

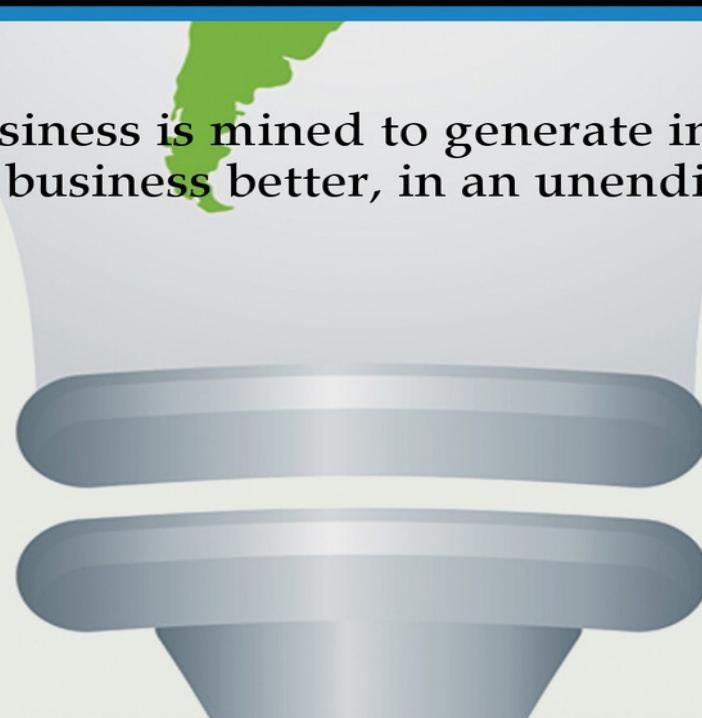


# Data Analytics



## Made Accessible

Data from business is mined to generate intelligence.  
This makes business better, in an unending cycle.



Dr. Anil Maheshwari

# **Data Analytics Made Accessible**

Copyright © 2014-20 by Anil K. Maheshwari, Ph.D.

By purchasing this book, you agree not to copy the book by any means, mechanical or electronic.

No part of this book may be copied or transmitted without written permission.

Other Books by the same author:

[Big Data Made Accessible](#)

[Moksha: Liberation Through Transcendence](#)

[Transformational Leadership: Spirit In Action](#)

[Marketing Made Accessible](#)

## **Select list of Universities that use this book as a Textbook**

1. James Madison University, Virginia
2. San Diego State University, California
3. Boston University, Massachusetts
4. Boise State University, Idaho
5. Maharishi University of Management, Iowa
6. Morehead State University, Kentucky
7. Regent University, Iowa
8. University of Alaska, Anchorage
9. Victoria University of Wellington, New Zealand
10. Letterkenny Institute of Technology, Ireland
11. University of Los Andes, Chile
12. Adolfo Ibáñez University, Chile

## Select Reviewer Comments for this book

“Dr. Maheshwari's book is a very nice introduction to analytics. He explains concepts very clear and to the point. I like particularly his chapter on decision trees and the process to obtain them. His explanation is very clear.” – Dr. Ramon A. Mata Toledo, Professor of Computer Science, James Madison University, Virginia.

“This book is a splendid and valuable addition to this subject. The whole book is well-written and I have no hesitation to recommend that this can be adapted as a textbook for graduate courses in Business Intelligence and Data Mining.” Dr. Edi Shivaji, St. Louis, Missouri.

“Really well written and timely as the world gets in the Big Data mode! I think this can be a good bridge and primer for the uninitiated manager who knows Big Data is the future but doesn't know where to begin!” – Dr. Alok Mishra, Singapore.

“This book has done a great job of taking a complex, highly important subject area and making it accessible to everyone. It begins by simply connecting to what you know, and then bang - you've suddenly found out about Decision Trees, Regression Models and Artificial Neural Networks, not to mention cluster analysis, web mining and Big Data.” – Ms. Charmaine Oak, United Kingdom.

“Bottom line is anyone interested in learning about Data Analytics this is THE book to start your learning path with and hope it will spur your interest in the field and enable you to grasp more in-depth topics and further your skills.” – Keith S Safford, Maryland.

## Preface to 2020 edition

Data Science continues to be the hottest discipline, along with Artificial Intelligence, as a new decade emerges. *This 2020 edition includes a new chapter on Data Privacy, which has emerged as a hot issue.* I have been teaching courses in Data Analytics and Big Data for many years. My students found other textbooks seem too long, too technical, too complex, or too focused on specific tools. My goal was to write a conversational book that feels easy and informative. This is an accessible book that covers everything important, with concrete examples, and invites the reader to join this field. Many young people have entered this field upon reading this book.

This book can be easily read in a short period by anyone who wants to understand data-based decision-making for their business or other organizations, without any expertise with software tools. The text is almost entirely devoid of complex jargon or programming code. It includes a primer on Statistics, Databases, Big Data, and Artificial Intelligence for those interested in these topics. A simple yet extensive tutorial on R, a popular, free, open-source, high-level programming language that is very popular with data scientists, is included at the end of the book. The book also includes a short tutorial on Python, another very popular high-level general-purpose programming language.

This book developed from my own class notes. It reflects my three decades of global IT experience in academia and industry. The chapters are organized for a typical one-semester graduate course. There are case-studies from real-world stories at the beginning of every chapter. There is a running case study across the chapters as exercises. There are review questions at the end of each chapter. Many universities around the world have adopted this as a textbook for their courses. The book has continued to evolve in response to the thoughts and suggestions expressed by the reviewers and students. Thanks to many reviewers for sharing many ideas for improvement. This book has also been translated in Chinese language.

My family had encouraged and supported me in first writing this book in 2014. I am also grateful to my students and others who read this book and provide feedback and suggestions. Finally, thanks to Maharishi Mahesh Yogi for providing a wonderful environment of consciousness-based education that provided the inspiration for writing this book from wholeness. This book is a ‘*guru-dakshina*’, or homage, to Maharishi.

Dr. Anil K. Maheshwari  
Fairfield, IA;  
December 2019

## Contents

[Preface to 2020 edition](#)

[Chapter 1: Wholeness of Data Analytics](#)

[Introduction](#)

[Business Intelligence](#)

[Caselet: MoneyBall - Data Mining in Sports](#)

[Pattern Recognition](#)

[Types of Patterns](#)

[Finding a Pattern](#)

[Uses of Patterns](#)

[Data Processing Chain](#)

[Data](#)

[Database](#)

[Data Warehouse](#)

[Data Mining](#)

[Data Visualization](#)

[Terminology and Careers](#)

[Organization of the book](#)

[Review Questions](#)

[Section 1](#)

[Chapter 2: Business Intelligence Concepts and Applications](#)

[Introduction](#)

[Caselet: Khan Academy – BI in Education](#)

[BI for better decisions](#)

[Decision types](#)

[BI Tools](#)

[BI Skills](#)

[BI Applications](#)

[Customer Relationship Management](#)

[Healthcare and Wellness](#)

[Education](#)

[Retail](#)

[Banking](#)

[Financial Services](#)

[Insurance](#)

[Manufacturing](#)

[Telecom](#)

[Public Sector](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 1](#)

### [Chapter 3: Data Warehousing](#)

[Introduction](#)

[Caselet: University Health System – BI in Healthcare](#)

[Design Considerations for DW](#)

[DW Development Approaches](#)

[DW Architecture](#)

[Data Sources](#)

[Data Loading Processes](#)

[Data Warehouse Design](#)

[DW Access](#)

[DW Best Practices](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 2](#)

### [Chapter 4: Data Mining](#)

[Introduction](#)

[Caselet: Target Corp – Data Mining in Retail](#)

[Gathering and selecting data](#)

[Data cleansing and preparation](#)

[Outputs of Data Mining](#)

[Evaluating Data Mining Results](#)

[Data Mining Techniques](#)

[Tools and Platforms for Data Mining](#)

[Data Mining Best Practices](#)

[Myths about data mining](#)

[Data Mining Mistakes](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 3](#)

### [Chapter 5: Data Visualization](#)

[Introduction](#)

[Caselet: Dr Hans Gosling - Visualizing Global Public Health](#)

[Excellence in Visualization](#)

[Types of Charts](#)

[Visualization Example](#)

[Visualization Example phase -2](#)

[Tips for Data Visualization](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 4](#)

[Section 2 – Popular Data Mining Techniques](#)

[Chapter 6: Decision Trees](#)

[Introduction](#)

[Caselet: Predicting Heart Attacks using Decision Trees](#)

[Decision Tree problem](#)

[Decision Tree Construction](#)

[Lessons from constructing trees](#)

[Decision Tree Algorithms](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 5](#)

[Chapter 7: Regression](#)

[Introduction](#)

[Caselet: Data driven Prediction Markets](#)

[Correlations and Relationships](#)

[Visual look at relationships](#)

[Regression Exercise](#)

[Non-linear regression exercise](#)

[Logistic Regression](#)

[Advantages and Disadvantages of Regression Models](#)

[Conclusion](#)

[Review Exercises:](#)

[Liberty Stores Case Exercise: Step 6](#)

[Chapter 8: Artificial Neural Networks](#)

[Introduction](#)

[Caselet: IBM Watson - Analytics in Medicine](#)

[Business Applications of ANN](#)

[Design Principles of an Artificial Neural Network](#)

[Representation of a Neural Network](#)

[Architecting a Neural Network](#)

[Developing an ANN](#)

[Advantages and Disadvantages of using ANNs](#)

[Conclusion](#)

[Review Exercises](#)

[Chapter 9: Cluster Analysis](#)

[Introduction](#)

[Caselet: Cluster Analysis](#)

[Applications of Cluster Analysis](#)

[Definition of a Cluster](#)

[Representing clusters](#)

[Clustering techniques](#)

[Clustering Exercise](#)

[K-Means Algorithm for clustering](#)

[Selecting the number of clusters](#)

[Advantages and Disadvantages of K-Means algorithm](#)

[Conclusion](#)

[Review Exercises](#)

[Liberty Stores Case Exercise: Step 7](#)

[Chapter 10: Association Rule Mining](#)

[Introduction](#)

[Caselet: Netflix: Data Mining in Entertainment](#)

[Business Applications of Association Rules](#)

[Representing Association Rules](#)

[Algorithms for Association Rule](#)

[Apriori Algorithm](#)

[Association rules exercise](#)

[Creating Association Rules](#)

[Conclusion](#)

[Review Exercises](#)

[Liberty Stores Case Exercise: Step 8](#)

[Section 3 – Advanced Mining](#)

[Chapter 11: Text Mining](#)

[Introduction](#)

[Caselet: WhatsApp and Private Security](#)

[Text Mining Applications](#)

[Text Mining Process](#)

[Term Document Matrix](#)

[Mining the TDM](#)

[Comparing Text Mining and Data Mining](#)

[Text Mining Best Practices](#)

[Conclusion](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step 9](#)

[Chapter 12: Naïve Bayes Analysis](#)

[Introduction](#)

[CASELET: Fraud detection in government contracts](#)

[Probability](#)

[Naïve-Bayes model](#)

[Simple classification example](#)

[Text Classification Example](#)

[Advantages and Disadvantages of Naïve Bayes](#)

[Summary](#)

[Review Questions](#)

[Chapter 13: Support Vector Machines](#)

[Introduction](#)

[SVM model](#)

[The Kernel Method](#)

[Advantages and disadvantages](#)

[Summary](#)

[Review Questions](#)

[Chapter 14: Web Mining](#)

[Introduction](#)

[Web content mining](#)

[Web structure mining](#)

[Web usage mining](#)

[Web Mining Algorithms](#)

[Conclusion](#)

[Review Questions](#)

[Chapter 15: Social Network Analysis](#)

[Introduction](#)

[Caselet: The Social Life of Books](#)

[Applications of SNA](#)

[Network topologies](#)

[Techniques and algorithms](#)

[Finding Sub-networks](#)

[Computing importance of nodes](#)

[PageRank](#)

[Practical considerations](#)

[Comparing SNA with Data Analytics](#)

[Conclusion](#)

[Review Questions](#)

[Section 4 - Primers](#)

[Chapter 16: Big Data Primer](#)

[Introduction](#)

[Understanding Big Data](#)

[CASELET: IBM Watson: A Big Data system](#)

[Capturing Big Data](#)

[Volume of Data](#)

[Velocity of Data](#)

[Variety of Data](#)

[Veracity of Data](#)

[Benefitting from Big Data](#)

[Management of Big Data](#)

[Organizing Big Data](#)

[Analyzing Big Data](#)

[Technology Challenges for Big Data](#)

[Storing Huge Volumes](#)

[Ingesting streams at an extremely fast pace](#)

[Handling a variety of forms and functions of data](#)

[Processing data at huge speeds](#)

[Conclusion and Summary](#)

[Review Questions](#)

[Liberty Stores Case Exercise: Step P1](#)

[Chapter 17: Data Modeling Primer](#)

[Introduction](#)

[Evolution of data management systems](#)

[Relational Data Model](#)

[Implementing the Relational Data Model](#)

[Database management systems \(DBMS\)](#)

[Structured Query Language](#)

[Conclusion](#)

[Review Questions](#)

[Chapter 18: Statistics Primer](#)

[Introduction](#)

[Descriptive Statistics](#)

[Example data set](#)

[Computing Mean, Median, Mode](#)

[Computing the range and variance](#)

[Histograms](#)

[Normal Distribution and Bell Curve](#)

[Inferential Statistics](#)

[Random sampling](#)

[Confidence Interval](#)

[Predictive Statistics](#)

[Summary](#)

[Review Questions](#)

[Chapter 19 - Artificial Intelligence Primer](#)

[CASELET: Apple Siri Voice-activated personal assistant](#)

[AI, Machine Learning, and Deep Learning](#)

[The Industrial Revolution](#)

[The Information Revolution](#)

[The Cognitive \(or AI\) revolution](#)

[Jobs Losses and Gains](#)

[AI and Existential Threat](#)

[Conclusion](#)

[Review Questions](#)

[Chapter 20: Data Ownership and Privacy](#)

[Data Ownership](#)

[Data Privacy](#)

[Data Privacy Models](#)

[Chinese Model](#)

[US Model](#)

[European Model](#)

[Summary](#)

[Chapter 21: Data Science Careers](#)

[Data Scientist](#)

[Data Engineer](#)

[Data Science aptitude](#)

[Popular Skills](#)

[Appendix: R Tutorial for Data Mining](#)

[Getting Started with R](#)

[Installing R](#)

[Working on R](#)

[Import a Dataset in R](#)

[Data visualization](#)

[Plotting a Histogram](#)

[Plotting a Bar Chart](#)

[Ploting charts side by side](#)

[Data Mining Techniques](#)

[Decision Tree](#)

[Correlation](#)

[Regression](#)

[Clustering – Kmeans \(Unsupervised Learning\)](#)

[Big Data Mining](#)

[WordCloud](#)

[Twitter Mining](#)

[Steps on Twitter side](#)

[R Script](#)

[Page Rank](#)

[Additional Documentation](#)

[Appendix: Python Tutorial for Data Mining](#)

[1 About this Tutorial](#)

[2 Getting Started](#)

[3 Installation](#)

[4 Working on Python](#)

[4.1 Windows 7](#)

[4.2 Windows 10](#)

[4.3 Python Help and Tutorial](#)

[4.4 Import a Dataset in Python](#)

[4.5 Data visualization –](#)

[4.5.1 Ploting a Histogram](#)

[4.5.2 Plotting a Bar Chart](#)

[4.5.3 Ploting charts side by side](#)

[5 Data Mining Techniques](#)

[5.1 Decision Tree \(Supervised Learning\)](#)

[5.2 Regression \(Supervised Learning\)](#)

[5.3 Correlation \(Supervised Learning\)](#)

[5.4 Clustering – Kmeans \(Unsupervised Learning\)](#)

[6 Big Data Mining](#)

[6.1 WordCloud - directory FWordCloud and look at code module WordCloud.py.](#)

[6.2 Twitter Mining](#)

[6.2.1 Steps \(Twitter side\)](#)

[6.2.2 Python code](#)

[6.3 Page Rank](#)

[7 Additional Documentation](#)

[Additional Resources](#)

[About the Author](#)

# Chapter 1: Wholeness of Data Analytics

## Introduction

Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place. Business activities are recorded on paper or using electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole. All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning. These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues on (Figure 1.1).

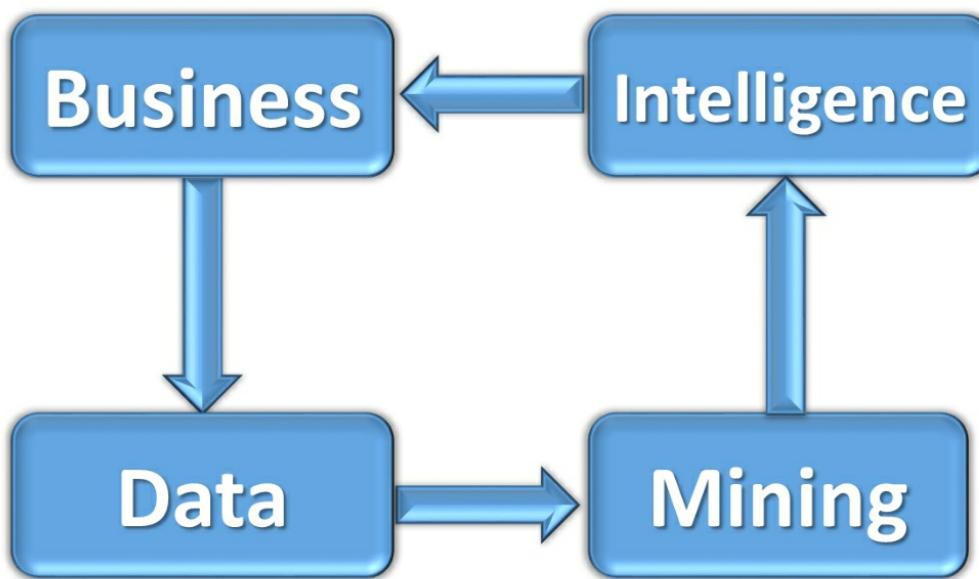


Figure 1.1: Business Intelligence and Data Mining (BIDM) Cycle

## Business Intelligence

Any business organization needs to continually monitor its business environment and its own performance, and then rapidly adjust its future plans. This includes monitoring the industry, the competitors, the suppliers, and the customers. The organization needs to also develop a balanced scorecard to track its own health and vitality. Executives typically determine what they want to track based on their key performance Indexes (KPIs) or key result areas (KRAs). Customized reports need to be designed to deliver the required information to every executive. These reports can be converted into customized dashboards that deliver the information rapidly and in easy-

to-grasp formats.

### **Caselet: MoneyBall - Data Mining in Sports**

*Analytics in sports was made popular by the book and movie, Moneyball. Statistician Bill James and Oakland A's general manager, Billy Bean, placed emphasis on crunching numbers and data instead of watching an athlete's style and looks. Their goal was to make a team better while using fewer resources. The key action plan was to pick important role players at a lower cost while avoiding the famous players who demand higher salaries but may provide a low return on a team's investment. Rather than relying on the scouts' experience and intuition Bean selected players based almost exclusively on their on-base percentage (OBP). By finding players with a high OBP but, with characteristics that lead scouts to dismiss them, Bean assembled a team of undervalued players with far more potential than the A's hamstrung finances would otherwise allow.*

*Using this strategy, they proved that even small market teams can be competitive — a case in point, the Oakland A's. In 2004, two years after adopting the same sabermetric model, the Boston Red Sox won their first World Series since 1918. (Source: Moneyball, 2004).*

**Q:** Could similar techniques apply to the games of soccer, or cricket? If so, how?

**Q2:** What are the general lessons from this story?

Business intelligence is a broad set of information technology (IT) solutions that includes tools for gathering, analyzing, and reporting information to the users about performance of the organization and its environment. These IT solutions are among the most highly prioritized solutions for investment.

Consider a retail business chain that sells many kinds of goods and services around the world, online and in physical stores. It generates data about sales, purchases, and expenses from multiple locations and time frames. Analyzing this data could help identify fast-selling items, regional-selling items, seasonal items, fast-growing customer segments, and so on. It might also help generate ideas about what products sell together, which people tend to buy which products, and so on. These insights and intelligence can help design better promotion plans, product bundles, and store layouts, which in turn lead

to a better-performing business.

The vice president of sales of a retail company would want to track the sales to date against monthly targets, the performance of each store and product category, and the top store managers that month. The vice president of finance would be interested in tracking daily revenue, expense, and cash flows by store; comparing them against plans; measuring cost of capital; and so on.

## Pattern Recognition

A pattern is a design or model that helps grasp something. Patterns help connect things that may not appear to be connected. Patterns help cut through complexity and reveal simpler understandable trends. Patterns can be as definitive as hard scientific rules, like the rule that the sun always rises in the east. They can also be simple generalizations, such as the Pareto principle, which states that 80 percent of effects come from 20 percent of the causes.

A perfect pattern or model is one that (a) accurately describes a situation, (b) is broadly applicable, and (c) can be described in a simple manner.  $E=MC^2$  would be such a general, accurate, and simple (GAS) model. Very often, all three qualities are not achievable in a single model, and one has to settle for two of three qualities in the model.

### Types of Patterns

Patterns can be temporal, which is something that regularly occurs over time. Patterns can also be spatial, such as things being organized in a certain way. Patterns can be functional, in that doing certain things leads to certain effects. Good patterns are often symmetric. They echo basic structures and patterns that we are already aware of.

A temporal rule would be that “some people are always late,” no matter what the occasion or time. Some people may be aware of this pattern and some may not be. Understanding a pattern like this would help dissipate a lot of unnecessary frustration and anger. One can just joke that some people are born “10 minutes late,” and laugh it away. Similarly, Parkinson’s law states that work expands to fill up all the time available to do it.

A spatial pattern, following the 80–20 rule, could be that the top 20 percent of customers lead to 80 percent of the business. Or 20 percent of products generate 80 percent of the business. Or 80 percent of incoming customer service calls are related to just 20 percent of the products. This last pattern

may simply reveal a discrepancy between a product's features and what the customers believe about the product. The business can then decide to invest in educating the customers better so that the customer service calls can be significantly reduced.

A functional pattern may involve test-taking skills. Some students perform well on essay-type questions. Others do well in multiple-choice questions. Yet other students excel in doing hands-on projects, or in oral presentations. An awareness of such a pattern in a class of students can help the teacher design a balanced testing mechanism that is fair to all.

However, long-established patterns can also be broken. The past cannot always predict the future. A pattern like "all swans are white" does not mean that there may not be a black swan. Once enough anomalies are discovered, the underlying pattern itself can shift. The economic meltdown in 2008 to 2009 was because of the collapse of the accepted pattern, that is, "housing prices always go up." A deregulated financial environment made markets more volatile and led to greater swings in markets, leading to the eventual collapse of the entire financial system.

### Finding a Pattern

Diamond mining is the act of digging into large amounts of unrefined ore to discover precious gems or nuggets. Similarly, data mining is the act of digging into large amounts of raw data to discover unique nontrivial useful patterns. Data is cleaned up, and then special tools and techniques can be applied to search for patterns. Diving into clean and nicely organized data from the right perspectives can increase the chances of making the right discoveries.

A skilled diamond miner knows what a diamond looks like. Similarly, a skilled data miner should know what kinds of patterns to look for. The patterns are essentially about what hangs together and what is separate. Therefore, knowing the business domain well is very important. It takes knowledge and skill to discover the patterns. It is like finding a needle in a haystack. Sometimes the pattern may be hiding in plain sight. At other times, it may take a lot of work, and looking far and wide, to find surprising useful patterns. Thus, a systematic approach to mining data is necessary to efficiently reveal valuable insights.

### Uses of Patterns

When used in clever ways, data mining can lead to interesting insights and be

a source of new ideas and initiatives.

Retaining students is an ongoing challenge for universities. Recent data-based research shows that students leave a school for social reasons more than they do for academic reasons. This pattern/insight can instigate schools to pay closer attention to students engaging in extracurricular activities and developing stronger bonds at school. The school can invest in entertainment activities, sports activities, camping trips, and other activities. The school can also begin to actively gather data about every student's participation in those activities, to predict at-risk students and take corrective action.

As another instance, the attitude of employees toward their employer may be hypothesized to be determined by a large number of factors, such as level of education, income, tenure in the company, and gender. It may be surprising if the data reveals that the attitudes are determined first and foremost by their age bracket. Such a simple insight could be powerful in designing organizations effectively. The data miner has to be open to any and all possibilities.

One can predict the traffic pattern on highways from the movement of cell phone (in the car) locations on the highway. If the locations of cell phones on a highway or roadway are not moving fast enough, it may be a sign of traffic congestion. Telecom companies can thus provide real-time traffic information to the drivers on their cell phones, or on their GPS devices, without the need of any video cameras or traffic reporters.

Similarly, organizations can find out an employee's arrival time at the office by when their cell phone shows up in the parking lot. Observing the record of the swipe of the parking permit card in the company parking garage can inform the organization whether an employee is in the office building or out of the office at any moment in time.

Some patterns may be so sparse that a very large amount of diverse data has to be seen together to notice any connections. For instance, locating the debris of a flight that may have vanished midcourse would require bringing together data from many sources, such as satellites, ships, and navigation systems. The raw data may come with various levels of quality, and may even be conflicting. The data at hand may or may not be adequate for finding good patterns. Additional dimensions of data may need to be added to help solve the problem.

## Data Processing Chain

Data is the new natural resource. Implicit in this statement is the recognition of hidden value in data. Data lies at the heart of business intelligence. There is a sequence of steps to be followed to benefit from the data in a systematic way. Data can be modeled and stored in a database. Relevant data can be extracted from the operational data stores according to certain reporting and analyzing purposes, and stored in a data warehouse. The data from the warehouse can be combined with other sources of data, and mined using data mining techniques to generate new insights. The insights need to be visualized and communicated to the right audience in real time for competitive advantage. Figure 1.2 explains the progression of data processing activities. The rest of this chapter will cover these five elements in the data processing chain.

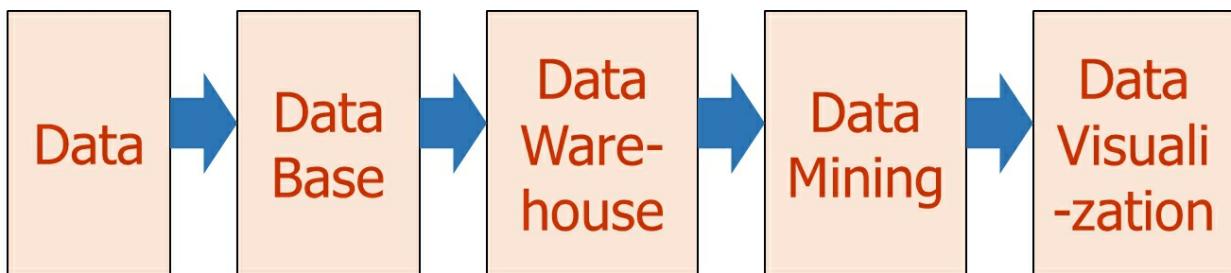


Figure 1.2: Data Processing Chain

### Data

Anything that is recorded is data. Observations and facts are data. Anecdotes and opinions are also data, of a different kind. Data can be numbers, like the record of daily weather, or daily sales. Data can be alphanumeric, such as the names of employees and customers.

1. Data could come from any number of sources. It could come from operational records inside an organization, and it can come from records compiled by the industry bodies and government agencies. Data could come from individuals telling stories from memory and from people's interaction in social contexts. Data could come from machines reporting their own status or from logs of web usage.
2. Data can come in many ways. It may come as paper reports. It may come as a file stored on a computer. It may be words spoken over the phone. It may be e-mail or chat on the Internet. It may

come as movies and songs in DVDs, and so on.

3. There is also data about data. It is called metadata. For example, people regularly upload videos on YouTube. The format of the video file (whether it was a high resolution file or lower resolution) is metadata. The information about the time of uploading is metadata. The account from which it was uploaded is also metadata. The record of downloads of the video is also metadata.

Data can be of different types.

1. Data could be an unordered collection of values. For example, a retailer sells shirts of red, blue, and green colors. There is no intrinsic ordering among these color values. One can hardly argue that any one color is higher or lower than the other. This is called nominal (means names) data.
2. Data could be ordered values like small, medium and large. For example, the sizes of shirts could be extra-small, small, medium, and large. There is clarity that medium is bigger than small, and large is bigger than medium. But the differences may not be equal. This is called ordinal (ordered) data.
3. Another type of data has discrete numeric values defined in a certain range, with the assumption of equal distance between the values. Customer satisfaction score may be ranked on a 10-point scale with 1 being lowest and 10 being highest. This requires the respondent to carefully calibrate the entire range as objectively as possible and place his own measurement in that scale. This is called interval (equal intervals) data.
4. The highest level of numeric data is ratio data which can take on any numeric value. The weights and heights of all employees would be exact numeric values. The price of a shirt will also take any numeric value. It is called ratio (any fraction) data.
5. There is another kind of data that does not lend itself to much mathematical analysis, at least not directly. Such data needs to be first structured and then analyzed. This includes data like audio, video, and graphs files, often called BLOBs (Binary Large Objects). These kinds of data lend themselves to different forms of analysis and mining. Songs can be described as happy or sad,

fast-paced or slow, and so on. They may contain sentiment and intention, but these are not quantitatively precise.

The precision of analysis increases as data becomes more numeric. Ratio data could be subjected to rigorous mathematical analysis. For example, precise weather data about temperature, pressure, and humidity can be used to create rigorous mathematical models that can accurately predict future weather.

Data may be publicly available and sharable, or it may be marked private. Traditionally, the law allows the right to privacy concerning one's personal data. There is a big debate on whether the personal data shared on social media conversations is private or can be used for commercial purposes.

*Datafication* is a new term that means that almost every phenomenon is now being observed and stored. More devices are connected to the Internet. More people are constantly connected to “the grid,” by their phone network or the Internet, and so on. Every click on the web, and every movement of the mobile devices, is being recorded. Machines are generating data. The “Internet of things” is growing faster than the Internet of people. All of this is generating an exponentially growing volume of data, at high velocity. Kryder’s law predicts that the density and capability of hard drive storage media will double every 18 months. As storage costs keep coming down at a rapid rate, there is a greater incentive to record and store more events and activities at a higher resolution. Data is getting stored in more detailed resolution, and many more variables are being captured and stored.

## Database

A database is a modeled collection of data that is accessible in many ways. A data model can be designed to integrate the operational data of the organization. The data model abstracts the key entities involved in an action and their relationships. Most databases today follow the relational data model and its variants. Each data modeling technique imposes rigorous rules and constraints to ensure the integrity and consistency of data over time.

Take the example of a sales organization. A data model for managing customer orders will involve data about customers, orders, products, and their interrelationships. The relationship between the customers and orders would be such that one customer can place many orders, but one order will be placed by one and only one customer. It is called a one-to-many relationship. The relationship between orders and products is a little more complex. One

order may contain many products. And one product may be contained in many different orders. This is called a many-to-many relationship. Different types of relationships can be modeled in a database.

Databases have grown tremendously over time. They have grown in complexity in terms of number of the objects and their properties being recorded. They have also grown in the quantity of data being stored. A decade ago, a terabyte-sized database was considered big. Today databases are in petabytes and exabytes. Video and other media files have greatly contributed to the growth of databases. E-commerce and other web-based activities also generate huge amounts of data. Data generated through social media has also generated large databases. The e-mail archives, including attached documents of organizations, are in similar large sizes.

Many database management software systems (DBMSs) are available to help store and manage this data. These include commercial systems, such as Oracle and DB2 system. There are also open-source, free DBMS, such as MySQL and Postgres. These DBMSs help process and store millions of transactions worth of data every second.

Here is a simple database of the sales of movies worldwide for a retail organization. It shows sales transactions of movies over three quarters. Using such a file, data can be added, accessed, and updated as needed.

Movies Transactions Database				
Order #	Date sold	Product name	Location	Amount
1	April 2018	Monty Python	US	\$9
2	May 2018	Gone With the Wind	US	\$15
3	June 2018	Monty Python	India	\$9
4	June 2018	Monty Python	UK	\$12
5	July 2018	Matrix	US	\$12
6	July 2018	Monty Python	US	\$12
7	July 2018	Gone With the Wind	US	\$15
8	Aug 2018	Matrix	US	\$12
9	Sept 2018	Matrix	India	\$12
10				

	Sept 2018	Monty Python	US	\$9
11	Sept 2018	Gone With the Wind	US	\$15
12	Sept 2018	Monty Python	India	\$9
13	Nov 2018	Gone With the Wind	US	\$15
14	Dec 2018	Monty Python	US	\$9
15	Dec 2018	Monty Python	US	\$9

## Data Warehouse

A data warehouse is an organized store of data from all over the organization, specially designed to help make management decisions. Data can be extracted from operational database to answer a particular set of queries. This data, combined with other data, can be rolled up to a consistent granularity and uploaded to a separate data store called the data warehouse. Therefore, the data warehouse is a simpler version of the operational data base, with the purpose of addressing reporting and decision-making needs only. The data in the warehouse cumulatively grows as more operational data becomes available and is extracted and appended to the data warehouse. Unlike in the operational database, the data values in the warehouse are not updated.

To create a simple data warehouse for the movies sales data, assume a simple objective of tracking sales of movies and making decisions about managing inventory. In creating this data warehouse, all the sales transaction data will be extracted from the operational data files. The data will be rolled up for all combinations of time period and product number. Thus, there will be one row for every combination of time period and product. The resulting data warehouse will look like the table that follows.

Movies Sales Data Warehouse			
Row#	Qtr sold	Product name	Amount
1	Q2	Gone With the Wind	\$15
2	Q2	Monty Python	\$30
3	Q3	Gone With the Wind	\$30
4	Q3	Matrix	\$36

5	Q3	Monty Python	\$30
6	Q4	Gone With the Wind	\$15
7	Q4	Monty Python	\$18

The data in the data warehouse is at much less detail than the transaction database. The data warehouse could have been designed at a lower or higher level of detail, or granularity. If the data warehouse were designed on a monthly level, instead of a quarterly level, there would be many more rows of data. When the number of transactions approaches millions and higher, with dozens of attributes in each transaction, the data warehouse can be large and rich with potential insights. One can then mine the data (slice and dice) in many different ways and discover unique meaningful patterns. Aggregating the data helps improve the speed of analysis. A separate data warehouse allows analysis to go on separately in parallel, without burdening the operational database systems (Table 1.1).

Dimension	Database	Data Warehouse
Purpose	Data stored in databases can be used for many purposes including day-to-day operations	Data stored in DW is cleansed data useful for reporting and analysis
Granularity	Highly granular data including all activity and transaction details	Lower granularity data; rolled up to certain key dimensions of interest
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields	Typically organized around a large fact tables, and many lookup tables
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size.	Grows as data from operational databases is rolled-up and appended every day. Data is retained for long-term trend analyses
Architectural	Relational, and object-	Star schema, or Snowflake

choices	oriented, databases	schema
Data Access mechanisms	Primarily through high level languages such as SQL. Traditional programming access DB through Open DataBase Connectivity (ODBC) interfaces	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

Table 1.1: Comparing Database systems with Data Warehousing systems

## Data Mining

Data Mining is the art and science of discovering useful innovative patterns from data. There is a wide variety of patterns that can be found in the data. There are many techniques, simple or complex, that help with finding patterns.

In this example, a simple data analysis technique can be applied to the data in the data warehouse above. A simple cross-tabulation of results by quarter and products will reveal some easily visible patterns.

Movies Sales by Quarters – Cross-tabulation				
Qtr/Product	Gone With the Wind	Matrix	Monty Python	Total Sales Amount
Q2	\$15	0	\$30	\$45
Q3	\$30	\$36	\$30	\$96
Q4	\$15	0	\$18	\$33
<b>Total Revenue</b>	\$60	\$36	<b>\$78</b>	\$174

Based on the cross-tabulation above, one can readily answer some product sales questions, like:

1. What is the best-selling movie by revenue? – *Monty Python*.
2. What is the best quarter by revenue this year? – *Q3*
3. Any other patterns? – *Matrix movie sells only in Q3 (seasonal item)*.

These simple insights can help plan marketing promotions and manage inventory of various movies.

If a cross tabulation was designed to include customer location data, one

could answer other questions, such as

1. What is the best-selling geography? – US
2. What is the worst selling geography? – UK
3. Any other patterns? – Monty Python sells globally, while Gone with the Wind sells only in the US.

If the data mining was done at the monthly level of data, it would be easy to miss the seasonality of the movies. However, one would have observed that September is the highest selling month.

The previous example shows that many differences and patterns can be noticed by analyzing data in different ways. However, some insights are more important than others. The value of the insight depends upon the problem being solved. The insight that there are more sales of a product in a certain quarter helps a manager plan what products to focus on. In this case, the store manager should stock up on the movie Matrix in the 3<sup>rd</sup> Quarter (Q3). Similarly, knowing which quarter has the highest overall sales allows for different resource decisions in that quarter. In this case, if Q3 is bringing more than half of total revenue, this requires greater attention on the e-commerce website in the third quarter.

### *Selecting Data Mining Projects*

Data mining should be done to solve high-priority, high-value problems. Much effort is required to gather data, clean and organize it, mine it with many techniques, interpret the results, and find the right insight. It is important that there be a large expected payoff from finding the insight. One should select the right data (and ignore the rest), organize it into a nice and imaginative framework that brings relevant data together, and then apply data mining techniques to deduce the right insight.

A retail company may use data mining techniques to determine which new product categories to add to which of their stores; how to increase sales of existing products; which new locations to open stores in; how to segment the customers for more effective communication; and so on.

Data can be analyzed at multiple levels of granularity and could lead to a large number of interesting combinations of data and interesting patterns. Some of the patterns may be more meaningful than the others. Such highly granular data is often used, especially in finance and high-tech areas, so that

one can gain even the slightest edge over the competition.

## *Data Mining Techniques*

Here are brief descriptions of some of the most important data mining techniques used to generate insights from data.

*Decision Trees:* They help classify populations into classes. It is said that 70% of all data mining work is about classification solutions; and that 70% of all classification work uses decision trees. Thus, decision trees are the most popular and important data mining technique. There are many popular algorithms to make decision trees. They differ in terms of their mechanisms and each technique works well for different situations. It is possible to try multiple decision-tree algorithms on a data set and compare the predictive accuracy of each tree.

*Regression:* This is a well-understood technique from the field of statistics. The goal is to find a best fitting curve through the many data points. The best fitting curve is that which minimizes the (error) distance between the actual data points and the values predicted by the curve. Regression models can be projected into the future for prediction and forecasting purposes.

*Artificial Neural Networks:* Originating in the field of artificial intelligence and machine learning, ANNs are multi-layer non-linear information processing models that learn from past data and predict future values. These models predict well, leading to their popularity. The model's parameters may not be very intuitive. Thus, neural networks are opaque like a black-box. These systems also require a large amount of past data to adequately train the system.

*Cluster analysis:* This is an important data mining technique for dividing and conquering large data sets. The data set is divided into a certain number of clusters, by discerning similarities and dissimilarities within the data. There is no one right answer for the number of clusters in the data. The user needs to make a decision by looking at how well the number of clusters chosen fit the data. This is most commonly used for market segmentation. Unlike decision trees and regression, there is no one right answer for cluster analysis.

*Association Rule Mining:* Also called Market Basket Analysis when used in retail industry, these techniques look for associations between data values. An analysis of items frequently found together in a market basket can help

cross-sell products, and also create product bundles.

## Data Visualization

As data and insights grow in number, a new requirement is the ability of the executives and decision makers to absorb this information in real time. There is a limit to human comprehension and visualization capacity. That is a good reason to prioritize and manage with fewer but key variables that relate directly to the Key Result Areas (KRAs) of a role.

Here are few considerations when making presentations using data:

1. Present the conclusions and not just report the data.
2. Choose wisely from a palette of graphs to suit the data.
3. Organize the results to make the central point stand out.
4. Ensure that the visuals accurately reflect the numbers.  
Inappropriate visuals can create misinterpretations and misunderstandings.
5. Make the presentation unique, imaginative and memorable.

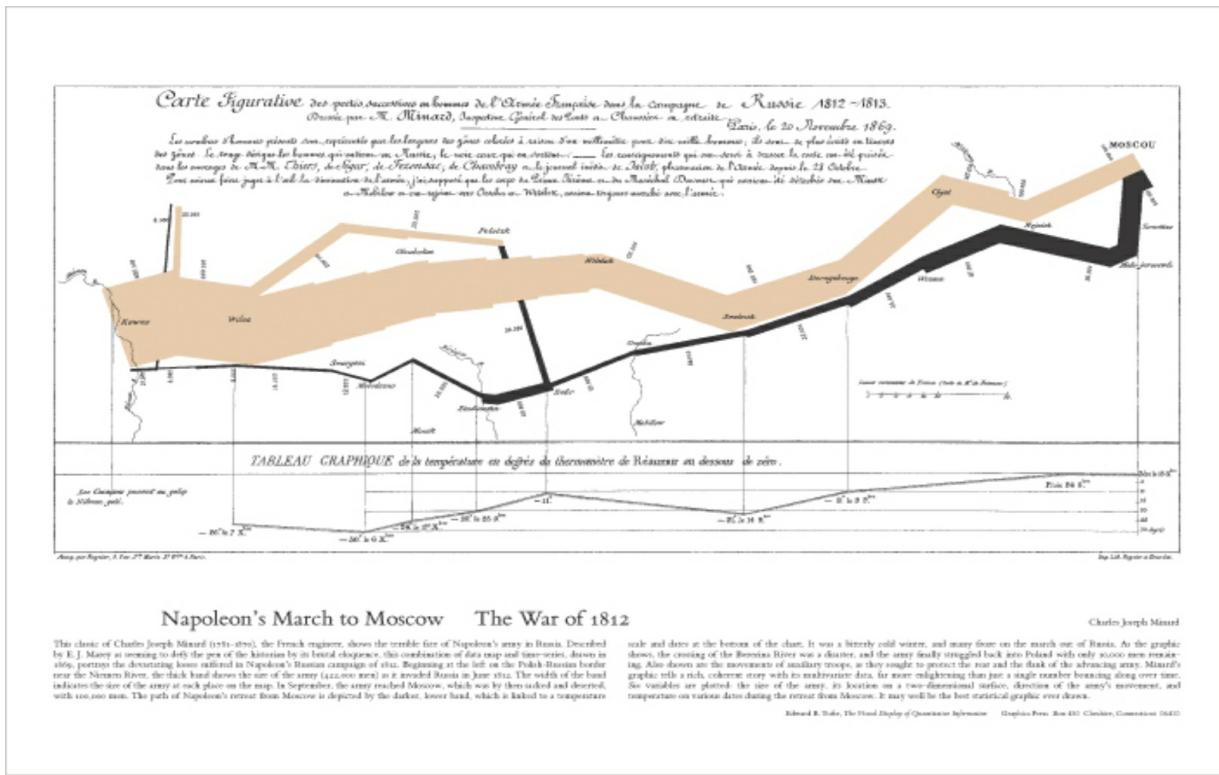
Executive dashboards are designed to provide information on select few variables for every executive. They use graphs, dials, and lists to show the status of important parameters. These dashboards also have a drill-down capability to enable a root-cause analysis of exception situations (Figure 1.3).



**Figure 1.3: Sample Executive Dashboard**

Data visualization has been an interesting problem across the disciplines. Many dimensions of data can be effectively displayed on a two-dimensional surface to give a rich and more insightful description of the totality of the story.

The classic presentation of the story of Napoleon's march to Russia in 1812, by French cartographer Joseph Minard, is shown in Figure 1.4. It covers about six dimensions. Time is on horizontal axis. The geographical coordinates and rivers are mapped in. The thickness of the bar shows the number of troops at any point of time that is mapped. One color is used for the onward march and another for the retreat. The weather temperature at each time is shown in the line graph at the bottom.



**Figure 1.4: Sample Data Visualization**

## Terminology and Careers

Many overlapping terms are used in the marketplace. Here is a brief look at the most popular terms used.

Business Intelligence is a business oriented term. So is Decision science. This is the field of using data-driven insights to make superior business decisions.

This is a broad category and in some sense includes data analytics or data mining as a component.

Data Analytics is a technology oriented term. So is Data Mining. Both involve the use of techniques and tools to find novel useful patterns from data. It involves organizing and modeling the data to test a particular hypothesis or answer a question. Knowledge of machine learning and statistics can be useful in this role.

Big Data is special type of data ... very large, fast and complex. Storing and managing all this data is a slightly more technical task and it uses more recent and innovative tools. This field is more attractive to the more technical people. Cloud computing is an attractive solution to store and process Big Data.

Data Science is a new discipline born in the early 2000s. The scope includes the entire data processing chain. A data scientist would ideally be familiar with all aspects of the discipline while specializing in a part of the area. This field was seeded by IBM and a few other companies who need many data scientists in the coming years, for themselves as well for their clients. They reached out to a select few top universities and encouraged them, with knowledge and resources, to create courses in this area. This helps increase the supply of data scientists, which has been called the ‘sexiest’ or hottest job of the decade.

## **Organization of the book**

This chapter is designed to provide the wholeness of business intelligence and data mining, to provide the reader with an intuition for this area of knowledge. The rest of the book can be considered in four sections.

Section 1 covers the data processing chain. Chapter 2 covers the field of business intelligence and its applications across industries and functions. Chapter 3 explains what is data warehousing and how does it help with data mining. Chapter 4 then describes data mining in some detail with an overview of its major tools and techniques. Chapter 5 shows insights from data can be visualized for better communication and consumption.

Section 2 focuses on core data mining techniques. Every technique is shown through solving an example in detail. Chapter 6 shows the power and ease of decision trees, which are the most popular data mining technique. Chapter 7

describes statistical regression modeling techniques. Chapter 8 provides an overview of artificial neural networks, a versatile machine learning technique that is so much in news lately. Chapter 9 describes cluster analysis and how it can help with market segmentation. Finally, chapter 10 describes association rule mining technique, also called market basket analysis, that helps find shopping patterns.

Section 3 covers advanced topics. Chapter 11 introduces the concepts and techniques of text mining, that helps discover insights from text data including social media data. Chapter 12 covers Naïve-Bayes, a probability-based classification technique used for text mining applications such as spam filtering. Chapter 13 covers support vector machines, a mathematically rigorous classification technique, also used for spam filtering and other applications with high dimensional data. Chapter 14 provides an overview of the growing field of web mining, which includes mining the structure, content and usage of web sites. Chapter 15 covers social network analysis, which helps analyze both text communication and web structure.

Section 4 provides primers of related topics. Chapter 16 will provide an overview of the field of Big Data. Chapter 17 is a primer on Data Modeling, for those who do not have any background in databases. Chapter 18 is a primer on statistics, for those who wish to refresh their knowledge of statistics. Chapter 19 is a primer on Artificial Intelligence, the hottest and most controversial topic of the current times. Chapter 20 offers a brief view of Data Science careers and the skills needed.

Appendix contains an extensive R tutorial to help with many of the techniques shown in the book. No prior experience with R is needed for using this tutorial.

## Review Questions

- 1: Describe the Business Intelligence and Data Mining (BIDM) cycle.
- 2: Describe the data processing chain.
- 3: What are the similarities between diamond mining and data mining?
- 4: What are the different data mining techniques? Which of these would be relevant in your current work?
- 5: What is a dashboard? How does it help?

6: Create a visual to show the weather pattern in your city. Could you show together temperature, humidity, wind, and rain/snow over a period of time.

## **Section 1**

This section covers four important high-level topics.

Chapter 2 will cover business intelligence concepts, and its applications in many industries.

Chapter 3 will describe data warehousing systems, and ways of creating and managing them.

Chapter 4 will describe data mining as a whole, its many techniques, and many do's and don'ts of effective data mining.

Chapter 5 will describe data visualization as a whole, with techniques and examples, and many thumb rules of effective data visualizations.

## Chapter 2: Business Intelligence Concepts and Applications

### Introduction

Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users. (Figure 2.1).

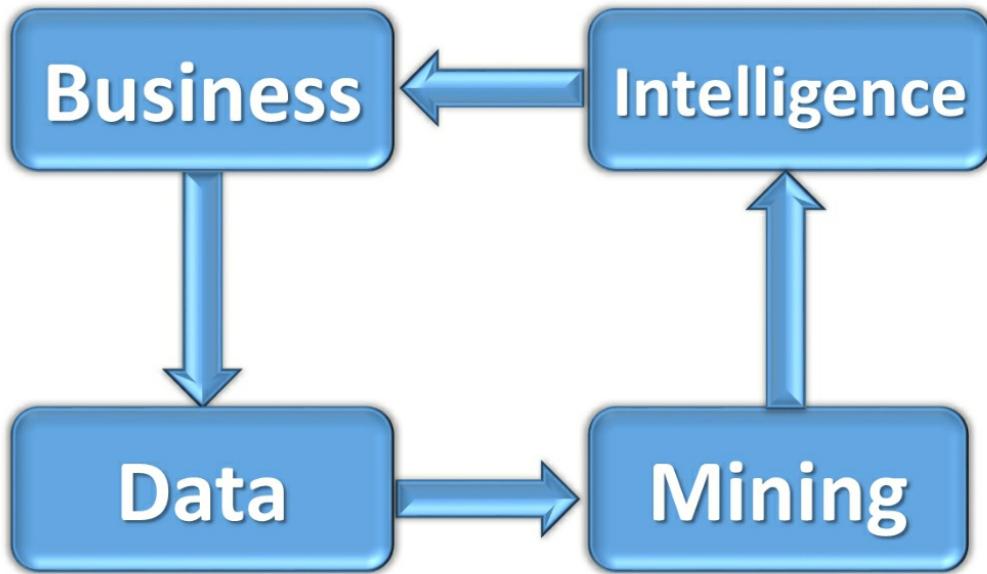


Figure 2.1: BIDM cycle

The nature of life and businesses is to grow. Information is the life-blood of business. Businesses use many techniques for understanding their environment and predicting the future for their own benefit and growth. Decisions are made from facts and feelings. Data-based decisions are more effective than those based on feelings alone. Actions based on accurate data, information, knowledge, experimentation, and testing, using fresh insights, can more likely succeed and lead to sustained growth. One's own data can be the most effective teacher. Therefore, organizations should gather data, sift through it, analyze and mine it, find insights, and then embed those insights into their operating procedures.

There is a new sense of importance and urgency around data as it is being viewed as a new natural resource. It can be mined for value, insights, and competitive advantage. In a hyperconnected world, where everything is potentially connected to everything else, with potentially infinite correlations, data represents the impulses of nature in the form of certain events and

attributes. A skilled business person is motivated to use this cache of data to harness nature, and to find new niches of unserved opportunities that could become profitable ventures.

### **Caselet: Khan Academy – BI in Education**

*Khan Academy is an innovative non-profit educational organization that is turning the K-12 education system upside down. It provides short YouTube based video lessons on thousands of topics for free. It shot into prominence when Bill Gates promoted it as a resource that he used to teach his own children. With this kind of a resource classrooms are being flipped ... i.e. student do their basic lecture-type learning at home using those videos, while the class time is used for more one-on-one problem solving and coaching. Students can access the lessons at any time to learn at their own pace. The students' progress is recorded including what videos they watched how many times, which problems they stumbled on, and what scores they got on online tests.*

*Khan Academy has developed tools to help teachers get a pulse on what's happening in the classroom. Teachers are provided a set of real-time dashboards to give them information from the macro level ("How is my class doing on geometry?") to the micro level ("How is Jane doing on mastering polygons?") Armed with this information, teachers can place selective focus on the students that need certain help. (Source: KhanAcademy.org)*



*Q1: How does a dashboard improve the teaching experience? And the student's learning experience?*

*Q2: Design a dashboard for tracking your own career.*

## BI for better decisions

The future is inherently uncertain. Risk is the result of a probabilistic world where there are no certainties and complexities abound. People use crystal balls, astrology, palmistry, ground hogs, and also mathematics and numbers to mitigate risk in decision-making. The goal is to make effective decisions, while reducing risk. Businesses calculate risks and make decisions based on a broad set of facts and insights. Reliable knowledge about the future can help managers make the right decisions with lower levels of risk.

The speed of action has risen exponentially with the growth of the Internet. In a hypercompetitive world, the speed of a decision and the consequent action can be a key advantage. The Internet and mobile technologies allow decisions to be made anytime, anywhere. Ignoring fast-moving changes can threaten the organization's future. Research has shown that an unfavorable comment about the company and its products on social media should not go unaddressed for long. Banks have had to pay huge penalties to Consumer Financial Protection Bureau (CFPB) in United States in 2013 for complaints made on CFPB's websites. On the other hand, a positive sentiment expressed on social media should also be utilized as a potential sales and promotion

opportunity, while the opportunity lasts.

## Decision types

There are two main kinds of decisions: strategic decisions and operational decisions. BI can help make both better. Strategic decisions are those that impact the direction of the company. The decision to reach out to a new customer set would be a strategic decision. Operational decisions are more routine and tactical decisions, focused on developing greater efficiency. Updating an old website with new features will be an operational decision.

In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal. The consequences of the decision would be apparent some time later. Thus, one is constantly scanning for new possibilities and new paths to achieve the goals. BI can help with what-if analysis of many possible scenarios. BI can also help create new ideas based on new patterns found from data mining.

Operational decisions can be made more efficient using an analysis of past data. A classification system can be created and modeled using the data of past instances to develop a good model of the domain. This model can help improve operational decisions in the future. BI can help automate operations level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way. For example, a bank might want to make decisions about making financial loans in a more scientific way using data-based models. A decision-tree-based model could provide a consistently accurate loan decisions. Developing such decision tree models is one of the main applications of data mining techniques.

Effective BI has an evolutionary component, as business models evolve. When people and organizations act, new facts (data) are generated. Current business models can be tested against the new data, and it is possible that those models will not hold up well. In that case, decision models should be revised and new insights should be incorporated. An unending process of generating fresh new insights in real time can help make better decisions, and thus can be a significant competitive advantage.

## BI Tools

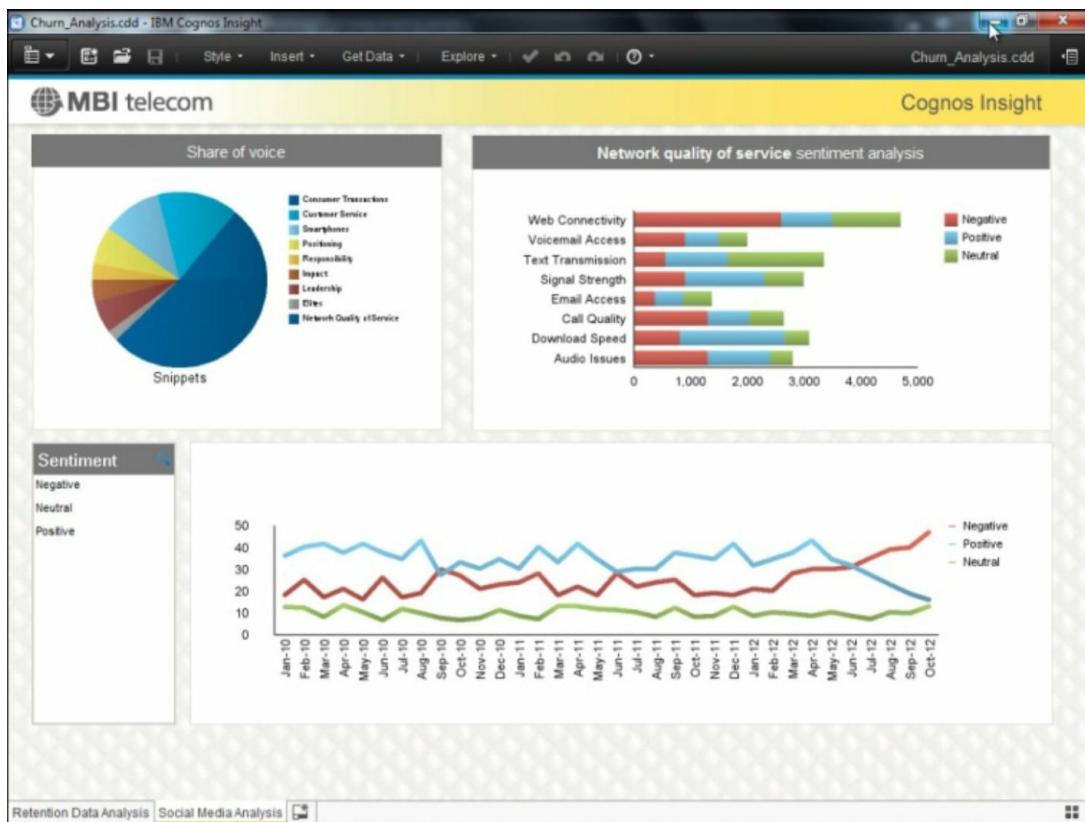
BI includes a variety of software tools and techniques to provide the managers with the information and insights needed to run the business. Information can be provided about the current state of affairs with the

capability to drill down into details, and also insights about emerging patterns which lead to projections into the future. BI tools include data warehousing, online analytical processing, social media analytics, reporting, dashboards, querying, and data mining.

BI tools can range from very simple tools that could be considered end-user tools, to very sophisticated tools that offer a very broad and complex set of functionality. Thus, Even executives can be their own BI experts, or they can rely on BI specialists to set up the BI mechanisms for them. Thus, large organizations invest in expensive sophisticated BI solutions that provide good information in real time.

A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated statistical analysis.

A dashboarding system, such as IBM Cognos or Tableau, can offer a sophisticated set of tools for gathering, analyzing, and presenting data. At the user end, modular dashboards can be designed and redesigned easily with a graphical user interface. The back-end data analytical capabilities include many statistical functions. The dashboards are linked to data warehouses at the back end to ensure that the tables and graphs and other elements of the dashboard are updated in real time (Figure 2.2).



**Figure 2.2: Sample Executive Dashboard**

Data mining systems, such as IBM SPSS Modeler, are industrial strength systems that provide capabilities to apply a wide range of analytical models on large data sets. Open source systems, such as Weka, are popular platforms designed to help mine large amounts of data to discover patterns.

## BI Skills

As data grows and exceeds our capacity to make sense of it, the tools need to evolve, and so should the imagination of the BI specialist. “Data Scientist” has been called as the hottest job of this decade.

A skilled and experienced BI specialist should be open enough to go outside the box, open the aperture and see a wider perspective that includes more dimensions and variables, in order to find important patterns and insights. The problem needs to be looked at from a wider perspective to consider many more angles that may not be immediately obvious. An imaginative solution should be proposed for the problem so that interesting and useful results can emerge.

A good data mining project begins with an interesting problem to solve.

Selecting the right data mining problem is an important skill. The problem should be valuable enough that solving it would be worth the time and expense. It takes a lot of time and energy to gather, organize, cleanse, and prepare the data for mining and other analysis. The data miner needs to persist with the exploration of patterns in the data. The skill level has to be deep enough to engage with the data and make it yield new useful insights.

## BI Applications

BI tools are required in almost all industries and functions. The nature of the information and the speed of action may be different across businesses, but every manager today needs access to BI tools to have up-to-date metrics about business performance. Businesses need to embed new insights into their operating processes to ensure that their activities continue to evolve with more efficient practices. The following are some areas of applications of BI and data mining.

### Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer, sell more of its offerings to the existing customers, and also, expand the pool of customers it serves. BI applications can impact many aspects of marketing.

1. *Maximize the return on marketing campaigns:* Understanding the customer's pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.
2. *Improve customer retention (churn analysis):* It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit, can help the business design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.
3. *Maximize customer value (cross-, up-selling):* Every contact with the customer should be seen as an opportunity to gauge their current needs. Offering a customer new products and solutions

based on those imputed needs can help increase revenue per customer. Even a customer complaint can be seen as an opportunity to wow the customer. Using the knowledge of the customer's history and value, the business can choose to sell a premium service to the customer.

4. *Identify and delight highly-valued customers.* By segmenting the customers, the best customers can be identified. They can be proactively contacted, and delighted, with greater attention and better service. Loyalty programs can be managed more effectively.
5. *Manage brand image.* A business can create a listening post to listen to social media chatter about itself. It can then do sentiment analysis of the text to understand the nature of comments, and respond appropriately to the prospects and customers.

## **Healthcare and Wellness**

Health care is one of the biggest sectors in advanced economies. Evidence-based medicine is the newest trend in data-based health care management. BI applications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.

1. *Diagnose disease in patients:* Diagnosing the cause of a medical condition is the critical first step in a medical engagement. Accurately diagnosing cases of cancer or diabetes can be a matter of life and death for the patient. In addition to the patient's own current situation, many other factors can be considered, including the patient's health history, medication history, family's history, and other environmental factors. This makes diagnosis as much of an art form as it is science. Systems, such as IBM Watson, absorb all the medical research to date and make probabilistic diagnoses in the form of a decision tree, along with a full explanation for their recommendations. These systems take away most of the guess work done by doctors in diagnosing ailments.
2. *Treatment effectiveness:* The prescription of medication and

treatment is also a difficult choice out of so many possibilities. For example, there are more than 100 medications for hypertension (high blood pressure) alone. There are also interactions in terms of which drugs work well with others and which drugs do not. Decision trees can help doctors learn about and prescribe more effective treatments. Thus, the patients could recover their health faster with a lower risk of complications and cost.

3. *Wellness management*: This includes keeping track of patient health records, analyzing customer health trends and proactively advising them to take any needed precautions.
4. *Manage fraud and abuse*: Some medical practitioners have unfortunately been found to conduct unnecessary tests, and/or overbill the government and health insurance companies. Exception reporting systems can identify such providers and action can be taken against them.
5. *Public health management*: The management of public health is one of the important responsibilities of any government. By using effective forecasting tools and techniques, governments can better predict the onset of disease in certain areas in real time. They can thus be better prepared to fight the diseases. Google has been known to predict the movement of certain diseases by tracking the search terms (like flu, vaccine) used in different parts of the world.

## **Education**

As higher education becomes more expensive and competitive, it becomes a great user of data-based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

1. *Student Enrollment (Recruitment and Retention)*: Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend. Schools can develop models of what kinds of students are attracted to the school, and

then reach out to those students. The students at risk of not returning can be flagged, and corrective measures can be taken in time.

2. *Course offerings:* Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students. This can help increase class size, reduce costs, and improve student satisfaction.
3. *Fund-raising from Alumni and other donors:* Schools can develop predictive models of which alumni are most likely to pledge financial support to the school. Schools can create a profile for alumni more likely to pledge donations to the school. This could lead to a reduction in the cost of mailings and other forms of outreach to alumni.

## Retail

Retail organizations grow by meeting customer needs with quality products, in a convenient, timely, and cost-effective manner. Understanding emerging customer shopping patterns can help retailers organize their products, inventory, store layout, and web presence in order to delight their customers, which in turn would help increase revenue and profits. Retailers generate a lot of transaction and logistics data that can be used to diagnose and solve problems.

1. *Optimize inventory levels at different locations:* Retailers need to manage their inventories carefully. Carrying too much inventory imposes carrying costs, while carrying too little inventory can cause stock-outs and lost sales opportunities. Predicting sales trends dynamically can help retailers move inventory to where it is most in demand. Retail organizations can provide their suppliers with real time information about sales of their items, so the suppliers can deliver their product to the right locations and minimize stock-outs.
2. *Improve store layout and sales promotions:* A market basket analysis can develop predictive models of which products sell together often. This knowledge of affinities between products can

help retailers co-locate those products. Alternatively, those affinity products could be located farther apart to make the customer walk the length and breadth of the store, and thus be exposed to other products. Promotional discounted product bundles can be created to push a nonselling item along with a set of products that sell well together.

3. *Optimize logistics for seasonal effects:* Seasonal products offer tremendously profitable short-term sales opportunities, yet they also offer the risk of unsold inventories at the end of the season. Understanding which products are in season in which market can help retailers dynamically manage prices to ensure their inventory is sold during the season. If it is raining in a certain area, then the inventory of umbrella and ponchos could be rapidly moved there from nonrainy areas to help increase sales.
4. *Minimize losses due to limited shelf life:* Perishable goods offer challenges in terms of disposing off the inventory in time. By tracking sales trends, the perishable products at risk of not selling before the sell-by date, can be suitably discounted and promoted.

## Banking

Banks make loans and offer credit cards to millions of customers. They are most interested in improving the quality of loans and reducing bad debts. They also want to retain more good customers, and sell more services to them.

1. *Automate the loan application process:* Decision models can be generated from past data that predict the likelihood of a loan proving successful. These can be inserted in business processes to automate the financial loan approval process.
2. *Detect fraudulent transactions:* Billions of financial transactions happen around the world every day. Exception-seeking models can identify patterns of fraudulent transactions. For example, if money is being transferred to an unrelated account for the first time, it could be a fraudulent transaction.

3. *Maximize customer value (cross-, up-selling).* Selling more products and services to existing customers is often the easiest way to increase revenue. A checking account customer in good standing could be offered home, auto, or educational loans on more favorable terms than other customers, and thus, the value generated from that customer could be increased.
4. *Optimize cash reserves with forecasting.* Banks have to maintain certain liquidity to meet the needs of depositors who may like to withdraw money. Using past data and trend analysis, banks can forecast how much to keep and invest the rest to earn interest.

## Financial Services

Stock brokerages are an intensive user of BI systems. Fortunes can be made or lost based on access to accurate and timely information.

1. *Predict changes in bond and stock prices:* Forecasting the price of stocks and bonds is a favorite pastime of financial experts as well as lay people. Stock transaction data from the past, along with other variables, can be used to predict future price patterns. This can help traders develop long-term trading strategies.
2. *Assess the effect of events on market movements.* Decision models using decision trees can be created to assess the impact of events on changes in market volume and prices. Monetary policy changes (such as Federal Reserve interest rate change) or geopolitical changes (such as war in a part of the world) can be factored into the predictive model to help take action with greater confidence and less risk.
3. *Identify and prevent fraudulent activities in trading:* There have unfortunately been many cases of insider trading, leading to many prominent financial industry stalwarts going to jail. Fraud detection models seek out-of-the-ordinary activities, and help identify and flag fraudulent activity patterns.

## Insurance

This industry is a prolific user of prediction models in pricing insurance

proposals and managing losses from claims against insured assets.

1. *Forecast claim costs for better business planning:* When natural disasters, such as hurricanes and earthquakes strike, loss of life and property occurs. By using the best available data to model the likelihood (or risk) of such events happening, the insurer can plan for losses and manage resources and profits effectively.
2. *Determine optimal rate plans:* Pricing an insurance rate plan requires covering the potential losses and making a profit. Insurers use actuary tables to project life spans and disease tables to project mortality rates, and thus price themselves competitively yet profitably.
3. *Optimize marketing to specific customers:* By micro-segmenting potential customers, a data-savvy insurer can cherry pick the best customers and leave the less profitable customers to its competitors. Progressive Insurance is a US-based company that is known to actively use data mining to cherry pick customers and increase its profitability.
4. *Identify and prevent fraudulent claim activities.* Patterns can be identified as to where and what kinds of fraud are more likely to occur. Decision-tree-based models can be used to identify and flag fraudulent claims.

## Manufacturing

Manufacturing operations are complex systems with inter-related subsystems. From machines working right, to workers having the right skills, to the right components arriving with the right quality at the right time, to money to source the components, many things have to go right. Toyota's famous lean manufacturing company works on just-in-time inventory systems to optimize investments in inventory and to improve flexibility in their product-mix.

1. *Discover novel patterns to improve product quality:* Quality of a product can also be tracked, and this data can be used to create a predictive model of product quality deteriorating. Many

companies, such as automobile companies, have to recall their products if they have found defects that have a public safety implication. Data mining can help with root cause analysis that can be used to identify sources of errors and help improve product quality in the future.

2. *Predict/prevent machinery failures:* Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failures could be constructed using past data. Preventive maintenance can be planned, and manufacturing capacity can be adjusted, to account for such maintenance activities.

## Telecom

BI in telecom can help with the customer side as well as network side of the operations. Key BI applications include churn management, marketing/customer profiling, network failure, and fraud detection.

1. *Churn management:* Telecom customers have shown a tendency to switch their providers in search for better deals. Telecom companies tend to respond with many incentives and discounts to hold on to customers. However, they need to determine which customers are at a real risk of switching and which others are just negotiating for a better deal. The level of risk should to be factored into the kind of deals and discounts that should be given. Millions of such customer calls happen every month. The telecom companies need to provide a consistent and data-based way to predict the risk of the customer switching, and then make an operational decision in real time while the customer call is taking place. A decision-tree- or a neural network-based system can be used to guide the customer-service call operator to make the right decisions for the company, in a consistent manner.
2. *Marketing and product creation.* In addition to customer data, telecom companies also store call detail records (CDRs), which can be analyzed to precisely describe the calling behavior of each customer. This unique data can be used to profile customers and

then can be used for creating new products/services bundles for marketing purposes. An American telecom company, MCI, created a program called Friends & Family that allowed free calls with one's friends and family on that network, and thus, effectively locked many people into their network.

3. *Network failure management*: Failure of telecom networks for technical failures or malicious attacks can have devastating impacts on people, businesses, and society. In telecom infrastructure, some equipment will likely fail with certain mean time between failures. Modeling the failure pattern of various components of the network can help with preventive maintenance and capacity planning.
4. *Fraud Management*: There are many kinds of fraud in consumer transactions. Subscription fraud occurs when a customer opens an account with the intention of never paying for the services. Superimposition fraud involves illegitimate activity by a person other than the legitimate account holder. Decision rules can be developed to analyze each CDR in real time to identify chances of fraud and take effective action.

## Public Sector

Government gathers a large amount of data by virtue of their regulatory function. That data could be analyzed for developing models of effective functioning. There are innumerable applications that can benefit from mining that data. A couple of sample applications are shown here.

1. *Law enforcement*: Social behavior is a lot more patterned and predictable than one would imagine. For example, Los Angeles Police Department (LAPD) mined the data from its 13 million crime records over 80 years and developed models of what kind of crime going to happen when and where. By increasing patrolling in those particular areas, LAPD was able to reduce property crime by 27 percent. Internet chatter can be analyzed to learn of and prevent any evil designs.

2. *Scientific research:* Any large collection of research data is amenable to being mined for patterns and insights. Protein folding (microbiology), nuclear reaction analysis (sub-atomic physics), disease control (public health) are some examples where data mining can yield powerful new insights.

## Conclusion

Business Intelligence is a comprehensive set of IT tools to support decision making with imaginative solutions for a variety of problems. BI can help improve the performance in nearly all industries and applications.

## Review Questions

1. Why should organizations invest in business intelligence solutions? Are BI tools more important than IT security solutions?
2. List 3 business intelligence applications in the hospitality industry.
3. Describe 2 BI tools used in your organization.
4. Businesses need a ‘two-second advantage’ to succeed. What does that mean to you?

### Liberty Stores Case Exercise: Step 1

*Liberty Stores Inc is a specialized global retail chain that sells organic food, organic clothing, wellness products, and education products to enlightened LOHAS (Lifestyles of the Healthy and Sustainable) citizens worldwide. The company is 20 years old, and is growing rapidly. It now operates in 5 continents, 50 countries, 150 cities, and has 500 stores. It sells 20000 products and has 10000 employees. The company has revenues of over \$5 billion and has a profit of about 5% of revenue. The company pays special attention to the conditions under which the products are grown and produced. It donates about one-fifth (20%) of its pre-tax profits from global local charitable causes.*

- 1: *Create a comprehensive dashboard for the CEO of the company.*
- 2: *Create another dashboard for a country head.*

## Chapter 3: Data Warehousing

### Introduction

A data warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to support decision support functions. DW is organized at the right level of granularity to provide clean enterprise-wide data in a standardized format for reports, queries, and analysis. DW is physically and functionally separate from an operational and transactional database. Creating a DW for analysis and queries represents significant investment in time and effort. It has to be constantly kept up-to-date for it to be useful. DW offers many business and technical benefits.

DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service. DW can present a competitive advantage by facilitating decision making and helping reform business processes.

DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself. DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis. It enhances overall IT performance by not burdening the operational databases used by Enterprise Resource Planning (ERP) and other systems.

#### **Caselet: University Health System – BI in Healthcare**

*Indiana University Health (IUH), a large academic health care system, decided to build an enterprise data warehouse (EDW) to foster a genuinely data-driven management culture. IUH hired a data warehousing vendor to develop an EDW which also integrates with their Electronic Health Records (EHR) system. They loaded 14 billion rows of data into the EDW—fully 10 years of clinical data from across IUH's network. Clinical events, patient encounters, lab and radiology, and other patient data were included, as were IUH's performance management, revenue cycle, and patient satisfaction data. They soon put in a new interactive dashboard using the EDW that provided IUH's leadership with the daily operational insights they need to solve the quality/cost equation. It offers visibility into key operational*

*metrics and trends to easily track the performance measures critical to controlling costs and maintaining quality. The EDW can easily be used across IUH's departments to analyze, track and measure clinical, financial, and patient experience outcomes. (Source: healthcatalyst.com)*

*Q1: What are the benefits of a single large comprehensive EDW?*

*Q2: What kinds of data would be needed for an EDW for an airline company?*

## Design Considerations for DW

The objective of DW is to provide business knowledge to support decision making. For DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up-to-date. Here are some requirements for a good DW:

1. *Subject oriented*: To be effective, a DW should be designed around a subject domain, i.e. to help solve a certain category of problems.
2. *Integrated*: The DW should include data from many functions that can shed light on a particular subject area. Thus the organization can benefit from a comprehensive view of the subject area.
3. *Time-variant (time series)*: The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.
4. *Nonvolatile*: DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.
5. *Summarized*: DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reduces the number of variables or dimensions of the data to make them more meaningful for the decision makers.
6. *Not normalized*: DW often uses a star schema, which is a

rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries.

7. *Metadata*: Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.
8. *Near Real-time and/or right-time (active)*: DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

## DW Development Approaches

There are two fundamentally different approaches to developing DW: top down and bottom up. The top-down approach is to make a comprehensive DW that covers all the reporting needs of the enterprise. The bottom-up approach is to produce small data marts, for the reporting needs of different departments or functions, as needed. The smaller data marts will eventually align to deliver comprehensive EDW capabilities. The top-down approach provides consistency but takes more time and resources. The bottom-up approach leads to healthy local ownership and maintainability of data (Table 3.1).

