# THE REALIZATION OF VISUALIZATION AND PREDICTION MODELS

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yijun Wang

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

April 2015

Fargo, North Dakota

# North Dakota State University
## Graduate School

**Title**

The Realization of Visualization and Prediction Models

**By**

Yijun Wang

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

**MASTER OF SCIENCE**

SUPERVISORY COMMITTEE:

Dr. Kendall Nygard
Chair

Dr. Kenneth Magel

Dr.Yarong Yang

4/17/2015
Date

Dr. Brain M. Slator
Department Chair

## ABSTRACT

Visualization is a specific technique for building images, animations or diagrams to communicate a message [1]. Nowadays, it effective to communicate both abstract and concrete ideas for big data by visual imagery. Visualization examples from history include cave paintings, Egyptian hieroglyphs, Greek geometry, and Leonardo da Vinci's revolutionary methods of technical drawing for engineering and scientific purposes [2]. In the work of this paper, we are describe the development of a web-based online visualization system and introduce an existing prediction model called - Markov Chains, which may be applied to specific data sets. The development processes, design structure, and testing results are presented in this paper.

## ACKNOWLEDGEMENTS

I would like to acknowledge the help pf many people who made this paper possible. First of all, I would like to thank my advisor, Dr. Kendall Nygard, for his continuous support, help, and direction. My sincere thanks to Dr. Magel and Dr. Yarong Yang for serving on committee. Also, I want to thank my friends Songtao Zheng and Qianwen Yan who encouraged me to complete my paper.

## DEDICATION

This paper is dedicated to my parents.

For their endless love, support and encouragement.

## TABLE OF CONTENTS

ABSTRACT ..... iii

ACKNOWLEDGEMENTS ..... iv

DEDICATION ..... v

LIST OF TABLES ..... viii

LIST OF FIGURES ..... ix

1. INTRODUCTION ..... 1

1.1. Visualization ..... 1

1.2. Big Data and Visualization ..... 3

1.3. Data Mining and Visualization ..... 4

1.4. Objectives and Technical Approach ..... 5

1.5. Structure of the Paper ..... 6

2. RELATED WORK AND BACKGROUND ..... 7

2.1. Motivation ..... 7

2.2. Related Work on Visualization ..... 7

3. FUNCTIONAL SPECIFICATION FOR VISULIZATION ..... 9

3.1. Introduction ..... 9

3.2. System Functional Categorization ..... 10

3.2.1. Realization of Functional Requirements ..... 13

3.2.2. Details of Technical Approaches. ..... 14

3.2.3.System Security and Privacy ..... 16

3.2.4. Markov Chain Model ..... 16

4. REAL-DATA VISUALIZATION TESTING ..... 21

4.1. Introduction .....21

4.2. Visualization Frameworks and Display .....21

4.3. Junit Test .....22

5. Test Result .....24

5.1. Evaluation of Test Result .....24

6. Conclusion and Future Work .....27

6.1. Conclusion .....27

6.2. Future Work .....27

References .....30

## LIST OF TABLES

Table Page

1. High-Level Functional Requirements....10

2. High-Level Non-Functional Requirements....10

3. MC Calculation of Probability....19

4. Statistics of Self-Satisfaction....24

5. Functional Requirements Self-Evaluation Result....24

6. Non-Functional Requirements Self-Evaluation Result....25

7. Result of Self-Evaluation of Real-Data....25

8. Result of Junit Tests....26

## LIST OF FIGURES

Figure Page

1. Traditional Visualization Charts ..... 2

2. Frequency of Sales and Customer Satisfaction ..... 3

3. More Fancy Visualization Charts ..... 4

4. Sample Bar Chart ..... 9

5. Use Case Diagram ..... 11

6. Class Components of the System ..... 12

7. Connections between Classes ..... 13

8. Server Architecture ..... 16

9. Nodes and Emission and Transition Probability ..... 17

10. Sample Transition and Emission Probability of Dice ..... 19

11. Data from EIA ..... 20

12. Charts of the Test Data ..... 22

13. Motion Chart ..... 22

14. Future Charts ..... 28

15. Future Pie Chart of MC Model Result ..... 29

# 1. INTRODUCTION

## 1.1. Visualization

Visualization refers to specific technique for building images, animations or diagrams to communicate a message [1].Visualization methods are particularly effective for understanding and communicating information concerning the big data sets encountered today. Examples of visualization from history include cave paintings, Egyptian hieroglyphs, Greek geometry, and Leonardo da Vinci's revolutionary methods of technical drawing for engineering and scientific purposes [2].

Applications of visualization abound, particularly in science, education, engineering, interactive multimedia, and medicine. Information visualization is the broadest term that could be used to subsume the developments to be described here. Tables, graphs, maps and even text, whether static or dynamic, provide many means to see what lies within, determine the answer about a question, find relations, and perhaps understand things which could not be seen so readily in other forms. As used today, the term information visualization is generally applied to the visual representation of large-scale collections of non-numerical information, such as files and lines of code in software systems [3], library and bibliographic databases, and networks of relations on the internet, from Software Engineering, unified modeling language (UML) provides a visual system that uses special symbols for representing phases of software development across the development cycle. Certain symbolic elements (such as class, contact, aggregate, inheritance) are utilized in the analysis. Other symbols of elements (such as those to implement identity and attributes) are introduced in the design. There are three kinds of symbolic roles: The first one is symbolic serves as a language, to convey decisions it cannot

obviously infer from the code. Secondly, it provides the semantics to capture all important strategic and tactical decisions. Thirdly, it provides a concrete form, enough for people to think, as well as tools to operate. Figure 1 Provides examples of how traditional the visualizations.

