WHAT IS THE MAJOR POWER LINKING STATISTICS & DATA MINING ?

M.E. Abd El-Monsef[a], E. A. Rady[b],
A. M. Kozea[c], W. A. Hassanein[d], S. Abd El-Badie[e]

[a,c,d,e]Mathematics Department, Faculty of Science, Tanta University, Tanta, Egypt
[b]Institute of Statistical Studies & Research (ISSR), Cairo University, Cairo, Egypt

ABSTRACT

In the recent years, numerous scientific research studies which stand for the intersecting disciplines between statistics and data mining (DM) are obtained [17, 18, 19, 24, 27, 30, 35]. This paper is devoted to answer the titled suggested question which is based on five reply trends, the 1st trend based on an updated historical vision for each of statistics and DM. The 2nd trend is concerned with modern theoretical significant reply between statistics and DM. The major power linking statistics and DM is established in the 3rd trend. Lastly, the 4th trend represents a significant comparison between statistics & DM. A conceptual classification about Statistical Data Mining (SDM) process in Egypt will be represented in the 5th reply trend. Finally, the conclusion and the future work are represented.

KEYWORDS

Statistics, Data Mining, Significant, Power, History, Theoretic, Reply

1. INTRODUCTION

Statistics, Data Mining (DM) and Knowledge Discovery form a featured and appropriate group of experts to deal with the recent developments in data analysis techniques for DM and knowledge extraction [17]. This awareness group provides a practical, multidisciplinary approach on using statistical techniques in various areas such as Business, Economics, Stock Market, Communications and Medical Diagnosis.

It can be observed that there is mutual ignorance between statisticians and data miners, Ganesh; S. (2002) discusses this point in details [13]. Actually, the statistician and data miners' analysis trend has the same manner. The most recent studies of statistical data mining introduced an ideal discussion of the historical and theoretical background for statistical analysis and DM and integrate them with the data discovery and data preparation operations. [30]

The sequence of the paper organized as follows. Section 1 presents the 1st direction of the answer of the titled question subtitled into two directions Statistics history and DM history reply. Section 2 presents the 2nd reply trend depending on a theoretical reply between statistics & DM. Section 3 discusses the 3rd trend of the paper which is the linking power between statistics & DM. Section 4 investigates a significant comparison between statistics and DM. Section 5 focuses on Statistics and DM in Egypt which presents the 5th reply trend. The conclusion and future work introduced in section 6.

DOI : 10.5121/ijdkp.2013.3609

1st Trend: Historical Reply

In this section a historical view for both of statistics and DM will be represented as follows.

1.1 Statistics History Reply

“Statistics are human beings with the tears wiped off.” Paul Brodeur. Statistics is the process for converting the dust to gold, which make any problem easier. The statistics name history has many stages to have “Statistics” name across different civilizations [1, 35, 36,]. Figure (1) presents the hierarchal stages for Statistics name across civilizations.

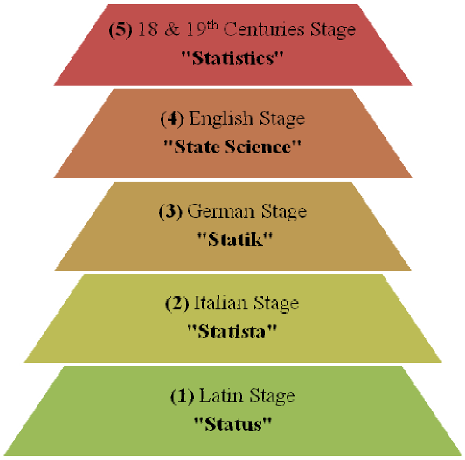

Figure 1: Hierarchal Statistics Name Stages across Civilizations

By the 18th century, the systematic collection of demographic and economic data by states was the "statistics” meaning. In the early 19th century, the statistics implication becomes more inclusive to have many operations, collection, summary, and analysis of data [38].

1.1.1 Statistics Definition

Across this long history of the statistics science, it takes a lot of definitions from different perspectives of views and the following some of these definitions, as some of them find its meaning in the tables and numbers on the life and events , and the others see its meaning through their newspapers and magazines from a variety of data .The truth is that such a view on the concept of Statistics and means deficient in perception , the fact that the statistical concept meaning has a greater view than this , especially if we know that the Statistics is as old as human beings , recalling the date that the ancient Egyptians had used Statistics in the majority of their activities such as building the pyramids, the rest of civilizations of the other countries aren’t different from its predecessor in the use of Statistics approach as a tool to count , census. Consequently, Statistics became word synonymous with the work of the state, it is in some sense the process of collecting the data, and facts relating to the affairs of the state called the knowledge of the count, or knowledge of the media, or the science of large numbers.

