AIT614 Meeting 18 (4/20/2022)

* Finish as much analysis as possible:
  + K-means clustering
  + Decision Tree
* Begin work on Progress Report:
  + K-means
  + Time Series
  + Decision Tree
  + MARs
  + Need visualization and interpretations/analysis for each to begin writing for progress report
* Meet tomorrow, 4/21/2022
  + Begin to work on documentation and visualization tomorrow

AIT614 Meeting 17 (4/16/2022)

* Set up a pipeline for the project
* K-means clustering progress
  + Advance that if possible, may have enough analysis without it
* Time Series Analysis
  + Fix bugs if possible
* Decision Trees
  + Get progress update from Neethu
* Progress Report:
  + Begin writing by Wednesday along with final presentation
* Establish a time to meet for the project presentation

Select average speed for current hour, current hour - 1, and current hour -2 for additional cluster features

AIT614 Meeting 16 (4/13/2022)

* Fetching Data from NYC DoT
  + API (We can totally skip this part)
* ~~Data Cleaning~~
  + ~~Filtering Speeds [0 < SPEED < 68] (Filtering data to 2 standard deviations)~~
  + ~~Filtering LINK\_ID's [ count > 45000 ] (Removing LINK\_IDs with fewer records)~~
  + Rename Boroughs from "Staten island" to "Staten Island"
  + Filtering Data by timedate [maybe 2017-2022 (further analysis required)]
  + Drop 153 Link ID count to roughly ~100 link IDs, and then graph roughly ~25 link IDs in roughly ~4 charts
* Data Wrangling
  + ~~Creating the schema~~
  + ~~Loading the data into the schema~~
  + ~~Mutating hour, minute, weekday, day, month, year, dayofyear information from the DATA\_AS\_OF column.~~
  + Sampling the dataset for development, training, and testing.
  + Create a new dataframe for predictions in future.
* Data Exploration
  + Databricks display() plots
  + Linear Regression Insights file
* Data Analysis
  + ~~Linear Regression~~
    - ~~Models~~
      * ~~years~~
      * ~~month-years~~
      * ~~hour model~~
        + ~~Sequentially assign hours as numbers~~
        + ~~Aggregate average speed or median speed for each hour~~
    - ~~Predictions~~
      * ~~Future years~~
      * ~~Future month-years~~
    - ~~Visualization~~
      * ~~Group by boroughs~~
      * ~~Show the training data, testing data (and predictions), as well as predictions for the future.~~
  + MARS - Multivariate Adaptive Regression Splines
    - ~~Model~~
      * ~~Entire Dataset~~
      * Grouped by LinkId’s
    - ~~Predictions~~
    - Visualization
      * Line Chart
        + Entire Dataset
        + Grouped by LinkId’s (facet wrap by Boroughs)
  + Time-Series
    - ~~Preparation (expecting to present the preparation)~~
    - ~~Model (can't make promises)~~
    - ~~Prediction (not possible)~~
    - Visualization
  + K-means cluster
    - Model
      * Speed, LinkID, Borough
      * Speed(T), Speed(T-1), Speed(T-2), ~~LinkPoints (location)~~, timedate
    - Visulaization
      * 3D Plots
      * Maps
  + ~~Association Rule Mining~~
    - ~~Preparation~~
      * ~~Generate Congestion value (1 if Speed<20~25, else 0)~~
      * ~~Group By hour and congestion, and string\_agg(concat([LINK\_ID]), ', ')~~
    - ~~Model (based on Lab 5)~~
  + Decision Tree
    - Preparation
      * ~~If speed is less than 10, that counts as congestion~~
      * ~~If speed is greater than 10, that does not count as congestion~~
      * Aggregate speed for every hour
    - Model
    - Predictions
    - Check for average speed in NYC from another source
  + ~~Random Forest Model~~
    - ~~Preparation~~
    - ~~Model~~
    - ~~Prediction~~
* Drop the data down to 10 GB or even less to run on Databricks Community Server (suggestion for Professor Liao)
* Forest Model (if linear regression models fail)
  + 3,000 records per model
  + Time issue, takes a long time to run
* Next Meeting: 4/14/2022 Fenwick Library
  + Fon help Ewin with Association Rule mining and linear regression
  + Cross help Neethu

AIT614 Meeting 15 (4/6/2022)

