Project Name: SF Investment Analysis

Team members and contact information:

Anya Stepnova	astepnova@berkeley.edu
Chad Wakamiya	cwakamiya@berkeley.edu
Nhut Nguyen	nhutnguyen@berkeley.edu

San Francisco Safety & Crime Analysis

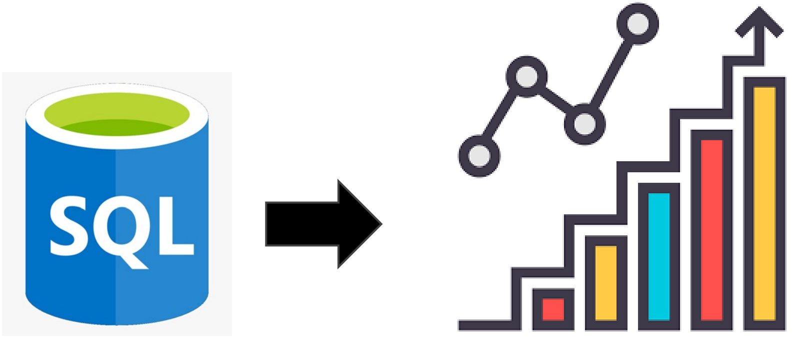
Anya Stepnova
Chad Wakamiya
Nhut Nguyen

IEOR 135 | Data-x
Spring 2020

Forecasting Crime in San Francisco

Why?

- Everyday ~150 crimes are reported in SF.
- Where should you live in SF?
- What type of crimes are most common?



