Predictive Modeling of Opioid Prescription Fraud by CMS Providers

DATA606 Capstone Project - Part 4

William Rubin – Spring 2020

PROJECT TIMELINE

Part 1 — **Problem Definition**

Part 2 — **Existing Research / Initial EDA**

Part 3 — **Final EDA / Predictive Model Preparation**

Part 4 — **Execution, Interpretation, and Results**

Brief Project Overview

- Problem Statement
  - Opioid Abuse Declared Public Health Emergency by HHS in 2017
  - Estimated 11.4 million people misuse prescriptions
  - Fostered environment for unscrupulous providers to profit from opioid over-prescription or encourage over-use

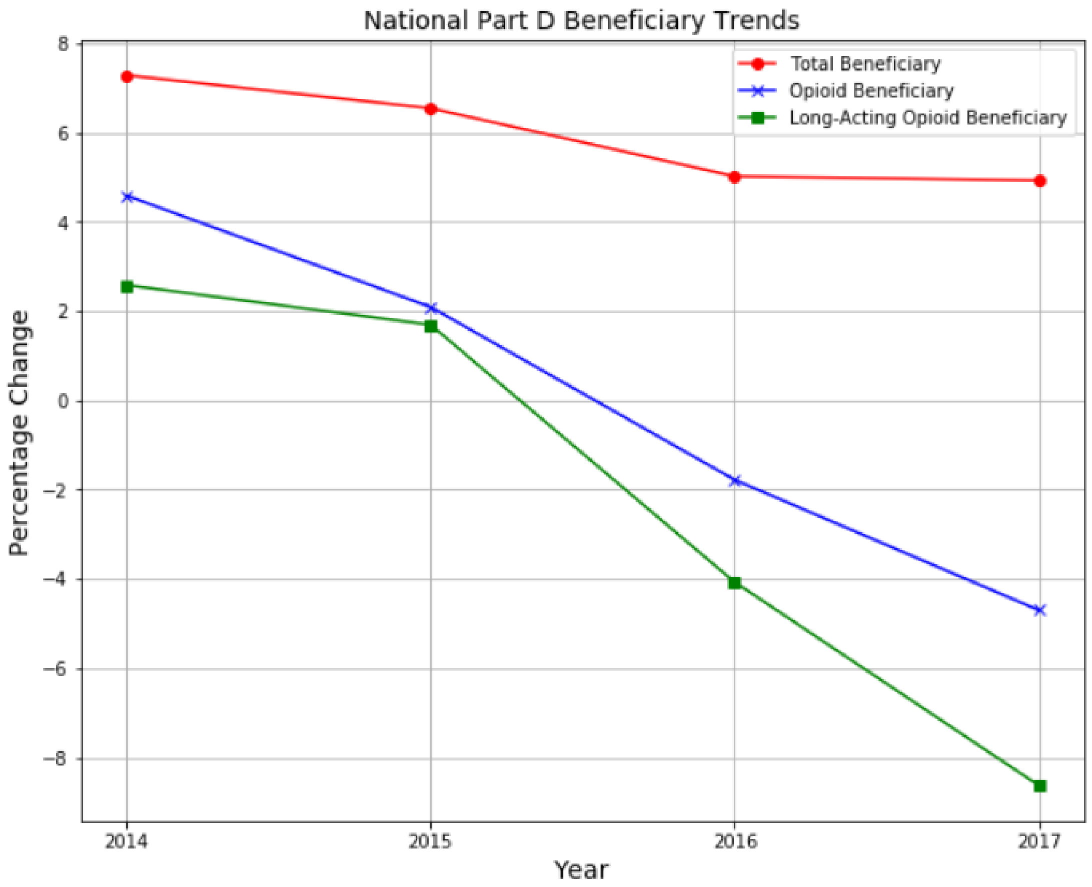
- Project Goal
  - Evaluate opioid prescription patterns for CMS Medicare Part D Providers
  - Create labeled dataset of fraudulent and legitimate opioid providers
  - Develop predictive modeling capabilities to assist in detection of potential prescription fraud

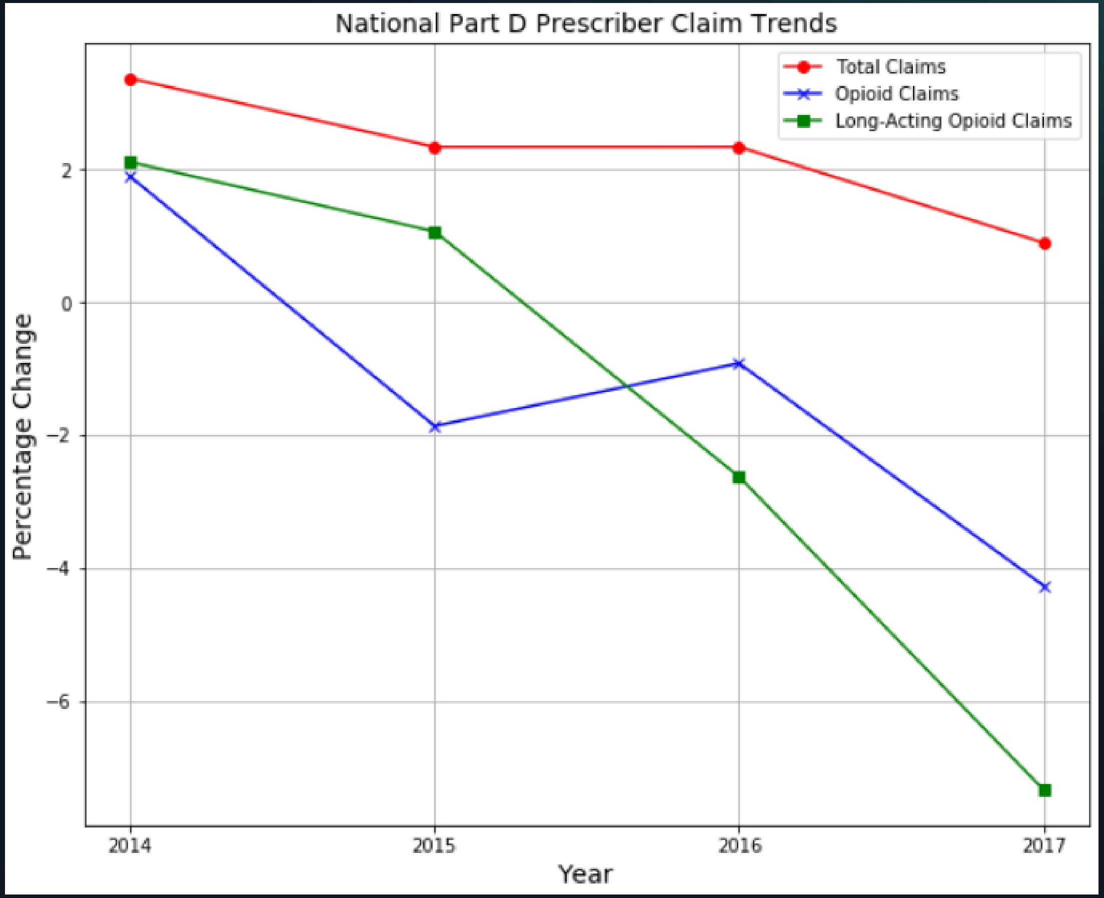
Data Sources Overview

- Department of Justice (DOJ) Press Releases
  - Opioid-Related Official Press Releases (Full Text)
- CMS National Provider Identifier (NPI) Registry
  - Repository of unique provider identifiers (NPI) for all registered Medicare / Medicaid eligible providers
- CMS Part D Prescriber Summary Tables
  - ~5.5M yearly, NPI-based provider records
- Physician Compare 2017 Individual Eligible Clinician Public Reporting - Overall MIPS Performance
  - ~376k NPI-based provider performance metric records

