6. Dashboards and Data Pipelines

Jonathan Hersh (Chapman University Argyros School of Business)

12/8/2021

Outline

1. Data Lakes and Data Pipelines
2. What is Open Data?
3. Dashboards for Model Displays

Why Do Public Policy Data Projects Fail?

1. Models can only be as good as your data pipeline (Bad data = bad model)
2. Build a consistent data lake first (excel files are bad)
3. All data and models should be an API (i.e. Application Programming Interface or structured way to access data)

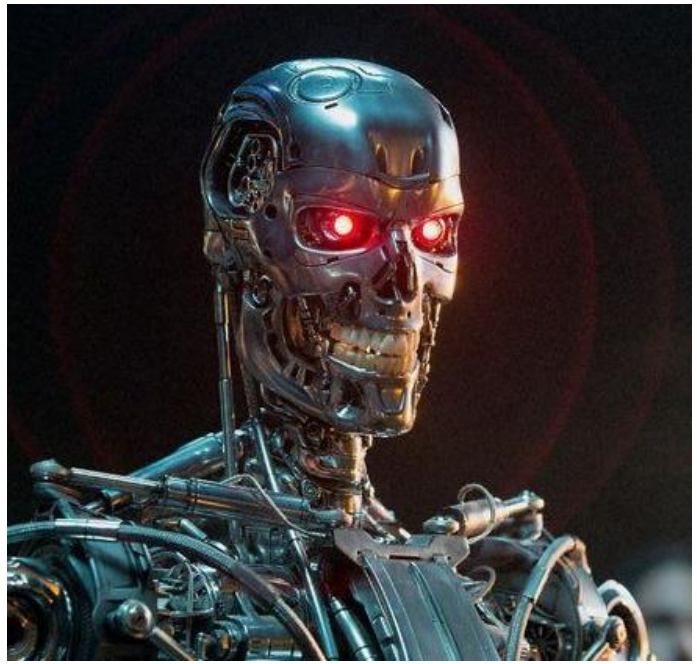
Class of Cultures Within Public Policy Organizations

- Development organizations: academic culture, focused on publications, journals. Slow, careful, hierarchical.

Unavoidable fact: development organizations produce knowledge products. Tech leads in how to organize production for knowledge products

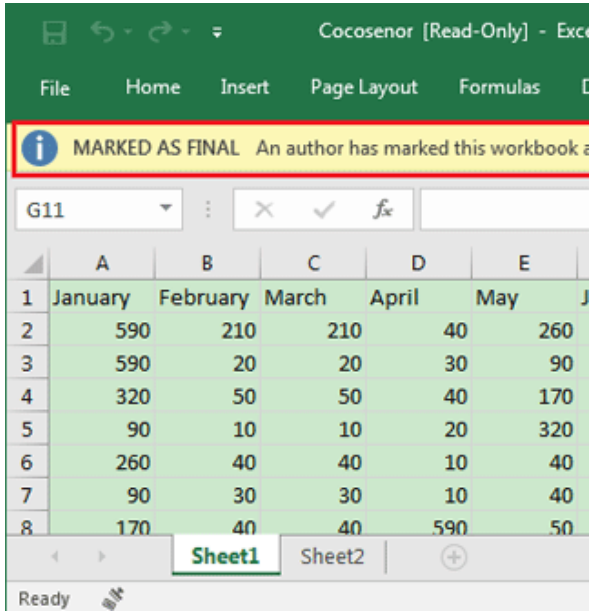
- Tech culture: build fast, design and revise. Experimental and see “what works.” Flat organization structure.

