NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

1.0 Introduction to Data Mining

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

@thephdataminer

Module 3 of the Business Intelligence and Analytics Certification of UP NEC and the UP Center for Business Intelligence

UP NEC BI Modules

- Analyst Level
  1. Introduction to Business Intelligence and Data Mining
  2. Data Warehousing
  3. **Data Mining**
- Professional Level
  4. Optimization Analysis
  5. Time Series Analysis and Forecasting
  6. R For Business Analytics

eljalao@up.edu.ph

Outline for This Training

1. Introduction to Data Mining
2. Data Preprocessing
   - Case Study on Big Data Preprocessing using R
3. Classification Methodologies
   - Case Study on Classification using R
4. Regression Methodologies
   - Case Study: Regression Analysis using R
5. Unsupervised Learning
   - Case Study: Social Media Sentiment Analysis using R

Outline for this Session

- Review of BA
- Introduction to Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Review of BA

- Timely
- Accurate
- High-Value
- Actionable

DECISIONS

Via organizational (and sometimes external) data

eljalao@up.edu.ph

Review of BA

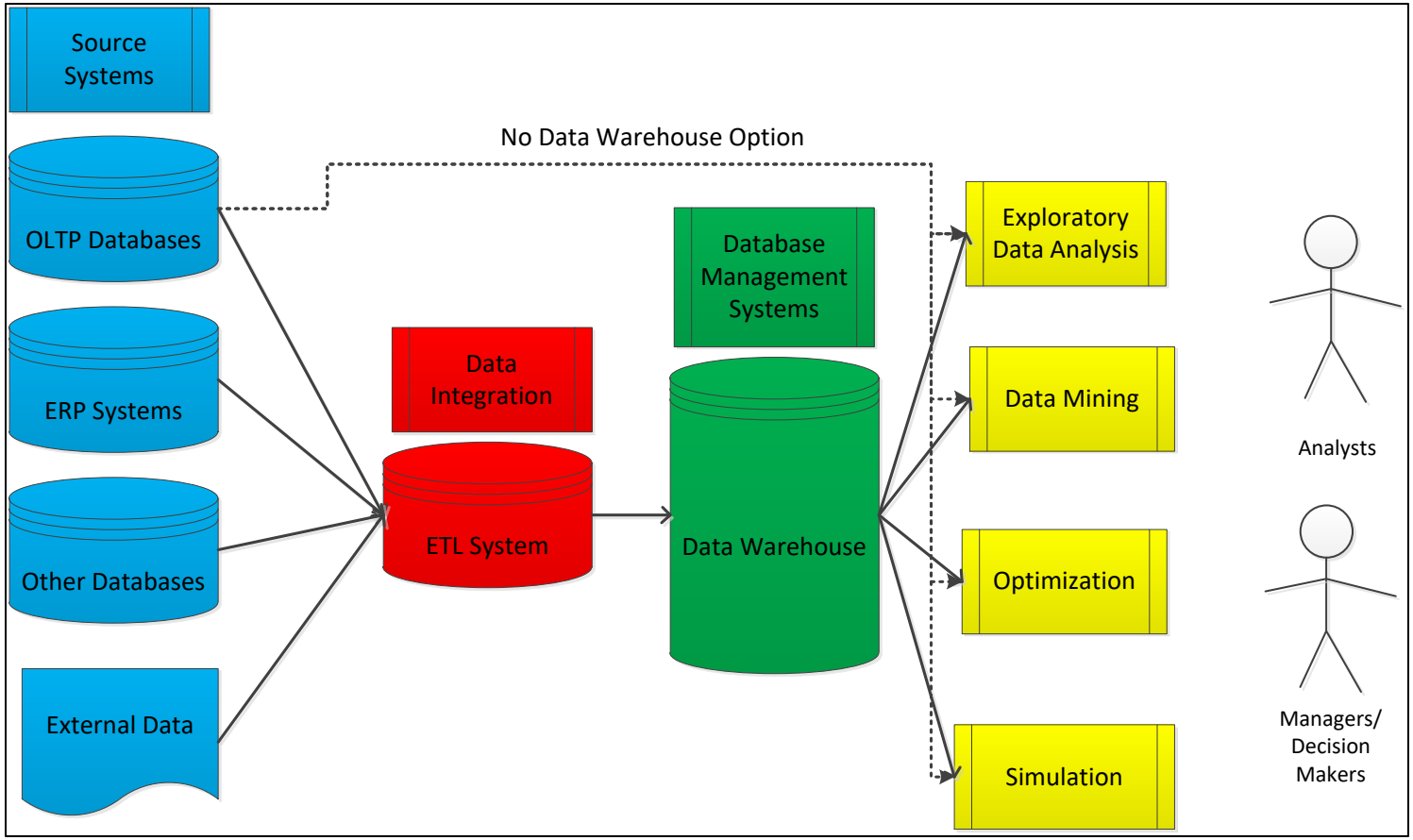


Figure 1.1: BA Framework

Review of BA

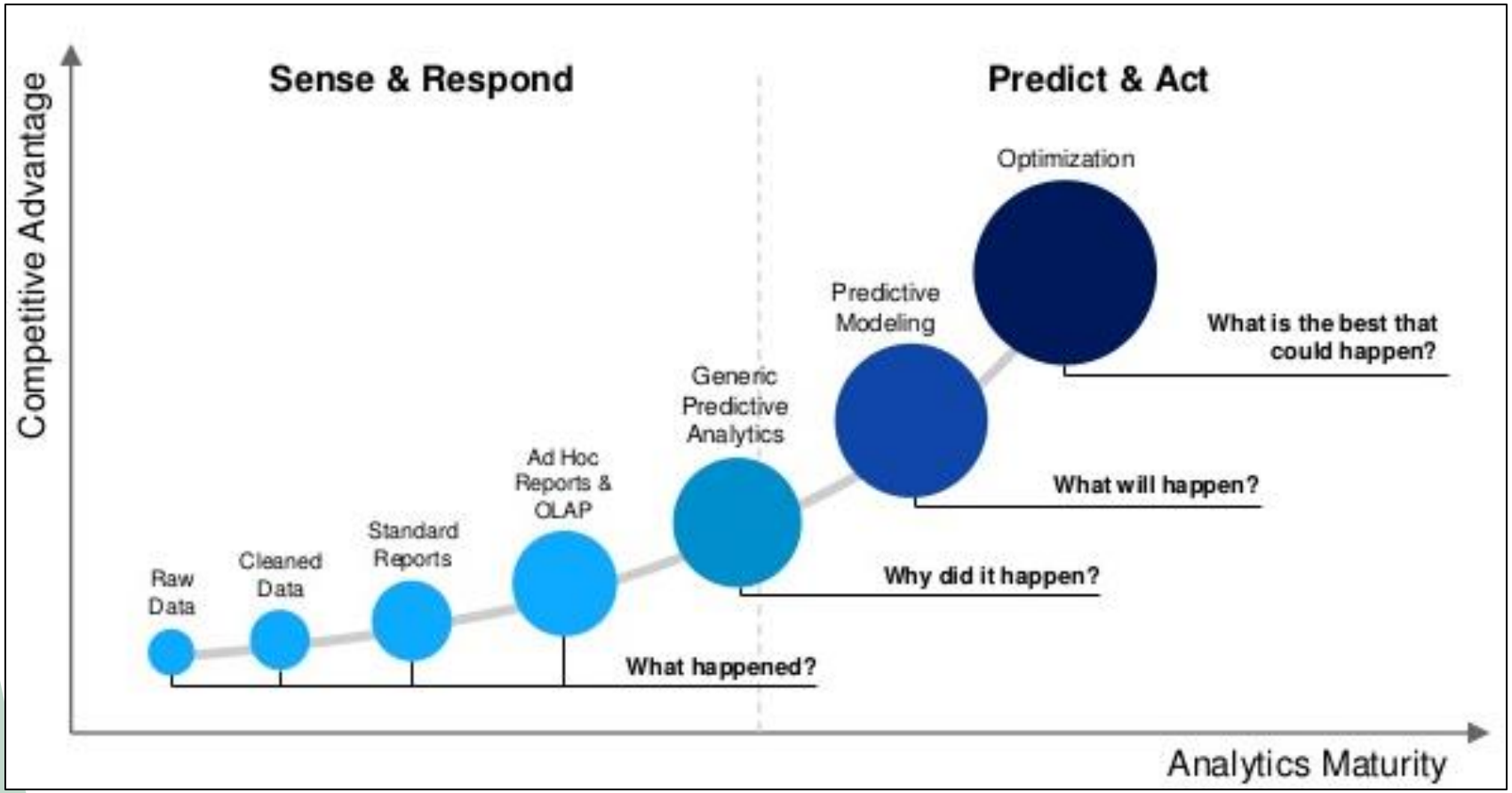


Figure 1.2: Types of BA According to Sophistication

Review of BA

