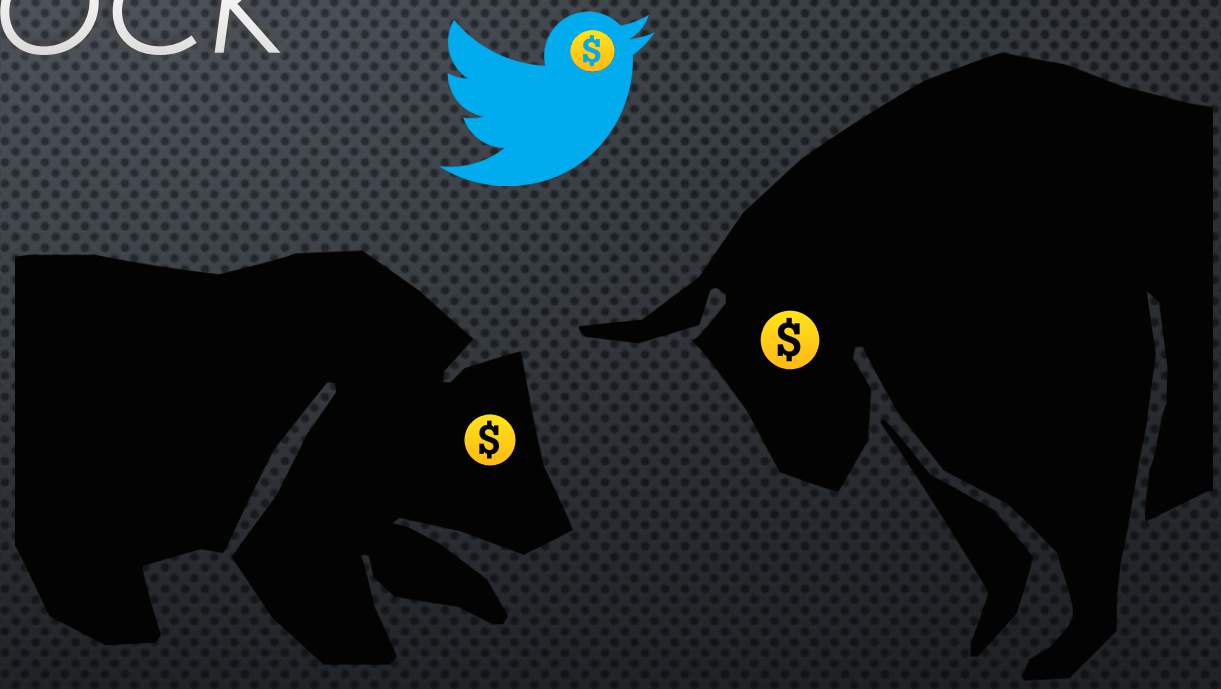
DS501 CASE STUDY 4 - *TWITTER & THE STOCK MARKET*

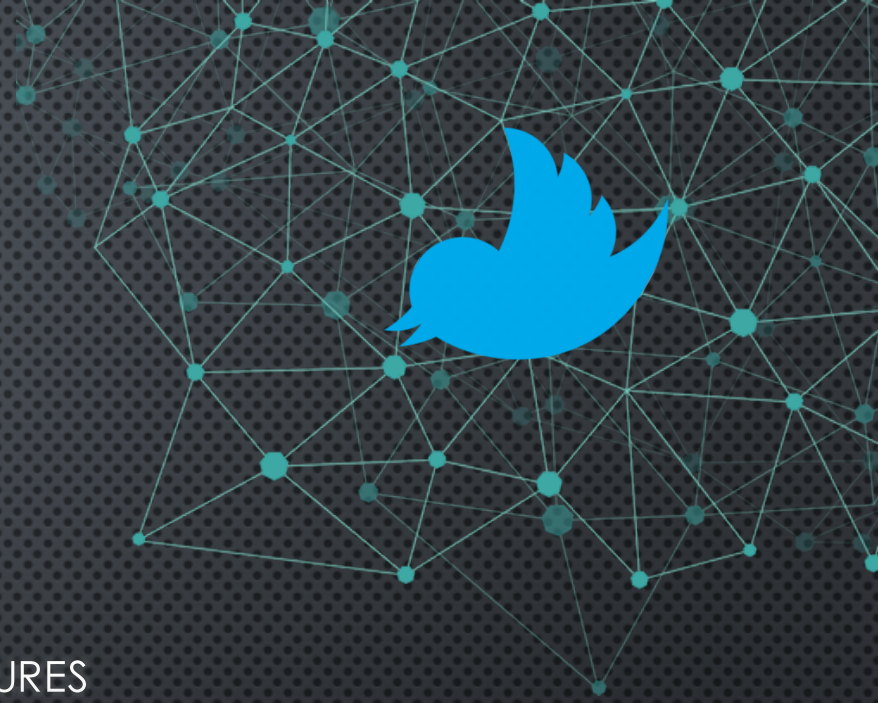


GROUP 6: QIAN WANG, ELIZABETH KARPINSKI, JINGNAN XU, XIAOSHUAI LI

INTRODUCTION

- **Social media**: the easiest & fastest way to transmit and receive information.
- **Twitter**: Aggregate of tweets could be seen as an indicator of collective mood.
- **data scientists** have been made several attempts to examine Twitter's predictive potential of consumer purchasing by observing users' mood.
- "*Twitter mood predicts the stock market*" by Bollen, Mao, and Zeng in 2011.
- **Our Analysis: Analyze Twitter data through sentiment analysis & compare it with instead with the SP500 and Nasdaq index (4 different tests)**

DATA COLLECTION



- DATA SOURCE: BOTH TWITTER & VARIOUS STOCK MARKET MEASURES
- TIME: EVERY 10 MIN, FROM 9:30 AM TO 4:00 PM EST (DEC. 1 ~ DEC. 3)
- DATA FORMAT: RATIO OF POSITIVE VS. NEGATIVE WORDS / DIFFERENCE BETWEEN THE NUMBER OF POSITIVE AND NEGATIVE WORDS. *(A FORM OF SENTIMENT ANALYSIS)*
- USE THE NASDAQ AND SP500 INDICES: TO EVALUATE CHANGES IN THE STOCK MARKET
- *4 DATA SETS ➡ 2 OVERARCHING CATEGORIES (TWITTER MOOD & STOCK MARKET INDICES)*

