Application of Big Data Analytics via Soft Computing

UTSA
ELECTRICAL & COMPUTER
ENGINEERING

Yunus Yetis

TECHLAV

INTRODUCTION

- System of Systems (SoS) and cyberphysic are integrated, independently operating systems working in a cooperative mode to achieve a higher performance.

- SoSs are generating "Big Data" which makes modeling of such complex systems a challenge indeed

- Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications.

What is BIG DATA?

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

- The challenges include capture, storage, search, sharing, transfer, analysis, and visualization.

- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data.

What is BIG DATA?

Air Bus A380
- 1 billion line of code
- each engine generate 10 TB every 30 min

640TB per Flight

Twitter Generate approximately 12 TB of data per day

New York Stock Exchange 1TB of data everyday

storage capacity has doubled roughly every three years since the 1980s

How big is the Big Data?

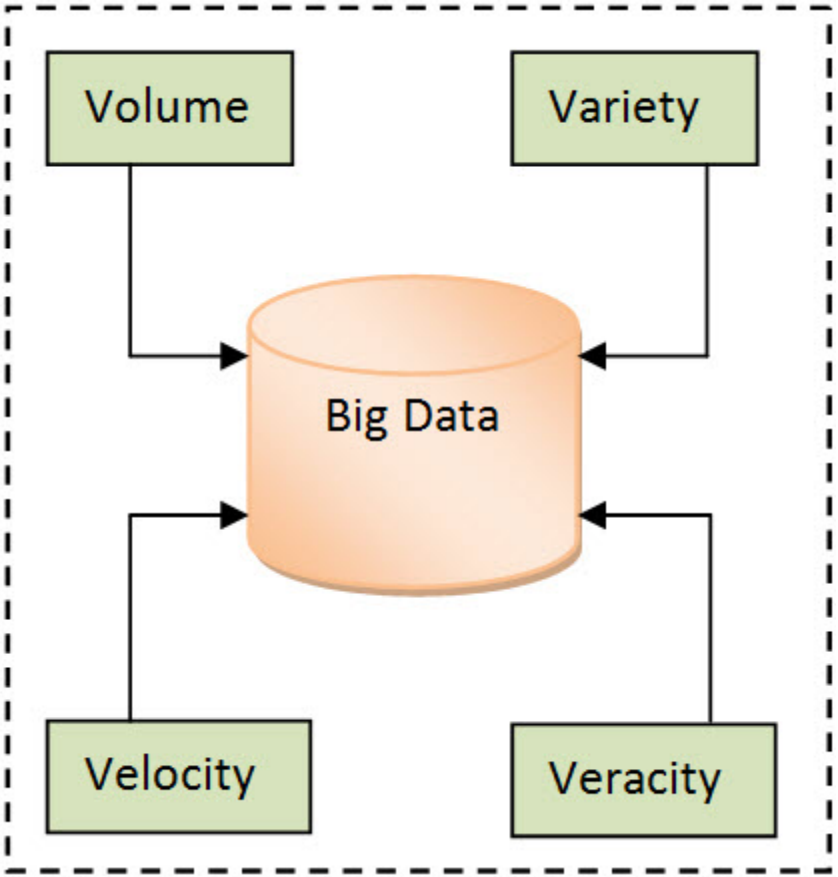
- What is big today maybe not big tomorrow

- Any data that can challenge our current technology in some manner can consider as Big Data
  - Volume
  - Communication
  - Speed of Generating
  - Meaningful Analysis

ELECTRICAL & COMPUTER ENGINEERING

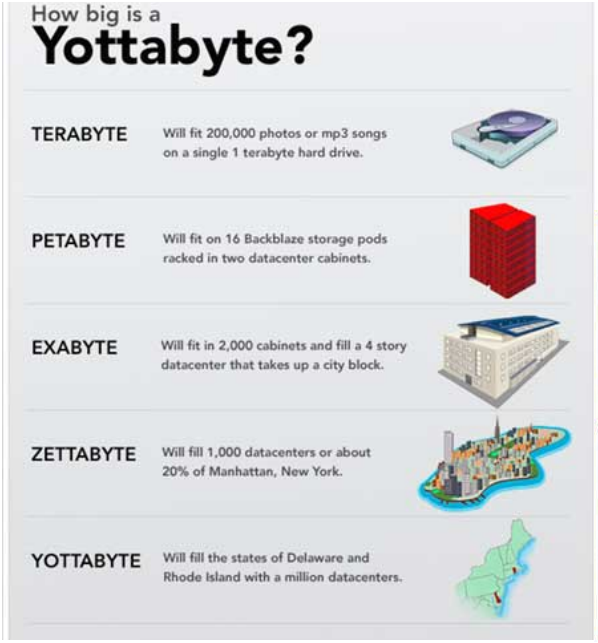
Big data can be described by the following characteristics

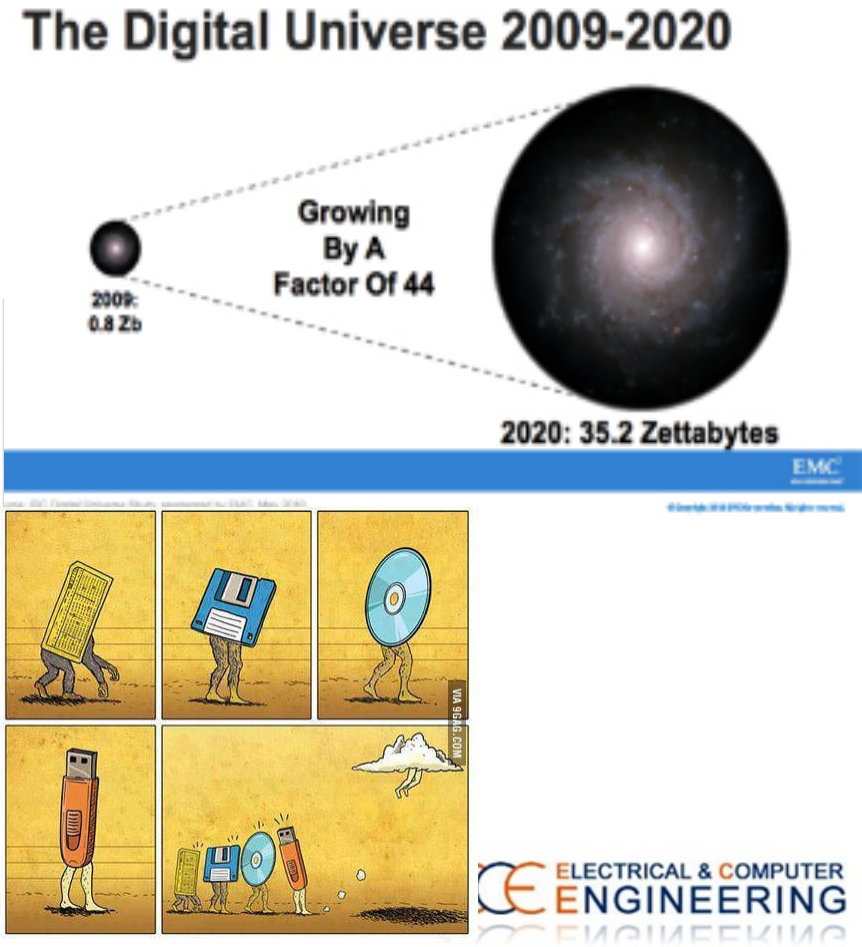
- Volume
- Variety
- Velocity



Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 to 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





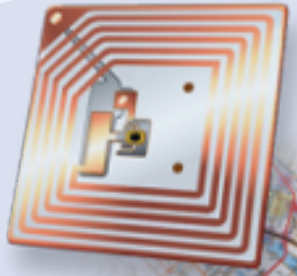
ELECTRICAL & COMPUTER ENGINEERING

12+ TBs **of tweet data every day**

? TBs of data every day

25+ TBs **of log data every day**

30 billion RFID tags today (1.3B in 2005)



4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

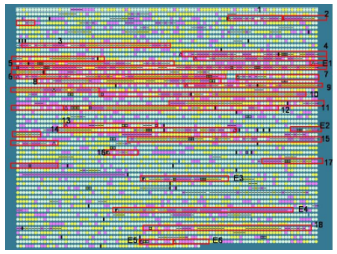
76 million smart meters in 2009… 200M by 2014

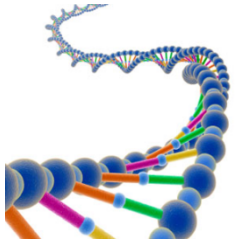
2+ billion people on the Web by end 2011

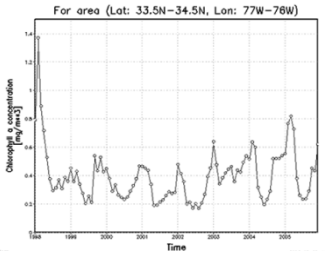
ELECTRICAL & COMPUTER ENGINEERING

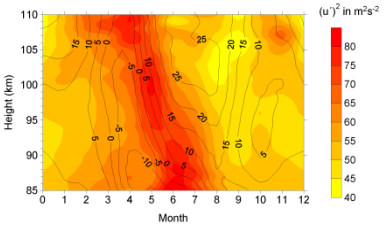
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network,
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)









