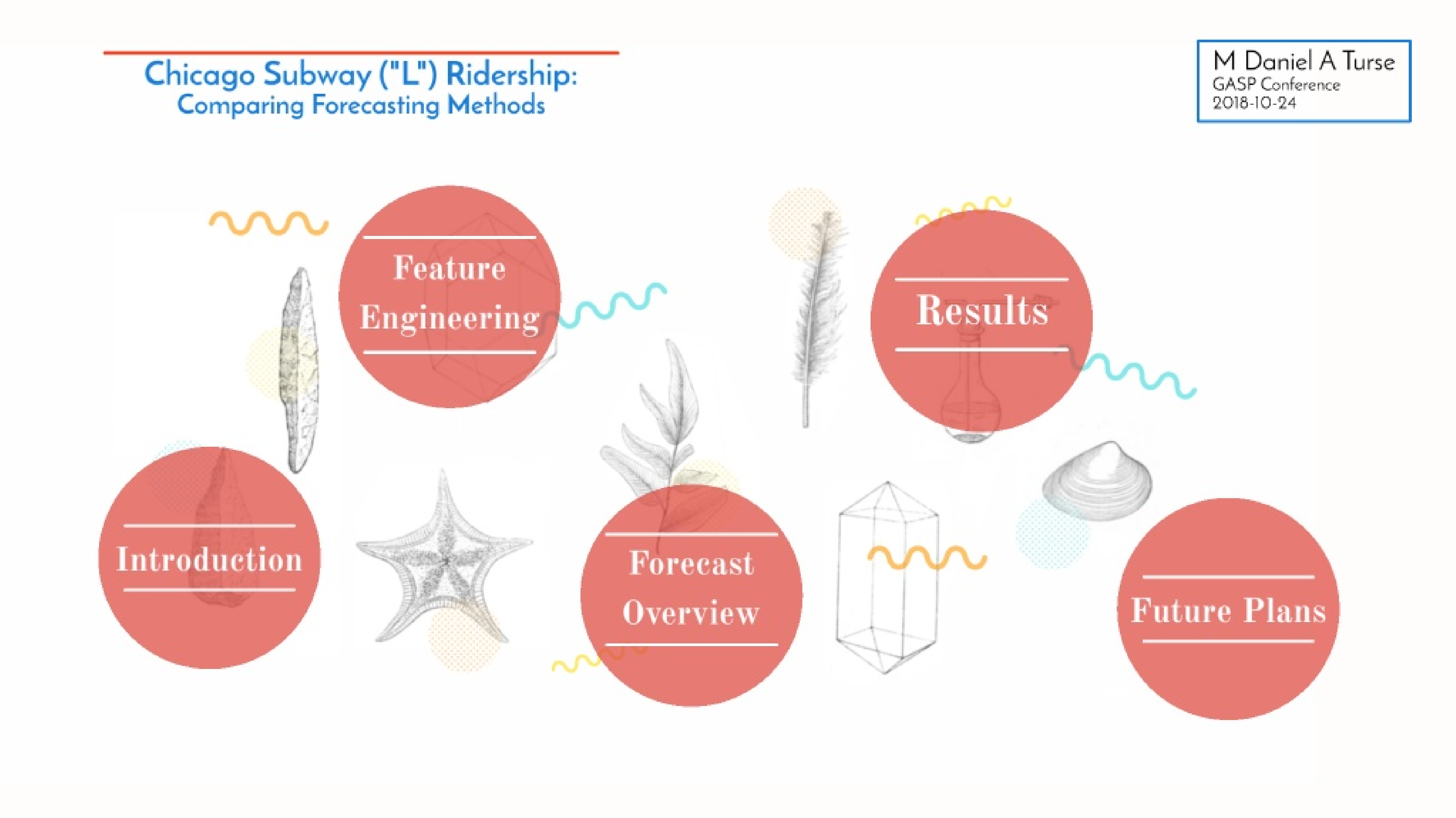


Chicago Subway ("L") Ridership:
Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

