# Crypto Recommender Final Update

Sean Hung, Yu-Kai Wang, Matthew Uryga, Jonathan Li

# Initial Goals

# Initial Goals

- Use crypto price history and real-time price and sentiment analysis to identify similar cryptocurrencies.
  - Construct a graph using this data
  - Use this to predict future price action
- Scrape data from Twitter and Reddit for real-time news and sentiment
  - Using BERT for sentiment analysis
- Construct an aggregate similarity metric that takes all of the data into account
- Determine if each crypto is bullish/neutral/bearish
  - Possibly output a price target for a given crypto

# Data Collection and Processing

# Data Collection and Processing

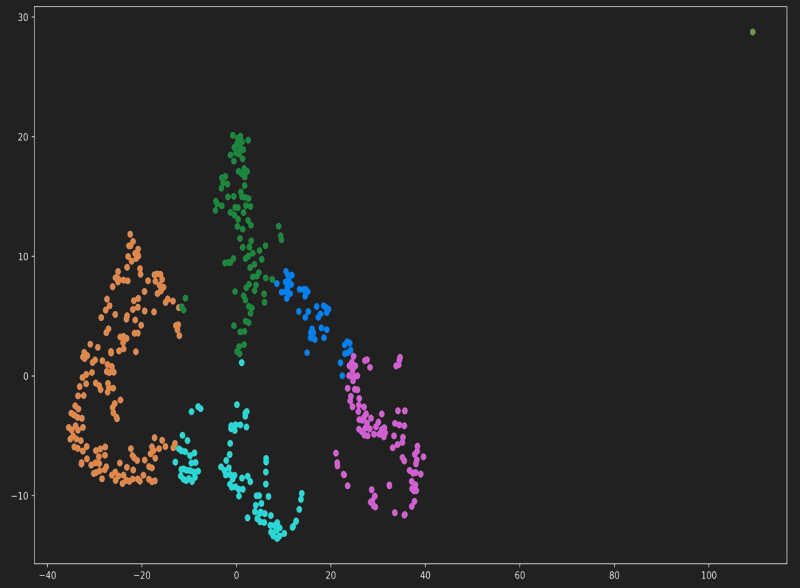
- Scraping data from various social media and crypto information sites
  - Twitter
  - Reddit
  - Binance API
  - Coin Market Cap
- Applying various clustering algorithms to create similarity metrics
- Sentiment Analysis
  - Bert-uncased model from hugging face
  - Evaluating stock performance on a given day based on reddit headline posts
- Calculating price data with technical analysis indicator

# Similarity Algorithms

# T-SNE

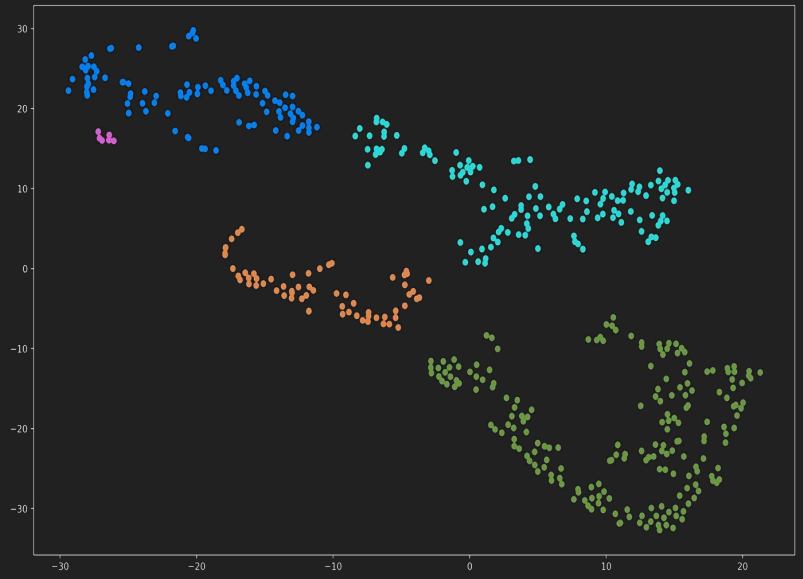
## T-distributed Stochastic Neighbor Embedding

- For each crypto assets, we have a list of metadata associated with it
  - Max supply, market cap, circulating supply, etc.
- To aggregate these values, we trained a T-SNE model to compress these features into 2-dimensional space
- We then use K-means clustering to categorize the crypto assets with the T-SNE embedding
- This artificial feature can represent the metadata of each crypto to a certain degree

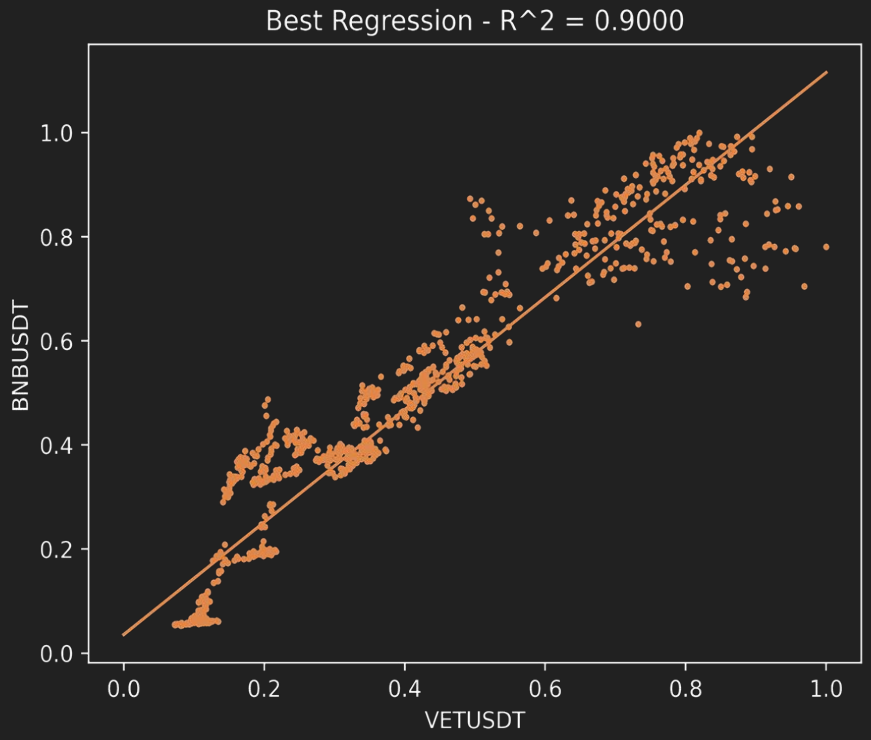


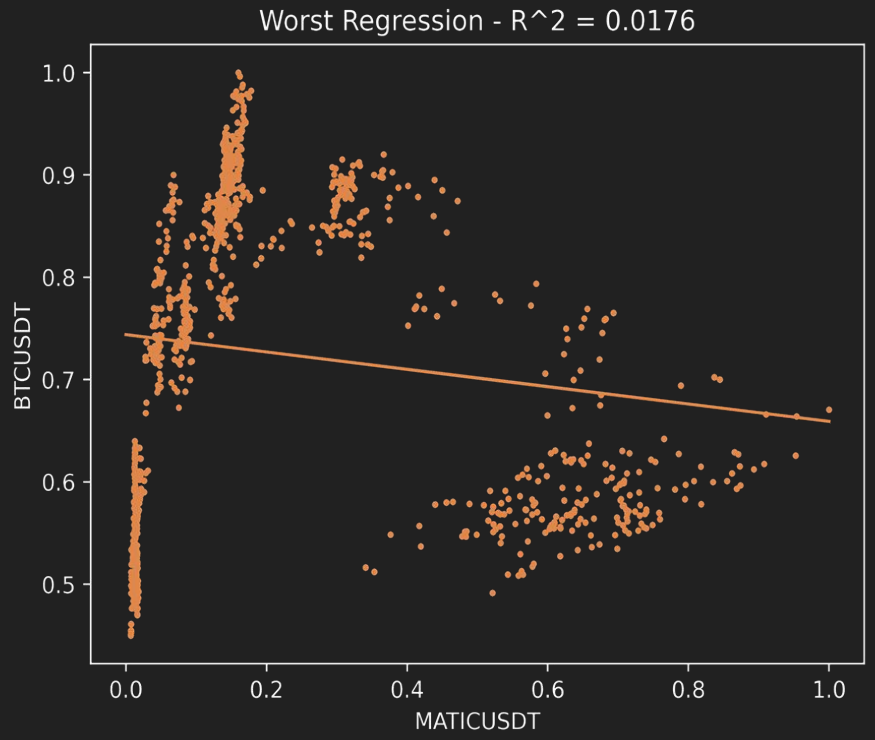
# DBSCAN Clustering

- Due to the volatility of the crypto market, there will inevitably be outliers in the data
- Clustering methods such as k mean clustering sensitive to outliers
- DBSCAN fixes this issue by clustering points based on density
- Outliers have low density, thus they are treated as noise and disregarded



# Regression Similarity



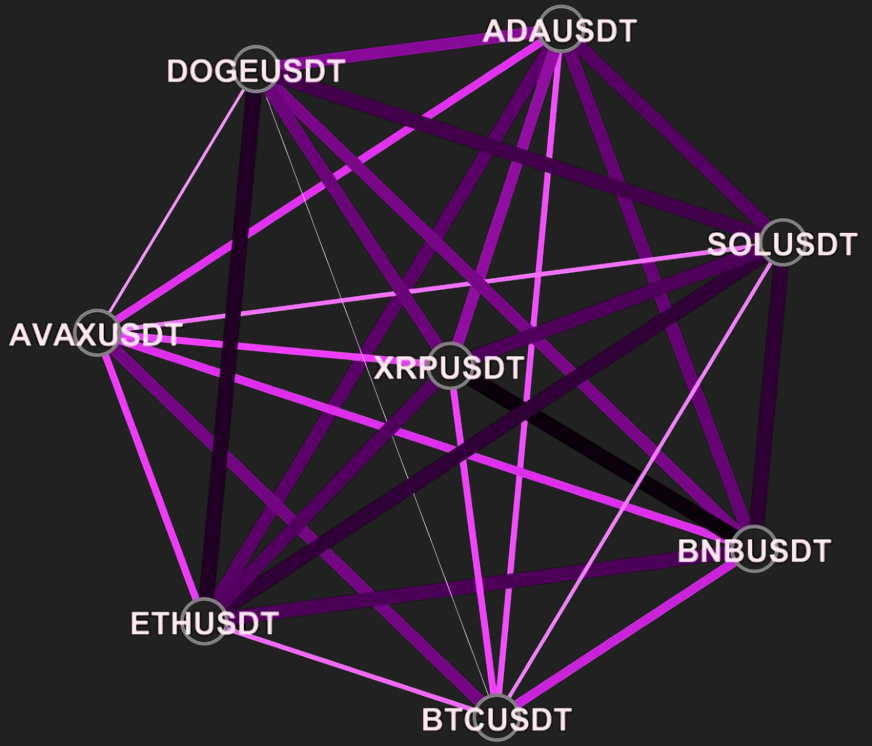



# Regression Similarity (cont'd)

- This is a visualization of 8 different cryptocurrencies

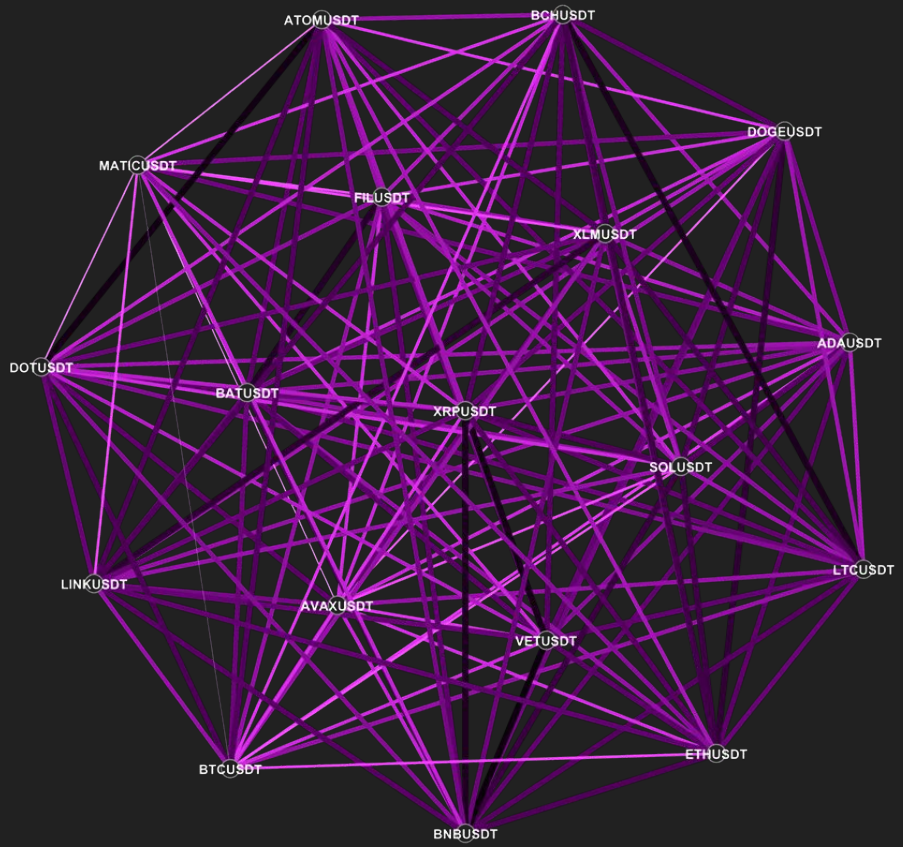
  A linear regression has been computed for every edge, and the edge's weight is the r^2 value for the regression