SF City Public Data

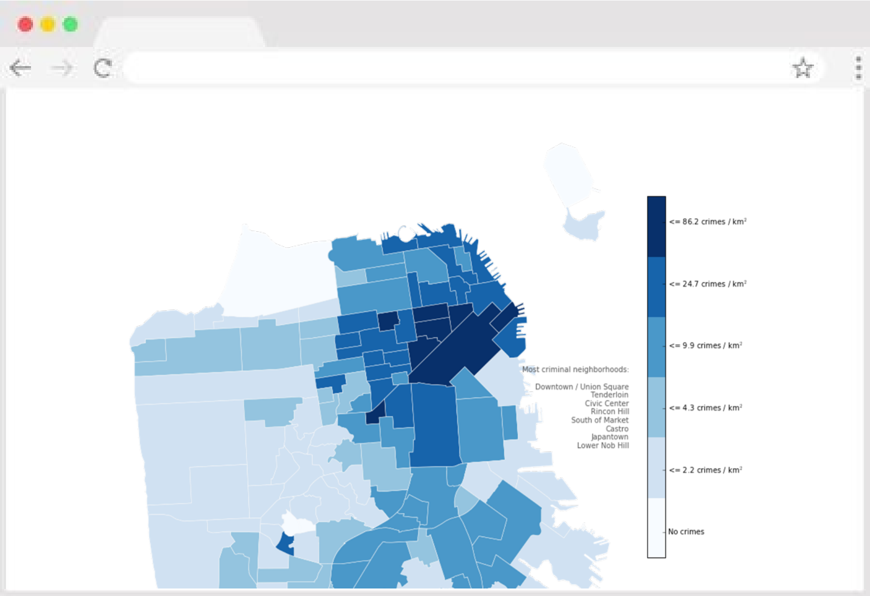
Graphs + Forecast + Insights

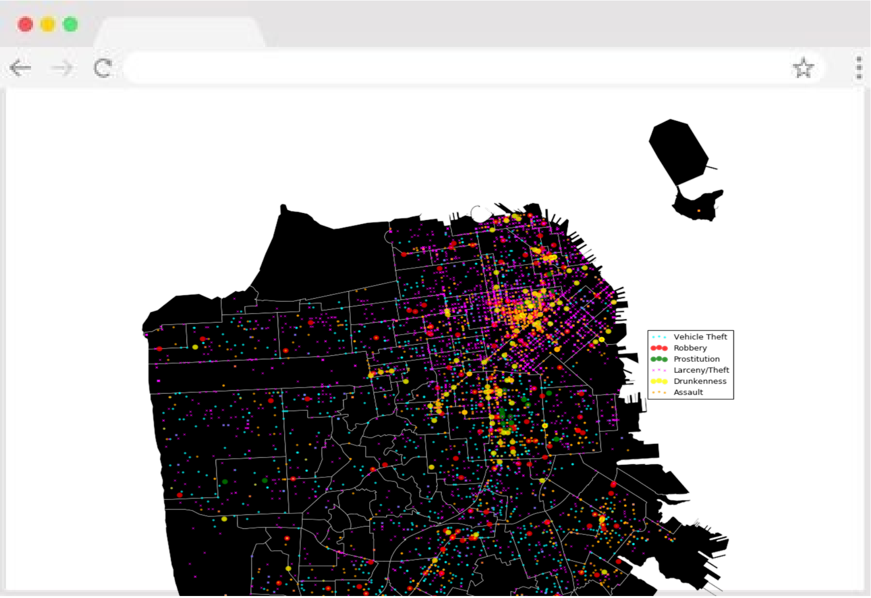
Our Approach

An **interactive interface** for investors, police, and people living or planning to live in SF.

+ Organize public data into easily accessible and understandable views
+ Help police decide where to deploy resources

