Multi Data Source Stock Market Prediction

Yuhan Su, Ruichuang Cao, Wei Xu

Tsinghua University

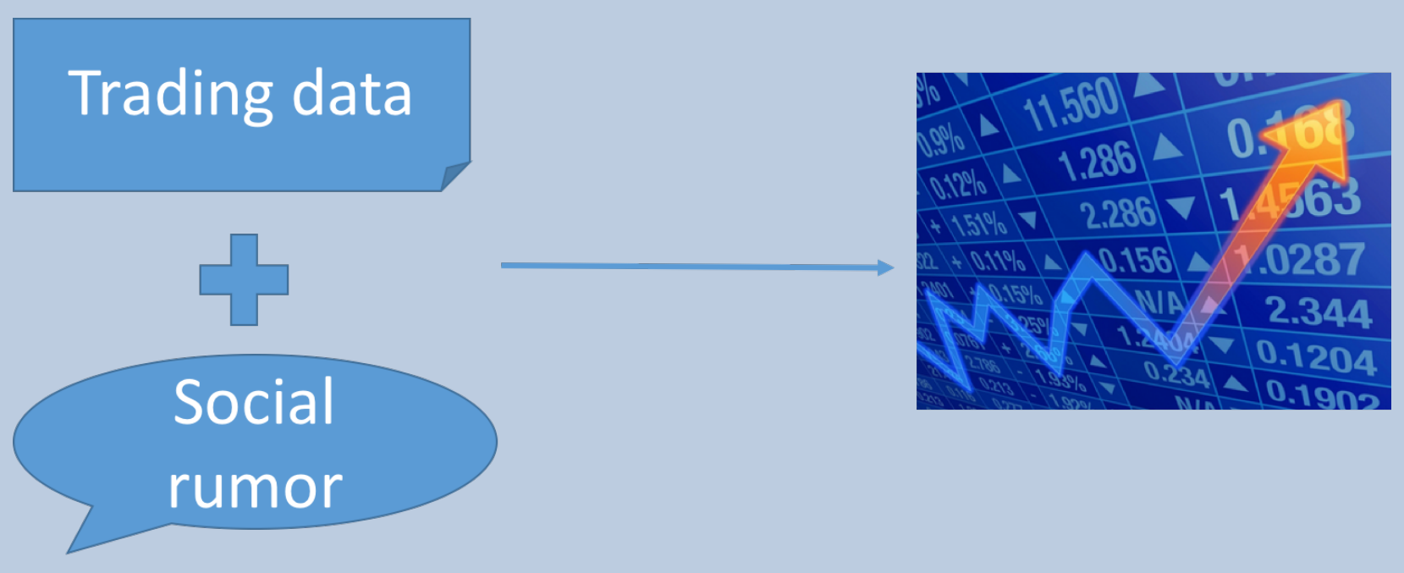
Introduction

- For hundreds of years, everyone dreams to predict stock price changes. Numerous studies have shown that stock could be predicted to some degree.
- China stock market is influenced by rumors on social media.

- We design and implement a real-time data stream stock prediction system based on IBM SuperVessel Cloud.
- With the system, we can perform stock prediction based on multiple data sources, including trading data and social media rumors.

Contribution

- Build a system to process real-time data stream from multiple sources.
- Use multi data source, including trading data and social rumors, to predict stock market in China.
- Leverage the state-of-the-art cloud technology to provide a scalable system.
- Provide an intuitive web UI to let users edit and analyze related information.



Data Source

Currently we have the following datasets and we are adding more.

- SSE50 index. We use its daily closing price. This is also our predicting target.
- Sina Stock Forum. We use posts and comments from Sina Guba.
- Financial news. These news are collected by Tushare website.
- NASQAF index. We use its daily closing price.
- RMB Exchange rate. We use its daily value.