- Darker edges have a stronger correlation than lighter edges



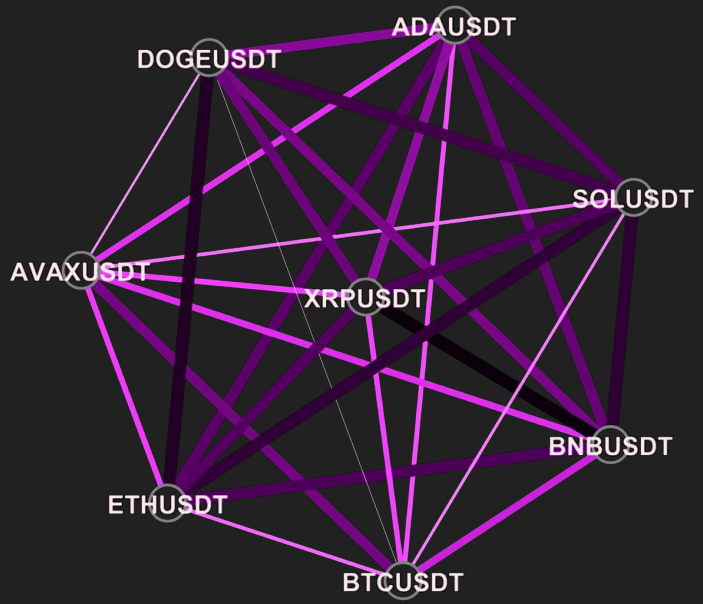
# Regression Similarity (cont'd)

- This is the same visualization but with 18 cryptocurrencies



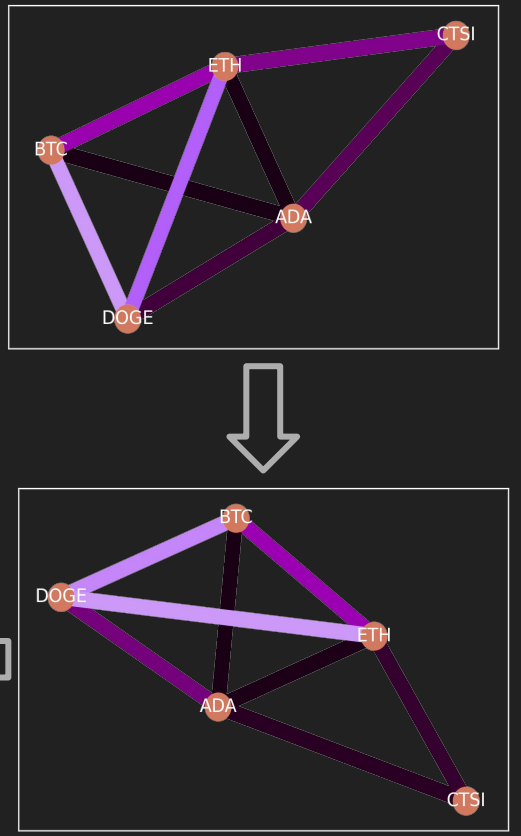
# Regression Similarity (cont'd)