* Everyone situated with GitHub
* Ewin progress in MongoDB
* Non-spark Linear Regression: <https://github.com/animesh-agarwal/Machine-Learning-Datasets/blob/master/boston-housing/Linear_Regression.ipynb>
* Can do time series analysis for certain link IDs
  + Make a for loop to figure out the coefficients for each Link ID for the Linear Regression, and then extract those coefficients to use for the time series analysis
  + Figure out which Link IDs are most prominent in the Linear Regression
  + Narrowing down which Link IDs to show in the Final Report/Submission:
    - Link IDs which are most prominent (for congestions)
    - Link\_ID’s which are major choke-points in NYC
    - LINK\_ID’s based on if something interesting pops up from LR Analysis.
* <https://github.com/maria-alphonsa-thomas/Multi-Time-Series-Pyspark-Pandas-UDF/blob/master/main.py>
* Start working on progress report due 4/17
* Linear Regression Model
  + Extract hour data
  + Split the day into four parts
  + Get average speed for that hour/quartile
  + Create line graph for data
  + Run the linear regression model with
* Data Cleaning
  + Drop null values
  + Drop values that are 0?
    - Plot 0 speed against datetime column
  + Look for when outliers appears, create a histogram, and only keep data points within one to two standard deviations
    - Anything going above 68mph should be removed, average is 38mph, standard deviations is 15mph. Keep everything within 2 standard deviations
      * Keep everything from 0mph to 68mph
      * Find a pattern for when the speed is 0mph
        + Keep these data points based on that patterns
      * Export 0mph to a CSV and then find times where it’s 0mph
        + Count by year, month, etc.
* Natural Language Processing for directions:
  + Extract words to see if a certain word appears often within dataframe for link ids and link name
* Ask for an example of a final working copy of a project to have an idea of what we’re aiming for in terms of presentation

AIT614 Meeting 14 (4/2/2022)

* Azure vs. AWS as Big Data Platform
  + Leaning towards AWS
    - 14 day trial/limit on data storage without a membership

AIT614 Meeting 13 (4/1/2022)

* Project Requirements (NoSQL Databases, Big Data Engines/Machine Learning Engines/Platforms)
  + Host on Databricks/AWS/Azure/Google Data Factory
  + Save the data on a MongoDB Cluster running on Spark, hosted by Databricks
  + Recommendation for df2, since the Link Points is being trimmed due to strong character limit
* Everyone runs a linear regression model based on Sagar’s cleaned data
  + Potentially add an asterisk (note) for running the model on different machines with different specs
* Drop the data down to 10 GB or even less to run on Databricks Community Server (suggestion for Professor Liao)
* Forest Model (if linear regression models fail)
  + 3,000 records per model
  + Time issue, takes a long time to run
* GWR (if time allows)
  + Clean the data to focus on Link Points
  + Clipping issue
  + Repeat polylines
  + Find the shortest polyline in terms of characters
* Share link to Google Drive containing all project files
* Calculate how far away the origin is from the endpoint for Link Point Distances
* Go over how to use GitHub
* Github Link to push files to:
  + <https://github.com/minoseah629/ait614project.git>
* Next meeting Saturday, 4/2/2022 or Sunday, 4/3/2022

AIT614 Meeting 12 (3/30/2022)

* Short Status Report:
  + Neethu & Fon:
    - Attempt to load trimmed down dataset
    - Work Around (Sagar): Random sample of data with 5000 records, possibly not ideal, but could make it easier to upload the data and run models on it
* In-Person Meeting:
  + 3/31/2022: Fenwick, 2nd Floor,
  + Overall: Meet based on availability, do tasks alone when possible
    - Ewin: Weekdays after 5pm; weekend anytime
    - Sagar: Anytime - [MW 4-7 Th 7-10]
    - Fon:
      * Work on Friday, Saturday, Sunday,
      * Lecture on Tuesday, Wednesday
    - Cross:
      * Anytime after 1pm and before 6pm on Monday, Tuesday
      * After 2-3pm to 5pm to 6pm on Thursday, Friday
      * After 1pm Saturday, Sunday
    - Neethu: lectures on Mon,Wed(can meet before 4:30)

Thursday,Friday ,Sunday (Anytime)

* Ask professor if we can host the project on the local spark server
  + Speed issue
  + Recreating the cluster every time
* Sample data for Fon (500,000 records)
  + Possibly post it to Teams
* Group By: Getting list by all boroughs
  + For every link ID
  + Direction of link ID for each record

