



MANAGING MACHINE LEARNING PROJECTS IN INTERNATIONAL DEVELOPMENT

A PRACTICAL GUIDE



ACKNOWLEDGEMENTS

This report is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of DAI and do not necessarily reflect the views of USAID or the United States Government.

This guide was written under DAI's Digital Frontiers project (USAID Cooperative Agreement No.AID-OAA-A-17-00033) with support from the Digital Frontiers team, overseen by Stephanie Creed. The guide was authored by Vital Wave, a professional services firm that focuses on the use of digital for development. Special thanks to the team at Vital Wave, including Jacqui Watson, Liz Grubin, Oswald Jumira, and Brendan Smith for their dedication and enthusiasm for this work.

The report's development was managed by the Strategy and Research team in USAID's Innovation, Technology, and Research Hub. Many thanks to Amy Paul, Shachee Doshi, Craig Jolley, Andrew Merluzzi and Aubra Anthony for their technical inputs and feedback.

The authors also extend their appreciation to all those who shared key insights and experiences to help us improve this report, including Alex Beyers, Mandeep Birdi, Ben Brockman, Erica Bustinza, Allan Cheboi, Talia Dweck, Andrea Falso, Steve Hellen, Jussi Hinkkanen, Karthi Keyan, Wanjiku Kiragu, Anine Kriegler, Benji Meltzer, Lungi Okoko, Alpan Raval, Amar Shah, Kieran Sharpey-Schafer, Patrick Singa, Radina Soebiyanto, Amanda Strydom, Bhavin Vadera, and Samhir Vasdev.

TABLE OF CONTENTS

HOW YOU CAN HELP IMPLEMENT ML/AI RESPONSIBLY	4
COMMONLY USED TERMS	6
ML AND AI: THE BASICS	8
Defining and Differentiating ML and AI	9
The Different Types of ML Tasks	10
How ML Works	11
ML IN THE CONTEXT OF THE PROJECT LIFECYCLE	12
MODULE 1: EVALUATE FEASIBILITY	15
Matching Your Problem to ML/AI Capabilities	16
Identifying Potential Data Sources	17
Building a Project Team	20
Learnings from the Field	27
MODULE 2: MODEL DESIGN AND BUILD	29
Considerations When Building the ML/AI Model	30
Designing for Sustainability	31
Data Laws and Regulations	34
Preparing to Assess Model Performance	43
MODULE 3: IMPLEMENTATION	50
Evaluating Your Project's Application of ML/AI	51
Incorporating ML/AI into the Decision Processes	55
Learnings from the Field	59
MODULE 4: POST IMPLEMENTATION	61
Long Term Sustainability	62
Ensure Responsible, Shared Learning	62
END-TO-END CASE STUDIES	63
Youth Employment	64
Media Integrity	66
CONCLUSION	68
ADDITIONAL RESOURCES	69
LIST OF DECISION AIDS	
Assessing Feasibility of Using ML/AI	24
Identifying Risks and Incorporating Safeguards	36
Building Your Project's ML/AI Model	44

HOW YOU CAN HELP IMPLEMENT ML/AI RESPONSIBLY

The authors have developed this guide to support practical decision-making by project managers throughout the development sector. The guide's focus is on promoting the responsible use of ML/AI in development and humanitarian work, without assuming background expertise in data science or digital programming. Earlier this year, USAID published its Digital Strategy. The aim of this guide is to provide practical resources for the development community that engender an informed, thoughtful approach to using ML/AI and to support development practitioners in actively shaping effective and responsible digital programming.

Who Should Read this Guide

This practical guide has been designed for **development practitioners who may not be trained technologists but are involved with or responsible for implementing projects that might have a technical machine learning/artificial intelligence component.** It is informed by USAID's previously released [*Reflecting the Past, Shaping the Future: Making AI Work for International Development*](#). In some instances, this guide incorporates images and language from that original USAID report.

How to Use this Guide

This guide is less about development of a technical ML model and more about management of a project that includes ML. Following the phases of a project lifecycle, this guide provides **practical guidance and examples of implementations of artificial intelligence (AI) and machine learning (ML)** with the goal of strengthening your understanding of how these technologies can be appropriately applied, integrated, and managed for impact.

Along with the project lifecycle modules, four key thematic areas are woven throughout, providing a framework for enhancing positive, responsible impact and sustainability when using ML/AI:

THEMATIC AREA 1

Responsible, Equitable and Inclusive Design

This theme is the cornerstone of ensuring positive impact and is important to consider throughout the project lifecycle. It explores three principles:

- **Responsible:** How can you take appropriate measures to increase the likelihood of your project having the desired outcomes and proactively **identify and mitigate any harmful outcomes** to individuals or groups?
- **Equitable:** How can you **ensure your project isn't disproportionately benefiting or harming** individuals or groups more than others?
- **Inclusive:** How can you **include end-users and relevant stakeholders** in conversation, observation, or co-design to ensure their perspectives are represented, their needs are properly addressed, and that ML/AI technology is designed with their interests in mind?

THEMATIC AREA 2

Strategic Partnerships and Human Capital

Technology is given meaning through the people who build and use it. This theme focuses on identifying in-house skills and strategic partnerships, from local knowledge experts to technical consultants, that you may need for your implementation.

THEMATIC AREA 3

Adaptive Management for ML Projects

This theme focuses on mechanisms for allowing continuous learning and iterative development given the unique needs and challenges of implementing ML/AI. By building in mechanisms for adaptation, you will be better placed to achieve your project goals.

THEMATIC AREA 4

Enabling Environment for ML/AI

Lastly, this theme highlights enablers for sustainably implementing ML/AI. These enablers may be externalities such as local infrastructure, legislation, or availability of data. This thematic area focuses on helping you identify whether the right elements are in place for your project to succeed.

ML/AI is a relatively new field, and in many cases, we will be learning while actively shaping and developing the field as we go. In some cases, you may find that working to strengthen the enabling environment is a prerequisite or concurrent objective of your project that will facilitate responsible use of your model in the long-term.



Wondering if ML/AI is the Right Tool for Your Project?

Development practitioners like yourself may be wondering if and how you might use ML/AI applications in your projects. This guide will help you answer these questions as well as highlight whether it's an appropriate and relevant tool in your context and how you can ensure its effective and fair use.

As ML/AI applications become more widely adopted, so does optimism for their ability to reshape the development sector; from automating labor-intensive tasks to offering new insights from complex datasets. However, with this growing opportunity comes the need to carefully assess when and if ML/AI is a good fit for any given project.

Each problem and context is unique, meaning only your project team can answer whether ML/AI is the right tool for your project(s). Keep an eye out for this icon  to understand the critical decisions and questions to consider; and this icon  for decision aids to assist you along the way.

COMMONLY USED TERMS

Accuracy: In general, accuracy refers to a measure of how well the results of the model compare to an observed value (e.g. how well forecasted-cases of malaria compare to those that actually occur) or to an external gold standard (how well a computer vision diagnostic performs relative to a human expert). There are multiple measures of accuracy that might be used, such as the fraction of correct classifications or the magnitude of difference between the model-predicted value and the observed value. Accuracy doesn't distinguish between false positives and false negatives, so two models could have the same overall accuracy but make different types of errors.

Advanced Analytics: The use of one or more machine learning algorithms to carry out sophisticated analysis of multiple data sources and structures.

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

Artificial intelligence (AI): See page 9 for the definition.

Bias: Systematically favoring one group relative to another. For the purposes of this document, bias is defined in terms of specific categories or attributes (e.g., gender, race, education level).

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."

Classification: Assignment of data points to one of two or more qualitatively different categories. Classification algorithms are used when the outputs are restricted to a limited set of values (e.g. low/medium/high for risk assessment or classifying images into a predetermined set of categories).

Computer Vision (CV): Analogous to natural language processing (NLP), computer vision algorithms process image data, such as satellite imagery, photographs, and videos. Common CV tasks include identifying and locating objects, interpreting writing in photographs, and facial recognition.

Data Augmentation: A technique used to increase the size of data used for training a model. This is often done through random transformations of existing data, such as shuffling words in text to make new sentences or rotating or cropping images to make new images. If there is insufficient data available, this strategy increases the diversity of data without actually collecting new data.

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

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

Data Scraping: The process of automating the extraction of data from a format meant for human consumption to a format that can be easily used by computers (referred to as machine-readable data). Web scraping automates the extraction of data (often text) from websites, while PDF scraping can "unlock" data from written reports.

Decision Systems: The means by which people plan or choose between options. Most decision systems use technology in some form. While simpler technologies 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.

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

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

Feature Selection: Feature selection, also called variable selection or attribute selection, is the selection of attributes in data (such as columns in a spreadsheet) that are most relevant to the predictive modeling output. Feature selection methods reduce the input dimensions (potentially avoiding model over-fitting), which further aid in creating a robust predictive model.

Human in the Loop: Leverages both human and machine intelligence to create machine learning models. Humans participate in the ML process by training, tuning, and testing a particular algorithm as well as controlling how results of ML models are used in organizational decision-making.

Interoperability: The ability of two or more information systems or components to exchange information based on specific standards, where each system is able to use the information that has been exchanged.

Machine Learning: See page 9 for the definition.

Model: A simplified depiction of reality. ML models are built by combining a training data set with an algorithm. Data scientists routinely try out many models and adjust model parameters in an effort to develop a model that works best for the desired task. Once a model is built, it will periodically need to be retrained on new data as performance will often degrade over time.

Model Evaluation: Quantitative assessment of a model's performance, according to predefined criteria.

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

Natural Language Processing (NLP): A form of language recognition and comprehension in which computers use "natural" language text as inputs or outputs (e.g., English, French, Arabic). NLP applications include machine translation, text summarization, and the "autocomplete" feature in many popular email clients.

Output 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 a target variable or dependent variable.

Out-of-Sample Accuracy: A model validation technique for assessing whether the results of a predictive ML model will perform against a new, independent data set.

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

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

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

Regression: Quantifies the relationship between one or more predictor variables and one output variable.

Supervised ML: Algorithms that require training data to be labeled with values of the output variable. Supervised algorithms need to know the "right" answer to develop prediction rules, in contrast to unsupervised ML, where output variables are not labeled in the training data.

Training Data: Machine learning algorithms build a mathematical model based on sample data, known as "training data", to make predictions without being explicitly programmed to perform the task. A learning algorithm will find patterns and relationships in training data and use them to define rules for new predictions.

Test Data: Data used to build machine learning models are usually split into a "training set" (see Training Data definition) and "test set," where the training set is used to develop the model, and the test set is withheld from the initial training process so that a model can be "tested" on it later to in order to evaluate how the model performs with new data.

Unsupervised ML: Algorithms that do not require pre-labeling of the output variable. Rather than predicting the "right" answer, unsupervised ML finds latent patterns in data. Examples of unsupervised ML include clustering and outlier detection.

ML AND AI: THE BASICS

Objective of this Module

This module introduces concepts and terminology to orient you for the rest of the guide. It defines and differentiates ML and AI, provides an overview of how they work, and gives practical examples of how they can be applied to different projects and contexts.

Topics Covered in this Module

- [Defining and Differentiating ML and AI](#)
- [The Different Types of ML Tasks](#)
- [How ML Works](#)

Defining and Differentiating ML and AI

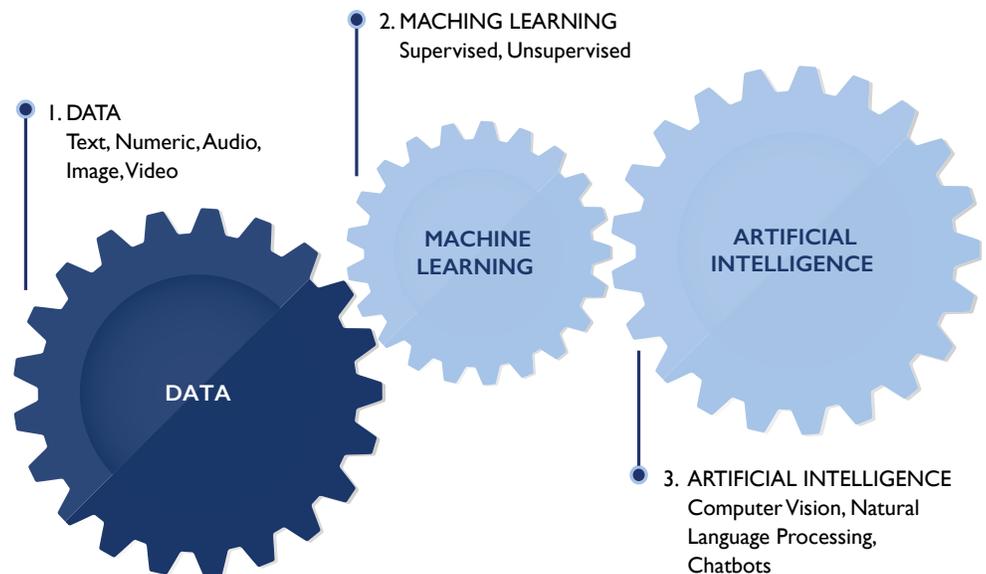
Machine learning (ML) and artificial intelligence (AI) are often used interchangeably, but they are slightly different.

- **Machine learning (ML)** is a set of methods for getting computers to recognize patterns in data and use these patterns to make predictions. It focuses on the development of computer programs (**models**) that can access and learn from data. You can think of ML as “data-driven predictions.” **ML is a subset of AI and enables data-driven predictions to inform decisions.**
- **Artificial intelligence (AI)** uses computers for making decisions or recommendations in an automated way. Automated decisions might be directly implemented or suggested to a human decision maker. You can think of AI as “smart automation.”



For a more in-depth technical overview, you can refer to page 80 of USAID’s report [Reflecting the Past, Shaping the Future: Making AI Work for International Development](#) or MIT’s guide [Exploring Fairness in Machine Learning for International Development](#).

Figure 1: The relationship between data, ML, 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. ML is a subset of AI that uses models to make data-driven predictions. AI applications can rely on a ML model to translate data into usable predictions to make, plan, or do something in the real world.



What Makes ML/AI Projects Different from Other Digital Projects?

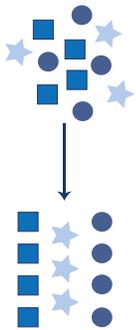
As with most tech projects, implementing an ML model is an iterative process. As new data becomes available or the context changes, the model needs to be tweaked and updated. Like many statistical approaches, ML models can easily produce errors unless very deliberate efforts are made to monitor performance, check for biases, and retrain models over time. Errors are a statistical reality and can result from embedded bias, poor quality data, or false associations between the data and reality. Implementing an ML model appropriately is largely about understanding the types of errors it will produce and the associated tradeoffs.

This guide highlights some of the anticipated errors and risks related to implementing ML, with the intent of helping you understand the decisions and tradeoffs you and your project team will need to make.

The Different Types of ML Tasks

Machine learning (ML) allows computers to generalize from existing data and make **predictions** for new data. They find patterns in **training data** and return a model to **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.

ML applications can be roughly organized into four “tasks”. These tasks can be used discretely or in combination (more on this in the section titled *Matching Your Problem to ML/AI Capabilities* on page 16).



Sort

Sorting - or **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, ML could be used to sort those that contain huts from those that do not.

Sorting models often separate things into two categories: typical and atypical. Atypical instances are often anomalies, such as malaria-infected cells in a rapid diagnostic test. Anomaly detection can use either **supervised** or **unsupervised** ML. Supervised ML requires model-builders to specify the “right” and “wrong” answers (referred to as training data), which the model will then learn to imitate. In unsupervised ML, the model separates instances into clusters or finds instances that differ from the majority.



Score

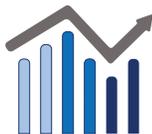
Scoring - or **regression** – uses patterns in the data to predict a quantity. For example, given a series of aerial images containing huts, ML could be used to predict the likely population density of the area.

