# Stock Market Prediction Using Online Data: Fundamental and Technical Approaches

Nikhil Bakshi
Master's Thesis in Computer Science
ETH Zurich, August 2008

# Acknowledgement

I would like to thank my advisor Prof. Gaston Gonnet and my mentor Prof. Friedemann Mattern for the opportunity to work on this topic and for their guidance during my thesis. I am also very grateful to Gina Cannarozzi for her valuable support and to Alex Schicker for his helpful advice and ideas.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Overview

## 1.1 Introduction

From mainstream books offering investing advice to research papers analyzing mathematical prediction models, the stock market has always been a topic of public and academic interest. Countless publications propose strategies with above-average profits, while others demonstrate the random and unpredictable behaviour of share prices. The debate on the predictability of the stock market recently piqued my interest and led me to choose a Master's thesis topic within this area of research. In particular, the following observations influenced my decision:

- The increasing amount of financially relevant data available on the internet
- The possibility to apply skills from a wide range of computer science disciplines including software engineering, distributed systems, databases and machine learning
- The opportunity to expand my knowledge in finance and investing, as I had only little prior exposure to these fields

The following sections define the goal of the thesis and give an overview of the system that was built.

## 1.2 Basics

In order to clarify the goal of the thesis, two dominant schools of thought on investing must first be introduced.

**Fundamental analysis**
This approach tries to identify promising companies by analyzing their fundamental attributes. This includes characteristics such as financial results, growth forecasts and anticipated product development. It is important to note that this type of analysis is not static; newly released financial information, corporate announcements and other news can influence the fundamental outlook of a company. Fundamental analysis requires expertise in a particular sector and is often conducted by professional analysts. Their investment recommendations are regularly published and updated.

**Technical analysis**
In contrast to fundamental analysis, technical analysis does not try to gain deep insight into a company's business. It assumes the available public information does not offer a competitive trading advantage. Instead, it focuses on studying a company's historical share price and on identifying patterns in the chart. The intention is to recognize trends in advance and to capitalize on them. More details on technical analysis will be presented in chapter 3.

## 1.3 Thesis Goal

The goal of the thesis was to build a system capable of the following tasks:

1. **Collecting fundamental and technical data from the internet**
   The system should be able to crawl specific websites to extract fundamental data like news articles and analyst recommendations. Furthermore, it should be able to collect technical data in the form of historical share prices.

2. **Simulating trading strategies**
   The system should offer ways to specify and simulate fundamental and technical trading strategies. Additionally, combining the two approaches should be possible.

3. **Evaluating and visualizing trading strategies**
   The system should evaluate and visualize the financial performance of the simulated strategies. This allows a comparison to be made between technical, fundamental and combined approaches.

## 1.4 Thesis Scope

The scope of the project was restricted to a particular time frame and sector. This was done because different sectors often behave differently; focusing on one reduces possible noise.

- **Biotechnology**
  The US biotechnology sector was focused on and the companies in the Nasdaq Biotech Index were selected. A detailed listing of the index can be found in appendix A.

- **2002 - 2008**
  The time period from January 2002 to January 2008 was selected, because it contains several phases of varied behavior. Figure 1.1 depicts the significant falling and rising phases of the Nasdaq Biotech Index during 2002 and 2003, as well as several smaller oscillations in the years that followed.


Figure 1.1: Nasdaq Biotech Index (2002 - 2008)

## 1.5 System Overview

The system consists of three main components: a crawler, a simulation server and a client interface. Figure 1.2 visualizes the system architecture and interactions between the components. By designing loosely-coupled components, the system became rather flexible and extensible. Basic descriptions of the individual components are listed below; more detailed explanations can be found in chapters 2, 3 and 4 of this thesis report.

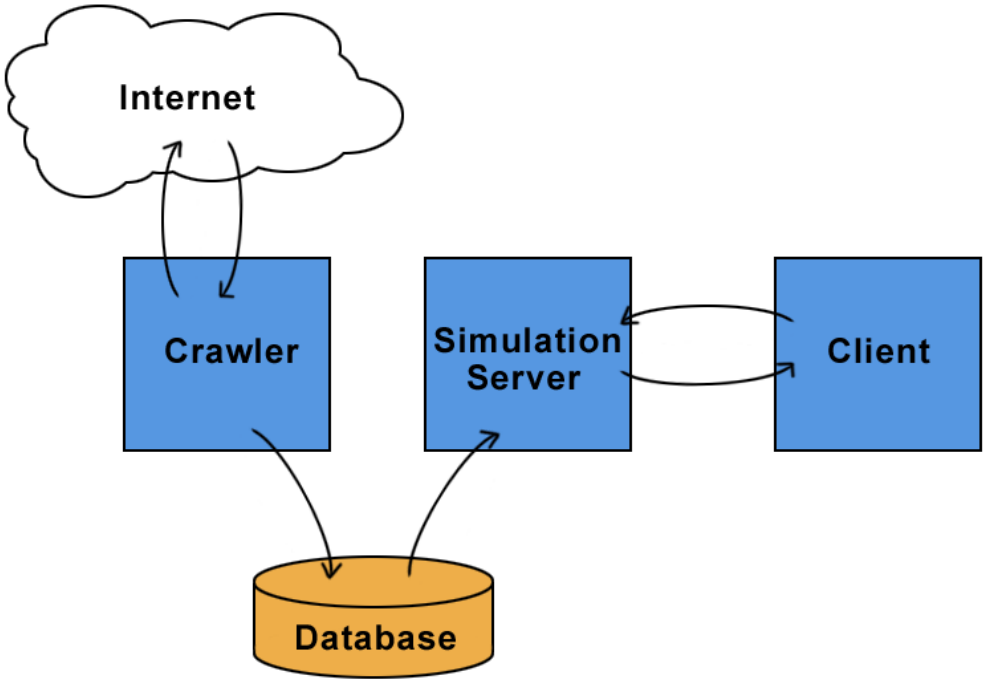


Figure 1.2: System architecture

1. **Crawler**
   The crawler is a Java-based application that extracts data from online sources. The crawled websites and parsed information are stored in a MySQL database. The crawler's frequency and data sources are configurable.

2. **Simulation server**
   The simulation server loads information from the database and runs

trading strategies when instructed by the client. The server's logic is written in Java and is equipped with unit tests.

3. **Client**
   The client offers users a browser-based graphical interface to specify strategies to simulate. The results returned by the simulation server are then visualized appropriately.

## 1.6 Source Code and Documentation

The CD supplied at the end of this thesis report contains the source code and test cases of the project. The classes are documented using Javadoc; this documentation is stored on the CD and can be viewed using a web browser. Sections 2.5, 3.7 and 4.3 explain the source code package structure for each of the three main components.

## 1.7 Thesis Schedule

The six months allotted to the thesis project were roughly spent as follows.

| | |
|---|---|
| 1 month | Studying stock market basics and identifying data sources |
| 1 month | Developing the crawler and collecting online data |
| 1.5 months | Implementing the simulation logic on the server |
| 1 month | Designing the client interface |
| 0.5 months | Simulating and comparing several trading strategies |
| 1 month | Finalizing the thesis report |

## 1.8 Thesis Organization

The rest of this thesis report is organized as follows: chapter 2 focuses on the crawler, chapter 3 explains the simulation server's logic and chapter 4 describes the client interface. Chapter 5 presents the results of several simulations. Chapter 6 summarizes the results and conclusions.

# Chapter 2

# The Crawler

## 2.1 Data Sources

In an initial phase, a large number of websites were studied and the ones most suitable for the project were identified (see table 2.1). The following sections outline characteristics of each data source and list some examples.

| **Data Source** | **Type** | **URL** |
|---|---|---|
| Marketwatch News | Fundamental | marketwatch.com |
| Reuters 'Key Developments' News | Fundamental | reuters.com |
| Yahoo Finance Analyst Recommendations | Fundamental | finance.yahoo.com |
| Yahoo Finance Historical Prices | Technical | finance.yahoo.com |

Table 2.1: Data sources

### 2.1.1 Marketwatch and Reuters News

Both Marketwatch and Reuters offer timely news and good coverage of the biotechnology industry. In contrast to many other websites, their company-specific news archives are easily traversable and date back many years. This was an important criterion for the thesis, as stock market simulations require large historical datasets to be reliable. Furthermore, neither of these news websites relies heavily on Javascript, thus simplifying the crawling task. Below are some sample news headlines from various points in time.

**Marketwatch**

- *Regeneron reports favorable data from obesity trial [9:47am 05/19/03]*
- *Incyte to cut 57% of jobs, close Calif. facility [4:24pm 02/02/04]*
- *Biogen Idec quarterly profit rises to $163.1 million [7:30am 04/23/08]*

**Reuters**

- *Cephalon, Inc. and Novuspharma Form R&D Collaboration for the Treatment of Cancer [Tuesday, 7 May 2002 03:01am EDT]*
- *Celgene Corporation Receives FDA Approvable Letter for Thalomid sNDA [Friday, 22 Oct 2004 05:59pm EDT]*
- *ImClone Systems Inc. Announces Resignation of Chief Executive Officer [Thursday, 10 Nov 2005 05:15pm EST]*

### 2.1.2 Yahoo Finance Analyst Recommendations

Yahoo's financial portal includes current and past analyst recommendations for each company. This makes it possible to track the changing sentiment of analysts by following the upgrades and downgrades over time. A data sample for the company Amgen is displayed in table 2.2 below.

| Date | Research Firm | Action | From | To |
|---|---|---|---|---|
| 2008-06-19 | Deutsche Securities | Initiated | - | Buy |
| 2008-03-20 | Wachovia | Downgrade | Outperform | Mkt Perform |
| 2008-01-02 | Lazard Capital | Upgrade | Sell | Hold |
| 2007-12-10 | Bernstein | Downgrade | Outperform | Mkt Perform |
| 2007-11-15 | Lehman Brothers | Upgrade | Equal-weight | Overweight |
| 2007-10-24 | Credit Suisse | Upgrade | Neutral | Outperform |
| 2007-10-01 | Stifel Nicolaus | Downgrade | Buy | Hold |
| 2007-09-12 | UBS | Upgrade | Sell | Neutral |
| 2007-08-13 | Bernstein | Upgrade | Mkt Perform | Outperform |
| 2007-08-01 | William Blair | Downgrade | Outperform | Mkt Perform |
| 2007-07-23 | Citigroup | Upgrade | Sell | Hold |
| ⋮ | | | | |

Table 2.2: Sample analyst recommendations for Amgen

### 2.1.3 Yahoo Finance Historical Prices

After analyzing OpenTick[1] and Yahoo Finance, Yahoo's historical stock quotes were selected. They consist of daily opening, high, low and closing prices and have been adjusted for stock splits and dividends. The more fine-grained resolution of OpenTick (including minute-frequency historical data) was more desirable, but was abandoned because of periods of missing prices and some price inconsistencies when compared to services like Yahoo and Google.

## 2.2 Architecture

The crawler is composed of one subcrawler per data source (see figure 2.1), which makes it easy to plug in new sources if required. All subcrawlers use the open source Apache httpclient[2] library for making HTTP requests.

