DATASTAX

ACCELERATE

Predicting Restaurant Inspection Failures

https://github.com/angelok1/chicago-inspections-dse/

DATASTAX®

About Us

Me

- Help organizations implement Health IT (FHIR), Data Science (ML, DNN), Big Data (Spark, DSE, Kafka)
- Graduate researcher at Harvard Data Systems Laboratory

github.com/angelok1

https://www.linkedin.com/in/angelok1/

akastroulis@carrera.io

@carraio

You

- Data Scientists (build Models)?
- Engineers (implement Models)?
- New to Data Science (Just want to know more about what this thing is and how it works)?

The Project

https://github.com/angelok1/chicago-inspections-dse

CHICAGO
DATA PORTAL

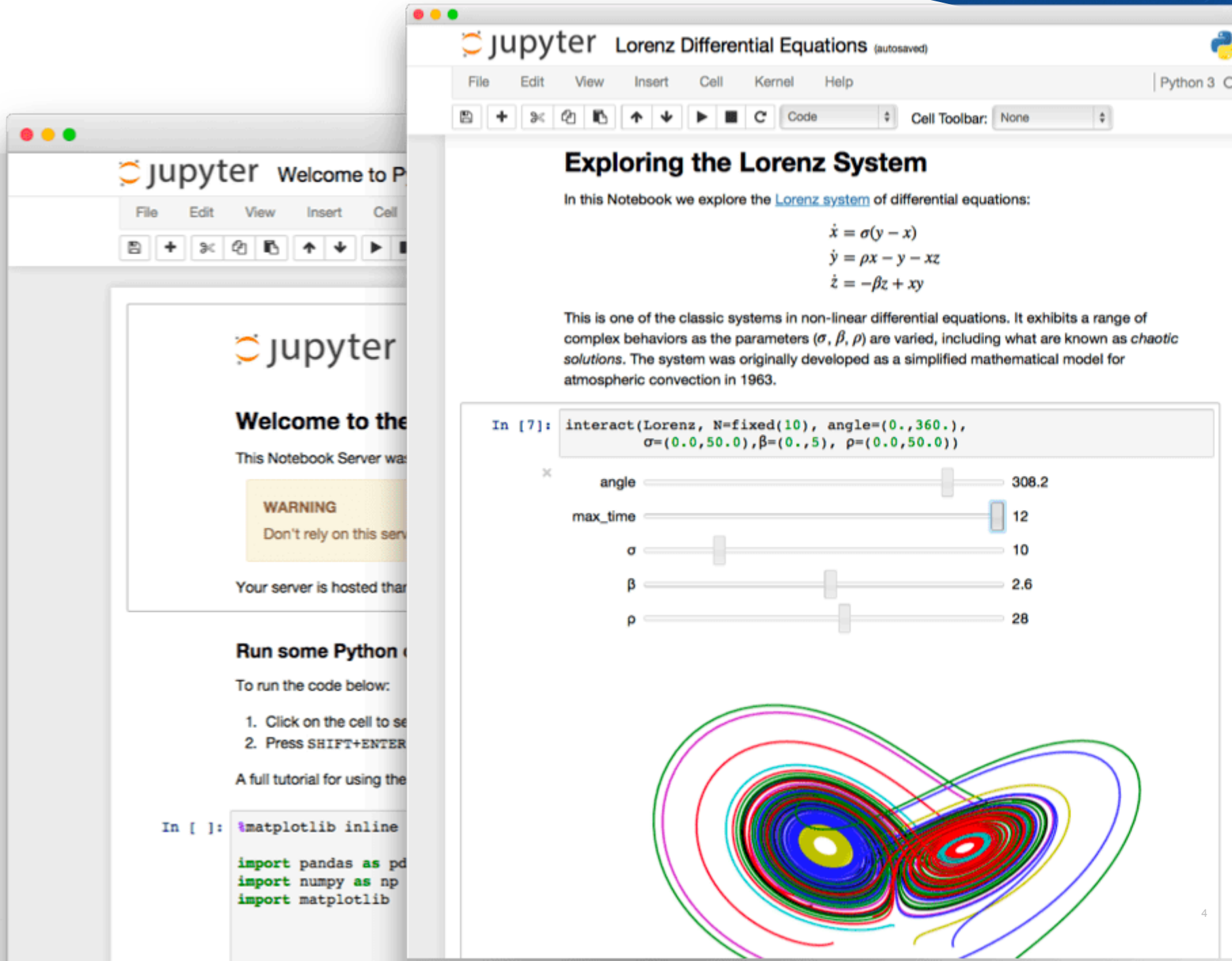
https://data.cityofchicago.org/

The Toolbox

scikit learn

TensorFlow

APACHE Spark™



Carrera GROUP INC

The Problem

- According to the CDC, 47.8M cases of foodborne illness annually
- Leading to 3,000 deaths
- 31 known pathogens but 38.4M come from unspecified agents
- Norovirus, Salmonella, Staph, E. Coli and a bunch of stuff we have no idea what they are.
- Seems like an ideal ML problem

Building Intuition

If we could discover food borne illness quicker, we'd have less exposure.

Where do we get our food?

Data?

github.com/angelok1/chicago-inspections-dse/

Building Intuition

If we could discover food borne illness quicker, we'd have less exposure.

Where do we get our food?

Restaurants, Grocery

Data?

Inspections, Business Licenses

github.com/angelok1/chicago-inspections-dse/

Building Intuition

Building Intuition - QUIZ

Which of these pairs do you think will be most relevant?

Complaint of food poisoning?	Did it snow?
Number of burglaries in the area?	Do they serve sushi?
Garbage in the alley?	How often they are inspected?
Close to a Walmart Supercenter?	Close to the water?

Building Intuition - QUIZ

Which of these pairs do you think will be most relevant?

Complaint of food poisoning?	**Did it snow?**
Number of burglaries in the area?	Do they serve sushi?
Garbage in the alley?	**How often they are inspected?**
Close to a Walmart Supercenter?	Close to the water?

Building Intuition

> “All models are wrong, but some are useful”
>
> — George Box

Carrera GROUP INC

Building Intuition

- Do **sanitation** complaints have anything to do with failures?
  - Where are rodents or bugs likely to infest?
  - If the alley is full of trash, how clean is the neighborhood?
- Does **crime** have anything to do with failures?
  - What if it was a burglary and affected the power or equipment?
- How does **weather** have an effect on equipment?