Exploratory Data Analysis

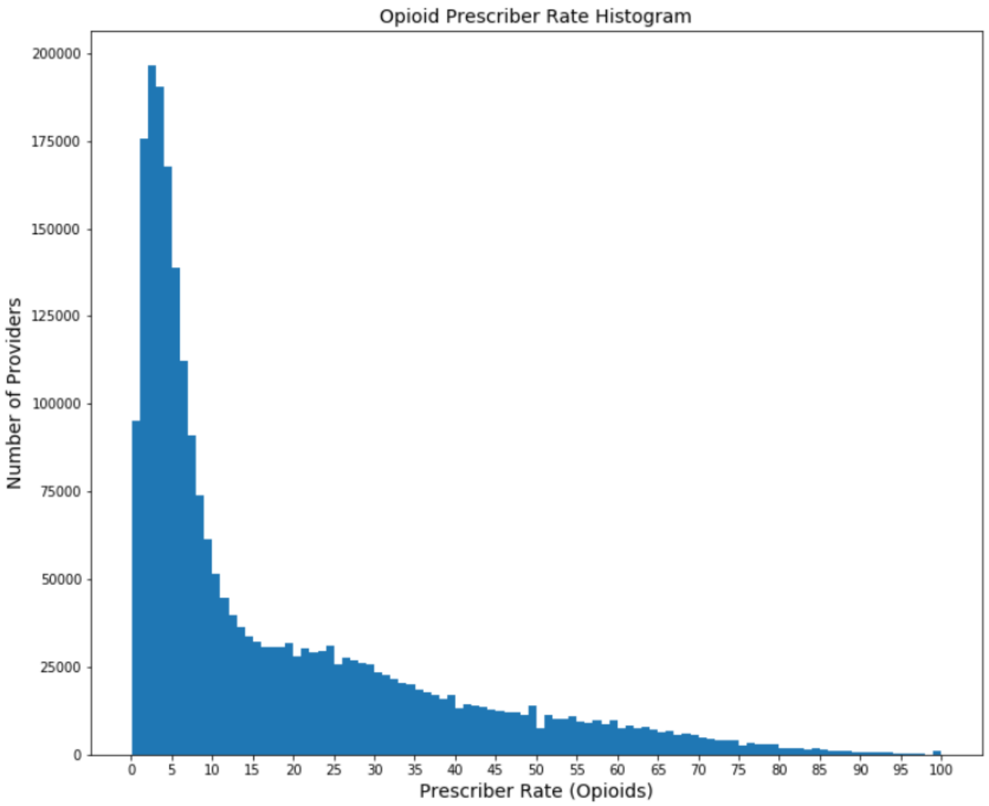
High-level Opioid-Specific Beneficiary / Claims Trend Analysis

