THE UNIVERSITY OF HONG KONG (HKU)

Final Year Project 2015/2016

Project Plan

Sibie Arunmozhi
UID: 2012555916

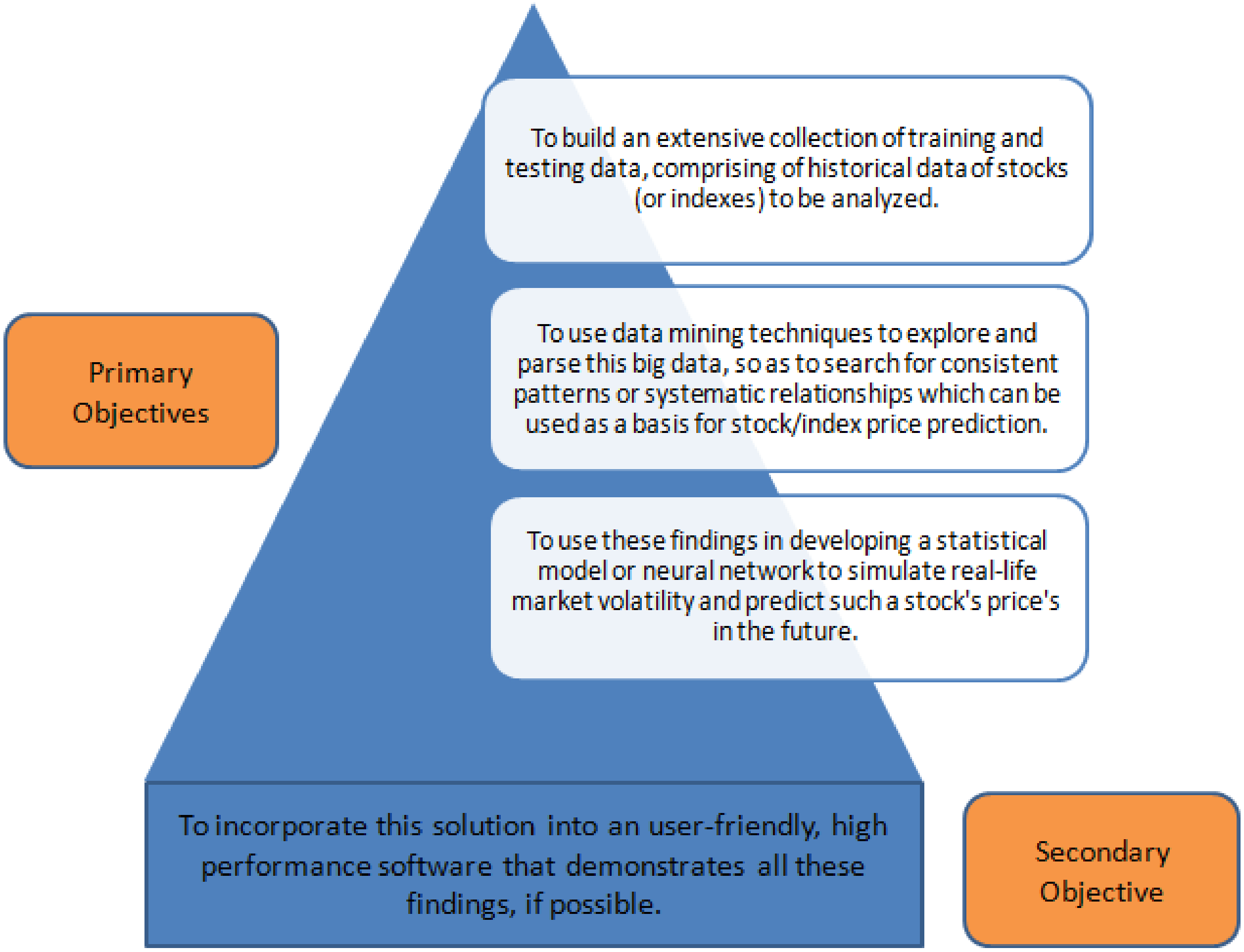
This FYP, offered by the Department of Computer Science, is on the development of a Financial Data Forecaster which can be used to predict the future prices of any given stock. This solution is to be developing using novel concepts in Computer Science such as data mining, neural networks and more.

