Temporal data mining

Students: Vedashree Bagade,
Ankit Gupta
Mentor: Jay Pujara

Motivation

- In most of the real world datasets like scientific data (treatment history of a patient, astronomy) or stock market, the future depends on the past.
- We study various techniques used to analyze a time-series datasets from different domains like:
  - health
  - stocks
  - automation
  - astrophysics

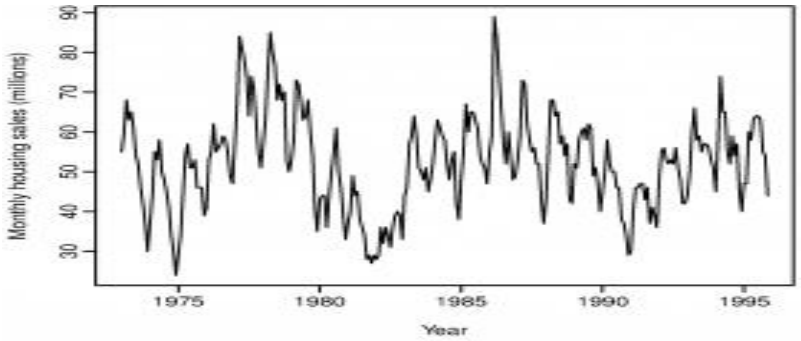
Temporal Data Mining

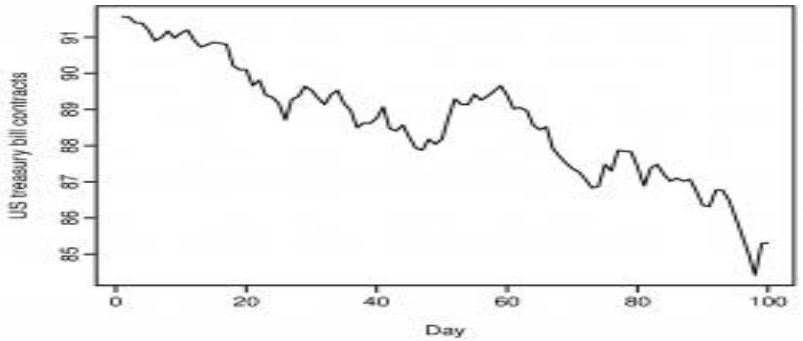
- **What is Time Series Data?**
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Dow Jones Dataset
- Our results: Mhealth Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

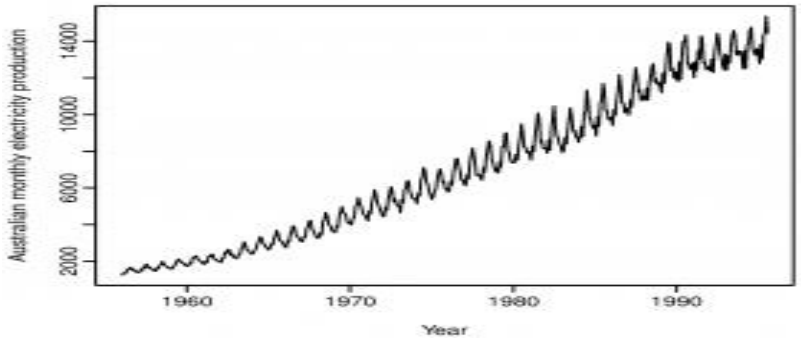
What is Time-Series Data?

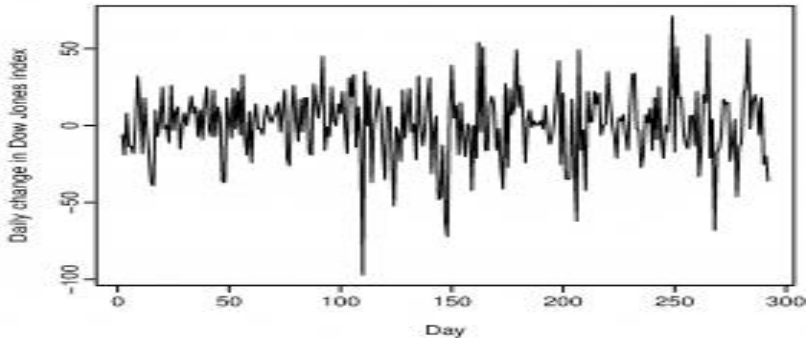
- Readings taken at different time intervals of a certain quantity.
- Each row corresponds to feature data read at different times.
- Examples:
    - Temperature data collected at different spots in a city at some interval (say 1 hour) to calculate city data and find pattern in it.
    - Stock market data
    - Electricity usage data

Time Series Data: Examples







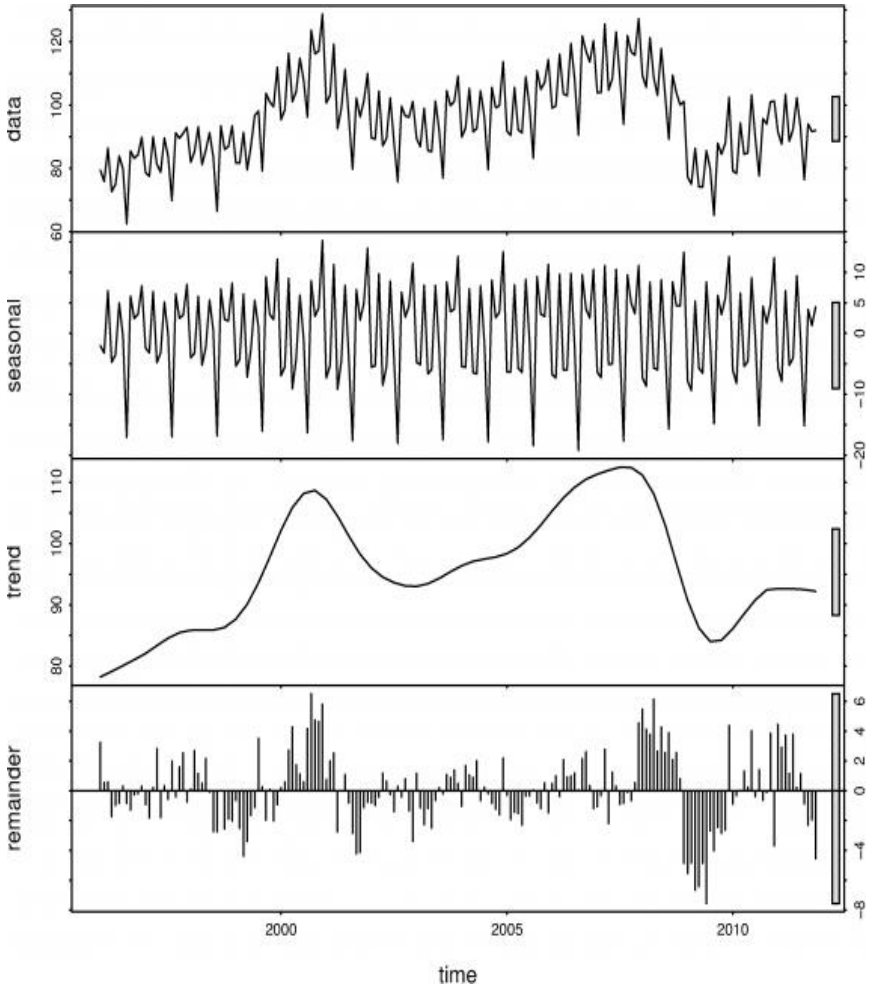


Temporal Data Mining

- What is Time Series Data?
- **Patterns: Trends, Seasonal and residual**
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Dow Jones Dataset
- Our results: Mhealth Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

Patterns

- Trend (T)
  - Indicates the long term increasing or decreasing movement in the function.
  - Can be thought of as a “smoothing” function over the original signal.
- Seasonal (S)
  - Indicates the influence of periodic time interval on the curve. Period is fixed and known.
- Cyclic (C)
  - Indicates that the curve repeats behavior after a certain unknown period.
  - Cyclic periods are generally more than seasonal periods



Temporal Data Mining

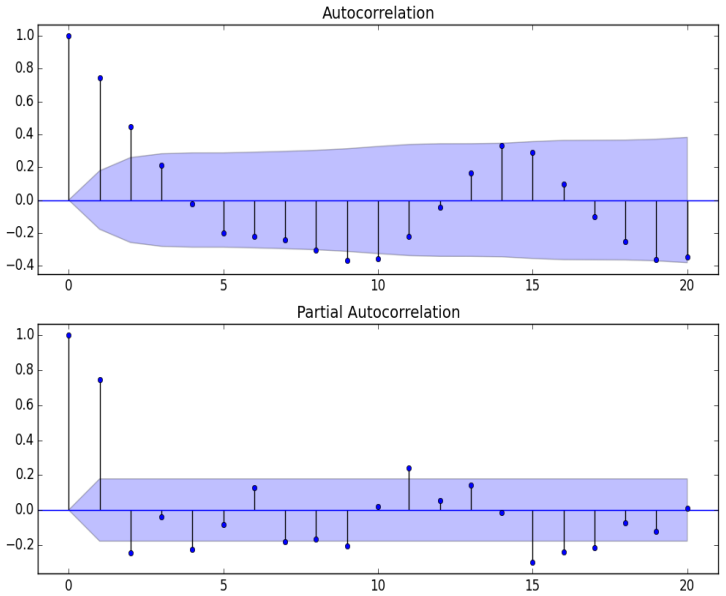
- What is Time Series Data?
- Components: Trends, Seasonal and residual
- **Decomposition: Additive and Multiplicative**
- Techniques: AR, MA, AR+MA
- Our results: Dow Jones Dataset
- Our results: Mhealth Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

Decomposition

- Any signal consists of 3 features: trend (T), season (S) and the remaining, residual (R).
- Additive
  - Y = T + S + R
  - Used when the fluctuations of trend or seasonal component not related to time series level
  - Example: Temperature Time series
- Multiplicative
  - Y = T * S * R
  - Used when the fluctuations of trend or seasonal component is related to time series level.
  - Example: Economics time series

Autocorrelation and partial correlation

- Autocorrelation is the similarity between observations as a function of the time lag between them
- Partial correlation will look for correlation between two signals
- Model order can be determined using Autocorrelation and partial correlation.