| TIME | XRPUSDT | BNBUSDT | TIME | XRPUSDT | BNBUSDT |
|---|---|---|---|---|---|
| 2021-06-13 00:00:00 | sell | sell | 2021-06-14 20:00:00 | buy | buy |
| 2021-06-13 04:00:00 | sell | sell | 2021-06-15 00:00:00 | buy | buy |
| 2021-06-13 08:00:00 | sell | sell | 2021-06-15 04:00:00 | buy | buy |
| 2021-06-13 12:00:00 | sell | sell | 2021-06-15 08:00:00 | buy | buy |
| 2021-06-13 16:00:00 | buy | buy | 2021-06-15 12:00:00 | buy | buy |
| 2021-06-13 20:00:00 | buy | buy | 2021-06-15 16:00:00 | buy | buy |
| 2021-06-14 00:00:00 | buy | buy | 2021-06-15 20:00:00 | buy | buy |
| 2021-06-14 04:00:00 | buy | buy | 2021-06-16 00:00:00 | buy | buy |
| 2021-06-14 08:00:00 | buy | buy | 2021-06-16 04:00:00 | buy | buy |
| 2021-06-14 12:00:00 | buy | buy | 2021-06-16 08:00:00 | sell | sell |
| 2021-06-14 16:00:00 | buy | buy | 2021-06-16 12:00:00 | sell | sell |



# Correlation Prediction

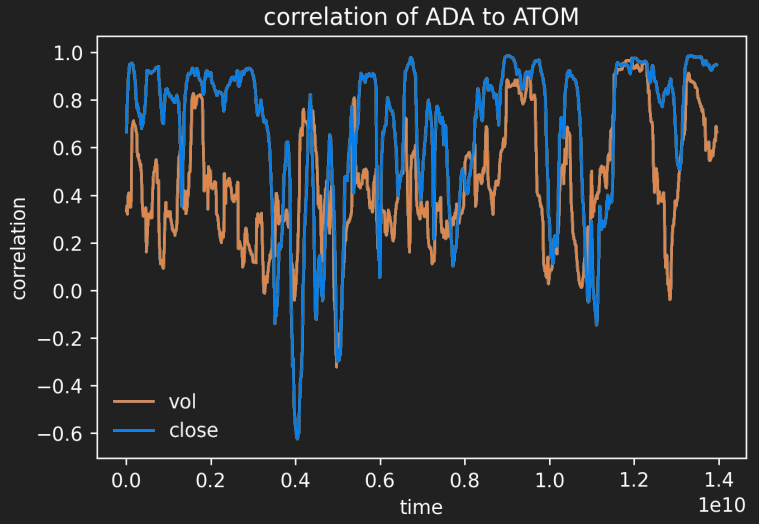
- Graph
  - Weight: Correlation
  - Nodes: Crypto assets
- Time series data prediction
  - Volume correlation prediction
  - Price correlation prediction
- Multi-head self-attention with CNN



# Challenges

# Challenges

- The volatile nature of the crypto market
- Getting historical data from Twitter
  - Twitter API has a 7 day limit on tweet fetches
- Various cryptocurrencies are relatively new
  - No standard start time
  - Don’t have enough information to backtest
- Real-time prediction is more complex than initially thought
  - All predictions will be made at runtime
- Unable to get the latest information in the first place
  - Crypto project teams usually post information on Twitter and discord



# Summary

# Summary

- As of now, we are able to construct a comprehensive graph:
  - Nodes are unique cryptocurrencies
  - Edges hold values for each similarity metric
    - T-SNE
    - k-means
    - DBSCAN
    - Regression similarity
    - Cosine similarity

- Sentiment analysis is also complete
  - Using BERT, we can achieve > 90% accuracy on determining bullish/neutral/bearish sentiment

# Work Distribution

# Work Distribution

- Sean Hung
  - Data scraping from twitter, reddit
  - Price and volume data from Binance API
  - Sentiment analysis
  - Implementing technical analysis indicator
- Steven Wang
  - Data scraping from CoinMarketCap
  - Metadata clustering (T-SNE, k-means)
  - Correlation graph prediction
- Matt Uryga
  - Regression similarity
  - Real-time price fetching (not used)
  - Graph visualizations
  - Implemented graph and crypto object framework
- Jonathan Li
  - DBSCAN
  - Cosine Similarity
  - Data Clustering and visualization

# Looking Ahead

# Looking Ahead

- We need to implement graph similarity algorithms to determine which cryptocurrencies are similar within the constructed graph

- This, combined with the sentiment analysis will be the backbone of the aggregate similarity algorithm that will output the final signal/price target for each cryptocurrency

- Evaluating the reliability of different metrics