Dublin Business School

excellence through learning

Student Number	**:**	**10399137**
Course Title	**:**	**M.Sc. Data Analytics**
Lecturer's Name	**:**	**Terri Hoare**
Student Name	**:**	**Poonam Dhoot**
Module/Subject Title	**:**	**Data Mining**
Assignment Title	**:**	**Big Data Mining Process and Application**

Date of Submission: 04/03/2019

PART A –

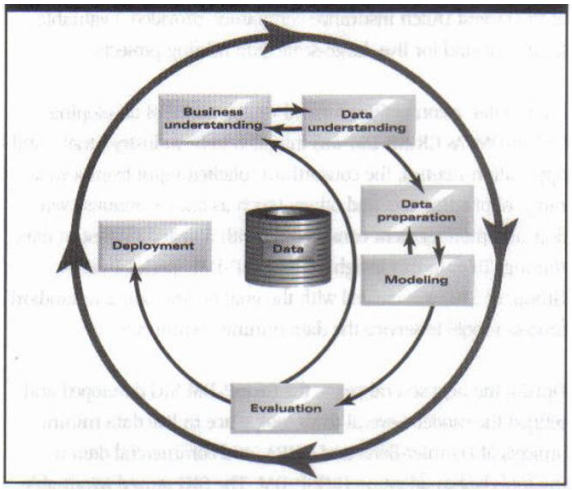
Read the journal article available on Moodle "The CRISP-DM Model: The New Blueprint for Data Mining" Shearer 2000. Write a critique of this article as it applies to the mining of 'Big Data' in 2019. Your appraisal should include a review of a minimum of 3 related journal articles, all published on or after 2012, and be no longer than 1500 words.

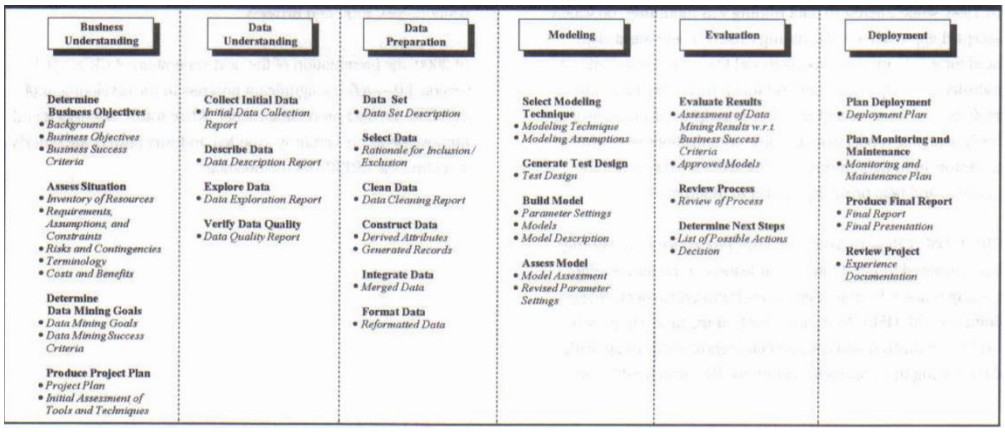
Solution:

"The CRISP-DM Model: The New Blueprint for Data Mining", Shearer (2000) describes the 6 phases as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

There are important and frequent dependency between the phases which is indicated by arrows in the figure below. Each phase has been divided into multiple steps for better understanding and elaboration about what needs to be done in the phase.





CRISP-DM was conceived by 4 leaders of nascent data mining market:

1. Daimler-Benz(DaimlerChrysler)
2. Integral Solution Ltd. (ISL)
3. NCR
4. OHRA

The following is the types of data mining techniques used for solving a business problem:

1. Data Description and Summarization
2. Segmentation
3. Concept Descriptions
4. Classification
5. Prediction
6. Dependency Analysis

Specifically, the study sought to determine:
(a) to what extent does CRISP-DM applies to Big data in 2019
(b) the improvement made in CRISP DM over the period
(c) various methodology which are built or used apart for Big Data mining
(d) the pros and cons of CRISP - DM