Scores are often probabilities, such as whether a loan will be repaid or if 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. They rely on large volumes of data to learn nuanced prediction rules.



Discover

Discovery models allow us to draw novel insights that are not apparent in traditional statistical methods from large and seemingly disparate datasets. They 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.



Forecast

Forecasting models make predictions about complex events, often drawing from machine learning models as one of several inputs. Forecasting often predicts the likelihood of events in the future using historical patterns. Common examples of forecasting in development are with respect to climate events, famine, disease outbreaks, or civil unrest.

Please Keep in Mind

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

How Machine Learning Works

Whether sorting, scoring, discovering, or forecasting, ML models aim to estimate values of a target variable based on a set of predictors. The example in Figure 2 is based on an ML application developed by South African startup Raphta and used by hospitals in Kenya to assist in managing and preventing the spread of COVID-19. It used **computer vision** to measure use of personal protective equipment (PPE) requirements (e.g., use of masks), social distancing, and the size of gatherings in health facilities and public places, and then used this data to develop a risk “score.”

The risk score is an indication of the level of risk of exposure to COVID-19 infections in the facility and helps decision makers to determine whether they should shut down the facility.

Figure 2: Illustration of data terminology for a sample dataset in which the target variable is a risk score, while the predictors describe a hospital’s adherence to COVID-19 protocols and historical cases.

- 1 **FEATURES** (columns): Values that describe instances. Target variables and predictors are both types of features. Also referred to as variables.
- 2 **PREDICTORS**: Values used to generate a prediction. Also referred to as independent variables.
- 3 **TARGET VARIABLE**: The target variable, or dependent variable, is the output of the model prediction.
- 4 **INSTANCES** (rows): The individual people, places, things, or events described by a dataset.

1 FEATURES					
2 PREDICTORS					3 TARGET VARIABLE
Organisation	Average Social Distancing (metres)	Mask wearing adoption (%) <i>(% of people in the facility wearing masks in a month)</i>	Number of people gathering in an area within the hospital	Number of past Covid19 Infections at the institution	Risk Score <i>(1 being least risk of infection and 10 being high infection risk, resulting in complete shutdown of facility)</i>
Hospital 1	1.0	90	20	164	9
Hospital 2	1.4	100	9	42	8
Hospital 3	2.0	70	8	24	6
Hospital 4	1.5	80	5	0	2
Hospital 5	1.1	60	3	5	1
Hospital 6	2.1	95	6	20	3
Hospital 7	1.3	96	3	32	5

4 INSTANCES

The set of predictors should be diverse enough to capture different relevant aspects of what they describe. For example, a dataset that only contains the number of previous COVID cases is less diverse than one that also describes adherence to COVID-19 protocols (such as PPE requirements and social distancing guidelines).

Broadly speaking, ML systems seek to use a set of predictors (input variables) to estimate a target variable. When these variables are hard to measure (because they are difficult, expensive, or even dangerous to measure directly), proxies are used to estimate target variables.

ML IN THE CONTEXT OF THE PROJECT LIFECYCLE

Overview

The modules that follow aim to equip you with the knowledge and tools to understand how to assess, design, implement, and evaluate the use of ML in your project(s), as illustrated in the figure on the next page. Each module provides guidance through highlighting thematic areas, critical decisions and examples of questions to ask your team, data scientists, and stakeholders.

CRITICAL DECISION

As you move through the project lifecycle there will be **critical decisions** to make collaboratively with other stakeholders. It's crucial that you play an active role in shaping these decisions and that you discuss the decisions with your data scientist as they're advancing through the **data science workflow**.

These critical decisions cut across all the thematic areas, and you'll need to return to some of them as you learn new information or as the context changes. By continuously iterating, you'll have a greater likelihood of building a responsible, appropriate and sustainable solution.

Critical Decisions
These are tools you can use to help your decision making.

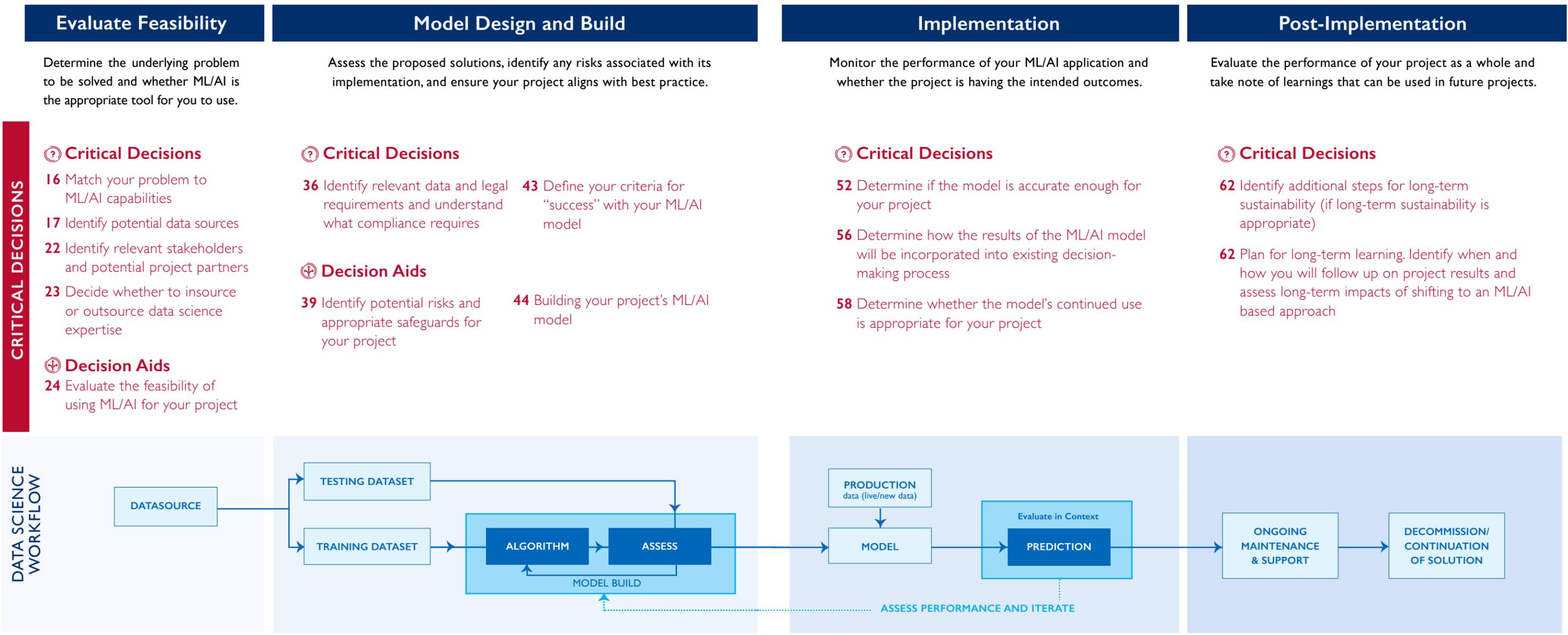
Decision Aids
These are tools you can use to help your decision making.

XX Guide Page Number

Data Science Workflow

This is the overall workflow your data scientist/model builder will be going through during the project lifecycle.

Figure 3: An overview of the ML workflow and the related critical decisions within a typical project lifecycle.



THROUGHOUT
Collaborate, collect feedback, and iterate the ML model and implementation

MODULE 1: EVALUATE FEASIBILITY

Objective of this Module

This module aims to help you identify whether ML/AI is an appropriate tool to use for your chosen problem. It discusses different data sources and their limitations as well as the composition of your team and strategic partners.

Making Decisions About

- [Matching Your Problem to ML/AI Capabilities](#)
- [Identifying Potential Data Sources](#)
- [Building a Project Team](#)
- [Assessing Feasibility of Using ML/AI](#)

At the end of this module, you'll conduct a feasibility assessment for your project based on the sections that follow.

Matching Your Problem To ML/AI Capabilities

While it may seem obvious, it's worth stating: it's good practice to ensure your problem is well matched for the use of ML/AI technology. Consult Table 1 and background materials as needed to get a sense of whether your problem aligns with the types of applications best suited for this technology.

Table 1: The uses and examples of different ML tasks

ML Task	Type of Application	Specific Examples (sector)	Commonly Used ML/AI Capabilities
Sort 	<p>Sorting images into different categories</p> <p>Interpreting text data to triage incoming information</p> <p>Processing language to deliver targeted information</p>	<ul style="list-style-type: none"> • Triageing satellite images of damage after disaster (<i>Humanitarian Assistance</i>) • Predicting presence or absence of disease/health issue based on image (<i>Health: Point of Service Diagnostics</i>) • Automating interpretation of text in social media posts to identify need requests after disaster (<i>Humanitarian Assistance</i>) • Automating answers to common questions that guide users through a series of decisions via conversational interfaces (<i>Diagnostics assessment for Health, Education, or Hiring</i>) 	Computer Vision (CV) , Natural Language Processing (NLP)
Score 	<p>Predicting likelihood of a specific, usually individual level outcome based on patterns in historical data</p>	<ul style="list-style-type: none"> • Predicting loan repayment for an individual (<i>Financial Inclusion</i>) • Predicting loss to follow up for patient (<i>Health</i>) • Predicting school drop out for student (<i>Education</i>) 	Advanced Analytics , NLP, CV
Forecast 	<p>Identifying likelihood of a system-level event based on historical trends</p>	<ul style="list-style-type: none"> • Predicting likelihood of disease outbreaks, famine, flood, civil unrest (e.g. from satellite imagery and/or additional data sources) (<i>Resilience</i>) 	Advanced Analytics, CV
Discover 	<p>Identifying new relationships between data points (or confirming suspected relationships)</p>	<ul style="list-style-type: none"> • Identifying which combination of crop management techniques and climate conditions are related to optimal crop yield in areas affected by climate change (<i>Agriculture</i>) • Identifying household characteristics that predict resilience to shocks in order to identify new points of intervention (<i>Humanitarian Relief</i>) 	Advanced Analytics

Identifying Potential Data Sources

Data for ML models can come from general data sources (e.g., national census data) and/or internal data collected by an organization as part of project implementation or business operations. Regardless of its source, the quantity and quality of data will impact whether a model will work for you or not. In this stage, review potential data sources to understand the strengths and weaknesses of each and to help set expectations for how much effort will be needed to have data ready to start using machine learning analyses. As you go through the project, you may revise or add to data sources you identify now.

Common Data Sources

- **Internal Data Sources:** Development practitioners may have access to existing operational data. This includes monitoring and evaluation (M&E) data, financial records, travel logs, or other operational data collected by the organization. An organization with years' worth of such data may turn to ML for new insights. Developing trusting relationships with local organizations can also open doors to more local, context-specific, accurate, and timely data. The data that your ML projects need may come from pre-existing partnerships that will be further strengthened by productive, responsible data sharing.
- **General Data Sources:** There may be publicly or privately held general datasets that can be **scraped** or repurposed for ML model building. These include household surveys, censuses, mobile phone metadata, satellite imagery, and social media posts.

However, these data sources can have various, common limitations. They might be outdated, focused on problems not relevant to the one you are trying to solve, or reflect **bias** in how they represent certain people, places, or things. They might also be difficult to access due to concerns over data privacy or intellectual property.

 A more in-depth discussion on these data sources is available on page 48 of USAID's [Reflecting the Past, Shaping the Future: Making AI Work for International Development](#).

Local Context and Data

The foundation of all ML tools is data. It is necessary to understand who or what is represented, overrepresented, underrepresented, or misrepresented by your data. Context is essential. 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.

KEY THEMATIC AREA

Adaptive Management

It is easy to underestimate the amount of time and effort that will be necessary to get data ready to start machine learning-based analyses. Take time to understand in detail how much data is available, how standardized it is, and how much time and effort it will take to get into the correct format before you can start.

You may want to build in an “exploratory” period where you work with a partner for only a few months to allow time for you and a data science partner to dig into available data before committing to a long-term project. Allowing adequate time for this process can help better manage project expectations, budget, and timelines going forward.

USAID offers additional guidance on data management in ADS 579 that may be applicable to the projects or activities of development professionals.

Context can also affect people's willingness to share data or provide accurate, reliable data. In countries where power is dominated by one ethnic or religious group, mistrust among the disenfranchised can run deep. People also 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 avoid participation and are thus excluded from datasets.

Common Challenges with Accessing Data

Data related to development challenges are often scarce and difficult to obtain, especially in a digital, machine-readable format conducive to ML. There is no definitive answer on how much data is considered enough. Some algorithms can produce sufficiently accurate predictions based on incomplete or limited data, while others may require millions of data points. You and your data science expert will likely need to experiment with different algorithms and augmentation techniques to determine whether the data you have has merit going forward with machine-learning projects. This is discussed in more detail in the [Model Design and Build](#) module.

Good Practices for Data Management

Development project and activity managers can work with implementing partners to help plan for high-quality data management. Some tools and best practices that organizations have found helpful for managing data include:

- Creating and maintaining an **inventory of datasets and documentation** that are required deliverables per award provisions and guidelines;
- Ensuring that data-related legal agreements and informed consent procedures **document data access and re-use rights**;
- Validating that you as the partner have the **capabilities to store and manage data responsibly** and to create rich documentation that describe data and analyses;
- Ensuring that you as the partner have the **capabilities to document and manage any privacy and security risks** associated with the data;
- **Drafting a Data Management Plan (DMP)** as a part of the Monitoring, Evaluation, and Learning (MEL) Plan that includes the inventory of datasets and describes the information outlined above.

More practical guidance and considerations on data protection are available in the following documents, which may be applicable to the work of development professionals:

- [Handbook on Data Protection in Humanitarian Action](#) by the ICRC
- [Considerations for Using Data Responsibly at USAID](#)

For an overview of global data privacy laws and policies around AI, refer to the following resources:

- [Global Data Privacy Laws 2019: 132 National Laws & Many Bills](#) by Graham Greenleaf
- [OECD AI Policy Observatory](#)

Table 2: Common challenges with accessing data for use in ML models

Risk	Description	Mitigation Strategies
Scarce Data	Often, data that are relevant and appropriate for addressing development challenges are not available.	Reuse data collected for other purposes to overcome data scarcity or invest in collecting new data to create a foundation for future ML work. Alternatively, you may decide not to pursue ML approaches at this time.
Repurposed Data	Variables in a repurposed dataset may be only indirectly related to the relevant quantities of interest for your project.	Learn what you can about the data collection methodology, limitations, and original intent of collecting the data to decide whether it is appropriate for your use case. Identify to what extent the people, communities, or things you're interested in are adequately represented in the data.
Biased Data	Some of the most abundant development data sources may be biased by, for example, underrepresenting poor, rural, and/or minority populations.	Discuss with your data science experts and other team members which biases are important to consider given the local context and intentionally look for how these biases affect the data. Consider how the effects of bias may be mitigated by model design choices and explore the implications for how the model should be used in practice.
Lack of Access to Data	Some data may be available, but you won't have ownership, use rights, or user consent to use it for your project.	Consider collecting consent to use the data from the population described in the data, exploring data access and sharing agreements, or finding additional partners with such agreements in place.
Old Data	Data that may be considered a gold standard, such as health diagnostic protocols/guidelines, could be outdated and not accurately reflect current best practices.	Speak to subject matter experts where possible, to understand whether there are more recent data available from reliable sources. Talk to your team about the difference between the old data and what's being observed in practice and determine the best way forward.

 **CRITICAL DECISION**

Where will the Data Come From?

- What relevant data are available to address the problem?
These could be internal or general data sources or data from a strategic partner.
- Are there any organizations we could partner with to access additional data?
These could be organizations who could benefit from the outputs of the ML model.
- 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?
This data would be used to compare the performance of the ML predictions to actual outcomes (or proxy data).
- If needed, can we generate data in another way?
For example, by collaborating with the data scientist to create synthetic data or identifying good proxies for the data we are lacking?