Introduction

Feature Engineering

Forecast Overview

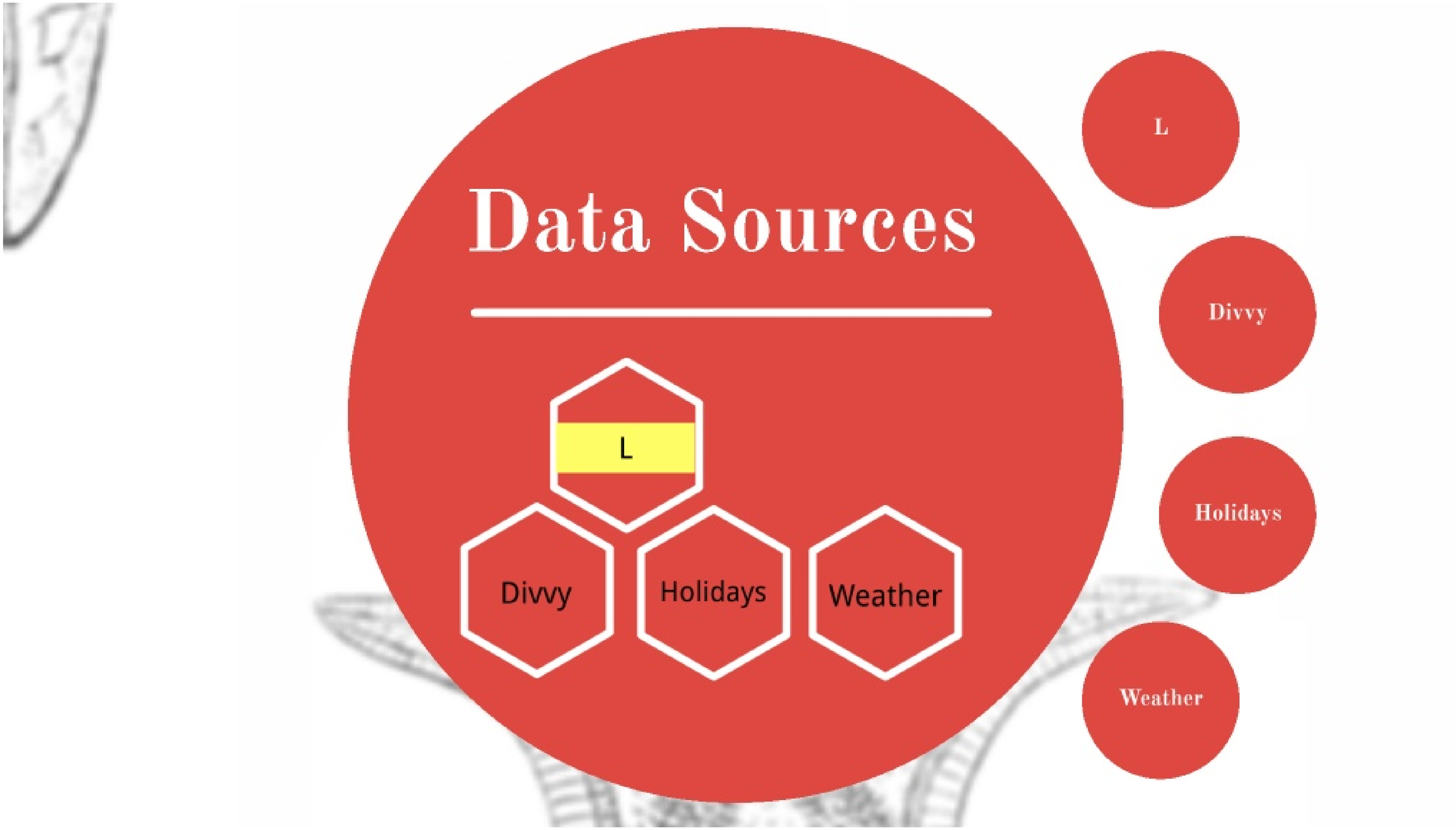
Results

Future Plans

Introduction

Data Sources

Objectives



Data Sources

L

Divvy

Holidays

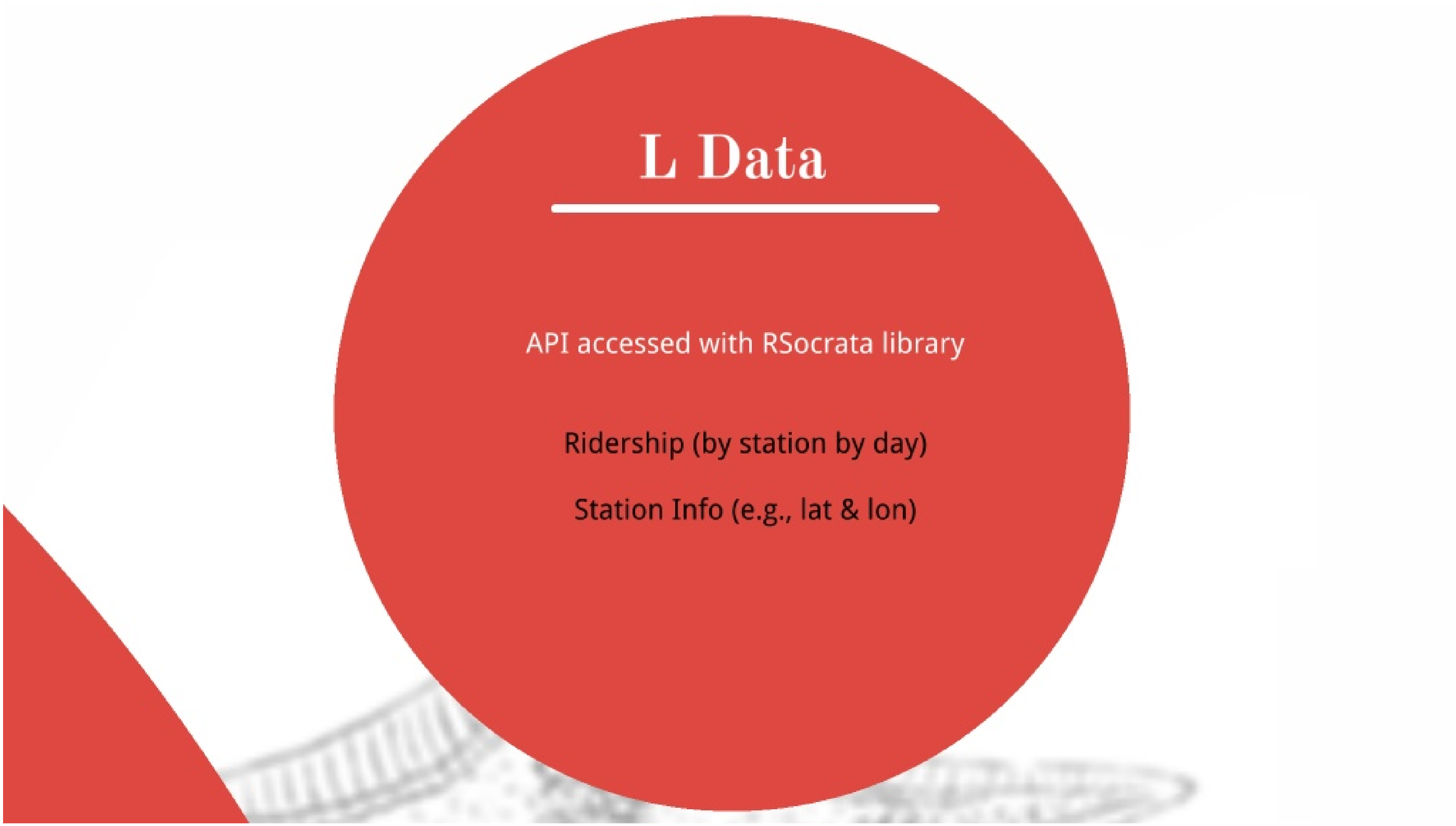
Weather

L

Divvy

Holidays

Weather



L Data

API accessed with RSocrata library

Ridership (by station by day)

Station Info (e.g., lat & lon)

Divvy Data

Programmatically downloaded .csv & .xlsx files from the Divvy website (https://www.divvybikes.com/system-data)

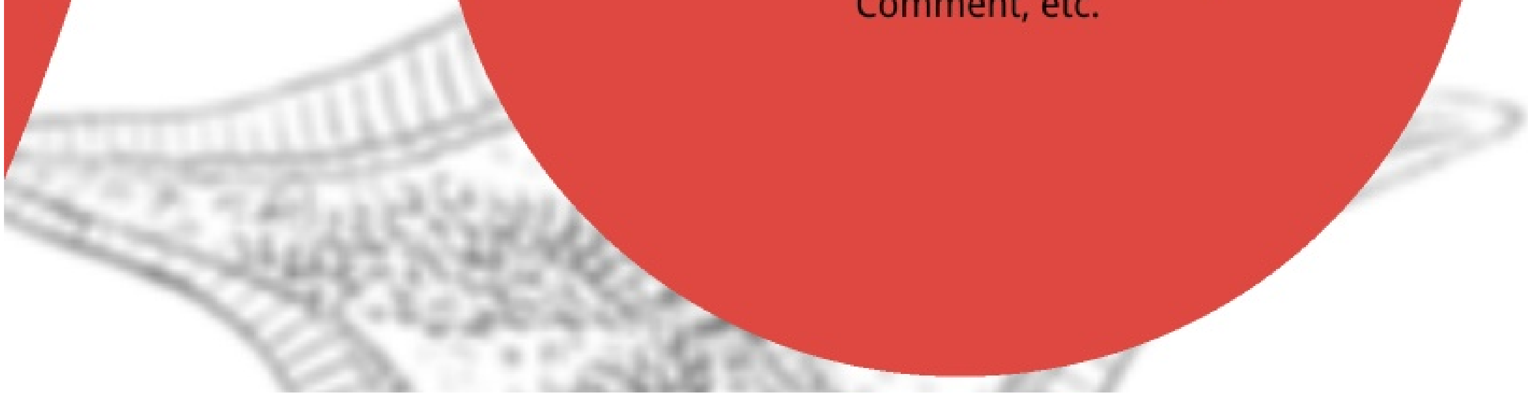
Trip Details (start location, stop location, datetime start, datetime stop, user type, etc)

Station Info (e.g., lat & lon, station "in use" date, etc.)

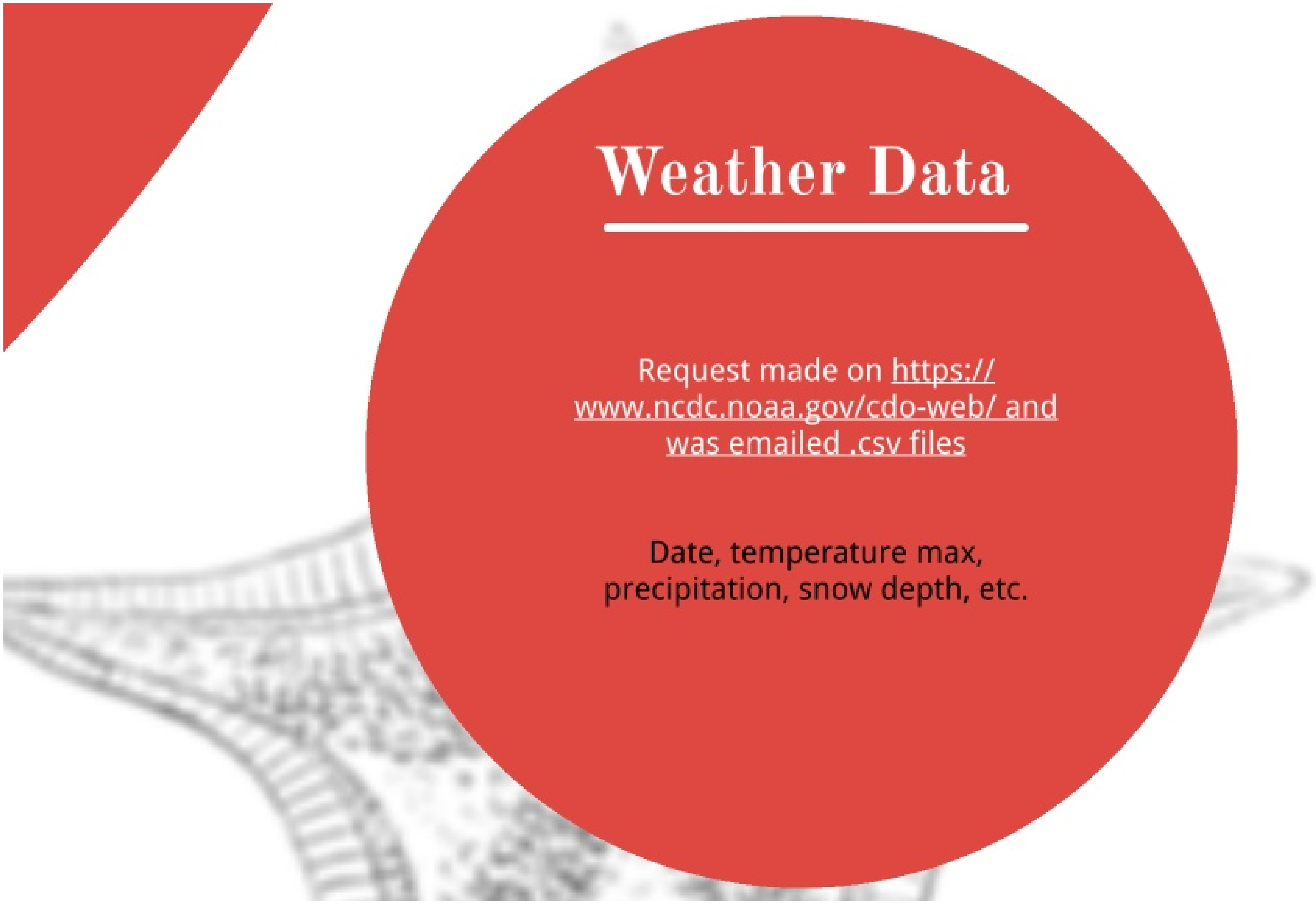
Holiday Data

Data scraped from https://www.officeholidays.com with rvest library

Holiday Date, Holiday Name, Comment, etc.