- Exploratory Data Analysis (Descriptive Analytics)
  - *Tell Me What has Happened and Why*
  - *Tell Me What is Happening Right Now*
- Data Mining (Predictive Analytics)
  - *Tell Me What is Likely to Happen*
  - *Tell Me Something Interesting Without Me Asking*
- Optimization/Simulation (Prescriptive Analytics)
  - *Tell Me What Might Have Happened*
  - *Tell Me the Best Solution*

eljalao@up.edu.ph

Outline for this Session

- Review of BA
- **Introduction to Data Mining**
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

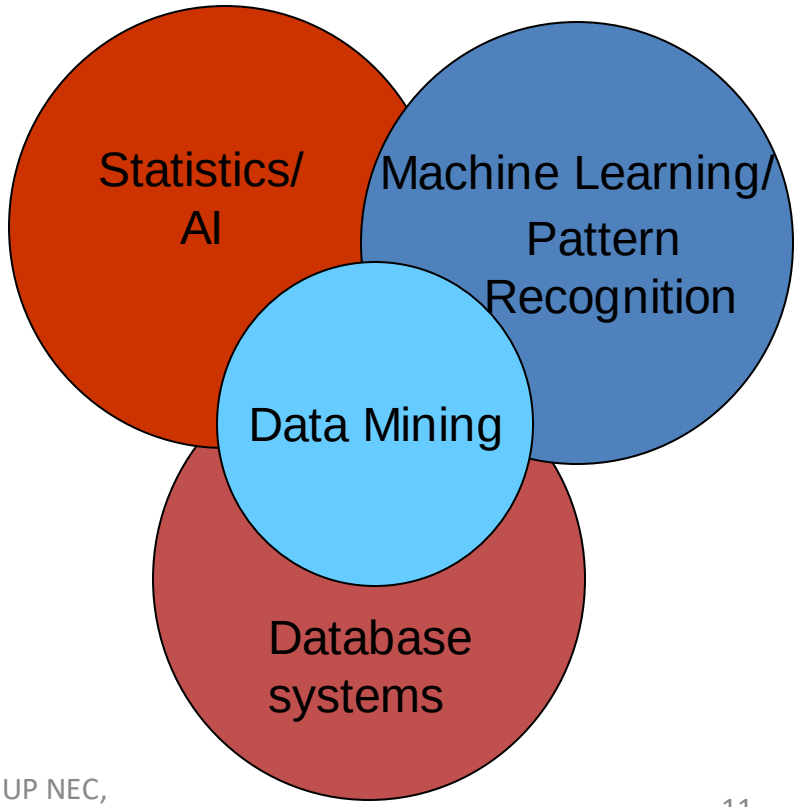
Introduction to Data Mining

Definition 1.1: Data Mining

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- Data Mining is about explaining the past and predicting the future by means of data analysis.

Introduction to Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Introduction to Data Mining

- Types of Data Mining Algorithms
  - Supervised Learning
    - *Classification*
    - *Regression*
  - Unsupervised Learning
    - *Association Analysis*
    - *Sequential Pattern Analysis*
    - *Clustering*
    - *Text Mining/Social Media Sentiment Analysis*

Introduction to Data Mining

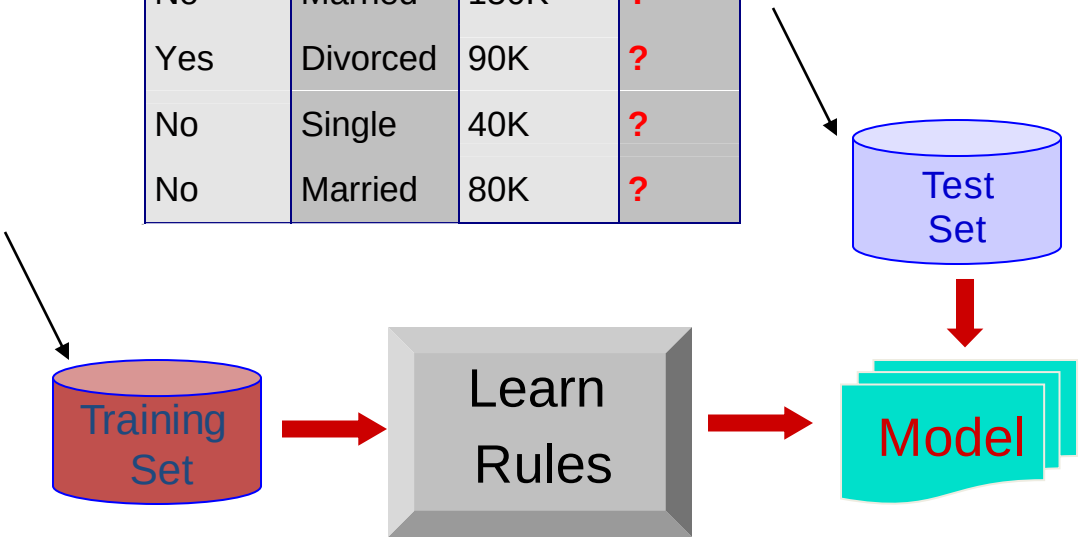
Definition 1.2: Classification

- Classification is a data mining task of predicting the value of a categorical variable by building a model based on one or more numerical and/or categorical variables.