Building a Project Team

The Value of Diverse Perspectives

As people with different perspectives and experiences, we accumulate different assumptions about how the world works. Acknowledging this reality is the first step to avoiding bias, followed by collaborating with a diverse group of stakeholders to help us understand those assumptions and to build more effective, inclusive solutions.

Development experts and technologists generally have different experiences, skill sets, and priorities. This diversity of perspectives is both enriching and challenging. However, **it increases your team's ability to understand, anticipate, and identify potential impacts of the ML model and how it's used.**

Bringing diverse perspectives together requires intentionally including different stakeholders throughout the project lifecycle. **The perspectives may come in the form of professional expertise or in the lived experiences of the people you are working to impact.** This guide identifies key stakeholders to consider engaging in different phases of the project lifecycle.

Table 3 (on the next page) is not prescriptive or exhaustive but aims to highlight when and how different stakeholders can be included in the project. Some stakeholders can take on a combination of roles – for example, your project coordinator might also be the subject matter expert. This table is meant to describe the complementary roles required to deliver your ML project.

KEY THEMATIC AREA

Responsible, Equitable and Inclusive Design

More perspectives are generally better than fewer. Those with technical training will likely be best positioned to make technical design choices, but you can still **ensure that people with diverse backgrounds, subject matter expertise, and context awareness have meaningful participation in the design process, including those likely to be affected by the predictions ML tools make.**

It's always worth asking who isn't at the table and what they might be able to contribute. Bringing in local voices 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 and feedback sessions as you move through model design can help you integrate local perspective and knowledge.



Table 3: What roles should different stakeholders have on the project team?

Typical Stakeholder(s) in Role	Role	Considerations
<p>Development Experts</p> <p><i>These could be subject matter experts, M&E professionals, or individuals focused on specific development problems, such as health outcomes</i></p>	<ul style="list-style-type: none"> • Manage and participate in the design and build of the ML model • Develop problem definition, assist with defining “representative data” for this context, and identify potential risks (e.g., where there are likely to be embedded assumptions and biases) • Ensure due consideration for local context in critical decisions throughout the project • Identify the local experts and anticipated users to be consulted throughout process 	<p>These stakeholders are critical for ensuring the context of the project is understood by various stakeholders and, in particular, the data science advisors.</p> <p>Sometimes ML model design goals (for highest efficiency or accuracy) can cause a model to evolve in such a way that its outcomes begin to diverge from its original purpose. The development experts should engage actively to ensure that model outcomes continue to align with development priorities.</p>
<p>Data Science Advisors/Data Scientists</p> <p><i>These could be someone in-house or outsourced for the project</i></p>	<ul style="list-style-type: none"> • Co-design and build the ML model to fit the problem context • Be transparent with development experts about design choices made • Prepare and provide clear explanation of exploratory data analyses, model design choices, and model performance metrics 	<p>Data science advisors may often be less familiar with the development problem and project context than development experts, which could impact the overall strength of the ML model. While they ultimately will be the partners with the expertise to build a model, their design decisions should be informed by those of other project team members.</p>
<p>Partner Organizations</p> <p><i>These could be implementers, training partners, local organizations involved in initial data collection, or M&E organizations</i></p>	<ul style="list-style-type: none"> • Collaborate on the model design and build • Facilitate access to data • Provide input on model iterations • Evaluate project implementation 	<p>ML projects often benefit from the expertise and connections of many organizations. You may have a variety of different partners, formal or otherwise, at various points of the project.</p>

Project Coordinator

This could be a distinct role, such as a project manager, or an additional role for one of the other stakeholders

- Coordinates the various moving parts of the project and manages relationships among different stakeholders

Not only does this role focus on coordinating various aspects of the project, but it also helps bridge the cultural, language, contextual divides that may exist between the various stakeholders. Even if the coordinator isn't a subject matter or technical expert, he or she should ensure various perspectives and voices are included throughout the project.

End-User

These are the people or communities who would use or be impacted by the output of the ML model

- Participate in consultations and co-design sessions
- Provide feedback on the effects of the ML application

Hearing from these stakeholders is important for context building and buy-in. Sometimes the feedback and suggestions they make may not be technically possible, so it will be important to manage realistic expectations.

Decision Makers

These are the people within the community or organization where the ML model may be integrated who will decide what happens with the use of ML after the project and who shape how the results of the model inform organizational operations

- Stay informed of how the model is being built and its accuracy
- Provide context for how ML model may be integrated into broader organizational goals

Those who will ultimately be acting upon the results of the model should have a good understanding of how it was developed and how it will need to evolve. This helps appropriately set expectations, build trust where merited, and retain skepticism where needed to ensure ML models are not used in ways at odds with how they were intended.

? CRITICAL DECISION

Who are the Strategic Partners and Project Beneficiaries that Should be Involved in this Project and How Can They be Included Throughout the Process?

Partnerships

- What partnerships should we develop for the project?
These could be organizations that have access to data or an interest group that would benefit from or be adversely impacted because of an ML implementation.
- What practices will we adopt to encourage collaboration across data science experts, development experts, and other key members of the team?
These may be anything from weekly meetings to workshops or co-design sessions.

Project Stakeholders

- Who is/will be impacted by the ML application, and how can we make sure we hear from them?
These are people/a community who will be directly or indirectly impacted by the ML application.
- How can we create opportunities for consultations with different stakeholders during the design and implementation of the project? (*refer to the section titled [Human Centered Design](#) on page 32*)

In-Sourcing vs Out-Sourcing Specific Data Skills

You may consider whether your organization should hire technical expertise, such as software developers or data scientists, in-house rather than partnering with an external firm. This depends on what role ML plays in your organization's strategic direction and core offerings. If ML is expected to be used across numerous projects and/or for an extended period, it may be beneficial to hire certain roles. However, if its intended use is for a limited time or a single project, partnering with an external data science team would be recommended.

 An additional resource that may be helpful for data scientists new to working on machine-learning problems in the development sector is: [Exploring Fairness in Machine Learning for International Development](#) by MIT.



CRITICAL DECISION

Should We Outsource the Technical Skills?

There are numerous factors to consider when thinking about outsourcing or building your own team. The following are not exhaustive and should be discussed within your organization and project team.

Outsourcing

Advantages	Challenges
<ul style="list-style-type: none"> • Works well when ML/AI needs are short term • Works well if project would benefit from broader expertise across data science, user design, and coding that may all be accessible through a firm 	<ul style="list-style-type: none"> • Limited awareness of the context • Potential pressure to use a global solution

ML Software as a Service (SAAS)

Advantages	Challenges
<ul style="list-style-type: none"> • Out-of-the-box solution can be deployed quickly • May not need to hire or outsource ML technical skills 	<ul style="list-style-type: none"> • Limited customization • Model may not be appropriate for the context • Subject to changes outside of your control

In-sourcing

Advantages	Challenges
<ul style="list-style-type: none"> • Works well when ML/AI projects will be a long-term endeavor for the organization • Data scientists may be more accessible and responsive • Allows for greater awareness of the context 	<ul style="list-style-type: none"> • Added costs to organization • Not having a clear sense of the types of skills and competencies needed or how to hire for them



DECISION AID

Assessing Feasibility of Using ML/AI

Before beginning an ML-backed project, you should assess the extent to which it is a feasible approach for your problem in context. First, you should understand how the problem is currently being solved. Not only does this create a benchmark for future evaluation, but it also provides insight to the current failures and where ML could play a role in improving/replacing the solution.



CRITICAL DECISION

Is ML Worth Trying Out as an Approach to Solving Our Problem?

Rate the following statements based on your project's context to determine whether ML is the right tool for you to use.

Rate the Following Statements from 1 - 5

1 = strongly disagree, 5 = strongly agree

1. The problem we want to solve is one that aligns with capabilities of ML/AI (e.g., predicts, classifies, or explores new relationships in data or would otherwise improve the status quo approach).

If you rated it below 3, consider: Exploring alternative approaches to solving your problem that do not rely on ML.

2. There are relatively large data sets that adequately represent the population of interest; these data sets are available to my project.

If you rated it below 3, consider: Reusing data from new data sources or partners, collecting data, and/or identifying partners with mutual interest in data sharing and collaboration.

3. Available data sets are machine-readable. (Handwritten or data stored in PDF files will not be as readily available for machine learning based approaches.)

If you rated it below 3, consider: Exploring whether there are alternative data sets or whether there are sufficient time and resources available to get data into a machine-readable form.

Rate the Following Statements from 1 - 5*1 = strongly disagree, 5 = strongly agree*

4. The output (target variable) we want to predict is well defined, relevant, and measurable.

If you rated it below 3, consider: Redefining the problem and consulting with domain experts to determine the most appropriate output variable for your context.

5. There are local stakeholders with familiarity and access to local data, knowledge of local context, and ability to facilitate input from the project's target population.

If you rated it below 3, consider: Approaching different local organizations or community initiatives you can partner with.

6. There are local stakeholders with technical expertise in data science with whom we could work.

If you rated it below 3, consider: Consulting local networks or universities and/or issuing an RFI to identify new partners.

7. There are resources available for a data science advisor/team to build and maintain/update the ML model(s) over the life of its use.

If you rated it below 3, consider: Applying for additional funding/grants or pursuing arrangements for mechanisms to access necessary computing infrastructure/services and/or "in kind" support from tech partners.

8. There is potential to secure resources and develop infrastructure sufficient to support the ML/AI application if sustainability is desired.

If you rated it below 3, consider: Evaluating how critical long-term sustainability will be, and at what point the ability to scale a potential ML application should influence investments in exploring ML approaches.

Rate the Following Statements from 1 - 5 *1 = strongly disagree, 5 = strongly agree*

9. We can clearly and concisely articulate what we hope to achieve with our machine learning approach.

If you rated it below 3, consider: Consulting with non-technical team members, such as a communications specialist, to help refine and describe what you are aiming to do. By accurately and effectively communicating how ML will be used for your project, you will be able to set and manage the expectations of various stakeholders.

TOTAL

Calculate your total and map it to the scale below to help you understand how well-suited ML is for your project. While these are not hard-cutoffs, **low scores should prompt greater attention to the identified alternatives** and reconsideration of whether ML is a worthwhile and appropriate approach for your project at this time.



Note: You may notice this Decision Aid doesn't include a cost analysis. ML is a burgeoning field, and with its wide range of applications, it's tricky to pinpoint exact costs or budget line items. These will vary based largely (but not exclusively) on the underlying maturity of existing digital systems, the anticipated scale, hiring decisions, cloud computing considerations, and multiple variables related to how ML models will be incorporated into existing processes.

To assist with determining costs related to the technical workflow of building and implementing your model, refer to [this article](#) by Cognifield. Actual costs will vary depending on the location of your technical experts and data labelers, as well as the computational needs of your model.



LEARNINGS FROM THE FIELD

Using Machine Learning To Identify Out-of-School Girls In Rural India

PHOTO: IDINSIGHT

Partners: Educate Girls and IDinsight

Problem to be Solved

Identifying villages with the highest concentration of out-of-school girls in four states in India to inform where program activities were most needed to (re)enroll girls in school.

Summary of the Project

Educate Girls works to locate out of school (OOS) girls in a bid to (re)enroll them in school and provide remedial education. These OOS girls are typically concentrated in highly disadvantaged clusters, with 95% of them living in 50% of villages.

To identify target beneficiaries, Educate Girls would conduct door-to-door censuses across thousands of villages, which was both time consuming and expensive. As the program scales, machine learning offers a more efficient and informed mechanism of identifying key villages to target.

IDinsight worked with Educate Girls to build and test various ML models to determine an appropriate method of predicting areas of high concentration of OOS girls. The models were built using existing programmatic data from 8,000 villages, which were linked to publicly available predictor data for the same group of villages. The predictor data came from sources such as the 2011 census and the annual census of school facilities and included values such as caste, literacy rates, and poverty indicators. These were then used by the ML model to predict the number of OOS girls in a village (the target variable).

KEY THEMATIC AREA

Strategic Partnerships and Human Resources

An important aspect of the success of this project was the longstanding working partnership between Educate Girls and IDinsight. IDinsight has worked on numerous technical projects for Educate Girls, from monitoring and evaluation to building digital tools for data collection. Having worked together before meant there was an existing foundation of trust and comfort with each other's way of working.

Having co-located teams helped build context and enable clear communication. When this wasn't possible, the IDinsight team would work remotely, then meet in person once or twice a month.

By establishing a trusted relationship between team members, IDinsight and Educate Girls were able to ensure that frequent, open, and productive communication helped contribute to a viable, problem-aligned application of ML.

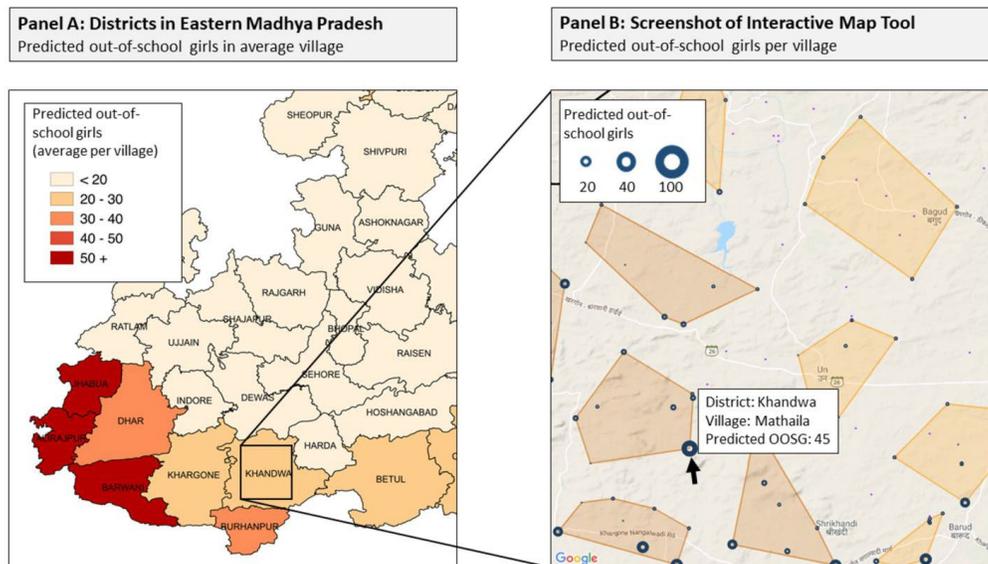


PHOTO: DINSIGHT

Results

IDinsight was able to validate the accuracy of the model by **comparing the target variable predicted by the model (the number of OOS girls per village) to actual data collected by Educate Girls** after the predictions were made for a given village.

[Find out more about this project.](#)



In this map, larger circles denote villages with more predicted out-of-school girls, small dots denote villages with few out-of-school girls (not recommended for targeting), and polygons show optimized clusters of villages that a single field team could viably cover.

Source: https://ssir.org/articles/entry/can_machine_learning_double_your_social_impact#

MODULE 2: MODEL DESIGN AND BUILD

Objective of this Module

This module aims to help you understand the different considerations around model building, data practices, and designing for sustainability so that you are familiar with these aspects in the context of an ML implementation.

At the end of this module is a decision aid which provides questions and considerations you can discuss with your data scientist and stakeholders while the model is being designed and built. It's important for you to be comfortable with these aspects of the model design so that you're equipped to mitigate risks of unintentional harms and increase the likelihood that the ML model will work in context.

Making Decisions About

- [Considerations when Building the ML/AI Model](#)
- [Designing for Sustainability](#)
- [Data Laws and Regulations](#)
- [Identifying Risks and Incorporating Safeguards](#)
- [Preparing to Assess Model Performance](#)



PHOTO: NKECHI CHARLES

Considerations When Building the ML/AI Model

Model building and selection involve decisions about how best to represent different aspects of the real world in a computational framing. 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 potentially more accurate but harder to interpret, and they must clearly recognize tradeoffs among the choices they've made. Similarly, modelers 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.