AIT614 Meeting 11 (3/24/2022)

* Link to API:
  + <https://api-portal.nyc.gov/docs/services/dot-datafeeds/operations/bikeinfo>
    - [**https://api.nyc.gov/dot/bikeInfo/{fileName**](https://api.nyc.gov/dot/bikeInfo/%7BfileName)**}**
  + [**https://api-portal.nyc.gov/docs/services/dot-datafeeds/operations/trafficSpeed**](https://api-portal.nyc.gov/docs/services/dot-datafeeds/operations/trafficSpeed)**?** 
    - [**https://api.nyc.gov/dot/trafficSpeed**](https://api.nyc.gov/dot/trafficSpeed)
* Show professor dataset and ask if clustering can work on the data (or if it would be useful)
* Note everyone’s availability again:
  + <https://www.when2meet.com/?14813212-oE7cc>
* Data cleaning
* Linear regression model (with raw and cleaned data)
  + Discuss how to carry out our particular models with the professor
* Data visualization
* Next Meeting: Next Wednesday/Thursday/Saturday (March30, March 31, April 2)?

AIT614 Meeting 10 (3/23/2022)

* Meet tomorrow, 3/24/2022, after 2pm in Fenwick Library (Cross, Sagar, Fon, Neethu)

Project Proposal Edits:

* Data Preparation:
  + Store unnecessary columns in an array
  + Mutate new columns for certain data like seconds, minutes, hour, etc.
    - May not need to do this for SQL Spark functions, YEAR, DATETIME
* Data Analytics:
  + Linear Regression Model: Cross, Neethu
  + Clusterting: Sagar
    - Experiment with what variables to cluster
    - Attempt 3/24/2022
* Data Visualization:
  + Back-End: Everyone (likely have to learn as we go)
  + Front-End: Everyone
* Have models run on Ewin’s machine to have outputs come out faster
* Store project proposal in Microsoft Team Files (with editing access)
* Possibly have a Google Drive additionally for all project files (original links to them)
* Ask for examples of final projects to get a better idea of expectations, either from TA or professor

AIT614 Meeting 9 (3/19/2022)

# Project Proposal Skeleton / Outline

* **Project Goals**
* **Project Requirements**
* **Proposed Selected Dataset**
  + 24 Gb csv data
  + APIs
* **Description of Proposed System**
  + Conceptual System Architecture Diagram
  + Data Analytical Methods
    - Data Ingestion and Cleaning
    - Data Exploration
    - Linear Regression Models to predict Congestion  
      (Maybe Geographically Weighted Regression)
      * With respect to location and time
        + Hourly Predictions
        + Daily Prediction
        + Weekly Predictions
        + Monthly Predictions
        + Annual Predictions
    - Random Forest / Feature Extraction
      * Figure out bottlenecks by feature extraction of the average traffic speed for neighboring locations
    - Clustering locations and sensors
      * Based on Individual Sensor values
      * Based on Individual w.r.t. values for neighboring sensors / localities
        + Values of frequencies
        + Values of Congestion Intervals
    - Predictions Dashboard (Heatmap / Bubble Map) (if time allows)
      * Runs based on weekly data
        + Back-end Scheduled Script

Data Extraction | Extracts data from the original dataset source/API to display data

Running the model | Training the model based on the new data

* + - * + Front-end Dashboard Design

Design of the dashboard

File that keeps updating with current coefficients

* + - * Hourly predictions
      * Weekly predictions
      * Monthly predictions
      * Annual predictions
    - Case Studies
    - Documentation and Deliverables
      * Reports
      * Presentation
        + Slide Deck: Cross
        + Share Screen for working demonstration: Cross
* **Proposed Development Platforms**
  + Python | Spark
  + R | Spark
* **Project Tasks and Timeline**
  + A
* References
* Appendix

# Project Proposal | Assigning Sections

* + **Tasks Member**
  + Introduction Cross
  + Project Goals Cross
  + Project Requirements Cross
  + Proposed Selected Dataset Ewin
  + Description of Proposed System Sagar
  + Proposed Development Platforms Neethu
  + Project Tasks and Timeline Fon
  + References Everyone
* Distribute roles

AIT614 Meeting 8 (3/9/2022)