Then the statistical concept have grown to become nowadays the science which depends on formulas, mathematical laws and quantity; the statistical methods become important pillar in the way of scientific research, help researcher in the development of plans and designs for his research or experience to be able to eventually achieve results which it seeks. As well as the statistical work is the best in finding solutions and achieving goals.

Overall statistics has become one of the branches of Pure Applied Mathematics, its own rules, laws, symbols, terminology and theories which make statistics uses the numbers to analyze the qualities and phenomena as reflected in the data to be examined and has what sets it apart from other sciences in the methods and techniques.

Def.$_1$ "The art and science which examines the principles and methods implemented in collecting, presenting, analyzing and interpreting the numerical data on a research field" [38]

Def.$_2$ The Branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability).It is subdivided into descriptive statistics and inferential statistics. [31]

Def.$_3$ The science of kings, political and science of statecraft" The kings and rulers in the ancient times. [28]

Def.$_4$ The most important science in the whole world: for upon it depends the practical application of every other science and every art: the one science essential to all political and social administration, all education, and all organization based on experience, for it only gives results of our experience" Florence Nightingale.[11]

Def.$_5$ The science of counting. [28]

Def.$_6$ The Science of averages. [28]

Def.$_7$ The Science of estimate and probabilities. [28]

Def.$_8$ The method of judging collection, natural or social phenomena from the results obtained from the analysis or enumeration or collection of estimates. [28]

Def.$_9$ The numerical statement of facts capable of analysis and interpretation and the science of statistics is the study of the principles and the methods applied in collecting, presenting, analysis and interpreting the numerical data in any field of inquiry. [28]

1.1.2 Statistics History across Centuries (16 -20)

Statistics science has a great historical evolution, from the latest of the sixteenth century followed by the beginning of the seventeenth century. Statistics simply means counting which is an old idea standing to the history of civil humanity, the need to obtain digital information or descriptive for communities and circumstances and material conditions of their existence was an urgent need since found organized human societies, the ancient Egyptians, Chinese and Greeks have some statistics belong to their communities in terms of population and the amount of agricultural and mineral wealth gathered to guide in the conduct of the affairs of state and policy-making. We should not lose sight of what is stated in the Quran mention of the word count as a sign to the idea of counting and is the oldest inventory of several centuries.

Figure (2) represents the growth of the statistical history across the centuries from the 16-century to 20-century according to the appearance of the statistical scientists during this period. Starting from *Sir W. Petty (1532)* to *B. Efron (1979)*. It is clear that there were a huge scientific jump between centuries especially in the 20-century.

It is known that there are the beginnings of well known in the field of possibilities have emerged in the sixteenth century where Cardano (1501-1571) presented some of the ideas in the odds associated with throwing dice table. Then a development work in the field of probability and statistical methods appeared theoretical and practical dimensions. The letters and discussions that were taking place between Pascal (1623-1662) and Fermat an indication of the emergence of the assets of the odds when some issues associated with games of luck. Pascal had made in 1665 founded expectation and discussed the issue of bankruptcy bears.

However, the theoretical and the mathematical sense of the statistical dimension have a great jump in the eighteenth century and spread to the first third of the twentieth century. At the start, it was not in the development theories of probability and statistical methods, but in response to the practical needs of the real issues in science and society. In general, statistical methods were developed to suit the analytical work in the field of science. As well as Laplace (1749-1827) established the concept of general application of statistical methods in general and proved that probability theory approach is necessary to improve all kinds of human knowledge, Quetelet (1796-1874) an astronomer and statistical learn something about the logical scientific possibilities. Moreover, the work of both of Galton (1857-1936) and Pearon (1822-1911) for applications in the fields of genetics and life sciences. Then it was developed by Fisher (1890-1962) in the fields of genetics and agricultural field trials included in this framework. In addition, the work of those on the application of statistical methods in these areas led them to develop new statistical methods.

When talking about statistics evolution in 20th & 21st centuries we have to refer to the great effect of the computer. While, Statistical tables and tables of random numbers first became much easier to produce and then they disappeared as their function was subsumed into statistical packages. Huge data sets could be assembled and analyzed. In-depth the necessity of using DM in many problems helps the statisticians to take the suitable decision. Models that are largely more complex and methods could be used. Methods have designed with computer implementation in mind, like the family of generalized linear models linked to the program GLIM. Monte Carlo methods have used directly in data analysis. In classical statistical inference, the bootstrap has been very prominent. In Bayesian analysis Markov Chain Monte-Carlo methods have been used extensively; previously conjugate priors and non-informative priors had been used because of computational limitations.

