AI Applications in Finance

iCSE Research Day

Ali Hirsa

Industrial Engineering & Operations Research

Data Science Institute

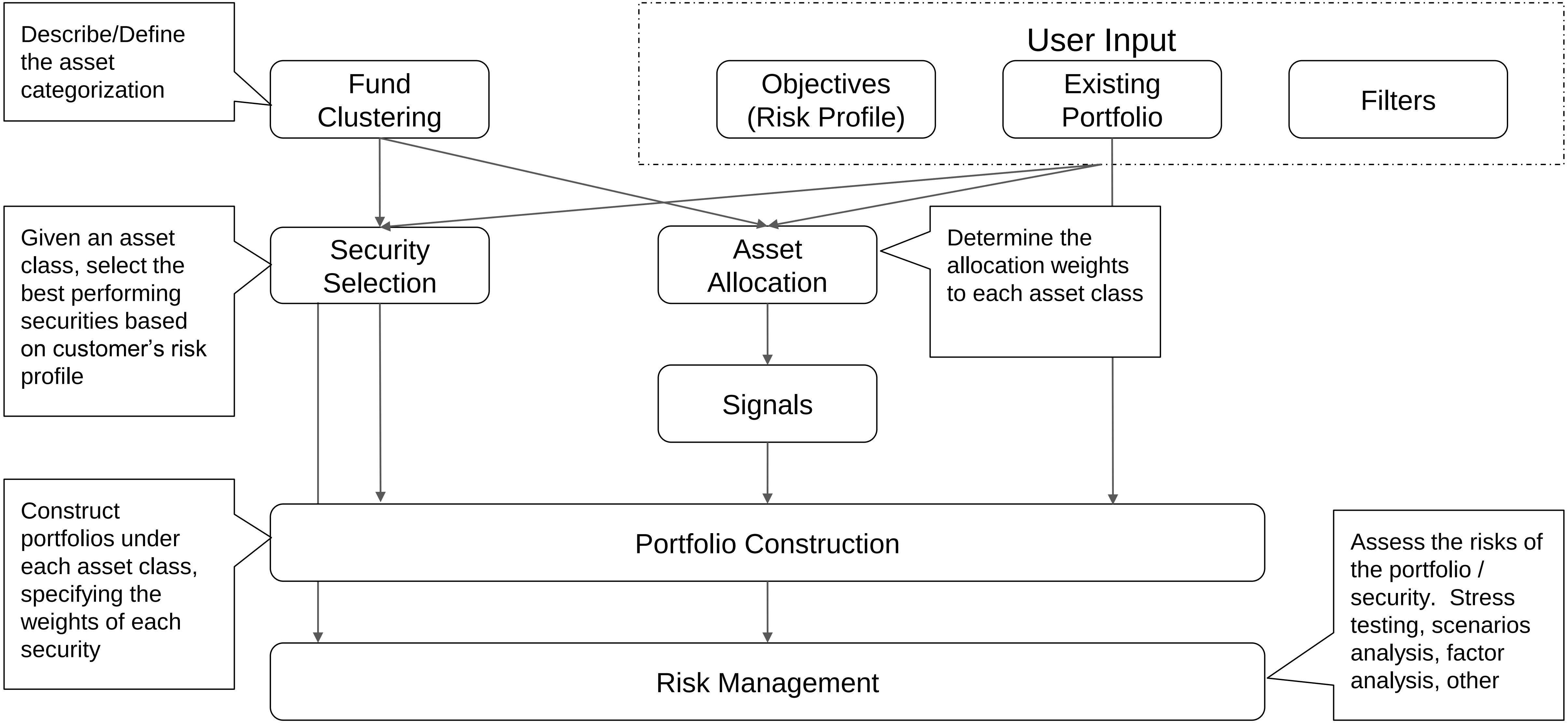
Columbia University

Industry/Research Topics

- Wealth Management
- Private Equity
- Real Estate
- Climate Change
- Fraud Detection
- Interpretability & Adversarial Attacks
- Predicting Market Moves using News Traffic
- Building Classification for Insurance Pricing
- Replication