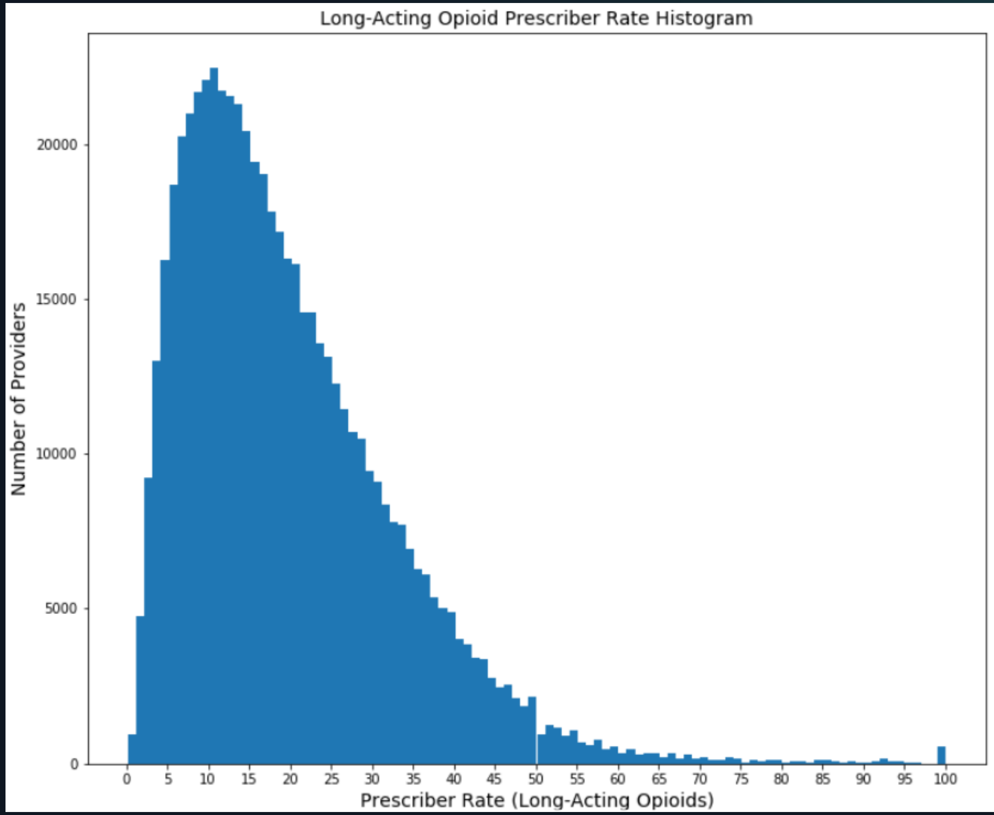




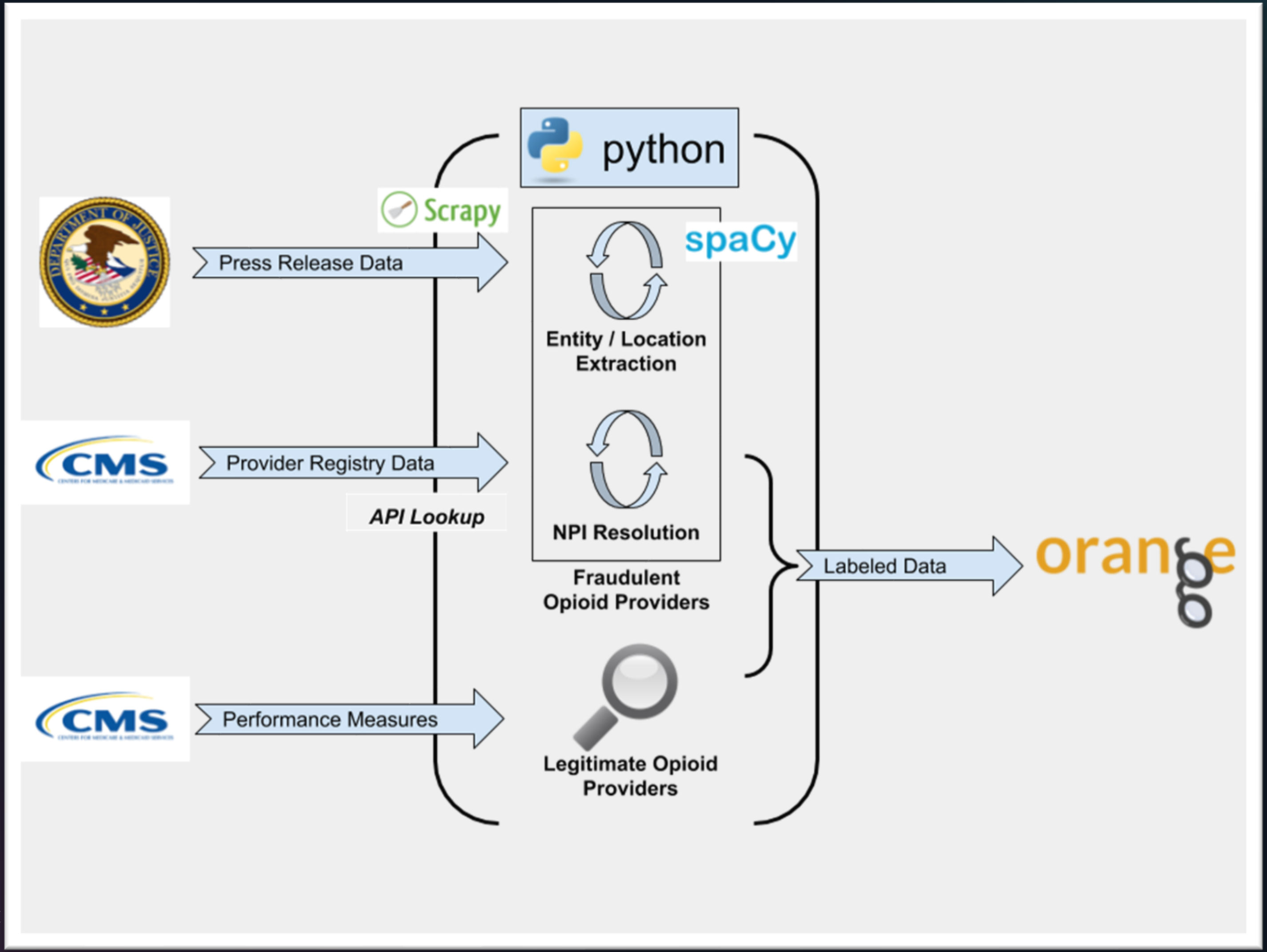
Exploratory Data Analysis

Opioid and Long-acting Opioid Prescriber Rate Histogram

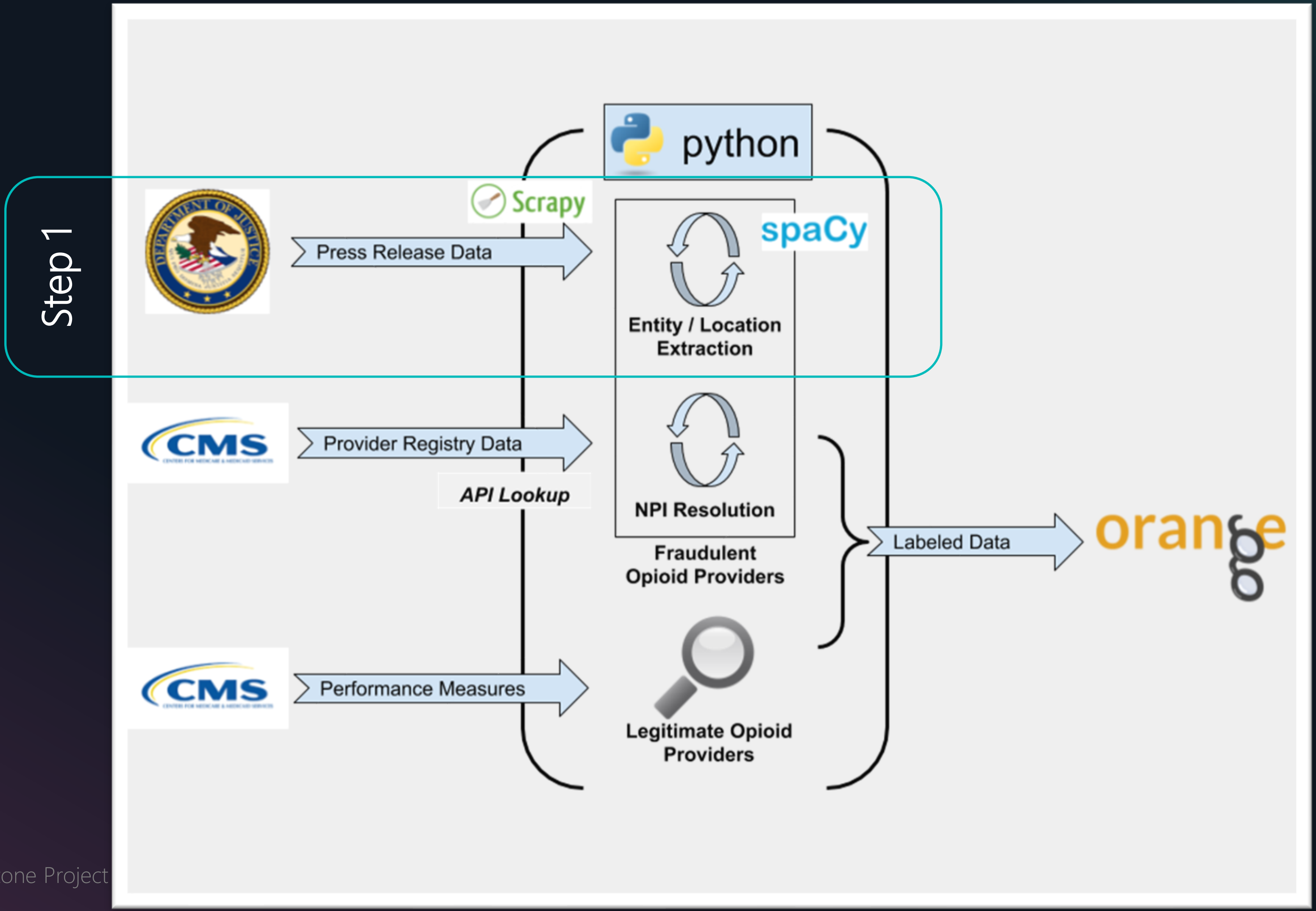




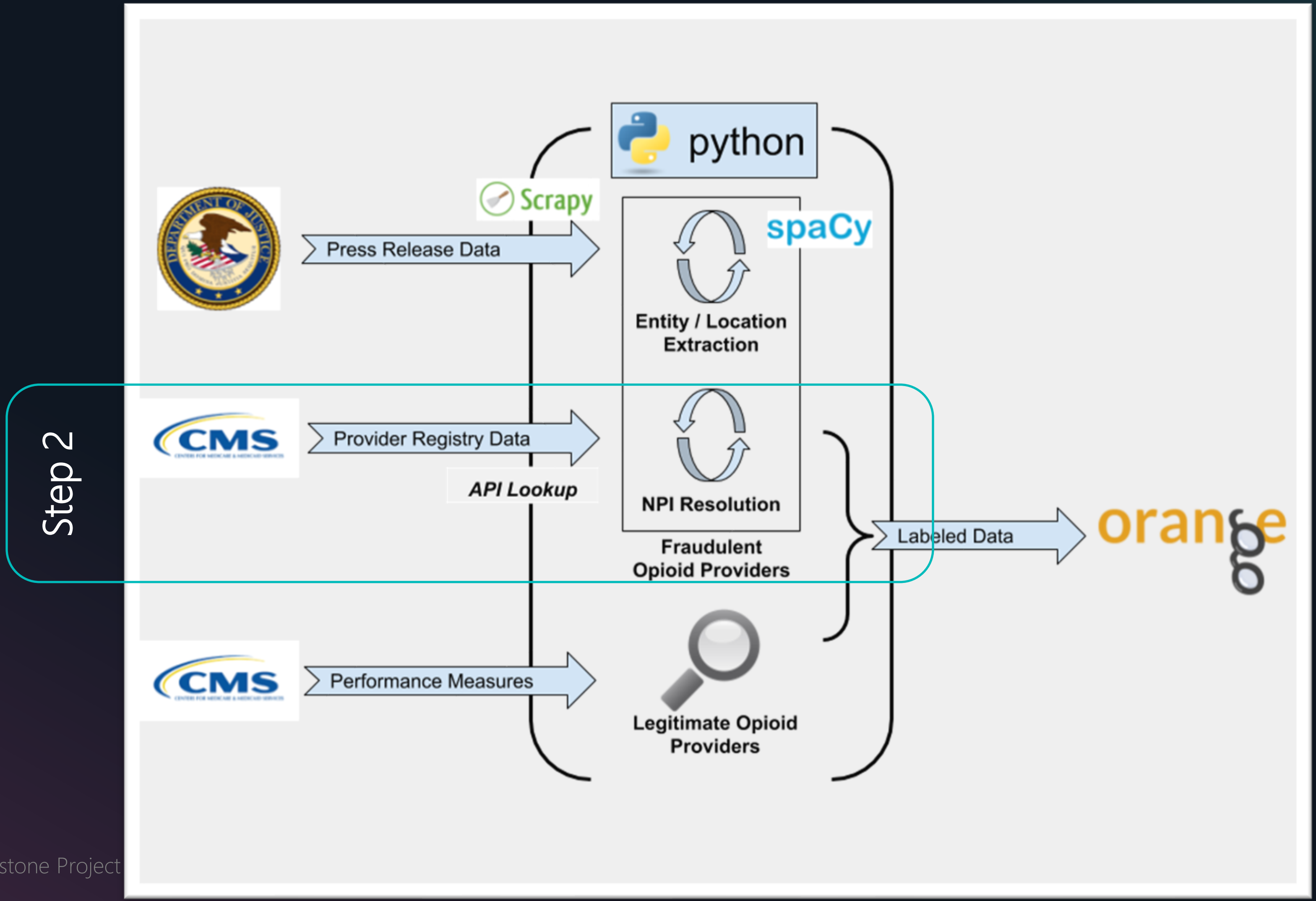
Opioid Provider Labeled Dataset Process



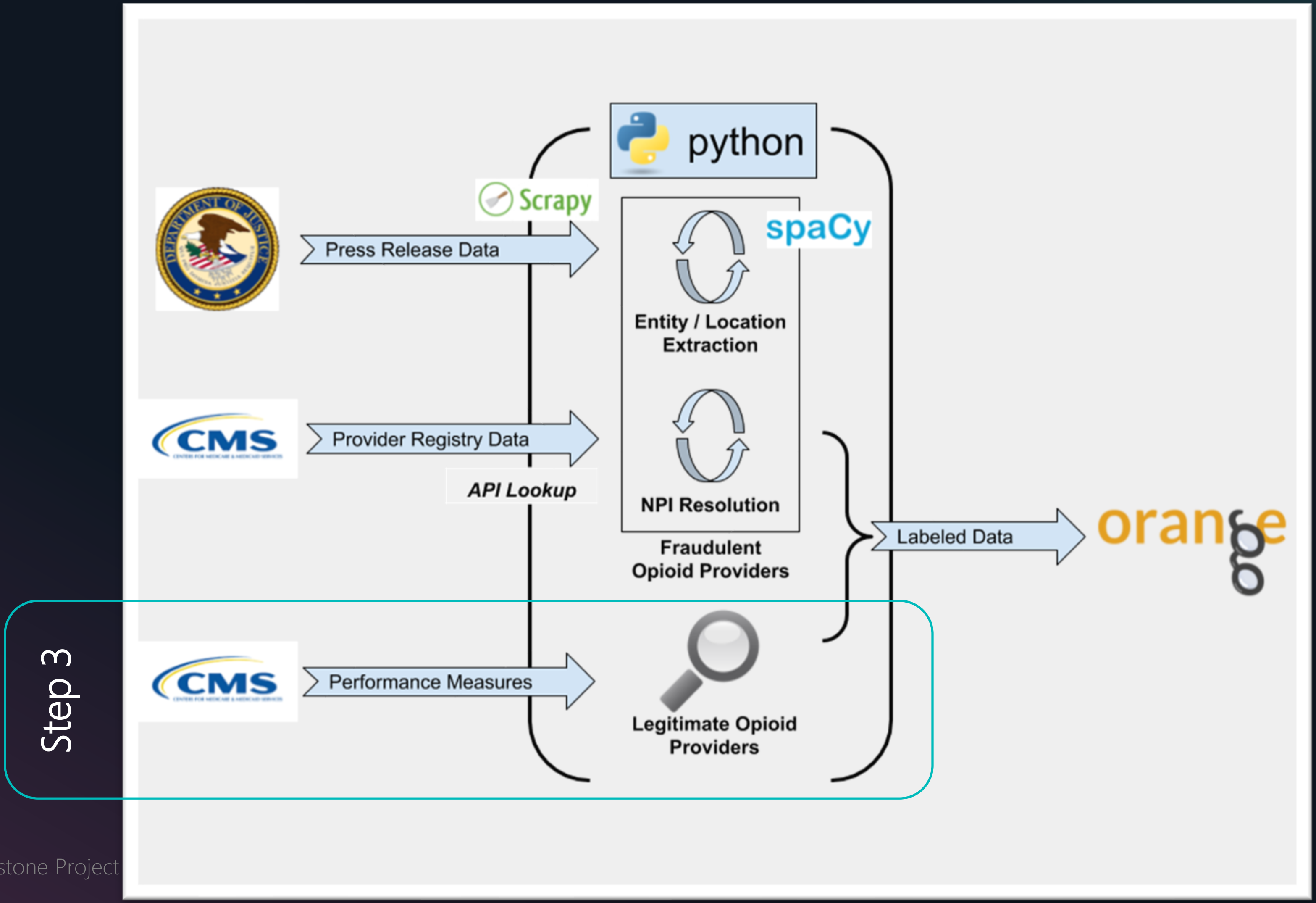
Opioid Provider Labeled Dataset Process



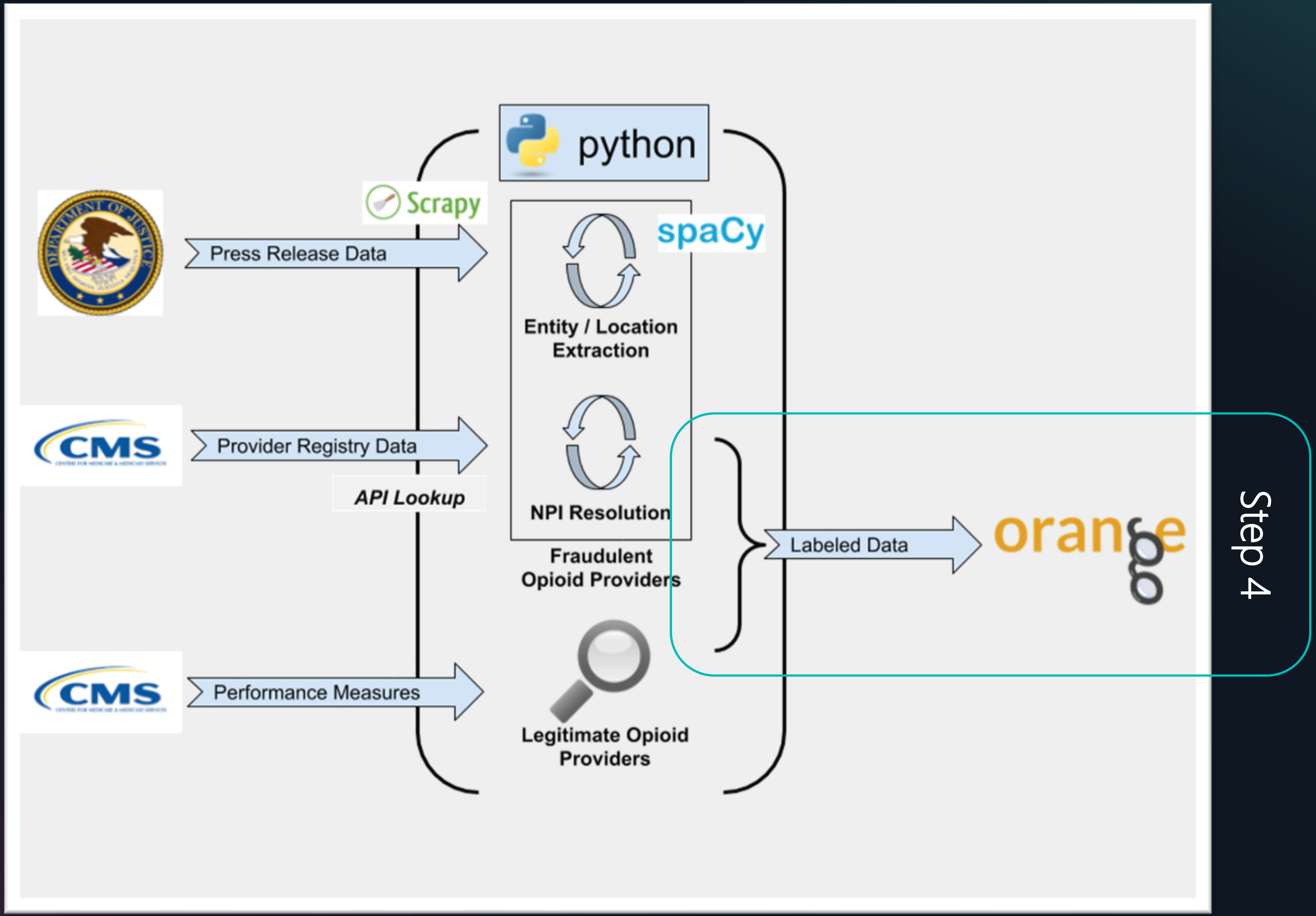
Opioid Provider Labeled Dataset Process



Opioid Provider Labeled Dataset Process



Opioid Provider Labeled Dataset Process



Data Pipeline Modifications (PART 4)

- Addressed Mis-identified Fraudulent Providers
  - Extracted Verbs from Press Releases Text
  - Filtered Press Release to Include Only Legal Action
  - 9 Press Releases Removed
- Resolved Instances of Multiple NPI Results
  - Extracted Likely City Related GPE Entities to Identify Solitary NPI Match
- Included URL / Press Release Date to Output for Traceability
- Finalized Fraudulent Provider Labeled Dataset

Data Pipeline Modified Results

DOJ Press Releases	111
Extracted Entities / Terms	21,113
Resolved NPIs	221

Likely Legitimate Providers

- Merit-based Incentive Payment System (MIPS) Data Source Identified
- Component of CMS Quality Payment Program (QPP)
- Rewards participant providers with adjusted payments based on cost efficiency, quality of care, and health outcomes
- Highly unlikely fraudulent provider would participate and draw additional attention/scrutiny
- No overlap with identified fraudulent providers
- Criteria Utilized: Final MIPS Score = 100

Opioid Provider Labeled Dataset Composition

All Opioid Providers	704,463
Identified Fraudulent Opioid Providers	85*
Likely Legitimate Opioid Providers	38,229

