THE WEB CONFERENCE
April 20-24
TAIPEI
2020

Domain Adaptive Multi-Modality Neural Attention Network for Financial Forecasting

Presenter: Dawei Zhou

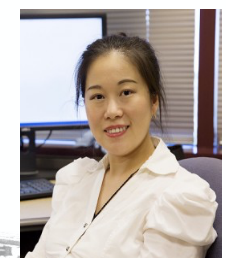
Contact: dzhou21@Illinois.edu

Dawei Zhou*
(UIUC)

Lecheng Zheng
(UIUC)

Jianbo Li
(Three Bridges Capital)

Yada Zhu
(IBM)



Jingrui He
(UIUC)

ILLINOIS

Stock Market

- **Fact 1: Stock market is the aggregation of buyers and sellers (a loose network of economic transactions) of stocks.**

Stock Market

- **Fact 2: The asset prices represent the summarized expectation of stocks from every players in the stock market.**
  - Any asset prices P_t are expectations of the future.
  - **Efficient Market Theory**[1] states a hypothesis in financial economics that the asset prices P_t reflect all available information.

Stock prices at t

Stock prices at $t + 1$

$$\log P_t = \log M + \mathbb{E}_{P_{t+1} \sim S}(P_{t+1})$$

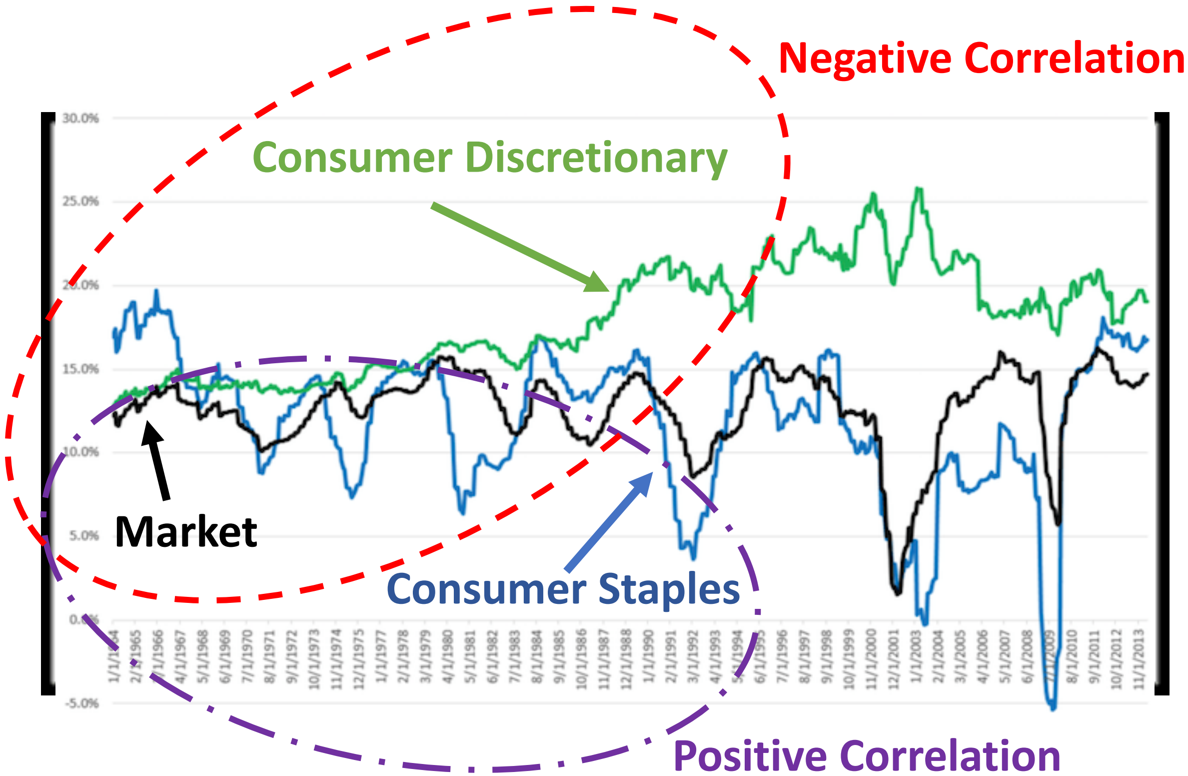
Stochastic discount factor

Expectation over $|S|$ information sources

[1] Malkiel, Burton G., and Eugene F. Fama. "Efficient capital markets: A review of theory and empirical work." The journal of Finance 25.2 (1970): 383-417.