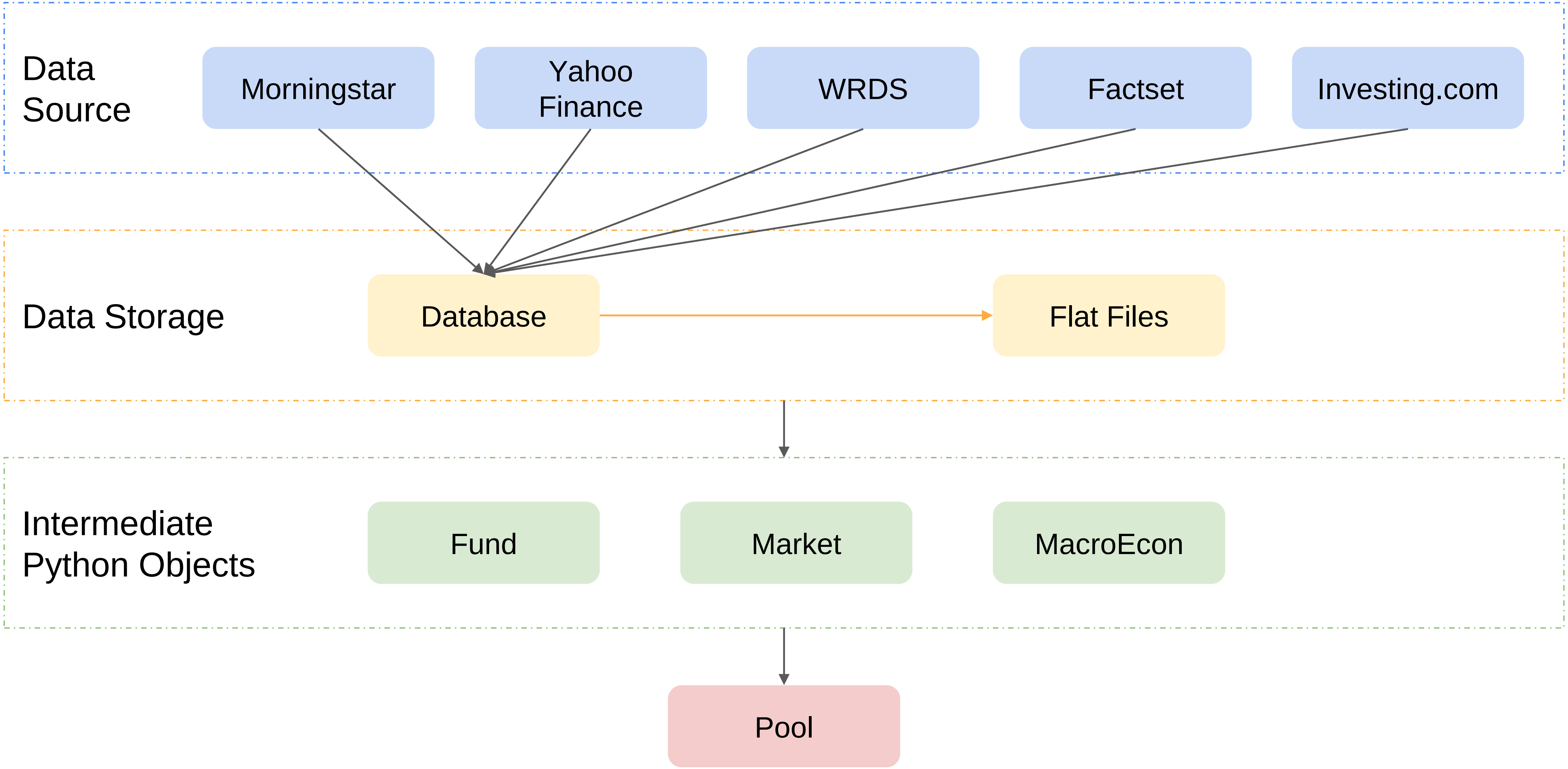
Wealth Management’s Overview

Team members: Miao Wang, Sikun Xu



Data – Overview

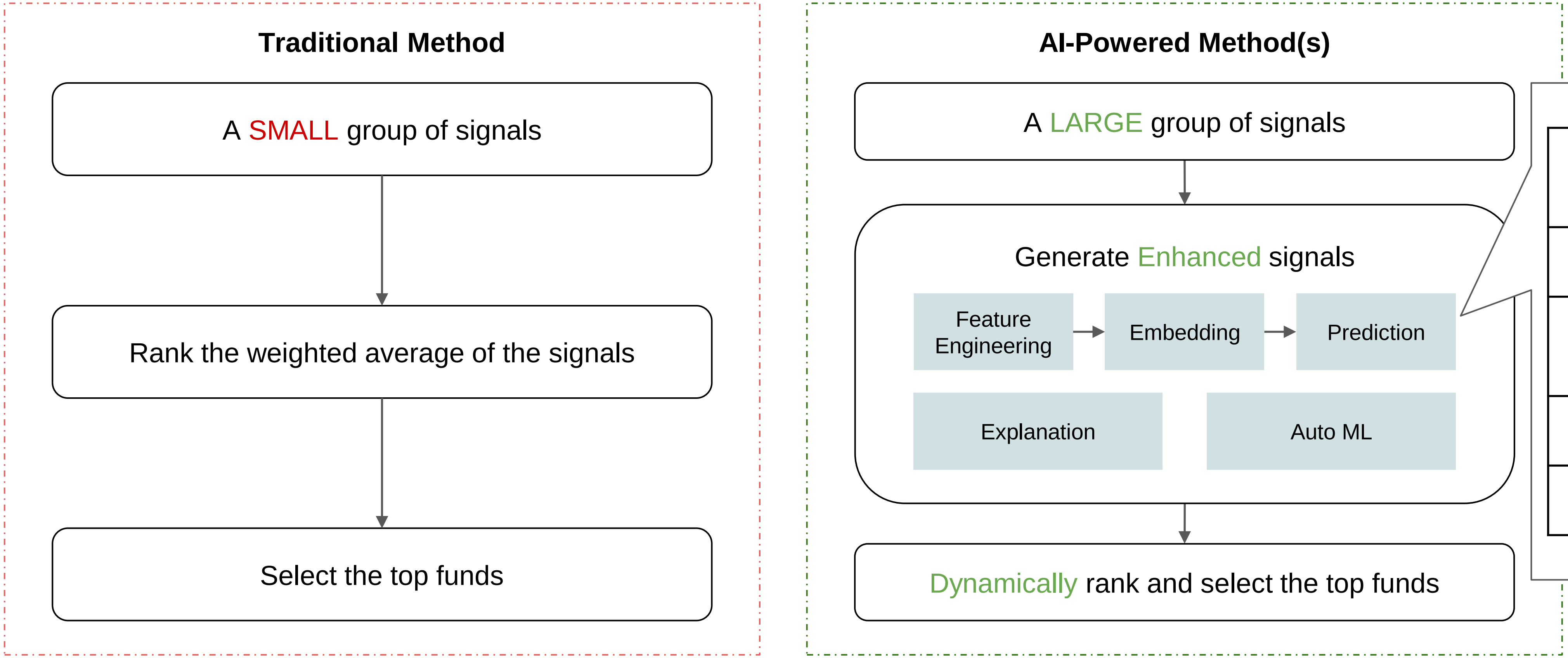
Team members: Miao Wang, Sikun Xu



Security Selection

Team members: Miao Wang, Sikun Xu, Joao Ferraz, Mengyao He

The security selection project aims at designing AI-powered methods to **recommend the top-performing securities** in a given asset category. It utilizes a **large number of features** and deploys various **dynamical models** to make intelligent security selection decisions.

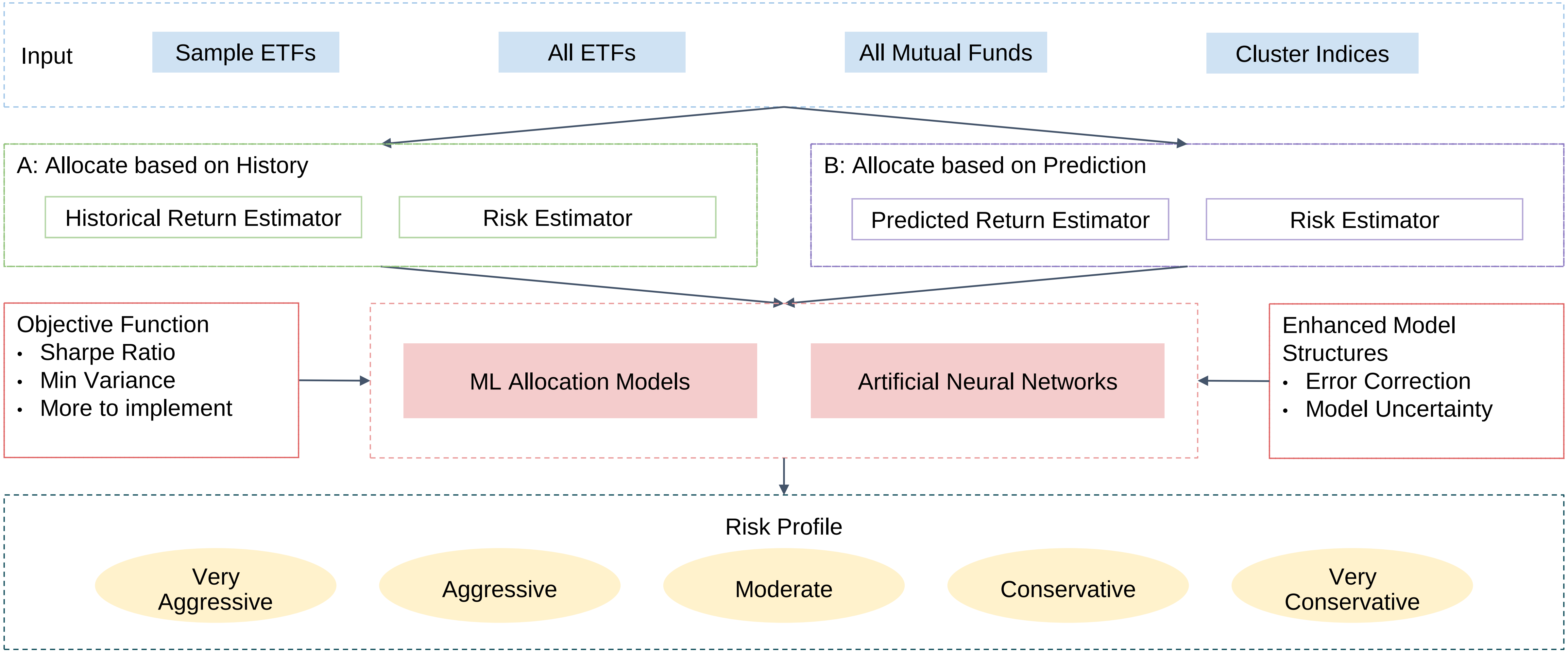


Model for Return Forecast	RMSE
LSTM	0.0914
LSTM w/ Time2Vec	0.0866
Bi-LSTM	0.0697
Transformer	0.1081

Asset Allocation (1 of 2)

Team members: Miao Wang, Sikun Xu

This module gives the ability to **ascertain how much of the total capital to allocate to each asset type** given time horizon and risk tolerance/profile.

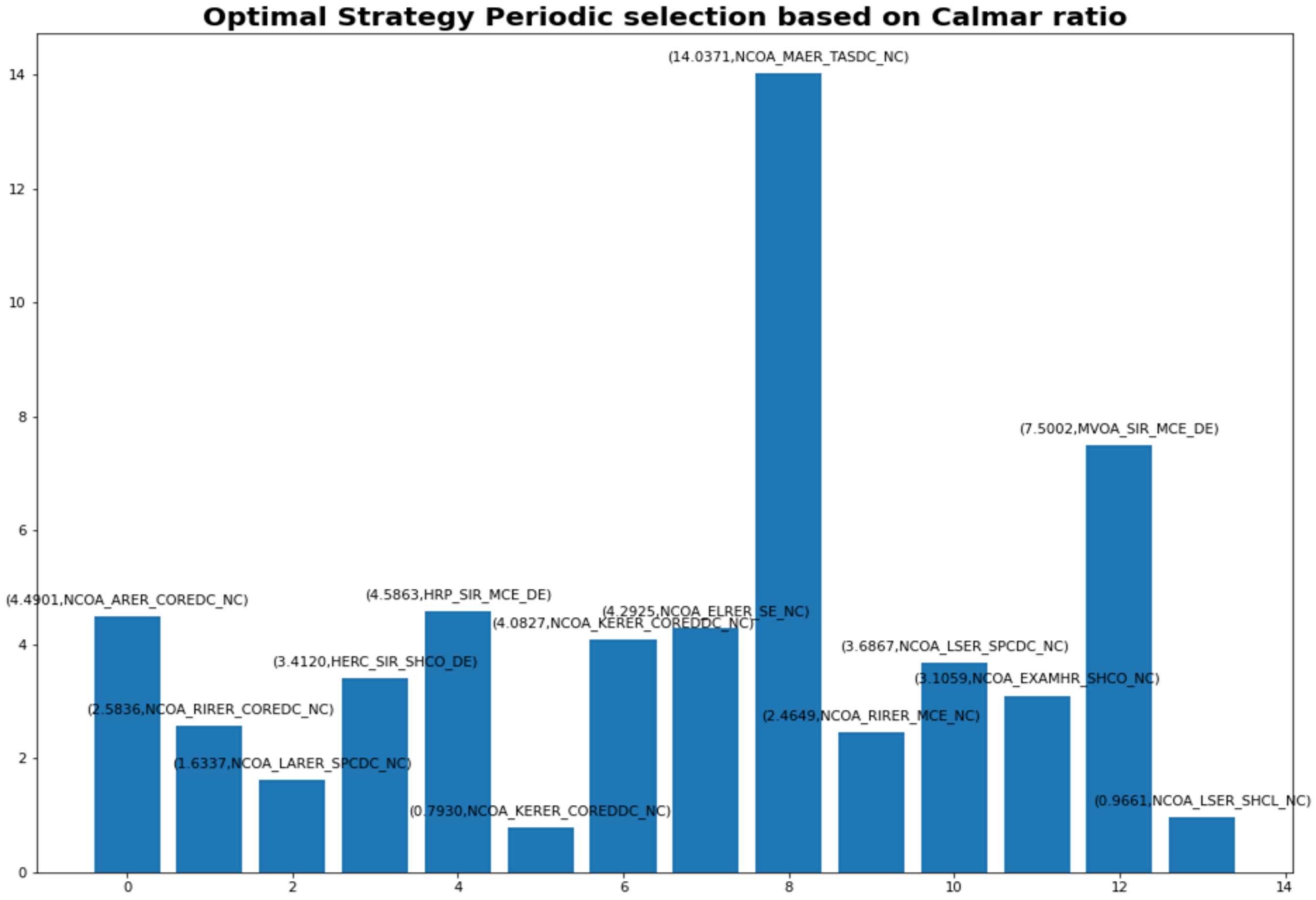


Asset Allocation (2 of 2)

Team members: Zongyuan Chen, Zhiqing Fan, Srihari Dammalapati

Performing different ML allocation models using ML return estimators and Risk estimators for 3 month investment periods across 5 years.