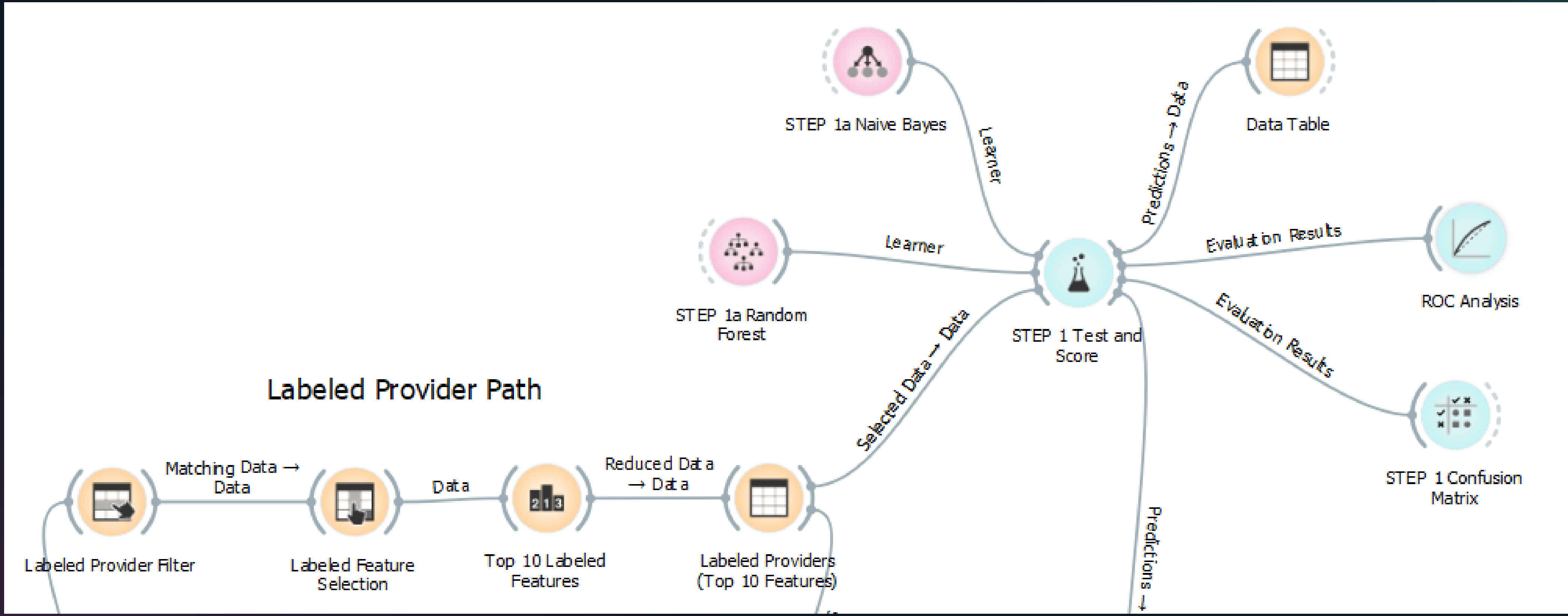
* 221 identified overall but only 85 remained after cross-reference to opioid provider dataset, likely due to inactive status

Prediction Modeling

- Semi-Supervised Learning Approach
  - Practical Approach Between Supervised and Unsupervised Learning
  - Utilizes Small Labeled Data Subset to Predict Unlabeled Records (Pseudo-Labeling)

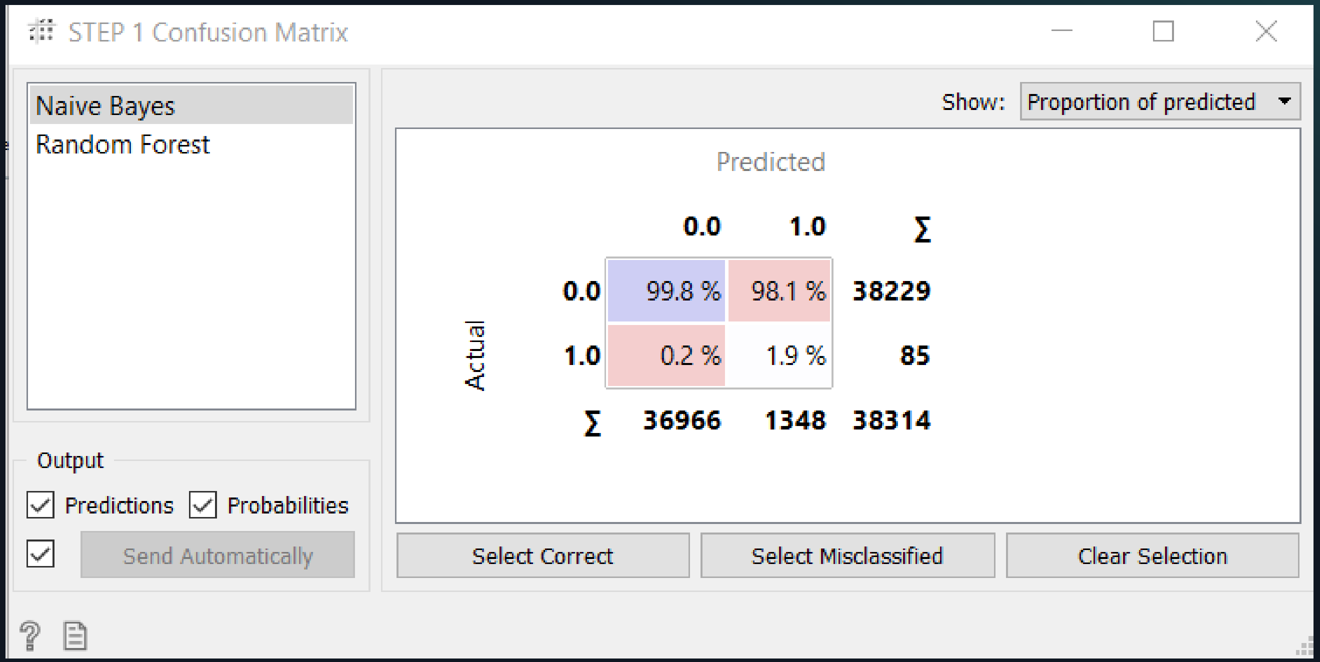
- Implementation
  - Labeled Provider Modeling
  - Pseudo-Labeling of Unlabeled Providers
  - Combined Dataset
  - Fraudulent Provider Prediction Modeling