Introduction to Data Mining

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Introduction to Data Mining

Example 1.1: Churn Analysis in Telcos

- Sample model framework for predicting probability of churn of subscribers

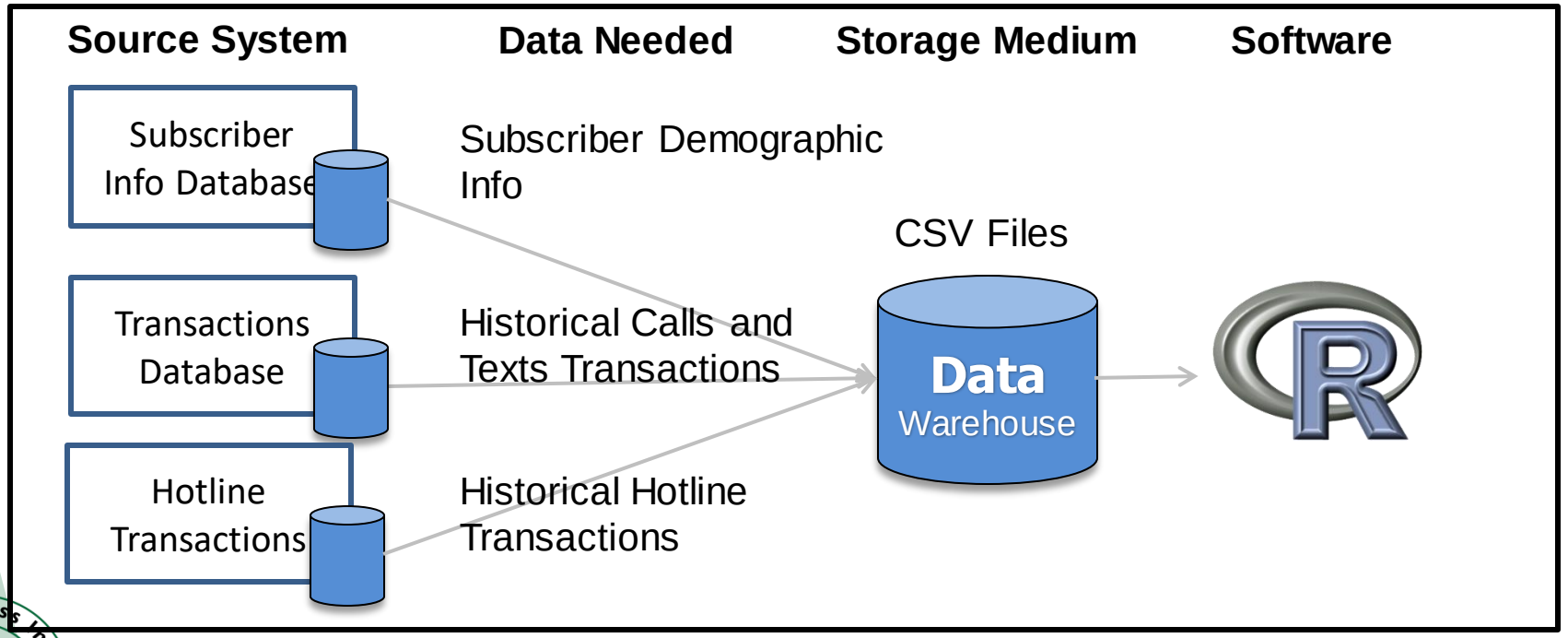


Figure 1.9: Churn Analysis Framework

Introduction to Data Mining

Definition 1.3: Regression

- Regression is a data mining task of predicting the value of target (numerical variable) by building a model based on one or more predictors (numerical and categorical variables).

Introduction to Data Mining

Example 1.2: Manpower Headcount in an FMCG Company

- Create a regression model to predict the headcount of the merchandisers of a supermarket