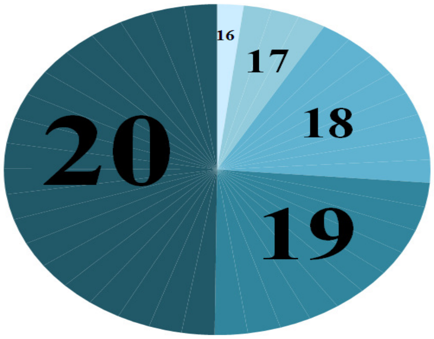

Figure 2: Statistical Growth History across Centuries (16-20)

1.1.3 What about the 21- Century?

A huge statistical evolution occurred at the end of the 20-century to the early 21-century to utilize statistics with the computer technology by obtaining much statistical software (SPSS, SAS, Minitab, Excel, STATA, STATISTICA ...).

The wide variety of data collection methods in the 21st century caused many complex problems which are considered as a big challenge to achieve the greatest benefit from this data [40, 26]. As the volumes of many commercial, industrial, and scientific datasets have exceeded the terabyte range and are approaching petabytes, exabyte, zettabyte and yottabyte. Statistical techniques have long been employed to find the decision core in any type data. [15]

Many 21st century opportunities and challenges for statistical analysis lie in the effective management and compression of massive datasets, motivation and justification of DM algorithms, support of the transition from data exploration to data and result explanation, and evaluation of DM results against reality. In addition, statistical analysis may well be useful in creating value from DM results by yielding new insights, motivating decisions, and justifying actions.

The 18th of November of each year is the **African Statistics Day (ASD) which is** initiated in 1990 by the Subsidiary Body of the United Nations Economic Commission for Africa **(UNECA),** ASD is a great opportunity to emphasize the realization of statistics in the daily life worldwide.
The United Nations Statistical Commission **(UNSC)** declared a special day to celebrate by statistics which is called **World Statistics Day (WSD).** This Day was celebrated for the first time on Wed, 20th October 2010 (20-10-2010) worldwide. [38]

To highlight the important role of Statistics in our daily life, the year 2013 is determined to be the International Year of Statistics by the American Statistical Association (ASA). All the continents all over the world will celebrate this year to be the International Year of Statistics.

1.2 DM History Reply

Starting 1989 till now, a novel scientific direction to analyze the data is obtained which is called DM. DM is a combination of computational and statistical techniques to perform exploratory data analysis (EDA) on rather large and mostly not very well cleaned data sets (or data bases). DM history started nearly from 40 years ago but it was not called that then. SAS and SPSS companies were the 1st to promote DM as statistical analysis.

For 21-century [15, 16], the problem isn't accessing data but ignoring irrelevant data. Most modern problems can electronically deal with the cumulative data from many years ago [37]. This leads to a requirement for training data miners in statistics or statistics graduates in data mining.

Web Mining and text mining are the most recent advances directions for DM process. Applying DM to these data adds a great depth to the patterns already uncovered through DM process. [3, 33]

1.2.1 DM Definitions History

Although the short history of DM, it takes many definitions from many sources and the reason for this that the DM depends on different scientific directions during its process, the following are some of these definitions,
Def.$_1$ The analysis of (often large) observational (as opposed to experimental) data sets to find unsuspected relationships and to summarize the data in novel ways which are both understandable and useful to the data owner. "Appears in 12 books from 2001-2006"

Def.$_2$ The extraction of hidden predictive information from large databases. [25]

Def.$_3$ The process of analyzing data from different perspectives and summarizing it into useful information within a particular context. [4]

Def.$_4$ The process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. [2]

Def.$_5$ Finding interesting structure (patterns, statistical models, relationships) in databases. [39]

Def.$_6$ The application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets. (Insightful Miner 3.0 User Guide)

Def.$_7$ The process of semi-automatically analyzing large databases to find patterns which are:

- valid: hold on new data with some certainty
- novel: non-obvious to the system
- useful: should be possible to act on the item
- understandable: humans should be able to interpret the pattern[25]

Def.$_8$ The process of knowledge discovery in databases (KDD). [41]

1.2.2. DM Names Sequence History

Although, DM has been appeared with a short history, DM has many names; the following figure represents DM names sequence history. Statisticians have used some terms like “Data Fishing”, “Data Dredging” or “Data snooping” for some times. These names are used to refer to what they considered a bad practice of analyzing data without an a priori hypothesis.

The database community used DM term in 1990. Briefly, there was a phrase “database mining”, then the researchers turned it into “data mining”.

“Knowledge Discovery in Databases” term was used for the 1st time by Gregory Piatetsky-Shapiro in 1989 and this term became more popular in AI and Machine Learning Community. However, the term DM became more accepted in business community and in the press.

