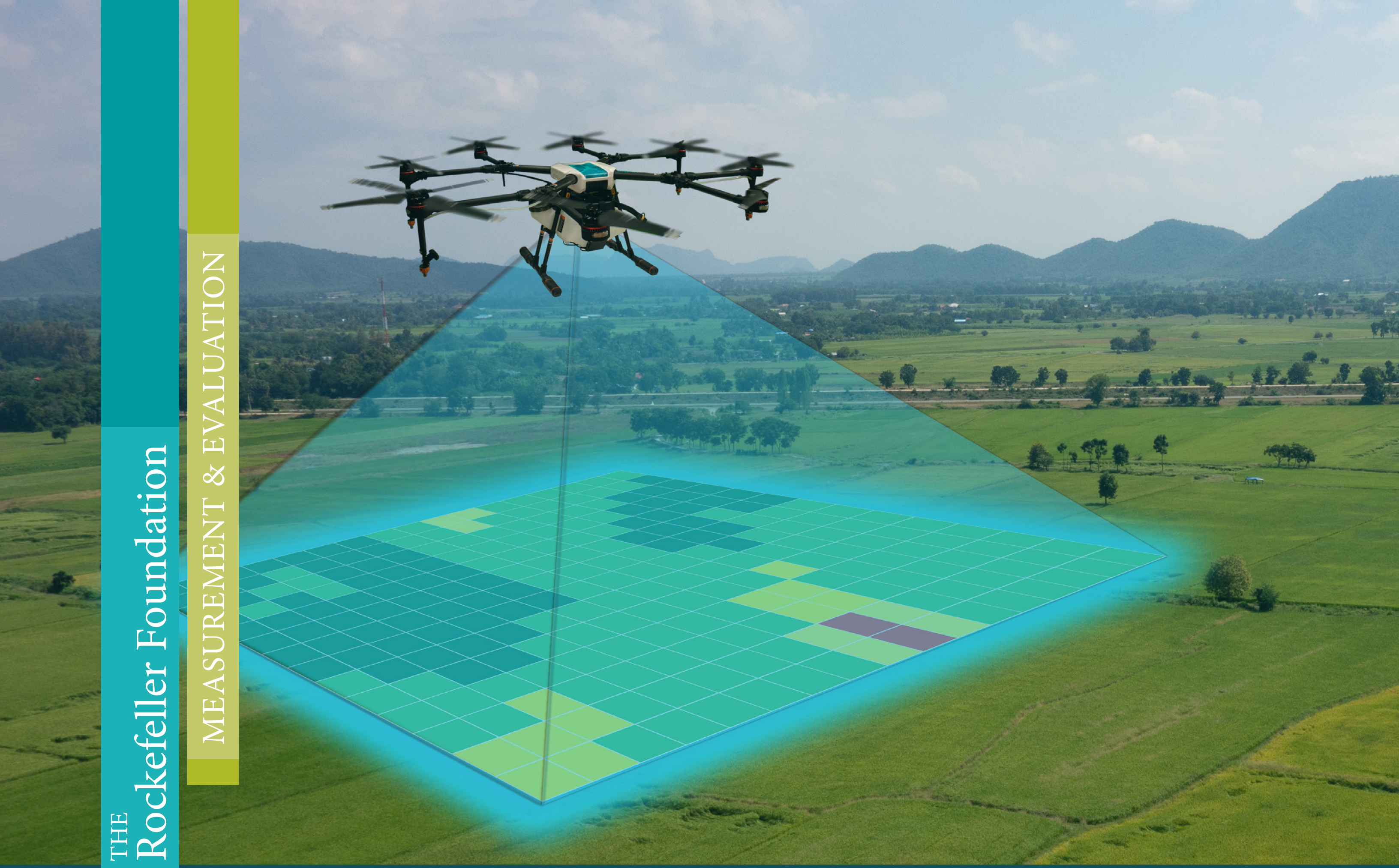


THE Rockefeller Foundation

MEASUREMENT & EVALUATION

Measuring results and impact in the age of big data:

The nexus of evaluation, analytics, and digital technology

Pete York and Michael Bamberger

Supported by

The ROCKEFELLER FOUNDATION

March 2020

About the authors

Pete York has over 20 years experience as an evaluation consultant and researcher working with government agencies, philanthropies, corporations, and nonprofit organizations. For the past eight years, he has built predictive, prescriptive, and causal evaluation models using large administrative datasets and machine-learning algorithms in the fields of child welfare, juvenile justice, mental health, workforce development, and others. He has recently co-authored a chapter of *The Application of Predictive Analytics and Machine Learning to Risk Assessment in Juvenile Justice: The Florida Experience*, and wrote a peer-reviewed article, *Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience*, and a Scattergood Foundation paper, *A New Way to Use Data: Precision Care for Better Outcomes in Residential Treatment for Children*. He is currently a Principal and Chief Data Scientist at BCT Partners, where he applies his evaluation and data science skills to advance the organization's mission of "providing insights about diverse people that lead to equity."

Michael Bamberger has been involved in development evaluation for over 40 years. After working in development and evaluation in Latin America for over a decade, he joined the World Bank, where he worked on urban research, evaluation, and gender and development. Since retirement, he has consulted with UN agencies, development banks, foundations, bilateral aid agencies, NGOs, and developing country governments. He has studied the potential applications of big data and data science in the evaluation of development programs, and has researched the reasons why evaluators have been slower than other development practitioners to adopt big data and data science approaches. Recent publications include: *Integrating big data into the monitoring and evaluation of development programs, Dealing with complexity in development evaluation,* and *RealWorld Evaluation: Working under budget, time, data and political constraints*.

About The Rockefeller Foundation

The Rockefeller Foundation advances new frontiers of science, data, policy and innovation to solve global challenges related to health, food, power and equity & economic opportunity. As a science-driven philanthropy focused on building collaborative relationships with partners and grantees, The Rockefeller Foundation seeks to inspire and foster large-scale human impact that promotes the well-being of humanity throughout the world by identifying and accelerating breakthrough solutions, ideas and conversations. For more information, visit www.rockefellerfoundation.org.

THE Rockefeller Foundation

MEASUREMENT & EVALUATION

Measuring results and impact in the age of big data:

The nexus of evaluation, analytics, and digital technology

March 2020

Pete York and Michael Bamberger

Supported by

The ROCKEFELLER FOUNDATION

The contents of this report are the views of the authors and do not necessarily reflect the views or policies of The Rockefeller Foundation.

© STARS,Kristen Buus, 2012

Table of Contents

BOXES

TABLES

FIGURES

Acronyms

AEA	American Evaluation Association
API	Application programming interface software
AWS	Amazon web services
BD	Big data
EES	European Evaluation Society
GBV	Gender-based violence
GIS	Geographic information system
GPS	Global positioning system
ICT	Information and communication technology
IOT	Internet of things
ML	Machine learning
NIT	New information technology
OFW	Overseas foreign worker
PSM	Propensity score matching
QED	Quasi-experimental design
QCA	Qualitative comparative analysis
RCT	Randomized controlled trials
ToC	Theory of change
ToR	Terms of reference

Preface

The world today is more connected, interdependent, and data-rich than at any time in human history. Yet we increasingly see populations divided into those who benefit from the policies, products, and services driven by advances in data science, and those who are left behind or actively harmed.

At its best, the global development sector creates tangible improvements in people's lives, with evaluation among the sector's most critical tools for knowing what is and is not working. By taking advantage of new thinking emerging from the field of data science, development sector practitioners and professionals can make their toolkit even more powerful.

The benefits – and challenges – of big data are now spreading rapidly throughout the world, increasingly reaching the poorest and most inaccessible areas and, in turn, revolutionizing the way global challenges can be solved. This includes providing new opportunities to understand and track problems in society, designing and scaling new solutions to address them, and enabling evaluators to move more rapidly in measuring and assessing the impact that development programs have on poor and vulnerable people.

With our history of pushing the envelope on innovation, The Rockefeller Foundation works to secure the fundamentals of human well-being for everyone, everywhere. Grounded in what we've seen work for more than a century, our approach is inspired by science and rigorous about data, it brings together and empowers others, and is focused on real results that improve people's lives.

To this end, we are pleased to have supported this report, as it brings together the data and social sciences – two distinct schools of thought. In doing so, it offers an opportunity to explore how data science and new digital technology can be used to measure the results and impact of development assistance interventions.

We are thankful to Michael Bamberger and Peter York for their forward-looking input, deep technical guidance, and leadership in managing this important report through completion. We hope their insights will help organizations rethink what is possible when it comes to using data science to improve people's lives.

Veronica Olazabal
Senior Advisor & Director
Measurement, Evaluation & Organizational Performance

Tariq Khokhar
Managing Director (former)
Data and Technology

Executive Summary

We are living in a world that is increasingly dependent on big data and data science in every aspect of our personal lives and our economic, political, and social systems. Big data also plays an ever more important role in research, in large part because there are powerful new user-friendly analytic methods that make all of the world's rapidly growing data accessible and more meaningful to an increasingly wider range of audiences. These trends began in industrialized nations not too long ago and now are expanding at an exponential rate in middle-income and even low-income countries. With all of this rapid expansion of big data and analytics, it is time for the two fields of program evaluation and data science to come together in order to more rapidly and cost-effectively learn what works, improve social solutions, and scale positive impact as never before.

Chapter 1 of the paper looks at how data science vastly broadens the range of data and analytical tools, including predictive models, that evaluators can draw upon. These tools can reduce the time and cost of data collection while greatly increasing the range and depth of analysis. Integration can free up evaluators' time and energy, allowing them to focus on the important aspects of evaluation design, strengthen the quality and validity of data sources, and dig deeper into explaining what happened as a result of the program, why it happened, who benefited, and who did not.

Yet, despite this exciting potential, evaluators have been much slower than other development practitioners in adopting the tools and techniques of data science. Thus, in addition to discussing how evaluators can use the new data sources and analytical tools, the paper also examines the methodological, economic, organizational, and even political reasons for the slower take-up, and argues that all development partners need to work together to build bridges between data scientists and evaluation – to ensure the promise of integration is realized.

Chapter 2 shows how big data is increasingly used in all aspects of our personal and professional lives, and throughout the commercial and public sectors. In defining big data, it discusses the significant differences between big data and the kinds of conventional data currently used in most evaluations. It also identifies three main categories of big data as they relate to evaluation: i) human-generated (centered) data, ii) administrative (transactional) data, and iii) geospatial data, all of which play important but distinct roles in evaluation. Problems arise when the differences between the data science ecosystem and the conventional evaluation ecosystem are not well understood by evaluators, making it difficult for them to work with big data. The chapter further identifies some of the issues and the methodological, political, and organizational challenges of managing big data for social good. Attention is also drawn to a set of important ethical issues which often are not fully understood or are ignored.

Chapter 3 discusses the increasing use of big data in social and economic programs in both industrialized and developing countries. Agriculture, health, and education are among the sectors in which big data has had the most impact. Developing countries use all three kinds of big data described in Chapter 2 – in their emergency programs and for disaster relief, program design, management and monitoring, and information dissemination. However, big data tools and techniques have been used much less extensively in program evaluation than in other program activities. Evaluation offices' slow take-up is often because: i) evaluators and data scientists traditionally work in different contexts; ii) the two professions tend to use different methodological frameworks and analytical tools; iii) there are weak institutional linkages between data science and evaluation offices, even when both offices are in the same organization; and iv) training for big data and training for evaluation often do not cover the methodological approaches and philosophies of the other. The slow progress of integration has potentially serious consequences for both evaluators and data scientists. At the technical level, there are important approaches, tools, and techniques that each could learn from the other. At the broader level, there is a risk that evaluation philosophy, approaches, and methods could become marginalized as organizations start to introduce integrated data management systems driven by data science – under which program performance assessment would be increasingly conducted using only data science methods.

Chapter 4 begins by introducing some of the big data tools that can help strengthen evaluations by reducing the time, cost, and effort required for data collection, sample design, and data analysis. Data science makes it possible to collect a vastly increased range and volume of data more easily, quickly, and economically. The ability of big data to include all of those in an entire population with a particular attribute, rather than just a relatively small sample, makes it possible to avoid many kinds of selection bias, and enables disaggregation of the sample to cover many different sub-samples and categories. Big data also means data analysis can be conducted more rapidly and cheaply, and makes it possible to conduct much more sophisticated and complex analysis.

Machine learning, one of the most important developments, brings a completely new focus to hypothesis development and to integrated analysis, by combining multiple kinds of quantitative and qualitative data in ways which were not previously possible. The chapter includes a case study of how the performance of a large child welfare program was improved, illustrating how machine learning, predictive analytics, and other big data techniques can be applied in the real world. The chapter ends with a discussion of some of the necessary conditions for integration of data science and evaluation to take place.

Chapter 5 has a more technical focus. It discusses some of the areas of disagreement or misunderstanding between evaluators and data scientists, and categorizes them into issues relating to theory, data quality and validity, and sampling and sample selection bias. In each case, the main areas of debate are described, and the authors distinguish between: i) misunderstandings, many of which could be relatively easily clarified, and ii) disagreements, which focus on the strengths and limitations of different approaches and methods. There is, of course, a fine line between misunderstanding and disagreement, and proponents of a particular approach, who believe it is the "best" or "correct" way to conduct a study, will often argue that those on the other side would agree with them if they had a better understanding of their approach.

The distinction between misunderstanding and disagreement is important, because different approaches may be required for bridge-building between data scientists and evaluators. The chapter concludes that discussions about theory are likely to be the most difficult to resolve, as theoretical positions tend to combine philosophy and values as well as methodology. In contrast, greater understanding of different methodological approaches is likely to provide a basis for broader agreement on issue relating to data quality. Issues of sample design probably lie between theory and data quality, as some issues relate directly to specific questions of methodology, but sampling also tends to involve some broader issues of philosophy and broader approaches to research.

Chapter 6 draws on the previous two chapters to illustrate how data science tools and techniques can be applied to strengthen evaluations. The chapter begins by discussing some of the challenges that conventional evaluation approaches face and to which big data could potentially contribute. Challenges are discussed with respect to: evaluation design, data collection, sample design, and data analysis and dissemination of findings.

- ***Evaluation design challenges*** include defining the counterfactual when experimental designs are not possible, evaluating complex programs, and identifying unintended outcomes.
- ***Data collection challenges*** start with the cost and time required for the data collection, but also include collecting information on difficult-to-reach groups, monitoring implementation processes and behavioral change, integrating different kinds of data, and collecting information on the spatial dimension of programs which includes changes that take place outside the immediate project area.
- ***Sample design challenges*** include reducing sample selection bias and ensuring the sample is sufficiently large to ensure statistical significance and to permit disaggregated analysis.
- ***Data analysis and dissemination challenges*** include working with very large data sets, integrating data from multiple sources and in different formats, analyzing complex programs, and providing actionable predictions of the likely effects of different interventions on different groups. The dissemination of evaluation findings – in a timely manner and in formats that are understandable to different groups – has also proved to be a challenge.

The chapter then illustrates how the incorporation of some of the data science techniques and approaches discussed in the two previous chapters can help address these challenges. The message throughout is that data science should be used to complement, not to replace, conventional evaluation approaches.

Chapter 7 uses the example of gender differences in the impacts of big data and other new information technologies to illustrate the importance of understanding how different groups – based on gender, age, income, or geographical location – have access to, use, and are affected by the new technologies. It begins by illustrating how women and men have different experiences with the new information technology. There are many social, economic, and cultural factors that affect women's access to mobile phones and other information technologies, as well as examples of the negative consequences women may face related to mobile phones, including increased gender-based violence. The chapter then identifies some important new research areas that should be explored, and it concludes by identifying a list of factors that can affect access to new information technologies by women and vulnerable groups.

The concluding Chapter 8 summarizes lessons about the potential benefits and challenges at the nexus of data science and evaluation practice, and builds on identifying a set of recommendations on ways to move forward to create a conducive environment for integration. The lessons include: recognition that the exponential increase in the availability of, and applications for, big data is creating a new and complex information ecosystem that is fundamentally changing how data is generated and used; and there is increasing recognition of the wide range of potential benefits from integrating data science and evaluation. There are also a number of lessons concerning conditions required for integration to occur, including the need for creating a conducive policy environment, ensuring open data access for all sectors and not just for a few powerful commercial and government agencies, facilitating wider access to advanced computing facilities and analytical expertise, and creating organizational structures, at the organizational, national, and international levels, that promote cooperation and convergence. The lessons conclude by identifying a number of challenges, such as deciding how big data should be controlled and regulated, and determining who has access, and how to avoid economic and ethnic bias in access and in addressing issues of privacy and security. It also recognizes that while big data has the capacity to empower and give voice to poor and vulnerable groups, it can also be used "extractively" by decision-makers who use information collected from and about poor people to make decisions about priority programs for these groups without having to consult them.

The chapter also includes a set of recommendations that identifies priority issues to be addressed in moving towards solutions at the nexus data science and evaluation, and the benefits it can produce. These include: establishing priority measures to build bridges between the two groups; developing capacity development programs that combine data science and evaluation approaches, tools and techniques; and promoting landscaping research to improve understanding of the data science and evaluation ecosystems and how they can be better integrated. Finally, it emphasizes the potential key role for funding agencies in creating space for dialog and collaboration, and providing critical seed funding in areas such as collaborative research and training programs.

Many actors must be involved in promoting and facilitating the changes required to achieve integration, including: the institutions that train data scientists and evaluators; the organizations that plan and fund social and development programs; program managers in the public, business, and non-profit sectors; and, of course, the data scientists and evaluation practitioners. Foundations will play an important role in achieving this change, as they have both the resources and the flexibility to create spaces for innovation and new forms of collaboration.

Glossary of terminology:

Recognizing that this is a new field and those involved have adopted or created a vernacular for what they do – the problem is that different groups use different terms to describe the same issues. Thus we call attention to the Glossary of technical terms found at the conclusion of this paper. The authors have compiled definitions of the terms that are part of the field to ensure that readers have equal understanding of their nuance. And, as with the rest of this paper, it remains a work-in-progress and further suggestions would be more than welcome.

Chapter 1 makes the case for the importance of working towards the integration of data science and program evaluation. As a part of data science, big data vastly broadens the range of data and analytical tools (including predictive models) that evaluators can draw upon. These tools can reduce the time and costs of data collection while greatly increasing the range and depth of analysis. Use or integration of big data can free up evaluators' time and energy, thus allowing them to focus on the important aspects of evaluation design, strengthen the quality and validity of data sources, and dig deeper into explaining what happened as a result of a program, why it happened, who benefited, and who did not. Yet, despite this exciting potential, evaluators have been much slower than other development practitioners in adopting the tools and techniques of data science. Thus, in addition to discussing how evaluators can use the new big data sources and analytical tools, the paper also examines the methodological, economic, organizational, and even political reasons for the slower take-up, and argues that all development partners need to work together to build bridges between data scientists and evaluation – to ensure the promise of convergence is realized.

1

Promoting the integration of data science and evaluation: New frontiers and remaining challenges

We are living in a world that is increasingly dependent on big data and data science in every aspect of our personal lives and our economic, political, and social systems. Data science also plays an increasingly important role in research and in how all kinds of economic and social programs are designed and managed. Evaluators now have cost-effective, fast, and simple access to vast new sources of data – economic, social, political, demographic, environmental, attitudinal, and behavioral – many of which could not have been imaginable even a few years ago. There are also powerful new, user-friendly analytical tools, predictive models for the analysis of all of this data, and user-friendly ways to make the data accessible to an increasingly wide range of audiences. Although these trends mostly began in industrialized nations, they now are expanding at an exponential rate in middle-income and even low-income countries. In fact, more people now have access to cell phones than to potable water – although we could debate the implications of examples such as this.

This paper presents the case for integrating the tools of the big data world – in particular data science – into the field of program evaluation. It illuminates and documents the tremendous benefits, already in evidence, that can be achieved in equity-oriented social and economic development in both developing and industrialized countries. The paper also documents the technical, political, economic, organizational, and even philosophical factors that have slowed, and some feel could even derail, the achievement of integration and its potential multiple benefits.

In addition to presenting evidence of the benefits that have already been achieved in improving the quality and effectiveness of evaluation and the resulting economic and social benefits for global development programs, the paper also examines the factors that have slowed integration and proposes practical steps that could be taken by a wide range of actors to build the bridges necessary to achieve the many benefits of integration.

For a number of reasons, the agencies responsible for evaluating social programs in both industrialized and developing countries have been much slower in adopting data science approaches than their colleagues who work in research and planning. That said, it must be recognized that data science and program evaluation are built on different traditions and use different tools and techniques. A future that embraces working together will require both groups to move out of their comfort zones.

Looking toward this ideal, the paper discusses the many exciting ways that data science can strengthen evaluation and that evaluation approaches can strengthen some of the data science approaches. Yet, despite many promising approaches, there remain a number of challenges that must be overcome before integration can be achieved and the full benefits of data science enjoyed. All parties must be involved in bridge-building.

Some of the promising areas in which data science can make the greatest potential contributions to evaluation include the following (for more detail, see Chapter 6).

- **Reduce the time and cost of data collection so that evaluators can focus on the critical evaluation tasks of defining evaluation questions and, in turn, develop a theoretical framework for the evaluation and for the analysis and interpretation of its findings.** Many evaluators have to spend so much time and effort on the collection and analysis of data that they have very little time or lack resources to focus on the critical elements of the evaluation process. Using data science to free up time will allow evaluators to focus on the areas of data quality, enabling them to spend more time focusing on the communities being studied, triangulation, ground truthing, and mixed methods – how many evaluators lament not having the time to address these issues properly?
- **Expand the kinds of data that can be collected and analyzed.** This potential for dramatic expansion includes: i) access to artificial intelligence (AI) which makes it possible to identify patterns in huge volumes of multiple kinds of data, and ii) access to a range of predictive analytics tools, which makes it possible to develop models and analytical tools for evaluating complex programs. Expanding the kinds of data collected and analyzed also advances the possibility of studying longitudinal trends, in some cases over periods as long as 20 years. This makes it possible to observe historical trends before a program is launched while also tracking sustainability of program-induced changes, maintenance of program infrastructure, and continued delivery of services. All of these are virtually impossible with conventional evaluations that have defined start-and-end dates.
- **Use algorithms incorporating AI and data mining to process huge volumes of data and, in turn, improve decision-making and prediction of the best treatments for different groups affected by a program.** Their use generates a very powerful set of tools for evaluation managers and policy makers that can: i) improve their ability to analyze the factors affecting outcomes for individuals or small groups, and ii) provide specific real-time recommendations on the best treatment or combination of treatments for each small group or individual. The use of these tools contrasts with conventional evaluation designs that usually only make recommendations on how to improve the average outcome for the whole population. That said, many of these algorithms are based on complex predictive models which often are not well understood by users because they are complex, but also because they are proprietary and not usually made available to clients. Consequently, there is a danger that some algorithms can have unintended negative outcomes that clients may not even be aware of (for more detail, see Sections 2.1, 2.2, and 5.1).
- **Use AI to combine multiple data sources into a single integrated data platform.** Although this idea has received less attention and appears less exciting than the changes described above, doing so makes it possible to explore relationships among the different data sets, which was not previously possible. For example, a program to combat modern slavery in the Philippines used an integrated data base to bring together multiple and quite different data sources, and identified relationships among variables that had previously been very difficult to detect (for more detail, see Section 2.2).

