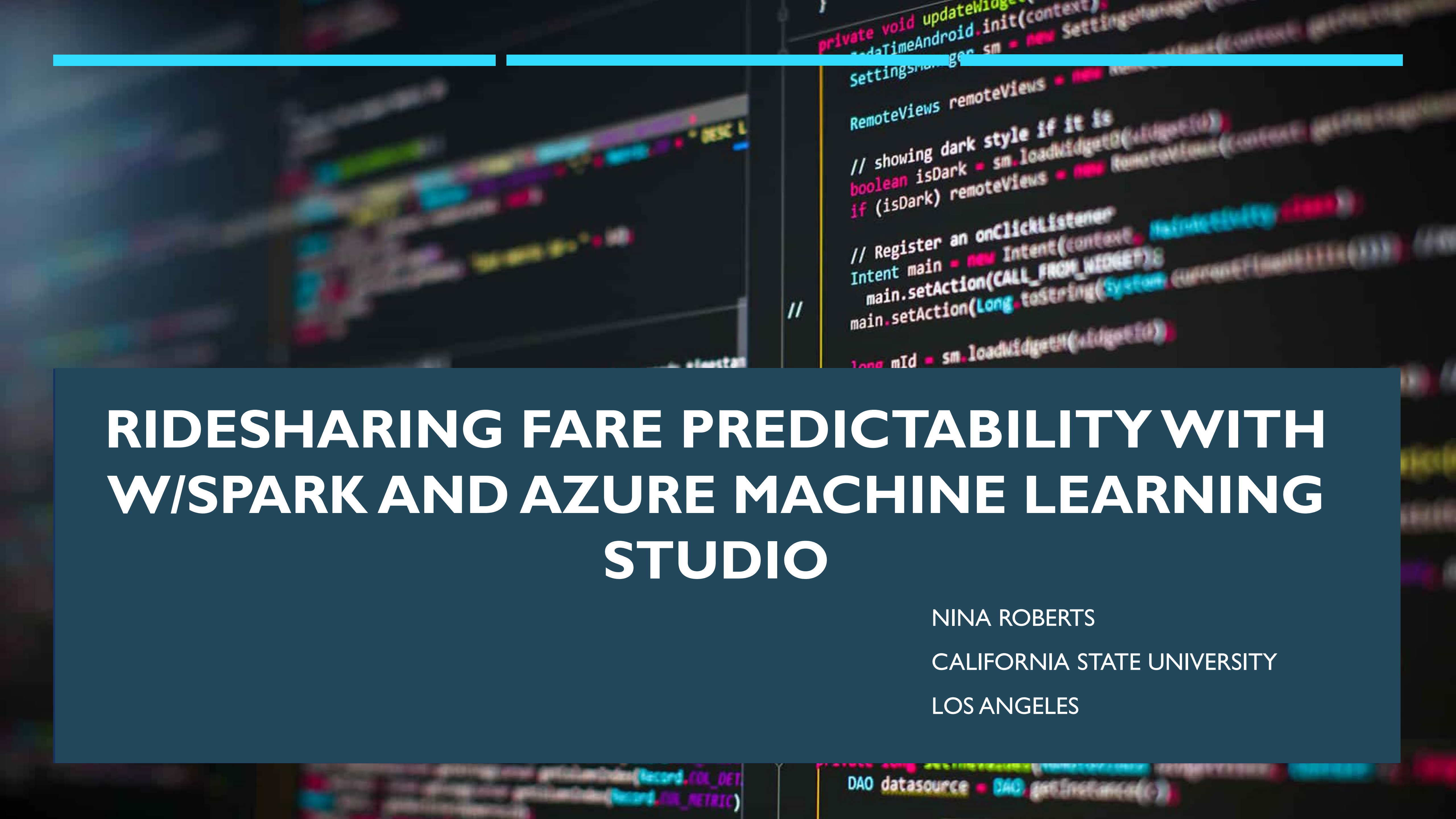


RIDESHARING FARE PREDICTABILITY WITH W/SPARK AND AZURE MACHINE LEARNING STUDIO

NINA ROBERTS

CALIFORNIA STATE UNIVERSITY

LOS ANGELES

INTRODUCTION

- Uber has recently been in the news for several topics:
  - The use of data science to create the algorithm for surge pricing created to incentivize drivers.
  - Lawsuits by cities due to unfair pricing practices.
- Can we use Data Science to predict Uber's surge pricing algorithm?