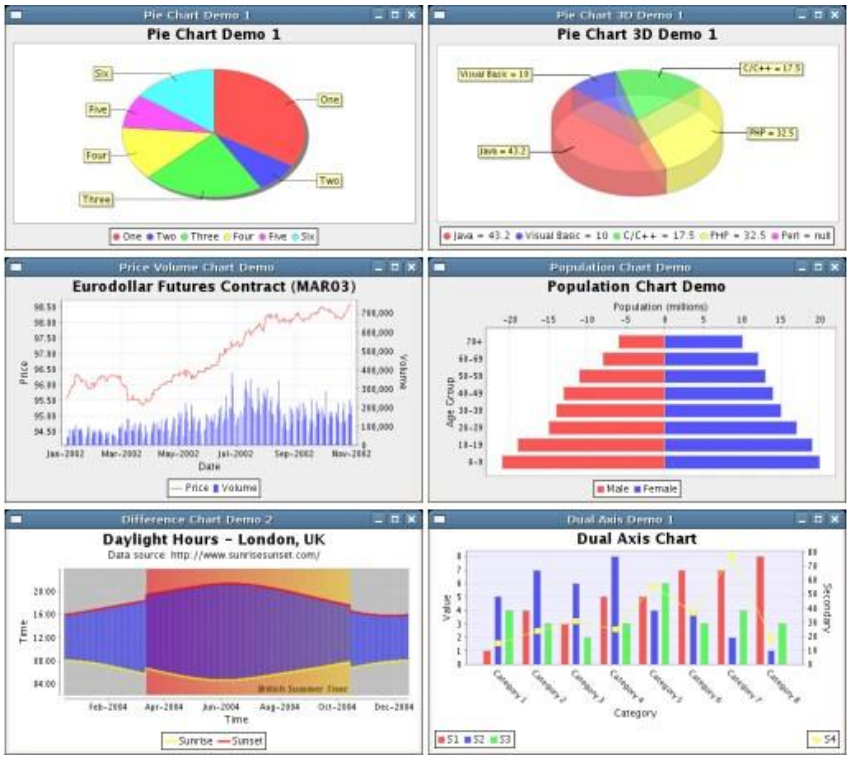


Figure 1. Traditional Visualization Charts

Why is data visualization important? The simple reason is that visualizations help people see things that were previously not obvious to them. Even when data volumes are very large, patterns can be spotted through visualizations. Visualizations convey information in a universal manner and make it simple to share ideas with others. It lets people ask others, “Do you see what I see?” And it can even answer questions like “What would happen if we made an adjustment to that area?”[26]

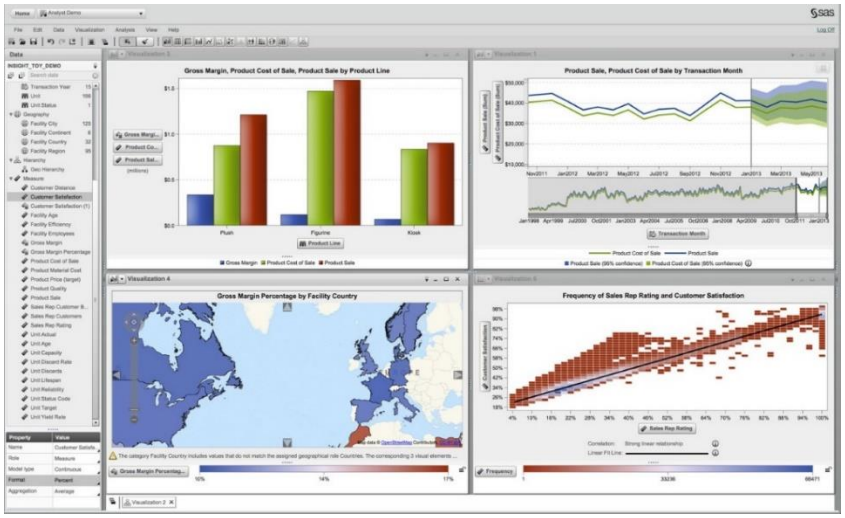


Figure 2. Frequency of Sales and Customer Satisfaction

Furthermore, Data visualization presents the data in a way that the director can easily interpret, saving time and energy. For example, Figure 2 shows the frequency of sales that correspond to customer satisfaction as well as the sales rep rating as customer satisfaction increases. [26]

### 1.2. Big Data and Visualization

Under the background of big data and data mining, a newer technology of visualization is needed for different requirements from different industrial areas.

“Big data” is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization and privacy violations. 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, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on."[5] A better visualization technology will help human beings gain understanding about their data and

as a result, a detailed analysis phase would be easier. In addition, the final visualization of the data can be more acceptable to most people. (Figure 3)

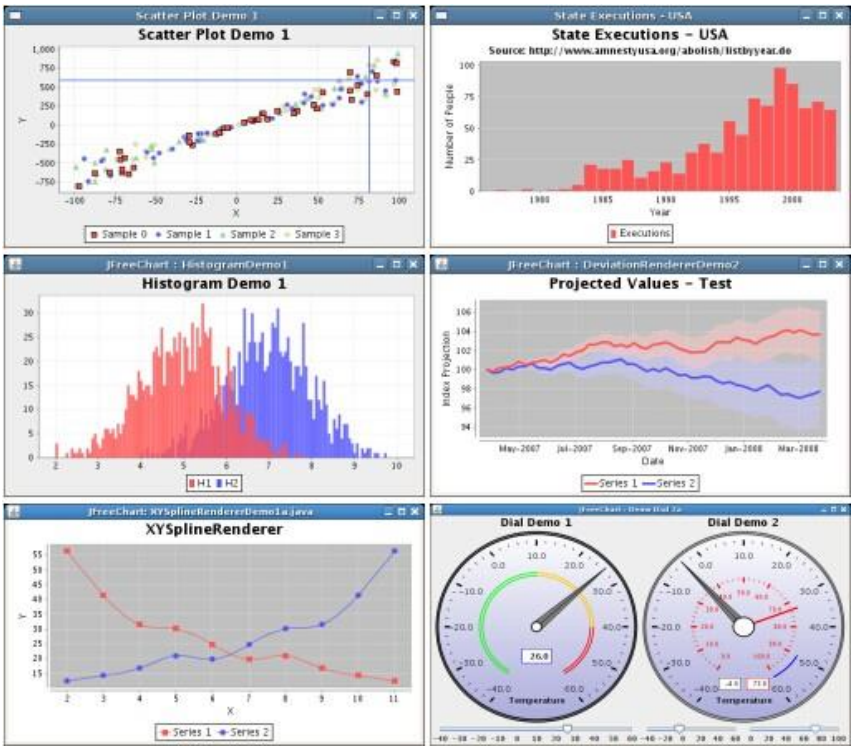


Figure 3. More Fancy Visualization Charts

Big data brings new challenges to visualization because large volumes, different varieties and varying velocities must be taken into account. And, in many cases today, data is just being generated faster than it can be digested. There are many factors to consider.

For example, the cardinality of the data set to be visualized is a factor. High cardinality means there is a large proportion of unique values (e.g., bank account numbers, because each item should be unique). Low cardinality means a set of data contains a large percentage of repeated values (as might be seen in a “gender” column).