CRISP-DM was well built in 1996, in real-world environment hence it was well suited. But over the years the data paradigm changed, big data came into picture. Big Data - So far, there is no accepted definition for big data. However, the authors are going to use the definition proposed by Jin et al. Big data refers to a "*bond that connects and integrates the physical world, the human society, and cyberspace in a subtle way*" and can be classified into two categories, concretely,
data from the **physical** world (e.g. sensors, scientific experiments and observations) and data from the **human society**, which (e.g. social networks, Internet, health, finance, economics and transportation)[3]

By doing a comparison of the KDD and SEMMA stages we would, on a first approach, affirm that they are equivalent:

- Sample can be identified with Selection,
- Explore can be identified with Pre-processing
- Modify can be identified with Transformation
- Model can be identified with Data Mining
- Assess can be identified with Interpretation/Evaluation.

On examining it, we see that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS Enterprise Miner software.

Comparing the KDD phases with the CRISP-DM phases is not as straightforward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP-DM methodology incorporates the steps that, as referred above, must precede and follow the KDD process that is to say:

- The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user
- The Deployment phase can be identified with the consolidation by incorporating this knowledge into the system. Concerning the remaining stages, we can say that:
- The Data Understanding phase can be identified as the combination of Selection and Pre-processing
- The Data Preparation phase can be identified with Transformation
- The Modeling phase can be identified with Data Mining
- The Evaluation phase can be identified with Interpretation/Evaluation.

In the following table, presents a summary of the presented correspondences.

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

According to KDNuggets4 polls, results are presented in the Table 1, the leading methodology for data mining process is CRISP-DM, followed by SEMMA and KDD. [7]

Poll Years	**2002**	**2004**	**2007**	**2014**
CRISP-DM	51%	42%	42%	43%
SEMMA	12%	10%	13%	8.5%
KDD process			7%	7.5%
My organization's	7%	6%	5%	3.5%
My own	23%	28%	19%	27.5%
Other (incl. domain specific)	4%	6%	9% (5%)	10% (2%)
None	4%	7%	5%	0%

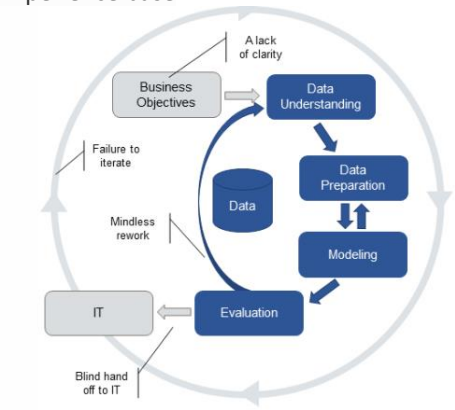
However, the usage of CRIPS-DM has reached plateau while others are steadily declining. Importantly, data scientists own methodologies usage stays above 25% rate and coupled with other ones (domain and non-domain specific) is steadily increasing reaching usage rate of over 30%. [7] This indicates decline in adoption rates of CRISP-DM and potential need for revision and modification. Indeed, this methodology though widely used was not updated since 2000 while data mining usage, methods and tools have developed exponentially.

The Analytics Solutions Unified Method for Data Mining/predictive analytics (ASUM-DM) is an extended and refined version of CRISP-DM for implementing data mining and predictive analytics projects, which was created by IBM in 2015. ASUM-DM tried to compensate the weaknesses of CRISP-DM by adding new activities,

incorporating templates and guidelines, and enhancing existing activities. There are currently two versions of ASUM-DM: an external version, which is offered in the web for free, and a proprietary version used by IBM internally [5]. This proposal is built upon the external version of ASUM-DM.