ELECTRICAL & COMPUTER ENGINEERING

Velocity (Speed)

- Data is generated fast and need to be processed fast

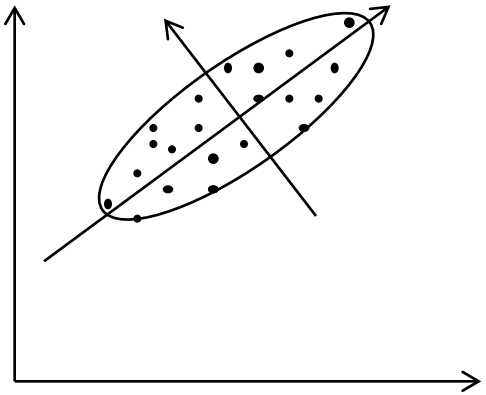
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Brief Description of Machine Learning

- *Principal Component Analysis (PCA)*
- *Artificial Neural Networks (ANN)*
- *Genetic Algorithm*

Principal Component Analysis

- Eigen Vectors show the direction of axes of a fitted ellipsoid
- Eigen Values show the significance of the corresponding axis
- The larger the Eigen value, the more separation between mapped data
- For high dimensional data, only few of Eigen values are significant



- Finding Eigen Values and Eigen Vectors
- Deciding on which are significant
- Forming a new coordinate system defined by the significant Eigen vectors

  (→lower dimensions for new coordinates)
- Mapping data to the new space

→Compressed Data

Case study: Principal Component Analysis (PCA)

PCA is used abundantly in all forms of analysis because it is a simple, non-parametric method of extracting relevant information from confusing data sets.

PCA provides us a roadmap for how to reduce a complex data set to a lower dimension to save time and data storage.

It covers standard deviation, covariance, eigenvectors and eigenvalues.

First, it is the optimal (in terms of mse) linear scheme for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing

Second, the model parameters(covariance, eigenvectors and eigenvalues) can be computed directly from the data.

Another approaches to PCA is that it is not obvious how to deal properly with incomplete data set, in which some of the points are missing.

station	valid	(GMT	timezone	Air Temperature	Humidity in %	Wind Direction	Wind speed	Pressure altimeter	Sea Level Pressure	Sky level coverage	Sky level Altitide
IOW	12/10/2012	13:52	21.02	77.45	300	16	29.93	1014.4	0	1400	M
IOW	12/10/2012	14:52	19.94	81.09	290	13.7	29.95	1015.3	0	1600	M
IOW	12/10/2012	15:52	19.94	77.35	300	12.5	29.96	1015.6	0	1600	3500
IOW	12/10/2012	16:20	21.2	79.31	300	11.4	29.96	M	0	1600	3500
IOW	12/10/2012	16:52	21.92	74.56	310	10.3	29.96	1015.5	0	3500	M
IOW	12/10/2012	17:13	23	73.51	300	11.4	29.95	M	0	1600	3700
IOW	12/10/2012	17:52	24.08	70.81	310	11.4	29.94	1014.9	0	1600	M
IOW	12/10/2012	18:09	24.8	68.18	300	13.7	29.94	M	0	1600	4000
IOW	12/10/2012	18:45	24.8	68.18	310	12.5	29.94	M	0	2900	4000
IOW	12/10/2012	18:52	24.08	70.81	300	12.5	29.94	1014.6	0	2900	4000
IOW	12/10/2012	19:52	24.98	71.47	310	12.5	29.93	1014.5	0	2900	M
IOW	12/10/2012	20:20	24.8	73.7	330	12.5	29.93	M	0	3100	M
IOW	12/10/2012	20:52	26.06	71.04	300	12.5	29.93	1014.4	0	1700	3100
IOW	12/10/2012	21:02	26.6	73.89	320	11.4	29.93	M	0	1700	M
IOW	12/10/2012	21:52	26.06	74.41	320	12.5	29.95	1015	0	1700	M
IOW	12/10/2012	22:13	24.8	79.62	310	8	29.95	M	0	1700	4000
IOW	12/10/2012	22:52	24.98	77.82	320	8	29.96	1015.4	0	4000	M

Problem Statement

- Create Neural Network to Wind Speed Prediction using large datasets which includes pattern of wind speed.
- We have been encountered some issues;

1. The datasets sometimes may have missing values like wind datasets.

2. Analyzing of large datasets take much time.

3. Error and results are not stable because of that initial weights are randomly chosen, with typical values between -1.0 and 1.0 in Neural Network structure.

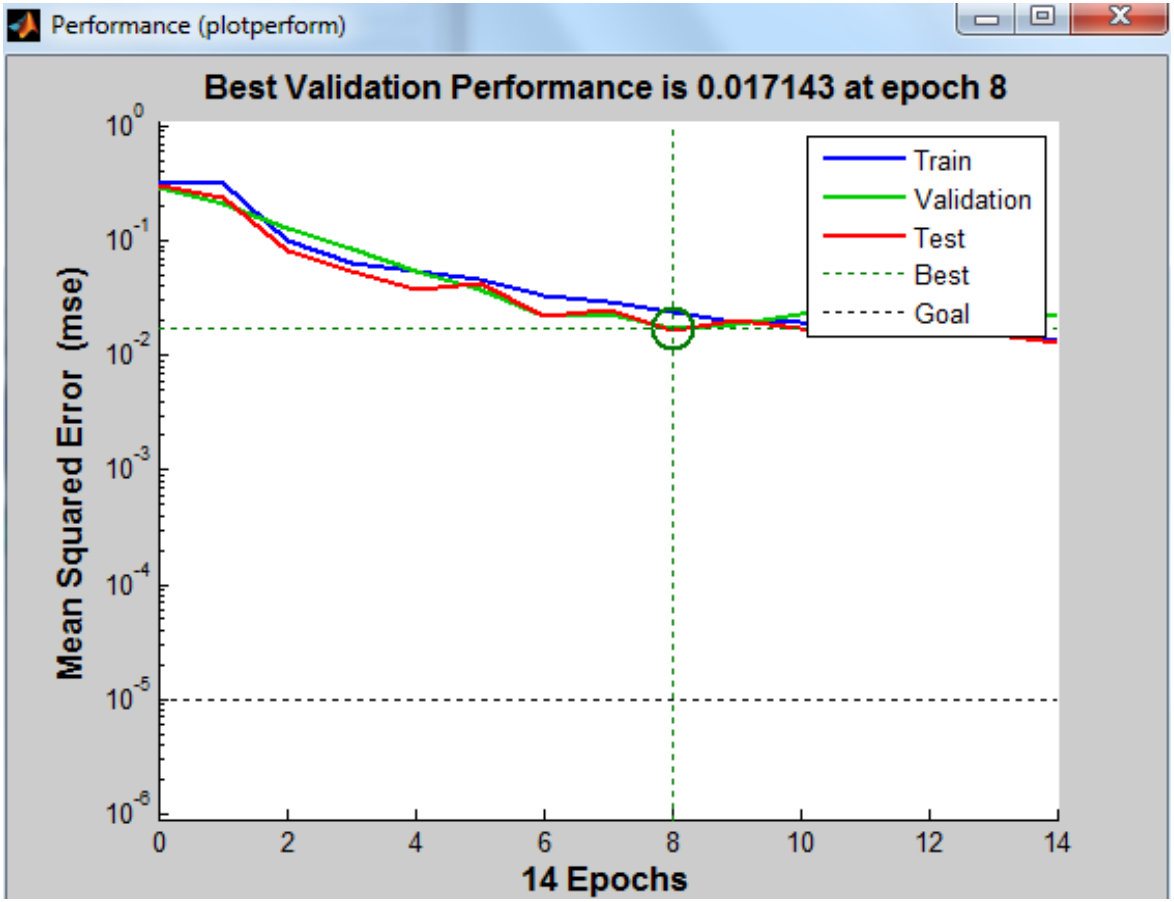
Solution and Implementation

- Creating Neural network and PCA toolbox to get less error.
  - **Output** is Wind Speed
  - **Inputs** are;

Air temperature
Humidity
Wind direction
Pressure altimeter
Sea Level Pressure
Sky Level Coverage
Sky Level Altitude
Time Zone

http://mesonet.agron.iastate.edu/request/download.phtml?network=TR_ASOS

- Check error before trying to correct (Without PCA)



There is missing values and weights are randomly chosen, it looks worst results

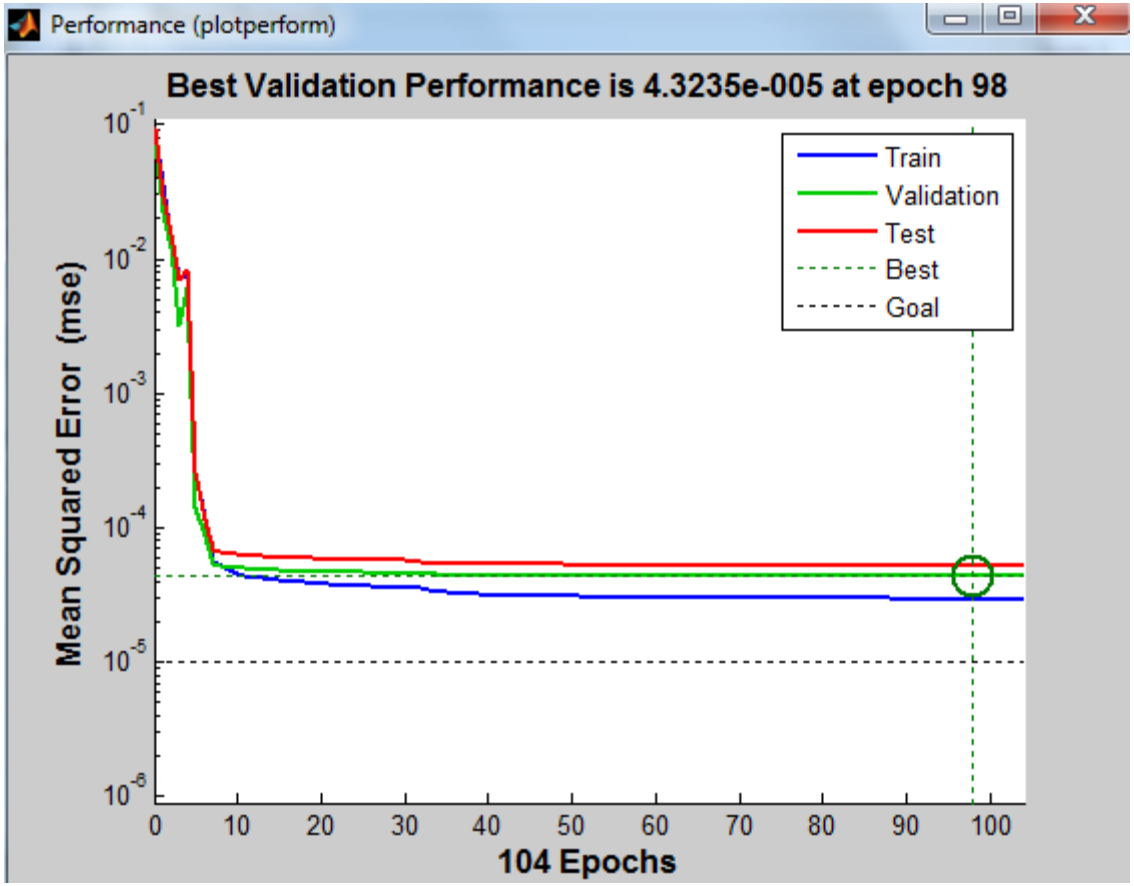
PCA using ALS for Missing data

station	valid	(GMT	timezone	Air Temperature	Humidity in %	Wind Direction	Wind speed	Pressure altimeter	Sea Level Pressure	Sky level cover age	Sky level Altitide
IOW	12/10/2012	13:52	21.02	77.45	300	16	29.93	1014.4	0	1400	M
IOW	12/10/2012	14:52	19.94	81.09	290	13.7	29.95	1015.3	0	1600	M
IOW	12/10/2012	15:52	19.94	77.35	300	12.5	29.96	1015.6	0	1600	3500
IOW	12/10/2012	16:20	21.2	79.31	300	11.4	29.96	M	0	1600	3500
IOW	12/10/2012	16:52	21.92	74.56	310	10.3	29.96	1015.5	0	3500	M

When there are missing values in the data,find the principal components using the alternating least squares (ALS) algorithm.

Then reconstruct data matrix without Missing value

PCA using for Missing data

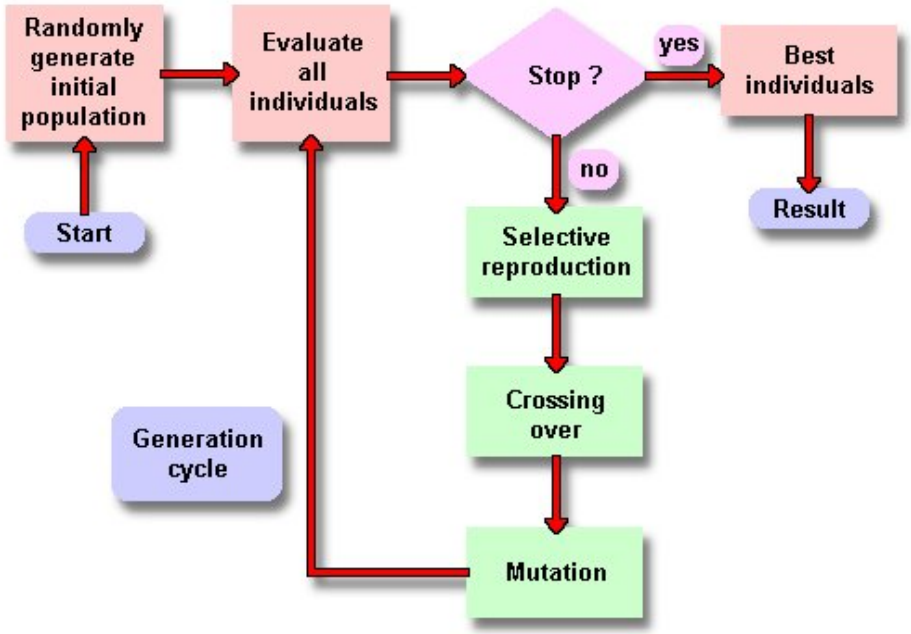


Results

- It is necessary to get rid of missing value while we are forecasting with large datasets.
- Preprocessing with PCA is very important to get less error(4.323e-005<< 0.01714).

Genetic Algorithm

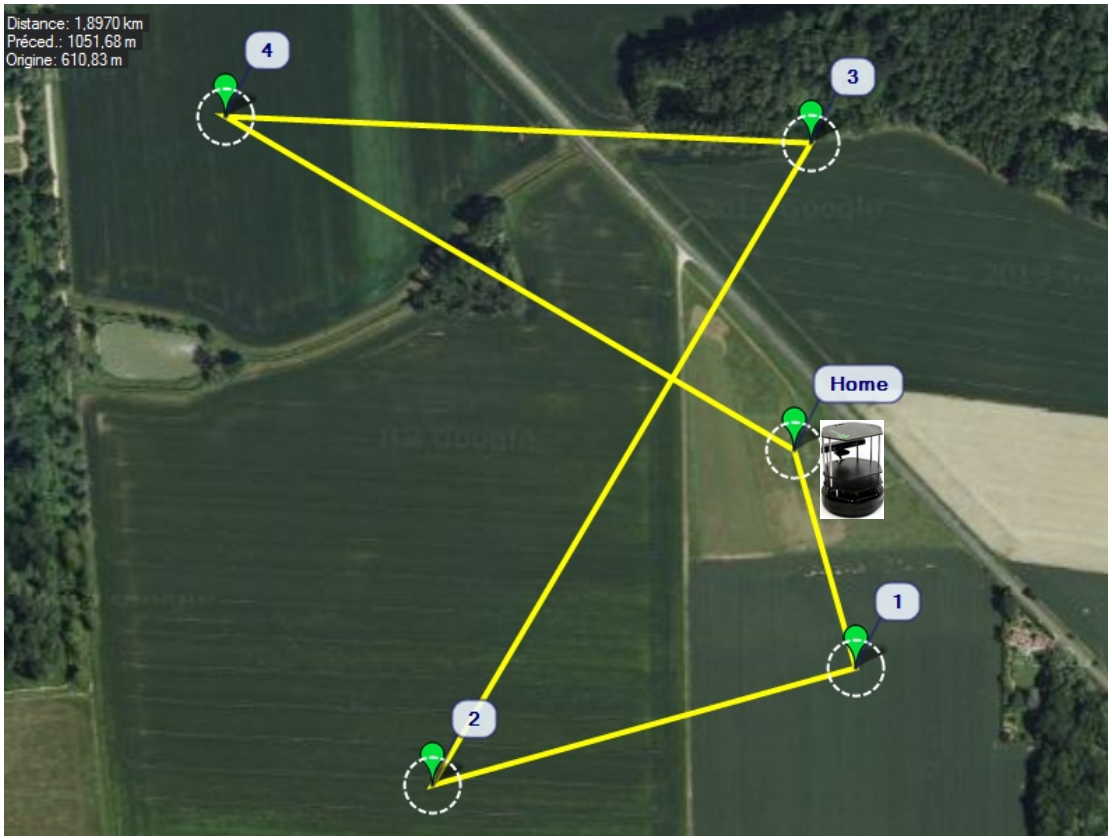
• It is started with a set of randomly generated solutions and recombine pairs of them at random to produce offspring.
• Only the best offspring and parents are kept to produce the next generation