Perceptions of capabilities of Artificial Intelligence

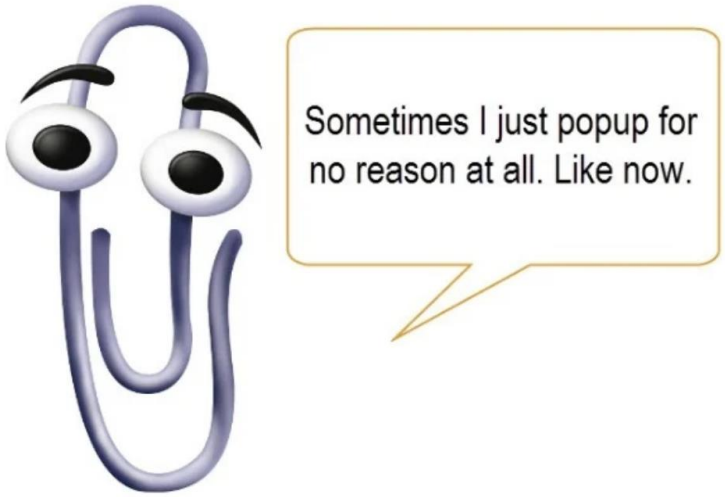


+

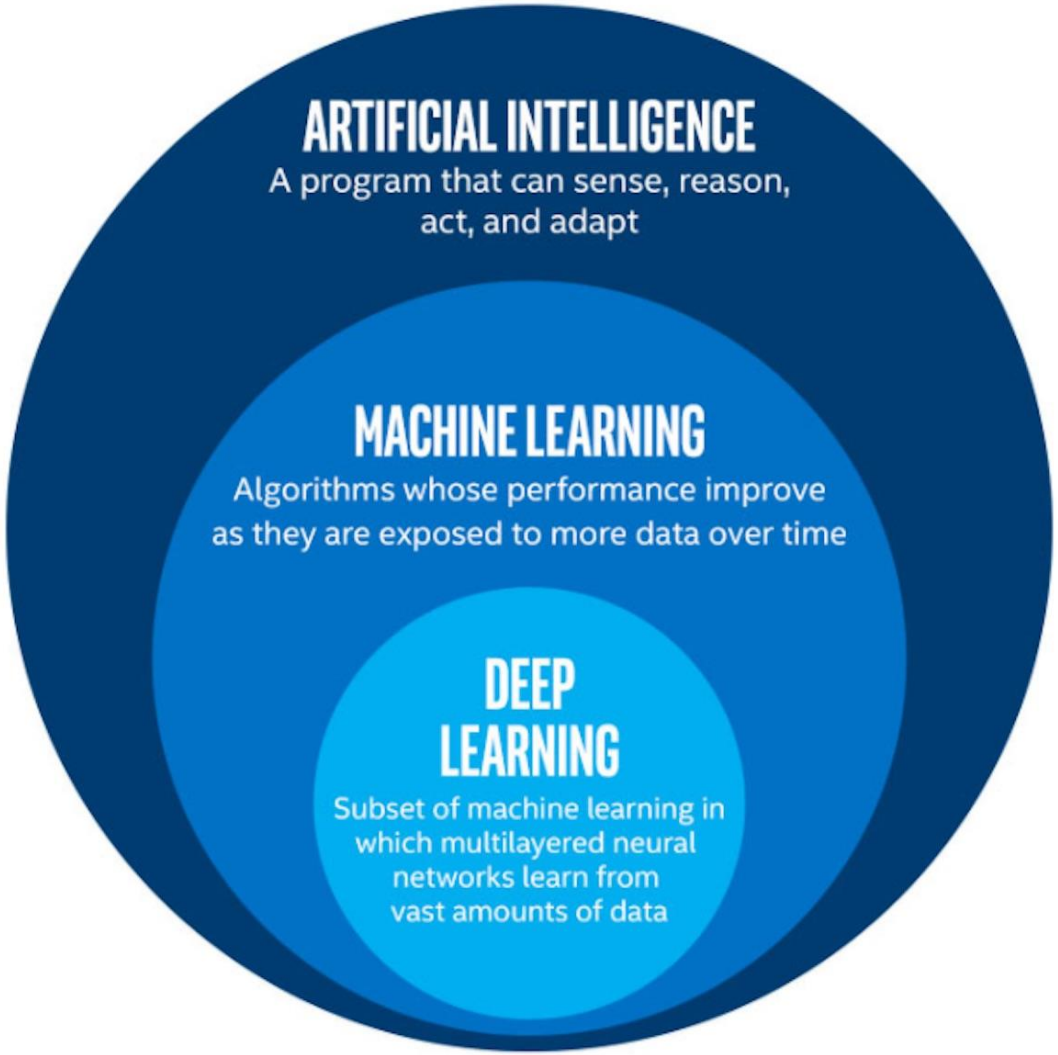
Existing data infrastructure



=

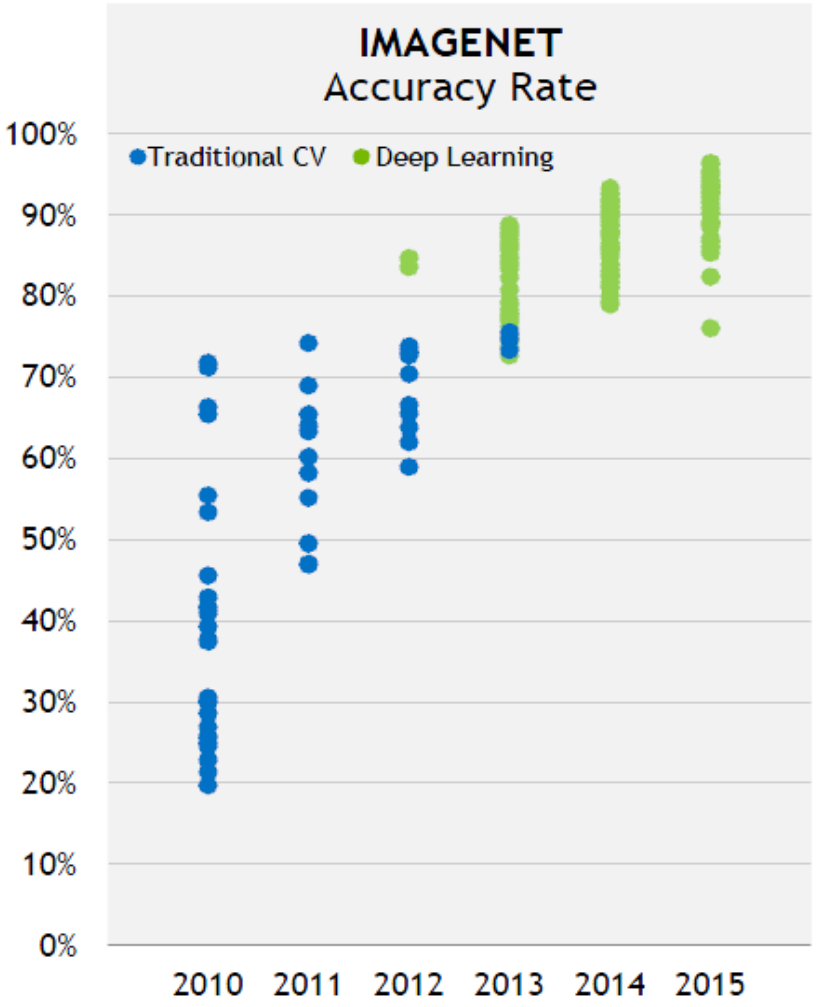


Artificial Intelligence vs Machine Learning vs Deep Learning



"**The deep in deep learning isn't a reference to any kind of deeper understanding achieved by the approach**; rather, it stands for this idea of successive layers of representations" – Francois Chollet

One Reason for The Deep Learning Hype: Images



- Before deep learning, computers couldn't recognize objects from images.
- In 2012 researchers first used a deep learning model on the ImageNet competition, which tests an AI to recognize objects in images