User Interface Demo





List Top 3 User Requirements

1. Accuracy
2. Easy Navigation
3. Query Speed

Interface Features

- Crime heat maps
- Interactive drop down menus
- Buttons to switch views

Technical Components of Project

Top Components Required for the Project in Order of Importance

Highest Priority

⬆

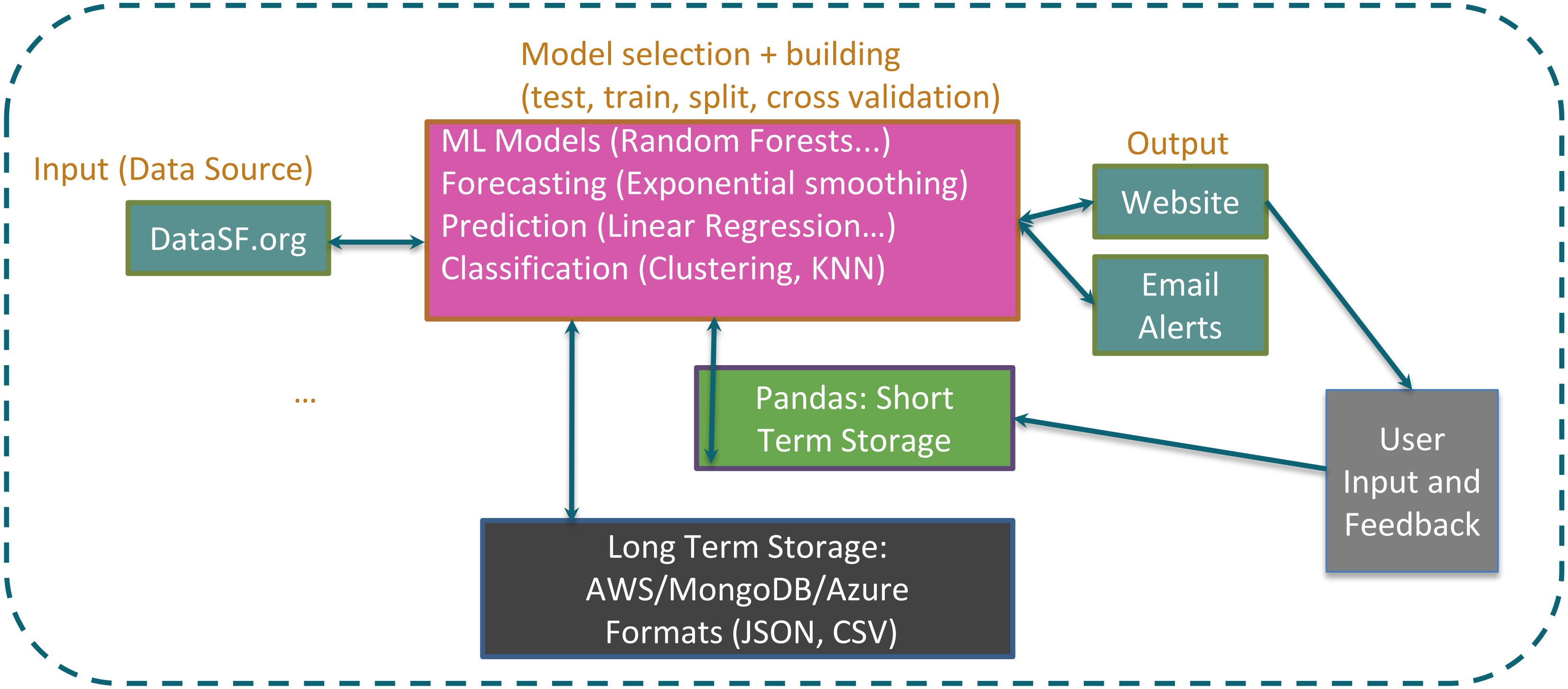
Lowest Priority

1. Selecting best ML, classification, and forecasting algorithms
2. Joining and cleaning data
3. Comprehensive, accurate, reliable data
4. Developing user interface
5. Creating and selecting best visuals
6. Ability to refresh current data

Legend

- Red: High difficulty to develop
- Orange: Medium difficulty to develop
- Green: Easy to develop

Data Model



Model selection + building
(test, train, split, cross validation)

ML Models (Random Forests...)
Forecasting (Exponential smoothing)
Prediction (Linear Regression...)
Classification (Clustering, KNN)

Input (Data Source)

DataSF.org

...

Output

Website

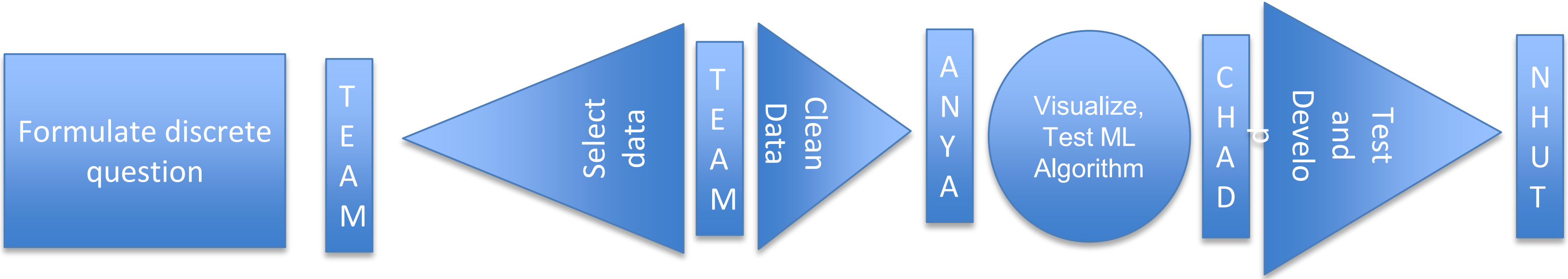
Email Alerts

Pandas: Short Term Storage

User Input and Feedback

Long Term Storage:
AWS/MongoDB/Azure
Formats (JSON, CSV)