Labeled Provider Modeling

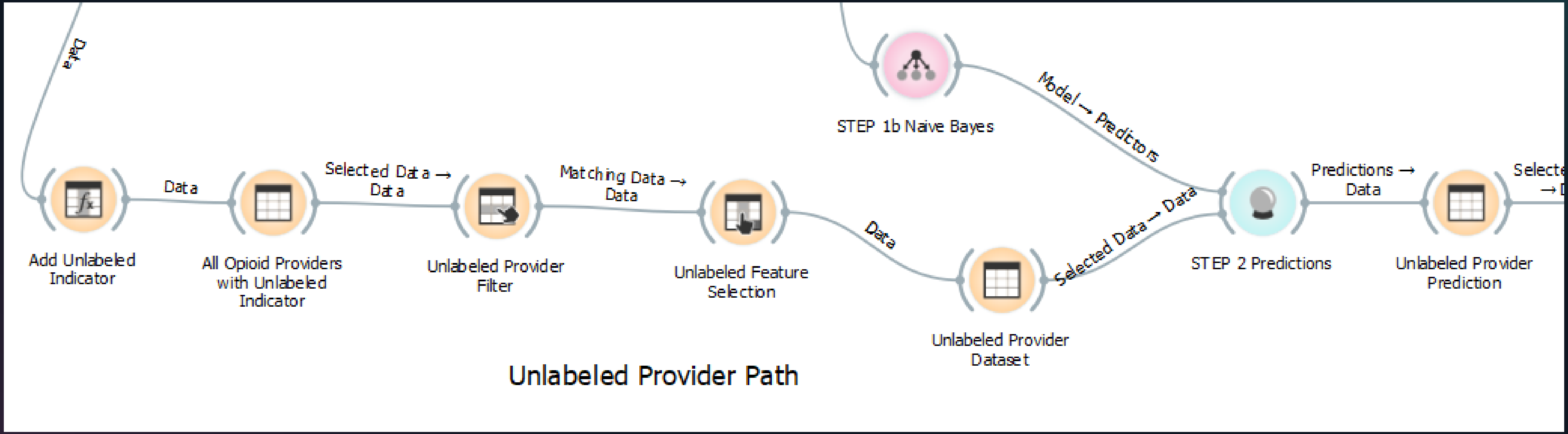


Labeled Provider Modeling Results

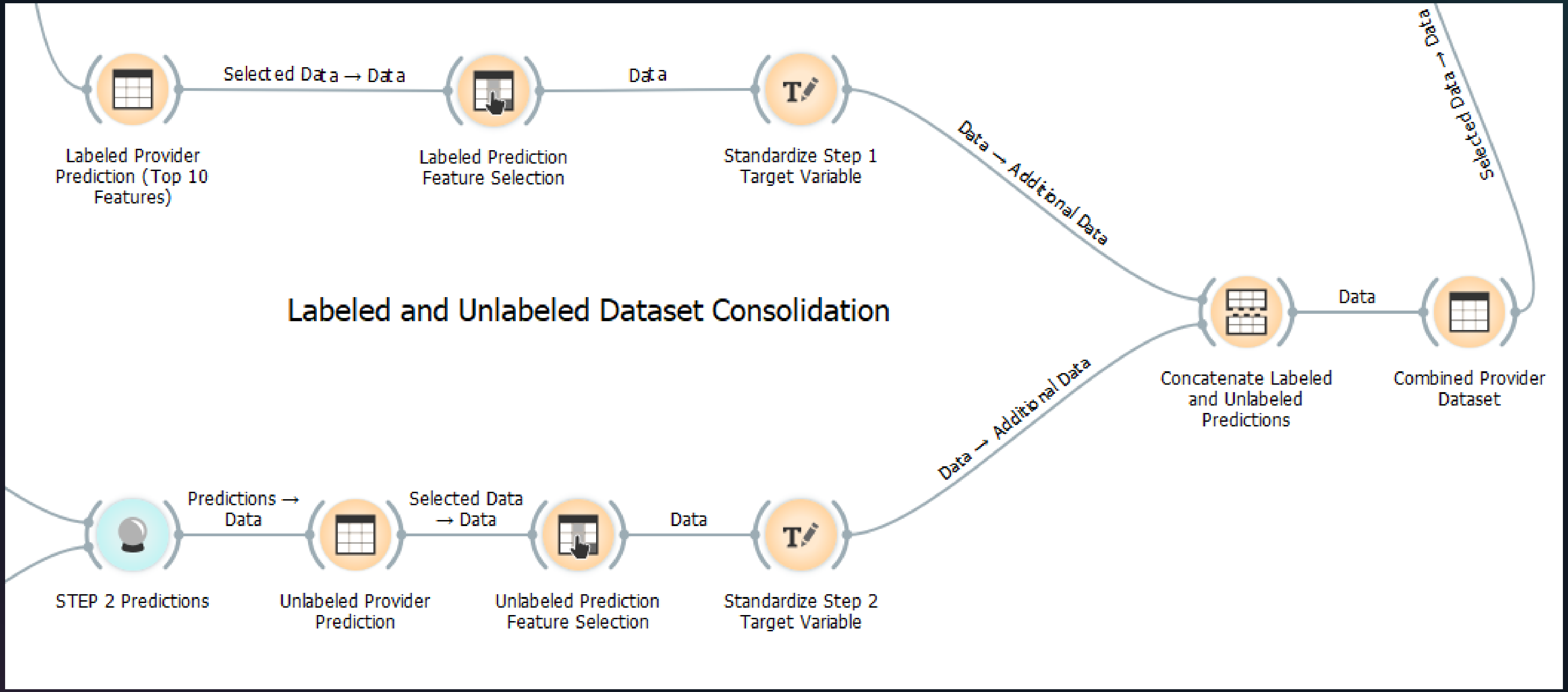
Model Outcomes			
Model	AUC	F1	Precision
Naïve Bayes	0.733	0.980	0.996
Random Forest	0.612	0.997	0.996