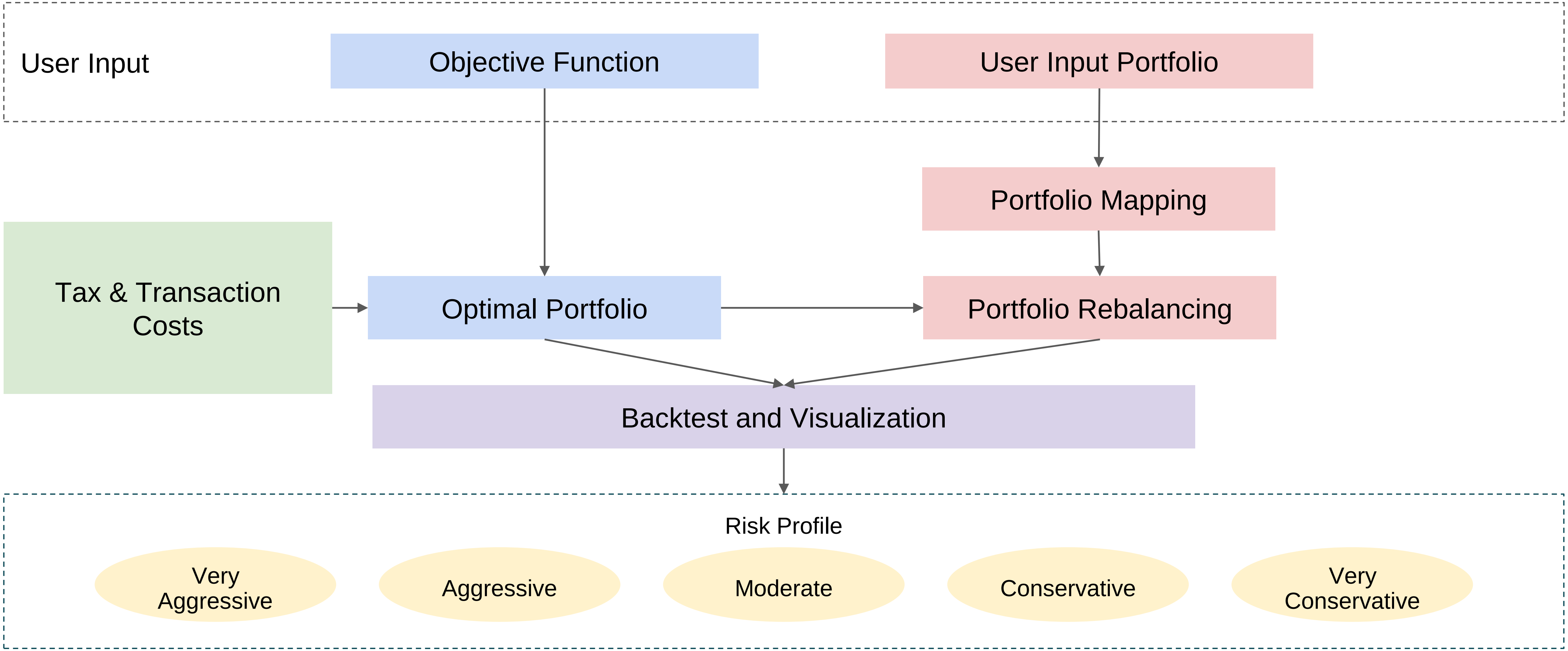
In general ML return estimators showed better performance compared to Simple Return estimator.



Portfolio Construction

Team members: Miao Wang, Sikun Xu

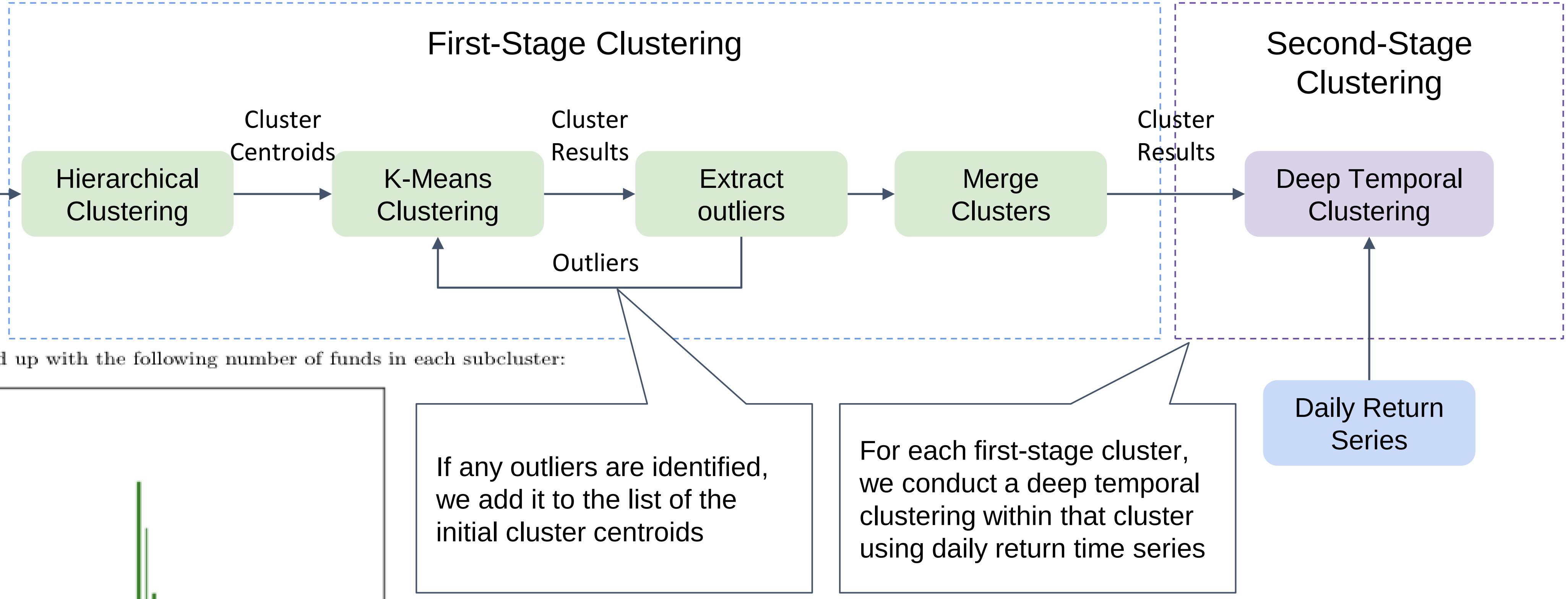
This module gives the optimal mix or composition of securities (e.g. funds, ETFs, stocks) that can populate each of the risk profile based on the percentage allocations driven by the allocation module given time horizon and other objectives



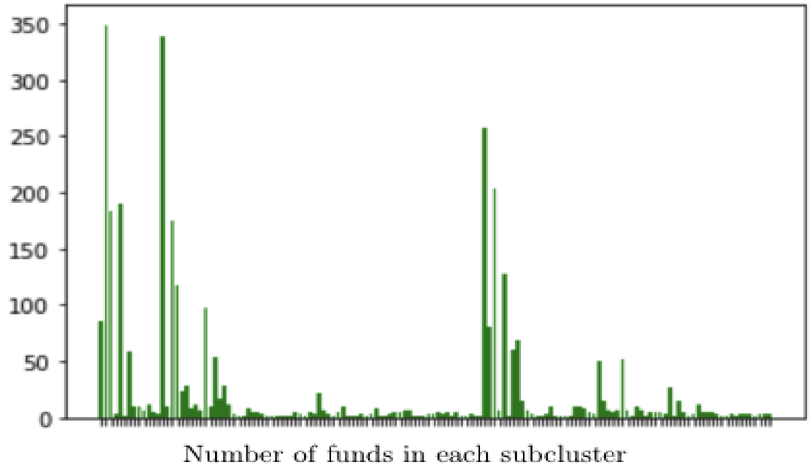
Fund Clustering

Team member: Penelope Lafeuille

The fund clustering project aims at grouping mutual funds that share some commonalities can generate clusters based on their holdings, risk/return profile and other types of features.



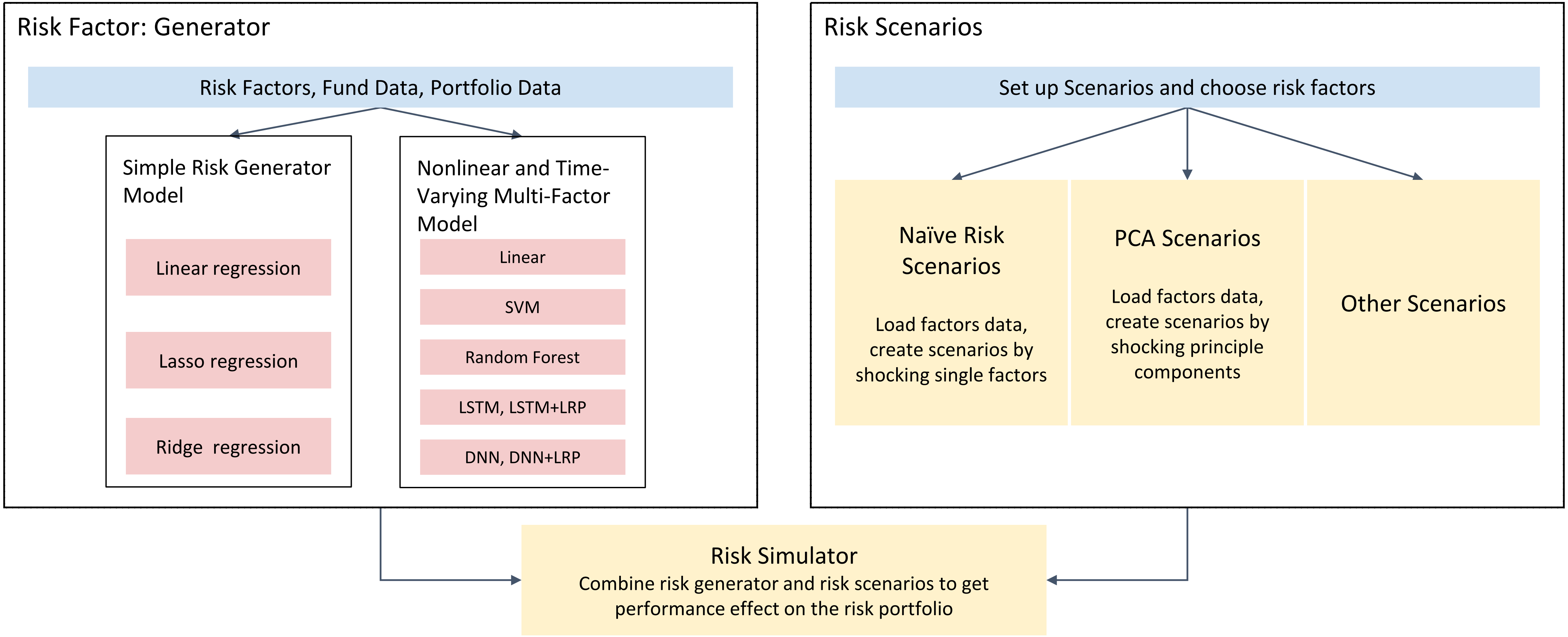
with 5 years of data, we end up with the following number of funds in each subcluster:



Risk Management

Team member: Miao Wang

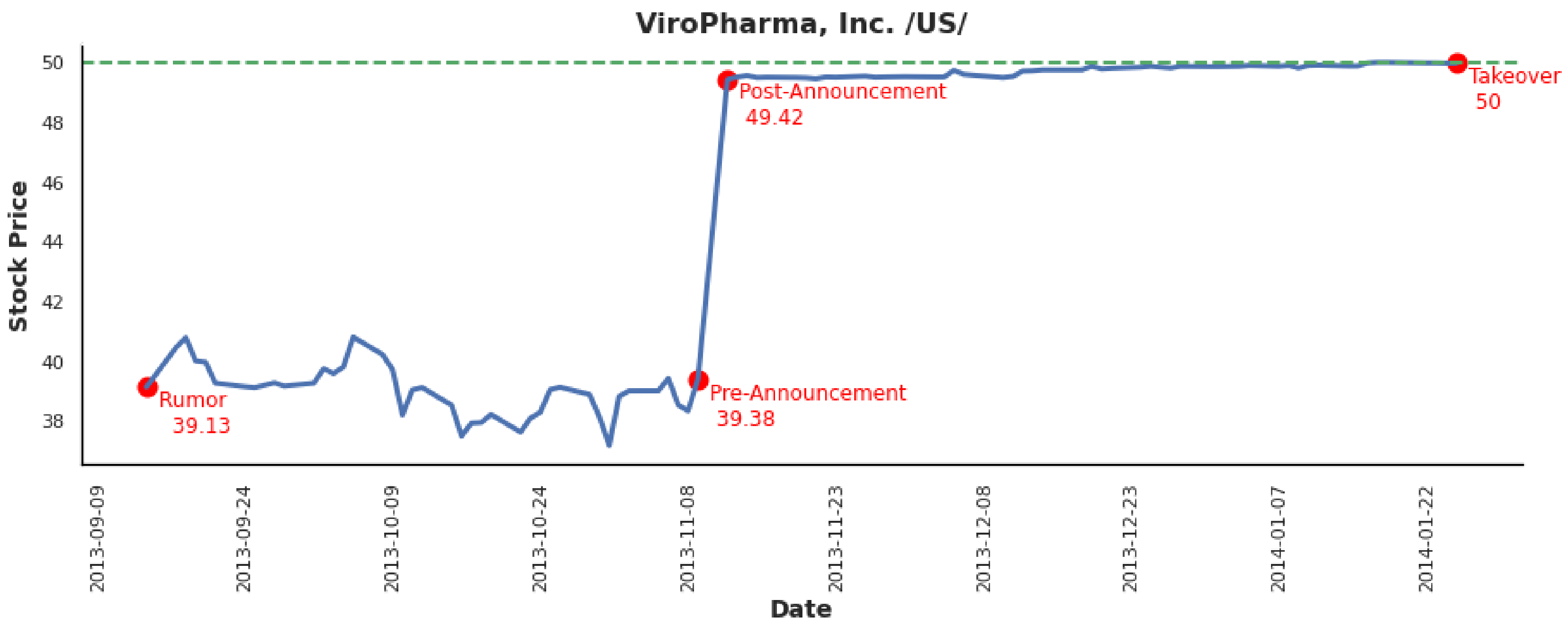
Based on the risk scenarios generated, and risk factor profile of risk portfolio, simulate the price and return movement of the portfolio based on risk shock



Predicting Takeover Success via Machine Learning Techniques

Team member: Tugce Karatas

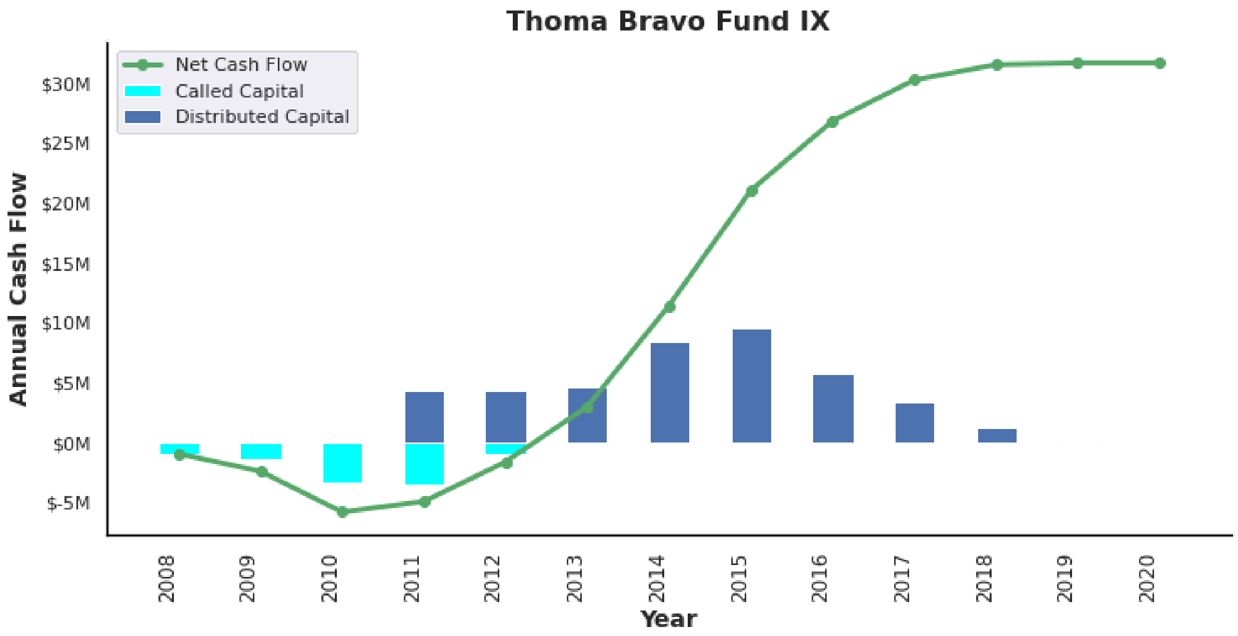
Predicting the status of M&A deals in advance is a vital problem for arbitrageurs.
We aim at building a robust classification methodology for deal success.
We investigate the problem in two stages:
(1) predicting deal announcement after rumor
(2) predicting takeover success after deal announcement.

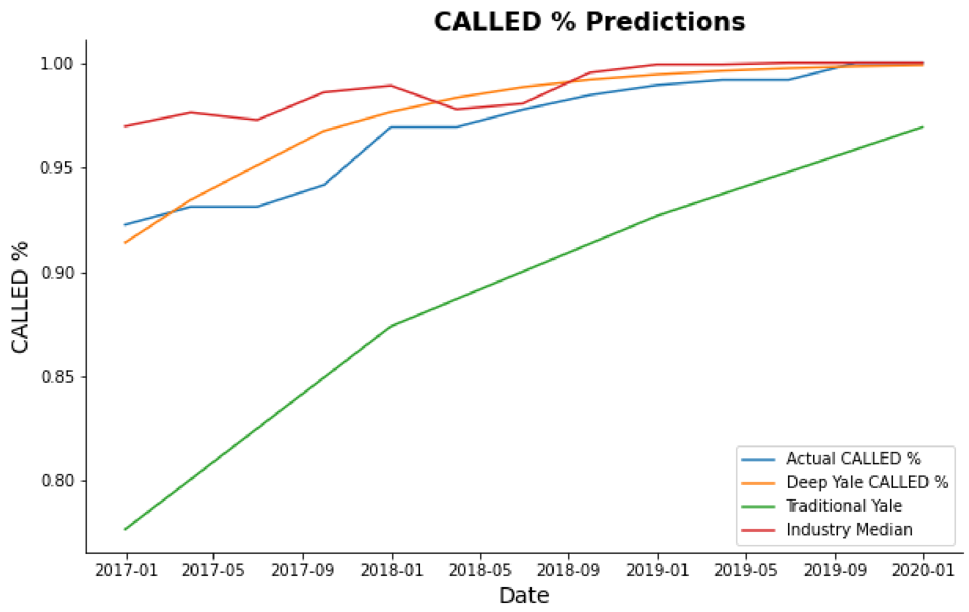


Cash Flow Forecasting of Private Equity Funds via Machine Learning Techniques

Team members: Tugce Karatas, Federico Klinkert, Wen Cheng

Cash flow forecasting of private equity funds is a challenging yet an interesting problem. The time and size of distributions and contributions are unknown. We aim at predicting distributions, contributions, and NAV of private equity funds in advance using machine learning and deep learning techniques.