	<b>Functional Data Mart</b>	<b>Enterprise Data Warehouse</b>
Scope	One subject or functional area	Complete enterprise data needs
Value	Functional area reporting and insights	Deeper insights connecting multiple functional areas
Target organization	Decentralized management	Centralized management
Time	Low to medium	High
Cost	Low	High
Size	Small to medium	Medium to large
Approach	Bottom up	Top down
Complexity	Low (fewer data transformations)	High (data standardization)

Technology	Smaller scale servers and databases	Industrial strength
------------	-------------------------------------	---------------------

Table 3.1: Comparing Data Mart and Data Warehouse

## DW Architecture

DW has four key elements (Figure 3.1). The first element is the data sources that provide the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the methods of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

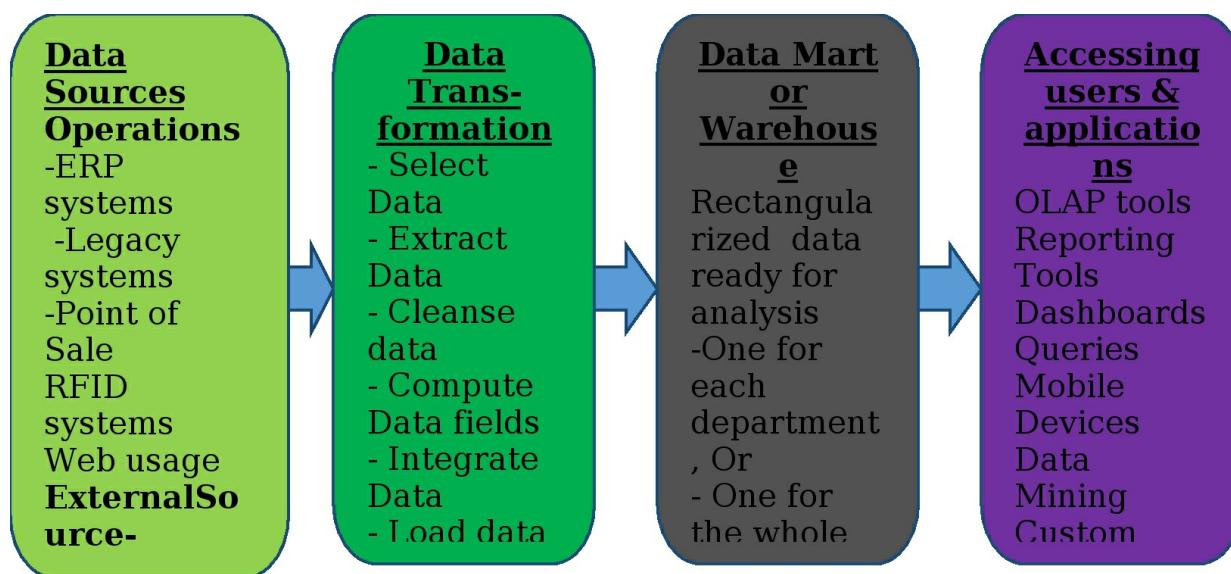


Figure 3.1: Data Warehousing Architecture

## Data Sources

Data Warehouses are created from structured data sources. Unstructured data such as text data would need to be structured before inserted into the DW.

1. *Operations data:* This includes data from all business applications, including from ERPs systems that form the backbone of an organization's IT systems. The data to be extracted will depend upon the subject matter of the data warehouse. For example, for a sales/marketing data mart, only the data about customers, orders, customer service, and so on would be extracted.

2. *Specialized applications*: This includes applications such as Point of Sale (POS) terminals, and e-commerce applications, that also provide customer-facing data. Supplier data could come from Supply Chain Management systems. Planning and budget data should also be added as needed for making comparisons against targets.
3. *External syndicated data*: This includes publicly available data such as weather or economic activity data. It could also be added to the DW, as needed, to provide good contextual information to decision makers.

## **Data Loading Processes**

The heart of a useful DW is the processes to populate the DW with good quality data. This is called the Extract-Transform-Load (ETL) cycle.

1. Data should be extracted from the operational (transactional) database sources, as well as from other applications, on a regular basis.
2. The extracted data should be aligned together by key fields and integrated into a single data set. It should be cleansed of any irregularities or missing values. It should be rolled-up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.
3. This transformed data should then be uploaded into the DW.

This ETL process should be run at a regular frequency. Daily transaction data can be extracted from ERPs, transformed, and uploaded to the database the same night. Thus, the DW is up to date every morning. If a DW is needed for near-real-time information access, then the ETL processes would need to be executed more frequently. ETL work is usually done using automated using programming scripts that are written, tested, and then deployed for periodically updating the DW.

## **Data Warehouse Design**

Star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There are lookup tables that provide detailed values for codes used in the central table.

For example, the central table may use digits to represent a sales person. The lookup table will help provide the name for that sales person code. Here is an example of a star schema for a data mart for monitoring sales performance (Figure 3.2).

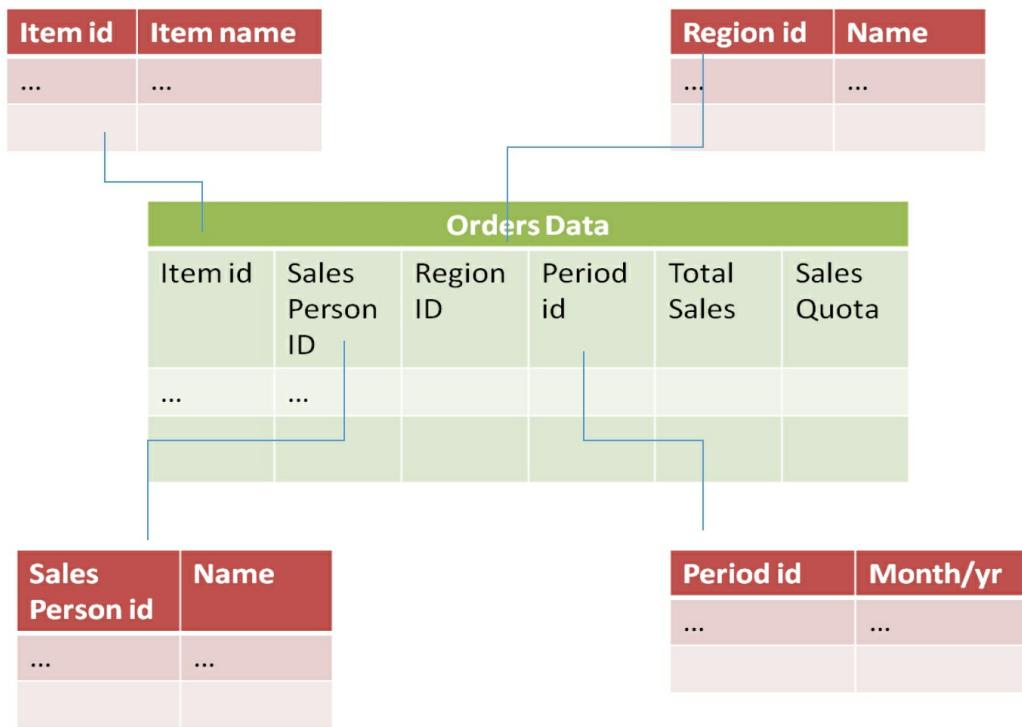


Figure 3.2: Star Schema Architecture for DW

Other schemas include the snowflake architecture. The difference between a star and snowflake is that in the latter, the look-up tables can have their own further look up tables.

There are many technology choices for developing DW. This includes selecting the right database management system and the right set of data management tools. There are a few big and reliable providers of DW systems. The provider of the operational DBMS may be chosen for DW also. Alternatively, a best-of-breed DW vendor could be used. There are also a variety of tools out there for data migration, data upload, data retrieval, and data analysis.

## DW Access

Data from the DW could be accessed for many purposes, by many users, through many devices, and many kinds of systems. A primary use of DW is

to produce routine management and monitoring reports.

1. OLAP (online analytical processing) systems are query and reporting tools that help to slice the multi-dimensional data in the DW, and produce desired reports. For example, a sales performance report would show sales by many dimensions, and compared with plan. Or one could request a list of all products sold in a region in a certain price range during a certain time period, and have them displayed in a report format.
2. A dashboarding system will use data from the warehouse and present analysis to users. The data from DW can be used to populate customized performance dashboards for executives. The dashboard could include drill-down capabilities to analyze the performance data for root cause analysis.
3. The data from the DW could be used for ad-hoc queries and any other applications that make use of the internal data.
4. Data from DW is used to provide data for mining purposes. Parts of the data would be extracted, and then combined with other relevant data, for data mining.

## DW Best Practices

A data warehousing project reflects a significant investment into information technology (IT). All of the best practices in implementing any IT project should be followed.

1. The DW project should *align with the corporate strategy*. Top management should be consulted for setting objectives. Financial viability (ROI) should be established. The project must be managed by both IT and business professionals. The DW design should be carefully tested before beginning development work. It is often much more expensive to redesign after development work has begun.
2. It is important to *manage user expectations*. The data warehouse should be built incrementally. Users should be trained in using the system so they can absorb the many features of the system.

3. *Quality and adaptability* should be built in from the start. Only relevant, cleansed, and high-quality data should be loaded. The system should be able to adapt to new tools for access. As business needs change, new data marts may need to be created for new needs.

## Conclusion

Data Warehouses are special data management facilities intended for creating reports and analysis to support managerial decision making. They are designed to make reporting and querying simple and efficient. The sources of data are operational systems, and external data sources. The DW needs to be updated with new data regularly to keep it useful. Data from DW provides a useful input for data mining activities.

## Review Questions

- 1: What is the purpose of a data warehouse?
- 2: What are the key elements of a data warehouse? Describe each one.
- 3: What are the sources and types of data for a data warehouse?
- 4: How will data warehousing evolve in the age of social media?

## Liberty Stores Case Exercise: Step 2

*The Liberty Stores company wants to be fully informed about its sales of products and take advantage of growth opportunities as they arise. It wants to analyze sales of all its products by all store locations. The newly hired Chief Knowledge Officer has decided to build a Data Warehouse.*

1. *Design a DW structure for the company to monitor its sales performance. (Hint: Design the central table and look-up tables).*
2. *Design another DW for the company's sustainability and charitable activities.*

## Chapter 4: Data Mining

### Introduction

Data mining is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data. Patterns must be valid, novel, potentially useful, and understandable. The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.

Data mining is a multidisciplinary field that borrows techniques from a variety of fields. It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

The field of data mining emerged in the context of pattern recognition in defense, such as identifying a friend-or-foe on a battlefield. Like many other defense-inspired technologies, it has evolved to help gain a competitive advantage in business.

For example, “customers who buy cheese and milk also buy bread 90 percent of the time” would be a useful pattern for a grocery store, which can then stock the products appropriately. Similarly, “people with blood pressure greater than 160 and an age greater than 65 were at a high risk of dying from a heart stroke” is of great diagnostic value for doctors, who can then focus on treating such patients with urgent care and great sensitivity.

Past data can be of predictive value in many complex situations, especially where the pattern may not be so easily visible without the modeling technique. Here is a dramatic case of a data-driven decision-making system that beats the best of human experts. Using past data, a decision tree model was developed to predict votes for Justice Sandra Day O’Connor, who had a swing vote in a 5–4 divided US Supreme Court. All her previous decisions were coded on a few variables. What emerged from data mining was a simple four-step decision tree that was able to accurately predict her votes 71 percent of the time. In contrast, the legal analysts could at best predict correctly 59 percent of the time. (Source: Martin et al. 2004)

### **Caselet: Target Corp – Data Mining in Retail**

*Target is a large retail chain that crunches data to develop insights that help target marketing and advertising campaigns. Target analysts managed to develop a pregnancy prediction score based on a customer's purchasing history of 25 products. In a widely publicized story, they figured out that a teenage girl was pregnant before her father did. The targeting can be quite successful and dramatic as this example published in the New York Times illustrates.*

*About a year after Target created their pregnancy-prediction model, a man walked into a Target store and demanded to see the manager. He was clutching coupons that had been sent to his daughter and he was angry, according to an employee who participated in the conversation. "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"*

*The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.*

*On the phone, though, the father was somewhat subdued. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. I owe you an apology." (Source: New York Times).*

*1: Do Target and other retailers have full rights to use their acquired data as it sees fit, and to contact desired consumers with all legally admissible means and messages? What are the issues involved here?*

*2: FaceBook and Google provide many services for free. In return they mine our email and blogs and send us targeted ads. Is that a fair deal?*

## **Gathering and selecting data**

The total amount of data in the world is doubling every 18 months. There is an ever-growing avalanche of data coming with higher velocity, volume, and variety. One has to quickly use it or lose it. Smart data mining requires choosing where to play. One has to make judicious decisions about what to

gather and what to ignore, based on the purpose of the data mining exercises. It is like deciding where to fish; as not all streams of data will be equally rich in potential insights.

To learn from data, quality data needs to be effectively gathered, cleaned and organized, and then efficiently mined. One requires the skills and technologies for consolidation and integration of data elements from many sources. Most organizations develop an enterprise data model (EDM) to organize their data. An EDM is a unified, high-level model of all the data stored in an organization's databases. The EDM is usually inclusive of the data generated from all internal systems. The EDM provides the basic menu of data to create a data warehouse for a particular decision-making purpose. DWs help organize all this data in an easy and usable manner so that it can be selected and deployed for mining. The EDM can also help imagine what relevant external data should be gathered to provide context and develop good predictive relationships with the internal data. In the United States, the various federal and local governments and their regulatory agencies make a vast variety and quantity of data available at data.gov.

Gathering and curating data takes time and effort, particularly when it is unstructured or semistructured. Unstructured data can come in many forms like databases, blogs, images, videos, audio, and chats. There are streams of unstructured social media data from blogs, chats, and tweets. There are streams of machine-generated data from connected machines, RFID tags, the internet of things, and so on. Eventually the data should be *rectangularized*, that is, put in rectangular data shapes with clear columns and rows, before submitting it to data mining.

Knowledge of the business domain helps select the right streams of data for pursuing new insights. Only the data that suits the nature of the problem being solved should be gathered. The data elements should be relevant, and suitably address the problem being solved. They could directly impact the problem, or they could be a suitable proxy for the effect being measured. Select data could also be gathered from the data warehouse. Every industry and function will have its own requirements and constraints. The health care industry will provide a different type of data with different data names. The HR function would provide different kinds of data. There would be different issues of quality and privacy for these data.

## Data cleansing and preparation

The quality of data is critical to the success and value of the data mining project. Otherwise, the situation will be of the kind of garbage in and garbage out (GIGO). The quality of incoming data varies by the source and nature of data. Data from internal operations is likely to be of higher quality, as it will be accurate and consistent. Data from social media and other public sources is less under the control of business, and is less likely to be reliable.

Data almost certainly needs to be cleansed and transformed before it can be used for data mining. There are many ways in what data may need to be cleansed – filling missing values, reigning in the effects of outliers, transforming fields, binning continuous variables, and much more – before it can be ready for analysis. Data cleansing and preparation is a labor-intensive or semi-automated activity that can take up to 60-80% of the time needed for a data mining project.

1. *Duplicate data needs to be removed.* The same data may be received from multiple sources. When merging the data sets, data must be de-duped.
2. *Missing values need to be filled in,* or those rows should be removed from analysis. Missing values can be filled in with average or modal or default values.
3. *Data elements should be comparable.* They may need to be (a) transformed from one unit to another. For example, total costs of health care and the total number of patients may need to be reduced to cost/patient to allow comparability of that value. Data elements may need to be adjusted to make them (b) comparable over time also. For example, currency values may need to be adjusted for inflation; they would need to be converted to the same base year for comparability. They may need to be converted to a common currency. Data should be (c) stored at the same granularity to ensure comparability. For example, sales data may be available daily, but the sales person compensation data may only be available monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.
4. *Continuous values may need to be binned* into a few buckets to help with some analyses. For instance, work experience could be

- binned as low, medium, and high.
5. *Outlier data elements need to be removed* after careful review, to avoid the skewing of results. For example, one big donor could skew the analysis of alumni donors in an educational setting.
  6. *Ensure that the data is representative of the phenomena* under analysis by correcting for any biases in the selection of data. For example, if the data includes many more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.
  7. Data may need to be selected to *increase information density*. Some data may not show much variability, because it was not properly recorded or for other reasons. This data may dull the effects of other differences in the data and should be removed to improve the information density of the data.

## Outputs of Data Mining

Data mining techniques can serve different types of objectives. The outputs of data mining will reflect the objective being served. There are many ways of representing the outputs of data mining.

One popular form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually follow the steps to make a model-based decision. The tree may have certain attributes, such as probabilities assigned to each branch. A related format is a set of business rules, which are if-then statements that show causality. A decision tree can be mapped to business rules. If the objective function is prediction, then a decision tree or business rules are the most appropriate mode of representing the output.

The output can be in the form of a regression equation or mathematical function that represents the best fitting curve to represent the data. This equation may include linear and nonlinear terms. Regression equations are a good way of representing the output of classification exercises. These are also a good representation of forecasting formulae.

Population “centroid” is a statistical measure for describing central tendencies of a collection of data points. These might be defined in a multidimensional space. For example, a centroid could be “middle-aged, highly educated, high-net worth professionals, married with two children,

living in the coastal areas”. Or a population of “20-something, ivy-league-educated, tech entrepreneurs based in Silicon Valley”. Or it could be a collection of “vehicles more than 20 years old, giving low mileage per gallon, which failed environmental inspection”. These are typical representations of the output of a cluster analysis exercise.

Business rules are an appropriate representation of the output of a market basket analysis exercise. These rules are if-then statements with some probability parameters associated with each rule. For example, those that buy milk and bread will also buy butter (with 80 percent probability).

## Evaluating Data Mining Results

There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.

### Predictive Accuracy = (Correct Predictions) / Total Predictions

Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created. The model is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the other hand, when a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). Similarly, when a true-negative data point is classified as positive, that is classified as a false positive (FP). This is represented using the confusion matrix (Figure 4.1).

ConfusionMatrix		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Predicted class	(TP)		True Negative (TN)
	Negative	False Negative (FN)	

Figure 4.1: Confusion Matrix

Thus the predictive accuracy can be specified by the following formula.

$$\text{Predictive Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}).$$

All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100%. In practice, predictive models with more than 70% accuracy can be considered usable in business domains, depending upon the nature of the business.

There are no good objective measures to judge the accuracy of unsupervised learning techniques such as Cluster Analysis. There is no single right answer for the results of these techniques. For example, the value of the segmentation model depends upon the value the decision-maker sees in those results.

## Data Mining Techniques

Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved (Figure 4.2).

Data Mining Techniques		
<b>Supervised Learning</b> (Predictive ability based on past data)	Classification – Machine Learning	Decision Trees
		Neural Networks
	Classification - Statistics	Regression
<b>Unsupervised Learning</b> (Exploratory analysis to discover patterns)	Clustering Analysis	
	Association Rules	

Figure 4.2: Important Data Mining Techniques

The most important class of problems solved using data mining are classification problems. Classification techniques are called supervised learning as there is a way to supervise whether the model is providing the right or wrong answers. These are problems where data from past decisions is

mined to extract the few rules and patterns that would improve the accuracy of the decision making process in the future. The data of past decisions is organized and mined for decision rules or equations, that are then codified to produce more accurate decisions.

*Decision trees* are the most popular data mining technique, for many reasons.

1. Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
2. Decision trees select the most relevant variables automatically out of all the available variables for decision making.
3. Decision trees are tolerant of data quality issues and do not require much data preparation from the users.
4. Even non-linear relationships can be handled well by decision trees.

There are many algorithms to implement decision trees. Some of the popular ones are C5, CART and CHAID.

*Regression* is a most popular statistical data mining technique. The goal of regression is to derive a smooth well-defined curve to best fit the data.

Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature. Simply plotting the data may show a non-linear curve. Applying a non-linear regression equation will fit the data very well with high accuracy. Once such a regression model has been developed, the energy consumption on any future day can be predicted using this equation. The accuracy of the regression model depends entirely upon the dataset used and not at all on the algorithm or tools used.

*Artificial Neural Networks* (ANN) is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure: Neurons receive stimuli, process them, and communicate their results to other neurons successively, and eventually a neuron outputs a decision. A decision task may be processed by just one neuron and the result may be communicated soon. Alternatively, there could be many layers of neurons involved in a decision task, depending upon the complexity of the domain. The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation and communication parameters based

on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make any intuitive sense to an observer. Thus, the neural networks are considered a black-box system.

At some point, the neural network will have learned enough and begin to match the predictive accuracy of a human expert or alternative classification techniques. The predictions of some ANNs that have been trained over a long period of time with a large amount of data have become decisively more accurate than human experts. At that point, the ANNs can begin to be seriously considered for deployment, in real situations in real time. ANNs are popular because they are eventually able to reach a high predictive accuracy. ANNs are also relatively simple to implement and do not have any issues with data quality. However, ANNs require a lot of data to train it to develop good predictive ability.

*Cluster Analysis* is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural groupings of things. Data instances that are similar to (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters. There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number (K) of clusters from the data.

Clustering is also known as the segmentation technique. It helps divide and conquer large data sets. The technique shows the clusters of things from past data. The output is the centroids for each cluster and the allocation of data points to their cluster. The centroid definition is used to assign new data instances that can be assigned to their cluster homes. Clustering is also a part of the artificial intelligence family of techniques.

*Association rules* are a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities. This is the heart of the personalization engine used by ecommerce sites like Amazon.com and streaming movie sites like Netflix.com. The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form  $X \text{ } \textcircled{R} \text{ } Y$ , where  $X$  and  $Y$  are sets of data items. A form of unsupervised learning, it has no dependent variable; and there are

no right or wrong answers. There are just stronger and weaker affinities. Thus, each rule has a confidence level assigned to it. A part of the machine learning family, this technique achieved legendary status when a fascinating relationship was found in the sales of diapers and beers.

## Tools and Platforms for Data Mining

Data Mining tools have existed for many decades. However, they have recently become more important as the values of data have grown and the field of big data analytics has come into prominence. There are a wide range of data mining platforms available in the market today.

1. *Simple or sophisticated*: There are simple end-user data mining tools such as MS Excel, and there are more sophisticated tools such as IBM SPSS Modeler.
2. *Stand-alone or Embedded*: There are stand alone tools and there are tools embedded in an existing transaction processing or data warehousing or ERP system.
3. *Open source or Commercial*: There are open source and freely available tools such as Weka, and there are commercial products.
4. *User interface*: There are text-based tools that require some programming skills, and there are GUI-based drag-and-drop format tools.
5. *Data formats*: There are tools that work only on proprietary data formats and there are those directly accept data from a host of popular data management tools formats.

Here we compare three platforms that we have used extensively and effectively for many data mining projects.

**Table 4.1: Comparison of Popular Data Mining Platforms**

Feature	Excel	Weka	R	IBM SPSS Modeler
Ownership	Commercial	Open-source, free	Open-source, free	Commercial, expensive
Data Mining Features	Limited; extensible with add-on modules	Extensive features, issues with large data sets	Extensive features	Extensive features, unlimited data sizes
Stand-alone	Stand-alone	Stand-alone	Embeddable in	Embedded in BI

			other systems	software suites
User skills needed	End-users	Skilled BI analysts	Programming skills	For skilled BI analysts
User interface	Text and click, Easy	GUI, mostly b&w text output	Integrated Development Environment	Drag & Drop use, colorful, beautiful GUI
Data formats	Industry-standard	Proprietary, CSV	CSV	Variety of data sources accepted

MS Excel is a relatively simple and easy data mining tool. It can get quite versatile once Analyst Pack and some other add-on products are installed on it.

Weka is an open-source GUI based tool that offers a large number of data mining algorithms.

R is an extensive and extensible, versatile open-source statistical programming language with 600+ libraries and 120,000 functions. It is very popular with startup companies, and increasingly so in the large organizations.

IBM's SPSS Modeler is an industry-leading data mining platform. It offers a powerful set of tools and algorithms for most popular data mining capabilities. It has colorful GUI format with drag-and-drop capabilities. It can accept data in multiple formats including reading Excel files directly.

ERP systems include some data analytic capabilities, too. SAP has its Business Objects (BO) software. BO is considered one of the leading BI suites in the industry, and is often used by organizations that use SAP.

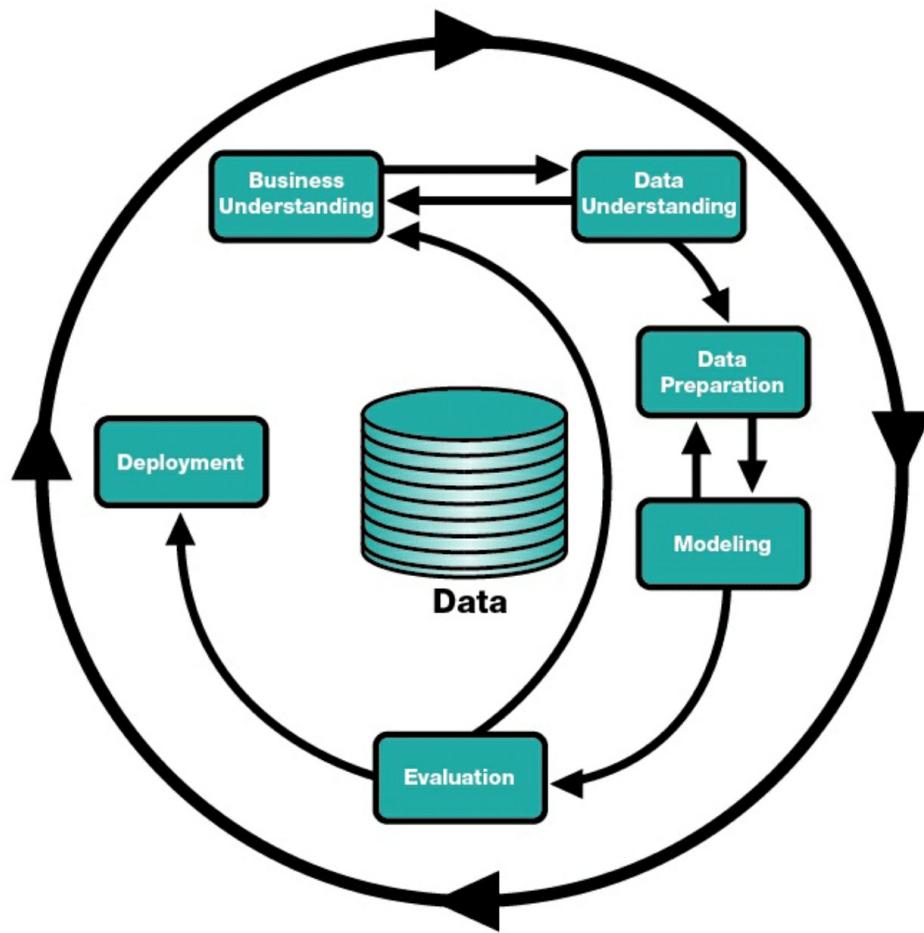
## **Data Mining Best Practices**

Effective and successful use of data mining activity requires both business and technology skills. The business aspects help understand the domain and the key questions. It also helps one imagine possible relationships in the data, and create hypotheses to test it. The IT aspects help fetch the data from many sources, clean up the data, assemble it to meet the needs of the business problem, and then run the data mining techniques on the platform.

An important element is to go after the problem iteratively. It is better to divide and conquer the problem with smaller amounts of data, and get closer to the heart of the solution in an iterative sequence of steps. There are several

best practices learned from the use of data mining techniques over a long period of time. The Data Mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM). It has six essential steps (Figure 4.3):

1. *Business Understanding*: The first and most important step in data mining is asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise. In other words, selecting a data mining project is like any other project, in that it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy. A related important step is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in terms of a proposed model as well in the data sets available and required.



### Figure 4.3: CRISP-DM Data Mining cycle

2. *Data Understanding:* A related important step is to understand the data available for mining. One needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem. Without relevant data, the hypotheses cannot be tested.
3. *Data Preparation:* The data should be relevant, clean and of high quality. It's important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. Data cleaning can take 60-70% of the time in a data mining project. It may be desirable to continue to experiment and add new data elements from external sources of data that could help improve predictive accuracy.
4. *Modeling:* This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported. Patience is required in continuously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used. A tool could be tried with different options, such as running different decision tree algorithms.
5. *Model Evaluation:* One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques, and conducting many what-if scenarios, to build confidence in the solution. One should evaluate and improve the model's predictive accuracy with more test data. When the accuracy has reached some satisfactory level, then the model should be deployed.
6. *Dissemination and rollout:* It is important that the data mining solution is presented to the key stakeholders, and is deployed in the organization. Otherwise the project will be a waste of time and will be a setback for establishing and supporting a data-based decision-process culture in the organization. The model should be eventually embedded in the organization's business processes.

### Myths about data mining

There are many myths about this area, scaring away many business executives from using data mining. Data Mining is a mindset that

presupposes a faith in the ability to reveal insights. By itself, data mining is not too hard, nor is it too easy. It does require a disciplined approach and some cross-disciplinary skills.

*Myth #1:* Data Mining is about algorithms. Data mining is used by business to answer important and practical business questions. Formulating the problem statement correctly and identifying imaginative solutions for testing are far more important before the data mining algorithms gets called in. Understanding the relative strengths of various algorithms is helpful but not mandatory.

*Myth #2:* Data Mining is about predictive accuracy. While important, predictive accuracy is a feature of the algorithm. As in myth#1, the quality of output is a strong function of the right problem, right hypothesis, and the right data.

*Myth #3:* Data Mining requires a data warehouse. While the presence of a data warehouse assists in the gathering of information, sometimes the creation of the data warehouse itself can benefit from some exploratory data mining. Some data mining problems may benefit from clean data available directly from the DW, but a DW is not mandatory.

*Myth #4:* Data Mining requires large quantities of data. Many interesting data mining exercises are done using small or medium sized data sets, at low costs, using end-user tools.

*Myth #5:* Data Mining requires a technology expert. Many interesting data mining exercises are done by end-users and executives using simple everyday tools like spreadsheets.

## **Data Mining Mistakes**

Data mining is an exercise in extracting non-trivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide. Much domain knowledge, tools and skill is required to find such patterns. Here are some of the more common mistakes in doing data mining, and should be avoided.

*Mistake #1: Selecting the wrong problem for data mining:* Without the right goals or having no goals, data mining leads to a waste of time. Getting the right answer to an irrelevant question could be interesting, but it would be pointless from a business perspective. A good goal would be one that would

deliver a good ROI to the organization.

*Mistake #2: Buried under mountains of data without clear metadata:* It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought. There may be insufficient knowledge about the data, or metadata. Examine the data with a critical eye and do not naively believe everything you are told about the data.

*Mistake #3: Disorganized data mining:* Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy. This can come from being sloppy about keeping track of the data mining procedure and results. Not leaving sufficient time for data acquisition, selection and preparation can lead to data quality issues, and GIGO. Similarly not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

*Mistake #4: Insufficient business knowledge:* Without a deep understanding of the business domain, the results would be gibberish and meaningless. Don't make erroneous assumptions, courtesy of experts. Don't rule out anything when observing data analysis results. Don't ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even when insights emerge at one level, it is important to slice and dice the data at other levels to see if more powerful insights can be extracted.

*Mistake #5: Incompatibility of data mining tools and datasets.* All the tools from data gathering, preparation, mining, and visualization, should work together. Use tools that can work with data from multiple sources in multiple industry standard formats.

*Mistake #6: Looking only at aggregated results and not at individual records/predictions.* It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level. Diving into the data at the right angle can yield insights at many levels of data.

*Mistake #7: Not measuring your results differently from the way your sponsor measures them.* If the data mining team loses its sense of business objectives, and beginning to mine data for its own sake, it will lose respect and executive support very quickly. The BIDM cycle (Figure 1.1) should be remembered.

## Conclusion

Data Mining is like diving into the rough material to discover a valuable finished nugget. While the technique is important, domain knowledge is also important to provide imaginative solutions that can then be tested with data mining. The business objective should be well understood and should always be kept in mind to ensure that the results are beneficial to the sponsor of the exercise.

## Review Questions

1. What is data mining? What are supervised and unsupervised learning techniques?
2. Describe the key steps in the data mining process. Why is it important to follow these processes?
3. What is a confusion matrix?
4. Why is data preparation so important and time consuming?
5. What are some of the most popular data mining techniques?
6. What are the major mistakes to be avoided when doing data mining?
7. What are the key requirements for a skilled data analyst?

### Liberty Stores Case Exercise: Step 3

*Liberty is constantly evaluating opportunities for improving efficiencies in all its operations, including the commercial operations as well its charitable activities.*

1. *What data mining techniques would you use to analyze and predict sales patterns?*
2. *What data mining technique would you use to categorize its customers*

## Chapter 5: Data Visualization

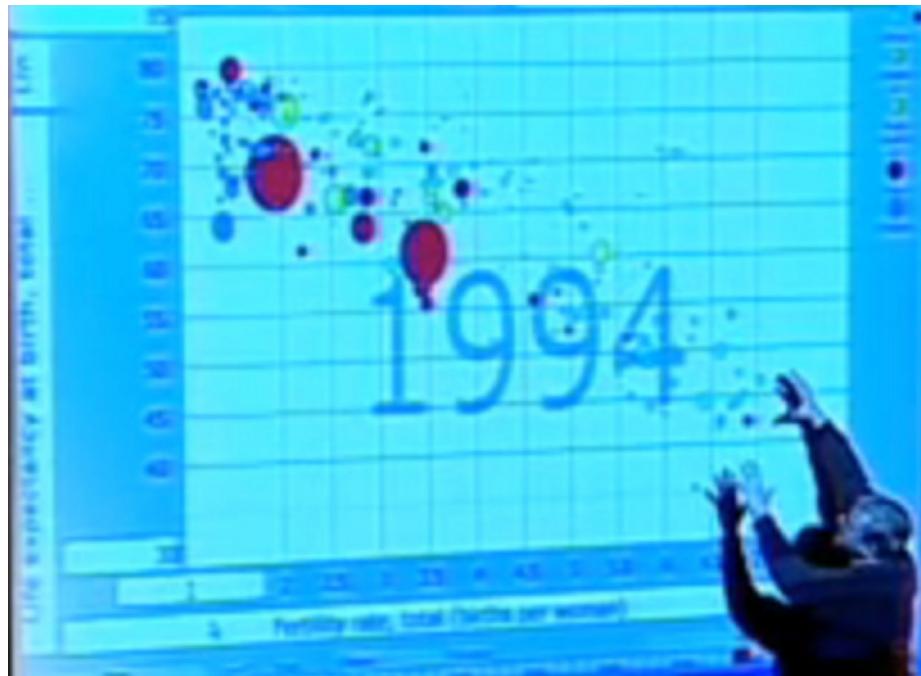
### Introduction

Data Visualization is the art and science of making data easy to understand and consume, for the end user. Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information. The right visualization requires an understanding of the consumer's needs, nature of the data, and the many tools and techniques available to present data. The right visualization arises from a complete understanding of the totality of the situation. One should use visuals to tell a true, complete and fast-paced story.

Data visualization is the last step in the data life cycle. This is where the data is processed for presentation in an easy-to-consume manner to the right audience for the right purpose. The data should be converted into a language and format that is best preferred and understood by the consumer of data. The presentation should aim to highlight the insights from the data in an actionable manner. If the data is presented in too much detail, then the consumer of that data might lose interest and the insight.

### **Caselet: Dr Hans Gosling - Visualizing Global Public Health**

*Dr. Hans Rosling is a master at data visualization. He has perfected the art of showing data in novel ways to highlight unexpected truths. He has become an online star by using data visualizations to make serious points about global health policy and development. Using novel ways to illustrate data obtained from UN agencies, he has helped demonstrate the progress that the world has made in improving public health on many dimensions. The best way to grasp the power of his work is to click [here to see this TED video](#), where Life Expectancy is mapped along with Fertility Rate for all countries from 1962 to 2003. Figure 5.1 shows a one graphic from this video.*



**Figure 5.1: Visualizing Global Health Data (source: ted.com)**

*“THE biggest myth is that if we save all the poor kids, we will destroy the planet,” says Hans Rosling, a doctor and professor of international health at the Karolinska Institute in Sweden. “But you can't stop population growth by letting poor children die.” He has the computerised graphs to prove it: colourful visuals with circles that swarm, swell and shrink like living creatures. Dr Rosling's mesmerizing graphics have been impressing audiences on the international lecture circuit, from the TED conferences to the World Economic Forum at Davos. Instead of bar charts and histograms, Dr Rosling uses Lego bricks, IKEA boxes and data-visualization software developed by his Gapminder Foundation to transform reams of economic and public-health data into gripping stories. His aim is ambitious. “I produce a road map for the modern world,” he says. “Where people want to drive is up to them. But I have the idea that if they have a proper road map and know what the global realities are, they'll make better decisions.” (source: economist.com).*

*Q1: What are the business and social implications of this kind of data visualization?*

*Q2: How could these techniques be applied in your organization and area of work?*

## **Excellence in Visualization**

Data can be presented in the form of rectangular *tables*, or it can be presented

in colorful graphs of various types. “Small, non-comparative, highly-labeled data sets usually belong in tables” – (Ed Tufte, 2001, p 33). However, as the amount of data grows, graphs are preferable. Graphics help give shape to data. Tufte, a pioneering expert on data visualization, presents the following objectives for graphical excellence:

1. *Show, and even reveal, the data:* The data should tell a story, especially a story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.
2. *Induce the viewer to think of the substance of the data:* The format of the graph should be so natural to the data, that it hides itself and lets data shine.
3. *Avoid distorting what the data have to say:* Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.
4. *Make large data sets coherent:* By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.
5. *Encourage the eyes to compare different pieces of data:* Organize the chart in ways the eyes would naturally move to derive insights from the graph.
6. *Reveal the data at several levels of detail:* Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.
7. *Serve a reasonably clear purpose* – informing or decision-making.
8. *Closely integrate with the statistical and verbal descriptions of the dataset:* There should be no separation of charts and text in presentation. Each mode should tell a complete story. Intersperse text with the map/graphic to highlight the main insights.

Context is important in interpreting graphics. Perception of the chart is as important as the actual charts. Do not ignore the intelligence or the biases of the reader. Keep the template consistent, and only show variations in data. There can be many excuses for graphical distortion. E.g. “we are just approximating.” Quality of information transmission comes prior to aesthetics of chart. Leaving out the contextual data can be misleading.

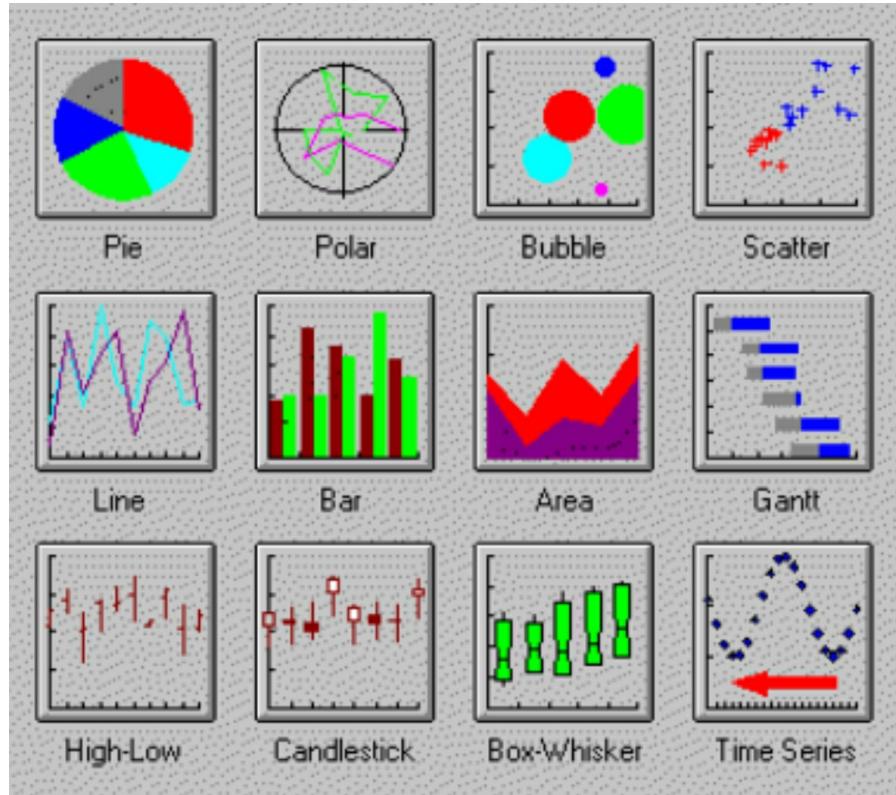
A lot of graphics are published because they serve a particular cause or a point of view. It is particularly important when in a for-profit or politically contested environments. Many related dimensions can be folded into a graph. The more the dimensions that are represented in a graph, the richer and more useful the chart become. The data visualizer should understand the client's objects and present the data for accurate perception of the totality of the situation.

## Types of Charts

There are many kinds of data as seen in the caselet above. Time series data is the most popular form of data. It helps reveal patterns over time. However, data could be organized around alphabetical list of things, such as countries or products or salespeople. Figure 5.2 shows some of the popular chart types and their usage.

1. *Line graph.* This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.
2. *Scatter plot:* This is another very basic and useful graphic form. It helps reveal the relationship between two variables. In the above caselet, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.
3. *Bar graph:* A bar graph shows thin colorful [rectangular](#) bars with their [lengths](#) being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.
4. *Stacked Bar graphs:* These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.
5. *Histograms:* These are like bar graphs, except that they are useful

in showing data frequencies or data values on classes (or ranges) of a numerical variable.

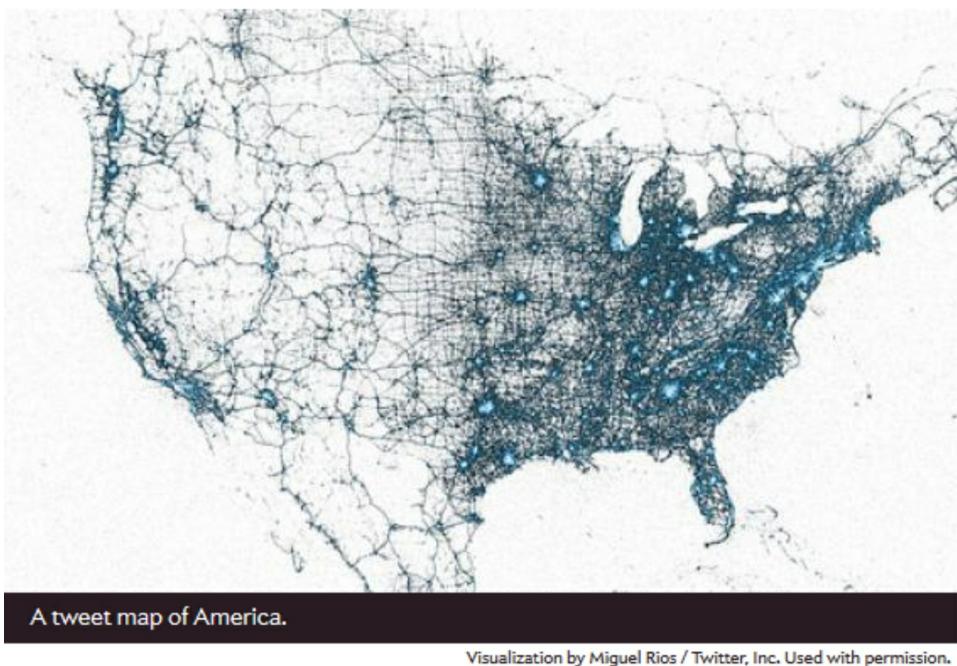


**Figure 5.1: Many types of graphs**

6. *Pie charts:* These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.
7. *Box charts:* These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.
8. *Bubble Graph:* This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) ... the size of the circle and the color fill in the circle could represent two additional dimensions.
9. *Dials:* These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low

range, medium range, or high range. These ranges could be colored red, yellow and green to give an instant view of the data.

10. *Geographical Data* maps are particularly useful maps to denote statistics. Figure 5.3 shows a tweet density map of the US. It shows where the tweets emerge from in the US.



**Figure 5.3: US tweet map (Source: Slate.com)**

11. *Pictographs:* One can use pictures to represent data. E.g. Figure 5.2 shows the data of liters of water needed to produce one pound of each of the products, where images are used to show the product for easy reference. Each droplet of water also represents 50 liters of water.

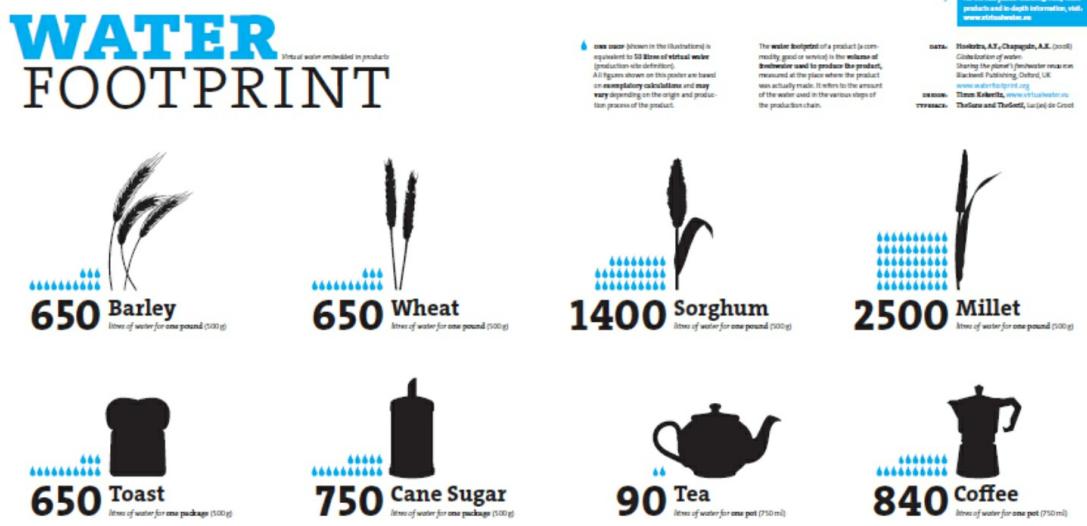


Figure 5.4: Pictograph of Water footprint (source : [waterfootprint.org](http://waterfootprint.org))

## Visualization Example

To demonstrate how each of the visualization tools could be used, imagine an executive for a company who wants to analyze the sales performance of his division. Figure 5.1 show the important raw sales data for the current year, alphabetically sorted by Product names.

Product	Revenue	Orders	SalesPers
AA	9731	131	23
BB	355	43	8
CC	992	32	6
DD	125	31	4
EE	933	30	7
FF	676	35	6
GG	1411	128	13
HH	5116	132	38
JJ	215	7	2
KK	3833	122	50
LL	1348	15	7
MM	1201	28	13

Table 5.1: Raw Performance Data

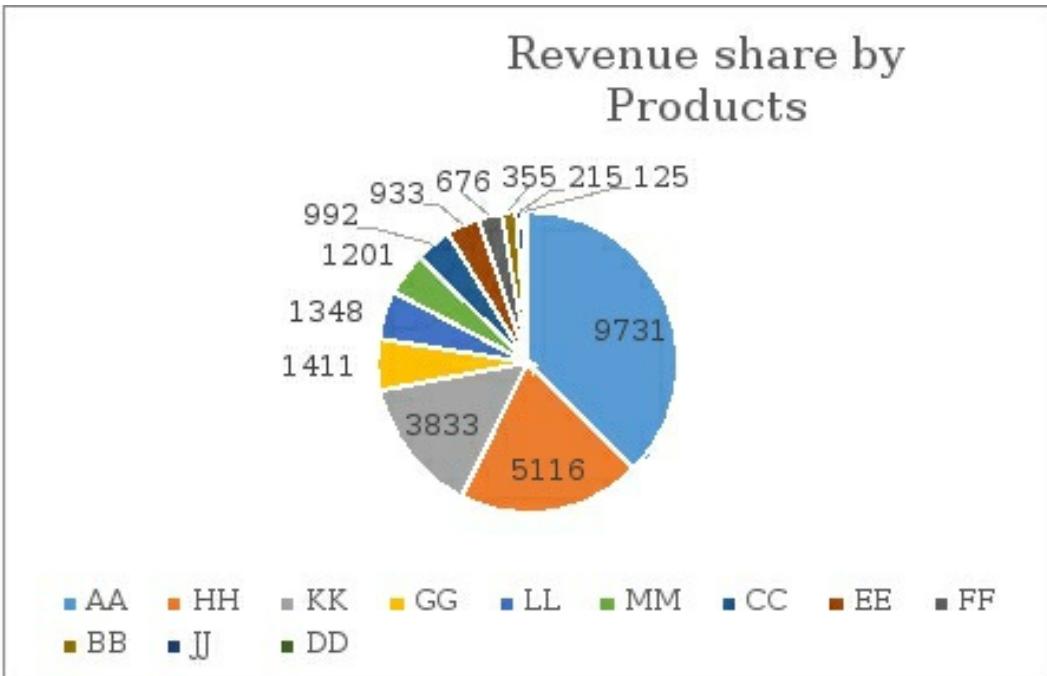
To reveal some meaningful pattern, a good first step would be to sort the table by Product revenue, with highest revenue first. We could total up the values of Revenue, Orders, and Sales persons for all products. We can also add some important ratios to the right of the table (Table 5.2).

<b>Product</b>	<b>Revenue</b>	<b>Orders</b>	<b>SalesPers</b>	<i>Rev/Order</i>	<i>Rev/SalesP</i>	<i>Orders/SalesP</i>
AA	9731	131	23	74.3	423.1	5.7
HH	5116	132	38	38.8	134.6	3.5
KK	3833	122	50	31.4	76.7	2.4
GG	1411	128	13	11.0	108.5	9.8
LL	1348	15	7	89.9	192.6	2.1
MM	1201	28	13	42.9	92.4	2.2
CC	992	32	6	31.0	165.3	5.3
EE	933	30	7	31.1	133.3	4.3
FF	676	35	6	19.3	112.7	5.8
BB	355	43	8	8.3	44.4	5.4
JJ	215	7	2	30.7	107.5	3.5
DD	125	31	4	4.0	31.3	7.8
<b>Total</b>	<b>25936</b>	<b>734</b>	<b>177</b>	<b>35.3</b>	<b>146.5</b>	<b>4.1</b>

**Table 5.2: Sorted data, with additional ratios**

There are too many numbers on this table to visualize any trends in them. The numbers are in different scales so plotting them on the same chart would not be easy. E.g. the Revenue numbers are in thousands while the SalesPers numbers and Orders/SalesPers are in the single or double digit.

One could start by visualizing the revenue as a pie-chart. The revenue proportion drops significantly from the first product to the next. (Figure 5.5). It is interesting to note that the top 3 products produce almost 75% of the revenue.



**Figure 5.5: Revenue Share by Product**

The number of orders for each product can be plotted as a bar graph (Figure 5.2). This shows that while the revenue is widely different for the top four products, they have approximately the same number of orders.



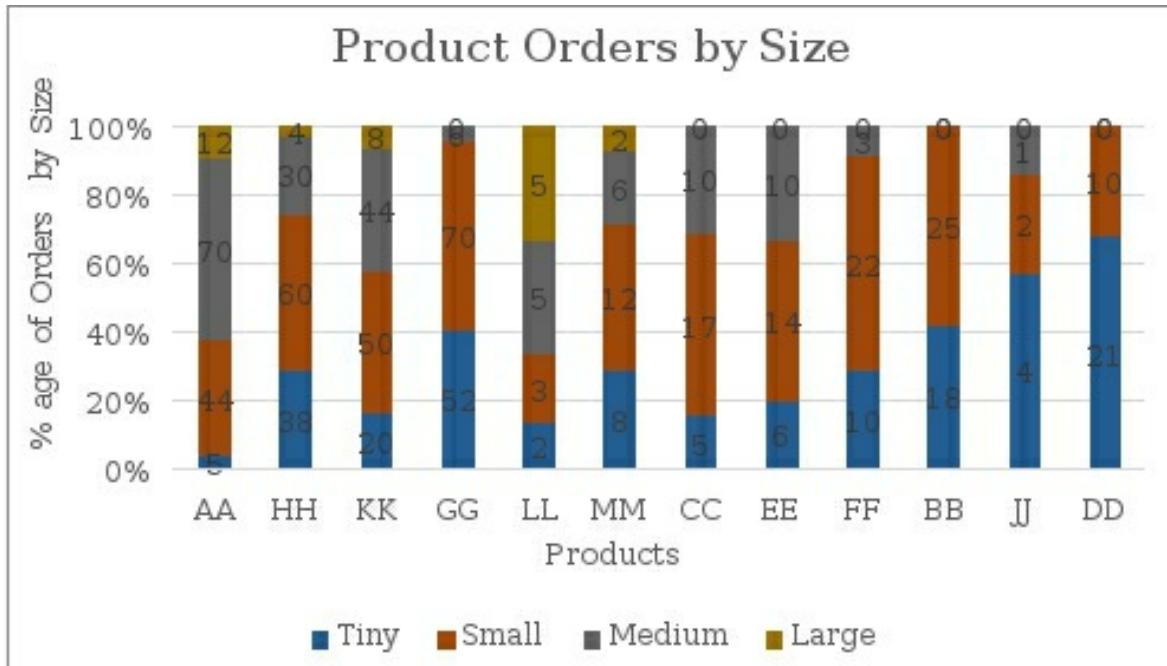
**Figure 5.6: Orders by Products**

Therefore, the orders data could be investigated further to see order patterns. Suppose additional data is made available for Orders by their size. Suppose the orders are chunked into 4 sizes: Tiny, Small, Medium, and Large. Additional data is shown in Table 5.3.

Product	Total Orders	Tiny	Small	Medium	Large
AA	131	5	44	70	12
HH	132	38	60	30	4
KK	122	20	50	44	8
GG	128	52	70	6	0
LL	15	2	3	5	5
MM	28	8	12	6	2
CC	32	5	17	10	0
EE	30	6	14	10	0
FF	35	10	22	3	0
BB	43	18	25	0	0
JJ	7	4	2	1	0
DD	31	21	10	0	0
Total	734	189	329	185	31

**Table 5.3: Additional data on order sizes**

Figure 5.7 is a stacked bar graph that shows the percentage of Orders by size for each product. This chart (Figure 5.7) brings a different set of insights. It shows that the product HH has a larger proportion of tiny orders. The products at the far right have a large number of tiny orders and very few large orders.



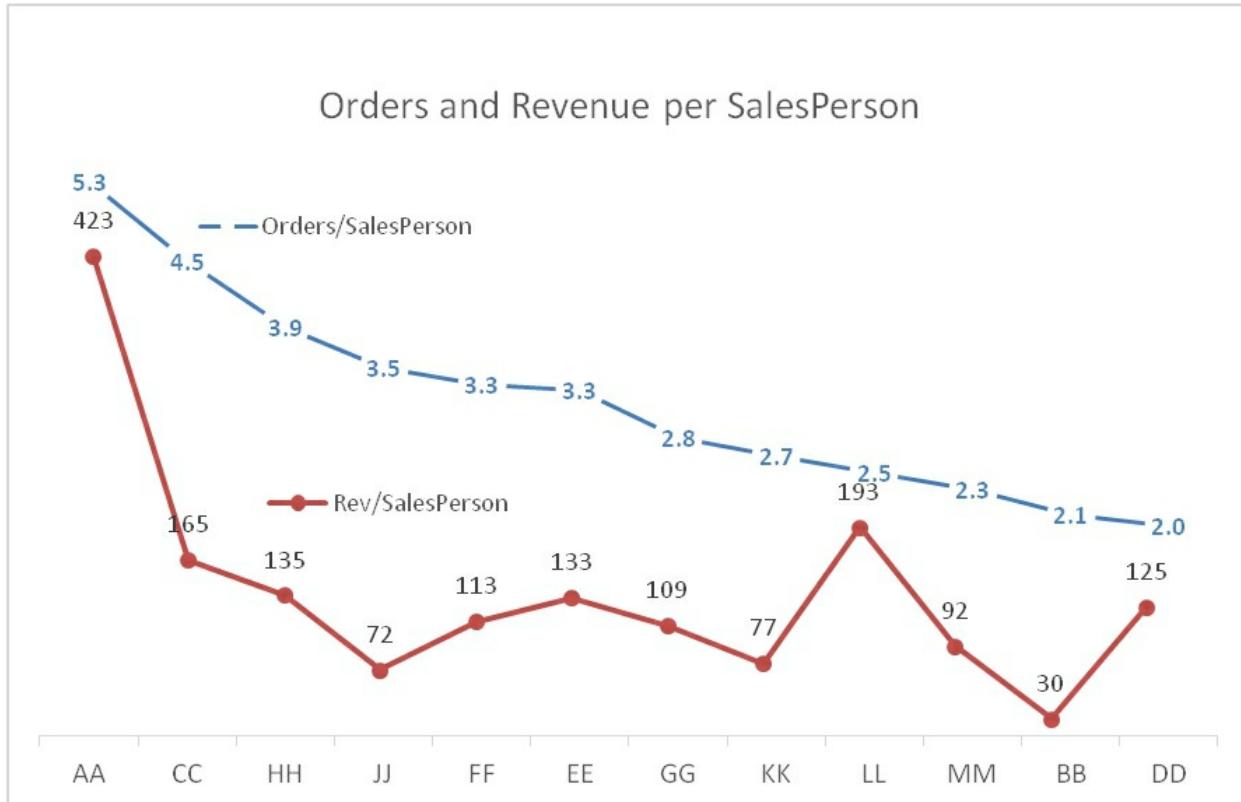
**Figure 5.7: Product Orders by Order Size**

## Visualization Example phase -2

The executive wants to understand the productivity of salespersons. This analysis could be done both in terms of the number of orders, or revenue, per salesperson. There could be two separate graphs, one for the number of orders per salesperson, and the other for the revenue per salesperson. However, an interesting way is to plot both measures on the same graph to give a more complete picture. This can be done even when the two data have different scales. The data is here resorted by number of orders per salesperson.

Figure 5.8 shows two line graphs superimposed upon each other. One line shows the revenue per salesperson, while the other shows the number of orders per salesperson. It shows that the highest productivity of 5.3 orders per sales person, down to 2.1 orders per salesperson. The second line, the blue line shows the revenue per sales person for each for the products. The revenue per salesperson is highest at 630, while it is lowest at just 30.

And thus additional layers of data visualization can go on for this data set.



**Figure 5.8: Salesperson productivity by product**

## Tips for Data Visualization

To help the client in understanding the situation, the following considerations are important:

1. *Fetch appropriate and correct data for analysis.* This requires some understanding of the domain of the client and what is important for the client. E.g. in a business setting, one may need to understand the many measures of profitability and productivity.
2. *Sort the data in the most appropriate manner.* It could be sorted by numerical variables, or alphabetically by name.
3. *Choose appropriate method to present the data.* The data could be presented as a table, or it could be presented as any of the graph types.
4. *The data set could be pruned* to include only the more significant elements. More data is not necessarily better, unless it makes the most significant impact on the situation.
5. *The visualization could show additional dimension for reference*

- such as the expectations or targets with which to compare the results.
- 6. *The numerical data may need to be binned into a few categories.* E.g. the orders per person were plotted as actual values, while the order sizes were binned into 4 categorical choices.
  - 7. *High-level visualization could be backed by more detailed analysis.* For the most significant results, a drill-down may be required.
  - 8. *There may be need to present additional textual information* to tell the whole story. For example, one may require notes to explain some extraordinary results.

## Conclusion

Data Visualization is the last phase of the data lifecycle, and leads to the consumption of data by the end user. It should tell an accurate, complete and simple story backed by data, while keeping it insightful and engaging. There are innumerable types of visual graphing techniques available for visualizing data. The choice of the right tools requires a good understanding of the business domain, the data set and the client needs. There is ample room for creativity to design ever more compelling data visualization to most efficiently convey the insights from the data.

## Review Questions

1. What is data visualization?
2. How would you judge the quality of data visualizations?
3. What are the data visualization techniques? When would you use tables or graphs?
4. Describe some key steps in data visualization.
5. What are some key requirements for good visualization.

### Liberty Stores Case Exercise: Step 4

*Liberty is constantly evaluating its performance for improving efficiencies in all its operations, including the commercial operations as well its charitable activities.*

1. *What data visualization techniques would you use to help understand sales patterns?*
2. *What data visualization technique would you use to categorize its customers?*

## **Section 2 – Popular Data Mining Techniques**

This section covers five important data mining techniques.

The first three techniques are examples of supervised learning, consisting of classification techniques.

Chapter 6 covers decision trees, which are the most popular form of data mining techniques. There are many algorithms to develop decision trees.

Chapter 7 describes regression modeling techniques. These are statistical techniques.

Chapter 8 covers artificial neural networks, which are a machine learning technique.

The next two techniques are examples of unsupervised learning, consisting of data exploration techniques.

Chapter 9 covers Cluster Analysis. This is also called Market Segmentation analysis.

Chapter 10 covers the Association Rule Mining technique, also called Market Basket Analysis.

# Chapter 6: Decision Trees

## Introduction

Decision trees are a simple way to guide one's path to a decision. The decision may be a simple binary one, whether to approve a loan or not. Or it may be a complex multi-valued decision, as to what may be the diagnosis for a particular sickness. Decision trees are hierarchically branched structures that help one come to a decision based on asking certain questions in a particular sequence. Decision trees are one of the most widely used techniques for classification. A good decision tree should be short and ask only a few meaningful questions. They are very efficient to use, easy to explain, and their classification accuracy is competitive with other methods. Decision trees can generate knowledge from a few test instances that can then be applied to a broad population. Decision trees are used mostly to answer relatively simple binary decisions.

### Caselet: Predicting Heart Attacks using Decision Trees

*A study was done at UC San Diego concerning heart disease patient data. The patients were diagnosed with a heart attack from chest pain, diagnosed by EKG, high enzyme levels in their heart muscles, etc. The objective was to predict which of these patients was at risk of dying from a second heart attack within the next 30 days. The prediction would determine the treatment plan, such as whether to keep the patient in intensive care or not. For each patient more than 100 variables were collected, including demographics, medical history and lab data. Using that data, and the CART algorithm, a decision tree was constructed.*

*The decision tree showed that if Blood Pressure was low ( $<=90$ ), the chance of another heart attack was very high (70%). If the patient's BP was ok, the next question to ask was the patient's age. If the age was low ( $<=62$ ), then the patient's survival was almost guaranteed (98%). If the age was higher, then the next question to ask was about sinus problems. If their sinus was ok, the chances of survival were 89%. Otherwise, the chance of survival dropped to 50%. This decision tree predicts 86.5% of the cases correctly. (Source: Salford Systems).*

- 1: Is a decision tree good enough in terms of accuracy, design, readability, for this data etc.*
- 2: Identify the benefits from creating such a decision tree. Can these be quantified?*

## Decision Tree problem

Imagine a conversation between a doctor and a patient. The doctor asks questions to determine the cause of the ailment. The doctor would continue to ask questions, till she is able to arrive at a reasonable decision. If nothing seems plausible, she might recommend some tests to generate more data and options.

This is how experts in any field solve problems. They use decision trees or decision rules. For every question they ask, the potential answers create separate branches for further questioning. For each branch, the expert would know how to proceed ahead. The process continues until the end of the tree is reached, which means a leaf node is reached.

Human experts learn from past experiences or data points. Similarly, a machine can be trained to learn from the past data points and extract some knowledge or rules from it. Decision trees use machine learning algorithms to abstract knowledge from data. A decision tree would have a predictive accuracy based on how often it makes correct decisions.

- The more data is available for training the decision tree, the more accurate its knowledge extraction will be, and thus, it will make more accurate decisions.**
- The more variables the tree can choose from, the greater is the likely of the accuracy of the decision tree.
- In addition, a good decision tree should also be frugal so that it takes the least number of questions, and thus, the least amount of effort, to get to the right decision.

Here is an exercise to create a decision tree that helps make decisions about approving the play of an outdoor game. The objective is to predict the play decision given the atmospheric conditions out there. The decision is: Should the game be allowed or not? Here is the decision problem.

Outlook	Temp	Humidity	Windy	Play

Sunny	Hot	Normal	True	??
-------	-----	--------	------	----

To answer that question, one should look at past experience, and see what decision was made in a similar instance, if such an instance exists. One could look up the database of past decisions to find the answer and try to come to an answer. Here is a list of the decisions taken in 14 instances of past soccer game situations. (Dataset courtesy: Witten, Frank, and Hall, 2010).

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

If there were a row for *Sunny/Hot/Normal/Windy* condition in the data table, it would match the current problem; and the decision from that row could be used to answer the current problem. However, there is no such past instance in this case. There are three disadvantages of looking up the data table:

1. As mentioned earlier, how to decide if there isn't a row that corresponds to the exact situation today? If there is no exact matching instance available in the database, the past experience cannot guide the decision.

2. Searching through the entire past database may be time consuming, depending on the number of variables and the organization of the database.
3. What if the data values are not available for all the variables? In this instance, if the data for humidity variable was not available, looking up the past data would not help.

A better way of solving the problem may be to abstract the knowledge from the past data into decision tree or rules. These rules can be represented in a decision tree, and then that tree can be used make the decisions. The decision tree may not need values for all the variables.

## **Decision Tree Construction**

A decision tree is a hierarchically branched structure. What should be the first question asked in creating the tree? One should ask the more important question first, and the less important questions later. What is the most important question that should be asked to solve the problem? How is the importance of the questions determined? Thus, how should the root node of the tree be determined?

*Determining root node of the tree:* In this example, there are four choices based on the four variables. One could begin by asking one of the following questions: what is the outlook, what is the temperature, what is the humidity, and what is the wind speed? A criterion should be used to evaluate these choices. The key criterion would be that: which one of these questions gives the most insight about the situation? Another way to look at it would be the criterion of frugality. That is, which question will provide us the shortest ultimate decision tree? Another way to look at this is that if one is allowed to ask one and only one question, which one would one ask? In this case, the most important question should be the one that, by itself, helps make the most correct decisions with the fewest errors. The four questions can now be systematically compared, to see which variable by itself will help make the most correct decisions. One should systematically calculate the correctness of decisions based on each question. Then one can select the question with the most correct predictions, or the fewest errors.

Start with the first variable, in this case outlook. It can take three values, sunny, overcast, and rainy.

Start with the sunny value of outlook. There are five instances where the

outlook is sunny. In 2 of the 5 instances the *play* decision was yes, and in the other three, the decision was No. Thus, if the decision rule was that Outlook:sunny → No, then 3 out of 5 decisions would be correct, while 2 out of 5 such decisions would be incorrect. There are 2 errors out of 5. This can be recorded in Row 1.

<u>Attribute</u>	<u>Rules</u>	<u>Error</u>	<u>Total Error</u>
Outlook	Sunny → No	2/5	

Similar analysis would be done for other values of the outlook variable. There are four instances where the outlook is overcast. In all 4 out 4 instances the Play decision was yes. Thus, if the decision rule was that Outlook:overcast → Yes, then 4 out of 4 decisions would be correct, while none of decisions would be incorrect. There are 0 errors out of 4. This can be recorded in the next row.

<u>Attribute</u>	<u>Rules</u>	<u>Error</u>	<u>Total Error</u>
Outlook	Sunny → No	2/5	
	Overcast → yes	0/4	

There are five instances where the outlook is rainy. In 3 of the 5 instances the *play* decision was yes, and in the other three, the decision was no. Thus, if the decision rule was that Outlook:rainy → Yes, then 3 out of 5 decisions would be correct, while 2 out of 5 decisions would be incorrect. There will be 2/5 errors. This can be recorded in next row.

<u>Attribute</u>	<u>Rules</u>	<u>Error</u>	<u>Total Error</u>
Outlook	Sunny → No	2/5	4/14
	Overcast → yes	0/4	
	Rainy → yes	2/5	

Adding up errors for all values of outlook, there are 4 errors out of 14. In other words, Outlook gives 10 correct decisions out of 14, and 4 incorrect ones.

A similar analysis can be done for the other three variables. At the end of that

analytical exercise, the following Error table will be constructed.

<u>Attribute</u>	<u>Rules</u>	<u>Error</u>	<u>Total Error</u>
Outlook	Sunny → No	2/5	4/14
	Overcast → yes	0/4	
	Rainy → yes	2/5	
Temp	Hot → No	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No	3/6	

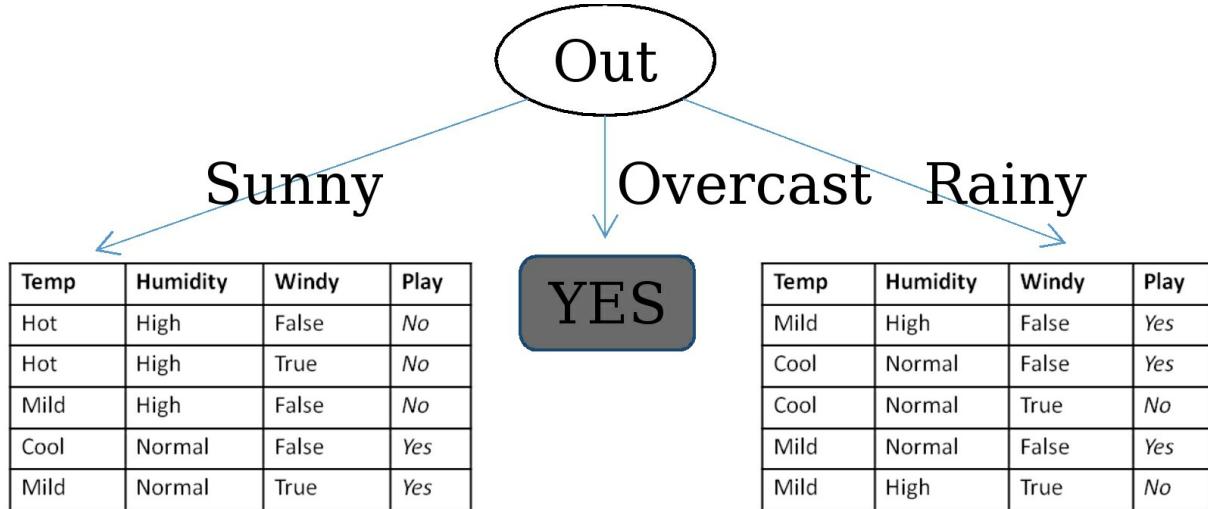
The variable that leads to the least number of errors (and thus the most number of correct decisions) should be chosen as the first node. In this case, two variables have the least number of errors. There is a tie between outlook and humidity, as both have 4 errors out of 14 instances. The tie can be broken using another criterion, the purity of resulting sub-trees.

If all the errors were concentrated in a few of the subtrees, and some of the branches were completely free of error, that is preferred from a usability perspective. Outlook has one error-free branch, for the overcast value, while there is no such pure sub-class for humidity variable. Thus the tie is broken in favor of outlook. The decision tree will use outlook as the first node, or the first splitting variable. The first question that should be asked to solve the Play problem, is ‘What is the value of outlook’?

*Splitting the Tree:* From the root node, the decision tree will be split into three branches or sub-trees, one for each of the three values of outlook. Data for the root node (the entire data) will be divided into the three segments, one for each of the value of outlook. The sunny branch will inherit the data for the instances that had sunny as the value of outlook. These will be used for further building of that sub-tree. Similarly, the rainy branch will inherit data for the instances that had rainy as the value of outlook. These will be used for

further building of that sub-tree. The overcast branch will inherit the data for the instances that had overcast as the outlook. However, there will be no need to build further on that branch. There is a clear decision, yes, for all instances when outlook value is overcast.

The decision tree will look like this after the first level of splitting.



*Determining the next nodes of the tree:* A similar recursive logic of tree building should be applied to each branch. For the sunny branch on the left, error values will be calculated for the three other variables – temp, humidity and windy. Final comparison looks like this:

<u>Attribute</u>	<u>Rules</u>	<u>Error</u>	<u>Total Error</u>
Temp	Hot->No	0/2	1/5
	Mild ->No	1/2	
	Cool -> yes	0/1	
Humidity	High->No	0/3	0/5
	Normal->Yes	0/2	
Windy	False->No	1/3	2/5
	True->Yes	1/2	

The variable of humidity shows the least amount of error, i.e. zero error. The other two variables have non-zero errors. Thus the Outlook:sunny branch on the left will use humidity as the next splitting variable.

Similar analysis should be done for the ‘rainy’ value of the tree. The analysis would look like this.

<b>Attribute</b>	<b>Rules</b>	<b>Error</b>	<b>Total Error</b>
Temp	Mild->Yes	1/3	2/5
	Cool->yes	1/2	
Humidity	High->No	1/2	2/5
	Normal->Yes	1/3	
Windy	False->Yes	0/3	0/5
	True-No	0/2	

For the Rainy branch, it can similarly be seen that the variable Windy gives all the correct answers, while none of the other two variables makes all the correct decisions.

This is how the final decision tree looks like. Here it is produced using Weka open-source data mining platform (Figure 6.1). This is the model that abstracts the knowledge of the past data of decision.

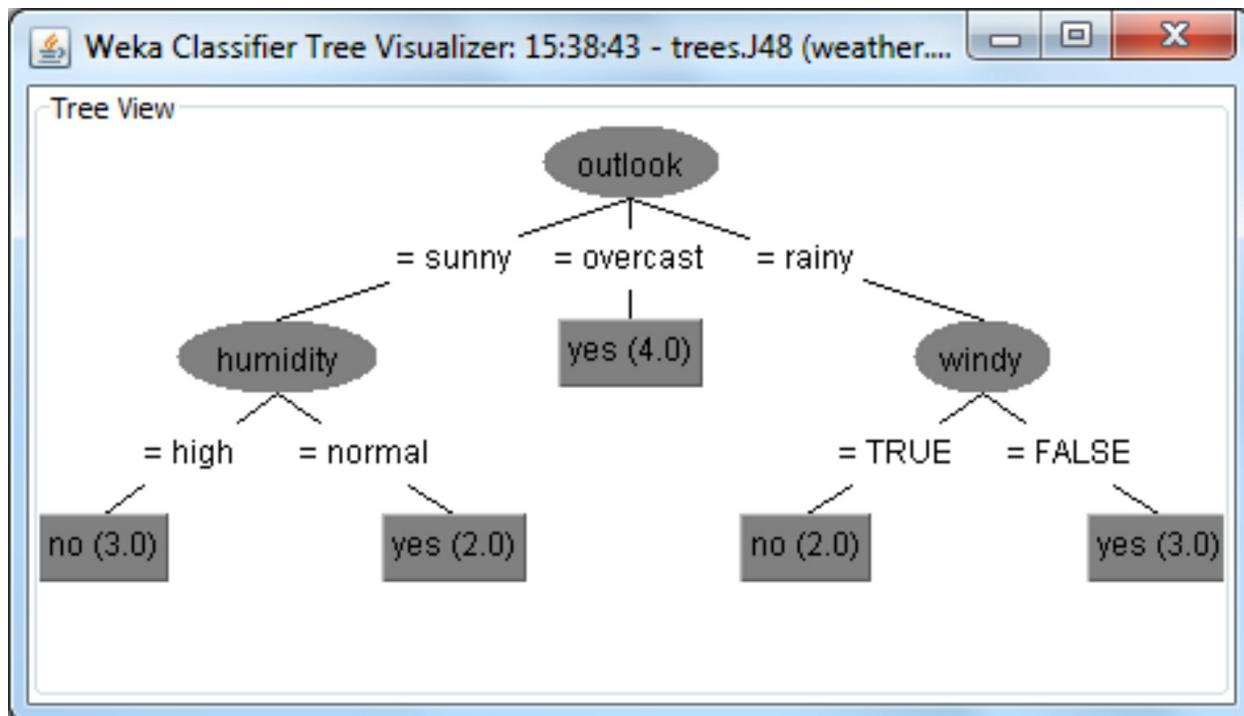


Figure 6.1: Decision Tree for the weather problem

This decision tree can be used to solve the current problem. Here is the problem again.

<b>Outlook</b>	<b>Temp</b>	<b>Humidity</b>	<b>Windy</b>	<b>Play</b>
Sunny	Hot	Normal	True	??

According to the tree, the first question to ask is about outlook. In this problem the outlook is sunny. So, the decision problem moves to the Sunny branch of the tree. The node in that sub-tree is humidity. In the problem, Humidity is Normal. That branch leads to an answer Yes. Thus, the answer to the play problem is Yes.

<b>Outlook</b>	<b>Temp</b>	<b>Humidity</b>	<b>Windy</b>	<b>Play</b>
Sunny	Hot	Normal	True	<b>Yes</b>

## Lessons from constructing trees

Here are some benefits of using this decision tree compared with looking up the answers from the data table (Figure 6.1)

	<b>Decision Tree</b>	<b>Table Lookup</b>
Accuracy	Varied level of accuracy	100% accurate
Generality	General. Applies to all situations	Applies only when a similar case had occurred earlier
Frugality	Only three variables needed	All four variables are needed
Simple	Only one, or max two variable values are needed	All four variable values are needed
Easy	Logical, and easy to understand	Can be cumbersome to look up; no understanding of the logic behind the decision

Figure 6.1: Comparing Decision Tree with Table Look-up

Here are a few observations about how the tree was constructed:

1. The final decision tree has zero errors in mapping to the prior data. In other words, the tree has a *predictive accuracy of 100%*. The tree

completely fits the data. In real life situations, such perfect predictive accuracy is not possible when making decision trees. When there are larger, complicated data sets, with many more variables, a perfect fit is unachievable. This is especially true in business and social contexts, where things are not always fully clear and consistent.

2. The decision tree algorithm *selected the minimum number of variables* that are needed to solve the problem. Thus, one can start with all available data variables, and let the decision-tree algorithm select the ones that are useful, and discard the rest.
3. This tree is *almost symmetric* with all branches being of almost similar lengths. However, in real life situations, some of the branches may be much longer than the others, and the tree may need to be pruned to make it more balanced and usable.
4. It may be possible to *increase predictive accuracy by making more sub-trees* and making the tree longer. However, the marginal accuracy gained from each subsequent level in the tree will be less, and may not be worth the loss in ease and interpretability of the tree. If the branches are long and complicated, it will be difficult to understand and use. The longer branches may need to be trimmed to keep the tree easy to use.
5. A perfectly fitting tree has the *danger of over-fitting the data*, thus capturing all the random variations in the data. It may fit the training data well, but may not do well in predicting the future real instances.
6. There was a *single best tree* for this data. There could however be two or more equally efficient decision trees of similar length with similar predictive accuracy for the same data set. Decision trees are *based strictly on patterns within the data*, and do not rely on any underlying theory of the problem domain. When multiple candidate trees are available, one could choose whichever is easier to understand, communicate or implement.

## Decision Tree Algorithms

As we saw, decision trees employ the divide and conquer method. The data is branched at each node according to certain criteria until all the data is assigned to leaf nodes. It recursively divides a training set until each division consists of examples from one class.

The following is a pseudo code for making decision trees:

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute according to certain criteria.
3. Add a branch to the root node for each value of the split.
4. Split the data into mutually exclusive subsets along the lines of the specific split.
5. Repeat steps 2 and 3 for each and every leaf node until a stopping criteria is reached.

There are many algorithms for making decision trees. Decision tree algorithms differ on three key elements:

1. Splitting criteria
  - a. Which variable to use for the first split? How should one determine the most important variable for the first branch, and subsequently, for each sub-tree?

Algorithms use different measures like least errors, information gain, Gini's coefficient, etc to compute the splitting variable that provides the most benefit. Information gain is a mathematical construct to compute the reduction in [information entropy](#) from a prior state to the next state that takes some information as given. The greater the reduction in entropy, the better it is. The Gini coefficient is a statistical concept that measures the inequality among values of a frequency distribution. The lower the Gini's coefficient, the better it is.

  - b. What values to use for the split? If the variables have continuous values such as for age or blood pressure, what value-ranges should be used to make bins?
  - c. How many branches should be allowed for each node? There could be binary trees, with just two branches at each node. Or there could be more branches allowed.
2. Stopping criteria: When to stop building the tree? There are two major ways to make that determination. The tree building could be stopped when a certain depth of the branches has been reached and the tree becomes unreadable after that. The tree could also be stopped when the error level at any node is within predefined

tolerable levels.

3. Pruning: This is the act of reducing the size of decision trees by removing sections of the tree that provide little value. The decision tree could be trimmed to make it more balanced, more general, and more easily usable. The symptoms of an over-fitted tree are that it is too deep, with too many branches, which may reflect anomalies due to random noise or outliers instead of mainly the underlying relationship. The pruning is often done after the tree is constructed. There are two approaches to avoid over-fitting.
  - Pre-pruning means to halt the tree construction early, when certain criteria are met. The downside is that it is difficult to decide what criteria to use for halting the construction, because we do not know what may happen subsequently, if we keep growing the tree.
  - Post-pruning: Remove branches or sub-trees from a “fully grown” tree. This method is commonly used. C4.5 algorithm uses a statistical method to estimate the errors at each node for pruning. A validation set may be used for pruning as well.

The most popular decision tree algorithms are C5, CART and CHAID (Table 6.2)

**Figure 6.2: Comparing popular Decision Tree algorithms**

Decision-Tree	C4.5	CART	CHAID
Full Name	Iterative Dichotomiser (ID3)	Classification and Regression Trees	Chi-square Automatic Interaction Detector
Basic algorithm	Hunt's algorithm	Hunt's algorithm	adjusted significance testing
Developer	Ross Quinlan	Breiman	Gordon Kass
When developed	1986	1984	1980
Types of trees	Classification	Classification & Regression	Classification & regression

		trees	
Serial implementation	Tree-growth & Tree-pruning	Tree-growth & Tree-pruning	Tree-growth & Tree-pruning
Type of data	Discrete & Continuous; Incomplete data	Discrete and Continuous	Non-normal data also accepted
Types of splits	Multi-way splits	Binary splits only; Clever surrogate splits to reduce tree depth	Multi-way splits as default
Splitting criteria	Information gain	Gini's coefficient, and others	<i>Chi-square test</i>
Pruning Criteria	Clever bottom-up technique avoids overfitting	Remove weakest links first	Trees can become very large
Implementation	Publicly available	Publicly available in most packages	Popular in market research, for segmentation

## Conclusion

Decision trees are the most popular, versatile, and easy to use data mining technique with high predictive accuracy. They are also very useful as communication tools with executives. There are many successful decision tree algorithms. All publicly available data mining software platforms offer multiple decision tree implementations.

## Review Questions

1: What is a decision tree? Why are decision trees the most popular classification technique?

2: What is a splitting variable? Describe three criteria for choosing splitting variable.

3: What is pruning? What are pre-pruning and post-pruning? Why choose one over the other?

4: What are gini's coefficient, and information gain? (Hint: google it).

Hands-on Exercise: Create a decision tree for the following data set. The objective is to predict the class category. (loan approved or not).

Age	Job	House	Credit	LoanApproved
Young	False	No	Fair	No
Young	False	No	Good	No
Young	True	No	Good	Yes
Young	True	Yes	Fair	Yes
Young	False	No	Fair	No
Middle	False	No	Fair	No
Middle	False	No	Good	No
Middle	True	Yes	Good	Yes
Middle	False	Yes	Excellent	Yes
Middle	False	Yes	Excellent	Yes
Old	False	Yes	Excellent	Yes
Old	False	Yes	Good	Yes
Old	True	No	Good	Yes
Old	True	No	Excellent	Yes
Old	False	No	Fair	No

Then solve the following problem using the model.

Age	Job	House	Credit	LoanApproved
Young	False	False	Good	??

### Liberty Stores Case Exercise: Step 5

Liberty is constantly evaluating requests for opening new stores. They would like to formalize the process for handling many requests, so that the best candidates are selected for detailed evaluation.

Develop a decision tree for evaluating new stores options. Here is the training data:

--	--	--	--	--

<b>City-size</b>	<b>Avg Income</b>	<b>Local investors</b>	<b>LOHAS awareness</b>	<b>Decision</b>
Big	High	yes	High	yes
Med	Med	no	Med	no
Small	Low	yes	Low	no
Big	High	no	High	Yes
Small	med	yes	High	No
Med	high	yes	med	Yes
Med	med	yes	med	No
Big	med	no	med	No
Med	high	yes	low	No
Small	High	no	High	Yes
Small	med	no	High	No
Med	high	no	med	No

Use the decision tree to answer the following question?

<b>City-size</b>	<b>Avg Income</b>	<b>Local investors</b>	<b>LOHAS awareness</b>	<b>Decision</b>
Med	med	no	med	??

## Chapter 7: Regression

### Introduction

Regression is a well-known statistical technique to model the predictive relationship between several independent variables (DVs) and one dependent variable. The objective is to find the best-fitting curve for a dependent variable in a multidimensional space, with each independent variable being a dimension. The curve could be a straight line, or it could be a nonlinear curve. The quality of fit of the curve to the data can be measured by a coefficient of correlation ( $r$ ), which is the square root of the amount of variance explained by the curve.

The key steps for regression are simple:

1. List all the variables available for making the model.
2. Establish a Dependent Variable (DV) of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict DV using the other variables.

#### **Caselet: Data driven Prediction Markets**

*Traditional pollsters still seem to be using methodologies that worked well a decade or two ago. Nate Silver is a new breed of data-based political forecasters who are seeped in big data and advanced analytics. In the 2012 elections, he predicted that Obama would win the election with 291 electoral votes, compared to 247 for Mitt Romney, giving the President a 62% lead and re-election. He stunned the political forecasting world by correctly predicting the Presidential winner in all 50 states, including all nine swing states. He also, correctly predicted the winner in 31 of the 33 US Senate races.*

*Nate Silver brings a different view to the world of forecasting political elections, viewing it as a scientific discipline. State the hypothesis scientifically, gather all available information, analyze the data and extract insights using sophisticated models and algorithms and finally, apply human judgment to interpret those insights. The results are likely to be much more grounded and successful. (Source: The Signal and the Noise: Why Most Predictions Fail but Some Don't, by Nate Silver, 2012)*

*Q1: What is the impact of this story on traditional pollsters & commentators?*

## Correlations and Relationships

Statistical relationships are about which elements of data hang together, and which ones hang separately. It is about categorizing variables that have a relationship with one another, and categorizing variables that are distinct and unrelated to other variables. It is about describing significant positive relationships and significant negative differences.

The first and foremost measure of the strength of a relationship is co-relation (or correlation). The strength of a correlation is a quantitative measure that is measured in a normalized range between 0 (zero) and 1. A correlation of 1 indicates a perfect relationship, where the two variables are in perfect sync. A correlation of 0 indicates that there is no relationship between the variables.

The relationship can be positive, or it can be an inverse relationship, that is, the variables may move together in the same direction or in the opposite direction. Therefore, a good measure of correlation is the correlation coefficient, which is the square root of correlation. This coefficient, called  $r$ , can thus range from  $-1$  to  $+1$ . An  $r$  value of 0 signifies no relationship. An  $r$  value of 1 shows perfect relationship in the same direction, and an  $r$  value of  $-1$  shows a perfect relationship but moving in opposite directions.

Given two numeric variables  $x$  and  $y$ , the coefficient of correlation  $r$  is mathematically computed by the following equation.  $\bar{x}$  (called  $x$ -bar) is the mean of  $x$ , and  $\bar{y}$  ( $y$ -bar) is the mean of  $y$ .

$$r = \frac{[(x - \bar{x})(y - \bar{y})]}{\sqrt{[(x - \bar{x})^2][(y - \bar{y})^2]}}$$

## Visual look at relationships

A scatter plot (or scatter diagram) is a simple exercise for plotting all data points between two variables on a two-dimensional graph. It provides a visual layout of where all the data points are placed in that two-dimensional space. The scatter plot can be useful for graphically intuiting the relationship between two variables.

Here is a picture (Figure 7.1) that shows many possible patterns in scatter

diagrams.

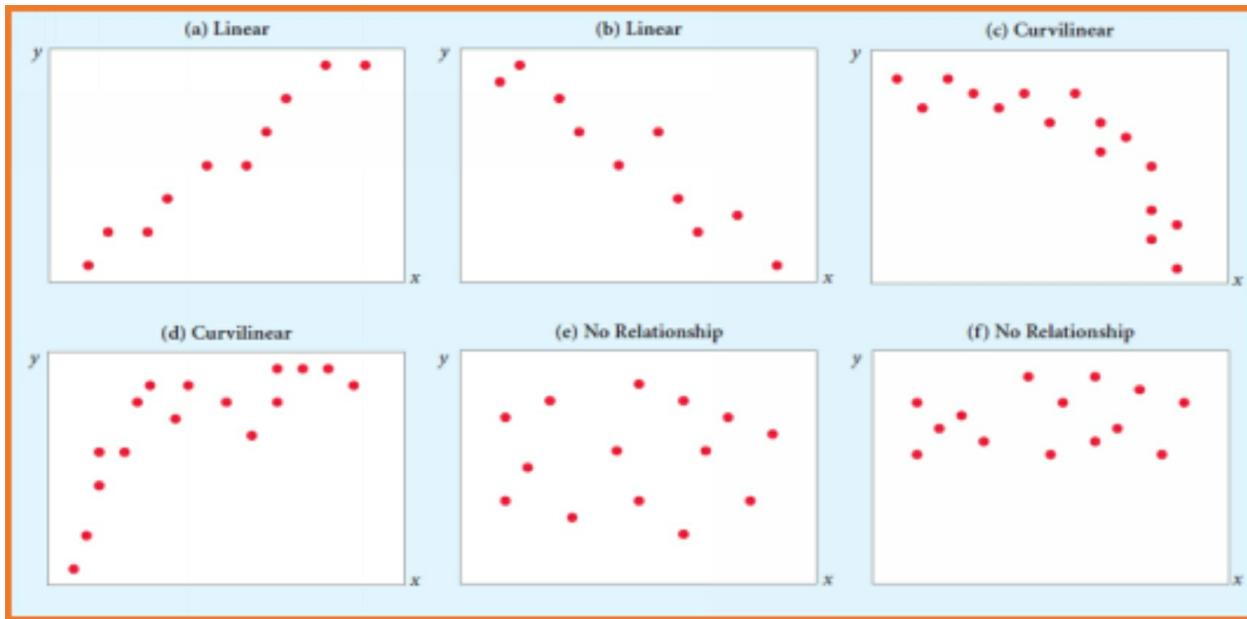


Figure 7.1: Scatter plots showing types of relationships among two variables  
(Source: Groebner et al. 2013)

Chart (a) shows a very strong linear relationship between the variables  $x$  and  $y$ . That means the value of  $y$  increases proportionally with  $x$ . Chart (b) also shows a strong linear relationship between the variables  $x$  and  $y$ . Here it is an inverse relationship. That means the value of  $y$  decreases proportionally with  $x$ .

Chart (c) shows a curvilinear relationship. It is an inverse relationship, which means that the value of  $y$  decreases proportionally with  $x$ . However, it seems a relatively well-defined relationship, like an arc of a circle, which can be represented by a simple quadratic equation (quadratic means the power of two, that is, using terms like  $x^2$  and  $y^2$ ). Chart (d) shows a positive curvilinear relationship. However, it does not seem to resemble a regular shape, and thus would not be a strong relationship. Charts (e) and (f) show no relationship. That means variables  $x$  and  $y$  are independent of each other.

Charts (a) and (b) are good candidates that model a simple linear regression model (the terms regression model and regression equation can be used interchangeably). Chart (c) too could be modeled with a little more complex, quadratic regression equation. Chart (d) might require an even higher order polynomial regression equation to represent the data. Charts (e) and (f) have

no relationship, thus, they cannot be modeled together, by regression or using any other modeling tool.

## Regression Exercise

The regression model is described as a linear equation that follows.  $y$  is the dependent variable, that is, the variable being predicted.  $x$  is the independent variable, or the predictor variable. There could be many predictor variables (such as  $x_1, x_2, \dots$ ) in a regression equation. However, there can be only one dependent variable ( $y$ ) in the regression equation.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where  $\beta_0$  and  $\beta_1$  are the constant, and the co-efficient for the  $x$  variable; and  $\varepsilon$  is the random error variable.

A simple example of a regression equation would be to predict a house price from the size of the house. Here is a sample house prices data:

House Price	Size (sqft)
\$229,500	1850
\$273,300	2190
\$247,000	2100
\$195,100	1930
\$261,000	2300
\$179,700	1710
\$168,500	1550
\$234,400	1920
\$168,800	1840
\$180,400	1720
\$156,200	1660
\$288,350	2405
\$186,750	1525
\$202,100	2030
\$256,800	2240

The two dimensions of (one predictor, one outcome variable) data can be plotted on a scatter diagram. A scatter plot with a best-fitting line looks like the graph that follows (Figure 7.2).

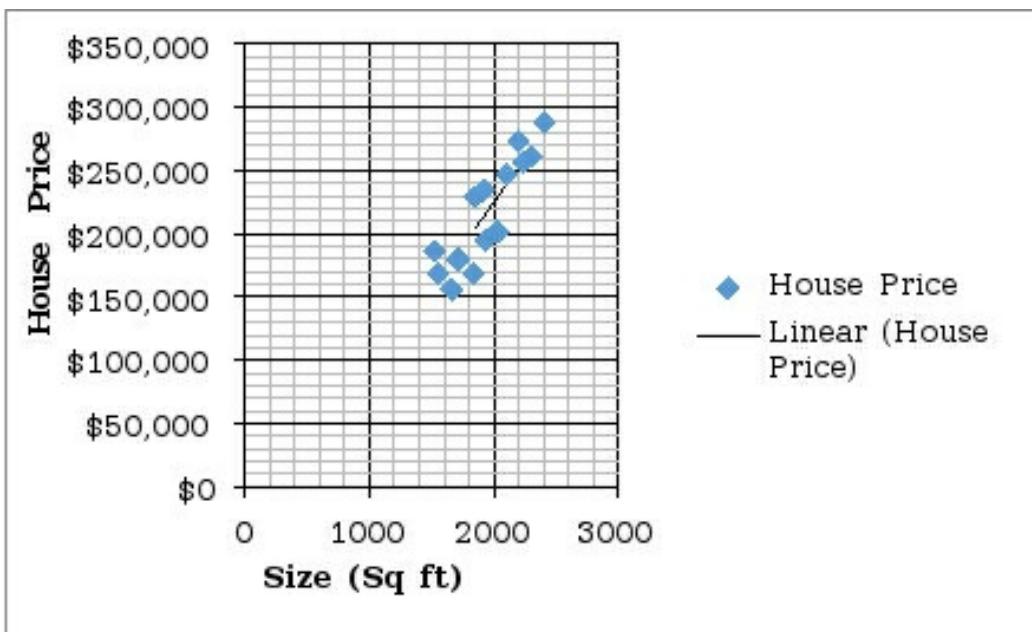


Figure 7.2: Scatter plot and regression equation between House price and house size.

Visually, one can see a positive correlation between House Price and Size (sqft). However, the relationship is not perfect. Running a regression model between the two variables produces the following output (truncated).

Regression Statistics	
r	<b>0.891</b>
r <sup>2</sup>	<b>0.794</b>
Coefficients	
Intercept	<b>-54191</b>
Size (sqft)	<b>139.48</b>

It shows the coefficient of correlation is 0.891. r<sup>2</sup>, the measure of total variance explained by the equation, is 0.794, or 79%. That means the two variables are moderately and positively correlated. Regression coefficients help create the following equation for predicting house prices.

$$\text{House Price (\$)} = 139.48 * \text{Size(sqft)} - 54191$$

This equation explains only 79% of the variance in house prices. Suppose

other predictor variables are made available, such as the number of rooms in the house. It might help improve the regression model.

The house data now looks like this:

House Price	Size (sqft)	#Rooms
\$229,500	1850	4
\$273,300	2190	5
\$247,000	2100	4
\$195,100	1930	3
\$261,000	2300	4
\$179,700	1710	2
\$168,500	1550	2
\$234,400	1920	4
\$168,800	1840	2
\$180,400	1720	2
\$156,200	1660	2
\$288,350	2405	5
\$186,750	1525	3
\$202,100	2030	2
\$256,800	2240	4

While it is possible to make a 3-dimensional scatter plot, one can alternatively examine the correlation matrix among the variables.

	<i>House</i>		
	<i>Price</i>	<i>Size (sqft)</i>	<i>#Rooms</i>
House			
Price	1		
Size (sqft)	0.891	1	
Rooms	0.944	0.748	1

It shows that the House price has a strong correlation with number of rooms (0.944) as well. Thus, it is likely that adding this variable to the regression model will add to the strength of the model.

Running a regression model between these three variables produces the following output (truncated).

<i>Regression Statistics</i>	
<b>r</b>	<b>0.984</b>
<b>r<sup>2</sup></b>	<b>0.968</b>

	<i>Coefficients</i>
<b>Intercept</b>	<b>12923</b>
<b>Size(sqft)</b>	<b>65.60</b>
<b>Rooms</b>	<b>23613</b>

It shows the co-efficient of correlation of this regression model is 0.984.  $R^2$ , the total variance explained by the equation, is 0.968 or 97%. That means the variables are positively and very strongly correlated. Adding a new relevant variable has helped improve the strength of the regression model.

Using the regression coefficients helps create the following equation for predicting house prices.

$$\text{House Price (\$)} = 65.6 * \text{Size (sqft)} + 23613 * \text{Rooms} + 12924$$

This equation shows a 97% goodness of fit with the data, which is very good for business and economic data. There is always some random variation in naturally occurring business data, and it is not desirable to overfit the model to the data.

This predictive equation should be used for future transactions. Given a situation as below, it will be possible to predict the price of the house with 2000 sq ft and 3 rooms.

House Price	Size (sqft)	#Rooms
??	2000	3

$$\text{House Price (\$)} = 65.6 * 2000 (\text{sqft}) + 23613 * 3 + 12924 = \$214,963$$

The predicted values should be compared with the actual values to see how close the model is able to predict the actual value. As new data points become available, there are opportunities to fine-tune and improve the model.

### Non-linear regression exercise

The relationship between the variables may also be curvilinear. For example, given past data from electricity consumption (KwH) and temperature (temp), the objective is to predict the electrical consumption from the temperature value. Here are a few past observations.

KWatts	Temp (F)
12530	46.8

10800	52.1
10180	55.1
9730	59.2
9750	61.9
10230	66.2
11160	69.9
13910	76.8
15690	79.3
15110	79.7
17020	80.2
17880	83.3

In two dimensions (one predictor, one outcome variable) data can be plotted on a scatter diagram. A scatter plot with a best-fitting line looks like the graph below (Figure 7.3).

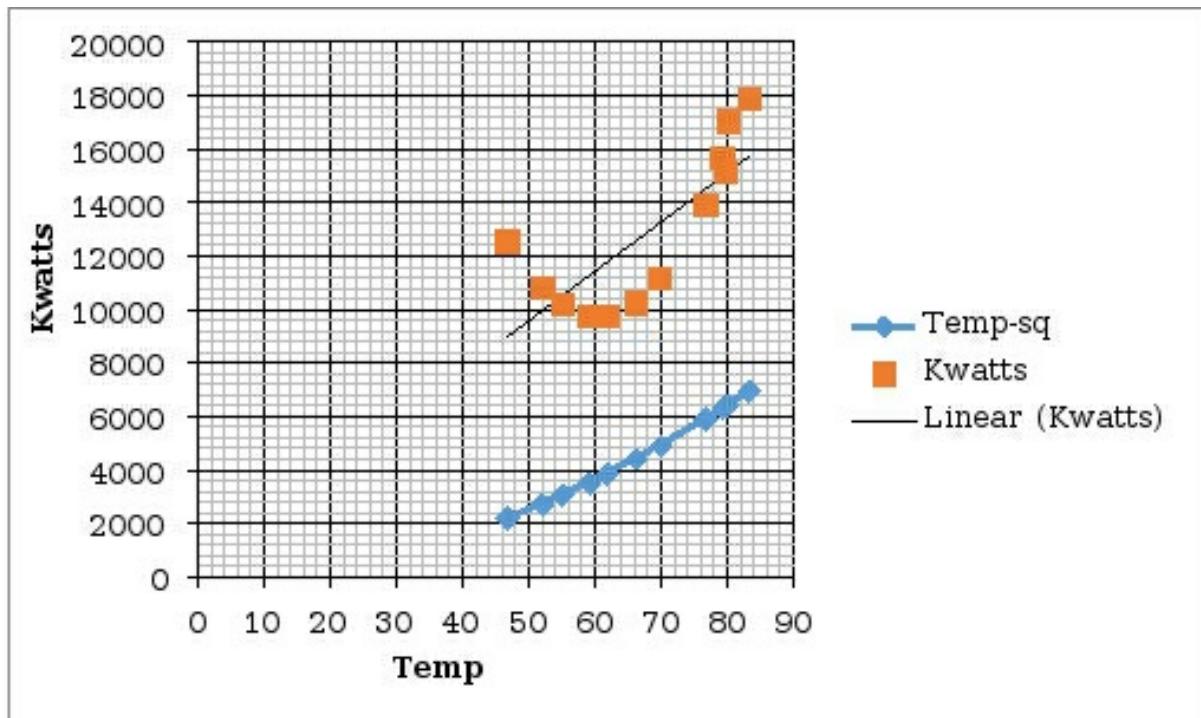


Figure 6.2: Scatter plots showing regression between (a) kwatts and temp, and (b) kwatts and temp square

It is visually clear that the straight line does not fit the data well. The relationship between temperature and Kwatts follows a curvilinear model, where it hits bottom at a certain value of temperature. The regression model confirms the relationship since R is only 0.77 and R-square is also only 60%.

Thus, only 60% of the variance is explained.

This regression model can be enhanced by introducing a non-linear variable (such as a quadratic variable Temp<sup>2</sup>) in the equation. The second line is the relationship between KWH and Temp<sup>2</sup>. The scatter plot shows that the Energy consumption shows a strong linear relationship with Temp<sup>2</sup>. Computing the regression model after adding the Temp<sup>2</sup> variable, leads to the following results:

Regression Statistics	
r	0.992
r <sup>2</sup>	0.984
Coefficients	
Intercept	67245
Temp (F)	-1911
Temp-sq	15.87

It shows that the co-efficient of correlation of the regression model is now 0.99. R<sup>2</sup>, the total variance explained by the equation is 0.985, or 98.5%. That means the variables are very strongly and positively correlated. The regression coefficients help create the following equation for

$$\text{Energy Consumption (Kwatts)} = 15.87 * \text{Temp}^2 - 1911 * \text{Temp} + 67245$$

This equation shows a 98.5% fit which is very good for business and economic contexts. Now one can predict the Kwatts value for when the temperature is 72-degrees.

$$\text{Energy consumption} = (15.87 * 72*72) - (1911 * 72) + 67245 = 11923 \text{ Kwatts}$$

## Logistic Regression

Regression models traditionally work with continuous numeric value data for dependent and independent variables. Logistic regression models can, however, work with dependent variables that have categorical values, such as whether a loan is approved (yes or no). Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables. For example, Logistic regression might be used to predict whether a patient has a given disease (e.g. [diabetes](#)), based on

observed characteristics of the patient (age, gender, [body mass index](#), results of [blood tests](#), etc.).

Logistical regression models use probability scores as the predicted values of the dependent variable. Logistic regression takes the [natural logarithm](#) of the probability of the dependent variable being a case (referred to as the [logit](#) function), and creates a continuous criterion as a transformed version of the dependent variable. Thus the logit transformation is used in logistic regression as the dependent variable. The net effect is that although the dependent variable in logistic regression is binomial (or categorical, i.e. has only two possible values), the logit is the continuous function upon which linear regression is conducted. Here is the general logistic function, with independent variable on the horizontal axis and the logit dependent variable on the vertical axis (Figure 7.3).

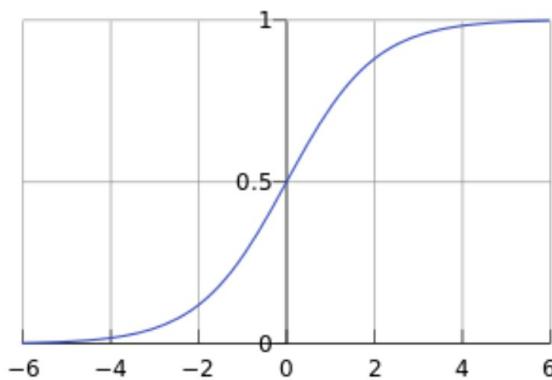


Figure 7.3: General Logit function

All popular data mining platforms provide support for regular multiple regression models, as well as options for Logistic Regression.

## Advantages and Disadvantages of Regression Models

Regression Models are very popular because they offer many advantages.

1. Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error.
2. Regression models provide simple algebraic equations that are easy to understand and use.
3. The strength (or the goodness of fit) of the regression model is measured in terms of the correlation coefficients, and other

related statistical parameters that are well understood.

4. Regression models can match and beat the predictive power of other modeling techniques.
5. Regression models can include all the variables that one wants to include in the model.
6. Regression modeling tools are pervasive. They are found in statistical packages as well as data mining packages. MS Excel spreadsheets can also provide simple regression modeling capabilities.

Regression models can however prove inadequate under many circumstances.

1. Regression models can not cover for poor data quality issues. If the data is not prepared well to remove missing values, or is not well-behaved in terms of a normal distribution, the validity of the model suffers.
2. Regression models suffer from collinearity problems (meaning strong linear correlations among some independent variables). If the independent variables have strong correlations among themselves, then they will eat into each other's predictive power and the regression coefficients will lose their ruggedness.  
Regression models will not automatically choose between highly collinear variables, although some packages attempt to do that.
3. Regression models can be unwieldy and unreliable if a large number of variables are included in the model. All variables entered into the model will be reflected in the regression equation, irrespective of their contribution to the predictive power of the model. There is no concept of automatic pruning of the regression model.
4. Regression models do not automatically take care of non-linearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit.
5. Regression models work only with numeric data and not with categorical variables. There are ways to deal with categorical variables though by creating multiple new variables with a yes/no value.

## Conclusion

Regression models are simple, versatile, visual/graphical tools with high predictive ability. They include non-linear as well as binary predictions. Regression models should be used in conjunction with other data mining techniques to confirm the findings.

\*\*\*

### Review Exercises:

Q1: What is a regression model?

Q2: What is a scatter plot? How does it help?

Q3: Compare and contrast decision trees with regression models?

Q4: Using the data below, create a regression model to predict the Test2 from the Test1 score. Then predict the score for one who got a 46 in Test1.

Test1	Test2
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	50

### Liberty Stores Case Exercise: Step 6

Liberty wants to forecast its sales for next year, for financial budgeting.

Year	Global GDP index per capita	# cust serv calls('000s)	# employees ('000)	# Items ('000)	Revenue (\$M)
1	100	25	45	11	2000

2	112	27	53	11	2400
3	115	22	54	12	2700
4	123	27	58	14	2900
5	122	32	60	14	3200
6	132	33	65	15	3500
7	143	40	72	16	4000
8	126	30	65	16	4200
9	166	34	85	17	4500
10	157	47	97	18	4700
11	176	33	98	18	4900
12	180	45	100	20	5000

Check the correlations. Which variables are strongly correlated?

Create a regression model that best predicts the revenue.

## Chapter 8: Artificial Neural Networks

### Introduction

Artificial Neural Networks (ANN) are inspired by the information processing model of the mind/brain. The human brain consists of billions of neurons that link with one another in an intricate pattern. Every neuron receives information from many other neurons, processes it, gets excited or not, and passes its state information to other neurons.

Just like the brain is a multipurpose system, so also the ANNs are very versatile systems. They can be used for many kinds of pattern recognition and prediction. They are also used for classification, regression, clustering, association, and optimization activities. They are used in finance, marketing, manufacturing, operations, information systems applications, and so on.

ANNs are composed of a large number of highly interconnected processing elements (neurons) working in a multi-layered structures that receive inputs, process the inputs, and produce an output. An ANN is designed for a specific application, such as pattern recognition or data classification, and trained through a learning process. Just like in biological systems, ANNs make adjustments to the synaptic connections with each learning instance.

ANNs are like a black box trained into solving a particular type of problem, and they can develop high predictive powers. Their intermediate synaptic parameter values evolve as the system obtains feedback on its predictions, and thus an ANN learns from more training data (Figure 8.1).



Figure 8.1: General ANN model

#### **Caselet: IBM Watson - Analytics in Medicine**

*The amount of medical information available is doubling every five years and much of this data is unstructured. Physicians simply don't have time to read every journal that can help them keep up to date with the latest advances. Mistakes in diagnosis are likely to happen*

*and clients have become more aware of the evidence. Analytics will transform the field of medicine into Evidence-based medicine. How can healthcare providers address these problems?*

*IBM's Watson cognitive computing system can analyze large amounts of unstructured text and develop hypotheses based on that analysis. Physicians can use Watson to assist in diagnosing and treating patients. First, the physician might describe symptoms and other related factors to the system. Watson can then identify the key pieces of information and mine the patient's data to find relevant facts about family history, current medications and other existing conditions. It combines this information with current findings from tests, and then forms and tests a hypotheses by examining a variety of data sources—treatment guidelines, electronic medical record data and doctors' and nurses' notes, as well as peer-reviewed research and clinical studies. From here, Watson can provide potential treatment options and its confidence rating for each suggestion.*

*Watson has been deployed at many leading healthcare institutions to improve the quality and efficiency of healthcare decisions; to help clinicians uncover insights from its patient information in electronic medical records (EMR); among other benefits.*

*Q1: How would IBM Watson change medical practices in the future?*

*Q2: In what other industries & functions could this technology be applied?*

## **Business Applications of ANN**

Neural networks are used most often when the objective function is complex, and where there exists plenty of data, and the model is expected to improve over a period of time. A few sample applications:

1. They are used in stock price prediction where the rules of the game are extremely complicated, and a lot of data needs to be processed very quickly.
2. They are used for character recognition, as in recognizing hand-written text, or damaged or mangled text. They are used in recognizing finger prints. These are complicated patterns and are unique for each person. Layers of neurons can progressively clarify the pattern leading to a remarkably accurate result.
3. They are also used in traditional classification problems, like

approving a financial loan application.

## Design Principles of an Artificial Neural Network

1. A neuron is the basic processing unit of the network. The neuron (or processing element) receives inputs from its preceding neurons (or PEs), does some nonlinear weighted computation on the basis of those inputs, transforms the result into its output value, and then passes on the output to the next neuron in the network (Figure 8.2). X's are the inputs, w's are the weights for each input, and y is the output.

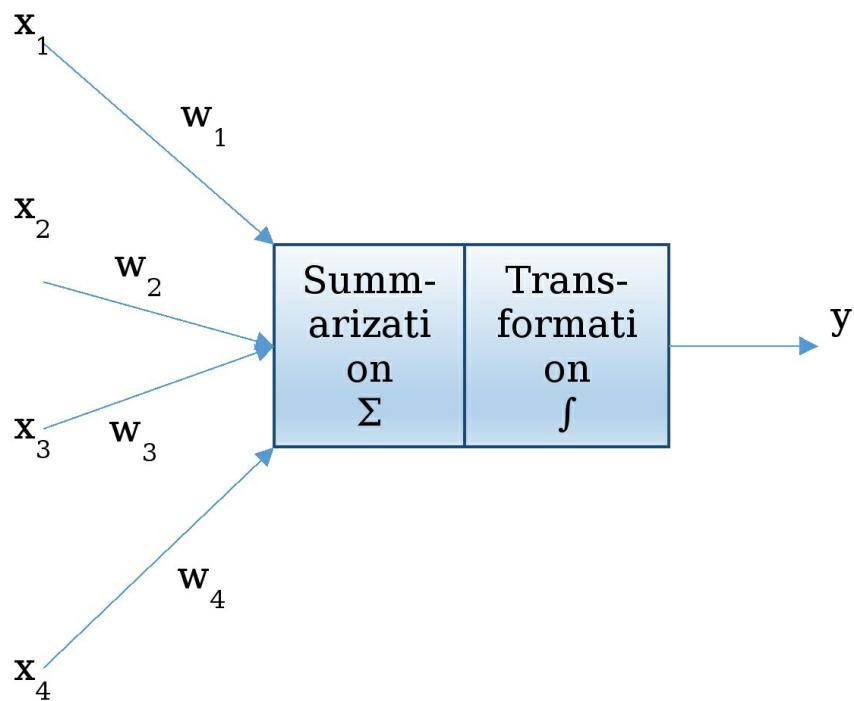


Figure 8.2: Model for a single artificial neuron

2. A Neural network is a multi-layered model. There is at least one input neuron, one output neuron, and at least one processing neuron. An ANN with just this basic structure would be a simple, single-stage computational unit. A simple task may be processed by just that one neuron and the result may be communicated soon. ANNs however, may have multiple layers of processing elements

in sequence. There could be many neurons involved in a sequence depending upon the complexity of the predictive action. The layers of PEs could work in sequence, or they could work in parallel (Figure 8.3).

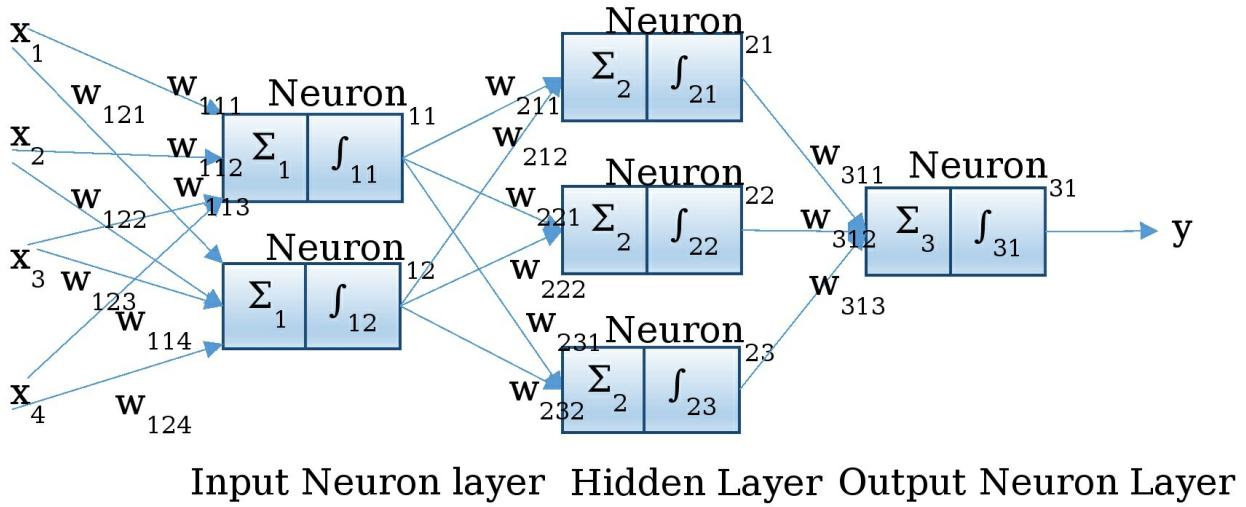


Figure 8.3: Model for a multi-layer ANN

3. The processing logic of each neuron may assign different weights to the various incoming input streams. The processing logic may also use non-linear transformation, such as a sigmoid function, from the processed values to the output value. This processing logic and the intermediate weight and processing functions are just what works for the system as a whole, in its objective of solving a problem collectively. Thus, neural networks are considered to be an opaque and a black-box system.
4. The neural network can be trained by making similar decisions over and over again with many training cases. It will continue to learn by adjusting its internal computation and communication based on feedback about its previous decisions. Thus, the neural networks become better at making a decision as they handle more and more decisions.

Depending upon the nature of the problem and the availability of good training data, at some point the neural network will learn enough and begin to match the predictive accuracy of a human expert. In many practical situations, the predictions of ANN, trained over a long period of time with a

large number of training data, have begun to decisively become more accurate than human experts. At that point ANN can begin to be seriously considered for deployment in real situations in real time.

## Representation of a Neural Network

A neural network is a series of neurons that receive inputs from other neurons. They do a weighted summation function of all the inputs, using different weights (or importance) for each input. The weighted sum is then transformed into an output value using a transfer function.

Learning in ANN occurs when the various processing elements in the neural network adjust the underlying relationship (weights, transfer function, etc) between input and outputs, in response to the feedback on their predictions. If the prediction made was correct, then the weights would remain the same, but if the prediction was incorrect, then the parameter values would change.

The Transformation (Transfer) Function is any function suitable for the task at hand. The transfer function for ANNs is usually a non-linear sigmoid function. Thus, if the normalized computed value is less than some value (say 0.5) then the output value will be zero. If the computed value is at the cut-off threshold, then the output value will be a 1. It could be a nonlinear hyperbolic function in which the output is either a -1 or a 1. Many other functions could be designed for any or all of the processing elements.

Thus, in a neural network, every processing element can potentially have a different number of input values, a different set of weights for those inputs, and a different transformation function. Those values support and compensate for one another until the neural network as a whole learns to provide the correct output, as desired by the user.

## Architecting a Neural Network

There are many ways to architect the functioning of an ANN using fairly simple and open rules with a tremendous amount of flexibility at each stage. The most popular architecture is a Feed-forward, multi-layered perceptron with back-propagation learning algorithm. That means there are multiple layers of PEs in the system and the output of neurons are fed forward to the PEs in the next layers; and the feedback on the prediction is fed back into the neural network for learning to occur. This is essentially what was described in the earlier paragraphs. ANN architectures for different applications are shown in Table 8.1.

Classification	Feedforward networks (MLP), radial basis function, and probabilistic
Regression	Feedforward networks (MLP), radial basis function
Clustering	Adaptive resonance theory (ART), Self-organizing maps (SOMs)
Association Rule Mining	Hopfield networks

**Table 8.1: ANN architectures for different applications**

## Developing an ANN

It takes resources, training data, skill and time to develop a neural network. Most data mining platforms offer at least the Multi-Layer-Perceptron (MLP) algorithm to implement a neural network. Other neural network architectures include Probabilistic networks and Self-organizing feature maps.

The steps required to build an ANN are as follows:

1. Gather data. Divide into training data and test data. The training data needs to be further divided into training data and validation data.
2. Select the network architecture, such as Feedforward network.
3. Select the algorithm, such as Multi-layer Perception.
4. Set network parameters.
5. Train the ANN with training data.
6. Validate the model with validation data.
7. Freeze the weights and other parameters.
8. Test the trained network with test data.
9. Deploy the ANN when it achieves good predictive accuracy.

Training an ANN requires that the training data be split into three parts (Table 8.2):

<b>Training set</b>	This data set is used to adjust the weights on the neural network (~ 60%).
<b>Validation set</b>	This data set is used to minimize overfitting and verifying accuracy (~ 20%).

<b>Testing set</b>	This data set is used only for testing the final solution in order to confirm the actual predictive power of the network (~ 20%).
<b>k-fold cross-validation</b>	This approach means that the data is divided into k equal pieces, and the learning process is repeated k-times with each piece becoming the training set. This process leads to less bias and more accuracy, but is more time consuming.

**Table 8.2: ANN Training datasets**

## Advantages and Disadvantages of using ANNs

There are many benefits of using ANN.

1. ANNs impose very little restrictions on their use. ANN can deal with (identify/model) highly nonlinear relationships on their own, without much work from the user or analyst. They help find practical data-driven solutions where algorithmic solutions are non-existent or too complicated.
2. There is no need to program neural networks, as they learn from examples. They get better with use, without much programming effort.
3. They can handle a variety of problem types, including classification, clustering, associations, etc.
4. ANN are tolerant of data quality issues and they do not restrict the data to follow strict normality and/or independence assumptions.
5. They can handle both numerical and categorical variables.
6. ANNs can be much faster than other techniques.
7. Most importantly, they usually provide better results (prediction and/or clustering) compared to statistical counterparts, once they have been trained enough.

The key disadvantages arise from the fact that they are not easy to interpret or explain or compute.

1. They are deemed to be black-box solutions, lacking explainability. Thus they are difficult to communicate about,

except through the strength of their results.

2. Optimal design of ANN is still an art: it requires expertise and extensive experimentation.
3. It can be difficult to handle a large number of variables (especially the rich nominal attributes).
4. It takes large data sets to train an ANN.

## Conclusion

Artificial neural networks are complex systems that mirror the functioning of the human brain. They are versatile enough to solve many data mining tasks with high accuracy. However, they are like black boxes and they provide little guidance on the intuitive logic behind their predictions.

## Review Exercises

- 1: What is a neural network? How does it work?
- 2: Compare a neural network with a decision tree.
- 3: What makes a neural network versatile enough for supervised as well as non-supervised learning tasks?
- 4: Examine the steps in developing a neural network for predicting stock prices. What kind of objective function and what kind of data would be required for a good stock price predictor system using ANN?

# Chapter 9: Cluster Analysis

## Introduction

Cluster analysis is used for automatic identification of natural groupings of things. It is also known as the segmentation technique. In this technique, data instances that are similar to (or near) each other are categorized into one cluster. Similarly, data instances that are very different (or far away) from each other are moved into different clusters.

Clustering is an unsupervised learning technique as there is no output or dependent variable for which a right or wrong answer can be computed. The correct number of clusters or the definition of those clusters is not known ahead of time. Clustering techniques can only suggest to the user how many clusters would make sense from the characteristics of the data. The user can specify a different, larger or smaller, number of desired clusters based on their making business sense. The cluster analysis technique will then define many distinct clusters from analysis of the data, with cluster definitions for each of those clusters. However, there are good cluster definitions, depending on how closely the cluster parameters fit the data.

### Caselet: Cluster Analysis

*A national insurance company distributes its personal and small commercial insurance products through independent agents. They wanted to increase their sales by better understanding their customers. They were interested in increasing their market share by doing some direct marketing campaigns, however without creating a channel conflict with the independent agents. They were also interested in examining different customer segments based on their needs, and the profitability of each of those segments.*

*They gathered attitudinal, behavioral, and demographic data using a mail survey of 2000 U.S. households that own auto insurance. Additional geo-demographic and credit information was added to the survey data. Cluster analysis of the data revealed five roughly equal segments:*

- *Non-Traditionals: interested in using the Internet and/or buying insurance at work.*
- *Direct Buyers: interested in buying via direct mail or telephone.*
- *Budget Conscious: interested in minimal coverage and finding the best deal.*

- *Agent Loyals: expressed strong loyalty to their agents and high levels of personal service.*
- *Hassle-Free: similar to Agent Loyals but less interested in face-to-face service.*

(Source: greenbook.org)

*Q1. Which customer segments would you choose for direct marketing? Will these create a channel conflict?*

*Q2. Could this segmentation apply to other service businesses? Which ones?*

## Applications of Cluster Analysis

Cluster analysis is used in almost every field where there is a large variety of transactions. It helps provide characterization, definition, and labels for populations. It can help identify natural groupings of customers, products, patients, and so on. It can also help identify outliers in a specific domain and thus decrease the size and complexity of problems. A prominent business application of cluster analysis is in market research. Customers are segmented into clusters based on their characteristics—wants and needs, geography, price sensitivity, and so on. Here are some examples of clustering:

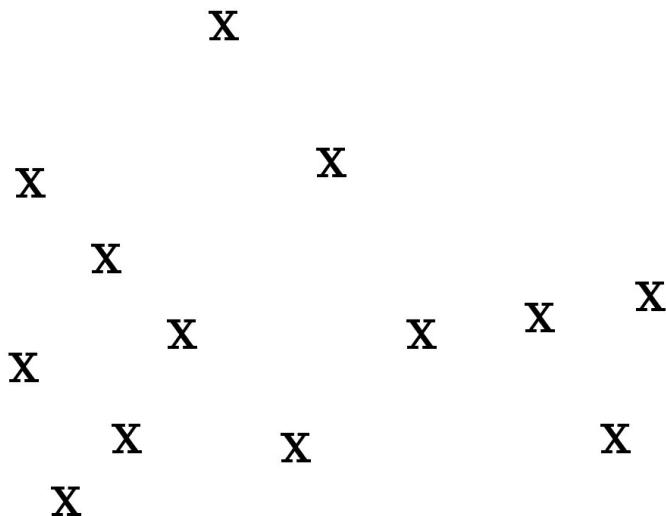
1. *Market Segmentation:* Categorizing customers according to their similarities, for instance by their common wants and needs, and propensity to pay, can help with targeted marketing.
2. *Product portfolio:* People of similar sizes can be grouped together to make small, medium and large sizes for clothing items.
3. *Text Mining:* Clustering can help organize a given collection of text documents according to their content similarities into clusters of related topics.

## Definition of a Cluster

An operational definition of a cluster is that, given a representation of  $n$  objects, find  $K$  groups based on a measure of similarity, such that objects within the same group are alike but the objects in different groups are not alike.

However, the notion of similarity can be interpreted in many ways. Clusters can differ in terms of their shape, size, and density. Clusters are patterns, and there can be many kinds of patterns. Some clusters are the traditional types, such as data points hanging together. However, there are other clusters, such as all points representing the circumference of a circle. There may be concentric circles with points of different circles representing different clusters. The presence of noise in the data makes the detection of the clusters even more difficult.

An ideal cluster can be defined as a set of points that is compact and isolated. In reality, a cluster is a subjective entity whose significance and interpretation requires domain knowledge. In the sample data below (Figure 9.1), how many clusters can one visualize?



**Figure 9.1: Visual cluster example**

It seems like there are two clusters of approximately equal sizes. However, they can be seen as three clusters, depending on how we draw the dividing lines. There is not a truly optimal way to calculate it. Heuristics are often used to define the number of clusters.

## Representing clusters

The clusters can be represented by a central or modal value. A cluster can be defined as the *centroid* of the collection of points belonging to it. A *centroid* is a measure of central tendency. It is the point from where the sum total of squared distance from all the points is the minimum. A real-life equivalent

would be the city center as the point that is considered the most easy to use by all constituents of the city. Thus all cities are defined by their centers or downtown areas.

A cluster can also be represented by the most frequently occurring value in the cluster, i.e. the cluster can be defined by its modal value. Thus, a particular cluster representing a social point of view could be called the ‘soccer moms’, even though not all members of that cluster need currently be a mom with soccer-playing children.

## Clustering techniques

Cluster analysis is a machine-learning technique. The quality of a clustering result depends on the *algorithm*, the *distance* function, and the *application*. First, consider the distance function. Most cluster analysis methods use a distance measure to calculate the closeness between pairs of items. There are two major measures of distances: Euclidian distance (“as the crow flies” or straight line) is the most intuitive measure. The other popular measure is the Manhattan (rectilinear) distance, where one can go only in orthogonal directions. The Euclidian distance is the hypotenuse of a right triangle, while the Manhattan distance is the sum of the two legs of the right triangle. There are other measures of distance like Jaccard distance (to measure similarity of sets), or Edit distance (similarity of texts), and others.

In either case, the key objective of the clustering algorithm is the same:

- Inter-clusters distance  $\rightarrow$  maximized; and
- Intra-clusters distance  $\rightarrow$  minimized

There are many algorithms to produce clusters. There are top-down, hierarchical methods that start with creating a given number of best-fitting clusters. There are also bottom-up methods that begin with identifying naturally occurring clusters.

The most popular clustering algorithm is the K-means algorithm. It is a top-down, statistical technique, based on the method of minimizing the least squared distance from the center points of the clusters. Other techniques, such as neural networks, are also used for clustering. Comparing cluster algorithms is a difficult task as there is no single right number of clusters. However, the speed of the algorithm and its versatility in terms of different dataset are important criteria.

*Here is the generic pseudocode for clustering*

1. *Pick an arbitrary number of groups/segments to be created*
2. *Start with some initial randomly-chosen center values for groups*
3. *Classify instances to closest groups*
4. *Compute new values for the group centers*
5. *Repeat step 3 & 4 till groups converge*
6. *If clusters are not satisfactory, go to step 1 and pick a different number of groups/segments*

The clustering exercise can be continued with a different number of clusters and different location of those points. Clusters are considered good if the cluster definitions stabilize, and the stabilized definitions prove useful for the purpose at hand. Else, repeat the clustering exercise with a different number of clusters, and different starting points for group means.

## **Clustering Exercise**

Here is a simple exercise to visually and intuitive identify clusters from data. X and Y are two dimensions of interest. The objective is to determine the number of clusters, and the center points of those clusters.

X	Y
2	4
2	6
5	6
4	7
8	3
6	6
5	2
5	7
6	3
4	4

A scatter plot of 10 items in 2 dimensions shows them distributed fairly

randomly. As a bottom-up technique, the number of clusters and their centroids can be intuited (Figure 9.2).

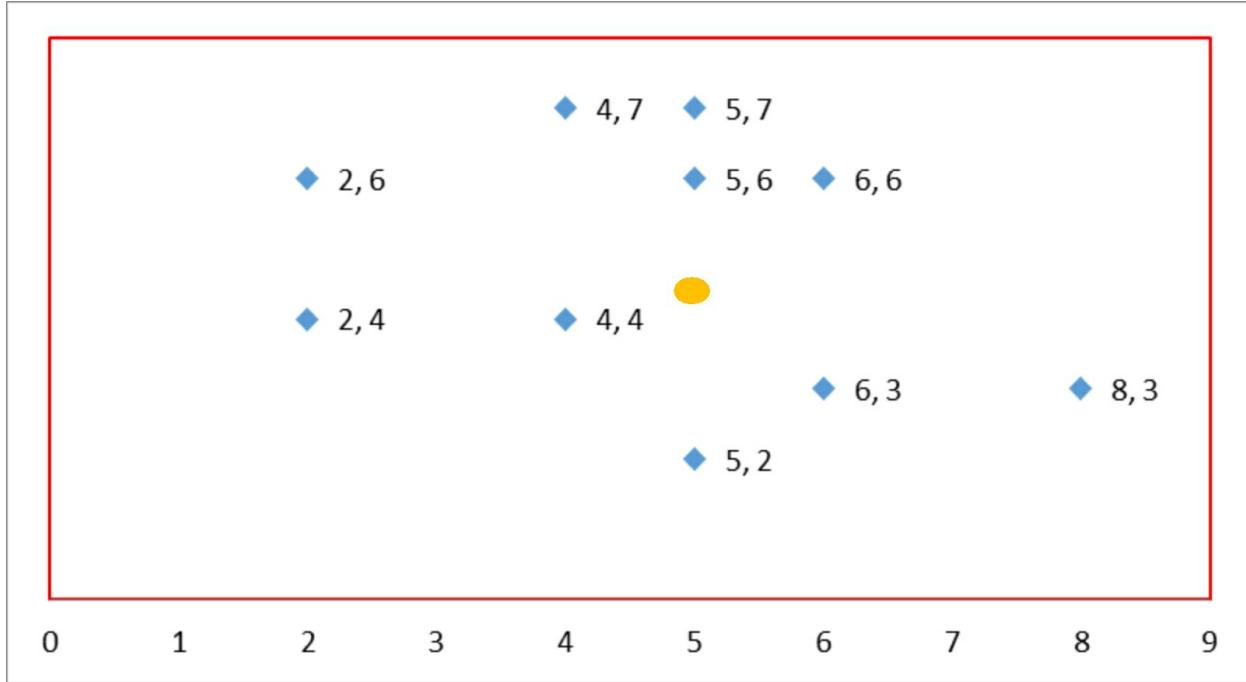


Figure 9.2: Initial data points and the centroid (shown as thick dot)

The points are distributed randomly enough that it could be considered one cluster. The solid circle would represent the central point (centroid) of these points.

However, there is a big distance between the points (2,6) and (8,3). So, this data could be broken into 2 clusters. The three points at the bottom right could form one cluster and the other seven could form the other cluster. The two clusters would look like this (Figure 9.3). The two circles will be the new centroids.

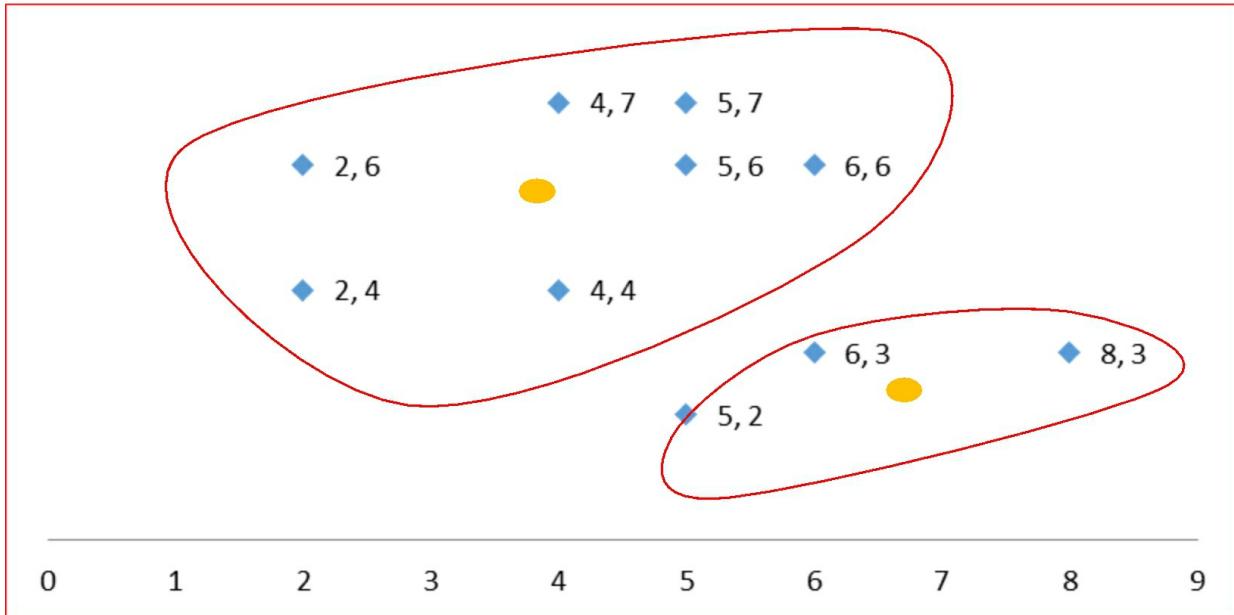


Figure 9.3: Dividing into two clusters (centroids shown as thick dots)

The bigger cluster seems too far apart. So, it seems like the 4 points on the top will form a separate cluster. The three clusters could look like this (Figure 9.4).

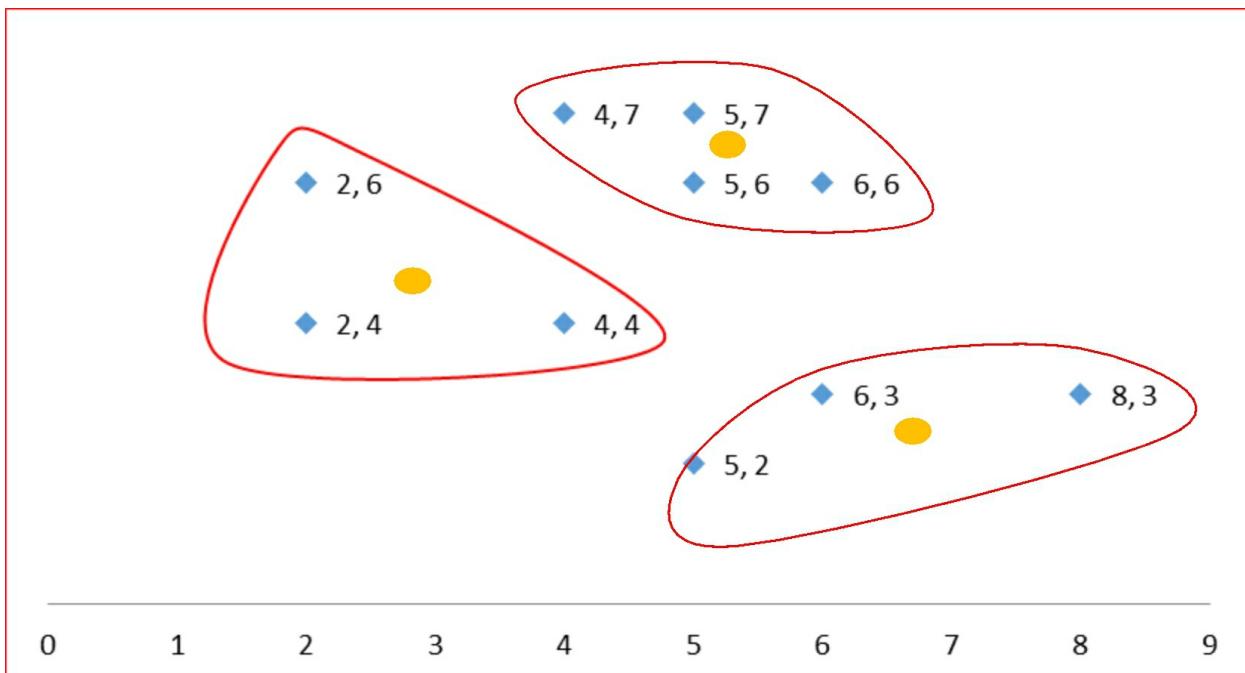


Figure 9.4: Dividing into three clusters (centroids shown as thick dots)

This solution has three clusters. The cluster on the right is far from the other two clusters. However, its centroid is not too close to all the data points. The

cluster at the top looks very tight-fitting, with a nice centroid. The third cluster, at the left, is spread out and may not be of much usefulness.

This was a bottom-up exercise in visually producing three best-fitting cluster definitions from the given data. The right number of clusters will depend on the data and the application for which the data would be used.

## K-Means Algorithm for clustering

K-means is the most popular clustering algorithm. It iteratively computes the clusters and their centroids. It is a top down approach to clustering. Starting with a given number of K clusters, say 3 clusters. Thus three random centroids will be created as starting points of the centers of three clusters. The circles are initial cluster centroids (Figure 9.5).

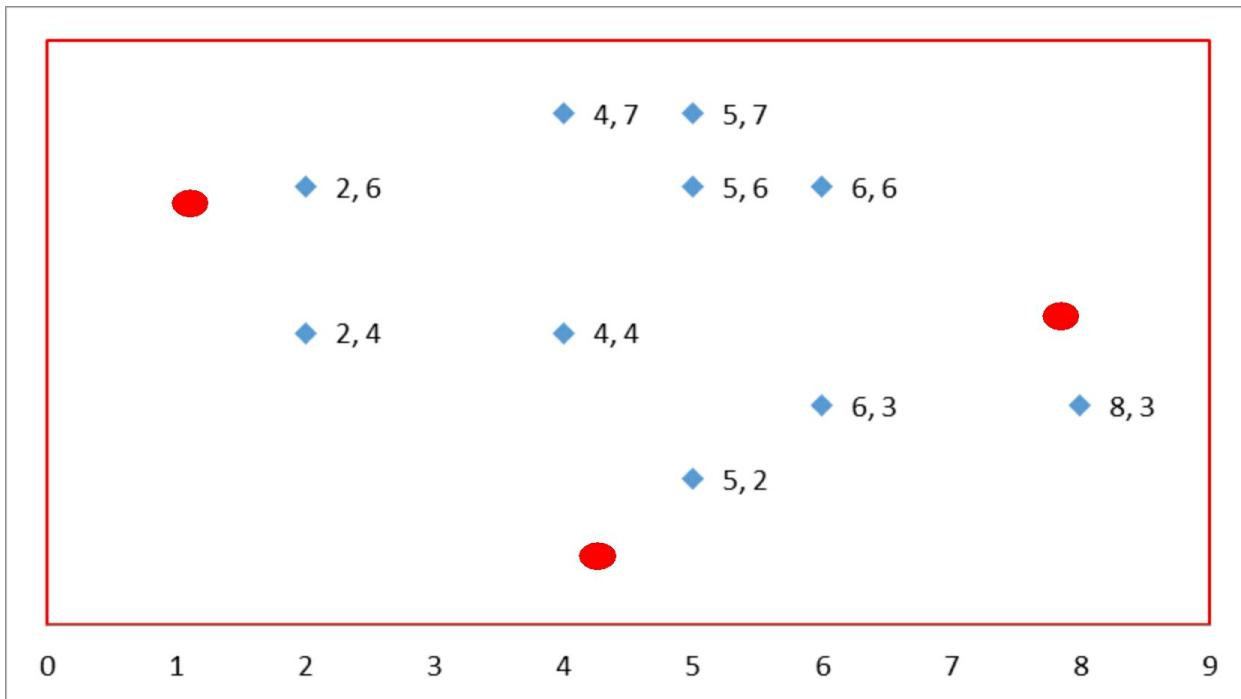


Figure 9.5: Randomly assigning three centroids for three data clusters

*Step 1:* For a data point, distance values will be from each of the three centroids. The data point will be assigned to the cluster with the shortest distance to the centroid. All data points will thus, be assigned to one data point or the other (Figure 9.6). The arrows from each data element shows the centroid that the point is assigned to.

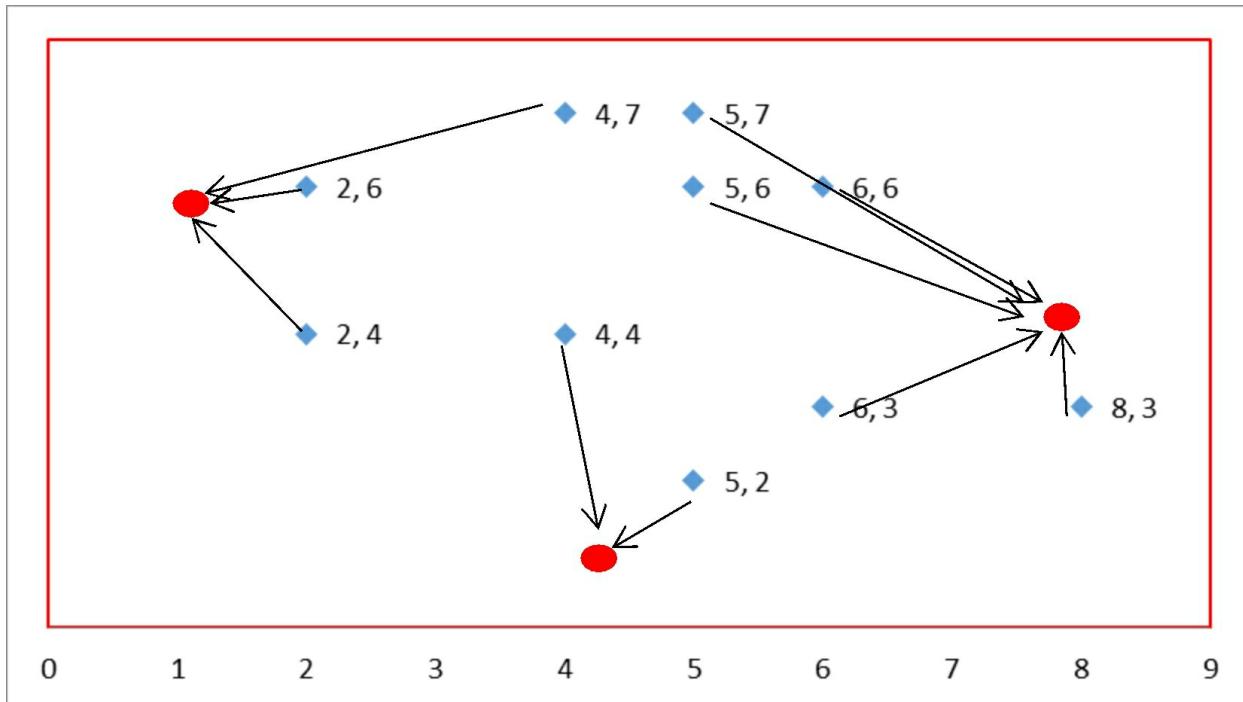


Figure 9.6: Assigning data points to closest centroid

*Step 2:* The centroid for each cluster will now be recalculated such that it is closest to all the data points allocated to that cluster. The dashed arrows show the centroids being moved from their old (shaded) values to the revised new values (Figure 9.7).

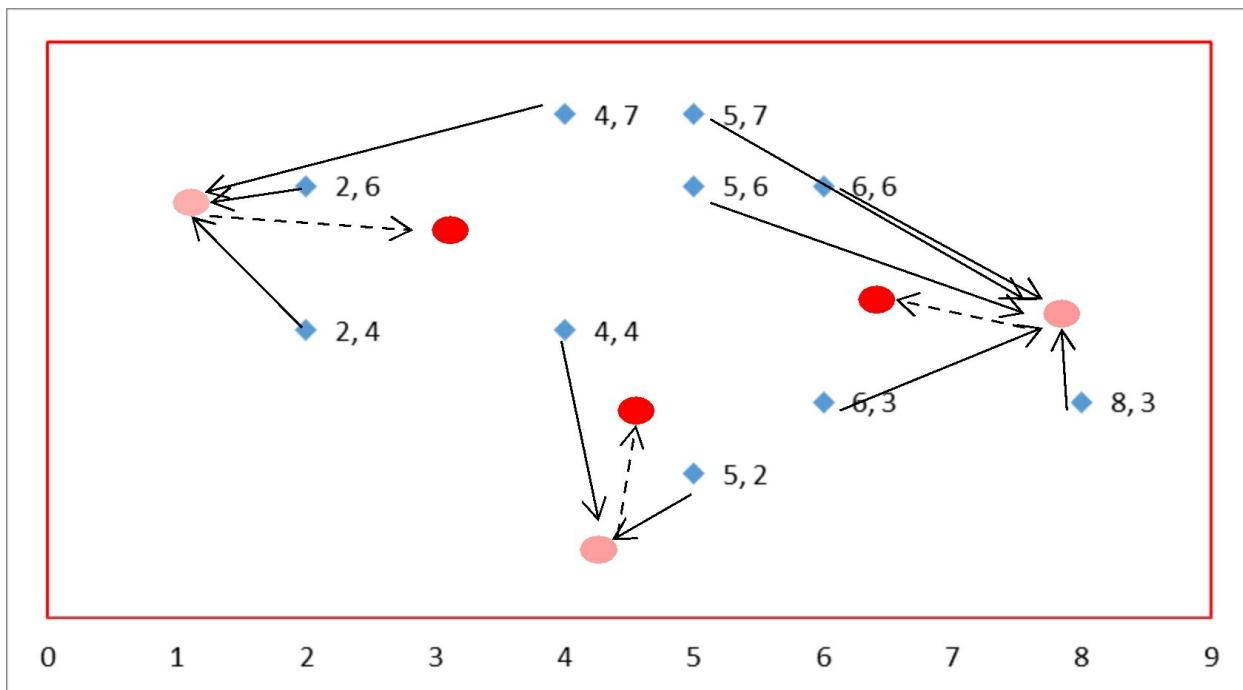


Figure 9.7: Recomputing centroids for each cluster

Step 3: Once again, data points are assigned to the three centroids closest to it (Figure 9.8).

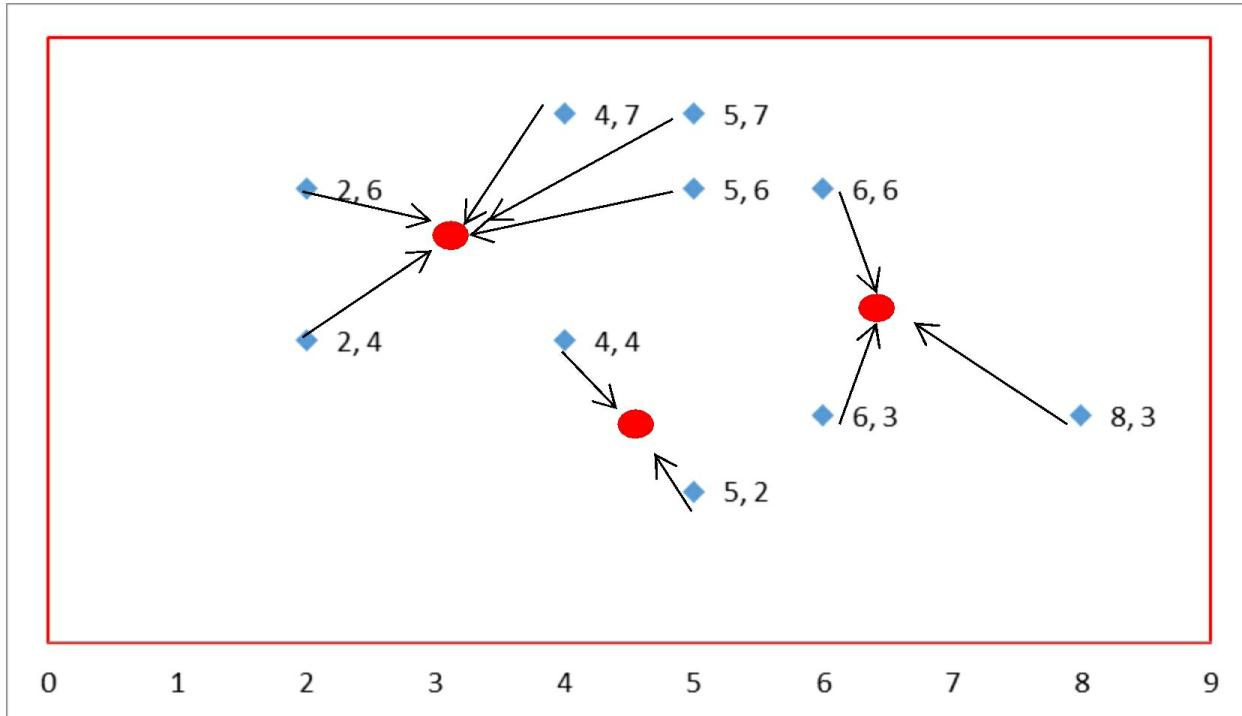
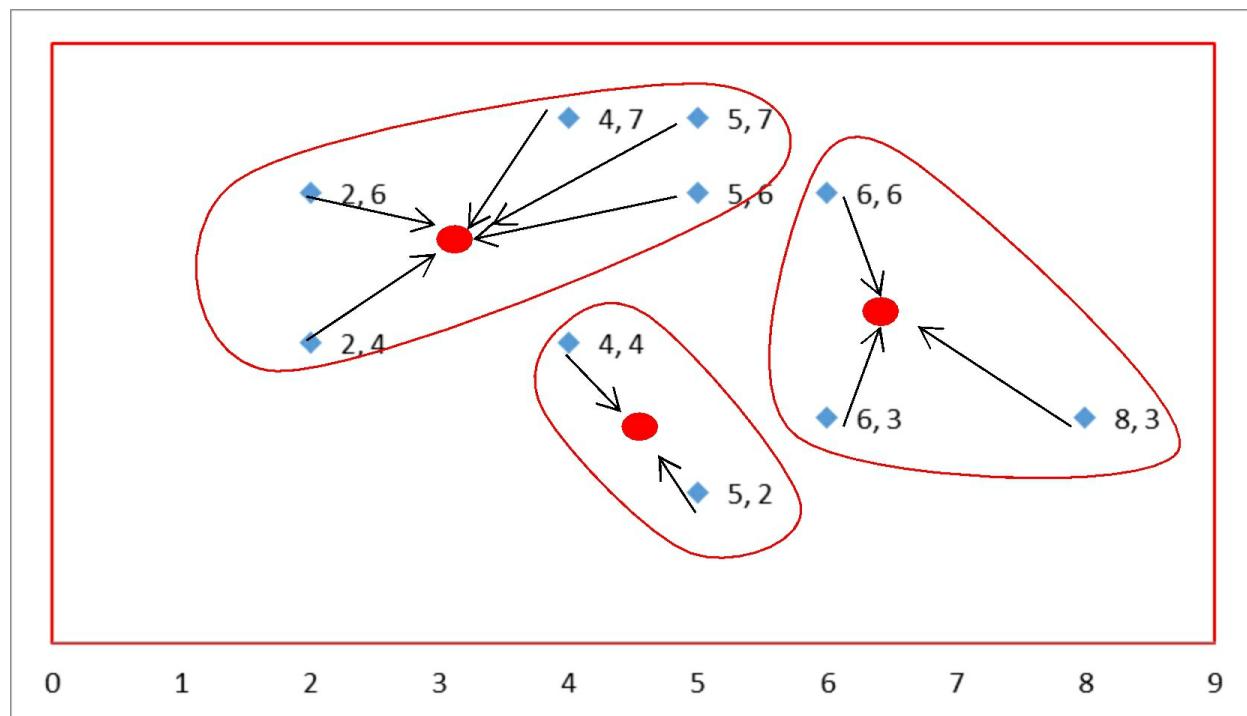


Figure 9.8: Assigning data points to recomputed centroids

The new centroids will be computed from the data points in the cluster until finally, the centroids stabilize in their locations. These are the three clusters computed by this algorithm.