STATISTICAL METHODS

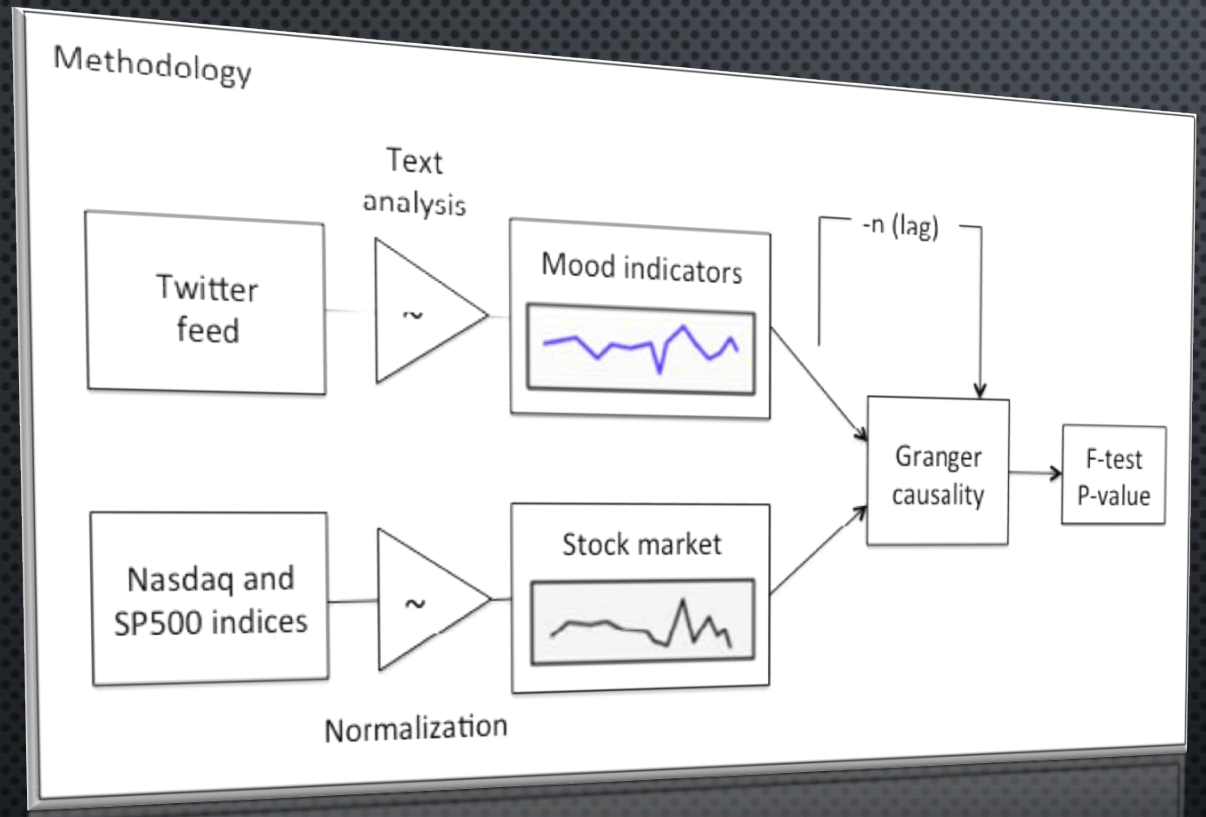
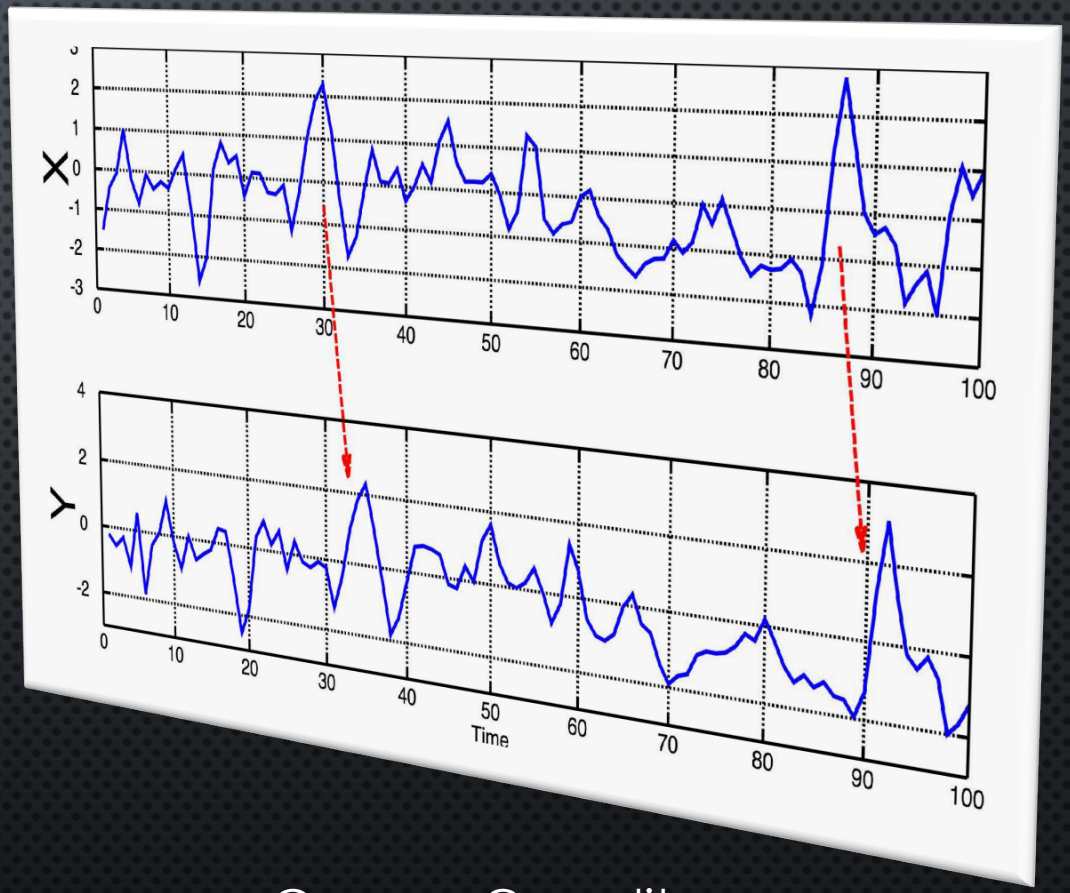


Diagram outlining 3 phases of methodology & corresponding datasets

In order to make linear predictions based on the data collected, we elected to use **Granger Causality testing**.

STATISTICAL METHODS

- Past vales of X can be used for the prediction of future values of Y.
- A time series X is said to Granger-cause Y if it can be shown, usually through a series of F-tests on lagged values of X, that those X values provide statically significant information about future values of Y.



Granger Causality

STATISTICAL METHODS

$$Y_t = \alpha + \sum_{i=1}^{n} \beta_i Y_{t-i} + \varepsilon_t \quad (1)$$

$$Y_t = \alpha + \sum_{i=1}^{n} \beta_i Y_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \varepsilon_t \quad (2)$$

To examine whether Twitter mood time series predicts changes in stock market, we compared the variance explained by 2 linear used to help compare and explain the variance:

- Model 1: Use only n lagged values of Yt to predict the future of stock market.
- Model 2: Use the n lagged values of both Yt and the Twitter mood time series denoted Xt.