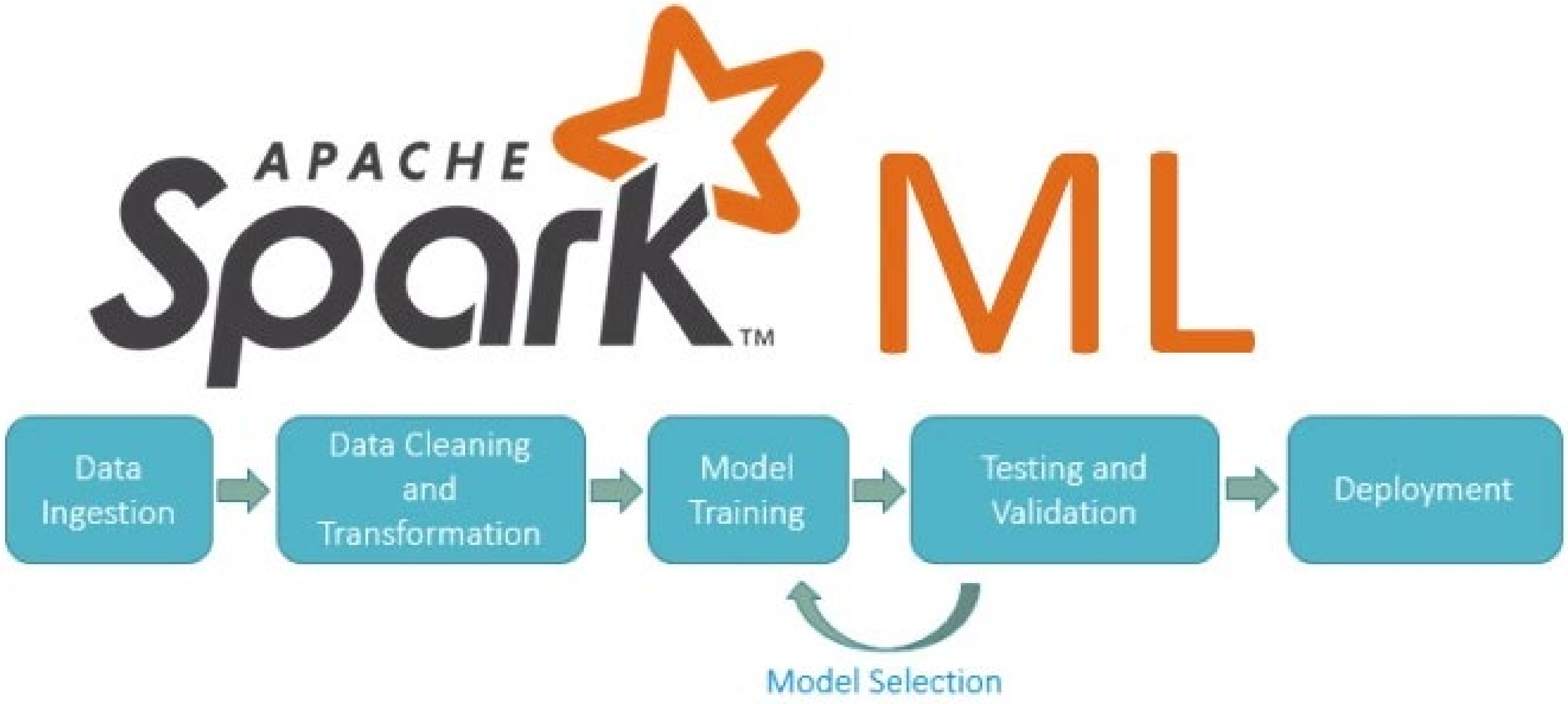
GITHUB

- https://github.com/NinaRo2/CIS5560

TECHNOLOGY STACK

- JupyterLab
- DataBricks
- Hadoop and Oracle Cloud
- Azure ML Studio
- AWS S3

SOFTWARE AND TOOLS



