technical challenges along with suggestions of novel ML lines of inquiry that they could motivate. While this list is far from exhaustive, it provides a wide breadth of technical challenges that can inspire novel research in numerous ML disciplines. Additionally, in Figure 2, we link each of the seven topics to the ML4D road map proposed in Section 4.2.

(1) **Learning with Small Data.** Lack of digitized historical records, as well as limited capacity to record data (such as in the health care domain) is a common challenge in developing countries. Thus, ML4D problems often have only relatively small datasets, e.g., hundreds or thousands of data records rather than millions or billions. In some cases the number of data records is small compared to the number of attributes, resulting in high potential for overfitting if complex models are used. While handling massive amounts of data in the order of petabytes is an important contemporary research problem in ML, developing optimization techniques for small data is a similarly vexing problem. Some research has been pursued on this topic (Forman and Cohen 2004), but ML4D can inspire new models and general optimization routines that specifically concentrate on small data settings.

(2) **Learning from Multiple Messy Datasets.** Due to a dearth of official datasets in the developing world, it is often the case that citizens, journalists, and non-profit organizations lead initiatives to collect data. In many cases, those collecting the data have good intentions, but limited funds. Additionally, adverse environmental conditions and lack of technical expertise prevents them from constructing the types of "clean" datasets ML researchers have come to expect. Instead, we are left with multiple sources of biased, incomplete data, as in the case of, e.g., Lum et al. (2010). Creating methods that can seamlessly integrate multiple diverse datasets and properly account for potential biases would enable researchers to leverage vast troves of often neglected information sources, while simultaneously advancing the study of fields such as bootstrapping, data imputation, and machine fairness.

(3) **Intelligent Data Acquisition.** Given that limited resources and infrastructure are often a constraint in developing world settings, prioritizing which new data to collect may be a critical question. Hence, active learning and active feature-value acquisition models that can be deployed successfully in settings with limited infrastructure are key. Understanding the context in which such models are required can give rise to novel algorithms in this realm.

(4) **Transfer Learning for Low-Resource Languages.** As discussed by Barnard et al. (2010) and Farrell et al. (2010), enabling Natural Language Processing (NLP) for low-resource languages is critical for those developing countries that host dozens of regional dialects and tribal languages. Considering how to use knowledge distillation in this field could be of particular importance since data on these languages is extremely sparse. Such advances could spur novel transfer learning models as well as enable peoples from across the globe to access information and communicate with each other.

(5) **Learning with Limited Memory and Computation.** In order to enable active learning and other in situ methods in the developing world, it is important to enable ML routines to function with potentially limited memory and computational capacity. This relates to research on ML in embedded systems as well as to recent advances in knowledge distillation (Hinton et al. 2015). Developing general methods for enabling ML in these technologically limited settings could produce research applicable to both important developed and developing world settings.

(6) **Intelligent Compression Algorithms.** Telecommunication networks may have limited bandwidth in the developing world, but are critical for advanced technologies, such as telemedicine, as well as general access to global information on the Internet. While traditional compression algorithms for video, text, and voice are well studied, a pioneering body of work in ML considers using learning algorithms to further compress data. For example, Isola et al. (2016) use adversarial deep learning networks for intelligent image-to-image translation. While this specific solution requires significant computational

capacity, exploring this field more deeply may yield intelligent compression algorithms that can perform well in developing world settings with limited computational power. This would represent a huge advancement toward bringing telecommunications to such regions while also substantially advancing deep learning research.