Currently, DM and Knowledge Discovery are used interchangeably, and we use these terms as synonyms. [7]

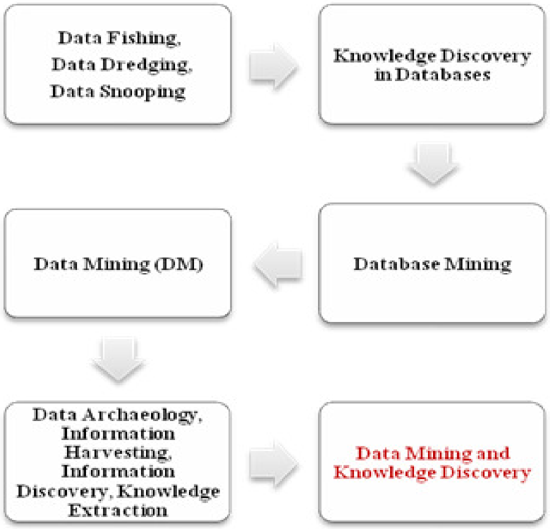
Data Fishing,
Data Dredging,
Data Snooping
Knowledge Discovery in Databases
Database Mining
Data Mining (DM)
Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction
Data Mining and Knowledge Discovery

Figure 3: DM Names Sequence History

1.2.3 DM Dependency Table History

SAS and SPSS companies were the 1st to promote DM as statistical analysis in the early 1960s. By the late 1980s, the traditional techniques had been augmented by new methods such as fuzzy logic, rough set, heuristics and neural networks. [30]

Applying DM techniques in the industry field started from the 1990s. DM dependency table history can be described by three basic scientific techniques, the 1st is the classical statistics but during DM process, some problems in the complex business requirements area come into view. So, the 2nd technique on the DM table which is AI deals with this type of problems, but some commercial problems faced the analysts using this derivation. So, the analysts need to use machine learning (3rd technique) which is more accurately described as the union of advanced statistics and AI and this short history can be summarized in Figure (4).

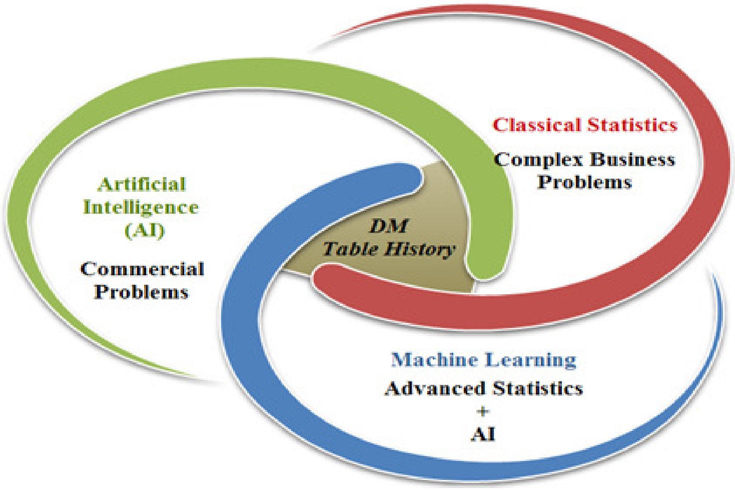

Figure 4: DM Dependency Table History

Most analysts separate data mining process into two groups: DM tools and DM applications. DM tools can solve any business problem. In DM applications there is a very large sets of data that are collected through the use of some non-automated controlled methods completely; These data represent the analysis process income, and therefore can contain data values that fall outside the areas of knowledge or unreasonably compilation of some of the values do not agree terms.; and you should pay attention to the analysis of data that have not been collected or carefully selected, can lead to misleading results specifically in the predictive DM.

The choice of the appropriate DM techniques depends on the nature of the data under study and on the data size. The idea is that DM process extracts the hidden patterns which were not noticeable before. Such as in the medical treatment process the computer helps us how to provide accurate information extracted for specialists in the field of medicine and cancer treatment with high efficiency and the least damage to the patient.

2. 2nd TREND : THEORETICAL REPLY

The 2nd reply trend of this paper is concerned with modern theoretical significant reply between Statistics and DM will be obtained in this section. Starting the idea of this section by the statement said by *Z.-H. Zhou* “It is still clear that without the solid theoretical foundation donated by the Statistics community, DM will be building a castle in the air" and this is true that DM without Statistics cannot be a real science.

Figure (4) represents a brief description for the DM cycle which represents the DM as a stage of the KDD process.

The theoretical relation between Statistics and DM will be summarized in the following points,

1) Methodology Similarity: This Concept leads that most statisticians to consider DM as one of the Statistics branches. So, most of the DM softwares now were invented by statisticians.

2) Statistics Provides the Theoretical Basis of DM Process: The previous studies of DM process focus on the statistical prospective as a measure of the DM validity.

3) DM Process used many statistical algorithms such as Cluster Analysis, Bayes Networks and Regression.