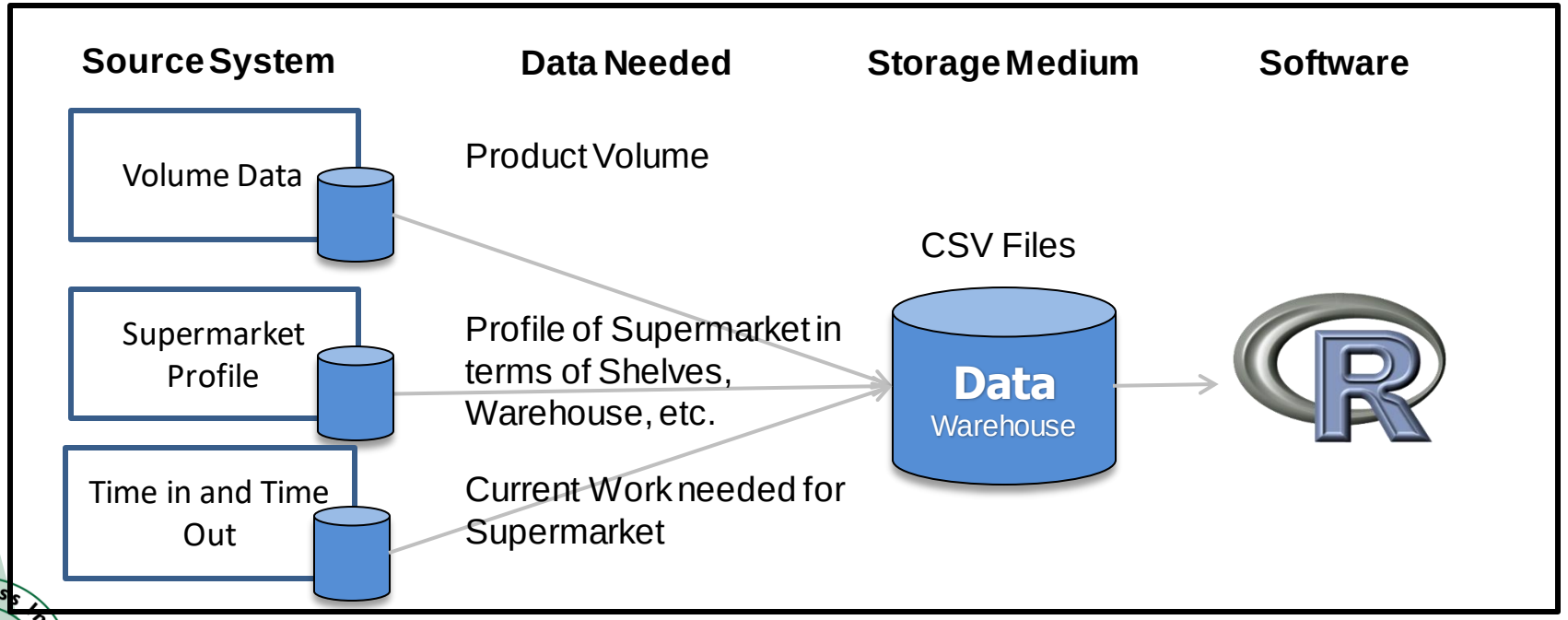


Figure 1.10: Regression Model Framework

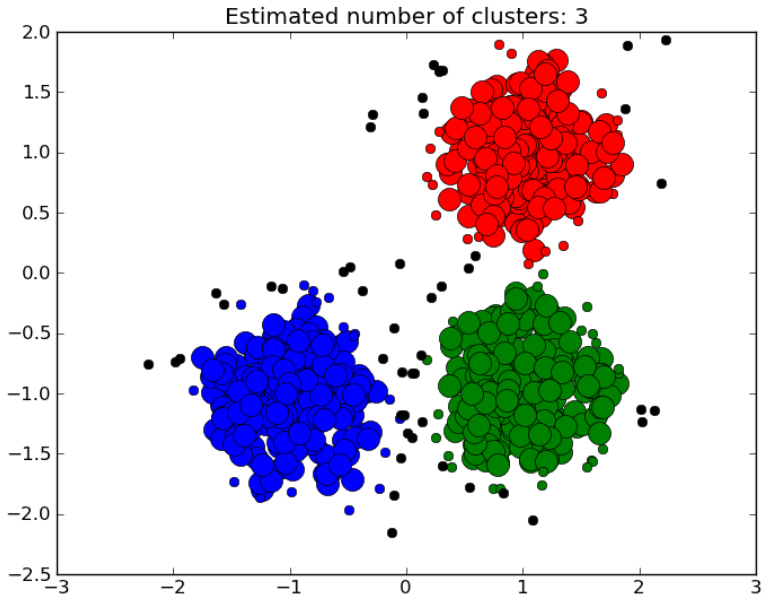
Introduction to Data Mining

- Types of Data Mining Algorithms
  - Supervised Learning
    - *Classification*
    - *Regression*
  - Unsupervised Learning
    - *Association Analysis*
    - *Sequential Pattern Analysis*
    - *Clustering*
    - *Text Mining/Social Media Sentiment Analysis*

Introduction to Data Mining

Definition 1.4: Clustering

- Clustering is the process of dividing a dataset into groups such that the members of each group are as similar (close) as possible to one another, and different groups are as dissimilar (far) as possible from one another.



http://scikit-learn.org/0.10/modules/clustering.html

Introduction to Data Mining

Example 1.3: Market Segmentation

- Goal: **subdivide** a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Introduction to Data Mining

Definition 1.5: Association Rule Analysis

- Definition: Is a data mining task used to identify strong rules that associate elements **together** in datasets using different measures of **interestingness**

Introduction to Data Mining

Example 1.4: Supermarket Basket Analysis

- Given a set of records each of which contain some number of items from a given collection;
- Produce **dependency rules** which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Figure 1.10: Basket Transactions

Introduction to Data Mining

Example 1.5: Promo Pairings for a Restaurant Chain

- Identify which Menu items are ordered **frequently with each other** such that a promo meal can be launched.

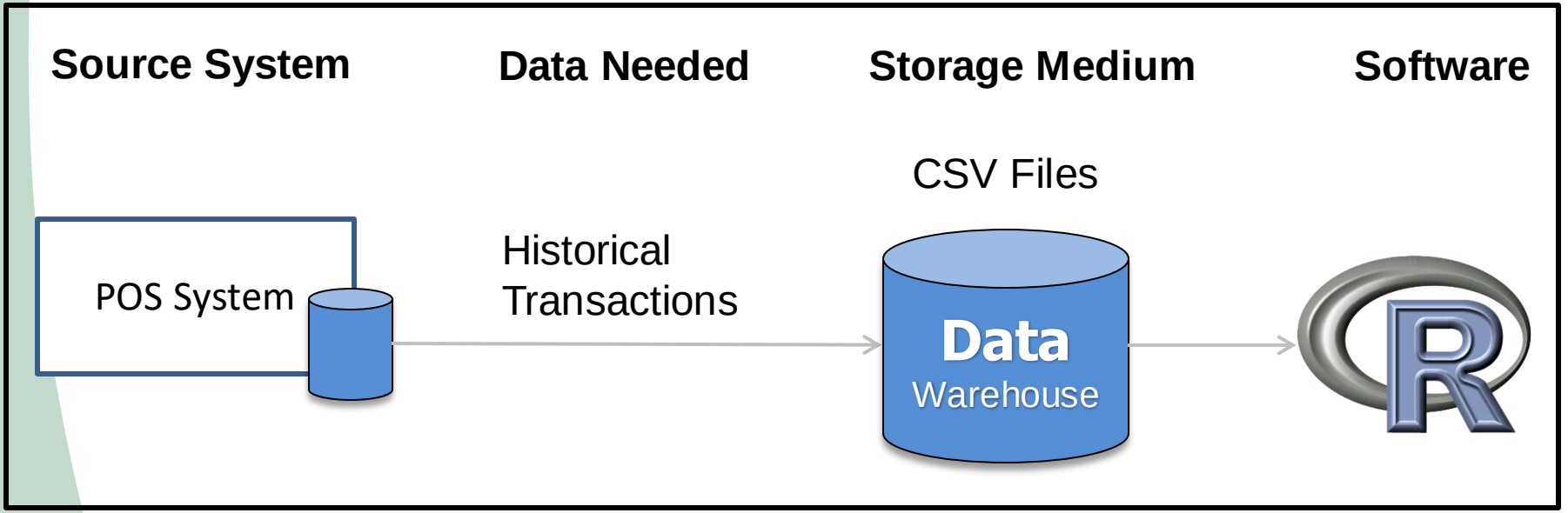


Figure 1.11: Restaurant Promo Pairings

Introduction to Data Mining

Definition 1.6: Sequential Pattern Analysis

- Given a set of objects, with each object associated with its own timeline of events, find rules that predict strong **sequential dependencies among different** events.

(A B) (C) (D E)

- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

Introduction to Data Mining

Example 1.6: Sequence of Calls in a Call Center Hotline

- Identify which Menu items are ordered **frequently with each other** such that a promo meal can be launched.