**Figure 9.9: Recomputing centroids for each cluster till clusters stabilize**

The three clusters shown are: a 3-datapoints cluster with centroid (6.5,4.5), a 2- datapoint cluster with centroid (4.5,3) and a 5-datapoint cluster with centroid (3.5,3) (Figure 9.9).

These cluster definitions are different from the ones derived visually. This is a function of the random starting centroid values. The centroid points used earlier in the visual exercise were different from that chosen with the K-means clustering algorithm. The K-means clustering exercise should therefore, be run again with this data, but with new random centroid starting values. With many runs, the cluster definitions are likely to stabilize. If the cluster definitions do not stabilize, that may be a sign that the number of clusters chosen is too high or too low. The algorithm should also be run with different values of K.

Here is the pseudo code for implementing a K-means algorithm.

**Algorithm K-Means (K number of clusters, D list of data points)**

1. Choose K number of random data points as initial centroids (cluster-centers)
2. Repeat till cluster-centers stabilize
  - a. *{ Allocate each point in D to the nearest of K centroids;*
  - b. *Compute centroid for the cluster using all points in*

### Selecting the number of clusters

The correct choice of the value of  $k$  is often ambiguous. It depends on the shape and scale of the distribution points in a data set and the desired clustering resolution of the user. Heuristics are needed to pick the right number. One can graph the percentage of variance explained by the clusters against the number of clusters (Fig 9.10). The first clusters will add more information (explain a lot of variance), but at some point the marginal gain in variance will fall, giving a sharp angle to the graph, looking like an elbow. Beyond that elbow point, adding more clusters will not add much incremental value. That would be the desired  $K$ .

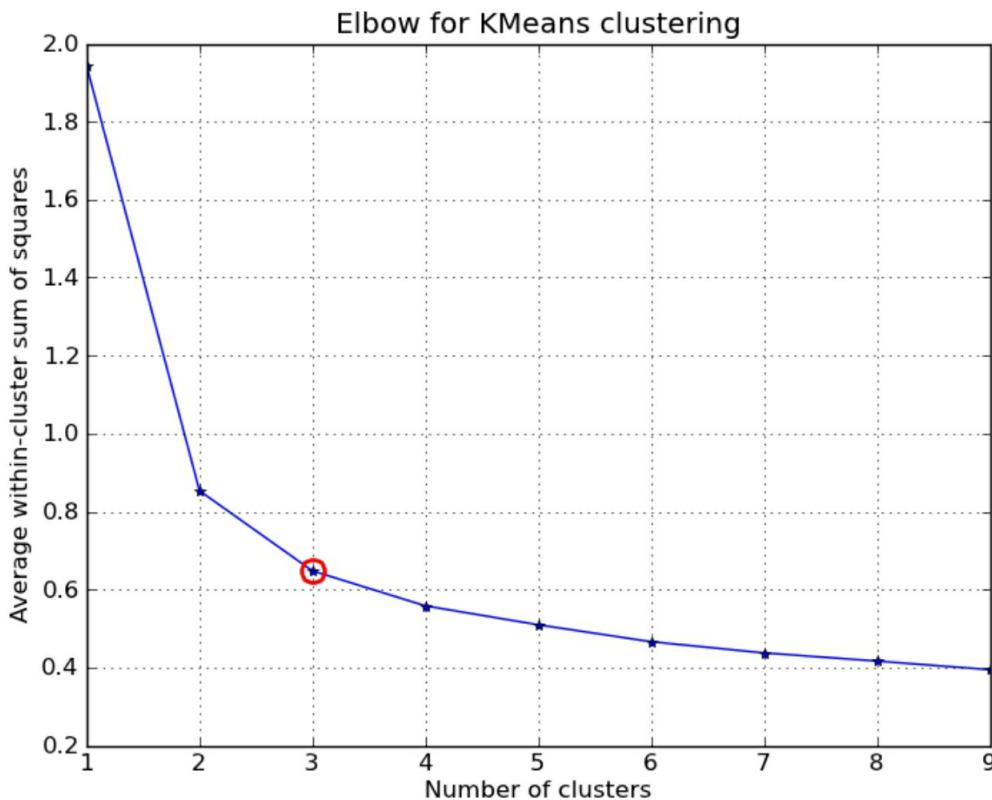


Figure 9.10: Elbow method for determining number of clusters in a data set

To engage with the data and to understand the clusters better, it is often better to start with a small number of clusters such as 2 or 3, depending upon the data set and the application domain. The number can be increased subsequently, as needed from an application point of view. This helps understand the data and the clusters progressively better.

## Advantages and Disadvantages of K-Means algorithm

There are many advantages of K-Means Algorithm

1. K-Means algorithm is simple, easy to understand and easy to implement.
2. It is also efficient, in that the time taken to cluster k-means, rises linearly with the number of data points.
3. No other clustering algorithm performs better than K-Means, in general.

There are a few disadvantages too:

1. The user needs to specify an initial value of K.
2. The process of finding the clusters may not converge.
3. It is not suitable for discovering clusters shapes that are not hyper-ellipsoids (or hyper-spheres).

Neural networks can also be deployed for clustering, using the appropriate objective function. The neural network will produce the appropriate cluster centroids and cluster population for each cluster.

## Conclusion

Cluster analysis is a useful, unsupervised learning technique that is used in many business situations to segment the data into meaningful small groups. K-Means algorithm is an easy statistical technique to iteratively segment the data. However, there is only a heuristic technique to select the right number of clusters.

## Review Exercises

1: What is unsupervised learning? When is it used?

2: Describe three business applications in your industry where cluster analysis will be useful.

3: Data about height and weight for a few volunteers is available. Create a set of clusters for the following data, to decide how many sizes of T-shirts should be ordered.

Height	Weight
71	165
68	165
72	180
67	113
72	178
62	101
70	150
69	172
72	185

63	149
69	132
61	115

### Liberty Stores Case Exercise: Step 7

Liberty wants to find suitable number of segments for its customers, for targeted marketing. Here is a list of representative customers.

Cust	# of trans-actions	Total Purchase (\$)	Income (\$ K)
1	5	450	90
2	10	800	82
3	15	900	77
4	2	50	30
5	18	900	60
6	9	200	45
7	14	500	82
8	8	300	22
9	7	250	90
10	9	1000	80
11	1	30	60
12	6	700	80

1. What is the right number of clusters for Liberty?
2. What are their centroids for the clusters?

# Chapter 10: Association Rule Mining

## Introduction

Associate rule mining is a popular, unsupervised learning technique, used in business to help identify shopping patterns. It is also known as market basket analysis. It helps find interesting relationships (affinities) between variables (items or events). Thus, it can help cross-sell related items and increase the size of a sale.

All data used in this technique is of categorical type. There is no dependent variable. It uses machine learning algorithms. The fascinating “relationship between sales of diapers and beers’ is how it is often explained in popular literature. This technique accepts as input the raw point-of-sale transaction data. The output produced is the description of the most frequent affinities among items. An example of an association rule would be, “A Customer who bought a flight tickets and a hotel reservation also bought a rental car plan 60 percent of the time.”

### Caselet: Netflix: Data Mining in Entertainment

*Netflix suggestions and recommendation engines are powered by a suite of algorithms using data about millions of customer ratings about thousands of movies. Most of these algorithms are based on the premise that similar viewing patterns represent similar user tastes. This suite of algorithms, called CineMatch, instructs Netflix's servers to process information from its databases to determine which movies a customer is likely to enjoy. The algorithm takes into account many factors about the films themselves, the customers' ratings, and the combined ratings of all Netflix users. The company estimates that a whopping 75 percent of viewer activity is driven by recommendations. According to Netflix, these predictions were valid around 75 percent of the time and half of Netflix users who rented CineMatch-recommended movies gave them a five-star rating.*

*To make matches, a computer*

1. *Searches the CineMatch database for people who have rated the same movie - for example, "The Return of the Jedi".*
2. *Determines which of those people have also rated a second*

- movie, such as "The Matrix".*
- 3. *Calculates the statistical likelihood that people who liked "Return of the Jedi" will also like "The Matrix".*
  - 4. *Continues this process to establish a pattern of correlations between subscribers' ratings of many different films.*

*Netflix launched a contest in 2006 to find an algorithm that could beat CineMatch. The contest, called the Netflix Prize, promised \$1 million to the first person or team to meet the accuracy goals for recommending movies based on users' personal preferences. Each of these algorithm submissions was required to demonstrate a 10 percent improvement over CineMatch. Three years later, the \$1 million prize was awarded to a team of seven people. (source: <http://electronics.howstuffworks.com>).*

*1: Are Netflix customers being manipulated into seeing what Netflix wants them to see?*

*2: Compare this story with Amazon's personalization engine.*

## **Business Applications of Association Rules**

In business environments a pattern or knowledge can be used for many purposes. In sales and marketing, it is used for cross-marketing and cross-selling, catalog design, e-commerce site design, online advertising optimization, product pricing, and sales/promotion configurations. This analysis can suggest not to put one item on sale at a time, and instead to create a bundle of products promoted as a package to sell other non-selling items.

In retail environments, it can be used for store design. Strongly associated items can be kept close together for customer convenience. Or they could be placed far from each other so that the customer has to walk the aisles and by doing so is potentially exposed to other items.

In medicine, this technique can be used for relationships between symptoms and illnesses; diagnosis and patient characteristics/treatments; genes and their functions; etc.

## **Representing Association Rules**

A generic Association Rule is represented between a set X and Y: **X  $\rightarrow$  Y**

## [S%, C%]

**X, Y:** products and/or services

**X:** Left-hand-side (LHS)

**Y:** Right-hand-side (RHS)

**S:** Support: how often X and Y go together in the dataset – i.e.  $P(X \cup Y)$

**C:** Confidence: how often Y is found, given X – i.e.  $P(Y | X)$

*Example:* {Hotel booking, Flight booking}  $\bowtie$  {Rental Car} [30%, 60%]

[Note:  $P(X)$  is the mathematical representation of the probability or chance of X occurring in the data set.]

Computation example:

Suppose there are 1000 transactions in a data set. There are 300 occurrences of X, and 150 occurrences of (X, Y) in the data set.

Support S for X  $\bowtie$  Y will be  $P(X \cup Y) = 150/1000 = 15\%$ .

Confidence for X  $\bowtie$  Y will be  $P(Y | X)$ ; or  $P(X \cup Y) / P(X) = 150/300 = 50\%$

## Algorithms for Association Rule

Not all association rules are interesting and useful, only those that are strong rules and also those that occur frequently. In association rule mining, the goal is to find all rules that satisfy the user-specified *minimum support* and *minimum confidence*. The resulting sets of rules are all the same irrespective of the algorithm used, that is, given a transaction data set T, a minimum support and a minimum confidence, the set of association rules existing in T is *uniquely determined*.

Fortunately, there is a large number of algorithms that are available for generating association rules. The most popular algorithms are Apriori, Eclat, FP-Growth, along with various derivatives and hybrids of the three. All the algorithms help identify the frequent item sets, which are then converted to association rules.

## Apriori Algorithm

This is the most popular algorithm used for association rule mining. The objective is to find subsets that are common to at least a minimum number of

the itemsets. A frequent itemset is an itemset whose support is greater than or equal to minimum support threshold. The Apriori property is a downward closure property, which means that any subsets of a frequent itemset are also frequent itemsets. Thus, if (A,B,C,D) is a frequent itemset, then any subset such as (A,B,C) or (B,D) are also frequent itemsets.

It uses a bottom-up approach; and the size of frequent subsets is gradually increased, from one-item subsets to two-item subsets, then three-item subsets, and so on. Groups of candidates at each level are tested against the data for minimum support.

### **Association rules exercise**

Here are a dozen sales transactions. There are six products being sold: Milk, Bread, Butter, Eggs, Cookies, and Ketchup. Transaction#1 sold Milk, Eggs, Bread and Butter. Transaction#2 sold Milk, Butter, Egg & Ketchup. And so on. The objective is to use this transaction data to find affinities between products, i.e. which products sell together often.

The support level will be set at 33 percent; the confidence level will be set at 50 percent. That means that we have decided to consider rules from only those itemsets that occur at least 33 percent of the time in the total set of transactions. Confidence level means that within those itemsets, the rules of the form X → Y should be such that there is at least 50 percent chance of Y occurring based on X occurring.

Transactions List			
1	Milk	Egg	Bread
2	Milk	Butter	Egg
3	Bread	Butter	Ketchup
4	Milk	Bread	Butter
5	Bread	Butter	Cookies
6	Milk	Bread	Butter
7	Milk	Cookies	
8	Milk	Bread	Butter
9	Bread	Butter	Egg
10	Milk	Butter	Bread

11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

First step is to compute 1-item Itemsets. i.e. How often does any product individually sell.

1-item Sets	Freq
Milk	9
Bread	10
Butter	10
Egg	3
Ketchup	3
Cookies	5

Thus, Milk sells in 9 out of 12 transactions. Bread sells in 10 out of 12 transactions. And so on.

At every point, there is an opportunity to select itemsets of interest, and thus further analysis. Other itemsets that occur very infrequently may be removed. If itemsets that occur 4 or more times out of 12 are selected, that corresponds to meeting a minimum support level of 33 percent (4 out of 12). Only 4 items make the cut. The frequent items that meet the support level of 33 percent are:

Frequent 1-item Sets	Freq
Milk	9
Bread	10
Butter	10
Cookies	5

The next step is to go for the next level of itemsets using items selected earlier: 2-item itemsets.

--	--

<b>2-item Sets</b>	<b>Freq</b>
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

Thus (Milk, Bread) sell 7 times out of 12. (Milk, Butter) sell together 7 times, (Bread, Butter) together 9 times, and (Bread, Cookies) sell 4 times.

However only four of these transactions meet the minimum support level of 33%.

<b>2-item Sets</b>	<b>Freq</b>
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

The next step is to list the next higher level of itemsets: 3-item itemsets.

<b>3-item Sets</b>	<b>Freq</b>
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3

Thus (Milk, Bread, Butter) sell 6 times out of 12. (Bread, Butter, Cookies) sell 3 times out of 12. One one 3-item itemset meets the minimum support requirements.

<b>3-item Sets</b>	<b>Freq</b>
Milk, Bread, Butter	6

There is no room to create a 4-item itemset for this support level.

## Creating Association Rules

The most interesting and complex rules at higher size itemsets start top-down with the most frequent itemsets of higher size-numbers. Association rules are created that meet the support level ( $>33\%$ ) and confidence levels ( $> 50\%$ ).

The highest level itemset that meets the support requirements is the three-item itemset. The following itemset has a support level of 50% (6 out of 12).

Milk, Bread, Butter	6
---------------------	---

This itemset could lead to multiple candidate Association rules.

Start with the following rule: (Bread, Butter)  $\rightarrow$  Milk.

There are a total of total 12 transactions.

X (in this case Bread, Butter) occurs 9 times;

X,Y (in this case Bread, Butter, Milk) occurs 6 times.

The support level for this rule is  $6/12 = 50\%$ . The confidence level for this rule is  $6/9 = 67\%$ . This rule meets our thresholds for support ( $>33\%$ ) and confidence ( $>50\%$ ).

Thus, the first valid Association rule from this data is: **(Bread, Butter)  $\rightarrow$  Milk {S=50%, C=67%}**.

In exactly the same way, other rules can be considered for their validity.

Consider the rule: (Milk, Bread)  $\rightarrow$  Butter. Out of total 12 transactions, (Milk, Bread) occur 7 times; and (Milk, Bread, Butter) occurs 6 times.

The support level for this rule is  $6/12 = 50\%$ . The confidence level for this rule is  $6/7 = 86\%$ . This rule meets our thresholds for support ( $>33\%$ ) and confidence ( $>50\%$ ).

Thus, the second valid Association rule from this data is: **(Milk, Bread)  $\rightarrow$  Butter {S=50%, C=67%}**.

Consider the rule (Milk, Butter)  $\rightarrow$  Bread. Out of total 12 transactions (Milk, Butter) occurs 7 times while (Milk, Butter, Bread) occur 6 times.

The support level for this rule is  $6/12 = 50\%$ . The confidence level for this rule is  $6/7 = 86\%$ . This rule meets our thresholds for support ( $>33\%$ ) and confidence ( $>50\%$ ).

Thus, the next valid Association rule is: **Milk,Butter → Bread {S=50%, C=86%}**.

Thus, there were only three possible rules at the 3-item itemset level, and all were found to be valid.

One can get to the next lower level and generate association rules at the 2-item itemset level.

Consider the rule Milk → Bread. Out of total 12 transactions Milk occurs 9 times while (Milk, Bread) occur 7 times.

The support level for this rule is  $7/12 = 58\%$ . The confidence level for this rule is  $7/9 = 78\%$ . This rule meets our thresholds for support ( $>33\%$ ) and confidence ( $>50\%$ ).

Thus, the next valid Association rule is:

**Milk -> Bread {58%, 78%}**.

Many such rules could be derived if needed.

Not all such association rules are interesting. The client may be interested in only the top few rules that they want to implement. The number of association rules depends upon business need. Implementing every rule in business will require some cost and effort, with some potential of gains. The strongest of rules, with the higher support and confidence rates, should be used first, and the others should be progressively implemented later.

## Conclusion

Association Rules help discover affinities between products in transactions. It helps make cross-selling recommendations much more targeted and effective. Apriori technique is the most popular technique, and it is a machine learning technique.

## Review Exercises

Q1: What are association rules? How do they help?

Q2: How many association rules should be used?

Q3: What are frequent item-sets?

Q4: How does the Apriori algorithm work?

## Liberty Stores Case Exercise: Step 8

*Here is a list of Transactions from Liberty's stores. Create association rules for the following data. With 33% support level and 66% confidence.*

	Barley	Corn	Gram	Millet	Rice	Wheat
1	1	1	1	1	1	1
2		1		1	1	1
3	1		1		1	1
4		1	1		1	1
5	1		1	1	1	
6			1		1	1
7	1			1	1	1
8				1	1	1
9	1	1	1	1		
10	1	1	1		1	1
11		1		1	1	1
12	1	1		1	1	1

## **Section 3 – Advanced Mining**

This section covers advanced mining techniques that also include Text and Web Mining.

Chapter 11 covers Text Mining, the art and science of generating insights from text. It is very important in the age of social media.

Chapter 12 covers Naïve-Bayes classification technique used often for text data mining problems such as spam filtering. It includes two solved problems.

Chapter 13 covers Support Vector Machines, a machine learning technique, also used often for text data mining problems such as spam filtering.

Chapter 14 covers Web Mining, the art and science of generating insights from the world-wide web, its content and usage. It is very important in the digital age where a lot of advertising and selling is moving to the web.

Chapter 15 covers Social Network Analysis, the art and science of discovering patterns of interaction and influence within the participants in a network. It can be applied in a wide range of situations from communication networks among human beings to the relative importance of websites with the worldwide web,

# Chapter 11: Text Mining

## Introduction

Text mining is the art and science of discovering knowledge, insights and patterns from an organized collection of textual databases. Textual mining can help with frequency analysis of important terms, and their semantic relationships.

Text is an important part of the growing data in the world. Social media technologies have enabled users to become producers of text and images and other kinds of information. Text mining can be applied to large-scale social media data for gathering preferences, and measuring emotional sentiments. It can also be applied to societal, organizational and individual scales.

### Caselet: WhatsApp and Private Security

*Do you think that what you post on social media remains private? Think again. A new dashboard shows how much personal information is out there, and how companies are able to construct ways to make use of it for commercial benefits. Here is a dashboard of conversations between two people Jennifer and Nicole over 45 days.*

*There is a variety of categories that Nicole and Jennifer speak about such as computers, politics, laundry, desserts. The polarity of Jennifer's personal thoughts and tone is overwhelmingly positive, and Jennifer responds to Nicole much more than vice versa, identifying Nicole as the influencer in their relationship.*

*The data visualization reveals the waking hours of Jennifer, showing that she is most active around 8:00pm and heads to bed around midnight. 53% of her conversation is about food – and 15% about desserts . Maybe she's a strategic person to push restaurant or weight loss ads.*

*The most intimate detail exposed during this conversation is that Nicole and Jennifer discuss right wing populism, radical parties, and conservative politics. It exemplifies that the amount of private information obtained from your WhatsApp conversations is limitless and potentially dangerous.*

*WhatsApp is the world's largest messaging service that has over 450 million users. FaceBook recently bought this three year old company for a whopping \$19 billion. People share a lot of sensitive personal information on WhatsApp that they may not even share with their family members.*

*(Sources: What Facebook Knows About You From One WhatsApp Conv, by Adi Azaria, on LinkedIn, April 10, 2014).*

- 1: *What are the business and social implications of this kind of analysis?*
- 2: *Are you worried? Should you be worried?*

Text mining works on texts from practically any kind of sources from any

business or non-business domains, in any formats including Word documents, PDF files, XML files, text messages, etc. Here are some representative examples:

1. In the *legal profession*, text sources would include law, court deliberations, court orders, etc.
2. In *academic research*, it would include texts of interviews, published research articles, etc.
3. The world of *finance* will include statutory reports, internal reports, CFO statements, and more.
4. In *medicine*, it would include medical journals, patient histories, discharge summaries, etc.
5. In *marketing*, it would include advertisements, customer comments, etc.
6. In the world of *technology and search*, it would include patent applications, the whole of information on the world-wide web, and more.

## Text Mining Applications

Text mining is a useful tool in the hands of chief knowledge officers to extract knowledge relevant to an organization. Text mining can be used across industry sectors and application areas, including decision support, sentiment analysis, fraud detection, survey analysis, and many more.

1. *Marketing*: The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.
  - a. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used towards anticipating and predicting consumer behavior.
  - b. A ‘listening platform’ is a text mining application, that in real time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiment. The insights can lead to more effective product marketing and better customer service.

- c. The customer call center conversations and records can be analyzed for patterns of customer complaints. Decision trees can organize this data to create decision choices that could help with product management activities and to become proactive in avoiding those complaints.
- 2. *Business operations*: Many aspects of business functioning can be accurately gauged from analyzing text.
  - a. Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the emotional states and the mood of employee populations. Sentiment analysis can reveal early signs of employee dissatisfaction which can then be proactively managed.
  - b. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.
- 3. *Legal*: In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.
  - a. Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents.
  - b. Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in a healthcare situation that can help better predict high-cost injuries and prevent costs.
- 4. *Governance and Politics*: Governments can be overturned based on a tweet originating from a self-immolating fruit-vendor in Tunisia.
  - a. Social network analysis and text mining of large-scale social media data can be used for measuring the emotional states and the mood of constituent populations. Micro-targeting constituents with specific messages gleaned from social media analysis can be a more efficient use of resources when fighting democratic elections.

- b. In geopolitical security, internet chatter can be processed for real-time information and to connect the dots on any emerging threats.
- c. In academic, research streams could be meta-analyzed for underlying research trends.

## **Text Mining Process**

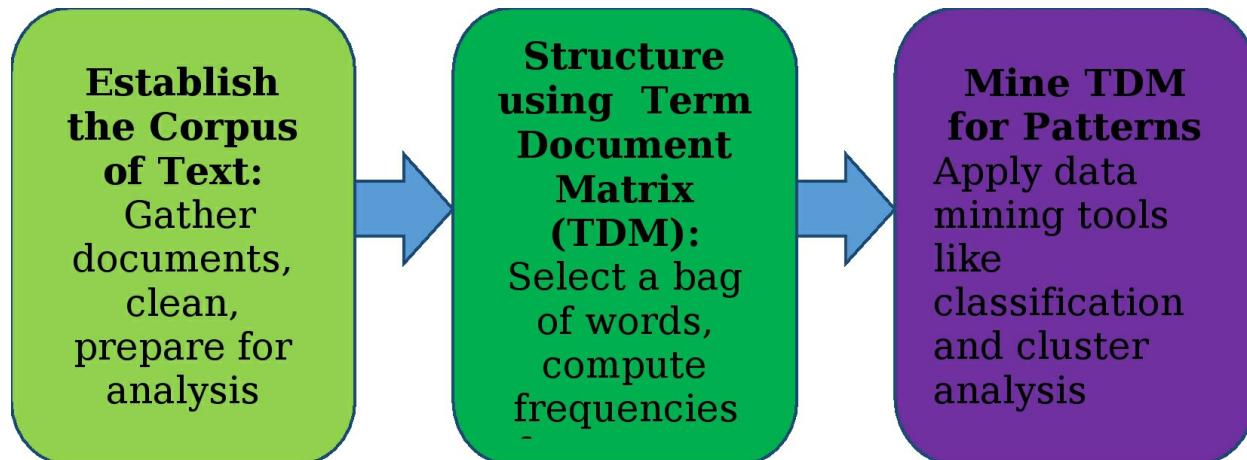
Text Mining is a rapidly evolving area of research. As the amount of social media and other text data grows, there is need for efficient abstraction and categorization of meaningful information from the text.

The first level of analysis is identifying frequent words. This creates a bag of important words. Texts – documents or smaller messages – can then be ranked on how they match to a particular bag-of-words. However, there are challenges with this approach. For example, the words may be spelled a little differently. Or there may be different words with similar meanings.

The next level is at the level of identifying meaningful phrases from words. Thus ‘ice’ and ‘cream’ will be two different key words that often come together. However, there is a more meaningful phrase by combining the two words into ‘ice cream’. There might be similarly meaningful phrases like ‘Apple Pie’.

The next higher level is that of Topics. Multiple phrases could be combined into Topic area. Thus the two phrases above could be put into a common basket, and this bucket could be called ‘Desserts’.

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3-step process (Figure 11.1)



**Figure 11.1: Text Mining Architecture**

1. The text and documents are first gathered into a corpus, and organized.
2. The corpus is then analyzed for structure. The result is a matrix mapping important terms to source documents.
3. The structured data is then analyzed for word structures, sequences, and frequency.

## Term Document Matrix

This is the heart of the structuring process. Free flowing text can be transformed into numeric data in a TDM, which can then be mined using regular data mining techniques.

1. There are several efficient techniques for identifying key terms from a text. There are less efficient techniques available for creating topics out of them. For the purpose of this discussion, one could call key words, phrases or topics as a term of interest. This approach measures the frequencies of select important terms occurring in each document. This creates a  $t \times d$  Term-by-Document Matrix (TDM) where  $t$  is the number of terms and  $d$  is the number of documents (Table 11.1).
2. Creating a TDM requires making choices of which terms to include. The terms chosen should reflect the stated purpose of the text mining exercise. The list of terms should be as extensive as needed, but should not include unnecessary stuff that will serve to confuse the analysis, or slow the computation.

Term Document Matrix					
Document / Terms	investment	Profit	happy	Success	...
Doc 1	10	4	3	4	
Doc 2	7	2	2		
Doc 3			2	6	
Doc 4	1	5	3		
Doc 5		6		2	
Doc 6	4		2		
...					

**Table 11.1: Term-Document Matrix**

Here are some considerations in creating a TDM.

1. A large collection of documents mapped to a large bag of words will likely lead to a very sparse matrix if they have few common words. Reducing dimensionality of data will help improve the speed of analysis and meaningfulness of the results. Synonyms, or terms with similar meaning, should be combined and should be counted together, as a common term. This would help reduce the number of distinct terms of words or ‘tokens’.
2. Data should be cleaned for spelling errors. Common spelling errors should be ignored and the terms should be combined. Uppercase-lowercase terms should also be combined.
3. When many variants of the same term are used, just the stem of the word would be used to reduce the number of terms. For instance, terms like customer order, ordering, order data, should be combined into a single token word, called ‘Order’.
4. On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately. This would enhance the quality of analysis. For example, the term order can mean a customer order, or the ranking of certain choices. These two should be treated separately. “The boss ordered that the customer orders data analysis be presented in chronological order”. This statement shows three different meanings for the word ‘order’. Thus, there will be a need for a manual review of the TD matrix.
5. Terms with very few occurrences in very few documents should be eliminated from the matrix. This would help increase the density of the matrix and the quality of analysis.
6. The measures in each cell of the matrix could be one of several possibilities. It could be a simple count of the number of occurrences of each term in a document. It could also be the log of that number. It could be the fraction number computed by dividing the frequency count by the total number of words in the document. Or there may be binary values in the matrix to represent whether a term is mentioned or not. The choice of value in the cells will depend upon the purpose of the text analysis.

At the end of this analysis and cleansing, a well-formed, densely populated, rectangular, TDM will be ready for analysis. The TDM could be mined using all the available data mining techniques.

## Mining the TDM

The TDM can be mined to extract patterns/knowledge. A variety of techniques could be applied to the TDM to extract new knowledge.

A simple application will be to visualize the highest frequency terms. This can be done very attractively and colorfully in the form of a Wordcloud. The wordcloud could be created after removing the common words like prepositions. It can be done for the top n words such as top 100 words, to focus on the key terms used in the document. The attached wordcloud represents the speech by US President Barack Obama on the topic of terrorism.

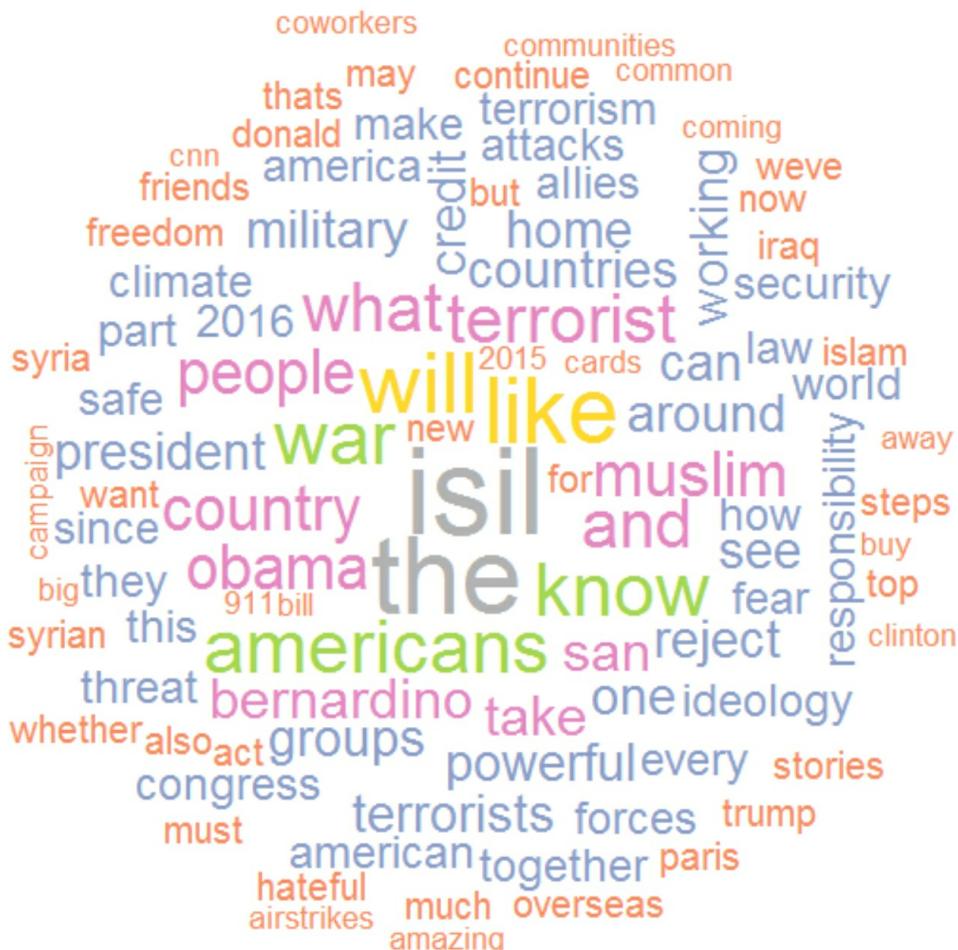


Figure 11-2: WordCloud of top 100 words from US President's speech

Predictors of desirable terms could be discovered through predictive techniques, such as regression analysis. Suppose the word profit is a desirable word in a document. The number of occurrences of the word profit in a document could be regressed against many other terms in the TDM. The relative strengths of the coefficients of various predictor variables would show the relative impact of those terms on creating a profit discussion.

Predicting the chances of a document being liked is another form of analysis. For example, important speeches made by the CEO or the CFO to investors could be evaluated for quality. If the classification of those documents (such as good or poor speeches) was available, then the terms of TDM could be used to predict the speech class. A decision tree could be constructed that makes a simple tree with a few decision points that predicts the success of a speech 80 percent of the time. This tree could be trained with more data to become better over time.

Clustering techniques can help categorize documents by common profile. For example, documents containing the words investment and profit more often could be bundled together. Similarly, documents containing the words, customer orders and marketing, more often could be bundled together. Thus, a few strongly demarcated bundles could capture the essence of the entire TDM. These bundles could thus help with further processing, such as handing over select documents to others for legal discovery.

Association rule analysis could show relationships of coexistence. Thus, one could say that the words, tasty and sweet, occur together often (say 5 percent of the time); and further, when these two words are present, 70 percent of the time, the word happy, is also present in the document.

## Comparing Text Mining and Data Mining

Text Mining is a form of data mining. There are many common elements between Text and Data Mining. However, there are some key differences (Table 11.2). The key difference is that text mining requires conversion of text data into frequency data, before data mining techniques can be applied.

Dimension	Text Mining	Data Mining
Nature of data	Unstructured data: Words, phrases, sentences	Numbers; alphabetical and logical values

Language used	Many languages and dialects used in the world; many languages are extinct, new documents are discovered	Similar numerical systems across the world
Clarity and precision	Sentences can be ambiguous; sentiment may contradict the words	Numbers are precise.
Consistency	Different parts of the text can contradict each other	Different parts of data can be inconsistent, thus, requiring statistical significance analysis
Sentiment	Text may present a clear and consistent or mixed sentiment, across a continuum. Spoken words adds further sentiment	Not applicable
Quality	Spelling errors. Differing values of proper nouns such as names. Varying quality of language translation	Issues with missing values, outliers, etc
Nature of analysis	Keyword based search; co-existence of themes; Sentiment mining;	A full wide range of statistical and machine learning analysis for relationships and differences

Table 11.2: Comparing Text Mining and Data Mining

## Text Mining Best Practices

Many of the best practices that apply to the use of data mining techniques will also apply to text mining.

1. The first and most important practice is to ask the right question. A good question is one which gives an answer and would lead to large payoffs for the organization. The purpose and the key

question will define how and at what levels of granularity the TDM would be made. For example, TDM defined for simpler searches would be different from those used for complex semantic analysis or network analysis.

2. A second important practice is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in the quality of the proposed solution as well as in finding the high quality data sets required to test the hypothesized solution. For example, a TDM of consumer sentiment data should be combined with customer order data in order to develop a comprehensive view of customer behavior. It's important to assemble a team that has a healthy mix of technical and business skills.
3. Another important element is to pursue the problem iteratively. Too much data can overwhelm the infrastructure and also befuddle the mind. It is better to divide and conquer the problem with a simpler TDM, with fewer terms and fewer documents and data sources. Expand as needed, in an iterative sequence of steps. In the future, add new terms to help improve predictive accuracy.
4. A variety of data mining tools should be used to test the relationships in the TDM. Different decision tree algorithms could be run alongside cluster analysis and other techniques. Triangulating the findings with multiple techniques, and many what-if scenarios, helps build confidence in the solution. Test the solution in many ways before committing to deploy it.

## Conclusion

Text Mining is diving into the unstructured text to discover valuable insights about the business. The text is gathered and then structured into a term-document matrix based on the frequency of a bag of words in a corpus of documents. The TDM can then be mined for useful, novel patterns, and insights. While the technique is important, the business objective should be well understood and should always be kept in mind.

\*\*\*

## Review Questions

1: Why is text mining useful in the age of social media?

2: What kinds of problems can be addressed using text mining?

3: What kinds of sentiments can be found in the text?

Do a Text mining analysis of sales speeches by three salesmen.

1. *Did you know your team can build Powerpoint muscles? Yes, I help build PowerPoint muscles. I teach people how to use PowerPoint more effectively in business. Now, for instance, I'm working with a global consulting firm to train all their senior consultants to give better sales presentations so they can close more business.*
2. *I train people how to make sure their PowerPoint slides aren't a complete disaster. Those who attend my workshop can create slides that are 50% more clear and 50% more convincing by the end of the training, based on scores students give each other before and after the workshop. I'm not sure if my training could work at your company. But I'd be happy to talk to you about it.*
3. *You know how most business people use PowerPoint but most use it pretty poorly? Well, bad PowerPoint has all kinds of consequences – sales that don't close, good ideas that get ignored, time wasted building slides that could have been used developing or executing strategies. My company shows businesses how to use PowerPoint to capture those sales, bring attention to those great ideas and use those wasted hours on more important projects.*

The purpose is to select the best speech.

1: How would you select the right bag of words?

2: If speech #1 was the best speech, use the TDM to create a rule for good speeches.

### Liberty Stores Case Exercise: Step 9

Here are a few comments from customer service calls received by Liberty.

1. *I loved the design of the shirt. The size fitted me very well. However, the fabric seemed flimsy. I am calling to see if you can replace the shirt with a different one. Or please refund my money.*
2. *I was running late from work, and I stopped by to pick up some groceries. I did not like the way the manager closed the store while I was still shopping.*

3. *I stopped by to pick up flowers. The checkout line was very long. The manager was polite but did not open new cashiers. I got late for my appointment.*
4. *The manager promised that the product will be there, but when I went there the product was not there. The visit was a waste. The manager should have compensated me for my trouble.*
5. *When there was a problem with my catering order, the store manager promptly contacted me and quickly got the kinks out to send me replacement food immediately. There are very courteous.*

*Create a TDM with not more than 6 key terms. [Hint: Treat each comment as a document]*

## Chapter 12: Naïve Bayes Analysis

### Introduction

Naïve Bayes (NB) technique is a supervised learning technique that uses probability-theory-based analysis. It is a machine-learning technique that computes the probabilities of an instance belonging of each one of many target classes, given the prior probabilities of classification using individual factors. Naïve Bayes technique is used often in classifying text documents into one of multiple predefined categories.

### **CASELET: Fraud detection in government contracts**

TCU, the Brazilian Court of Audit, wanted to use data mining for more efficient and effective use of resources for fraud detection in government transactions. They wanted to find ways to identify projects that had high probability of fraud. A NB classifier was designed to assign a risk factor to each case.

The audit universe was the set of purchases, contracts, assets, programs, public employees or contractors. The audit universe was the set of pairs of public and private parties of all public contracts signed by the Brazilian federal administration between 1997 and 2011, totalizing almost 800,000 pairs. allowing, and assisting audit agencies in transitioning from reactive to proactive fraud detection model.

Second, the risk factors of an audit universe were identified that could relate to the entire entities set. The risk factors included the project related elements as well as external factors.

Competitors: less competition = higher risk.

Value: more money = higher risk

Bid protests: more protests = higher risk

Bid winner profile: no qualified winner = higher risk

Bid winner previous sanctions: sanctioned winner = higher risk

Factors associated with different concerns than the risk itself were also added:

Political/economic impact: related to policy goals = higher economic impact

Social impact: public health spend = higher social impact

Timing and effectiveness: older purchases = less auditing effectiveness

Naïve Bayes algorithm analysis was used to compute the probability distribution of entities belonging into risk classes. They have only two classes (high and low risk). So two values for each entity were obtained: Probability of high risk, and probability of low risk. These probabilities were used to compute the conditional probability of fraud for the set of 800,000 pairs of public and private parties. The probability of high risk was used to sort the auditable units by total risk score. The result was that almost 2500 pairs had a high risk probability score of higher than 99%. This result included parties with obvious links, like private not-for-profit foundations for which the Brazilian public procurement law gives special treatment. Moreover, a number of high risk pairs were known by the auditors from previous investigations where corruption was effectively found. The auditors found the derivation of the high risk list as very reasonable. The ranked audit universe was then used to develop an audit plan.

Q1: What are the advantages to the government and people of Brazil from this analysis? How could this analysis be improved?

Q2: Identify another environment where a similar NB analysis could be beneficial?

## Probability

Probability is defined as a chance of something happening. The probability values thus range from zero to one; with a value of zero representing no chance, and to one representing total certainty. Using past event records, the probability of something happening in the future can be reliably assessed. For example, one can assess the probability of dying from an airline accident, but dividing the total number of airline accident related deaths in a time period by the total number of people flying during that period. These probabilities can then be compared to come to conclusions, such as the safety levels of various event types. E.g. past data may show that the probability of dying from an airline accident is less than that of dying from being hit by lightning.

The Naïve-Bayes algorithm is special in that it takes into consideration the

prior-probability of an instance belonging to a class, in addition to the recent track record of the instance belonging to that class.

- (a) The word Bayes refers to Bayesian analysis (based on the work of mathematician Thomas Bayes) which computes the probability of a new occurrence not only on the recent record, but also on the basis of prior experience.
- (b) The word Naïve represents the strong assumption that all parameters/features of an instance are independent variables with little or no correlation. Thus if people are identified by their height, weight, age, and gender, all these variables are assumed to be independent of each other.

NB algorithm is easy to understand and works fast. It also performs well in multi class prediction, such as when the target class has multiple options beyond binary yes/no classification. NB can perform well even in case of categorical input variables compared to numerical variable(s).

## Naïve-Bayes model

In the abstract, Naïve-Bayes is a conditional probability model for classification purposes.

The goal is to find a way to predict the class variable (Y) using a vector of independent variables (X). i.e. find the function  $f : X \rightarrow Y$ . In probability terms, the goal is to find  $P(Y|X)$ , the probability of Y belonging to a certain class, given X. Y is generally assumed to be a categorical variable with two or more discrete values.

Given an instance to be classified, represented by a vector  $\mathbf{x} = (x_1, \dots, x_n)$  representing some n features (independent variables), the Naïve Bayes model assigns, to an instance, probabilities of belonging to any of K classes. The class k with the highest posterior probability, is the label assigned to the instance.

The posterior probability (of belonging to a Class k) is calculated as a function of prior probabilities and current likelihood value, as shown in equation below:

$$p(C_k|\mathbf{x}) = \frac{p(C_k) p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

$P(C_k | \mathbf{x})$  is the posterior probability of *class k* given *predictor X*.

$P(C_k)$  is the prior probability of *class k*.

$P(\mathbf{x})$  is the prior probability of *predictor*.

And  $P(\mathbf{x}|C_k)$  is the current likelihood of *predictor* given *class*.

## Simple classification example

Suppose a barber salon needs to predict the service required by the incoming customer. Suppose there are only two services offered: Cutting-Hair (R) and Manicure-Pedicure (M). Thus the value to be predicted is whether the next customer will be for R or M. The number of classes (K) is 2.

First step is to compute prior probability. Suppose the data gathered for the last one year showed that during that period there were 2500 customers for R, and 1500 customers for M. Thus the default (or prior) probability for the next customer to be for R is  $2500/4000$  or  $5/8$ . Similarly, the default probability for the next customer to be for M is  $1500/4000$  or  $3/8$ . Based on this information alone, the next customer would likely be for R.

Another way to predict the service requirement by the next customer would be to look at the most recent data. One could look at the last few (choose a number) customers, to predict the next customer. Suppose the last five customers were for the services ... R, M, R, M, M, in that order. Thus the data shows the recent probability of R is  $2/5$  and the recent probability of M is  $3/5$ . Based on just this information, the next customer will likely be for M.

Thomas Bayes suggested that the prior probability should be informed by the more recent data. Naïve Bayes posterior probability for a class is computed by multiplying the prior probability and the recent probability.

Thus in this example, the NB posterior probability P (R), is  $5/8 \times 2/5 = 10/40$ . Similarly, the NB probability P (M) is  $3/8 \times 3/5 = 9/40$ . Since P (R) is greater than P (M), there is a greater probability of the next customer be for R. Thus the expected class label assigned to the next customer would be R.

Suppose however that the next customer coming in was for M service. The last five customer sequence now becomes M, R, M, M, M. Thus the recent data shows the recent probability for R is  $1/5$  and the recent probability of M is  $4/5$ .

Now the NB probability for R is  $5/8 \times 1/5 = 5/40$ . Similarly, the NB probability for M is  $3/8 \times 4/5 = 12/40$ . Since  $P(M)$  is greater than  $P(R)$ , there is a greater probability of the next customer being for M. Thus the expected class label assigned to the next customer is M.

The NB predictor thus dynamically changes its prediction value based on recent data.

## Text Classification Example

The probability of a document d being in class c is computed as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

where  $P(t_k|c)$  is the conditional probability of term  $t_k$  occurring in a document of class c.

Here is text classification training and test data. The goal is to classify the test data into the right class as h or  $\sim h$  (read as not h).

Training set	Document ID	Keywords in the document	Class= h (Healthy)
	1	Love happy joy joy love	Yes
	2	Happy love kick joy happy	Yes
	3	Love move joy good	Yes
	4	love happy joy pain love	Yes
	5	joy love pain kick pain	No
	6	Pain pain love kick	No
Test Data	7	Love pain joy love kick	?

The prior probabilities of a document being classified using the six documents are:

$$P(h) = 4/6 = 2/3$$

$$P(\sim h) = 2/6 = 1/3$$

i.e. there is  $2/3$  prior probability that a document will be classified as h and  $1/3$  probability of not h.

The conditional probabilities for each term is the relative frequency of the

term occurring in each class of documents the h class and the not h class.

Conditional Probabilities	
Class h	Class $\sim h$
$P(\text{Love}   h) = 5 / 19$	$P(\text{Love}   \sim h) = 2 / 9$
$P(\text{pain}   h) = 1 / 19$	$P(\text{pain}   \sim h) = 4 / 9$
$P(\text{joy}   h) = 5 / 19$	$P(\text{joy}   \sim h) = 1 / 9$
$P(\text{kick}   h) = 1 / 19$	$P(\text{kick}   \sim h) = 2 / 9$

The probability of the test instance belonging to class h, can be computed by multiplying the prior probability of the instance belonging to class h, with the conditional probabilities for each of the terms in the document for class h.

Thus:

$$\begin{aligned} P(h | d7) &= P(h) * (P(\text{love} | h))^2 * P(\text{pain} | h) * P(\text{joy} | h) * P(\text{kick} | h) \\ &= (2/3) * (5/19) * (5/19) * (1/19) * (5/19) * (1/19) = \sim 0.0000067 \end{aligned}$$

Similarly, the probability of the test instance being (not h) can be computed using conditional probabilities for not h.

$$\begin{aligned} P(\sim h | d7) &= P(\sim h) * P(\text{love} | \sim h) * P(\text{love} | \sim h) * P(\text{pain} | \sim h) * P(\text{joy} | \sim h) * P(\text{kick} | \sim h) \\ &= (1/3) * (2/9) * (2/9) * (4/9) * (1/9) * (2/9) = 0.00018 \end{aligned}$$

The NB probability of the test instance being (not h) is much higher than its being h. Thus the test document will be classified as (not h).

## Advantages and Disadvantages of Naïve Bayes

1. The NB logic is simple. The NB posterior probability computation is also simple.
2. Conditional probabilities can be computed for discrete data and for probabilistic distributions. When there is number of variables in the vector X, then the problem can be modeled using

probability functions to simulate the incoming values. A variety of methods exist for modeling the conditional distributions of the X variables, including normal, lognormal, gamma, and Poisson

3. Naive Bayes assumes that all the features are independent. For most instances, that works fine. However, it can be a limitation. If there are no joint occurrences at all of a class label with a certain attribute, then the frequency-based conditional probability will be zero. When all the probabilities are multiplied, it will make the entire posterior probability estimate to be zero. This can be rectified by adding 1 to all the numerators and adding n, the number of variables in X, to all the denominators. This will be making those probabilities very small but not zero.
4. A limitation of NB is that the posterior probability computations are good for comparison and classification of the instances. However, the probability values themselves are not good estimates of the event happening.

## Summary

Naïve Bayes is a probability-based machine learning technique used for classifications. It is a mathematically simple way to include the contributions of many factors in predicting the class of the next data instance. It is often used to classifying texts.

## Review Questions

Q1: What is Naïve-Bayes technique? What do Naïve and Bayes stand for?

Q2: In what ways is NB better than other classifications techniques?

Compare with Decision trees.

Q3: What are the most popular applications for NB technique?

# Chapter 13: Support Vector Machines

## Introduction

Support Vector Machine (SVM) is a mathematically rigorous, machine learning technique to build a linear binary classifier. It creates a hyperplane in a high-dimensional space that can accurately slice a dataset into two segments according to the desired objective. The algorithms for developing the classifier can be mathematically challenging though. SVM's are popular since they are the state-of-the-art for many practical problems, such as identifying spam emails and other text mining applications.

### CASELET: Prostate cancer detection using protein biomarkers

Mass spectrometry technology is being used to generate a fast and efficient profile of the protein content of complex samples. It can generate data on up to 40,000 variables that constitute the profile for a protein sample. Data analytics can be used for narrowing the search space for protein biomarker candidates. These spectra can be used to identify peaks on certain biomarkers whose intensities correlate with a particular outcome variable, e.g. prostate cancer. Thorough cross-validation studies and randomization tests are performed on a prostate cancer dataset with over 300 patients, obtained at the Eastern Virginia Medical School using SELDI-TOF mass spectrometry. A two-stage linear SVM-based procedure led to average classification accuracies of 87% on a four-group classification problem, using only 13 variables (or peaks).

	# of peaks used							
	10	15	20	25	30	35	50	70
Quadr. Discr.	74.7 (7.4)	74.7 (9.6)	74.1 (8.4)	74.7 (7.1)	78.2 (6.8)	77.8 (7.3)	78.7 (6.6)	76.8 (7.1)
Nonpar (Kernel)	76.7 (7.1)	77.4 (8.4)	77.7 (6.9)	78.6 (6.6)	80.0 (6.3)	79.9 (7.3)	78.1 (6.5)	76.1 (7.6)
kNN	73.4 (7.4)	76.4 (6.9)	76.9 (6.0)	76.6 (6.1)	75.8 (6.7)	77.2 (6.9)	73.9 (7.5)	69.8 (6.7)
Fisher Linear	72.4 (7.3)	77.3 (6.9)	80.8 (6.5)	80.1 (5.8)	81.8 (6.0)	84.6 (5.2)	85.5 (6.1)	84.3 (5.1)
Linear SVM	75.4 (6.4)	79.3 (7.4)	81.7 (7.2)	81.3 (5.7)	83.7 (6.8)	83.1 (6.6)	83.5 (6.1)	84.0 (6.2)

Source:

<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-26>

Q1: What can be said about the accuracy of various techniques? Does the accuracy increase by increasing the number of variable used in the

analysis?

Q2: As the predictive models get more accurate, what will be the role of doctors? What other skills will they need to learn?

## SVM model

An SVM is a classifier function in a high-dimensional space that defines the decision boundary between two classes. The support vectors are the data points that define the ‘gutters’, or the boundary condition, on either side of the hyperplane, for each of the two classes. The SVM model is thus conceptually easy to understand.

Suppose there is a labeled set of points classified into two classes. The goal is to find the best classifier between the points of the two types.

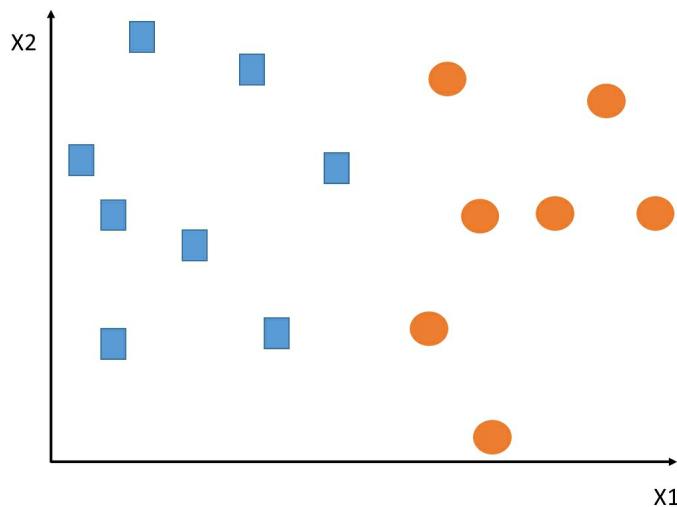
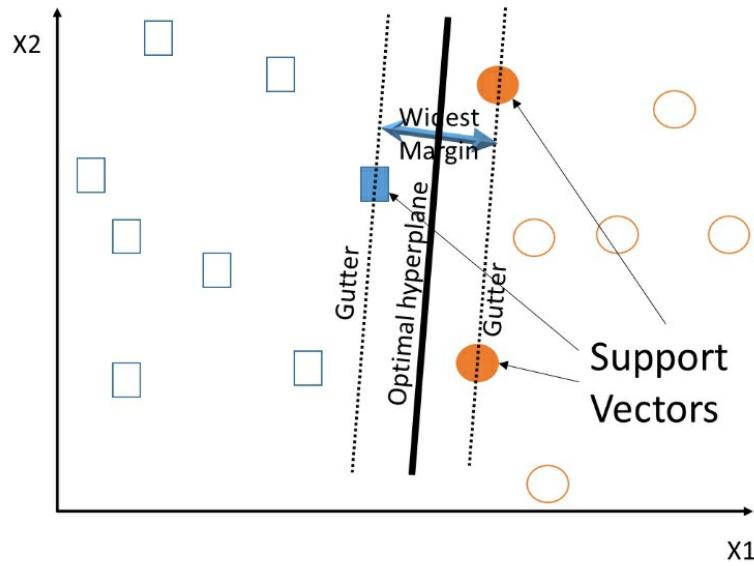


Figure 13.1: Data points for classification

SVM takes the widest ‘street’ (a vector) approach to demarcate the two classes, and thus finds the hyperplane that has the **widest margin**: i.e. largest distance to the nearest training data points of either class (Figure 13.2)

In Figure 13.2, the hard line is the optimal hyperplane. The dotted lines are the gutters on the sides of the two classes. The gap between the gutters is the maximum or widest margin. The classifier (hyperplane) is defined by only those points that fall on the gutters on both sides. These points are called the support vectors (shown in their bold shape). The rest of the data points that safely fit in their class are irrelevant for defining the classifier (now shown unfilled).



**Figure 13.2 Support Vector machine classifier**

Abstractly, suppose that the training data of n points is:

$$(\mathbf{X}_1, y_1), \dots, (\mathbf{X}_n, y_n)$$

Where  $\mathbf{x}_i$  represents the p-value vector for point i, and  $y_i$  is its binary class value of 1 or -1. Thus there are two classes represented as 1 and -1.

Assuming that the data is indeed linearly separable, the classifier hyperplane is defined as a set of points (which is a subset of the training data), that satisfy the equation

$$\mathbf{W} \cdot \mathbf{X} + b = 0$$

Where  $\mathbf{W}$  is the normal vector to the hyperplane.

The hard margins can be defined by the hyperplanes:

$$\mathbf{W} \cdot \mathbf{X} + b = 1 \quad \text{and} \quad \mathbf{W} \cdot \mathbf{X} + b = -1$$

The width of the hard margin is  $(2 / |\mathbf{W}|)$

For all points not on the hyperplane, they will be safely in their own class. Thus the y values will have to be either greater than 1 (for point in class 1), or less than -1 (for points in class -1).

The SVM algorithm finds the weights vector ( $\mathbf{W}$ ) for the features, such that there is a widest margin between the two categories.

Computing an SVM using these equations is a hill-climbing process problem in a convex space. However, by working only with points only nearest to the classifying boundary, it reduces sharply the number of data instances to work with. This approach reduces its memory requirements for computation. This is possible because of using kernel methods.

## The Kernel Method

The heart of an SVM algorithm is kernel methods. Most kernel algorithms are based on optimization in a convex space, and are statistically well-founded.

Kernel stands for the core, or the germ in a fruit. Kernel methods operate using what is called the ‘kernel trick’. This trick involves to compute and work with the inner products of only the relevant pairs of data in the feature space; they do no need to compute all the data in a high-dimensional feature space. The kernel trick makes the algorithm much less demanding in computational and memory resources.

Kernel methods achieve this by learning from instances. They do not apply some standard computational logic to all the features of each input. Instead they remember each training example, and associate a weight representing its relevance to the achievement of the objective. This could be called instance-based learning. There are several types of Support Vector models including linear, polynomial, RBF, and sigmoid.

It is somewhat akin to how human beings learn, especially about domains with a broad sets of features, such as ethics. We learn lessons from tough situations, and discard the ‘normal’ instances. In kernel parlance, we assign high weights to the abnormal situations, and very low weight to the normal situations. The normal situations are almost completely forgotten.

SVMs have evolved to be more flexible and be able to tolerate some amount of misclassification. The margin of separation between the categories is thus a ‘soft margin’, as against a hard margin.

## Advantages and disadvantages

The main strength of SVMs is that it works well even when the number of features is much larger than the number of instances. It can work on datasets with huge feature space, such as is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.

Another advantage of SVMs is that even when the optimal decision boundary is a non-linear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data.

SVMs are conceptually easy to understand. They create an easy-to-understand linear classifier. By working on only a subset of relevant data, they are computationally efficient. SVMs are now available with almost all data analytics toolsets.

The SVM technique has two major constraints. (a) It works well only with real numbers. I.e. all the data points in all dimensions must be defined by numeric values only (b) It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.

Training the SVMs is an inefficient and time-consuming process, when the data is large. It does not work well when there is much noise in the data, and thus has to compute soft margins. The SVMs will also not provide a probability estimate of classification, i.e. the confidence level for classifying an instance.

## Summary

Support Vector machines is a machine learning technique for classifying high-dimensional data into two classes. It creates a hyperplane, with the largest amount of separation between the two classes. The classifier is made linear, by transforming the original features of input data into new features. SVMs use kernel methods to learn from specific instances, that are close to the decision boundary. SVMs are used for text mining, such as spam filtering and outlier detection.

## Review Questions

Q1: What is a Support Vector Machine?

Q2: What are support vectors?

Q3: What is a kernel method?

Q4: Name two advantages and two limitations of SVMs.

## Chapter 14: Web Mining

### Introduction

Web mining is the art and science of discovering patterns and insights from the World-wide [web](#) so as to improve it. The world-wide web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via Web crawlers, web logs, and other means.

Here are some characteristics of optimized websites:

1. *Appearance*: Aesthetic design. Well-formatted content, easy to scan and navigate. Good color contrasts.
2. *Content*: Well planned information architecture with useful content. Fresh content. Search-engine optimized. Links to other good sites.
3. *Functionality*: Accessible to all authorized users. Fast loading times. Usable forms. Mobile enabled.

This type of content and its structure is of interest to ensure the web is easy to use. The analysis of web usage provides feedback on the web content, and also the consumer's browsing habits. This data can be of immense use for commercial advertising, and even for social engineering.

The web could be analyzed for its structure as well as content. The usage pattern of web pages could also be analyzed. Depending upon objectives, web mining can be divided into three different types: Web usage mining, Web content mining and Web structure mining (Figure 12.1).

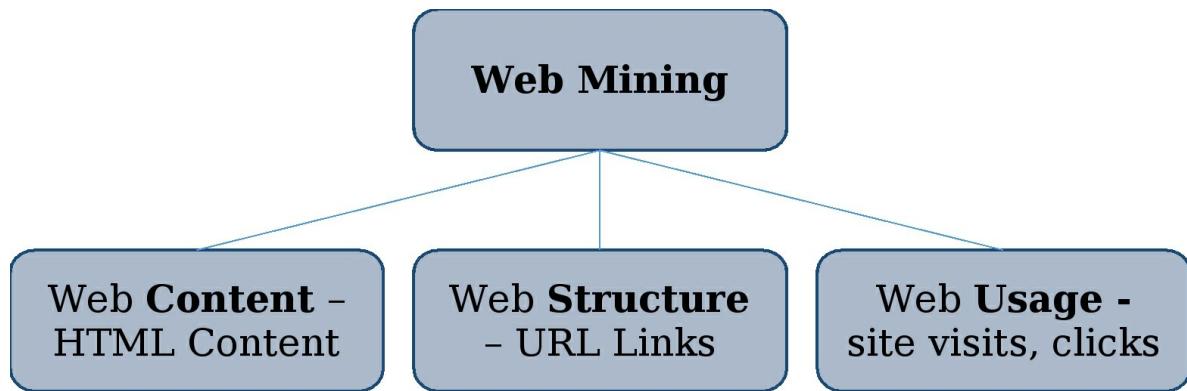


Figure: 12.1 Web Mining structure

## Web content mining

A website is designed in the form of pages with a distinct URL (universal resource locator). A large website may contain thousands of pages. These pages and their content is managed using specialized software systems called Content Management Systems. Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user generated content.

The websites keep a record of all requests received for its page/URLs, including the requester information using ‘cookies’. The log of these requests could be analyzed to gauge the popularity of those pages among different segments of the population. The text and application content on the pages could be analyzed for its usage by visit counts. The pages on a website themselves could be analyzed for quality of content that attracts most users. Thus the unwanted or unpopular pages could be weeded out, or they can be transformed with different content and style. Similarly, more resources could be assigned to keep the more popular pages more fresh and inviting.

## Web structure mining

The Web works through a system of hyperlinks using the hypertext protocol (http). Any page can create a hyperlink to any other page, it can be linked to by another page. The intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages. There are two basic strategic models for successful websites: Hubs and Authorities.

1. *Hubs*: These are pages with a large number of interesting links.

They serve as a hub, or a gathering point, where people visit to access a variety of information. Media sites like Yahoo.com, or government sites would serve that purpose. More focused sites like Traveladvisor.com and yelp.com could aspire to becoming hubs for new emerging areas.

2. *Authorities*: Ultimately, people would gravitate towards pages that provide the most complete and authoritative information on a particular subject. This could be factual information, news, advice, user reviews etc. These websites would have the most number of inbound links from other websites. Thus Mayo clinic.com would serve as an authoritative page for expert medical opinion. NYtimes.com would serve as an authoritative page for daily news.

## Web usage mining

As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click, and the web server providing the content would also make a record of the pages served and the user activity on those pages. The entities between the client and the server, such as the [router](#), [proxy server](#), or [ad server](#), too would record that click.

The goal of web usage mining is to extract useful information and patterns from data generated through Web page visits and transactions. The activity data comes from data stored in server access logs, referrer logs, agent logs, and client-side cookies. The user characteristics and usage profiles are also gathered directly, or indirectly, through syndicated data. Further, metadata, such as page attributes, content attributes, and usage data are also gathered.

The web content could be analyzed at multiple levels (Figure 12.2).

1. The *server side analysis* would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
2. The *client side analysis* could focus on the usage pattern or the actual content consumed and created by users.
  - a. Usage pattern could be analyzed using ‘clickstream’ analysis, i.e. analyzing web activity for patterns of

sequence of clicks, and the location and duration of visits on websites. Clickstream analysis can be useful for web activity analysis, software testing, market research, and analyzing employee productivity.

- b. Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques. The text would be gathered and structured using the bag-of-words technique to build a Term-document matrix. This matrix could then be mined using cluster analysis and association rules for patterns such as popular topics, user segmentation, and sentiment analysis.

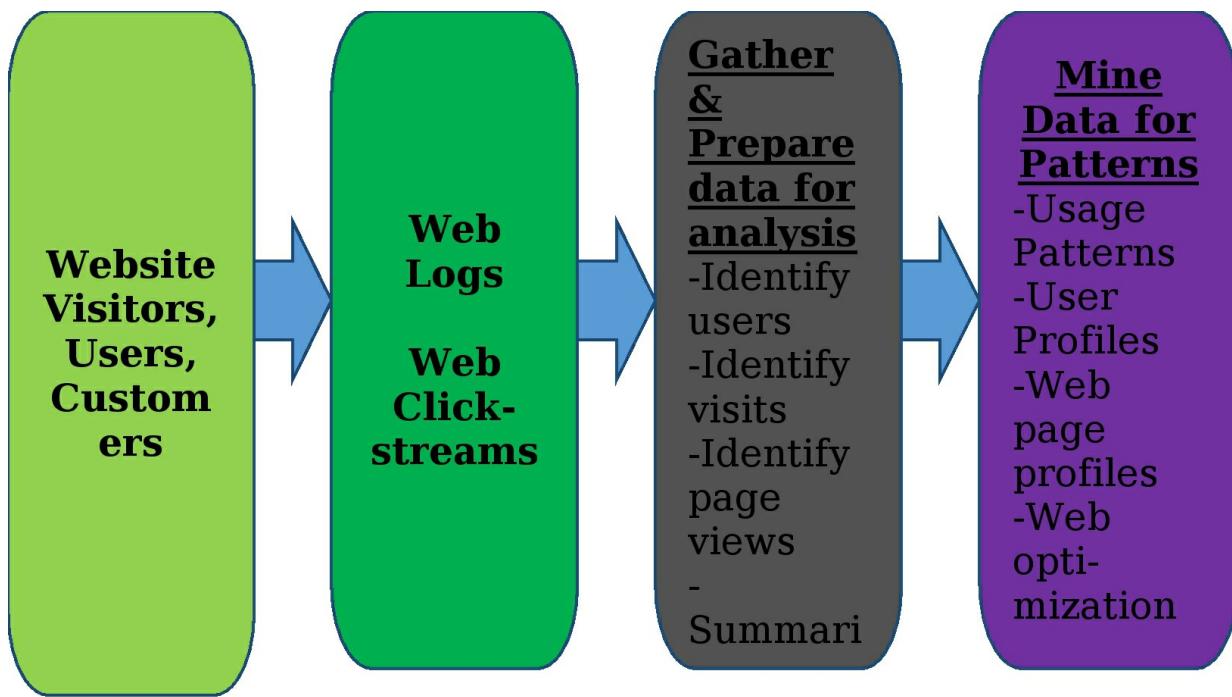


Figure: 12.2 Web Usage Mining architecture

Web usage mining has many business applications. It can help predict user behavior based on previously learned rules and users' profiles, and can help determine lifetime value of clients. It can also help design cross-marketing strategies across products, by observing association rules among the pages on the website. Web usage can help evaluate promotional campaigns and see if the users were attracted to the website and used the pages relevant to the campaign. Web usage mining could be used to present dynamic information to users based on their interests and profiles. This includes targeted online ads

and coupons at user groups based on user access patterns.

## Web Mining Algorithms

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities. Many other HITS-based algorithms have also been published. The most famous and powerful of these algorithms is the PageRank algorithm. Invented by Google co-founder Larry Page, this algorithm is used by Google to organize the results of its search function. This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page. The websites with more number of links, and/or more links from higher-quality websites, will be ranked higher. It works in a similar way as determining the status of a person in a society of people. Those with relations to more people and/or relations to people of higher status will be accorded a higher status.

PageRank is the algorithm that helps determine the order of pages listed upon a Google Search query. The original PageRank algorithm formulation has been updated in many ways and the latest algorithm is kept a secret so other websites cannot take advantage of the algorithm and manipulate their website according to it. However, there are many standard elements that remain unchanged. These elements lead to the principles for a good website. This process is also called Search Engine Optimization (SEO).

## Conclusion

The web has growing resources, with more content everyday and more users visiting it for many purposes. A good website should be useful, easy to use, and flexible for evolution. From the insights gleaned using web mining, websites should be constantly optimized.

Web usage mining can help discover what content users really like and consume, and help prioritize that for improvement. Web structure can help improve traffic to those sites, by building authority for the sites.

## Review Questions

- 1: What are the three types of web mining?
- 2: What is clickstream analysis?
- 3: What are the two major ways that a website can become popular?

4: What are the privacy issues in web mining?

5: A user spends 60 minutes on the web, visiting 10 webpages in all. Given the clickstream data, what kind of an analysis would you do?

## Chapter 15: Social Network Analysis

### Introduction

Social Networks are a graphical representation of relationships among people and/or entities. Social network analysis (SNA) is the art and science of discovering patterns of interaction and influence within the participants in a network. These participants could be people, organizations, machines, concepts, or any other kinds of entities. An ideal application of social network analysis will discover essential characteristics of a network including its central nodes and its sub-network structure. Sub-networks are clusters of nodes where the within-sub-network connections are stronger than the connections with nodes outside the sub-network. SNA is accomplished through graphically representing social relationships into a network of nodes and links, and applying iterative computational techniques to measure the strengths of relationships. The social network analysis ultimately helps relate the totality of network to the Unified Field which is the ultimate entity with infinite relationships among everything.

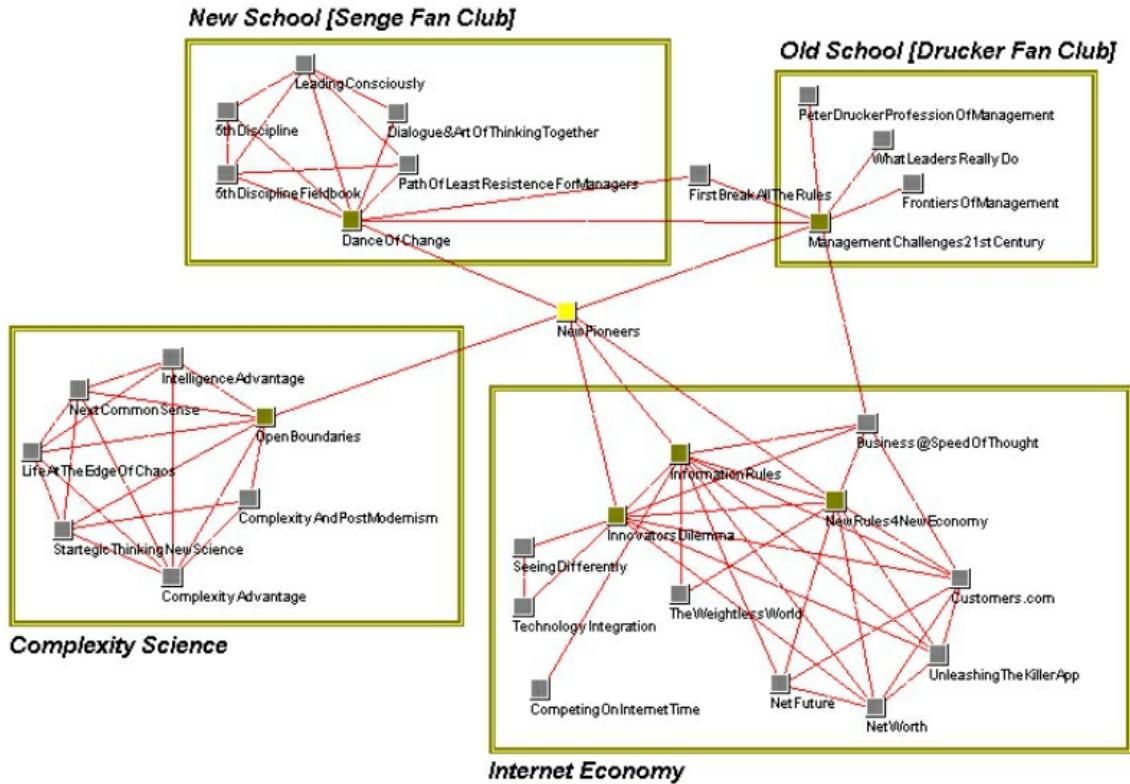
**Caselet: The Social Life of Books: Visualizing Communities of Interest via Purchase Patterns on the WWW (source: [orgnet.com/booknet.html](http://orgnet.com/booknet.html))**

*One of Amazon's value-added service is called -- Customers who bought this book also bought these books. Amazon lists the top 6 books that were bought by individuals who also bought the book currently being browsed. Let's choose Tom Petzinger's The New Pioneers. What themes would we see... in the recommended books? What other topics are Tom's readers interested in? Will Tom's book end up in the center of one large, massively interconnected cluster -- a single community of interest? Or, will it end up linking together otherwise disconnected clusters -- diverse communities of interest?*

*Below is the network surrounding 'The New Pioneers'. Each node represents a book. A red line links books that were purchased together. The buying pattern of the books has self-organized into emergent clusters that I have named for the content of each cluster. It is obvious that Tom's book does span a diversity of interests!*

*The most common measure in social networks is network centrality.*

*To assess 'positional advantage' we measure each node's network centrality. We have two parts of the network 1) the dense Complexity Science cluster and 2) the dense Internet economy clusters, and 3) the other 2 interconnected clusters forming a large network component.*



### Questions:

*Q1: Can these maps could provide an opportunity to the businesses that Amazon competes against? To authors and publicists?*

*Q2: Could a similar map of services be developed for a healthcare situation? How?*

## Applications of SNA

Self-awareness: Visualizing his/her social network can help a person organize their relationships and their support network.

Communities: Social Network Analysis can help identification, construction, and strengthening of networks within communities to build wellness, comfort and resilience. Analysis of joint authoring relationships and citations help identify sub-networks of specializations of knowledge in an academic field. Researchers at Northwestern University found that the most determinant

factor in the success of a Broadway play was the strength of relationships amongst the crew and cast.

**Marketing:** There is a popular network insight that any two people are related to each other through at most seven degrees of links. Organizations can use this insight to reach out with their message to large numbers of people, and also to listen actively to opinion leaders as ways to understand their customers' needs and behaviors. Politicians can reach out to opinion leaders to get their message out.

**Public Health:** Awareness of networks can help identify the paths that certain diseases take to spread. Public health professionals can isolate and contain diseases before they expand to other networks.

## **Network topologies**

There are two primary types of network topologies: the ring-type and hub-spoke topologies. Each of the topologies has different characteristics and benefits.

