SoLi: *Socioeconomics of Light*

Associating Nights Lights and Consumer Sophistication

Saifullah Rais, Nic Vogler and Heather Koo.

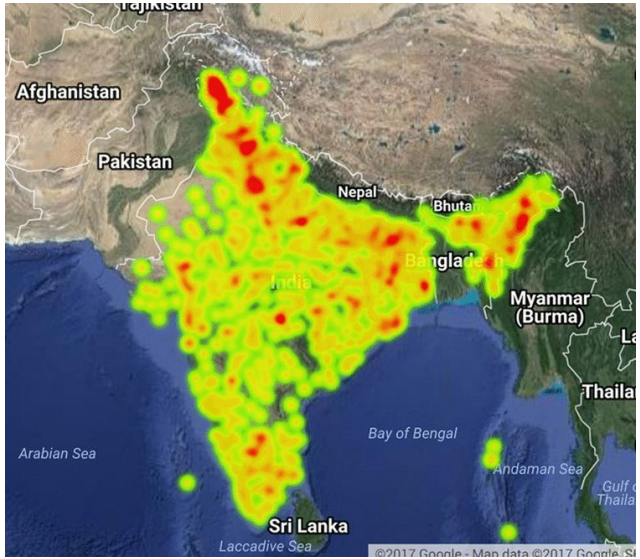
W205-2: Final Project

Socioeconomics of Light

Shine on: associating night-time lights and consumer sophistication

Status: Complete

Night Light Development Index



Brief summary

We use **night light** images to nowcast **socioeconomic characteristics** for India on a district-level.

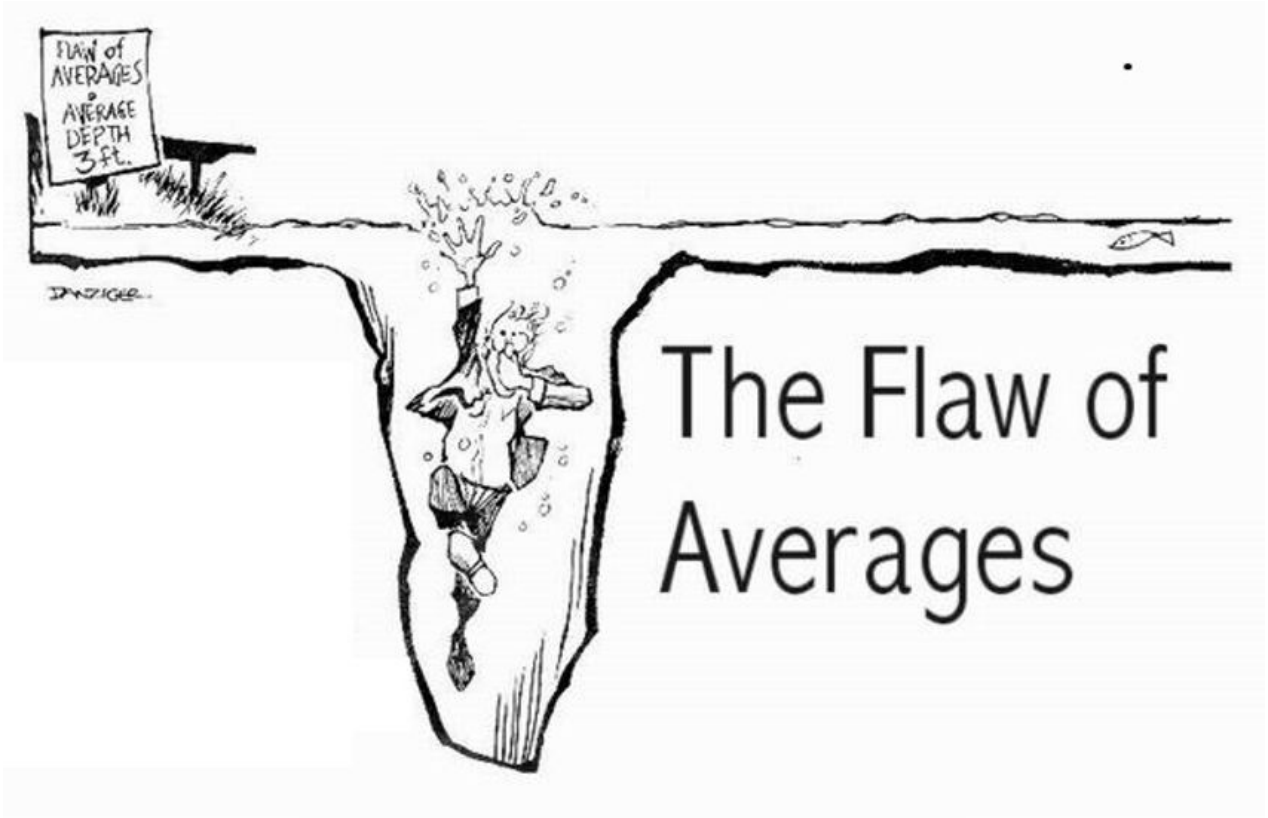
Unlike other research in the space, the predictions are made available real-time on Google's **Earth Engine** Platform.

