



# World Bank Loans

**Team Moonstone: Xiaoyan Ding, Jakob Hansen, Shasha Liao, Shuo Xu**

*Data insights from unstructured PDF files*

**<https://github.com/xushuo0/Qarik-Project>**



# Problem and Goals

## PROBLEM

Vast amounts of legacy documents and information are only available as PDFs

- Not an easily machine-readable format

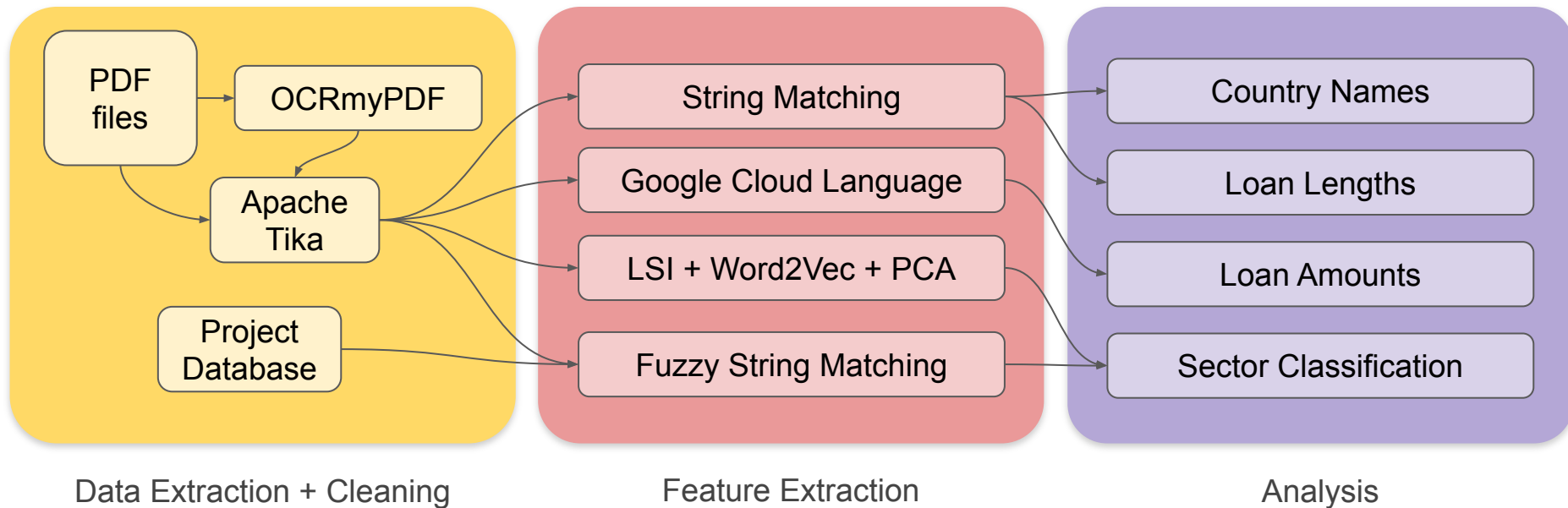
Incorporating PDFs into data science and decision pipelines is an important challenge.

## GOAL

Demonstrate the ability to extract interesting, useful information and insights from PDF files. Do this by using a collection of freely available loan documents from the World Bank.