(7) **Decision Support Systems.** As health care and government services expand in developing countries, ML decision support systems can be created to help scale these services. This use of ML could be particularly relevant where medical and administrative expertise is rare. Such decision support systems could also address needs that are particularly pressing in developing countries, such as corruption (Olken and Pande 2012). Grace et al. (2016) and Cantú and Saiegh (2011) have shown that ML can detect subtle patterns of fraud in complex data. New methods may embed such anomaly detection within decision support technologies to ensure robust and flexible ML systems, potentially decreasing the governmental inefficiency that may result from widespread corruption (Anechiarico and Jacobs 1996). An emerging methodological field focused on fairness, accountability, and transparency in machine learning (Hardt et al. 2016; Kearns 2017; Zemel et al. 2013) is well situated to aid these systems. Ideally, systems can be designed from the ground up, employing these tools to support fair and equitable decision making. As an alternative, approaches to auditing black-box algorithms for fairness, such as Zhang and Neill (2016), might be used to discover and correct systematic biases in existing decision processes. Additionally, decision support systems capable of discovering and leveraging causal relationships could guide interventions and policy decisions. Such needs could motivate novel research in the growing subfield of ML and causality (Athey and Imbens 2015; Schölkopf et al. 2012).

6 CONCLUSIONS

ML4D is a broad and growing area of academic research with tremendous potential for real-world impact. In this article, we explored prominent topics in the literature and analyzed how best practices in both ML and development studies should inform ML4D projects. Additionally, we argued that computational or data limitations in the developing world, which are often lamented as deterrents for ML tools in these regions, should instead be considered cutting-edge ML research questions. Researchers should be encouraged to explore how such challenges in the developing world can inspire new ML paradigms. By considering how the diverse fields of ML and development studies can reinforce each other, ML4D researchers have the opportunity to create cutting-edge ML methods while addressing critical issues in the developing world.

REFERENCES

Irasema Alcántara-Ayala. 2002. Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. *Geomorphology* 47, 2 (2002), 107–124.

Anwaar Ali, Junaid Qadir, Raihan ur Rasool, Arjuna Sathiaseelan, Andrej Zwitter, and Jon Crowcroft. 2016. Big data for development: Applications and techniques. *Big Data Analytics* 1, 1 (2016), 2.

Frank Anechiarico and James B. Jacobs. 1996. *The Pursuit of Absolute Integrity: How Corruption Control Makes Government Ineffective*. University of Chicago Press.

Susan Athey and Guido W. Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *Stat* 1050, 5 (2015).

Yukino Baba, Hisashi Kashima, Yasunobu Nohara, Eiko Kai, Partha Ghosh, Rafiqul Islam, Ashir Ahmed, Masahiro Kuroda, Sozo Inoue, Tatsuo Hiramatsu, and others. 2015. Predictive approaches for low-cost preventive medicine program in developing countries. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1681–1690.

Etienne Barnard, Marelie H. Davel, and Gerhard B. Van Huyssteen. 2010. Speech technology for information access: A South African case study. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. 13–15.

Matthew Benigni and Reinhard Furrer. 2012. Spatio-temporal improvised explosive device monitoring: Improving detection to minimise attacks. *Journal of Applied Statistics* 39, 11 (2012), 2493–2508.

Joshua E. Blumenstock, Dan Gillick, and Nathan Eagle. 2010. Who's calling? Demographics of mobile phone use in Rwanda. *Transportation* 32 (2010), 2–5.

Joshua E. Blumenstock, Niall C. Keleher, and Joseph Reisinger. 2016. The premise of local information: Building reliable economic indicators from a decentralized network of contributors. In *Proceedings of the 8th International Conference on Information and Communication Technologies and Development*. ACM, 61.

Emma Brunskill, Sunil Garg, Clint Tseng, Joyojeet Pal, and Leah Findlater. 2010. Evaluating an adaptive multi-user educational tool for low-resource environments. In *Proceedings of the IEEE/ACM International Conference on Information and Communication Technologies and Development*. 13–16.

Emma Brunskill and Neal Lesh. 2010. Routing for rural health: Optimizing community health worker visit schedules. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

William Caicedo-Torres, Ángel Paternina, and Hernando Pinzón. 2016. Machine learning models for early dengue severity prediction. In *Ibero-American Conference on Artificial Intelligence*. Springer, 247–258.