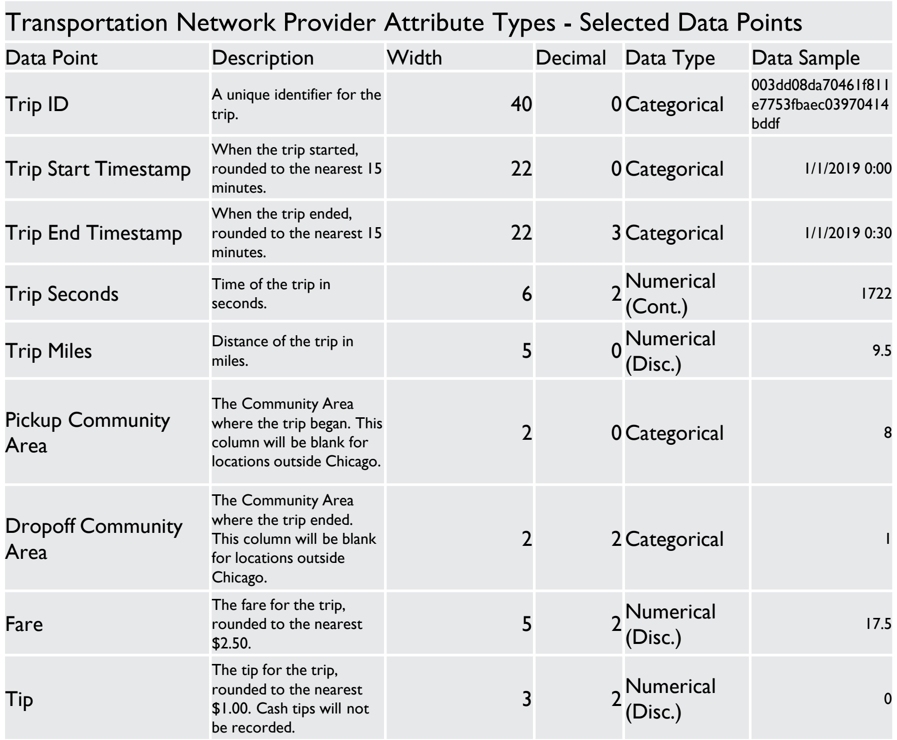
DATASET

CHICAGO
DATA PORTAL

- https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips-2019/iu3g-qa69

- 2.2 gigabytes, 8,675,393 rows and 21 columns.

