Bhoomi

U.C. Berkeley School of Information

Master's of Information Management and Science (MIMS)

Capstone Project

May 5, 2017

Team Members:

Cristián Garay

Brian Goodness

Eric Hagen

Deepa Kalpathi

Surendran Subbiah

Project Advisor:

Joshua Blumenstock

Table of Contents

Acknowledgements

The Bhoomi team would like to thank all individuals we worked with over the course of this study for the time, effort and support they placed in our project and our team.

We would like to thank our advisor, Professor Joshua Blumenstock, for his invaluable support and advice throughout the project.

1. Abstract

Information scarcity is a real problem faced by decision makers working in fields related to International Development. The cost of conducting regular censuses is often prohibitively high, and governments in many developing nations are thus unable to maintain current records of many metrics that are essential to informing international development and international aid. This project seeks to address this acute information deficiency. By leveraging the power of machine learning techniques on satellite imagery, a team of researchers led by Neal Jean has demonstrated an ability to make predictions related to poverty in certain Sub Saharan African nations[1]. Our group has worked to reproduce the Jean team's machine learning efforts, and to build an interactive website that will make this information more accessible to stakeholders in the International Development community. Furthermore, we have worked to establish a pipeline that would make this process easily replicable for those interested in reproducing these results for other nations.

[1] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

2.Introduction

2.1 Problem Statement

"In a world increasingly awash with data, it is shocking how little is known about some people and some parts of our environment." This statement, taken from a United Nations report[2] about sustainable development, served to guide our team through the process of this capstone project. Living as graduate students at UC Berkeley, it is exceedingly easy to become inoculated against the idea that there exists, anywhere, a scarcity of information. However, there are many parts of the world in which reliable information regarding essential aspects of a population's well-being are either incomplete or altogether missing. Indeed, "indigenous populations and slum dwellers are...consistently left out of datasets"[3].

In 2015, the United Nations (UN) adopted 17 'sustainable development goals.' The three encompassing themes of these development goals are as follows: End Poverty; Protect the Planet; and Ensure Prosperity for All[4].

---

[2] Revolution, U. D. (2014). A World That Counts: Mobilising the Data Revolution for Sustainable Development, United Nations Independent Expert Advisory Group on a Data Revolution for Sustainable Development.

[3] Revolution, U. D. (2014). A World That Counts: Mobilising the Data Revolution for Sustainable Development, United Nations Independent Expert Advisory Group on a Data Revolution for Sustainable Development.

[4] Sustainable Development Goals. Retrieved from http://www.un.org/sustainabledevelopment/sustainable-development-goals/

Figure 1 - UN Sustainable Development Goals

The members of our project group, to a person, have an interest in the developing world. These goals served as a starting point for our exploration of the development space, and our inquiry into parts of this space where we believed we could contribute to addressing information shortcomings.

In order to address these goals, it is essential that those involved in international development have access to trustworthy information about the conditions on the ground in developing countries. Unfortunately, in many cases, the cost of acquiring this information is prohibitively high[5]. Conducting a nationwide census requires resources that many developing countries cannot spare. Thus, there is a great need for alternative sources of information to 'stand in' for these lacking data sources. Our problem statement: how can we address information shortcomings in a manner that allows for a better understanding of economic conditions in the developing world, especially in impoverished areas.

[5] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

2.2 Project Description

Coalescing around a clear, defined, and attainable project turned out to be a thoroughly challenging exercise. For example, at the time of preliminary presentation in January, our project was focused around the concept of using satellite imagery to build a tool that could be used remotely by managers of infrastructure projects that were being carried out in rural areas. The idea was to reduce corruption and improve the transparency of projects in which proper monitoring might be inconvenient or prohibitively expensive. Ideas such as this one dominated the early weeks of our meetings and discussions.

After some time it became clear that while many of our ideas were novel and even interesting, we had serious concerns about the feasibility of their implementation, especially considering the project timeframe. However, most of our team had been exposed (through our advisor's class) to the work of a team of researchers led by Neal Jean. This team had recently published a paper in *Science* magazine titled "Combining satellite imagery and machine learning to predict poverty."[6] The novel concept behind their work involved the use of daylight satellite imagery in order to predict the wealth of areas in Sub Saharan Africa. Much work had been done previously to demonstrate the ability of nightlight satellite imagery in understanding variation in wealth between different regions. However, a shortcoming of nightlights imagery was found in its inability to differentiate between the most extreme levels of poverty. Jean's team addressed this shortfall, essentially, by deploying a pre-trained convolutional neural network on daylight satellite imagery in order to extract features found in the daylight imagery that can be related to the intensity of nightlight luminosity.

The work of the Jean team addressed an area of central concern to our team - addressing the issue of poverty in the developing world. Our project goal became reproduce and potentially improve upon the machine learning aspect of Jean's work, while constructing an information architecture centered on a data pipeline that would allow the work of Jean et al. to be replicated on any country for which data is available (discussion of 'data' can be found in section 5). Furthermore, we would develop an interactive tool so that this potential solution to a pressing information shortcoming could be presented to the development community at large.

---

[6] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

This report will detail the efforts made to achieve our project goals.

3. Background

3.1 Motivation

Planning for developmental projects involves a top down, bottom up or a hybrid approach. A top down approach is used when budgets have been allocated to certain projects and these need to be executed[7]. For example: $50 million is allocated towards building schools in Sub Saharan Africa. The project planning team would determine the number of schools the budget would entail and the specific locations for the schools.

A bottom up approach is built from the ground up; starting with an assessment of the needs and of the people, determining the resources required and the goals, creating an outline of the plan, preparing a detailed formulation of sectoral programs, regional and local plans, drawing up the complete plan, getting it approved by a chief executive and then building up to budget requests and approvals[8].

Many projects also follow a hybrid approach, in which the budget for the overall goal or objective is pre-determined but the local teams decide the allocation to various projects based on regional need[9].

Our project aims to support the decision making process, especially in top down or hybrid approaches by arming the decision makers (policy makers, project managers, etc.) with centrally available accurate data to help plan projects with greater granularity.

---

[7] Based on in depth interview
[8] Nath, V. (2011). *Administration and Development Planning in India*. Concept Publishing Company.
[9] Based on in depth interview

3.2 Existing Landscape

Agencies use different means of targeting to deliver benefits to different groups of people. There are three broad approaches to targeting - Household or individual targeting, Categorical targeting and self selection targeting[10].

3.2.1 Household or Individual Targeting[11]

With household assessment, an official (usually a government employee) directly assesses, household by household, whether the applicant is eligible for the specific program(s). Unified household targeting systems are often designed to serve multiple social programs. While the actual design and implementation of household targeting systems varies significantly by country, most systems involve the following basic steps:

- Collecting data on specific (potentially eligible) households via interviews (and sometimes home visits) using pre-designed questionnaires (which depend on the type of household assessment mechanism);
- Entering these data into a unified household information registry (with varying degrees of verification and consistency checks);
- Comparing household characteristics with pre-established eligibility criteria (program-specific); and
- Establishing program-specific beneficiary lists (sub-registries) for the purposes of program implementation and payroll.

Household or individual targeting is often used when the beneficiary is the individual. For example: Cash transfers, food stamps or scholarships. This targeting method is not used when the beneficiary is the entire district or region such as for infrastructure projects such as construction of roads, provision for water, etc[12].

---

[10] Design and Implementation: Targeting & Selection (2013, Jun 6). Retrieved from http://www.worldbank.org/en/topic/communitydrivendevelopment/brief/cdd-targeting-selection

[11] Castañeda, T., Lindert, K., de la Brière, B., Fernandez, L., Hubert, C., Larrañaga, O., ... & Viquez, R. (2005). Designing and implementing household targeting systems: lessons from Latin America and the United States. *World Bank Social Protection Discussion Paper Series*, *526*.

[12] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.

Household targeting is expensive since it relies on income or poverty surveys in order to develop robust measures and also to measure the impact of such targeting. Often, proxy measures such as level of education, number of children in the family or housing quality are used to determine a household's eligibility[13].

3.2.2 Categorical Targeting

Categorical targeting refers to selecting broad groups of households, based on a common characteristic, such as geographic location[14]. Characteristics are usually chosen for their correlation with poverty levels. Geographical targeting belongs to this class of targeting method, with place of residence a proxy for poverty level.

Tools used to guide geographic targeting decisions include[15]:

- geographic maps of poverty and inequality (such as those which combine household survey and census data);
- geographic data on particular social indicators (e.g., educational outcomes, Human Development Index);
- geographic data on coverage of specific infrastructure (e.g., water systems, health coverage, school systems); and/or
- geographic mappings of basic needs indices.