Weather Data



Request made on https://www.ncdc.noaa.gov/cdo-web/ and was emailed .csv files

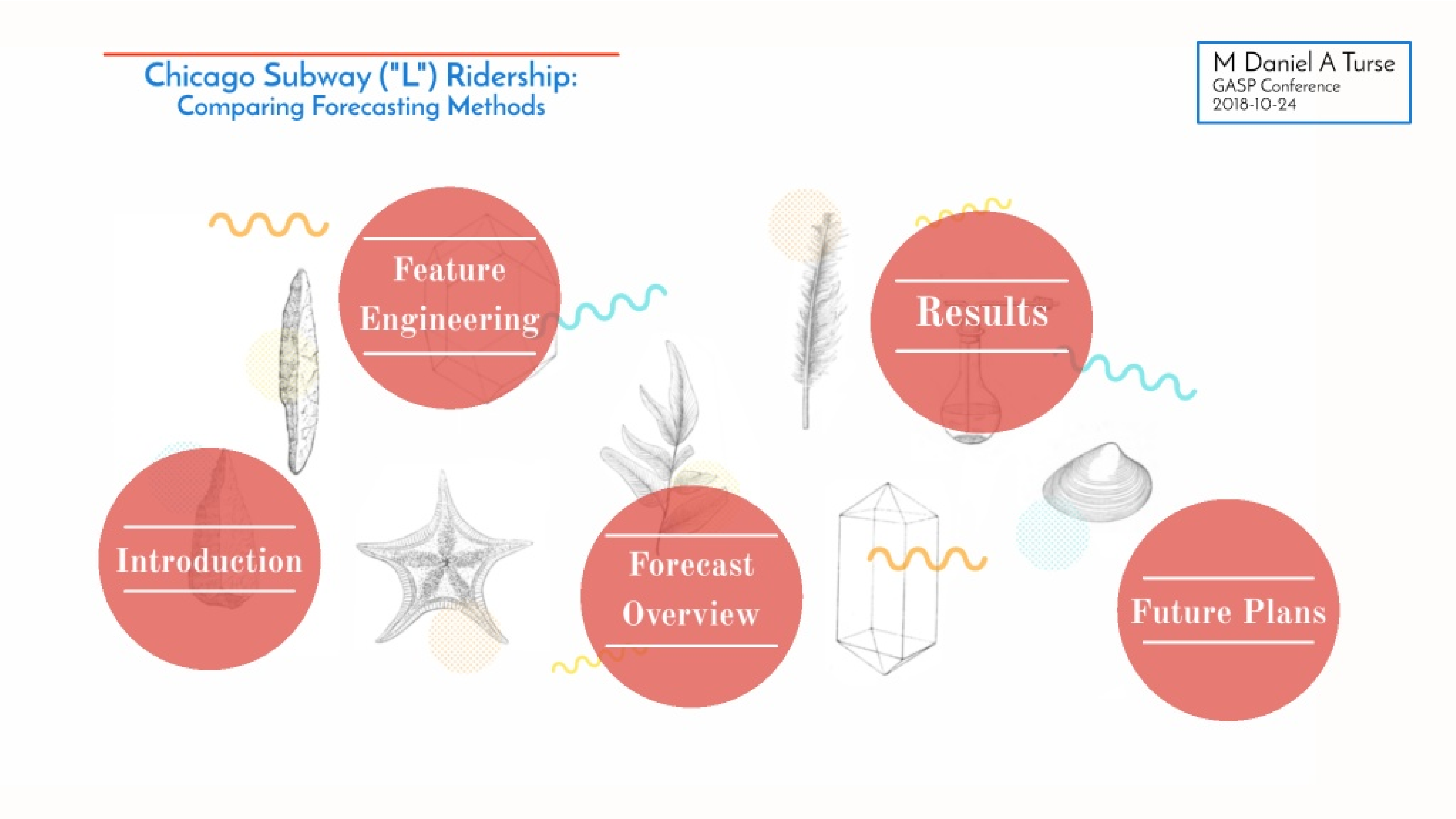
Date, temperature max, precipitation, snow depth, etc.

Objectives

1 Week Ahead L Forecasts
(by station by day)

Compare Forecasts By Algorithm

Explore Variable Effects



Chicago Subway ("L") Ridership:
Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

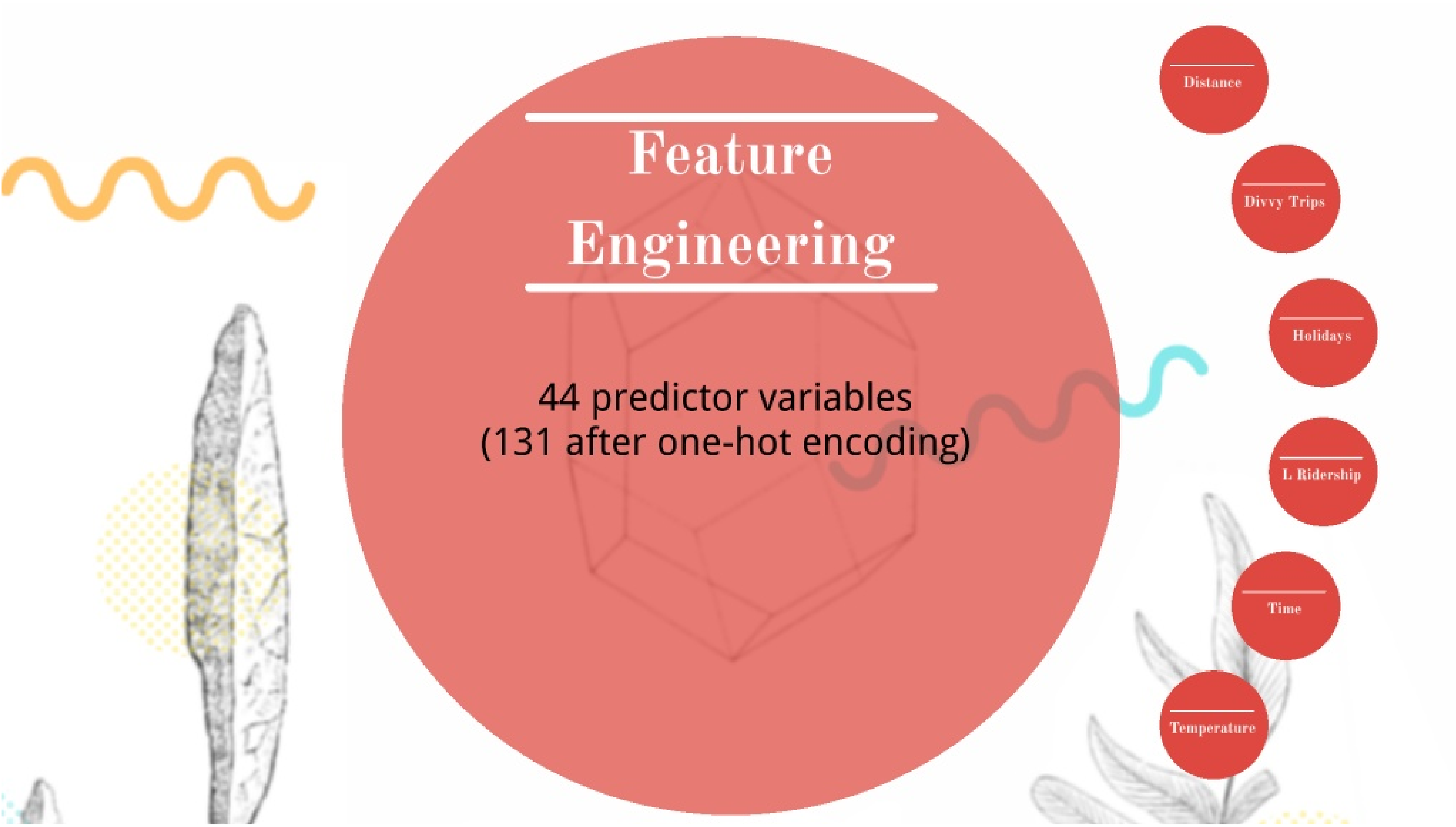
Introduction

Feature Engineering

Forecast Overview

Results

Future Plans



Feature Engineering

44 predictor variables
(131 after one-hot encoding)

Distance

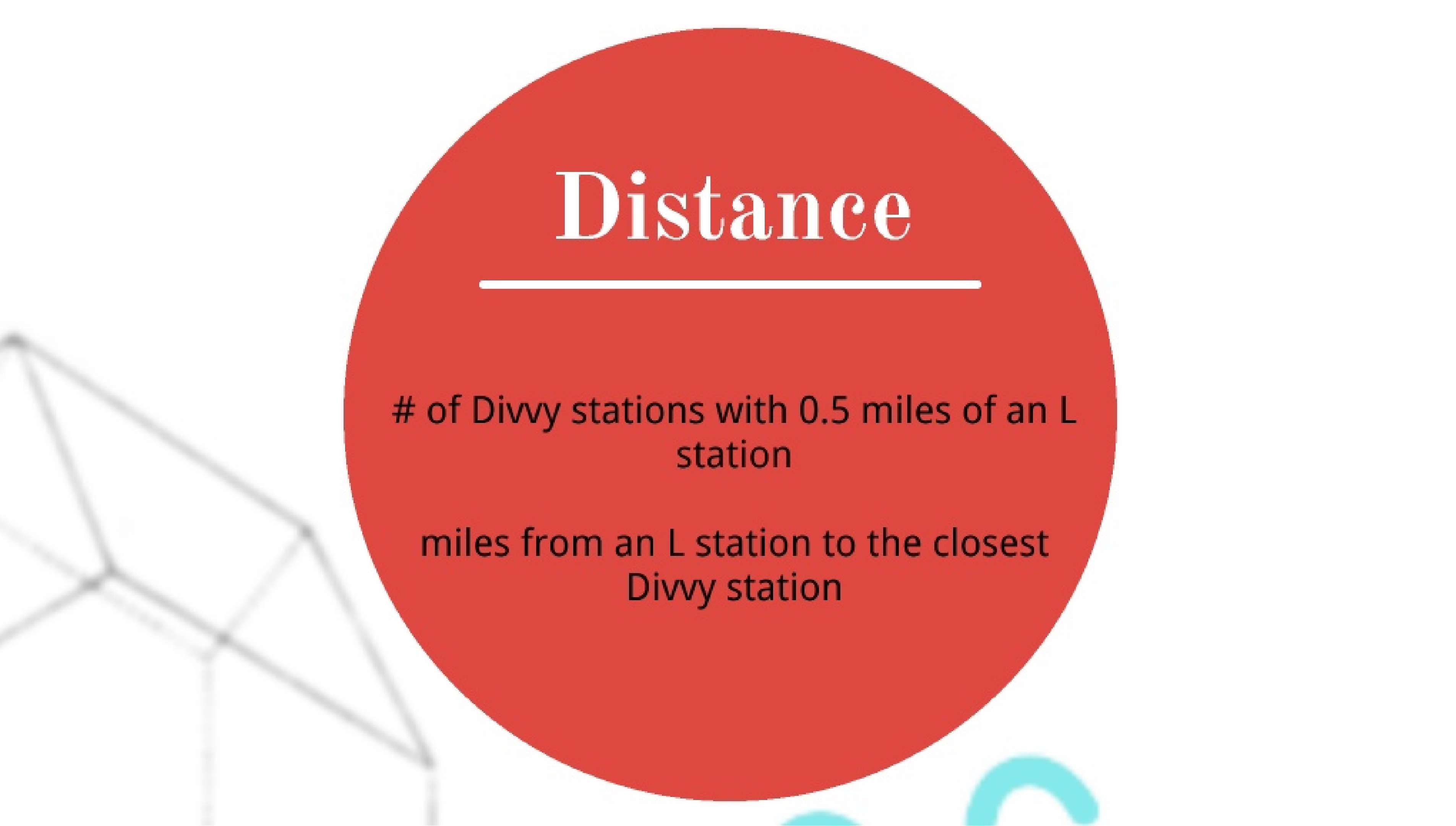
Divvy Trips

Holidays

L Ridership

Time

Temperature



Distance

# of Divvy stations with 0.5 miles of an L station

miles from an L station to the closest Divvy station

Divvy Trips

(done daily for each of the three types of customers)