### 1.3. Data mining and Visualization

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) [6], an interdisciplinary subfield of computer science,[7][8][9], is the

computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [7]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for subsequent use [7]. Aside from the raw analysis step, data mining involves database and data management, pre-processing, model and inference considerations, interestingness metrics, complexity, post-processing of discovered structures, visualization, and online updating [7]. As a part of the raw step of data mining, visualization not only helps find the interconnections among independent data in a visual way, but can also help to indicate the deviations in the data and define noisy points. The task of designing a software model is analogous to a building needing a blueprint. A good model should be able to identify needs and communicate information, reveal components of system interaction, enable understanding of the relationships among design components and improve cross-team communication. These software engineering design issues are all helped through visualization.

### 1.4. Objectives and Technical Approach

In this paper, we describe a methodology for visualizing data that utilizes xml frames and a popular web-based visualization package called FusionChart to support diagrams in a web-based system. To incorporate possibilities for prediction, we also introduce a famous technique called Markov Chains into our system.

Our web based system currently provides a single page which contains four different styles of charts. It allows users to upload their original data from an excel file and choose the diagram from the four pre-made diagrams, then either hide or reveal items in each diagram. For each column of data we also show basic statistics, such as average values.

A Markov Chain model provides accurate and complicated predictions. The goal is to calculate quantitative values called the “emission probabilities” and “transition probabilities” between two continuous nodes (states). Theoretically, those possibilities can help in predicting the likelihoods for specific states following given initial states.

### 1.5. Structure of the Paper

The paper is organized as follows: The first chapter gives the introduction, definition of the problem, related research, tools and research objectives. The second chapter explains the literature overview. The third chapter discusses the functional specification of the software. The fourth and fifth chapters discuss the design and implementation of the data testing. The sixth chapter discusses the testing results and provides a conclusion and description of future work.

# 2. RELATED WORK AND BACKGROUND

This section outlines the purpose of developing the web-based visualization system and presents related works. This section also cites relevant background to provide context for the research that will be presented in the remainder of the document.

## 2.1. Motivation

Business Analysis is defined as the practice of enabling change in an organizational context by defining needs and recommending solutions that deliver value to stakeholders.

A successful business analysis relies on massive data analysis. A very efficient way to do the data analysis is data mining. Visualization plays a vital role not only in data mining, but also provides an excellent way to understand the data itself. For one real-world example of data mining, consider Lending Club, a US peer-to-peer lending company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club operates an online lending platform that enables borrowers to obtain a loan and investors to purchase notes backed by payments made on loans. Data mining is a powerful way to control risk. Data analytics can provide support for loan trading and visualization can directly support applications such as this one. [24]

## 2.2. Related Work on Visualization

Visualization technology is a broad term that appears in many historical accounts of development within many fields, including probability [10], statistics [11, 12, 13], astronomy [14], and cartography [15]. There are other more specialized accounts, which focus on the early history of graphic recording [16, 17], statistical graphs [18, 19], and fitting equations to empirical data [20], cartography [21] and thematic mapping [22].

Robinson [22] gives an overview of some of the important intellectual, scientific and technical developments of the 15th–18th centuries which lead to the thinking of thematic cartography and statistical analyses.

In this paper we describe our software that provides a contribution to build free-visualization. The primary goal is to provide a flexible, and useful web-based visualization system, containing prediction functions and illustrative images. We also include examples using real-world data (electric power consumption and other resource consumption data). Related work are as follows: (a) Model Design by collecting different teams and roles in the business in order to impact and analyze the demand of the model, (b) the method and principle of the construction of the unified model, (c) delimited the business service application system model of the system boundary, (d) after preliminary information and acquisition, building up the business service application about the system model, (e) the concept of the system with the related team subdivision, (f) the definition and correlation of the particle size, (g) confirmed and concentrated the model for the operations team, (h) formed the basis of elaborating the carding process data.

## 3. FUNCTIONAL SPECIFICATION FOR VISUALIZATION

This section introduces the basic structure of our web-based visualization system and shows the codes that explain the ideas employed.

### 3.1. Introduction

Our goal of doing research is to provide a web based visualization system, which may be used online for multiple kinds of facilities such as a personal computer, mobile phone or tablet. To illustrate the system, four popular types of charts have been developed, including a Bar chart, Line charts, Lattice charts, and a Pie chart. (Figure 4) Also, this system is based on four technologies, namely java + html + xml + tomcat server. All the charts and buttons are justified by using the html and xml files and the bootstrap frame.

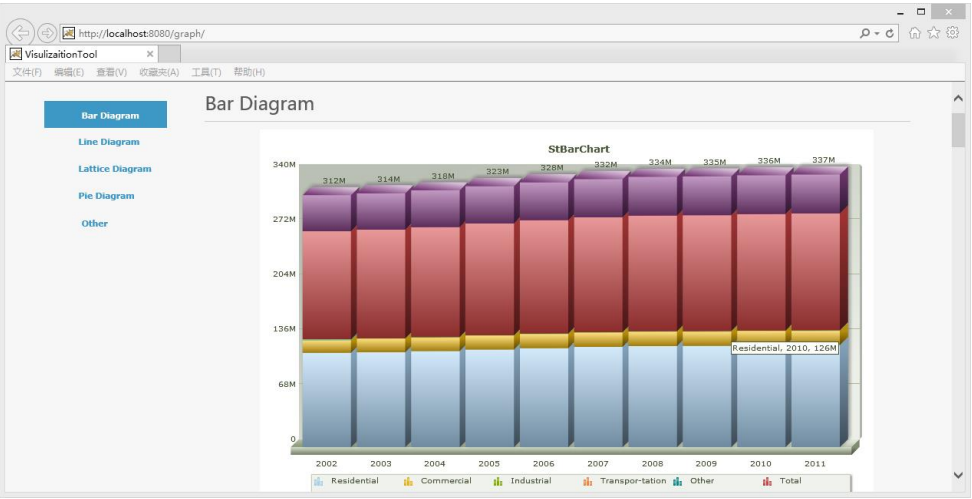


Figure 4. Sample Bar Chart

In order to support flexibility in the system, all the elements at the bottom of the chart in our visualization system can be easily hidden or revealed by one click, and the shape or the color of the data can be predefined while storing data in the xml based data store.

The left part is the navigation bar of charts. Users can find the chart they want and click on the specific button.

### 3.2. System Functional Categorization

Table 1 and 2 shows the main functional requirement and non-functional requirement of the system.