The advantages of geographic targeting[16] include its simplicity. If reliable data by way of maps are available, then these can be used for targeting. The process is transparent and minimizes political interference by way of preferred regions. In case of developmental projects that are allocated on a first-come first-served basis, or those that are quick to organize and implement the projects; this method of targeting ensures that the resources are reserved for those in need

---

[13] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.
[14] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.
[15] Castañeda, T., Lindert, K., de la Brière, B., Fernandez, L., Hubert, C., Larrañaga, O., ... & Viquez, R. (2005). Designing and implementing household targeting systems: lessons from Latin America and the United States. *World Bank Social Protection Discussion Paper Series*, *526*.
[16] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.

but that take time to organize and carry out planning. Geographic targeting is often combined with household targeting.

The limitations of geographic targeting include its information requirements (accurate, current, and spatially disaggregated data on living conditions), its weak performance when poverty is not spatially correlated, and the potential for political controversy of including and excluding certain areas[17].

Geographic targeting is particularly appropriate in circumstances where[18] considerable variations exist in living conditions across regions, administrative capacity is sufficiently limited so as to preclude use of individual or household assessment, and/or delivery of the intervention will use a fixed site, such as a school, clinic, or ration shop.

3.2.3 Self-selection Targeting

Self selection relies on the potential beneficiaries, such as the poor, self selecting themselves into the developmental programs. Some of the means for this are by offering below market wages, which would deter the non poor from participation or specifying eligibility requirements[19].

In a study on the efficacy of self targeting vs. other methods such as automatic screening, it was seen that while self-selection does improve targeting, there are still exclusion errors[20].

3.3 Challenges in Geographic Targeting

Of the various targeting methods, Geographic Targeting is most popular[21]. One of the key inputs for geographic targeting is the availability of data. Household surveys are seen as one of

---

[17] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.
[18] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.
[19] Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. *Network HD. Washington DC: World Bank*, 1-52.
[20] Alatas, V., Purnamasari, R., Wai-Poi, M., Banerjee, A., Olken, B. A., & Hanna, R. (2016). Self-targeting: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, *124*(2), 371-427.
[21] Kumar, N. (2005). *The Effectiveness of World Bank Support for Community-Based and-Driven Development: An OED Evaluation*. World Bank Publications.

the major tools to collect data on income and poverty. However, these surveys are considered deficient in many ways, especially in developing countries. Some of the problems are[22]: inappropriately defined concepts - of household and income - often lead to inconsistent data collected, outdated or inappropriate sampling frames and lack of resources especially trained manpower, leading to poor data being collected. In addition, bottlenecks in software, hardware and personnel add to the decreased quality of data.

One of the ways to overcome these challenges in developing countries is to use data from standardized surveys, if available. Some of the standardized surveys are country census data - periodicity and coverage varies by country, Demographic and Health surveys (DHS surveys) - that is a part of the DHS program and Living Standards Measurement Study - that is a part of the Survey unit of World Bank's Development Data Group.

Even for standardized surveys, there has been skepticism with respect to its sampling and data [23].

While quality of this data will be high, policy makers and project planners would still be reliant on data that is updated only sporadically. For instance, the last two cycles of the DHS data for Rwanda are 2010 and 2014-15.

3.4 Supplementing data for Geographic Targeting

Advances in computing and easy access to satellite imagery has resulted in remote sensing data being used for a variety of purposes[24]. Applications of remote sensing data in economics include use of night lights data, climate and weather data, topography, agricultural land use and crop choice, urban land use, crop choice, natural resources, pollution monitoring and a combination of these sources[25].

---

[22] Chander, R., Grootaert, C., & Pyatt, G. (1980). *Living standards surveys in developing countries* (No. 1). World Bank, Development Research Center.
[23] Jerven, M. (2015). Measuring African development: past and present.
[24] Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *The Journal of Economic Perspectives*, *30*(4), 171-198.
[25] Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *The Journal of Economic Perspectives*, *30*(4), 171-198.

We aim to show that by supplementing survey data with remote sensing data, specifically night lights data, granular estimates of poverty can be achieved.

4. Research & Methodology

4.1 Overview

We employed both secondary and primary research as a part of our methodology. The objective of secondary research was to gain an understanding of the landscape of developmental projects. The objective of primary research was understanding the on ground planning and implementation of developmental projects. Our goals were as follows:

- To understand the challenges faced by policy makers and project planners with respect to projects in the developing world
- To study the tools for planning used currently for such projects
- To understand the audience for this data and how they used it

4.2 Methodology

4.2.1 Secondary Research

In order to understand the landscape of developmental projects, we researched objectives, projects and processes of organizations such as World Bank, International Monetary Fund; as well as NGOs that work on infrastructure development in developing nations such as Room to Read and M.S. Swaminathan Research Foundation.

4.2.2 Competitive Analysis

The existing tool being used is Excel. This tool is popular due to familiarity with and universality of the tool. Excel is used for both planning as well as ongoing monitoring.

One example of Excel use is from Room to Read. Room to Read builds schools in developing nations, and use Excel to track the progress of every project, along with photographs of the actual project. They also use Excel to plan site selection for school buildings.

Another example of using Excel is combining survey data with local field level information and then exploring the data using charts to identify trends.

4.2.3 In Depth Interviews

To understand the planning and execution of developmental projects, we spoke to representatives from International Finance Corporation, Room to Read, M.S. Swaminathan Research Center, United Nations and Association for India's development. The discussion guide used for these interviews is in Appendix A.

We also spoke to experts in the field of Remote sensing to understanding the applications of remote sensing data.

4.2.4 Contextual Inquiry

To understand the tools currently being used in these projects, we also conducted contextual inquiries with a few stakeholders in this field. We also built upon the experience of a team member who was involved in project planning for a developmental research organization in India.

The contextual inquiries were around the use of excel for analysis. As an analogy to using a web tool to retrieve information, we also used the website Global Forest Watch (Link: http://www.globalforestwatch.org/), specifically the country profiles, to observe how users interacted with the website.

4.2.5 Usability Testing

We used usability testing on our paper prototype, low fidelity wireframe as well as the site mockup. While the paper prototype and the low fidelity prototype was used to test for the overall design and understanding of the data; the site mockup was used for user preferences such as color of the background, etc.

4.3 Findings

Some of our key findings are:

- An opportunistic approach to projects is being used, in that the availability of data determines which project or region gets funded
- Maps as a feature are extremely important for project planning purposes
- Since the project planners and policy makers often use multiple sources of data, there must be options to either add user input data onto the map, or allow the data to be downloaded so that it can be used for analysis along with other sources of data.
- Existing tools are insufficient in providing the granularity required
- Use of satellite images is considered interesting, but has not been validated as an approach
- Project monitoring is still heavily dependent on manual intervention
- One of the features that planners would find most useful is the ability to select a part of the data, and then download that data, for further analysis or layering with other other data sources they would use.

4.4 Personas

The in depth interviews and contextual inquiries led to the creation of the following personas, one of which was used as a primary persona for the development of the prototype.

Persona 1 - Nick Anderson, 45 years

Image Source

About Nick:

Nick Anderson is the Director of Projects, Developing Nations at the United Nations. He is responsible for an annual development budget of $1 billion for use in projects across the world. Nick is based out of Geneva, Switzerland, and manages a global team of 100.

A day in the life of Nick:

Nick's day is filled with meetings. Almost 80% of Nick's work day is spent with colleagues, discussing budgets and projects. The focus of these meetings are new projects to invest in and the monitoring of existing projects. His days also involve meeting other top management officials from around the world including government officials, leaders of firms implementing the projects on the ground and other officials within the United Nations. Nick recommends projects to invest in based on the fit with United Nations' Millennium Development Goals. Nick's team creates a summary of project progress reports on a monthly basis; and Nick leads quarterly reviews with the project managers.

Nick is a great manager. His subordinates have voted him as best manager for three years in a row now. Across the organization, Nick is known for his strength in strategic thinking as well as attention to detail.

Persona 2 - Ann Blake, 28 years

Image Source

About Ann:

Ann is a graduate of London School of Economics and worked for Morgan Stanley as a financial analyst based out of the London, UK, office before moving to Paris, France, to work with the International Monetary Fund (IMF). At IMF, Ann's role is as a project analyst in which she is responsible for monitoring IMF's funds in various projects.

A day in the life of Ann:

Anne's day involves creating various reports for her manager, who manages a team of data analysts. Some of the reports Ann works on are budget allocation, project progress and project completion. In a typical day, she uses multiple separate reports to create an aggregate or

summary report. Anne is very data savvy and can work with multiple software tools. However, she uses Excel the most since that was the software tool used at Morgan Stanley and the one that she is most comfortable in.

Anne has consistently scored ‘Exceeds Expectations’ during her time at the IMF and is the go-to data person on her team for anything data related.

Ann is our primary persona, and we have incorporated features in our prototype based on her requirements.

Persona 3 - Ramesh Kumar, 38 years

Image Source

About Ramesh:

Ramesh is a project manager with the Shakthi foundation and is responsible for implementing on ground infrastructure projects for the non profit organization – Shakthi, based in Chennai, India. Ramesh was previously a project executive responsible for an infrastructure project in Thiruvedagam village in Madurai district, Tamil Nadu, India.