Most discussions in the literature see data science as an exciting new frontier that can assist evaluators – evaluators who are often portrayed as having fallen behind the times with respect to the use of new technology. However, it is important to recognize there are also potential weaknesses in data science approaches. Many data science approaches were originally developed in and for much simpler and less demanding environments, such as marketing analysis and on-line advertising. In these cases, an on-line advertiser may only be interested in correlations in correlation. For example, if the font size and color are changed, will more visitors to the site click on the ad? Or, are men who purchase diapers in the supermarket also likely to purchase beer? In these cases, the client does not need to know why this relationship exists. Also, the quality of the data does not matter too much, as the mantra of many data scientists is that even if "all data is bad and all data is biased, it does not matter as we will have new data tomorrow." Because of the limited demands of clients, many data scientists do not have to develop the kinds of theoretical frameworks and theories of change used by most evaluators.

For all of these reasons, when data scientists and app developers venture into the new world of community development – designing complex programs for disadvantaged communities and trying to explain why a program produces certain outcomes for some groups and not for others – there are many lessons that data scientists can learn from their evaluation colleagues, such as the need to:

- have greater concern for the quality and validity of data
- understand the importance of *construct validity,* which calls for interpreting those indicators extracted from social media, phone call records, or satellite images; and determine, for example, how changes in the number of references to hunger or sickness can be used as indicators of changes in short-term poverty levels, or what satellite counts of the proportion of roofs constructed of straw compared with zinc, tell us about changes in poverty levels
- .address issues of social exclusion and sample bias
- insure causality by creating or observing counterfactuals (comparison groups) and controlling for selection and other biases inherent to all types of human-centered transactional data
- rethink the role of theory and the need to base an evaluation on a theory of change
- recognize the importance of *ground-truthing*, which calls for checking on-the-ground hypotheses generated from the analysis of remote, non-reactive data.

The paper has several goals. First, it aims to identify the many potential benefits that the integration of data science and evaluation can contribute to achieving greater social good, particularly for poor and vulnerable groups. Second, it provides multiple examples that show how integration is already happening, although at a rate slower than had been hoped. Third, it identifies the challenges – methodological and philosophical, political, economic, organizational, educational and cultural – that are slowing progress towards full convergence of the two disciplines. Finally, its ultimate goal is to identify the practical actions that must be taken to build the bridges, create a conducive environment for integration to take place, and provide exciting opportunities for improving how social and development programs and policies are designed, implemented, and evaluated in both industrial and developing countries.

Many actors must be involved in promoting and facilitating the changes required to achieve integration, including the institutions that train data scientists and evaluators; the organizations that plan and fund social and development programs; program managers in the public, business, and non-profit sectors; and, of course, the data scientists and evaluation practitioners. Foundations will play an important role in achieving this change, as they have both the resources and the flexibility to create spaces for innovation and new forms of collaboration.

Chapter 2 shows that big data is increasingly used in all aspects of our personal and professional lives, and throughout the commercial and public sectors. It defines big data and discusses the significant differences between big data and the kinds of conventional data currently used in most evaluations. Further, it identifies three main categories of big data as they relate to evaluation, namely: human-generated (centered), administrative (transactional), and geospatial. Each of these types of data plays an important but distinct role in evaluation. There are important differences between the big data ecosystem and the conventional evaluation ecosystem, but they are often not well understood by evaluators, which is one of the deterrents for evaluators to take-up big data. Chapter 2 also identifies some of the methodological, political, and organizational issues and challenges associated with managing big data for social good, as well as a set of important ethical issues which are often not fully understood or are ignored.

2

Big data in an interconnected world

2.1. The increasing role of big data in our daily lives

The world today is more connected, interdependent, and data-rich than at any time in human history. Exponential growth in the volume of data produced globally means that 90 percent of all the data in existence today was generated in just the past two years. An explosion of digital services over the past decade has allowed many new actors to become producers, owners, and consumers of data. Between 2005 and 2015, the number of internet users more than tripled – from 1 billion to 3.2 billion – and more households now own mobile phones than have access to electricity or clean water (World Bank, 2016).

The exponential growth of big data and data analytics provides information and analytical capacity that would have been unimaginable even a few years ago. Digital data brings a trove of real-time information on many issues, such as the cost of food, availability of jobs, access to health care, quality of education, and reports of natural disasters. The 2016 World Development Report was dedicated to the analysis of big data, the potential benefits – or "digital dividends" – that they offer, and the major challenges of the continued digital divide (World Bank, 2016).

Just a decade ago, much of this information was available to the relatively small number of organizations that had access to large computing capacity. Today, a number of developments with major political and cultural implications are evolving.

- Individuals and small organizations with limited resources now have access to increasing amounts of information about themselves, their communities and the wider world.
- Individuals and groups now generate much of this information, rather than being passive consumers.
- Commercial, political, and other kinds of agencies have increasing amounts of data about individuals and organizations. This data can be used to manipulate information and motivate behavior of consumers, workers, voters, and members of social, religious, and cultural organizations.
- Social media can be used to communicate with, and hold accountable, political, economic, and other kinds of organizations.

BOX 1:

We are all leaving huge traceable trails of our lives

With people leaving so many swipes and key entries on so many tech platforms, they create traceable trails which have become the real logs of their lives and identities.

- Google searches – what's on their minds
- Facebook – how they want their friends and family to view them
- Twitter – what's in the news or community "air" that others should care about
- Credit card swipes – what they buy
- GPS – where they go
- In-person connection apps – with whom they interact
- LinkedIn – what they do, and how well and with whom they do it
- Podcasts/media streams – what their interests are and how they spend their leisure
- Work productivity apps – how much time they spend on personal communications or social media during working hours

Taken together, it is obvious we are all generating huge streams of traceable data about all aspects of our lives, over which most people have very little knowledge or control (see Box 1).

All of these developments have important applications for research and planning. For example:

- biometric data generated from the Internet of Things (see Glossary) have produced a rapidly evolving research area on the Quantified Self (Wolf, 2014)
- sentiment analysis, socio-metric analysis, and digital spatial analysis, among others, have made research on the Quantified Community possible (Kontokosta, 2012)
- satellite images, social media analysis, and cellphone data on traffic patterns, among many other sources, have contributed to new fields of research on city planning, urban development, migration, and poverty analysis (Ashton et al., 2017)
- satellite images and remote sensors have greatly advanced climate research.

2.2. Defining big data and what it means for impact measurement

Multiple sources of big data

The multiple sources of big data introduced in Boxes 1 and 2 are, or could be, used for program evaluation. Even though none of this data was generated for the purpose of evaluation, it can be manipulated and synthesized into a format that can be used for evaluation. It is also important to point out that while many discussions of big data state that it is fast and cost-effective to use, gathering most original data – such as satellite and remote sensor images, digital financial transaction data, social media streams, internet searches, GPS location data, and .pdf data files – is in fact very expensive and difficult to process, as it requires accessible and powerful computer facilities. However, processed (transformed) data – such as phone and other apps, publicly available social media files, and data visualization outputs – is economical and easy to use, as it is available in the form required for data analysis.

Defining big data

Big data has huge volume, is generated very fast and often in real time, is too large to be analyzed on a single computer, and in most cases, is collected around the clock in perpetuity (see Table 1). It is also non-reactive, as it is collected digitally and for a different purpose than an evaluation. This means that evaluators cannot design information for the specific purpose of the evaluation and, instead, will usually use proxy variables they assume to be relevant and reliable for the purpose of the evaluation.

Big data has many benefits for evaluation research, including features such as:

- having ability to obtain data on the total population, rather than from a relatively small sample – which increases the granularity of the data and makes it possible to compare small groups or even individuals

TABLE 1. Defining big data

BASIC CHARACTERISTICS	
Generated very fast – often in real-time	Some kinds of data are accessible almost immediately (traffic movements) while others may be available hourly, daily, or over slightly longer periods.
Primary data is too large to analyze on a single computer	Processed data is accessible to organizations with limited computing capacity.
Always on	Some kinds of data are available at any hour and are never turned off.
Non-reactive	Data has normally been collected for some other purpose and, consequently, the act of accessing the data for the evaluation does not in any way affect the data. This contrasts with most evaluation data, where respondents are aware that the data is being collected and may change their response, depending on how they perceive the purpose of the evaluation or based on their attitude toward the data collector.
Networked/connected computers and systems	Networked and connected computers and systems permit contextual learning and evaluation.
ADVANTAGES FOR EVALUATION RESEARCH	
Complete population coverage	Most big data analysis covers the total population of interest, such as: everyone using an ATM machine or posting on a particular app. However, it is important to recognize that all users of an app or everyone using an ATM machine present a biased sample of a wider population, such as: all young people living in a particular geographical area.
Granular	Data can be broken down to the level of small groups or even to the individual level.
Longitudinal data sets increasingly available	Data from satellites or social media sites now cover increasingly long periods of time, so it is becoming possible to collect data before a program begins as well as after it ends.
Multiple kinds of data integrated into a platform with common metrics	An integrated platform enables combining quantitative, qualitative, location data, narrative, pictures, and video. It offers a huge expansion of the kinds of evidence that can be used.
Economic data collection for vulnerable and difficult-to-reach populations	This data can be more representative of the poor.
Qualitative and behavioral data	This data can be easily captured.
SOME LIMITATIONS	
Incomplete	Easily accessible data rarely cover the total target population and often cover a short period of time, such as publicly available Twitter data which usually only covers 7 days.
Inaccessible	Only app developers and a few large organizations have direct access to the primary data sources.
Non-representative	Data are only obtained from people who have mobile phones, who use a particular app and who chose to use that app.
Dirty	Many miss data points, lack variability within key metrics, and there are, for example, misspellings, inconsistent formatting of data points, data entry errors, lack of standardization of response categories and types, and changes in variable response codes over time.
Drifting	The population covered will often change over time, and often in ways that are difficult for the user to understand. For example, the kinds of people using a particular social media site can gradually change as the site becomes more popular or goes out of fashion. These changes are difficult to track.
Algorithmically confounded	Behavior in big data systems is not natural; it is driven by the engineering goals of the system (Salganik (2018:35).
Sensitive	Some data sources contain sensitive personal, financial, or political data, and the researcher should respect confidentiality.
Indirect	Using data collected for other purposes means assumptions must be made about how well the proxy data captures the research constructs.

Source: the authors and Salganik, 2018.

BOX 2:

Examples of big data sources that are, or could be, used for evaluation

- Satellite images
- Remote sensor data (e.g. film photography, video, infrared)
- GPS location data (e.g. tracking traffic or pedestrian travel patterns)
- Social media streams (e.g. Facebook, Twitter)
- Social media profile data
- Call-in radio data
- Internet search data
- PDF files (often including millions of organization digital files)
- Integrated data platforms that merge many secondary data files (e.g. surveys, agency reports)
- Biometric data (e.g. smart watch data on heart rate, calorie expenditure)

- reducing the cost and difficulty of reaching vulnerable and inaccessible groups, which makes sample coverage more complete and representative
- integrating different kinds of data into a single database with a common metric, so that all can be combined in the analysis
- increasing availability of longitudinal data.

Added together, this indicates that with big data, it becomes possible to evaluate programs that operate in complex contexts and to capture data on processes and behavioral change.

Comparing big data and conventional evaluation data

Table 2 identifies some of the common differences between big data and the kind of data currently used in most evaluations. For example, big data can be collected more economically and rapidly, and it usually covers most of the population of interest compared with the relatively small samples used by most evaluations. Also, because of the large samples, it is more granular, permitting more disaggregated analyses. That said, it is also important to consider the criticism that few big data sets cover the total population – although the same can be said of many of the sampling frames used by evaluators.

The contribution of big data is further enhanced by the increasing availability of longitudinal data sets, such as 20-year time series showing changes in forest cover, and social media streams that cover decades or more. In addition, it is now possible to include much more contextual data. New information technologies also permit faster and more user-friendly dissemination.

With data convergence, a new class of "big data" is emerging from the non-profit, philanthropy, and NGO social sectors. The old and always present tracking and monitoring data of the past are now growing and evolving at a record pace due to low-cost, cloud-based computing data systems, and connected case-management applications that all work together to capture the details of every transaction, unit of service, or activity. As a result, the largest government agencies and NGOs no longer have the best and biggest data – their small and medium counterparts are beginning to generate more data, sometimes in collaboration with one another.

Additionally, all of these tracking and monitoring datasets now include historical data that cover, for example, every case or situation and all assessments, diagnostics, or tests taken by or captured from a case, plus the data points captured by each case using

TABLE 2: Comparing big data and conventional evaluation data

COMPARISON	BIG DATA	CONVENTIONAL EVALUATION SURVEY DATA
Costs of data collection and analysis	While the initial costs of collection and analysis of primary big data can be high, processed data will often be available to researchers and evaluators (end users) at a relatively low cost.	Costs of data collection and analysis are high.
Speed of data collection, analysis	Fast	Time-consuming
Coverage of population and ease of disaggregated analysis (granularity)	Data often covers the whole population. Low cost of data permits disaggregated analysis to the level of the individual.	Data collection costs produce pressures to keep sample size as small as possible, but consistent with power analysis requirements.
Collecting contextual data covering a broader population	Big data can collect and synthesize a broad set of variables that cover regions, or national and international data. It also combines multiple types of information.	Most evaluations only collect a limited amount of contextual data as it is expensive and difficult to collect.
Potential sample bias	Many indicators only cover people who have and use a phone or app, or who use ATMs, which is usually a biased sample of the total population. Procedures exist for estimating and controlling bias in small phone surveys.	Surveys have procedures for controlling sample selection bias or response bias.
Ease of dissemination	Dissemination is fast, sometimes real-time, to users who have access to required digital technology.	Evaluation findings can be expensive and time-consuming to produce and distribute, and often are only distributed to priority stakeholders.
Relevance and appropriateness for a particular evaluation purpose	Most big data was collected for a different purpose and assumed to be appropriate proxy indicators.	Data collection instruments and questions are designed for a special evaluation purpose.
Longitudinal and time-series data	Longitudinal data sets already exist and the number is increasing – some already cover more than a decade.	Longitudinal data is difficult and expensive to collect.
Combining multiple data sources	Integrated data platforms can be developed, and they are very powerful tools. However, they are very time-consuming to construct.	Integrated data platforms are difficult to construct. Mixed methods are widely used but usually they have to be analyzed separately and compared manually.
Creation and analysis of qualitative data	The technology is rapidly improving for the fast and economic analysis of qualitative data such as narrative text, and video and audio files	Qualitative data is expensive and time-consuming to analyze. Case studies and in-depth interviews can still explore lived experience that is difficult to achieve with big data.

BOX 3:

The three categories of big data.

Big data can be classified in three main categories, depending on how it is generated and used. And each can be used in different kinds of program evaluation.

- **Human-generated (centered)** data includes social media, phone call records, ATM transactions, internet searches, audio and visual material, and text.
- **Administrative (transactional) data** includes program administration data such as supervision reports, monitoring data, outreach worker reports, government publications and reports, and survey data collected and stored by governments, donor agencies, and others. It also includes government administrative data such as migration records or employment data. Of note, most administrative data is underutilized or, if used, each kind of data is used separately. Data analytics now makes it possible to combine multiple, very different kinds of data into an integrated data platform, where the relationship between different kinds of data can be examined for the first time. Box 4 illustrates how an integrated data base was created to help understand the multiple and complex factors affecting modern slavery in the Philippines.
- **Geospatial data** includes satellites, drones, and remote sensing. Using this kind of data makes it possible to study large geographical areas and, for example, to monitor characteristics of the natural environment or human patterns of movement and settlements.

external datasets. It is these individual and case-level big data sets – rather than grouped aggregations of data – that will be required for evaluators to determine if a program, service, or activity contributed or led to the desired outcome. These "enhanced" programmatic big datasets are where the greatest evaluation opportunity lies. This trend is rapidly evolving among US-based NGOs, with middle-income countries likely to follow. The evolution may be slower in many low income countries.

In the US, many non-profit organizations have adopted big data. Most of the medium and large organizations filing 990 tax forms – those that have less than $200,000 in annual revenue and less than $500,000 in assets – and for sure those receiving government funding of any kind, are using software tools such as Social Solutions Apricot and Salesforce IoT, with more applications coming online every day. These systems have mostly moved to the cloud, and vendors are helping them make all kinds of connections. This is also happening in philanthropy and has expanded to many parts of Europe as well. Many international organizations working in developing countries are starting to incorporate and connect these systems, at least within their affiliate networks. These systems are the foundation for tying into other large data systems, thanks, in part, to the development of application programming interface software (APIs).

Big data challenges

In terms of the relationship between big data and evaluation, Box 3 details the three main types of big data sources that impact evaluation and research, each of which has its own set of challenges as shown in the list below. That said, it is important to remember that many of the identified challenges refer to only one of the three categories of big data sources.

Big data from apps, platforms, sensors, household surveys, and the Internet of Things (IOT) are criticized by some, such as Matthew Salganik (2018), for sometimes being incomplete, non-representative of the total population, or "dirty," meaning loaded with junk and spam. It becomes difficult to make comparisons over time, as data tend to drift due to changes in the population and in how "data" is defined and measured. However, these criticisms can also be applied to many conventional survey data sets. Overall, it is not clear that big data is any worse than traditional survey data, and supporters would argue big data is on average cleaner. However, users must be aware that, as with any data, a considerable amount of time may be required for cleaning and organizing it. Some kinds of big data can be difficult to access, as they are owned by companies or governments, or they can be too confidential, identifiable, or sensitive to be made available for research. That said, these concerns can also apply to survey data. Finally, data are often processed through algorithms, which are designed for a specific purpose. This means the data that emerges are not "natural" but highly engineered. Because many algorithms are also proprietary, users – and the public – often do not know what variables were included or how decisions were made.

- **Big data from program administrative databases** often has many missing data points, or incorrect scaling of responses due to variations in how providers or implementers score responses to specific questions. In addition, many variables or metrics do not vary enough across the whole population to be valuable, and there is the need to remove identifying or confidential information if it is to be analyzed. As with other big data sources, program data will change over time in response to requests by funders and other stakeholders for new or revised metrics or data points. Additionally, most current administrative and transactional data systems have multiple sets of data tables, which all must be extracted, transformed, connected, and loaded in order to be analytically useful.
- **Big data from satellite images and remote sensors** (geospatial data) can be very powerful for evaluating programs that cover large geographic areas, but using this type of data has the disadvantage of often being relatively expensive and requiring a higher level of technical expertise and computing capacity for analysis.

BOX 4:

Using an integrated data platform to help end modern slavery

The massive and largely unregulated migration of overseas foreign workers (OFWs) from the Philippines leaves many workers vulnerable to exploitation and in conditions that can be considered slavery. Social Impact Consultants, in collaboration with Novametrics LLC, used data analytics modeling to create an integrated data platform that combined available data on the main sectors employing OFWs in different countries, with: i) socio-economic indicators on the characteristics of the vulnerable populations; ii) data on the (limited) social safeguards, such as labor laws and their enforcement, social safety nets, and institutional mechanisms for access to money and land; and iii) risk perceptions and knowledge of OFWs. The analysis of the data platform provided policy makers with a better understanding of the supply, demand, and contextual factors driving OFWs, the areas of greatest concentration of servitude and slavery, and some of the policy options. This project illustrated the great potential of data analytics for addressing complex policy issues, but also showed the significant time and resource investment needed to create the data platform and conduct the analytics.

Source: Bruce, 2017

2.3. The big data ecosystem and links to the evaluation ecosystem

The big data ecosystem

Big data operates in an ecosystem that is completely different from the ecosystem with which evaluators are familiar. This can be seen in data's interlinked subsystems that include data generators, filterers, regulators, users, shapers, and scientists.

- **Data generators.** This comprises data generated through the multiple sources described in Box 2. While as recently as a decade ago, almost all big data was generated by large organizations, increasing amounts of data are now generated by individuals, and small to mid-sized organizations, communities, and groups.
- **Data filterers, brokers, and gatekeepers.** Big data users rarely have access to the original data sources. Rather, they access what has been selected, synthesized, and marketed by app developers, universities, and other data science institutions. While this makes the data easy to use, it is important to understand that the organization doing the synthesis selects what information will be made available and how it will be calculated and interpreted. This means that, in many cases, an intermediary is deciding what information is useful, either for a commercial purpose or to promote a particular political or social objective. Many of the algorithms on which the apps are based are proprietary, so it is difficult to know what criteria were used, for example, in the selection of job candidates, or for consumers requesting mortgages or other forms of credit, students applying to college, or people filing on-line applications for government benefits. There is also an active market of data brokers that can provide information which is intended to be confidential or is not easily available from public records.
- **Data regulators.** Because of its reach and influence, the data science ecosystem is regulated differently from the evaluation ecosystem. Data security and privacy are subject to different levels of regulation in different countries. For example, the European Union often has more stringent regulations than the United States. There is also increasing pressure for stricter regulations on how user data can be accessed, sold, and used. Issues such as political influence, the production or distribution of "false news," excessive portrayal of violence and child pornography, and cyber-bullying and stalking are other areas where there is active debate on appropriate levels of regulation.
- **Data users.** Big data users can be classified into: i) large institutional users, such as governments, development agencies, corporations, and research institutions; and ii) small users, such as civil society, local government agencies, local NGOs, and individuals. These two categories vary significantly in terms of their direct access to all types of big data and in how much influence they have over how data is generated and used. Mass media can be considered users of big data but also producers and disseminators. The risks come from having so many different media able to create "new" without any clear system of accountability.
- **Data scientists.** Data scientists transform and analyze data into user-friendly apps for government agencies, companies, and individuals. They offer a range of approaches to analytics other than the conventional approaches used by evaluators, particularly for prediction. At the same time, there is an increasing awareness, at least among the informed public, of the dangers of excessive reliance on opaque and non-transparent algorithms for making critical choices about who is hired, given a loan, or selected for college. These opaque and problematic algorithms probably represent only a very small proportion of the huge number of decision-making apps, but they have received a great deal of attention. A few have caused significant problems, such as the wrongful elimination of a million people from the welfare rolls of Cincinnati, Ohio (Eubanks, 2017), and they merit discussion (see Section 2.4).

The big data ecosystem is incredibly complex. It involves so many powerful economic and political groups that most of the population have a limited understanding of how the system works and all of its potential benefits and costs. Looking ahead, complex interactions will have to be navigated as the big data world converges with the research and evaluation world and begins to take hold. The ecosystem of the big data world and its data scientists will create both opportunities and challenges when joining forces with the social impact world and its evaluators and researchers.