What will you do next



Formulate discrete question

TEAM

Select data

TEAM

Clean Data

ANYA

Visualize, Test ML Algorithm

CHAD

Test and Develop

NHUT

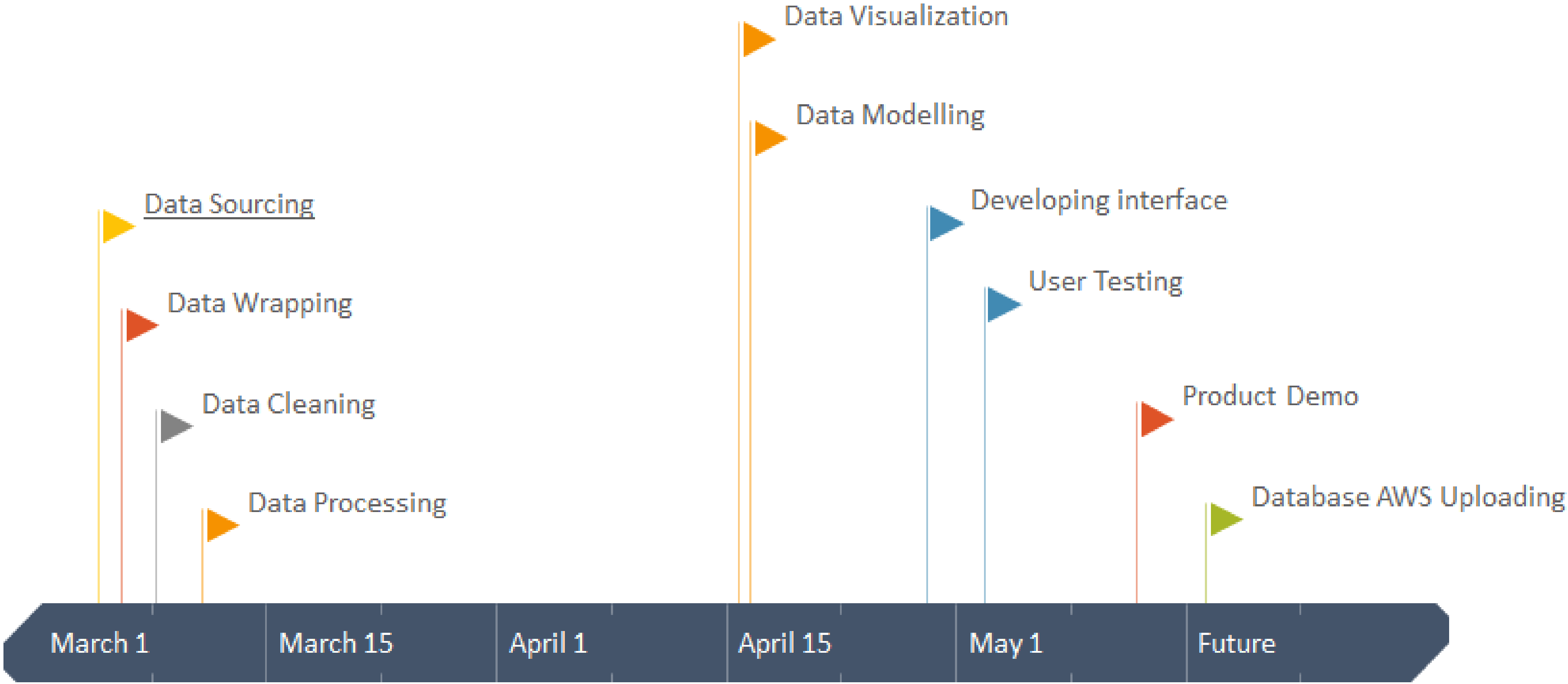
Storm
1-2 Week

Norm
1
Week

Minimal
Viable

Scrum
1-2 Week
Code +
Test

Project Timeline



Thanks for listening.
Questions?

Need

Who needs it and why?

People investing in SF property, business owners, low-medium budget entrepreneurs looking for potential to grow

Approach

What approach will be used?

Merging multiple public datasets from data.sfgov.org on crime, property tax, locale, and street art. Analyze data and build an interactive interface to view the data. Our goal is to provide a tool that will be able to predict trends in SF infrastructure development based on current data provided by the city.

What is the benefit?

Help potential investors predict and analyze crime around SF neighborhoods.
Providing the tool for individuals interested in direction of infrastructure development in certain area in order to expedite their decision upon investing into real estate or business location.
Identifying emerging “under the radar” “hot spots”.

Benefit

What is the next best alternative if this project was not done?

Online articles may providing similar statistics and data visualizations, however, they are not centralized and difficult to find.

Why is this a good space?

How will others in the industry react?

The purpose of the project is to expose publically available data through an interactive interface. There are no competitors in this industry as we strive to empower the community through providing new views of data that may be of interest to current and potential SF property owners/investors.

Online project folders,

Add links here

- Project Folder
- Online Public SF Database

How will you win?

- Answer:

1). We are coordinating our team resources in the most effective ways (previous skills bank and early identification of methods and resources to learn).
2). We are starting data exploration early.
3). We are exploring what has been already done and make sure to offer more extensive information source.

What is Working/Known:

- Answer:

1). We are working on additional data collection (unknown).
2). We have access to the public data set of crime in San Francisco dating from 2003 to 2019 (known).
3). We are working on available data cleaning and visualization in Jupyter Notebook (known/unknown).