Detailed links

The Need:

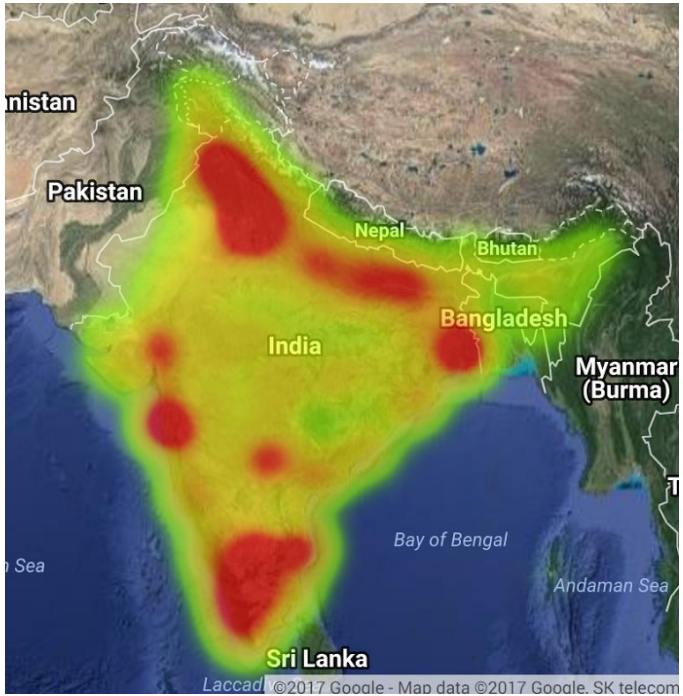
Tackling ecological fallacy

Developing markets: Lands of great divide



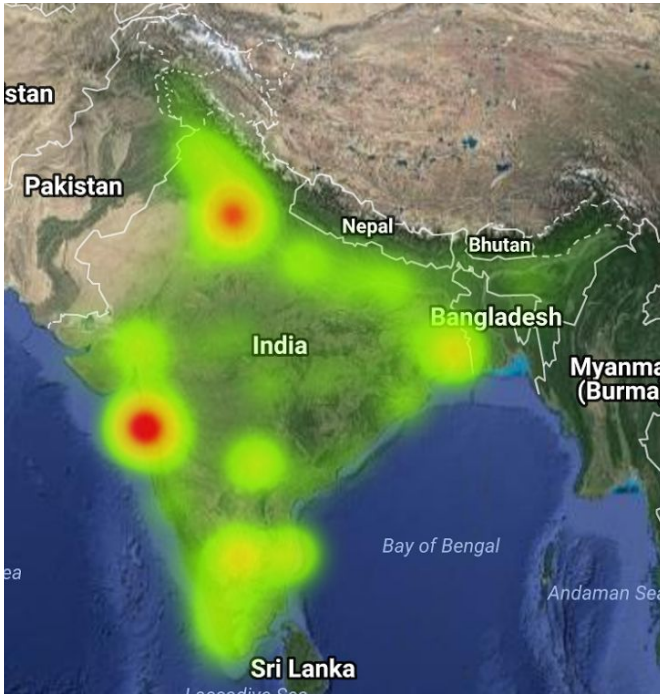
An illustration: India's geographic disparity

Distribution of Bank Branches, 2014



Source: RBI

Distribution of Deposits (in value), 2014



Source: RBI

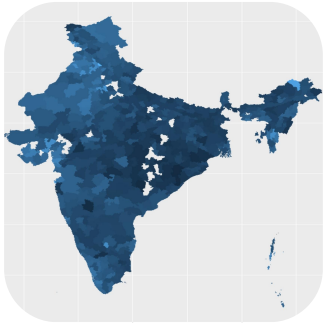
Project SoLi:

Socioeconomics of Light

SoLi: Remote sensing consumer expenditure

- **What? SoLi** is a real-time economic monitoring dashboard using Google's Earth Engine Platform
- **How?** SoLi is trained using Tensor Flow to associate spending patterns & night lights
- **Why?** SoLi offers real-time monitoring for policy decisions
- What is **so big**? It integrates Google's ***Tensor Flow*** and ***Earth Engine*** offering **scale**

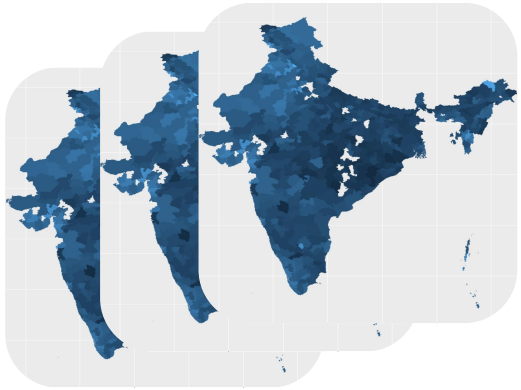
Expenditure Survey (2014)



TensorFlow

VIIRS Night-time (2014)

Estimated MPCE (2014-current)