FIGURE 1: The data science ecosystem and the linkages to the evaluation ecosystem

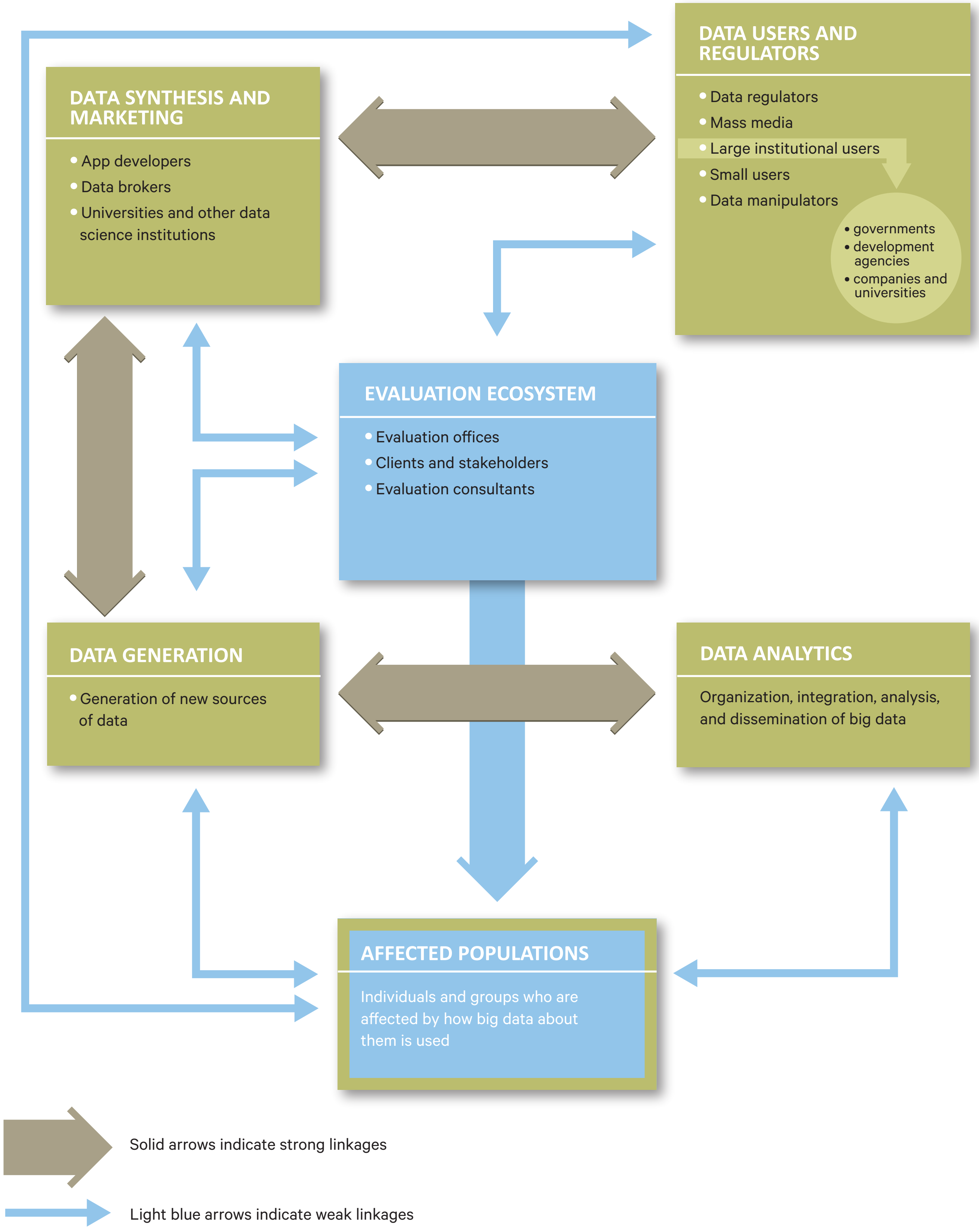
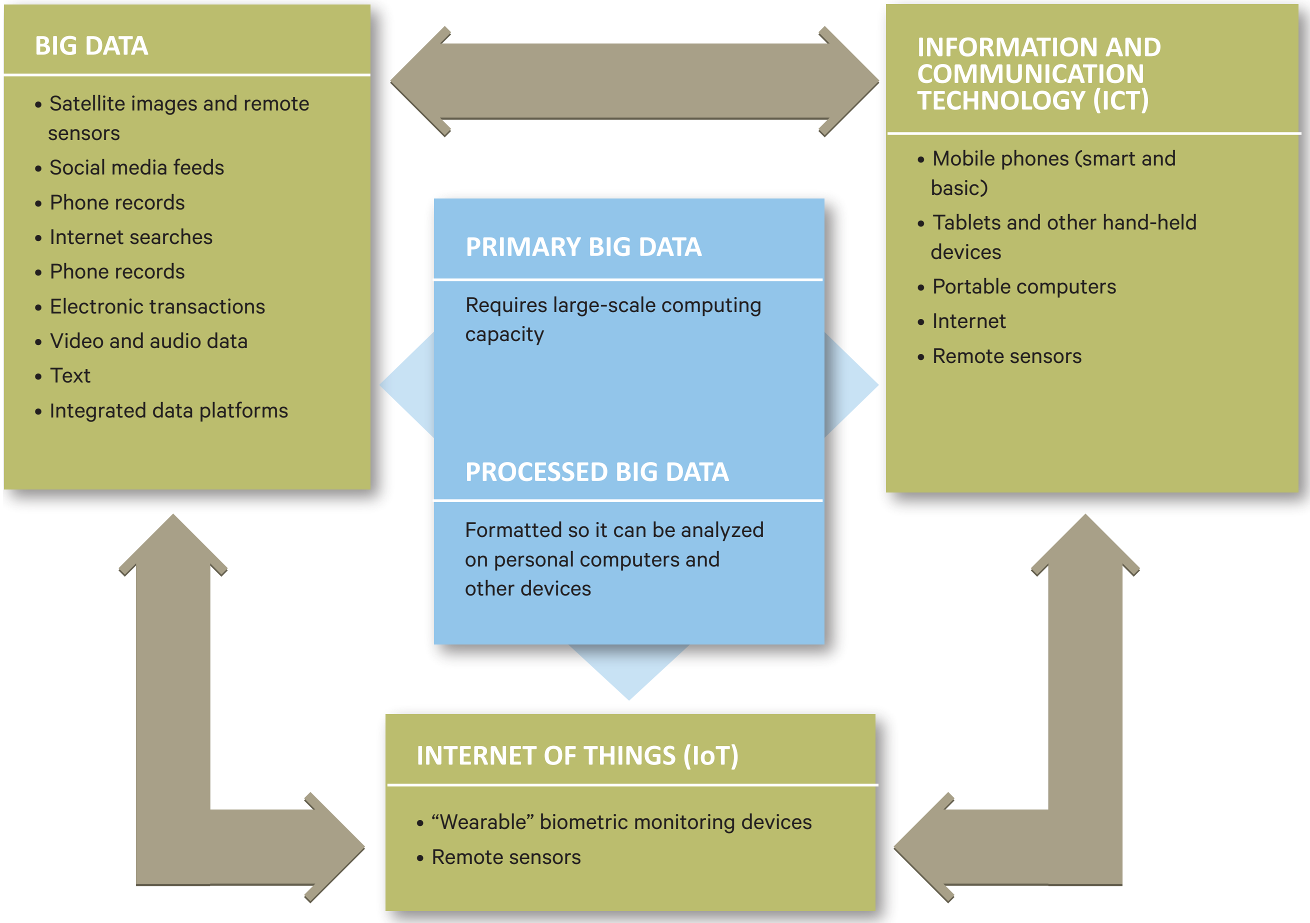

DATA USERS AND REGULATORS
• Data regulators
• Mass media
• Large institutional users
• Small users
• Data manipulators
• governments
• development agencies
• companies and universities
DATA SYNTHESIS AND MARKETING
• App developers
• Data brokers
• Universities and other data science institutions
EVALUATION ECOSYSTEM
• Evaluation offices
• Clients and stakeholders
• Evaluation consultants
DATA GENERATION
• Generation of new sources of data
DATA ANALYTICS
Organization, integration, analysis, and dissemination of big data
AFFECTED POPULATIONS
Individuals and groups who are affected by how big data about them is used
Solid arrows indicate strong linkages
Light blue arrows indicate weak linkages

FIGURE 2: The Components of New Information Technology (NIT)


BIG DATA
• Satellite images and remote sensors
• Social media feeds
• Phone records
• Internet searches
• Phone records
• Electronic transactions
• Video and audio data
• Text
• Integrated data platforms
INFORMATION AND COMMUNICATION TECHNOLOGY (ICT)
• Mobile phones (smart and basic)
• Tablets and other hand-held devices
• Portable computers
• Internet
• Remote sensors
PRIMARY BIG DATA
Requires large-scale computing capacity
PROCESSED BIG DATA
Formatted so it can be analyzed on personal computers and other devices
INTERNET OF THINGS (IoT)
• "Wearable" biometric monitoring devices
• Remote sensors

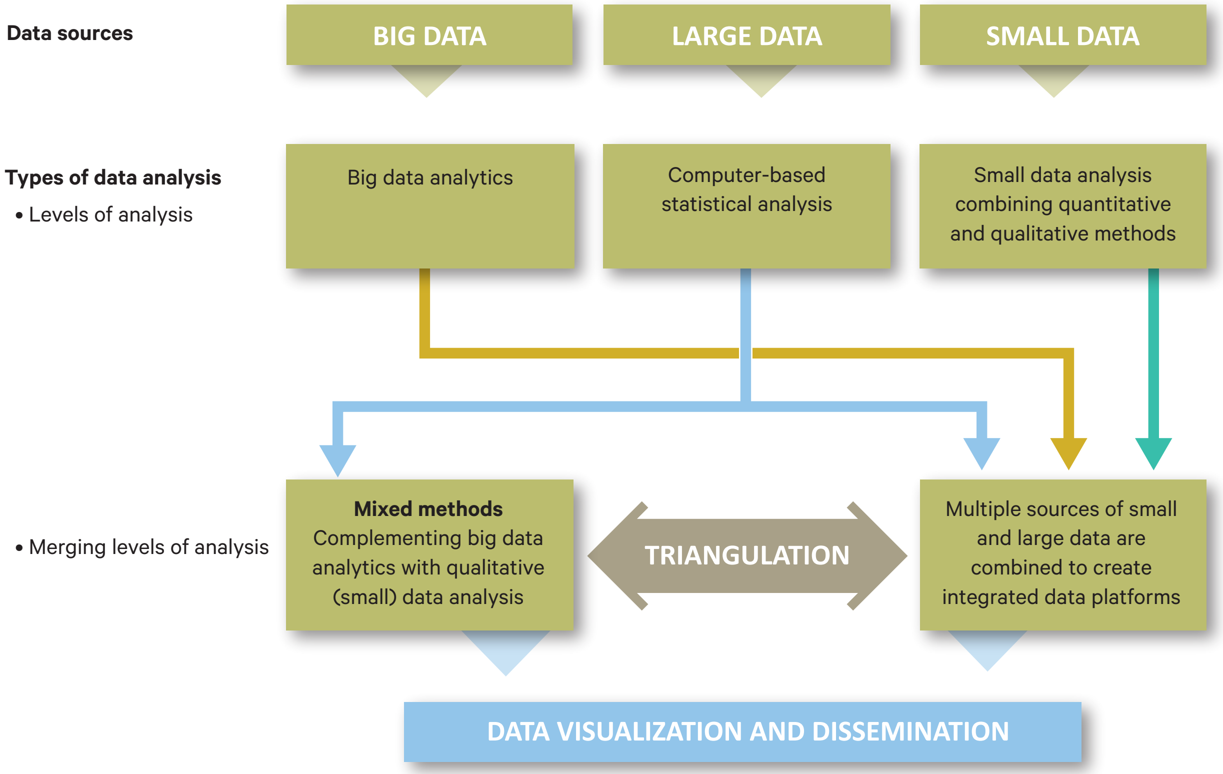
It is helpful to consider big data as one of the three main components of the new information technology (NIT), as illustrated in Figure 2. The other two components, information and communication technology (ICT) which comprises mobile phones and other handheld technology such as laptop computers and tablets, and the rapidly evolving field of the Internet of Things (IOT), which includes "wearable" devices such as smart watches, which are personal monitors and remote sensors. The boundaries between the three categories of NIT are not always clear, but it is quite common that two, or all three of these categories provide data incorporated into data analytics.

In data science, there is a data continuum that can initiate with big data, range through large data, such as census data, large surveys, and government records, and can also include small data such as community surveys, case studies, and in-depth qualitative interviews (see Figure 3). All of these can be incorporated into data science analytics.

The evaluation ecosystem

In presenting a simplified illustration of the evaluation ecosystem and how it relates to the big data ecosystem, Figure 1 uses broken lines to suggest that the linkages between the two ecosystems are not well defined or understood.

FIGURE 3: The data continuum


Data sources
BIG DATA
LARGE DATA
SMALL DATA
Types of data analysis
• Levels of analysis
Big data analytics
Computer-based statistical analysis
Small data analysis combining quantitative and qualitative methods
• Merging levels of analysis
Mixed methods
Complementing big data analytics with qualitative (small) data analysis
TRIANGULATION
Multiple sources of small and large data are combined to create integrated data platforms
DATA VISUALIZATION AND DISSEMINATION

This representation suggests that the evaluation ecosystem comprises: i) evaluation offices, ii) consultants who help design and conduct evaluations, iii) stakeholders, including managers, policy makers and others within the agency where the evaluation department operates, iv) funders, v) government agencies, and vi) civil society. There is also an important relationship – sometimes close and sometimes more distant – with the populations likely to be affected by the outcomes of the evaluation by the process of conducting the studies. Figure 1 also illustrates that the evaluation ecosystem is linked to all elements of the data science ecosystem, and that the relationships are often not very strong or clearly defined. As this paper argues, although there are some areas of convergence, there are often organizational, logistical, economic, and methodological challenges to data science and evaluation working together smoothly.

2.4. Managing big data for social good

The world of data science contains powerful new tools for the collection, analysis, and dissemination of data. Examples throughout this paper, examples throughout this paper indicate the many present and potential benefits of big data for society in general and for improving understanding of the strengths and limitations of global development programs. As with all tools, the outcomes of the application of data science will depend on the purposes for which it is used but, also, how carefully it is used.

The following briefly discusses methodological issues, political and organizational challenges, security and privacy, and what is sometimes called the “dark side” of big data. It is important to remember the distinction between the three types of big data (introduced in Box 3), and that many of the following comments about big data may not necessarily refer to all three categories.

Being aware of – and addressing – methodological challenges

Working with the big data of human-centered apps, platforms and IoT involves a number of methodological challenges. Some are common to any kind of data, but some are fairly unique to the special characteristics of big data. The fact that many data scientists do not have a background in evaluation can mean that issues familiar to evaluators may not always be addressed. The following briefly introduces these issues, which will be further discussed in Section 6.1.

- **Selection bias.** Many kinds of big data are only generated from sub-samples of the population of interest, such as people who have access to mobile phones, who use a particular app, or who discuss a particular issue on this app. The fact that data can be economically and rapidly obtained on millions of people can lead to the "fallacy of large numbers" – which means it is implicitly assumed that access to huge numbers makes issues of bias less serious.
- **Poor quality data.** Data may require extensive cleaning to remove junk and spam, or data may be incomplete or not accurately recorded.
- **Construct validity.** Big data collected for a transactional, documentation or tracking purpose, rather than a research purpose, is very different from evaluation data that was collected to investigate and accurately measure specific key concepts, variables, or constructs. This indicates that big data may serve better as a proxy than as an actual measure where it is assumed, for example, that Twitter references to food or hunger, or reduced ATM withdrawals can be used as proxy indicators of increased poverty.
- **Issues using and interpreting algorithms.** Much processing of big data is done through "black box" machine-learning algorithms. In many cases, these algorithms – or their results – are considered proprietary. This means the user cannot see what variables have been included in the analysis or how they were interpreted. For apps used to automate the selection of candidates for issues that impact their lives, such as jobs, credit, or college entrance, it might be useful to go beyond the conventional algorithm-testing procedures (comparing learning and test groups) and follow the two groups over a longer period of time to compare performance. For example, were there differences in job performance, job history, academic and post-academic performance indicators between the groups selected and those rejected by the apps? A number of recent publications have warned about the serious negative consequences of some widely used apps (O'Neil, 2016; Eubanks, 2017). While those with such consequences are probably a very small minority of the millions of decision-making apps, according to O'Neil (2016) and Eubanks (2017), some of these apps have caused significant harm. At this point, we do not have solid data to assess either the seriousness of these reported problems or how widespread they are, but it is an area where further research is needed.
- **Researcher bias.** There is an ongoing debate among both evaluators and data scientists as to whether researcher bias is incorporated into how data analytics are designed and used. On one side, it is argued that techniques such as artificial intelligence – because of their iterative processes and often very large numbers of cases – are able to detect and eliminate bias. However, on the other side, it is argued that at some point, there is human intervention. For example, in supervised learning models, a human adds what are considered the "correct" answers into the AI program.
- **Confusing correlation and causality.** This is discussed in detail in Section 5.1.

Political and organizational challenges

- **Big data can be used in a participatory or an extractive way.** While big data can be a powerful tool for enhancing the participation of vulnerable groups and local-level organizations in the collection, use, and dissemination of research and evaluation findings, there is a significant risk that big data will be used extractively. In other words, big data can be seen as a set of convenient tools that reduces the cost and time of collecting data for the planning and evaluation of development projects by eliminating the time "wasted" by going to the field.

- **Big data can exacerbate social exclusion by ignoring groups that are difficult to reach or that are not included in the publicly available data.** While machine-learning algorithms can theoretically identify missing groups, it will only do this if the research is based on a theoretical framework that focuses on issues of exclusion. Many data scientists have a much narrower perception of the role of theory in guiding the design of a study. Suthors such as Cathy O'Neil (2016) have documented many widely used selection apps that completely ignored the fact that poor groups or ethnic minorities (often identified through their postal codes) were often excluded for access to loans or jobs. The apps were marketed on their ability to save the client money, rather than as a way to promote social justice.
- **Big data need for resources.** There is a danger that in many circumstances, big data will only be available to the powerful groups – governments, commercial groups, investors, or donor agencies – that can pay for the data or that have the sophisticated technology needed for the analysis.

New ethical challenges

The world of big data introduces new ethical challenges. There have been separate discussions of these issues among policy makers in Europe and the United States (with the EU inclining towards tighter regulation), on the one hand, and among development practitioners and the media on the other. In much of this discussion, it has been implicitly assumed that the challenges are more serious in many developing countries due to less sophisticated information technologies and regulatory frameworks, and certainly some of the issues are more extreme in some non-democratic regimes. However, the increasing evidence of political manipulation of the media in the United States and Europe, the widespread acknowledgement of fake news and the sale – for political or commercial purposes – of confidential consumer information, shows that technical sophistication is no guarantee of higher ethical standards. Policy makers, governments, and funding and implementing agencies are only beginning to understand many of these challenges, as they involve questions about new and unfamiliar information technologies, such as the role of proprietary algorithms in deciding who does and does not have access to a wide range of social and economic services as well as how different groups will be treated by police departments and the justice system. There are also major challenges in understanding and interpreting the findings of complex predictive models with respect to the distribution of benefits. These are critical issues.

There is growing concern about the challenges of ensuring the privacy of big data, much of it including very sensitive and politically dangerous information. There are two levels that need to be addressed with due diligence: i) how organizations collect, protect, and use data on the populations with whom they work; and ii) commercial and political information that can be extracted from the multiple sources of information that are collected on individuals and groups (see Chapter 2 and Box 1). With respect to the first issue, many agencies that conduct surveys and collect sensitive information are not fully aware of the physical and economic dangers, or the social embarrassment that such information can contain, such as revealing that someone is HIV positive in cultures where this can lead to extreme social ostracism. They also are not fully aware of the technical challenge of keeping data secure and private.

Many agencies take minimal levels of precaution, which frequently prove inadequate to protect people's privacy as technology advances (Raftree, in press).

Many of these issues have perhaps been discussed more by international evaluators working in countries with repressive governments or where technological issues of data security may be more challenging. However, current concerns in countries such as the United States about data theft for political as well as financial reasons, as well as the manipulation of news sources, show that these concerns do not just relate to developing countries.

In these cases, the specific form of security and privacy regulations may vary. Even if they have received a high level of attention in Europe, they may not be systematically applied in many developing countries. In fact, there are a number of countries where privacy concerns are completely ignored and where governments have developed software to identify people who, for example, participate in protests.

Finally, there are evolving conversations about the need to develop codes of conduct, to decide what kinds of regulatory mechanisms should be put in place, and to determine what role the evaluation profession should play in formulating and governing the implementation of these regulations.

The dark side of big data

Big data comprises a complex set of tools and technologies that are already demonstrating their huge potential for social good. However, in the wrong hands, the same tools can produce serious social harm. As is already being demonstrated in the ongoing evidence about digital manipulation of political and cultural attitudes and behavior, and the increasing ability to create "fake news," the negative power of big data must be recognized and addressed. Society and politicians are just beginning to recognize and discuss these multiple political, technical, and social issues. The following shares some of the more critical issues.

- **Misuse of social media.** There is accumulating evidence of how the media can be manipulated for commercial or political purposes, or to create and heighten social differences and tension. Social media can intentionally use these methods to raise their prominence and, in turn, increase profits, or it can contribute to deepening cultural divisions by only providing the public with "news" that supports their opinions and biases.
- **The consequences of inaccurate, opaque, or internationally biased algorithms**. Influential publications such as *Weapons of Math Destruction* (O'Neil, 2016) and *Automating inequality* (Eubanks, 2017) illustrate how algorithms that are used with insufficient understanding or control by clients can result in huge numbers of applicants being unfairly eliminated from welfare rolls – such as Cincinnati, Ohio's wrongful elimination of a million people from its welfare program – or people of color and ethnic minorities being disproportionately denied loans, mortgages, or college entrance, or being targeted by the police. The danger is exaggerated by the appeal to managers and budget departments of significant cost-reductions.
- **Faulty data mining modelling and simplistic or biased interpretation of findings.** Many evaluators are critical of the limited use of theory in the design of data mining and artificial intelligence (see Section 5.1). This, combined with the widespread lack of understanding of how to interpret probability estimates, and the limitations of many research projects, can result in faulty or misleading policy applications of data analysis.
- **Excessive digital control and manipulation of the work-force.** The behavior and performance of factory and office workers can be increasingly monitored and manipulated digitally so that, according to critics, workers are starting to be treated and controlled in the same way as robots (*The Economist*, 2018).
- **Extractive potential of big data for research and evaluation.** Big data is often promoted as a way to reduce the cost and time of data collection, and to avoid the headaches and potential political challenges of going to communities to collect data – communities that are sometimes hostile. This

indicates the danger if policy makers are able to make important decisions affecting people's lives, without the need to consult with them and listen to their concerns, or without even having to let them know that decisions are being made that can affect their lives in major ways.

Chapter 3 discusses the increasing use of big data in social and economic programs in both industrialized and developing countries. Developing countries are using all three kinds of big data described in Chapter 2, Box 3, in, for example, their emergency programs, and for disaster relief, program design, management and monitoring, and for information dissemination. Agriculture, health and education are some of the sectors where big data has had most impact. However, a central issue for this paper is the fact that the big data tools and techniques have been used much less extensively in program evaluation than in the other program activities mentioned. Some of the factors affecting the slow take-up by evaluation offices include: the different contexts in which evaluators and data scientists traditionally work; the different methodological frameworks and analytical tools the two professions tend to use; the weak institutional linkages between data science and evaluation offices, even when both offices are in the same organization; and the fact that big data and evaluation training tend not to cover the methodological approaches and philosophies of the other. The slow progress of convergence has potentially serious consequences for both evaluators and data scientists. At the technical level, there are important approaches, tools, and techniques that each could learn from the other. At the broader level, there is a risk that evaluation philosophy, approaches, and methods could become marginalized as organizations start to introduce data-science-driven integrated data management systems – where program performance assessment would be increasingly conducted using data science methods.

3

New frontiers: The increasing application of big data in social and economic development programs in industrial and developing countries

3.1. The rapid adoption of new information technologies in social and economic programs in developing countries

With big data making it possible to advance a vast range of data collection and analysis techniques few could have imagined even a decade ago, development programs are rapidly adopting these technologies in areas such as development research, program design, management and information, and extension. The take-up has been particularly dramatic in the field of emergency relief.