- But ImageNet had 1M+ images!
- Translation: models and data are close complements

Modern Data Pipelines

Raw Data

amazon S3

Raw Data Transformation

python

R

Analysis Dataset

amazon S3

Model Output Dataset

amazon S3

API = structured way to access your model

API

tableau Shiny plotly | Dash

Data Viz Dashboard

Report Few Will Read

python

R

Variable Creation

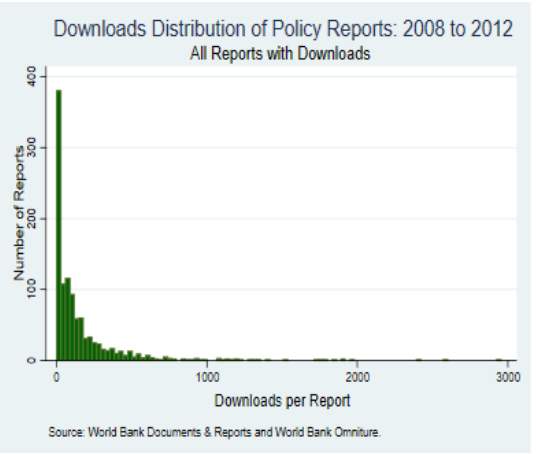
python

R

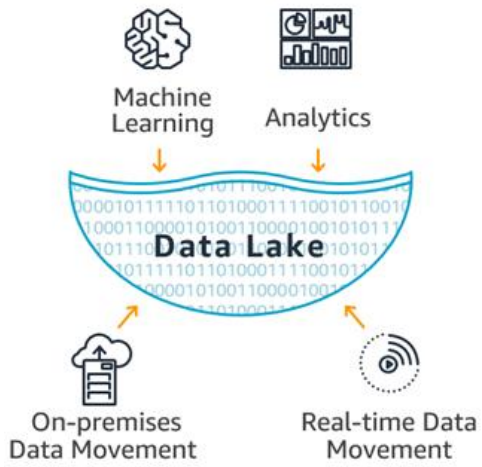
Modeling & Exploratory Data Analysis

Data Lake: a centralized repository that stores all your structured and unstructured data on demand and accessible by everyone

