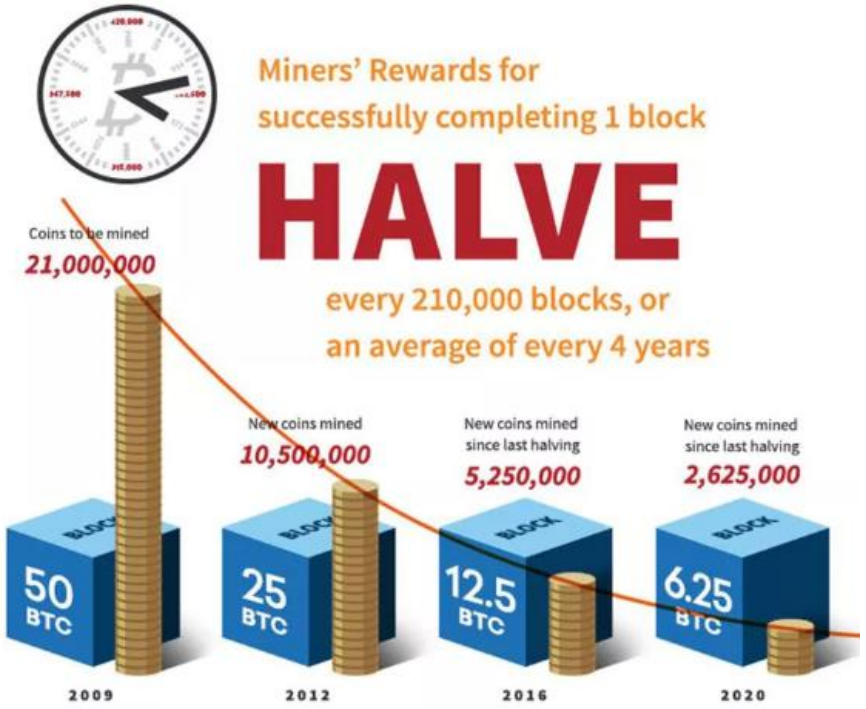
DATA MINING AND PREDICTIVE MODELING FOR CRYPTO-CURRENCIES

LAUREN DEANS, JAMES BEEDLE, THEO SHIN, INDERPAL DHILLON

INTRODUCTION: BITCOIN MINING

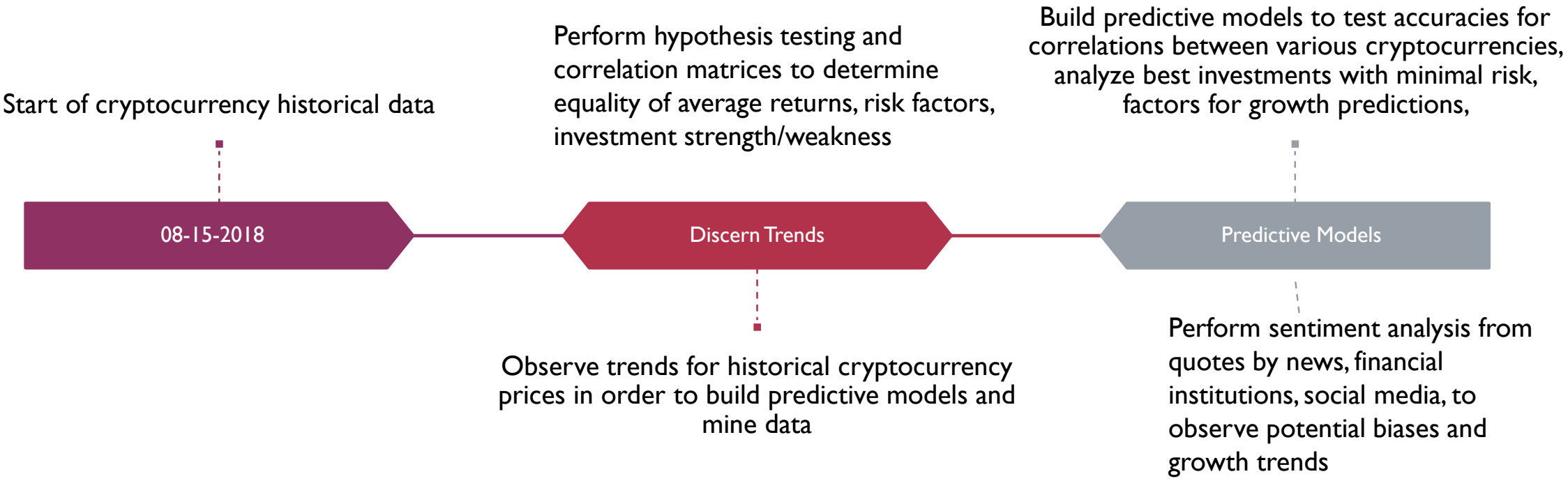
KEY TAKEAWAYS

- By mining, you can earn cryptocurrency without having to put down money for it.
- Bitcoin is mined in units called "blocks."
- Double spending means, as the name suggests, that a Bitcoin user is illicitly spending the same money twice.
- You need either a GPU (graphics processing unit) miner or an application-specific integrated circuit (ASIC) miner.
- Mining rewards are paid to the miner who discovers a solution to the puzzle first, and the probability that a participant will be the one to discover the solution is equal to the portion of the total mining power on the network.