DATA ANALYSIS & RESULTS

F-testing for Data of December 2, 2015

* P-value < 0.05
** P-value < 0.01

(Positives/Negatives) vs. SP500 index					
Number of lags	1	2	3	4	5
F-statistic	3.7735	2.1788	1.2376	1.3980	0.8668
P-value	0.0602	0.1297	0.3140	0.2623	0.5182

(Positives/Negatives) vs. Nasdaq index					
Number of lags	1	2	3	4	5
F-statistic	8.1114	4.4031	1.4935	2.8027	1.7895
P-value	**0.0073****	**0.0205***	0.2371	**0.0465***	0.1548

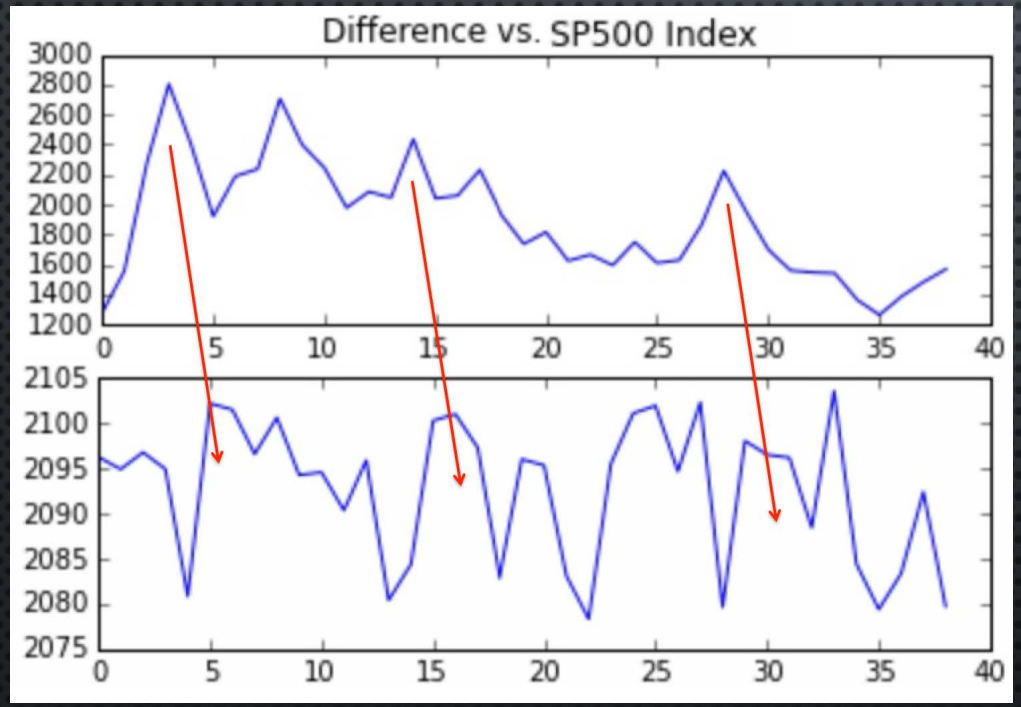
(Positives - Negatives) vs. SP500 index					
Number of lags	1	2	3	4	5
F-statistic	7.0728	4.6334	2.1526	1.9519	1.3555
P-value	**0.0117***	**0.0171***	0.1152	0.1318	0.2772

(Positives - Negatives) vs. Nasdaq index					
Number of lags	1	2	3	4	5
F-statistic	10.1884	7.6876	2.8832	3.1745	2.4234
P-value	**0.0030****	**0.0019****	0.0527	**0.0299***	0.0664

Given results of our Granger causality above, we can reject the null hypothesis that **Twitter mood time series do not predict stock price with a high level of confidence**.

The ratio as Twitter mood **doesn't have significant causal relations** with SP500 index, but **does shows Granger causality with changes** in Nasdaq index.