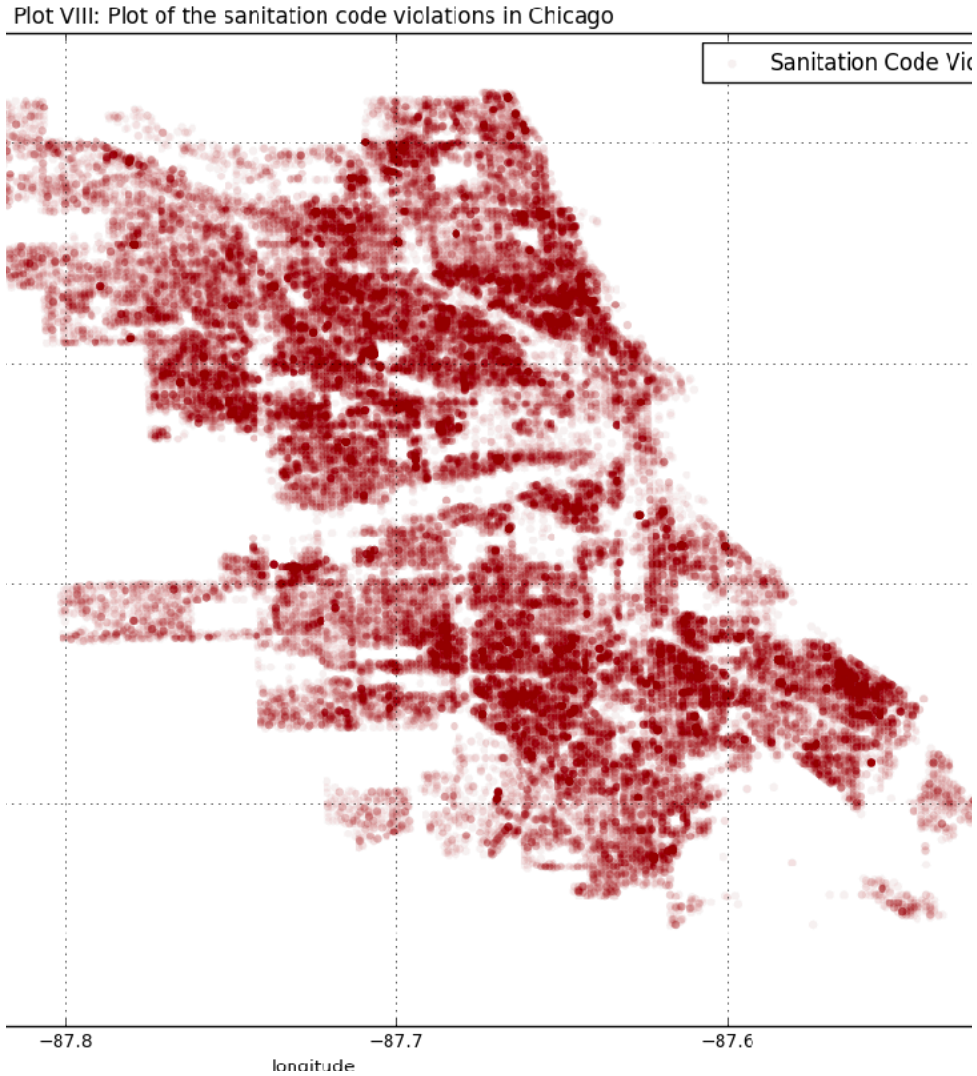
Carrera GROUP INC

github.com/angelok1/chicago-inspections-dse/

DATASTAX®

Exploring Sanitation

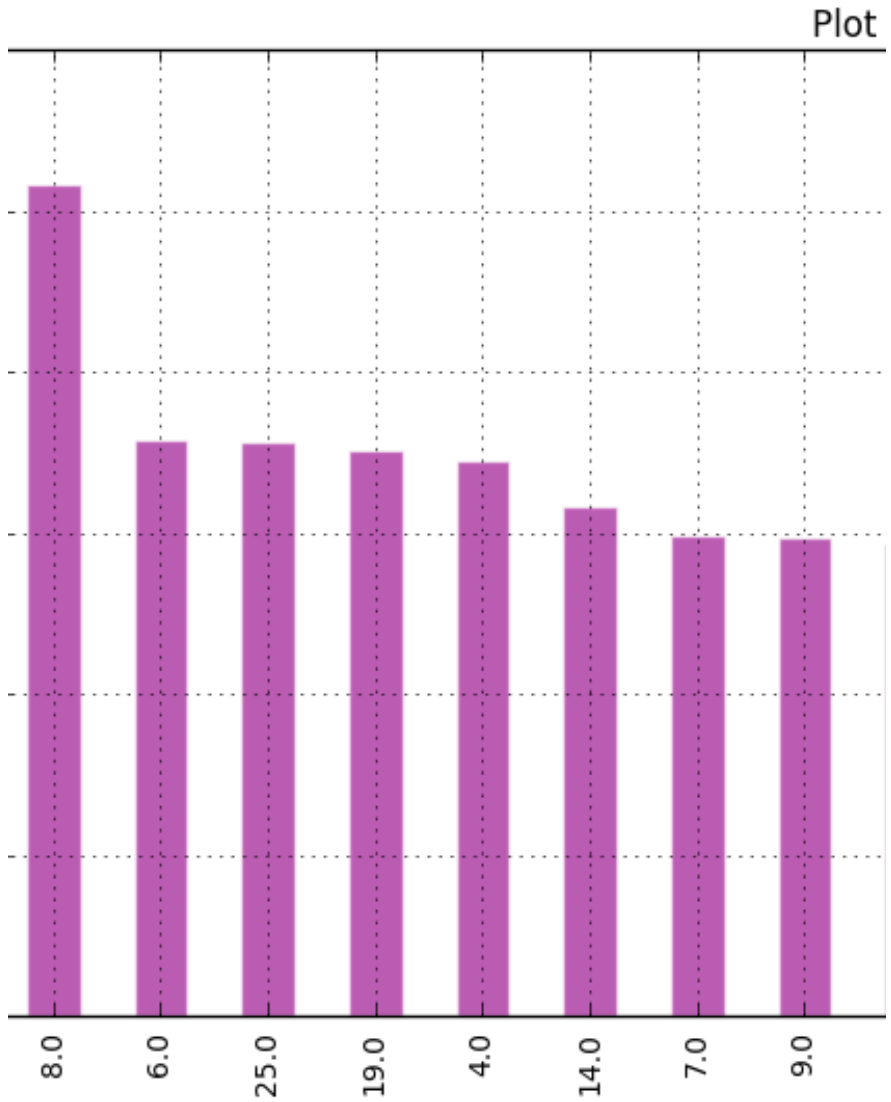
```
scatter_plot(x='longitude', y='latitude', my_df=df_sanitation,
       group='type_of_service_request',
       col=["darkred"], alpha=0.05, size=20, marker='o',
       title="Plot of the sanitation code violations in Chicago”)
```



github.com/angelok1/chicago-inspections-dse/Step4-SanitationData.ipynb

Exploring Sanitation

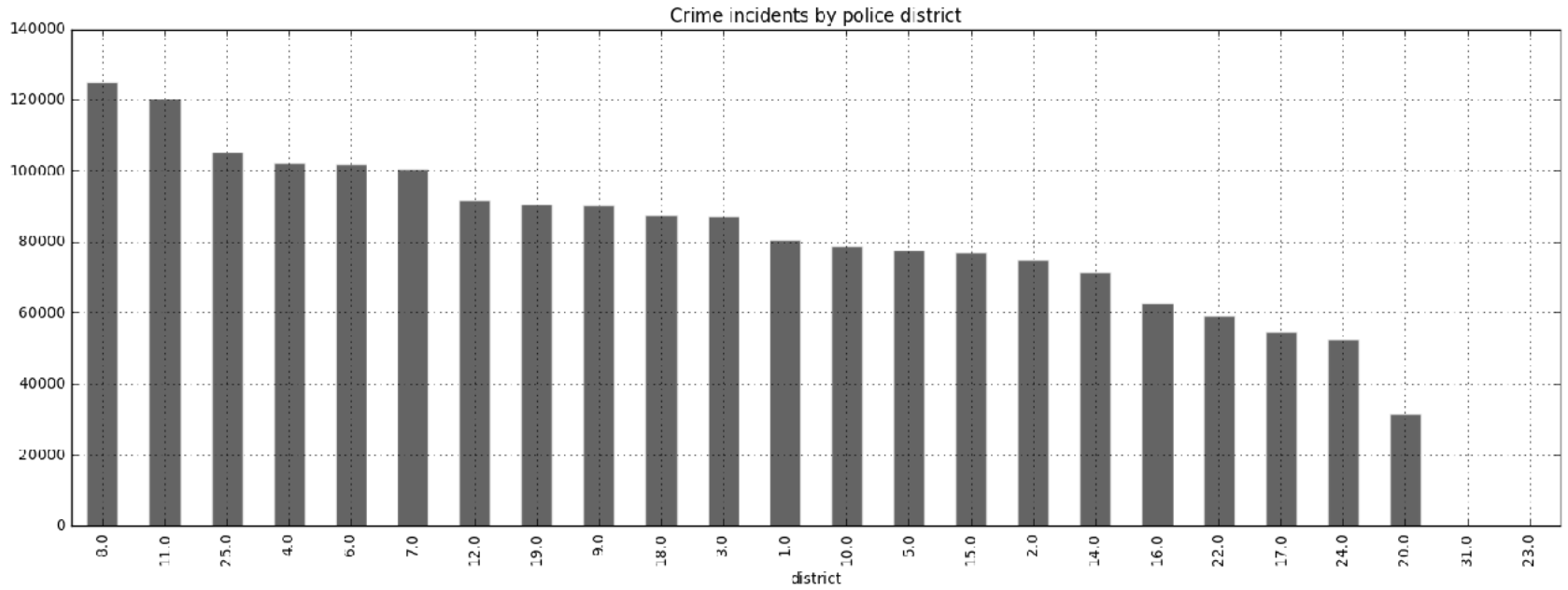
```
bar_plot(my_df=df_sanitation, var_name='police_district',
    color='purple', alpha=0.6,
    title='Number of sanitation code complaints by police district',
edgecolor='white', figsize=(18, 6))
```