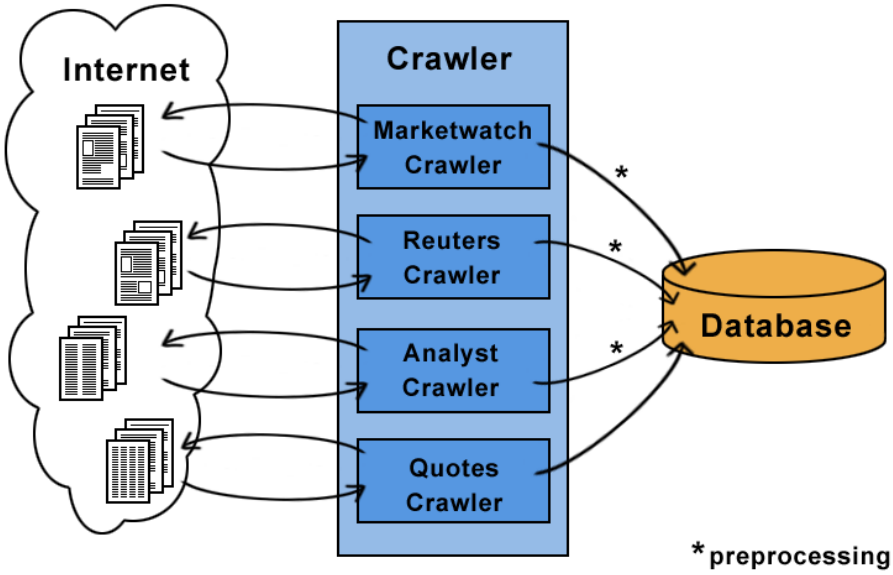


Figure 2.1: Crawler architecture

Before storing data into the database, the news and analyst crawlers perform some preprocessing in order to extract the relevant information from the raw HTML (see sections 2.2.1 and 2.2.2). The quotes crawler does not need this phase, as Yahoo's historical quotes are conveniently available in CSV format.

### 2.2.1 Preprocessing the News

The goal of the new preprocessing phase is to parse headlines and their exact timestamps from the raw HTML. Below are some practical considerations that came up during the implementation of this phase.

**Parsing HTML**
Initial approaches using primitive string pattern matching became rather unwieldy and error-prone when encountering the occasional non-standard-conforming HTML page. This problem was solved by 1) fixing faulty HTML with the open source NekoHTML[3] library and 2) replacing the string pattern matching techniques with more robust XPath expressions. For example, the XPath expression `//A[@class="pageNext"]` extracts all hyperlinks of the class pageNext, thus simplifying the task of crawling a series of linked news pages. The news crawlers' source code can be studied for more examples of XPath usage (see section 2.5).

**Parsing headlines**
Marketwatch's company-specific news occasionally includes headlines that are not necessarily directly related to a company (e.g. general news regarding inflation rates). Such news is detected by removing headlines that do not contain the name of a company. For this purpose a collection of common company names was created; for example, the firm *Biogen Idec, Inc.* is commonly referred to as *Biogen Idec* and sometimes simply *Biogen*. The entire list of common company names can be found in the crawler source code accompanying this report (see section 2.5).

**Parsing timestamps**
The timestamps on news articles could be extracted as strings such as *4:56pm 10/30/06* and *Thursday, 10 Nov 2005 05:15pm EST*. They were parsed with Java's SimpleDateFormat class and stored in Coordinated Universal Time (UTC). This time zone standardization simplified all future timestamp comparisons.

### 2.2.2 Preprocessing the Analyst Recommendations

Two issues surfaced while parsing analyst recommendations.

**Parsing research firms**
Several notations were being used for the same research firm (e.g. *CSFB* and *CS First Boston*). A map was manually created to ensure the different expressions were mapped to the same firm. The entire list of associations can be found in the crawler source code accompanying this report (see section 2.5).

**Parsing analyst recommendations**
Different research firms tend to use different vocabulary for recommendations. For example, some use *Market Outperform*, while others use *Overweight* or simply *Buy* to suggest a buying opportunity. In order to compare recommendations, all 96 different phrases found in the dataset were manually mapped to the three expressions *Buy*, *Neutral* and *Sell*. Appendix B lists the various phrases and their mappings.

## 2.3 Storage

As mentioned in the overview, a MySQL database was used to store all extracted data. Additionally, a copy of the raw HTML pages was kept for reference purposes. Java Database Connectivity (JDBC) was used as a communication layer between MySQL and Java. Table 2.3 lists the database tables used for storage; appendix C can be consulted for details regarding their schema.

| Database Table Name | Description |
|---|---|
| marketwatch_news_pages | Raw Marketwatch HTML pages |
| marketwatch_news_articles | Parsed Marketwatch articles |
| reuters_news_pages | Raw Reuters HTML pages |
| reuters_news_articles | Parsed Reuters articles |
| yahoo_analyst_pages | Raw Yahoo analyst HTML pages |
| yahoo_analyst_recommendations | Parsed analyst recommendations |
| yahoo_quotes | Historical share prices |

Table 2.3: Database tables

## 2.4 Data Statistics

A total of 381'479 historical quotes, 4'222 analyst recommendations, 31'651 Marketwatch and 13'907 Reuters news articles were collected. Figure 2.2 depicts the distribution of monthly news article frequencies for Marketwatch and Reuters. For many companies, Reuters publishes on average between 0.5 and 1.5 articles per month. In the case of Marketwatch, the variance between individual companies is larger (with several companies even receiving an average of over 5 articles per month).

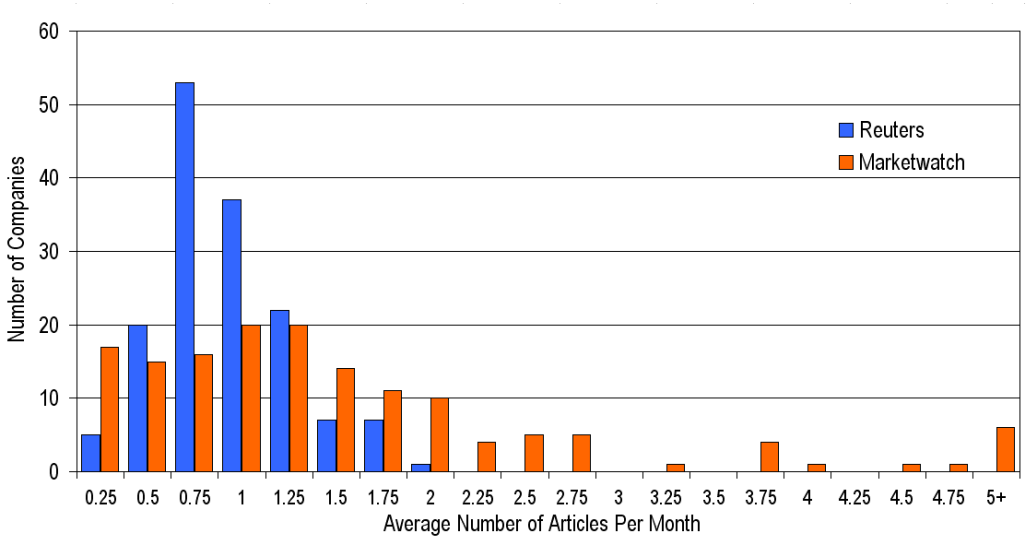


Figure 2.2: Frequency of Reuters and Marketwatch news articles

## 2.5 Source Code Organization

The crawler's source code packages are listed below (see table 2.4).

| Java Package | Description |
|---|---|
| `server.data.crawler` | The implementation of the crawler |
| `server.data.crawler.analyst` | The analyst recommendations subcrawler, including research firms map |
| `server.data.crawler.news` | The Marketwatch and Reuters subcrawlers, including parsing logic |
| `server.data.crawler.company` | Companies and their common names |
| `server.data.crawler.quote` | The quotes subcrawler |
| `server.data.crawler.support` | Support classes common to all subcrawlers |
| `server.data.storage` | The database logic and settings |

Table 2.4: Crawler source code packages

# Chapter 3

# The Simulation Server

## 3.1 Introduction

The simulation server's job is to simulate trading strategies on the data collected by the crawler. Table 3.1 shows the simplified behavior of a trading strategy. Each step is explained in detail in the following sections.

| | |
|---|---|
| **Step 1** | Begin on a starting date with an initial amount of cash. |
| **Step 2** | Evaluate all available biotechnology companies based on selected fundamental and/or technical signals. Identify the<br>- Most promising companies (share price expected to rise)<br>- Least promising companies (share price expected to fall) |
| **Step 3** | Invest part of the available cash in the identified companies.<br>- Buy shares of the most promising companies<br>- Short sell the least promising companies |
| **Step 4** | Trigger any stop losses if necessary. |
| **Step 5** | Move to the next point in time and go to Step 2. |

Table 3.1: Trading strategy behavior

### 3.1.1 Step 1: Initialization

The starting date, initial amount of cash and duration are flexible parameters for the simulation. The currency of all transactions is the US Dollar (USD).

### 3.1.2 Step 2: Computing Trading Signals

The fundamental and technical signals for evaluating companies are described in sections 3.2 and 3.3. A company is deemed promising when all the specified signals indicate an upward price trend. Accordingly, a company is deemed failing when all specified signals predict a downward price trend. Typically, several companies meet the criteria on a given day.

### 3.1.3 Step 3: Investing

A configurable percentage of the available cash is set aside for investments. It is distributed evenly amongst the investments identified in step 2. Additionally, one can specify a limit for any individual investment (e.g. USD 500). Note that transaction costs are not taken into account.

**Long positions**
Shares of promising companies are bought; this is known as entering a long position. If shares of the company are already being held, nothing is done. If a short position of the company is being held, it is covered before entering a long position.

**Short positions**
Shares of failing companies are shorted. Short selling allows profits to be made when share prices fall. If a short position is already being held, nothing is done. If a long position of the company is being held, it is exited before entering a short position.

### 3.1.4 Step 4: Stop Loss

A trading strategy can specify a stop loss threshold if desired. The goal is to limit the losses of trading by exiting any position that has made a specified loss. For example, a stop loss of 30% exits any trade that has made a loss of 30%, thus preventing potential further loss.

### 3.1.5 Step 5: Iteration

Two points per day are considered for trading: Just after the stock market opens and just before the stock market closes. Steps 2 through 4 are executed for each such point in time.

## 3.2 Fundamental Trading Signals

### 3.2.1 News

Initially, the Text Mining Handbook[4], the crawled news articles and existing research papers in the area of news-based financial prediction were studied. During this phase two important observations were made.

1. **Most news occurs when the market is closed**
   A close analysis of the news articles dataset revealed that 78% of all news occurred outside the stock market trading hours[1]. This means that using natural language processing (NLP) techniques to instantly distinguish good from bad news is of limited practical value, as there is not much of an opportunity to trade on the result. One could just as well wait till the market reopens and observe the change in stock price, which should reflect the impact of important news.

2. **Existing research contains few promising results**
   Appendix D lists a summary of the relevant research papers that were studied. Most of them attempt to classify news into categories like *good*, *neutral* and *bad* using machine learning / NLP. As explained in appendix D, no result stood out as being clearly promising. For example, Mittermayer[21] manages to achieve 60% recall in classifying articles, but the precision for identifying good articles is only 6%.

Therefore, a simplified approach was taken. Instead of trying to classify news the instant it is released, the reaction of the share price is observed at the next available time point. For news occurring after 4:00pm, the current closing price $p_1$ and the next day's opening price $p_2$ are studied. For news occurring between 9:30am and 4:00pm, the current day's opening price $p_1$ and closing price $p_2$ are studied. The following signal is then computed:

$$signal = \begin{cases} 1.0 & \frac{p_2 - p_1}{p_1} > threshold \\ 0.0 & \frac{p_2 - p_1}{p_1} < -threshold \\ 0.5 & \text{else} \end{cases}$$

The *threshold* is typically a value like 5% or 10%. A signal of 1.0 can be regarded as an opportunity to enter the market with a long position, in the

[1] The Nasdaq stock exchange's trading hours are from 9:30am to 4:00pm.

hope that the good news will be followed by a continued positive price trend. Accordingly, 0.0 can be regarded as a signal for short selling due to bad news. 0.5 is a neutral signal.