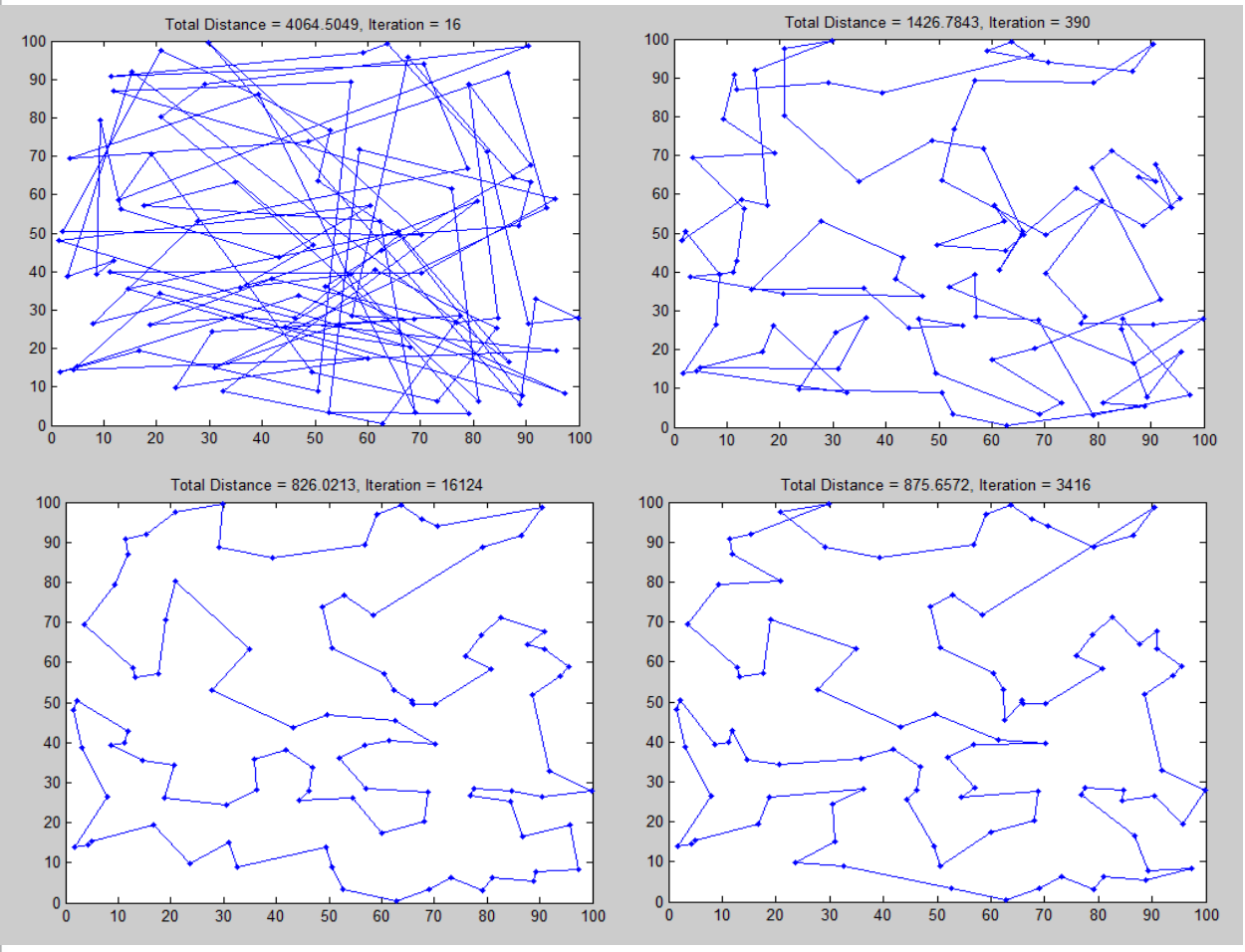
Applications

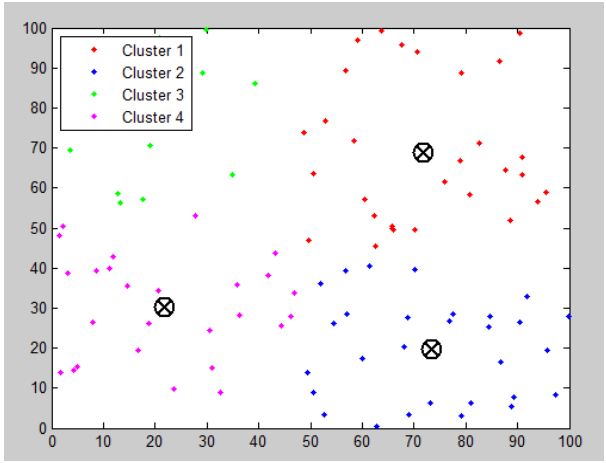
- Design of water distribution systems.
- Distributed computer network topologies.
- Electronic circuit design, known as Evolvable hardware.
- File allocation for a distributed system
- Mobile communications infrastructure optimization

Genetic Algorithm

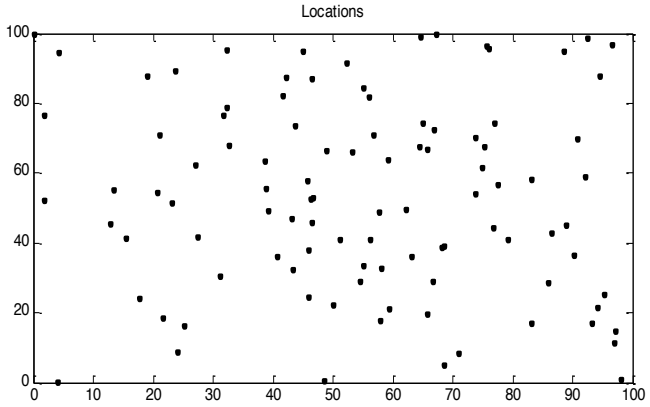


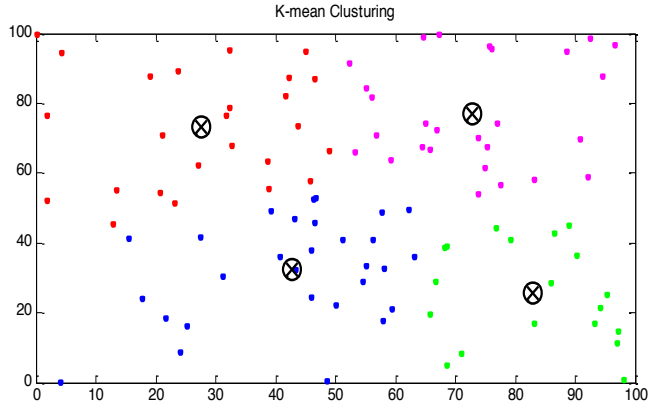
Ref: https://github.com/jlnaudin/x-drone/wiki/x-drone:-MaxiSwift,-mission-35---comparison-of-FPL-path-of-Real-flight-Vs-HIL-simulation