Francisco Cantú and Sebastián M. Saiegh. 2011. Fraudulent democracy? An analysis of Argentina's "infamous decade" using supervised machine learning. *Political Analysis* (2011), 409–433.

Sunandan Chakraborty, Zohaib Jabbar, Lakshminarayanan Subramanian, and Yaw Nyarko. 2016. Satellite image analytics, land change and food security. In *ACM SIGKDD Workshop on Data Science for Food, Energy and Water, Co-Located with KDD*. San Francisco.

Feng Chen and Daniel B. Neill. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1166–1175.

Feng Chen and Daniel B. Neill. 2015. Human rights event detection from heterogeneous social media graphs. *Big Data* 3, 1 (2015), 34–40.

Lujie Chen, Artur Dubrawski, Nuwan Waidyanatha, and Chamindu Weerasinghe. 2011. Automated detection of data entry errors in a real time surveillance system. *Emerging Health Threats Journal* 4, s69 (2011), 9–10.

Jean-Paul Chretien, Howard S. Burkom, Endang R. Sedyaningsih, Ria P. Larasati, Andres G. Lescano, Carmen C. Mundaca, David L. Blazes, Cesar V. Munayco, Jacqueline S. Coberly, Raj J. Ashar, and others. 2008. Syndromic surveillance: Adapting innovations to developing settings. *PLoS Medicine* 5, 3 (2008), e72.

Rumi Chunara, Jason R. Andrews, and John S. Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene* 86, 1 (2012), 39–45.

Maria De-Arteaga and Artur Dubrawski. 2017. Discovery of complex anomalous patterns of sexual violence in El Salvador. In *Data for Policy*. Zenodo. DOI: http://dx.doi.org/10.5281/zenodo.571551

M. Bernardine Dias and Eric Brewer. 2009. How computer science serves the developing world. *Communications of the ACM* 52, 6 (2009), 74–80.

Artur Dubrawski, Maheshkumar Sabhnani, Michael Knight, Michael Baysek, Daniel Neill, Saswati Ray, Anna Michalska, and Nuwan Waidyanatha. 2009. T-Cube web interface in support of real-time bio-surveillance program. In *Proceedings of the 2009 International Conference on Information and Communication Technologies and Development (ICTD'09)*. IEEE, 495–495.

Stefano Ermon, Yexiang Xue, Russell Toth, Bistra Dilkina, Richard Bernstein, Theodoros Damoulas, Andrew G. Mude, Patrick Clark, Steve DeGloria, Christopher Barrett, and others. 2015. Learning large-scale dynamic discrete choice models of spatio-temporal preferences with application to migratory pastoralism in East Africa. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*. ACM, 644–650.

Robert G. Farrell, Rajarshi Das, and Nitendra Rajput. 2010. Social navigation through the spoken web: Improving audio access through collaborative filtering in Gujarat, India. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. Citeseer.

Elizabeth Ferris. 2010. Natural disasters, conflict, and human rights: Tracing the connections. *The Brookings Institution. Presented at Brookings-Bern Project on Internal Displacement. Texas, March 3, 2010.*

George Forman and Ira Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 161–172.

Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. 2010. A gender-centric analysis of calling behavior in a developing economy using call detail records. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Emily Grace, Ankit Rai, Elissa Redmiles, and Rayid Ghani. 2016. Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data'16)*. IEEE, 1444–1453.

Cheryl W. Gray and Daniel Kaufmann. 1998. Corruption and development. *Finance and Development* 35, 1 (1998), 7.

Sean T. Green. 2009. *Machine Learning Methods for Decision Support in Health Policy for Developing Countries*. Ph.D. dissertation. Carnegie Mellon University.

Sean T. Green and Abraham D. Flaxman. 2010. Machine learning methods for verbal autopsy in developing countries. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Isabelle Guyon, Jean-Philippe Pellet, and Alexander R. Statnikov. 2010. Development projects for the causality workbench. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. Citeseer.