A day in the life of Ramesh:

Ramesh starts his day on various project sites, at around 10am. Post meeting with the on ground project coordinator and monitoring progress, he either visits another site or goes back to the administrative office in Chennai that he is based out of to fill out the day's reports. In some cases, the project sites are far apart and he stays overnight at a hotel near one of the project sites so that he can cover a few project sites on one trip. Despite traveling for 15-20 days in a month, Ramesh is unable to visit some of the project locations regularly due to the distances involved and hence project progress is slower in these cases, since his sign off is required on critical steps. While Ramesh is grateful for his promotion, he also feels that he is spending too much time away from home and not enough time with his family. Ramesh has had very little experience with data tools. He uses Excel to create reports and uses PowerPoint to make presentations to his managers. He uses the basic functions of these software tools and his presentations are usually to the point and do not contain any fancy graphics.

Ramesh is well respected among his peers and his subordinates since he is seen as empathetic as he has been one of them.

Persona 4 - Otome Kamwendo, 26 years

Image Source

About Otome:

Otome completed his undergraduate degree from University of Pennsylvania, majoring in Economics. During his last year at Penn, he attended London Business School as an International Exchange Student and thereafter has been working with Tony Blair Associates on various projects in Africa. Otome spent a year in Vietnam supporting the government with improving the climate for foreign direct investment (FDI) and accelerating public-private partnerships in infrastructure. Otome is based out of Kigali, Rwanda, and now works closely with the Rwandan government, as a part of the Tony Blair Africa Governance Initiative. He supports Finance ministry projects in the areas of payments, credit ratings, financial literacy and access to financial services. Otome recommends locations for ATMs based on location analytics. One of the issues he grapples with is the access to quality data on ground in Rwanda.

A day in the life of Otome:

Otome spends about half his day coordinating with his counterpart in the Finance ministry on data and project specifics. The other half is spent in recommending way forward for the various initiatives and obtaining buy-in, both from the Tony Blair initiative as well as the Rwandan government, specifically the Finance ministry.

Otome has a lot of respect among his counterparts in the finance ministry due to his educational background and business savviness.

4.5 Scenarios

Based on the user personas identified above, the user scenarios have been crafted. These are based on our discussions with the various stakeholders.

Nick Anderson

Nick has recently won a grant for a developmental project in Malawi and is meeting with the Millennium Development Goals committee to recommend an investment plan and to provide an update on the status of current projects. Nick relies on project summaries from his subordinates and since each project team works separately, he often has multiple project reports for one country. He wishes he could access a tool that would enable him to choose a country and then see the projects within that country. He could also foresee using this tool to explain his investments so that he can show new project geographies as well as show build on existing projects.

Ann Blake

Ann is working on a project to determine the poverty levels across countries in Sub Saharan Africa, South America and Asia. Ann is working with disparate data sources such as country census data, world bank reports and private market research reports. Since she is using different reports, she is not sure of the comparability of data points across the countries, and wishes she could have a report that allows her to look at poverty levels across the different areas in a country.

Ramesh Kumar

Ramesh is preparing for the quarterly update of his projects with the Head of Operations of the NGO, Shakthi. The pace of his projects have slowed and Rohan wants to showcase to the management the various on ground issues his team faces such as accessibility of these projects and manpower availability to name a few. He feels showing individual reports would not drive home the common points (both positive and negative) across all his projects. He wishes he could have a report that would show the management the geographical locations of his projects and the progress at each site.

Otome Kamwendo

Otome has just heard from his manager, the Director of the Tony Blair Africa Governance Initiative, that the payments company Visa is partnering with Tony Blair Associates to explore adding ATMs to remote locations in Rwanda. Otome has been tasked with identifying 100 potential ATM sites based on current access levels in that area and also determining the cost to serve these locations. For this, Otome needs an estimate of the poverty levels and financial literacy levels at various locations. Moreover, in order to estimate costs of serving, he would need to look at measures such as frequency of trips to ATM, cost of these trips, etc. He wishes he could have a centralized tool to help him evaluate these metrics for the different location options.

4.6 Limitations

One of the limitations is that the specialized nature of these projects made it difficult to find the relevant stakeholders for contextual enquiries. Another limitation is that the sampling for usability testing of the various prototypes was based on convenient sampling, and hence might have introduced biases.

5. Data Collection

5.1 DHS Data

Demographic and Health surveys are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition[26]. DHS surveys provide information on wealth, nutrition, education, HIV prevalence, among others. The DHS program has conducted over 300 surveys in over 90 countries. We acquired DHS version VI data on all countries for all countries considered in our project. This data would serve as a form of 'ground truth' data against which we could compare the predictions made by our model.

5.2 Downloading Daytime Satellite Imagery (Google Static Maps API)

We relied on the Google Static Maps API[27] for the daytime satellite imagery used in our neural net methodology. The starter code to develop this pipeline was inspired and based upon an assignment for Josh Blumenstock's class at the I School, Data-Intensive International Development (Info 290). The final version of our code implementation is available here.

The pipeline for satellite imagery collection and storage consisted of the following activities:

- *Obtaining a global base map reference*. A raster image of the entire world was loaded and processed, in order to construct a two-dimensional matrix of equally-spaced cells, of which each cell represented a 1km by 1km grid. The raster image was taken from worldwide night lights data, from the National Oceanic and Atmospheric Administration DMSP-OLS dataset, for 2010 (see "Version 4 DMSP-OLS Nighttime Lights Time Series"). The raster contains cloud-free composites of average visible, stable lights, and

[26] DHS Overview. Retrieved from http://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm
[27] https://developers.google.com/maps/documentation/static-maps/

cloud free coverages, with a smooth resolution of 30 arc-second grids, spanning from -180 to 180 degrees longitude and -65 to 75 degrees latitude.

- *Processing the shape files of each country for which images were retrieved*. The shape files, which were downloaded from the *GADM Database of Global Administrative Areas*[28] , were loaded and processed, and then a rectangular area was constructed to border the country's shape. The array of column and row indices of cells, based on the grid for the entire world, were sub-selected from the worldwide raster so that we could work with a complete that covered the individual country's geographic boundaries.

- *Iterating across the rows, and down the columns, for each of the cells, for a given country.* During each step of the iteration, the Google Maps Static API was called to retrieve a 400x400 pixel image, which corresponded with the centroid (specified by latitude, longitude) of the cell in the nightlights raster. During each iteration, the luminosity value from the night lights raster is stored with this image.

  Images were downloaded at *zoom level=16* resolution, where each image pixel corresponds to a square grid of 2.38865 meters on each side.[29] We queried for images with dimensions of 400 by 400 pixels, so that each daytime satellite image roughly represents a 1km by 1km square, to align with each cell in the nightlights raster.

5.3 AWS configuration

In order to efficiently obtain the data our models required, we needed to set up a data-retrieving pipeline that could not only serve our immediate purpose to get the images we needed but could also scale in the analysis. We could have done if more time and computation could have been possible.

---

[28] http://gadm.org/about

[29] https://groups.google.com/forum/#!searchin/google-maps-js-api-v3/2.38865|sort:relevance/google-maps-js-api-v3/kfd3n5wbwo0/hApc1qmOGJkJ

In this case we set up an Amazon Web Services (AWS) elastic computing instance[30] that served as the main system to download all of the images for all the countries we decided to analyze.

The main AWS instance we used consisted on a *t2.2xlarge* system, with:

- 8 vCPUs
- 32Gb RAM
- +200GB on Elastic Block Store (EBS)[31] volume.

More storage was used to store the downloaded images on the Amazon Simple Storage Service (Amazon S3)[32] and later moved onto a Google Cloud[33] platform storage.

5.4 Adaptation, Scaling, and Parallelization

One key challenge was the vast scale of images that we sought to retrieve from the Google Maps API. Rwanda, the country focused on for the class assignment, is a relatively small country. It still required additional work in order to sufficiently generalize and modularize the script in order to flexibly afford the retrieval of images for different countries.

We retrieved nearly 2.7 Million images from the following four countries of Ghana, Malawi, Rwanda, and Tanzania:

[30] https://aws.amazon.com/ec2/
[31] https://aws.amazon.com/ebs/
[32] https://aws.amazon.com/s3/
[33] https://cloud.google.com/storage/

Data Years (DHS)	Country	Left Cell Index	Top Cell Index	Right Cell Index	Bottom Cell Index	Number of Images	Percent of Total (%)
2014	Ghana	21,216	7,668	21,735	9,000	693,160	25.9%
2015 - 2016	Malawi	25,539	10,137	25,900	11,052	331,592	12.4%
2014 - 2015	Rwanda	25,067	9,133	25,301	9,336	47,940	1.8%
2015 - 2016	Tanzania	25,158	9,124	26,438	10,373	1,601,250	59.9%

FIGURE: Table of cell indices (in worldwide nightlights raster) and number of images retrieved for each country from the Google Static Maps API.