Sina Stock Forum

SSE50 Index (Prediction Target)

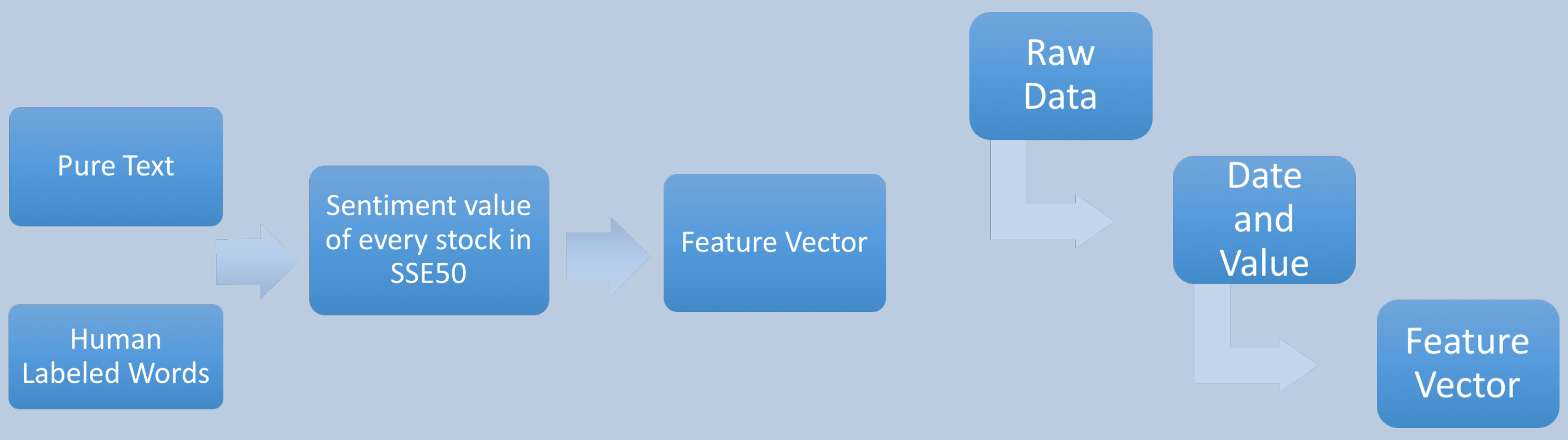
NASDAQ Index

RMB Exchange Rate

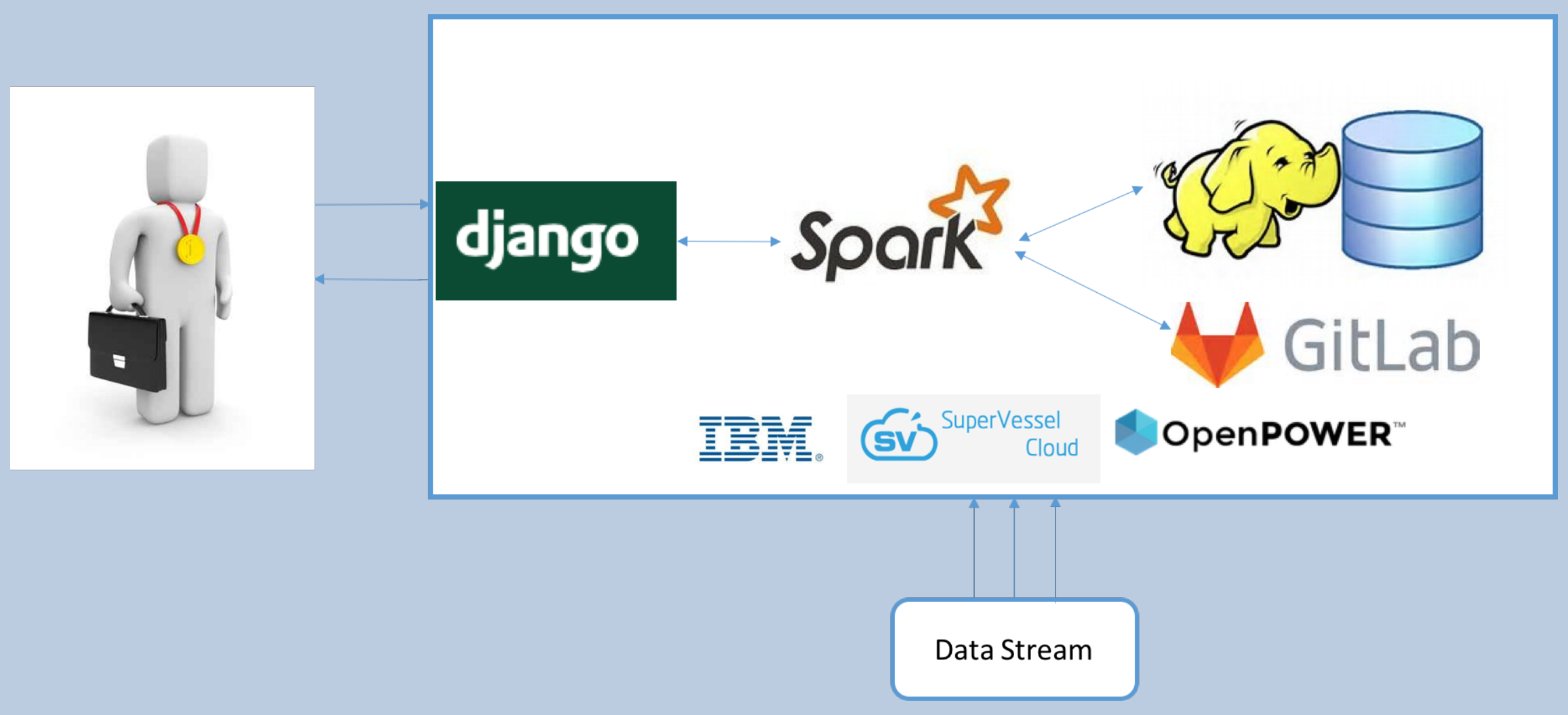
Financial News in China

Data Pre-Processing

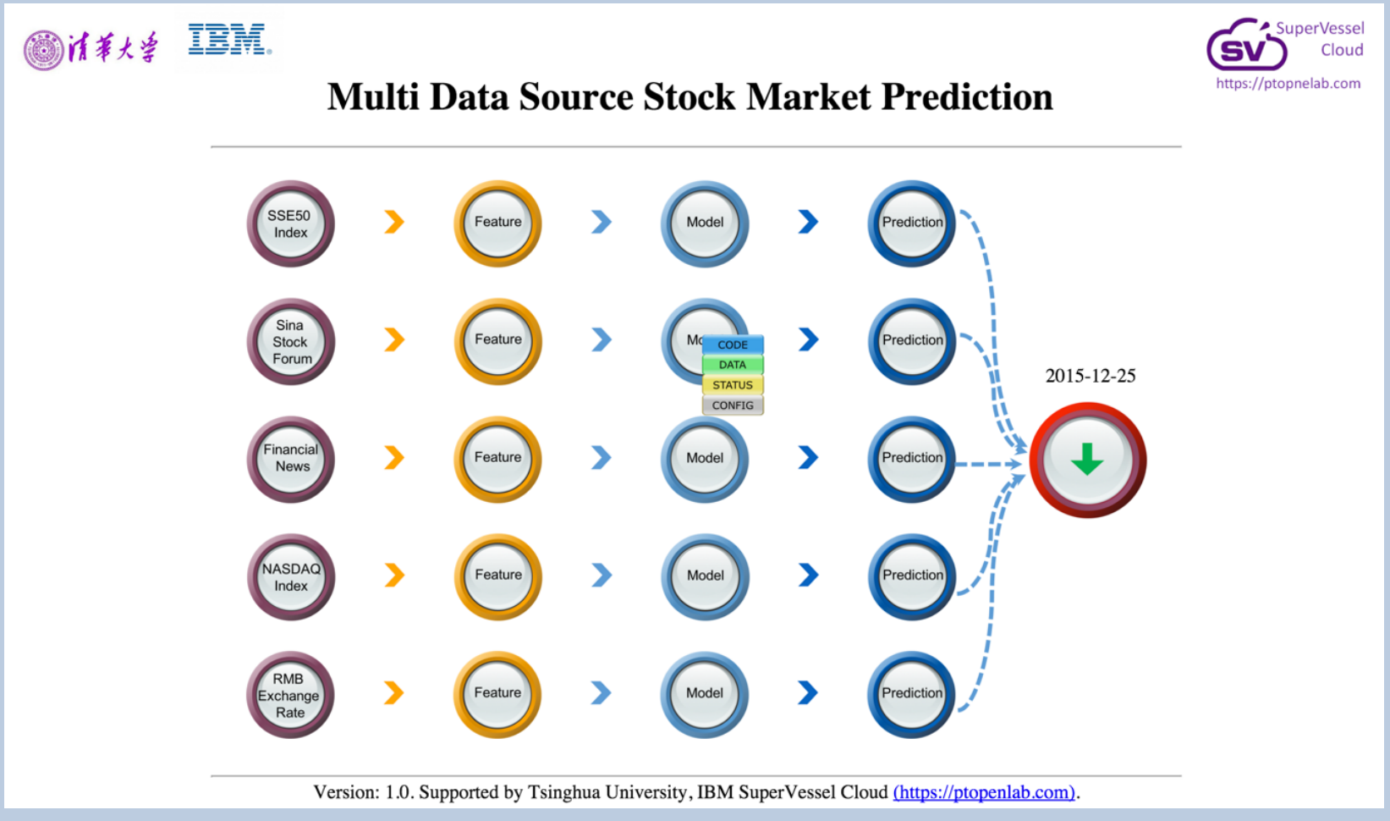
- We perform multi-step data pre-processing on cloud-based streaming system. For data like rumors and news, we extract their sentiment as features, just like the left figure shows. For trading history data, we use their values within a certain window, as the right figure shows.