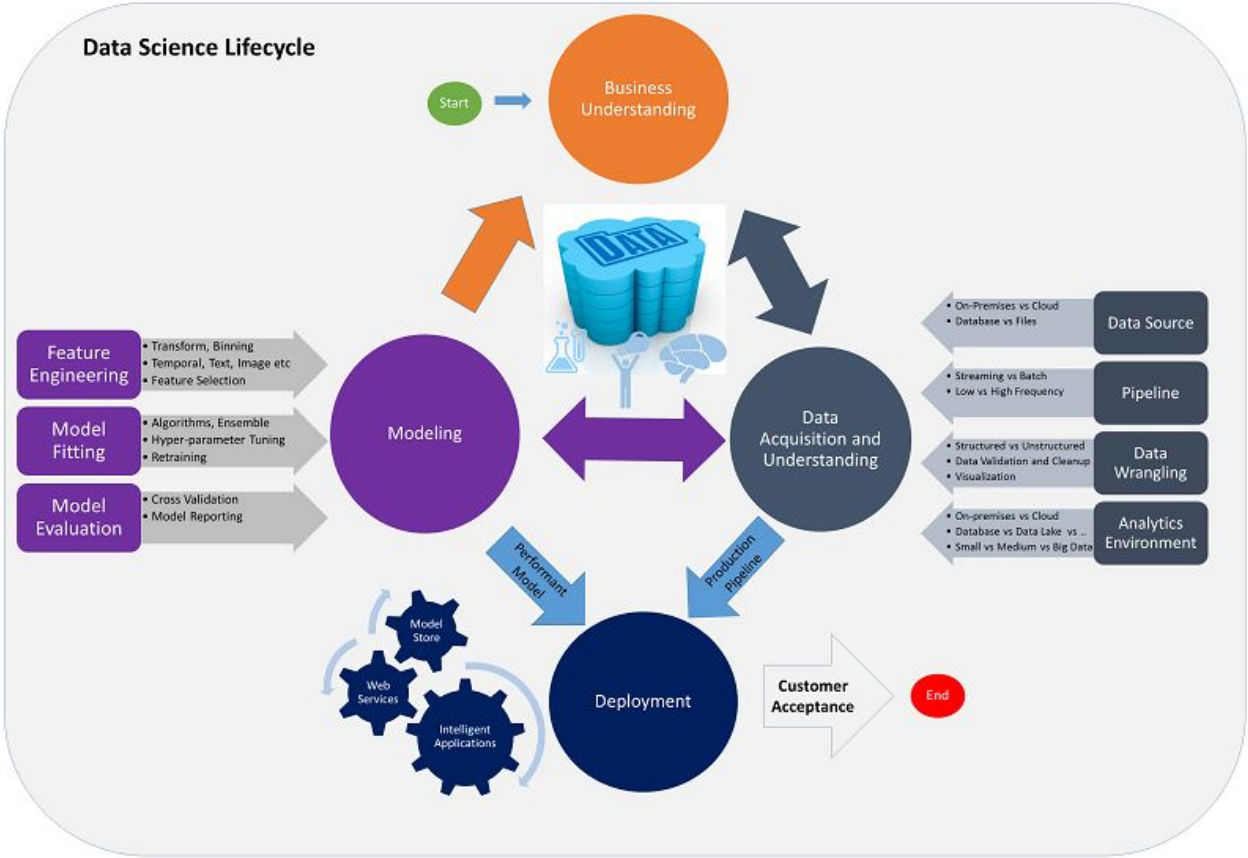
TRANSFER LEARNING

NOWCASTING

Our Data Processing Solution:

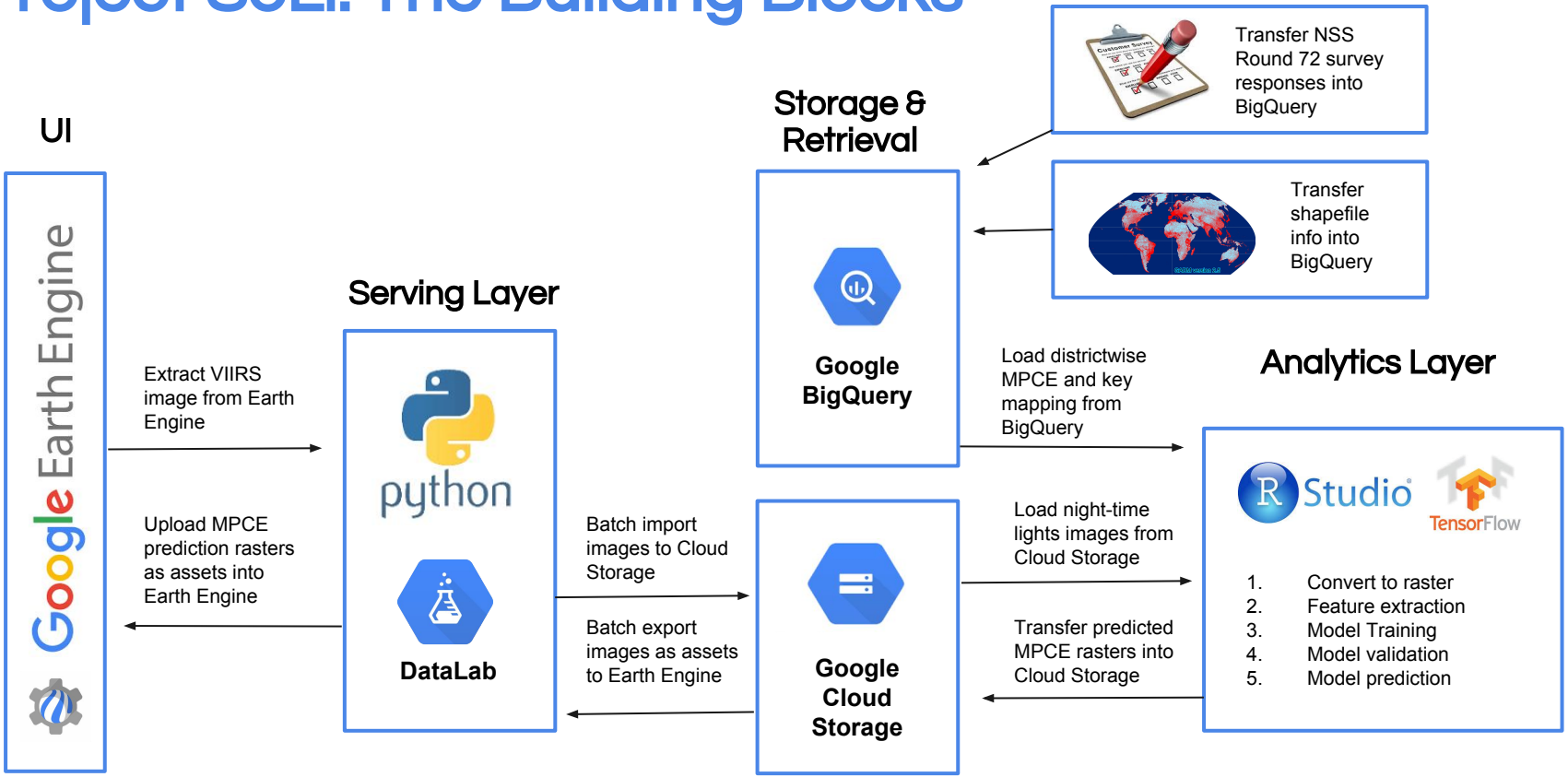
Data & Computation

Typical Data Science WorkFlow



Project SoLi: The Building Blocks

Project SoLi: The Building Blocks



The Data: Three sources, three keys!

National Sample Survey Office

- unit-level data on consumer expenditure
- **Structured**
- Frequency: 3-4 years
- Latest year: 2014

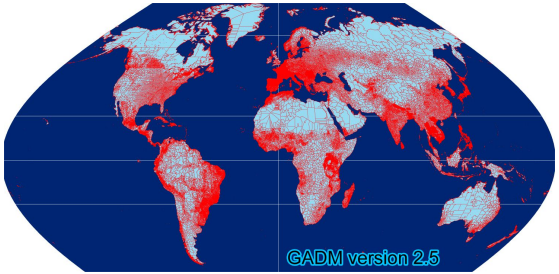
VIIRS Satellite Imagery

Google Earth Engine

- Visible Infrared Imaging Radiometer Suite
- **Semi-structured**
- Frequency: Monthly (2 month lag)
- Latest month: September 2017

GADM (UCB) Shapefiles

Berkeley
UNIVERSITY OF CALIPORNIA



- Global Administrative Areas
- **Structured**
- Developed by Robert Hijman (UCB)
- Looks to map regional boundaries

The Data: Three sources, three keys!

National Sample Survey Office

- **Structured:** Yes
- **Size:** 200 MB
- **Velocity**
  - **Sink Latency:** Low
  - **Source Latency:** Low
- **Quality:** Medium
- **Complete:** Incomplete

VIIRS Satellite Imagery

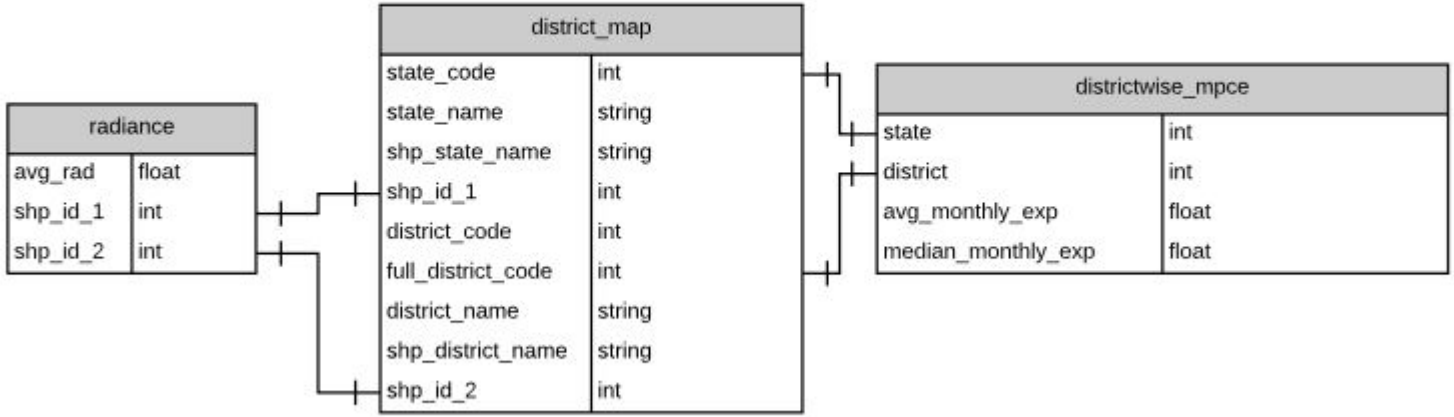
- **Structure:** Semi-structured
- **Size:** 50 MB/image
- **Velocity**
  - **Sink Latency:** NA
  - **Source Latency:** Low
- **Quality:** High
- **Complete:** Incomplete

GADM (UCB) Shapefiles

- **Structure:** Structured
- **Size:** 50 MB
- **Velocity**
  - **Sink Latency:** NA
  - **Source Latency:** NA
- **Quality:** Low (*)
- **Complete:** Incomplete