What is Not

- Answer:

1). Variables of interest
2). Additional data sets to convert data into reasonable format (address to coordinates, coordinates to zip code, etc.)
3). The type of clustering algorithm we should apply (didn't explore visualizations yet)

Reflection 1:

- Answer:

As we has started our data explorations early, we run into some technical issues we are solving as we go. That highly justifies our early start as difficulties are expected.

Reflection 2:

- Answer:

Accelerating our data exploration shifts many unknown aspects of our data and ways of application into a realm of known that helps to outline our further actions. We learn as we go.

- Answer: Deirdre Quillen

Log Date: 3/9/2020

How will you win?

- Answer:

We will narrow down the data that make sense in decision making for our project. We will group certain data points into clusters and create one hot encoded variables that will empower us to implement different predictive models in order to beat the baseline.

What is Working/Known:

- Answer:

We created a variety of visualizations in Jupyter notebook including a barplot showing which types of crime are the most common and line graphs showing the number of weekly crimes for different crimes. We succeeded in identifying patterns and selecting a few types of crimes that appear to have seasonal patterns that could be forecasted. Next, we will join this data with weather, housing, geographical, and income data.

What is Not :

- Answer:

We planned to merge all datasets but they do not have the same format. The dataset were collected and recorded in different methods between 2003-2018 and 2018 to present. Moreover, it will be time consuming to train the merged dataset and our computers are not strong enough to handle the job.

Reflection 1:

- Answer:

We are equipped with numerous Jupiter libraries and methods and online resources in order to implement our intended winning scenario.

Reflection 2:

- Answer:

We were able to plot and visualize trends in the data this week. This will be helpful for us to visualize the time series that we will be forecasting. We also have validated that there is enough clean data to work with and build robust models.

- Answer: Deirdre Quillen

Log Date: 3/20/20

How will you win?

- Answer:

We will look at the potential of our data and models from the point of view of people that might use this analysis (police department, medical workers, regular citizens) and create a useful and interactive interface to aid their human resources and logistics decision making.

What is Working/Known:

- Answer:

We continued exploring various factors that are correlated with crime including weather and wind speed. We also implemented some forecasting models to predict the number of crime per month. Some of the models include SARIMA, Triple Exponential Smoothing (TES), Random Forest, Adaboosting. Our models have decent accuracy with around a 10% mean absolute percentage error (MAPE).

What is Not :

- Answer:

The scope of our project is very broad now. We have to define our user again to adjust our model. The visualization is not precise enough and we are working to put some more layout and rendering to make it better. The Logistic Regression model has very low accuracy at the moment so we are thinking to adjust the feature selections. The idea of UI/UX for this project is also pending.

Reflection 1:

- Answer:

We will identify the type of data each user of interest may be capable of providing in order to narrow the output to the immediate interest of the user.

Reflection 2:

- Answer:

We were able to learn about various forecasting techniques (SARIMA, TES) and about applying machine learning models. Since the models we created are decently accurate, we can now focus on defining the use case and applying them.

- Answer: Deirdre Quillen

Log Date 4/6/20

How will you win?

- Answer:

We are introducing a very comprehensible interface in order to make our crime prediction model useful and simple for an end-user. The most reasonable user for our model is law enforcement agency that will be able to predict distribution of crime per district, per type of crime, etc. and will be able to calculate their internal forces allocations based on predictions.

What is Working/Known:

- Answer:

We used Voila to create a basic UI. The interface allows a user to select a specific type of crime and then visualize the historical crime patterns since 2003. Our machine learning model then forecasts the number of monthly crime incidents for the next few months. The UI contains a number slider that allows a user to specify how many months into the future they would like the forecast to predict.

What is Not

- Answer:

We tried to customize our project for specific end users such as police or homebuyer. However, the data is not comprehensive and the our tool is also not strong enough to make it. We also used different time series model to make the prediction but the run time is too long and the result is not our expectation

Reflection 1:

- Answer:

It is good to learn more explore and apply different time series model during this phase. We also learn how to restructure our notebook in the most efficient way and prepare for our final delivery.

Reflection 2:

- Answer:

We taught ourselves how to use the Viola and Plotly Python packages to make the graphs and the interactive UI. So far both seem to be working well. In the next few weeks, we are going to learn new features in those packages to improve how the UI looks and user functionality.

- Answer: Deirdre Quillen

Log Date: 4/20/20