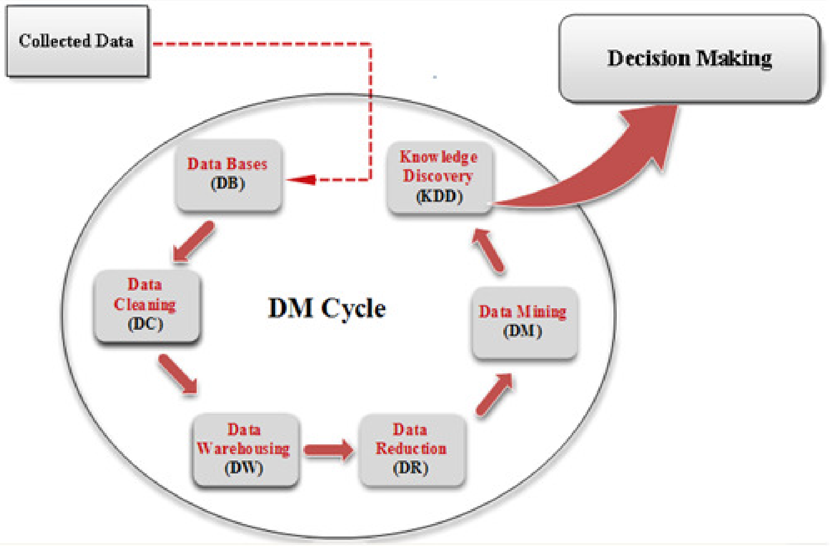

Figure 5: DM Cycle (Brief Description)

3. 3rd TREND : STATISTICS & DM LINKING POWER

The connection strength between Statistics and DM has a very physically powerful in the recent years especially that Statistics is considered the validity power part in the process of DM [5, 6, 8, 9, 10, 12]. Using statistical analysis techniques has a great impact on DM process as follows,

1) Data Preparation: beginning with data preprocessing step, passing through data cleaning, data integration and transformation, data reduction, mining frequent patterns, association, and correlation until prediction and modeling [24]. All of these steps aren't independent of each other and are common in many techniques like the outliers techniques are used in the data description and in the data cleaning.

2) Find Patterns: Applying data analysis algorithms and find the patterns from them.

3) Pattern Evaluation: Measuring the error, accuracy percentage in the output patterns, the pattern is considered as knowledge if it has a specific percentage of accuracy.

Comparing Statistics Def.$_2$ in Section (1.1) by the above DM steps, we will not find a big difference between the aims of the two processes because both of Statistics and DM are interested in the data collection and its analysis. This Similarity between them make every stage in the DM process derives its theoretical and technical concepts from its counterparts in Statistics.

There were many Statistics techniques which are included in the DM process [14], these techniques can be summarized in Table (1) which is considered the linking power between Statistics & DM. [23] Data preparation stage is considered the most important and the longer time stage in DM process.

The connection analysis between Statistics and DM can be described by statistical DM (SDM) process which divided into two directions according to the researcher demand as in Figure (7),

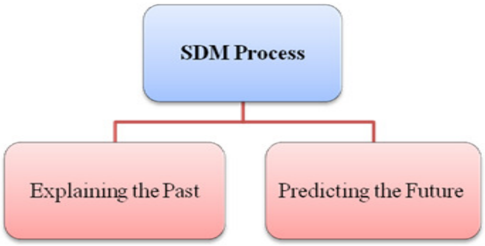

Figure 7: Statistical DM Process (SDM)

4. 4th TREND : STATISTICS & DM SIGNIFICANT COMPARISON

Although, data miners and statisticians use similar techniques to solve similar problems, but the DM approach differs from the standard statistical approach in several areas [18, 16], and Table (2) clarify these differences.

5. 5th TREND : STATISTICS AND DM IN EGYPT

In this section as *a* conceptual classification *about the statistical DM(SDM) process in Egypt* will be represented in Figure (8), we divide the sectors in Egypt into two types: Governmental Sectors and Private Sectors.

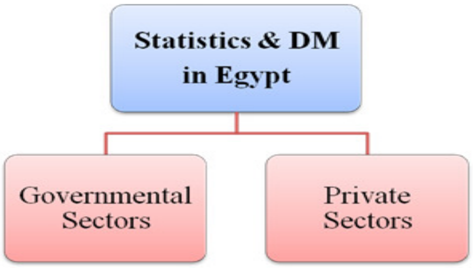

Figure 8: SDM in Egypt

Governmental sectors in Egypt can be represented as follows,

1. **Central Agency for Public Mobilization and Statistics** (**CAPMAS**), the official statistical organization in Egypt that makes all statistical analysis and the Census.

2. **The Information and Decision Support Center (IDSC),** Its mission is to support the government decisions through advice on best policy scenario mix and analytical research to improve the socio-economic well-being of the Egyptian society.

3. **Institute of Statistical Studies and Research (ISSR),** It is an educational institute and also it has a center for analyzing the data statistically.