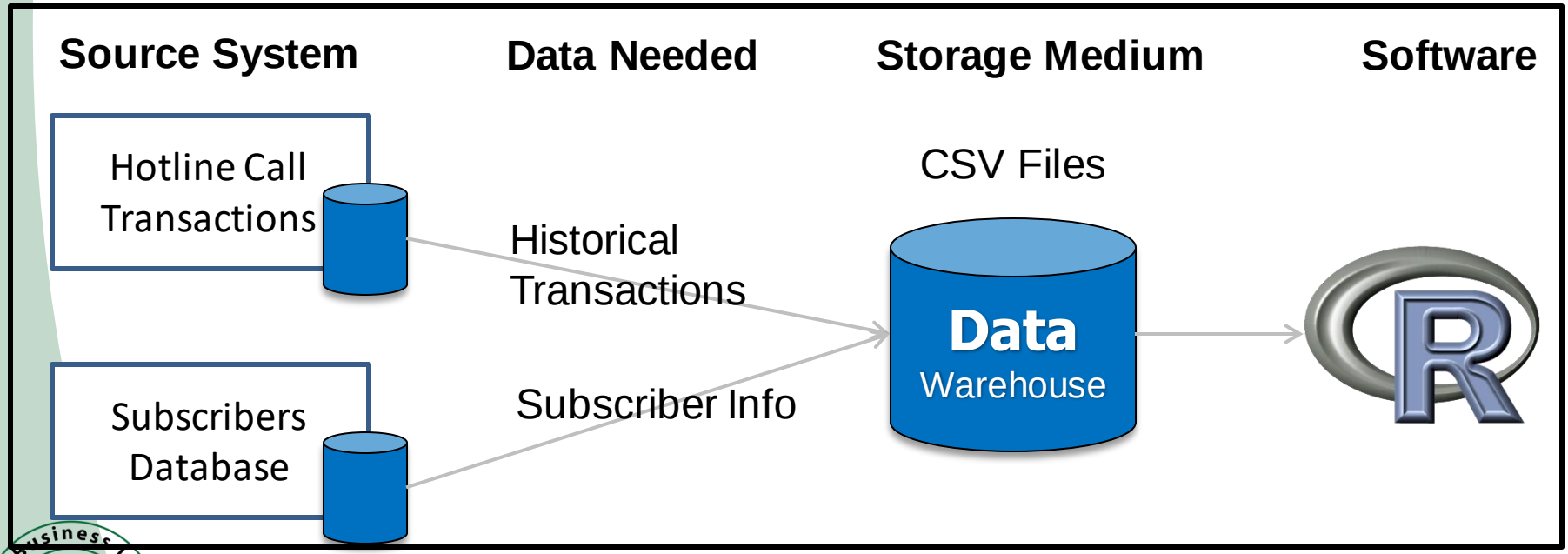


Figure 1.12: Hotline Sequence Model

eljalao@up.edu.ph

Introduction to Data Mining

Example 1.6: Sequence of Calls in a Call Center Hotline

```
21 <{"DEVICE CONFIGURATION"},{"SUCCESSFUL NOT INTERESTED"}>
22     <{"DEVICE CONFIGURATION"},{"SUCCESSFUL INTERESTED"}>
23      <{"MECHANICS PROCEDURE"},{"SUCCESSFUL INTERESTED"}>
24                              <{"SHORT CALL"},{"SHORT CALL"}>
25        <{"MECHANICS PROCEDURE"},{"MECHANICS PROCEDURE"}>
26      <{"DEVICE CONFIGURATION"},{"DEVICE CONFIGURATION"}>
27     <{"SUCCESSFUL INTERESTED"},{"DEVICE CONFIGURATION"}>
28          <{"UNCOMPLETED CALL"},{"DEVICE CONFIGURATION"}>
29                <{"ACCOUNT DETAILS"},{"BILLING INQUIRY"}>
30                <{"BILLING INQUIRY"},{"BILLING INQUIRY"}>
31          <{"AFTERSALES REQUEST"},{"AFTERSALES REQUEST"}>
32             <{"BILLING INQUIRY"},{"AFTERSALES REQUEST"}>
33                <{"BILLING INQUIRY"},{"ACCOUNT DETAILS"}>
```

Figure 1.13: Hotline Sequence Transactions

The Business Analytics Framework

Definition 1.7: Text Mining

- Finding **frequently** occurring words from unstructured data, e.g. word files, reviews, journals, articles.

The Business Analytics Framework

Example 1.7: Word Cloud

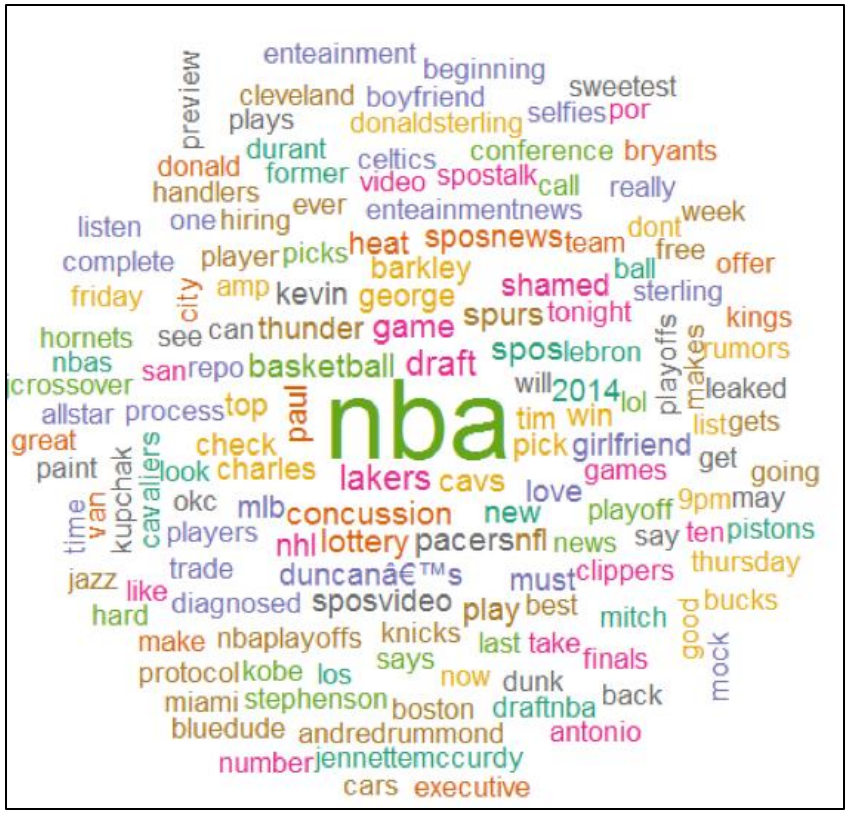


Figure 1.14: Word Cloud

The Business Analytics Framework

Definition 1.8: Social Media Sentiment Analysis

- Identifying **sentiment** of a customer on a specific product using social media or text mining