### 3.2.2 Analyst Recommendations

Due to the preprocessing described in section 2.2.2, the analyst recommendations were easily comparable across research firms. Thus, they could be aggregated to an analyst sentiment. At any given time, the number of analysts recommending *Buy*, *Neutral* or *Sell* could be computed ($n_{Buy}$, $n_{Neutral}$ and $n_{Sell}$ accordingly). This resulted in the following signal:

$$signal = \begin{cases} 1.0 & sentiment > threshold_1, n \geq min \\ 0.0 & sentiment < threshold_2, n \geq min \\ 0.5 & \text{else} \end{cases}$$

where $n = n_{Buy} + n_{Neutral} + n_{Sell}$,

$sentiment = \dfrac{n_{Buy}}{n}$,

and $threshold_1 \geq threshold_2$

The values $threshold_1$ and $threshold_2$ represent levels of analyst sentiment that must be met to trigger buy or sell signals; e.g. selecting a value of 0.8 for $threshold_1$ means 80% of the analysts must be recommending a *Buy*. The parameter *min* specifies the least number of analysts required to compute a signal. As an example, figure 3.1 visualizes the changing analyst *sentiment* of the company Cephalon Inc over time.

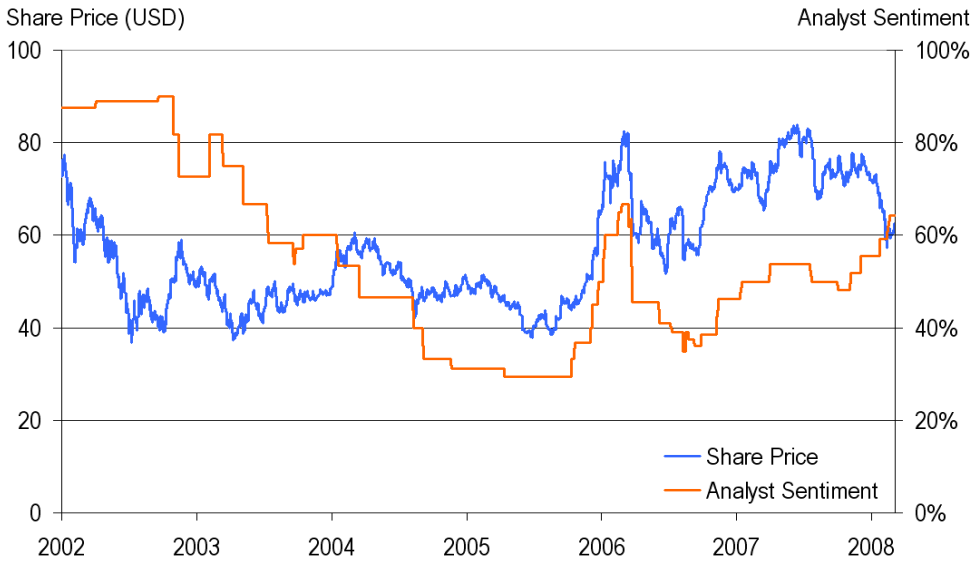


Figure 3.1: Share price and analyst sentiment (Cephalon Inc.)

## 3.3 Technical Trading Signals

The book 'New Trading Systems and Methods'[5] covers technical analysis in detail. After studying the book, the following technical signals seemed promising and were implemented:

1. Moving Average
2. Bollinger Bands
3. MACD (Moving Average Convergence Divergence)
4. RSI (Relative Strength Index)
5. Stochastic

To give the reader a flavour of technical analysis, the Moving Average and Bollinger Bands will be elaborated in the following sections. The other signals are summarized in appendix E and described in detail in [5].

### 3.3.1 Moving Average

A moving average is a simple technique to suggest buying and selling points on a stock price chart. For this purpose, the average share price in a trailing window is computed. Common values for the window size are 20 days, 63 days and 200 days. When the current price rises above the moving average, a buy signal is triggered. When the current price drops below the moving average, a sell signal is triggered. Figure 3.2 visualizes this using an example.

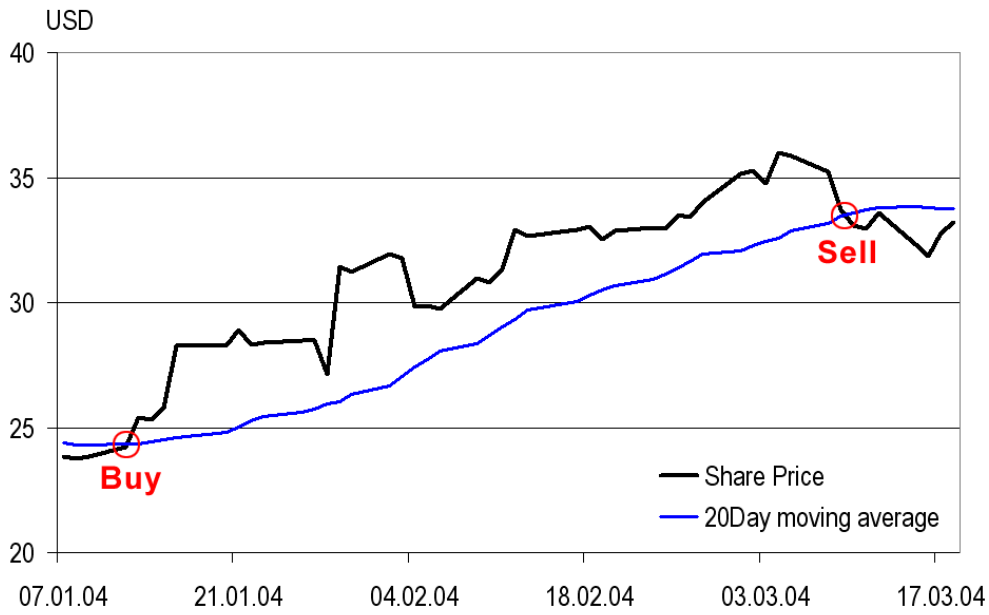


Figure 3.2: Share price and 20-day moving average (Affymetrix Inc.)

With $p_t$ symbolizing the share price at time t, a moving average signal can simply be expressed as:

$$signal = \begin{cases} 1.0 & p_t > movingAverage(n) \\ 0.0 & p_t < movingAverage(n) \\ 0.5 & \text{else} \end{cases}$$

$$\text{where } movingAverage(n) = \frac{1}{n} \sum_{i=1}^{n} p_{t-i}$$

### 3.3.2 Bollinger Bands

Bollinger Bands are volatility-based upper and lower bands around the Moving Average. Buy and sell signals are only triggered when the share price crosses outside one of the bands (see figure 3.3 for an example).

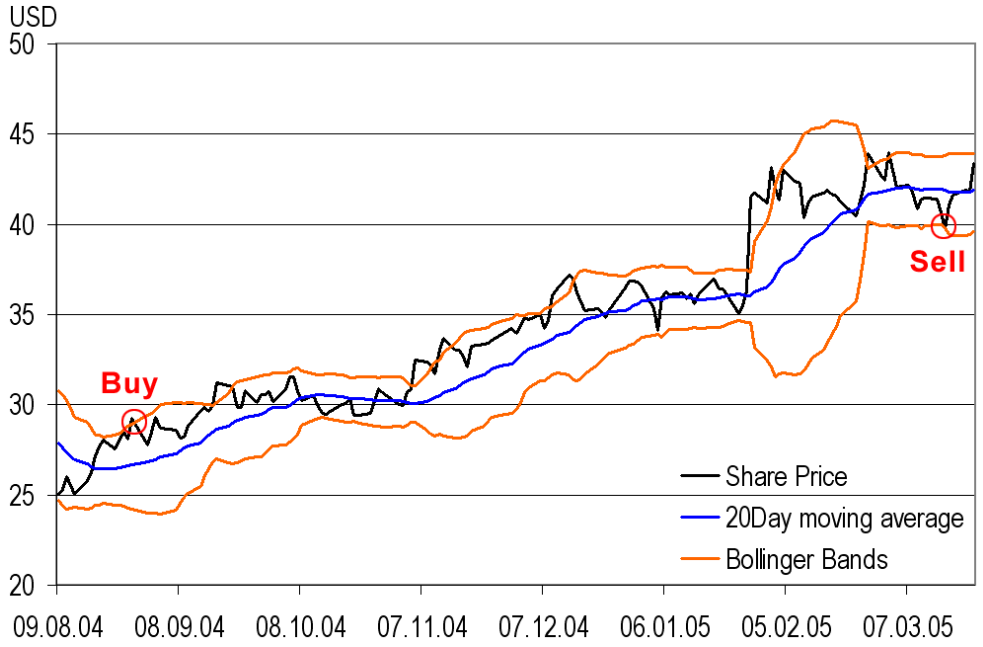


Figure 3.3: 20-day moving average and Bollinger Bands (Affymetrix Inc.)

The formula for this signal is

$$signal = \begin{cases} 1.0 & p_t > movingAverage(n) + 2 \cdot stddev(n) \\ 0.0 & p_t < movingAverage(n) - 2 \cdot stddev(n) \\ 0.5 & \text{else} \end{cases}$$

where $stddev(n)$ is the standard deviation of the prices in the last n days.

## 3.4 Combining Trading Signals

A trading strategy can use one or more of the signals specified in sections 3.2 and 3.3. When using more than one signal, a scheme for combining them is required. The following sections describe the two possible combination techniques that were implemented.

### 3.4.1 Simple Combinations

A simple way to combine the output of several signals is to only signal a buy or sell when all specified signals do so. This can be expressed as follows:

$$signal_{combined} = \begin{cases} 1.0 & \text{if all individual signals return 1.0} \\ 0.0 & \text{if all individual signals return 0.0} \\ 0.5 & \text{else} \end{cases}$$

### 3.4.2 Combinations using Neural Networks

Several papers[6, 7, 8, 9] describe how neural networks can be used for financial prediction. In order to get a better grasp of neural network fundamentals, part of the course 'Information Processing with Neural Networks'[10] was attended and the relevant chapters in the book 'Elements of Statistical Learning'[11] were studied.

#### Approach

Using historical data, a neural network can be learned that describes how trading signals are related to subsequent price movements. Figure 3.4 shows how the implemented fundamental and technical signals can be used as input values and how the expected future price trend is the desired output. The trained neural network can then be applied to new data to predict future price movements and make investments. Details are explained in the following sections.

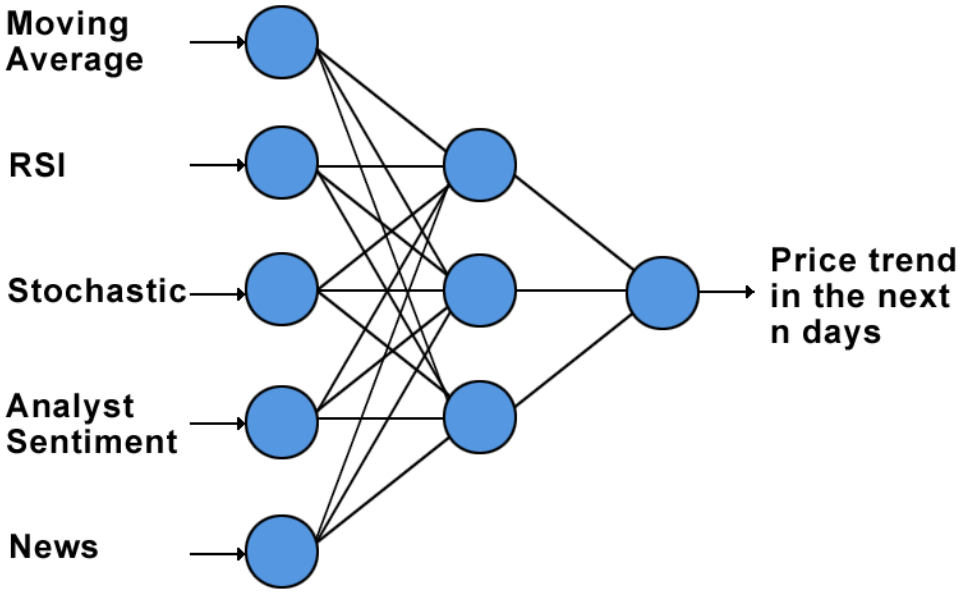


Figure 3.4: The neural network setup

**Inputs**

- **Moving Average**
  The Moving Average signal can be expressed in a price-independent way by computing $\frac{price}{movingAverage}$.
- **RSI**
  The Relative Strength Index uses the $RSI(n)$ formula described in appendix E.
- **Stochastic**
  The Stochastic uses the Stochastic%D-slow expression described in appendix E.
- **Analyst Sentiment**
  This input uses the *sentiment* formula described in section 3.2.2.
- **News**
  This input uses the news signal described in section 3.2.1.