Table 1. High-Level Functional Requirements

| | |
|---|---|
| Data Converting | The one thing a user needs to do is to upload an .xls files. The system automatically converts this file into our predefined xml file in order to show the data as charts. |
| Chart selection | There are 4 different types of charts in our system. Once the data file has been submitted, our user can easily choose the chart they want at the left of the webpage. |
| Element selection | Users can hide or reveal element in the displayed chart, as long as the data file has been submitted. |
| Average Value | The average value for each column should be shown as a new item of each chart. |

Table 2. High-Level Non-Functional Requirements

| | |
|---|---|
| Efficient Structure | All the elements of our system should be well structured which means our user can find what they want in no more than 5 seconds. |
| Availability | The web server should run 24 hours a day, 7 days a week. |
| Easy used buttons | Buttons should be clear to see, and the name of the button should be easily understood. Users must be able to find the button they want in no more than 5 seconds |
| Autoing justify data | Data should be shown within the screen size. |

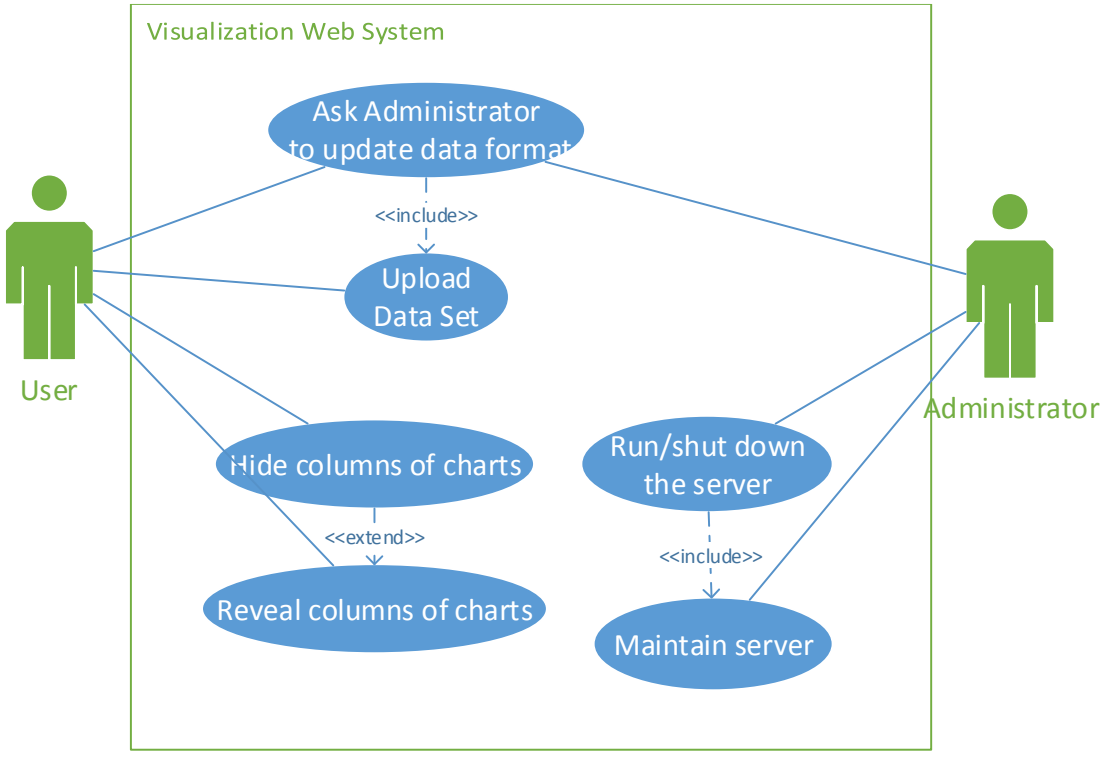


Figure 5. Use Case Diagram

Figure 5 shows the interactions between our web-based visualization system and users. Our user can simply upload their data set or original date by clicking the upload button. Before that, an xml data format based on the original data set should be defined.


Figure 6. Class Components of the System

The Figure 6 indicates the base components of our system. We have 6 different packages and more than 10 different classes in our system. Each of them does a certain job such as converting a data set to a pre-defined data format, translating those data from xml to map, and converting data into a proper diagram in Fusionchart.

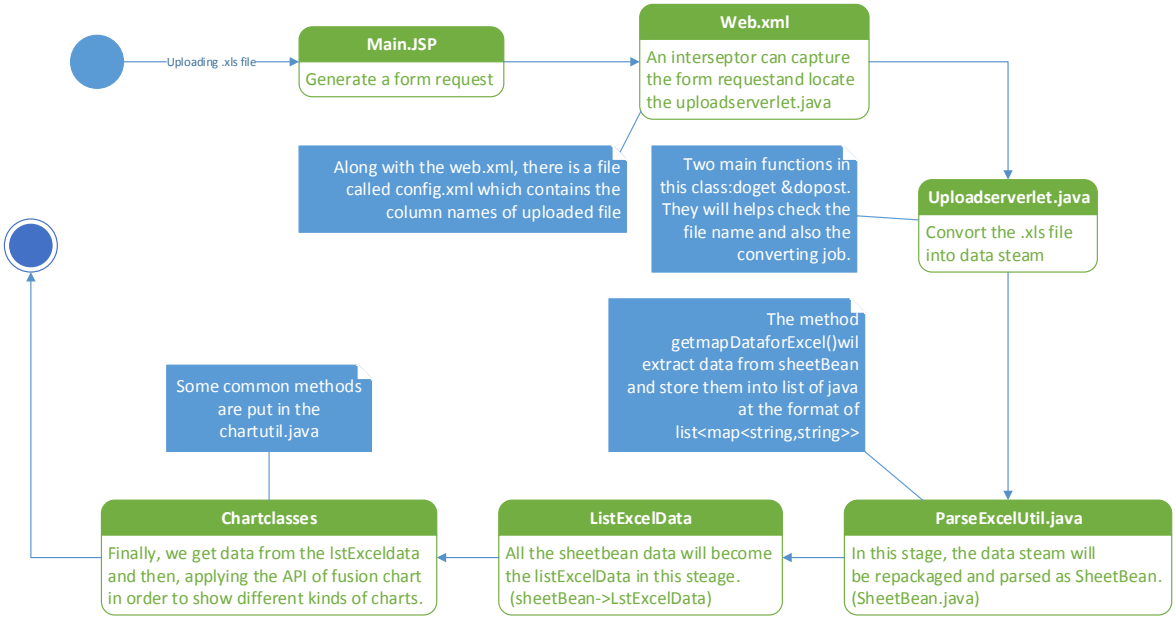


Figure 7. Connections between Classes

The Figure 7 details the connections among various classes and the ways in which they work together. This whole process starts with the file-uploading and ends with showing the chart. Once our users upload the .xls file at the main webpage, an interceptor captures the form request and locates the method in uploadserver.java. In this method, the .xls file will be converted into a data stream. Thus, in the next phase, the data stream can be repackaged and parsed as a SheetBean object. The SheetBean object will be repackaged as a ListExcelData. Finally, we extract data from the ListExcelData object and apply those data into a specific format which fits the APIs of different chart classes such as Line chart, Lattice Chart, Bar Chart, etc.