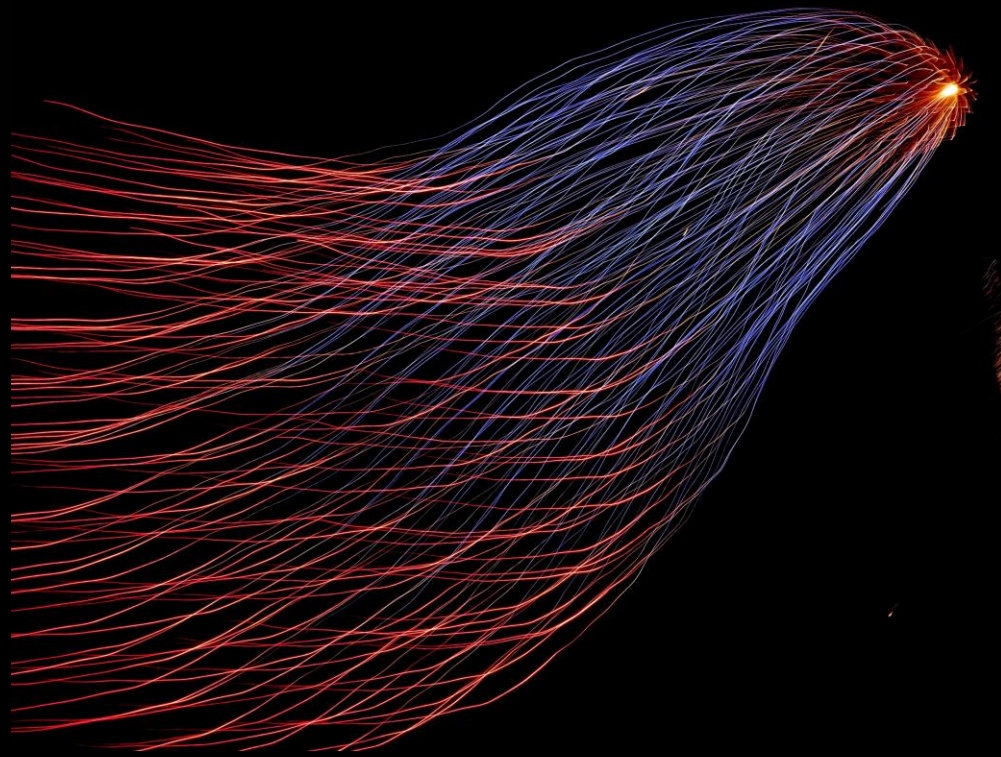
Most economists only consider this portion of the data pipeline!



