Preface: Data Science for Social Good

A new movement has slowly been building momentum within the community of data scientists. Increasingly, individuals and groups with data science expertise are working with social change organizations on the toughest challenges facing humanity. This issue of the *IBM Journal of Research and Development* presents 12 research papers that capture the zeitgeist of the data-for-good movement.

Social responsibility is as old a concept as human civilization, and technological development has always been with an eye toward societal benefit. With industrialization, however, there came a division between for-profit companies leading the way in developing technology and nonprofits leading the way in serving society. The for-profits created corporate social responsibility arms that donated money, products, and volunteer labor to nonprofits but did not directly engage with them to jointly understand social problems and jointly work on the science to create solutions. The data-for-good movement is attempting to change the equation by convening collaborations among leading technologists and leading social-impact workers, with a focus on the ubiquitous resource of today: digital information.

Certain industries and sectors have taken advantage of the big data era more quickly and reaped the benefits through descriptive, predictive, and prescriptive analytics. However, despite dealing with the problems most important for humanity, being the leaders in open data, and having a fact-based heritage, organizations and agencies in the social and public sectors are not yet utilizing advanced data science techniques as much as they could. The papers herein describe some initial steps to change the status quo.

What subject matter areas constitute social good? In September 2015, the member states of the United Nations adopted an ambitious set of 17 global goals to be achieved by the year 2030 that provide a comprehensive framework for answering the question. This sustainable development agenda includes (using the shorthand notation at the un.org website): no poverty; zero hunger; good health and well-being; quality education; gender equality; clean water and sanitation; affordable and clean energy; decent work and economic growth; industry, innovation, and infrastructure; reduced inequalities, sustainable cities and communities, responsible consumption and production, climate action, life below water, life on land; peace, justice, and strong institutions; and partnerships for the goals.

What technical problems arise when using data science to investigate these areas? Monitoring and evaluating initiatives via surveys and basic statistical tests has a long tradition in social policy and practice. Supervised machine learning methodologies may be used for prediction and early warning. Causal inference and operations research have a role in policy design, prioritization, and resource allocation. Artificial intelligence approaches reduce the burden on people for performing low-level information processing tasks so that they may be more efficient. Tools from the discipline of computer-supported collaborative work, enhanced with cognitive capabilities, improve knowledge exchange, problem identification, and the development of actionable solutions when working on difficult social good issues that require multi-party inputs. Finally, despite the ubiquity of digital data, there remain phenomena that are difficult to observe or are unobservable, requiring the construction of data proxies via descriptive and predictive analytics.

The first paper, by Tao et al., is an application of statistical testing in the monitoring and evaluation of a diarrhea treatment program in Nigeria. In particular, the authors note that diarrhea causes over 500,000 deaths of children younger than 5 years of age, globally, each year. A combination of oral rehydration salts (ORS) and zinc is regarded as the most effective treatment to prevent diarrhea-related deaths among children. The authors study the effectiveness of a Clinton Health Access Initiative (CHAI) program in Nigeria, aimed at increasing usage of ORS and zinc. Peer detailers (educators) reinforce the benefits of the treatment to patent-and-proprietary medicine vendors (PPMVs), a main source of healthcare in the country. Two aspects of program effectiveness are analyzed: 1) PPMV knowledge that this combination is the most effective treatment and 2) PPMV inventory of ORS and zinc. Hypothesis tests reveal that the percentage of PPMVs with knowledge of the most effective treatment increases significantly in most of the tested Nigerian states after peer detailing. Surprisingly, no significant patterns were detected regarding the percentage of PPMVs with inventory of ORS and zinc. Logistic regression results suggest that PPMVs that were detailed have significantly higher odds ratios for both knowledge and inventory. The authors conclude with a discussion of their findings regarding the overall effectiveness of the CHAI program and the limitations of the study, as well as suggestions for future activities for CHAI to support their information dissemination program.