### 3.2.1. Realization of Functional Requirements

In this paper, there are five main technical approaches: Bootstrap frame, XML, POI package, FusionChart, JSP in order to complete our goals.

### 3.2.2. Details of Technical Approaches

The key technical approach that we are using in our system is the FusionChart package. Comparing to its competitor JFreeChart, it contains more types of charts and better user-interface design than JFreeChart. For commercial purposes, the user must purchase a license. A free download trial package is available and can be used for research purposes. The FusionChart Company has 23,000 customers and 500,000 users in 120 countries, including technology giants such as Apple, Google, Facebook, Intel. [27]

Also, as a part of the user interface (UI) design, the open source frame called the Bootstrap frame, is applied in our system. It is usually regarded as an excellent front-end frame work because it provides a consistent and low maintenance UI. Also, comparing to another famous framework called Foundation, it supports internet explorer 8 well and it provides wide choices for plugins and widgets. [28]

Another well-known technology called XML is used in this system as a standard data frame for the source data that is uploaded by users. Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable. Comparing to HTML, XML is much better in the schema for storing data because new tags can be defined at will and tags can be nested to arbitrary depth. Also, it can contain an optional description of its grammar. [29]

The following code is code segment from our system for re-format the original data:

```
<?xml version="1.0" encoding="UTF-8"?>
<sheetName name="dataSheet" startRowNum="1">
<column name="Year" value="Year" columnNum="1" category="true"/>
```

<column name="Residential" value="Residential" columnNum="2" />

<column name="Commer-cial" value="Commercial" columnNum="3" />

<column name="Industrial" value="Industrial" columnNum="4" />

<column name="Transpor-tation" value="Transpor-tation" columnNum="5" />

<column name="Other" value="Other" columnNum="6" />

<column name="Total" value="Total" columnNum="7" />

</sheetName>

Technically, each line of this xml codes defines a specific column in the .xsl file. The column name in the xml file is the same as the corresponding column name in the uploaded .xsl file. For example, if the first column name in the .xls file is ‘student name’, then we should write <column name = student name> along with the index ‘1’ in our xml file. As you can see the second attribute ‘value’ is the same as the column name, which will be shown in our chart. ’

In order to convert the .xls file into a Java data stream, the open-source package called POI is utilized in our system. This is a project run by the Apache Software Foundation, and previously was a sub-project of the Jakarta Project. It provides pure Java libraries for reading and writing files in Microsoft Office formats, such as Word, PowerPoint and Excel. [30] This package is clearly the best choice for converting excel files at the present time.

Finally, the popular technology Java Server Pages (JSP) supports the logical layer of our system. JSP was originally designed for dynamic web pages. The primary reason that we use this technology is that JSP can be translated and compiled into Java serverlets. Also, it uses a simple scripting language based syntax for embedding HTML into JSP.


Figure 8. Server Architecture

Because our system will be run on the tomcat server, JSP is an excellent choice for doing logical based works. (Figure 8)

### 3.2.3. System Security and Privacy

In the basement of design, our system does not contain any database layer for the purpose of storing data from users, thus the most important thing we need to do in order to improve the system security is protect the privacy of data from users while they upload them and defend the potential ping attack (DDoS).

For the first one, we encourage our user to install antivirus software for the purpose of defending the leak of credential information due to malicious software. For the second problem here, generally, we will write some scripts that try to filter out the bad traffic or we will try to build firewalls to block the traffic. As an alternative, Internet Service Provider (ISP) can be another solution. We can buy an ISP to provide DDoS mitigation.

### 3.2.4. Markov Chain Model

Theoretically, we will introduce the Markov Chain model in this section and along with some possible ways to showing prediction result on visualization chart.

#### 3.2.4.1. Basic Background

A Markov chain, named after Andrey Markov, is a mathematical system that undergoes transitions from one state to another on a state space. It is a random process usually characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it [25].

Generally, there are three main key terminologies for a Markov Chain. They are states, emission probabilities, and transition probabilities between states. Taking the electronic power consumption data of the USA as an example (The target of this prediction can be determine whether our consumption increasing is in a good or bad developing mode). Each year is modeled as a node ant the total power consumption is a state of this node (Figure 9).

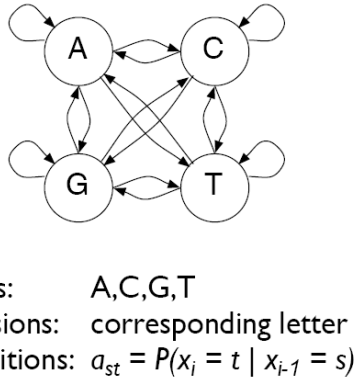


Figure 9. Nodes and Emission and Transition Probability

The emission probability and transition probity can be obtained from the following formulas:

a. The maximum likelihood (ML) approach is used to estimate the transition probabilities. (Cab is the data that b follows a, and the bottom are the counts of occurrences of b following of a).

$$a_{ab} = \frac{c_{ab}}{\sum_{b`} c_{ab`}}$$

b. The maximum likelihood (ML) approach is also used to estimate the emission probabilities. (Where Ca is the count of showing times and the denominator is the count of showing times in the past).

$$a_a = \frac{c_a}{\sum_{a`} c_{a`}}$$

After finding the data that we want, the probability that a sequence x is generated by a Markov chain model:

$$\begin{aligned} P(x) &= P(x_1, x_2, ..., x_n) \\ &= P(x_1, x_2, ..., x_{n-1}) \cdot P(x_n \mid x_1, x_2, ..., x_{n-1}) \\ &= P(x_1, x_2, ..., x_{n-2}) \cdot P(x_{n-1} \mid x_1, x_2, ..., x_{n-2}) \cdot P(x_n \mid x_1, x_2, ..., x_{n-1}) \\ &... \\ &= P(x_1) \cdot P(x_2 \mid x_1) ... P(x_n \mid x_1, x_2, ..., x_{n-1}) \end{aligned}$$

Which also can be taken as by applying many times of

P(X, Y) = P(X) * P(Y|X)

The next question is to determine the increasing consumption is in a good or bad developing mode:

$$\log \frac{P(x \mid \text{model} +)}{P(x \mid \text{model} -)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1} x_i}}{a^{-}_{x_{i-1} x_i}}$$

If the log likelihood ratio >0, then we will say the consumption of this year is in a good developing mode.