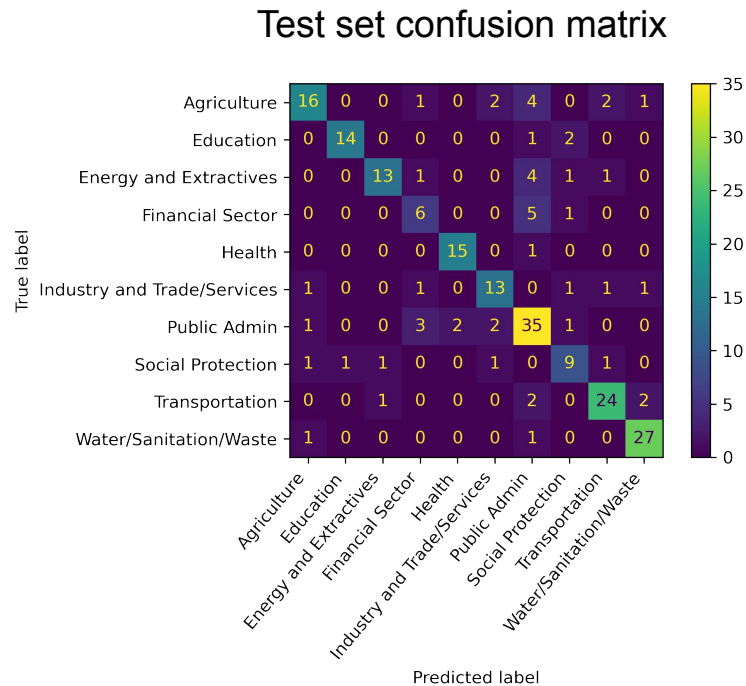


# Our Approach and Teamwork

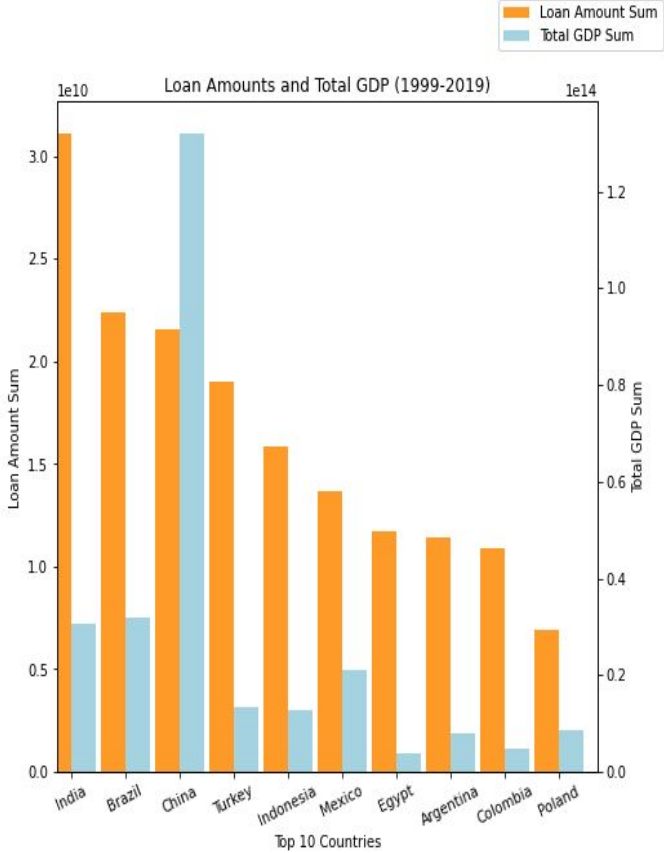
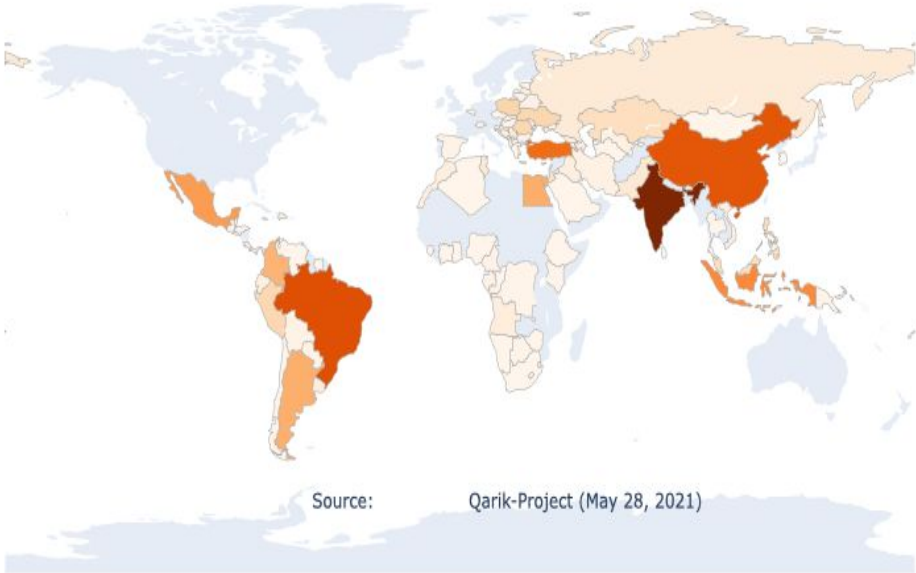