Data Structure - E-R Diagram

- **Transform:** unit-level information into district-wise mean and median values through aggregation
- **Key mapping:** connecting three different sources and their respective naming conventions for districts



Computation: Serving and Analytics

- **Serving Layer:** communicates with Earth Engine & Cloud Storage using a Datalab Docker Container
- **Analytics Layer:** communicates with Cloud Storage, BigQuery and Tensor Flow using R server

Serving Layer

- **Selectivity:** Low
- **Processing Time:** Medium
- **Query Execution:** Low
- **Precision:** Low
- **Joins:** Medium
- **Aggregation:** Incomplete

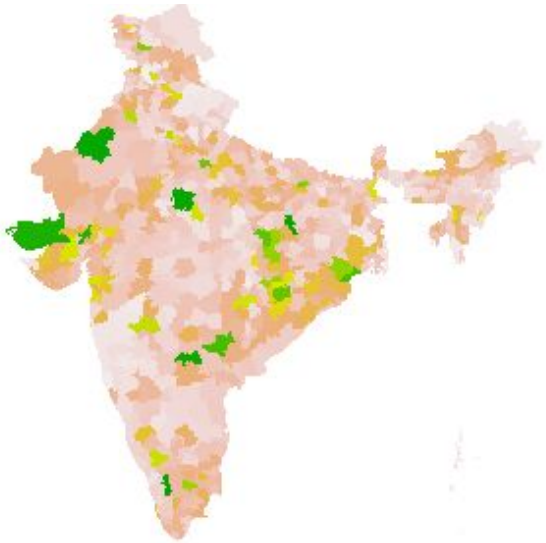
Analytics Layer

- **Selectivity:** High
- **Processing Time:** Long
- **Query Execution:** Low
- **Precision:** Approximate
- **Joins:** Medium
- **Aggregation:** Advanced

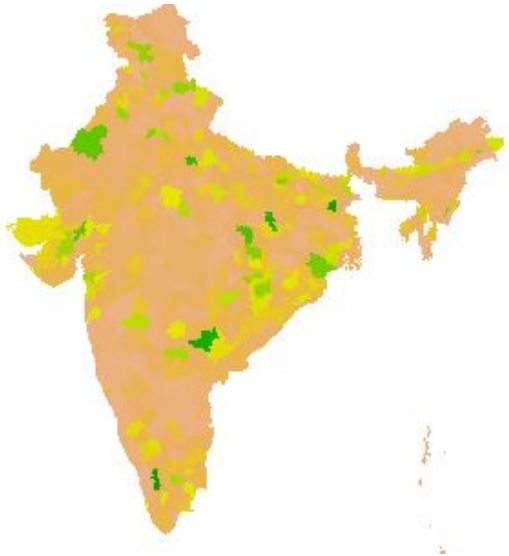
Nowcasting Illustration: estimated MPCE 2015

- **Output:** Geotiff files uploaded as assets into Google's Earth Engine Platform
- **Nowcasting:** incorporating variances in space and time across 642 districts in India