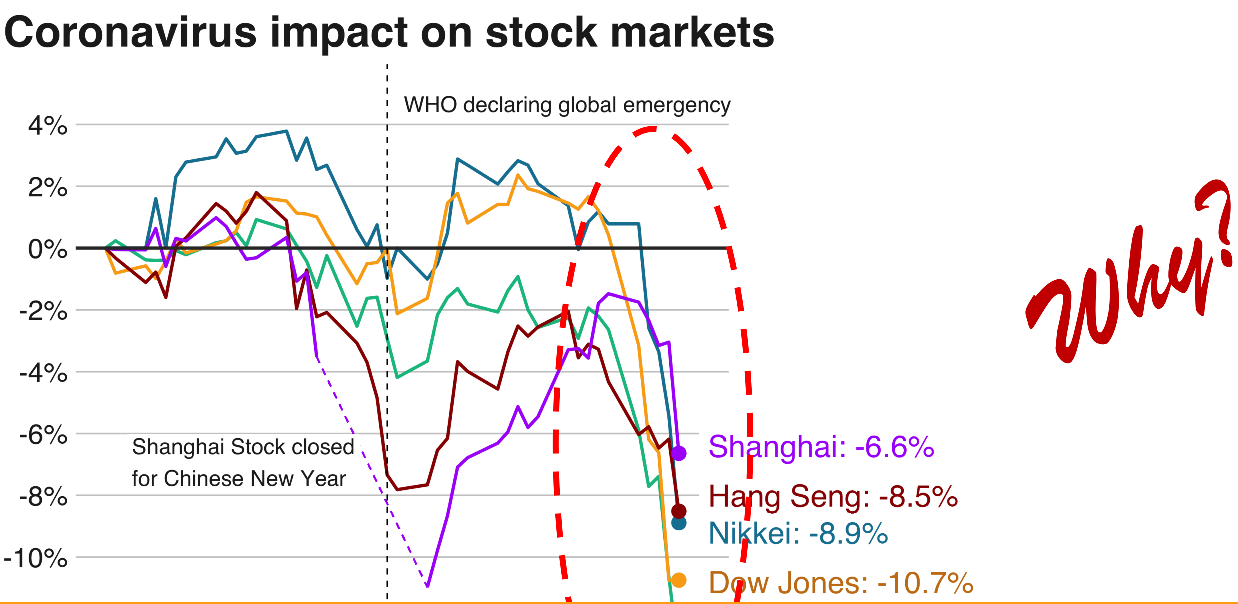
Stock Market

- **Fact 3: Stock in different domains exhibit multi-modal behaviors.**



Financial Forecasting

- Can we forecast the “circuit breaker” due to COVID-19?



The environment is chaotic. Macro-economic forecasts are normally too inexact to have. *--Lars Tvede*

Challenges

- **Challenge 1: Data Heterogeneity**
  - Q1: How to capture and incorporate various key factors into account which might affect stock prices?

Efficient Market Theory

$$\log P_t = \log M + \mathbb{E}_{P_{t+1}\sim s}(P_{t+1})$$

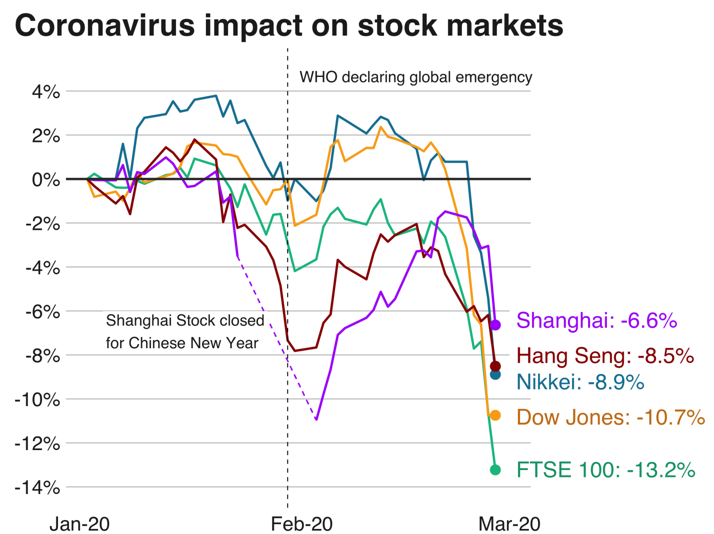
Relevant stocks

Social media

News

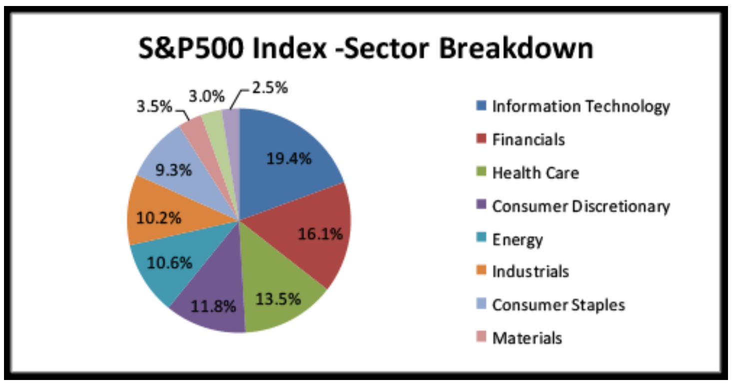
Google Trend

Multiple Sources

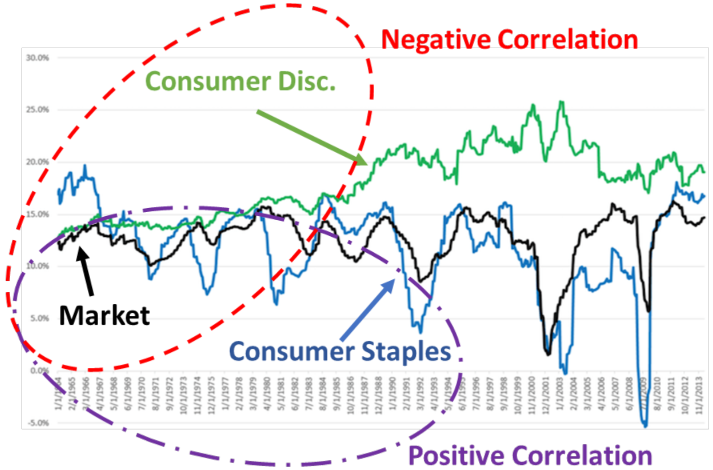


Challenges

- **Challenge 2: Task Heterogeneity**
  - Q2: How can we leverage the potentially noisy input data from various domains to construct models with a satisfactory performance?

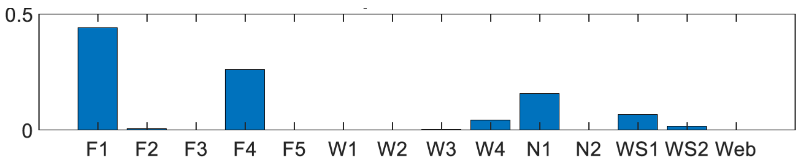


Multiple Behaviors

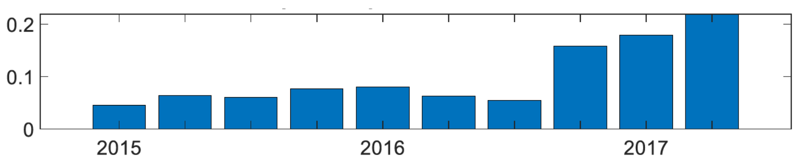


Challenges

- **Challenge 3: Data Interpretability**
  - Q3: How do we interpret the output results to the analysts by providing the relevant clues?