Many Options of Data Lakes to Choose From



Questions at This Point

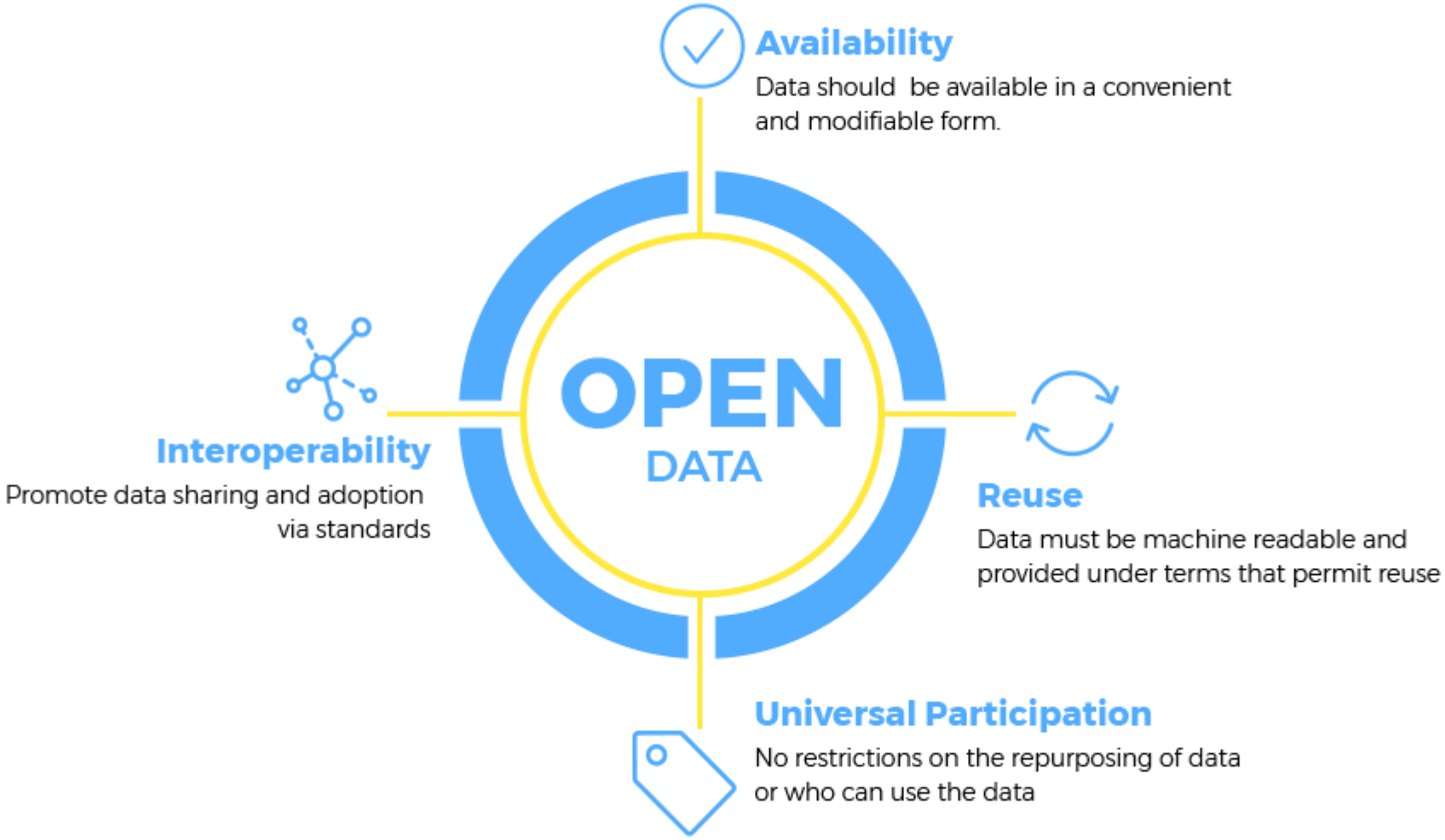


- What's best data lake?
  - What's the best food when you're starving?
  - Cloud for sure, open source if you can
- What programming languages do I need to learn?
  - None really. Version control (git). R and Python will make your life easier. SQL variants (or just dbplyr)
- **Theme: Have a consistent data strategy and ensure everyone abides by it**

Outline

1. Data Lakes and Data Pipelines
2. What is Open Data?
3. Dashboards for Model Displays

What is Open Data?



Why is Open Data Important?

INFORMATION TECHNOLOGY FOR DEVELOPMENT
2021, VOL. 27, NO. 2, 263–292
https://doi.org/10.1080/02681102.2020.1811945

Open data for algorithms: mapping poverty in Belize using open satellite derived features and machine learning*

Jonathan Hersh [a], Ryan Engstrom[b] and Michael Mann[b]

[a]Argyros School of Business, Chapman University, Orange, CA, USA; [b]Department of Geography, George Washington University, Washington DC, USA

ABSTRACT
Several methods have been proposed for using satellite imagery to model poverty. These include poverty mapping using convolutional neural networks applied either directly or using transfer learning to high resolution satellite images, or combinations of methods that combine satellite imagery with standard methods. However, these methods require proprietary imagery which, given their cost and infrequent acquisition, may render these advances impractical for most applications. The authors investigate how satellite-derived poverty maps may improve when incorporating features derived from Sentinel-2 and MODIS imagery, which are both open-source and freely and readily available. The authors estimate a poverty map for Belize which incorporates spatial and time series features derived from these sensors, with and without survey derived variables. They document an 8% percent improvement in model performance when including these satellite features and conclude by arguing that Open Data for Development should include open data pipelines where possible.

imagery in perpetuity. In comparison, a statistical agency that incorporates proprietary data into their statistical pipeline opens themselves to price gouging as proprietary data providers have pricing power due to 'lock-in' type effects (Arthur, 1989). It is possible that even with competition among data providers, any surplus from using Big Data at statistical agencies may eventually be captured by proprietary data providers because of lock-in effects due to the difficulty of moving from established data pipelines. Thus, it is crucial to consider open-source alternatives to proprietary data providers.

https://publications.iadb.org/en/mapping-income-poverty-in-belize-using-satellite-features-and-machine-learning

https://www.tandfonline.com/doi/full/10.1080/02681102.2020.1811945?scroll=top&needAccess=true

