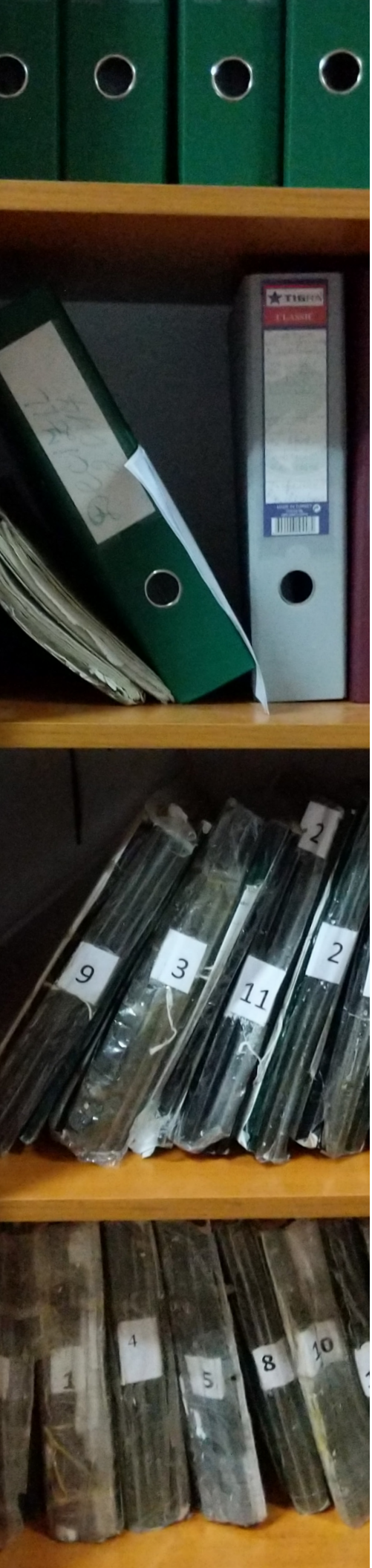

HOW TO ENSURE THAT YOUR DATA SCIENCE IS INCLUSIVE

As advocates of data science for social good, we have an opportunity—and an obligation—to ensure that our efforts lead to more sustainable, equitable, inclusive, and indigenous research and action.

As a new generation of data scientists emerges in Africa, they will encounter relatively little trusted, accurate, and accessible data upon which to apply their skills. It's time to acknowledge the limitations of the data sources upon which data science relies, particularly in lower-income countries.

The potential of data science to support, measure, and amplify sustainable development is undeniable. As public, private, and civic institutions around the world recognize the role that data science can play in advancing their growth, an increasingly robust array of efforts has emerged to foster data science in lower-income countries.

This phenomenon is particularly salient in Sub-Saharan Africa. There, foundations are investing millions into building data literacy and data science skills across the continent. Multilaterals and national governments are pioneering new investments into data science, artificial intelligence, and smart cities. Private and public donors are building data science centers to build cohorts of local, indigenous data science talent. Local universities are launching graduate-level data science courses.

Despite this progress, among the hype surrounding data science rests an unpopular and inconvenient truth: **As a new generation of data scientists emerges in Africa, they will encounter relatively little trusted, accurate, and accessible data that they can use for data science.**

We hear promises of how data science can help teachers tailor curricula according to students' performances, but many school systems don't collect or

How useful can data science be if so much important information is analog, as it is in this mayor's office in Moldova?

track that performance data with enough accuracy and timeliness to perform those data science–enabled tweaks. We believe that data science can help us catch disease outbreaks early, but health care facilities often lack the specific data, like patient origin or digitized information, that is needed to discern those insights.

These fundamental data gaps invite the question: **Precisely what data would we perform data science on to achieve sustainable development?**

Data possessed by mobile operators, produced by expensive technologies, or generated only by those fortunate enough to be connected online will never be as diverse, expansive, accessible, or representative as it should be.

There are, of course, compelling examples of data science being put to use for the public good. Emerging use cases include exploring **call detail records** to improve mobility and urban planning, using **remote sensors** to measure agricultural or economic growth, and mining **online content** to monitor election violence. These and other examples prove beyond any doubt that data science has a role to play in advancing sustainable development.

But call detail records take time, money, and (often) political connections to obtain. Online content, like tweets, only reflects the small number of people in lower-income countries who have internet access and use those platforms. **We're working hard to make data science accessible to everyone, but those data scientists are left to work with information that remains either inaccessible to everyday technologists, or unrepresentative of the marginalized.**

Often, data is not properly structured for data science, despite containing invaluable information, like these notes about citizens' priorities in Tanzania.

There are numerous consequences of these approaches. As leaders and influencers increasingly rely on data science to guide their decision-making, they risk making flawed decisions that ignore the needs, perspectives,

or values of the people they serve who aren't online (over half the world's population), or who aren't using mobile devices (which are used more by men than by women).

These leaders also **risk disenfranchising a new generation of African data scientists** who lack the financial resources to access large and reliable datasets to put their skills to use, or who stand by to watch as better funded organizations—such as universities in the "Global North"—conduct data science and analytics about their communities from oceans and continents away.

As advocates of data science for social good, we have an opportunity—and an obligation—to ensure that our efforts lead to more sustainable, equitable, inclusive, and indigenous research and action.

There are specific steps we can take for data science to achieve its full potential in the realm of sustainable development:

1. **Be wary of producing a generation of data scientists who must rely on expensive, hard-to-access data to meaningfully apply their skills.** We should couple our data science trainings with efforts to build skills on collecting or producing data through methods like community mapping, or through data-sharing initiatives like data collaboratives.

2. **Be conscious of reinforcing dependencies on foreign companies whose technologies and platforms compose the bulk of today's data science case studies.** We should intentionally pair our investments into data science with investments into indigenous innovations that *produce* statistics for data science. Low-cost, locally-built technologies like unmanned aerial vehicles (UAVs), and initiatives that produce locally relevant training data sets, can help avoid these dependencies.

3. **Be mindful of focusing too much on data science and not enough on fundamental data literacy.** We should double down on building fundamental data skills—collecting, cleaning, analyzing, and sharing data—among staff of health clinics, schools, local governments, and elsewhere where so much valuable information is actually produced. This will improve the availability and reliability of large datasets for data

scientists to use.

Fortunately, momentum is beginning to shift in favor of indigenous data science. Independent innovators are addressing gaps in African languages that can be used for natural language processing. Initiatives like Data Science Africa and the Deep Learning Indaba are nurturing communities of machine learning experts. These are steps in the right direction.

Five years from now, a new generation of socially conscious, impact-driven data scientists in Africa will emerge, and many of them will be driven to use their skills to address sustainable development challenges. **We must ensure that the information that powers their efforts isn't limited to expensive, inaccessible, or unrepresentative data that sits primarily in the hands of a few mobile operators, banks, or tech companies.**

Getting to this level means complementing the hype of data science for global good with the long, hard hauls of improving data quality at local levels, investing in indigenous technology and content, and building fundamental data skills. Only then will the data science revolution reach its full potential.

For more information, contact Samhir Vasdev at samhir@irex.org.

Staff at Sikika, one of Tanzania's largest nongovernmental health organizations, practice analyzing and managing datasets in order to improve data quality.

IREX
1275 K Street, NW, Suite 600
Washington, DC 20005
+1 (202) 628-8188
irex.org | communications@irex.org