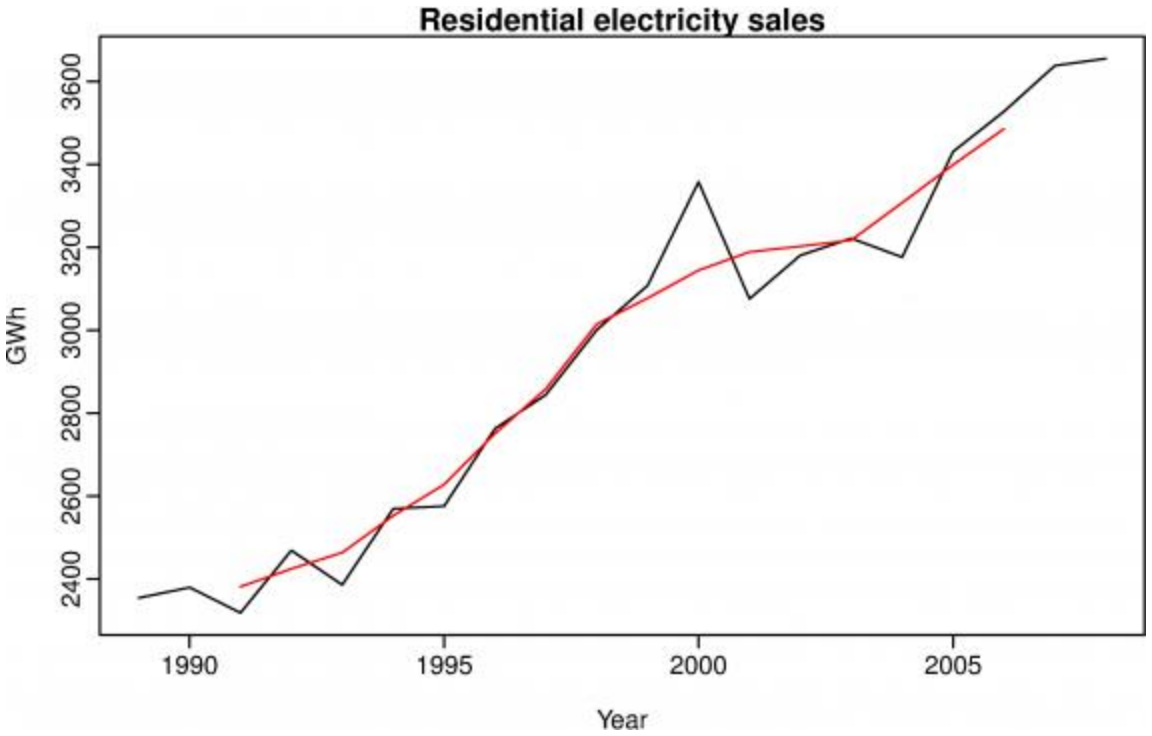
Temporal Data Mining

- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- **Analysis Tools: AR, MA, AR+MA**
- Our results: Dow Jones Dataset
- Our results: Mhealth Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

Moving Averages (MA)

- A moving average of order m means that we take average of (m-1)/2 components each from the past and future readings, along with the current reading.
- So, a 5-MA will mean we add 2 past, 1 current value and 2 future values and divide the result by 5.
- Generates a trend cycle
- Smooths the original signal
- Reduces impact of noisy and random elements

Moving Average visualization



Auto-regression (AR)

- Just like Linear Regression
- Only difference being that now the classifier “fits” on the past values of the same features.
- We can predict the future values based on the weights and constants obtained.
- Just like Linear Regression, we can have a quadratic dependency of future value on past values.
- Accuracy depends on order of autoregression.

Autoregression - Moving Average (ARMA)

- Combination of AR and MA
- Returns the weights and constants for AR
- Returns the coefficients for MA
- All this newly obtained features can be used to predict future values.
- Appropriate order of AR and MA can be decided by autocorrelation and partial autocorrelation graphs.

Temporal Data Mining

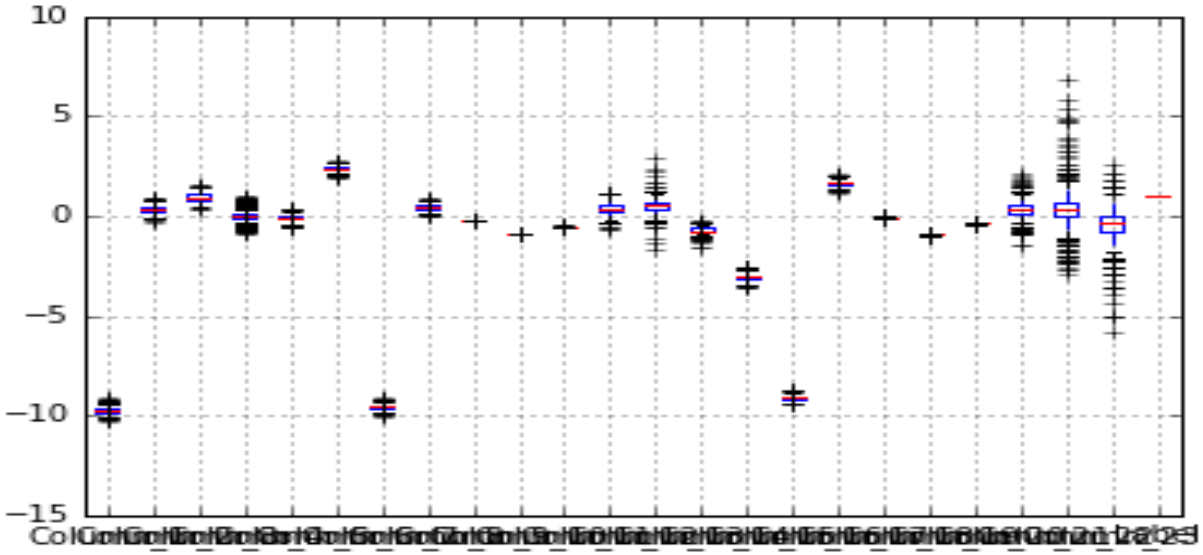
- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- **Our results: Mhealth Dataset**
- Our results: Dow Jones Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

MHealth Dataset (Multivariate, Time Series)

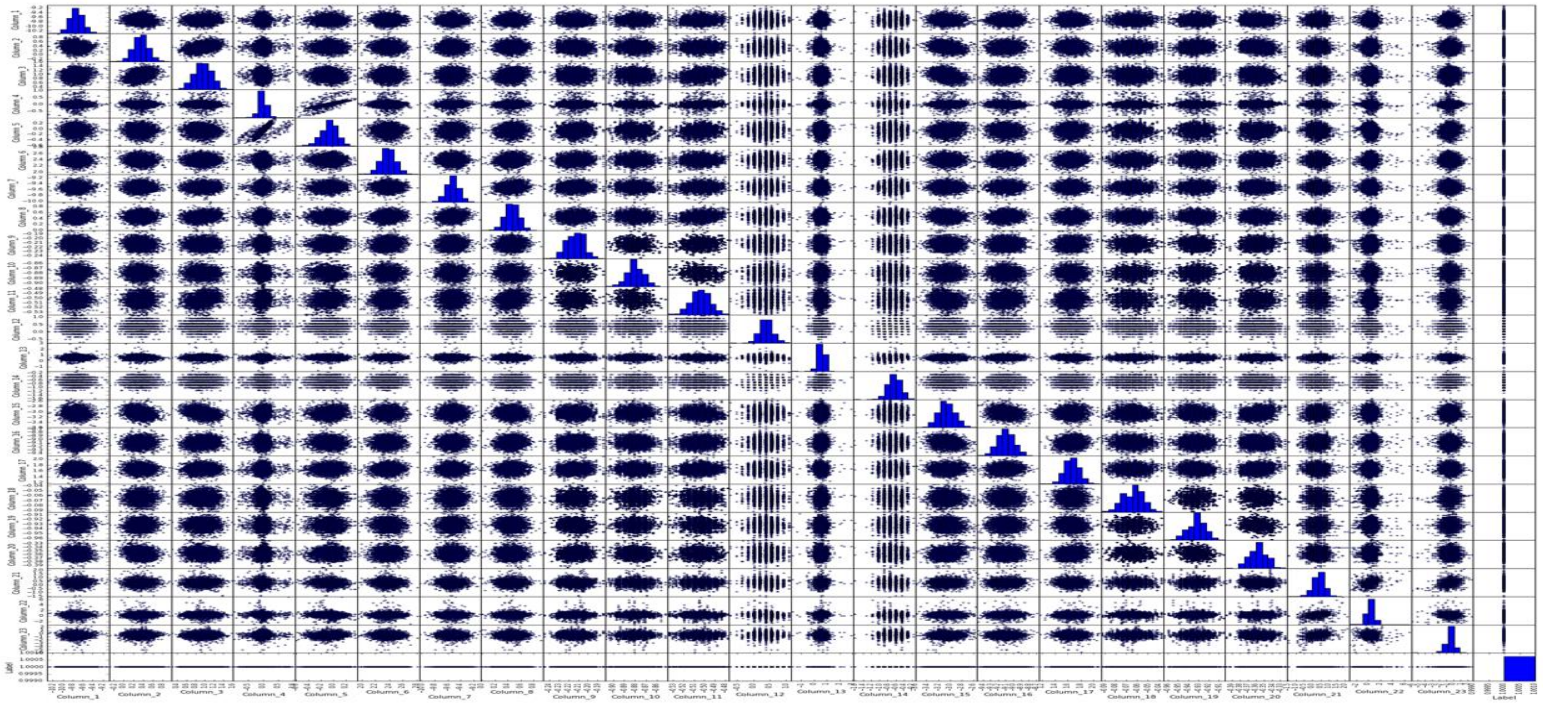
- 10 Subjects
- 24 readings from 3 sensors
- Each reading taken at 50 HZ Freq for duration of 1 min or 20 repetitions of each task.