Total Number of Images: 2,673,942

The free tier of the Google Static Maps API allows up to 25,000 map loads per API key, per 24 hour period, with a 640 x 640 maximum image resolution.[34]

At a maximum daily limit of 25,000 images each day, with only one Google Static Maps API key it would have taken 107 days to download and save the nearly 2.7 million images that we collected for this project.

We acknowledged that more images would (1) improve the sample size and generalizability of our machine learning model training and fitting, and (2) expand the reach of our product to a greater number of countries. We considered and weighed against several alternative approaches: for instance, obtaining images at lower-resolution, such as zoom level 15 or 14 (where each image pixel represented 4.8 meters or 9.6 meters on each side, respectively), which would expand our geographical breadth.

We weighed this option (retrieving images at lower resolution) with obtaining higher-resolution images, at zoom level 16, which covered a relatively smaller overall area. Kim et al (2016)[35] found that there was marginal differences in model performance when relying upon satellite imagery at zoom levels 14, 15, or 16. We also considered potentially retrieving high-resolution (i.e., zoom level 16) imagery, but larger in dimension, at 1600x1600 pixel resolution-- 16 times

[34] https://developers.google.com/maps/documentation/static-maps/usage-limits
[35] Kim, J., Xie, M., Jean, N., Ermon, S. (2016). Incorporating Spatial Context and Fine-grained Detail from Satellite Imagery to Predict Poverty. Working Paper. DOI: 10.13140.

the area of a 400x400 image-- but the free tier of Google's API constrained image sizes to a maximum of 640x640 image resolution.

Dense Area

Level 16

Level 15

Level 14

Rural Area

Level 16

Level 15

Level 14

FIGURE: Comparison of daytime satellite imagery retrieved from the Google Static Maps API at different resolutions, in rural and dense areas: zoom levels 16 (2.4m), 15 (4.8m), and 14 (9.6m).

In the end, we decided to use higher-resolution images, at zoom level 16, in spite of the findings in Kim et al. Upon visual inspection of the images at zoom levels 14, 15, and 16 (see figure above), we decided that levels 14 and 15 were not sufficiently-detailed to be able to discern meaningful differences in the physical features.

We mitigated the downside of using higher-resolution imagery acquiring multiple API account keys from amongst our team's members, and, from classmate peers, to download daytime satellite imagery from each API account, in parallel. This allowed us to still manage to obtain a large set of satellite imagery, but for a broader range of countries.

A pipeline was created in order to partition the cells for each country into smaller regions that were traversed using each API key / account. A spreadsheet was used to calculate the width and height for each country's boundaries, and split the country's area into into evenly-spaced rows. Our download script accepted arguments to execute based on a different country shape file (e.g., Ghana, Malawi, Rwanda, Tanzania) and starting latitude in the country's geography.

Once the 14 parallel scripts -- one script execution for each API key-- were calculated to cover all cell indices for a given country, they were then configured to run in parallel, for consecutive days, using a terminal multiplexer. tmux was used to execute all 14 scripts concurrently on the same AWS EC2 instance. Moreover, Bash shell scripts were written to run all 14 jobs each day, and then scheduled via CRON jobs, a time-based job scheduler, for multiple days in succession.

5.5 Quality Control and Testing

At the outset, there was a large degree of testing, trial, and error upon deployment of the scripts. Due to the large number of files that were being worked with, it was not readily apparent the extent to which the downloaded images were correct or complete. In spite of best efforts to ensure that the code was correct, there inevitably were necessary tweaks. However, each cycle

of running the script can take many hours at a time, which delayed feedback. For instance, it was not until after implementing our pipeline and executing that we realized that the iterator for rows and columns were being swapped, which affected the completeness of collection due to the script requiring a starting row position. In another instance, we later noticed that inconsistent versions of DHS clusters were being used, and had to be replaced later-on. It became essential to to incorporate tools into our pipeline that afforded us the ability to assess the quality and completeness of our geographic data inventory, to ensure that our processes were as smoothed out and generalized as early on as possible. We also relied upon peer review.

In some cases earlier on, we noticed that our download script was missing parts of countries--i.e., it was not retrieving the number of images that we anticipated. Therefore, we ended up writing a script to traverse each of the resultant image files and directory names, and then count the number of images in each row of the worldwide reference grid that were downloaded from the Google Maps server. This was to ensure that the total count of images should be the same as the anticipated number of images for the given country, but, to also ensure that the number of images in each row is uniform, so that the two-dimensional array of images completely covers the rectangular area that encompasses the DHS clusters in the given country. This was run as a quality check during retrievals for other countries.

6. Machine Learning

6.1 Setting

Given that direct measurement of wealth and economic indicators is difficult to scale up, alternate methods for indirect inference of measuring wealth at a given location have been actively explored and include leveraging mobile phone data and the luminosity of a location during night time as observed by satellites as proxies for human economic activities. Among the above, leveraging satellite images of night light luminosities have been gaining traction over recent years[28-31] . Mellander et. al found that there was a strong enough correlation between night lights and economic activity for it serve as a good proxy for population and establishment density[32].

While night lights have been shown to hold promise in improving predictions concerning country level economic status, it has become more apparent that since luminosity levels in poor areas are generally uniform, night lights data is not effective in identifying differences in economic activities in poor areas, where the population is living under the international poverty line. The same problem also holds when it comes to distinguishing between densely populated poor areas and sparsely populated wealthy areas, both of which, one would expect, have low luminosity emanating. Thus night lights alone are a rough and noisy, but consistently available proxy for human activity. Other features such as agro-ecological conditions, built environments, distance from cities etc. have shown to have linear relationship with the economic well being of an area.

Features such as the built environments and habitations that can be directly measured from space as daylight images have correspondence to the economic conditions of a given locations as dictated by conventional wisdom. Humans are usually capable of distinguishing and visually estimating the wealth of a location. However programmatically it is hard to understand what features in an image correspond to a region's wealth from raw pixel values. Day time images of a location encapsulating such features are available consistently across the globe. Since night-lights have been shown to distinguish wealth, albeit effectively at higher income settings, learning a representation of daylight images by training them against night lights can help

extract features from a built environment that is capable of distinguishing economic activity across the gamut. Jean et al. in their work on using satellite imagery to predict poverty demonstrated the effectiveness of such an approach in Rwanda[33]. Here we extend upon Jean et al's work scaling it across multiple countries in Africa.

We first solve the problem of learning a representation from daylights by training a deep Convolutional Neural Network (CNN) to predict the night light intensity of a location given a daylight image. We further treat the trained CNN as a feature extractor and extract a representation of the image and regress over the representation to fit a model to predict the night wealth index.

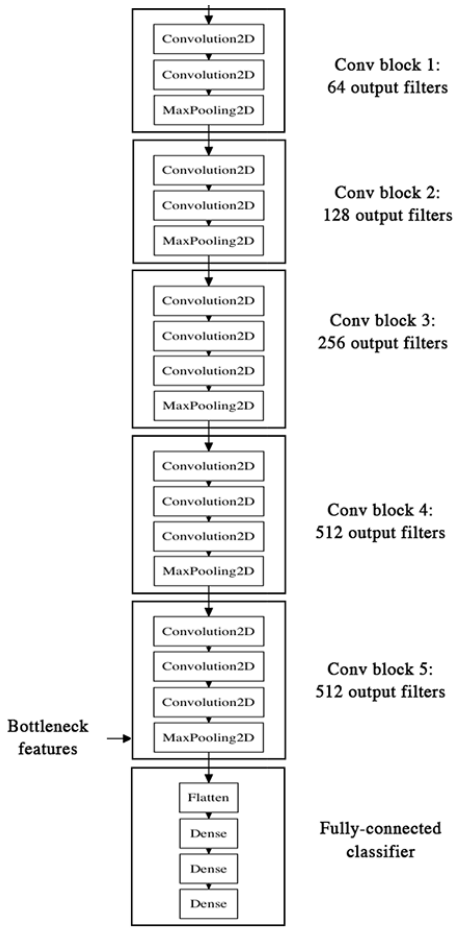
6.2 Transfer learning on Convolutional Neural Networks:

Considering the constraints with data, which include, but are not limited to:

a. Lack of labels
b. Paucity of training samples
c. Large feature space

Training a deep Learning model from scratch becomes infeasible due to the above constraints and the lack of computing power at our disposal. We instead use a transfer learning approach, where we use a pre-trained Convolutional Neural Net (CNN) and transfer the learning acquired by being trained on a certain classes of images to our problem. CNN models that have been trained on the ImageNet dataset, a gold standard image dataset covering 1000 categories, have been known to be good feature extractors, being able to generalize to many new tasks[34]. In our approach we use a pre trained VGG16 Net, that has been trained on ImageNet dataset as a feature extractor. However we are cognizant of the fact that satellite images, which are taken from a bird's eye perspective are completely different from the ImageNet dataset, which is as seen from human eye. In order to solve this problem, instead of using the VGG16 as a generic feature extractor, we use an approach known as fine tuning, wherein we freeze the weights of the convolutional blocks (There are 5 convolutional blocks in VGG16. Refer to architecture) and

extract the features out of them, a process also known as bottleneck. We then proceed to modify the final fully connected layers of the VGG16 to predict three class and initialize and train the fully connected layer with the features extracted as inputs. This fine tuning approach is computationally less intensive requiring relatively less infrastructure, while at the same time leverages the full depth and richness of the convolutional blocks as feature extractor and simultaneously fine tuned for a different three class prediction problem.