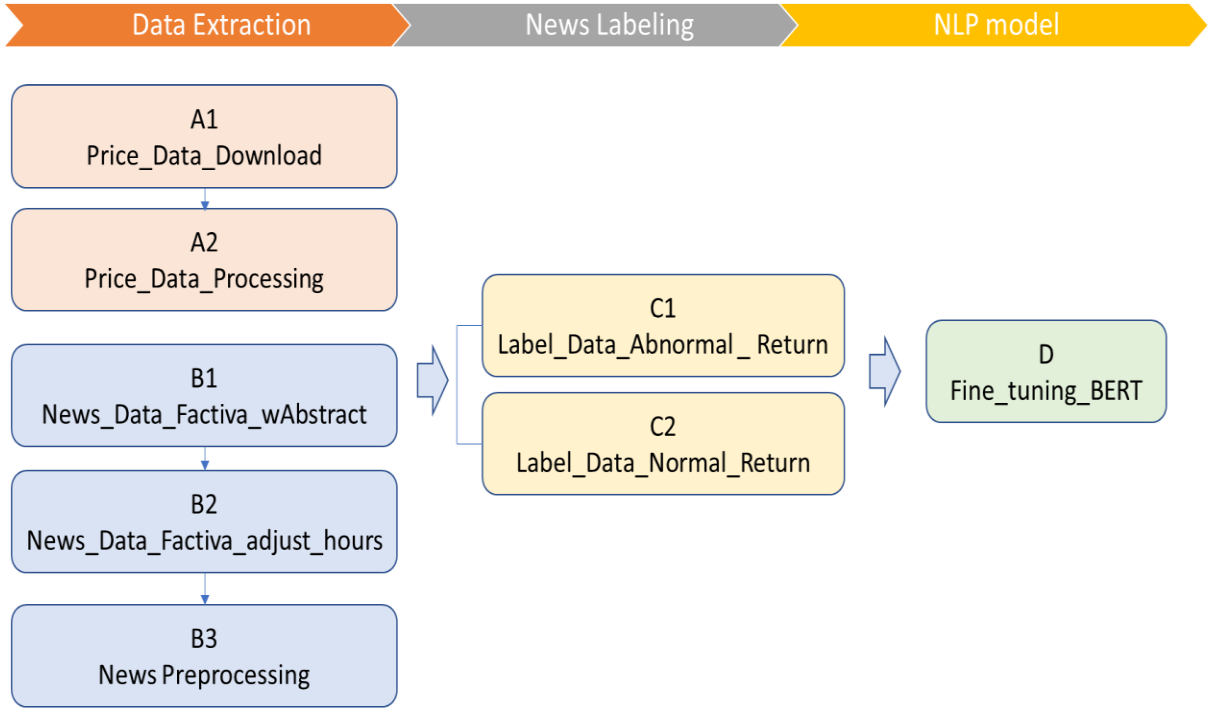
Market Indicator: Sentiment Analysis Using News Data

team members: Arun Varghese, Chia-Yi Wei

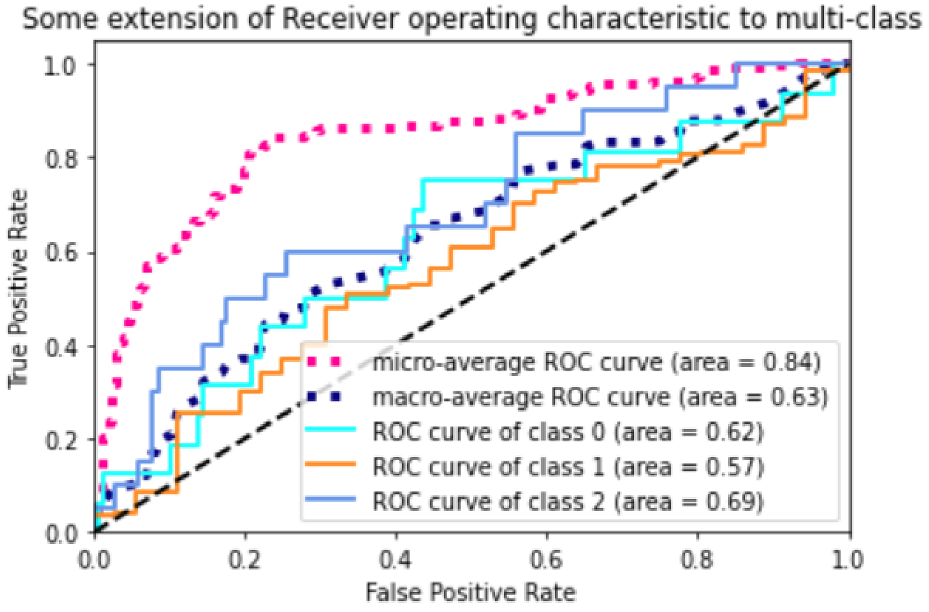
To predict the return of sector ETFs based on the sentiment of unstructured news articles.

We extracted news articles + abstract, cleaned and label them as {-1,0,1}with 15 mins ETF return. We use BERT to obtain word embedding and classify with three hidden layer feedforward neural networks.

Workflow:



Numerical result on XLE (Energy ETF):

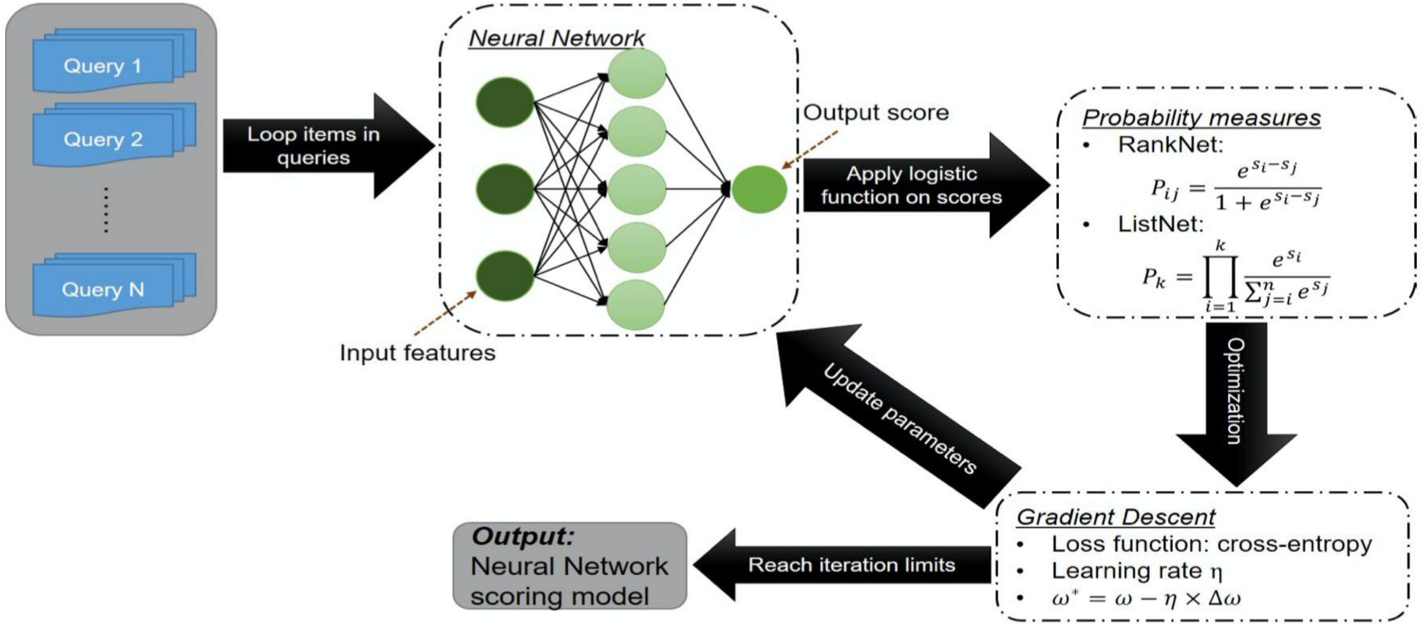



```
Accuracy on the test set: 68.39%
Macro auc: 0.631735
Micro auc: 0.842416
```

Market Indicators: Sector Performance

team members: Irene Qinyi Lin & Sheldon Allen

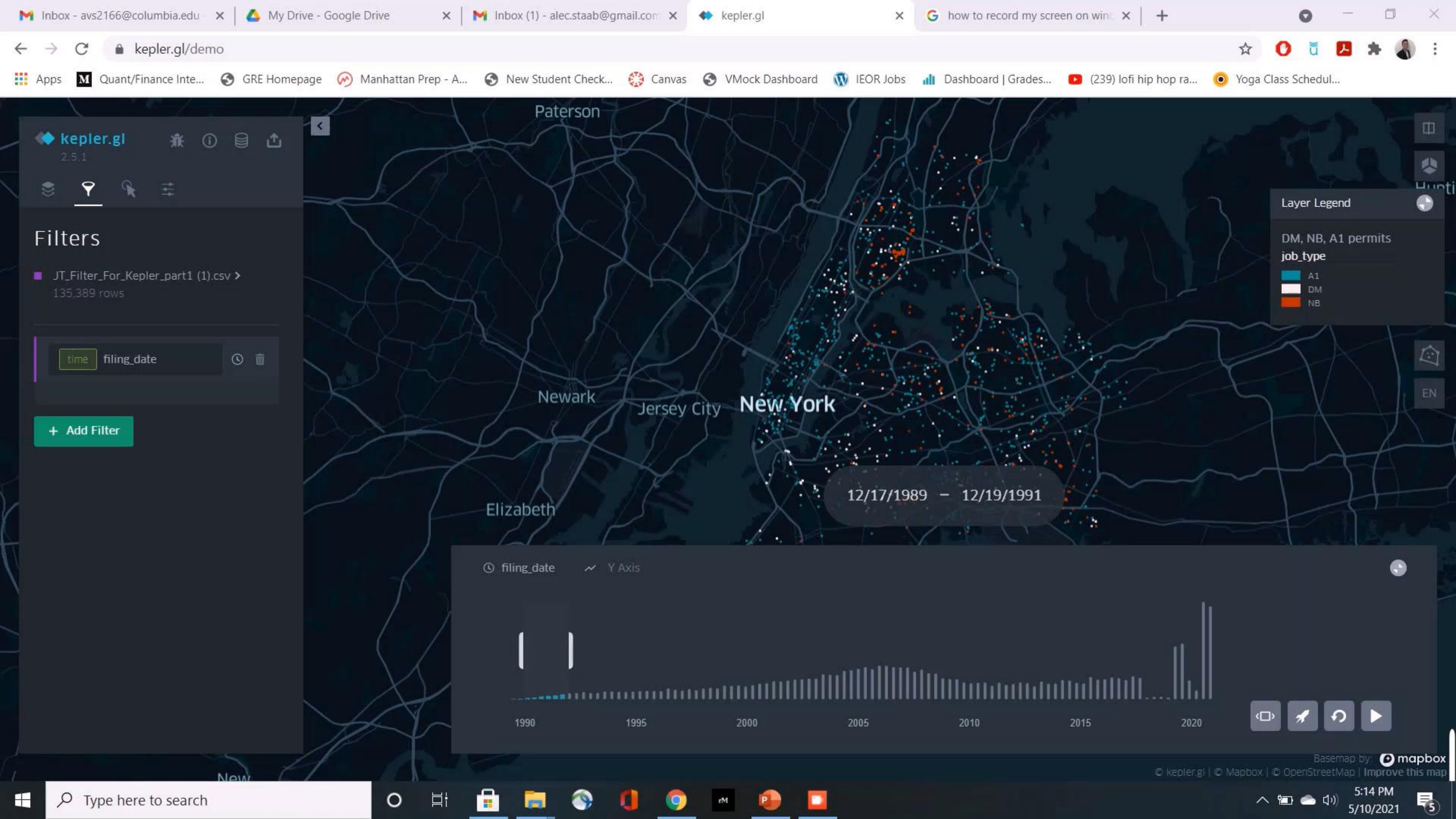
- Leveraging the huge progress in search engine techniques inspired by the Web's explosion, we extended Information Retrieval algorithms and adapted "Learn To Rank" ML algorithms (such as those developed and used by Microsoft, Yahoo, Salesforce, etc.) to "query" macroeconomic and other indicators for clues about future sector performance.