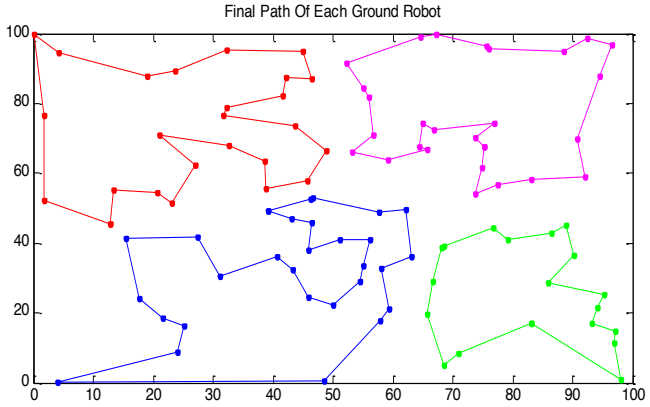


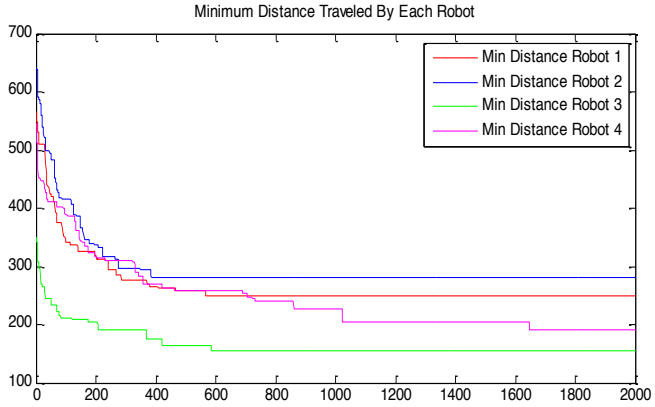


ELECTRICAL & COMPUTER ENGINEERING



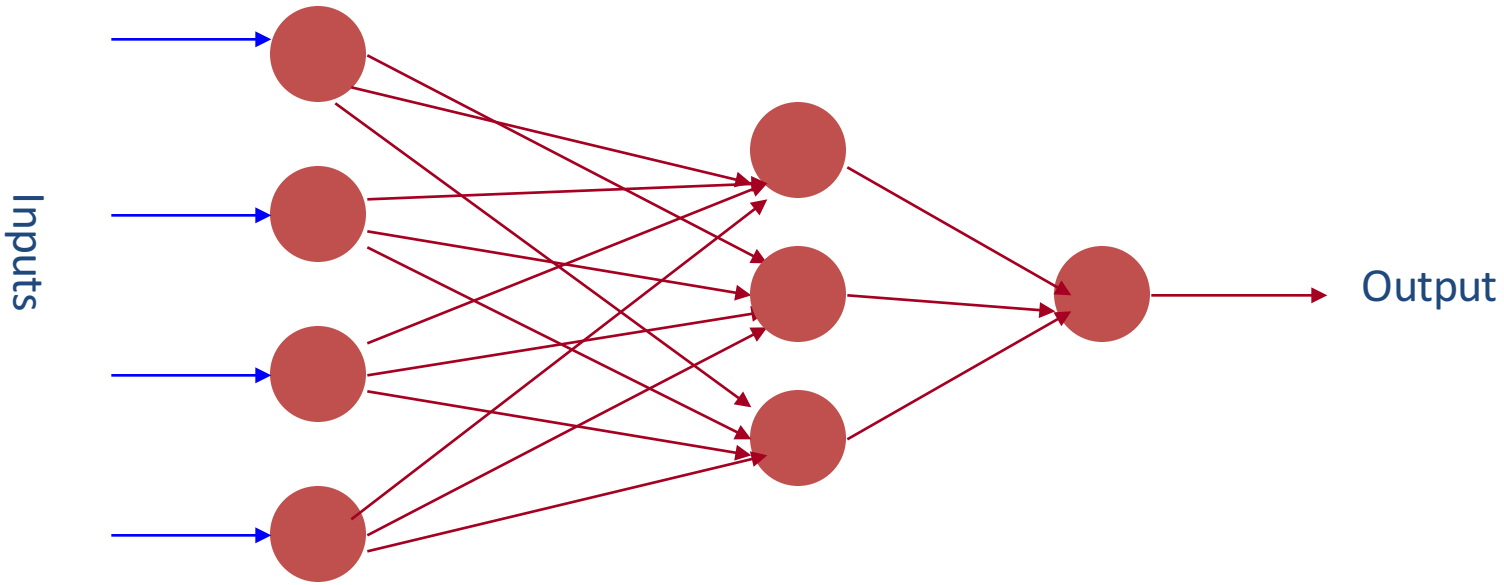






ELECTRICAL & COMPUTER ENGINEERING

Artificial Neural Network



An artificial neural network is composed of many artificial neurons that are linked together according to a specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.

Tasks to be solved by artificial neural networks:

- controlling the movements of a robot based on self-perception and other information (e.g., visual information);
- deciding the category of potential food items (e.g., edible or non-edible) in an artificial world;
- recognizing a visual object (e.g., a familiar face);
- predicting where a moving object goes, when a robot wants to catch it.

Neural network tasks

- control
- classification
- prediction
- approximation

These can be reformulated in general as

FUNCTION APPROXIMATION

tasks.

Approximation: given a set of values of a function g(x) build a neural network that approximates the g(x) values for any input x.

Artificial Neural Network

Problem Statement

- To develop a graphical user interface which given the open price, high, low, volume of the day and the previous day's closing price; outputs the estimated closing price of the day based on the previous data.

- Collect amount of historical stock data

- Using this data, train a neural network

- Once trained, the neural network can be used to predict stock behavior

- Need to some way to gauge value of results – we will compare with www.finance.yahoo.com as well as compare with what actually happened

Advantages & Disadvantages

✓ ***Advantages***

>> Neural network can be trained with a very large amount of data. Years, decades, even centuries

>> Able to consider a "lifetime" worth of data when making a prediction

>> Completely unbiased

✓ ***Disadvantages***

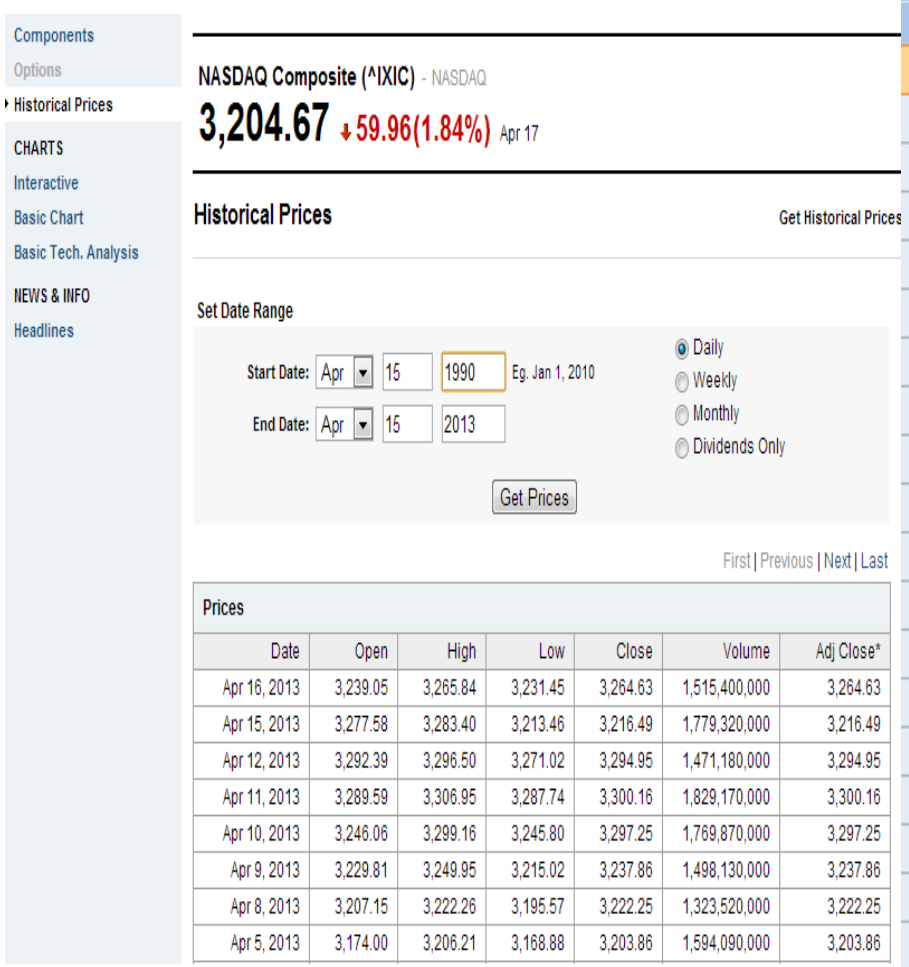
>> No way to predict unexpected factors, i.e. natural disaster, legal problems, etc.

✓ Neural networks are used to predict stock market prices because they are able to learn nonlinear mappings between inputs and outputs.

✓ Several researchers claim the stock market and other complex systems exhibit chaos.

✓ With the neural networks' ability to learn nonlinear, chaotic systems, it may be possible to outperform traditional analysis and other computer-based methods.

Download the Spreadsheet from http://finance.yahoo.com/q/hp?s=%5EIXIC+Historical+Prices



Components
Options
Historical Prices
CHARTS
Interactive
Basic Chart
Basic Tech. Analysis
NEWS & INFO
Headlines