github.com/angelok1/chicago-inspections-dse/Step4-SanitationData.ipynb

Exploring Crime

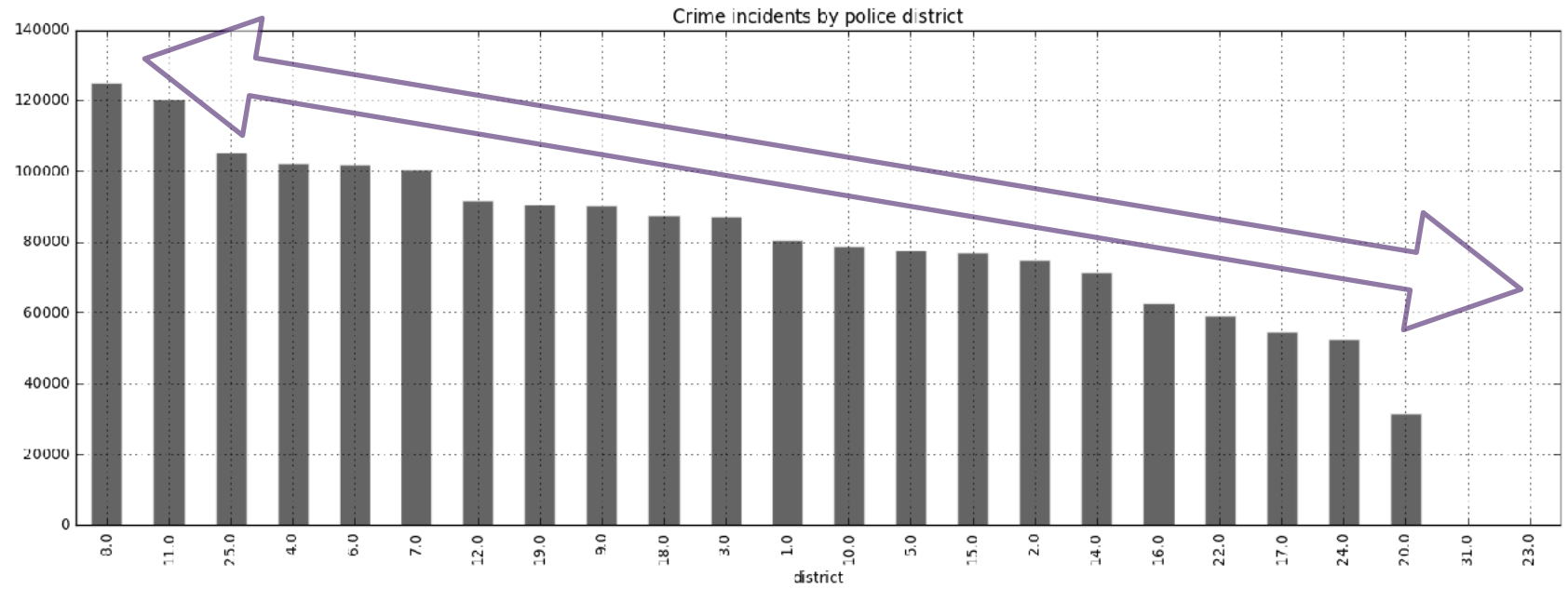
```
bar_plot(my_df=df_crimes, var_name='district', color='black', alpha=0.6,
  title='Crime incidents by police district', edgecolor='white', figsize=(18, 6))
```



github.com/angelok1/chicago-inspections-dse/Step3-CrimeData.ipynb

Exploring Crime

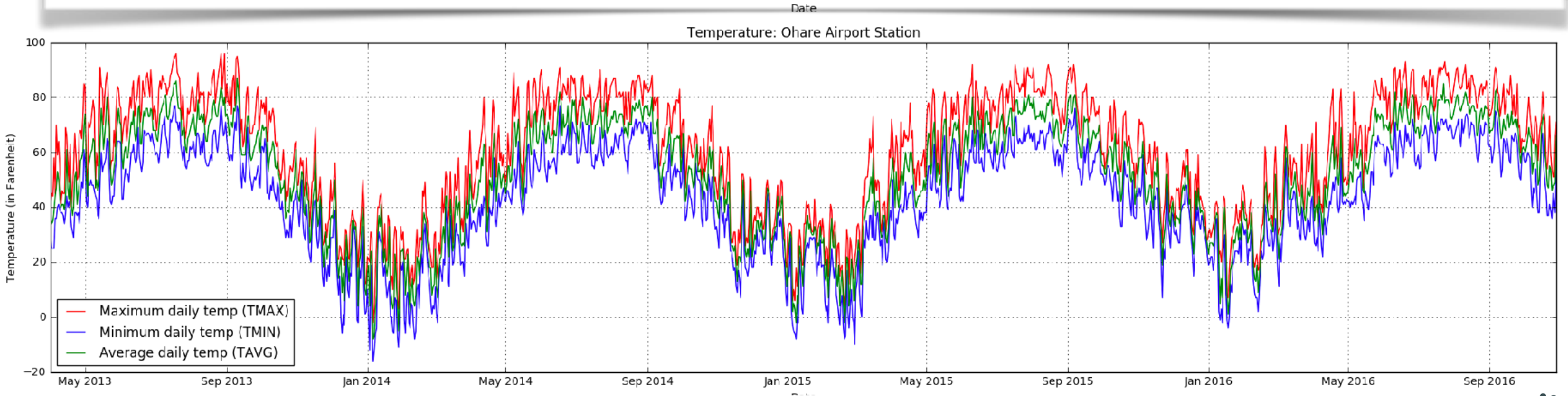
```
bar_plot(my_df=df_crimes, var_name='district', color='black', alpha=0.6,
  title='Crime incidents by police district', edgecolor='white', figsize=(18, 6))
```



github.com/angelok1/chicago-inspections-dse/Step3-CrimeData.ipynb

Exploring Weather

```
ax2.plot(df_ohareairport['DATE'], df_ohareairport['TMAX'], c='r', label='Maximum daily temp (TMAX)')

ax2.plot(df_ohareairport['DATE'], df_ohareairport['TMIN'], c='b', label='Minimum daily temp (TMIN)')

ax2.plot(df_ohareairport['DATE'], df_ohareairport['TAVG'], c='g', label='Average daily temp (TAVG)')

ax2.set_title('Temperature: Ohare Airport Station')
ax2.set_xlabel('Date')
ax2.set_ylabel('Temperature (in Farenheit)')
ax2.legend(loc='best')
ax2.grid()
```