trip counts

trip time stats (e.g., mean, median)

Holidays

holiday name and date

L Ridership

lags of ridership

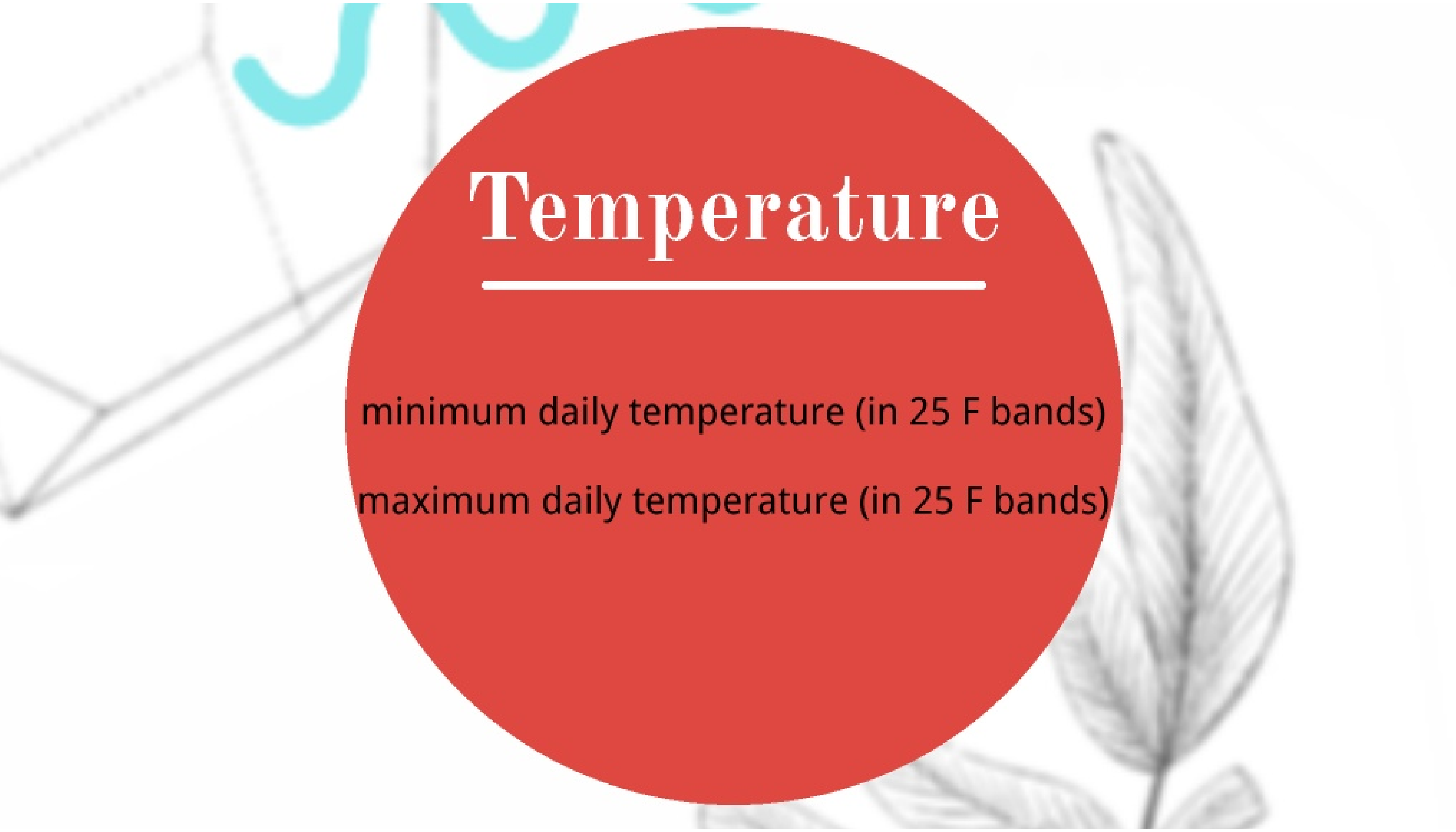
moving averages of ridership

Time

day of the week

month

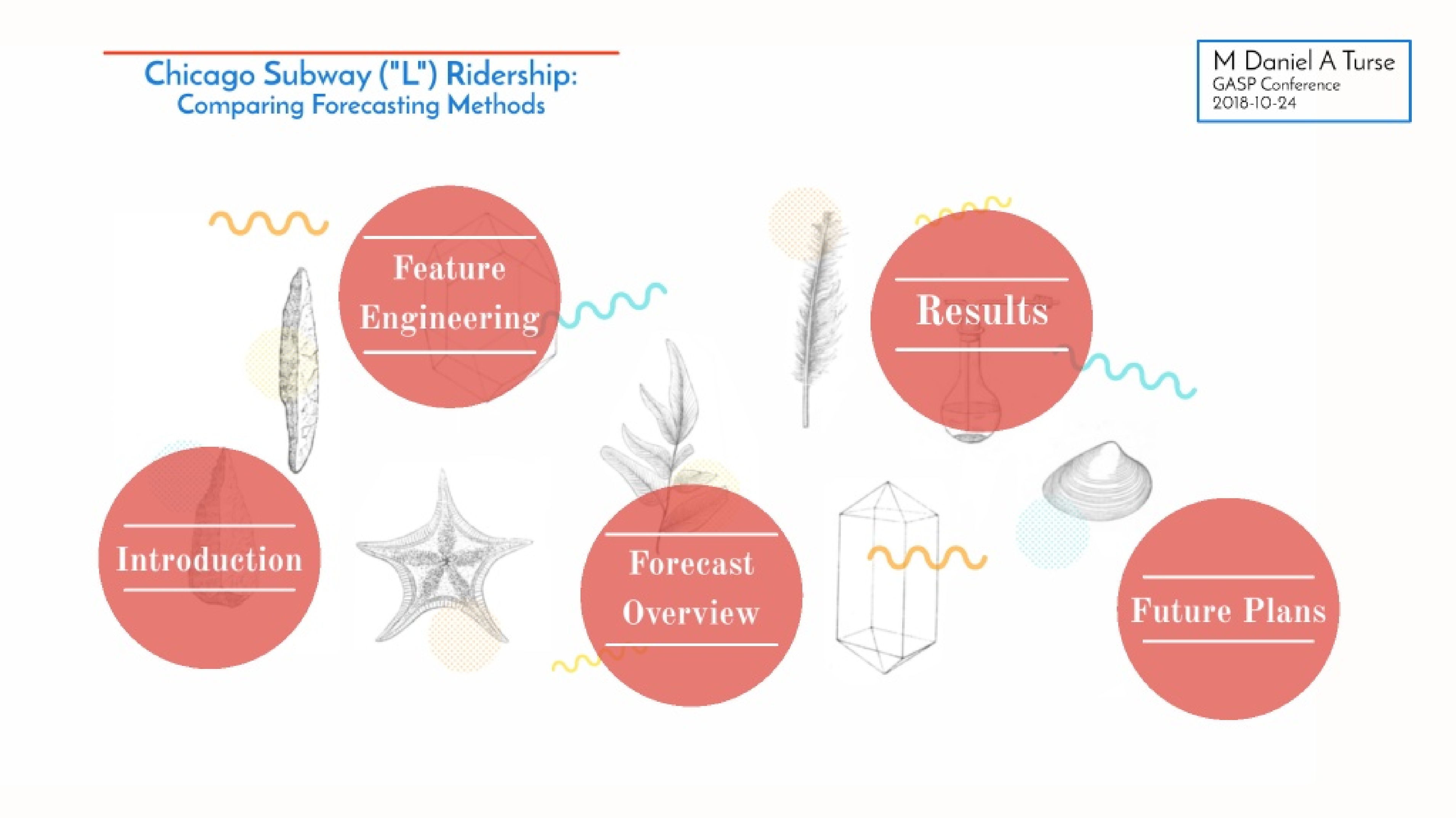
week of the month



Temperature

minimum daily temperature (in 25 F bands)

maximum daily temperature (in 25 F bands)



Chicago Subway ("L") Ridership:
Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