The exponential increase in use of mobile phones, digital banking, satellites, and remote sensors provides examples of how digital technologies are spreading rapidly, even in very poor countries and in remote and inaccessible regions. It had been predicted that the main driver of information technology in developing countries would be the expansion of internet, largely promoted through donor agencies and the public sector. However, in reality, the main impetus has come largely from private sector initiatives, particularly through the spread of mobile phones. Consequently, until now, those in the development community have been the users rather than the drivers of most of these initiatives.

Over the past decade, big data has played an increasingly important role in international development. The use of mobile phones and GPS, social media analyses of, for example, Twitter or Facebook, radio call-in programs, satellite images and remote sensors, internet searches, and early-warning systems are all rapidly. That said there continue to be digital divides between urban and rural areas, and among groups and countries of different economic levels. Development applications include early-warning systems and disaster response (Meier, 2015), program design, management, monitoring, and data dissemination, including data visualization. Agriculture, health, and education are often cited as the development sectors that have made the most progress in the use of big data. On a broader level, the increased ability to quantify and explain the dynamics of poverty is one of the areas where big data can potentially contribute to improving the wellbeing of humanity.

Some of the current applications of big data and analytics in international development, as illustrated in Table 3, indicate that rapid use of big data technologies currently focuses on benefitting the implementation, coordination, and management of programs, strategies, and services, but not the evaluation of these efforts or programs. To put it simply, these big data applications, platforms, and data science techniques focus almost exclusively on building program-support applications.

Further, the data science behind these uses is not yet considering or using big data for evaluation and research, often due to lack of training on and integration of social science methods. Yet, the opposite is true as well – social scientists are not trained in the methods and techniques of data science.

These two data camps – data scientists and social scientists – have been separately trained. Their methods are either unknown to professional camps or unknown to each other, or they are out of alignment with one another. Although this once again infers that there is a need – and it is time – for integration, there are also a number of organizational, economic, and political factors that have so far limited the ability and willingness of many agencies to take advantage of the potential benefits of working closer together. These other factors are further discussed in Chapter 5.

An important consequence of big data for development is that more data is becoming available on difficult-to-access populations. For example, a recent census conducted in the Islamic Republic of Afghanistan

TABLE 3: Widely used big data/data analytics applications in international development

APPLICATION	BIG DATA/DATA ANALYTICS TOOLS
Early warning systems for natural and human-made disasters	Analysis of Twitter, Facebook, and other social media Analysis of radio call-in programs Satellite images and remote sensors Electronic transaction records, e.g. ATMs, on-line purchases
Emergency relief	GPS mapping and tracking Crowd-sourcing Satellite images
Dissemination of information on questions such as health, agricultural or marketing information to mothers, small farmers, fishermen and other traders	Mobile phones Internet
Feedback from marginal and vulnerable groups and on sensitive topics	Crowd-sourcing Secure hand-held devices, e.g. UNICEF's U-Report device
Rapid analysis of poverty and identification of low-income groups	Analysis of phone records Social media analysis Satellite images, e.g. using thatched roofs as a proxy indicator of low-income households Electronic transaction records
Creation of an integrated database synthesizing all the multiple sources of data on a development topic	Integration of data from multiple sources on questions such as: national water resources human trafficking agricultural conditions in a particular region

combined an on-going demographic survey, satellite imagery and other remote sensing data, and urban data with geographic information system (GIS) statistical modelling. It used data analytics to integrate the different data sources into a common platform, which was then used to generate information on the country's population (UNFPA, 2016, cited in Bamberger, 2017).

3.2. Implications for how future development programs will be evaluated

It is likely that many of the data sources used for program monitoring and evaluation (M&E) soon will be generated or synthesized through new information technologies, rather than collected through the stand-alone M&E studies commonly used today. These are often referred to as transactional data sets. While each individual data set – comprising monitoring data, sample surveys, and management reports – would be considered large (or even small) on its own, when factored into an integrated database, the computational possibilities will often require greater computing capacity. Future M&E systems are likely to be closely linked to new types of management information systems that integrate program identification, design, management, and M&E into a single system. For development evaluation, this implies that program evaluation may gradually become one of several outputs of the integrated management information system, rather than being a separate function that collects and analyzes specially generated data from a range of quantitative and qualitative, and primary and secondary data. Under this latter scenario, many evaluations will be based on data not collected specifically for evaluation, and they will often be designed and analyzed by data scientists rather than by conventional evaluators.

While some of the data analysts may be familiar with conventional evaluation methods, many will not. Further, many of the evaluations will use methods such as integration of multiple data sources, dashboards, data-mining, predictive analytics, and machine learning. There are opportunities here for conventional evaluation to play a lead role in designing the systems for data collection, synthesis, and analysis.

3.3. The slower take-up of big data by many evaluators

Both industrialized and developing countries have seen rapid expansion of information technology in the planning and management of social and economic programs. However, less progress has been made in the application of these new tools in the evaluation of their programs. Interviews with development agencies conducted in 2016 – as part of a study on the integration of big data into the monitoring and evaluation of development programs (Bamberger, 2017) – found that in many agencies, institutional linkages between the newly created data development centers and the evaluation offices were not well established, and evaluation offices often were not actively involved in the work of these centers. The study reported:

> *"The interviews conducted during the preparation of this report revealed that one of the challenges to integrate big data into development M&E is that most data scientists operate within a very different research and management paradigm than do most evaluators. Both groups use different terminology and have a very different approach to questions such as: how to evaluate development interventions, the nature of data and how to assess the quality and utility of different types of data, approaches to data analysis and the role of theory. The limited familiarity of each with the approaches used by the other sometimes means that questions that could easily be clarified on the basis of a discussion can become quite confrontational or misunderstood"* (Bamberger, 2017:34).

The findings of this report were consistent with an internet survey of LinkedIn groups affiliated with the American Evaluation Association (AEA) and the European Evaluation Society (EES), which were assumed to represent evaluation practitioners. The study determined that only about half of the 324 respondents conceptually understood big data, and only about 10 percent claimed professional experience working with big data (Hojlund *et al.*, 2017).

However, there are a number of entities, including UN agencies, universities, and consulting firms, that are actively working on these questions. Given the speed with which new technologies are developed, many of these new initiatives are likely to come to fruition within the next five years.

As Josh Blumenstock, Director of the Data Science and Analytics Lab at the University of Washington observed in an interview with the on-line journal, Devex:

> *"At least on the academic side, the two communities have different traditions, and generally approach problems differently. I don't think this is necessarily an insurmountable problem but it does mean that that you cannot just put a social scientist and a data scientist in a room and assume that magic will ensue. There are some very real obstacles that stand in the way of collaborations happening at scale, but also some value that each side can bring to the table. As long as they can learn to get along"* (Devex, 2016).

In the same article, Emmanuel Letouzé, Director of the Data Pop Alliance stated:

> *"What statisticians, demographers and economists need to realize is that data science is not just a fad, and what computer scientists and engineers need to acknowledge is that they cannot solve global poverty by crunching numbers alone"* (Devex, 2016).

All of the evidence, from both the increasingly data-driven industrial economies and the beginning of similar trends in developing countries, suggests that big data will come to play an increasingly important role in the design and implementation of development programs. Letouzé estimated that "Big Data will affect development work somewhere between significantly and radically" (UN Global Pulse, 2012).

This indicates the likelihood that development programs will increasingly use the new data analytics platforms to create integrated databases. These will be databases that combine multiple sources of big data including large data, such as large household surveys with tens of thousands of cases, and small data, such as surveys with only a few hundred cases, and, in turn, use them for project and program selection, design, management, monitoring, evaluation, interpretation, and dissemination. Much of the data generated from big data sources that will be used in evaluations will be data originally generated for other purposes, such as phone messages, call-in radio stations, digital financial payments, social media, and satellite images.

Many decisions on evaluation designs and the data to use in the evaluations will be made by data scientists, not by professional evaluators. Further, many of these evaluation designs will use approaches currently unfamiliar to most evaluators, such as data mining, predictive analytics based on Bayesian probability models (see Glossary), and machine learning. The question for the evaluation profession is whether evaluators will become sufficiently familiar with the logic – as well as the strengths and limitations – of big data approaches. Doing so will enable them to contribute to the formulation of the methodologies used to evaluate future development programs.

For now, it seems likely that data science will play an increasingly important role in how program performance is assessed. The challenge for evaluators will be to adapt to, and incorporate, these new approaches into their evaluation practice. There is debate as to how quickly evaluators will adapt to these new approaches. The optimistic scenario is that data science and evaluation will converge so that program evaluation will draw upon a greatly expanded tool kit. A more pessimistic scenario is that many evaluators may be slow to adapt, and evaluation as a profession may lose ground to data science. This is an important area for research.

Factors contributing to the slow take-up of big data by development evaluators

The following looks at a number of factors contributing to the slow adoption of big data by many evaluators.

Evaluators and data scientists work in different contexts. Many quantitative evaluation designs tend to make evaluators conservative in their approach – they rigidly control the metrics and experimental conditions to ensure findings are valid, reliable, and generalizable. Evaluators invest considerable time and resources in developing and testing sampling frames and data collection instruments that will be replicated over time in pre- and post-test comparison designs. The logic of these designs requires applying the same data collection instruments and the same sampling frame at two or more points in time, such as pre- and post-test. Evaluators make efforts to ensure data quality and reliability, and to avoid selection bias. So inevitably, they are resistant to changing indicators and data collection methods, and sample selection procedures. In contrast, big data is constantly changing and will often have issues of selection bias, data quality, and construct validity. At the same time, evaluators often believe that data scientists do not take these issues seriously, claiming that issues of data quality and bias are not important, as new data will soon become available and large sample sizes can compensate for bias.

Further, most development programs have a much longer decision-making cycle than is often the case for the situations in which data analysts typically work. For example, programs that provide infrastructure – such as houses, water supply, or sanitation services – often have construction cycles of at least six months and up to several years. This means that only limited use can be made of many kinds of real-time data, as even agile programs cannot make short-term adjustments.

Evaluators and data scientists use different frameworks and analytical tools. Evaluators and data scientists tend to work with different research frameworks, and data collection and analysis tools. This means that many evaluators and data scientists have limited understanding of each other's approaches.

- **Real-time data.** Data analytics makes extensive use of real-time data while evaluators are more familiar with data generated from surveys, project records, and official statistics. The two groups also have different approaches to issues of bias, data quality, and construct validity.
- **Theory.** While most (but not all) evaluators use a theory-based approach to evaluation design, the role of theory is less clear for many data scientists. Some argue that the high speed data mining and iterative correlation analysis of very large data sets eliminates the need for theory, while others argue that any kind of data mining must be based on an implicit theory. However, the perceived lack of a theoretical framework that could result in reliance on possibly spurious correlations is one of the criticisms that evaluators often level at data scientists.
- **Attribution, correlation, and causality.** Evaluation approaches, including experimental and quasi-experimental, are designed to assess the causal relationships between project design and outcomes by controlling for other factors that could explain the outcomes. Evaluators argue that correlation-based methods cannot explain causality, which seriously limits their practical utility for policy makers and program managers. However, data analysts argue that, with sufficiently large data sets covering a much wider range of variables, constant updating of the data, and techniques such as predictive analytics, it is possible to provide operationally useful guidance for identifying groups that are most at risk and likely to respond to different kinds of intervention. This approach has the advantage of providing real-time guidance on the current situation. In contrast, experimental designs seek to explain causality related to the point in time when the project began. Yet, in a rapidly changing world, the operational utility of this historical data may be quite limited. At present, much of the debate

continues to focus on experimental designs versus predictive analytics, but there is clearly a need to find ways to combine and take advantage of the learnings of both approaches.

- **Samples.** With big data analytics, it is often possible to work with the total population, while most evaluation designs usually work with relatively small samples, due to the costs and time involved in data collection. However, evaluators argue that claims of working with total population data can be misleading, as it is often difficult to ensure complete population coverage, which means the issue of selection bias must be addressed. Also, survey researchers spend considerable time cleaning data and trying to ensure a high level of data quality. As this kind of data quality assurance is usually not possible with big data, the quality of the big data may be questionable in many cases. Also, as much big data is generated through proprietary algorithms, it often is not possible to check how data findings were generated.
- **Time period.** Many sources of big data have the potential to cover much longer time periods than conventional evaluations, often including periods before a program began, and potentially continuing after its completion. For example, satellite images of forest cover, infrastructure, and human settlements may cover periods up to 20 years, and the period covered by social media feeds is increasing year by year. This time dimension is particularly valuable for assessing project sustainability as most conventional evaluations do not have the capacity to continue collecting data after a project ends. Some authors, such as Mylynn Felt (2016), have suggested that many social media analyses may focus intensively on a short period of time, because data may be available.

Big data centers and evaluation offices have weak institutional linkages. Many development agencies have begun to establish data development centers. However, these centers often do not work closely with the evaluation office, and institutional linkages between the two can be quite weak. The data centers tend to be staffed by professionals with a background in data science or mathematics, but little training in evaluation methods. In some cases, the evaluation office is not familiar with the work of the data center, and management does not see the need to encourage the two offices to work together more closely.

In looking at the disconnect between data sciences and evaluation offices, Forss and Noren (2017) conducted a review of the terms of reference (TORs) for 25 evaluations commissioned by donor agencies, including UNDP, Sida, UNICEF, DFID, Norad, Danida, and the Millennium Challenge Corporation. They found that none of these TORs included a requirement for the evaluators to consider the possibility of using big data.

Evaluation capacity development does not normally include an introduction to big data and vice versa for data science. Training programs for evaluators rarely include an introduction to big data, and the tools and techniques of data analytics. For example, 2017 was the first year that the American Evaluation Association (AEA) offered an introduction to big data in professional development workshops. Equally, data scientists can have a number of different professional backgrounds, but they rarely have training in program evaluation.

Slower take-up of new information technology by evaluators has potential consequences. As data science comes to play an increasingly important role in development programs and their evaluation, there will be consequences if evaluators do not have a seat at the table when the data science systems are being developed. For example, evaluators and evaluation offices may become marginalized, meaning that the important methodological lessons that evaluation has learned may not be taken into consideration.

On the other hand, data science approaches, many of which were not developed in social service delivery contexts, will not benefit from the extensive on-the-ground experience of evaluators. Some of the important areas of evaluation experience that might not be built into these new management and program performance assessment systems might include:

- consulting with stakeholders to identify the key evaluation questions and ensuring that the evaluation is designed to address these questions
- building the evaluation on a theoretical framework to ensure the right questions are asked and to strengthen the validity of the interpretation of findings
- ensuring that questions are asked to the right people who can provide the information required
- recognizing that no single evaluation design can address all of the important questions
- being aware of selection bias and potential exclusion of vulnerable and difficult-to-reach groups, and taking measures to include all of these groups
- applying mixed methods and triangulation to correct for bias and strengthen the interpretation of findings.

Chapter 4 begins by illustrating some of the big data tools that can help strengthen evaluations by reducing the time, cost, and effort of data collection, sample design, and data analysis. Big data makes it possible to collect a vastly increased range and volume of data more easily, quickly, and economically. The ability of big data to include an entire population with a particular attribute – rather than just a relatively small sample – makes it possible to avoid many kinds of selection bias, and to disaggregate the sample to cover many different sub-samples and categories. It enables data analysis to be conducted more rapidly and cheaply, and also makes it possible to conduct a much more sophisticated and complex analysis. Machine learning has emerged as one of the most important developments, as it brings a completely new focus to hypothesis development and the integrated analysis by combining multiple kinds of quantitative and qualitative data in ways which were not previously possible. A case study of a large child welfare program illustrates how machine learning, predictive analytics, and other big data techniques can be applied in the real world. The section ends with a discussion of some of the conditions necessary for integration of data science and evaluation to take place.

4

Advances in the application of data science for impact evaluation

4.1. How tools and methods of big data can strengthen evaluation

Evaluations require a significant length of time, expertise, and labor to plan, conduct, and control. Most well-developed evaluations take months to years to implement through to the point of generating conclusive findings. Because of the time and expertise required, especially when it comes to the labor of data collection, transformation, and analysis, the cost of evaluation is quite high relative to the program effort, and the cost is even higher if using a mixed-methods approach that relies on both quantitative and qualitative data. Because of the need to test theories and their hypotheses, the data collection instruments must be constructed expertly to ensure they are both valid and reliable. As such, secondary datasets don't easily meet the measurement standards of good evaluation. Additionally, sampling plans must be carefully researched and planned to ensure that the study subjects are in high enough numbers to be representative of the target population. That said, most evaluation studies can't afford the size and scale needed to sample beyond a very narrow population within a specific community or neighborhood. As a result, it is often very difficult to "generalize" findings to other communities, settings, or locations, so that a program found to be effective can, in fact, be replicated and produce the same results elsewhere.

Data analysis is also slow, laborious, and challenging, with quantitative data requiring a high level of effort to clean, prune, and prepare it for the very rigid requirements necessary for accurate statistical analysis. Much data is left on the proverbial chopping block. Qualitative data preparation and analysis is similarly time consuming, as narratives must be cleaned, sorted, tagged, coded, and analyzed to be useful. Integrating quantitative and qualitative conclusions into a comprehensive finding report requires a lot of closing effort. Many of these challenges, once considered necessary for ensuring rigorous and generalizable conclusions, can now be mitigated or even eliminated by the advances of the science behind the big data era.

Table 4 summarizes challenges to current evaluation methods that big data can potentially help resolve. Some of these are discussed in the following pages.

TABLE 4. Challenges to current approaches to the design, implementation, and analysis of development evaluations where data science can potentially

1. EVALUATION DESIGN CHALLENGES AND UNDERSTANDING THE PROGRAM CONTEXT
Strengthening the counterfactual
Evaluating complex programs and capturing contextual data
Identifying unintended outcomes
2. DATA COLLECTION CHALLENGES
Working with the high cost and time required for data collection
Collecting data on difficult to reach groups
Monitoring project implementation and processes of behavioral change
Collecting qualitative data
Collecting and integrating different sources of data
Enhancing quality control of data collection and analysis, and reducing the time and cost for these controls
Collecting information on the spatial dimensions of programs
Recognizing the need to solve sample design challenges
3. DATA ANALYSIS AND DISSEMINATION CHALLENGES
Analyzing data is time-consuming and expensive
Analyzing large data sets with standard computers is difficult
Disseminating findings is time-consuming and expensive, and often only reaches the main stakeholders
Disseminating findings in a long, written report is typical, but it not user-friendly for many groups

Tools of big data – a salve for the pain of collecting data

One of the greatest boons of the tools and methods of big data is reducing the high cost and increasing the speed of collecting data. Data can now be collected anywhere at any time from anyone.

- **Data collection.** Tracking and monitoring technologies undergird all types of consumer applications, enabling feedback between consumers and companies which can make or break a business. Evaluators now can access free, low-cost or open-source data collection tools that use the web, mobile phones, and even new tools such as chatbots to speed up the process of gathering information.
- **Data-storage.** Transactional data collection has proliferated through the advancement of data-storage technologies, specifically cloud-based database managers, such as Amazon Web Services (AWS). Which no longer require users to purchase and maintain one or more servers to hold a lot of data.
- **Secondary data collection.** Researchers and evaluators, particularly in the fields of economics, healthcare, and political science, have long tapped into, for example, the datasets of national household studies conducted by governments, electronic health records, and voting datasets.

Now, with open-source data movements spurred on by the U.S. Freedom of Information Act in combination with more robust web- and cloud-based portals for these datasets, and with increasingly connected administrative datasets, the collection of secondary data is growing exponentially.

- **Objective data.** What many evaluators and researchers refer to as "objective" data, particularly behavioral and environmental data, is now available through the use of GPS, satellite, and the Internet of Things. While these data may initially appear messy to many evaluators and researchers, they ameliorate the need to develop complex measures and scales to serve as proxies for behaviors. Rather than designing reliable and valid questions and response patterns for people to self-report behaviors, it is now possible to collect actual behavioral data.
- **Mobile phone application.** Mobile phones have made it possible to collect information, actively or passively, from anyone almost anywhere at any time. Smart phones make it possible to build applications that can notify individuals at specific points in time or at locations, gathering information passively as a part of quid pro quo service or even actively as a part of a study. Service delivery applications used on mobile phones, such as case management and distribution of resources (food, water, medicine, or social service and benefit programs), are providing much more real time and accurate data on dosage, frequency, and quality than previous efforts which required entering data as an administrative task at the end of a day. Even "dumb" mobile phones can be used to gather movement data and to send people quick surveys for real-time response, minimizing the need for the traditional door-knocking survey research approaches, as long as there are not too many questions.
- **Data history access.** All individuals have access to their complete data history on almost all current social network platforms, such as Facebook, Twitter and LinkedIn, as well as most other web-based/applications, such as Google search histories and Amazon purchase histories. Individuals can download their individual data, which allows evaluators to further enhance their datasets, both qualitatively and quantitatively.

Big data will help alleviate the pain of sampling plans that often go awry

Evaluators need to develop a sampling plan that ensures the target population is accurately represented across subjects chosen for a study. In theory, this is a relatively straightforward and statistical procedure that ideally calls for randomly sampling enough cases from among all program participants. However, when the realities of an ideal sampling plan meet the real-world time, place, and interpersonal constraints of the program setting, problems arise, sampling errors rise, and generalizability of findings diminish. More specifically, there are many reasons why a randomly selected individual cannot participate when and how the sampling plan require. Often, it is an issue of not being able to be in the right place at the right time.

Evaluators have many hacks, such as incentives, which they can deploy to remedy this situation. However, there are always those selected who cannot participate or participate fully, which results in the loss of data and diminishes trust that findings are generalizable. Often, the reasons for not being able to participate in the study – such as when some members of the population do not have transportation to get to the program – are the same reasons that a population needs to be represented in the dataset – because, in fact, the program focuses on helping poor people. If poor people cannot pay for or access the transportation needed to attend a program, then the choice of people who participate in the experimental program are unintentionally being selected in a way that does not represent those who need it most. After all, those without transportation – often the true poor – cannot attend, meaning the "less poor" are the ones selected for the program.

This is what is meant by the term "selection bias." It is one of the most invasive and difficult to control biases in the real world of program evaluations – which are always in the experimental context as opposed to a laboratory setting. Without the removal of selection bias, we cannot generalize findings to another time or place.

The era of big data and, specifically, technology-enhanced data collection – through mobile and web-based applications and platforms that people use for productivity and pleasure – is removing the time and place barriers to gathering information from all voices. The time and location barriers that prevent specific and, therefore, potentially unrepresented persons from providing their data are ameliorated when the vehicle for all communication is in their pocket, and data are collected via an application they carry with them that is required as a part of service delivery. As more support applications are developed,such as for case management, service delivery, communication, and coordination, and then integrated with strategy and program implementation, there will be fewer data collection barriers, because data collection will be an add-on and more a part of the experience. In essence, the tools of big data collection are making the ideal sampling plan something that is feasible without much compromise. As an obvious result, this means generalizability – the important goal of a strong sampling plan – is now more scalable.

Data-shaping tools of big data reduce the time it takes to get to analysis and results

There are now many open-source (free) and for-pay software programs and platforms that allow those who are comfortable working with data to extract, clean, transform, and load data into their final analysis tools or process. In their early development, these tools required programming skills in computer languages such as R and Python, but this is no longer the case. These tools, such as PowerBI, Talend, and Knime, now allow for anyone comfortable with data to conduct all of the data shaping needed to get to analysis, and also to do so efficiently, in a way that concomitantly builds a "recipe" that can be automated for deployment with all new incoming data. Gone are the days when evaluators and researchers would have to redo their data cleaning and shaping efforts every time new data arrived. In fact, evaluators won't have to wait until enough data arrive to make data cleaning and data shaping worth the time and effort – the data will be prepped using these automated "recipes" as soon as each new case enters the database.

Machine-learning algorithms alleviate the pain of messy, disconnected, unusable data

Machine-learning algorithms, the widespread tools of the data science world, are being used in the private sector to make all types of data useful for analysis. This is not yet the case for the evaluation community, but it is coming.

Evaluators who gather and analyze both quantitative and qualitative data must prepare the respective datasets separately, following strict guidelines and rules.

Quantitative data. Due to the use of mathematically based statistical modeling requirements, quantitative data must meet specific statistical standards, such as distribution assumptions, to be analyzed. As such, when quantitative data that do not meet these requirements are collected, the data either need to be removed or manipulated to make it conform.

Qualitative data. Qualitative data must be prepared separately, and there are many discretionary techniques and methods for doing so.

There are many software programs to help with quantitative and qualitative data preparation, but these tools still require a lot of labor and expert judgment. There is as much of an art to the process as there is a science.

Evaluators conduct these data manipulation acrobatics to find associative and causative patterns in the data. Using the tools of statistics for quantitative data pattern-finding and human coding for qualitative data

BOX 5:

Practical examples: How big data, analytics, and social science are converging: Evaluating the effectiveness of a child welfare system in Broward County, Florida

In a recently published study on the use of algorithms for building predictive and prescriptive models that can determine the best course of action for an incident of child abuse or neglect, social scientists and data scientists applied a joint, rigorous evaluation method using machine learning to draw conclusions (Schwartz, *et al.*, 2017).

This project began by combining the administrative datasets from the Broward County child welfare system, including its child abuse and neglect hotline and investigation data, as well as the administrative datasets of a number of community-based service providers that served children and their families who were involved in the system. The final evaluation data set combined over 80 data sets and more than 80,000 cases spanning a five-year period. The data included all types of intake, historical, contextual, transactional, assessment – including validated assessment instruments – and outcome data, of which a subsequent incident of child abuse or neglect was the target/dependent variable (the outcome to be predicted and evaluated).

The researchers on the project included social scientists with subject matter expertise and data scientists with machine-learning expertise who collaborated to build and rigorously test a theory of what works to prevent a subsequent incident of abuse or neglect.

The first step in the project was to build a theory of what works, using a combination of subject matter expertise in child abuse and neglect, the body of prior research and evaluation evidence, as well as machine learning algorithms. The cross-discipline social and data science team conducted a literature review to develop an evidence-based understanding of the variables that matter and how, theoretically, they interact to affect a positive or negative outcome. Concomitantly, the first step also included the use of expert-guided machine learning algorithms to develop a "data-driven" theory of what works. This theory-building process was an iterative process whereby experts determined the variables to put into the algorithms, viewed the results, assessed the level of alignment of these results with prior research (the historical evidence base), and adjusted and reran the algorithms, until achieving both the highest level of analytic accuracy and alignment with prior rigorous research.

The next step was to train machine-learning algorithms to find naturally occurring experiments in the historical dataset, whereby matched cases on background, context, and prior history of child abuse and neglect were "randomly" treated differently by system actors, such as investigators, judges, and providers. As with the theory-building step above, this process of finding the naturally occurring random experiments was iterative, guided and shaped by subject matter experts, and supported by machine-learning algorithms. This causal analytic process – using observational data studies to find counterfactuals of different treatments for similar cases – has been used in, for example, the fields of medicine, health, education, and economics for tens of thousands of studies, although only recently using machine-learning algorithms. This study was able find a number of matched comparison groups within the large dataset of cases, and then identify the differences in treatment that matched/like-cases received. For example, there were similar "low-risk" cases who were or were not removed from the home at the time of the first incident, providing counterfactuals that were then studied to determine if it was best to pull a specific "type" of case from the home, or not. Since these cases had the same histories and levels of incident severity at their time of intake, they theoretically should have received the same treatment, but in spite of these matching characteristics, system actors, such as judges, sometimes removed a child and sometimes did not. Herein is the naturally occurring "random" experiment.

This example shows how this technique can help identify counterfactuals and thereby evaluate cause

(cont.)

BOX 5: *(cont.)*

Practical examples: How big data, analytics, and social science are converging

and effect by removing selection bias. That said, it is important to note that matches can only be found based on metrics/variables that were in the dataset. There may be exogenous reasons not captured in the dataset motivating a judge to remove a low-risk case from the home. This is why this causal modeling technique should be considered "quasi-experimental" and not a true "experimental" study. However, due to the size of the data and the number of variables in these big administrative datasets, and given that the data are derived from ongoing real-world transactions, this method's feasibility and the generalizability level of its findings are arguably more useful, scalable, and adaptable than the findings from highly controlled randomized experiments. And, if subject matter experts continue to guide the analytics process, they can test new hypotheses by adding new variables into the rapidly growing administrative dataset, measuring any exogenous or other possible explanations, thereby providing checks and balances for potential threats to validity.

The final step reflected in this study was to evaluate each of the matched comparison groups, as well as the whole system, to determine attributable success – i.e. how many "like" (similar) cases received what worked best for their given history and incident, and for preventing the next incident of abuse and/or neglect. This study was able to provide the Broward County Sheriff's Office with an accurate number and proportion of cases for which their treatment choices attributably achieved a significantly lower likelihood for a repeat incident. This study also showed how many treatment decisions had, in fact, led to significantly worse outcomes, an often overlooked result of social programs. For example, the resulting model from the Broward County study determined that at least 40 percent of the cases that were pulled from the home did not contain hot line or investigative data that warranted the decision. As a result, these inappropriate referrals to the court were 30 percent more likely to return to the system.

This data-driven convergent approach is leading to lower-cost evaluations that can be automated to provide more timely, rigorous results to program leaders, while also producing on-demand predictive and prescriptive insights to front line staff. Once evaluation models have been developed using the above machine-learning techniques, a repeatable algorithm and code gets produced that can be attached to an existing administrative data system in order to streamline and automate ongoing evaluation, while also generating predictive and prescriptive insights for front line staff.

For example, an article by Gay and York (2018) published by the Scattergood Foundation, shared how a residential treatment center for children with severe mental health problems engaged in an analytic project similar to the Broward County example. The center leveraged its administrative data and subject-matter expertise, in collaboration with data science experts, to build evaluation, prediction and treatment prescription models to reduce the likelihood that children receiving residential treatment would be re-hospitalized subsequent to their discharge. It was able to generate an accurate evaluation of their attributable success with each of eight matched comparison groups of children they serve, as well as evaluate the overall success of their services across all children.

The center is now refining and attaching its evaluation algorithm to its administrative data system, including adding new questions to test unaddressed hypotheses. When this algorithm is automated alongside the data system, it will provide predictive and prescriptive insights and program planning (prediction and prescription) and monitoring tools for individual providers working with each child and his or her family. The cost of the original evaluation modeling process was affordable to this independent non-profit organization, and the ongoing cost will be even less because the evaluation algorithm can be inserted and automated, including updating the findings and what works, within the organization's existing administrative data system – in this case, the organization's electronic health record data system.

pattern-finding is laborious. Additionally, all of these manipulations require human judgments during the data-shaping process, which introduces the high likelihood of making biased decisions that will lead to untrustworthy or erroneous conclusions.

Machine-learning algorithms, the pattern finders of the big data world, ameliorate much of the data-transformation pain suffered in the evaluation and research community. Unlike statistical modeling methods for quantitative data, machine-learning algorithms do not have the same mathematical requirements for identifying patterns. Statistics look for patterns by aggregating data and running mathematical summaries, frequencies, and associations based on these aggregations. This aggregation approach requires that data not be skewed or unevenly distributed, or else the statistical conclusions cannot be trusted. This process requires a lot of human judgment as to how much unevenness is tolerable – deferential position evaluators are trained that it is best to remove data close to the line of being skewed or uneven. Machine-learning algorithms are not so inflexible and, therefore, do not require losing so much valuable data. Because of machine learning's computational power and its ability to hold and analyze every possible combination of data elements together in its memory at one time, algorithms can find all the best associations, or patterns, without the need for rigid statistical rules. Machine-learning algorithms are like a super-powered human brain that has studied every member of a population, has a photographic memory, and has Spock-like expertise in all forms of objective logic and analysis. Because of these powers, so to speak, machine-learning algorithms can use data that are skewed, unevenly distributed, categorical, true/false, or ranked.

In addition to machine-learning algorithms allowing users to consume and learn from all types of numeric and ordered data, they are getting better and better at drawing concomitant learning from all types of qualitative data. Advances in natural language processing and tools, such as sentiment analysis, are making it possible to combine numeric, categorical, ranked, and narrative data into one analytic or "modeling" process. These tools will require human collaboration, especially for using algorithms to appropriately tag different types of language and messages. Once these algorithms begin to identify patterns in the tagging process, they can continue to refine the tagging on their own, making them more and more accurate at finding patterns among the qualitative data, as well as combining this data with all the other data to identify more refined and precise patterns.

The reality is here – machine-learning algorithms are able to transform the evaluation community's view of qualitative and quantitative data transformation and analytic methods as distinct from one another and require further judgment to draw mixed-methods conclusions. Because of machine-learning, datasets and analytic mixed methods are converging, becoming cleaner, faster, cheaper, and smarter (see Box 6 for practical examples that demystify machine learning).

4.2. Necessary conditions for integrating data science and evaluation

The use of big data and the tools and methods behind it are slow to be adopted by the evaluation community. There are a number of possible reasons for the slow adoption.

- **Lack of education.** The primary reason the evaluation community has not yet embraced the tools of the big data world is its lack of introduction to and education about them. The tools of big data sit within the computer engineering or data science departments of colleges and universities, where education focuses on teaching computer programming skills, statistical and machine-learning analytics for prediction, and data architecture development and management. These departments focus on producing products and solutions rather than knowledge. However, when it comes to the tools of the "data science" that is being conducted, including how to collect data, process data, prepare data, analyze data, and make predictions, there are techniques and methods that would benefit social scientists and evaluators. These tools, techniques, and methods are also being built for wide-scale use and adoption, without having to learn raw programming. With more education, the social science community could not only take advantage of these tools and methods, it could also inform their further development and refinement – such that they could not only serve product development, they could adopt the use of data for research and evaluation. Universities, such as the University of Chicago, are starting to develop centers that integrate the faculty and students of these respective departments.

- **Philosophical differences.** There is a fundamental underlying philosophical difference that will need to be bridged if a integration is to occur: social scientists and data scientists are implicitly at odds as to the purpose of data. Think about the question: "is data to be used to make probabilistic case-by-case decisions and recommendations, or to rigorously, and with certainty, study and evaluate what works?" Data scientists would answer that the purpose of data is to improve individual, case-by-case decision-making through best guesses (probabilities), even if we aren't certain. However, evaluators would answer that data are for understanding and judging what works for a group or population through trial-and-error experiments until we are as certain as possible (95 to 99 percent certain) that the average member of a whole group will benefit. Data scientists don't worry about whether their algorithm's predictions might be biased, as long as they are as accurate as possible. Evaluators acculturated values are tuned to believe that correlational predictions aren't socially okay, and that very confident (rigorous) causative conclusions should be the standard for prediction. These differences, such as frequentist vs. Bayesian analytics, have affected the analytics frameworks, methods of data collection, instrumentation, and analysis, such that two very different skill sets have evolved (discussed in more detail in Section 5.1). These differences aren't something the two camps are mindful of, but they show up in many of the debates and misunderstandings that come up when these two groups meet.
- **Weak organizational linkages.** Although hard data is not easy to find, anecdotal evidence suggests that in many large development agencies that have both an evaluation office and a data science office – such as a data center and an innovation lab – the linkages between the two offices tend to be quite weak. The data science office often does not consider evaluation to be part of its mandate, and the evaluation office is often not very familiar with the work of the data center. There may also be a concern on the part of the evaluation office that, in the future, it will be competing for funds with the data center.
- **Economic pressures to move to big data.** One of the attractions of big data is that it is seen by some agencies as being able to collect information on local communities faster and more economically. It may be difficult to convince managers who are working under budget constraints of the benefits of actually visiting communities and meeting people face-to-face. This is expressed in the concern that big data is becoming "extractive," as it makes it possible to collect information about communities, and to make decisions about "what they really need" without their being consulted or even knowing that data is being collected.

These barriers will need to be worked through for a beneficial integration to occur. It is difficult to know whether to begin with the education or philosophical challenges. It seems most forward thinking to begin with both – through joint educational experiences. Combined education could orient and teach data scientists about social science theory, methods, and tools, their purpose, and the ethos of removing bias from analytic processes and conclusions. And, the social scientists could become oriented to data science theory, methods, and tools, their purpose, and the ethos of data being applied in real time on a case-by-case basis, rather than drawing conclusions about population averages that are hard, if not impossible, to translate to useful information on the front lines of social change. The starting place is agreed by both camps: using data to improve the likelihood that every member of a target population will achieve a positive result.

It will be key to start with those prone to early adoption. Among evaluators, a large group seems ready for convergence efforts, namely those evaluators who are not experimentalists, but instead formative evaluators who: i) value participatory processes and more inductive evaluation methods; ii) tend to place high value on mixed methods, tools, and analytics; and iii) see emergence of pathways to success and what works as an ever-changing and evolving process, which means they are open to finding many pathways to success rather than one experimental condition or a one-size-fits-all solution that can be replicated and scaled everywhere. These types of evaluators embrace complexity, and acknowledge and accommodate the constantly morphing realities. Their analytic processes, while not the same as the data science world, seek similar goals with respect to individualized decision-making. These evaluators represent a "ready" group of professionals – social scientists who can add the tools and methods of big data science to their tool boxes in ways that will make program data analytics not only more useful, but also more rigorous and accountable for causation.

The same holds true among data scientists. Many data scientists are now engaging and partnering with non-profits, NGOs, government agencies, and philanthropies to apply their training and tools to build decision-making applications to improve social impact. Through these projects, they are becoming educated about the worries of bias in their techniques, beginning to collaborate with social scientists and researchers, and attending each other's professional conferences, to put their heads together to solve these problems. This group is growing, non-profit and for-profit data science organizations are proliferating, and many are partnering with the social impact sector.

BOX 6:

Demystifying machine-learning approaches:

Beyond the misperception that machine learning only works with quantitative data

It is important to shift beyond the widespread perception that machine learning (ML) only works with quantitative data. One of ML's greatest strengths is that it can work with, and combine, any kind of data, including large volumes of PDF files, any kind of text, satellite images, audio and visual data, and any kind of numerical data. In fact, any kind of data can be reduced to a matrix of 0s and 1s, and thus can be incorporated into an integrated data platform. So ML is a bridge – and a key part of bridging the quantitative/qualitative divide, negating the idea that ML can only work with quantitative data. The confusion comes from the fact that while any kinds of data can be input, the processes of analysis involve mathematical procedures. Although there is likely to be some continued resistance from some qualitative folk, this will gradually be overcome once evaluators gain some hands-on experience seeing how qualitative data can be used and analyzed. The beauty of ML is hidden to the untrained eye, so the challenge for data scientists will be to demystify the approaches.

Bringing these two groups together more formally – for co-education and the development of shared skills, tools, technologies, and standards of practice – would go a long way toward facilitating a rapid and improved change in how the social impact sector benefits from the tools and methods of the big data world.

Chapter 5 discusses some of the areas of disagreement or misunderstanding between evaluators and data scientists. These are categorized into issues relating to theory, to data quality and validity, and to sampling and sample selection bias. In each case the main areas of debate are described, and the authors then try to distinguish between misunderstandings, at least some of which could be relatively easily clarified, and disagreements on the strengths and limitations of different approaches and methods. There is, of course, a fine line between misunderstanding and disagreement. Proponents of a particular approach, who may believe this is the “best” or “correct” way to conduct a study, will often argue that the other side would agree with them if only they had a better understanding of their approach. The distinction between misunderstanding and disagreement is important, because different approaches may be required for bridge-building between data scientists and evaluators. The section concludes that the discussions about theory are likely to be the most difficult to resolve, as theoretical positions tend to combine philosophy and values as well as methodology. In contrast, greater understanding of different methodological approaches is likely to provide a basis for broader agreement on issues relating to data quality. Issues of sample design probably lie between theory and data quality, as some issues relate directly to specific questions of methodology, while sampling can involve broader issues of philosophy and broader approaches to research.

5

A closer look at methodology, areas of integration, and perceived disagreements

When discussing areas of potential disagreement between data scientists and evaluators, it is important to keep in mind that there is no universally accepted "evaluation approach." Many authors, such as Stern *et al.* (2012) and Bamberger, *et al.* (2016), have identified six or eight quite distinct approaches, while others, such as Patton (2008) and Stuffelbeam (2001), have proposed much larger numbers of distinct approaches. Widely used evaluation designs – such as experimental and quasi-experimental designs; statistical (econometric); case studies, including qualitative comparative analysis (QCA); mixed-method, qualitative and participatory designs; complexity-responsive and gender-responsive designs; and systematic reviews – are all quite distinctive. The disagreements among these different approaches are as deep and heated as any of the debates between data scientists and evaluators.

So it is important to appreciate that many data science critiques of evaluation focus mainly on experimental ("frequentist" or "randomista") approaches and, particularly, randomized control trials (RCTs). Many evaluation practitioners who use qualitative or mixed-methods approaches, or QCA are just as critical of the RCT approach as are many data scientists. So in the following discussions concerning points of convergence and disagreement, it is important to clarify which issues concern only specific evaluation methods and which may be more general. Similarly, although space does not permit discussion in this paper, there is no single data science approach. For example, there are discussions within data science on the issues relating to the role of theory.

5.1. The role of theory

Theory is important to many evaluators and data scientists, but it does not play the same role for each. Within evaluation, theory also has a very different role in quantitative evaluations which mainly use deductive, theory-driven approaches, for which the evaluation is designed to test a specific, pre-determined theory, Also, there are many qualitative evaluations which use inductive approaches that have no predetermined theory, and where theory – if it is used at all – is gradually developed as the researcher comes to gain a better understanding of the context in which the program being evaluated is immersed. Many in the evaluation field might argue that theory does not play a role in data science at all, but this would not be completely true. While many data scientists do not conduct their work by developing or testing theory, many others would state that theory plays an important role in their analytic work, but in a different way. If we are to achieve a convergence of evaluation and data science, it is very important to understand and then reconcile how theory plays different roles for evaluation studies than for data science projects, and vice versa. What follows are statements of how theory is viewed most often by evaluators and data scientists, respectively.

- **Deductive vs. inductive discovery.** Many quantitatively oriented evaluators consider theory to be a certain truth – to be discovered through deductive experiments. There are, however, other branches of qualitative and mixed-methods evaluation that use inductive methods, whereby the theoretical framework evolves as the researcher gains a deeper understanding of the community or organization being studied. Specifically, when beginning a study, evaluators trained in quantitative social science tend to: i) formulate an expert-developed theory, specifically a program theory or "theory of change," based on prior research or field knowledge; and then: ii) design a deductive experiment to test the theory, gathering and analyzing new and original data using inferential statistics and controlling for all other variables to determine if a treatment or program caused an outcome and to conclude, with 95 to 99 percent certainty, that the result was not due to chance. If the null hypothesis that the program did not make a statistically significant difference is rejected, then the theory is supported and the program is deemed "evidence-based" and worth replicating. The most accurate conclusion would be that the program, if repeated on a similar population with fidelity to the experimental conditions, will statistically significantly improve the average score of an entire target population.

  The emphasis on the average improvement score is a critical feature of the experimental approach. It also offers an important distinction from many data science approaches which seek to predict specific outcomes for different sub-groups of the population, or even for individuals.

  It should be added that adherents to the increasingly influential mixed-methods approaches would not completely follow this deductive approach. A common variation of mixed methods begins with exploratory studies to gain an understanding of the local culture and context before designing quantitative survey research. In another variant, quantitative studies are used to develop typologies of outcomes, which are then explored through case studies or in-depth interviews to understand the variations of outcomes.

- **Ever-changing work-in-progress.** Data scientists view theory as an ever-updating probability to be discovered inductively through historical observations. Put another way, data scientists don't view their research as testing and refining a predetermined and unchanging theory through experiments, but rather as an ever-changing work-in-progress to be updated and improved upon with new observations and analysis. The most accurate conclusion for a data science investigation would be that, based on prior observations of what had happened so far, "X" is the probability or likelihood of a particular outcome for this program. In fact there will often be a set of probabilities (Xs) of expected outcomes for different sub-groups in the population. It should be noted that many evaluators who work within a complexity-responsive framework recognize the importance of "emergence" – and the fact that the program being evaluated and the context within which it is operating are in a state of continual flux, so that a much more flexible or adaptive approach must be used. In operational terms, this is often referred to as adaptive (or agile) management.

Theoretical differences lead to different uses of findings

These views on the different roles of theory lead to very different approaches to how data are analyzed and, therefore, very different uses of the findings. With respect to the estimation of changes in population means, most experimental evaluators (who most commonly use RCTs) are frequentists (see Glossary), while many data scientists are Bayesians (see Glossary). While recalling that many, and perhaps the majority of, evaluators do not use or even believe in RCTs, these distinct differences in approach to theory and use of different analytic approaches hold true for quantitatively, or at least experimentally oriented evaluators.

Frequentists deductively test new, experimental data and ultimately determine – through analysis and controlling for all other variables – if an intervention made a statistically significant difference. The most important point here is that the social scientist's use of deductive theory testing through experimentation must determine if and how a program does or does not work to move the average of the treated population. In this way, theories are being tested for entire populations, although regression analysis can also test for differences in outcomes between groups, such as women and men, or those of different education levels. Theory testing does not concern itself with each case's treatment, but rather whether the overall population average improves. It is this approach to theory that makes frequentists so relevant to policy-making, as the concern is with a whole population.

Bayesians inductively discover and refine their theory as more and more individual observations are made, in order to make more accurate predictions on a case-by-case basis. While in some cases, Bayesian analytics do use statistical modeling, the results are deployed on a case-by-case basis. The data science community has also developed and advanced machine learning, again to maximize the prediction accuracy of a conclusion for each case based on all similar cases from the past. It is for this reason that Bayesian use of findings is literally "applications." The reason many data scientists are trained within schools of computer engineering is that the results are used to drive decision-making and recommendations for the person, situation, or event that is about to be acted upon. The best tech-era vehicle for delivering predictive and prescriptive insights or recommendations is the software application.

Just because Bayesian data scientists are inductively developing and updating their theories as more data come in, instead of deductively testing theories to confirm what is believed to be true, doesn't mean they are not using theory at all. In fact, these Bayesian data science approaches are building and improving upon theories in fields such as health care, education, and employment, and driving innovation and personalized solutions that address many of the downfalls of treating everyone the same way – all based on frequentist population studies that require rigid fidelity to the implementation methods used during the experiments that proved a treatment could move the overall population average. For example, there is a "science of the individual" movement afoot in the education field, led by Harvard University educational neuroscientist Todd Rose, to build models (theories) that can help fine-tune educational experiences to the different types of strengths and needs of each student. In his book, *The End of Average: How We Succeed in a World that Values Sameness,* Rose (2016) argued for a Bayesian inductive science approach to help students learn the way they prefer to learn, rather than be taught the same way en masse, simply to move the "average" test score. This leaves so many individuals with different learning styles out.

Data science can be applied without theory, leading to all kinds of biases

While being a Bayesian data scientist doesn't mean you don't use theory in your work, there are way too many data scientists using analytic approaches and machine-learning algorithms that make predictions without any consideration of theory at all. The problem with a Bayesian approach is that it doesn't require theory building, which means it doesn't require results to reflect a "true" or "valid" understanding of why or how something works, leaving it blind to all types of potential biases.