DATASET - SELECTED DATA POINTS

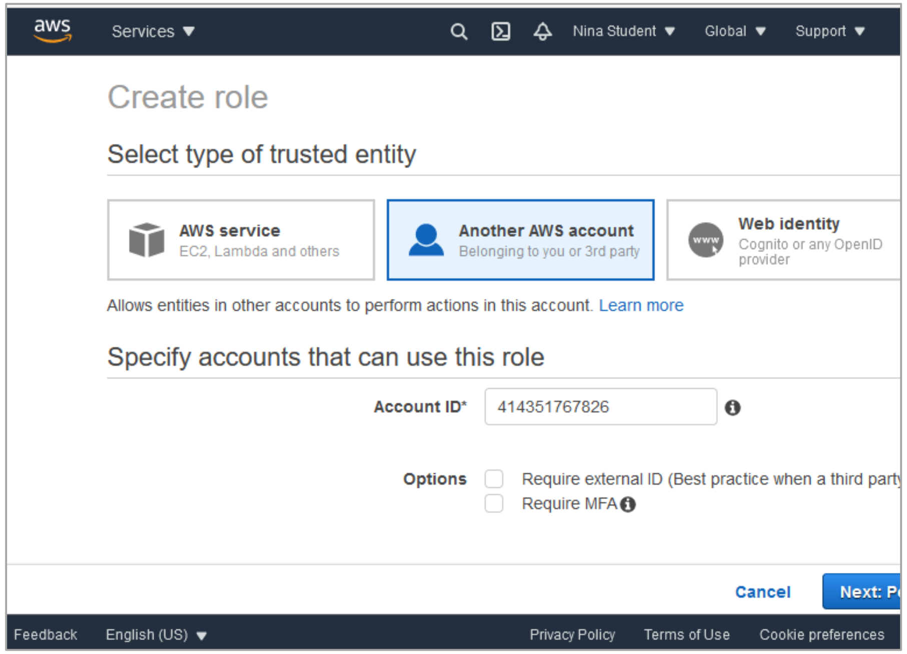


Transportation Network Provider Attribute Types - Selected Data Points					
Data Point	Description	Width	Decimal	Data Type	Data Sample
Trip ID	A unique identifier for the trip.	40	0	Categorical	003dd08da70461f811e7753fbaec03970414bddf
Trip Start Timestamp	When the trip started, rounded to the nearest 15 minutes.	22	0	Categorical	1/1/2019 0:00
Trip End Timestamp	When the trip ended, rounded to the nearest 15 minutes.	22	3	Categorical	1/1/2019 0:30
Trip Seconds	Time of the trip in seconds.	6	2	Numerical (Cont.)	1722
Trip Miles	Distance of the trip in miles.	5	0	Numerical (Disc.)	9.5
Pickup Community Area	The Community Area where the trip began. This column will be blank for locations outside Chicago.	2	0	Categorical	8
Dropoff Community Area	The Community Area where the trip ended. This column will be blank for locations outside Chicago.	2	2	Categorical	1
Fare	The fare for the trip, rounded to the nearest $2.50.	5	2	Numerical (Disc.)	17.5
Tip	The tip for the trip, rounded to the nearest $1.00. Cash tips will not be recorded.	3	2	Numerical (Disc.)	0

- You have two options to divide the presentation into smaller modules.
- One option is to divide the presentation into one module for each member of the group. In this case, you may have submodules as well.
- The other option is to divide the presentation into smaller modules and choose and select the modules that would be presented by each member of the group.

IMPLEMENTATION

- Worked in JupyterLab to split of sample dataset to use in Azure ML Studio and DataBricks Community Edition
- Imported sample files to both Azure ML Studio DataBricks Community Edition
- Created AWS Role and Policy and connected to DataBricks – however trial account did not connect



Create role

Select type of trusted entity

AWS service
EC2, Lambda and others

Another AWS account
Belonging to you or 3rd party

Web identity
Cognito or any OpenID provider

Allows entities in other accounts to perform actions in this account. Learn more

Specify accounts that can use this role

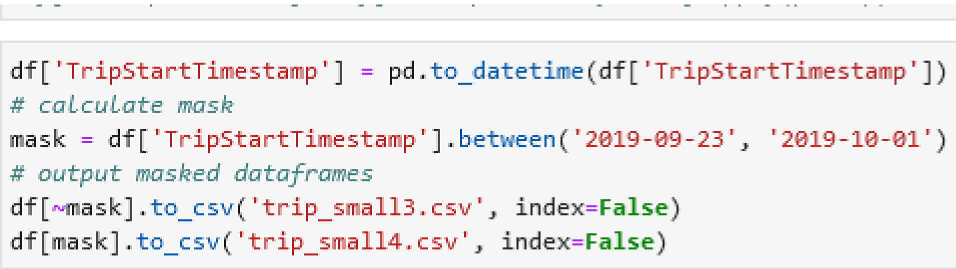
Account ID* 414351767826

Options □ Require external ID (Best practice when a third party
□ Require MFA

DATA ENGINEERING

- Used Python to add a calculated column in JupyterLabs IDE using Lambda function
  - df.assign(AvgFareMile=lambda x: x.Fare / x.TripMiles)