Data Visualization

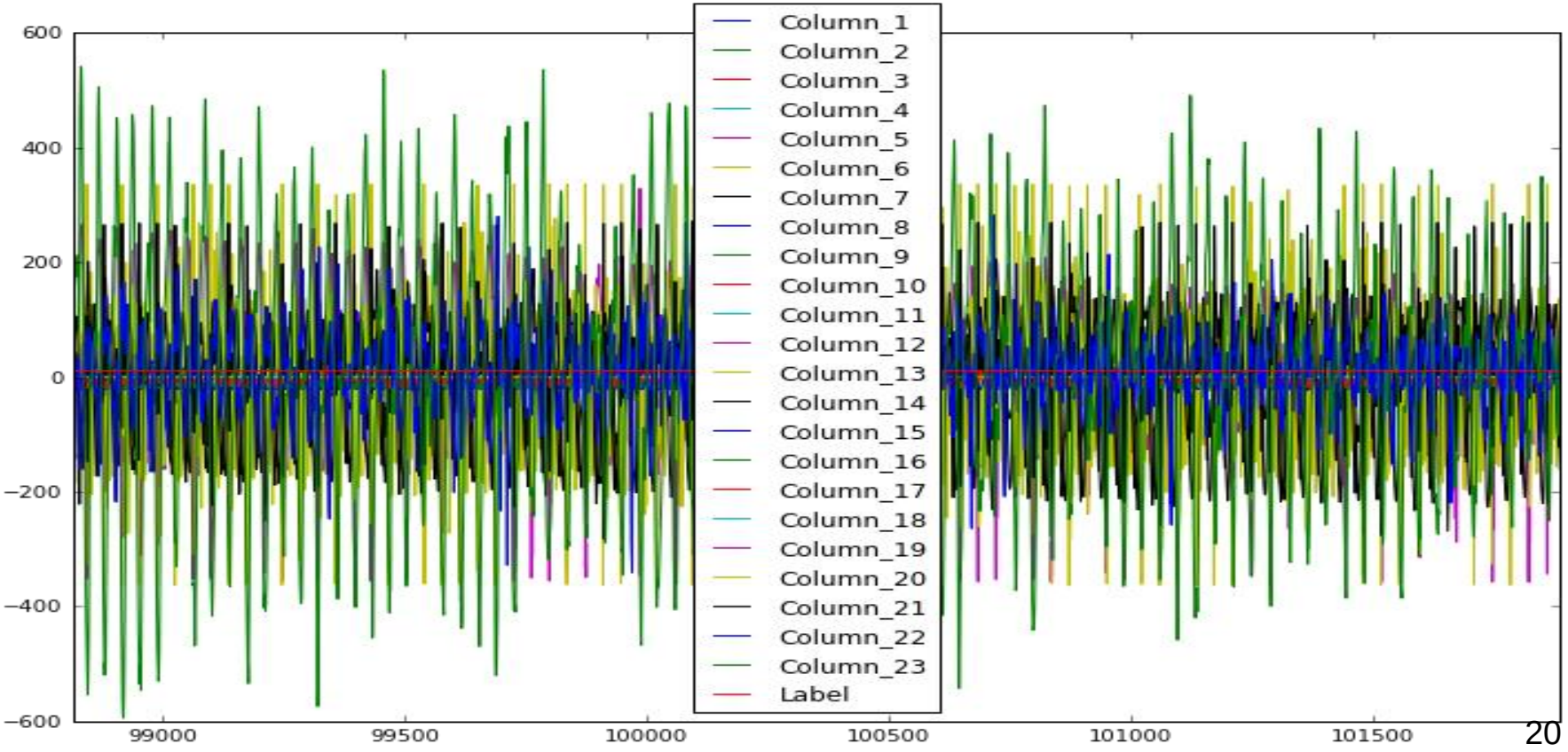
Box Plot



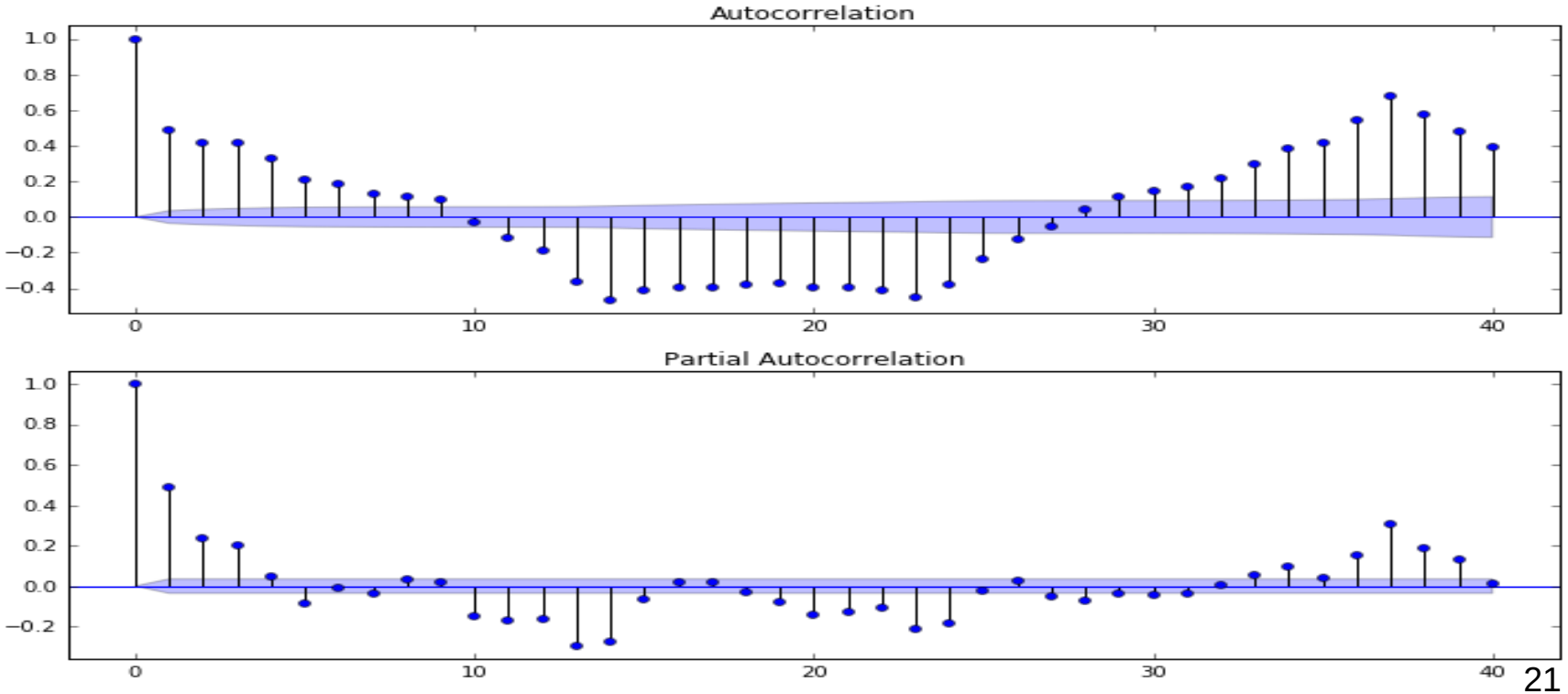
Scatter Plot



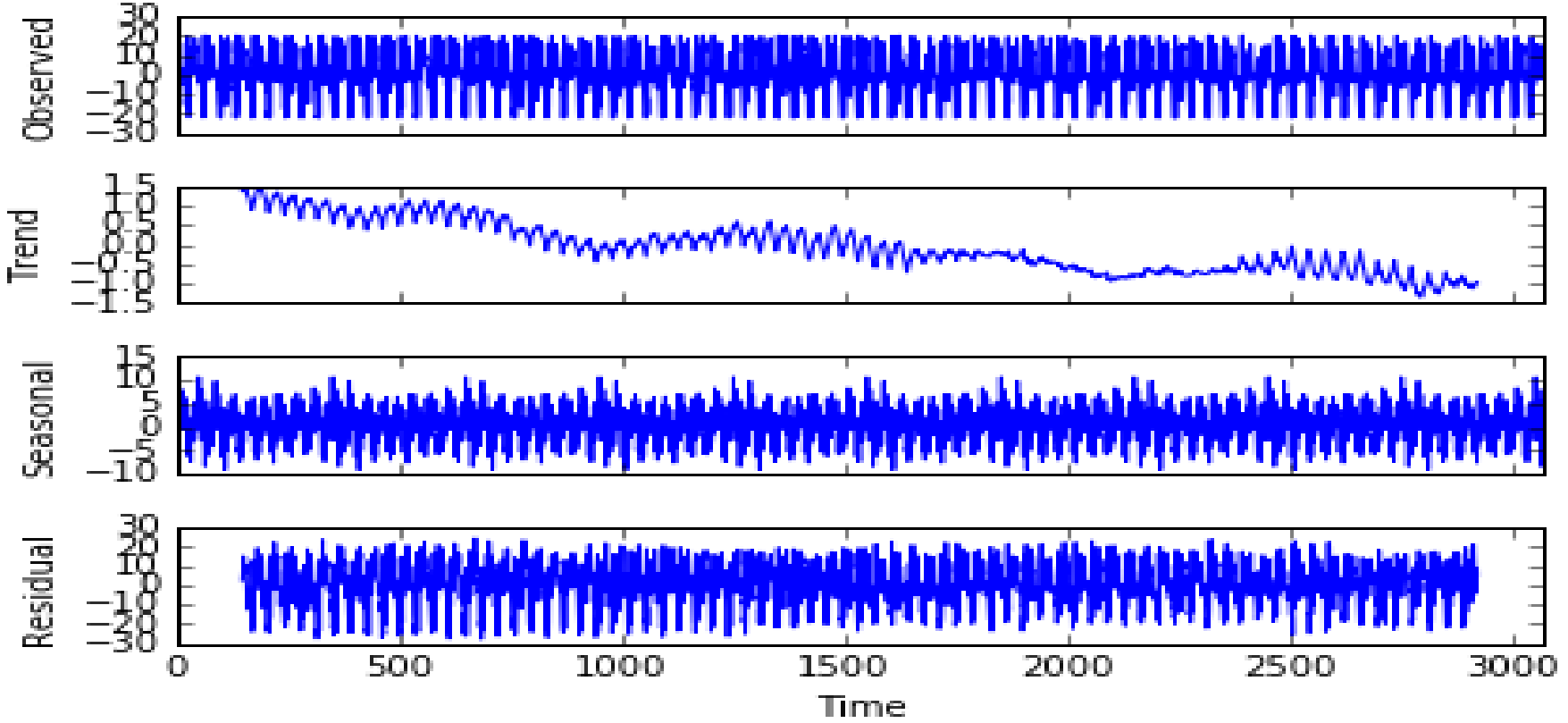
Plot for variations in all the columns for one label