* DOT Traffic Speeds First Impressions
  + 154 sensors
    - Null value, something off in calculation
  + Tableau map graph
    - Took all points in link point column, split them up, and then put them into the longitude and latitude graph
  + ID Number, average speed. Group dataset by ID number, then find average speed for ID number
    - Column: ID
    - Rows: AVG(Speed)
  + API for NYC to see if it can work
* API for NYC to see if it can work
* Feedback for proposal:
* Project Proposal Outline: <https://docs.google.com/document/d/1pYJuclcpnxOi1FupuL-KACfxB4xdDqK4wODDeZpSRSw/edit?usp=sharing>
* Get NYC Public Transportation usage data (Sagar)

AIT614 Meeting 7 (3/2/2022)

* Farewell to Cheng
* Welcome Neethu
* Ewin loading records into dataframe
  + Extract minute of day from model, then construct model for average traffic speed for day-to-day
  + https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1927099756038143/3898273970404299/7738811756334167/latest.html
* Next thing to do on the project:
  + Main Project Goal: Analyze traffic records to create a model to predict what day(s) experience the most traffic
    - Future Considerations: Geospatial analysis, see where most of the traffic is (regions)
  + Basic data exploration/explore the datasets
    - Distribution of data (e.g. less data on Sundays)
  + Have Ewin give everyone access to the databook
  + Access to CSV files?
    - <https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9> (all of the raw data)
  + Format to work in:
    - JSON
    - XML
  + Integrate all links to project files into wiki for ease of access

AIT 614 Meeting 6 (2/27/2022)

* Finish project topic paper
  + Submit and then have everyone check if it was submitted for them
* Project Topic
  + Domain of the project
* Project Proposal
  + Solving the traffic is a much bigger scope than we may be able to achieve. May still want to give suggestions based on insight into data
* Draft Email to TA on Neethu Situation:

Good afternoon Rajeev,

I hope you are doing well. We have a few queries regarding the Group Project. Neethu Battula is requesting to join our team due to certain circumstances in her team that caused a division within her group. Would you be okay with her joining our team, Project Team 1 of section 002, for the semester?

Additionally, for any project deliverables we turn in like the project topic submission for tonight, 2/27/2022, does only one person need to submit for the whole group, or does everyone from the group need to submit individually? Thanks for your time and consideration.

Best,

Project Team 1

AIT 614 Meeting 5/5.5 (2/25/2022 & 2/26/2022)

1. Discussed NYC traffic dataset for traffic prediction (Multiple Linear Regression Models, K-means Clustering, Geo-Spatial Analysis (if possible))
   1. Issue with missing data
   2. Finding insights to improve the NYC traffic system
      1. Choke points/Bottlenecks in the current transportation infrastructure
   3. Potential points/regions to improve upon
   4. Audiences
      1. Residents can use it to know when there will be high volumes of traffic so they know when to leave for optimal commute times
      2. Emergency services can use it to know what roads to travel to improve their efficiency
      3. Stakeholders in transportation sector could use outcomes to provide better services, such as taxis, to increase profits
      4. Assist Department of Transportation in framing better policies and provide them with insights for better infrastructure planning.
   5. Link to dataset: <https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9>
   6. Tools
      1. Use Python/R to clean up some null value.
      2. Excel power query?
   7. Post-Project
      1. Government Case Studies
      2. White Papers by Stakeholders in Transportation Industry
2. Used Car with time series dataset (regression trees, linear regression models, Time-Series Analysis and Feature Selection/Analysis)
   1. Most achievable?
3. Marital Status - problematic dataset in process (Marital Status Child Outcome data)
   1. Choose a certain approach
   2. Split parents’ status into roughly four categories, then judge children’s educational achievements and economic outcomes from this
   3. Description of a group of families localized to one region
      1. Use an API script to analyze this data
4. Transportation - Airline delays (predict flight delay to assist travelers experiences, real time weather conditions)

# To-Do:

* Project topic 1 is the selected topic for project topic submission by 2/27/2022, change subject if need be in the future for the proposal
* New York City traffic submission posted on discussion board 2 by Sagar for AIT614, can be used for project submission, look for if the posted dataset can be used for submission
* Sagar and Ewin confirm if datasets are good to go or not
* Have Neethu Battula join the group as the sixth member
  + Confirm with TA and/or teacher
  + Solidify roles within the group to prevent too much redundancy
    - Lead coders
    - Documentation (people who write documentation for the code)
    - Lead writers (may overlap with coders)
    - Other roles?
* Confirm that one person submitting the project deliverables submits it for everyone else as well on Blackboard
* Have a valid backup topic in case we change our minds or one project falls through