### 3.2.4.2. Prediction Part

The following MC algorithm follows the classic Markov Chain model. This classic model can be illustrated in the action of rolling dice and may be directly applied in our visualization model. In the classic dice rolling problem, we have one fair die which shares the same probability of rolling numbers (1/6 for each number) and one loaded die which

has 1/2 for 6 and 1/10 for each of the remaining numbers. If we give a sequence of numbers like 6, 2, 6, 3 we ask the question of what is the probability of a fair or loaded dice.

Returning to the model in the background chapter, we can find the emission probability and transition probability for the different states F and L (Figure 10).

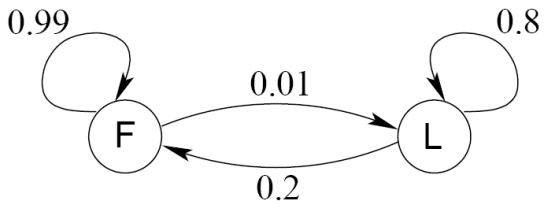


Figure 10. Sample Transition and Emission Probability of Dice

As a classic way to calculate the probability, the following table shows the result of the probability of obtaining each of the numbers (Figure 10).

Table 3. MC Calculation of Probability

| | E | 6 | 2 | 6 |
|---|---|---|---|---|
| F | 0 | (1/2) x (1/6) = **1/12** | (1/6) x max{(1/12) x 0.99, (1/4) x 0.2} = **0.01375** | (1/6) x max{0.01375 x 0.99, 0.02 x 0.2} = **0.00226875** |
| L | 0 | (1/2) x (1/2) = **1/4** | (1/10) x max{ (1/12) x 0.99, (1/4) x 0.2} = **0.01375** | (1/2) x max {0.01375 x 0.01, 0.02 x 0.8} = **0.008** |

Thus, the one that has the highest probability is more likely to be the best prediction number (Table 3).

In our data sink, each year will be taken as a single state and the total consumption will be taken as the numbers in the dice-rolling problem.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Residential | Commer-cial | Industria | Transpor- | Other | Total |
| 2 | 2002 | 116,622,037 | 15,333,700 | 601,744 | N/A | 1,066,554 | 133,624,035 |
| 3 | 2003 | 117,280,481 | 16,549,519 | 713,221 | 1,127 | N/A | 134,544,348 |
| 4 | 2004 | 118,763,768 | 16,606,783 | 747,600 | 1,025 | N/A | 136,119,176 |
| 5 | 2005 | 120,760,839 | 16,871,940 | 733,862 | 518 | N/A | 138,367,159 |
| 6 | 2006 | 122,471,071 | 17,172,499 | 759,604 | 791 | N/A | 140,403,965 |
| 7 | 2007 | 123,949,916 | 17,377,219 | 793,767 | 750 | N/A | 142,121,652 |
| 8 | 2008 | 124,937,469 | 17,562,726 | 774,713 | 727 | N/A | 143,275,635 |
| 9 | 2009 | 125,177,175 | 17,561,661 | 757,519 | 705 | N/A | 143,497,060 |
| 10 | 2010 | 125,717,935 | 17,674,338 | 747,746 | 239 | N/A | 144,140,258 |
| 11 | 2011 | 126,143,072 | 17,638,062 | 727,920 | 92 | N/A | 144,509,146 |

Figure 11. Data from EIA

According to the theory we talked above, there are two steps to needs to be done before we can apply this sample data with the MC model. The first step is determine the data groups and groups different columns into those intervals such as low status, normal status or high status. The second steps is calculate the emission and transition probabilities for each state. The intent is for our visualization system to support graphical representations of all the relevant possible states and their probabilities.

# 4. REAL-DATA VISULIZATION TESTING

In this phase, we illustrate the result of our visualization software, which is combined with the data in EIA.

## 4.1. Introduction

Our data is downloaded from the official website of EIA (U.S Energy Information Administration.). The mechanism we were using in the testing was to build a C/S based system, which has its own domain name and pre-uploaded data that can be visited through Internet Explorer.

The first test is run in our personal computer. We ran the apache server on our computer and visited local host: 8080/charts in order to visit the package, which was generated by our system in the app folder.

## 4.2. Visualization Frameworks and Display

There are four main steps to apply the system:

a. Set up the apache server with the right settings.

b. Copy and paste the .war file into the apache app folder

c. Run the apache server in the local host

d. Input the local host into your web browser and choose your resource data.

As designed, all titles and values can be revealed or hidden by one simple click on the name at the bottom of the diagram. Meanwhile, all diagrams will be automatically justified depending on the data size. In our testing, usually, it takes no more than 10 seconds to read the data.

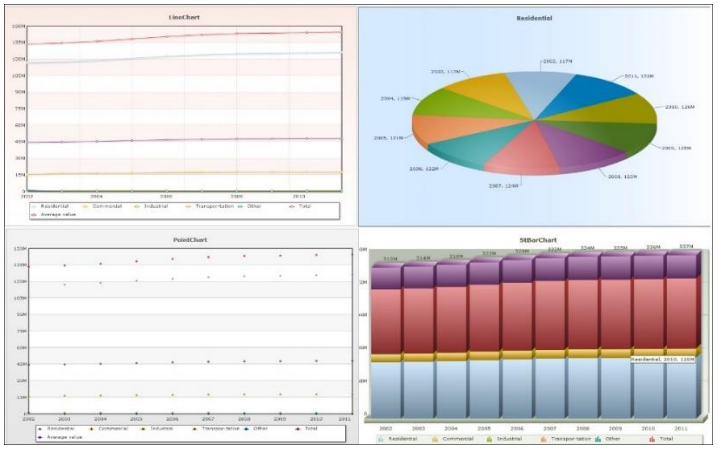


Figure 12. Charts of the Test Data

The Figure 12 shows the Bar chart, Line chart, Lattice chat, Pie chart after the user uploaded the source data from EIA.