Outline

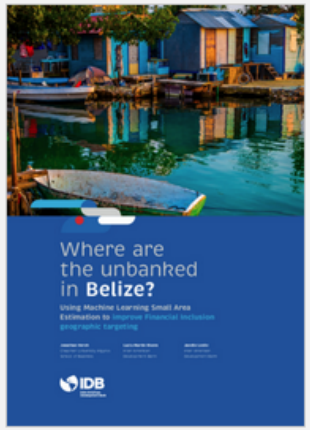
1. Data Lakes and Data Pipelines
2. What is Open Data?
3. Dashboards for Model Displays

Outline

1. Data Lakes and Data Pipelines
2. What is Open Data?
3. **Dashboards for Model Displays**

Example Dashboard Project: Financial Inclusion

Where are the Unbanked in Belize?: Using Machine Learning Small Area Estimation to Improve Financial Inclusion Geographic Targeting



AUTHOR: Hersh, Jonathan; Martin Rivero, Lucia; Leslie, Janelle

DATE: Jul 2021

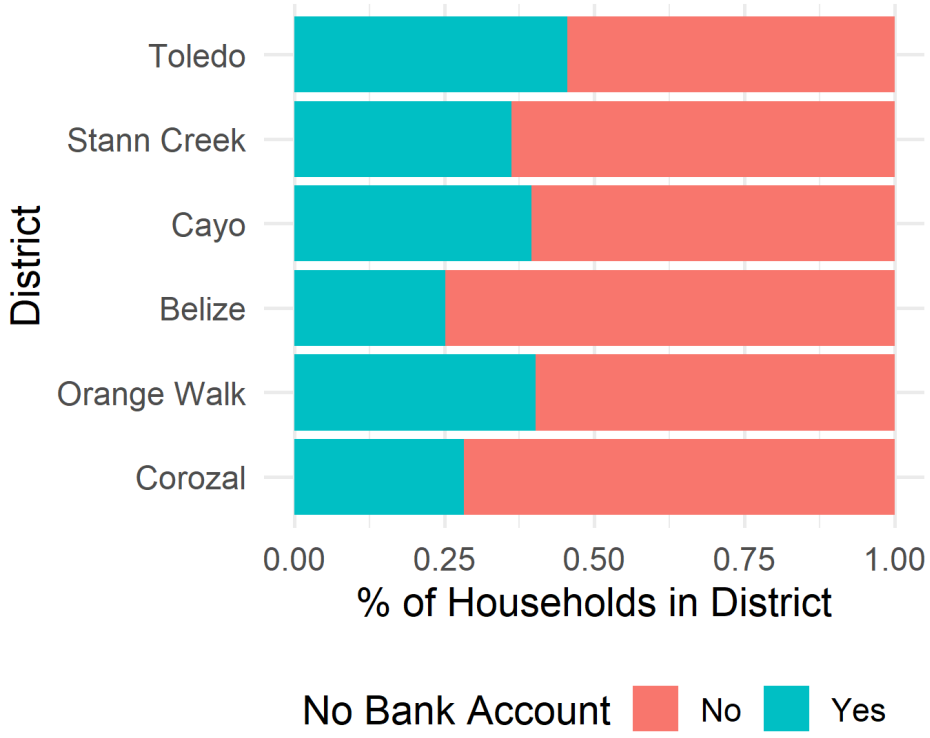
DOWNLOAD: English (353 downloads)