- Split off sample file using parser

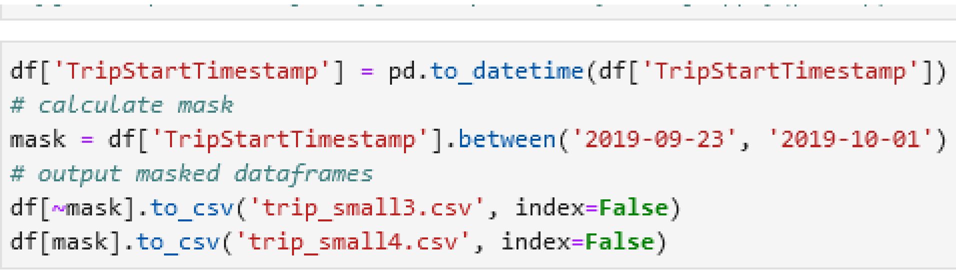


```
df['TripStartTimestamp'] = pd.to_datetime(df['TripStartTimestamp'])
# calculate mask
mask = df['TripStartTimestamp'].between('2019-09-23', '2019-10-01')
# output masked dataframes
df[~mask].to_csv('trip_small3.csv', index=False)
df[mask].to_csv('trip_small4.csv', index=False)
```

DATA DISCOVERIES

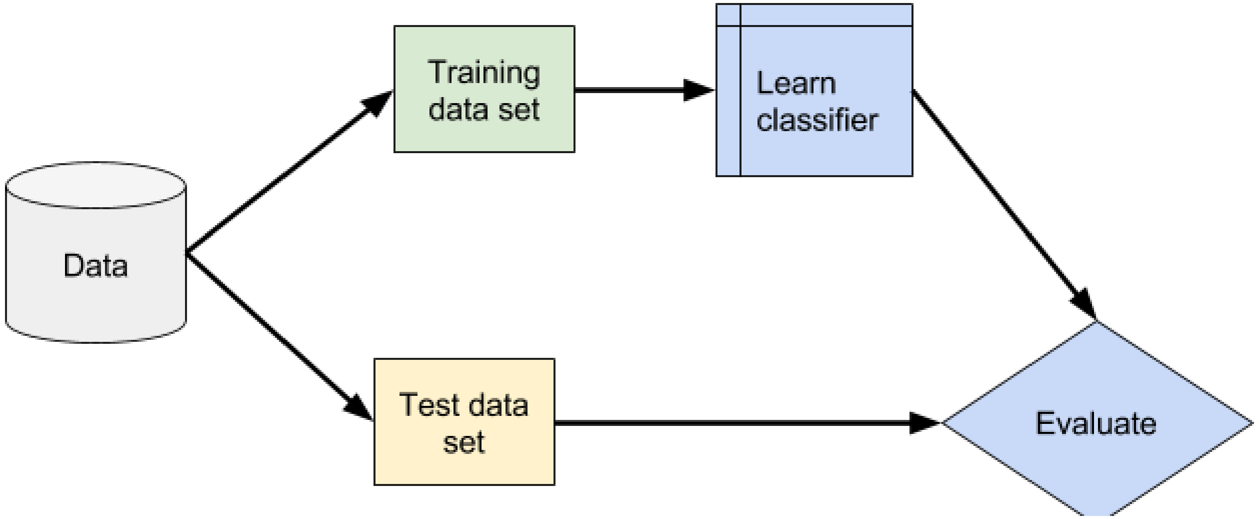
- Used Python to add a calculated column in JupyterLabs IDE using Lambda function
  - df.assign(AvgFareMile=lambda x: x.Fare / x.TripMiles)

- Split off sample file using parser



```
df['TripStartTimestamp'] = pd.to_datetime(df['TripStartTimestamp'])
# calculate mask
mask = df['TripStartTimestamp'].between('2019-09-23', '2019-10-01')
# output masked dataframes
df[~mask].to_csv('trip_small3.csv', index=False)
df[mask].to_csv('trip_small4.csv', index=False)
```

METHODOLOGY



two options to divide the
tion into smaller modules.

ion is to divide the presentation
module for each member of
p. In this case, you may have
les as well.

er option is to divide the presentation into smaller modules and choose and select the modules that would be presented by each member of the group.