github.com/angelok1/chicago-inspections-dse/Step5-WeatherData.ipynb

Modeling

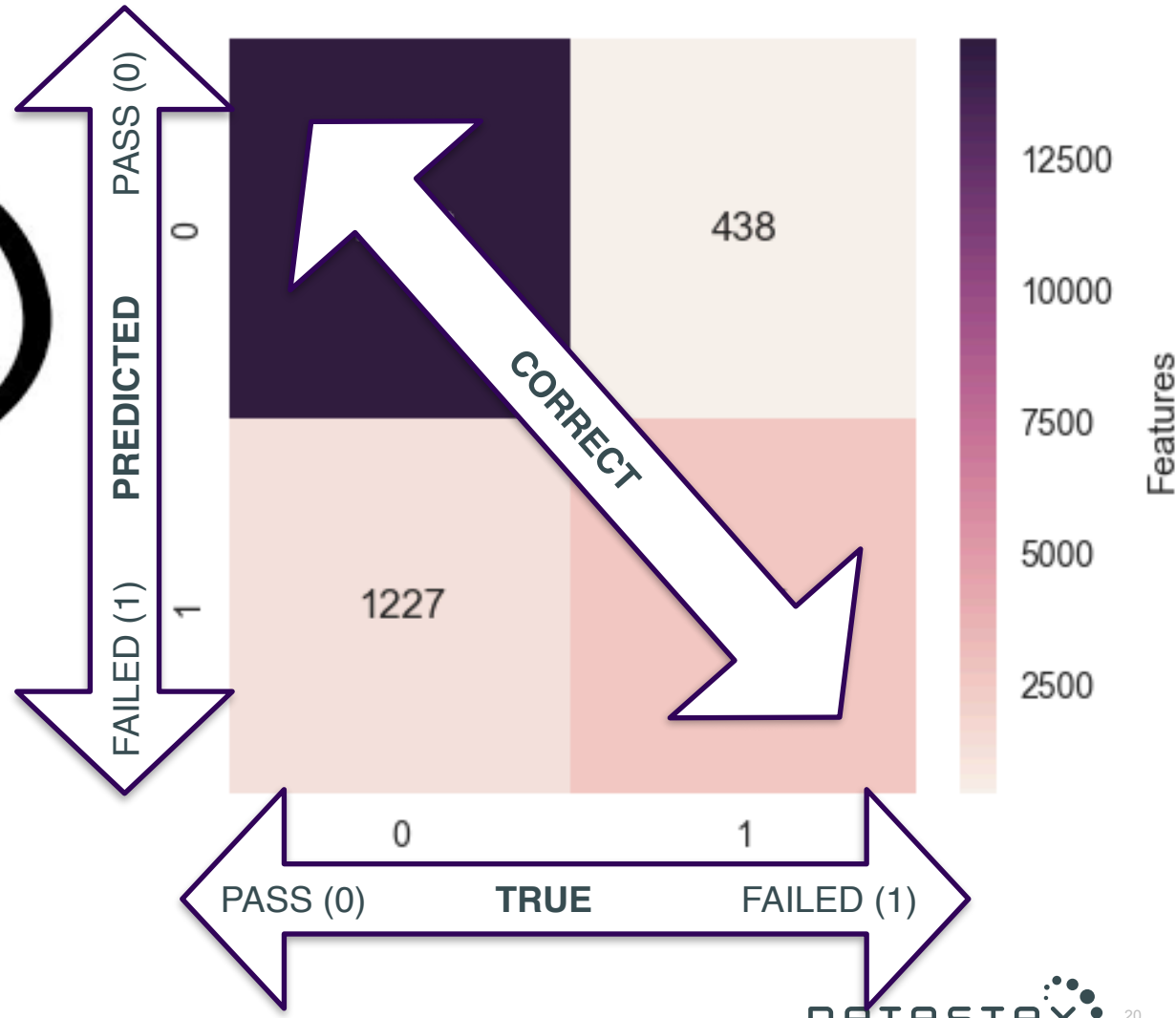
- What are we trying to **accomplish**?
  - We want something to sort on.
- How do we **model** this?
  - NN? time series? ML?
- What kinds of **decisions** do we make?
  - How engineer features?
    Preprocess? Where to store data?
    How will we run this model?

Model accuracy

- Do we use accuracy as the measure of quality (i.e. actual - predicted)
- Remember **everyone** gets inspected.

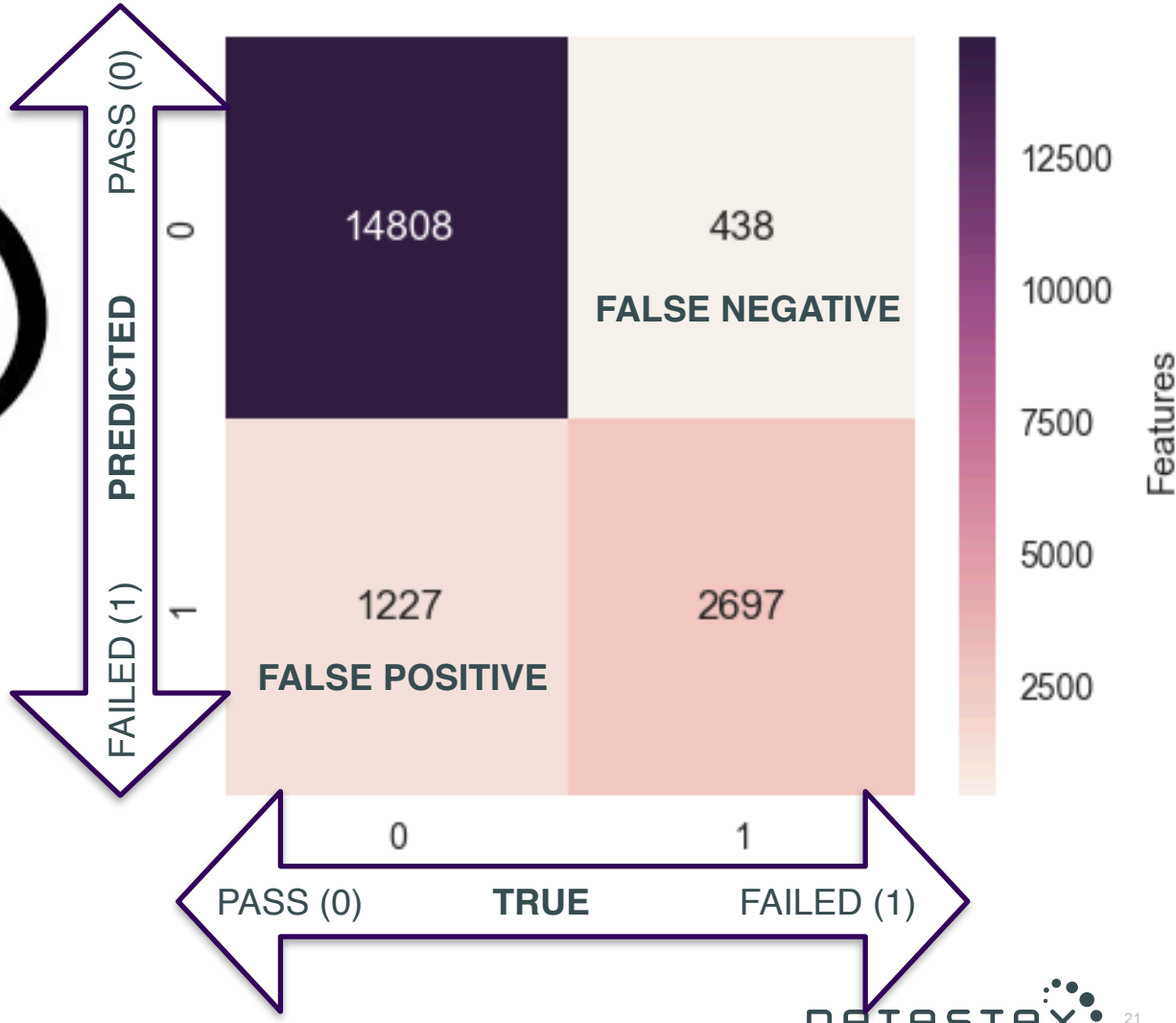
Model accuracy

- Do we use accuracy as the measure of quality (i.e. actual - predicted)
- Remember **everyone** gets inspected.