Introduction

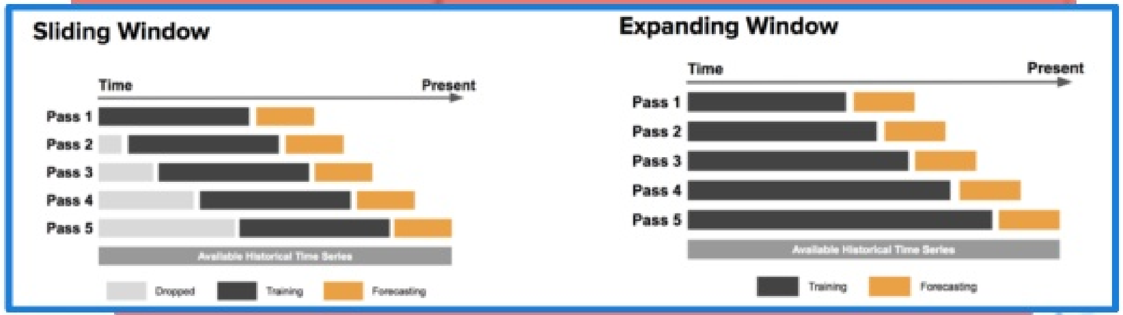
Feature Engineering

Forecast Overview

Results

Future Plans

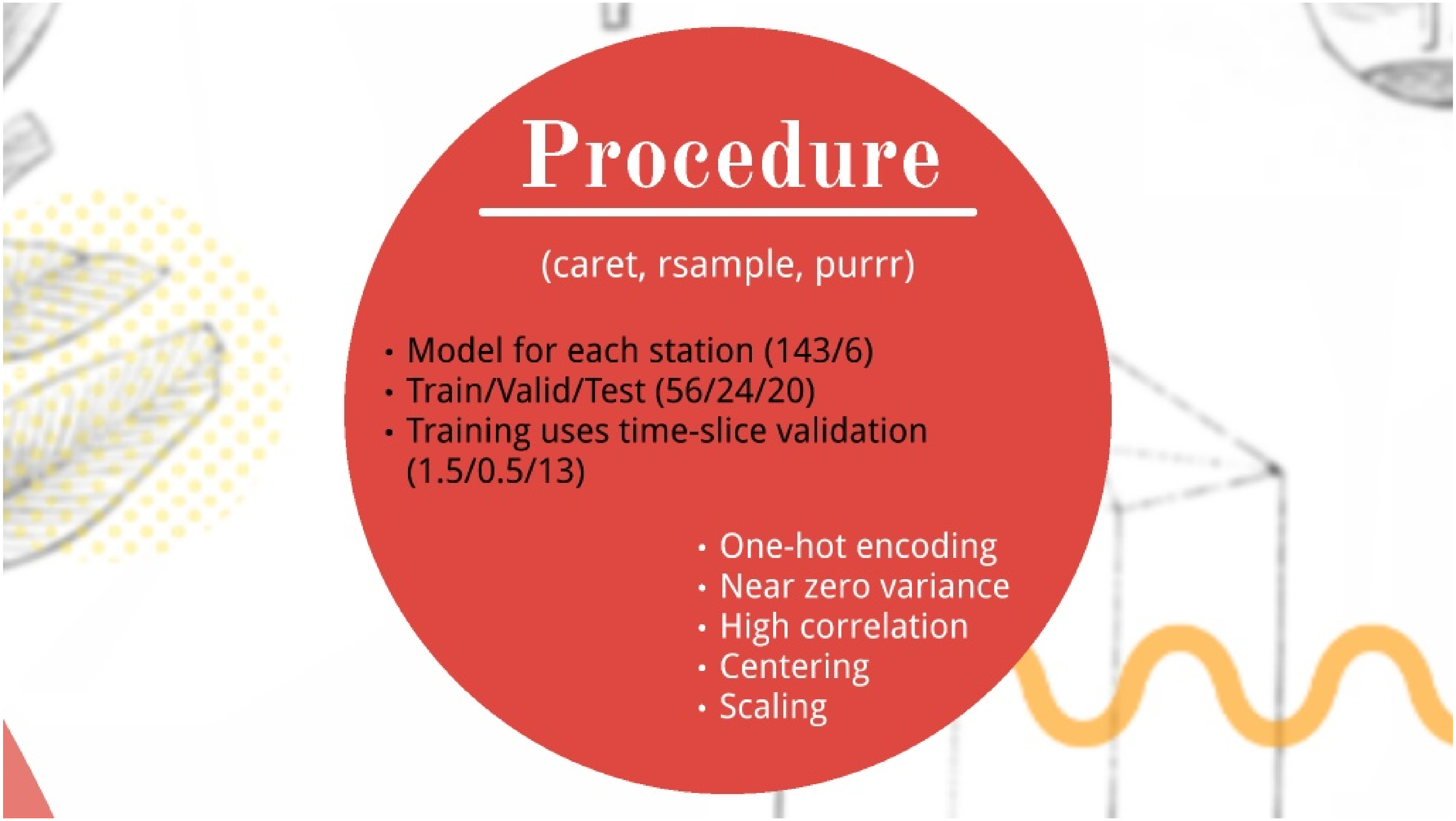
Forecast Overview



Procedure

Models

Comparison Criteria



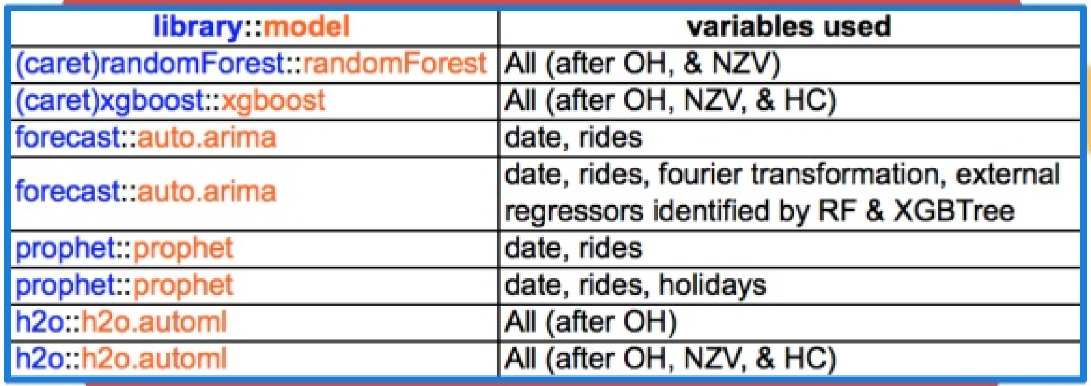
Procedure

(caret, rsample, purrr)

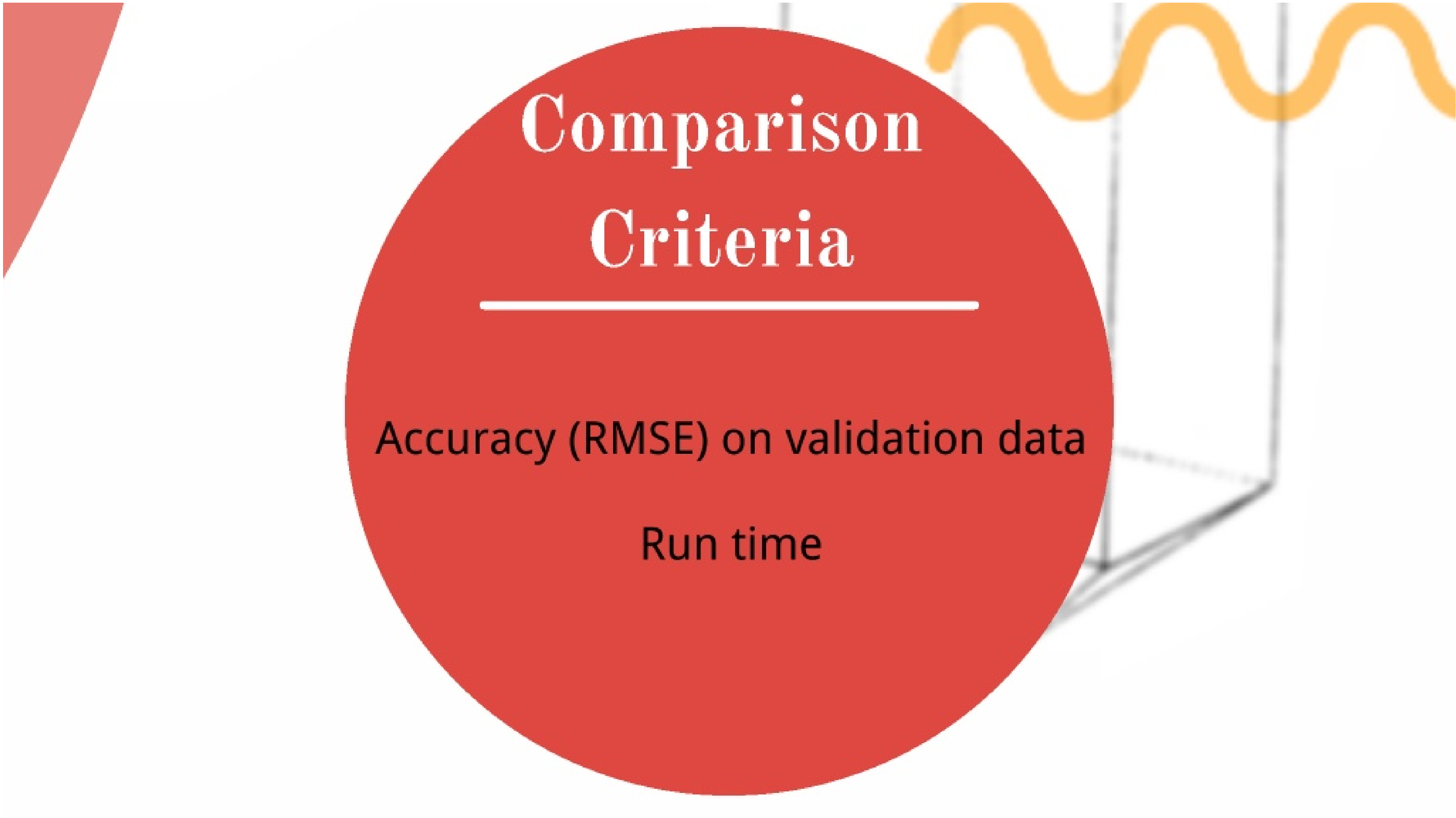
- Model for each station (143/6)
- Train/Valid/Test (56/24/20)
- Training uses time-slice validation (1.5/0.5/13)