NASDAQ Composite (^IXIC) - NASDAQ
3,204.67 ↓59.96(1.84%) Apr 17

Historical Prices

Get Historical Prices

Set Date Range

Start Date: Apr 15 1990 Eg. Jan 1, 2010
End Date: Apr 15 2013

Daily
Weekly
Monthly
Dividends Only

Get Prices

First | Previous | Next | Last

Prices

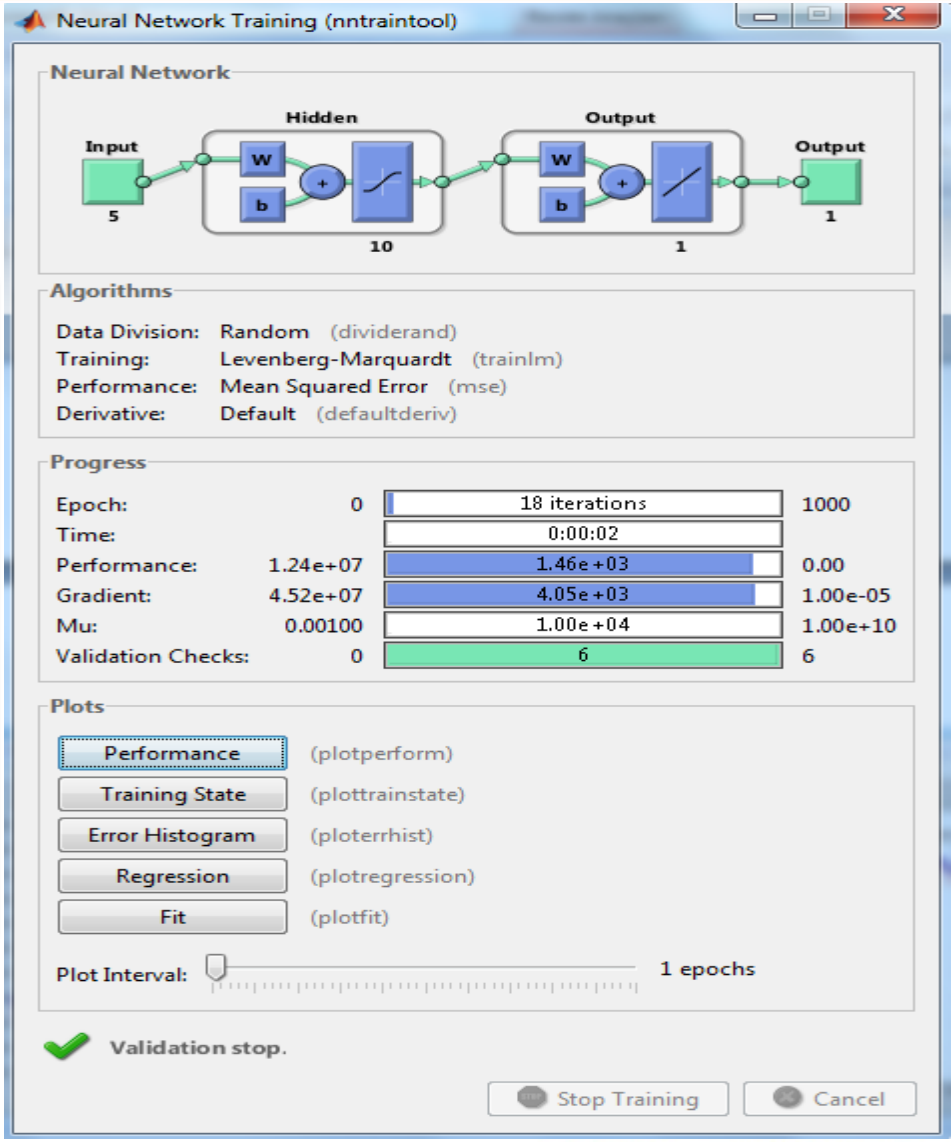
Date	Open	High	Low	Close	Volume	Adj Close*
Apr 16, 2013	3,239.05	3,265.84	3,231.45	3,264.63	1,515,400,000	3,264.63
Apr 15, 2013	3,277.58	3,283.40	3,213.46	3,216.49	1,779,320,000	3,216.49
Apr 12, 2013	3,292.39	3,296.50	3,271.02	3,294.95	1,471,180,000	3,294.95
Apr 11, 2013	3,289.59	3,306.95	3,287.74	3,300.16	1,829,170,000	3,300.16
Apr 10, 2013	3,246.06	3,299.16	3,245.80	3,297.25	1,769,870,000	3,297.25
Apr 9, 2013	3,229.81	3,249.95	3,215.02	3,237.86	1,498,130,000	3,237.86
Apr 8, 2013	3,207.15	3,222.26	3,195.57	3,222.25	1,323,520,000	3,222.25
Apr 5, 2013	3,174.00	3,206.21	3,168.88	3,203.86	1,594,090,000	3,203.86

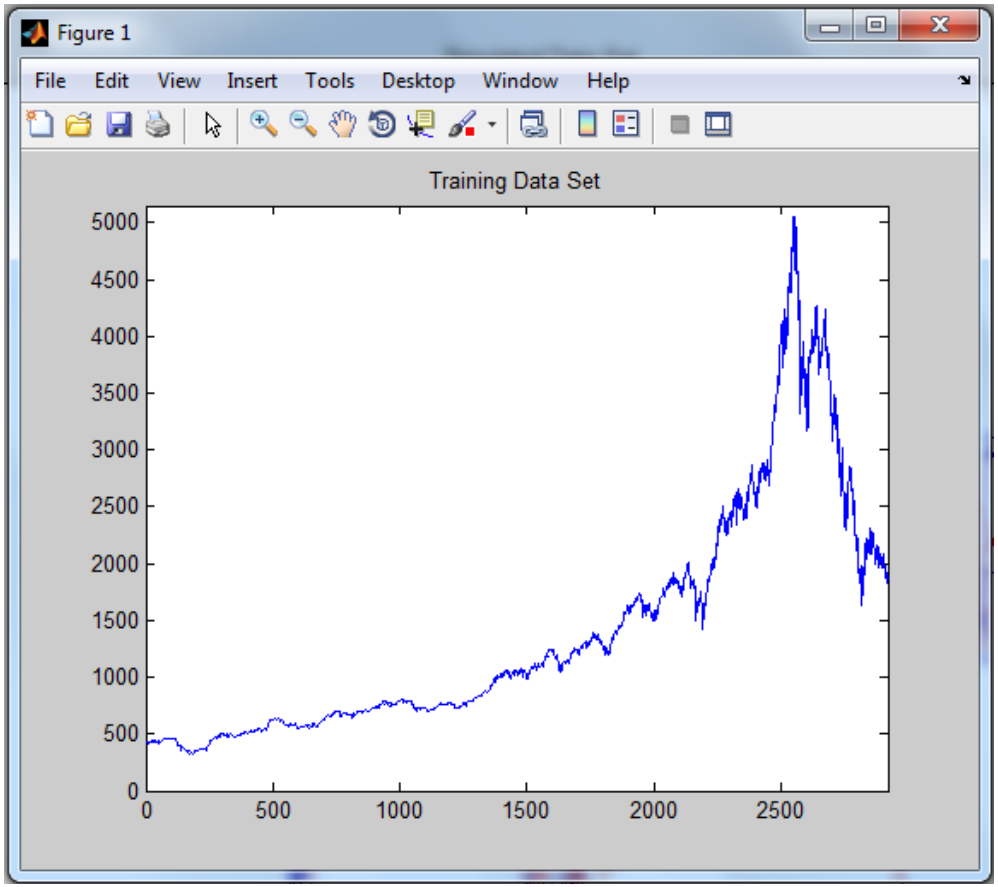
	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Volume	Adj Clo
2	4/15/2013	3277.58	3283.4	3213.46	3216.49	1779320000	3216.49
3	4/12/2013	3292.39	3296.5	3271.02	3294.95	1471180000	3294.95
4	4/11/2013	3289.59	3306.95	3287.74	3300.16	1829170000	3300.16
5	4/10/2013	3246.06	3299.16	3245.8	3297.25	1769870000	3297.25
6	4/9/2013	3229.81	3249.95	3215.02	3237.86	1498130000	3237.86
7	4/8/2013	3207.15	3222.26	3195.57	3222.25	1323520000	3222.25
8	4/5/2013	3174	3206.21	3168.88	3203.86	1594090000	3203.86
9	4/4/2013	3219.11	3226.24	3206.02	3224.98	1475720000	3224.98
10	4/3/2013	3257.38	3260.15	3210.39	3218.6	1813910000	3218.6
11	4/2/2013	3252.55	3267.93	3245.41	3254.86	1580800000	3254.86
12	4/1/2013	3268.63	3270.23	3230.57	3239.17	1481360000	3239.17
13	3/28/2013	3257.32	3270.3	3253.21	3267.52	1636800000	3267.52
14	3/27/2013	3230.76	3258.26	3227.02	3256.52	1420130000	3256.52
15	3/26/2013	3249.95	3252.93	3239.92	3252.48	1444500000	3252.48
16	3/25/2013	3255.85	3263.63	3222.48	3235.3	1666010000	3235.3
17	3/22/2013	3235.3	3247.94	3230.86	3245	1681360000	3245
18	3/21/2013	3228.17	3237.57	3215.69	3222.6	1692260000	3222.6
19	3/20/2013	3251.91	3257.99	3240.9	3254.19	1599120000	3254.19

