# CRYPTEX:
# fine-grained CRYPTocurrency datasets EXploration


**Lucas Raicu**[1] **Stefan Donisa**[2] **Lucas Ciobanu**[3] **Lan Nguyen**[4] **Ioan Raicu**[4]

266520@glenbrook225.org stefandonisa08@gmail.com lucasciobanu@gmail.com lnguyen18@hawk.iit.edu iraicu@cs.iit.edu

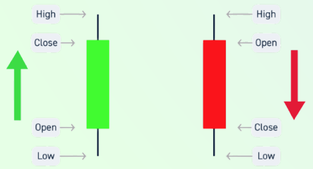
[1]Glenbrook South High School [2]John Hersey High School [3]Lisle High School [4]Computer Science Department, Illinois Institute of Technology


## ABSTRACT


Bitcoin is a decentralized digital currency and the first-ever cryptocurrency created in 2009. Analyzing cryptocurrency pricing data is going to be of high interest to the Fintech community. Getting access to this price data is readily available through both graphs and numerical datasets. The sources that are available make it incredibly difficult to get OHLC and volume data at fine-grained resolution below 1-min intervals. When simulating financial models, fine-grained datasets can significantly impact model effectiveness. Our research aimed to extract cryptocurrency time-series data from the largest cryptocurrency exchange in the world, Binance. However, obtaining high-quality, one-second granularity data from exchange APIs and even online websites proved difficult, unintuitive, impossible, or expensive. This challenge inspired us to create an efficient Python-based framework that extracts the transaction history for 153 cryptocurrency trading pairs from Binance.us since September 2019. This data is then cleaned and summarized into a variety of sub-datasets ranging from yearly to one-second granularity candlesticks. To enable ease of data-sharing, we published a sample (4-year BTC-USDT trading pair data) dataset on Kaggle and the entirety of the 261GB datasets in CSV formatted files on a publicly accessible web server that updates nightly.


## DATA REPRESENTATION

- Time-series data for cryptocurrencies is a representation of events or measurement (price over time)
- Candlestick [1] data and plots is widely used in finance industry to condense time series data into only the key values




The Basics of Candlesticks


Example Candlestick Chart

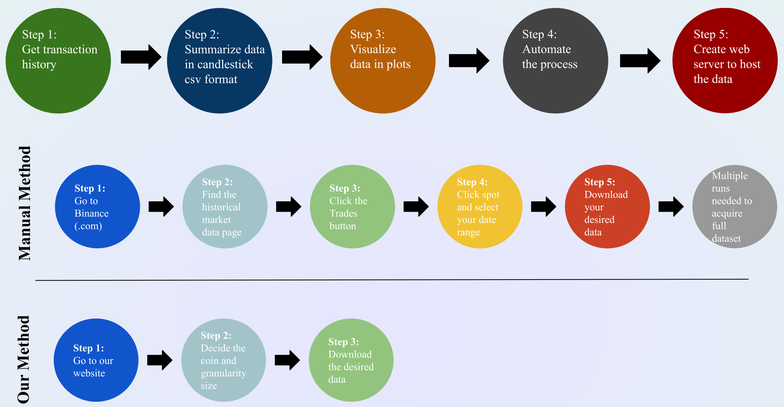
## MAKING A CASE FOR FINE-GRAINED DATA

Smaller candlestick data granularity means more opportunities to buy and sell, leading to greater potential profit. For BTCUSDT in 2021, investing one bitcoin on 1/1/21 showed the highest profit with the 1-second granularity.

| Period | Potential Profit | Opportunities |
|---|---|---|
| 1-sec | $170M | 28.9M |
| 1-min | $37M | 524K |
| 1-hour | $5.4M | 8.7K |
| 1-day | $1.1M | 365 |
| 1-week | $431K | 52 |

## PROPOSED SOLUTION

Our aim was to make fine-grain candlestick data readily accessible and easy to download. To achieve this goal, we implemented the following pipeline of events:




## RESULTS

The below table compares the performance of extracting fine-grain data with existing methods one could have used to extract data from Binance, with the Python API approach being ours.

| Source | Approach | # of files | Time (s) | Info |
|---|---|---|---|---|
| binance.com | Manual | 48 | 3,600 | User-friendly, long run time |
| binance.us | Manual | N/A | N/A | Not possible |
| binance.com | API | N/A | N/A | Not possible without VPN |
| binance.us | API | 1 | 660 | Automated, requires programming |
| crypto.cs.iit.edu | XStore | 1 | 70 | Low latency, slower than wget |
| crypto.cs.iit.edu | wget | 1 | 51 | Fastest, no date range |

### Result Statement:

- Finer-grained data provides the necessary precision to build models for high-frequency trading

### Automation for Transaction History Extraction:

- Allows user to save time, as manually updating data for all 153 coins supported by USDT (currently 261 GB) would be time-consuming
- Top 10 Cryptocurrency Trading Pairs (by volume) published online: BTCUSDT [2], ETHUSDT, USDCUSTD, XRPUSDT, BNBUSDT, BUSDUSDT, LTCUSTD, SOLUSDT, DOGEUSDT, CRVUSDT

## CHALLENGES

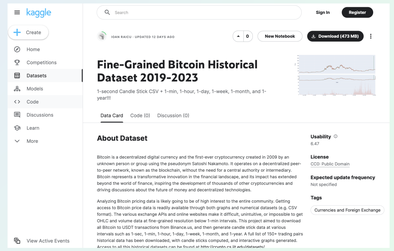
We have identified five main challenges in developing an automated Python program to enable easy access to cryptocurrency candlestick data:

- Jupyter Notebook environment compatability with large files/datasets
- Binance.com vs Binance.us
- Limited requests bandwidth to Binance
- Unsynced timezones between datasets and long update wait times
- Limited GitHub Pages storage (5 GB)

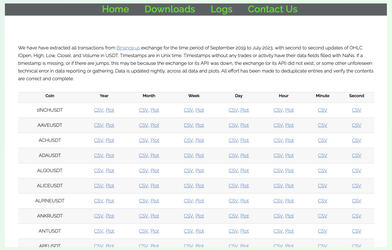
## CONCLUSIONS AND FUTURE WORK

### Contributions

- Developed different granularity candle stick datasets (1-sec, 1-min, 1-hour, 1-day, 1-week, 1-month, and 1-year) that are updated daily
- Made datasets from 153 trading pairs since 09/2019 available to the public through a Kaggle page and a website hosted in Mystic [3]




https://www.kaggle.com/iraicu/datasets




http://crypto.cs.iit.edu/datasets/

### Future Work

- Collect global transaction data from Binance.com
- Conduct backtesting analysis for trading strategies using the datasets collected
- Explore data correlation between trading pairs to improve backtesting
- Explore sentiment analysis to augment time-series pricing information

## REFERENCES


[1] B. R. Marshall, M. R. Young, and L. C. Rose, "Candlestick technical trading strategies: Can they create value for investors?" *Journal of Banking & Finance*, vol. 30, no. 8, pp. 2303–2323, 2006.

[2] S. Nakamoto, "Bitcoin whitepaper," *URL: https://bitcoin. org/bitcoin. pdf-(: 17.07. 2019)*, 2008.

[3] A. Orhean, A Ballmer, T Koehring, *et al.*, "Mystic: Programmable systems research testbed to explore a stack-wide adaptive system fabric," in *8th Greater Chicago Area Systems Research Workshop (GCASR)*, 2019.