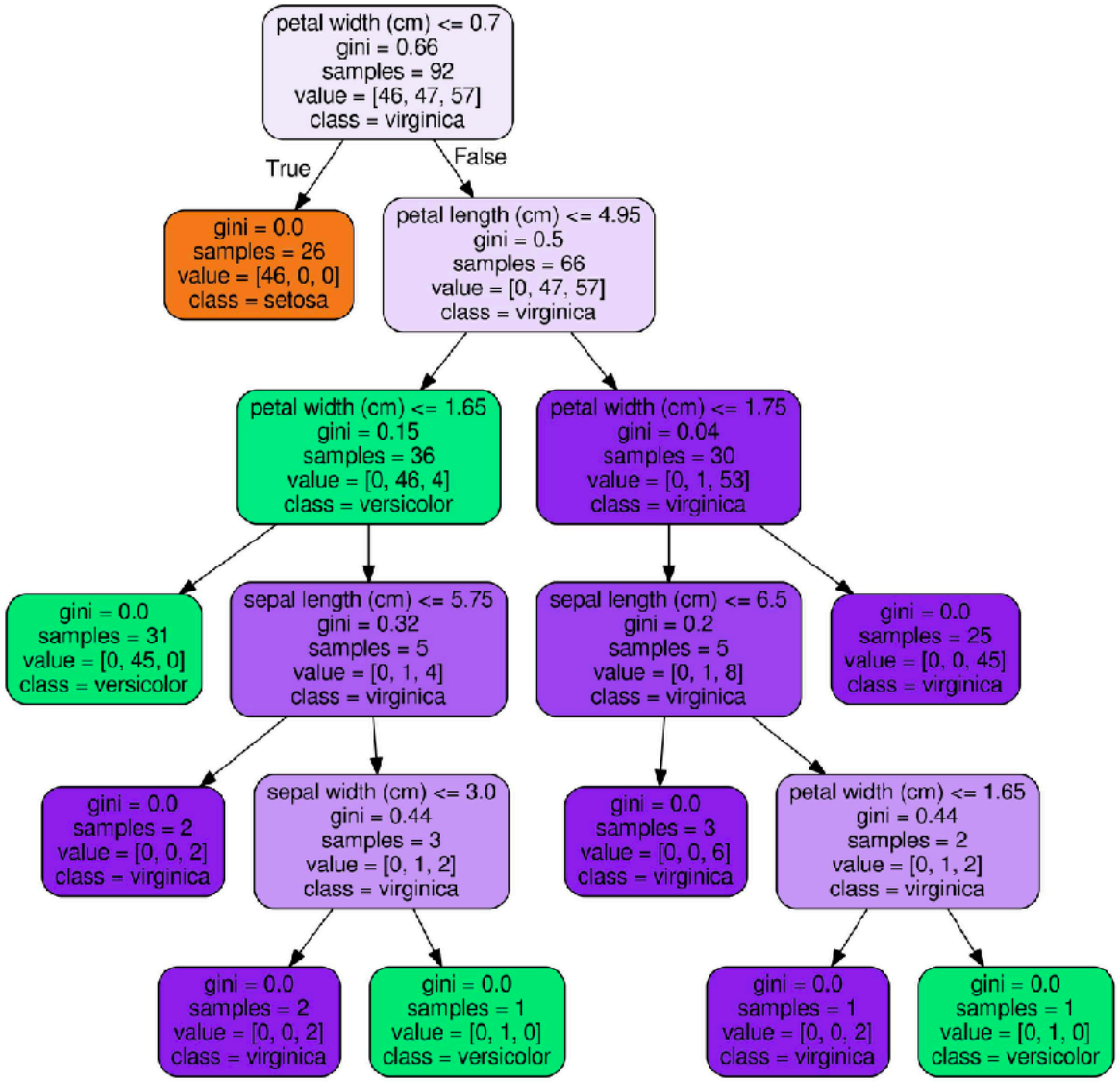


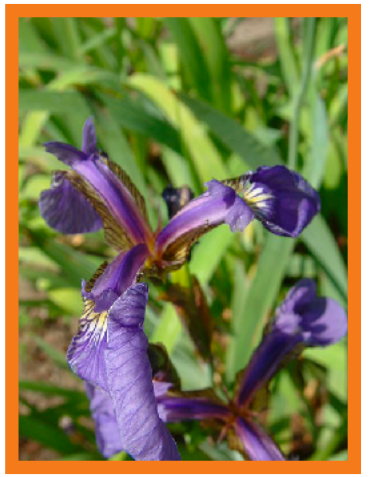
Model accuracy

- Do we use accuracy as the measure of quality (i.e. actual - predicted)
- Remember **everyone** gets inspected.



Decision Tree Classification

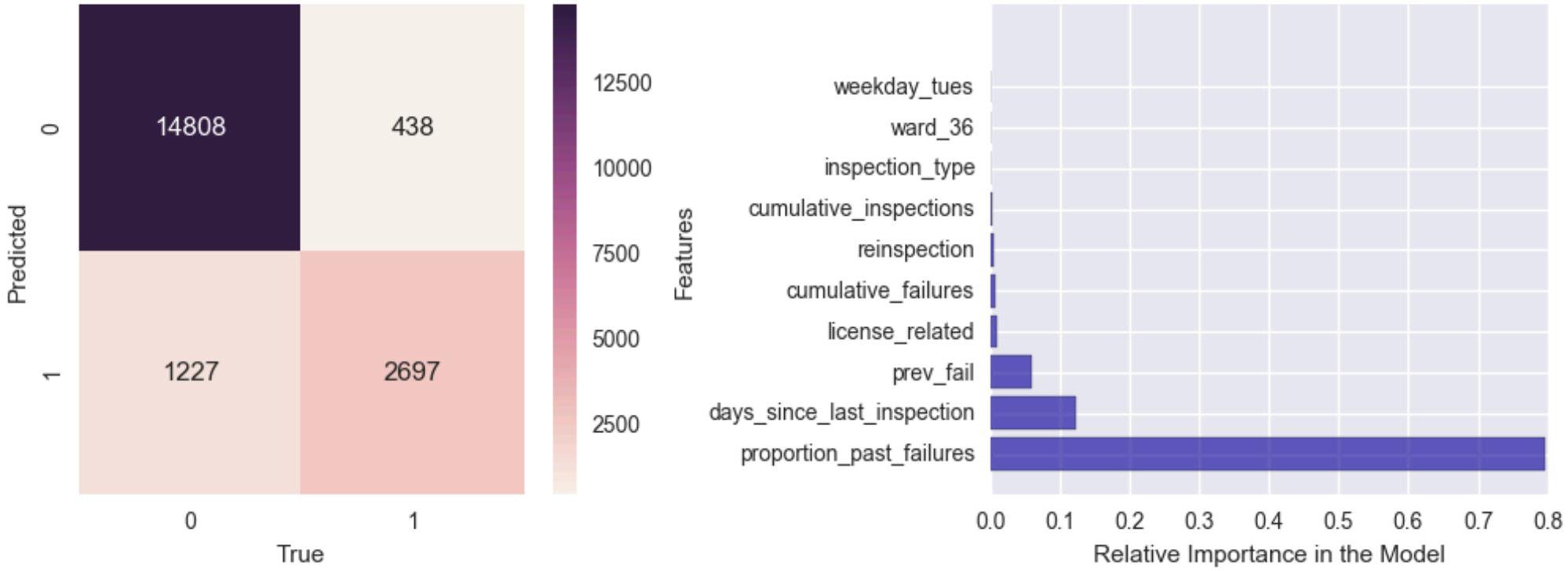




https://en.wikipedia.org/wiki/Iris_flower_data_set

Modeling - Decision Trees

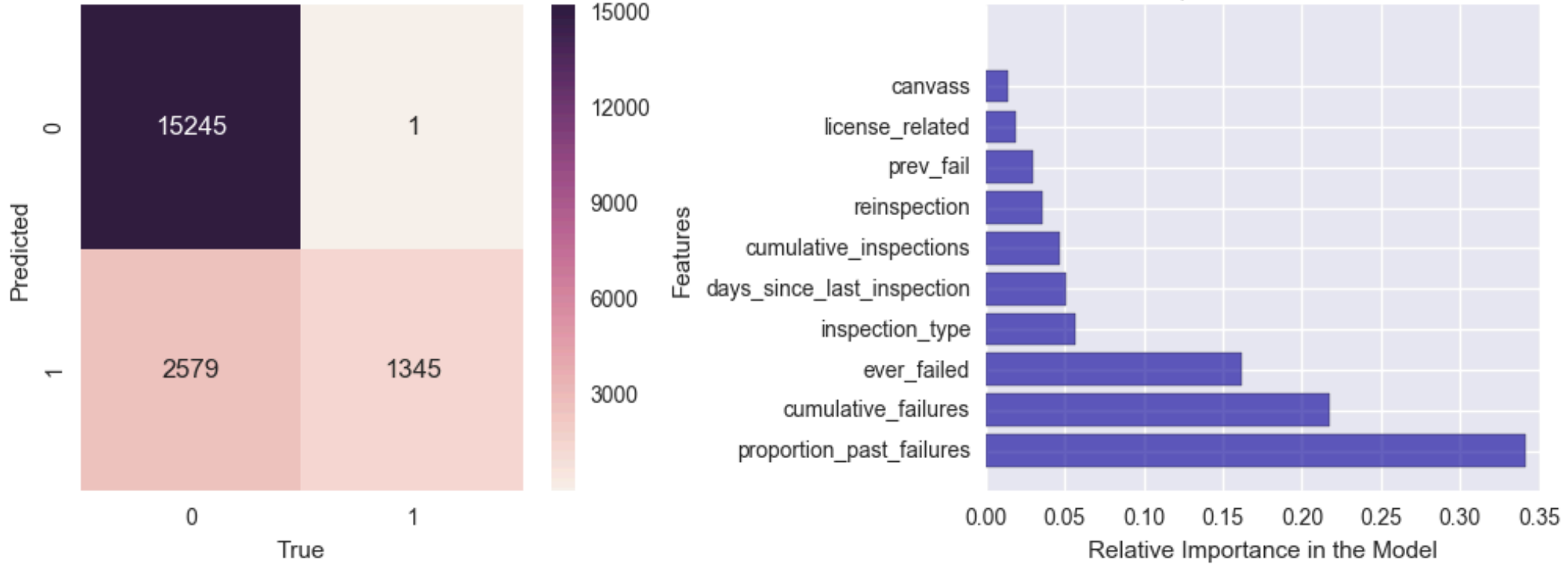
```
dt = DecisionTreeClassifier(labelCol="label", featuresCol="features")
model = dt.fit(train)
predictions = model.transform(test)
reportResults(predictions.select("prediction", "label").rdd, model)
```



