Data for sustainable development

Earth Systems 162/262, Computer Science 325B
Instructors: David Lobell (Earth Systems), Marshall Burke (Earth Systems), Stefano Ermon (Computer Science)
Course Assistants: Anthony Perez, Christopher Yeh
Meetings: Tuesday 2-4:30 pm, Shriram 108
Units: 3-5
Office Hours:
Stefano: Friday 11am-12pm, Gates 228 except this week - wednesday 4-5pm
David: Friday 1-2pm, Y2E2 room 367
Marshall: TBD
Class website: https://web.stanford.edu/class/cs325b/
Piazza: cs325b

Overview:

The sustainable development goals (SDGs) encompass many important aspects of human and ecosystem well-being that are traditionally difficult to measure. This project-based course will focus on ways to use inexpensive, unconventional data streams to measure outcomes relevant to SDGs, including poverty, hunger, health, governance, and economic activity. Students will apply machine learning techniques to various projects outlined at the beginning of the quarter. The main learning goals are to gain experience conducting and communicating original research. Prior knowledge of machine learning techniques, such as from CS 221, CS 229, CS 231N, STATS 202, or STATS 216 is required. Open to both undergraduate and graduate students. Enrollment limited to 24. Students must apply for the class by filling out the form at https://goo.gl/forms/9LSZF7IPkHadix5D3. A permission code will be given to admitted students to register for the class.

By the end of the quarter, students will be able to:

- Identify and explain the SDGs and the relevant outcome variables for each.
- Identify various sources of geospatial data, their pros and cons, and connect data sources to specific development problems
- Implement basic spatial operations in GDAL and basic spatial data visualization
- Develop ML applications in the cloud (scripting, batching, etc)
- Demonstrate improved skills in building ML models (avoiding overfitting, model development and selection, combining multiple data sources, modeling spatio-temporal dependence, etc.)
- Effectively communicate their problem and results in both oral and written form

Class schedule:

The class will consist of three main parts: the first week will be background info, the next seven will focus on team development of their projects with regular presentations, and the last two weeks will focus on wrapping up and communicating results. Guidelines for each presentation will be given later.

Week	**Topic(s)**	**Items Due**
Part 1: Intro		
1	-Introduction to the SDGs -Overview of project choices -Review of syllabus -Overview of common datasets and tools you might want to use -Examples of prior projects	
2	Group presentations: summary of -- what others have done on this topic -- what benchmarks are for performance on this or related tasks -- what other sources of data might be useful?	Literature Review Slides for presentation
3	Group presentations: -Data visualization. Show basic summary plots/maps of your data. E.g. what are typical images for high/low values of infrastructure -Discuss possible ideas for modeling	Slides for presentation
4	Group presentations: -Simple models. Show results from some baseline models using some simple reference model, e.g., regression	Slides for presentation
5	Group presentations:	Slides for presentation
6	Group presentations:	Slides for presentation

7	Group presentations:	Slides for presentation
8	Group presentations:	Slides for presentation Draft of final paper
9	-Peer feedback session	Written review of another team’s paper
10	Group presentations: -Final presentations	Slides for presentation Final paper

Grading components:

Individual Lit Review: 5 pts
Weekly Group Presentations (week 2-8): 6 pts each
Peer-review report of another team’s paper: 8 pts
Final presentation and paper (45%) (to be written in the form of a long blog post or a short conference paper)

Students ***will not*** be graded on whether they can successfully achieve their desired accuracies in predicting outcomes, given that most projects will be risky and not guaranteed to work. Students ***will*** be graded on devoting sufficient time to the project, clearly explaining progress and challenges, correctly applying techniques, and clearly writing up results. Successful projects will have the reward of paid trips to conferences (if the paper is accepted).

Length of weekly presentations will be determined by the number of projects. All students are expected to attend all sessions, and to give full attention and feedback to their classmates or instructors (no open laptops except for presenters).

Students will work in groups of 3 people. We expect that each member of the team contribute in both technical and non-technical components. At the end of quarter, we will solicit feedback on your teammates and reserve the right to give individuals in the group higher or lower grades than the group average.

Project topics (tentative):

Fall 2017:

1) Mapping infrastructure in Africa

This project will test the ability of deep learning models that use a combination of high (~1-3m) and moderate (10-30m) resolution optical and radar imagery to predict measures of infrastructure in Africa. Training data on measures such as access to electricity, quality roads, and piped water will be from the recently georeferenced Afrobaromter surveys of multiple countries, as well as detailed field data from Addis Ababa. The goal is to produce reliable maps that can be updated over time to track the provision of basic public services.

2) Mapping poverty in Uganda, Bangladesh, and India

This project will build upon past work that mapped poverty using CNNs and high resolution imagery (Jean et al. 2016). Three new datasets that include household-level measures of assets and expenditures will allow further refinement and testing of past approaches. In addition, the team will use new sources of imagery, including Sentinel-1 radar data, that could be useful for poverty prediction. The goal is to produce reliable maps that can be updated over time, in order to track the progress of communities in building assets and wealth, and test hypotheses about which factors speed up or slow down progress.

3) Forecasting crop production around the world (esp. Africa, Latin America)

This project will use primarily satellite data from MODIS (both surface reflectance and temperature) with CNNs and Gaussian processes to forecast crop yields. This approach was first developed using U.S. data for soybean and maize in You et al. (2017). This project will start with that model and then extend it for application to sub-national crop datasets in Argentina and for several countries in Africa. The goal is to produce accurate estimates of final yield at various lead times, from several months before to the month of harvest.

4) Forecasting food prices in India

This project will test the ability of geolocated tweets from India to track fluctuations in the prices of major food staples in India. The core datasets will be 2.5 years of tweets obtained as part of Stanford's Data Science Initiative, and local weekly price data in India from the World Food Program. Past work has suggested people discuss food more when prices are rising, such as in an Indonesia study by the UN Pulse Lab, but the concept has not been widely proven. The goal is to produce timely warnings of where prices are changing, particularly if they are moving very rapidly, as a way for governments, NGO's, and the private sector to cheaply monitor and respond to these situations.

5) Mapping land cover around the world

This project will develop methods to map the occurrence of cultivated croplands around the world at high spatial resolution. The core dataset will be ~50,000 high resolution images with crowdsourced labels of whether or not cropland is present in the image, as well as coarser 10m resolution images from Sentinel-2. The first step will be to see whether deep learning models can reproduce the human labels on the high-res imagery, and the second step to see whether the 10m data work nearly as well. The goal would be to use the 10m data, which is available for free globally, to produce a global map of where crops are. This information would be useful for a wide range of applications, including developing a mask to apply to more sophisticated analyses of crop yield (such as in project #3).

6) Tracking displaced peoples in humanitarian crises
   TBD

7) Tracking changes in population around the world

   TBD