Pros of CRISP-DM:

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business & technical issues
- Framework for guidance
- Experience base



The problems with CRISP DM are as follows:

- **Lack of Clarity** - Rather than drill down into the details and really get clarity on both the business problem and exactly how an analytic might help, the project team make do with the business goals and some metrics to measure success. Now they "understand" the business objective, they want to minimize "overhead" and leap into the "interesting" bit of the project, analyzing the data. Too often this results in interesting models that don't meet a real business need.
- **Mindless Rework** - Some analytic teams simply assess their project results in analytic terms – if the model is predictive then it must be good. Most realize that this is not necessarily true and try and check their analytic results against the business objective. This is difficult without real clarity on the business problem. If the analytic they have developed does not seem to meet the business objectives, the team has few options. Most try to find new data or new modeling techniques rather than working with their business partners to re-evaluate the business problem.
- **Blind hand-off to IT** - Some analytic teams don't think about deployment and operationalization of their models at all. Most do better than that, though, recognizing that the models they build will have to be applied to live data in operational data stores or embedded in operational systems. Even these teams have typically not engaged with IT prior to this point don't have clarity on how the analytic needs to be deployed and don't really regard deployment as "analytic" work. The end result is a model that is thrown over the wall to IT. Whether the model is easy to implement or hard (or impossible) and whether it's really usable once deployed is someone else's problem. This increases the time and cost of deploying a model and contributes to the huge percentage of models that never have a business impact.
- **Failure to iterate** - Analytic professionals know that models age and that models need to be kept up to date if they are to continue to be valuable. They know that business circumstances can change and undermine the value of a model. They know that the data patterns that drove the model may change in the future. But they think of that as a problem for another day – they don't have enough clarity on the business problem to determine how to track the model's business performance nor do they invest in

thinking about to make revision of the model less work than the initial creation. After all, it's much more interesting to tackle another new problem. This leaves aging models unmonitored and unmaintained, undermining the long term value of analytics.

Each of these problems adds to the likelihood that the team will build an impressive *analytic* solution that will not add *business* value. Organizations that want to really exploit analytics – especially more advanced analytics like data mining, predictive analytics and machine learning – cannot afford these problems to persist. [6]

Not every analyst is enthusiastic about process. At one talk, a data science graduate student, raised strong objections to the recommendation to use CRISP-DM. "This is so ten years ago," he said. At another, a young man complained that the process seemed slow, and required a lot of writing from the start. People who are trained in programming and math are not always trained to care about business process and documentation.[8]

The CRISP-DM is used worldwide today, across sectors in government, non-profit and many industries. CRISP-DM is a flexible standard, makes sense for today's datasets and applications. It's mentioned in some current textbooks and some analytics tools offer special feature to support it. But there's no organization to maintain and promote use of the standard. The consortium that created the standard disbanded when EU funding ran out.[8]

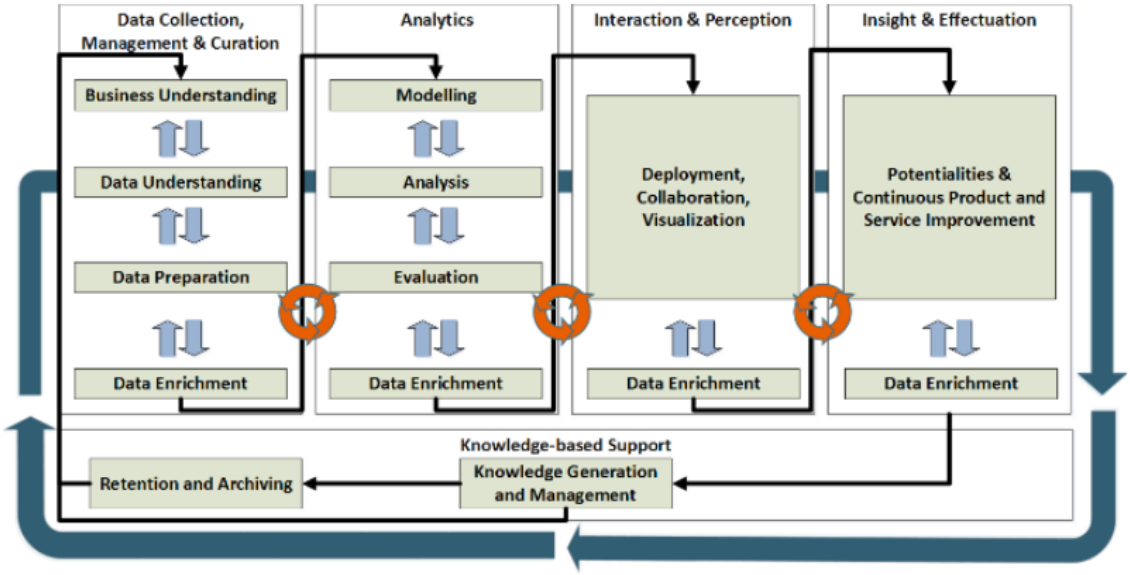
CRISP4BigData is an enhancement of CRISP-DM. The CRISP4BigData Reference Model is based on Kaufmann`s 5 phases of Big Data Management such as