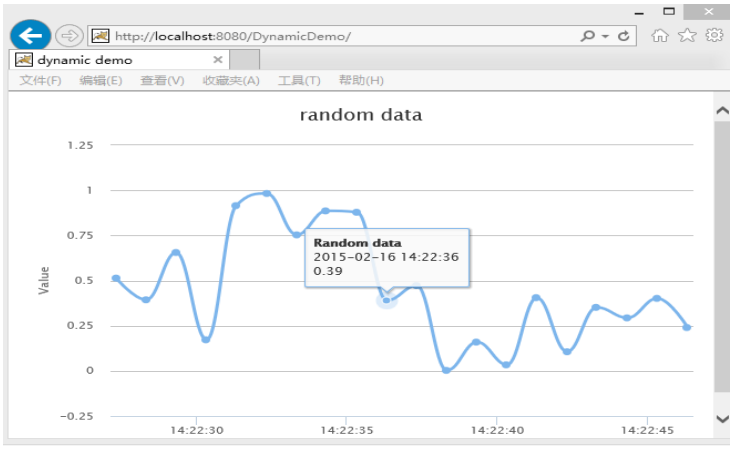


Figure 13. Motion Chart

The Figure 13 shows the conceptual motion diagram along with random data.

### 4.3. Junit Test

JUnit is a unit testing framework for the Java programming language. JUnit has been important in the development of test-driven development, and is one of a family of unit-testing frameworks which is collectively known as xUnit that originated with SUnit. [31]

Regarding the system itself, some JUnit tests have been applied to our system for the purpose of confirming the correctness of the functionalities of each method and class.

## 5. TEST RESULT

In this section, we will describe the evaluation of our system and show the results.

### 5.1. Evaluation of Test Result

To use the surrounding reality idea, “visual modeling” is the creation of an organizational model to help the developed devise a way to do something. The vital target of modeling is to promote a better understanding of requirements, a better design and to make an easily maintained system. A self-evaluated frame is shown below (Table 4 and 5):

Table 4. Statistics of Self-Satisfaction

| | Displeasure | Adequate | Satisfied | Perfect |
|---|---|---|---|---|
| UI Design | 13% | 56% | 17% | 14% |
| Functionality | 5% | 34% | 42% | 19% |
| Easy to Use | 11% | 40% | 37% | 12% |
| Over all | 3% | 59% | 21% | 17% |

On the other hand, the evaluation which depends on our original requirements is another vital factor of evaluation. As a designer of the system, I configured the following evaluation table:

Table 5. Functional Requirements Self-Evaluation Result

| | Realized | Follow the doc | Self-evaluation | Overall |
|---|---|---|---|---|
| Data transferring | 4 | 5 | 5 | 4 |
| Chart selection | 5 | 5 | 5 | 5 |
| Element selection | 5 | 5 | 4 | 5 |
| Data prediction | 3 | 4 | 4 | 3 |

Table 6. Non-Functional Requirements Self-Evaluation Result

| | Realized | Follow the doc | Future work | Overall |
|---|---|---|---|---|
| Friendly Structure | 5 | 5 | 5 | 5 |
| Easy Color selection | 2 | 5 | 5 | 5 |
| Easy used buttons | 5 | 4 | 4 | 5 |
| Autoing justify data | 5 | 5 | 5 | 5 |

In real data testing, we determined that everything we downloaded from the EIA website can be properly shown on the screen. In addition, all data are shown with correct values. However, there is an inconvenience in our system in that we cannot upload a data file without first justifying it. Thus, every time before we upload the data, we need to first justify them by the administrator. This deficiency can be corrected in the future. We tested 10 different data sets from the EIA website. The table below shows data for the self-evaluation.

Table 7. Result of Self-Evaluation of Real-Data

| File Name | Correctness | Auto Justifying | Time-Cost(in 10 secs) |
|---|---|---|---|
| Summary electricity statistics 2002–2012 | 100% | 95% | 100% |
| Supply and disposition of electricity 2002–2012 | 100% | 100% | 95% |
| Electricity overview | 100% | 100% | 100% |
| Consumption for electricity generation | 100% | 100% | 100% |
| Generating capacity | 100% | 100% | 100% |

Table 8. Result of NUnit Tests

| Class Name | Correctness of Assertions |
|---|---|
| ColumnBean.java | 100% |
| SheetBean.java | 100% |
| GlobalUtils.java | 100% |
| UploadServlet.java | 100% |
| ParseExcelUtil.java | 100% |

# 6. CONCLUSION AND FUTURE WORK

In this chapter, an overall conclusion is described. A description of the advantages and defects of our system is presented. In addition, future work ideas and methods are described.

## 6.1. Conclusion

In this study, our initial objective of building a web-based online visualization system that meets functional requirements was met. The resultant system is shown to be correct. A small number of deficiencies, such as limitations in accepted data formats, somewhat rigid user interface, and limited styles of charts are addressable in the future and do fit within our established system architecture. Below we present some approaches for improving the system and eliminating identified defects.

## 6.2. Future Work

In the work described in this paper, we established a web-based visualization system with pre-designed functionalities. However, some deficiencies remain. In improve the product, a prototype Software Design Life Cycle is applied in this study. As a result, we claim only that the system is a prototype that follows a prescribed architecture, and plan to continue to improve the system and to add useful new features. The description below presents some concepts for proceeding to address future functionalities:

The capability to read different file formats, more chart options, flexibility of styles and outputs the primary area to address.

Automatic data conversion is important because there are many and varied kinds of data sources. Among the many types of data, five of them are very common, namely pdf, xls, word, txt and database files. It is a significant challenge to extract all the keywords

from those documents and justify them into xml files. However, this requirement is important and will be needed in a more advanced system. In the next generation of the system, we expect to use data mining models to help determine the titles of each column in those files and automatically put them into a proper xml structure. In addition, outputting is another significant functionality. In the next generation of the system, users will be able to output their visualization data as jpg and png files or save them as a xls file.

To improve the flexibility of our system, more chart styles should be added. In the package of FusionChart, we have a large amount of charts that can be applied into our web-system. The following figure 14 shows some other choices for our system in the future.