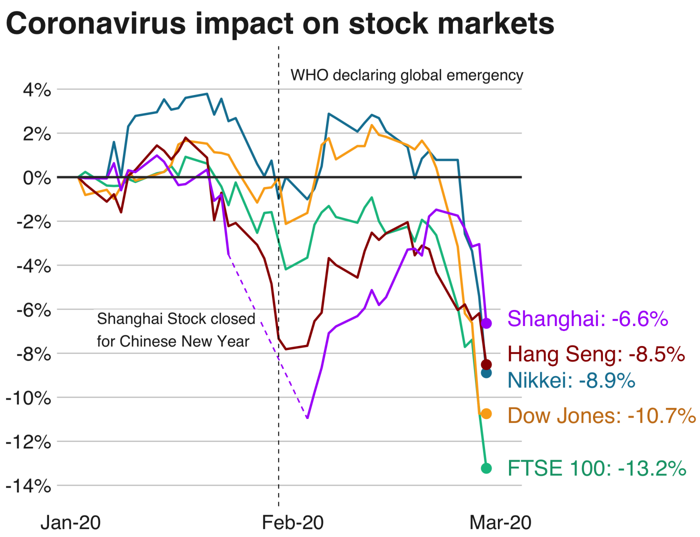
Clue 1: Important Factors



Clue 2: Important Timestamps

⋮

Interpretation Requirement

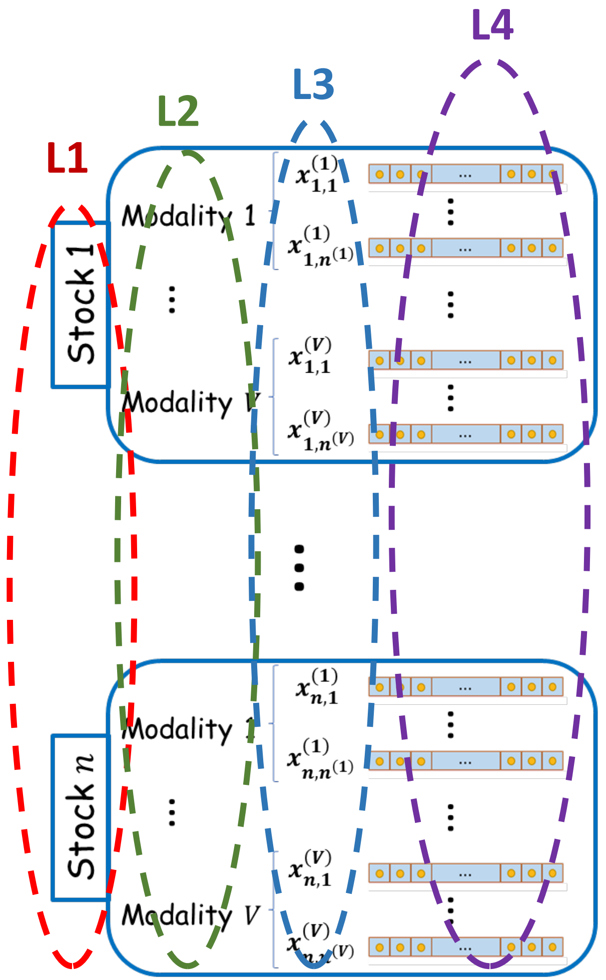


Outline

- Background
- Problem Definition
- Proposed *Dandelion* Framework
- Experiments
- Conclusion