DOI: http://dx.doi.org/10.18235/0003381

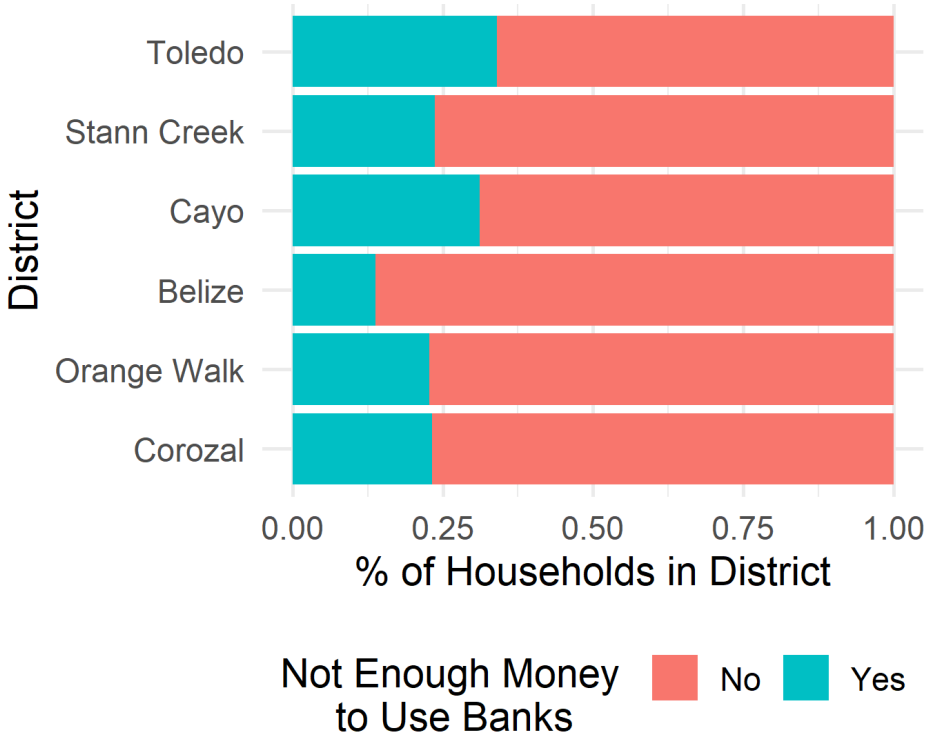
This study aims to contribute to the efficient and effective implementation of Belize's National Financial Inclusion Strategy (NFIS) that was launched by the Central Bank of Belize in 2019. It employs Machine Learning Based Small Area Estimation to develop granular estimates of Financial Inclusion at the smallest geographical level know as Enumeration Districts (ED) that were previously unavailable for Belize. To gain deeper understanding of the populations financial characteristics at the ED level, we build five measures of access to banking and financial services. Significant clustering of financial inclusion metrics that are not apparent in the district level averages are identified. This study also analyzes the factors that influence the use of financial services and instruments in order to propose appropriate adjustments in the strategies implemented by authorities in each geographical area. Both the spatial distribution of Financial Inclusion indicators and the factors influencing the adoption of financial services shed light on specific recommendations relevant to each of the four Thematic Financial Inclusion Task Forces included in the NFIS.

https://publications.iadb.org/publications/english/document/Where-are-the-Unbanked-in-Belize-Using-Maching-Learning-Small-Area-Estimation-to-Improve-Financial-Inclusion-Geographic-Targeting.pdf

Measures of Financial Inclusion – Unbanked + Barriers to Banking



1) Does anyone in the household have an account at a credit union or a bank? (FI module question 1)



2) Is the reason you do not have an account at a bank or credit union because you don't have enough money to use them? (FI module question 3-F)

Modeling Approach

2010 Belize Census
- 75,000 households

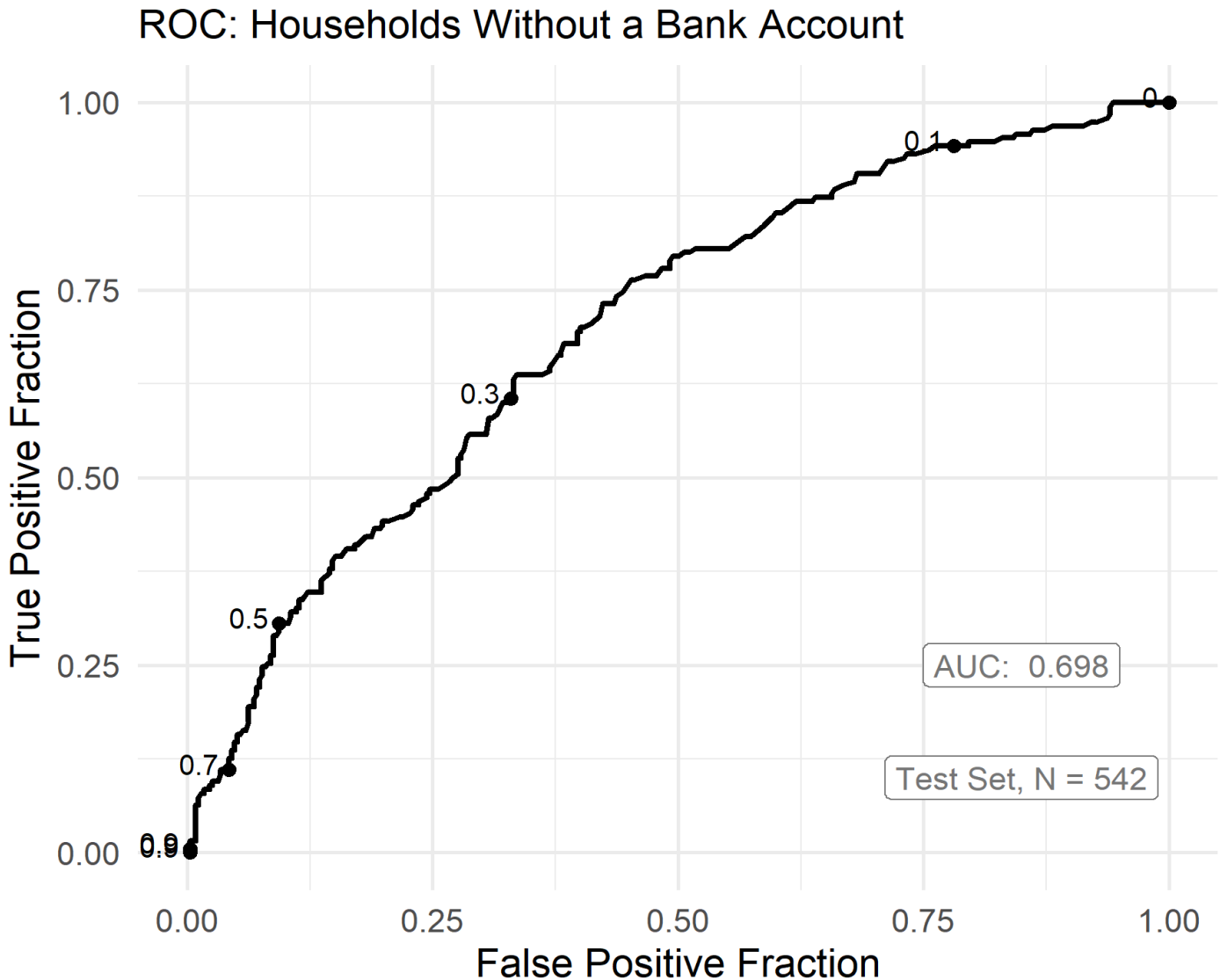
Labor Force Survey
- April 2019 Wave
- 2,216 unique households