Moritz Hardt, Eric Price, Nati Srebro, and others. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

César A. Hidalgo, Bailey Klinger, A.-L. Barabási, and Ricardo Hausmann. 2007. The product space conditions the development of nations. *Science* 317, 5837 (2007), 482–487.

Martin Hilbert. 2016. Big data for development: A review of promises and challenges. *Development Policy Review* 34, 1 (2016), 135–174.

Shawndra Hill, Anita Banser, Getachew Berhan, and Nathan Eagle. 2010. Reality mining Africa. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. Citeseer.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531*.

Johan Huysmans, David Martens, Bart Baesens, Jan Vanthienen, and Tony Van Gestel. 2006. Country corruption analysis with self organizing maps and support vector machines. In *Intelligence and Security Informatics*. Springer, 103–114.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004*.

ITU UNESCO. 2016. *The State of Broadband: Broadband Catalyzing Sustainable Development*. Technical Report, September.

Anil K. Jain, Sunpreet S. Arora, Lacey Best-Rowden, Kai Cao, Prem S. Sudhish, Anjoo Bhatnagar, and Yoshinori Koda. 2016. Giving infants an identity: Fingerprint sensing and recognition. In *Proceedings of the 8th International Conference on Information and Communication Technologies and Development*. ACM, 29.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Predict land covers with transition modeling and incremental learning. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 171–179.

Ashish Kapoor, Nathan Eagle, and Eric Horvitz. 2010. People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Michael Kearns. 2017. Fair algorithms for machine learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 1–1.

Uma Kothari. 2005. From colonial administration to development studies: A post-colonial critique of the history of development studies. *A Radical History of Development Studies: Individuals, Institutions and Ideologies* (2005), 47–66.

Nir Kshetri. 2014. The emerging role of big data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society* 1, 2 (2014), 2053951714564227.

Kristian Lum, Megan Price, Tamy Guberek, and Patrick Ball. 2010. Measuring elusive populations with Bayesian model averaging for multiple systems estimation: A case study on lethal violations in Casanare, 1998–2007. *Statistics, Politics, and Policy* (2010).

Linden McBride and Austin Nichols. 2015. Improved poverty targeting through machine learning: An application to the USAID poverty assessment tools. Retrieved April 2015 from econthatmatters.com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf.

Ted McCarthy, Brian DeRenzi, Joshua Blumenstock, and Emma Brunskill. 2013. Towards operationalizing outlier detection in community health programs. In *Proceedings of the 6th International Conference on Information and Communications Technologies and Development: Notes-Volume 2*. ACM, 88–91.

Mvurya Mgala and Audrey Mbogho. 2015. Data-driven intervention-level prediction modeling for academic performance. In *Proceedings of the 7th International Conference on Information and Communication Technologies and Development*. ACM, 2.

Massoud Moussavi and Noel McGinn. 2010. A model for quality of schooling. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Ernest Mwebaze, Washington Okori, and John Alexander Quinn. 2010. Causal structure learning for famine prediction. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Gunnar Myrdal. 1974. What is development? *Journal of Economic Issues* 8, 4 (1974), 729–736.

Daniel B. Neill. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 2 (2012), 337–360.

Benjamin D. Nye. 2015. Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education* 25, 2 (2015), 177–203.

Benjamin A. Olken and Rohini Pande. 2012. Corruption in developing countries. *Annual Review of Economics*. 4, 1 (2012), 479–509.

Tapan S. Parikh. 2010. Voice as data: Learning from what people say. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Md Rizwan Parvez, Turash Mosharraf, and Mohammed Eunus Ali. 2016. A novel approach to identify spatio-temporal crime pattern in Dhaka City. In *Proceedings of the 8th International Conference on Information and Communication Technologies and Development*. ACM, 41.

Shashwat Pathak and Basant Kumar. 2016. A robust automated cataract detection algorithm using diagnostic opinion based parameter thresholding for telemedicine application. *Electronics* 5, 3 (2016), 57.