2014

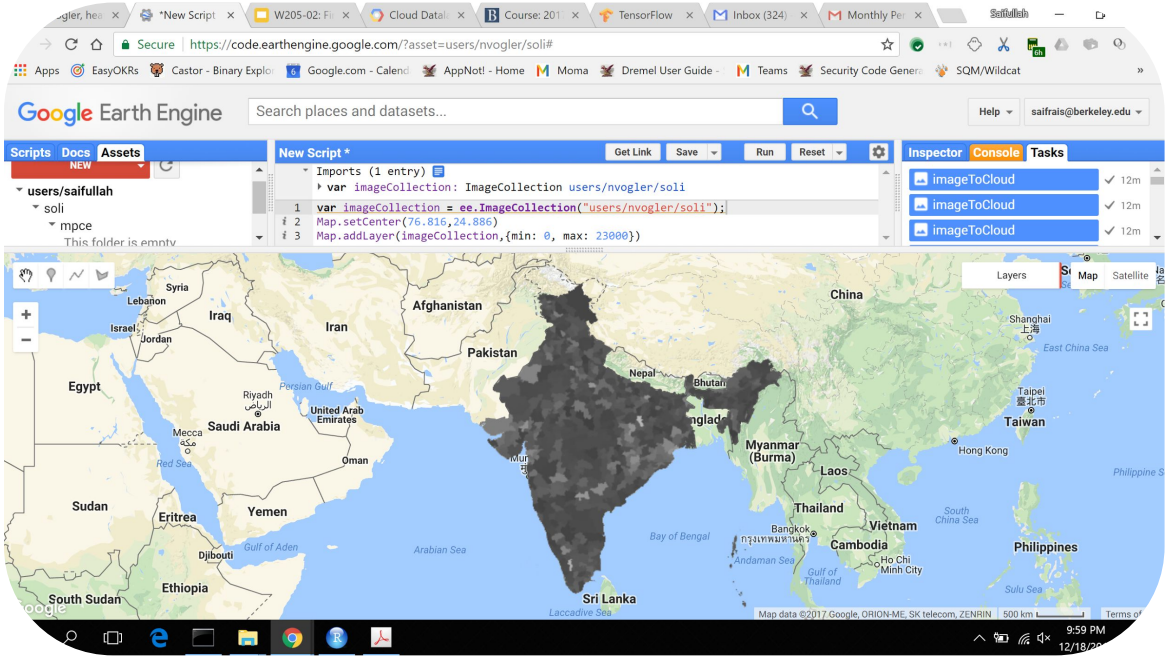


2015



Real-time availability for all

- **Image Collection @** users/nvogler/soli

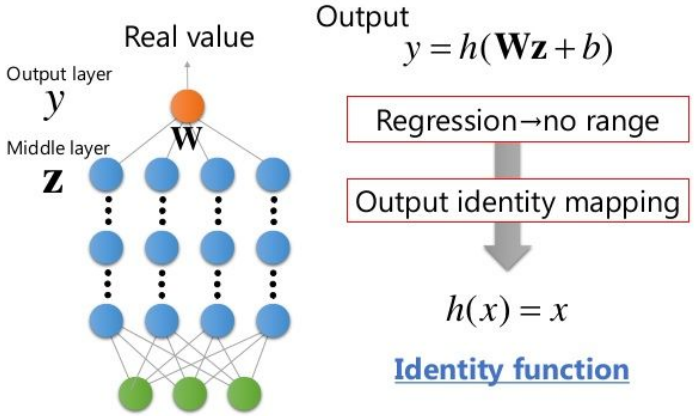


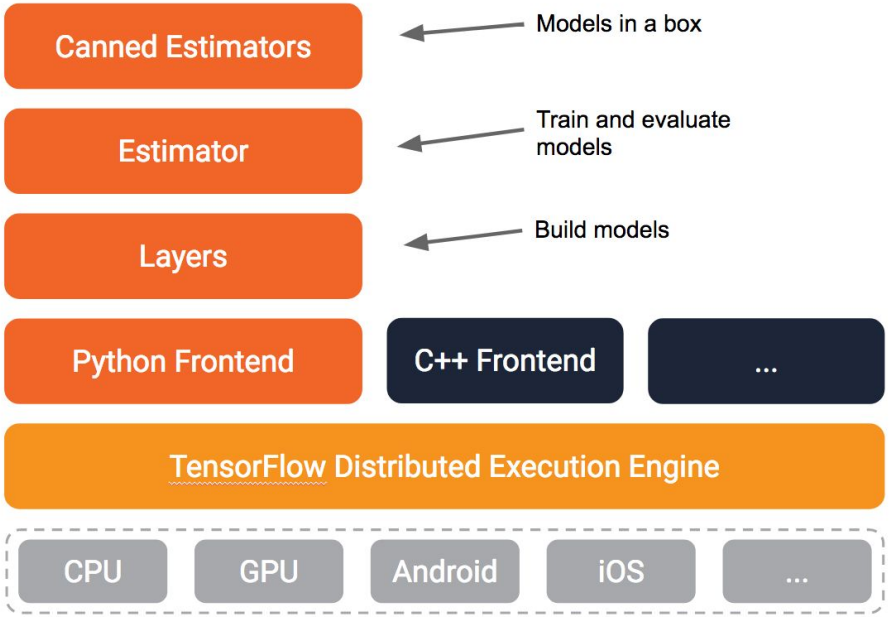
Training the Model: with Tensor Flow Estimators

- **What?** It is an open source software library for numerical computation using data flow graphs
- **Why?** It offers the flexibility to scale and deploy computation across servers with an API
- **Complex?** Offers an R interface to Tensor Flow Estimators, a high-level API like scikit learn

Deep Neural Network Regressor

Regression by DNN





Tensor Flow: Finding Complex Patterns in Images

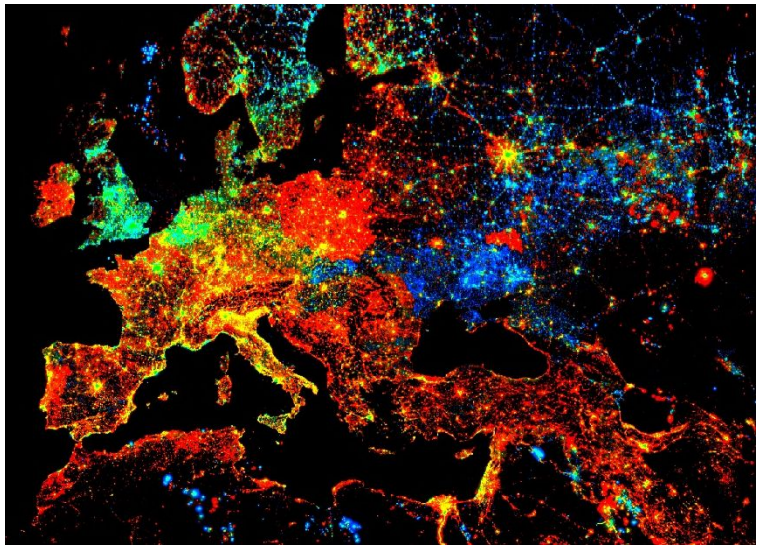
- **MPCE variation:** how much temporal variation can be explained it using satellite imagery?
- **Train:** a Deep Neural Network regressor using satellite imagery and 2014 NSSO survey
- **Predict:** estimate temporal variations in monthly expenditure

Monthly Per Capita Expenditure (MPCE)

MODEL

TensorFlow

VIIRS SATELLITE DATA



Analytics Layer: Model Performance

- **Training:** Currently trained for ~30,000 steps.
- **Feature set:** merely comprise of 16 descriptive statistics of average radiance in a district
- **Explanatory Power:** best model explained ~35% of variance before cross-validation deteriorated.