The Business Analytics Framework

Example 1.12: Sentiment Analysis Map

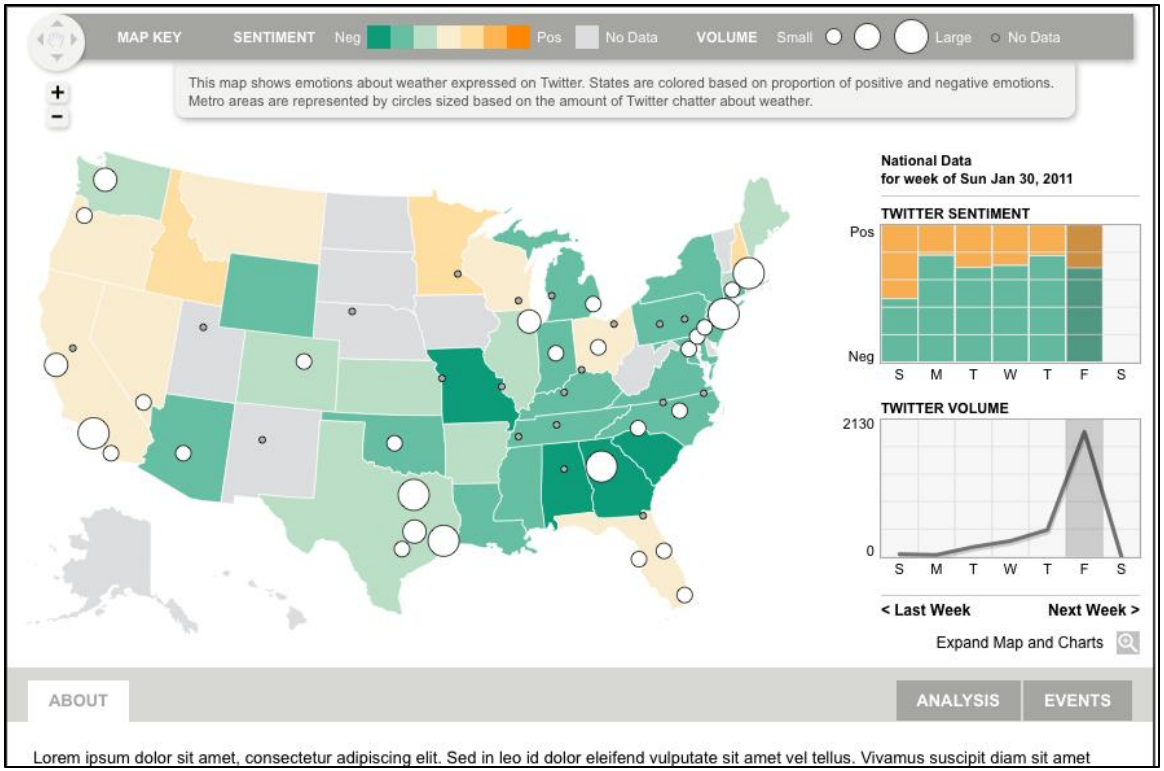
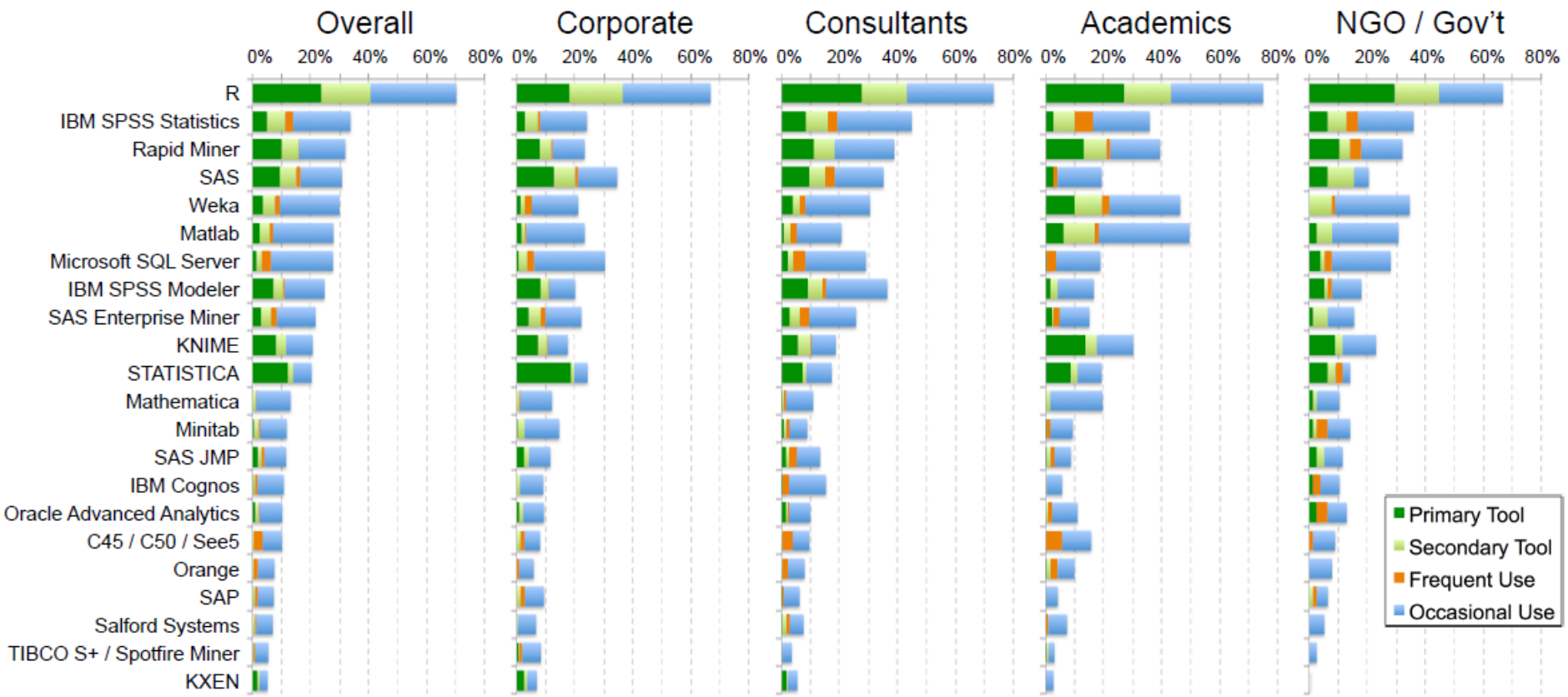


