Twitter Sentiment & External Factors
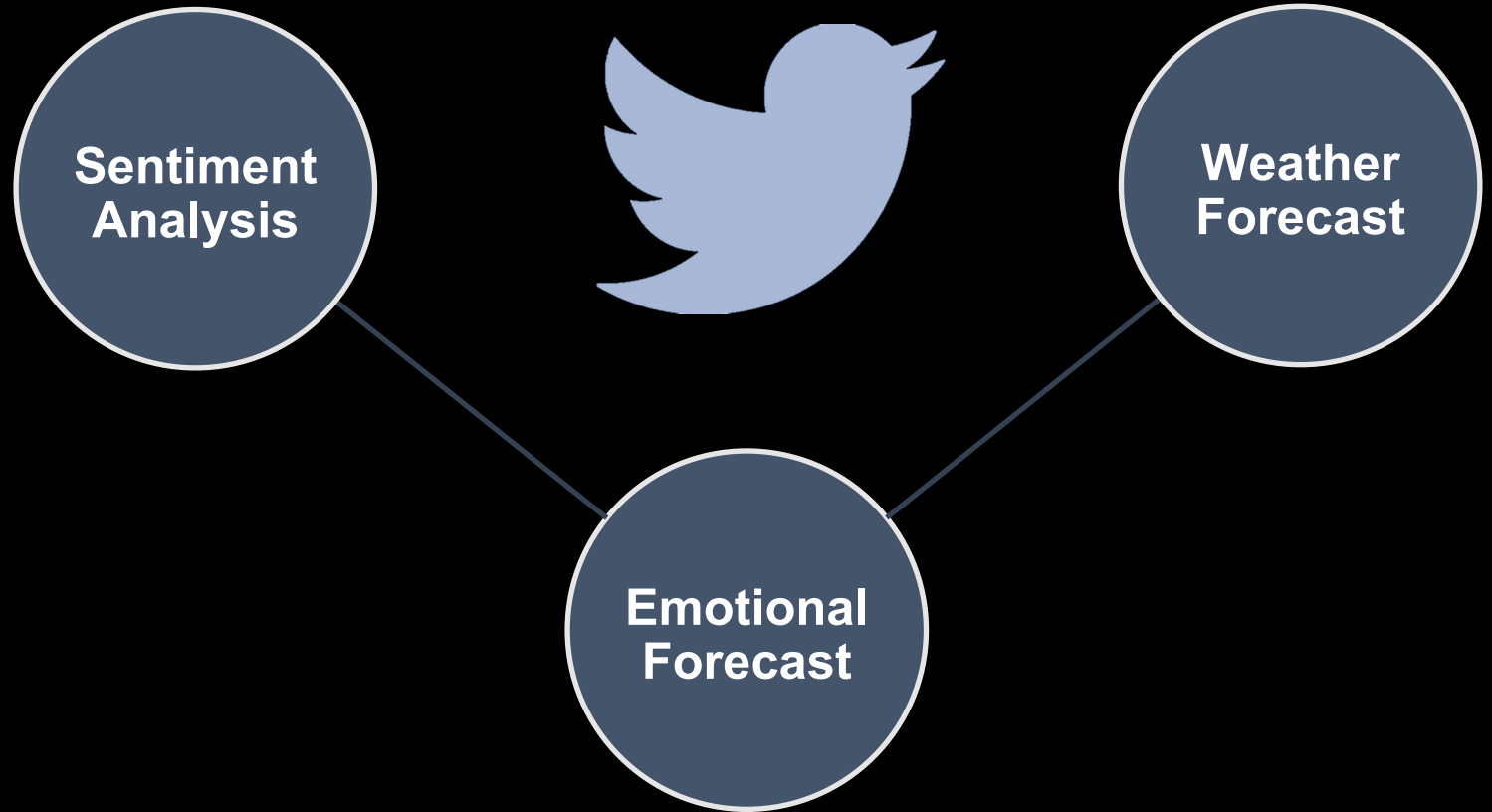
Projects in Data Science: Python

Zach Bogart, Josh Feldman, Joe Gamse, Ramy Jaber, Pit Kauffmann

INTRO

What external factors influence how people tweet?

Eg can we create a weekly 'emotion forecast' for Twitter based on the weather forecast?

DATA

TWEETS

- Use tweepy streamer to stream tweets from specified locations – running on Google Cloud
- ~200k total tweets from 3 cities
- 18500-word list with sentiment scores between -1 and 1

WEATHER

- Get weather data for specific weather stations from NOAA (ftp.ncdc.noaa.gov), corresponding with specified locations for Tweets
- Includes temperature, wind speed, cloud coverage, precipitation

METHODOLOGY

TWEETS – Stream in for each location

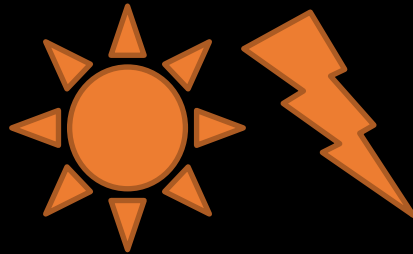**SENTIMENT SCORE – calculate sentiment of each tweet using wordlist**

WEATHER – add weather data for each tweet from closest weather station

TRAIN – train model on weather/tweet data using Random Forest, Extra Trees and Bagging Classifiers

