IDinsight

Humanizing Artificial Intelligence

Putting People at the Center of Data-Driven Decisions

datadelta

Social sector insights driving change.

Humanizing Artificial intelligence

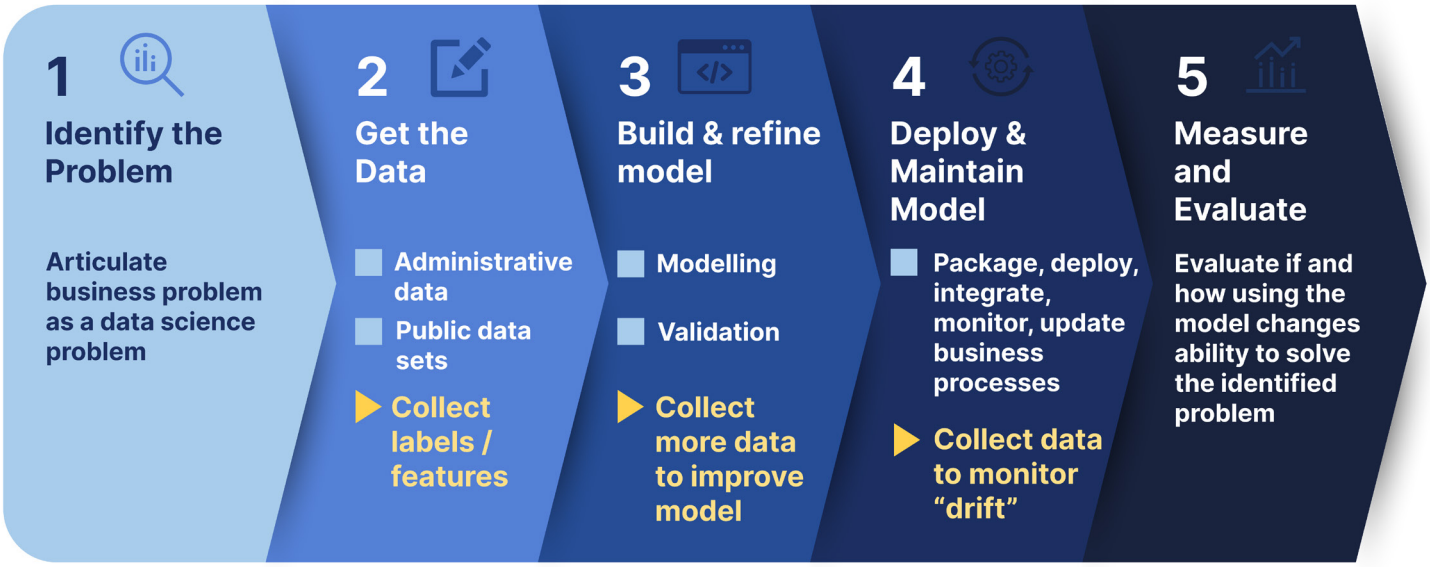
Putting People at the Center of Data-Driven Decisions

Government and social sector leaders often make high-stakes decisions that require prediction, for example, where to target programs to reach those most in need, or which entitlement programs a given household is eligible for. Machine learning (ML) models, a subset of Artificial Intelligence, can make these decisions more efficient. However, for decisions to be accurate, equitable, and unbiased, the models that underpin them must be based on high-quality data that truly represent the people the programs aim to serve.

Over the last five years, IDinsight has been evolving how it uses data science to support partners' decision-making. For example, IDinsight has worked with a non-governmental partner in India to build, test, and refine models to predict where out-of-school girls are to serve them more cost-effectively. We have provided training data for satellite imagery models predicting crop types and boundaries in India. We have worked with a government partner in Africa as they developed an ML model to predict households' ability to pay for contributions to a national insurance scheme. In each case, we strive to improve the reliability and equity of partners' models to inform decisions that directly affect people's lives and well-being.

For partners interested in using machine learning, IDinsight can support the entire life cycle of model development and deployment – from identifying the problem, developing, testing, and refining the model to evaluating the impact of its implementation (see Figure 1). This note focuses on how IDinsight's DataDelta and Data Science teams can help collect and use data to create and maintain equitable machine learning models (see ▶ yellow triangles in Figure 1).

Figure 1: Life cycle of using machine learning to improve impact

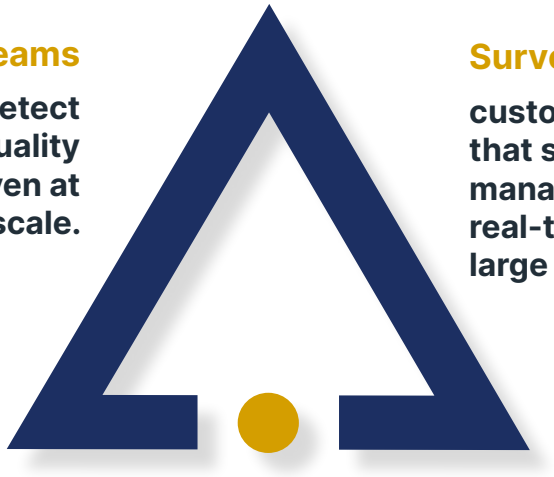


IDinsight's DataDelta delivers large-scale, high-quality, representative survey data and insights to government and social sector leaders when they need it, enabling more informed decisions to improve people's lives. DataDelta uses innovations in statistical methodology, data collection operations, and data science to **make it easier and faster to collect high-quality data directly from people, including underrepresented subgroups in the population**.

What sets DataDelta apart

Data collection teams

expertly trained to detect and correct data quality issues quickly, even at large scale.



SurveyStream

custom-built software that streamlines survey management and enables real-time quality control of large amounts of data.

Innovative sampling

approaches that allow for cost-effective representation of small geographic areas and hard-to-reach populations.

- **Tailored:** Each project is customized to ensure timely data generation tailored to specific decisions.
- **High Quality:** Innovative data quality techniques facilitate real-time feedback loops with field teams to maintain data integrity, even at scale.
- **Representative:** Innovative sampling approaches represent diverse populations that can be hard to reach.
- **Rapid turn-around:** DataDelta's pillars save time so we can quickly test, refine, and rapidly roll out new projects and get results.
- **Scale:** SurveyStream enables us to maintain quality while processing large amounts of data. Our largest survey captured 60,000 households.

IDinsight's Data Science team supports social sector leaders to better predict, optimize, or model important program outcomes to improve decisions and amplify social impact. All solutions built for our partners are tested for bias and performance with careful documentation of code such that solutions are maintainable by IDinsight, our partners, and any future collaborators.

Together, these teams help ensure that tech-enabled decisions are equitable, representative, and decision-relevant. This means that partners can build machine learning models tailored to the decisions they're making and validate and refine the models using data collected directly from the communities they serve.

IDinsight teams work with diverse partners, including government agencies, non-governmental organizations, international agencies, philanthropic institutions, and AI experts. In this document, we offer a glossary of ML terms to clarify terminology (see Box 5).

1. Get the Data

Machine Learning models often incorporate data from many different sources, such as official statistics or administrative datasets - census, facilities data, transaction data, etc. or use remote sensing (satellite imagery) to make predictions about an outcome, for example, people's income, type of crops, where demand for a given service would be highest, and so on. But sometimes, crucial information about the populations or places the model is designed to understand is missing, making them less accurate for decisions and less equitable. Missing data could be related to the variable you want to predict, such as income (label) or the variables you use to get an accurate prediction such as years of schooling (features). DataDelta can fill these data gaps.

DataDelta could conduct a comprehensive survey to collect all the features needed for a model or collect specific indicators for labels or additional features. Likewise, if partners know all the variables required for a model but only have them for a limited population, DataDelta could expand coverage with a new survey. For example, if a dataset is primarily in an urban area, but requires data on rural households to be accurate, DataDelta could collect data on the same variable from these additional geographies and households. Finally, DataDelta can add data layers for models relying on satellite imagery, such as collecting on-the-ground data about crops and plot boundaries to inform models that aim to predict these indicators (see Box 1).

BOX 1. Predicting crop types and boundaries in India

In 2022, the DataDelta team at IDinsight collaborated with RadiantEarth Foundation to carry out high-quality data collection of crop types and field boundaries (for ~ 3000 farmers and ~ 7000 plots) in four states in India. Radiant Earth used these data to create a benchmark training dataset and to develop a baseline model focused on field boundary detection. A competition was organized to classify crop types in agricultural fields across northern India using multi-spectral observations from Sentinel-2 satellite.

2. Build and refine the model

Once we have the data (labels and features), it is time to build and test ML models. While creating the model itself, we may discover a need for more data. For example, as part of the model training process, we may discover that the model is not making accurate predictions either overall or for sub-groups of interest such as women, certain geographies, or ethnicities. In this case, we have two options: make algorithmic adjustments or get more data to improve the model. IDinsight can determine the required solution and help with either or both.

Once we and our partners have a trained model that is sufficiently accurate - with acceptable error rates - it is essential to test the model, especially if the training data is not precisely representative of the population about which the model is making predictions. For example, if we train a model to predict income using data from three states in India, there is no guarantee that it will work well in a fourth state. In this case, we could use the model to predict the label (e.g. people with a certain income level) for the fourth state and then use DataDelta to collect ground-truth values to ensure the model is making accurate predictions.

BOX 2. Finding out-of-schools girls for Educate Girls

IDinsight developed a machine learning model to help Educate Girls find and cost-effectively serve more out-of-school girls. Educate Girls had ambitious plans to reach millions of children but could not practically expand its program to every one of India's 650,000 villages. They needed to prioritize villages to reach as many out-of-school girls as possible. No up-to-date, comprehensive, reliable data sources indicated where the most out-of-school girls were concentrated, and knocking on tens of millions of doors would be prohibitively costly. Hence, Educate Girls had a data-void problem that machine learning could help overcome. We generated village-level predictions using census data, DISE survey, and data from ASER and updated the model over time to allow Educate to serve 600,000 additional out-of-school girls on the same budget.

BOX 3. Estimating the ability to pay for insurance in Kenya

Kenya's Ministry of Health (MoH) is rolling out a new nationwide Social Health Insurance Fund requiring means-based contributions. This requires accurate predictions about household income and ability to pay, even for Kenyans in the informal sector for whom there are no formal records to determine income (like tax or employment records). In collaboration with key stakeholders, MoH is developing an algorithm to estimate households' ability to contribute. MoH has developed an algorithm for the means testing and IDinsight is partnering with MoH to validate the algorithm in sampled households, including through additional data collection. The aim is to correctly determine the income levels of households with a high confidence level by fine-tuning the algorithm with the best-fit predictors of income. This will ensure household incomes are accurately determined for millions of households in the informal sector such that families are asked to pay what they can truly afford for health insurance.

BOX 4. Improving the accuracy of gridded population maps

High-resolution population density maps are increasingly used to predict population in countries where census data is outdated, unreliable, or insufficiently granular. These maps are often created by sophisticated ML models combining satellite imagery and population census data. However, the accuracy of these maps is frequently poor or unknown, thus limiting their usage. Currently, we are conducting a project to understand the accuracy of gridded population density maps in India. We have rolled out a mini census in several 30X30m grids in urban areas of Telangana state. We are comparing the actual population with the ML predictions of population density in these areas. Early results indicate significant discrepancies between the population and the ML algorithm-predicted densities. This suggests that "ground-truthed" data could be important to increase the models' accuracy and thereby improve their utility for estimating population, which is critical for providing reliable sampling frames for large-scale surveys and targeting social services.

3. Deploy and maintain the model

Models built to predict labels accurately at a specific point in time begin to degrade in their predictive performance as real-world conditions change. Therefore, models must be updated with new data to keep them accurate over time.

It can be challenging to observe model accuracy through implementation without active data collection. For example, say hospitals in Kenya use a model to predict whether or not to administer HIV tests to patients. As doctors use this model to select patients for testing, they can gather data about HIV positivity rates from tested patients. This tells them the false positive rate of the model.

However, without active data collection, there is no way to track HIV positivity rates in patients for whom the model did not recommend a test; the false negative rate cannot be estimated. False negatives lead to exclusion errors. Measuring the false positive and false negative rates is important to understand model accuracy

Monitoring degrading model accuracy over time, or "drift," requires having a mechanism to collect features or labels regularly. DataDelta can collect data on a rapid turn-around, recurring basis to maintain the accuracy of models. We can quickly update sampling approaches, questionnaires, and survey operations as needed to ensure the data is of high quality, representative of target populations, and granular enough for appropriate geographic representativeness.

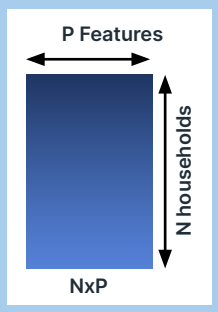
Please reach out to Sarah Lucas, Global Lead for DataDelta (sarah.lucas@idinsight.org) or Sid Ravinutala, Director, Data Science (sid.ravinutala@idinsight.org) if you are interested in working with IDinsight to get data, build and refine your models, or maintain the accuracy of your deployed models.

Machine Learning Basics

Labels

"Labels" are the variables we want to predict, e.g., the poverty rate at a sub-national level or the number of out-of-school children in a village. It can be a discrete variable, like "what crop is grown on the farm," or a continuous variable, like "the number of out-of-school children in the village". The terms "outcome variable" or "dependent variable" are sometimes also used for this, but we use "labels" in this document.

Features



"Features" are the variables we will use to predict the label. The variables are based on the level of prediction. Suppose we are predicting at a village level, for example, the number of out-of-school girls in a village. In that case, the features will be at the village level, such as the number of schools, proximity to roads, electricity access, percentage of the population engaged in agriculture, etc. If we are predicting at the individual level, such as income, then features will be at the individual level, such as level of education, primary language, or number of children. For models using imagery, features could be things like the intensity (amount of red, blue, or green) in each pixel for an image. The terms "predictors" or "independent variables" are also used for these, but we use "features" throughout this document.

Model

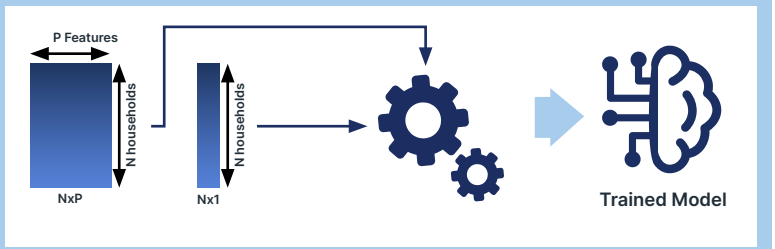
The "model" refers to the algorithm best suited to address the specific problem, question, or decision. It could range from linear regression (e.g., ordinary least squares) to Convolutional Neural Networks (used for image processing) to Transformers (used by GPT4).

Error

The "error" is the difference between the model's predictions and the actual ground truth. The mathematical definition is also called the "loss function," but we'll call it the error in this document.

Trained Model

A model has several parameters (think of these as knobs) that can be tuned to minimize the error in the model.



We feed a model features and labels, and it learns the optimal parameters to minimize the error. Once this learning process is over, you have a trained model. Trained models can be given a new set of features, and it can predict labels for it.

False positive and false negative

There are many different ways to evaluate the performance of a machine-learning model. *False positives and false negatives* are two important metrics for model evaluation.

False Positives are positive outcomes that the model predicted incorrectly. In standard statistical terms, they are also called Type I errors. In a hypothetical example where a machine learning model predicts cancer, a false positive prediction means that patients predicted to have cancer were healthy.

False Negatives are negative outcomes that the model predicted incorrectly. This is also known as Type II error. In our hypothetical example, this means that patients who were predicted to be healthy had cancer.

About IDinsight

IDinsight uses data and evidence to help leaders combat poverty worldwide. Our collaborations deploy a large analytical toolkit to help clients design better policies, rigorously test what works, and use evidence to implement effectively at scale. We place special emphasis on using the right tool for the right question, and tailor our rigorous methods to the real-world constraints of decision-makers.

IDinsight works with governments, foundations, NGOs, multilaterals and businesses across Africa and Asia.

We work in all major sectors including health, education, agriculture, governance, digital ID, financial access, and sanitation.

We have offices in Dakar, Lusaka, Manila, Nairobi, New Delhi, Rabat, and Remote.

Visit www.idinsight.org and follow on Twitter @IDinsight to learn more.