6.3 Training

We trained and fine tuned a pretrained VGG16 model on a subsample of satellite images. Since there was extensive class imbalance in the data with 95% of satellite images corresponding to zero light intensity, we subsampled the images from zero night light intensity uniformly through contiguous geography. We further upsampled the number of images for training corresponding

to higher luminosity by performing rotations, flips and adding random noise to the images. We then proceeded to train the model. We did not perform fine tuning as the goal here was to learn a generic representation and as such stopped training as we reached and passed the baseline accuracy.

6.4 Representations as features

Once features have been extracted we carry out Principal Components Analysis to reduce the feature space of the representations to their corresponding subspace that explains maximum variance in data in order to reduce overfitting. We further carry-out ridge regression with L2 regularization to predict the average wealth of a location.

CNN models trained on the ImageNet dataset are recognized as good generic feature extractors, with low-level and mid-level features such as edges and corners that are able to generalize to many new tasks (Donahue et al. 2013; Oquab et al. 2014). Arvix NJ

Previous work in domains with images fundamentally different from normal “human-eye view” images typically resort to curating a new, specific dataset such as Places205 (Zhou et al. 2014). Additionally, unsupervised approaches such as autoencoders may waste representational capacity on irrelevant features, while the nighttime light labels guide learning towards features relevant to wealth and economic development.

We find that the correlation between NTL and economic activity is strong enough to make it a relatively good proxy for population and establishment density, but the correlation is weaker in relation to wages. In general, we find a stronger relationship between light and density values, than with light and total values. We also find a closer connection between radiance light and economic activity, than with saturated light. Further, we find the link between light and economic activity, especially estimated by wages, to be slightly overestimated in large urban areas and underestimated in rural areas.

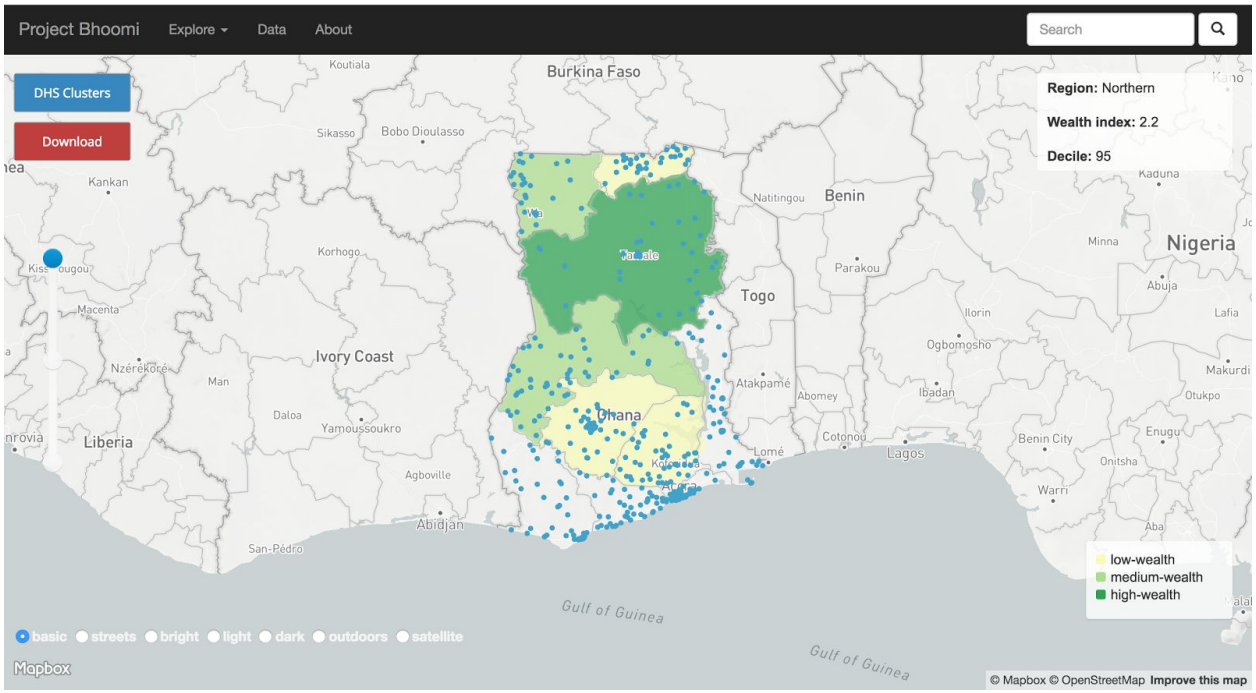
7. Website Architecture

Our web application has been launched at http://www.project-bhoomi.co/

On top of all the analysis we made on the satellite imagery and DHS data, we built a tool that could make all this information more accessible and could expose our stakeholders to easy-to-digest information about wealth for several African countries. Moreover, the tool would allow power users to export the information they would need in order to incorporate into other work products and decision-making processes.

7.1 Objective

The main idea of the web application is to expose users to an exploration of wealth regions on several African countries. For this purpose, we displayed wealth index predictions, along with DHS data and basic geographic information in a map interface, with color-coded symbols to provide a quick visualization of the different wealth concentrations within the country.



Within these visualizations it was also critical, based on our user research, to provide a quick and easy way to download all the information the users may need for their posterior analysis and decision-making. Every screen provides a download feature, affording users the ability to obtain, in CSV format, all of the information they are currently visualizing.

A search tool is provided to ease the process of moving between different countries and regions. Also, controls to allow people to filter certain information or layers, DHS cluster centers for example, or to modify the level of aggregation and detail people would like to see on the maps, like administrative regions and such.

Every map visualization can be displayed on seven different styles, from satellite and street view to a more clean and basic styling.

7.2 Technology stack

The web application was built around the following technology stack:

7.2.1 Backend

- Django 1.11 (relies upon Django-GeoJson)
- PostgreSQL/PostGIS (Hosted on Heroku)
    - Stores geometries for DHS cluster points, regional polygons, 1km x 1km cell polygons, and associated data (e.g., household level data income). At the DHS cluster level, we are concerned with obtaining the median of these metrics in order to represent the cluster.
- GeoJSON API
- Amazon S3 file storage
- Heroku deployment, with continuous integration via Github

7.2.2 Frontend

- Mapbox GL API
- Bootstrap CSS

8. Evaluation & Next Steps

8.1 Usability Testing

Each of the prototypes was tested using convenient sampling and think aloud sessions. In each of these sessions, we gave users three tasks and analyzed their workflow and their task completion.

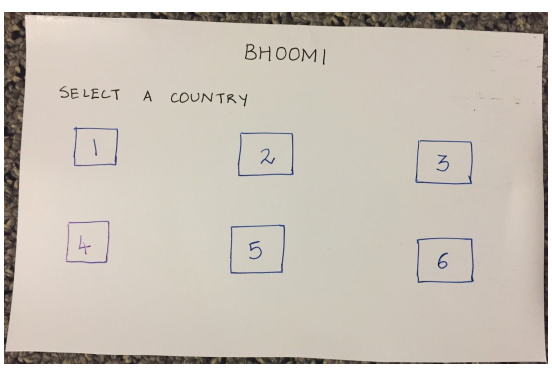
The tasks that we had the users complete were:

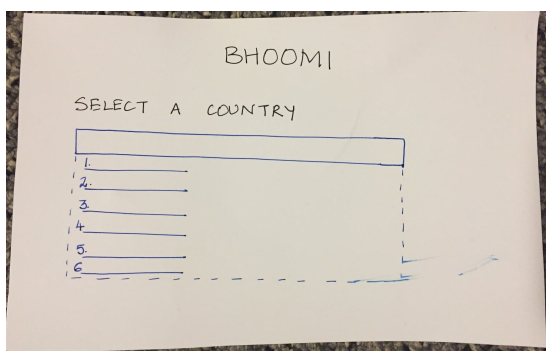
1) To find the wealth index of a DHS cluster in Rwanda
2) To find the population of a DHS cluster in Rwanda
3) To download the entire dataset for Rwanda