When these tools fail unevenly for different groups of people, the people affected may be unfairly denied services or singled out for scrutiny. **The cumulative burden of this “selective” failure can compound existing marginalization or inequity.**

To understand the different steps involved with building the model as well as questions and decisions to make throughout the process, refer to the decision aid [Building Your Project's ML/AI Model](#) on page 44.

While this process may seem purely technical, it **has real implications for a model's fairness and development impact**, which is why it is **important for non-technical experts to participate in the process of model design and development.**

Example: Unintended Consequences of ML/AI Models

Choices made during model building and development can have far-reaching consequences. The U.S. example of the [Allegheny Family Screening Tool](#) illustrates such consequences. The tool is a model designed to assist humans in deciding whether a child should be removed from her/his family due to abusive circumstances. It was noted that referrals occur three times more for African-American and biracial families than white families. The unwanted biases of the model stem from a public dataset that reflects broader societal inequities – a number of observers argued that middle to upper class families, who tend to be white, received lower risk scores because their use of the same services that functioned as risk factors in the model was

captured in data from private providers and not included in the model. By relying on publicly available data, those who had no alternative to publicly funded social services, predominantly low-income families, were overrepresented in the data. Despite the underlying intent of development projects, they, too, can result in disproportionate harm to minority groups. This could be caused by training data which contains embedded prejudices around (for example) ethnicity, caste, or religion. By interrogating how the model is being built and incorporating safeguards to mitigate risks, development experts can reduce or remove harmful outcomes. Refer to the section [Identifying Risks and Incorporating Safeguards](#) on page 37 for more.

Designing for Sustainability

Ensuring sustainability can be time and resource intensive and is context dependent. The points below introduce enablers for the sustainability of your ML project. They have largely been taken from [Principles for Digital Development](#) and are neither prescriptive nor exhaustive.

 For a more in depth approach to identifying and implementing mechanisms for sustainability, refer to DIAL's guide [Beyond Scale: How to Make Your Digital Development Program Sustainable](#).

Human-Centered Design

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**.

Often technologies are developed by people and organizations who do not reflect the people and communities who are influenced by the system. Assumptions and biases can become embedded into the system, resulting in ineffective technologies or harmful consequences.

By asking your data scientist to incorporate Human Centered Design (sometimes referred to as User-Centered Design or Design Thinking), you can increase the likelihood that **the system will be rooted in an understanding of the characteristics, needs, and challenges the technology is meant to solve for**. This is also a way to include the perspectives and voices of the people who will be most impacted by the implementation of your ML application.

Conducting Human Centered Design (HCD) **means partnering with stakeholders throughout the project lifecycle, co-creating solutions, and continuously seeking out and incorporating their feedback**. To understand who should be involved at different points in the project lifecycle, refer to [Table 3](#) on page 21.

There are HCD tools and techniques available to help gather information through conversation, observation, and co-creation. You can refer to the [Principles for Digital Development: Designing with the User](#) for resources and tips around conducting HCD.

Open Source Vs Proprietary Tools

There are several open source options to access ML algorithms. Well-established packages like [scikit-learn](#), [TensorFlow](#), and [PyTorch](#) are available to download. Newly emerging algorithms and complete models are often shared through sites such as [GitHub](#), though they may not be as thoroughly maintained. By making use of open source technology, your project can benefit from the work done by other organizations without needing to invest heavily in software development costs. High-quality open source software is supported by a community that produces documentation, bug fixes, and new versions. Your data scientist may even contribute to global resources by adding new features and sharing them with the community.

KEY THEMATIC AREA

Strategic Partnerships and Human Resources

Push for transparency from your data scientists and 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 building and participate in the key choices they are making**. This will give you insight into how and why the tools work as they do so you can communicate these with other key stakeholders. At the same time, external ML experts can learn about the priorities and needs of development practitioners.

A common misconception about Open Source tools is that they are completely free. Users can access the software for free, yet there remain costs associated with customization, hosting, and ongoing maintenance of the system. (Developers sometimes joke that open source is “free like puppies” rather than “free like pizza.”) The [total cost of ownership](#) should be considered alongside the costs of a licensed proprietary solution.

Interestingly, many software development companies leverage open source tools to develop commercial applications that they sell through licensing models based on specific use cases, their business model, or type of customer (to name a few).

KEY THEMATIC AREA**Responsible, Equitable and Inclusive Design**

ML finds patterns in data, but ML models do not “know” anything about the patterns identified – including whether or not these are patterns we want to replicate in the future. Including subject matter experts, people who understand the local context, and diverse perspectives enhances the alignment of ML-based tools with the norms and values we want to promote.

Table 4: The costs and benefits of open source and proprietary licensing models*

Open Source

Pros	Cons
<ul style="list-style-type: none"> • Free to use, update and distribute • Global community of practice that contributes to ongoing system enhancements • Greater degree of control over changes • Better security 	<ul style="list-style-type: none"> • May require customization • Need for hosting infrastructure • Requires technical skills for customization and maintenance (manual upgrades), which may require hiring this skill set if it's not already on team

Proprietary

Pros	Cons
<ul style="list-style-type: none"> • Out-of-the-box solution • Dedicated team and customer support • Includes hosting infrastructure 	<ul style="list-style-type: none"> • Subject to ongoing maintenance fees • May result in vendor lock-in • No access to the source code • Limited ability to share with partners

*Note: this is a simplification of the amount of licencing structures available for software development

Interoperability

Often digital solutions are built and implemented in isolation of other systems, resulting in disparate systems unable to communicate with each other. This can have implications for the effective use of a technology, where instead of creating efficiencies, additional workflows must be created to transfer information between systems.

By understanding and integrating with existing systems, your team can allow for effective reporting, planning, and decision-making, leading to increased buy-in from key stakeholders such as the government.

Strengthening Local Technical Capacity

Ideally, ML tools for development should be built and maintained by local data scientists. By working with local companies, you can help grow fledgling technology sectors and leverage the local knowledge and experience of technologists.

Unfortunately, local talent isn't always available. In the absence of local expertise, many development ML projects may rely on data scientists who live and work in another country or continent. These long-distance partnerships make it even more crucial to prioritize open and frequent communication and build capacity among local project stakeholders.

Strengthening Governance Structures

Governments around the world are wrestling with the policy implications of AI, and in-house data science expertise is often in short supply. Even developed countries struggle to find the right balance between promoting innovation and avoiding risk.

Strong governance requires robust laws for the protection of personal data and adequate resources and expertise to enforce these laws. In many communities where development projects work, the weakness or absence of personal data protection laws is a widespread problem that can create opportunities for malicious actors to surveil and manipulate with impunity. To address this, you can consult with your data scientist on developing your own data management policy (around access, use, storage, and ownership) to govern your ML application.

KEY THEMATIC AREA

Enabling Environment

Data Management and Data Hosting in the Context of ML Projects

ML initiatives, as with all software technology projects, **require server hosting infrastructure** for delivering services. Most technology companies default to cloud hosting which has a number of benefits, especially for companies exploring ML based projects, including:

- ability to handle large volumes of data and transactions;
- flexible pricing options;
- quick and easy scalability;
- built-in redundancy;
- high levels of uptime and availability; and
- strong security.

With an increase in the number of countries adopting data sovereignty laws – which set legal conditions stating that data about citizens or residents must be collected, processed, and stored inside the country – **governments have started to favor local hosting over the cloud.** Without the presence of big hosting services (such as AWS and Microsoft Azure) within each country or region, projects using ML rely on local servers to host and maintain project-related data and services.

Typically, local hosting is costly because it requires considerable human resources and infrastructure to set up and maintain. **Companies that offer local hosting have to buy hardware, ensure its physical safety, and have the expertise to support the servers – all in addition to maintaining their software.** These costs can add up, presenting barriers for ML projects considering local hosting, to the point of making an ML project infeasible in countries with data sovereignty laws and no major cloud hosting service providers.

It will take a long time for all countries to have appropriate **and reliable local hosting services, and where there are data sovereignty laws in place, there will continue to be a tension between compliance and project feasibility and performance.**

Data Laws and Regulations

ML/AI projects are often data-intensive and should be subject to appropriate data policies and compliance regulation. Regulations such as the European General Data Protection Regulation (GDPR) **provide guidance on the collection, storage, use, and ownership of data.** These regulations are designed to protect and limit the use of people's personal information to safeguard them against misuse. Despite being considered a new standard, GDPR may not be practical to implement in its entirety within your context.

As countries develop and implement laws, you should work with your legal adviser and information communications technology (ICT) partners to understand what is required for your project. Regulations to pay attention to include those that govern data privacy, payment and settlement systems, digital marketing, use and transmission of personal data (such as medical records), user consent, data ownership, localization, and sharing. Most national governments require that such laws and regulations be published on government websites with a point of contact for both national and sub-national governments.

Data laws and regulations are subject to change. **It's important that you and your data science advisor stay up to date with these laws** because they may have significant implications for the design and delivery of your digital solution.

In the absence of local or project-specific guidance, consider existing frameworks to determine the best data handling approach for your project and create Standard Operating Procedures (SOPs) around data access and sharing.



Additional considerations can be found in USAID's [Considerations for Using Data Responsibly at USAID](#).

This has been demonstrated by recent events in South Africa, whose Parliament approved the Protection of Personal Information (POPI) Act in 2013. This Act aims to:

- promote the protection of personal information processed by public and private bodies;
- establish minimum requirements for the processing of personal information;
- provide for the rights of persons regarding unsolicited electronic communications and automated decision-making; and
- regulate the flow of personal information across the borders of the Republic.

At the time the Act was approved, **South Africa had no cloud hosting service providers which could handle large volumes of data and transactions locally**, with most technology companies using cloud services in Europe or America. Technology companies, including those focused on ML, had to explore other avenues of hosting – from setting up the infrastructure themselves, to partnering with academic institutions, or hosting it within government departments.

Fortunately, by the time the Act took effect in July 2020, two major companies had launched hosting services in South Africa. By accessing these services, barriers to achieving compliance were substantially reduced.

However, **many projects using ML tools had not yet developed or implemented appropriate data sharing and management protocols.** These protocols should have detailed appropriate and responsible sharing of data with various stakeholders, such as funders (to meet reporting requirements) and partners (for service delivery). Their absence resulted in project delays where essential services, such as the distribution of medication for chronic conditions, were suspended and uncertainty about the long-term viability of using ML was introduced.

Data policies should focus on all aspects of the data management process and exist to safeguard the end users of digital systems. Software technology projects need to approach compliance to these policies holistically, with each piece, from data hosting to data sharing processes, equally prioritized and implemented.

**CRITICAL DECISION**

What Data and Legal Requirements Do We Need to Comply with?

- What are the data privacy laws in the country/region?
If none, where can you borrow best practices to inform your own approach?
- How will data be collected and meaningful consent instituted, when necessary and appropriate?
- If the data are being reused or repurposed, do you need to re-collect consent?
e.g., determined based on whether consent to use one's data in an ML model was obtained at the time of data collection
- Where does data collected in-country need to be stored?
How will this data be protected from breaches?
e.g., locally, or cross-border cloud hosting
- Who will own the data?
e.g., local government, our organization, the consortium of partners
- What other parties will be interested in this data, and what are their motivations? What would happen if data were unintentionally shared with them?
- For people-specific data: What is the extent of users' data ownership? Can they access, remove, or edit their data?

KEY THEMATIC AREA

Responsible, Inclusive and Equitable Design

The use and application of ML in development contexts is still nascent. Following the best practice of collecting consent isn't always achieved in practice nor will it, even if practiced, be meaningful in many contexts where digital literacy is low. There is an imperative for development experts to consider risks, learn together, and actively shape norms in the field in acknowledgement of where existing paradigms are not effective.





DECISION AID

Identifying Risks and Incorporating Safeguards

There are various risks to consider when implementing your ML model. This section describes these risks and provides tools to support identifying possible mitigations as your team develops a ML model.

Table 5: Common risks with implementing ML models

Risk	Description	Case Example(s)
Systemic Exclusion	Structural inequities and biases often result in some groups being less represented in the training data used to develop ML models, which in turn results in ML models working less well for those groups. Groups subject to existing inequities and bias may also miss opportunities to benefit from ML/AI tools more than other groups.	Computer vision models for assessing skin cancer work less well on dark-skinned patients, who are less represented in medical image databases. Learn more here.
Disproportionate Harm	Sometimes models will work better (in terms of rate of “correct” predictions) for some people, communities, or regions. This can stem from many factors, such as under- or misrepresentation in data sets or between-group differences in the relationship between the input and output variables. Where the real-world cost of errors in model prediction are high, this can lead to significant and disproportionate harms experienced by the affected group.	<p>Credit scoring models in Latin America that were developed using a single model for both women and men showed more rejections of credit applications from women than models that estimated creditworthiness separately for men and women, using the same cutoff for credit determinations.</p> <p>When separate models were developed for men and women, the results showed more women would be offered credit, including some who were rejected under the gender-agnostic model. This suggests that the combined model may not be modeling creditworthiness for women optimally, resulting in more false-negative predictions (i.e., denying credit to credit-worthy women). Creating separate models for subgroups in the data may be one way of mitigating errors that otherwise have harmful consequences. Find out more here.</p>

Risk	Description	Case Example(s)
Codified Social and Structural Bias	<p>Rather than provide fair and objective decisions, ML models can instead serve to reinforce existing social biases.</p> <p>For example, gender bias is widespread in existing data about hiring, educational achievement, and access to financial capital in most countries and may easily be replicated in ML models.</p>	<p>Plan International's initial attempt at developing an avatar for a virtual chat-assistant to help young Filipino women identify job opportunities matched with their skills was found to embed gender norms and make gendered job recommendations (e.g. hairdresser, caregiver).</p> <p>Subsequent changes were made to intentionally combat gender bias by including women in the tech development team, in addition to building in use of gender transformative assessments and agile integration of findings to improve inclusion.</p> <p>For more on this, watch NetHope's AI Ethics Webinar Series: Part I.</p>
Inability to Interpret Predictions and/or Explain Decisions	<p>Most algorithms rely on processes so complex there is little ability for humans to decipher what it means in real-world terms or explain decisions to those affected by them. This also makes it difficult to determine whether biases which influence predictions exist or to identify interventions that would influence the model outcome.</p>	<p>CIAT's data-driven agronomy work intentionally uses ML models that are easier to interpret so they can discover new insights about how crop management practices affect crop yield. If they had chosen more "black box" algorithms, they might be able to predict crop yield but would not be able to identify the specific inputs that most strongly determined yield and, therefore, wouldn't have actionable insights for farmers.</p>
Excessive Trust in ML/AI Technology	<p>People often have unrealistic expectations for algorithmic systems.</p> <p>This can lead to overlooking model failures, under-investment in monitoring model performance over time, and letting biases go unaddressed.</p>	<p>Models optimized for one context may be expected (incorrectly) to work equally well in others, inflating expectations of what may actually be possible. For example, a health diagnostic tool developed in Europe may be adopted for use in parts of Africa but shouldn't be trusted to perform with the same level of accuracy. Ensuring a slow introduction to a new context, including dedicated efforts to monitor performance and allow for humans to override automated decisions, can help mitigate risks of excessive trust.</p>

A key advantage of digital systems is their ability to scale rapidly. But this is also a source of risk as their effects can reach millions of people before they are fully understood, causing **ineffective or inequitable results to scale rapidly as well.**

See the section [How Models Fail](#) on page 53 for more information on how these risks manifest.



PHOTO: MICHAEL DAWSON/FHI 360



CRITICAL DECISION

What Level of Risk Exists for My ML Project?

Determine the level of risk with implementing an ML model by rating the statements for **each risk type**. High risk doesn't necessarily mean you should not proceed with your project. The aim is to identify risk and explore what actions you can take to lower risk, then determine whether those safeguards will be enough to alleviate concerns.