- One-hot encoding
- Near zero variance
- High correlation
- Centering
- Scaling

Models



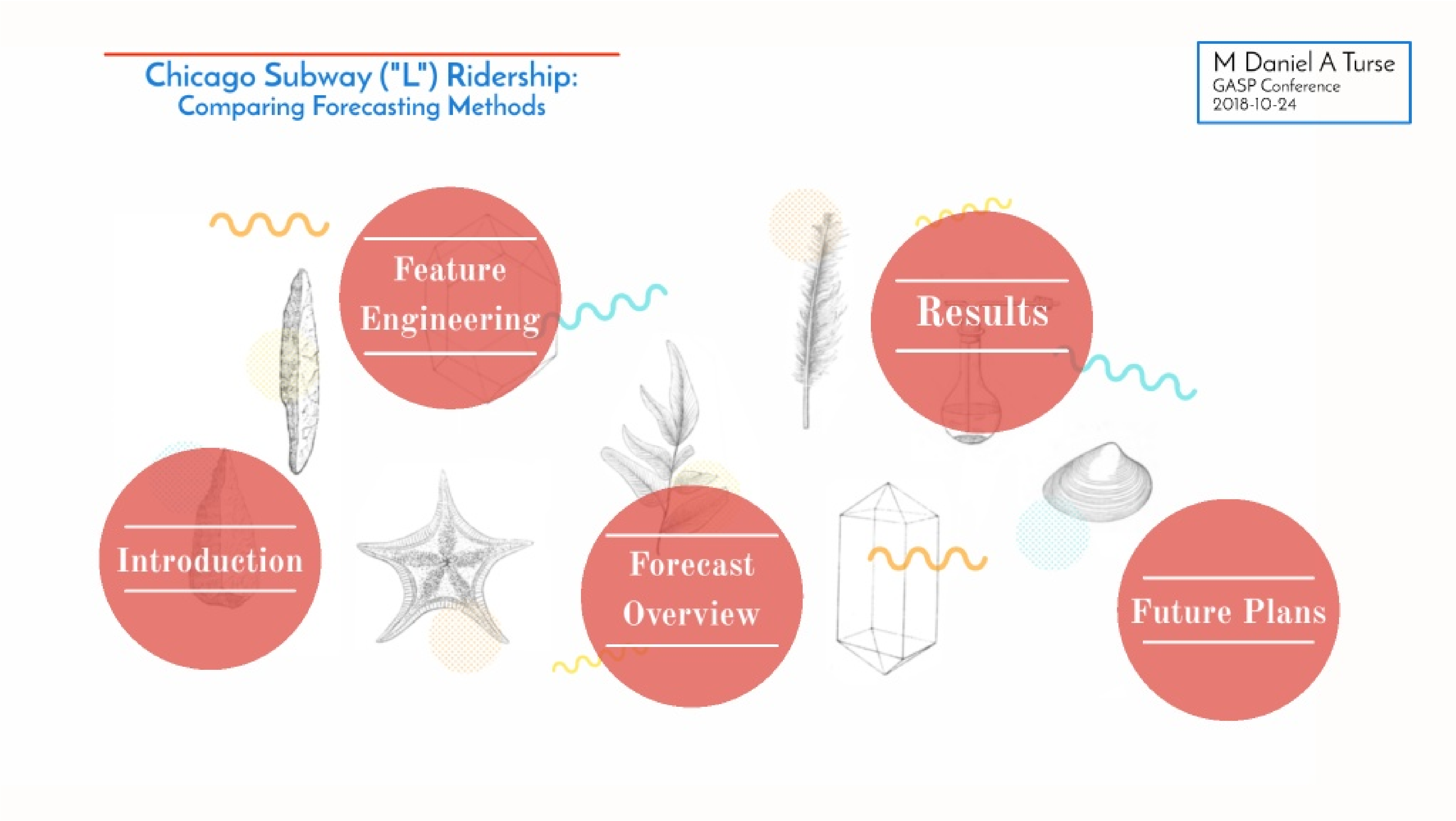
library::model	variables used
(caret)randomForest::randomForest	All (after OH, & NZV)
(caret)xgboost::xgboost	All (after OH, NZV, & HC)
forecast::auto.arima	date, rides
forecast::auto.arima	date, rides, fourier transformation, external regressors identified by RF & XGBTree
prophet::prophet	date, rides
prophet::prophet	date, rides, holidays
h2o::h2o.automl	All (after OH)
h2o::h2o.automl	All (after OH, NZV, & HC)



Comparison Criteria

Accuracy (RMSE) on validation data

Run time



Chicago Subway ("L") Ridership: Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

Introduction

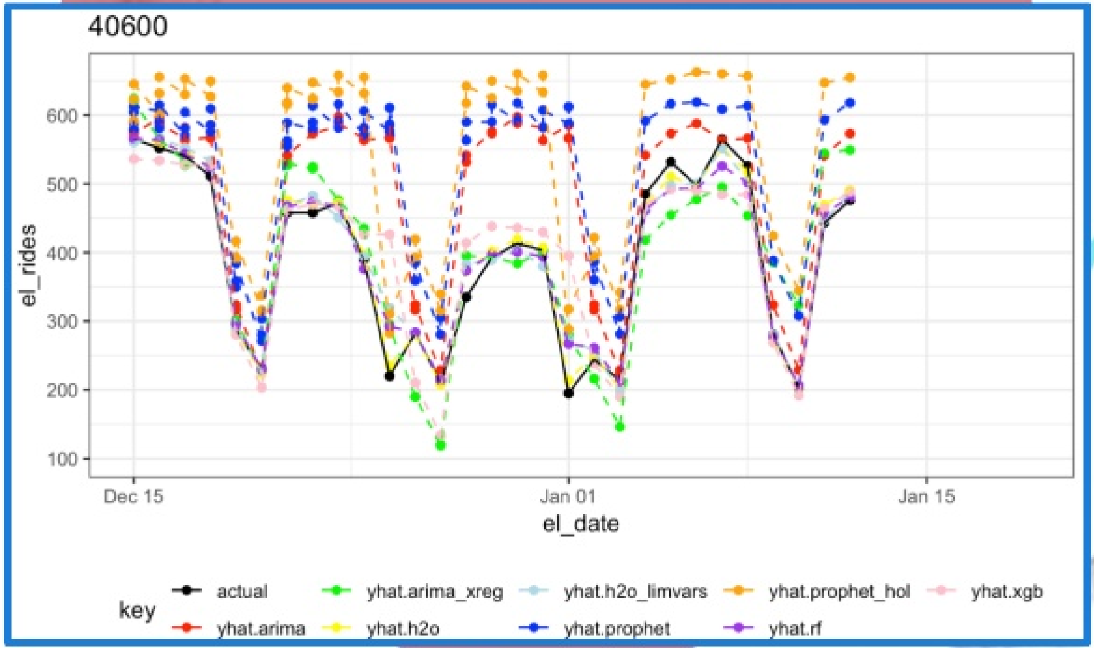
Feature Engineering

Forecast Overview

Results

Future Plans

Results



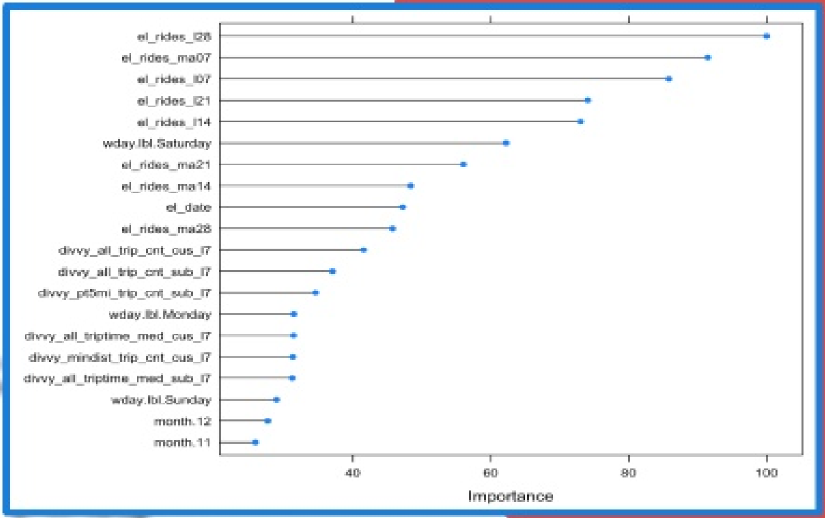
Variable Importance

Extrap vs. Interp

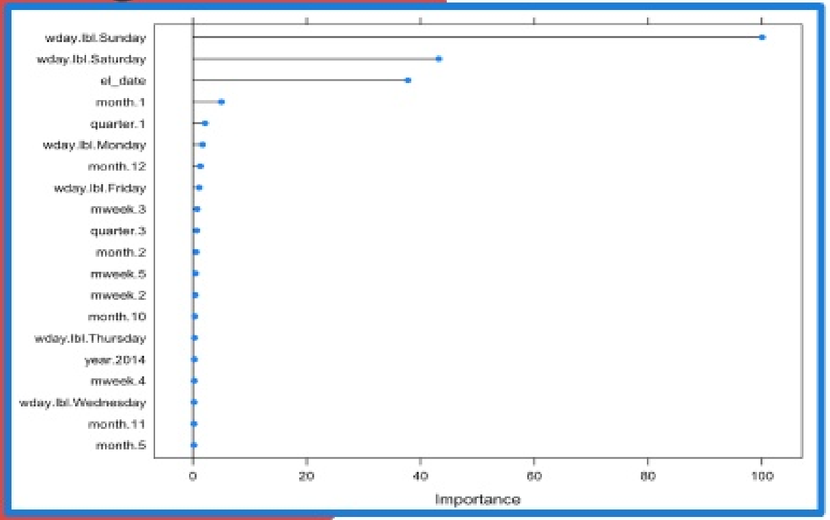
Champion?