ELECTRICAL & COMPUTER ENGINEERING

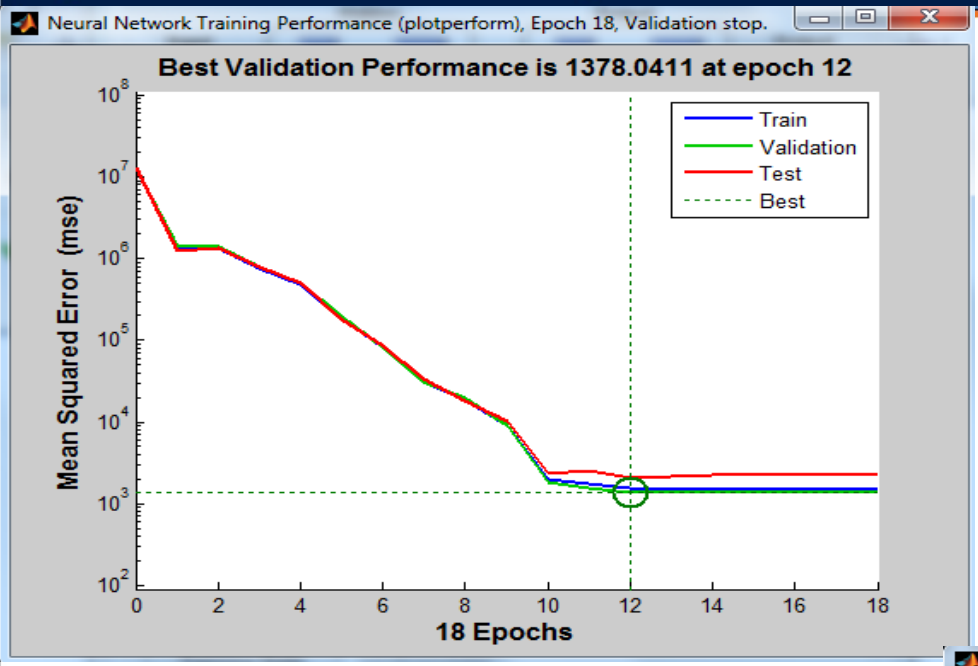
Backpropagation is the process of backpropagating errors through the system from the output layer towards the input layer during training.
Backpropagation is necessary because hidden units have no training target value that can be used, so they must be trained based on errors from previous layers.

The output layer is the only layer which has a target value for which to compare.





With these settings, the input vectors and target vectors will be randomly divided into three sets as follows:
70% will be used for training.
15% will be used to validate that the network is generalizing and to stop training before overfitting.
The last 15% will be used as a completely independent test of network generalization.



The result is reasonable because of the following considerations:
The train set error , the validation set error and test set error have similar characteristics.

Error histogram to obtain additional verification of network performance.

You can see that while most errors fall between -120 and 100.

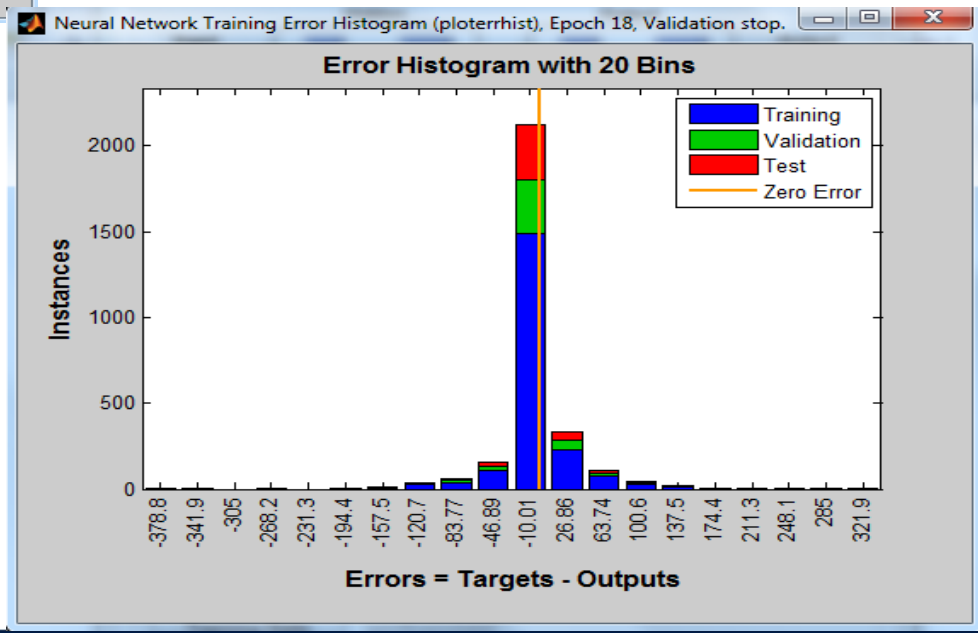


Figure 2

File Edit View Insert Tools Desktop Window Help

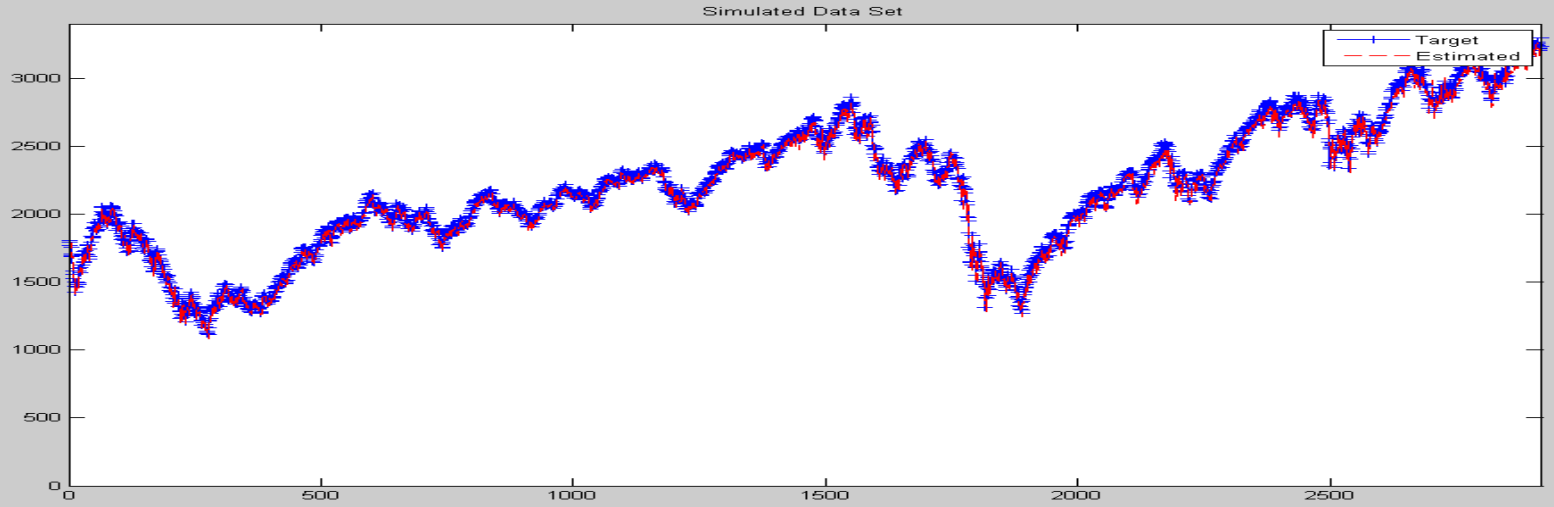
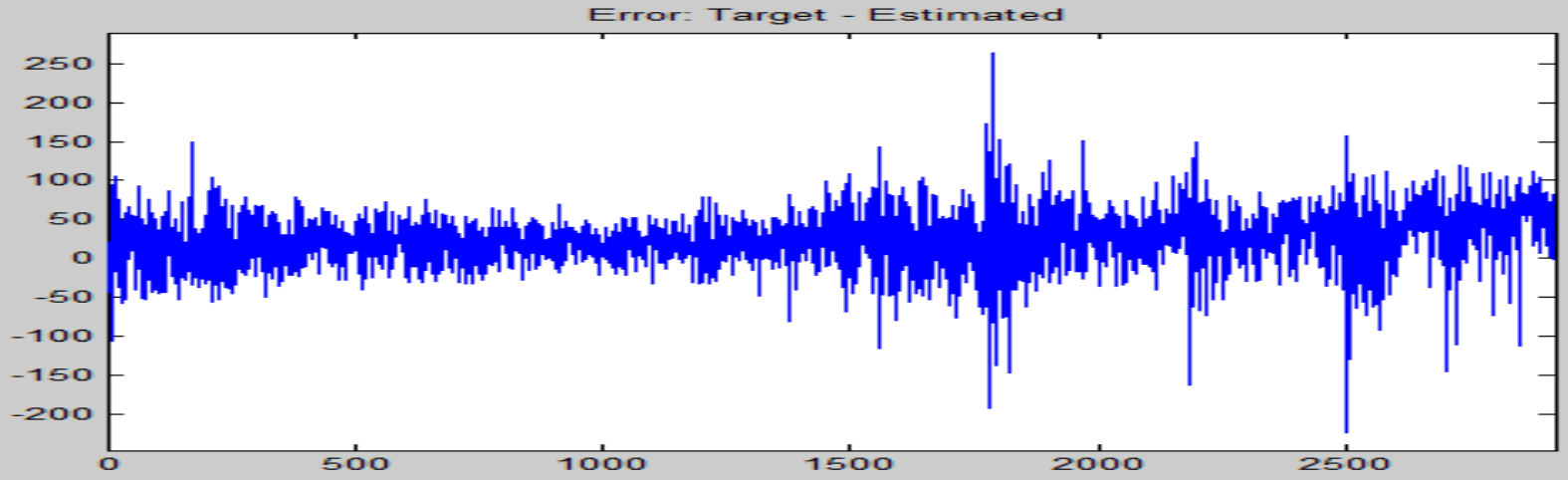