DATA ANALYSIS & RESULTS



Twitter mood and SP500 index of December 2, 2015

- Correlation between Twitter mood difference and the stock market indexes: *Changes in past value of Twitter mood (t-2) predicts a similar rise in SP500 values (t-0)*
- **The Twitter mood thus has *predictive value* regarding to the stock market indexes.**

DATA ANALYSIS & RESULTS

F-testing for 3 days period

* P-value < 0.05
** P-value < 0.01

(Positives/Negatives) vs. SP500 index					
Number of lags	1	2	3	4	5
F-statistic	4.1715	1.9049	1.2195	0.4457	0.5761
P-value	**0.0435***	0.1538	0.3064	0.7753	0.7182

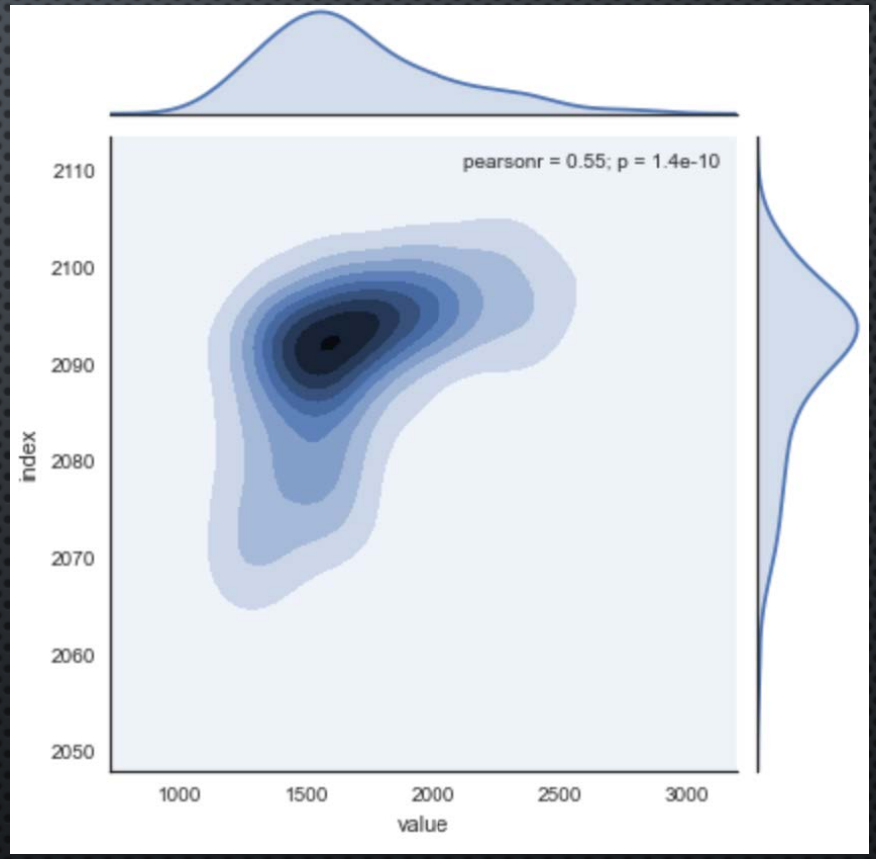
(Positives/Negatives) vs. Nasdaq index					
Number of lags	1	2	3	4	5
F-statistic	5.2212	2.4186	1.4228	0.7359	0.7431
P-value	**0.0242***	0.0939	0.2403	0.5696	0.5931

(Positives - Negatives) vs. SP500 index					
Number of lags	1	2	3	4	5
F-statistic	12.9210	6.7096	3.6875	1.6029	1.2570
P-value	**0.0005****	**0.0018****	**0.0143***	0.1793	0.2887

(Positives - Negatives) vs. Nasdaq index					
Number of lags	1	2	3	4	5
F-statistic	11.4039	6.1914	3.6221	2.0130	1.5756
P-value	**0.0010****	**0.0028****	**0.0155***	0.0982	0.1740

DATA ANALYSIS & RESULTS

- Difference between positive and negative words as the indicator of Twitter mood has the **highest Granger causality** relation with SP500 index and Nasdaq index for lags ranging from 10 to 30 minutes (p-value < 0.05).
- Take Twitter mood difference and Nasdaq index as an example, when the number of lags is 3, the p-value of F-test is only 1.55%, which means in 95% significant level, we can reject the null hypothesis of the two series data do not have causality.



Twitter mood and SP500 index

CONCLUSION

- CHANGES IN PUBLIC MOOD COLLECTED FROM TWITTER CAN BE TRACKED FROM CONTENT OF TWEETS BY MEANS OF TEXT PROCESSING TECHNIQUE.
- BETWEEN TWO TWITTER MOOD INDICATORS, THE DIFFERENCE BETWEEN POSITIVE AND NEGATIVE WORDS IS GRANGER CAUSATIVE OF THE STOCK MARKET INDICES, BECAUSE CHANGES OF THIS MOOD INDICATOR MATCH SHIFTS IN THE SP500 INDEX AND NASDAQ INDEX THAT OCCUR 10-30 MINS LATER.


THANK
YOU