It might only be possible to assess some of these risks as the data science and/or tech expert completes different steps in the model building process. For example, the risk of systemic exclusion might only become apparent during the exploratory data analysis (EDA) phase. These assessments should ideally be done in collaboration with others in your team to ensure that a variety of perspectives are considered.

See the [Decision Aid: Building Your Project's ML/AI Model](#) on page 44 for more details on the steps your data scientist will follow and how to work with your data scientist to mitigate risks.



DECISION AID

Evaluating Risk within Your Project

Consider where your project sits with respect to each of the following risks:

Risk 1: Systemic Exclusion

Statement	Rate (high/medium/low)
There is a risk of some people, communities, or regions of interest being misrepresented or underrepresented in the training data.	
The project team lacks diversity and may have collective blind spots around inclusion.	

Approaches for Lowering Risk

- Talk to your data scientists early and often about the data they are using to develop the model and who is represented in it.
- Identify which groups are important in your context and ask your data scientists to share with you an analysis that describes how and to what extent they are represented in the training data.
- Even when you do your best to create representative datasets in the model build phase, it is likely that you may identify additional exclusions during implementation. As implementation moves forward, continue to talk with your data scientist to understand options for addressing exclusion. These could include **data augmentation**, introduction of a new model, or decommissioning the use of ML altogether.

Risk 2: Disproportionate Harm

Statement	Rate (high/medium/low)
There is a risk of some people, communities, and regions being disproportionately harmed by the outputs of the ML model.	

Approaches for Lowering Risk

- Ask your data scientist to share with you an analysis of model accuracy across different sub-groups in the data. For whom or in what contexts does the model perform best? For whom or in what context does it fail more often?
- Consider what impact this will have in your context and discuss it with your data science advisor. Ask your partner to talk through different approaches that would minimize discrepancies across groups and work together to find an approach that is acceptable from both technical and programmatic viewpoints.

Risk 3: Codified Social and Structural Bias

Statement	Rate (high/medium/low)
There is a risk of bias (e.g. with respect to gender, ethnicity, economic class, etc.) embedded in the training data about certain people, communities, or regions.	

Approaches for Lowering Risk

- This kind of bias may not be evident from technical measures of model performance. For example, a ML model that uses measures of individual skill to predict jobs in which an individual is likely to succeed may be trained using historical employment records. The model may produce technically accurate results, providing predictions that closely follow historical trends, yet still embed gender bias. Identify the structural biases that are most evident in your context and be intentional in understanding how they may be affecting the patterns you see in your data.
- Talk to your data scientist about these in advance and often. For variables you don't want to form the basis of model discrimination, such as gender, race, and income, ask for exploratory data analysis that will allow you to examine patterns in the data.
- You can also discuss intentional efforts to mitigate the influence of bias in model performance. This often requires including the same variables whose bias you are trying to reduce, e.g., race or gender (sometimes called protected attributes) so that their influence on model results can be assessed.

Risk 4: Inability to Interpret Predictions and/or Explain Decisions

Statement	Rate (high/medium/low)
There is a risk of the outputs from the model being disputed where there is no clear understanding (or explanation) of how the output was generated.	

Approaches for Lowering Risk

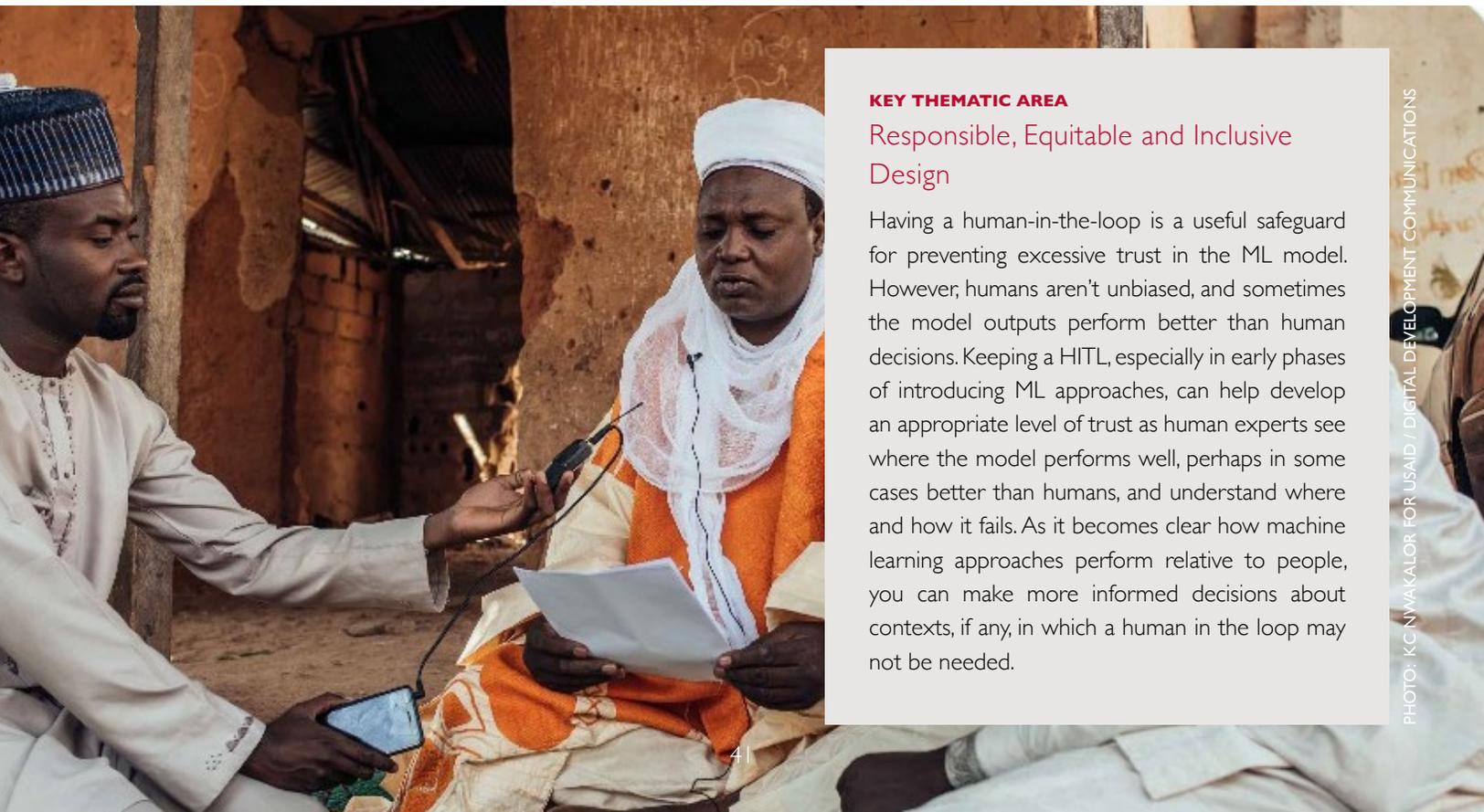
- Accountability in ML/AI systems can often be strengthened if we understand how input variables influence the outcome. There are multiple ways to build for interpretability, such as using less complex algorithms like decision trees or taking steps to test how models respond to certain changes using approaches like Local Interpretable Model-Agnostic Explanations ([LIME](#)). See *Decision Aid: Building Your Project's ML/AI Model* for suggestions on talking with your data scientist about how to improve interpretability of ML models.
- If the ML output influencing or automating a decision impacts an individual (e.g. prevents them from receiving a loan), the model should be more interpretable. Some regulatory requirements might contain a clause for the "right to explanation." In these cases, your model will need to be interpretable. If interpretability is essential for your project, you should strongly consider alternatives to opaque "black box" systems.

Risk 5: Excessive Trust in ML/AI Technology

Statement	Rate (high/medium/low)
There is a risk of the ML model outputs being acted on without adequate monitoring, unknowingly causing harm.	
There is a risk of inflated expectations for model performance.	

Approaches for Lowering Risk

- Ensuring those using the outputs of ML/AI models feel empowered to override results if they suspect an error can minimize the risk of over-reliance on an automated decision. This approach is sometimes called keeping a **human-in-the-loop (HITL)**.
- Ensuring adequate resources to monitor model outcomes over time. If ML/AI tools are filling a gap in available human resources, the ML tools may become de-facto decision makers even if human decision makers know they are likely to make errors.
- Keeping healthy skepticism of model results can help in remaining vigilant in checking for bias and help prevent unintended consequences. It can improve the overall accuracy and relevance of the ML model through continuous training.
- Refer to the *Implementation* module for more on including a human in the decision-making process.



KEY THEMATIC AREA

Responsible, Equitable and Inclusive Design

Having a human-in-the-loop is a useful safeguard for preventing excessive trust in the ML model. However, humans aren't unbiased, and sometimes the model outputs perform better than human decisions. Keeping a HITL, especially in early phases of introducing ML approaches, can help develop an appropriate level of trust as human experts see where the model performs well, perhaps in some cases better than humans, and understand where and how it fails. As it becomes clear how machine learning approaches perform relative to people, you can make more informed decisions about contexts, if any, in which a human in the loop may not be needed.



CRITICAL DECISION

What Safeguard(s) Should be Implemented to Prevent Harmful Consequences for Stakeholders?

Complete this table to begin action-planning for the identified safeguards.

Risk <i>What kind of risk exists?</i>	Safeguard/Mitigation Description <i>How will you lower the risk?</i>	Stakeholders <i>Which stakeholders should be involved in this?</i>
<i>(Illustrative example)</i> Disproportionately failing to work for ethnic minority populations	<i>(Illustrative example)</i> - Assess extent to which the minority group is present in training data - Evaluate model performance across groups to check for discrepancies - Seek feedback with members of the group during initial implementation to understand their experience of model's use	<i>(Illustrative example)</i> - Data science advisors - Local partners - Representatives from the minority group

Preparing to Assess Model Performance

Ultimately, the adoption and retention of ML-based approaches will involve **comparing the model's performance to existing (or other) methods of solving the specific problem**. This includes assessing the model's technical performance (decision process accuracy, speed, and transparency) as well as its performance in context (fairness, inclusiveness, risk to vulnerable populations, etc.). **ML/AI tools, at least at this level in their evolution, will not be the best choice for all problems.** Sometimes low-tech human judgement will offer equally good or better outcomes.

As with any development project, it's important to define the indicators and measurements of success. These will help with your evaluation of the ML model during the implementation phase. Your evaluation should consider both the model's accuracy and its performance in the context of your project's activities.



CRITICAL DECISION

How Will We Measure Success?

Discuss the following questions with your project stakeholders to inform how you plan to evaluate success.

Model Accuracy

- What level of accuracy (or failure) are we willing to accept?
- Across which subsets of the population will it be most important to compare model performance?
- How will we test the solution once it's been deployed?
e.g., collect and compare actual outcome data to the predicted value.

Performance in Context

- What would be "enough" of an improvement over the status quo approach in efficiency, accuracy, fairness, and scope to make the implementation a success?
- What mechanisms can we incorporate to gather concerns and questions and get feedback about model performance?
e.g., from local experts, end users, strategic partners and stakeholders.

KEY THEMATIC AREA

Responsible, Equitable, and Inclusive Design

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

KEY THEMATIC AREA

Adaptive Management

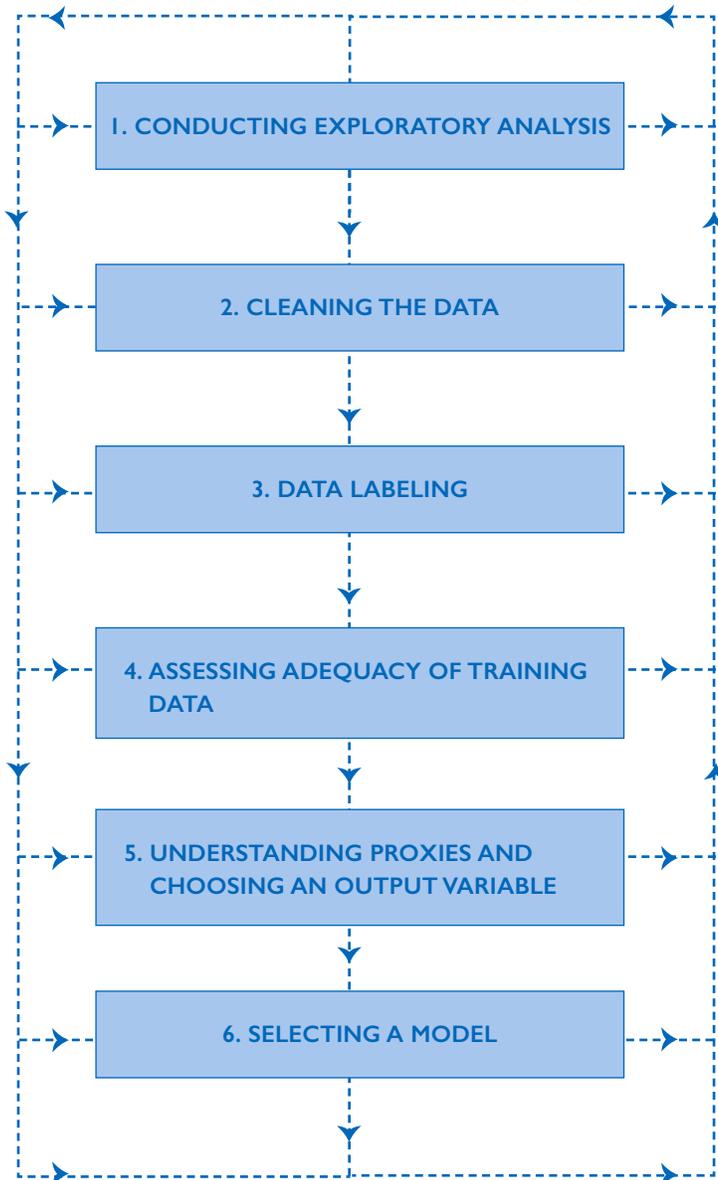
You should plan to go through an iterative process of measuring performance and determining how and whether to use your model in context – you may revisit decisions made earlier about data, your target variable, and how you ultimately use the ML model depending on how well (or poorly) the model works.



DECISION AID

Building Your Project's ML/AI Model

Figure 4: The different steps involved in building a ML model. Your starting point largely depends on the state of your dataset, e.g., if you already have clean or labeled data, those steps will be much easier.



This decision aid is intended to help a project manager understand the technical steps in the process of building a ML model. If your data science experts will be building a model, this can help you understand the technical steps they will be going through in the process. Knowing this can help facilitate a collaborative process, in which non-data science experts can engage and participate in key decisions in the design process that will have implications for model performance and use in context.

Building your model should be an iterative process. The steps outlined in this tool are not prescriptive and don't need to be completed in this order. They can be repeated as often as necessary. **Discuss the best approach with your data science expert and create opportunities to collectively review how the revisions made during each step impact the training data and model outputs.**

Before your data science experts get started, use the questions in this tool to talk with them about the model design. As they work through these steps, stay engaged and work together to understand the key choices and tradeoffs that may have to be made and to understand their implications.

Note: For the purposes of this tool, we use the term "modeler" to describe the data scientist building the ML model.

1. Conducting Exploratory Data Analysis

The modeler will conduct exploratory data analysis (EDA) to learn about the data's contents, structure, and potential biases. Socially aware EDA processes can help mitigate problems before they arise, and this is an ideal stage at which to question how existing social inequities might be reflected in data. At this point, the modeler and development expert(s) should be working closely together to ensure the context is well understood.



More technical details on Model Building can be found in the following resources:

- Appendix: Peering under the hood of USAID's [Reflecting the Past, Shaping the Future: Making AI Work for International Development](#)
- MIT's [Exploring Fairness in Machine Learning for International Development](#)

2. Cleaning the Data

Data cleaning is the process of preparing your dataset for analysis using different techniques such as standardizing definitions, changing units, imputing missing values, and removing implausible values. This can result in the final dataset looking very different to the one you started with, and you may need to re-evaluate how to implement your ML solution.