The next paper, by Chen et al., concerns evaluating the impact of air pollution on different diseases in Shenzhen, China. Ambient air pollution has been a worldwide concern with a devastating impact on the health of populations, while increasing the burden on public health systems. Assessing the adverse health effects of air pollution is vital for forming disease control policies. This study investigates the excess risk of six air pollutants for 21 disease groups (observed in outpatient visits) through Poisson regression modeling. Daily air quality data and 1.6 million outpatient visit records from Shenzhen, China are used in the study. The outpatient visits are classified into 21 disease groups according to the *International Classification of Diseases*,

Digital Object Identifier: 10.1147/JRD.2017.2742707

Tenth Revision (ICD-10). The results show that associations between air pollutants and diseases vary across different disease groups. Specifically, the following disease classes are significantly associated with air pollution: blood, metabolic, ophthalmological, circulatory, respiratory, digestive, musculoskeletal, connective-tissue, and genitourinary diseases. Nitrogen dioxide (NO_2), particulate matter less than 10 μm in diameter (PM_{10}), particular matter less than 2.5 μm in diameter ($PM_{2.5}$), and air quality index (AQI) have the most extensive impact on more than 10 disease groups. A health effect graph is built to support public health management decision-making and provide residents with information about health effects of air pollution.

Fang et al. investigate the prediction problem, as well as policy design and resource allocation, for protecting wildlife from poaching. Wildlife species such as tigers and elephants are under the threat of poaching. To combat poaching, conservation agencies (defenders) need to (1) anticipate where the poachers are likely to poach and (2) plan effective patrols. The authors propose an anti-poaching tool CAPTURE (Comprehensive Anti-Poaching tool with Temporal and observation Uncertainty REasoning), which helps the defenders achieve both goals. CAPTURE builds a novel hierarchical model for poacher-patroller interaction. It considers the patroller's imperfect detection of signs of poaching, the complex temporal dependencies in the poacher's behaviors, and the defender's lack of knowledge of the number of poachers. Further, CAPTURE uses a new game-theoretic algorithm to compute the optimal patrolling strategies and plan effective patrols. This paper investigates the computational challenges that CAPTURE faces. First, the authors present a detailed analysis of parameter separation and target abstraction, two novel approaches used by CAPTURE to efficiently learn the parameters in the hierarchical model. Second, they propose two heuristics—piecewise linear approximation and greedy planning—to speed up the computation of the optimal patrolling strategies. They also discuss the lessons learned from using CAPTURE to analyze real-world poaching data collected over 12 years in Queen Elizabeth National Park in Uganda.

Yadav et al. also look at policy design and prioritization, focusing on using social networks to raise HIV (human immunodeficiency virus) awareness among homeless youth. Many homeless shelters conduct interventions to raise awareness about HIV infection among homeless youth. Due to human and financial resource shortages, these shelters need to choose intervention attendees strategically, in order to maximize awareness through the homeless youth social network. In this work, the authors propose HEALER (hierarchical ensembling-based agent, which plans for effective reduction in HIV spread), an agent that recommends sequential intervention plans for use by homeless shelters. HEALER's sequential plans (built using knowledge of homeless youth social networks) strategically select intervention participants to maximize influence spread, by solving POMDPs (partially observable Markov decision processes) on social networks using heuristic ensemble methods. This paper explores the motivations behind the design of HEALER and analyzes the performance of HEALER in simulations on real-world networks. First, the authors provide a theoretical analysis of the dynamic influence maximization under uncertainty problem (DIME), the main computational problem solved by HEALER. Second, the authors explain why heuristics used within HEALER work well on real-world networks. Third, they present results comparing HEALER to baseline algorithms augmented by the heuristics of HEALER. HEALER is currently being tested in real-world pilot studies with homeless youth in Los Angeles.

The next paper, by Goudey et al., discusses a framework for optimal health worker allocation in under-resourced regions. The effectiveness of health systems is dependent on the availability of health workers and their distribution across a healthcare system. Decision support systems for health workforce planning can play an important role in achieving an effective and equitable allocation of workers. However, existing methodologies rely on expert panels to determine the relationship between facility performance and staff levels rather than empirical evidence, require large amounts facility-specific data collection, and frequently fail to account for geographic, social, and economic differences between health facilities. The authors propose a framework for health worker allocation that overcomes some of these limitations. By integrating multiple sources of publicly available data with key facility-specific measures, statistical modeling can be used to estimate the relationship between staff allocations and facility performance. The resulting model can then be used in an optimization framework to explore how changes in policy scenarios and demographics can affect optimal staffing allocation. The authors explore this framework in a case study of South African health facilities, demonstrating the effectiveness of even this limited application of their framework, despite the challenges posed, and discuss the implications for future policy decisions and data collection.