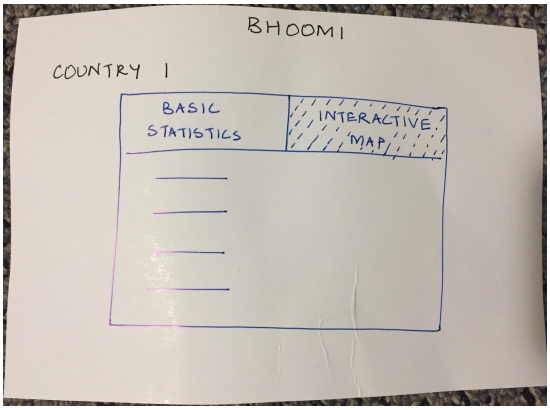
8.1.1 Paper Prototype

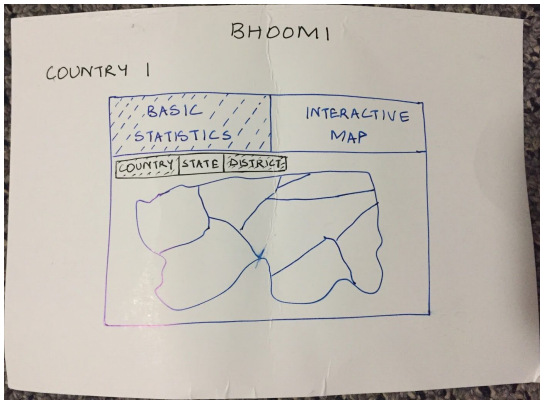
The initial paper prototype was developed to introduce the user to the web tool to get information about the countries. At this level, we included options such as basic statistics for a country and the ability to get more detailed information about the chosen level of granularity - country, state or district.

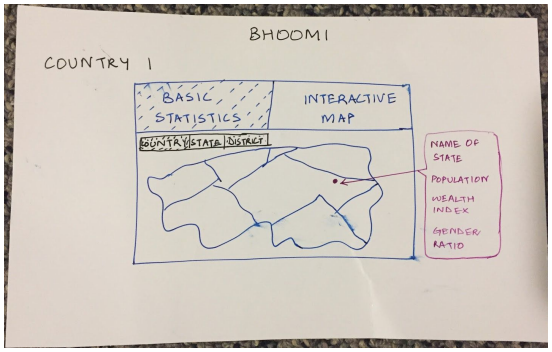
We also looked at two options of choosing countries - that of displaying them all or using a dropdown.









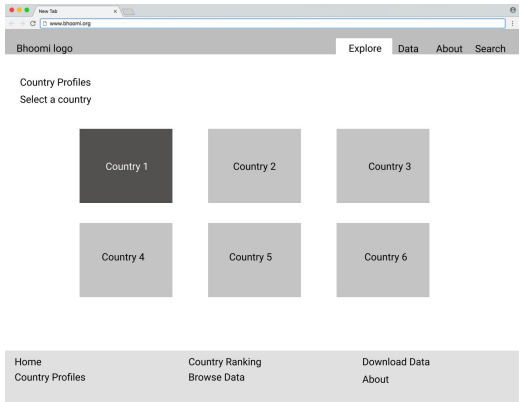


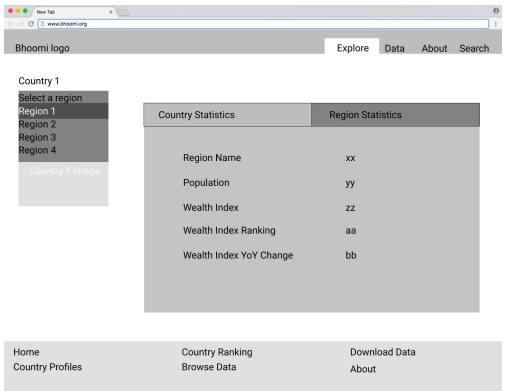
The usability tests with the paper prototypes revealed that users liked the interactivity of the map. Being able to choose the different levels of granularity from country to district was also found useful as a feature.

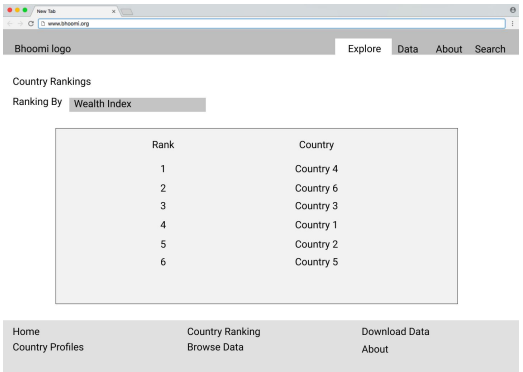
We had the users complete the tasks mentioned in the previous section, and we found that in finding the wealth index, there was some confusion among users whether basic statistics would reveal the wealth index or whether they would need to use the map for that information. Finding the population was straightforward and all users were able to navigate and find population information. We did not test task 3 since we had not incorporated the data download option in this paper prototype.

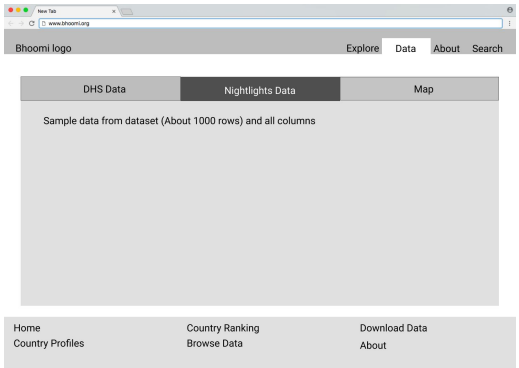
8.1.2 Wireframe

Post the paper prototypes, we developed two versions of wireframes. The first was developed similar to the paper prototype, with additional features such as data download options and country profile and ranking information. Some of the frames of this prototype are given below:



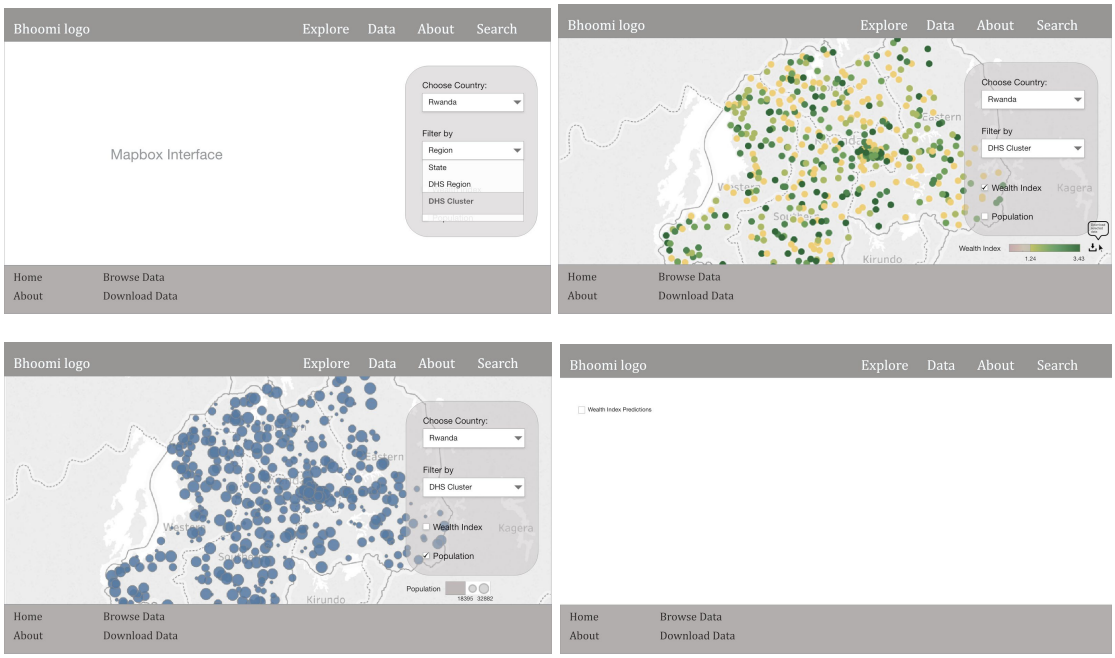






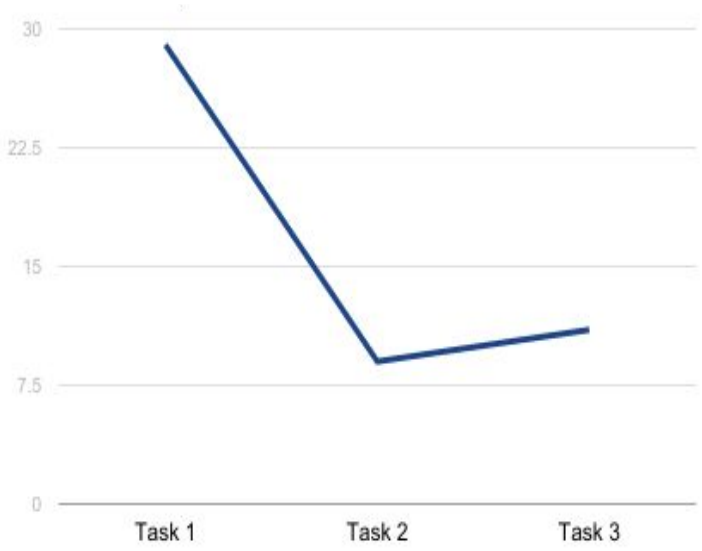
The feedback on this from our project advisor, Joshua Blumenstock, was to focus on the interactivity of the map and not on aggregate statistics since that information is available from other sources.