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 original data subjects, as the process of re-interpretation may unintentionally overlook or distort local concerns or nuances. Discuss the impact of your data cleaning with relevant stakeholders.

3. Data Labeling

Data labeling is the process through which the modeler provides values of the output variable for each instance in a training data set. For example, labels might indicate whether a photo contains a hut or which words were spoken in a recording. Data labeling may require additional data collection, crowdsourcing, or expert curation.

There are concerns about the assumptions and biases that arise from having data labeled by people who are not familiar with the local context. Where possible, engage with local communities to assist with labeling your training data. This can be a powerful way to build local capacity, improve appropriateness of labeling for your context, and garner buy-in.

Data cleaning and labeling can result in a reinterpretation of the data and embed underlying assumptions and biases, causing a misrepresentation of the original dataset. Additionally, if the data are irreversibly grouped or reclassified in the cleaning process, it could have implications for equity or make it difficult to separate by sub-groups in later analysis. **To mitigate this, you should plan out your bias analysis before you begin cleaning and ensure that the modeler is working with a team with good knowledge of the context during these steps.**

4. Assessing Adequacy of Training Data

In the *Evaluate Feasibility* module, we discussed accessing different data sources. Once this is done, the modeler will need to **determine whether the available data are adequate to build a reliable, fair, and accurate (enough) ML model.** Training data should be of high enough quality to have trustworthy instances to train the model.

KEY THEMATIC AREA

Adaptive Management and Responsible, Equitable and Inclusive Design

During each step of model building, the data used for the ML model is subject to change. Sometimes data cleaning encodes the data cleaner's assumptions about which data points deserve more attention or the reliability of data 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.

In some instances, different groups or categories might be underrepresented in the training data. Here, the data scientist can adopt methods (such as weighted regression) to compensate by increasing their influence relative to other groups present in the data.

Predictions will be based on whatever patterns are in the data. Those patterns may reflect aspects of the real world we seek to change, in which case predictions will reflect the same, unsatisfactory status quo. For example, [Amazon had to scrap an automated recruiting tool](#) that was favoring male candidates over female candidates because the model was trained on historical patterns that included a larger share of men being hired. Decisions about which training data to use shape a model's impact on the world.

It may be impossible to know what appropriate input variables will be in advance of trying them in your model. Modelers often collect as much data as possible – including things that may seem irrelevant – in the hopes of finding a good set of predictors. This approach can be problematic when the data include poorly understood biases or omissions. On the other hand, **careful analysis can sometimes reveal new variables that are good predictors** (and perhaps more equitable) than traditional ones.



CRITICAL DECISION

Does the Data Need to be Cleaned, Grouped and Labelled Differently?

Discuss with Your Development Experts

- What are the implications of the way in which data are grouped or classified?
- Which variables do we need to make sure remain in the data set in order to check for bias?
- What characteristics of the data would make us uncomfortable moving forward? What are our standards for timeliness, quality, and representativeness?
- What local people or organizations can assist with data labeling?

Ask Your Data Science Advisors

- How will you test whether there is enough data that is of adequate quality and representative of the population we're working with?
- What are some of the issues you're finding in the data cleaning process?
- How were variables combined, relabeled, or changed in the data cleaning process?
- How familiar are you with how the data was labeled? What kinds of kinds of labeling errors are you seeing in the data?

While data are rarely perfect, low-quality data may limit the use of ML tools. **Gaps or inconsistencies in data could lead to poor predictions.** Should this happen, you would need to source additional data. This could be done by integrating with an existing dataset or collecting new data.

Most data will likely reflect some bias, and the important part is for you and your team to recognize it and take steps to manage it.

**CRITICAL DECISION**

Do We Need to Find Additional Data Sources?

Discuss with Your Development Experts

- What does “timely” data mean for us? How recent do data need to be to be meaningful?
- Who or what is most likely to be excluded or overlooked in our context? Are there any key groups or people missing from the data?
- 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*
- What social, political, cultural, or economic biases might have affected the data collection and who/what is represented in it?
- Who else should we consult or partner with to better assess or strengthen representativeness of data?
- If we cannot address issues around data representation, how will this affect our implementation plan for this AI tool?

Ask Your Data Science Advisors

- How recently were data collected?
- How clear are the definitions of the variables in the data? Are there any variables where you're not sure what is actually being measured?
- How complete are the data? Are there concerns about quality or accuracy of any of the variables in the data?
- Are there any gaps or inconsistencies in the data?
- What concerns do you have about data quality?

5. Understanding Proxies and Choosing an Output Variable

Possibly the most crucial decision to discuss with your data scientist in the development of a **supervised learning model** is the choice of output variable. An **output variable** is the value you are trying to predict, and **should be easily quantifiable, unambiguous, and closely related to the problem** at hand. You will need to carefully select the output variable to avoid predicting the wrong thing.

Because output variables are often hard to define or measure, we have to rely on **proxies** to approximate what we are trying to predict. But not all proxies are good substitutes; sometimes they can be only weakly associated with what you really want to understand.

Poor proxies for input and output variables can bias model output if they reflect an underlying bias or are weakly associated with what you are trying to understand. Should this happen, you would need to determine whether another variable is more appropriate (and available) to use.

At this point, **it is crucial for your development experts to speak to your data scientists** to ensure they clearly understand what the model should predict. This will help them provide better recommendations on which input variables and output variables are appropriate.

Example: Choosing an Output Variable

Employers often want to understand whether candidates will be successful at their jobs. ML might be considered as an approach to try to predict success in a job from a number of other data points about a candidate. For example, one might choose to predict job success based on indicators of skill and personality of a candidate.

Modelers have to decide what data to use as inputs into the model (predictors of “success”) as well as the output variable that will designate “success.”

For input variables, we might wish to use things like “ability” and “personality,” which are inherently difficult to measure. We may rely on proxies for these inputs, such as the results from aptitude tests and personality assessments. These will each have some set of limitations, and potential bias, in how well they measure aptitude and personality.

Similarly, the output variable of job success is hard to measure. Without being able to place candidates in an environment that closely matches the job reality and assess their performance in advance of hiring them, we need to select a variable to act as a proxy for job success.

Does success mean that a candidate gets an interview, or is actually hired, or is retained in the position for at least 12 months? Does it mean the candidate reports a certain level of job satisfaction? Choosing an output variable requires you to have a clear understanding of what you want to know. You often will not have the opportunity to measure what you think the most appropriate output variable will be and will have to choose a proxy that may be less well aligned with “success.” In these cases, it is important to have a clear view of what your model is predicting and think through the limitations in deciding how to use the model. In this example, if all you have the ability to measure is whether candidates were hired, you still won’t know about their actual performance or whether it was perceived to be a good “fit” from the perspective of employer or employee. You also won’t know if the reason they were not hired is because they genuinely were not well suited in terms of aptitude or personality or because the input variables you chose were not a good proxy for aptitude and personality.



CRITICAL DECISION

What Should Our Output Variable be?

Discuss with Your Development Experts

- What assumptions are embedded in the variables we will use?
- Given what we know about context, how closely do the variables we’ve chosen reflect what we really want to know?
- What will the output or target variable tell us? What will it not tell us?

Ask Your Data Science Advisors

- How does the inclusion of different input variables affect model performance? Are there any you’ve found particularly influential?
- What are the limitations of using the variables we’ve initially chosen to use?

6. Selecting a Model

Once the data have been cleaned, assessed, and labeled, and the data scientist has a good understanding of what you want to predict and how you will use the results, you will be able to select an appropriate ML model. This will be determined based on factors such as:

- which data are available,
- what the model’s intended use will be, and
- estimates of model performance.

There are dozens of popular ML algorithms ranging from very interpretable methods to more complex and opaque algorithms. **Understanding which one to use depends on the goals of your project.**

Different models have different levels of interpretability, accuracy, and performance. Along with your team, you will need to determine whether the model's performance meets your project's requirements.

Example: Considerations on Interpretability

In some cases, a significant goal of building a ML model is to **identify the features that most strongly influence the output 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, the modeler must **ensure that the model is interpretable enough to determine which practices should be recommended**.

Sometimes the primary goal will be to get the right or most accurate answer, and you're not as interested in how it was

determined. For example, an agricultural cooperative may want to estimate crop yields for their region for planning and logistics. **Here, it may be worth trading off some interpretability to get higher accuracy in the prediction of crop yields.** For some tasks, such as many audio and image classification systems that rely on deep neural networks, it may not be possible to complete the task with an interpretable model, in which case you should carefully consider whether such interpretability is necessary.



CRITICAL DECISION

Which ML Model Should We Implement Based on its Predictions and Level of Interpretability?

Discuss with Your Development Experts

- How might users be excluded or harmed by the implementation of ML/AI? Where or for whom is this most likely to happen?
- What level of inaccuracy are we comfortable with?
- Under what circumstances might we need to be able to explain model predictions?
- Does the level of interpretability meet the needs of our problem?
- How do differences in accuracy or interpretability affect how we might implement the ML/AI tool?

Ask Your Data Science Advisors

- What are the error/misclassification rates for different groups/categories?
- Which models are most accurate? Which are the most transparent? What are the pros and cons of each model?
- Can you identify the factors that most significantly influence the outcome of the model?

MODULE 3: IMPLEMENTATION

Objective of this Module

This module aims to emphasize important considerations in preparing for and executing the implementation of your new ML model in the context of your project. It's important that you can identify some key choices and risks in order to ensure that you effectively and equitably integrate the ML model into your project and goals. It's a good idea to plan for a period of iteration in model development as you work with data science experts to "tune" or adjust a model for the optimal performance in your context. Further, plan for a "pilot" period or phased rollout to allow your team to test the model in context, further evaluate the model, and develop a plan to integrate it into practice. During this phase, you may decide to make changes to design decisions made when the team was initially building the model.

Making Decisions About

- [Evaluating Your Project's Application of the ML/AI Model](#)
- [Incorporating ML into Decision Making processes](#)

Evaluating Your Project's Application of ML/AI

Once the model has started making predictions, development experts need to evaluate the model for accuracy as well as application in the context of the project itself. Both levels of evaluation are needed to ensure ML is being applied accurately, fairly, and effectively to the project.

Model Evaluation

As you move past the initial model design and build stage, there remains a need for continuous evaluation and vigilance around certain common risks.

Model evaluation begins with getting a clear sense of how well the model performs – essentially, **how much decision makers can trust its predictions**. At a minimum, this means getting an estimate of its **out-of-sample accuracy**.

There's no single formula for model evaluation. We know ML models can be assessed using multiple, different measures of accuracy, and their performance will change over time. Traditional pre-and post-project assessment may be insufficient to monitor progress through adaptations or support interim benchmarks that may incentivize evaluating not only for model performance, but also aspects of representativeness, transparency, and tradeoffs that may become apparent through model design and use. ML tools differ from traditional software technology in that they undergo multiple iterations over the course of the project – they are never static – and we need to account for these iterations in our evaluation and create monitoring practices that encourage learning, adaptation, and improvement to models over time. Data scientists and users must creatively interrogate the model with across-group comparisons, sensitivity analyses, and other contextually rooted performance tests. This is discussed in greater detail in MIT's guide [Exploring Fairness in Machine Learning for International Development](#).

KEY THEMATIC AREA

Adaptive Management

Evaluating a decision system's performance requires continuous monitoring of its prediction accuracy. The real world is never static, and **many systems experience model drift, in which the real-world 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.

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. **This often requires an independent process for collecting and labeling new data**. To assess model drift, the model's predictions should be stress-tested against "live" outcome data after implementation.

Relying on a single number to characterize a model's performance can be a dangerous oversimplification. When a model is deployed in a context marked by structural inequity (around gender, age, ethnicity, geography, or other factors), it is essential to compare error rates across these categories.

Ask your data scientist about uneven failure rates and make sure you understand how these were evaluated. Start with how a model's errors can be quantified and what types of bias you're most concerned about given your project context. Identify subsets of the population (e.g., male/female, urban/rural) across which error rates can be compared. Consider the possible real-world consequences of uneven failure rates. Error testing and performance monitoring should continue after deployment, once the model begins evaluating "live" data. We will further discuss the problem of uneven failure rates below, in the next section *How Models Fail*.



CRITICAL DECISION

Is the Model Accurate Enough for the Needs of Our Project?

If not, discuss with your data scientist whether new training data should be added, a different model used, or the model decommissioned.

Questions for Consideration

- To what extent has the model improved efficiency, accuracy, or scale?
- What options exist to ground-truth predictions? How does the ML model compare to alternative ways to get at the same result?
- Does the model fail more often for some people, communities, or geographic contexts than others?
- Do additional measures to ensure fairness need to be incorporated into the ML model?

A team might build a functional ML model, only to **determine that it adds no immediate value over alternative methods**. In these instances, it may be better to defer to existing methods.

KEY THEMATIC AREA

Responsible, Equitable and Inclusive Design

Machine learning models may not work equally well for everyone. Although ML might lead to faster decision-making, it could result in unintentionally harmful outcomes for your project stakeholders. Ultimately, staying true to your development problems means being prepared to walk away from ML. Consider this if you aren't comfortable with the level of accuracy of the model predictions, if you're concerned about the potential risks for stakeholders, or if you're unable to make decisions based on the predictions.

TIP

Consider building a model card with your data science advisor to document and monitor your model's performance. Model cards provide a structured framework for reporting on ML model provenance, usage, and ethics-informed evaluation and give a detailed overview of a model's suggested uses and limitations.

Using this type of tool can help you understand the risks of implementing your model, from recognizing codified/structural bias causing disproportionate harm to preventing excessive trust in the model.

Google has created a publicly available [Model Card Toolkit](#).

How Models Fail

There is potential for ML-based tools to improve efficiency, precision, and effectiveness in development work and humanitarian assistance. But, as discussed in the Model Design and Build module and in this one, this impact is not guaranteed to be positive for everyone. Below are some of the most common ML failure points that researchers are working to better understand and mitigate. Recognizing these common failures and addressing them as they arise in your project is important for ensuring responsible, equitable, and inclusive design.

- **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.**

- **Uneven Error Balance**

Accuracy can be broken down into different types of errors, e.g., 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**. A model may have similar accuracy across two sub-populations but show a different balance of false positives and false negatives in different groups. This can create an uneven playing field.

- **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 a model that scores resumes based on past hiring records will discriminate against women.

- **Imperfect Proxies**

In many cases, the quantity we'd like to model isn't available, and we must settle for a related value, or proxy. Maybe we're interested in actual levels of crime committed but only have data about arrests. If the alignment between the "real" outcome of interest and the proxy isn't perfect, then models can develop blind spots – like missing un-arrested criminals or the arrests of innocent people. When that blind spot overlaps with existing disparities, it can compound existing bias. For example, when poor people have less access to healthcare, their needs will be under-represented in medical records.

- **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, though it would be important to keep the shortcoming in mind and not over-rely on the model. They may be fair in the sense that they are equally inaccurate for everyone.

Understanding False Positives and False Negatives: Real-World Implications

Although we care about accuracy and how accuracy may differ across groups of people affected by machine learning applications, accuracy alone may not be enough. It is important to look at the type of error that is occurring. This is important for understanding the cost of errors and determining what level of performance is acceptable for your program, as well as identifying whether there are systematic patterns in the type of error that occurs across sub-groups in your data.

In many cases, the real-world implications of a false positive are quite different than those of a false negative, and differences in types of errors can be another source of inequity, unintended consequence, or disproportionate harm.