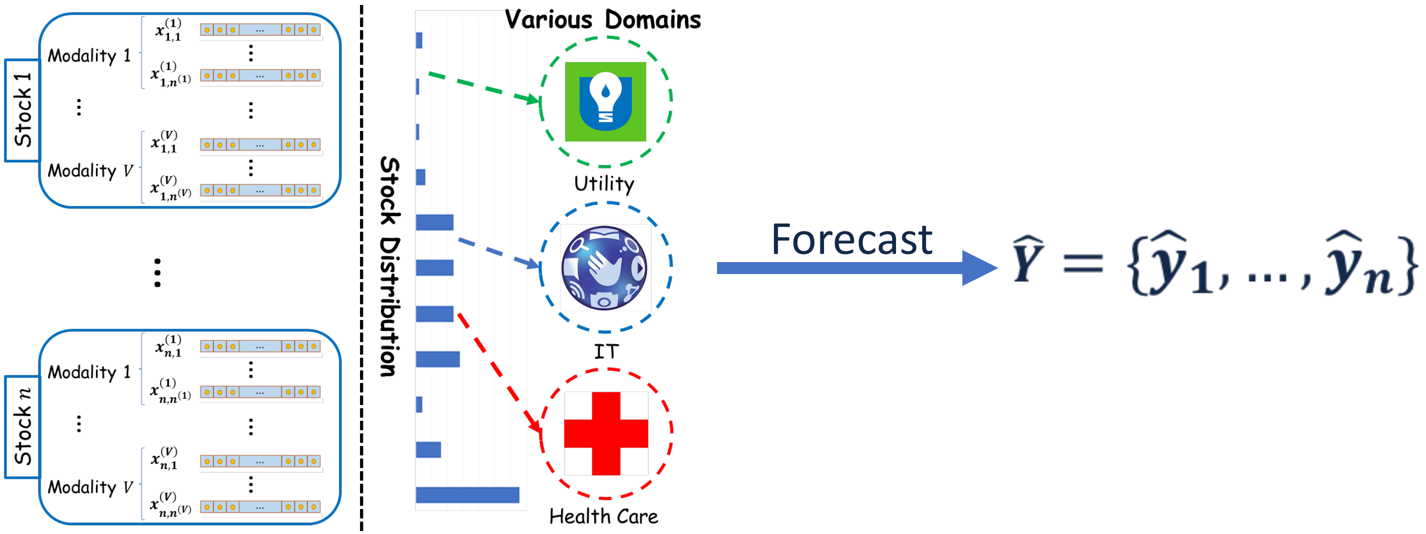
Problem Definition

- **Multi-Modality Multi-Variable Time Series**
  - **L1. Database-level**: The given time series database $\boldsymbol{X} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ consists of n stocks.
  - **L2. Instance-level**: Each observation $\boldsymbol{X}_i \in \boldsymbol{X}$ is composed of m modalities, i.e., $\boldsymbol{X}_i = \{\boldsymbol{X}_i^{(1)}, \boldsymbol{X}_i^{(2)}, \dots \dots, \boldsymbol{X}_i^{(m)}\}$.
  - **L3. Modality-level**: Each modality $\boldsymbol{X}_i^{(v)} \in \boldsymbol{X}_i$ consists of $n^{(v)}$ variables, i.e., $\boldsymbol{X}_i^{(v)} = \{\boldsymbol{x}_{i,1}^{(v)}, \boldsymbol{x}_{i,2}^{(v)}, \dots \dots, \boldsymbol{x}_{i,n^{(v)}}^{(v)}\}$.
  - **L4. Variable-level**: Each variable $x_{i,f}^{(v)} = \{x_{i,f}^{(v)}(1), x_{i,f}^{(v)}(2), \dots \dots, x_{i,f}^{(v)}(T)\}$ is a T length temporal sequence.



Problem Definition

- **Multi-Modality Multi-Task Time Series Forecasting**
  - **Given**: (i) a multi-modality time series $\boldsymbol{X} = \{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$ from time $t = 1$ to $t = T$; (ii) the target signal $\boldsymbol{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\}$ from time $t = 1$ to $t = T$.
  - **Find**: the prediction $\widehat{Y} = \{\widehat{\boldsymbol{y}}_1, \ldots, \widehat{\boldsymbol{y}}_n\}$ from time $t = T + 1$ to $t = T + T'$.

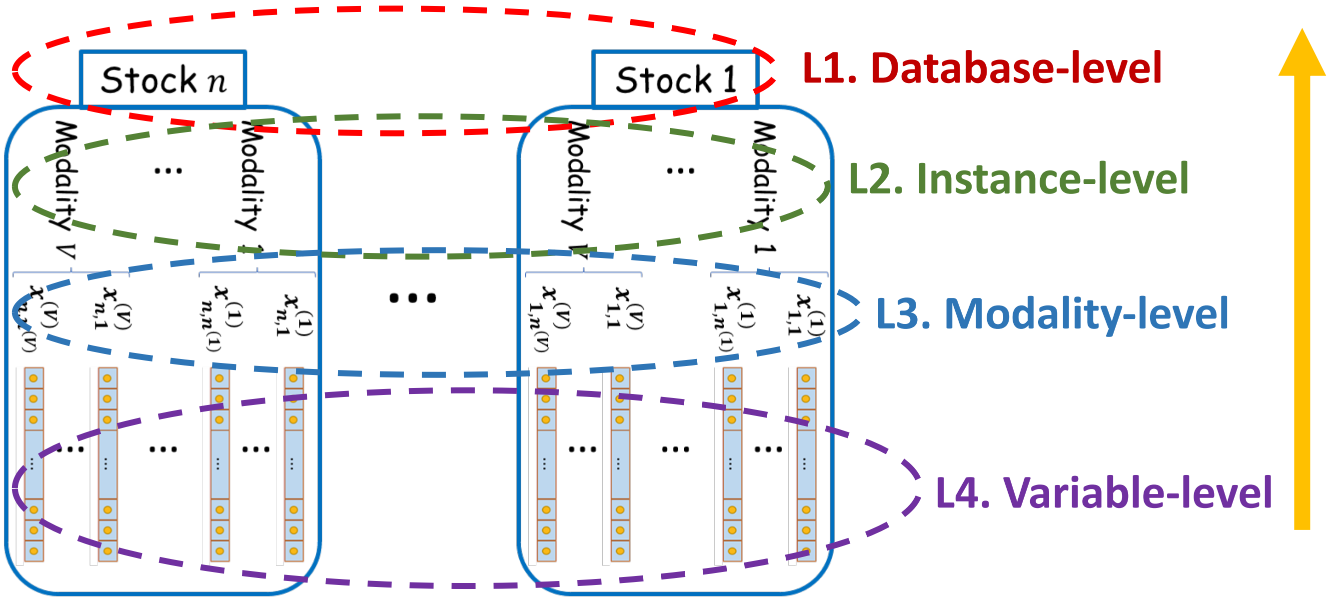


Outline

- ➢Background
- ➢Problem Definition
- ➢Proposed *Dandelion* Framework
- ➢Experiments
- ➢Conclusion

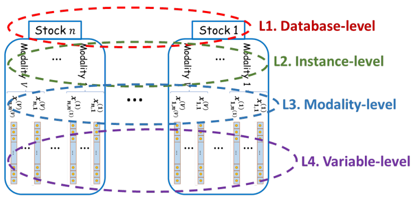
A Generic Framework *Dandelion*

- **A Generic Joint Learning Framework for modeling Multi-Modality Multi-Variable Time Series**

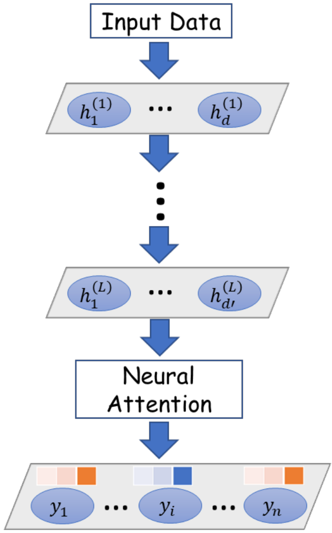


We will present ***Dandelion*** in a bottom-up fashion

Dandelion – Variable-Level

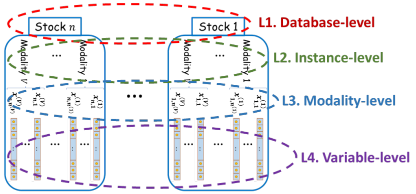


- **Multivariate Forecasting Neural Network with Soft Attention.**
  - **Assumption**[1]: Each observation $\boldsymbol{X_i}$ for factor f at time τ,i.e., $x_{if}(\tau)$, is assumed to have independent effect on $y_{if}(\tau)$.
  - Soft attention mechanism
    - Prediction: $y_{if}(\tau) = \beta_{if}^{(v)}(\tau) h_{if}^{(v)}(\tau)$
    - Hidden layer: $h_{if}^{(v)}(\tau) = \tanh(W_{hf} x_{if}^{(v)}(\tau) + b_h)$
    - Attention: $a_{if}^{(v)}(\tau) = \tanh(W_{df} h_{if}^{(v)}(\tau) + b_a)$
    - Normalized Attention: $\beta_{if}^{(v)}(\tau) = \dfrac{a_{if}^{(v)}(\tau)}{\sum_i^n \sum_v^m \sum_f^{n^v} \sum_t^T a_{if}(\tau)}$



[1] Riemer, Matthew, et al. "Correcting forecasts with multifactor neural attention." International Conference on Machine Learning. 2016.

Dandelion –Modality-Level



- **Learning from multi-modality time series data.**
  - Observations[1]
    - O1: Only a relatively small subset of variables are relevant to making the prediction at a certain timestamp.
    - O2: The different modalities is complementary, whereas the variables within the same modality are redundant.
  - Formulation

$$\mathcal{L}(\tau) = \mathcal{L}_Y(\tau) + \mathcal{L}_s(\tau) + \mathcal{L}_c(\tau)$$

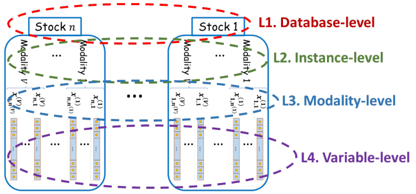
$$= \underbrace{\sum_{i=1}^{n} |y_i(\tau) - \hat{y}_i(\tau)|}_{\mathcal{L}_Y:\text{ prediction loss}} \underbrace{+\gamma \sum_{i=1}^{n}\sum_{v=1}^{m}\sum_{f=1}^{n^{(v)}}\sum_{t=1}^{T} |a_{i,f}^{(v)}(t)|}_{\mathcal{L}_s:\text{ sparse attention regularizer}} \underbrace{+\eta \sum_{i=1}^{n}\sum_{v=1}^{m}\sum_{f=1}^{n^{(v)}} |z_i^{(v)}(\tau) - \beta_{i,f}^{(v)}(\tau)|}_{\mathcal{L}_c:\text{ consensus regularizer}}$$

Address O1

Address O2

[1] Li, Jianboi, Jingrui He, and Yada Zhu. "HiMuV: Hierarchical framework for modeling multi-modality multi-resolution data." 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017.

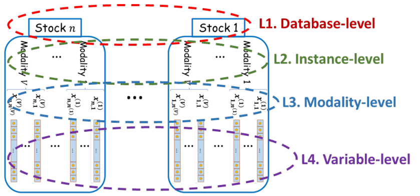
Dandelion – Instance-Level



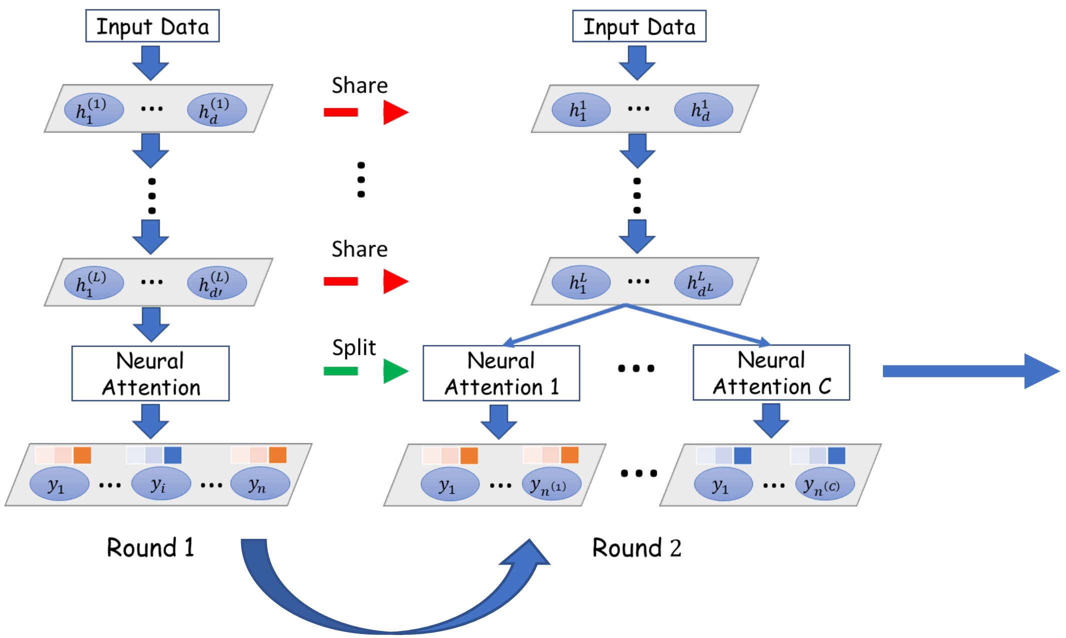
- **Fully-adaptive hierarchical multi-task learning.**
  - **Intuition**: different stocks from the same domain may exhibit similar behaviors.
  - **EX**: most healthcare stocks rely on the news from Food and Drug Administration.
  - **Our Approach**: Explore the domain relatedness via neural network *split* and *widen* procedure[1,2] at each layer l.
    - S1: Group the neurons with similar attention vectors into c clusters by spectral clustering.
    - S2: Split layer l into c branches and back link to layer $l-1$.
    - S3: Initialize each branches by directly cloning the original layer l.

[1] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris. 2017. Fully-Adaptive Feature Sharing in Multi-Task Networks with Applications in Person Attribute Classification. (2017).