Figure 1.15: Sentiment Analysis

Outline for this Session

- Review of BA
- Introduction to Data Mining
- **Tools of Data Mining**
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Tools of Data Mining



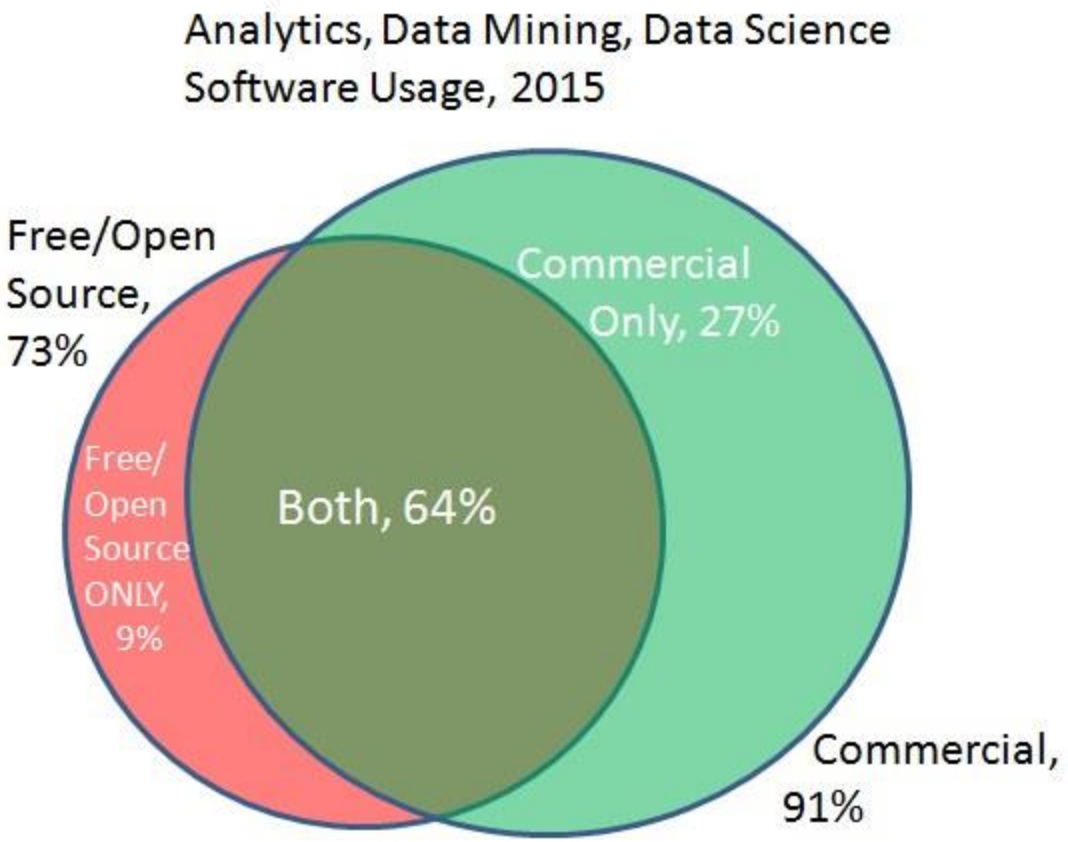
http://www.rexeranalytics.com/Data-Miner-Survey-2013-Intro.html

Tools of Data Mining

- **The top 10 tools by share of users were**
  - **R**, 46.9% share (38.5% in 2014)
  - **RapidMiner**, 31.5% (44.2% in 2014)
  - **SQL**, 30.9% (25.3% in 2014)
  - **Python**, 30.3% (19.5% in 2014)
  - **Excel**, 22.9% (25.8% in 2014)
  - **KNIME**, 20.0% (15.0% in 2014)
  - **Hadoop**, 18.4% (12.7% in 2014)
  - **Tableau**, 12.4% (9.1% in 2014)
  - **SAS**, 11.3 (10.9% in 2014)
  - **Spark**, 11.3% (2.6% in 2014)

http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Tools of Data Mining



http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- **CRISP-DM Framework**
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

CRISP-DM

Definition 1.9: CRISP-DM

- CRoss-Industry Standard Process for Data Mining
- Why Should There be a Standard Process?
  - The data mining process must be reliable and repeatable by people with little data mining background.
- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on experts

CRISP-DM

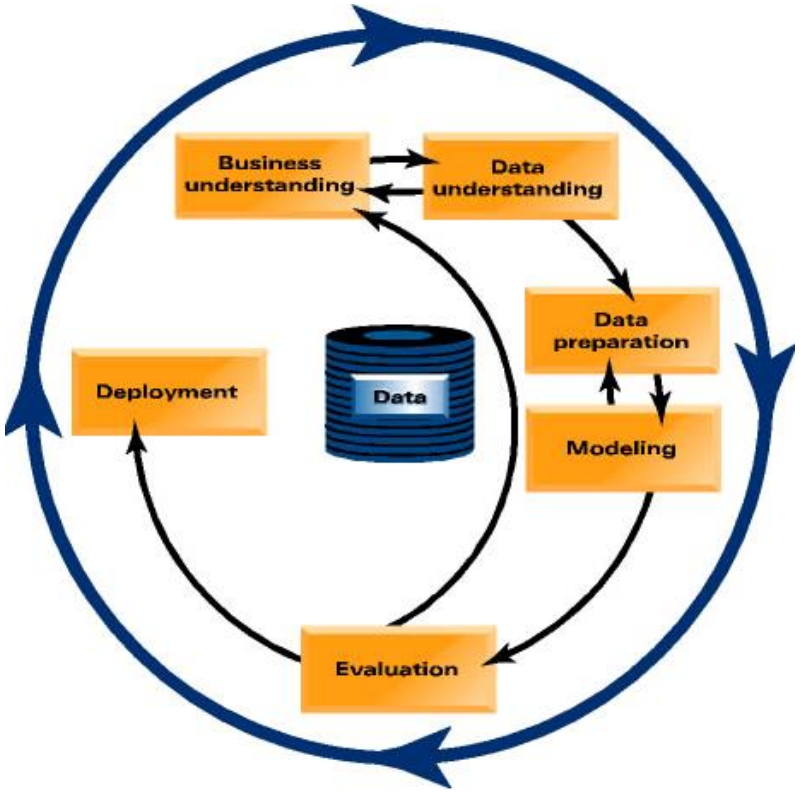
- Initiative launched in late 1996 by three "veterans" of data mining market.
  - Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR
- Developed and refined through series of workshops (from 1997-1999)
- Over 300 organization contributed to the process model
- Published CRISP-DM 1.0 (1999)
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, etc.
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, etc.