Table of Contents

Problem Statement

"Given the historical data of a number of stock or index prices at different points in time, design an algorithm that would predict their values in the future. Some factors the algorithm can take into consideration include the day in month, weekday of day, time of day, various financial indicators and correlations between data from different time series. Time series data of at least 15 stocks, futures and/or indices are expected to be analyzed."

Objectives



Primary Objectives

To build an extensive collection of training and testing data, comprising of historical data of stocks (or indexes) to be analyzed.

To use data mining techniques to explore and parse this big data, so as to search for consistent patterns or systematic relationships which can be used as a basis for stock/index price prediction.

To use these findings in developing a statistical model or neural network to simulate real-life market volatility and predict such a stock's price's in the future.

Secondary Objective

To incorporate this solution into an user-friendly, high performance software that demonstrates all these findings, if possible.

Theoretical Background

There are three key concepts studied in this project:

Data Mining

Data mining is a novel concept used to intelligently read patterns from vast amounts of raw data, in this case historical data of stocks analyzed. Key uses of data mining aside from pattern discovery are pattern association, path analysis and clustering (documenting groups of facts not previously known). Thus the power of data mining can be used to forecast stock performance and predict their potential future values. Using this concept with Neural Networks could result in powerful applications, not limited to the financial domain, but other industries as well.

Neural Networks

An Artificial Neural Network (ANN) is a data-processing entity that is inspired by the biological nervous systems like the brain process information. ANNs are made up of a vast number of highly interconnected processing nodes (neurones) which work together to solve complex problems. A key characteristic of ANNs is that they learn by example, for example pattern recognition (sound familiar?). Thus by harnessing the remarkable ability of neural networks to derive meaning from huge repositories of data and detect trends that may be too complex for the average human to observe, together with the technique of data mining, the aim is to develop an algorithmic model that predicts stock prices to an accurate degree. It is important to note here that this is a problem to which there can never be a perfect solution, as real-life markets are impossible to faultlessly simulate; however such a combination of neural networks with data mining could be as close as one could get to one.

Software Design

As always, the Software Development Life Cycle (SDLC) is a tedious process. As this project is an individual one, there is no project management to be done. After developing the neural network, the right software architecture should be chosen to adequately code this model. Testing will also be crucial in maintaining a high performance while processing such vast amounts of historical data.

Methodology

Research Design

The purpose of this study is to develop an algorithmic model that can simulate market volatility effectively, so as to predict future prices of a selected set of stocks (or indexes). Data mining is to be used to discover any patterns that may be present in historical data, which can then be used in developing a neural network to simulate the market. This neural network, possibly with the use of Monte Carlo simulations to make market volatility more realistic, is to be used to predict future stock prices, and incorporated into a software for demonstration.

Research Approach

The stocks to be studied will have to be selected after careful consideration. A good variety in stocks would be desirable as the solution developed must be tested with different types of commodities to observe how well it works. For example, stocks from different industries and different markets may be selected as price-determining factors would vary extensively from case to case.

Sampling Method

A limit of 15 stocks has been decided upon for this project. These stocks will be predominantly from the U.S. and Hong Kong markets. Industries which will be considered include but are not limited to Technology, Natural Resources, Aviation amongst others. Historical Data will be collected from financial data repositories such as Bloomberg, as well as the company or index websites. Key financial statistics can be acquired from financial statements which are available to the public.

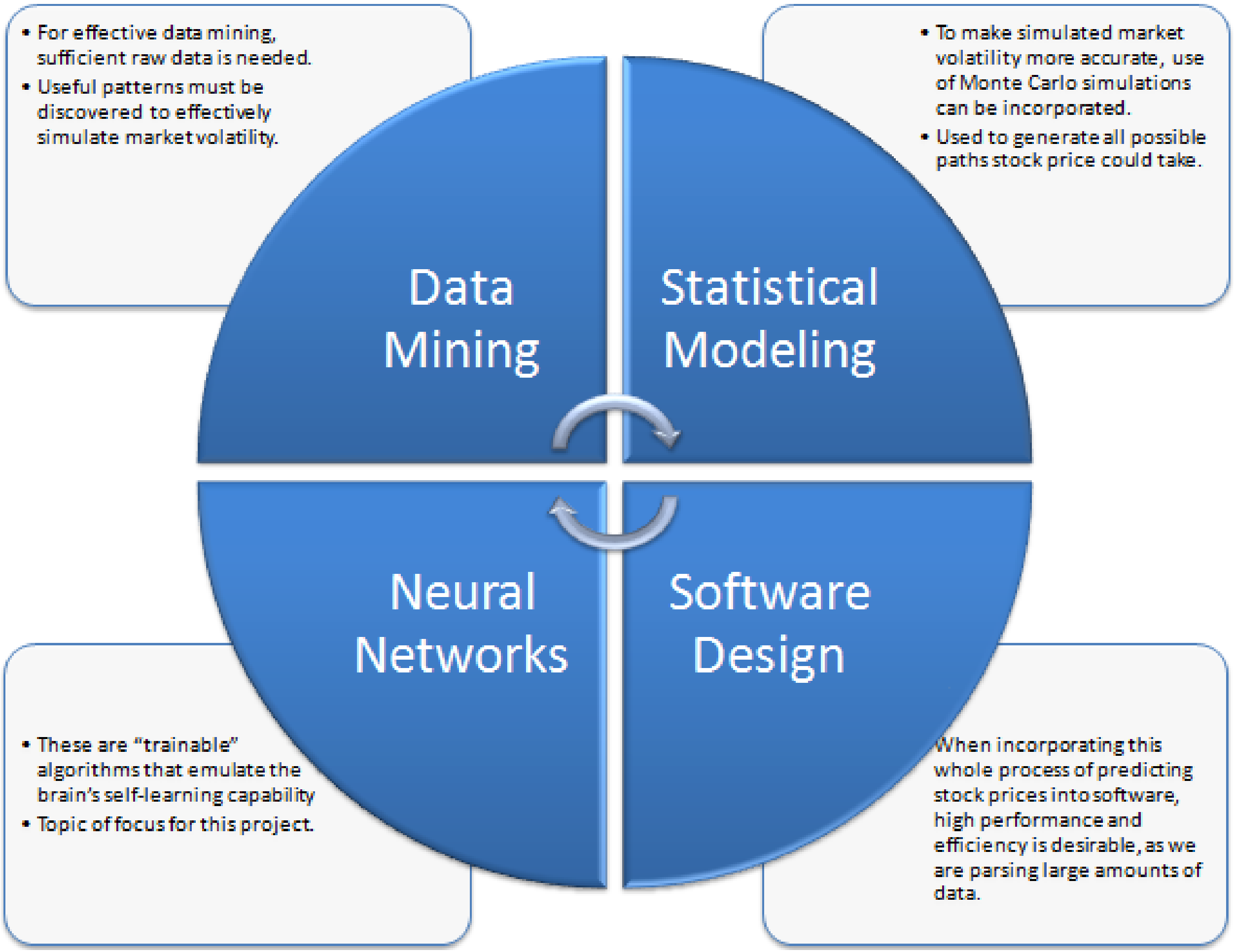
Data Collection Method

Data will be collected directly from these online sources and organized into a centralized repository (possibly a "warehouse") where it can be effectively mined for patterns and key information. This data will then be tested in the designed neural network, unless an alternate solution is developed.

Data Analysis Method

The data analysis of this research will be mostly represented on a quantitative basis. Data mining is to be documented in depth, and results to be factored into neural network design. Data will then be tested with the neural network so as to make stock price predictions, the accuracy of which is the key success criteria of this project. However there may be certain qualitative aspects to the research as well.

Scope



- For effective data mining, sufficient raw data is needed.
- Useful patterns must be discovered to effectively simulate market volatility.
- To make simulated market volatility more accurate, use of Monte Carlo simulations can be incorporated.
- Used to generate all possible paths stock price could take.
Data Mining
Statistical Modeling
Neural Networks
Software Design
- These are "trainable" algorithms that emulate the brain's self-learning capability
- Topic of focus for this project.
When incorporating this whole process of predicting stock prices into software, high performance and efficiency is desirable, as we are parsing large amounts of data.

Pre-requisites

Before the actual implementation, there are some pre-requisites that need to be covered:

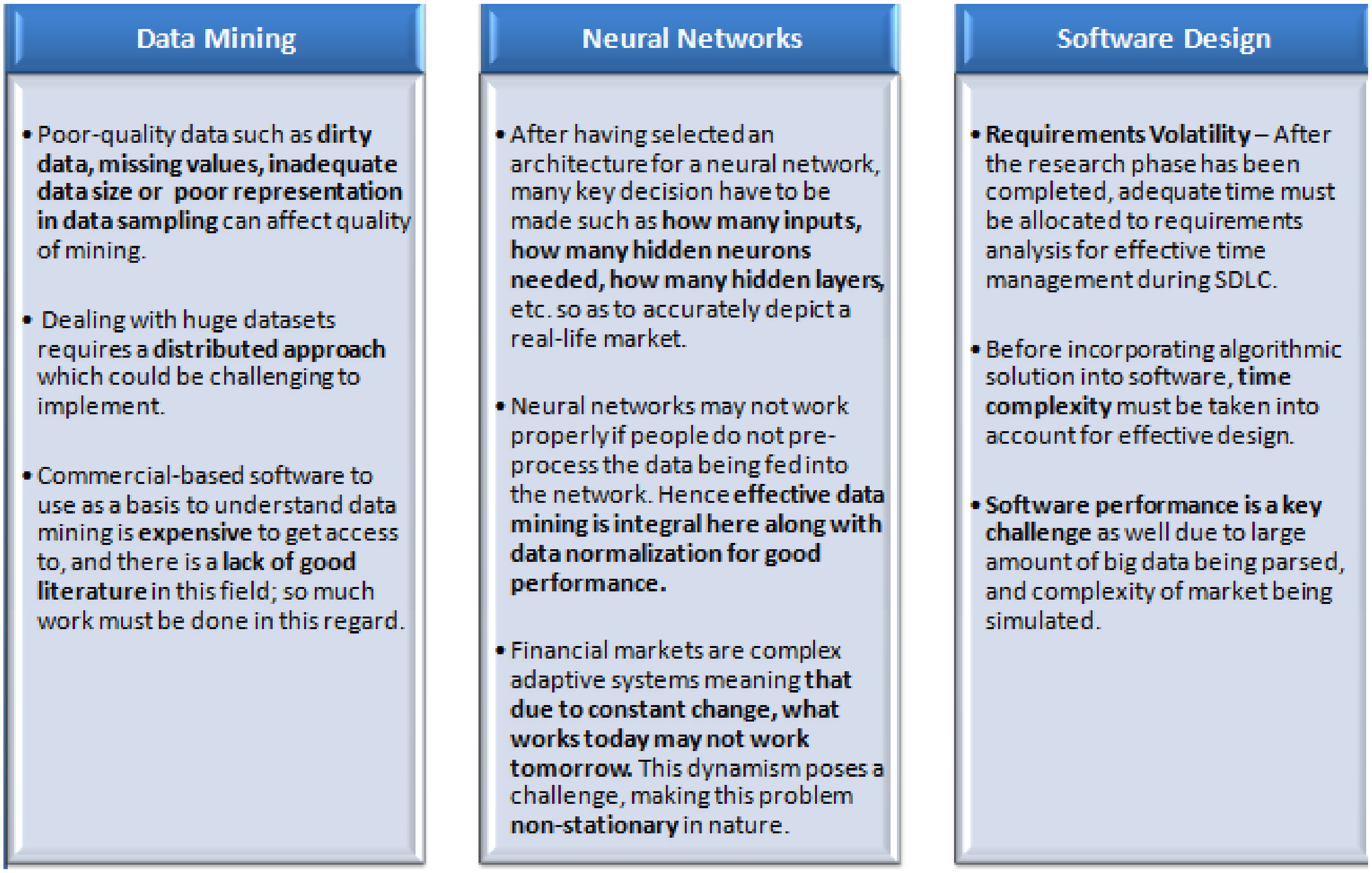
- Research on data mining methods will be so as to effectively analyze historical data.
- Research on neural networks and how to factor data mining results into network design.
- SDLC process to be decided upon in advance to avoid hindrance entering into the implementation phase. SDLC includes testing as well.
- Software pre-requisites include but are not limited to Java, Eclipse or IntelliJ as an IDE for software design, web-design essentials such as HTML, JavaScript, CSS, XML and possibly PHP or Python. There are no particular Hardware pre-requisites. Some existing forecasting software will be assumed to be the standard.