The input values are all normalized to the continuous [0,1] range.

### Hidden Layer

One hidden layer is used with a configurable number of neurons. A regular sigmoid function is used as a transfer function.

### Output

The output is the expected price change in a window of days. The value is in the continuous [0,1] range with 1.0 representing a 10% rise, 0.5 representing no price change and 0.0 representing a 10% price drop.

### Training

The neural network's weights are learned using the backpropagation algorithm with a configurable learning rate and number of epochs.

### Implementation

The open source neural network library Joone[12] was used to build and train neural networks.

### Trading

Once a neural network is built, it can be used by inserting current technical and fundamental input values and computing the predicted output value. If the output crosses a certain upper threshold (e.g. 0.7), an upward price trend can be predicted and shares can be bought. Similiarly, a downward price trend can be signalized by an output value below a lower threshold and short-selling can take place.

## 3.5 Architecture

The simulation server underwent several iterations during the course of the thesis. The final architecture was designed with two primary goals in mind:

1. The ability to run subsequent simulations with minimum time loss in between
2. The ability to easily add new trading signals

To achieve the first goal, the server initially loads data from the database and builds the `TimeLine` data structure in memory (see figure 3.5). Details on loading the data are explained in section 3.5.2. Once built, all simulations then use this shared `TimeLine` without depending on the database any longer, thus saving time. On a single-core 2GHz machine with 512MB RAM, building the initial `TimeLine` roughly takes two minutes and subsequent simulations take ten seconds. Note that the Java Virtual Machine's heap space must be increased for this purpose (e.g. by passing the flag `-Xmx256m`).

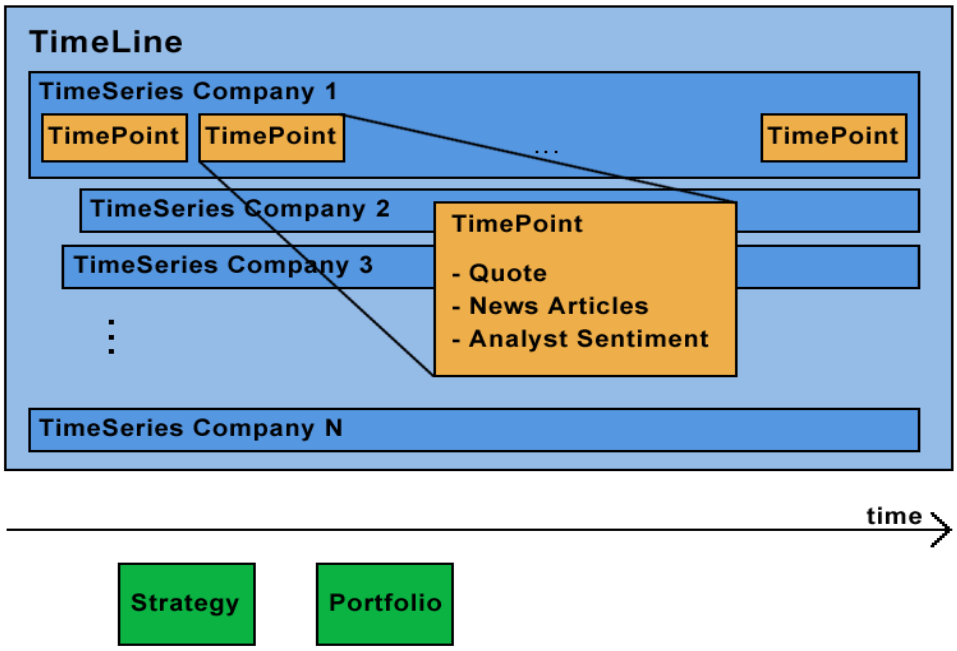


Figure 3.5: The TimeLine data structure

### 3.5.1 TimeLine

Figure 3.5 shows the components of the timeline. It consists of a `TimeSeries` per company containing `TimePoints`. Each `TimePoint` holds the stock price information, news articles and analyst recommendations for that date.

When running a simulation, a `Strategy` (consisting of one or more trading `Signals`) is specified. While iterating over the `TimeLine`, the `Signals` are evaluated and investments are made; the class `Portfolio` tracks the performance over time.

### 3.5.2 Building the TimeLine

When loading data from the database to build the `TimeLine`, two optimizations are made in order to keep the memory usage within 256MB:

1. Quotes before the year 2000 are not loaded
2. News articles triggering less than 5% change in stock price are not loaded

### 3.5.3 Adding a Signal

Additional fundamental or technical signals can be added by implementing the `Signal` interface found in the package `server.simulation.signal`. The following two methods must be implemented:

```
public interface Signal {

   public String getName();

   public Double computeSignal(TimePoint timePoint, boolean morning);

}
```

The method `computeSignal()` returns a double between 0.0 and 1.0, where 0.0 symbolizes a clear sell signal, 1.0 a buy signal and 0.5 a neutral signal. The method `getName()` returns the signal's name.

## 3.6 Unit Testing

When working on a large project, small bugs can creep in and easily go unnoticed for some time (e.g. array indices off by one). Particularly when running simulations, the results may be greatly affected and the error may be hard to track down. In order to prevent this to a certain extent, unit tests were written using the JUnit4 framework. The behavior of all relevant simulation server classes could be checked; when refactoring parts of the server, the behavior could be revalidated. The CD accompanying this report contains the test cases in the directory `/test`. The package structure is the same as the one used in the source code.

## 3.7 Source Code Organization

Below is a list of the simulation server's source code packages.

| **Java Package** | **Description** |
|---|---|
| `server.common.model` | Key datastructures including TimeLine, TimeSeries, TimePoint, Company, Quote, NewsArticle, AnalystSentiment, Portfolio |
| `server.common.finance` | Implementation of financial computations (moving averages, standard deviations, etc.) |
| `server.common.utilities` | Basic time and math utilities |
| `server.simulation.build` | Logic for building the TimeLine and each of its components |
| `server.simulation.signal` | Fundamental and technical trading signals |
| `server.simulation.strategy` | Logic for specifying a trading strategy |
| `server.simulation.neural` | Neural network related classes |
| `server.simulation.visualization` | Simulation visualization logic (will be explained in more detail in chapter 4) |
| `server.simulation` | The Simulation control class for running a simulation |

Table 3.2: Simulation server source code packages

# Chapter 4

# The Client

## 4.1 Architecture

The client lets users specify and simulate trading strategies. It is a browser-based interface built using the open source GWT[13] framework. GWT allows you to develop a web front-end in Java and automatically compiles it to JavaScript, thus simplifying the development of interactive web applications. Client-server communication is asynchronous and is achieved using AJAX. Figure 4.1 depicts the interaction between the client and the server.

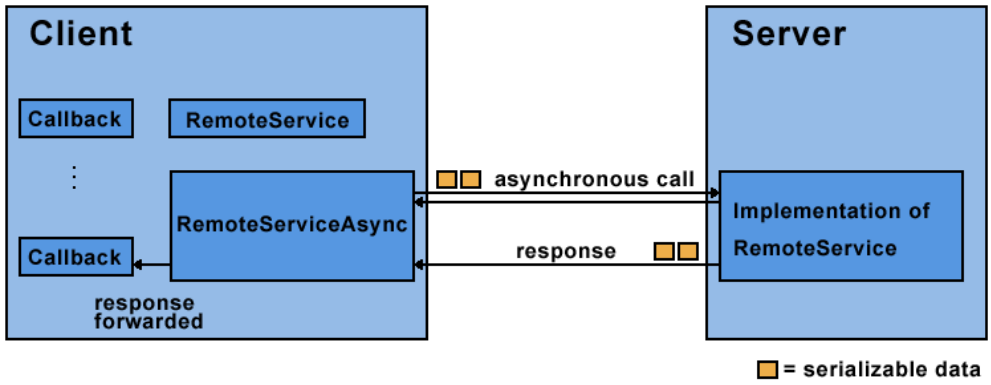


Figure 4.1: Asynchronous client-server communication

The interface between client and server extends GWT's `RemoteService` interface. The asynchronous server responses are handled by callbacks that

implement the `AsyncCallback` interface. All data transferred between the client and the simulation server must be serializable and must implement the `IsSerializable` interface.

Apart from triggering simulations, the client must also visualize the results. This was achieved by extending the simulation server to also create charts using JFreeChart[14]. The resulting .png images are then transferred to the client for display.

## 4.2 User Interface

The client web interface was codenamed StreetSmart; figure 4.2 shows the initial screen. The following sections explain the available functionality.

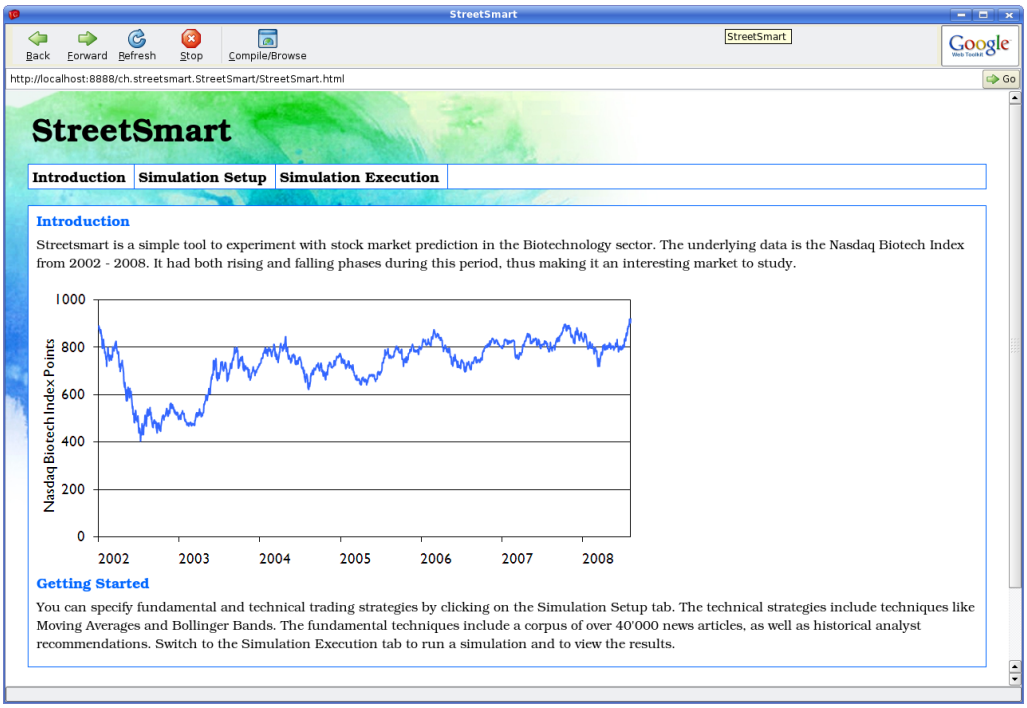


Figure 4.2: Initial screen

### 4.2.1 General Settings

Clicking the tab 'Simulation Setup' brings up some general settings (see figure 4.3). This includes the simulation time period (starting date and number of simulation days), as well as the initial cash amount. It is also possible to specify the percentage of cash to be invested (see section 3.1.3, e.g. 0.5 means half the available cash is invested). Furthermore, the maximum size for any individual trade can be limited (e.g. USD 500).

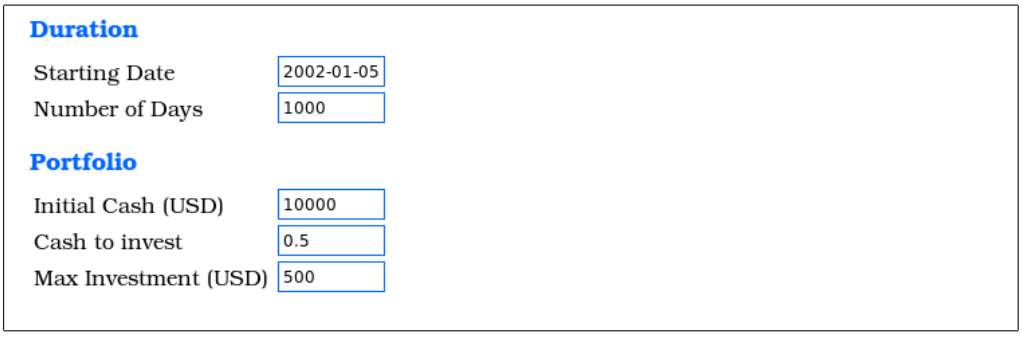


Figure 4.3: General settings

### 4.2.2 Specifying Trading Strategies

Specifying trading strategies is also done within the 'Simulation Setup' tab. Any number of trading strategies can be specified using the 'Add' and 'Remove' buttons.

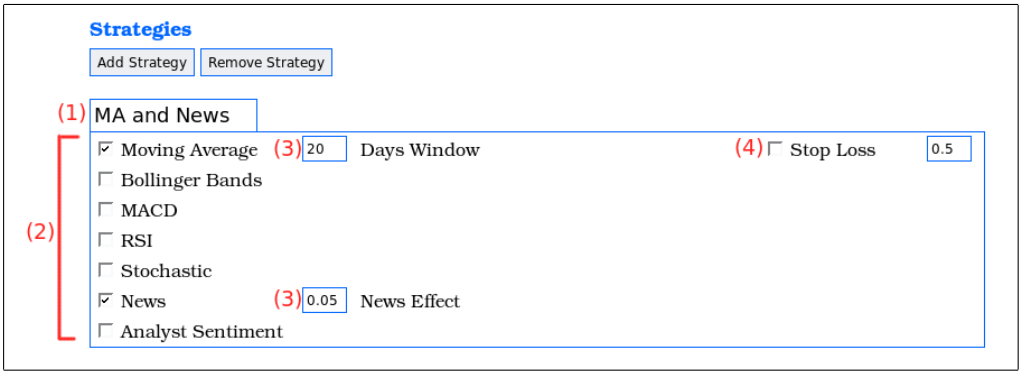


Figure 4.4: Specifying trading strategies

Figure 4.4 shows the available options for a strategy. Each strategy can be given a name (1) and a list of trading signals (2). Each selected signal's parameters can be configured further (3). Stop loss can be specified with (4) if desired. Note that the signals are combined using the simple technique explained in section 3.4.1. Neural network approaches are not available through the client - they are run directly on the server using the command line, as the simulations take considerably longer.

### 4.2.3 Building the TimeLine

Once the simulation has been set up, it can be executed by switching to the 'Simulation Execution' tab. The first step here is to build the server side `TimeLine` (see sections 3.5.1 and 3.5.2 for more details). Clicking 'Build Time Line' triggers an AJAX call to the simulation server, during which the `TimeLine` is built. If completed successfully, this is displayed (see figure 4.5).

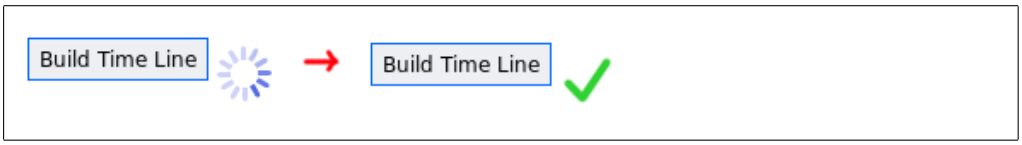


Figure 4.5: Building the TimeLine

### 4.2.4 Viewing Simulation Results

The simulation can then be started by clicking 'Run Simulation'. On completion, the results are displayed.

#### Portfolio Value Chart

Results include a time line of the portfolio value. This is the value of cash on hand and the long and short positions at any given time. If several strategies are specified, their portfolio values are displayed using different colors. Figure 4.6 shows an example.

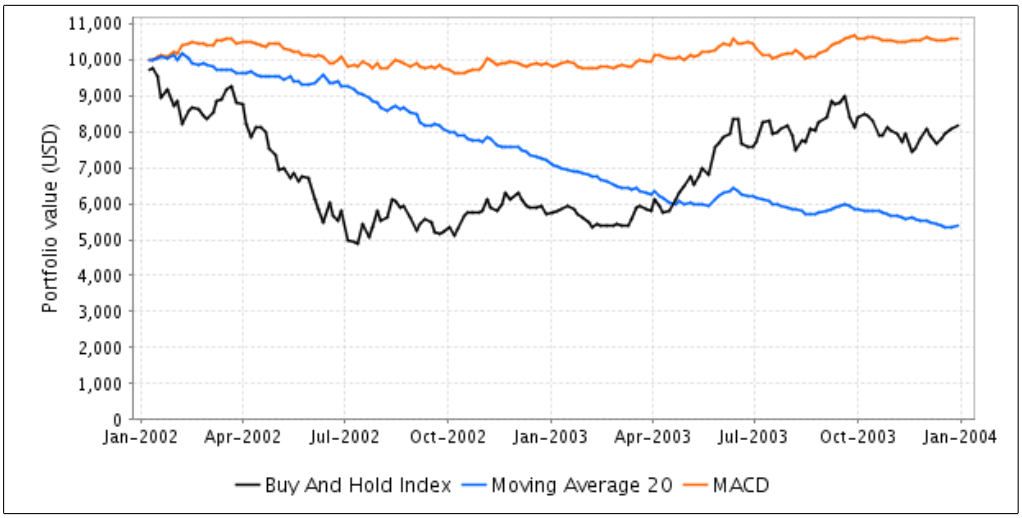


Figure 4.6: Portfolio value chart

### Strategy Details

For each of the strategies, following details are listed (see figure 4.7 for an example).

**(1) Annual ROI**
The annual return on investment is the annual percentage gain/loss on the initial cash amount during the simulation period.

**(2) Number of Positions Held**
The number of long/short positions held during the simulation period.

**(3) Average Position Duration**
The average number of days a position was held during the simulation period.

**(4) ROI Distribution**
The distribution of the return on investment amongst the positions held.

**(5) Buy And Hold Index**
The strategy 'Buy and Hold Index' is the approach of investing in the Nasdaq Biotech Index as a whole and not doing individual trades during the entire time period. This is displayed as a comparison to the specified strategies, as it reflects the general trend of the market.

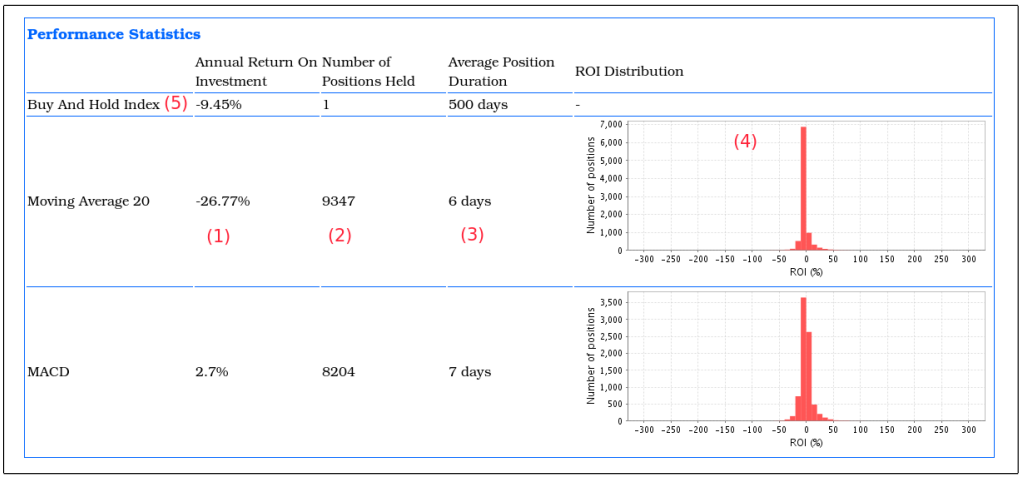


Figure 4.7: Performance statistics for each strategy

## 4.3 Source Code Organization

Below is a list of the client's source code packages.

| **Java Package** | **Description** |
|---|---|
| `client.callbacks` | The callback classes for each type of asynchronous request |
| `client.data` | Serializable data transfer classes |
| `client.layout` | Support logic for the client user interface |
| `client` | Contains StreetSmart, the main user interface class |

Table 4.1: Client source code packages

# Chapter 5

# Experimental Results

## 5.1 Experiment Design

The dataset was split into two disjoint sets and used for two experiment phases I and II. The following sections describe the approach and present the results.

### 5.1.1 Phase I

The initial four years of data (January 2002 - January 2006) were used in this phase. Several experiments were conducted to observe how well individual technical and fundamental strategies performed. Combinations of the two were also studied. See section 5.2 for the simulation results.

### 5.1.2 Phase II

This phase was begun after completion of phase I. Based on the results of phase I, three promising strategies were identified. Their performance was then tested on the two remaining years of data (January 2006 - January 2008). See section 5.3 for the simulation results.

## 5.2 Phase I Results

### 5.2.1 Moving Average and Bollinger Bands

This simulation compares the Moving Average and Bollinger Bands signals with different window sizes.

#### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

#### Strategies and Results

| Strategy | Annual ROI | # Positions | Average Duration |
|---|---|---|---|
| Buy And Hold Index | -2.93% | - | - |
| 20-Day Moving Average | -19.24% | 18124 | 6 days |
| 63-Day Moving Average | -7.24% | 9818 | 12 days |
| 200-Day Moving Average | +4.06% | 4952 | 23 days |
| 20-Day Bollinger Bands | -3.89% | 3296 | 36 days |
| 63-Day Bollinger Bands | +26.64% | 1122 | 104 days |
| 200-Day Bollinger Bands | +19.78% | 441 | 241 days |

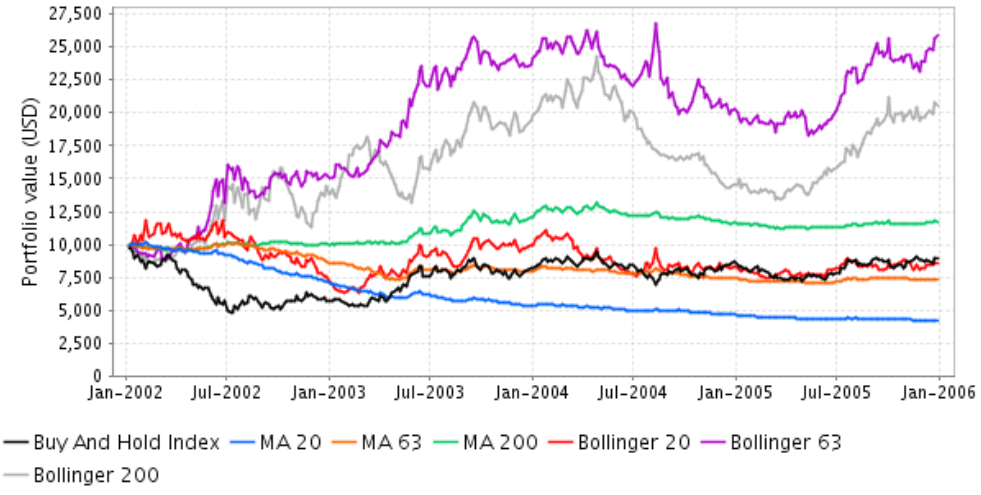


Figure 5.1: Simulation of Moving Average and Bollinger Bands

#### Observation

Several observations can be made. Firstly, using Bollinger Bands reduces the number of signals compared to the Moving Average. This is reflected in the number of positions traded and the average duration of the positions. For example, the 20-Day Moving Average had 18124 positions and the 20-Day Bollinger Bands signal only had 3296. Secondly, Bollinger Bands tend to perform better in the simulation than the Moving Average counterpart. This can be observed in the higher annual ROI for all window sizes. Thirdly, for both signals, the larger 63-day and 200-day windows perform better than the smaller 20-day window.

