Duke
NICHOLAS INSTITUTE
FOR ENVIRONMENTAL POLICY SOLUTIONS

Policy Brief
May 2020

Nicholas Institute for Environmental Policy Solutions nicholasinstitute.duke.edu

Harnessing Data Analytics to Accelerate Energy Access: Reflections from a Duke-RTI Convening on Data for Development

Rob Fetter and Justin S. Baker

CONTENTS

Author Affiliations
Rob Fetter, Senior Policy Associate, Duke University, Energy Access Project. Justin S. Baker, Senior Economist, Center for Applied Economics and Strategy, RTI International.

Citation
Fetter, Rob, and Justin S. Baker. 2020. “Harnessing Data Analytics to Accelerate Energy Access.” NI PB 20-03. Durham, NC: Duke University. http://nicholasinstitute.duke.edu/publications.

Acknowledgments
The authors thank the participants of the event on “Data for Development: Using Data Analytics to Accelerate Global Energy Access” in Washington, DC, as well as Andrew Allee, Kyle Bradbury, Madeleine Gleave, Julian Glucroft, Marc Jeuland, Jordan Malof, Arif Mamun, Emily McAteer, Jonathan Phillips, Jay Rineer, James Smith, Dan Sweeney, John Tkacik, Melia Ungson, and Nathan Williams for insightful comments and suggestions.

Publication Number: NI PB 20-03

Executive Summary

One of the defining features of our current era is the proliferation of innovative technologies that constantly generate data and information. Earth observation satellites, ground-based tools such as vehicle-mounted cameras, smart meters, and crowdsourced platforms all collect and gather data with applications for the energy sector. Whether such data can be fully utilized for accelerating access to energy and its myriad services, however, remains an open question, as the usefulness of these applications is hindered by data-sharing and coordination challenges, market power, and, ironically, lack of data relevant to remote and underserved populations.

This document presents a vision and summary of how development and scaling of new tools, and application of big data analysis principles, have the potential to transform energy systems planning, policy design and implementation, and investment decisions. Drawing from ideas presented in a December 2019 workshop co-organized by the Energy Access Project at Duke (Duke EAP) and RTI International, it describes a set of applications that include, for example, assessment of existing energy infrastructure, demand, and consumption; electrification system planning; measurement of electricity reliability; and characterization of customer types, willingness and ability to pay, and identification of locations or populations with the greatest potential for productive uses of energy. The methods presented have broad relevance and could be extended to cover many

problems inhibiting progress on energy access, thereby revolutionizing the way the sector and its various stakeholders—policymakers, donors, investors, private companies, and consumers—do business.

This brief summarizes the novel methods presented in the Duke-RTI convening, which serves to highlight a diverse set of applications in the energy access domain, and comments on how big-data methods can most effectively complement traditional data collection methods, e.g., population-based surveys. It then lays out a vision for a data commons that would assemble key stakeholders united around goals of increasing energy data accessibility and interoperability with clear energy access use cases prioritized to meet specific needs. This idea is motivated by the observation that despite so many promising applications, significant challenges and barriers remain, related to data availability or usability; issues of interoperability; unclear provenance and quality; decentralization; and unclear applicability to the most pressing policy problems. Stakeholders would work to co-create an energy data research agenda based on specific real-world priorities by promoting standards and establishing a coordinated network of analysts and practitioners. The network of stakeholders would also facilitate improved access to data and insights through curating databases, providing online visualization tools, and disseminating results to decision makers and the wider public.

A public data commons could provide broad public benefits. However, the financial sustainability of such a platform, as well as limits on how much data individual companies or public utility operators are willing or able to release into the public domain, represent practical limits on how far such an approach can extend. Thus, we also reflect on the possibilities that arise from restricted-access data sharing that would govern access to some types of data (e.g., protecting privacy and confidentiality). For example, we suggest a role for restricted-access data competitions, whereby data providers offer snippets of data to researchers, solicit proposals in an open competition, then grant complete access to a limited group of winning teams. Such arrangements benefit data providers (allowing them to understand how data could be used to generate new insights), decision makers and donors (who could sponsor competitions in exchange for policy-relevant advice), and researchers (who gain by contributing their expertise to real-world problems and gaining prestige from publications related to these unique data applications).

We conclude with recommendations for focused investment by donors, researchers, and practitioners, highlighting four domains that would benefit from additional resources to support collaboration, research, and integration into practice:

(1) Investments in new data sources and algorithms to facilitate interpretation.

(2) Integrating traditional surveys, such as the World Bank's Multi-Tier Framework survey, with remote sensing and big-data approaches.

(3) A managed data commons to advance practice on emerging and heretofore unidentified topics.

(4) Restricted-access data sharing that matches data generators with users, protects confidentiality and privacy, and creates a living record of a wide range of energy access data projects.

INTRODUCTION

The United Nations' Sustainable Development Goal 7 (SDG7) highlights the importance of access to affordable, reliable, sustainable, and modern energy for widespread and sustainable economic development. Accurate and timely data on energy services, outcomes, and willingness-to-pay is essential for monitoring progress, evaluating the success or failure of interventions, and making decisions about what will ultimately amount to hundreds of billions of dollars in investments. Data from traditional administrative sources—such as periodic population surveys—provide some insights but suffer from important limitations: data collection and compilation take months to years, standardization of measures can be difficult, and some items are notoriously hard to track (e.g., energy use over time) using standard surveys.

Data analytics and machine learning offer great promise for helping to fill knowledge gaps that emerge from reliance on traditional data sources. For instance, the application of machine learning to satellite imagery enables scalable, high-frequency and spatially well-resolved data collection on topics such as access to grid power, the presence of solar panels, air pollution concentrations, and biomass stocks. Similarly, smart meter data can help pinpoint the extent and effects of outages or provide insights for on- and off-grid utilities seeking to reduce peak use and balance demand across the day. Many other applications are possible as the cost and flexibility of data collection increases, and object recognition algorithms and other machine learning methods grow ever more sophisticated. Some data sources have already been well used by the energy sector and energy researchers: for instance, nighttime lights are commonly used to assess electricity access, and satellite imagery is being processed to detect distributed generation infrastructure (Rolnick et al. 2019, Malof et al. 2016, 2019). Other data sources, such as street-view images recorded by vehicle-mounted cameras, are more emergent, with energy sector applications still in early stages of development.

Recognizing that we have barely scratched the surface on useful applications of data analytics and machine learning for energy access, the Energy Access Project at Duke University (Duke EAP) and RTI International convened a day-long workshop on the topic in Washington, DC, in December 2019. The Duke-RTI convening—"Data for Development: Using Data Analytics to Accelerate Global Energy Access"—brought together key players from research and practice communities working on novel data collection methods and machine learning tools for improving energy access in the developing world.

This brief summarizes key takeaways from this convening and implications for this type of work going forward. We begin by describing novel methods of data collection and interpretation, highlighting their potential to contribute to enhanced understanding of the energy access landscape. We then discuss additional applications of big-data methods, including how they can be integrated with conventional survey methods. Next, we reflect on the value of creating a data commons that would advance energy data accessibility and interoperability, focusing on prioritized use cases that are based on specific needs. We conclude with recommendations for focused investment by donors, researchers, and practitioners, highlighting four domains that would benefit from additional resources to support collaboration, research, and integration into policy and practice.

WHAT MAKES DATA BIG?

The terms "big data" and "data analytics" have become ever more common over the recent past, but it is difficult to define precisely what makes data "big," or how to distinguish data analysis from data analytics. Many users define big data as referring to a volume or complexity of data that is difficult to process using standard analytical techniques and hardware. In line with this notion, the term data analytics typically refers to all of the processes that are necessary to manage big data sets, including collection, assembly and organization, and storage. Data analytics thus also includes data analysis: that is, the process of examining, transforming, and arranging a dataset in specific ways or using standard and reproducible algorithms to identify and study features and generate useful insights. A distinguishing feature of most big data is that the marginal cost of collecting additional observations is very low, once the infrastructure is in place to do so (although the fixed cost of setting up the infrastructure may be very high).

NOVEL METHODS FOR DATA COLLECTION AND ANALYSIS

The experts who presented at the Duke-RTI convening discussed innovative methods for scalable data collection and analysis including from satellites, ground-based sensors, smart meters, and crowdsourcing. They highlighted the value of these tools for all phases of project planning, design, and development. Specific applications described included baseline assessments of energy infrastructure and energy consumption, electrification system planning, measuring the reliability of electrical power, characterizing the creditworthiness of residential or enterprise customers, and identifying which neighborhoods or customers are most likely to engage in the productive economic use of electric power. The methods have broad relevance to many problems facing those intervening to accelerate energy access—policymakers, donors, investors, and private companies—and could thus be transformative for the sector.

Satellite Data and Automated Object Identification

Over 150 Earth observation satellites are currently in orbit, carrying sensors that measure visible light and nonvisible (e.g., infrared) regions of the electromagnetic spectrum by day and night. Most satellites carry "passive" sensors that capture reflected or emitted energy from the Earth's surface or atmosphere, while some newer satellites also carry "active" sensors that beam energy toward Earth and record the reflected or backscattered signal (Tatem et al. 2008). Whole scientific journals are devoted to sharing and developing knowledge about remote observation, and numerous researchers have used remotely observed data to characterize elements of energy access and socioeconomic status, among other things. Perhaps the best-known application is the "lights at night" dataset, which is based on the idea that access to electricity and greater economic activity correlate positively with the amount of lighting on structures and streets. Scientists also use satellite imagery to identify built infrastructure, including that for electricity generation, transmission, and distribution, and to characterize visible landscape features in ways that help indicate energy use. For instance, circles of green vegetation surrounded by brown land commonly indicate areas of crop irrigation, which typically requires energy to pump water and thus signifies access to some form of energy to drive pumps.

Several presenters at the December 2019 event described efforts using satellite data for advancing SDG7, with applications including baseline assessment, monitoring and evaluation, and improved targeting. Jordan Malof (Duke University) presented an overview of work from the Duke University Energy Data Analytics Lab, where researchers are working to create tools for automated mapping of energy infrastructure— including assets for generation (solar photovoltaic panels) and transmission and distribution (towers, lines, and substations), as well as generating high-resolution estimates of electricity access based on nighttime lighting, land cover, and other signals. Malof also spoke about estimating building-level energy demand using satellite imagery, based on estimated building footprints and other information, which could assist planners seeking to better predict electricity demand.