Autocorrelation and partial correlation for one attribute of a class



Trend Analysis



Results for Mhealth DataSet

	One vs One Classifier (Logistic Regression)	One vs Rest Classifier (Logistic Regression)	KNN	Decision Tree
Considering Coefficients only	67.5%	73.33%	70.8%	74.16%
Considering constants as well as co-efficients	91.66%	87.5%	92.5%	92.5%

Temporal Data Mining

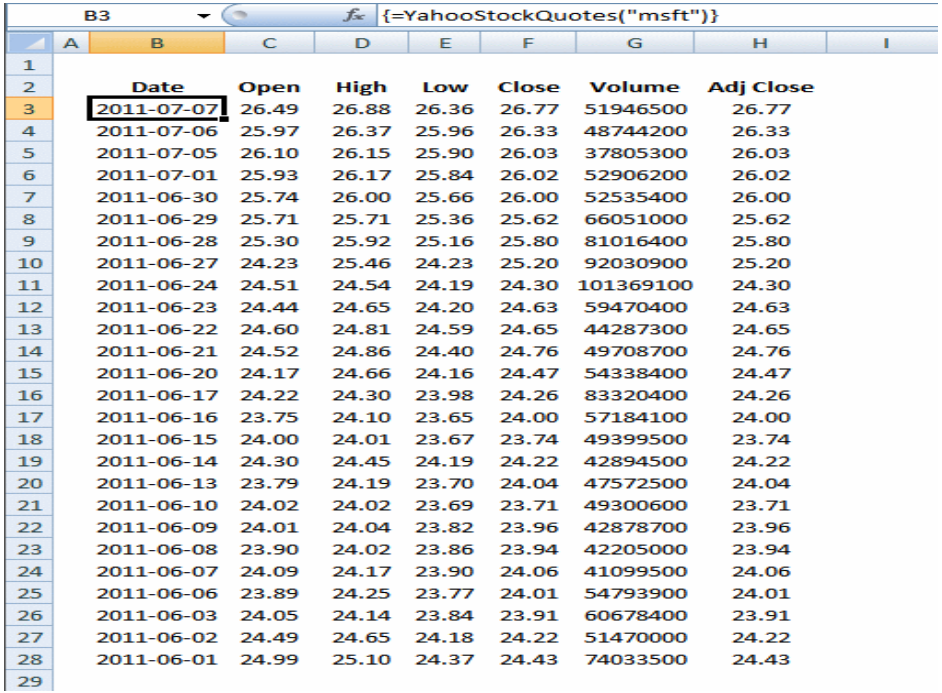
- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Mhealth Dataset
- **Our results: Dow Jones Dataset**
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

Dow Jones Dataset

- Dataset:
  - Each row corresponds to data for 1 week for a company
  - 25 weeks
  - 2 quarters
  - 30 companies
- Shape:
  - 25 * 30 = 750 instances (rows)
  - 16 features
- Task:
  - Predict the gain that can be expected from investing in a particular stock in next week

Dow Jones: Features

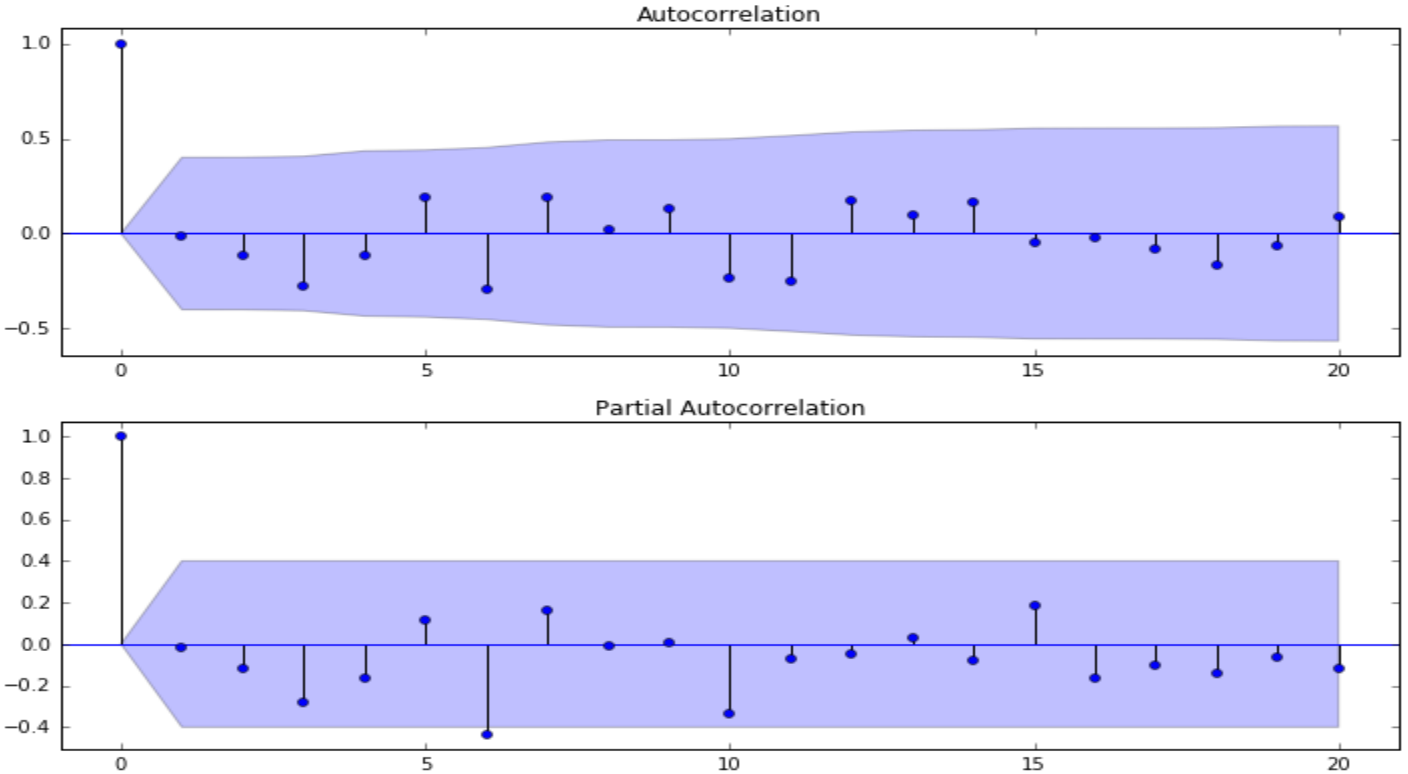
- True Features:
  - Opening stock price for the week
  - Closing
  - Week's high
  - Week's low
  - Volume
  - Next dividend
- Transformed features
  - % change in price for current week
  - % change in volume
  - % change in price for next week
  - % gain from dividend