PREDICT – use forecast weather data to predict change in sentiment by location
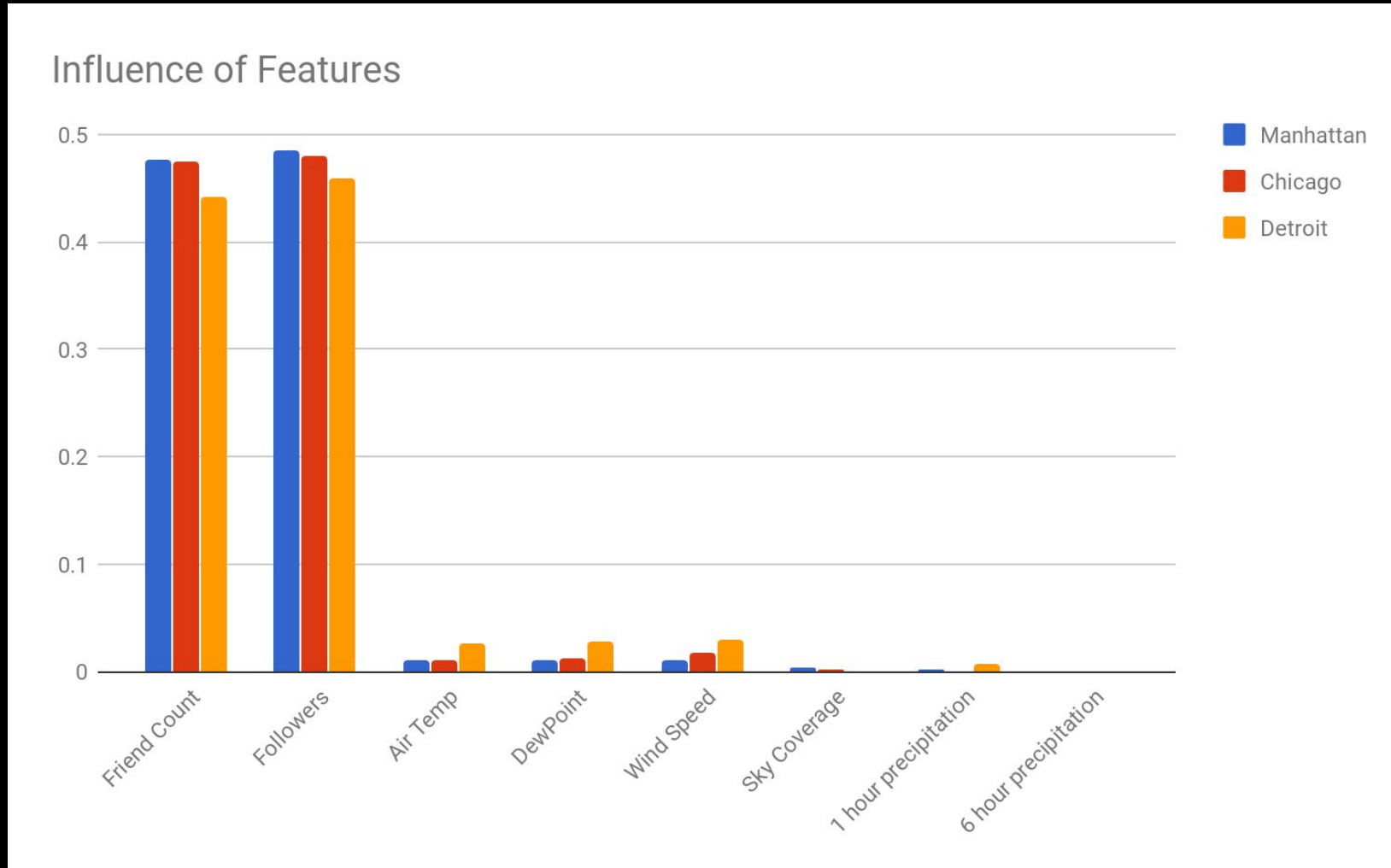
PRELIMINARY RESULTS

Goal: Classify each tweet as either positive, negative, or neutral sentiment

Features: Friend Count, Followers, Air Temp, Dew Point, Wind Speed, 1 hour precipitation, 6 hour precipitation

Test Accuracy	Training Accuracy		
Manhattan	~85K	0.845	0.425
Chicago	~65K	0.818	0.4
Detroit	~12K	0.815	0.405

PRELIMINARY RESULTS (cont'd)

PRELIMINARY RESULTS (cont'd)

Confusion Matrixes (Manhattan - testing on 25% of tweets)

ALL FACTORS	Predictions			
		Neutral	Positive	Negative	Total
Actual	Neutral	1987	1966	1540	5493
	Positive	1715	4129	2607	8451
	Negative	1493	2977	2978	7448
	Total	5195	9072	7125	

WEATHER ONLY	Predictions			
		Neutral	Positive	Negative	Total
Actual	Neutral	0	5171	388	5559
	Positive	0	7919	608	8527
	Negative	0	6782	524	7306
	Total	0	19872	1520	

CONCLUSION & NEXT STEPS

- Collect tweets for even more locations and across a longer time period

- Run each classifier with more parameters to find the best fit

- Run more classifiers: K-means, Support Vector Machines, K-nearest neighbours, Naive Bayes

- Make sentiment analysis more granular & precise

- How do other factors, such as crime rates, sports events, etc. affect overall sentiment?

QUESTIONS?