Nathan Williams (Rochester Institute of Technology) presented several projects in progress under the rubric of the e-GUIDE (Electricity Growth and Use in Developing Economies) initiative. Launched in January 2019, e-GUIDE leverages electricity consumption, nightlights, and other data sources, and focuses on four aspects: predicting electricity consumption, co-planning electricity and agricultural infrastructure, improving energy system planning and operation, and mapping power quality and reliability. Williams presented several ongoing projects on these themes, including work from Chris Elvidge and Kim Baugh (Colorado School of Mines) that uses nighttime lights data to assess grid reliability, and a project led by Simone Fobi (Columbia University), Jay Taneja (University of Massachusetts Amherst), and Vijay Modi (Columbia) that works with satellite imagery—landscape and built environment features—to predict electricity consumption. This work builds on high-quality, spatially granular consumption data in Kenya (Fobi et al. 2018) and attempts to discern how satellite-observable features such as roads, building footprints, and land cover correlate with consumption patterns. The ultimate aim is to create a tool that uses such correlations to create a predictive model for latent electricity demand, in settings where detailed consumption data are not available. Since high-resolution satellite imagery is available for virtually all of the globe, such a tool would offer invaluable support for donors and national governments aiming to select the most appropriate electrification technologies (e.g., grid versus stand-alone systems) in their settings, and to optimize system design and siting. As Williams emphasized, sophisticated electricity planning tools are being developed by leading academic institutions, but the reliability of their outputs is limited by a lack of high quality, spatially resolved demand data.

On a similar note, Melia Ungson presented work by Fraym, a company whose core product is high-quality, fine-resolution data on social and economic characteristics produced using machine learning of satellite imagery, household survey, and other data sources. National statistical offices conduct periodic household and business surveys, often in collaboration with international agencies such as the World Bank and USAID.[1] These surveys serve as the backbone for many analyses conducted by researchers in public health, economic development, and related fields, and are critical for donors and governments seeking to develop programs and infrastructure. However, such survey data are rarely representative at spatial scales below the country or region even in the rare instances where they contain geographic identifiers (which are often masked due to confidentiality concerns), challenging their usefulness for many desirable applications. One of Fraym's innovations is to "downscale" household survey data using remotely sensed data and applied machine learning algorithms. As Ungson explained, the company uses data on a huge range of features—including Earth observation data, gridded population information, and biophysical surfaces—and develops machine learning algorithms to create spatial layers from household survey data. The methodology builds on existing methods for interpolating spatial data, and Fraym has performed validation exercises using census data. Fraym's spatial layers include estimates of socio-economic characteristics, including access to electricity and ability-to-pay for electricity services, at a high spatial resolution. Ungson shared a recent example in which Fraym provided granular information on electricity access across Côte d'Ivoire, illustrating how the algorithms can provide clarity on electricity connections as well as energy needs in unconnected communities close to grid infrastructure. Fraym also uses data analytics to identify and map potential customers, thereby helping private firms, investors, and donors to segment customers and set priorities for targeting investments.

Ground-Based Sensing

While Earth observation data allow identification of infrastructure such as rooftop solar and electricity distribution systems, satellite imagery itself has limitations: the best commercially available imagery has a resolution of about nine pixels per square meter, which is insufficient to accurately identify small objects (e.g., many PV panels).[2] Cameras mounted on aerial vehicles (planes or drones) can collect higher-resolution data, though both options are relatively costly, and flying aircraft may not be politically feasible in certain regions.

Another option is ground-based sensing, such as vehicle-mounted cameras on cars. This technology is able to capture images of energy and related infrastructure with very high resolution. Google Street View is perhaps the most commonly known example of ground-based monitoring; it provides global street-level imagery, though usually with a significant time lag between the times of image capture. Such vehicle-based cameras can be deployed in most human settlements worldwide, from dense urban spaces to sparsely populated rural settings, and—when combined with object identification algorithms—can be used to conduct rapid asset inventories. Jay Rineer (RTI International) presented on RTI's Corylus tool, which facilitates rapid asset inventories by using object identification algorithms to analyze 360-degree vehicle-based and smart phone imagery through Mapillary, a street-level imagery platform that aggregates crowdsourced and georeferenced photos. This work remains in development, but offers significant potential, in many contexts, for baseline characterization or before-and-after assessments aimed at evaluating energy-related program implementation.[3]

[1] The Living Standards and Measurement Survey (LSMS) and the Demographic and Health Surveys (DHS) are two examples.

[2] This is commonly referred to as 0.3m resolution, meaning that an individual pixel represents a square that is 0.3m on a side.

[3] While ground-based images theoretically offer a dramatic increase in the ability to automatically identify infrastructure, interpreting these images presents substantial challenges due to differing camera perspectives (angle from lens to object) and camera resolution. State-of-the-art deep learning methods are sensitive to changes in resolution and camera perspective, and so novel methods must be developed to interpret such ground-based images.

Smart Meters

Smart meters automatically record data on electricity consumption and other information, and communicate those data to utilities or other system managers, enabling remote observation of customer use. Depending on the model and application, smart meters can collect data as frequently as multiple times per second, although in most applications they are set to record information every 15 to 60 minutes. Smart meter installations have increased dramatically over the past decade: in the U.S., for example, about 79 million smart meters had been installed by 2017 (EIA 2018). System operators in developing countries are increasingly using this technology as well, partly due to a belief in their usefulness for reducing commercial losses. Several speakers at the Duke-RTI convening highlighted useful energy access applications of smart meter data.

Dan Sweeney, of the Renewable Electricity and Energy Efficiency Partnership (REEEP), presented on REEEP's efforts to assist the Swedish international development agency (SIDA) with a results-based financing (RBF) initiative that would measure the impacts of providing new electrical connections to households in Zambia. Serving as the implementing agency for SIDA's Beyond the Grid Fund, REEEP built a custom tool—the Energy Data and Intelligence System for Off-grid Networks, or EDISON—to monitor the consumption patterns of over 110,000 newly connected households. The solar home system (SHS) and minigrid connections were funded by grants from SIDA and REEEP, who required in exchange that the four participating companies share at least daily updates on customer consumption and transactions. The data collected through EDISON—which also includes customer profiles and locations, and technical engineering aspects relating to energy services—allows careful tracking of customer payment behavior, power demand (load), and power production.[4] Companies connected their databases to an automated interface to allow REEEP to analyze the usage data in EDISON. The primary purpose is to monitor the RBF initiative, providing accountability for the publicly financed subsidy and ensuring that connections remain live and in use. As Sweeney noted, this helps to (1) demonstrate that customers value the energy services they receive (which is vital for future market development), and (2) ensure that companies offer reliable equipment alongside maintenance or repair services as needed. REEEP and SIDA are also using the system, along with periodic surveys, to measure impacts on employment, productive use, displaced emissions from generators, and to improve customer targeting for future efforts.