# Economic Sector Classification Results

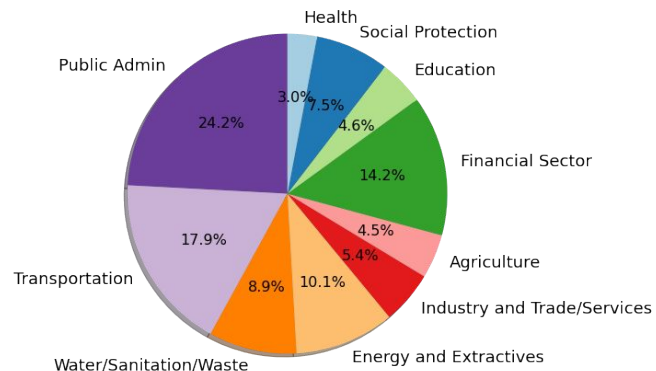
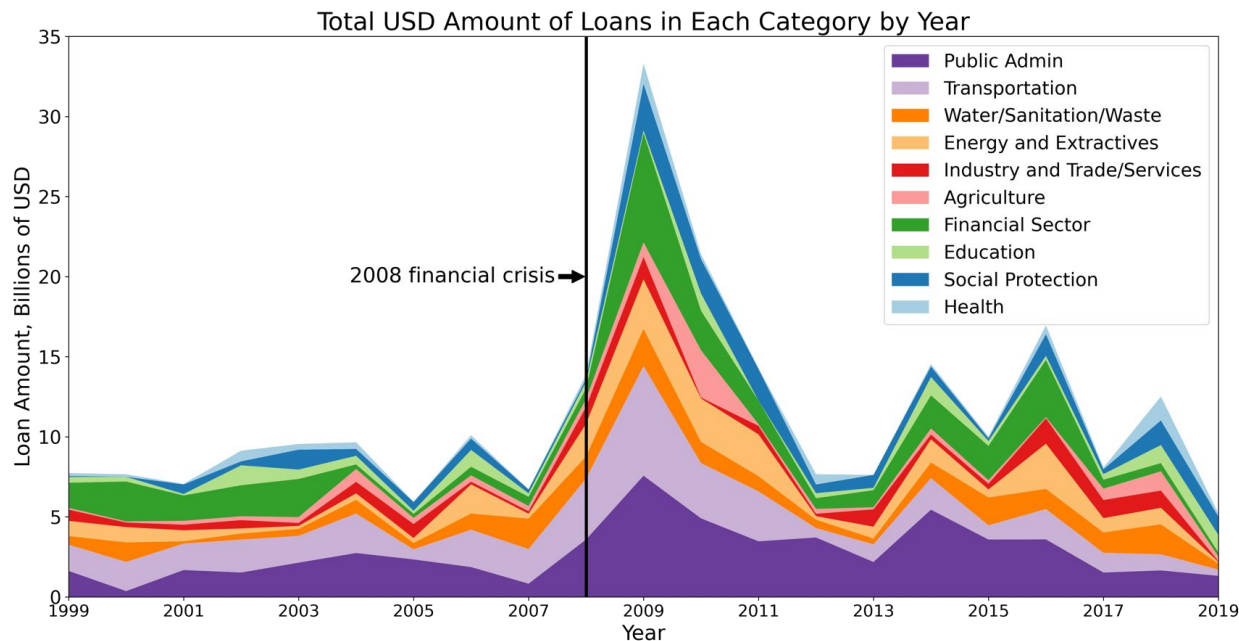
- Unsupervised classification predicts sectors with **47% accuracy** based on project description keywords alone, using similarity of Word2Vec embeddings.
- Supervised classification predicts sectors with **76% accuracy** using an ensemble of Random Forest, Logistic, and kernel SVM models on LSI topic vectors.



World Bank (IBRD) Loans 1999-2019



# Total Loans in Current US Dollars





# Conclusions

- We extracted meaningful information from unstructured data sources.
- We predicted the sectors from the project descriptions.
- We showed that loans are conditioned on repayment ability, not necessarily on need.
- We showed that the World Bank helped developing countries in times of crisis.

# Thank You!

## TEAM 42: Moonstone

<https://github.com/xushuo0/Qarik-Project>

**Xiaoyan Ding**

[dingx@umich.edu](mailto:dingx@umich.edu)

University of  
Michigan

**Jakob Hansen**

[hansen.612@osu.edu](mailto:hansen.612@osu.edu)

The Ohio State  
University

**Shasha Liao**

[sliao7@gatech.edu](mailto:sliao7@gatech.edu)

Georgia Institute of  
Technology

**Shuo Xu**

[xushuo0@gmail.com](mailto:xushuo0@gmail.com)

The Ohio State  
University