Dandelion –Instance-Level



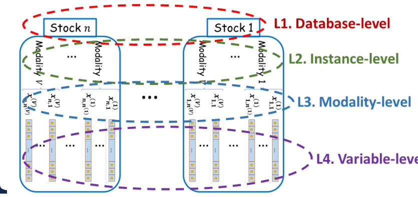
- **Fully-adaptive hierarchical multi-task learning.**



S2&S3: Split and widen

S1: Grouping

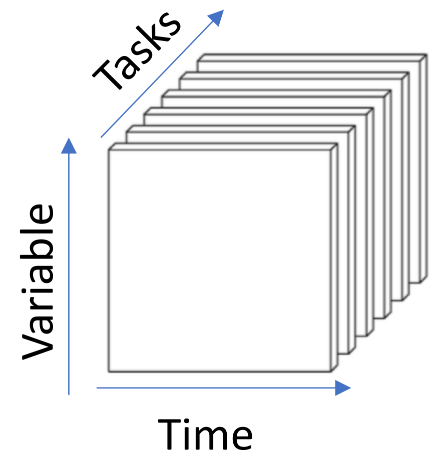
Dandelion – Database-Level



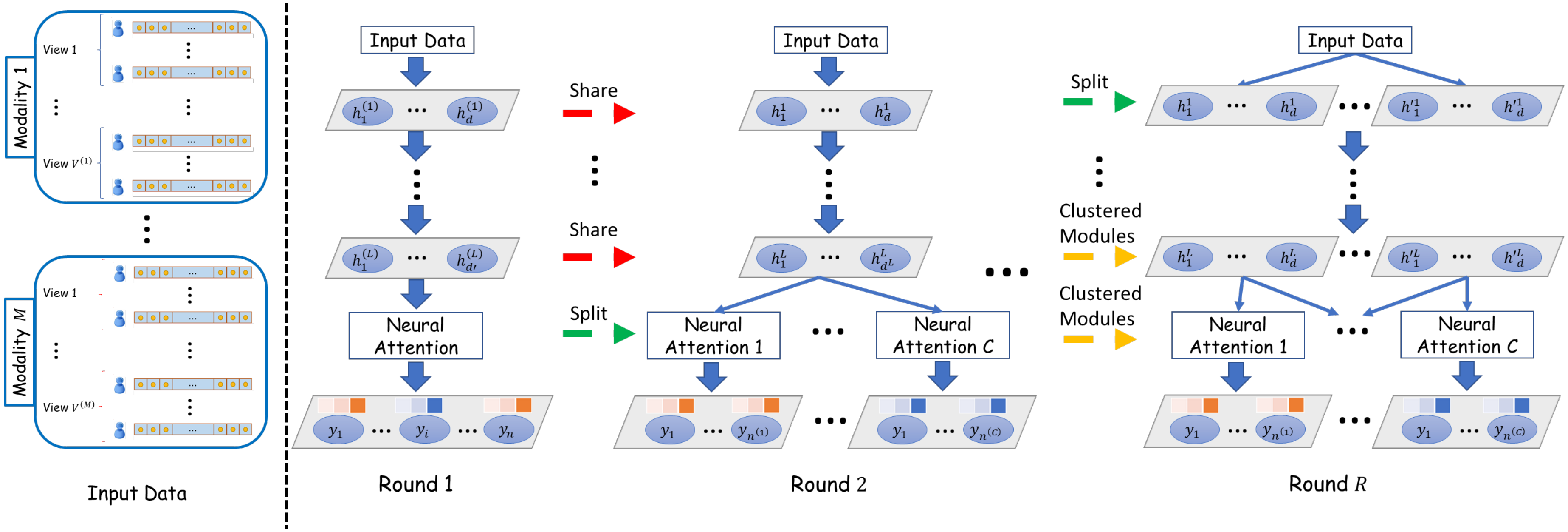
- **End-User Oriented Interpretation via Trinity Attention.**
  - The interpretability of the predictive model is critical for end users to understand and evaluate the model outputs.
  - Interpretation over tasks, time and variables via summarization function $f_{agg}(\cdot)$.

$$\boldsymbol{\beta}_{var} = f_{agg}(\boldsymbol{\beta}) = \left[\frac{\sum_t^T \beta(t,1)}{\sum_t^T \sum_f^F \beta(t,f)}, \ldots, \frac{\sum_t^T \beta(t,F)}{\sum_t^T \sum_f^F \beta(t,f)} \right]$$

$$\boldsymbol{\beta}_{temp} = f_{agg}(\boldsymbol{\beta}^T) = \left[\frac{\sum_f^F \beta(1,f)}{\sum_t^T \sum_f^F \beta(t,f)}, \ldots, \frac{\sum_f^F \beta(T,f)}{\sum_t^T \sum_f^F \beta(t,f)} \right]$$



Dandelion – An Overview



Outline

- Background
- Problem Definition
- Proposed *Dandelion* Framework
- Experiments
- Conclusion

Experiment Setup

- **Data set**
  - 396 Stocks of public US companies
  - 4 modalities, including finance data, news, Google Trends and weather data
  - 4 stock sectors
  - 14 years

Sector	# of stocks	Starting time stamp	Ending time stamp
Consumer Cyclical	90	5-6-2004	6-26-2018
Healthcare	105	5-3-2004	5-20-2018
Industrial	98	5-4-2004	6-27-2018
Technology	103	5-3-2004	6-25-2018

Experiment Setup

- **Comparison Methods**
  - **ConEst**: the Wall Street consensus estimates.
  - **ARIMAX**: an Auto Regressive Integrated Moving Average based method.
  - **MVR**: a multi-view regression approach that uses canonical correlation analysis) to make predictions via ridge regression..
  - **Bi-LSTM**: a bi-directional LSTM architecture.
  - **MNA**: a neural attention network that is designed for demand forecasting using multi-modality event data.
  - **Dandelion-M**: a variation of Dandelion framework, which ignores the task heterogeneity.
  - **Dandelion-D**: a variation of Dandelion framework, which ignores the data heterogeneity but adopts the hierarchical multitask learning mechanism.

Experimental Results

- **Sector-Level Prediction Performance**
  - We compare prediction accuracy based on the median absolute deviation (Med-abs).
  - Med−abs $= median(|X_i - \bar{X}|)$, where $\bar{X} = median(X)$
  - The lower the better!