John Quinn, Vanessa Frias-Martinez, and Lakshminarayan Subramanian. 2014. Computational sustainability and artificial intelligence in the developing world. *AI Magazine* 35, 3 (2014).

Nabeel Abdur Rehman, Shankar Kalyanaraman, Talal Ahmad, Fahad Pervaiz, Umar Saif, and Lakshminarayanan Subramanian. 2016. Fine-grained dengue forecasting using telephone triage services. *Science Advances* 2, 7 (2016), e1501215.

Joel Robertson and Del DeHart. 2010. An agile and accessible adaptation of Bayesian inference to medical diagnostics for rural health extension workers. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. Citeseer.

Alberto Rubio, Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. 2010. Human mobility in advanced and developing economies: A comparative analysis. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Santiago Saavedra and Mauricio Romero. 2016. Local incentives and national tax evasion: The response of illegal mining to a tax reform in Colombia.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv:1206.6471*.

Amartya Sen. 1988. The concept of development. *Handbook of Development Economics* 1 (1988), 9–26.

Nathan Silberman, Kristy Ahrlich, Rob Fergus, and Lakshminarayanan Subramanian. 2010. Case for automated detection of diabetic retinopathy. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Romesh Silva, Jeff Klingner, and Scott Weikart. 2010. Measuring lethal counterinsurgency violence in Amritsar District, India using a referral-based sampling technique. In *Joint Statistical Meetings*. 552–580.

Andrew J. Tatem, Youliang Qiu, David L. Smith, Oliver Sabot, Abdullah S. Ali, and Bruno Moonen. 2009. The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among Zanzibar residents. *Malaria Journal* 8, 1 (2009), 287.

Mahyat Shafapour Tehrany, Biswajeet Pradhan, and Mustafa Neamah Jebur. 2014. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of Hydrology* 512 (2014), 332–343.

Dieu Tien Bui, Biswajeet Pradhan, Owe Lofman, and Inge Revhaug. 2012. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and naive Bayes models. *Mathematical Problems in Engineering* (2012), 26 pages.

UNESCO. 2015. *The Challenge of Teacher Shortage and Quality: Have We Succeeded in Getting Enough Quality Teachers into Classrooms?* Technical Report. UNESCO. http://unesdoc.unesco.org/images/0023/002327/232721E.pdf.

United Nations. 2014. World Economic Situation and Prospects 2014. Retrieved July 15, 2017 from http://www.un.org/en/development/desa/policy/wesp/wesp_current/wesp2014.pdf.

N. Waidyanatha, C. Sampath, Artur Dubrawski, S. Prashant, M. Ganesan, and Gordon Gow. 2013. Affordable system for rapid detection and mitigation of emerging diseases. *Digital Advances in Medicine, E-Health, and Communication Technologies* (2013), 271.

Julie Sage Weber and Kentaro Toyama. 2010. Remembering the past for meaningful AI-D. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

Amy Wesolowski and Nathan Eagle. 2010. Parameterizing the dynamics of slums. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*. Citeseer.

World Bank. 2014. *Big Data in Action for Development*. World Bank Other Operational Studies 21325. The World Bank. http://EconPapers.repec.org/RePEc:wbk:wboper:21325.

Mariye Yigzaw, Shawndra Hill, Anita Banser, and Lemma F. Lessa. 2010. Using data mining to combat infrastructure inefficiencies: The case of predicting nonpayment for Ethiopian telecom. In *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*.

YouGov Omnibus. 2016. Poll of robots.

Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning fair representations. *ICML (3)* 28 (2013), 325–333.

Zhe Zhang and Daniel B. Neill. 2016. Identifying significant predictive bias in classifiers. *Presented at NIPS Interpretable Machine Learning in Complex Systems Workshop* (2016).

Jen Ziemke. 2012. Crisis mapping: The construction of a new interdisciplinary field? *Journal of Map & Geography Libraries* 8, 2 (2012), 101–117.

Received November 2017; revised March 2018; accepted April 2018