github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

Modeling - Random Forest

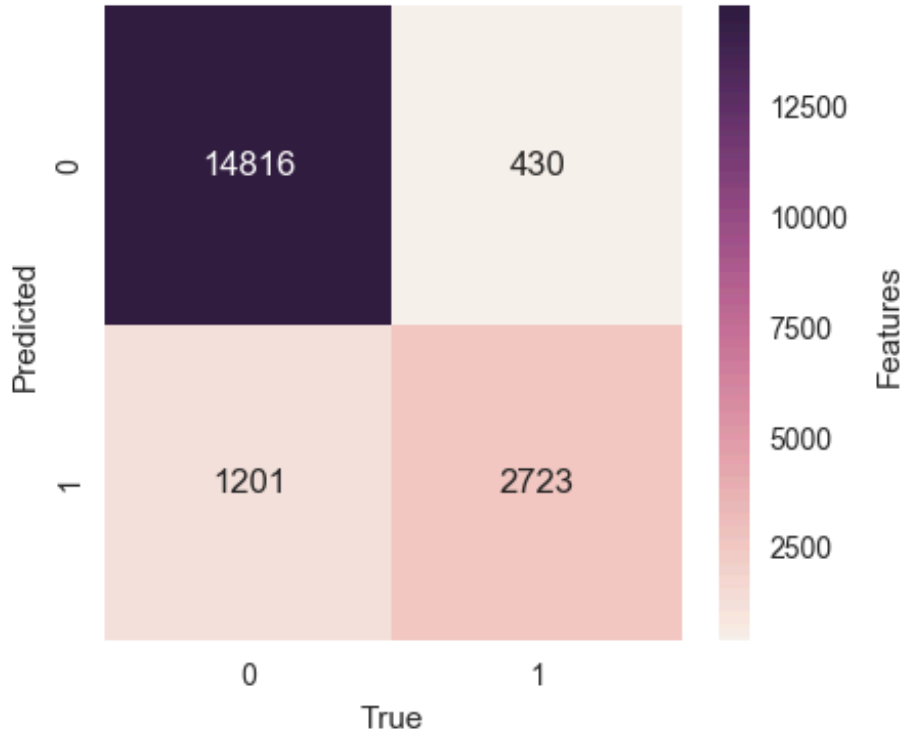
```
rf = RandomForestClassifier(labelCol="label", featuresCol="features", numTrees=500)
model = rf.fit(train)
predictions = model.transform(test)
reportResults(predictions.select("prediction", "label").rdd, model)
```

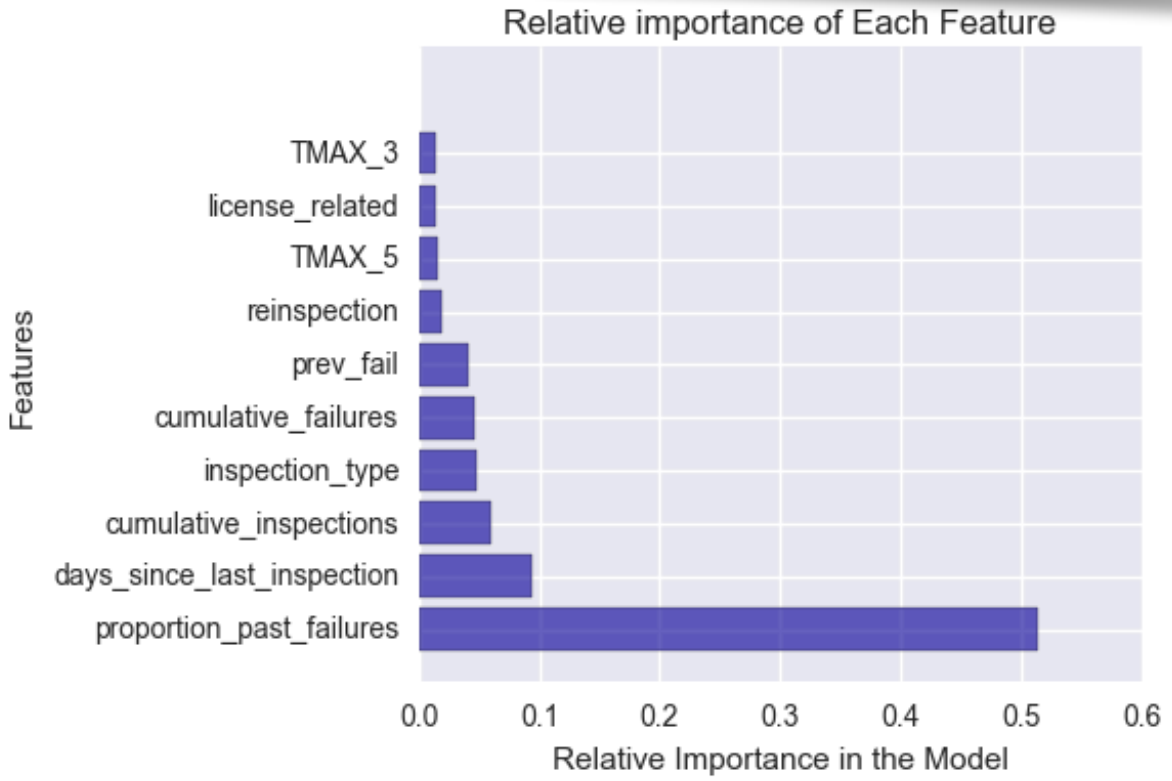


github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

Modeling - Gradient Boosted Trees

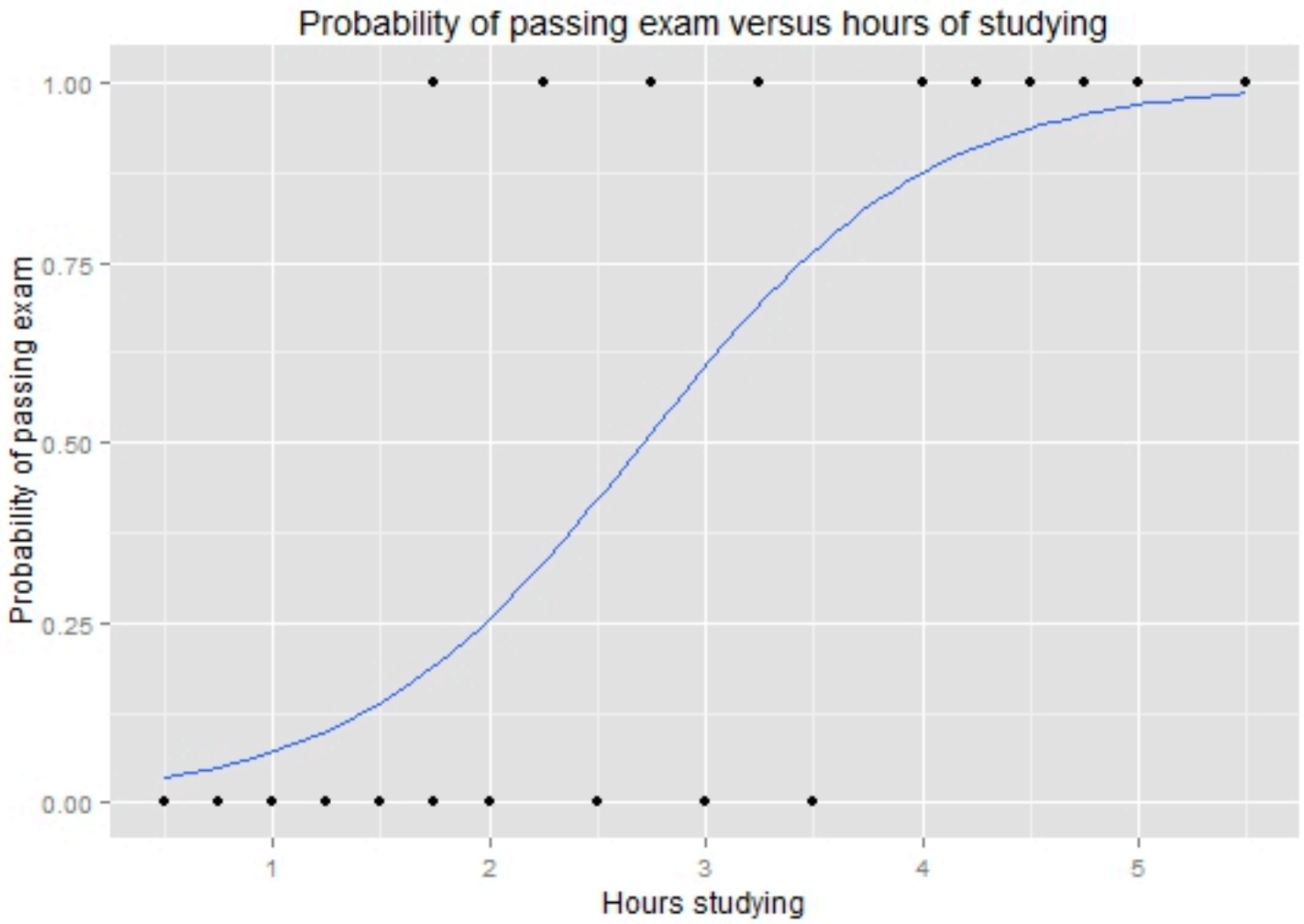
```
gbt = GBTClassifier(labelCol="label", featuresCol="features", maxIter=100)
gbt_model = gbt.fit(train)
predictions = gbt_model.transform(test)
reportResults(predictions.select("prediction", "label").rdd, gbt_model)
```