As the EDISON example demonstrates, there is both a demand and capacity for processing and analyzing large scale energy data in the development sector. As SIDA, with assistance from REEEP, prepares to expand the program beyond Zambia, smart meters are well suited to provide even more precise data for service providers, investors, and donors to understand consumption patterns across multiple users.

Further demonstrating the value of integrating centralized tracking systems with smart meters, Emily McAteer (Odyssey Energy Solutions) presented her company's work to provide platforms for minigrid and SHS providers to manage data across the portfolio lifecycle. Odyssey's integration feature, which collects consumption and payment data directly from SHS back-ends and smart meters on minigrids, helps developers and donors to assess site feasibility, monitor impacts to support RBF, and provide real-time insights on customer usage and payment patterns. Odyssey is working with donors and DFIs such as the World Bank to streamline RBF efforts, and supporting sector initiatives with partners such as the African Minigrid Developers Association (AMDA) to create a sector-wide data platform that will share key impact elements with the public, thereby promoting industry knowledge and project investment.

As discussed above in the section on satellite imagery, a major challenge for minigrid developers and investors—where smart meter data can also be leveraged—is the accurate prediction of electricity consumption in communities that have yet to benefit from access to power. Developers must typically design and engineer their systems to meet an unknown demand, and investors face significant uncertainty about revenues that will be generated from electricity sales. Bottom-up approaches to predicting consumption using assumptions about

[4] For SHS users, power consumption and production is not measured directly, but rather is imputed based on whether the system is active (i.e., whether the user is making payments), the system specifications, and solar insolation data.

appliance use and behavior often overpredict electricity consumption by a large factor, leading to wasteful capital outlays and economically unsustainable business models. Andrew Allee (Rocky Mountain Institute and Dartmouth College) presented work carried out in collaboration with Paulina Jaramillo (Carnegie Mellon University) and Nathan Williams to improve the prediction of electricity consumption in to-be-electrified communities. Machine learning algorithms that relate customer characteristics (collected using surveys) to smart meter data on electricity consumption can improve accuracy of demand predictions by an order of magnitude. This leads to more appropriate sizing of minigrids and lowers energy costs to consumers. The same machine learning model results can also be used to streamline future surveys of prospective customers by informing companies which survey fields are most useful for predicting demand. Furthermore, the ability to estimate prediction accuracy aids investors in assessing risk.

Robyn Meeks (Duke University) presented research on a somewhat different, but also potentially groundbreaking, application of smart meters. Working with a utility in the Kyrgyz Republic in central Asia, her team installed smart meters in households throughout the distribution network to assess whether improved monitoring helps to improve service quality and reduce electricity theft. When utilities receive consumption data through smart meters and bill customers accordingly, this removes human meter readers from the billing system and reduces opportunities for theft. Some smart meters also permit utilities to disconnect nonpaying households remotely, which reduces payment enforcement costs. Smart meters also provide valuable information to customers and utilities about electricity outages and voltage spikes. Without smart meters, for example, utilities may have incentives to under-report outages and spikes, while customers may over-report them, but smart meters provide both parties with the same transparent information so that negotiations over service quality can proceed more easily. In preliminary analysis, Meeks and her colleagues find that installing smart meters results in better cost recovery for the utility, a reduction in voltage spikes, and more frequent utility maintenance and repair, but that they have no impact on outages.

Crowdsourcing

Citizen science and crowdsourcing initiatives can help fill critical data gaps relevant to a wide range of sustainable development goals (Fritz et al. 2019). Because of the expense and logistical difficulty of conducting more traditional household surveys, such surveys are typically conducted only every five to ten years; crowdsourcing can enable more frequent data collection. Such initiatives could be used to delineate energy infrastructure and document consumption patterns, for example, in ways that would supplement traditional data sources. Applications to date have been rare, however, except for collection of valuable information on energy consumption from the grid (Flostrand et al. 2019). Similar efforts for off-grid energy consumption and technology adoption by citizen scientists could inform planning of electrification strategies and private investors' planning, by documenting and verifying current energy access and consumption behavior in remote rural areas. Verification could include geo-referenced smartphone images of household energy technology (e.g., solar panels or diesel generators).

Integrating Traditional Surveys

Although the tools described above offer promising improvements to improve energy access, they cannot fully replace traditional surveys. Administrative data such as the LSMS and DHS surveys provide foundational insights for many plans, programs, and infrastructure projects. These surveys have been subject to critique (Jerven 2013), but nonetheless typically remain the first source of trustworthy information that practitioners and researchers can use when scoping a project or comparing trends over time or across space. Of particular interest for energy access, the World Bank's Multi-Tier Framework (MTF) data series, which Bryan Koo described at the Duke-RTI convening, provides high-quality survey data on the availability of energy services in a dozen countries, with plans to include additional countries in the coming years. The MTF surveys provide many more energy-related variables than have been available in the past, and permit comparisons across countries by including common questions across multiple countries.

Researchers and practitioners also routinely conduct targeted surveys that are specifically designed to support evaluations of particular projects or planning processes. These surveys are idiosyncratic in both breadth and depth, and are crafted to produce both quantitative and qualitative insights that support monitoring and evaluation as well as market assessments or that are used for exploratory research purposes. They are also usually costly to deploy, though mobile phone-based data collection is reducing costs. Garlick et al. (2019) provide evidence that high-frequency phone interviews of microenterprises produce data on business outcomes that is of comparable quality to that obtained from less frequent in-person interviews, although they find evidence of potentially increasing inaccuracy over time. This suggests that phone surveys may be better for measuring broad-scale effects while in-person interviews may be superior for assessing within-enterprise dynamics and changes.[5] The private firm 60 Decibels has also advanced the design and implementation of "lean survey" methods using brief phone- or text-based surveys. This allows for a lower-cost and more scalable approach with response rates and accuracy that is comparable to in-person surveys, albeit for a limited number of questions. One limitation is that respondents are limited to those with mobile phones, but this limitation is lessening as this technology achieves wider penetration.

ADDITIONAL METHODS AND APPLICATIONS OF DATA ANALYTICS

Downscaled Administrative Data

Above, we discussed some microlevel household datasets that are available publicly (e.g., LSMS and DHS), but most administrative data based on household surveys are only available in highly aggregated or averaged form. This limits the spatial, temporal, and activity-scale information present in the data, and limits the extent of potential applications of those data. Recent statistical and geospatial applications—like those deployed by Fraym—are breaking new ground by transforming aggregated administrative data to a more spatially-explicit scale. For example, synthetic population, a method for translating aggregate statistics to a smaller subpopulation scale, offers the ability to populate existing households (with known locations) with demographic characteristics that represent sample populations. Synthetic population, when combined with agent-based modeling and other analytical techniques, can inform analysis of technology adoption (Shafei et al. 2012), energy demand (Bustos-Turo et al. 2016), or infrastructure development, though limited energy access research to date has applied such methods.

Similarly, downscaling administrative data to a household scale creates opportunities to combine downscaled information with physical data to conduct spatial assessments. For example, integrating off-grid energy generation potential with different technologies (based on physical factors) with demand (using downscaled administrative data). Furthermore, downscaled administrative data can be linked with financial data on contract terms and repayment of in-home off-grid solar or other technology. Researchers at Nithio have recently developed approaches to assess the creditworthiness of individual homes based on location, demographic characteristics and other factors, to inform private sector investment for in-home energy service offerings. At the Duke-RTI convening, Madeleine Gleave, Chief Data Scientist at Nithio, discussed an application that links disaggregated household characteristics from Fraym with utility- and household-scale information on consumer adoption of in-home solar, including pricing, energy utilization, loan repayments, and payment delinquency. The machine-learning powered application can be integrated with existing sales applications, customer relationship management platforms, or planning tools to assist with underwriting, portfolio management, or customer targeting. Third-party firms such as Nithio have the ability to assemble data from multiple private firms and public utilities and develop those data into sophisticated analytical products and services—allowing practitioners to obtain actionable and externally validated insights from a wider universe of information than their data alone could provide.

[5] Though not definitive, their analysis suggests that the variation within enterprises over time results from "social desirability bias," which refers to respondents' desire to report outcomes they believe are more "acceptable" to the enumerator.

Demand Modeling and Electricity System Planning

Electricity and energy systems models are widely used for infrastructure development planning and prospective policy analysis. Modeling and scenario analysis for electricity planning purposes is typically conducted relative to a baseline scenario that reflects current electricity system infrastructure, generation capacity, and energy demands by sector. Improved or updated information on existing energy sector infrastructure allows for a more robust representation of baseline conditions in such models. This helps to avoid the risk of misleading baseline assessments or misguided policy recommendations. Moreover, several communities of practice would benefit from improved information on existing electricity infrastructure and consumption, including the Energy Modeling Forum (Fawcett et al. 2018). Furthermore, recent advances in electricity sector modeling include greater spatial representation of energy supply and distribution systems (Wu et al. 2017) and demands delineated by household type (Diawuo et al. 2019). Spatially explicit and downscaled administrative data outlined in this policy brief can be linked with this new generation of planning models to offer greater insight on the potential benefits and trade-offs of different policy and technology pathways for increasing electricity access.

Of course, the benefits of access to electricity and the financial sustainability of rural electricity infrastructure depend critically on infrastructure systems and services outside the energy sector. For example, electricity has the potential to transform agricultural productivity in rural Africa and generate levels of electricity demand that ultimately reduce energy costs and improve cost recovery for utilities. Nathan Williams presented work under the e-GUIDE Initiative, carried out in collaboration with Paulina Jaramillo and Jorge Izar (Carnegie Mellon University), that is using remotely sensed data on climate, soil, and water resources to identify opportunities where co-planning of electricity and small scale irrigation infrastructure could result in large productivity gains and improved utility sustainability. Bryan Koo presented similar work by the World Bank's Energy Sector Management Assistance Program (ESMAP) that uses data on biophysical conditions, climate, and agricultural yields in conjunction with downscaled administrative data to develop energy demand models for irrigation and post-harvest processing activities. The unifying idea in both applications is that understanding energy demand in the agricultural sector is critical to providing the right energy solutions for rural communities, and that improving agricultural productivity could help develop more resilient rural economies and societies.

TOWARD AN ENERGY ACCESS DATA COMMONS: BENEFITS AND PROSPECTS

Despite the many promising applications of big data and data analytics to enhance energy access policy and investments, a number of challenges remain. These include gaps in data availability or usability (data are not always adequate for users' needs, and some data can be difficult to access and use); issues of interoperability (energy data often lack standardization, making it difficult to build standard models that reach across developers and countries because the data collected is not uniform, and linking datasets often requires expert knowledge); provenance (not all data are equally trustworthy, and frequently lack metadata); prioritization (there is no definitive source on priority use cases); and decentralization (data sources are often dispersed over too many separate data caches to be readily accessible). All prospective users face these challenges, but they are particularly acute for researchers and practitioners in developing countries, who also face structural barriers to full participation in communities of research and practice. These barriers often include, for instance, higher costs to obtain specialized training, reduced interaction with experts through formal and informal networks, and less access to funding (e.g., for professional skill development, travel to conferences, or data purchases).

To respond to these interrelated challenges, Duke experts (including Jordan Malof, who introduced the idea at the December 2019 event, and Kyle Bradbury, Managing Director of the Energy Data Analytics Lab at Duke) have proposed the creation of a Global Energy Data Commons (GEDC). The Duke team—along with partners at the World Resources Institute, the Electric Power Research Institute, and the National Renewable Energy Laboratory—received a grant in 2019 to advance the vision for a GEDC as an open knowledge network under the National Science Foundation Convergence Accelerator Program. The project has three primary goals: to inform an energy data research agenda based on priorities identified by prospective users with real-world problems; to enhance data interoperability through the use of consensus standards and a more closely coordinated research network; and to reduce barriers to data access and insights through curated databases and online tools for data access and visualization.

Malof described how the GEDC team has been assessing user needs and contributions of potential collaborating partners, with an eye toward identifying specific use cases such as increased power sector resilience. The GEDC collaboration also aims to proactively encourage contributors of data to conform to specific standards that enhance easier access, use, and interoperability, with the overall aim to bring together partners with specific needs and unanswered questions with experts in developing tools for gathering, analyzing, and visualizing data could help.

Challenges to Sustaining an Energy Access Data Commons

Energy access practitioners and researchers could also benefit from such a data commons, especially one focused on use cases that would advance the particular needs of developing country stakeholders, such as tracking the progress of SDG7 or accurately forecasting electricity demand in off-grid communities. Among the challenges that a data commons would face, two are arguably critical: the need to create a sustainable, long-term funding model, and the difficulty of gathering privately held data or information that businesses or governments may prefer to keep confidential.

Models for sustainable financing. Although startup processes can be funded with grants, financing public knowledge goods over the long term can be a challenge. Three broad models exist for creating a self-sustaining program. The first is a "freemium" model, in which basic data or tools are free, but full access is granted only to paying subscribers. For instance, subscribers might get exclusive access to more recent data, the most geographically granular data, application programming interface (API) access (as opposed to a point-and-click interface), downloadable data, or specialized visualization and data processing tools. A second model is to provide data and tools on a fee basis, with a sliding scale, so that all users must pay for access, but some organizations—academic organizations, for example, or institutions based in developing countries—can request access at reduced or no cost. A third option is to fund commons-related organizational activities from implementation contracts with public entities that are mandated to collect data for tracking purposes. For instance, REEEP developed its EDISON tracking tool as part of a contract with SIDA to implement their RBF program in Zambia. Another example of this option, used in water resources, comes from the Internet of Water team at Duke's Nicholas Institute for Environmental Policy Solutions (NIEPS). This group, which advances interoperability and data standards for water quality data in the U.S., enters into service agreements with federal agencies—which are mandated to collect some monitoring data from U.S. states—to help state agencies set up data caches and APIs in ways that facilitate federal use.

Gathering sensitive information. Another challenge to building a data commons is that entities with access to some of the most valuable data, such as energy consumption, may not want or be able to make those data public, for reasons of privacy protection or to protect competitive advantage. The EDISON project funded by SIDA and implemented by REEEP provides a notable example of how donors can require beneficiaries of subsidies to provide more detailed data than usual, but even so, privacy requirements mean that even anonymized EDISON microdata cannot be put in the public domain.

Given these challenges, it seems clear that a data commons would serve as a complement to, but not a replacement for, commercial initiatives such as those advanced by Fraym, Nithio, Odyssey, and other companies offering data analytics services. It could facilitate better and more efficient decision-making for public planners, donors, investors, and others, and could help to overcome the structural barriers faced by researchers and practitioners in developing countries who stand to benefit from access to data that is more interoperable, better documented, collected or referenced in a central location, and includes easier access to analytical and visualization tools. But ultimately there are limits on how far a public-goods approach can extend.

The Role of Restricted-Access Data Sharing

By its nature, a data commons promotes open, public access to data. However, prioritizing public access may sometimes limit or even preclude collaborations that would have broader social value. Several models exist for restricted-access data sharing—providing select researchers with full data access (anonymized as necessary to comply with privacy regulations), under the provision that researchers may document insights in the public sphere but must keep the underlying data confidential (Verhulst et al. 2019).

One common approach involves creating a "data brokerage" with a trusted intermediary, in which a third party facilitates connections between data suppliers who could individually and jointly benefit from analysis of a combined dataset. The Duke EAP has used this model in the past, combining sales data from several off-grid providers in East Africa to analyze the effects of solar panel import tariffs on electricity access (see Fetter and Phillips (2018)). Another model for restricted-access data sharing involves bilateral partnerships between owners of proprietary data and external researchers. For instance, Fobi et al. (2018) use proprietary electricity consumption data from Kenya Power & Light (KPLC) to generate and publish key insights on consumption patterns among newly connected households over time, while maintaining confidentiality of the underlying data.

Despite such solutions to data sharing, a "commons" idea remains attractive, because others who are not privy to such partnerships may produce wholly different and useful insights that are not readily apparent to the limited group with exclusive access or may be able to provide them faster. A restricted-access data competition provides a different model for broadening access. In this setup, data suppliers provide a codebook or a small snippet of data and then solicit research proposals in an open competition.[6] The data suppliers review the proposals received and select a few winners, perhaps prioritizing proposals based on the quality or value of the idea, or their ability to link related datasets in a way that generates novel insights. Winning proposal teams are then granted access to the full dataset.

For instance, in a "Data for Development Challenge" organized by Orange Telecom, one team of researchers was granted access to anonymized mobile phone user data in Senegal, which they used to estimate demand for electricity and design a least-cost electrification plan (Martinez-Cesena et al. 2015). This anecdote provides a good example of how data providers can benefit from sharing data and crowdsourcing solutions; the original call for proposals had solicited ideas on a broad set of topics related to infrastructure, environment, and health. Orange Telecom benefited by reviewing proposals from expert teams, since even proposals that are not ultimately selected can generate useful ideas for the data provider. Orange also likely enjoyed some reputational benefits. Community planners benefited from receiving advice on infrastructure planning. And researchers benefited by obtaining access to unique proprietary data, contributing their expertise to real-world problems, and for successful teams, a gain in prestige.

The Orange Telecom story is only one recent example; the New York University Government Lab (GovLab) has catalogued a more complete set of data collaborations, documenting data challenges, and providing guidance on the formation of collaborative projects that create public value by exchanging data (Verhulst et al. 2019). One concern about the competition model is that it likely favors research teams with structural advantages, potentially exacerbating the structural barriers faced by developing-country researchers and practitioners. This could be addressed by reserving one or more awards for developing-country teams, thus enhancing the ability of these teams to contribute to improved practice, building stronger global networks, and ensuring opportunities for these researchers to build their reputations and expertise.

[6] In large organizations such as utility companies, individual departments often lack knowledge of the data held by other departments within the same organization; these organizations can often benefit from having defined, cross-cutting roles for storing and managing data. To the extent that competitions help to formally create channels for data sharing, they could facilitate data sharing within organizations, providing additional benefits.

RECOMMENDATIONS AND NEXT STEPS

Big data approaches offer considerable promise for improving energy access policy, planning, and investment. We have discussed here a number of exciting projects focused on energy emerging from partnerships of academic researchers, government policymakers, civil society, and companies. Some such applications, such as using satellite-detected nighttime lights to monitor community-level access, are relatively well-established and mature, but many others are an early stage of development. We believe there are four high-priority focus areas that would benefit from additional resources to support collaboration, research, and integration of insights into policy and practice.

(1) **Investments in new data sources and algorithms to facilitate interpretation.** For instance, smart meters enable lower utility management costs (e.g., through automated billing and reduced commercial losses), high-frequency data analytics to optimize system planning and operations, and more efficient and timely data collection for donors or decision-makers seeking to implement RBF. As another example, automated object identification applied to street-level imagery could help to identify energy infrastructure (e.g., diesel generators) that is important but not always visible from above or is difficult to distinguish in satellite images.

(2) **Integrating traditional surveys, such as the World Bank's MTF survey, with remote sensing and big-data approaches**. As Fraym's work demonstrates, using satellite images to infer traffic flows, building footprints, or other data at fine spatial resolution could help in downscaling (estimated) Census data that enables measurement of social and economic conditions in communities over time. Integration with recent, high-quality survey data can also help to advance big-data analytical methods. For instance, the MTF survey provides granular information on household access to relatively low- or high-tier energy in select communities; this data could be combined with data from satellite images to predict low- or high-tier access in communities not sampled in the MTF. Integrative approaches can also help both public and private entities better target and evaluate investments, by first using satellite images to identify clusters of infrastructure or land uses that are indicative of significant latent demand for energy—e.g., unirrigated agricultural fields near surface water sources, or health clinics or schools—and then deploying targeted surveys alongside secondary data sources to quantify demand and build appropriate energy infrastructure, and finally assessing the extent to which these predictions and the investments prove successful.

(3) **A managed data commons to advance practice on emerging and heretofore unidentified topics**. Enhancing the accessibility and interoperability of energy access data, including through curated databases organized around specific priority use cases, would benefit current and prospective users all around the world—especially those based in developing countries, who face additional structural barriers. Providing online tools for data visualization that respond to identified needs of practitioners and researchers would help to promote the generation of new ideas, and to enhance collaborations on new or existing projects. Although a data commons would require startup investment, proven models exist for maintaining operations once the building blocks are in place. This is more likely to the extent that a commons can take advantage of network effects (e.g., if it becomes recognized as the most promising or prominent solution), which suggests that the best path forward for building one would center on the needs of prominent or influential stakeholders.

(4) **Restricted-access data sharing that matches data generators with users, protects confidentiality and privacy, and creates a living record of a wide range of energy access data projects**. Restricted-access data sharing through bilateral partnerships, data brokerage arrangements, or data competitions that offer data access as part of a "prize" hold considerable potential for producing game-changing insights. Data providers in these arrangements can gain new insights from the research conducted while also contributing public goods and positively engaging their broader stakeholder community (as in the case of the Data for Development Challenge). Researchers also benefit from being able to engage with unique and nonstandard data sources. Donor and philanthropic resources can help to enable high-quality research and the creation of public knowledge goods from such privately held data.

REFERENCES

Bustos-Turu, G., K.H. van Dam, S. Acha, C.N. Markides, and N. Shah. 2016. “Simulating Residential Electricity and Heat Demand in Urban Areas Using an Agent-Based Modelling Approach.” In 2016 IEEE International Energy Conference (ENERGYCON), pp. 1–6.

Diawuo, F.A., M. Sakah, A. Pina, P.C. Baptista, and C.A. Silva. 2019. “Disaggregation and Characterization of Residential Electricity Use: Analysis for Ghana.” *Sustainable Cities and Society* 48: 101586.

Fawcett, A.A., J.R. McFarland, A.C. Morris, and J.P. Weyant. 2018. “Introduction to the EMF 32 Study on US Carbon Tax Scenarios.” *Climate Change Economics* 9(01): 1840001.

Fetter, R., and J. Phillips. 2019. “The True Cost of Solar Tariffs in East Africa.” NI PB 19-01. Durham, NC: Duke University. http://nicholasinstitute.duke.edu/publications.

Flostrand, A., T. Eriksson, and T.E. Brown. 2019. “Better Together—Harnessing Motivations for Energy Utility Crowdsourcing Activities.” *Energy Research and Social Science* 48: 57–65.

Fritz, S., L. See, T. Carlson, M.M. Haklay, J.L. Oliver, D. Fraisl, et al. 2019. “Citizen Science and the United Nations Sustainable Development Goals.” *Nature Sustainability* 2(10): 922–930.

Malof, J.M., K. Bradbury, L.M. Collins, and R.G. Newell. 2016. “Automatic Detection of Solar Photovoltaic Arrays in High Resolution Aerial Imagery.” *Applied Energy* 183: 229–240.

Malof, J.M., B. Li, B. Huang, K. Bradbury, and A. Stretslov. 2019. “Mapping Solar Array Location, Size, and Capacity Using Deep Learning and Overhead Imagery.” https://arxiv.org/abs/1902.10895.

Martinez-Cesena, E.A., P. Mancarella, M. Ndiaye, and M. Schläpfer. 2015. “Using Mobile Phone Data for Electricity Infrastructure Planning.” https://arxiv.org/abs/1504.03899.

Shafiei, E., H. Thorkelsson, E.I. Ásgeirsson, B. Davidsdottir, M. Raberto, and H. Stefansson. 2012. “An Agent-Based Modeling Approach to Predict the Evolution of Market Share of Electric Vehicles: A Case Study from Iceland.” *Technological Forecasting and Social Change* 79(9): 1638–1653.

Tatem, A.J., S.J. Goetz, and S.I. Hay. 2008. “Fifty Years of Earth-Observation Satellites.” *American Scientist* 96(5): 390.

Verhulst, S.G., A. Young, M. Winowatan, and A.J. Zahuranec. 2019. “Leveraging Private Data for Public Good: A Descriptive Analysis and Typology of Existing Practices.” https://datacollaboratives.org/static/files/existing-practices-report.pdf.

Wu, G. C., R. Deshmukh, K. Ndhlukula, T. Radojicic, J. Reilly-Moman, A. Phadke, et al. 2017. “Strategic Siting and Regional Grid Interconnections Key to Low-Carbon Futures in African Countries.” *Proceedings of the National Academy of Sciences* 114(15): E3004–E3012.

Rolnick, D., P.L. Donti, L.H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, et al. 2019. “Tackling Climate Change with Machine Learning.” https://arxiv.org/abs/1906.05433.

APPENDIX: ATTENDEES AT THE DUKE-RTI CONVENING (* = panelist or moderator)

Vivian Agbegha, Millennium Challenge Corporation

Andrew Allee, Rocky Mountain Institute*

Ruba Amarin, RTI International

Allison Archambault, EarthSpark International

Katie Auth, USAID / Power Africa

Justin Baker, RTI International*

Fabiola Baltodano, Interamerican Development Bank

Amer Bargouth, RTI International*

Duncan Chaplin, Mathematica

Terence M Conlon, Columbia University

Molly Dean, USAID

Rob Fetter, Duke University*

Michael Gallaher, RTI International

Piyush Gambhir, Duke University

Hannah Girardeau, SEforALL

Madeleine Gleave, Nithio*

Julian Glucroft, Millennium Challenge Corporation*

Ziting Huang, World Bank

Prathibha Juturu, Johns Hopkins University

Njeri Kara, Duke University

Kadeem Khan, Facebook

Bryan Koo, World Bank*

Rachel Kriegsman, Clean Energy Leadership Institute

Elisa Lai, CLASP

Santiago Sinclair Lecaros, World Resources Institute

Ruoshui Li, Duke University

Jordan Malof, Duke University*

Arif Mamun, Mathematica

Subodh Mathur, Johns Hopkins SAIS

Emily McAteer, Odyssey*

Robyn Meeks, Duke University*

Bethyn Merrick-Nguyen, Duke University

Muchiri Nyaggah, Africa Open Data Network*

Alicia Oberholzer, Duke University

Shreena Patel, Millennium Challenge Corporation

Catherine Pham, USAID

Jonathan Phillips, Duke University*

Victoria Plutshack, Duke University

Jem Porcaro, SEforALL

Anthony Randazzo, OPIC

Pauline Ravillard, IDB

James (Jay) Rineer, RTI International*

Maria Hilda Rivera, Power Africa

Jon Saiger, Millennium Challenge Corporation*

Rajah Saparapa, Duke University

Dan Schnitzer, SparkMeter

Laura Seidman, American Forest & Paper Association

Myriam Sekkat, Duke University

Nicole Silvya Bouris, IFC

Julius Svoboda, USAID

Dan Sweeney, REEEP*

Melia Ungson, FRAYM*

Nathan Williams, Rochester Institute of Technology*