Microeconomic Real Estate Pricing Utilizing Public Data Sources

Team members: Josh Panknin, Alec Staab, Shi Jie Koh, Cyrus Moazami

- The video in the next slide runs a 2-year & 5-year rolling time series about 3 types of building permits.
- Permits are NB for New Building, DM for Demolition, and A1 for major change permit to a building.
- Goal: to find patterns from this data visualization to identify changing areas based on major permit type.
- Through visualizations of all types of permits, the distributions are relatively the same regardless of permit type.
- We are using LODES economic data, Building Permits, Property Valuation data for tax assessments, and hopefully an automated Yelp business data scrapper to find trends in changing economic areas within New York City.
- We hope to later model and perform analysis to provide further evidence that visually changing areas are changing economically, and have various leading indicators based on the public data sources that we have acquired.


Inbox - avs2166@columbia.edu
My Drive - Google Drive
Inbox (1) - alec.staab@gmail.com
kepler.gl
how to record my screen on wind
kepler.gl/demo
Apps
Quant/Finance Inte...
GRE Homepage
Manhattan Prep - A...
New Student Check...
Canvas
VMock Dashboard
IEOR Jobs
Dashboard | Grades...
(239) lofi hip hop ra...
Yoga Class Schedul...
kepler.gl
2.5.1
Filters
JT_Filter_For_Kepler_part1 (1).csv
135,389 rows
time
filing_date
Add Filter
Paterson
Newark
Jersey City
New York
Elizabeth
Layer Legend
DM, NB, A1 permits
job_type
A1
DM
NB
EN
12/17/1989 – 12/19/1991
filing_date
Y Axis
1990
1995
2000
2005
2010
2015
2020
Basemap by: mapbox
© kepler.gl | © Mapbox | © OpenStreetMap | Improve this map
Type here to search
5:14 PM
5/10/2021

Value drivers in real estate with major US bank

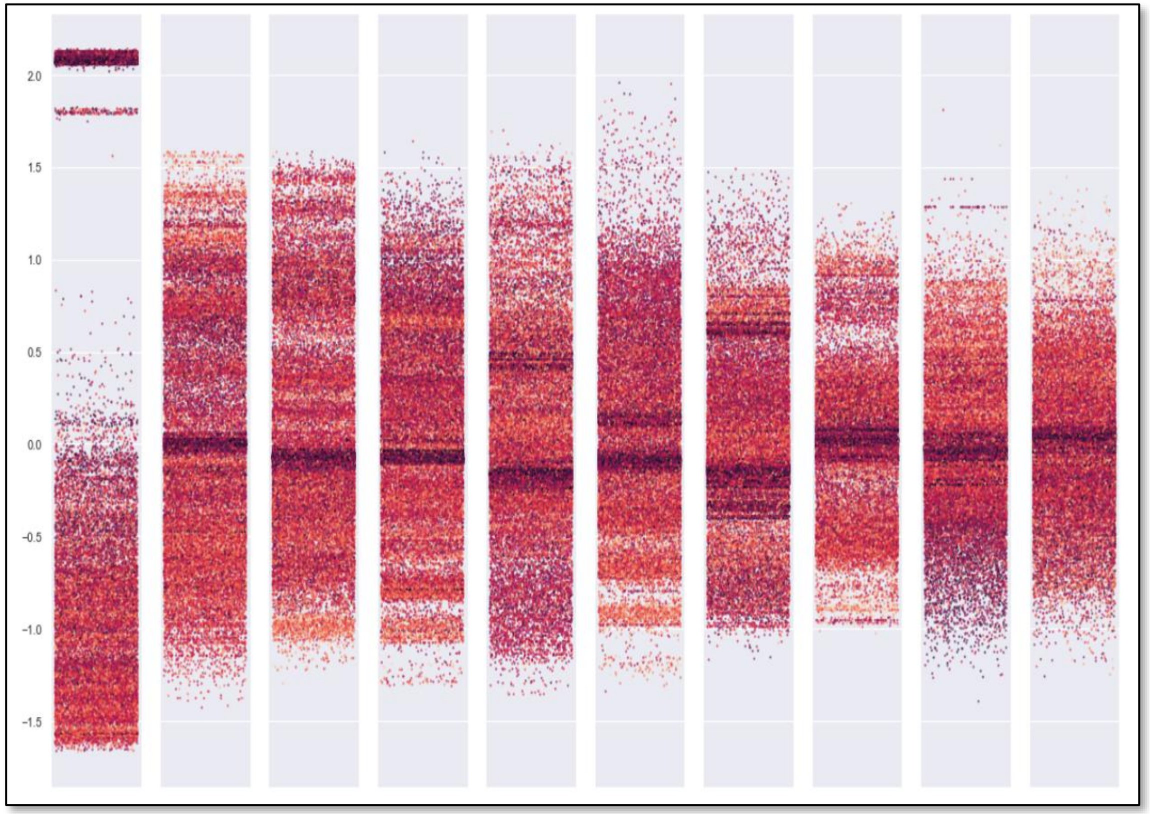
Team members: Josh Panknin, Tang Tianyi, Yang Liu, Lim Guizong Isaac, Cyrus Moazami

The goal of our project is to understand the drivers of real estate value at a local level for a major bank's real estate business.

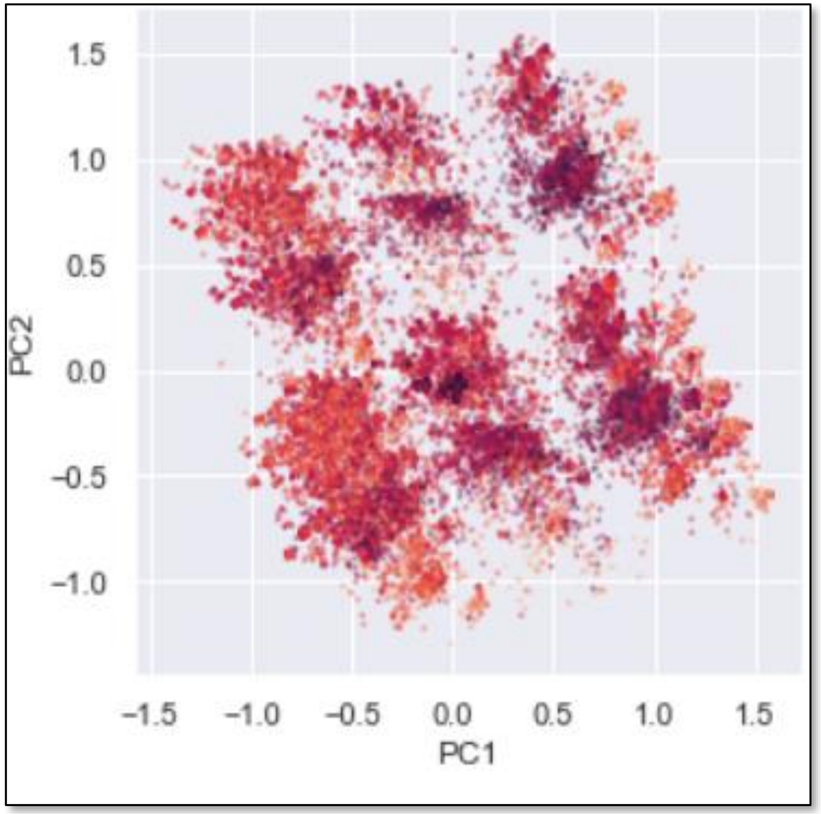
Our work is premised on theories in urban economics, which state that value in urban areas is driven not just by broad macroeconomic factors, but also by localized spatial factors relating to liveability and consumption.

Focusing on Dallas County, the project requires to integrate data from sources like Dallas County Appraisal District and Yelp, and to identify drivers of real estate value from extremely high dimensional data.

To that end, we have sought to use tools like PCA, embeddings and autoencoders to reduce the dimensionality of our data and render it more interpretable.



A plot of individual principal components, colored by log(value). There are clear regions of concentration in low and high valued properties.



Principal components plotted against each other, colored by log(value). Depicts obvious clusters in our data - goal is to understand the factors underlying these clusters.

Climate Change Performance Classification

Sophia Wang, Nathan Therien, Shirley Shen

- Use ML methods to classify countries worldwide in terms of their climate change performance ratings from 1 to 5 (1 is the best)
- Best model: LGBM (validation accuracy score 72.25%)

	country	2022 yr ratings
0	Algeria	5
1	Argentina	5
2	Australia	2
3	Austria	3
4	Belarus	5
5	Belgium	4
6	Brazil	2
7	Bulgaria	2
8	Canada	2
9	Chile	3
10	China	5
11	Croatia	5
12	Cyprus	3
13	Czech Republic	2
14	Denmark	2
15	Estonia	3
16	Finland	2
17	France	2
18	Germany	3
19	Greece	3
20	Hungary	2
21	India	5
22	Indonesia	5
23	Ireland	4
24	Italy	2
25	Japan	3
26	Kazakhstan	3
27	Latvia	5
28	Lithuania	2
29	Luxembourg	3
30	Malaysia	3
31	Malta	3
32	Mexico	5
33	Morocco	5
34	Netherlands	4
35	New Zealand	2
36	Norway	2
37	Poland	2
38	Portugal	4
39	Romania	5
40	Saudi Arabia	2
41	Slovenia	3
42	South Africa	2
43	Spain	4
44	Sweden	2
45	Switzerland	2
46	Thailand	5
47	Turkey	2
48	Ukraine	5
49	United Kingdom	4
50	United States	2

Fraud Detection based on Large Scale Graph Analysis

Team members: Wenqi Wang, Shuai Zhang

The aim of this project is to detect default loan applications based on large scale graph analysis.