Based on this feedback, we revised the wireframe to remove aggregate level statistics and rankings and keep the focus on the interactive map. We added various levels of granularity (from Region to DHS cluster) and layers for population and wealth index. We also added a download feature that enables a download of the selected level of granularity on the map. A few sample frames are shown below.



We tested this design among users. Since this tool is targeted at specifically those in the field of international development and since the sample we chose for testing was based on convenience sampling, we gave the users a context of the data, and then asked them to complete the three tasks.

We saw that on average, the task completion time reduced as the user progressed from one task to the other. The steep decline in task completion time between tasks 1 and 2 could be attributed to the similarity of the tasks as well as familiarity with the interface. Task 3 takes slightly longer than task 2, since it involves a different feature of the interface.



Another qualitative feedback on the interface was that users wanted more map “real-estate” on display rather than the options to make the selections such as country, region, etc. We’ve incorporated this feedback in our website by removing the selection box on the side and focusing on the map interface.

8.2 Limitations

The sample chosen was based on convenience and this could have led to some biases. In addition, providing the users with context might have added to the bias.
While the decreased task completion time for tasks 2 and 3 could be attributed to familiarity with the interface; some of it could also be attributed to the similarity between tasks 2 and 3.

8.3 Next Steps

Our prototype is a working demo of the web tool. As next steps, we could add the population layer so that population for each granular region is available along with the wealth index. We would also like to test the prototype with stakeholders such as planners and policy makers to get their perspective.

While we have limited our results to a few countries, the information architecture (pipeline) that has been set up makes it possible to replicate the process for any country for which there exists the required data.

9. References

- J. V. Henderson, A. Storeygard, D. N. Weil, Am. Econ. Rev. 102, 994–1028 (2012).
- X. Chen, W. D. Nordhaus, Proc. Natl. Acad. Sci. U.S.A. 108, 8589–8594 (2011)
- S. Michalopoulos, E. Papaioannou, Q. J. Econ. 129, 151–213 (2013).
- M. Pinkovskiy, X. Sala-i-Martin, Q. J. Econ. 131, 579–631 (2016)
- Mellander C, Lobo J, Stolarick K, Matheson Z (2015) Night-Time Light Data: A Good Proxy Measure for Economic Activity? PLoS ONE 10(10): e0139779. https://doi.org/10.1371/journal.pone.0139779
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.
- Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2013. DeCAF: A deep convolutional activation feature for generic visual recognition. CoRR abs/1310.1531.
- Nath, V. (2011). Administration and Development Planning in India. Concept Publishing Company.
- Design and Implementation: Targeting & Selection (2013, Jun 6). Retrieved from http://www.worldbank.org/en/topic/communitydrivendevelopment/brief/cdd-targeting-selection
- Castañeda, T., Lindert, K., de la Brière, B., Fernandez, L., Hubert, C., Larrañaga, O., ... & Viquez, R. (2005). Designing and implementing household targeting systems: lessons from Latin America and the United States. World Bank Social Protection Discussion Paper Series, 526.
- Van Domelen, J. (2007). Reaching the poor and vulnerable: Targeting strategies for social funds and other community-driven programs. Network HD. Washington DC: World Bank, 1-52.
- Alatas, V., Purnamasari, R., Wai-Poi, M., Banerjee, A., Olken, B. A., & Hanna, R. (2016). Self-targeting: Evidence from a field experiment in Indonesia. Journal of Political Economy, 124(2), 371-427.
- Kumar, N. (2005). The Effectiveness of World Bank Support for Community-Based and-Driven Development: An OED Evaluation. World Bank Publications.

- Chander, R., Grootaert, C., & Pyatt, G. (1980). Living standards surveys in developing countries (No. 1). World Bank, Development Research Center.
- Jerven, M. (2015). Measuring African development: past and present.
- Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. The Journal of Economic Perspectives, 30(4), 171-198.
- DHS Overview. Retrieved from http://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm
- http://www.globalforestwatch.org/
- https://developers.google.com/maps/documentation/static-maps/
- http://gadm.org/about
- https://groups.google.com/forum/#!searchin/google-maps-js-api-v3/2.38865|sort:relevance/google-maps-js-api-v3/kfd3n5wbwo0/hApc1qmOGJkJ
- https://developers.google.com/maps/documentation/static-maps/usage-limits