B3 {=YahooStockQuotes("msft")}

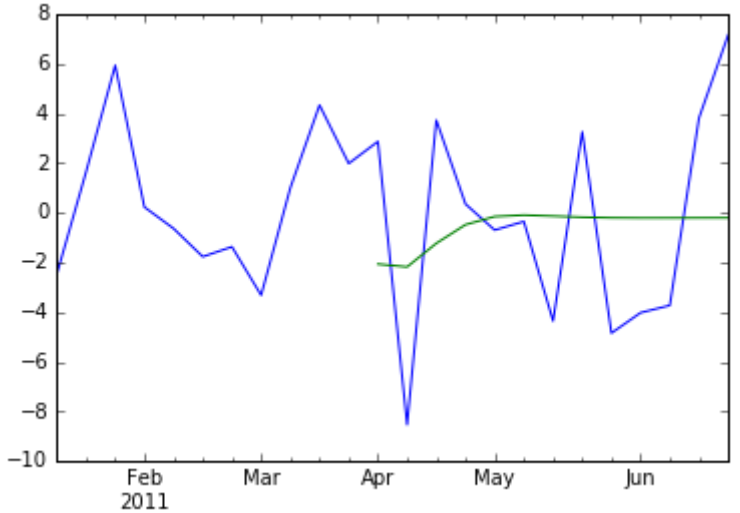
Date	Open	High	Low	Close	Volume	Adj Close
2011-07-07	26.49	26.88	26.36	26.77	51946500	26.77
2011-07-06	25.97	26.37	25.96	26.33	48744200	26.33
2011-07-05	26.10	26.15	25.90	26.03	37805300	26.03
2011-07-01	25.93	26.17	25.84	26.02	52906200	26.02
2011-06-30	25.74	26.00	25.66	26.00	52535400	26.00
2011-06-29	25.71	25.71	25.36	25.62	66051000	25.62
2011-06-28	25.30	25.92	25.16	25.80	81016400	25.80
2011-06-27	24.23	25.46	24.23	25.20	92030900	25.20
2011-06-24	24.51	24.54	24.19	24.30	101369100	24.30
2011-06-23	24.44	24.65	24.20	24.63	59470400	24.63
2011-06-22	24.60	24.81	24.59	24.65	44287300	24.65
2011-06-21	24.52	24.86	24.40	24.76	49708700	24.76
2011-06-20	24.17	24.66	24.16	24.47	54338400	24.47
2011-06-17	24.22	24.30	23.98	24.26	83320400	24.26
2011-06-16	23.75	24.10	23.65	24.00	57184100	24.00
2011-06-15	24.00	24.01	23.67	23.74	49399500	23.74
2011-06-14	24.30	24.45	24.19	24.22	42894500	24.22
2011-06-13	23.79	24.19	23.70	24.04	47572500	24.04
2011-06-10	24.02	24.02	23.69	23.71	49300600	23.71
2011-06-09	24.01	24.04	23.82	23.96	42878700	23.96
2011-06-08	23.90	24.02	23.86	23.94	42205000	23.94
2011-06-07	24.09	24.17	23.90	24.06	41099500	24.06
2011-06-06	23.89	24.25	23.77	24.01	54793900	24.01
2011-06-03	24.05	24.14	23.84	23.91	60678400	23.91
2011-06-02	24.49	24.65	24.18	24.22	51470000	24.22
2011-06-01	24.99	25.10	24.37	24.43	74033500	24.43

Dow Jones visual analysis

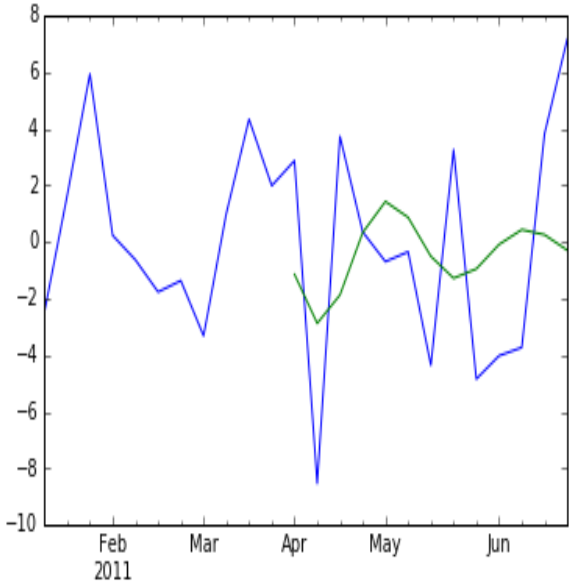


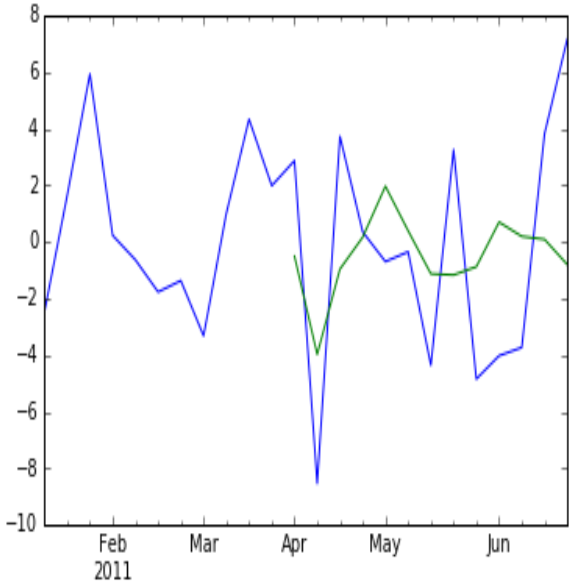
First Approach: ARMA only

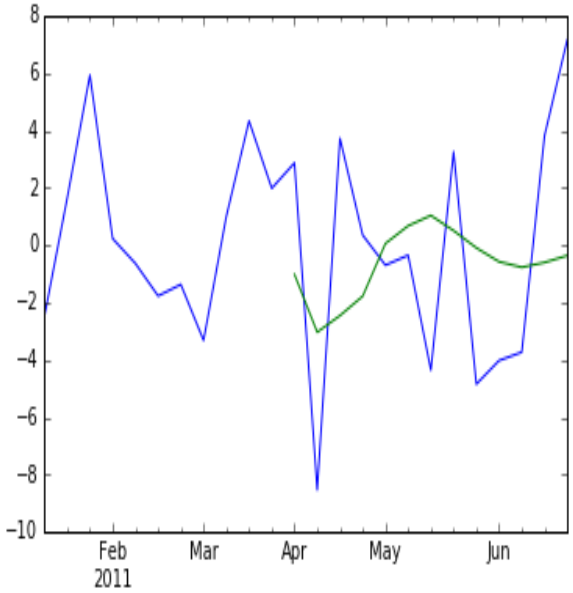
- Consider the feature we want to predict and use ARMA model for only that.
- Percent_change_price_next_week
- Train over first quarter
- Predict the second quarter
- Test different orders for AR and MA
- Compare different models



ARMA Model comparisons: First Approach





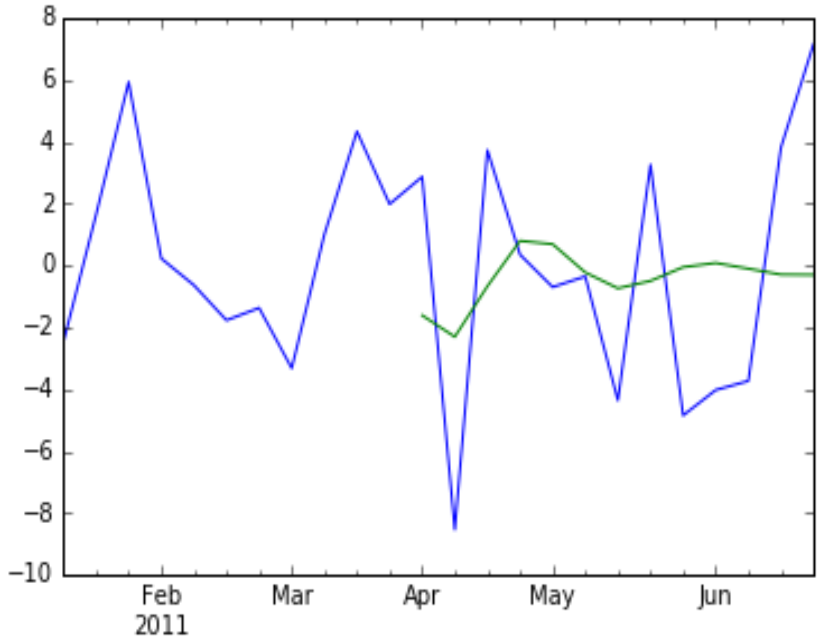


Shortcomings of First approach

- Each company has its own classifier
  - Fails if market sentiment too positive
  - The industry with which a company is associated might be doing really well
- The generated classifier only depends on one feature

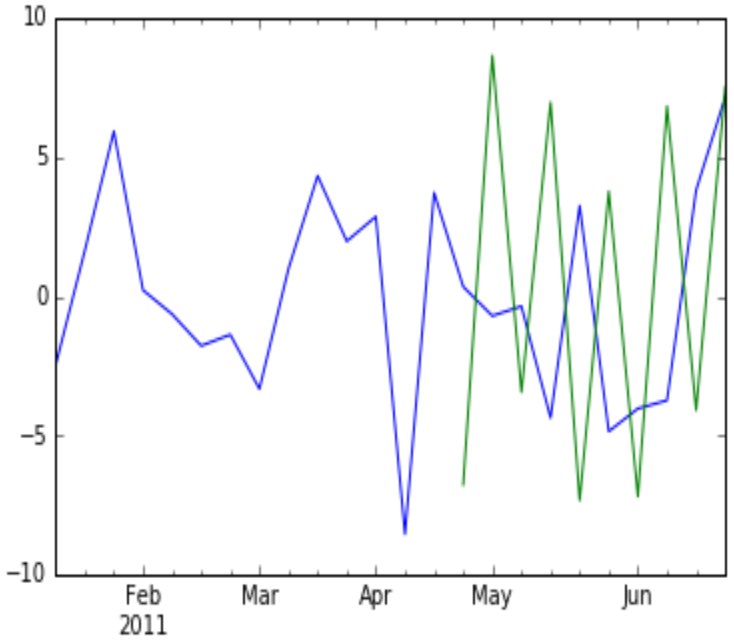
Second Approach: ARMA + Linear Regression

- Construct a new dataset consisting of ARMA coefficients of relevant features from the previous values in original dataset and target.
- Do this company wise (group by)
- Feed this transformed dataset to linear regression model
- Calculate ARMA coefficients again by adding the current week's values
- Predict next week's gain



Second Approach

- Setup:
  - Since data is limited for each company, we train for first 4 weeks and predict for next 4.
  - Then we add the next 4 weeks to dataset, so now we train on first 8 weeks and predict next 4 weeks.
  - Can also train on 1st quarter data and predict 2nd quarter
- Difference:
  - Here, we generate a more general classifier that considers all the features for all companies
- Results:
  - Best accuracy = 78%



Temporal Data Mining

- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Mhealth Dataset
- Our results: Dow Jones Dataset
- **Our results: Robot Execution Failure Dataset**
- Case study: LIGO Gravitational Wave detection
- Conclusions
- Future Work

Robot Dataset
(Multivariate, Time Series)

- 5 datasets defining different problems of failure
- 15 readings per failure in a period for force and torque
- 90 readings per Label
- Each reading taken at 50 HZ Freq for duration of 1 min

Robot Dataset

- Autoregression used as a feature transformation strategy
- This gave the trend and residual signal associated with the data
- Different classifiers then applied on the extracted features for training and testing purposes.

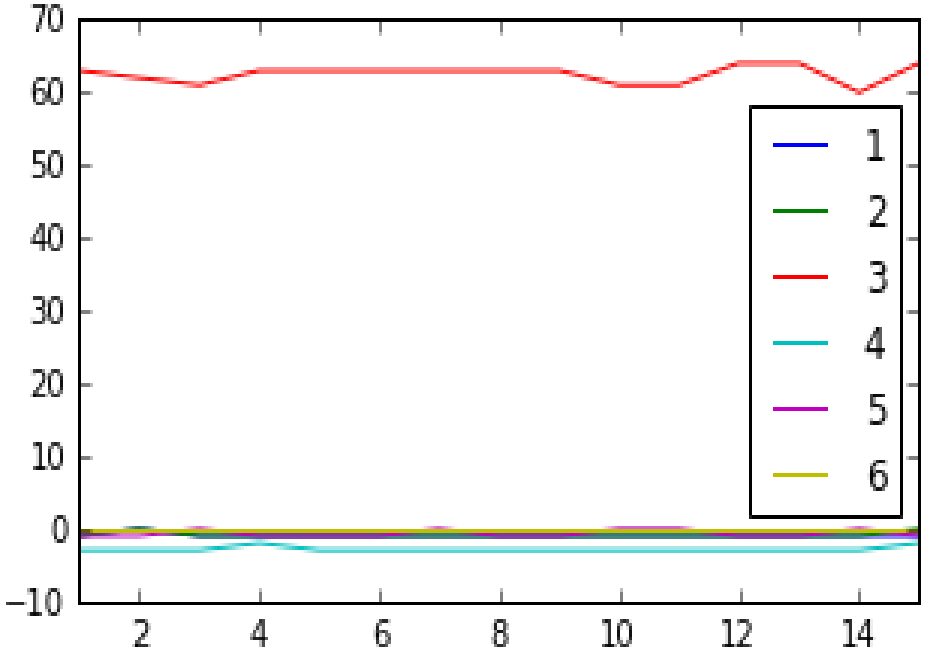
normal						
	-1	-1	63	-3	-1	0
	0	0	62	-3	-1	0
	-1	-1	61	-3	0	0
	-1	-1	63	-2	-1	0
	-1	-1	63	-3	-1	0
	-1	-1	63	-3	-1	0
	-1	-1	63	-3	0	0
	-1	-1	63	-3	-1	0
	-1	-1	63	-3	-1	0
	-1	-1	61	-3	0	0
	-1	-1	61	-3	0	0
	-1	-1	64	-3	-1	0
	-1	-1	64	-3	-1	0
	-1	-1	60	-3	0	0
	-1	0	64	-2	-1	0
normal						
	-1	-1	63	-2	-1	0
	-1	-1	63	-3	-1	0
	-1	-1	61	-3	0	0

	1	2	3	4	5	6
1	-1	-1	63	-3	-1	0
2	0	0	62	-3	-1	0
3	-1	-1	61	-3	0	0
4	-1	-1	63	-2	-1	0
5	-1	-1	63	-3	-1	0
6	-1	-1	63	-3	-1	0
7	-1	-1	63	-3	0	0
8	-1	-1	63	-3	-1	0
9	-1	-1	63	-3	-1	0
10	-1	-1	61	-3	0	0
11	-1	-1	61	-3	0	0
12	-1	-1	64	-3	-1	0
13	-1	-1	64	-3	-1	0
14	-1	-1	60	-3	0	0
15	-1	0	64	-2	-1	0

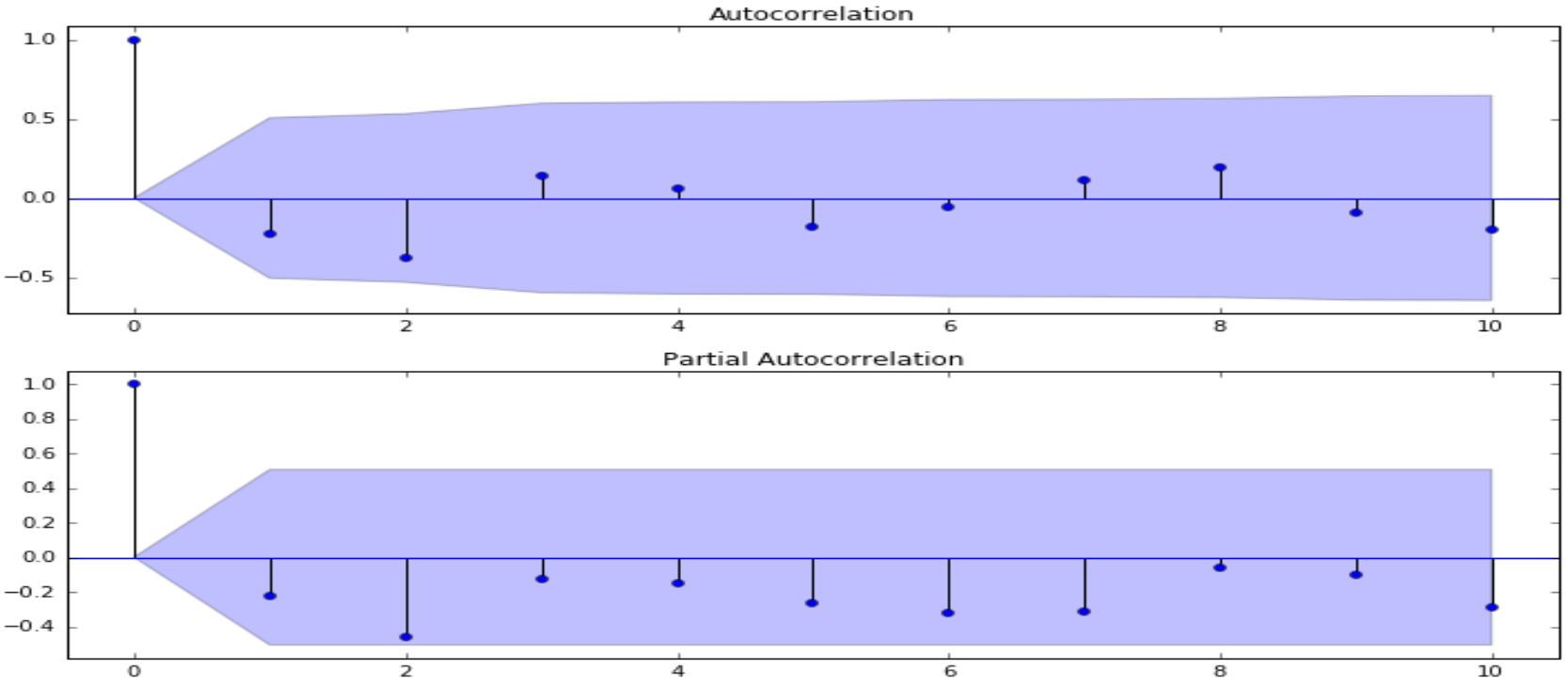
	class	0	1	2	3	4	5 \
0	normal	-0.932730	-0.072163	-0.872954	-0.144732	62.507648	-0.240887
1	normal	-0.865214	-0.087930	-0.666801	0.003015	62.163023	-0.236802

	6	7	8	9	10	11
0	-2.872954	-0.144732	-0.657972	-0.235557	0.000000	0.000000
1	-3.532647	0.004962	-1.299998	-0.263175	-0.252861	0.290513

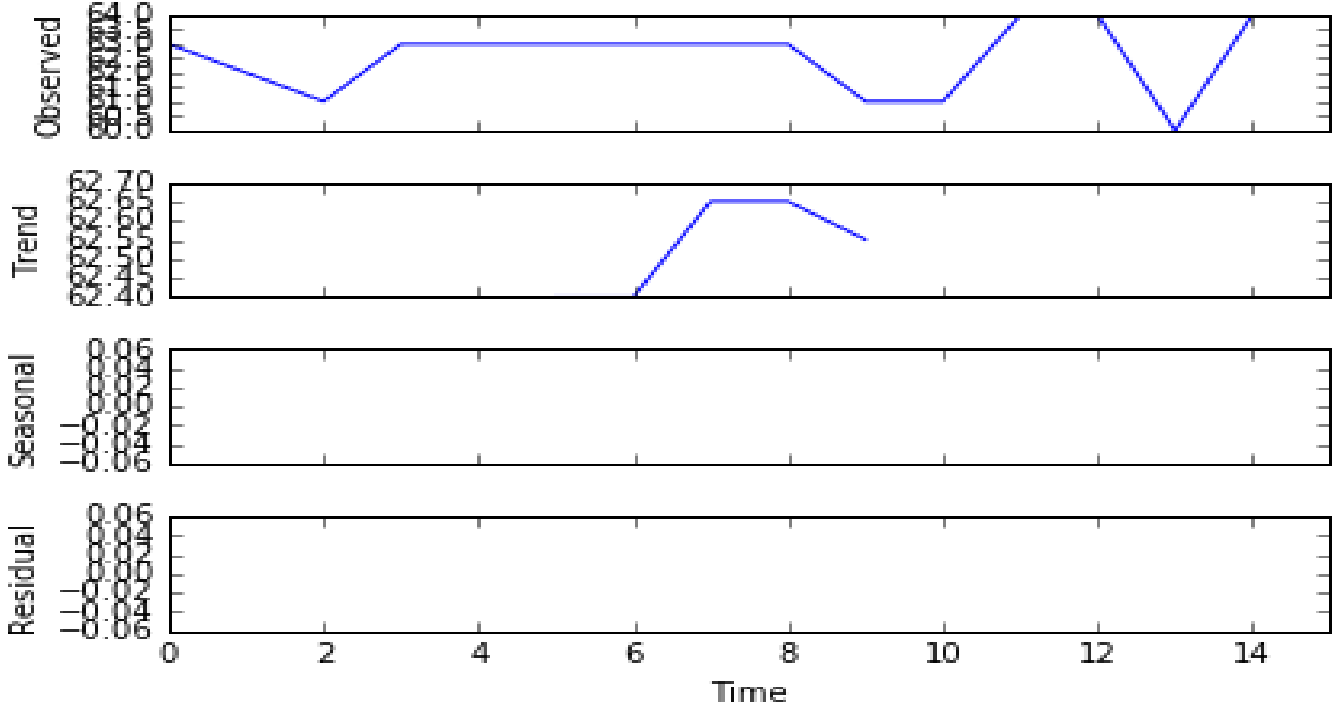
Failure to grasp position: Class-> Normal