# Project Preference Order:

Ewin - 1234

Sagar - (31)42

Fon - 1243

James - 12

Cross - 1342

# Project To-Do:

* Create a database hierarchy diagram for Project 3. (Sagar)

AIT614 Meeting 4 (2/23/2022)

* Select project topic by 2/26/2022
  + Needs to be submitted for approval in the project topic document
* Get link to where autos(1).csv comes from (https://data.world/data-society/used-cars-data)
  + Ask professor if we could do the project on this database
  + Hypothesis/Question:
    - Figuring out which factor is the most important to determining the cost of a used car
      * Which factor determines the cost of a used car the most?
    - Predicting the price of used cars?
    - After a certain amount of time, used cars should cost this much?
      * Could also be “after a certain number of miles, the car should cost this much”
      * Compare the same models but with different colors, does that affect the price
        + Or any other condition
    - Certain brands of cars will cost more (after x amount of time) when being categorized as a used car.
      * Which brand has the highest price?
* Marital Status dataset:
  + Focus on Virginia, D.C., Maryland
* May have to abandon the FBI dataset
  + Varying number of characters per line
  + Read over the “Help” manual to parse it better.
    - Is it formatted with various column headings that we can parse through
* Is R2 based on our individual hypothesis, or a team submission?
* Ask the professor if Spark or other platforms handle machine learning algorithms well, and if they may be appropriate for this project
  + Certain machines may not be able to run all of this data, make a note of which ones can’t
  + Running projects on cloud or PCs, be careful of the amount of power we have access to
* List the primary topic along with every other topic to get approval of them in the email to the professor
  + FBI Incident data
    - To what degree are there racial disparities in incident reporting for crimes within the U.S.?
  + FBI Arrests data ()
    - To what degree are there racial disparities in arrests for crimes within the U.S.?
  + Marital Status Child Outcome data (70k records, 15 columns, 12 years worth of data)
    - “Children in single parent households have worse outcomes educationally” or “Children in single parent households have worse outcomes educationally even when controlling for income”
  + Used Automobile pricing data (600k records, 10+ columns)
    - Figuring out which factor is the most important to determining the cost of a used car
      * Which factor determines the cost of a used car the most?
  + Flight delay prediction system

Possible data set: (https://www.transportation.gov/policy/aviation-policy/us-international-air-passenger-and-freight-statistics-report)

(<https://www.transtats.bts.gov/OT_Delay/ot_delaycause1.asp?qv52ynB=qn6n&20=E>) (310k records, 10+ columns) (use R to get the number of rows)

* + - We can use big data to predict when a flight will be delayed to improve the travel of patrons
* Ask the professor if the autos(1).csv is appropriate and if the project hypothesis is appropriate
  + Need to get approval before working on project topic file? Or is the project topic approved after submission?
* If the professor approves of none of the datasets, what do we do?
  + Try to find another dataset by ourselves
  + Ask the professor for help in finding a topic
    - What’s an appropriate topic?
    - Hypotheses we should be developing?

AIT614 Meeting 3 (2/19/2022)

* Project Topic: Due 2/27
* Ask about the FBI topics and if they’re doable this semester and appropriate
  + In terms of big data (does it count as big data)
* What analytical tools can we/do we need to use: R, Python, etc.?
* Are we using the NoSQL to make the query or something else?
* Find a way to parse and make sense of the .txt file (Flatfile Format) (https://crime-data-explorer.fr.cloud.gov/pages/downloads#:~:text=Master%20File%20Downloads) to get the desired data
* Find out if the project topics are too simple or complex
* Meet next Wednesday, 2/23/2022

Draft Email to Professor Liao (about FBI topics):

To: dliao2@gmu.edu

CC: [sgoswam@gmu.edu](mailto:sgoswam@gmu.edu) [clee94@gmu.edu](mailto:clee94@gmu.edu) [ehong8@gmu.edu](mailto:ehong8@gmu.edu) [shapikul@gmu.edu](mailto:shapikul@gmu.edu)

Title: AIT614 002 Project Topic Questions

Hello Professor Liao,

I hope you’re doing well. Our team, project team 1, of section 002, had several questions about the project. First, does our data set need to be of a certain size in terms of number of records for the project (i.e. being big data)?

Secondly, we have several potential topics. One is about the outcomes of children in single-parent vs. still married couples. Another is about how Big Data analytics can be used to optimize transportation operation and boost customer satisfaction by predicting flight delays.