Source: Investopedia

PROJECT DESCRIPTION

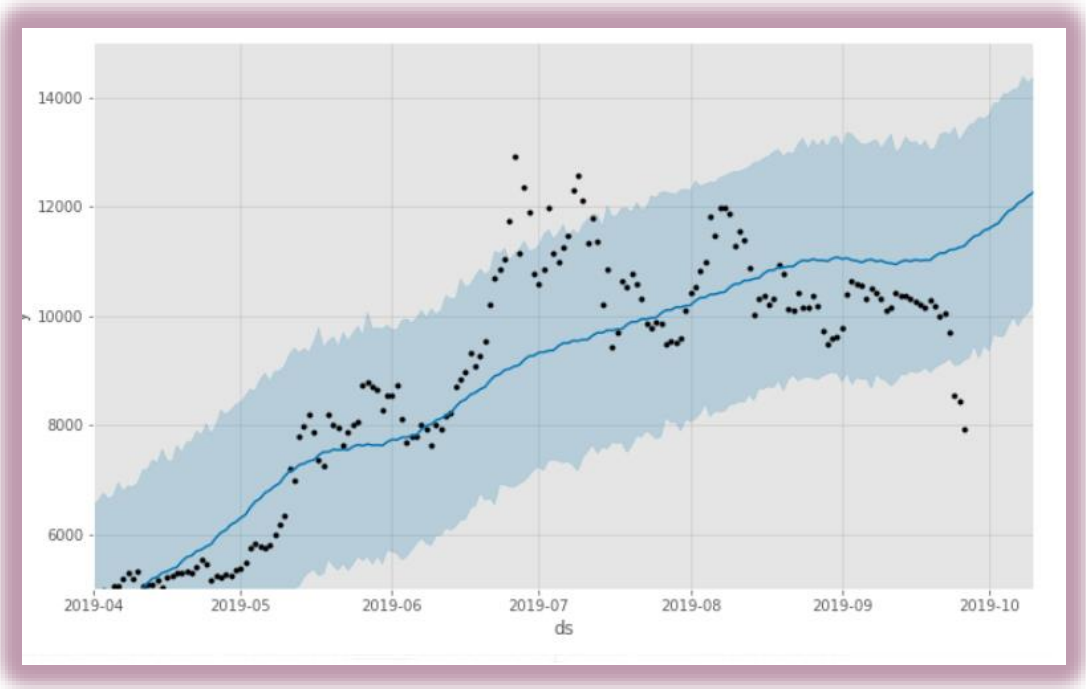


PROJECT DESCRIPTION (CON'T)

- • As the timeline suggests, we are focused on using historical crypto-currency quotes to discover real-time and future trends by employing predictive models (i.e. multilinear regression and classification with sentiment analysis). Questions and deeper meanings we would like to uncover include which currency provides the best average returns for investments, are trends consistent amongst various groups/countries, unique correlations which can potentially be linked to predictions, can we accurately predict future trends? Similarly, does a USD dollar increase in BitCoin lead to a similar increase across basket of currency indices.