Household level outcome to model: One of four Financial Inclusion Question

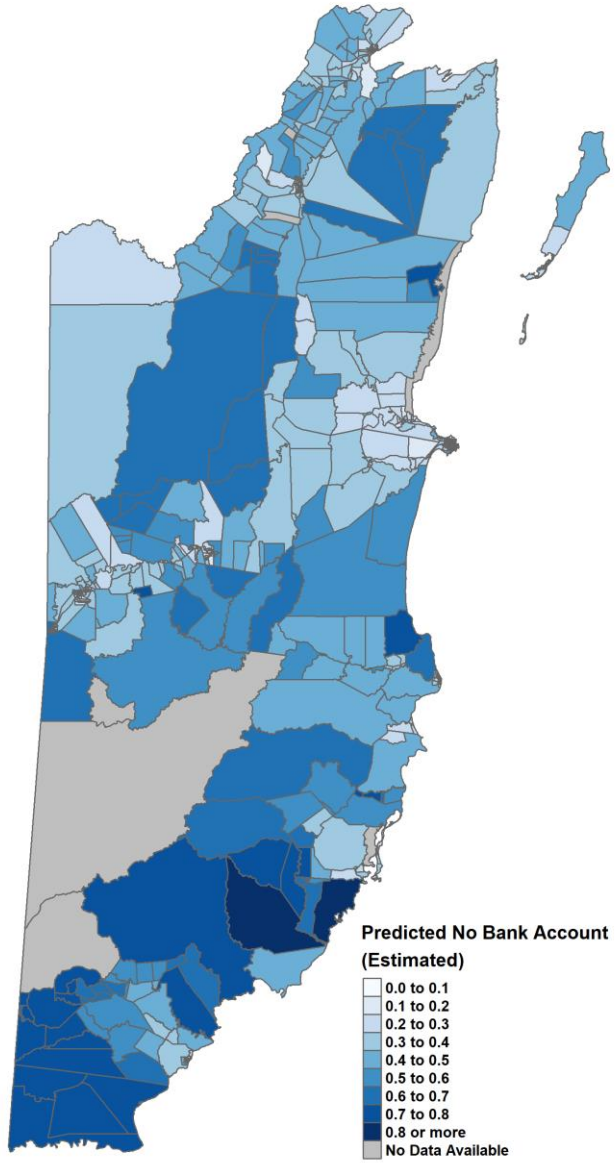
Test/Validation approach
- 75% data training (estimation) sample and 25% test (validation)

Models
- Random Forest each made up of 500 decision trees

Intersection of Census and LFS Surveys had 26 transformed variables


ROC: Households Without a Bank Account
True Positive Fraction
False Positive Fraction
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0
0.1
0.3
0.5
0.7
AUC: 0.698
Test Set, N = 542

Estimated % of Households in ED Without Bank Accounts



Dashboard To Interactively View Results