Frequentists are trained to begin with theory, thereby ensuring that their experimental design and analytics have the clear goal of determining cause-and-effect. Frequentists hold a philosophical view that understanding causation is the goal, but their methods and analytics have it baked in. The job of their analytics, in fact, is to refute the theory and its hypotheses. This analytic skepticism improves the likelihood that a conclusion is not spurious, whereby a correlation gets misunderstood as causation.

Unfortunately, many, if not most, data scientists are not trained to always build theory. Therefore, their methods and analytics do not have the theory-based ethos of trying to figure out what "truly" works rather than what is correlated with an outcome. Just because murder rates go up when ice cream consumption goes up does not mean ice cream consumption causes more murders. The problem is that data scientists trained within computer engineering departments are taught to make the most accurate predictions possible, regardless of whether the conclusion was based on selection bias or on the correlations of inequities inherent to the way systems treat people who are different or in the minority. Predictive accuracy is to the data scientist what the p-value – the level of confidence that there is a statistically significant difference between the mean score for two samples[1] – is to the experimental evaluator. For example, data scientists have determined that race is a strong predictor of whether a student will complete college and, therefore, will often leave race in their models for "predicting" graduation in order for colleges to prescreen applications for admission. Just because race is a predictor certainly does not mean that the color of one's skin is a determinant, or cause, of staying or dropping out. Now, non-theoretical data scientists concerned with prediction accuracy above all else wouldn't want to remove race as a variable in their models because it would, in fact, weaken the metric they care most about – predictive accuracy.

As Cathy O'Neil (2016) pointed out in *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* when accurate predictions are the goal, any variable, regardless of whether it is a correlate or a cause of an outcome, is fair game. What she is implying is that predictions without theory-building or testing is unethical, unjust, and discriminatory. There are many examples in the data science world provided by O'Neil's book, and many more in the news every day. In "Black Box Justice," Popp (2017) cited the work of data scientist Richard Berk, who is building algorithms used by many criminal justice systems to predict future crimes. His algorithms use all variables – from race to zip codes – because, as he puts it, "I'm not trying to explain criminal behavior, I'm trying to forecast it. If shoe size or sunspots predict that someone's going to commit a homicide, I want to use that information – even if I have no idea why it works." This very dangerous attitude and resulting applications will promulgate injustice. This is the direct result of theory-less data science.

That said, it is important to realize the power of data science tools and methods, and to use them to build and iteratively test cause-and-effect theories, rather than perpetuate inequities, prejudices, and injustice. Many in the research and evaluation community cite O'Neil (2016) and Eubanks (2017) and other studies and articles as proof that the algorithms of data science should not be used at all. This would be an injustice as well, as we have always learned that technology can be used for good or bad. What matters is in whose hands we put and entrust a technology and how we have trained them and imbued their techniques with the ethics and values we want.

5.2. Disagreements on data quality and validity

Disagreements on data quality are really just misunderstandings

There are misunderstandings and thus disagreements on issues of data quality when it comes to the use of big data for evaluation and research. Evaluators

[1] For more detail on the P-value see en.wikipedia.org/wiki/P-value

quantitatively trained in deductive experimental research methods and the use of frequentist analytics desire normally distributed data using numeric metrics (variables). There are plenty of data analyses that also use non-numeric data when applying non-parametric analyses. However, these techniques aggregate non-numeric data for mathematical and statistical calculations and, as such, they also require specific rules, stipulations, and manipulations to make these non-numeric data statistically analyzable. Because of the frequentist approach and statistical requirements for both parametric and non-parametric analyses, much of big data would understandably be considered too messy to be usable. However, statistical aggregation is not the only analytic lens through which data can be viewed.

In the data science world, the use of Bayesian inference and tools, such as machine-learning algorithms, allow for the identification of patterns that are not based exclusively on aggregation methods, such as the use of means, medians, and frequencies. As a result, data considered too messy for statistical modeling are now available for algorithmic modeling. It isn't that there are no constraints within the data science world, but for the most part, they are limited by: i) too much missing data; or ii) lack of variance within a measure (variable). Additionally, most data scientists are trained in how to blend both statistical models and other algorithmic models, such as machine learning, thereby maximizing the usability of all data and analytic techniques. While concerns of data quality by the evaluation and research communities are very real for their analytic toolbox, learning about and combining the tools of the data science world would diminish these concerns considerably. What remains is a need to focus on missing data and data that provide no variance or – put another way – data that are so flat as to make them worthless.

Other evaluation concerns about data quality and validity

While some evaluators' concerns about how data scientists address data quality are at least partly based on misunderstandings about the data science approach, others relate to the more fundamental differences of approach. Quantitative, qualitative, and mixed-methods evaluators all invest considerable time and effort in developing indicators that are culturally meaningful, measurable, and reliable.

Evaluators who devote effort to ensuring the quality and cultural appropriateness of their data tend to be skeptical of assurances from data scientists that data quality need not be a concern, because artificial intelligence and data mining (see Glossary) can detect all of the cultural nuances and issues relating to how language is used.

BOX 7:

Capturing the cultural essence before preparing a survey

For an exploratory study to understand the survival strategies of low-income households in Cartagena, Colombia, in preparation for a household sample survey, an anthropologist spent a month living in some of the communities to understand how concepts such as "household" or "vulnerability" were understood, as well as to identify how the many forms of social debt and reciprocity were used. It was found, for example, that a household was defined as all of the people who eat from the same pot, so there could be people living in the same room (these were very poor communities) who were not considered part of the same household. Consequently, the basic sampling unit – the household – was defined very differently than it would have been for conventional demographic and household surveys which would have classified everyone living in the same dwelling as part of the same household. More importantly, the concepts and social support mechanisms relating to reciprocity are very culturally sensitive – something most previous surveys had failed to capture.

Source: Bamberger

Evaluators are also concerned about data that is collected remotely, and without understanding of the social context in which it is generated. Consider social media posts by people – particularly but not exclusively young people – when others were present who may have influenced the post in ways that are difficult to understand or assess remotely. Gender researchers also report that the phone calls or social media posts of young women in many cultures are produced under the watchful eye of parents, mothers-in-law, or suspicious husbands. Sociologists and anthropologists argue that communication is a form of social interaction, and that it is essential to understand the social context in which communications or forms of social behavior take place. Can the remote analysis of data capture these social influences?

Construct validity

Many of the key concepts and constructs used in evaluation are complex and require combining a number of different indicators. For example, such constructs can include poverty, gender equality, empowerment, quality of life, or mobility. A danger of working with secondary data, including big data, is that the analysis must rely on proxy variables that use data collected for a different purpose. The challenges are increased when combining variables to form complex constructs. Evaluators argue that the challenges for data science to create valid constructs is further complicated by the fact that many data scientists would argue that a theoretical framework is not required to develop or validate constructs, but that the constructs will gradually emerge as the data is mined and analyzed. But how will the analysts know if the constructs adequately capture what is being measured if there is no theory or it is not possible to assess the quality of the data that are used in the analysis? As an example, consider the many studies evaluating the impact of microcredit on women's empowerment in countries such as Bangladesh and India. Many such studies measured women's empowerment in terms of women's control over household decision-making, which called for asking women to estimate their percentage of control over major household budget decisions. This is something that big data analysis could measure. However, studies conducted by anthropologists and others familiar with village life found that most women felt this was not a meaningful measure of empowerment (Hashemi, *et al.*, 1996). Although this might be a desirable long-term goal, many women felt that battles over budget control could have very negative outcomes for women, given the very restricted environment in which they lived. When asked to construct their own empowerment scales, they began with much more modest steps, such as being allowed to walk to the edge of the village without a male escort, or not having to jump into the water to allow a male to pass on the narrow footpaths traversing the rice paddies. While to western eyes these may seem shamefully small steps towards empowerment, these were the perspectives of the women in these villages. It is difficult to see how data mining could capture these humble local perspectives. In fact, very few of the donor-funded studies had investigated these perspectives. So there could be a danger of constructing an empowerment analysis using constructs with arguably low validity. It should, of course, be acknowledged that many quantitatively oriented (and frequently cited) evaluation studies had similarly low construct validity.

As shown, there are strengths and weaknesses of both data scientists and evaluators on most methodological issues. It could be argued that many quantitative evaluators also tend to use narrow and rigid approaches in developing complex constructs, and often fail to incorporate qualitative dimensions or consider the broader context. Also, as shown in the example of measuring women's empowerment in Bangladesh and India, it was quantitative evaluators, rather than data scientists, who were using what many considered culturally inappropriate western concepts of empowerment.

What this shows, in a big picture, is that these challenges could be at least partially remedied by joint training for data scientists and evaluators when conducting big data/analytic studies. Everyone needs training on construct measurement within a specific context. It is probably true that data scientists are currently no more (or less) prone to make these construct measurement mistakes than evaluators or other social scientists. The social science world is also replete with these kinds of errors, in part because of the ethos of researcher objectivity, which creates rules that implicitly prevent the kinds of contextual investigation that is required to do measurement correctly. This is also a function of the desire of many (but not all) quantitative researchers for population-based estimates and conclusions, as contrasted with data science approaches that usually focus on disaggregated data that compares different groups, or in some cases reaches to the level of the individual (see the Broward County, Florida, case study in Section 4, Box 5).

Messages from this discussion: both data science and evaluation approaches have strengths and weaknesses, and there is a critical need for convergence. Neither data science nor evaluation is "better" or more "rigorous" than the other, but rather each has complementary strengths and limitations. Furthermore, the current lack of strong links between the two approaches means that both sides spend a great deal of effort debating which approach is "best." For example, should evaluation be replaced by the much more sophisticated approaches of artificial intelligence? Or, is data science fundamentally flawed by its inability to understand the cultural and social realities of the world it is trying to study? In response, both approaches could be greatly strengthened by integration, and more time should be spent finding points of agreement on which to build.

5.3. Concerns about selection bias

A brief primer on selection bias

Selection bias occurs when a randomly selected sample is not representative of the target population, or if the program is not fully administered to all those selected for treatment, which thereby makes the results ungeneralizable to the whole population. In lay terms, anyone close to a problem or solution that was studied or researched could refer to findings where selection bias occurred as, "you don't have enough people like [fill in the blank], so your findings don't apply to them," or, "sure it worked, but only for those who completed it, fully. Those who did not complete it did not do so because they are different. So, your findings don't apply to everyone."

The first concern of selection bias – not being able to randomly sample the entire target population – is a problem of access, cost, and feasibility. Most studies are simply too small or too local to be broadly generalizable. The practical solution would call for narrowing the generalizability to the local target population, and perhaps to specific segments or sub-populations within the local target population. However, taking this step to prevent selection bias leads to conclusions that cannot or should not be applied to other settings or unstudied sub-populations within the larger target population. Selection bias due to weak target population sampling is replete in studies among the social science community. As such, while the research advances knowledge, it is very difficult to generalize these findings from the local scope and the segment of the population actually selected to the general target population. Even so, public media is fond of doing this, but it is seriously problematic.

The second concern of selection bias arises after sampling a population, at the point of assigning subjects to treatment and no-treatment groups within a study. Once selected through sampling, subjects from the target population will naturally differ on the personal, interpersonal, situational, and environmental variables in their lives, outside of, or exogenous to, what makes them a part of the target population. These exogenous variables can impact if and how different subjects respond to a program, including how they are treated by those providing the program. Due to these differences in engagement and treatment, each person's likelihood for completing and receiving all the key program ingredients in the exact same dosage will be low, no matter how rigidly the instruction for how they should be delivered is prescribed. So, if a study doesn't control for these "life" differences occurring outside of and uncontrolled by the program experience, selection bias will arise in the form of unintended treatment variations, attrition problems, and unmeasured factors in people's lives which actually cause or hinder the outcome, not the program. In lay terms, this type of selection bias pertains to how people who look or act differently, and come from different circumstances and places engage in treatment, as well as get treated by the implementers.

Controlling selection bias with random assignment to treatment groups. Selection bias is the social science community being mindful that the individual conditions of all beneficiaries' lives play a role in how they individually "select" whether and how much they can and will engage in all aspects of a treatment. Additionally, it is the recognition that the human beings who implement a program bring their own implicit and explicit biases in terms of how fully they treat different types of people from different backgrounds. The method for controlling this type of selection bias is the random assignment of persons to treatment groups. Doing so spreads the "life" differences equally across all treatment groups, eliminating life variables as a possible reason why some might or might not fully engage, and therefore benefit. Additionally, to address implementer selection bias, studies will "blind" implementers from people's life circumstances, particularly those that might unconsciously bias their treatment decisions. Of course, visible life circumstances, such as race and gender, cannot be blinded. Since these are two of the most publicly biased factors in all aspects of life, removing selection bias at the point of treatment through random assignment will never be possible.

Will the world of big data hurt or help the problem of selection bias?

The evaluation and research community has rigorous methods to address selection bias, but they are extremely difficult and impractical to implement. They can even be unethical, as randomly assigning people in need to a no-treatment group would be considered "unethical" by many. Additionally, selection bias in the form of treatments being affected by the race or gender of a subject cannot be removed through "blinding" the implementers.

Removing selection bias is possible, but it is most often only with small numbers of subjects within local settings, which compromises generalizability to the point of irrelevance to the entire target population. To get bigger, more representative data through an evaluation or research study would have an exorbitant cost that, in almost all cases, would only be affordable to big governments. Of course, this would only be for studies where race and gender either can be completely ruled out as a factor in selection bias, or where studies only focus on one race or gender. The next question is: can the tools and techniques of big data help?

When it comes to selection bias, data science often makes it much worse. For many data science efforts seeking to predict an outcome accurately, there are no considerations or deliberate methods for addressing selection bias. In fact, if selection bias improves the accuracy of many algorithmic models (e.g. how race powerfully correlates with and is a strong predictor of a college student dropping out of school), then there are many incentives not to exclude it. Of course, this pertains exclusively to data science for accurate outcome prediction, not data science for studying what works. It is here where the opportunity lies.

While many applications and uses of data science raise real and serious concerns with respect to selection bias, there is plenty of experience with using these tools to address it. Arguments are now being made that data science can do a better job of controlling for selection bias while drawing more generalizable conclusions than traditional experimental and quasi-experimental approaches. For example, Judea Pearl, a Turing Award winner, expert in machine learning, and grandfather of Bayesian Networks, and co-author Dana

Mackenzie provided a mathematically provable method for conducting causal modeling using data, including controlling for selection bias in their book, *The Book of Why: The New Science of Cause and Effect* (Pearl and Mackenzie, 2018). Other data scientists are advancing this field, as a new kind of data-driven experimental design. Many "social data scientists" have long concerned themselves with selection bias in fields such as medicine, political science, public health, economics, and sociology, applying rigorous observational study methods that leverage existing big, representative, or complete datasets, and applying analytic techniques to control for selection bias.

First, they take advantage of the representativeness of bigger data sets, such as electronic health records from government, and large health systems, such as Medicare data in the United States, voting datasets, census data, nationally funded household survey data, and business transactional datasets. Since many of these datasets are complete target population datasets or truly randomly sampled at regional and national levels, they are able to ameliorate the first selection bias problem noted above.

Second, they have developed statistical and algorithmic solutions to match people in historical datasets based on their likelihoods to engage in and get treated within a program or intervention. Over 45,000 studies have been published using these selection bias removal techniques with big observational data. Meta studies comparing the validity of findings from these studies and the removal of selection bias have been found to do as well as – and at times even better than – experimental studies (*New England Journal of Medicine*, 2000a; 2000b). Selection bias factors, such as how people get treated based on factors such as race and gender, are beginning to be addressed through these techniques as well as with larger more representative cases. There are also segmentation methods that can find naturally occurring experiments within every segment.

Third, more and more modeling projects are removing race and gender from the causal analytics process, first learning what works for all, then adding factors such as race and gender into the model at the end to determine if there had been a significant difference between races and genders as to whether they received all of the treatment methods found to have caused an outcome. In this way, the ability of statistical and algorithmic methods for controlling for selection bias will continue to advance, evolve, and improve.

To remove selection bias using data science techniques doesn't require really big data

The statistical and algorithmic (machine-learning) methods of: i) matching based on likelihood to engage and get treated; ii) finding naturally occurring experiments in historical data; and iii) testing, post facto, for inequities in treatment, can be applied and conducted starting with a few hundred cases in a program administration dataset. More and more, evaluation and modeling studies are applying the big data observational study methods for controlling selection bias. There have been observational studies, using selection-bias-removal techniques with organizations' program administrative data of a few hundred to a few thousand cases for those providing programs in, for example, child welfare, juvenile justice, residential psychiatric services, and workforce development. The data science techniques deriving from applied observational studies and the data of the big data world can be applied to the smaller but expanding world of administrative program data. It is important to note that when these smaller, albeit complete, datasets are used, the findings only apply to the target population of that organization. However, it is now becoming possible to connect datasets or the metrics and variables gathered across organizations that provide the same type of programming.

Chapter 6 draws on the previous two chapters to illustrate how some of the data science tools and techniques can be applied to strengthen evaluations. The section begins by discussing some of the main challenges that conventional evaluation approaches face and to which big data could potentially contribute. Challenges are discussed with respect to: evaluation design, data collection, sample design, and data analysis and dissemination of findings.

Evaluation design challenges include determining how to define the counterfactual, when experimental designs are not possible, for evaluating complex programs and identifying unintended outcomes.

Data collection challenges include the cost and time required for data collection and also collecting information on difficult-to-reach groups, monitoring implementation processes and behavioral change, integrating different kinds of data, and collecting information on the spatial dimension of programs (changes that take place outside the immediate project area).

Sample design challenges include reducing sample selection bias and ensuring the sample is sufficiently large to ensure statistical significance and to permit disaggregated analysis.

Data analysis and dissemination challenges include working with very large data sets, integrating data from multiple sources and in different formats, analysing complex programs, and providing actionable predictions of the likely effects of different interventions on different groups. The dissemination of evaluation findings in a timely manner and in formats that are understandable to different groups has also proved to be a challenge.

The section then illustrates how the incorporation of some of the data science techniques and approaches discussed in the two previous chapters can help address these challenges. The message throughout is that data science should be used to complement – not to replace – conventional evaluation approaches.

6

Current and potential applications of data science for development evaluation

6.1. Challenges facing the design and implementation of all evaluations, and potential contributions of data science

There are three types of challenges that many evaluations face – evaluation design, data collection, and data analysis. This chapter looks at how data science can contribute to addressing each of these challenges.

Evaluation design challenges

Strengthening the counterfactual

Most development evaluations do not have sufficient control over how programs are implemented to permit the use of randomized control trials (RCTs) – even if they were considered to have the most appropriate design. Consequently, many evaluators will decide to use a quasi-experimental design that matches the project group with a comparison group that matches the project population as closely as possible. In some cases, the two groups can be matched statistically using techniques such as propensity score matching, while, in other cases, the only option is to use judgmental matching.

Evaluations conducted retrospectively face an additional challenge – the availability of baseline data capturing the conditions of the two groups at the start of the project. A final challenge is the difficulty of obtaining contextual data to match the groups on the local and wider socio-economic, political, demographic, and ecological factors that may affect program performance. The availability of reliable, affordable, and accessible comparison group data to address these three sets of factors is frequently a challenge – and all are areas for which big data can potentially strengthen the evaluation design.

For area-based sampling, such as environmentally protected areas, it is sometimes possible to use satellite images to identify characteristics on which the project and comparison group samples can be matched. For example, it could include distance from the protected area's boundaries, density of roads and services, or density of forest cover. In some cases, density of phone coverage can also be used as a proxy for level of economic development. Matching will be strengthened when satellite data can be complemented by ICT with GPS mapping or remote sensors, or by survey data. The different sources of indicators can be combined using propensity score matching (see Glossary) to strengthen matching.

Evaluating complex programs

The evaluation of complex programs normally requires the collection of information on a large number of program components, contextual factors, interactions among multiple stakeholders, integrating often inconsistent monitoring data collected by different agencies, and tracking complex, non-linear processes of change (Bamberger *et al.*, 2016). All of these require collection of larger and more complex data sets than what is required for the evaluation of simple or complicated programs. Big data, often complemented by ICTs, can contribute by:

- combining a range of big data and ICT techniques to collect a wider range of contextual data
- using systems mapping to map the interactions among the different components of the intervention and its context
- using social media to track attitudes and behavioral change
- using software to develop scales and indices, such as concept mapping, for the different dimensions of complexity.

Identifying unintended outcomes

Many widely used evaluation designs fail to capture unintended outcomes. Most quantitative designs, including randomized control trials, are designed to test whether intended outcomes have been achieved – for example, if there was a statistically significant difference in the change of specific outcomes between the project and control groups over the life of the project. They are not designed to identify outcomes that were not included in the original project design and research hypothesis. While qualitative designs, such as a theory of change, can potentially identify unintended outcomes, evaluation clients are often only interested in knowing whether their project has achieved its intended outcomes (Bamberger *et al.*, 2016).

Big data and ICTs can potentially provide real-time or rapid feedback on changes in a range of key indicators so that the process of project implementation – the time when many unintended outcomes occur – can be tracked. Big data and ICTs can also provide feedback on the influence of a wider range of contextual factors that can contribute to unintended outcomes. More importantly, actionable feedback can be provided to managers and other groups so that early signals of potential problems can be explored. An on-line theory of change can provide a useful framework for identifying, tracking, and updating unintended outcomes. This can be complemented by the analysis of Twitter and other social media to track potential problems that might produce unintended outcomes.

Data collection challenges

The high cost and time required for data collection

One of the serious constraints for most evaluations is the high cost and time required to collect data. Consequently, many evaluations have to work with relatively small samples, struggling within their budgets to generate the smallest possible sample that will provide adequate statistical power to test for statistical significance between two groups for the total population. Disaggregated statistically significant comparisons among sub-groups is frequently not possible. Data collection costs can also result in pressure to exclude marginal and difficult-to-reach groups where the time and, consequently, the cost for each interview can be significantly higher. Big data is well suited to address these challenges.

As most big data has already been collected for different purposes, it can usually be accessed at a relatively low cost for monitoring and evaluation purposes. Much of the data can also be delivered in near real time and updated continuously. Smart phones can also provide data quickly and economically.

Collecting data on difficult-to-reach groups

Groups such as drug dealers or illegal immigrants try to avoid being identified, while other groups can be difficult to reach because of security situations or because of their remote and inaccessible locations. In some of these cases, smart phones can be used for contact by, for example, interviewing people by phone rather than in person, or by automatically monitoring if they receive automated phone messages, such as reminders for medical appointments, and, if so, how they follow up. Women and other groups that do not have voice in a particular community may speak more freely on a phone. People in high risk zones can sometimes send out video and audio recordings of the situation in these zones, and satellites can also track population groups that would otherwise be difficult to locate or contact, such as refugees.

Monitoring project implementation and processes of behavioral change

Big data can often provide real-time and continuous data, which is helpful for observing the processes through which a program evolves. Smart phones and remote sensors can also provide different kinds of continuous monitoring data. Studying behavioral change also requires capturing information on processes, rather than just comparing two points in time. A variety of big data and ICT resources can assist – for example, phones can capture video and audio records of meetings, work groups, and different aspects of community life that can be helpful. Social media is a rich source of information as are analyses of Twitter feeds and social networks.