Methods		Con. Cyc.	Healthcare	Indus.	Tech.	All
Industry Benchmark	ConEst	0.01575	0.02247	0.01587	0.02133	0.01857
Regression	ARIMAX	1.22291	1.95461	1.28935	2.08068	1.55457
	MVR	0.48691	0.48922	0.51235	0.57606	0.51599
Neural Networks	Bi-LSTM	0.93098	1.54184	0.97901	1.44376	1.19222
	MNA	0.01692	0.02251	0.01695	0.02132	0.01960
Our Approaches (v.s ConEst)	*Dandelion*	0.01430 (↓ 9.2%)	**0.02119** (↓ 5.7%)	**0.01560** (↓ 1.7%)	**0.01883** (↓ 11.7%)	**0.01731** (↓ 6.8%)
	Dandelion-M	0.01582 (↑ 0.4%)	0.02173 (↓ 3.2%)	0.01579 (↓ 0.5%)	0.02032 (↓ 4.8%)	0.01806 (↓ 2.8%)
	Dandelion-D	**0.01387** (↓ 11.9%)	0.02127 (↓ 5.3%)	0.01567 (↓ 1.3%)	0.01970 (↓ 7.6%)	0.01753 (↓ 5.6%)

Table 3: Results of four sector companies. *Dandelion* and its variations (i.e., *Dandelion*-M, *Dandelion*-D) achieve smaller Med-abs values than all benchmark methods on each individual sector as well as the overall performance. (The lower the better)

Experimental Results

- **Stock-Level Prediction Performance over Time.**

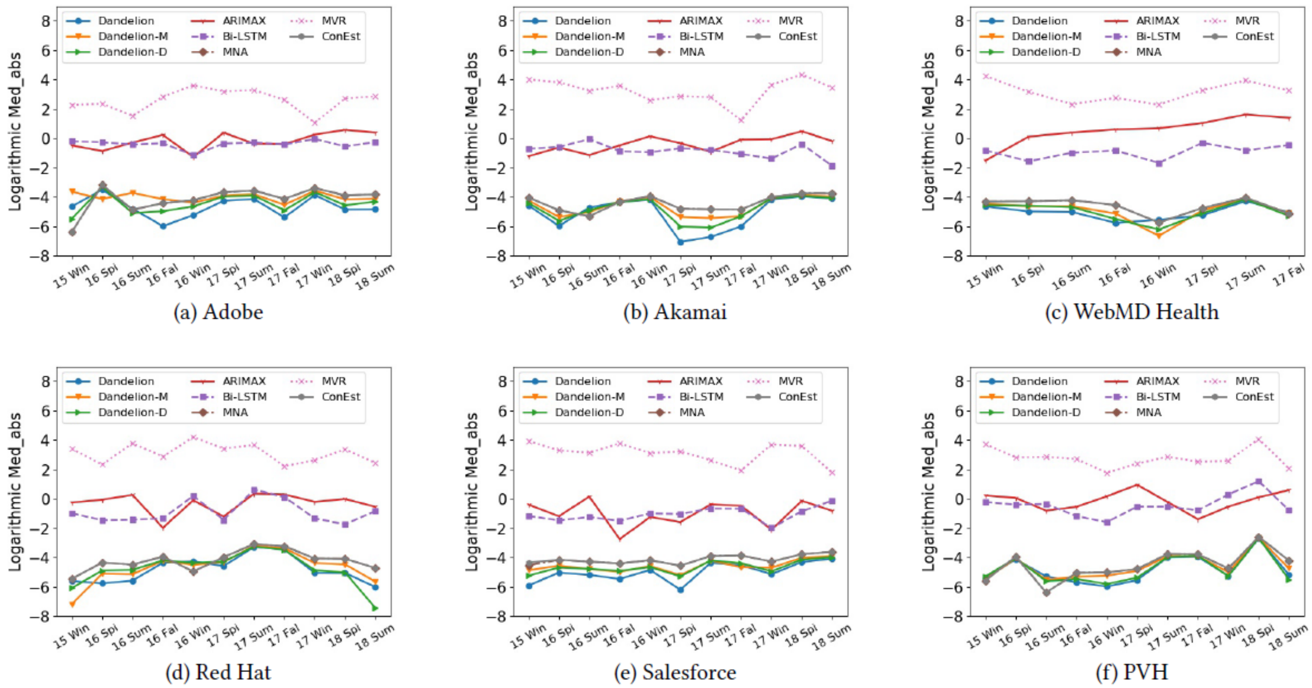


Figure 3: Individual prediction performance of six companies over time. Dandelion consistently performs better than all other methods in most of the time. (The lower the better)

The lower the better!