Explanatory Power of Model

```
raph events.  Overwriting the graph with the newest event.
W1218 11:24:42.420541 Reloader plugin_event_accumulator.py:311] Found more than one metagraph event p
W1218 11:24:42.578456 Reloader plugin_event_accumulator.py:303] Found more than one graph event per r
raph events.  Overwriting the graph with the newest event.
W1218 11:24:42.579092 Reloader plugin_event_accumulator.py:311] Found more than one metagraph event p
Started TensorBoard at http://127.0.0.1:3329
# A tibble: 1 x 3
  average_loss global_step      loss
         <dbl>       <dbl>     <dbl>
1      6197208       40000 762267712

Call:
lm(formula = unlist(res) ~ data[, response()])

Residuals:
    Min      1Q  Median      3Q     Max
-3397.3  -600.6  -114.0   369.0  8500.5

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.566e+03  1.481e+02   37.59   <2e-16 ***
data[, response()] 2.876e-01  1.803e-02   15.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1124 on 618 degrees of freedom
Multiple R-squared:  0.2918,    Adjusted R-squared:  0.2906
F-statistic: 254.6 on 1 and 618 DF,  p-value: < 2.2e-16

Called from: build_regressor(data, response = "median_mpce", model_loc = model_loc,
    save_loc = save_loc, no_epochs = no_epochs, no_steps = no_steps)
Browse[1]>
```

Tensor Board Monitoring

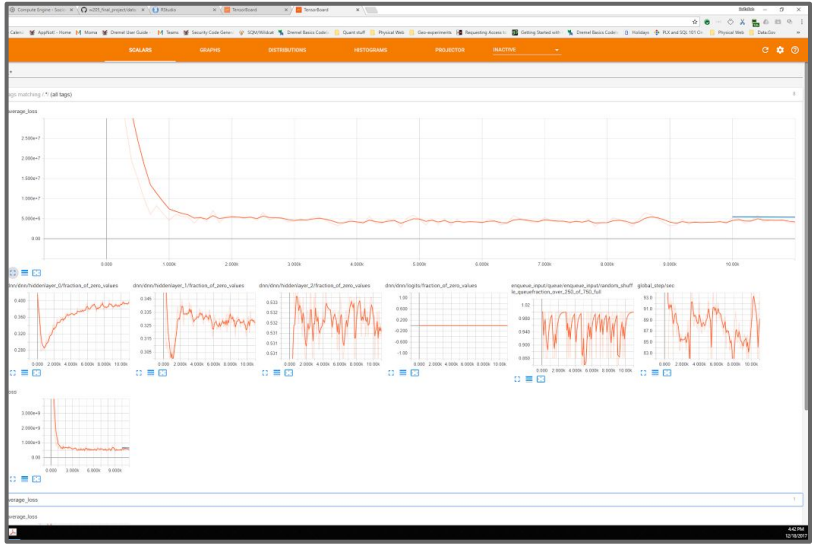
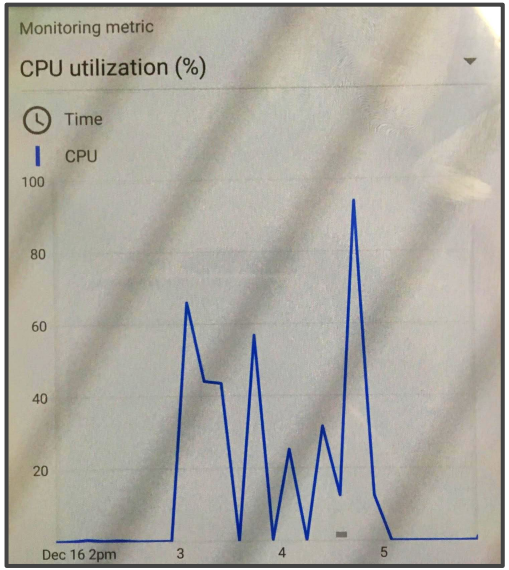


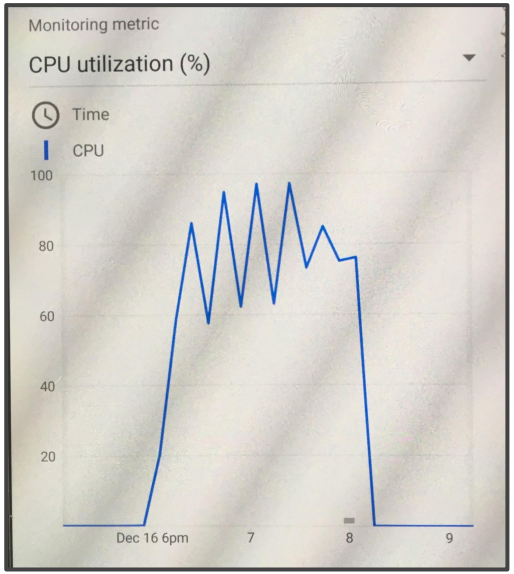
Illustration: compromise on output image precision

- **Traditionally single-core processes:** parallel processing can improve performance by 7X
- **Parallel Processing for aggregation:** Raster aggregation runs across 7 cores (2 mins/image)
- **Rasterizing predictions still single-core activity:** therefore drop in computation utilization
- **Speed over precision:** given district-wise analysis, we reduced the dimensions of the output image