4. **The Cairo Demographic Center (CDC),** An educational institute that nurtures a new generation of specialists in demography in the developing world, who are concerned with the study and analysis of critical population issues. It fosters innovative interdisciplinary approaches to population studies and helps policy- makers design and implement appropriate population and development policies.

5. **Information and Communication Technology (ICT) Indicators Project,** This project provides the necessary, accurate and meaningful data about ICT sector in Egypt.

6. **Cairo University Theses Mining System (CUTMS)** Cairo University**,** is in possession of large volumes of graduated students theses data (master and doctor theses) to which they are looking to add value in various ways and it expresses a vast, rich range of information. Cairo University Theses Mining System **(CUTMS)** targets two main challenges predominant for the theses mining system; the first challenge is at Cairo University Governance level, and the second one is at researcher level.

7. **Data Mining and Computer Modeling (DMCM)** Center, DMCM is a virtual center that creates Data Mining and Computer Modeling techniques, theories and products with innovation, partnerships and collaboration at its core strategies.
8. **Centers of Excellence Program (CoEP)**, CoEP supported the Center of Excellence in Data Mining and Computer Modeling (DMCM) during the period from 2008 to 2010.

While Some Private Sectors in Egypt can be represented as follows,

1. **Vodafone Company:** The largest mobile phone company in Egypt in terms of active subscribers.

2. **Mobinil Company:** The Egyptian Company for Mobile Services which is one of Egypt's three mobile phone operators.

3. **Etisalat Company:** Etisalat Egypt is one of 15 service providers managed by Etisalat in the Middle East, Asia and Africa. Etisalat group currently has access to a potential market of just below 1 billion subscribers and today Etisalat services over 130 million subscribers including the total number of fixed-line, Internet, mobile and television from each of its subsidiaries.

4. **Orange Company:** is a French multinational telecommunications corporation and represents the flagship brand of the France Telecom group. It is a global provider for phone, landline, Internet, mobile internet, and IP television services

5. **Mo’men Restaurants:** is a chain of fast food restaurants based in Cairo, Egypt, specializing in sandwiches. Mo'men is a wholly owned subsidiary of the Mo'men Group.

After a lot of discussions with different employees in the above companies, we got that, using Statistics and DM especially in the governmental sectors in Egypt still needs more improvement to use it to predict the future not just for presenting the past. In addition, some of the private sectors are making their DM process using only statistical techniques for their analysis of data.

6. CONCLUSION & FUTURE WORK

The answer of the suggested titled question, what is the major power linking Statistics & DM? Of this paper gives us some new significant spots on the relation between statistics and DM based on the historical, the theoretical reply and the linking power between them, also, many novel directions to be as a future work. These points can be summarized as follows; the old history of the statistics science gives the conclusion that statistics has the priority to be the major power of the DM process especially after determining this year to be the International Statistics Year in WSD2013 by ASA. From the other hand, the short history of DM shows that there are many statistical techniques and measures which aren't used till now in DM process especially nonparametric and semi parametric statistical tests which it will be as a future work. From the other hand using Statistics and DM especially in the governmental sectors in Egypt still needs more improvement to use it to predict the future not just for presenting the past. And some of the private sectors in Egypt are making their DM process using only statistical techniques for their analysis of data. So, this gives us motive forces to our future work to be a project for implementing of the current and real situation of using SDM analysis of in Egypt Also, as an upcoming work, Rough Set Theory (RST) and its generalized models will be used in modeling the connection between statistics and DM.

Table (1): Statistics & DM Linking Power

	Statistical Technique		Description
1	**Descriptive Statistics**		- Central Tendency - Dispersion - Graphical Display
2	**Missing Values** **Noisy Data** **Outlier Analysis**	}	Data Cleaning
3	**Regression** - **Linear** - **Logistic** - **Robust** **Correlation Analysis** - Cascade Correlation - Pearson Correlation - Spearman Correlation	}	- Prediction - Modeling - Association
4	**Probability Theory** **Distributions Theory**	}	Prediction of the behavior of defined systems.
5	**Bayesian Classification**		Bayes' theorem and Naïve Bayesian.
6	**Estimation Theory**		- Model selection - Estimating confidence Intervals - Roc curves
7	**Analysis of Variance ANOVA**		Tests the equality of two group means or not
8	**Factor Analysis (FA)**		Reduce the variables by combining them to generate some factors

9	**Discriminate Analysis (DA)**		Predict a categorical response variable.
10	**Time Series Analysis** - Auto-Regression Methods - Univariate ARIMA Modeling - Long-Memory Time-Series Modeling.	}	Time Series Analysis
11	**Quality Control** - Shewhart Charts - Cusum Charts		Display group summary statistics
12	**Principle Components Analysis (PCA)**	}	Data Reduction
	Canonical Correlation Analysis		
	Cluster Analysis (CA) - **Hierarchal** - **Partitioning** - **Density Based** - **Model Based**		
	Sampling - **Simple random sample without replacement (SRSWOR)** - **Simple random sample with replacement (SRSWR)**		
13	**"Probably Approximately Correct" PAC Learning**		Determining how much data needed for a given classifier to achieve a given probability of correct predictions on a given fraction of future test data.
14	**Spatial Statistics**		- Analyzing spatial data - Exploring geographic information.
15	**Spatial Surveillance**		- Discusses scan statistics - Discovering over densities of disease cases.