Variable Importance

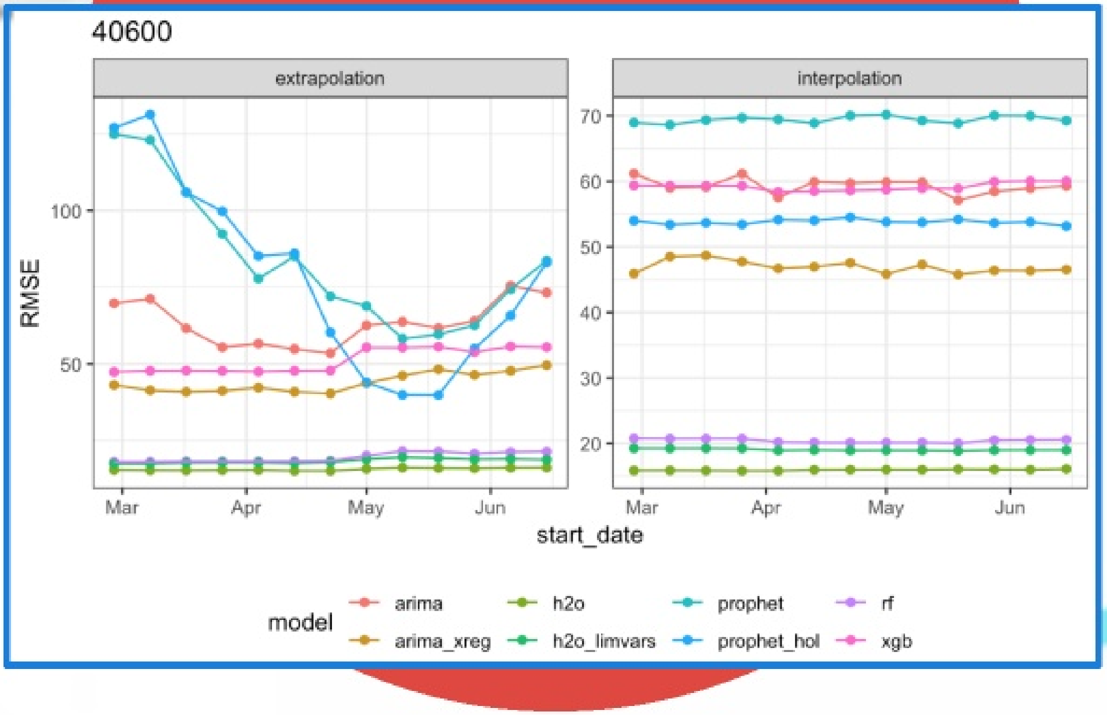
RF



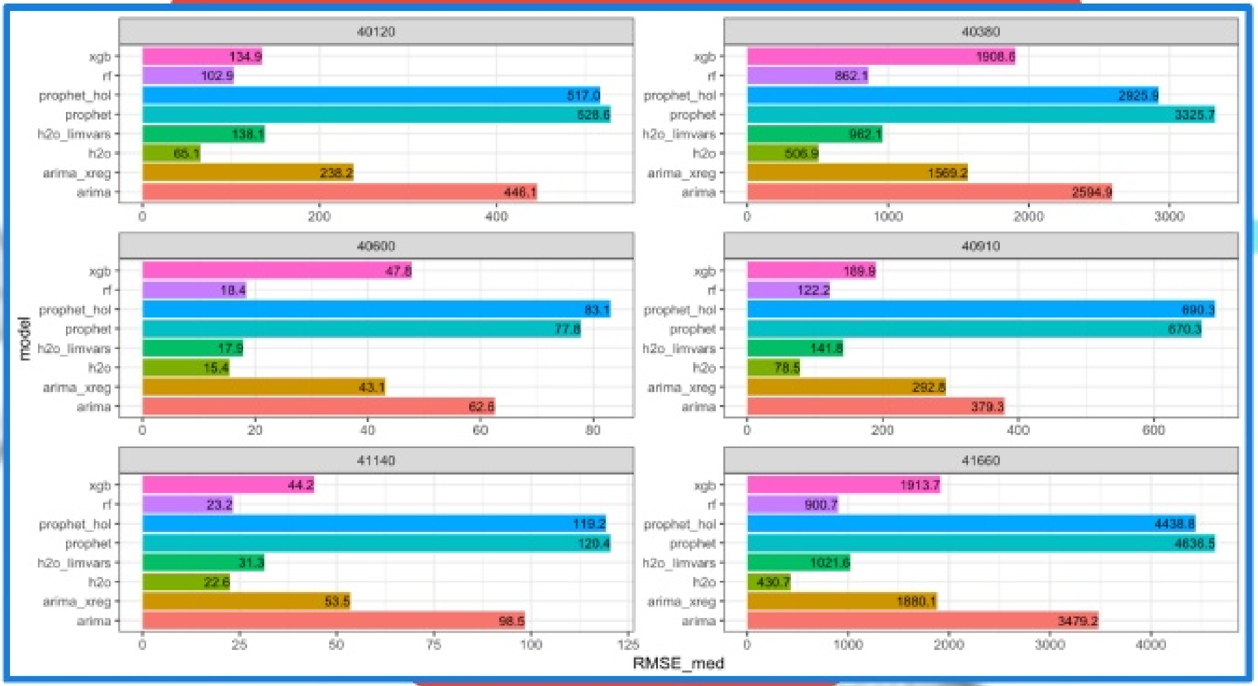
XgbTree

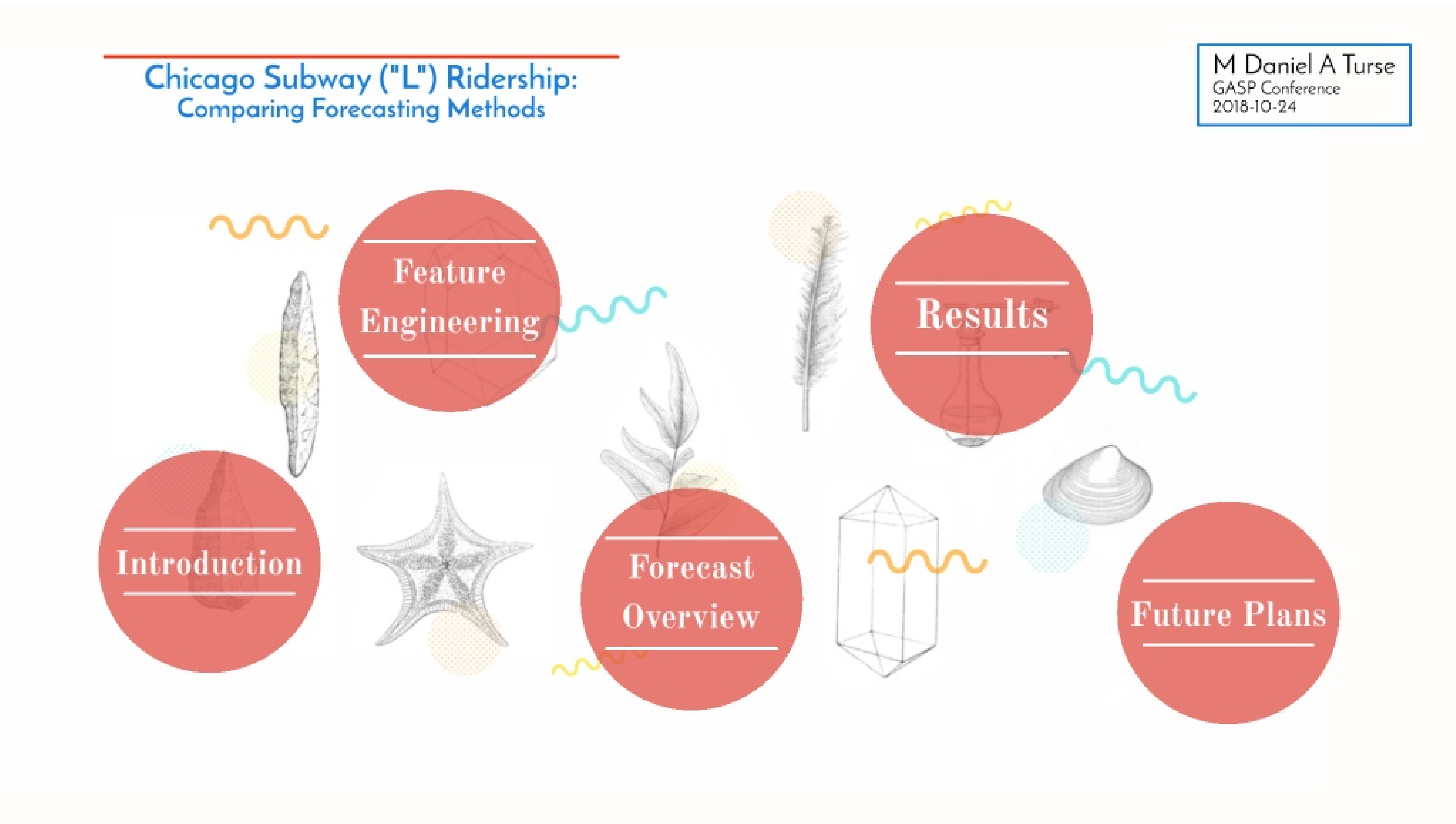


Extrap vs. Interp



Champion?