### 5.2.2 MACD, RSI and Stochastic

This simulation compares the technical signals MACD, RSI and Stochastic.

#### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

#### Strategies and Results

| Strategy | Annual ROI | # Positions | Average Duration |
|---|---|---|---|
| Buy And Hold Index | -2.93% | - | - |
| MACD | +2.73% | 16592 | 7 days |
| 7-Day RSI | -14.66% | 4815 | 25 days |
| 14-Day RSI | -14.57% | 2870 | 42 days |
| 10-Day Stochastic | -5.36% | 28667 | 4 days |
| 20-Day Stochastic | -6.09% | 27559 | 4 days |

#### Observation

One can observe that both the MACD and Stochastic signals enter a large number of positions. The MACD achieves an annual ROI of +2.73%, whereas the Stochastic signals both make minor losses. Figure 5.2 displays the usual behavior of the RSI signals in that they experience a sudden and significant downturn.

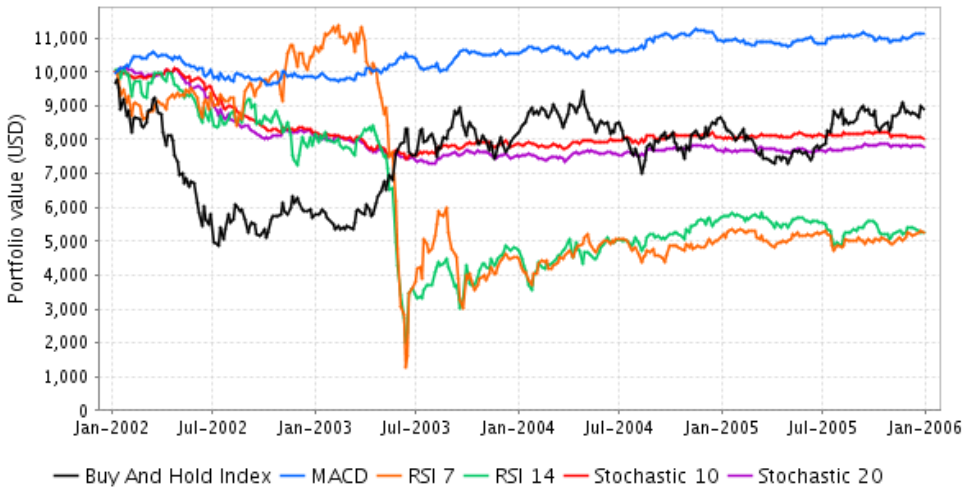


Figure 5.2: Simulation of MACD, RSI and Stochastic

### 5.2.3 Analyst Sentiment

This simulation tests analyst sentiment signals. Different values for the minimum number of analysts and threshold values are tested.

#### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

#### Strategies and Results

| Strategy | Annual ROI | # Positions | Average Duration |
|---|---|---|---|
| Buy And Hold Index | -2.93% | - | - |
| 5 Analysts [20%,40%] | +0.32% | 107 | 722 days |
| 5 Analysts [40%,60%] | -1.23% | 122 | 630 days |
| 5 Analysts [60%,80%] | -6.3% | 144 | 529 days |
| 10 Analysts [20%,40%] | +1.48% | 65 | 634 days |
| 10 Analysts [40%,60%] | -1.49% | 74 | 554 days |
| 10 Analysts [60%,80%] | -4.38% | 85 | 465 days |

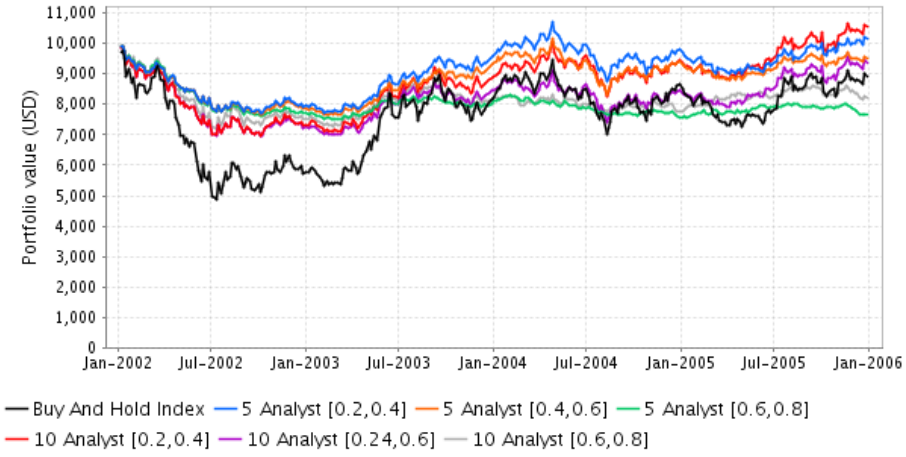


Figure 5.3: Simulation of the analyst sentiment signal

#### Observation

Based on figure 5.3, one can observe that the strategies tend to follow the general trend of the market. Changing the sentiment thresholds slightly influences the portfolio performance: for both 5 and 10 analysts, the thresholds 20% and 40% have the highest annual ROI.

### 5.2.4 News

This simulation tests fundamental signals based on news. Different thresholds values for the stock price reaction are tested.

#### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

### Strategies and Results

| Strategy | Annual ROI | # Positions | Average Duration |
|---|---|---|---|
| Buy And Hold Index | -2.93% | - | - |
| 5% News | +40.86% | 804 | 119 days |
| 7% News | +49.71% | 551 | 168 days |
| 10% News | +41.53% | 346 | 226 days |
| 15% News | +37.97% | 204 | 349 days |
| 20% News | +26.05% | 137 | 419 days |
| 30% News | +23.97% | 77 | 472 days |

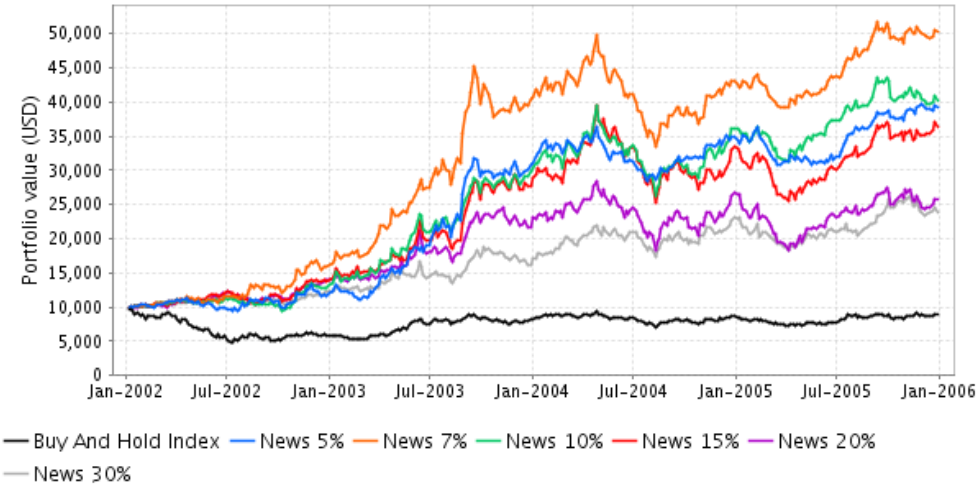


Figure 5.4: Simulation of the news signal

### Observation

One can observe that all the news strategies performed well, with an annual ROI above the market. The strategies using only strong news (15%, 20% and 30% stock price effect) performed worse than the strategies using weaker news (5%, 7% and 10% stock price effect).

### 5.2.5 Simple Combinations

These simulations test strategies that combine a technical and a fundamental signal. The resulting annual ROI values are listed below. The values in bold represent combinations that achieved an annual ROI that was higher than each of the individual signals alone.

#### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

#### Strategies and Results

| | News 7% | News 10% | News 20% | 5 Analysts [20%,40%] | 10 Analysts [20%,40%] |
|---|---|---|---|---|---|
| Moving Average 63 | +31.29% | +25.64% | +9.37% | -6.23% | -4.2% |
| Moving Average 200 | +15.43% | +14.56% | -15.07% | +5.17% | +5.22% |
| Bollinger Bands 63 | +26.8% | +26.36% | +2.65% | +22.14% | **+35.48%** |
| Bollinger Bands 200 | -22.5% | +5.58% | -28.13% | +18.34% | **+20.69%** |
| MACD | +44.48% | +33.36% | +8.01% | +0.23% | **+2.79%** |
| RSI 7 | +8.55% | -4.53% | +1.08% | -9.94% | -9.96% |
| RSI 14 | +32.82% | +20.79% | +2.78% | -4.01% | -5.06% |
| Stochastic 10 | +42.07% | +29.79% | +10.83% | -4.26% | -4.62% |
| Stochastic 20 | +43.01% | +32.75% | +13.65% | -3.75% | -5.13% |

#### Observation

Only three of the combinations result in strategies that outperform their individual signals. In all other cases, combining signals does not increase the annual ROI.

## 5.3 Phase II Results

Based on the annual ROI values of the simulations in phase I, the following signals were selected for phase II.

- 63-Day Bollinger Bands
  This signal achieved the highest performance amongst the technical signals.
- News 7%
  This signal achieved the highest performance amongst the fundamental signals.
- Combined 63-Day Bollinger Bands and 10 analysts [20%,40%]
  This was the best combination strategy that improved upon the individual signals.
- 14-Day RSI
  This signal was selected as a low performance validation signal.

### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

### Strategies and Results

| Strategy | Annual ROI during phase I | Annual ROI during phase II |
|---|---|---|
| Buy And Hold Index | -2.93% | +1.67% |
| 63-Day Bollinger Bands | +26.64% | +3.35% |
| News 7% | +49.71% | +10.57% |
| Bollinger + Analysts | +35.48% | +4.82% |
| 14-Day RSI | -14.57% | -11.51% |

### Observation

The annual ROI values during phase II are much smaller than during phase I. Thus, only modest profits could have been made using the strategies during phase II. However, one can observe that the rank order of the tested strategies is consistent over both phases: the news signal has the best performance in both cases, followed by the combined strategy and the Bollinger Bands. The

RSI is a low performance strategy in both phases. Figure 5.5 visualizes the portfolio value over time for the strategies during phase II.

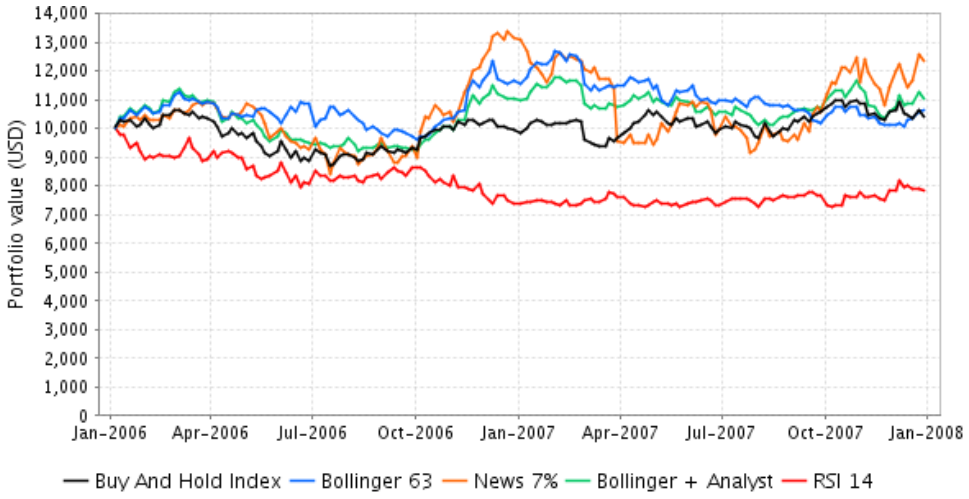


Figure 5.5: Simulation of phase II

## 5.4 Combinations using Neural Networks

The neural network approach described in section 3.4.2 was also evaluated for data from phases I and II. Figure 5.6 shows how the neural network was trained and used: every year, a new common model was built using the last two years of data from all available biotech companies. The output neuron represented the price movement in the next 20 days; values crossing above the threshold 0.7 triggered a buy, values below 0.3 triggered a sell.