Table (2): Statistics & DM Highlight Points

Issue	Statistics	Data Mining (DM)
Data Number	Hundreds to thousands	Millions or billions
Data Type	Experimental	Observational
Sampling	Yes (Statistical Reasoning)	Sometimes
Hypotheses	Yes (conceptual Model)	No
Experiment Based	Collect data to answer a specific question (Validate of hypotheses)	Secondary data analysis (Mining Algorithm based on interestingness)
Analysis Type	Hypotheses Types 1. Null Hypothesis. 2. Alternative Hypothesis	Interesting Types **1.** Frequency **2.** Rarity **3.** Correlation **4.** Length of occurrence (for sequence of temporal data) **5.** Consistency **6.** Repeating / Periodicity **7.** “Abnormal” Behavior **8.** Other Patterns

Observation No.	Depend on the desired model and the power of the test.	No Restriction
Characteristics	- Consolidation of realities. - Expressed in numerical values. - Affect Statistics by scaling the variables up. - Ordered by accuracy with a realistic standard and estimated. - Is compiled for a predetermined purpose. - Is compiled in a systematic way. - Should be comparable. - Deals with grouping and consolidation. - Efficient & Correct Statistical Analysis - Collection of appropriate numeric data - Complicated data clarification through tables, diagrams and graphics. - Understanding the structure and changing patterns of a fact through quantitative observations. - Enables correct inference on a certain reliability level, on the variables of the population through sampling.	- Explain or categorize some particular objective - Find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. - Classification: predicting an item class - Clustering: finding clusters in data - Associations: e.g. A & B & C occur frequently - Visualization: to facilitate human discovery - Summarization: describing a group - Deviation Detection: finding changes - Estimation: predicting a continuous value - Link Analysis: finding relationships - Data representation in transactional databases for data mining - Data reduction - Data transformation - Data cleaning - Data sparsity - Data rarity
Boundaries	- Increasing no. of variables leads to the curse of dimensionality. - Cannot be applied to heterogeneous data. - Single observations are not statistics.	- Focus on training Rely on one technique - Ask the wrong question - Listen (only) to the data - Accept leads from the future - Discount peaky cases - Extrapolate Answer every Inquiry - Sample casually - Believe the best model

Applications	- Actuarial science - Biostatistics - Business Analytics - Chemo metrics - Demography - Econometrics - Environmental statistics - Epidemiology - Geo statistics - Operations research - Population ecology - Quantitative psychology - Psychometrics - Quality control - Statistical finance - Statistical mechanics - Statistical physics - Statistical thermodynamics	- Banking: loan/credit card approval - Customer relationship - Management - Targeted marketing - Fraud detection - Telecommunications - Financial transactions - Manufacturing and production - Medicine - Technical data analysis - Network Analysis - Educational data mining - Quantitative structure-activity relationship - Surveillance / Mass surveillance - National Security Agency - DM in Agriculture - DM in Meteorology

REFERENCES

[1] Anders Hald, A history of probability and statistics and their applications before 1750, Wiley-IEEE, ISBN 0471471291, (2003).

[2] Berry, J.A.M., and Linoff, G., Data mining techniques-for marketing, sales and customer support", New York, Wiley, (1997).

[3] Berry, M.J.A. and Linoff, G.S., Mastering Data Mining -The Art and Science of Customer Relationship Management, New York, (2000).

[4] Bill Palace, Data Mining: What is Data Mining?, (1996), http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[5] Chatfield C., Data Mining, Royal statistical society news, Vol. 25, (1997), 1-2.

[6] Clark Glymour Et. Al, Statistical Themes and lessons for Data Mining, Data Mining and Knowledge Discovery,Vol. l, Kluwer Academic publishers, (1997), 11- 28.

[7] Data Mining Community's Top Resource , Data Mining, and Knowledge Discovery: An Introduction, (2011),http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html

[8] David J. Hand, Statistics and Data Mining: Intersecting Disciplines, copyright @ ACM SIGKDD, Vol. 1, Issue 1, (1999).

[9] David J. HAND, Data Mining: Statistics and More? The American Statistician, Vol. 52, No. 2. , (1998).

