Reflecting the Past, Shaping the Future:

Making AI Work for International Development

USAID
FROM THE AMERICAN PEOPLE

This report was written by Amy Paul, Craig Jolley, and Aubra Anthony of the Strategy & Research team within the Center for Digital Development at USAID. The team would like to thank every organization and individual interviewed during the course of the research, and those who offered feedback and counsel along the way. Without your expertise, insight, and support, this report would not have been possible.

Cover photo: USAID's Responsible Engaged and Loving (REAL) Fathers Initiative aims to build positive partnerships and parenting practices among young fathers. ***Credit:*** Save the Children

For details on the cover and other artwork in this report, *see the "About the artwork" section on pg. 90–91.*

Contents

Introduction

We are in the midst of an unprecedented surge of interest in machine learning (ML) and artificial intelligence (AI) technologies. These tools, which allow computers to make data-derived predictions and automate decisions, have become part of daily life for billions of people. Ubiquitous digital services such as interactive maps, tailored advertisements, and voice-activated personal assistants are likely only the beginning. Some AI advocates even claim that AI's impact will be as profound as "electricity or fire[1]" that it will revolutionize nearly every field of human activity. This enthusiasm has reached international development as well[2,3]. **Emerging ML/AI applications promise to reshape healthcare, agriculture, and democracy in the developing world.** ML and AI show tremendous potential for helping to achieve sustainable development objectives globally. They can improve efficiency by automating labor-intensive tasks, or offer new insights by finding patterns in large, complex datasets. A recent report suggests that AI advances could double economic growth rates and increase labor productivity 40% by 2035[4]. **At the same time, the very nature of these tools — their ability to codify and reproduce patterns they detect — introduces significant concerns alongside promise.**

In developed countries, ML tools have sometimes been found to automate racial profiling,[5] to foster surveillance,[6] and to perpetuate racial stereotypes[7]. Algorithms may be used, either intentionally or unintentionally, in ways that result in disparate or unfair outcomes between minority and majority populations[8]. Complex models can make it difficult to establish accountability or seek redress when models make mistakes[9]. These shortcomings are not restricted to developed countries. They can manifest in any setting, especially in places with histories of ethnic conflict or inequality. **As the development community adopts tools enabled by ML and AI, we need a clear-eyed understanding of how to ensure their application is effective, inclusive, and fair.** This requires knowing when ML and AI offer a suitable solution to the challenge at hand. It also requires appreciating that these technologies can do harm — and committing to addressing and mitigating these harms.

ML and AI applications may sometimes seem like science fiction, and the technical intricacies of ML and AI can be off-putting for those who haven't been formally trained in the field. However, there is a critical role for development actors to play as we begin to lean on these tools more and more in our work. **Even without technical training in ML, development professionals have the ability — and the responsibility — to meaningfully influence how these technologies impact people.**

You don't need to be an ML or AI expert to shape the development and use of these tools. All of us can learn to ask the hard questions that will keep solutions working for, and not against, the development challenges we care about. Development practitioners already have deep expertise in their respective sectors or regions. They bring necessary experience in engaging local stakeholders, working with complex social systems, and identifying structural inequities that undermine inclusive progress. Unless this expert perspective informs the construction and adoption of ML/AI technologies, ML and AI will fail to reach their transformative potential in development.

This document aims to inform and empower those who may have limited technical experience as they navigate an emerging ML/AI landscape in developing countries. Donors, implementers, and other development partners should expect to come away with a basic grasp of common ML techniques and the problems ML is uniquely well-suited to solve. We will also explore some of the ways in which ML/AI may fail or be ill-suited for deployment in developing-country contexts. Awareness of these risks, and acknowledgement of our role in perpetuating or minimizing them, will help us work together to protect against harmful outcomes and ensure that AI and ML are contributing to a fair, equitable, and empowering future.

Roadmap: How to use this document

TERMINOLOGY

Throughout this document, you'll see "definition boxes" in the page margins that explain key technical terms. Some of these terms will recur throughout this document, while others provide background information that may be useful for future discussions or reading. There are two pieces of jargon in particular that you should start with an understanding of:

- **Machine learning (ML)** is a set of methods for getting computers to recognize patterns in data and use these patterns to make future predictions. For shorthand, you could think of ML as "data-driven predictions."
- **Artificial intelligence (AI)** uses computers for automated decision-making that is meant to mimic human-like intelligence. Automated decisions might be directly implemented (e.g., in robotics) or suggested to a human decision-maker (e.g., product recommendations in online shopping); the most important thing for our purpose is that some decision process is being automated. AI often incorporates ML (when using data-driven predictions to make better decisions) but doesn't have to. For shorthand, you can think of AI as "smart automation."

These definitions are rough and informal, and will probably not be very satisfying to some experts in ML and AI. Our goal here is to provide non-experts with enough context to understand what's going on, without getting bogged down in nuance. Although we'll use words like "learning" and "intelligence," keep in mind that we're not ascribing consciousness to computers. They're just machines.

! CAUTION

The line between ML and AI, especially in some of the examples we cite, may be blurry. We will limit ourselves in this report to only those AI systems that incorporate a ML component, rather than the broader field of Artificial Intelligence[10]. Because of this, we'll often default to using the term "machine learning" to describe applications that are both purely ML as well as those that may justifiably be called AI, but that are built on ML or have an ML component.

WHO SHOULD READ THIS REPORT?

This document is aimed at development practitioners who may find themselves funding, managing, or advising on projects that involve ML or AI. Our goal is to provide enough technical background to help "non-technical people" to ask hard questions and insist on answers they can understand. On the other hand, if you're already an expert in ML, this report can help you see how your development colleagues can contribute to your work.

WHICH PARTS ARE IMPORTANT FOR ME?

If you're pressed for time and want to prioritize your reading efforts, the following table can help guide your attention. Our hope is that readers can focus on the sections that are most important for their needs or interest, with the understanding that the implications of the field and the final recommendations will make more sense if you've read the full document.

Report navigation guide

ML and AI: What are they?

Purpose: Give an introduction to ML and AI.

Intended Audience: People interested in a slightly deeper understanding of how these fields relate to each other.

How ML works: The basics

Purpose: Explain what ML does, using a simple example of a line fitted to points.

Intended Audience: People looking for an explanation of ML at a level similar to Microsoft Excel or other spreadsheet software.

How people influence the design and use of ML tools

Purpose: Explain how ML tools are shaped by the decisions of people who design, build, and test them.

Intended Audience: Anyone who wants to know how all technical design choices matter in the real world.

Action suggestions

Purpose: Share concrete recommendations for how development experts can help keep ML/AI applications on the right track.

Intended Audience: Development practitioners who want to help make sure that their technology projects are effective, responsible, and safe.

Looking forward

Purpose: Offer some concluding thoughts on how to prepare for better ML/AI applications in the near future.

Intended Audience: Development practitioners who may not have a current ML/AI project, but want to take steps now to support future success.

PHOTO: ROBERT SAUERS FOR USAID

Example applications in development

Purpose: Provide a sampling of ML and AI applications in development sectors including humanitarian relief, health, and agriculture.

Intended Audience: Anyone looking for examples of what ML and AI can do, or ideas about how to apply them in their own work.

Case studies

Purpose: Share two in-depth case studies of ML and AI in development.

Intended Audience: Readers who want examples of how ML and AI can be used to improve development outcomes.

What can go wrong?

Purpose: Illustrate some of the ways that ML can imperil development outcomes.

Intended Audience: People concerned about the risks of using ML in development programs.

PHOTO: LESLIE DETWILER FOR USAID

Quick reference: Guiding questions

Purpose: Start a conversation with a list of questions meant to help target issues of fairness in projects incorporating ML or AI.

Intended Audience: Development actors with technical and non-technical expertise who want to explore the fairness of the ML/AI tools that are or may be incorporated in their projects.

Appendix: Peering under the hood

Purpose: Give a more detailed description of how ML models are built and integrated into decision processes. Provides technical background for understanding "Social impacts" section.

Intended Audience: Development practitioners who want to know more about how ML tools are built and used, and to understand the design choices that impact a model's behavior.

Machine Learning: Where we are and where we might be going

ML AND AI: WHAT ARE THEY?

Put simply, machine learning (ML) is a set of methods for training computers to learn from data, where "learning" generally amounts to the detection of patterns or structures in data. This differs from how statistical analysis has traditionally been done. The usual method is to first develop a model based on mathematical rules and then apply this model to data. ML approaches flip this process (*See FIGURE 1*). They begin by finding patterns in ***training data*** and return a ***model*** that can make ***predictions*** for new, unseen data. ML techniques can be especially effective at finding complex, nonlinear relationships, and for making sense of large amounts of unstructured image, audio, and text data. *A more detailed overview of how ML works is provided in an appendix of this document. See "Appendix: Peering under the hood."*

Artificial Intelligence (AI) is the science and technology of creating intelligent systems. AI systems are often enabled by ML, and apply data-derived predictions to automate decisions. While ML focuses on learning and prediction, AI applications often create, plan, or do something in the real world. For example, an ML model might be used to predict driving time between two places, while an AI application would plan routes (or even drive the car).

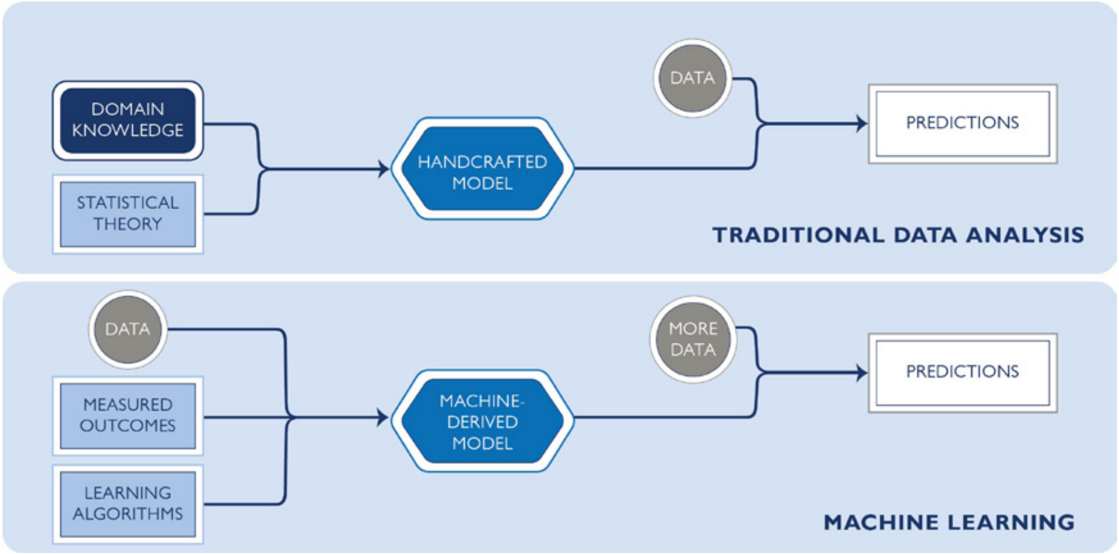


FIGURE 1: In traditional data analysis, one begins with data and a predefined model, and uses these to calculate an output. Machine learning requires some outputs to be specified in advance, but can use these to build a predictive model from the data.

Many sources use the terms "Machine Learning" and "Artificial Intelligence" interchangeably, and it can be helpful to put them in context. In general, AI is the field of science and technology concerned with building machines that act "intelligently*." For the purposes of this report, ML can be considered a sub-field of AI that is concerned with learning — building computer systems that can generalize from past experience to form expectations about new experiences.

A–Z

Training Data:
Data used to develop a ML model. A learning algorithm will find patterns and relationships in training data and use them to define rules for new predictions.

Model:
A simplified depiction of reality. ML models consist of an algorithm and parameters that were learned from training data. When an algorithm is combined with training data, we get a predictive model.

Prediction:
Guessing an unknown attribute or quality based on known information. ML predictions are not always about the future; they are estimates based on measurable features. Features are often predicted because direct measurement is difficult, dangerous, or expensive.

* Defining intelligence is tricky, but there is general consensus that it requires the ability to learn. Natural intelligence also involves other things, like attention, memory, and creativity.

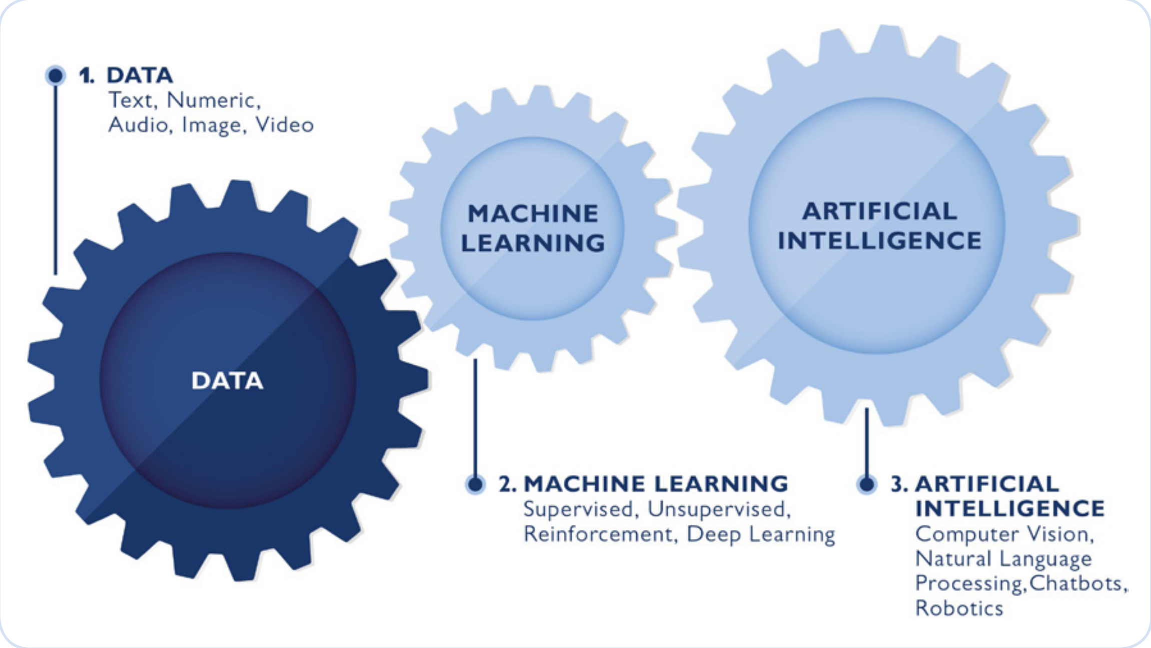

1. DATA
Text, Numeric, Audio, Image, Video
DATA
MACHINE LEARNING
ARTIFICIAL INTELLIGENCE
2. MACHINE LEARNING
Supervised, Unsupervised, Reinforcement, Deep Learning
3. ARTIFICIAL INTELLIGENCE
Computer Vision, Natural Language Processing, Chatbots, Robotics

FIGURE 2: The relationship between data, machine learning, and AI applications is shown as a set of three interlocking gears. Data serve as the foundation of ML/AI systems, and decisions about data affect the function of higher-level systems. Rather than working directly with data, AI applications typically rely on a machine learning algorithm to translate data into usable predictions. Finally, AI applications use those predictions to make, plan, or do something in the real world.

One difference between natural, or human, intelligence and artificial intelligence is that humans absorb and process data (especially visual data) in the context of the surrounding environment. If the interpretation of the data (for example identifying an image) doesn't fit the context of the situation, the human can recognize that something is not quite right. Presently ML and AI systems lack the ability to recognize if the "answer" the machine arrives at agrees with the context.

Another difference is that humans have the ability to learn from very small data sets: a child, for example, can be told something only a few times and learn a new word or behavior. Typically ML and AI systems require very large data sets; sometimes thousands of images or words and examples are required to "learn" how to provide an answer. The result can be systems that are statistically very good, but that for individual cases fail spectacularly. Many ML and AI systems may provide wrong or inappropriate answers if used in a context different from their training environment.

This section emphasizes taxonomy to provide some basic orientation for the reader. *See the "Roadmap" above for how to use this document.* Later in this report we will have more to say about exactly how machine learning actually gets done. The figure shown here FIGURE 2 depicts how common AI and ML tools and terms fit together.

BOX 1: What are machine learning and artificial intelligence?

Machine learning (ML) allows computers to generalize from existing data and make predictions for new data. This differs from traditional statistics, which specifies a theoretical model and assesses the model by fitting it to data. ML approaches flip this process: they find patterns in "training" data and return an empirical model that can make predictions for new, unseen data.

ML models can be especially effective at finding complex, nonlinear relationships, and for making sense of unstructured image, audio, and text data.

Artificial Intelligence (AI) is the science and technology of creating intelligent systems. AI systems are often enabled by ML, but go beyond learning and prediction to create, plan, or do something in the real world. For example, an ML model might predict driving time between two places, while an AI application would plan routes (or even drive the car).

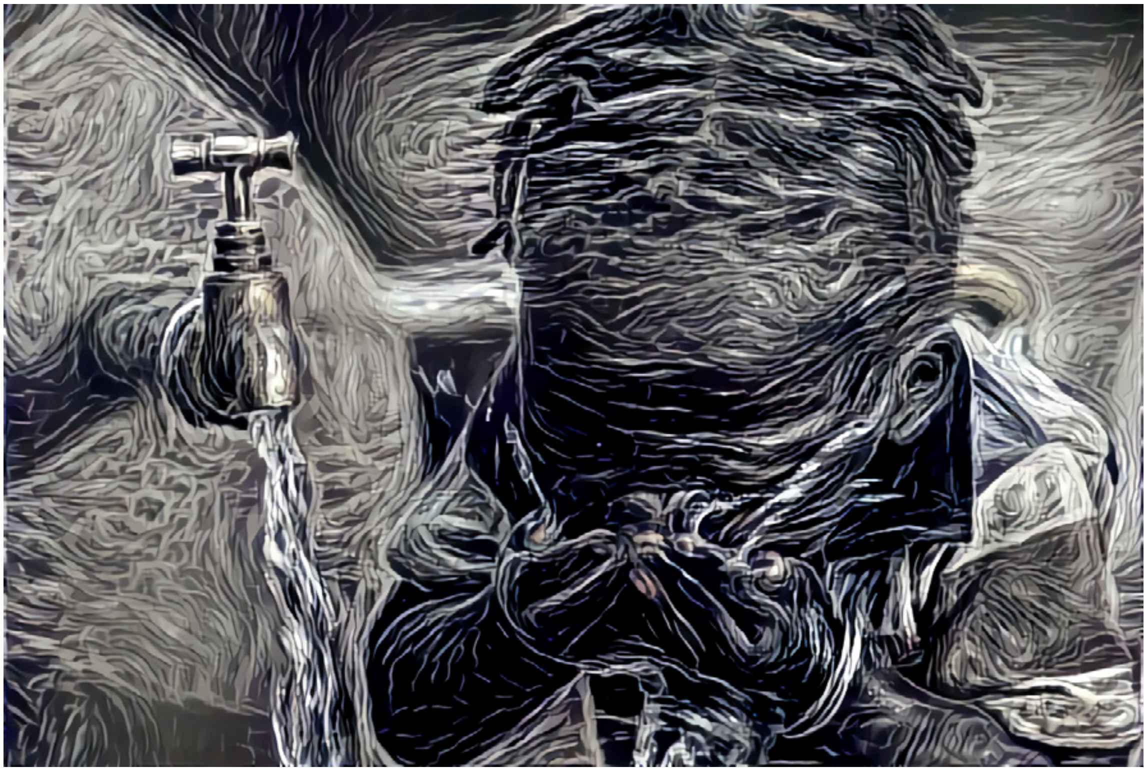


PHOTO: AECOM

All ML/AI systems are built on data. This can refer to numeric data (e.g. tables with rows and columns of numbers), but can also include other types of data: images, audio, text, etc. Non-numeric formats usually require additional pre-processing to be converted into a format that can work with ML *algorithms.* In some cases, such as computer vision for image data or *natural language pxrocessing* for text data, these pre-processing steps can be complex and sophisticated — and can even themselves be augmented by ML.

Once data have been adequately prepared, they are input into a machine learning model. ML models typically involve either *classification or regression. Classification* aims to assign an instance to one of several categories, based on learning from past observations. For example, given a series of aerial images, which contain huts? *Regression* uses patterns in the data to predict a quantity. For example, given the same aerial images, what is the likely population density of the area?

ML algorithms can also be broadly categorized as *supervised or unsupervised**. Supervised ML requires model-builders to specify the "right" answers (referred to as training data), which the algorithm will then learn to imitate. For example, a credit-scoring algorithm might analyze the repayment history of past borrowers to determine which future borrowers are likely to default. In unsupervised ML, the algorithm detects patterns or clusters in the data without being told what to look for. An unsupervised credit algorithm might identify clusters of similar borrowers, but would not make individual predictions about repayment.

Sometimes, a system will automatically use the predictions from an ML model to plan, create, or do something in the real world. These are the AI applications at the top of FIGURE 2 — tools that apply the data-derived learning of ML algorithms. AI applications can incorporate other aspects of intelligence, like creativity (image/text generation) or autonomy (robotics, control systems, etc.). As mentioned in the *"Roadmap"* section: for simplicity, we will default to refer to ML-backed systems as "ML" even when that ML model is part of a broader AI system.

A–Z

Algorithm:
A systematic procedure for performing a task or solving a problem, often implemented by a computer.

Natural Language Processing:
Using computers to process a "natural" language spoken and written by humans (e.g., English, French, Arabic).

Classification:
Assignment of data points to one of two or more qualitatively-different categories.

Regression:
Predicting a numeric value or score for each data point.

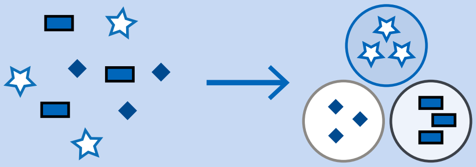
Supervised ML:
Algorithms that require training data to be labeled with values of the outcome variable. Supervised algorithms need to know the "right" answer to develop prediction rules.

Unsupervised ML:
Algorithms that do not require pre-labeling of the outcome variable. Rather than predicting the "right" answer, unsupervised ML finds latent patterns in data.

* Other ML approaches exist outside these two major categories. One important example is reinforcement learning, in which a computer learns to achieve an objective through trial and error[11]

BOX 2: Machine Learning Tasks

While specific applications of ML in development vary considerably, they can be roughly organized into three "tasks."

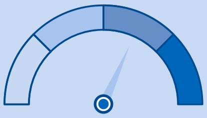


SORT

Machine learning algorithms often must sift through large volumes of data to find specific instances of interest. In humanitarian response scenarios, this might involve looking through social media data to provide real-time situational awareness in a disaster[12]. In global health, ML tools provide tailored surveillance data to improve risk assessment for Zika[13]. Applications like these often involve a function of "sorting" — separating instances into one of several qualitatively distinct categories. Machine learning experts also refer to this process as classification.

Sorting applications often separate things into two categories: typical and atypical. Atypical instances are often anomalies, such as malaria-infected cells in a rapid diagnostic test,[14] or sudden departures from normal-looking forest cover in deforestation monitoring[15]. Anomaly detection can use either supervised or unsupervised ML. Unsupervised approaches aim to tell when something is different from anything seen before. Supervised anomaly detection requires a curated set of "anomalies" and finds things that match them.

In other cases, ML tools are used to sort things into a larger number of categories. For example, many image-recognition algorithms are built on the ImageNet database,[16] which associates thousands of digital photographs with text descriptions. Google has released open-source code for language detection[17] that can sort text samples into one of more than 200 languages.



SCORE

ML models can also be used to give specific instances a numerical score. Scores are often probabilities, such as whether a loan will be repaid or a new hire will succeed in a job. They can also be quantitative estimates such as a person's age or a household's annual income. Machine learning experts typically refer to scoring applications as *regressions,* similar to the way the term is used in statistics.

Scoring applications are most useful when an individual decision must be made about each item in the dataset. They rely on large volumes of data to learn nuanced prediction rules. Many two-category "sort" applications actually use a "score" method under the surface. The model generates a probability of belonging to one category and then makes binary decisions based on a probability cutoff (e.g., one category for scores less than 50% and the other category for scores greater than 50%). The important difference is in the type of decision being made — if the choice is all-or-nothing, a "sort" application is often used.

DISCOVER

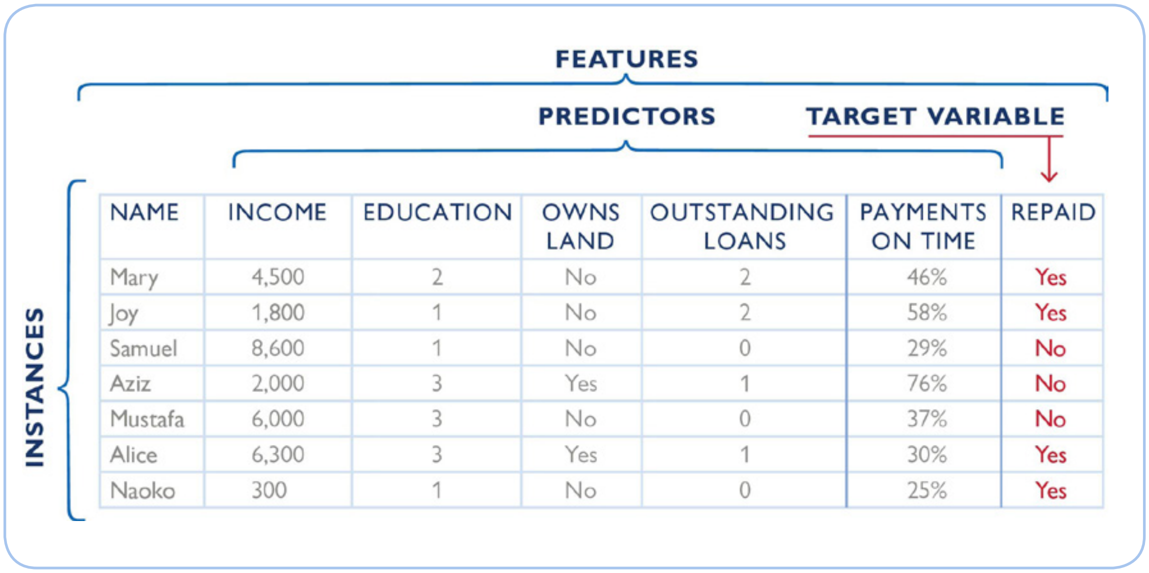
Machine learning can also be used to understand trends and identify patterns in data. Instead of returning predicted scores or classifications, discovery applications are pursued to uncover correlations that offer testable hypotheses about the causal relationship between input and output variables. This requires algorithms to be at least somewhat interpretable to the people who use them. Some algorithms, such as linear regression or simple partition trees, are designed for easy interpretation. For more complex algorithms, other techniques can aid in interpretation. *See BOX 5: Opacity and explainability.* Knowing *how* an algorithm sorts or scores can help us generate new hypotheses or discern which of several factors most strongly influences an outcome. In some cases, we care less about the predictions made by a model than about the variables that are most important in making those predictions.

Algorithmic discovery can also inform "offline" decision rules. In one example, researchers developing a test for Zika infection detected a variety of viral fragments in blood from infected patients. An ML algorithm was used to construct a profile of viral fragments that could distinguish Zika infection from Dengue or other viruses. This algorithm won't be used directly for Zika diagnosis; instead the fragments with the most predictive power will be incorporated into a cheaper, simpler Zika test. Other discovery-oriented uses of ML have helped to make crop-management recommendations for smallholder farmers[18].

HOW ML WORKS: THE BASICS

For readers who are interested in the basics of machine learning, this section introduces some terminology and walks through a simplistic credit-scoring example. *For more detail, see "Appendix: Peering under the hood".*

ML models aim to estimate values of a target variable based on a set of predictors. For example, FIGURE 3 shows a sample dataset in which the ***target variable*** is loan repayment, while the ***predictors*** describe a borrower's financial situation. In terms of the taxonomy introduced in BOX 1, this is a "sort" application — borrowers are being separated into two discrete categories, depending on whether they repaid their loans. If you've ever worked with a spreadsheet, you can imagine the format that most ML algorithms are designed for — a set of rows (often called ***instances***) and columns (often called ***features***). One particular feature will be our target variable, while others may be used as predictors.


FEATURES
PREDICTORS
TARGET VARIABLE
INSTANCES

NAME	INCOME	EDUCATION	OWNS LAND	OUTSTANDING LOANS	PAYMENTS ON TIME	REPAID
Mary	4,500	2	No	2	46%	Yes
Joy	1,800	1	No	2	58%	Yes
Samuel	8,600	1	No	0	29%	No
Aziz	2,000	3	Yes	1	76%	No
Mustafa	6,000	3	No	0	37%	No
Alice	6,300	3	Yes	1	30%	Yes
Naoko	300	1	No	0	25%	Yes

FIGURE 3: Illustration of data terminology for a sample dataset in which the target variable is loan repayment, while the predictors describe a borrower's financial situation.

Ideally, the set of predictors should be diverse enough to capture different aspects of the things they describe*. For example, a dataset that contains a person's repayment history on several loans is much less diverse than one that also describes her employment history, social contacts, and education. Broadly speaking, ML systems seek to find appropriate ***proxies*** to estimate target variables that can be difficult, expensive, or even dangerous to measure directly.

Target Variable: The value being predicted by an ML model. This can be either a number (for regression) or a category label (for classification). Also referred to as an outcome variable or dependent variable.

Predictors: Values used to generate a prediction. Also referred to as independent variables.

Instances: The individual people, places, things, or events described by a dataset.

Features: Values that describe instances. Target variables and predictors are both types of features. Also referred to as variables.

Proxy: Value that is measured as a substitute for the real quantity of interest. Proxies may be used to make predictions, or as a direct stand-in for things that are hard to quantify (e.g., potential or risk).

* Combining multiple data sources can be a good diversification strategy, but isn't always straightforward. Datasets can only be combined successfully if they contain unique identifiers — items like names or location that appear in several datasets and can be used to ensure that feature values belong to the same instance across all the datasets.

ML always relies on training data which have been collected in advance. Training data are used to optimize the model's *parameters.* These parameters, in turn, are what determine the model's predictions. For example, one of the simplest ML models finds a straight line that best fits a collection of points (much like would be done with traditional statistical approaches, for example). The figure above/to the right FIGURE 4 shows a very simple example, in which household income (the predictor) can be used to predict the fraction of on-time loan payments (the target variable). In terms of the taxonomy given in BOX 2, this is a "score" application, because we are predicting a numeric value: the fraction of on-time payments. The points are the training data, and each one affects where the line is located. The parameters of the model are the line's slope (0.0001) and intercept (0.2).

A–Z

Parameters: An ML model's parameters specify the rules on how it will make predictions for new data. Parameters are set during the training process.

This is, of course, a contrived example for demonstration purposes. A case like this could easily be handled with traditional statistics, and doesn't showcase the power and flexibility of ML. A more realistic ML model may involve hundreds or thousands of each variables and parameters, rather than just two.

Once training data have been used to set an ML model's parameters (0.0001 and 0.2, in this case), that model can then be used to make predictions. This is where machine learning begins to go beyond the data-fitting functions that might be familiar from spreadsheet software. Suppose, for example, that we want to predict the fraction of on-time payments (y) for a specific value of the household income, say x = 6000. In this simple model, we would simply calculate 0.0001*6000 + 0.2 = 0.8. This means we predict that 80% of payments will be made on-time. This precise value was never actually observed; instead it's the most likely output we would expect to see with the specified input. The calculations done in many ML models are often much more complex, but the idea is the same — we start with known or assumed values of the predictors and predict a value of the target variable.

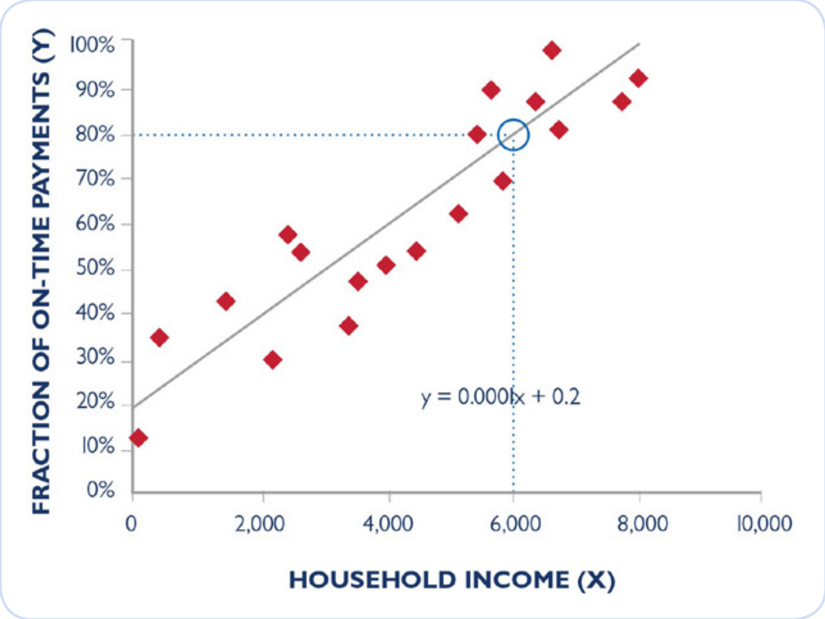


FIGURE 4: Illustration of a simple model that predicts on-time loan payments based on household income. In this very simple case, the model consists of a single equation, describing a linear relationship between income (x) and on-time loan payments (y). This model predicts that a household with an income of 6,000 will make 80% of their loan payments on time.

A–Z

Labeling: Providing values of the outcome variable for each instance in a training data set. This may require additional data collection, crowdsourcing, or expert curation.

Note that these models will Note that these models will necessarily have limitations. For example, with our simplistic model, an income greater than 8,000 won't guarantee 100% likelihood of repayment. It is up to a person to interpret this prediction and understand when the ML model's output is appropriate to apply to decision making.

Many machine learning models, known as supervised ML models, require the target variable to be *labeled* in advance. For example, a borrower's records would be annotated according to whether she actually repaid her loan. The model can then be applied to new data for which the label is not available. Unsupervised ML models search for patterns in the training data without requiring them to be labeled in advance. For example, a clustering algorithm might be used to find groups of borrowers that behave similarly. Unsupervised outlier detection could be used to spot potential mistakes in loan applications, where the information entered is very different from all others.

APPLICATIONS IN DEVELOPMENT

Applications of ML in international development are relatively new, and many projects are still at the exploratory phase. Despite this, there are promising signs of activity in nearly all sectors of development. The following sections touch on some common example applications of ML in development across all sectors; sector-specific breakdowns can be found in TABLE 1.

EARLY-WARNING SYSTEMS

Many ML algorithms are geared toward making predictions, making them useful in early-warning systems. The goal is to monitor whether conditions are similar to those that have preceded a crisis in the past. This allows attention and resources to be directed toward rapid response. Machine learning has been explored as a means of strengthening early warning systems for nutrition,[19] conflict,[20] food security,[21] and others. In most cases, early warning systems will base predictions on diverse data types. For example, a platform might integrate satellite imagery with economic, demographic, and health data.

The USAID-funded startup Grillo[22] processes vast amounts of ground motion data to generate real-time warnings about incoming earthquakes. A network of sensors and rapid trigger algorithms enable Grillo to issue warnings up to two minutes faster than existing official methods. Another example is HealthMap, an initiative that combines both expert data (e.g. reports from clinicians) and informal sources such as news reports to generate a global map of emerging disease threats in real time. Similar approaches could be used to provide early warnings of political instability, crop pest infestations, or commodity price shocks.

Of course, not all early-warning systems rely on ML. It is common for people to analyze geospatial, economic, or health data and make predictions about what might happen. One major difference is that human analysts tend to make predictions based on a small number of strong signals, such as anticipating a famine if rainfall is low and food prices are high. In contrast, ML methods excel at combining a large number of weak signals, each of which might have escaped human notice. This gives ML-based early warning systems the potential to find the "needle in a haystack" and spot emerging problems more quickly than traditional methods.

SITUATIONAL AWARENESS

ML-enabled computer vision has been applied in humanitarian settings, such as using satellite imagery to identify possible human rights violations. In one application, an algorithm was built to locate tukuls, traditional thatched-roof dwellings that are common throughout East Africa. By comparing images of the same area over time, it is possible to document the burning or destruction of tukuls[23]. Conflict monitors can then direct attention to areas of active violence.

Similar algorithmic approaches have been used in non-conflict settings as well. For example, one project aims to predict deforestation[24] by analyzing images from forested areas shortly before they were cleared. Knowing where illegal logging is about to begin could help guide law enforcement interventions. Human geospatial analysts might know to look for early indicators such as road construction in previously-undisturbed areas. The potential advantage of ML methods is their ability to quickly filter through large image databases and spot weaker, harder-to-define signals that might otherwise have been missed. In addition to working directly with image data, ML tools are uniquely well-suited to integrate geospatial data with other information, such as text-based reports. Researchers are also applying ML tools to images and text to spot evidence of wildlife trafficking on social media[25].

Other situational awareness applications rely on social media data. Crisis-response platforms such as AIDR[26] and PetaJakarta[27] sort social media posts from crisis-affected areas. This allows them to target response efforts and map affected areas in space and time. The key role of ML in these applications is to prioritize messages for follow-up action. They need to distinguish between a first-hand report ("my street is flooded") and an indirect reference ("praying for my family in flood zone"), as well as the level of severity of message. Social media data have also been used to support infectious disease surveillance, pharmacovigilance (tracking the safety of medications) and behavioral medicine[28]. Law enforcement agencies are exploring how ML-based tools can process images and text from online advertisements and message boards to detect human trafficking[29]. ML researchers have also examined linguistic cues in extremist web forums to find early warning signs of "lone-wolf" terrorism[30].

A–Z

Image Recognition: Identifying or describing an object based on a picture of it. Humans can often recognize a new or unusual object or phenomenon after seeing it only a few times. In contrast, today's ML models must be trained on hundreds or thousands of images before they can add a new item to their classification repertoire. This means that image recognition algorithms depend on large human-annotated image datasets.

Decision Trees: A type of machine learning algorithm in which predictions are made by answering a series of yes/no questions.

SUPPLEMENTING DEVELOPMENT DATA

ML techniques have also shown potential to fill gaps in data related to poverty, population density, or basic infrastructure. For example, census data in many developing countries may be decades old. This makes it difficult to plan interventions or design representative surveys. By making predictions about things that are difficult to measure, ML methods can fill some of these gaps[31]. ML can help to infer poverty levels based on structural features such as roofing material and proximity to roads and other buildings,[32] or by mobile phone usage data[33]. Computer vision algorithms can show where electric grids exist by picking out features like electric towers and power lines[34]. Algorithms can work more quickly than people, allowing larger regions to be mapped more efficiently. Similar approaches are being used to map road networks[35] from satellite imagery.

Data from mobile phone usage have been used to map climate-driven migration in Bangladesh[36] and population displacement after the 2015 Nepal earthquake[37]. In most cases, filling these data gaps requires beginning with scarce, high-value data (e.g., household surveys, electricity usage, or disease burden) and using cheap, abundant mobile metadata to predict these target data. Just because CDRs cost less than surveys doesn't make them a panacea, however. These sensitive data can be difficult to obtain and come with significant legal and privacy challenges[38].

BOX 3: Data and metadata

Metadata are data about data. While most data used in development describe something in the real world (e.g., a person, an event, a location, etc.) metadata describe a collection of data. Metadata may include information about who produced a dataset, its time and place of origin, or the meaning of variables. One important class of metadata is mobile call detail records (CDRs). In this case, the data are the actual content of voice calls or text messages. The CDR metadata does not include the content of calls or messages but can include information about the caller's location, the time and date of the call, and the number dialed. Even without the content of calls or texts, CDRs are among the most informative (and sensitive) large-scale datasets on human behavior. Metadata analysis can be powerful because it helps us bypass irrelevant detail (in this case, the contents of calls and texts) in favor of higher-level insights about where, when, and with whom people are communicating. CDRs are becoming increasingly popular in development applications. In the near future, other types of metadata — regarding purchases, media consumption, or other digital behaviors — may prove just as valuable.

POINT-OF-SERVICE DIAGNOSTICS

Computer vision has also been used for rapid, point-of-care diagnosis of diseases such as malaria,[39] hookworms and schistosomiasis[40]. Images are labeled with their disease status and used to train a supervised ML algorithm to spot infections.

Similar computer vision algorithms are used to diagnose plant disease. For example, Plantix[41] is a mobile app that provides diagnostic information to smallholder farmers around the world. The app uses an *image recognition* algorithm that can diagnose over 240 plant diseases, pests, and nutrient deficiencies. Once Plantix identifies plant damage, it returns simple information on disease symptoms, management and prevention techniques.

MARKET SEGMENTATION

Machine learning algorithms known as *decision trees* have been used to precisely target different interventions. In the health domain, the Surgo Foundation used this approach in a program promoting medical male circumcision[42]. They identified different sub-populations of uncircumcised men in Zambia and Zimbabwe and tailored messages for each group. Segmentation allowed community health workers to quickly assign individual men to groups and deliver targeted messaging.

Market segmentation has also shown promise for financial inclusion efforts. Start-ups such as Chile-based Destacame[43] have used machine learning to improve their services of providing alternative ways of assessing credit worthiness. Incorporating ML for market segmentation allowed them to improve their predictive profiles for potential users[44].

CUSTOMER AND CITIZEN SERVICE

USAID is exploring how conversational interfaces like *chatbots* may be used to fill gaps in user-facing services. The USAID-funded RegTech Accelerator — a joint initiative by USAID, the Gates Foundation, and the Omidyar network that aims to sync market movement with regulators — partnered with the *Bangko Sentral ng Pilipinas* (the Philippine Central Bank, or BSP, and the country's monetary and financial sector regulator) to support the development of a prototype chatbot.
An online platform accessible by any handheld phone via app or SMS, the chatbot is intended to field and address consumer complaints more efficiently[45]. BSP's chatbot is meant to strengthen customer communications, improve response time, and reduce the workload of bank employees. It will also facilitate more timely, efficient visibility over complaints by the BSP. Interfaces like these are increasingly common in the financial sector worldwide. For example, Teller,[46] a New York-based startup, uses messaging apps (Facebook Messenger, WhatsApp, SMS, etc.) to provide automated account assistance and financial advice. Their platform has been used in both the U.S. and Africa, and offers services for banks, microfinance institutions, and development organizations.

A–Z

Chatbot:
A computational system that engages with human users using natural language. Chatbots typically use text messages or messaging apps (e.g., Facebook Messenger or WhatsApp). Also referred to as "conversational interfaces."

PHOTO: JASON HOUSTON FOR USAID

Another USAID-funded financial chatbot is Mr. Finance, the first financial education app designed for Burmese users[47]. Mr. Finance communicates with users via Facebook Messenger in both Burmese and English. By using Messenger, Mr. Finance responds to the social media preferences of users and uses less data or smartphone memory than a standalone app would require. Mr. Finance includes a “gamified novel” to convey financial management concepts in a realistic manner, troubleshooting tips based on common business challenges, and a suite of reminders based on individual circumstances.

Conversational interfaces can also address shortcomings in service delivery due to workforce shortages. This is especially important in areas such as mental health. For example, the San Francisco-based startup company X2AI has developed Karim, an Arabic-language chatbot that acts as a mental health counselor for refugees[48]. Karim is one of many bots that offer objective “listening” and simple strategies to improve mood[49]. When people begin to express intentions of self-harm, bots are triggered to respond with prompts to reach trained professionals. While still in the very early stages of development, chatbots may be able to fill health care gaps in many developing countries.

Table 1: Illustrative Sectoral ML Applications

SECTOR	ILLUSTRATIVE ML APPLICATIONS
HEALTH	• Market segmentation approaches to inform behavioral health (Surgo Foundation) • Point of Service diagnostics using computer vision (Parasight, Excelscope) • Disease outbreak forecasting (Dalberg) • Chatbots for mental health (X2AI, Woebot) • Chatbots for reproductive health education (Girl Effect)
AGRICULTURE	• Predicting crop yields (Stanford U.) • Site specific agriculture (CIAT, Apollo) • Digital credit scoring for agricultural input loans (Apollo, FarmDrive, Ricult) • Project evaluation (World Bank)
DEMOCRACY & GOVERNANCE	• Detecting tax evasion (India, OECD) • Evaluation of crime reduction policies in Colombia (NYU) • Quantifying women's participation in community governance forums (World Bank) • Tracking media reports of violence against women (Bangladesh)
ENVIRONMENT	• Habitat monitoring (Terra-i, Rainforest Connection) • Spotting illegal fish catches (link) • Identifying wildlife trafficking on social media (link)
ECONOMIC GROWTH	• Credit scoring (Branch, Tala, Lenddo/EFL) • Improving financial regulation — suspicious activity reporting (Hummingbird) • Market information (Premise) • Chatbots for consumer complaint/service navigation (R2A)
HUMANITARIAN RESPONSE	• Social media analysis for situational awareness (AIDR) • Famine/food insecurity forecasting (NEWS) • Chatbots for service navigation (DoNotPay, Refugee Text)
EDUCATION	• Predicting drop out for need-based targeting of education intervention (Preliminary IDInsight work) • Chatbot teaching assistants (GA Tech) • Teaching tools for children with Autism Spectrum Disorder (link)
INTERNAL OPERATIONS	• Streamlining government procurement (blog) • Monitoring & Evaluation (blog)

PHOTO: SAVE THE CHILDREN

Case studies: *Machine Learning in Context*

Case study: *Data-driven agronomy and machine learning at the International Center for Tropical Agriculture*

Every morning, I get up and check the salt. It's a ritual — like going to eat breakfast.

For Alberto, a farmer in rural Colombia, decisions about what to plant and when to plant and harvest can be complex. For generations, smallholder farmers have relied on traditional practices to forecast rain, floods, and drought. Checking moisture levels in carefully-placed salt mounds is just one practice to predict rainfall and make decisions about when to plant and harvest.

More recently, scientists, growers' associations, agricultural technicians, and an increasing number of farmers have been looking to new data-driven methods to complement traditional knowledge. If applied well, machine learning (ML) can help farmers align planting, sowing, and management practices to specific local conditions. For example, an ML model can recommend crop management practices that are tailored to local soil type, plant varieties, and climate forecasts. Having data, however, is just the beginning. Getting to the point of following data-driven decisions — potentially in place of salt mounds — is a long journey. It's one that the Decision and Policy Analysis (DAPA) team at the International Center for Tropical Agriculture (CIAT), knows well. In their work to promote climate- and site-specific agriculture for Colombian smallholders, they have applied ML in several ways. Their work charts a path for those looking to leverage ML to help farmers adapt to climate variability and change and improve food security.

Why turn to machine learning as a tool?

The DAPA team at CIAT has utilized ML in several collaborations. In each case, ML has played a slightly different role.

Working with the national fruit growers' association[50], CIAT researchers used ML to identify how to maximize the yield of plantains for different soil types. The relationship between different management practices and crop yield was not known in advance. ML allowed them to identify these relationships from local data rather than beginning with a more general, theoretical model. Together with the national rice growers' association, Fedearroz, they worked to identify which varieties grow best under specific climatic conditions. This enabled Fedearroz to provide tailored recommendations about what to grow in specific regions.

CIAT also uses ML techniques to link agronomic decisions with local climate forecasts. These are combined with insights about which varieties grow best under specific climate conditions. The climate and crop information can be combined to give local seasonal forecasts. Farmers use the forecasts to determine when and what to plant.

Where did the data come from?

In CIAT's case, site-specific agriculture requires data from local farm plots and associated weather data. Fortunately, Fedearroz already had decades of records on planting dates and yields. This presented an opportunity to leverage ML for site-specific insights. But data sharing can be tricky. Any organization that has invested resources in collecting data may be reluctant to share it. Data can be a source of commercial value or proprietary advantage. Privacy can also be a concern. However, organizations can be frustrated when they don't get much value from the data they've put so much effort into collecting. This frustration can sometimes counter the reluctance to open the data and provide an opportunity to collaborate for more fruitful data use.

To explore the potential for ML, CIAT needed to build partnerships with the organizations that had local data. In working to build these partnership, CIAT researchers found that starting a conversation around their partners' current needs is one effective strategy for collaboration. Rather than simply asking for data, the team aims to understand the bottlenecks that keep partners from reaching their own goals. By first identifying impediments or knowledge gaps, CIAT's team can home in on more purposeful entry points for collaboration.

Building strong relationships isn't just important for getting access to data. It also helps set realistic expectations about what can be offered. If partners are bought into the model-building process, they can help interpret and disseminate insights. After all, ML is not magic, nor is it the right solution to all questions or problems. There's no guarantee partners will find the outputs of ML useful, and they need to be open-minded about the results of the analysis and be willing to find ways to improve it. Although Fedearroz's interest stemmed from a desire to explain low yields to their members, there were several times during the collaboration when CIAT's models performed poorly for some geographical regions. The research team needed to emphasize that ML models won't always provide valuable insights; sometimes the resulting model will only pick up noise in the data rather than true patterns.

How were ML models developed?

Supporting agricultural goals with ML requires a broad range of skills. Often, teams must be interdisciplinary. In addition to having staff trained in computer science and ML, the CIAT team draws upon many other areas of expertise to build and validate models and disseminate results.

Subject matter expertise — in this case, local agronomy expertise — is key to refining and interpreting ML models. Machine learning algorithms find patterns based on associations, some of which will not always be meaningful. Variables that appear correlated may actually be redundant; others may be related to an additional underlying factor that may not be apparent from the model itself. In these cases, it's critical to have domain expertise that can help distinguish meaningful results. Having reviewers with technical knowledge is critical. For example, when investigating limiting variables for maize, the ML team found correlations between slope and runoff. Local agronomists helped them recognize that, rather than independent factors, both variables were closely tied to water balance.

At each step of model building, it is important to bring in perspectives of those who are "closer" to the realities in the field. They can help validate model results and lend credibility to the insights gained from the work. In addition to subject matter experts, this includes those who will ultimately be affected by the outputs of the model. The DAPA team sought a range of perspectives, including other CIAT scientists, field technicians, and farmers. One scientist made the analogy of having multiple filters. First the statisticians review models, then ML and agriculture experts check for obvious errors. Finally, they are reviewed by technicians and those closest to the field to make sure it makes sense based on their experience.

CIAT has also learned that the process doesn't end with model review and validation. Making information meaningful and accessible to end users requires attention to communications and behavior change. These dissemination skills may be part of the core ML team or part of a partner organization. People who can convey complex information to non-technical audiences, visualize data, and assess how information is being understood and used are critical. They can help translate information to practice, evaluate performance, and collect feedback to improve both the models and the way in which they are shared with farmers.

What happens with the outputs of the model?

Ultimately, site-specific predictions should enable farmers to make more informed decisions and improve their harvests. As one way to translate ML outputs into useful recommendations, the DAPA team works with a research program on Climate Change, Agriculture, and Food Security (CCAFS), to support Local Technical Agroclimatic Committees (LTACs). The committees meet monthly, serving as a "roundtable" that brings together scientific knowledge (seasonal climate forecasts and outputs from crop modeling) with local knowledge on production, infrastructure, and markets (farmers, indigenous groups, technicians, local traders, local politicians). Committee members share information and make agronomic recommendations based on the seasonal climate forecast data. In this way, the seasonal forecast is "brought to the table" as one piece of information that farmers can consider in a broader context. Site-specific recommendations can then be made using either traditional or novel ML-based knowledge about which agronomic practices work best.

The participatory format of the LTACs is an important part of sharing accountability. If CIAT simply issued guidelines without local input, they might lose credibility if forecasts are wrong. By engaging others in the process, CIAT helped farmers have greater ownership over the results and develop the necessary skills to critically review the forecast and use it responsibly. To help farmers engage as full partners, CIAT supported workshops to teach about probability, uncertainty, and how to interpret seasonal forecasts. As new members join the roundtables, veteran participants help them get up to speed so that they can also understand and engage in the process. The LTACs issue a jointly-authored bulletin which includes a seasonal forecast and recommends sowing and planting dates for the region. In addition to the printed bulletin, recommendations are shared digitally through a WhatsApp group that includes technicians and farmers. Technicians also visit individual farms to share recommendations.

Through these roundtables, individual farmers can learn about climate, access the seasonal forecast, and introduce another source of information into their crop management decisions. Ultimately, whether or not farmers choose to act on the information is up to them, and farmers will use multiple methods for deciding what to do. The recommendations from the roundtable may reinforce decisions that are consistent with traditional methods and provide greater confidence in a decision they were already planning to make. When the forecasts contradict traditional methods, it's up to farmers to choose how to reconcile them. For farmers like Alberto, it really comes down to experience. Where traditional methods are observed to fail, such as during El Niño and La Niña years, the information from the LTACs is a welcome addition. But it's also important to recognize that ML-based analysis and climate predictions may not fully replace traditional methods — methods that, for some farmers, are as routine as eating breakfast.

Case study: *Harambee Youth Employment Accelerator*

Harambee Youth Employment Accelerator (Harambee) works to break down barriers that traditionally exclude low-income youths in South Africa from participating in formal employment. Headquartered in Johannesburg, Harambee has opened its doors to tens of thousands of youths to help them find employment.

Harambee aims to match unemployed youth with job opportunities appropriate for their skills and potential. Highly aware of the structural barriers that have traditionally disadvantaged non-white South African youth, Harambee recognizes that many low-income job seekers may not meet traditional job qualifications. During apartheid, many non-white families were forcibly displaced to townships. This has created educational and economic disparities that still exist today. Many youths living in townships struggle to meet requirements that are based on high school graduation, literacy and numeracy scores. Instead, Harambee uses a variety of alternative methods to assess candidates' potential; they aim to assess innate ability and identify the types of environments and activities in which a specific candidate may thrive. This enables them to provide targeted training and skill-building programs to prepare candidates for successful interviews and job placement.

Corporate partnerships are a critical piece of Harambee's work. Harambee partners with South African businesses to source candidates for their hiring needs. They learn from employers which skills are needed for a particular job and work to identify candidates who would be a good match. At the same time, they deliberately separate job competencies from traditional expectations about the backgrounds of people who have them. They ask corporate partners to trust them to source good candidates even if they don't have the usual qualifications. In this partnership, a corporate partner will inform Harambee how many candidates they would like to hire by what date.

Harambee then works to identify a group of qualified candidates, deliver high quality work readiness interventions to address the risks identified by employers, and facilitate an interview process. The employer can still hire whomever they like, but working with Harambee helps youths who may have previously been overlooked get interviews and be hired into jobs in which they are prepared to succeed.

Since opening its doors in 2011, Harambee has helped more than 50,000 youths find their first job. Today, they hope to expand their services to reach more of the estimated seven million unemployed youths in South Africa. Harambee is looking to machine learning (ML) to better leverage the data they have collected over the past seven years.

Why turn to machine learning as a tool?

Employment matching in South Africa is complex. Trying to identify job opportunities that match the skills and location of a specific candidate requires working with many different variables. Millions of unemployed South Africans are spread across a wide and varied geography. Each person will demonstrate some of over 500 identified job competencies spread across seven unique ‘job families’. In addition, each candidate may benefit from some of hundreds of possible learning opportunities. Any specific individual needs a job match that accounts for geography, job competencies, and job type. Harambee also seeks to identify the specific learning opportunities that will help her succeed in her new job. This is an enormous computational task. After six years of collecting data on their clients, they need more sophisticated analytics to make the best use of their data.

Harambee is looking to ML to help solve several problems. One aim is to generate new insights about the features of a candidate that best predict success in certain types of jobs. Their current suite of assessments has enabled them be much more precise in their matching than simply relying on numeracy and literacy scores. At the same time, there is a limit to how many features traditional matching algorithms can handle. ML can help identify new factors, create more precise matches, and quantify the relative importance of different factors. Better matches will hopefully reduce the proportion of interviews that don’t result in hiring. The integration of ML into their matching process is still nascent, but offers high potential to enable Harambee to scale its services and serve more young people more efficiently.

They are also using machine learning to fill in specific aspects of successful employment for which they don’t have good data. For example, one of the biggest barriers to youth employment is transportation. The apartheid-era policy of forced relocation to townships moved many families outside the economic centers of South Africa. Today, this has resulted in long and expensive commutes for township residents, which is costly both during job searching after job placement. Transportation to and from townships can cost more than half of what an entry-level employee earns. Harambee wanted to understand how job candidates get to work so that they can avoid matching candidates to jobs for which transportation costs would be prohibitive. However, it was not straightforward to get data about the taxi routes that many candidates take. The taxi industry in South Africa is not well regulated and has complicated, unintuitive routes. Harambee was able to leverage ML to predict likely taxi routes based in part on the self-reported origin of the employee and the job location.

Where do the data come from?

Harambee is an example of an organization that is applying ML to data they have already collected in the normal course of business. Harambee's various assessments were originally independent of each other, and they have only recently been brought into a single environment for analysis. The time and effort to prepare data for analysis can be significant, and this consolidation has been a major accomplishment.

Harambee's data includes basic demographic data reported by each candidate at registration. These include, for example, name, gender, age, address, and household size. Although Harambee downplays the importance of literacy and numeracy scores, employers may still ask for them. As a result, these legacy metrics are also measured by Harambee. However, these scores are balanced by additional data points that are less tightly linked to educational background. Harambee has candidates complete an assessment of learning potential, as well as another assessment intended to measure candidates' work preferences. In some cases, Harambee may also collect additional assessment data specific to the nature of the job or job family. Harambee also collects data on how often a candidate is matched for a position, and how many are placed in a job. Finally, Harambee collects information about the experiences of their "alumni," regardless of job placement. Every couple of months, Harambee calls candidates who have registered with Harambee to deliver a phone survey. This survey asks candidates about their current job status and tries to better understand their personal "employment journey" — how they search for jobs, whether and when they are hired or why they leave employment, and what their experience is during employment.

How are ML models being developed and used?

Harambee partnered with an external technology firm in order to access machine learning expertise. One key criteria for their partner was having a local presence, as Harambee wanted to be able to have their technology partners on site. They work with DotModus, a local Google Partner, and have several machine learning experts working on site at Harambee a few days a week. This creates a very strong relationship between the "tech" team and those who run the "business" side of Harambee.

Harambee is in the early stages of utilizing ML, yet they are already grappling with the difficult questions that can arise when new insights are discovered. For example, given that transportation can be a significant barrier to retaining a job, it is a key consideration for Harambee in matching candidates to job opportunities. Harambee tries not to match candidates to positions for which the cost of commuting would be prohibitive; if it's more than two taxi rides away, most candidates can't afford to keep the job long-term. In many ways, this is a great efficiency. Candidates are set up to place in jobs where they can keep most of what they earn, rather than spend it on transportation. However, it may also mean that those who live far from any economic center will rarely be called for interviews. Harambee's work is confronting another structural barrier that places some youths at greater disadvantage than others. More than this, Harambee is testing algorithms that take into account transportation and multiple other attributes of a candidate so that they have a composite "suitability" score for the candidate relative to other candidates. This way, Harambee is able to actively adjust a candidate's position as one or more of their attributes change.

Although it's not Harambee's responsibility to "fix" transportation routes and urban planning in South Africa, they still face hard questions about what their responsibility to unemployed youths requires. Using transportation data, they are now able to identify communities that suffer from "employment deserts" and whose residents will likely be excluded from many job opportunities because of the prohibitive costs of commuting. Having data about barriers to employment gives Harambee the evidence base to also start to advocate for change in the ecosystem. As they discover similar insights, Harambee will have to continually assess their role in the broader system and figure out how data should inform their strategy going forward.

New insights can also raise questions about integration into existing practices. Right now, new insights are reviewed by the management team before they are built-in to the process for matching candidates to jobs. This allows those with many different perspectives and background to lend expertise to the interpretation of new insights. For example, household size has been identified as a relatively strong predictor of one's ability to find a job. This is just an association, for which there are multiple possible explanations. Is it because those with larger families have stronger networks? Or because they may be more desperate to find work and end up trying "harder"? Deciding how to act on these insights raises important, value-laden questions. Should those with larger families be ranked lower than those with smaller families by Harambee's system because they have a better chance of finding work without Harambee's intervention? Or should Harambee's process remain neutral to family size? These decisions involve value judgements that a ML algorithm, if simply optimized for efficiency, might gloss over without deliberation. As Harambee advances their ML work, the effort they have put into developing inclusive review processes and ensuring that both "business people" and "tech people" are involved in decision-making will help ensure ML is a tool that supports their overarching mission and organizational values.

BOX 4: Suitability: When does ML work best?

Not all possible problems are equally amenable to ML approaches. Machine decision processes are much narrower and more fragile than human ones, and a recent review[51] identifies key questions for determining a problem's suitability for ML-enabled automation.

- WELL-DEFINED INPUTS AND OUTPUTS: ML will be easier when the quantity to be predicted is clear and unambiguous (e.g., monthly rainfall levels) than when it is more subjective (quality of governance). Similarly, while it may be possible to predict food scarcity based on Twitter posts, it will be much easier to predict it based on less-ambiguous inputs like crop yields and commodity prices.

- CLEAR FEEDBACK AND DEFINABLE GOALS: If a model's predictions can be tested against something in the real world, then deficiencies can be identified and corrected. In some cases (e.g., estimating the risk of rare events) it is difficult to know whether a model is truly accurate.

- LARGE, DIVERSE DATASETS: In general, algorithms will be more accurate and less biased if training data are larger and more diverse.

- STABILITY OF LEARNING PROBLEM: ML makes predictions based on training data, and is always extrapolating from the past. If the phenomena being modeled change quickly (e.g., attempts to evade a security system), ML will only be able to keep up if new training data can also be acquired quickly.

- NO NEED FOR DETAILED EXPLANATIONS: While explainable ML is an active area of research,[52] the most-accurate models are still often the most opaque. In situations where there is a compelling need for explainability, it may require sacrificing some degree of model accuracy in order to retain interpretability. When sufficient accuracy cannot be achieved without compromising explainability, ML may not be a good option. *See BOX 6: Opacity and explainability.*

- NO REQUIREMENT OF BACKGROUND KNOWLEDGE OR COMMON SENSE: ML researchers frequently cite Andrew Ng's "one-second rule" — a task is best-suited for automation if a normal person could do it with less than one second of thinking[53]. For example, we recognize the face or voice of a familiar person immediately, without much conscious thinking. By contrast, evaluating the logic of a written argument takes more cognitive effort, and is likely to rely on information from outside the text. Many "one-second" tasks remain un-automatable,[54] because they still rely heavily on people making common sense judgments.

- TOLERANCE FOR ERROR: All decision systems make mistakes, and decisions made by machines can be just as fallible as those made by people. Relying on machines to make decisions requires honestly assessing the expected rates at which machine outputs will be incorrect — and whether those rates are acceptable. Automation may sometimes require tolerating more errors in order to reduce costs or achieve greater scale.

If a task is a candidate for ML-driven automation, we should also consider how important it is to a project's broader goals. Development of ML systems can be expensive and time-consuming, and they will yield the largest benefit if they target a process that is part of a project's "critical path." In contrast, solutions to less-important problems (e.g., detecting fraud at the end of a supply chain when most losses occur earlier) will likely have a lesser impact.

Machine Learning: What can go wrong?

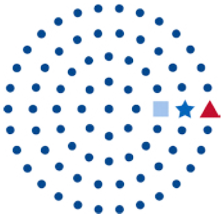
A–Z

Facial Recognition: Identifying a person based on a photograph of her face.

Face Detection: Determining whether or not a photograph contains a human face. Face detection is a building block of facial recognition systems.

There is undeniable potential for ML-based tools to bring greater efficiency, precision, and effectiveness to development work and humanitarian assistance. But while there is tremendous potential for impact, this impact is not guaranteed to be positive. Experience from higher-income contexts warns us that ML-based tools can actually result in significant harm. Relying on these systems risks unfairly targeting or excluding people. Algorithmic decisions may be faulty, and the people impacted may be unable to hold anyone accountable for the results. While the examples below are mostly drawn from a U.S. context, the underlying drivers will look familiar to development experts. These problems will likely exist anywhere with entrenched patterns of inequality or discrimination.

INVISIBLE MINORITIES



Algorithmic mistakes often fall disproportionately on minorities and marginalized groups. Consider *facial recognition,* which is used in applications such as smart phone unlocking[55] and mobile payments[56]. Researchers have documented how commercial *face detection* systems often fail to notice dark-skinned faces[57]. These failures trace back to the algorithms' training sets, which contain predominantly Caucasian faces. In addition, many observers have pointed out the relative homogeneity of the tech industry, noting that AI has a "white guy problem[58]." Lack of diversity on engineering teams may have created a situation where no one might think to test how well the system would work for skin tones that differ from those on the team. Oversights like this have real consequences. Facial-recognition technologies are becoming more widespread in the U.S. criminal justice system,[59] amid concerns that Black people could be systematically ignored, undercounted, or misidentified. Adoption of similar systems in developing countries must build on a critical analysis of the underlying ML models' construction and performance. Otherwise, we may see even more discriminatory or ineffective outcomes.

PREDICTING THE WRONG THING

In another example, researchers adapted an algorithm for earthquake risk modeling to predict crime "hotspots." The parallel was motivated by the fact that both earthquakes and criminal activity tend to cluster spatially and temporally. One important difference is that earthquakes virtually never go undetected, while criminal activity is only partially captured by arrests.

Put another way, so-called "criminal activity" data does not capture all criminal activity. When data come from arrest records, they reflect information about criminals who have been *arrested,* not necessarily all crimes that have *occurred*. Arrests require the presence of police. In some communities in the U.S., for example, police patrols are concentrated in poor and minority neighborhoods. This biases the geographic distribution of arrests. Rather than being a neutral sample of criminal activity, arrest data reflect both crime and policing. When these biased observations are fed back into the model, they reinforce the association between poor neighborhoods and criminal activity. Meanwhile, other neighborhoods are neglected[60].

A high-performing model that predicts arrests instead of crime may actually worsen the correlation between arrests and crimes. This happens because police presence is focused on an ever-shrinking area, and crimes elsewhere go unnoticed. As a result, more people in over-policed neighborhoods are arrested for relatively minor crimes, while their peers elsewhere in the city fly under the algorithmic radar. Those additional arrests can contribute to a cycle of employment challenges, more serious criminal behavior, and community mistrust of law enforcement. These dynamics can undermine progress in the places that need it most.

BUNDLING ASSISTANCE AND SURVEILLANCE

In 2016, a local government in the U.S. launched an algorithmic ***decision system*** to identify children at risk for abuse[61]. The model is based on information from law enforcement and the county welfare system. It predicts whether a call to a reporting hotline will lead to removal from the home, based on a family's past history. Its designers have published an extensive description of how they audited their model for accuracy, fairness, and racial bias[62]. While the system's goals are laudable, it has a major shortcoming. It only accesses data on families that have received public assistance, which generally correlates with poverty. When wealthy families pursue childcare, marital counseling, or drug rehabilitation, they primarily do so outside of the public assistance system. Instead, they choose privately-offered services. As a result, their experiences are never recorded and their children's risk scores are not affected. Poor families often do not have access to (more expensive) privately offered services. This means that they cannot "opt out" of the public system's tracking unless they forego assistance. Critics have charged that the system unfairly criminalizes the receipt of assistance and serves to surveil and intimidate the poor. Parents may face an untenable choice: either forego the help your children need, or risk having them taken away from you.

A–Z

Decision Systems: Decision systems are the means by which people plan or choose between options. Most decision systems use technology in some form. While simpler technologies (such as reference books) might give people general guidance, ML or AI-enabled decision systems can make recommendations that are tailored to a specific situation. Decision systems include both social and technological components.

MALICIOUS USE

The preceding discussion of pitfalls in the use of ML has assumed that people turn to these tools with good intentions. This can, however, be a naive and dangerous assumption. As with any other technology, repressive governments, unscrupulous corporations, and foreign adversaries will seek to use ML for their own ends, even if at the expense of others. For example, there are serious allegations that a UK-based company used micro-targeted internet advertisements to inflame and exploit ethnic tensions during the 2017 Kenyan election[63]. This is a global concern, and computational propaganda activities have also been documented far beyond East Africa[64]. The weaponization of ML-enabled content generation tools, such as chatbots, will likely make these persuasion campaigns increasingly more effective and harder to detect[65].

The full implications of AI for cyber, physical and cognitive security[66] are beyond the scope of this report. However, development actors should be aware of the potential for weaponization or malicious repurposing of AI and ML, even for systems that are built with neutral or beneficial aims.

BOX 5: How models fail

There has been increasing attention not only to how ML systems might fail us, but also how they can be designed to be more fair, accountable, and transparent[67]. Below are some of the most common ML failure points that researchers are working to better understand and mitigate.

- FAIR BUT INACCURATE: Some prediction tasks are just really difficult, and models may not end up being very accurate. Such models can still be useful, especially if the previous decision method wasn't any better. It's also possible for them to be fair in the sense that they are equally inaccurate for everyone.

- LESS ACCURATE FOR MINORITY GROUPS: Sometimes the relationships that are used to make predictions will be different for minority groups than for the majority population. Models that do not account for this may have impressive performance for the population as a whole, but exhibit high error rates for the minority group. For example, winning entries in a recent competition to detect buildings from satellite images achieved 89% accuracy for images of Las Vegas, but only 42% for images of Khartoum, Sudan[68]. If most of a city looks like Las Vegas (with paved roads and perpendicular streets) and a few neighborhoods look like Khartoum (with fewer paved roads and more irregular buildings), then the less-developed neighborhoods will be misrepresented.

- UNEVEN ERROR BALANCE: *"Accuracy"* can be broken down into different types of errors — for example, *false positives* or false negatives. If a model predicts loan repayment, false positives are cases where a borrower was predicted to repay, but then defaulted. If the model predicted non-payment but the loan was repaid, then the error is a *false negative.* It is possible for a model to have similar accuracy across two sub-populations, but for the balance of false positives and false negatives to change between different groups. A model that grants more false positives to one population and more false negatives to another creates an uneven playing field and systematically disadvantages one group.

- REPRODUCING EXISTING INEQUITIES: Training data used in machine learning are always data about the past. If we aim to change an unjust status quo, predictions based on what happened in the past might be unhelpful, even if they are highly accurate. For example, if women have traditionally faced discrimination in hiring, then an algorithm that scores resumes based on past hiring records will discriminate against women.

- DOUBLING DOWN ON BIAS: In many cases, the quantity we'd like to model isn't available and we must settle for a related value, known as a proxy. Maybe we're interested in actual levels of crime committed but only have data about arrests. Or we'd like to predict disease rates but only have data about hospitalizations. If the alignment between the "real" outcome of interest and the proxy isn't perfect, then models can develop blind spots – for example, missing un-arrested criminals or un-hospitalized sick people. When that blind spot overlaps with existing disparities, it can compound existing bias. For example, police may be over-concentrated in poor neighborhoods, due to broader societal inequities. This would mean that poor criminals are more likely to be arrested than rich ones. This disparity would then be reflected in arrest records, and reproduced in any model based on those records. Similarly, if poor people enjoy less access to healthcare, then their needs will be under-represented in medical records.

- MODEL DRIFT: Another potential problem with modeling based on the past is that the real world changes. Models that infer human behavior from mobile call detail records can be upended by changes in billing plans or service improvements. A model to predict flu cases based on Google searches eventually lost its accuracy,[69] in part due to improvements in the search interface.

A–Z

Accuracy: The fraction of correct predictions made by a model. Accuracy doesn't distinguish between false positives and false negatives, so two models could have the same overall accuracy but make very different types of errors.

False Positive: When a model falsely predicts that something will happen.

False Negative: When a model falsely predicts that something will not happen.

A–Z

Bias: Systematically favoring one group relative to another. Bias is always defined in terms of specific categories or attributes (e.g., gender, race, education level). Some types of bias are socially or ethically undesirable.

UNEVEN FAILURES AND WHY THEY MATTER

Choices made during model development can have real and far-reaching consequences. Because of this, even those without an immediate role in developing the technical workings of the model should feel empowered to play a role in engaging and understanding how these impactful choices are made. We've seen how use of data with poor representation of minority groups can contribute to facial recognition systems that underperform for dark-skinned faces. When these tools fail unevenly for different groups of people, the people affected may be unable to use a payment system, singled out for enhanced screening at border crossings, or wrongly called into a police station for questioning. The cumulative burden of this "selective" failure can be substantial, effectively compounding existing marginalization or inequity.

"Uneven failure" doesn't necessarily mean that the model doesn't work as designed. Predictions will be based on whatever patterns are in the data. Those patterns may reflect aspects of the real world that we seek to change, in which case predictions will reflect the unsatisfactory status quo. For example, if an algorithm that rates school admission applications is trained on past decisions that were colored by gender or racial ***bias,*** then the algorithm will "correctly" reproduce those same patterns. Decisions about which training data to use, and recognizing who or what may be missing — or all-too-present — will shape the model's impact on the world.

RISK FACTOR: EXCESSIVE TRUST

Although many factors contribute to the production of harms, the disproportionate trust that is often placed in ML-based tools is particularly worrisome. When these tools are not only fallible, but used at scale, they can be sources of significant harm. Excessive trust can be dangerous when it leads to unquestioning acceptance of model results, which can result in misinformed choices when models do get it wrong.

People often have unrealistic expectations for algorithmic systems. In situations where models make the same number of mistakes as people, we will often tend to forgive the humans and give up on the algorithms[70]. Even if people know better than to act on model mistakes, receiving useless or irrelevant advice may lead them to disengage. By association, people may also come to lose trust in organizations that prematurely implement a model's use. The layers of trust between development practitioners and the communities they serve are complex and made no less so with the introduction of AI and ML technology. If development actors place unmerited trust in a model, this may ultimately lead to irreparable loss of trust elsewhere.

RISK FACTOR: SYSTEMATIC EXCLUSION

Many ML applications aim to improve the efficiency of resource allocation through more precise targeting of services. For example, by predicting who is likely to pay back a loan, creditors can limit losses, lower costs, and serve more customers. However, when predictions systematically disfavor some groups of people, they can reinforce exclusions and deepen marginalization. The three examples cited at the beginning of this section could easily overlap to impact the same people with reinforcing algorithmic failures. ML mistakes often further stack the deck against people who were already vulnerable. Many developing countries suffer from inequality and marginalization that could be compounded by poorly-executed ML tools.

Another way that algorithms can exclude is through biased *feedback.* Imagine that an algorithm used in hiring happens to give unfairly low scores to qualified women. This will result in fewer women being hired at the company. If decision rules are updated based on the success rates of recent (predominantly male) hires, women will be missing from the new training data set. Over time, this kind of biased feedback can lead to a situation where users never know that the predictions are wrong, because all of their data are filtered by past predictions — the model generates self-fulfilling prophecies that can't be disproven. In this example, training-set challenges would include data representing a lack of women employees in a historically sexist hiring environment.

A–Z

Feedback:
A system exhibits feedback when its outputs influence its inputs. In a ML context, this can happen when model predictions influence what data are available to train future iterations of a model.

Although we often hope for ML models to provide more fair and objective decisions, the dynamics that arise can lead to a very different effect: automating the status quo. The same people who experienced discrimination before model development are still shortchanged, but the new discrimination is hidden under a veneer of computational impartiality. This is a challenge to be particularly sensitive to in development contexts, where societal inequalities may be long-standing or structural, making them difficult, if not impossible, to "correct" for in a model.

RISK FACTOR: OPAQUE DECISION-MAKING

In general, we prefer decision systems that are accountable. This means both that one can explain why a particular decision has been made, as well as assign responsibility for actions taken based on those decisions. ML models can undermine accountability in both of these senses. For example, many U.S. jurisdictions have adopted algorithmic tools that aim to predict whether criminals will reoffend when released from jail[71]. One popular algorithm is so complex and opaque that a corrections official described it as a "giant correctional pinball machine." For an individual case, it can be nearly impossible to point to the precise factors that led to a low or high score.

Of course, people also demonstrate bias and can misjudge character. The difference is that individuals can (sometimes) be identified, complained about, and called to account for their actions.

ML adoption is sometimes driven by a sense that technology will make difficult choices easier. In development programs, need often exceeds resources, and choices about who should receive help are uncomfortable to make. At times, decisions that impact individuals — about offering credit, or granting parole, or admission into a school — will negatively impact some while benefiting others. It might seem that technology can ease some of the moral discomfort of deciding who does or does not benefit by taking difficult or controversial decisions out of our hands and making them quantifiable and ostensibly objective. If things go wrong or someone complains, it is easier to blame a computer than to own a decision that caused harm or undue burden. Less-scrupulous actors may even desire to use so-called "impartial" technology as a smokescreen for their real intentions. In reality, algorithmic decision systems never free us from making uncomfortable choices; they simply displace or mask those choices.

Some researchers have pointed to "moral crumple zones" in automated decision systems[72]. Just as cars are designed with crumple zones that absorb the shock of an impact, the frontline users of partially-automated systems can be blamed when any part of the system fails. These operators often aren't the model designers — they're the social workers, police officers, or humanitarian workers who end up implementing algorithmic decisions. Algorithms may seem to simplify decision systems by making decisions more formalized, consistent, and impersonal. In reality, decision systems become more complex as the influence of human discretion becomes less visible, pushed into the gaps between people, machines, and policies. In the next section, we'll see some of the places where this human influence can be found, buried in the inner workings of ML systems.

BOX 6: Opacity and explainability

Many ML decision systems are opaque, in the sense that people cannot easily understand the process by which they make decisions. ML models whose inner workings are inscrutable even to their designers are typically referred to as black-box models. White-box models are those in which decision rules can easily be interpreted (and even checked by hand). Opacity stems from a few sources, related to both the models themselves and the context in which they are used .

Model owners may intentionally keep their systems opaque, in the interest of security or competitive advantage. Opacity may also result from technical illiteracy, where users lack the capacity or interest to understand how their tools work. There are also inherent features of ML models that can make them hard to understand, even for those with the right incentives and skills. Human brains tend to prefer explanations that involve a small number of causal factors. Models with hundreds or even thousands of parameters and features — something typical of common ML applications — can defy our efforts at explanation.

How people influence the design and use of ML tools

A–Z

Socio-technical Systems (STS): STS include technology, along with the people who make, use, and are affected by it, and the policies that govern its production and use. The different parts of a STS will often affect each other, and can't be fully understood independently of the whole.

Without question, people play a very important role in the development and use of ML tools. All of the steps of ML implementation — reviewing data, building a model, and integrating it into practice — are influenced by human decisions*. **Each step of building a ML model requires making choices that can reflect personal biases and judgments, as well as expertise and insight.** As ML models are developed into tools that inform decision-making, they become part of a larger system, interacting with people, organizations, social norms and policies. This social influence is reflected in the data the models consume, in choices that are made about how models are developed and refined, and in decisions about how the outputs of the model are used.

ML-enabled decision systems are not merely a technological tool, but part of a *socio-technical system* — a system in which technologies shape and are shaped by people, organizations, and policies.

As ML systems increasingly augment or displace the role of people in decision-making, we must understand how blind trust in models can lead to ineffective, unfair or exclusionary results. **Computational systems can scale rapidly, reaching millions of people before their effects are fully understood. For these reasons, it's crucial to recognize how individual and social bias enters ML models — and to address these points of "hard-coded bias" before models become integrated into development work.**

REVIEWING DATA: HOW IT CAN MAKE ALL THE DIFFERENCE

Data for ML models in development typically come from one of a few sources. Many draw on various types of "big data", such as satellite imagery, call detail records, or survey data. In other cases, internal data are collected by a development organization as part of project implementation or regular business operations. Examples of internal resources could include M&E data, financial records, travel logs, or electronic medical records[74]. An organization with years' worth of data on hand may turn to ML seeking programmatic or business insights. These data might have been collected expressly for the construction of an ML model, or they may be "repurposed data" drawn from reporting, billing, or other operations. Regardless of its source, the quantity and quality of data will impact for whom a model will work — and for whom it will not.

CHOOSING DATA IN DEVELOPMENT CONTEXTS

Data related to development challenges are often scarce and difficult to obtain, especially in a digital format conducive to ML. There are a variety of reasons for this. Compared to developed countries, developing regions exhibit less ownership and usage of digital devices. Connectivity is slower, more expensive, and

* For more detail on these three steps and a more in-depth treatment of the way ML models are generated and applied, *see the Appendix "Peering under the Hood"*

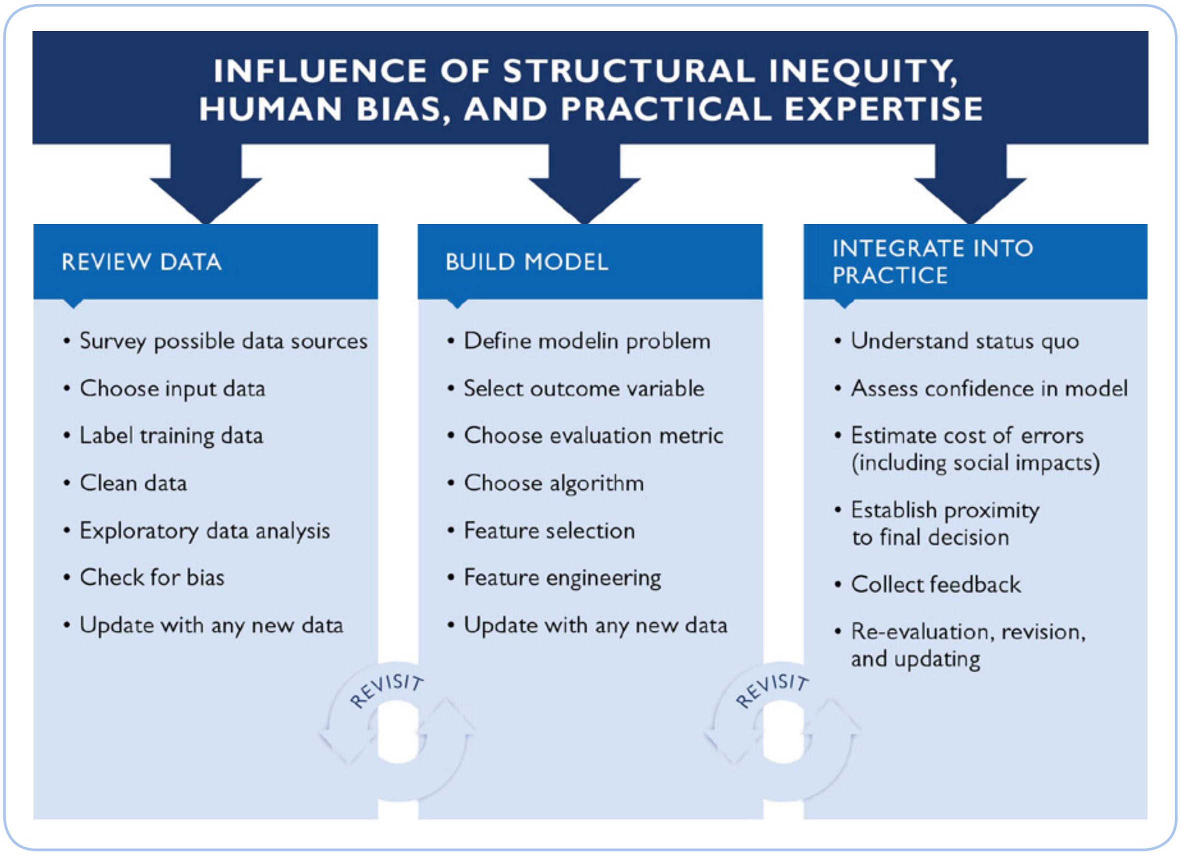

INFLUENCE OF STRUCTURAL INEQUITY, HUMAN BIAS, AND PRACTICAL EXPERTISE
REVIEW DATA
• Survey possible data sources
• Choose input data
• Label training data
• Clean data
• Exploratory data analysis
• Check for bias
• Update with any new data
BUILD MODEL
• Define modelin problem
• Select outcome variable
• Choose evaluation metric
• Choose algorithm
• Feature selection
• Feature engineering
• Update with any new data
INTEGRATE INTO PRACTICE
• Understand status quo
• Assess confidence in model
• Estimate cost of errors (including social impacts)
• Establish proximity to final decision
• Collect feedback
• Re-evaluation, revision, and updating
REVISIT
REVISIT

FIGURE 5: Depiction of the ML modeling process. In addition to the three key stages of model building (Review data, Build model, Integrate into practice), human influence (top) is important at all stages of the modeling process. One must also continually re-assess a problem's suitability for ML (bottom), based on experience at each stage of the modeling process.

A–Z

Data Exhaust: Data that are generated as a by-product of digital activities such as communication, commerce, or media consumption.

more geographically restricted. As a result of this digital divide, the *"data exhaust"* that has driven ML advances elsewhere — like online browsing habits or digital purchase records — is much less available. In addition, more formal data sources (such as censuses and birth registries) are too often incomplete or absent. All of this is exacerbated in countries that suffer from armed conflict or fragility.

When it is not possible to collect or work with data that is derived from the context you're working in, there are several "general" data sets that can be a resource. For example, satellite images that capture nighttime illumination may be used as a proxy for electricity access[75]. Call detail records can be used to fill gaps in epidemiological data[76]. Data science has great potential to help us address "data deserts." At the same time, **proxies are imperfect stand-ins for the values we actually want to measure, and they can introduce distortions.** The section below on "Choosing Proxies" includes some discussion of the value and limitations of proxies.

BOX 7: Common data issues in development

SCARCE DATA: Much of the discourse around ML assumes organizations are awash in data. In international development we often face the opposite problem — we are most concerned about the areas of the world where data availability is lowest. When data are scarce and expensive, then their collection and analysis can be in tension with the primary goals of a development project.

REPURPOSED DATA: One solution to data scarcity is to reuse data that were collected for another purpose. USAID's Open Data policy[91] actively promotes data reuse, as existing programmatic data can help future projects be better-informed and more impactful. At the same time, variables in a recycled dataset may be only indirectly related to the real quantities of interest. Data may not represent the current context, or the context of data collection may be poorly understood. It's also important to consider that when data from human subjects is repurposed, their initial consent may no longer apply.

BIASED DATA: Some of the most abundant development data sources, such as satellite imagery and call detail records, may be subtly biased. For example, mobile phone metadata can be a valuable source of information about people's activity, mobility, and social networks. At the same time, women are underrepresented, as are poor and rural populations. *For more details, see BOX 8: Common data sources and their limitations.*

CHOOSING PROXIES: KNOW WHAT YOU WANT TO KNOW

The goal of supervised ML is often to use something known or measurable as a proxy for something unknown or unmeasurable. *(See the "How ML works" section above for a more detailed definition.)* Before a model is built, it may be impossible to know which possible predictors will act as effective proxies. Modelers will often collect as much data as possible — including things that may seem irrelevant — in the hopes of finding a good set of predictors. This "throw it all in" approach can be problematic, however, when the data include poorly-understood biases or omissions. On the other hand, careful analysis can sometimes reveal new proxies that are more powerful (and perhaps more equitable) than more traditional options.

For example, Harambee has focused on leveraging their own internal data to develop better proxies. One of the most important efforts that has set them up to do this, however, came long before their interest ML. Harambee recognized early on that traditional proxies for employability — literacy and numeracy scores, along with the presence of a high school diploma — were weak predictors of workplace success. Harambee used their own data to show that people who scored poorly on school tests could nonetheless perform the duties of a griller, call center agent, or other job responsibilities successfully. Rather than relying on traditional proxies, Harambee has worked for years to develop customized assessments of learning capacity, rather than accumulated knowledge. This crucial step has created a data set that separates the systemic inequities of the public education system from the future employability young adults in post-apartheid South Africa. Reliance on institutional measures of literacy and numeracy in a ML model to predict employment success would have reproduced the same social bias that pervades the educational system. Understanding the flaws in traditional proxies has set Harambee up to collect data that more fairly reflects the potential of each candidate and forms a better foundation for subsequent ML analysis.

In general, there is no simple test to reveal the best proxy for a given modeling problem. Proxies should be chosen carefully, with an understanding of the local context and the relationship of the proxy to the true outcome of interest.

Labeling often relies on *crowdsourcing* approaches to access human judgement on demand[94]. One widely-used platform for data labeling is Amazon's MTurk. While MTurk workers can be located anywhere, only those in the U.S. and India are paid in local currency[95]. Others are paid in Amazon gift cards, a currency that's not equally valuable globally. This incentive structure means that MTurk customers in other locations may struggle to get a local perspective. The AI for Disaster Response (AIDR) platform[96] uses the volunteer-based, open-source PyBossa[97] crowdsourcing platform to develop its models of disaster-related social media posts. Because human judgement is subjective, AIDR's workflow assigns each tagging task to several people to label; final labels reflect the majority opinion on how a post should be classified.

Concerns about perspective and local context can also arise with labeled data sets that are released as a public good. For example, the UC Merced Land Use Dataset[98] has been invaluable in the construction of ML tools that classify satellite imagery according to land use type. It consists of 100 images for each of 21 land-use classes. While many land-use classes (e.g., forest, beach, chaparral) will be similar across the globe, others (e.g., baseball diamond, golf course, tennis court) are much more specific to the U.S. context. Models trained on this dataset will likely underperform on images from developing countries. Similarly, the current state of the art in *sentiment analysis*[99] (the algorithmic labeling of text as being "positive" or "negative" in tone) relies on a dataset of movie reviews. Development actors interested in democracy or health should consider whether the text they will be encountering is similar enough to movie reviews to warrant using the same models.

A–Z

Crowdsourcing: Using voluntary input from a large number of people (typically non-experts) as a source of data.

Sentiment Analysis: Algorithms that attach an emotional label to natural-language text.

BOX 8: Common data sources and their limitations

ML-backed tools for development projects often use a few general data sources that are more widely available. These more-accessible sources, however, also come with their own “bias baggage” that we should be aware of as we look to use them in an ML context.

HOUSEHOLD SURVEYS

Household surveys such as the Demographic and Health Surveys[77] (DHS) or the Living Standards Measurement Survey[78] (LSMS) function as a gold standard in development data. These surveys are often designed to be statistically-representative at a national level. Most include responses from thousands of households on hundreds of questions, often with approximate geolocation. These detailed, carefully-vetted data are often used as ground truth when developing ML methods that rely on more indirect proxies.

Despite their advantages, household surveys face some logistical challenges. The long time periods and high staffing levels required for data collection make household surveys expensive, which means that they are not done often and results may be outdated. Stale data can have implications for data quality, as they may not capture relationships that are relevant in the present. Inclusion can also be a concern: younger cohorts or new immigrants are missing, for example, and household surveys are often impossible in areas that are fragile or conflict-affected.

Households to be interviewed are generally selected in a way that ensures statistically-valid results at a certain geographic scale (often national or provincial). Geo-referenced results may be available at a finer scale. Given this disparity, over-interpreting the survey results for smaller geographic areas is a common mistake, effectively reducing data quality in the areas of interest. This can be problematic when survey data are used to train ML models that rely on more-granular satellite imagery or mobile metadata.

Most importantly, household surveys tend to be oriented toward sectoral concerns. As one might guess from its name, the DHS survey focuses on health issues, with many specific questions about malaria and HIV prevention. While DHS also collects economic data, the questions about income and asset ownership are less nuanced than those in LSMS surveys. ML model builders needing relatively up-to-date information will often need to rely on surveys with sectoral interests that may differ from their own. This can increase their reliance on proxies, as the true variables of interest may be unavailable.

MOBILE PHONE METADATA

In many developing countries, mobile phones are the most widespread source of on-the-ground location, timing, and volume of mobile usage can be a rich data source. The multinational mobile network operator (MNO) Orange recently conducted a series of "Data for Development" competitions. Orange shared anonymized data[79] about tower-to-tower traffic (voice and text) and individual mobility patterns for customers in Cote d'Ivoire and Senegal. Researchers returned submissions[80] applying ML methods in areas including agriculture, health, transportation, demographic mapping, and energy infrastructure planning.

While CDRs can be a useful source of data on people's activity and mobility, they also come with limitations and potential sources of exclusion bias. Most importantly, not everyone is equally likely to own or use a mobile phone. In low and middle income countries, women are estimated to be, on average, 26% less likely to use mobile internet than men[81]. This gap increases to 70% in South Asia. In addition, mobile ownership is less likely among populations that are poorer or rural. Household survey data (such as DHS or LSMS) can be valuable in establishing the geographic and demographic correlates of mobile ownership. At the same time, infrequent (and expensive) surveys often cannot keep up with rapid growth in ownership rates.

Research has shown that typical patterns of usage differ between developed and emerging mobile markets. Rather than the subscription plans common in rich countries, most mobile customers in developing countries use prepaid plans that lead them to ration their usage. Many developing countries allow mobile operators to charge higher fees for out-of-network calls. As a result, many customers will carry multiple prepaid SIM cards and swap them out based on whom they plan to call. All of this undermines the notion of a one-to-one match between mobile numbers and people. In such cases, SIM card-based tracking may not be a high-quality source for user behavior and mobility.

Finally, there is no universally-accepted format for CDR data, and combining datasets from different operators can be extremely difficult. Because CDRs contain sensitive customer information, mobile network operators may be reluctant to share them, either due to legal risk or protection of corporate data assets. (One proposed solution[82] emphasizes open *algorithms* rather than *open data*. The idea is to keep data in place on secure MNO servers while allowing researchers to submit code that is run behind the corporate firewall).

BOX 8: Common data sources and their limitations (Continued)

All of this has implications for the use of CDRs in ML applications. CDR-derived estimates of human mobility and social networks may not represent the entire population. For example, if mobility information is used to plan transportation infrastructure, then the exclusion of rural users may lead them to be (further) neglected. If information about social networks is used to model disease spread, then ignoring women may lead to inaccurate models. Any applications that require detailed information on individual movement patterns can be foiled by frequent SIM card swapping. When different MNOs within a country specialize in serving different populations, relying on data volunteered by a single operator may also lead to biased results.

SATELLITE IMAGERY

One of the most well-developed use cases for ML in international development is the automated analysis of satellite imagery. This is an ideal use case for several reasons. The data required are abundant, have global coverage, and are often too large for people to analyze without technological tools. New technologies such as the "CubeSats" used by Planet aim to offer daily updates on the Earth's entire land surface. ML techniques for image processing can scan through these huge image datasets to precisely locate objects or identify trends. Satellite imagery can provide invaluable information about human settlement patterns, land use, and infrastructure.

While satellite imaging platforms aim for equal, unbiased coverage, several technical issues get in the way. Cloudy conditions preclude imaging, meaning that certain regions or seasons (e.g. monsoon conditions) may be absent from image databases. While satellite imagery offers a rich (and literal) "bird's-eye" view of things, some regions or conditions might be under-imaged.

More importantly, bias can creep into remote sensing models when they are applied outside of the circumstances under which they were trained. For example, the color of roads may vary based on paving material or soil characteristics. As a result, road-detection models would be less accurate when applied in a new region. More subtly, the appearance of background features will change when moving between ecozones (e.g., from a forested to a desert region), possibly compromising model performance. Buildings and infrastructure can also change with geography. For example, a model designed to detect round tukuls with conical roofs will likely fail on rectangular tukuls with pitched roofs. This can result in exclusion of people whose geographic context or building practices differ from the majority population[84].

SOCIAL MEDIA

Another voluminous, readily-available data source comes from messages posted to social media platforms. For example, the AIDR platform[85] aims to use social media reports of disaster conditions to aid in real-time response. In Uganda, a team with U.N. Global Pulse analyzed social media messages to better understand public reactions to the country's first televised presidential debate[86]. In addition to digital social media platforms like Facebook and Twitter, it is possible to "listen in" on less-connected rural populations through automated transcriptions of radio broadcasts[87]. Social media can offer a low-cost proxy for public sentiment in real time, enabling development actors to be more informed and responsive.

Social media posts can be a rich and accessible source of data. Unfortunately, they can also over-represent the voices of the wealthy, urban, literate, and male. Social media uptake varies widely across geographies and cultures. By over-relying on data sources that neglect women and poor or rural populations, development actors may seriously misjudge public sentiment.

Social content is typically voluntary and uncoerced, but the topics we are interested in may not feature prominently in social data sets. Even in countries where water and sanitation are pressing issues, these are unlikely to attract as much social media buzz as celebrity scandals and soccer. Finally, there is significant evidence[88] that social media platforms are vulnerable to disinformation campaigns, sometimes using automated tools. Both of these factors can lead to a lower-quality signal for studying topics of interest for development.

ELECTRONIC HEALTH RECORDS

Recent years have seen a push to digitize health information systems in both developed and developing countries. In some countries, routinely-collected patient healthcare data are beginning to rival the human genome in scale and complexity[89]. Machine learning algorithms have been applied to patient data from U.S. clinics, with impressive results[90].

While health information systems in developing countries may not yet generate the same volume of data, important advances are taking place. In recent years, open-source health information systems designed for the development context have become more prevalent. These include iHRIS (for health workforce management), DHIS2 (for health information management), and OpenHIE (for health information sharing). As digital health information systems become more widespread and capable, it will likely become more feasible to build advanced ML applications on top of them. In particular, the push for standardization and interoperability in health data may enable the rapid expansion of ML-based tools. Consistent, widely-adopted standards allow software developers to developing products that can be used in many different health systems, rather than creating a bespoke tool for a specific client.

Data labeling can also be a powerful way to engage local communities and build both models and capacity at the same time. For example, the USAID-funded YouthMappers project[100] mobilizes a network of universities around the world to get young people involved in mapping their own communities. USAID projects are able to get the hyper-local knowledge and awareness that they need, while mappers receive leadership experience and training in open-source geospatial tools. The Humanitarian OpenStreetMap Team[101] leverages both local mapping teams and an international network of volunteer mappers. These efforts are not directly tied to ML model development, but create open, locally-informed geospatial datasets that can be used in ML applications.

ASSESSING DATA QUALITY: GARBAGE IN, GARBAGE OUT

While data are rarely perfect, low-quality data may limit the use of ML tools. For example, in CIAT's work with site-specific agriculture in Colombia, the maize growers' association was interested in collaborating to develop site-specific recommendations for maize farmers. When they shared their data with CIAT researchers, it became clear that the existing data had too many gaps and inconsistencies to be used in modeling. Relying on high-quality agronomic data from elsewhere was also a poor option — it would defeat the purpose of site-specific recommendations. CIAT had to communicate that they could not develop predictive ML models without more complete and standardized data. They worked with the growers' association to improve their data-collection and management capacity. This collaboration led to the development of an online platform where farmers can directly enter data on their crops, management practices, and outcomes. In this way, they are cultivating strong partnerships and a robust data set for future ML endeavors.

Deciding whether available data are adequate to move forward with a machine-learning endeavor is an important first step. Training data should be both of high enough quality to have complete, trustworthy instances to train an algorithm. While most data will likely reflect some bias, the important part is to recognize and manage it going forward.

MODEL-BUILDING: WHY THE DETAILS MATTER

Model building and selection involve decisions about how to best represent different aspects of the real world in a computational framing. While these choices may seem purely technical, they have very real implications for a model's fairness and development impact.

Modelers must decide whether the goals of a development project would be better-served by a simple, easy-to-interpret model or one that is more accurate but harder to interpret. Similarly, they will have to make choices about where it may be important to improve accuracy for some groups, even if it comes at the expense of accuracy for others. Choices must be made about how to define groups and which features should be lumped together or considered separately. Many other assumptions must be made in the process of building a functioning model.

All of these choices will be informed by things like which data are available, what the model's intended use will be, and estimates of model performance. In this section, we highlight several key decisions made during the process of building, tuning, and evaluating a ML model. These details include choices regarding the output variable, *exploratory data analysis, data cleaning, model selection,* and *model evaluation (for more technical details on each of these, see "Appendix: Peering under the hood").*

These model-building details might seem technical and arcane, and their effect on development outcomes may not be readily apparent. However, these are precisely the details which development practitioners should feel entitled to inquire about, interrogate, and ultimately inform. If we are to be fair, effective, and inclusive with our use of AI and ML tools, it is critical that "non-technical" development experts engage with the model development process. This section aims to provide you with what you need to enrich technology conversations with your awareness of context and sensitivity to development impact.

CHOOSING AN OUTPUT VARIABLE

Possibly the most crucial decision in the development of a supervised learning model is the choice of output variable. In general an output variable should be easily quantifiable, unambiguous, and closely related to the problem at hand. Unfortunately, there are often trade-offs between these goals. In the case of an employment matching service like Harambee, the ultimate goal is more than just getting a job. It's about putting young people on a pathway to success in the formal economy. Rather than simply relying on job placement, this broader goal might be better served by looking at factors like job satisfaction, growth, and retention. These outcomes, however, would likely be difficult to measure directly. For example, "job satisfaction" can be subjective and long-term follow-up can be challenging. Instead, an outcome like "hired" or "not hired" will be clear and unambiguous, though potentially less aligned to long-term goals.

A–Z

Exploratory Data Analysis: Preliminary analysis aimed at understanding the contents and limitations of a dataset. This is typically done before constructing more sophisticated models.

Data Cleaning: Preparing a dataset for analysis. This may involve standardizing definitions, changing units, removing implausible values, etc.

Model Selection: Rather than building a single model, ML workflows typically build several models and choose one that best matches their design requirements.

Model Evaluation: Quantitative assessment of a model's performance, according to pre-defined |criteria.

Similarly, a model designed to detect fraud might be trained to detect patterns of behavior that have led to convictions in the past[102]. While convictions are quantifiable and unambiguous, they are an imperfect proxy for fraud. Other factors (such as the cost and difficulty of investigation and prosecution) will influence whether suspects are convicted. Instead, it would be better to train a model to predict something with less dependence on these exogenous factors, like, for example, the opening of an investigation. In general, it will be important to choose output variables so that the model isn't learning the (potentially inequitable) features of the broader system.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a process by which data scientists become familiar with a dataset. Through EDA, they learn about its contents, structure, and potential biases. EDA may not seem like a step at which crucial modeling decisions are made. However, socially-aware EDA processes can help mitigate problems before they arise. For example, EDA is the ideal stage at which to question how existing social inequities might be reflected in data. It can be a good time to test for under-representation of important demographic groups. It is much easier to design an algorithm for fairness when potential problems are understood early. For example, if some demographic groups are underrepresented in training data, methods such as weighted regression can be used to compensate by increasing their influence relative to other groups present in the data.

DATA CLEANING

Distortion or bias can be introduced at each step of data cleaning and preparation. When data are collected from people, data cleaning may amount to re-interpreting their responses or attempting to fill gaps. These steps may misrepresent the views of the original data subjects. For example, the Americas Barometer public opinion surveys[103] routinely ask Latin Americans to name the biggest problems facing their countries. Questions like these usually elicit a few popular responses, along with a diverse group of rare, unique responses that are hard to group together. To avoid overly complex models, analysts will need to reduce the number of unique responses. This may mean grouping some of the most similar responses together (e.g., do "Economy", "Unemployment", and "High Prices" refer to the same problem?), while the rare responses might be combined into an "Other" category. This process of re-interpretation may unintentionally silence localized concerns or nuanced distinctions.

Data cleaning unavoidably encodes some of the data cleaner's assumptions about which data points deserve more attention, or the reliability of sources. As with EDA, this can be an opportunity to document assumptions and explore the effects of making different choices. This process is iterative — analysts will often return to EDA after data have been cleaned and prepared to see how the dataset as a whole has changed.

MODEL SELECTION

Model selection offers another opportunity to explicitly design for fairness. Developers might choose to evaluate a model according to anti-discrimination measures such as distinguishing error rates and types for different sub-populations. Model evaluators can test counterfactual propositions by assessing how outcomes might change if the data are modified. For example, one could invert the gender of all people in the model, to see if their individual predictions change. Error estimates are important for assessing the model's limitations, defining appropriate scopes of use, and managing user expectations.

For example, the Center for Effective Global Action is supporting research to better understand the impacts of gender-differentiated credit scoring models in the Dominican Republic[104]. They have found that explicitly including gender as a feature in the model would give significantly more women access to credit than models that omit gender data. **When the pattern that best predicts an outcome differs between two groups (e.g., when the predictors of loan repayment are different for men and women) members of the minority group tend to be misclassified more often. In these cases, including features that distinguish between the two groups can improve equity across them.** This requires a deliberate choice to check how misclassification rates compare between groups, and whether to incorporate such features in the model.

OPACITY AND INTERPRETABILITY

The choice of which machine learning algorithm to use in the modeling problem affects whether the model can be interpreted later on. For some algorithms it can be difficult or even impossible to explain why a particular outcome was obtained *(See BOX 6: Opacity and explainability).*

In some cases, a significant goal of building a ML model is to identify the features that most strongly influence the outcome variable. For example, imagine a scenario in which several crop management practices are known to affect yield, but their relative importance is unknown. In this case, interpretable model parameters are necessary to determine which practices to recommend.

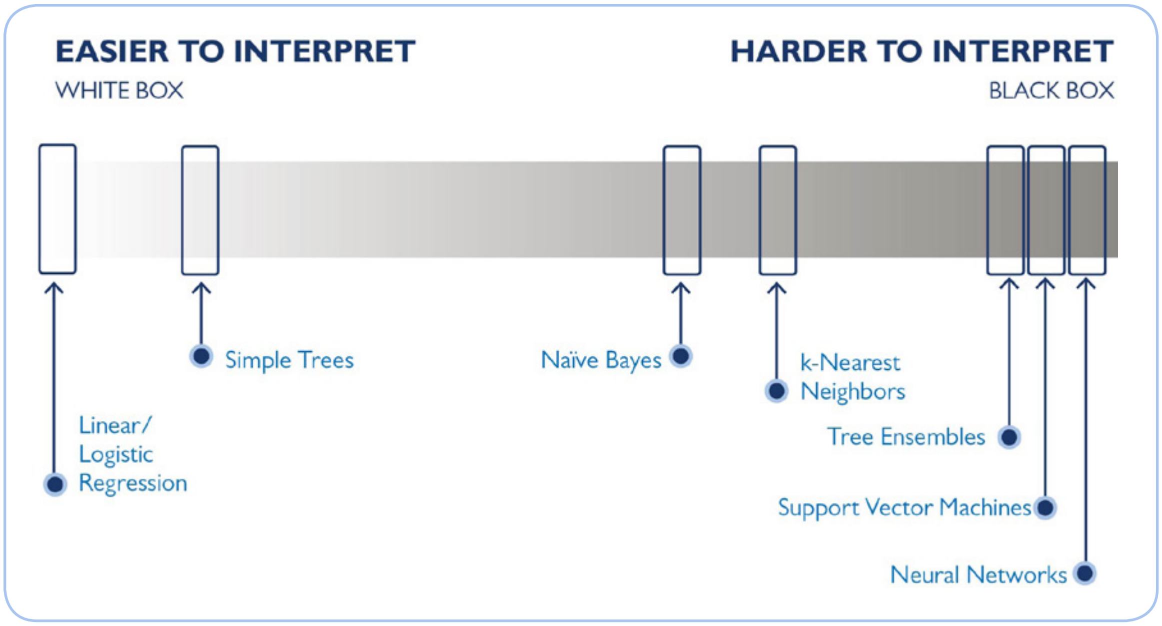


FIGURE 6: Rough categorization of some ML model types by their ease of interpretation. Some are highly-interpretable (so-called "white box" models) while, others are inscrutable "black boxes."

A complicated model with uninterpretable parameters might provide accurate yield estimates but little actionable information. Similarly, a loan applicant who was denied credit by an algorithm might want to know how to improve her score in the future. Such explanations require some degree of model interpretability if they are to be incorporated into decision making.

There are dozens of popular machine learning algorithms, ranging from very interpretable methods, like *logistic regression* and *decision trees,* to more complex and opaque algorithms used in *deep learning*. In their work with site-specific agriculture, CIAT considers a number of factors in determining which modeling approach is appropriate for their problem[105]. For example, some models excel in cases where many values are missing, where multiple types of data are used, or where there are many *outliers*. Some models are better-suited for *non-linear* relationships. Some allow for easier interpretation of parameters than others. For the work with growers' associations, seeking to provide recommendations required having some level of interpretability of models so that CIAT could explain what factors led to particular outcomes.

Logistic Regression:
Statistical model to describe the effect of predictors on a binary variable. Output variables in logistic regression take on only two discrete values.

Decision Tree:
A model that describes instances by asking a series of yes/no questions. Each question forms a branching point in the "tree." *For a visual depiction of a decision tree,* ***see "Appendix: Peering under the hood."***

Deep Learning:
A variety of artificial neural networks — models inspired by connections in the animal brain. Simple processing units are connected in layers, with outputs from one layer being input to the next. The number of layers is the "depth" of the network.

Outlier:
Data point that differs from others in the dataset so much that it is likely to have been generated by a different process. Outliers sometimes result from faulty sensors or data-entry errors. In other cases, outlier data is a "correct" measurement of something very different from the rest of the population.

Non-linear:
In a linear relationship, doubling the value of an input leads to doubling the output. Any other connection between inputs and outputs is non-linear.

MODEL EVALUATION

Once a model is built, people must make decisions about how to integrate the model into existing decision processes. This requires having a clear sense of how well the model performs — essentially, how much decision makers can trust its predictions. At a minimum level, this means getting an estimate of its *out-of-sample accuracy.* At the same time, relying on a single number to characterize a model's performance can be a dangerous oversimplification. If a model is deployed in a context marked by structural inequity (around gender, age, ethnicity, geography, or other factors), it will be important to compare error rates explicitly across these categories. Even if overall accuracy remains the same, a change in the balance of *false positives and false negatives* can lead to systematic discrimination. *The "What can go wrong?" section describes this in more detail.*

Using a consistent evaluation metric in the model selection stage also allows us to compare alternative models. If, for example, the highest-performing model is only slightly better than another that is more interpretable or easier to update with new data, this might justify choosing the second-place model.

Even during the model-building stage, **model evaluation is not purely a mathematical exercise. As always, domain expertise, diversity and awareness of local context can help avoid blind spots. There's no single formula for an adequate model evaluation. Model developers and users need to creatively interrogate the model with** *across-group comparisons, sensitivity analyses,* **and other contextually-rooted performance tests**

A–Z

Accuracy:
The fraction of correct predictions made by a model. Accuracy doesn't distinguish between false positives and false negatives, so two models could have the same overall accuracy but make very different types of errors.

Out-of-Sample Accuracy:
The accuracy of a model when applied to data that were not used to train the model. This is typically lower than the in-sample accuracy, which measures a model's accuracy on the data used in training.

False Positive:
When a model falsely predicts that something will happen.

False Negative:
When a model falsely predicts that something will not happen.

Sensitivity:
The degree to which outputs change as a single input is changed. Many models will show much higher sensitivity to some features than to others.

Across-group Comparison:
Estimating the accuracy of a model separately for different sub-populations (gender, ethnicity, geographic, etc.). This can be important to ensure that the model performs equitably across these categories.

INTEGRATING INTO PRACTICE: IT'S NOT JUST "PLUG AND PLAY"

Any ML model is only as good as its integration into practice; a mediocre but well-applied model is better than a superb model that is unused or misused. **Determining whether, when, and how to integrate ML models into decision-making processes can significantly influence how effective, fair, and inclusive the model is in practice.** This section will briefly introduce an "automation spectrum" that illustrates the relationship of models to decisions. It then focuses on three elements of model integration:

1. Evaluating the model in context,
2. Determining how a ML tool should augment, displace, or replace existing decision-making processes,
3. Considering how the introduction of ML will influence future decision-making processes.

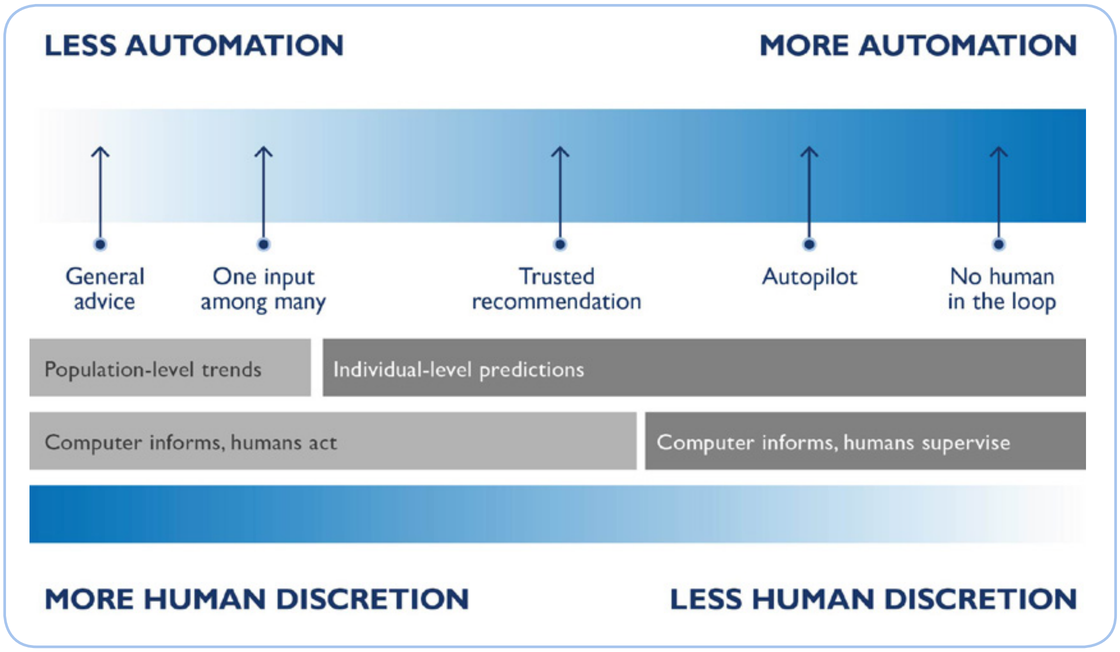


FIGURE 7: Stages of automation in AI tools. These range from models that offer general advice based on population-level findings, through those that make individual-level predictions, to full automation with no human input. This high level of automation is likely to remain very rare in development applications.

THE AUTOMATION SPECTRUM

ML and AI tools are often used to automate tasks, allowing them to be done more quickly, cheaply, or safely than by human workers. Task automation can be arranged along a spectrum, as shown in FIGURE 7. *(For more discussion of this figure and the nuances of automation choices, see "Appendix: Peering under the hood.")*

This spectrum ranges from analyses that only offer general advice to systems that make and implement decisions without people "in the loop." A recent study of automated decision systems[106] defines their "constitution" as a mixture of software, human discretion, and policies. Most automation is partial, and the interaction of these three components is key to a system's real-world impact.

Automation can bring risks. Most obviously, systems without a person in the loop may take dangerous or nonsensical actions that no human would approve. However, people can also be a source of risk — they suffer from unconscious bias, become fatigued, and can be manipulated or corrupted. Which source of risk (human- or machine-borne) is more important to avoid depends on what kinds of decisions will be made and the in-context tolerance for different kinds of mistakes.

Automation also brings opportunities, and "perfect" shouldn't be the enemy of "good." Whether they are made by people or machines, decisions are often made based on whatever inputs are available at the time. An imperfect system with limited inputs may still offer significant improvements over the status quo. Data and the models they support are rarely perfect, but something may well be better than nothing.

As described in the case studies above, both CIAT and Harambee are retaining a significant role for human judgement. CIAT relies on the review of models by agronomists and field agents, and additionally engages farmers in discussions about how to translate the findings of ML models into actionable recommendations. Harambee, while earlier in the process of adopting ML into their operations, nevertheless retains significant control over their employment matching process and ensures that new insights from ML analysis are reviewed collaboratively before incorporating into their process.

A–Z

Model Drift: Gradual loss of model accuracy, due to changes in the relationships used as the basis for predictions.

EVALUATION IN CONTEXT

At the most basic level, evaluation of a decision system's performance requires continued monitoring of its prediction accuracy. The real world is never static, and many systems experience ***model drift,*** in which the relationships between model variables shift over time. This can lead to deteriorating model performance as a model that was optimized based on a static "snapshot" becomes increasingly out-of-date as the real world evolves beyond the point where the snapshot was taken. It's impossible to know how quickly a model might become "stale" — the only protection against model drift is to have an independent source of information about its accuracy. For example, the failure of Google Flu Trends was noticed because the Centers for Disease Control maintain and publish accurate records of flu cases. If something like Google Flu Trends were used as a substitute for CDC reporting, its model drift may never have been noticed. When ML/AI systems are proposed as substitutes for more traditional data systems, model drift becomes a greater risk. One strategy to guard against model drift is to periodically retrain the model with fresh data. This often requires having an independent process for labeling new data, in order to avoid runaway feedback loops. *See the "What can go wrong?" section above.*

One example from the development context is the USAID-funded startup Apollo Agriculture. Apollo uses an algorithmic credit-scoring model to determine which smallholder farmers are likely to repay loans. While they allow the model to guide most of their lending decisions, Apollo reserves a small fraction of their portfolio for "high-risk" loans. This allows them to learn more about populations who might otherwise be excluded by their model and to test whether the credit risk model is maintaining accuracy. Without this external source of learning they might risk undermining their social mission by directing ever-larger volumes of their business to the safest borrowers.

At a higher level, algorithmic evaluation can include gathering information about whether a model is actually used in decision-making. Rather than just dispensing advice, model builders can engage in dialogue with users, learning which information products are helpful and exactly how they are being used. This can help avoid a situation in which model predictions are disregarded or misinterpreted.

Feedback is an important mechanism for learning about how people are interpreting and acting on the results of ML models. As described in the case studies earlier, part of the evaluation of the site-specific agronomic and climate models CIAT develops includes regular feedback through the roundtables held with farmers. Each month, farmers can report back on whether the predicted forecast was correct, and discuss how they might update models.

Ultimately, an ML application succeeds only if it contributes to the success of the larger effort it serves. For example, Harambee's efforts to collect information on the employment experience of all the candidates that they served, regardless of whether or not they have been hired, helps them understand for whom their model is (or isn't) working. Algorithmic matching is a small part of Harambee's services, but their matches are only considered a success if they help improve employment outcomes.

The "right" role of ML always depends on context *(see BOX 9: What influences automation?)*. Context can include the application type, possible alternatives to ML, the model's accuracy and fairness in context, and how mistakes will be discovered and rectified. In some cases, relying almost entirely on the ML model may be appropriate, as there is no way people could otherwise perform the same task. One example would be quickly categorizing thousands of images of disaster-hit areas. In other cases, especially when people will be directly impacted by decisions, full reliance on models may be unwise — for example, choosing where to allocate resources in a disaster response. If people remain engaged in the decision process, they can weigh model results against broader contextual factors and use their own judgement.

It is also possible that a team might develop a functional ML model yet determine that its performance adds no immediate value over alternative methods. For example, the USAID-funded Famine Early Warning System Network (FEWS NET)[107] provides near-term and medium-term predictions of food insecurity. FEWS NET relies on many sources, including data on commodity prices, remote sensing, and agroclimatic modeling. According to interviews with FEWS NET staff, however, they do not yet employ any ML-based predictive modeling. Their concern is that data quality is often too low to support complex predictive models. In addition to traditional statistics and data visualization, they rely on human judgement to interpret noisy or conflicting signals. Recently, the Netherlands Red Cross has explored the application of ML tools to predict famine[108]. They found that models trained on open data could not yet outperform FEWS NET's expert judgement. In this case, despite developing a functioning model, the clear course of action was to defer to current methods.

BOX 9: What influences automation?

The relationship between technology, people, and policy is shaped by a variety of factors.

TEAM COMPOSITION AND PROFESSIONAL IDENTITY: Technologists may be inclined to push for more decision automation, claiming that ML tools are more efficient and objective. Subject-matter experts, on the other hand, may seek a more limited role for ML models, emphasizing the nuance and contextual awareness of a "human touch." In particular, workers with specialized training may resent the intrusion of outsiders into their sphere of expert discretion.

ORGANIZATIONAL CHANGE: Institutions sometimes change processes, policies, and staffing patterns in ways that increase reliance on models before their effects are known[109]. When top-down mandates clash with bottom-up resistance, pro-automation managers may try to force adoption by redesigning their organizations around it.

MODEL PRESENTATION: When modelers present their work to customers or teammates, they must choose how to convey its limitations. This will affect how much a model is trusted (or over-trusted). This can include subtle choices such as whether confidence intervals are reported or how error measures are explained.

MODEL PERFORMANCE: All else being equal, people are likely to place more confidence in a model that they perceive as highly accurate. It is critical that model limitations are presented honestly and accuracy is evaluated in context, so that this trust isn't misplaced.

DECISION SPEED: Often, people simply cannot keep pace with the need for rapid information processing. For example, manual analysis of drone images taken over a disaster-affected area may simply take too long. Even if a model's accuracy is inferior to a well-trained person, the speed advantage may be more important.

COGNITIVE BURDEN: There are some things people would rather not do. For example, social media platforms employ thousands of content moderators, who are responsible for reviewing reports of violent, disturbing and offensive content[110]. Some of these workers experience symptoms of post-traumatic stress disorder from constant viewing of disturbing images. It's not hard to imagine that they might welcome more algorithmic assistance.

DECISION COMPLEXITY: Some decisions are just too nuanced for computers. Recent reporting on Facebook's internal training materials[111] has revealed the maddening complexity and ambiguity of their moderation rules. These rules are complicated because they have to be; content moderation requires a level of context-awareness and cultural sensitivity that is challenging for people and impossible for today's AI algorithms.

USER BEHAVIORS: User behaviors and preferences also play a role. A study of algorithm use in journalism and criminal justice has found[112] that the enthusiasm of managers often contrasts with resistance from users. Reluctant users employ various buffering strategies to keep algorithms from influencing their work, including foot-dragging, gaming, and open critique. Instead of replacing human subjectivity, algorithmic tools may simply force it underground.

HOW ML AFFECTS EXISTING DECISION PROCESSES: AUGMENTING, DISPLACING, OR REPLACING

Many development applications of ML will not require full automation. Instead, they rely on ML models to improve some part of a larger decision-making process. Rather than being taken immediately as credible, action-guiding recommendations, ML predictions may be subject to review and scrutiny by a variety of stakeholders.

For example, the management team at Harambee consistently questions and deliberates how and when to integrate ML-derived insights into their job-matching decisions. Some are enthusiastic about letting ML 'do its work' and automatically feeding insights from ML analysis into their matching process. Others advocate for careful review of new insights before integrating them into a matching algorithm. This appears to result in a healthy push-and-pull that appeals to fundamental questions about Harambee's goals, values, and business model.

In CIAT's work on site-specific agriculture, the Decision and Policy Analysis team regularly questions how much to "trust" model results. This decision determines whether these results will be passed on as recommendations for farmers. The CIAT team judges trustworthiness by looking for convergence between multiple lines of evidence. In addition to the model, other sources of evidence could include observations made by technicians in the field and formal experiments run by growers' associations. When these sources agree, the CIAT team shares with greater confidence and supports using them as the basis for actionable recommendations.

This triangulation requires the use of multiple prediction tools. Algorithmic predictions are an additional signal to be combined with other methods. In some cases, no single approach gave sufficient confidence on its own. Instead, comparing results to understand the relationships between plant cultivars, site-specific conditions, and output led to more trustworthy recommendations. The ML model was a valuable input, but it was not given much credibility when considered in isolation.

These examples are in contrast to applications where ML insights are automatically fed into decision-making processes without review. **While automating decision processes using ML can sometimes be seen as the fastest path to efficiency gains,** these examples underline that **it's important to proceed slowly and build in opportunities for people to retain control over decision-making.**

INFLUENCING FUTURE DECISIONS

One common complaint about automated decision systems is the lack of a meaningful process to appeal poor decisions. When people make decisions, there are often mechanisms to trace the decision back to a person, who can be blamed, punished, praised, or even bribed to change their ruling — for better or worse. One possible shift that can occur as ML tools are introduced into a decision process is that decisions become less collaborative. If the people impacted by algorithmic decisions view them as fixed and unquestionable, they may feel their options are limited to either compliance or opting out. This is often undesirable, especially in the context of government services, where fair and accountable systems would provide citizens with options of explanation and appeal.

Although ML tools often have high up-front development costs, the barriers to scaling a functioning tool are low. This has allowed some ML-fueled services to reach global scale with astonishing speed. For example, the ride-hailing service Uber reached 58 countries and a $50 billion valuation after its first five years[113] — the lifespan of a typical USAID project. Even if development applications aim for more modest scaling, rapid growth can bring risks. As tools are rolled out in new contexts and for new populations, there is greater potential for misalignment between the underlying assumptions of a model and the context in which it is used. Evaluating ML tools in context and taking time to tailor them to local conditions are critically important as they are scaled.

Some proposed applications of ML in the developing world aim to fill gaps in human capacity and mitigate system failures. For example, medical image analysis algorithms can accelerate testing and treatment by bringing diagnosis from the laboratory to the clinic[114]. Chatbots are providing counseling where there aren't enough trained mental health professionals to go around[115]. This is an area where ML and AI can augment development objectives and offer clear benefits if pursued where feasible and culturally appropriate. Yet, if algorithms are used as a substitute for scarce human capital, premature automation may lead to the loss other benefits that people bring to those roles, especially in relationship-driven activities like caregiving. While ML may be a useful, even life-saving, stopgap, we should exercise caution in seeing them as a self-contained, long-term solution.

Action suggestions: What development practitioners can do today

Even without formal ML training, development practitioners can, and should, still play a key role. Our actions can help support the development and use of effective, inclusive, and fair ML tools*. We must collaborate with technology experts to develop these tools for the contexts in which we work.

Development experts and technologists generally have different experiences, skill sets, and priorities. This diversity of perspectives is both enriching and challenging, as it can increase the risk of siloing, and "translating" across disciplines may not be straightforward. When partners don't communicate, problems may go unnoticed until it is too late to remedy them. Even if we design algorithms that group similar things together, we should avoid doing the same to ourselves.

Many of the projects discussed in this report have involved collaboration between a "technology partner" and a "development partner." In some cases, the development partner may be based in a donor agency or implementing partner (e.g., as an activity or grant manager), while the technology partner is contracted to deliver an ML-dependent tool. Development-technology partnerships can also arise from situations with less formal distinctions. These include academic collaborations, co-creation efforts, or within an in-house interdisciplinary team.

PHOTO: AMY FOWLER FOR USAID

* As a reminder, *see the "Roadmap" section* for clarification of what we mean by machine learning (ML), and how that relates to the broader field of artificial intelligence (AI).

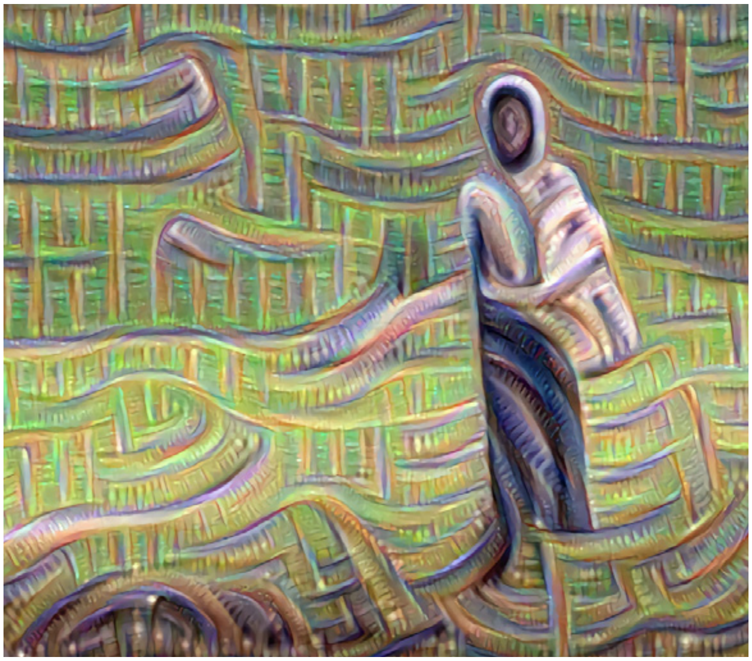


PHOTO: KELLEY LYNCH

The suggestions below focus on development partners who are exploring or collaborating with ML projects. The adoption of ML-backed tools in development projects is likely to increase. Development actors can take concrete steps now to help their organizations make the best use of these new technologies.

ADVOCATE FOR YOUR PROBLEM

Technology-development partnerships often pair "solution people" with "problem people." As a development practitioner, you can help others to stay focused on the problem and ensure that solutions don't become self-justifying. Effective technology solutions require those familiar with the problem to be outspoken, well-informed, and focused on development challenges rather than exclusively on solutions. Even if you're not actively managing a project incorporating ML, a deep understanding of your project and where new technologies will (or won't) help will set you up for future success.

One concrete way to be problem-focused is to pay close attention to which proxies are used in an ML model. Don't just settle for the proxies that are easiest to obtain or predict. Instead, work with your technology partners to find proxies that are as close as possible to what you really care about. For example, a nutrition program might be interested in caloric consumption, but only have access to data about household income. While there is often a general correlation between lack of money and hunger, factors other than income may also influence how much people eat. Rather than building a sophisticated model to predict a dubious proxy, it may be better to look for other data sources. A development partner with a deep knowledge of the model's sectoral and programmatic context can work with technology partners to make informed decisions about proxies.

Ultimately, staying true to your development problems means being prepared to walk away from ML. Consider this if you don't have the right proxies in your data, if your data isn't good enough, or if the model doesn't perform well enough to provide useful decision support.

PHOTO: ROBERT SAUERS FOR USAID

BRING CONTEXT TO THE FORE

Technology experts are often new to international development. Even when technology partners are local to the region, development practitioners have a unique and helpful perspective. They can bring much-needed awareness of some of the ways that the development context makes ML deployment more challenging than in well-known "textbook" applications.

LOCAL CONTEXT AND DATA

Given the foundational role of data in all ML tools, it is necessary to understand who or what is represented by available data, and who or what isn't. Context can influence what is recorded in data sets that may be used for machine learning. For example, structural inequities are present in nearly all societies. Large segments of society, such as those who lack official ID and work informally, may be left out of formal systems that supply census or demographic information. Reliance on data that come from formal systems may mean missing a large part of the picture — often the most important part.

Context can also affect people's willingness to share data. Especially in countries where government is dominated by one ethnic or religious group, mistrust among the disenfranchised can run deep. People may fear (often justifiably) that any collection or use of personal data could link to a government surveillance system. This can exacerbate problems with bias in data, as the most vulnerable populations often avoid participation and are thus excluded from datasets.

Context can also shape data from other sources of routine data collection. Gender norms may bias who has access to health or education systems. As a result, not everyone will be equally included in data from these systems. Rural populations, children, and elderly or disabled populations are often less represented in routine data sets. Household survey data may disproportionately reflect the perspective of men and under-represent rural populations and minority groups. Development practitioners who understand these social and institutional structures can provide valuable insight in how to assess and interpret the representativeness of data.

LOCAL CONTEXT AND APPLICATION OF ML TOOLS

The ability to adopt and maintain ML tools depends strongly on the capacity of partner organizations. Leveraging ML tools requires capacity to use as well as maintain the models from which they are built. Long-term use can be enhanced by aligning the requirements of model use and maintenance with the capacities of the organizations who will ultimately use the tools routinely as part of their work.

For example, in CIAT's work on site-specific agriculture *(See "Case studies,")* the ultimate goal was to put information into the hands of farmers. While local growers' associations were not staffed with ML experts, CIAT has found that at least some staff are often eager and able to pick up new skills. By investing in training partner staff, explaining the basics of coding environments and how ML and climate prediction models work, they can relatively quickly train others to reproduce models. CIAT has prioritized helping build their partners' capacity to replicate analyses, troubleshoot, and work with CIAT to resolve problems when they arise. CIAT's willingness to help partners learn to run their models was key in extending the life of their work.

Understanding context can also help determine when it may be possible to rely on a tool developed elsewhere. Many developing countries exhibit more internal diversity than areas where ML applications are currently being deployed. For example, most American cities have a consistent appearance in satellite images — rectangular street grids, similar road surfaces, similarly-sized houses and lots, etc. Elsewhere, building materials and settlement designs may have broader regional variation, making it more difficult to build satellite image-analysis algorithms with broad geographic applicability. Development practitioners familiar with local context can identify early on whether off-the-shelf tools are a poor match for context.

INVEST IN RELATIONSHIPS

Building effective ML-backed tools requires listening to many voices and perspectives. Development practitioners can be key advocates for investing in respectful, productive relationships over the course of both the development and use of ML models. In an ideal situation, ML tools for development projects can be built and maintained by local technology partners. By working with local companies, we can help to grow fledgling technology sectors and leverage the local knowledge and experience of technologists.

PHOTO: KAARLI SUNDSMO FOR USAID

Unfortunately, depending on local talent isn't always practical. In the absence of local expertise, many development ML projects may rely on software developers who work remotely from "tech hubs", for example in major metropolitan areas across North America or Europe. These long-distance partnerships make it even more crucial that open and frequent communication be prioritized. Take the opportunity to push for transparency from your technology partners, working to ensure that they can, and do, explain their decisions to you in terms that you can understand. Emphasize that you need to understand their approach to model development and the key choices they are making. This will give you insight into how and why the tools work as they do. At the same time, external ML experts can learn about the priorities and needs of development practitioners.

It's also important to recognize that more perspectives are likely better than fewer. While those with technical training will likely be best positioned to make technical design choices, we can still ensure that people with diverse backgrounds, subject matter expertise, and context awareness still have channels that allow for their participation in the process[116]. This can help the group involved in developing or testing a model to question assumptions and avoid blind spots. This both improves the model itself and helps create buy-in for its eventual use, assuming a good result. It's always worth asking who isn't at the table, and what they might be able to contribute. Bringing in local voices — from civil society, local governments, and affected communities — can help you become more aware of structural inequities and possible sources of bias. Even when there isn't much local ML expertise to draw on, turning to local communities for things like data labeling tasks can help you integrate local perspective and knowledge.

When planning an ML-enabled project, it's useful to think through what structures or processes could help make sure that all voices are heard. This could include regular meetings between software developers, subject matter experts, and other stakeholders. If your technology partners are working remotely, it may be helpful to bring them to the implementation site periodically, so that they can see the project context first-hand. In general, well-designed organizational processes can go a long way toward achieving more fair and representative outcomes.

Development practitioners also have an important role in investing in relationships with those who are not directly involved in the development of an ML tool, but could have influence over its shaping or be affected by the tool's use. Without trust in a model's outputs, decision makers — ranging, for example, from smallholder farmers, to frontline health workers, to policy makers — are unlikely to incorporate model-based recommendations into their routine decision making process.

Developing trusting relationships with local organizations can also open doors to more local, accurate, and timely data that is essential for ML success. In the CIAT case study, their analytical models were only possible because of existing partnerships with local growers' associations. CIAT's long-term presence in Colombia's agricultural sector enabled them to build trusting relationships that encouraged farmers to take a chance by sharing data and testing out their recommendations. Similarly, Harambee *(See "Case studies")* is investing in robust relationships with their corporate partners. They will need these connections if they want to better understand the experience of candidates after they are hired. The data that your ML projects need may come from pre-existing partnerships that will be further strengthened by productive data sharing. Especially when data is being repurposed, conversations with its original caretakers will help you understand whether it's being taken too far out of context.

Finally, invest in relationships with those who are ultimately affected by the use of ML tools. Development practitioners often seek to know their end-users better, and ML should be no different. Getting feedback directly from those who are intended to use or benefit from ML tools is an important part of testing models in context. Development practitioners have a key role to play in closing this feedback loop.

CRITICALLY ASSESS ML TOOLS

Especially when managing a grant or a contract, the development partner fills the role of a customer on whose behalf a technology tool is being developed. Understanding both how ML tools are built and how to assess their performance and suitability will help you to be an informed customer.

One of the most important actions you can take is to ask about model errors and potential bias, and make sure you understand how these were evaluated. If you're not sure what to ask for, then start with a candid discussion about how a model's errors can be quantified and what types of bias you're most concerned about given the context in which the tool is likely to be used. In particular, identify subsets of the population (e.g., male/female, urban/rural) across which error rates can be compared. If there are uneven failure rates, what real-world consequences might these have? For models that will be evaluating "live" data after an initial testing phase, it's important to ensure that error testing and performance monitoring continues after deployment.

Some ML algorithms generate models that are more easily interpretable than others, and not all applications require an interpretable model. Even for more opaque algorithms, it is possible to estimate the influence of different features on model outputs. Your technology partners should be able to estimate which features are most influential and which data sources could be omitted without compromising model accuracy. For example, if some variables present privacy concerns or are expensive to collect but don't add much predictive accuracy, you can probably do without them. Understanding which variables a model relies on most heavily will help you anticipate possible problems when the model is deployed.

As a development partner, you can also guide a more contextually-rooted integration of the resulting tool into existing decision-making processes. You probably understand the deployment context more deeply than your technology partners can be expected to, and it's up to you to be sure that you're getting something usable given that context. This integration needs to be done carefully, to ensure that ML tools are not ignored, over-trusted, or otherwise misused. This requires taking a long look at how decisions are currently made and how users are likely to interact with new technologies.

PHOTO: U.S. AIR FORCE

You may need to estimate the accuracy of status-quo decision-making processes to see how much of an improvement ML can deliver. Similarly, you should think about how much error the existing decision-making process is able to tolerate, and whether the ML-backed tool will be able to meet expectations.

Finally, for ML models that inform decisions about individual people, the development partner may need to view the model as part of a two-way communication process. If someone receives a score (e.g., for credit risk) and wants to know what she can do to improve it in the future, is the model interpretable enough to provide her an answer? If someone feels he has been wrongly evaluated, is there a way for him to seek redress? These feedback processes are often missing, even when decisions are made without algorithmic help, and correcting this is likely to be more about institutional processes and priorities than about technology. When it comes to receiving feedback, providing explanations, or correcting mistakes, it is often better to create formal channels than to rely on ad hoc improvisation. **Listening to the people impacted by our programs is always good development practice — no high-tech tool will change that.**

Looking forward: How to cultivate fair & inclusive ML for the future

ML offers significant potential to help us achieve development goals if developed and used appropriately. But we're not there yet. This guide has highlighted numerous ways in which development practitioners can help shape the use of ML in development to be effective, inclusive, and fair. For many organizations, current limitations in capacity and data availability may make any significant use of ML-based tools seem a distant reality. However, there are important investments that can be prioritized now in order to ensure we can responsibly leverage ML in the future.

STRENGTHEN LOCAL TECHNICAL CAPACITY

Given the importance of local perspectives in developing and using ML tools, we must work to consistently involve individuals who are experts in local context in addition to those who have machine learning expertise. The development of indigenous research talent can tailor new technologies to local needs, a critical enabler of innovation and sustainable progress. In many development contexts, data scientists and individuals with background in machine learning are scarce; local universities may not have appropriate specialized programs and departments. Strengthening training programs for data science and machine learning in local development contexts can help create a pipeline of individuals who are "bilingual" in the sense of understanding local context and having the technical skills to take an active role in developing ML tools.

STRENGTHEN RELEVANT GOVERNANCE STRUCTURES

At the same time, there is more to successful AI adoption than technical capacity. To help partner countries become self-reliant AI users, we must also help develop capacity for AI governance. Governments around the world are wrestling with the policy implications of AI, and in-house technology expertise is often in short supply. Even developed countries struggle to find the right balance between promoting innovation and avoiding risk. We should expect that this will be hard work for our developing-country counterparts as well.

Strong governance also requires robust laws for the protection of personal data, and the adequate resources and expertise necessary to enforce these laws. The weakness or absence of personal data protection laws is a widespread problem that can create opportunities for malicious actors to surveill and manipulate with impunity. As we strengthen local capacity for technology adoption and use, we must not neglect this critical piece of technology governance.

ENSURE RESPONSIBLE DATA PRACTICES

Data protection and privacy are likely to become even more important for development work in the coming years. Today, this dynamic is shaped by the interaction of two powerful trends. ML tools are data-hungry, and their adoption (both in development programs and in society more broadly) will increase the demand for data. Developed economies already feature a thriving market of data brokers who generate personal-level profiles and sell them to marketers and others. We should expect this phenomenon to spread to developing countries, as increasingly-connected people generate more and more data. At the same time, backlash against data-fueled technology companies is growing, with calls for more government action to protect privacy. New laws such as the European Union's General Data Protection Regulation (GDPR) have aroused concerns within the emerging AI industry[117]. Development agencies and practitioners will need to balance the utility and risk of data, in an increasingly complex legal and regulatory environment[118].

With data as the foundation of all ML/AL tools, we can always work toward increasing the quality of data that's available to development actors. We want data that are robust, inclusive, and representative of the contexts in which we work. Many of the examples of the potential harms of ML-backed tools described here expose the numerous ways in which available data can be misleading. We can take advantage of this moment of high interest and hype in ML to reflect on whether the data we collect is inclusive, representative, and trustworthy, and invest in ways that strengthen routine data collection. This will provide a stronger foundation for ML projects in the future.

ENSURE RESPONSIBLE, SHARED LEARNING

We can also invest in becoming savvy consumers of emergent technological tools. As the development community works to make our interventions more effective and efficient, it's critical that we actively investigate the appropriate use of new technologies, like ML and AI, that show promise for enhancing our effectiveness and our efficiency. For AI and ML in particular, this includes learning about the data from which tools are developed, inquiring into the process of testing and validating tools, and identifying embedded assumptions. We must actively research these tools, understanding their powers and limitations across contexts and geographies if we hope to effectively leverage them in our work.

Our research efforts must help us learn about these tools' failure as much as success. As donors, we need to recognize that requiring success with emerging technologies only makes it harder to learn and improve. Failing will be a necessary part of learning how to use ML/AI tools well; we must acknowledge this and construct appropriate safeguards that allow us to fail responsibly, transparently, and in a way that ensures failures will be learned from, not repeated. We should aim to create the mechanisms and the incentives to honestly explore these tools transparently and with support for evaluations that capture both the good and the bad, before they are rolled out at scale.

TRACK WORKFORCE IMPLICATIONS

This report has focused on the ethical implications of ML adoption in development programs. This is far from the whole story of ML and development, however. The global adoption of ML-backed AI tools will have profound implications for the ways in which countries can hope to overcome poverty. For example, increasing automation may lead to a decoupling of labor costs from overall manufacturing costs. As low labor costs become less important, some have predicted a transition from offshoring to "re-shoring", as automated production is moved closer to rich-country markets. At the same time, emerging technologies will likely also create new ways for people and countries to generate income. The changing nature of production and employment will likely be one of the key challenges for development in the 21st century.

PHOTO: NICHOLE GRABER FOR USAID

CONCLUSION

ML and AI are an increasingly important part of our digital infrastructure. As with roads and bridges,[121] the builders of digital infrastructure make choices about equity, access, and justice, and their choices will have long-term consequences. Automated decision systems can encode human priorities, ignorance, or biases — sometimes in ways that can undermine development gains. And hype can ultimately yield to distrust or disillusionment if ML and AI fail to meet inflated expectations, slowing technological investment and progress.

It may be tempting to see technology as a shortcut around political or social change. Indeed, the breathless commentary of ML proponents may seem to imply that non-digital realities will soon be a thing of the past. But even when technology-led changes are rapid and dramatic, international development is often concerned with the members of society for whom these changes take the longest to reach. By engaging with ML technologies at an early stage, development practitioners can help ensure that the people we serve aren't left behind as our global community embraces the promise offered by ML and AI.

Quick Reference: Guiding questions

Effective, fair, and inclusive machine learning and artificial intelligence based tools require careful development. The questions below should not be considered a checklist that certifies a given ML application as "good" or "bad." Instead, these questions can help us have better conversations across disciplines and fields of expertise. The goal is to engage both funders and implementers in a collaborative discussion around the process of designing, building, and ultimately using ML-backed tools in development contexts.

GAUGING SUITABILITY, FEASIBILITY, AND APPROPRIATENESS OF ML/AI

Before beginning an ML-backed project, try to assess whether it is the right tool for your problem, as well as whether it is feasible and appropriate in context.

- Why is my problem a good fit for machine learning? Would this problem benefit from a tool that could help with predicting, classifying or discovering a new relationship?
- What relevant data are available to address the problem?
- How might vulnerable or marginalized populations be affected by this tool?

CONSIDERING REPRESENTATIVENESS OF TRAINING DATA

ML models can only learn from the data used to train them. If certain populations or contexts are left out of that data or misrepresented in that data, the resulting tools may fail to work equitably for those populations.

- Are there people, communities, or geographies underrepresented or excluded from the training data set who will be affected by the outcomes of the model (e.g. speakers of minority languages, rural populations, women)?
- How might locally-collected data be used to validate the outputs of the model? Which local partners could be engaged to help validate the tool?

ASSESSING APPROPRIATENESS OF PROXIES

We often want to know about things that are hard to measure directly. Proxies are alternative indicators or variables that can be used to "approximate" what we're really interested in. But not all proxies are good substitutes; sometimes they can be only weakly associated with what you really want, and sometimes they can reflect underlying biases in how they were measured. Poor proxies can bias model output.

- Are you using proxies?
- What assumptions are embedded in your proxies?
- Given what you know about context, do your proxy choices seem reasonable?
- Might there be other variables that would be a better proxy?

BRINGING DIVERSE PERSPECTIVES INTO MODEL BUILDING

ML relies on finding patterns in the data, but ML models do not "know" anything about the patterns identified. Including subject matter experts, people who understand the local context, and diverse perspectives can enhance the quality of ML-based tools.

- How will those developing the machine learning model incorporate inputs from relevant domain experts?
- How will those developing the machine learning model incorporate inputs from those representative of the local context in which the tool will be used?
- What locally important perspectives might be missing?

DESIGNING FOR MODEL INTERPRETABILITY

Some machine learning algorithms are more complex than others. If it's important to understand which variables are informing decisions or know why a model reached a certain outcome, choosing models that are easier to interpret will be important.

- Can you identify the factors that most significantly influence the outcome of the model?
- Under what circumstances might you need to be able to explain model predictions?
- Does the level of interpretability meet the needs of your problem?

EVALUATING THE MODEL FOR FAIRNESS

Machine learning models may not work equally well for everyone. We can improve fairness by assessing model performance for inequities in failure rates and error types.

- Across which subsets of the population will it be most important to compare model performance?
- Does the model fail more often for some people than others?
- What are the consequences of differential failure rates in this context?

INTEGRATING THE MODEL INTO PRACTICE

ML models can be just one input into decisions, or, alternatively, they can be the deciding factor. It's important to consider what role the tool should play in decision-making in order to understand how well it will need to perform.

- Does the model perform better, in practice, than the existing decision-making process?
- What will using the model add in terms of efficiency, accuracy, or scope? What, if anything, will be diminished? Where might human judgement still play a valuable role?
- How will we safeguard against the malicious use of the model?

ENSURING LOCAL FEEDBACK MECHANISMS

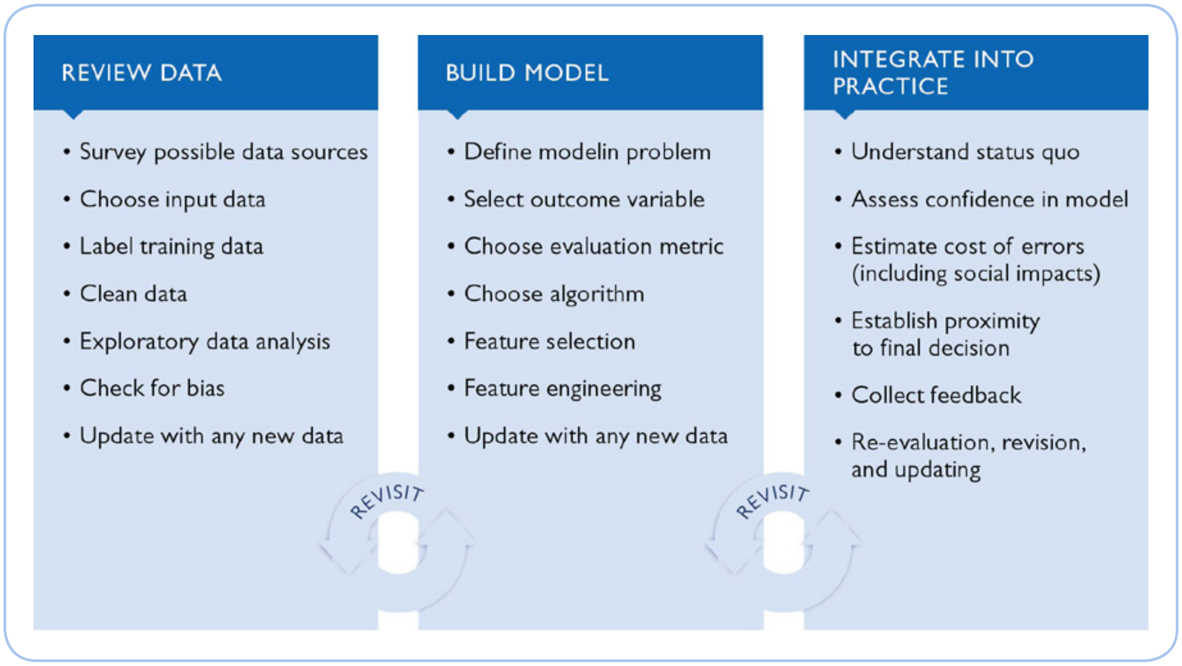
Hearing from those who use and are affected by ML-based tools is critical for ensuring they are and remain effective, inclusive, and fair.

- What mechanisms are in place to get feedback about model performance in context, and to update, maintain, or improve the model over time?
- How can local experts or end users provide feedback into model function?
- What mechanisms are there for end-users to question or contest results of ML-based tools?

Appendix: Peering under the hood

This section will provide a more detailed discussion of how ML models are built and integrated into decision processes. It is more technically-oriented than the rest of the report, and is designed for readers who want more detail on the inner workings of ML technologies. Reading this section won't turn you into a ML expert, but it should prepare you to understand the social impact of choices made during model building and implementation — topics which are the focus of the section *(How People Influence ML tools)*.

While machine learning is inherently an iterative process, it's useful to think about the development of a machine learning-based tool in three general stages: choosing data, building a model, and integrating the tool into practice. The following sections will illustrate these steps using the hypothetical example of a credit-scoring algorithm that estimates the default risk of borrowers. Text that is specific to this example will be highlighted in dark red. The process described will be similar for many other applications as well.



REVIEW DATA

- Survey possible data sources
- Choose input data
- Label training data
- Clean data
- Exploratory data analysis
- Check for bias
- Update with any new data

BUILD MODEL

- Define modelin problem
- Select outcome variable
- Choose evaluation metric
- Choose algorithm
- Feature selection
- Feature engineering
- Update with any new data

INTEGRATE INTO PRACTICE

- Understand status quo
- Assess confidence in model
- Estimate cost of errors (including social impacts)
- Establish proximity to final decision
- Collect feedback
- Re-evaluation, revision, and updating

FIGURE 8: The development of a ML model can be separated into three phases: reviewing data, building a model, and integrating the model into practice. These phases are rarely strictly sequential, as results and challenges encountered at later phases may prompt model-builders to revisit the earlier stages in the process.

CHOOSING DATA

An ML model will make predictions on the assumption that all new data it encounters are like the data used to train it. Because of this, a critical first step in the ML process centers on the data used to build the model. At this stage, model developers must answer questions around what types of information are likely to be useful for prediction. What data sources exist? Which data sources do the developers have access to? Beyond the initial data-landscaping piece, developers will make choices around basic **data cleaning and preparation**. Data cleaning steps might include:

- Standardizing formats *(e.g., converting the dates of loan issuance and repayment into MM/DD/YYYY format).*
- Geocoding *(e.g., converting a home address field into latitude-longitude pairs).*
- Simplifying a continuous range into discrete bins *(e.g., converting raw annual income into a low-medium-high scale).*
- Guessing the values of missing entries based on existing data *(e.g., using non-financial information to estimate monthly income for borrowers who declined to share it).*
- Adding new attributes *(e.g., inferring a borrower's gender based on his/her first name).*

When using supervised ML models, it's important to think about how to obtain or generate labeled training data. Often this involves combining high-volume proxy data (such as mobile call detail records, or CDRs) with a smaller set of expensive, hard-to-get data (such as surveys)[122]. *In our credit-scoring example, one might combine CDRs with borrowing records to see which patterns of mobility or communication correlate with repayment — does someone with a stable contact list and predictable calling patterns tend to be more regular with loan repayments, for example?* Another possibility for labeling is to crowdsource hand-annotation of data, as was done with the ImageNet database for computer vision. *For example, if a credit application involves handwritten forms, crowdsourced labeling of anonymized snippets could help improve computer interpretation of local languages and scripts.*

When choosing datasets for use, modelers will often use **data exploration** to examine which data are suitable for the modeling task. The goal of exploratory data analysis (EDA) is to understand the data that will be used to develop the model. This requires getting a feel for the quantity and type of data (numeric, text, video, etc.). Exploratory data analysis may involve the identification of outliers or clusters. Many datasets include missing values (e.g., from non-responses on surveys or faulty sensors), and EDA explores how missing values are distributed within the dataset. *In the credit-scoring example, one may be concerned with whether female borrowers are less likely to share their income or whether low-income borrowers don't provide addresses.*

A–Z

Imputation:
Filling in missing values, often by making algorithmic guesses based on non-missing data.

Dimensionality Reduction:
For a dataset with a large number of features, combining these to create a smaller number of features that still capture most of the useful information.

Test Data:
Labeled data that are not used to tune model parameters. Instead, they are set aside in order to evaluate model performance.

Train-test Split:
Randomly choosing which instances in a labeled dataset to use for training, and which to use for testing.

Hyperparameters:
Values that control how an algorithm learns. Often these influence the complexity of the resulting model.

Overfitting:
A modeling error in which an overly-complex model is fitted to training data. As a result, the model "learns" idiosyncratic features of the training data and cannot generalize well to new data.

While the methods used in EDA are quantitative, ML analysts will use it to gain intuition about what they're dealing with. For example, EDA can clarify the quality of a dataset and how deficiencies might be mitigated. It presents an opportunity to build for transparency and accountability by logging potential sources of error or bias and documenting how existing inequities are represented in a particular dataset.

BUILDING A MODEL

Model-building begins with **defining the modeling task.** This is much more specific and quantitative than outlining the goals of an ML system or the development challenge it aims to address. Model developers must precisely specify what the model should predict (the target variable), as well as how to quantify the accuracy of those predictions (*the evaluation metric*[124]). *Building a credit-scoring algorithm might require deciding whether to predict whether a loan was repaid, the maximum number of days that any payment was late, or the fraction of the principal that was paid back on time.* We may be more concerned about false negatives (denying credit to good borrowers) than about false positives *(lending to a risky borrower).* An evaluation metric may also include explicit anti-discrimination goals[125], such as ensuring that the false negative rates for men and women are identical. Successfully integrating such goals into the model development process requires precisely quantifying fairness and encoding it in an evaluation metric.

A well-defined modeling task should also specify a plan for judging a model's accuracy during development[126]. One common approach is to train the model on a subset (e.g., 70%) of data, then evaluate its performance on the remaining *test data* (e.g., 30%). In cross-validation, modelers will make this random split repeatedly and average the results for a better estimate of model accuracy.

Most algorithms have one or more *hyperparameters.* Despite their intimidating name, hyperparameters are easy to conceptualize. Think of them as the "knobs" that can be adjusted to control model training. While an algorithm's parameters are the outcome of learning from data, the hyperparameters control how an algorithm learns, and must be set before learning begins. Even when a model has many parameters, these can't be chosen arbitrarily by model builders. Most algorithms have only a handful of hyperparameters. The parameters, on the other hand, can't be controlled directly, and are the outcome of training an algorithm with a specific training data set and specific hyperparameter values.

Many hyperparameters affect model complexity — they can help avoid *overfitting* by constraining models to be simple. For example a decision tree could give perfect accuracy on training data by using overly-detailed rules. In FIGURE 9, the tree on the left shows some decision rules that might generalize well *(e.g., lending to people older than 35 and with more than 6 years of education)* and others that likely reflect quirks of the training data *(e.g., making an exception for older, less-educated, rural, experienced borrowers named Mary).* This is an example of where ML's strength can also become its weakness. Algorithms are unconstrained by causal assumptions and can find unintuitive correlations and decision rules. In some cases, though, this can lead to spurious conclusions that defy common sense. *In this hypothetical example, having a few exemplary borrowers named Mary and fitting this precise demographic could be enough to "convince" an algorithm that this first name (in combination with other factors) somehow enhances a person's ability to repay loans.* To avoid such problems, one could use a hyperparameter that limits the depth of trees (e.g., to only two layers) so that decision rules are less precise, but simple enough to generalize. Modelers will typically use systematic hyperparameter tuning to find the optimal balance of flexibility and generalizability.

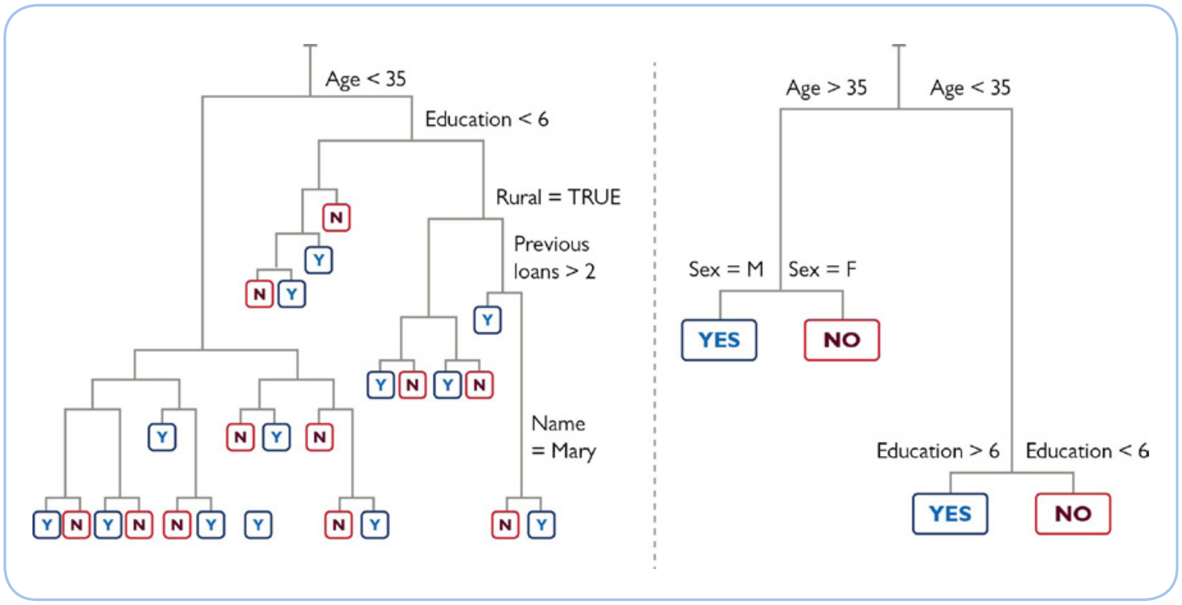
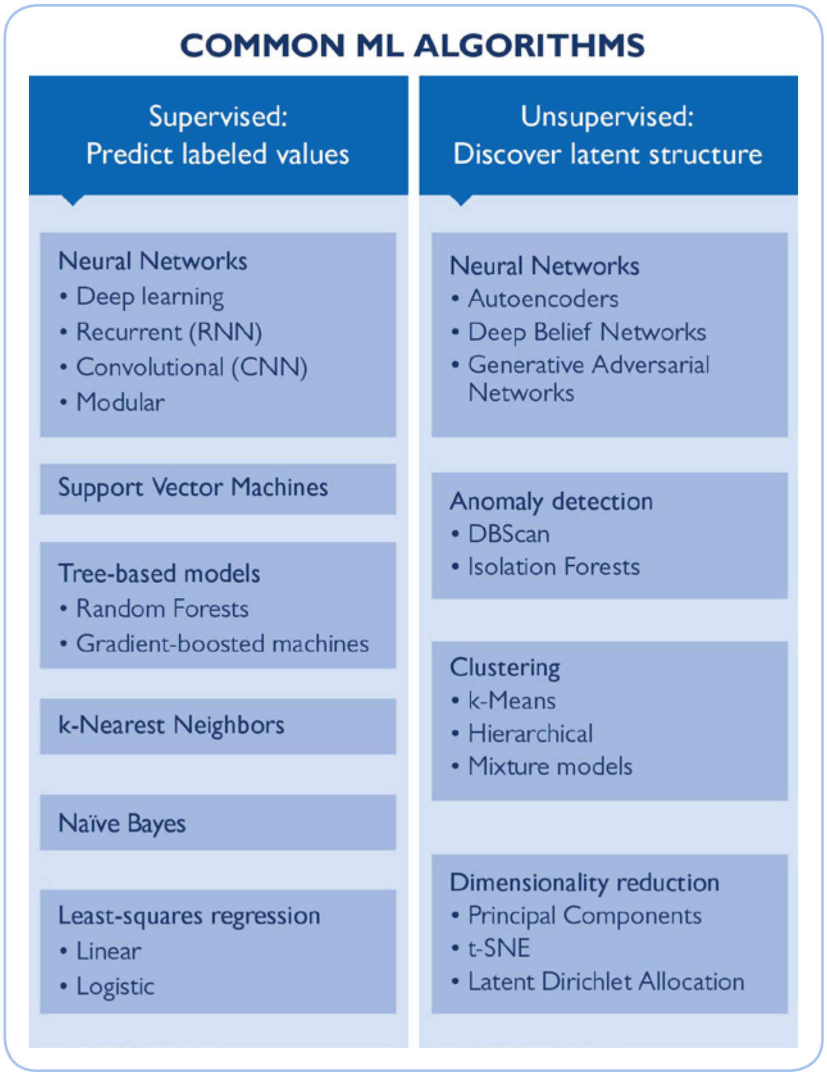


FIGURE 9: Simple illustration of decision trees that might be used in a credit scoring algorithm. The tree on the left is deep and detailed, containing decision rules that might not generalize well. In the tree on the right, regularization was used to limit tree depth so that rules remain simple and broadly applicable.

The final stage in model development is **model training and selection.** This involves choosing an algorithm that is well-suited to your modeling task and optimizing its hyperparameters. Some common ML algorithms are listed in FIGURE 10 for reference. Some algorithms work better with narrow datasets (i.e., fewer features) while others have an advantage with wider datasets (i.e., more features). Some are designed to work only with numerical data, while others can handle categorical variables (e.g., multiple-choice questions). Algorithm choices may also be constrained by dataset size or computational cost.

A–Z

Feature Engineering: Creating new predictors based on existing ones, in an effort to improve model performance.



COMMON ML ALGORITHMS

Supervised: Predict labeled values

- Neural Networks
  - Deep learning
  - Recurrent (RNN)
  - Convolutional (CNN)
  - Modular
- Support Vector Machines
- Tree-based models
  - Random Forests
  - Gradient-boosted machines
- k-Nearest Neighbors
- Naïve Bayes
- Least-squares regression
  - Linear
  - Logistic

Unsupervised: Discover latent structure

- Neural Networks
  - Autoencoders
  - Deep Belief Networks
  - Generative Adversarial Networks
- Anomaly detection
  - DBScan
  - Isolation Forests
- Clustering
  - k-Means
  - Hierarchical
  - Mixture models
- Dimensionality reduction
  - Principal Components
  - t-SNE
  - Latent Dirichlet Allocation

FIGURE 10: Overview of some popular ML algorithms. Detailed discussion of these algorithms is beyond the scope of this report, but most algorithms currently in use are similar to one of these.

Not all algorithms are equally interpretable. For example, linear regression provides easily-interpretable parameters for which rigorous confidence intervals can be derived. *For example, a regression coefficient could estimate how the probability of loan repayment is affected by a borrower's family size.* Simple decision trees return a set of rules that can be used for future classifications. *For example, a decision tree might predict that repayment rates are high for borrowers over 35 years old with more than 6 years of schooling.* For more complex algorithms, it's possible to estimate variable importance or test hypothetical scenarios, but this will likely be unable to "explain" an individual prediction *(See BOX 6: Opacity and explainability). In the credit example, variable importance calculations might show that income has a strong effect on repayment, but leave the details of that effect unexplained.*

Model training is part of an iterative process. The evaluation scheme chosen in the task-definition stage is used to tune hyperparameters and choose between algorithms. In a process known as *feature engineering,* modelers will often revisit data cleaning and preparation to optimize model performance. *For example, if a credit-scoring model uses the loan-issuance date as an input but is more accurate when loans are made after the harvest, deriving a "season" variable, or feature, may improve accuracy.* If a desired level of performance proves to be unattainable, it may be necessary to reconsider the original task definition and question whether the choices of target variable and training data were appropriate.

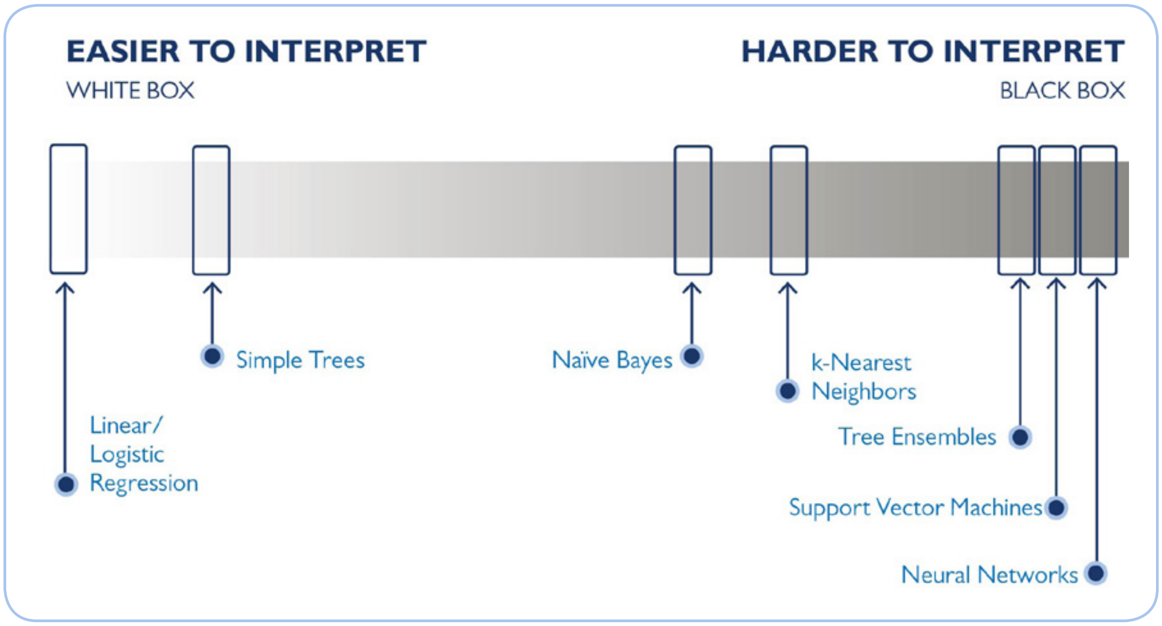


FIGURE 11: [Duplicate of Figure 6] Rough categorization of some ML model types by their ease of interpretation. Some are highly-interpretable (so-called "white box" models) while, others are inscrutable "black boxes."

INTEGRATING ML INTO PRACTICE

Along with the nuts-and-bolts work of model building, people also have to figure out how the model might inform or change existing operational processes. Whether the results of a ML model suggest hypotheses, triage data for human attention, or make predictions to inform operational decisions, decisions about how to interpret and use model results are always made by people.

UNDERSTANDING THE STATUS QUO

Integration begins with understanding the existing decision making process, both from a technical and social lens. The predictions made by a model need to be well-matched to the needs of the decision system already in place. This includes assessing the format of model predictions, as well as their precision and timeliness, but also assessing the specific needs and capacities of the people who will be

A–Z

In-sample Error: The rate at which a model makes mistakes on the data that were used in training.

Out-of-sample Error: The rate at which a model makes mistakes on data that were not used in training. Techniques such as cross-validation give estimates of the out-of-sample error.

expected to act upon model outputs. *For example, a credit-scoring algorithm would need to present results in a way loan officers can interpret easily, such as a score that is similar to existing credit scoring systems that are familiar to the officers.* The introduction of new technologies into decision systems can lead to long-term changes in the way decisions are made, but they will be taken up most effectively if they are congruent with current processes.

ESTIMATING THE COST OF ERRORS

Errors always "cost" something. These costs may be financial, environmental, or social, but understanding them is key to knowing how much to demand from a model. Often, different types of errors will have different costs. For example, in a facial-recognition system that controls access to a secure facility, false positives (letting an intruder in) are much more costly than false negatives (making an employee try a second time). *An algorithmic mistake that denies a microloan to someone living in poverty may prove more "costly" than a mistake that denies a loan to a wealthy retiree.* Estimates of error costs will inform the choice of evaluation metric during model development, as well as the appropriate role of the resulting model in a decision process. The costs and consequences of errors are independent of any particular model, and can be assessed even before a model has been developed.

ASSESSING CONFIDENCE IN MODEL

The next step in ML integration is to understand the strengths and weaknesses of the tool itself. The process of model selection and hyperparameter tuning should have generated several different estimates of error. The most important is an estimate of *out-of-sample error* — how the model will perform on data it has never seen before. *For a credit-scoring algorithm, the in-sample error would measure the agreement between actual and predicted repayment in the training data. The out-of-sample error would measure whether predictions were correct for new customers, outside the training set.* Depending on the application, it may be important to measure the relative accuracy of the model across different sub-populations. These measures of accuracy can be compared against the needs of the existing decision process — models with an accuracy rate that's too low to provide useful advice are often abandoned.

If a model allows estimates of variable importance or sensitivity, these should be weighed against the reliability of different sources of data. Models that rely heavily on low-confidence inputs may not provide useful decision support. *For example, smallholder farmers may have only a rough estimate of their annual income, and a model that relies heavily on this number could be unreliable.* Understanding the reliability of data sources is essential in a development context. We often do the best we can with the limited data that are available. Investing in better data collection and Data Quality Assessments[127] can improve confidence in development data, which can in turn improve the level of confidence that can be placed in resulting data-driven models.

The model training and development process can also provide information on other measures of performance, such as execution speed and *computational cost*. These can be compared against the number of new predictions that will need to be made, the expected timeliness of those predictions, and the financial resources available. Optimizing for fast decisions may require either greater investments in computing capacity or relying on simpler models in order to provide decisions in time.

A–Z

Computational Cost: The computational resources (including CPU time and memory) needed to perform a task. Ultimately, this translates into financial costs for either hardware or cloud services.

ESTABLISHING PROXIMITY TO DECISION

ML-enabled decision systems are often described as being automated. In general, automation can increase as tools becomes more trusted. Most automatable tasks, however, exist in the context of a larger process or workflow. A trusted ML model may be used to automate one specific task, while human discretion remains key to subsequent steps. For example, a program to track forest fires might fully automate the detection of burned areas in aerial imagery. In this example, those making decisions would never need to look at images of burned forest; instead they would see a summary map of burned and unburned areas. The ultimate decisions made — about where to evacuate residents or deploy firefighting, for example — would be based on the summary map along with their broader contextual understanding.

Other applications of ML are aimed at discovering new relationships, *such as understanding the social and economic factors that correlate with loan repayment.* Here, decision-makers might trust an ML model to identify new and important features, yet they would rely on people to use those insights in intervention design. In such cases, the ML tool is identifying population-level trends rather than making individual-level predictions. Even if such an analysis is highly trusted, it would not be used to automate decisions.

In contrast, when ML applications make predictions about individual cases, automation of a single task can lead to automation of the process as a whole. This happens when predictions are more proximal to decision-making. Different situations can be arranged along a spectrum, as shown in FIGURE 12. At one end, people retain much more discretion, and rely on ML/AI tools as one source among many consulted in the process of making a decision. At the opposite end, people may be removed from the system entirely, as ML/AI tools perform tasks with minimal scrutiny.

In the least-automated scenario, data-derived insights are used to give general advice about population-level trends, but not to make individual-level predictions. This, for example, is the traditional role of statistics in informing macro-level policies. At the next step, a ML algorithm might give suggestions tailored to specific cases, and decision-makers would need to weigh this against other sources of information. Automation increases as specific algorithmic recommendations become more trusted, so that people who make decisions rely on them more than other sources

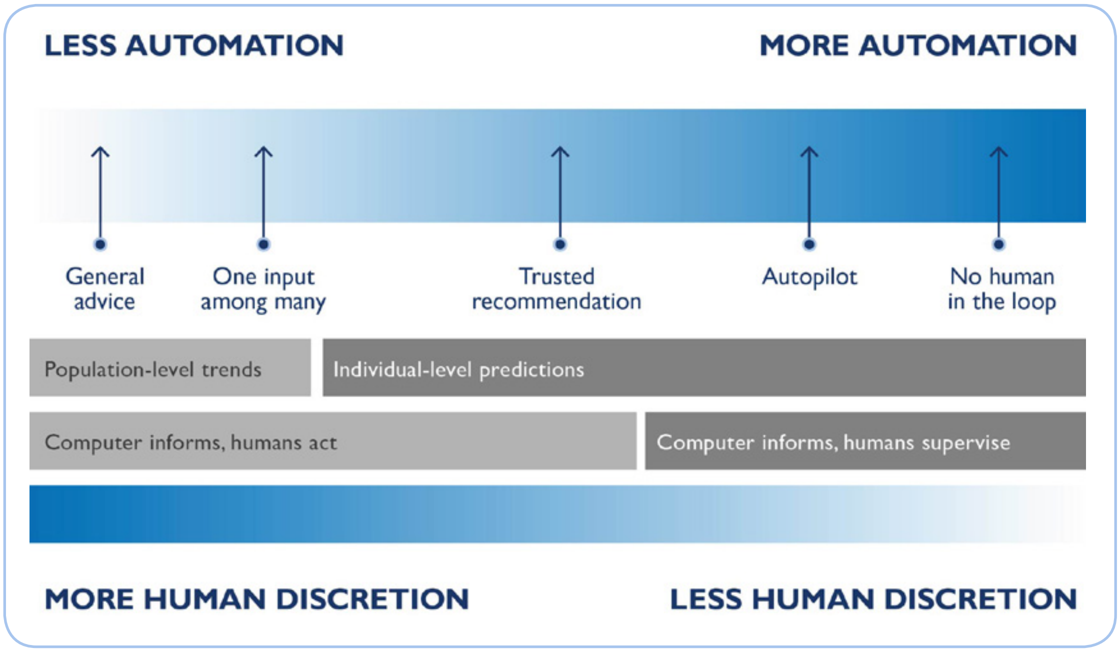


FIGURE 12: [Duplicate of Figure 7] Stages of automation in AI tools. These range from models that offer general advice based on population-level findings, through those that make individual-level predictions, to full automation with no human input. This high level of automation is likely to remain very rare in development applications.

of input. For example, the navigation app used by Uber drivers provides driving directions that are typically followed, although drivers are free to take another route. *Similarly, if loan officers receive an algorithmic credit score but not much other information about an applicant, they may have little choice but to follow the recommendation.* People exercise even less autonomy in autopilot systems, which will take action on their own unless the operator intervenes. Finally, a fully-automated system takes people entirely out of the decision process. This is most feasible in settings (such as online advertising) where decisions must be made quickly and the cost of mistakes is relatively low. *The equivalent in credit scoring would be a system that automatically disburses microloans via mobile money, without sign-off by a human loan officer.* The appropriate level of automation in higher-stakes systems is an active topic of legal and policy debate,[128] but is largely beyond the scope of this report.

Choosing the right degree of automation will be important for AI applications in development. As we seek to integrate algorithmic tools into decision processes, we need to consider the reliability and specificity of predictions, as well as the risk that might be introduced by automation — especially with highly vulnerable groups or in high-stakes contexts. Although automation often promises to save time or money, extensive automation may not always be suitable for development problems.

GETTING FEEDBACK

Once an ML model has been deployed as part of a decision system, its maintainers will need to collect information on how well it is working. One source of information is *online testing*, in which new ground-truth information is collected for comparison to model predictions. *In the case of credit scoring, this could involve a comparison of predicted loan repayment to actual repayment rates.* Many AI-enabled services collect usage statistics, so that model owners can understand how their tool is being used. *For example, one might track the actual decisions of loan officers, to see when they choose not to follow the algorithm's recommendations.* It is also possible to elicit qualitative feedback or complaints from users or others impacted by the model.

A–Z

Online Testing: Testing of a model's accuracy using predictions made after deployment. The word "online" refers to the decision tool being actively used, as opposed to sequestered for development and testing. "Online" a in this context does not refer specifically to the internet.

RE-EVALUATING, REVISING, AND UPDATING

Feedback about model performance, usage, and impact is only useful when it is acted upon. Performance loss may require retraining the model to incorporate newly-available data. If a model is being misused or under-used, its user interface may need adjustment. If more fundamental problems arise, a re-examination of the model's goals and underlying assumptions may be needed. In any case, maintaining a ML-enabled system requires much more than keeping servers switched on. It means continually adapting the entire decision process to constantly-changing needs and context.

About the artwork

The artwork in this document was created using two different methods. The images on pages 68 and 70 use a version of Google's "deep dream" methodology[129]. In this method, an image is enhanced by "inverting" an object-recognition algorithm. For example, an algorithm trained to detect pictures of cats can be used to modify an image, making non-cat objects (such as clouds, trees, or people) look increasingly "cat-like."

The remaining artwork uses a "style transfer" algorithm[130] to render each photograph in the style of a different image (often a painting). In this approach, a convolutional neural network is used to create a multi-scale representation of a painting's texture that is independent of its content. This "style representation" can then be transferred to a photograph, giving it the texture of the original painting.

Cover: USAID's Responsible Engaged and Loving (REAL) Fathers Initiative aims to build positive partnerships and parenting practices among young fathers. **Credit:** Save the Children

Page 8: In October 2016, USAID launched the READ Community Outreach activity at Soyama Primary School. The activity will reach students in nearly 2,500 schools throughout Ethiopia. **Credit:** Robert Sauers for USAID

Page 9: Ambassador Osius and Assistant Secretary Garber visit a climate-smart rice field supported by USAID's Vietnam Forests and Deltas project. **Credit:** Leslie Detwiler for USAID

Page 12: A young girl takes a drink of water from a newly constructed water tank. In 2017, USAID provided clean drinking water for more than 300,000 people in Ethiopia. **Credit:** AECOM

Page 22: Lokta bark paper rhododendron flowers being made for Aveda at Himalayan Bio Trade Pvt. Ltd. Kathmandu, Nepal. **Credit:** Jason Houston for USAID

Page 24: A study in Nepal found that birth attendant and maternal handwashing were associated with a 41 percent reduction in newborn mortality. Handwashing with soap also reduces infections in mothers and children during pregnancy and childbirth. **Credit:** Save the Children

Page 66: Breastfeeding is an important component in USAID's maternal and child health and nutrition efforts. **Credit:** Amy Fowler for USAID

Page 67: Kassa Mulualem is one of the first women in her area to take up plowing, an activity that is traditionally reserved for men. She is helping to raise awareness about gender equality and encouraging others to change their understanding of the division of labor between men and women. **Credit:** Kelley Lynch

Page 68: In October 2016, USAID launched the READ Community Outreach activity at Soyama Primary School. The activity will reach students in nearly 2,500 schools throughout Ethiopia **Credit:** Robert Sauers for USAID

Page 70: Participants in a Ugandan cash-for-work program build a road to link their community with the nearest market as part of the 2009 Horn Food Price Crisis Response (HFPCR). **Credit:** Kaarli Sundsmo for USAID

Page 73: Urban search and rescue teams working with USAID's Office of U.S. Foreign Disaster Assistance help search for survivors after a March 2011 magnitude 9.0 earthquake and subsequent tsunami in Japan. **Credit:** U.S. Air Force

Page 76: Women dancing in a competition put on by USAID's Northern Uganda Transition Initiative, which encourages northerners to celebrate their culture, return home, and take pride in their communities after a 23-year civil war. **Credit:** Nichole Graber for USAID

Endnotes

[1] Clifford, Catherine (2018) “Google CEO: A.I. is more important than fire or electricity.” CNBC.com

[2] AAAI Spring Symposium Report (2010)” Artificial Intelligence for Development.”

[3] Smith & Neupane (2018) “Artificial intelligence and human development: toward a research agenda.” IDRC Digital Library

[4] Accenture (2016). “Why Artificial Intelligence is the Future of Growth.”

[5] Angwin et al. (2016). “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.”
Lum & Isaac (2016). “To predict and serve?” Significance 13: 14–19

[6] Eubanks, Virginia (2018). “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.” St. Martin’s Press

[7] Zhang, Maggie (2015). “Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software.” Forbes.com
Lum, Jessica (2010). “‘Racist’ Camera Phenomenon Explained — Almost.” PetaPixel Blog

[8] Lomas, Natasha (2018). “UK report urges action to combat AI bias.” TechCrunch

[9] AI Now Report (2016). “The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term.”

[10] Russell & Norvig (2009). “Artificial Intelligence: A Modern Approach.”

[11] Sutton & Barto (2018). “Reinforcement Learning: An Introduction.” MIT Press

[12] http://aidr.qcri.org/

[13] “Intelligent Trap to Enhance Zika Surveillance.” Combating Zika and Future Threats: A Grand Challenge for Development

[14] Mahadevan, Karthik (2017). “Excelscope - Design for Primary Healthcare in Uganda.” YouTube

[15] http://terra-i.org/terra-i/about.html

[16] Deng et al. (2009). “ImageNet: A Large-Scale Hierarchical Image Database.” 2009 Conference on Computer Vision and Pattern Recognition

[17] McCandless, Michael (2011) “Language detection with Google’s Compact Language Detector.” Changing Bits Blog

[18] Araba, Debisi (2017). “The Digital Revolution and African Agriculture: why we need to seize this moment.” CIAT Blog

[19] CIAT (2017). “Good NEWS for the Fight Against Malnutrition: Using Big Data and Machine Learning to Power a Nutrition Early Warning System (NEWS) for Africa.”

[20] Perry, Chris, (2013). “Machine Learning and Conflict Prediction: A Use Case.” Stability: International Journal of Security and Development 2(3):56.

[21] Kaplan, Melanie (2017). “Precision Agriculture Predicts Civil Unrest.” Trajectory 2017(1)

[22] https://grillo.io/

[23] Harvard Humanitarian Initiative (2015). “Satellite Imagery Interpretation Guide: Intentional Burning of Tukuls.”

[24] World Resources Institute (2015). “RELEASE: Orbital Insight and World Resources Institute Partner on Satellite Imagery to Curb

Deforestation." Also: http://www.terra-i.org/terra-i/about.html

[25] Di Minin et al. (2018) "Machine learning for tracking illegal wildlife trade on social media." Nature Ecology & Evolution 2:406–407

[26] Imran et al. (2014) "AIDR: Artificial Intelligence for Disaster Response." International Conference on World Wide Web (WWW), Seoul, Korea.

[27] Eilander et al. (2016) "Harvesting Social Media for Generation of Near Real-time Flood Maps." Proc. Eng. 176–183

[28] Paul et al. (2016) "Social media mining for public health monitoring and surveillance." Pacific Symposium on Biocomputing

[29] Wang et al. (2012) "Data integration from open internet sources to combat sex trafficking of minors." 13th Annual International Conference on Digital Government Research

[30] Cohen et al. (2013) "Detecting Linguistic Markers for Radical Violence in Social Media." Terrorism and Political Violence 26(1):246–256

[31] World Bank (2018). "Can State-of-the-Art Machine Learning Tools Give New Life to Household Survey Data?"

[32] Jean et al. (2016). "Combining satellite imagery and machine learning to predict poverty." Science 353 (6301):790–794

[33] Blumenstock et al. (2015). "Predicting poverty and wealth from mobile phone metadata." Science 350(6264):1073–1076

[34] Wronkiewicz M. (2018). "Mapping the electric grid." Development Seed Blog

[35] Wang et al. (2015). "Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine." Int. J. Remote Sensing 36 (12):3144–3169

[36] Lu et al. (2016). "Unveiling Hidden Migration and Mobility Patterns in Climate Stressed Regions: A Longitudinal Study of Six Million Anonymous Mobile Phone Users in Bangladesh." Global Environmental Change 38:1–7

[37] Wilson et al. (2016). "Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake." PLoS Currents Disasters

[38] McDonald, Sean (2016). "Ebola: A Big Data Disaster."

[39] Pollak et al. (2017). "Computer Vision Malaria Diagnostic Systems — Progress and Prospects." Frontiers in Public Health

[40] Holmstrom et al. (2017) " Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and *Schistosoma haematobium*." Global Health Action 1337325

[41] https://plantix.net/

[42] Sgaier et al. (2017). "A case study for a psychographic-behavioral segmentation approach for targeted demand generation in voluntary medical male circumcision." eLife 6:e25923.

[43] https://destacame.cl/

[44] Grasser, Matt (2017). "The Fourth Industrial Revolution: How Big Data and Machine Learning Can Boost Inclusive Fintech." Next Billion Blog. January 30, 2017.

[45] https://www.r2accelerator.org/bsp/

[46] https://www.textteller.com/

[47] Wallace, Matt (2017). “Revolutionizing Financial Inclusion—with Chat.” Medium http://www.onowmyanmar.org/mr-finance-bot/

[48] Solon, Olivia (2016). “Karim the AI delivers psychological support to Syrian refugees.” The Guardian.

[49] Bayana, Neha (2017). “When your shrink is a bot.” The Times of India. November 2, 2017.

[50] Jiménez et al. (2016) “From Observation to Information: Data-Driven Understanding of on Farm Yield Variation.” PLoS ONE

[51] Brynjolffson & Mitchell (2017). “What can machine learning do? Workforce implications.” Science 358(6370):1530–1534

[52] https://www.darpa.mil/program/explainable-artificial-intelligence

[53] Ng, Andrew (2016). “What artificial intelligence can and can’t do right now.” Harvard Business Rev.

[54] Marcus, Gary (2018). “Deep learning: A critical appraisal.”

[55] De Luca et al. (2015). “I Feel Like I’m Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones.” Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 1411–1414

[56] Knight, Will (2017). “Paying with Your Face: Face-detecting systems in China now authorize payments, provide access to facilities, and track down criminals. Will other countries follow?” MIT Technology Review.

[57] Buolamwini & Gebru (2018). “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of Machine Learning Research 81:77–91

[58] Crawford, Kate (2016). “Artificial Intelligence’s White Guy Problem.” New York Times, June 25, 2016.

[59] https://www.perpetuallineup.org/

[60] Lum & Isaac (2016). “To predict and serve?” Significance 13: 14–19

Ensign et al. (2018). “Runaway feedback loops in predictive policing.” Proceedings of Machine Learning Research 81:1–12

[61] Eubanks, Virginia (2018). “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.” St. Martin’s Press

[62] Chouldechova et al. (2018). “A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions.” Proceedings of Machine Learning Research 81:134–148

[63] Privacy International (2018). “Further questions on Cambridge Analytica’s involvement in the 2017 Kenyan Elections and Privacy International’s investigations.”

[64] Woolley & Howard (2017). “Computational Propaganda Worldwide: Executive Summary.” Project on Computational Propaganda Working Paper 2017.11

[65] Chessen, Matt (2017). “The MADCOM Future.” Atlantic Council

[66] Brundage et al. (2018). “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.”

[67] FAT/ML “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.”

[68] Lindenbaum, David (2017). “2nd SpaceNet Competition Winners Code Release.” Medium

[69] Lazer and Kennedy (2015). “What We Can Learn From the Epic Failure of Google Flu Trends.” Wired 10/01/15

[70] Coren, Michael (2016). “People dump AI advisors that give bad advice, while they forgive humans for doing the same.” Quartz

Prahl & Van Swol (2017). “Understanding algorithm aversion: When is advice from automation discounted?” Journal of Forecasting 36(6):691–702.

[71] Angwin et al. (2016). “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.”

[72] Elish, M.C. (2016). “Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.” We Robot 2016 Working Paper

[73] Burrell, Jenna (2016). “How the machine ‘thinks’: Understanding opacity in machine learning algorithms.” Big Data & Society

[74] Rajkomar et al. (2018) “Scalable and accurate deep learning with electronic health records.” npj Digital Medicine 1:18

[75] Gaba, Kwawu Mensa (2016). “Eyes in the sky help track rural electrification.” World Bank Sustainable Energy for All Blog

[76] Wesolowski et al (2017). “Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics.” Nature Communications 8, 2069

[77] https://dhsprogram.com/

[78] http://surveys.worldbank.org/lsms

[79] Blondel et al. (2013) “Data for Development: the D4D Challenge on Mobile Phone Data.”

De Montjoye et al. (2014) “D4D-Senegal: The Second Mobile Phone Data for Development Challenge.”

[80] Orange (2014) “Data for Development Challenge Senegal Book of Abstracts: Scientific Papers.”

[81] GSMA. (2018). “Connected Women: The Mobile Gender Gap Report.“

[82] https://www.opalproject.org/

[83] https://www.planet.com/

[84] Kim, Annette (2018). “Satellite Images Can Harm the Poorest Citizens.” The Atlantic

[85] Imran et al. (2014) "AIDR: Artificial Intelligence for Disaster Response." International Conference on World Wide Web (WWW), Seoul, Korea.

[86] U.N. Global Pulse (2016). “Informing governance with social media mining.”

[87] U.N. Global Pulse (2016). “Making Ugandan Community Radio Machine-Readable Using Speech Recognition Technology.”

[88] Oxford Internet Institute (2016). “The Computational Propaganda Project.”

[89] EMC (2014). “Vertical Industry Brief: Digital Universe Driving Data Growth in Healthcare.”

[90] Rajkomar et al. (2018) “Scalable and accurate deep learning with electronic health records.” npj Digital Medicine 1:18

[91] USAID (2015). “ADS Chapter 579: USAID Development Data.”

[92] Kim, Annette (2018). “Satellite Images Can Harm the Poorest Citizens.” The Atlantic

[93] Holder, Sarah, “Who Maps the World?” CityLab, March 14, 2018.

[94] Irani, Lily (2015). “Justice for ‘data janitors.’” Public Books Blog

[95] https://www.mturk.com/worker/help

[96] Imran et al. (2014) AIDR: Artificial Intelligence for Disaster Response. International Conference on World Wide Web (WWW), Seoul, Korea.

[97] https://pybossa.com/

[98] Yang & Newsam (2010). “Bag-of-visual-words and spatial extensions for land-use classification.” Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems 270–279. Data available at: http://weegee.vision.ucmerced.edu/datasets/landuse.html

[99] Socher et al. (2013). “Recursive deep models for semantic compositionality over a sentiment treebank.” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 1631–1642. Data available at: https://nlp.stanford.edu/sentiment/

[100] http://www.youthmappers.org/

[101] https://www.hotosm.org/

[102] Abbott, D. (2012). “Why defining the target variable in predictive analytics is critical.”

[103] https://www.vanderbilt.edu/lapop/

[104] Center for Effective Global Action. (2017). “Gen pproaches.” PLoS ONE

[106] Upturn and Omidyar Network (2018). “Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods.”

[107] http://www.fews.net/

[108] Marijnis, Martijn (2018). “Can machine learning help us better predict hunger?” ICCO Cooperation Blog

[109] Eubanks, Virginia (2018). “Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.” St. Martin’s Press

[110] Solon, Olivia (2017). “Facebook is hiring moderators. But is the job too gruesome to handle?” The Guardian, May 4 2017.

[111] Hopkins, Nick (2017). “Revealed: Facebook’s internal rulebook on sex, terrorism and violence.” The Guardian, May 21 2017.

[112] Christin, Angèle (2017). “Algorithms in practice: Comparing web journalism and criminal justice.” Big Data & Society 4(2):1–14

[113] McAlone, Nathan (2015). “Here’s how Uber got its start and grew to become the most valuable startup in the world.” Business Insider

[114] Eshel et al. (2017). “Evaluation of the Parasight Platform for Malaria Diagnosis.” Journal of Clinical Microbiology 55(3):768–775

[115] Solon, Olivia (2016). “Karim the AI delivers psychological support to Syrian refugees.” The Guardian

Bayana, Neha (2017). “When your shrink is a bot.” The Times of India November 2, 2017.

[116] Jimenez, D. (2018). “Everyone’s looking for a data unicorn. There’s no such thing.” Medium.

[117] Wallace & Castro (2018). “The Impact of the EU’s New Data Protection Regulation on AI.” Center for Data Innovation.

[118] USAID (2018). "Considerations for Using Data Responsibly at USAID."

[119] Arntz, M., T. Gregory and U. Zierahn (2016). “The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis”, OECD Social, Employment and Migration Working Papers, No. 189, OECD Publishing, Paris.

[120] Banga & te Velde (2018). “Digitalisation and the Future of Manufacturing in Africa.” Supporting Economic Transformation (ODI/ DFID Report).

[121] Winner, Langdon (1980). “Do Artifacts Have Politics?” Daedalus 109(1):121–136.

[122] Blumenstock et al. (2015). “Predicting poverty and wealth from mobile phone metadata.” Science 350 (6264):1073–1076

[123] Deng et al. (2009). “ImageNet: A Large-Scale Hierarchical Image Database.” 2009 Conference on Computer Vision and Pattern Recognition

[124] Zheng, Alice (2015). “Evaluating Machine Learning Models: A Beginner’s Guide to Key Concepts and Pitfalls.” O’Reilly Media.

[125] World Economic Forum (2018). “How to Prevent Discriminatory Outcomes in Machine Learning.”

Barocas & Selbst (2016). “Big Data’s Disparate Impact.” California Law Review 104:671–732

Friedler et al. (2016) “On the (im)possibility of fairness.”

[126] Abu-Mostafa et al. (2012). “Learning from data.” AMLBook

[127] USAID (2018). "Considerations for Using Data Responsibly at USAID."

[128] Elish, M.C. (2016). “Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.” We Robot 2016 Working Paper

[129] Mordvintsev et al. (2015). “Inceptionism: Going Deeper into Neural Networks.” Google AI Blog

[130] Gatys et al. (2015). “A Neural Algorithm of Artistic Style.” arXiv:1508.06576 [cs.CV]

UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT

USAID

USAID

FROM THE AMERICAN PEOPLE