Lamba et al. develop an approach for recommending allocations and designs of large philanthropic projects. The authors note that U.S. citizens donated an estimated \$373.25 billion to charity in 2015. These donations came from individuals, corporations, and various foundations. Most of the funds were used for launching projects focused on topics in specific regions of the world. However, there is infrequent formal knowledge transfer between the wide array of projects, and often no singular, unifying historical database. Therefore, an organization initiating a new project may not be aware of what organizations it can partner with, what the estimated value (or the budget) should be, and what learning can be derived from projects that have happened in the past. In this paper, the authors study data from the Clinton Global Initiative's Commitment to Action directory, a philanthropic project portfolio comprising 3,200 projects, 10,000 organizations, and multiple topics such as healthcare and education (as of June 2016). The authors propose a kernel-based tensor factorization approach that provides recommendations to organizations starting on a new project, based on the lessons learned from all previous projects.

Pham et al. discuss a prototype artificial intelligence tool that automates many of the tasks required for a real-time understanding of humanitarian crises. Humanitarian relief agencies must assess humanitarian crises occurring in the world in order to prioritize the aid that can be offered. While the rapidly growing availability of relevant information enables better decisions to be made, it also creates an important challenge: how to find, collect, and categorize this information in a timely manner. To address the problem, the authors propose a targeted retrieval system that automates these tasks. The system makes use of historical data collected and labeled by subject matter experts (SMEs) to train a classifier that identifies relevant content. Using this classifier, it deploys a focused crawler to locate and retrieve data at scale. The system also incorporates feedback from SMEs in order to adapt to new concepts and information sources. A novel component of the system is an algorithm for re-crawling that improves the crawler efficiency in retrieving recent data. The preliminary result of the authors shows that the algorithm can increase the freshness of collected data while simultaneously decreasing crawling effort. Furthermore, the authors show that focused crawling outperforms general crawling in this domain. Their initial prototype has received positive feedback from analysts at the Assessment Capacities Project (ACAPS), a humanitarian response agency.

Sherchan et al. similarly show how artificial intelligence can make disaster management more efficient. Social media has become a critical source of information for the public, government authorities, and other stakeholders both during and after large-scale emergencies. However, the sheer volume of data and the low signal-to-noise ratio, with respect to information, limit the effectiveness and the efficiency of using social media as an intelligence resource. This paper describes Australian Crisis Tracker (ACT), a tool designed to facilitate the understanding of critical information available in social media channels for people and agencies responding to natural disasters. ACT harnesses the Twitter streaming application program interface (API) by processing each tweet through a pipeline of analytic components, including filtering, metadata parsing, image extraction, and clustering of relevant tweets into events. Each of these events is then geocoded, categorized, and augmented with images from Instagram. The pipeline of these analytics is coupled to a web user interface that allows stakeholders to better access information during natural disasters. In this paper, the authors describe the ACT pipeline, analytics, and pilot by the Australian Red Cross during the 2013-2014 Australian bushfire season.

Patterson et al. broach the subject of collaboration platforms for working on difficult social good problems, with an artificial intelligence twist. Data science plays an increasingly important role in solving today's scientific and social challenges. To promote progress towards a cure for multiple sclerosis, the Accelerated Cure Project has created an open repository of biological and survey data on multiple sclerosis (MS) patients. Similar large-scale repositories are being created in other domains. As the open, data-driven model of science proliferates, the research community faces a growing need for a cloud platform for collaborative data science. Such a platform should facilitate collaboration between domain experts and data scientists and possess artificial intelligence capabilities for organizing, recommending, and manipulating data analyses. In this paper, the authors present some foundational technologies motivated by this vision. Their system automatically extracts a high-level dataflow graph from a data analysis. This graph describes how data flows through an analysis pipeline, including which statistical methods are used and how they fit together. The system requires no special annotations from the data analyst and consumes analyses written in Python using standard tools, such as Scikit-learn and StatsModels. In this paper, the authors explain how their system works and how it fits into their larger vision for a collaborative data science platform.