System Architecture



- Components in our system include Spark, HDFS, Gitlab, Django. They all run efficiently on SuperVessel Cloud.
- Our system integrates the acquisition, preprocessing and learning of data source streams, and shows the final prediction result. Users can monitor the whole process all from a single web UI. They can change the model parameters, modify the codes, and check the running status online.



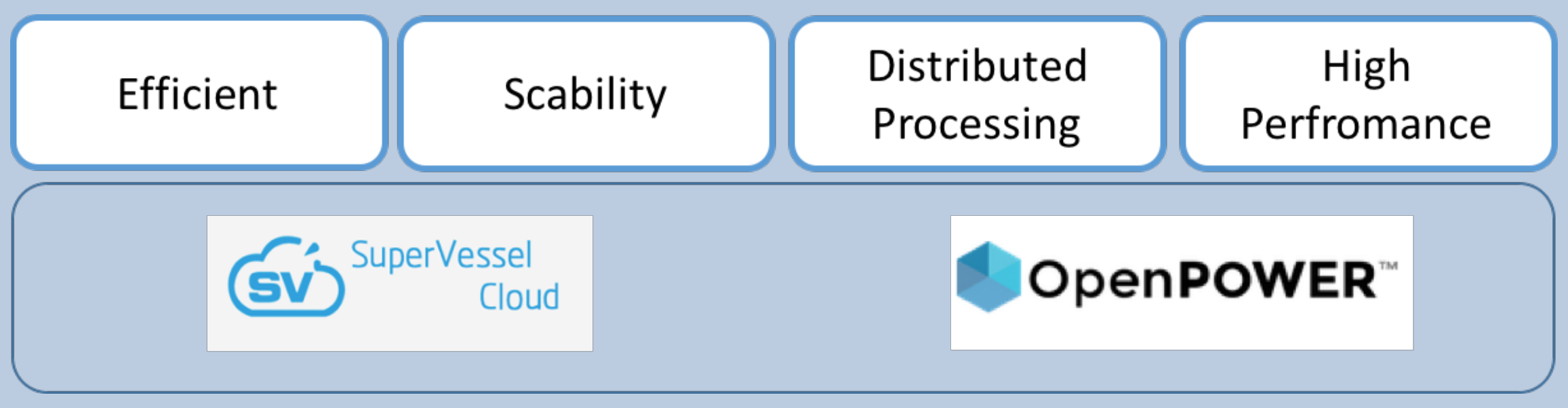
Multi Data Source Stock Market Prediction

Version: 1.0. Supported by Tsinghua University, IBM SuperVessel Cloud (https://ptopenlab.com).

SuperVessel Cloud

https://ptopnelab.com

- This Project is developed on SuperVessel Cloud, which is based on OpenPOWER technology and provides the high efficiency cognitive computing infrastructure for frontier science with high performance heterogenous platform (GPU/FPGA).
- SuperVessel consists of three parts: basic cognitive cloud service, cognitive computing service platform, and application acceleration store for new technology sharing.
- SuperVessel provides us with a scalable and easy-to-maintain cloud infrastructure to build our systems on.



Conclusion

- Using multiple data sources improves stock prediction.
- Cloud technology provides us with high efficiency cognitive computing infrastructure, makes it simple for users to analyze in an intuitive web UI.

Acknowledgments

- Thank IBM China Research Lab for providing computing resources on SuperVessel Cloud.
- Thank Tushare website for providing datasets.

Contact Information

- Web: http://iiis.tsinghua.edu.cn https://ptopenlab.com
- Email: Yuhan Su syhmartin@yeah.net
  Ruichuang Cao create0818@163.com
  Prof. Wei Xu wei.xu.0@gmail.com