We built a Semi-GNN model and achieved marginal improvements compared to the GCN model.

We improved it further by experimenting with different sampling algorithms such as random walk with restart, random jump, random degree node selection, random PageRank node selection.

Our evaluation metrics showed that our algorithms perform better in terms of capturing the original graphs' structures.

Our algorithms achieved marginal improvements in capturing the degree distributions of the original graphs

Sampling methods	calls	contacts	device	idfa	idfv	imsi	phone	user
Random Walk (previous group)	0.21	0.57	0.40	0.11	0.01	0.05	0.09	0.03
Random Walk with Restart (1 sample)	0.19	0.52	0.40	0.11	0.001	0.04	0.08	0.002
Random Walk with Restart (4 samples)	0.12	0.30	0.40	0.10	0.00	0.02	0.06	0.00
Random Jump (1 sample)	0.19	0.51	0.40	0.11	0.00	0.04	0.08	0.003
Random Jump (4 samples)	0.12	0.26	0.40	0.10	0.00	0.02	0.06	0.00
Random PageRank Node	0.20	0.54	0.40	0.11	0.004	0.04	0.08	0.004
Random Degree Node	0.18	0.47	0.40	0.11	0.002	0.04	0.08	0.002

Table 9: *The table displays the D-statistics of the degree distributions with different sampling algorithms.*

Fraud Detection with User Sequential Behaviors

Team members: Zongqian Wu , Jiahao Yan

To utilize customer's behavior sequences before submission of the loan application to make fraud detection in the online lending business. The group tries to develop some deep learning methods to extract features from those sequential data thus being effective for fraud detection.

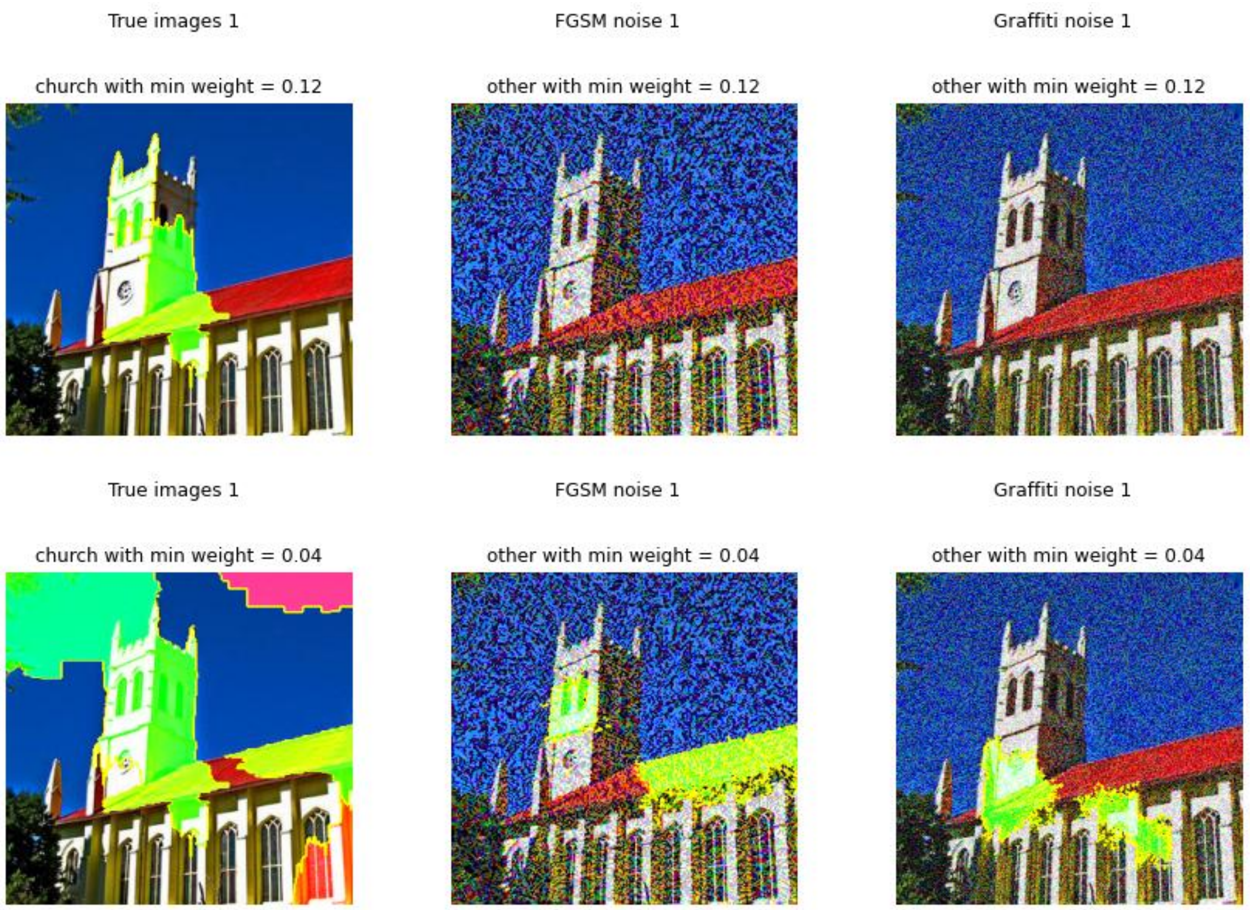
Model	AUC	KS
BERT， MLM 15% with no NAs data	0.6202	0.1758
CatBoost, SGT features by single & combined sequences	0.6493	0.2200

Interpretability against adversarial attacks

Team members: Augustin Laruelle & Hanze Sun

In order to help users decide when to trust or not to trust a black-box model's predictions, we use the LIME interpretation model on original and adversarial images to understand the rationale behind the predictions.

With the hypothesis that the output distribution of the LIME model will be understandable and be focused for the original images whereas it would look random for adversarial images.



Adversarial Attack on WordLSTM and Soft Pattern (Sopa) Models

Team members: Ruilin Xiao & Wenxin Mu

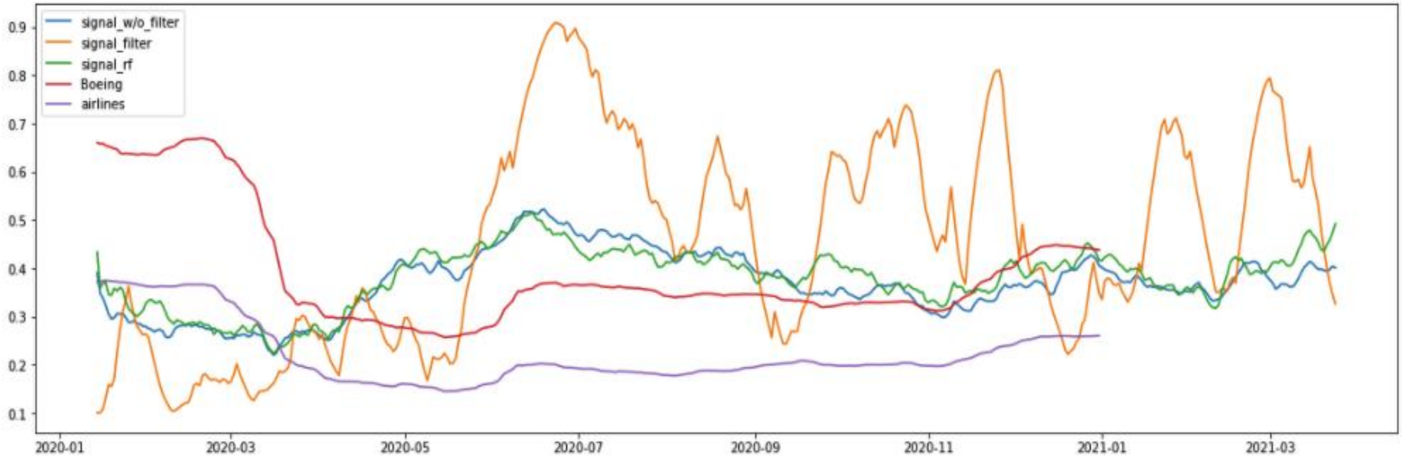
Conduct experiments on using an attacking algorithm to generate adversarial texts and fool the WordLSTM and Sopa models

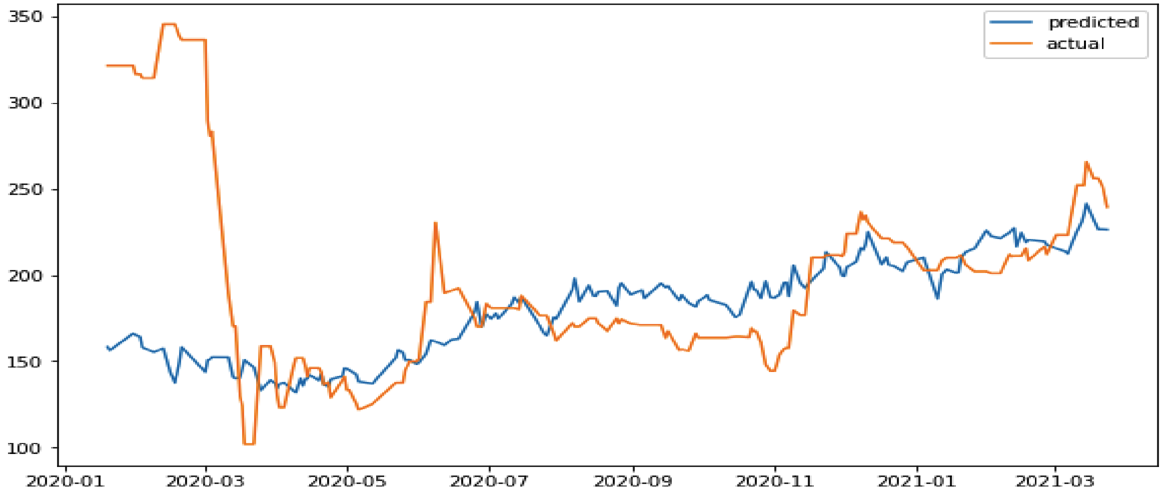
	Original Accuracy	Adversarial Accuracy	perturbed word percentage
WordLSTM	84.5%	0.26%	3.465%
Sopa	85.881%	11.936%	13.874%