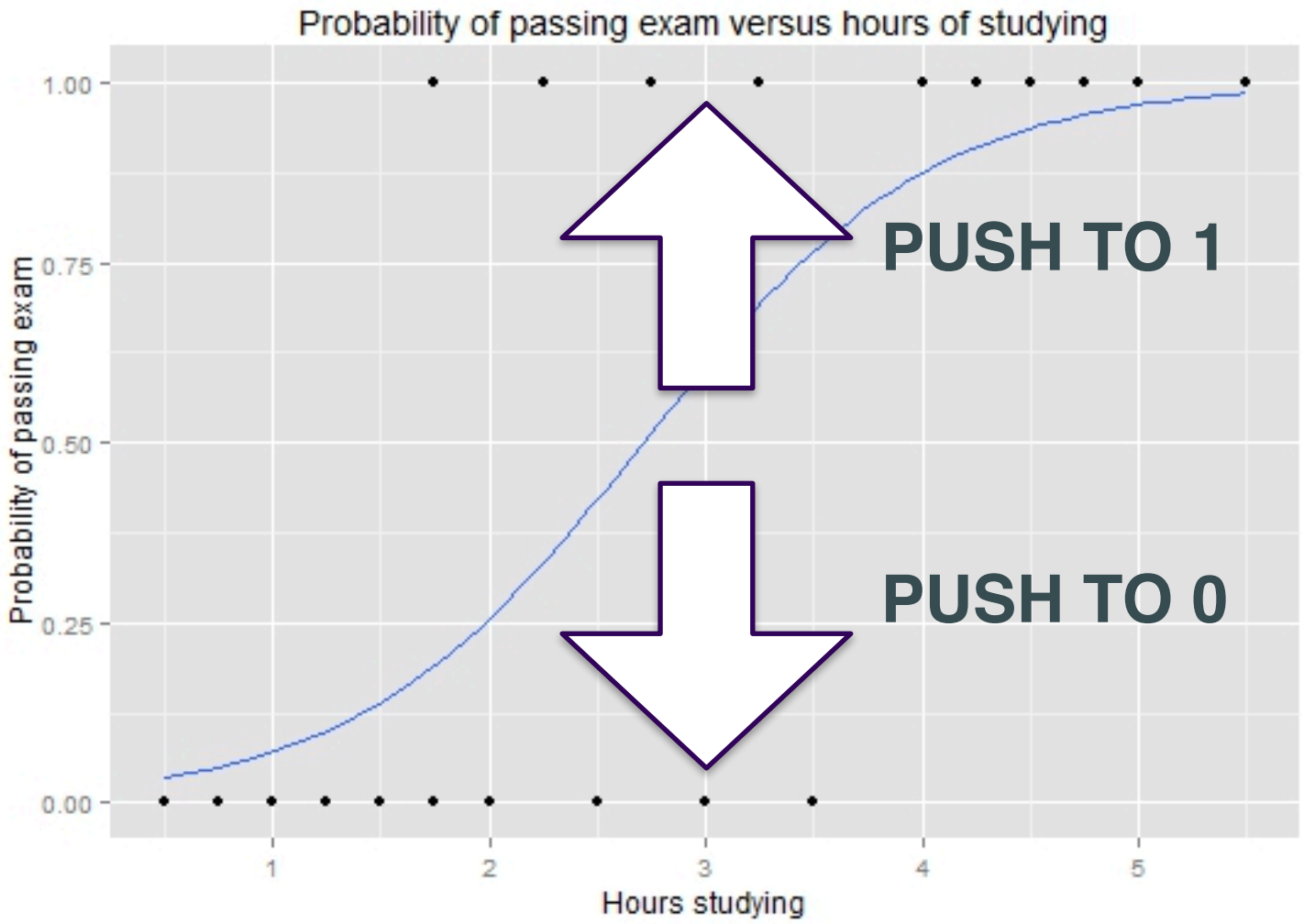
github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

Modeling - Weighted Logistic Regression



github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

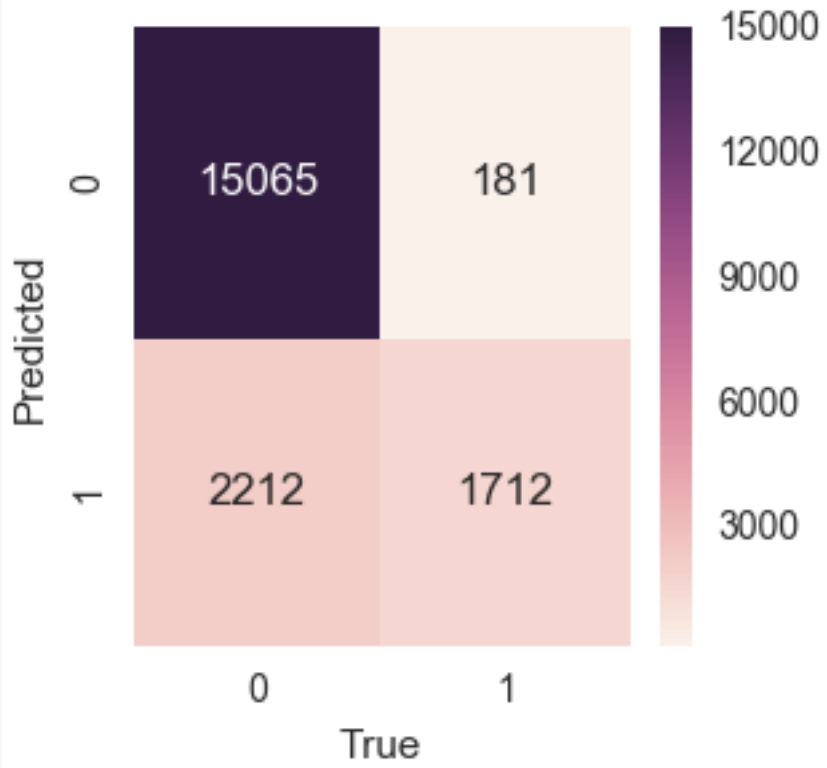
Modeling - Weighted Logistic Regression



github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

Modeling - Weighted Logistic Regression

```
weighted_train = train.withColumn("logistic_weights",
   when(col("label")==0.0, 0.25).otherwise(0.75))

lr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8,
   weightCol="logistic_weights")

model = lr.fit(train)

predictions = model.transform(test)

metrics = MulticlassMetrics(predictions.select("prediction", "label").rdd)

reportAccuracy(metrics)
```