40120
xgb
rf
prophet_hol
prophet
h2o_limvars
h2o
arima_xreg
arima
134.9
102.9
517.0
528.6
138.1
65.1
238.2
446.1
0
200
400
40380
1908.6
862.1
2925.9
3325.7
962.1
506.9
1569.2
2594.9
0
1000
2000
3000
40600
47.8
18.4
83.1
77.8
17.9
15.4
43.1
62.6
0
20
40
60
80
40910
189.9
122.2
690.3
670.3
141.8
78.5
292.8
379.3
0
200
400
600
41140
44.2
23.2
119.2
120.4
31.3
22.6
53.5
98.5
0
25
50
75
100
125
41660
1913.7
900.7
4438.8
4636.5
1021.6
430.7
1880.1
3479.2
0
1000
2000
3000
4000
model
RMSE_med



Chicago Subway ("L") Ridership:
Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

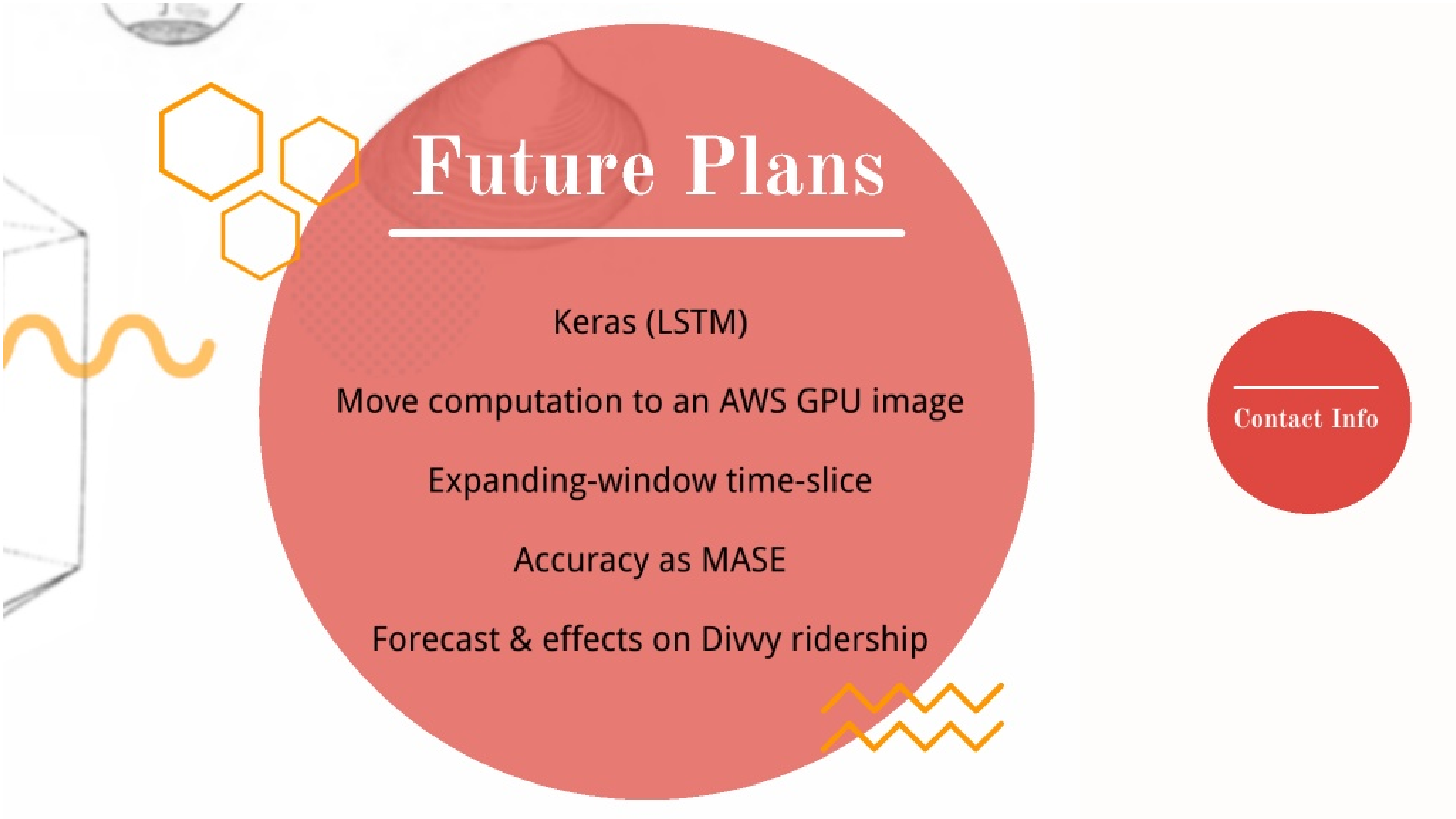
Introduction

Feature Engineering

Forecast Overview

Results

Future Plans



Future Plans

Keras (LSTM)

Move computation to an AWS GPU image

Expanding-window time-slice

Accuracy as MASE

Forecast & effects on Divvy ridership

Contact Info

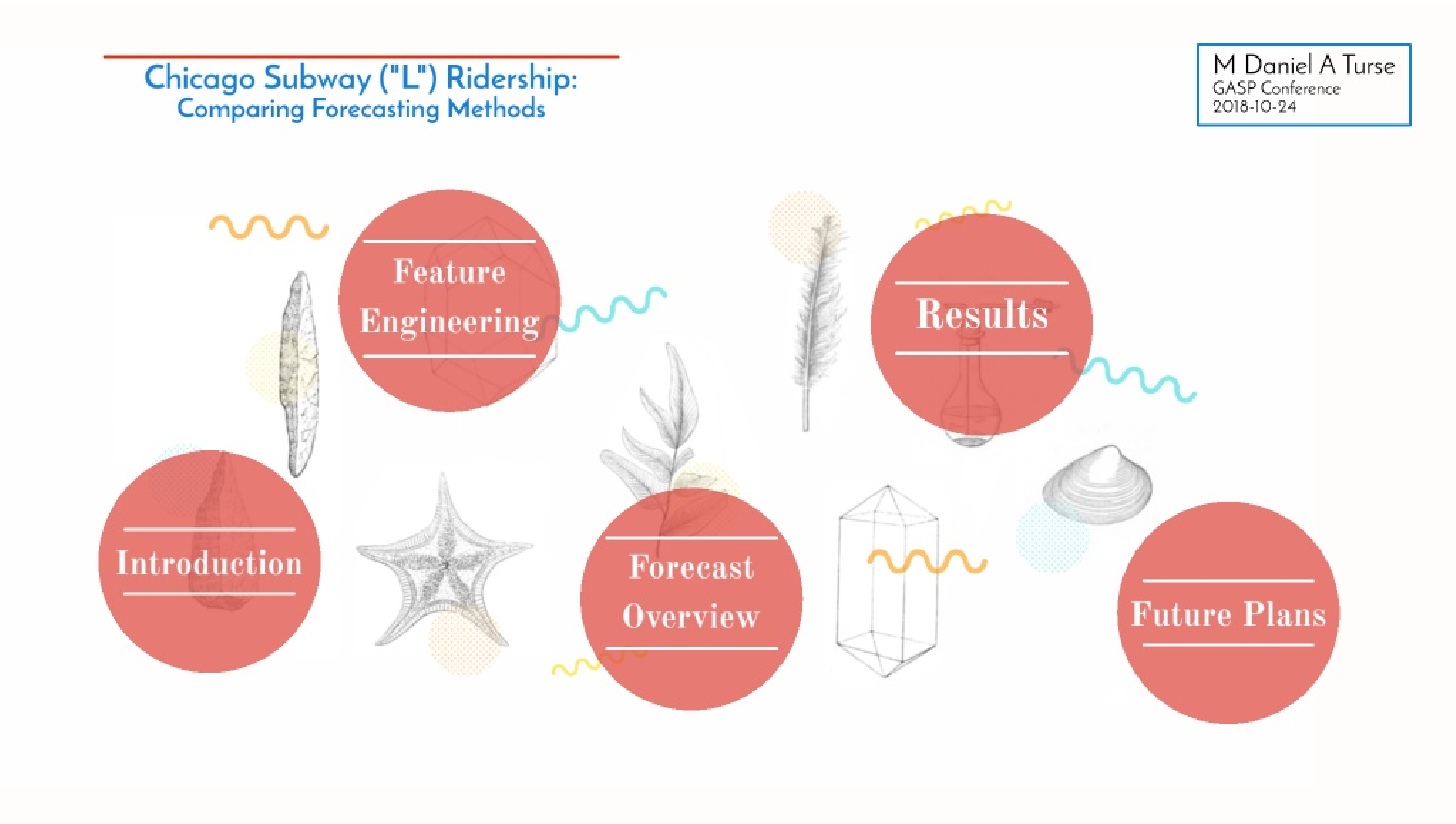
Contact Info

turse.mda@gmail.com

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

https://github.com/supermdat/Chicago_El_Divvy

http://rpubs.com/mdat/



Chicago Subway ("L") Ridership:
Comparing Forecasting Methods

M Daniel A Turse
GASP Conference
2018-10-24

Introduction

Feature Engineering

Forecast Overview

Results

Future Plans