Tables 6 - 8: Illustration of confusion matrices for understanding accuracy errors

Credit Scoring Models		Prediction	
		Will Repay	Won't Repay
Actual Outcome	Did Repay	True Positive Implication: individuals who are given credit and repay on time	False Negative Implication: individuals who are not given credit but would have repaid on time if you had given them credit
	Didn't Repay	False Positive Implication: individuals who are given credit but end up defaulting on a loan	True Negative Implication: individuals who are not given credit and wouldn't have repaid if you had given them credit

Supply Chain Forecasting Models		Prediction	
		Stock Needed	Stock Not Needed
Actual Outcome	Stock Was Needed	True Positive Implication: stock supplies were adequate to meet demand	False Negative Implication: demand was higher than predicted and stock outs occurred; health outcomes suffer
	Stock Wasn't Needed	False Positive Implication: stock over-supplied and risks expiration; inefficient use of a scarce resource	True Negative Implication: stock wasn't needed; no wasted stock nor inability to meet demand

Job Hiring/Skills Matching		Prediction	
		Individual Will Succeed Job	Individual Will Not Succeed Job
Actual Outcome	Individual Succeeded in Job	True Positive Implication: individual is offered a position in which they will succeed	False Negative Implication: individual is denied a job in which they would have succeeded (missed opportunity for success)
	Individual Did Not Succeed in Job	False Positive Implication: individual is given a job in which they won't succeed (set up for failure)	True Negative Implication: individual is denied a job in which they would not have succeeded

Incorporating ML into Decision Processes

Regardless of the ingenuity or accuracy of the model, it is up to people to determine how best to incorporate it into existing decision processes.

Many development applications of ML will not involve full automation, but rather use ML models to improve part of a larger decision-making process, subject to review by stakeholders. There is value in proceeding deliberately and building in opportunities for people to review decision-making.

The “right” role of ML in the decision process always depends on context, which **includes 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, such as rapid categorization of thousands of images of disaster-hit areas, relying almost entirely on the ML model may be the best, or only, viable approach. In other cases, full reliance on models may be unwise, and it may be preferable to have the model assist but retain a human perspective in making ultimate resource allocation decisions. When people remain **engaged in the decision process, they can weigh model results against contextual factors and use their own judgement.**

As a development expert, you’ll understand the context better than your data science advisors, and **it’s up to you to ensure that the degree of automation is appropriate for that context.** This requires thinking about how decisions are currently made and how users are likely to interact with new technologies. 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 consider how much error the existing decision-making process is able to tolerate and whether the ML-backed tool will be able to meet expectations.

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 – offering credit or granting parole or admission into a school – will negatively impact some while benefiting others. Technology might seem to ease some moral discomfort by taking difficult or controversial decisions out of our hands and making them quantifiable and ostensibly objective. However, if things go wrong or someone complains, it is all too easy to blame the front-line people relaying the results of a complex system for decision-making, even if they didn’t actually make the determination.

ML models 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.

KEY THEMATIC AREA

Responsible, Equitable and Inclusive Design

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.



More information on the factors influencing automation can be found on 62-63 of USAID’s [Reflecting the Past, Shaping the Future: Making AI Work for International Development](#).

For ML models that inform decisions about individual people, the development expert 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 how to improve it, is the model interpretable enough to provide an answer? If someone objects to the evaluation, is there a way 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.** As mentioned in the Model Design and Build module, it may be better to create formal channels for receiving feedback, providing explanations that end-users can understand, or correcting mistakes, rather than relying on ad hoc improvisation. Listening to the people impacted by development programs is always best practice – no high-tech tool will change that.



CRITICAL DECISION

How Should the ML Predictions be Incorporated into Our Decision-making Processes?

Questions for Consideration

- How would integrating ML predictions complement or disrupt the existing processes for decision-making?
- Where might human judgement still play a valuable role?
- Given the level of interpretability of the model, how much should it influence ultimate decision-making, and who should be involved in making that decision?
- What, if anything, has been (or will be) diminished by incorporating model predictions into our decision process?
- How should processes be changed (or created) to accommodate the inclusion of ML, if we decide to integrate it into decision-making?

Example: Missing Feedback Loops

ML innovations have been introduced in the mobile financial service sector across Africa. Traditionally, there were fixed credit scores (tables) that would be used to decide on an individual's loan value based on non-transactional data (e.g., salary, home ownership, demographic information). Now, ML models are able to make decisions about disbursing loans by using multiple data inputs such as call records, mobile money transactions, demographics, and income.

Most mobile money service providers have witnessed an **increase in call center traffic from subscribers** who are questioning the value of the loans they were awarded. These subscribers often blame the call center agents despite the fact they weren't involved in assessing and awarding the loan. These call center staff aren't the model designers; the models are generally designed by the senior managers from the finance, mobile money, commercial and business analyst teams.

Evaluating Performance in Your Project Context

Besides a model's technical performance, longer-term model evaluation can include **investigating how a model is actually applied to support decision-making in the context of project activities**. Data science advisors can engage in dialogue with users, learning which information products are helpful and exactly how they are being interpreted and used.

Feedback is an important mechanism for learning about how the results of ML models are being implemented in practice. Ultimately, **an ML application succeeds only if it contributes to the success of the larger effort it serves**.

There are a number of factors to consider when evaluating the implementations of ML. One of these is whether the use of ML provides an improved solution compared to how it was previously being solved.

Example: Evaluating ML Use in Context

An ML model was designed in the UK to triage and diagnose patients. This model was replicated in an East African country where the types of diseases differ, as does patient access to healthcare services. These differences resulted in the model producing predictions with a lower accuracy in the new context. The speed of triaging took three times longer with the ML application because the original app was designed to diagnose as well, which requires more specificity than triage. Diagnosing also took three times longer than clinicians making the same decisions.

At this point, the team could have deferred to the previous method of triaging without the ML model. However, the East African project team instead decided to work with local clinicians and the data scientists in the UK to update the model and make it more applicable to their context. While triage still takes longer than it had previously, the new approach offers other benefits, such as standardization of approach.



? CRITICAL DECISION

Is the Model's Continued Use Appropriate for the Context of Our Project?

EVALUATING IN CONTEXT

Upon adequately instituting risk mitigation strategies and implementing your A/ML project, you should start collecting data on how your "live" model is performing. At this stage, it is important to decide whether it's appropriate, within the context of your project, to continue to rely on the outputs of the model.

If not, consider whether (and what) additional safeguards need to be included or if the model needs to be updated or decommissioned.

- Does the model perform better in practice than the existing decision-making process?
- Does the model reduce or reproduce existing inequalities?

KEY THEMATIC AREA

Adaptive Management

Developing a Monitoring, Evaluation, and Learning Plan for projects with a ML component can be challenging given how many unknowns you'll have when you start. Therefore, building flexibility into your plan may be a key consideration.

At minimum, consider including in your plan key checkpoints to assess data representativeness, model accuracy in the context of your problem, and impact on project goals, as well as opportunities to revisit original indicators and metrics you may have chosen. Also consider building in ways to seek feedback from those affected by the use of the model over time.

Finally, identifying learning questions appropriate for your project can help your team not only better manage the project, but also help set expectations up front by laying out areas of uncertainty that you will be exploring through the project.

As development experts gain experience integrating ML/AI approaches in their work, building an evidence base and lessons learned can help strengthen our collective capacity to use technology responsibly and effectively.

Illustrative Learning Questions

- What are the data quality and data preparation requirements for effective use of ML/AI in our project context?
- What are the data integration and workflow transformation requirements for our use case?
- What safeguards will most effectively limit potential bias or unintended harm from the use of ML/AI in this context?
- How might we monitor and safeguard effectiveness?
- How can we ensure sustained capacity such that these data and model considerations will continue to be monitored and evaluated beyond the life of the project?

LEARNINGS FROM THE FIELD

Using Machine Learning to Deliver Agronomy and Financial Literacy Content at Scale



PHOTO: DIGIFARM

Partners: Safaricom (DigiFarm) and Arifu

Problem to be Solved

Increase agricultural smallholders' income and productivity by (amongst other things) delivering knowledge resources using ML.

Summary of the Project

DigiFarm, a digital platform for Kenyan farmers run by Safaricom, provides smallholder farmers with convenient, one-stop access to a suite of products, including financial and credit services, quality farm products and customized information on farming best practices. DigiFarm helps agribusinesses and small holding farmers share information and transact more easily. DigiFarm's vision is to change farmers' lives in a commercially sustainable way by addressing the following gaps:

1. Knowledge on best farming practices
2. Quality inputs
3. Access to financial services – credit & insurance
4. Access to markets.

The platform is accessible on basic feature phones, making SMS a primary form of interaction and engagement.

The majority of the learning content is focused on vocational agricultural skills - from agronomy to farming techniques - as well as financial literacy. This content plays a crucial role in addressing skills gaps and provides farmers with transferable technical skills that can improve yields.

The learning platform is powered by Arifu, a SaaS platform which uses ML/AI for interactive content delivery. To meet the learning needs of DigiFarm's users, DigiFarm and Arifu invested in the development of conversational AI algorithms using **NLP** and text analysis. Arifu collects the engagement data generated by farmers, using it to build their data pipeline and the algorithms used for content delivery. Through ML, the data can be used to discover relationships in how users engage with materials and to better target retention and reactivation campaigns to encourage continued use of the learning platform.

Results

Since DigiFarm launched in Kenya, the platform has reached over 1 Million farmers, of which over 300,000 farmers have accessed Arifu's learning content with more than 2 Million interactions on the service.

The experience of accessing Arifu's platform has helped farmers increase their volume of sales, increased in their yield, increased market access and increased access to credit.

This is showcased by farmer Sammy who said, "I had never heard about kikuyu grass before. I got to learn about it in Arifu trainings. Before I would feed my cows on the grass around my compound and maize stems when they were available. But after discovering kikuyu grass, my cows produce more milk than before."

[Find out more about this project.](#)



KEY THEMATIC AREA

Adaptive Management

Arifu measures the usage and impact of their ML/AI learning modules in various ways. They calculate usage by analyzing what content gets the most traffic and at which points users drop off in the user journey. Knowledge retention is evaluated through responses to quizzes, questions, and surveys. Arifu also measures customer satisfaction through a ratings system.

The Arifu team is also in the process of bringing to market the Arifu Skills Score, which uses proprietary algorithms to measure users' proficiency in and comprehension of concepts.

They further make efforts to ground-truth their predictive models with survey or partner data, measure impact as reported by end-users through lean data surveys, or by employing rigorous evaluation methodologies, including, but not restricted to, ongoing A/B testing, Randomized Controlled Trials, Regression Discontinuity Designs, and Difference-in-Difference models, which help causally attribute changes in the digital learning and associated behavior change to Arifu's services.

Operational Learnings with Implementing an ML/AI Project

Although the barriers to entry for ML/AI projects have been lowered due to availability of open source tools, there remains a need for highly experienced and skilled teams of data scientists and engineers who have previously worked in the data science field across different industries to support ML/AI projects.

The data scientists worked closely with the software engineers to build the learning **chatbots**. They developed the text analysis algorithms and continue to analyze the data from the platform to ensure the service aligns with customer requirements, while meeting the product objectives.

MODULE 4: POST IMPLEMENTATION

Objective of this Module

This module briefly discusses the long-term sustainability of using ML, and the need for ongoing evaluation of your implementation.

Making Decisions About

- [Long Term Sustainability](#)
- [Ensuring Responsible, Shared Learning](#)

Long Term Sustainability

The ability to adopt and maintain ML tools depends strongly on the decisions you made when designing and implementing the model as well as the capacity of partner organizations (as discussed in the [Designing for Sustainability](#) section on page 31). Leveraging ML tools requires capacity to use and maintain the models from which they are built. Decisions such as open source vs. proprietary models can influence the model's long-term use. This can be enhanced by aligning the requirements of the model use and maintenance with the capacities of the organizations who will eventually be using the tools as a routine part of their work.



CRITICAL DECISION

How Can We Create Conditions for the Long-Term Sustainability of the Project?

Questions for Consideration

- What limitations are there (if any) on making ongoing adjustments to the model?
- Who will be responsible for the ongoing updates and evaluation of the model?
- What organizational capacities would need to be strengthened or developed to maintain this technology in the future?

Should You Decide to Retire or Decommission Your ML Implementation, Consider the Following

- What will we do with any data collected during implementation?
- What will we do with the solution's code?

Ensuring Responsible, Shared Learning

As the development community works to make our interventions more effective and efficient, **it's critical that we become savvy consumers of emergent technological tools.** ML and AI show considerable promise, so **we have a responsibility to learn about how these tools are developed, tested, validated, and shaped by data, assumptions, policies, people, and relationships.** If we hope to leverage them in our work, we must understand their powers, limitations, and risks across different contexts.

This includes learning and sharing experiences about both failures and successes. These are emerging technologies which will not always work as intended. Failing is a necessary part of progress. The *Design and Implementation* modules described many of the known risks and failure points. **We must construct appropriate safeguards to allow us to fail responsibly, transparently, and in a way that ensures failures will produce shared learning, not repetition.** We should create mechanisms and incentives to explore these tools with integrity, acknowledging both the good and the bad, before they are rolled out at scale.



CRITICAL DECISION

How Will We Capture Results of the ML Model and Unanticipated Outcomes?

Questions for Consideration

- What mechanisms will we build in to ensure opportunities to hear from those affected?
- What steps will facilitate learning from the successes and failures of this project?

END-TO-END CASE STUDIES

Bringing it All Together

We hope that this guide has given you a deeper understanding of the concepts, key steps, and critical decisions related to implementing ML/AI in the context of development initiatives. As you prepare to apply these leanings to your own activities, let's consider how these components manifest in the context of two hypothetical case studies. Each of these is framed in the context of the project lifecycle to see how the project team considers the critical decisions associated with each phase. As you read this final section, we encourage you to think about your own projects and how your team would handle similar scenarios and critical decisions.

We hope you find this guide a useful tool as you consider applying ML/AI to your own project activities.

CASE EXAMPLE: YOUTH EMPLOYMENT

Problem Statement: Youth unemployment in the region is high; it is especially challenging for youth lacking good educational opportunities and living in informal settlements far from economic hubs of the country to find work. Many young people may be well-suited for entry-level jobs, but limited educational opportunities have left them without documentation attesting to their skills. Traditional hiring criteria of job history and educational achievement consistently exclude these youth from opportunities to succeed.

Background: A team of researchers in a civil society organization wanted to determine a better way to match youth with fulfilling, stable jobs. Knowing that reliance on job history and educational records would embed bias against many of the youth they hoped to serve, the project team wanted to explore other potential predictors of job attainment, satisfaction, and retention.

Evaluate Feasibility Model Design and Build

- CRITICAL DECISIONS**

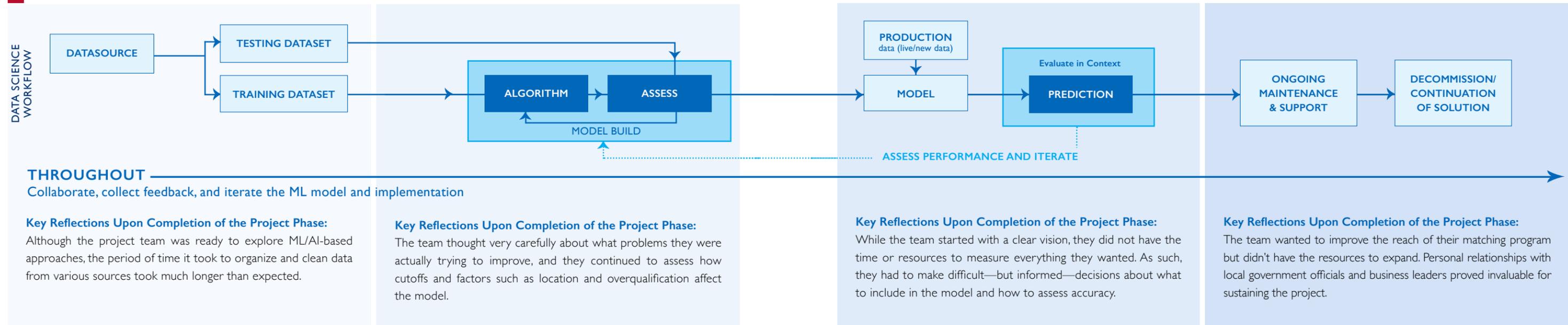
 - Match your problem to ML/AI capabilities**
Two key ML/AI capabilities are useful for addressing this problem. First, the project team can discover new relationships for employment fit by using alternative data sources (besides just educational achievement and job history). Second, the team can score how likely a given candidate is to succeed in a position.
 - Identify potential data sources**
The team has years' of data, including aptitude assessments, personality traits, candidate preferences, location data, and data from previous interviews. However, this large volume of data is not organized into an easily-manipulated database. The data are generated by the population they intend to work with, and therefore it is representative. At each step in the process, the team is reevaluating who might be excluded or underrepresented in the data. If the team considers expanding to other countries or providing insights into different labor markets, more consideration would be needed to assess appropriateness of the data.
 - Identify relevant stakeholders and project partners**
The team does not have the in-house data science expertise to develop a digital matching platform using machine learning and, therefore, chose to partner with an external firm. Their key criteria were that the firm have a local presence, ability to closely collaborate, and top-notch skills. The team is moving forward with a local subsidiary of a global technology company headquartered across town from their office and is consistently working to break down the silos that isolated the "tech" people from "development" people.