Bolten et al. also examine collaboration platforms, but with a focus on a pedestrian-centered data approach for equitable access to urban infrastructure environments. Crucial for a barrier-free city, *equitable pedestrian access* allows people with different abilities to independently access streets and services using relevant information. Pedestrians require both static and transient information regarding the street environment. Government stakeholders—such as municipalities, transportation agencies, and city planners—require accurate descriptions of the urban pedestrian environment to equitably carry out their mandates. However, pedestrian-centric data are generally unavailable in a standardized format, making it challenging to maintain and disseminate relevant information. This paper describes these challenges in the context of AccessMap, a customizable routing solution for pedestrians with limited mobility. Because existing routing solutions do not account for most barriers to accessibility, the information needs of these users are largely unmet. Using AccessMap as a case study, the authors demonstrate that a data model for equitable access to pedestrian information should: (1) include an annotated pedestrian transportation network, (2) be openly accessible, and (3) allow for the selective sharing of information to address the needs of all stakeholders. Finally, the authors generalize their experience to showcase a model of a *community-mediated data commons* that can contribute to better public sector functioning.

Kuhlman et al. discuss the development of a data proxy for economic competitiveness and innovation. Innovation is a key factor driving economic growth in countries worldwide. However, innovation is hard to define, and therefore even harder to measure. To help policy makers and business leaders better understand how to foster innovation, the authors need robust ways to quantify innovation at local and global scales. In this work, the authors use a data-driven, machine-learning approach to measuring innovation. Analyzing a large number of country-level metrics, the authors aim to automatically discover actionable levers of innovation. Using unsupervised learning methods the authors determine groups of related world development indicators among a collection compiled by the World Bank. The authors then train a Group Lasso predictive model using data from the World Economic Forum (WEF) that captures the perceived level of innovation in 150 countries. Aside from providing high predictive accuracy, the Group Lasso also provides a model that is easily interpretable. The result is the Open Innovation Index (OII), an automatic global model for measuring innovation using machine learning algorithms and open data. The authors present case studies for which the innovation levers of a few representative countries are uncovered automatically by the proposed model.

The final paper in the issue concerns the extraction of information from newspaper archives in Africa that may be used as a proxy for public safety, education, and healthcare. In this paper, Zeni and Weldemariam note that in sub-Saharan Africa, lack of useful information for the social good is one obstacle to the development of public services. This makes the extraction of data from digital archives [e.g., analog sources such as printed newspaper archives and born-digital sources like native portable document format (PDF)] an interesting alternative source of data to increase the amount and diversity of potentially useful information. Printed newspapers contain a variety of multi-article page layouts, wherein articles in the newspaper are designed to allow readers to define their own reading. The title of an article, the introductory story of the title, and related images are mostly grouped together. However, subsequent paragraphs and images are spread across various pages of the newspaper in a somewhat unpredictable manner. This, together with the poor quality of existing archives, makes the extracting of data from archived newspapers a daunting research problem. To solve these challenges, the authors present a system that extracts, detects, and clusters articles in newspapers from digital archives (mainly containing scanned newspaper archives from which the information is extracted). Finally, the authors also describe their proof-of-concept service using the extracted data.

In closing, we note that the papers in this issue only scratch the surface of a vast space of research and development in data science for social good. One of our goals has been to provide a compelling view of several solutions, models, capabilities, and collaborations that represent both innovative and practical steps in this area. If the enthusiasm for this special issue is any indication, the hybridization of data science expertise and technology with initiatives working towards the social good is poised to increasingly make the world a better place for ourselves and our future generations.

The cover art for this special issue is from Peshkova/iStockphoto.

A. Mojsilović
IBM Fellow
IBM T. J. Watson Research Center

K. R. Varshney
Research Staff Member and Manager
IBM T. J. Watson Research Center