PRIOR WORK

FORECAST MODEL



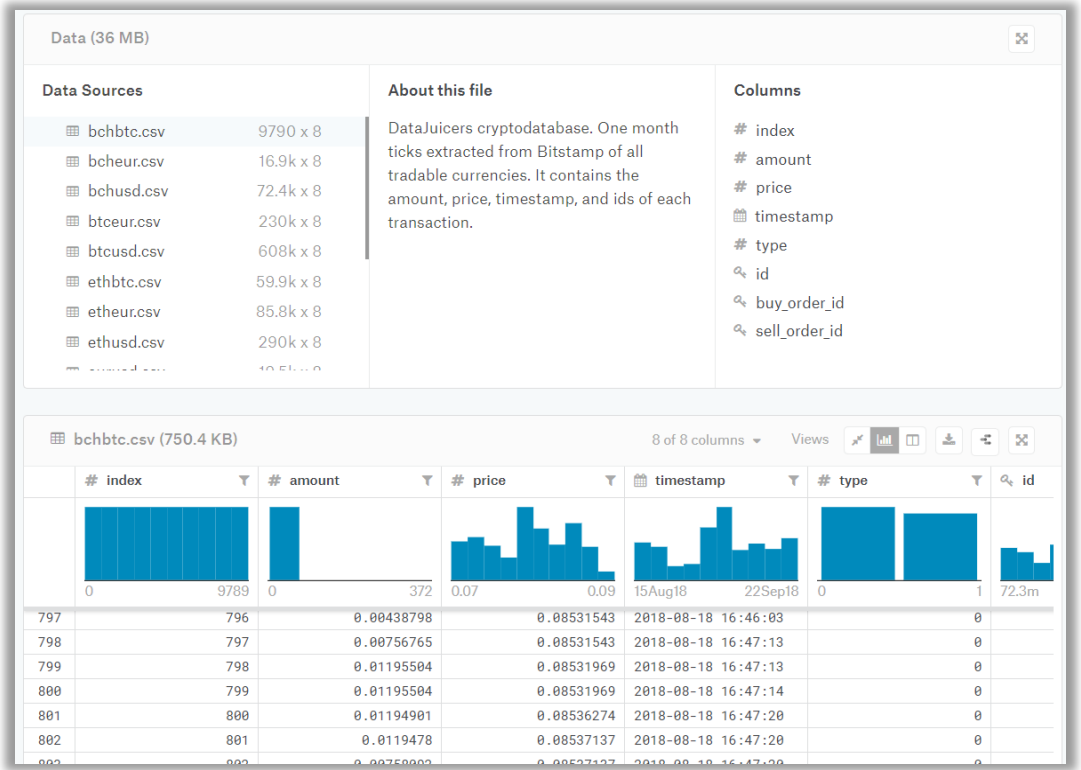
https://towardsdatascience.com

- From Kaggle .csv download @ https://www.kaggle.com/kerneler/starter-ticks-bitcoin-8fbbc257-6. Previous exploratory analysis by combining data and preliminary analysis using a correlation and scatter matrix.
- From https://towardsdatascience.com, forecasted future prices using Time Series models. Author concludes "it is close to impossible to predict the future of Bitcoin, but with **machine learning**, (by fitting and training models) we can understand where it might go with a high degree of confidence."

DATASETS: CRYPTO-CURRENCY

- Historic Crypto Ticks: Bitcoin, Ethereum, Litecoin, Ripple:

https://www.kaggle.com/albala/ticks-bitcoin-ethereumlitecoin-ripple



- Cryptocurrency Market Data:

Historical Cryptocurrency Prices For ALL Tokens!

https://www.kaggle.com/jessevent/all-crypto-currencies

Summary

> Observations: 758,534
> Variables: 13
> Crypto Tokens: 1,584
> Start Date: 28/04/2017
> End Date: 21/05/2018

Description

All historic open, high, low, close, trading volume and market cap info for all cryptocurrencies.

I've had to go over the code with a fine tooth comb to get it compatible with CRAN so there have been significant enhancements to how some of the field conversions have been undertaken and the data being cleaned. This should eliminate a few issues around number formatting or unexpected handling of scientific notations.

Data Structure

Observations: 649,051
Variables: 13
$ slug

(Both datasets hosted on Group GitHub Page)

- Data Cleaning: We want to aggregate cryptocurrency data for several months at various times. Remove any unnecessary or redundant information along with filling in any incomplete columns using appropriate means.
- Data Pre-Processing: reorganize the integration of both datasets to simplify correlation analysis and group the different cryptocurrencies by most effective trends. Employ visualizations and preliminary analysis to determine potential future growth. Utilize effective visualization tools typically used with stock market analysis (i.e. candlestick chart, correlation matrices).
- Integrate Data: Combine data from downloaded datasets and create uniform patterns for usability ease.

PROPOSED WORK:

What do we need to do?

LIST OF INTENDED TOOLS

Programs

- Python
- SQL
- Price Converter (Excel)

Tools

- Sklearn
- Seaborn
- nltk
- Wordcloud
- Candlestick
- Matplotlib
- Statistical methods: Naïve Bayes, supervised learning
- Pandas

EVALUATION

- Given the dataset start date in 2015, we can compare our forecasted predictions to actual patterns
- Determine which hypotheses are accurate based on trends
- Use forecasting models to observe correlations between predictions and real-time analysis
- Compare our prediction models to relevant data science accuracy measures (TBD)