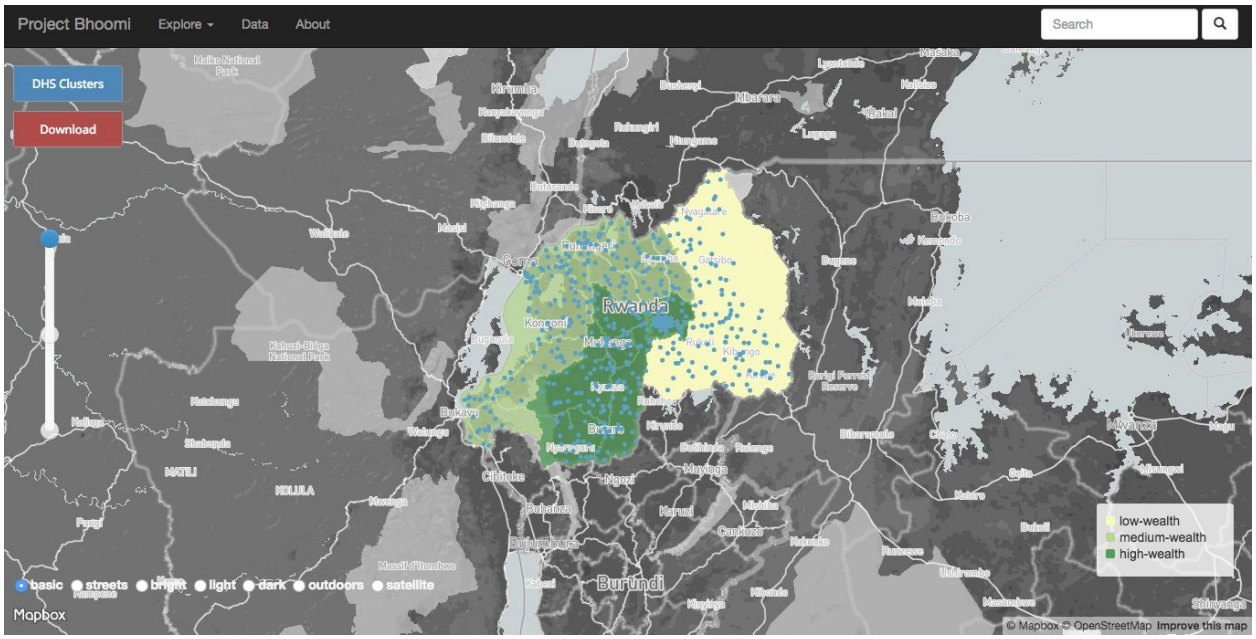
10. Appendix

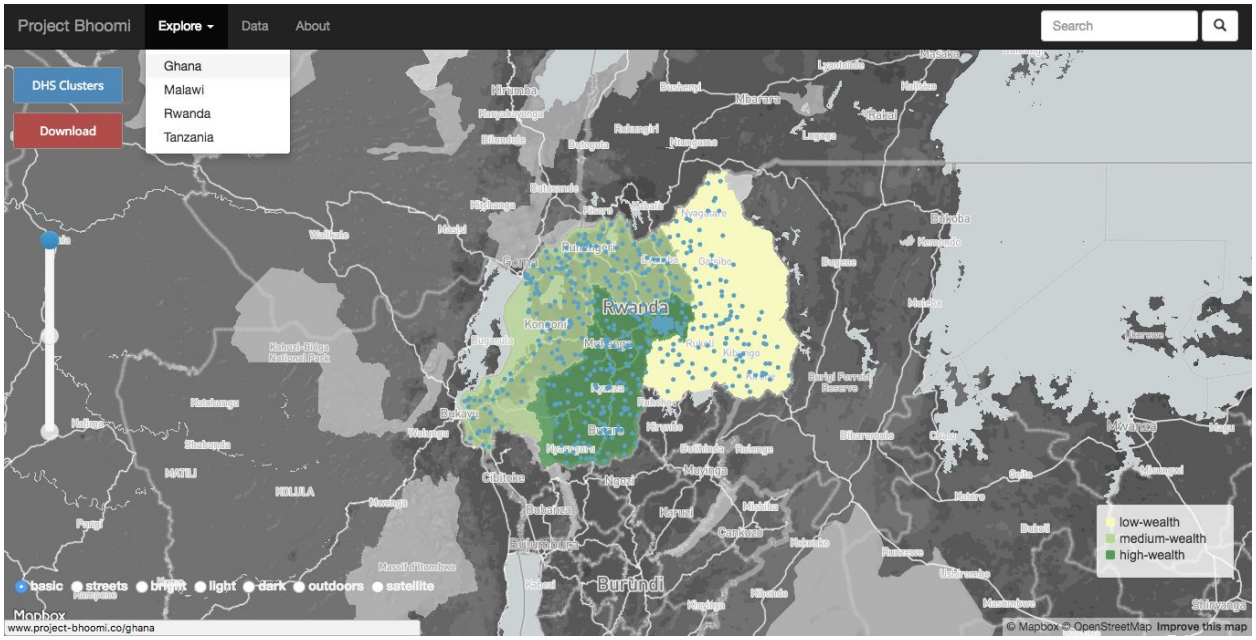
10.1 Appendix A: Discussion Guide - In Depth Interviews

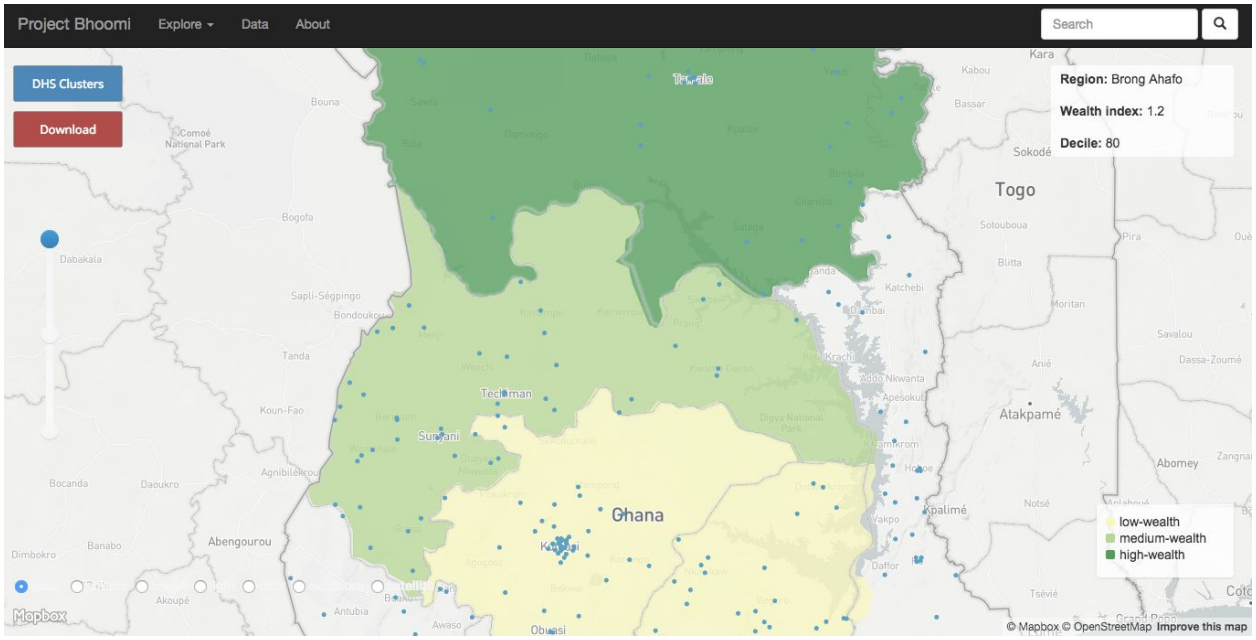
- What are examples of some projects that your organization typically works on?
- How are areas for development currently identified?
- What are the economic metrics currently used in your projects?
- How is consumption expenditure and asset wealth currently captured across countries?
    - What are the challenges with this method?
- Are these measures of wealth the same for all countries? Is there a variation in metrics between developing and developed countries?
    - Is there a variation in the method of data collection?
- What are the challenges with respect to data collection?
    - What kind of data do you find most useful?
    - What level of granularity would you want to achieve with the data?
    - What data manipulations would you find most useful?
- What is the kind of analysis you work on with the data collected?
    - Are there specific metrics?
    - Do you use a specific software or tool for this analysis?
    - What are the current challenges of this software or tool?
    - Have you ever considered using satellite image data for this?
- What kind of projects does this sort of data enable?
- Is the data more critical for planning or for analysis (pre or post)?
- What are the challenges in implementing these projects?
- How are the projects monitored?
    - Are there specific metrics you use to monitor?
    - Do you use a specific software or tool for monitoring progress?
    - What are the current challenges of this software or tool?
    - Have you ever considered using satellite image data for this?
- What are the other challenges you face?
    - Is corruption a challenge? How do you manage/mitigate it?

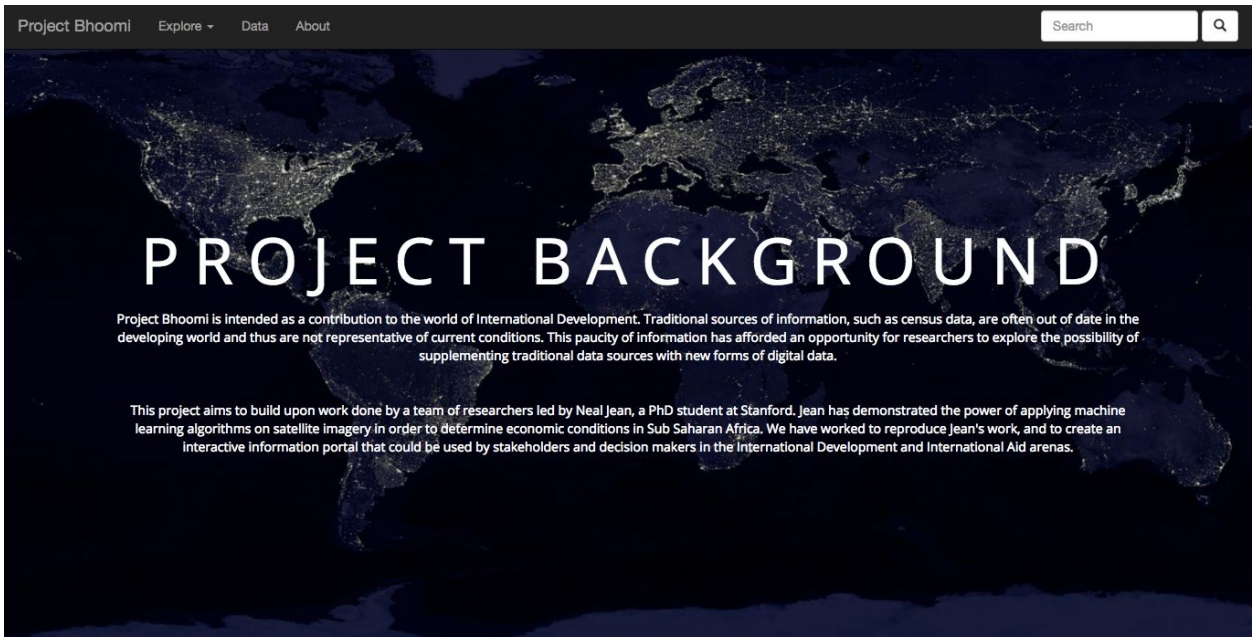
- In the link for Global Forest Watch (here), what features are most important for you? Why or why not?

10.2 Appendix B - Screenshots of Website






Project Bhoomi
Explore
Data
About
Search
DHS Clusters
Download
Region: Brong Ahafo
Wealth index: 1.2
Decile: 80
Togo
Ghana
Techiman
Sunyani
Atakpamé
low-wealth
medium-wealth
high-wealth
© Mapbox © OpenStreetMap Improve this map


Project Bhoomi
Explore
Data
About
Search
PROJECT BACKGROUND
Project Bhoomi is intended as a contribution to the world of International Development. Traditional sources of information, such as census data, are often out of date in the developing world and thus are not representative of current conditions. This paucity of information has afforded an opportunity for researchers to explore the possibility of supplementing traditional data sources with new forms of digital data.
This project aims to build upon work done by a team of researchers led by Neal Jean, a PhD student at Stanford. Jean has demonstrated the power of applying machine learning algorithms on satellite imagery in order to determine economic conditions in Sub Saharan Africa. We have worked to reproduce Jean's work, and to create an interactive information portal that could be used by stakeholders and decision makers in the International Development and International Aid arenas.