- Data Collection, Management & Curation,
- Analytics,
- Interaction & Perception,
- Insight & Effectuation,
- Knowledge-based Support

and the standard four layer methodology of the CRISP-DM model.

The CRISP4BigData Methodology (based on CRISP-DM Methodology) describes the hierarchical process model, consisting of a set of tasks disposed to the four layers

- Phase,
- Generic Task,
- Specialized Task, and
- Process Instance.



The CRISP4BigData will be additionally implemented with an Apache based workflow-engine to automate and describe the whole analysis process. The description of the analysis process based on CRISP4BigData is going to be implemented with a graphical user interface.[9]

References:

[1] Shearer, The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*,(1996).

[2] Amiya, The Criticism of Data Mining Applications and Methodologies. *International Journal of Advance Research in Computer Science,*(2016)

[3] Santiago, Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects.*13th International Conference, KMO 2018, Žilina, Slovakia, (2018)*

[4] Gandomi, A., Haider, M.,Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*,(2015)

[5] Haffar, J., Have you seen ASUM-DM? (2015), *https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm*

[6] KDNuggets Homepage, *https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html*

[7] KDNuggets Homepage, *https://www.kdnuggets.com/2014/10/crisp-dm-top-methodologyanalytics-data-mining-data-science-projects.html*

[8] Forbes Homepage, *https://www.forbes.com/sites/metabrown/2016/03/31/open-standard-process-yields-best-big-data-analytics-results/#717391ff3fae*

[9] M. Bornschlegl, Towards a Cross Industry Standard Process to support Big Data Applications in Virtual Research Environments, (2017)

PART B –

Offer a critical analysis of a 'Big Data' mining problem domain and select and propose the implementation of appropriate data mining tools, and techniques to meet an organization's needs for business intelligence. Highlight the benefits to the business together with measurable implementation success criteria. Your report should be no longer than 1500 words. Cite all reference material.

Solution:

'Big Data' mining problem in Telecom industry

Indian telecom industry has crossed two decades post privatization of this sector. Since then a lot of innovation, consolidation, and maturation have happened in the industry, and today we have 12 major mobile telecom operators operating in the country. Presently, the total revenue of telecom operators is about ₹ 1.8 trillion with a burden of ₹ 2.5 trillion debt with a dwindling voice and SMS revenue. This is happening due to severe tariff competition in case of voice and declining SMS revenue due to the advent of new instant messaging applications, etc. Operators are facing disruptive technologies, rapidly changing business rules, intensified regulatory environment leading to eroding service margins.

In this scenario, the only stabilizing factor for telecom operators is revenue generated from data provisioning and driving value from this data. Telecom operators can use advanced analytics on customer and network data to generate a real-time view of customer preferences and network efficiency. This could empower them to make near real-time and fact-based decisions and hence enable a forward looking, focused, decisive, and action-oriented culture in the company. For customers, this brings faster results, predictive power, and new depth to analytics in the following aspects:

Big data promises to promote growth and increase efficiency and profitability across the entire telecom value chain. Entire telecom value chain can be benefited by leveraging the Big Data solutions and below listed are the proposed areas:

Challenges Faced by Telecom Operators

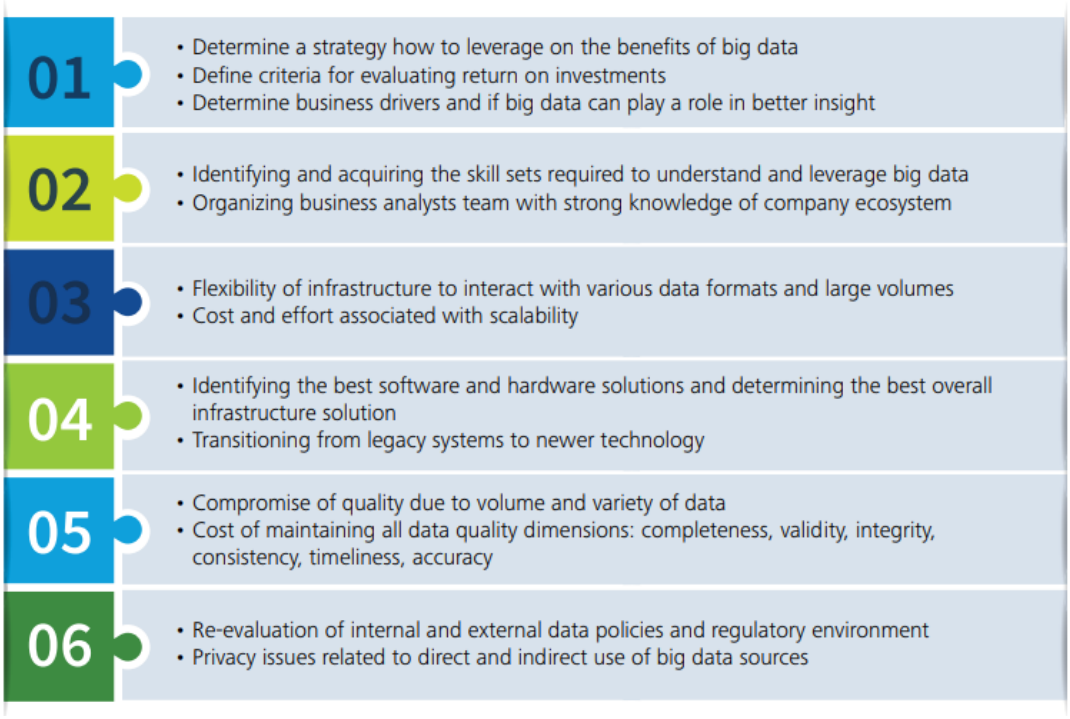
At a **macro level**, the big data analytics throws many challenges for communications services providers (CSPs) because of its variety, velocity, and complexity.

• **Variety**: The social media networks, connected devices, government portals, call data records, billing information, etc., produce huge amount of data as shown in figure above. Most of the data coming from different sources is unstructured. The telecom players need to enrich their Call Data Records (CDR) with other information like location based service, financial information, etc., to standardize the data for business intelligence platforms before the analysis can be done on it.

• **Velocity**: Every minute, Indians spend ₹ 1.85 million to shop online. Almost 100 hours of video is shared on YouTube every minute.4 The average time spent by a social media user in India is 2.5 hours a day.5 All this points to the fact that the data generation speed is tremendously high and to gain value from this data, it needs to be processed in proper timeframe. This volume of data requires new real-time operational capabilities for various functions and that in turn demands increased data storage for compliance and potential future uses as well as new tools for mediating, managing, and archiving data within available time frames.

• **Complexity**: The user generated data is mostly unstructured and complex because of the lack of standard format to store data. The legacy network and storage devices do not have any specific format to store data which can be relevant for advanced analytics. The data varies with demographics, geography, life style, etc. Analytics may provide unwanted results if the data is not filtered properly.

At a **micro level**, telecom companies faces other challenges while adopting big data for advanced analytics.