### Neural Network Setup

The technical inputs were 63-Day Moving Average, 14-Day RSI and 10-Day Stochastic. The fundamental inputs were the 5% news signal and the analyst sentiment. Three hidden neurons were used and training was done using a learning rate of 0.2 for 1000 epochs.

### Common Setup

| Initial cash | Cash to invest | Maximum per trade | Stop loss |
|---|---|---|---|
| 10'000 | 50% | 500 | - |

### Strategies and Results

| Strategy | Annual ROI during phase I | Annual ROI during phase II |
|---|---|---|
| Buy And Hold Index | -2.93% | +1.67% |
| Neural network with only technical inputs | -3.27% | +0.75% |
| Neural network with fundamental and technical inputs | +0.47% | +0.35% |

### Observation

In phase I, the neural network using additional fundamental inputs performs slightly better than the purely technical one. However this is reversed in phase II. In general, the differences are marginal and neural networks do not show promising results in this simulation.

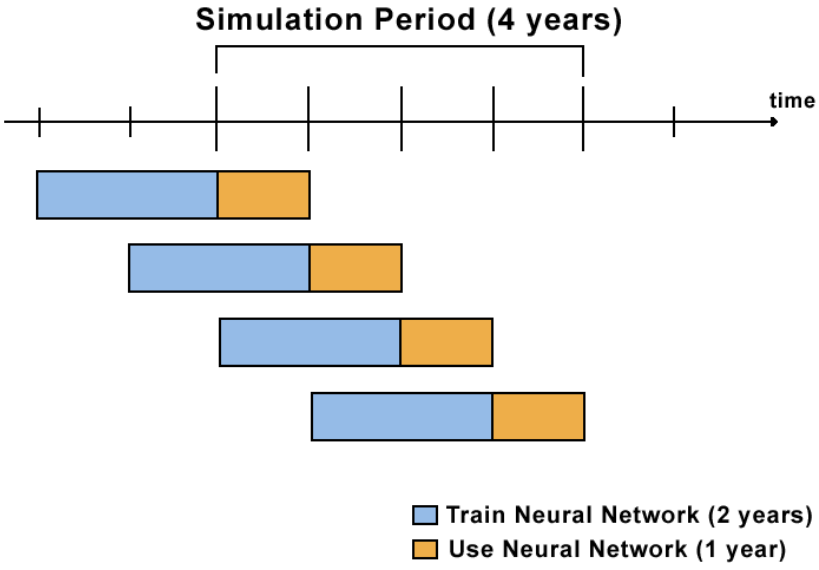


Figure 5.6: Simulation using neural networks

# Chapter 6

# Conclusion

The simulation results and observations in chapter 5 can be summarized as follows. Note that these observations are restricted to the scope of the project, namely the biotechnology sector during the period 2002 to 2008.

- From all technical signals, the Bollinger Bands signal exhibited the highest annual return on investment during phase I of the data. However, this performance could not be replicated during phase II of the data, during which the return on investment was significantly lower.
- Trading based on aggregated analyst recommendations achieved annual returns on investment between -6.3% and +1.48% during phase I. The portfolio value tends to follow market trends, thus suggesting that analyst recommendations primarily reflect the current state of the market and do not have strong predictive power.
- The best news-based trading strategy achieved an annual return on investment of +49.71% during phase I. During phase II of the simulation, this value was considerably lower at +10.57%. However, it still exhibited the highest performance of the tested strategies. This was a promising result, as news is considered public information and should not offer a substantial competitive trading advantage according to the Efficient Market Hypothesis. A possible explanation is the fact that the biotechnology sector consists of many small firms that may not be tracked closely by large investors, thus making the market not fully information efficient. Future work could include focusing on news-based strategies and studying price reactions in greater detail.

- The combined technical and fundamental strategies that were simulated did not consistently show better results than using individual signals separately. The neural network approach to combine signals also led to an unclear result.
- All strategies tested in phase II achieved a lower annual return on investment than in phase I. A possible explanation is the fact that the biotechnology sector exhibited higher volatility during phase I, thus offering potentially larger profits.

Future work could study news-based trading strategies in more detail, as well as test news-based approaches in different sectors.

# Appendix A

# The Nasdaq Biotech Index

The following 152 companies make up the Nasdaq Biotech Index as of July 2008 [Source: Yahoo Finance].

| Symbol | Company Name | Symbol | Company Name |
|---|---|---|---|
| ACAD | Acadia Pharmaceuticals Inc. | ACOR | Acorda Therapeutics, Inc. |
| ADLR | Adolor Corp. | AFFX | Affymetrix Inc. |
| AGEN | Antigenics Inc. | AKRX | Akorn Inc. |
| ALKS | Alkermes, Inc. | ALNY | Alnylam Pharmaceuticals Inc. |
| ALTH | Allos Therapeutics, Inc. | ALTU | Altus Pharmaceuticals Inc. |
| ALXA | Alexza Pharmaceuticals Inc. | ALXN | Alexion Pharmaceuticals, Inc. |
| AMAG | AMAG Pharmaceuticals, Inc. | AMGN | Amgen Inc. |
| AMLN | Amylin Pharmaceuticals Inc. | AMRI | Albany Molecular Research Inc. |
| ANPI | Angiotech Pharmaceuticals Inc. | APPX | APP Pharmaceuticals, Inc. |
| ARIA | Ariad Pharmaceuticals Inc. | ARNA | Arena Pharmaceuticals, Inc. |
| ARQL | ArQule Inc. | ARRY | Array BioPharma, Inc. |
| AUXL | Auxilium Pharmaceuticals Inc. | AVII | AVI Biopharma, Inc. |
| BCRX | BioCryst Pharmaceuticals Inc. | BIIB | Biogen Idec Inc. |
| BIOD | Biodel Inc. | BJGP | BMP Sunstone Corporation |
| BLUS | BELLUS Health Inc. | BMRN | BioMarin Pharmaceutical Inc. |
| BMTI | BioMimetic Therapeutics Inc. | BTRX | Barrier Therapeutics Inc. |
| CADX | Cadence Pharmaceuticals Inc. | CALP | Caliper Life Sciences, Inc. |
| CBRX | Columbia Laboratories Inc. | CBST | Cubist Pharmaceuticals Inc. |
| CEGE | Cell Genesys, Inc. | CELG | Celgene Corporation |
| CEPH | Cephalon Inc. | CERS | Cerus Corporation |
| CRME | Cardiome Pharma Corp. | CRXL | Crucell NV |
| CRXX | CombinatoRx, Incorporated | CTIC | Cell Therapeutics, Inc. |
| CVTX | CV Therapeutics, Inc. | CYPB | Cypress Bioscience, Inc. |
| CYTK | Cytokinetics Inc. | DCGN | deCODE genetics Inc. |
| DEPO | DepoMed Inc. | DNDN | Dendreon Corp. |

*Continued on next page*

| Symbol | Company Name | Symbol | Company Name |
|---|---|---|---|
| DRRX | Durect Corp. | DSCO | Discovery Laboratories Inc. |
| DVAX | Dynavax Technologies Corp. | DYAX | Dyax Corp. |
| ENDP | Endo Pharmaceuticals Holdings | ENZN | Enzon Pharmaceuticals Inc. |
| EPIX | EPIX Pharmaceuticals, Inc. | EURX | Eurand N.V. |
| EXEL | Exelixis, Inc. | FLML | Flamel Technologies SA |
| GENZ | Genzyme Corp. | GERN | Geron Corporation |
| GHDX | Genomic Health Inc. | GILD | Gilead Sciences Inc. |
| GNVC | GenVec Inc. | GPRO | Gen-Probe Inc. |
| GTOP | Genitope Corp. | GTXI | GTX Inc. |
| HALO | Halozyme Therapeutics, Inc. | HBIO | Harvard Bioscience Inc. |
| HGSI | Human Genome Sciences Inc. | HITK | Hi Tech Pharmacal Co. Inc. |
| IDEV | Indevus Pharmaceuticals Inc. | IDIX | Idenix Pharmaceuticals Inc. |
| ILMN | Illumina Inc. | IMCL | Imclone Systems Inc. |
| IMGN | Immunogen Inc. | IMMU | Immunomedics Inc. |
| INCY | Incyte Corporation | INGN | Introgen Therapeutics Inc. |
| ISIS | ISIS Pharmaceuticals Inc. | ISPH | Inspire Pharmaceuticals, Inc. |
| ISTA | ISTA Pharmaceuticals Inc. | ITMN | InterMune Inc. |
| IVGN | Invitrogen Corp. | LGND | Ligand Pharmaceuticals Inc. |
| LMNX | Luminex Corporation | LXRX | Lexicon Pharmaceuticals, Inc. |
| MAXY | Maxygen Inc. | MBRK | MiddleBrook Pharmaceuticals |
| MDCO | Medicines Co. | MDVN | Medivation, Inc. |
| MEDX | Medarex Inc. | MGRM | Monogram Biosciences, Inc. |
| MNKD | MannKind Corp. | MNTA | Momenta Pharmaceuticals Inc. |
| MRNA | MDRNA, Inc. | MTXX | Matrixx Initiatives Inc. |
| MYGN | Myriad Genetics Inc. | NABI | Nabi Biopharmaceuticals |
| NBIX | Neurocrine Biosciences Inc. | NKTR | Nektar Therapeutics |
| NOVN | Noven Pharmaceuticals Inc. | NPSP | NPS Pharmaceuticals, Inc. |
| NUVO | Nuvelo Inc. | NVAX | Novavax, Inc. |
| OMPI | Obagi Medical Products, Inc. | OMRI | Omrix Biopharmaceuticals, Inc. |
| ONXX | Onyx Pharmaceuticals Inc. | ORCH | Orchid Cellmark Inc. |
| OREX | Orexigen Therapeutics, Inc. | OSIP | OSI Pharmaceuticals Inc. |
| OSIR | Osiris Therapeutics, Inc. | PDLI | PDL BioPharma Inc. |
| PGNX | Progenics Pharmaceuticals Inc. | POZN | Pozen, Inc. |
| PPCO | Penwest Pharmaceuticals Co. | PRGO | Perrigo Co. |
| PTIE | Pain Therapeutics Inc. | QGEN | Qiagen NV |
| QLTI | QLT Inc. | REGN | Regeneron Pharmaceuticals Inc. |
| RIGL | Rigel Pharmaceuticals, Inc. | SCRX | Sciele Pharma Inc. |
| SEPR | Sepracor, Inc. | SGEN | Seattle Genetics Inc. |
| SGMO | Sangamo Biosciences Inc. | SHPGY | Shire Limited |
| SLXP | Salix Pharmaceuticals Ltd. | SNTA | Synta Pharmaceuticals Corp. |
| SNTS | Santarus Inc. | SQNM | Sequenom Inc. |
| STEM | StemCells Inc. | SUPG | SuperGen Inc. |
| SVNT | Savient Pharmaceuticals Inc. | TECH | Techne Corp. |
| TELK | Telik Inc. | TEVA | Teva Pharmaceutical Industries |