MODEL 1

Classification – Two Class Decision Forest w/Permutation Feature Importance

- Text Here

MODEL 2

Multiple Linear Regression w/Parameter Tuning

- Text Here

MODEL 3

TBA

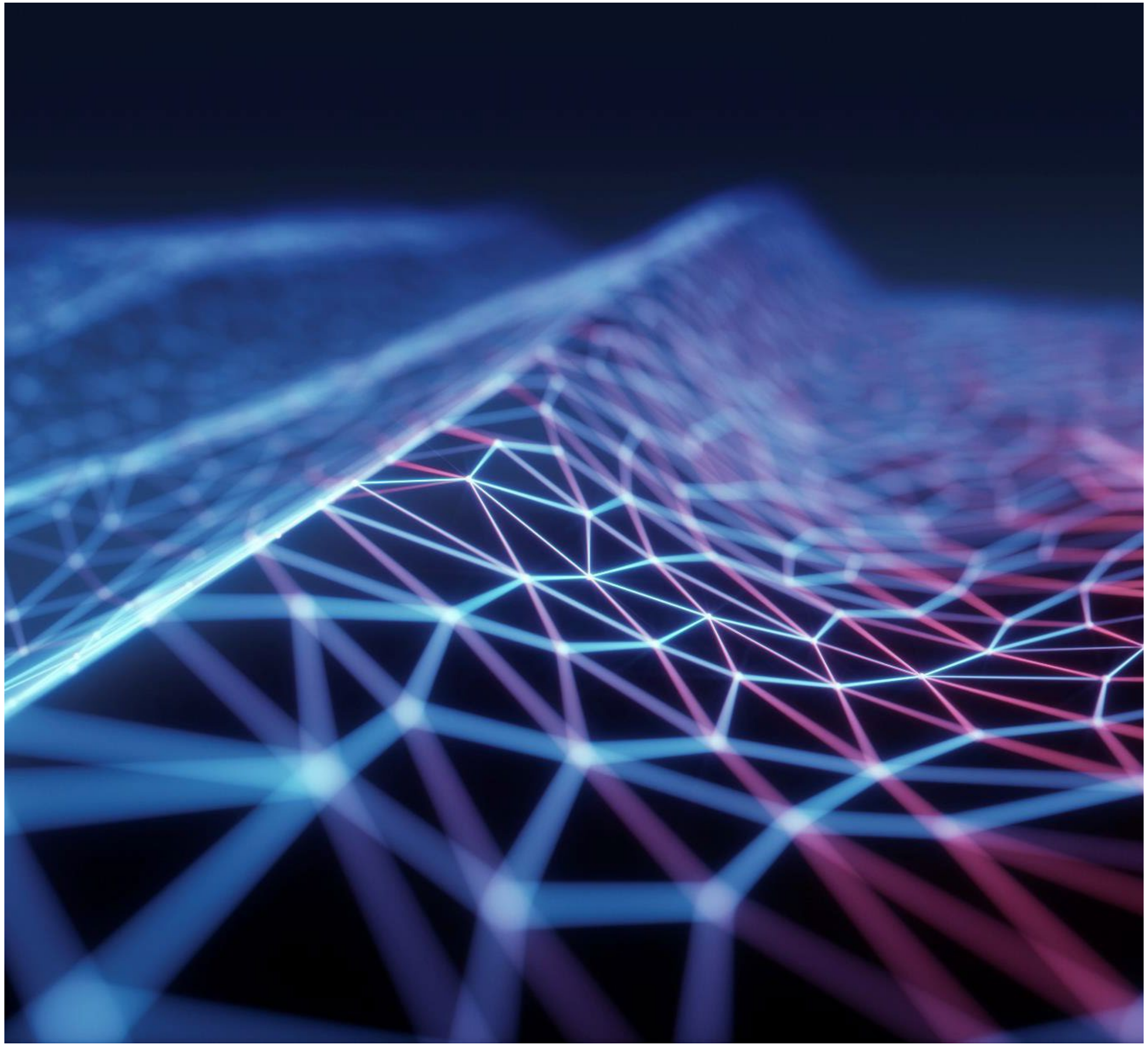
- Text Here

SUCCESSES

- Working with DataFrames is better than working with code in IDEs such as Spyder
- Learned a lot about Data Science

CHALLENGES

- Working with a large sample file of one week in both Azure ML Studio and DataBricks was difficult. File uploads crashed multiple times.
- No background in data science



Q & A