"Analyze data taken from more than 600,000 bicycle commuter trips made in Los Angeles and **develop business solutions** for LA Bike Share, a for-profit company in Los Angeles."
1st Place in Grad
Texas A&M 2019

Collaborators:
- Josiah Coad
- Chinmay Phulse
- Brandon Walker
- Sheelabhadra Dey

April 2019
2nd Place in Grad Texas A&M 2018

Collaborators:
- Josiah Coad
- Erica
- Savinay

April 2018
"Where should we place new bike share stations in LA city?"

- How successful are current stations? Why?
- Where do people want new stations? Why?

- Think outside the box!
- What questions aren't being asked?
- What are people wasting time on that could be automated?
- Think outside the box!
- Start collecting data now.
- Streamline the data collection process for more efficient use.

- Ride data (trip time and location)
- External data (census data)
- Web-scraped data (forums)
WEB SCRAPED DATA

INSTRUCTIONS:
1. Click the button to suggest a station location or add a comment.
2. Place a new pin on the map, or click on an existing pin.
3. Submit the short survey.

Notes

- Think outside the box!
- Start collecting data now.
- Streamline the data collection process for more efficient use.
- Think outside the box!
- Start collecting data now.
- Streamline the data collection process for more efficient use.
- How to fill in missing values?
- How to format data?
- Feature engineering
  *Python, Jupyter and Pandas*

**Notes**

- This is time consuming! Try to get data as clean as possible from the beginning.
- Bias introduced when cleaning. Document this!
comment = "More transportation options for students getting to school and from school."

extract_topics(comment)

students, school

<table>
<thead>
<tr>
<th>comment</th>
<th>timestamp</th>
<th>location_lat</th>
<th>location_lgn</th>
<th>topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>More...</td>
<td>2343243.2</td>
<td>118.2437</td>
<td>34.0522</td>
<td>[student, school]</td>
</tr>
</tbody>
</table>
Data Science

Research Question ➔ Data Collection ➔ Data Cleaning ➔ Exploration & Visualization

Update Domain Knowledge ➔ Interpretation of Result ➔ Post Hoc Analysis ➔ Model Fitting "Machine Learning"

TRIP DATA

Our riders are commuters.

Downtown LA usage is on the rise!

Notes

- Visualizing is the most important part of the process.
- Convey a message.
- Test your assumptions.
**Notes**

- Visualizing is the most important part of the process.
- Convey a message.
- Test your assumptions.
Data Science

[Diagram showing steps: Research Question → Data Collection → Data Cleaning → Exploration & Visualization]

- Visualizing is the most important part of the process.
- Convey a message.
- Test your assumptions.

WEB SCRAPED DATA

Word Cloud of LA Bikeshare Comments

The bigger the word, the more it was mentioned.
Data Science

Research Question ➔ Data Collection ➔ Data Cleaning ➔ Exploration & Visualization

Update Domain Knowledge

WEB SCRAPED DATA

Heatmap of LA Bikeshare Comment Activity

Notes
- Visualizing is the most important part of the process.
- Convey a message.
- Test your assumptions.
- So many types of ML. Focus on how to pick the right one.

- The theory here is heavy. Software packages that do this for you! `sklearn`
- So many types of ML. Focus on how to pick the right one.

- The theory here is heavy. Software packages that do this for you! *sklearn*
**Feature Importance**

*Predict:* **Station Popularity**

*Using (in order of imp.):*
1. Number of rides in surrounding stations
2. Metro distance
3. Population Density
4. Percent Ppl Single
5. Percent Ppl Renting
6. Median Age

---

**Notes**
- Feature importance
- Results of prediction, forecasting
- Report performance of machine learning systems
- Feature importance
- Results of prediction, forecasting
- Report performance of machine learning systems
Station Placement Using Machine Learning

- Feature importance
- Results of prediction, forecasting
- Report performance of machine learning systems
Research Question → Data Collection → Data Cleaning → Exploration & Visualization

Data Science

Update Domain Knowledge ← Interpretation of Result ← Post Hoc Analysis ← Model Fitting

"Machine Learning"

Proximity
Metro Expo Line, Koreatown, UCLA, Bus Stops and Bike Paths

DTLA
Downtown LA is popping. Put more stations here.

User Base
Young Single Commuting

Culver City
Your next big move should be Culver City

Notes
- Answer "So what?"
- Make it actionable
- All your conclusions must be supported by your data analysis
Tools if Using Python

- Code environment:
  - Jupyter notebook (google colab)
- Data Wrangling
  - Pandas
- Data Viz
  - High Level: Tableau/Data Studio
  - Mid Level: Plotly, Seaborn
  - Low Level: Matplotlib
- Machine Learning
  - Sklearn
- Presenting Options
  - (Static) Slides
  - (Dynamic) Plots on Colab
  - Interactive Website (bonus points)