Experimental Results

- **Profitability Performance**

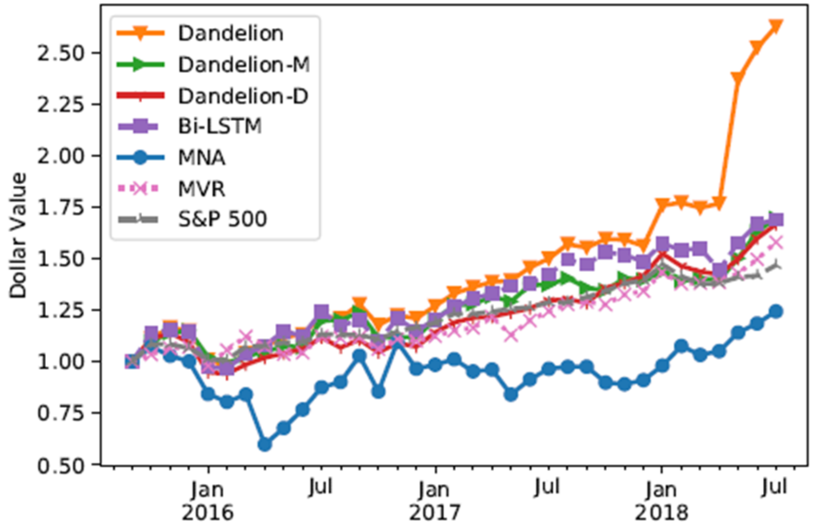
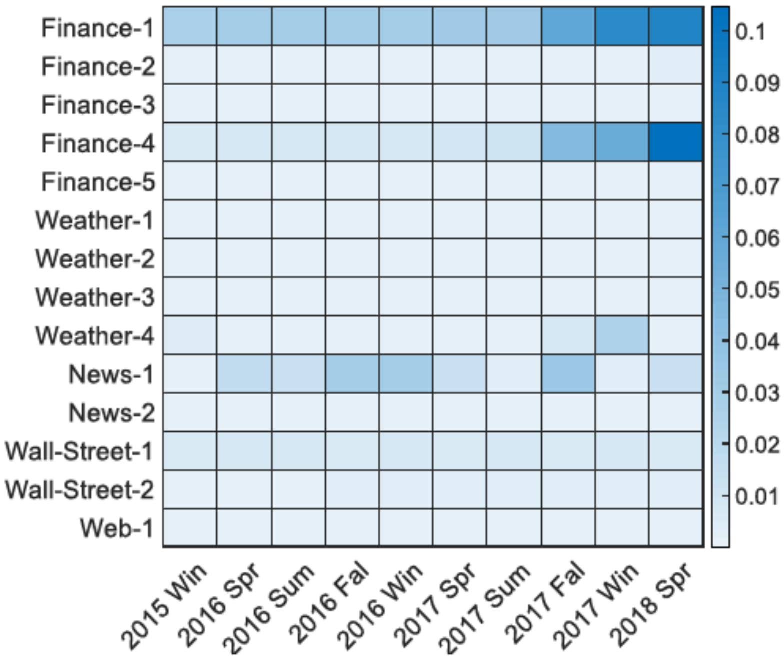


Figure 4: Portfolio value across the testing period if starting with $1. *Dandelion* outperforms all the benchmark portfolios and increased more than 1.6 times in less than 3 years. (The larger the better)

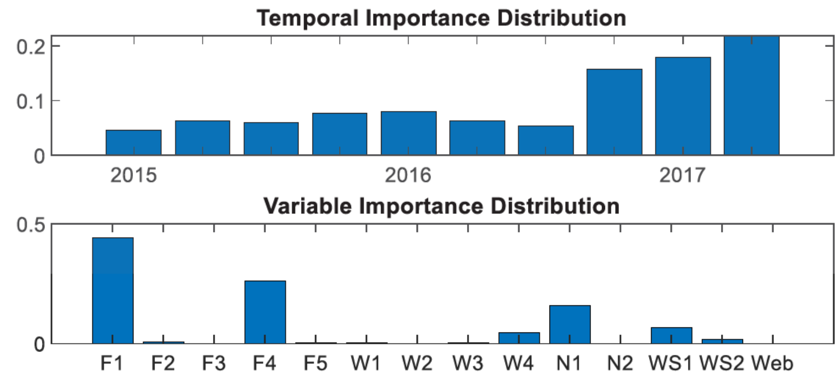
The higher the better!

Experimental Results

- **Data Interpretation**
  - Amgen: a biotechnology company



(a) Attention heat map (the darker, the higher importance)

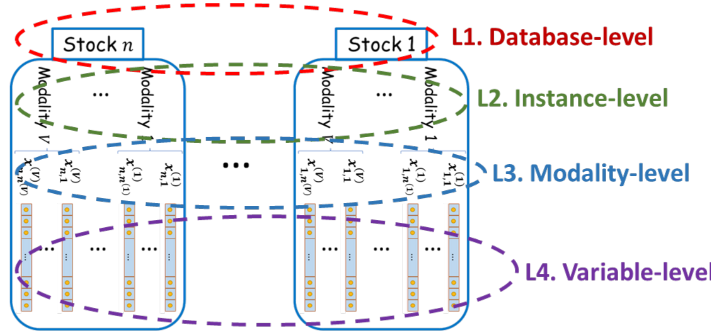


(b) Summary attention

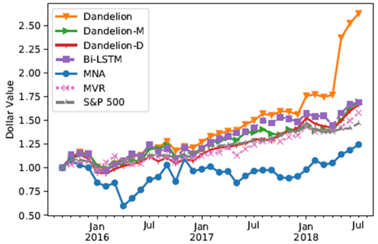
Outline

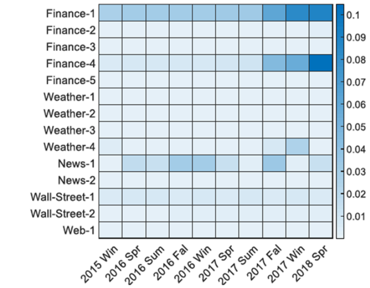
- Background
- Problem Definition
- Proposed *Dandelion* Framework
- Experiments
- Conclusion

Conclusion



- **Learning from Multi-Modality Multi-Variable Time Series**
  - Challenge #1: Data Heterogeneity (L4. and L3.)
  - Solution #1: Multi-modality multi-variable learning.
  - Challenge #2: Task Heterogeneity (L2.)
  - Solution #2: Fully-adaptive hierarchical multi-task learning.
  - Challenge #3: Data Interpretation (L1.)
  - Solution #3: Trinity attention.





(a) Attention heat map (the darker, the higher importance)

- **Results**
  - ***Dandelion*** outperforms other baseline methods in financial forecasting.
  - ***Dandelion*** outperforms other baseline methods in a case study of profitability analysis.
  - ***Dandelion*** provides interpretation w.r.t. tasks, variables, and time.

Back Up Slides

- **Multi-Modality Multi-Task Time Series Forecasting**
  - **Given**: (i) a multi-modality time series $\boldsymbol{X} = \{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$ from time $t = 1$ to $t = T$; (ii) the target signal $\boldsymbol{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\}$ from time $t = 1$ to $t = T$.
  - **Find**: the prediction $\widehat{\boldsymbol{Y}} = \{\widehat{\boldsymbol{y}}_1, \ldots, \widehat{\boldsymbol{y}}_n\}$ from time $t = T + 1$ to $t = T + T'$.