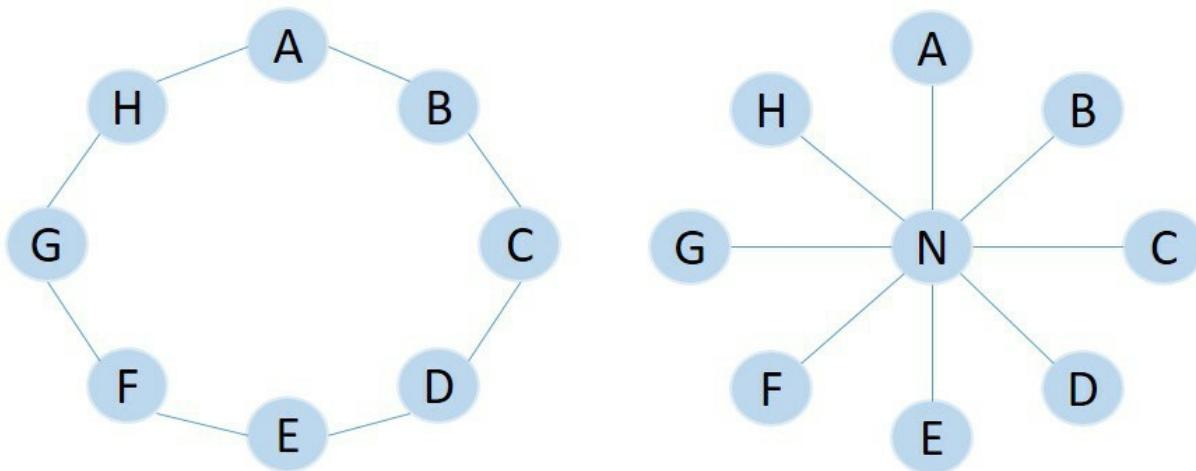
In the ring network, nodes typically connect to their adjacent nodes in the network. Potentially all nodes can be connected to each other. A ring network could be dense where every node has a direct connection with practically every node. Or it could be sparse where every node connects to a small subset of the nodes. The number of connections needed to reach from one node to another could vary tremendously. A dense network, with more connections, will allow many direct connections between pairs of nodes. In a sparse network, one may need to traverse many connections to reach the desired destination. A peer-to-peer email (or messaging) network is an example on the ring model, as anyone can potentially directly connect with anyone else. In practice the email networks are sparse as people connect directly with only a subset of people.

In the hub-spoke model, there is one central hub node to which all the other nodes are connected. There are no direct relationships between the nodes. Nodes connect with each other through the hub node. This is a hierarchical network structure since the hub node is central to the network. The hub node is structurally more important, as it is central to all communications between the other peripheral nodes. The good thing about hub-spoke network is that one could predictably reach from any node to any other node through traversing exactly just two connections. As an example, modern airlines

operate on this model to maintain hub networks from which they operate flights to a large number of airports.

The density of a network can be defined as the average number of connections per node. The cohesiveness of the network is a related concept, which is the average number of connections needed to reach from one node to the other.

Another way of analyzing networks is to define the centrality (or importance) of every node. The number of links associated with a node is a sign of centrality of the node. In the Ring network in the figure below, each node has exactly 2 links. Thus, there is no central node. However, in the hub-spoke network, the hub-node N has 8 links while all other nodes have only 1 link each. Thus node N has a high centrality.



**Figure 13-0-1: Network Topologies: Ring (left) and Hub-spoke (right)**

A variant that combines both the above types, is the network of networks, in which each participating network will connect with other networks at select points of contact. For example, the internet is a network of networks. Here all the commercial and university and government and similar computer networks connect to one other at certain designated nodes (called gateways) to exchange information and conduct business.

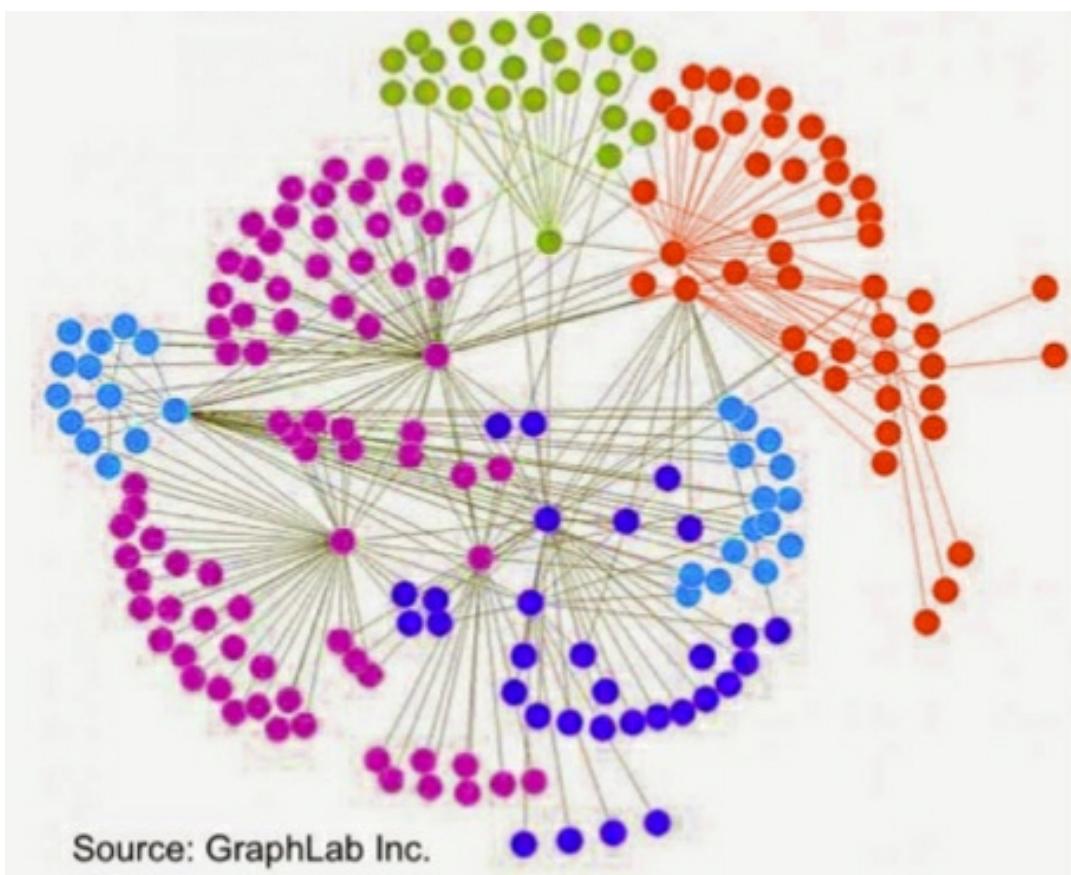
## Techniques and algorithms

There are two major levels of social network analysis: discovering sub-networks within the network, and ranking the nodes to find more important

nodes or hubs.

#### Finding Sub-networks

A large network could be better analyzed and managed if it can be seen as an inter-connected set of distinct sub-networks each with its own distinct identity and unique characteristics. This is like doing a cluster analysis of nodes. Nodes with strong ties between them would belong to the same sub-network, while those with weak or no ties would belong to separate sub-networks. This is unsupervised learning technique, as apriori there is no correct number of sub-networks in a network. The usefulness of the sub-network structure for decision-making is the main criterion for adopting a particular structure.



**Figure 13-0-2: A network with distinct sub-networks**

Visual representations of networks can help identify sub-networks. Use of color can help differentiate different types of nodes. Representing strong ties with thicker or bolder lines could help visually identify the stronger relationships. A sub-network could be a collection of strong relationships.

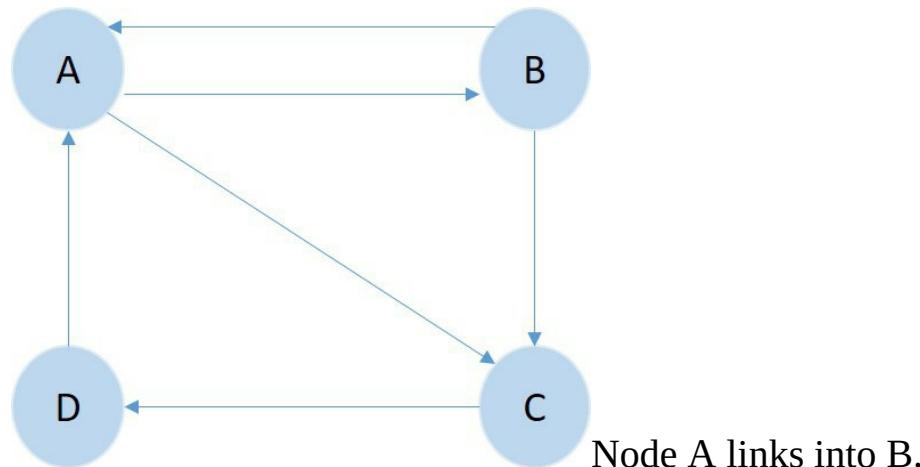
around a hub node. In that case, the hub node could represent a distinct sub-network. A sub-network could also be a subset of nodes with dense relationships between them. In that case, one or more nodes will act as gateways to the rest of the network.

#### Computing importance of nodes

When the connections between nodes in the network has a direction to them, then the nodes can be compared for their relative influence or rank. (The terms importance and influence and rank of a node will be used interchangeably for the purpose of this discussion.) This is done using a ‘Influence Flow model’. Everyone outbound link from a node can be considered an outflow of influence., Every incoming link is similarly an inflow of influence. More in-links to a node means greater importance. Thus there will be many direct and indirect flows of influence between any two nodes in the network.

Computing the relative influence of each node is done on the basis of an input-output matrix of flows of influence among the nodes. Assume each node has a influence value. The computational task is to identify a set of rank values that satisfies the set of links between the nodes. It is an iterative task where we begin with some initial values and continue to iterate till the rank values stabilize.

Consider the following simple network with 4 nodes (A, B,C,D) and 6 directed links between them. Note that there is a bidirectional link. Here are the links:



Node B links into C.

Node C links into D.

Node D links into A.

Node A links into C.

Node B links into A.

The goal is to find the relative importance, or rank, of every node in the network. This will help identify the most important node(s) in the network.

We begin by assigning the variables for influence (or rank) value for each node, as Ra, Rb, Rc, and Rd. The goal is to find the relative values of these variables.

There are two outbound links from node A to nodes B and C. Thus both B and C receive half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A. Thus both C and A receive half of node B's influence. There is only outbound link from node D, into node A. Thus Node A gets all the influence of Node D. There is only outbound link from node C, into node D. Thus Node D gets all the influence of Node C.

Node A gets all of the influence of node D, and half the influence of Node B.

$$\text{Thus } Ra = 0.5 * Rb + Rd$$

Node B gets half the influence of node A.

$$\text{Thus } Rb = 0.5 \times Ra$$

Node C gets half the influence of Node A, and half the influence of Node B.

$$\text{Thus } Rc = 0.5 \times Ra + 0.5 \times Rb$$

Node D gets all of the influence of Node C, and half the influence of Node B.

$$\text{Thus } Rd = Rc$$

We have 4 equations using 4 variables. These can be solved mathematically.

We can represent the coefficients of these 4 equations in a matrix form as below. This is the Influence Matrix. The zero values represent that the term is not represented in an equation.

	Ra	Rb	Rc	Rd
Ra	0	0.50	0	1.00

Rb	0.50	0	0	0
Rc	0.50	0.50	0	0
Rd	0	0	1.00	0

For simplification, let us also state that all the Rank values add up to 1. Thus each node has a fraction as the Rank value. Let us start with an initial set of Rank values, and then iteratively compute new Rank values till they stabilize. One can start with any initial Rank values, such as  $1/n$  or  $\frac{1}{4}$  for each of the nodes.

Variable	InitVal
Ra	0.250
Rb	0.250
Rc	0.250
Rd	0.250

Computing the revised values using the equations stated earlier, we get a revised set of values shown as Iteration1. [This can be computed easily by creating formulae using the Influence matrix in a spreadsheet such as Excel].

Variable	InitVal	Iteration1
Ra	0.250	0.375
Rb	0.250	0.125
Rc	0.250	0.250
Rd	0.250	0.250

Using the Rank values from Iteration1 as the new starting values, we can compute new values for these variables, shown as Iteration 2. Rank values will continue to change.

Variable	InitVal	Iteration1	Iteration2
Ra	0.250	0.375	0.3125
Rb	0.250	0.125	0.1875
Rc	0.250	0.250	0.250

Rd	0.250	0.250	0.250
----	-------	-------	-------

Working from values of Iteration2, and so, we can do a few more iterations till the values stabilize. Here are the final values after the 8<sup>th</sup> iteration.

Variable	InitVal	Iteration1	Iteration2	....	Iteration8
Ra	0.250	0.375	0.313	....	0.333
Rb	0.250	0.125	0.188	....	0.167
Rc	0.250	0.250	0.250	....	0.250
Rd	0.250	0.250	0.250	....	0.250

The final rank shows that Rank of node A is the highest at 0.333. Thus the most important node is A. The lowest rank is 0.167, of Rb. Thus B is the least important node. Nodes C and D are in the middle. In this case, their ranks did not change at all.

The relative scores of the nodes in this network would have been the same irrespective of the initial values chosen for the computations. It may take longer or shorter number of iterations for the results to stabilize for different sets of initial values.

## PageRank

PageRank is a particular application of the social network analysis techniques above to compute the relative importance of websites in the overall world wide web. The data on websites and their links is gathered through web crawler bots that traverse through the webpages at frequent intervals. Every webpage is a node in a social network, and all the hyperlinks from that page become directed links to other webpages. Every outbound link from a webpage is considered an outflow of influence of that webpage. An iterative computational technique is applied to compute a relative importance to each page. That score is called PageRank, according to an eponymous algorithm invented by the founders of Google, the web search company.

PageRank is used by Google for ordering the display of websites in response to search queries. To be shown higher in the search results, many website owners try to artificially boost their PageRank by creating many dummy websites whose ranks can be made to flow into their desired website. Also,

many websites can be designed to cyclical sets of links from where the web crawler may not be able to break out. These are called spider traps.

To overcome these and other challenges, Google includes a Teleporting factor into computing the PageRank. Teleporting assumed that there is a potential link from any node to any other node, irrespective of whether it actually exists. Thus the Influence matrix is multiple by a weighting factor called Beta, with a typical value of 0.85 or 85%. The remaining weight of 0.15 or 15% is given to Teleportation. In the Teleportation matrix, each cell is given a rank of  $1/n$ , where n is the number of nodes the web. The two matrices are added to compute the final Influence matrix. This matrix can be used to iteratively compute the PageRanks of all the nodes, just as shown in the example earlier.

## Practical considerations

1. Network size: Most SNA research is done using small networks. Collecting data about large networks can be very challenging. This is because the number of links is the order of the square of the number of nodes. Thus in a network of 1000 nodes there are potentially 1 million possible pairs of links.
2. Gathering Data: Electronic communication records (emails, chats, etc) can be harnessed to gather social network data more easily. Data on the nature and quality of relationships need to be collected using survey documents. Capturing and cleansing and organizing the data can take a lot of time and effort, just like in a typical data analytics project.
3. Computation and Visualization: Modeling large networks can be computationally challenging, and visualizing them also would require special skills. Big data analytical tools may be needed to compute large networks.
4. Dynamic networks: Relationships between nodes in a social network can be fluid. They can change in strength and functional nature. For example, there could be multiple relationships between two people ... they could simultaneously be co-workers, co-authors, and spouses. The network should be modeled frequently to see the dynamics of the network.

## Comparing SNA with Data Analytics

Dimension	Social Network Analysis	Traditional Data Mining
Nature of learning	Unsupervised learning	Supervised and unsupervised learning
Analysis goals	Hub nodes, important nodes, and sub-networks	Key decision rules, cluster centroids
Dataset structure	A graph of nodes and (directed) links	Rectangular data of variables and instances
Analysis techniques	Visualization with statistics; iterative graphical computation	Machine learning, statistics
Quality measurement	Usefulness is key criterion	Predictive accuracy for classification techniques

**Table 3: Social Network Analysis vs Traditional Data Analytics**

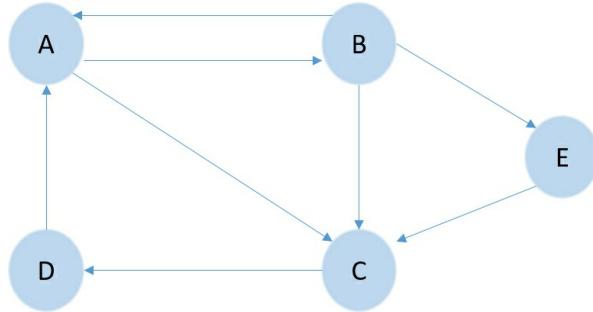
## Conclusion

Social network analysis is a powerful method of analyzing relationships among entities to identify strong patterns. There could be sub-networks in a network based on strong ties within the network. A computationally rigorous set of techniques can be used to rank every node in a network for its influence and importance. PageRank is the implementation of this algorithm in the context of ranking websites.

## Review Questions

1. What is social network analysis? How is it different from other data mining techniques such as clustering or decision trees?
2. How can SNA help with improving gross national happiness of any country or society?
3. What kind of pitfalls should one guard against in doing SNA?
4. Data preparation takes more than 2/3 of the total time of an analytics project. Would that be true for SNA also?
5. Compute the rank values for the nodes for the following network which is a modified version of the exercise solved earlier. Which

is the highest ranked node now.



## **Section 4 - Primers**

This section covers high-level primes on additional topics.

Chapter 16 covers Big Data. This is a new moniker created to describe the phenomenon of large amounts of data being generated from many data sources, and which cannot be handled with the traditional data management tools.

Chapter 17 covers Data Modeling. This is useful as a ramp-up to data mining, especially for those who have not had much exposure to traditional data management or may need a refresher.

Chapter 18 covers Statistics. This is useful as a ramp-up to data mining, for those who may not have had much exposure to statistics, or may need a refresher.

Chapter 19 is a new primer on Artificial Intelligence. AI includes Machine learning which is the source of many Data Mining techniques. AI also includes other fields like Natural language understand, Expert Systems, and more.

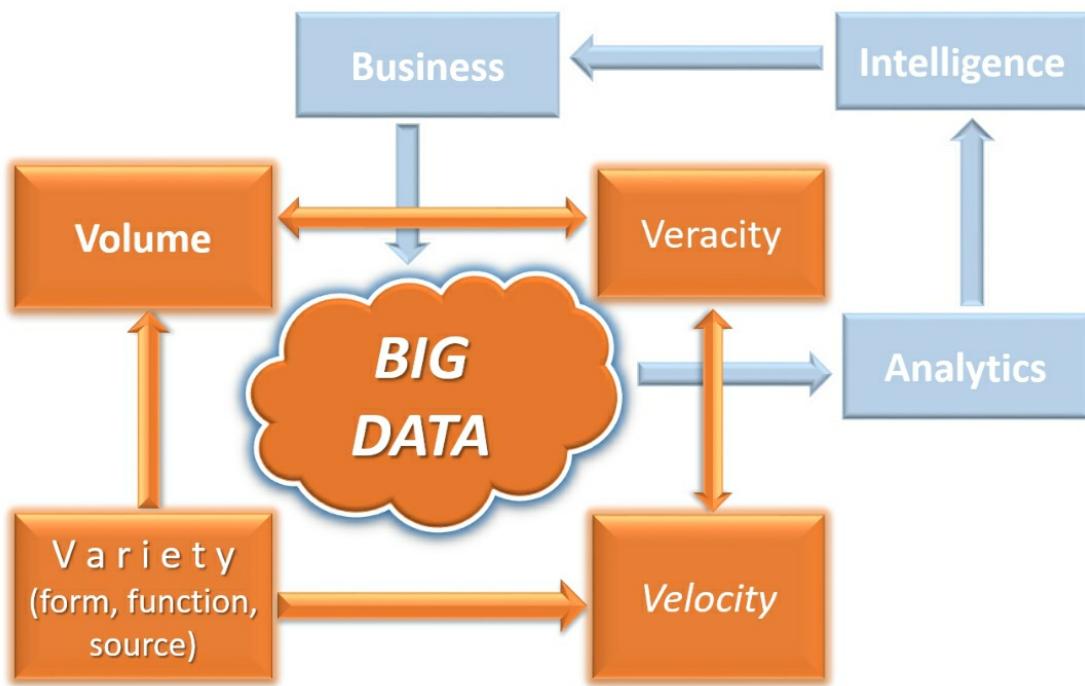
Chapter 20 offers a brief view of Data Science careers and the skills needed. It also includes some additional data analytics case studies.

Appendix A include an extensive R tutorial to help with many of the techniques shown in the book. No prior experience with R is needed for using this tutorial.

# Chapter 16: Big Data Primer

## Introduction

Big Data is an all-inclusive term that refers to extremely large, very fast, highly diverse, and complex data that cannot be managed with traditional data management tools. Ideally, Big Data includes all kinds of data, and helps deliver the right information, to the right person, in the right quantity, at the right time, to help make the right decisions. Big Data can be harnessed by developing infinitely scalable, totally flexible, and evolutionary data architectures, coupled with the use of cost-effective computing machines. The infinite potential knowledge embedded within this Big Data cosmic computer can help connect with and enjoy the support of all the laws of nature.



**Figure 16-1: Big Data Context**

This book will provide a complete overview of Big Data for the executive and the data scientist. This opening chapter will cover the key challenges and benefits of Big Data, and the essential tools and technologies for organizing and benefitting from Big Data.

## Understanding Big Data

Big Data can be examined on two levels (Figure 16.1). At a fundamental level, it is just another collection of data that can be analyzed and utilized for the benefit of the business. On another level, it is a special kind of data that poses unique challenges and offers unique benefits. This is the level that this book will focus on.

At the level of business, data generated by business operations, can be analyzed to generate insights

that can help the business make better decisions. This makes the business grow bigger, and generate even more data, and the cycle continues. This is represented by the cycle on the top-right of Figure 1.1. This aspect is discussed in Chapter 10, a primer on Data Analytics.

On another level, Big Data is different from traditional data in every way: space, time, and function. The quantity of Big Data is 1000s of times greater than that of traditional data. The speed of data generation and transmission is 1,000 times faster. The forms and functions of Big Data are 10 times more diverse: from numbers to text, pictures, audio, videos, web logs, machine data, and more. There are many more sources of data, from individuals to organizations to governments, using a range of devices from mobile phones to computers to industrial machines. Not all of Big Data is of equal quality and value. Big Data is represented by the part of the picture on the bottom left of Figure 16.1. This aspect of Big Data, and its new technologies, is the main focus of this book.

Big Data is mostly, over 90%, unstructured data. Every type of Big Data is structured differently, and therefore has to be dealt with its own appropriate way. There are huge opportunities for technology providers to innovate and manage the entire life cycle of Big Data ... to generate, gather, store, organize, analyze, and visualize this data.

## CASELET: IBM Watson: A Big Data system

*IBM created the Watson system as a way of pushing the boundaries of Artificial Intelligence and natural language understanding technologies. Watson beat the world champion human players of Jeopardy (quiz style TV show) in Feb 2011. Watson reads up data about everything on the web including the entire Wikipedia. It digests and absorbs the data based on simple generic rules such as: books have authors; stories have heroes; and drugs treat ailments. A Jeopardy clue, received in the form of a cryptic phrase, is broken down into many possible potential sub-clues of the correct answer. Each sub-clue is examined to see the likeliness of its answer being the correct answer for the main problem. Watson calculates the confidence level of each possible answer. If the confidence level reaches more than a desirable threshold level, Watson decides to offer the answer to the clue. It manages to do all this in a mere 3 seconds.*

*Watson is now being applied to many important applications such as diagnosing diseases, especially cancer. Watson can read all the new research published in the medical journals to update its knowledge base. It can diagnose the probability of various diseases, by applying factors such as patient's current symptoms, health history, genetic history, medication records, and other factors to recommend a particular diagnosis. (Source: Smartest machines on Earth: youtube.com/watch?*

v=TCOhyaw5bwg)



Figure 16.2: IBM Watson playing Jeopardy (Source: IBM)

*Q1: What kinds of Big Data knowledge, technologies and skills are required to build a system like Watson? What other resources are needed?*

*Q2: Will doctors be able to compete with Watson in diagnosing diseases and prescribing medications? Who else could benefit from a system like Watson?*

## Capturing Big Data

If data was only growing in size, OR only moving at greater speeds, OR only becoming too diverse, it would have been relatively easy. However, when three Vs (Volume, Velocity, Variety) arrive together in an interactive manner, it creates a perfect storm. While the Volume and Velocity of data drive the major technological concerns and the costs of managing Big Data, these two Vs are themselves driven by the 3<sup>rd</sup> V, the Variety of forms and functions and sources of data. The varying accuracy (veracity) and value of data complicate the situation further.

### Volume of Data

The quantity of data generated in the world has been relentlessly doubling every 12-18 months. Traditional data is measured in Gigabytes (GB) and Terabytes (TB), but Big Data is measured in Petabytes (PB) and Exabytes (1 Exabyte = 1 Million TB). This data is so huge that it is almost a miracle that one can find any specific thing in it, in a reasonable period of time. Searching the world-

wide web was the first true Big Data application. Google perfected the art of this application, and developed many of the path-breaking Big Data technologies we see in use today.

The primary reason for the growth of data is the dramatic reduction in the cost of storing data. The costs of storing data have decreased by 30-40% every year. Therefore, there is an incentive to record everything that can be observed. It is called ‘datafication’ of the world. The costs of computation and communication of data have also been coming down. Another reason for the growth of data is the increase in the number of forms and functions of data. That will be discuss further in the Variety section.

### Velocity of Data

If traditional data is like a calm lake, Big Data is like a fast and furious river. Big Data is being generated by billions of devices, and communicated at the speed of the light, through the internet. Ingesting all this data is like drinking from a fire hose. One does not have any control over how fast the data will come. A huge unpredictable data-stream is the new metaphor for thinking about Big Data.

The primary reason for the increased velocity of data is the increase in internet speed. Internet speeds available to homes and offices are now increasing from 10MB/sec to 1 GB/sec (100 times faster). More people are getting access to high-speed internet around the world. Another important reason is the increased variety of sources, such as mobile devices, that can generate and communicate data from anywhere, at any time.

### Variety of Data

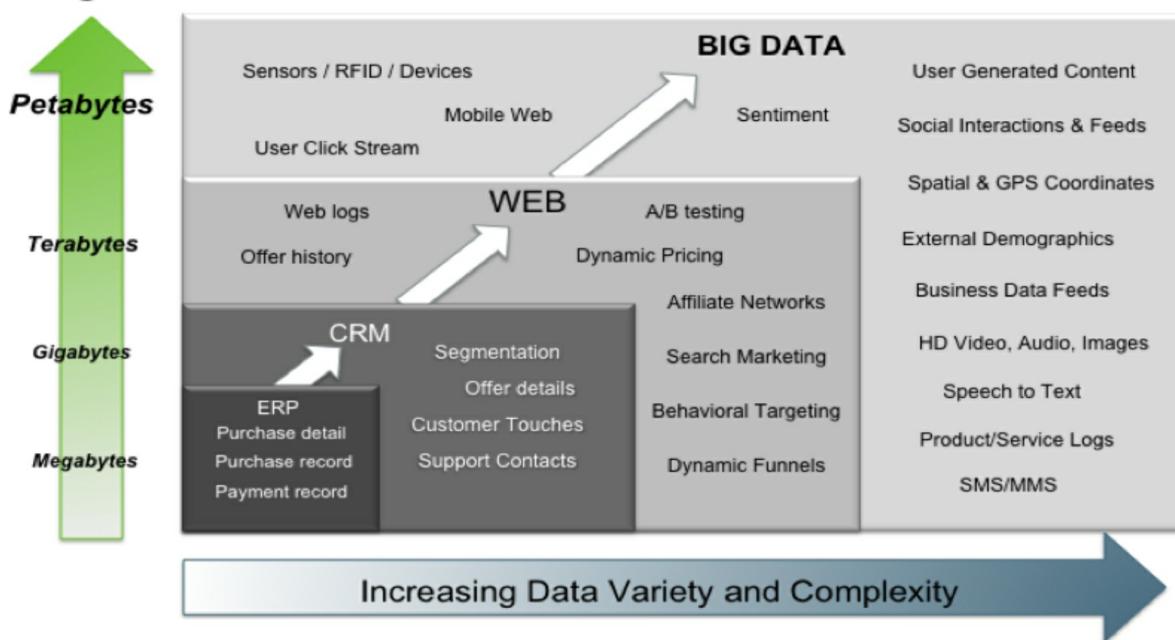
Big Data is inclusive of all forms of data, for all kinds of functions, from all sources and devices. If traditional data from as invoices and ledgers were like a small room full of data, Big Data is like the biggest imaginable shopping mall that offers unlimited variety. There are three major kinds of variety of data.

1. The first aspect of variety is the **Form** of data. Data types range from numbers to text, graph, map, audio, video, and others. Some of these types of data are simple while others are very complex. There are also composite data types that includes many elements in a single file. For example, text documents include graphs and pictures embedded in them. Video movies include audio songs embedded in them. Audio and video have different and more complex storage formats than numbers and text. Numbers and text can be more easily analyzed than an audio or video file.
2. The second aspect of variety is the **Function** of data. There is data from human conversations, songs and movies, business transaction records, machine and operations performance data, new product design data, old archived data, etc. Human communication data would need to be processed very differently from operational performance data, with totally different expectations and objectives. Big Data technologies could be used to recognize people’s faces in pictures; and compare voices

to identify the speaker; and compare handwritings to identify the writer.

- The third aspect of variety is the **Source** of data. Broadly speaking, there are three types of sources of data: Human-human communications; human-machine communications; and machine-to-machine communications. Mobile phone and tablet devices enable a wide series of applications (or apps) to access data. They also generate data from anytime anywhere. Web access and search logs are another new and huge source of data. Business systems generate massive amounts of structured business transactional information. Temperature and pressure sensors on machines, and Radio Frequency (RFID) tags on assets, generate incessant and repetitive data. The sources of Big Data, and their business applications, will be discussed in the next chapter.

## Big Data = Transactions + Interactions + Observations



*Source:* Contents of above graphic created in partnership with Teradata, Inc.

Figure 16.3 Sources of Big Data (Source: Hortonworks.com)

### Veracity of Data

Veracity relates to the truthfulness, believability and quality of data. Big Data is messy. There is a lot of misinformation and disinformation out there. The reasons for poor quality of data can range from technical error, to human error, to malicious intent.

- The source of information may not be authoritative. For example, all websites are not equally trustworthy. Any information from whitehouse.gov or from nytimes.com is more likely to be authentic and complete. Wikipedia is useful, but not all pages are equally reliable. In each case, the communicator may have an agenda or a point of view.

2. The data may not be communicated and received correctly because of human or technical failure. Sensors and machines for gathering and communicating data may malfunction and may record and transmit incorrect data. Urgency may require the transmission of the best data available at a point in time. Such data makes reconciliation with later, accurate, records more problematic.
3. The data provided and received, may however also be intentionally wrong, for competitive or security reasons. There could be disinformation and malicious information spread for strategic reasons.

Big Data needs to be sifted and organized by quality, for it to be put to any great use.

## Benefitting from Big Data

Data is the new natural resource. Traditionally, data usually belongs to the organization that generates it. There are other types of data, such as social media data, that are freely accessible under an open general license. Organizations can use this data to learn about their consumers, improve their service delivery, and design new products to delight their customers and gain a competitive advantage. Data is also being used to create new digital products, such as on-demand entertainment and learning.

Organizations cannot afford to ignore Big Data. They may choose to gather and store this data for later analysis, or to sell it to other organizations, that might benefit from it. They may also legitimately choose to discard parts of their data for privacy or legal reasons. However, organizations that do not learn to engage with Big Data, could find themselves left far behind their competition, landing in the dustbin of history. Innovative and nimble small organizations can use Big Data to quickly scale up and beat larger and more mature organizations.

Big Data applications exist in all industries and aspects of life. There are three major types of Big Data applications: Monitoring and Tracking, Analysis and Insight, and Digital Product Development.

**Monitoring and Tracking Applications:** Consumer goods producers use monitoring and tracking applications to understand the sentiments and needs of their customers. Industrial organizations use Big Data to track inventory in massive interlinked global supply chains. Factory owners use it to monitor machine performance and do preventive maintenance. Utility companies use it to predict energy consumption, and manage demand and supply. Information Technology companies use it to track website performance and improve its usefulness. Financial organizations use it to project trends better and make more effective and profitable bets, etc.

**Analysis and Insight:** Political organizations use Big Data to micro-target voters and win elections. Police use Big Data to predict and prevent crime. Hospitals use it to better diagnose diseases and make medicine prescriptions. Advertisement agencies use it to design more targeted marketing campaigns more quickly. Fashion designers use it to track trends and create more innovative products.



**Figure16.4: The first Big Data President (Source: washingtonpost.com)**

**New Product Development:** Incoming data could be used to design new products such as reality TV entertainment. Stock market feeds could be a digital product. Imagination is the limit on how new products and services can be developed and delivered at the speed of thought.

## Management of Big Data

Big Data is a new phase representing a digital world. Business and society are not immune from its impacts. Many organizations have started initiatives around the use of Big Data. However, most organizations do not necessarily have a good grip on it. Here are some emerging insights into making better use of Big Data.

1. Across all industries, the business case for Big Data is strongly focused on addressing customer-centric objectives. The first focus on deploying Big Data initiatives is to protect and enhance customer relationships and customer experience.
2. Big Data should be used to solve a real pain-point. It should be deployed for specific business objectives in order to have management avoid being overwhelmed by the enormity of Big Data.
3. Organizations are beginning their pilot implementations by using existing and newly accessible internal sources of data. It is better to begin with data under one's control and where one has a superior understanding of the data. One should use more diverse data, and not just more data. This would permit tuning into a broader perspective of reality, leading to better insights.
4. Putting humans and data together leads to the most insights. Combining data-based analysis with human intuition and perspectives is better than going just one way. Most

organizations lack the advanced analytical capabilities required to get the most value from Big Data. There is a growing awareness of building or hiring those skills and capabilities.

5. The faster one analyzes the data, the more will be its predictive value. The value of data depreciates rapidly with time. Many kinds of data need to be processed within seconds, or else the immediate competitive advantage is lost.
6. One should not throw away data if no immediate use can be seen for it. Data usually has value beyond what one initially anticipates. Data can add perspective to other data later in a multiplicative manner. One should maintain just one copy of your data, not multiple. This would help avoid confusion and increase efficiency.
7. Big Data is growing exponentially, so one should plan for exponential growth. Storage costs continue to fall, data generation continues to grow, data-based applications continue to grow in capability and functionality.
8. Big Data builds upon a resilient, secure, efficient, flexible, and real-time information processing environment. Big Data is transforming business, just like IT did.

## Organizing Big Data

Good organization depends upon the purpose of the organization.

Given huge quantities of data, it would be desirable to organize the data to speed up the search process for finding a specific desired thing in the entire data. The cost of storing and processing the data, too, would be a major driver for the choice of an organizing pattern.

Given fast and variable speed of data, it would be desirable to create a scalable number of ingest points. It will also be desirable to create at least some element of control over the data by maintaining count and averages over time, unique values received, etc.

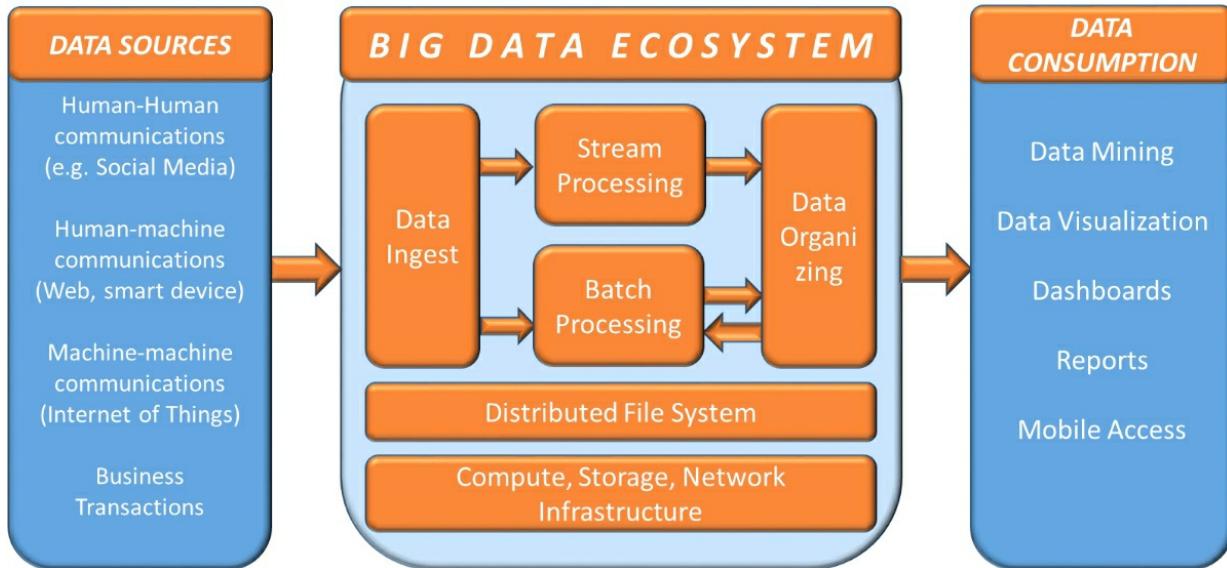
Given wide variety in form factors, data will need to be stored and analyzed differently. Videos need to be stored separately and used for serving in a streaming mode. Text data may be stored separately, so it can be combined, cleaned, and analyzed for themes and sentiments.

Given different quality of data, various data sources may need to be ranked and prioritized before serving them to the audience. For example, the quality of a webpage and its data may be evaluated using its PageRank value.

## Analyzing Big Data

Big Data can be analyzed in two ways. Big Data can be utilized to visualize a flowing or a static situation. These are called analyzing Big Data in motion or Big Data at rest. First way is to process the incoming stream of data in real time for quick and effective statistics about the data. The second way is

to store and structure batches of data and apply standard analytical techniques on for generating insights. This could then be visualized using real-time dashboards. The nature of processing this huge, diverse, and largely unstructured data, can be limited only by one's imagination.



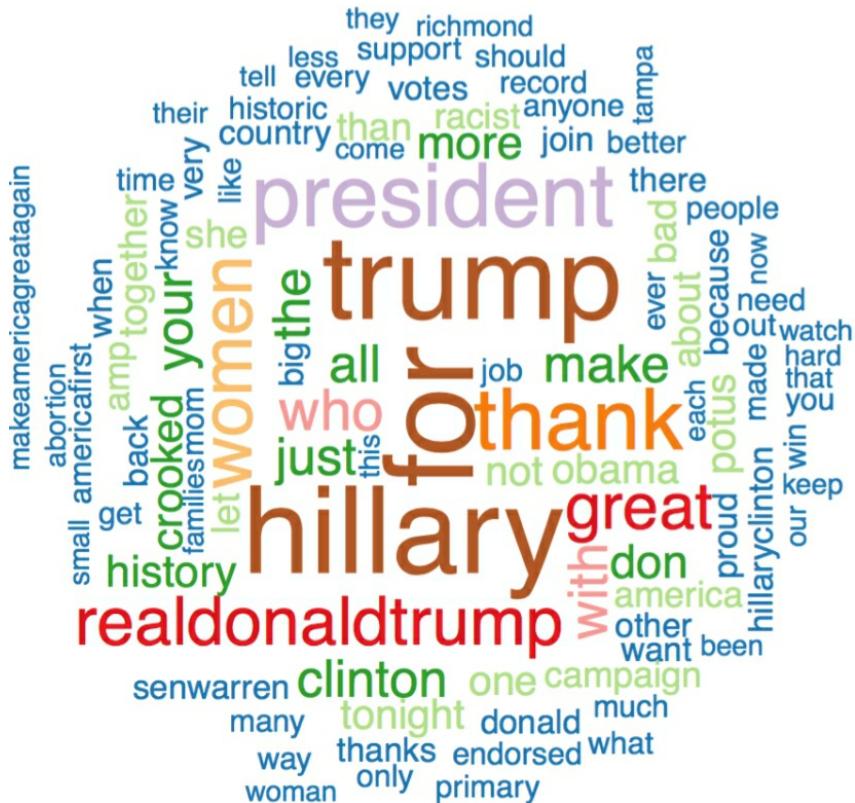
**Figure 16.5: Big Data Architecture**

A million points of data can be plotted in a graph and offer a view of the density of data. However, plotting a million points on the graph may produce a blurred image which may hide, rather than highlight the distinctions. In such a case, binning the data would help, or selecting the top few frequent categories may deliver greater insights. Streaming data can also be visualized by simple counts and averages over time. For example, below is a dynamically updated chart that shows up-to-date statistics of visitor traffic to the author's blogsite, anilmah.com. The bar shows the number of page views, and the inner darker bar shows the number of unique visitors. The dashboard could show the view by days, weeks or years also.



**Figure 16.6: Real-time Dashboard for website performance for the author's blog**

Text Data could be combined, filtered, cleaned, thematically analyzed, and visualized in a wordcloud. Here is wordcloud (Figure 16.7) from a recent stream of tweets (ie Twitter messages) from 2016 US Presidential candidates Hillary Clinton and Donald Trump. The larger size of words implies greater frequency of occurrence in the tweets. This can help understand the major topics of discussion between the two.



**Figure 16.7: A wordcloud of Hillary Clinton's and Donald Trump's tweets**

## Technology Challenges for Big Data

There are four major technological challenges in managing Big Data. There are four matching layers of technologies to overcome those challenges.

### Storing Huge Volumes

The first challenge relates to storing huge quantities of data. No storage machine is big enough to store the relentlessly growing quantity of data. Therefore, data needs to be stored in a large number of smaller inexpensive machines. However, with a large number of machines, there is the inevitable challenge of machine failure. Each of these commodity machines will fail at some point or another. Failure of a machine could entail a loss of data stored on it.

The first layer of Big Data technology thus helps store huge volumes of data, at an affordable cost,

while avoiding the risk of data loss. It distributes data across a large cluster of inexpensive commodity machines. It ensures that every piece of data is systematically replicated on multiple machines to guarantee that at least one copy is always available.

Apache Hadoop is the most well-known clustering technology for Big Data. Its data storage system is called Hadoop Distributed File System (HDFS). This system is built on the patterns of Google's Big File system, designed to store billions of pages and sort them to answer user search queries.

### **Ingesting streams at an extremely fast pace**

The second challenge relates to the Velocity of data, i.e. handling unpredictable and torrential streams of data. Some of the data streams may be too large to store, but must still be monitored. The solution lies in creating scalable ingest systems that can open an unlimited number of channels for receiving data. These systems can hold data in queues, from which business applications can read and process data at their own pace and convenience.

The second layer of Big Data technology manages this velocity challenge. It uses a special stream-processing engine, where all incoming data is fed into a central queueing system. From there, a fork-shaped system sends data to batch storage as well as to stream processing directions. The stream processing engine can do its work while the batch processing does its work. Apache Spark is the most popular system for streaming applications.

### **Handling a variety of forms and functions of data**

The third challenge relates to the structuring and access of all varieties of data that comprise Big Data. Storing them in traditional flat or relational structures would be too impractical, wasteful and slow. Accessing and analyzing them requires different capabilities.

The third layer of Big Data technology solves this problem by storing the data in non-relational systems that relax many of the stringent conditions of the relational model. These are called NoSQL (Not Only SQL) databases. These databases are optimized for certain tasks such as query processing, or graph processing, document processing, etc.

HBase and Cassandra are two of the better known NoSQL databases systems. HBase, for example, stores each data element separately along with its key identifying information. This is called a key-value pair format. Cassandra stores data in a columnar format. There are many other variants of NoSQL databases. NoSQL languages, such as Pig and Hive, are used to access this data.

### **Processing data at huge speeds**

The fourth challenge relates to moving large amounts of data from storage to the processor, as this would consume enormous network capacity and choke the network. The alternative and innovative mode would be to do just the opposite, i.e. to move the processing to where the data is stored.

The fourth layer of Big Data technology avoids the choking of the network. It distributes the task logic throughout the cluster of machines where the data is stored. Those machines work, in parallel, on the

data assigned to them, respectively. A follow-up process consolidates the outputs of all the small tasks and delivers the final results. MapReduce, also invented by Google, is the best-known technology for parallel processing of distributed Big Data.

**Table 16.1: Technological challenges and solutions for Big Data**

Challenge	Description	Solution	Technology
<b>Volume</b>	Avoid risk of data loss from machine failure in clusters of commodity machines	Replicate segments of data in multiple machines; master node keeps track of segment location	HDFS
<b>Volume &amp; Velocity</b>	Avoid choking of network bandwidth by moving large volumes of data	Move processing logic to where the data is stored; manage using parallel processing algorithms	Map-Reduce
<b>Variety</b>	Efficient storage of large and small data objects	Columnar databases using key-pair values format	HBase, Cassandra
<b>Velocity</b>	Monitoring streams too large to store	Fork-shaped architecture to process data as stream and as batch	Spark

Once these major technological challenges arising from the 4 Vs of data are met, all traditional analytical and presentation tools, such as machine learning and statistics, can be reliably applied to Big Data. There are many additional technologies to make the task of managing Big Data easier. For example, there are technologies to monitor the resource usage and load balancing of the machines in the cluster.

## Conclusion and Summary

Big Data is a major social and technological phenomenon that impacts everyone. It also provides an opportunity to create new ways of knowing and working. Big Data is extremely large, complex, fast, and not always clean as it is data that comes from many sources such as people, web, and machine communications. Big Data needs to be gathered, organized and processed in a cost-effective way that manages the volume, velocity, variety and veracity of Big Data. Hadoop, MapReduce, NoSQL, and Spark systems are popular technological platforms for this purpose.

To summarize, Table 16.2 shows the many differences between traditional and Big Data.

**Table 16.2: Comparing Big Data with Traditional Data**

Feature	Traditional Data	Big Data
<b>Representative Structure</b>	Lake / Pool	Flowing Stream / river
<b>Primary Purpose</b>	Manage business activities	Communicate, Monitor
	Business transactions,	Social media, Web access logs, machine

<b>Source of data</b>	documents	generated
<b>Volume of data</b>	Gigabytes, Terabytes	Petabytes, Exabytes
<b>Velocity of data</b>	Ingest level is controlled	Real-time unpredictable ingest
<b>Variety of data</b>	Alphanumeric	Audio, Video, Graphs, Text
<b>Veracity of data</b>	Clean, more trustworthy	Varies depending on source
<b>Structure of data</b>	Well-Structured	Semi- or Un-structured
<b>Physical Storage of Data</b>	In a Storage Area Network	Distributed clusters of commodity computers
<b>Database organization</b>	Relational databases	NoSQL databases
<b>Data Access</b>	SQL	NoSQL such as Pig
<b>Data Manipulation</b>	Conventional data processing	Parallel processing
<b>Data Visualization</b>	Variety of tools	Dynamic dashboards with simple measures
<b>Database Tools</b>	Commercial systems	Open-source - Hadoop, Spark
<b>Total Cost of System</b>	Medium to High	high

## Review Questions

- Q1. What is Big Data? Why should anyone care?
- Q2. Describe the 4V model of Big Data.
- Q3. What are the major technological challenges in managing Big Data?
- Q4: What are the technologies available to manage Big Data?
- Q5. What kind of analyses can be done on Big Data?
- Q6: Why did people pay attention to Hadoop/MapReduce when it was introduced?

### Liberty Stores Case Exercise: Step P1

Liberty Stores Inc. is a specialized global retail chain that sells organic food, organic clothing, wellness products, and education products to enlightened LOHAS (Lifestyles of the Healthy and Sustainable) citizens worldwide. The company is 20 years old, and is growing rapidly. It now operates in 5 continents, 50 countries, 150 cities, and has 500 stores. It sells 20000 products and has 10000 employees. The company has revenues of over \$5 billion and has a profit of about 5% of its revenue. The company pays special attention to the conditions under which the products are grown and produced. It donates about one-fifth (20%) from its pre-tax profits from global local charitable causes.

Q1: Suggest a comprehensive Big Data strategy for the CEO of the company.

Q2: How can a Big Data system such as IBM Watson help this company?

# Chapter 17: Data Modeling Primer

## Introduction

Data needs to be efficiently structured and stored so that it includes all the information needed for decision making, without duplication and loss of integrity. Here are top ten qualities of good data.

Data should be:

1. *Accurate*: Data should retain consistent values across data stores, users and applications. This is the most important aspect of data. Any use of inaccurate or corrupted data to do any analysis is known as the garbage-in-garbage-out (GIGO) condition.
2. *Persistent*: Data should be available for all times, now and later. It should thus be nonvolatile, stored and managed for later access.
3. *Available*: Data should be made available to authorized users, when, where, and how they want to access it, within policy constraints.
4. *Accessible*: Not only should data be available to user, it should also be easy to use. Thus, data should be made available in desired formats, with easy tools. MS Excel is a popular medium to access numeric data, and then transfer to other formats.
5. *Comprehensive*: Data should be gathered from all relevant sources to provide a complete and holistic view of the situation. New dimensions should be added to data as and when they become available.
6. *Analyzable*: Data should be available for analysis, for historical and predictive purposes. Thus, data should be organized such that it can be used by analytical tools, such as OLAP, data cube, or data mining.
7. *Flexible*: Data is growing in variety of types. Thus, data stores should be able to store a variety of data types: small/large, text/video, and so on
8. *Scalable*: Data is growing in volume. Data storage should be organized to meet emergent demands.
9. *Secure*: Data should be doubly and triply backed up, and protected against loss and damage. There is no bigger IT nightmare than corrupted data. Inconsistent data has to be manually sorted out which leads to loss of face, loss of business, downtime, and sometimes the business never

recovers.

10. *Cost-effective*: The cost of collecting data and storing it is coming down rapidly. However, still the total cost of gathering, organizing, and storing a type of data should be proportional to the estimated value from its use.

## **Evolution of data management systems**

Data management has evolved from manual filing systems to the most advanced online systems capable of handling millions of data processing and access requests every second.

The first data management systems were called file systems. These mimicked paper files and folders. Everything was stored chronologically. Access to this data was sequential.

The next step in data modeling was to find ways to access any random record quickly. Thus hierarchical database systems appeared. They were able to connect all items for an order, given an order number.

The next step was to traverse the linkages both ways, from top of the hierarchy to the bottom, and from the bottom to the top. Given an item sold, one should be able to find its order number, and list all the other items sold in that order. Thus there were networks of links established in the data to track those relationships.

The major leap came when the relationship between data elements itself became the center of attention. The relationship between data values was the key element of storage. Relationships were established through matching values of common attributes, rather than by location of the record in a file. This led to data modeling using relational algebra. Relations could be joined and subtracted, with set operations like union and intersection. Searching the data became an easier task by declaring the values of a variable of interest.

The relational model was enhanced to include variables with non-comparable values like binary objects (such as pictures) which had to be processed differently. Thus emerged the idea of encapsulating the procedures along with the data elements they worked on. The data and its methods were encapsulated into an *object*. Those objects could be further specialized. For example, a vehicle is an object with certain attributes. A car and a truck are more specialized versions of a vehicle. They inherited the data structure of

the vehicle, but had their own additional attributes. Similarly the specialized object inherited all the procedures and programs associated with the more general entity. This became the object-oriented model.

## Relational Data Model

The first mathematical-theory-driven model for data management was designed by Ed Codd of IBM in 1970.

1. A relational database is composed of a set of relations (data tables), which can be joined using shared attributes. A “data table” is a collection of instances (or records), with a key attribute to uniquely identify each instance.
2. Data tables can be JOINed using the shared “key” attributes to create larger temporary tables, which can be queried to fetch information across tables. Joins can be simple ones as between two tables. Joins can also be complex with AND, OR, UNION or INTERSECTION, and more operations.
3. High-level commands in Structured Query Language (SQL) can be used to perform joins, selection, and organizing of records.

Relational data models flow from conceptual models, to logical models to physical implementations. Data can be conceived of as being about entities, and relationships among entities. A relationship between entities may be hierarchy between entities, or transactions involving multiple entities. These can be graphically represented as an entity–relationship diagram (ERD).

In Figure 16.1, the rectangle reflects the entities students and courses. The relationship is enrolment. In the example below the rectangle reflects the entities Students and Courses. The diamond shows the Enrolment relationship.



Figure: 16.1 Simple relationship between two entities

Here are some fundamental concepts on ERD:

1. An **entity** is any object or event about which someone chooses to

collect data, which may be a person, place, or thing (e.g., sales person, city, product, vehicle, employee).

2. Entities have **attributes**. Attributes are data items that have something in common with the entity. For example, student id, student name, and student address represent details for a student entity. Attributes can be single-valued (e.g., student name) or multi-valued (list of past addresses for the student). Attribute can be simple (e.g., student name) or composite (e.g., student address, composed of street, city, and state).
3. Every entity must have a **key attribute(s)** that can be used to identify an instance. E.g. Student ID can identify a student. A primary key is a unique attribute value for the instance (e.g. Student ID). Any attribute that can serve as a primary key (e.g. Student Address) is a candidate key. A secondary key—a key which may not be unique, may be used to select a group of records (Student city). Some entities will have a composite key—a combination of two or more attributes that together uniquely represent the key (e.g. Flight number and Flight date). A **foreign key** is useful in representing a one-to-many relationship. The primary key of the file at the one end of the relationship should be contained as a foreign key on the file at the many end of the relationship.
4. **Relationships** have many characteristics: degree, cardinality, and participation.
5. **Degree of relationship** depends upon the number of entities participating in a relationship. Relationships can be unary (e.g., employee and manager-as-employee), binary (e.g., student and course), and ternary (e.g., vendor, part, warehouse)
6. **Cardinality** represents the extent of participation of each entity in a relationship.
  - a. One-to-one (e.g., employee and parking space)
  - b. One-to-many (e.g., customer and orders)
  - c. Many-to-many (e.g., student and course)
7. **Participation** indicates the optional or mandatory nature of relationship.
  - a. Customer and order (mandatory)
  - b. Employee and course (optional)

8. There are also **weak entities** that are dependent on another entity for its existence (e.g., employees and dependents). If an employee data is removed, then the dependent data must also be removed.
9. There are **associative entities** used to represent **many-to-many relationship** relationships (e.g., student-course enrolment). There are two ways to implement a many-many relationship. It could be converted into two one-to-many relationships with an associative entity in the middle. Alternatively, the combination of primary keys of the entities participating in the relationship will form the primary key for the associative entity.
10. There are also **super sub type entities**. These help represent additional attributes, on a subset of the records. For example, vehicle is a supertype and passenger car is its subtype.

## Implementing the Relational Data Model

Once the logical data model has been created, it is easy to translate it into a physical data model, which can then be implemented using any publicly available DBMS. Every entity should be implemented by creating a database table. Every table will be a specific data field (key) that would uniquely identify each relation (or row) in that table. Each master table or database relation should have programs to create, read, update, and delete the records.

The databases should follow 3 Integrity Constraints.

1. *Entity integrity* ensures that the entity or a table is healthy. The primary key cannot have a null value. Every row must have a unique value. Or else that row should be deleted. As a corollary, if the primary key is a composite key, none of the fields participating in the key can contain a null value. Every key must be unique.
2. *Domain integrity* is enforced by using rules to validate the data as being of the appropriate range and type.
3. *Referential integrity* governs the nature of records in a one-to-many relationship. This ensures that the value of a foreign key should have a matching value in primary keys of the table referred to by the foreign key.

## **Database management systems (DBMS)**

These are many database management software systems that help manage the activities related to storing the data model, the data itself, and doing the operations on the data and relations. The data in the DBMS grows, and it serves many users of the data concurrently. The DBMS typically runs on a computer called a database server – in an n-tier application architecture. Thus in an airline reservation system, millions of transactions might simultaneously try to access the same set of data. The database is constantly monitored and managed to provide data access to all authorized users, securely and speedily, while keeping the database consistent and useful. Content management systems are special purpose DBMS, or just features within standard DBMS, that help people manage their own data on a website. There are object-oriented and other more complex ways of managing data.

## **Structured Query Language**

SQL is a very easy and powerful language to access relational databases. There are two essential components of SQL: the Data Definition Language (DDL) and Data Manipulation Language.

DDL provides instructions to create new database, and to create new tables within a database. Further it provides instructions to delete a database, or just a few tables within a database. There are other ancillary commands to define indexes etc for efficient access to the database.

DML is the heart of SQL. It provides instructions to add, read, modify and delete data from the database and any of its tables. The data can selectively accessed, and then formatted, to answer a specific question. For example, to find the sales of movies by quarter, the SQL query would be:

```
SELECT Product-Name, SUM(Amount)  
FROM Movies-Transactions  
GROUP BY Product-Name
```

## **Conclusion**

Data should be modeled to achieve the business objectives. Good data should be accurate and accessible, so that it can be used for business operations. Relational data model is the two most popular way of managing data today.

## **Review Questions**

- 1: Who invented relational model and when?
- 2: How does relational model mark a clear break from previous database models?
- 3: What is an Entity-Relationship diagram?
- 4: What kinds of attributes can an entity have?
- 5: What are the different kinds of relationships?

# Chapter 18: Statistics Primer

## Introduction

Statistics is a mathematically elegant and efficient way of understanding the characteristics of a whole population by projecting from the understanding of a sampled part. Ideally, the statistical analysis would give a completely accurate and parsimonious description of the whole population with the least amount of effort. This is achieved by accessing a truly random representative sample of the population, gathering data on sample observations, analyzing the data to discover patterns, and projecting the results to the whole with a certain high minimum level of confidence. Statistical analyses ultimately lead to a deeper understanding of the Totality, the unified field of all the laws of nature.

There are many applications of Statistics. It is mostly useful in understanding human populations. Politicians conduct polling surveys to understand the needs of voters. Marketers use statistics to understand the needs and wants of consumers. Statistics is also useful in assessing quality in production processes, by measuring the quality of a sample of products.

## Descriptive Statistics

These are tools and techniques to describe a collection of data. Data is often described by its central tendency and its spread. The primary central tendency is the mean, or the average, of the values. There are other central tendencies such as the median and mode. The spread within the data is called Variance, and is often described by ‘standard deviation’.

### Example data set

For example, use this data set of people as shown below.

Gender	Age	Height	Weight
M	24	71	165
M	29	68	165
M	34	72	180
F	21	67	113
M	32	72	178
F	25	62	101
M	26	70	150
M	34	69	172
M	31	72	185

F	49	63	149
M	30	69	132
M	34	69	140
F	28	61	115
M	30	71	140

A person might be described by their gender, age, height, and weight, for example. Those are the attributes for a person. A collection of persons can be described by their total count, and their average age, height and weight. The averages could be computed separately for different values of gender.

Averages need to be complemented by the spread, or variance in the values of the attributes. For example, two groups may have the same average age. However, in one group all members have the same exact age. While in the other group, the age ranges from low values to high values. In the first group, the variability, or spread, is zero. In the other group, there is a non-zero spread.

#### Computing Mean, Median, Mode

The count of instances in this data set is 14. The average age is 30.5 years. (It can be computed by using the function ‘Average’ in MS Excel). The average height is 68.3 inches. The average weight is 149.9 lbs.

To appreciate the concept of spread, it helps to sort the data by one of the attributes. This data is sorted by Age and shown below.

Gender	Age	Height	Weight
F	21	67	113
M	24	71	165
F	25	62	101
M	26	70	150
F	28	61	115
M	29	68	165
M	30	69	132
M	30	71	140
M	31	72	185
M	32	72	178
M	34	72	180
M	34	69	172
M	34	69	140

F	49	63	149
---	----	----	-----

The age of the members in this group ranges from 21 years to 49 years. The average may not tell the full story about the group. Thus another way to describe the central tendency of the age of this group, is to describe the age of the person in the middle of this group. The age of the middle member in this group of 14 people is 30 years. That is called the median value of the age of the group. (It can be computed by using the function ‘Median’ in MS Excel). Median is particularly useful when the distribution of values is highly skewed. Thus the median height of the group is 69 inches. And the median weight is 149.5 lbs.

Another way of describing central tendency is of particular relevance when the values are non-numeric, such as their gender, native language, etc. It is not possible to average out the non-numeric data. Thus the concept of ‘mode’ becomes important. A mode is the value that shows the highest frequency of occurrence in the data. In this data, the age value of 34 occurs three times, the age value of 30 occurs two times, and all the other values occur just one time each. Thus the mode value of the age of this group is 34 years. (It can be computed by using the function ‘Mode’ in MS Excel). Similarly, the modal height of the group is 69 inches and 72 inches as both occur 3 times each. In such cases it is called a bi-modal data. The modal value of weight is 140 lbs.

The concept of mode can be useful in describing the dominant gender of this group. The M value occurs 10 times, and the F value occurs 4 times. Thus one can say that the modal value of the gender for the group is M.

#### Computing the range and variance

The spread of the data can be defined in two primary ways. One is the range, the other is variance.

The range is simply defined by the minimum and maximum values of the variable. Thus the range of values of age in this dataset is from 21 to 49. Similarly the range of height is from 61 to 72 inches, and the range of weight is from 101 to 185 lbs. The greater the range, the more there is variation in the data. The lower the range, the greater is the homogeneity in the data. The limitation of the range measure is that it depends upon only the extreme values on either side, and ignores the distribution of the rest of the data in the middle.

The variance is a more inclusive measure of the spread in the data. Variance

is defined as the sum of squares of the distance from the mean. Thus to compute the variance for the age variable for this group, first calculate for each instance the difference between the age and the average value for the group (which is 30.5). Make a square of those difference values by multiplying each value with itself. Summing up those squared values leads to the total variance. Dividing the total variance by the number of instances, leads to the average variance of the data. The standard deviation is defined as the square root of the average variance. Mathematically speaking, here is the formula for standard deviation:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Where S is the standard deviation.  $\Sigma$  means the sum of all the values. X is the value of the variable for an instance in the dataset.  $\bar{X}$  is the mean for that variable in the data set. N is the number of instances in the dataset.

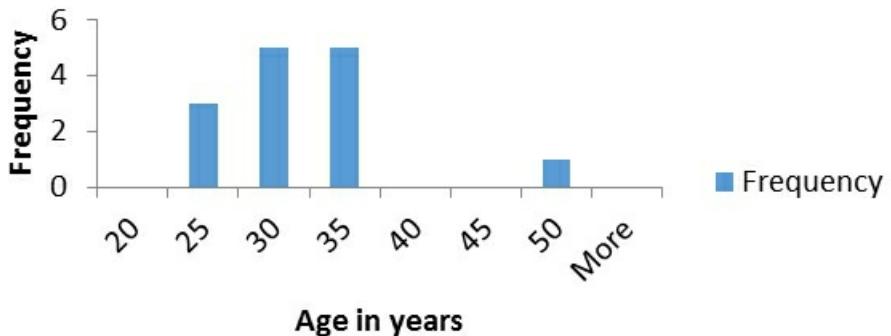
Thus computing the average variance for age in this group will lead to a value of 6.64 years. (It can be computed with the function ‘Stdev’ in MS Excel). Similarly, the standard deviation can be computed for all the other numeric variables. The standard deviation for height is 3.75 years, and the standard deviation for weight is 26.8 lbs.

Thus, any dataset can be defined quite meaningfully in terms of its mean and standard deviation.

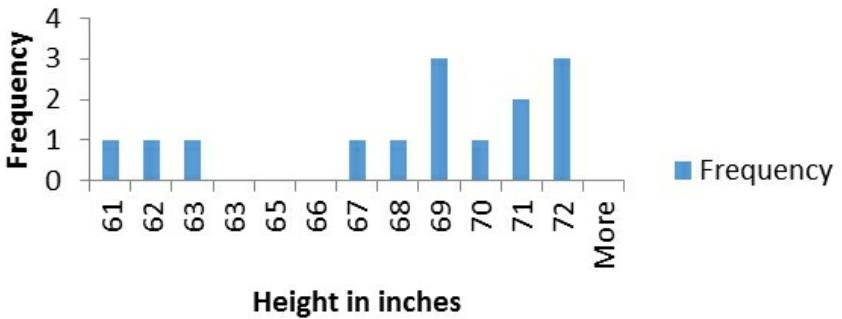
### Histograms

Histograms are a particularly useful way of visually describing the distribution of data. Here below are histograms for each of the variables.

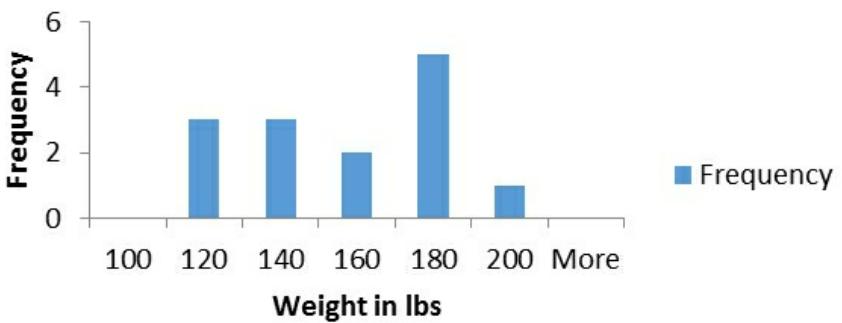
## Histogram of Age



## Histogram of Height



## Histogram of Weight



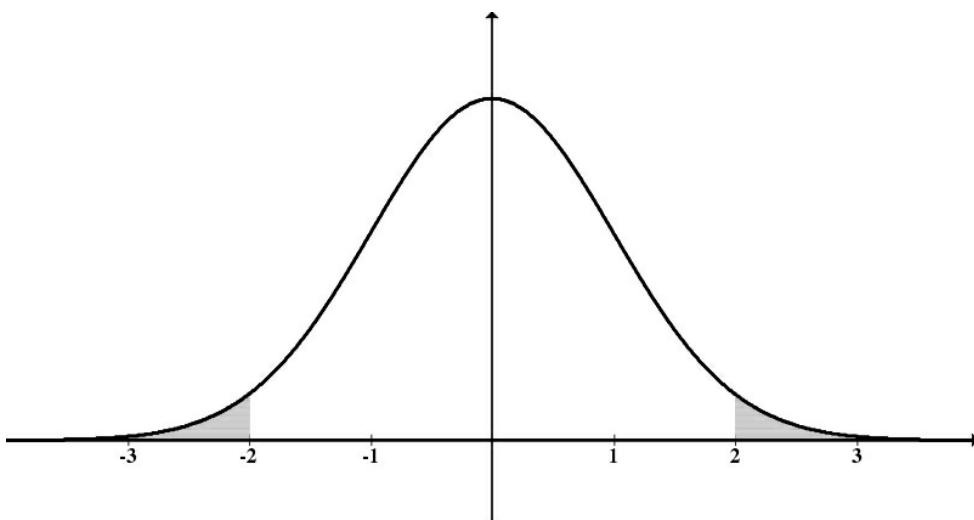
Histograms are usually developed for numeric variables, though in principle they can also be developed for non-numeric variables such as Gender.

These histograms show a wide range of data values, with major gaps in the

middle. Therefore, the range of each variable should be specified in addition to the mean value to give a more accurate picture of the data. And more importantly, the gaps suggest that there are probably different subgroups within the data. In this case, it is likely that the values vary significantly based on the value of Gender. The Analysis of Variance (ANOVA) technique helps determine the impact of such variables.

#### Normal Distribution and Bell Curve

The data is normally distributed if the data values are evenly distributed like a symmetric bell-shaped curve.



The central tendency is the middle point of the distribution in data. The perfect symmetry ensures that this central point is the mean, as well as the median, and the mode for the data distribution. The frequency of the values decreases on either sides of the central value. The numbers on the x-axis represent the number of standard deviations away from the central tendency, or the mean. Approximately 95% of the data values, the area under the curve, fall within 2 standard deviations (s.d.) on each side.

If the data set were normally distributed, 95% of the age values in the data set will fall within  $\text{Mean} \pm 2 \times \text{s.d.}$ . Assuming the age is normally distributed, 95% of the data values would fall within that 2 sd range, i.e.  $(30.5 - 2 \times 6.64)$  to  $(30.5 + 2 \times 6.64)$ . Or from about 17 to 44 years. This adequately reflects this data set.

A normal distribution assumes a natural clustering of values around the central value. It is considered a naturally occurring pattern in most situations. In most data sets, data is assumed to be normally distributed unless otherwise

specified. As the dataset gets larger, it would more and more begin to resemble the bell curve.

The normal distribution can be mathematically modeled in terms of two parameters ... mean and standard deviation. Parametric curves have many properties and benefits. . For example, normal curves can be instantly compared. Also, the area under the normal curve can be used using calculus functions, and will follow another symmetric curve called the S-curve. That can be used in further computations. A normal distribution of data helps infer the properties of the population from a randomized sample, with a certain high degree of confidence.

## Inferential Statistics

The primary purpose of statistics is to infer the properties of the population without having to reach out to and measure the entire population. If a suitably randomized sample is selected, such that it can reasonably be assumed to be representative of the overall population, then the properties of the population can be inferred from it, with a certain high degree of confidence. The larger the sample size is, the greater would be the confidence.

### Random sampling

A random sample is a mathematically rigorous technique of selecting a sample that is representative of the population. It should be contrasted with other forms of sampling, such as convenience sampling, which finds the easiest way to gather data without caring much about its representativeness of the entire population. The random sampling process uses a random number generating mechanism to select a random a certain number of instances from the total population, such that every instance has an equal chance of being selected. An example of random selection would be like how lottery winners are selected from the entire pool of lottery tickets sold. The size of the random sample is chosen by the statistician based on the accuracy needs, and the time and resources available.

The random sample can be analyzed using any of the descriptive techniques described above. And even analytical techniques such as Analysis of Variance.

### Confidence Interval

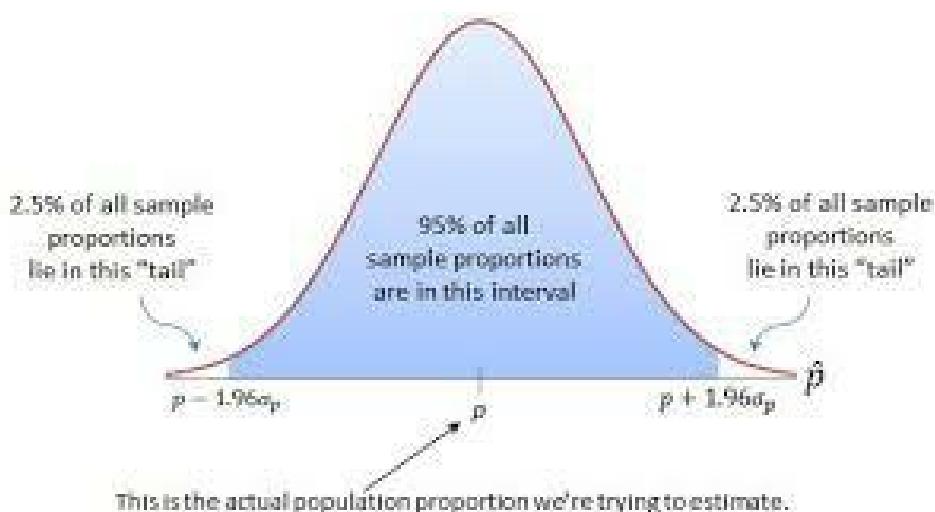
Unless the entire population has been measured, a completely true value of the variable for the population cannot be provided. The random sampling

method can provide an estimate of the true mean of the population, with a certain degree of confidence. The true population mean will fall within a range centered about the sample. This is called the confidence interval. The larger the sample size, the narrower will be the confidence interval. The greater the sample size, the greater will be the confidence in the projection from the sample mean.

Suppose that the data set used above was actually representative of the people in a certain community. How confidently can one claim then the average age of population of the community is 30.5 years (the average age of the sample)?

The range, around the sample mean, within which the true population mean ( $P$ ) is likely to fall, it called confidence interval. The higher the confidence level that is desired, the wider will be the confidence interval.

Based on the normal distribution of data, the confidence interval of 95% translates into 2 standard deviations worth of distance on either side of the mean. It means that if such randomized sampling and analysis was done often, then 95% of the time the confidence interval of mean  $\pm 2$  s.d. will bracket (i.e. include) the true population mean. The confidence level is usually specified by its inverse, the chance or error, i.e.  $1 - \alpha$ , where  $\alpha$  is called the significance level, or simply the p-value. For a 95% confidence level, the p-value should be less than or equal to 0.05.



## Predictive Statistics

Statistical techniques such as regression analysis can be used for predicting

the values of variables of interest. Correlations among variables shows which variables might influence the values of others. There are many variations of regression, such as linear, non-linear and logit regression. Time series analysis is a special case of regression analysis where the key independent variable is time. These are described in full detail in chapter 7.

## Summary

Statistical tools are time-tested ways of understanding large populations using small randomized samples. Data is described using central tendencies such as mean, median and mode; and its distribution using range and variance. Predictive statistical methods like regression and others are prominent tools in data analytics. These are fully covered in chapter on Regression.

## Review Questions

Q1: What is statistics? What are its different types?

Q2: What is a histogram? How does it help?

Q3: What is the difference between central tendency and spread?

Q4: what is the difference between mean, mode, and median?

Q5: What is standard deviation? How is it computed? How does it help?

Q6: What is random sampling? How does it help?

Q7: what is the difference between confidence level and confidence interval?

## Chapter 19 - Artificial Intelligence Primer

Artificial Intelligence (AI) is the way of abstracting intelligence capabilities out of the human mind, and embedding them into non-sentient objects. Ideally, AI will be able to express intelligent behavior surpassing human intelligence in every way. This is achieved by modeling human intelligence through the whole spectrum of perspectives from genetic, to neurological, to cognitive, to behavioral, and the physical. AI ultimately models and implements the universal laws of nature.

Human Beings have sought to improve their quality of lives. These have been revolutions in human life from hunter-gatherer caveman, to agricultural societies, to industrial cities, to information society. At every stage human beings have tried to understand the phenomena in and around themselves and made tools to make their life easier and simpler. Artificial Intelligence can be seen as the latest link in a series of innovations over the last many centuries that helped improve lives.

### CASELET: Apple Siri Voice-activated personal assistant

**Apple's Siri** is a computer program that works as an intelligent personal assistant and knowledge navigator. It uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. It is a voice-activated assistant that mimics human intelligence and natural conversation. It can open apps, give you movie times and sports scores, make dinner reservations, call or send messages to people in your contact list and perform many other useful tasks. It also asks what she's doing and prompts you with questions so that you have a chance to make choices and correct her if she misinterprets you.

Siri is an offshoot of the DARPA-funded project, CALO, which was a part of DARPA's PAL initiative (Personalized Assistant that Learns). Siri involves a number of technologies, including natural language processing, question analysis, data mashups, and machine learning. Siri's main algorithm, at a high level, is as follows:

1. Use Automatic speech recognition to transcribe human speech into text.
2. Use natural language processing to translate transcribed text into 'parsed text'.
3. Use question & intent analysis to analyze parsed text, detecting user commands and actions such as "What is the weather".
4. Use data mashup technologies to interface with 3rd-party web services to perform actions, search operations, and question answering
5. Transform output of 3rd party web services back into natural language text (e.g, "The weather will be sunny")
6. Using TTS (text-to-speech) technologies to transform the answer into synthesized speech.

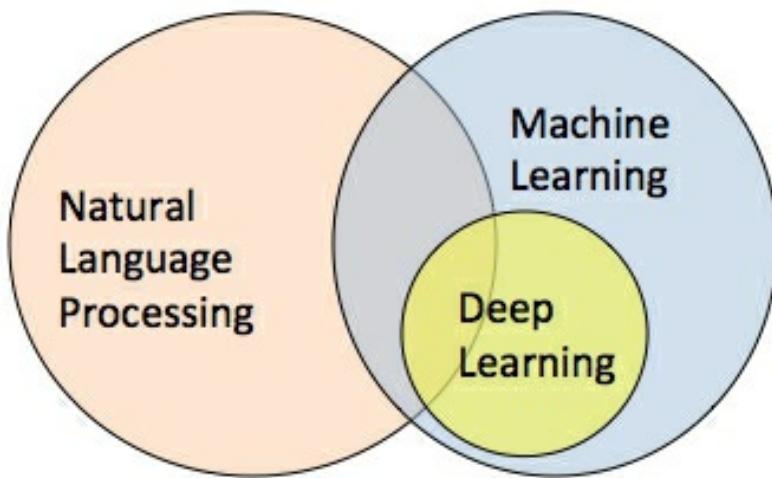
Q1: What are some of the benefits of Siri?

Q2: How does Big Data support AI?

## AI, Machine Learning, and Deep Learning

Artificial Intelligence is a broad term that includes all manner of intelligence machines. Patrick Winston of MIT describes AI as representations that support models targeted at thinking, perception and action. AI includes Machine Learning, Natural Language Understanding, Expert Systems, Computer vision, Robotics, and other capabilities (Figure 19.1). Machine learning systems are those

that helps learn patterns from data using Neural Networks, Genetic Algorithms, and more. Deep Learning systems are prediction systems based on Neural Networks. They are used to implement Decision Trees, Clustering systems, Recommender systems, Text mining, Web mining, Social Network mining, and more. They have been used for playing games like chess, Go, Pac-man and others at great level of expertise.



**Figure 19.1: Components of AI and Machine Learning**

To understand the impact of AI on life and work, it may help to understand the revolutionary technologies that have come before it. Industrial revolution came about 200 years ago, and information revolution came about 100 years ago. The AI idea started about 50 years ago and has come into full bloom in recent years.

## The Industrial Revolution

A few hundred years ago, work mostly meant physical work such as lifting, pushing, carrying or otherwise manipulating physical stuff. The invention of steam engine in the 18<sup>th</sup> century, was the result of human creativity. At first people did not recognize or appreciate it. Human and horse muscle was prided. Gradually the steam engine began to be deployed in a variety of uses. This launched the Industrial revolution, which made Britain an industrial power and a global empire. The creation of trans-continental railway transport systems in the 19<sup>th</sup> century was possible because of the steam engine. The invention of the internal combustion engine helped harness the physical motion in machines. The invention of flying machines began to further revolutionize transportation. This was the beginning of the end of the value of brute force work by human muscle. The power of horses was also being surpassed. Horse riders and whip makers worried for their profession. Science fiction imagined tele-portionation as a possible way to travel instantly, and of traveling into space. The steam and internal-combustion engines, however, created new industries in the 20<sup>th</sup> century, such as steel and rubber and fuel and machine maintenance and highways and tourism and fast-food and motel industries. Millions of new jobs were created.

## The Information Revolution

In the 1850s, the tele-graph enabled writing at a distance. In the late 1890s, the tele-phone made voice possible at a distance. In the 1930, the tele-vision enabled seeing a picture at a distance. In about 1910s, electric light created the ability to work all day and improve public safety. In the 1950s, the IBM corporation invented the computer. In the 1960s, the powerful IBM ‘mainframe’ computer was launched. In 1976, Apple-II was launched a personal computer. It enabled people to do their own writing and printing. In 1981, the launch of the IBM PC further democratized personal computing.

The spreadsheet was the killer-app for the PC. The spreadsheet enabled an ordinary person to do many kinds of sophisticated calculations. Word processing software made typing long documents such as books and doctoral dissertations easy. The combination of the PC and the internet democratized email. Powerful software called the Enterprise Resource Planning suites began to transform how businesses were run, providing up-to-date information to every manager. Business processes were re-engineered and people feared massive loss of well-paying jobs.

In the 1960s, the US Defense Advanced Projects Agency funded the creation of a distributed computer network that would survive a nuclear attack. This evolved into the internet. Soon the internet was used for personal communications and email became the killer-app for the internet. In 1990s the world-wide web was born as an internet application. It used a system of hyperlinks to link all information. In 1993, e-commerce, or the trading of goods and services on the web, emerged as a new industry. Amazon began selling books online in 1993. In 1998, Google began as a platform to “organize the world’s information”, and soon became a verb to search the web. In 2004, Facebook helped people communicate across distances and launched the social media.

In 2007, Apple launched the iPhone as a smart phone; it launched the mobile web browsing experience. Millions of software applications (or apps) now run on billions of smart phones of many types to fetch any information anywhere and anytime. Access to information has become democratized and cheap. Siri system on the iPhone represents an advanced AI-based voice assisted tool for easy user interaction.

## The Cognitive (or AI) revolution

In 1950 a British mathematician Alan Turing who worked to computationally crack communication codes in the second World War, expressed hope that computers someday will be able work like a human being. Turing’s test for a computer was to engage in a conversation with a human being, such that the human being could not decipher whether he was conversing with a machine or a human being. Turing’s implicit aim was extract intelligence, and place it into disembodied device such as a computer. In 1956 a small seminar was held in Dartmouth college in New Hampshire, with the aim of developing Artificial Intelligence. Many successful projects spawned from this academic gathering. These include natural language processing, Artificial Neural Networks, and Expert systems.

Expert systems captured human knowledge as logical if-then statements. Expert systems developed well to do some routine complex tasks with great efficiency and accuracy. Expert systems were also sought for playing games such as chess. In 1997, IBM’s Deep Blue computer, an expert system with enormous computing power behind it, beat world chess champion Gary Kasparov decisively in a 6-game match. Many of these systems worked on deterministic hard-coded logic. However, as the world’s store of information kept growing, hard-coding deterministic logic became tenuous. Another class of techniques called stochastic or probabilistic systems came in handy. Genetic Algorithms are a class of probabilistic computational techniques that mimic genetic evolution systems. Genes mutate randomly as they evolve in their next generation. Useful mutations, that enhance survival, get passed on to the next generation while the unhelpful mutations die out. Natural language understanding systems also use a probabilistic approach to interpret a voice signal.

Neural networks need large amount of data to train the systems. As the world's store of data grew, neural networks learned from those data sets to better achieve their goals. Now, deep learning systems based on neural networks have demonstrated enormous predictive powers. In 2011, IBM's Watson system beat the world champion in the Jeopardy quiz challenge. In 2012, Neural networks were demonstrated to accurately classify images in various categories. In 2017, Google's AlphaGo system based on TensorFlow algorithms (advanced neural networks) beat the world's Go champion. FaceBook and other companies have developed Face Recognition systems to tag the pictures posted on its system with the right people. These AI based systems are being used in China to identify people in the crowd, if they were doing something undesirable.

IBM Watson and Google TensorFlow are called Cognitive systems. These systems have capabilities to use relatively simple but clever algorithms in a recursive manner to achieve their objective functions. They can also learn from new instances of activities and can add that learning in their repertoire. These systems can easily engage in relatively intelligent and humorous conversations with human beings, in many languages. Disembodied voice-based interaction systems such as Apple's Siri and Amazon's Alexa have been deployed in smart phones and other devices. These systems understand human language commands and respond appropriately such as play music, switch on lights, check on weather and directions, and more. The achievement of Turing Test is almost here. By 2029, Ray Kurzweil has predicted, machines will be able to achieve the intelligence of human beings.

## **Jobs Losses and Gains**

There is genuine fear that AI will take away the jobs of even white-collar professionals. However the concept of a job is a relatively new concept in the last century or two. The advent of Industrial revolution created the concept of 'jobs' to work with machines. Jobs developed into careers as people advanced from a lower level job to the higher level. Industrial and information societies also created the concept of life styles as we know ourselves, with our clean and prosperous life, with resources for engaging in leisure and self-development activities.

What will be the future of work in the age of smart machines? No one can tell for sure. One does know that human ingenuity creates new possibilities. One lesson is to stop leaning too much on our jobs for our self-esteem. Today a job is considered to be not only a source of income, but also a source of meaning and social status. If income element could be taken care of, through Universal Basic Income, how important would a job be. The human need for challenge and status leads to many kinds of competitions to challenge ourselves, extend our perceived capabilities, and achieve recognition.

The media is full of stories about AI and its socio-political and economic impact. Many conferences are being devoted to AI. What kinds of challenges will remain in a world where machines would be able to do almost all the work, and be able to learn almost everything about our likes and desires and fears? What would we want to do? Ask a super-rich kid today what would he want to do? Perhaps he would want to engage in hedonism of the pleasures-seeking type. That approach could lead to aristocratic society that peaked just prior to the French Revolution. Or perhaps she would want to become more self-aware of her own divine nature and enjoy continuous bliss. That approach could lead to Ram Rajya where no one would need to worry about survival or being taken advantage of. Or she could want to become a creative person, engaged in creating art or knowledge or joyful environments for others.

## **AI and Existential Threat**

A very important perspective on AI is existential; whether an AI that is superior to human intelligence will ignore its human creators and masters. Will it honor its principles and not harm human beings? Will it even need human beings? Could superior AI do to the human beings what human beings did to

the apes and to elephants? The Future of Life Institute has been holding very informative conferences on this area with participation from scientists and entrepreneurs. People like Elon Musk have said that AI poses an existential threat to the human race. Others like Mark Zuckerberg of FaceBook have countered with his confidence in the ability to master AI. And finally, Mark Cuban says that whoever is able to master AI will be the world's first trillionaire!

## Conclusion

Artificial Intelligence is a way of understanding intelligence and embedding it in software and devices. AI is able to do many routine knowledge works. It poses a threat to the jobs we see today. However, the Industrial revolution before it, created many new industries even as it destroyed old jobs. AI has the potential to match and exceed human intelligence in the next 10-15 years. It has the potential to pose an existential threat to humanity.

## Review Questions

Q1: What is Artificial Intelligence? Why should we care?

Q2: Compare Neural Networks with Genetic Algorithms, listing two similarities and two differences. How does Big data relate to these techniques?

Q3: What are the lessons from the Industrial revolution that helps understand Cognitive or AI revolution? And the lessons from the Information Revolution?

Q4: Imagine and list 4 new types of jobs that will be commonplace in the next 5-7 years. Does any of them appeal to you?

Q5: What are the ethical issues involved in widespread deployment of Face Recognition systems.

## Chapter 20: Data Ownership and Privacy

Data is a digital resource that can be generated and shared by and with many entities simultaneously. Data revolution is so recent that most countries have had little time to develop ownership and privacy regimes for data. Data privacy begins with the expectation that data belongs to the subject of the activity that generated the data. The field of data privacy includes all the legal frameworks and technological tools related to gathering and using the data.

### CASELET: Cambridge Analytica

*Cambridge Analytica, a consulting company based in the UK, misused data of more than 50 million FaceBook users for political marketing, without the consent of those individuals. The initial activity was to collect data from a few hundred thousand Facebook users for academic purposes only, through a Facebook app by a researcher at Cambridge University. However, the app was able to collect the data from the entire personal networks of these individuals. That data was somehow shared with Cambridge Analytica, which used this data on tens of millions of individuals for political marketing in the USA. There were public hearings in the UK parliament, and FaceBook was called a ‘digital gangster’ for its role in these activities. FaceBook was also fined by the UK regulator its failure to safeguard people’s information. In July 2019, the US Federal Trade Commission FaceBook around \$5 billion for failing to act on the improper data gathering methods of Cambridge Analytica.*

*Q1: Is this a case of an illegal ‘data breach’? Who is responsible? What would be the downside to the companies involved?*

*Q2: Would Europe’s GDPR have stopped this activity?*

## Data Ownership

Data is generated by human beings as well as by machines. Many devices and technologies are involved in the creation and dissemination of data. The question arises whether every entity involved in the data production and communication chain can make a copy of the data and own it. Here are some instances:

- First, the person or entity about whom the data is generated, should have the right to control over its dissemination and use. It may even be covered by copyright law. For example, a video of a dance should give the dancer the right over the ultimate use of the video. The performer of the activity may have commissioned the recording in the first place.
- Secondly, the owner of the data generating device, such as a video or audio recording device, may have used skill and presence of mind to generate that data, and would have some ownership of that data. However, if a mobile device such as a smart phone is used for making the recording of the dance above, and uploaded to a public cloud account, the ownership should not belong to the maker of the device or the owner of the cloud service.
- Third, if the data is posted on a social media application such as FaceBook or Twitter, then the social media company would like to use that data for its own benefit. If the dance video is posted on a social media account, it becomes publicly available. However, it should be available for certain uses and not for other uses without the consent of the owner.
- Fourth, a collection of such data may be needed to develop and train an Artificial Intelligence application that may be developed by a national agency, and they may claim mandatory access to that data for training purpose.

Data ownership rights will almost necessarily work on top of copyright laws. Thus, anyone who can

demonstrated clear copyright ownership of a video, or image, or graph, etc. will naturally own that data. However, the terms and conditions of the use of a social media site such as FaceBook are such that the data may jointly belong to the social media site itself, in return for free sharing service.

## Data Privacy

Europe introduced GDPR (General Data Protection Rights) to provide a first comprehensive law about data protection and privacy. Many countries are still developing their data privacy regimes. US, in particular, is far behind Europe in guaranteeing data privacy rights to its citizens. India is developing a data privacy policy that essentially imitates GDPR.

### Data Privacy Models

There are three different models of data privacy. At one end is the Chinese model where there is not much privacy. At the other end, there is the US model which is commerce-driven where the consumer has essentially no inherent data rights, and vendor organizations are responsible for managing privacy through consent agreements. Then there is the European model where the state takes cognizance of data privacy and enshrines certain rights for the consumer. Many other countries are taking this middle of the road approach to balance the rights of the individuals and the support for continuous innovation. All three models will be discussed next.

### Chinese Model

There is little or no expectation or provision of data privacy. Most young people in China express no interest or expectation of data privacy. Why fear anything, if one has done nothing wrong, say these people. There has been an intense movement of the society towards using the digital platforms, most notably WeChat. This mobile platform enables a wide range of activity and payments whereby all data is captured in digital form. Face-based recognition by the government and other organizations deployed to monitor individuals for security purposes. Only time will tell if this is the right approach to take.

### US Model

Privacy is very important to the average US citizen. However, data privacy has been difficult to enforce. Financial data is used for credit scoring of individuals, and such data is regulated by a few organizations. Financial organizations take great pains to protect the identity and financial details of all individuals. Whenever data breaches happen, there is a big hue and cry, and the organizations lose reputation and market value. The Big Technology companies such as Google and Facebook collect extremely large amounts of data about individuals and their online activity. This data becomes the platform to generate new and profitable customized services for individuals. Mobile apps collect huge amount of data without telling the users, or that consent is buried in the small print of the terms and conditions of use.

### European Model

Europe has taken great cognizance to need to protect individual's privacy and codified certain rights as General Data Protection Rights (GDPR). This law came into force in May 2018, and it applies to all entities operating in any or all parts of Europe. It provides for stiff penalties for companies violating those data privacy rights. Thus, it is important to understand the approach and the specifics of this law.

The key principles relate to (a) what data can be collected and stored, (b) how that data might be processed and used.

Data should be collected and stored only for explicit, specific and legitimate reasons. It should be collected with the consent of an individual. The individual has the right to withdraw consent for the storage of data. . Data owners (about whom the data was) have the right to be forgotten. i.e. data about them must be deleted after a certain time, if they so demand. The individual has the right to see that the

data is accurate, and if needed the data should be corrected. The data should be stored in the geography of the individual. The data should be retained only for the duration of the specific purpose and then discarded.

Data should be processed in a lawful and transparent manner. It should be processed to meet specific purposes such as to meet legal or contractual obligations. It may be processed when certain vital public interest is involved. It should serve the transparent and legitimate interests of the individual of the organization.

GDPR prescribes a penalty of US\$20 million or 4% of global revenue of the company, whichever is greater. This amount should be prove to be a good deterrent, and the regulators have shown the propensity to enforce these laws. No major such penalty has been issued under GDPR so far though. There is a growing stream of data privacy professionals who advise organizations on being on the right side of law, to protect themselves from financial and reputational loss.

## Summary

Data can be put together, using software tools, to identify and target individuals without their consent. This requires a new data privacy regime to protect individuals against giant organizations. Different countries and regions follow different regimes in keeping with their cultures and political philosophies. Europe's GDPR system is the new gold standard in striking a good balance between individual's privacy and economic innovation. Data privacy professionals are in increasing demand to guide organizations to be ready for the future.

## Chapter 21: Data Science Careers

Data Science is rapidly growing and maturing field of work. A useful parallel is to compare its rise with the rise of Computer science. Just like Computer Science took a couple of decades before it became mainstream, so also Data Science is becoming mainstream now. Computer Science had its first course offerings in the early 1960s. At this moment, every university is already offering, or developing, a Data Science curriculum.

A data science major takes of a wide range of jobs such as Data Engineer, Data Analyst, Data visualizer, Data Warehousing specialist, Machine learning specialist, and so on. Data scientists are earning high salaries. Fortunately, many online data science courses are available for just in time learning. This field evolves very rapidly, so one should continue to enhance one's skills.

### Data Scientist

Everyone can become a data scientist. A data scientist works through the entire BIDM cycle (see Figure 1.1) to help grow the business. A good data scientist knows how to identify and focus on the high-priority problems. He/she finds the right angle to approach the problem, dives into the data, and seeks new insights and models that the data may throw up, that can solve the problem.

The data scientist works through the entire data processing chain from gathering, preparing, modeling, warehousing, analyzing, and visualizing the data for easy consumption by business experts. The ideal data scientist is a multi-disciplinary person, persistent in pursuing the solution, and well-versed in business management, databases, statistics, artificial intelligence, and communication skills. The data scientist finds bliss in immersing into the data, and finding patterns that ultimately align with the unified field of all the laws of nature.

### Data Engineer

A Data Engineer focuses on ownership of data. It includes continuous gathering data from many data pipelines, organizing it to suit business needs, and then manipulating it to make it ready for analytics. A Data Engineer works with business analysts, software engineers, and data scientists to modify the database design and performance as needed. The Data Engineer may also work as a Data Warehouse engineer and work with ETL scripts to

keep the data warehouse up to date. The Data Engineer defines and builds new data pipelines for faster, better, data-informed decision-making within the business.

## Data Science aptitude

There is no one really defining skill for a data scientist. The real test is how strongly do you want to do data science. Are you willing to put your heart and effort into becoming good at it?

The first self-test is your relationship with data. Reflect upon yourself and see how much do you love working with data to solve problems or develop insights? What have you done with data lately? Collect a small dataset and start playing with it.

Analyze and visualize a new dataset. See what kind of insights you can generate out of the data. Can you visualize and communicate those insights? Does that process excite you? Visit kaggle.com to see free data sets, and find out the kind of data science challenges are being offered for prize money.

Data preparation takes about 70-80% of the time of a data science project. It can be tedious work but very critical. Without good data, one gets GIGO (garbage in, garbage out). It requires patience and attention to detail.

Executives will not trust the data-based insights immediately. They will ask for the data to be looked at in many more ways before they will accept the results. Curiosity, persistence and creativity are needed in any work, and more so in data science.

Download R and RStudio. Use it to complete the exercises from the tutorial in the Appendix. Does that process look easy and joyful to you? Do you enjoy programming in R?

## Popular Skills

Data Science is a truly multidisciplinary discipline. The popular programming languages for data analytics are the open source languages such as R, Python, and Scala. There are many commercial GUI platforms from major providers such as IBM SPSS, SAS, and others. There are also open-source GUI-based platforms such as Weka.

There are many good sources on the web for advice on data science careers,

and education opportunities. Here is a good one:  
[www.mastersindatascience.org/careers/data-scientist/](http://www.mastersindatascience.org/careers/data-scientist/)

## **Appendix: R Tutorial for Data Mining**

**This R Tutorial for Data Mining was developed for academic  
use**

**by Dr. Anil Maheshwari**

**with contributions from graduate students:**

**Mr. Tonmay Bhattacharjee**

**Ms. Marie Claire Niyirora**

**Mr. Mauro Nogueira**

## Getting Started with R

As the R Project website describes, “*R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.*”.

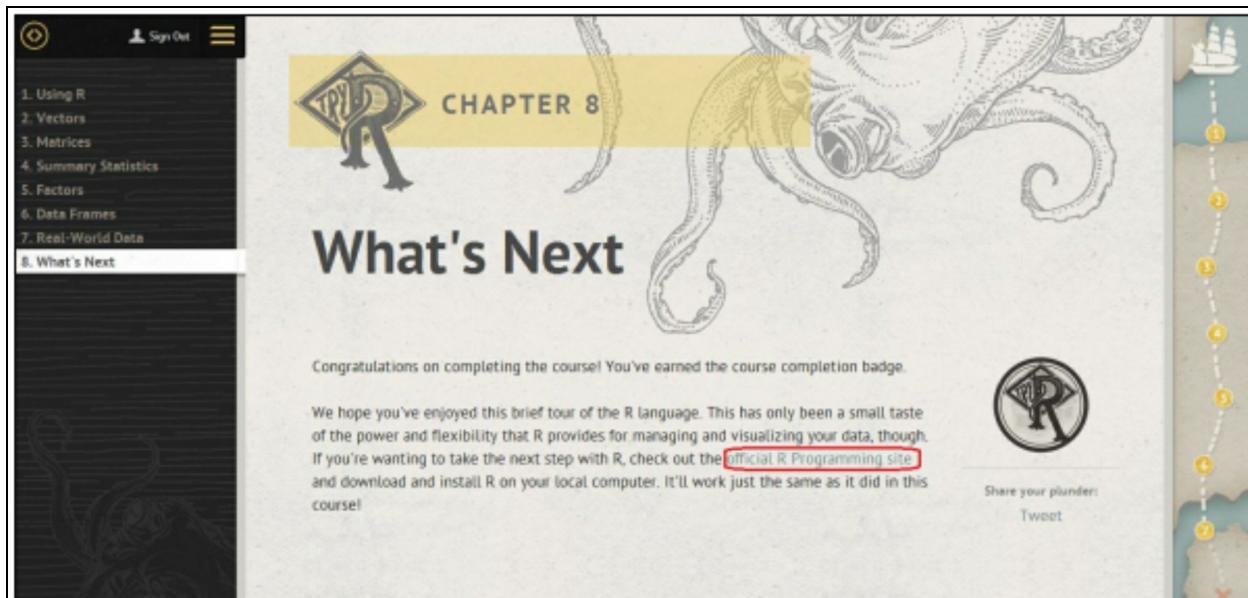
The first step to understand how R works, type at Google “code R” and go to the **R code school website**. You can directly go to <http://tryr.codeschool.com> too.

Sign up/register providing the simple information.

Follow the simple instruction and practice on the given code window.

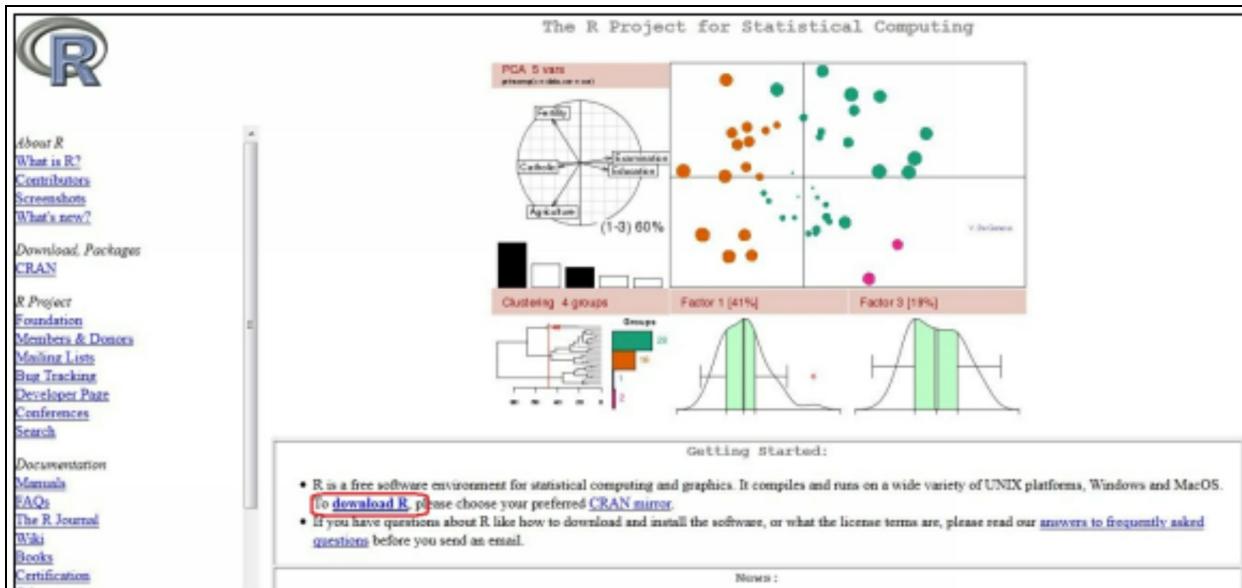
Finish the step and unlock the next steps.

Finish all seven steps and you’ll see a congratulation page like bellow.



## Installing R

Click on the official R programming site (<http://www.r-project.org>). Search for the Download section



Click “download R” to get the proper mirror. This should take you to a page something like bellow.



Choose the link for Iowa State University or any other mirror you like.



The Comprehensive R Archive Network

[Download and Install R](#)

Precompiled binary distributions of the base system and contributed packages. Windows and Mac users most likely want one of these versions of R.

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

[Source Code for all Platforms](#)

[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)  
[About R](#)  
[R Homepage](#)

Choose your operating system. In the tutorial case it was selected Microsoft Windows.



**R for Windows**

Subdirectories:

<a href="#">base</a>	Binaries for base distribution (managed by Duncan Murdoch). This is what you want to <a href="#">install R for the first time</a> .
<a href="#">contrib</a>	Binaries of contributed packages (managed by Uwe Ligges). There is also information on <a href="#">third party software</a> available for CRAN Windows services and corresponding environment and make variables.
<a href="#">Rtools</a>	Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)  
[About R](#)  
[R Homepage](#)  
[The R Journal](#)  
[Software](#)  
[R Sources](#)

Click “install R for the first time”.



**R-3.1.2 for Windows (32/64 bit)**

[Download R 3.1.2 for Windows \(54 megabytes, 32/64 bit\)](#)

[Installation and other instructions](#)  
[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

[Fingerprint](#) [1w...asked](#) [envelope](#) [Log in](#)

[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

Click on “download” to download the installer (for Windows ).

[Download R 3.1.2 for Windows](#) (54 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check the [fingerprint](#), You will need a ...

- [How do I install R without administrator rights?](#)
- [How do I update packages?](#)
- [Should I run 32-bit or 64-bit R?](#)

Opening R-3.1.2-win.exe

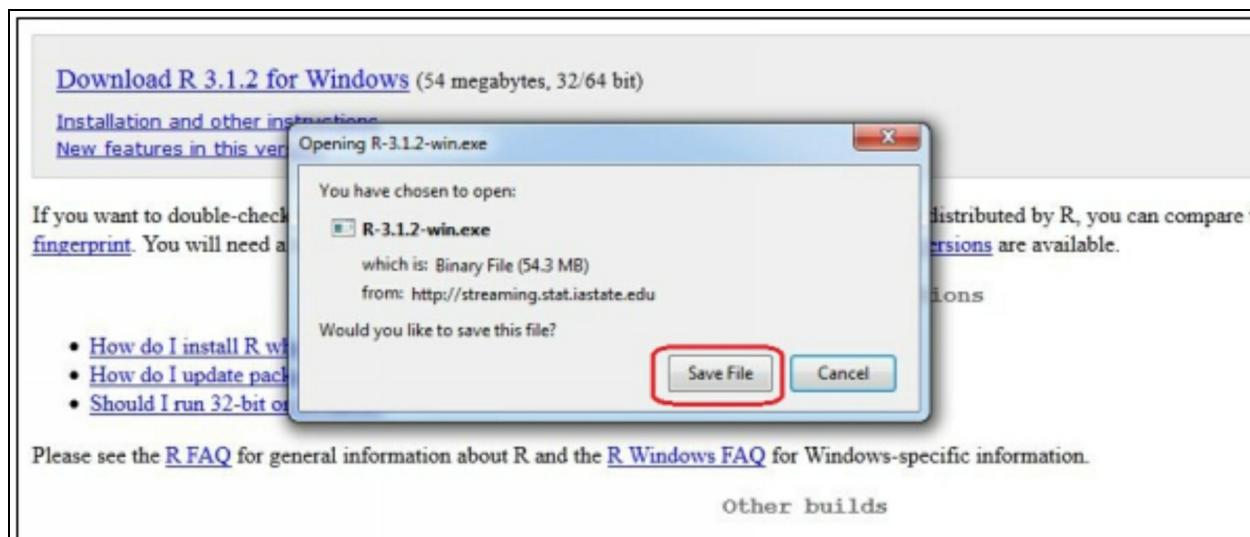
You have chosen to open:

R-3.1.2-win.exe  
which is: Binary File (54.3 MB)  
from: http://streaming.stat.iastate.edu

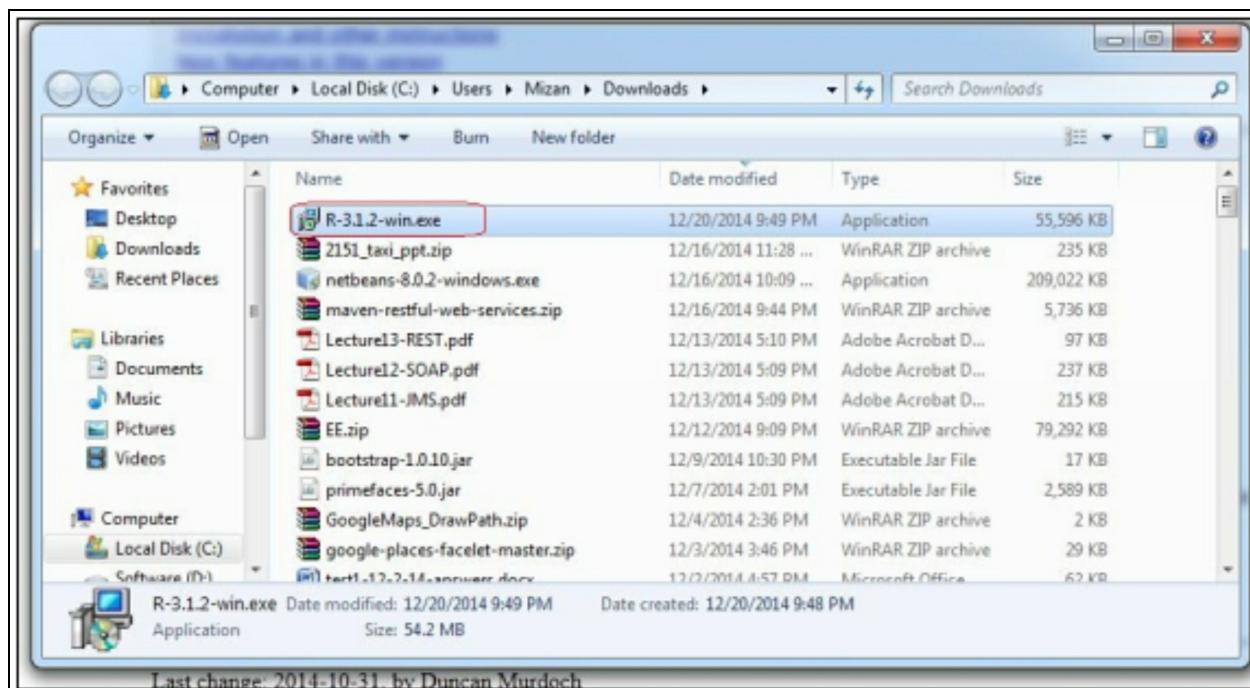
Would you like to save this file?

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

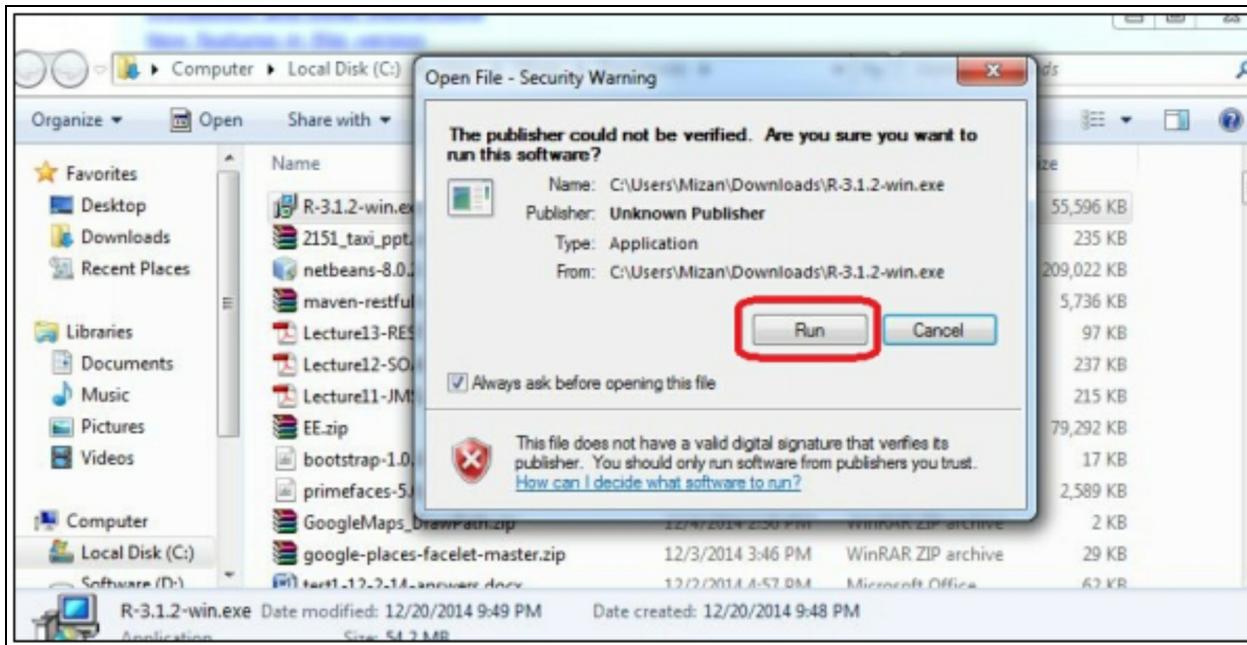
[Other builds](#)



Click on “Save file” to save the installer to your computer.



Double click the .exe file for installation.



Click “Run” to start the installation. Follow the steps. Click next, accept agreement, select your installation folder and finish the installation.

## Working on R

Select the R application from your start menu. All coding style should be same what you practiced on R code school.

Get and set working directory.

```
getwd()
setwd("/path/to/your/working/directory")
```

```
R Console
> #check working directory
> getwd()
[1] "C:/Users/nimacilare/Documents"
> #set working directory to another directory (ex:r_document) Note:Replace backslash by forward slash
> setwd("F:/Maharishi University of Management/Courses/Big_data_analytics/r_document")
> getwd()
[1] "F:/Maharishi University of Management/Courses/Big_data_analytics/r_document"
> |
```

List of R commands

Here are some commands that you can use in your R script.

Command/word	Description	Example
#	The characters after are considered comment	#this is a comment

<b>cat</b>	Print a message to the console (like System.out.println in Java)	cat("Hello World!\n")
<b>source</b>	Execute a R script file previously created	source("myScript.r")
<b>&lt;-</b>	Assign a value to a variable	one <- 1 Hello <- "Hello"
<b>install.packages</b>	Download and install a package	install.packages("rpart")
<b>library</b>	Load a installed library to your R script	load(rpart)

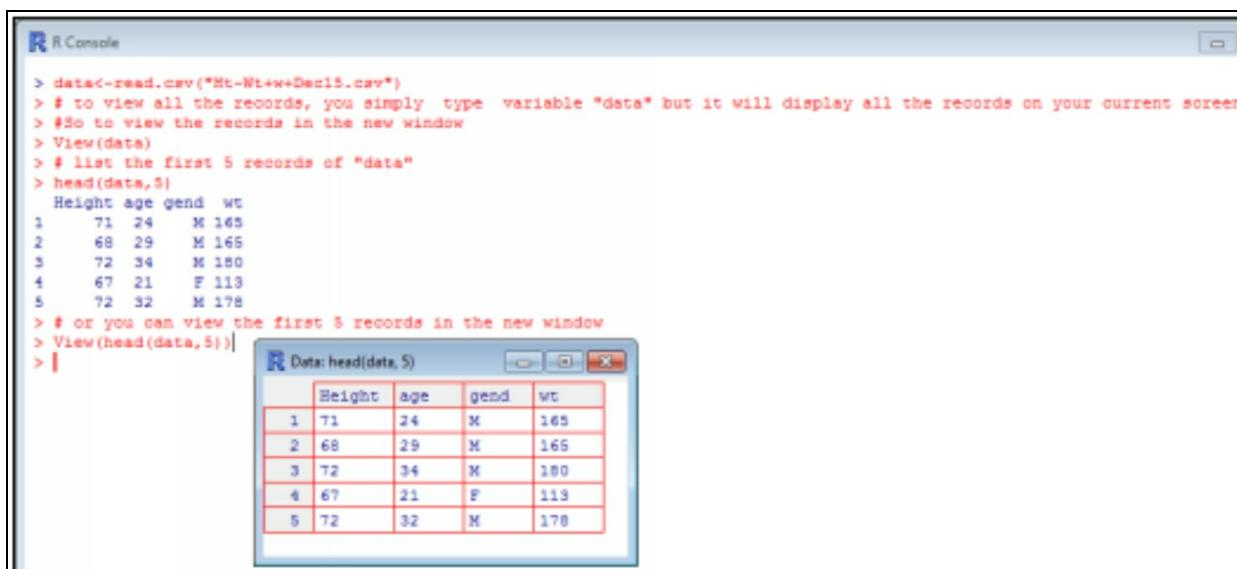
Note: R is case sensitive, so **Variable** is different from **variable**.

## Import a Dataset in R

Before importing your dataset you must convert your .xls or .xlsx file to .csv file and save it in the working directory. Otherwise you will be required to specify its path when you are importing the data in R.

Load the data to a variable using `read.csv` command. For example, for the screen below we used the following command:

```
data <- read.csv("Ht-Wt+w+Dec15.csv") #Read data from the csv file and assign to variable
# "data"
View(data) #show all values from the "data" variable
head(data,5) #print in the console the first 5 rows from "data" variable
View(head(data,5)) #show the data loaded into "data" variable (5 first records)
```



The screenshot shows the R Console window. The user has run the following R code:

```

> data<-read.csv("Ht-Wt+w+Dec15.csv")
> # to view all the records, you simply type variable "data" but it will display all the records on your current screen
> #So to view the records in the new window
> View(data)
> # list the first 5 records of "data"
> head(data,5)
  Height age gend wt
1    71   24     M 165
2    68   29     M 165
3    72   34     M 180
4    67   21     F 113
5    72   32     M 178
> # or you can view the first 5 records in the new window
> View(head(data,5))
>

```

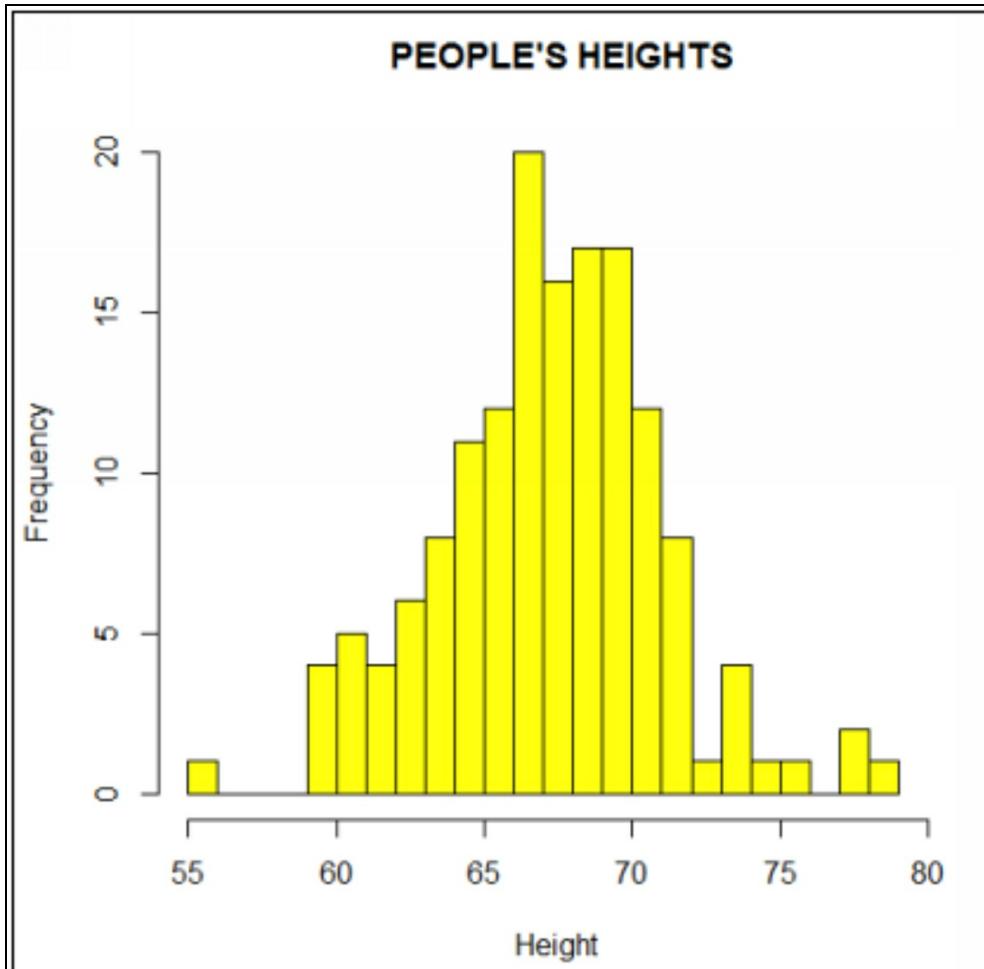
A modal dialog box titled "Data: head(data, 5)" is displayed, showing the first five rows of the dataset:

	Height	age	gend	wt
1	71	24	M	165
2	68	29	M	165
3	72	34	M	180
4	67	21	F	113
5	72	32	M	178

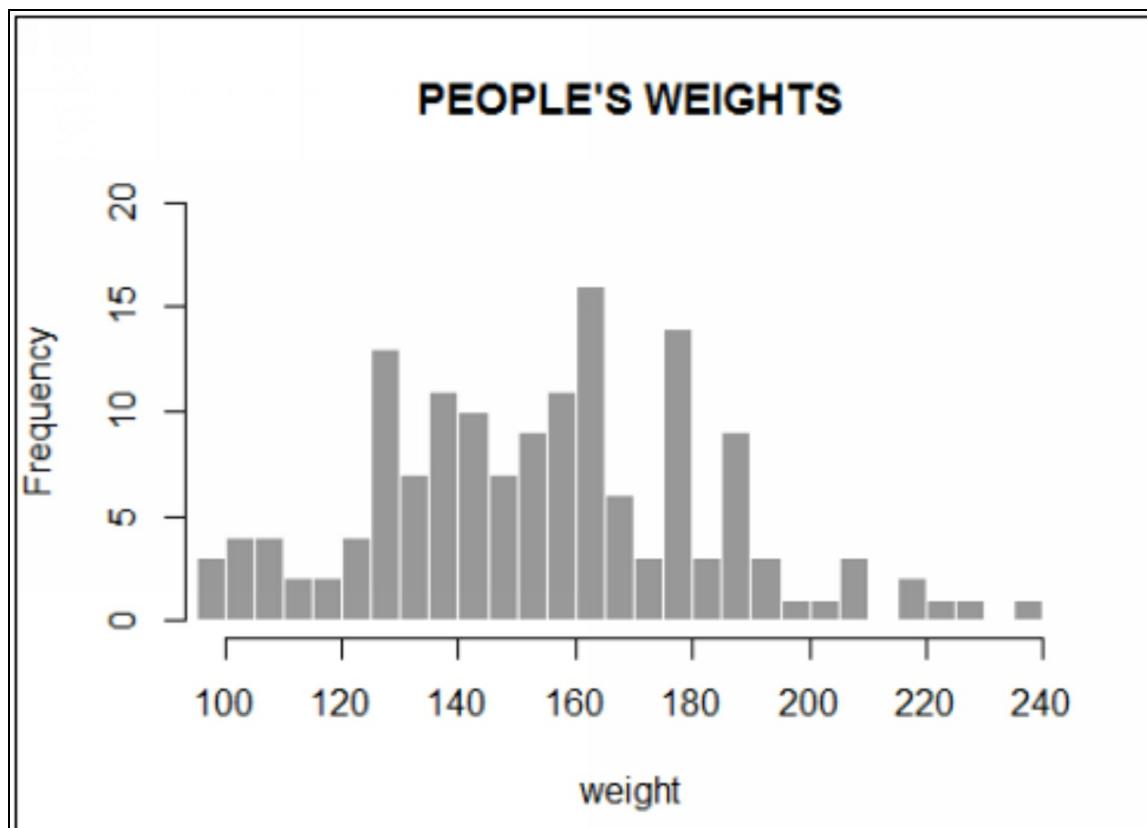
# Data visualization

## Plotting a Histogram

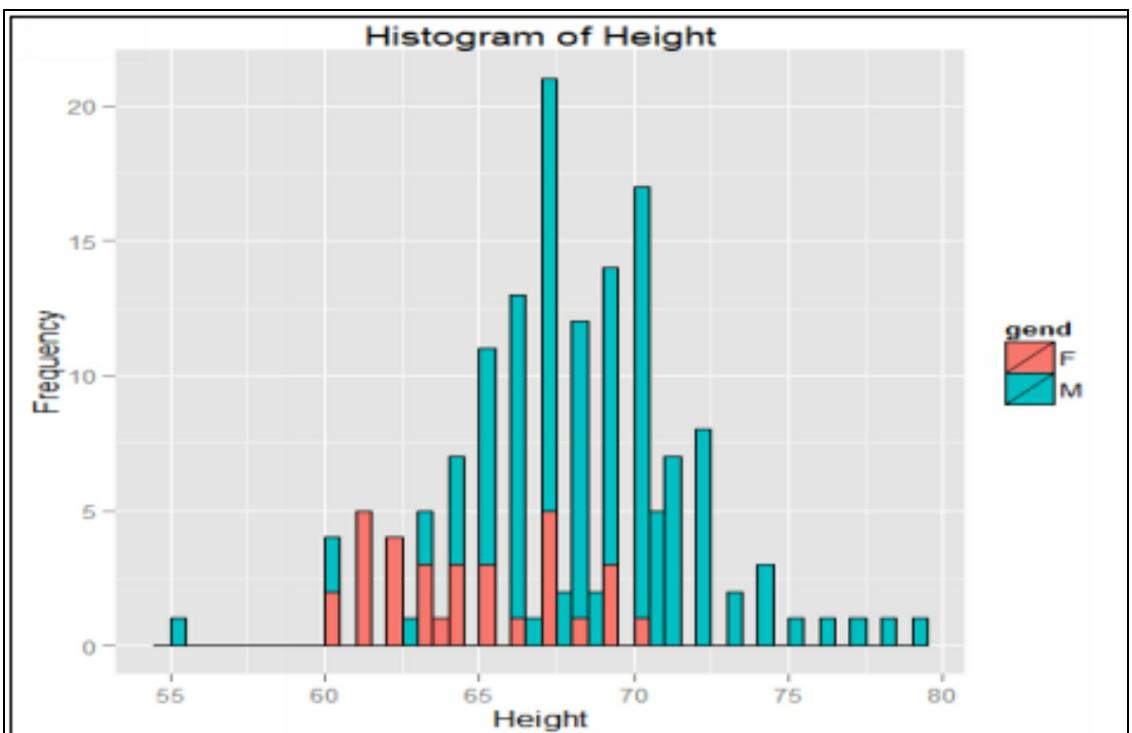
```
df <- read.csv("Ht-Wt+w+Dec15.csv") #Read data from the csv file and assign to variable  
"df"  
hist(df$Height,ylim(c(0,20)),xlim=c(55,80),main= "PEOPLE'S HEIGHTS",  
breaks=32,xlab="Heighth", col="#FFFF0F",border="#000000") #plot the histogram height
```



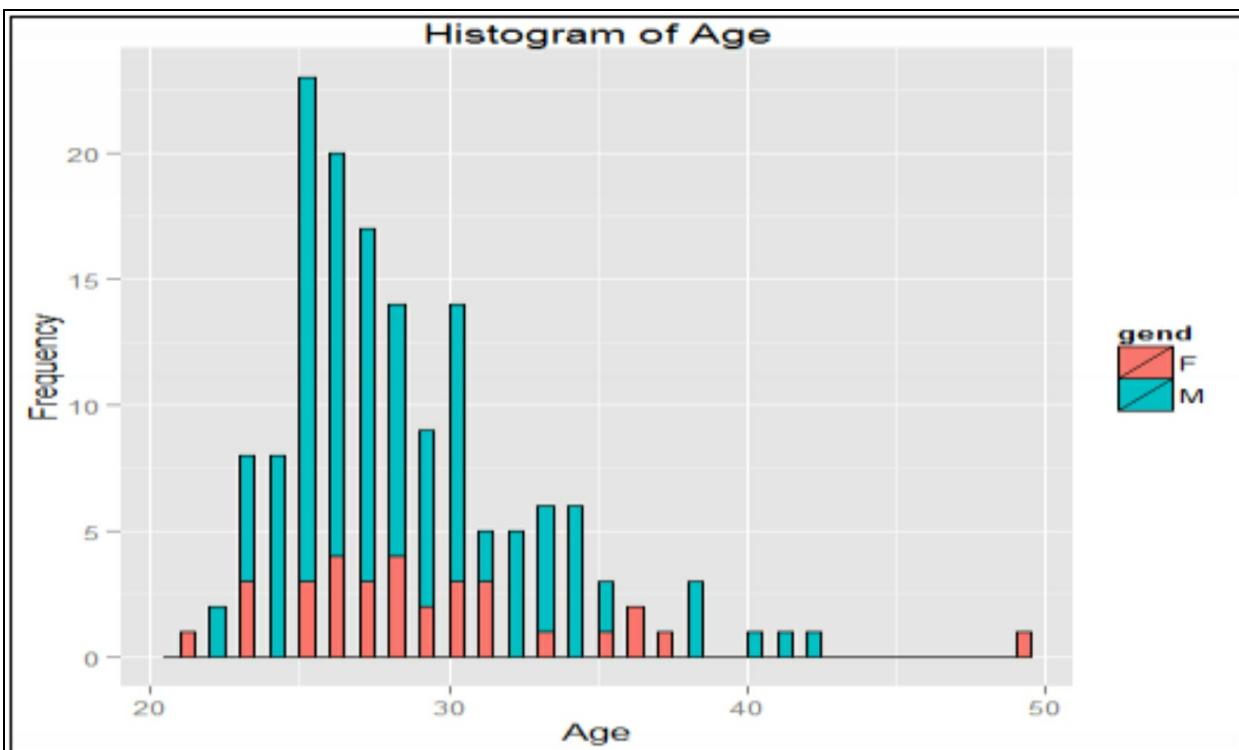
```
hist(df$wt,ylim(c(0,20)),xlim=c(99,240),main= "PEOPLE'S WEIGHTS",  
breaks=32,xlab="weight", col="#999999",border="#ffffff") #plot the histogram for weight
```



```
library(ggplot2) #load the library ggplot2, needed for ggplot command. If not installed, try  
install.packages("ggplot2")  
ggplot(df,aes(x=Height,fill=gend))+geom_bar(position="stack",binwidth=0.5,col="black")+sc  
of Height")
```

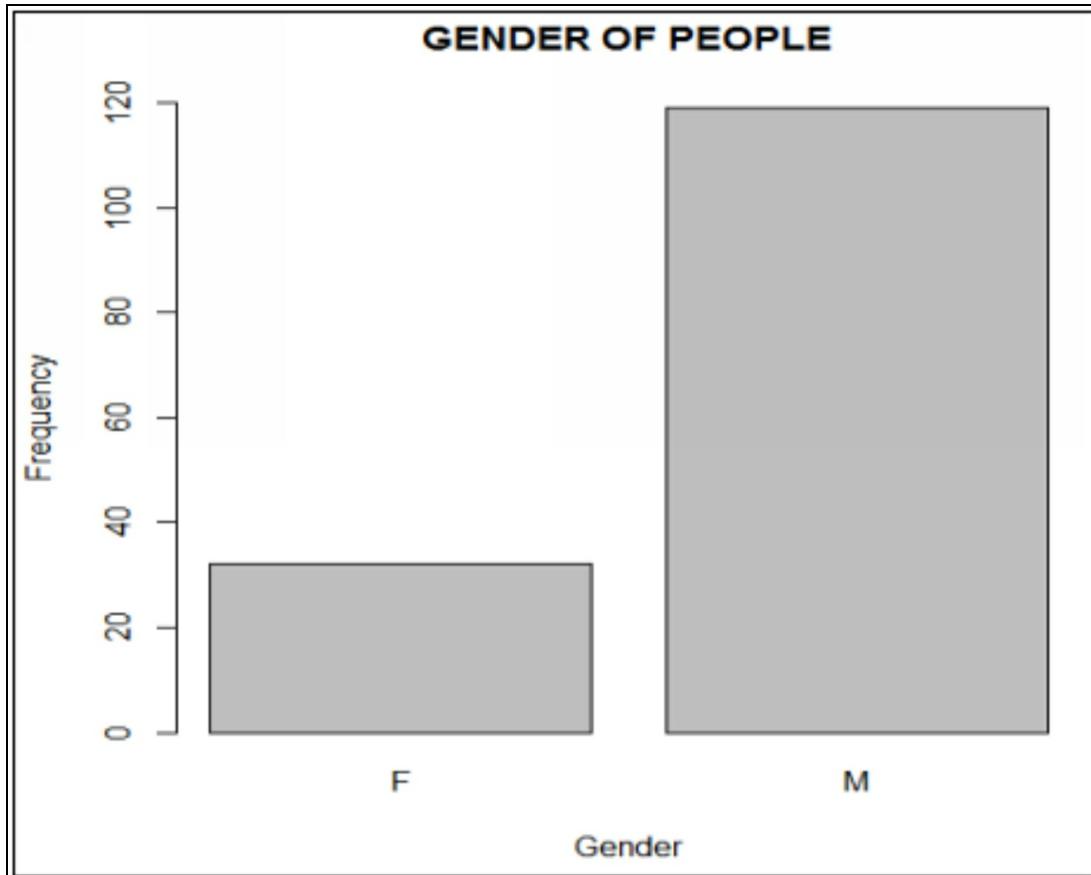


```
ggplot(df,aes(x=age,fill=gend))+geom_bar(position="stack",binwidth=0.5,col="black") + scale_x_continuous("Age")
```



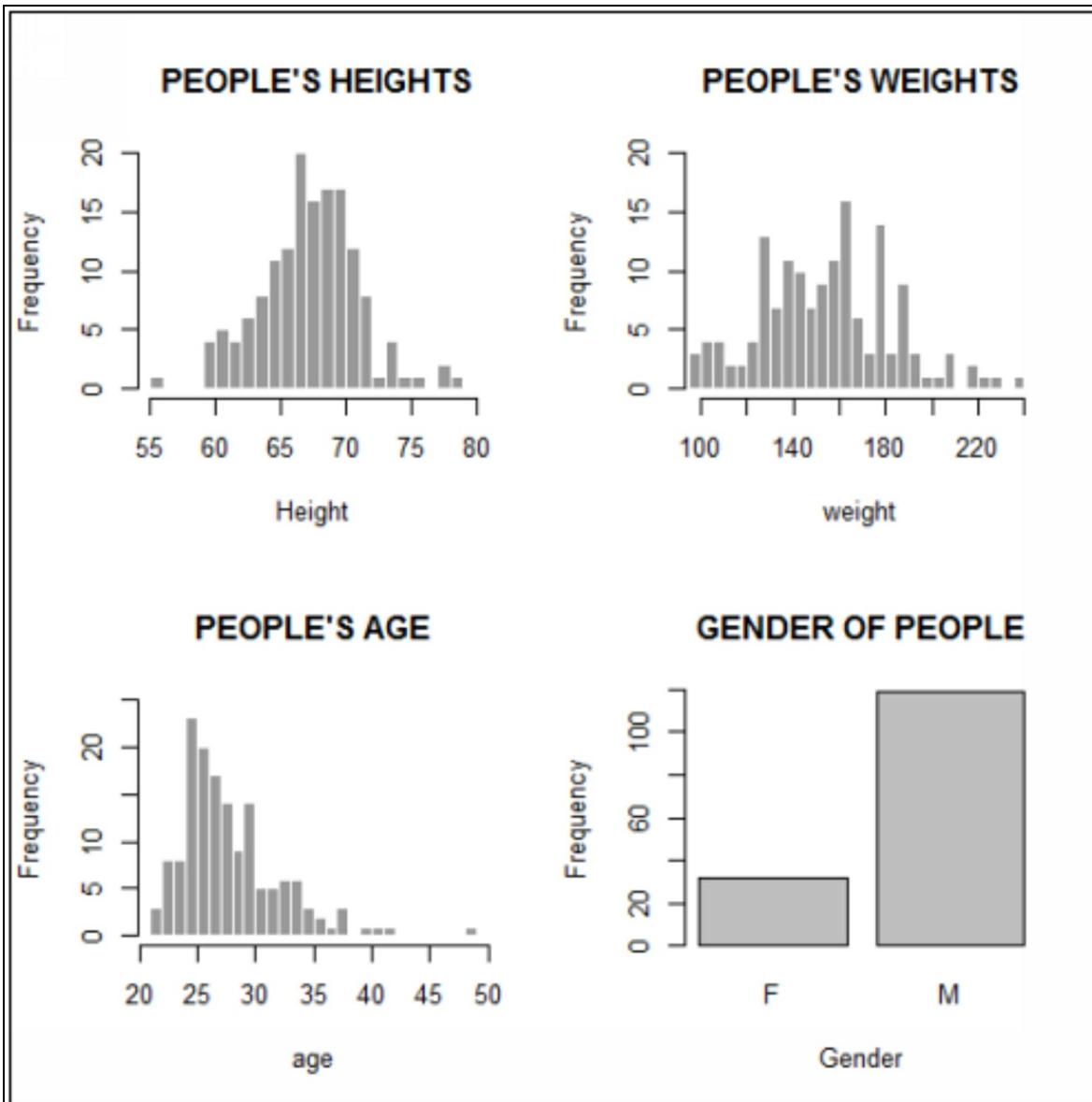
### Ploting a Bar Chart

```
barplot(table(df$gend),ylab="Frequency",main="GENDER OF  
PEOPLE",xlab="Gender",ylim(c(0,120)) #plot a bar chart for gender
```



### Ploting charts side by side

```
Par(mfrow=c(2,2))  
hist(df$Height,ylim(c(0,20)),xlim=c(55,80),main= "PEOPLE'S HEIGHTS",  
breaks=32,xlab="Heighth", col="#FFFF0F",border="#000000") #plot the histogram height  
hist(df$wt,ylim(c(0,20)),xlim=c(99,240),main= "PEOPLE'S WEIGHTS",  
breaks=32,xlab="weight", col="#999999",border="#ffffff") #plot the histogram for weight  
hist(df$age,ylim(c(0,25)),xlim=c(21,49),main= "PEOPLE'S AGE", breaks=32,xlab="age",  
col="#999999",border="#ffffff") #plot the histogram for age  
barplot(table(df$gend),ylab="Frequency",main="GENDER OF  
PEOPLE",xlab="Gender",ylim(c(0,120)) #plot a bar chart for gender
```



## Data Mining Techniques

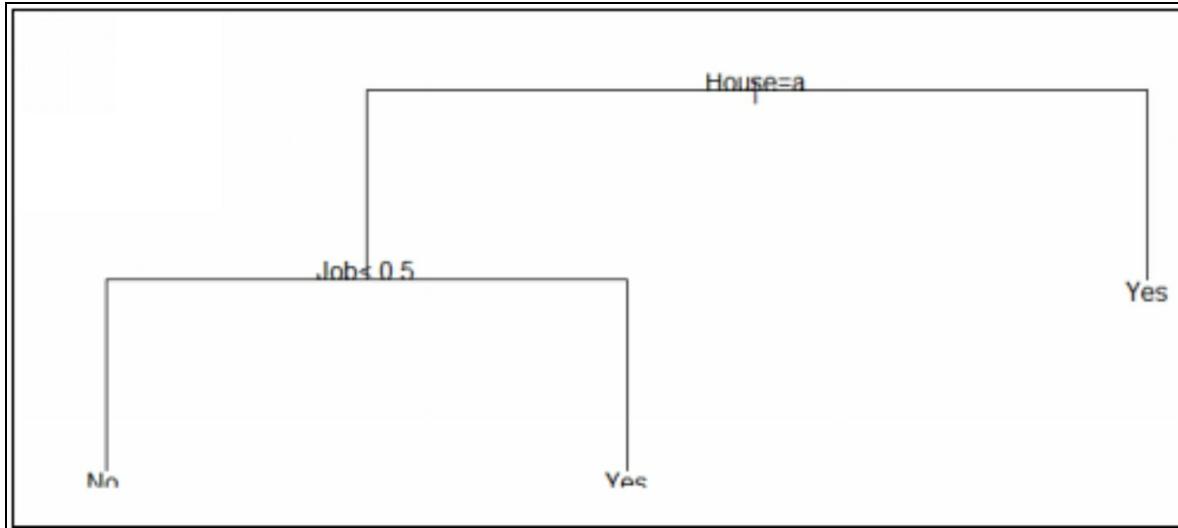
### Decision Tree

```

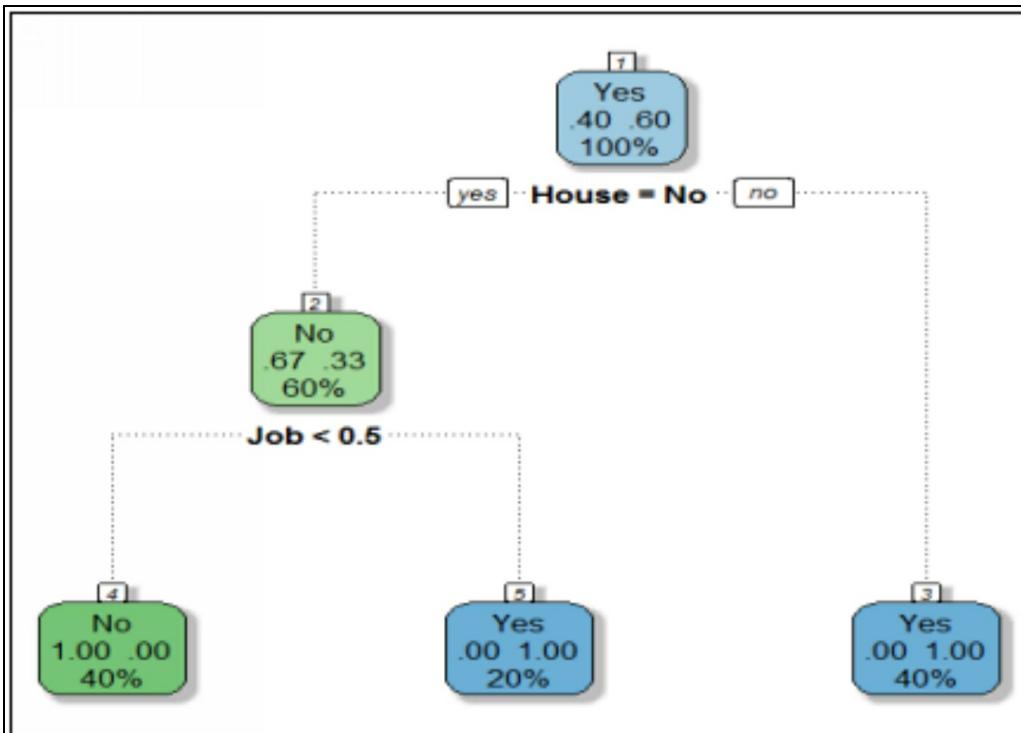
library("rpart") #load the library for decision tree
# Draw the tree and assign to any variable (in our case, "myTree")
# Here Age, Job, House and Credit are column names and we are using decision tree to
predict "LoanApproved" based on those columns, loanData is the variable name and
# method=class stands for classification.
loanData <- read.csv("Loan-Approval+dataset.csv") # read dataset
loan_tree = rpart(formula = LoanApproved ~ Age + Job + House + Credit, data = loanData,
method = "class", control = rpart.control(minsplit = 5))

```

```
plot(loan_tree) # plot the tree  
text(loan_tree) # Insert labels on tree
```



```
# To make the tree little bit fancy, install some extra packages  
install.packages("rattle")  
install.packages("RColorBrewer") # Don't forget that R is case sensitive  
install.packages("rpart.plot")  
library(rattle) # Load rattle library  
library(RColorBrewer)  
library(rpart.plot)  
fancyRpartPlot(loan_tree) # We are using the same "loan_tree" variable previously assigned
```



### Correlation

In the same way described in decision tree, you can install the necessary library and load the data.

Using `cov(data)` you can see relation.

Using `pairs(data)` you can see the regression.

```

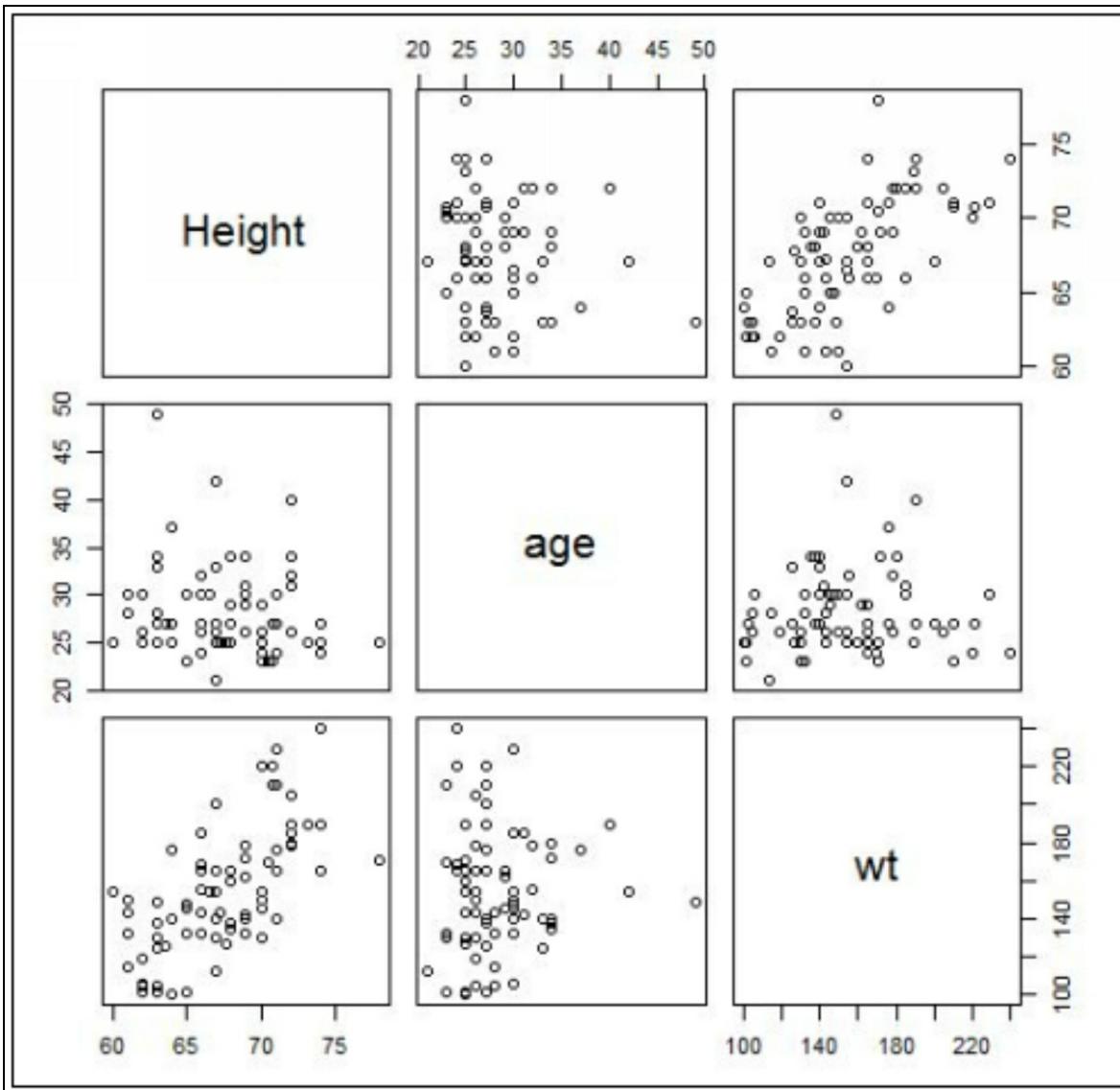
library(rattle) # If not loaded previously
library(rpart.plot)
library(RColorBrewer)
data <- read.csv("height.csv")
cov(data)
pairs(data)

```

```

> data<-read.csv("height.csv")
> cov(data)
      Height         age          wt
Height 14.750943 -1.892988   79.413229
age     -1.892988 21.559608   5.685839
wt      79.413229  5.685839 1028.802734

```



## Regression

With R is possible work with multiple linear regression.

```
myData <- read.csv("Ht_Wt_wMarch16.csv")
head(myData,5)
  Height age gend wt
1   71  24   M 165
2   68  29   M 165
3   72  34   M 180
4   67  21   F 113
5   72  32   M 178
```

```
x <- lm(myData$wt ~ myData$Height + myData$age +
myData$gend) #Runs the correlation
summary(x)
```

Call:

```
lm(formula = myData$wt ~ myData$Height + myData$age +
myData$gend)
```

Residuals:

Min	1Q	Median	3Q	Max
-46.405	-19.329	-1.992	16.141	65.594

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-59.7711	37.7863	-1.582	0.115678
myData\$Height	2.7103	0.5571	4.865	2.73e-06 ***
myData\$age	0.5700	0.4251	1.341	0.181844
myData\$gendM	19.9375	5.1898	3.842	0.000176 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.61 on 159 degrees of freedom

Multiple R-squared: 0.3024, Adjusted R-squared: 0.2893

F-statistic: 22.98 on 3 and 159 DF, p-value: 2.068e-12

coefficients(x)

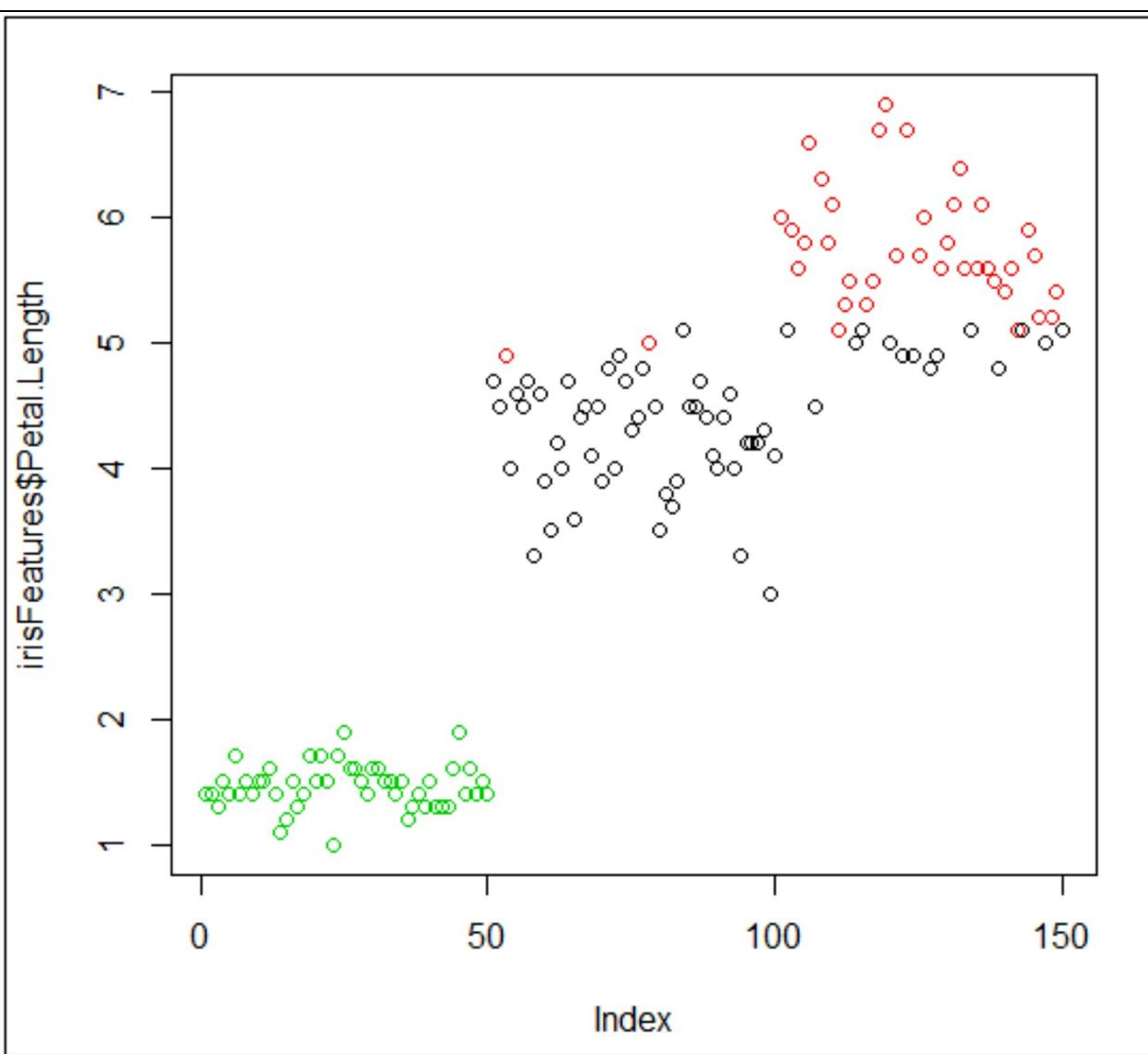
(Intercept)	myData\$Height	myData\$age	myData\$gendM
-59.7710660	2.7102542	0.5700159	19.9374996

## Clustering – Kmeans (Unsupervised Learning)

There are different ways of plotting k-means clustering results

### Method 1

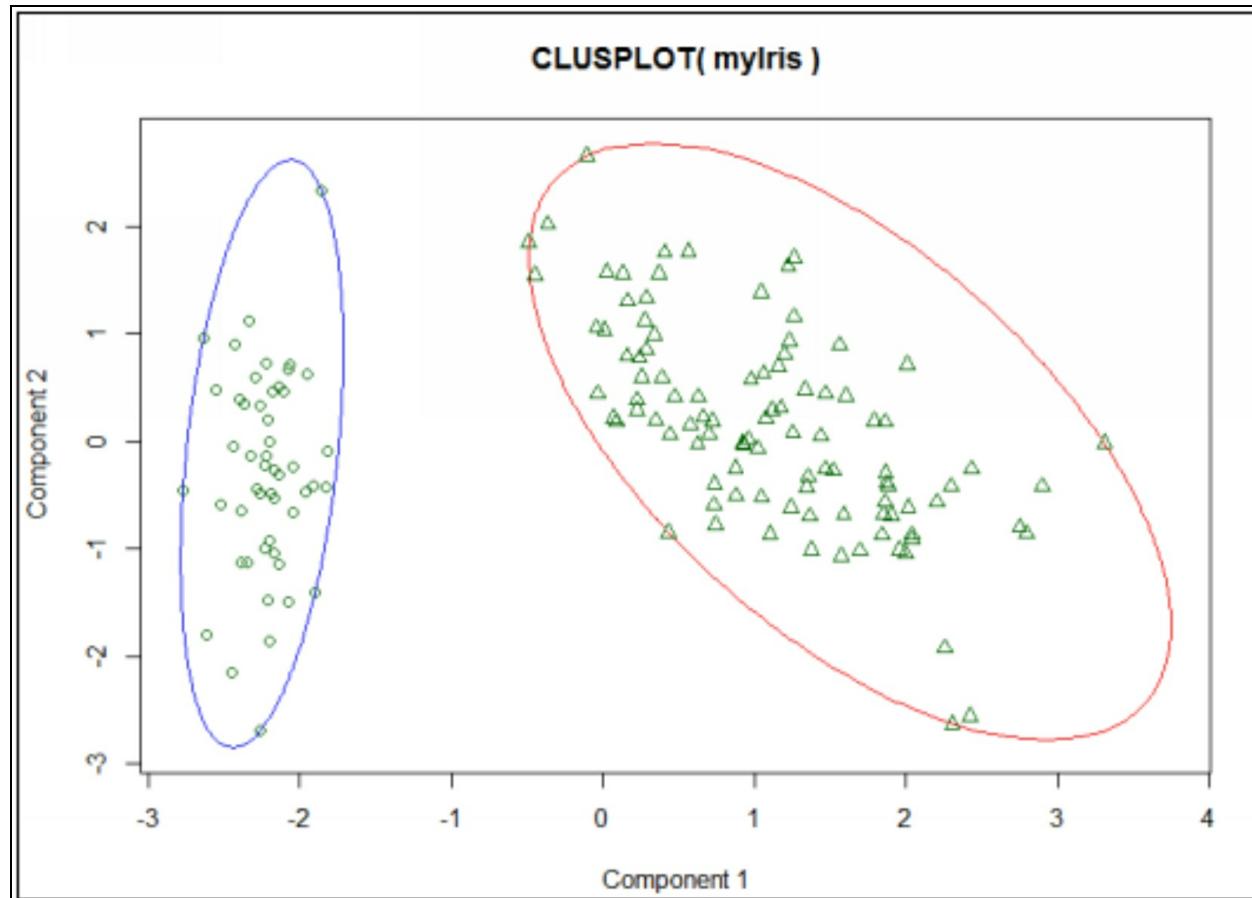
```
library(datasets)
irisFeatures <- iris[-5]
km <- kmeans(irisFeatures,3)
km$size
km$cluster
plot(irisFeatures$Petal.Length, col = km$cluster)
```



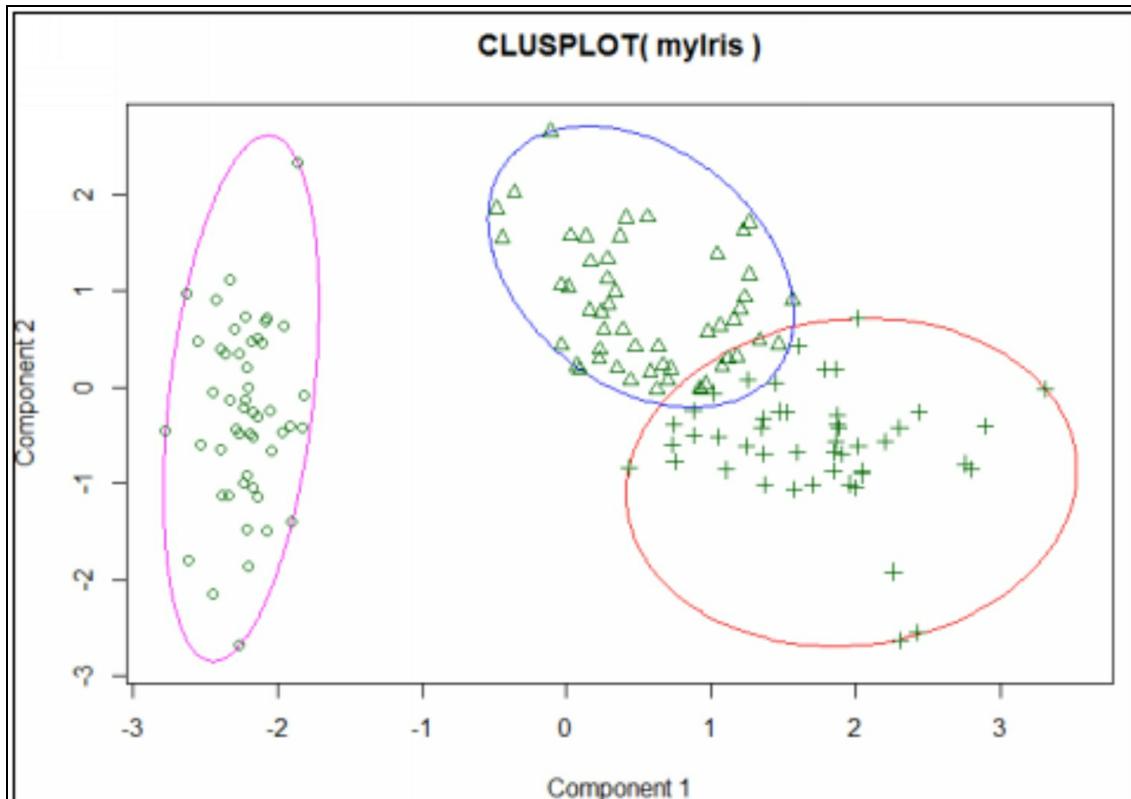
### Method 2

```
km # Show all the information about km  
data(iris) # Using R Iris dataset  
head(iris,3) # Just checking the first 3 rows of the dataset
```





```
#Now, using different parameters to plot...
# k = 3 with 5 iterations
irisKM <- kmeans(myIris, 3, iter.max = 5)
clusplot(myIris,IrisKM$cluster, color = TRUE, lines = 0)
```



## Big Data Mining

### WordCloud

Using file contents

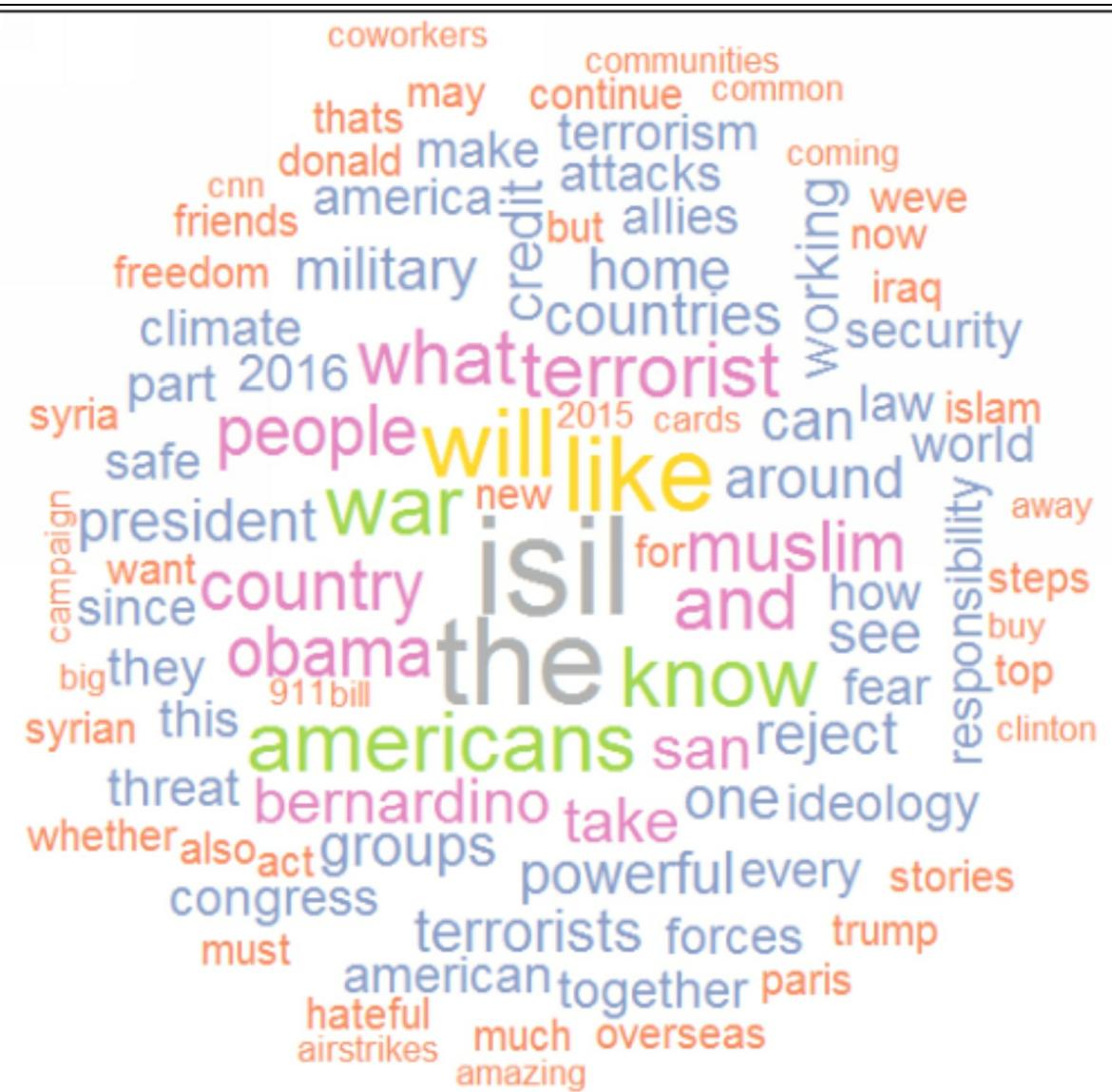
```
# Install the packages TM, wordcloud, RColorBrewer
# install.packages(c("tm", "wordcloud", "RColorBrewer"))
# Import the libraries
library(tm)
library(wordcloud)
library(RColorBrewer)

# select your file name below
#In my case I used obama speech from this link
#http://www.cnn.com/2015/12/06/politics/transcript-obama-san-bernardino-isis-address/
articleCorpus <- readLines("obamaSpeech.txt")
# Read text as a vector
# Create corpus
mycorpus = Corpus(VectorSource(articleCorpus))
# Transformation
mycorpus1 = tm_map(mycorpus, removePunctuation)
mycorpus2 = tm_map(mycorpus1, removeWords, stopwords("english"))
mycorpus3 = tm_map(mycorpus2, tolower)
```

```

mycorpus4 = tm_map(mycorpus3, stripWhitespace)
mycorpus5 = tm_map(mycorpus4, PlainTextDocument)
# Create Term document matrix. A document-term matrix or
# term-document matrix is a mathematical matrix that describes
# the frequency of terms that occur in a collection of documents.
tdm <- TermDocumentMatrix(mycorpus5)
m1 <- as.matrix(tdm)
# Sort data by frequency
v1<- sort(rowSums(m1), decreasing=TRUE)
# create dataframe with data and count/frequency
d1 <- data.frame(word=names(v1), freq=v1)
d1 <- head(d1,100)
# Create word cloud
wordcloud(d1$word,d1$freq,col=brewer.pal(8,"Set2")), min.freq="5",random.order=FALSE)

```



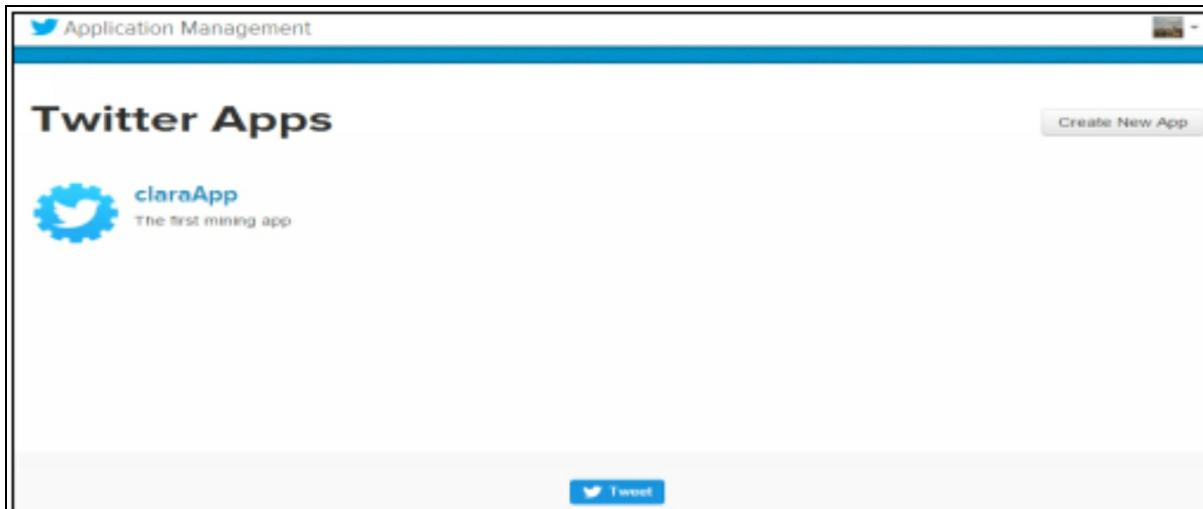
**OBS:** type `?wordcloud` in the R Console to get documentation about wordcloud

## Twitter Mining

- Using twitter hashTag mining

### Steps on Twitter side

- Sign in using your twitter account
- <https://twitter.com/apps>



- Click on create a New App

## Create an application

### Application Details

**Name \***

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

**Description \***

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

**Website \***

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.  
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

**Callback URL**

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

- Under "keys and Access Tokens" tab, check the consumer key (API key), consumer secret (API secret), access Token and Access Token Secret.

# TwitterMiningApp

Details    Settings    Keys and Access Tokens    **Permissions**

## Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) T3OfRlpWVgjIPVu1LARuQVEqf

Consumer Secret (API Secret) Ijjd1FN1JmvqGyiKH2YAcSLqemZcbf9NGWxP0kanwwKpPMPBCp

Access Level Read and write (modify app permissions)

Owner nimaclara

Owner ID 84304937

## Application Actions

Regenerate Consumer Key and Secret    Change App Permissions

- Under "permissions" tab, authorize you application. By default, it is “read and write”. Select “Read, write and Access direct messages”. And then click update settings

# TwitterMiningApp

Details    Settings    Keys and Access Tokens    Permissions    Test OAuth

## Access

What type of access does your application need?

Read more about our [Application Permission Model](#)

Read only  
 Read and Write  
 Read, Write and Access direct messages

**Note:**  
Changes to the application permission model will only reflect in access tokens obtained after the permission model change is saved. You will need to re-negotiate existing access tokens to alter the permission level associated with each of your application's users.

Update Settings

## R Script

```

#twitter mining
install.packages("RCurl")
install.packages("twitteR")
install.packages("ROAuth")
install.packages("base64enc")

library(RCurl)
library(twitteR)
library(ROAuth)
library(base64enc)

# make sure to remove the blank spaces (if any) both at the end of your
# apiKey or apiSecret between the quotations

apiKey <- "T3OfRIpmWVgiPVu1LARuQVEqf"
apiSecret <- "Ijjd1FN1JmvqGyiKH2YAcSLqemZcbf9NGWxP0kanwwKpPMPBCp"
accessToken <- "84304937-Vi1CYqNRbjjsoS7ShrjSY7SZHXfmtIqvS4qdXT3w6"
accessSecret <- "HyAdI4yWEYe00Jk1xdjzCEcxIzCNrH1vmdnvkTuL1tzjP";

setup_twitter_oauth(apiKey,apiSecret,access_token=accessToken,access_secret=accessSecret)

#install.packages("tm")
#install.packages("wordcloud")
library(tm)
library(wordcloud)
library(RColorBrewer)
hillaryClinton <- searchTwitter("#hillaryclinton",n = 500,lang="en",resultType="recent")
hillaryClinton_text = sapply(hillaryClinton, function(x) x$getText())
#usableText=str_replace_all(tweets$text,"[^[:graph:]]", " ")
hillaryClinton_corpus = Corpus(VectorSource(hillaryClinton_text))

#tdm = TermDocumentMatrix(
#  hillaryClinton_corpus,
#  control = list(
#    removePunctuation = TRUE,
#    stopwords = c("hillaryclinton", "HillaryClinton", "hillaryClinton", stopwords("english")),
#    removeNumbers = TRUE, tolower = TRUE)
#)

hillary_clean <- tm_map(hillaryClinton_corpus,removePunctuation)
hillary_clean <- tm_map(hillary_clean,content_transformer(tolower))
hillary_clean <- tm_map(hillary_clean, removeWords, stopwords("english"))
hillary_clean <- tm_map(hillary_clean,removeNumbers)
hillary_clean <- tm_map(hillary_clean,stripWhitespace)
hillary_clean <- tm_map(hillary_clean, removeWords, c("hillaryclinton", "HillaryClinton",
"hillaryClinton"))

tdm <- TermDocumentMatrix(hillary_clean)
m = as.matrix(tdm)

# get word counts in decreasing order

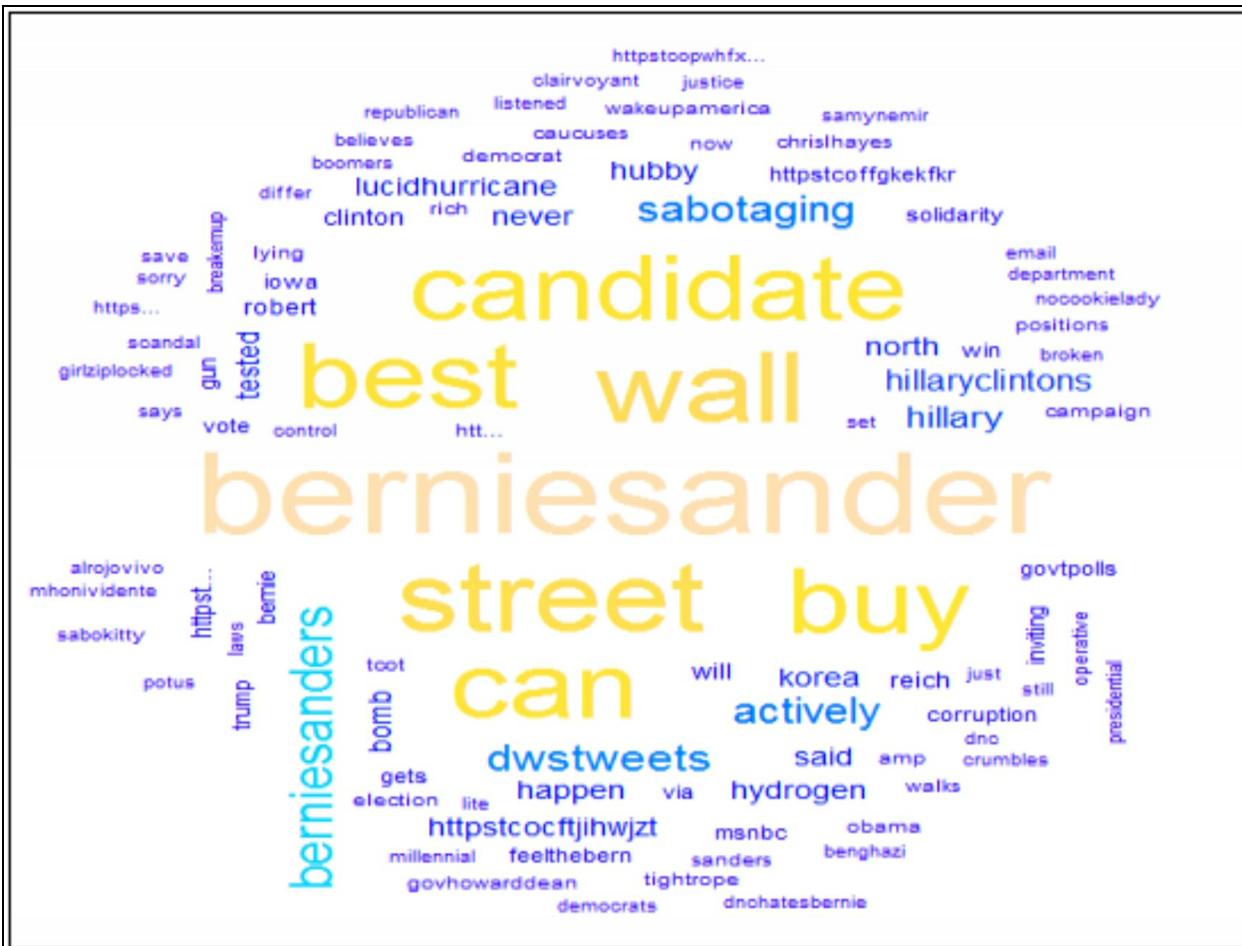
```

```

word_freqs = sort(rowSums(m), decreasing = TRUE)

# create a data frame with words and their frequencies
dm = data.frame(word = names(word_freqs), freq = word_freqs)
dm <- dm[1:100,]
wordcloud(dm$word, dm$freq, random.order = FALSE, colors = topo.colors(100))

```

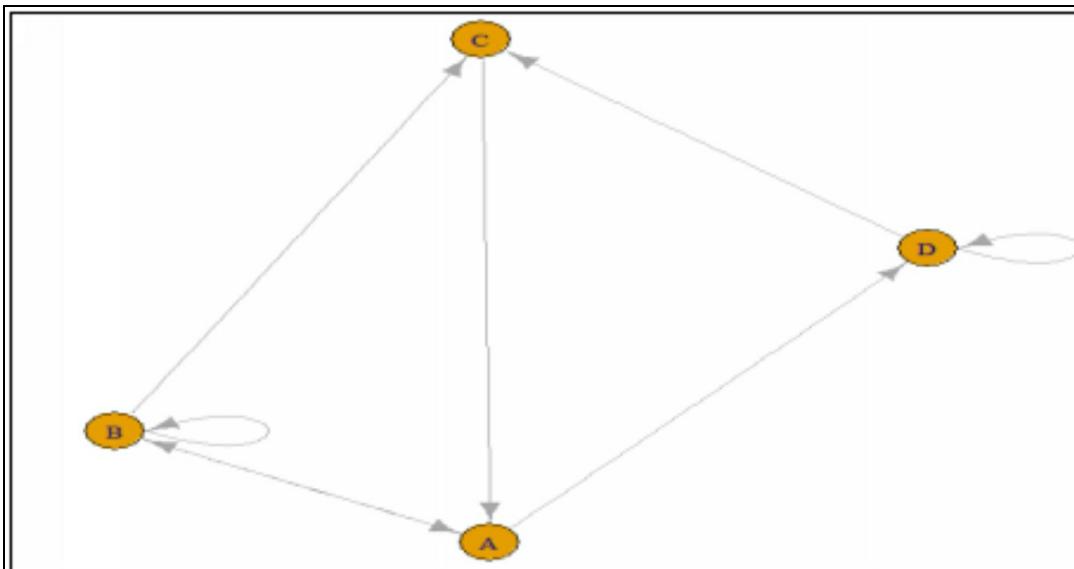


## Page Rank

```

#install.packages("igraph")
library(igraph)
myGraph <- graph(c("A", "B", "A", "D", "B", "A", "B", "B", "B", "C", "C", "A", "D", "C",
"D", "D"), directed = TRUE)
plot(myGraph)

```



page.rank(myGraph)\$vector

```
> page.rank(myGraph)$vector
      A          B          D          C
0.2851196 0.2214081 0.2759579 0.2175144
> |
```

## Additional Documentation

<https://www.r-project.org>

<http://www.rdatamining.com/docs/introduction-to-data-mining-with-r>

<https://www.rstudio.com>

<https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>



## Appendix P: Python Tutorial for Data Mining



# Tutorial for Data Mining

Version 1; January 2017

Keith S. Safford

for

Dr Anil K. Maheshwari

### 1 About this Tutorial

This tutorial was created for Dr. Anil K. Maheshwari to complement the R Tutorial in his book “Data Analytics Made Accessible.” One of the differences between the Python and the R Tutorial is there is a considerable amount of code so all code will not be printed in the tutorial. I will put in the tutorial the first major code segment Data Visualization and explain what needs to be changed in subsequent code segments to make the code work. All the rest of the code will be available for download so that you can execute it, modify it, and try different things as I feel that is the best way to learn.

The tutorial will start out by showing you how to download and install Anaconda 2 which utilizes Python 2.7.12 and numerous scientific and statistical libraries. The Anaconda environment is the ideal

environment to do Data Analytics/Data Science/Machine Learning with Python due to it installing the majority of necessary libraries to do what has to be done. You may ask “why didn’t I use Anaconda 3 which uses Python 3.5.2?” Suffice it to say, Python 2.7.12 is a stable product as is all the libraries utilized with it in Anaconda 2. At some point in the future, all the libraries will be converted and re-written to use Python 3.x, but at this point in time, many organizations and businesses are utilizing Python 2.7.x

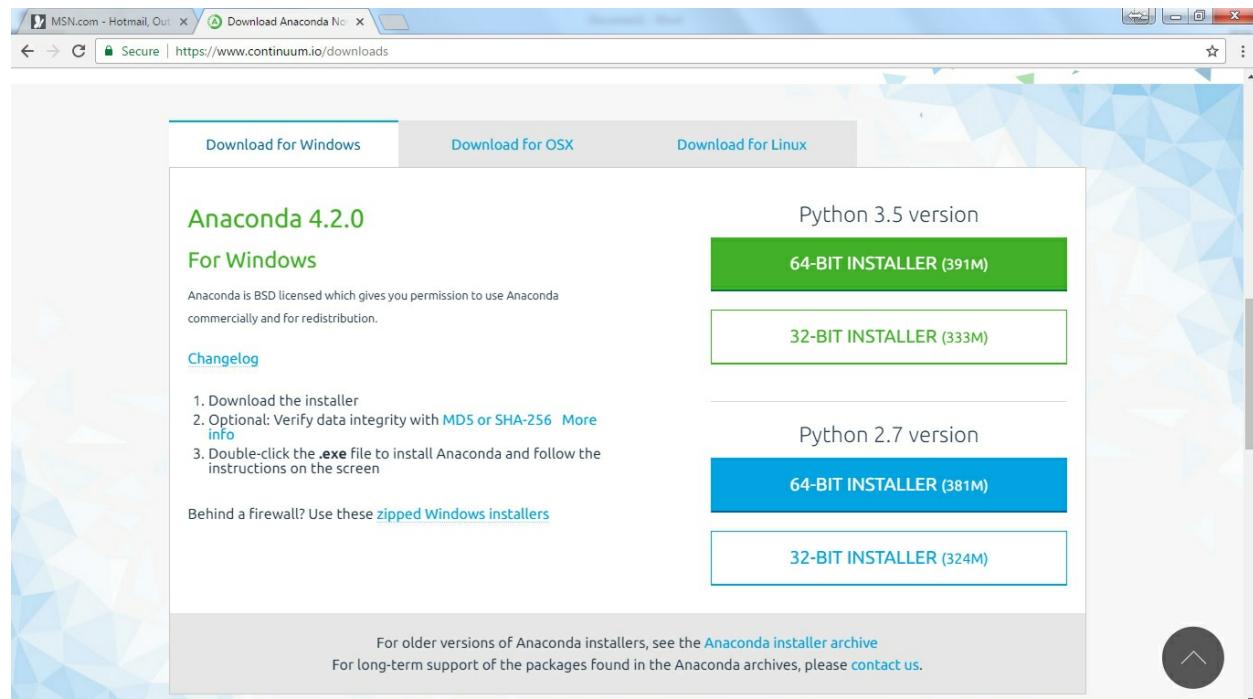
## 2 Getting Started

As is the case with R, Python is a free and open source language. “Python is a programming language that lets you work quickly and integrate systems more effectively.” I will start off by saying R is not better than Python and Python is not better than R. They both are very powerful and useful languages to accomplish the task that they were meant to do. In my research for completing this tutorial and learning Python and R, I came across many blogs where Data Analysts were using Python to munge and string process data to do analytics on in R. So it is my opinion that it is best to learn both languages and utilize their individual strengths to accomplish

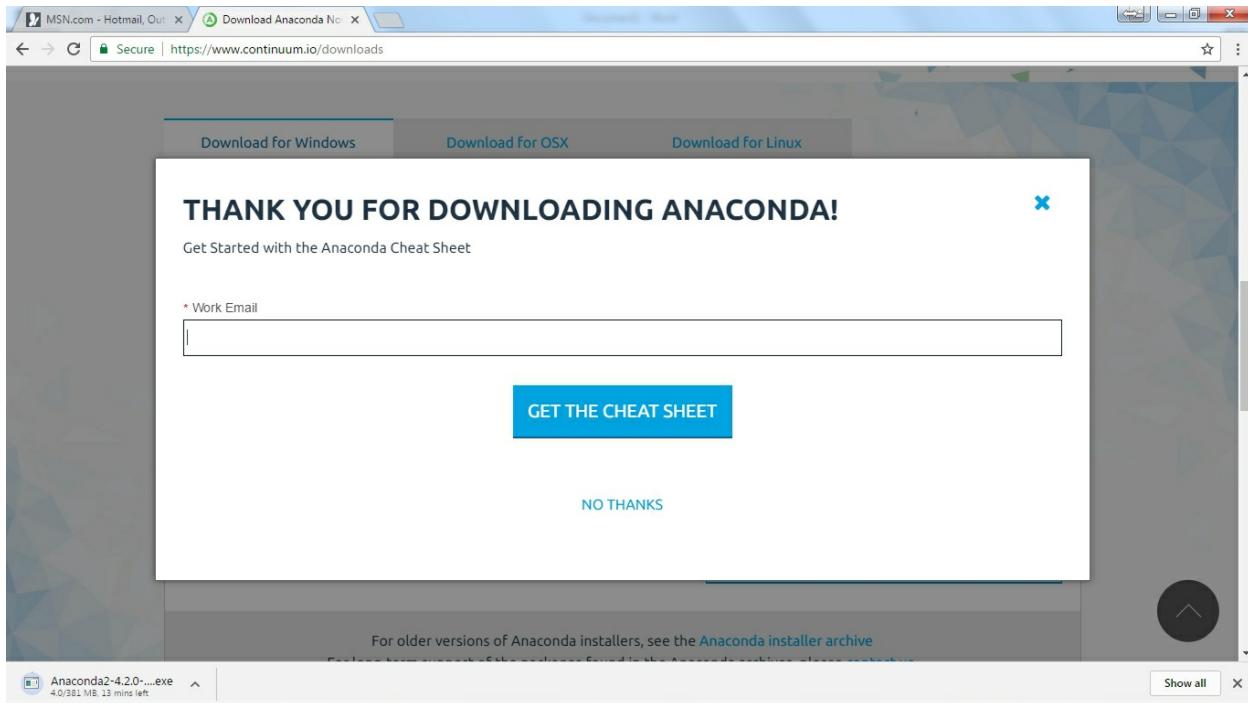
The first step to understand how Python works is to go to [https://www.continuum.io/downloads\\_and\\_download\\_Anaconda\\_2](https://www.continuum.io/downloads_and_download_Anaconda_2), install it, set-up some shortcuts on your desktop to Idle (a quick and dirty editor for Python), Spyder (a full blown IDE for developing Analytics code in Python), and the help file for Python 2.7.12 as there are a number of tutorials on Python that can help you “get your feet wet.”

## 3 Installation

Click on the official Anaconda site (<https://www.continuum.io/downloads>).

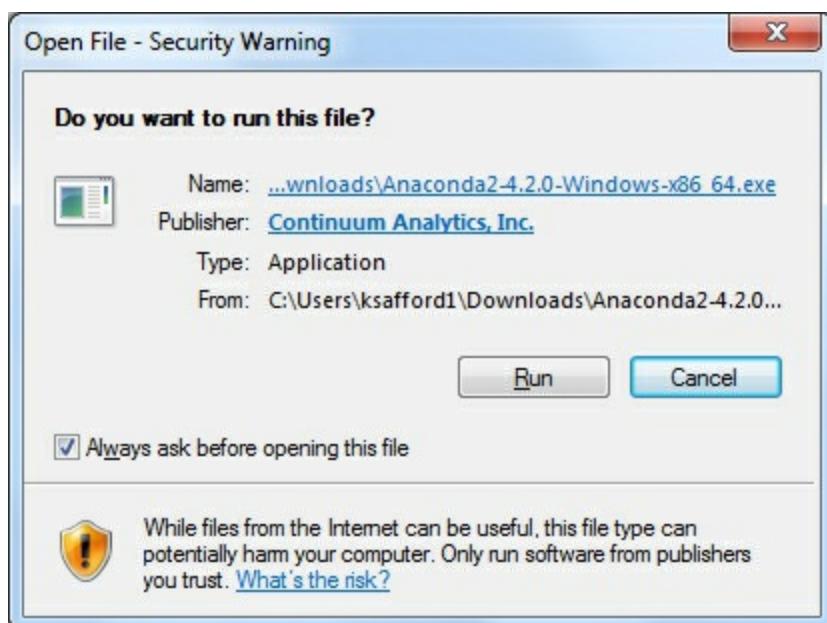
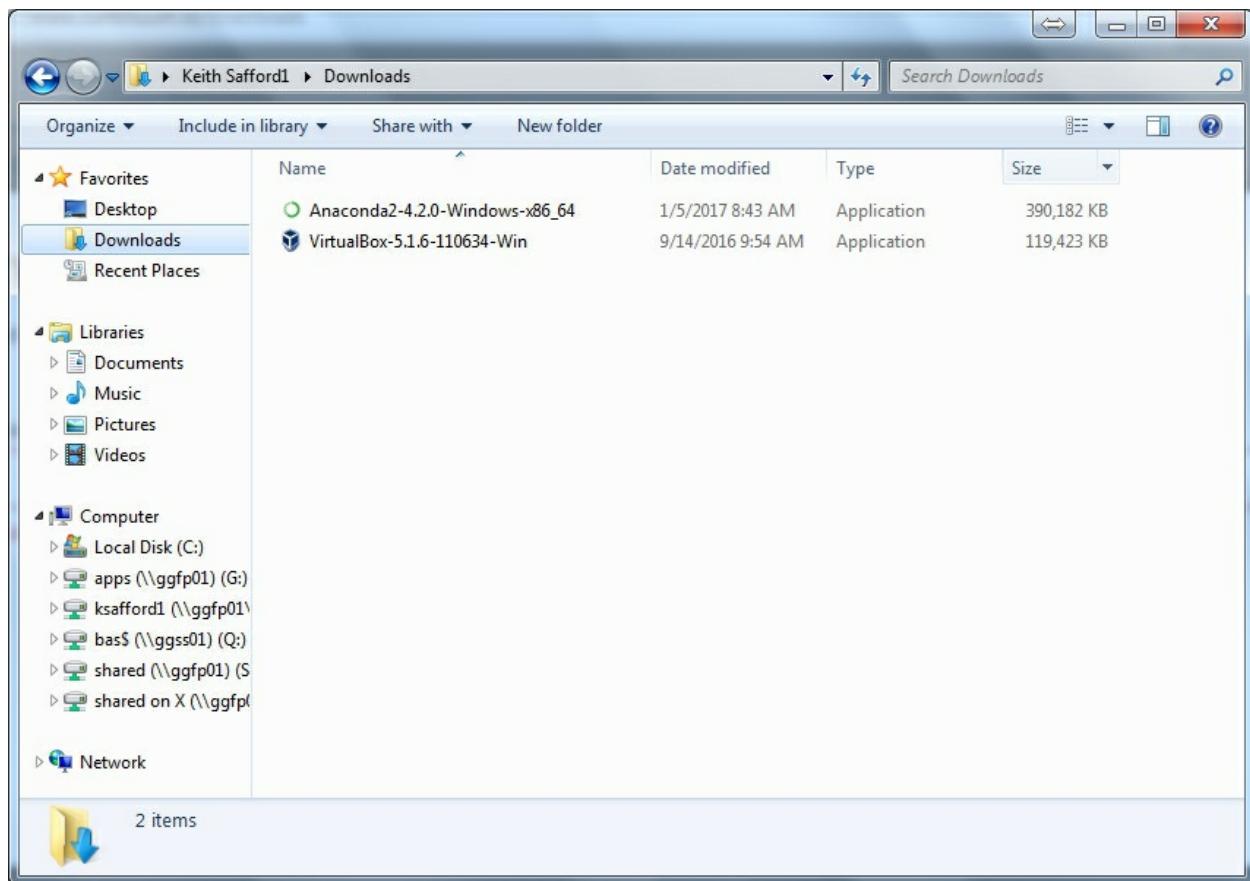


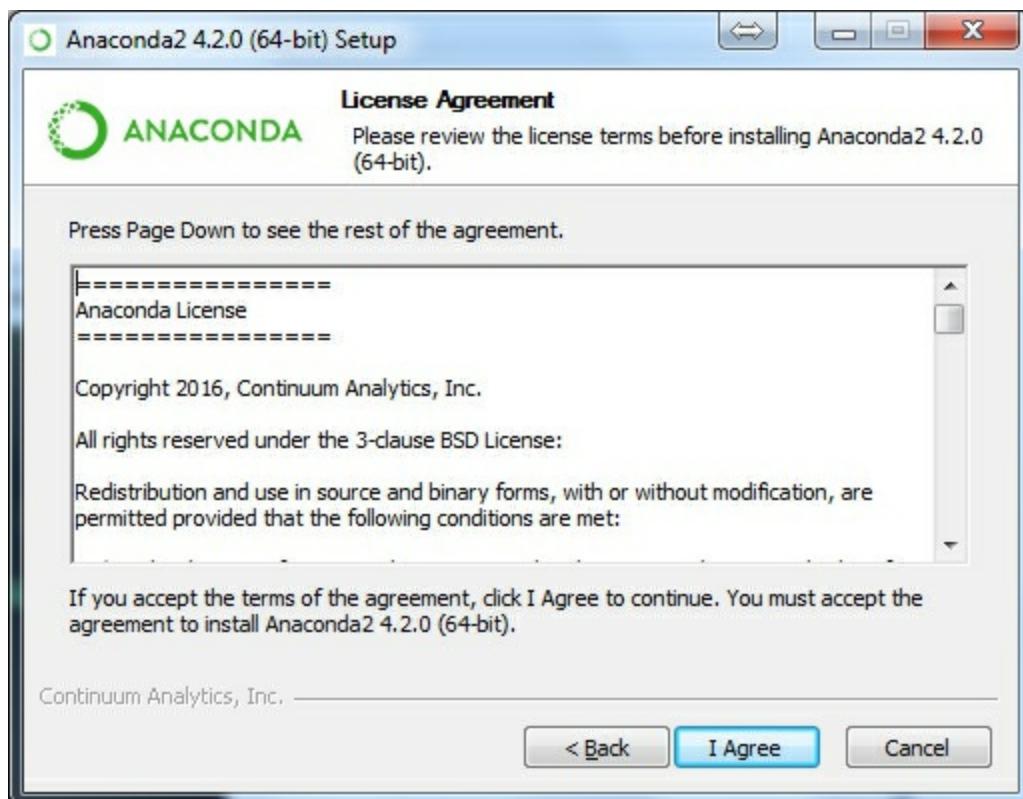
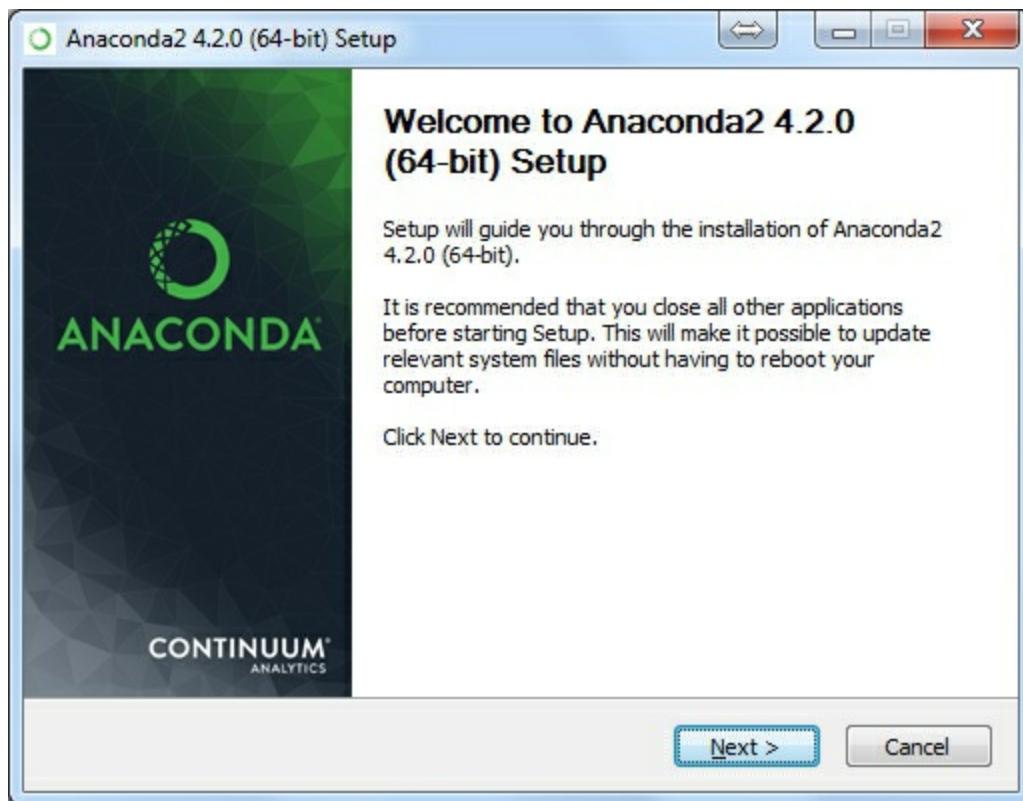
Click either the 64-BIT INSTALLER or the 32-BIT INSTALLER. Note: I installed and utilized the 64-BIT INSTALLER. Everything should work with the 32-BIT INSTALLER. HOWEVER, being a developer and knowing how “Murphy” is everywhere, there could be problems. I developed and tested everything utilizing the 64-bit version as most developers have 64-bit OS’s.

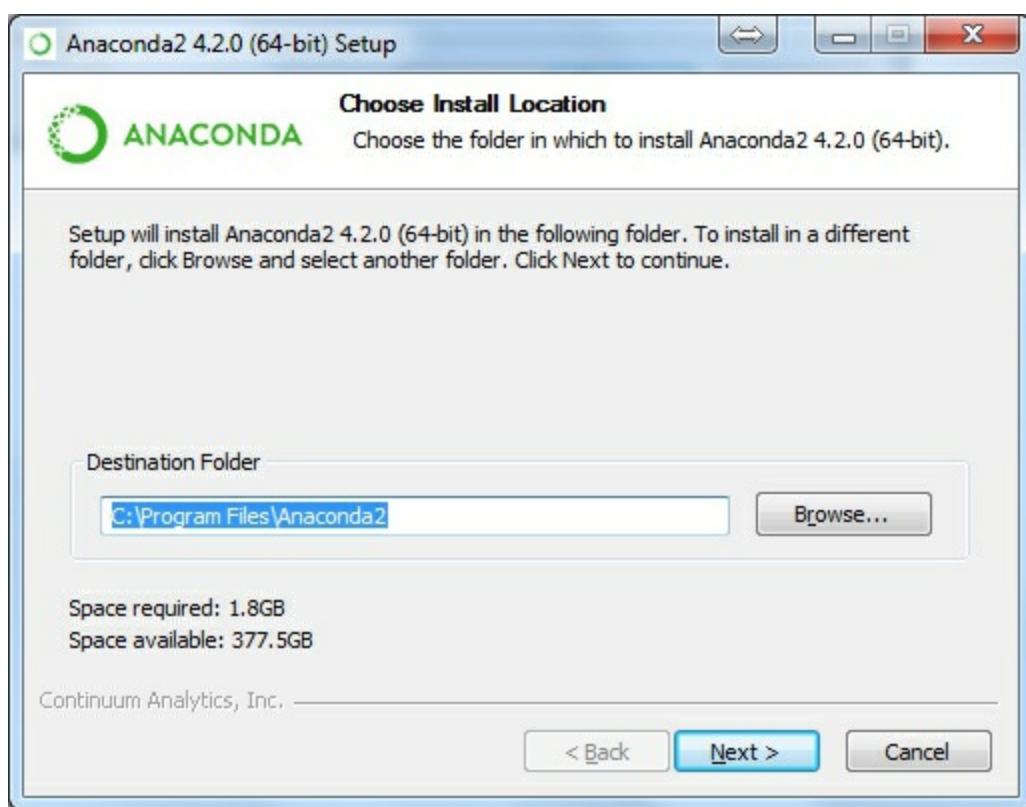
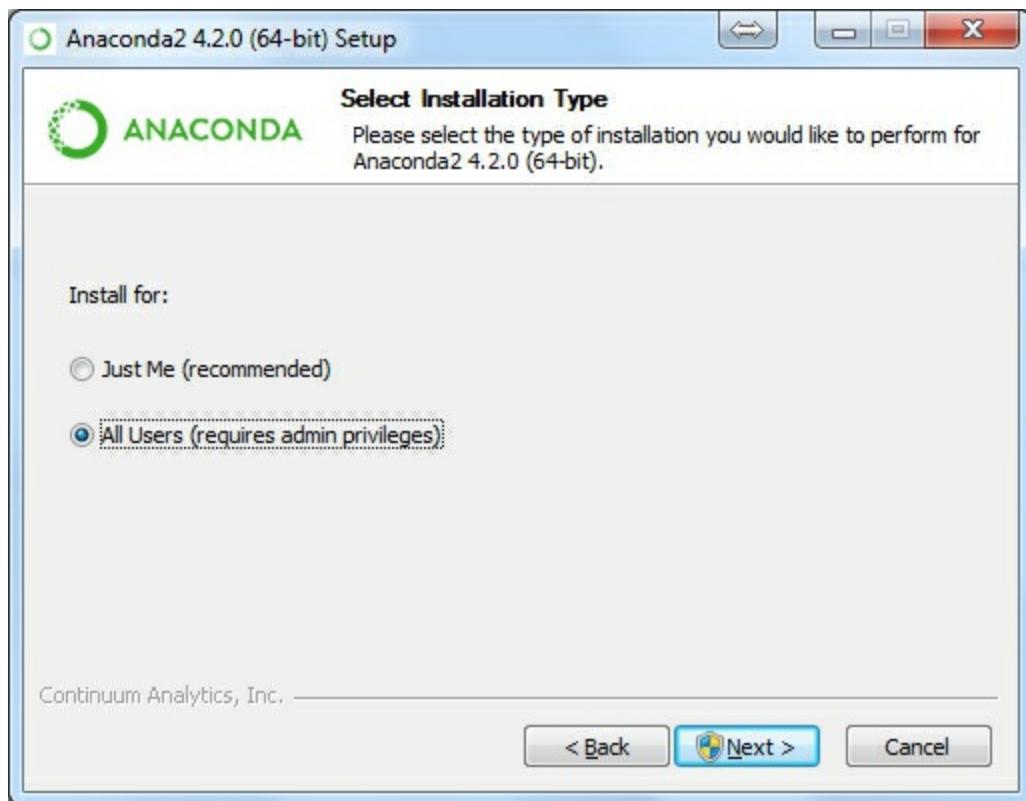


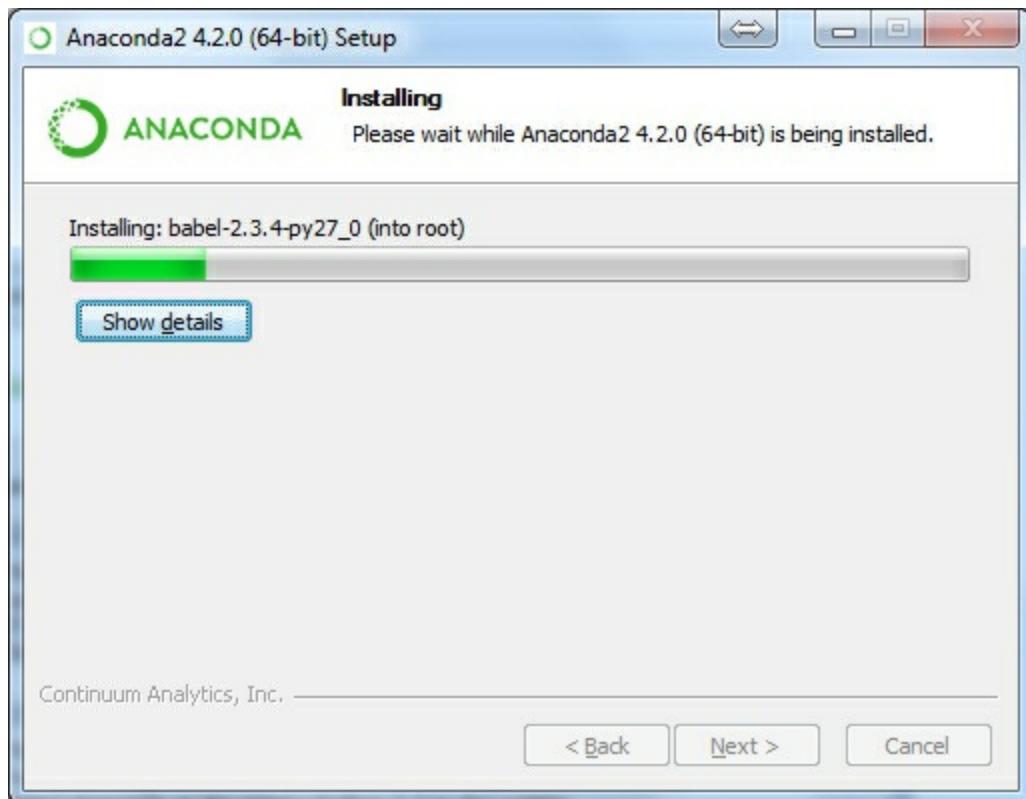
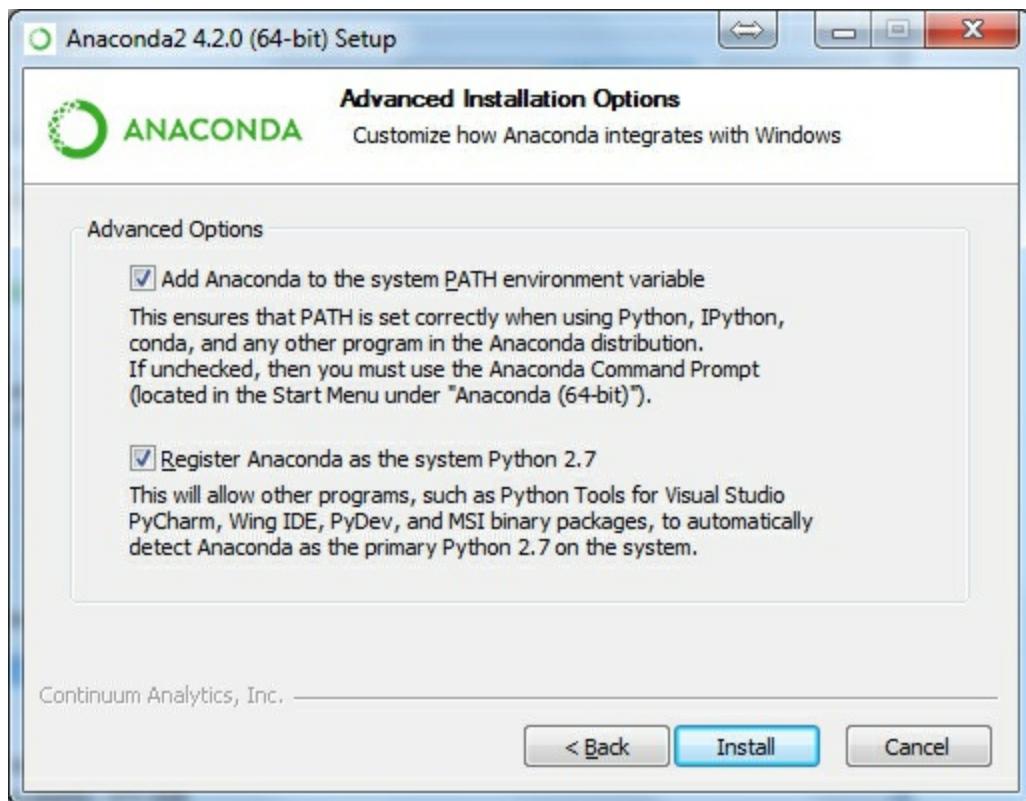
Go ahead and enter your email address as you will get a “cheat sheet” for Anaconda which is most useful. I have included numerous “cheat sheets” in the directory “PythonCheatSheets” in the directory “PythonTutorial.”

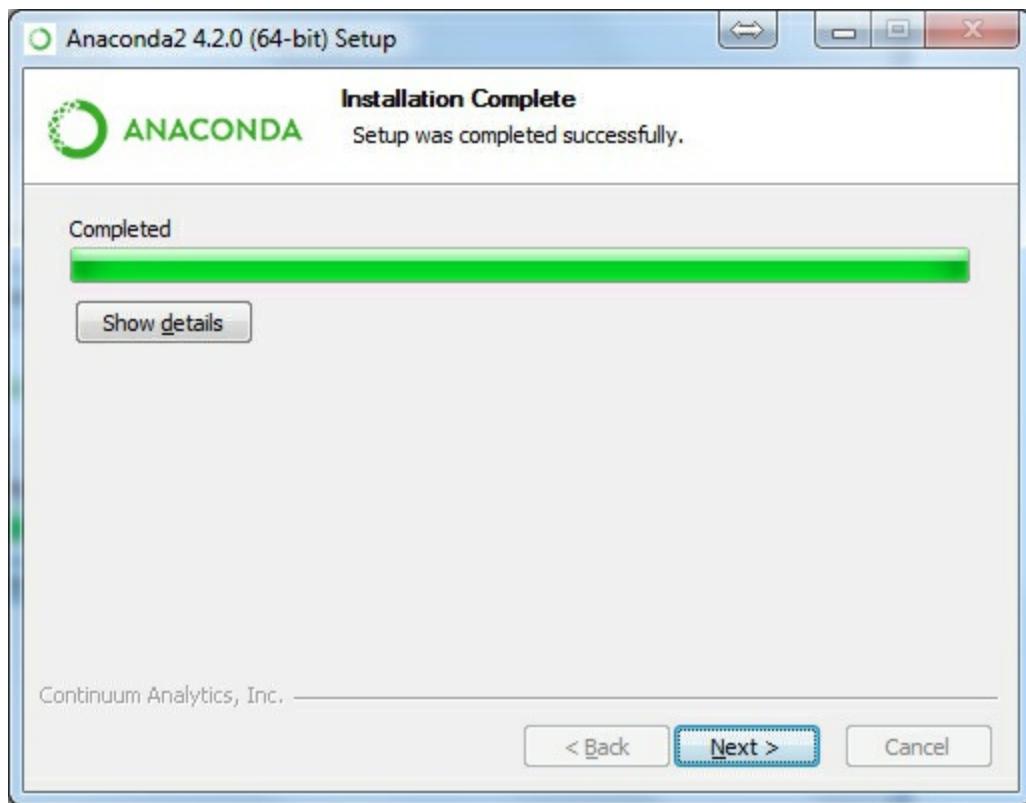
Double-click on “Anaconda2-4.2.0-Windows-x86\_64 to start the installation and follow the prompts.













## Why you'll love Anaconda Cloud

Go ahead and sign up for Anaconda Cloud if you want to or not – it is up to you!

Get into command prompt as Administrator and get to Anaconda2 directory as follows

A screenshot of a Windows Command Prompt window titled "Administrator: Command Prompt". The window shows the following text:

```
C:\Administrator: Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\windows\system32>cd\
C:\>cd program files\anaconda2
C:\Program Files\Anaconda2>_
```

The window has a blue title bar and a black background. It includes standard window controls like minimize, maximize, and close buttons.

Update Anaconda2 package as follows by typing “conda update anaconda” at the command prompt.

```
Administrator: Command Prompt
C:\windows\system32>cd\
C:\>cd program files\anaconda2
C:\Program Files\Anaconda2>conda update anaconda
```

```
Administrator: Command Prompt - conda update anaconda
The following packages will be downloaded:
package                                | build
-----|-----
conda-env-2.6.0                         | 0          498 B
anaconda-custom                          | py27_0    13 KB
requests-2.12.4                           | py27_0    755 KB
pyopenssl-16.2.0                          | py27_0    68 KB
conda-4.3.7                               | py27_0    522 KB
                                         | Total:   1.3 MB

The following NEW packages will be INSTALLED:
  conda-env: 2.6.0-0

The following packages will be UPDATED:
  anaconda:  4.2.0-np111py27_0 --> custom-py27_0
  conda:     4.2.9-py27_0      --> 4.3.7-py27_0
  pyopenssl: 16.0.0-py27_0      --> 16.2.0-py27_0
  requests:  2.11.1-py27_0     --> 2.12.4-py27_0

Proceed <[y]/n>?
```

Press “y”

The above packages will be updated. This is good to do as maybe some packages have been updated since the install was created for Anaconda 2.

Then proceed to update and install other necessary packages that will be needed/required to execute the tutorials.

Type the following commands in the command prompt:

conda update spyder

conda update pip

conda install nltk

For wordcloud – necessary for the WordCloud display

Anaconda2

<https://blogs.msdn.microsoft.com/pythonengineering/2016/04/11/unable-to-find-vcvarsall-bat/>

Download and run VCForPython27.msi

pip install wordcloud

pip install tweepy

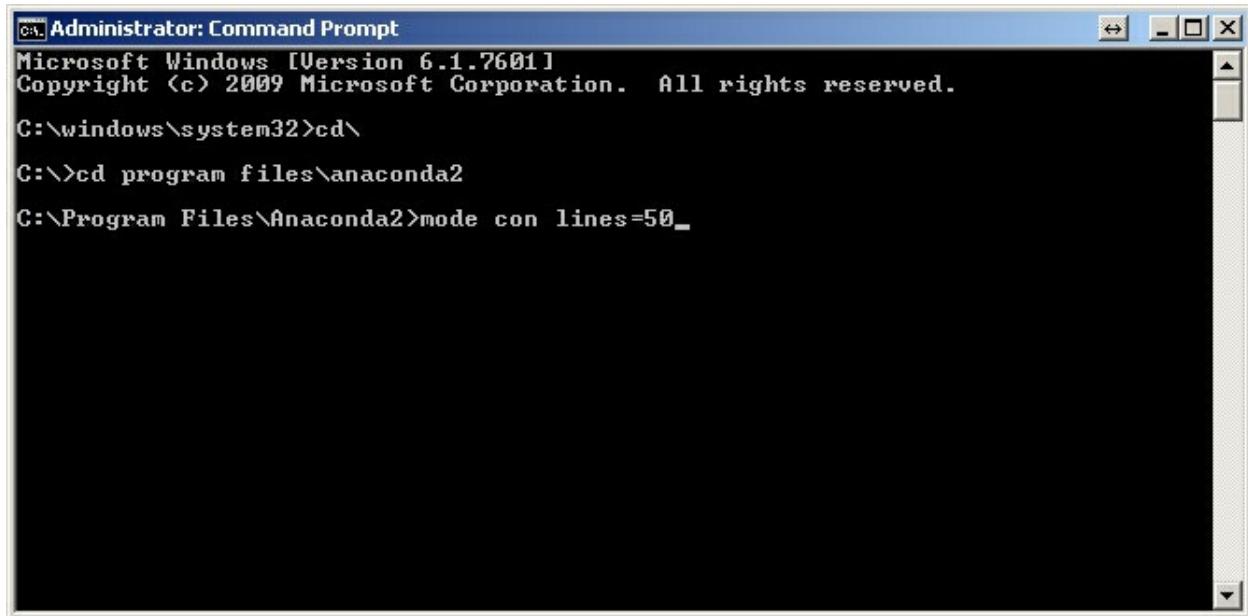
pip install pydotplus

For Graphviz – necessary for the Decision Tree display

<http://graphviz.org/Download..php>

Download and install the graphviz-2.38.exe and it will install to your C:\Program Files (x86) directory. THEN ADD THE PATH TO YOUR WINDOWS PATH STATEMENT!!! On my machine I added "C:\Program Files (x86)\Graphviz2.38\bin" to the path statement. NOTE: If you had Spyder running when you did this, you need to restart Spyder to be able to get the new path into the environment. If Spyder is running and you do NOT restart, you will get an error message "Graphviz executable not found". *It took me a while to get all this figured out!!!*

Make the command line window larger – mode con lines=50



```
Administrator: Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\windows\system32>cd\
C:\>cd program files\anaconda2
C:\Program Files\Anaconda2>mode con lines=50_
```

Then enter the command “conda list” as this will display ALL packages installed in the Anaconda environment and their version.

You can also do a conda list > e:\analytics\conda\_pkgs.txt which will create a text file that contains a list of all the packages installed in Anaconda. You can then print this out to keep track of what you have installed in your Anaconda environment and see what version you have installed.

```
Administrator: Command Prompt
python-dateutil           2.5.3          py27_0
pytz                      2016.6.1       py27_0
pywin32                   220            py27_1
pyyaml                    3.12           py27_0
pyzmq                     15.4.0          py27_0
qt                        5.6.0           vc9_0 [vc9]
qtawesome                 0.4.1           py27_0
qtconsole                 4.2.1           py27_2
qtpy                      1.1.2           py27_0
requests                  2.12.4          py27_0
requests-oauthlib         0.7.0           <pip>
rope                      0.9.4           py27_1
ruamel_yaml                0.11.14        py27_0
scikit-image               0.12.3           np111py27_1
scikit-learn               0.17.1           np111py27_1
scipy                     0.18.1           np111py27_0
setuptools                27.2.0          py27_1
simplegeneric              0.8.1           py27_1
singledispatch             3.4.0.3        py27_0
sip                        4.18            py27_0
six                       1.10.0          py27_0
snowballstemmer            1.2.1           py27_0
sockjs-tornado             1.0.3           py27_0
sphinx                    1.4.6           py27_0
spyder                    3.1.0           py27_0
sqlalchemy                 1.0.13          py27_0
ssl_match_hostname         3.4.0.2          py27_1
statsmodels                0.6.1           np111py27_1
sympy                     1.0             py27_0
tk                         8.5.18           vc9_0 [vc9]
toolz                     0.8.0           py27_0
tornado                   4.4.1           py27_0
traitlets                 4.3.0           py27_0
tweepy                     3.5.0           <pip>
unicodecsv                 0.14.1          py27_0
vs2008_runtime             9.00.30729.1      2
wcwidth                    0.1.7           py27_0
werkzeug                  0.11.11          py27_0
wheel                     0.29.0          py27_0
widgetsnbextension          1.2.6           py27_0
win_unicode_console        0.5             py27_0
wordcloud                  1.2.1           <pip>
wrapt                     1.10.6          py27_0
xlrd                      1.0.0           py27_0
xlsxwriter                 0.9.3           py27_0
xlwings                   0.10.0          py27_0
xlwt                      1.1.2           py27_0
zlib                      1.2.8           vc9_3 [vc9]

C:\Program Files\Anaconda2>
```

It is a good idea to do a “conda update conda” frequently, I do it every day, as it ensures all your packages necessary for Anaconda are updated and current.

## 4 Working on Python

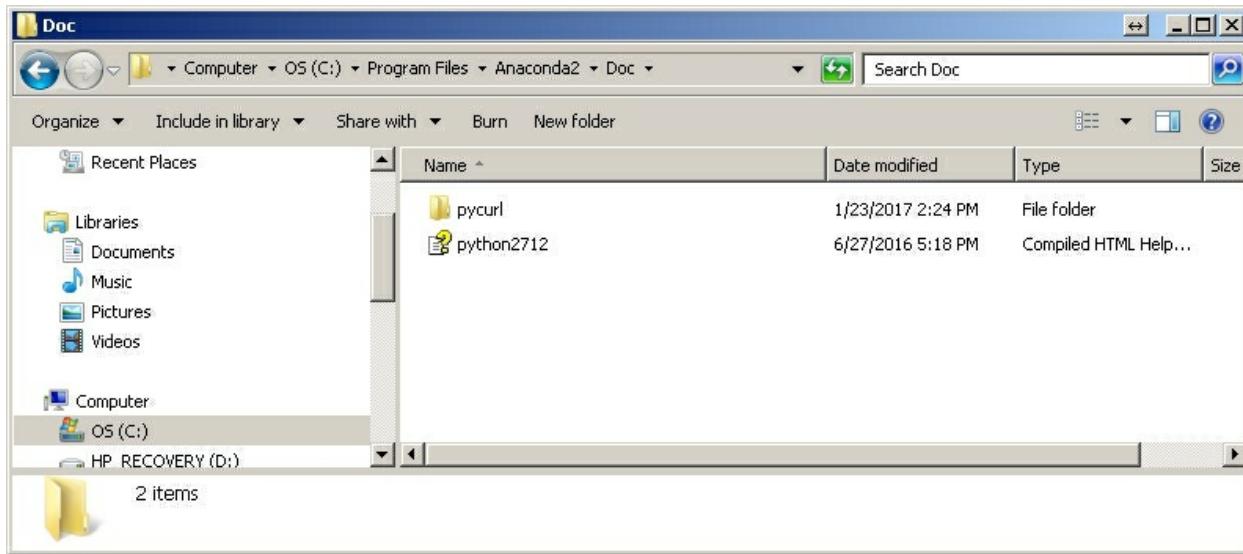
Note: this is based on Windows 7, Windows 10 is different.

## 4.1 Windows 7

Select the Anaconda application from your start menu, right click your mouse for Spyder and select “Send to” and then “Desktop”. That will create a short-cut to Spyder, the IDE for Analytics Development.

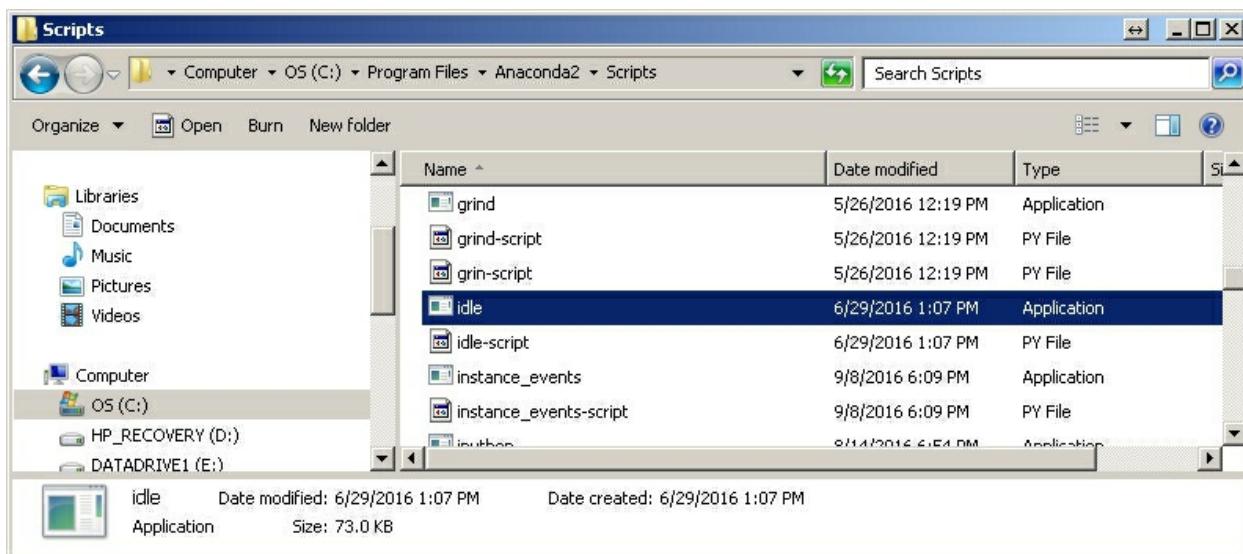


Go to Windows Explorer and select directory “C:\Program Files\Anaconda2\Doc”



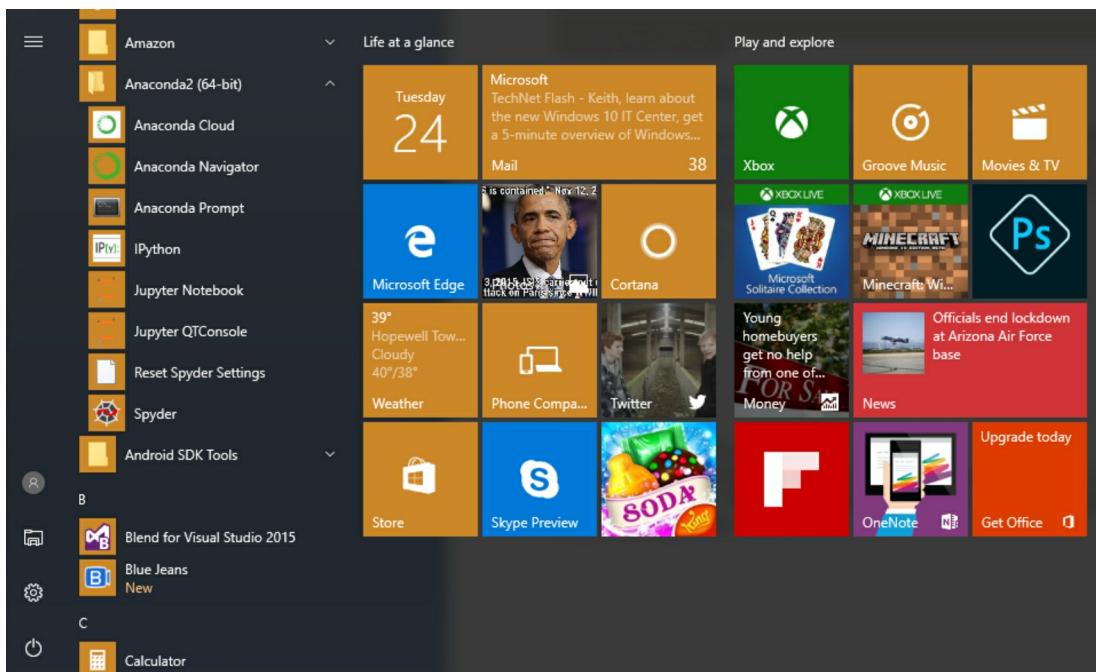
Right click your mouse on Python712.chm file “Send to” and then “Desktop”. That will create a short-cut to the Python 2.7.12 help file. You want this ready and easily accessible because you WILL have lots of questions when learning Python!

Go to Windows Explorer and select directory “C:\Program Files\Anaconda2\Scripts”

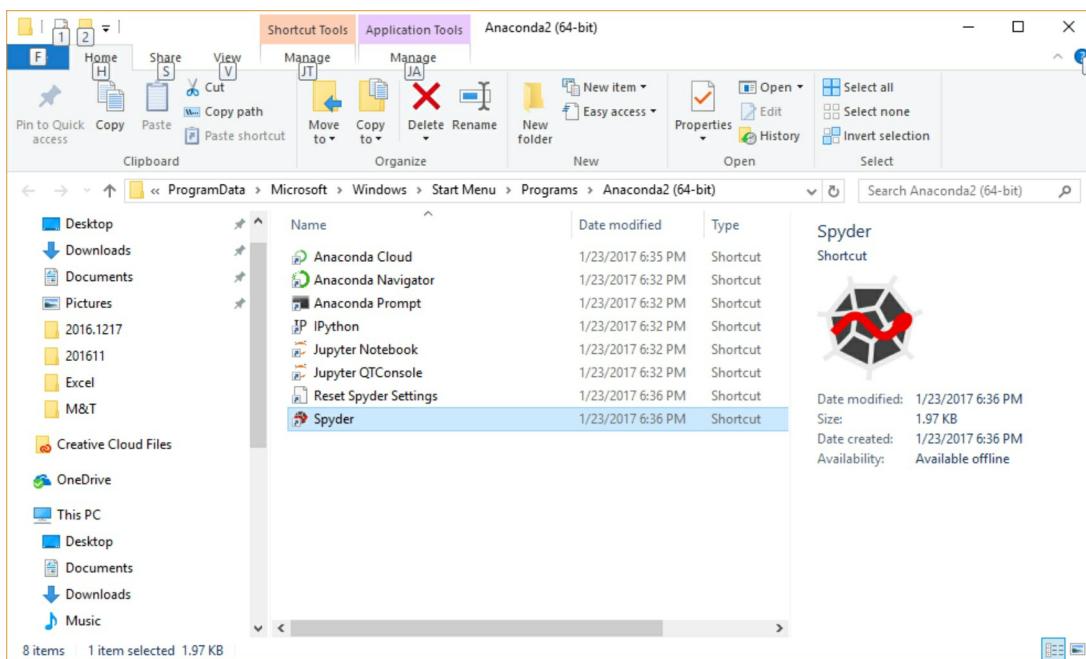


Right click your mouse on Idle file “Send to” and then “Desktop”. That will create a short-cut to the Python Idle editor which is an ideal editor to go through the tutorials in the Python 2.7.12 help file. Idle is an excellent editor to “get your feet wet” in Python and it is easy to use.

## 4.2 Windows 10

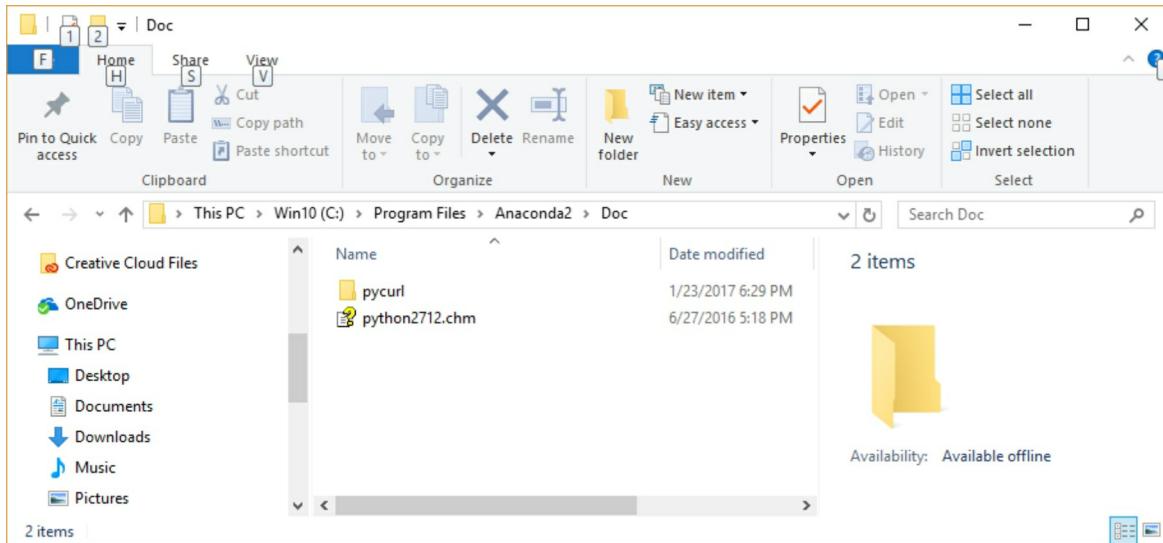


Right click on Spyder -> select “More” -> select “Open File Location”



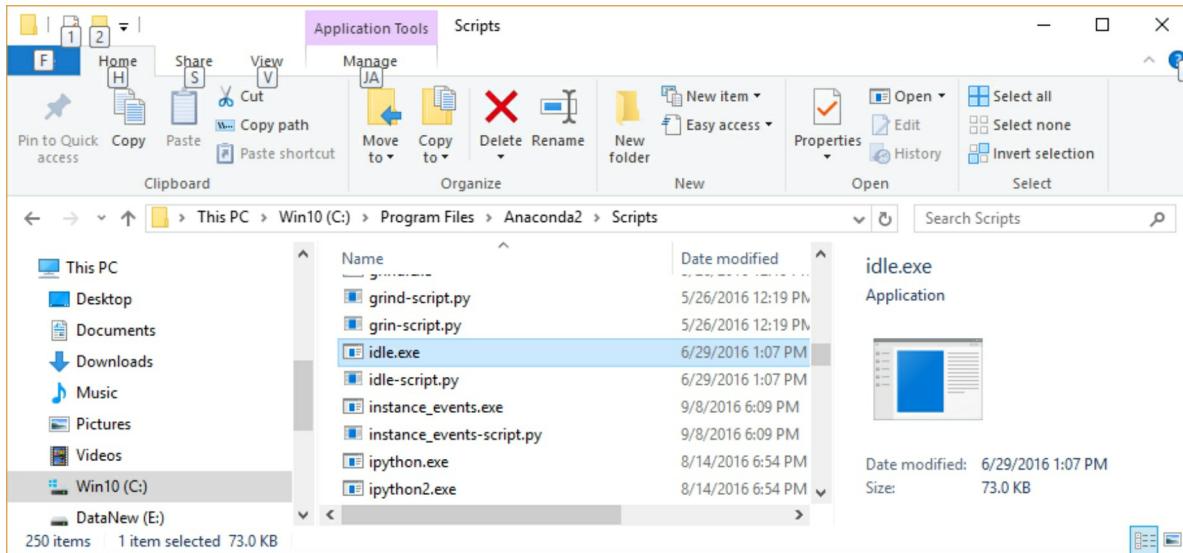
Right click on “Spyder” -> select “Send to” -> select “Desktop (create shortcut)”

Go to Windows Explorer and select directory “C:\Program Files\Anaconda2\Doc”



Right click your mouse on Python712.chm file “Send to” and then “Desktop”. That will create a short-cut to the Python 2.7.12 help file. You want this ready and easily accessible because you WILL have lots of questions when learning Python!

Go to Windows Explorer and select directory “C:\Program Files\Anaconda2\Scripts”



Right click your mouse on Idle file “Send to” and then “Desktop”. That will create a short-cut to the Python Idle editor which is an ideal editor to go through the tutorials in the Python 2.7.12 help file. Idle is an excellent editor to “get your feet wet” in Python and it is easy to use.

**NOTE: I do have a laptop with Windows 8.1 and I had problems installing Anaconda 2 on it so I am only using Anaconda with Windows 7 and Windows 10 machines.**

## 4.3 Python Help and Tutorial

Click on the help file on the desktop and select “The Python Tutorial” and you will get the following window:

The screenshot shows a Windows-style application window titled "Python 2.7.12 documentation". The menu bar includes "File", "Edit", "View", "Insert", "Format", "Tools", "Help", and "About". Below the menu is a toolbar with icons for Hide, Locate, Back, Forward, Home, Font, Print, and Options. The left sidebar contains a "Contents" tree view with the following structure:

- Python 2.7.12 documentation
- Python Module Index
- What's New in Python
- The Python Tutorial** (selected)
- Whetting Your Appetite
- Using the Python Interpreter
- An Informal Introduction to P.
- More Control Flow Tools
- Data Structures
- Modules
- Input and Output
- Errors and Exceptions
- Classes
- Brief Tour of the Standard Li
- Brief Tour of the Standard Li
- What Now?
- Interactive Input Editing and
- Floating Point Arithmetic: Iss
- Appendix
- Python Setup and Usage
- The Python Language Reference
- The Python Standard Library
- Extending and Embedding the P
- Python/C API Reference Manual
- Distributing Python Modules
- Installing Python Modules
- Python HOWTOs
- Python Frequently Asked Questi
- Glossary
- About these documents
- Reporting Bugs
- Copyright
- History and License

The main content area displays the "The Python Tutorial" page with the following text:

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, <https://www.python.org/>, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

This tutorial introduces the reader informally to the basic concepts and features of the Python language and system. It helps to have a Python interpreter handy for hands-on experience, but all examples are self-contained, so the tutorial can be read off-line as well.

For a description of standard objects and modules, see [The Python Standard Library](#). The [Python Language Reference](#) gives a more formal definition of the language. To write extensions in C or C++, read [Extending and Embedding the Python Interpreter](#) and [Python/C API Reference Manual](#). There are also several books covering Python in depth.

This tutorial does not attempt to be comprehensive and cover every single feature, or even every

A quick and dirty example of using Python as a calculator:

Click on the Idle icon on the desktop and you will see as follows (it may take a little while to come up the first time):

The screenshot shows the Python 2.7.12 Shell window. The title bar reads "Python 2.7.12 Shell". The menu bar includes "File", "Edit", "Shell", "Debug", "Options", "Window", and "Help". The main window displays the Python welcome message:  
Python 2.7.12 |Anaconda custom (64-bit)| (default, Jun 29 2016, 11:07:13) [MSC v .1500 64 bit (AMD64)] on win32  
Type "copyright", "credits" or "license()" for more information.  
>>>

Type the following commands in to Idle and you are on your way to becoming a Pythonista!

The screenshot shows the Python 2.7.12 Shell window. The title bar reads "Python 2.7.12 Shell". The menu bar includes "File", "Edit", "Shell", "Debug", "Options", "Window", and "Help". The main window displays the following Python session:  
Python 2.7.12 |Anaconda custom (64-bit)| (default, Jun 29 2016, 11:07:13) [MSC v .1500 64 bit (AMD64)] on win32  
Type "copyright", "credits" or "license()" for more information.  
>>> 2 + 2  
4  
>>> 2\*\*2  
4  
>>> 2 / 2  
1  
>>> 2 % 2  
0  
>>>

Just type exit() to exit Idle.

**OBS:** Python is case sensitive, so **Variable** is different from **variable**.

As for Spyder, the IDE that is included with Anaconda, here is a url with some tutorials on how to get going with Spyder: <http://datasciencesource.com/python-with-spyder-tutorial/>. Spyder is a full blown development environment for doing Data Analytics utilizing Python and it will take a while to get familiar with but will be well worth your effort.

## 4.4 Import a Dataset in Python

Before importing your dataset you must convert your .xls or .xlsx file to .csv file and save it in the working directory. There are libraries for reading in .xls and/or .xlsx files, but for simplicity, I kept everything in .csv format.

Load the data to a variable using pd.read.csv command. For example, for the code below we used the following code:

```
import os.path
import pandas as pd
import sys

DEBUG = True
DRIVE = "E:\\"
DATA_DIR_PATH = "PythonTutorial\\ADataVisualization"
IN_FILE_NAME = "HeightWeight2016.csv"

def get_data(drive, data_dir_path, in_file):
    file_path = os.path.join(drive + data_dir_path, in_file)
    try:
        return pd.read_csv(file_path, delimiter=",", names=None)
    except IOError:
        msg = "An error occurred trying to read file: %s" % file_path
        print(msg)
        sys.exit()
```

*and it is called as follows:*

```
ht_wt_df = get_data(DRIVE + DATA_DIR_PATH, IN_FILE_NAME)
if DEBUG == True:
    print('*** ht_wt_df ***')
```

```
print(ht_wt_df.head(n = DISPLAY_RECORDS))
print(ht_wt_df.tail(n = DISPLAY_RECORDS))
```

Code displayed in IPython Console in Spyder as follows

```
*** ht_wt_df ***
   height  age  gender  weight
0    71.0   24      M   165.0
1    68.0   29      M   165.0
2    72.0   34      M   180.0
3    67.0   21      F   113.0
4    72.0   32      M   178.0
   height  age  gender  weight
175   65.0   31      M   120.0
176   70.0   33      M   178.2
177   68.0   34      M   140.0
178   71.0   26      M   130.0
179   69.0   27      M   170.0
```

A number of important items to note about the above code

1. import – imports a specific library which is important and has methods that are used throughout the code segment. It is always recommended to put these at the top of the module then you can see ALL imports that are used in the code.
2. Constants – these are variables that are constants used in the program. I put the DEBUG constant in there as I am developing so as things go, I can see what the results of my code are and I can turn this off and on as need be.
3. def defines a function in Python
4. Executing a function ht\_wt\_df = get\_data(DRIVE + DATA\_DIR\_PATH, IN\_FILE\_NAME) creates a DataFrame variable named ht\_wt\_df from the results of the get\_data function
5. if DEBUG == True: prints data that you want to see printed in the IPython Consold in Spyder

## 4.5 Data visualization –

see directory ADataVisualization and look at code modules DataVisualization.py and DataVisualizationOO.py. The first example used Python code and functions and the second example utilized a strictly object-oriented approach.

### 4.5.1 Ploting a Histogram

```
# function to create histogram
```

```
def histogram_it(df, color, title, xlabel, ylabel, img_name, orient=None):
```

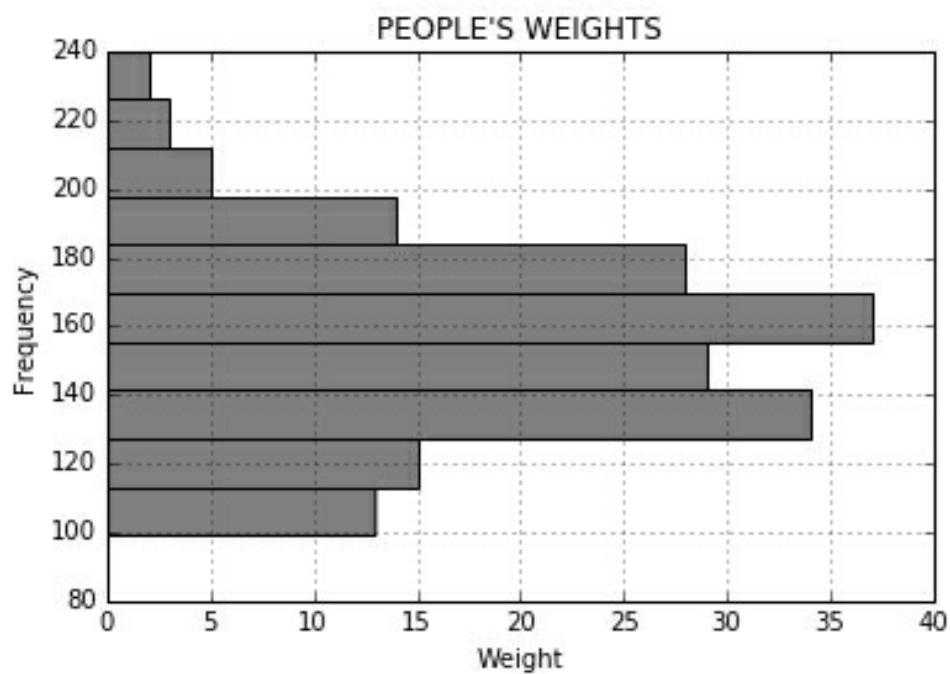
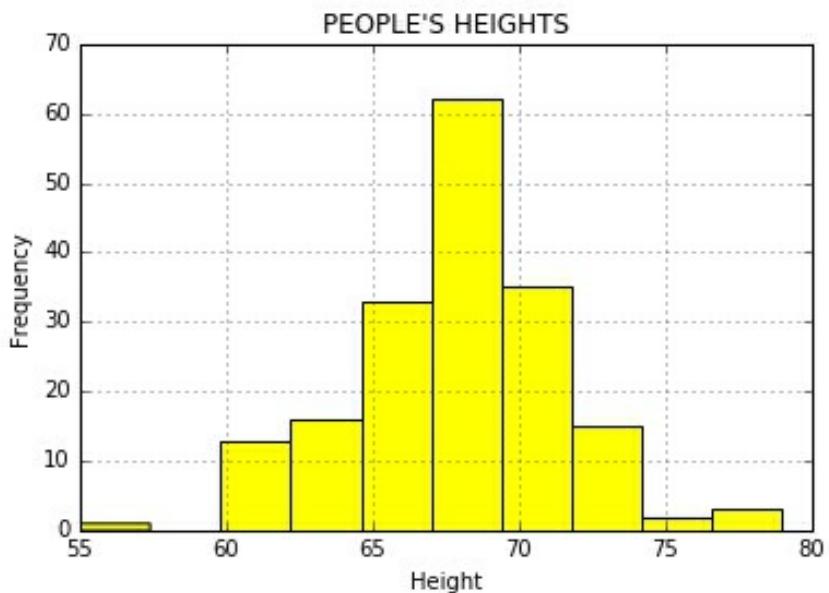
```

plt.clf()
plt.figure()
    if orient == None:
        df.hist(color=color)
    else:
        df.hist(color=color, orientation=orient)
plt.title(title)
plt.xlabel(xlabel)
plt.ylabel(ylabel)
write_image(DRIVE, DATA_DIR_PATH, img_name)
plt.show()

# Histogram of height – call the histogram_it function
histogram_it(ht_wt_df['height'],
             'yellow',
             "PEOPLE'S HEIGHTS",
             'Height',
             'Frequency',
             OUT_IMG_NAME + "_height_hist.png")

# Histogram of weight
histogram_it(ht_wt_df['weight'],
             'grey',
             "PEOPLE'S WEIGHTS",
             'Weight',
             'Frequency',
             OUT_IMG_NAME + "_weight_hist.png",
             'horizontal')

```

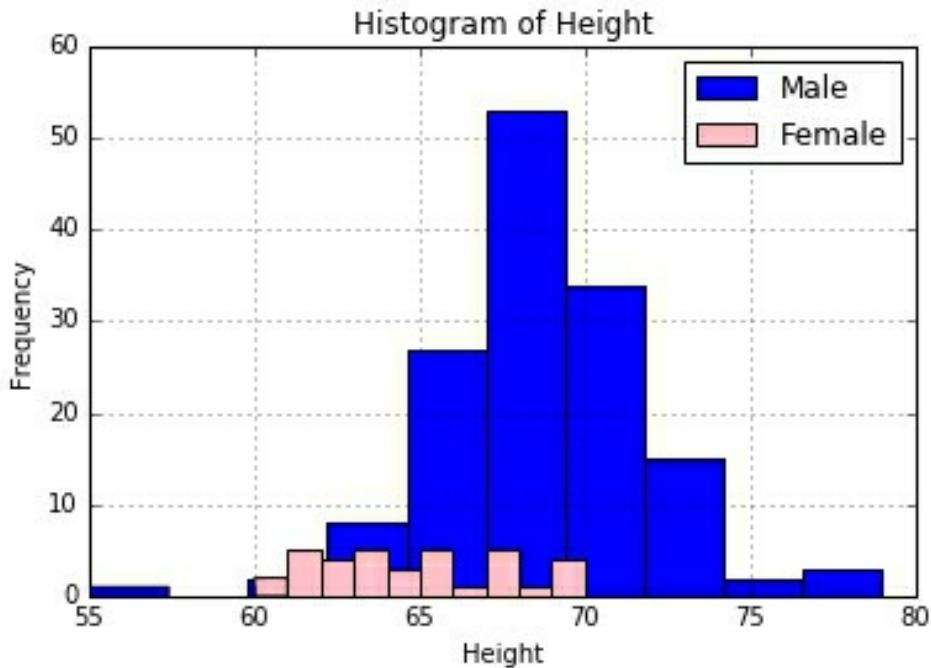


```
#Histogram of height by gender
gend_m_df = ht_wt_df.loc[(ht_wt_df['gender']=="M")]
gend_f_df = ht_wt_df.loc[(ht_wt_df['gender']=="F")]
if DEBUG == True:
    print(gend_f_df)
    print(gend_m_df)
plt.clf()
```

```

plt.figure()
gend_m_df['height'].hist(color='blue')
gend_f_df['height'].hist(color='pink')
plt.title("Histogram of Height")
plt.xlabel("Height")
plt.ylabel("Frequency")
plt.legend(['Male', 'Female'], loc=0)
out_dir_file_image = OUT_IMG_NAME + "_ht_gend_hist.png"
write_image(DRIVE, DATA_DIR_PATH, out_dir_file_image)
plt.show()

```

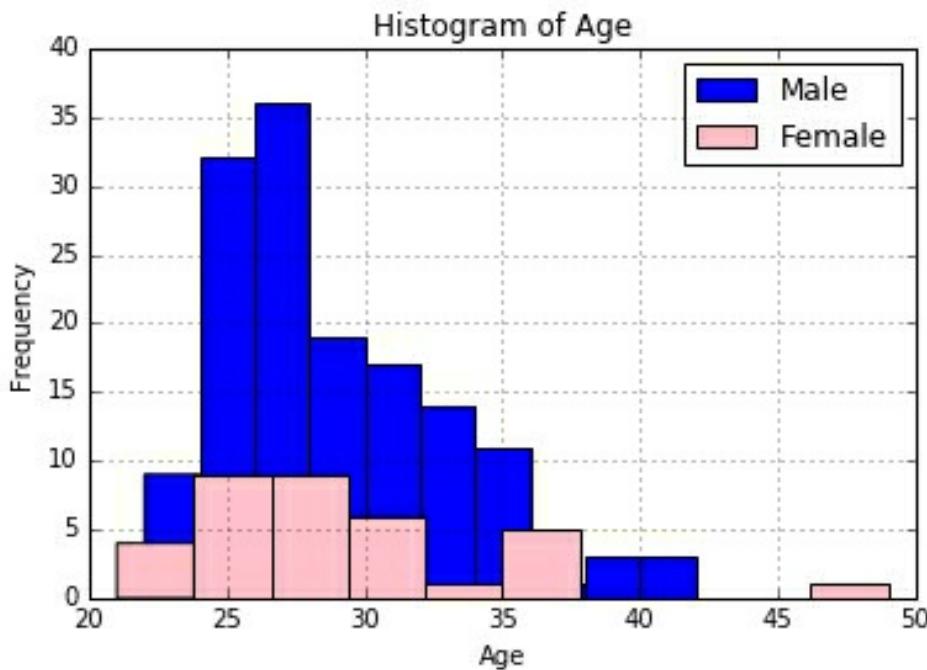


```

# Histogram of age by gender
plt.clf()
plt.figure()
gend_m_df['age'].hist(color='blue')
gend_f_df['age'].hist(color='pink')
plt.title("Histogram of Age")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.legend(['Male', 'Female'], loc=0)
out_dir_file_image = OUT_IMG_NAME + "_age_gend_hist.png"
write_image(DRIVE, DATA_DIR_PATH, out_dir_file_image)

```

```
plt.show()
```



#### 4.5.2 Plotting a Bar Chart

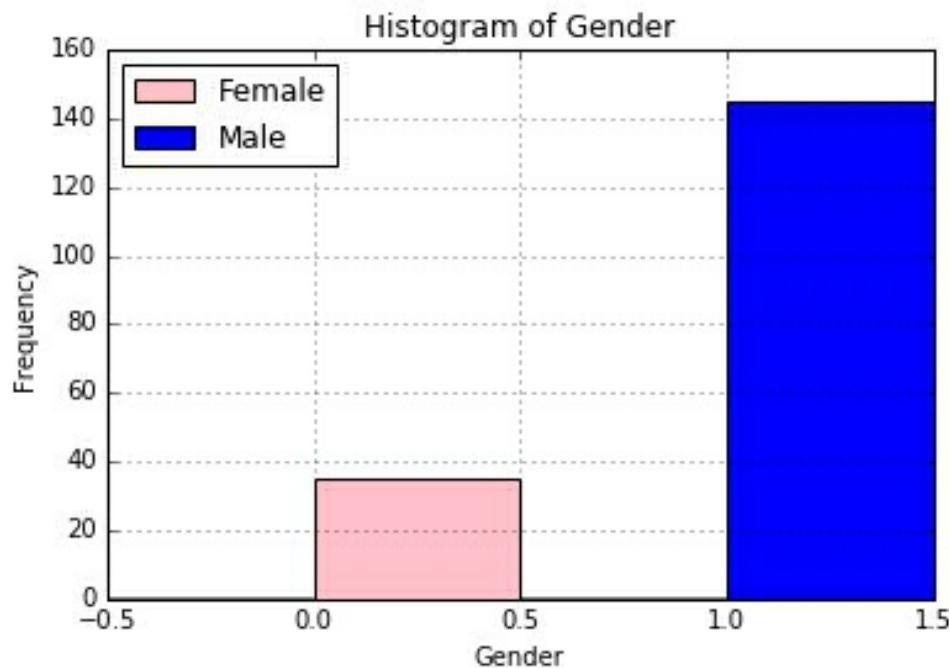
```
# Histogram of gender  
# Convert 'F' and 'M' to 0 and 1 to create histogram  
# (data has to be numeric so it is essential to convert F and M to 0 and 1)  
female = []  
for row in ht_wt_df['gender']:  
    if row == 'F':  
        female.append(0)  
    else:  
        female.append(1)  
  
    # Append female array to ht_wt_df dataframe  
ht_wt_df['female'] = female  
  
    # Repurpose gend_ dataframes for this histogram  
gend_m_df = ht_wt_df.loc[(ht_wt_df['female']==1)]  
gend_f_df = ht_wt_df.loc[(ht_wt_df['female']==0)]  
  
if DEBUG == True:
```

```

print(gend_f_df)
print(gend_m_df)

plt.clf()
plt.figure()
gend_f_df['female'].hist(color='pink', bins=2)
gend_m_df['female'].hist(color='blue', bins=2)
plt.title("Histogram of Gender")
plt.xlabel('Gender')
plt.ylabel('Frequency')
plt.legend(('Female', 'Male'), loc=0)
out_dir_file_image = OUT_IMG_NAME + "_gend.png"
write_image(DRIVE, DATA_DIR_PATH, out_dir_file_image)
plt.show()

```



#### 4.5.3 Ploting charts side by side

# Multiple Plots

```

# Set up for multiple plots 2 rows and 2 columns
fig, axes = plt.subplots(nrows=2, ncols=2)
ax0, ax1, ax2, ax3 = axes.flat

```

```

# Histogram of height
ax0.hist(ht_wt_df['height'], histtype='bar', color='yellow')
ax0.set_title("PEOPLE'S HEIGHTS")
ax0.set_xlabel('Height')
ax0.set_ylabel('Frequency')

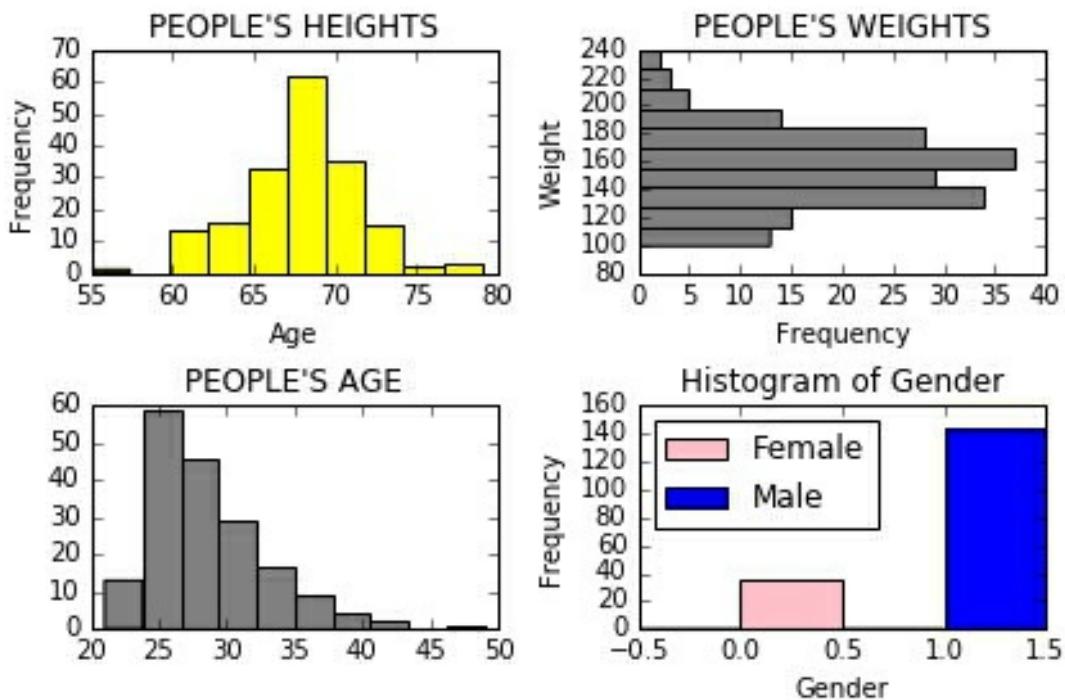
# Histogram of weight
ax1.hist(ht_wt_df['weight'], histtype='bar',
         orientation='horizontal', color='grey')
ax1.set_title("PEOPLE'S WEIGHTS")
ax1.set_xlabel('Frequency')
ax1.set_ylabel('Weight')

# Histogram of age
ax2.hist(ht_wt_df['age'], histtype='bar', color='gray')
ax2.set_title("PEOPLE'S AGE")
ax0.set_xlabel('Age')
ax0.set_ylabel('Frequency')

# Histogram of gender
ax3.hist(gend_f_df['female'], color='pink', bins=2)
ax3.hist(gend_m_df['female'], color='blue', bins=2)
ax3.set_title("Histogram of Gender")
ax3.set_xlabel('Gender')
ax3.set_ylabel('Frequency')
ax3.legend(['Female', 'Male'], loc=0)

plt.tight_layout()
out_dir_file_image = OUT_IMG_NAME + "_multi.png"
write_image(DRIVE, DATA_DIR_PATH, out_dir_file_image)
plt.show()

```



## 5 Data Mining Techniques

### 5.1 Decision Tree (Supervised Learning)

```
# Important libraries
from sklearn.externals.six import StringIO
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree

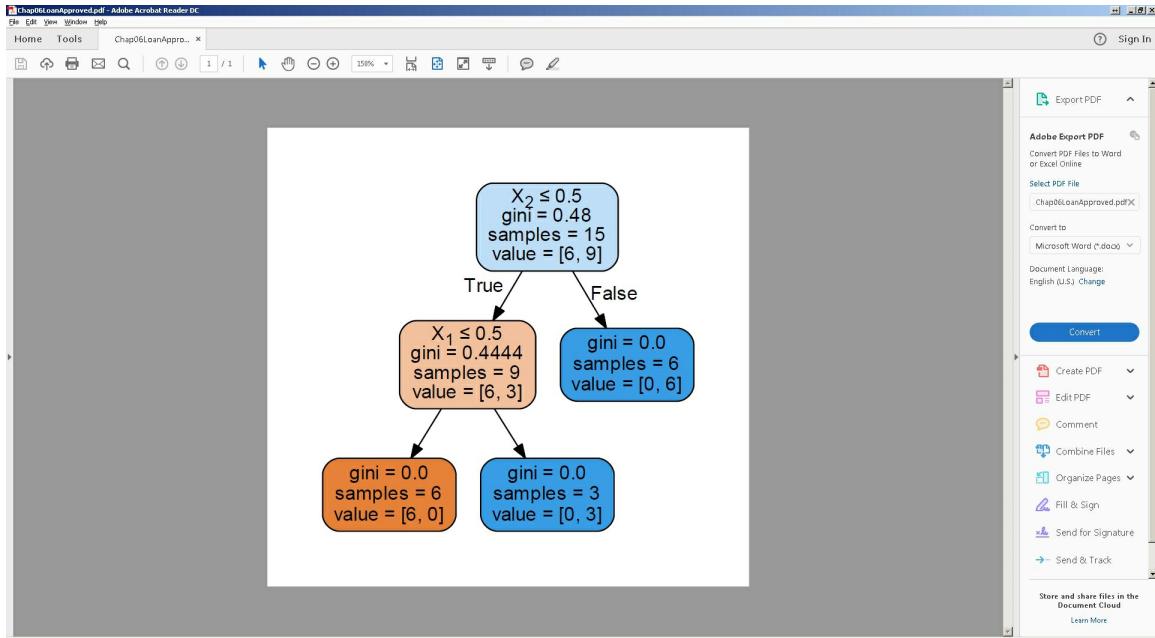
def visualize_tree(self, dec_tree_cls, feature_names):
    drive_file_path = os.path.join(self.drive + self.data_dir_path,
                                   self.out_image)
    dot_data = StringIO()
    try:
        tree.export_graphviz(dec_tree_cls, out_file=dot_data,
                             filled=True, rounded=True,
                             special_characters=True)
        graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
        graph.write_pdf(drive_file_path)
    except:
```

```

exit("Could not run dot, ie graphviz, to "
    "produce visualization")

```

Above code created .pdf file in the current working directory.



```

def get_code(self, dec_tree_cls, feature_names, target_names,
            spacer_base="  "):
    left    = dec_tree_cls.tree_.children_left
    right   = dec_tree_cls.tree_.children_right
    threshold = dec_tree_cls.tree_.threshold
    features = [feature_names[i] for i in dec_tree_cls.tree_.feature]
    value = dec_tree_cls.tree_.value

    def recurse(left, right, threshold, features, node, depth):
        spacer = spacer_base * depth
        if (threshold[node] != -2):
            print(spacer + "if ( " + features[node] + " <= " + \
                  str(threshold[node]) + " ) {")
            if left[node] != -1:
                recurse(left, right, threshold, features,
                        left[node], depth+1)

```

```

print(spacer + "}\n" + spacer +"else {")
if right[node] != -1:
    recurse(left, right, threshold, features,
            right[node], depth+1)
    print(spacer + "}")
else:
    target = value[node]
    for i, v in zip(np.nonzero(target)[1],
                    target[np.nonzero(target)]):
        target_name = target_names[i]
        target_count = int(v)
        print(spacer + "return " + str(target_name) + \
              " (" + str(target_count) + " examples )")
    recurse(left, right, threshold, features, 0, 0)

```

Above code created this this group of statements in the IPython Console.

```

if ( House <= 0.5 ) {
    if ( Job <= 0.5 ) {
        return 0 ( 6 examples )
    }
    else {
        return 1 ( 3 examples )
    }
}
else {
    return 1 ( 6 examples )
}

```

## 5.2 Regression (Supervised Learning)

With Python it is possible work with multiple linear regression.

```
self.ht_wt_df = self.get_data()
```

\*\*\* Formulas \*\*\*

\*\*\* ht\_wt\_df \*\*\*

height age gender weight

0	71.0	24	M	165.0
1	68.0	29	M	165.0
2	72.0	34	M	180.0
3	67.0	21	F	113.0
4	72.0	32	M	178.0

height age gender weight

175	65.0	31	M	120.0
176	70.0	33	M	178.2
177	68.0	34	M	140.0
178	71.0	26	M	130.0
179	69.0	27	M	170.0

\*\*\* OLS Regression Results Parameters (formulas) \*\*\*

Intercept -58.703262  
gender[T.M] 20.076624  
height 2.734531  
age 0.431700

\*\*\* OLS Regression Results Summary (formulas \*\*\*

OLS Regression Results

---

Dep. Variable: weight R-squared: 0.305  
Model: OLS Adj. R-squared: 0.293  
Method: Least Squares F-statistic: 25.72  
Date: Tue, 24 Jan 2017 Prob (F-statistic): 7.58e-14  
Time: 15:26:40 Log-Likelihood: -819.88

No. Observations:	180	AIC:	1648.		
Df Residuals:	176	BIC:	1661.		
Df Model:	3				
Covariance Type:	nonrobust				
<hr/>					
	coef	std err	t	P> t	[95.0% Conf. Int.]
<hr/>					
Intercept	-58.7033	36.826	-1.594	0.113	-131.380 13.973
gender[T.M]	20.0766	4.928	4.074	0.000	10.350 29.803
height	2.7345	0.542	5.050	0.000	1.666 3.803
age	0.4317	0.402	1.075	0.284	-0.361 1.224
<hr/>					
Omnibus:	8.771	Durbin-Watson:	1.961		
Prob(Omnibus):	0.012	Jarque-Bera (JB):	9.235		
Skew:	0.553	Prob(JB):	0.00988		
Kurtosis:	2.905	Cond. No.	1.56e+03		
<hr/>					

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.56e+03. This might indicate that there are strong multicollinearity or other numerical problems.

\*\*\* Rainbow test for linearity \*\*\*

(1.1947660651585406, 0.2034517105727254)

*There are a number of other runs: categorization of variables (multiple*

*regression) and contrast coding systems for categorical variables: (Patsy) in the code segment but I did not want to paste the tables and results in the Word Document as it does not format nicely as displayed above!*

### 5.3 Correlation (Supervised Learning)

In the same way described in decision tree, you can install the necessary library and load the data.

Using self.ht\_wt\_df.cov() you can see relation. Ran with and without gender.

\*\*\* Covariance Data 1 \*\*\*

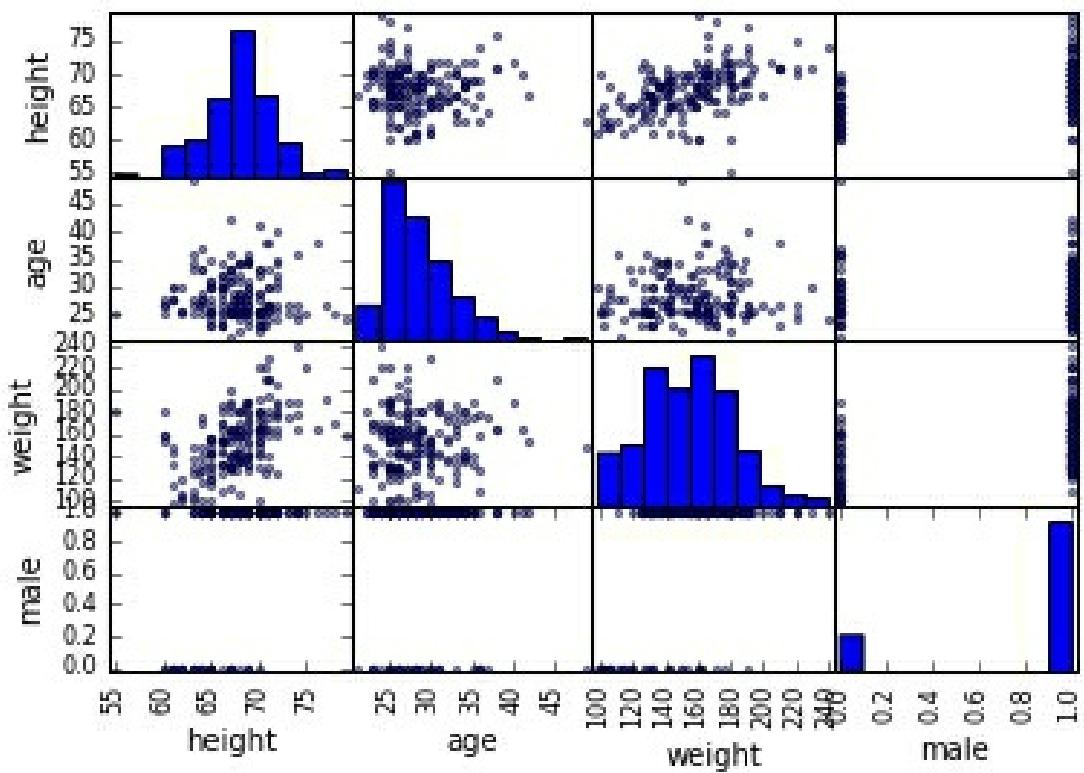
	height	age	weight	male
height	13.008211	-0.261263	48.528256	0.650990
age	-0.261263	18.820453	5.450666	-0.097610
weight	48.528256	5.450666	765.871420	4.900301
male	0.650990	-0.097610	4.900301	0.157511

\*\*\* Covariance Data 2 \*\*\*

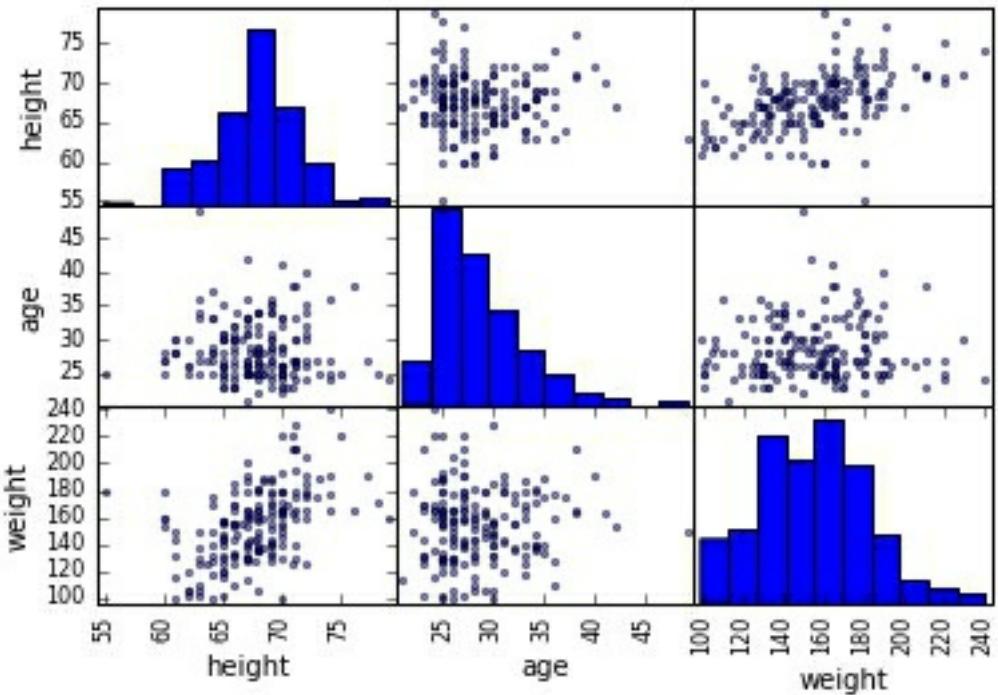
	height	age	weight
height	13.008211	-0.261263	48.528256
age	-0.261263	18.820453	5.450666
weight	48.528256	5.450666	765.871420

Scatter plotting the data, you can see the regression

With gender:



Without gender:



## 5.4 Clustering – Kmeans (Unsupervised Learning)

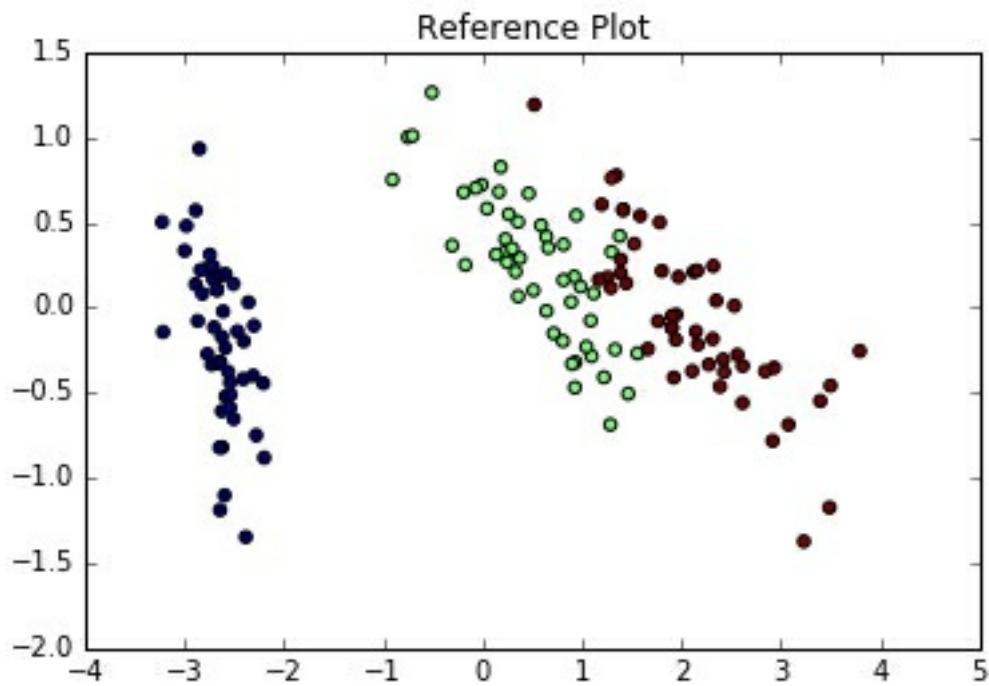
.....  
The following 2 lines reduce the number of dimensions by applying  
a dimensionality reduction algorithm that operates on 4 numbers  
and outputs two new numbers so you can plot.

.....

```
self.pca = PCA(n_components=2).fit(self.iris.data)
self.pca_2d = self.pca.transform(self.iris.data)

if DEBUG == True:
    print(self.pca_2d)

self.title = 'Reference Plot'
pl.clf()
pl.figure(self.title)
pl.title(self.title)
pl.scatter(self.pca_2d[:, 0], self.pca_2d[:, 1], c=self.iris.target)
self.write_image(self.out_image + "_ref_plot.png")
pl.show()
```

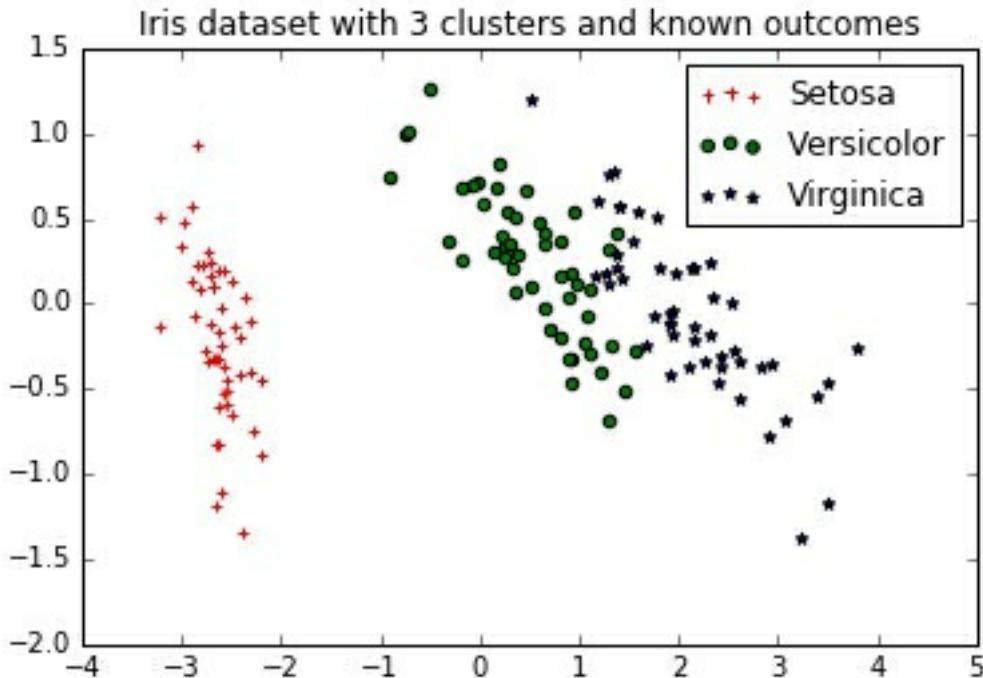


```
for self.idx in range(0, self.pca_2d.shape[0]):  
    if self.iris.target[self.idx] == 0:  
        self.c1 = pl.scatter(self.pca_2d[self.idx, 0],  
                            self.pca_2d[self.idx, 1],  
                            c='r',  
                            marker='+')  
  
    elif self.iris.target[self.idx] == 1:  
        self.c2 = pl.scatter(self.pca_2d[self.idx, 0],  
                            self.pca_2d[self.idx, 1],  
                            c='g',  
                            marker='o')  
  
    elif self.iris.target[self.idx] == 2:  
        self.c3 = pl.scatter(self.pca_2d[self.idx, 0],  
                            self.pca_2d[self.idx, 1],  
                            c='b',  
                            marker='*')  
  
self.title = 'Iris dataset with 3 clusters and known outcomes'
```

```

    pl.legend([self.c1, self.c2, self.c3],
              ['Setosa', 'Versicolor', 'Virginica'])
    pl.title(self.title)
    self.write_image(self.out_image + "_3_clusters_known.png")
    pl.show()

```



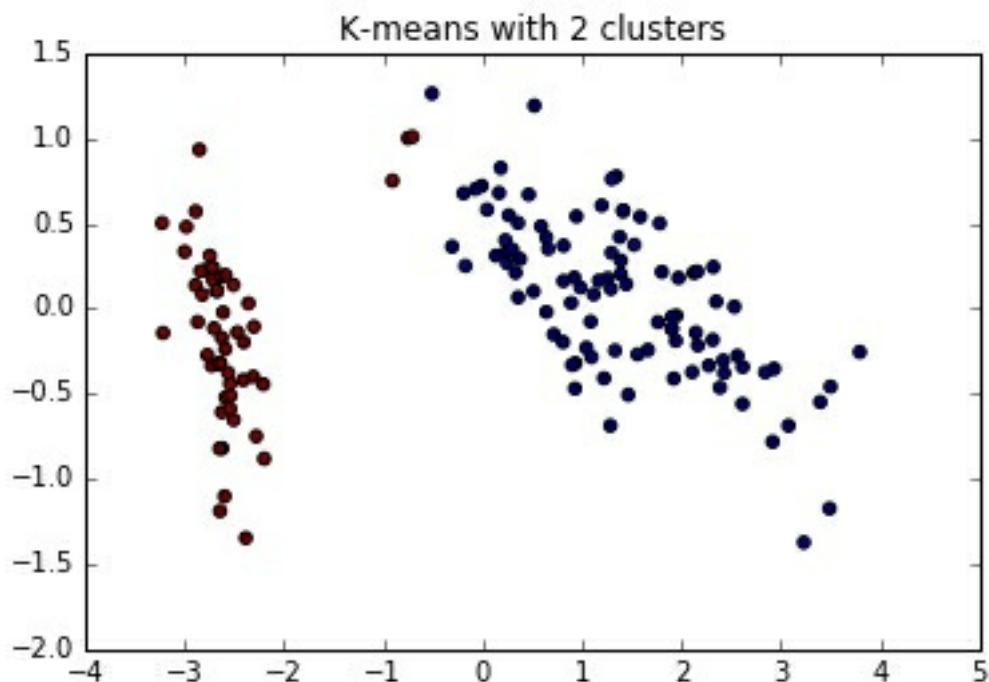
```

Scatter plots with a variety of k-means ranges
for self.k_range in range(self.start_k_range,
                           self.end_k_range):
    if DEBUG == True:
        print('k = %d' % self.k_range)
        self.kmeans = KMeans(n_clusters=self.k_range, random_state=111)
    self.kmeans.fit(self.iris.data)

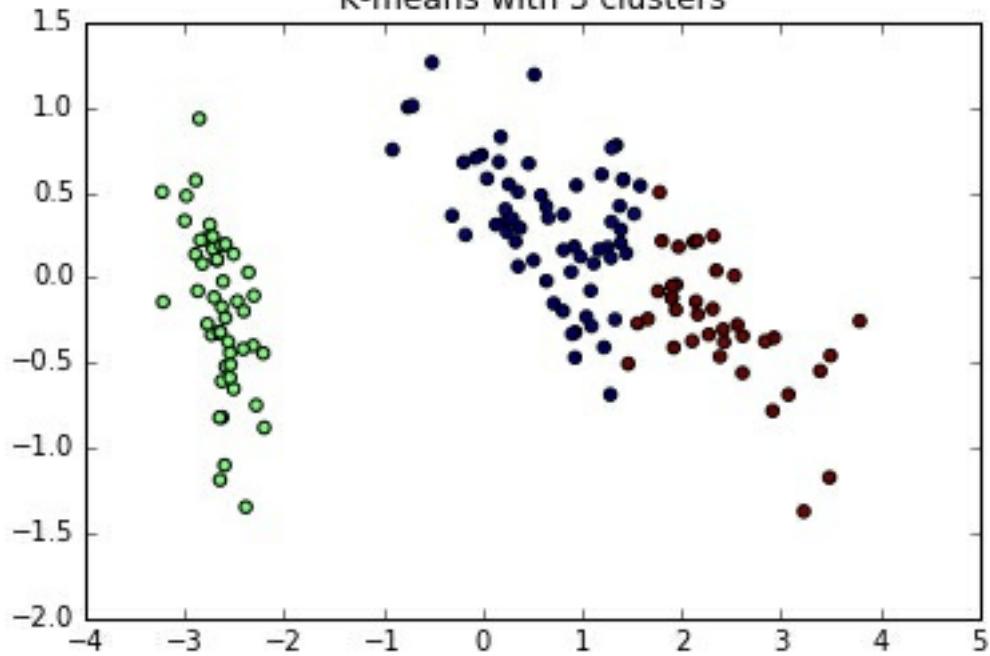
    self.title = 'K-means with ' + str(self.k_range) + ' clusters'
    pl.clf()
    pl.figure(self.title)
    pl.title(self.title)

```

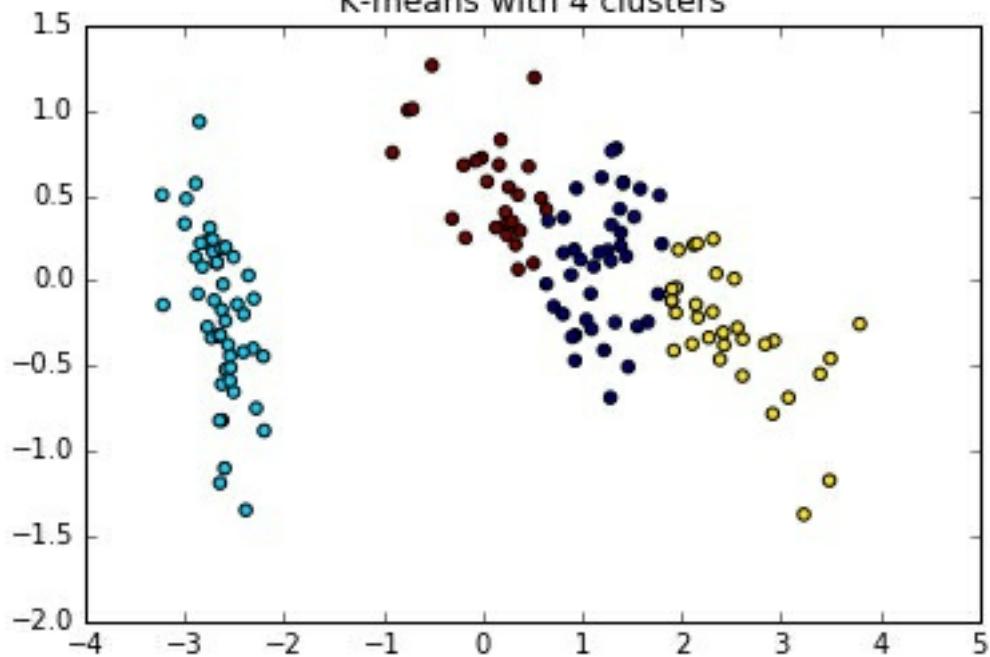
```
pl.scatter(self.pca_2d[:, 0],  
          self.pca_2d[:, 1],  
          c=self.kmeans.labels_)  
  
self.write_image(self.out_image + "_" \  
                 + str(self.k_range) \  
                 + "_clusters.png")  
  
pl.show()
```

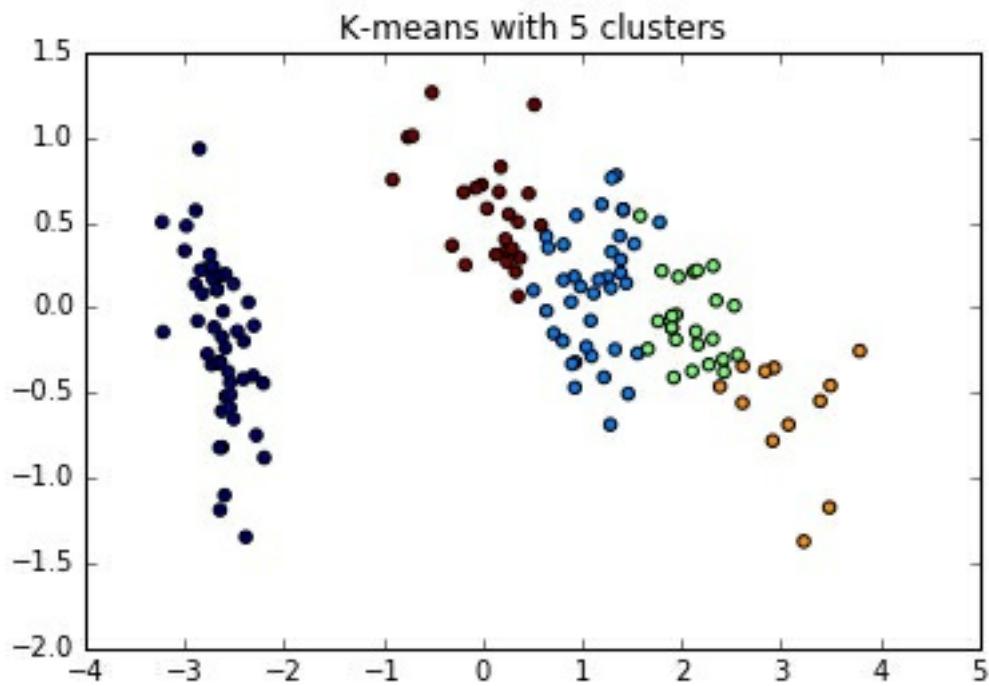


K-means with 3 clusters



K-means with 4 clusters





## 6 Big Data Mining

### 6.1 WordCloud - directory FWordCloud and look at code module WordCloud.py.

```

import string

def remove_punctuation(self, s):
    table = string.maketrans("", "")
    return s.translate(table, string.punctuation)

self.text = self.get_data()
self.text = self.remove_punctuation(self.text)
self.word_cloud = WordCloud().generate(self.text)

plt.imshow(self.word_cloud)
plt.axis("off")
self.out_dir_file_image = self.out_image + "_1.png"
self.write_image(self.out_dir_file_image)

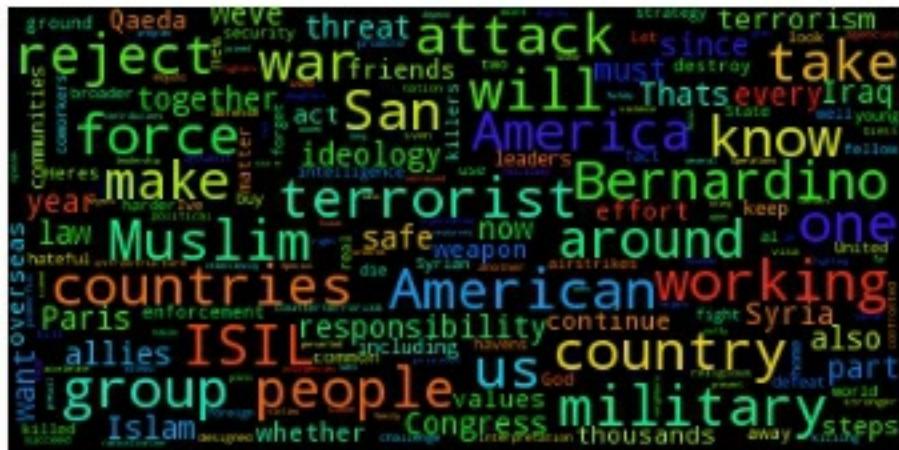
```

```
plt.show()

self.word_cloud = WordCloud(max_font_size=self.max_font_size) \
    .generate(self.text)

plt.figure()
plt.imshow(self.word_cloud)
plt.axis("off")
self.out_dir_file_image = self.out_image + "_2.png"
self.write_image(self.out_dir_file_image)
plt.show()
```



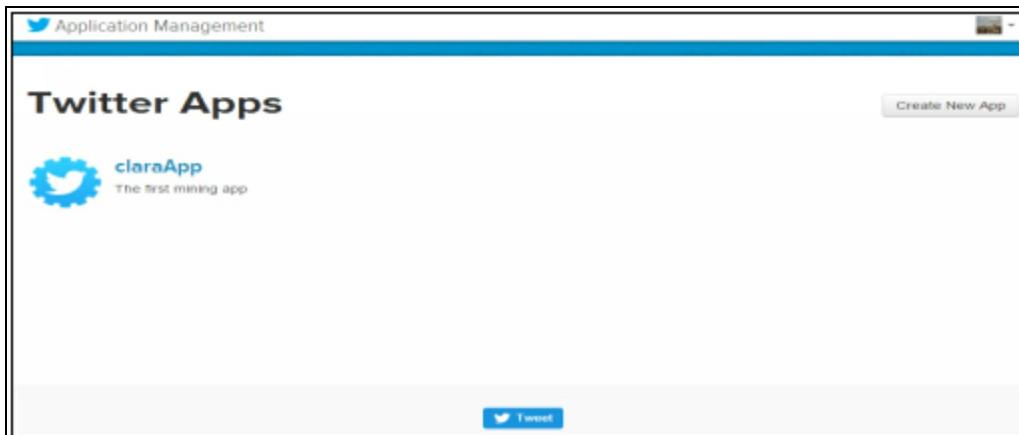


## 6.2 Twitter Mining

- Using twitter hashTag mining

### 6.2.1 Steps (Twitter side)

- Sign in using your twitter account
- <https://twitter.com/apps>



- Click on create a New App

## Create an application

Application Details

Name \*

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 52 characters max.

Description \*

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website \*

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

- Under "keys and Access Tokens" tab, check the consumer key (API key), consumer secret (API secret), access Token and Access Token Secret.

## TwitterMiningApp

Details    Settings    **Keys and Access Tokens**    Permissions

### Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	T3OfRlpmWVgiPVu1LARuQVEqf
Consumer Secret (API Secret)	Ijjd1FN1JmvqGyiKH2YAcSLqemZcbf9NGWxP0kanwwKpPMPBCp
Access Level	Read and write ( <a href="#">modify app permissions</a> )
Owner	nimaclara
Owner ID	84304937

### Application Actions

[Regenerate Consumer Key and Secret](#)    [Change App Permissions](#)

- Under "permissions" tab, authorize your application. By default, it is “read and write”. Select “Read, write and Access direct messages”. And then click update settings

The screenshot shows the Twitter Mining App's settings interface. The top navigation bar includes 'Details', 'Settings', 'Keys and Access Tokens', and 'Permissions'. The 'Permissions' tab is active. Below it, the 'Access' section asks what type of access the application needs, with three options: 'Read only', 'Read and Write', and 'Read, Write and Access direct messages'. The third option is selected. A note below states: 'Changes to the application permission model will only reflect in access tokens obtained after the permission model change is saved. You will need to re-negotiate existing access tokens to alter the permission level associated with each of your application's users.' At the bottom left is a 'Update Settings' button.

### 6.2.2 Python code

# Lots of imports!!

```
from nltk.tokenize import TweetTokenizer  
from tweepy import OAuthHandler  
from tweepy import Stream  
from tweepy.streaming import StreamListener  
from wordcloud import WordCloud  
import json  
import matplotlib.pyplot as plt  
import os.path  
import pandas as pd  
import re  
import sys  
import time  
import webbrowser  
# Lots of Constants  
DEBUG = False  
APPEND_TO_FILE = False  
DISPLAY_TWEETS_FILE = True
```

```

OPEN_LINKS = False
DRIVE = "E:\\"
DATA_DIR_PATH = "PythonTutorial\\GTwitterMining"
FILE_NAME = "twitter_jobs.json"
TWEET_TIME_LIMIT = 2400
DISPLAY_RECORDS = 5
MAX_FONT_SIZE = 25
# Twitter keys
CONSUMER_KEY = "T3OfRIpmWVgiPVu1LARuQVEqf"
CONSUMER_SECRET =
"Ijjd1FN1JmvqGyiKH2YAcSLqemZcbf9NGWxP0kanwwKpPMPBCp"
ACCESS_TOKEN = "84304937-Vi1CYqNRbjjsoS7ShrjSY7SZHXfmtIqvS4qdXT3w6"
ACCESS_TOKEN_SECRET = "HyAdI4yWEXe00Jk1xdjzCEcxIzCNRH1vmdnvkTuL1tzjP"

auth = OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN, ACCESS_TOKEN_SECRET)

# To acquire tweets for n seconds (2400 based on the constant TWEET_TIME_LIMIT)
class MyStreamListener(StreamListener):
    def __init__(self,
                 time_limit,
                 drive,
                 data_dir,
                 out_file, ):
        self.start_time = time.time()
        self.limit = time_limit
        self.drive = drive
        self.data_dir_path = data_dir
        self.out_file_name = out_file

        tweets_data_path = os.path.join(self.drive + self.data_dir_path,
                                         self.out_file_name)
        self.writeFile = open(tweets_data_path, 'a')
        super(MyStreamListener, self).__init__()

    def on_data(self, data):

```

```

"""
Subtract the current time from the start time and check if it is
less than the time limit
"""

if (time.time() - self.start_time) < self.limit:
    self.writeFile.write(data)
    self.writeFile.write("\n")
    return True
else:
    self.writeFile.close()
    return False

def on_error(self, status):
    print(status)

myStream = Stream(auth, listener=MyStreamListener(TWEET_TIME_LIMIT,
                                                 DRIVE,
                                                 DATA_DIR_PATH,
                                                 FILE_NAME))

myStream.filter(track=['postdoc', 'postdoctoral'])

# To process acquired tweets

def get_data(self):
    tweets_data_path = os.path.join(self.drive + self.data_dir_path,
                                    self.in_file_name)

    tweets_data = []
    try:
        tweets_file = open(tweets_data_path, "r")
        for line in tweets_file:
            try:
                tweet = json.loads(line)
                tweets_data.append(tweet)
            except:
                continue
    
```

```

if DEBUG == True:
    print(len(tweets_data))

tweets = pd.DataFrame()

tweets['text'] = list(map(lambda tweet: tweet.get('text', None), tweets_data))
tweets['lang'] = list(map(lambda tweet: tweet.get('lang', None), tweets_data))
tweets['country'] = list(map(lambda tweet: tweet.get('place',
                                                       {}).get('country', None) if tweet.get('place') != None
                                                       else
                                                       None, tweets_data))

return tweets

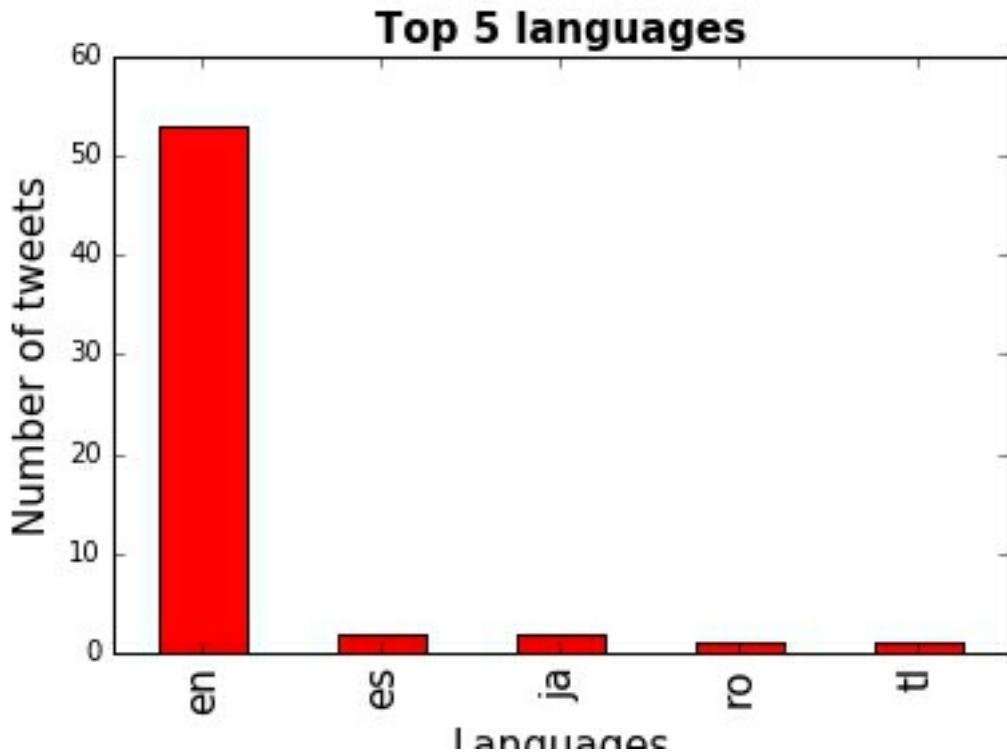
self.the_tweets = self.get_data()

if DEBUG == True:
    print("\n*** head and tail of tweets ***")
    print(self.the_tweets.head(n=DISPLAY_RECORDS))
    print(self.the_tweets.tail(n=DISPLAY_RECORDS))

# Tweets by language
self.tweets_by_lang = self.the_tweets['lang'].value_counts()

fig, ax = plt.subplots()
ax.tick_params(axis='x', labelsize=15)
ax.tick_params(axis='y', labelsize=10)
ax.set_xlabel('Languages', fontsize=15)
ax.set_ylabel('Number of tweets', fontsize=15)
ax.set_title('Top 5 languages', fontsize=15, fontweight='bold')
self.tweets_by_lang[:5].plot(ax=ax, kind='bar', color='red')
self.out_dir_file_image = "Top5Languages.png"
self.write_image("Top5Languages.png")

```



There are also bar graphs of the “Top 5 Countries”, “Ranking of postdoc and postdoctoral (Raw Data)”, “Ranking of postdoc and postdoctoral (Relevant Data)”

Now the code listing for the WordCloud. I filtered only on English as there was a number of tweets that were in Chinese and Japanese and I wasn’t sure how to handle these in a WordCloud.

.....

Generate a word cloud image

.....

```
self.the_tweets = self.the_tweets[self.the_tweets['lang'] == 'en']
self.tweets_text = " "
if DEBUG == True:
    print("\n*** rows ***")
self.tknzr = TweetTokenizer(strip_handles=True, reduce_len=True)
for row in self.the_tweets.itertuples():
    if row != None:
        words = self.tknzr.tokenize(self.remove_link(row[1]))
        for word in words:
            word = word.encode('utf-8') # remove u' (unicode) stuff
    if DEBUG == True:
        print("\n*** word ***")
        print(word)
```

```

        self.tweets_text = self.tweets_text + " " + str(word)

if DEBUG == True:
    print("\n*** tweets_text ***")
    print(self.tweets_text)
self.word_cloud = WordCloud(max_font_size=self.max_font_size) \
    .generate(self.tweets_text)

plt.figure()
plt.imshow(self.word_cloud)
plt.axis("off")
self.write_image("TweetsWordCloud.png")
plt.show()

```



## 6.3 Page Rank

```

# Imports

import matplotlib.pyplot as plt

import networkx as nx

import os.path

DEBUG = True

```

```

DRIVE = "E:\\"
DATA_DIR_PATH = "PythonTutorial\\HPageRank"
OUT_IMG_NAME = "PageRank.png"

G = nx.DiGraph()

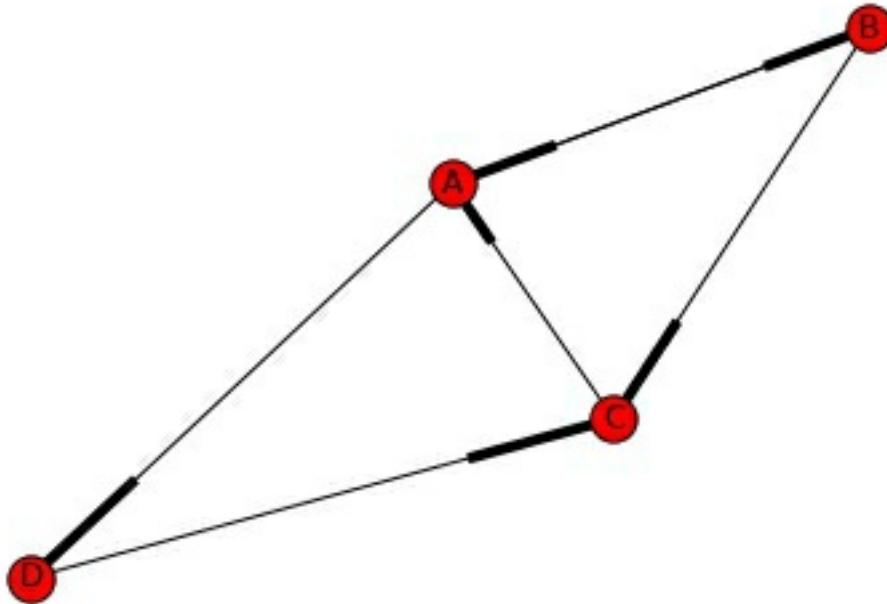
G.add_edges_from([('A', "B"), ("B", "A"), ("B", "B"), ("B", "C"), ("C", "A")])
G.add_node("D")
edge1=("A", "D")
G.add_edge(*edge1)
edge2=("D", "C")
G.add_edge(*edge2)
edge3=("D","D")
G.add_edge(*edge3)

pr = nx.pagerank(G)

nx.draw(G, with_labels=True)
img = os.path.join(DRIVE + DATA_DIR_PATH, OUT_IMG_NAME)
plt.savefig(img)
plt.show()

*** Page Ranks ***
{'A': 0.28511854592348873, 'C': 0.2175145358898481, 'B': 0.22140856006497459, 'D': 0.2759583581216885}

```



## 7 Additional Documentation

At the top of each .py file in all the tutorials are numerous references that I used in creating the appropriate tutorial for that section. Without the hard work of the people who created the code that I used for references I would still be on the first tutorial – Data Visualization! As a developer it is of the utmost importance that you are able to search all available resources and find an algorithm, and/or code that you can utilize. It is also of the utmost importance to give credit to the individual(s) that helped you along the way.

<http://datasciencesource.com/python-with-spyder-tutorial/>

<https://pythontips.com/2013/07/30/20-python-libraries-you-cant-live-without/>

<http://programminghistorian.org/lessons/text-mining-with-extracted-features>

<http://www.nltk.org/>

<http://pandas.pydata.org/>

<http://scikit-learn.org/stable/index.html>

<http://matplotlib.org/>

<http://www.tweepy.org/>

<http://statsmodels.sourceforge.net/devel/contrasts.html>

## **Additional Resources**

Here are some other books, papers and other resources, for a deeper dive into the topics covered in this book.

1. Ayres, I. (2007) **SuperCrunchers: Why Thinking-by-Numbers Is the New Way to be Smart.** Random House Publishing.
2. Davenport, T. & J. Harris (2007). **Competing on Analytics: The New Science of Winning.** HBS Press.
3. Gartner (2012). Business Implications of Big Data.
4. Gartner (2012). Technology Implications of Big Data.
5. Gordon Linoff & Michael Berry (2011). **Data Mining Techniques.** 3<sup>rd</sup> edition. Wiley.
6. Groebner, David F,P.W. Shannon, P.C. Fry. (2013). *Business Statistics* (9<sup>th</sup> edition). Pearson.
7. Jain, Anil K. (2008). “Data Clustering: 50 years beyond K-Means.” 19<sup>th</sup> International Conference on Pattern Recognition.
8. Lewis, Michael (2004). **Moneyball: The Art of Winning an Unfair Game.** Norton & Co.
9. Andrew D Martin et al (2004). “*Competing Approaches to Predicting Supreme Court Decision making*”, *Perspective in Politics*.
10. Mayer-Schonberger, Viktor; Cukier, Kenneth (2013). **Big Data: A Revolution That Will Transform How We Live, Work, and Think** . Houghton Mifflin Harcourt.
11. McKinsey Global Institute Report (2011). **Big data: The next frontier for innovation, competition, and productivity.** Mckinsey.com
12. Sathi, Arvind (2011). **Customer Experience Analytics: The Key to Real-Time, Adaptive Customer Relationships.** Independent Publishers Group.
13. Sharda, R., D. Dusen, and E. Turban. (2014). **Business Intelligence and Data Analytics.** 10<sup>th</sup> edition. Pearson.
14. Shmueli, G, N. Patel, & P. Bruce (2010). **Data Mining for Business Intelligence.** Wiley.
15. Siegel, Eric, (2013). **Predictive Analytics.** Wiley.
16. Silver, N. (2012). **The Signal and the Noise: Why So**

**Many Predictions Fail but Some Don't.** Penguin Press.

17. Statsoft. [www.statsoft/textbook](http://www.statsoft/textbook)
18. Taylor, James (2011). **Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics** (IBM Press). Pearson Education.
19. Weka system. <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
20. Witten, I., E. Frank, M. Hall (2009). **Data Mining**. 3<sup>rd</sup> edition. Morgan Kauffman.

## About the Author



**Anil Maheshwari, Ph.D.**, is a professor of Management at the Maharishi University of Management, Fairfield, Iowa, since 2012. He did B.Tech. from IIT Delhi, MBA from IIM Ahmedabad, and Ph.D. from Case Western Reserve University. He has previously been on the faculty of the University of Cincinnati, City University of New York, and others. He has published over a dozen papers and has also delivered dozens of seminars and talks around the world on technology, management, and spirituality.

He is Director of the MBA in MIS program and teaches courses in data analytics and information systems. He is the author of five books including global #1 bestseller [Data Analytics Made Accessible](#). He has 20 years of IT industry experience including leadership roles at IBM Corporation in Austin,

TX, for 9 years. His recent research focuses on Bliss, Vedas, Leadership, and World Peace. For his leadership in organizing blissful community programs, he received the coveted Maharishi Award in 2015. He blogs on Blissful Living at [anilmah.com](http://anilmah.com). He ran a marathon in 2006. He can be contacted for any seminars / speaking engagements at [akm2030@gmail.com](mailto:akm2030@gmail.com).