A) Join bottleneck

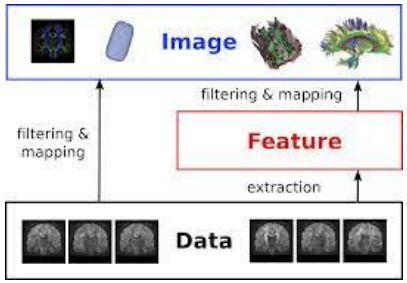


B) Compromise Precision

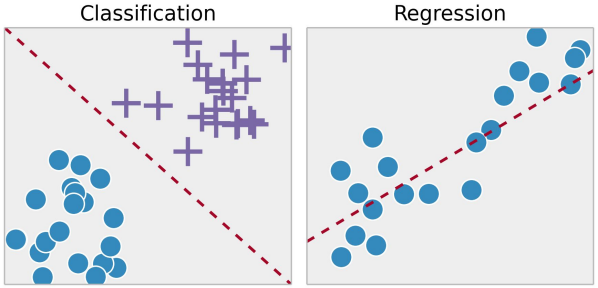


Further Work: Train-Test-Repeat

Feature Extraction



Classification vs Regression



Asset ownership

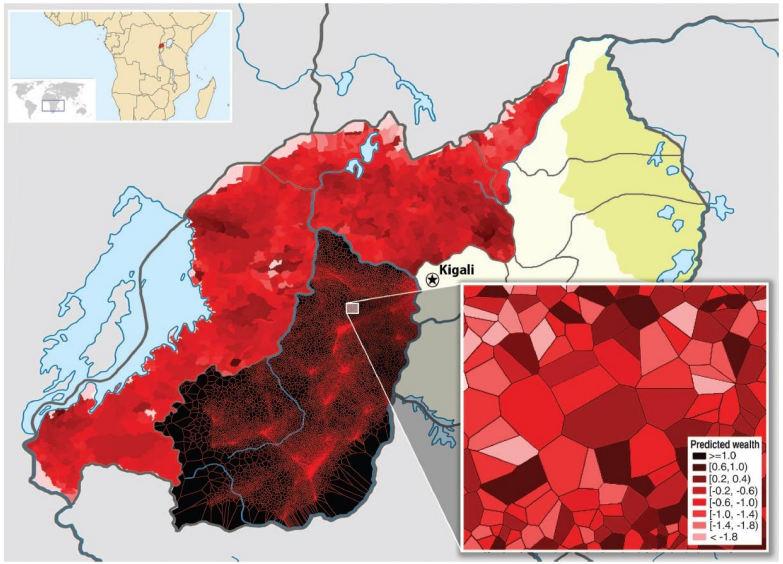
- **Feature Extraction**
  - Total number of features: 16
  - Possible Convolution Neural Network
- **Classification vs. Regression**
  - Current Model: DNN Regressor
  - Literature review: Development economics have generally converted response into categories
  - Potential: Convert continuous MPCE into categories
- **Asset ownership vs. MPCE**
  - Night Light Development Index: surrogate for Gini coefficients
  - Fighting Poverty with Data
  - Combining Satellite Imagery with Machine Learning to fight poverty

DATA SCIENCE FOR SOCIAL GOOD

Current State Of The Art

Fighting Poverty with Data: UC Berkeley

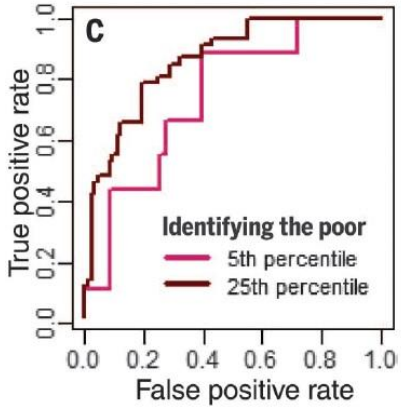
Identify Poverty and Wealth from call records.

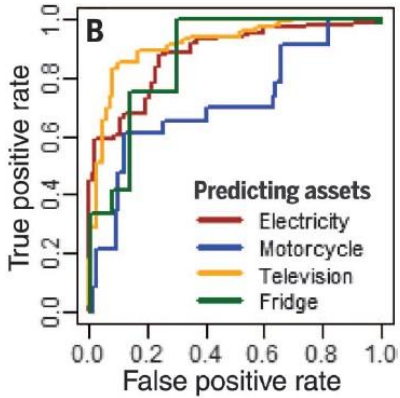


Source: Blumenstock & all (2015), Fighting Poverty with Data

That this may prove fruitful is motivated by the fact that mobile phone data capture rich information, not only on the frequency and timing of communication events (*12*) but also reflecting the intricate structure of an individual's social network (*13*, *14*), patterns of travel and location choice (*15*–*17*), and histories of consumption and expenditure. Regionally aggregated measures of phone penetration and use have also been shown

Classification performance (using AUC)





New State-Of-The-Art: *using real-world footprints*

INTENT-BASED FEATURES

Travel-related searches are intent, not action

ACTION-BASED FEATURES

Foursquare check-ins showcase citywise flows

STATIC SIGNALS

Possession of a passport is a static signal

DYNAMIC SIGNAL

Immigration Stamps are constantly evolving

Questions