Predict market moves using news traffic from wikipedia pages

Team members: Shiqi Chen & Ranchao Yang

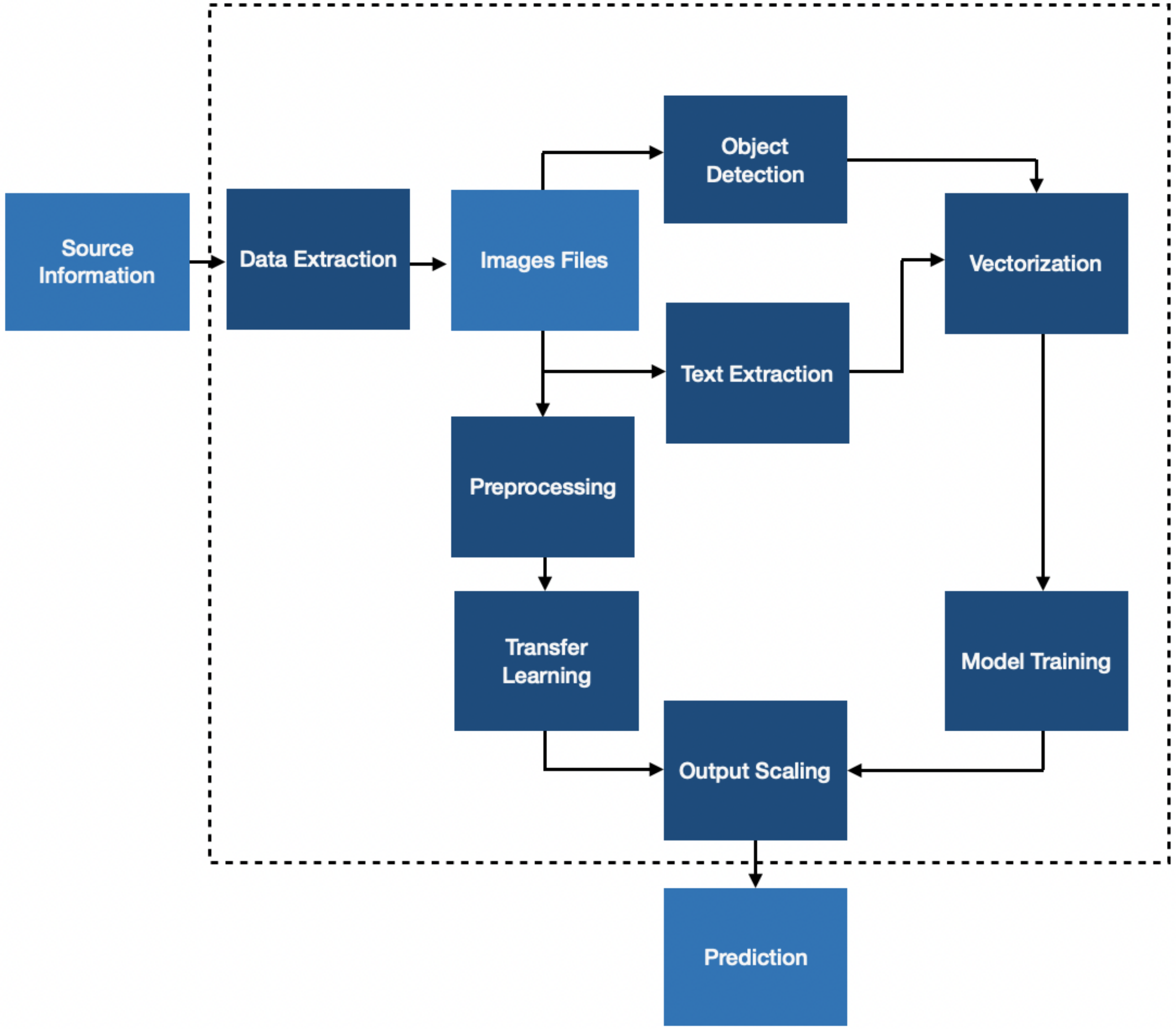
Investigate the relationship between Covid-19 signal and stock prices during pandemic period





Building Instance classification using Google images

- Occupancy and construction attributes are very important for insurance pricing and risk models
- Although this information is provided by the customer, the data is inconsistent appearing in free-form text, categorical or numeric data
- A model that augments image data with text, categorical, etc. to predict the occupancy with the highest accuracy is the goal of this project
- Used the publicly available images of buildings like Hospitals, Hotels, Schools, etc. and built a Deep Learning model to predict the building class
- Explored and implemented different ways to extract information (object detection and text recognition) from building images
- Scraped google images from various US cities (reduce geographical bias) to train a model that predicts building class



Team members: Manohar Anantha, Gino Castellucci

Simulating financial time series using RegGAN & QuantGAN

Team members Weilong Fu, Michael Xiong, Ruilong Zhuang

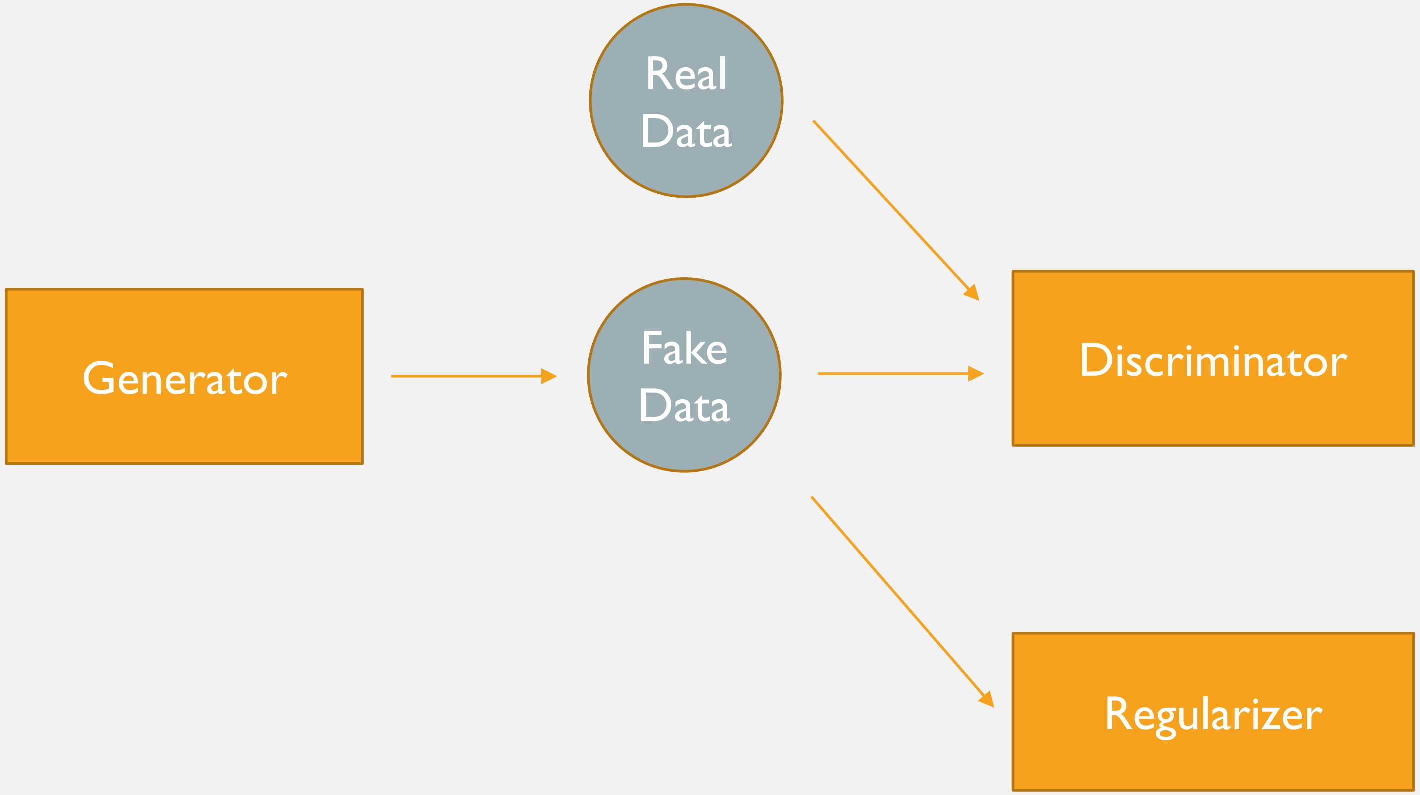
focuses on implementing the RegGAN framework into the QuantGAN architecture to simulate financial time series.

The QuantGAN architecture allows the model to capture some long-range non-linear dependencies such as the presence of volatility clustering.

The RegGAN framework allows us to influence specific characteristics of the samples the generator is creating.

In our case, we focus on matching the skew of the log returns to that of the long-term log returns.

High level overview



1. Pretrain the Regularizer to represent the skew() function.

2. Generator creates fake financial time series data.

3. Discriminator tries to predict real data as real, and fake data as fake. Then Discriminator gets an update step.

4. Generator tries to trick Discriminator into classifying fake data as real. Then Generator gets an update step.

5. Fake data is passed into the Regularizer which represents skew(). We set the labels to be the skew we would like the generator samples to be. Update the Generator again towards the skew we desire.