Deliverables

There will be three core deliverables for this project aside from standard FYP submissions:

- A fully-functional, user-friendly, well-tested software to demonstrate the neural network, or alternate algorithmic solution, in action.
- Complete documentation on how data mining is done on historical data by software, and factored into neural network. Essential information on the software design will also be included here, along with standard SDLC documentation.
- An in-depth white paper on how the neural network/alternate algorithmic solution was developed, how it works, as well as a short study on any alternate solutions that could have been explored.

Key Challenges



Data Mining

- Poor-quality data such as **dirty data, missing values, inadequate data size or poor representation in data sampling** can affect quality of mining.
- Dealing with huge datasets requires a **distributed approach** which could be challenging to implement.
- Commercial-based software to use as a basis to understand data mining is **expensive** to get access to, and there is a **lack of good literature** in this field; so much work must be done in this regard.

Neural Networks

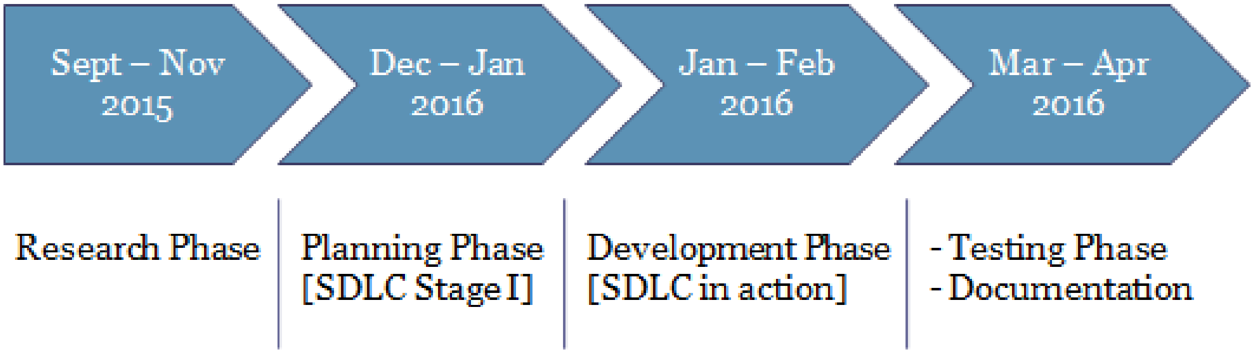
- After having selected an architecture for a neural network, many key decision have to be made such as **how many inputs, how many hidden neurons needed, how many hidden layers,** etc. so as to accurately depict a real-life market.
- Neural networks may not work properly if people do not pre-process the data being fed into the network. Hence **effective data mining is integral here along with data normalization for good performance.**
- Financial markets are complex adaptive systems meaning **that due to constant change, what works today may not work tomorrow.** This dynamism poses a challenge, making this problem **non-stationary** in nature.

Software Design

- **Requirements Volatility** – After the research phase has been completed, adequate time must be allocated to requirements analysis for effective time management during SDLC.
- Before incorporating algorithmic solution into software, **time complexity** must be taken into account for effective design.
- **Software performance is a key challenge** as well due to large amount of big data being parsed, and complexity of market being simulated.

Schedule

There is a tentative four phase plan for this project. After the pre-requisites outlined before have been covered during the research phase and the solution has been developed, this project's SDLC will go on for a time-span of approximately ten to twelve weeks. Standard industry norms with respect to SDLC documentation to be followed. After development, a minimum of one month will be dedicated towards software testing and documentation, moving into the final weeks before FYP presentations.



* SDLC – Software Development Life Cycle

Summary

To conclude, the above are the key aspects to this project. This project is to be completed individually by myself under the guidance of Professor Chi Lap Yip of the Department of Computer Science. As of the start of October, the research phase is ongoing. The key success criterion can be assumed to be of two types: model design and software design. Model design focuses on data mining and neural networks (how well it was designed, critical factors, etc.). Software design has to do purely with how well SDLC was followed, and the performance quality of this core deliverable. A tentative completion date can be marked for early April, although depending on research phase, this time period could be slightly shorter.