Collecting qualitative data

High quality qualitative data is often difficult to collect, plus the recording and interpretation process often introduces a level of subjective interpretation that is difficult to control. Smart phones can now collect high quality audio and visual data, and software for analysis and interpretation is also improving rapidly. Employing these collection methods can remove certain kinds of reporting bias or subjective interpretation.

Collecting and integrating different sources of data

Both big data and ICT offer a range of ways to integrate data from multiple sources and in multiple formats. Big data can do this for very large data sets, and ICT can do the same for smaller data sets.

Enhancing quality control and reducing the cost of data collection and analysis

ICT software is now available to control the quality of data at all stages of the collection and analysis process. This includes a series of consistency checks on how data is inputted, such as the ability of a supervisor to turn on audio remotely and listen to a mobile phone interview while it is in progress. GPS can also check that the right household is being interviewed and, for random route sampling, the phone can ensure that appropriate selection procedures are used.

Collecting information on the spatial dimensions of programs

Satellite images, remote sensors, and GPS mapping can help analyze the spatial dimensions of programs. These tools also make it much easier to include a more in-depth analysis of the contextual factors – transport networks, access to services and markets, population movements, soil quality, and crop production – that are essential for a full understanding of the wide range of factors affecting program outcomes.

Sample design challenges

Recognizing and avoiding sample selection bias is a challenge for all non-experimental evaluation designs. Post-project differences between the project and comparison groups – usually assumed to be due to the effects of the intervention – are actually often due to differences in how the two groups were selected. A related challenge, which receives less attention, is that sampling frames often do not cover all of the sample population and, frequently, there are important differences between the population that is sampled and the population that is excluded. Very often, the excluded group is poorer or has less access to services. Big data and ICT can help address these problems in a number of ways using data from satellite images, GPS maps, and phone company records.

- Satellite images and GPS maps can provide images of the total target population that can be overlaid with the population that is actually sampled, to determine if there are important differences.
- Phone companies keep detailed information on their customers, which can be used to ensure that selected samples of phone users are representative of all phone users. It is also possible to determine how closely a sample of phone users matches the total population.
- Combining satellite images with GPS mapping data and information from household, farm, and other kinds of surveys can improve the matching of comparison and project samples by using techniques such as propensity score matching.

Data analysis and dissemination challenges

Big data analytics offer a number of powerful tools for the analysis of data sets that are too large and complex for analysis using conventional data analysis programs and computers. These can be broadly classified into basic, advanced, and operationalized data analytics.

- **Basic.** Basic analytical tools use data mining to break data into smaller units that are easier to explore. Data visualization is used to present the findings in an easily understandable manner and can provide data for program monitoring. These techniques are often used to identify trends, relationships, and patterns that can later be explored with more advanced analytics.
- **Advanced.** Advanced analytics include predictive modelling and text analytics that analyze unstructured text and transform it into structured information that can be analyzed quantitatively.
- **Operationalized.** Operationalizing data analytics for an organization or set of organizations calls for developing models to design the particular applications required by a particular organization.

6.2. Examples of promising areas for collaboration between big data and development evaluation

Incorporation. Big data has been incorporated into most of the widely used evaluation methods, as illustrated in the examples presented in Table 5. UN Global Pulse, which focuses on harnessing big data safely and responsibly as a public good, has conducted more than 100 proof-of-concept projects in cooperation with national development agencies around the world. Results have shown that a wide range of big data techniques for data collection, analysis, and dissemination can be applied in developing countries. Many of these techniques are already being used in other development fields, such as emergency relief, early warning, and development research. Consequently, an extensive range of techniques has already demonstrated viability in development contexts. What remains is to find ways to encourage evaluators to make use of these techniques.

Dissemination. Dissemination of evaluation findings faces challenges, as the process of printing and distributing reports is expensive and time consuming. Often, printed copies only go to the main stakeholders, meaning civil society organizations and community groups may be excluded. For many groups, long, written reports with multiple tables are not user-friendly, and, if only aggregate data are presented, it is not possible to find information on particular districts or communities. Table 6 provides examples of how data visualization has been used to make evaluation findings available on tablets or laptops, presented as understandable maps or other graphics.

In conclusion, Table 7 identifies some of the conditions where the integration of data science and evaluation are likely to be most promising.

TABLE 5: How big data and ICTs have been used to strengthen widely used evaluation designs

EVALUATION DESIGN	EXAMPLE
Design 1: Experimental and quasi-experimental designs	
• Randomized control trial	High frequency metering data used for high quality information about energy consumption and demand in rural solar micro-grids in India. (Source: Poverty Action Lab)
• Randomized control trial [a second example]	Tablet-based financial education in Colombia uses savings and transaction data combined with survey and telemetric tablet data (Source: Poverty Action Lab)
• Strong quasi-experimental design	The Global Environment Facility [GEF] has used quasi-experimental designs to assess the impact of its programs to protect forest cover and mangrove swamps. In the case of protected forest programs, satellite images and remote sensor data were combined with conventional survey and secondary data to construct comparison groups using propensity score matching. Time series data obtained from the satellite images permits the use of longitudinal analysis. Pre- and post-test comparison group design uses propensity score matching to strengthen the comparison group. (Source: GEF, 2015)
• Natural experiments	Effects of a government tax increase on smoking used changes in search query volume to assess the effects of a major increase in cigarette smoking in the US. Canada, which did not have a similar increase, was used as the comparison group. (Source: Letouzé, *et al.*, 2016: 237-8)
Design 2: Statistical modelling	Causal interactions between labor market shocks and internal mobility evaluated using mobile phone data (Source: World Bank, Latin American region)
Design 3: Theory-based evaluation	The Robert Wood Johnson Foundation conducted an evaluation of a 10-year program to assess the impacts of its program to improve health and safety outcomes in distressed US cities. This combined a quasi-experimental design, including comparison cities, with a theory of change. Given the size, complexity, and duration of the program. Very large data sets had to be managed. (Source: cited in Leeuw, 2016)
Design 4: Case-based evaluation	QCA country-level data assessing factors determining impacts of women's economic empowerment programs at the national level (UN Women. 2014. Independent Evaluation "An empowered future: corporate evaluation of UN women's contribution to women's economic empowerment."
Design 5: Participatory evaluation	The World Bank India Social Observatory uses a participatory approach to involve women in the identification of the key questions that should be included in large scale community surveys to identify priority development issues. Community women are involved in conducting the surveys and in the interpretation of findings. The surveys have been administered to over 800,000 households, so data analytics are required for the analysis and synthesis of the findings. (World Bank, India Social Observatory)[2]
Design 6: Review and synthesis approaches	A review and synthesis study was conducted to assess the effects of microcredit on women's empowerment. The study used data analytic search mechanisms with customized key-word sequence to cover academic databases and on-line portals (Vaessen *et al.*, 2016).

Source: Adapted from Bamberger, 2017

[2] www.worldbank.org/en/programs/social-observatory/brief/setting-up-the-social-observatory-in-bihar

TABLE 6. Examples of data visualization that make complex analysis easily accessible to managers and local communities

Feedback from more than 1 million women in India obtained through tablet-based surveys that the women design and interpret through data visualization (Source: World Bank)[3]
Response to disease outbreaks (Source: FAO in collaboration with UN Global Pulse)[5]
Crop disease detection automated with maps that allow farmers and agencies to drill-down on specific locations (Source: Makerere University in collaboration with UN Global Pulse)[6]
Radio mining conducted in Uganda, with maps permitting users to pin-point locations where particular problems were identified through analysis of radio programs (Source: UN Global Pulse)[7]
Data compilation on Boston disaggregated by city vitality, culture, economy, education, environment, health, housing, public safety, technology and transportation (Source: Boston Indicators Project)[8]
Earthquakes and other emergencies mapped based on crowdsourcing and analysis of social media (Source: Patrick Meier, 2015)

TABLE 7. Factors determining the applicability of big data and data analytics in program evaluation

HIGH APPLICABILITY	LOW APPLICABILITY OF BIG DATA
Large, “complex” interventions	Small, “simple” projects
Programs where conventional evaluation designs are considered methodologically weak	Programs where conventional evaluation designs are considered perfectly adequate, and there is no obvious need for a new approach
Programs that use easily measurable (and readily available) physical measurement, such as climate change, urban growth, traffic patterns	Programs that rely on social and behavioral indicators such as domestic violence, or on community organizations for which there is no readily available data and would require special data collection
Availability of big data indicators with high construct validity (indicators were collected for a purpose relevant to the evaluation)	Big data indicators with low construct validity (proxy indicators generated for a different purpose and their relevance is not clear)
Programs with a relatively long duration and where time series data can be generated in real time	Programs with a relatively short duration and where time series data cannot be generated or are not relevant
Programs that will continue to operate after the initial proof of concept so that prediction is possible	Proof of concept programs that will end after the initial program assumptions have been tested and prediction is not possible
Programs that have a large number of potential variables that might affect outcomes and that have no articulated theory of how outcomes are expected to be achieved.	Programs that require a carefully articulated theory of change to understand the behavioral, socio-cultural and organizational processes through which outcomes are expected to be achieved
Programs that have no political concerns about ownership, control, access, or privacy	Programs that have political concerns about ownership, control, access, or privacy

[3] http://socialobservatory.worldbank.org/articles/participatory-tracking-customizing-visualizations

[4] www.proving.it/

[5] http://unglobalpulse.org/mapping-infectious-diseases

[6] http://unglobalpulse.org/mapping-infectious-diseases

[7] www.unglobalpulse.org/document/using-machine-learning-to-analyse-radio-content-in-uganda/

[8] www.bostonindicators.org/

Chapter 7 uses the example of gender differences in the impacts of big data and other forms of new information technology to illustrate the importance of understanding how different groups – based on gender, age, income, or geographical location – have access to, use, and are affected by the new technologies. It begins by illustrating some of the dimensions of differences in the new information technology experiences of women and men. There are many social, economic, and cultural factors that affect women's access to mobile phones and other information technologies. Examples are also given of negative consequences of mobile phones for women, including increased gender-based violence. It then identifies some important new research areas that should be explored, and concludes by identifying a list of factors that can affect the access to new information technologies by women and vulnerable groups.

7

Gender dimensions of big data and ICT: An example of a deeper exploration of important sectoral issues

7.1. Big data and other new information technologies impact women and men differently

When information technologies and related tools were introduced, it was assumed that they would benefit everyone. However, the term "digital divide" soon entered the conversation, as it became clear that access to these tools was strongly influenced by income, location, education, and gender. Mobile phones offered great potential benefits to women, but it also became clear that there was a potential downside: women's access to mobile phones sometimes resulted in domestic violence as husbands became jealous of women's ability to communicate unsupervised outside the immediate family. The new medium also opened new opportunities for sexual harassment, stalking, cyber bullying and gender-based violence (see Box 8). Yet, even among gender activists, there is still a low level of awareness of these issues.

For example, Wakanuma (2013) described the negative consequences of mobile phone ownership for women in Zambia: husband/partner checks wife's phone, calls numbers he does not recognize, and beats his wife if a male voice answers. Many men also limit their wife's air-time or the services she can use. Wakanuma pointed out that there was so much enthusiasm concerning the liberating effect of phone ownership by women, it was some time before these negative outcomes began to be recognized. Of course, the many positive benefits must also be recognized. Consequently, it is important for a review of the opportunities and challenges offered by the new technologies to look carefully at who benefits and who does not, and identify the factors that determine the distribution of benefits. At this point, most research has been conducted on the differential impacts of the new information technologies on women and men. But many of the issues identified in feminist research also reflect differential benefits in terms of class, income, education, ethnicity, disadvantage, and geographical location.

A reflection on the gender dimensions of new information technologies can also serve as a proxy for a wider discussion on how benefits are influenced by cultural practices and power relations in a given society. In pointing out potential risk, O'Donnell and Sweetman

BOX 8:

Mobile phones, internet and gender-based violence (GBV)

How ICTs are used is strongly influenced by the social norms, prejudices, and power relations of each society. Despite the widespread incidence of gender-based violence related to ICT, Thakur (2018) reported a low level of awareness of GBV even among women activists. Two-thirds of a sample of 909 respondents in Jamaica reported having observed online abuse. The following items are examples of how ICT encourages or condones GBV.

- Mobile phones and internet can promote new forms of online violence, and can increase control and surveillance of women [Thakur 2018]. Online has become home to new forms of violence: stalking, trolling, invasion of privacy.
- Online social norms justify and encourage abuse due to anonymity, the speed at which images and messages can be sent and received, and the fact that electronic exchanges occur at a distance reduces empathy.
- Phone tracking devices increase the risk of violence by making it easier to physically locate callers.
- Being female increases the likelihood of abuse but reduces the likelihood of reporting.

Source: O'Donnell and Sweetman, 2018

(2018) found that the growth and uptake of information and communications technologies "have the potential to improve access to information and services or enable collective action for social justice. But there is also the risk this revolution will carve stark inequalities in terms of who benefits and whose voice is heard."

An extensive literature already exists that emphasizes the increasing accessibility of mobile phones and IOTs which now reach remote rural areas and, increasingly, low-income groups. However, the advances and benefits of the new technologies are often assessed in terms of the four "A"s: access, affordability, availability, and awareness. As noted by feminists and other social activists, these are all quantitative, and largely technical measures of performance and benefits. O'Donnell and Sweetman (2018) emphasized this, explaining that "Technology mirrors the societies that create it, and access to (and effective use of) technologies is affected by intersecting spectrums of exclusion, including gender, ethnicity, age, social class, geography and disability."

Further, according to Gurumurthy (2014), "Existing power relations in society determine the enjoyment of benefits from ICTs, hence these technologies are not gender neutral." O'Donnell and Sweetman (2018) also pointed out that digital tools and methodologies need to be fit for the purpose of challenging and ending the inequalities and injustices that shape poverty, concluding that: "if ICTs are to reach their full potential as a force for change, a feminist and social justice approach is needed. At the very start of the field of 'international development', technologies designed by men in the global North were assumed to be equally useful to all humanity. Now, we know better."

7.2. Research areas for better understanding the gender dimensions of new information technology

In an issue devoted to ICTs, the journal Gender and Development (Thakur, 2018) identified a number of important research areas on the impact of mobile phones and other information technology on women. Research on gender and communication technology can also provide a framework for understanding other dimensions of the digital divide. The research areas include the following.

Complex inequalities: discuss the larger context

Access to, and benefits from, digital technology are the result of the interactions among multiple dimensions. In addition to gender, these include age, education, income, civil status, urban-versus-rural locations, physical and mental handicaps, freedom of information, and infrastructure and regulatory regimes in different countries. So it is clearly an oversimplification to discuss the impact of ICT on women without the larger context.

Digital communication technology: recognize its potential to change gender relations and to empower women

Kabeer's 2015 study – on the role of ICT in helping urban Afghan women cope with daily life – stressed the need to have a broader focus. ICT must enable women to achieve "transformative agency" and to challenge gender inequality. Kabeer argued there is a need to focus on how women use ICT and how men prevent them from achieving transformative agency at the household level, controlling and policing women's use of ICT. Further, Hussain and Amin (2018) warned that a focus on access, not feminist empowerment, means international development policy makers will ignore this critical distinction.

Power analysis: determine whose tools and whose knowledge

O'Donnell and Sweetman (2018) argued that for ICT to be empowering, more on-line content needs to challenge gender biases and fill in the gaps in history, offering an alternative account or "herstory." The digital world is heavily male dominated. Women hold only 17 percent of tech jobs in the UK and only 25 percent in the US. This means technologies are more likely to be designed by men and tailored to the needs of male users. Messages that could cater to women's needs are often slower to be developed. We need to understand how digital content affects attitudes, beliefs, and ideas.

7.3. Implications for evaluating the potential impacts of ICT on women and other vulnerable groups

The increased reliance on this new technology underscores the importance of ensuring that all evaluations of ICT programs anticipate that access to and impacts of ICTs will vary for different socio-economic and cultural groups, and that vulnerable groups, which often include women, are likely to have more limited access. It also indicates how use of ICTs may have negative consequences, including giving rise to new forms of gender-based violence. In the list below, Raftree *et al.* (2013) identified some of the factors that may affect access to and use of ICTs, and can result in biased or incomplete information from the analysis of digital data.

Access. What direct access do different groups have to ICTs? Do individuals own a device by which they can receive, share, or connect to information? Do they share a device? With whom, and how does sharing affect how they use the device? Can they easily borrow a mobile or computer? How often? Do some members of the family or community have more access than others?

Age. What age group is involved or targeted? Does information need to be adapted to certain age groups? Do the very young or very old have access? Do they have resources to cover costs of accessing information or communicating via ICTs? How does age impact on privacy?

Capacity. What skills are needed to access and use a device? Does the target population have these skills?

Conflict and emergencies. Will conditions of conflict or emergency affect access and use? How will they impact on willingness to share information or consequences of doing so?

Connectivity. Is there a network? Is it reliable? Is it steady or intermittent? Slow or fast? How do the connectivity speed and device type affect what information can be accessed or shared? Does this impact how a data-gathering exercise is designed? How does it impact whose "data exhaust" might be available for big data analysis?

Cost. How much does it cost to own a device? To borrow one? To use one? To access information on one? How does cost impact access and use? Who is left out because of cost?

Disability. Do ICTs hinder or facilitate those with a disability from accessing or using a device or the internet, or from participating in an information-gathering exercise? Can ICTs help make information more accessible to those with a disability?

Economic status. Will those with greater economic capacity have more of an opportunity to communicate their points of view and influence a program or the outcomes of an evaluation? How might their disproportionate input bias the data?

Language. How does content in a particular language create data bias? Who is left out or ignored due to language?

Literacy. What are the levels of literacy of the target population? How do they vary and who is left out because of it? If literacy is low, what other alternatives have been considered? Face-to-face contact? Radio?

Power. Will the more powerful community members be more privileged because of access, literacy, or the language of the information shared? Who is being left out by a particular choice of data gathering method or tool? How will this be addressed in the evaluation design, and who will keep an eye on this during the data collection and analysis?

Protection. Does access to a device or to content put people at risk in any way, whether because of the value of the device, the information they may access via the device, the fact that others can reach them through the device, perceptions about their having a device, or access to particular information or communication channels? How does this introduce bias into responses or big data being used to draw conclusions about a group or geographic area?

Privacy. What are people's perceptions of privacy as it relates to a particular device, application, or their organization? How will this affect responses to an information-gathering exercise or their overall internet use?

Security. Has there been any type of perception (or actual incident) of a digital data breach? How might that affect willingness to share information with an evaluator or on a digital platform? Is consent language plain and clear?

Trust. Do people trust your organization, the platform or information source, or the entity that generated the request to participate in an information-gathering exercise? Are you able to detect instances in which mistrust affects responses or response rates without a physical presence?

In their article *Modern Mobility: The role of ICTs in child and youth migration,* Raftree, *et al.* (2013) discussed these issue more broadly as affecting children and youth, but they are all equally applicable to the analysis of data on women's use of ICTs. Many of these are difficult to detect, particularly as women will often be reluctant to report many forms of violence.

For all of these reasons, it is important to recognize that access to and use of ICTs, as well as the generation of content disseminated through these media, are likely to be influenced by the social, cultural, and political norms and practices of each society. Existing patterns of gender relations and social control should also be considered in the evaluation design.

Chapter 8 summarizes lessons about the potential benefits and challenges of a integration of data science and evaluation practice, and builds on this to identify a set of recommendations on ways to move forward to create an environment that is conducive for convergence. The lessons include recognition that the exponential increase in the availability of, and applications for, big data, is creating a new and complex information ecosystem that is fundamentally changing how data is generated and used, along with increasing recognition of the wide range of potential benefits from convergence of big data and evaluation. There are also a number of lessons concerning conditions required for integration to occur, including: the creation of a conducive policy environment, ensuring open data access for all sectors and not just for a few powerful commercial and government agencies, facilitating wider access to advanced computing facilities and analytical expertise, and the creation of organizational structures at the organizational, national, and international levels that promote cooperation. The lessons conclude by identifying a number of challenges, including: deciding how big data should be controlled and regulated, and who has access; and recognizing that while big data has the capacity to empower and give voice to poor and vulnerable groups, it can also be used "extractively" by decision makers. This means they can use information collected from and about poor people when making decisions about priority programs for these groups without having to consult them – or even letting them know that decisions are being made.

It also presents a set of recommendations that identify priority issues to be addressed in moving ahead towards convergence and the benefits this can produce. These include: prioritizing measures to build bridges; developing integrated capacity development programs that combine data science and evaluation approaches, tools and techniques; and promoting landscaping research to better understand the data science and evaluation ecosystems and how they can be better integrated. Finally, it emphasizes that there is a key role for funding agencies in creating space for dialog and collaboration, and providing critical seed funding in areas such as collaborative research and training programs.

8

Conclusions: Lessons learned and recommendations for moving forward

8.1. Lessons learned

The new, complex, and unfamiliar big data ecosystem

Many policy makers, development agencies, and evaluators are still unfamiliar with the complex nature of the big data ecosystem. The many ways in which data are generated, transformed, marketed, used, and regulated are completely different from the much simpler and familiar evaluation ecosystem. The high profile of big data in the media – as an almost magical way to solve world problems but also as a source of threats, fraud, and invasion of privacy – further complicates obtaining an objective understanding.

From the perspective of the evaluator, big data can seem unprofessional, as it does not conform to conventional evaluation practice, and also threatening, due to the concern that big data, which is the exciting new tool, will begin to compete for evaluation budgets and authority in the assessment of development programs. This lack of understanding makes it more difficult to develop a dialog on how to promote integration of data science and evaluation.

The benefits of integrating data science and evaluation

Continued integration of data science and evaluation offers many benefits for both professions as well as for the promotion of social good. From the perspective of evaluators, access to new sources of data and to the new analytical approaches helps resolve many of the challenges discussed in Section 6.1. Many of these benefits relate to the economical and rapid access to a wide new range of data sources, and also to an escape from the constraints imposed by the small sample sizes with which evaluators frequently have to work. With convergence, it becomes possible to incorporate contextual variables, access longitudinal data sets, provide more robust estimates of baseline conditions when using retrospective evaluations, and measure processes and behavioral change. It also becomes easier to identify and include vulnerable and difficult-to-reach groups in evaluations – thus reducing an important cause of selection bias. Data visualization also makes it easier to disseminate findings to a wider audience and in a user-friendly manner.

Once evaluators become more familiar with new tools of data analytics, it will become possible to conduct more sophisticated analyses by working with many more variables and conducting dynamic analysis over time. These new tools have great potential for the evaluation of complex programs, which are particularly difficult to model and evaluate with conventional evaluation tools. Once the two professions begin to work more closely, it will also become possible to integrate predictive analytics with experimental designs and mixed methods, so as to strengthen and broaden the understanding of causality.

From the perspective of the data scientist, closer cooperation with evaluators can help address a number of the perceived weaknesses of many widely used approaches which were originally developed for the assessment of commercial activities such as on-line marketing. Theory-based evaluation can potentially strengthen some of the limitations of data mining by providing clearer guidelines on how to define key evaluation questions and a framework for interpreting findings.

Evaluators also have detailed procedures for assessing the quality of data and for assessing construct validity. Poor data quality is often a major issue for data science. This issue is often not fully addressed by many data scientists who are perceived by evaluators as having a mantra that all data is bad and biased. Mixed methods and the complementary strategies of triangulation have the potential to broaden and deepen the interpretation of findings. On a more philosophical or ideological level, many evaluators are concerned with issues such as social justice, empowerment, and equity, while many – but certainly not all – data scientists do not perceive the need or value of incorporating values into their analyses.