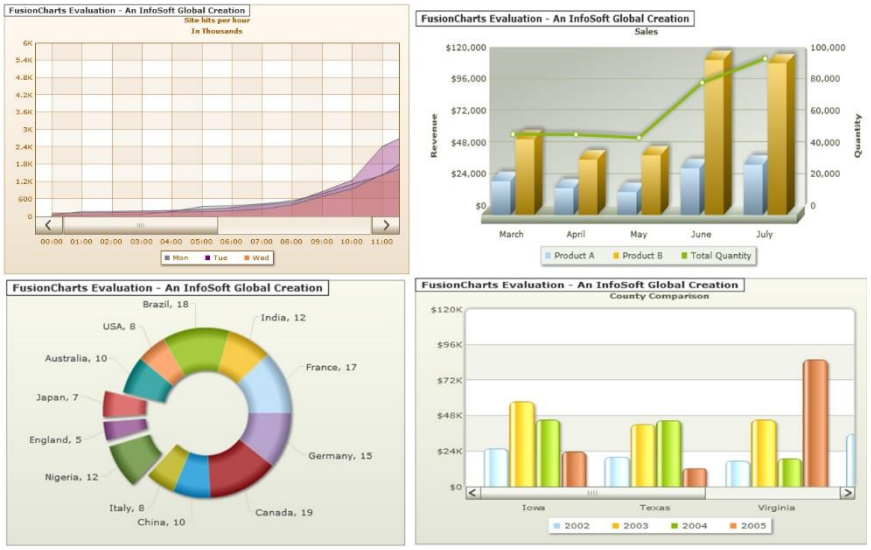


Figure 14. Future Charts

Reusability and robustness can help improve the efficiency and security of our system. The idea is that we plan to follow is to combine our system with a cloud driver and thereby help users to store useful visualization charts. Also, an authorization of user ID and password will be needed in order to improve the robustness of our system. The added flexibility will mean that a user can change the color and style of lines in each type of chart

in the next generation of the system. For the purpose of flexibility, simple buttons will be created to help our user easily find their needs.

Regarding the prediction aspect in the future version of the system, the result of MC model can be shown as an extra pie diagram or an extra element in charts for predicting the next state. The Figure 15 shows this idea.

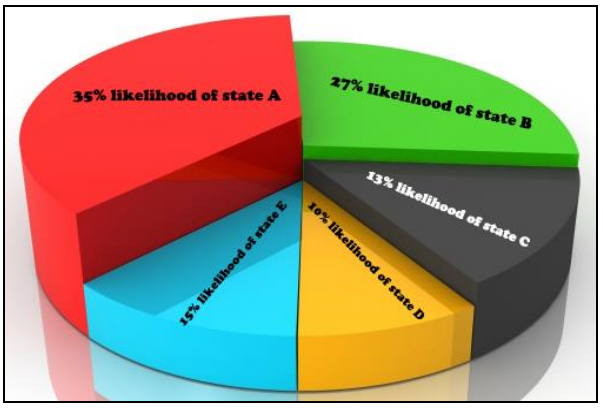


Figure 15. Future Pie Chart of MC Model Result

The highly competitive and changing business environment has led to increasing complexity, which brought unique challenges for system developers. Thus, after adding some new features, consultation meetings will be needed in order to improve this system.

## REFERENCES

[1] Retrieved from: http://www.sciencedaily.com/articles/s/scientific_visualization.htm. Retrieved 2015-3-12

[2] Retrieved from: http://en.wikipedia.org/wiki/Visualization. Retrieved 2015-3-12

[3] Eick, S. G. (1994). Graphically displaying text. Journal of Computational and Graphical Statistics, 3:127–142. 2

[4] Michael Friendly (2008). "Milestones in the history of thematic cartography, statistical graphics, and data visualization".

[5] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[6] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[7] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28

[8] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[9] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[10] Hald, A. (1990). A History of Probability and Statistics and their Application before 1750. New York: John Wiley and Sons.

[11] Pearson, Egon S., ed. (1978). The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religeous Thought. London: Griffin & Co. Ltd. ISBN 85264 250 4. Lectures by Karl Pearson given at University College London during the academic sessions 1921–1933.

[12] Porter, T. M. (1986). The Rise of Statistical Thinking 1820–1900. Princeton, NJ: Princeton University Press.

[13] Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty before 1900. Cambridge, MA: Harvard University Press.

[14] Riddell, R. C. (1980). Parameter disposition in pre-Newtonain planetary theories. Archives Hist. Exact Sci., 23:87–157.

[15] Wallis, Helen M. and Robinson, Arthur H. (1987). Cartographical Innovations: An International Handbook of Mapping Terms to 1900. Tring, Herts: Map Collector Publications. ISBN 0-906430-04-6.

[16] Hoff, Hebbel E. and Geddes, L. A. (1959). Graphic recording before Carl Ludwig: An historical summary. Archives Internationales d'Histoire des Sciences, 12:3–25.

[17] Hoff, Hebbel E. and Geddes, L. A. (1962). The beginnings of graphic recording. Isis, 53:287–324. Pt. 3.

[18] Funkhouser, H. Gray (1936). A note on a tenth century graph. Osiris, 1:260–262. URL http:// tinyurl.com/2czmqc.

[19] Funkhouser, H. Gray (1937). Historical development of the graphical representation of statistical data. Osiris, 3(1):269–405. URL http://tinyurl.com/32ema9. Reprinted Brugge, Belgium: Stz Catherine Press, 1937.

[20] Farebrother, R. W. (1999). Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900. New York: Springer. ISBN 0-387-98598-0.

[21] Friis, H. R. (1974). Statistical cartography in the United States prior to 1870 and the role of Joseph C. G. Kennedy and the U.S. Census Office. American Cartographer, 1:131–157.

[22] Robinson, Arthur H. (1982). Early Thematic Mapping in the History of Cartography. Chicago: University of Chicago Press. ISBN 0-226-72285-6.

[23] Wheeler, John Archibald (1982). Bohr, Einstein, and the strange lesson of the quantum. In R. Q. Elvee, ed., Mind in Nature. San Francisco: Harper and Row.

[24] Lending Club. Retrieved from: http://en.wikipedia.org/wiki/Lending_Club. Retrieved 2015-3-01

[25] Markov Chain. Retrieved from: http://en.wikipedia.org/wiki/Markov_chain. Retrieved 2015-2-20

[26] SAS, Data Visualization. Retrieved from: http://www.sas.com/en_us/insights/big-data/data-visualization.html. Retrieved 2015-3-20

[27] Fusion Chart. Retrieved from: http://en.wikipedia.org/wiki/FusionCharts. Retrieved 2015-3-25

[28] Bootstrap vs. Foundation. Retrieved from: https://bootstrapbay.com/blog/bootstrap-vs-foundation. Retrieved 2015-2-15

[29] HTML versus XML. Retrieved from: http://courses.cs.vt.edu/~cs1204/XML/htmlVxml.html. Retrieved 2015-4-1

[30] Apache POI. Retrieved from: http://en.wikipedia.org/wiki/Apache_POI. Retrieved 2015-4-1

[31] Junit. Retrieved from: http://en.wikipedia.org/wiki/JUnit. Retrieved 2015-3-20