- Identify relevant data and legal requirements**
The team is collecting substantial amounts of data on each person, including identifiable and personally sensitive information. As such, they're building a secure database and gaining truly informed consent from each person. Because the data is their own, no data-sharing agreement was required.
 - Define your criteria for "success" with your ML/AI model**
The team initially wanted to measure youth job satisfaction, but they're realizing this would be difficult to accurately measure over time. Instead, they've settled on "being hired" as a target variable because it is feasible to track. The team knows that while this measure does indicate formal participation in the economy, it does not measure other things that might be important (e.g., subsequent career growth, employer satisfaction, retention).
 - Identify potential risks and appropriate safeguards**
The team is building in several assumptions to create an appropriate model for the situation. For instance, they're tuning the model so that those who would do well in more complex jobs don't "get stuck" in less complex ones; that is, candidates won't be matched with jobs for which they are technically suited but likely overqualified.

With respect to a candidate's commute, the team knows that structural biases are often the reason that candidates live far away from job opportunities. Rather than hard-coding this structural bias into the model, the team is making sure that applicants understand when they're matched with a long or dangerous commute; at that point, they have a discussion about the financial and personal implications of taking that job.

Implementation Post Implementation

- Determine if the model is accurate enough for your project**
The team needed to decide what an "optimal" model would be: a "threshold" model, where individuals are shown all jobs that they might be qualified for (even jobs they're overqualified for); or a true matching model, where individuals are matched precisely to jobs at their skill level. The team recognized that the first model could leave lower-skilled job applicants at a disadvantage, as jobs may be filled by higher-skilled applicants. On the other hand, the second model does not display all jobs that higher-skilled applicants might be qualified for, technically limiting their options. The team worked with their data science experts to understand how the model would be different under these two designs. Eventually, they decided on a true matching model—one that is designed to improve the overall system by precisely matching the skill level of applicants to jobs.
 - Determine how the results of the ML/AI model will be incorporated into the decision-making process**
In this case, the ultimate goal is to have the ML/AI model fully automate the matching process in order to quickly provide candidates with a suitable job opportunity. Knowing that this would take time, the team realized that they could also collect data on whether youth were interviewed for a particular position as an interim proxy metric of whether that organization might be likely to hire that individual. The team is iterating the model with this new information, while keeping in mind that hiring is the ultimate goal—not just interviewing.
- Identify additional steps for long-term sustainability**
As the team is working to improve the accuracy of their matching process, they're also developing relationships with local business leaders to socialize the matching program and demonstrate its value. In addition, the team is working with the Ministry of Labor to make the case for the merits of more innovative and fair hiring practices.
 - Plan for long-term learning**
After iterating with the model and engaging in reflection throughout the course of this project, the team has become keenly aware that no model is perfect. In fact, they continue to be up-front with youth, employers, business leaders, and the government that this model is not intended to predict retention or job satisfaction. Instead, it only predicts whether a youth is hired, and additional feedback on performance and retention would be important to gather over time to improve the model. The team is now developing ways to identify individuals who are interviewed but ultimately never hired. Unpacking these cases could help the team understand the root of the problem and how they might better serve youth who clearly demonstrate skill but are not finding work.



CASE EXAMPLE: MEDIA INTEGRITY

Problem Statement: Media that is ostensibly “fact-based” can often contain hidden opinions or biases, making it difficult for the public to assess the accuracy of the news in their state, region, or country. Identifying instances of opinion in journalism can be an important part of measuring media integrity, yet it is time intensive and subjective.

Background: A team of researchers wanted to use ML/AI to detect the presence of journalist opinion in news articles. The team recognizes that author opinion or bias in news articles is only one aspect of media quality and that other media-strengthening efforts would be important as well. This project is exploratory; the team hopes that the tool could be useful but is approaching it as a learning activity, with tempered expectations about exactly who might use the tool and for what purposes.

Evaluate Feasibility

Model Design and Build

Implementation

Post Implementation

CRITICAL DECISIONS

❓ Match your problem to ML/AI capabilities
The application involves interpretation of text and is well-suited to a natural language processing (NLP) capability. The application sorts strings of text into two groups based on whether they include author opinion. If the ML/AI approach performs this task successfully, it could free up humans to perform more complex tasks related to measuring quality of media (e.g., reliability of sources).

❓ Identify potential data sources
The data are from local media sources in the local language and were readily available on the internet. They are, therefore, representative of the articles the internet-using public was reading on a day-to-day basis. However, the data are not well standardized for NLP analysis, and the team has to do significant cleaning and organizing of the articles into similar formats.

❓ Identify relevant stakeholders and project partners
The team does not have in-house data science capacity and explored the idea with several partners. Budgets are limited for this pilot effort, ruling out many global tech leaders in ML/AI. As such, the team is moving forward with a research group at a U.S.-based university that has previously researched text classification.

❓ Identify relevant data and legal requirements
The data (i.e., news articles) are publicly available; the team doesn't have to worry about data or legal requirements.

⚙️ Define your criteria for “success” with your ML/AI model
The model sorts strings of text into two groups based on whether they contain opinion. The target variable of “opinion” thus had to be labeled in the training data by a human and is therefore subject to labeler bias. The team knows this step takes a lot of time and that labelers have to work to reduce their personal bias when labeling something as “opinion.” They are holding several rounds of reviews to ensure the training data are as robust as possible.

❓ Identify potential risks and appropriate safeguards
This is a low-risk project. The model is not making a direct decision about a person's life, career, or well-being. Nonetheless, the team is aware that personal bias is widespread, and their model could influence the way in which bias (in media) is measured. The team is transparent in acknowledging that it may be impossible to eliminate subjectivity and bias altogether; rather, they're aiming to standardize the definition of bias through an NLP algorithm instead of haphazardly relying on local labelers whose biases vary substantially. They are also considering hiring local students to help with the labeling, as they may be more familiar with local vernacular and be more attuned to bias language in this region. Still, no labeling strategy—including this team's—can be 100% accurate.

❓ Determine if the model is accurate enough for your project
The development experts and data science team are planning to look at several measures of accuracy. Because a key aim of the project is to determine whether instances of opinion could be identified, the data science team suggested a metric known as “recall,” which measures the proportion of all instances of opinion in the media articles that are detected by the ML model.

The data science team also suggests looking at other metrics; a model could have perfect recall simply by predicting all instances as instances of opinion - that is, if everything is labeled opinion, it won't miss any! To complement this, the team also decides to measure the proportion of articles detected by the model that were true instances of opinion, a metric known as precision.

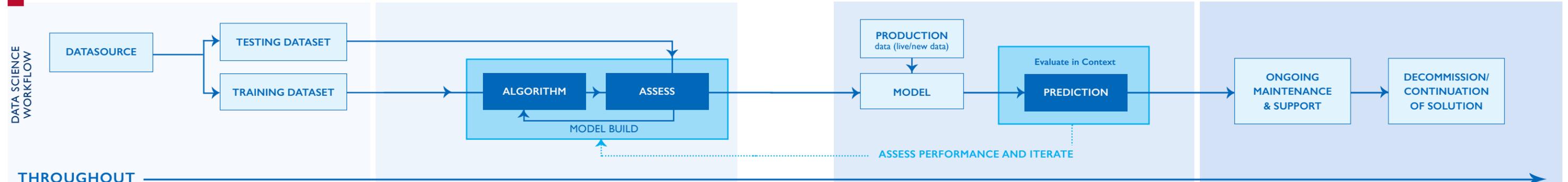
The team recognizes that they have to make a value judgment about what to optimize for: ensuring more instances of opinion are detected by the model, knowing there may be more false positives included, or potentially missing instances of opinion, but knowing most of those detected are in fact true instances of opinion.

❓ Determine how the results of the ML/AI model will be incorporated into the decision-making process
If the project is eventually implemented at scale or within an organization, the team would need to decide whether the process should be fully automated or whether to retain a “human in the loop” to make final determinations about bias.

❓ Identify additional steps for long-term sustainability
Though this project is a pilot, the team is discussing how it could be used in practice. Though the data (i.e., news articles) are publicly available, there is a possibility that the outcomes of the project would be unfavorable to the news outlets themselves. As such, the team is in communication with the news outlets, keeping them apprised of the project throughout and listening to any concerns they have. At the same time, such technology could be useful to the news outlets as a quality control metric for their publications. If the project moves forward, the team would need to consider and advise on how to build data science capacity within the implementing organization.

❓ Plan for long-term learning
As the project is progressing, the team is routinely comparing model results against the manual process to determine how the model outputs compare - are certain instances of opinion the model has trouble identifying? In what ways is it performing better, and where might a human determination still be needed?

If a model like this were to be deployed by media companies, watchdog groups, or government regulators, the team would need to consider how the algorithm interacts with other assessments of media quality and at what point a “human in the loop” would no longer be required.



THROUGHOUT Collaborate, collect feedback, and iterate the ML model and implementation

Key Reflections Upon Completion of the Project Phase:
Development experts and data scientists often have to be judicious in scoping their project due to budget constraints. It is important to find a partner with the right skill set and with whom you can develop a good working relationship.

Key Reflections Upon Completion of the Project Phase:
Bias is unavoidable when building an ML/AI model. Practitioners should aim to effectively identify, communicate, and build safeguards against biases and assumptions. In addition, using publicly available data has pros (easy to access, free) and cons (often not cleaned or organized).

Key Reflections Upon Completion of the Project Phase:
There is no “magic” number when it comes to assessing model accuracy. Rather, model accuracy is balanced against other factors, including the potential that an inaccurate model could cause harm.

Key Reflections Upon Completion of the Project Phase:
Not every project will be successful or sustainable. Sometimes, the model itself is of limited utility. Other times, the cultural or political environment proves prohibitive.

CONCLUSION

This guide has explored some of the key considerations that should inform the conceptualization and implementation of ML and AI components within a development project. New, automated decision systems can offer considerable and rapid efficiency gains, but we must always remember that they embed numerous and ongoing human decisions. These may be intentional or unintentional, benevolent or malicious, general or highly context specific. As with physical infrastructure such as roads and bridges, digital infrastructure can all too easily encode unexamined bias – sometimes in ways that can undermine development gains.

As outlined in this guide, a wide variety of decisions need to be made at different stages of the project lifecycle: from which stakeholders should be involved and how, to measuring model accuracy and success, to determining overall whether ML is an appropriate tool to use for your development context. There is no one-size-fits all answer to these questions.

But whatever the specific ML/AI technologies and applications you consider, broad guidance is offered in the four thematic areas woven throughout this guide:

- Responsible, equitable, and inclusive design
- Strategic partnerships and human capital
- Adaptive management
- Enabling environment for ML/AI

These focal points should help you and your project team make the best possible choices at each stage of the project life cycle.

ADDITIONAL RESOURCES

Useful Tools and Guides

- Artificial Intelligence in Global Health: Defining a Collective Path Forward
<https://www.usaid.gov/cii/ai-in-global-health>
- Artificial Intelligence (AI) Suitability Toolkit for Nonprofits
<https://solutionscenter.nethope.org/artificial-intelligence-suitability-toolkit>
- Beyond Scale: How to Make Your Digital Development Program Sustainable
<https://digitalimpactalliance.org/research/beyond-scale-how-to-make-your-digital-development-program-sustainable/>
- Exploring Fairness in Machine Learning for International Development
https://d-lab.mit.edu/sites/default/files/inline-files/Exploring_fairness_in_machine_learning_for_international_development_03242020_pages.pdf
- Handbook on Data Protection in Humanitarian Action
<https://www.icrc.org/en/data-protection-humanitarian-action-handbook>
- Introducing the Model Card Toolkit for Easier Model Transparency Reporting
<https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html>
- OECD AI Policy Observatory
<https://oecd.ai/>
- Principles for Digital Development
<https://digitalprinciples.org/>
- Reflecting the Past, Shaping the Future: Making AI Work for International Development
<https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf>

Examples and Case Studies

- AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind
<https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/>
- Algorithms and Artificial Intelligence in Latin America, A Study of Implementation by Governments in Argentina and Uruguay
http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf
- The Allegheny Family Screening Tool: Predictive Risk Modeling in Child Welfare in Allegheny County
<https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx>
- Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women
<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- Applying Machine Learning to Laboratory Data: Predicting Suppression of Next HIV Viral Load In South Africa
<http://www.heroza.org/wp-content/uploads/2019/12/Policy-brief-32-Predicting-VL-suppression-pubready-2019-12-10.pdf>
- Can Machine Learning Double Your Social Impact?
https://ssir.org/articles/entry/can_machine_learning_double_your_social_impact
- DigiFarm: A Digital Platform for Farmers
https://mercycorpsagrifin.org/wp-content/uploads/2019/05/DigiFarm-Platform-Case_Final_.pdf

- Gender-Differentiated Credit Scoring: A Potential Game-Changer for Women
<https://financialallianceforwomen.org/news-events/gender-differentiated-credit-scoring-a-potential-game-changer-for-women/>
- Innovations in Platform-led Upskilling: Digifarm Trains its Farmers Through SMS with the Support of a Face-to-Face Advisory Network
<https://www.transformationalupskilling.org/digifarm>
- Putting Data at the Service of Agriculture: A Case Study of CIAT
<https://www.usaid.gov/digitalag/ciat-case-study>

More on ML and AI

- A Gentle Introduction to Computer Vision
<https://machinelearningmastery.com/what-is-computer-vision/>
- How to Explain Natural Language Processing (NLP) in Plain English
<https://enterpriseproject.com/article/2019/9/natural-language-processing-nlp-explained-plain-english>
- Understanding Model Predictions with LIME
<https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b>
- What is Advanced and Predictive Analytics?
<https://bi-survey.com/predictive-analytics#def>

Data

- ADS 579 – USAID Development Data
<https://www.usaid.gov/sites/default/files/documents/1868/579.pdf>
- Considerations for Using Data Responsibly at USAID
<https://www.usaid.gov/responsibledata>
- Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda
<https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda>
- Global Data Privacy Laws 2019: 132 National Laws and Many Bills
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=338159
- Unpacking the Issue of Missed Use and Misuse of Data
<https://www.unglobalpulse.org/2019/03/unpacking-the-issue-of-missed-use-and-misuse-of-data/>

Budgeting

- The Cost of Machine Learning Projects
<https://medium.com/cognifeed/the-cost-of-machine-learning-projects-7ca3aea03a5c>
- The Total Cost of Ownership of Open Source Software
<https://www.ipeg.com/the-total-cost-of-ownership-of-open-source-software/>

Open Source Technical Tools

- Github www.github.com
- PyTorch <https://pytorch.org/>
- Scikit-learn <https://scikit-learn.org/stable/>
- TensorFlow <https://www.tensorflow.org/>