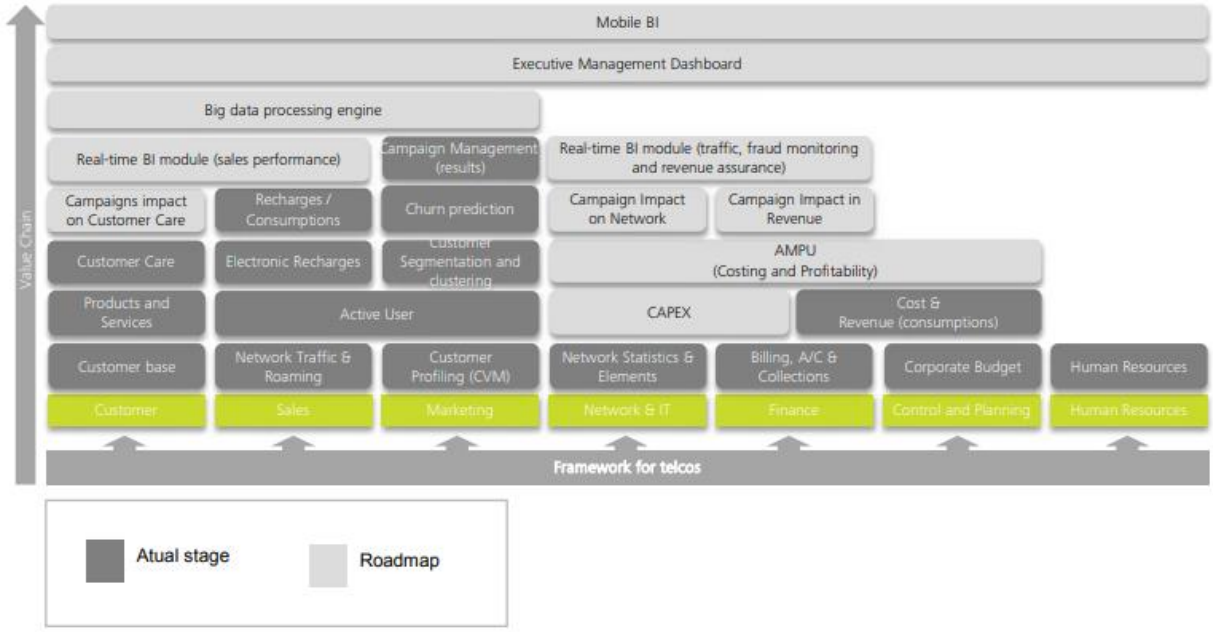


Implementation of Data Mining Tools and Techniques:

There are many data mining tools for handling large datasets. The main open source tools are scikit-learn, R, WEKA, KNIME, Orange, MOA, RapidMiner and SAMOA as shown below.

- **Scikit-learn** has emerged as one of the most popular open source machine learning toolkits, now widely used in academia and industry. Scikit-learn provides easy-to-use interfaces to perform advanced analysis and build powerful predictive models. The tutorial will cover basic concepts of machine learning, such as supervised and unsupervised learning, cross validation, and model selection. We will see how to prepare data for machine learning, and go from applying a single algorithm to building a machine learning pipeline. We will also cover how to build machine learning models on text data, and how to handle very large data sets.[5]
- **R open source programming language** is designed for statistical computing and visualization. R is the successor of S, a statistical language originally developed by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. Interface to R is command line and use through scripting. It also offers simple GUI for input. From data mining user's perspective, R offers very fast implementations of many machine learning algorithms and statistical data visualizations methods. It has specific data types for handling big data, supporting parallelization, web mining, data streams, graph mining, spatial mining, and many other advanced tasks [3]. R's main problem is its language; it is highly extendable but a difficult one to learn regarding data mining.
- **Waikato Environment for Knowledge Analysis (WEKA)** is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. Weka is open source software issued under the **GNU General Public License**.[4]
- **KNIME**, the **Konstanz Information Miner**[6], is a free and open-source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface and use of JDBC allows assembly of nodes blending different data sources, including preprocessing (ETL: Extraction, Transformation, Loading), for modeling, data analysis and visualization without, or with only minimal, programming. To some extent as advanced analytics tool KNIME can be considered as a SAS alternative. Since 2006, KNIME has been used in pharmaceutical research,[7] it also used in other areas like CRM customer data analysis, business intelligence and financial data analysis.
- **Orange** Developed by Bioinformatics Lab at University of Ljubljana, Slovenia, in collaboration with open source community. Provides data visualization and data analysis for novice and expert, through interactive workflows. Large widget toolbox and several add-ons. Possibility to use it programmatically o via GUI (Orange canvas, PyQT). Open source project (GPL license)[8]
- **Massive Online Analysis** Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm[9]
- **RapidMiner** Turbo Prep makes it easy to get data ready for predictive modeling. Interactively explore data to evaluate its health, completeness, and quality. Quickly fix common issues like missing values and outliers. Blend multiple datasets together and create new columns using a simple expression editor. When the data is finally ready, create predictive models using RapidMiner Studio and Auto Model, or just export it to popular business applications like Excel. RapidMiner Turbo Prep makes it easy to get data ready for predictive modeling**.** Interactively explore data to evaluate its health, completeness, and quality. Quickly fix common issues like missing values and outliers. Blend multiple datasets together and create new columns using a simple expression editor. When the data is finally ready, create predictive models using RapidMiner Studio and Auto Model, or just export it to popular business applications like Excel.
- **SAMOA** is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.[9]