*Continued on next page*

| Symbol | Company Name | Symbol | Company Name |
|---|---|---|---|
| THRX | Theravance Inc. | TRCA | Tercica Inc. |
| TRMS | Trimeris Inc. | TWTI | Third Wave Technologies Inc. |
| UTHR | United Therapeutics Corp. | VNDA | Vanda Pharmaceuticals, Inc. |
| VPHM | ViroPharma Inc. | VRTX | Vertex Pharmaceuticals Inc. |
| VRUS | Pharmasset, Inc. | VVUS | VIVUS Inc. |
| WCRX | Warner Chilcott Limited | XNPT | Xenoport, Inc. |
| XOMA | XOMA Ltd. | ZGEN | ZymoGenetics, Inc. |

Table A.1: Companies in the Nasdaq Biotech Index

# Appendix B

# Recommendation Phrases

The following 96 phrases were found in the analyst recommendations dataset.

| **Buy** |
| --- |
| Above Average, Accumulate, Add, Attractive, Buy, Buy Aggressive, Buy Speculative, IT Outperform, LT Accum, LT Accumulate, LT Attractive, LT Buy, LT Mkt Outperformer, LT Outperform, LT Strong Buy, Market Outperform, Mkt Outperform, Mkt Outperformer, Net Positive, NT Accum, NT Accum/LT Accum, NT Accum/LT Buy, NT Accumulate, NT Buy, NT Buy/LT Buy, NT Buy/LT Strong Buy, NT Mkt Outperformer, NT Outperformer, NT Strong Buy, NT Strong Buy/LT Strong Buy, NT/LT Accum, NT/LT Buy, NT/LT Outperformer, NT/LT Strong Buy, Outperform, Outperform/Buy, Over Weight, Overweight, Positive, Recomm List, Recomm. List, Recommended List, SB, Sector Outperform, Speculative Buy, ST Buy, ST Buy/LT Buy, Strong Buy, Strong Buy Aggress, Strong Buy Spec, Top Pick |
| **Neutral** |
| Average, Equal Weight, Equal-weight, Hold, In-line, IT Mkt Perform, LT Market Perform, LT Mkt Performer, LT Mkt Perforn, Maintain, Maintain Position, Market Perform, Market Weight, Mkt Perform, Neutral, NT Mkt Performer, NT Mkt Prfm/LT Outprfm, NT Neutral, NT Neutral/LT Buy, NT Reduce/LT Neut, NT/LT Mkt Performer, NT/LT Ntrl, Peer Perform, Perform, Sector Perform, ST Mkt Perform, ST Mkt Perform/LT Mkt Perform, Under Review |
| **Sell** |
| Avoid, Below Average, Market Underperform, Mkt Underperform, Negative, NT Reduce-Sell/LT Neutral, NT Reduce/Sell, Reduce, Sector Underperform, Sell, ST Avoid, ST Avoid/LT Avoid, Strong Sell, Unattractive, Under Weight, Underperform, Underweight |

Table B.1: Analyst recommendation phrases

# Appendix C

# Database Schema

The diagrams below are simple representations of the database schema.

| **marketwatch_news_pages** | | | |
|---|---|---|---|
| id | symbol | url | html |
| bigint(20) | varchar(8) | varchar(1024) | mediumtext |

| **marketwatch_news_articles** | | | | |
|---|---|---|---|---|
| id | symbol | timestampseconds | headline | ... |
| bigint(20) | varchar(8) | int(10) unsigned | varchar(512) | ... |

| ... | | | |
|---|---|---|---|
| .. | source | pageurl | articleurl |
| .. | varchar(128) | varchar(1024) | varchar(1024) |

| **reuters_news_pages** | | | |
|---|---|---|---|
| id | symbol | url | html |
| bigint(20) | varchar(8) | varchar(1024) | mediumtext |

| **reuters_news_articles** | | | | |
|---|---|---|---|---|
| id | symbol | timestampseconds | headline | pageurl |
| bigint(20) | varchar(8) | int(0) unsigned | varchar(512) | varchar(1024) |

| **yahoo_analyst_pages** | | | |
|---|---|---|---|
| id | symbol | url | html |
| bigint(20) | varchar(8) | varchar(1024) | mediumtext |

| **yahoo_analyst_recommendations** | | | | | | |
|---|---|---|---|---|---|---|
| id | symbol | date | researchfirm | action | fromOpinion | ... |
| bigint(20) | varchar(8) | date | varchar(128) | varchar(128) | varchar(128) | ... |

| ... | | | | |
|---|---|---|---|---|
| ... | toOpinion | mappedFirm | mappedFromOpinion | mappedToOpinion |
| ... | varchar(128) | varchar(128) | varchar(128) | varchar(128) |

| **yahoo_quotes** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| id | symbol | date | open | high | low | close | volume | ... |
| bigint(20) | varchar(8) | date | float | float | float | float | bigint(20) | ... |

| ... | |
|---|---|
| ... | adjustedclose |
| ... | float |

Figure C.1: Database schema

# Appendix D

# Research Papers using News-Based Prediction

Below is a list of the relevant research papers that were studied; they are sorted chronologically and accompanied by a short summary.

**1998. Daily Stock Market Forecast from Textual Web Data[15]**
This paper is credited as the first to mine online news for market prediction. To begin, a list of influential phrases (e.g. 'bond strong') was created by domain experts. Probabilistic rules based on the frequency of the terms appearing in news articles are learned using the last 100 days of data. The rules are then applied to current news articles in order to predict the next day's movement of several stock indices (*up*, *steady* or *down*). The best classification accuracy that was achieved was 46.7%. A follow-up paper[16] extends the approach by combining several news sources.

**2000. Language Models for Financial News Recommendation[17]**
This approach tries to filter out the most relevant articles from a stream of news. This is done by associating historical news with five types of stock price trends and building a Bayesian classifier for terms. A trading strategy based on the classification achieved a gain of 0.23% per trade.

**2002. News Sensitive Stock Trend Prediction[18]**
This approach attempts to classify news articles into the three categories *rise*, *drop* and *neutral*. Text classification is achieved with Support Vector Machines (SVM). A trading strategy based on the news classifier is tested,

but the performance is not clearly documented. Two other research papers describing SVM-based approaches are [19] and [20].

**2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques[21]**
The NewsCATS system described in this paper analyzes official press releases instead of regular news articles. SVMs are used for text classification and result in a recall of 60%. However, the precision value for good(bad) news is 6%(5%), which is rather low.

# Appendix E

# Technical Trading Signals

In addition to the Moving Average and Bollinger Bands covered in section 3.3, here are summaries of the other three technical signals. Detailed explanations can be found in [5].

## E.1 MACD

The MACD (Moving Average Convergence Divergence) is a momentum trading signal based on two indicators. It can be expressed as the following signal:

$$signal = \begin{cases} 1.0 & \text{MACD} > \text{MACD line} \\ 0.0 & \text{MACD} < \text{MACD line} \\ 0.5 & \text{else} \end{cases}$$

where MACD = EMA(close, 12) - EMA(close, 26),
MACD line = EMA(MACD, 9)
and EMA is an exponentially smoothed moving average.

## E.2 Relative Strength Index (RSI)

The Relative Strength Index indicates an overbought/oversold situation. It is based on cumulating the upward and downward price movements in a window. Low RSI values suggest an oversold situation and suggest a buy signal. High RSI values suggest a sell signal.

$$signal = \begin{cases} 1.0 & RSI(n) < 0.3 \\ 0.0 & RSI(n) > 0.7 \\ 0.5 & \text{else} \end{cases}$$

where $RSI(n) = \dfrac{RS(n)}{1 + RS(n)}$

and $RS(n) = \dfrac{\text{total upward price movements in the last } n \text{ days}}{\text{total downward price movements in the last } n \text{ days}}$

## E.3 Stochastic

The Stochastic is an oscillating signal.

$$signal = \begin{cases} 1.0 & \text{Stochastic\%D}_t > \text{Stochastic\%D-slow}_t \\ 0.0 & \text{Stochastic\%D}_t < \text{Stochastic\%D-slow}_t \\ 0.5 & \text{else} \end{cases}$$

where $\text{Stochastic\%D-slow}_t = \dfrac{1}{3}\sum_{i=1}^{3} \text{Stochastic\%D}_{t-i}$,

$\text{Stochastic\%D}_t = \dfrac{1}{3}\sum_{i=1}^{3} \text{Stochastic\%K}_{t-i}$,

$\text{Stochastic\%K}_t = \dfrac{p_t - min(p_t..p_{t-5})}{max(p_t..p_{t-5}) - min(p_t..p_{t-5})}$

and $p_t =$ share price at time $t$

# References


[1] OpenTick is a project that offers free historical stock market data.
http://www.opentick.com

[2] The Apache httpclient library is an open source Java library for working with HTTP.
http://hc.apache.org/httpcomponents-client/

[3] NekoHTML is an open source Java library for fixing HTML.
http://nekohtml.sourceforge.net

[4] The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data
Ronan Feldman and James Sanger
Cambridge University Press, 11 Dec 2006

[5] New Trading Systems And Methods
Perry J. Kaufman
Wiley, 4th Edition, 28 Feb 2005

[6] Stoch Market Prediction with Backpropagation Networks
Freisleben, B.
Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, vol.604, pp.451-460, 1992

[7] An Intelligent Forecasting System of Stock Price Using Neural Networks
Baba, N.; Kozaki, M.
International Joint Conference on Neural Networks, vol.1, pp.371-377, 1992

[8] Neural Networks for Technical Analysis: A Study on KLCI
Yao, J.; Tan, C.; Poh, H.-L.
International Journal of Theoretical and Applied Finance, 1999

[9] Combining Technical Analysis and Neural Networks in the Australian Stockmarket
Vanstone, B; Finnie, G.
Artificial Intelligence and Soft Computing, 2006

[10] Information Processing with Neural Networks
Lecture at ETH Zurich, summer term 2008
Dr. J. Bernasconi

[11] The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Trevor Hastie, Robert Tibshirani and Jerome Friedman
Springer, 2001

[12] Joone (Java Object Oriented Neural Engine) is an open source Java neural network framework.
http://jooneworld.com

[13] Google Web Toolkit (GWT) is an open source Java library for developing AJAX-based web applications.
http://code.google.com/webtoolkit

[14] JFreeChart is a Java library for creating charts.
http://jfree.org/jfreechart

[15] Daily Stock Market Forecast from Textual Web Data
Wüthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J.; Lam, W.
IEEE International Conference on Systems, Man, and Cybernetics, vol.3, pp.2720-2725 vol.3, 11-14 Oct 1998

[16] Combining Forecasts from Multiple Textual Data Sources
Cho, V.; Wüthrich, B.
Methodologies for Knowledge Discovery and Data Mining, vol.1574, pp.174-179, 1999

[17] Language Models for Financial News Recommendation
Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.
Ninth International Conference on Information and Knowledge Management, pp.389-396, 2000

[18] News Sensitive Stock Trend Prediction
Fung, G.; Yu, J.; Lam, W.
Advances in Knowledge Discovery and Data Mining, vol.2336, pp.481-493, 2002

[19] Textual Analysis of Stock Market Prediction Using Financial News Articles
Schumaker, R.; Chen, H.
AI Lab, University of Arizone, 2006

[20] Predicting the Short-Term Market Reaction to Asset Specific News: Is Time Against Us?
Robertson, C.; Geva, S.; Wolff, R.
Emerging Technologies in Knowledge Discovery and Data Mining, vol.4819, pp.15-26, 2007

[21] Forecasting Intraday Stock Price Trends with Text Mining Techniques
Mittermayer, M.-A.
International Conference on System Sciences, 5-8 Jan 2004

# Project Source Code (CD)

- /src contains the source code
- /test contains the test cases
- /doc contains the documentation
- /lib contains the libraries