Many development evaluators assume that the design and implementation of programs often include intended or unintended biases against the poor and minorities – an assumption which indicates the potential benefit that could come from a value orientation. A related assumption is that many of the evaluation datasets may exclude important sectors of the population, usually the poorest. Consequently, many evaluators will seek to assess the adequacy and inclusiveness of the data they are using. In contrast, many data scientists do not have skepticism about their data, or they believe that machine learning can teach the computer to identify and correct for these kinds of limitations. Some evaluators will argue that this will only happen if the researchers' experience in the field makes them aware of this potential limitation and of the political, organizational, and economic reasons why these gaps may occur and how they may be concealed. This skeptical approach has already proven useful in assessing some of the social exclusion biases in some of the widely used algorithms mentioned in Section 2.4.

The requirements for integration

Unfortunately, integration of the two disciplines does not occur automatically. In fact, experience has shown there are many situations in which it has not occurred, and where a number of factors can mitigate against integration. There are a number of conditions which are necessary for this to occur, including a conducive policy environment, data access, access to computer and tech facilities, and appropriate organizational structures.

- **Conducive policy environment.** Policy and regulatory requirements may be necessary to permit or promote convergence. These may include regulations concerning access to public data, privacy and confidentiality regulations, and rules concerning issues such as algorithmic transparency. In many countries that have large stocks of survey data stored in sectoral or geographic silos, a major investment may be required to create accessible data platforms so that data from different sources can be combined and accessed. In some countries, governments consider their data confidential, or they are reluctant to share with civil society and commercial agencies. In these cases, a fundamental change in attitudes towards the purpose and use of publicly generated data may be required.

- **Data access.** Even assuming a more conducive policy environment, access to many kinds of data can be expensive or difficult. In many cases, only a few large and influential government, international, academic, or commercial institutions may have access to important data sources such as social media, ATM and phone records, or satellite images. There are also proprietary, reputational, privacy, and sensitivity issues affecting data access.
- **Access to computing facilities and technical analysis expertise.** While some kinds of data, such as social media's Facebook and Twitter, may have been processed to make them easily accessible to the general public, other kinds of data, such as satellite images or phone records, may require access to large computing facilities. Many kinds of data analytics may also require access to large computing facilities or to specialized analytical expertise. These are all considerations that may significantly limit access and use.
- **Organizational structures that promote integration.** In many development agencies, linking the data center and the evaluation office, or supporting evaluation may not be part of the data center mandate. Similarly, the evaluation office may not be familiar with the work of the data center. Effective coordination between the two offices is essential to the integration of data science and evaluation. It is essential for the collaboration to be institutionalized, with regular meetings, sharing of plans, and perhaps joint budgets and joint training for some activities. Management should also identify pilot programs where the two offices can assess the value of collaboration.

Challenges and concerns

Facilitating convergence will require a number of challenges and concerns to be addressed. The following lists some of the most important steps to take.

- **Determine who controls big data,** how is it used, and who has access, and identify the barriers to greater access
- **Recognize that big data has the potential to empower poor and vulnerable groups,** and to be used to hold government and other powerful groups to account. There is already extensive evidence that big data can be used by governments, donor agencies, and other powerful public and private groups to strengthen top-down control. Big data can be used "extractively" to obtain information on and about different groups and to make decisions on what services they need, without having to "waste time and money" going to affected communities to consult with them. However, citizens are often not even aware that this information is being collected about them.
- **Address privacy and security.** Privacy and security are increasingly recognized as important and complex issues which the public and many development agencies do not fully understand and are not able to fully address.
- **Avoid economic and ethnic bias.** Economic and ethnic biases are built into many widely used algorithms, and the incomplete coverage of many big data information sets often excludes the poorest and most vulnerable groups. While the low cost and speed with which data can be collected makes it possible to overcome many of the conventional evaluation challenges to including remote or difficult-to-reach groups, other biases related to the nature of big data have to be addressed.

8.2. Moving forward

We have argued that big data and data analytics have a demonstrated potential in the design, analysis, and use of development evaluations. A wide range of tools, in use in the private sector for at least a decade, is already being used by some development agencies for planning, research, and emergency programs. However, to date, most of these methods have not been widely used in program evaluation. We also discussed some of the reasons why integration has been slower among evaluators. The following looks at some possible steps or actions that have potential to promote and facilitate the integration of big data and development evaluation.

Building bridges

- **Strengthen organizational structures.** Large agencies that have both an evaluation office and a data development office should strengthen the linkages between the two. Support to evaluation activities should be included in the mandate of existing data centers, and mechanisms for cooperation should be clearly defined. These might include: attending each other's management or operational meetings; involving the data center in the planning of evaluations; and involving the evaluation office in the discussions of the data center work program and the kinds of databases they will generate or integrate.
- **Identify opportunities for pilot collaborative activities.** Collaboration on selected evaluation programs should be considered following a careful assessment of the value-added or case for expanding collaboration. The evaluation staff could reciprocate with data scientists and data centers, by utilizing its expertise to assess data quality and collaborating in strengthening the quality of data centers' data.
- **Provide analytical support to selected evaluations.** Opportunities should be identified to apply data analytical techniques to the analysis of selected evaluations.
- **Collaborate on the creation of integrated databases.** Many potentially useful databases, available within an agency or organization, or from its country partners, are not utilized for evaluations because they have never been linked and integrated. The tools for creating these integrated databases are well understood and tested, and could be a practical way to strengthen evaluation capacity. This collaboration for integrating databases could also be considered by a few carefully selected larger scale operations in countries – those that have large volumes of under-utilized survey data from different sectors that could be integrated into an extremely useful resource for many evaluations and many different research agencies. India is often cited as one example of a country with huge under-utilized data potential, and Box 4's example of the program to combat human trafficking in the Philippines illustrates how previously untapped data could be integrated into a single data platform and used effectively. The Broward County Youth Protection Program in Florida case study in Box 5 illustrates that similar opportunities exist in countries such as the United States.

Integrated evaluation and research capacity development

Developing a common set of tools and approaches to the evaluation of development programs is essential for building the base for collaboration. At present, common understanding is largely lacking, as many evaluators are not familiar with the sources of big data or the analytic approaches, and, similarly, many data scientists do not use, and often are not familiar with, evaluation tools and approaches. Promoting the base of common understanding requires the incorporation of big data approaches into the training curriculum of evaluators and vice versa. This, in turn, requires collaboration to develop the common curriculum through:

- setting up workshops, conferences, and other forms of exchange to identify common approaches, areas of perceived disagreement, and practical tools that can be applied in the field
- inviting data scientists to contribute to evaluation journals (particularly on-line journals) and conferences, and vice versa for data science journals and conferences
- drawing lessons from pilot collaboration to assess what works and identify any barriers
- developing capacity through training for new professionals and for the staff at all levels of experience
- organizing exchanges and on-the-job training
- including a data scientist in the evaluation team – if the organization is sufficiently large – either as regular staff or as a consultant.

Critical need for landscaping research

Systematic documentation on how widely big data is used by evaluators is currently lacking, along with understanding of what has worked well and what are the challenges. At present, the few studies that have been conducted and mainly anecdotal evidence suggest a low level of data science utilization by evaluators who also present significant questioning of big data approaches. To expand from this level of anecdotal evidence, there is an urgent need to conduct basic landscaping research which calls for:

- documenting how effectively evaluation and data development centers coordinate or work together in different organizations and sectors
- filling in knowledge gaps on the levels of consensus, differences, and tensions in different sectors and organizations
- producing case studies on examples of cooperation.

Critical roles for funding and grant-making agencies

Given the great potential for convergence, combined with the slow rate of progress on the ground, funding agencies can play a critical role in creating the space for dialog and collaboration, and provide the seed funding in critical areas. Funding can provide the critical impetus in all of the steps for moving forward: bridge building, capacity developing, landscaping, and supporting pilot initiatives to implement convergence on the ground.

Glossary of technical terms

Algorithm	A set of rules or problem-solving operations followed by a computer program to make predictions, recommendations, and/or conclusions.
Artificial intelligence (AI)	Any device that learns from data to act, independent of humans. AI devices continuously gather data, analyze data, revise and update their predicted odds of success, and, accordingly, adapt and evolve their actions or rule sets on their own. AI can be embedded in software applications, machines, robots, and other devices that act in the world. AI is mimicking the cognitive functioning of the human brain, and as such, when combined with an engineered program, machine, or tool, AI can make decisions and act in the world just like humans.
Bayesian analytics	An analytic paradigm that uses prior information to assign a probability to any hypothesis of interest, updating these probabilities as more evidence becomes available.
Big data	Data which is generated very fast, huge in volume, and too large to be analyzed on a single computer. It is collected continuously, often over long periods of time and is non-reactive. The information collected is not affected by the process of analysis – which contrasts with most kinds of evaluation data where respondents may adapt their responses based on their perception of the purpose of the evaluation. The three sources of big data include: human generated (e.g. social media, internet searches, phone messages), administrative or transactional data (e.g. project reports, government publications and records); and geospatial data (e.g. satellite images, drones and remote sensors).
Big data ecosystem	The different processes and agencies or actors involved in the generation, filtering and distribution, regulation, use, and analysis of the different kinds of big data. The big data ecosystem operates in a very different way than the conventional evaluation ecosystem.
Black box	A subset of machine-learning algorithms (e.g. neural network and deep learning algorithms) that learn from data without being able to show how their conclusions were derived. The term is also used to refer to evaluations that only assess linkages between program inputs and outcomes without analyzing the processes of program implementation.
Concept mapping	A technique for developing scales and indices based on asking experts to classify words or statements into different categories. A computer program such as principle component analysis sorts the statements into groups, which are then converted into scales that can be used to rate programs or to compare changes over time.
Construct validity	An indicator or set of indicators used to define and measure often complex program outputs and impacts. An assessment is then made of how well the indicator(s) capture all of the dimensions of the construct (high or low construct validity).
Convergence	The joining of the theories, tools, and methods of two fields of data-driven inquiry – evaluation/social science and data science – for the purposes of increasing knowledge about what works to achieve positive social impact.
Data analytics	The application of statistical modeling or the training of machine-learning algorithms to derive meaningful insights from data.
Data mining	The iterative process of analyzing data until the resulting models align with the applied purpose they are to serve.
Data scientists	Statisticians and machine-learning experts who analyze data and build algorithms that can be used to learn, plan, or make decisions.
Data shaping	Software programs and platforms that streamline the processes of extracting, cleaning, transforming, and loading data into their final analysis tools or process. These tools can build a “recipe” so the analysis plan can be developed on the early data and then updated as new data arrives.

Deductive theories	Theories that are put forth as up-front hypotheses based on expertise, not raw data, whereby future experimentally designed data collection and analysis serve to prove or disprove the pre-supposed theories (hypotheses).
Experimental designs: Randomized controlled trials (RCTs)	The most rigorous evaluation design whereby future program participants are randomly assigned to different treatment groups – often a "no treatment" or a "treatment" group – in order to determine if a program truly makes a significant difference for those who participate in it versus those who do not. This is the only research design that can completely eliminate the potential for selection bias, as long as the population sample is representative of the target population. Researchers using experimental deigns are sometimes referred to as "randomistas" or "frequentists."
Frequentists	See experimental designs.
Global positioning system (GPS)	A radio navigation system that allows land, sea, and airborne users to determine their exact position. Mobile phones and other devices now use GPS to map communities, to locate areas of risk, or determine where maintenance is required, as well as for tracking the location of services such as food stores or clinics. GPS is also used in emergency and humanitarian programs to monitor movement and to locate groups such as refugees or groups force to flee to escape violence.
Human-generated data	Big data that is generated by humans, such as social media posts, internet searches, and phone calls.
Inductive theories	Theories that derive from the analysis and identification of patterns from prior or past data, and that also are updated and evolve as new data come in.
Integrated data platforms	Data platforms that make it possible to combine many different data sources (e.g. text, video, numbers, sound) into a format that makes it possible to compare different kinds of data. Often artificial intelligence (AI) will be used to find complex patterns of relationships in the data.
Internet of Things [IoT]	A network of remote sensors that measure and feed information into an internet network from devices embedded in appliances (e.g. a refrigerator or house thermostat) and vehicles, or from individuals (see human-generated data definition). As systems become more interconnected, IoTs begin to automate or regulate an increasing range of activities such as ordering food and domestic products, adjusting heating systems, and automating vehicles.
Natural language (processing)	A subfield of machine-learning algorithms created to process, analyze, and make meaning from natural language (text/narrative) data.
Machine learning	A branch of artificial intelligence whereby algorithms are trained to learn from data, identify patterns, and make decisions with minimal human intervention.
Propensity score matching [PSM]	A statistical matching technique applied to observational data that analyzes the observed (real-world documentation) effect of a treatment or program by controlling for conditions (covariates) – such as socioeconomic status, religious beliefs, address, community conditions, and historical family functioning – that a program cannot control but nonetheless will affect the outcomes that can be achieved. PSM attempts to reduce one of the biggest experimental threats to the validity of a study's findings: selection bias (see definition). As such, PSM is a technique for finding matched comparison groups of similar cases (segments of the overall population) based on sharing a common set of circumstances that make all group members equally likely to receive a treatment. When those in the same "propensity" group do and don't in fact receive different levels of treatment due to naturally occurring real world variances in delivery, a study can determine if these randomly occurring counterfactual experiences result in a statistically significant difference between those who got none, got some, got a lot, or got all of a treatment. PSM has been used in tens of thousands of peer-reviewed, published observational studies in all types of fields, with a predominance applied to research in medicine, health, economics, and education.

New information technology (NIT)	A term that encompasses big data, information and communication technology (principally data from mobile phones), and the Internet of Things.
Qualitative comparative analysis (QCA)	The unit of analysis for QCA is the case, rather than an individual subject. Cases may be individuals, households, organizations, or even countries. Every case must include the same set of variables which include attributes of the case. For example, this could be education, income for households, or indicators relating to democratic behavior or high or low performing education systems when the country is the case. It could also be an outcome or impact indicator. For example, is the country classified as democratic or is a school rated high or low on the inclusion of minorities. The analysis identifies the configuration of attributes that are linked to either the presence or absence of the outcome being studied. These configurations are defined as "necessary" or "sufficient" conditions for the outcome to occur or not occur. As the analysis is based on a configuration of attributes, it is considered a useful tool for the evaluation of complex programs.
Quasi-experimental designs (QEDs)	When it is not possible to use true experimental designs (such as RCTs) with random assignment of subjects to the project and control groups, QEDs are often used to match the project group with a comparison group. In some cases, it is possible to create a statistically strong comparison group using techniques such as propensity score matching (see definition), but, in many cases, only judgmental matching is possible. In this latter case, experts or stakeholders are asked to identify communities or groups that match the project group as closely as possible.
Quantified self	A movement that tries to incorporate technology into data acquisition on different aspects of a person's daily life. For example, data can be collected on food consumed, air quality, mood, skin conductivity, or blood oxygen level. The purpose is to increase a person's level of self-awareness.
Quantified community	A network of instrumented urban neighborhoods where a network of remote sensors collect, measure, and analyze data on physical and environmental conditions and human behavior to better understand how neighborhoods and built environments affect individual and social wellbeing.
Randomistas	See experimental designs.
Randomized controlled trials (RCTs)	See experimental designs.
Sample selection bias	Selection bias happens when the people selected to participate in a study aren't properly randomized and, therefore, are not representative of the overall population. Two examples of selection bias: i) using a first-come-first-served selection method, randomly assigning each person at the front of the line to a treatment or control group, where individuals with the proclivities to arrive early could skew (bias) the sample; or ii) after being randomly assigned to a program/treatment or not, participants without transportation to the program drop out (self-select out, after being "randomly selected" to be in the study), resulting in the treatment/program group being biased by attrition. In lay terms, selection bias can be explained as the reality that not everyone in the overall target population is equally likely to engage in, receive, or complete a program or treatment due to circumstances, factors, and conditions (covariates) beyond the control of the program or treatment. Selection bias is one of the biggest threats to validity for any program evaluation study, especially because program delivery in the real world is almost never "controllable."
Theory of Change (ToC)	A model that represents graphically all of the inputs, outputs, and, outcomes through which a program is intended to achieve its intended impacts. The model also defines a set of hypotheses that explain the processes through which inputs will be transformed into outputs, outputs into outcomes, and outcomes into impacts. A ToC can be used both as a tool to help program design and to provide a framework to design the evaluation. ToCs are widely used in the evaluation of international development programs.

References

Anderson, C. (2008) The end of theory: The data revolution makes the scientific method obsolete. Wired Magazine 6/23/08

Ashton, P., Weber, R. & Zook, M. (2017) The cloud, the crowd, and the city: How new data practices reconfigure urban governance? Big Data and Society. May 16, 2017.

Bamberger, M. (2017) Integrating big data into the monitoring and evaluation of development programs. UN Global Pulse with support from the Rockefeller Foundation.

Bamberger, M., Vaessen, J. & Raimondo, E. (2016) Dealing with complexity in development evaluation: A practical approach. Thousand Oaks, CA: Sage.

Bamberger, M., Raftree, L. & Olazabal, V. (2016) The role of new information and communication technologies in equity–focused evaluation: opportunities and challenges. *Evaluation*. Vol 22(2) 228–244.

Bamberger, M., Tarsilla, M. & Hesse-Biber, S. (2016). Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning* 55: 155-162

Bruce, K. (2017) How data analytics can help us to end modern slavery: Examples from the Philippines. Professional Development Workshop. American Evaluation Association. November 2017. Washington D.C.

Devex (2016) Inside Development: Data driven development needs both social and computer scientists, by C. Cheney. Devex, July 29, 2016.

Eubanks, V. (2017) Automating inequality: How high-tech tools profile, police and punish the poor. London. St. Martin's Press.

Felt, M. (2016) Social media and the social sciences: How researchers employ Big Data Analytics. Big Data and Society, research article Sage publications, available at journals.sagepub.com/doi/abs/10.1177/2053951716645828.

Forss, K. & Noren, J. (2017) Using big data for equity-focused evaluation: understanding and utilizing the dynamics of data ecosystems. In G.K. Petersson and J.D. Breul (editors) Big data and evaluation. London: Transaction Publishers.

Gay,K. & York, P. 2018. A New Way to Use Data. Precision care for better outcomes in psychiatric residential treatment for children. Scattergood Foundation. Available at scattergood-foundation.org/sites/default/files/A_New_Way_to_Use_Data.pdf.

Global Environment Facility (2015) Impact evaluation of GEF support to protected areas and protected area systems. 49th GEF Council meeting October 20–22, 2015. Washington D.C.

Gurumurthy, A. (2004). Gender and ICTs: Overview report. BRIDGE Cutting Edge Pack. Brighton: Institute of Development Studies.

Hashemi, S.M, Schuler, S.R. & Riley, A.P. (1996) Rural Credit Programs and Women's Empowerment in Bangladesh. *World Development*, 24(4):636–653.

Hussain, F. and Amin, S. (2018) I don't care about their reactions: agency and ICTs in women's empowerment in Afghanistan. *Gender and Development* (26(2): 249-266.

Kabeer, N. (2005) Gender equality and women's empowerment: a critical analysis of the third Millennium Development Goal. *Gender and developement.* 13(1): 13-24.

Kontokosta, C. (2012) Quantified Community: The QC Sensor Array. Available at urbanintelligencelab.org/quantified-community

Letouzé, E., Areais, A. & Jackson, S. (2016) The evaluation of complex development interventions in the age of Big data. In Bamberger, Vaessen and Raimondo. Op cit.

Meier, P. (2015) Digital Humanitarians: How big data is changing the face of humanitarian response. CRC Press:Chapters 5, 7.

New England Journal of Medicine (2000a) A Comparison of Observational Studies and Randomized, Controlled Trials, by Benson, K. & Hartz, A. *NEJM 2000.* Available at www.nejm.org/doi/full/10.1056/NEJM200006223422506.

New England Journal of Medicine (2000b) Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs, by Concato, J., Shah, N. & Horwitz, R., *NEJM 2000.* Available at www.nejm.org/doi/full/10.1056/nejm200006223422507

O'Donnell, A. & Sweetman, C. (2018) Introduction: Gender, development and ICTs. *Gender and Development 2018.* Vol. 26 No. 2, 217–219. Special number of ICTs.

O'Neil, C. (2016) Weapons of Math Destruction: *How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

Patton, M.Q (2008). Utilization Focused Evaluation 4th ed. Thousand Oaks: Sage.

Pearl, J. & Mackenzie, D. (2018) *The Book of Why: The New Science of Cause and Effect*. New York, NY. Basic Books, Inc.

Petersson, G.K. & Breul, J.D. (editors) (2017) *Cyber society, big data and evaluation. Comparative policy evaluation.* Volume 24. Transaction Publications.

Popp, T. (2017) Black Box Justice, *The Pennsylvania Gazette:* Sept/Oct 2017. Available at thepenngazette.com/pdfs/PennGaz0917_feature1.pdf.

Raftree, L. (in press) Exploring the soft side: Ethics, inclusion, bias and privacy in ICT-enabled evaluation. ICT4Eval: Threats and opportunities. IFAD.

Raftree, L., Appel, K. & Ganness, A. (2013) Modern Mobility: The role of ICTs in child and youth migration. Plan International USA and Oak Foundation.

Rose, T. (2016) *The End of Average: How We Succeed in a World that Values Sameness*. San Francisco, CA, HarperOne.

Salganik, M. (2018) Social research in the digital age: Bit by Bit. Princeton: Princeton University Press.

Schwartz, M., York, P., Nowakowski-Sims, E. & Ramos-Hernandez, A. (2017), Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County Experience, in Children and Youth Services Review, 81: 309-320. Available at www.sciencedirect.com/science/article/pii/S0190740917303523.

Stuffelbeam, D (2001) Evaluation models. New Directions for Evaluation, No. 89. San Francisco: Jossey-Bass.

Siegel, E. (2013) *Predictive analytics: The power to predict who will click, buy, lie or die*. New Jersey: Wiley Foundation.

Stern, E., Stam, N., Mayne, J., Forss, K., Davies, R. & Befani, B. (2012) Broadening the range of designs and methods for impact evaluations. Department for International Development. London.

Thakur, D. (2018) How do ICTs mediate gender-based violence in Jamaica? *Gender and Development,* 26(2). July 2018.

The Economist.(2018) AI-Spy. The workplace of the future. Special Report: AI in business. March 31, 2018.

UN Global Pulse (2012) Big Data for Development Challenges and Opportunities. Article by E. Letouzé, May 2012.

United Nations Population Fund (2016) Big data and the sustainable development goals. Background paper prepared jointly by UNDP, UNFPA (coordinator), UNOPS, UNICEV, UN WOMEN and WFP (coordinator) for the Joint Meeting of the executive boards of UNDP/UNFPA/UNOPS, UNICEF, UN Women and WFP June 3, 2016 (unpublished).

Vaessen, J., Rivas, A. & Leeuw, F. “Microcredit and women’s empowerment: Complexity in systematic review” in Bamberger, Vaessen and Raimondo, 2016: Dealing with complexity in development evaluation: A practical approach. Thousand Oaks, CA: Sage.

Wakanuma, K.J. (2013) New media technology, gender and the power paradox of women’s empowerment in K.Prasad (ed) *New media and pathways to social change: shifting development discourses* (pp. 314–362). Delhi, India. B.R. Publishing Corporation.

Wolf, G. (2015) The quantified self. Ted Talk, available at: www.ted.com/talks/gary_wolf_the_quantified_self?language=en.

World Bank (2016) *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank.

©Hadynyah, 2018

Supported by

The ROCKEFELLER FOUNDATION

www.rockefellerfoundation.org