Carrera GROUP INC

github.com/angelok1/chicago-inspections-dse/Step8-Modeling.ipynb

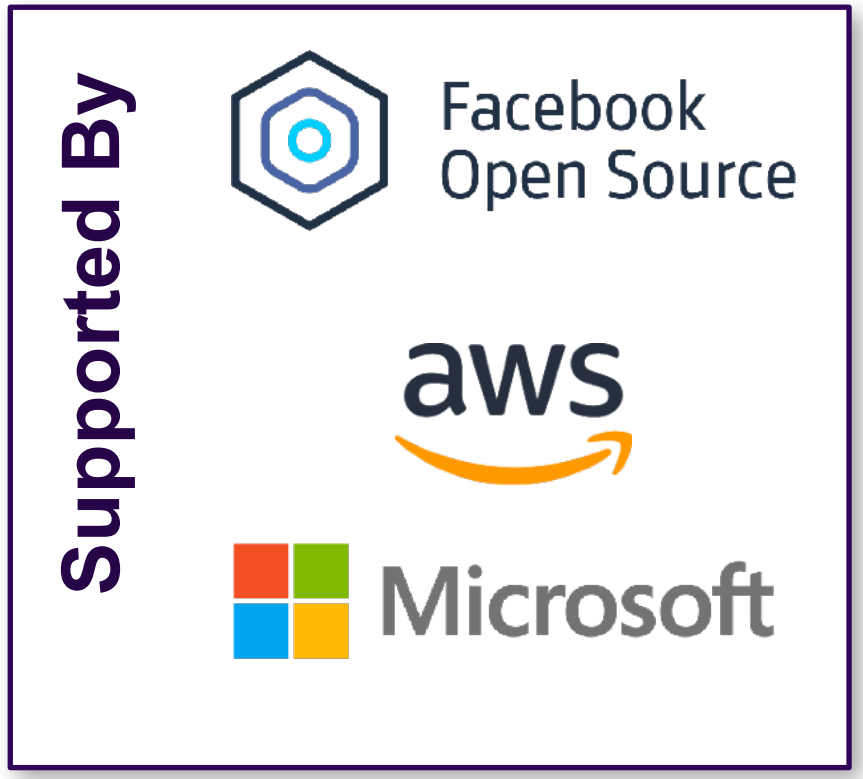
DATASTAX

Production

- How do we get this to production?
- What if there's not parity between tools?
- Where does the data live? DFS? DB?

ONNX

- An open format to represent deep learning models



https://onnx.ai/

ONNX

```
pipeline_model = pipeline.fit(training_data)

onnx_model = convert_sparkml(pipeline_model, 'My Sparkml Pipeline', initial_types)


with open("model.onnx", "wb") as f:

    f.write(onnx_model.SerializeToString())
```

SparkML

ONNX Framework

```
onnx_model = onnx.load('path/to/the/model.onnx')


sess = onnxruntime.InferenceSession(onnx_model)

output = sess.run(None, input_data)
```

Carrera GROUP INC

https://onnx.ai/

DATASTAX®

Koalas

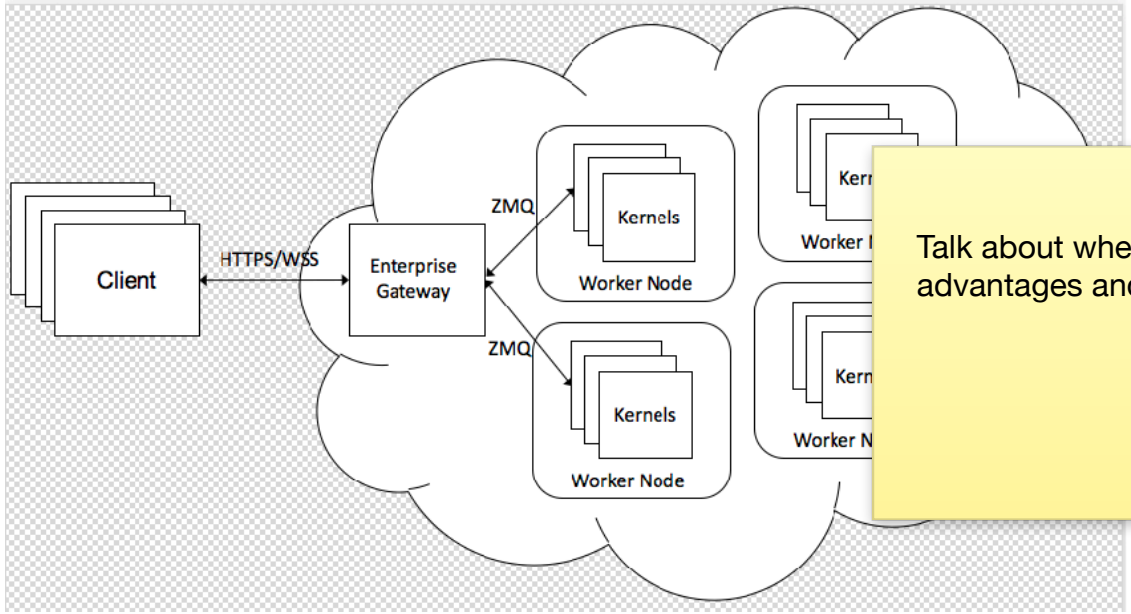
```
import pandas as pd
import databricks.koalas as ks

pdf = pd.DataFrame({'x':range(3), 'y':['a','b','b'],
'z':['a','b','b']})


# Create a Koalas DataFrame from pandas DataFrame
df = ks.from_pandas(pdf)

# Rename the columns
df.columns = ['x', 'y', 'z1']

# Do some operations in place:
df['x2'] = df.x * df.x
```

- Parity with pandas!
- It's New (VERY new)
- Support Spark 2.4, not tested on older versions

https://github.com/databricks/koalas

Jupyter with Spark

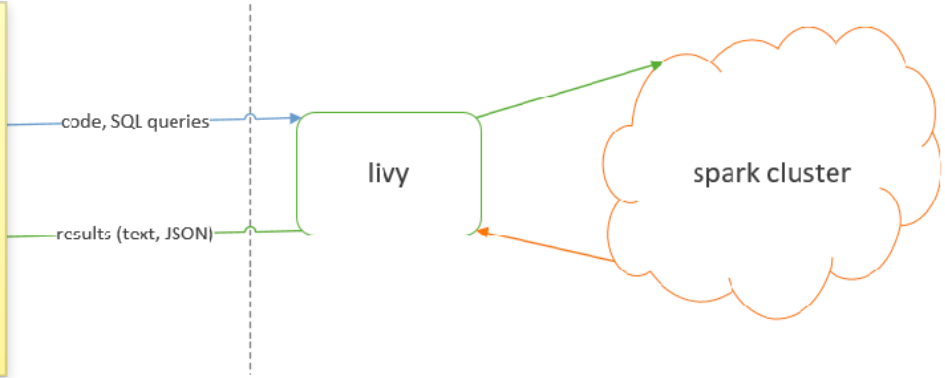
Jupyter Enterprise Gateway



Talk about where the kernel runs, advantages and disadvantages

https://github.com/jupyter/enterprise_gateway

Jupyter, Sparkmagic and Apache Livy



https://github.com/jupyter-incubator/sparkmagic

The Tools

- JEG Backed by Jupyter and containerizes the notebooks, too. It's also well supported by IBM.
- Livy backed by Apache and Sparkmagic is incubated by Jupyter. It's also a bit older.

	JEG	SparkMagic/ Livy
Works With DSE	Yes*	Yes*
Vanilla Spark	Yes*	Yes*
Kubernetes	Yes	No
Security	Yes*	Yes*

* Installation needs a bit of work to make seamless

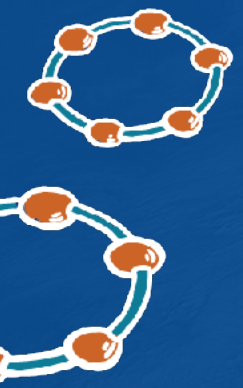
Final Thoughts

- Challenge your intuition
- The models we make do not live in a vacuum
- Don’t make everyone use the same tools for everything. The platform should be flexible

DATASTAX

ACCELERATE

THANK YOU



https://github.com/angelok1/chicago-inspections-dse/
https://www.linkedin.com/in/angelok1/
akastroulis@carrera.io