Implementation framework – in Telecom Domain



Source: Deloitte

Scope of Big Data Implementation:

Following is the implementation roadmap that aims to create additional value for telecom operators to more actively support the financial and commercial strategies and to better understand customers through inclusion of social unstructured data using a big data engine.

1. **Call Drop Analysis**

Reasons for call drops are varied across:

- Network failures
- Credit limits/ outstanding balance
- Handover issues

2. **Network Analysis**

- Address Activation and Provisioning
- Change Management
- Inventory Management
- Performance Optimization

3. **Churn Prediction**

Retaining customers is one of the most critical challenges in the maturing mobile telecommunications service industry. Prediction of customers who are at risk of leaving a company is called as churn prediction in telecommunication.

Source: Deloitte

4. **Customer Segmentation**

The process of segmenting the market or customer base into groups that behave similarly is known as customer segmentation.[11]

Some of the benefits that telecom companies can derive through segmentation are listed below:

- Customer Value Segmentation
- Create tailored products for customers
- Identifying high-value and long-term customers
- Identify potential customers

5. **Predictive Campaign**

- Renewed focus on the individual
- Shift to omni-directional
- Event-driven communication

6. **Location-Based Services**

Geographical location data related to mobile devices provide great insights of real-time information. These data sets can be utilized for various analytical services and representations.

References:

[1] https://www.mckinsey.com/industries/telecommunications/our-insights/telcos-the-untapped-promise-of-big-data

[2] https://www.reuters.com/article/us-mobile-world-bigdata/telecom-firms-mine-for-gold-in-big-data-despite-privacy-concerns-idUSBREA1M09F20140223

[3] A. Jovic, K. Brkic and N. Bogunovic, "An overview of free software tools for general data mining", 37th International Conventiion on Information & Communication Technology Electronics & Microelectronics., 2014, PP. 1112-1117.

[4] https://www.cs.waikato.ac.nz/ml/weka/

[5] https://conferences.oreilly.com/strata/big-data-conference-sg-2015/public/schedule/detail/46058

[6] Berthold, Michael R.; Cebron, Nicolas; Dill, Fabian; Gabriel, Thomas R.; Kötter, Tobias; Meinl, Thorsten; Ohl, Peter; Thiel, Kilian; Wiswedel, Bernd (16 November 2009). "KNIME - the Konstanz information miner". ACM SIGKDD Explorations Newsletter. **11** (1): 26. *doi:10.1145/1656274.1656280*.

[7] *Tiwari, Abhishek; Sekhar, Arvind K.T. (October 2007). "Workflow based framework for life science informatics". Computational Biology and Chemistry.* ***31*** *(5–6): 305–319. doi:10.1016/j.compbiolchem.2007.08.009*

[8] https://www.pycon.it/media/conference/slides/introduzione-a-orange-data-mining.pdf

[9] http://albertbifet.com/big-data-mining-tools/

[10] https://rapidminer.com/

[11] Process of segmenting. http://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technologymedia-telecommunications/deloitte-uk-a%20people-based-telecom-business%20-%20Breathing-new-life-intosegmentation-strategies.pdf