Autocorrelation and partial correlation for one attribute of a class



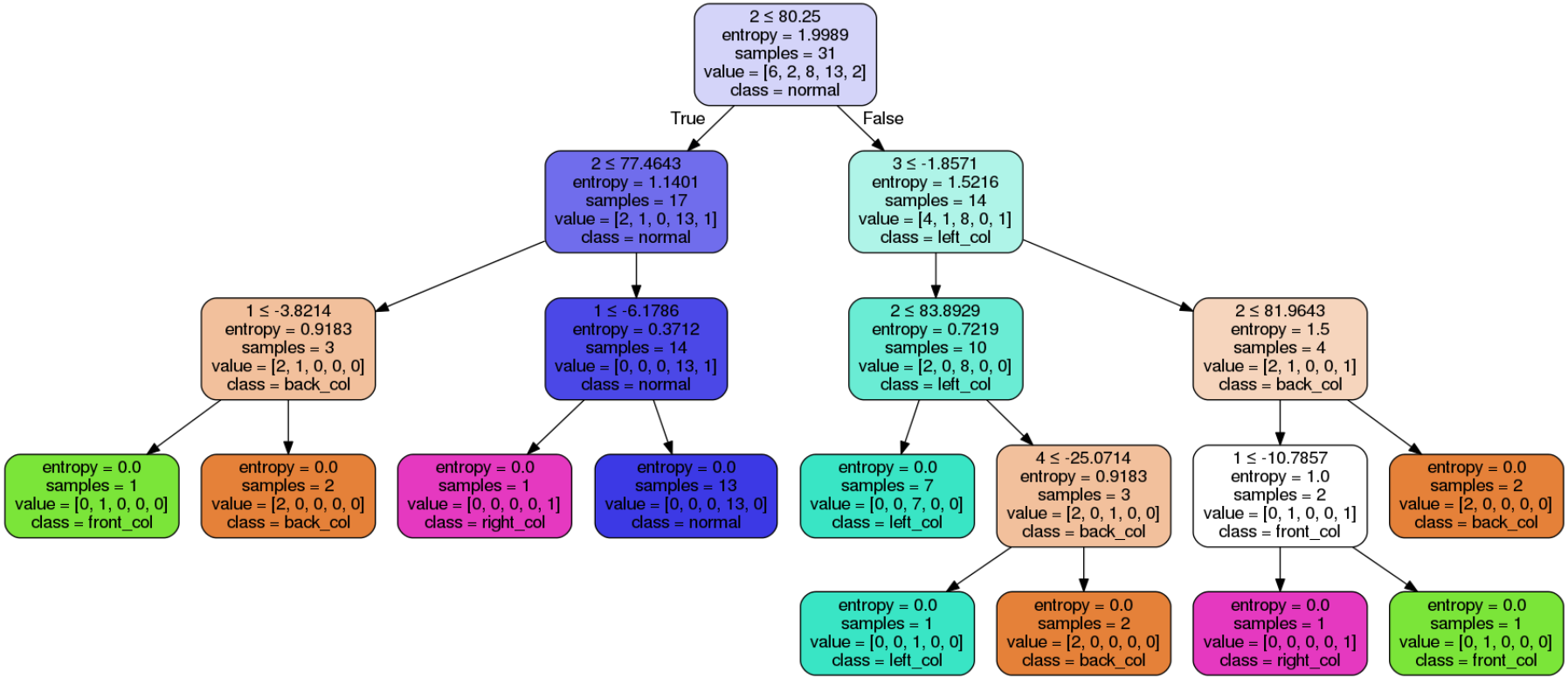
Trend Analysis



Robot dataset results

Classifier:	One vs One (Logistic Regression)	One vs Rest (Logistic regression)	KNN	Decision trees
Failure to grasp position	74.73%	65.93%	89.3%	91.11%
Failure to transfer part	60.33%	62.33%	71.2%	61.83%
Position of part after transfer failure	55.55%	63%	71.5%	79.66%
Failure in approach to ungrasp position	89.85%	82.99%	95.0%	89.099%
Failures in motion with part	68.7%	62.68%	71.0%	73.98%

Decision tree classifier for one Failure



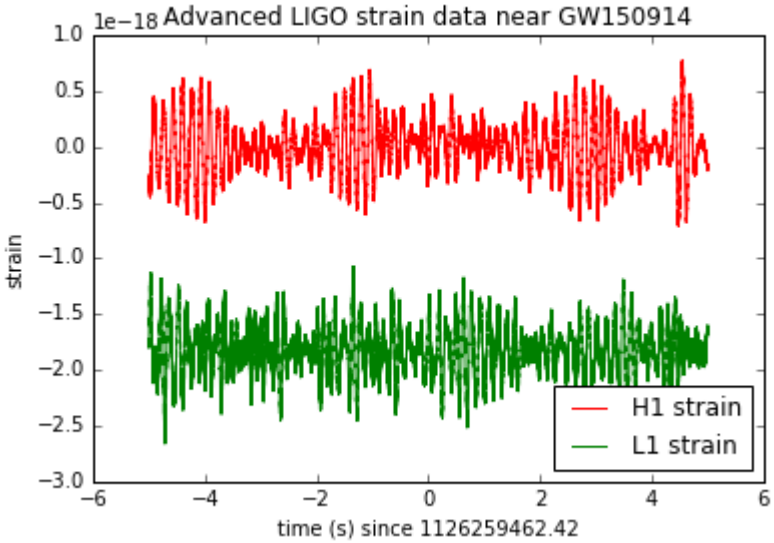
Temporal Data Mining

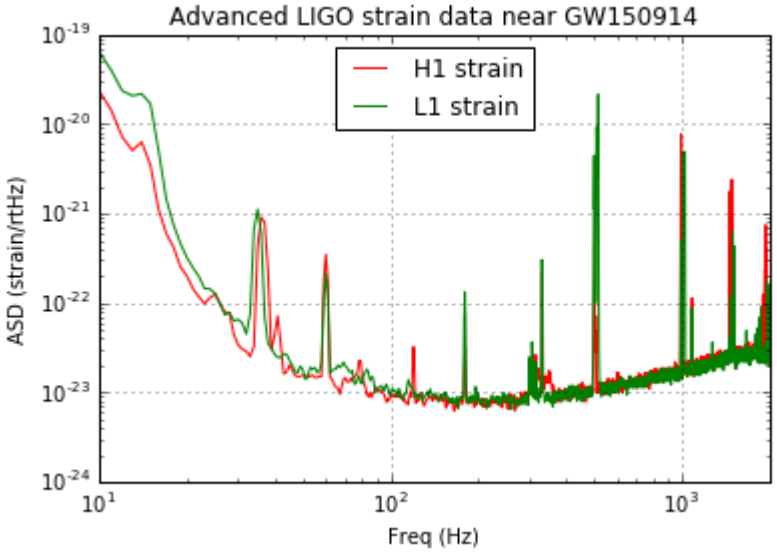
- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Mhealth Dataset
- Our results: Dow Jones Dataset
- Our results: Robot Execution Failure Dataset
- **Case study: LIGO Gravitational Wave detection**
- Conclusions
- Future Work

LIGO Gravitational Wave detection

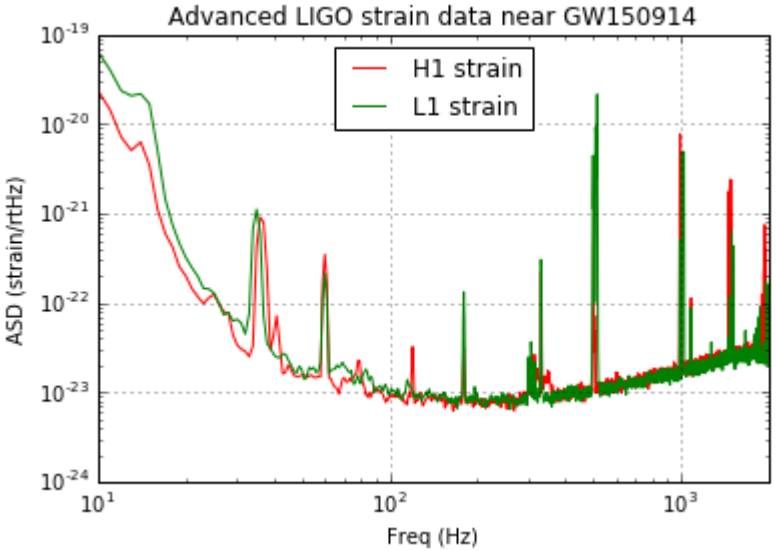
- Dataset and code open-sourced by the LIGO observatory
- First direct detection of gravitational waves
- There is no learning in this dataset (at least as of now!)
    - Since this was the first detection, we actually just got our first training sample!
- Pre-processing and Digital Signal processing techniques
- Data sampled at 16384 and 4096 Hertz.
- Readings from 2 different sensors (H and L)