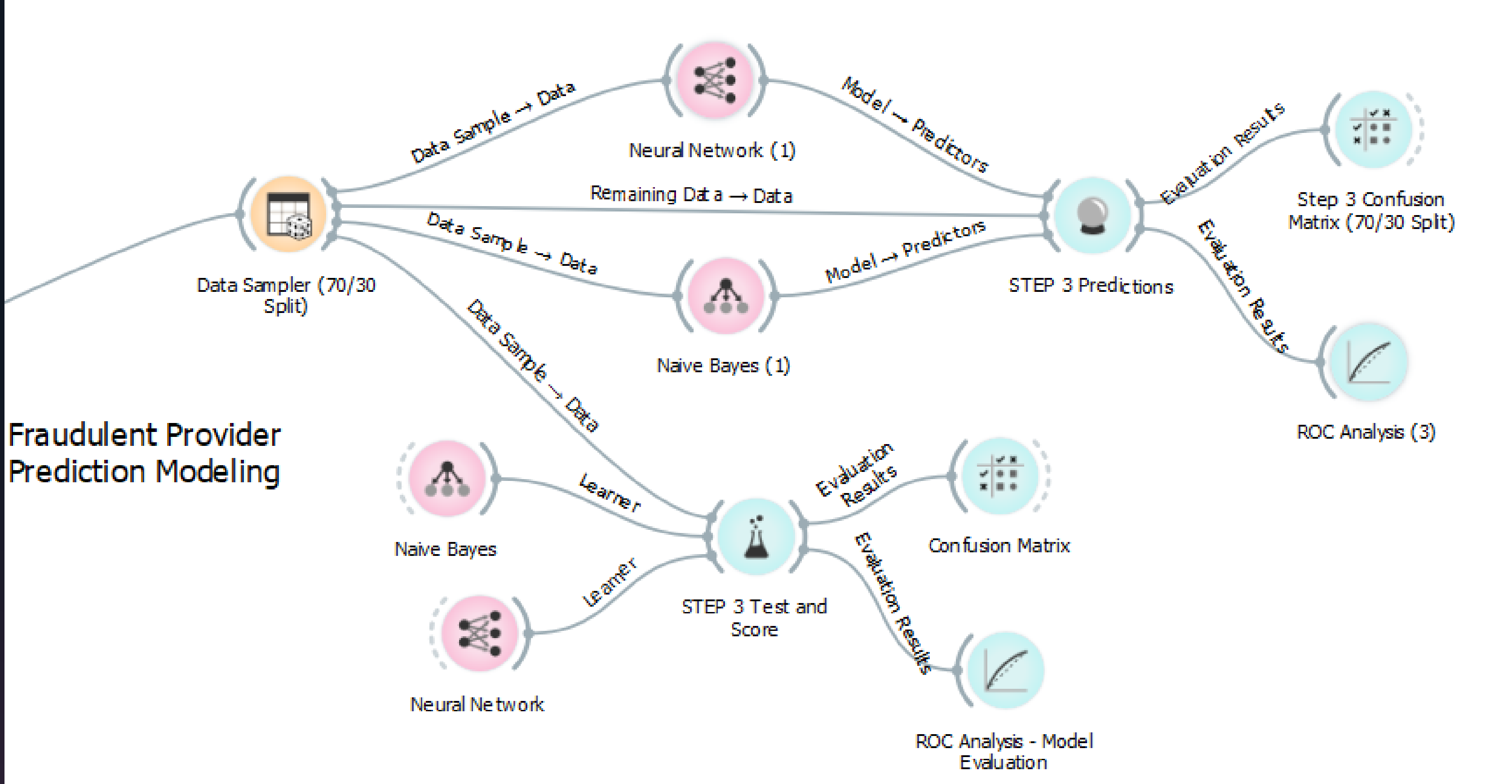
Pseudo-Labeling of Unlabeled Providers



Combined Dataset

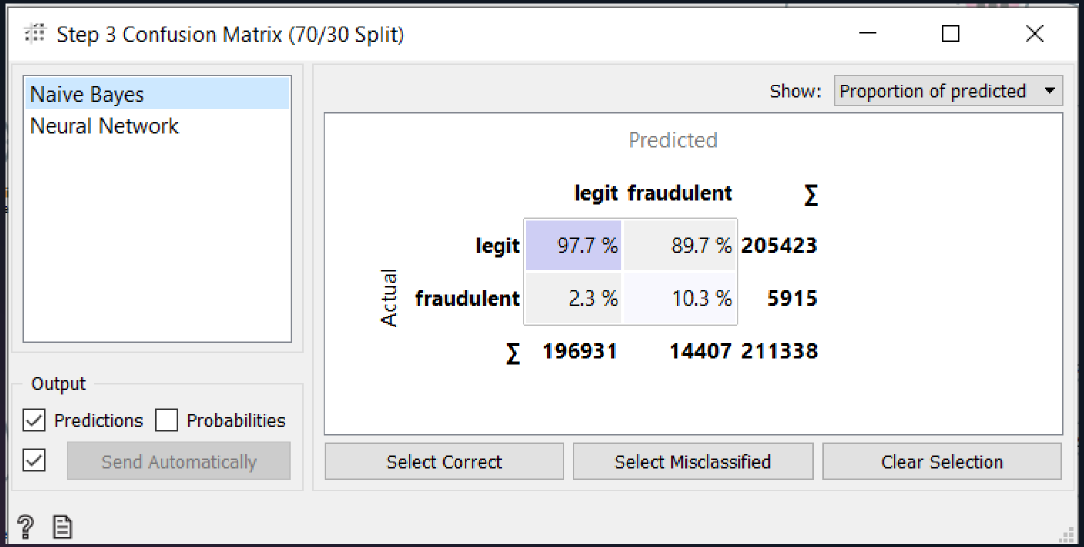


Fraudulent Provider Prediction Modeling



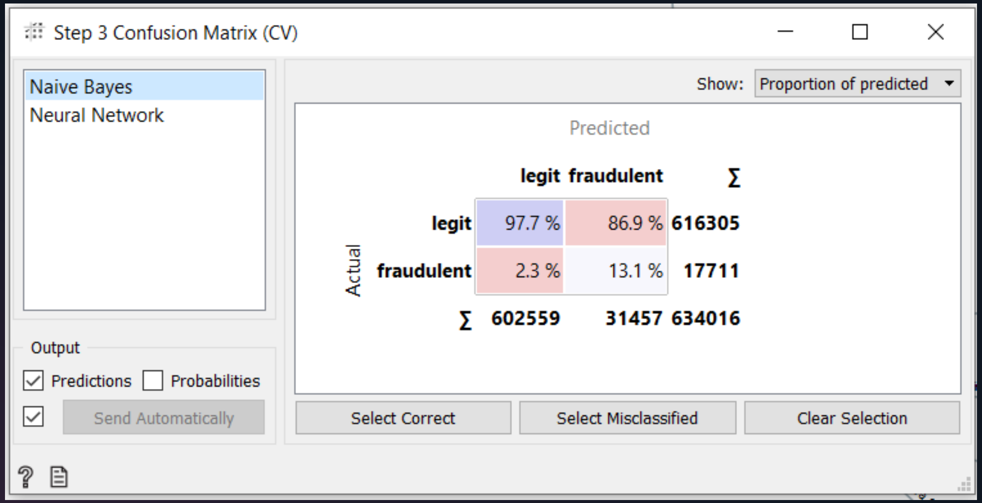
Fraudulent Provider Prediction Modeling Results

Model Outcomes				
Sampling Type	Model	AUC	F1	Precision
70/30 Split	Naïve Bayes	0.934	0.918	0.191
70/30 Split	Neural Network	0.958	0.972	0.121



Fraudulent Provider Prediction Modeling Results

Model Outcomes				
Sampling Type	Model	AUC	F1	Precision
Cross Validation (10 Folds)	Naïve Bayes	0.944	0.935	0.165
Cross Validation (10 Folds)	Neural Network	0.958	0.972	0.121



Conclusion

- Semi-supervised learning with pseudo-labeling yielded results
- Overall model prediction hampered by small proportion of fraudulent labeling
- Additional data sources and enhancements to the NLP analysis needed to increase fraudulent labeling
- Established Value in Developing CMS Part D Opioid Provider Labeled Dataset Process and Prediction Workflow

Lessons Learned

- Inaccurate data mining assumption caused delays and mis-identification of fraudulent providers
- Basic NLP-based entity / location extraction not sufficient to effectively perform data mining
- Early identification and proper understanding of machine learning requirements needed

Future Considerations

- Opioid Provider Analysis
  - Incorporate time component and provider trend analysis
  - Extract fraudulent activity date / timeframe to perform more granular analysis
- Labeled Dataset Construction
  - Identify additional fraudulent provider data sources to increase labeled instances
  - Improve NLP analysis to consider entity position within published text
  - Improve likely legitimate opioid providers determination
  - Develop weighting mechanism to strongly classify providers charged on multiple occasions
- Modeling
  - Incorporate feature engineering to improve results

THANK YOU!

William Rubin

Email:
wrubin1@umbc.edu

GitHub:
https://github.com/warubin410/DATA606Spring2020

Google Site:
https://sites.google.com/umbc.edu/data606/home/willie-rubin