Figure 3

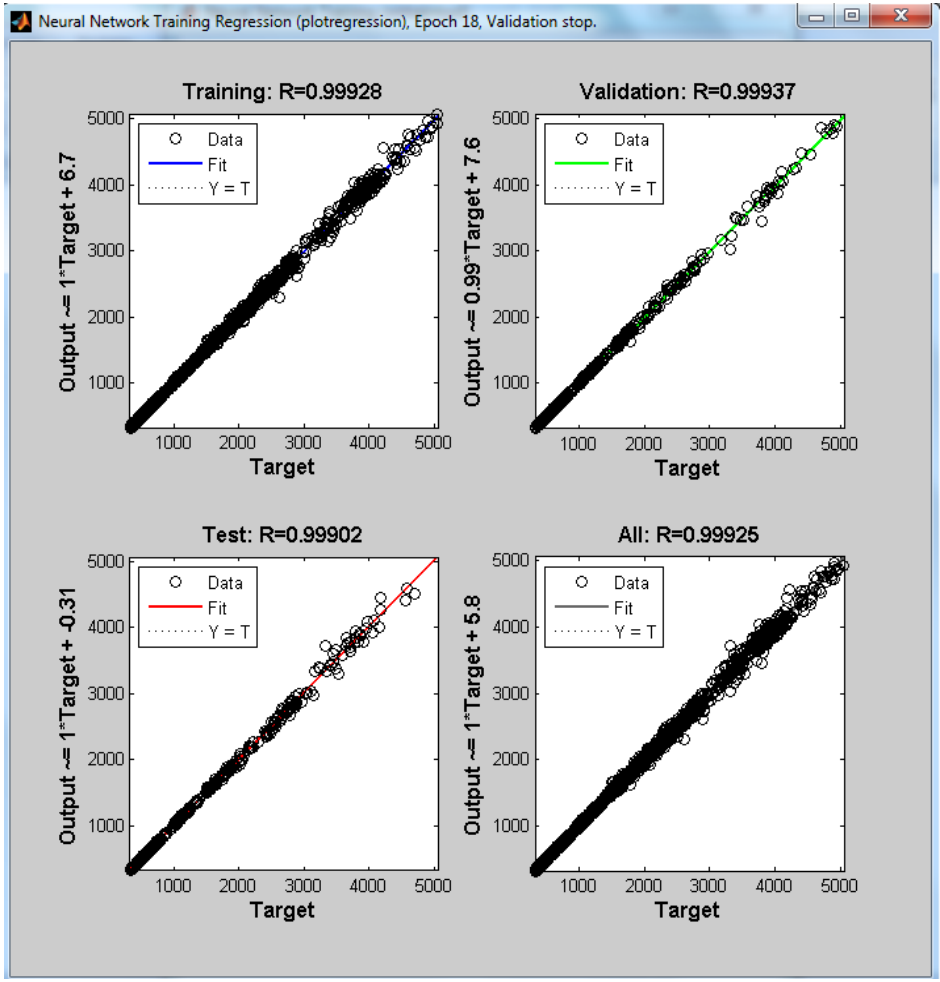
File Edit View Insert Tools Desktop Window Help



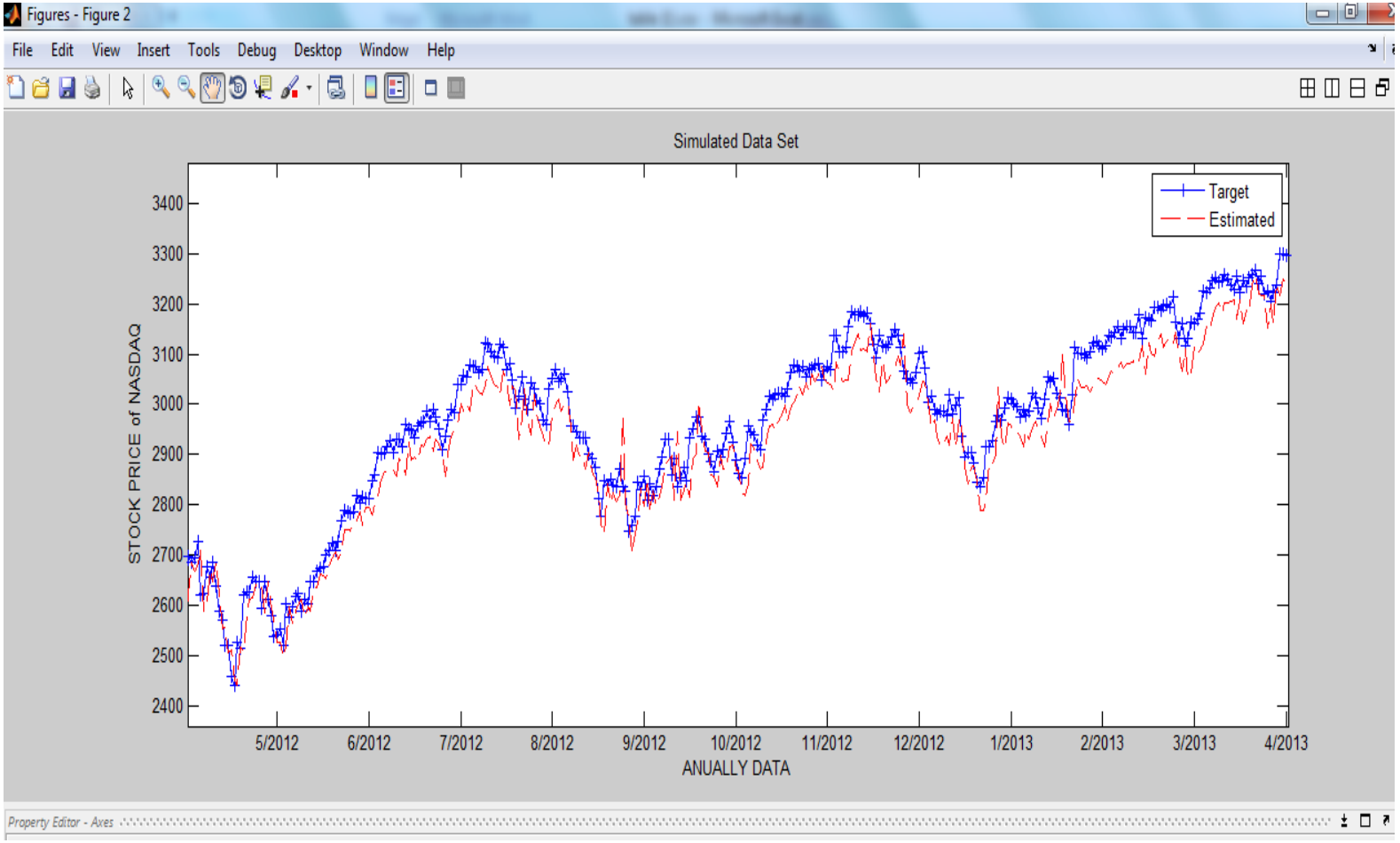
Regression is used to validate the network performance.
The following regression plots display the network outputs with respect to targets for training, validation, and test sets.

For a perfect fit, the data should fall along a 45 degree line, where the network outputs are equal to the targets.

For this problem, the fit is reasonably good for all data sets, with R values in each case of 0.99 or above.



VISUALIZATIONS



Conclusion

✓ Our model shows promise, but needs improvement before becoming an effective aid.

Needs more data, possibly more types of data

✓ No human or computer can perfectly predict the volatile stock market

✓ Under "normal" conditions, in most cases, a good neural network will outperform most other current stock market predictors and be a very worthwhile, and potentially profitable aid to investors

References

[1] M. Jamshidi (ed.), *Systems of Systems Engineering—Principles and Applications* (CRC/Taylor & Francis, London, 2008) (also in Mandarin language, China Machine Press, ISBN 978-7- 111-38955-2, Beijing, 2013)

[2] M. Jamshidi (ed.), *System of Systems Engineering—Innovations for the 21st Century* (Wiley, NewYork, 2009)

[3] Jamshidi, Mo, Barney Tannahill, Yunus Yetis, and Halid Kaplan. "Big Data Analytic via Soft Computing Paradigms." In *Frontiers of Higher Order Fuzzy Sets*, pp. 229-258. Springer New York, 2015.

[4] Yetis, Y., Kaplan, H., & Jamshidi, M. (2014). Stock market prediction by using artificial neural network. In *World Automation Congress Proceedings.* (pp. 718-722).

THANK YOU FOR TIME


UTSA