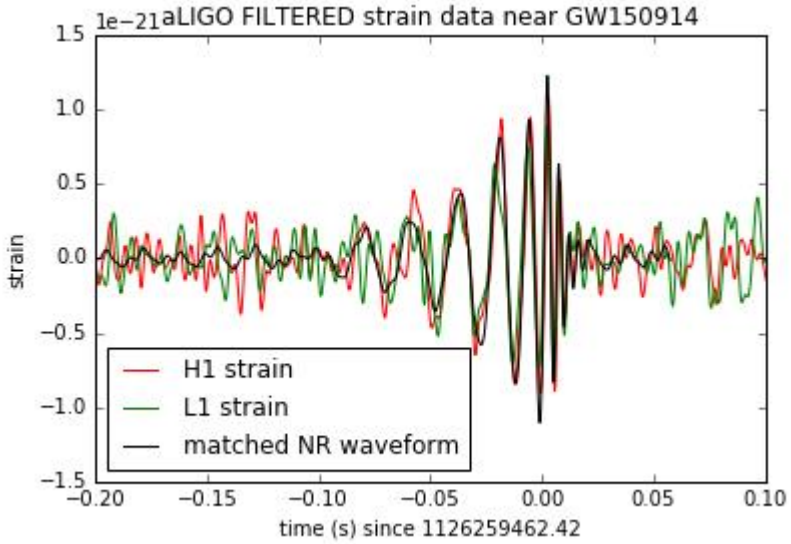
Signals in time and frequency domain



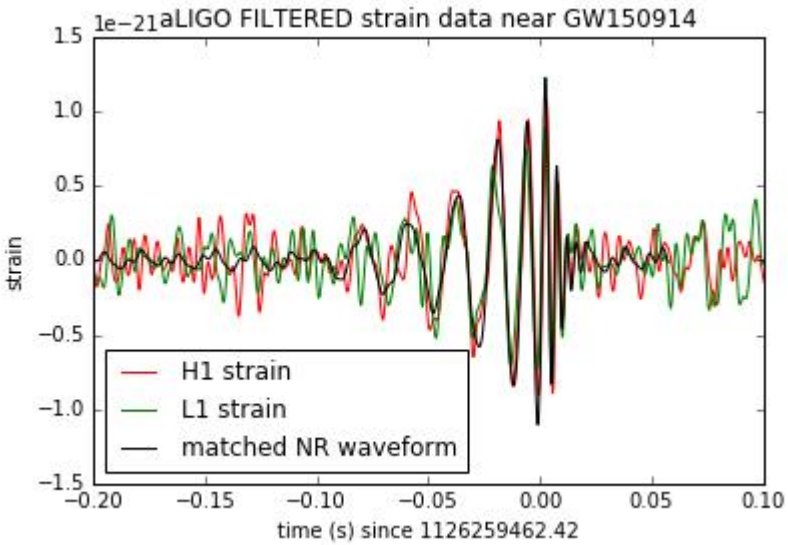


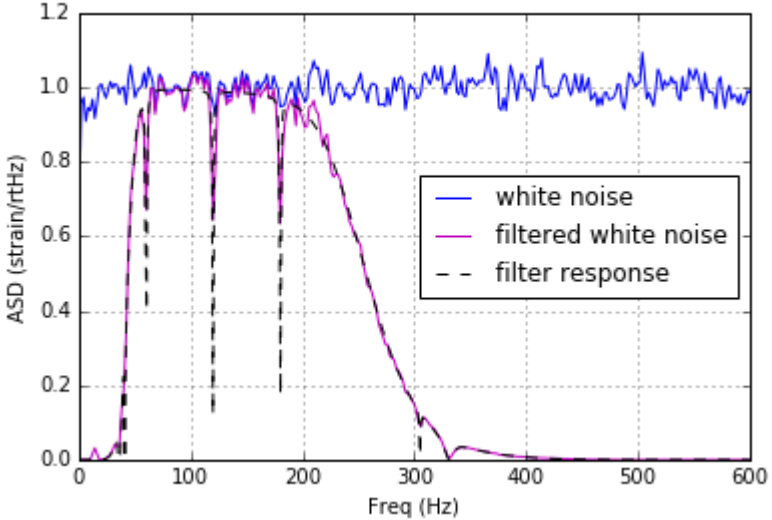
Signal Whitening



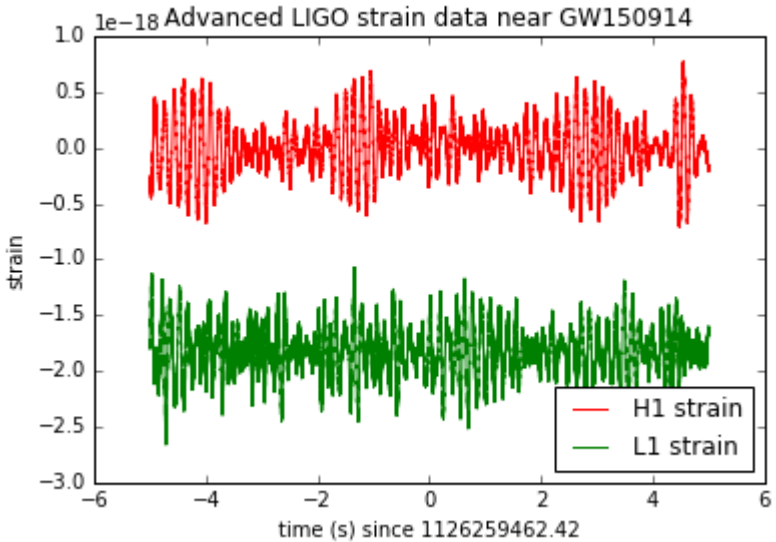


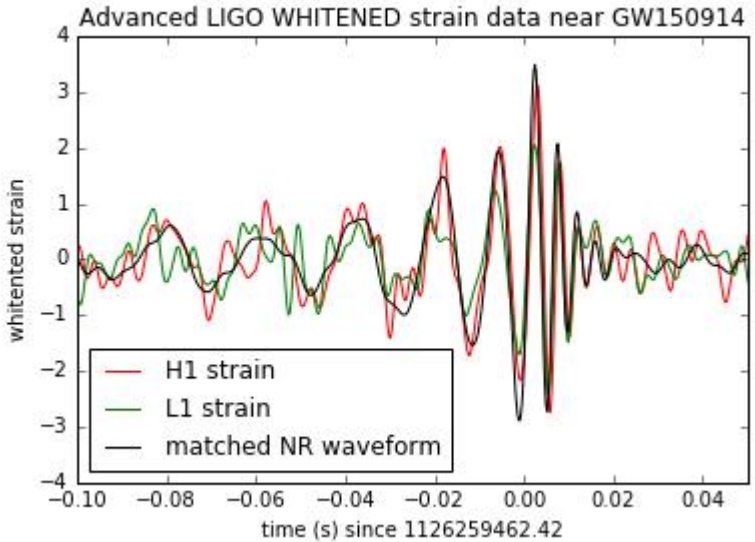
Band passing (40 to 300 Hz)





The GW signal





Sound file

Temporal Data Mining

- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Mhealth Dataset
- Our results: Dow Jones Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- **Conclusions**
- Future Work

Conclusions

- R seems like a better option for time-series data
- The behavior of the “target” feature might be dependent on external factors not included in the dataset.
    - Example: Tesla’s stock price fell almost 30% despite the fact that all their products are doing really well. Reason: Decrease in oil price.
- Finding the correct order for ARMA model requires more practice (and better understanding)

Temporal Data Mining

- What is Time Series Data?
- Components: Trends, Seasonal and residual
- Decomposition: Additive and Multiplicative
- Techniques: AR, MA, AR+MA
- Our results: Mhealth Dataset
- Our results: Dow Jones Dataset
- Our results: Robot Execution Failure Dataset
- Case study: LIGO Gravitational Wave detection
- Conclusions
- **Future Work**

Future Work

- For Dow Jones, the stock price in next week can depend on external factors like some recent big announcement from the company which might make people buy more stock of that company and result in greater spike in stock prices compared to what we obtained from just trend analysis. So, we can add a “company’s repo” kind of feature which indicates company’s reputation in media based on current event.
- For Mhealth and robot, we can integrate the code we have developed with a sensor that provides real time sensor readings. Example: Google Fit, autonomous cars

Time series analysis: DIY

- We will be uploading the ipython notebooks for the 3 datasets on github and share the link on course ecommons site by tomorrow
- Please feel free to try it out yourself and provide feedback
- The ipython notebooks for Gravitational wave detection is already available online

References

- Wikipedia

- https://www.otexts.org/fpp

- http://www.stats.ox.ac.uk/~burke/Autocorrelation/Time%20Series%20Graphs.pdf
- http://nbviewer.jupyter.org/github/bashtage/arch/blob/master/examples/univariate_volatility_modeling.ipynb#Specifying-Common-Models
- http://www.quantatrisk.com/2014/10/23/garch11-model-in-python/
- https://onlinecourses.science.psu.edu/stat510/node/62
- http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/tsa_arma.html
- http://cims.nyu.edu/~almgren/timeseries/Vol_Forecast1.pdf
- http://www.rigtorp.se/2011/01/01/rolling-statistics-numpy.html
- http://bicorner.com/2015/11/16/time-series-analysis-using-ipython/

Thanks to:

- Jay Pujara
- Ryan Hausen
- Keshav Mathur

Questions?