[10] Emanuel Parzen, Data Mining, Statistical Methods Mining and History of Statistics, Interface Symposium on Computing Science and Statistics, Proceedings, ed. D. Scott., (1998).

[11] Florence Nightingale, Statistics,(2011), http://jwilson.coe.uga.edu/emt668/EMAT6680.Folders/Brooks/6690stuff/Statistics/Statistics.htm

[12] Friedman J.H, Data mining and Statistics-What's the Connection, 29th Symposium on the interface, (1998).

[13] Ganesh, S., Data mining: Should it be Included in The Statistics Curriculum? The 6th international conference on teaching statistics (ICOTS 6), Cape Town, South Africa, (2002).

[14] Glymour Et al., Statistical Inference and Data Mining, Communications of the ACM, Vol. 39, No. 11, (1996).

[15] Goodman A. , Kamath C. And Kumar V., Data Analysis in the Twenty-First Century, Vol. 1, No. 1, Journal Volume: 1; Journal Issue: 1, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, (2008), 1-3.

[16] Gorunescu F., Data Mining Concepts, Models and Techniques, Vol. 12, intelligent systems reference library, Springer, (2011).

[17] Hamparsum Bozdogan Et al, Statistical Data Mining and Knowledge Discovery. 2nd edition, London, (2004).

[18] Hand, D.J., Data mining-statistics and more?" American Statistician, Vol. 52, (1998), 112-118.

[19] Hastie, T., Tibshirani R. and Friedman J. H., Elements of statistical learning-data mining inference and prediction, Springer Verlag, New York, (2001).

[20] I. Krishna Murthy, Data Mining- Statistics Applications: A Key to Managerial Decision Making" article, indiastat.com, (2010)

[21] J. Hosking, E. Pednault, and M. Sudan, A Statistical Perspective on Data Mining, Future Generation Computing Systems, special issue on Data Mining, (1997).

[22] Jure Leskovec, Data Mining : Introduction, (2010), http://www.stanford.edu/class/cs345a/slides/01-intro.pdf

[23] Klamber M. and Han J."Data Mining: Concepts and Techniques, 2nd Edition, Elsevier Inc., USA, (2006).

[24] Kuonen, D., Data mining and Statistics: What is the connection? The Data Administrative Newsletter, Switzerland, (2004).

[25] Kurt Thearling, An Introduction to Data Mining, (2010),
http://www.thearling.com/text/dmwhite/dmwhite.htm

[26] Lomax, R. G., An Introduction to Statistical Concepts for Education and Behavioral Sciences (2nd ed.). New York: Routledge, (2007).

[27]Lovleen Kumar Grover and Rajni Mehra, The Lure of Statistics in Data Mining, Journal of Statistics Education Volume 16, Number 1, (2008),www.amstat.org/publications/jse/v16n1/grover.html/

[28] Math Zone, Definition of Statistics, (2011),
http://www.emathzone.com/tutorials/basic-statistics/definition-of-statistics.html

[29] Neal Leavitt, Data Mining Corroborate Masses, (2011), http://www.leavcom.com/ieee_may02.htm

[30] Robert Nisbet Et al, The Handbook of Statistical Analysis and Data Mining Applicants, Academic Press, ISBN: 0123747651, (2009), www.elsevierdirect.com/datamining

[31] SAS Analytics, Statistics Definitions, http://www.businessdictionary.com/definition/statistics.html

[32] Siva Ganesh, Data Mining: Should It Be Included In The 'Statistics' Curriculum? ICOTS6, (2002).

[33] SPSS Inc., SPSS Data Mining Tips, ISBN 1-56827-282-0 Printed in the U.S.A., (2005).

[34] STASTICA Data Analysis Software and Services, Stat Soft Electronic Statistics Textbook, Data Mining Techniques, (2011), http://www.statsoft.com/textbook/data-mining-techniques/

[35] Stephen M. Stigler, Statistics on the table: the history of statistical concepts and methods, Cambridge, Mass: Harvard University Press, (2002).

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

[37] Tim Menzies and Ying Hu, Computing Practices Data Mining for Very Busy People, (2009), http://biblioteca.universia.net/html_bura/ficha/params/title/computing-practices-data-mining-for-very-busy-people/id/47808919.html

[38] Turkish Statistical Institute, World Statistics Day 2010, (2010), http://www.turkstat.gov.tr/digEn/istTarihi.html

[39] U. Fayyad, S. Chaudhuri and P. Bradley, Data mining and its rule in database systems, Proceeding of 26th VLDB Conference. Cairo, Egypt, Morgan Kaufmanu, (2000), 63 – 124.

[40] Wiley Inter Science, Data Analysis in the 21st Century, (2007), www.interscience.wiley.com/.

[41] Wikipedia, Data Mining, (2011), http://en.wikipedia.org/wiki/Data_mining

Tables:

Figures: