Vision Statement

Some of the most pressing sustainability challenges of our times, including climate change, poverty mitigation, and food security involve global phenomena that are unique in scale and complexity [24, 26]. Successful quantitative approaches to these global problems will require an ability to gather and model global data while simultaneously being able to capture the rich temporal and spatial heterogeneity in responses that can occur locally. Remote sensing is perhaps the only technology providing data at a global scale at reasonable costs, that also has the temporal and spatial resolution to quantify local-scale developments. Thanks to recent technological developments, remote sensing is becoming increasingly more economical and accurate. In ten years, it is expected that commercial services will provide sub-meter resolutions of any area in the world at a fraction of the current cost [22]. In principle, this type of data contains a wealth of information, including details on economic activities, land cover changes, and urbanization. Unfortunately, this information cannot be easily extracted from the streams of noisy raw digital data produced by satellites. The sheer size and the unstructured nature of the data precludes traditional analysis techniques. Big Data computational techniques, on the other hand, have a successful track record in dealing with these challenges [6, 15]. Machine learning techniques, for example, have been shown to be extremely effective at extracting relevant information from large amounts of raw data, including in genetics [2], astronomy [4], and particle physics [3]. We believe that research bridging sustainability science, remote sensing and machine learning will open the way for entirely new approaches for modeling and understanding global scale phenomena involving the balancing of economic, societal and environmental needs of present and future generations.

In the past decades, many developing countries have witnessed dramatic improvements in some key indicators of economic welfare. Sustainable development, however, has remained elusive for many African countries [19]. Poverty is still a major challenge in Sub-Saharan Africa, where countries face persistent food insecurity and millions of people face regular risks due to limited resource availability, uncertainty and armed conflicts, as well as potential shocks from droughts, disease, and climate change. Developing a deeper understanding of the economic, societal and environmental drivers of poverty and contributing to its reduction is one of the grand challenges of sustainability science [19]. To this end, we can identify three main sub-challenges: 1) describing and mapping poverty and human livelihoods, 2) understanding its driving forces, and 3) providing policies and solutions. Successful approaches will require contributions from multiple disciplines, and we believe computational techniques and models can and should play a role.

While there is a vast scientific literature on these topics, most models and analyses rely on census and survey data, collected through small scale field surveys or provided by government agencies. In the developing world, however, institutional data is often scarce and when available, it does not exist at village-level granularity and at sufficiently fine-grained time scales. These limitations have severely limited researchers’ ability to both characterize the distribution of poverty over time and space, as well as to understand what drives this variation. Analyses are often limited to annual data at the country level, again based on government figures that are often thought untrustworthy. Remote sensing data, on the other hand, is available at very high spatial and temporal resolutions, and provides a wealth of information on human and natural capital. Unfortunately, such data is also extremely unstructured and noisy, largely preventing its use in mapping and modeling human livelihoods. Efficient computational methods are therefore needed for transforming large amounts of unstructured data into actionable insights.
Modern statistical **machine learning** approaches have been quite effective in extracting relevant information from large amounts of raw data [15]. For example, machine-learning techniques have recently shown enormous progress in tasks such as object recognition from images [5, 18, 21]. Even though similar tasks are key for the interpretation of satellite imagery, so far there has been little interaction between the two communities. We propose to fill this gap by developing machine learning techniques for the extraction of features relevant for predicting socio-economic outcomes such as well-being, assets and mortality indexes from remote sensing data. This will require the development of new techniques to deal with challenges that are unique to remote sensing data, such as hyperspectral and multiresolution data. A successful approach will also need to explicitly capture the spatial and dynamic nature of the problem. While this spatio-temporal structure can be in principle captured using **graphical models** formalisms [20], the scale is unprecedented and will require the development of new models and efficient and scalable inference and learning techniques.

We propose the formation of an innovative interdisciplinary research group to advance our understanding of the environmental and economic drivers of poverty in the developing world through **data-driven** models and the development of new **computational lenses**. We propose research bridging computer science, remote sensing, and economics, centered on the development of new machine learning methods for the automated construction of large scale poverty maps from unstructured data, including remote sensing, weather data, and surveys. The models will track both **spatial** and **temporal** variations of wellbeing, thus offering policymakers a better tool to target assistance where it is needed, and researchers a better platform to study the macro- and micro-determinants of human livelihoods. This project has the potential to provide broadly applicable methodologies for developing large scale spatio-temporal models, including in ecology and conservation planning.

**Background and Significance**

Sub-Saharan Africa lags substantially behind the rest of the world on most livelihood indicators: its people are more likely to live in poverty, its children are more likely to die at a young age, and its households are less likely to have secure access to food. This overall poor performance masks substantial heterogeneity at both the national and sub-national level. For instance, at the national level, the percentage of people estimated to be living on under $1.25 per day (the current World Bank definition of poverty) ranges from a high of roughly 80% in the Democratic Republic of the Congo to below 10% in countries such as Botswana. Similarly, over 16% of children do not live past their 5th birthday in Angola, while only 5% die before age five in Namibia and Eritrea. [14] Subnational variation, while much more poorly characterized, is probably just as large or larger. For instance, under-5 mortality is three times higher in rural western Kenya than in more urban central Kenya, and is more than twice as high in parts of northern Nigeria than in parts of southern Nigeria [7, 23].

What explains this remarkable variation? The answer is we don’t really know, and a primary reason for this is that despite the statistics just presented, variation in livelihoods over space and time remains remarkably poorly characterized in Africa. This presents a basic impediment to
understanding why some countries or locations might be doing better than others, and how to improve outcomes in regions that are lagging. Local-level data on economic output or on poverty are essentially unavailable for Africa, and the more aggregate estimates that are available (e.g. at the country level) are commonly thought to be highly untrustworthy [8]. Data that are thought to perhaps be more trustworthy - e.g. the national household surveys used to construct the poverty measures cited above - are only available in a subset of countries, and even in those countries only once or twice a decade. This leaves researchers with disturbingly little firepower to understand African poverty and its determinants, and leaves policymakers with little way to target assistance to the most needy or to understand whether chosen policies were effective. Put more broadly, understanding the economic and environmental roots of poverty – one of the grandest challenges of sustainability science, and arguably one of the most important questions facing mankind – is also one on which we have some of the worst data.

Our proposed work aims to fill this data void by harnessing the wealth of new remote-sensing data that is becoming available, and by developing new computational tools to make sense of it. Our goals are to provide (i) a better understanding of how livelihoods have evolved over space and time in Africa, and (ii) a tool to help understand why things have evolved as they have. These goals are intertwined, in the sense that our proposed approach – mapping the evolution of livelihoods through the development of large-scale remote-sensing-based spatio-temporal models – will also necessarily shed light on what features in these data (e.g. transport infrastructure, natural resources such as water or climate, farmland productivity, etc) are important drivers of livelihood improvements.

We believe the models we develop will be extremely valuable from a scientific and practical perspective. From a scientific perspective, these new models will be used to identify which features are most relevant, providing new insights into the leading causes of poverty. From a practical perspective, such models can be used by both government and non-government organizations to plan their efforts in a more targeted way. For example, an NGO such as GiveDirectly needs to identify regions of the country with highest poverty rates before starting a donor campaign. A fine-grained poverty model can improve the effectiveness of a campaign and significantly reduce the costs of identifying suitable targets, for example reducing logistic costs.

The models we develop can also be used to perform what if simulations. For example, we can simulate poverty dynamics under new circumstances such as changes in resource availability, climate, or new policy interventions. This approach will be extremely useful to support decision support tools, including decision making under uncertainty frameworks. For example, one could estimate costs for possible interventions and automatically identify the most effective measures that are compatible with given financial budget.

**Research Plan**

We propose to use large-scale data driven models to characterize poverty and its determinants at two scales that are relevant to both research and policy: the country scale and the village scale. Existing data at these scales are either poor (country scale) or incredibly sparse (village scale), hampering efforts to understand either the macro trends or micro-determinants of poverty. We propose research bridging computer science, remote sensing, and economics, centered on the development of new machine learning methods for the automated construction of data-driven large scale poverty maps, leveraging remote sensing, weather data, and survey data.
1 Data

Our proposed work utilizes two main types of data: high temporal and spatial resolution remote sensing data, and sparser ground-based datasets on socio-economic outcomes that we will use as training data. We plan to use the following six sources of remote-sensed data:

1. High resolution imagery in Google Earth, which can be accessed through the Google Earth API. The resolution of these images are typically high enough to observe relevant objects that impact the economy like rooftops, roads, terrain, barns, and vehicles.

2. Weekly imagery from Skybox and Planet Labs for selected sites. Both are start-up companies in the Bay Area operating low-cost, high resolution satellite sensors with high temporal frequency, and Co-I Burke and Lobell have established relationships with both companies. The higher frequency could improve estimates by providing indicators of assets that move around (e.g., vehicles or livestock) or by better characterizing dynamics of agricultural crops.

3. Digital Globe imagery with $\sim 1m$ resolution is available for many sites throughout Africa, and we have obtained these in the past for free through the Digital Globe research foundation.

4. Landsat provides 15m resolution pan-sharpened data for free, with both Landsat 7 and 8 currently providing 16-day repeat observations for any place in the world. Although coarser, these data could still be useful for identifying many relevant landscape features, such as productivity of crops or roads.

5. Satellite-derived data on night-time lights, available at 1km globally since the early 1990s from the United States Air Force Defense Meteorological Satellite Program. Existing work suggests that nightlight intensity can improve existing (ground-based) measures of economic productivity at the country level [17], but its ability to provide more localized data on economic activity is probably even higher – and currently unexplored.

6. Finally, we will use a combination of satellite-derived and ground-based weather measurements. These include Global Precipitation Climatology Center, NOAA, and the growing Trans-African Hydro Meteorological Observatory (TAHMO) weather data repository. The TAHMO project is planning to develop a dense network of hydro-meteorological monitoring stations in sub-Saharan Africa, with a station every 30 kilometers.

Our ground based datasets used to train the models (as described below) will also come from a variety of sources. Three main sources of data will give us temporal and spatial variation to train with at a local level:

1. Demographic and Health Surveys (DHS). These data will constitute our primary village-level training dataset. The DHS are large, nationally representative surveys mainly focusing on health outcomes. As shown in the figure, the DHS cover a large swath of Africa, surveying a random sample of roughly 25 households per village in over 29,000 villages across the continent. Our main outcome of interest in these surveys is the village-level rate of child mortality. Child mortality is thought to be a summary measure of socio-economic well-being, and is the DHS, with each interviewed mother providing the history (and survival) of each of her births. This allows us to construct a village-level time series of child mortality, giving us both spatial and temporal variation in this key outcome at a local level.
2. Living Standards Measurement Surveys (LSMS): nationally representative surveys conducted by the World Bank, based on smaller samples sizes and conducted more infrequently than the DHS but providing more detailed economic information on livelihoods, including consumption and income. LSMS surveys in a growing number of countries contain a panel element (i.e. interview the same households over time), and we will focus on these surveys and their ability to give a detailed local picture of changes in poverty over time.

3. Various (single or multi-year) surveys from field experiments and other research projects. The coverage is sporadic and the samples sizes are often small, but many provide information on livelihoods. Datasets in this category that we will use include geo-referenced data on assets (e.g. dwelling size, roof type) that we and collaborators have collected in Kenya and Uganda, and detailed data we have collected on crop yields in Kenya, Uganda, and Rwanda.

Finally, because economic output at a national level is perhaps the key indicator of interest for policymakers – and one that is thought to be measured with substantial error in Africa – we will also use our models to try to improve upon existing national-level estimates of economic output over space and time. The “training” data available at the national level are annual time series at the country level that are easily available from international financial institutions such as the World Bank.

2 Automatic Feature Extraction

The first stage of the project will involve the development of a machine learning pipeline to extract relevant features from the raw data. This will involve a combination of image processing and machine learning techniques. For example, in Kenya and Uganda identifying and counting houses with metal rooftops (as opposed to thatched ones) provides a strong signal on the wellbeing of the households and villages [1]. Another example of features we will consider is the presence and intensity of lights during night hours, as access to electricity correlates strongly with economic wellbeing. Other features we plan to extract involve human infrastructure, e.g. the presence and quality of roads, and features of the natural landscape, such as the presence of water resources (rivers, lakes, etc.), elevation and terrain type. From the point of view of economic activities, we also plan to identify crops and estimate their productivity (Figure 2). For most agricultural crops – particularly the staple crops grown by African smallholders, such as maize – high productive crops crop can be distinguished from low-yielding crops based on color. Healthy vegetation reflects light differently than less-healthy vegetation. Sensors on satellites can discern these differences in the visual wavelengths as well as at other wavelengths, providing additional signal for crop yield estimation. Co-I Lobell has extensive experience on designing appropriate yield metrics, as illustrated in his biographical sketch.

Developing an accurate and scalable feature extraction phase presents numerous computational challenges. Traditional techniques from the computer vision literature cannot be directly
applied to remote sensing imagery data, and will have to be adapted to multispectral and multiresolution data. While there is little variability in the images due to changes in viewpoint, our techniques need to be robust to clouds and changing illumination and weather conditions, which are the major sources of variability in satellite imagery. Automatic feature extraction using deep learning techniques is at the core of the recent dramatic improvements on many computer vision benchmarks [5, 18, 21]. These techniques have the potential to have a major impact in the automatic extraction of features for remote sensing data. Unfortunately, deep learning techniques shine in a regime where training data is abundant, and it is not uncommon to use millions or even billions of training examples. Successful application of these techniques in our data regime, where unlabeled data is abundant but there is little training data, will likely require revisiting and adapting unsupervised pre-training techniques [27]. Techniques we will develop will be applicable beyond this domain and will be useful for other remote imagery classification tasks, including land cover classification, tracking deforestation, human development and desertification.

This machine learning pipeline will be evaluated and validated using hand-labeled data and survey data collected on the ground. For example, we have access to a small set of images where rooftops have been manually identified and classified. If necessary, we will generate more training data using crowdsourcing platforms, e.g. Amazon Mechanical Turk [10]. PIs Marshall Burke and David Lobell are collecting ground-based yield data in Kenya, Uganda and Rwanda that will be used to train and evaluate the yield models.

Figure 2: Field boundaries in Western Kenya, as mapped by the authors.

3 Fine-grained spatial and temporal modeling

As a first step, the features extracted will be used to train a predictive model of economic well-being for individual villages and other regions of interest. The model will be trained and evaluated (e.g., in cross-validation) using a set of ground truth estimates as summarized above. We will focus on three outcomes that are relevant to wellbeing and where sufficient ground data exists to train the model: child mortality at the village scale, and economic output (GDP) and expenditure-based poverty measures at the national level.

To further improve the accuracy of the model, we propose to construct a more complex spatio-temporal statistical model to jointly describe socioeconomic outcomes of an entire region of interest over time. Rather than classifying separately each individual land patch or village (based on the extracted features), we propose a model which explicitly takes into account spatial and temporal relationships. For example, we expect a village’s wellbeing to change slowly over time. These temporal dependencies (dynamics) can be included into the model to improve the accuracy. This will be especially useful in our setting, where features extracted with our machine learning pipeline will inherently be noisy. For example, no useful visual features can be extracted from a satellite image occluded by clouds. The temporal model we plan to construct will automatically take
into account all the snapshots available for the same region, propagating the relevant information across time steps. Similarly, we plan to explicitly model spatial correlations, as we expect the distribution to be relatively smooth at a sufficiently fine grained scale.

The spatio-temporal nature of the problem can be naturally captured using graphical models formalisms, a flexible machine learning framework for modeling large collections of random variables, where complex global behavior emerges from simpler local (spatial and temporal) interactions [20]. In Figure 3, we show a prototype graphical model capturing both spatial interactions between neighboring patches and temporal dependencies. These dependencies can either be incorporated based on prior domain knowledge or learned from data, given enough samples.

From a computational perspective, developing a spatio-temporal model with high spatial resolution is extremely challenging. Adding temporal structure is required to track changes over time, but makes model even more intractable. In fact, the resulting model is going to be very high dimensional. Furthermore, the models we envision are hybrid, i.e. they involve a combination of discrete and continuous variables. Such high-dimensional models are notoriously difficult to train and to do inference with [25]. The resulting model will have high-treewidth, precluding exact inference and learning algorithms. New, efficient approximate probabilistic inference algorithms will be required to train the models and make accurate predictions to track poverty dynamics over time. These algorithms will likely involve a combination of sampling and variational approaches, which will have to be tailored to the specific application domain. The computational efficiency of these algorithms will dictate the level of spatial and temporal granularity we can achieve for a given region of interest. Some of the new approximate inference techniques that we recently introduced will likely be useful to overcome these computational challenges [11, 12]. In our new probabilistic inference approach, we compute statistics of interest by considering only a small set of representative states from the original high-dimensional statistical model, like in a sampling strategy. However, these states are not drawn at random from the underlying probability distribution by simulating a Markov Chain, rather they are very particular states that can be discovered using state-of-the-art optimization tools and random projections. Quite surprisingly, we can show that a small collection of such extreme states is with high probability representative of the overall probability distribution and can be used to answer a range of queries about the original statistical model with provable accuracy guarantees. Furthermore, our approach can directly leverage state of the art combinatorial optimization techniques, such as mixed integer programming (MIP). These techniques are promising but will have to be adapted to the poverty mapping domain, for example to handle hybrid models with continuous random variables.

We will also develop nested hierarchical models, where the region of interest is modeled at different spatial granularities, for example at the district, county and village level. Imposing a hierarchical tree-like structure will simplify the interactions between the random variables, reducing the complexity of the model, e.g., ignoring dependencies between patches from different counties or
districts. We can thus achieve a tradeoff between computational complexity and expressive power of the model. Another benefit of a hierarchical model is that it will be easier to validate predictions at an aggregate level (e.g., at the district granularity) using estimates from survey data, which tend to be more accurate because of the larger sample sizes.

4 Model interpretation and policy implications

The model described above will provide an unprecedented ability to predict outcomes of mortality and poverty from readily available sources (satellites, weather, etc.). These predictions in themselves will be useful for achieving sustainability goals, for instance by helping a NGO to target their efforts to areas that are most in need and by tracking progress over time. But the model and predictions could also be used to develop hypotheses and insights into the mechanisms by which these outcomes improve.

We plan to do this in a few ways. First, sensitivity tests can reveal which factors are most important in the model. For example, if outcomes are much more sensitive to a change in night lights than roof type, then it suggests (but does not prove) that rural electrification may be more important than other forms of infrastructure. Or conversely if night lights are unimportant, it suggests this factor may be less critical than some have argued. Second, we will compare the cross-validated accuracy for models with and without certain factors. For example, if excluding weather variables does not reduce model performance, this implies that either weather is relatively unimportant for the outcomes, or that it acts via impacts on other factors in the model – a possibility we can in turn evaluate. Third, as a measure of intervention efficiency, we will use the model to explore scenarios of applying a fixed budget to different types of interventions, such as improving roads vs. electrification. None of these will be able to prove causality, given the colinearity of different variables and the likely feedbacks between outcomes and predictors. However, these exercises will help to generate hypotheses that can then be tested in the future, such as by using randomized control field trials.

5 Research Team

The team combines expertise in computer science and machine learning, economics and sustainable development, and remote sensing data analysis. Stefano Ermon is an expert in machine learning and statistical modeling of data, and has considerable experience in applying machine learning techniques to novel sustainability domains. He is a fellow of the Woods Institute for the Environment at Stanford. Marshall Burke is an expert in the study of global scale phenomena involving the relationship between society and the environment using statistical analyses of real world data. He has considerable previous experience working in Africa and with datasets such as Demographic and Health Surveys (DHS). David Lobell is an expert in remote sensing, GIS, and crop and climate models. He is the Deputy Director of the Center on Food Security and the Environment at Stanford University. Burke and Lobell have already a partnership in place with Skybox, and ongoing discussions with Planet Labs, relationships that we will leverage for this project. All PIs are located at Stanford University, enabling frequent face-to-face communication which will simplify collaboration efforts.