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

CRISP-DM

- Data Mining methodology
- Process Model
- For anyone
- Provides a complete blueprint
- Life cycle: 6 phases



CRISP-DM

- Phases
  - Business Understanding
    - Project objectives and requirements understanding, Data mining problem definition
  - Data Understanding
    - Initial data collection and familiarization, Data quality problems identification
  - Data Preparation
    - Table, record and attribute selection, Data transformation and cleaning

CRISP-DM

- Phases
  - Modeling
    - Modeling techniques selection and application, Parameters calibration
  - Evaluation
    - Business objectives & issues achievement evaluation
  - Deployment
    - Result model deployment, Repeatable data mining process implementation

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

Generic tasks (bold) and outputs (italic) of the CRISP-DM Model

CRISP-DM

Example 1.13: Hotline Decongestion

- Problem Statement: A leading telecommunications company is facing a congestion problem in their support hotline. Average waiting time for subscribers reaches 7 minutes.
- On average a call for specific transactions (Balance Inquiry) costs PhP 40. On the hand, the same transaction in other self care channels (e.g. Online) costs only PhP 5.

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - **Business Understanding**
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Phase 1. Business Understanding

- Consists of:
  - Statement of Business Objective
  - Statement of Data Mining Objective
  - Statement of Success Criteria
- Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

Phase 1. Business Understanding

- Determine business objectives
  - thoroughly understand, from a business perspective, what the client really wants to accomplish
  - uncover important factors, at the beginning, that can influence the outcome of the project
  - neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions
- Determine data mining objective
  - a business goal states objectives in business terminology
  - a data mining goal states project objectives in technical terms
- Produce the Project Plan

Example: Hotline Call Reduction

- Business Objective
  - To reduce the amount of calls in hotline and correctly recommend to subscribers to utilize less costly alternative channels to do transactions.
- Data Mining Objective
  - Analyze call inter-arrival time, frequency and type of the calls. Do a comprehensive profiling of the customers and to accurately predict if a certain customer is likely to call again. Determine frequent transactions sequences that subscribers historically transact.

Example: Hotline Call Reduction

- Project Plan
  - Get management approval for access of databases
  - Extract 1 Year's Worth of Data from the Hotline Call Databases
  - Extract Demographic Data about the customers that called the Hotline
  - Merge the datasets
  - Do a quality control on the data
  - Run prediction methodologies to profile customers that call
  - Run sequential analysis to determine sequence of transactions
  - Design a marketing campaign to target a specific customer profile to reduce calls and transfer them to other low cost channels
  - Implement the recommendation

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - **Data Understanding**
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

Phase 2. Data Understanding

- Consists of:
  - Data Exploration
  - Verify the Quality
  - Find Outliers
- Starts with an initial data collection and proceeds with activities in order to get familiar with the data,
- to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Example: Hotline Call Reduction

- Data Exploration/Data Profiling/Data Quality
  - Use a simple ETL to gather sample data from the source database
  - Simple aggregation showed that 95% of the subscribers transacted at most 10 times within the year while the remaining 5% called more than 10 times to at most 240 times with approximately 1.2 Million calls.
  - Analysis of the data showed that 25% of the calls, agents were not able to collect enough info or it was an uncompleted call
  - 2% of the calls where prank calls.

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - **Data Preparation**
  - Modeling
  - Evaluation
  - Deployment

Phase 3. Data Preparation

- Takes usually over 90% of the time
  - Collection
  - Assessment
  - Consolidation and Cleaning
  - Data selection
  - Transformations
- Covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and *not in any prescribed order*.
- Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Example: Hotline Call Reduction

- Data Preparation
  - Select all calls made last year
  - Merge demographic data
  - Remove Prank Calls, Incomplete Calls but Gather Statistics
  - Derive Inter-arrival Time (Time Between Calls)
  - Format calls in MM/DD/YYYY Format
  - Format time in HH:MM:SS

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - **Modeling**
  - Evaluation
  - Deployment

Phase 4. Modeling

- Select modeling technique
  - select the actual modeling technique that is to be used
  - ex) decision tree, neural network
  - if multiple techniques are applied, perform this task for each technique separately
- Generate test design
  - before actually building a model, generate a procedure or mechanism to test the model's quality and validity
  - ex) In classification, it is common to use error rates as quality measures for data mining models.
  - Therefore, typically separate the dataset into train and test set, build the model on the train set and estimate its quality on the separate test set

Phase 4. Modeling

- Build model
  - run the modeling tool on the prepared dataset to create one or more models
- Assess model
  - interprets the models according to his domain knowledge, the data mining success criteria and the desired test design
  - judges the success of the application of modeling and discovery techniques more technically
  - contacts business analysts and domain experts later in order to discuss the data mining results in the business context
  - only consider models whereas the evaluation phase also takes into account all other results that were produced in the course of the project

Example: Hotline Call Reduction

- Modeling
  - Sample data to get a training and test set
  - Utilize Decision Trees to Generate Rules for Profiling Customers using the training set
  - Utilize Validation techniques on the test set to determine accuracy of predictions
  - Perform sequential analysis to determine the sequence of call transactions made

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - **Evaluation**
  - Deployment

Phase 5. Evaluation

- Evaluate results
  - assesses the degree to which the model meets the business objectives
  - seeks to determine if there is some business reason why this model is deficient
  - test the model(s) on test applications in the real application if time and budget constraints permit
  - also assesses other data mining results generated
  - unveil additional challenges, information or hints for future directions

Phase 5. Evaluation

- Review process
  - do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked
  - review the quality assurance issues
  - ex) “Did we correctly build the model?”
- Determine next steps
  - decides how to proceed at this stage
  - decides whether to finish the project and move on to deployment if appropriate or whether to initiate further iterations or set up new data mining projects
  - include analyses of remaining resources and budget that influences the decisions

Example: Hotline Call Reduction

- Evaluation of Model
  - Model has a 65.4 % accuracy in terms of predicting whether a customer will call again
  - A total of 255 sequential rules where gathered and validated

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - **Deployment**

Phase 6. Deployment

- Determine how the results need to be utilized
- Who needs to use them?
- How often do they need to be used
- Deploy Data Mining results by
  - Scoring a database, utilizing results as business rules, interactive scoring on-line
  - The knowledge gained will need to be organized and presented in a way that the customer can use it. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Phase 6. Deployment

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

Example: Hotline Call Reduction

- Deployment
  - Volume of calls reduced by at most 25%.
  - Repeat analytics to continue

References

- Simon, Alan. CIS 391 PPT Slides
- Tan et al. Intro to Data Mining Notes
- Runger, G. IEE 520 notes
- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
- Chapman, P. et al. CRISP-DM 1.0 Reference Model
- http://www.ams.org/samplings/feature-column/fcarc-svd#sthash.pX2XSmkz.dpuf

Outline for this Session

- Definitions of Data Mining
- Examples of Data Mining
- Tools of Data Mining
- CRISP-DM Framework
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment