

## TABLE OF CONTENTS

<i>Preface</i>	vii
<b>Chapter No. 1 What Is Data Mining?</b>	<b>1</b>
1.1 Introduction.....	1
1.2 Keywords Associated With Data Mining.....	5
1.3 Why Data Mining?.....	8
1.4 How is Data Mining Different?.....	13
1.4.1 Data Mining vs. Statistics.....	15
1.4.2 Knowledge Discovery Using Statistics.....	17
1.5 What Constitutes Data Mining?.....	19
1.6 Where Does Data Mining Fit in?.....	21
Questions .....	23
<b>Chapter No. 2 Data Pre-processing ETL (Extract Transform Load)</b>	<b>25</b>
2.1 Introduction.....	25
2.2 The ETL Cycle.....	25
2.3 The ETL Challenge.....	28
2.4 Some ETL Challenges.....	29
2.4.1 Work is underestimated.....	30
2.4.2 Diversity in Source System Platforms.....	31
2.4.3 Same Data But Different Representations.....	31
2.4.4 Multiple Sources of the Same Data.....	32
2.4.5 Complexity of Required Transformations.....	34
2.4.6 Rigidity and Unavailability of Legacy systems.....	35
2.4.7 Volume of Legacy Data.....	37
2.4.8 Web Scrapping.....	38
2.5 Data Transformation.....	39
2.5.1 Conversion.....	40
2.5.2 Enrichment.....	42
2.6 Data Cleansing.....	43
2.6.1 The Lighter Side of ‘Dirty’ Data.....	44
2.6.2 Serious Problems Due to ‘Dirty’ Data.....	45
2.7 Two Classes Of Anomalies.....	46
2.7.1 Coverage Problems .....	46
2.7.2 Key-Based Classification Problems.....	48
2.7.2.1 Primary Key Problems.....	48

---

2.7.2.2 Non-Primary Key Problems.....	49
2.8 Data Quality.....	50
Questions .....	51
<b>Chapter No. 3 Data And Data Quality Management</b>	<b>53</b>
3.1 Introduction.....	53
3.2 What is Data?.....	53
3.2.1 Attribute Values.....	54
3.2.2 Basic Data Types.....	55
3.2.3 Discrete and Continuous Attributes.....	56
3.2.4 Types of Data Sets .....	57
3.3 What is Quality?.....	60
3.3.1 Types of Data Quality.....	62
3.3.2 Orr's Laws of Data Quality .....	63
3.3.3 Dimensions of Data Quality.....	64
3.3.4 Data Quality Validation Techniques.....	65
3.3.4.1 Referential Integrity Validation .....	65
3.3.4.2 Attribute Domain Validation .....	67
3.3.4.3 Data Histograming .....	69
3.3.5 Total Quality Management (TQM).....	70
3.3.6 Where is Data Quality Critical?.....	72
Questions .....	74
<b>Chapter No. 4 What Can Data Mining do?</b>	<b>75</b>
4.1 Introduction.....	75
4.2 Classification.....	76
4.3 Estimation.....	77
4.4 Prediction.....	78
4.5 Market-Basket Analysis.....	79
4.6 Clustering.....	82
4.7 Description .....	84
4.8 Comparing The Methods.....	85
4.9 Some applications of Data Mining.....	87
4.9.1 Telecommunication.....	87
4.9.2 Insurance.....	88
4.9.3 Banking.....	89
4.9.4 Customer Acquisition.....	90
4.9.5 CRM.....	90
4.9.6 E-Commerce.....	90

4.9.7 Bioinformatics.....	91
Questions .....	92
<b>Chapter No. 5 Supervised And Unsupervised Learning</b>	<b>93</b>
5.1 Introduction .....	93
5.2 Data Structures in Data Mining.....	93
5.3 Main Types of Data Mining.....	95
5.3.1 Supervised Learning.....	95
5.3.2 Unsupervised Learning.....	96
5.3.3 Min-Max Distance.....	96
5.4 How Clustering Works?.....	98
5.5 Main Types of Clustering.....	99
5.5.1 One-way Clustering Using Simulated Data.....	99
5.5.2 One-way Clustering Using Real Data .....	100
5.5.3 Two-way Clustering Using Simulated Data .....	102
5.6 Classification .....	102
5.6.1 How Classification Works? .....	103
5.6.2 Using Classification for Prediction-Simulated Data.....	104
5.6.3 Using Classification for Prediction-Real Data .....	107
5.7 Clustering vs. Cluster Detection.....	108
5.7.1 K-means Clustering .....	109
5.7.2 Comments on K-means Clustering.....	111
Questions .....	112
<b>Chapter No. 6 Visual Data Mining</b>	<b>113</b>
6.1 Introduction.....	113
6.2 Human Factor.....	115
6.3 Pre-attentive Processing For Different Data Types.....	117
6.4 Visualization vs. Traditional Charting.....	119
6.5 Data reduction.....	123
6.5.1 Reducing The Number Of Dimension.....	123
6.5.1.1 Factor Analysis.....	123
6.5.1.2 Multi-Dimensional Scaling (MDS).....	124
6.5.1.3 Principal Component Analysis (PCA)...	126
6.5.2 Reducing the Number of Rows or Data Pre- processing.....	127
6.5.2.1 Sub-setting Techniques.....	127
6.5.2.2 Aggregation Techniques.....	129

6.6 The Cost of Generating Visualization.....	129
6.7 Some Problems With Data Visualization.....	132
6.7.1 Monitor Screen Resolution .....	132
6.7.2 How to Evaluate the Solution? .....	133
6.7.3 The Lie Factor.....	133
6.7.4 Lack of 1-to-1 Mapping.....	134
6.8 Visualization Techniques.....	135
6.8.1 Pixel-Oriented.....	135
6.8.2 Geometric Projection.....	135
6.8.2.1 Parallel Coordinates.....	136
6.8.2.2 Strengths and Weaknesses of Parallel Coordinates.....	137
6.8.3 Icon-Based Technique.....	137
6.8.4 Hierarchical & Graph Drawing-Based Techniques	138
6.9 Comparison of Techniques.....	141
Questions .....	142
<b>Chapter No. 7 Case Study :Data Mining Agriculture Extension Data Warehouse</b>	<b>143</b>
7.1 Introduction.....	143
7.2 Background.....	144
7.3 Need for an Extension of Data Warehouse.....	146
7.4 Related Work.....	147
7.5 Pilot AE-DWH (PAE-DWH) System.....	148
7.5.1 Step 1 : Determine User’s Need.....	149
7.5.1.1 Availability of Data .....	149
7.5.1.2 Cost/Benefit Analysis, Project Estimation and Risk Assessment.....	150
7.5.2 Step 2 & 3: Determine DBMS Server & Hardware Platform .....	150
7.5.3 Step 4:Information and Data Modeling.....	151
7.5.3.1 Dimensional Modeling .....	151
7.5.3.2 Logical and Physical Design of Data Warehouse.....	152
7.5.4 Step 5: Construct Meta Data Repository.....	153
7.5.4.1 Building a Meta Data Repository.....	153
7.5.4.2 Business User’s View of Meta Data.....	155
7.5.4.3 Steps to Develop an Effective Meta Data.....	155

---

7.5.5 Step 6 :Data Acquisition & Cleansing .....	155
7.5.5.1 Data Cleansing and Reconciliation .....	157
7.5.6 Step 7: Data Transform, Transport and Populate ...	158
7.5.6.1 Motivation .....	158
7.5.6.2 Methods .....	158
7.5.6.3 Results .....	159
7.5.6.4 Transporting the data .....	159
7.5.6.5 Populating and Loading the Data Warehouse .....	159
7.5.6.6 Data Validation .....	160
7.5.7 Step 8: Determine Middleware Connectivity .....	161
7.5.8 Step 9: Prototyping, querying and Reporting .....	161
7.5.9 Step 10 & 11: OLAP and Data Mining ...	161
7.5.10 Step 12: Deployment and System Management .....	161
7.6 Decision Support using PAE-DWH .....	162
7.6.1 Working Behavior at Field Level, Spray Dates ....	162
7.6.2 Working Behavior at Field Level, Sowing Dates	163
7.6.3 Sowing vs. Weather .....	165
7.7 Data Mining PAE-DWH .....	166
7.7.1 Data Considered .....	167
7.7.2 Clustering of Agro-Met Data .....	167
7.7.3 Discussion .....	168
7.8 Conclusions .....	171
Questions .....	174
<b>INDEX</b> .....	175 - 179

This page intentionally left blank

## Preface

In Dec. 2001 faculty members from four leading universities of the country flew to Cyprus for a week-long crash course on “High Performance Data Warehouse Design”, I was among those faculty members. We were very fortunate to have Dr. Stephen Brobst (CTO Teradata) as our instructor. One of the lectures in the course was about Data mining, that topic caught my imagination; I was so much impressed and motivated that finally I did my PhD in Data Mining.

To me Data Mining is very close to Algorithms; my favorite subject. Therefore, for several years I have been teaching Data Mining to my BSc and MSc students as part of my different Algorithm courses, and as part of my Data Warehouse course too. This book is based on my academic, research and professional experience in this particular field.

### To the Teacher

The book is divided into seven chapters. Chapters 1, 4 and 5 cover classical Data Mining but treated with a modern approach such as covering one-way and two-way clustering, noise, current applications, while Chapter 6 covers the emerging domain of Visual Data Mining. Chapters 2 and 3 cover the Data Processing or ETL (Extract Transform Load) and Data Quality Management. Chapter 7 is based on my journal paper and is a real-life case study using real Agriculture Data.

### To the Student

Data Mining is an exciting area, and I hope the reader will not only enjoy the book, but also benefit from it. The book is written primarily for those who would like to venture into the field of Data Mining. I have included several examples and tried to make it an easy reading. Review questions have been provided at the end of every chapter, for a better understanding of the material in the book. The last chapter consists of a case study, which shows how the techniques discussed in this book are used in day-to-day-life. I can say without doubt that Data Mining has changed my life, I am sure this book can agitate your mind to enter into this wonderful world of knowledge discovery and its interpretation.

Ahsan Abdullah

**URL:** [www.ahsanabdullah.com](http://www.ahsanabdullah.com)

### Note

In spite of all the care taken in editing of the book, there are sure to be errors and mistakes in the book. I will be grateful if the readers inform me on discovering those errors and mistakes along with suggestions for improvement for the next edition of the book. My email address is [ahsan1010@yahoo.com](mailto:ahsan1010@yahoo.com)