Our other two topics are about FBI incidents about crime reporting and another concerned FBI arrests, both dealing with racial disparities. Are all of these appropriate topics to pursue? Which one in your professional opinion would you suggest we choose for our project?

For the FBI incidents reporting, would this be an appropriate topic to pursue for the semester in terms of big data as the given .txt file, which is a Flatfile, is 3.59 GB in size but we haven’t been able to parse through it quite yet to figure out how many records it has.

And are we using NoSQL for the queries or for something else?

For our analysis, do we need to use R, Python, NoSQL, or can we use any tool of our choosing? Are we supposed to host our project Database on Spark or Flink?

Also, is the due date for the project topic 2/20/2022 or 2/27/2022?

Thanks for your time and consideration.

Best,

XXX

AIT614 Meeting 2

* Research Questions (based on specialized knowledge from your experiences)
* Databases that can be integrated with Sagar’s databases, or work on Saga
* Survey form for our working experience
  + Look for common experiences and interests
  + Look for certain expertise like machine learning
* Project requirement: >= 10000 and nosql
* Favorite Topics (pick a favorite by 2/18 or 2/19, coin flip otherwise):
  + Big Data analytics can be used to optimize transportation operation and boost customer satisfaction by predicting flight delays. (Suchada, Ewin)
    - Real time data
    - Consider weather conditions
* “Children in single parent households have worse outcomes educationally” or “Children in single parent households have worse outcomes educationally even when controlling for income” (Sagar, Ewin, Cross)
  + Consider other categories, such as together but never married, never married and separated,
* FBI Incidents Reporting (Cheng-yuan)
* FBI Arrests Reporting (Sagar)

AIT614 Meeting 1

[Names]

1. Cross Zeigler (leader, submit to teacher)
2. Suchada Hapikul
3. Ewin Hong
4. Cheng Yuan Lee (moved to another section)
5. Sagar Deepakgiri Goswami
6. Neethu Battula

Agenda:

* Communication
  + When2Meet for times outside of regular class time
  + Zoom vs. Microsoft Teams for video chats in the future
  + Microsoft Teams for general communication
* Backgrounds
  + Interests
  + Skills
    - Things Listed in Class: Python, SQL, NoSQL, MongoDB, MapReduce, Hadoop, Spark
    - Cross
      * SQL, Python (Pandas)
      * Analytics (Maryland car crashes Database, D.C. gentrification, CO2 emissions by country over a decade)
      * Data visualization (CO2 emissions by country over a decade)
    - Suchada
      * SQL, (Beginner in) R, Python, Tableau, MongoDB Compass
    - Ewin
      * SQL, Python and C#/powershell, tableau
      * Couple machine learning projects and nlp visualizations
    - Cheng(James)
      * SQL, R
      * Linear regression,CV, visualization
    - Sagar
      * R, Python, basic SQL,
    - Neethu
  + General Topics
    - Medicine
    - Transportation
    - Infrastructure
    - Family/Childcare
      * Impact of single parent households on children’s performance/outcomes (educational achievement, happiness, stress, salary/job mobility) (Cross)
      * How do siblings (or the absence thereof) affect the socialization of young children (number of friends, social anxiety, etc.)
    - Education
      * Impact of education on wages (there being a plateau at a certain point?)
* Potential project topics
  + How does big data help people understand the relationship between the inflation rate and some US largest technology companies? (Cheng)
  + Big Data analytics can be used to optimize transportation operation and boost customer satisfaction by predicting flight delays. (Suchada)
    - Real time data
    - Consider weather conditions
  + How big data predicts the U.S. used car market
* Past Topics:
  + What social, political, and economic factors have led to the stagnation of wages and increasing wealth inequality in the U.S.? (Cross)
  + What has been the impact of social media algorithms spreading fake news in the past decade (alternatively: How have social media algorithms been abused to spread fake news in the past decade?) (Cross)
  + Blockchain tweet analysis in blockchain market cap (Ewin)
* Future Considerations
  + Decide on a project topic by 2/18 or 2/19
    - Pick your favorite topic
  + Ask if we need to use Big Data for our dataset or if we only need to use big data tools (databases, engines)
    - Focus on velocity or volume
  + Survey form for expertise on machine language processing, APIs, complex relational databases, etc. for the team
    - Industry experience
  + Meet before or after class sessions to go over project